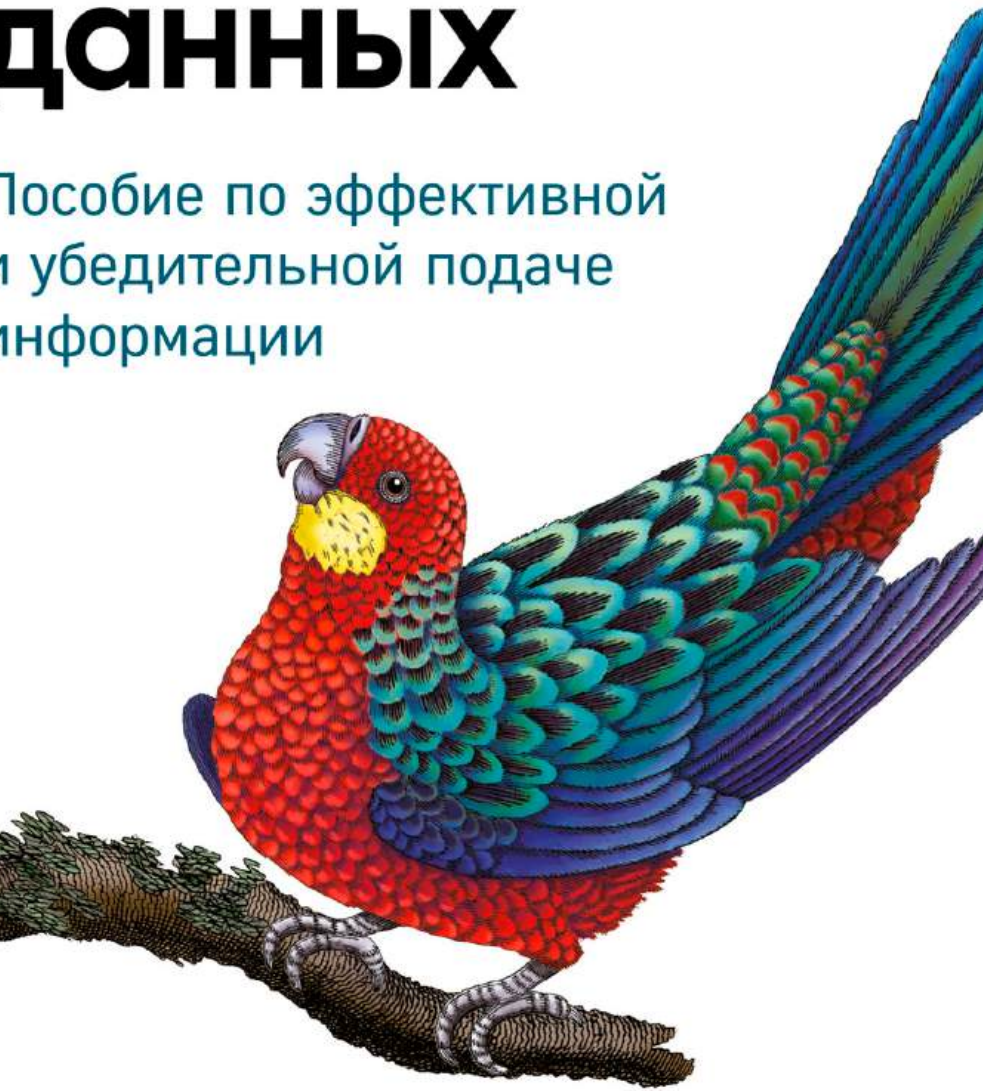
O'REILLY®

Основы визуализации данных

Пособие по эффективной и убедительной подаче информации



Клаус Уилк

Библиотека цифровой трансформации

O'REILLY®

Claus Wilke

Fundamentals of Data Visualization

A Primer on Making Informative and Compelling Figures

O'REILLY®

Клаус Уилке

Основы визуализации данных

Пособие по эффективной и убедительной подаче информации

БОМБОРА
ИЗДАТЕЛЬСТВО
Москва 2024

УДК 004.4
ББК 32.973.2-018.2
У36

FUNDAMENTALS OF DATA VISUALIZATION
Claus O. Wilke

Authorized Russian translation of the English edition of Fundamentals of Data Visualization ISBN 9781492031086

Уилке, Клаус.

У36 Основы визуализации данных : пособие по эффективной и убедительной подаче информации / Клаус Уилке ; [перевод с английского М.А. Райтмана]. — Москва : Эксмо, 2024. — 352 с. — (Библиотека цифровой трансформации).

ISBN 978-5-04-106457-0

Без визуализации информации сейчас не обходится ни один бизнес. Итоги продаж, месячные отчеты, презентации новых проектов — все это примеры того, что важно и нужно правильно визуализировать. Благодаря этой книге вы научитесь делать различные форматы визуализации для своих презентаций, собирать собственный инструментарий для построения графиков и уделять внимание мелким деталям, зачастую упускаемым другими людьми.

УДК 004.4
ББК 32.973.2-018.2

ISBN 978-5-04-106457-0

Оглавление

ЧАСТЬ II
ПРИНЦИПЫ ДИЗАЙНА ВИЗУАЛИЗАЦИЙ

От научного редактора русского издания

Перед вами прекрасная книга, которая в сжатой и очень доступной форме рассказывает о том, как можно подавать и визуализировать самые разноплановые данные. Книга хороша тем, что написана профессионалом своего дела и человеком, который не просто собрал несколько статей, а по-настоящему уже много лет использует визуализацию данных в своей работе.

Оригинал книги написан на английском языке и содержит очень большое количество названий различных диаграмм, графиков и других форм визуализации, у части которых нет эквивалентов в русском языке, а часть может в различных изданиях и приложениях быть переведена по-другому.

Мы старались пользоваться как можно более широкой базой для того, чтобы максимально качественно перевести все названия, но заранее приносим извинения, если вы привыкли к другим названиям форм представления и визуализации данных, нежели к тому, как они переведены в этой книге.

Андрей Бояршинов,
Director of Production, General Arcade

Предисловие

Если вы ученый, аналитик, консультант или любой другой специалист, чьи обязанности включают в себя подготовку технических документов или отчетов, то вы не понаслышке знаете, как важно уметь убедительно визуализировать данные в формате изображений. Чем нагляднее графики, тем весомее смотрятся аргументы. Изображения должны быть ясными, привлекательными и убедительными. Разница между хорошим и плохим графиком подобна разнице между влиятельной и малоизвестной газетой, выигранными и упущенными грантами или контрактами, удачным и провальным собеседованием. При этом, несмотря на, казалось бы, очевидную востребованность методических материалов по этой теме, существует на удивление мало ресурсов, которые рассказывают о том, как качественно и красиво иллюстрировать данные. Высших учебных заведений, которые предлагают курсы по этой теме, совсем немного, да и специализированную литературу придется поискать. (Но что-то, разумеется, есть.) Учебные материалы, посвященные программным средствам построения графиков, обычно рассказывают лишь о создании некоторых визуальных эффектов, однако объяснений, почему следует выбрать тот или иной вариант, не дается. При этом в повседневной рабочей рутине предполагается, что вы знаете, как создавать хорошие графики. А если вам повезет, у вас появится терпеливый научный руководитель, который научит нескольким приемам визуализации при написании ваших первых научных статей.

Что касается писательской работы, опытные редакторы говорят о так называемом «слухе» — способности слышать (внутренне, когда вы читаете отрывок прозы), хорош ли этот текст. Думаю, что для графиков и других визуализаций нам тоже нужно «зрение», то есть способность смотреть на картинку и видеть, является ли этот график сбалансированным, ясным и убедительным. Как и в случае с текстом, умению отличать эффективный график от нерабочего можно научиться. Наличие «зрения» — это прежде всего знакомство с более широким спектром простых законов и принципов хорошей визуализации, а также внимание к мелким деталям, зачастую упускаемым другими людьми.

По своему опыту я знаю, что, как и в случае с текстом, невозможно развить «зрение», всего лишь прочитав одну книжку на выходных. Это длительный процесс, и он будет с вами всю вашу жизнь, и может случиться так, что концепции, которые сейчас вам кажутся слишком сложными или, наоборот, малозначимыми, через пять лет станут для вас куда более существенными. Что касается меня, то я и сегодня продолжаю совершенствоваться в искусстве создания графиков. Я постоянно ищу новые подходы и обращаю внимание на визуальные и дизайнерские решения других людей. Не исключаю, что буду менять свое мнение, если появятся убедительные аргументы в пользу иной точки зрения. Сегодня я могу считать один график отличным, а уже через месяц у меня появится повод его покритиковать. Помните об этом, читая книгу, и не принимайте все мои слова за истину в последней инстанции. Будьте критичны к моим аргументам в пользу определенных решений и выбирайте сами, принимать их или нет.

Материал книги выстроен в логической последовательности, однако большинство глав можно читать как самостоятельный текст: штудировать книгу от корки до корки вовсе не обязательно. Не стесняйтесь пропускать шаги, чтобы выбрать наиболее интересный для вас в данный момент раздел или тот, который посвящен тому типу дизайна, над которым вы сейчас размышляете. Более того, думаю, что вы извлечете из этой книги максимум пользы только в том случае, если будете читать ее не всю целиком за раз, а по частям и в течение длительного периода времени. Попробуйте применить парочку-другую концепций из книги, а потом вернитесь к ней, чтобы узнать о других принципах или освежить в памяти уже изученные разделы. Вполне возможно, что одна и та же глава предстанет перед вами совсем в другом свете, если вы вновь вернетесь к ней через несколько месяцев.

Несмотря на то, что почти все рисунки в этой книге сделаны с помощью R и ggplot2, я не считаю эту книгу справочником по созданию графиков при помощи R. Моя цель — рассказать об общих принципах создания графиков. Выбор программного обеспечения, используемого для создания изображений, зависит от ваших конкретных потребностей. Если вы захотите воспроизвести визуализации, приведенные в этой книге, то можете использовать любое ПО для построения графиков. Тем не менее хочу отметить, что многое из того, что я демонстрирую, сделать посредством ggplot2 и аналогичных пакетов будет гораздо проще, чем с помощью других библиотек для построения графиков. Важно отметить, что, поскольку данная книга не справочник по созданию графиков при помощи R, вы не найдете здесь обсуждений кода или методов программирования. Я хочу, чтобы вы сосредоточились на концепциях и графиках, а не на их реализациях. Если вам интересно, как были сделаны те или иные рисунки, обратитесь к исходному коду по ссылке на с. 15.

Мнение автора о ПО и процессах для построения графиков

За моими плечами больше двух десятилетий опыта подготовки графиков для научных публикаций, я являюсь автором тысяч рисунков. Если в течение этого времени что-то и оставалось незыблемым, так это постоянные изменения рабочего процесса их подготовки. Каждые несколько лет появляется новая библиотека для построения графиков или даже новая парадигма, после чего огромное количество ученых переключается на более актуальный инструментарий. Я делал рисунки с помощью gnuplot, Xfig, Mathematica, Matlab, matplotlib в Python, base R, ggplot2 в R и, возможно, других инструментов, названия которых я уже и не вспомню. В настоящее время я предпочитаю подход ggplot2 в R, но не ожидаю, что он дотянет до моей пенсии.

Постоянное изменение программных платформ — одна из основных причин того, что данная книга не является учебником по программированию и что в ней нет ни единого примера кода. Я хочу, чтобы эта книга была полезной вне зависимости от того, какое ПО вы используете, и хочу, чтобы она оставалась таковой даже после того, как все уйдут с ggplot2 и перейдут к следующему поколению программ визуализации данных. Я понимаю, что выбранный мною подход наверняка расстроит некоторых пользователей ggplot2, которые хотели бы знать, как я сделал тот или иной рисунок, и поэтому те, кому интересно узнать о моих методах программирования, могут обратиться к исходному коду книги: он находится в открытом доступе. Кроме того, вероятно, в будущем появится дополнительный материал, посвященный исключительно вопросам программирования.

За прошедшие годы я понял одну вещь: автоматизация — ваш друг. Мое мнение — графики должны генерироваться автоматически как часть процесса анализа данных (который также должен быть автоматизирован), и они должны в результате создаваться уже готовыми к отправке на принтер, без необходимости ручной доработки. Я часто вижу, как стажеры сначала делают грубый набросок будущего графика, а затем импортируют его в Illustrator, чтобы привести в порядок. Эта идея плоха в силу нескольких причин. Во-первых, если вы вручную редактируете график, конечный результат становится невоспроизводимым. Никто другой не сможет сгенерировать график, идентичный созданному вами. Даже если вы всего лишь изменили шрифт подписей к засечкам осей, линии могут получиться размытыми, и уже одно это может нанести ущерб информативности изображения. Например, вы решили вручную заменить некоторые непонятные метки на более читаемые — другой человек вряд ли сможет проверить правильность этой замены. Во-вторых, если

вы добавите много ручной постобработки в процесс подготовки графиков, вы будете менее охотно вносить какие-либо изменения в свою работу или переделывать ее. В результате может случиться так, что вы просто проигнорируете разумные предложения ваших соавторов или коллег об изменении графиков или у вас может возникнуть соблазн повторно использовать старую визуализацию, даже если данные уже обновились. В-третьих, вы можете попросту забыть, что именно вы делали, и не сможете создать аналогичный график, но с другим наполнением. Все эти примеры взяты из жизни: я лично видел, как такое происходило с реальными людьми и настоящими публикациями.

Поэтому использование интерактивных программ по созданию графиков — плохая идея. По сути, они заставляют вас создавать графики вручную. На самом деле, вероятно, лучше автоматически создать эскиз фигуры и украсить его в Illustrator, чем создавать всю фигуру вручную в какой-нибудь интерактивной программе. Имейте в виду, что Excel тоже является интерактивной программой построения графиков и не рекомендуется для подготовки рисунков (или анализа данных).

Одним из важнейших компонентов книги по визуализации данных является возможность воспроизведения предлагаемых графиков. Приятно изобрести какой-нибудь новый изящный тип визуализации, но, если воссоздать график, используя ваш метод, невозможно, пользы от вашей идеи будет немного. Например, когда Эдвард Тафти предложил так называемые спарклайны, то первое время никто толком не понимал, как их делать. И хотя нам безусловно нужны гении, которые заставляют мир двигаться вперед, расширяя границы возможного, эта книга посвящена практике и может использоваться в повседневной работе специалистов по данным, готовящих графики для своих публикаций. Поэтому предлагаемые мной визуализации могут быть созданы с помощью нескольких строк кода R, ggplot2 и легкодоступных пакетов расширений. Почти все изображения в этой книге, за исключением тех, что находятся в главах 26–28, были автоматически сгенерированы именно в том виде, в каком они здесь представлены.

Условные обозначения

В этой книге используются следующие условные обозначения.

Курсивный шрифт

Обозначает новые термины, имена файлов и их расширения.

`Моноширинный шрифт`

Используется для обозначения элементов кода, таких как имена переменных или функций, операторов и ключевых слов.

Совет или подсказка

Общее замечание

Предостережение

Использование примеров кода

Дополнительные материалы можно скачать по следующему адресу: *https://addons.eksmo.ru/it/Data_Visualization.zip*.

Цель этого руководства — помочь вам в решении ваших задач. В работе над своими программами или документацией вы можете пользоваться фрагментами кода из данной книги.

Благодарности

Появление этого проекта на свет было бы невозможно без той фантастической работы, которую проделала команда RStudio, превратив вселенную R в первоклассную платформу для подготовки оригинал-макетов. В частности, я должен поблагодарить Хэдли Уикхэма за создание ggplot2 — программы для построения графиков, — которая использовалась для создания всех изображений в этой книге. Я также хотел бы поблагодарить Се Ихуэй за создание R Markdown и за написание пакетов `knitr` и `bookdown`. Не думаю, что я решился бы взяться за настоящий проект, не будь у меня под рукой этих инструментов. Писать R Markdown-файлы — одно удовольствие, собирать материал и наращивать темп работы — легче легкого. Особую благодарность я хочу выразить Ахиму Зейлеису и Рето Штауфферу за `colorspace`, Томасу Лину Педерсену за `ggforce` и `gganimate`, Камилу Словиковски за `ggrepel`, Эдзеру Пебесме за `sf` и Клэр Маквайт за ее работу над пакетами `colorspace` и `colorblindr` для имитации дальтонизма при просмотре готовых иллюстраций.

Отдельно хочу поблагодарить людей, которые предоставили полезные отзывы о черновых версиях этой книги. Наибольший вклад внесли Майк Лукидес, мой редактор в O'Reilly, и Стив Хароз — они прочитали и прокомментировали каждую главу. Я также получил полезные комментарии от Карла Бергстрома, Джессики Халлман, Мэтью Кея, Тристана Мара, Эдзера Пебесмы, Джона Швабиша и Хэдли Уикхэма. Блог Лена Кифера, а также книги и сообщения Кирана Хили послужили источником вдохновения для создания графиков и наборов данных для использования. Ряд людей указали на незначительные проблемы или опечатки: Тьяго Аррайс, Малкольм Барретт, Джессика Бернетт, Джон Колдер, Антонио Педро Камарго, Дарен Кард, Ким Крессман, Акос Хайду, Томас Йохманн, Эндрю Кинсман, Уилл Керсен, Алекс Лаледжини, Джон Лидли, Катрин Лайнвебер, Микель Мадина, Клэр Маквайт, С'бусисо Мхондване, Хосе Назарио, Стив Путман, Маэль Салмон, Кристиан Шудома, Джеймс Скотт-Браун, Энрико Спиниелли, Воутер ван дер Бейл и Рон Юрко.

Я также хотел бы поблагодарить всех остальных участников tidyverse и R-сообщества в целом. Действительно, для любой задачи по визуализации, с которой вы можете столкнуться, существует R-пакет. Все эти приложения были разработаны силами большого сообщества, состоящего из тысяч специалистов по обработке данных и статистике, и многие из них в той или иной форме внесли свой вклад в создание этой книги.

Наконец, я хотел бы поблагодарить мою жену Стефанию за ее терпение в течение несметного количества вечеров и выходных, когда я часами сидел перед компьютером, писал код `ggplot2` и был полностью погружен в мельчайшие детали графиков и проработку деталей глав.

Введение

Визуализация данных — это отчасти наука, а отчасти искусство. Самое сложное в этом деле — сделать так, чтобы искусство получилось хорошим, при этом не переврав науку, и наоборот. Визуализация данных — это прежде всего точная передача информации. Недопустимо даже малейшее искажение данных. Если вы заметили, что одно число в два раза больше другого, но при этом на схеме соответствующие этим числам элементы имеют одинаковый размер, знайте: вся работа по визуализации пошла под откос. В то же время нельзя забывать, что визуализация должна быть приятна глазу. Качественное визуальное преподнесение данных обычно повышает их информативность. Если у зрителя рябит в глазах от ярких цветов, элементы графика несбалансированы или на нем множество отвлекающих внимание объектов, зрителю будет сложнее рассмотреть изображение и верно считать его смысл.

Исходя из своего опыта, хочу отметить, что в большинстве случаев (но не всегда!) ученые умеют визуализировать данные, не вводя читателей в заблуждение. К сожалению, они не всегда обладают развитым художественным вкусом и периодически используют вещи, сильно отвлекающие внимание от передаваемой информации. Дизайнеры, в свою очередь, делают все невероятно красиво, но при этом крайне небрежно относятся к самим данным. Посыл моей книги состоит в том, чтобы донести до каждой из этих групп полезную для них информацию.

В этой книге я концентрируюсь на базовых принципах, методах и концепциях, которые применяются при иллюстрации публикаций, докладов или презентаций. Поскольку визуализация данных — обширное поле, которое в широчайшем толковании может включать в себя такие разнообразные вещи, как схемотехника, 3D-анимации и пользовательские интерфейсы, я вынужден сузить спектр. Данная книга посвящена исключительно способам визуализации, представленным в печатном виде, онлайн или в виде слайдов. Здесь вы не встретите информацию об интерактивных визуальных элементах или видео, кроме как в небольшом разделе главы 25. Поэтому в рамках данной книги я буду достаточно вольно использовать слова «визуализация» и «изображение», имея в виду одно и то же. Помимо перечисленного, в этой книге не идет речь о том, *как* создавать изображения при помощи существующих

программ для визуализации и программных библиотек. Библиография в конце книги содержит ссылки на подходящие для изучения этих тем статьи.

Данная книга состоит из трех частей. Первая — «От данных до визуализации» — описывает различные виды графиков и диаграмм, такие как гистограммы, диаграммы рассеяния и круговые диаграммы. Основной упор в этой части делается на научный аспект визуализации. Однако вместо того, чтобы писать обширную энциклопедию со статьями, посвященными каждому мыслимому подходу к визуализации, я расскажу вам о тех способах визуализации, которые вы наверняка встретите в публикациях или будете использовать в своих работах. При написании данной части я постарался сгруппировать подходы к визуализации по тому, какой посыл они несут зрителям, а не по типам используемых данных. Учебники статистики обычно описывают анализ и визуализацию данных в привязке к типам данных и организуют материал по количеству и типу переменных (одна непрерывная переменная, дискретная переменная, две непрерывные переменные, одна непрерывная и одна дискретная переменная и т. д.). Мое мнение на этот счет состоит в том, что найти в таких текстах что-то полезное смогут только ученые-статистики. Большинство людей воспринимает данные через призму информативной составляющей сообщения: например, насколько что-то велико или мало, каковы его составные части, как оно соотносится с чем-то другим и т. д.

Вторая часть, «Принципы дизайна визуализаций», посвящена вопросам дизайна, которые зачастую возникают в процессе создания схем или диаграмм. Основная, но не единственная тема, поднимаемая в данной главе, — эстетический аспект визуализации данных. После того как мы выбрали подходящий для наших данных график или диаграмму, перед нами встает проблема выбора визуальных элементов: цвет, символы, размеры и гарнитуры шрифтов. От выбора этих элементов зависит то, насколько наш способ визуализации будет понятен и эстетичен. Из второй части вы узнаете о наиболее частых проблемах, с которыми мне приходится сталкиваться во время непосредственной работы над визуализацией.

Третья часть, «Разное», посвящена темам, которые не вошли ни в первую, ни во вторую часть. Сюда относятся, например, форматы файлов, в которых обычно хранятся рисунки и графики, а также подсказки по выбору ПО. Кроме того, эта глава рассказывает о том, как правильно встраивать диаграммы в большие документы с учетом контекста.

Некрасивые, плохие и ложные изображения

На протяжении этой книги вам будут встречаться разные варианты одних и тех же изображений. Часть из них будет использоваться в качестве примеров того, как надо и не надо делать. Чтобы вам было понятнее, какие примеры

стоит взять на заметку, а каких лучше избегать, я разделил неудачные варианты на три категории: «некрасиво», «плохо» и «ложно» (рис. В.1).

Некрасиво

Изображение, в котором есть эстетические проблемы, но при этом оно все же является понятным и информативным.

Плохо

Изображение, в котором есть проблемы, связанные с восприятием информации. Такое изображение может быть непонятным, сбивающим с толку, чересчур сложным или обманчивым.

Ложно

Изображение, в котором есть математический изъян и которое объективно неверно.

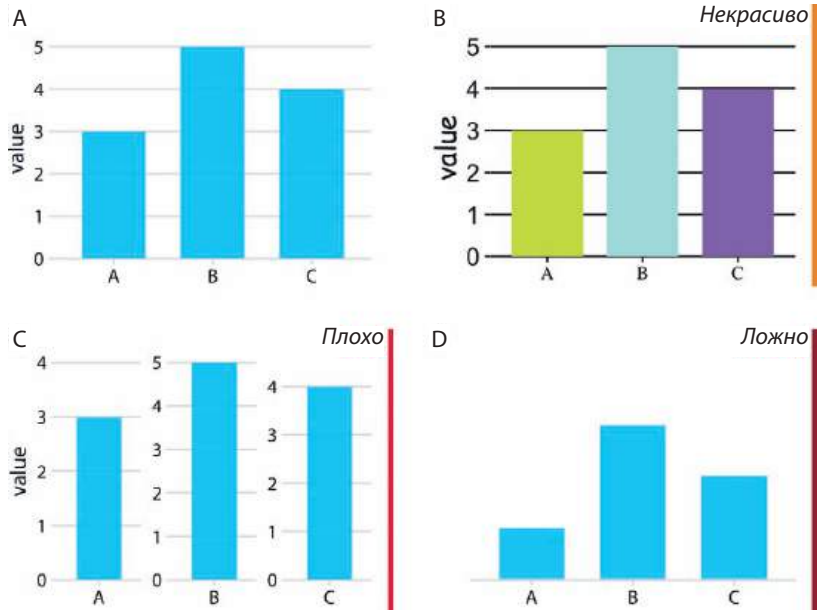


Рис. В.1. Примеры некрасивых, плохих и ложных изображений. A. Столбиковая диаграмма, на которой указаны три значения (A = 3; B = 5; C = 4). Подобный способ являет собой пример грамотной визуализации без существенных изъянов. B. Некрасивая версия графика A. Хотя с точки зрения информативности график является корректным, эстетически он менее удачен. Цвета слишком яркие и не несут никакого смысла. Линии на заднем плане отвлекают внимание. Текст набран тремя разными шрифтами и в трех разных размерах. C. Плохая версия графика A. Масштаб оси ординат меняется в зависимости от столбца. Поскольку каждый из столбцов имеет свой масштаб, понять график становится сложно. Складывается впечатление, что значения на графике ближе друг к другу, чем они есть на самом деле. D. Ложная версия графика A. Так как на графике нет четко выделенной оси ординат, невозможно установить точные числовые соответствия столбцов. Кажется, что значения столбцов равны 1, 3 и 2, хотя на самом деле они — 3, 5 и 4

В этой книге я не буду обращать особое внимание на «хорошие» изображения. Все изображения, не помеченные как «неудачные», должны считаться как минимум приемлемыми. Такие изображения информативны, выглядят привлекательно и могут быть напечатаны как есть. Однако не стоит забывать, что даже среди хороших вариантов бывают те, которые различаются по качеству выполнения, поэтому одни хорошие изображения могут быть лучше, чем другие, тоже хорошие.

В большинстве случаев я буду объяснять, почему то или иное изображение неудачно, но некоторые мои аргументы являются вкусовщиной. Честно говоря, понятие «некрасивых» изображений более субъективно, нежели понятие «плохих» или «некорректных» графиков. И вообще, граница между «некрасивыми» и «плохими» изображениями довольно размыта. В ряде случаев неудачные дизайнерские решения могут настолько сильно мешать восприятию информации, что изображение будет относиться скорее к категории «плохих», нежели просто «некрасивых». В любом случае я призываю вас, дорогие читатели, вырабатывать свою позицию по этому вопросу, а также развивать собственное видение, чтобы критически оценивать мою точку зрения.

Часть I

От данных до визуализации

Глава 1

Визуализация данных: соответствие данных и эстетики

Мы говорим, что занимаемся визуализацией данных, когда мы берем набор значений и, пользуясь принципами системности и логики, делаем из них графические элементы будущего окончательного изображения.

И хотя существует огромное количество типов визуализации данных, а различные графики, такие как диаграммы рассеяния, круговые диаграммы и тепловые карты, на первый взгляд, имеют между собой мало общего, их все можно описать общим языком, определяющим то, каким образом значения данных могут быть преобразованы в капли чернил на бумаге или цветные пиксели на экране. Главный вывод таков: любая визуализация преобразует имеющиеся данные и значения в отдельные, измеримые элементы готового изображения. В книге я называю эти элементы *эстетикой.*

Эстетика и типы данных

Эстетика описывает каждый аспект любого графического элемента. Несколько примеров можно увидеть на рис. 1.1. Ключевым компонентом любого графического элемента является его *положение,* которое говорит о том, где он находится. В стандартной двухмерной графике мы определяем его относительно осей *x* и *y*, однако возможны одно- и трехмерные визуализации, равно как и визуализации в других системах координат. Далее, нельзя не отметить тот факт, что все графические элементы имеют *форму, размер и цвет.* Даже если мы создаем черно-белое изображение, у графических элементов все равно будет цвет: черный, если фон белый, и наоборот. И наконец, для создания визуализации используются линии. Они могут различаться толщиной или состоять из различных последовательностей точек и тире. Помимо примеров, приведенных на рис. 1.1, существует множество других эстетических решений визуализации. Например, если на нашем изображении должен быть текст, нам придется выбрать гарнитуру шрифта, его начертание

и размер, а также в случае, если графические объекты накладываются друг на друга, нам может потребоваться задать некоторым из них показатель прозрачности.

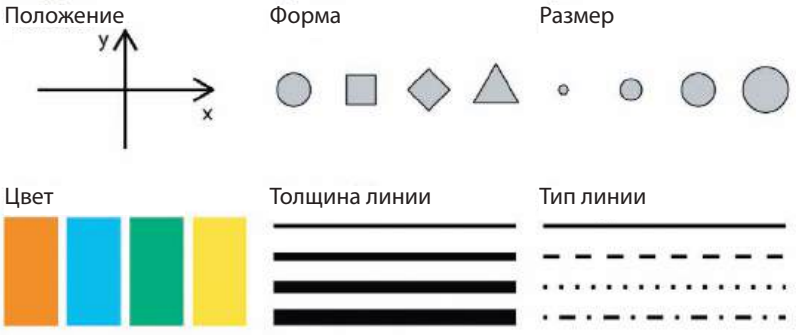


Рис. 1.1. Наиболее распространенные эстетические элементы в визуализации данных: позиция, форма, размер, цвет, ширина линии и тип линии. Некоторые из них могут представлять как дискретные, так и непрерывные данные (положение, размер, ширина линии, цвет), остальные же — только дискретные (форма, тип линии)

Эстетические решения можно разбить на две группы: те, которые могут представлять непрерывные данные, и те, которые не могут. Непрерывные значения — это те, для которых существуют сколь угодно точные промежуточные элементы. К примеру, протяженность во времени можно считать непрерывной величиной. Между двумя значениями, например 50 секундами и 51 секундой, существуют такие промежуточные элементы, как 50,5 секунды, 50,51 секунды, 50,50001 секунды и т. д. Для сравнения, количество человек в комнате — дискретная величина. Землекопов может быть один, два или три, но никак не полтора. Если говорить об элементах, показанных на рис. 1.1, то положение, размер, цвет и толщина линии могут использоваться для отображения непрерывных данных, а форма и тип линии — как правило, только для дискретных.

Далее мы рассмотрим типы данных, которые мы можем захотеть визуализировать. Вы можете думать о данных как о числах, однако числовые значения — это лишь два из нескольких типов данных, которые нам встречаются в жизни. В дополнение к непрерывным и дискретным числовым значениям данные могут быть выражены в виде дискретных категорий, таких как дата или время, или в форме текста (табл. 1.1). Когда данные представлены в виде числовых значений, мы также называем их *количественными*, а в противном случае — *качественными*. Переменные, содержащие количественные данные, называются *факторами*, а различные категории — *уровнями*. Уровни фактора обычно являются неупорядоченными (например, *собака, кот, рыбка*

в табл. 1.1), но, если уровни фактора содержат некую внутреннюю градацию* (например, *хороший, приемлемый, плохой* в табл. 1.1), они считаются упорядоченными.

Таблица 1.1. Типы переменных, используемые в стандартных сценариях визуализации данных

Тип переменной	Примеры	Тип данных	Описание
Количественная / непрерывная числовая	1,3, 5,7, 83, $1{,}5 \times 10^{-2}$	Непрерывный	Произвольные числовые значения. Могут быть целыми, рациональными или действительными числами
Количественная / дискретная числовая	1, 2, 3, 4	Дискретный	Числа в виде дискретных единиц. Как правило, это целые числа, однако бывают и исключения. Примером могут служить числа 0,5, 1,0, 1,5, которые будут считаться дискретными, если в выбранном нами наборе данных между этими величинами нет промежуточных значений
Качественная / категориальная неупорядоченная	собака, кот, рыбка	Дискретный	Категории без определенного порядка. Это уникальные дискретные категории, не имеющие какого-либо внутреннего ранжирования. Такие переменные также называются *факторами*
Качественная переменная / категориальная упорядоченная переменная	хороший, приемлемый, плохой	Дискретный	Категории, имеющие порядок. Это дискретные и уникальные категории, обладающие определенными правилами ранжирования. К примеру, «приемлемый» всегда будет находиться между «хороший» и «плохой». Подобные переменные называются *упорядоченными факторами*
Дата или время	5 января 2018, 8:03	Непрерывный или дискретный	Конкретный день или время. Это также относится к более общей записи дат, как, например, 4 июля или 25 декабря (без указания года)
«Текст»	«Съешь еще этих мягких французских булок, да выпей же чаю»	Отсутствует или дискретная	Текст, написанный в свободной форме. Может считаться категориальным в случае надобности

Для того чтобы увидеть конкретные примеры этих различных типов данных, обратите внимание на табл. 1.2. В таблице приведены первые несколько строк массива данных, описывающего среднюю дневную температуру (среднесуточную температуру на временном промежутке в 30 лет), зафиксированную на четырех погодных станциях США. Таблица содержит значения пяти видов

* Тут речь о свойственной определенной группе градации — естественно присущей этой группе.

переменных: месяц, день, место, ID станции и температура (по шкале Фаренгейта). Здесь месяц является упорядоченным фактором, день — дискретной числовой переменной, место — неупорядоченным фактором, ID станции — также неупорядоченный фактор, а температура — непрерывная числовая переменная.

Таблица 1.2. Первые восемь строк массива данных, описывающего среднюю дневную температуру для четырех метеостанций. Источник: National Oceanic and Atmospheric Administration (NOAA, Национальное управление океанических и атмосферных исследований США)

Месяц	День	Место	ID станции	Температура (в градусах Фаренгейта)
Январь	1	Чикаго	USW00014819	25,6
Январь	1	Сан-Диего	USW00093107	55,2
Январь	1	Хьюстон	USW00012918	53,9
Январь	1	Долина Смерти	USC00042319	51,0
Январь	2	Чикаго	USW00014819	25,5
Январь	2	Сан-Диего	USW00093107	55,3
Январь	2	Хьюстон	USW00012918	53,8
Январь	2	Долина Смерти	USC00042319	51,2

Использование шкал для отображения данных на эстетические элементы

Для того чтобы представить значения данных в виде графических элементов, необходимо задать между ними соответствие. Например, если у нашего графика есть ось *x*, то нам нужно правило, описывающее, какие значения соответствуют каким точкам на оси. Аналогичным образом следует указывать, какие значения соответствуют тем или иным цветам или формам. Подобное отображение значений данных на эстетические элементы достигается при помощи *шкал*. Шкала устанавливает уникальное правило отображения данных на нашу эстетику (рис. 1.2). Обратите внимание: шкала непременно должна задавать взаимно однозначное соответствие (чтобы каждому значению данных соответствовал ровно один эстетический элемент и наоборот). Если однозначность соответствия нарушается, визуализация становится непонятной.

Давайте применим полученные знания на практике. В качестве примера возьмем массив данных, часть которого приведена в табл. 1.2, и разместим температуру по оси *y*, дни — по оси *x*, а место покажем цветом. В качестве визуальной составляющей эстетики выберем сплошные линии. В результате у нас получится стандартная линейная диаграмма, показывающая изменение средней дневной температуры в четырех городах (рис. 1.3).

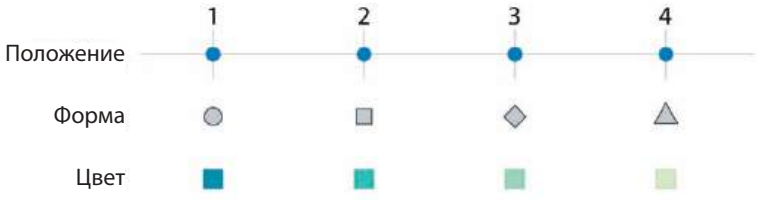


Рис. 1.2. Шкалы связывают значения данных и эстетические элементы. В данном примере цифры от 1 до 4 отображаются на шкалы положения, формы и цвета. На каждой шкале каждой цифре соответствует своя уникальная характеристика (положение, форма или цвет), и наоборот

Рис. 1.3 — это наиболее типичный пример визуализации температурной кривой. Именно такой способ большинство экспертов по работе с данными используют в первую очередь. Однако то, какие именно данные и на какие шкалы мы отображаем, зависит целиком и полностью от нас. К примеру, вместо того чтобы разместить температуру на оси *y*, а место показать цветом, мы можем сделать все наоборот. Если мы выделим наиболее интересную для нас переменную (температура) цветом, количество цветов на нашей диаграмме резко возрастет: так делают, чтобы как можно более точно передать полезную информацию [Stone, Albers Szafir, and Setlur, 2014]. Поэтому для нашей визуализации я заменил линии квадратами и раскрасил их сообразно значениям средней температуры для каждого месяца (рис. 1.4).

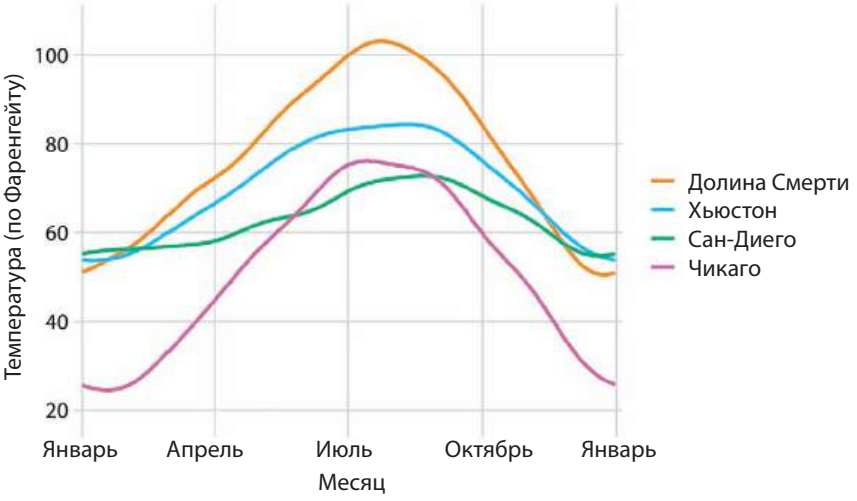


Рис. 1.3. Средняя дневная температура в четырех точках наблюдения в США. Значения температуры отложены по оси *y*, дни — по оси *x*, а цветами обозначены точки наблюдения. Источник: NOAA

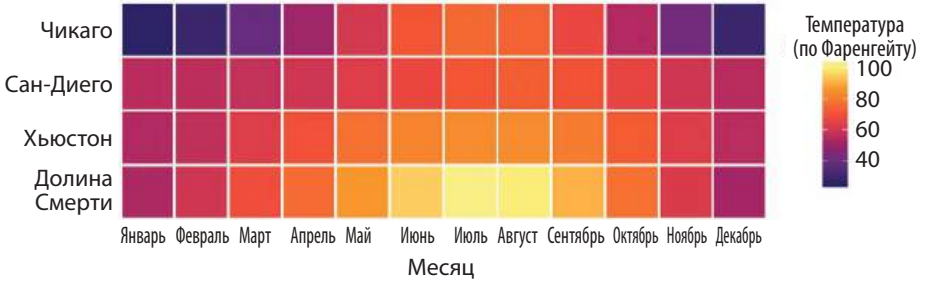


Рис. 1.4. Среднемесячные температуры в четырех точках наблюдения в США. Источник: NOAA

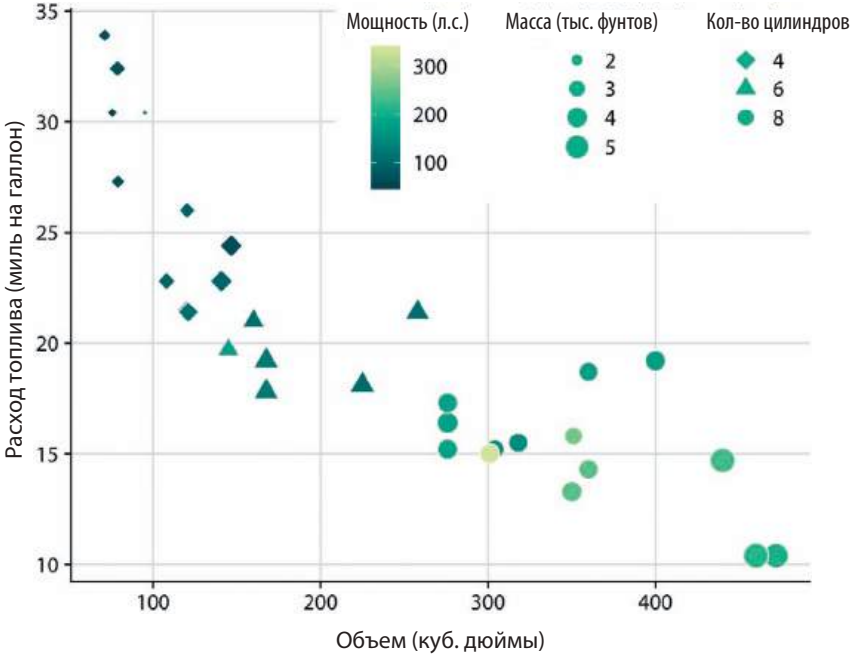


Рис. 1.5. Отношение топливной экономичности к объему двигателя для 32 моделей автомобилей (периода 1973–1974 годов). На данном изображении используется пять различных шкал: 1) ось *x* (объем двигателя); 2) ось *y* (расход топлива); 3) выделение точек с помощью цвета (мощность); 4) выделение точек с помощью размера (масса); 5) выделение точек с помощью формы (количество цилиндров двигателя). Четыре из этих шкал (объем, расход топлива, мощность и масса) являются непрерывными числовыми переменными. Оставшаяся шкала (количество цилиндров) может быть как дискретной числовой переменной, так и упорядоченной качественной. Источник: Motor Trend 1974

Хочу обратить ваше внимание на то, что на рис. 1.4 используются две шкалы положения (месяц по оси *x* и место по оси *y*), но при этом ни одна

из них не является непрерывной. Месяц — это упорядоченный фактор, состоящий из 12 элементов. Место, в свою очередь, является неупорядоченным фактором. Следовательно, обе эти шкалы являются дискретными. При использовании дискретных шкал положения мы располагаем различные уровни на одинаковом друг от друга расстоянии вдоль оси. В случае, если фактор является упорядоченным (на графике таким является месяц), уровни должны располагаться в соответствующем порядке. Если же фактор не является упорядоченным (на графике таким является место), то порядок расположения уровней может быть произвольным. На рис. 1.4 я расположил места наблюдения в порядке повышения температуры (от Чикаго до Долины Смерти), чтобы создать изящный цветовой переход. Разумеется, я мог бы выбрать любой другой порядок мест, и это бы никак не отразилось на точности подаваемой информации.

Как на рис. 1.3, так и на рис. 1.4 мы использовали три шкалы: две шкалы положения и одну цветовую. Такое количество шкал довольно типично при создании простой визуализации, однако никто не запрещает использовать и гораздо больше. Например, на рис. 1.5 используется пять шкал: две шкалы положения, одна цветовая, одна шкала размера и одна шкала формы — и каждая из них соответствует одной из переменных массива данных.

Глава 2

Оси и системы координат

Вне зависимости от вида создаваемой нами визуализации, чтобы придать ей хоть какой-то смысл, мы должны задать шкалы положения, которые определяют расположение различных значений данных на графике. Визуализировать что-либо, не задав для каждой точки данных ее положение в пространстве, невозможно: даже если мы хотим просто разместить все точки одну за другой на одной прямой, без шкалы положения мы не обойдемся. При использовании стандартных двухмерных методов визуализации (на плоскости) для определения положения каждой конкретной точки нам требуются два числа, а значит — две шкалы положения. Эти шкалы обычно, но не обязательно, являются осями *x* и *y* плоского графика. Также мы должны определить геометрическое расположение описанных шкал относительно друг друга: обычно ось *x* располагается горизонтально, а ось *y* — вертикально, но существуют и другие варианты. К примеру, угол между осями *x* и *y* может быть острым, или одна ось может иметь форму окружности, а вторая быть радиальной. Комбинация набора шкал положения и их расположения относительно друг друга называется *системой координат*.

Прямоугольная (декартова) система координат

Наиболее популярной системой координат для визуализации данных является двухмерная *прямоугольная (декартова) система координат*, в которой положение каждой точки однозначно определяется значениями *x* и *y*. Оси абсцисс (*x*) и ординат (*y*) расположены перпендикулярно друг другу, а значения данных находятся на равномерном расстоянии вдоль них (рис. 2.1). Эти оси являются непрерывными шкалами положения и могут отображать как положительные, так и отрицательные действительные числа. Чтобы полноценно задать систему координат, нам необходимо определить промежуток, который покрывает каждая из осей. На рис. 2.1 ось *x* отображает промежуток от –2,2 до 3,2, а ось *y*, в свою очередь, от –2,2 до 2,2. Любые значения внутри этих промежутков располагаются в соответствующей точке графика, а значения, выходящие за границы данных промежутков, отбрасываются.

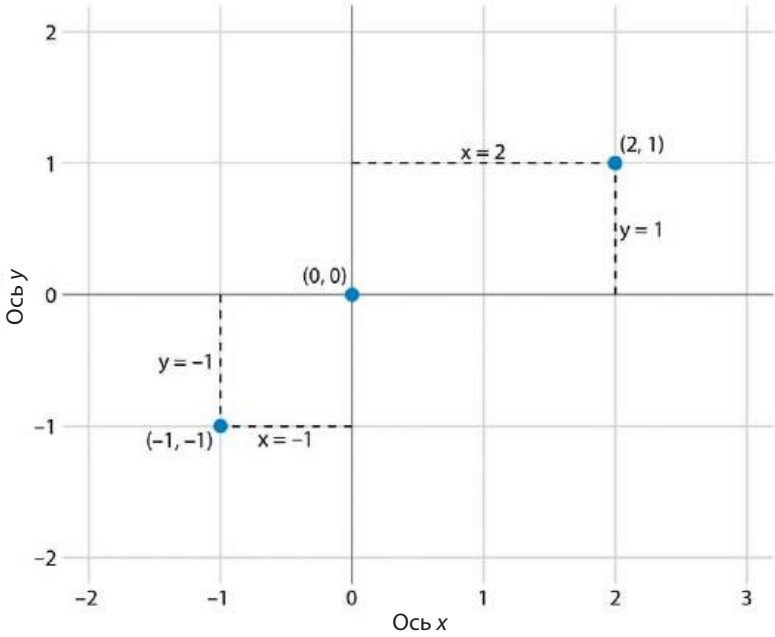


Рис. 2.1. Стандартная прямоугольная система координат. Горизонтальная ось обычно называется *x*, а вертикальная — *y*. На этих осях построена сетка из равноудаленных горизонтальных и вертикальных линий, расстояние между которыми равно 1. Точка (2, 1) расположена на две клетки правее по оси *x* и на одну клетку выше по оси *y* от начала координат (0, 0). Точка (–1, –1) расположена от начала координат на одну клетку левее по *x* и одну клетку ниже по *y*

Стоит отметить, что значения данных зачастую являются не просто числами. Как правило, они сопровождаются единицами измерения. К примеру, если речь идет о температуре, то значения могут измеряться в градусах по Цельсию или Фаренгейту. Если мы будем измерять расстояние, то значения могут подразумевать километры или мили, а при измерении длительности значения будут в минутах, часах или днях. В прямоугольной системе координат расстояние между линиями сетки вдоль осей координат соответствует единичным шагам, выраженным в единицах измерения данных. Так, например, при отображении температуры единичный отрезок может быть равен 10 градусам по Фаренгейту, а при измерении расстояния — 5 километрам.

В прямоугольной системе координат на разных осях могут быть отложены разные единицы измерения. Довольно часто возникают ситуации, в которых на осях абсцисс и ординат отображаются переменные разных типов. Например, на рис. 1.3 одна ось представляет температуру, а вторая — дни. Единичный отрезок оси *y* на рис. 1.3 (по которой отложена температура в градусах

Фаренгейта) равен 20. По оси *x* отложены месяцы, а ее единичный отрезок равен 3. Несмотря на то что у осей разные единицы измерения, мы можем вытянуть или сжать их таким образом, чтобы создать понятную визуализацию данных (рис. 2.2). Каким именно способом будет оформлен график, зависит от того, на чем мы хотим сделать акцент. Узкое и высокое изображение фокусирует внимание на изменения вдоль оси *y*, а широкое и низкое — ровно наоборот. В идеале нам нужно такое соотношение сторон, которое сделает заметными именно важные изменения.

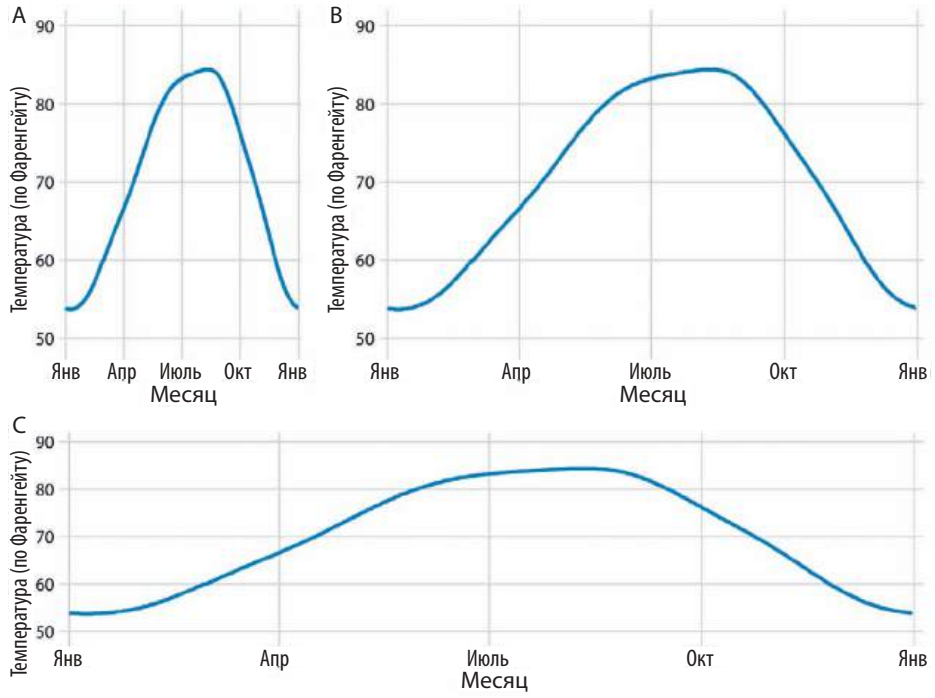


Рис. 2.2. Средняя дневная температура города Хьюстон, штат Техас. Температура отложена по оси *y*, а день — по оси *x*. Части A, B и C показывают один и тот же рисунок с разным соотношением сторон. При этом все три изображения корректно передают данные о температуре. Источник: NOAA

С другой стороны, если значения на осях *x* и *y* имеют одинаковые единицы измерения, это обычно значит, что их единичные отрезки должны быть одинаковыми. В качестве примера мы можем сравнить температуру в Хьюстоне, штат Техас, и в Сан-Диего, штат Калифорния, в каждый день года (рис. 2.3A). Поскольку по обеим осям мы откладываем значения одного и того же диапазона, мы должны убедиться, что линии сетки формируют идеально ровные квадраты, как на рис. 2.3A.

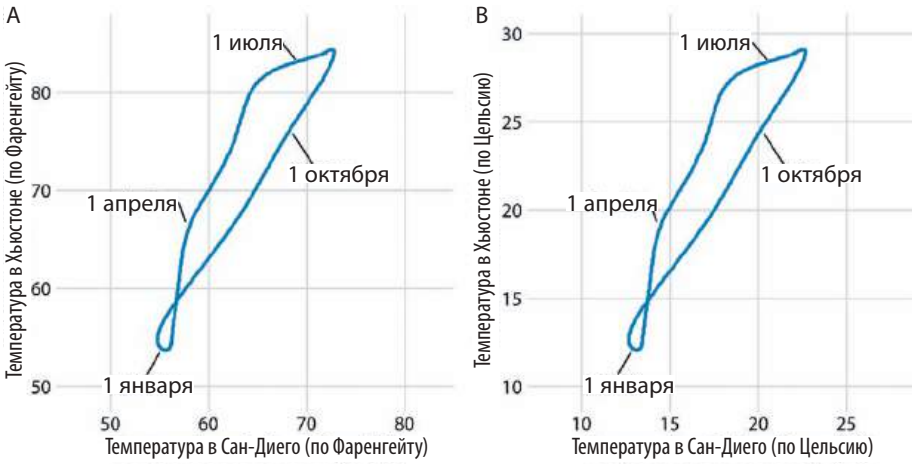


Рис. 2.3. Средняя дневная температура в городе Хьюстон, штат Техас, относительно средней дневной температуры в Сан-Диего, штат Калифорния. Первые дни таких месяцев, как январь, апрель, июль и октябрь, подсвечены, чтобы обозначить привязку ко времени. A. Температура выражена в градусах Фаренгейта. B. Температура выражена в градусах Цельсия. Источник: NOAA

Вам, наверное, интересно, что произойдет, если поменять единицы измерения данных. В конце концов, единицы измерения выбираются произвольно, и ваши предпочтения могут отличаться от предпочтений других людей. Смена единиц измерения — это линейное преобразование, при котором мы добавляем или вычитаем некоторую величину относительно всех значений и/или умножаем все значения на определенное число. К счастью, подобные преобразования не оказывают влияния на прямоугольные системы координат. Поэтому при смене единиц измерения конечный результат останется неизменным. Например, сравните рис. 2.3A и рис. 2.3B. Оба графика показывают одни и те же данные, но на варианте A единицей измерения температуры является градус Фаренгейта, а на варианте B — градус Цельсия. Несмотря на то что линии сетки расположены в разных местах, а единицы измерения различны, обе визуализации данных выглядят одинаково.

Нелинейные оси

В декартовой системе координат линии сетки находятся на одинаковом друг от друга расстоянии: как с точки зрения единиц измерения, так и на получаемой визуализации. Такие шкалы положения называются *линейными*. Несмотря на то, что обычно линейные шкалы обеспечивают точное представление данных, существуют ситуации, когда предпочтительнее использовать нелинейную шкалу. В нелинейной шкале равномерное расстояние между

единичными отрезками единиц измерения соответствует неравномерному расстоянию в визуализации и наоборот.

Наиболее распространенная нелинейная шкала — логарифмическая. Логарифмические шкалы являются линейными при умножении, то есть единичный отрезок на шкале соответствует умножению на фиксированное значение. Для создания такой шкалы к значениям данных следует применить логарифмическое преобразование, а потом возвести в степень числа, которые расположены вдоль линий сетки координат. Этот процесс показан на рис. 2.4, где числа 1; 3,16; 10; 31,6; 100 размещены на линейных и логарифмических шкалах. Может возникнуть вопрос, почему именно 3,16 и 31,6: эти числа выбраны потому, что они находятся на половине пути между 1 и 10 и между 10 и 100 на логарифмической шкале. То, что это действительно так, следует из того, что $10^{0,5} = \sqrt{10} \approx 3,16$ и, соответственно, $3,16 \times 3,16 \approx 10$. Также $10^{1,5} = 10 \times 10^{0,5} \approx 31,6$.

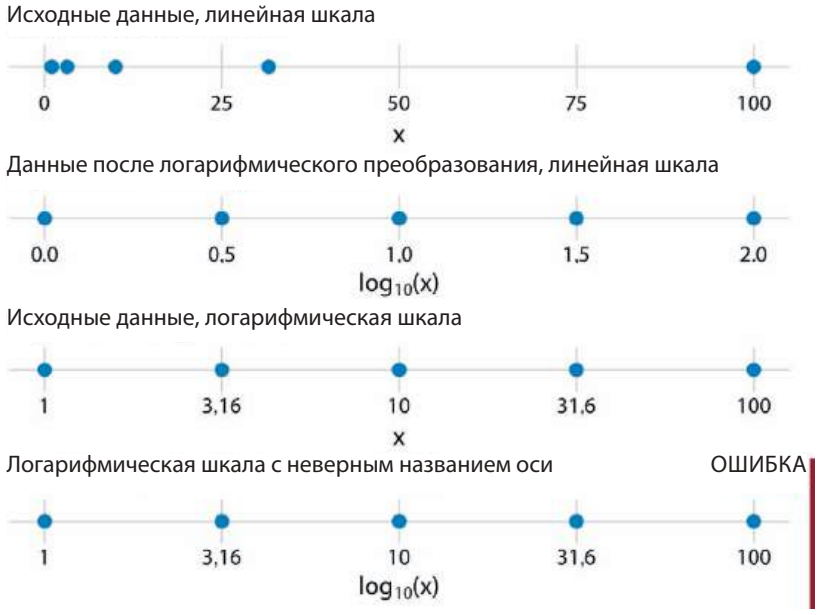


Рис. 2.4. Соотношение между линейной и логарифмической шкалами. Точки соответствуют значениям данных 1; 3,16; 10; 31,6 и 100, которые являются числами, равномерно расположенными в логарифмическом масштабе. Данные точки можно как отобразить на линейной шкале, так и преобразовать их в логарифмы, после чего поместить на линейную или на логарифмическую шкалу. Важно отметить, что правильным названием оси для логарифмической шкалы является имя отображаемой переменной, а не ее логарифм

С точки зрения математики нет никакой разницы между построением логарифмически преобразованных данных на линейной шкале и исходных

данных на логарифмической (рис. 2.4). Единственная разница заключается в обозначении засечек единичных отрезков и самой оси.

В большинстве случаев обозначение шкалы как логарифмической имеет смысл, поскольку так читателю будет проще интерпретировать числа, показанные в качестве меток на оси. Кроме того, снижается вероятность неверного определения основания логарифма. Имея дело с логарифмическими преобразованиями, легко запутаться в том, какой логарифм использовался: натуральный или десятичный. Помимо этого, метки зачастую бывают неопределенными — например, log(x), где вообще не указано основание. Всегда проверяйте наличие основания при работе с логарифмическими данными. При нанесении логарифмически преобразованных данных на график обязательно указывайте основание в маркировке оси.

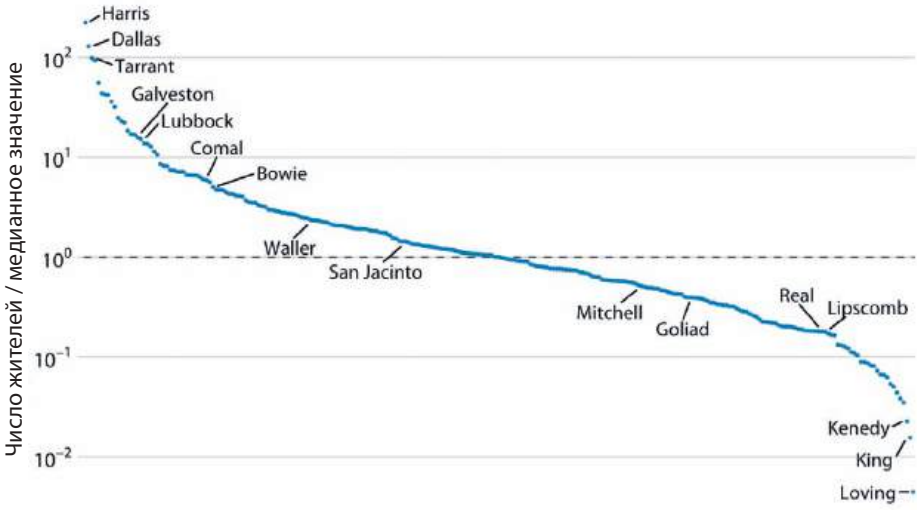


Рис. 2.5. Отношение численности населения округов штата Техас к медианной численности населения штата. Название выбранных округов выделено. Пунктирная линия показывает значение соотношения, равное 1, которое соответствует медиане. В наиболее густонаселенных округах проживает примерно в 100 раз больше жителей, чем в этом округе, а в наименее населенных округах проживает примерно в 100 раз меньше человек. Источник: Перепись населения США, 2010 год

Поскольку умножение на логарифмической шкале выглядит как сложение на линейной, логарифмические шкалы являются логичным выбором для любых данных, полученных путем умножения или деления. В частности, хорошо смотрятся в логарифмическом масштабе доли или соотношения. Рассмотрим следующий пример: я взял число жителей в каждом округе Техаса и разделил его на медианное значение числа жителей по всем округам Техаса.

Результирующее отношение — это число, которое может быть больше или меньше 1. Отношение, равное единице, означает, что в соответствующем округе число жителей равно медианному. При нанесении этих отношений на логарифмическую шкалу мы можем увидеть, что численность населения в округах Техаса симметрично распределена относительно медианы и что в наиболее густонаселенных округах проживает в 100 раз больше людей, чем по медиане, а в наименее населенных округах — в 100 раз меньше (рис. 2.5).

Важно отметить, что расположение тех же данных на линейной шкале будет скрывать различия между округом с медианным числом населения и округом с гораздо меньшим числом (рис. 2.6).

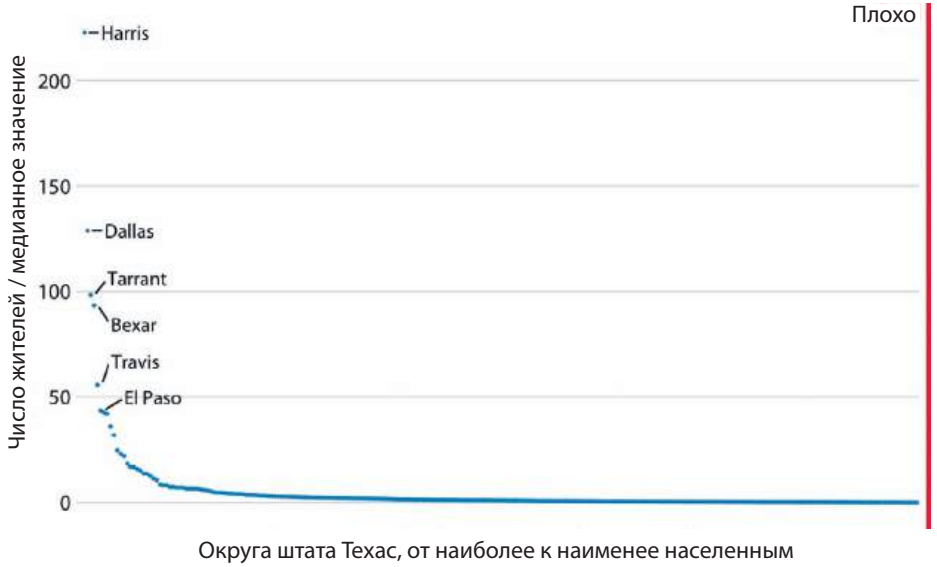


Рис. 2.6. Отношение численности населения округов штата Техас к медианному значению численности населения. Нанеся соотношение на линейную шкалу, мы уделим чрезмерное внимание коэффициентам больше 1 и проигнорируем коэффициенты меньше 1. В большинстве случаев использовать линейную шкалу для отображения соотношений не следует. Источник: Перепись населения США в очередном десятилетии, 2010 год

На логарифмической шкале единица (1) является естественной точкой отсчета, аналогично значению 0 на линейной шкале. Мы можем думать о значениях, превышающих 1, как об умножении, а о значениях, меньших 1, как о делении. Так, мы можем написать: $10 = 1 \times 10$, а $0{,}1 = 1/10$. С другой стороны, значение 0 на логарифмической шкале недостижимо ввиду того, что оно бесконечно далеко от 1. Чтобы понять, почему так, просто взгляните на равенство $\log(0) = -\infty$. Или представьте, что для перехода от 1 к 0 требуется либо бесконечное число делений на конечное значение (например, 1/10/10/10/10/10/10 … = 0), либо одно деление на бесконечность (то есть $1 / \infty = 0$).

Логарифмические шкалы часто используются, когда набор данных содержит значения очень разных порядков. Что касается округов штата Техас, показанных на рис. 2.5 и 2.6, то в наиболее густонаселенном (Harris) проживает 4 092 459 жителей, а в наименее населенном (Loving) — 82. Поэтому логарифмическая шкала здесь будет к месту, хоть мы и не делим численность населения на медиану, чтобы получить соотношения. Но как быть, если появится округ численностью в 0 жителей? Такой округ нельзя отобразить на логарифмической шкале, потому что он будет лежать на минус бесконечности. В подобной ситуации иногда рекомендуют использовать *квадратичную шкалу*, которая подразумевает использование преобразования взятием квадратного корня вместо логарифмического (рис. 2.7). Квадратичная шкала, как и логарифмическая, сжимает большие числа в меньший диапазон, но, в отличие от логарифмической шкалы, допускает наличие 0.

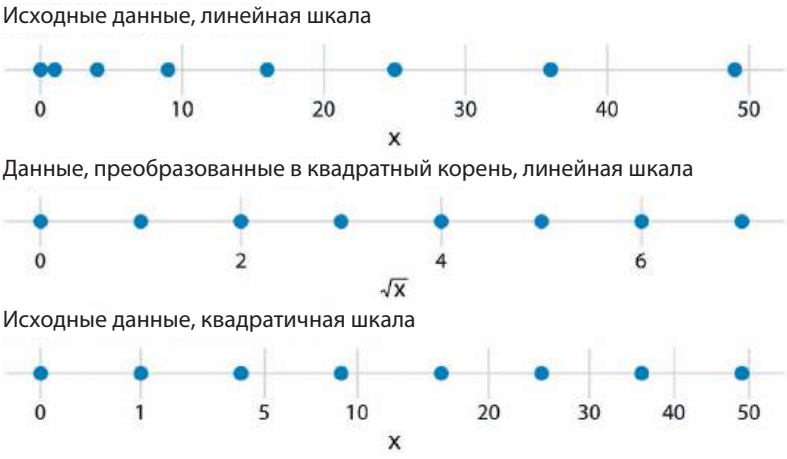


Рис. 2.7. Соотношение между линейной и квадратичной шкалами. Точки соответствуют значениям данных 0, 1, 4, 9, 16, 25, 36 и 49, которые являются равномерно расположенными числами на квадратичной шкале, поскольку они представляют собой квадраты целых чисел от 0 до 7. Мы можем нанести эти точки на линейную шкалу, можем преобразовать их в квадратный корень и затем показать их в линейном масштабе или нанести их на квадратичную шкалу

У квадратичных шкал в глаза бросаются две проблемы. Во-первых, в то время как на линейной шкале один единичный отрезок соответствует прибавлению или вычитанию постоянного значения, а в логарифмической соответствует умножению или делению на постоянное значение, для квадратичной шкалы такого правила не существует. Значение шага квадратичной шкалы зависит от значения масштаба, с которого мы начинаем. Во-вторых, неясно, каким образом лучше всего размещать отметки на квадратичной

оси. Чтобы получить равномерно расположенные отметки, мы должны разместить их у квадратов чисел, но если разместить засечки, например, в позициях 0, 4, 25, 49 и 81 (каждый второй квадрат натурального числа), то это будет совершенно не интуитивно*. В качестве альтернативного решения мы могли бы разместить их через линейные интервалы (10, 20, 30 и т. д.), но это привело бы либо к слишком малому числу отметок у начала шкалы, либо к их слишком большому количеству в конце. На рис. 2.7 я разместил отметки оси в позициях 0, 1, 5, 10, 20, 30, 40 и 50 на квадратичной шкале. Значения выбраны без четкой системы, но обеспечивают адекватный охват нашего диапазона данных.

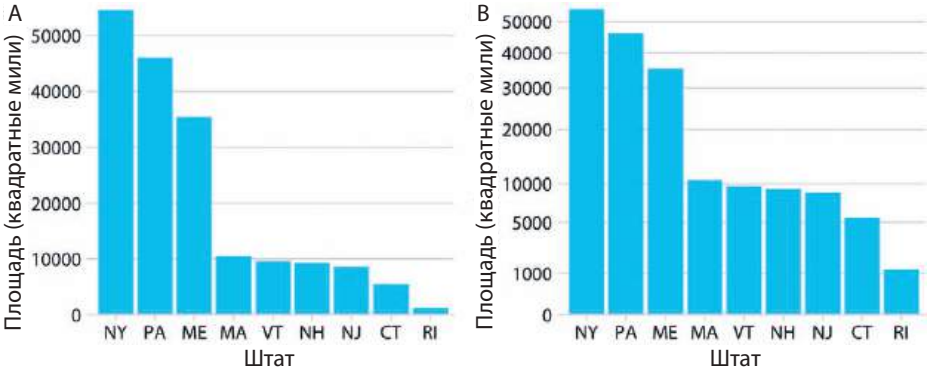


Рис. 2.8. Площадь северо-восточных штатов США. A. Отображение на линейной шкале. B. Отображение на квадратичной шкале. Источник: Google

Несмотря на все вопросы к квадратичным шкалам, как шкалы положения они работают хорошо, и вполне возможно, что вы и для них сможете найти применение. Например, точно так же как то, что логарифмическая шкала является лучшим видом шкалы для соотношений, можно утверждать, что квадратичная шкала является естественной шкалой для данных, представленных в виде квадратов чисел. В качестве примера, естественным образом представляют собой квадраты чисел сведения о географических регионах. Размещая площади регионов на квадратичной шкале, мы, по сути, подсвечиваем линейную протяженность регионов с востока на запад или с севера на юг. Такое представление имеет смысл, например, если нам нужно понять, сколько времени потребуется, чтобы пересечь регион на автомобиле. На рис. 2.8 показаны площади штатов на северо-востоке США на линейной и квадратичной шкалах. Несмотря на существенные различия между площадями (рис. 2.8A), относительное время, необходимое для проезда через

* Эти метки не считываются при восприятии графика и вызывают много вопросов, как правильно читать данные и нет ли какого нарушения в легенде.

каждый из штатов, можно более точно представить в случае отображения данных о площадях на квадратичной шкале (рис. 2.8B), нежели на линейной (рис. 2.8A).

Системы координат с изогнутыми осями

Все системы координат, с которыми мы сталкивались до сих пор, состояли из двух прямых осей, расположенных под прямым углом друг к другу, даже если сами оси нелинейно отображали значения данных на позиции. Однако существуют и другие виды систем координат: с изогнутыми осями. Например, в *полярной* системе координат местоположение точки задается через угол и радиус от начала координат, и поэтому угловая ось является окружностью (рис. 2.9).

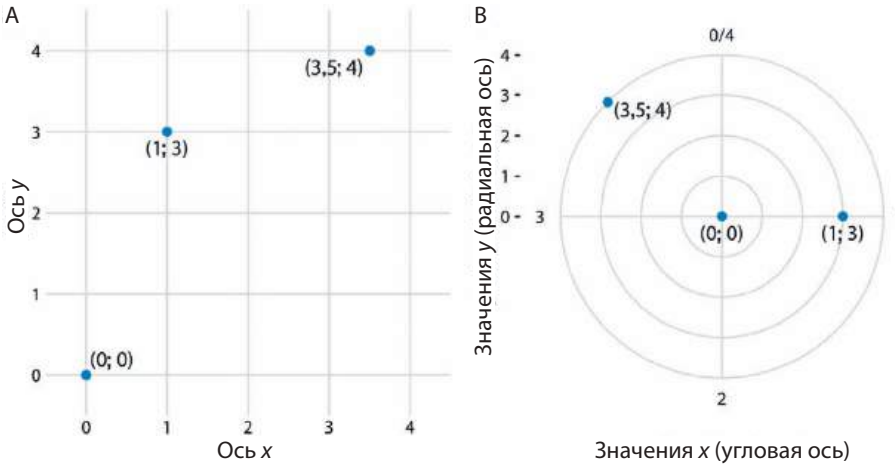


Рис. 2.9. Связь между прямоугольной и полярной системами координат.
A. Три точки данных, показанные в прямоугольной системе координат.
B. Те же три точки данных показаны в полярной системе координат.
Мы взяли координаты по оси *x* из части A и использовали их как угловые координаты, а координаты по оси *y* из части A — как радиальные.
В этом примере угловые координаты имеют диапазон от 0 до 4, и, следовательно, в этой системе координат $x = 0$ и $x = 4$ находятся в одной и той же точке

Полярные координаты могут быть полезны для демонстрации данных периодического характера, таких, где значения данных на одном конце шкалы могут быть логически связаны со значениями данных на другом конце. В качестве иллюстрации рассмотрим дни в году: 31 декабря — последний день года, но это также день, за которым следует первый день следующего года. Если мы хотим отобразить изменение некоторых количественных данных в течение года, хорошим решением может быть использование полярных

координат, в которых координата угла наклона указывает на день в году. Давайте применим эту концепцию к статистике температур с рис. 1.3. Поскольку температурные нормы — это средние температуры, которые не привязаны к какому-либо конкретному году, можно считать, что 31 декабря находится на 366 дней позже 1 января (данные включают 29 февраля) и в то же время на один день раньше.

Нанося температурные данные на график в полярной системе координат, мы подчеркиваем их цикличный характер (рис. 2.10). По сравнению с рис. 1.3 версия с полярной системой координат показывает близость температур в Долине Смерти, Хьюстоне, Чикаго и Сан-Диего с конца осени и до начала весны. На графике в декартовой системе координат этот факт обнаружить сложнее, потому что значения температуры в конце декабря и в начале января показаны в противоположных частях рисунка и поэтому визуально не образуют единое целое.

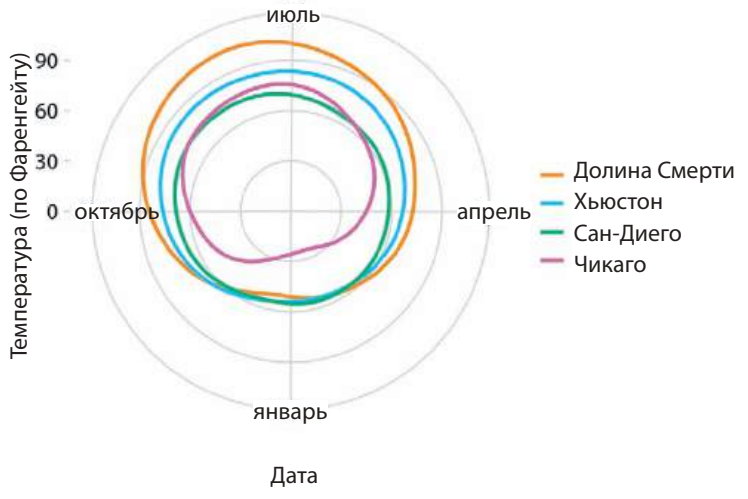


Рис. 2.10. Нормы дневной температуры для четырех пунктов наблюдения в США, визуализированные в полярной системе координат. Расстояние от центра показывает величину средней дневной температуры в градусах Фаренгейта, а дни откладываются против часовой стрелки, начиная с 1 января в позиции 6 часов на циферблате. Источник: NOAA

Еще одним примером, в котором мы встречаемся с изогнутыми осями, являются географические данные, например карты. Положение точек на земном шаре определяется их долготой и широтой, а поскольку Земля — это (в некотором приближении) сфероид, отображение широты и долготы в виде прямоугольной системы координат вводит в заблуждение и не рекомендуется для использования (рис. 2.11). В данном случае вместо прямоугольных систем следует использовать различные типы нелинейных

проекций, которые минимизируют искажения и обеспечивают баланс между сохранением площадей и углами относительно истинных изогнутых линий земного шара.

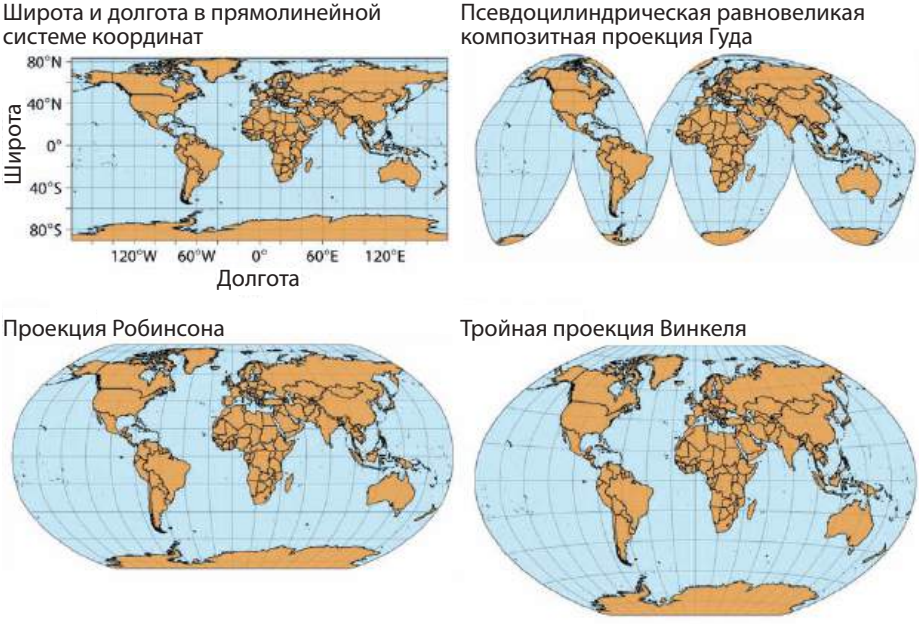


Рис. 2.11. Карта мира, представленная в четырех разных проекциях. Прямоугольная карта отображает долготу и широту каждого местоположения в стандартной прямоугольной системе координат. Это отображение вызывает существенные искажения как площадей, так и углов относительно их истинных величин на трехмерном сфероиде. Проекция Гуда прекрасно отображает истинные площади поверхности, но ценой разделения некоторых массивов суши на отдельные части (в особенности Гренландии и Антарктики). Проекция Робинсона и тройная проекция Винкеля сохраняют баланс между искажениями углов и площадей и обычно используются для отрисовки карт всей поверхности Земли

Глава 3

Цветовые шкалы

Существует три основных способа использования цвета при визуализации данных: мы можем использовать цвета, чтобы различать группы данных, представлять значения данных, а также для выделения. В зависимости от цели типы используемых цветов и способы их применения будут различаться.

Цвет как средство различения

Мы часто используем цвет, чтобы отделить друг от друга предметы или группы, которые не имеют собственного ранжирования, такие как, например, страны на карте или производители определенного продукта. В этих случаях применяется *качественная*, или *категориальная*, цветовая шкала. Шкала такого типа содержит конечный набор цветов, выбираемых таким образом, чтобы они явно отличались друг от друга, но при этом оставались эквивалентными. Второе условие означает, что ни один цвет не должен выделяться из ряда остальных. Кроме того, выбор цветов не должен создавать впечатление упорядоченности, как может быть в случае с набором оттенков с выраженным последовательным осветлением: такой подбор цветов создает иллюзию порядка среди окрашиваемых элементов, хотя они по определению не упорядочены.

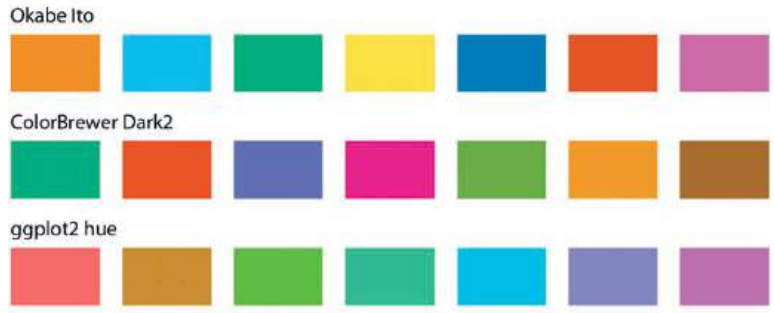


Рис. 3.1. Примеры категориальных цветовых шкал. Шкала Okabe Ito является шкалой, используемой по умолчанию в этой книге [Okabe and Ito, 2008]. Шкала ColorBrewer Dark2 предоставлена проектом ColorBrewer [Brewer, 2017]. Шкала оттенков ggplot2 hue — шкала, используемая по умолчанию в ggplot2, широко распространенном программном обеспечении для построения графиков

В свободном доступе есть немало категориальных цветовых шкал. На рис. 3.1 приведены три примера. В частности, проект ColorBrewer предоставляет хороший выбор категориальных цветовых шкал, включая как довольно светлые, так и довольно темные [Brewer, 2017].

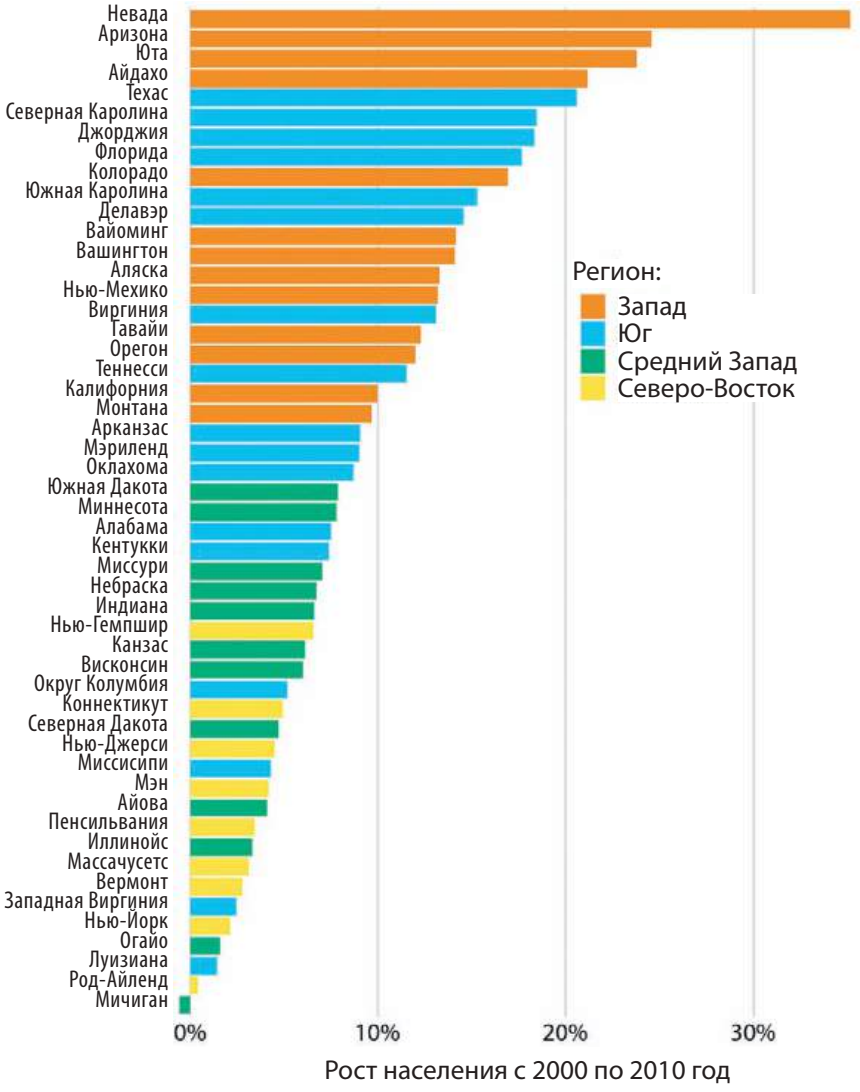


Рис. 3.2. Прирост населения в США с 2000 по 2010 год. В западных и южных штатах наблюдается наибольший прирост, тогда как в штатах на Среднем Западе и Северо-Востоке рост был значительно слабее (или, в случае Мичигана, даже наблюдалось снижение). Источник: Бюро переписи населения США*

* Административный орган, отвечающий за проведение переписи населения в США. — *Прим. науч. ред.*

В качестве примера использования категориальной шкалы рассмотрим рис. 3.2. График демонстрирует рост населения в процентах с 2000 по 2010 год в различных штатах США. Я расположил штаты в порядке возрастания объемов роста и раскрасил их в зависимости от географического региона. Раскраска делает акцент на то, что в штатах, расположенных в одних и тех же регионах, наблюдается схожий рост населения. В частности, самый большой прирост населения виден в южных и западных штатах США, в то время как в штатах, расположенных на Среднем Западе и Северо-Востоке, прирост значительно ниже.

Цвет как средство представления значений данных

Цвет также можно использовать для представления количественных значений данных, таких как доход, температура или скорость. В таких случаях применяется *последовательная* цветовая шкала. Такой тип шкалы представляет собой последовательность цветов и четко указывает, какие значения больше или меньше других, а также то, насколько удалены друг от друга два конкретных значения. Второе подразумевает, что изменения цветовой шкалы должны быть однородными на всем диапазоне значений.

Последовательные шкалы могут основываться на одном основном цвете (например, от темно-синего до светло-голубого) или на нескольких основных цветах (например, от темно-красного до светло-желтого) (рис. 3.3). Многоцветные шкалы обычно строятся на цветовых градиентах, которые встречаются в природе: например, от темно-красного к зеленому, от синего до светло-желтого или от темно-фиолетового до светло-зеленого. Обратная последовательность цветов (например, от темно-желтого до светло-голубого) выглядит неестественно, из-за чего шкала теряет осмысленность.

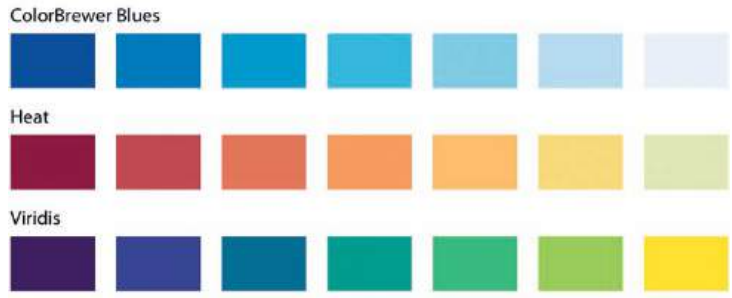


Рис. 3.3. Примеры последовательных цветовых шкал. Шкала ColorBrewer Blues представляет собой одноцветную шкалу, состоящую из оттенков от темно-синего до светло-голубого. Шкалы Heat и Viridis представляют собой многоцветные шкалы, которые состоят из оттенков в диапазоне от темно-красного до светло-желтого и от темно-фиолетового к зеленому и желтому соответственно

Цветовое представление данных особенно хорошо подходит для тех случаев, когда мы хотим продемонстрировать различия в данных географических областей: мы можем нарисовать карту регионов и раскрасить ее в соответствии со значениями данных. Такие карты называются хороплетами, или *фоновыми картограммами*. Пример этой визуализации приведен на рис. 3.4, где на карте округов штата Техас отображен медианный годовой доход в каждом из этих округов.

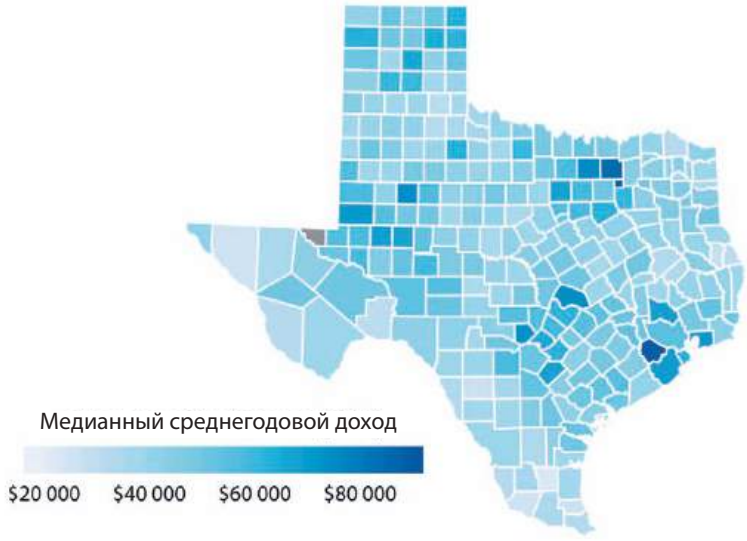


Рис. 3.4. Медианный годовой доход в округах штата Техас. Самые высокие средние доходы наблюдаются в крупных мегаполисах Техаса, в частности вблизи Хьюстона и Далласа. Показатели среднего дохода в округе Ловинг отсутствуют, поэтому этот округ отображается серым цветом. Источник: Исследование американских сообществ от 2015 года

Бывают случаи, когда требуется визуализировать отклонение значений данных в одном из двух направлений относительно некоторой нейтральной средней точки. В качестве примера можно привести любой набор данных, содержащий как положительные, так и отрицательные числа. Чтобы с первого взгляда было понятно, что за число перед нами — положительное или отрицательное, а также насколько далеко оно от нуля, — мы можем раскрасить значения разными цветами. В этой ситуации будет уместно использовать *расходящуюся* цветовую шкалу. Расходящуюся шкалу можно представить себе как две последовательные шкалы, «сшитые» в средней точке, которая обычно имеет светлый оттенок (рис. 3.5). Расходящиеся шкалы должны быть сбалансированы таким образом, чтобы переход от светлых цветов в центре к темным цветам по краям был примерно одинаковым

в обоих направлениях. В противном случае восприятие величины значения данных будет зависеть только от того, где она находится — выше или ниже значения средней точки.

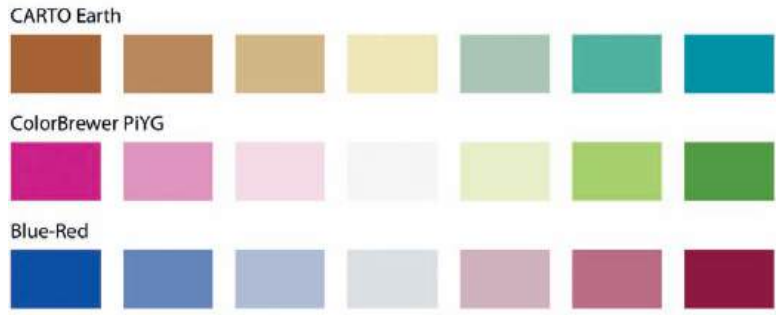


Рис. 3.5. Примеры расходящихся цветовых шкал*. Расходящиеся шкалы можно рассматривать как две последовательные шкалы, сходящиеся в средней точке в общий цвет. Распространенные цветовые решения для расходящихся шкал включают в себя гаммы от коричневого до аквамаринового, от розового до желто-зеленого и от синего до красного

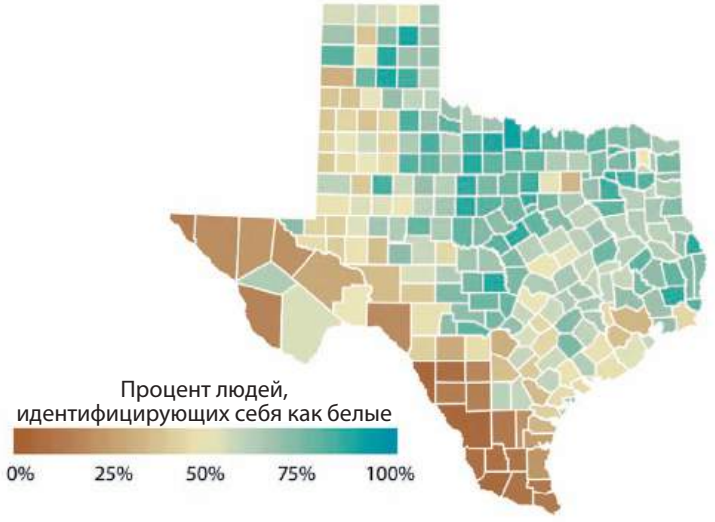


Рис. 3.6. Процент людей, идентифицирующих себя как белых, в округах штата Техас. В Северном и Восточном Техасе белые составляют большинство, а в Южном и Западном Техасе — меньшинство. Источник: Перепись населения США 2010 года

* Стоит отметить, что любое цветовое кодирование должно учитывать известные нарушения цветовосприятия: наиболее частым из них является красно-зеленый дальтонизм, реже встречаются сине-желтый, сине-фиолетовый и полный дальтонизм. На практике это означает, что в расходящейся шкале и при любой цветовой кодификации не стоит использовать пары цветов, соответствующие дальтоническим парам. — *Прим. науч. ред.*

В качестве примера использования расходящейся цветовой шкалы рассмотрим рис. 3.6, на котором показан процент людей, идентифицирующих себя как белых, в различных округах штата Техас. Несмотря на то что процент всегда является положительным числом, расходящаяся шкала вполне здесь уместна, если в качестве срединной точки выбрать 50%. Числа выше 50% указывают на то, что белые составляют большинство, а значения ниже 50% указывают на обратное. Визуализация ясно показывает, в каких округах белые составляют большинство, в каких они составляют меньшинство, а также те, в которых белые и небелые граждане встречаются примерно в равных пропорциях.

Цвет как средство выделения данных

Цвет также может быть эффективным инструментом для выделения определенных данных. В наборе данных могут быть некоторые категории или значения, содержащие ключевую информацию об истории, которую мы хотим рассказать, и мы можем усилить ее, выделив нужные нам элементы изображения. Простой способ добиться этого — раскрасить элементы визуализации цветом или набором цветов, выделяющимся на фоне остальных элементов. Этот эффект может быть достигнут с помощью *акцентирующих* цветовых шкал, которые представляют собой цветовые шкалы, содержащие одновременно наборы приглушенных и более ярких, темных и/или насыщенных цветов (рис. 3.7).

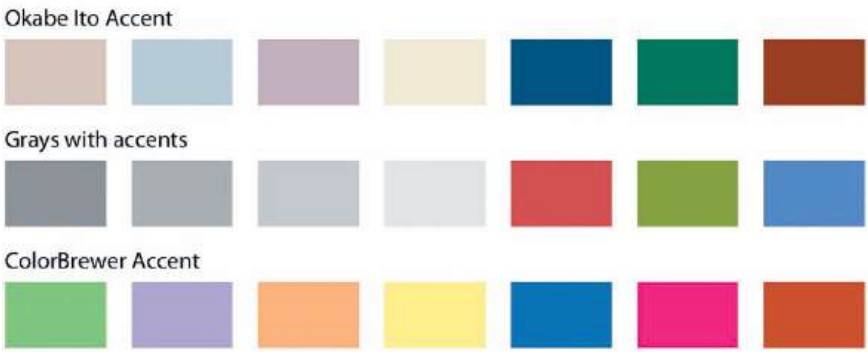


Рис. 3.7. Примеры акцентирующих цветовых шкал, в каждой из которых есть четыре основных цвета и три акцентных цвета. Акцентирующие цветовые шкалы могут быть получены несколькими способами: (сверху) можно взять существующую цветовую шкалу (например, шкалу Okabe Ito, см. рис. 3.1) и осветлить и/или частично обесцветить некоторые цвета, затемняя другие; (посередине) можно взять оттенки серого и добавить к ним цвета; (внизу) можно использовать уже готовую акцентирующую цветовую шкалу (например, из проекта ColorBrewer)

В качестве примера того, как к одним и тем же данным можно применить различные подходы окрашивания, был создан вариант рис. 3.2, где выделены

два конкретных штата: Техас и Луизиана (рис. 3.8). Оба штата находятся на юге и являются соседями, но при этом Техас — пятый самый быстрорастущий штат США в период с 2000 по 2010 год, тогда как Луизиана — всего лишь третий с конца.

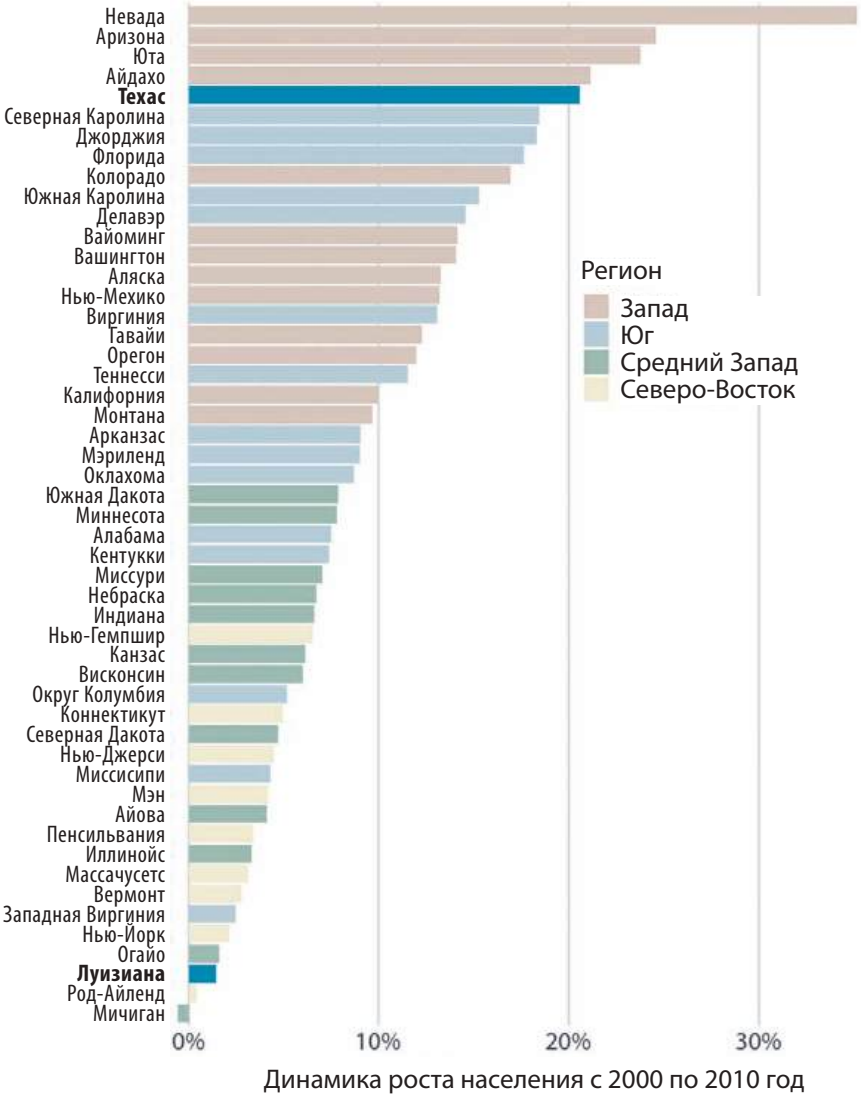


Рис. 3.8. С 2000 по 2010 год в двух соседних южных штатах, Техасе и Луизиане, наблюдался один из самых высоких и самых низких темпов роста населения в США соответственно. Источник: Бюро переписи населения США

Работая с цветовыми акцентами, необходимо помнить, что базовые цвета не должны конкурировать за внимание зрителя. Взгляните на рис. 3.8: цвета

смотрятся скучно, но зато они хорошо оттеняют акцентный цвет. Если базовые цвета будут слишком яркими, они в конечном итоге начнут конкурировать с акцентными цветами за внимание читателя. Эта проблема решается просто: можно вообще удалить цвет со всех элементов на рисунке, кроме выделенных категорий или точек данных. Пример подобной стратегии представлен на рис. 3.9.

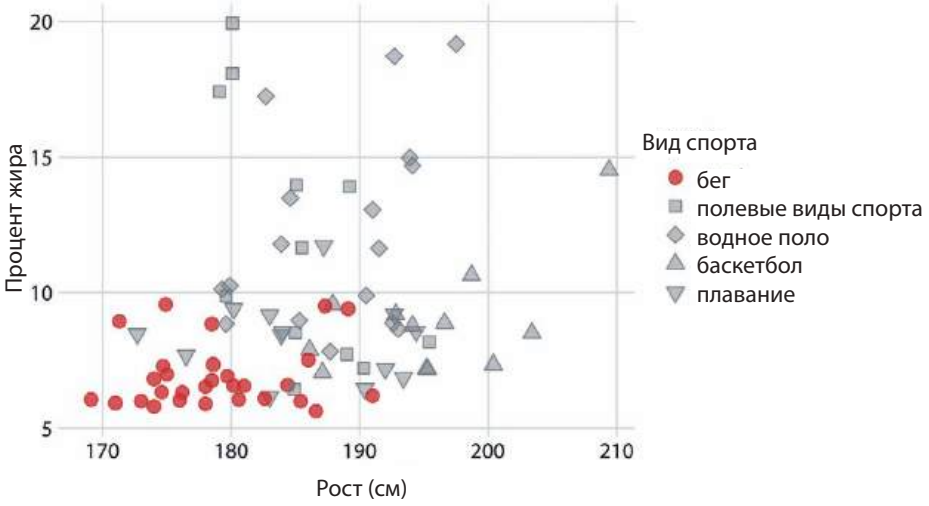


Рис. 3.9. Бегуны — одни из наиболее низкорослых и худых профессиональных спортсменов мужского пола, занимающихся популярными видами спорта. Источник данных: [Telford and Cunningham, 1991]

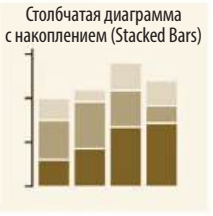
Глава 4

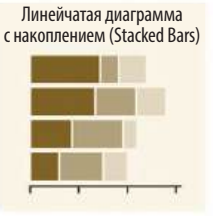
Каталог визуализаций

В этой главе вас ждет краткий обзор с примерами графиков и диаграмм, которые обычно используются для визуализации различных типов данных. Данный раздел можно использовать как каталог, если вам нужна конкретная визуализация, название которой вы не знаете, или же как источник вдохновения, когда вы пытаетесь найти альтернативу набившему оскомину представлению.

Количественные диаграммы

Наиболее распространенный подход к визуализации количественных данных (то есть числовых значений, заданных для некоторого набора категорий) заключается в использовании вертикально или горизонтально расположенных столбцов (см. главу 5).

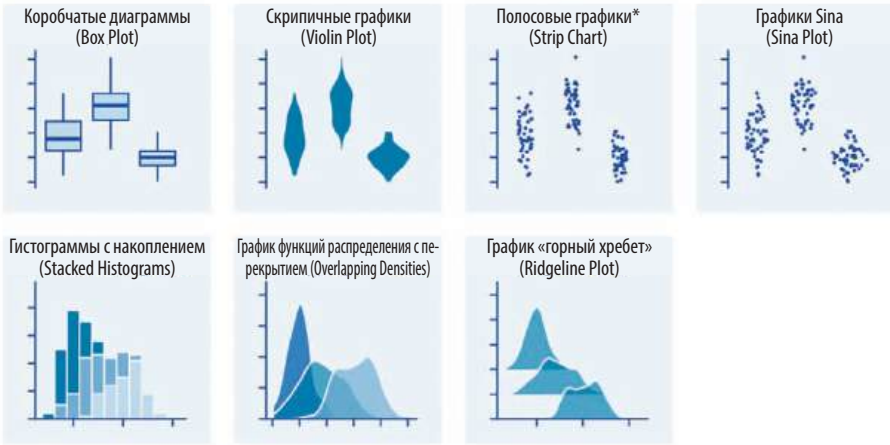




При этом вместо рисования целых столбцов можно просто ставить точки в тех местах, где должны заканчиваться соответствующие столбцы (см. главу 5). Если есть два или более набора категорий, для которых нужно отобразить количественные данные, мы можем сгруппировать столбцы рядом или сложить их один на другой (см. главу 5). Еще вариант — нанести категории на оси *x* и *y*, а значения показать цветами с помощью тепловой карты (см. главу 5).

Диаграммы распределения

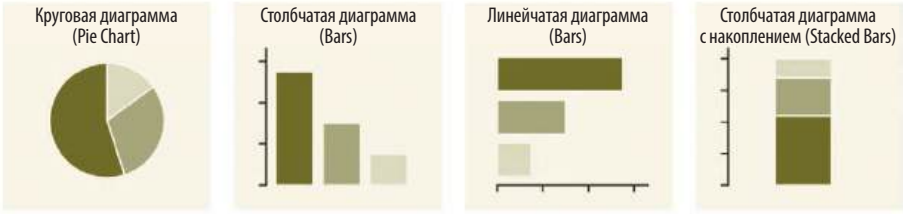
Гистограммы и графики плотности (см. главу 6) обеспечивают наиболее интуитивную визуализацию распределения, но они оба требуют подбора параметров и могут ввести зрителя в заблуждение. Интегральные функции распределения и графики «квантиль-квантиль» (см. главу 7) всегда точно представляют данные, но диаграммы такого типа трудны для восприятия.



* Термин strip chart не имеет однозначного русского перевода. Названия «полосовой» и «ленточной» диаграмм заняты в русскоязычной литературе другими типами диаграмм. В данной книге мы будем называть этот график полосовым, так как «классические» полосовые диаграммы — это то же, что и различные виды линейчатых. — *Прим. науч. ред.*

Коробчатые диаграммы, скрипичные графики, полосовые графики и Sinz полезны тогда, когда мы хотим визуализировать несколько распределений одновременно или если нас в первую очередь интересуют различия между распределениями (см. раздел «Визуализация распределений вдоль вертикальной оси» на с. 88). Гистограммы с накоплением и графики функций распределения с перекрытием дают возможность более глубокого сравнения меньшего количества распределений, однако гистограммы с накоплением не так-то просто считать с первого взгляда, в связи с чем их лучше избегать (см. раздел «Визуализация нескольких распределений одновременно» на с. 75). Графики типа «горный хребет» являются хорошей альтернативой скрипичным графикам и часто бывают полезны при визуализации либо очень большого количества распределений, либо временны́х изменений в распределениях (см. раздел «Визуализация распределений на горизонтальной оси» на с. 95).

Пропорциональные диаграммы



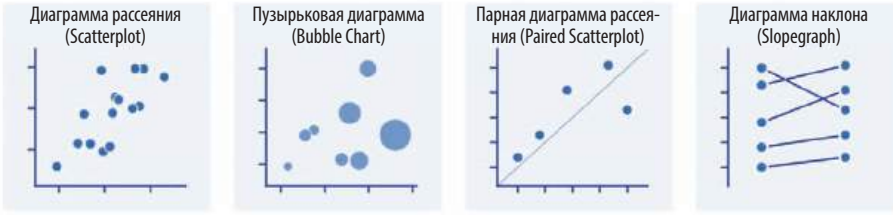
Пропорции можно визуализировать в виде круговых диаграмм, а также столбцов, расположенных рядом друг с другом или сложенных один на другом (см. главу 9). Как и в случае с количественными данными, столбцы можно располагать вертикально или горизонтально. Круговые диаграммы подчеркивают, что целое состоит из частей, и подсвечивают их. При этом сами отдельные части легче сравнивать между собой, используя параллельные столбцы. Столбцы с накоплением же больше подходят для сравнения между собой нескольких наборов пропорций, чем для визуализации одного.

При визуализации набора данных, содержащего несколько групп пропорций или изменений в пропорциях между категориями, круговые диаграммы,

несмотря на свою привлекательность, занимают слишком много места и недостаточно хорошо отражают взаимосвязи. При этом столбчатая диаграмма с группировкой хорошо показывает себя на небольшом количестве факторов сравнения, а когда их становится много, ее место может занять столбчатая (или нормированная столбчатая) диаграмма с накоплением. Нормированный график распределения с накоплением (см. главу 9) работает в случаях, когда изменения происходят не по категориальной, а по непрерывной переменной.

В случаях, когда пропорции представлены в разрезе нескольких группирующих переменных, рекомендуется использовать такие типы визуализаций, как мозаичные графики, деревья и графики в параллельных координатах (см. главу 10). Мозаичный график предполагает, что каждый уровень одной группирующей переменной можно комбинировать с каждым уровнем другой группирующей переменной, тогда как деревья не используют такого предположения. Деревья работают хорошо даже в тех случаях, когда группы несопоставимы между собой по составу подгрупп. Когда же группирующих переменных становится много, лучше других себя показывает график в параллельных координатах: пусть и не такой наглядный, как два предыдущих, он практически не имеет ограничений при отображении наборов данных с большим количеством группирующих переменных.

Диаграммы двух переменных



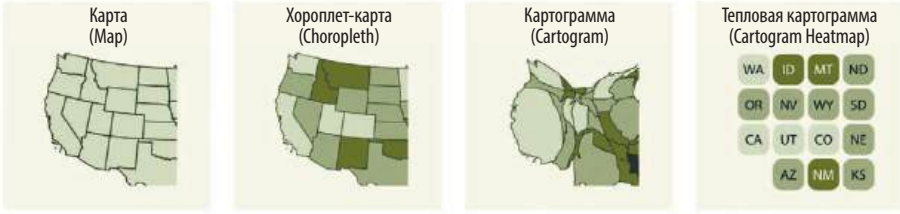
Диаграммы рассеяния (см. главу 11) представляют собой классическую визуализацию, отображающую взаимосвязь между двумя количественными переменными. Если же у нас есть три количественные переменные, мы

можем отобразить одну из них на размер точки, таким образом получив вариант диаграммы рассеяния, называемый *пузырьковой диаграммой*. Для случаев, когда переменные по осям x и y измеряются в одних и тех же единицах, бывает полезно добавить линию, обозначающую $x = y$ (см. «Парные выборки» на с. 130), получая таким образом частный случай, называемый парной диаграммой рассеяния. Парные данные можно также представить в виде наклонного графика, состоящего из пар точек, соединенных прямыми линиями.

В случаях, когда набор данных состоит из очень большого количества точек, стандартные диаграммы рассеяния могут стать неинформативными из-за чрезмерной детализации. На помощь приходят диаграммы с группировкой: изолинии плотности, плоские или шестигранные ячейки (см. главу 17). С другой стороны, если нужно визуализировать более двух количественных факторов, то вместо исходных данных мы можем отобразить коэффициенты их корреляции в форме коррелограммы (см. «Коррелограммы» на с. 124).

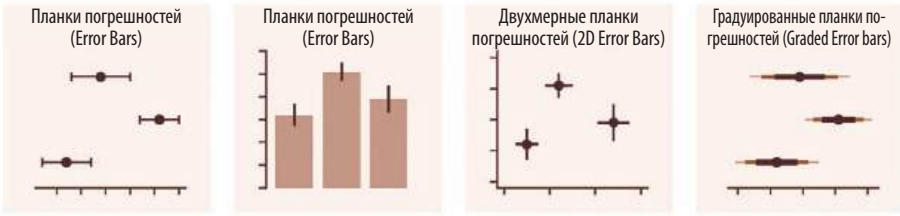
Когда на оси x графика откладывается переменная, характеризующая время или какую-либо строго монотонно возрастающую величину (например, дозу лекарства), для визуализации, как правило, используется линейный график (см. главу 12). Если у нас есть последовательность значений двух объясняемых переменных во времени, то мы можем нарисовать диаграмму рассеяния с отрезками, соединив линиями последовательные по времени точки (см. «Временной ряд двух или более объясняемых переменных» на с. 140). Для демонстрации трендов в большом наборе данных можно использовать сглаженные графики (см. главу 13).

Геопространственные диаграммы



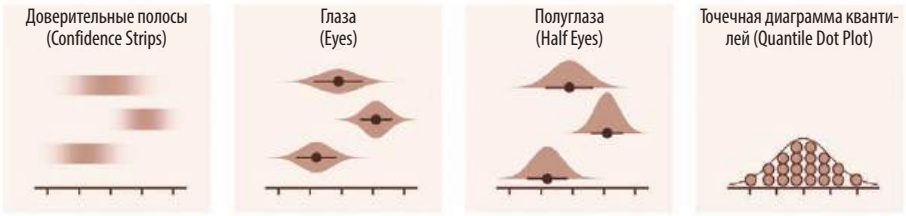
Основной способ отображения геопространственных данных — это, конечно же, карта (см. главу 14). Карта представляет собой проекцию координат земного шара на плоскую поверхность таким образом, чтобы формы и расстояния на земном шаре приблизительно соответствовали формам и расстояниям в 2D-пространстве. Кроме того, карта позволяет показать значения данных в разных регионах, раскрасив регионы в соответствии со значениями. Такая карта называется хороплетом (см. «Фоновые картограммы» на с. 172). В случаях, когда естественные форма и размер региона значения не имеют, их можно сопоставить с какими-либо другими величинами (например, численностью населения или среднедушевым доходом) и соответствующим образом исказить или упростить до квадратов или многоугольников. Такие визуализации называются картограммами (см. «Картограммы» на с. 176).

Неопределенность на диаграммах



Отображение на диаграмме планок погрешностей предназначено для указания диапазона возможных значений при оценке или измерении чего-либо. Визуально планки погрешностей представляют собой границы, размещенные вокруг некоторой опорной точки, которая представляет собой матожидание оценки или измерения (см. главу 15). Сами опорные точки могут изображаться различными способами: например, собственно точками или столбиками. Градуированные планки погрешностей на диаграммах показывают несколько диапазонов одновременно, где каждый диапазон соответствует разной доверительной вероятности. Фактически они представляют собой несколько планок погрешностей,

изображенных одна поверх другой линиями разной толщины. Для получения более детальной картины, чем позволяют планки погрешностей или градуированные планки погрешностей, можно добавить к визуализации уровни доверительной вероятности или апостериорное распределение. Доверительные полосы достаточно хорошо передают суть неопределенности, но интерпретировать их сложно, а вот «глаза» и «полуглаза» объединяют доверительные полосы с подходами визуализации распределений (соответственно, скрипичные графики и «горные хребты») и, таким образом, показывают как точные интервалы для уровней значимости, так и общее распределение неопределенности.



Точечная диаграмма квантилей может служить альтернативной визуализацией распределения неопределенности (см. «“Кадрирование” вероятностей в виде частот» на с. 179), поскольку она показывает распределение в дискретных единицах. Этот вид графика не очень точен, но зато считывается значительно легче, чем непрерывное распределение в виде скрипичного графика или диаграммы «горного хребта».

Для сглаженных графиков эквивалентом доверительных интервалов в точках является доверительный диапазон (см. «Визуализация неопределенности подгонки кривых» на с. 197). Он показывает диапазон значений, через которые может пройти линия с заданной доверительной вероятностью. Как и в случае с доверительными интервалами, мы можем одновременно нарисовать несколько доверительных диапазонов, которые будут соответствовать различным доверительным вероятностям, получив ранжированные доверительные диапазоны. А также вместо самих диапазонов или в дополнение к ним мы можем отобразить несколько «возможных» линий, соответствующих этим диапазонам.

Глава 5

Визуализация количественных данных

При создании визуализаций нас часто интересует суммарная величина какого-либо набора данных: например, общий объем продаж различных брендов автомобилей, общее количество людей, живущих в разных городах, или возраст участников Олимпийских игр в зависимости от вида спорта, которым они занимаются. Во всех этих случаях мы имеем дело с категориями (бренды, города или виды спорта), каждой из которых соответствует количественное значение. Я называю такие случаи количественными визуализациями, потому что основной фокус этих визуализаций — столбчатая диаграмма и ее вариации, включающие как простые столбики, так и столбики сгруппированные или с накоплением. Альтернативами столбчатой диаграмме являются точечная диаграмма и тепловая карта.

Столбчатые диаграммы

Давайте рассмотрим использование столбчатых диаграмм на следующем примере: возьмем общий объем продаж билетов на фильмы, которые были наиболее популярны в выбранный уик-энд.

Таблица 5.1. Самые кассовые фильмы уик-энда с 22 по 24 декабря 2017 года. Источник: Box Office Mojo. Использовано с разрешения источника

Место	Название	Объем продаж за уик-энд
1	«Звездные войны: Последние джедаи»	$71 565 498
2	«Джуманджи: Зов джунглей»	$36 169 328
3	«Идеальный голос 3»	$19 928 525
4	«Величайший шоумен»	$8 805 843
5	«Фердинанд»	$7 316 746

Таблица 5.1 показывает пять наиболее финансово успешных фильмов в кинопрокате в уик-энд перед Рождеством 2017 года. Наиболее популярным стал фильм «Звездные войны: Последние джедаи», обогнавший по объему

продаж билетов фильмы с четвертой и пятой строчек таблицы («Величайший шоумен» и «Фердинанд») практически на порядок.

Данные такого типа чаще всего визуализируются с помощью вертикальных столбцов. Для каждого фильма рисуется столбец, который начинается в точке 0, а заканчивается в точке, соответствующей величине объема продаж билетов (рис. 5.1). Такого рода визуализации называются *столбчатыми (или столбиковыми) диаграммами.*

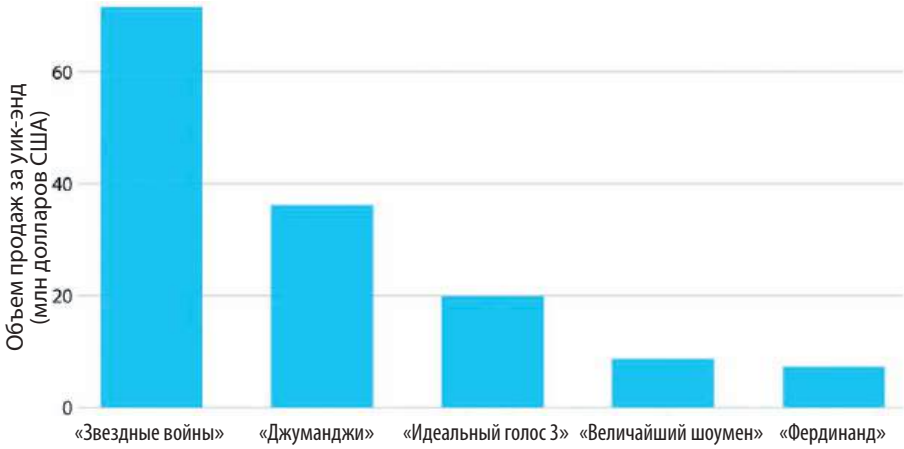


Рис. 5.1. Самые кассовые фильмы уик-энда с 22 по 24 декабря 2017 года, отображенные в виде столбчатой диаграммы. Источник: Box Office Mojo. Использовано с разрешения источника

Одна из наиболее распространенных проблем, связанных с использованием вертикальных столбцов, заключается в том, что подписи к ним занимают слишком много места по горизонтали. Фактически, для того чтобы создать рис. 5.1, мне пришлось сильно расширить столбцы, а также увеличить расстояние между ними. Это было сделано для того, чтобы названиям попросту хватило места. Для экономии пространства мы могли бы разместить столбцы ближе друг к другу, а подписи слегка повернуть вокруг своей оси (рис. 5.2). Однако хочу отметить, что я не большой фанат такого размещения подписей. На мой взгляд, смотрится это некрасиво, да и воспринимать информацию становится труднее. Кроме того, если подписи занимают слишком много места по горизонтали, то и в развернутом виде они тоже не будут хорошо смотреться.

Решением этой проблемы является перемена местами осей координат, так чтобы столбцы расположились горизонтально (рис. 5.3). После подобной рокировки мы получаем удобное для восприятия горизонтальное размещение всех переменных. Как вы могли заметить, такой способ отображения данных куда более приятен глазу, нежели рис. 5.2 и даже рис. 5.1.

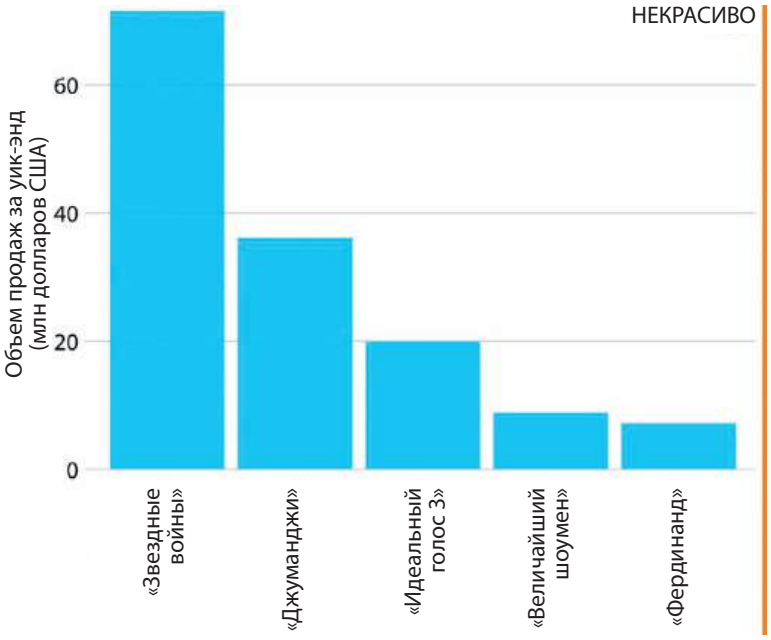


Рис. 5.2. Самые кассовые фильмы уик-энда с 22 по 24 декабря 2017 года, отображенные в виде гистограммы с подписями, повернутыми относительно собственной оси. Строки, расположенные подобным образом, тяжело читать, а кроме того, пространство под графиком можно было бы использовать более рационально. Именно поэтому я считаю такие диаграммы некрасивыми. Источник: Box Office Mojo. Использовано с разрешения источника

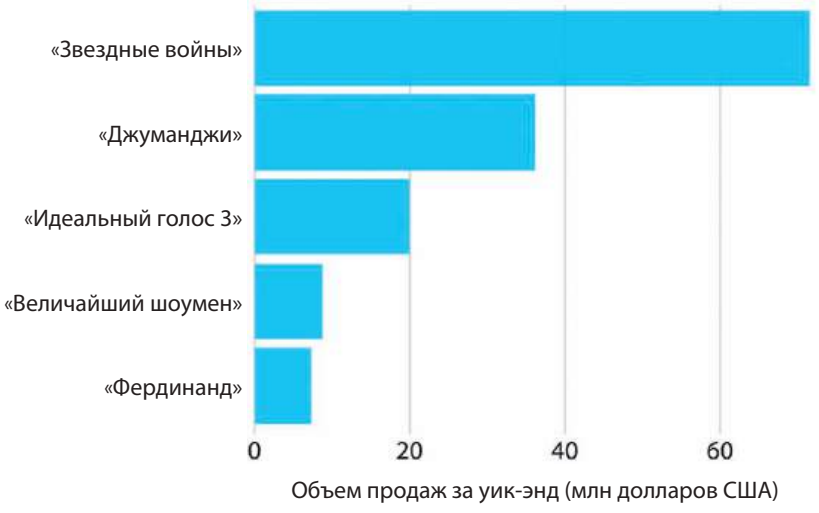


Рис. 5.3. Самые кассовые фильмы уик-энда с 22 по 24 декабря 2017 года, отображенные в виде полосовой диаграммы. Источник: Box Office Mojo. Использовано с разрешения источника

Но как бы ни были расположены столбцы, куда важнее порядок их следования. Я часто встречаю гистограммы, в которых столбцы расположены либо в случайном порядке, либо по каким-то критериям, несущественным для цели визуализации.

Иногда в таких диаграммах данные отсортированы в алфавитном порядке, а иногда по какому-либо другому незначимому для визуализации признаку. Один такой пример можно увидеть на рис. 5.4. В результате такие изображения сложнее считываются глазом и сильнее сбивают с толку, нежели те, в которых столбцы расположены в порядке увеличения или убывания размера.

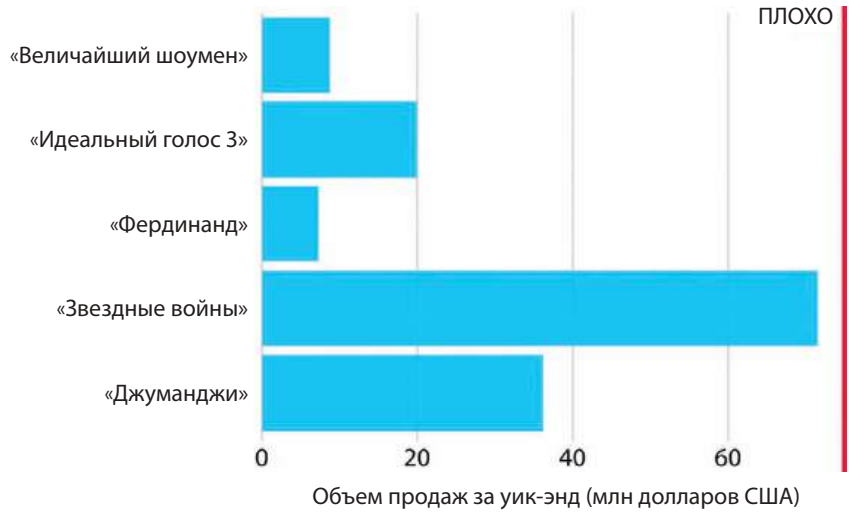


Рис. 5.4. Самые кассовые фильмы уик-энда с 22 по 24 декабря 2017 года, отображенные в виде горизонтальной столбиковой диаграммы. На данном рисунке столбики расположены в порядке убывания длины названия фильма на английском языке. Подобное расположение является произвольным, не несет какой-либо смысловой нагрузки, а также делает изображение менее понятным, нежели рис. 5.3. Источник: Box Office Mojo. Использовано с разрешения источника

Порядок столбцов стоит менять только в тех случаях, когда категории, отображаемые этими столбцами, не имеют никакого внутреннего ранжирования. В случаях же, когда внутреннее ранжирование есть (например, когда наша категориальная переменная является упорядоченным фактором), исходный порядок на уровне визуализации менять не следует.

К примеру, на рис. 5.5 показан средний годовой доход жителей США, разбитый по возрастным группам. В этом случае столбцы должны идти в порядке увеличения возраста группы. Сортировка столбцов по высоте, из-за чего возрасты идут произвольным образом, не имеет никакого смысла (рис. 5.6).

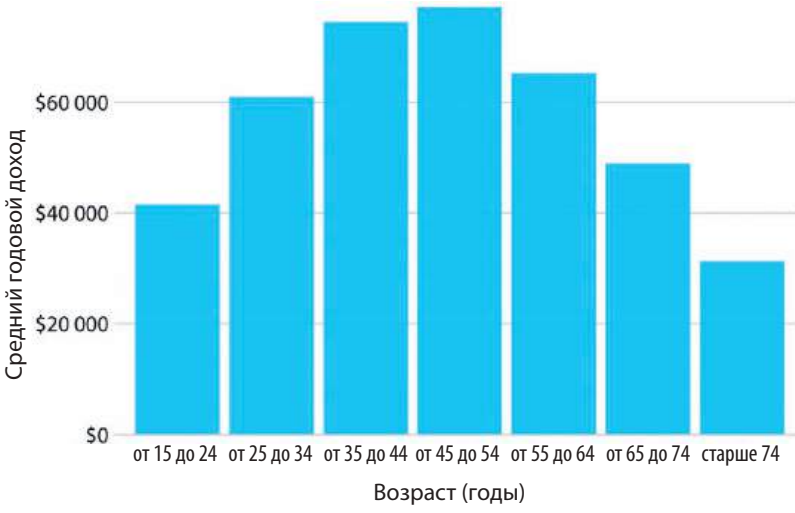


Рис. 5.5. Среднегодовой доход жителей США в 2016 году в зависимости от возрастной группы. Наивысший доход демонстрирует возрастная группа от 45 до 54 лет. Источник: Бюро переписи населения США

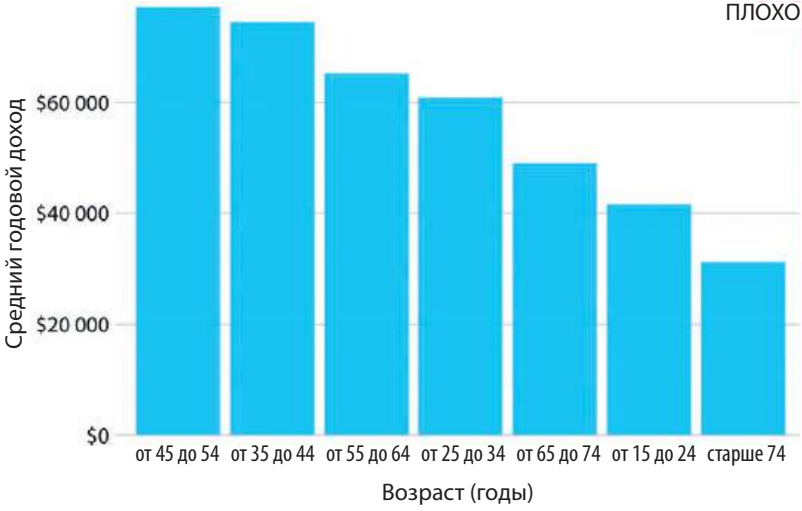


Рис. 5.6. Среднегодовой доход жителей США в 2016 году в зависимости от возрастной группы. График отсортирован в порядке уменьшения уровня доходов. Несмотря на то что визуально график стал более привлекательным, порядок следования возрастных групп сильно сбивает с толку. Источник: Бюро переписи населения США

Всегда обращайте внимание на порядок следования столбцов. В случае, если они представляют неупорядоченные категории, располагайте их в порядке возрастания или уменьшения числовых значений данных.

Столбчатые диаграммы с группировкой и накоплением

Примеры предыдущего раздела иллюстрируют изменение количественной величины в зависимости от одной категориальной переменной. Однако гораздо чаще встречается ситуация, когда на графике нужно отобразить две такие переменные одновременно. К примеру, Бюро переписи населения США предоставляет данные о среднем годовом доходе не только в разрезе возраста, но и расы. Мы можем визуализировать этот набор данных при помощи *столбчатой диаграммы с группировкой* (рис. 5.7).

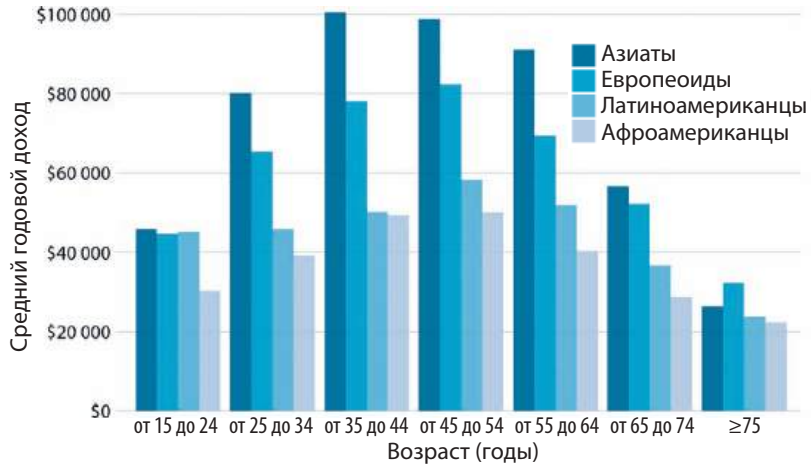


Рис. 5.7. Средний годовой уровень дохода граждан США в 2016 году в зависимости от возраста и расы. Возрастные группы отложены вдоль оси *x,* и каждая из них содержит четыре столбца, отражающие средние годовые доходы четырех расовых групп: азиатов, европеоидов, латиноамериканцев и афроамериканцев соответственно. Источник: Бюро переписи населения США

При создании столбчатой диаграммы с группировкой сначала каждому значению одной категориальной переменной ставится в соответствие отметка по оси *x*, а затем на каждой из отметок откладываются столбцы для каждого из значений другой категориальной переменной. Плотность информации в подобных диаграммах очень высока, и это может сбивать с толку. Честно говоря, несмотря на то, что рис. 5.7 не помечен мной как «плохой» или «некрасивый», я считаю, что он является довольно сложным для восприятия. В частности, очень трудно сравнивать значения среднего дохода различных возрастных групп в пределах одной расовой группы. Отсюда следует вывод, что данное изображение подходит только для тех случаев, когда нас интересует разница в доходах различных рас в зависимости от возрастной группы. Если бы перед нами стояла задача сравнения уровня дохода в зависимости

от расы, удобнее было бы показать расовую группу как основную по оси *x*, а возраст задать в виде отдельных столбцов внутри каждой расовой группы (рис. 5.8).

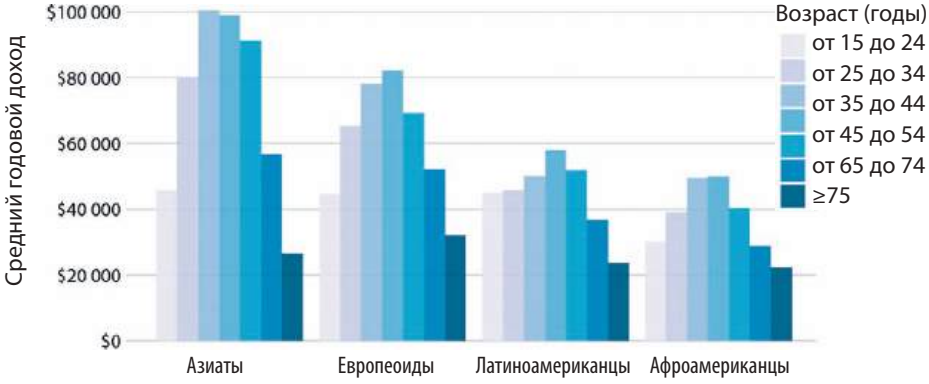


Рис. 5.8. Средний годовой уровень дохода граждан США в 2016 году в зависимости от возраста и расы. В отличие от рис. 5.7, здесь вдоль оси *x* отложены расовые группы, каждая из которых содержит по семь столбцов, отражающих средние годовые доходы семи возрастных групп. Источник: Бюро переписи населения США

Как на рис. 5.7, так и на рис. 5.8 одна категориальная переменная показывается положением на оси *x*, а вторая — цветом столбца. В обоих случаях переменные, что показаны при помощи положения на системе координат, воспринимаются легче, нежели те, что показаны цветом. Так происходит потому, что нам все время приходится сравнивать цвета с легендой. Мы можем избежать лишних умственных усилий, нарисовав четыре обычных столбиковых диаграммы вместо одной с группировкой (рис. 5.9). Выбор типа графика — дело вкуса. Я бы выбрал рис. 5.9, так как он избавляет нас от нужды раскрашивать столбцы в различные цвета.

Совсем необязательно рисовать столбцы рядом, иногда бывает удобнее расположить их сверху друг на друге. Данный метод очень удобен для тех случаев, когда сумма значений столбцов сама по себе является важным числовым значением. Например, нет никакого смысла складывать средние значения, которые мы видим на рис. 5.7 (сумма нескольких средних годовых уровней дохода не является значимой величиной), но при этом значения прибыли за уик-энд с рис. 5.1 можно попробовать сложить (сумма прибыли за уик-энд двух фильмов является общим объемом прибыли за оба фильма). Накопление также будет иметь смысл, если значения столбцов отражают количества. К примеру, если мы работаем с данными о людях, то мы можем посчитать и отобразить данные о мужчинах и женщинах как по отдельности, так и вместе. Если мы располагаем столбец, показывающий какую-либо информацию о количестве женщин, над столбцом, показывающим информацию

о количестве мужчин, то общий столбец будет показывать суммарное количество вне зависимости от пола.

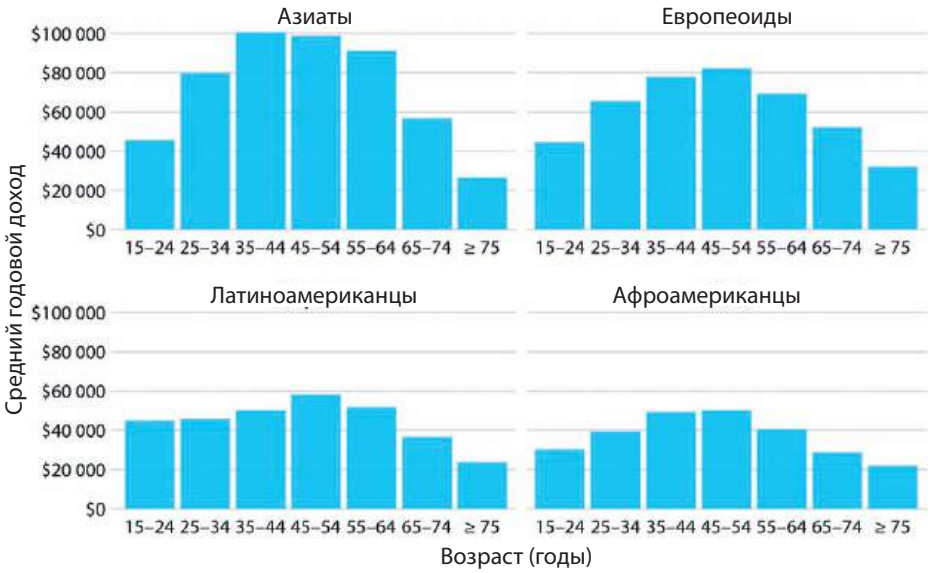


Рис. 5.9. Средний годовой уровень дохода в 2016 году в зависимости от расы и возраста. Вместо того чтобы изобразить эти данные на столбиковой диаграмме с группировкой, как на рис 5.7 и 5.8, мы изображаем их на четырех разных диаграммах. Благодаря подобному представлению нам не нужно выделять цветом ни одну из категориальных переменных. Источник: Бюро переписи населения США

Продемонстрирую вышесказанное с помощью набора данных о пассажирах трансатлантического лайнера «Титаник», который затонул 15 апреля 1912 года. На борту было около 1300 пассажиров без учета экипажа. Пассажиры были распределены по каютам одного из трех классов (первого, второго или третьего), а мужчин среди пассажиров было в два раза больше, чем женщин. Для того чтобы визуализировать распределение пассажиров по классу и полу, нарисуем отдельные столбцы для каждого из них, а затем расположим столбцы, представляющие женщин, над столбцами, представляющими мужчин, отдельно для каждого класса (рис. 5.10). Накопленные значения в столбцах показывают общее количество людей в каютах определенного класса.

В отличие от всех приведенных ранее столбчатых диаграмм, рис. 5.10 не имеет явно выраженной оси *y*. Вместо этого я записал значения внутри самих столбцов. Если на графике предполагается отразить небольшое количество разных значений, попробуйте сделать так же и нанести числовые значения непосредственно на элементы графика, чтобы не перегружать картинку визуальными элементами и избавиться от вертикальной оси.

Рис. 5.10. Количество пассажиров мужского и женского пола на «Титанике», путешествовавших в первом, втором и третьем классах. Источник: Encyclopedia Titanica

Точечные графики и тепловые карты

Столбчатые диаграммы не единственный способ визуализации количественных данных. Существенным их ограничением является то, что длина столбцов должна быть пропорциональна отображаемым значениям, а начинаться они должны в нуле. Для некоторых наборов данных это непрактично, а в каких-то случаях такое представление может просто скрыть ключевой смысл. В подобных случаях количественные значения можно обозначить с помощью размещения точек в заданных координатах осей *x* и *y*.

На рис. 5.11 продемонстрирован способ визуализации набора данных, посвященного ожидаемой продолжительности жизни в 25 странах Северной, Центральной и Южной Америки. Этот показатель варьируется от 60 до 81 и для каждой страны обозначен синей точкой на определенном расстоянии по оси *x* от начала координат. Ограничение отображаемых на оси *x* значений до интервала от 60 до 81 позволяет выделить ключевые особенности этого набора данных. Как видно, жители Канады имеют самую высокую ожидаемую продолжительность жизни среди всех перечисленных стран, в то время как Боливия и Гаити сильно отстают от большинства. Если бы мы воспользовались столбцами вместо точек (рис. 5.12), то наша визуализация выглядела бы куда менее информативно. Поскольку все столбцы имеют довольно большую длину и при этом различия между ними минимальны, то внимание читателей наверняка будет сфокусировано на середине графика, нежели на его концах. Отсюда можно сделать вывод, что данная визуализация не справляется с поставленной перед ней задачей по передаче информации.

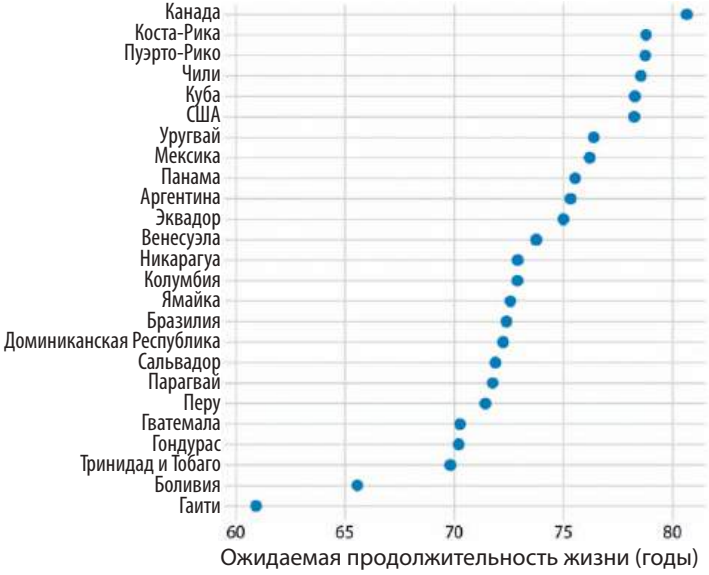


Рис. 5.11. Ожидаемая продолжительность жизни в странах Северной, Центральной и Южной Америки. Информация актуальна на 2007 год. Источник: Gapminder

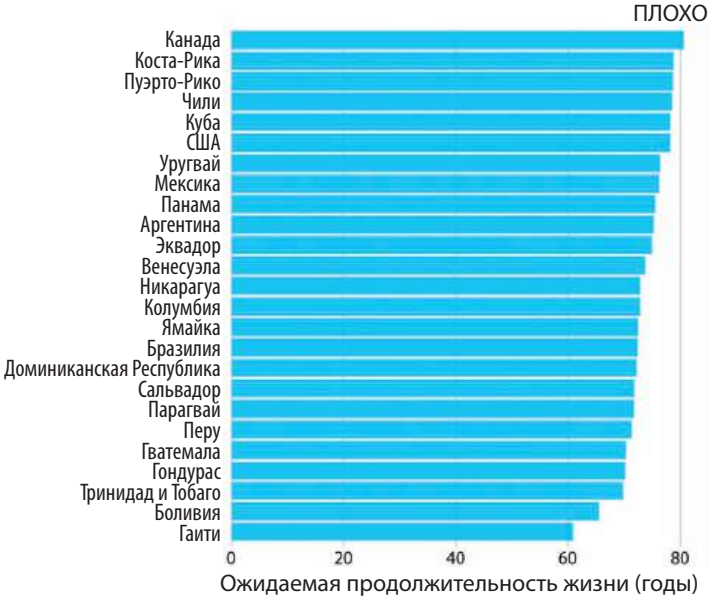


Рис. 5.12. Ожидаемая продолжительность жизни в странах Северной, Центральной и Южной Америки. Информация актуальна на 2007 год. Данные показаны в виде столбчатой диаграммы. Можно увидеть, что выбранный тип визуализации не подходит для данных. Столбцы имеют слишком большую длину, что отвлекает внимание от основного посыла графика — разницы в ожидаемой продолжительности жизни. Источник: Gapminder

Оформление графика может быть любым: столбцы или точки, однако куда более важным параметром является порядок следования величин. На рис. 5.11 и 5.12 страны расположены в порядке убывания ожидаемой продолжительности жизни. Если бы мы расположили эти страны в алфавитном порядке, изображение стало бы просто «кашей» из точек, сбивающей с толку и мешающей воспринимать информацию (рис. 5.13).

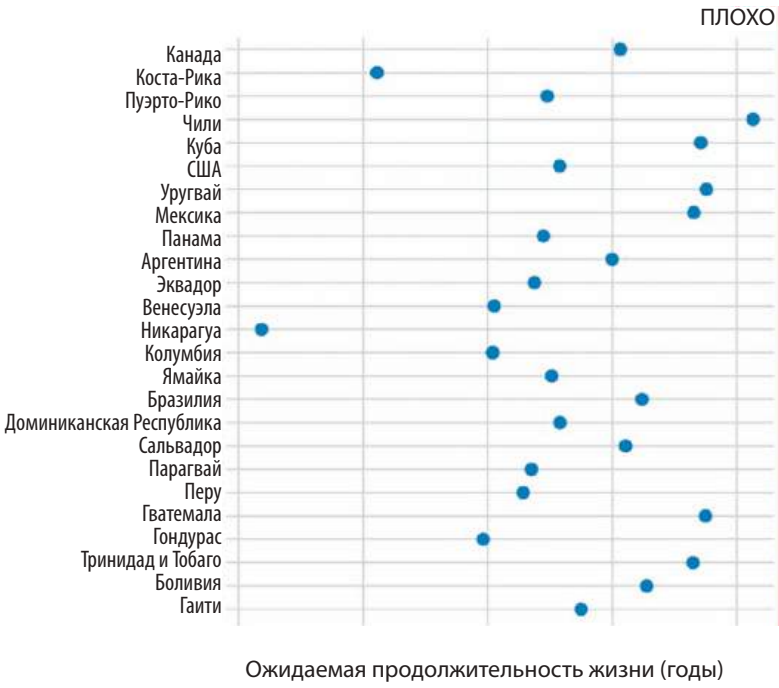


Рис. 5.13. Ожидаемая продолжительность жизни в странах Северной, Центральной и Южной Америки. Информация актуальна на 2007 год. На данном изображении страны расположены в случайном порядке, что делает из точек на графике неупорядоченное облако данных. Это ухудшает информативность изображения, и поэтому данная визуализация относится к категории «плохих». Источник: Gapminder

Во всех ранее приведенных примерах количественные данные были представлены на позиционных шкалах либо как длины столбцов, либо как расположения точек. Для больших наборов данных ни один из вариантов может оказаться непригоден, так как изображение будет сильно перегружено. С этой проблемой мы уже сталкивались: на рис. 5.7 всего семь групп из четырех значений данных превращают визуализацию в сложную и трудную для восприятия вещь. Если бы у нас было 20 групп по 20 значений данных, полученная в результате картина имела бы чрезвычайно запутанный вид.

В качестве альтернативы отображения значений данных посредством столбцов или точек мы можем отображать эти значения с помощью цвета.

Подобного рода визуализации называются *тепловыми картами*. Рис. 5.14 иллюстрирует этот подход на примере данных о проценте пользователей интернета для 20 стран на промежутке в 23 года, с 1994 по 2016 год. Несмотря на то что подобная визуализация не позволяет точно определить значения данных (например, какой конкретно процент пользователей был в США в 2015 году?), она отлично справляется с заданием показать тенденцию развития. На графике хорошо видно, какие страны начали пользоваться интернетом раньше всех и в какой стране наибольшее количество пользователей, имевших доступ в интернет за последний год, показанный на графике (2016).

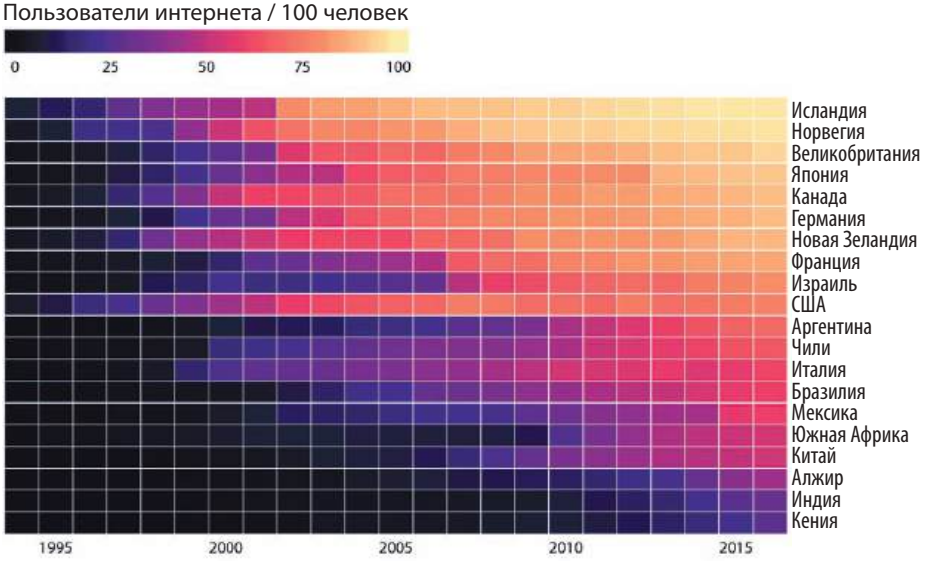


Рис. 5.14. Тенденция роста количества пользователей интернета в разных странах. Цветом показан процент пользователей в определенной стране в определенный год. Страны расположены в порядке убывания количества пользователей в 2016 году. Источник: World Bank

Как и в случае с другими подходами к визуализации из тех, что обсуждались в этой главе, при составлении тепловых карт нельзя упускать из виду порядок следования значений категориальных данных. На рис. 5.14 страны расположены в порядке убывания интернет-пользователей в 2016 году. Благодаря этому Великобритания, Япония, Канада и Германия находятся выше Соединенных Штатов, поскольку во всех этих странах в 2016 году использование интернета было шире, чем в США, несмотря на то что в Соединенных Штатах наблюдался значительный рост доли интернет-пользователей в более ранний период. В качестве альтернативы мы могли бы расположить страны в зависимости от того, как рано в них начался существенный рост

использования интернета. На рис. 5.15 страны упорядочены по годам, когда использование интернета впервые превысило 20%. На этом рисунке США занимают третью позицию сверху и выделяются относительно низким процентом использования интернета в 2016 году по сравнению со временем начала роста. Аналогичная картина наблюдается в Италии. Израиль и Франция, напротив, стартовали относительно поздно, но быстро заняли высокие места в рейтинге.

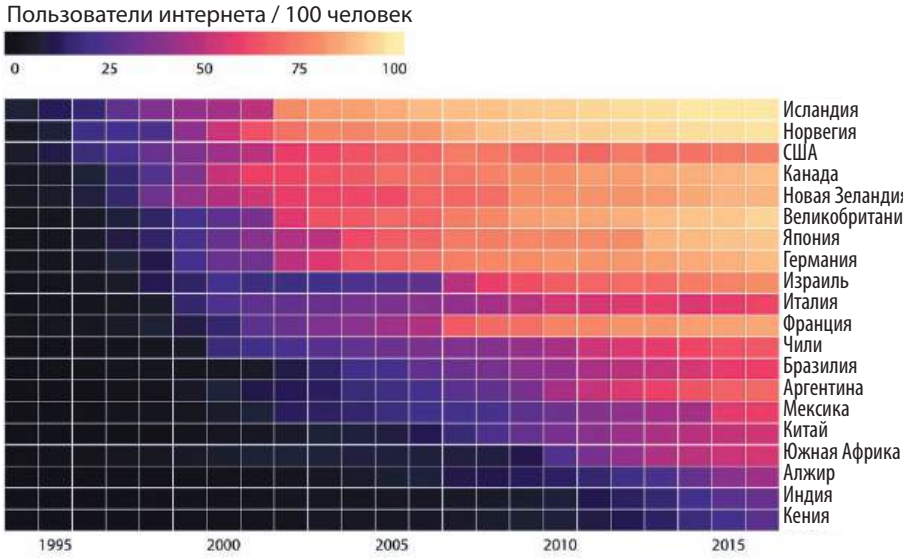


Рис. 5.15. Тенденция роста количества пользователей интернета в разных странах. Порядок сортировки стран выбран по году, когда процент пользователей впервые превысил 20% населения страны. Источник: World Bank

И рис. 5.14, и рис. 5.15 являются корректными способами представления данных. В зависимости от того, какую информацию нам нужно отобразить, мы можем выбрать любой из этих подходов. Если наша история посвящена тому, как широко была распространена сеть Интернет в 2016 году, то наш выбор — рис. 5.14. Если речь идет о зависимости текущих показателей использования интернета от момента начала распространения, наилучшим вариантом будет рис. 5.15.

Глава 6

Визуализация распределений: гистограммы и графики плотности

Одной из распространенных задач на практике является ситуация, когда нужно понять, как распределена та или иная переменная в наборе данных. Для примера вновь обратимся к массиву данных о пассажирах «Титаника», использовавшемуся в главе 5. Итак, на корабле присутствовало примерно 1300 пассажиров (не считая экипажа), возраст 756 из них известен. Пусть мы хотим узнать, сколько пассажиров какого возраста было на борту «Титаника», то есть сколько на корабле было детей, подростков, пассажиров среднего возраста, пенсионеров и т. д. Относительные пропорции возрастов пассажиров мы назовем *возрастным распределением (age distribution)*.

Визуализация одного распределения (Single Distribution)

Для получения информации о возрастном распределении пассажиров мы сгруппируем всех пассажиров по возрастам, а затем подсчитаем количество пассажиров в каждой возрастной группе. Результат данной процедуры приведен в табл. 6.1.

Мы можем визуализировать эту таблицу подобно столбчатой диаграмме: путем рисования закрашенных прямоугольников, высота которых соответствует количеству пассажиров, а ширина — возрастному диапазону (рис. 6.1). Подобного рода визуализации называются *гистограммами*. (Очень важно, чтобы все группы имели одинаковую ширину. Только в этом случае гистограмму можно считать корректной.)

Поскольку гистограммы создаются путем разбиения данных на интервалы, результат визуализации сильно зависит от выбранной ширины интервала. Большинство программ, используемых для визуализации, сами могут предложить некоторую ширину по умолчанию, однако стандартные настройки подходят далеко не для всех диаграмм.

Таблица 6.1. Количество пассажиров «Титаника», чей возраст известен

Возрастной диапазон	Количество
0–5	36
6–10	19
11–15	18
16–20	99
21–25	139
26–30	121
31–35	76
36–40	74
41–45	54
46–50	50
51–55	26
56–60	22
61–65	16
66–70	3
71–75	3

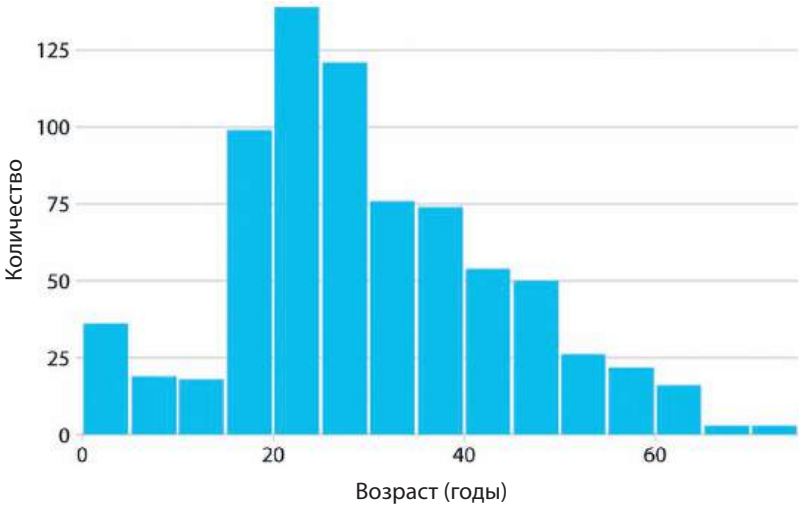


Рис. 6.1. Гистограмма возрастного распределения пассажиров «Титаника». Источник: Encyclopedia Titanica

Ширина интервала гистограммы играет крайне важную роль в правильном отображении визуализируемой информации. Если ширина будет слишком мала, гистограмма будет подобна частоколу и станет перегруженной,

а доносимая информация в ней — сложной для восприятия. Однако если ширина будет слишком велика, небольшие особенности гистограммы (например, резкое снижение количества пассажиров в промежутке от 5 до 15 лет) могут потеряться.

Говоря о возрастном распределении пассажиров «Титаника», видно, что единичный отрезок возраста в 1 год слишком мал, а отрезки в 15 лет, наоборот, слишком велики. Размер единичных отрезков от 3 до 5 является идеальным для данной визуализации (рис. 6.2).

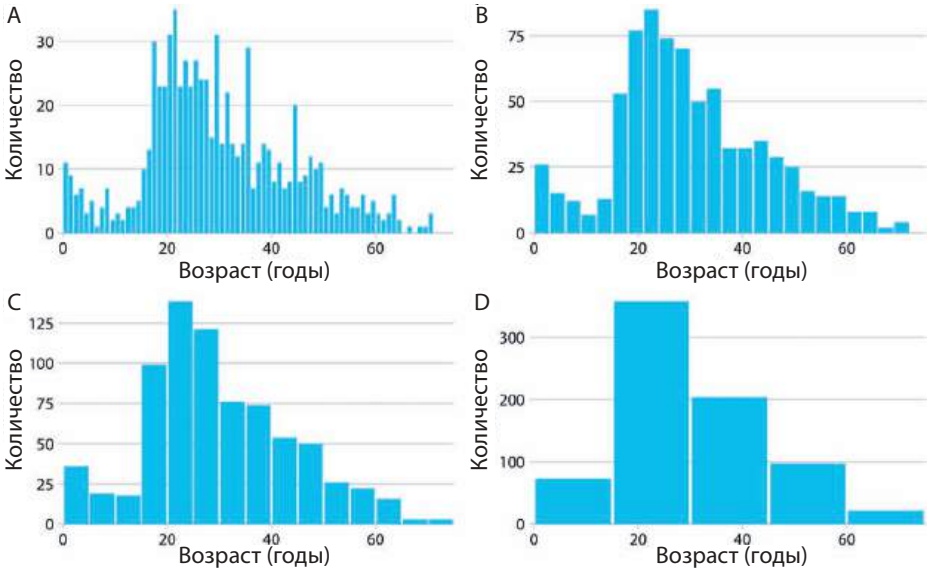


Рис. 6.2. Гистограммы с различной шириной интервала. На данном изображении представлены возрастные распределения пассажиров «Титаника», созданные с использованием различных значений ширины интервала: A) 1 год; B) 3 года; C) 5 лет; D) 15 лет. Источник: Encyclopedia Titanica

Создавая гистограмму, всегда экспериментируйте с шириной интервала.

Гистограммы были одним из наиболее популярных способов визуализации информации начиная с XVIII столетия, если не раньше. Отчасти так произошло потому, что график этого типа прост в построении и не требует никаких специальных средств. В настоящее время, когда вычислительные устройства стали доступны широкой публике, а их мощность достигла высокого уровня, гистограммы стали заменяться *графиками плотности*

распределения. При создании графика плотности мы пытаемся визуализировать вероятностное распределение данных путем рисования подходящей непрерывной кривой (рис. 6.3).

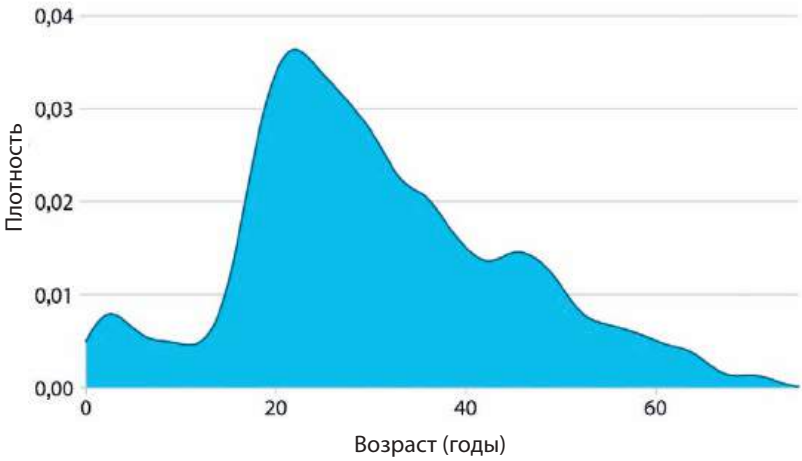


Рис. 6.3. Ядерная оценка плотности возрастного распределения пассажиров «Титаника». Масштаб графика выбран такой, что площадь фигуры под ним равна 1. Оценка плотности проведена с использованием функции Гаусса и пропускной способности, равной 2. Источник: Encylopedia Titanica

Значения для построения кривой вычисляются на основе имеющихся данных, а сам процесс вычисления называется *ядерной оценкой плотности.* В процессе мы рисуем непрерывные кривые (так называемые «ядра») малой ширины (где ширина управляется параметром под названием *пропускная способность,* или *полоса пропускания*) в каждой точке данных, после чего объединяем все эти кривые для получения конечной оценки плотности. Наиболее часто в ядерной оценке плотности используют гауссово ядро (кривую плотности распределения Гаусса), но существует и много других вариантов.

Как и в случае с гистограммами, внешний вид графика плотности определяется ядерной функцией и шириной полосы (рис. 6.4). Величина пропускной способности по своей сути подобна размеру ячейки для гистограммы: если ее размер будет слишком мал, график плотности может приобрести вид частокола и стать перегруженным, при этом основной посыл графика пройдет мимо внимания читателя. А если пропускная способность будет слишком велика, мелкие особенности распределения данных будут незаметны. Также стоит отметить, что выбор функции распределения оказывает большое влияние на форму кривой плотности. К примеру, функция Гаусса характеризуется «фирменным» стилем, состоящим из плавных линий и имеющим «хвосты» в начале и конце графика. А, например, прямоугольная функция

распределения сделает кривую плотности распределения похожей на лесенку (рис. 6.4D). В целом чем больше данных представлено на графике, тем меньшее значение имеет выбор ядерной функции. Вообще, из вышеописанного следует, что использование графиков плотности имеет смысл лишь при большом объеме набора данных, а в случае, если массив данных содержит всего несколько точек, график плотности легко может ввести в заблуждение.

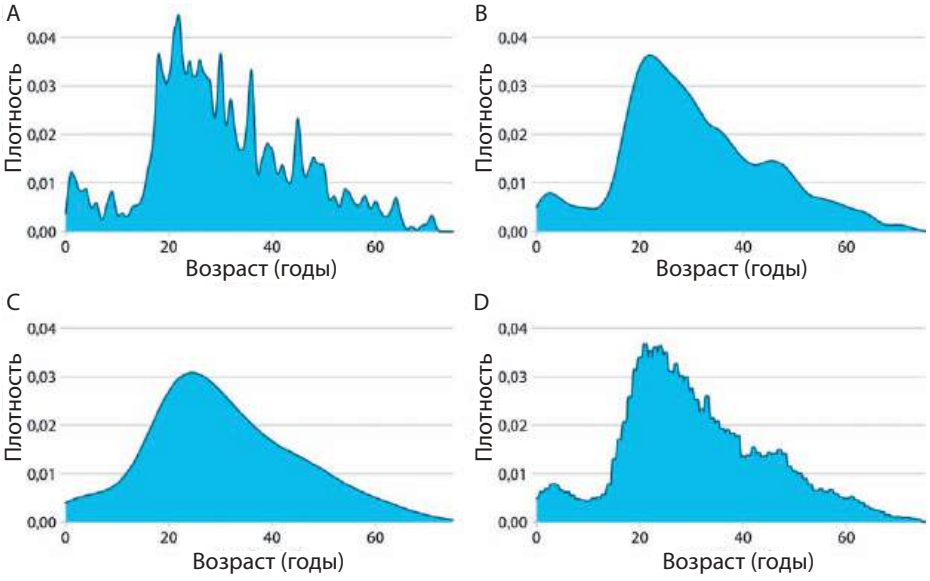


Рис. 6.4. Ядерные оценки плотности, различающиеся величиной пропускной способности и функцией распределения. На данном изображении показано возрастное распределение пассажиров «Титаника» для четырех различных комбинаций следующих параметров: A) гауссово ядро; пропускная способность равна 0,5; B) гауссово ядро; пропускная способность равна 2; C) гауссово ядро; пропускная способность равна 5; D) прямоугольное ядро; полоса пропускания равна 2. Источник: Encyclopedia Titanica

Кривые плотности обычно строятся таким образом, чтобы площадь под графиком равнялась 1. Из-за этого правила значения оси *y* могут казаться весьма странными, поскольку они целиком зависят от значений по оси *x*. К примеру, в случае возрастного распределения пассажиров «Титаника» диапазон данных по оси *x* начинается в 0 и заканчивается на отметке в 75. Таким образом, средняя высота кривой плотности должна составлять $1/75 \approx 0{,}013$. Если взглянуть на кривые возрастного распределения (то есть рис. 6.4), можно заметить, что значения по оси *y* находятся в промежутке от 0 до примерно 0,04, а среднее — где-то в районе отметки 0,01.

При использовании ядерной оценки плотности важно не забывать об одной особенности, присущей данному типу графиков: они имеют тенденцию

создавать видимость наличия данных там, где их нет, особенно в «хвостах» распределений. Поэтому необдуманное использование кривых плотности может привести к появлению изображений, не имеющих никакого смысла. Достаточно небольшой оплошности при создании визуализации возрастного распределения, и в нее запросто могут попасть отрицательные значения возрастов (рис. 6.5).

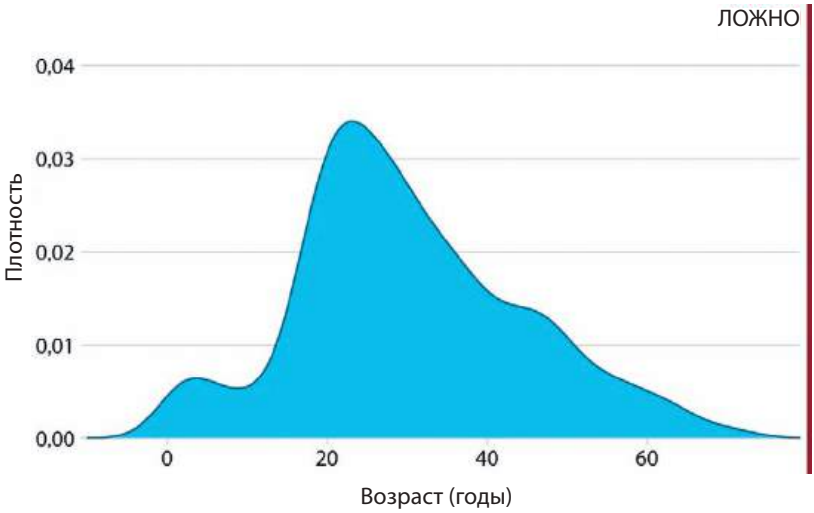


Рис. 6.5. Ядерная оценка плотности может продлевать «хвосты» распределения в тех областях, где данных быть не должно. На этом изображении график плотности возрастного распределения, построенный на основе гауссова ядра, не был «обрублен» в нулевой точке (как это сделано на рис. 6.4), а продлен до отрицательных значений, что, очевидно, не имеет смысла. Подобных ошибок при создании визуализаций необходимо избегать. Источник: Encyclopedia Titanica

Итак, в каких же случаях следует выбирать гистограмму, а в каких — график плотности? На эту тему можно провести немало жарких дискуссий. Некоторые специалисты не признают использование графиков плотности и считают их условными и необъективными. Другие осознаю́т, что гистограммы тоже могут быть условными и необъективными. Я же придерживаюсь мнения, что выбор между этими видами визуализации — дело вкуса, однако в зависимости от ситуации тот или иной тип графика способен лучше отобразить интересующие нас свойства данных, нежели его «собрат». А бывает и так, что не подходит ни один из них, и в таких случаях лучше будет воспользоваться эмпирической интегральной функцией распределения или графиками типа «квантиль-квантиль» (см. главу 7). В завершение хочу все же отметить, что я считаю, что визуализации оценок плотности имеют преимущество перед гистограммами в случаях, когда необходимо отобразить более одного распределения за раз.

Визуализация нескольких распределений одновременно

На практике нередко встречаются ситуации, когда имеется несколько распределений, которые нужно визуализировать вместе. Например, пусть нам нужно узнать, как распределяется возраст пассажиров «Титаника» в зависимости от их пола. Мужчины и женщины были приблизительно одного возраста или между ними была некоторая разница? Одним из способов это узнать является создание гистограммы с накоплением. Для создания такого типа графика мы располагаем столбцы гистограммы с данными о женщинах над столбцами с данными о мужчинах и раскрашиваем их в разные цвета (рис. 6.6).

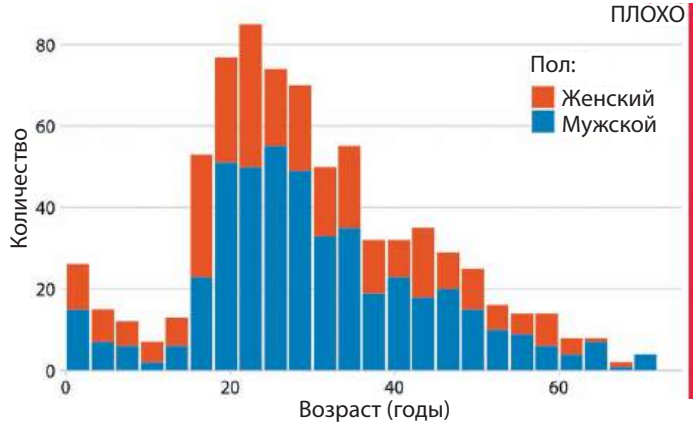


Рис. 6.6. Гистограмма возрастов пассажиров «Титаника» с использованием наслоения в зависимости от пола. Данное изображение было отмечено как «плохое», потому что гистограмму с накоплением очень легко перепутать с гистограммой с наложением (рис. 6.7). Кроме того, высоты столбцов, показывающие значения о женщинах-пассажирах, сложно сравнить между собой. Источник: Encyclopedia Titanica

На мой взгляд, данного типа визуализации следует избегать. На это есть как минимум две причины. Первая состоит в том, что при беглом взгляде на фигуру очень сложно понять, где по вертикали начинаются столбцы ряда, отображающего данные о женщинах: в нуле или там, где заканчиваются столбцы с данными о мужчинах? Другими словами, на корабле было 25 женщин в возрасте 18–20 лет или почти 80? Правильный — первый ответ. Вторая проблема — столбцы, показывающие количество женщин на борту, нельзя сравнить друг с другом из-за того, что все они начинаются на разной высоте. Например, фактически средний возраст мужчин больше среднего возраста женщин, однако из рис. 6.6 понять это нельзя.

Можно попробовать избежать этих проблем, если начать столбцы с нуля и сделать их частично прозрачными (рис. 6.7).

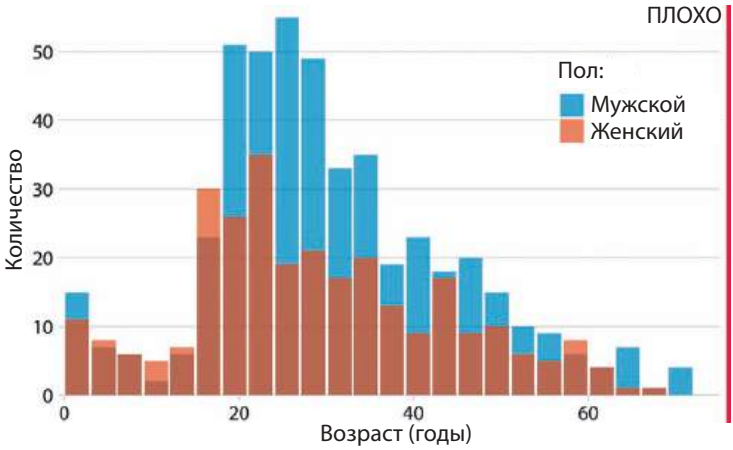


Рис. 6.7. Возрастное распределение женщин и мужчин — пассажиров «Титаника», показанное в виде двух наложенных друг на друга гистограмм. Это изображение относится к категории «плохих», так как неясно, все ли синие столбцы начинаются с отметки 0. Источник: Encyclopedia Titanica

Однако и у этого подхода есть недостатки. Теперь на нашем изображении вместо двух групп три, и мы все еще не знаем, где начинается и заканчивается каждый из столбцов. Накладывающиеся диаграммы здесь тоже не подходят, потому что два полупрозрачных столбца, наложенные друг на друга, больше похожи на непрозрачный столбец совершенно другого цвета.

А вот накладывающиеся графики плотности лишены этого недостатка, потому что непрерывные линии плотности позволяют глазу легко отличать одно распределение от другого. Как видно из графика, в этом наборе данных возрастные распределения для пассажиров мужского и женского пола практически идентичны до возраста 17 лет, но после этой точки расходятся. Таким образом, данная визуализация все еще не полностью достигает нужной нам цели (рис. 6.8).

Идеальное визуализационное решение для этого набора данных заключается в том, чтобы по отдельности показать возрастное распределение пассажиров и пассажирок «Титаника» с учетом соблюдения пропорций относительно общего количества пассажиров (рис. 6.9). Эта визуализация наглядно показывает, что в возрастной группе 20–50 лет женщин на «Титанике» было гораздо меньше, чем мужчин.

Ну и наконец, если мы хотим точно визуализировать два возрастных распределения, можно сделать две отдельные гистограммы, повернуть их на 90 градусов, а потом сделать так, чтобы столбцы отходили от линии их соприкосновения и расходились в противоположные стороны. Данный метод часто используется при визуализации возрастных распределений, а такой график называется *половозрастная пирамида* (рис. 6.10).

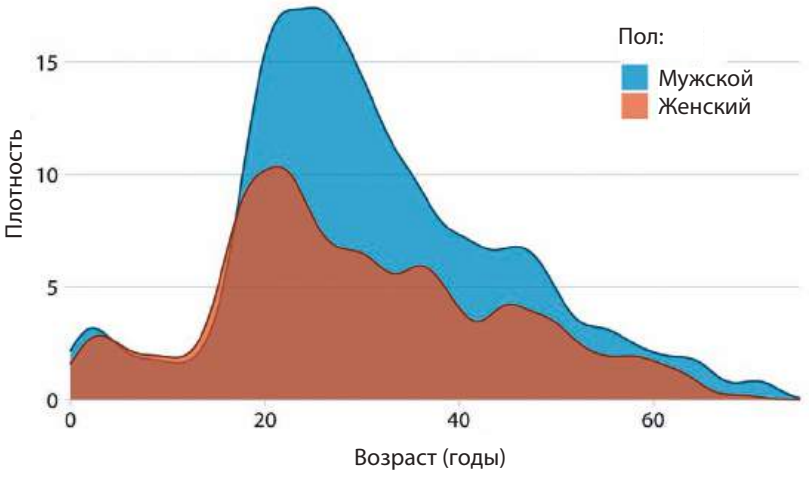


Рис. 6.8. Ядерные оценки плотности возрастных распределений женщин и мужчин — пассажиров «Титаника». Для отражения того факта, что мужчин на корабле было значительно больше, чем женщин, кривые плотности были сформированы таким образом, что площади под каждой из них соответствуют общим числам мужчин и женщин, чей возраст нам известен (468 и 288 соответственно). Источник: Encyclopedia Titanica

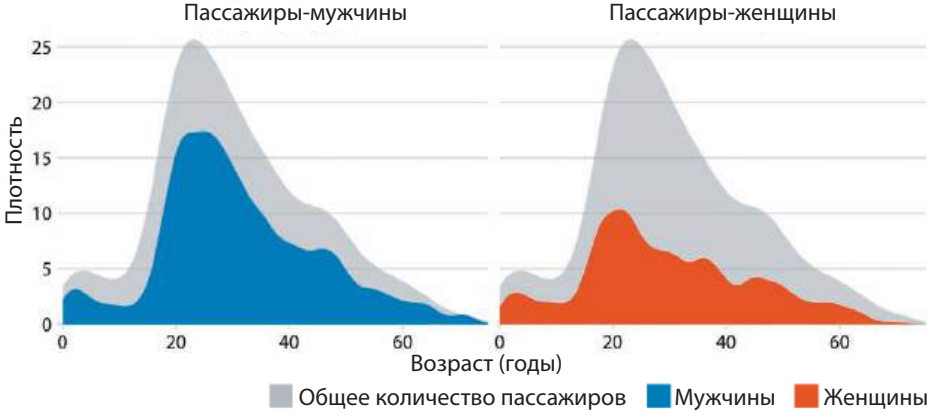


Рис. 6.9. Возрастные распределения мужчин и женщин — пассажиров «Титаника», представленные как доли относительно общего количества пассажиров. Раскрашенные области графиков показывают ядерные оценки плотности пассажиров-мужчин и пассажиров-женщин соответственно. Серые области показывают общее возрастное распределение пассажиров. Источник: Encyclopedia Titanica

Важно понимать, что данный подход не сработает в случае визуализации более чем двух распределений. В случае нескольких распределений гистограммы сильно сбивают с толку, а графики плотности работают хорошо только в том случае, если наборы данных не пересекающиеся, но смежные.

Например, графики плотности с наложением отлично подходят для визуализации данных о жирности молока, производимого четырьмя различными породами коров (рис. 6.11).

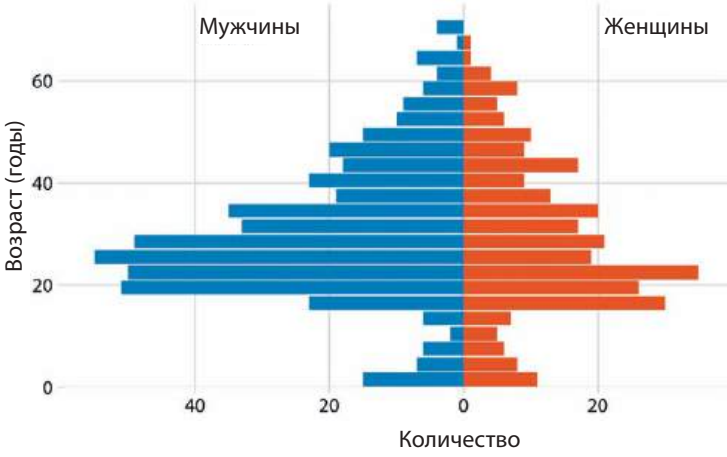


Рис. 6.10. Возрастное распределение мужчин и женщин — пассажиров «Титаника», визуализированное в виде половозрастной пирамиды. Источник: Encyclopedia Titanica

Ядерные оценки плотности являются более подходящим методом для одновременной визуализации нескольких распределений, нежели гистограммы.

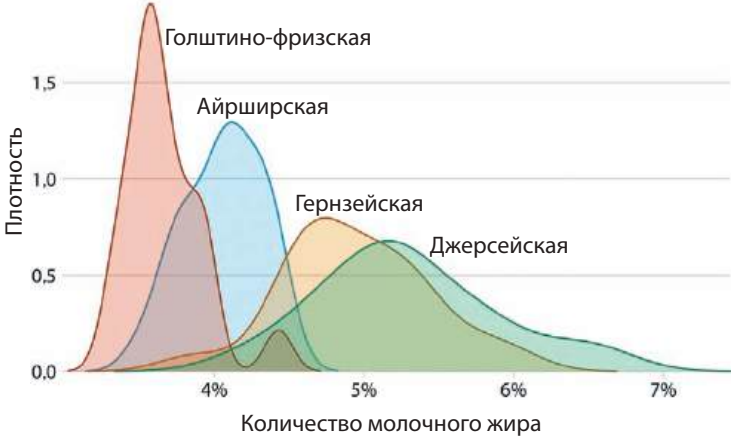


Рис. 6.11. График оценки плотности распределения процента молочного жира в молоке, производимом четырьмя различными породами коров. Источник: Canadian Record of Performance for Purebred Dairy Cattle

Глава 7

Визуализация распределений: функции распределения и графики «квантиль-квантиль»

В главе 6 мы разобрались, как можно визуализировать распределения при помощи гистограмм и графиков плотности. Оба этих варианта являются интуитивно понятными и визуально привлекательными. Поговорили мы и о недостатках этих методов, одним из которых является большое количество параметров, которые необходимо учитывать: в частности, ширину столбцов гистограммы и пропускную способность графиков плотности. В результате можно сделать вывод, что эти методы скорее интерпретируют данные, а не визуализируют их.

Одной из альтернатив описанным выше способам является показ всех точек по отдельности, в виде облака. Однако в ситуации очень больших наборов данных такой подход становится громоздким, а ценность методов визуализации агрегированных данных все-таки состоит в том, что они подчеркивают особенности распределения, а не отдельных точек. Для решения данной проблемы были придуманы визуализации эмпирических функций распределения и графики «квантиль-квантиль». Эти визуализации не требуют подбора значений параметров, плюс они показывают все данные одновременно. Но, к сожалению, такие графики не настолько интуитивно понятны, как гистограммы или графики плотности, да и используют их где-то кроме технической литературы очень редко. Данные типы визуализаций пользуются большой популярностью в основном у ученых-статистиков, но мне кажется, что об этих методах должен знать каждый, кто интересуется визуализацией данных.

Функции распределения

Чтобы продемонстрировать, что же такое функция распределения, я воспользуюсь примером, с которым как преподаватель сталкиваюсь очень часто: набором данных с оценками успеваемости студентов. Пусть в нашем классе 50 студентов и все они только что сдали экзамен, оценка за который колеблется

в пределах от 0 до 100 баллов. Как лучше всего визуализировать успеваемость учащихся, чтобы, например, определить подходящие границы оценок?

Можно выбрать следующий подход: построить график, где каждому возможному значению набранных баллов (в нашем случае — от 0 до 100) будет поставлено в соответствие количество студентов, набравших ровно столько баллов или меньше. Такой график будет выглядеть как растущая линия, исходящая из 0 для 0 баллов и заканчивающаяся в 50 для 100 баллов. Или можно сделать по-другому: ранжировать всех учеников по количеству набранных баллов в порядке возрастания (так ученик с наименьшим количеством баллов занимает низшую позицию, а ученик с наибольшим количеством баллов — высшую), а затем построить график зависимости позиции студента в общем рейтинге от реально набранных им баллов. В результате получится эмпирическая функция распределения. Каждая точка представляет одного студента, а линии отображают самую высокую позицию в рейтинге класса для каждой возможной суммы набранных баллов (рис. 7.1).

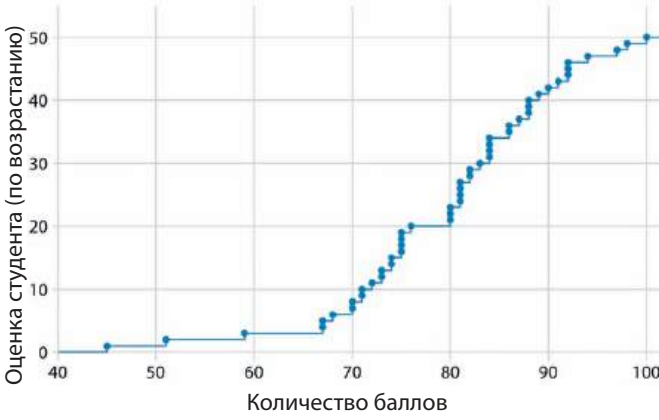


Рис. 7.1. Эмпирическая функция распределения оценок гипотетического класса из 50 человек

Вам, наверное, интересно, что будет, если ранжировать студентов не в порядке возрастания оценки, а наоборот. В данном случае график просто перевернется вверх ногами: в результате получится все та же эмпирическая функция распределения, однако теперь для каждого возможного значения набранного числа баллов она будет показывать низшую наблюдаемую позицию в рейтинге (рис. 7.2).

Возрастающие функции распределения используются значительно чаще, чем убывающие. Однако не стоит забывать, что у каждой из них есть свои преимущества. В следующем разделе мы увидим, что убывающие функции распределения хорошо подходят для случая визуализации сильно искаженных распределений.

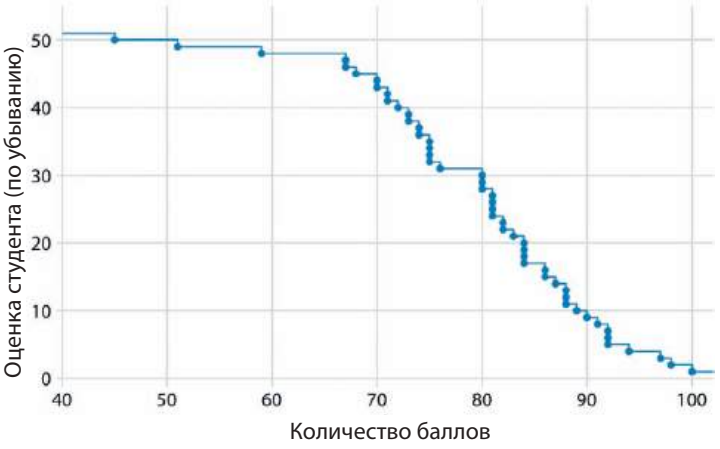


Рис. 7.2. Убывающая эмпирическая функция распределения оценок гипотетического класса из 50 человек

Весьма распространенной практикой является определение функции распределения без выделения индивидуальных точек с нормализацией значений по максимальному. В таком случае ось *y* будет показывать накопленную частоту получения определенного количества баллов или меньше (рис. 7.3).

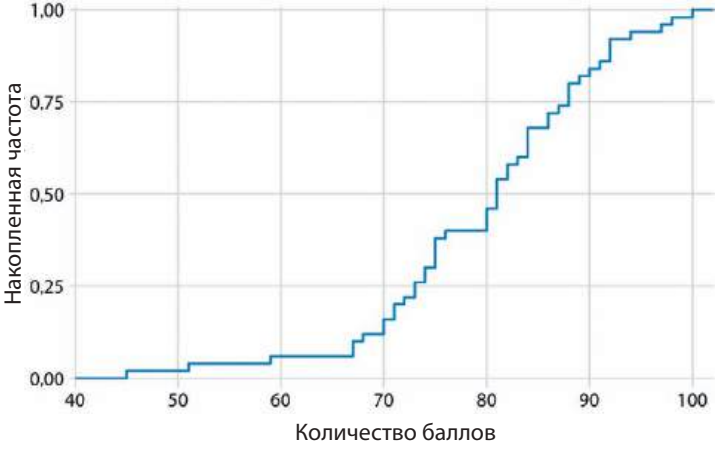


Рис. 7.3. Эмпирическая функция распределения оценок гипотетического класса из 50 человек. Положения студентов в рейтинге группы нормализованы по отношению к общему числу студентов таким образом, что значения по оси *y* соответствуют доле студентов в классе, набравших указанное по оси *x* или меньшее количество баллов

Этот график позволяет быстро считать основные свойства распределения результатов студентов. Например, примерно четверть из них (25%) получили менее 75 баллов. Медианное значение (точка, соответствующая накопленной частоте 0,5) равно 81. Примерно 20% студентов получили 90 и более баллов.

Я нахожу эмпирические функции распределения очень полезными с точки зрения распределения оценок внутри студенческих групп, так как они дают конкретные границы отсечения, которые помогают минимизировать общее расстройство. К примеру, в наших данных на графике видна достаточно длинная горизонтальная линия перед результатом в 80 баллов и резкий подъем сразу после него. Это вызвано тем, что сразу трое студентов набрали 80 баллов на экзамене, а предыдущий результат составил 76 баллов. Так, я бы поставил всем набравшим 80 баллов и выше оценку 4, а набравшим 79 и менее — оценку 3. Три студента с 80 баллами были бы очень довольны оценкой 4, а студенты, набравшие 76 и меньше, увидели бы, что до следующей оценки им еще работать и работать. Если бы я решил, например, сделать порог отсечения в 77 баллов, то студент с 76 баллами наверняка бы пришел ко мне в кабинет в надежде договориться. То же самое бы произошло, если бы я установил границу в 81 балл: трое 80-балльников точно бы попытались каким-нибудь образом получить дополнительный балл.

Сильно искаженные распределения

Многие эмпирические наборы данных представляют собой сильно искаженные распределения, в частности с тяжелыми хвостами справа, и визуализировать эти распределения бывает нелегко. Примерами таких распределений являются: количество людей, живущих в разных городах или округах, количество контактов в социальных сетях, частота появления отдельных слов в книге, количество научных работ, написанных различными авторами, чистые активы отдельных лиц, а также количество партнеров по взаимодействию отдельных белков в сетях белкового взаимодействия [Clauset, Shalizi, and Newman, 2009]. Все эти распределения объединяет то, что правый хвост затухает медленнее, чем экспоненциальная функция. На практике это означает, что очень большие значения не так уж и редки, даже если математическое ожидание распределения невелико. Одним из широких классов таких распределений являются степенные распределения, где вероятность наблюдения значения, в x раз превышающего некоторую опорную точку, уменьшается со скоростью степени x. Например, давайте рассмотрим чистые активы в США, распределение которых соответствует степенному закону с параметром, равным 2. Если взять любую сумму располагаемых активов (скажем, 1 миллион долларов США), то люди, которые владеют активами в половину такой суммы, встречаются в четыре раза чаще, а люди с вдвое большим объемом чистых активов — в четыре раза реже. Важно отметить, что эта зависимость сохраняется, если в качестве точки отсчета мы используем 10 000 долларов или даже 100 миллионов долларов. По этой причине степенные распределения также иногда называются масштабно-инвариантными.

Для иллюстрации вышесказанного рассмотрим и визуализируем следующий набор данных: количество людей, живущих в разных округах США, по данным переписи 2010 года. Это распределение обладает крайне тяжелым правым хвостом: несмотря на то что в большинстве округов численность жителей относительно невысока (медиана составляет 25 857 человек), некоторые из них весьма густонаселены (например, в округе Лос-Анджелес население составляло 9 818 605 человек). Если мы попытаемся визуализировать распределение количества жителей по округам в виде графика плотности или функции распределения, мы получим изображения, толку от которых будет немного (рис. 7.4).

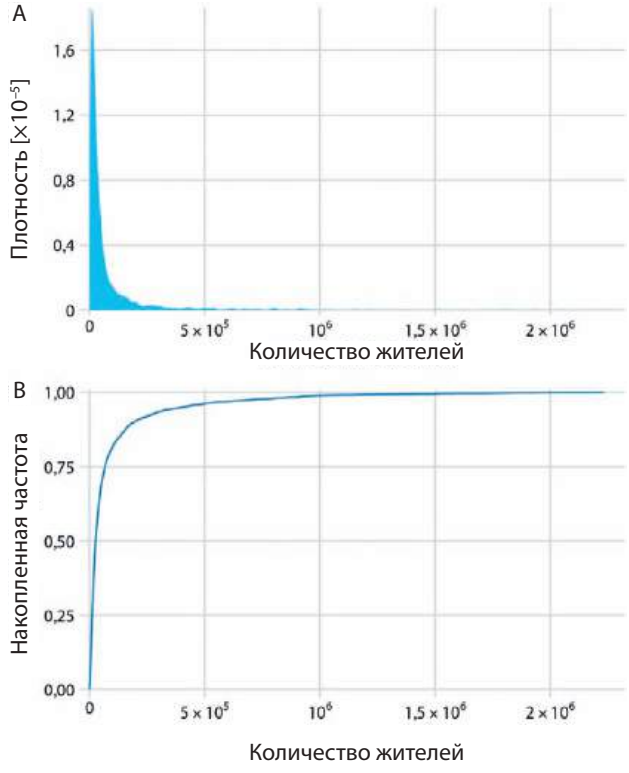


Рис. 7.4. Распределение числа жителей округов Соединенных Штатов Америки: A) график плотности; B) эмпирическая функция распределения. Источник: Перепись населения США, 2010 год

График плотности (рис. 7.4A) показывает резкий пик в ближайшей окрестности нуля, но каких-либо деталей распределения практически не видно. Аналогично функция распределения (рис. 7.4B) демонстрирует резкий рост в окрестности нуля, а детали распределения вновь теряются. Чтобы визуализировать этот конкретный набор данных, мы можем преобразовать данные

в логарифмические значения и нанести на график уже их. Эта трансформация работает здесь по той причине, что распределение численности населения в округах соответствует не степенному закону, а является почти идеальным логнормальным распределением (см. раздел «Графики “квантиль-квантиль” на с. 86). И на самом деле, график плотности преобразованных по логарифму значений выглядит как красивая кривая в виде колокола, а соответствующая функция распределения имеет форму почти идеальной сигмоидальной кривой (рис. 7.5).

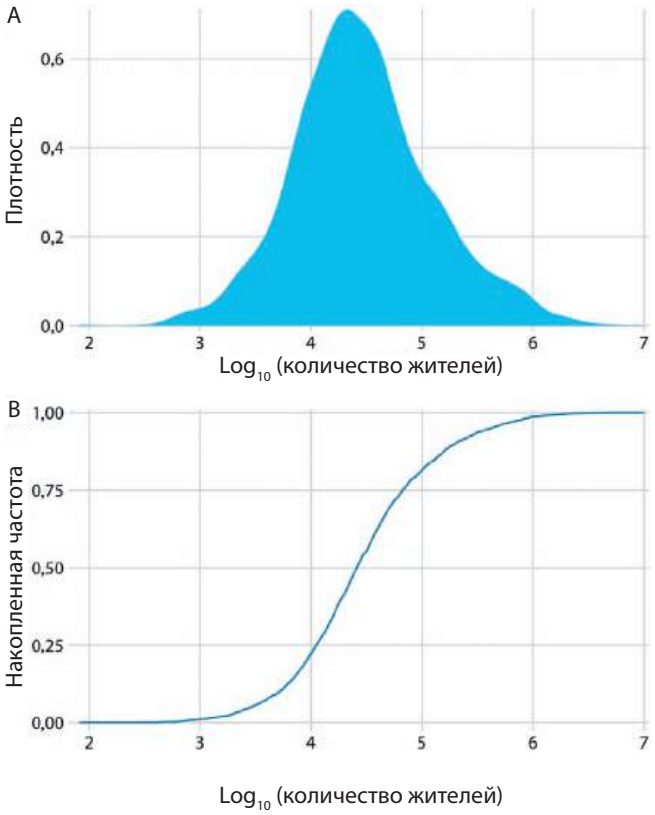


Рис. 7.5. Распределение логарифма численности населения в округах Соединенных Штатов Америки. A. График плотности. B. Функция распределения. Источник: Перепись населения США, 2010 год

Убедиться в том, что это распределение не является степенным законом, нам поможет нисходящий график эмпирической функции распределения с логарифмическими осями x и y. В такой визуализации степенной закон выглядит как идеальная прямая линия. Что касается численности населения в округах, то на нисходящем графике функции распределения лишь правый хвост образует почти идеальную прямую линию (рис. 7.6).

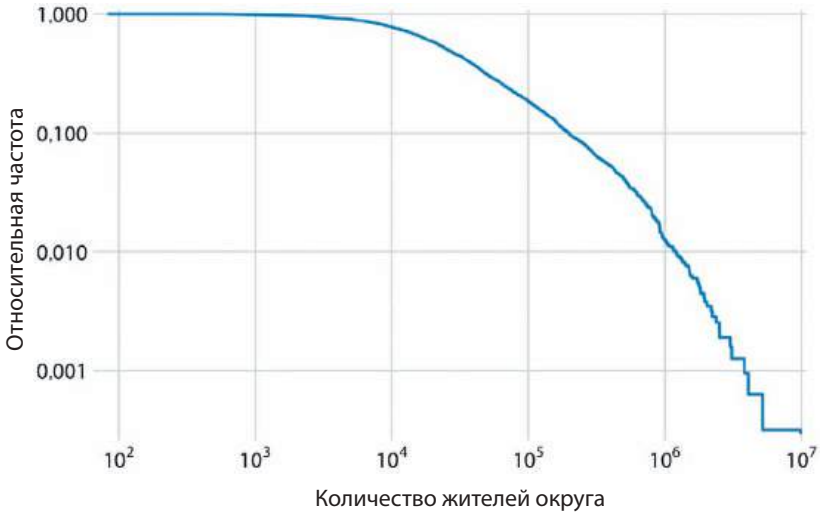


Рис. 7.6. Относительная частота округов с таким же или меньшим количеством жителей по отношению к количеству жителей округов. Источник данных: Перепись населения США, 2010 год

Рассмотрим еще один пример — частотное распределение слов в романе «Моби Дик». Вот это распределение идеально соответствует степенному закону: если построить нисходящий график функции распределения с логарифмическими осями, то мы увидим почти идеальную прямую линию (рис. 7.7).

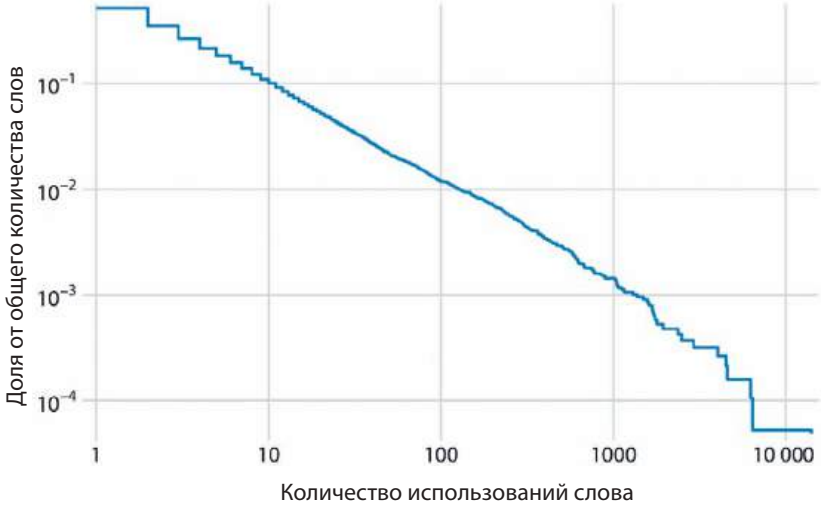


Рис. 7.7. Частотное распределение слов в романе «Моби Дик». Показана относительная частота слов, многократно встречающихся в романе, по отношению к общему количеству имеющихся в книге слов. Источник: [Clauset, Shalizi, and Newman, 2009]

Графики «квантиль-квантиль»

Графики «квантиль-квантиль» (q-q) являются полезным способом визуализации в тех случаях, когда нам нужно определить, насколько данные, полученные в результате наблюдений, соответствуют или не соответствуют заданному распределению. Как и при построении функций распределения, графики «квантиль-квантиль» также основаны на ранжировании данных и визуализации связи между рангами и фактическими значениями. Однако в этом случае мы вместо отображения рангов на графике используем их для прогнозирования того, куда попадет данная точка, если данные будут распределены в соответствии с заданным распределением. Чаще всего q-q-графики строятся с использованием нормального распределения в качестве эталонного. Чтобы привести конкретный пример, предположим, что фактические значения данных имеют математическое ожидание 10 со стандартным отклонением 3.

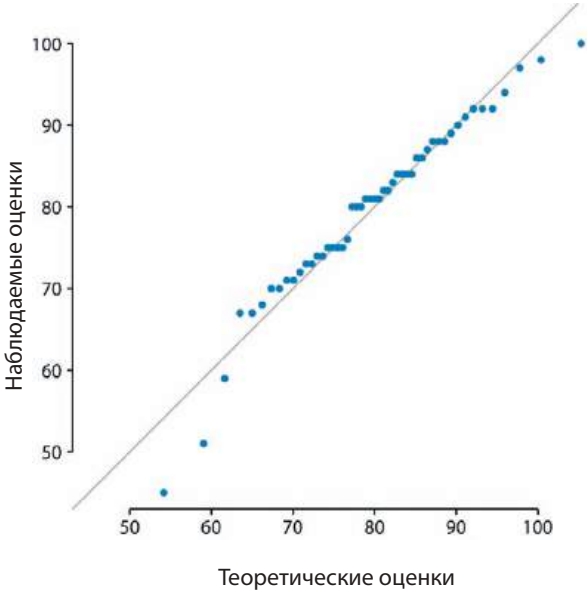


Рис. 7.8. График «квантиль-квантиль» гипотетических оценок студентов

Предполагая, что наши данные распределены по нормальному закону, мы ожидаем, что в ранжированных данных точка, попавшая ровно в 50-й процентиль, будет иметь значение 10 (матожидание), точка на 84-м процентиле — значение 13 (на одно стандартное отклонение выше матожидания), а точка в 2,3-м процентиле будет иметь значение 4 (на два стандартных отклонения ниже матожидания). Эти расчеты мы можем выполнить для всех точек набора данных и затем отобразить на графике наблюдаемые значения относительно их теоретических эквивалентов (значений, которые должны

принимать точки, находящиеся на тех же позициях, если ранжировать значения теоретического распределения). Применив эту процедуру к массиву данных об оценках студентов из начала этой главы, мы получим рис. 7.8.

Сплошная линия здесь не является линией регрессии, а лишь указывает на точки, где x равняется y, то есть где наблюдаемые значения совпадают с теоретическими. В той степени, в какой точки попадают на эту линию, эмпирическое распределение данных и соответствует теоретическому (здесь — нормальному). Мы видим, что оценки учеников в основном распределяются нормально, с небольшими отклонениями в нижней и верхней части (несколько учеников показали результаты хуже, чем ожидалось, — на обоих концах). Отклонения от распределения в верхней части обусловлены максимальным значением в 100 баллов на гипотетическом экзамене: независимо от того, насколько хорош лучший студент, он не сможет получить больше 100 баллов.

График «квантиль-квантиль» можно использовать и для проверки моего предположения, сделанного ранее в этой главе: распределение населения по округам США соответствует логнормальному распределению. Если эти значения логнормально распределены, то преобразованные по логарифму значения являются нормально распределенными и поэтому должны находиться прямо на линии $x = y$. При построении этого графика мы видим исключительное соответствие между наблюдаемыми и теоретическими значениями (рис. 7.9), что свидетельствует о том, что распределение населения в округах США действительно является логнормальным.

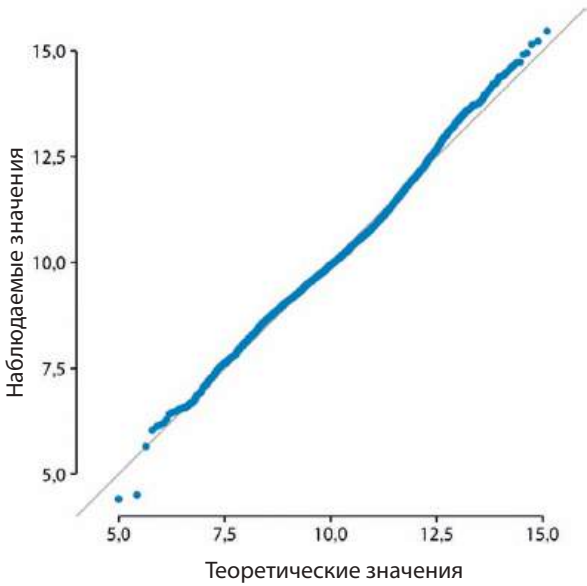


Рис. 7.9. График «квантиль-квантиль» логарифма числа жителей в округах США. Источник: Перепись населения США в очередном десятилетии, 2010 год

Глава 8

Одновременная визуализация множества распределений

Существует множество сценариев, в которых требуется визуализация нескольких распределений одновременно: например, это могут быть метеорологические данные. Может быть, нам захочется узнать, как меняется температура в разные месяцы, а также показать распределение наблюдаемых температур для каждого месяца. Этот сценарий требует одновременного отображения дюжины распределений (температуры), по одному для каждого месяца. Ни одна из визуализаций, рассмотренных в главах 6 и 7, тут не сработает. Здесь нам понадобятся такие способы отображения данных, как коробчатая диаграмма, скрипичный график и график «горный хребет».

Всякий раз, когда мы имеем дело с большим количеством распределений, стоит подумать о наличии объясняемой переменной и одной или нескольких группирующих переменных. Объясняемая переменная — это та величина, чье распределение мы хотим показать. Группирующие переменные определяют подмножества данных с разными распределениями объясняемой переменной. Например, для распределения температуры по месяцам объясняемая переменная — это температура, а группирующей переменной является месяц. Все техники, рассмотренные в этой главе, располагают объясняемую переменную вдоль одной оси, а группирующие переменные — вдоль другой. Ниже я сначала опишу подходы, в которых объясняемую переменную располагают вдоль вертикальной оси, а затем подходы, которые располагают ее по горизонтальной оси. В каждом из этих случаев мы можем поменять оси местами и получить альтернативную жизнеспособную визуализацию. Здесь же я продемонстрирую канонические формы различных визуализаций.

Визуализация распределений вдоль вертикальной оси

Самый простой подход к одновременному отображению большого количества распределений заключается в отображении их математических ожиданий (или медиан) в виде точек, с указанием отклонений от средних в виде планок

погрешностей. На рис. 8.1, где показано распределение месячных температур в Линкольне (штат Небраска) в 2016 году, используется именно такой подход. Я отнес это изображение к категории «плохих», потому что у этого подхода множество недостатков. Во-первых, представляя каждое распределение только одной точкой и двумя границами диапазона погрешности, мы теряем значительную часть данных. Во-вторых, сразу понять, что представляют собой эти точки, невозможно, хотя большинство читателей, скорее всего, догадаются, что перед ними либо медиана, либо средняя величина. В-третьих, совершенно неочевидно, что означают планки погрешностей. Представляют ли они среднеквадратическое отклонение данных, среднеквадратическую ошибку, 95%-ный доверительный интервал или что-либо еще? Общепринятого стандарта не существует. Читая подписи к рис. 8.1, мы видим, что на этой диаграмме отображается двойное среднеквадратическое отклонение средних суточных температур, что дает доверительную вероятность порядка 95%.

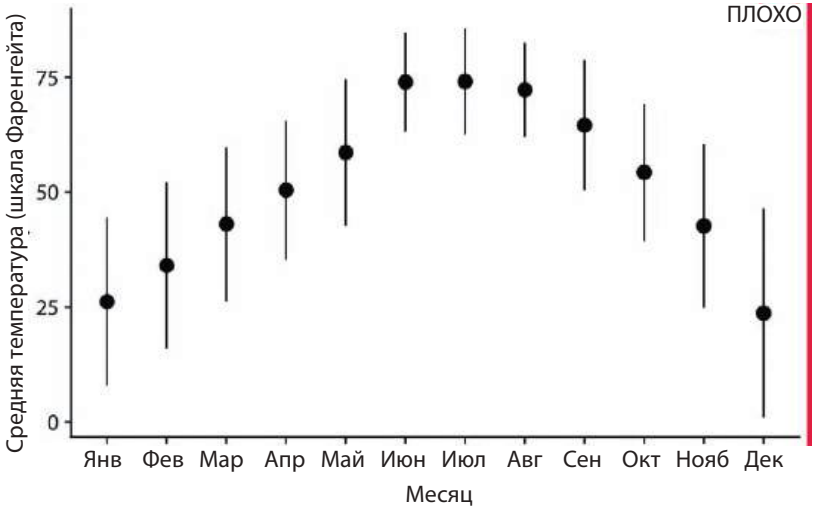


Рис. 8.1. Среднесуточные температуры в Линкольне (штат Небраска) в 2016 году. Точки представляют собой среднесуточные температуры для каждого месяца, усредненные за все дни месяца, а планки погрешностей отображают двойное стандартное отклонение среднедневных температур в течение каждого месяца. Это изображение относится к категории «плохих», поскольку планки погрешностей обычно используются для визуализации неопределенности оценки, а не изменчивости данных генеральной совокупности. Источник: Weather Underground

Однако планки погрешностей чаще всего используются для визуализации стандартной ошибки (или удвоенной стандартной ошибки для получения 95%-ного доверительного интервала), поэтому при считывании графика высок риск перепутать стандартную ошибку со стандартным отклонением. Первая показывает, насколько точна наша оценка среднего значения, в то время

как второе оценивает разброс данных относительно среднего значения. Набор данных может одновременно иметь как очень небольшую стандартную ошибку, так и весьма большое стандартное отклонение. В-четвертых, если в данных есть асимметрия, симметричные планки погрешностей введут в заблуждение, как это происходит и в данном случае, и почти всегда для наборов данных о реальном мире.

Все четыре недостатка рис. 8.1 можно устранить с помощью традиционного и широко используемого метода визуализации распределений — *коробчатой диаграммы (иногда называемой «ящик с усами»)*. Такая форма визуализации разделяет данные на квартили и визуализирует их в едином формате (рис. 8.2).

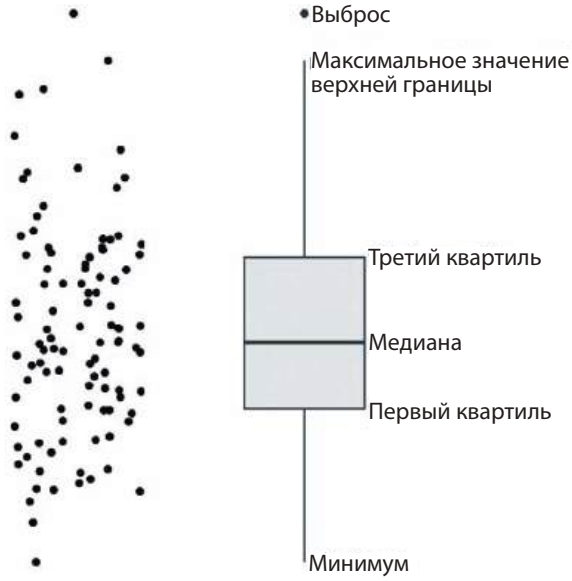


Рис. 8.2. «Анатомия» коробчатой диаграммы. Слева показано облако точек, а справа — представление той же информации в виде «ящика с усами»

В коробчатой диаграмме на рис. 8.2 показаны только значения точек по оси *y*. Линия посередине представляет собой медиану, а «коробка» очерчивает границы средних 50% точек. Вертикальные линии, простирающиеся вверх и вниз от коробки, называются усами.

Верхние и нижние усы растягиваются от границ коробки либо до максимального и минимального значений, либо до максимальных значений, не превышающих 1,5 высоты соответствующей стороны коробки. При этом значения в 1,5 высоты соответствующей стороны коробки называются пределами. Точки данных, лежащие вне пределов, обычно наносятся на график как отдельные точки в своих значениях.

Коробчатые представления — это простые и в то же время информативные графики. Они хорошо выглядят рядом друг с другом, если нужно визуализировать сразу несколько распределений. Взгляните на рис. 8.3, на котором приведены данные о температуре в городе Линкольн с использованием коробчатых представлений. Из рисунка видно, что в декабре температура сильно выделяется из общей картины (большинство дней умеренно холодные, а некоторые очень холодные), тогда как в другие месяцы, например в июле, перекоса в данных нет.

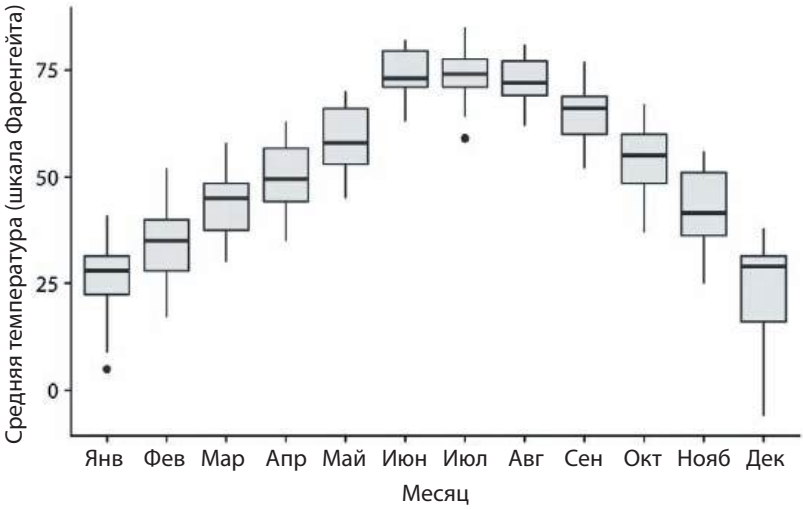


Рис. 8.3. Среднесуточная температура в Линкольне (штат Небраска) в 2016 году, визуализированная в виде коробчатой диаграммы. Источник: Weather Underground

Коробчатые диаграммы были изобретены статистиком Джоном Тьюки в начале 1970-х годов и очень скоро стали популярными благодаря своей информативности и тому, что их было несложно нарисовать от руки (в те годы большинство графиков создавалось именно так). Однако сейчас, когда к услугам каждого специалиста по работе с данными имеется масса современных вычислительных и графических возможностей, мы не ограничиваемся только теми визуализациями, которые легко нарисовать вручную. Именно поэтому в настоящее время на смену коробчатым диаграммам приходят *скрипичные графики* (рис. 8.4). Их можно использовать во всех тех случаях, что и коробчатые представления, а картина данных, которую они дают, более щедра на нюансы. В частности, скрипичные графики точно отображают бимодальные данные, в то время как «ящики с усами» — нет.

Скрипичные графики точно так же отображают только y-координаты точек, а вот ширина «скрипки» при заданном значении y равна плотности распределения в точке y. Технически скрипичный график представляет собой

график функции плотности, отраженный относительно горизонтальной оси и повернутый на 90° (см. главу 6), благодаря чему «скрипки» всегда выглядят симметрично. «Скрипки» начинаются и заканчиваются на минимальном и максимальном значениях данных соответственно. Самая широкая часть «скрипки» соответствует самой высокой плотности точек в наборе данных.

Рис. 8.4. «Анатомия» скрипичного графика. Слева показано облако точек, а справа — соответствующий ему скрипичный график

Прежде чем использовать «скрипки» для визуализации распределений, убедитесь, что каждая группа содержит достаточное количество точек данных, чтобы изобразить графики плотности в виде сплошных линий.

Если визуализировать температурные данные Линкольна при помощи скрипичных графиков, результат будет выглядеть как рис. 8.5. На картинке видно, что некоторые месяцы действительно характеризуются умеренно бимодальными данными. Например, в ноябре, похоже, было два температурных кластера: один около 50 градусов и один около 35 градусов по Фаренгейту.

Поскольку скрипичные графики строятся на основе оценок плотности, они обладают всеми их недостатками. В частности, графики такого типа могут создать иллюзию наличия данных там, где их нет, или что набор данных очень плотный, хотя на самом деле он весьма скудный. Мы можем попробовать перехитрить природу графика и обойти эти проблемы, просто изобразив все отдельные точки данных именно как точки (рис. 8.6). Такая фигура

называется *полосовым графиком*. Полосовые графики — неплохой метод визуализации, но только если не увлекаться и не накладывать друг на друга слишком много точек. Избежать этой проблемы можно, если слегка разнести точки вдоль оси *x*, добавив по этой оси случайный шум (рис. 8.7). Такая техника называется *разбросом (или джиттерингом)*.

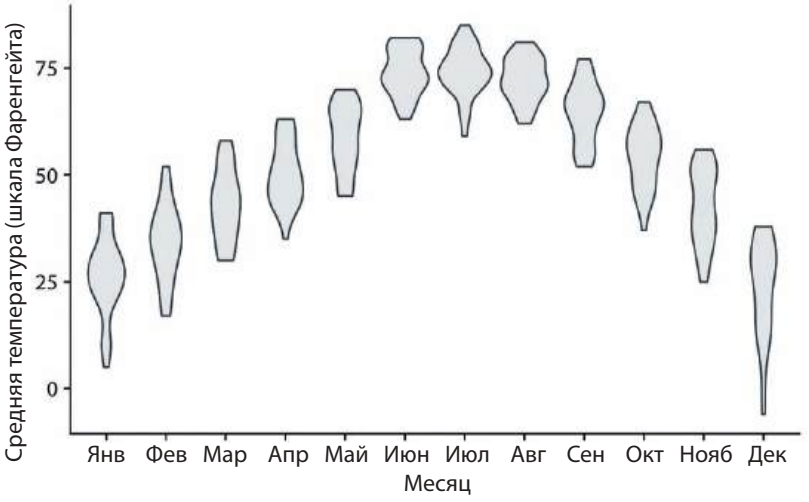


Рис. 8.5. Среднесуточные температуры в Линкольне (штат Небраска), визуализированные в виде скрипичного графика. Источник: Weather Underground

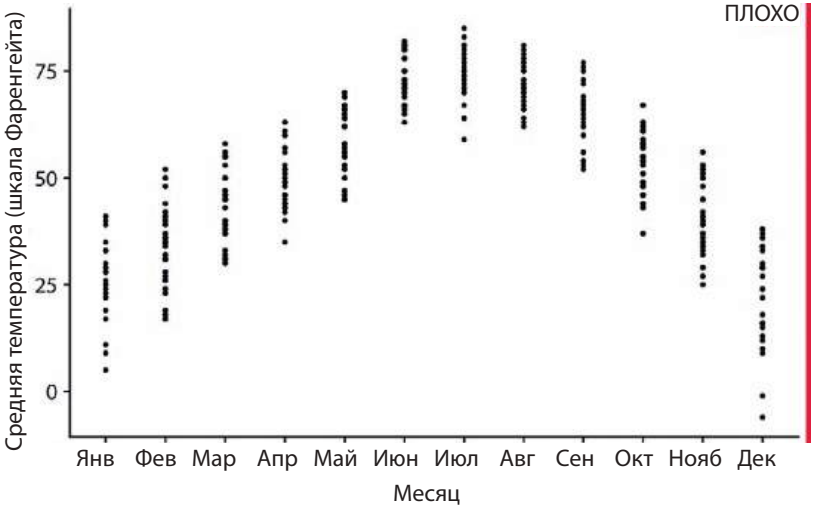


Рис. 8.6. Среднесуточные температуры в Линкольне (штат Небраска), визуализированные в виде полосовой диаграммы. Каждая точка показывает среднюю температуру за конкретный день. Данное изображение я отнес к категории «плохих» из-за слишком большого количества точек, расположенных друг над другом. Сложно понять, какая температура была наиболее частой в каждом из месяцев. Источник: Weather Underground

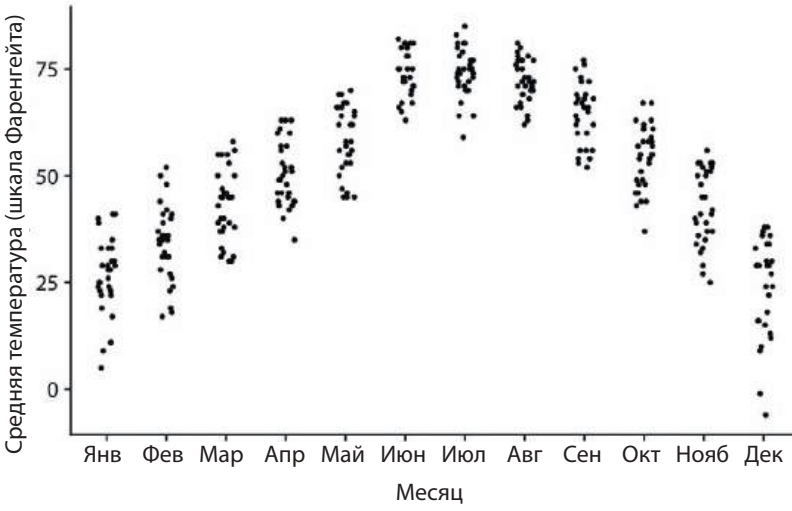


Рис. 8.7. Средние дневные температуры в Линкольне (штат Небраска), визуализированные в виде полосовой диаграммы. Точки разбросаны вдоль оси *x*, благодаря чему плотность точек на каждом температурном значении становится более очерченной. Источник: Weather Underground

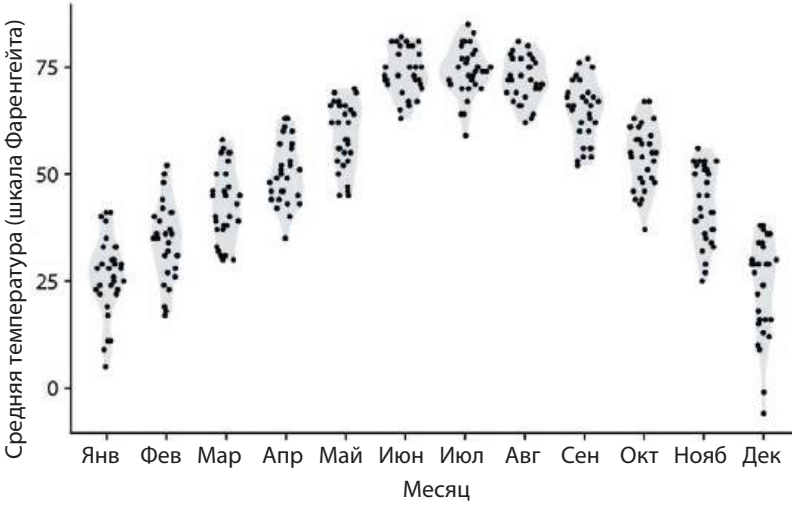


Рис. 8.8. Среднесуточные температуры в Линкольне (штат Небраска), визуализированные в виде графика Sinz (комбинация отдельных точек и скрипичного графика). Точки распределены вдоль оси *x* согласно их плотности при соответствующих значениях температур. Графики Sinz на этом изображении наложены на скрипичные графики. Источник: Weather Underground

Если у вас не хватает данных для построения скрипичного графика, можно просто нанести на график все имеющиеся точки.

Наконец, мы можем взять все самое лучшее из обоих миров, если распределим точки пропорционально их плотности в заданной *y*-координате. Этот метод, получивший название графика Sinz [Sidiropoulos et al., 2018]*, можно рассматривать как некий гибрид скрипичного графика и разбросанных точек. Данный график показывает каждую отдельную точку, одновременно визуализируя распределения. На рис. 8.8 я нарисовал графики Sinz поверх скрипичных графиков, чтобы подчеркнуть взаимосвязь между этими двумя подходами.

Визуализация распределений на горизонтальной оси

В главе 6 мы визуализировали распределения по горизонтальной оси при помощи гистограмм и графиков плотности. В данном разделе мы расширим эту идею, распределив несколько диаграмм распределения по вертикали. Полученная в результате визуализация называется *«горный хребет»***, потому что она очень напоминает скалистую горную гряду. «Горные хребты» показывают себя лучше всего в тех случаях, когда нужно продемонстрировать временны́е тенденции распределений.

Стандартный график «горный хребет» использует для визуализации оценки плотности (рис. 8.9). Он тесно связан со скрипичным графиком, но зачастую выглядит более интуитивно понятным. Например, два кластера ноябрьских температур — около 35 градусов и 50 градусов по Фаренгейту — гораздо более очевидны на рис. 8.9, нежели на рис. 8.5.

Так как по оси *x* отложена объясняемая переменная, а ось *y* визуализирует группирующую переменную, отдельной оси для оценок плотности на «горном хребте» не существует. Оценки плотности показаны вместе с группирующей переменной, что ничем не отличается от скрипичного графика, на котором оценки плотности тоже находятся вместе, без указания отдельной явной шкалы. В обоих случаях цель графика состоит не в том, чтобы показать значения плотности, а в том, чтобы читатель мог легко сравнить между собой формы и высоты графиков плотности распределений различных групп.

В общем-то для визуализации «горного хребта» мы могли бы вместо графиков плотности использовать гистограммы, но, к сожалению,

---

* График Sina Plot назван в честь Sina Hadi Sohi, студента университета Копенгагена (Дания), который написал первую версию кода для создания этих графиков (Frederick O. Bagger, personal communication).

** Название графика Ridgeline Plot не имеет в доступной литературе стандартизированного перевода на русский язык. В данной книге мы будем называть этот график «горный хребет». — *Прим. ред.*

результат такой замены зачастую оставляет желать лучшего (рис. 8.10). Проблемы здесь аналогичны проблемам гистограмм с наложением или перекрытием (см. «Визуализация нескольких распределений одновременно» на с. 75).

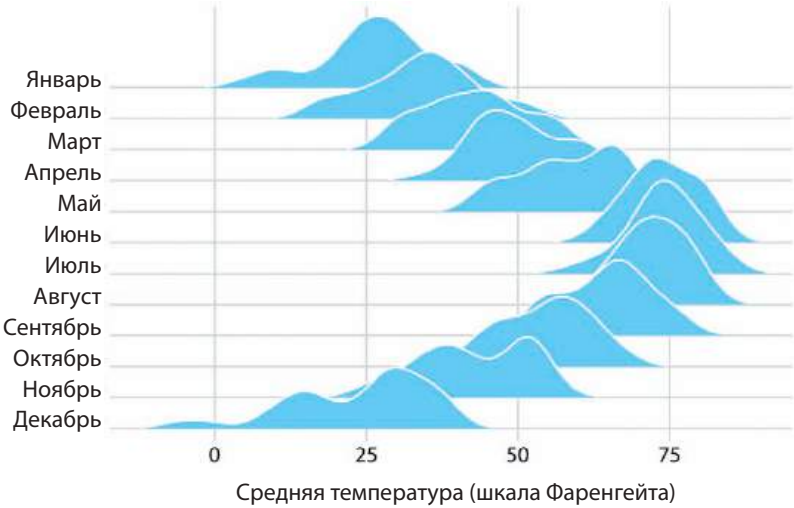


Рис. 8.9. Среднесуточные температуры в Линкольне (штат Небраска), визуализированные в виде графика «горный хребет». Для каждого месяца указано распределение средних дневных температур в градусах Фаренгейта. Изначальная концепция изображения: [Wehrwein, 2017]. Источник: Weather Underground

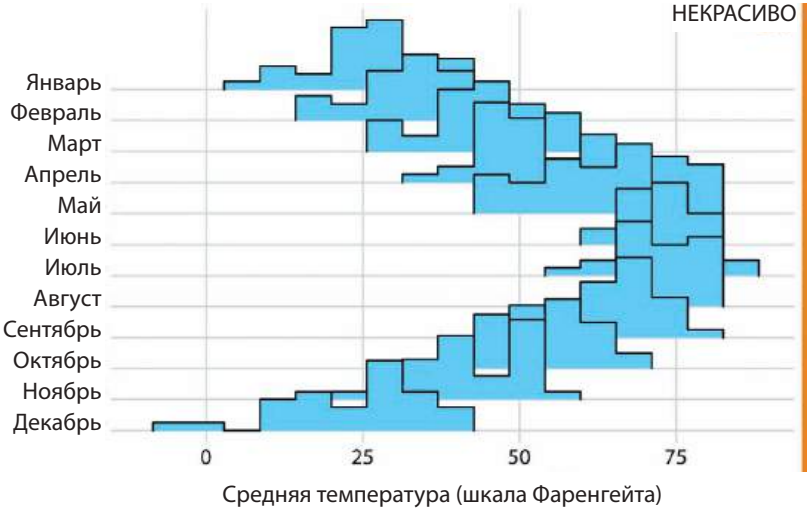


Рис. 8.10. Среднесуточные температуры в Линкольне (штат Небраска), визуализированные в виде графика «горный хребет», составленного из гистограмм. Отдельные гистограммы сложно отделить друг от друга, поэтому общая картина сбивает с толку и перегружает информацией. Источник: Weather Underground

Поскольку вертикальные линии в этих гистограммах всегда отображаются в одних и тех же значениях *x*, столбцы из разных гистограмм визуально смешиваются друг с другом. На мой взгляд, таких накладывающихся гистограмм лучше избегать.

Графики типа «горный хребет» можно легко масштабировать до очень больших количеств распределений. К примеру, на рис. 8.11 показаны распределения продолжительности фильмов с 1913 по 2005 год. На графике содержится почти 100 самостоятельных распределений, однако читается он очень легко.

Мы видим, что в 1920-е годы длительность фильмов могла быть самой разной, но примерно в 1960 году она пришла к некоему стандарту на уровне около 90 минут.

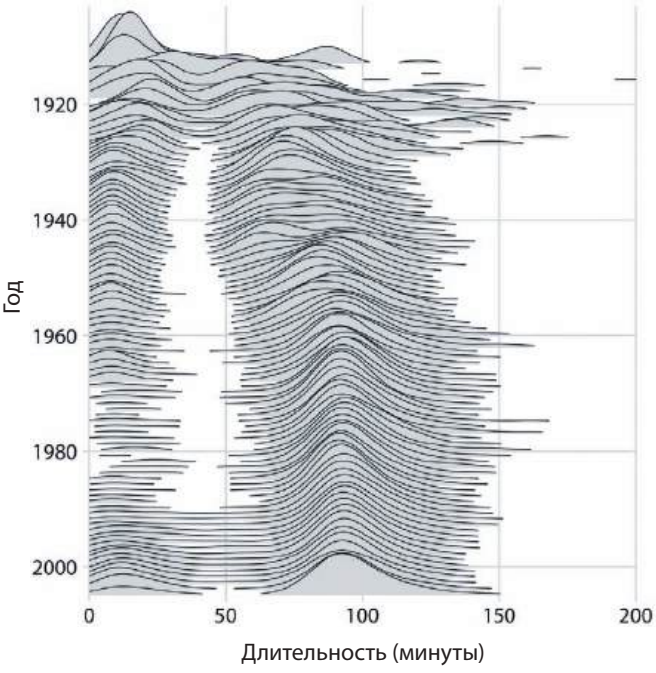


Рис. 8.11. Динамика продолжительности фильмов с течением времени. Начиная с 1960-х годов большинство фильмов стало идти около 90 минут. Источник: Internet Movie Database (IMDB)

«Горные хребты» хорошо работают и для случаев сравнения двух временны́х тенденций. Такой сценарий обычно возникает, когда нам нужно проанализировать распределение голосов между членами двух разных партий. Провести такое сравнение можно, если расположить распределения вертикально по времени и нарисовать разным цветом два распределения, представляющих обе партии в каждый момент времени (рис. 8.12).

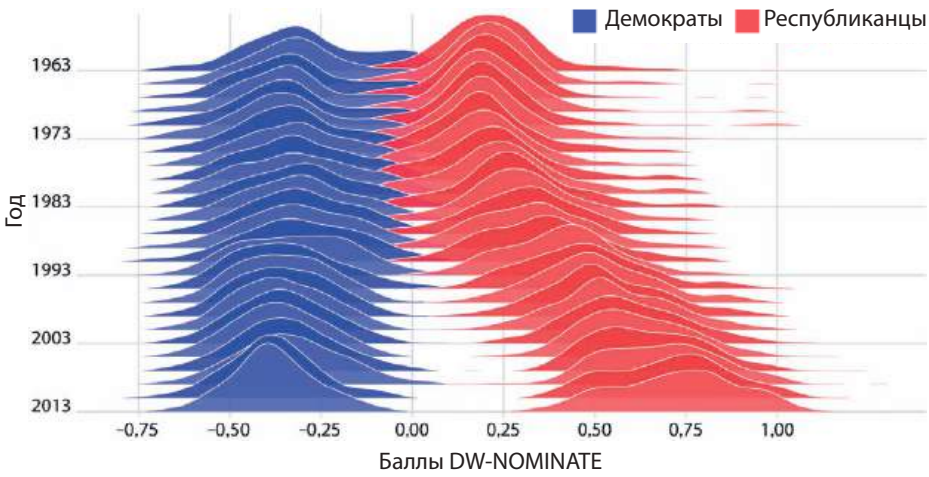


Рис. 8.12. Модель голосования в Палате представителей Конгресса США становится все более поляризованной. Баллы DW-NOMINATE часто используются для сравнения распределения голосов между партиями с течением времени. На этом графике показаны распределения баллов для каждого Конгресса с 1963 по 2013 год отдельно для демократов и республиканцев. Данные по каждому Конгрессу приведены за первый год его работы. Оригинальная концепция графика: [McDonald, 2017]. Источник: Keith Poole

Глава 9

Визуализация пропорций

Одним из распространенных сценариев визуализации является ситуация, когда какая-либо группа, сущность или сумма разбиваются на отдельные элементы, каждый из которых является частью целого. Наиболее очевидными примерами являются: распределение мужчин и женщин в той или иной группе, проценты людей, голосующих за различные политические партии на выборах, доли рынка компаний. Типичным визуальным воплощением таких данных является круговая диаграмма, которую можно увидеть практически в каждой бизнес-презентации и которую так сильно недолюбливают ученые. Дальше мы увидим, что визуализация пропорций может быть непростой задачей, особенно когда некоторое целое разбито на множество различных частей или когда нам нужно увидеть изменения в пропорциях с течением времени или при различных условиях. Не существует какой-то одной идеальной визуализации, которая подходит для любого случая. В доказательство этого утверждения я покажу вам несколько различных сценариев, для каждого из которых требуется свой подход к отображению.

Помните: всегда нужно выбирать ту визуализацию, которая лучше всего подходит для вашего конкретного набора данных и подчеркивает именно те ключевые особенности массива данных, которые вы хотите показать.

Время круговых диаграмм!

С 1961 по 1983 год парламент Германии (Бундестаг) состоял из членов трех различных партий: ХДС/ХСС, СДПГ и СвДП. Большую часть этого времени ХДС/ХСС и СДПГ имели примерно сопоставимое число мест, в то время как СвДП принадлежала лишь небольшая доля. Например, в восьмом Бундестаге в 1976–1980 годах ХДС/ХСС занимала 243 места, СДПГ — 214, а СвДП — 39, то есть в общей сложности 496. Подобные парламентские данные чаще всего представляются в виде круговой диаграммы (рис. 9.1).

Круговая диаграмма разбивает круг на части таким образом, чтобы площадь каждого сегмента была пропорциональна доле общей суммы, которую

он представляет. Аналогичное разбиение можно выполнить и на прямоугольнике, в результате чего получится столбчатая диаграмма с накоплением (рис. 9.2). В зависимости от способа нарезки график будет иметь вид или столбчатой (рис. 9.2A), или линейчатой (рис. 9.2B) диаграммы.

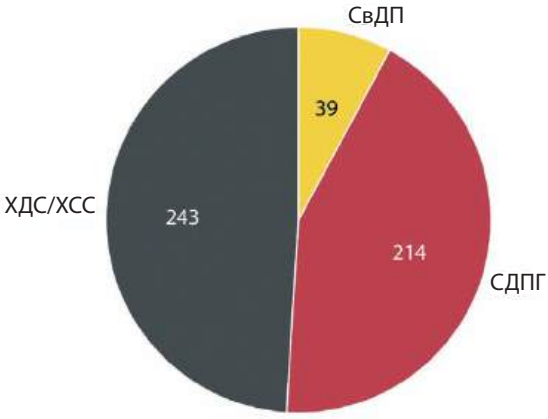


Рис. 9.1. Партийный состав восьмого немецкого Бундестага 1976–1980 годов, визуализированный в виде круговой диаграммы. Эта визуализация показывает, что правящая коалиция СДПГ и СвДП имела незначительное большинство голосов против оппозиционного ХДС/ХСС. Источник: Wikipedia

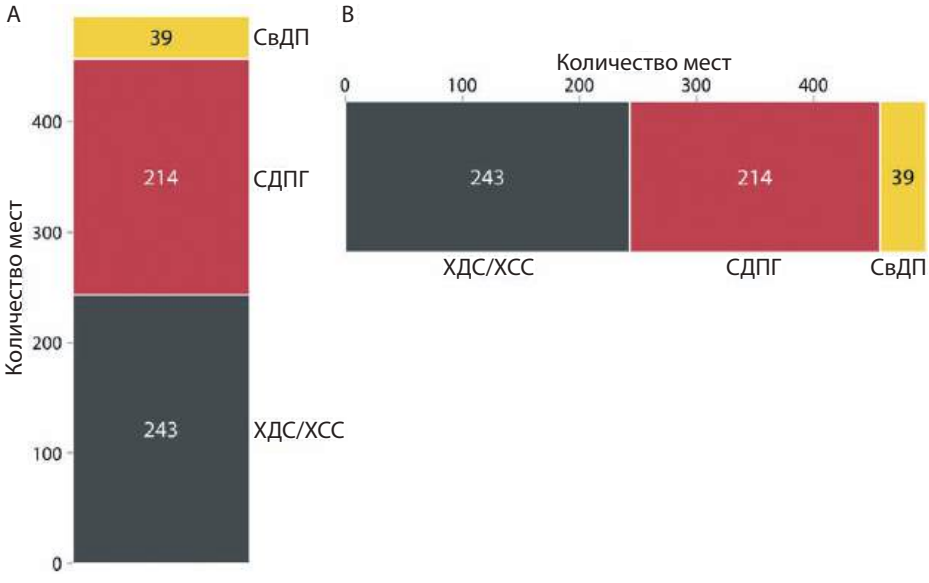


Рис. 9.2. Партийный состав восьмого немецкого Бундестага 1976–1980 годов, визуализированный в виде диаграмм с накоплением. A. Столбчатая диаграмма с накоплением. B. Линейчатая диаграмма с накоплением. Тот факт, что у СДПГ и СвДП в сумме больше мест, чем у ХДС/ХСС, заметен не сразу. Источник: Wikipedia

В качестве альтернативы мы можем расположить столбцы из рис. 9.2А рядом, а не накладывать их друг на друга. Такая визуализация упрощает прямое сравнение трех групп, но при этом скрывает другие аспекты данных (рис. 9.3). Наиболее существенный недостаток стандартной столбчатой диаграммы — это визуальная неочевидность связи каждого столбца с общей суммой.

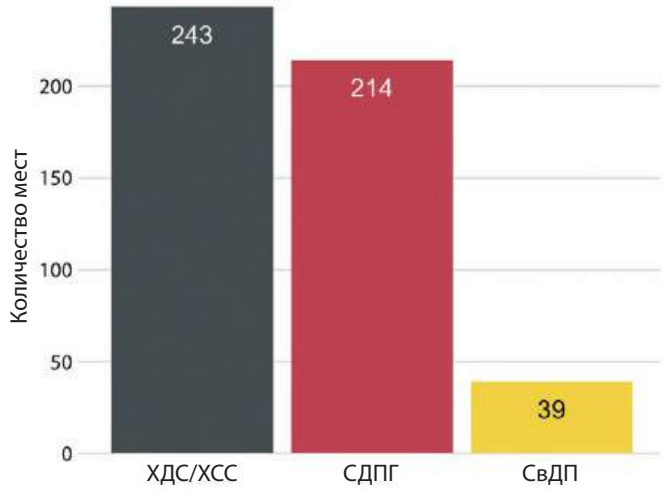


Рис. 9.3. Партийный состав восьмого немецкого Бундестага 1976–1980 годов, визуализированный в виде стандартной столбчатой диаграммы. Как и на рис. 9.2, здесь тоже не сразу видно, что СДПГ и СвДП совместно имели больше мест, чем ХДС/ХСС. Источник данных: Wikipedia

Многие авторы очень негативно относятся к использованию круговых диаграмм и считают столбчатые диаграммы (как стандартные, так и с наложением) более предпочтительными. Другие говорят, что круговые диаграммы допустимо использовать в определенных случаях. Лично я считаю, что ни одна из этих визуализаций не превосходит другую. Способ визуализации следует выбирать в зависимости от особенностей набора данных и цели, которую мы преследуем. На мой взгляд, для отображения данных о восьмом немецком Бундестаге круговая диаграмма является как раз лучшим вариантом. Этот тип графика подчеркивает, что у правящей коалиции СДПГ и СвДП в сумме было небольшое преимущество по сравнению с ХДС/ХСС (см. рис. 9.1). Ни один из других графиков не показывает это столь же наглядно (см. рис. 9.2 и 9.3).

В целом круговые диаграммы хорошо работают в тех случаях, когда нужно сделать акцент на элементарных частях целого, таких как половина, треть или четверть. Также круговые диаграммы хорошо подходят для малых наборов данных. Простая круговая диаграмма (см. рис. 9.1) смотрится отлично, но при этом столбчатая диаграмма с накоплением, построенная на тех же данных (см. рис. 9.2А), выглядит гораздо хуже. С другой стороны, диаграммы

с накоплением могут использоваться для одновременного сравнения частей нескольких целых или при изменении частей одного целого во времени. Они же являются предпочтительными, когда нужно просто сравнить между собой отдельные части целого. Краткий перечень плюсов и минусов круговых, обычных столбчатых и столбчатых с накоплением диаграмм приведен в табл. 9.1.

Таблица 9.1. Преимущества и недостатки общих подходов к визуализации пропорций: круговые диаграммы, обычные столбчатые диаграммы и столбчатые диаграммы с накоплением

	Круговая диаграмма	Столбчатая диаграмма с накоплением	Обычная столбчатая диаграмма
Четко визуализирует данные в виде частей единого целого	✓	✓	×
Обеспечивает легкое визуальное сравнение относительных пропорций	×	×	✓
Визуально подчеркивает простые разделения, такие как 1/2, 1/3, 1/4	✓	×	×
Визуальная привлекательность сохраняется даже в случае очень маленьких наборов данных	✓	×	✓
Хорошо работает для случаев разбиения целого на множество частей	×	×	✓
Хорошо работает в случае визуализации большого количества пропорций или изменения пропорций во времени	×	✓	×

Пример в пользу столбчатых диаграмм

Давайте посмотрим, в каких случаях круговым диаграммам следует сказать твердое «нет». Следующий пример создан на основе критики круговых диаграмм, которая изначально была опубликована в «Википедии» [Wikipedia, 2007]. Рассмотрим гипотетический сценарий для пяти компаний — A, B, C, D и E, — каждая из которых имеет сопоставимую долю рынка в размере около 20%. Наш гипотетический набор данных содержит сведения о доле рынка каждой компании за три года подряд. Визуализируя эту информацию с помощью круговых диаграмм, мы жертвуем ясностью отображения конкретных тенденций (рис. 9.4). Из рисунка видно, что доля компании A растет, а доля компании E уменьшается, однако это все, что мы можем сказать о происходящем. Например, совершенно непонятно, как именно сравниваются рыночные доли разных компаний в течение каждого года.

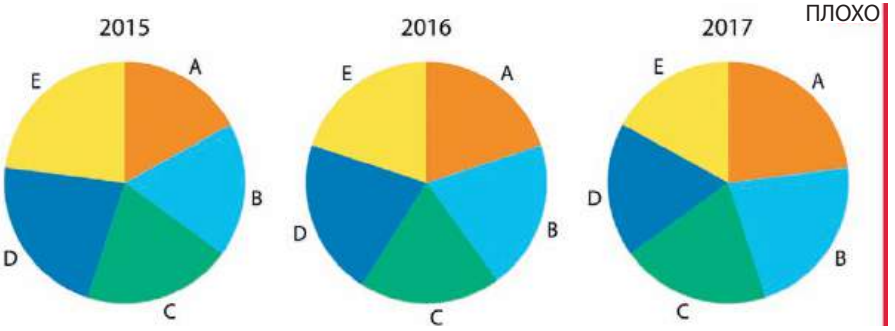


Рис. 9.4. Доля рынка пяти гипотетических компаний A…E за 2015–2017 годы представлена в виде круговых графиков. У этой визуализации есть две основные проблемы: 1) практически невозможно сравнивать относительные доли рынка на отрезке в несколько лет; 2) трудно понять, как менялись доли рынка компаний с течением времени

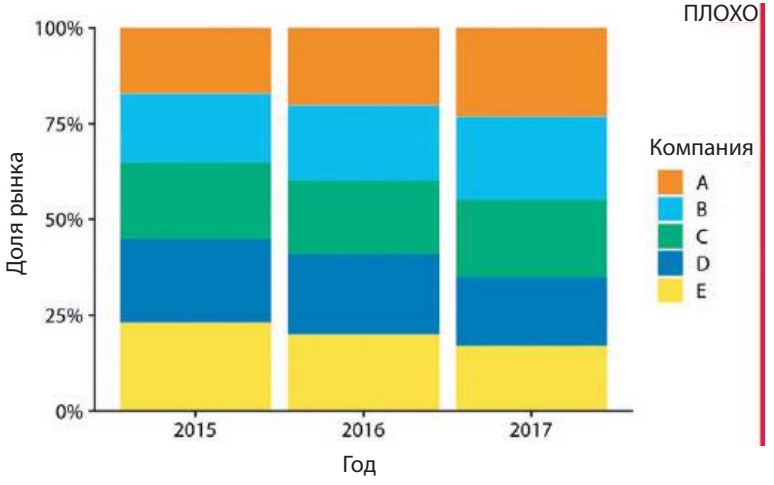


Рис. 9.5. Доля рынка пяти гипотетических компаний за период 2015–2017 годов, представленная в виде столбчатой диаграммы с накоплением. У этой визуализации есть две основные проблемы: 1) трудно провести сравнение относительной доли рынка на отрезке нескольких лет; 2) трудно заметить изменения доли рынка за период нескольких лет для компаний посередине графика (B, C и D), поскольку положение их столбцов меняется год от года

Если этот же набор данных мы отобразим в виде столбчатой диаграммы с накоплением, картина станет немного яснее (рис. 9.5). Теперь хорошо видны тенденции увеличения доли компании A и сокращения доли компании E на рынке. Тем не менее все еще очень трудно сравнивать относительную долю рынка этих компаний в течение каждого года. Кроме того, по-прежнему сложно сравнить между собой доли компаний B, C и D, потому что столбцы в каждый следующий год отображаются сдвинутыми относительно друг

друга. Это общая проблема столбчатых диаграмм с накоплением и основная причина, по которой я обычно не рекомендую этот способ визуализации.

Вообще, наилучшим вариантом визуализации этого гипотетического набора данных является обычная столбчатая диаграмма (рис. 9.6). Из графика видно, что компании A и B увеличили свою долю рынка в период с 2015 по 2017 год, в то время как компании D и E свою долю рынка сократили. Кроме того, на изображении хорошо видно, что рыночные доли последовательно увеличиваются, начиная с компании A и до компании E, в 2015 году, а в 2017 году таким же образом снижаются.

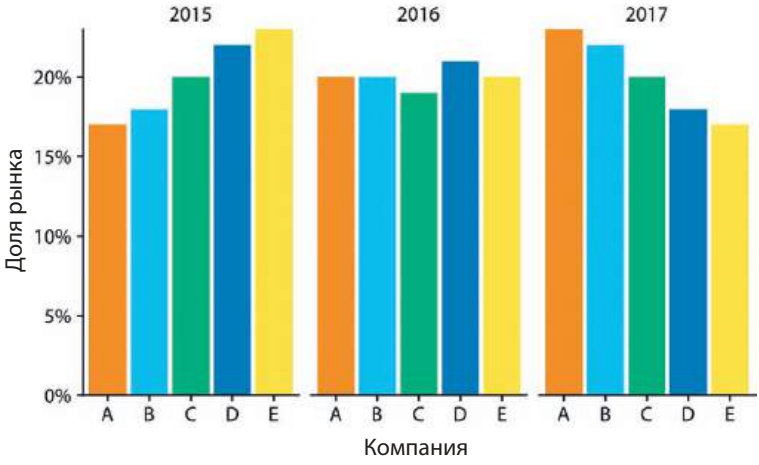


Рис. 9.6. Доля рынка пяти гипотетических компаний за период 2015–2017 годов, визуализированная в виде обычной столбчатой диаграммы

Пример в пользу столбчатых диаграмм и графиков плотности с наложением

В предыдущем разделе я говорил о том, что обычно не рекомендую использовать визуализацию с помощью столбчатых диаграмм с накоплением, потому что положение внутренних столбцов смещается от точки к точке. Однако если на графике изображены ровно два таких столбца, то упомянутая проблема исчезает, а итоговая визуализация становится достаточно ясной. В качестве примера рассмотрим долю женщин в национальном парламенте страны. Мы здесь будем говорить об африканской стране Руанда, которая по состоянию на 2016 год занимает первое место в списке стран с наибольшей долей женщин в парламенте. В 2008 году в парламенте Руанды начинают преобладать женщины, а в 2013 году их количество составляет почти две трети от всех депутатов. Чтобы показать этот процесс во времени, мы можем нарисовать столбчатую диаграмму с накоплением, где каждый столбец будет соответ-

ствовать точке во времени (рис. 9.7). Этот график дает наглядное представление об изменении пропорций год от года. Чтобы читателю было легче понять, с какого времени в парламенте Руанды стали преобладать женщины, я добавил в график пунктирную горизонтальную линию уровня для значения 50%. Без нее было бы невозможно понять, кто составлял большинство в период с 2003 по 2007 год — мужчины или женщины. Чтобы не перегружать изображение, я не стал добавлять подобные линии уровня для 25 и 75%.

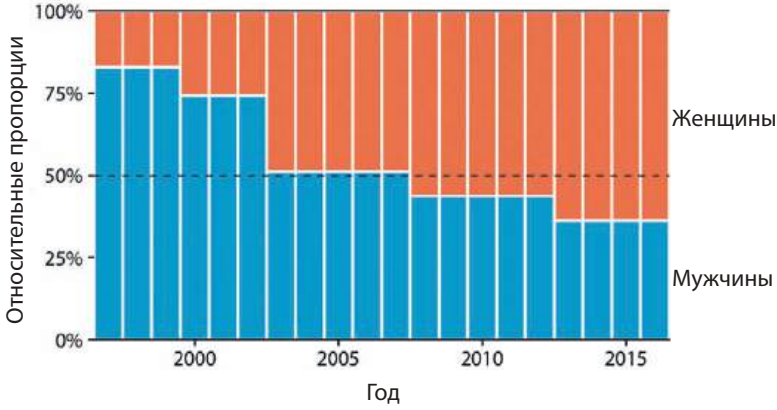


Рис. 9.7. Перемены в гендерной составляющей парламента Руанды с 1997 по 2016 год. Источник: Inter-Parliamentary Union (IPU)*

Если мы хотим проиллюстрировать изменение пропорций на основании значений непрерывной переменной, мы можем перейти от столбчатых диаграмм с накоплением к графикам плотности с накоплением. Подобный вид графиков можно рассматривать как предельный случай бесконечно большого количества столбчатых диаграмм с бесконечно малой шириной столбца, расположенных рядом друг с другом. Значения плотности для графиков плотности с накоплением, как правило, получают при помощи ядерной оценки плотности, как это описано в главе 6, и, если вы хотите лучше представлять себе сильные и слабые стороны данного метода, советую обратиться к указанной главе.

В качестве примера, когда может быть уместно использование графиков плотности с наложением, давайте посмотрим, каким образом состояние здоровья людей зависит от их возраста. Если рассматривать его как непрерывную переменную, мы получим довольно качественную визуализацию (рис. 9.8). Несмотря на то что мы используем четыре категории здоровья, а я, как правило, не люблю нагромождать условия, набор данных в этом случае отображен вполне приемлемо. Легко заметить, что общий уровень здоровья снижается по мере старения, но также видно, что, несмотря на эту тенденцию, более

* ipu.org.

половины населения до глубокой старости остается в хорошем или даже отличном состоянии здоровья.

Однако у этой визуализации есть серьезное ограничение: отображая пропорции четырех состояний здоровья в процентах от общего числа, изображение скрывает, что в наборе данных гораздо больше молодых людей, чем пожилых. Таким образом, даже несмотря на то, что процент людей, сообщивших о наличии хорошего здоровья, остается почти неизменным в возрасте, охватывающем семь десятилетий, *абсолютное число* людей, имеющих хорошее здоровье, уменьшается по мере сокращения общего числа людей в данном возрасте. В следующем разделе я покажу, как можно решить эту проблему.

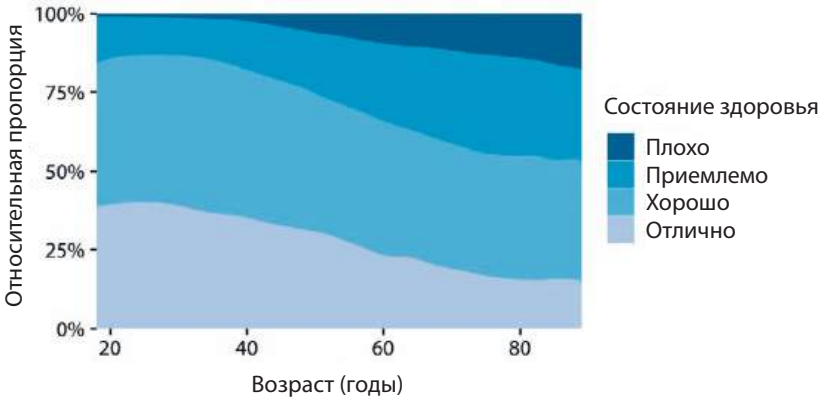


Рис. 9.8. Зависимость уровня здоровья людей от их возраста. Источник: General Social Survey (GSS)

Визуализация пропорций по отдельности как частей целого

Проблема обычных столбчатых диаграмм заключается в том, что они не дают четкого представления о размере отдельных частей относительно целого, а недостатком столбчатых диаграмм с накоплением является затрудненность сравнения наложенных друг на друга столбцов, потому что они начинаются в разных значениях. Для решения этих проблем мы можем для каждой части составить отдельный график, который будет показывать отношение этой части к целому. Применяя эту логику к рис. 9.8, мы получаем рис. 9.9. Общее возрастное распределение в наборе данных представлено серыми областями, а синим цветом показано распределение по возрасту для каждой категории здоровья. Этот показатель свидетельствует о том, что в абсолютном выражении число людей с отличным или хорошим здоровьем снижается после 30–40 лет, в то время как количество людей с посредственным здоровьем остается почти неизменным для всех возрастов.

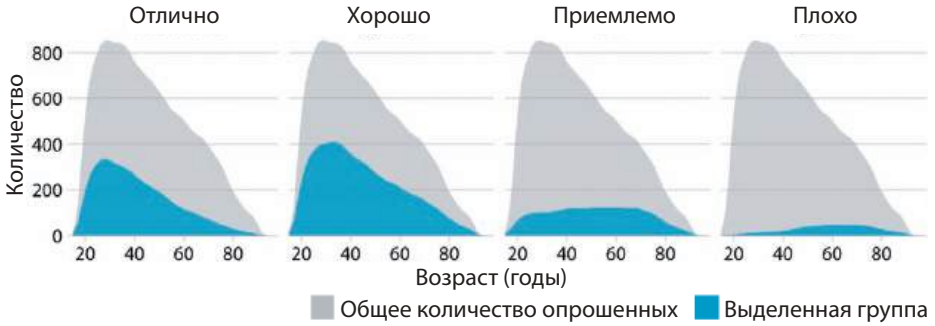


Рис. 9.9. Распределение состояния здоровья по возрастам, показанное в отношении к общему количеству опрошенных. Раскрашенные области показывают оценки плотности возраста людей, имеющих определенное состояние здоровья, а серые области показывают общее распределение по возрасту. Источник: GSS

В качестве следующего примера рассмотрим еще одну переменную из того же опроса: семейное положение. С возрастом семейный статус меняется гораздо сильнее, нежели состояние здоровья, поэтому график плотности с накоплением, отображающий эту зависимость, будет неинформативным (рис. 9.10).

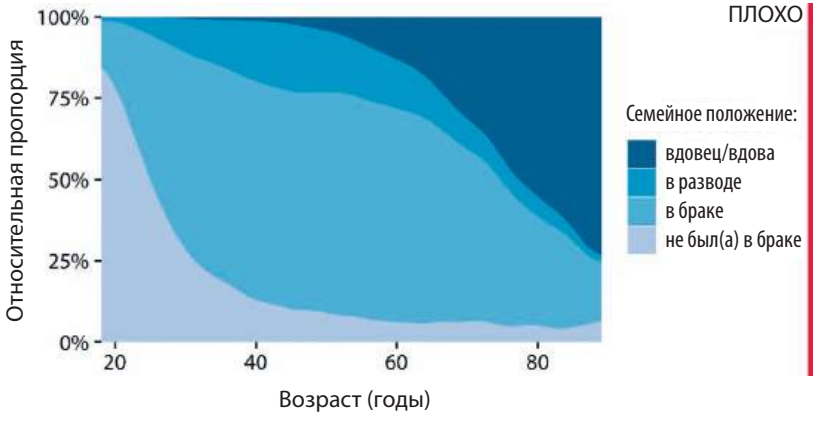


Рис. 9.10. Зависимость семейного положения от возраста. Чтобы читателю было проще воспринимать график, я удалил небольшое количество случаев, отмеченных как «живут раздельно». Данная визуализация относится к категории «плохих», потому что частота людей, никогда не состоявших в браке или ставших вдовцами/вдовами, с возрастом меняется очень резко, из-за чего сильно искажается возрастное распределение женатых и разведенных людей, поэтому эти данные становится сложно интерпретировать. Источник данных: GSS

Если тот же набор данных визуализировать в виде отдельных графиков плотности, картинка получается гораздо более понятной (рис. 9.11). В частности, мы видим, что пик числа людей, состоящих в браке, приходится на

возраст в районе 30 лет, пик числа разведенных людей приходится на 40 лет, а пик вдовства приходится на возраст около 75 лет.

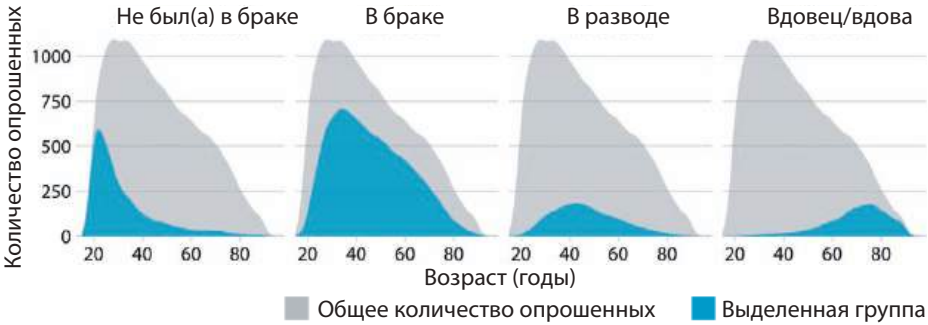


Рис. 9.11. Семейное положение в зависимости от возраста среди людей, участвовавших в опросе. Цветные области показывают оценки плотности распределения с соответствующим семейным положением, а серые области показывают общее распределение по возрасту. Источник: GSS

Одним из недостатков рис. 9.11 является то, что такое представление не позволяет легко определить относительные пропорции в любой момент времени. Например, если бы нам нужно было узнать, в каком возрасте более 50% всех опрошенных состоят в браке, рис. 9.11 мало чем помог бы в поисках ответа на этот вопрос. Для решения этой проблемы можно воспользоваться визуализацией того же типа, только по оси *y* вместо абсолютных чисел следует показывать относительные пропорции (рис. 9.12). Данное отображение демонстрирует, что начиная с возраста ближе к 30 годам большинство составляют женатые люди, а начиная с возраста в окрестности 75 лет преобладающей категорией становятся вдов(ц)ы.

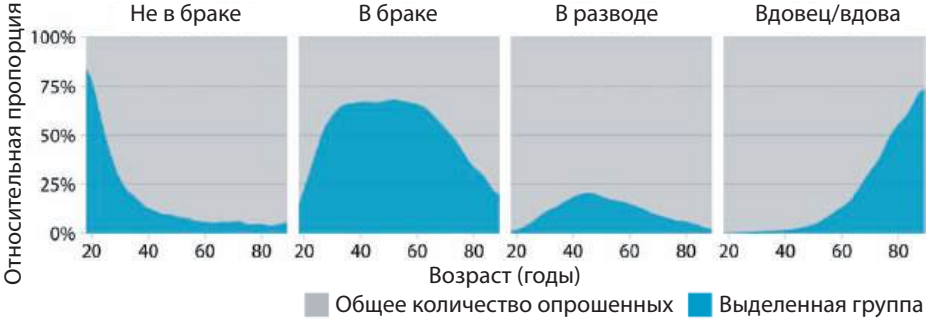


Рис. 9.12. Семейное положение в зависимости от возраста, выраженное в процентах от общего числа людей, участвовавших в опросе. Цветные области показывают процент людей данного возраста с соответствующим семейным положением, а области, выделенные серым цветом, показывают процент людей, имеющих другое семейное положение. Источник: GSS

Глава 10

Визуализация пропорций на нескольких уровнях

В предыдущей главе мы обсуждали сценарии, в которых набор данных разбивается на части в зависимости от какой-либо категориальной переменной: например, такой как политическая партия, компания или состояние здоровья. Однако нередко бывает так, что для решения задачи приходится углубляться в детали, из-за чего набор данных нужно разбить по нескольким категориальным переменным одновременно. Например, в случае с парламентом нас может заинтересовать распределение мест и пола депутатов в зависимости от партии. Аналогично, вспоминая пример с набором данных о состоянии здоровья людей, мы могли бы задаться вопросом, как выглядит зависимость уровня здоровья от семейного положения. Я называю такие сценарии многоуровневыми пропорциями, потому что каждая дополнительная категориальная переменная приводит к появлению нового уровня вложенности и к еще более гранулярному разбиению данных. Для визуализации таких многоуровневых пропорций существует несколько интересных подходов, к которым относятся мозаичные графики, деревья и графики в параллельных координатах.

Как не надо строить многоуровневые пропорции

Для начала давайте рассмотрим два ошибочных подхода к визуализации многоуровневых пропорций. Несмотря на то, что любой опытный исследователь с первого взгляда заметит бессмысленность таких подходов, мне приходилось на практике сталкиваться с этими примерами, поэтому я считаю, что их обязательно следует обсудить. В этой главе я буду работать с набором данных, который содержит информацию о 106 мостах города Питтсбург, как то: материал, из которого они были построены (сталь, железо или дерево), и год, когда они были воздвигнуты. В зависимости от года возведения мосты подразделяются на следующие категории: классические, возведенные до 1870 года, и современные мосты, возведенные после 1940 года.

Предположим, что нам нужно создать график, который показывает долю мостов в зависимости от материала изготовления (сталь, железо или дерево), а также долю классических/современных мостов. Если мы поддадимся соблазну отобразить данные в виде круговой диаграммы (рис. 10.1), толку от такой визуализации будет мало. Сумма всех фрагментов должна равняться 100%, а здесь получилось 135%. Причиной такого результата является двойной учет части мостов. Каждый мост изготовлен из стали, железа или дерева, поэтому эти три части круговой диаграммы уже представляют в сумме 100% мостов. При этом каждый классический или современный мост также является мостом, изготовленным из стали, железа или дерева, и, следовательно, повторно учитывается в круговой диаграмме.

Рис. 10.1. Круговая диаграмма, визуализирующая разбиение мостов города Питтсбург в зависимости от материала конструкции (сталь, дерево, железо) и от даты постройки (созданные по классическим технологиям, до 1870 года; современные, созданные после 1940-го). Значения показывают процентное соотношение мостов данного типа среди всех мостов. Поскольку общее процентное соотношение составляет более 100%, данное изображение не имеет никакого смысла. Информация о строительных материалах и дате постройки перекрывается: к примеру, все современные мосты изготавливаются из стали, а большинство старых мостов сделаны из дерева. Источник: Yoram Reich and Steven J. Fenves, via the UCI Machine Learning Repository [Dua and Karra Taniskidou, 2017]

Двойной учет одних и тех же данных не всегда является ошибкой, если, например, выбранная нами визуализация не требует, чтобы сумма всех пропорций составляла 100%. Как уже говорилось в предыдущей главе, простые столбчатые диаграммы полностью соответствуют этому критерию. Например, мы можем показать различные пропорции мостов в виде полос на одном графике, при этом изображение будет технически корректным (рис. 10.2). Тем не менее я отнес данный график к категории «плохих», потому что он

не сразу показывает, что некоторые из представленных категорий частично перекрываются. Глядя на рис. 10.2, случайный наблюдатель может сделать вывод, что существует пять отдельных категорий мостов и что, например, современные мосты не изготавливаются из стали, дерева или железа.

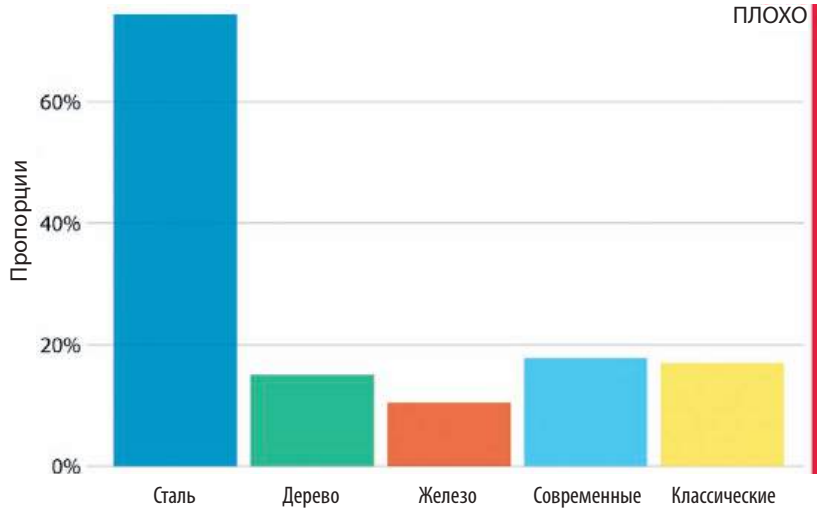


Рис. 10.2. Столбчатая диаграмма визуализирует разбиение мостов города Питтсбург в зависимости от материала конструкции (сталь, дерево, железо) и даты постройки (классические, построенные до 1870 года, и современные, созданные после 1940 года). В отличие от рис. 10.1, такая визуализация технически корректна, так как она не подразумевает, что общая высота столбцов должна составлять 100%. Тем не менее из этого графика нельзя понять, какие мосты и из каких групп были посчитаны дважды, поэтому данное изображение относится к категории «плохих». Источник: Yoram Reich and Steven J. Fenves

Мозаичные графики и древовидные карты

Всякий раз, когда мы имеем дело с пересекающимися категориями, следует четко показывать, как они соотносятся друг с другом. Это можно сделать с помощью *мозаичного графика* (рис. 10.3). На первый взгляд, графики этого типа напоминают столбчатые диаграммы с накоплением (например, рис. 9.5). Однако на мозаичном графике, в отличие от сложенных гистограмм, высота и ширина отдельных участков различаются. Обратите внимание, что на рис. 10.3 присутствуют две дополнительные строительные эпохи: «ранние» (с 1870 по 1889 год) и «зрелые» (с 1890 по 1939 год). В сочетании с категориями «классические» и «современные» эти строительные эпохи учитывают все мосты, которые имеются в наборе данных, равно как и все строительные материалы. Это условие чрезвычайно важно для построения мозаичного графика: каждая показанная категориальная переменная должна охватывать все данные из набора.

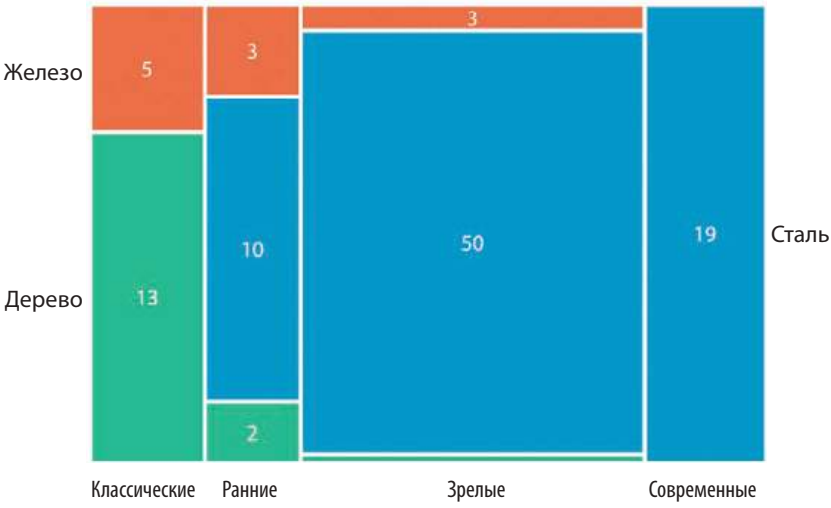


Рис. 10.3. Мозаичный график, отражающий разбиение мостов города Питтсбург в зависимости от материалов конструкции (сталь, дерево, железо) и даты постройки (классические, ранние, зрелые, современные). Ширина каждого прямоугольника пропорциональна количеству мостов, построенных в соответствующую эпоху, а высота пропорциональна количеству мостов, построенных из соответствующего материала. Значения отражают количество мостов в каждой из категорий. Источник: Yoram Reich and Steven J. Fenves

Создание нашего мозаичного графика начинается с размещения одной категориальной переменной вдоль оси *x* (здесь и далее — эпоха строительства моста) и деления оси *x* на относительные пропорции, составляющие эти категории. Затем мы помещаем другую категориальную переменную вдоль оси *y* (здесь — строительный материал) и внутри каждой категории вдоль оси *x* делим ось *y* на относительные пропорции, соответствующие значениям элементов категории *y* для этой категории по *x*. Результатом станет набор прямоугольников, площади которых пропорциональны количеству случаев, относящихся к каждой возможной комбинации двух категориальных переменных.

Существует еще один вариант визуализации данных о мостах — родственный мозаичному, однако имеющий собственный формат графика, который называется «деревом» (иногда — «древовидной картой»). На дереве, как и на мозаичном графике, мы берем ограничивающий прямоугольник и делим его на более мелкие прямоугольники, площадь которых представляет собой пропорции. Однако метод размещения вложенных прямоугольников в охватывающие отличается от метода, используемого в мозаичных графиках: в случае древовидной карты мы рекурсивно размещаем прямоугольники внутри друг друга. Например, говоря о мостах города Питтсбург, мы можем сначала разделить общую площадь на три части, каждая из которых представляет один из трех строительных материалов: дерево, железо и сталь. Затем мы

снова разбиваем каждую из этих областей, чтобы каждому строительному материалу соответствовала своя эпоха (рис. 10.4). Теоретически мы могли бы продолжить разбиение прямоугольников на более мелкие, однако в этом случае исходное изображение очень быстро станет громоздким и запутанным.

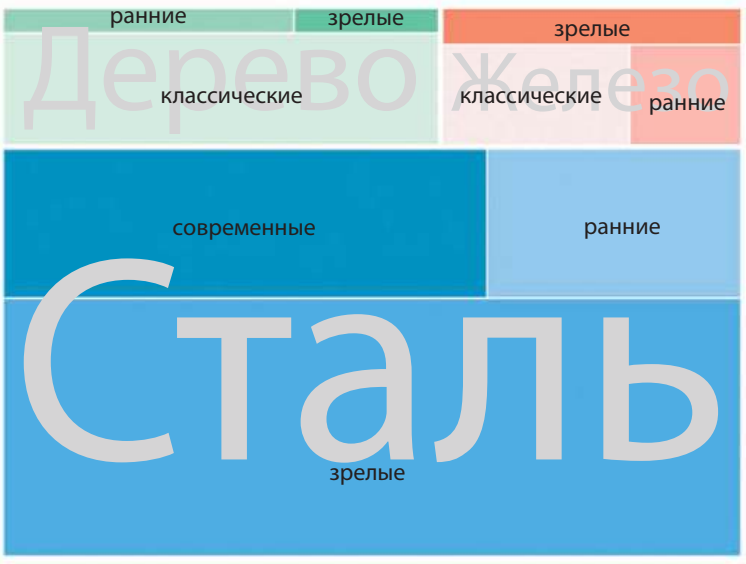


Рис. 10.4. Древовидная карта, которая визуализирует разбиение мостов города Питтсбург в зависимости от материала конструкции (сталь, дерево, железо) и эпохи постройки (классические, ранние, зрелые, современные). Площадь каждого прямоугольника пропорциональна количеству мостов этого типа. Источник: Yoram Reich and Steven J. Fenves

Несмотря на то что мозаичный график и древовидная карта тесно связаны между собой, у каждого из этих типов диаграмм имеется своя «специализация». В нашем случае мозаичный график (см. рис. 10.3) подчеркивает эволюцию использования строительных материалов в зависимости от эпохи, в то время как древовидная карта (см. рис. 10.4) делает акцент на общее количество стальных, железных и деревянных мостов.

Если брать шире, мозаичные графики предполагают, что все пропорции могут быть определены с помощью комбинаций из двух или более ортогональных категориальных переменных. К примеру, на рис. 10.3 каждый мост может быть описан выбором типа строительного материала (дерево, железо, сталь) и даты постройки (классические, ранние, зрелые, современные). Более того, возможна любая комбинация этих двух переменных, хотя на практике это происходит не всегда (у нас в наборе данных нет ни стальных классических мостов, ни деревянных или железных современных мостов). Однако для случая древовидных карт такого требования нет. По сути, древовидные карты

прекрасно работают даже в тех случаях, когда пропорции нельзя описать с помощью объединения нескольких категориальных переменных. К примеру, мы можем разделить США на четыре региона (Запад, Северо-Восток, Средний Запад и Юг), а каждый регион на отдельные штаты, при этом штаты в одном регионе не имеют отношения к штатам в другом регионе (рис. 10.5).

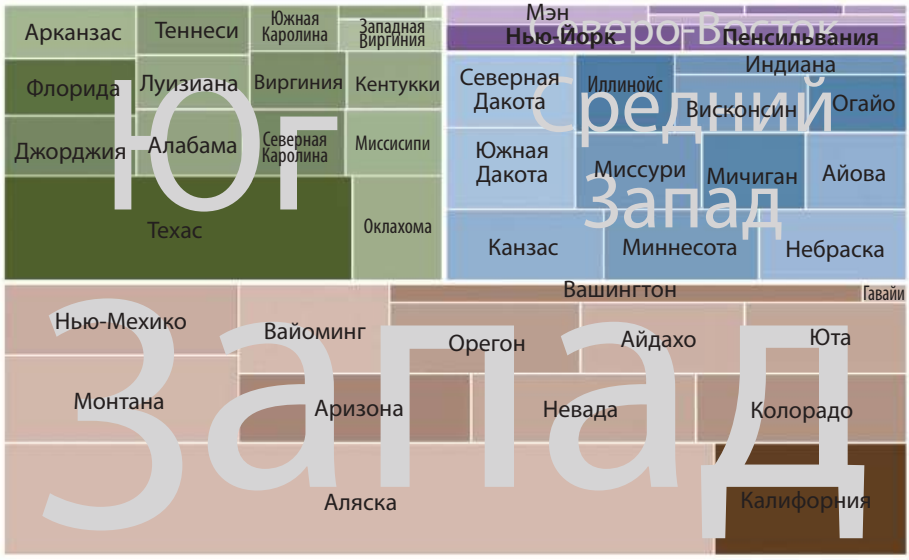


Рис. 10.5. Штаты США изображены в виде древовидной карты. Каждый прямоугольник представляет собой один штат, а площадь каждого прямоугольника пропорциональна площади территории штата. Штаты сгруппированы в четыре региона: Запад, Северо-Восток, Средний Запад и Юг. Окраска пропорциональна количеству жителей в каждом штате, при этом темные цвета означают большее количество жителей. Источник: Перепись населения США в очередном десятилетии, 2010 год

Мозаичные графики и древовидные карты широко распространены благодаря своей способности отображать большое количество информации, однако у этих типов графиков имеются ограничения, сходные с ограничениями столбчатых диаграмм с накоплением (табл. 9.1); в частности, может быть затруднено прямое сравнение данных, поскольку разные прямоугольники не всегда имеют общие основания, с помощью которых проводится визуальное сравнение. В случае мозаичных графиков и древовидных карт эта проблема усугубляется тем, что формы различных прямоугольников могут различаться. Например, среди ранних и зрелых мостов имеется одинаковое количество железных мостов (три), но увидеть этот факт на мозаичном графике (см. рис. 10.3) очень трудно, поскольку два прямоугольника, представляющие эти группы из трех мостов, имеют совершенно разные формы. Не факт, что эту проблему можно (и нужно ли?) решить, поскольку визуализация вложенных пропорций — задача непростая. А вообще,

когда это возможно, я просто рекомендую указывать на графике фактические значения или доли, чтобы читатели могли проверить правильность своей интуитивной интерпретации раскрашенных областей.

Многоуровневые круговые диаграммы

В начале этой главы я визуализировал набор данных о мостах с помощью неправильной круговой диаграммы (см. рис. 10.1), а затем привел аргументы в пользу того, что более подходящим типом визуализации является мозаичный график или древовидная карта. Следует отметить, однако, что оба эти вида графиков тесно связаны с круговыми диаграммами, поскольку все они используют площадь для представления значений данных. Основное различие заключается в типе системы координат: полярная в случае круговой диаграммы и декартова в случае мозаичного графика или древовидной карты. Такая тесная взаимосвязь между этими различными графиками наводит на мысль о том, что, возможно, есть какой-то вариант круговой диаграммы, с помощью которого можно было бы корректно визуализировать наш набор данных о мостах.

И действительно, существует два варианта решения данной проблемы именно с помощью круговых диаграмм. Первый вариант заключается в том, что мы можем нарисовать круговую диаграмму, состоящую из двух частей: внутренней и внешней (рис. 10.6). Внутренний круг показывает разбиение данных по категориям одной переменной (здесь — строительный материал), а внешний круг показывает разбиение на части каждого кусочка внутреннего круга по второй переменной (здесь — эпоха постройки моста). Такая визуализация выглядит логично, но я предвзято отношусь к этому подходу, поэтому отнес его к категории «некрасивых». Дело в том, что два отдельных круга скрывают тот факт, что каждый мост в наборе данных характеризуется и строительным материалом, и датой постройки. Чтобы убедиться в правильности моих слов, обратите внимание на рис. 10.6. На нем мы по-прежнему дважды учитываем каждый мост. Если взять сумму всех чисел, показанных в обеих окружностях, результат будет равен 212, что в два раза больше числа мостов, присутствующих в наборе данных.

В качестве альтернативы мы можем сначала разделить диаграмму на части, представляющие пропорции в соответствии с одной переменной (например, материал), а затем повторно разделить эти части в соответствии с другой переменной (дата постройки) (рис. 10.7). В результате получится привычная круговая диаграмма с большим количеством маленьких кусочков. При этом мы можем воспользоваться разными цветами для того, чтобы подчеркнуть наличие на нашей диаграмме нескольких уровней: на рис. 10.7 зеленым цветом показаны деревянные мосты, оранжевым — железные, а синим цветом — стальные. Интенсивность каждого цвета обозначает эпоху строительства: более темные цвета соответствуют мостам, построенным позже. Используя цветовую шкалу,

мы можем визуализировать разбиение данных как по первичной переменной (строительный материал), так и по вторичной (дата постройки).

Рис. 10.6. Вложенная круговая диаграмма иллюстрирует разбиение мостов города Питтсбург в зависимости от материалов конструкции (сталь, дерево, железо; внутренний круг) и эпохи постройки (классические, ранние, зрелые, современные; внешний круг). Источник: Yoram Reich and Steven J. Fenves

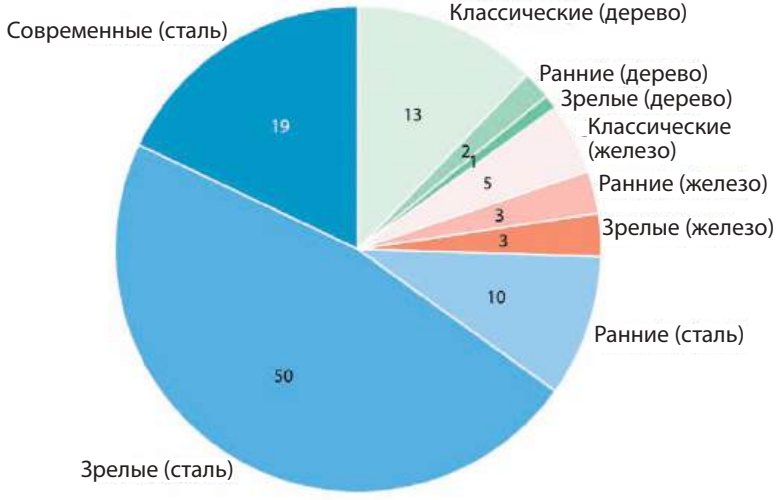


Рис. 10.7. Круговая диаграмма иллюстрирует разбиение мостов города Питтсбург в зависимости от материалов конструкции (сталь, дерево, железо) и даты постройки (классические, ранние, зрелые, современные). Цифры означают количество мостов в каждой из категорий. Источник: Yoram Reich and Steven J. Fenves

Круговая диаграмма на рис. 10.7 корректно иллюстрирует набор данных о мостах, но, если сравнивать ее напрямую с аналогичной древовидной картой (см. рис. 10.4), я выберу последний вариант и вот почему. Во-первых, прямоугольная форма древовидной карты позволяет более рационально использовать имеющееся пространство.

Размер рис. 10.4 и 10.7 одинаков, но на втором графике бóльшая часть пространства пуста, а древовидная карта на рис. 10.4 практически не содержит пустот. Это обстоятельство важно тем, что благодаря ему мы можем разместить подписи внутри раскрашенных областей: подписи, расположенные внутри элементов инфографики, воспринимаются как неотъемлемая часть данных, в отличие от подписей, расположенных снаружи изображения, и поэтому являются предпочтительными. Во-вторых, некоторые сектора круговой диаграммы, показанные на рис. 10.7, очень тонкие, из-за чего их трудно рассмотреть. Что же касается рис. 10.4, то каждый прямоугольник на нем имеет адекватный размер.

Диаграммы в параллельных координатах

При визуализации пропорций, описанных более чем двумя категориальными переменными, мозаичные графики, древовидные карты и круговые диаграммы зачастую очень быстро становятся громоздкими. В качестве приемлемой альтернативы в таких случаях можно использовать *диаграмму в параллельных координатах*. На такой диаграмме мы показываем, как общий набор данных разбивается в соответствии с каждой отдельной категориальной переменной, а затем рисуем цветные полосы, показывающие, как подгруппы соотносятся друг с другом. Пример такого графика приведен на рис. 10.8.

Здесь я разбил набор данных о мостах на следующие категории: материал конструкции (железо, сталь, дерево), длина моста (длинные, средние, короткие), дата постройки моста (классические, ранние, зрелые, современные) и река, через которую перекинут мост (Аллегейни, Мононгахила, Огайо). Ленты, соединяющие параллельные множества, окрашены в цвет их строительного материала.

Это показывает, например, что деревянные мосты в основном имеют среднюю длину (плюс несколько коротких мостов), относятся к категории классических (плюс несколько мостов средней длины, относящихся к категории ранних и зрелых) и в основном пересекают реку Аллегейни (плюс несколько мостов, перекинутых через реку Мононгахила). Напротив, все железные мосты средней длины, относящиеся в основном к категории классических, охватывают реки Аллегейни и Мононгахила примерно в равных пропорциях.

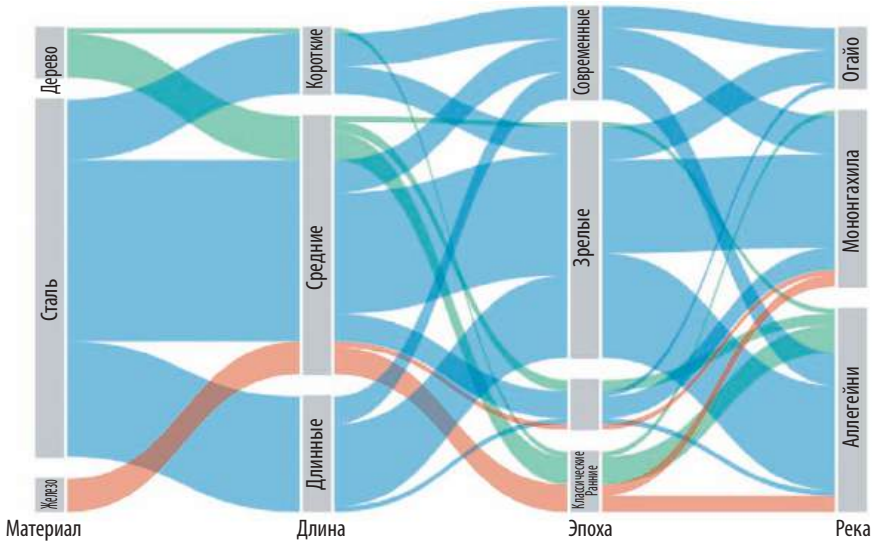


Рис. 10.8. Диаграмма в параллельных координатах описывает разбиение мостов города Питтсбург в зависимости от материала конструкции, длины моста, эпохи постройки и реки, через которую мост перекинут. Цвет полос означает тип строительного материала моста. Источник: Yoram Reich and Steven J. Fenves

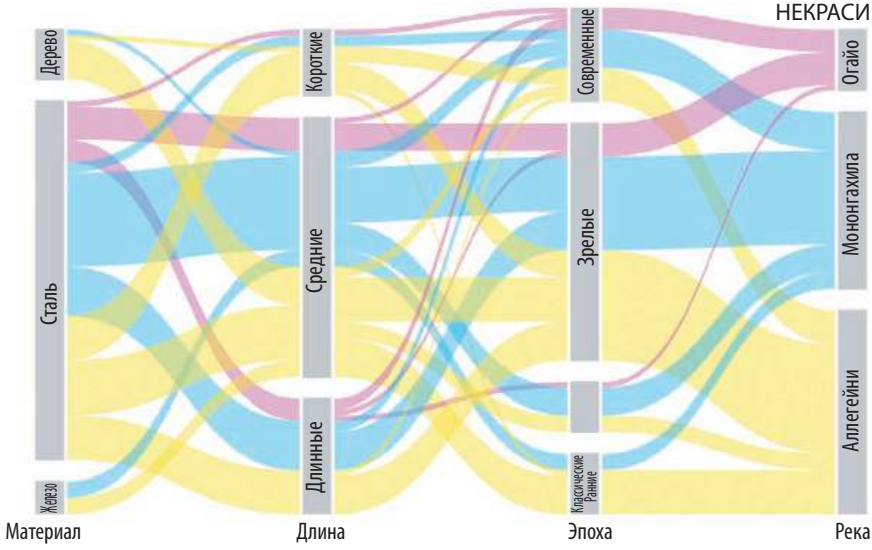


Рис. 10.9. Диаграмма в параллельных координатах иллюстрирует разбиение мостов города Питтсбург в зависимости от материалов конструкции, длины моста, эпохи постройки и реки, через которую мост перекинут. Данное изображение сходно с рис. 10.8, но теперь цветом выделены не строительные материалы, а пересекаемые мостами реки. Я отнес это изображение к категории «некрасивых», так как расположение цветных полос ближе к центру визуально перегружает рисунок, а также потому, что график должен читаться справа налево. Источник: Yoram Reich and Steven J. Fenves

Однако, если мы раскрасим элементы в соответствии с другим критерием, например в зависимости от реки, через которую перекинут мост (рис. 10.9), та же самая визуализация будет выглядеть совершенно иначе. График выглядит перегруженным из-за большого количества пересекающихся полос, но зато на нем видно, что для каждой реки можно найти практически любой мост любого типа.

Рисунок 10.9 я отнес к категории «некрасивых», потому что считаю его визуально перегруженным и запутанным. Во-первых, поскольку мы привыкли читать слева направо, я думаю, что цветные полосы должны исходить не из правого набора множеств, а из левого. Это упрощает визуальное восприятие, помогает понять, откуда берется цвет и как он проходит через набор данных. Во-вторых, желательно изменить категории таким образом, чтобы количество пересечений лент было сведено к минимуму. Следуя этим условиям, я создал рис. 10.10, который считаю более качественным, нежели рис. 10.9.

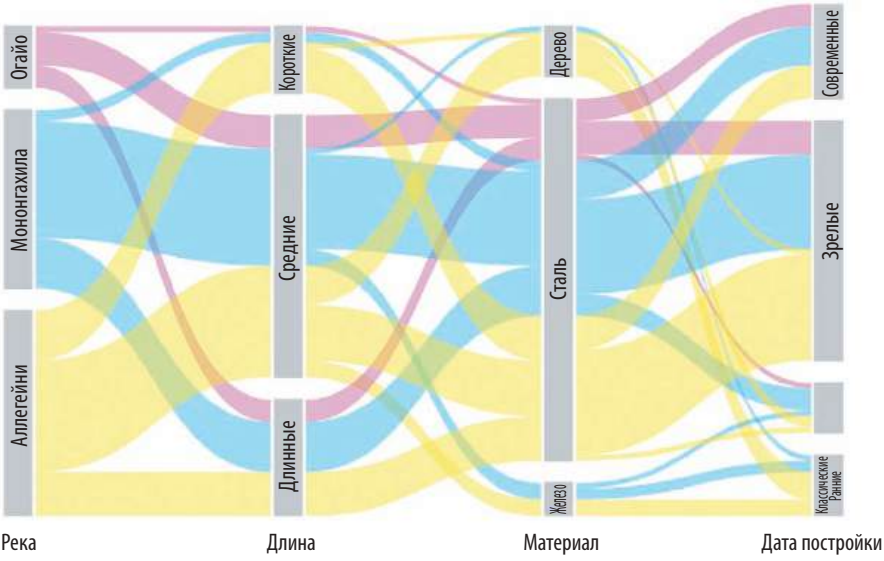


Рис. 10.10. Диаграмма в параллельных координатах иллюстрирует разбиение мостов города Питтсбург в зависимости от материала конструкции, длины моста, эпохи постройки и реки, через которую мост перекинут. Данная визуализация отличается от рис. 10.9 только порядком следования категорий. Изменение порядка позволило визуально разгрузить график и упростить его восприятие. Источник: Yoram Reich and Steven J. Fenves

Глава 11

Визуализация связей между двумя и более количественными переменными

Поскольку многие наборы данных содержат две или более количественных переменных, нередко возникает вопрос, как эти переменные соотносятся друг с другом. Рассмотрим пример с набором данных, который содержит такие количественные характеристики животных, как рост, вес, длина тела и суточная потребность в энергии. Для иллюстрации соотношения только двух переменных, таких как рост и вес, мы обычно используем диаграмму рассеяния. Если же требуется показать более двух переменных одновременно, мы можем выбрать пузырьковый график, матрицу диаграмм рассеяния или коррелограмму. Наконец, для наборов данных с большим количеством измерений может быть полезно понизить размерность, например, при помощи анализа главных компонент.

Диаграммы рассеяния

Как выглядят диаграмма рассеяния и ее вариации, я продемонстрирую на примере набора данных, который представляет собой результаты измерений 123 голубых соек. Массив данных содержит такую информацию, как длина головы (измеренная от кончика клюва до затылка), размер черепа (длина головы минус длина клюва) и масса тела каждой птицы. Предположительно, между этими переменными существуют связи. Например, ожидается, что у птиц с более длинным клювом череп будет иметь больший размер, а у птиц с более высокой массой тела — клюв и череп будут больше, чем у птиц с меньшей массой тела.

Чтобы исследовать взаимосвязь этих величин, я начну с диаграммы, которая показывает зависимость между длиной головы и массой тела сойки (рис. 11.1). На этом графике длина головы показана вдоль оси *y*, масса тела — вдоль оси *x*, а каждая птица представлена одной точкой (в случаях, когда

график призван отразить связь между переменными, обычно вдоль оси *y* откладывают предполагаемую зависимую переменную, а вдоль оси *x* — независимую). Точки образуют фигуру, похожую на рассеянное облако (отсюда термин «диаграмма рассеяния»), но на графике хорошо заметна тенденция, что у птиц с большей массой тела длина головы будет больше. Точка, отображающая птицу с самой длинной головой, располагается вблизи значения максимальной наблюдаемой массы тела, а точка птицы с самой короткой головой близка к величине минимальной наблюдаемой массы тела.

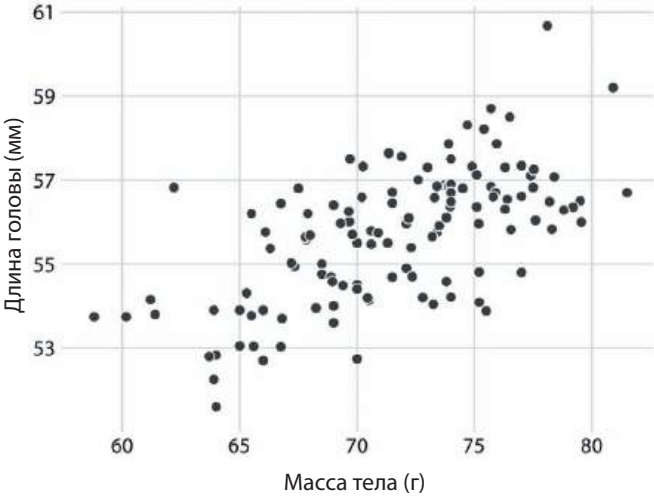


Рис. 11.1. Отношение длины головы (измеряется от кончика клюва до затылка) к массе тела (в граммах) для 123 голубых соек. Каждая точка на графике соответствует одной птице. Для более массивных птиц характерна умеренная тенденция иметь более длинные головы. Источник: Keith Tarvin, Oberlin College

Набор данных о голубых сойках содержит информацию как о мужских, так и о женских особях, и, вероятно, мы можем захотеть узнать, присутствует ли обнаруженная нами зависимость у обоих полов. Чтобы ответить на этот вопрос, мы можем раскрасить точки на диаграмме в зависимости от пола птицы (рис. 11.2). Полученный рисунок свидетельствует о том, что общая тенденция отношения длины головы к массе тела по крайней мере частично зависит от пола птиц: при равной массе тела самки имеют меньшую длину головы, нежели самцы. Кроме того, самки вообще характеризуются меньшим весом.

Поскольку длина головы определяется как расстояние от кончика клюва до затылка, большая длина головы может означать как бóльшую длину клюва, так и бóльшую длину черепа, или и то и другое одновременно. Чтобы прояснить этот момент, следует обратить внимание на другую переменную в наборе данных — размер черепа. Этот показатель схож с показателем длины головы, но при этом не включает в себя длину клюва. Поскольку мы уже

используем ось *x* для определения массы тела, ось *y* для определения длины головы и цвет точки для определения пола птицы, нам нужен еще какой-то визуальный элемент, с помощью которого мы сможем показать длину черепа. Таким элементом может стать, например, размер точек, задействуя который мы получим визуализацию, называемую *пузырьковой диаграммой* (рис. 11.3).

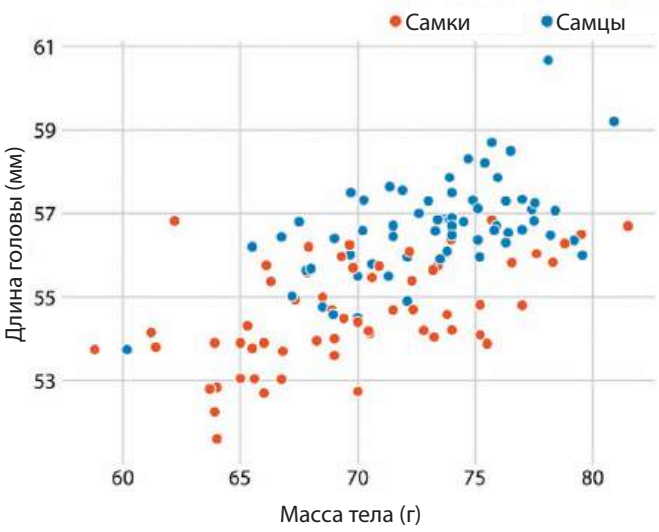


Рис. 11.2. Связь между длиной головы и массой тела для 123 голубых соек. Пол птиц выражен цветом точек. При одинаковой массе тела самцы имеют более длинные головы (а соответственно, и клювы), нежели самки. Источник: Keith Tarvin, Oberlin College

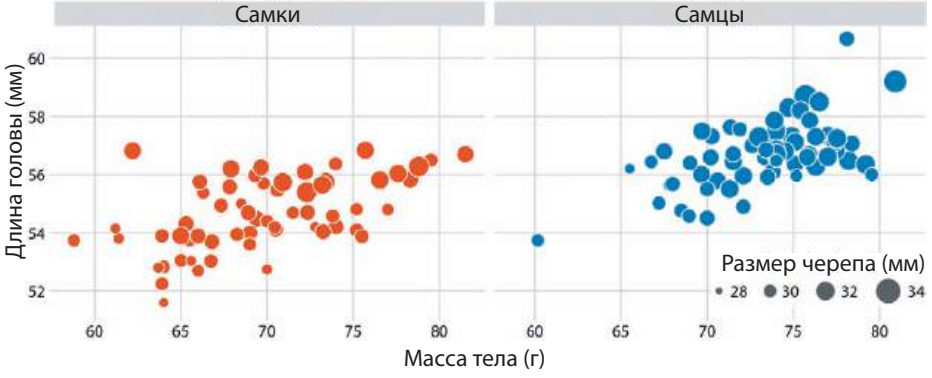


Рис. 11.3. Связь между длиной головы и массой тела для 123 голубых соек. Пол птицы обозначается цветом точки, а размер черепа птицы — размером точки. Показатель длины головы включает в себя длину клюва, в то время как показатель размера черепа — нет. Длина головы и размер черепа обычно коррелируют, однако в наборе данных есть птицы с необычно длинными или короткими клювами, учитывая размер их черепа. Источник: Keith Tarvin, Oberlin College

Недостатком пузырьковых диаграмм является то, что они показывают одни и те же типы переменных — количественные — с двумя различными типами шкал (положением и размером). Из-за этого бывает сложно оценить на взгляд, насколько существенна связь между разными переменными. Кроме того, различия в значениях данных, если они представлены размером точек, воспринимаются хуже, чем различия в значениях данных, представленных положением точек.

Поскольку даже самые большие пузырьки должны быть меньше общего размера изображения, разница в размерах между самыми большими и самыми маленькими из них будет незначительной. Отсюда следует вывод, что малые различия в значениях данных будут соответствовать очень малым различиям в размерах, заметить которые будет очень сложно. На рис. 11.3 я добавил легенду размеров, которая визуально усиливает разницу между наименьшими (около 28 мм) и наибольшими (около 34 мм) черепами, однако понять, существует ли какая-то связь между размером черепа и массой тела или длиной головы все еще очень сложно.

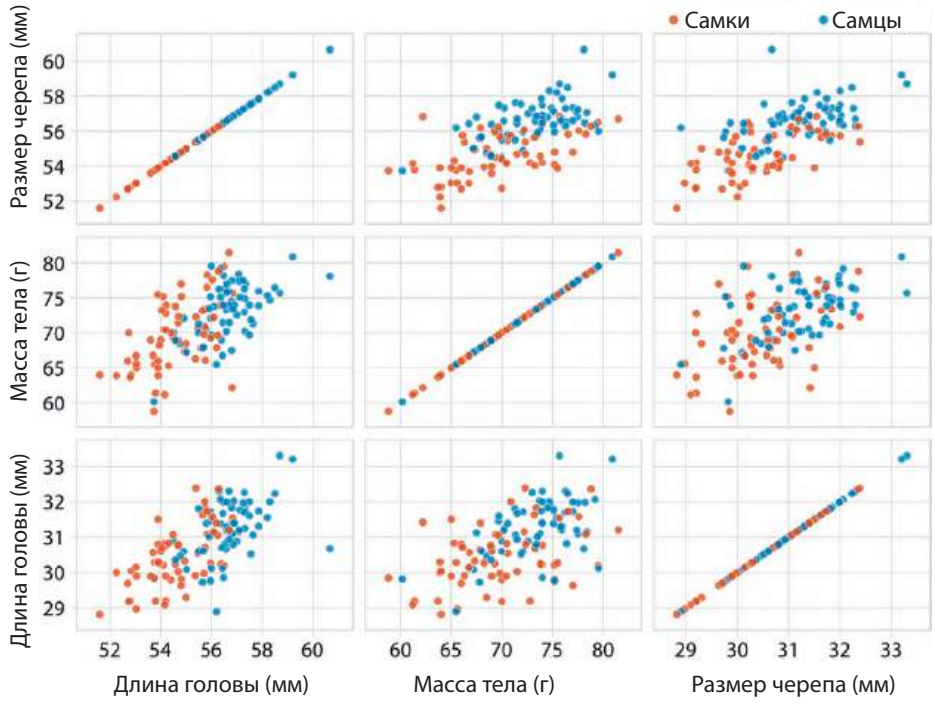


Рис. 11.4. Матрица диаграмм рассеяния, визуализирующая связи между длиной головы, массой тела и размером черепа для 123 голубых соек, построенная на том же массиве данных, что и рис. 11.2. Поскольку положение точки на графике оценить проще, чем ее размер, связь между размером черепа и двумя другими переменными видна здесь лучше, чем на рис. 11.2. Источник: Keith Tarvin, Oberlin College

Альтернативой пузырьковой диаграммы может служить матрица, содержащая графики рассеяния, построенные для всех возможных пар переменных. В этом случае каждый отдельный график будет отображать два измерения данных (рис. 11.4). По этим графикам легко заметить, что соотношение размера черепа и массы тела у самцов и самок сопоставимо, за исключением того, что самки, как правило, несколько меньше. Однако, когда мы говорим о соотношении длины головы и массы тела, это правило перестает работать, потому что появляется четкое разделение по полу. Самцы, как правило, имеют более длинные клювы, чем самки, при прочих равных условиях.

Коррелограммы

Когда мы сталкиваемся с набором данных, который содержит более трех-четырех количественных переменных, матрица диаграмм рассеяния, пример которой мы видели на рис. 11.4, становится слишком громоздкой. В этом случае имеет смысл рассчитать количественную оценку степени связи пар переменных и визуализировать уже эти данные, а не исходные. Чаще всего это делается с помощью расчета *коэффициентов корреляции*. Коэффициент корреляции r — это число на отрезке между –1 и 1, которое показывает степень взаимосвязи (взаимоизменчивости) двух переменных. Значение $r = 0$ означает полное отсутствие связи, а значение 1 или –1 указывает на линейную зависимость. Знак коэффициента корреляции показывает, является ли зависимость прямо пропорциональной (когда бóльшие значения одной переменной соответствуют бóльшим значениями другой) или обратно пропорциональной (когда бóльшие значения одной переменной соответствуют меньшим значениям другой). Наглядные примеры различных коррелограмм приведены на рис. 11.5, где показаны сгенерированные случайным образом наборы точек, которые сильно различаются по степени связи значений x и y.

Коэффициент корреляции рассчитывается следующим образом:

$$r = \frac{\Sigma_i \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\Sigma_i \left(x_i - \bar{x}\right)^2}\sqrt{\Sigma_i \left(y_i - \bar{y}\right)^2}},$$

где x_i и y_i — это два набора наблюдений, а $\bar{x}$ и $\bar{y}$ — соответствующие выборочные средние. Глядя на эту формулу, мы можем сделать следующие выводы. Во-первых, формула симметрична относительно x_i и y_i, поэтому отношение x и y равно отношению y и x. Во-вторых, индивидуальные значения x_i и y_i вводятся в формулу только в контексте разницы с соответствующим выборочным средним, поэтому если сдвинуть весь набор данных на фиксированное значение, например, заменить x_i на $x_i' = x_i + C$ для некоторой постоянной C, то

коэффициент корреляции останется неизменным. В-третьих, коэффициент корреляции останется прежним и в случае умножения данных на некоторую постоянную (например, $x_i' = Cx_i$), поскольку константа C будет присутствовать как в числителе, так и в знаменателе формулы и поэтому может быть сокращена.

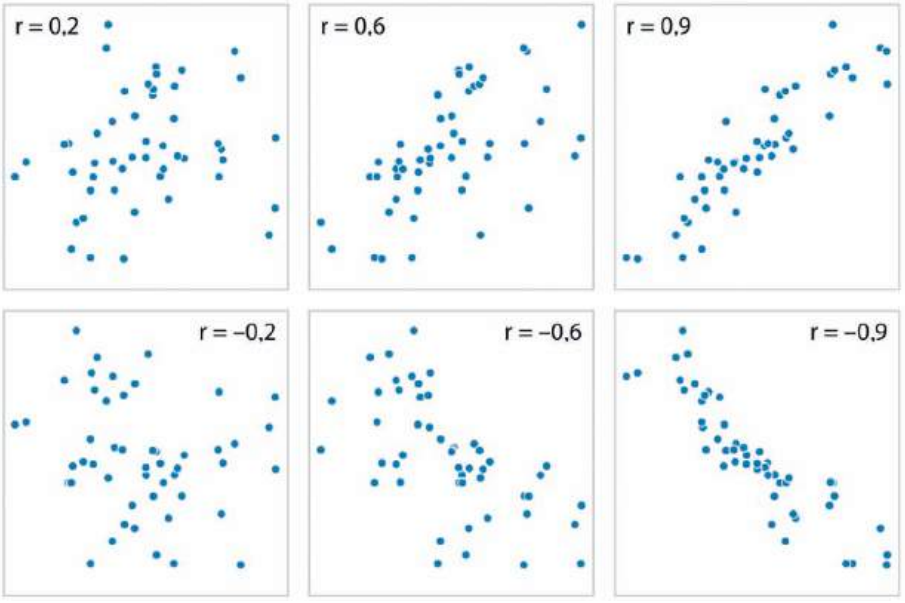


Рис. 11.5. Примеры корреляций различных величин и направлений с соответствующим коэффициентом корреляции *r*. В обоих рядах слева направо взаимосвязи переходят от слабых к сильным. В верхней строке переменные прямо пропорциональны (бóльшие значения одной величины связаны с бóльшими значениями другой), а в нижней строке обратно пропорциональны (бóльшие значения одной величины связаны с меньшими значениями для другой). На всех шести графиках наборы значений *x* и *y* идентичны, но пары некоторых значений *x* и *y* были перетасованы для получения указанных коэффициентов корреляции

Визуализации коэффициентов корреляции называются *коррелограммами*. Чтобы понять, каким образом их можно использовать, мы рассмотрим набор данных, который содержит информацию о более чем 200 осколках стекла, полученную в ходе криминалистической экспертизы. Нам известен состав каждого осколка стекла, который представляет собой набор массовых долей различных оксидов минералов. Измерения включают в себя доли 7 оксидов, которые дают в общей сложности 6 + 5 + 4 + 3 + 2 + 1 = 21 парную корреляцию. Мы можем отобразить эти взаимосвязи в виде матрицы, состоящей из цветных плиток, где каждая плитка представляет собой один коэффициент корреляции (рис. 11.6). Благодаря этой коррелограмме мы можем быстро

уловить следующие тенденции: показатели магния обратно пропорциональны показателям почти всех остальных оксидов, а алюминий и барий имеют выраженную прямую пропорциональность.

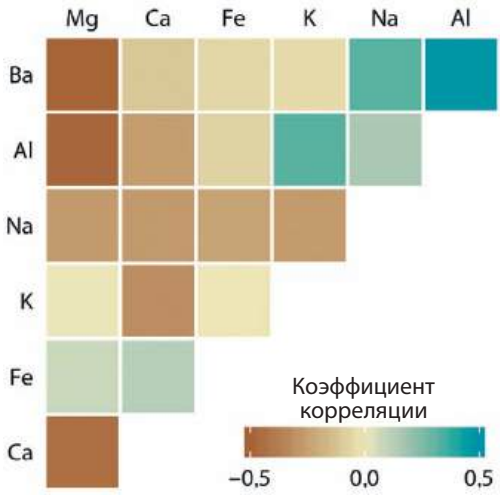


Рис. 11.6. Корреляции в минеральном составе 214 фрагментов стекла, изученных в ходе криминалистической экспертизы. Набор данных содержит семь переменных величин: количество магния (Mg), кальция (Ca), железа (Fe), калия (K), натрия (Na), алюминия (Al) и бария (Ba), обнаруженных в каждом фрагменте. Цветные плитки обозначают корреляции между парами этих переменных. Источник: B. German

Одной из слабых сторон коррелограммы на рис. 11.6 является то, что значения с низким коэффициентом корреляции, то есть корреляции с абсолютным значением вблизи нуля, заметны больше, чем следует. Например, магний (Mg) и калий (K) не имеют никакой взаимосвязи друг с другом, однако, глядя на рис. 11.6, понять это сложно. Для преодоления этого ограничения мы можем отобразить корреляции в виде цветных окружностей и масштабировать размер каждой из них согласно абсолютному (то есть по модулю) значению коэффициента корреляции (рис. 11.7). Таким образом, низкие коэффициенты корреляции будут менее заметны, а высокие — наоборот.

У всех коррелограмм есть существенный недостаток: они довольно абстрактны. С одной стороны, графики этого типа позволяют отразить наиболее существенные закономерности данных, но в то же время они скрывают часть важных показателей, из-за чего читатель может сделать неверные выводы. Безусловно, всегда лучше визуализировать исходные данные, а не производные от них показатели. К счастью, существует один интересный метод визуализации, который позволяет найти золотую середину между отображением на графике тенденций в определенном наборе данных и отображением исходных данных.

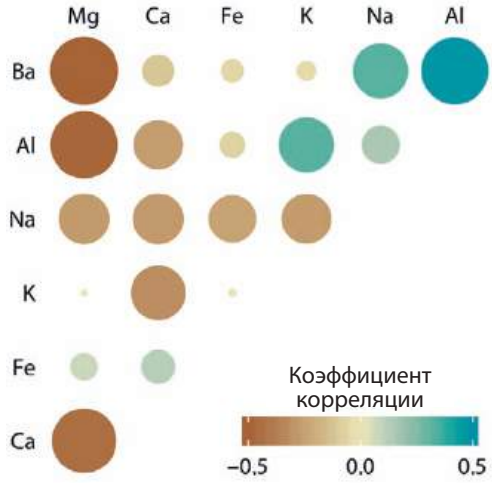


Рис. 11.7. Корреляции в минеральном составе 214 фрагментов стекла. Цветовая шкала идентична указанной на рис. 11.7, однако теперь величина коэффициента корреляции также выражена в размерах окружностей. Подобный подход позволяет сделать визуальный акцент на тех корреляциях, коэффициент которых много выше нуля, и убрать на второй план близкие к нулю. Источник: B. German

Снижение размерности

Метод под названием «снижение размерности» применим в предположении, что многомерный набор данных содержит множество взаимосвязанных в какой-то мере переменных, вследствие чего передаваемая ими информация частично дублируется. Если избавиться от этого дублирования, мы сможем уменьшить размер массива данных, не жертвуя при этом никакой важной информацией. В качестве простого и интуитивно понятного примера таких данных давайте рассмотрим набор сведений о физических особенностях человека: росте, весе, длине рук и ног, обхвате талии, бедер, груди и др. Априори понятно, что все эти характеристики зависят прежде всего от размера человека в целом. При прочих равных более крупный человек, как правило, имеет более высокий рост, бóльший вес, более длинные руки и ноги, бóльший обхват талии, бедер и груди. Следующим важным параметром данных является пол человека: мерки мужчин и женщин одного размера будут сильно различаться. К примеру, окружность бедер у женщин будет больше, чем у мужчин того же роста.

Существует огромный спектр техник снижения размерности. В этой книге мы поговорим только об одной из них — *методе главных компонент (МГК)** — как наиболее распространенной. Давайте посмотрим, как рабо-

* В зарубежной литературе обычно используется аббревиатура PCA — Principal Components Analysis. — *Прим. ред.*

тает этот метод. При использовании МГК создается набор переменных, называемых *главными компонентами*. Данные компоненты образуются путем линейной комбинации* исходных с приведением их к нулевому среднему (рис. 11.8 с примером для двух измерений). Главные компоненты выбираются таким образом, чтобы они не имели корреляционной связи друг с другом, и упорядочиваются так, чтобы первая компонента описывала максимальную долю вариации в данных, а каждая следующая — максимально возможную долю остаточной вариации. Как правило, ключевые особенности данных в достаточной степени содержатся в первых двух-трех компонентах.

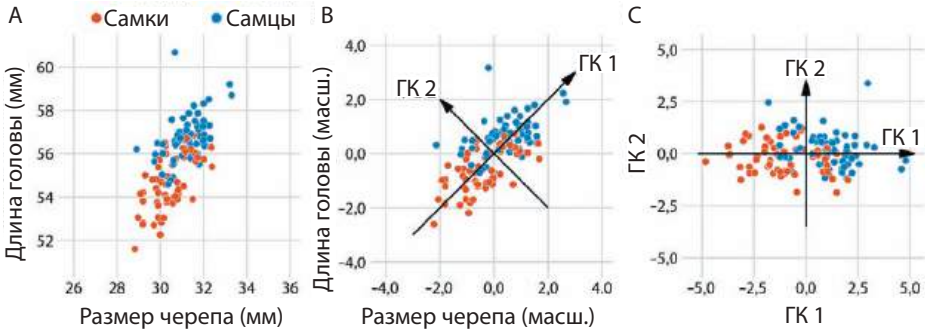


Рис. 11.8. Пример использования метода главных компонент для случая двух измерений. A. Исходные данные. В качестве примера я взял показатели длины головы и размера черепа из массива данных о голубых сойках. На графике самцы и самки различаются цветом, однако этот параметр не влияет на работу МГК. B. Первым шагом при использовании метода главных компонент будет масштабирование исходных данных для приведения их к нулевому среднему и единичной дисперсии. Далее мы определяем новый набор переменных (главных компонент) по направлениям максимальной вариации данных. C. Наконец, мы проецируем данные на новую систему координат. Математически эта проекция эквивалентна вращению точек данных вокруг начала координат. В приведенном здесь примере для двухмерного пространства точки данных повернуты по часовой стрелке на 45 градусов. Источник: Keith Tarvin, Oberlin College

При использовании метода главных компонент нас, как правило, интересуют два момента: состав главных компонент и расположение отдельных точек данных в пространстве главных компонент. Рассмотрим эти моменты на примере набора данных криминалистической экспертизы осколков стекла.

Начнем с состава главных компонент (рис. 11.9). В исходном массиве данных нас интересуют только две главные компоненты: ГК 1 и ГК 2. Так как они являются линейными комбинациями исходных переменных (после стандартизации), мы можем представить эти переменные в виде векторов (стрелок),

* Линейной комбинацией называется выражение, равное сумме элементов множества, умноженных на некоторые коэффициенты.

указывающих на степень их присутствия в главных компонентах. Как видно, барий и натрий находятся в основном в ГК 1, а не в ГК 2. Кальций и калий, наоборот, составляют ГК 2. Остальные переменные присутствуют в обоих компонентах в разном количестве. Стрелки имеют разную длину, поскольку из нашего массива данных можно вычленить более двух главных компонент. Например, длина стрелки железа очень мала, потому что железо большей частью присутствует в главных компонентах более высокого порядка (на рисунке не показаны).

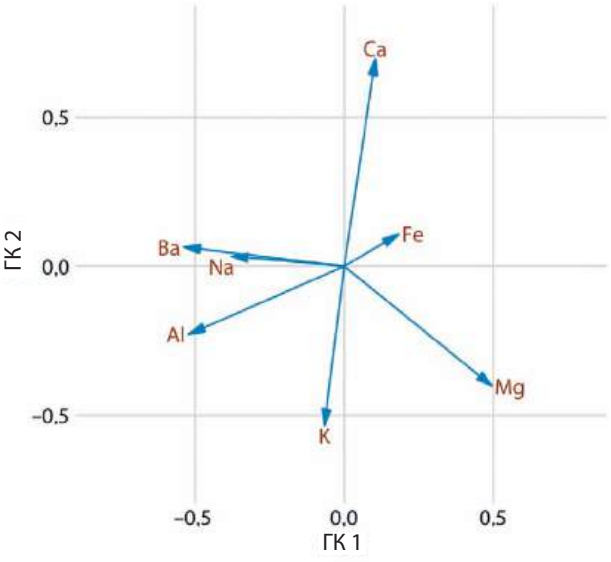


Рис. 11.9. Состав двух главных компонент массива данных криминалистической экспертизы стекла. Первая компонента (ГК 1) в основном показывает количество алюминия, бария, натрия и магния во фрагменте стекла, а вторая (ГК 2) — количество кальция и калия и лишь в некоторой степени — количество алюминия и магния. Источник: B. German

Далее, мы проецируем исходные данные на пространство главных компонент (рис. 11.10) и обогащаем информацией о том, от какого предмета взят осколок. На графике хорошо заметна выраженная группировка различных типов стекла: осколки фар и окон находятся на рисунке в четко очерченных областях с небольшими выбросами, осколки посуды и емкостей рассредоточены чуть сильнее, но тем не менее их легко отличить от осколков фар и окон. Сравнивая рис. 11.10 с рис. 11.9, можно сделать вывод, что в образцах оконных стекол содержание магния, как правило, выше среднего, а количество бария, алюминия и натрия — наоборот. При этом для остекления фар наблюдается прямо противоположная картина.

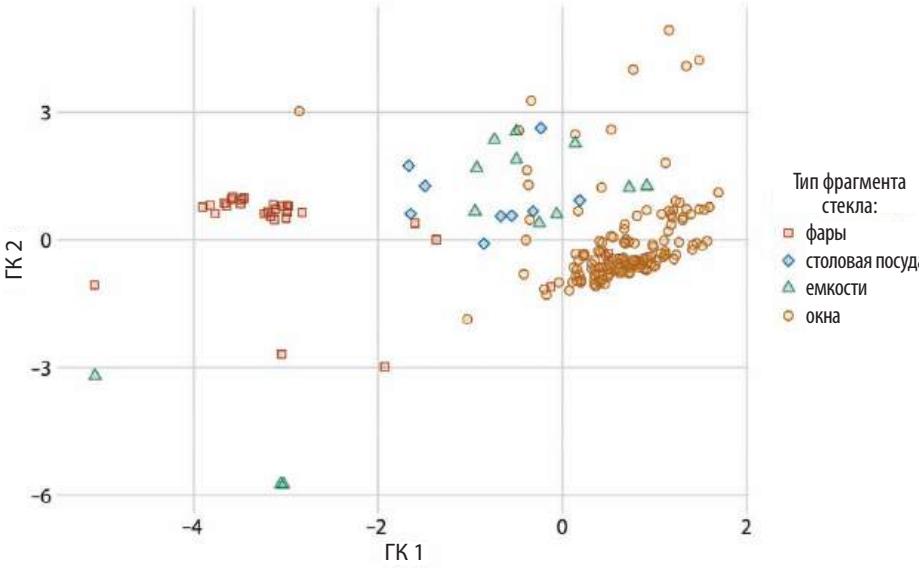


Рис. 11.10. Состав отдельных осколков стекла визуализирован в пространстве главных компонент, взятом с рис. 11.9. Мы видим, что различные типы стекол образуют группы в характерных значениях ГК 1 и 2. В частности, состав остекления фар характеризуется отрицательным значением ГК 1, в то время как состав оконного стекла, напротив, имеет положительное значение. Стеклянные посуда и емкости имеют значения ГК 1, близкие к нулю, и, как правило, положительные значения ГК 2. Следует отметить, что на графике присутствует несколько исключений, когда осколки емкостей имеют отрицательное значение и ГК 1, и ГК 2. Состав этих осколков кардинально отличается от состава всех остальных проанализированных осколков. Источник: B. German

Парные выборки

Особым случаем многомерных количественных данных являются *парные данные*: это данные, в которых присутствуют два или более результата измерений одной и той же величины при немного отличающихся условиях. Примерами таких массивов данных могут служить два сопоставимых результата измерения объекта (например, длина правой и левой руки человека), результаты повторных измерений одного и того же объекта в разные моменты времени (например, вес человека в два различных момента времени в течение года) или результаты измерения двух почти одинаковых объектов (например, рост двух близнецов). В случае парной выборки разумно предполагать, что два результата измерений, принадлежащие каждой из пар, похожи друг на друга больше, чем результаты, принадлежащие другим парам. Два близнеца будут иметь примерно один и тот же рост, однако значение роста других близнецов будет иным. Из этого следует, что для визуализации значений подобного

рода нужно выбирать такой способ, который подчеркнет даже самые малые различия между величинами. Наилучшим способом визуализации парных данных является простая диаграмма рассеяния, на которую дополнительно нанесена диагональная линия $x = y$. В случае, если единственным различием между двумя измерениями окажется случайный шум, точки на таком графике будут расположены симметрично относительно диагонали. Напротив, любые систематические различия в парных измерениях будут проявляться в массовом смещении точек выше или ниже диагонали. В качестве примера рассмотрим результаты замеров (сделанных в 1970 и 2010 годах) количества выбросов углекислого газа (CO_2), приходящихся на одного человека, в 166 странах (рис. 11.11). Этот образец данных хорошо иллюстрирует две общие черты парной выборки. Во-первых, большинство точек находится вблизи диагональной линии. Даже несмотря на то что выбросы CO_2 варьируются от страны к стране в пределах четырех порядков, объемы выбросов внутри каждой отдельной страны изменяются довольно последовательно на протяжении всех 40 лет. Во-вторых, хорошо заметно систематическое смещение точек выше диагонали. Это означает, что в течение рассмотренного периода времени в большинстве стран наблюдалось увеличение выбросов CO_2.

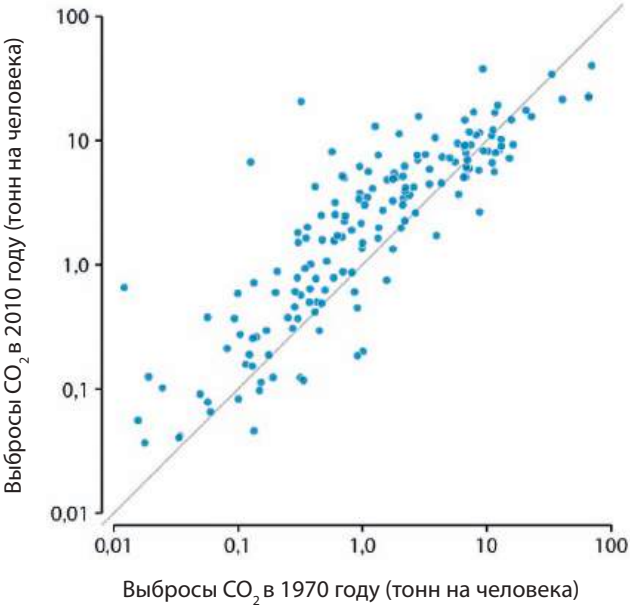


Рис. 11.11. Выбросы углекислого газа в расчете на одного человека в период с 1970 по 2010 год. Каждая точка символизирует одну страну. Диагональная линия символизирует одинаковое количество выбросов CO_2 в 1970 и в 2010 годах. Как видно, точки смещены заметно выше диагонали: в большинстве стран объем выбросов углекислого газа в 2010 году вырос по сравнению с 1970 годом. Источник: Carbon Dioxide Information Analysis Center

Диаграммы рассеяния, подобные той, что показана на рис. 11.11, хорошо работают только в тех случаях, когда массив изучаемых нами данных достаточно велик и/или нас интересует именно систематическое отклонение данных от нулевого ожидания. В тех случаях, когда мы имеем дело с небольшим объемом данных и нас интересует не общая картина, а каждый конкретный случай, лучшим способом визуализации, возможно, будет *диаграмма наклона (slopegraph)*. На графике такого типа отдельные данные изображаются точками, упорядоченными в виде двух столбцов и соединенными попарно. Наклон каждой линии означает величину и направление изменения. Этот подход используется на рис. 11.12 для того, чтобы показать 10 стран с наибольшими изменениями в выбросах CO_2 на человека за период с 2000 по 2010 год.

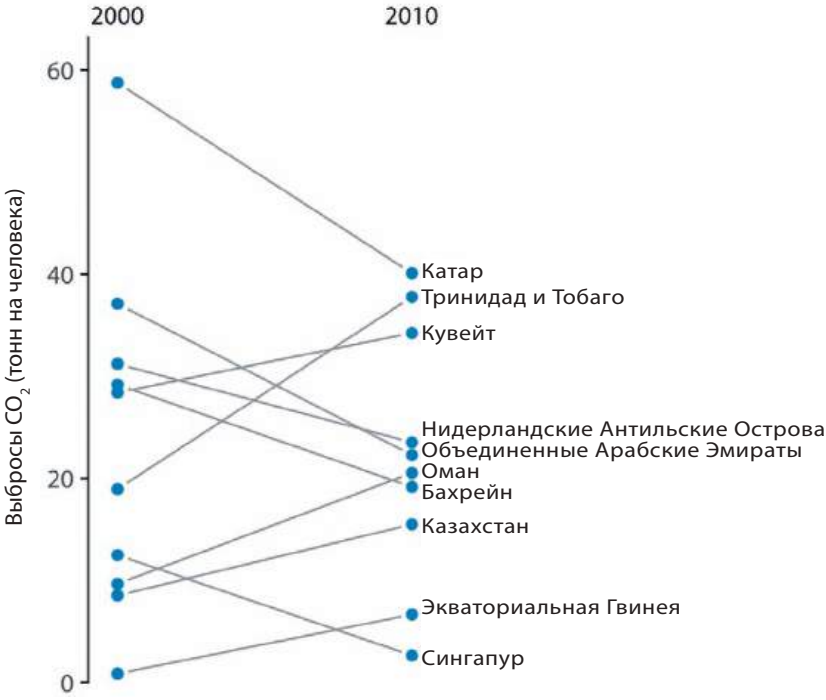


Рис. 11.12. Выбросы углекислого газа в расчете на одного человека за период с 2000 по 2010 год. На графике показаны 10 стран с наибольшими изменениями в объемах выбросов. Источник: Carbon Dioxide Information Analysis Center

У диаграмм наклона есть одно важное преимущество перед диаграммами рассеяния: их можно использовать для сравнения более чем двух результатов измерений одновременно. Например, мы можем изменить рис. 11.12 так, чтобы на нем отображались выбросы CO_2 в трех временных точках, например, в 2000, 2005 и 2010 годах (рис. 11.13). Таким образом мы сможем понять,

в каких странах количество выбросов больше всего изменилось за десять лет и в каких (Катар, Тринидад и Тобаго) наблюдается значительная разница в динамике изменения количества выбросов в течение первого пятилетнего периода и второго.

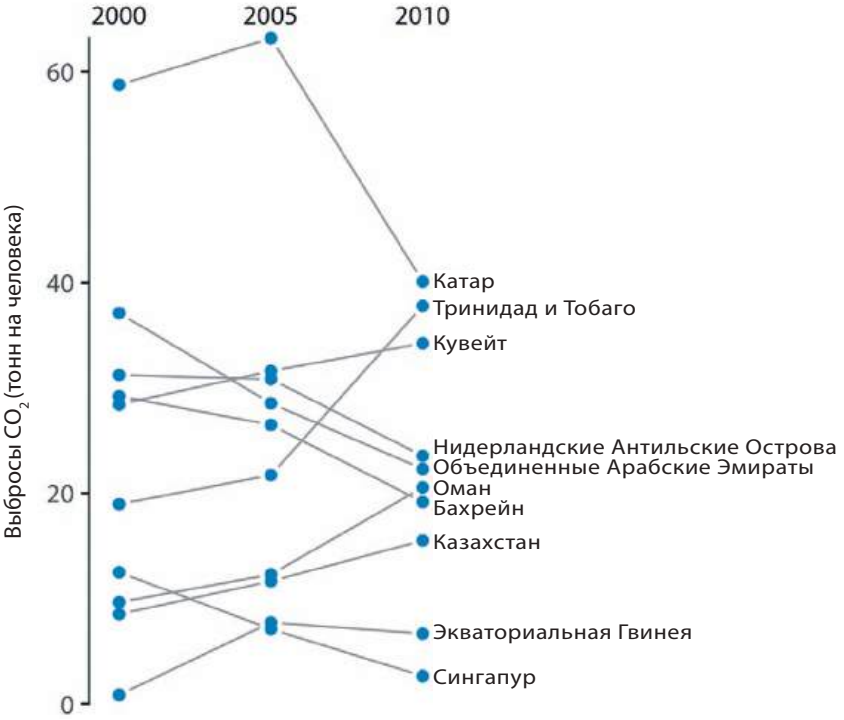


Рис. 11.13. Выбросы углекислого газа в расчете на одного человека в 2000, 2005 и 2010 годах. На графике показаны 10 стран с наибольшими изменениями в показателях в период с 2000 по 2010 год. Источник: Carbon Dioxide Information Analysis Center

Глава 12

Визуализация временных рядов и других функций независимой переменной

В предыдущей главе мы обсуждали диаграммы рассеяния, которые отображают связи между двумя количественными переменными. Ситуация с двумя переменными приобретает особые свойства, когда одна из переменных является временем, поскольку оно добавляет к данным дополнительную структуру. Данные приобретают определенный порядок: мы можем расставлять точки в порядке возрастания и убывания времени, чтобы определить предшественника и преемника для каждой точки на графике. Данные такого типа часто визуализируют с помощью диаграмм в виде ломаных линий. Хочу отметить, что использование таких графиков не ограничивается лишь временными рядами. Их можно использовать во всех тех случаях, когда одна переменная задает порядок для всего набора. Примером такого сценария является, например, управляемый эксперимент, в процессе которого изучаемому параметру намеренно задается диапазон различных значений. Если у нас есть несколько переменных, зависящих от времени, мы можем либо построить для каждой из них отдельный линейный график, либо нарисовать обычную диаграмму рассеяния, после чего соединить линиями соседние во времени точки.

Самостоятельные временные ряды

В качестве первой демонстрации временных рядов мы рассмотрим модель ежемесячного представления препринтов по биологии. Препринты — это научные статьи, которые исследователи размещают в интернете до официального рецензирования и публикации в научном журнале. Специализированный препринт-сервер bioRxiv, основанный в ноябре 2013 года и предназначенный для исследователей, работающих в области биологических наук, показал значительное увеличение количества ежемесячных заявок. Для визуализации этого процесса мы воспользуемся диаграммой рассеяния (см. главу 11), на которую нанесем точки, отражающие количество поданных в каждом месяце заявок (рис. 12.1).

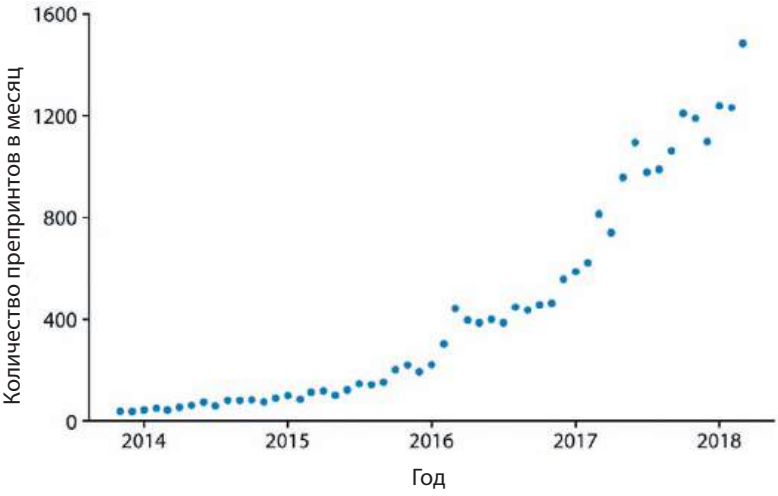


Рис. 12.1. Количество ежемесячных заявок, поданных на препринт-сервер bioRxiv за период с момента его создания в ноябре 2013 года по апрель 2018 года. Каждая точка соответствует количеству поданных за месяц заявок. Как видно из графика, на протяжении всех 4,5 года наблюдается стабильный рост количества заявок. Источник: Jordan Anaya, www.prepubmed.org

Между рис. 12.1 и диаграммами рассеяния, представленными в главе 11, существует важное различие: на рис. 12.1 точки расположены равномерно вдоль оси x в определенном порядке и у каждой точки есть только одна соседка с каждой стороны (за исключением крайних левых и крайних правых точек, которые имеют только одну соседнюю точку). Чтобы акцентировать внимание читателя на упорядочивании, можно соединить соседние точки линиями (рис. 12.2). Такая визуализация называется линейным графиком.

Некоторые специалисты по работе с данными считают, что следует избегать рисования декоративных линий, поскольку они не несут никакой смысловой нагрузки. Например, если у нас есть всего несколько измерений, которые расположены друг от друга на большом удалении, то в случае, если бы какие-то наблюдения были сделаны в периоды времени, которые находятся между наших точек, результат, скорее всего, не попал бы на проведенные линии. Таким образом, можно сказать, что подобного рода декоративные элементы являются в некотором смысле выдуманными данными. Тем не менее они могут помочь с восприятием графика, когда точки расположены далеко друг от друга или распределены неравномерно. В качестве частичного решения проблемы мы можем сообщить об этом факте в подписи к изображению, например: «Линии проведены для удобства восприятия» (см. подпись к рис. 12.2).

Использование линий для представления временных рядов является общепринятой практикой, и зачастую точки на графике просто опускаются (рис. 12.3). Таким образом внимание читателя больше сконцентрировано на

общей тенденции, которую сообщают данные, и меньше на отдельных наблюдениях.

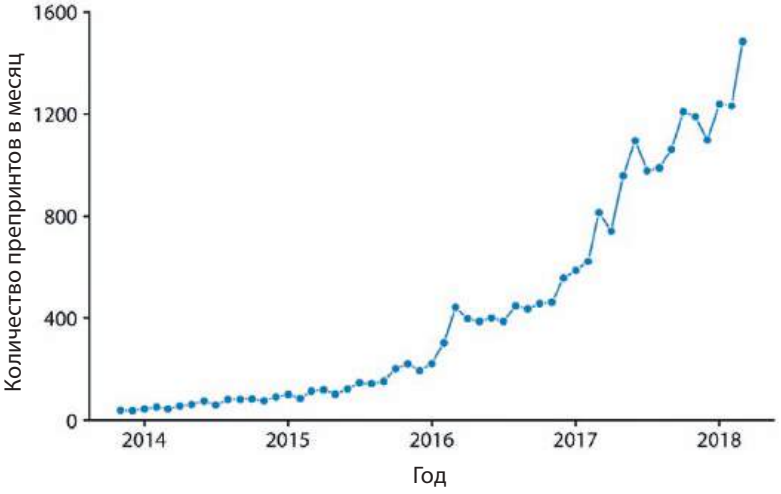


Рис. 12.2. Количество ежемесячных заявок, поданных на препринт-сервер bioRxiv, визуализированное в виде точек, соединенных друг с другом линиями. Сами линии не отражают какие-либо данные, я добавил их просто для удобства восприятия. Соединение точек означает, что между ними существует определенный порядок: у каждой точки есть одна последующая и одна предыдущая точка. Источник: Jordan Anaya, www.prepubmed.org

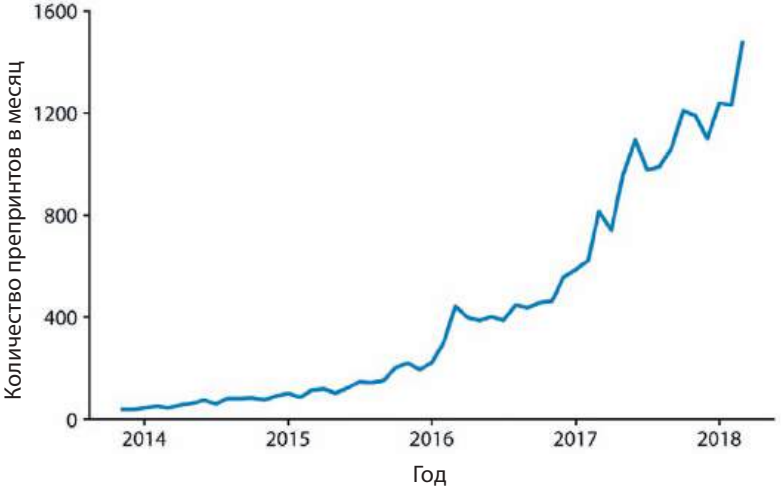


Рис. 12.3. Количество ежемесячных заявок, поданных на препринт-сервер bioRxiv, визуализированное в виде линейного графика без точек. Отсутствие точек подчеркивает общую временную тенденцию, позволяя «увидеть лес за деревьями». Такой прием особенно полезен в тех случаях, когда временные значения расположены очень плотно. Источник: Jordan Anaya, www.prepubmed.org

Кроме того, отсутствие точек делает график менее визуально перегруженным. Вообще говоря, чем плотнее временной ряд, тем менее важным является отображение отдельных наблюдений. Я считаю, что для приведенного здесь набора данных о препринтах отсутствие точек вполне приемлемо.

Также мы можем закрасить сплошным цветом область под кривой (рис. 12.4). Это еще больше подчеркнет общую тенденцию в данных, поскольку визуально отделяет область выше кривой от области ниже. Важно отметить, что подобный способ визуализации подходит только для тех графиков, на которых ось *y* начинается с нуля — в этом случае высота закрашенной области в каждой точке времени представляет собой значение данных в этот момент времени.

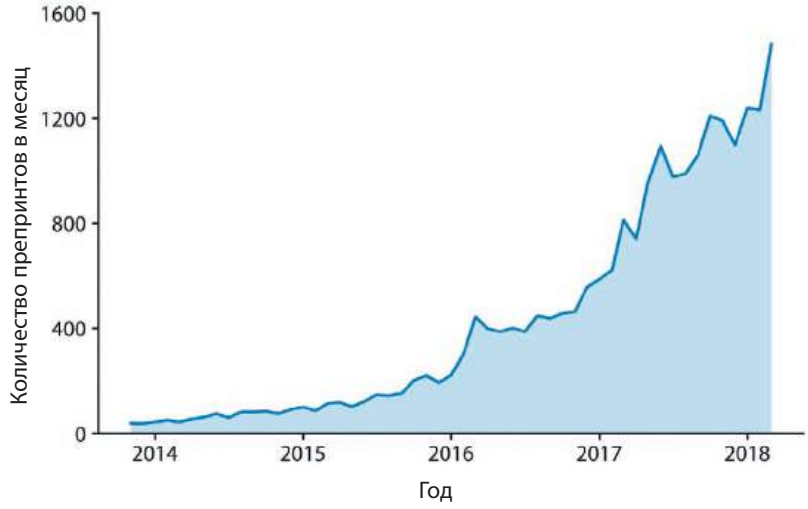


Рис. 12.4. Количество ежемесячных заявок, поданных на препринт-сервер bioRxiv, визуализированное в виде линейного графика с закрашенным внизу пространством (так называемой диаграммы с областями). Закрашенная под кривой область сильнее акцентирует внимание на тенденции изменения данных во времени, чем простая линия на рис. 12.3. Источник: Jordan Anaya, www.prepubmed.org

Множественные временные ряды и кривые «доза — эффект»

На практике часто бывает так, что на одном графике нужно показать несколько наборов данных во времени. В таких случаях следует быть особенно аккуратными при построении графика, чтобы рисунок не получился запутанным или трудночитаемым. К примеру, если мы хотим показать количество заявок, поданных на несколько препринт-серверов, использование диаграммы рассеяния будет не самой лучшей идеей, так как временные промежутки каждого

сервера будут пересекаться (рис. 12.5). Решить эту проблему можно, соединив точки линиями (рис. 12.6).

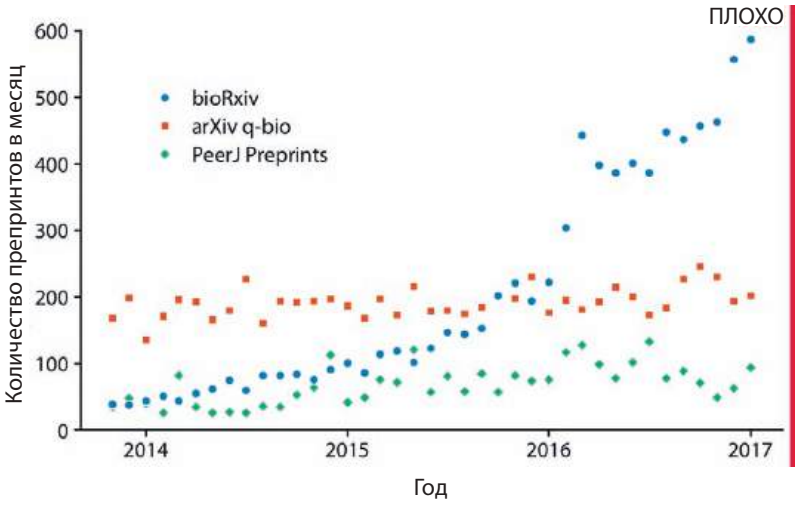


Рис. 12.5. Количество ежемесячных заявок, поданных на три препринт-сервера, созданных для научных сотрудников, занятых биомедицинскими исследованиями: bioRxiv, сектор q-bio сервера arXiv и PeerJ Preprints. Каждая точка представляет собой количество заявок, отправленных в течение одного месяца на соответствующий препринт-сервер. Данный рисунок относится к категории «плохих», потому что визуализации трех временных промежутков на одном графике мешают друг другу и затрудняют восприятие информации. Источник: Jordan Anaya, www.prepubmed.org

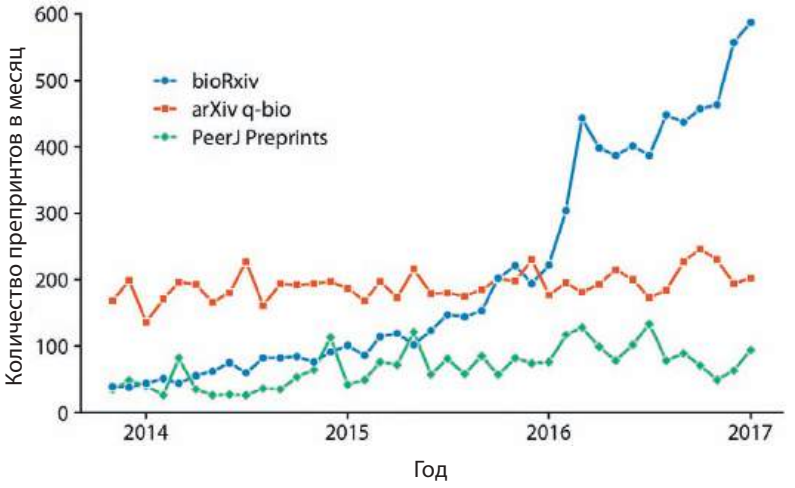


Рис. 12.6. Количество ежемесячных заявок, поданных на три препринт-сервера, созданных для научных сотрудников, занятых биомедицинскими исследованиями. Мы соединили точки на рис. 12.5, благодаря чему стало проще понять картину для каждого из серверов. Источник: Jordan Anaya, www.prepubmed.org

График на рис. 12.6 является приемлемой визуализацией набора данных о препринт-серверах. Однако легенда, расположенная отдельно от графика, излишне «нагружает» читателя. Чтобы избавиться от этого эффекта, мы можем разместить описания линий прямо рядом с ними (рис. 12.7). Чтобы результат выглядел более упорядоченным и удобным для чтения, я убрал с рисунка отдельные точки, благодаря чему график теперь смотрится более простым и понятным, нежели исходная диаграмма на рис. 12.5.

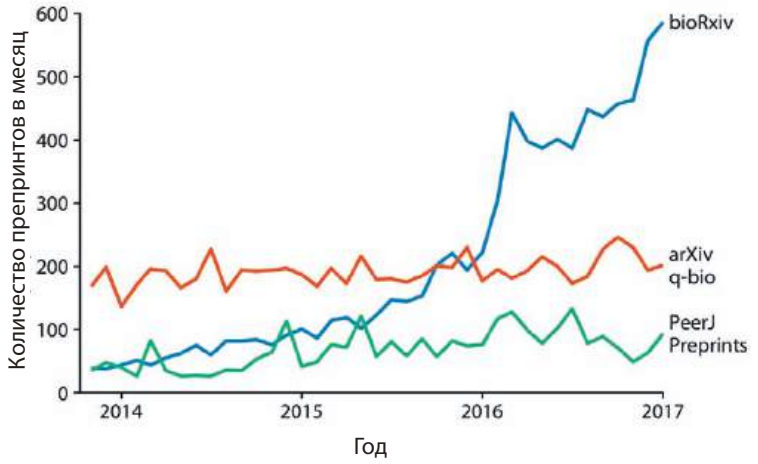


Рис. 12.7. Количество ежемесячных заявок, поданных на три препринт-сервера, созданных для научных сотрудников, занятых биомедицинскими исследованиями. Использование маркировки линий вместо отдельной легенды снижает уровень когнитивной нагрузки. Кроме того, теперь нет необходимости в использовании точек различной формы, благодаря чему можно полностью отказаться от точек на графике, чтобы оптимизировать рис. 12.6. Источник: Jordan Anaya, www.prepubmed.org

Использование линейных графиков не ограничивается одними временными рядами. Диаграммы этого типа можно использовать во всех тех случаях, когда точки данных располагаются в естественном порядке, который задается переменной, расположенной вдоль оси x таким образом, чтобы соседние точки можно было соединить друг с другом линией. Одним из примеров такого графика является кривая «доза — эффект», которая показывает, как изменение какого-либо численного параметра в эксперименте (доза) влияет на интересующий нас результат (эффект). На рис. 12.8 показан классический эксперимент такого рода, измеряющий урожайность овса в зависимости от количества вносимых удобрений. Данный график показывает, что кривые «доза — эффект» имеют одинаковую форму для трех рассматриваемых сортов овса, но различаются начальной точкой, которая означает полное отсутствие удобрения (то есть одни сорта имеют более высокую естественную урожайность, чем другие).

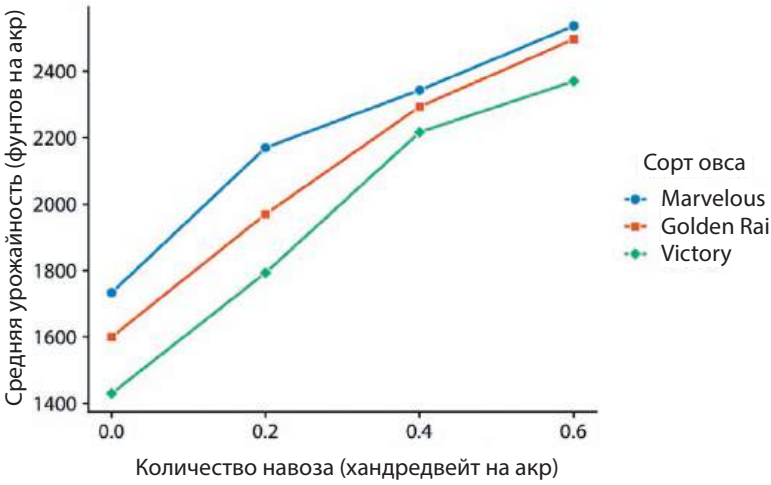


Рис. 12.8. Кривая «доза — эффект», показывающая среднюю урожайность сортов овса после внесения удобрений, которые содержат навоз (источник азота). Урожайность овса, как правило, увеличивается по мере поступления большего количества азота независимо от сорта овса. На графике внесение навоза измеряется в хандредвейтах на акр. Хандредвейт — это старая английская единица измерения, равная 112 фунтам, или 50,8 кг. Источник: [Yates, 1935]

Временной ряд двух или более объясняемых переменных

В предыдущих примерах мы рассматривали временные ряды только одной переменной (например, количество препринтов в месяц или урожайность овса). Однако на практике нередко встречаются ситуации, когда нужно учитывать несколько объясняемых переменных сразу. Чаще всего такие сценарии встречаются в макроэкономике. Например, нас может интересовать изменение цен на жилье по сравнению с предыдущим годом в зависимости от уровня безработицы. Логично предположить, что цены на жилье будут расти, когда уровень безработицы низкий, и наоборот.

С помощью инструментов, описанных в предыдущих главах, мы можем визуализировать данные такого типа в виде двух отдельных линейных графиков, расположенных друг над другом (рис. 12.9). На данном графике показаны обе интересующие нас переменные, и его легко интерпретировать. Но поскольку переменные представлены в виде отдельных линейных графиков, сравнить их между собой непросто. Например, если нам понадобится узнать, в какой период времени обе переменные движутся в одном и том же направлении или, наоборот, в противоположных направлениях, нам придется постоянно переключаться между графиками, чтобы сравнить наклоны обеих кривых в точках.

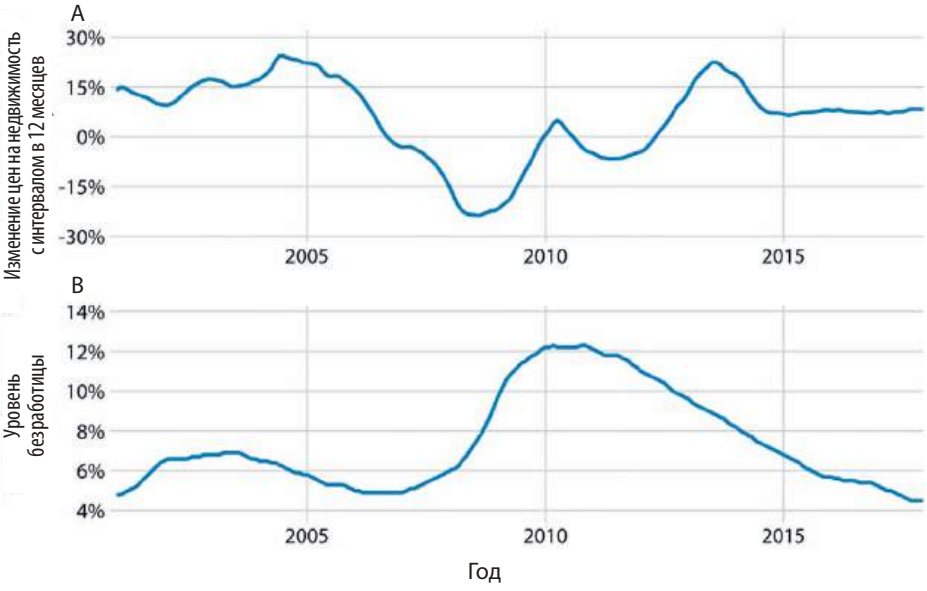


Рис. 12.9. 12-месячное изменение цен на недвижимость (A) и уровень безработицы (B) на временном периоде с января 2001-го по декабрь 2017 года. Источники: Freddie Mac House Prices Index, US Bureau of Labor Statistics

В качестве альтернативы изображению двух отдельных линейных графиков мы можем разместить две переменные друг напротив друга, нарисовав путь, ведущий от самой ранней по времени точки до самой поздней (рис. 12.10). Такая визуализация называется соединенной *диаграммой рассеяния*, потому что мы строим диаграмму рассеяния двух переменных относительно друг друга, а затем соединяем последовательные точки. Физики и инженеры часто называют такой график фазовой траекторией, потому что он обычно используется для представления движения состояния системы в фазовом пространстве. В главе 2 мы уже сталкивались со связанными диаграммами рассеяния, когда сравнивали суточные нормы температуры в Хьюстоне, штат Техас, с таковыми в Сан-Диего, штат Калифорния (см. рис. 2.3).

На приведенной фазовой траектории линии, идущие из нижнего левого угла в верхний правый, представляют собой прямо пропорциональное движение двух переменных (то есть когда растет одна переменная, растет и другая). Линии, идущие в перпендикулярном направлении, из левого верхнего угла в правый нижний, представляют обратно пропорциональное движение (когда значение одной переменной растет, второй — уменьшается). Циклическая связь между переменными выглядит на графике как круги или спирали. На рис. 12.10 показан один небольшой цикл на периоде с 2001 по 2005 год и один большой для остального временного промежутка. При построении

соединенной диаграммы рассеяния важно указывать направление и временной промежуток. Без этих пояснений график рискует превратиться в бессмысленные каракули (как на рис. 12.11). На рис. 12.10 направление обозначено через постепенное затемнение цвета. Альтернативным решением могут быть стрелки, нарисованные вдоль траектории.

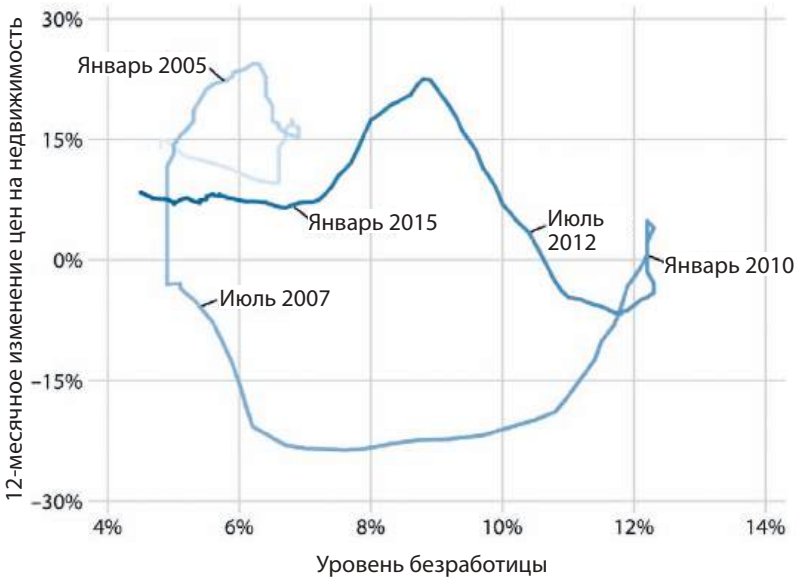


Рис. 12.10. Фазовая траектория визуализирует 12-месячное изменение цен на недвижимость в зависимости от уровня безработицы. Данный график отражает период с января 2001 года по декабрь 2017 года. Более темный оттенок означает более поздние месяцы. Обратная зависимость между ценами на недвижимость и уровнем безработицы, показанная на рис. 12.9, выглядит на графике как два круга, «идущих» против часовой стрелки. Оригинальная концепция графика: Len Kiefer. Источники: Freddie Mac House Price Index, US Bureau of Labor Statistics

Итак, какой вариант более предпочтителен: соединенная диаграмма рассеяния или два отдельных линейных графика? Отдельные графики, как правило, легче читаются, но как только люди привыкают к соединенным диаграммам рассеяния, они начинают замечать определенные закономерности (например, циклическое поведение с некоторыми отклонениями), которые может быть трудно обнаружить на линейных графиках. Действительно, заметить циклическую связь между изменением цен на жилье и уровнем безработицы на рис. 12.9 непросто, тогда как спираль, направленная против часовой стрелки на рис. 12.10, как раз помогает увидеть эту закономерность. Исследования показывают, что читатели чаще путают порядок и направ-

ление в соединенных диаграммах рассеяния, чем в линейных диаграммах, и поэтому реже замечают корреляцию между переменными [Haroz, Kosara and Franconeri, 2016]. С другой стороны, соединенные диаграммы рассеяния побуждают людей более внимательно рассматривать графики, благодаря чему можно эффективнее вовлекать читателя в рассказываемую на графике «историю».

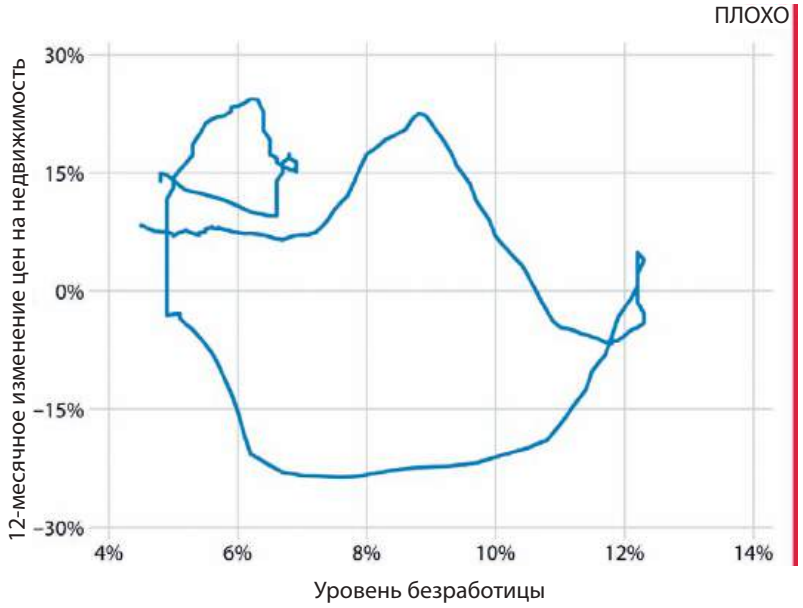


Рис. 12.11. 12-месячное изменение цен на жилье в сравнении с уровнем безработицы в период с января 2001-го по декабрь 2017 года. Данная визуализация относится к категории «плохих», так как на ней нет ни цветовых маркеров, ни временных меток, как на рис. 12.10, из-за чего невозможно понять ни направление линии, ни скорость изменения данных. Источники: Freddie Mac House Price Index, US Bureau of Labor Statistics

Несмотря на то что соединенные диаграммы рассеяния могут одновременно отображать только две переменные, графики такого рода все же можно использовать для визуализации многомерных наборов данных. Суть заключается в том, чтобы сначала применить методы снижения размерности (см. главу 11), а затем нарисовать фазовый портрет в пространстве с более низкой размерностью. В качестве примера такого подхода можно привести визуализацию базы данных ежемесячных наблюдений за более чем 100 макроэкономическими показателями, созданную Федеральным резервным банком Сент-Луиса. Давайте выполним анализ главных компонент для всех показателей, а затем нарисуем фазовый портрет ГК 2 по отношению к ГК 1 (рис. 12.12A) и по отношению к ГК 3 (рис. 12.12B).

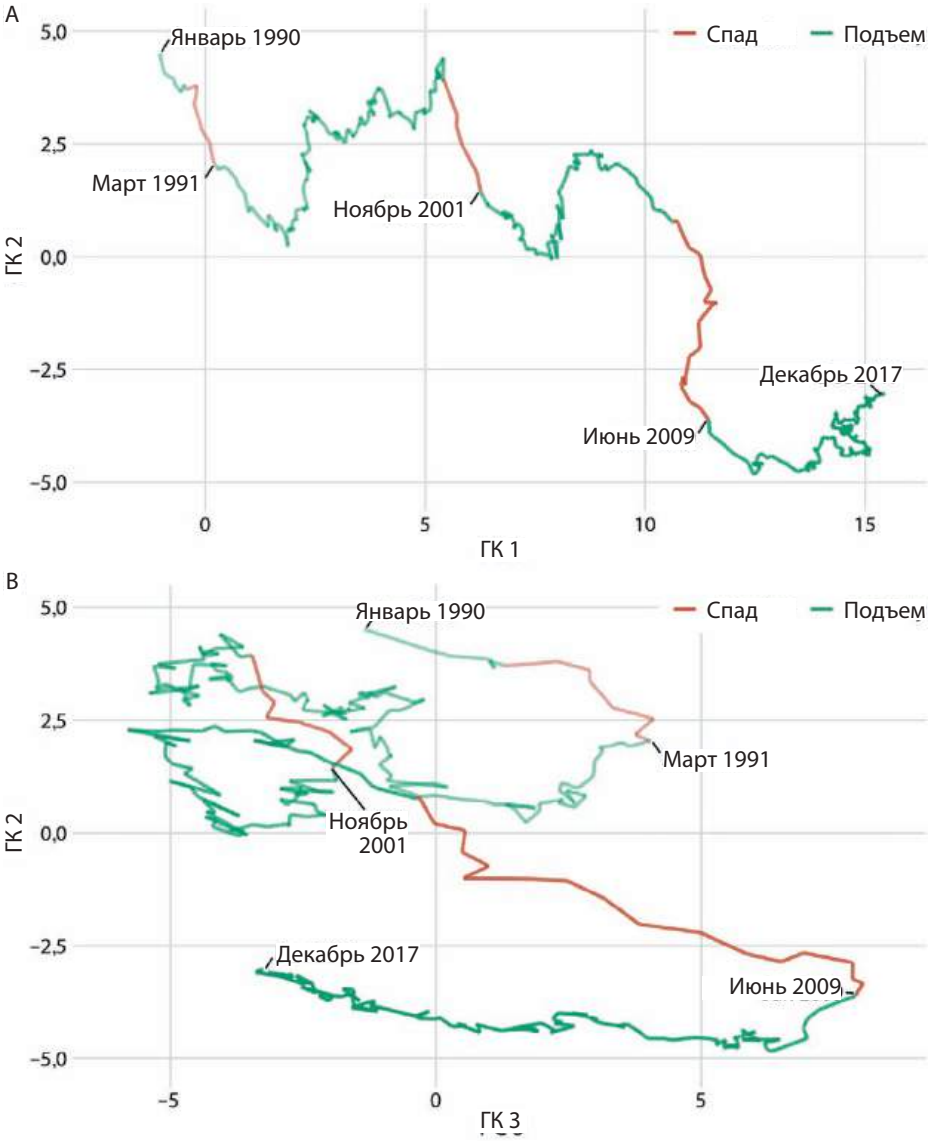


Рис. 12.12. Визуализация многомерного временного ряда в виде фазовой траектории в пространстве главных компонент. График отражает совместное движение более 100 макроэкономических показателей за время с января 1990 года по декабрь 2017 года. Цветом обозначены периоды спада и подъема экономики, а конечные точки трех рецессий подписаны (март 1991 года, ноябрь 2001 года и июнь 2009 года). A. ГК 2 по отношению к ГК 1. B. ГК 2 по отношению к ГК 3. Источник: M. W. McCracken, Федеральный резервный банк Сент-Луиса

Примечательно, что рис. 12.12A выглядит почти как обычный линейный график, где время «идет» слева направо. Такая закономерность обусловлена общей особенностью метода главных компонент: первая компонента часто измеряет общий «размер» системы. Здесь ГК 1 приблизительно измеряет общий размер экономики, который редко уменьшается с течением времени.

Раскрасив на соединенных диаграммах рассеяния времена спада и подъема экономики, мы увидим, что спады связаны с падением ГК 2, в то время как рост не соотносится с какими-либо свойствами ГК 1 и ГК 2 (см. рис. 12.12A). Скорее всего, рост связан с падением ГК 3 (см. рис. 12.12B). Более того, на графике, показывающем отношение ГК 2 к ГК 3, мы видим, что линия приобретает форму спирали, двигающейся по часовой стрелке. Спираль подчеркивает цикличный характер экономики, который представляет собой последовательную смену спадов подъемами, и наоборот.

Глава 13

Визуализация трендов

При построении диаграмм рассеяния (см. главу 11) или временных рядов (см. главу 12) нас зачастую больше интересует общая тенденция, нежели информация о том, где находится каждая отдельная точка данных. Нарисовав тенденцию поверх или вместо фактических точек данных (обычно в виде прямой или кривой линии), мы получим визуализацию, которая поможет читателю очень быстро понять ключевые особенности набора данных. Существует два фундаментальных подхода к поиску тенденций: можно либо сгладить данные при помощи какого-либо метода (например, скользящей средней), либо «подогнать» под значения данных кривую известной функциональной формы, а затем нанести ее на график. После нахождения тренда в наборе данных также следует обратить внимание на отклонения значений ряда от тренда или разложить сам ряд на компоненты, одной из которых будет та самая трендовая составляющая, а другие — представлять собой циклические компоненты, эпизодические включения или случайный шум.

Сглаживание

Рассмотрим временной ряд Dow Jones Industrial Average (сокращенно — просто Dow Jones или DJIA) — индекс фондового рынка, представляющий собой цену 30 наиболее крупных американских компаний, акции которых открыто торгуются на фондовой бирже. В частности, нас интересует 2009 год — сразу после кризиса 2008 года (рис. 13.1). Под конец спада, в период первых трех месяцев 2009 года, рынок потерял более 2400 пунктов (~27%), после чего шло медленное восстановление в течение оставшейся части года. Итак, как же нам визуализировать долгосрочные тенденции, не вдаваясь в подробности в виде менее существенных краткосрочных колебаний рынка?

Если бы мы задали этот вопрос специалисту по статистике, он бы ответил, что нам нужен метод *сглаживания* для данных фондового рынка. В результате сглаживания мы получим функцию, которая учитывает ключевые элементы данных и при этом отбрасывает незначительные детали или шумы. Финансовые аналитики обычно сглаживают данные фондового рынка с помощью *скользящих средних*. Для построения скользящей средней берется так

называемое «окно»: временной интервал фиксированной длины, например первые 20 дней временного ряда. После этого вычисляется средняя цена за эти 20 дней, далее окно перемещается на один день так, чтобы охватить период от 2 до 21 дня, и вычисляется среднее значение за эти 20 дней, затем окно снова перемещается и т. д. В результате получается новый временной ряд, состоящий из последовательности усредненных цен.

Рис. 13.1. Ежедневные значения индекса Dow Jones на момент закрытия торгов за 2009 год. Источник: Yahoo! Finance

Прежде чем строить последовательность скользящих средних, мы должны решить, какая именно точка во времени будет связана со средним значением для каждого временного окна. Финансовые аналитики часто располагают среднюю в конце соответствующего временного окна. Такое решение приводит к появлению кривых, которые отстают от исходных данных (рис. 13.2A), при этом бóльшее отставание соответствует бóльшей ширине временного окна. Ученые-статистики, наоборот, определяют среднее арифметическое по центру временного окна, в результате чего кривая хорошо накладывается на исходные данные (рис. 13.2B).

Вне зависимости от используемого типа сглаживания — с отставанием или без — можно заметить, что длительность промежутка времени, за который мы усредняем, задает масштаб колебаний, которые остаются видимыми на сглаженной кривой. Скользящая средняя за 20-дневный период устраняет только небольшие кратковременные скачки, а в остальном достаточно точно повторяет дневные значения. С другой стороны, скользящая средняя за 100-дневный период удаляет даже довольно значительные перепады или всплески, которые происходят в течение нескольких недель. Например, заметить значительное падение до уровня ниже 7000 пунктов в первом квартале

2009 года на такой скользящей средней нельзя, поскольку данное падение заменено пологой кривой, не опускающейся ниже 8000 пунктов (рис. 13.2). Аналогично, просадка в июле 2009 года полностью скрыта на 100-дневной скользящей средней.

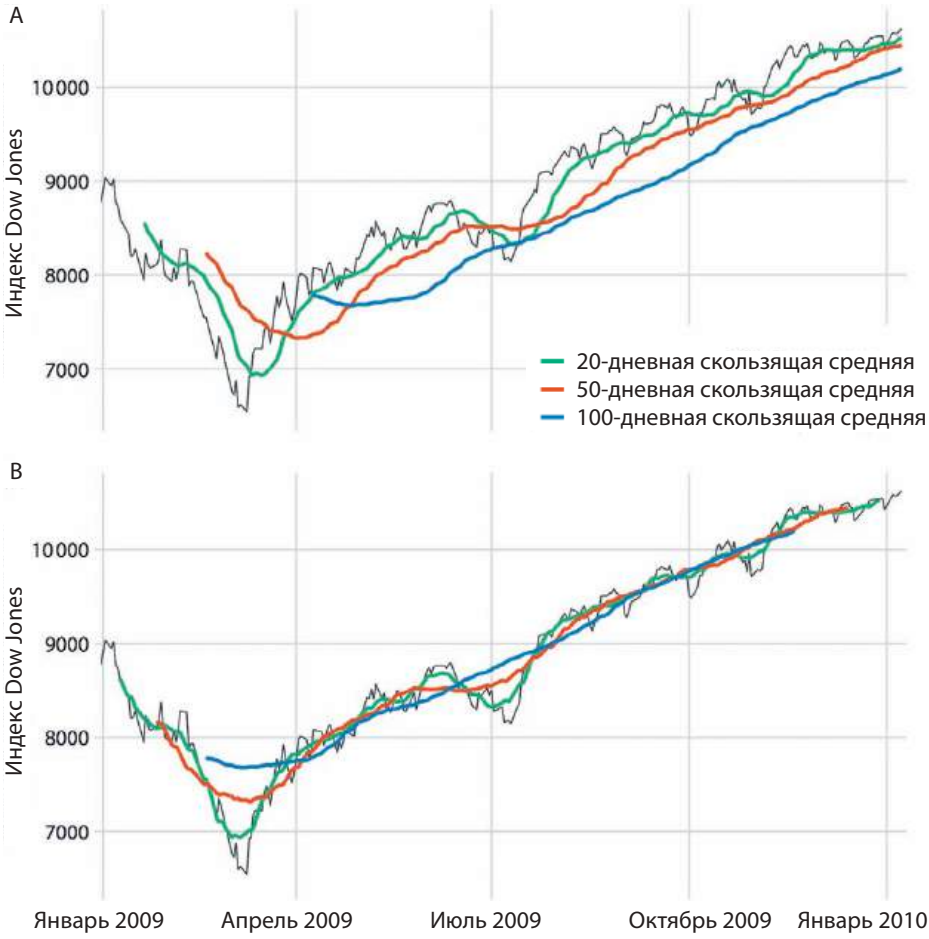


Рис. 13.2. Ежедневные значения индекса Dow Jones на момент закрытия торгов за 2009 год, показанные совместно с их скользящими средними за 20, 50 и 100 дней. A. Значения скользящих средних отображаются в конце соответствующих временных промежутков. B. Значения скользящих средних отображаются в центре соответствующих временных промежутков. Источник: Yahoo! Finance

Метод скользящей средней является наиболее простым способом сглаживания, и у него есть некоторые очевидные ограничения. Во-первых, кривая, получающаяся таким методом, значительно короче исходной (см. рис. 13.2). Детали могут отсутствовать где угодно — в начале или в конце, либо и там и там. И чем сильнее сглажен временной ряд (то есть чем шире временное

окно), тем короче будет итоговая кривая. Во-вторых, даже при использовании широкого окна усреднения скользящая средняя не всегда будет плавной линией. На ней могут быть небольшие кочки и завихрения даже при использовании длинного периода усреднения. Причиной этих завихрений являются отдельные точки данных, когда они попадают в окно усреднения или выходят из него. Поскольку все точки данных в окне имеют одинаковый вес, отдельные точки на границах окна могут сильно повлиять на значения средней.

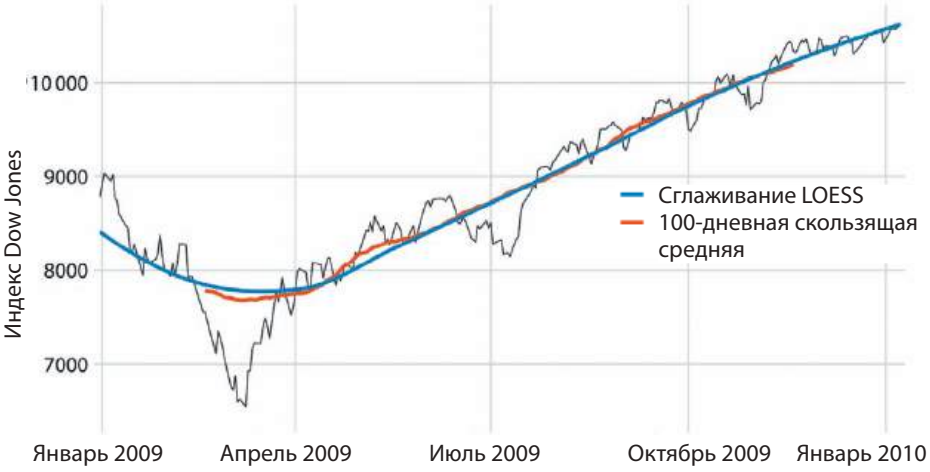


Рис. 13.3. Сравнение сглаживания LOESS и 100-дневной скользящей средней для графика индекса Dow Jones с рис. 13.2. Тенденция, которую демонстрирует кривая LOESS, практически идентична 100-дневной скользящей средней, при этом кривая сглаживания LOESS гораздо более плавная и охватывает весь диапазон данных. Источник: Yahoo! Finance

На сегодняшний день в статистике существует множество подходов к сглаживанию, которые лишены минусов, присущих скользящим средним. Разумеется, эти подходы более сложны в реализации и требуют бóльших вычислительных ресурсов, но современные статистические приложения и среды делают эти недостатки незаметными. Одним из широко используемых методов является *локально оцениваемое сглаживание диаграммы рассеяния* (LOESS) [Cleveland, 1979], которое сглаживает данные, разбивая временной ряд на отрезки, на каждом из которых данные аппроксимируются многочленом низкого порядка. Следует отметить, что точки в центре каждого отрезка имеют более высокий вес, чем точки на границах, поэтому данная схема взвешивания дает гораздо более гладкий результат, чем обычное взвешенное среднее. Кривая LOESS, показанная на рис. 13.3, похожа на график 100-дневной скользящей средней, приведенной на рис. 13.2, однако этому сходству не следует придавать большого значения. Гладкость LOESS-кривой можно настроить при помощи изменения значений параметров метода. Соответственно,

различные комбинации параметров позволяют создать кривые LOESS, похожие, например, на 20- или 50-дневные скользящие средние.

Что немаловажно, использование LOESS не ограничивается одними временными рядами: LOESS можно также применять к любым диаграммам рассеяния, как следует из названия метода — *локально оцениваемое сглаживание графика рассеяния*. К примеру, мы можем использовать LOESS для поиска тенденций во взаимосвязи между размером топливного бака автомобиля и его стоимостью (рис. 13.4). Линия LOESS показывает, что для сегмента дешевых автомобилей (стоимостью менее 20 000 долларов США) емкость бака увеличивается почти линейно, но стабилизируется для более дорогих машин. У автомобилей дороже 20 000 долларов США объем топливного бака не будет увеличиваться с ростом цены.

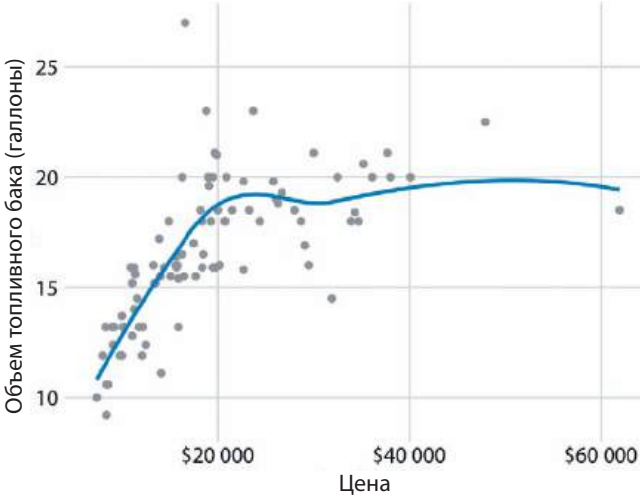


Рис. 13.4. Отношение объема топливного бака к цене автомобиля для 93 автомобилей модельного ряда 1993 года. Каждая точка соответствует одному автомобилю. Сплошная линия соответствует кривой сглаживания LOESS. Как можно видеть, размер топливного бака увеличивается прямо пропорционально цене автомобиля вплоть до показателя в 20 000 долларов США, после чего стабилизируется. Источник: Robin H. Lock, St. Lawrence University

LOESS является очень популярным методом сглаживания, поскольку результаты его работы выглядят «правильно». Однако использование этого метода требует подгонки множества отдельных регрессионных моделей. Из-за этого большие массивы данных будут обрабатываться долго, даже если в распоряжении исследователя находится мощное вычислительное оборудование.

Более производительной альтернативой LOESS являются сплайновые модели. *Сплайном* называется кусочно-полиномиальная функция, которая обладает высокой гибкостью и выглядит очень плавно. При работе со сплайнами

мы имеем дело с таким термином, как *узел*. Узлы в сплайне представляют собой конечные точки отдельных сегментов сплайна. Если мы создаем сплайн из k сегментов, нам понадобится указать $k + 1$ узел. Несмотря на то что создание сплайнов не требует большой вычислительной мощности (особенно если количество узлов невелико), у этого типа графиков есть свои особенности. В частности, существует огромное количество сплайнов различных типов, как то: кубические сплайны, B-сплайны, тонкие сплайны, сплайны гауссова процесса и многие другие. Отсюда следует, что выбор подходящего сплайна — дело крайне сложное. От выбора конкретного типа сплайна и количества используемых узлов сильно зависят результаты работы метода сглаживания, причем на одних и тех же данных (рис. 13.5).

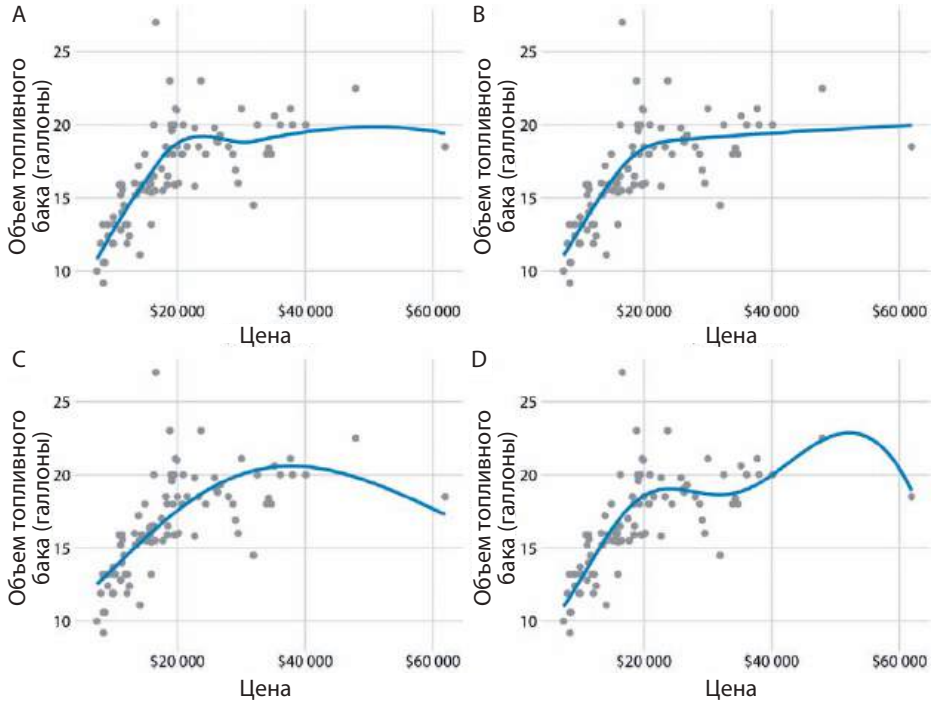


Рис. 13.5. Различные сглаживающие модели демонстрируют совершенно разное поведение, в особенности вблизи границ данных. A. Сглаживание LOESS, такое же как на рис. 13.4. B. Кубический сплайн с пятью узлами. C. Тонкий пластинчатый (thin-plate) сплайн с тремя узлами. D. Сплайн гауссова процесса с шестью узлами. Источник: Robin H. Lock, St. Lawrence University

Большинство программ, используемых для визуализации данных, предлагает сглаживание функций, реализованное либо в виде локальной регрессии (например, LOESS), либо в виде сплайна. Вообще, все методы сглаживания относятся к категории *обобщенной аддитивной модели* (GAM). Нельзя

забывать, что результаты сглаживания очень зависят от типа используемой модели, а также советую всегда пробовать как можно больше различных вариантов, иначе вы рискуете не увидеть, насколько сильно итоговый результат зависит от выбранной модели и настроек выбранного ПО.

Будьте осторожны при интерпретации результатов метода сглаживания. Один и тот же набор данных можно сгладить огромным количеством способов.

Подгонка трендов при помощи заданных функциональных форм

Рисунок 13.5 хорошо иллюстрирует тот факт, что для любого набора данных результаты работы общих методов сглаживания могут оказаться несколько непредсказуемыми. Кроме того, у этих методов нет выходных параметров, которые бы имели осмысленную интерпретацию. Поэтому, если ситуация располагает, выбирайте кривую с заданной функциональной формой, подходящей для конкретных данных, и использующую параметры с трактуемым смыслом.

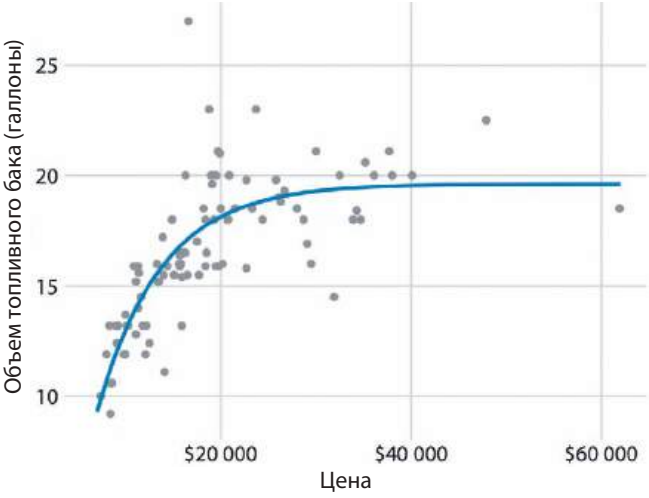


Рис. 13.6. Данные об объеме топливного бака, представленные в виде явной аналитической модели. Сплошная линия соответствует подобранной по методу наименьших квадратов кривой для функции $y = A - B \exp(-mx)$.
Подобранные значения параметров: $A = 19{,}6$; $B = 29{,}2$; $m = 0{,}00015$.
Источник: Robin H. Lock, St. Lawrence University

Рассмотрим следующий пример. Для визуализации набора данных о размерах топливного бака нам нужна кривая, которая сначала линейно

возрастает, а затем стабилизируется около постоянного значения. Данному условию удовлетворяет функция $y = A - B \exp(-mx)$. Здесь A, B и m — это константы, значения которых мы подгоняем, чтобы получившаяся в итоге кривая соответствовала данным. Функция близка к линейной для малых значений x: $y \approx A - B + Bmx$; для больших значений x функция близка к константе $y \approx A$. На всем диапазоне x функция строго возрастает. На рис. 13.6 видно, что эта функция аппроксимирует наш набор данных нисколько не хуже, чем ранее рассмотренные методы сглаживания (см. рис. 13.5).

Функциональной формой, применимой во многих различных контекстах, является простая линейная функция, $y = A + mx$. Близкие к линейным соотношения между двумя переменными удивительно часто встречаются в реальных наборах данных. Например, в главе 11 мы говорили о связи между длиной головы и массой тела голубых соек. Зависимость между этими переменными является практически линейной как для самцов, так и для самок, поэтому нанесение линий линейного тренда поверх точек на графике помогает читателю увидеть тенденции (рис. 13.7).

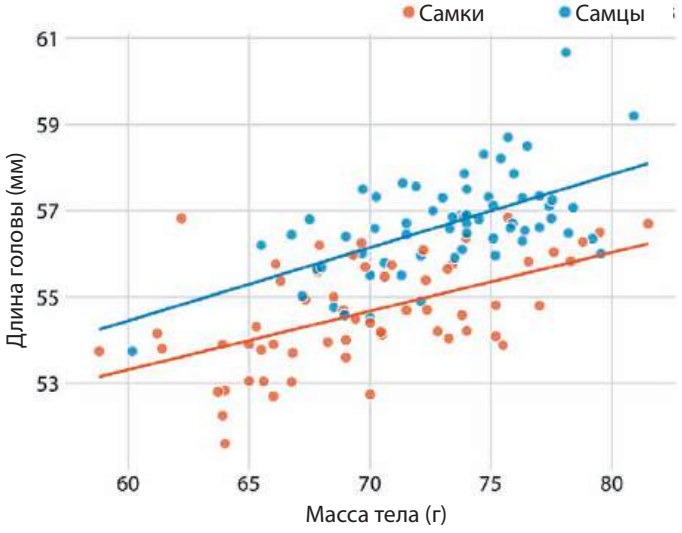


Рис. 13.7. Отношение длины головы к массе тела 123 голубых соек. Пол птиц выделен цветом. Данное изображение эквивалентно рис. 11.2 за тем исключением, что мы нарисовали прямые, показывающие линейные тренды, над индивидуальными значениями точек. Источник: Keith Tarvin, Oberlin College

В том же случае, когда в данных присутствует нелинейная зависимость, нередко приходится поломать голову над тем, какая функциональная форма будет наиболее подходящей. Чтобы оценить точность нашего предположения, мы можем преобразовать оси таким образом, чтобы появилась линейная зависимость. Для иллюстрации этого принципа давайте снова обратимся

к данным о количестве ежемесячных заявок, поданных на препринт-сервер bioRxiv, из главы 11. Если увеличение числа поданных заявок в каждом месяце пропорционально количеству поданных заявок за предыдущий месяц, то есть если количество заявок ежемесячно подрастает на фиксированный процент, результирующая кривая будет экспоненциальной. Подобное предположение, похоже, соответствует данным от bioRxiv, потому что кривая с экспоненциальной величиной $y = A \exp(mx)$ хорошо согласуется с данными об отправках на bioRxiv (рис. 13.8).

Рис. 13.8. Ежемесячное количество заявок, поданных на препринт-сервер bioRxiv. Сплошная синяя линия показывает фактическое количество заявок на препринт, а пунктирная черная — подогнанную экспоненциальную кривую вида $y = 60 \exp[0{,}77(x - 2014)]$. Источник: Jordan Anaya, www.prepubmed.org

Если исходная кривая является экспонентой вида $y = A \exp(mx)$, то логарифмическое преобразование значений оси y превратит ее в линейное соотношение, $\log(y) = \log(A) + mx$. Таким образом, построение графика данных с логарифмически преобразованными значениями y (или, что то же самое, с логарифмической осью y) и поиск линейной зависимости — это хороший способ подтвердить или опровергнуть наличие в наборе данных экспоненциального роста. Как можно заметить, при использовании логарифмической оси y для количества заявок, поданных на препринт-сервер bioRxiv, у нас действительно налицо линейная зависимость (рис. 13.9).

На рис. 13.9, в дополнение к фактическому количеству поданных заявок, приведено экспоненциальное приближение с рис. 13.8, а также линейное приближение преобразованных по логарифму данных.

Между этими вариантами есть сходство, однако они не идентичны. В частности, наклон пунктирной линии немного не соответствует графику:

линия находится выше отдельных точек данных на протяжении половины временного ряда.

Рис. 13.9. Ежемесячное количество заявок, поданных на препринт-сервер bioRxiv, показанное на логарифмической шкале. Сплошная синяя линия представляет собой фактическое количество препринтов, подаваемых ежемесячно, пунктирная черная линия — экспоненциальная зависимость с рис. 13.8, а сплошная черная линия представляет собой линейную зависимость, преобразованную в логарифмическую шкалу данных и соответствующую $y = 43 \exp[0{,}88(x - 2014)]$. Источник: Jordan Anaya, www.prepubmed.org

Такую картину можно часто увидеть при использовании экспоненциального приближения: квадраты отклонений значений данных от подогнанной кривой для больших значений данных настолько больше, чем для малых, что вклад отклонений на малых значениях практически ничтожен при вычислении суммы для минимизации среднеквадратического отклонения. Полученная в результате линия систематически недотягивает до или, наоборот, превышает значения данных в области, где они малы. По этой причине я советую избегать использования экспоненциального приближения и применять вместо него линейное приближение данных, преобразованных по логарифму.

Как правило, аппроксимация преобразованных данных линейной функцией дает лучшие результаты, чем аппроксимация исходных данных нелинейными функциями.

Графики, подобные тому, что показан на рис. 13.9, обычно называют *логлинейными*, поскольку ось y является логарифмической, а ось x — линейной. На практике могут встречаться и другие виды графиков, например со шкалой

вида «логарифм-логарифм», где оси y и x являются логарифмическими, или шкалой вида «линейный-логарифм», где ось y — линейная, а x — логарифмическая. На графиках типа «логарифм-логарифм» степенной закон формы $y \sim x^{\alpha}$ представлен в виде прямой линии (см. пример на рис. 7.7). На графиках типа «линейный-логарифм» логарифмические соотношения $y \sim \log(x)$ тоже отображаются в виде прямых линий. Другие функциональные формы также могут быть преобразованы в линейные соотношения при помощи более специфических преобразований координат, но при этом все приведенные выше (логлинейные, «логарифм-логарифм», «линейный-логарифм») подходят для очень широкого спектра практических сценариев.

Удаление трендов и декомпозиция временных рядов

Для любых временных рядов с выраженной долгосрочной тенденцией бывает полезно исключить информацию о тренде и сосредоточиться на отклонениях от него. Эта техника называется «детрендированием» (удалением тренда). Давайте посмотрим, как она работает, и в качестве подопытного воспользуемся набором данных о ценах на недвижимость. В США ипотечный кредитор Freddie Mac ежемесячно публикует тематическую сводку под названием Freddie Mac House Price Index, которая отражает изменение цен на жилье с течением времени. Индекс пытается как можно более полно отразить состояние рынка недвижимости в регионах, так, если в сводке сообщается о росте индекса в регионе на 10%, это означает, что средняя цена на жилье в данном регионе выросла на 10%. Значение индекса для декабря 2000 года было принято за 100.

В течение длительных периодов времени цены на жилье, как правило, демонстрируют устойчивый ежегодный рост, приблизительно равный показателю инфляции. Однако поверх этой тенденции часто возникают различного рода «пузыри», которые приводят к серьезному циклу подъемов и спадов. На рис. 13.10 показаны фактические значения цен на жилье, а также их долгосрочная тенденция для четырех отдельных штатов США. Мы видим, что в промежутке между 1980 и 2017 годами Калифорния пережила два «пузыря», один в 1990 году, а другой — в середине 2000-х. За тот же период в Неваде был только один «пузырь» — в середине 2000-х годов, а цены на жилье в Техасе и Западной Виргинии все это время почти совпадали со своими долгосрочными прогнозами. Поскольку цены на жилую недвижимость имеют тенденцию расти на проценты, то есть экспоненциально, на рис. 13.10 я использую логарифмическую ось y. Прямые линии соответствуют ежегодному повышению цен на 4,7% в Калифорнии и на 2,8% в Неваде, Техасе и Западной Виргинии.

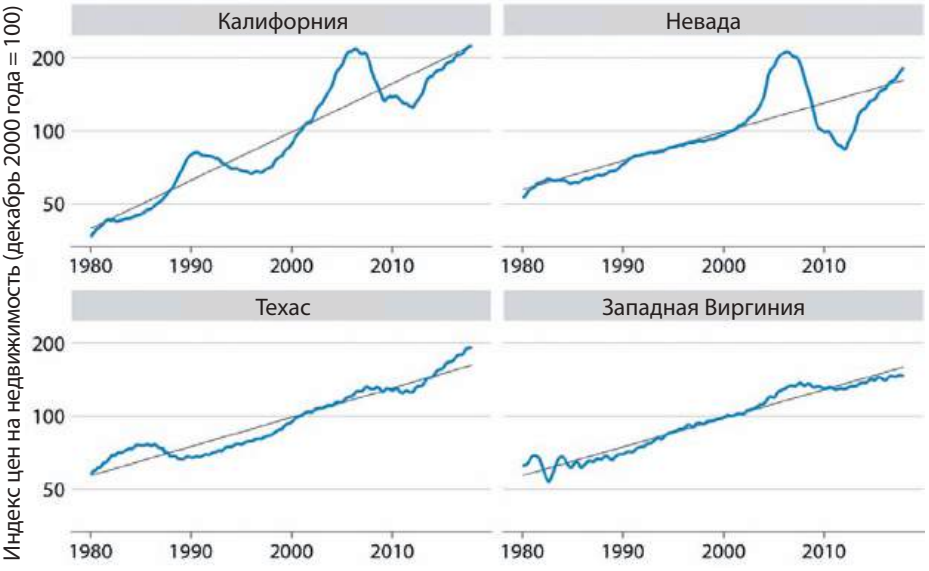


Рис. 13.10. Показатели Freddie Mac House Price Index в период с 1980 по 2017 год для четырех штатов (Калифорния, Невада, Техас и Западная Виргиния). Индекс цен на недвижимость — это число, не имеющее единиц измерения, но позволяющее отслеживать относительные цены на жилье в выбранном географическом регионе в течение определенного периода времени. Индекс отмасштабирован таким образом, чтобы значения для декабря 2000 года были равными 100 (для всех регионов). Синие линии показывают месячные колебания индекса, серые линии — долгосрочные тенденции роста цен в соответствующих штатах. Обратите внимание, что оси *y* являются логарифмическими, поэтому прямые серые линии означают стабильный экспоненциальный рост. Источник: Freddie Mac House Prices Index

Устранение тренда для индекса цен на недвижимость производится путем деления фактического индекса цен в каждый момент времени на соответствующее значение в долгосрочном тренде. Визуально это выглядит так, будто мы вычитаем серые линии из синих с рис. 13.10, потому что деление не преобразованных по логарифму значений эквивалентно вычитанию для преобразованных. После того как с графиков будет устранен тренд, мы сможем ясно увидеть «пузыри» (рис. 13.11), поскольку метод как раз и предназначен для выделения неожиданных изменений временного ряда. Например, в исходном временном ряду снижение цен на жилье в Калифорнии в период с 1990 по 1998 год выглядит скромно (рис. 13.10). Однако, если руководствоваться долгосрочным трендом, мы должны были бы увидеть на графике рост. Если сравнивать фактические показатели с ожидаемыми, мы получим существенное падение цен, которое равно 25% в нижней точке (см. рис. 13.11).

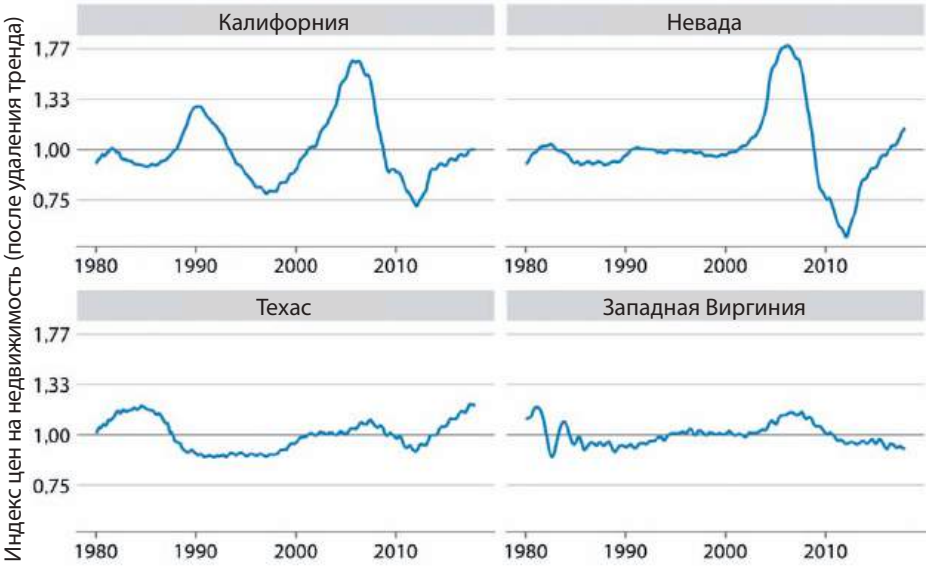


Рис. 13.11. Версия Freddie Mac House Price Index, представленного на рис. 13.10, после удаления тренда. Это сделано путем деления фактического индекса (синие линии с рис. 13.10) на ожидаемые значения, рассчитанные на основе долгосрочного тренда (прямые серые линии с рис. 13.10). Эта визуализация показывает, что Калифорния испытала на себе два «пузыря» цен: около 1990 года и в середине 2000-х. «Пузыри» проявляют себя в виде быстрого роста и последующего снижения фактических цен на жилье по сравнению с тем, что ожидалось в долгосрочной перспективе. Аналогичным образом можно увидеть, что такой же «пузырь» был в Неваде в середине 2000-х. Что касается Техаса и Западной Виргинии, то в этих штатах «пузырей» практически не было

Помимо удаления тренда мы также можем разделить временной ряд на компоненты, которые в сумме будут равны исходному временному ряду. В дополнение к долгосрочному тренду существует три отдельные компоненты, из которых может состоять временной ряд. Во-первых, это случайный шум, который вызывает небольшие и нерегулярные движения вверх и вниз. Этот шум можно заметить во всех временных рядах, приведенных в этой главе, но, возможно, лучше всего его видно на рис. 13.9. Во-вторых, на практике случаются уникальные внешние события, которые оставляют после себя следы во временных рядах, такие как, например, видимые «пузыри» на рис. 13.10. В-третьих, изменения могут иметь циклический характер. Так, например, график уличной температуры представляет собой набор ежедневных циклических изменений. Самые высокие температуры наблюдаются в середине дня, а самые низкие — ранним утром. Температуры, кстати, кроме ежедневного цикла, обладают также годичной цикличностью: весной, как правило, температура растет, летом она достигает

максимума, осенью начинается снижение, а зимой температура достигает минимума (см. рис. 2.2).

Чтобы продемонстрировать концепцию отдельных компонент временных рядов, давайте разложим график Килинга, который показывает изменение содержания углекислого газа в атмосфере с течением времени (рис. 13.12). С 1958 года в обсерватории Мауна-Лоа, расположенной на Гавайях, ведется постоянный мониторинг содержания CO_2 в атмосфере, начатый под руководством Чарльза Килинга.

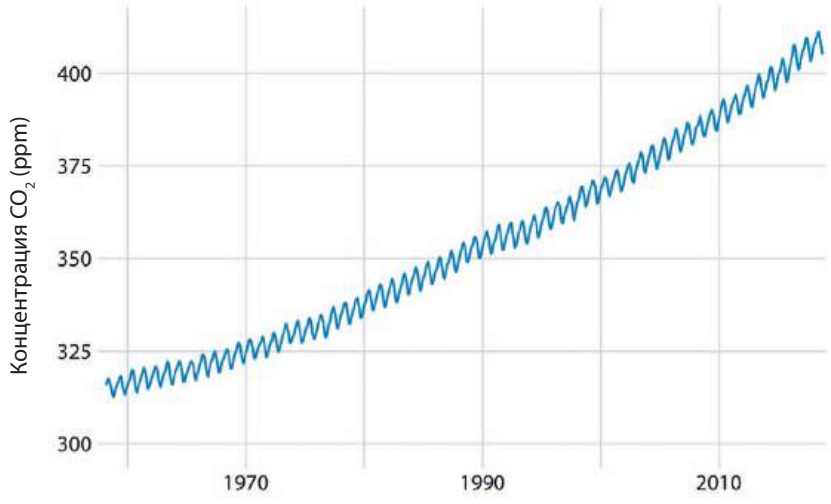


Рис. 13.12. График Килинга. Кривая Килинга показывает изменение концентрации CO_2 в атмосфере с течением времени. На графике приведены среднемесячные значения содержания CO_2, выраженные в миллионных долях (ppm). Показания CO_2 колеблются ежегодно в зависимости от сезона, демонстрируя рост в долгосрочной тенденции. Источник: Dr. Pieter Tans, NOAA/ESRL, и Dr. Ralph Keeling, Scripps Institution of Oceanography

Концентрация CO_2 измеряется в миллионных долях (ppm). На графике мы видим долгосрочное увеличение объема выбросов CO_2, скорость которого чуть выше линейной: с менее чем 325 ppm в 1960-е годы до 400 ppm во втором десятилетии XXI века (см. рис. 13.12). Кроме того, ежегодный объем выбросов CO_2 колеблется «вверх-вниз» относительно тенденции роста.

Годовые колебания обусловлены ростом растений в Северном полушарии. Растения потребляют CO_2 во время фотосинтеза. Поскольку бóльшая часть суши Земли расположена в Северном полушарии, а рост растений наиболее активно идет весной и летом, мы видим ежегодное глобальное снижение атмосферной концентрации CO_2, которое совпадает с летними месяцами в Северном полушарии.

Кривую Килинга можно разложить на долгосрочный тренд, сезонные колебания и остаток (рис. 13.13). Тот метод, которым я буду пользоваться, называется разложением временного ряда с выделением сезонной компоненты при помощи LOESS (seasonal decomposition of time series by LOESS, или STL) [Cleveland et al., 1990]. Стоит отметить, что существует множество других методов, с помощью которых можно достичь аналогичного результата.

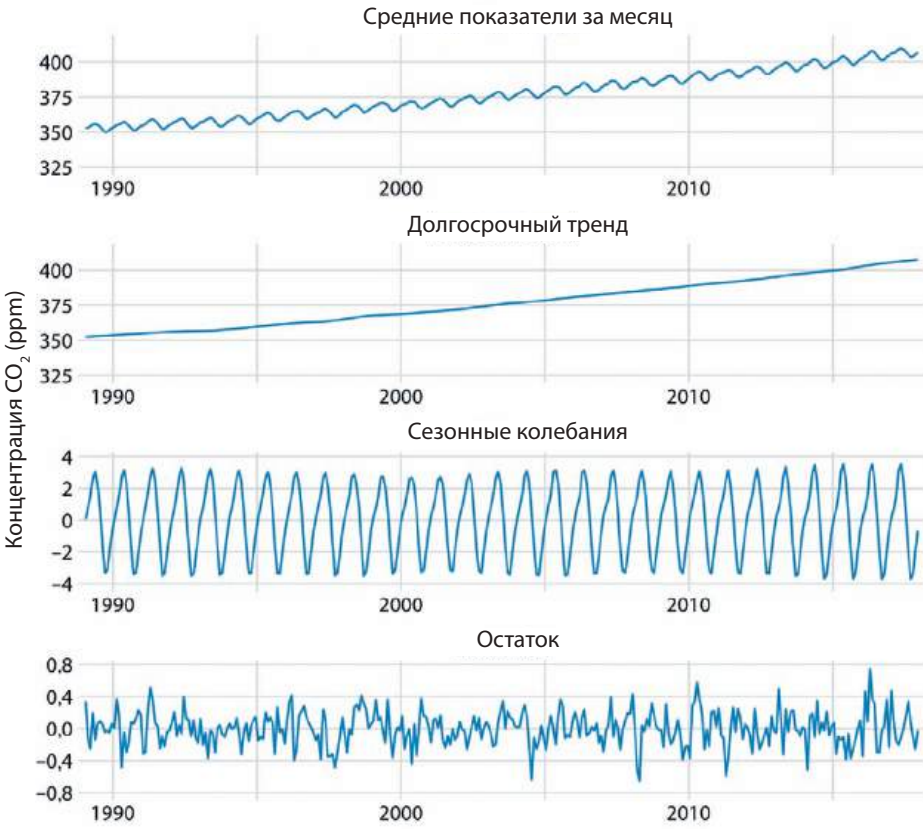


Рис. 13.13. Разложение временного ряда графика Килинга на отдельные компоненты. На графике показаны исходные данные о средних показателях (как на рис. 13.12), долгосрочный тренд, сезонные колебания и остаток. Остаток представляет собой разницу между фактическими показаниями и суммой долгосрочного тренда и сезонных колебаний и является случайным шумом. Я сделал акцент на данных за последние 30 лет, чтобы подчеркнуть форму ежегодных колебаний. Источник: Dr. Pieter Tans, NOAA/ESRL, и Dr. Ralph Keeling, Scripps Institution of Oceanography

Благодаря разложению временного ряда мы можем увидеть, что за последние три десятилетия содержание CO_2 возросло более чем на 50 ppm. Сравните этот показатель с сезонными колебаниями, которые составляют

менее 8 ppm (они никогда не отклоняются от долгосрочного тренда более чем на 4 ppm в любую из сторон), а колебания остатка, в свою очередь, составляют менее 1,6 ppm (см. рис. 13.13). Как уже было сказано ранее, остаток — это разница между фактическими показаниями и суммой долгосрочного тренда и сезонных колебаний, и в ежемесячных показаниях CO_2 он соответствует случайному шуму. Вообще говоря, остаток может служить также и признаком уникальных внешних событий. Например, если из-за мощного извержения вулкана в атмосферу попало значительное количество CO_2, на графике остатка такое событие может выглядеть как внезапный всплеск. Как следует из рис. 13.13, однако, ни одно из таких уникальных внешних событий не оказало существенного влияния на график Килинга за последние десятилетия.

Глава 14

Визуализация геопространственных данных

Многие наборы данных содержат информацию, связанную с местоположением в физическом мире. Например, в рамках экологического исследования набор данных может содержать перечень конкретных видов растений или животных с указанием мест, где они были найдены. Аналогично в социально-экономическом или политическом контексте набор данных может включать в себя информацию о том, где живут люди с определенными характеристиками (например, доход, возраст или уровень образования). Кроме того, такой набор данных может содержать информацию о том, где находятся объекты, созданные человеком (например, мосты, дороги, здания). Во всех этих случаях бывает полезно визуализировать данные в соответствующем геопространственном контексте: например, показать их на реалистичной карте или изобразить в виде диаграммы, похожей на карту.

Как правило, читатели хорошо воспринимают карты, однако их проектирование может оказаться непростой задачей. В частности, необходимо хорошо продумать такой параметр, как картографические проекции, а также решить, что более важно в нашем конкретном случае — точное отображение углов или площадей. Наиболее распространенной техникой создания карт является хороплет (или фоновая картограмма), суть которой заключается в представлении значений данных в виде областей разного цвета. В зависимости от ситуации хороплеты могут оказаться как удачным выбором визуализации, так и сбивающим читателя с толку. Альтернативой являются диаграммы, внешне напоминающие карты. Они называются *картограммами* и позволяют умышленно искажать области на карте или же представлять их в стилизованной форме, например в виде квадратов одинакового размера.

Проекции

Геометрически Земля больше всего похожа на сферу (рис. 14.1), а точнее на геоид, слегка сплющенный вдоль своей оси вращения. Две точки, где ось вращения пересекается со сферой, называются полюсами (различают Северный и Южный). Наш геоид традиционно делится на два полушария, Северное

и Южное, с помощью линии (она называется экватором), равноудаленной от обоих полюсов. Для однозначного задания местоположения чего-либо на Земле нам нужно знать три параметра: где мы находимся вдоль линии экватора (долгота), насколько близко мы находимся к любому полюсу при движении перпендикулярно экватору (широта) и как далеко мы находимся от центра Земли (высота). Долгота, широта и высота задаются относительно системы координат, называемой *датум*. Датум определяет такие свойства, как форма и размер Земли, а также расположение нулевой долготы, широты и высоты. Одним из наиболее широко распространенных датумов является World Geodetic System 1984 (WGS 84), которую использует Глобальная система позиционирования (GPS).

Рис. 14.1. Ортографическая проекция мира, показывающая Европу и Северную Африку в том виде, в котором они видны из космоса. Линии, исходящие из Северного полюса и идущие на юг, называются меридианами, а линии, перпендикулярные меридианам, называются параллелями. Все меридианы имеют одинаковую длину, но чем ближе мы подходим к любому из полюсов, тем короче становятся параллели

Несмотря на то что высота используется во многих геопространственных приложениях и является важной характеристикой, при визуализации данных нам больше важны два других измерения — долгота и широта. Оба эти

параметра представляют собой углы, выраженные в градусах. Долгота показывает, насколько далеко на востоке или западе находится точка. Линии, расположенные через равные промежутки долготы, называются меридианами, и все меридианы упираются в полюса (см. рис. 14.1). Основной меридиан, соответствующий 0° долготы, проходит через деревню Гринвич, расположенную в Великобритании. Меридиан, противоположный основному меридиану, находится в 180° от него (также обозначается как 180° восточной долготы), что эквивалентно –180° долготы (также обозначается как 180° западной долготы). Широта показывает, насколько далеко на севере или на юге находится точка. Экватор соответствует 0° широты, Северный полюс соответствует 90° широты (также называемый 90° с. ш.), а Южный полюс соответствует широте в –90° (также известной как 90° ю. ш.). Линии, расположенные через равные промежутки широты, называются параллелями, так как они идут параллельно экватору. Все меридианы имеют одинаковую длину, что соответствует половине круга вокруг Земли, тогда как длина параллелей зависит от их широты (см. рис. 14.1). Самой длинной параллелью является экватор, имеющий 0° широты, а самые короткие параллели — те, что находятся на Северном (90° с. ш.) и Южном (90° ю. ш.) полюсах и имеют нулевую длину.

Сложность создания карт заключается в том, что нам нужно «расплющить» сферическую поверхность Земли так, чтобы ее можно было отобразить на карте. Этот процесс, называемый проекцией, всегда приводит к искажениям, поскольку изогнутую поверхность нельзя в точности перенести на плоскость. Фактически проекция может сохранять либо углы, либо площади, но не то и другое вместе. Проекция, которая сохраняет углы, называется равноугольной, а проекция, которая сохраняет площади, — равновеликой. Существуют и другие проекции, которые вместо сохранения углов и площадей сохраняют какие-либо другие представляющие интерес величины, например расстояния до некоторой выбранной точки или линии. Наконец, существует отдельный класс проекций, который представляет собой попытку найти компромисс между сохранением углов и площадей. Такого рода проекции часто используются для отображения всего мира в эстетически привлекательной манере, допуская при этом некоторую степень угловых и пространственных искажений (см. рис. 2.11). Различные комитеты и учреждения, занимающиеся стандартизацией, ведут учет всех существующих методов проецирования с целью их систематизации и отслеживания. Такими организациями, в частности, являются European Petroleum Survey Group (EPSG) и Environmental Systems Research Institute (ESRI). Так, например, система координат EPSG:4326 представляет непроецированные значения долготы и широты в системе координат WGS 84, используемой GPS. Существует несколько веб-сайтов, которые обеспечивают удобный доступ к зарегистрированным системам проекций, например http://spatialreference.org/ и https://epsg.io/.

Одной из самых ранних проекций Земли является проекция Меркáтора, которая была изобретена в XVI веке для удобства морской навигации. Эта равноугольная проекция точно передает формы, но вносит серьезные пространственные искажения вблизи полюсов (рис. 14.2). Проекция Меркатора отображает земной шар на цилиндр, после чего цилиндр разворачивается, чтобы получилась прямоугольная карта. Меридианы в этой проекции — это равномерно расположенные вертикальные линии, а параллели — это горизонтальные линии, расстояние между которыми увеличивается по мере удаления от экватора. Расстояние между параллелями увеличивается пропорционально тому, насколько сильно сами параллели должны быть растянуты (чем ближе к полюсам, тем сильнее), чтобы меридианы были строго прямыми и вертикальными.

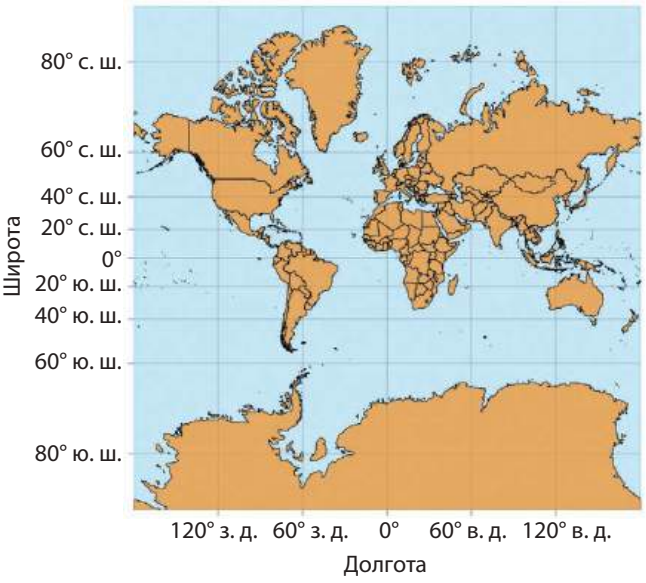


Рис. 14.2. Проекция Меркатора поверхности Земли. На данной проекции параллели представляют собой прямые горизонтальные линии, а меридианы — прямые вертикальные. Данная проекция относится к категории равноугольных, поэтому все углы на карте идентичны реальным, однако территории, находящиеся близко к полюсам, перенесены с очень серьезными искажениями. В частности, рассматривая данное изображение, можно решить, что Гренландия и Африка имеют одинаковый размер, но на самом деле площадь Африки в 14 раз больше площади Гренландии (см. рис. 14.1 и 14.3)

Из-за сильных искажений, которые проекция Меркатора вносит на карту, эта проекция давно вышла из употребления. Однако на практике все еще встречаются улучшенные варианты данного метода проецирования. Например, поперечная проекция Меркатора обычно используется для создания

крупномасштабных карт, на которых показаны относительно небольшие площади (менее нескольких градусов долготы) при большом увеличении. Другой вариант — веб-проекция Меркатора — был представлен компанией Google для своего сервиса Google Maps и используется в некоторых картографических приложениях, доступных онлайн.

Способ проецирования, который сохраняет площади без какого-либо искажения, называется проекцией Гуда (рис. 14.3). Как правило, она имеет прерывистую форму с одним разрезом в Северном полушарии и тремя в Южном, которые рассчитаны таким образом, чтобы крупные участки суши нигде не прерывались (см. рис. 14.3). Упомянутые разрезы позволяют сохранить и масштабы суши, и углы. Ценой такой точности является рассечение Гренландии пополам, Антарктики — на несколько частей, а также разрывы в океанах. Несмотря на то что проекция Гуда выглядит довольно своеобразно, она является отличным способом точно передать площади в масштабе Земли.

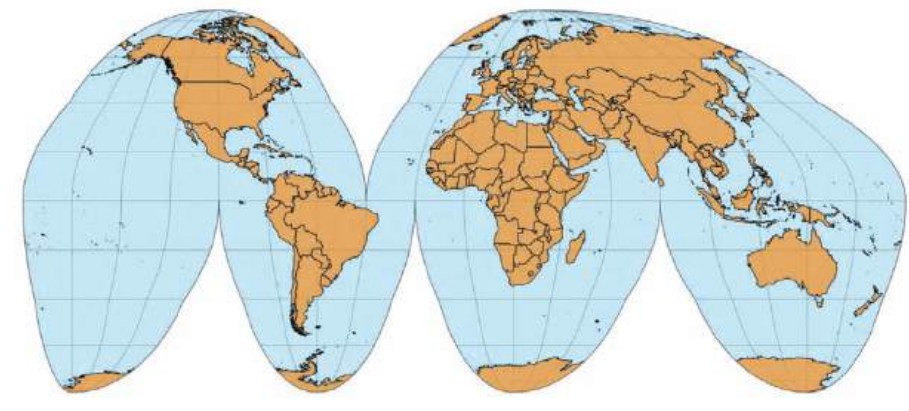


Рис. 14.3. Земля в проекции Гуда. Данная проекция прекрасно переносит на плоскость площади, при этом не сильно искажая углы. К сожалению, за точность приходится платить отсутствием целостности океанов и некоторых частей суши (Гренландия, Антарктика)

Искажения формы или площади, вызванные картографическими проекциями, особенно заметны в случае создания карты всего мира, однако они могут стать источником проблем даже на уровне отдельных континентов или стран. В качестве примера давайте рассмотрим Соединенные Штаты Америки, состоящие из 48 смежных штатов (иногда их называют континентальными), Гавайских островов и Аляски. Если нанесение на карту смежных штатов не представляет никакой проблемы, то в случае с Аляской и Гавайями, которые находятся на значительном удалении, проецирование всей страны целиком становится весьма нетривиальной задачей.

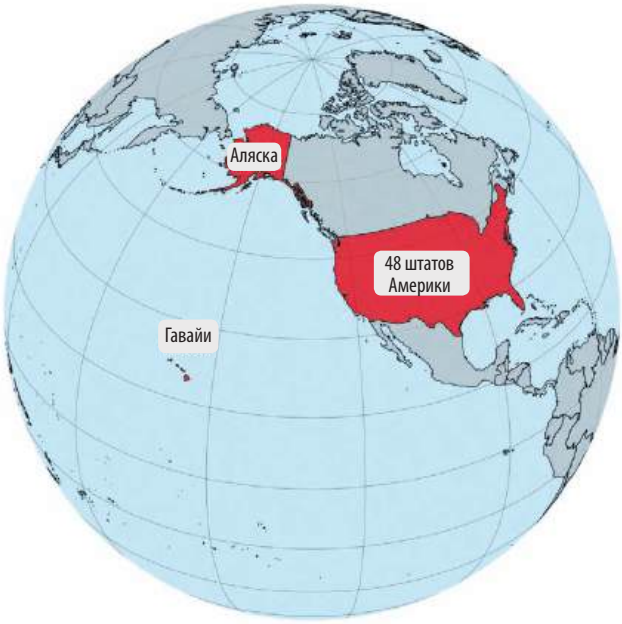


Рис. 14.4. Расположение 48 континентальных штатов США, Аляски и Гавайских островов относительно друг друга в масштабах земного шара

На рис. 14.5 показана карта всех 50 штатов США с помощью проекции Альберса. Данный вид проекции позволяет достаточно приемлемо отображать относительные размеры и положение всех 50 штатов, но эта проекция тоже не является идеальной. Во-первых, пропорции штата Аляска неестественно вытянуты в сравнении с тем, как это выглядит на рис. 14.2 и 14.4. Во-вторых, большую часть площади на карте занимает океан (по сути, пустое место). Кроме того, было бы неплохо увеличить масштаб территории, чтобы континентальные штаты занимали больше пространства на изображении.

Чтобы уменьшить пустое пространство на карте, Аляску и Гавайи стали изображать отдельно от основных штатов (чтобы избежать искажений), располагая 49-й и 50-й штаты внизу карты (рис. 14.6). Как можно заметить, на рис. 14.6 Аляска имеет меньший размер, чем на рис. 14.5. Так произошло потому, что изменению подверглось не только местоположение этого штата, но и его размер. Несмотря на то что данный подход к визуализации является общепринятым, я отношу его к категории «плохих».

Вместо того чтобы менять масштаб территории штата, одновременно двигая последний по карте, мы можем просто переместить его без изменения масштаба (рис. 14.7).

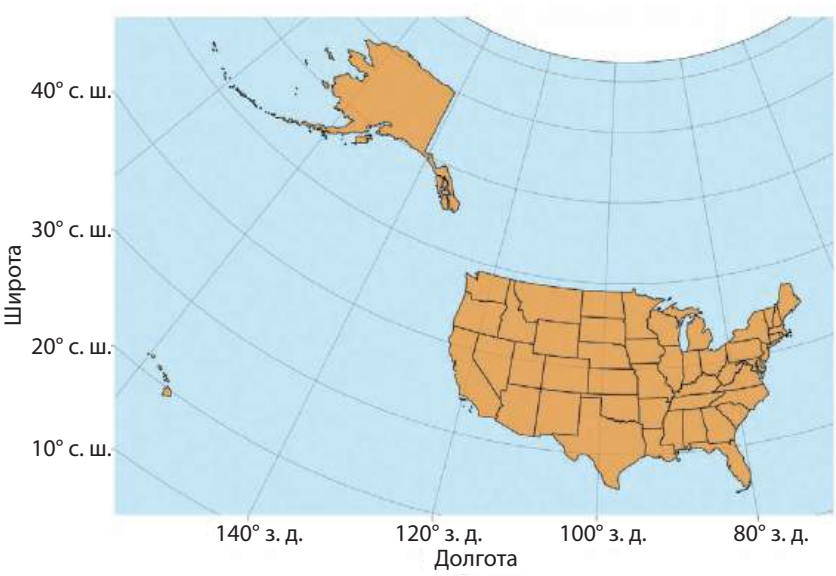


Рис. 14.5. Карта США, визуализированная посредством проекции Альберса (ESRI:102003 наиболее часто используется для проекции 48 континентальных штатов). Аляска и Гавайи показаны на своих истинных местах

Рис. 14.6. Визуализация США. Штаты Аляска и Гавайи находятся внизу карты. Обратите внимание, что линейные размеры Аляски составляют порядка 35% от истинных размеров штата (то есть площадь ее проекции составляет всего 12% от реальной). Подобного рода масштабирование часто применяется к Аляске, делая этот штат схожим по размеру с западными штатами. Несмотря на то что данный подход широко распространен, такое масштабирование сбивает с толку. Именно поэтому я отнес данное изображение к категории «плохих»

На этой визуализации хорошо видно, что Аляска является самым большим штатом Америки и имеет размер в два раза больше, чем Техас. Да, видеть США в таком виде, как показано на рис. 14.7, непривычно, однако он более корректно отображает все 50 штатов, чем рис. 14.6.

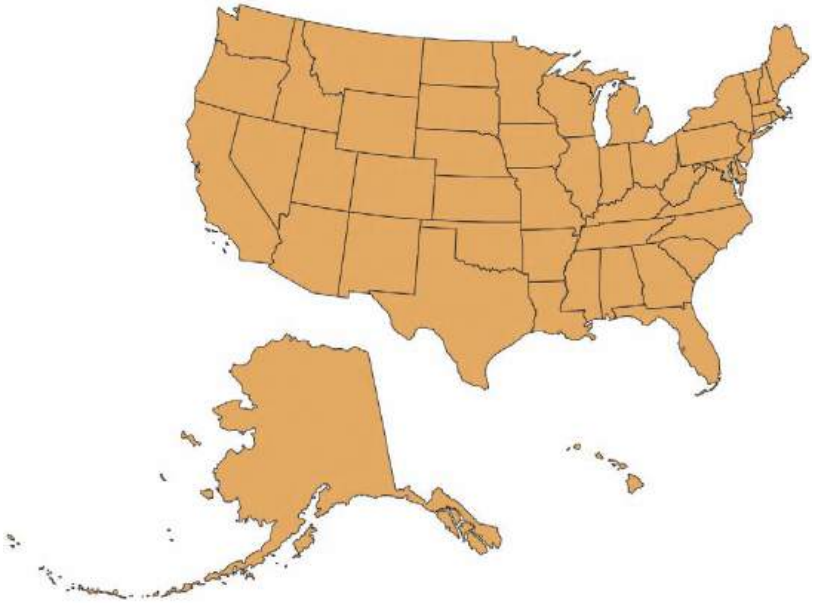


Рис. 14.7. Визуализация США. Штаты Аляска и Гавайи находятся под 48 континентальными штатами

Слои

Для визуализации геопространственных данных обычно создаются карты, состоящие из нескольких слоев, каждый из которых отображает свой тип информации. Чтобы понять, как это работает, давайте посмотрим на представление расположения ветрогенераторов в области залива Сан-Франциско. Турбины в данной местности сосредоточены в двух районах. Первый называется Ветряная ферма Шайло (Shiloh) и находится поблизости Рио-Виста, а второй район расположен неподалеку от городов Хэйворд и Трейси.

Рисунок 14.8 содержит четыре различных слоя. Первый и самый нижний слой — уровень земной поверхности, он показывает расположение холмов, долин и водных объектов. Следующий слой показывает расположение дорог на местности. Поверх слоя с дорогами я разместил слой, на котором нанесены индивидуальные ветрогенераторы, а также прямоугольники, указывающие расположение их крупных скоплений. Последний — самый

верхний — слой содержит названия и расположение городов. На рис. 14.9 все эти слои видны по отдельности. Когда мы хотим построить визуализацию на основе карты, мы можем как добавить новые слои, так и убрать часть существующих. Например, если мы захотим показать карту избирательных округов, то информация о холмах, озерах и прочих топографических объектах будет лишней. А если мы захотим сделать карту крыш, пригодных для размещения солнечных панелей, мы можем заменить слой с ландшафтом на снимки со спутника, где видны отдельные крыши и растительность. На большинстве онлайн-карт, например Google Maps, вы можете увидеть каждый из этих слоев. Вне зависимости от того, какие слои имеются на вашей карте, я настоятельно рекомендую добавлять на карту шкалу масштаба и стрелку, указывающую на север. Информация о масштабе позволяет понять размер объектов на карте, а стрелка уточняет ее ориентацию.

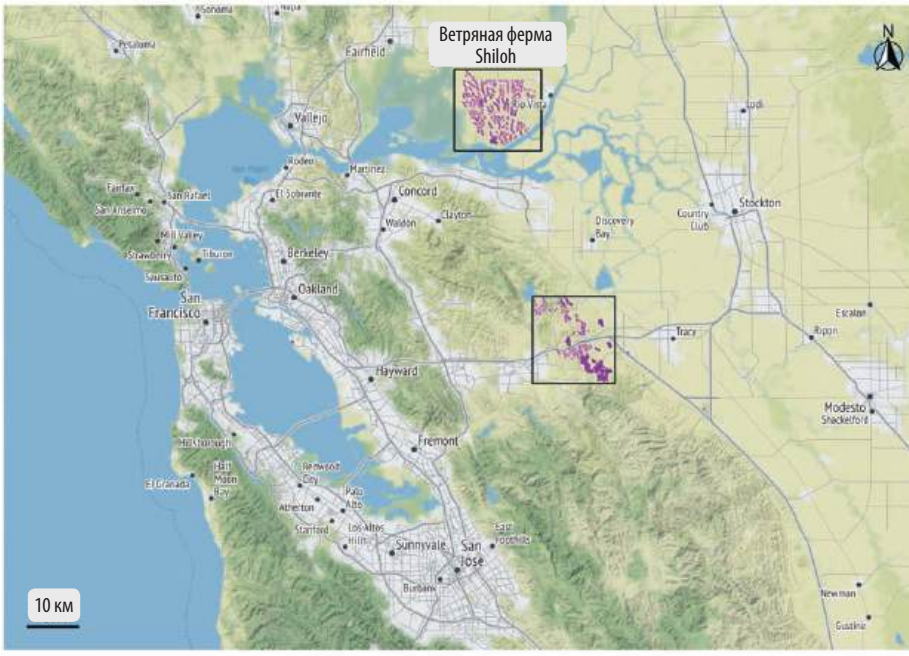


Рис. 14.8. Расположение ветрогенераторов в области залива Сан-Франциско. Отдельно стоящие генераторы обозначены фиолетовыми точками. Черные прямоугольники показывают две области с наибольшей плотностью размещения ветрогенераторов. Большое скопление ветрогенераторов, которое находится неподалеку от Рио-Виста, обозначено на карте как Ветряная ферма Шайло. Карта создана Stamen Design и предоставлена по лицензии CC BY 3.0. Данные карты представлены компанией OpenStreetMap по лицензии ODbL. Источник данных о местоположении ветрогенераторов: US Wind Turbine Database

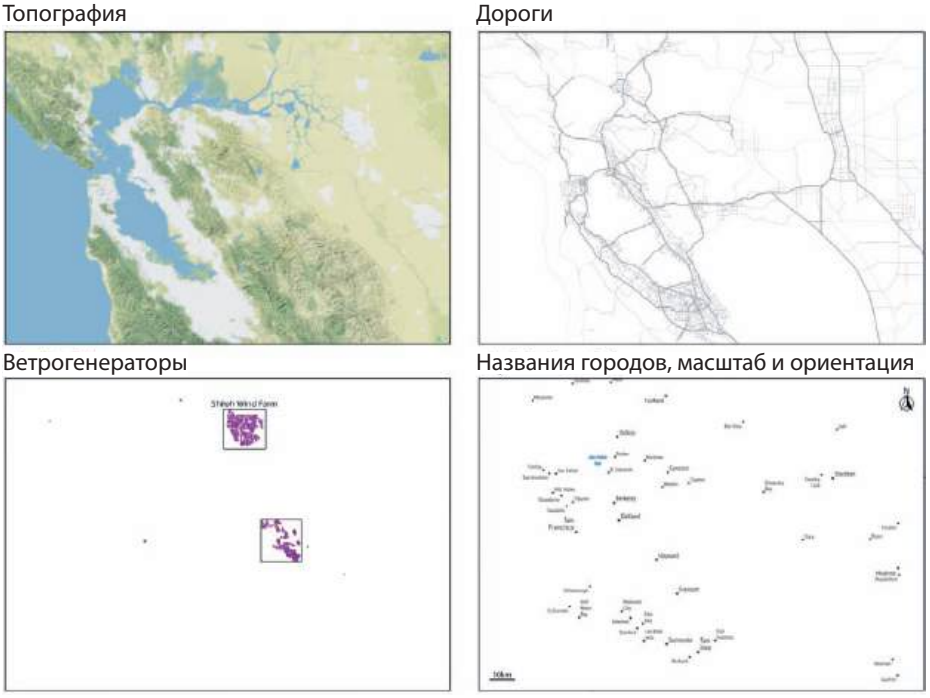


Рис. 14.9. Визуализированные по отдельности слои с рис. 14.8. Слева направо и сверху вниз на изображении показаны: топографический слой, слой с нанесенной дорожной сетью, слой с расположениями ветрогенераторов, а также слой с названиями городов, полосой масштаба и стрелкой, указывающей на север. Карта создана Stamen Design и предоставлена по лицензии CC BY 3.0. Данные карты представлены компанией OpenStreetMap по лицензии ODbL. Источник данных о местоположении ветрогенераторов: US Wind Turbine Database

В главе 1 мы говорили о различных концепциях, которые увязывают воедино данные и эстетику. Все эти рассуждения в полной мере относятся и к картам. Мы можем нанести точки данных на карту, а все остальное показать при помощи других визуальных элементов: например, цвета и формы. Взгляните на рис. 14.10, который представляет собой увеличенную карту Ветряной фермы Шайло с рис. 14.8. Здесь отдельные ветрогенераторы показаны в виде точек, цвет которых обозначает период постройки, а форма — принадлежность к проекту, в рамках которого турбина была построена. С помощью подобных карт можно получить представление о том, как шло развитие региона. Например, видно, что проект EDF Renewables был относительно небольшим, работа над ним шла до 2000 года. Проект High Winds имеет средний размер, период возведения — с 2000 по 2004 год. Наиболее крупными являются проекты Shiloh и Solano, соответственно, их создание заняло довольно много времени.

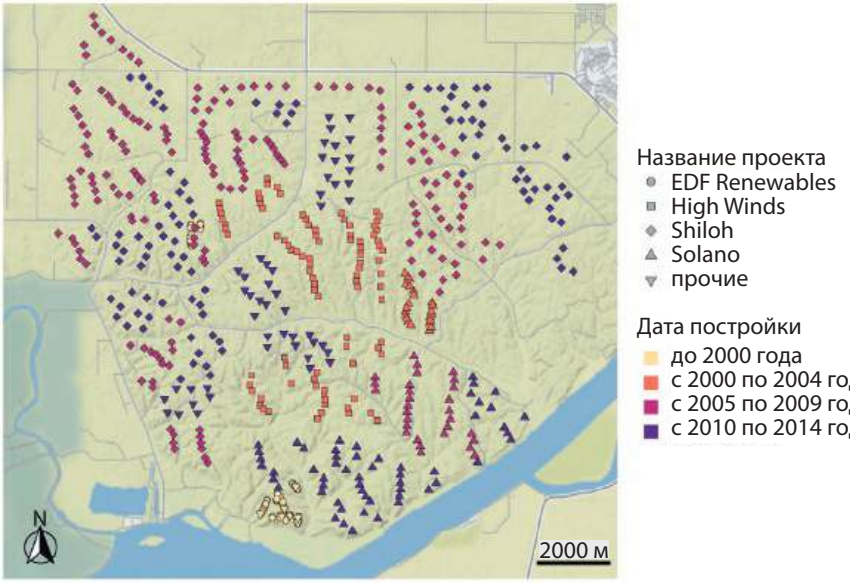


Рис. 14.10. Расположение отдельных ветрогенераторов на Ветряной ферме Шайло. Турбины обозначены точками. Охватываемая картой площадь соответствует территории верхнего правого прямоугольника с рис. 14.8. Цвет точек обозначает период постройки ветрогенератора, а форма — принадлежность к проекту. Карта создана Stamen Design и предоставлена по лицензии CC BY 3.0. Данные карты представлены компанией OpenStreetMap по лицензии ODbL. Источник данных о местоположении ветрогенераторов: US Wind Turbine Database

Фоновые картограммы

Широко распространенной практической задачей является визуализация изменения какой-либо величины в зависимости от местности. Одним из способов решения этой задачи является раскраска областей на карте в зависимости от значения показателя, который мы хотим визуализировать. Такие карты называются *фоновыми картограммами* или *хороплетами*.

В качестве простого примера рассмотрим информацию о плотности населения (человек/км^2) по всей территории США. Возьмем количество населения каждого из округов США, разделим его на площади соответствующих округов, а затем нарисуем карту, где цветом будет обозначено отношение численности населения к площади (рис. 14.11). Из данного изображения следует, что большие города на Восточном и Западном побережьях являются самыми густонаселенными регионами США, Великие равнины имеют низкую плотность населения, а штат Аляска — наименее населенный среди всех штатов США.

На рис. 14.11 светлые оттенки используются для обозначения областей с наименьшей плотностью населения, а темные оттенки — с наибольшей.

Вследствие этого городские агломерации выглядят как темные точки на светлом фоне. Если визуализация выполнена в светлых тонах, темные оттенки будут ассоциироваться с более высокими значениями плотности. Можно сделать и наоборот: более высокую плотность обозначить светлым цветом, а общей палитре графика придать темный оттенок (рис. 14.12). Если более светлые оттенки относятся к красной и желтой частям спектра, как будто они светятся, мы склонны воспринимать такие цвета как признак наибольшей плотности. Существует правило хорошего тона: если изображение будет печататься на бумаге, следует отдать предпочтение более светлому фону (как на рис. 14.11). В случае если вы планируете опубликовать вашу визуализацию в интернете или будете показывать ее на каком-то темном фоне, может быть лучше сделать основные области графика темного оттенка (как на рис. 14.12).

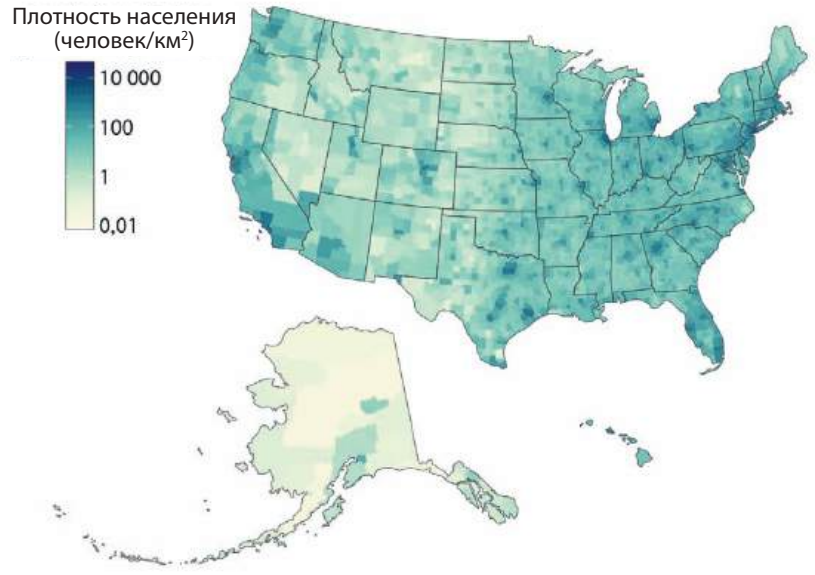


Рис. 14.11. Плотность населения в каждом регионе США, визуализированная в виде фоновой картограммы. Единица измерения плотности населения — количество человек на квадратный километр. Источник: 2015 Five-Year American Community Survey

Фоновые картограммы лучше всего подходят для тех случаев, когда цвет используется для отображения плотности (то есть речь идет о распределении некоторого количества объектов по какой-либо территории; примерами могут служить рис. 14.11 и 14.12). Наше восприятие говорит нам, что обычно наибольшая площадь территории соответствует наибольшему количеству (см. главу 16), а закрашивание областей в зависимости от плотности сводит этот эффект на нет. Однако на практике часто бывает так, что фоновые картограммы окрашены в соответствии с какой-то величиной, которая не является

плотностью. Взгляните на рис. 3.4, который представляет собой фоновую картограмму среднего годового дохода в регионах штата Техас. Такие графики не являются чем-то плохим, однако к их созданию следует подходить с повышенной аккуратностью. Существует два условия, при которых можно обозначать цветом количество, а не плотность: когда все области имеют примерно одинаковый размер и форму (в таком случае мы можем не бояться, что часть областей будут перетягивать внимание на себя ввиду их большего размера). Вторым условием является то, что каждая отдельная закрашиваемая область должна быть сравнительно мала по сравнению с общим размером карты, а количество, передаваемое цветом, должно меняться в масштабе, большем, чем отдельные раскрашенные области. На рис. 3.4 соблюдены оба этих условия.

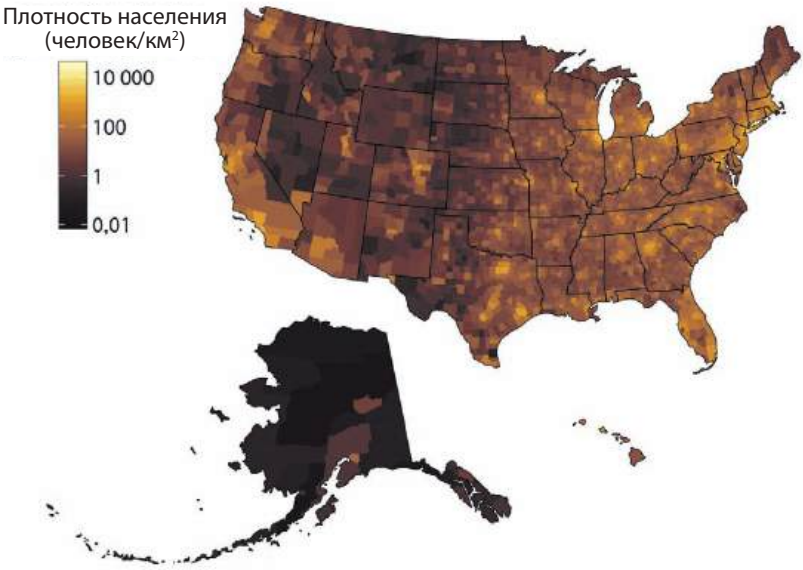


Рис. 14.12. Плотность населения в каждом регионе США, визуализированная в виде фоновой картограммы. Данный рисунок полностью идентичен рис. 14.11, за исключением того, что темный цвет используется для обозначения наименьшей плотности, а светлый — наоборот. Источник: 2015 Five-Year American Community Survey

При создании фоновых картограмм очень важно не забывать о разнице между непрерывными и дискретными цветовыми шкалами. Несмотря на то что непрерывные цветовые шкалы, как правило, выглядят красиво (взгляните на рис. 14.11 и 14.12), в то же время они могут быть тяжелыми для восприятия. Человеку сложно дается распознавание конкретного значения цвета и последующее его сопоставление с непрерывной шкалой. Поэтому часто бывает целесообразно сгруппировать данные в зависимости от цвета, которым они представлены. Хорошим вариантом является группировка данных

в 4–6 групп (ячеек). Подобного рода действие приносит в жертву некоторый объем полезной информации, однако воспринимать цвета, если они разбиты на небольшое количество групп, становится проще. В качестве иллюстрации ко всему сказанному можно привести рис. 14.13, на котором отображение среднего дохода в округах Техаса (см. рис. 3.4) расширено на все округа США. На данном изображении используется цветовая шкала, представляющая пять групп по объему среднего дохода.

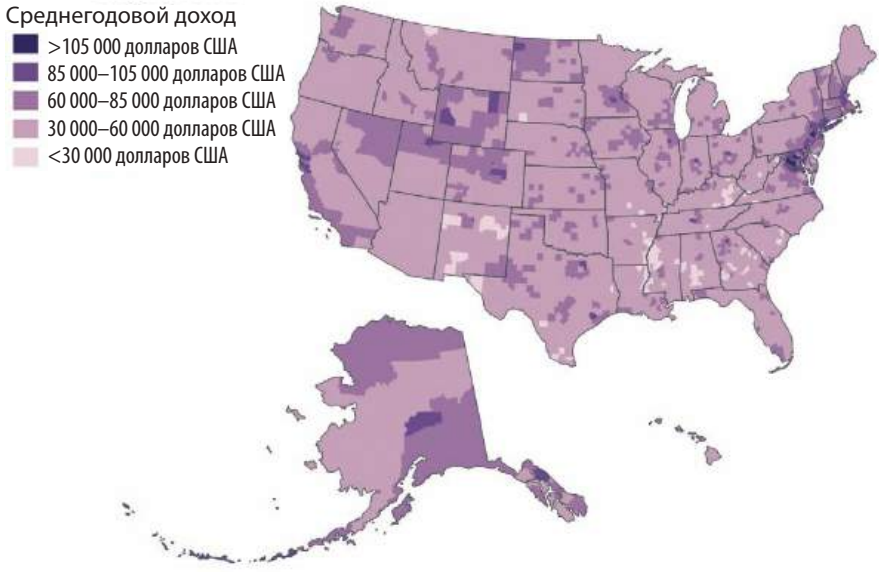


Рис. 14.13. Среднегодовой доход в каждом округе США, визуализированный в виде фоновой картограммы. Значения дохода разбиты на пять различных групп, поскольку шкалы, сгруппированные по цвету, воспринимаются легче, чем непрерывные. Источник: 2015 Five-Year American Community Survey

Несмотря на то что округа различаются по форме и размеру, я считаю, что рис. 14.13 как фоновая картограмма вполне жизнеспособен. Никакой из округов не доминирует на карте, однако картина будет выглядеть совершенно иначе, если территория будет разбита на штаты, а не на округа (рис. 14.14). В этом случае на карте появится очевидный доминант — штат Аляска, вследствие чего, учитывая размер этого штата, можно предположить, что для данного штата показатель среднего дохода свыше 70 000 долларов США является обычным явлением. Однако не стоит забывать о том, что Аляска чрезвычайно малонаселена (см. рис. 14.11 и 14.12). Таким образом, уровень среднегодового дохода на Аляске относится только к небольшой части населения США. Подавляющее большинство жителей регионов США, имеющих бóльшую по сравнению с Аляской численность населения, в среднем за год получают менее 60 000 долларов.

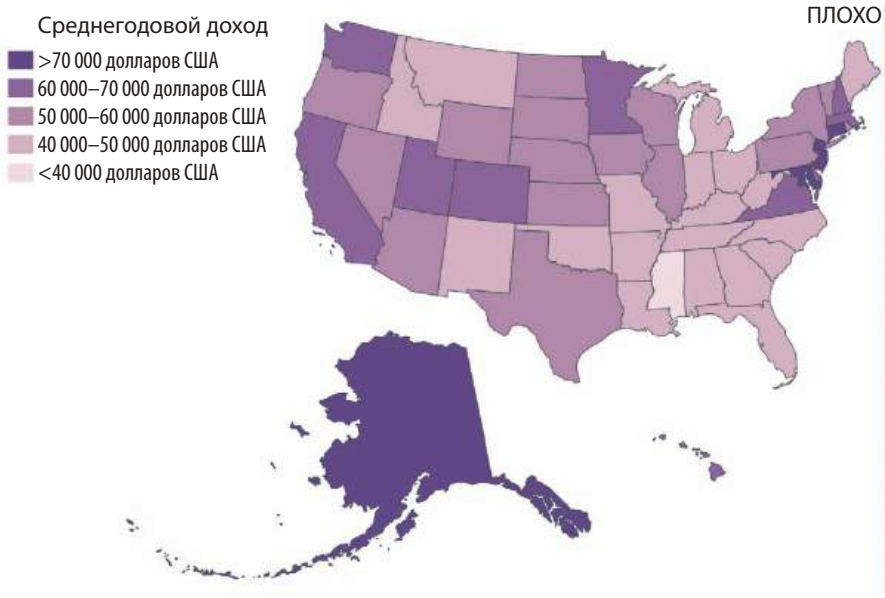


Рис. 14.14. Среднегодовой доход в каждом штате США, представленный в виде фоновой картограммы. На этой карте визуально доминирует штат Аляска, имеющий высокий средний доход и вместе с тем очень низкую плотность населения. В то же время густонаселенные штаты на Восточном побережье с их высоким уровнем дохода несколько «теряются» на этой карте. В целом этот график дает слабое представление о распределении доходов в США, поэтому я отнес его к категории «плохих». Источник: 2015 Five-Year American Community Survey

Картограммы

Вообще говоря, не каждая визуализация, выполненная в виде карты, обязана быть географически точной, чтобы читатель мог извлечь из нее пользу. Давайте еще раз посмотрим на рис. 14.14. Его проблема заключается в том, что некоторые штаты имеют большую площадь и вместе с тем низкий показатель численности населения, в то время как в других штатах присутствует большое количество жителей, но при этом площадь этих штатов довольно мала. Но что, если изменить форму штатов таким образом, чтобы их размер стал пропорционален количеству жителей? Такая модифицированная карта называется *картограммой*, а рис. 14.15 демонстрирует, как может выглядеть график этого типа для случая набора данных о среднегодовом доходе. На этой картограмме мы все еще можем узнать некоторые штаты, однако несомненно, что корректировка по численности населения значительно изменила вид графика. Штаты Восточного побережья, Флорида и Калифорния сильно выросли в размерах, в то время как западные штаты и Аляска сильно уменьшились.

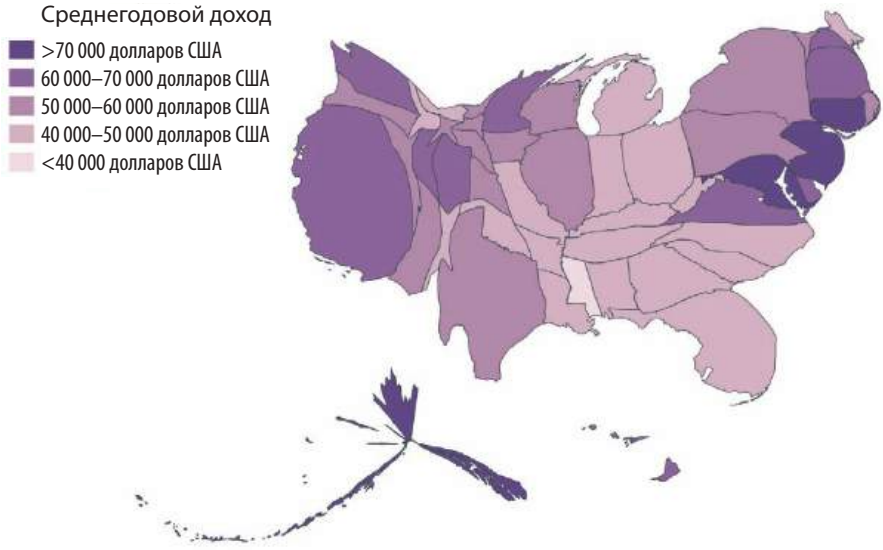


Рис. 14.15. Среднегодовой доход жителей в каждом штате США, визуализированный в виде картограммы. Формы отдельных штатов были изменены таким образом, чтобы их площадь была пропорциональна численности населения. Источник: 2015 Five-Year American Community Survey

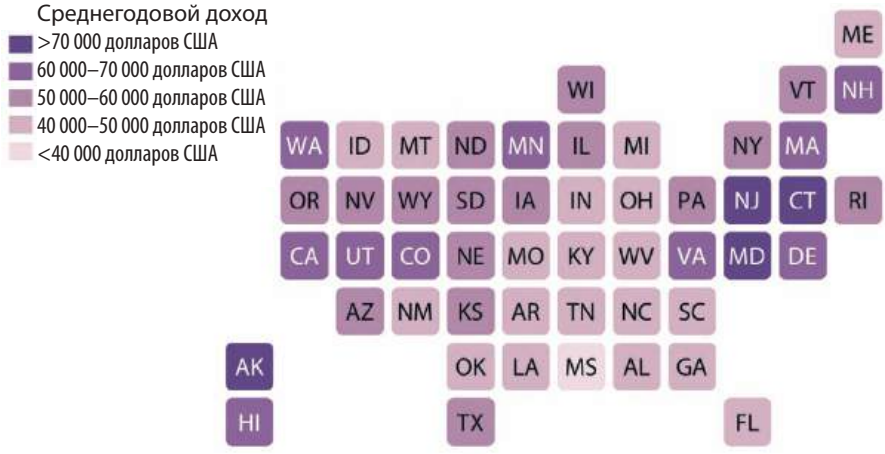


Рис. 14.16. Среднегодовой доход жителей в каждом штате США, визуализированный в виде тепловой картограммы. Штаты представлены в виде квадратов одинакового размера, которые приблизительно находятся на местах соответствующих штатов на карте. Такой способ визуализации придает каждому штату одинаковый визуальный вес

Альтернативой картограммы с искаженными формами является график под названием *тепловая картограмма*, на которой каждый штат представлен в виде цветного квадрата (рис. 14.16). Несмотря на то что этот график

не делает поправку на численность населения, из-за чего он в недостаточной степени представляет более густонаселенные штаты и в избыточной — менее густонаселенные, данная визуализация одинаково показывает все штаты, вне зависимости от их формы или размера.

Более того, мы можем создать еще более сложную картограмму, разместив отдельные графики на карте в тех местах, где находится каждый штат. Например, если мы хотим визуализировать изменение уровня безработицы с течением времени для каждого штата, мы можем нарисовать отдельный график для каждого из них, а затем поместить этот график на место соответствующего штата (рис. 14.17). Для тех, кто знаком с географией США, этот способ расположения должен упростить поиск графика конкретного штата по сравнению с ситуацией, когда штаты, например, расположены в алфавитном порядке. Кроме того, логично ожидать, что штаты, которые находятся по соседству, будут демонстрировать сходное положение дел. Рис. 14.17 подтверждает существование упомянутых тенденций.

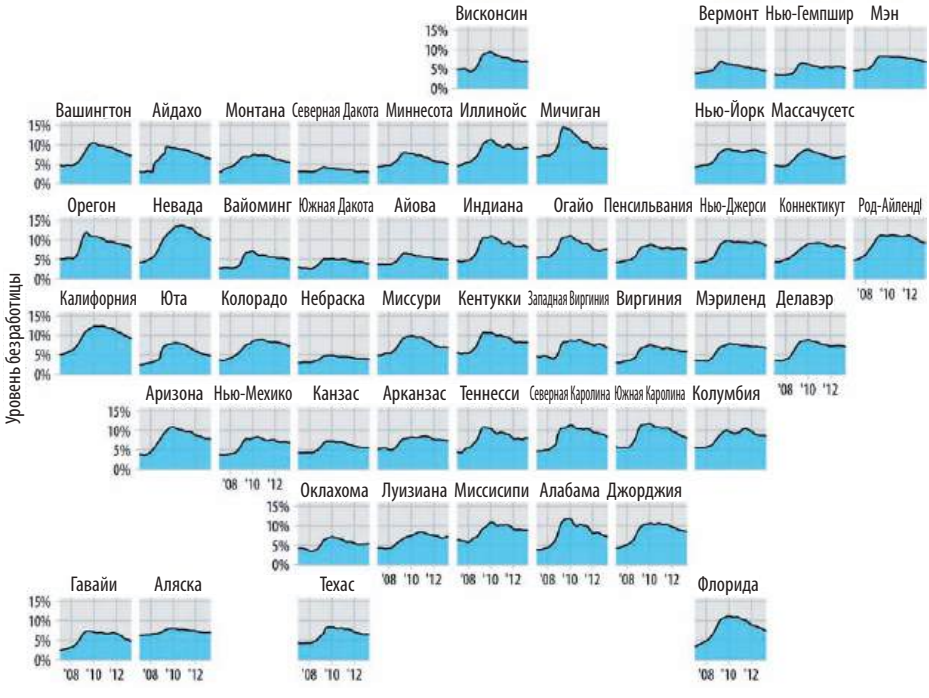


Рис. 14.17. Уровень безработицы в период, предшествующий финансовому кризису 2008 года и после него, в зависимости от штата. Каждый график показывает уровень безработицы в одном штате, включая округ Колумбия, с января 2007 года по май 2013 года. Вертикальными линиями обозначен январь 2008, 2010 и 2012 годов. В географически близких штатах наблюдаются схожие тенденции в уровне безработицы. Источник: US Bureau of Labor Statistics

Глава 15

Визуализация неопределенности

Одним из наиболее сложных аспектов визуализации данных является визуализация неопределенности. Когда мы видим на графике точку, нарисованную в определенном месте, мы склонны интерпретировать ее как точное представление истинного значения данных. Представить, что точка может соответствовать данным, которые не имеют отношения к ее расположению, довольно трудно. Однако этот сценарий встречается в визуализации данных повсеместно. Почти каждый набор данных обладает некоторой степенью неопределенности, и от того, как именно мы будем интерпретировать ее, напрямую зависит точность восприятия аудиторией визуализированной нами информации.

Двумя наиболее распространенными способами визуализации неопределенности являются диапазоны погрешности и *доверительные диапазоны*. Эти подходы были разработаны в контексте научных публикаций, и поэтому для правильной интерпретации изображений, построенных на основе данных методов, требуется определенный объем экспертных знаний. Следует отметить при этом, что эти подходы весьма точны и компактны (с точки зрения размера получаемых изображений). Например, используя планки погрешностей, мы можем на одном графике показать неопределенность множества значений различных параметров. При этом, если визуализация рассчитана на широкую публику, используйте методы, которые позволяют воспринимать неопределенность на интуитивном уровне, пусть и в ущерб точности или плотности отображения данных. Как вариант, можно прибегнуть к «кадрированию» частот (об этом мы поговорим чуть ниже), когда сценарии с различными вероятностями исхода изображаются на отдельных графиках. Или использовать анимации, которые будут перемежаться между всеми возможными сценариями.

«Кадрирование» вероятностей в виде частот

Прежде чем мы приступим к визуализации неопределенности, давайте разберемся, о чем вообще идет речь. Интуитивно понять концепцию неопределенности проще всего в контексте событий с неизвестной вероятностью исхода.

Если я собираюсь подбросить монетку, я не знаю заранее, каким будет исход этого события. То есть окончательный результат неясен. Однако я точно так же могу быть не уверен в событиях из прошлого: если бы вчера я дважды выглянул из окна своей кухни, один раз в восемь утра и один раз в четыре часа дня, и в первом случае увидел бы припаркованную через дорогу красную машину, но при этом не увидел ее в следующий раз, то я мог бы сделать вывод, что в какой-то момент внутри этого восьмичасового промежутка машина уехала, но когда именно, я бы сказать не смог. Это могло быть и в 8:01 утра, и в 9:30 утра, и в два часа дня, или в любое другое время.

Когда мы имеем дело с неопределенностью в математике, мы используем понятие вероятности. Точное определение этого понятия является сложным и не входит в список тем, которые рассматриваются в этой книге. Тем не менее мы вполне можем успешно использовать это понятие, даже не понимая всей математической подоплеки. В большинстве случаев, имеющих практическое значение, достаточно использовать понятие относительных частот. Рассмотрим следующий пример. Предположим, что вы проводите какое-то случайное испытание, например бросаете монету или кидаете игральную кость, пытаясь добиться какого-то конкретного результата (например, выпадения орла на монетке или шестерки на костях). Желаемый результат можно назвать *успехом*, а любой другой — *неудачей*. Тогда вероятность успеха приблизительно равна количеству успешных событий среди общего количества попыток. Например, если говорится, что некоторый исход происходит с вероятностью 10%, это означает, что среди многих повторных испытаний этот исход будет наблюдаться примерно в 1 случае из 10.

Визуализация конкретной вероятности — задача непростая. Например, как изобразить шанс выигрыша в лотерею или шанс выпадения 6 на равновесном кубике? В каждом из этих случаев вероятность — это одно-единственное число. Мы можем рассматривать его как сумму и отображать при помощи любой из техник, рассмотренных в главе 5 (например, гистограммы или точечного графика), однако полученный результат вряд ли нам будет чем-то полезен. Большинство людей не обладает интуитивным пониманием того, как значение вероятности отображается на реальный мир. К сожалению, эту проблему нельзя решить с помощью визуализации вероятности в виде столбца или точки на линии.

Чтобы сделать концепцию вероятности осязаемой, мы можем создать график, на котором будут подчеркнуты как частотный аспект, так и непредсказуемость случайного исследования: например, если нарисовать в случайном порядке квадраты разных цветов. На рис. 15.1 я использую этот метод для визуализации трех различных вероятностей: 1%-ный шанс на успех, 10%-ный шанс на успех и 40%-ный шанс на успех. Чтобы понять этот график, представьте, что вам нужно угадать местонахождение темного квадрата, причем до того, как вы укажете на тот или иной квадрат, его цвет вам известен не

будет. (Другими словами, выбор квадрата делается вслепую.) Интуитивно вы, скорее всего, понимаете, что угадать расположение темного квадрата в случае 1% вероятности почти наверняка не получится. В случае вероятности 10% угадать тоже было бы сложно. Однако в случае 40% вероятности шансы выглядят не так уж плохо. Такой стиль визуализации, когда мы показываем конкретные потенциальные результаты, называется *визуализацией дискретных исходов*, а визуализация вероятности как частоты называется *«кадрированием» частот (framing frequency)*. Мы представляем вероятностный характер результата в виде легко воспринимаемых частот исходов.

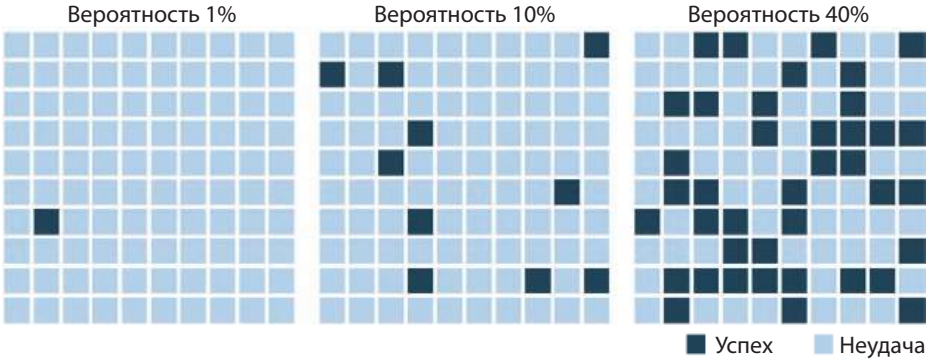


Рис. 15.1. Визуализация вероятности как частоты. Все сетки состоят из 100 квадратов, каждый квадрат представляет собой либо успех, либо неудачу в некотором случайном испытании. 1% вероятности успеха соответствует 1 темному и 99 светлым квадратам, 10% — 10 темным и 90 светлым квадратам, а вероятность 40% показана 40 темными и 60 светлыми квадратами. Если случайным образом расположить темные квадраты среди светлых, мы сумеем вызвать у зрителя ощущение случайности, которое подчеркнет неопределенность исхода каждой отдельной попытки из серии испытаний

В тех случаях, когда нас интересуют только два дискретных результата (успех или неудача), визуализация с рис. 15.1 работает хорошо. Однако на практике нам гораздо чаще приходится сталкиваться с более сложными сценариями, где результатом случайного испытания является числовая переменная. Одним из таких сценариев является ситуация предсказания результатов выборов, где нас интересует не только личность победителя, но и перевес, с которым он выиграл избирательную гонку. Рассмотрим гипотетический пример предстоящих выборов с двумя партиями — красной и синей. Предположим, что вы услышали по радио информацию, что синяя партия, по прогнозам, имеет преимущество в 1% по сравнению с желтой партией с погрешностью в 1,76%. Что эта информация говорит вам о вероятных результатах выборов? Человеческое восприятие понимает это как «синяя партия победит», однако реальность куда сложнее. Первым и наиболее значимым является тот факт, что существует целый ряд различных возможных исходов.

Синяя партия может выиграть с отрывом в 2%, или же желтая партия может выиграть с отрывом в 0,5%. Диапазон возможных исходов и связанных с ними вероятностей составляет так называемое *распределение вероятностей*, и для нашего случая мы можем изобразить его в виде гладкой кривой, которая сначала возрастает, а затем убывает в пределах диапазона возможных исходов (рис. 15.2). Чем выше конкретный результат находится на кривой, тем больше вероятность наступления данного события. Распределения вероятностей тесно связаны с гистограммами и ядерными оценками плотности, которые мы обсуждали в главе 6 (вы всегда можете вернуться к данной главе, если вам захочется освежить в памяти эту тему).

Рис. 15.2. Гипотетический прогноз результатов выборов. Ожидается, что синяя партия наберет на 1% больше голосов, чем желтая (это значение отмечено как «наиболее вероятная оценка»). Однако данный прогноз имеет погрешность: 1,76% в обоих направлениях от наиболее вероятной оценки. На графике эта планка погрешностей изображена так, чтобы покрывать 95% возможных исходов. Область, закрашенная синим цветом, соответствует 87,1% от площади под графиком и представляет собой все исходы, в которых выигрывает синяя партия. Аналогично область, закрашенная желтым, соответствует 12,9% от всех исходов и представляет все варианты исхода для случая победы желтой партии. Так, в данном примере вероятность победы синей партии составляет примерно 87%

Проделав некоторые математические вычисления, мы увидим, что в нашем вымышленном примере шанс на победу желтой партии составляет всего 12,9%. Таким образом, вероятность выигрыша желтой партии чуть выше, чем сценарий с рис. 15.1, где вероятность равна 10%. Если вам больше импонирует партия синих, вы можете не беспокоиться насчет их победы. Однако стоит отметить, что у партии желтых все же есть шансы на выигрыш в этой избирательной гонке. Сравнивая рис. 15.2 с рис. 15.1, можно заметить, что второй рисунок создает гораздо большее понимание неопределенности, чем первый, несмотря на то что заштрихованные области на рис. 15.2 полностью соответствуют

вероятностям победы партии синих и желтых. Именно этим и хороша визуализация дискретных результатов. Результаты исследования человеческого восприятия говорят о том, что нам гораздо проще воспринимать, подсчитывать и анализировать относительные частоты дискретных объектов (но только если их не очень много), чем относительные размеры различных областей.

Мы можем совместить дискретную природу рис. 15.1 с непрерывным распределением с рис. 15.2 путем построения точечной диаграммы квантилей [Kay et al., 2016]. На диаграмме такого типа площадь под кривой делится на доли одинакового размера, после чего каждая доля изображается в виде окружности. Эти окружности располагаются таким образом, чтобы их положение приблизительно соответствовало форме фигуры под исходной кривой распределения (рис. 15.3).

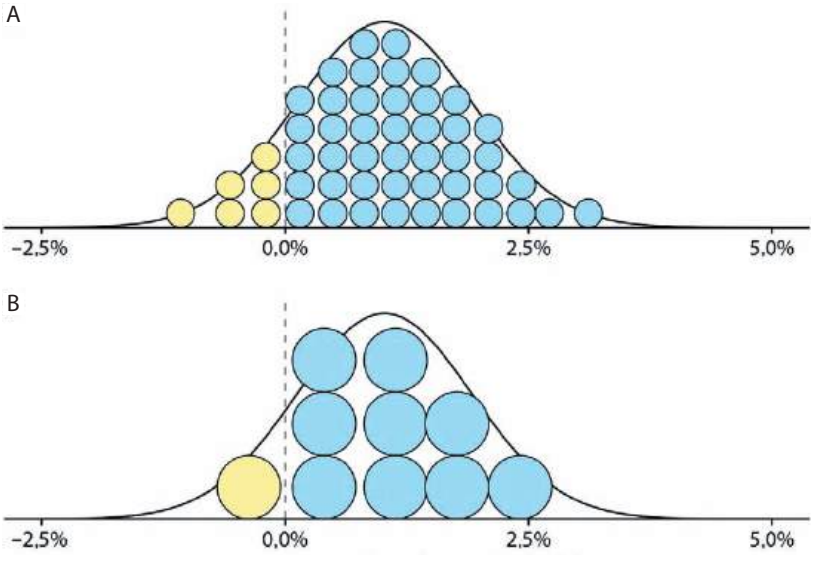


Рис. 15.3. Визуализация распределения результатов выборов с рис. 15.2 в виде точечной диаграммы квантилей. A. Сглаженное распределение аппроксимировано 50 точками, каждая из которых представляет вероятность 2%. Таким образом, 6 желтых точек соответствуют вероятности 12%, что достаточно близко к истинному значению 12,9%. B. Гладкое распределение аппроксимировано 10 точками, каждая из которых представляет вероятность 10%. Таким образом, 1 желтая точка соответствует вероятности 10%, что все еще близко к истинному значению. Точечные диаграммы квантилей с небольшим количеством точек, как правило, легче читаются, поэтому в данном конкретном примере 10-точечная версия может быть предпочтительнее 50-точечной

Из всего вышесказанного следует, что в диаграммах этого типа не следует использовать слишком много точек, иначе читатель будет склонен воспринимать вашу визуализацию больше как непрерывную, а не как набор отдельных,

дискретных элементов, что сводит на нет все преимущества этих графиков. На рис. 15.3 показаны варианты с 50 точками (рис. 15.3A) и с 10 точками (рис. 15.3B). Несмотря на то что версия с 50 точками более точно фиксирует истинное распределение вероятностей, количество точек слишком велико, поэтому читателю будет трудно выделить какую-то одну из них. Версия с 10 точками более точно передает относительные шансы на выигрыш синей или желтой партии. Единственным ее минусом является неточность. Мы недооцениваем вероятность победы желтых на 2,9%. Однако в большинстве случаев пренебрежение некоторой математической точностью ради более качественного восприятия, особенно широкой публикой, вполне оправданно. От визуализации, которая является математически правильной, но при этом плохо воспринимается зрителями, будет мало толку.

Визуализация неопределенности точечной оценки

На рис. 15.2 присутствуют метки «наилучшей оценки» и «погрешности», однако я пока ничего не сказал о том, что это за числа и откуда они берутся. Чтобы разобраться в этом вопросе, нам придется кратко ознакомиться с базовыми понятиями статистической выборки. Главная цель статистики — понять что-то об окружающем мире, основываясь на знаниях о небольшой его части. Продолжая пример с выборами, предположим, что существует множество различных избирательных округов и жители каждого из них будут голосовать либо за синюю, либо за желтую партию. Вероятно, было бы интересно заранее узнать, за кого планируют голосовать жители округов, а также общее среднее число голосов по округам. Но для получения предвыборного прогноза мы не можем опрашивать абсолютно всех жителей, чтобы узнать, за кого они собираются отдать свой голос. К счастью, нам это и не нужно — вполне достаточно опроса некоторого подмножества жителей из некоторого подмножества округов, чтобы затем использовать эти данные для получения результата, наиболее близкого к тому, что будет в реальности. Говоря языком статистики, совокупность возможных голосов всех граждан во всех округах называется *генеральной совокупностью*, а подмножество граждан и/или округов, которые мы опрашиваем, называется *выборкой*. Генеральная совокупность представляет собой истинное состояние мира, а выборка — это наше окно в мир.

Как правило, нас интересуют конкретные величины, которые отражают наиболее важные характеристики совокупности. В примере с выборами это могут быть средние результаты голосования по округам или стандартное отклонение между результатами округов. Значения, которые описывают генеральную совокупность, называются *параметрами*, и их, как правило,

узнать нельзя никак. Однако с помощью выборки мы можем сделать некоторое предположение об истинных значениях параметров — в статистике такие предположения называются *оценками*. Среднее значение по выборке является оценкой среднего значения по совокупности, которое является параметром. Оценки отдельных значений параметров также называются *точечными оценками*, поскольку каждая из них может быть представлена в виде точки на прямой.

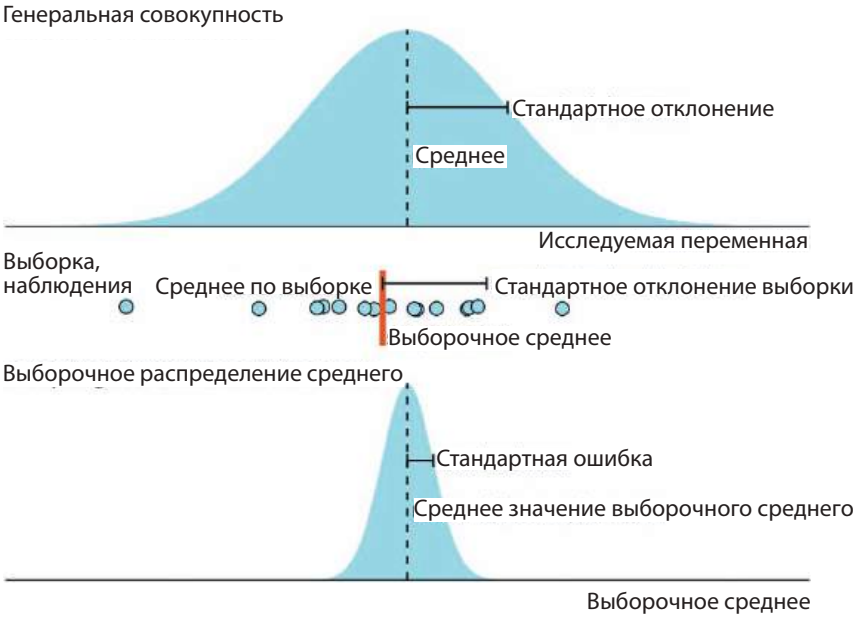


Рис. 15.4. Основные понятия статистической выборки. Исследуемая переменная, которую мы изучаем, имеет некоторое истинное распределение в генеральной совокупности, с «истинными» для генеральной совокупности средним и стандартным отклонением. Любая конечная выборка этой переменной будет иметь выборочное среднее значение и стандартное отклонение, которые отличаются от параметров генеральной совокупности. Если бы мы провели многократную выборку и всякий раз вычисляли ее среднее значение, то полученные средние были бы распределены в соответствии с выборочным распределением среднего. Стандартная ошибка дает информацию о ширине выборочного распределения и говорит нам о том, насколько точно мы оцениваем интересующий параметр (здесь — среднее значение по генеральной совокупности)

Рис. 15.4 иллюстрирует связь этих ключевых понятий друг с другом. Исследуемая переменная (например, результат голосования в каждом округе) имеет некоторое распределение среди генеральной совокупности, характеризуясь средним значением и стандартным отклонением генеральной совокупности. Выборка состоит из набора конкретных наблюдений. Количество отдельных наблюдений в выборке называется размером выборки. На основании

данных выборки мы можем рассчитать ее среднее значение и стандартное отклонение. Эти значения, как правило, отличаются от среднего значения и стандартного отклонения для генеральной совокупности. Наконец, мы можем определить *выборочное распределение*, представляющее собой распределение оценок, которое мы могли бы получить, если бы повторили процесс выборки много раз. Ширина выборочного распределения называется стандартной ошибкой, которая означает степень точности наших оценок. Другими словами, стандартная ошибка представляет собой меру неопределенности нашей оценки параметра. Как правило, чем больше размер выборки, тем меньше величина стандартной ошибки и, следовательно, неопределенности нашей оценки.

Между стандартным отклонением и стандартной ошибкой есть принципиальная разница. Стандартное отклонение — это свойство совокупности, которое говорит нам о том, насколько велик разброс между отдельными наблюдениями, которые мы могли бы сделать. Например, если мы изучаем генеральную совокупность избирательных округов, то стандартное отклонение покажет нам, насколько округа отличаются друг от друга. Стандартная ошибка же показывает, насколько точна оценка параметра, которую мы сделали по выборке. Если нам понадобится оценить средний результат голосования по всем округам, стандартная ошибка подскажет нам, насколько точной является наша оценка среднего.

Все специалисты в области статистики используют выборки для расчета оценок параметров и их неопределенностей, однако подходят они к этому поразному. Существует два различных подхода к этим вычислениям: байесовский и частотный. Сторонники байесовского подхода предполагают, что у них есть некоторые предварительные знания о мире, и используют полученную выборку для обновления этих знаний. Приверженцы частотных методов пытаются делать точные утверждения о мире, не опираясь на какие-либо уже имеющиеся знания. Тем не менее, когда дело доходит до визуализации неопределенности, и те и другие специалисты могут применять одни и те же типы стратегий. Далее, мы для начала рассмотрим частотный подход, а затем поговорим о некоторых проблемах, специфичных для байесовского метода.

Приверженцы частотного подхода чаще всего визуализируют неопределенность с помощью планок погрешностей. Последние, при всей их полезности как метода визуализации неопределенности, обладают некоторыми недостатками, о которых я уже упоминал в главе 8 (см. рис. 8.1). Читатели могут попросту не понять, что означают планки погрешностей. На рис. 15.5 эта проблема проиллюстрирована с помощью пяти различных вариантов использования планок погрешностей для одного и того же набора данных, который содержит экспертные оценки по шкале от 1 до 5 для шоколадных батончиков, изготовленных в разных странах. Для создания рис. 15.5 я использовал

оценки батончиков, произведенных в Канаде. Под исходными данными выборки, изображенными в виде полосовых графиков, мы видим среднее значение выборки плюс/минус стандартное отклонение для выборки, среднее значение выборки плюс/минус стандартная ошибка, а также доверительные интервалы в 80, 95 и 99%. Все пять планок погрешностей нарисованы на основе отклонений в выборке, все они математически связаны, но имеют разные значения. Кроме того, визуально их очень легко отличить друг от друга.

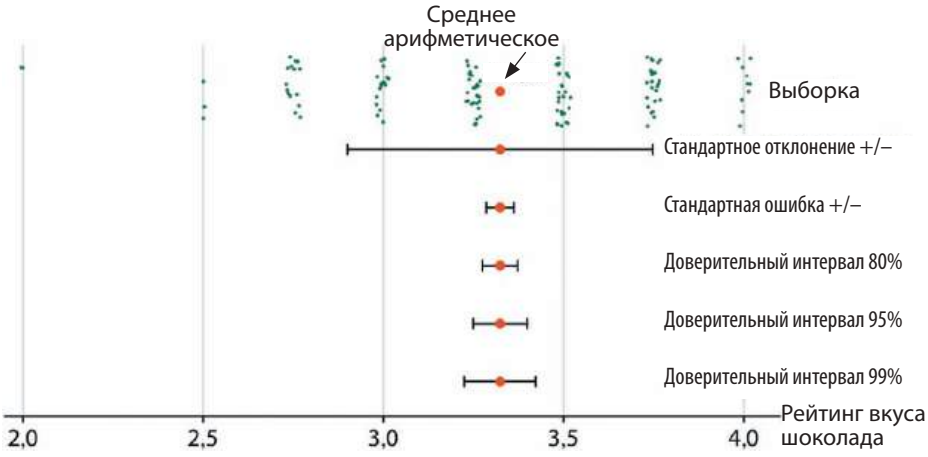


Рис. 15.5. Взаимосвязь между выборкой, средним по выборке, стандартным отклонением, стандартной ошибкой и доверительными интервалами на примере рейтинга шоколадных батончиков. Наблюдения (показанные в виде разбросанных групп зеленых точек), из которых состоит выборка, представляют собой экспертные оценки 125 шоколадных батончиков канадского производства, распределенные по дискретной шкале от 1 (неприятно) до 5 (идеально). Большая оранжевая точка обозначает среднее арифметическое значение рейтинга. Планки погрешностей показывают сверху вниз: двойное стандартное отклонение, двойную стандартную ошибку (стандартное отклонение среднего) и доверительные интервалы для вероятностей 80, 95 и 99% для среднего значения. Источник: Brady Brelinski, Manhattan Chocolate Society

Визуализируя неопределенность с помощью планок погрешностей, обязательно точно указывайте, какую величину и с какой доверительной вероятностью (если применимо) они отображают.

Примерное значение стандартной ошибки можно получить путем деления стандартного отклонения выборки на квадратный корень размера выборки, а доверительные интервалы рассчитываются путем умножения стандартной ошибки на небольшие постоянные значения. Например, доверительный интервал 95% примерно в два раза больше стандартной ошибки в любую сторону от среднего значения. Поэтому большие выборки, как правило,

характеризуются более узкими стандартными ошибками и доверительными интервалами даже в том случае, если их стандартное отклонение одинаково. Этот эффект хорошо заметен, если сравнивать рейтинги шоколадных батончиков из Канады и Швейцарии (рис. 15.6). Средняя оценка и стандартное отклонение выборки сопоставимы для шоколадных батончиков из этих двух стран, однако в нашем распоряжении имеются оценки для 125 канадских батончиков и лишь для 38 швейцарских, поэтому доверительные интервалы вокруг среднего значения намного шире для шоколадок, произведенных в Швейцарии.

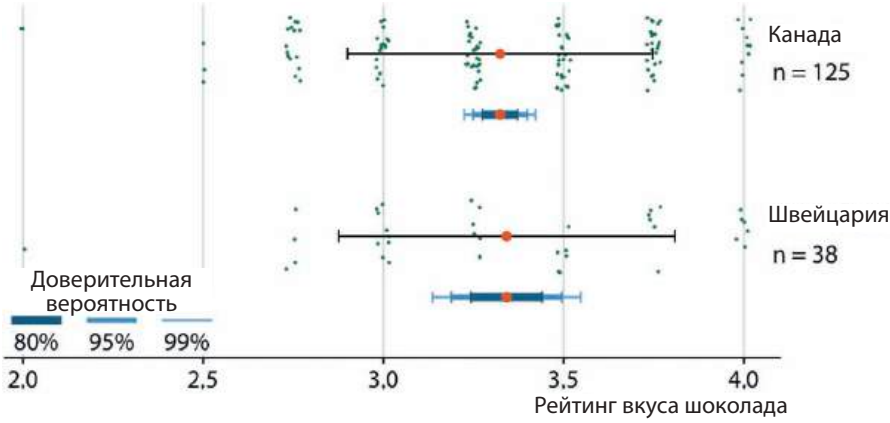


Рис. 15.6. Доверительные интервалы расширяются в соответствии с уменьшением размера выборки. Шоколадные батончики из Канады и Швейцарии имеют сравнимые средние оценки и сопоставимые стандартные отклонения (обозначены простыми черными планками погрешностей). Тем не менее количество канадских батончиков более чем в три раза превышает количество швейцарских, и поэтому доверительные интервалы (обозначенные планками погрешностей разных цветов и толщины, нанесенными друг на друга) значительно шире для среднего значения рейтинга швейцарских батончиков. Источник: Brady Brelinski, Manhattan Chocolate Society

На рис. 15.6 одновременно показаны три различных доверительных интервала посредством использования более темных цветов и более толстых линий для указания доверительных интервалов с более низкими доверительными вероятностями. Я называю этот способ визуализации *градуированными планками погрешностей*. Градуирование помогает читателю понять, что на графике отображены несколько возможных интервалов. Если группе людей показать простые планки погрешностей (без градуировки), то вполне возможно, что как минимум часть воспримет эти планки как отдельные элементы, например, представляющие минимальное и максимальное значения данных. Или же читатели могут подумать, что планки погрешностей обозначают диапазоны всех возможных оценок параметров, то есть что значение оценки не может находиться вне этих границ. Такая ошибка в интерпретации

называется *детерминистической ошибкой восприятия (deterministic construal error)**. Чем больше нам удастся снизить риск возникновения такой ошибки, тем лучше будет наша визуализация неопределенности.

Планки погрешностей удобны тем, что они позволяют нам одновременно показывать множество оценок с сопутствующими им неопределенностями. Именно поэтому данный способ визуализации часто используется в научных публикациях, где основной целью, как правило, является передача большого объема информации, предназначенной для экспертной аудитории. Рассмотрим в качестве примера рис. 15.7, на котором показаны средние оценки с доверительными интервалами для шоколадных батончиков, изготовленных в шести разных странах.

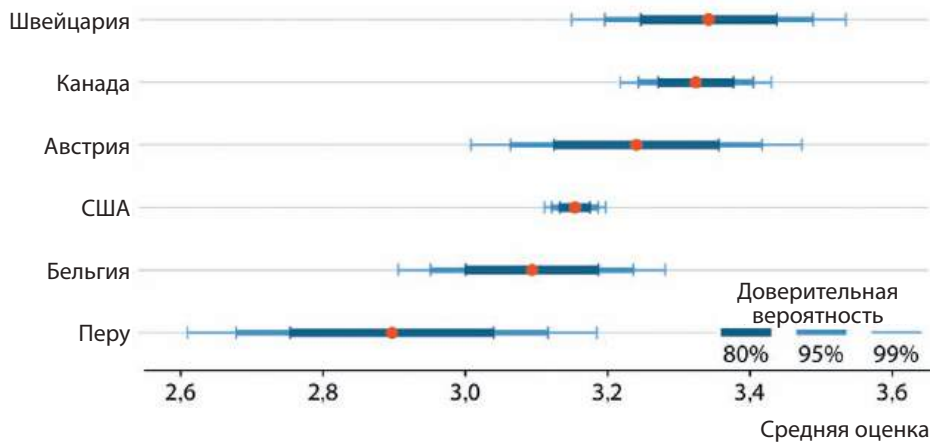


Рис. 15.7. Средние оценки шоколадных батончиков, произведенных в шести разных странах, с соответствующими доверительными интервалами. Источник: Brady Brelinski, Manhattan Chocolate Society

Взглянув на этот рисунок, вы можете задаться вопросом: «А что нам известно о средних значениях рейтинга разных батончиков?» Средние оценки канадских, швейцарских и австрийских экземпляров выше, чем у их американских собратьев. Однако являются ли различия значимыми, если учесть степень неопределенности средних оценок батончиков? Слово «значимый» здесь является техническим термином, используемым статистиками. Разница называется значимой, если мы можем с определенной доверительной вероятностью отвергнуть предположение, что наблюдаемая разница вызвана случайностью. Поскольку оценке подверглось лишь некоторое ограниченное число канадских и американских батончиков, могло быть так, что по чистой случайности

* Говорят, что такая ошибка происходит, когда человек, изучающий график, не осознает, что на графике изображена неопределенность, и вместо этого считает, что видит отображение неких количественных данных. — *Прим. ред.*

эксперты попробовали больше видов хороших канадских шоколадок и меньше хороших американских. Подобного рода случайность на графике может выглядеть как системное преимущество канадских батончиков над американскими.

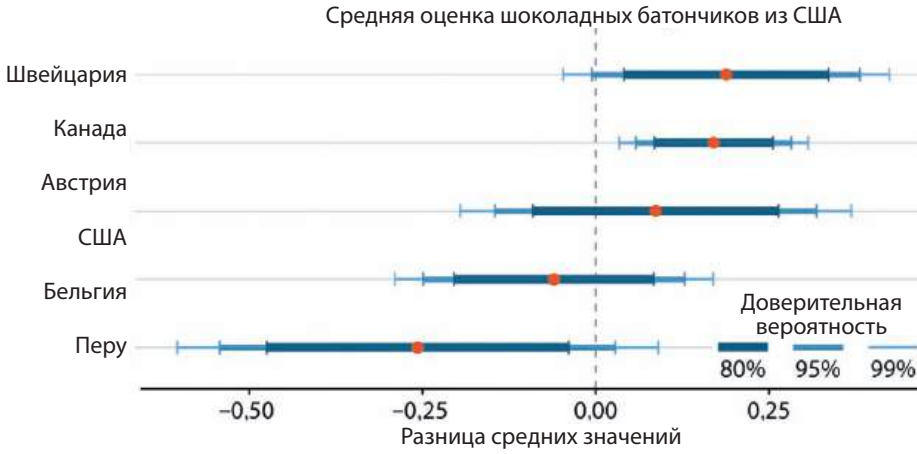


Рис. 15.8. Средние рейтинги шоколадных батончиков, произведенных в пяти разных странах, по отношению к среднему рейтингу американских шоколадных батончиков. Рейтинг канадских шоколадных батончиков значимо выше, чем американских. В среднем рейтинге батончиков, произведенных в остальных четырех странах, нет значимой разницы по сравнению с батончиками из США при доверительной вероятности 95%. Доверительные интервалы были скорректированы для множественных сравнений с использованием критерия Даннета. Источник: Brady Brelinski, Manhattan Chocolate Society

Оценить степень значимости, глядя на рис. 15.7, сложно, поскольку средние рейтинги Канады и США обладают неопределенностью. И та и другая неопределенность имеет значение, когда мы ищем ответ на вопрос, различаются ли средние. В учебниках по статистике или на онлайн-ресурсах, посвященных статистике, иногда встречается набор правил эмпирического характера, посвященных тому, как можно оценить значимость по степени наложения планок погрешностей друг на друга. Обращаю ваше внимание на то, что следует избегать использования этих правил ввиду их ненадежности. Правильным способом оценки различий в среднем рейтинге является расчет доверительных интервалов различий. Если окажется, что они не равны нулю, значит, разница является существенной с соответствующей доверительной вероятностью. Что касается набора данных о шоколадных батончиках, то мы видим, что только канадские сладости значимо лучше, чем американские (рис. 15.8). В свою очередь, 95%-ный доверительный интервал швейцарских батончиков лишь слегка заходит за нулевое значение разницы. Таким образом, вероятность того, что разница между средними оценками шоколадных батончиков из США и Швейцарии является не более чем статистической погрешностью,

составляет всего лишь чуть более 5%. Также нет никаких доказательств того, что австрийские батончики систематически получают более высокие средние оценки, чем американские.

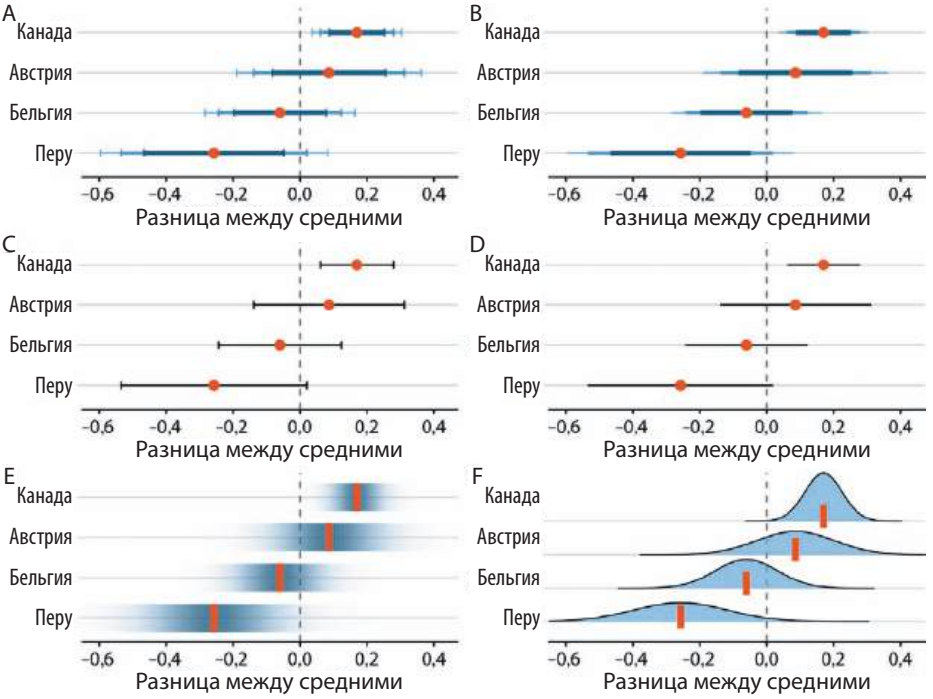


Рис. 15.9. Средние рейтинги шоколадных батончиков, произведенных в пяти разных странах, по отношению к среднему рейтингу американских шоколадных батончиков. Каждое изображение использует свой стиль для отображения одной и той же информации о неопределенности: A) градуированные планки погрешностей с засечками на концах; B) градуированные планки погрешностей без засечек; C) одиночные планки погрешностей с засечками; D) одиночные планки погрешностей без засечек; E) полосы значимости*; F) распределения значимости**. Источник: Brady Brelinski, Manhattan Chocolate Society

На предыдущих рисунках я использовал два различных типа планок погрешностей: градуированные и простые. Существуют и другие варианты. Например, мы можем нарисовать планки с засечками на концах (рис. 15.9A и 15.9C) или без них (рис. 15.9B и 15.9D). У каждого из этих вариантов есть

* Здесь так назван график, на котором отложены планки погрешностей для всех возможных доверительных вероятностей. Цвет полосы в точке указывает на долю доверительных интервалов, которые покрывают эту точку. Среднее значение по определению покрывается всеми доверительными интервалами. — *Прим. ред.*

** Здесь так названа функция распределения, которая описывает распределение доверительных интервалов для всех возможных доверительных вероятностей в пространстве значений параметра. — *Прим. ред.*

свои преимущества и недостатки. Градуированные планки погрешностей указывают на наличие разных диапазонов, соответствующих разным доверительным вероятностям. С другой стороны, чем больше на графике присутствует информации, тем тяжелее его воспринимать. В зависимости от того, насколько сложным и нагруженным информацией является изображение, простые планки погрешностей могут оказаться более предпочтительным вариантом. Наличие засечек на концах планок — вопрос вкуса. Смысл их использования в том, чтобы показать, где именно заканчивается планка погрешностей (см. рис. 15.9A и 15.9C), в то время как планки без засечек делают акцент на всем диапазоне интервала (см. рис. 15.9B и 15.9D). Кроме того, засечки на концах планок тоже могут быть визуальным шумом, поэтому на рисунке с большим количеством планок следует воздержаться от добавления к ним засечек.

Еще одной альтернативой планкам погрешностей являются постепенно «исчезающие» полосы значимости (рис. 15.9E). Они более точно передают возможные значения доверительных интервалов, нежели градуированные планки погрешностей, но они трудны для восприятия. Чтобы определить, где заканчивается тот или иной доверительный интервал, нам придется сравнивать различные оттенки цвета. Глядя на рис. 15.9E, можно сделать вывод, что средняя оценка перуанских шоколадных плиток значительно ниже, чем американских, однако это впечатление ошибочно. Подобные проблемы возникают и при отображении распределений значимости (рис. 15.9F). Визуально сложно оценить область под кривой и определить, где именно достигается желаемая доверительная вероятность. Однако эта проблема может быть решена с помощью точечного графика квантилей, показанного на рис. 15.3.

Для простых двухмерных рисунков планки погрешностей обладают одним важным преимуществом перед более сложными отображениями неопределенности: их можно комбинировать со многими другими типами графиков. При создании практически любого типа визуализации мы можем изобразить неопределенность с помощью планок погрешностей. Например, можно обозначить неопределенность количественного признака с помощью планок погрешностей на столбчатой диаграмме (рис. 15.10). Такой тип визуализации обычно используется в научных публикациях. Мы также можем изобразить планки погрешностей вдоль осей x и y на диаграмме рассеяния (рис. 15.11).

Чуть раньше в этом разделе я говорил о приверженцах частотного распределения и так называемых байесовцах. Первые оценивают неопределенность, используя доверительные интервалы, тогда как вторые работают с апостериорными распределениями и байесовскими доверительными интервалами. Байесовское апостериорное распределение говорит нам, насколько вероятны те или иные оценки параметров с учетом исходных данных. Байесовский доверительный интервал указывает диапазон значений, в которых значение параметра ожидается с заданной вероятностью, рассчитанной по

апостериорному распределению. Например, байесовский 95%-ный доверительный интервал соответствует центральным 95% апостериорного распределения. Истинное значение параметра с вероятностью 95% лежит в 95%-ном доверительном интервале.

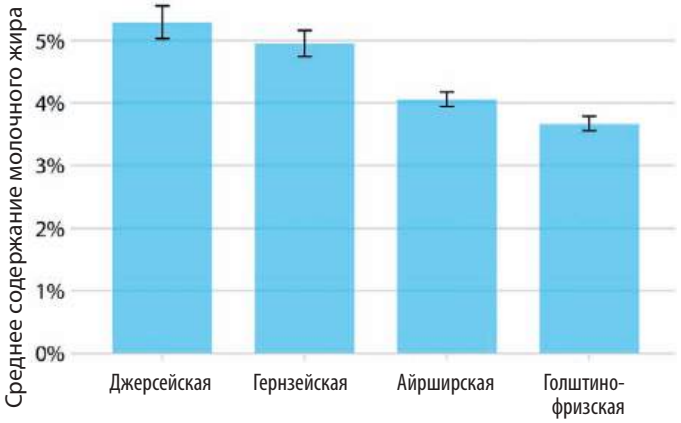


Рис. 15.10. Среднее процентное содержание жира в молоке, производимом четырьмя различными породами коров. Планки погрешностей отображают двусторонний диапазон стандартной ошибки. Визуализации подобного рода часто встречаются в научной литературе. Несмотря на их техническую правильность, они не показывают ни различия внутри каждой категории, ни неопределенность выборки. См. рис. 6.11, где показаны различия в содержании жира в молоке коров различных пород. Источник: Canadian Record of Performance for Purebred Dairy Cattle

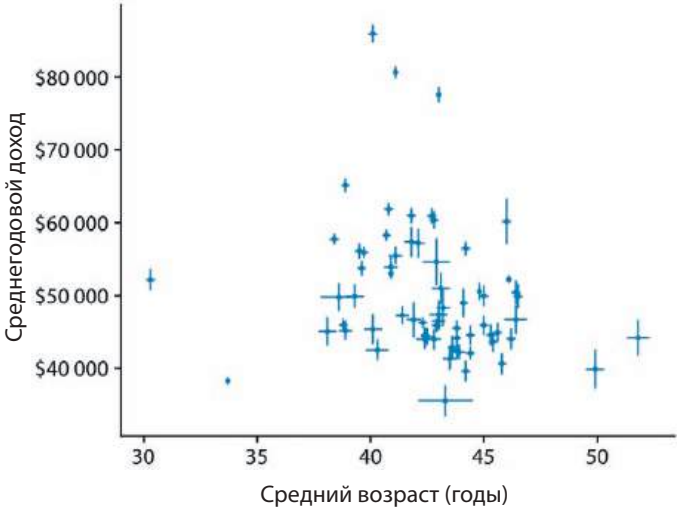


Рис. 15.11. Показатели среднего дохода жителей 67 округов штата Пенсильвания и средний возраст жителей округов. Планки погрешностей представляют собой доверительные интервалы для доверительной вероятности 90%. Источник: 2015 Five-Year American Community Survey

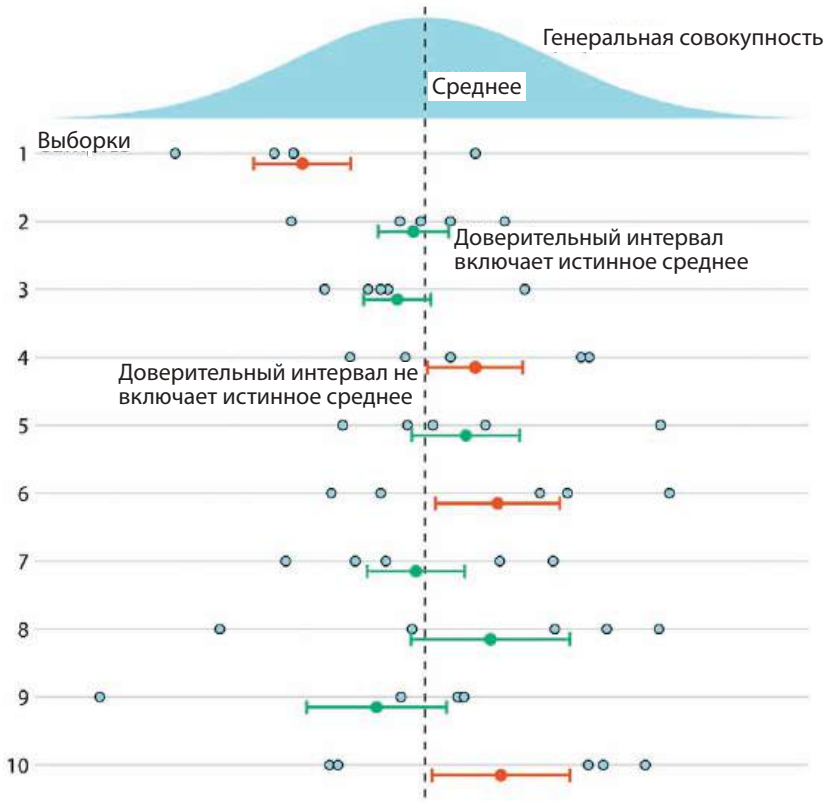


Рис. 15.12. Доверительный интервал в частотном представлении. Доверительные интервалы лучше всего понимаются в контексте повторных выборок. Для каждой выборки заданный доверительный интервал либо включает (зеленый), либо исключает (оранжевый) истинное значение параметра, которое здесь принято за среднее значение. Однако если мы проводим выборку несколько раз, то доверительные интервалы (здесь показаны доверительные интервалы 68%, соответствующие среднему значению выборки +/– стандартная ошибка) включают истинное среднее значение примерно в 68% случаев

Если вы не являетесь специалистом в статистике, вас может удивить мое определение байесовского доверительного интервала. Вы с легкостью могли подумать, что это было определением классического доверительного интервала, однако это не так. Байесовский доверительный интервал сообщает нам, где, вероятнее всего, находится истинный параметр, а частотный доверительный интервал говорит о том, где истинный параметр, скорее всего, находиться не будет. Хотя это различие может показаться простым формализмом, между этими двумя подходами существуют важные концептуальные различия. Байесовский подход означает использование как данных, так и своих априорных знаний об исследуемой системе для вычисления распределения вероятностей (апостериорного), которое показывает, где может находиться истинное

значение параметра. Напротив, при частотном подходе вы сначала делаете предположение, которое в дальнейшем будете пытаться опровергнуть. Это предположение называется «нулевой гипотезой», и зачастую это просто предположение о том, что параметр равен нулю (например, некоторые два условия равнозначны). Затем вы рассчитываете вероятность того, что в случайной выборке будут получены данные, аналогичные тем, которые бы наблюдались в том случае, если бы нулевая гипотеза была верной. Доверительный интервал является представлением этой вероятности. Если некоторый доверительный интервал исключает значения параметра, соответствующие нулевой гипотезе (то есть в нашем примере — нулевое значение), то вы можете отвергнуть нулевую гипотезу с соответствующей ей доверительной вероятностью. В качестве альтернативы вы можете рассматривать классический доверительный интервал как интервал, который захватывает истинное значение параметра с заданной вероятностью при повторной выборке (рис. 15.12). Таким образом, если истинное значение параметра равно нулю, доверительный интервал 95% будет исключать ноль только в 5% проанализированных выборок.

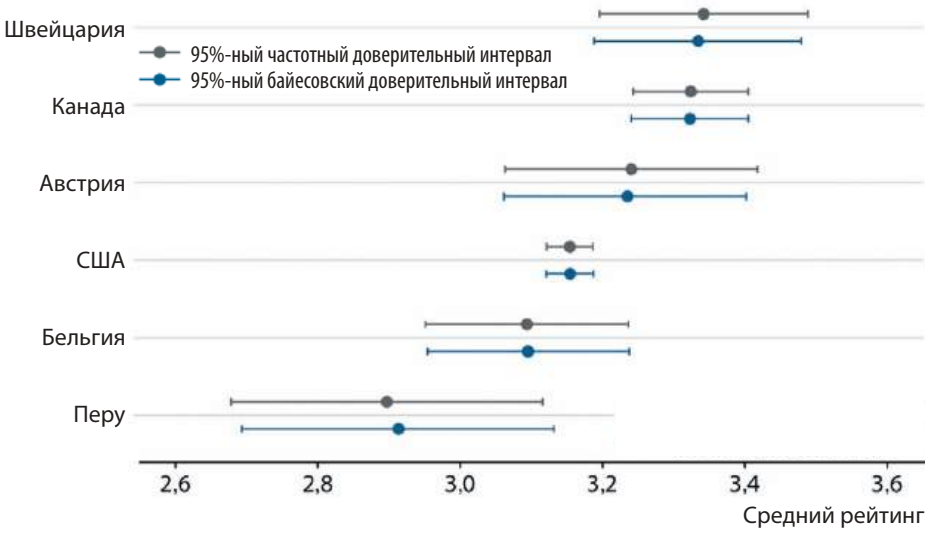


Рис. 15.13. Сравнение частотного и байесовского доверительных интервалов для рейтинга шоколадных батончиков. Мы видим, что оба подхода дают схожие, однако не совсем идентичные результаты. В частности, байесовские оценки показывают незначительный сдвиг и сужение интервалов, что является «корректировкой» наиболее экстремальных оценок параметров по отношению к среднему. (Обратите внимание, что байесовская оценка для Швейцарии слегка смещена влево, а для Перу слегка смещена вправо по отношению к соответствующим интервалам частотного подхода.) Показанные здесь оценки и доверительные интервалы частотного подхода идентичны результатам доверительного интервала 95%, показанного на рис. 15.7. Источник: Brady Brelinski, Manhattan Chocolate Society

Подводя итог, можно сказать, что байесовский доверительный интервал делает утверждение об истинном значении параметра, а частотный доверительный интервал делает утверждение о нулевой гипотезе. На практике, однако, байесовские и частотные оценки часто находятся очень близко (рис. 15.13). Концептуальное преимущество байесовского подхода заключается в том, что он подчеркивает выводы о величине некоего эффекта, в то время как частотный подход фокусируется на том, чтобы сделать вывод, существует эффект или нет.

Байесовский доверительный интервал отвечает на вопрос: «Где мы ожидаем увидеть истинное значение параметра?» Частотный доверительный интервал отвечает на вопрос: «Насколько мы уверены, что истинное значение параметра не равно нулю?»

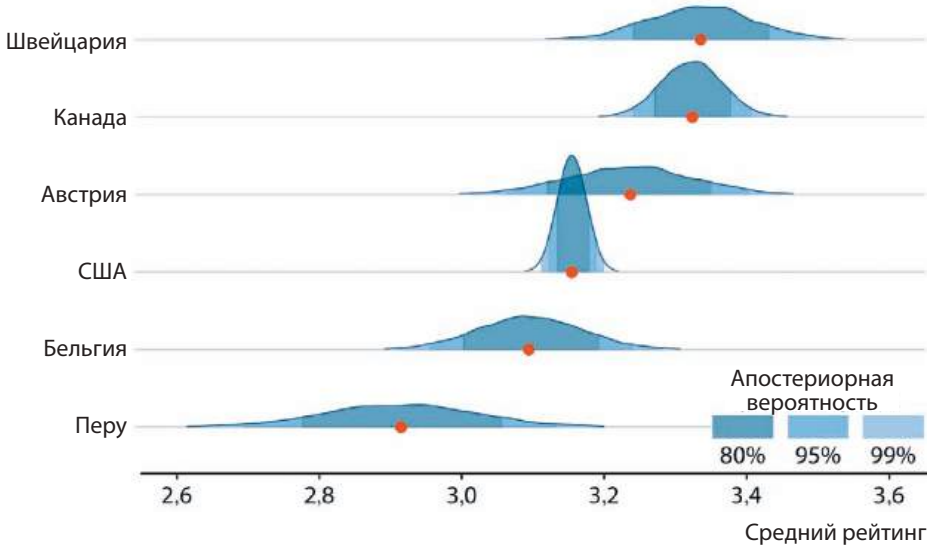


Рис. 15.14. Апостериорные распределения Байеса, визуализированные в виде графика «горный хребет». Красными точками указаны медианы каждого апостериорного распределения. Поскольку преобразовать непрерывное распределение в определенные доверительные области на глаз весьма непросто, я добавил затенение под кривыми, чтобы указать центральные апостериорные распределения для байесовских доверительных вероятностей 80, 95 и 99%. Источник: Brady Brelinski, Manhattan Chocolate Society

Основной целью байесовской оценки является получение апостериорного распределения. Поэтому байесовцы обычно визуализируют все распределение, а не упрощают его до доверительного интервала. Таким образом, мы можем визуализировать данные с помощью любого из подходов к визуализации

распределений, которые мы обсуждали в главах 6–8. В частности, для визуализации байесовского апостериорного распределения обычно используются гистограммы, графики плотности, коробчатые диаграммы, скрипичные графики и графики типа «горный хребет». Поскольку все эти способы были детально рассмотрены в соответствующих главах, я воспользуюсь только одним из них — графиком «горный хребет», — чтобы показать байесовские апостериорные распределения средних оценок шоколада (рис. 15.14). В данном случае я затемнил область под кривой различными оттенками цвета, чтобы обозначить определенные области апостериорных вероятностей. В качестве альтернативы затенению я мог бы также нарисовать точечные диаграммы квантилей или добавить градуированные планки погрешностей под каждым распределением. График «горный хребет» с расположенными внизу планками погрешностей называется «полуглаза» (half-eyes), а повернутые на 90° скрипичные графики с планками погрешностей называются «глаза» (eye plot) (см. раздел «Неопределенность на диаграммах» на с. 54).

Визуализация неопределенности подгонки кривых

В главе 13 мы обсуждали, как показать тренд в наборе данных путем подгонки прямой или кривой линии к данным. При оценке трендов мы вновь сталкиваемся с неопределенностью и обычно показываем ее в виде линии тренда с доверительной полосой — confidence band (рис. 15.15). Последняя сочетает в себе множество различных линий, которые будут соответствовать исходным данным. Когда студенты впервые сталкиваются с этим явлением, они часто удивляются, что даже идеально прямой график линейной аппроксимации создает изогнутую полосу уверенности. Причина кривизны заключается в том, что прямая линия может двигаться в двух разных направлениях: вверх и вниз (разные значения константы в уравнении прямой) и может вращаться (разные значения углового коэффициента в уравнении прямой). Чтобы наглядно показать, как получается доверительная полоса, мы можем нарисовать набор альтернативных подогнанных линий, случайно сгенерированных на основе апостериорного распределения параметров*. Рис. 15.16 иллюстрирует подобный подход с помощью 15 случайно выбранных альтернативных прямых. Несмотря на то что каждая линия является прямой, комбинация различных угловых коэффициентов и констант создает общую форму, выглядящую как доверительная полоса.

---

* Здесь для построения случайных линий мы после построения подогнанной прямой берем случайные значения углового коэффициента и константы в пределах их доверительных интервалов. — *Прим. науч. ред.*

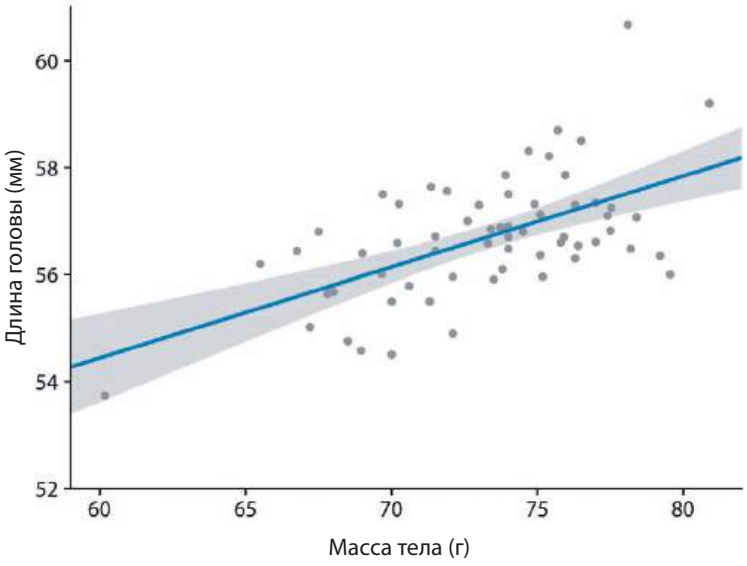


Рис. 15.15. Отношение длины головы к массе тела для самцов голубых соек, показанное на рис. 13.7. Прямая линия синего цвета обозначает наилучшую аппроксимацию данных прямой, а серая полоса вокруг показывает неопределенность линейной регрессии. Серая полоса нарисована для доверительной вероятности 95%. Источник: Keith Tarvin, Oberlin College

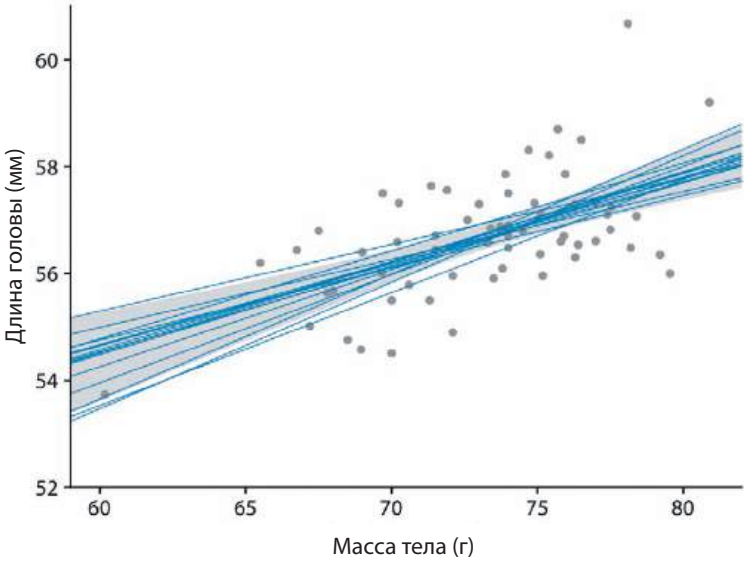


Рис. 15.16. Отношение длины головы к массе тела для самцов голубых соек. В отличие от рис. 15.15, здесь прямые линии синего цвета обозначают одинаково вероятные альтернативные линейные аппроксимации, построенные произвольно на основании апостериорного распределения значений параметров. Источник: Keith Tarvin, Oberlin College

Чтобы нарисовать доверительную полосу, нам нужно указать доверительную вероятность, и, как мы уже убедились при изучении планок погрешности и апостериорных вероятностей, может быть полезно выделить разные доверительные вероятности. Это приводит нас к созданию *градуированной доверительной полосы*, которая отображает несколько доверительных вероятностей одновременно (рис. 15.17). Градуированная доверительная полоса усиливает у читателя чувство неуверенности, поскольку ставит его перед фактом, что данные могут соответствовать нескольким различным альтернативным линиям тренда.

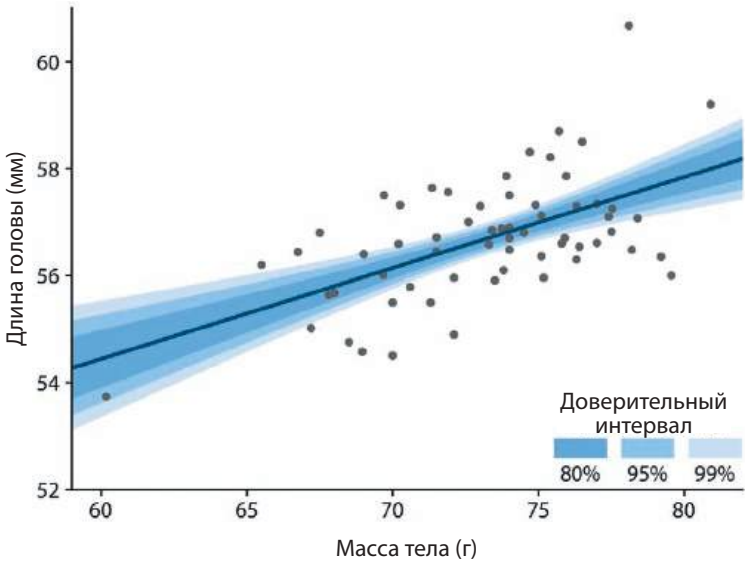


Рис. 15.17. Отношение длины головы к массе тела для самцов голубых соек. Как и в случае с планками погрешности, мы можем нарисовать градуированные доверительные полосы, чтобы акцентировать внимание читателя на наличии неопределенности. Источник: Keith Tarvin, Oberlin College

Мы также можем нарисовать доверительные полосы и для случая аппроксимации данных нелинейными кривыми. Такие доверительные полосы выглядят красиво, однако могут быть сложны в интерпретации (рис. 15.18). Глядя на рис. 15.18A, легко прийти к выводу, что доверительная полоса возникает при перемещении синей линии вверх и вниз и, возможно, ее незначительной деформации. Однако, как следует из графика на рис. 15.18B, доверительный интервал представляет собой семейство кривых, которые могут быть значительно более изогнутыми по сравнению с общей линией подгонки, показанной в части A. Таков общий принцип процесса приближения с помощью нелинейных кривых: неопределенность соответствует не только движению кривой вверх и вниз, но и увеличению степени кривизны графика.

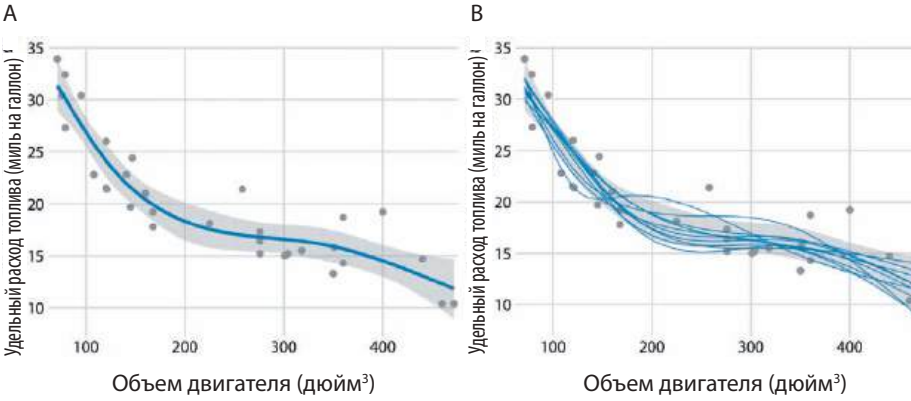


Рис. 15.18. Отношение удельного расхода топлива к объему двигателя для 32 различных автомобилей (1973–1974 годов выпуска). Каждая точка обозначает одну машину, а плавные линии были получены путем построения регрессии кубическим сплайном с пятью узлами. A. Лучшая аппроксимация сплайном и доверительная полоса. B. Одинаково вероятные альтернативные варианты подгонки, построенные на основании апостериорного распределения значений параметров. Источник: Motor Trend, 1974

Диаграммы гипотетических исходов

Общей проблемой всех визуализаций статистической неопределенности является то, что зрители могут интерпретировать некоторые аспекты визуализации неопределенности просто как особенность данных (как уже говорилось ранее, подобная ошибка называется детерминистической ошибкой восприятия). Мы можем избежать этой проблемы, визуализируя неопределенность с помощью анимации, циклически перебирая варианты различных, но одинаково вероятных графиков. Этот вид визуализации называется *диаграммами гипотетических исходов* (HOP, Hypothetical Outcome Plot) [Hullman, Resnick, Adar, 2015]. И хотя HOP нельзя использовать в печатных изданиях, у них есть своя ниша — онлайн-формат, например, в виде GIF-файлов или видео MP4. Кроме того, HOP может быть хорошим визуальным сопровождением для устной презентации.

Чтобы разобраться в сути концепции HOP, давайте еще раз вернемся к рейтингам шоколадных батончиков. Когда вы стоите в продуктовом магазине и думаете, какую шоколадку выбрать, вас, скорее всего, не очень волнует средняя оценка вкуса и связанная с этим неопределенность для конкретных групп шоколадных батончиков. Вместо этого, вероятно, вас больше будет интересовать ответ на такой простой вопрос, как: «Если я наугад возьму по одной шоколадной плитке канадского и американского производства, какая будет вкуснее?» Чтобы ответить на этот вопрос, мы могли бы случайным образом выбрать из набора данных пару батончиков, один из Канады и один из США, сравнить их рейтинги, записать результат, а затем повторить этот процесс много раз. Если

бы мы так и поступили, то обнаружили бы, что примерно в 53% случаев канадский батончик имеет более высокие показатели, а в оставшихся 47% случаев более вкусными (или на том же уровне) окажутся американские батончики. Мы можем визуализировать этот процесс, циклически переключаясь между некоторыми из выбранных пар и показывая относительные расположения их оценок друг относительно друга в виде двух засечек (рис. 15.19).

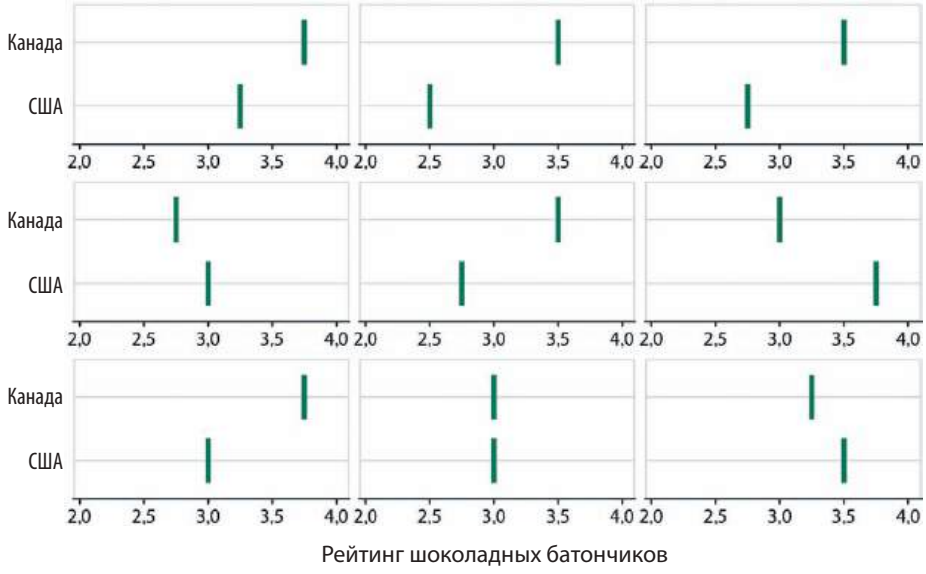


Рис. 15.19. Схема графика HOP рейтинга шоколадных батончиков, произведенных в Канаде и США. Зеленой засечкой обозначается рейтинг одного батончика. Прямоугольники показывают сравнение двух батончиков, выбранных произвольно (один батончик из США, один батончик из Канады). В реальном HOP изображение будет циклически переключаться между прямоугольниками, а не показывать их рядом друг с другом. Источник: Brady Brelinski, Manhattan Chocolate Society

В качестве второго примера рассмотрим разницу в форме среди одинаково вероятных линий тренда с рис. 15.18B. Поскольку все линии тренда нанесены друг на друга, мы прежде всего обращаем внимание на общую область, охваченную этими линиями, которая эквивалентна доверительной полосе. Очевидно, что восприятие отдельных линий тенденций затруднено. Чтобы решить эту проблему, нам нужно превратить рис. 15.18B в HOP (рис. 15.20).

В процессе создания HOP вы можете столкнуться с дилеммой: каким сделать переключение между различными исходами — жестким (как при просмотре слайдов на проекторе) или, наоборот, плавно анимированным (в виде, например, постепенной деформации линии тренда, превращения ее из одного варианта исхода в другой)? Несмотря на то что у сообщества нет однозначного ответа на этот вопрос и его поиски продолжаются, некоторые данные говорят

о том, что плавные переходы затрудняют оценку представленных вероятных кривых [Kale et al., 2018]. Поэтому, если вы планируете использовать плавную анимацию перехода между трендами, рекомендую сделать ее как можно более короткой или выбрать такой стиль, при котором изображения постепенно исчезают и появляются, а не перетекают из одного в другое.

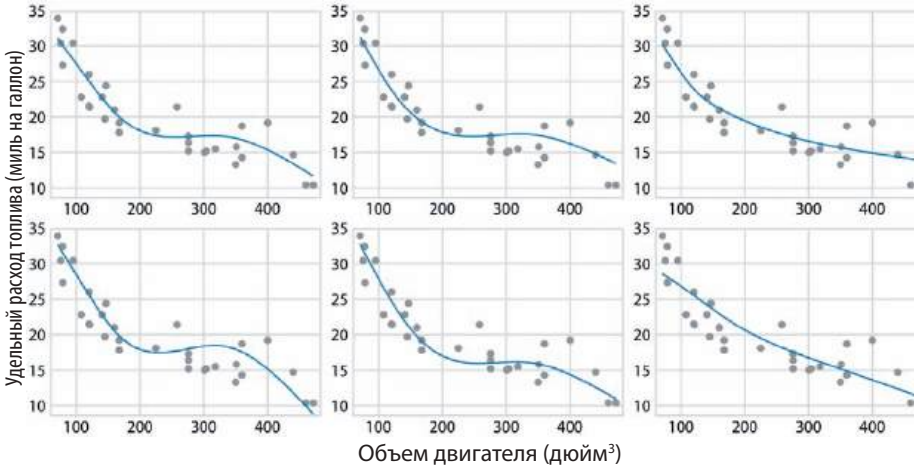


Рис. 15.20. Схематичное изображение HOP зависимости удельного расхода топлива от объема двигателя. Каждая точка представляет одну машину, а плавные линии получены путем подгонки кубического сплайна с пятью узлами. Каждая линия в прямоугольнике представляет один альтернативный результат аппроксимации, взятый из апостериорного распределения параметров подгонки. В реальном HOP будет происходить циклический переход между прямоугольниками, а не показ их рядом друг с другом

При подготовке HOP очень важно не упустить из виду один критически важный момент: результаты, которые мы показываем, должны отражать истинное распределение возможных результатов. В противном случае мы можем ввести читателя в заблуждение. Давайте снова вернемся к нашему исследованию рейтинга шоколадных батончиков. Если бы я случайным образом выбрал 10 пар результатов ранжирования шоколадных батончиков, среди которых в семи случаях у американских батончиков была бы более высокая оценка, чем у канадских, то я бы создал у читателя ошибочное мнение, что американские батончики имеют рейтинг выше канадских. Избежать ошибки можно двумя способами: либо набрать гораздо больше результатов, чтобы свести к минимуму вероятность отклонения выборки, либо каким-то образом удостовериться, что наши демонстрируемые результаты близки к истинным. При создании рис. 15.19 я убедился, что число прямоугольников, в которых значение рейтинга канадских батончиков выше, близко к истинному значению в 53%.

Часть II

Принципы дизайна визуализаций

Глава 16

Принцип пропорциональной заливки

Во многих различных сценариях визуализации мы представляем значения данных в виде каких-либо графических элементов, при этом значения еще и определяют характеристики этих элементов. Например, на столбчатой диаграмме индивидуальные столбцы начинаются в нуле и заканчиваются в представляемом значении данных. В этом случае значение данных отражается не только в координате верхней точки столбца, но и в его высоте или длине. Если бы столбец начинался со значения, отличного от 0, то его длина и конечная точка передавали бы противоречивую информацию. Подобного рода изображения содержат внутреннюю рассогласованность, потому что один и тот же графический элемент фактически соответствует двум разным значениям. Сравните эту ситуацию со сценарием, в котором мы визуализируем значение данных с помощью точки. В этом случае значение данных содержится только в месте положения точки, а не в ее размере или форме.

Подобные проблемы будут возникать всякий раз при использовании таких графических элементов, как столбцы, прямоугольники, закрашенные области произвольной формы или любые другие элементы, которые обладают четкими визуальными границами. Последние, в свою очередь, могут как соответствовать указанному значению данных, так и противоречить ему. Всегда проверяйте, что график не содержит подобных противоречий. Данная концепция носит название *принципа пропорциональной заливки (principle of proportional ink)* [Bergstrom and West, 2016].

Если закрашенная область представляет числовое значение, площадь этой закрашенной области должна быть прямо пропорциональна соответствующему значению.

На практике, особенно в массовой прессе или в мире финансов, этот принцип очень часто нарушается.

Визуализации на линейных шкалах

Для начала давайте рассмотрим наиболее распространенный сценарий — визуализацию количественных данных на линейной шкале. На рис. 16.1 показан

среднегодовой доход в пяти округах, которые все вместе составляют штат Гавайи. Этот график является типичным примером изображения, которое можно встретить в газетной статье. Бегло взглянув на рисунок, можно подумать, что округ Гавайи невероятно беден, в то время как Гонолулу гораздо богаче своих соседей по штату.

Тем не менее это впечатление ошибочно, потому что все столбцы на графике начинаются со значения 50 000 долларов. Таким образом, хоть конечная точка каждого столбца и верно отображает фактический средний доход в каждом округе, высота столбца показывает степень, в которой средний доход превышает стартовое значение столбцов — 50 000 долларов США, произвольное число. Человеческое восприятие склонно толковать изображение, выбирая ключевой характеристикой высоту столбца, а не положение его конечной точки на оси *y*.

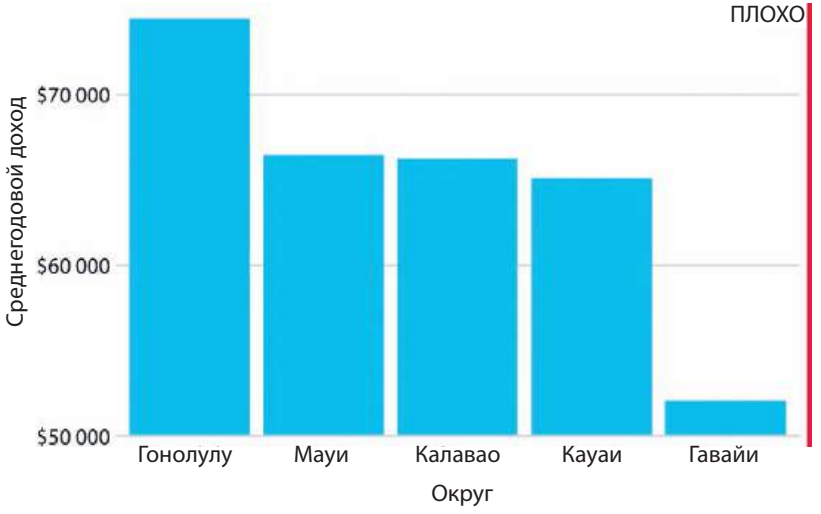


Рис. 16.1. Среднегодовой доход жителей в пяти округах штата Гавайи. Данное изображение вводит в заблуждение, потому что шкала оси *y* начинается с 50 000 долларов вместо 0. В результате высота столбцов оказывается непропорциональна показанным значениям, а разница в доходах между округом Гавайи и остальными округами штата выглядит намного большей, чем есть на самом деле. Источник: 2015 Five-Year American Community Survey

Адекватная данным визуализация выглядит, конечно, менее сенсационно (рис. 16.2). Несмотря на то что между округами действительно имеются различия, в реальности они не столь велики, как показано на рис. 16.1. Фактически значения среднего дохода округов вполне сопоставимы.

Столбцы на графике с линейной шкалой всегда должны начинаться с отметки 0.

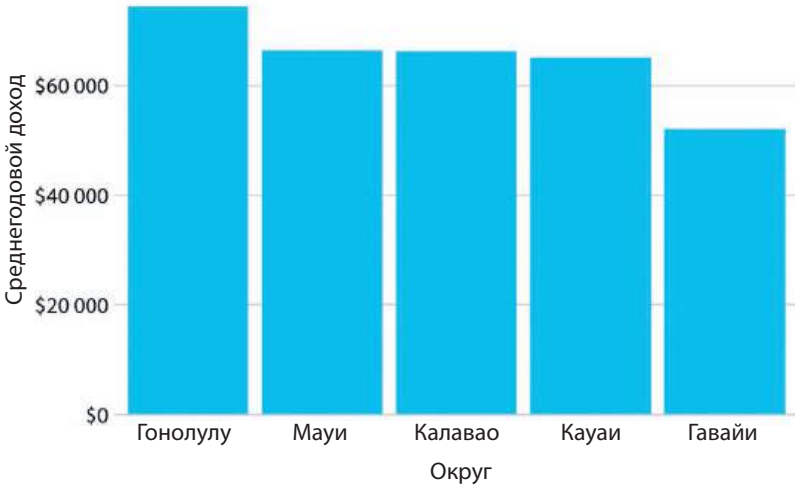


Рис. 16.2. Среднегодовой доход жителей в пяти округах штата Гавайи. На этом графике шкала оси *y* начинается с 0 долларов США, и поэтому значения относительных средних доходов в пяти округах показаны точно. Источник: Data source: 2015 Five-Year American Community Survey

Рис. 16.3. График изменения цены акций Facebook (FB) в период с 22 октября 2016 года по 21 января 2017 года. Из этого рисунка можно сделать кажущийся вывод, что примерно 1 ноября 2016 года цена акций FB сильно упала. Однако это впечатление ошибочно, поскольку ось *y* начинается со 110 долларов вместо 0. Источник: Yahoo! Finance

Подобные проблемы очень часто возникают при визуализации временных рядов, таких как цены на акции. Взглянув на рис. 16.3, можно подумать, что примерно 1 ноября 2016 года произошел мощный обвал курса акций

Facebook*. При этом на самом деле снижение цены по отношению к общей цене акций было умеренным (рис. 16.4). Даже если бы область под кривой не была затенена, диапазон оси *y* на рис. 16.3 все равно выглядел бы сомнительно. Однако с затенением график становится еще более сложным для восприятия: заливка области под кривой подчеркивает расстояния между точкой пересечения осей и конкретными показанными значениями *y*, создавая впечатление, что высота заштрихованной области в какой-либо день означает цену акций в этот день, тогда как в реальности высота области представляет собой изменение в цене акций по сравнению с базовым уровнем, который на рис. 16.3 находится на отметке в 110 долларов.

Рис. 16.4. График изменения цены акций Facebook (FB) в период с 22 октября 2016 года по 21 января 2017 года. Показывая цену акций в масштабе от 0 до 150 долларов США, мы точнее отражаем степень падения цены акций FB, которое наблюдалось примерно 1 ноября 2016 года. Источник: Yahoo! Finance

Примеры на рис. 16.2 и 16.4 могут натолкнуть на мысль, что столбцы и закрашенные области являются неподходящими вариантами для представления небольших изменений во времени или различий между условиями, поскольку нам всегда приходится рисовать столбец или область, начиная с 0. Однако этот вывод неверен. Вполне допустимой практикой является отображение различий между условиями при помощи столбцов и заштрихованных областей, но только в том случае, если вы ясно даете понять читателю, какие именно различия демонстрируются на вашем графике. Например, мы можем использовать столбцы, чтобы визуализировать изменение среднего дохода

* Принадлежит компании META, которая признана экстремистской на территории РФ. — *Прим. ред.*

в округах штата Гавайи с 2010 по 2015 год (рис. 16.5). Для всех округов, кроме Калавао, это изменение составляет менее 5000 долларов США. (Калавао — это особое графство, в нем проживает менее 100 человек, и поэтому для него вполне нормальны большие колебания среднего дохода в связи с небольшим количеством въезжающих и выезжающих людей.)

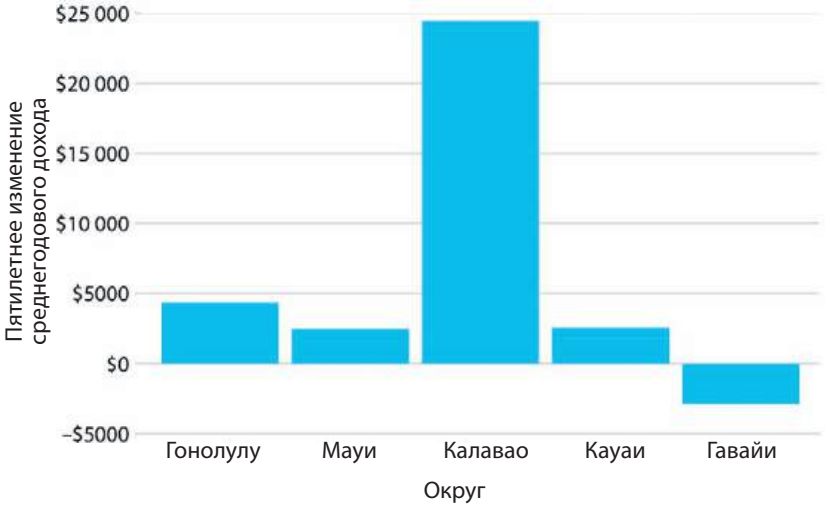


Рис. 16.5. Изменение среднегодового дохода в округах штата Гавайи с 2010 по 2015 год. Источник: 2010 и 2015 Five-Year American Community Surveys

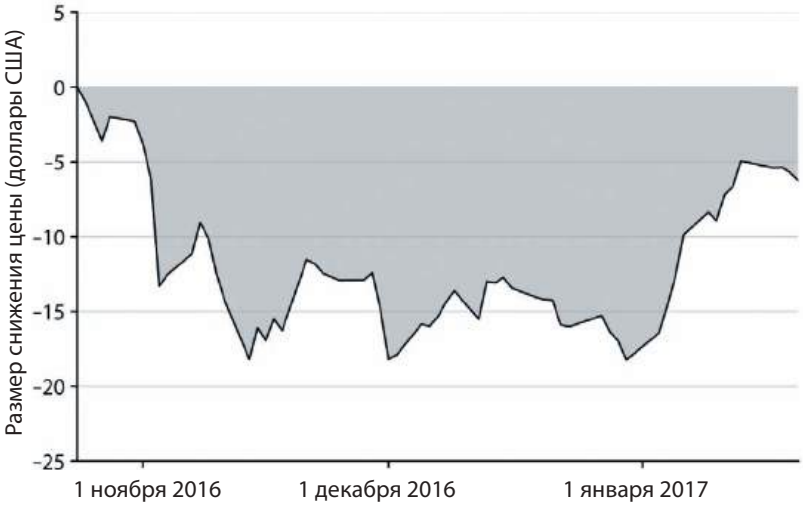


Рис. 16.6. Снижение цены акций Facebook (FB) по отношению к значению от 22 октября 2016 года. В период с 1 ноября 2016 года по 1 января 2017 года цена была примерно на 15 долларов меньше своего максимального значения, зафиксированного 22 октября 2016 года. С января цена вновь начала расти. Источник: Yahoo! Finance

Для округа Гавайи изменение является отрицательным; то есть средний доход в 2015 году был ниже, чем в 2010-м. Отрицательные значения представлены столбцами, которые идут в противоположном направлении, то есть вниз от нуля, а не вверх.

Аналогичным образом мы можем нарисовать изменение цены акций Facebook с течением времени как разницу от ее максимального значения, которое было достигнуто 22 октября 2016 года (рис. 16.6).

Затеняя область, представляющую собой расстояние от высшей точки, мы точно показываем абсолютную степень падения цены без каких-либо неявных утверждений об отношении величины падения цены акции к ее реальной цене.

Визуализации на логарифмических шкалах

Когда мы строим визуализацию на линейной шкале, то площади столбцов, прямоугольников или других фигур будут по построению пропорциональны значениям отображаемых данных. Однако это утверждение не работает для логарифмической шкалы, поскольку данные на такой шкале распределены неравномерно. Отсюда следует вывод, что столбчатые диаграммы, построенные на логарифмической шкале, по своей сути будут ущербными. С другой стороны, площадь каждого столбца такой диаграммы останется пропорциональна логарифму значения данных, следовательно, столбчатые диаграммы в логарифмическом масштабе удовлетворяют принципу пропорциональной заливки — в шкале, преобразованной по логарифму. Однако ни один из этих аргументов, по моему мнению, не будет на практике являться обоснованием использования столбчатых диаграмм на логарифмической шкале. В данном случае продуктивнее будет задаться вопросом, что мы хотим визуализировать — значения или доли?

В главе 2 мы узнали, что логарифмическая шкала является наиболее удобным и логичным инструментом для визуализации соотношений, поскольку единичный шаг вдоль логарифмической шкалы соответствует умножению или делению на постоянный коэффициент. Однако на практике логарифмические шкалы используют не специально для визуализации соотношений, а просто потому, что числа на диаграмме могут различаться на несколько порядков. В качестве примера рассмотрим валовой внутренний продукт (ВВП) стран Океании. В 2007 году значение этого показателя варьировалось в пределах от менее одного миллиарда долларов США до более 300 миллиардов долларов США (рис. 16.7). Если эти значения расположить на линейной шкале, мы не получим адекватного отображения данных, потому что весь график будет занят двумя странами с наибольшим значением ВВП (Новая Зеландия и Австралия).

Рис. 16.7. Показатели ВВП стран Океании в 2007 году. Значения данных, выраженные длиной столбцов, являются неточными, поскольку столбцы начинаются со значения в 0,3 миллиарда долларов США. Источник: Gapminder

Рис. 16.8. Показатели ВВП стран Океании в 2007 году. Длины столбцов не являются точным отражением визуализируемых значений данных, поскольку столбцы начинаются с произвольного значения 10^{-9} миллиардов долларов США. Источник: Gapminder

Хочу отметить, что визуализация в виде гистограммы, нанесенной на логарифмическую шкалу (см. рис. 16.7), тоже будет не слишком информативной. Поскольку значения на шкале начинаются с произвольного значения в 0,3 миллиарда долларов США, данный график обладает тем же изъяном, что и рис. 16.1: длины столбцов не являются адекватным отображением визуализируемых данных. Кроме того, процесс создания визуализации на логарифмической шкале осложнен еще и тем, что столбцы не могут начинаться со значения 0. На рис. 16.7 значение 0 будет расположено бесконечно далеко слева, поэтому мы можем нарисовать столбцы любой длины, просто отодвигая их начало все дальше и дальше, как показано на рис. 16.8. Эта проблема неизменно возникает

в тех случаях, когда мы пытаемся расположить количественные данные (чем и является показатель ВВП) на логарифмической шкале.

На мой взгляд, данные, визуализацию которых мы видим на рис. 16.7, не следует отображать в виде столбцов. Отказавшись от последних в пользу точек в соответствующих местах шкалы ВВП стран, мы полностью избавимся от проблем, связанных с длиной столбцов (рис. 16.9). Кроме того, размещая названия стран рядом с точками, а не вдоль оси *y*, мы убираем пустое пространство между названием страны и точкой, которое может быть ошибочно визуально воспринято как некоторая величина.

Рис. 16.9. Показатели ВВП стран Океании в 2007 году. Источник: Gapminder

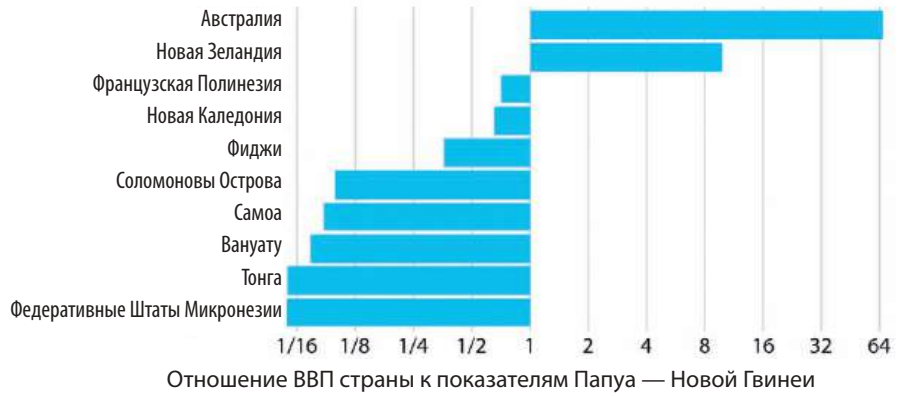


Рис. 16.10. Показатели ВВП в странах Океании в 2007 году по отношению к ВВП Папуа — Новой Гвинеи. Источник: Gapminder

Однако если речь заходит о визуализации соотношений, то столбцы на логарифмической шкале становятся вполне приемлемым решением. На самом деле в данном случае они даже являются предпочтительным вариантом

по отношению к столбцам на линейной шкале. Давайте в качестве примера попробуем визуализировать значения ВВП стран Океании относительно показателя ВВП Папуа — Новой Гвинеи. Получившийся график хорошо отражает основные взаимосвязи между ВВП разных стран (рис. 16.10). Мы видим, что ВВП Новой Зеландии в 8 раз выше, чем аналогичный показатель Папуа — Новой Гвинеи, Австралии — выше в 64 раза, а вот ВВП Тонга и Федеративных Штатов Микронезии составляет всего 1/16 от ВВП Папуа — Новой Гвинеи. Французская Полинезия и Новая Каледония имеют показатель ВВП лишь немногим меньше, чем у Папуа — Новой Гвинеи.

Рисунок 16.10 подчеркивает важный момент: при построении логарифмической шкалы исходной точкой должен быть не 0, а 1. Значения больше 1 должны идти в одну сторону от исходной точки, а значения меньше 1 — в противоположную. Столбцы на логарифмической шкале представляют отношения и поэтому должны начинаться с единицы, тогда как столбцы на линейной шкале отражают количества и, соответственно, должны начинаться с нуля.

Столбцы на логарифмической шкале представляют соотношения, поэтому они должны начинаться с единицы, а не с нуля.

Прямая визуализация площадей

Во всех предыдущих примерах данные визуализировались вдоль одного измерения, поэтому значения данных были отображены как в площади фигуры, так и в расположении фигуры относительно осей координат. Во всех этих случаях мы можем считать, что площадь имеет второстепенное значение в отображении величины по отношению к ее местоположению на координатной сетке. Однако существуют подходы к визуализации, которые представляют данные в первую очередь в виде площади (или вообще только так), не прибегая к каким-либо указаниям насчет местоположения элемента. Наиболее распространенным вариантом такой визуализации является круговая диаграмма (рис. 16.11). Хотя технически значения данных выражены при помощи углов, расположенных на круговой оси, на практике мы обычно не оцениваем углы круговой диаграммы. Куда большее внимание привлекают площади образуемых секторов.

Поскольку площадь каждого сектора окружности пропорциональна его углу, который, в свою очередь, пропорционален значению данных, представленных данным сектором, круговые диаграммы удовлетворяют принципу пропорциональной заливки. Однако мы воспринимаем область на круговой диаграмме иначе, нежели на гистограмме. Дело здесь в том,

что человек склонен обращать внимание на расстояния, а не на площади. Поэтому, если значение данных представлено только в виде расстояния (длины столбца в случае столбчатой диаграммы), мы воспринимаем величину более точно, в отличие от случаев, когда значение данных передается с помощью комбинации двух или более расстояний, которые вместе и составляют площадь.

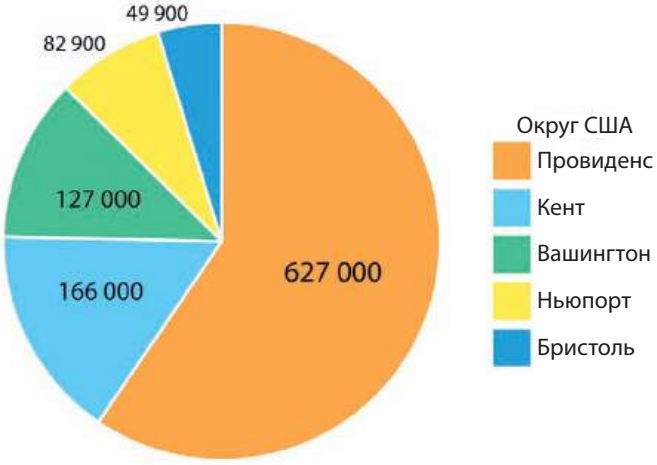


Рис. 16.11. Число жителей в округах Род-Айленда, представленное в виде круговой диаграммы. Как угол, так и площадь каждого сегмента окружности пропорциональны количеству жителей соответствующего округа. Источник данных: 2010 US Decennial Census

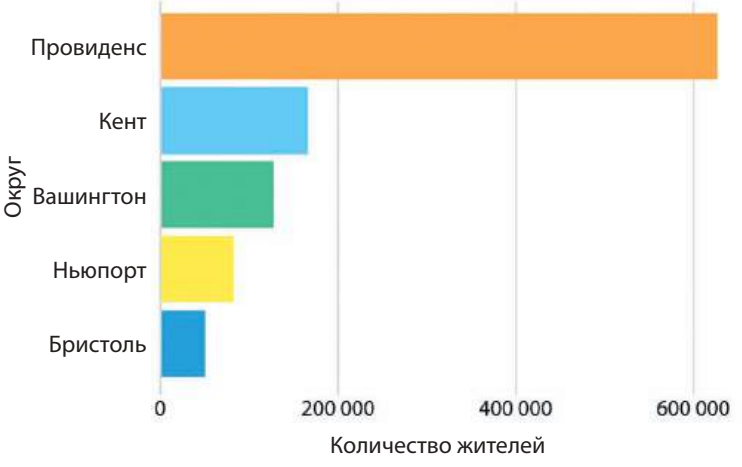


Рис. 16.12. Число жителей в округах Род-Айленда, визуализированное на столбчатой диаграмме. Длина каждого столбца пропорциональна количеству жителей в соответствующем округе. Источник: 2010 US Decennial Census

Сравните рис. 16.11 и 16.12, на которых представлены одни и те же данные. Как можно заметить, разница в количестве жителей между округом Провиденс и другими округами на рис. 16.12 кажется на взгляд больше, чем на рис. 16.11.

Особенность человеческого восприятия, выражающаяся в том, что расстояния мы оцениваем гораздо легче, чем площади, дает о себе знать и в случае древовидных карт (рис. 16.13). Последние фактически являются прямоугольными аналогами круговых диаграмм. И вновь, сравнивая рис. 16.12 с рис. 16.13, мы ощущаем, что на втором графике различия в численности населения округов кажутся нам менее выраженными.

Рис. 16.13. Количество жителей в округах Род-Айленда, визуализированное посредством древовидной карты. Площадь каждого прямоугольника пропорциональна количеству жителей в соответствующем округе.
Источник: 2010 US Decennial Census

Глава 17

Обработка накладывающихся точек

В процессе визуализации больших или очень больших наборов данных мы часто сталкиваемся с тем, что простые диаграммы рассеяния теряют эффективность, поскольку большое количество точек находится близко друг у другу, из-за чего они могут полностью или частично перекрываться. Подобные проблемы иногда возникают даже при визуализации небольших наборов данных, если значения имеют невысокую точность или округлены таким образом, что большое число наблюдений выражено одинаковыми величинами. Данную ситуацию часто описывают техническим термином *оверплоттинг* (или перегрузка графика), который означает, что мы имеем дело со множеством точек, нанесенных, по сути, одна поверх другой. Ниже мы поговорим о том, какие существуют методы решения этой проблемы.

Частичная прозрачность и джиттеринг (jittering)

Для начала рассмотрим сценарий с относительно небольшим количеством точек, которым соответствуют сильно округленные значения. В нашем наборе данных находятся сведения об экономии топлива при движении по городу и объемах двигателей для 234 популярных моделей автомобилей, выпущенных в период с 1999 по 2008 год (рис. 17.1). В качестве единицы измерения расхода топлива используются мили на галлон, значения округляются до ближайшего целого. Объем двигателя измеряется в литрах, значения округляются до ближайшего децилитра. Вследствие столь «размашистых» округлений большое количество моделей автомобилей характеризуются абсолютно одинаковыми значениями. Например, у нас есть 21 автомобиль с рабочим объемом двигателя 2 литра, но в совокупности все эти машины имеют лишь четыре различных значения экономии топлива: 19, 20, 21 или 22 мили на галлон. Таким образом, на рис. 17.1 эти автомобили представлены всего четырьмя различными точками, из-за чего читатель может сделать вывод,

что двигатели объемом 2 литра являются гораздо менее популярными, чем они есть на самом деле. Кроме того, в наборе данных присутствуют сведения о двух полноприводных автомобилях с 2-литровыми двигателями, на графике эти машины обозначены черными точками. Однако эти точки полностью закрыты точками желтого цвета, из-за чего складывается впечатление, что полноприводных автомобилей с 2-литровым двигателем не существует вообще.

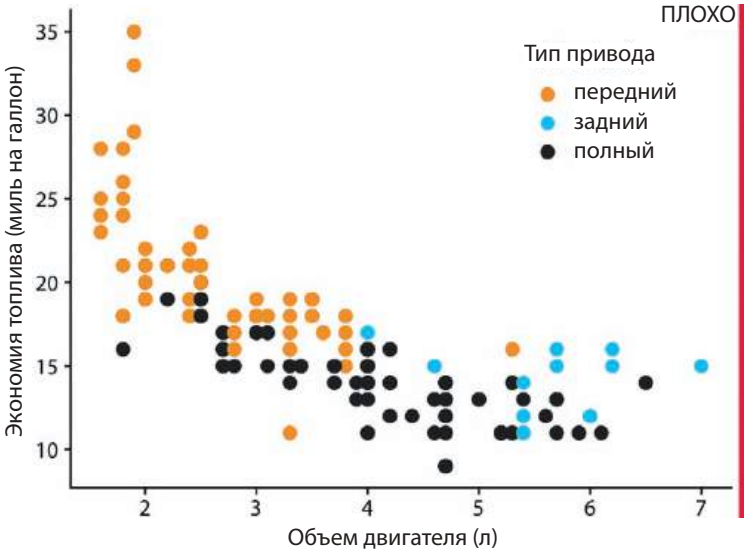


Рис. 17.1. Экономия топлива в городских условиях в зависимости от объема двигателя для популярных автомобилей, выпущенных в период с 1999 по 2008 год. Каждая точка представляет один автомобиль. Цветом обозначен тип привода автомобиля: передний, задний или полный. Данное изображение относится к категории «плохих», так как многие точки накладываются друг на друга, тем самым скрывая соседей. Источник: US Environmental Protection Agency (EPA), fueleconomy.gov

Одним из способов решения этой проблемы является использование частичной прозрачности. Если сделать некоторые точки частично прозрачными, то области наслоения точек будут выглядеть более темными, соответственно, более темный оттенок будет означать повышенную плотность точек в этой части графика (рис. 17.2).

Однако этот метод не всегда является достаточной мерой для предотвращения избыточного наложения точек. Если посмотреть на рис. 17.2, то можно заметить, что некоторые точки имеют более темный оттенок, чем другие, однако понять, сколько конкретно и каких точек было нанесено друг на друга, невозможно. Несмотря на то что различия в плотности цвета точек отчетливо видны, нет никаких пояснений того, в чем состоит суть этих различий. Читатель, впервые увидев данное изображение, вероятно,

задастся вопросом, почему некоторые точки темнее остальных. Тот факт, что на самом деле здесь находится несколько точек, расположенных друг поверх друга, не является интуитивно понятным. Простой прием, который позволяет обыграть данную ситуацию, заключается в том, чтобы добавить к точкам незначительный шум, или *джиттеринг** (jittering, буквально «дрожание»), то есть случайным образом сместить каждую точку на небольшую величину в направлении одной из осей или обеих сразу. Благодаря джиттерингу читатель ясно понимает, что более темные области состоят из точек, нанесенных друг на друга (рис. 17.3). Кроме того, теперь на графике видны и черные точки, изображающие полноприводные автомобили с 2-литровыми двигателями.

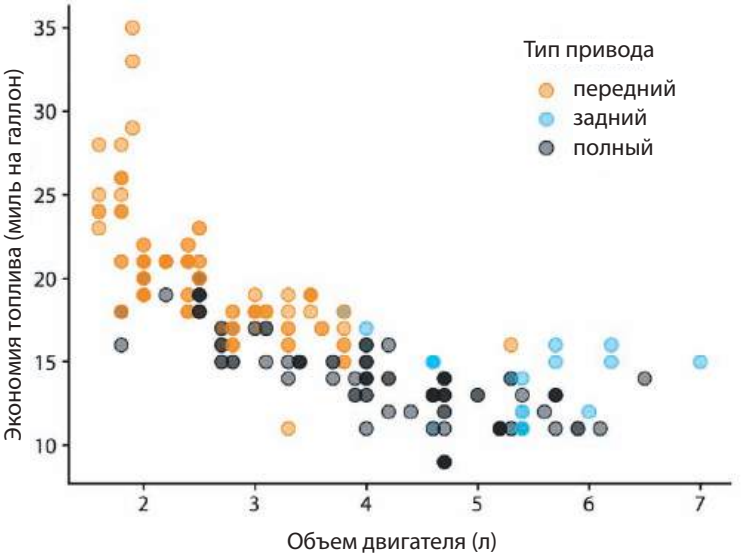


Рис. 17.2. Отношение экономии топлива в городских условиях к объему двигателя. Поскольку все точки сделаны частично прозрачными, точки, расположенные поверх других точек, теперь можно идентифицировать по их более темному оттенку. Источник: EPA

Одним из недостатков джиттеринга является то, что он, пусть и незначительно, но изменяет визуализируемые данные, из-за чего этот прием следует применять с осторожностью. Если мы сдвинем точки слишком сильно, то новое расположение точек исказит реальные значения, и график будет сбивать читателя с толку. Пример такой ситуации приведен на рис. 17.4.

---

* Так его называют в зарубежной литературе, что де-факто является стандартным термином. В русском языке слово «дрожание» применительно к данным не прижилось, а «шум» означает совсем другое, поэтому в данной книге мы будем использовать его английское название. — *Прим. ред.*

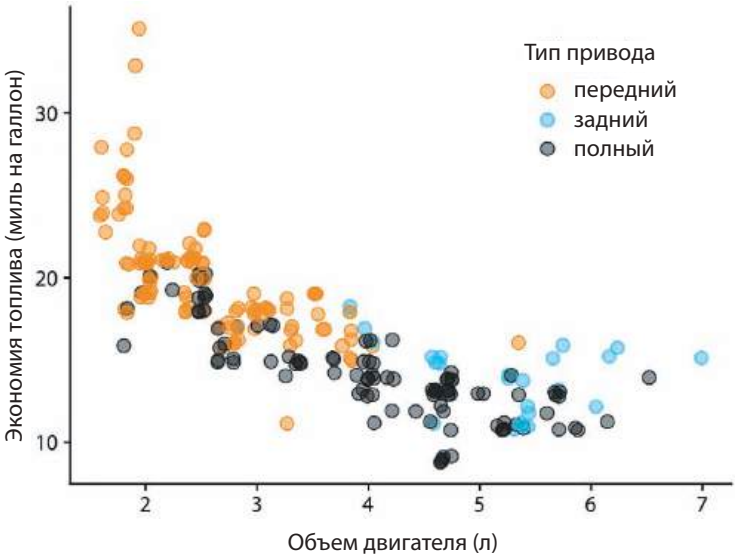


Рис. 17.3. Отношение экономии топлива в городских условиях к объему двигателя. На данном изображении используется прием джиттеринга (дрожания), благодаря чему на графике теперь видно больше точек с одинаковыми значениями. Подобный метод делает изображение более понятным, не отвлекая при этом зрителя от той информации, для визуализации которой и был создан график. Источник: EPA

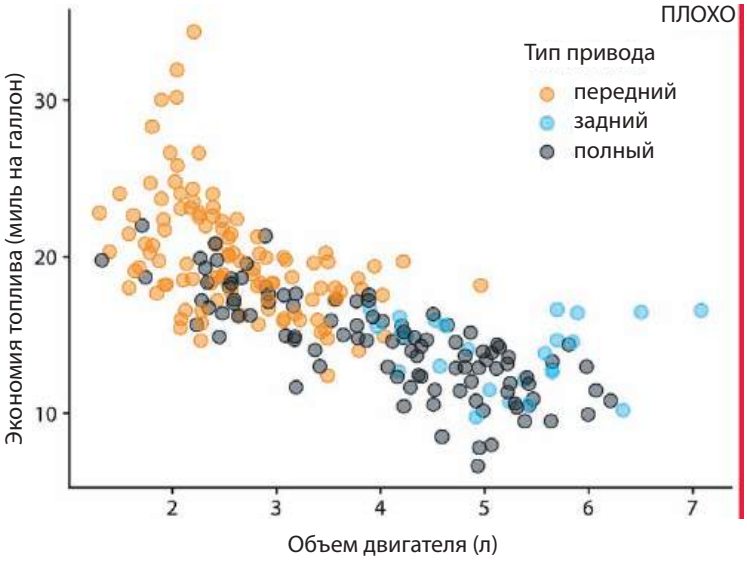


Рис. 17.4. Отношение экономии топлива в городских условиях к объему двигателя. Поскольку на данном изображении разброс точек в результате джиттеринга получился слишком большим, визуализация не является точным отображением исходных данных. Источник: EPA

Двухмерные гистограммы

В тех случаях, когда количество отдельных точек становится очень большим, метод частичной прозрачности (с джиттерингом или без) перестает быть решением проблемы оверплоттинга. Выглядит это так: области с высокой плотностью точек приобретают вид однородных пятен темного цвета, а в областях с низкой плотностью точек отдельные точки становятся едва заметны (рис. 17.5). Изменение уровня прозрачности отдельных точек приведет к смягчению одной проблемы и в то же время к усугублению второй, и наоборот: невозможно подобрать значение параметра, которое позволит избавиться от обеих проблем одновременно.

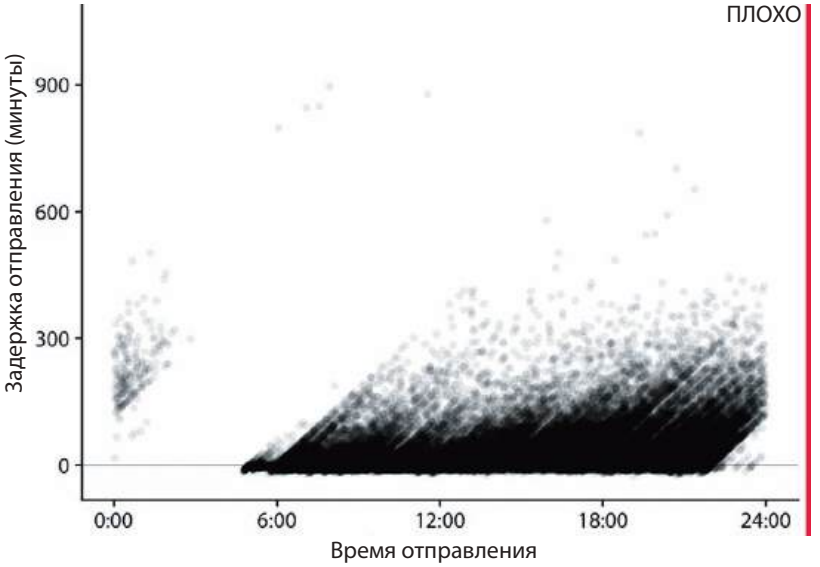


Рис. 17.5. Продолжительность задержек рейсов в зависимости от времени отправления. В качестве данных используется информация обо всех вылетах из аэропорта Ньюарк в 2013 году. Каждая точка соответствует одному вылету самолета. Источник: US Dept. of Transportation, Bureau of Transportation Statistics

На рис. 17.5 визуализированы задержки более 100 000 разных рейсов. Каждая точка представляет один рейс. Несмотря на то что на данном графике используется частичная прозрачность точек, большая часть наслоений выглядит как черная полоса, располагающаяся в границах от 0 до 300 минут задержки. Полностью закрашенная область не позволяет определить, был ли вылет большинства рейсов осуществлен по графику или же со значительной задержкой (скажем, 50 минут или более). Кроме того, из-за прозрачности точек рейсы с наибольшими задержками (400 и более минут) становятся практически незаметными.

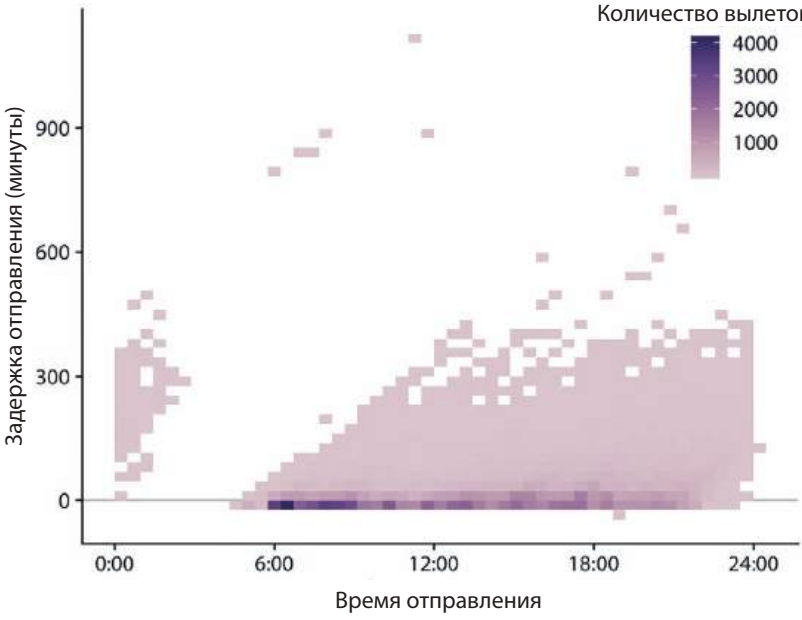


Рис. 17.6. Продолжительность задержек рейсов в зависимости от времени отправления. Цветные прямоугольники отображают рейсы, вылетевшие в указанное время с указанной задержкой. Цветом обозначено количество вылетов, представленных этим квадратом. Источник данных: US Dept. of Transportation, Bureau of Transportation Statistics

В таких случаях вместо построения отдельных точек следует прибегнуть к созданию *двухмерной гистограммы*. Этот вид графика концептуально похож на одномерную гистограмму, которая обсуждалась в главе 6, только сейчас мы будем использовать два измерения. Данный график строится следующим образом: плоскость *x*–*y* нужно разбить на малые прямоугольники, затем подсчитать, сколько значений приходится на каждый из них, а после окрасить их в тот или иной цвет в зависимости от количества значений. На рис. 17.6 показан результат данного подхода применительно к набору данных о задержках рейсов. На этой визуализации хорошо видны несколько важных нюансов. Во-первых, в течение дня (с шести утра до девяти вечера) подавляющее большинство вылетов отправляются фактически без задержки или даже раньше заявленного времени (отрицательная задержка). Небольшая часть вылетов состоялась с существенной задержкой. Более того, из графика видно, что отсрочка вылетов явным образом связана со временем суток: чем позже время вылета, тем больше может быть задержка. Важно отметить, что время вылета — это именно фактическое время вылета, а не запланированное, поэтому нельзя сказать, что самолеты, у которых вылет запланирован на раннее утро, никогда не задерживаются. Однако график позволяет сделать вывод

о том, что если самолет вылетает рано утром, то он вылетает либо с небольшой задержкой, либо — крайне редко — уже с задержкой около 900 минут. В качестве альтернативы объединению данных в прямоугольники мы можем объединить их в шестиугольники [Carr et al., 1987]. Преимущество этого подхода состоит в том, что точки в шестиугольнике в среднем ближе к его центру, чем точки прямоугольника — к его центру, поэтому цветные шестиугольники представляют данные несколько точнее, нежели цветные прямоугольники. На рис. 17.7 данные об отправлении рейсов объединены в шестиугольники.

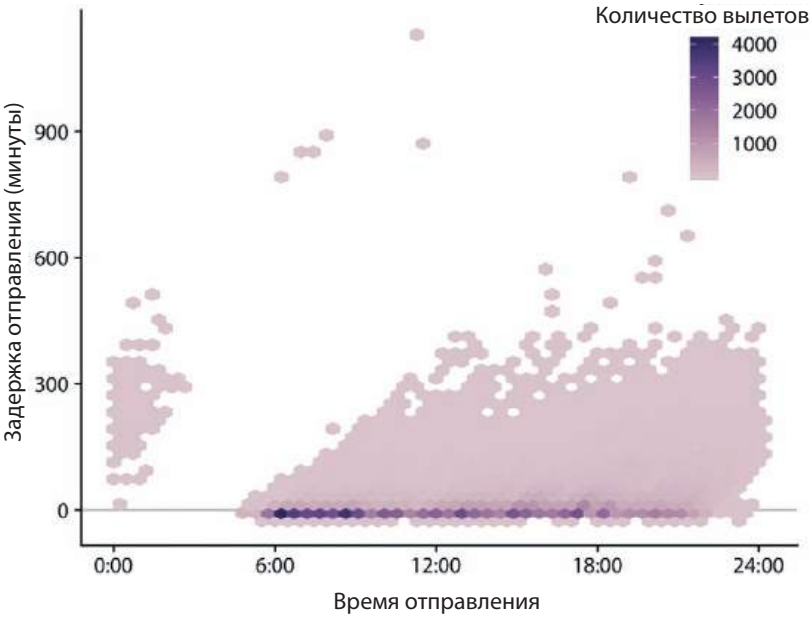


Рис. 17.7. Продолжительность задержек рейсов в зависимости от времени отправления. Цветные шестиугольники отображают рейсы, вылетевшие в указанное время с указанной задержкой. Цветом обозначено количество вылетов, представленных этим шестиугольником. Источник данных: US Dept. of Transportation, Bureau of Transportation Statistics

Изолинии

Альтернативой объединению точек в квадраты или шестиугольники может быть оценка плотности точек по всей площади графика с последующим выделением областей с одинаковой плотностью при помощи изолиний. Этот метод хорошо подходит для тех случаев, когда скорость изменения плотности точек по обеим осям — и по оси x, и по оси y — невелика.

Для иллюстрации этого метода нам опять пригодится набор данных о голубых сойках из главы 11. На рис. 11.1 показана взаимосвязь между длиной

головы и массой тела для 123 голубых соек, при этом часть точек на этом графике перекрывается. Для более четкого выделения распределения точек размер последних следует уменьшить, сделать их частично прозрачными и нанести точки поверх изолиний, очерчивающих области с близкой плотностью точек (рис. 17.8). А чтобы упростить восприятие изменений в плотности точек, области, окруженные линиями контура, можно раскрасить, а областям с более высокой плотностью точек придать более темный оттенок (рис. 17.9).

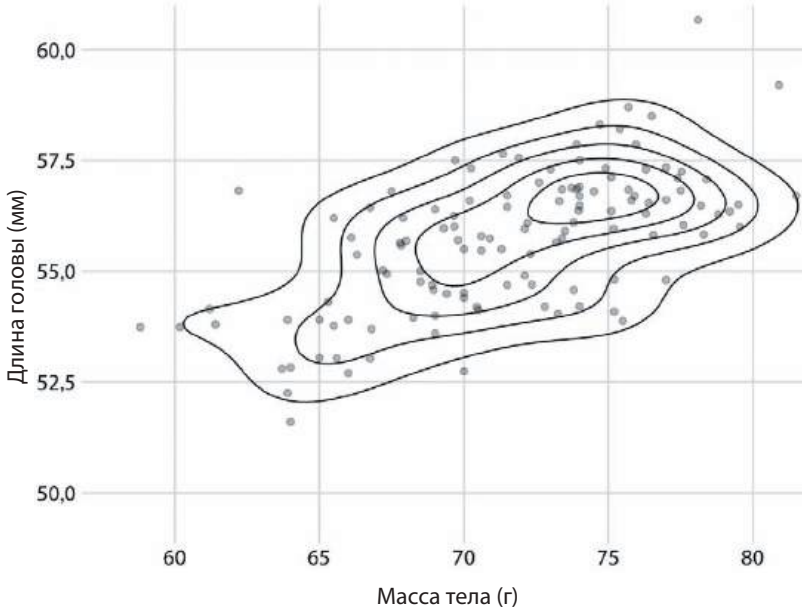


Рис. 17.8. Отношение длины головы к массе тела для набора данных о 123 голубых сойках с рис. 11.1. Каждая точка соответствует одной птице, а линии очерчивают области с одинаковой плотностью точек. Плотность точек растет ближе к центру графика — в окрестности массы тела 75 граммов и длины головы между 55 и 57,5 мм. Источник: Keith Tarvin, Oberlin College

В главе 11 мы также рассмотрели взаимосвязь между длиной головы и массой тела отдельно для самцов и самок голубых соек (см. рис. 11.2). Ту же самую задачу мы можем решить с помощью изолиний, покрасив их в отдельные цвета для самок и самцов птиц (рис. 17.10).

Рисование нескольких изолиний разными цветами может быть эффективной стратегией в деле отображения распределений сразу нескольких групп точек. Однако эту технику следует применять с осторожностью: она годится только для случаев небольшого количества выделяемых групп (от двух до трех), имеющих четкое разделение. В противном случае мы рискуем получить клубок из огромного количества пересекающихся разноцветных линий, не показывающих никакой конкретной картины.

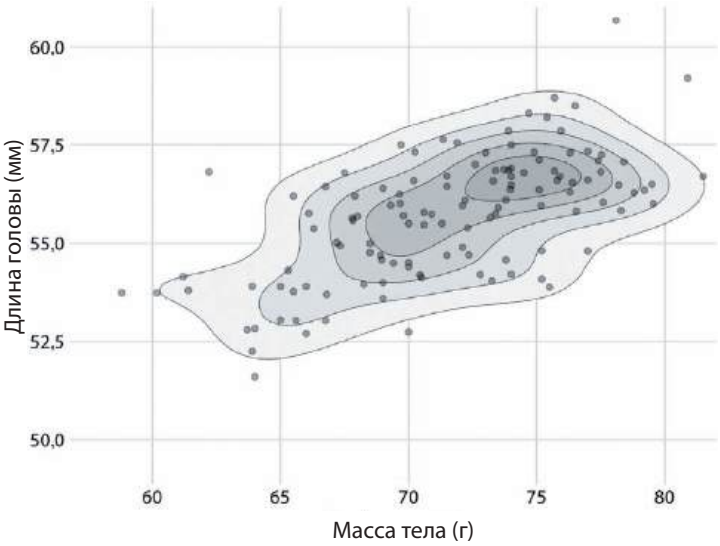


Рис. 17.9. Отношение длины головы к массе тела для набора данных о 123 голубых сойках с рис. 11.1. Данное изображение практически идентично рис. 17.8, за исключением того, что теперь области с очерченным контуром покрашены в различные оттенки серого цвета. Такое затенение дополнительно подчеркивает, что плотность точек растет по мере приближения к центру облака точек. Источник: Keith Tarvin, Oberlin College

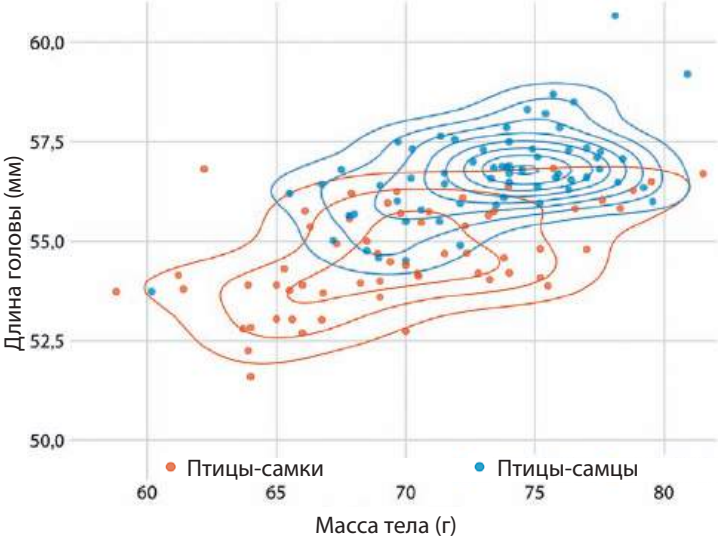


Рис. 17.10. Отношение длины головы к массе тела для набора данных о 123 голубых сойках. Как и на рис. 11.2, здесь мы тоже обозначаем половое разделение птиц, в данном случае с помощью рисования изолиний разного цвета. Из графика видно, что птицы-самцы более плотно сгруппированы в одном регионе диаграммы, тогда как среди птиц-самок, наоборот, заметен больший разброс в физических параметрах. Источник: Keith Tarvin, Oberlin College

Данную проблему я проиллюстрирую с помощью набора данных об алмазах, который содержит информацию о 53 940 драгоценных камнях, включая их цену, вес (карат) и огранку. На рис. 17.11 этот массив данных представлен в виде диаграммы рассеяния. На графике хорошо заметны следы, оставленные оверплоттингом. Из-за огромного количества разноцветных точек график становится почти неинформативным, за исключением разве что общих очертаний области спектра «цена — карат» алмазов.

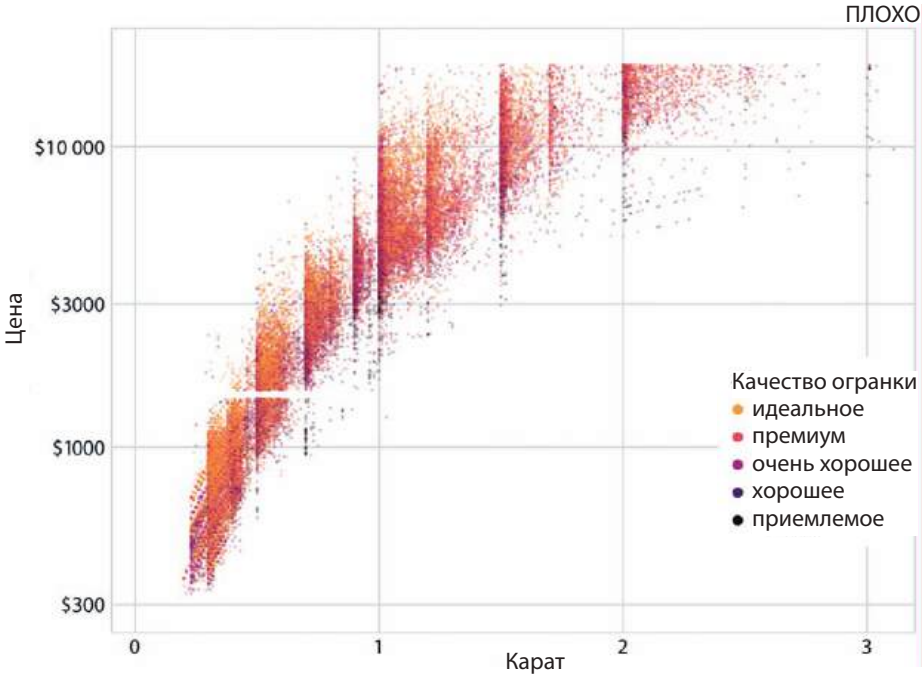


Рис. 17.11. Отношение стоимости бриллиантов к их массе в каратах. На данном графике содержится информация о 53 940 бриллиантах. Все бриллианты обозначены цветом в зависимости от качества огранки. Данное изображение можно отнести к категории «плохих», так как обширный оверплоттинг делает невозможным обнаружение каких-либо закономерностей. Источник: Hadley Wickham, ggplot2

Мы могли бы попытаться нарисовать изолинии для каждой степени качества огранки, как на рис. 17.10. Однако в наборе данных об алмазах у нас есть пять различных цветов, и на графике все цветовые группы в значительной степени перекрывают друг друга. Следовательно, у контурной диаграммы (рис. 17.12) не будет особых преимуществ перед исходной диаграммой рассеяния (рис. 17.11).

Чтобы визуализировать столь крупный массив данных, мы можем поступить следующим образом: для каждого качества огранки нарисовать изолинии на отдельной панели графика (рис. 17.13).

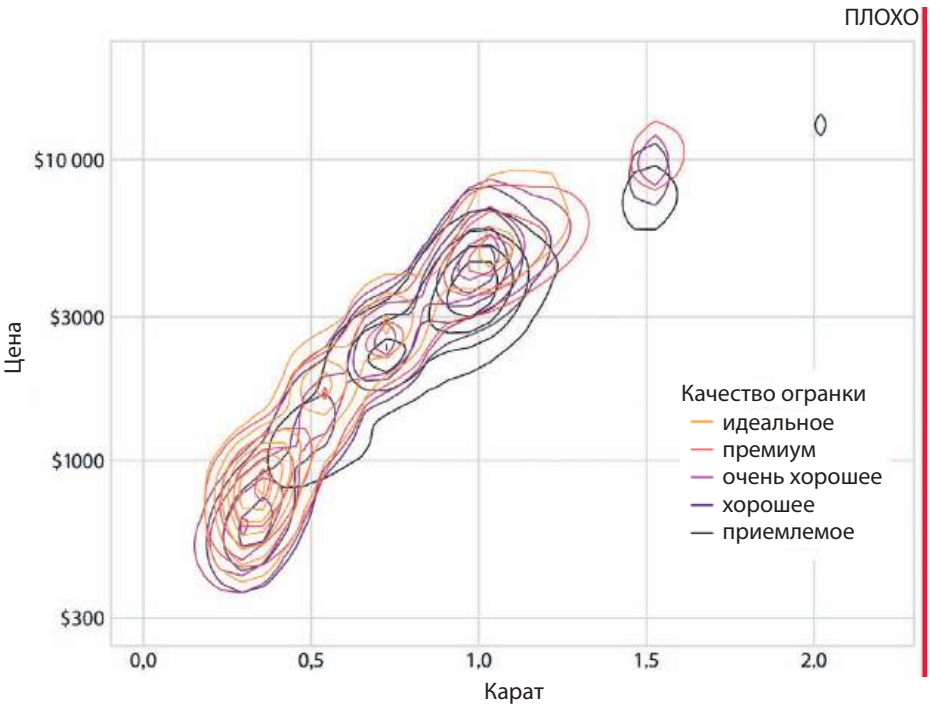


Рис. 17.12. Отношение стоимости бриллиантов к их массе в каратах. График построен по аналогии с диаграммой на рис. 17.11, с той разницей, что отдельные точки были заменены на изолинии плотности. Полученный график также относится к категории «плохих», поскольку все линии контура накладываются друг на друга. График не дает представления ни об общем распределении точек, ни о распределении по качеству огранки. Источник: Hadley Wickham, ggplot2

Данные, расположенные на одной панели, удобно сравнивать, разделив информацию на группы, однако рис. 17.12 настолько визуально перегружен, что сравнение становится невозможным. На рис. 17.13 используется другой подход: на заднем плане панелей присутствует одинаковая для всех панелей сетка, которая позволяет нам сравнивать различные группы огранки, фокусируясь на том, где именно линии контура попадают на линии сетки. Схожего эффекта можно было бы достичь с помощью нанесения частично прозрачных отдельных точек вместо изолиний на каждой панели.

Анализ графика позволяет выделить две основные тенденции. Во-первых, камни с более качественной огранкой (очень хорошая, премиум, идеальная), как правило, имеют более низкую массу в каратах, нежели бриллианты с менее качественной огранкой (хорошей и приемлемой). Напомню, что карат является мерой веса бриллианта (1 карат = 0,2 грамма). Более качественная огранка дает в итоге (в среднем) более легкие бриллианты, поскольку для

получения такой огранки необходимо снять больше материала. Во-вторых, при одной и той же массе в каратах бриллианты с более качественной огранкой имеют, как правило, более высокую стоимость.

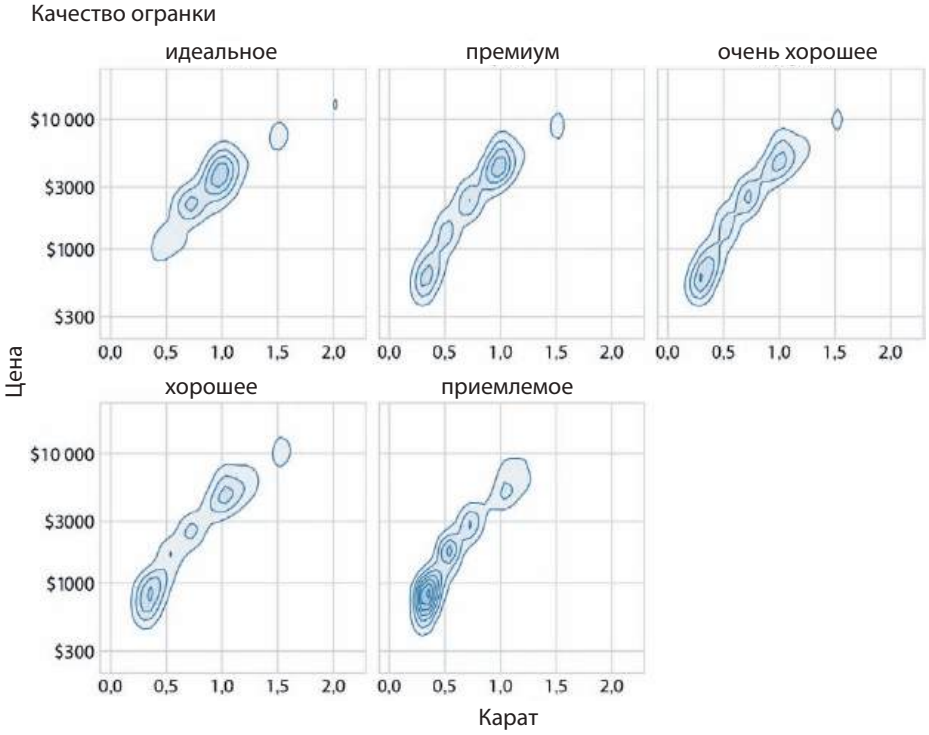


Рис. 17.13. Отношение стоимости бриллиантов к их массе в каратах. На данном изображении мы взяли изолинии плотности с рис. 17.12 и нарисовали их отдельно для каждого типа огранки. Как теперь видно, бриллианты с лучшим качеством огранки (очень хорошая, премиум, идеальная) имеют в целом меньшую массу, нежели бриллианты с худшим качеством огранки, но при этом их цена за карат выше. Источник: Hadley Wickham, ggplot2

Данная тенденция хорошо заметна на примере распределения цен бриллиантов массой в 0,5 карата. Распределение смещено вверх для бриллиантов с более качественной огранкой, и, в частности, оно гораздо выше для алмазов с идеальной огранкой, чем для алмазов с хорошей или приемлемой огранкой.

Глава 18

Распространенные ошибки при использовании цвета

Цвет может быть невероятно эффективным инструментом для улучшения качества визуализации данных. В то же время плохо подобранный цвет способен испортить даже отлично сделанный график. Добавляя в свои визуализации цвет, всегда знайте, для чего вы это делаете. Кроме того, цвет должен быть чистым и не отвлекать читателя от сути диаграммы.

Отображаем слишком много или ненужную информацию

Одной из наиболее распространенных ошибок является использование слишком большого количества элементов разных цветов. Взгляните на рис. 18.1. На этом графике показан рост численности населения по сравнению с количеством жителей во всех 50 штатах США, а также в округе Колумбия. Как видите, я присвоил каждому штату свой цвет, и результат вышел не самым удачным. И хотя мы всегда можем определить, на какой штат мы смотрим в данный момент (соотнеся точки на графике с легендой), это потребует от нас немалых усилий. Цветов слишком много, и многие из них очень похожи друг на друга. Даже если мы потратим силы на то, чтобы разобраться в представленных нам данных, факт остается фактом: цвет нам здесь скорее мешает, чем помогает. Цвет должен способствовать пониманию графика и делать изучение данных более удобным, а не быть источником визуальных головоломок.

Как правило, наиболее качественные цветовые шкалы получаются в тех случаях, когда мы имеем дело с 3–5 различными категориями, которые предполагаем раскрасить каждую в свой цвет. Но как только речь заходит о 8–10 категориях, задача сопоставления цветов резко усложняется и ее решение становится слишком обременительным и малополезным, даже если мы выберем хорошо различающиеся между собой цвета.

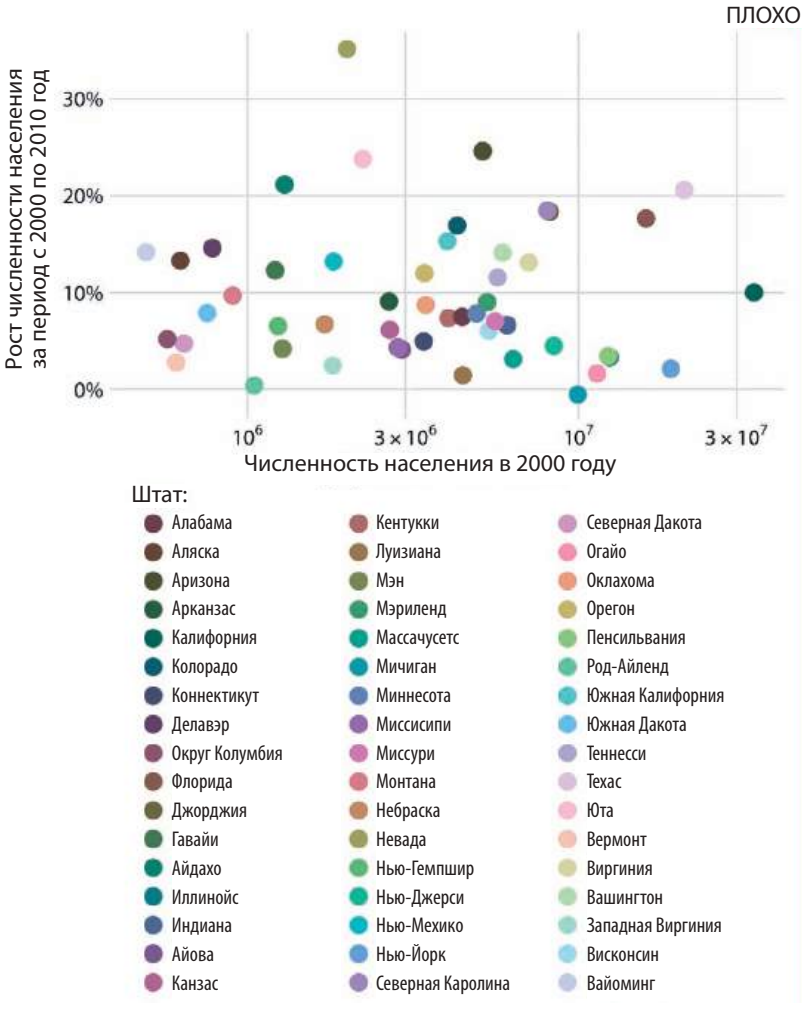


Рис. 18.1. Прирост населения за период 2000–2010 годов и общая численность населения в 2000 году в 50 штатах и округе Колумбия. Каждый штат окрашен в свой цвет. Поскольку штатов на карте довольно много, очень трудно сопоставить цвета в легенде с точками на диаграмме рассеяния. Источник: US Census Bureau

Для набора данных, показанного на рис. 18.1, кажется самым правильным использовать цвет только для указания географического региона каждого штата, а сами штаты просто промаркировать, то есть разместить соответствующие текстовые метки рядом с точками данных (рис. 18.2). Несмотря на то что промаркировать каждый штат, не перегрузив график информацией, невозможно, мы все же можем воспользоваться прямой маркировкой. Как правило, для визуализаций, подобных этой, нам не нужно отмечать выноской каждую точку данных. Достаточно обозначить репрезентативное

подмножество, например набор штатов, которые мы собираемся упомянуть в тексте, сопровождающем рисунок. Если мы захотим предоставить читателю доступ ко всей информации, то данные графика можно просто представить ниже в виде таблицы.

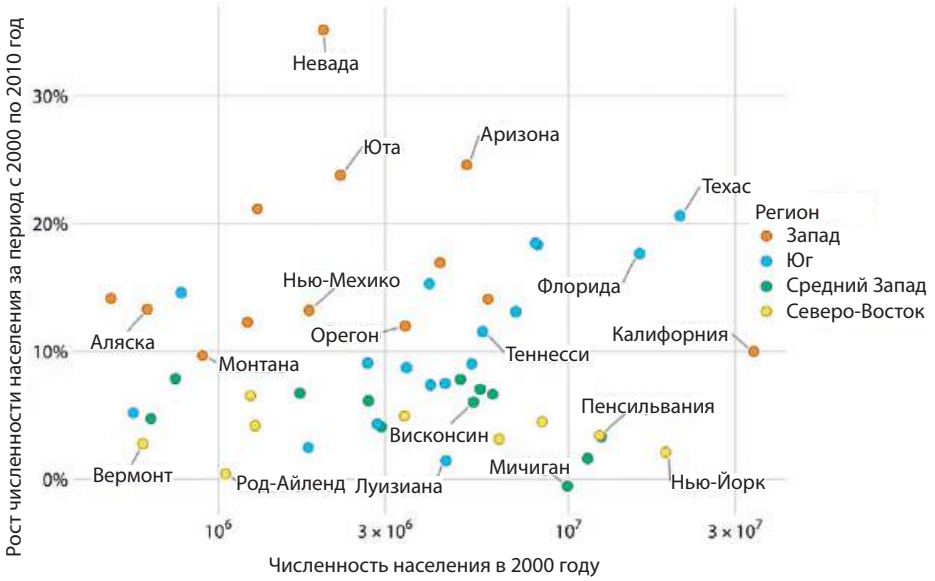


Рис. 18.2. Прирост населения за период с 2000 по 2010 год и общая численность населения в 2000 году в 50 штатах и округе Колумбия. В отличие от рис. 18.1, здесь я раскрасил штаты в зависимости от региона и выделил некоторые из них. Чтобы визуально не перегружать график, большую часть штатов я оставил без маркировки. Источник: US Census Bureau

Если на вашем графике присутствует более восьми категорий элементов, используйте прямую маркировку, а не цвет.

Другой распространенной ошибкой при использовании цвета является раскрашивание ради самого раскрашивания, без какой-либо конкретной цели. В качестве примера рассмотрим рис. 18.3, который представляет собой вариацию рис. 3.2. Здесь у каждого столбца есть свой цвет, который не имеет отношения к географическому региону. Цвета выбраны таким образом, чтобы получился эффект радуги.

Этот визуальный прием сам по себе смотрится красиво, однако он никак не влияет на интерпретирование графика читателем и не упрощает восприятие диаграммы.

Помимо избыточного количества цветов, на рис. 18.3 имеется еще одна проблема, тоже связанная с раскраской: оттенки цветов слишком яркие, из-за чего на рисунок трудно даже смотреть. Например, прочитать названия штатов, не отвлекаясь на большие ярко окрашенные области рядом с названиями, очень трудно. Аналогично соотнесение конечных точек столбцов с линиями сетки, расположенными ниже, тоже будет непростой задачей.

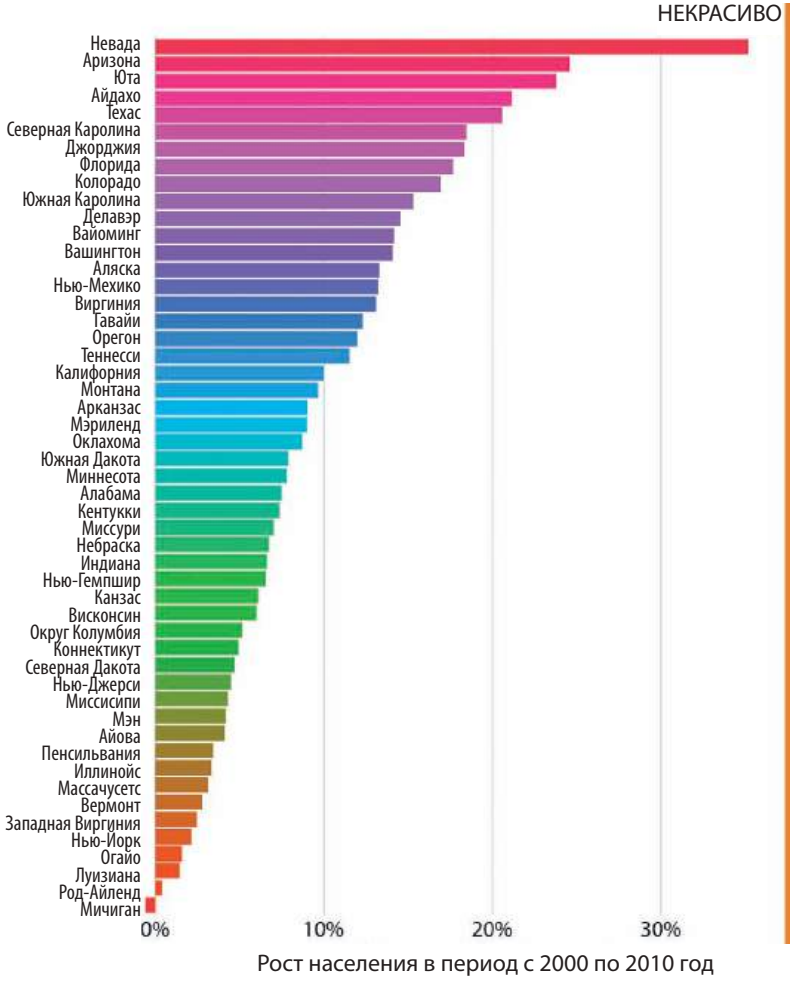


Рис. 18.3. Прирост населения США за период с 2000 по 2010 год. Создающие эффект радуги цвета столбцов не несут никакой смысловой нагрузки и попросту отвлекают зрителя. Кроме того, цвета слишком яркие. Источник: US Census Bureau

Не закрашивайте большие области слишком яркими цветами. Это мешает читателю внимательно рассматривать рисунок.

Использование немонотонных цветовых шкал для передачи значений данных

В главе 3 я говорил о существовании двух критических условий, которые должны соблюдаться при создании последовательных цветовых шкал, используемых для передачи данных: цвет должен однозначно показывать, какие значения меньше, а какие больше; кроме того, различия между цветами должны визуализировать соответствующие различия между значениями данных. К сожалению, некоторые цветовые шкалы, в том числе весьма популярные, нарушают как минимум одно из этих условий. Наиболее распространенной шкалой, в которой очень часто присутствует упомянутый изъян, является радуга (рис. 18.4). На этом графике цвет проходит последовательно всю видимую часть спектра. Это означает, что шкала фактически закольцована: цвет в начале и конце шкалы практически совпадает (темно-красный). Если эти два цвета соседствуют на графике, наше восприятие не считает, что они представляют значения данных, находящиеся максимально далеко друг от друга. Кроме того, шкала очень немонотонна. В некоторых ее областях цвета меняются очень медленно, а в некоторых — очень быстро. Отсутствие монотонности особенно хорошо заметно в случае преобразования этой шкалы в оттенки серого (см. рис. 18.4). Шкала начинается с оттенков средней яркости, переходит к ярким оттенкам, затем к темному участку и, наконец, обратно к средней яркости. А еще на шкале присутствуют протяженные участки, где яркость практически не меняется, за которыми следуют более узкие участки с гораздо более заметными перепадами яркости.

Рис. 18.4. Цветовая гамма радуги очень немонотонна. Это особенно хорошо заметно на шкале, преобразованной в оттенки серого. Слева направо шкала начинается с оттенков средней яркости, переходит к ярким оттенкам, затем к темному участку и, наконец, обратно к средней яркости. Кроме того, сама яркость меняется очень неравномерно. Самая светлая часть (соответствующая желтому, светло-зеленому и голубому цветам) занимает почти треть всей шкалы, в то время как самая темная часть (соответствующая темно-синему цвету) сосредоточена в достаточно узкой области

Визуализация данных с помощью радужной шкалы часто приводит к потере особенностей данных, а кроме того, такая шкала склонна произвольным

образом выделять те или иные аспекты (рис. 18.5) информации. К тому же цвета в радужной шкале тоже чрезмерно насыщенны. Рассматривание рис. 18.5 в течение продолжительного времени может вызвать зрительный дискомфорт.

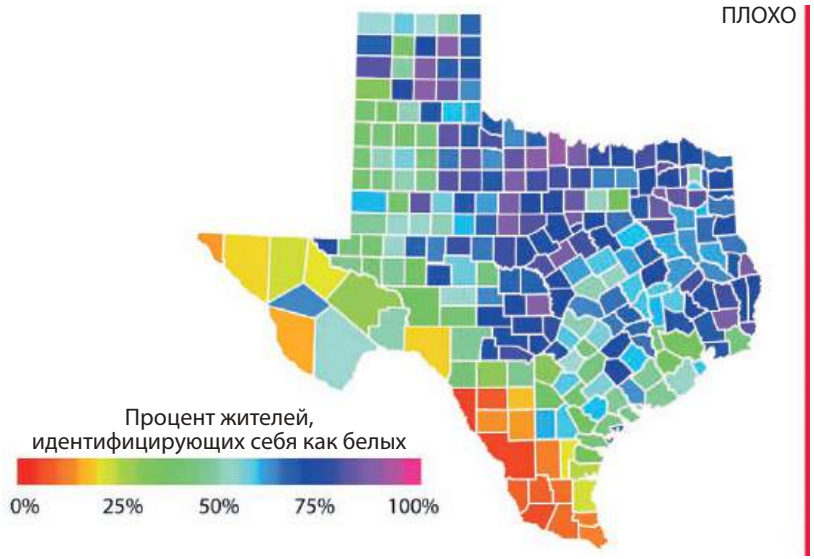


Рис. 18.5. Процент людей, идентифицирующих себя как белых, в округах Техаса. Радужная цветовая шкала не подходит для визуализации непрерывных значений данных, поскольку она склонна делать акцент на произвольных особенностях данных. На этом рисунке шкала выделила округа, в которых приблизительно 75% населения идентифицируют себя как белых. Источник данных: 2010 US Decennial Census

Игнорирование потребностей людей с нарушениями цветового зрения

При выборе цвета для визуализации следует учитывать тот факт, что у значимой части наших читателей может присутствовать та или иная форма нарушения цветового зрения (то есть какая-то из форм дальтонизма). Такие читатели могут быть не в состоянии различать цвета, которые другие люди различают без каких-либо проблем. Нарушение цветового зрения не означает, что человек вообще не видит никаких цветов. Как правило, в такой ситуации людям трудно различать определенные типы цветов, такие как красный и зеленый (красно-зеленое нарушение цветового зрения) или синий и зеленый (сине-желтое нарушение цветового зрения). Для указанных и родственных им состояний существуют специальные медицинские термины: дейтераномалия/дейтеранопия и протаномалия/протанопия для красно-зеленого варианта (когда люди испытывают трудности с восприятием зеленого или красного цвета соответственно) и тританомалия/тританопия для сине-желтого варианта (когда люди

испытывают трудности с восприятием синего варианта). Термины, оканчивающиеся на «-аномалия», подразумевают частично ухудшенное восприятие соответствующего цвета, а термины, оканчивающиеся на «-анопия», означают полное отсутствие восприятия указанного цвета. Приблизительно 8% мужчин и 0,5% женщин страдают от того или иного вида нарушений цветового зрения: наиболее распространенной формой является дейтераномалия, тогда как тританомалия встречается относительно редко.

Как мы уже знаем из главы 3, в визуализации данных используются три основных типа цветовых шкал: последовательные шкалы, расходящиеся шкалы и качественные шкалы. Последовательные шкалы, как правило, являются наиболее подходящими для людей с нарушениями цветового зрения, поскольку корректная последовательная шкала представляет собой непрерывный градиент от темных цветов к светлым. На рис. 18.6 показана тепловая шкала с рис. 3.3 в смоделированных версиях дейтераномалии, протаномалии и тританомалии. Несмотря на то что ни одна из этих шкал не похожа на оригинал, все они представляют собой четкий градиент от темного цвета к светлому и все они хорошо выполняют свою работу по передаче величин значений данных.

Рис. 18.6. Моделирование нарушений цветового зрения для последовательной цветовой шкалы, изменяющейся от темно-красного до светло-желтого. В направлениях слева направо и сверху вниз мы видим исходную шкалу и шкалы, моделирующие восприятие в условиях дейтераномалии, протаномалии и тританомалии. Несмотря на то что определенные цвета выглядят по-разному при всех трех типах нарушений цветового зрения, в каждом случае хорошо виден четкий переход от темного цвета к светлому. Отсюда следует вывод, что данная цветовая шкала подходит для зрителей с нарушениями цветового восприятия

Что касается расходящихся шкал, то здесь дела обстоят немного сложнее. Популярные цветовые контрасты могут быть неразличимы для людей с нарушениями цветового восприятия. В частности, красный и зеленый цвета обеспечивают наиболее сильный контраст для людей с нормальным цветовым зрением, но становятся почти неразличимыми для дейтанов (людей с дейтераномалией) или протанов (людей с протаномалией) (рис. 18.7). Аналогично сине-зеленые контрасты хорошо воспринимаются дейтанами и протанами, но становятся неразличимыми для тританов (людей с тританомалией) (рис. 18.8).

Рис. 18.7. Красно-зеленый контраст становится неразличимым при дейтераномалии и протаномалии

Рис. 18.8. Сине-зеленый контраст становится неразличимым при тританомалии

Рис. 18.9. Шкала ColorBrewer PiYG (от розового до желто-зеленого), используемая на рис. 3.5, выглядит почти как красно-зеленый контраст для людей с нормальным цветовым зрением, но при этом является различимой для людей со всеми формами дефицита цветового зрения. Так происходит потому, что красноватый цвет на самом деле является розовым (смесь красного и синего), в то время как зеленоватый цвет содержит желтый. Разница в синем компоненте между этими двумя цветами доступна для восприятия дейтанам и протанам, а разница в красном компоненте будет заметна тританам

Глядя на эти примеры, можно легко прийти к неверному выводу о невозможности подбора двух контрастных цветов, сочетание которых было бы безопасно при любом типе недостатка цветового зрения. Однако ситуация не так страшна, как кажется на первый взгляд. Зачастую достаточно внести небольшие изменения в цвета, чтобы они сохранили желаемые характеристики и при этом были бы различимы зрителями с нарушениями цветового зрения. Например, шкала ColorBrewer PiYG (от розового до желто-зеленого) на рис. 3.5 выглядит

красно-зеленой для людей с нормальным цветовым зрением и при этом остается различимой для людей с нарушениями цветового восприятия (рис. 18.9).

Наиболее сложным случаем являются качественные шкалы, поскольку они требуют для своего создания множества различных цветов, причем все они должны быть различимы при любом типе недостатка цветового зрения. Я считаю, что в этой ситуации предпочтение следует отдавать специализированной шкале, которая была разработана с целью решения этой проблемы (рис. 18.10). Палитра, содержащая восемь разных цветов, пригодна практически для любого сценария с разными цветами. Как уже говорилось в начале этой главы, вне зависимости от ситуации следует избегать передачи более восьми различных элементов на одном графике при помощи цвета.

Рис. 18.10. Цветовая палитра для качественной шкалы, учитывающая все особенности цветового восприятия [Okabe and Ito, 2008]. Буквенно-цифровые коды обозначают цвета в пространстве RGB, записанные в шестнадцатеричном формате. Во многих программах для работы с изображениями и графических библиотеках эти коды можно использовать как есть, либо же вы можете воспользоваться значениями из табл. 18.1

Таблица 18.1. Цветовая шкала, дружественная людям с нарушениями цветового зрения [Okabe and Ito, 2008]

Название	Hex-код	Hue	C, M, Y, K (%)	R, G, B (0–255)	R, G, B (%)
Оранжевый	#E69F00	41°	0, 50, 100, 0	230, 159, 0	90, 60, 0
Небесно-голубой	#56B4E9	202°	80, 0, 0, 0	86, 180, 233	35, 70, 90
Синевато-зеленый	#009E73	164°	97, 0, 75, 0	0, 158, 115	0, 60, 50
Желтый	#F0E442	56°	10, 5, 90, 0	240, 228, 66	95, 90, 25
Синий	#0072B2	202°	100, 50, 0, 0	0, 114, 178	0, 45, 70
Киноварь	#D55E00	27°	0, 80, 100, 0	213, 94, 0	80, 40, 0
Красновато-фиолетовый	#CC79A7	326°	10, 70, 0, 0	204, 121, 167	80, 60, 70
Черный	#000000	Н/п	0, 0, 0, 100	0, 0, 0	0, 0, 0

Несмотря на то что на данный момент существует несколько хороших шкал, способных нивелировать нарушения цветового зрения, следует признать, что они не являются панацеей. Бывает так, что, даже используя шкалу, различимую при нарушениях цветового зрения, мы можем получить визуализацию, которую человек с нарушением цветового зрения просто не сможет понять. Одним из наиболее важных параметров является размер цветных графических элементов. Различение цветов значительно упрощается, если

они применены к большим областям, а не к маленьким, или к тонким линиям [Stone, Albers Szafir, Setlur, 2014], и при нарушениях цветового зрения этот эффект только усиливается (рис. 18.11). В дополнение к различным аспектам цветового дизайна, которые обсуждались в этой главе и в главе 3, я рекомендую вам просматривать ваши цветные рисунки в режиме симуляции нарушений цветового зрения, чтобы понять, как разные люди будут видеть эти диаграммы. Для этого вы можете воспользоваться существующими онлайн-сервисами или настольными приложениями.

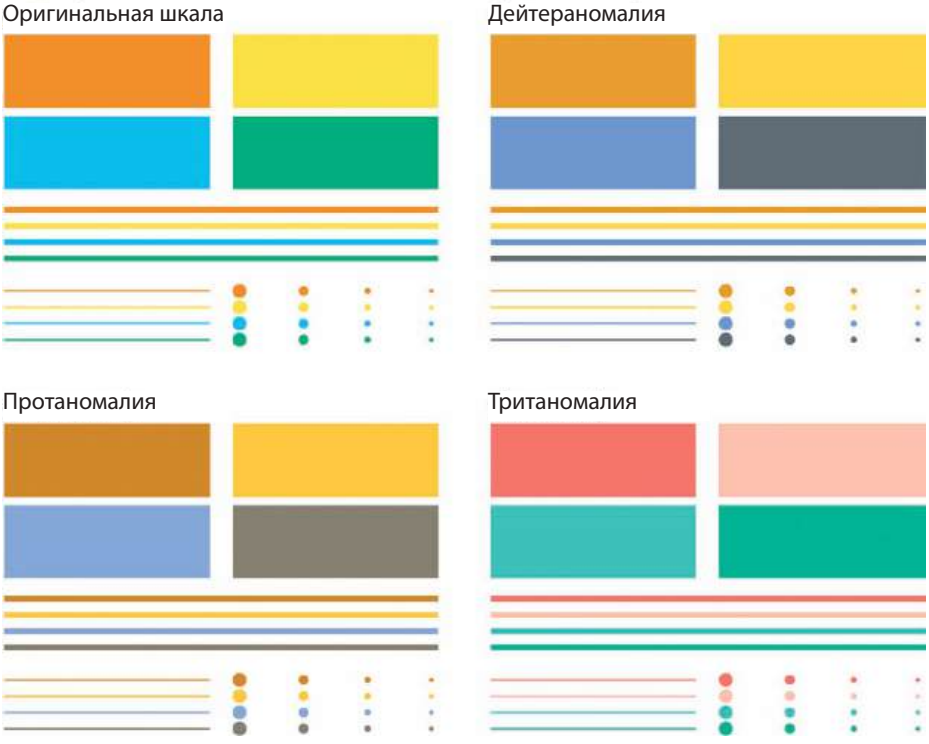


Рис. 18.11. Цветные элементы небольшого размера трудно отличимы друг от друга. Верхняя левая панель (помеченная как «оригинальная») показывает четыре прямоугольника, четыре толстые линии, четыре тонкие линии и четыре группы точек разного размера, которые окрашены в четыре одинаковых цвета. Хорошо заметно, что чем меньше или тоньше визуальные элементы, тем сложнее их различить. В симуляции нарушений цветового восприятия эта проблема только усугубляется, цвета становится сложно различить даже в случае крупных графических элементов

Чтобы ваши графики были доступны для восприятия людьми с недостатком цветового зрения, не полагайтесь лишь на специализированные цветовые шкалы. Обязательно проверяйте, как ваши диаграммы выглядят в симуляторах недостатков цветового зрения.

Глава 19

Избыточная передача данных

Примеры из главы 18 ясно показывают, что передача информации с помощью цвета не всегда приводит к желаемому результату. Если у нас есть большое количество различных элементов, которые должны быть выделены на графике, простое раскрашивание может не дать нужного результата, и мы лишь переусложним сопоставление цветов на графике с цветами в легенде (см. рис. 18.1). И даже если нам нужно выделить только два или три разных элемента, от цвета может оказаться мало толку, если эти элементы очень маленькие (см. рис. 18.11) и/или цвета выглядят одинаково для людей с нарушениями цветового восприятия (см. рис. 18.7 и 18.8). Простое и эффективное решение состоит в том, чтобы использовать цвет для улучшения визуального представления графика, при этом не делая его единственным средством передачи ключевой информации. Я называю этот принцип проектирования избыточной передачей, потому что он побуждает нас передавать информацию с помощью нескольких различных эстетических аспектов.

Проектирование легенд с применением принципа избыточной передачи данных

Зачастую диаграммы рассеяния для нескольких групп данных проектируются таким образом, чтобы точки, представляющие разные группы, отличались лишь цветом. В качестве примера рассмотрим рис. 19.1, на котором показана зависимость ширины чашелистика от его длины для трех разных видов ирисов. (Чашелистики — это внешние листья цветов у цветковых растений.) Точки, представляющие различные виды, имеют разный цвет, но в остальном выглядят совершенно одинаково. Несмотря на то что на этом графике присутствует только три группы точек, прочесть его будет трудно даже людям с полным цветовым восприятием. Причина этого в том, что точки данных для двух видов Iris virginica и Iris versicolor перемешаны на графике и их цвета, зеленый и синий, слабо отличаются друг от друга. Это может выглядеть странно, но люди с недостатками цветового зрения в красно-зеленом спектре (дейтераномалия или протаномалия) воспринимают зеленые и синие точки более четко, чем люди без нарушений цветового восприятия (сравните верхний ряд на рис. 19.2 с рис. 19.1).

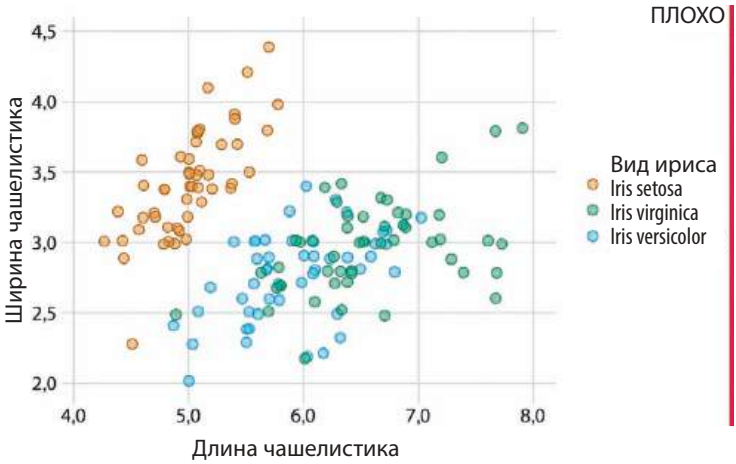


Рис. 19.1. Отношение ширины чашелистика к его длине для трех разных видов ирисов (Iris setosa, Iris virginica и Iris versicolor). Каждая точка представляет измерения для одного образца растения. Чтобы избежать оверплоттинга, ко всем точкам был применен небольшой джиттеринг. Данная визуализация относится к категории «плохих», потому что точки virginica зеленого цвета и точки versicolor синего цвета трудно отличить друг от друга. Источник: [Fisher, 1936]

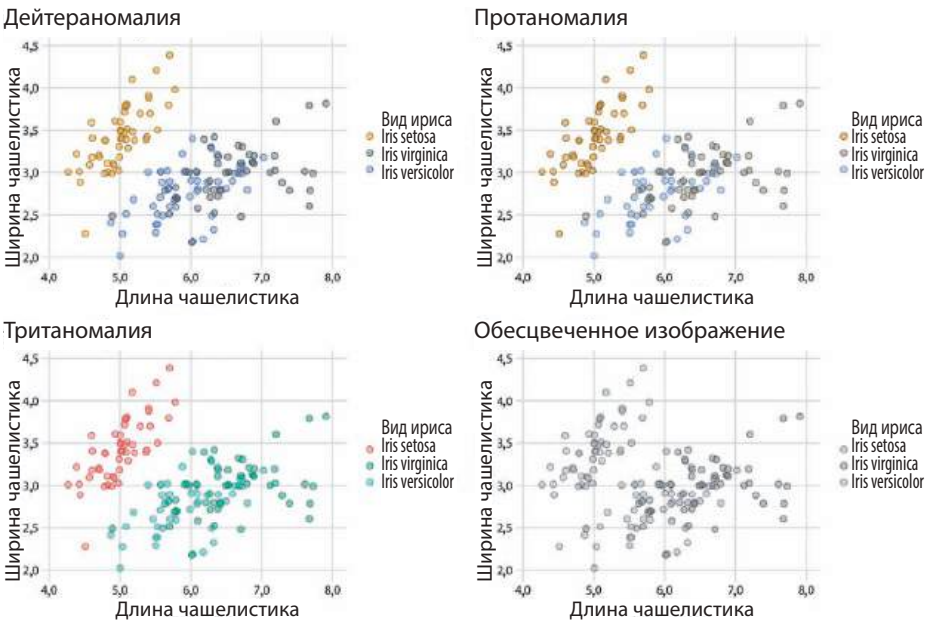


Рис. 19.2. Варианты рис. 19.1, полученные с использованием симуляции нарушений цветового восприятия. Источник: [Fisher, 1936]

С другой стороны, для людей с недостатками в сине-желтом спектре (тританомалия) синие и зеленые точки выглядят очень похожими (см. рис. 19.2,

внизу слева). А уж если мы распечатаем данное изображение в черно-белом режиме (то есть обесцветим рисунок), нам не удастся различить ни один из видов ириса (см. рис. 19.2, внизу справа).

Чтобы от этих проблем избавиться, можно улучшить наш график следующими двумя способами: во-первых, поменять местами цвета, используемые для Iris setosa и Iris versicolor, чтобы синий цвет не соседствовал с зеленым (рис. 19.3), а также использовать для отображения точек три разных символа, чтобы их было проще отличать друг от друга. Благодаря этим изменениям становятся понятными как исходная версия рисунка (см. рис. 19.3), так и его версии для людей с нарушениями цветового восприятия, а также обесцвеченный вариант диаграммы (рис. 19.4).

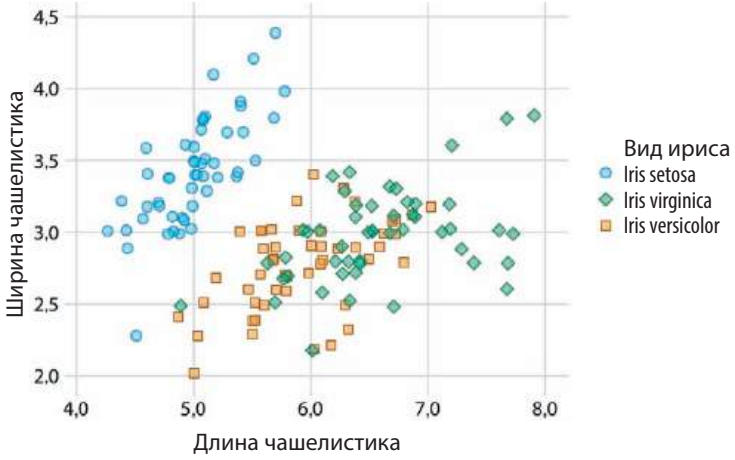


Рис. 19.3. Отношение ширины чашелистика к его длине для трех разных видов ирисов (Iris setosa, Iris virginica и Iris versicolor). По сравнению с рис. 19.1 мы поменяли местами цвета точек Iris setosa и Iris versicolor, а также изменили форму маркеров для всех видов ирисов таким образом, чтобы каждый вид имел собственный символ. Источник: [Fisher, 1936]

Изменение формы маркеров — это довольно простая стратегия, которая хорошо работает для диаграмм рассеяния, однако для других типов графиков этот метод может не подойти. Для линейных графиков мы можем изменить тип линии (сплошная, прерывистая, пунктирная и т. д.; см. рис. 1.1), но практика показывает, что использование прерывистых или пунктирных линий часто дает неоптимальные результаты. В частности, если пунктирные или прерывистые линии не являются идеально прямыми или только слегка изогнутыми, выглядят они не очень хорошо и создают визуальный шум. Кроме того, если на графике нарисовано множество линий с различной штриховкой, то сопоставление их с легендой будет весьма трудоемкой задачей. Итак, каким образом мы бы могли улучшить визуализацию на рис. 19.5? На этом

рисунке с помощью линий отображены изменения цен акций во времени для четырех крупных технологических компаний.

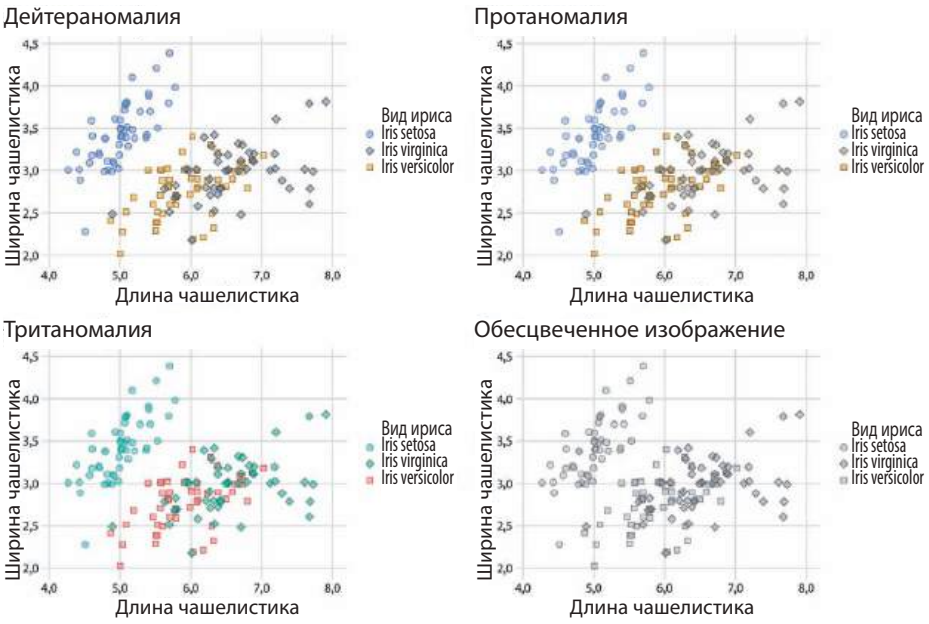


Рис. 19.4. Варианты рис. 19.3, полученные с использованием симуляции нарушений цветового зрения. Благодаря использованию различных форм маркеров даже полностью обесцвеченная версия рисунка позволяет полноценно читать график. Источник: [Fisher, 1936]

Рис. 19.5. Изменения во времени цен акций четырех крупных технологических компаний. В июне 2012 года цены акций этих компаний были нормализованы и приняты за 100. Это изображение относится к категории «плохих», поскольку для сопоставления названий компаний в легенде с кривыми требуется немало усилий. Источник: Yahoo! Finance

На рисунке представлены четыре линии, показывающие цены на акции четырех разных компаний. Линии имеют цветовую кодировку с использованием цветовой шкалы, подходящей для людей с нарушениями цветового зрения. На первый взгляд кажется, что связать каждую строку с соответствующей компанией должно быть относительно просто, однако это не так. Изъян данного рисунка заключается в том, что строки данных имеют визуальный порядок. Желтая линия, представляющая Facebook, воспринимается как самая высокая линия, черная линия, представляющая Apple, воспринимается как самая низкая, а компании Alphabet и Microsoft находятся между ними. При этом в легенде названия компаний перечислены в следующем порядке: Alphabet, Apple, Facebook, Microsoft (то есть по алфавиту). Таким образом, воспринимаемый порядок строк данных отличается от порядка перечисления компаний в легенде, из-за чего зрителю потребуется приложить неожиданно серьезные умственные усилия для сопоставления линий с названиями компаний.

Такая проблема обычно возникает при использовании программного обеспечения для построения графиков, которое автоматически генерирует легенды. Программы, строящие графики, ничего не знают о существовании визуального порядка и об удобстве его восприятия зрителем. Вместо этого программное обеспечение сортирует легенду по заданному по умолчанию порядку, чаще всего по алфавиту. Решение проблемы очевидно: нужно вручную переупорядочить записи в легенде, чтобы они соответствовали наблюдаемому порядку данных (рис. 19.6). Выполнив эти действия, мы получим рисунок, на котором сопоставлять легенду с данными будет гораздо проще.

Рис. 19.6. Изменения во времени цен акций четырех крупных технологических компаний. В отличие от рис. 19.5, записи в легенде упорядочены таким образом, чтобы они соответствовали воспринимаемому визуальному порядку линий данных. Как можно видеть, самый большой рост показали акции Facebook, а самый низкий — Apple. Источник: Yahoo! Finance

Если на графике данные имеют некий визуальный порядок, то легенда должна ему соответствовать.

Согласовывать легенду с порядком следования данных имеет смысл всегда, но плюсы этого согласования особенно очевидны в условиях симуляции недостатков цветового зрения (рис. 19.7). Например, на тританомальной версии рисунка, где синий и зеленый цвета трудно различимы, соблюдение этого правила будет очень полезным (см. рис. 19.7, внизу слева). То же самое касается и версии графика в оттенках серого (см. рис. 19.7, справа внизу). Несмотря на то что два цвета — для Facebook и для Alphabet — имеют практически одинаковое значение серого, мы видим, что Microsoft и Apple представлены более темными цветами и занимают два нижних места. Благодаря этому мы можем сделать правильный вывод, что самая высокая строка соответствует Facebook, а вторая строка по высоте — компании Alphabet.

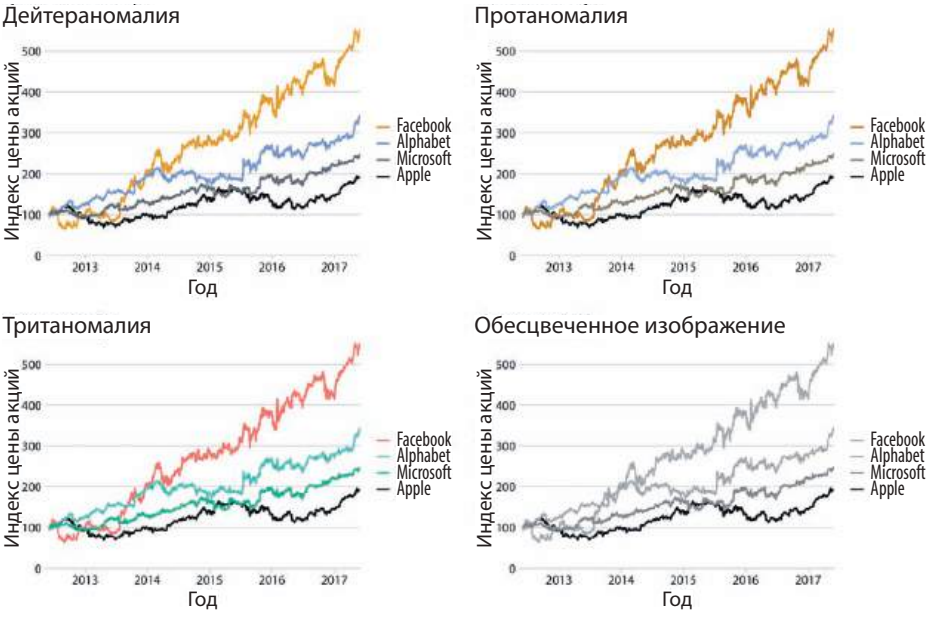


Рис. 19.7. Варианты рис. 19.6, полученные с использованием симуляции нарушений цветового восприятия. Источник: Yahoo! Finance

Проектирование визуализаций без легенды

Несмотря на то что удобочитаемость легенды можно улучшить при помощи принципа избыточной передачи данных, с эстетической точки зрения легенда всегда заставляет читателя тратить больше усилий на изучение графика.

При чтении легенды зрителю необходимо собрать информацию в одной части рисунка и затем использовать ее в другой части. Как правило, отказ от легенды может лишь упростить жизнь нашей аудитории. Тем не менее исключение легенды не означает, что взамен мы будем размещать какие-нибудь пояснения в подписи к изображению, например, «Желтые точки относятся к Iris versicolor». Исключение легенды означает, что мы проектируем визуализацию таким образом, чтобы при одном взгляде на нее было понятно, что обозначают те или иные графические элементы.

Наиболее общим подходом здесь являются *подписи данных*, что означает размещение соответствующих текстовых меток или других визуальных элементов в качестве ориентиров для всего остального графика. В главе 18 мы уже сталкивались с этим методом (см. рис. 18.2), когда использовали его в качестве альтернативы рисованию легенды с более чем 50 цветами. Чтобы добавить на график цен акций подписи данных, мы помещаем название каждой компании рядом с концом соответствующей линии данных (рис. 19.8).

Рис. 19.8. Изменения во времени цен акций четырех крупных технологических компаний. В июне 2012 года цены акций этих компаний были нормализованы и приняты за 100. Источник: Yahoo! Finance

Старайтесь оформлять графики таким образом, чтобы они не нуждались в отдельной легенде.

Давайте теперь попробуем добавить подписи данных к данным об ирисах, о которых шла речь в начале этой главы (см. рис. 19.3). Поскольку этот график является диаграммой рассеяния, состоящей из множества точек, которые, в свою очередь, образуют три разные группы, нам необходимо обозначить

именно группы, а не отдельные точки. Например, мы можем заключить большинство точек в эллипсы и добавить подписи к ним, а не к точкам (рис. 19.9).

Что касается графиков плотности, то здесь вместо использования легенды с цветовой кодировкой мы можем добавить подписи непосредственно к кривым (рис. 19.10). На рис. 19.9 и 19.10 я раскрасил текстовые метки в те же цвета, что и данные. Раскраска подписей способна значительно усилить эффект от маркировки, однако таким образом мы рискуем усложнить восприятие данных. Если метки будут иметь слишком светлый тон, прочесть их будет очень трудно. Поскольку текст обычно состоит из очень тонких линий, цветной текст часто кажется светлее, чем смежные области, залитые тем же цветом. Чтобы избежать этой проблемы, я обычно использую два разных оттенка каждого цвета: светлый для закрашенных областей и темный для линий, контуров и текста. Если вы внимательно рассмотрите рис. 19.9 или 19.10, то увидите, что каждая точка данных или заштрихованная область заполнены светлым цветом, а их контур выполнен в более темном цвете того же оттенка. Аналогично подписи рядов тоже выполнены в темном цвете.

Мы также можем использовать графики ядерной оценки плотности, как, например, тот, что показан на рис. 19.10, в качестве альтернативы легенде: мы можем поместить графики оценки плотности на поля диаграммы разброса (рис. 19.11). Подобные действия позволяют нам добавлять подписи к графикам оценок плотности, не перегружая саму диаграмму рассеяния, вследствие чего визуальная нагрузка графика становится меньше, чем на рис. 19.9, на котором подписи добавлены к огибающим эллипсам.

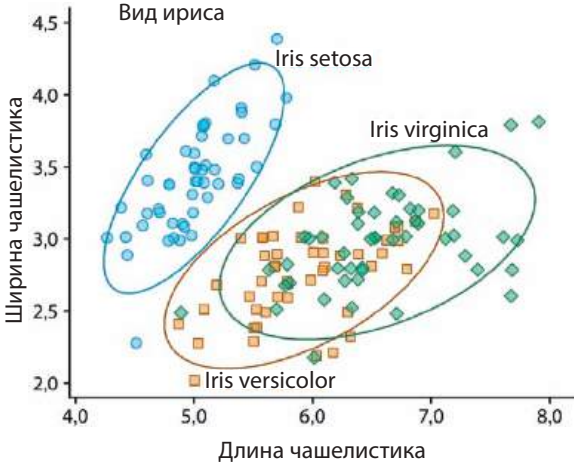


Рис. 19.9. Отношение ширины чашелистика к его длине для трех разных видов ирисов. Точки, представляющие различные виды ирисов, были очерчены цветными эллипсами и снабжены подписями. Если сравнить данную визуализацию с рис. 19.3, то можно заметить, что я удалил фоновую сетку, так как она делала график визуально перегруженным. Источник: [Fisher, 1936]

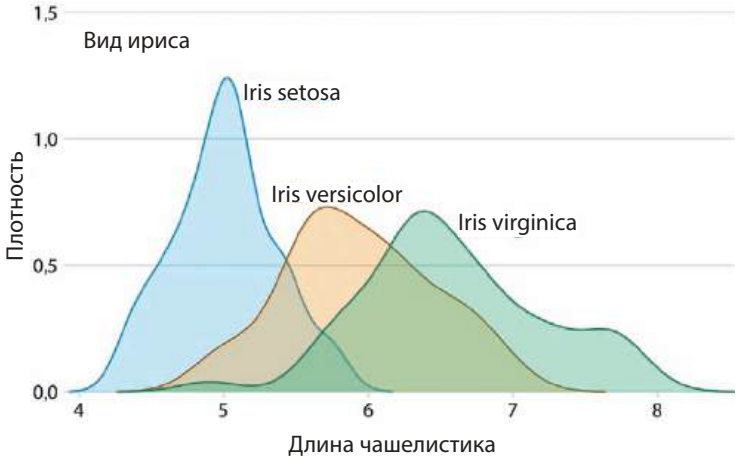


Рис. 19.10. Ядерная оценка плотности длин чашелистиков трех видов ирисов. Каждая оценка плотности снабжена подписью с названием соответствующего вида. Источник: [Fisher, 1936]

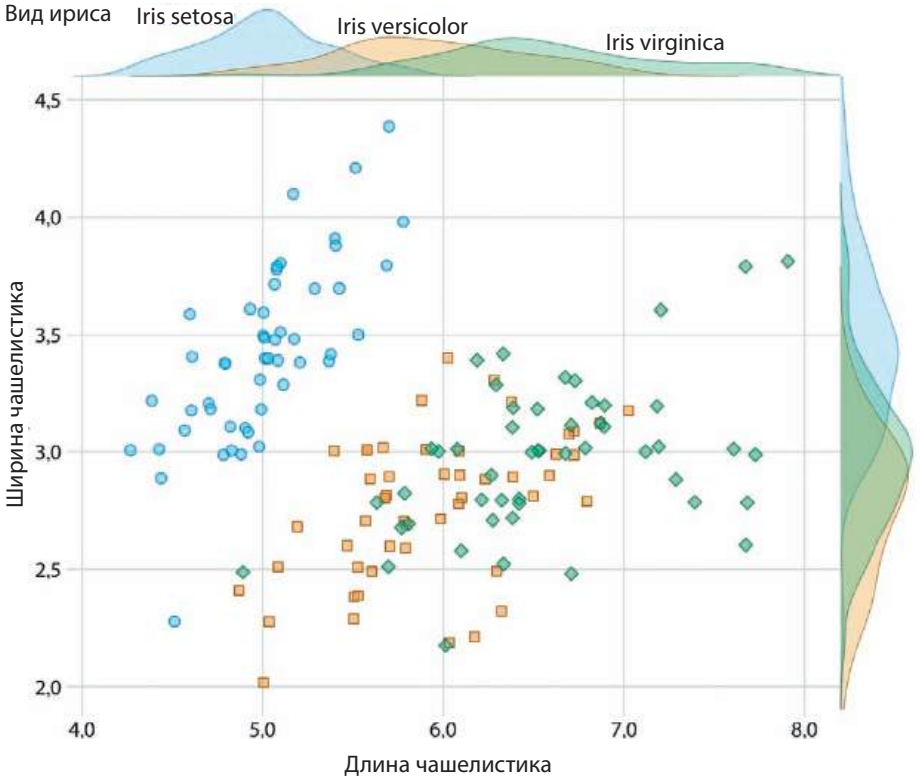


Рис. 19.11. Отношение ширины чашелистика к его длине для трех разных видов ирисов. На графике также расположены ядерные оценки плотности обеих переменных для каждого вида ирисов. Источник: [Fisher, 1936]

Наконец, когда мы представляем переменную одновременно в нескольких визуальных элементах, мы, как правило, не хотим делать несколько разных легенд. Куда лучше было бы придумать такой визуальный элемент, который обозначит все сопоставления одновременно. Если мы передаем одну и ту же переменную посредством цвета и положения вдоль главной оси, это означает, что эталонная цветовая полоса должна проходить строго вдоль той же оси и являться ее неотъемлемой частью. На рис. 19.12 температура обозначается одновременно положением на оси *x* и цветом, поэтому мы интегрировали цветовую легенду в ось *x*.

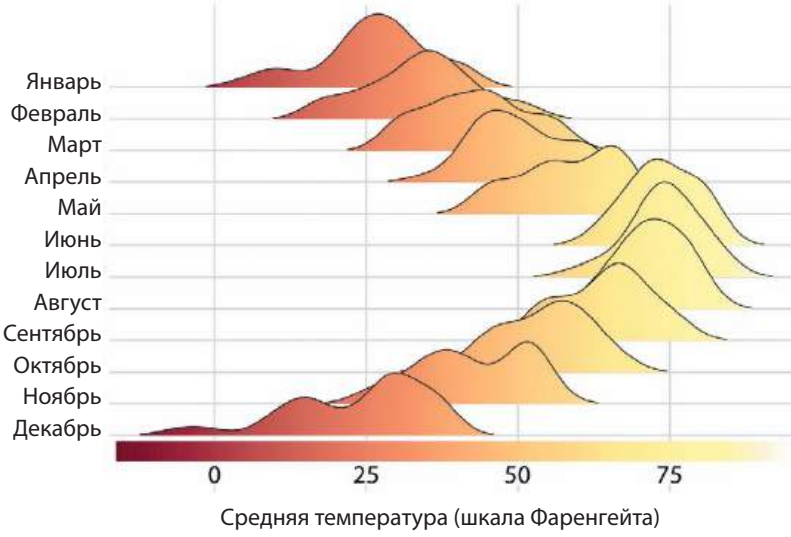


Рис. 19.12. Температуры в городе Линкольн, штат Небраска, в 2016 году. Данная визуализация является вариацией рис. 8.9. Значение температуры передано с помощью положения на оси *x* и цвета, а цветная полоса вдоль оси *x* является одновременно шкалой, согласно которой температура преобразуется в цвет. Источник: Weather Underground

Глава 20

Многопанельные визуализации

Когда наборы данных становятся большими и сложными, они часто содержат гораздо больше информации, чем имеет смысл показывать на одной-единственной диаграмме. Более рациональным способом отображения таких наборов данных являются многопанельные визуализации. Под термином «многопанельная визуализация» понимается изображение, которое состоит из нескольких панелей с графиками, и каждый из них отображает некоторое подмножество данных. Подобные визуализации можно разбить на две группы: малые панельные, состоящие из небольшого количества панелей, и составные графики. *Малые панельные* визуализации как раз и представляют собой графики, состоящие из нескольких панелей, расположенных в виде сетки. Каждая панель отображает отдельное подмножество данных, при этом на всех панелях используется один и тот же тип визуализации. *Составные* визуализации же представляют собой набор отдельных панелей, расположенных в произвольном порядке (который может как основываться на стандартной сетке, так и нет). Подобные графики могут состоять из совершенно разных типов визуализаций или даже отражать разные наборы данных.

Ранее в этой книге мы уже встречались с обоими типами многопанельных визуализаций. По своей сути эти изображения интуитивно понятны, и их довольно легко интерпретировать. Однако при создании таких рисунков необходимо учитывать несколько нюансов, таких как: подходящее масштабирование осей, выравнивание и согласованность между отдельными панелями.

Малые панельные визуализации

По-английски данный вид визуализаций называется small multiple (small — небольшой, multiple — множество). Широкому распространению данного термина значительно поспособствовал американский статистик Эдвард Тафти [Tufte, 1990]. Другой вариант термина, trellis plot, был популяризован учеными Кливлендом, Беккером и их коллегами из Bell Labs [Cleveland, 1993; Becker, Cleveland, and Shyu, 1996]. Вне зависимости от используемого названия суть этого способа визуализации заключается в следующем: сначала мы разбиваем исходные

данные на определенное количество частей по одному или нескольким измерениям (категориям), затем мы визуализируем каждую подгруппу данных по отдельности и в конце концов упорядочиваем эти визуализации в виде сетки. Столбцы, строки или отдельные панели в сетке помечаются значениями по измерениям, на основе которых они построены. Относительно недавно у этой техники появилось еще одно, третье название — faceting*, в честь методов, с помощью которых такие графики создаются в популярной библиотеке графиков ggplot2 (например, функция ggplot2 facet_grid ()) [Wickham, 2016].

Давайте перейдем к примерам и начнем с визуализации набора данных о пассажирах «Титаника», который мы уже использовали ранее в этой книге. Массив данных можно разбить на группы в зависимости от класса каюты, в которой путешествовал тот или иной пассажир, а также по тому, выжил человек или нет. В каждом из этих шести фрагментов данных есть пассажиры как мужского, так и женского пола, и мы можем визуализировать их количество при помощи столбчатых диаграмм. В результате мы получим шесть таких представлений, которые расположим в двух колонках: первая — погибшие пассажиры, вторая — выжившие. В каждом из этих столбцов содержится по три строки, по одной для каждого класса (рис. 20.1). Так как у столбцов и строк в решетке проставлены подписи, читатель может сразу увидеть, какой из шести графиков соответствует той или иной комбинации класса каюты и статуса выживаемости после катастрофы.

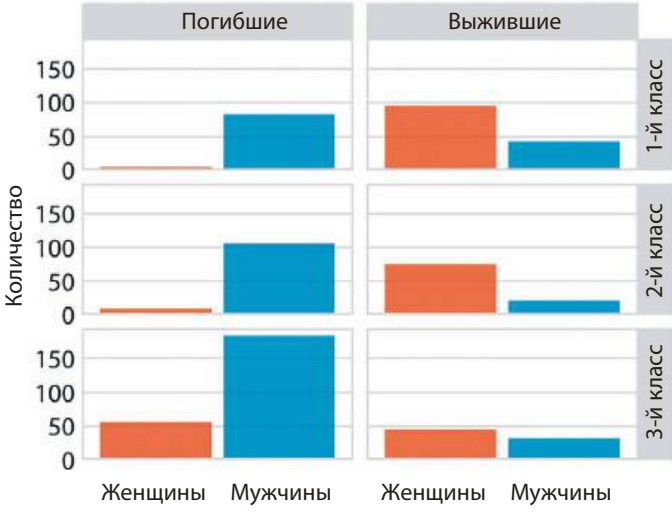


Рис. 20.1. Распределение пассажиров «Титаника» по полу, классу каюты (1-й, 2-й или 3-й) и статусу выживаемости. Источник: Encyclopedia Titanica

* На русский язык это переводится как «огранка», что соответствует смыслу: построение различных «граней» рассматриваемого множества данных; однако в литературе такое название не используется. — *Прим. ред.*

Этот график представляет собой интуитивно понятную визуализацию судеб пассажиров «Титаника». Мы сразу видим, что большинство мужчин умерли, а большинство женщин выжили. Кроме того, умершие женщины в большинстве своем были пассажирками третьего класса.

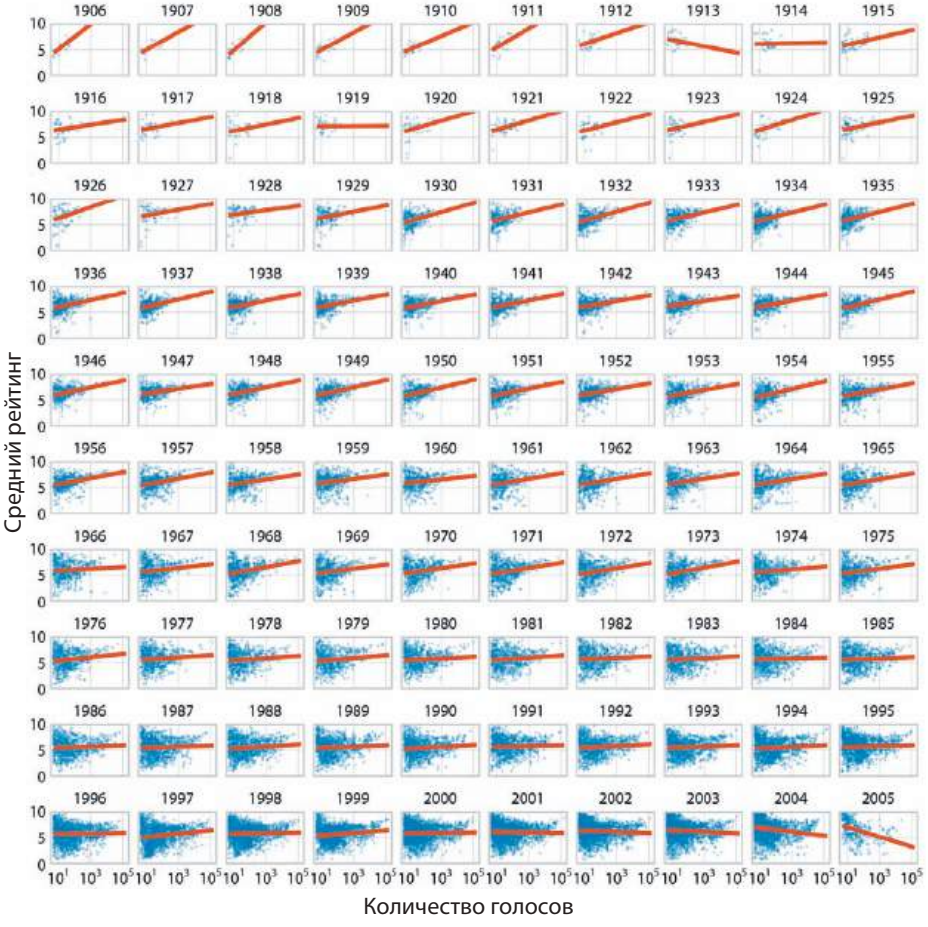


Рис. 20.2. Средние рейтинги в зависимости от количества голосов для фильмов, выпущенных в период с 1906 по 2005 год. Точками синего цвета обозначены отдельные фильмы, а оранжевые линии представляют линейную регрессию среднего рейтинга каждого фильма в зависимости от логарифма количества голосов, полученных фильмом. Как правило, фильмы с большим количеством голосов в среднем имеют более высокий средний рейтинг. Однако к концу XX века эта тенденция ослабла, и у фильмов, выпущенных в начале 2000-х годов, заметна обратная зависимость. Источник: IMDB

Малые панельные визуализации являются мощным инструментом для одновременной визуализации очень больших объемов данных. На рис. 20.1 используется шесть отдельных панелей, но при желании мы легко можем

добавить гораздо больше. На рис. 20.2 показана взаимосвязь между средним рейтингом фильма в сервисе Internet Movie Database (IMDB) и количеством голосов, полученных фильмом, за каждый год периода в 100 лет. Здесь мы выделили только одну категорию — год выпуска, а панели, соответствующие каждому году, расположили рядами, идущими слева направо и сверху вниз. Эта визуализация показывает, что существует общая взаимосвязь между средним рейтингом фильма и количеством отданных за него голосов, то есть фильмы с бóльшим количеством голосов, как правило, имеют более высокий рейтинг. Однако сила, с которой проявляется тенденция, зависит от года, а для фильмов, выпущенных в начале 2000-х годов, подобная связь отсутствует полностью или даже является отрицательной.

Рис. 20.3. Тенденции в присвоении степеней бакалавра в высших учебных заведениях США. Все показанные научные области в среднем составляют более 4% от общего количества выпускников. Данная визуализация относится к категории «плохих», так как на панелях используются разные масштабы оси *y*. Из-за этого относительные объемы присвоения степеней бакалавра по различным специальностям кажутся одинаковыми, а изменения, произошедшие в некоторых областях, сильно преувеличены.
Источник: National Center for Education Statistics

Чтобы графики такого размера можно было легко считывать, очень важно, чтобы на каждой панели использовались одинаковые диапазоны осей

и одинаковый масштаб. В противном случае есть большая вероятность, что читатель неверно истолкует показанное на рисунке. Например, давайте посмотрим на рис. 20.3, на котором показано, как со временем менялась доля бакалавров в различных областях науки. На рисунке показаны девять научных областей, на которые в среднем приходилось более 4% всех степеней, присвоенных в период с 1971 по 2015 год. На каждой панели ось *y* масштабирована таким образом, чтобы кривая покрывала весь диапазон значений по оси *y*. Как следствие, после достаточно беглого анализа легко сделать вывод, что все девять областей популярны одинаково и колебания этой популярности тоже имеют одинаковый диапазон.

Если на всех панелях масштаб оси *y* привести к одинаковой шкале, то мы сразу увидим, как сильно предыдущее изображение искажало исходные данные (рис. 20.4). Некоторые области науки намного популярнее других, а в части областей популярность выросла или сократилась намного больше остальных. Например, количество бакалавров в сфере образования резко снизилось, тогда как доля бакалавров в сфере художественных искусств оставалась приблизительно постоянной или слегка выросла.

Рис. 20.4. Тенденции в присвоении степеней бакалавра в высших учебных заведениях США. Все показанные научные области в среднем составляют более 4% от общего количества выпускников. Источник: National Center for Education Statistics

Вообще, я не рекомендую использовать разные масштабы осей в отдельных панелях малых панельных визуализаций. К сожалению, на практике встречаются такие ситуации, когда избежать этого невозможно. Если вы столкнетесь в своей работе с подобным сценарием, как минимум обратите внимание читателя на эту особенность в подписи к рисунку. Например, так: «Обратите внимание, что на разных панелях используется разный масштаб по оси *y*».

А еще важно учитывать порядок следования панелей. Если в нем будет присутствовать какая-то логика, вашу визуализацию будет гораздо проще воспринимать. На рис. 20.1 я расположил ряды в порядке убывания класса — от самого высокого (первый класс) до самого низкого (третий класс). На рис. 20.2 я расположил панели в порядке возрастания года: от верхнего левого угла к нижнему правому. На рис. 20.4 панели расположены в порядке уменьшения популярности научной области, поэтому самые популярные области находятся в верхнем ряду и/или слева, а наименее популярные — в нижнем ряду и/или справа.

Панели в малых панельных визуализациях следует располагать в осмысленном и логичном порядке.

Составные визуализации

Не каждая визуализация, состоящая из нескольких панелей, подходит под определение малой панельной. Порой все, что нам нужно, — простое объединение нескольких независимых панелей в один результирующий график. В этом случае мы берем отдельные графики и располагаем их в виде рядов, столбцов или любым другим более сложным способом, после чего эта схема приобретает статус единого рисунка. Хорошим примером является рис. 20.5, который продолжает анализ тенденций в присуждении степеней бакалавра высшими учебными заведениями США. На панели A рис. 20.5 показан рост общего количества присвоенных степеней с 1971 по 2015 год. Как можно видеть, это число выросло примерно в два раза. На панели B показано изменение долей присужденных степеней за тот же период времени в пяти наиболее популярных областях науки. Мы видим, что в период с 1971 по 2015 год количество степеней, выдаваемых в областях образования и социальных наук и истории, значительно сократилось, в то время как количество бакалавров в сферах бизнеса и здравоохранения существенно выросло.

Обратите внимание, что, в отличие от примеров с малыми панельными визуализациями, маркировка отдельных панелей составной визуализации имеет алфавитный порядок. Как правило, в таком случае используют

строчные или заглавные буквы латинского алфавита, чтобы у каждой панели был уникальный идентификатор. Например, если я захочу рассказать о той части рис. 20.5, которая показывает изменения в процентах присуждаемых степеней, я могу сослаться на панель B этого рисунка или просто на рис. 20.5B. В противном случае мне пришлось бы постоянно ссылаться на правую или левую панель рис. 20.5, что довольно неудобно, а кроме того, ссылки на эти панели выглядели бы очень громоздко. Что уж говорить о более сложных визуализациях! Фактически маркировку можно не проставлять лишь в том случае, когда мы имеем дело с малой панельной визуализацией, поскольку в ней каждая панель однозначно определяется переменной (переменными), значения которой переданы в подписях панелей.

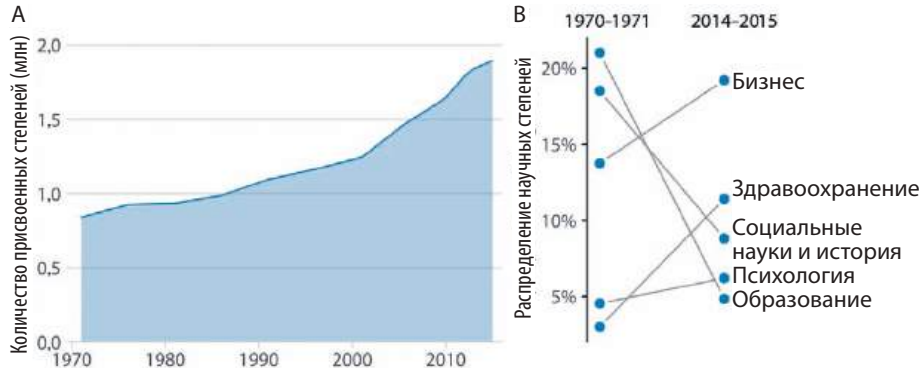


Рис. 20.5. Тенденции в присвоении степеней бакалавра в высших учебных заведениях США. A. В период с 1970 по 2015 год количество присвоенных степеней выросло вдвое. B. Среди наиболее популярных направлений бакалавриата заметный спад произошел в областях образования и социальных наук и истории, в то время как популярность профессий по направлениям бизнеса и здравоохранения возросла. Источник: National Center for Education Statistics

Обдумывая маркировку панелей составной визуализации, учитывайте тот факт, что подписи панелей должны вписываться в общий дизайн рисунка. Я часто вижу изображения, на которых подписи выглядят так, как будто они были намалеваны другим человеком. Нередко подписи наносятся буквами слишком большого размера, что чрезмерно отвлекает на себя внимание читателя; также бывает, что подписи находятся в неподходящем месте или набраны разными шрифтами (см. пример на рис. 20.6). Когда читатель смотрит на составную визуализацию, подписи и метки — не то, что должно бросаться в глаза в первую очередь. Да и вообще, метки и подписи просто не должны выделяться. В большинстве случаев и так понятно, какая подпись относится к какой панели, потому что правила маркировки подразумевают, что отсчет идет с буквы «A», которая располагается в верхнем левом углу, а далее маркировка наносится последовательно слева направо и сверху вниз.

Я смотрю на подписи к панелям как на эквиваленты номеров страниц. Вряд ли вы при чтении обращаете внимание на номера страниц, и само по себе явление нумерации вполне обыденное, однако в некоторых случаях номера страниц могут сыграть важную роль, если вам нужно сослаться на определенное место в книге или статье.

Также необходимо учитывать и то, как сочетаются друг с другом отдельные панели составной визуализации. Вы можете нарисовать набор панелей, каждая из которых сама по себе будет отлично смотреться, однако, объединив их в график, вы получите вместо цельного изображения разнобой. В частности, визуальный язык вашего рисунка должен быть непротиворечив. Под визуальным языком я подразумеваю цвета, символы, шрифты и т. д. — все, что мы используем для отображения данных. Если говорить в двух словах, согласованность визуального языка означает, что одни и те же вещи должны выглядеть одинаково или, по крайней мере, иметь существенное сходство.

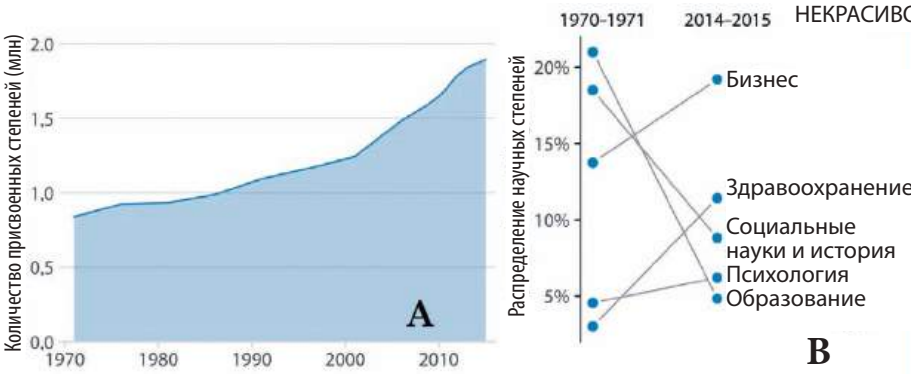


Рис. 20.6. Вариация рис. 20.5 с плохой маркировкой. Метки слишком большие и толстые, набраны неправильным шрифтом и расположены в неудобном месте. Маркировка на данном графике выполнена в виде заглавных букв, и, несмотря на то что это вполне допустимая и распространенная практика, на данном графике все же нарушен принцип визуальной согласованности меток. В этой книге я придерживаюсь правила, что на многопанельных графиках в подписях и метках должны использоваться строчные буквы, и поэтому данный график не согласуется с другими диаграммами из этой книги. Источник: National Center for Education Statistics

Давайте рассмотрим пример, нарушающий данный принцип. Рис. 20.7 представляет собой график, состоящий из трех панелей, которые совокупно визуализируют набор данных о физиологии и телосложении спортсменов мужского и женского пола. Панель A показывает количество мужчин и женщин в наборе данных, панель B показывает количество красных и белых кровяных телец у мужчин и женщин, а панель C показывает процентное содержание жира у мужчин и женщин в зависимости от вида спорта, которым они занимаются. По отдельности каждая панель выглядит вполне приемлемо.

Однако если мы объединим их в диаграмму, то увидим, что визуально каждая панель так и остается сама по себе, потому что использует свой, отличный от других панелей визуальный язык. Во-первых, панель A использует один и тот же синий цвет как для спортсменов, так и для спортсменок, панель B отмечает цветом только спортсменов-мужчин, а панель C выделяет цветом только спортсменок. Кроме того, панели B и C вводят дополнительные цвета, однако на разных панелях использованы разные варианты. Было бы лучше, если бы спортсменам обоих полов на каждой панели, включая и панель A, соответствовали одни и те же два цвета. Во-вторых, на графиках на панелях A и B женщины указаны слева, а мужчины — справа, однако на панели C все наоборот. Порядок расположения мужских и женских элементов коробчатой диаграммы на панели C должен быть изменен, чтобы соответствовать панелям A и B.

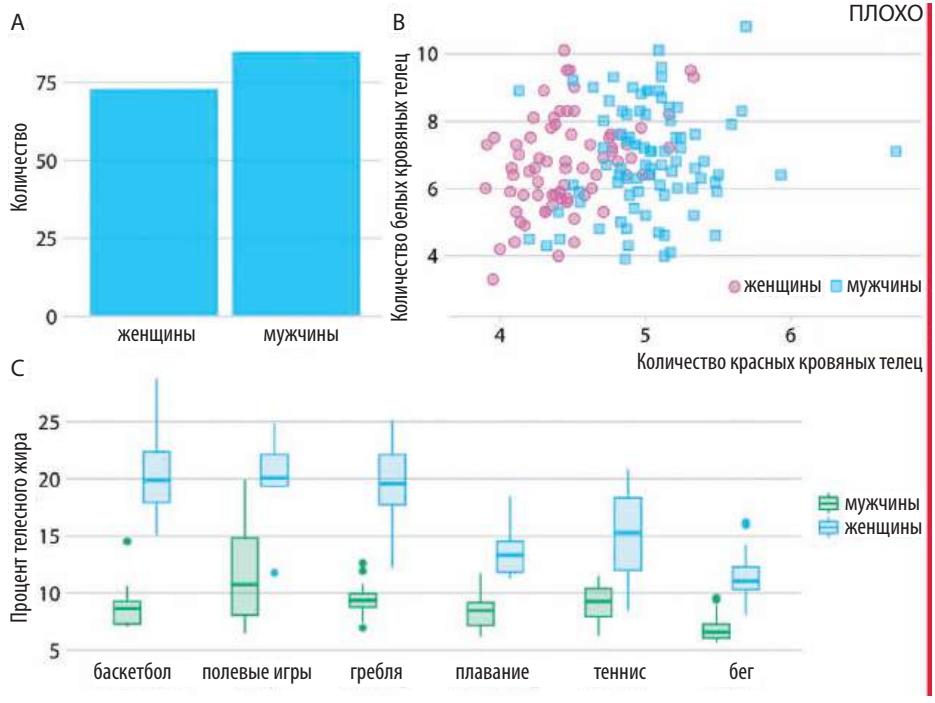


Рис. 20.7. Сравнение физиологии и телосложения спортсменов мужского и женского пола. A. Набор данных о профессиональных атлетах включает данные о 73 женщинах и 85 мужчинах. B. Атлеты мужского пола, как правило, имеют более высокое количество эритроцитов (красные кровяные тельца, единица измерения — 10^{12} на литр), чем женщины, однако в количестве лейкоцитов явных различий нет (белые кровяные тельца, единица измерения — 10^{9} на литр). C. Среди спортсменов, занимающихся одним и тем же видом спорта, у мужчин процент жира ниже, чем у женщин.
Это изображение относится к категории «плохих», так как части A, B и C не используют согласованный визуальный язык. Источник: [Telford and Cunningham, 1991]

На рис. 20.8 все указанные проблемы устранены: женщины стабильно обозначены оранжевым цветом, всегда расположены слева от мужчин, которые, в свою очередь, показаны синим цветом. Обратите внимание, насколько легче воспринимается этот рисунок, в отличие от рис. 20.7. Согласованный визуальный язык позволяет затрачивать минимум умственных усилий для определения, например, того, какие визуальные элементы на разных панелях представляют женщин, а какие — мужчин. Еще одной проблемой рис. 20.7 является то, что при первом взгляде на него можно сделать вывод, что мужчины, как правило, имеют более высокий процент жира в организме, чем женщины. Также обратите внимание, что рис. 20.8 для пояснения нужна только одна легенда, а рис. 20.7 — две. Благодаря использованию единого визуального языка эта же легенда подходит для панелей B и C.

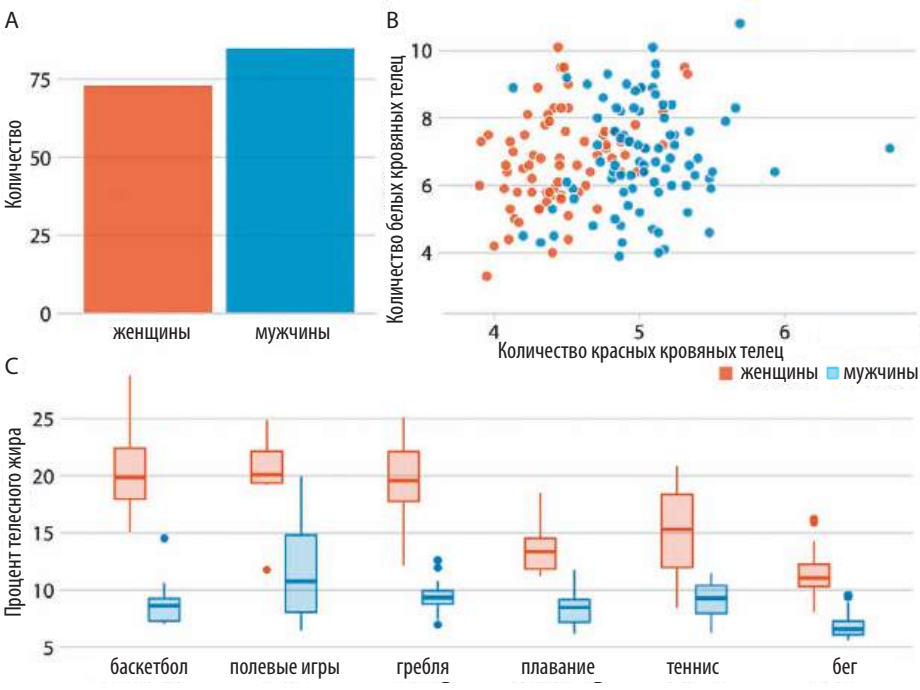


Рис. 20.8. Сравнение физиологии и телосложения спортсменов мужского и женского пола. На данном рисунке показаны те же самые данные, что и на рис. 20.7, однако здесь используется общий визуальный язык. Данные о спортсменах-женщинах всегда отображаются слева от соответствующих данных о спортсменах-мужчинах, а пол имеет одинаковую цветовую кодировку во всех элементах графика.
Источник: [Telford and Cunningham, 1991]

И напоследок следует упомянуть еще один немаловажный аспект — выравнивание панелей в составном графике. Оси и другие графические элементы отдельных панелей должны быть выровнены относительно друг друга.

Добиться правильного выравнивания может быть довольно сложно, особенно если все панели были созданы отдельно друг от друга, например, разными людьми и/или в разных программах, после чего склеены в программе обработки изображений. Чтобы подчеркнуть важность этого вопроса, я подготовил рис. 20.9, являющийся вариантом рис. 20.8, на котором все элементы немного сдвинуты. Чтобы акцентировать ваше внимание на проблеме выравнивания, я добавил осевые линии ко всем панелям рис. 20.9. Обратите внимание, что никакие две произвольно взятые осевые линии не выровнены относительно друг друга.

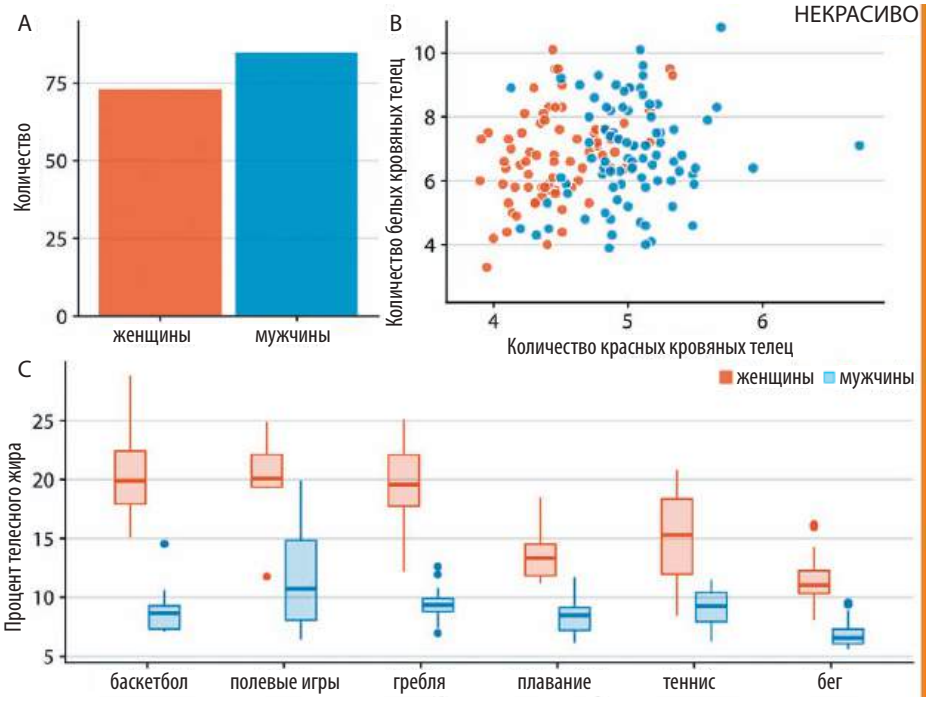


Рис. 20.9. Вариация рис. 20.8, где все элементы слегка сдвинуты друг относительно друга. Отсутствие выравнивания выглядит очень неэстетично, и этого следует всячески избегать. Источник: [Telford and Cunningham, 1991]

Глава 21

Заголовки, подписи и таблицы

Визуализация данных — это не произведение искусства, которое интересно только из-за его эстетических особенностей. Напротив, цель визуализации состоит в отображении информации и донесении ее смысла до зрителя. Чтобы гарантированно достичь этой цели, визуализируемые данные следует поместить в контекст, снабдив их заголовками, подписями и другими различными видами аннотаций. В этой главе я расскажу, как правильно называть и маркировать рисунки. Кроме того, мы поговорим о том, как представить данные в виде таблиц.

Заголовки и подписи к рисункам

Одним из важнейших компонентов любого графика является его заголовок. Каждый рисунок должен как-то называться. Собственно, прежде всего благодаря названию читатель понимает, о чем эта визуализация и в чем ее смысл. Однако заголовок рисунка не обязательно должен располагаться там, где вы ожидаете его увидеть. Взгляните на рис. 21.1. Его название — «Коррупция и человеческое развитие: в наиболее развитых странах наименьший уровень коррупции». В этом случае заголовок расположен не сверху рисунка, а идет самым первым элементом блока подписей, расположенного под графиком. Именно этот стиль я использую на протяжении всей книги. В моих визуализациях нет интегрированных заголовков, но все графики снабжены отдельными подписями. Исключение составляют примеры стилизованных графиков в главе 4, на которых присутствуют заголовки, но нет подписей.

Существует и другой подход: заголовок рисунка и такие элементы, как ссылка на источник данных, можно добавить в основную часть визуализации (рис. 21.2). Сравнивая между собой рис. 21.1 и 21.2, вы можете посчитать последний более привлекательным и удивиться, почему везде в этой книге я использую стиль рис. 21.1. Причина этого в том, что каждый из этих стилей должен использоваться в рамках своей области применения, а рисунки со встроенными заголовками не подходят для использования в книгах. Основной принцип, который следует неукоснительно соблюдать, создавая график, звучит очень просто: у рисунка может быть только один заголовок.

Он должен быть или интегрирован в основную часть рисунка, или быть первым элементом в подписи к диаграмме. И если публикация спроектирована таким образом, что основной блок информации о рисунке находится внизу графика, значит, и заголовок должен находиться там же. По этой причине при публикации печатных книг или статей мы обычно не помещаем названия в границы рисунков. При этом рисунки со «встроенными» заголовками, подзаголовками и ссылками на источники данных уместны, но лишь в тех случаях, когда они предназначены для использования в качестве самостоятельных единиц инфографики или для размещения в социальных сетях или на веб-страницах без сопроводительных подписей.

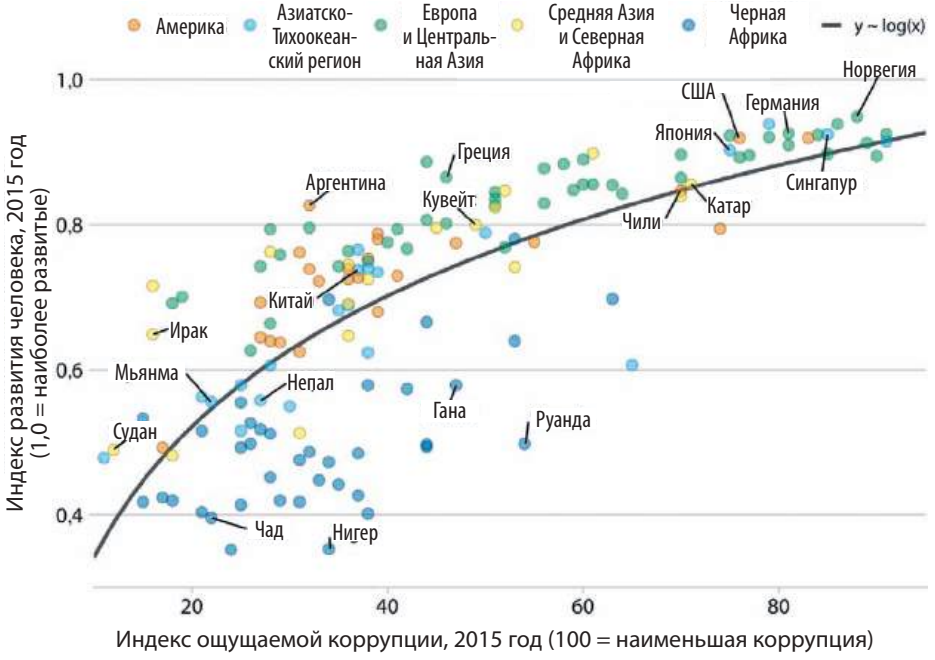


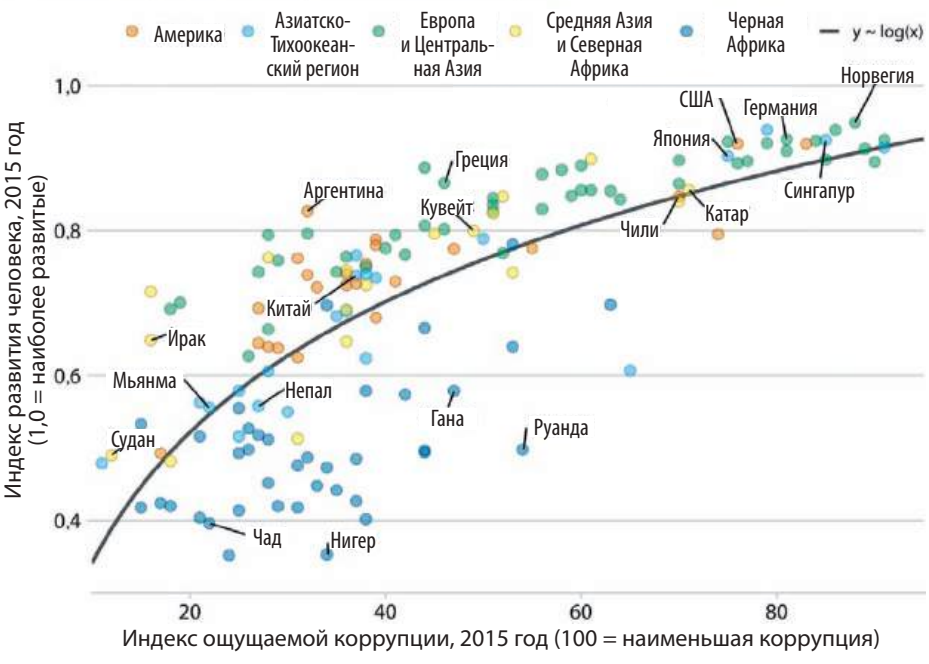
Рис. 21.1. Коррупция и человеческое развитие: в наиболее развитых странах наименьший уровень коррупции. Оригинальная концепция: [The Economist Online, 2011]. Источники: Transparency International & UN Human Development Report

Если в макете документа под каждым рисунком присутствует блок подписей, заголовки рисунков должны находиться в подписи в качестве первого элемента, а не размещаться над рисунком.

Одна из наиболее распространенных ошибок, которые мне встречаются в подписях к графикам, — это отсутствие правильного названия рисунка в качестве первого элемента подписи. Взгляните на рис. 21.1. Подпись

к диаграмме начинается с фразы «Коррупция и развитие человека». Обратите внимание, что она не начинается с «*Этот рисунок показывает*, как коррупция связана с развитием человека». Первая часть подписи — это всегда именно заголовок, а не описание содержимого рисунка. Заголовок не обязательно должен иметь форму полного предложения, однако короткие предложения, звучащие как ясное утверждение, будут в тему. Например, на рис. 21.1 такой заголовок, как «Самые развитые страны являются наименее коррумпированными», выглядел бы абсолютно уместно.

Коррупция и человеческое развитие:
в наиболее развитых странах наименьший уровень коррупции



Источники: Transparency International & UN Human Development Report

Рис. 21.2. Вариант рис. 21.1 в формате инфографики. В рисунок включены заголовок, подзаголовок и информация об источниках. Такой график можно разместить в интернете как есть или использовать любым другим способом без отдельного блока подписи

Названия осей и легенд

Как и самому графику, его осям и легендам тоже нужны заголовки. (В разговорной речи заголовки осей часто называются подписями или метками осей.) С помощью заголовков и меток осей и легенд читатель понимает, какие именно данные изображены на графике и каким визуальным элементам они соответствуют.

Чтобы привести пример графика, в котором все оси и легенды соответствующим образом помечены и озаглавлены, я взял набор данных о голубых сойках, который подробно обсуждался в главе 11, и визуализировал его в виде пузырьковой диаграммы (рис. 21.3). Из названий осей следует, что по оси *x* указывается масса тела в граммах, а по оси *y* — длина головы в миллиметрах. Аналогично заголовок легенды показывает, что цвет точек обозначает пол птицы, а размер точек — размер черепа птицы (в миллиметрах). Хочу подчеркнуть, что для всех числовых переменных (масса тела, длина головы и размер черепа) в соответствующих заголовках указываются не только названия переменных, но и их единицы измерения. Это хорошая практика, и ее следует придерживаться всегда, когда это возможно. У категориальных переменных (например, пол) единиц измерения нет.

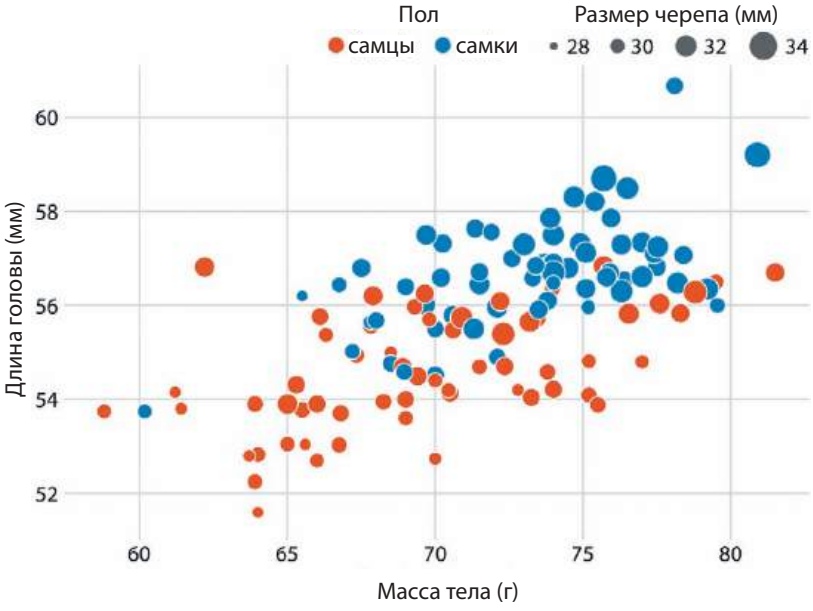


Рис. 21.3. Отношение длины головы к массе тела у 123 голубых соек. Пол птиц отображен цветом точки, а размер черепа птиц — размером точки. Измерения длины головы включают в себя длину клюва, в то время как измерения размера черепа — нет. Источник: Keith Tarvin, Oberlin College

В некоторых случаях, однако, названия осей и легенд можно не указывать: конкретно — когда метки являются полноценными пояснениями. Например, легенда, показывающая две разноцветные точки с надписями «женский» и «мужской», без дополнительных объяснений позволяет понять, что цвет кодирует пол. Нам не нужно добавлять надпись «пол», чтобы пояснить этот факт, и, как вы могли заметить, в этой книге я часто опускал название легенды для пола или гендера (см., например, рис. 5.10, 11.2 или 20.1). Точно так

же, как правило, не требуют пояснения названия стран (рис. 5.11), названия фильмов (рис. 5.1) или годы (рис. 21.4).

Рис. 21.4. Изменения во времени цен акций четырех крупных технологических компаний. Цены акций каждой компании были нормализованы и приняты за 100 в июне 2012 года. Эта визуализация является слегка измененной версией рис. 19.6 в главе 19. Здесь у оси *x*, представляющей время, не указан заголовок. Из контекста очевидно, что 2013, 2014 и т. д. обозначают годы. Источник: Yahoo! Finance

Тем не менее, принимая решение не указывать названия осей или легенды, следует быть осторожными, поскольку очень легко ошибиться в оценке того, что для читателя будет очевидно из контекста, а что нет. В популярной прессе очень часто встречаются графики, где названиями осей пренебрегли настолько, что мне становится не по себе. Например, в некоторых публикациях можно увидеть рисунок, подобный рис. 21.5, автор которого полагал, что смысл осей должен быть очевиден из названия и подзаголовка самой диаграммы (здесь: «Изменения во времени цен акций четырех крупных технологических компаний» и «Цены акций каждой компании были нормализованы и приняты за 100»).

Я не считаю, что значения осей можно понять из контекста. Поскольку заголовок, как правило, не включает такие слова, как «ось *x*/*y* показывает», для интерпретации рисунка всегда требуется сделать определенное количество догадок. По моему опыту, графики без должным образом обозначенных осей обычно оставляют у зрителя ноющее чувство неопределенности. Даже если я на 95% уверен, что понимаю, о чем этот график, это все же не полная уверенность. Я считаю плохой практикой заставлять читателей гадать, что автор хотел сказать своим рисунком. Зачем вам создавать чувство неопределенности у вашей аудитории?

Рис. 21.5. Изменения во времени цен акций четырех крупных технологических компаний. Цены акций каждой компании были нормализованы и приняты за 100 в июне 2012 года. Эта вариация рис. 21.4 относится к категории «плохих», потому что ось *y* не имеет заголовка, а значения по ней совершенно неочевидны. Источник: Yahoo! Finance

Рис. 21.6. Изменения во времени цен акций четырех крупных технологических компаний. Цены акций каждой компании были нормализованы и приняты за 100 в июне 2012 года. Данная вариация рис. 21.4 относится к категории «некрасивых», потому что на ней присутствует слишком много меток и подписей. В частности, указание единиц измерения («годы») для значений по оси *x* — громоздко и бесполезно. Источник: Yahoo! Finance

С другой стороны, переусердствовать с маркировкой тоже нетрудно. Если в легенде перечислены названия четырех известных компаний, то озаглавливать легенду «компании» совершенно излишне и абсолютно бесполезно

(рис. 21.6). Точно так же, хоть мы и должны указывать единицы измерения для всех количественных переменных, если ось *x* показывает несколько последних лет, называть ее «время (годы)» по большому счету нелепо.

Наконец, в некоторых случаях допустимо опускать не только заголовок оси, но и всю ось целиком. Так, например, у круговых диаграмм, как правило, нет явных осей (например, рис. 9.1), равно как и у древовидных карт (см. рис. 10.4). У мозаичных графиков или гистограмм может отсутствовать одна из осей или даже обе, если смысл графика понятен и так (см. рис. 5.10 и 10.3). Отсутствие на диаграмме явных осей с собственными метками и метками единичных отрезков сигнализирует читателю, что качественные характеристики графика важнее конкретных значений данных.

Таблицы

Таблицы являются важным инструментом визуализации данных. Однако из-за кажущейся простоты их зачастую незаслуженно игнорируют. В этой книге мы уже встречались с данным способом визуализации, например таблицы 5.1, 6.1 и 18.1. Уделите им немного времени и рассмотрите, как они отформатированы, а затем сравните их с таблицей, которую недавно создали вы или ваш(а) коллега. Наверняка вы увидите между ними существенные различия. По моему опыту, если вас никто специально не обучал форматированию таблиц, вам вряд ли удастся правильно оформить таблицу, полагаясь лишь на интуицию. В документах, которых не касалась рука опытного специалиста, плохо отформатированные таблицы встречаются даже чаще, чем плохо спроектированные графики. Кроме того, большинство программ, в которых обычно создаются таблицы, предлагают по умолчанию такие значения параметров, использовать которые я не рекомендую. Например, моя версия Microsoft Word содержит 105 предустановленных стилей таблиц, и по крайней мере 70 или 80 из них нарушают некоторое количество правил оформления таблиц, о которых мы и поговорим в этом разделе. Таким образом, если вы наугад выберете макет таблицы Microsoft Word, вероятность того, что вы выберете проблемный формат, составляет приблизительно 80%. А выбранный по умолчанию стиль таблицы означает, что на выходе вы гарантированно получите плохо отформатированную таблицу.

Далее приведены некоторые ключевые правила создания таблиц.

1. Не используйте вертикальные линии.
2. Между строками данных не должно быть горизонтальных линий. (Исключением являются горизонтальные линии, играющие роль разделителя между строкой заголовка и первой строкой данных или рамки для всей таблицы.)
3. Текстовые столбцы следует выравнивать по левому краю.

4. Числовые столбцы должны быть выровнены по правому краю и содержать одинаковое количество десятичных цифр.
5. Столбцы, содержащие одиночные символы, должны быть выровнены по центру.
6. Поля заголовка должны быть выровнены в соответствии с типом значения столбцов; то есть заголовок текстового столбца следует выравнивать по левому краю, а заголовок числового столбца — по правому краю.

На рис. 21.7 показаны четыре разных варианта таблицы 5.1, два из которых (A, B) нарушают часть описанных выше правил, а другие два (C, D) — нет.

A НЕКРАСИВО

Место	Название фильма	Сборы
1	«Звездные войны: Последние джедаи»	$71 565 498
2	«Джуманджи: Зов джунглей»	$36 169 328
3	«Идеальный голос 3»	$19 928 525
4	«Величайший шоумен»	$8 805 843
5	«Фердинанд»	$7 316 746

B НЕКРАСИВО

Место	Название фильма	Сборы
1	«Звездные войны: Последние джедаи»	$71 565 498
2	«Джуманджи: Зов джунглей»	$36 169 328
3	«Идеальный голос 3»	$19 928 525
4	«Величайший шоумен»	$8 805 843
5	«Фердинанд»	$7 316 746

C

Место	Название фильма	Сборы
1	«Звездные войны: Последние джедаи»	$71 565 498
2	«Джуманджи: Зов джунглей»	$36 169 328
3	«Идеальный голос 3»	$19 928 525
4	«Величайший шоумен»	$8 805 843
5	«Фердинанд»	$7 316 746

D

Место	Название фильма	Сборы
1	«Звездные войны: Последние джедаи»	$71 565 498
2	«Джуманджи: Зов джунглей»	$36 169 328
3	«Идеальный голос 3»	$19 928 525
4	«Величайший шоумен»	$8 805 843
5	«Фердинанд»	$7 316 746

Рис. 21.7. Примеры хорошего и плохого форматирования таблиц с использованием данных из табл. 5.1 из главы 5. A. Эта таблица нарушает большинство правил форматирования таблиц, включая использование вертикальных линий, горизонтальные линии между строками данных и использование центрированных столбцов данных. B. Эта таблица страдает от всех проблем (A), а также создает визуальный шум, чередуя строки с очень темным и очень светлым фоном. Кроме того, заголовок таблицы почти неотличим от тела таблицы. C. Это правильно отформатированная таблица с минималистичным дизайном. D. Использование цвета может быть эффективным инструментом для группировки данных в строки, однако различия в окраске строк должны быть незначительными. Заголовок таблицы может иметь более яркий цвет. Источник: Box Office Mojo. Используется с разрешения источника

Когда таблицы рисуют с разделением строк при помощи горизонтальных линий, конечной целью обычно является помощь читателю сосредоточиться на той или иной строке. Однако, если таблица не очень широкая и достаточно разреженная, подобный визуальный «костыль», как правило, не требуется. Сравните эту ситуацию с книгой, где нам вряд ли придет в голову нарисовать горизонтальные линии между строками в обычном тексте. Горизонтальные (или вертикальные) линии вносят визуальный беспорядок. Сравните

части A и C рис. 21.7. Часть C гораздо легче читать, чем часть A. Если при чтении таблицы вы ощущаете, что вам недостает разделения строк, то лучшим вариантом его реализации будет окраска чередующихся строк в более светлые и более темные оттенки, поскольку такой вариант не станет источником помех при чтении таблицы (рис. 21.7D).

Наконец, между рисунками и таблицами есть ключевое различие, которое заключается в расположении заголовка относительно элемента, к которому он относится. Для рисунков принято размещать заголовок внизу, а для таблиц — вверху. Данное правило не является простым формализмом. Расположение подписи обусловлено тем, как люди обрабатывают рисунки и таблицы. В случае рисунков читатели, как правило, сначала смотрят на изображение, а затем читают заголовок, поэтому будет логичнее поставить заголовок под картинкой. Таблицы же, как правило, обрабатываются подобно тексту — сверху вниз, и поэтому изучение содержимого таблицы до того, как вы ознакомитесь с ее заголовком, вряд ли будет иметь смысл. Именно поэтому подписи к таблице располагают над ней.

Глава 22

Баланс данных и контекста

В любой визуализации есть как графические элементы, которые отображают данные, так и вспомогательные элементы, которые с данными не ассоциированы. К первым относятся такие элементы, как точки на диаграмме рассеяния, столбцы на гистограмме или заштрихованные области на тепловой карте. Ко второй категории относятся оси графиков, отметки и подписи осей, условные обозначения и аннотации графиков. Эти элементы, как правило, описывают контекст данных и/или визуальной структуры графика. При проектировании графика лучше всего сразу определить, какая часть изображения будет отвечать за данные, а какая — за контекст (также давайте вспомним о принципе пропорциональной заливки — его можно найти в главе 16). Наиболее часто рекомендуют уменьшать количество элементов, не относящихся к данным, благодаря чему итоговая визуализация будет выглядеть более чисто и эстетично. Тем не менее не следует слишком увлекаться минимализацией графика, поскольку контекст и визуальная структура тоже важны, и, если чрезмерно снизить количество элементов, которые их представляют, может получиться так, что графики станут слишком сложными для интерпретации, запутанными или попросту неубедительными.

Предоставление подобающего объема контекста

Идея о том, что из различения элементов, отображающих данные, и элементов, не связанных с ними, можно извлечь пользу, была популяризирована Эдвардом Тафти в его книге «Визуальное отображение количественной информации» [Tufte, 2001]. Тафти вводит понятие «соотношение данных и чернил» (data-ink ratio), которое он определяет как «пропорцию чернил на графике, предназначенную для неповторяющегося отображения информации, содержащей данные». Далее он приводит следующий тезис (выделение мое):

«Максимизируйте соотношение данных и чернил *в пределах разумного*».

Я подчеркнул окончание фразы — «в пределах разумного», — потому что это очень важное уточнение, про которое часто забывают. На самом деле я думаю, что и сам Тафти забывает об этом в идущей далее части своей книги, где он выступает за чрезмерно минималистичные подходы, которые, на мой взгляд, никак

нельзя назвать ни изящными, ни простыми для понимания. Если мы интерпретируем фразу «максимизировать соотношение данных и чернил» как «убрать беспорядок и стремиться к чистому и элегантному дизайну», то я думаю, что этот совет абсолютно разумен. Но если мы интерпретируем это как «делайте все возможное, чтобы удалить элементы, не относящиеся к данным», то это повлечет за собой сомнительные дизайнерские решения. Если мы зайдем слишком далеко в этом направлении, наши визуализации станут откровенно некрасивыми. По счастью, между крайностями есть широкое пространство вариантов дизайна, которые могут быть вполне приемлемыми в зависимости от потребностей.

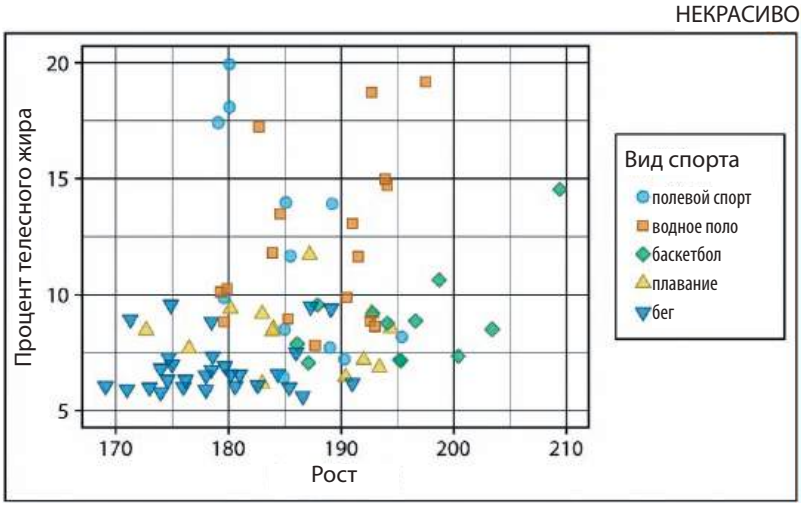


Рис. 22.1. Соотношение роста и процента жира у профессиональных австралийских спортсменов-мужчин. Каждая точка представляет одного спортсмена. Этот график отводит слишком много «чернил» элементам, не относящимся к визуализируемому набору данных: рамки вокруг графика, панели и легенды. Кроме того, координатная сетка слишком бросается в глаза, отвлекая внимание от точек данных.
Источник: [Telford and Cunningham, 1991]

Чтобы обозначить «границы крайностей», давайте рассмотрим визуализацию, которая содержит слишком много элементов, не имеющих отношения к отображаемым данным (рис. 22.1). Цветные точки в области графика (область графика — центральная область, очерченная рамкой и содержащая точки данных) — это «чернила данных». Все остальное — «чернила», которыми нанесена информация, не имеющая отношения к данным. В частности, к ним относятся рамка вокруг визуализации в целом, рамка вокруг области графика и рамка вокруг легенды. Ни один из этих элементов не нужен. Также на фоне присутствует хорошо заметная сетка, которая отвлекает внимание от содержимого графика. Рис. 22.2 представляет собой вариант рис. 22.1, на котором отсутствуют рамки и второстепенные линии сетки, а основные линии сетки

выделены светло-серым цветом. В этой версии рисунка точки данных видны гораздо лучше и воспринимаются как наиболее важная компонента графика.

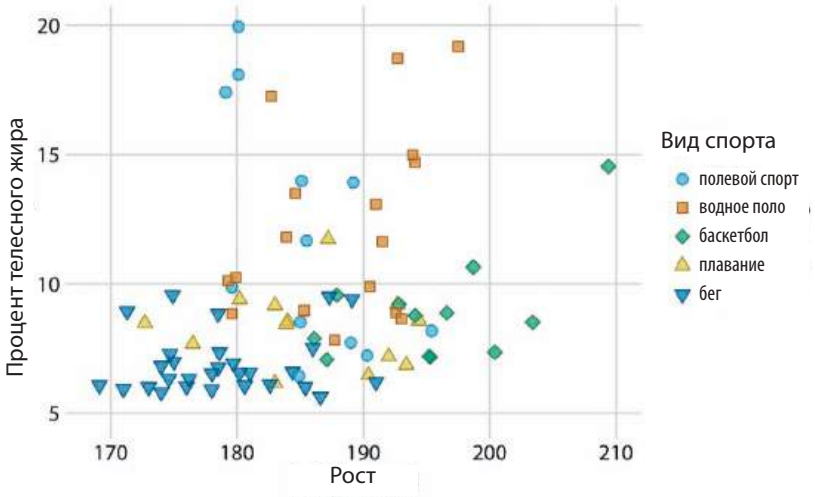


Рис. 22.2. Соотношение роста и процента жира у профессиональных австралийских спортсменов-мужчин. Данный график является визуально улучшенной версией рис. 22.1. Удалены ненужные рамки и второстепенные линии сетки, а основные линии сетки нанесены светло-серым цветом, чтобы сделать точки данных более заметными. Источник: [Telford and Cunningham, 1991]

Переходя к другой крайности, мы можем предельно минимизировать график на рис. 22.2, результатом чего станет диаграмма с рис. 22.3. На этом графике метки и заголовки осей выглядят настолько тускло, что прочитать их почти невозможно. Беглого взгляда на график недостаточно, чтобы понять, о чем эта диаграмма. Читатель увидит ее как набор точек, плавающих в пространстве. Более того, аннотации к легендам настолько плохо видны, что точки в легенде можно запросто принять за точки данных. Из-за того, что между областью графика и легендой нет визуального разделения, этот эффект становится еще сильнее. Обратите внимание, как фоновая сетка на рис. 22.2 визуально фиксирует точки в пространстве и отделяет область данных от области легенды. На рис. 22.3 эти эффекты отсутствуют.

На рис. 22.2 я использую открытую фоновую сетку без линий осей или рамки вокруг области графика. Мне нравится этот дизайн, поскольку он дает читателю понять, что диапазон возможных значений данных выходит за пределы видимых значений осей. Несмотря на то что на рис. 22.2 нет ни одного спортсмена с ростом выше 210 см, такой человек вполне может существовать. Однако некоторые авторы предпочитают обозначать границы панели графика, рисуя вокруг нее рамку (рис. 22.4). Оба варианта одинаково приемлемы, наличие или отсутствие рамки определяется личным мнением автора графика.

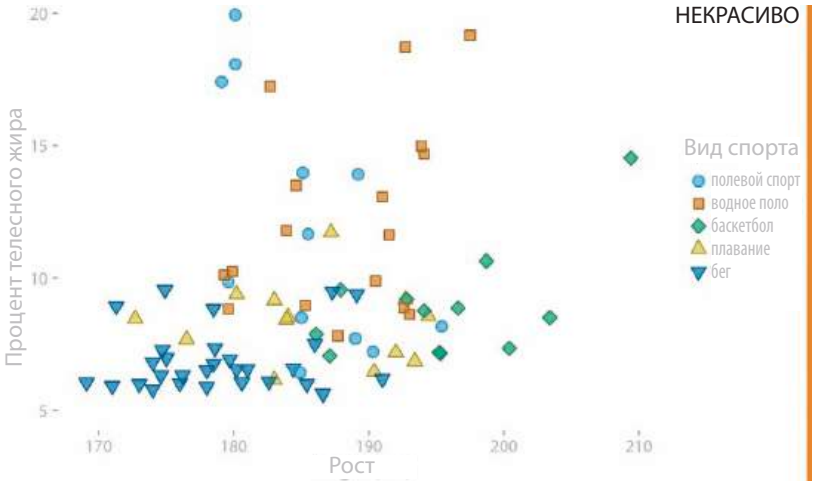


Рис. 22.3. Соотношение роста и процента жира у профессиональных австралийских спортсменов-мужчин. На этом примере применение правила исключения элементов, не связанных с отображаемыми данными, доведено до абсурда. Засечки, метки и подписи осей очень бледные и едва различимы. Точки данных будто плавают в пустом пространстве. Точки в легенде визуально плохо отделены от точек данных, из-за чего случайный наблюдатель может подумать, что они тоже относятся к графику. Источник: [Telford and Cunningham, 1991]

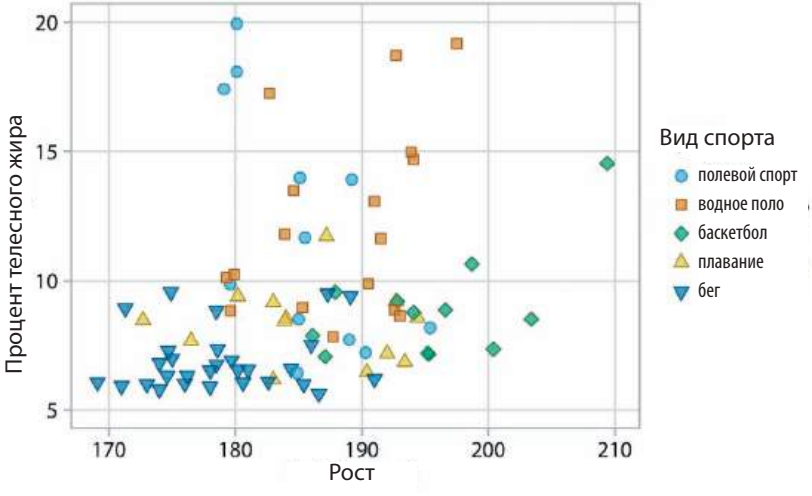


Рис. 22.4. Соотношение роста и процента жира у профессиональных австралийских спортсменов-мужчин. Этот рисунок полностью повторяет рис. 22.2, за исключением того, что область графика заключена в рамку, которая визуально отделяет легенду от данных. Источник: [Telford and Cunningham, 1991]

Следует отметить, что одно из преимуществ версии в рамке заключается в том, что она визуально отделяет легенду от панели самого графика.

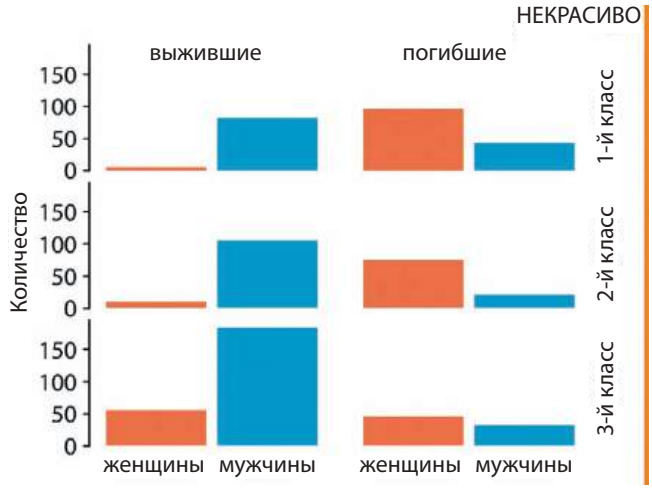


Рис. 22.5. Информация о выживаемости среди пассажиров «Титаника», разбитая по полам и классам кают. Данная малая панельная визуализация чересчур минималистична. Из-за того что у панелей нет явных границ, сложно понять, какие элементы к чему относятся. Кроме того, отдельные столбцы не привязаны к базовой линии, из-за чего кажется, что они плавают в пространстве. Источник: Encyclopedia Titanica

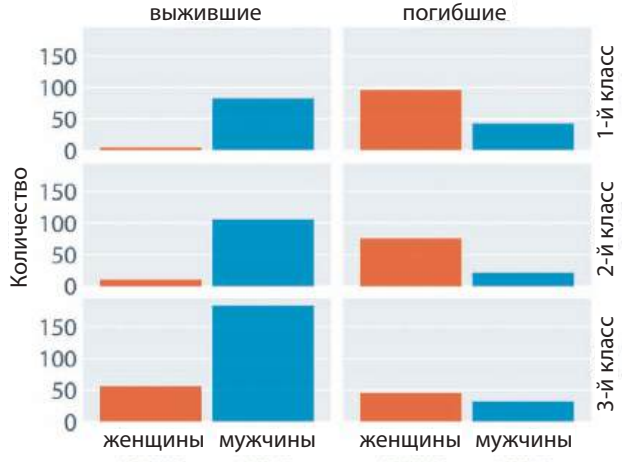


Рис. 22.6. Информация о выживаемости среди пассажиров «Титаника», разбитая по полам и классам кают. Данный график является улучшенной версией рис. 22.5. Поскольку теперь у графиков есть фон, все шесть панелей хорошо визуально различимы (выжившие или умершие; в 1-м, 2-м или 3-м классах). Тонкие горизонтальные линии на заднем плане позволяют определять высоту столбцов и облегчают сравнение высоты столбцов в различных панелях. В качестве альтернативы мы могли бы поместить рамку вокруг каждой из панелей графика и использовать серые полосы для выделения переменных группировки (см. рис. 20.1). Источник: Encyclopedia Titanica

Рисунок, который содержит слишком мало элементов, не имеющих отношения к данным, часто воспринимается зрителем в виде набора элементов, плавающих в пространстве, без какой-либо связи друг с другом или ссылки на что-либо вовне. Особенно остро эта проблема проявляется на малых панельных визуализациях. На рис. 22.5 показан пример такой визуализации, сравнивающий шесть различных столбчатых диаграмм, но, пожалуй, он больше похож на произведение современного искусства, нежели на осмысленную визуализацию данных. Столбцы не привязаны к базовой линии, а у отдельных панелей графика нет явно очерченных границ. Для решения этих проблем каждую отдельную панель можно покрасить в светло-серый фон и добавить на нее тонкие горизонтальные линии сетки (рис. 22.6).

Фоновые сетки

Линии сетки, расположенные на заднем плане графика, помогают читателю различать значения данных и сравнивать значения в одной части графика со значениями в другой. Однако они могут быть и источником визуального шума, особенно если они слишком выделяются или расположены слишком близко друг к другу. Единого мнения о том, стоит использовать линии сетки или нет, не существует, а в случаях, когда решено их использовать, также нет и рекомендаций о том, как их лучше форматировать и с какой густотой наносить. В этой книге я использую множество различных стилей сеток, чтобы подчеркнуть, что не существует какого-то универсального, пригодного на все случаи жизни варианта.

Программное обеспечение ggplot2 популяризировало стиль, основанный на заметно выраженной фоновой сетке из белых линий на сером фоне. На рис. 22.7 приведен пример использования этого стиля. График показывает изменения цен акций четырех крупных технологических компаний на пятилетнем периоде, с 2012 по 2017 год. Я прошу прощения у автора ggplot2 Хэдли Уикхэма, к которому я испытываю огромное уважение, за то, что не считаю вариант «белая сетка на сером фоне» привлекательным. На мой взгляд, серый фон может отвлекать от фактических данных, а сетка с основными и второстепенными линиями способна «зашумлять» пространство. Кроме того, меня смущают серые квадраты в легенде.

Серый фон хорош тем, что он как придает графику ощущение единой визуальной сущности, так и не позволяет диаграмме превратиться в белый прямоугольник, окруженный текстом темного цвета [Wickham, 2016]. Я полностью согласен с первым пунктом и именно по этой причине сделал на рис. 22.6 заливку серым цветом. Что касается второго пункта, я хотел бы обратить ваше внимание, что воспринимаемая темнота текста будет зависеть от шрифта, его размера и расстояния между строками, а воспринимаемая

темнота рисунка будет зависеть от абсолютного количества и цвета элементов, включая все «чернила», используемые для данных. Текст научной статьи, набранный убористым шрифтом Times New Roman размера 10, будет восприниматься более темным, чем текст бульварного романа, набранный шрифтом Palatino с интервалом в полторы строки и размером 14. Аналогично диаграмма рассеяния из пяти точек желтого цвета будет выглядеть намного светлее, чем диаграмма рассеяния, состоящая из 10 000 точек черного цвета. Если вы решили сделать на графике серый фон, обязательно учитывайте интенсивность цветов элементов, которые расположены на переднем плане, а также особенности раскладки элементов и типографику текста вашей будущей визуализации. Какой именно оттенок серого цвета выбрать для вашего фона, должно целиком определяться этими соображениями. В противном случае вы можете столкнуться с тем, что ваши рисунки будут выглядеть как темные прямоугольники, окруженные более светлым текстом. Кроме того, имейте в виду, что цвета, которые вы используете для построения графиков, должны сочетаться с серым фоном. Человек по-разному воспринимает один и тот же цвет в зависимости от фона, и серый фон требует использования на переднем плане более темных и более насыщенных цветов, чем фон белого цвета.

Рис. 22.7. Изменения во времени цен акций четырех крупных технологических компаний. Цены акций каждой компании были нормализованы и приняты за 100 в июне 2012 года. Этот рисунок имитирует стандартный дизайн, предлагаемый по умолчанию ggplot2: белые линии на сером фоне. На мой взгляд, линии сетки на этом графике более заметны, чем линии данных, и в результате получается плохо сбалансированное изображение, где основная информация теряется на общем фоне. Источник: Yahoo! Finance

Мы можем и сделать все наоборот — полностью убрать фон и линии сетки (рис. 22.8). Но в этом случае нам понадобятся видимые линии осей, чтобы

обрамить график и придать ему вид единого целого. В данном конкретном случае я считаю этот подход менее удачным и поэтому отнес эту диаграмму к категории «плохих». Если на графике не будет никакой фоновой сетки, кривые будут визуально «плыть», из-за чего будет сложнее сопоставлять крайние значения справа с отметками на оси слева.

Рис. 22.8. Индексированные цены акций четырех крупных технологических компаний. В этой вариации рис. 22.7 линии данных недостаточно сильно «привязаны». Из-за этого стало сложнее понять, насколько значения отклоняются от базового значения индекса (100) в конце рассматриваемого интервала времени. Источник: Yahoo! Finance

В самом крайнем случае нам нужна хотя бы одна горизонтальная линия для привязки данных. Поскольку в июне 2012 года цены на акции были приняты за 100, маркировка этого значения с помощью линии уровня $y = 100$ значительно упрощает восприятие рисунка (рис. 22.9). В качестве альтернативы мы можем попробовать воспользоваться «сеткой», состоящей только из горизонтальных линий. Для графика, в котором нас интересует только изменение значений по оси *y*, вертикальные линии сетки не нужны. Более того, во многих сценариях будет достаточно нарисовать только линии, расположенные на отметках по главной оси, поэтому мы можем или совсем убрать оси координат, или сделать их очень тонкими, поскольку горизонтальные линии уже сами по себе обозначают границы графика (рис. 22.10).

Столь минималистичная сетка координат обычно создается путем рисования линий, перпендикулярных направлению изменения интересующих нас значений. Поэтому если вместо графика изменения цены акций с течением времени мы визуализируем, например, пятилетнее изменение цены в виде горизонтальных столбцов, то наиболее подходящим вариантом будут вертикальные, а не горизонтальные линии сетки координат (рис. 22.11).

Рис. 22.9. Индексированные цены акций четырех крупных технологических компаний. Добавленная на рис. 22.8 тонкая горизонтальная линия уровня в значении индекса 100 создает важный ориентир на протяжении всего периода времени, который охватывает эта визуализация. Источник: Yahoo! Finance

Рис. 22.10. Индексированные цены акций четырех крупных технологических компаний. Добавление тонких горизонтальных линий в точках основных значений оси *y* делает визуализацию проще для восприятия по сравнению с рис. 22.9. Благодаря такому подходу график более не нуждается в осях *x* и *y*, поскольку равномерно расположенные горизонтальные линии в достаточной мере обрамляют рисунок. Источник: Yahoo! Finance

Линии сетки координат должны располагаться перпендикулярно направлению изменения основной объясняемой переменной.

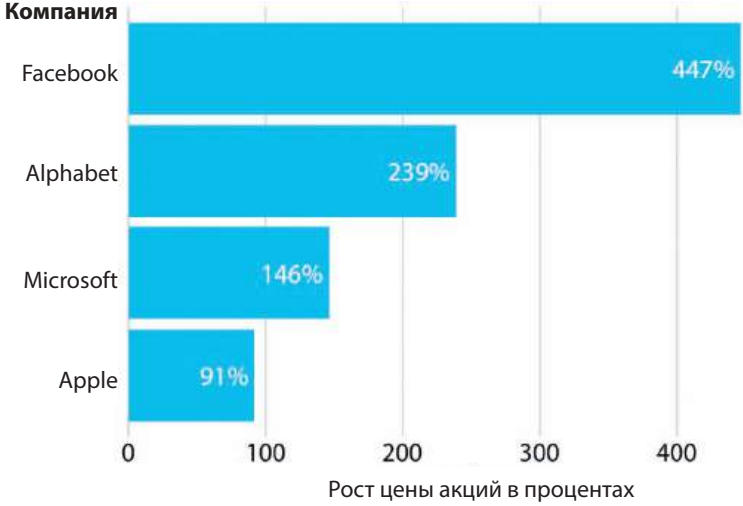


Рис. 22.11. Темпы прироста цен акций четырех крупных технологических компаний в период с июня 2012 года по июнь 2017 года. Поскольку столбцы с данными расположены горизонтально, то линии сетки координат должны быть вертикальными. Источник: Yahoo! Finance

Для гистограмм, подобной той, что показана на рис. 22.11, Тафти рекомендует рисовать поверх столбцов белые линии сетки координат вместо темных линий внизу [Tufte, 2001]. Подобного рода белые линии создают эффект разделения столбцов на отдельные сегменты одинаковой длины (рис. 22.12). Я смотрю на этот стиль двояко. С одной стороны, исследования человеческого восприятия показывают, что разбиение столбцов на отдельные сегменты помогает зрителю распознать их длину [Haroz, Kosara и Franconeri, 2015]. С другой стороны, на мой взгляд, столбцы выглядят так, как будто они разваливаются на части, и из-за этого график не воспринимается как единая визуальная единица. Я намеренно использовал этот стиль на рис. 5.10, чтобы визуально разделить столбики, представляющие пассажиров мужского и женского пола. То, какой из этих эффектов будет доминировать, может зависеть от конкретного выбора ширины столбцов, расстояния между ними и толщины белых линий координатной сетки. Поэтому, если вы собираетесь использовать этот стиль, как можно тщательнее подбирайте значения указанных выше параметров, чтобы итоговая визуализация имела желаемый визуальный эффект.

У рис. 22.12 есть еще один недостаток, который я хотел бы отметить. Мне пришлось вынести подписи значений за пределы столбцов, поскольку у части столбцов метки не помещались в их последние сегменты. Из-за этого длина столбцов неоправданно увеличилась, поэтому к такому решению стоит прибегать лишь в самых крайних случаях.

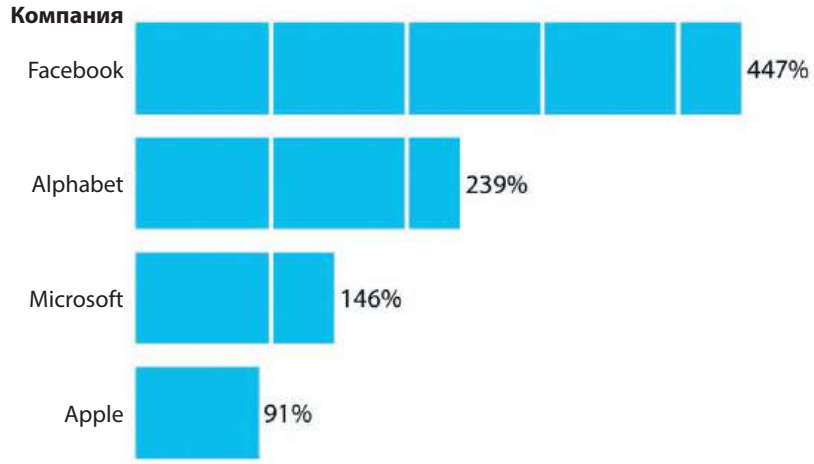


Рис. 22.12. Темпы прироста цен акций четырех крупных технологических компаний в период с июня 2012 года по июнь 2017 года. Расположенные поверх столбцов белые линии координатной сетки позволяют читателю оценить относительную длину столбцов. С другой стороны, эти линии могут создать впечатление, что столбцы как будто рассыпаются. Источник: Yahoo! Finance

Фоновая координатная сетка вдоль обеих осей наиболее подходит для диаграмм рассеяния, поскольку в них нет какой-то одной оси, значения на которой представляют для нас наибольший интерес. Рис. 22.2, приведенный в начале этой главы, являет собой подобный пример. Кроме того, вполне очевидно, что, если на изображении есть полная координатная сетка, линии осей обычно уже не нужны.

Парные данные

Для визуализаций, где основой сравнения является линия $x = y$ (например, на диаграммах рассеяния), я предпочитаю использовать саму диагональную линию, а не координатную сетку. Давайте в качестве примера рассмотрим рис. 22.13, на котором сравниваются уровни экспрессии генов в мутантном вирусе с его немутировавшим образцом (дикий тип). Благодаря диагональной линии мы сразу видим, какие гены экспрессируются выше или ниже у мутанта относительно дикого типа. Если бы визуализация имела фоновую координатную сетку без диагональной линии, сделать аналогичное наблюдение было бы гораздо сложнее (рис. 22.14). Таким образом, хоть данная визуализация и выглядит красиво, я отношу этот график к категории «плохих». В частности, на рис. 22.14 визуально не выделен ген 10A, который имеет очевидно сниженный уровень экспрессии в мутантной версии по сравнению нормальным вариантом (рис. 22.13).

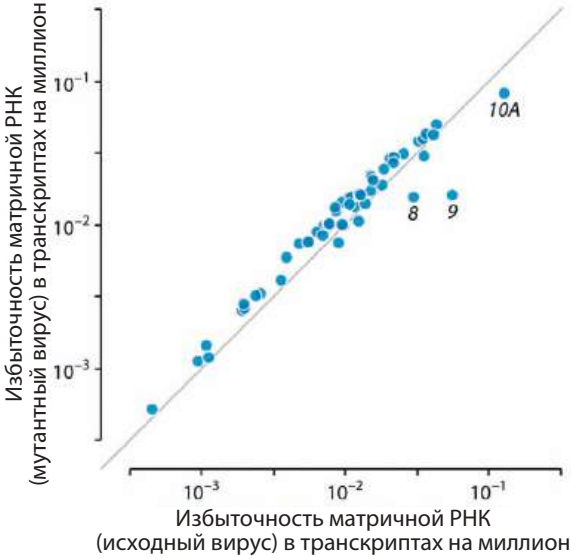


Рис. 22.13. Уровни экспрессии генов в мутантном бактериофаге T7 относительно дикого типа. Уровень экспрессии генов измеряется количеством мРНК в транскриптах на миллион. Каждая точка соответствует одному гену. В мутантном бактериофаге T7 промотор перед геном 9 был удален, что привело к уменьшению избыточности мРНК гена 9, а также соседних генов 8 и 10A (выделено). Источник: [Paff et al., 2018]

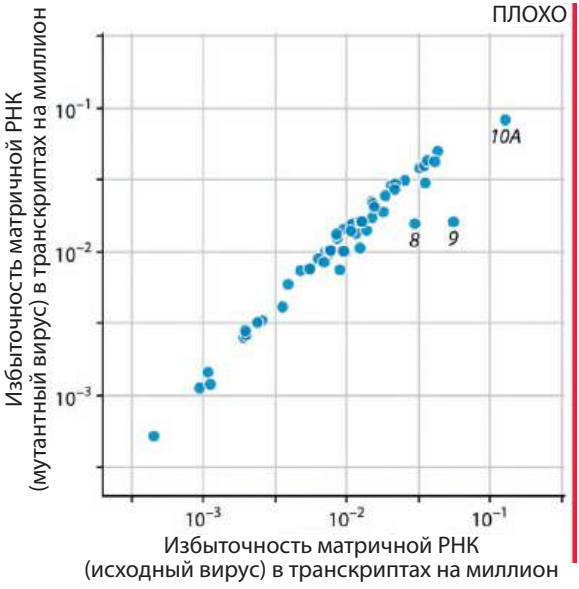


Рис. 22.14. Уровни экспрессии генов в мутантном бактериофаге T7 относительно дикого типа. Если разместить этот набор данных на фоне координатной сетки вместо диагональной линии, будет сложно понять, уровень экспрессии каких генов в мутанте выше или ниже, чем в бактериофаге дикого типа. Источник: [Paff et al., 2018]

Конечно, мы могли бы взять диагональную линию с рис. 22.13 и наложить ее поверх координатной сетки рис. 22.14, чтобы обеспечить наличие соответствующей визуальной отсылки. Однако в таком случае рисунок (рис. 22.15) становится визуально перегруженным. Мне пришлось сделать диагональную линию более темной, чтобы она выделялась на фоне решетки координат, правда, из-за этого точки данных начали сливаться с фоном. Эту проблему можно решить, увеличив размер точек или сделав их цвет более темным, но я считаю рис. 22.13 наилучшим решением.

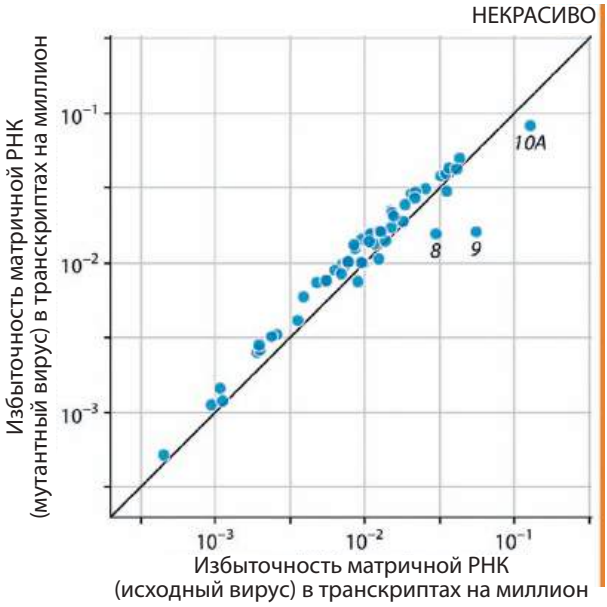


Рис. 22.15. Уровни экспрессии генов в мутантном бактериофаге T7 относительно дикого типа. Этот рисунок объединяет фоновую координатную сетку с рис. 22.14 и диагональную линию с рис. 22.13. На мой взгляд, данный рисунок визуально перегружен по сравнению с рис. 22.13, поэтому я отдаю предпочтение последнему. Источник: [Paff et al., 2018]

Вывод

Плохой дизайн графика может быть результатом как перенасыщения диаграммы визуальными элементами, так и их недостатка. Всякий раз, создавая диаграмму, мы должны стремиться найти золотую середину, которая подразумевает, что главный акцент графика делается на точках данных, однако при этом не упускается из виду и контекст, дающий понимание того, каким данным посвящен этот график, где расположены точки и что они означают.

Что касается фонов и фоновой сетки координат, не существует какого-то единственно правильного варианта, идеального для любого контекста.

Я рекомендую с осторожностью относиться к применению координатной сетки: следует как можно тщательнее подходить к выбору сетки или ее части, которая в каждом конкретном случае будет наиболее информативной, и использовать для визуализации именно ее. Я предпочитаю минималистичные светлые сетки на белом фоне, так как на бумаге белый по умолчанию нейтрален и на его фоне можно использовать практически любой цвет. Однако затемненный фон может помочь вам представить график в виде единого визуального объекта, что бывает особенно полезно для малых панельных визуализаций. Наконец, мы должны убедиться, что все изображения согласуются с шаблонами и правилами брендинга и общего визуального стиля. Многие журналы и веб-сайты стремятся к тому, чтобы их собственный визуальный стиль был мгновенно узнаваемым, и поэтому накладывают определенные требования на цвет фона и стиль фоновой сетки.

Глава 23

Подписи осей должны быть крупными

Если из всей этой книги вы вынесете лишь один какой-то урок, то пусть это будет «всегда обращайте внимание на подписи к осям, на метки основных и промежуточных делений и на любые другие пометки на графике». Практика показывает, что, скорее всего, они будут слишком маленькими. По моему опыту, почти все графические программы и графические библиотеки предлагают плохие настройки по умолчанию. Если вы будете на них полагаться, ваша визуализация почти наверняка выйдет неудачной.

Рассмотрим в качестве примера рис. 23.1. Я постоянно вижу подобного рода визуализации.

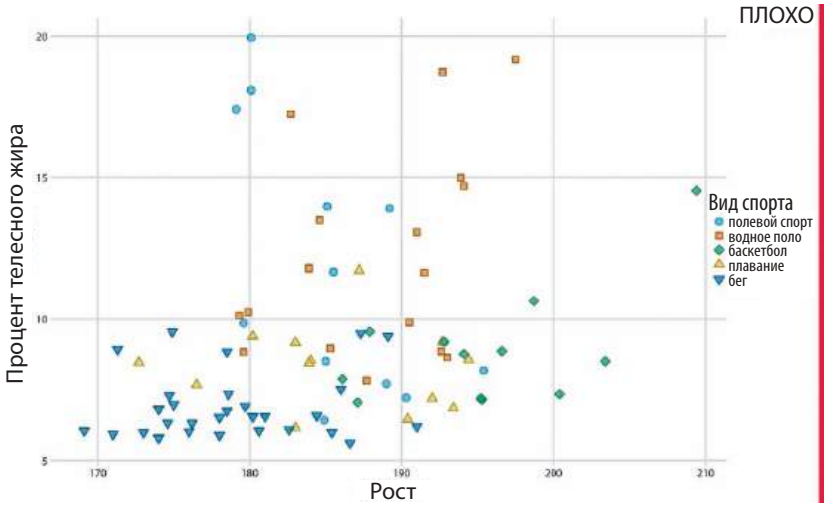


Рис. 23.1. Соотношение роста и процента жира у профессиональных австралийских спортсменов-мужчин (каждая точка соответствует одному спортсмену). Эта визуализация содержит распространенную ошибку: текстовые элементы слишком малы, и их сложно различить. Источник: [Telford and Cunningham, 1991]

Названия осей, метки делений на осях и подписи в легенде получились чрезвычайно крошечными. Чтобы прочитать их, нам, скорее всего, придется увеличить масштаб страницы.

На рис. 23.2 показана чуть более качественная версия этого графика. На мой взгляд, размеры шрифтов на нем все еще слишком малы, и поэтому я отнес данную визуализацию к категории «некрасивых», но тем не менее эта версия рисунка является шагом в правильном направлении. При определенных условиях эту диаграмму можно даже посчитать удовлетворительной. В данном случае моя критика больше направлена не на плохую читаемость текста, а на то, что этот вариант визуализации плохо сбалансирован: по сравнению с остальной частью рисунка подписи выглядят несоразмерно малыми.

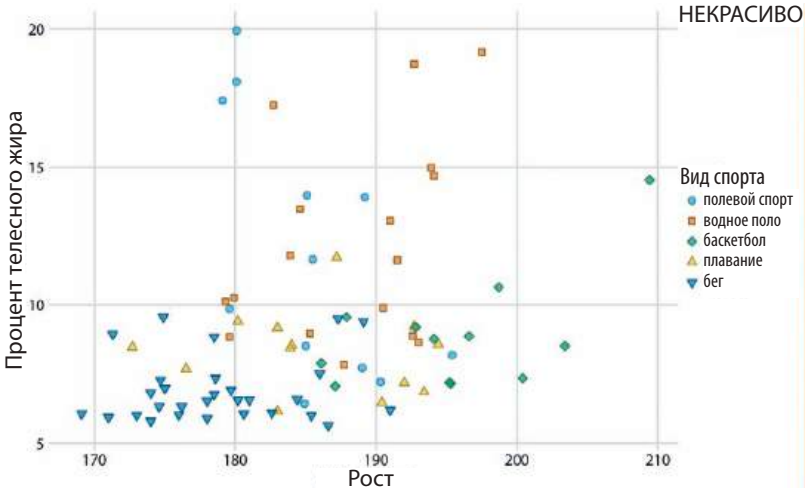


Рис. 23.2. Соотношение роста и процента жира среди спортсменов-мужчин. Данная визуализация является улучшенной версией рис. 23.1, однако текстовые элементы все еще слишком маленькие, а сама визуализация не сбалансирована. Источник: [Telford and Cunningham, 1991]

Настройки, использованные при создании рис. 23.3, применяются в этой книге повсеместно. Я считаю, что данный график является хорошо сбалансированным, потому что текст хорошо различим, а размер шрифта гармонирует с общим размером рисунка.

Следует помнить и о другой возможной крайности, которой тоже следует избегать: это слишком крупный текст (рис. 23.4). Бывает так, что большие подписи действительно нужны — например, если изображение планируется сделать маленьким целенаправленно, однако все остальные элементы рисунка (в частности, тексты подписи и символы на графике) должны соответствовать друг другу. На рис. 23.4 размер точек, используемых для

визуализации данных, слишком мал по отношению к размеру текста. Избавив рисунок от этого несоответствия, мы получим вполне приемлемую визуализацию (рис. 23.5).

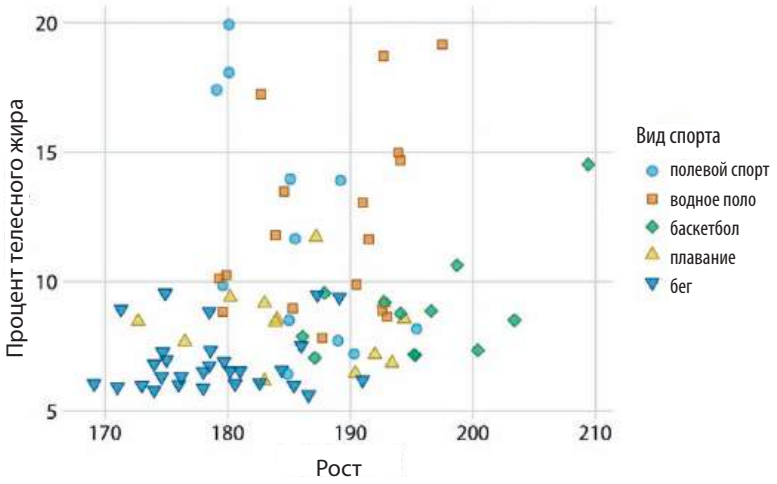


Рис. 23.3. Соотношение роста и процента жира среди спортсменов-мужчин. Элементы на этой диаграмме хорошо сбалансированы между собой. Источник: [Telford and Cunningham, 1991]

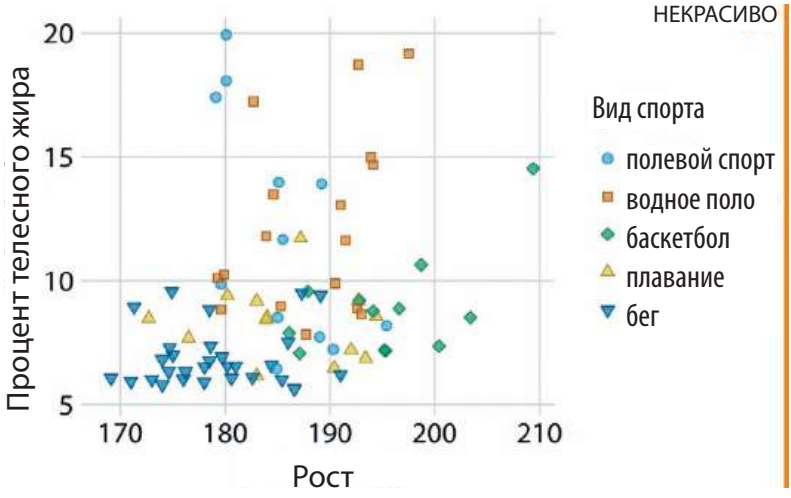


Рис. 23.4. Соотношение роста и процента жира среди спортсменов-мужчин. Текстовые элементы на данном графике довольно крупны, но это может быть оправданно, если рисунок предназначен для воспроизведения в гораздо меньшем, чем оригинал, масштабе. Тем не менее сама по себе визуализация плохо сбалансирована: точки данных на графике непропорционально малы по сравнению с размерами подписей. Источник: [Telford and Cunningham, 1991]

Посмотрев на рис. 23.5, вы можете подумать, что все элементы этого графика слишком большие. Однако данная диаграмма была задумана для использования в более мелком масштабе. Если вы уменьшите рисунок так, чтобы его ширина составляла всего пять — семь сантиметров, он будет смотреться прекрасно.

К слову сказать, эта визуализация — единственная во всей книге, которая будет хорошо выглядеть при таком масштабе.

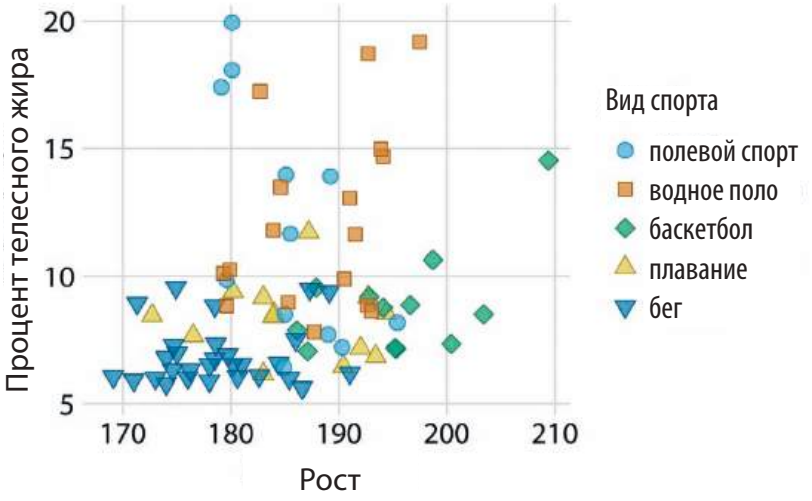


Рис. 23.5. Соотношение роста и процента жира среди спортсменов-мужчин. Все элементы изображения хорошо сбалансированы относительно друг друга, а сама визуализация пригодна для использования в уменьшенном масштабе. Источник: [Telford and Cunningham, 1991]

Всегда проверяйте, как ваш график выглядит в уменьшенном масштабе, чтобы убедиться в адекватности размеров подписей осей.

Я думаю, что существует простое психологическое объяснение того, почему мы склонны делать метки осей слишком маленькими, и связано оно с использованием мониторов большого размера и с высоким разрешением экранов.

Мы регулярно просматриваем графики на экране компьютера, при этом они зачастую занимают довольно много места. При таком способе просмотра даже сравнительно небольшой текст кажется абсолютно четким и совершенно разборчивым, а крупный текст, наоборот, — неуклюжим и громоздким. Вы можете сами в этом убедиться, если возьмете самое

первое изображение из этой главы и увеличите его до такой степени, чтобы оно заполнило весь экран. Вы наверняка подумаете, что график выглядит просто потрясающе. Таким образом, чтобы избежать подобного когнитивного искажения, всегда проверяйте, как выглядят ваши визуализации в том масштабе, в котором их планируется печатать. Для этого вы можете либо уменьшить масштаб, чтобы ширина графика на экране составляла всего несколько дюймов, либо попросту отойти от компьютера и взглянуть на диаграмму издалека, чтобы проверить, как выглядит график в уменьшенном из-за расстояния масштабе.

Глава 24

Избегайте лишних линий

Во всех тех случаях, где это возможно, старайтесь визуализировать данные при помощи залитых цветом геометрических фигур, а не их очертаний или простых линий. Цветные фигуры легче воспринимаются как целостные объекты, реже провоцируют появление визуальных артефактов или оптических иллюзий, а также лучше передают количества, в отличие от контуров. По моему опыту, визуализации, построенные из сплошных фигур, легче для понимания и приятнее для глаз, чем аналогичные изображения, в которых вместо сплошных фигур используются их контуры. И поэтому лично я стараюсь избегать рисования пустых контуров в принципе. Хочу, однако, подчеркнуть, что данная рекомендация не главнее, чем принцип пропорциональной заливки (см. главу 16).

Визуализации данных с помощью контуров очень много лет, поскольку на протяжении большей части XX века научные графики рисовались вручную и должны были быть воспроизводимы в черном и белом цветах без оттенков. Из-за этого было невозможно использовать элементы, закрашенные сплошными цветами, включая оттенки серого. Вместо этого «заполненные» области изображались с помощью различных видов штриховки или узоров. Ранние программные пакеты для черчения имитировали рисование от руки и также активно использовали контурные изображения фигур, штриховые или пунктирные линии, штриховку и узоры. Несмотря на то что современные инструменты визуализации и современные платформы для демонстрации и публикации лишены всех этих ограничений, многие графические приложения по-прежнему предлагают по умолчанию контуры и пустые формы, а не заполненные цветом области. Для лучшего понимания этой проблемы я покажу вам несколько примеров одних и тех же графиков, нарисованных как линиями, так и заполненными цветом фигурами.

Чаще всего от неуместного использования штриховых рисунков страдают гистограммы и столбчатые диаграммы. Столбцы, нарисованные в виде контуров, плохи тем, что читателю приходится прилагать усилия, чтобы понять, какая часть той или иной линии находится внутри столбца, а какая — снаружи. Из-за этого — особенно если между столбцами есть промежутки — визуализация приобретает запутанный вид, который отвлекает от основного посыла графика (рис. 24.1). Избежать указанной выше проблемы

можно с помощью заливки столбцов светлым оттенком какого-нибудь цвета или светло-серым, если носитель, на котором будет напечатан график, не предполагает цветопередачу (рис. 24.2).

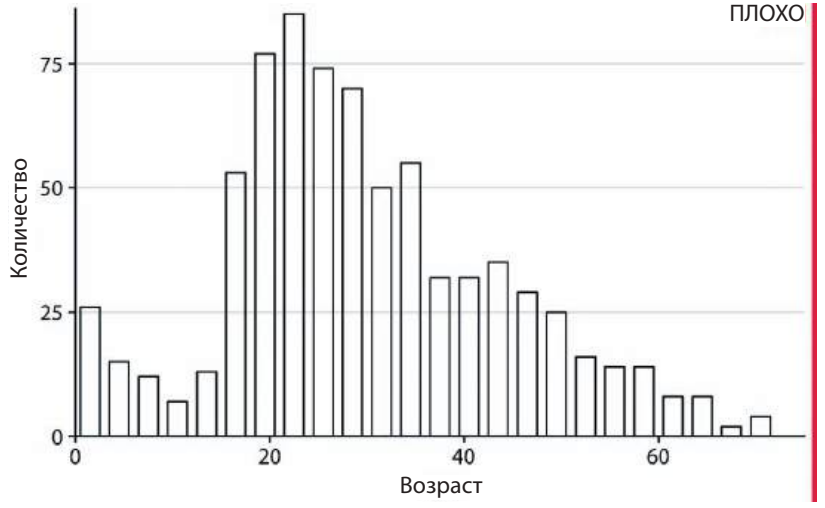


Рис. 24.1. Гистограмма распределения возрастов пассажиров «Титаника», нарисованная при помощи контуров столбцов. Контуры создают запутанный визуальный образ, затрудняя восприятие графика. Ближе к центру графика становится сложно понять, какие части столбцов находятся внутри, а какие — снаружи. Источник: Encyclopedia Titanica

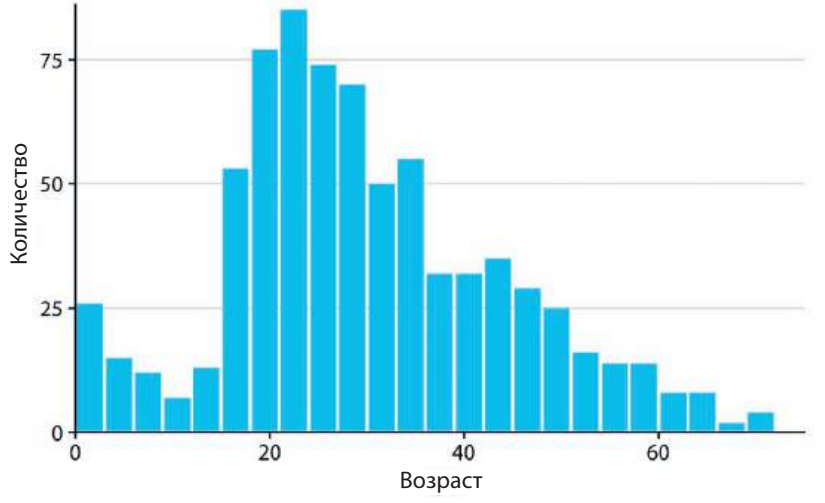


Рис. 24.2. Гистограмма распределения возрастов пассажиров «Титаника». Это та же самая гистограмма, что и на рис. 24.1, только здесь столбцы закрашены. На данном графике форма распределения воспринимается гораздо легче по сравнению с предыдущей визуализацией. Источник: Encyclopedia Titanica

Далее, давайте взглянем на диаграммы плотности распределений и как их строили раньше. В качестве примера я взял график, показывающий распределение длины чашелистика трех видов ирисов, изображенных в виде черных контуров на белом фоне (рис. 24.3). Распределения плотности представлены в виде контуров, а поскольку использовать можно только черный цвет, каждый график выполнен своим типом линий. На мой взгляд, у этой диаграммы есть два существенных недостатка. Во-первых, стили пунктирных линий не обеспечивают четкого разделения областей под кривой и над ней. Несмотря на то что зрительная система человека достаточно хорошо справляется с объединением отдельных элементов линии в одну непрерывную, пунктирные линии все равно выглядят прерывисто, и поэтому область, ими очерченная, не воспринимается однозначно как замкнутая. Во-вторых, поскольку эти линии пересекаются, а области, которые данные линии окружают, не затенены, визуально сложно понять, как шесть получающихся очерченных областей коррелируют с тремя отображаемыми распределениями. Этот эффект был бы еще сильнее, если бы для всех трех распределений я использовал сплошные линии, а не пунктирные разных стилей.

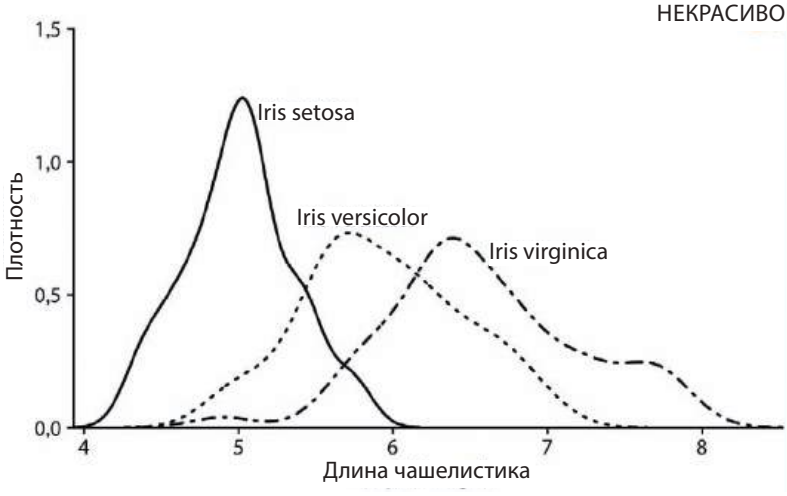


Рис. 24.3. Оценки плотности длин чашелистика для трех различных видов ирисов. Прерывистые линии, используемые для визуализации значений Iris versicolor и Iris virginica, отвлекают внимание зрителя от того, что области под кривыми и над ними не являются единым целым. Источник: [Fisher, 1936]

Мы можем попытаться решить проблему прерывистых границ, используя вместо пунктирных линий цветные (рис. 24.4). Пойдя этим путем, мы обнаружим, что графики плотности по-прежнему являются слабо различимыми. В общем, я считаю версию с закрашенными графиками плотности (рис. 24.5) наиболее простой и интуитивно понятной. Не стоит, однако, забывать, что

закрашенные области должны быть частично прозрачными, чтобы было видно полное распределение для каждого вида.

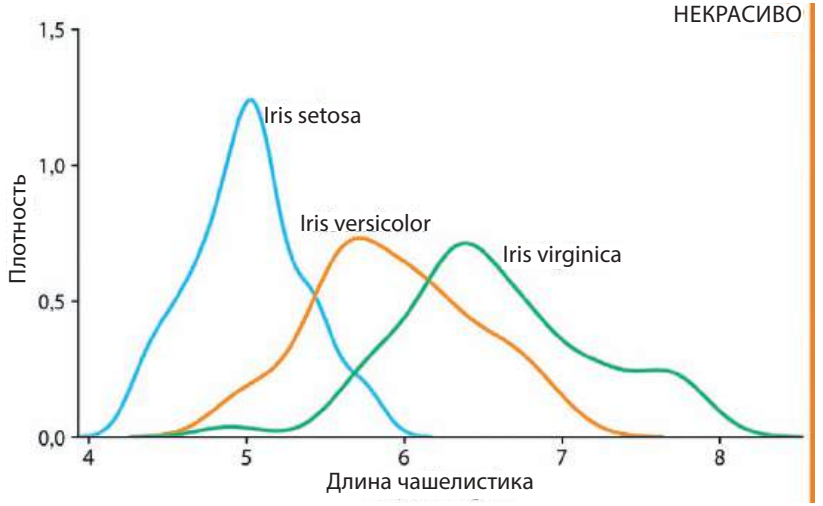


Рис. 24.4. Оценки плотности длин чашелистика для трех различных видов ирисов. На данном рисунке прерывистые линии заменены на цветные. Благодаря этому подходу мы смогли решить проблему рис. 24.3, связанную с тем, что области под кривыми и над ними выглядят связанными. Однако на данной визуализации все еще трудно понять размеры области под каждым графиком. Источник: [Fisher, 1936]

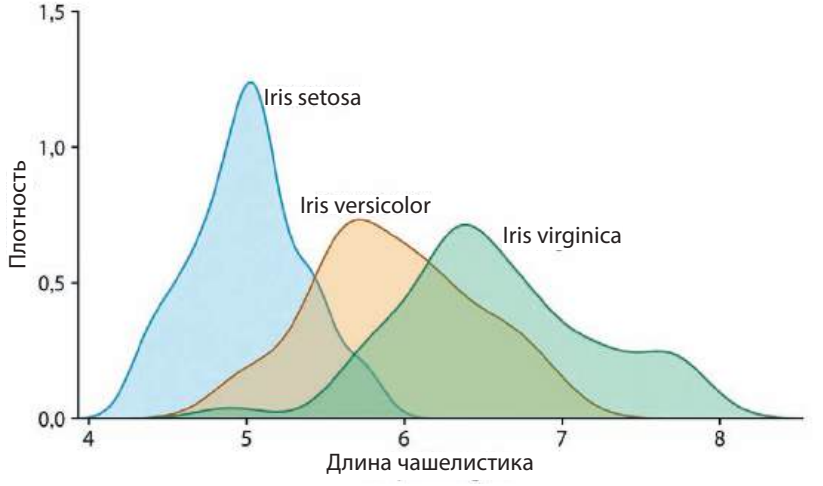


Рис. 24.5. Оценки плотности длин чашелистика для трех различных видов ирисов, визуализированные в виде частично прозрачных закрашенных областей. Благодаря тому, что область под каждой кривой покрашена в отдельный цвет, становится гораздо легче воспринимать три различных графика распределения плотности как отдельные объекты. Источник: [Fisher, 1936]

Контуры можно использовать и в диаграммах рассеяния, в которых для указания различных типов данных используются различные формы маркеров. К таким формам можно отнести незакрашенные круги, треугольники, крестики и тому подобные элементы. В качестве примера давайте рассмотрим рис. 24.6.

Данный рисунок очевидным образом страдает от слишком большого количества визуального шума, а кроме того, на этой визуализации довольно сложно отделить одни точки данных от других.

Если та же самая визуализация будет выполнена с помощью закрашенных фигур, отражающих точки данных, мы избавимся от указанной выше проблемы (рис. 24.7).

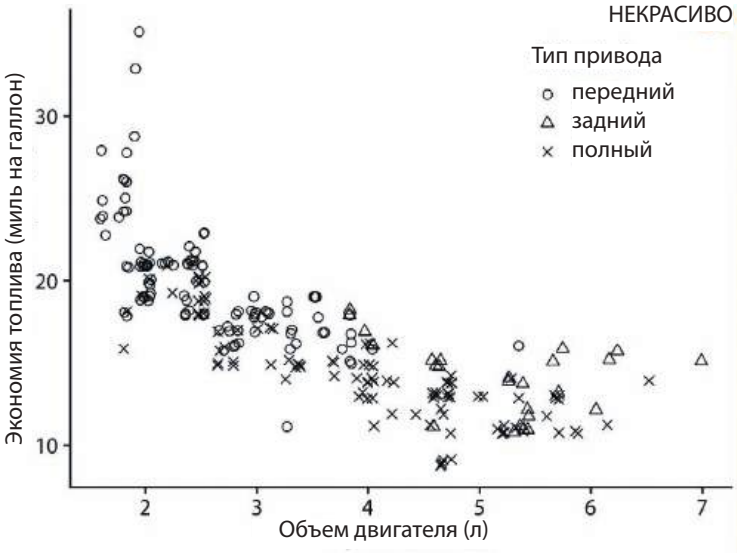


Рис. 24.6. Отношение экономии топлива в городских условиях к объему двигателя для автомобилей с передним, задним и полным приводом. Различные стили маркеров, нарисованных в виде черных контуров на белом фоне, зашумляют график и затрудняют его восприятие. Источник: US Environmental Protection Agency (EPA), fueleconomy.gov

Я предпочитаю закрашенные маркеры данных незакрашенным, поскольку первые визуально воспринимаются значительно лучше. Иногда применение незакрашенных маркеров аргументируют тем, что они решают проблему оверплоттинга, так как пустые области в середине каждой точки позволяют видеть другие точки, расположенные под ними. Этот довод, безусловно, разумен, однако, на мой взгляд, польза от этого эффекта не перевешивает вред, который причиняет графику визуальный шум от незакрашенных маркеров. Кроме того, есть и другие способы решения проблемы оверплоттинга, см. главу 17.

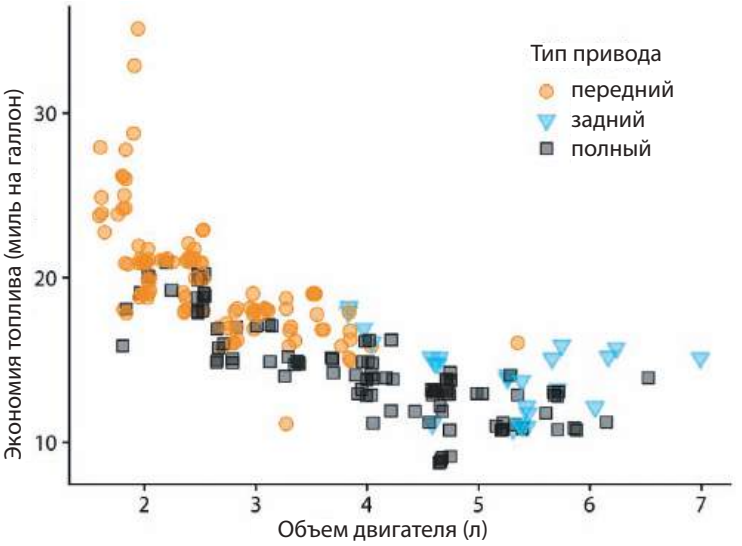


Рис. 24.7. Отношение экономии топлива в городских условиях к объему двигателя для автомобилей с передним, задним и полным приводом. Благодаря тому, что на данной визуализации для указания различных типов данных мы используем и цвет, и форму, то даже в случае обесцвечивания рисунка (перевода его в оттенки серого) мы все равно сможем легко понять распределение автомобилей по типу привода. Источник: US Environmental Protection Agency (EPA), fueleconomy.gov

В завершение главы давайте рассмотрим еще такой тип графика, как коробчатая диаграмма. Графики такого вида обычно рисуются с пустыми областями, как показано на рис. 24.8. Я предпочитаю затенять область внутри коробок светлым цветом, как это сделано на рис. 24.9. Затенение визуально отделяет прямоугольники от фона графика, что особенно полезно в том случае, когда мы показываем много прямоугольников рядом друг с другом (рис. 24.8 и 24.9).

Присутствие на рис. 24.8 большого количества прямоугольников и линий может ввести читателя в заблуждение, создав иллюзию наличия областей за пределами прямоугольников, как это было на рис. 24.1. На рис. 24.9 эта проблема устранена. Иногда мне говорят, что затенение внутренней части ящика придает слишком большой вес средним 50% данных, но я считаю данный аргумент ошибочным. Коробчатая диаграмма сама по себе склонна делать акцент на данных, которые попадают в центральные 50%. И поэтому заливка коробок цветом в данном случае не играет никакой роли. Если вы не хотите, чтобы в вашей визуализации присутствовал подобный эффект, просто откажитесь от использования данного типа графика. Вместо этого используйте скрипичную диаграмму, точки с джиттерингом или график Sina (см. главу 8).

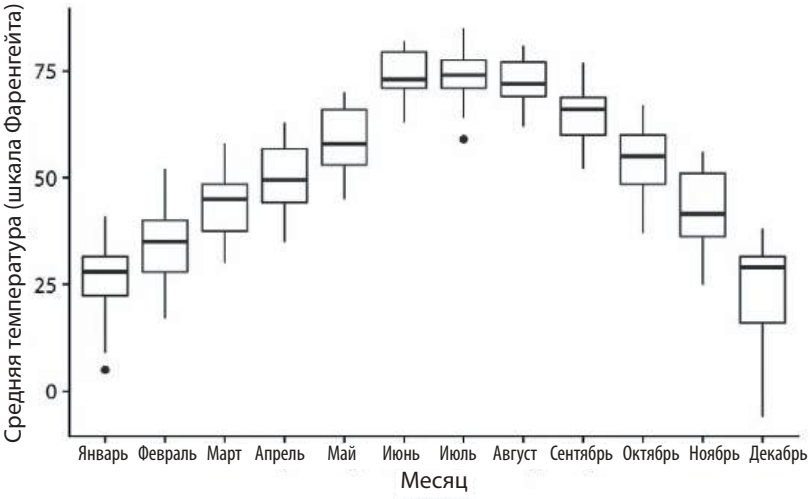


Рис. 24.8. Распределение среднесуточных температур в Линкольне, штат Небраска, в 2016 году. Все коробки имеют традиционный (незакрашенный) вид. Источник: Weather Underground

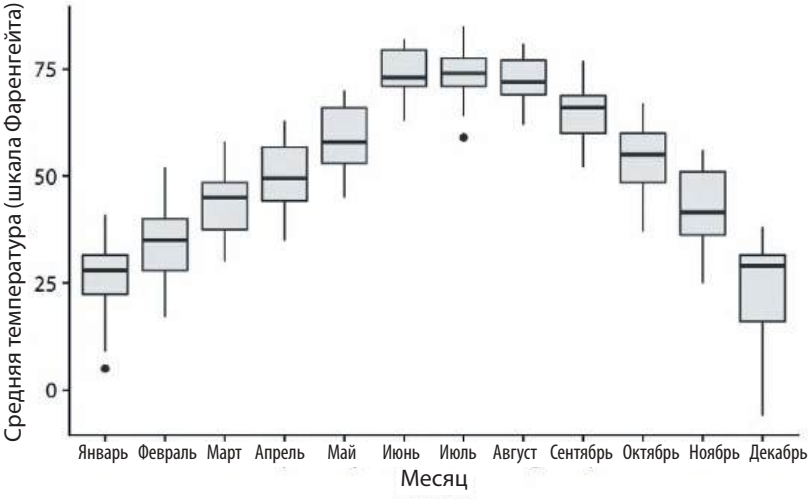


Рис. 24.9. Распределение среднесуточных температур в Линкольне, штат Небраска, в 2016 году. Если раскрасить ящики в светло-серый цвет, они будут лучше выделяться на белом фоне. Источник: Weather Underground

Глава 25

Не используйте 3D

3D-графики являются довольно популярным средством визуализации, которое можно встретить и в бизнес-среде, и в различных областях науки. При этом они практически всегда используются не к месту. Очень редко трехмерный график не становится значительно лучше, если его перевести в двухмерный вид. В этой главе мы поговорим о том, какие проблемы присущи 3D-графикам, почему использовать графики такого типа не стоит и в каких редких случаях они все же являются подходящим выбором для визуализации данных.

Избегайте неоправданного применения 3D

Многие инструменты визуализации предлагают вам улучшить вид ваших графиков с помощью преобразования элементов графиков в трехмерные объекты. Наиболее распространенными примерами подобного рода улучшений являются круговые диаграммы в виде дисков, повернутых в пространстве, столбчатые диаграммы с колоннами вместо столбцов и линейчатые диаграммы, состоящие из лент. Примечательно, что ни в одном из этих случаев третье измерение не сообщает зрителю никаких дополнительных сведений. 3D используется просто для украшения. На мой взгляд, такое использование 3D полностью лишено смысла. Подобный подход однозначно относится к категории «плохих», и его следует вычеркнуть из визуального словаря специалистов по работе с данными.

Проблема бесконтрольного и бессмысленного применения 3D заключается в том, что проекция трехмерных изображений на двухмерное пространство (такое как лист бумаги или монитор) всегда происходит с искажениями. Зрительная система человека автоматически пытается исправить эти искажения, однако полностью достичь этого невозможно. Давайте в качестве примера возьмем обыкновенную круговую диаграмму, разделенную на две части, 25 и 75%, и немного повернем ее в пространстве (рис. 25.1). По мере изменения угла обзора наше восприятие пропорций частей будет меняться. В частности, сегмент в 25%, расположенный в передней части диаграммы, будет выглядеть намного больше своего фактического размера, если смотреть на него под определенным углом (рис. 25.1A).

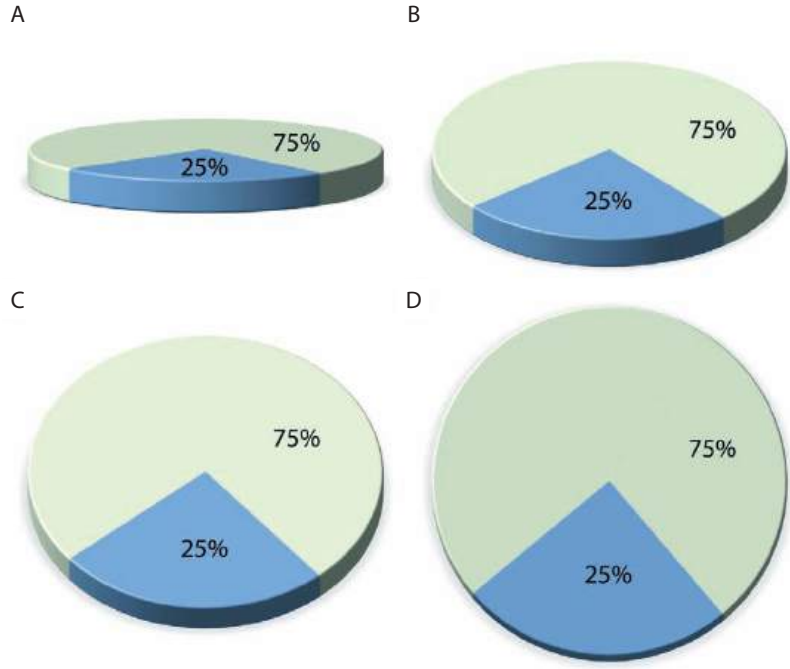


Рис. 25.1. Одна и та же трехмерная визуализация круговой диаграммы, показанная под разными углами. Вращение изображения в трехмерном пространстве приводит к тому, что сегменты в передней части выглядят крупнее, чем на самом деле, а сегменты в задней части графика — наоборот. Синий сегмент на изображениях A, B и C, соответствующий 25% данных, визуально занимает более 25% площади. Правильно представляет исходные данные только изображение D

Другие типы трехмерных графиков страдают теми же самыми недостатками. На рис. 25.2 показано распределение пассажиров «Титаника» по классам кают и полам с использованием трехмерных столбцов. Из-за того, как они расположены относительно осей координат, все столбцы выглядят короче, чем они есть на самом деле. Например, численность пассажиров первого класса составляет 322 человека, однако на рис. 25.2 кажется, что их меньше 300. Эта иллюзия возникает из-за того, что столбцы, представляющие данные, расположены на некотором расстоянии от задних стенок графика, на которых нарисованы серые горизонтальные линии уровня. Чтобы ощутить воздействие этого эффекта, попробуйте мысленно двигать нижний край одного из столбцов до тех пор, пока он не достигнет самой нижней серой линии, представляющей значение 0. Затем сделайте то же самое с любым из верхних краев, и вы поймете, что на самом деле все столбцы выше, чем кажутся на первый взгляд (см. рис. 5.10 в главе 5 как более корректную двухмерную версию данного графика).

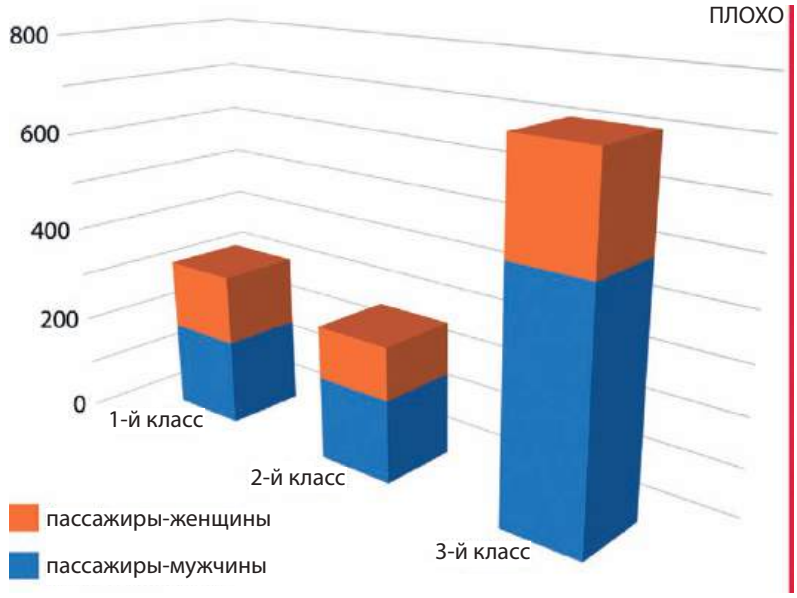


Рис. 25.2. Количество пассажиров и пассажирок на «Титанике», путешествовавших в первом, втором и третьем классах. Данный график представляет собой трехмерную столбчатую диаграмму. Общее количество пассажиров в первом, втором и третьем классах составляет 322, 279 и 711 человек соответственно (см. рис. 5.10). Эта 3D-диаграмма создает иллюзию, что в первом столбике показано менее 300 пассажиров, в третьем — менее 700 пассажиров, а во втором — около 210 пассажиров (хотя на самом деле их там 279). Более того, из-за того что третий столбец визуально доминирует над остальными, читателю кажется, что количество пассажиров, путешествовавших в третьем классе, выше, чем это было на самом деле

Не используйте трехмерную систему координат

В то время как визуализации с бессмысленным применением 3D можно однозначно отнести к категории плохих, визуализация данных в трех измерениях (с осями x, y и z) является несколько более спорной. В этом случае использование третьего измерения оправданно. Тем не менее такие графики часто трудны для восприятия, поэтому я считаю, что их следует избегать.

Рассмотрим трехмерную диаграмму рассеяния, показывающую зависимость расхода топлива от объема двигателя и его мощности для 32 автомобилей. В главе 1 мы уже встречались с этим набором данных (см. рис. 1.5). В этот раз ось x будет показывать объем двигателя, ось y — мощность, а ось z — расход топлива. Каждый автомобиль обозначается точкой (рис. 25.3). Несмотря на то что эта трехмерная визуализация показана с четырех разных перспектив, понять, как именно точки распределены в пространстве, все равно

сложно. Я нахожу часть D наиболее запутанной, поскольку создается впечатление, что на ней изображен совершенно другой набор данных, хотя мы всего лишь изменили угол обзора.

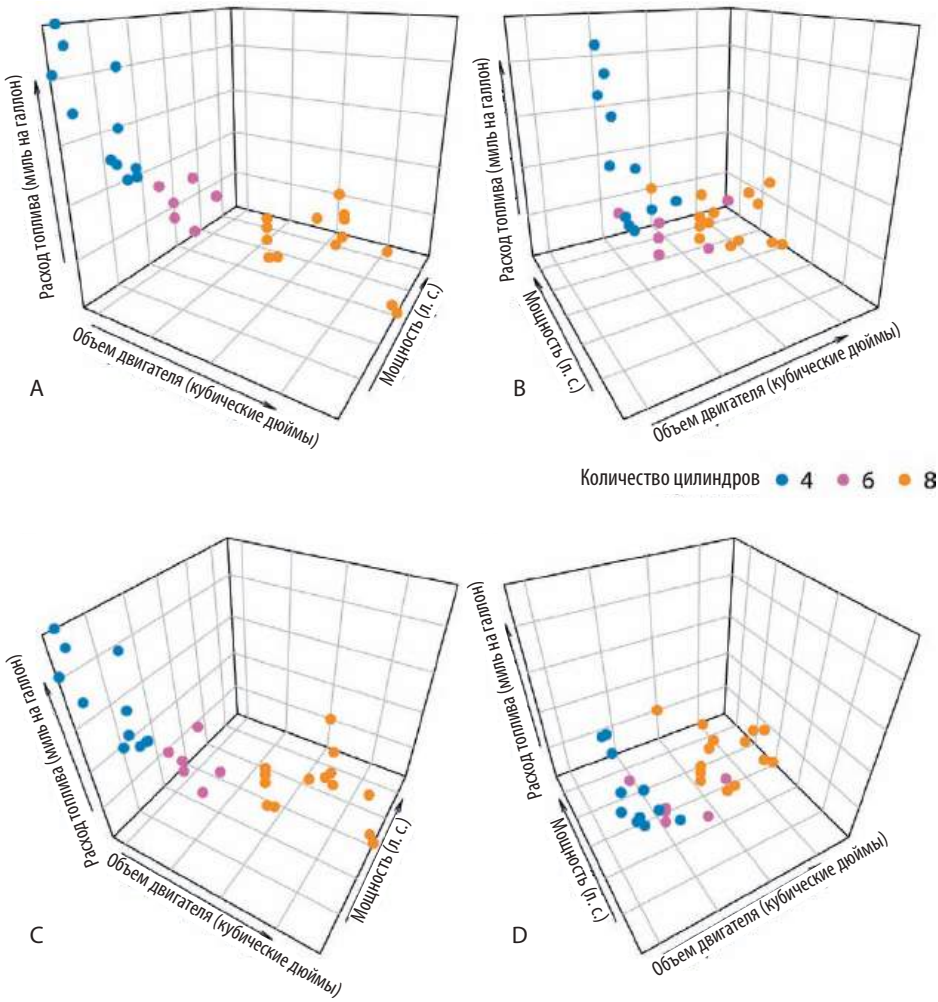


Рис. 25.3. Зависимость топливной экономичности от объема двигателя и его мощности для 32 автомобилей (модели 1973–1974 годов). Каждая точка представляет один автомобиль, а цвет обозначает количество цилиндров в двигателе. Каждая из четырех панелей данного графика показывает одни и те же данные, но с разных перспектив. Источник: Motor Trend, 1974

Основная проблема таких визуализаций заключается в том, что они требуют двух отдельных последовательных преобразований данных. Первое преобразование переводит данные из пространства данных в пространство

3D-визуализации. Ранее мы уже касались этой темы, когда обсуждали в главах 1 и 2 системы координат.

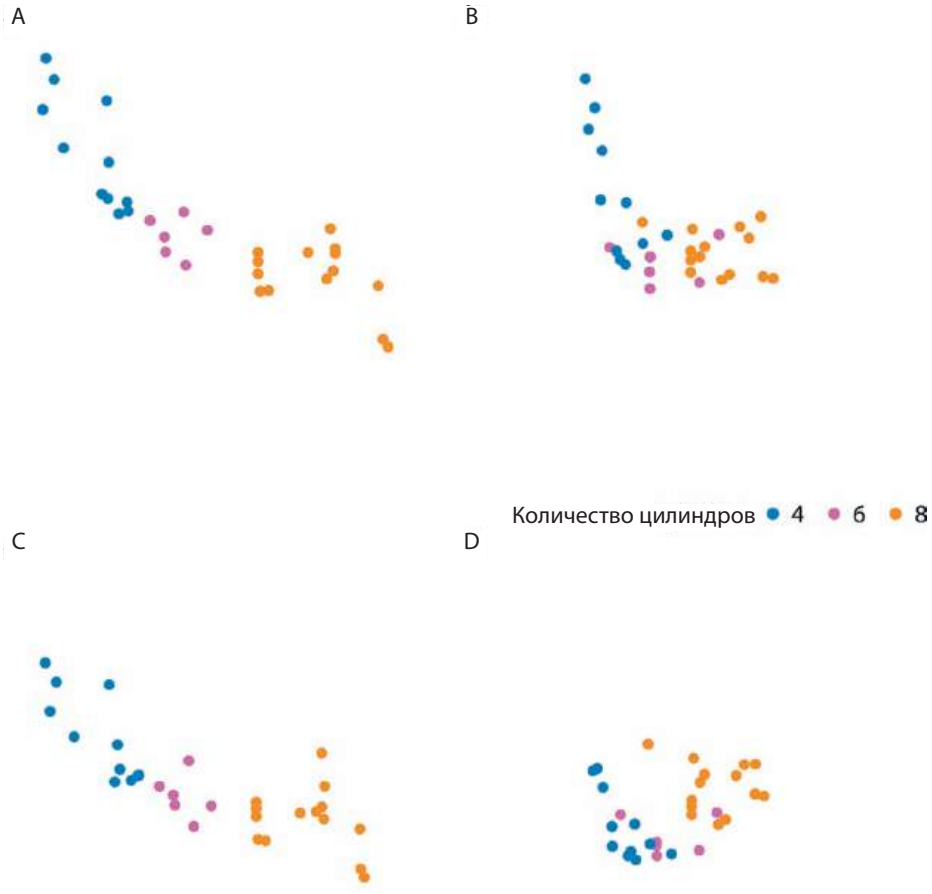


Рис. 25.4. Зависимость топливной экономичности от объема двигателя и его мощности для 32 автомобилей (модели 1973–1974 годов). Данная визуализация полностью идентична рис. 25.3, за исключением того, что все линии, указывающие на форму проекции трехмерного графика на плоскость, удалены. Источник: Motor Trend, 1974

Второе преобразование отображает данные из трехмерного пространства визуализации в двухмерном пространстве конечного изображения. (Очевидно, что второе преобразование не происходит для визуализаций, отображаемых в реальной трехмерной среде, например, когда речь идет о физических скульптурах или 3D-печати. В данной главе я прежде всего критикую отображение трехмерных визуализаций на плоскость.) Второе преобразование является необратимым, поскольку каждая точка на двухмерном дисплее

соответствует прямой в пространстве 3D-визуализации. По этой причине мы не можем однозначно определить, где в трехмерном пространстве находится та или иная точка данных.

Тем не менее наш глаз все же пытается инвертировать это преобразование. К сожалению, этот процесс чреват ошибками и сильно зависит от наличия на графике подсказок, которые передают ощущение трехмерности. Если убрать эти подсказки, инверсия станет абсолютно невозможной. Данный эффект продемонстрирован на рис. 25.4, который идентичен рис. 25.3, за исключением того, что все элементы, указывающие на глубину и трехмерность изображения, были удалены. В результате мы имеем дело с четырьмя наборами точек, которые невозможно ни интерпретировать, ни соотнести друг с другом. Например, можете ли вы сказать, какие точки на панели A соответствуют каким точкам на панели B? Я — нет.

По моему мнению, вместо двух отдельных преобразований данных, одно из которых является необратимым, будет лучше остановиться на одном обратимом преобразовании и сразу спроецировать данные на плоскость. Случаи, когда для визуализации данных действительно требуется третье измерение, крайне редки, поскольку, помимо откладывания переменных на координатной оси, их можно передать цветом, размером или формой. Например, в главе 1 я построил график по всем пяти переменным из набора данных об эффективности использования топлива, не выходя при этом за рамки двухмерной системы координат (см. рис. 1.5).

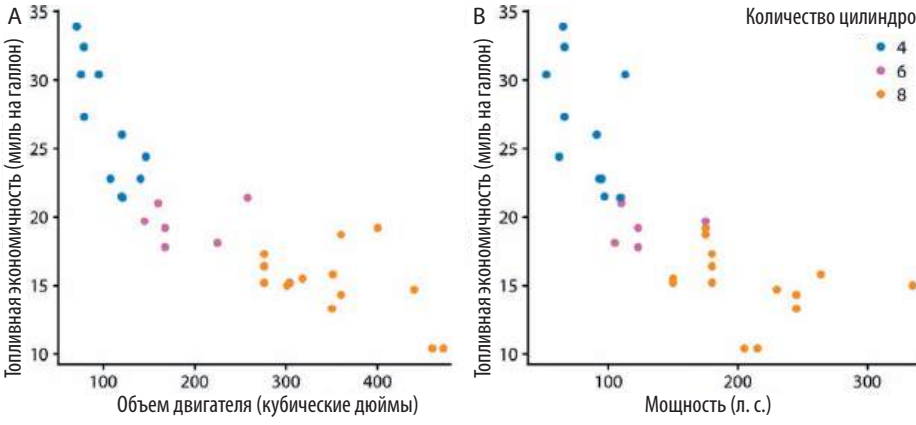


Рис. 25.5. Отношение топливной экономичности к объему двигателя (A) и мощности (B) для 32 автомобилей. Источник: Motor Trend, 1974

Давайте посмотрим, какими еще двумя способами мы могли бы построить график для всех тех переменных, что показаны на рис. 25.3. Во-первых, если в качестве объясняемой переменной нас больше всего интересует экономия топлива, мы можем построить ее дважды: один раз относительно объема

двигателя и другой раз относительно мощности двигателя (рис. 25.5). Во-вторых, если бы нас заинтересовало соотношение объема двигателя и его мощности, расход топлива отойдет на второй план. Исходя из данного условия, мы можем построить график зависимости мощности от объема двигателя и отобразить расход в виде размера точек (рис. 25.6). Оба графика являются более информативными и менее запутанными, чем рис. 25.3.

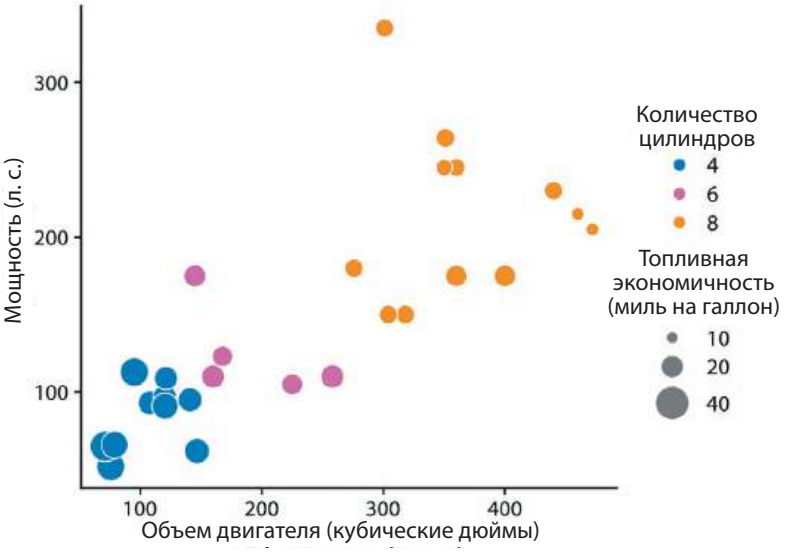


Рис. 25.6. Отношение мощности двигателя к его объему для 32 автомобилей. Топливная экономичность представлена размером точек. Источник: Motor Trend, 1974

Вероятно, вы уже задавались такими вопросами, как «Возможно ли, что проблема трехмерных диаграмм рассеяния вызвана тем, что точки — как способ представления данных — сами по себе не содержат никакой информации о третьем измерении?», «Что произойдет, если вместо точек мы задействуем, например, трехмерные столбцы?». На рис. 25.7 показан типичный набор данных, который можно визуализировать с помощью трехмерных столбцов. Данный набор содержит информацию об уровне смертности в штате Виргиния в 1940 году, разбитую по возрастным группам, полу и месту проживания. Мы видим, что трехмерные столбцы действительно помогают нам интерпретировать график. Маловероятно, что кто-то сможет перепутать столбец на переднем плане со столбцом на заднем плане, или наоборот. Тем не менее проблемы, о которых мы говорили в контексте рис. 25.2, никуда не делись. Довольно трудно точно понять, насколько высоки те или иные столбцы, и не менее сложно их сравнивать между собой. Например, как вы думаете, уровень смертности среди городского женского населения в возрастной группе 65–69 лет был выше или ниже, чем среди городского мужского населения в возрастной группе 60–64 лет?

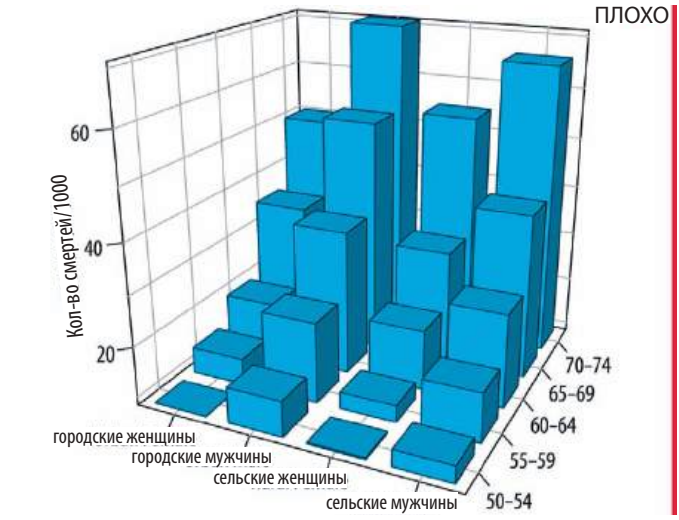


Рис. 25.7. Показатели смертности в штате Виргиния в 1940 году, визуализированные в виде трехмерной столбчатой диаграммы. Показатели смертности охватывают четыре группы людей (мужчины и женщины, проживавшие в городских и сельских условиях) в пяти возрастных категориях (50–54, 55–59, 60–64, 65–69, 70–74) и измеряются в количестве смертей на 1000 человек. Этот график относится к категории «плохих», потому что трехмерная перспектива затрудняет чтение графика. Источник: [Molyneaux, Gilliam, and Florant, 1947]

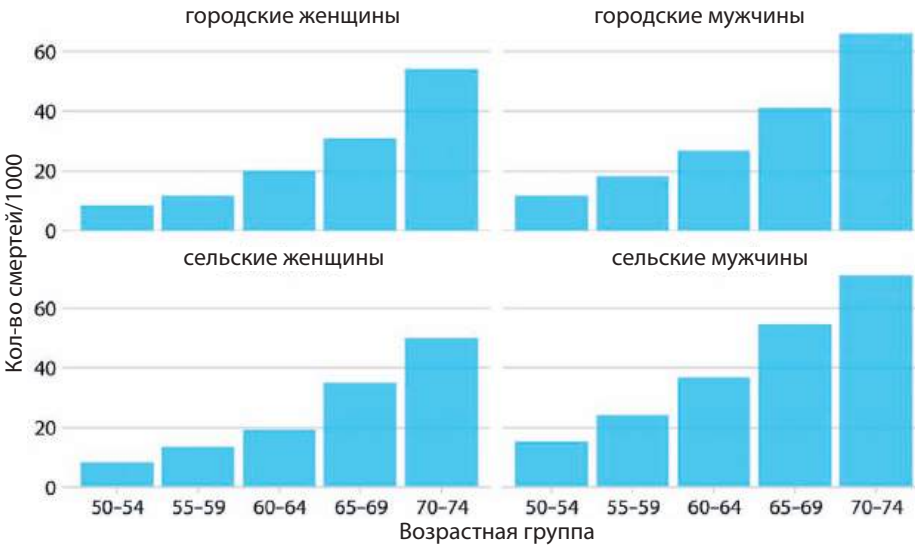


Рис. 25.8. Показатели смертности в штате Виргиния в 1940 году, визуализированные в виде малой панельной визуализации. Показатели смертности охватывают четыре группы людей (мужчины и женщины, проживавшие в городских и сельских условиях) в пяти возрастных категориях (50–54, 55–59, 60–64, 65–69, 70–74) и измеряются в количестве смертей на 1000 человек. Источник: [Molyneaux, Gilliam, and Florant, 1947]

Здесь куда более подходящим вариантом стали бы малые панельные визуализации (глава 20). В этом случае набор данных о смертности в штате Виргиния потребовал бы для своего отображения всего лишь четыре панели (рис. 25.8). На мой взгляд, данный график выглядит понятно и легко интерпретируется. Достаточно беглого взгляда, чтобы заметить, что смертность мужчин выше, чем женщин, а также что уровень смертности среди городского мужского населения, по-видимому, выше, чем среди деревенского, в то время как для городского и деревенского женского населения такой разницы не наблюдается.

Когда трехмерные визуализации уместны

В начале этой главы я сказал, что существует небольшой класс сценариев, в которых трехмерные визуализации используются вполне оправданно. Во-первых, проблемы, описанные в предыдущем разделе, частично теряют свою актуальность, если визуализация является интерактивной и зритель может ею манипулировать или если визуализация предназначена для отображения в среде виртуальной или дополненной реальности, где на диаграмму можно взглянуть под разными углами. Во-вторых, даже если визуализация не является интерактивной, но ее демонстрация предполагает медленное вращение графика, а не просто показ в виде статичного изображения под одним углом зрения, то это позволит зрителю определить, где в трехмерном пространстве находятся различные графические элементы.

Рис. 25.9. Рельеф острова Корсика в Средиземном море.
Источник: Copernicus Land Monitoring Service

Человеческий мозг очень хорошо умеет собирать трехмерную картинку из серии изображений, сделанных под разными углами, и прием с медленным вращением использует именно эту особенность человеческого восприятия.

Наконец, еще одним сценарием, в котором уместно использование 3D-визуализации, является демонстрация реальных 3D-объектов и/или данных, нанесенных на них. Например, выбор трехмерного подхода для визуализации топографического рельефа горного острова станет очень хорошим решением (рис. 25.9). Аналогично, если мы хотим визуализировать сохранность эволюционной последовательности белка, нанесенную на карту его структуры, показ структуры как трехмерного объекта будет наиболее разумным выбором (рис. 25.10). Однако следует отметить, что все эти визуализации могли бы смотреться еще лучше, если бы их изобразили в виде анимированных вращающихся изображений. И хотя такой способ визуализации неприменим в традиционных печатных изданиях, он как нельзя лучше подходит для публикации рисунков в интернете или для мультимедийных презентаций.

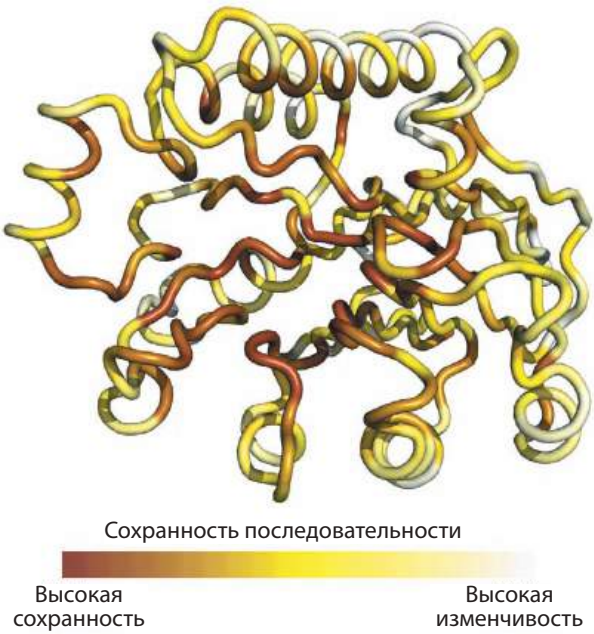


Рис. 25.10. Закономерности в процессе эволюционного изменения белка. Цветная трубка представляет собой основу белка экзонуклеазы III бактерии Escherichia coli. Окраска указывает на эволюционную сохранность отдельных участков в этом белке: темным цветом обозначены сохраняющиеся аминокислоты, а светлым — изменяющиеся аминокислоты. Источник: [Marcos and Echave, 2015]

Часть III

Разное

Глава 26

Наиболее распространенные форматы файлов изображений

Каждый, кто занимается подготовкой изображений для визуализации данных, должен рано или поздно разобраться в том, каким образом компьютер хранит графику. Существует множество различных форматов изображений, и у каждого из них есть свои преимущества и недостатки. Выбор правильного формата и правильного рабочего процесса позволит вам избежать множества проблем, которые часто возникают в процессе создания визуализации.

Лично я отдаю предпочтение PDF для случаев, когда мне надо подготовить документы высокого качества и пригодные для публикации, для онлайн-документов и других сценариев, где требуется растровая графика, — PNG, а если файлы PNG оказываются слишком велики — в крайнем случае я использую JPEG. Далее в этой главе мы поговорим о том, какие существуют ключевые различия между этими форматами файлов, об их преимуществах и недостатках.

Растровая и векторная графика

Графические форматы бывают двух типов — растровые и векторные (табл. 26.1). Растровое изображение представляет собой сетку, состоящую из отдельных точек (пикселей), каждая из которых имеет свой цвет. Векторный же формат представляет изображение как набор геометрических форм и хранит их положение друг относительно друга. Другими словами, векторное изображение содержит информацию типа «есть черная линия от верхнего левого угла до нижнего правого угла и красная линия от нижнего левого угла до верхнего правого угла», а сама картинка воссоздается непосредственно на месте, когда ее требуется отобразить на экране или распечатать.

Отличительной особенностью векторной графики является ее «независимость от разрешения», поскольку размер таких рисунков можно менять как угодно, не рискуя при этом потерять какие-либо детали или резкость изображения. Рис. 26.1 иллюстрирует вышесказанное.

Таблица 26.1. Наиболее распространенные форматы файлов

Сокращение	Имя	Тип	Способ применения
PDF	Portable Document Format	Векторная графика	Многоцелевой
EPS	Encapsulated PostScript	Векторная графика	Многоцелевой; устарел; предпочтительнее использовать PDF
SVG	Scalable Vector Graphics	Векторная графика	Онлайн
PNG	Portable Network Graphics	Растровая графика	Оптимизирован для хранения изображений, состоящих из линий
JPEG/JPG	Joint Photographic Experts Group	Растровая графика	Оптимизирован для фотографических изображений
TIFF	Tagged Image File Format	Растровая графика	Печатная продукция, точная цветопередача
RAW	Raw Image File	Растровая графика	Цифровая фотография, нуждается в постобработке
GIF	Graphics Interchange Format	Растровая графика	Устарел для статических изображений, приемлем для работы с анимацией

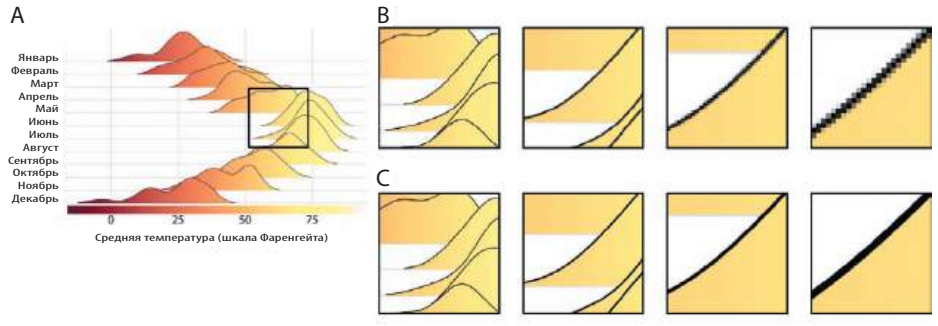


Рис. 26.1. Демонстрация ключевого различия между векторной графикой и растровой. A. Оригинальное изображение. Черный квадрат показывает область, которую мы увеличиваем на панелях B и C. B. Увеличенная версия оригинального изображения A в растровом формате. Как можно видеть, при увеличении пиксели становятся все более заметными. C. Увеличенная версия оригинального изображения A в векторном формате. Вне зависимости от степени увеличения четкость картинки остается идеальной

При всех свои преимуществах у векторной графики имеются два недостатка, которые зачастую становятся источником проблем при использовании такого типа изображений на практике. Во-первых, поскольку графические программы, в которых отображаются рисунки, рисуют векторную графику на лету, может случиться так, что один и тот же рисунок будет по-разному выглядеть в двух разных программах или на двух разных компьютерах. Чаще

всего эта проблема касается текста, например, когда требуемый шрифт недоступен и программное обеспечение для отрисовки графики заменяет его на какой-то другой шрифт. Обычно такая замена не влияет на читаемость текста, однако визуально результат выглядит не так, как это было задумано, и зачастую смотрится некрасиво. Существует несколько способов обойти эту проблему, таких как отрисовка символов шрифта как фигур (независимость от шрифтов) или встраивание нужных шрифтов в сам файл PDF, но для этого может потребоваться специальное программное обеспечение и/или специальные технические знания. Растровые изображения, напротив, всегда будут выглядеть одинаково.

Во-вторых, при создании больших и/или сложных визуализаций файл с векторной графикой может вырасти до огромных размеров, из-за чего его отрисовка будет идти очень медленно. Например, при отрисовке диаграммы рассеяния, состоящей из миллионов точек с координатами *x* и *y*, каждая точка будет отрисована в обязательном порядке, даже если она перекрыта и/или скрыта другими графическими элементами. Как следствие, размер файла может составлять много мегабайт, а появления рисунка на экране придется некоторое время ждать. Как-то раз, в начале 2000-х, когда я был молодым кандидатом наук, я создал файл PDF, для открытия которого Acrobat Reader требовался почти целый час. Несмотря на то что современные компьютеры работают намного быстрее, а время рендеринга измеряется миллисекундами, даже длящаяся несколько секунд отрисовка картинки может весьма негативно отражаться на рабочем процессе. Например, так бывает при размещении векторного рисунка в документе большого размера, из-за чего программа чтения PDF-файлов зависает каждый раз, когда вы переходите на страницу с этим рисунком. Конечно, если речь идет о простых векторных изображениях с небольшим количеством элементов (скажем, с несколькими точками данных и кратким текстом), то они, как правило, будут занимать намного меньше места, а их отрисовка программой для просмотра изображений будет происходить быстрее отрисовки растровых изображений того же размера.

Сжатие растровой графики с потерями и без

В большинстве форматов растровых файлов используется некоторая форма сжатия данных, с помощью которой можно управлять размером файлов. Сжатие бывает двух видов: без потерь и с потерями. Сжатие без потерь гарантирует, что сжатое изображение будет пиксель в пиксель совпадать с оригиналом, тогда как сжатие с потерями допускает некоторое ухудшение качества изображения в обмен на меньший размер файла.

Давайте кратко посмотрим, как работают алгоритмы сжатия в каждом из этих случаев, чтобы знать, когда следует применять тот или иной тип сжатия

при работе с изображениями. Начнем со сжатия без потерь. Представьте себе изображение с черным фоном, где большие области полностью покрашены в черный цвет и поэтому на картинке имеется множество черных пикселей, которые располагаются рядом друг с другом. Каждый такой пиксель может быть представлен строкой из трех нулей: 0 0 0. Эти цифры означают нулевую интенсивность в красном, зеленом и синем цветовых каналах изображения. Области черного фона на изображении соответствуют тысячам нулей в файле изображения. Пусть где-то на изображении находятся 1000 последовательных черных пикселей, что соответствует 3000 нулям. Вместо того чтобы записывать все эти нули, мы могли бы просто указать их количество, например 3000 0. Таким образом мы передали ту же самую информацию, использовав для этого всего лишь два числа: количество значений (3000) и само это значение (0). На сегодняшний день существует множество хитрых приемов, направленных на решение задачи сжатия рисунков, и современные форматы изображений, допускающие сжатие без потерь (такие как PNG), могут сохранять растровые данные с впечатляющей эффективностью. Однако следует отметить, что алгоритмы сжатия лучше всего работают в тех случаях, когда на картинках присутствуют большие области однородного цвета, и именно по этой причине в табл. 26.1 PNG указан как лучший формат для рисования линий.

Что касается фотографий, то здесь дела обстоят значительно сложнее. Изображения такого типа редко содержат последовательности пикселей одинакового цвета и яркости. Здесь вступают в игру иные характеристики: градиенты и некоторые другие закономерности в разных цветовых шкалах. Поэтому в случае фотографических изображений сжатие без потерь далеко не всегда работает лучшим образом, вследствие чего в качестве альтернативы был разработан метод сжатия с потерями. Его ключевая идея заключается в том, что некоторые детали на изображении слишком мелки для человеческого глаза, вплоть до неразличимости, и ими можно пренебречь без явного ухудшения качества изображения. Рассмотрим в качестве примера градиент из 1000 пикселей, каждый из которых имеет немного отличающееся от соседа значение цвета. Если мы уменьшим количество цветов до 200 и каждой группе из 5 смежных пикселей назначим один и тот же цвет, то, скорее всего, преобразованный градиент будет выглядеть почти так же, как исходный.

Наиболее распространенным форматом сжатия с потерями является JPEG (см. табл. 26.1), поэтому многие фотокамеры по умолчанию сохраняют файлы в этом формате. Сжатие JPEG лучше всего показывает себя в работе с фотографиями, и мы нередко можем добиться значительного уменьшения размера файла при очень незначительном ухудшении качества изображения. Наиболее неблагоприятным сценарием для применения сжатия JPEG является случай, когда изображения содержат острые углы, например созданные

нарисованными линиями или текстом. В этих случаях сжатие посредством JPEG может привести к появлению очень заметных артефактов (рис. 26.2).

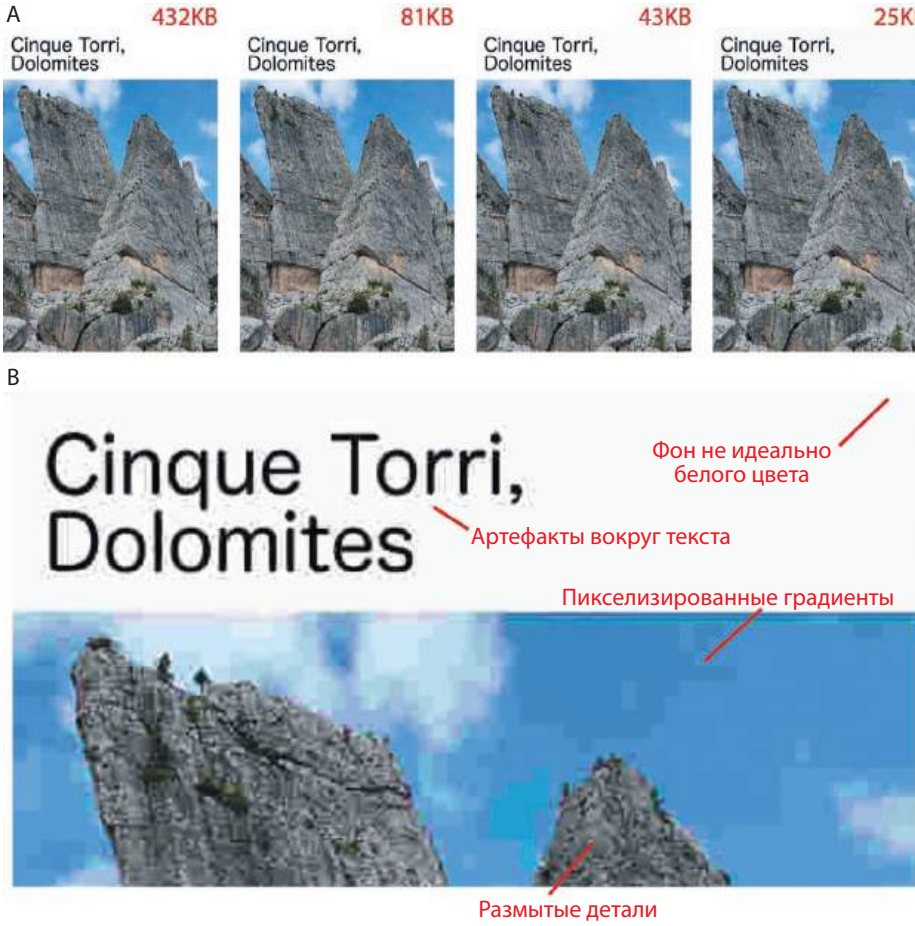


Рис. 26.2. Иллюстрация артефактов JPEG. A. Одно и то же изображение было сжато несколько раз, используя технологию сжатия JPEG. Выходной размер файла показан красным текстом над каждым изображением. Уменьшение размера файла в 10 раз, с 432 Кб в исходном виде до 43 Кб в сжатом, приводит лишь к незначительному заметному снижению качества изображения. Однако последующее уменьшение размера файла еще в два раза, всего до 25 Кб, приводит к появлению многочисленных видимых артефактов. B. При увеличении максимально сжатого изображения хорошо видны различные артефакты сжатия. Источник: Claus O. Wilke

При этом, даже если артефакты JPEG достаточно малы и не видны невооруженным глазом, они все равно могут стать причиной проблем, например, при печати. Поэтому следует избегать использования данного формата для рисунков, содержащих контурные изображения или текст, что чаще всего

встречается в подготовленных визуализациях или снимках экрана компьютера. Вместо этого используйте PNG или TIFF. Лично я использую формат JPEG только для фотографий. Если изображение содержит и фотографии, и контурные рисунки, и текст, все равно стоит отдать предпочтение PNG или TIFF. Самое плохое, что может случиться из-за использования этих форматов, — это большой размер выходных файлов, в то время как применение JPEG может в худшем случае попросту испортить вашу визуализацию.

Преобразования между форматами изображений

В большинстве случаев изображение из одного формата можно преобразовать в любой другой. Например, на Mac вы можете открыть картинку с помощью Preview, а затем экспортировать ее в несколько различных форматов. Однако в этом процессе может быть необратимо потеряна важная информация. Например, после сохранения векторной графики в растровом формате (скажем, преобразования PDF-файла в формат JPEG) независимость от разрешения, которая является ключевым преимуществом векторной графики, будет потеряна. И наоборот, сохранение изображения JPEG в файл PDF никоим образом не сделает из растровой картинку векторную. Вы получите все то же растровое изображение, только теперь оно хранится внутри файла PDF. Аналогично преобразование файла JPEG в файл PNG не позволит избавиться от артефактов, которые уже могли возникнуть при сжатии JPEG.

Существует эмпирическое правило: исходное изображение следует сохранять в формате, обеспечивающем максимальное разрешение, точность и гибкость. Поэтому для визуализации данных лучшим выбором будет либо PDF, который впоследствии можно преобразовать в PNG или JPEG, либо PNG с высоким разрешением. Аналогичный совет можно дать и для изображений, которые доступны только в виде растровой графики, такие как цифровые фотографии, — сохраняйте их в таком формате, который не использует сжатие с потерями, или, если это невозможно, выбирайте как можно меньшую степень сжатия. Кроме того, изображения следует сохранять в максимально высоком разрешении, чтобы впоследствии при необходимости вы могли уменьшить их до любого нужного размера.

Глава 27

Как выбрать подходящее программное обеспечение для визуализации

На протяжении всей этой книги я намеренно избегал обсуждения одного критически важного вопроса визуализации данных: с помощью каких инструментов следует создавать графики? Этот вопрос легко может вызвать ожесточенные споры, потому что многие люди эмоционально привязаны к тем инструментами, с которыми они постоянно работают. Я часто видел, как люди яростно защищают свои предпочтения, вместо того чтобы потратить это время на изучение какого-либо нового подхода, даже если у него есть очевидные преимущества. Разумеется, любой человек имеет право вести себя так, как считает нужным, и использование знакомых инструментов тоже вполне оправданно. Изучение любого нового инструмента требует времени и усилий, и вам придется пережить болезненный переходный период, во время которого вы будете испытывать дискомфорт, выполняя привычную работу непривычным способом. Стоит ли этот период затраченных усилий, можно понять лишь после того, как вы уже это сделаете и вложите силы и время в изучение нового инструмента. Поэтому, независимо от плюсов и минусов различных инструментов и подходов, главный принцип заключается в том, чтобы выбрать инструмент, подходящий именно для вас. Если с помощью имеющегося инструмента у вас получается без лишних усилий делать визуализации устраивающего вас качества, то это главное.

Лучшее программное обеспечение для визуализации — то, которое позволяет вам создавать необходимые вам изображения.

Принимая во внимание вышесказанное, я считаю, что существует набор общих критериев, по которым мы можем оценивать преимущества и недостатки различных подходов к созданию визуализаций. Эти критерии можно разбить на несколько групп: насколько получающиеся визуализации легко

воспроизводимы, насколько просто и быстро инструмент позволяет исследовать данные и насколько широким набором настроек внешнего вида выводимых изображений инструмент обладает.

Воспроизводимость и повторимость

В контексте научных экспериментов мы называем работу *воспроизводимой*, если результат эксперимента остается неизменным в том случае, когда другая исследовательская группа проводит аналогичное исследование. Например, если одна группа ученых обнаружит, что новое обезболивающее лекарство показывает значительное облегчение головной боли без возникновения заметных побочных эффектов, а другая группа, изучив впоследствии это же лекарство на другой группе пациентов, получит сходные результаты, то работу можно отнести к категории воспроизводимых. Работа считается *повторимой*, если очень похожее или идентичное измерение может быть получено одним и тем же лицом, повторяющим эту же процедуру измерения на одном и том же оборудовании. Например, если я взвешиваю свою собаку и обнаруживаю, что она весит 41 фунт, а затем еще раз взвешиваю ее на тех же весах и снова обнаруживаю, что она весит 41 фунт, то это измерение относится к категории повторимых.

С небольшими изменениями мы можем применить эти концепции к визуализации данных. Визуализация воспроизводима, если доступны исходные данные, нанесенные на график, а все преобразования, сделанные до построения графика, точно указаны. Например, если вы сделаете график, а затем отправите мне точные данные о том, как вы его построили, то я смогу сделать практически такой же. Поскольку каждый из нас выберет какой-то свой шрифт, цвета точек и их размеры, наши графики не будут выглядеть как близнецы, однако и ваш, и мой графики отображают одни и те же данные одним и тем же способом и поэтому фактически воспроизводят друг друга. С другой стороны, визуализацию можно назвать повторяемой, если из необработанных данных можно воссоздать график с точно таким же внешним видом, как у оригинальной версии, вплоть до последнего пикселя. Строго говоря, условие повторяемости требует, что должна существовать возможность воссоздания всех нюансов исходного варианта, вплоть до таких случайных элементов, как джиттеринг (глава 17). Для случайных данных повторяемость обычно требует, чтобы мы указали конкретный генератор случайных чисел и заданное для него начальное значение.

В этой книге мы видели много примеров рисунков, которые воспроизводят, но не повторяют другие визуализации. Например, в главе 24 представлены несколько наборов изображений, каждое из которых показывает одни и те же данные, но при этом несколько отличается от остальных рисунков.

Аналогично рис. 27.1A является повторением рис. 8.7, вплоть до случайного джиттеринга точек, который был применен ко всем точкам данных, тогда как рис. 27.1B является всего лишь его воспроизведением. На рис. 27.1B джиттеринг отличается от показанного на рис. 8.7, плюс ко всему, дизайн графиков довольно разный. Поэтому эти два рисунка совсем непохожи друг на друга, несмотря на то что они отображают одни и те же данные.

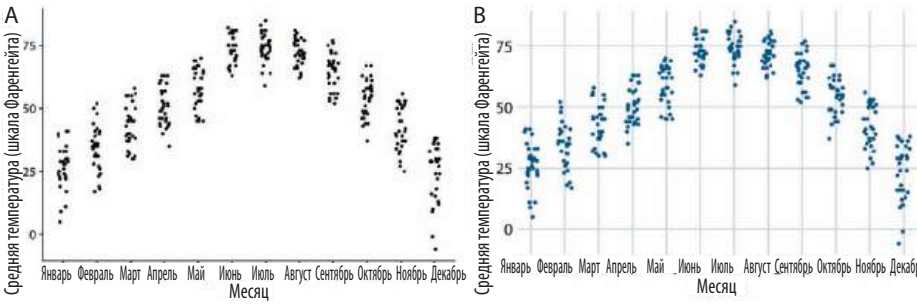


Рис. 27.1. Повторение и воспроизведение рисунка. Часть A является повторением рис. 8.7. Эти два графика идентичны вплоть до величин джиттеринга, который был применен ко всем точкам графика. Напротив, часть B является воспроизведением, а не повторением. В частности, джиттеринг в части B отличается от джиттеринга в части A или на рис. 8.7. Источник: Weather Underground

Порой достичь воспроизводимости и повторяемости может быть довольно сложно, особенно если речь идет о работе с интерактивным программным обеспечением для построения графиков. Большинство таких программ позволяют вам преобразовывать данные или каким-то иным способом манипулировать ими, но при этом не отслеживают каждую отдельно взятую операцию, а только сохраняют конечный продукт. Если вы сделаете визуализацию с помощью программы такого рода, а затем кто-то попросит вас воспроизвести ваш график или создать похожий, но на основе другого набора данных, у вас могут возникнуть трудности. Когда я, будучи еще молодым кандидатом наук и ассистентом на кафедре, создавал свои научные визуализации с помощью интерактивной программы, я часто сталкивался с упомянутой проблемой. Например, я сделал несколько визуализаций для научной рукописи. Спустя несколько месяцев я решил перечитать рукопись и в процессе чтения захотел сделать слегка измененную версию одного из изображений, но, приступив к работе, я понял, что не помню в точности все шаги, которые привели меня к созданию оригинального варианта. Этот опыт научил меня по возможности избегать интерактивных программ. Теперь я делаю визуализации программно, путем написания кода (скрипта), который генерирует диаграммы из необработанных данных. Сгенерированные программным способом рисунки, как правило, могут быть воспроизведены любым человеком,

у которого есть строящие диаграмму скрипты, а также доступ к среде разработки языка, на котором скрипты написаны, и необходимым для работы скрипта программным библиотекам.

Исследование данных и представление данных

Процесс визуализации данных состоит из двух этапов, значительно отличающихся друг от друга. Первый этап — это, собственно, изучение или исследование данных. Всякий раз, когда вы приступаете к работе с новым набором данных, вам нужно посмотреть на него под разными углами и опробовать несколько способов его визуализации просто для того, чтобы выделить ключевые особенности исследуемого набора. На этом этапе скорость и эффективность имеют первостепенное значение. Вам нужно попробовать разные типы визуализаций, разные преобразования данных и поработать с несколькими подмножествами данных. Чем быстрее вы сможете различными способами взглянуть на данные, тем больше информации о самом наборе вы сможете получить и тем выше вероятность того, что вы не пропустите какую-либо важную особенность этих данных. Второй этап — это представление данных. Переход к этому этапу должен происходить лишь после того, как вы почувствуете, что вы поняли основные свойства вашего набора данных и вы знаете, какие именно его аспекты вы хотите показать своей аудитории. Основная задача на этом этапе заключается в подготовке высококачественной, готовой к публикации диаграммы, которую можно напечатать в статье или книге, включить в презентацию или разместить в интернете.

На этапе исследования внешний вид графика имеет второстепенное значение. Ничего страшного, если подписи к осям отсутствуют, легенда не соответствует графику или символы слишком маленькие; главное на этом этапе — выявить основные закономерности в данных. Критичным для этого является то, насколько легко вы можете манипулировать способами отображения данных. Чтобы всесторонне исследовать данные, вы должны быть в состоянии, например, быстро перейти от диаграммы рассеяния к графикам плотности с перекрытием, от них к коробчатой диаграмме и далее к тепловой карте. В главе 1 мы говорили о том, что все визуализации представляют собой отображение данных на эстетические элементы. Удачно спроектированный инструмент исследования данных позволит вам легко и быстро настраивать отображение переменных на элементы эстетики и предоставит широкий выбор вариантов визуализации, связав все общей парадигмой. Однако, по моему опыту, многие инструменты визуализации (в особенности библиотеки для программной генерации изображений) не предоставляют таких возможностей. Вместо этого они чаще всего структурированы на основе типов графиков, где каждый тип графика для построения требует

передачи специфических данных, да еще и в специфичной только для него форме. Такие инструменты могут мешать эффективному исследованию данных, поскольку невозможно постоянно удерживать в памяти все особенности всех типов визуализаций. Я призываю аккуратно оценивать возможности вашего программного обеспечения для визуализации: позволяет оно быстро исследовать данные или, наоборот, мешает. Если вы находите, что оно часто мешает и путает вас, то это хороший повод присмотреться к альтернативным вариантам визуализации.

После того как мы определились со способом визуализизации наших данных, преобразованиями, которые мы хотим сделать, и типом выходного графика, мы обычно будем переходить к подготовке изображения, пригодного для публикации. Здесь перед нами открывается несколько путей. Во-первых, мы можем построить готовый рисунок с помощью той же программной платформы, которую мы использовали для первоначального исследования. Во-вторых, мы можем воспользоваться каким-то другим программным обеспечением, которое позволит нам более полно контролировать конечный продукт, пусть даже если эта платформа затрудняет процесс изучения данных. В-третьих, с помощью программного обеспечения для визуализации мы можем создать черновой вариант рисунка, а затем вручную обработать его с помощью программы обработки изображений или иллюстраций, такой как Photoshop или Illustrator. В-четвертых, мы можем вручную перерисовать весь график с нуля — ручкой на бумаге или в графическом редакторе.

Все эти варианты имеют право на жизнь. И все же я хотел бы предостеречь вас от соблазна вручную дооформлять графики в процессе анализа данных или для научных публикаций. Вмешательство руками в процесс подготовки графика в значительной мере затрудняет и замедляет последующее повторение или воспроизведение диаграммы. По моему опыту работы в области естественных наук, редко когда нам надо нарисовать какую-то диаграмму всего лишь один раз. В ходе исследования мы можем повторять эксперименты, расширять исходный набор данных или раз за разом по новой ставить эксперимент, каждый раз со слегка измененными условиями. На практике мне не раз приходилось сталкиваться с ситуацией, когда на этапе почти полностью завершенной работы вносятся небольшие изменения в то, как мы анализируем данные, после чего все графики приходится перерисовывать. И также я видел, что в таких ситуациях нередко принималось решение отказаться от внесения изменений в анализ или от перерисовки графиков либо из-за того, что жаль усилий, потраченных на предыдущую работу, либо потому, что авторы оригинальных изображений уже ушли из проекта. Во всех этих сценариях созданию наилучшего исследования помешал неоправданно сложный и невоспроизводимый процесс визуализации данных.

При этом я не считаю смертным грехом создание графиков вручную или ручную обработку диаграмм, например, в виде изменения меток осей, добавления аннотаций или изменения цветов. Так мы можем добиться создания красивых и уникальных графиков, которые невозможно получить никаким другим способом. Несмотря на то что сложные и предельно выверенные компьютерные визуализации становятся все более распространенным явлением, я наблюдаю заметное возрождение интереса к графикам, нарисованным вручную (рис. 27.2). Я думаю, что причиной этого является желание придать своим визуализациям уникальный и персонализированный образ в противовес стерильному и банальному способу представления данных.

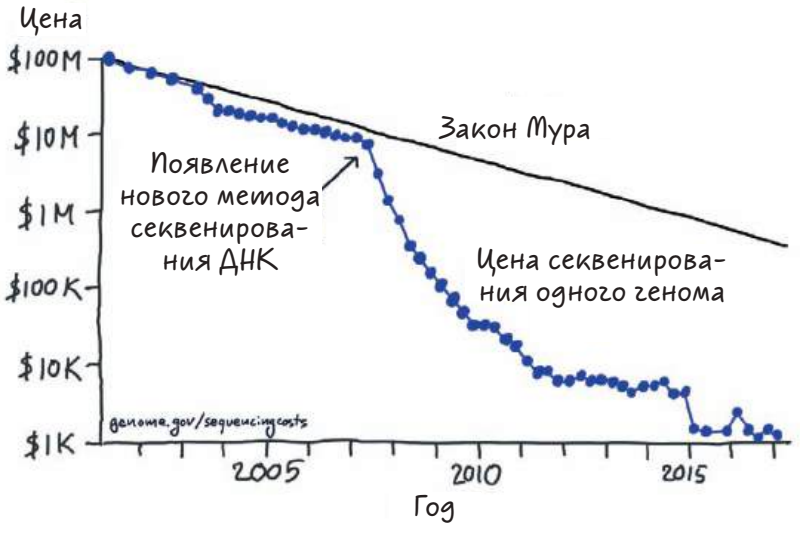


Рис. 27.2. После внедрения новых методов секвенирования стоимость секвенирования одного генома снижалась гораздо быстрее, чем это предсказывал закон Мура. Этот нарисованный от руки график воспроизводит широко распространенную визуализацию, подготовленную Национальными институтами здравоохранения США. Источник: National Human Genome Research Institute

Разделение содержания и дизайна

Хорошее программное обеспечение для визуализации данных должно позволять вам думать о дизайне ваших рисунков и об их содержании независимо друг от друга. Под содержанием я имею в виду конкретный отображаемый набор данных, произведенные над ним преобразования (если они имеются), указание на то, каким визуальным элементам на графике соответствуют какие данные, шкалы на диаграмме, диапазоны осей и тип графика (диаграмма рассеяния, линейчатый график, столбчатая диаграмма, коробчатая

диаграмма и т. д.). Под дизайном же понимаются цвета переднего плана и фона, параметры шрифта (например, размер шрифта, начертание и семейство), формы и размеры символов, наличие или отсутствие координатной сетки, а также расположение элементов рисунка (легенды, насечек на осях, названий осей и графиков). Приступая к работе над новой визуализацией, я, как правило, сначала определяю, каким должно быть содержание, используя тот метод быстрого исследования, о котором шла речь в предыдущем разделе. После того как содержание определено, я могу внести какие-либо изменения во внешний вид графика, хотя, скорее всего, воспользуюсь уже готовым вариантом дизайна, который мне понравился и/или который придает рисунку вид, согласующийся с другими диаграммами в контексте некоей большей научной работы.

В программном обеспечении ggplot2, с помощью которого я создал графики для этой книги, разделение содержания и дизайна достигается с помощью так называемых *тем*. Тема определяет внешний вид рисунка, и при этом темы очень легко применяются к уже существующим рисункам (рис. 27.3). Темы могут быть созданы другими пользователями или разработчиками и распространяться как пакеты R. Благодаря этому механизму вокруг ggplot2 выросла целая экосистема готовых загружаемых тем, которая охватывает широкий спектр различных стилей и сценариев применения. Если вы рисуете графики при помощи ggplot2, вы почти наверняка найдете тему, которая соответствует вашим потребностям в дизайне.

Разделение содержания и дизайна позволяет как ученым, так и дизайнерам сосредоточиться на том, что они делают лучше всего. Большинство специалистов по работе с данными не являются дизайнерами, и поэтому их главной заботой должна быть визуализируемая информация, а не внешний вид графика. Аналогично большинство дизайнеров не являются специалистами в исследовании данных, поэтому их задача заключается в предоставлении уникального и привлекательного визуального языка изображения, а не во вникании в конкретные данные, их преобразования и т. д. Тот же принцип разделения содержания и дизайна давно используется в издательском мире книг, журналов, газет и веб-сайтов, где авторы предоставляют контент, а верстской и дизайном занята отдельная группа людей — специалистов в области дизайна и полиграфии, — чьей задачей является создание визуально непротиворечивого и привлекательного стиля. Этот принцип логичен и полезен, но пока еще не так широко распространен в мире визуализации данных, как бы этого хотелось.

Подведу итог всему сказанному в этой главе: выбирая программное обеспечение для визуализации, думайте о том, сможете ли вы с его помощью легко воспроизводить рисунки и повторять их с обновленными или иным образом измененными наборами данных, сможете ли вы быстро исследовать

различные визуализации одних и тех же данных и в какой степени вы можете настраивать визуальный дизайн отдельно от содержания. В зависимости от уровня ваших навыков и умения программировать может быть полезно использовать различные инструменты визуализации на этапах исследования и представления данных, а окончательную визуальную настройку выполнять в интерактивном режиме или вручную — как вам будет удобнее. Если вам приходится создавать визуализации в интерактивном режиме, в частности с помощью программного обеспечения, которое не отслеживает примененные вами преобразования данных и визуальные настройки, подумайте о ведении некоего подобия журнала, где вы можете делать подробные записи обо всех изменениях рисунка, чтобы ваша работа оставалась полностью воспроизводимой.

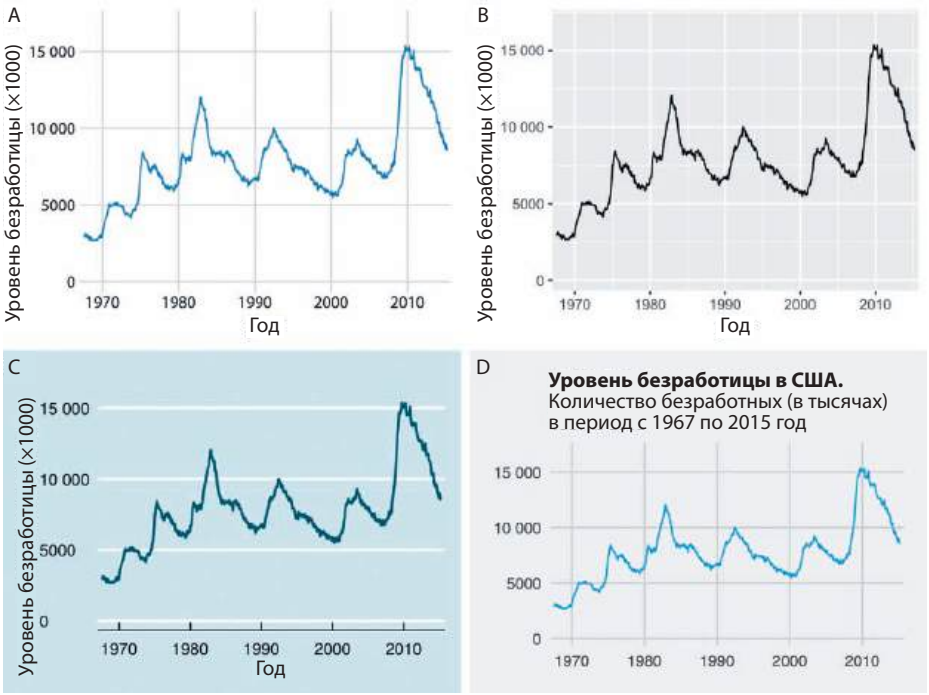


Рис. 27.3. Число безработных в США в период с 1970 по 2015 год. Один и тот же график отображен с помощью четырех разных тем ggplot2. A. Тема по умолчанию для этой книги. B. Тема по умолчанию для ggplot2, программного обеспечения для построения графиков, которое я использовал для создания всех изображений в этой книге. C. Тема, которая имитирует визуализации журнала Economist. D. Тема, которая имитирует визуализации сайта FiveThirtyEight. FiveThirtyEight часто отказывается от подписей к осям в пользу заголовков и подзаголовков, поэтому я скорректировал рисунок соответствующим образом

Глава 28

Как рассказать историю и донести свою мысль

В большинстве случаев целью визуализирования данных является коммуникация. У нас есть понимание некоторого набора данных, потенциальная аудитория, и мы хотели бы донести нашу информацию до нашей аудитории. Для успешного решения этой задачи нам нужно представить аудитории понятную и интересную для них историю. Необходимость рассказывать истории может смутить ученых и инженеров, которые могут посчитать, что речь идет о придумывании, приукрашивании или преувеличении результатов исследований. На самом деле это совсем не так: слово «история» здесь означает не приукрашивание, а связное повествование — человеческая память и механизмы рассуждения устроены так, что истории воспринимаются ими намного лучше сухих фактов. Хорошая история нас вдохновляет, а плохая или полное ее отсутствие — вызывает скуку. Более того, любая коммуникация уже сама по себе история, которая создается в умах слушателей. Если нам не удастся представить аудитории связный рассказ, то наши слушатели придумают его себе сами. В лучшем случае эта история будет достаточно похожа на то, что мы хотели донести. Однако зачастую все складывается далеко не так удачно. История, созданная восприятием слушателей, может выглядеть как «это скучно», «автор исследования ошибается» или даже «да он просто не понимает, о чем говорит».

Рассказывая историю, вы должны с помощью фактов и логических рассуждений заинтересовать и увлечь аудиторию вашим материалом. Позвольте мне рассказать вам историю о физике-теоретике Стивене Хокинге. В возрасте 21 года — через год после получения докторской степени — ему был поставлен диагноз «заболевание двигательных нейронов» и дано два года жизни. Хокинг отказался смиряться с таким положением дел и направил всю свою энергию в науку. В конечном итоге он дожил до 76 лет и стал одним из самых влиятельных физиков своего времени, создав все свои фундаментальные работы, будучи глубоким инвалидом. На мой взгляд, это весьма впечатляющая история. При этом она полностью основана на фактах и является абсолютно правдивой.

Что такое история?

Прежде чем мы перейдем к обсуждению стратегий превращения визуализаций в истории, мы должны разобраться, что вообще такое история. История — это набор наблюдений, фактов или событий, истинных или выдуманных, которые представлены в определенном порядке, вызывающем эмоциональный отклик у аудитории. Последний формируется через нагнетание напряжения в начале истории, за которым следует какое-то разрешение в конце. Переход от напряжения к разрешению называется сюжетной линией, и у каждой хорошей истории есть четкая, узнаваемая сюжетная линия.

Опытные авторы знают, что существуют стандартные схемы повествования, которые учитывают ход мысли большинства людей. Например, мы можем рассказать историю, используя формат «открытие — вызов — действие — решение». Именно этот формат я использовал, рассказывая историю Хокинга. В начале истории я обозначил ее тему — история физика Стивена Хокинга. Затем перешел к вызову — диагностика заболевания двигательных нейронов в возрасте 21 года. Далее последовало действие — неудержимая преданность ученого науке. Наконец, я представил решение, в котором Хокинг прожил долгую и успешную жизнь, став одним из самых влиятельных физиков своего времени. Существуют и другие сюжетные схемы, которые используются так же широко. Газетные статьи часто следуют формату «введение — развитие — решение» или попросту «введение — развитие», где введение сразу передает основной посыл, а последующий материал предоставляет дополнительную информацию. Если бы мы хотели рассказать историю Хокинга в этом формате, мы могли бы начать с предложения, такого как «Влиятельный физик Стивен Хокинг, который произвел революцию в нашем понимании черных дыр и космологии, пережил прогноз своих врачей на 53 года и сделал все свои наиболее влиятельные работы, будучи глубоким инвалидом». Это введение. В развитии мы могли бы более подробно описать жизнь, болезнь и преданность Хокинга науке. Еще один формат — «действие — предыстория — развитие — кульминация — конец», который развивает историю немного быстрее, чем «открытие — действие — решение», но медленнее, чем «введение — развитие». В этом случае мы могли бы начать с такого предложения, как «Молодой Стивен Хокинг, столкнувшийся с катастрофической инвалидностью и перспективой ранней смерти, решил посвятить свою жизнь науке и попытаться добиться в ней успеха, насколько хватит сил и времени». Цель этого формата — привлечь аудиторию и создать эмоциональную связь на ранней стадии, но без немедленного раскрытия финала.

Я не ставлю своей целью рассказать вам в этой главе обо всех стандартных формах повествования. Куда лучше меня с этой задачей справятся

прекрасные специализированные ресурсы; для ученых и аналитиков я особенно рекомендую книгу Джошуа Шимеля Writing Science [Schimel, 2011]. Вместо введения в теорию повествования я хочу поговорить о том, как мы можем включить визуализацию данных в сюжетную линию нашей истории. Самое главное здесь то, что в большинстве случаев мы не сможем с помощью лишь одной (статичной) визуализации рассказать всю историю целиком. Визуализация может служить иллюстрацией открытию, вызову, действию или решению, но не всем этим частям одновременно. Для рассказа всей истории целиком нам обычно требуется несколько визуализаций. Например, создавая презентацию, мы можем сначала показать какой-нибудь фон или мотивационный материал, затем график, который создает вызов, и в конечном итоге еще одну, другую визуализацию, которая показывает решение. Если работа носит исследовательский характер, мы можем развернуть перед зрителями последовательность рисунков, которые все вместе образуют убедительную сюжетную линию. Однако нет ничего невозможного и в том, чтобы «сжать» всю сюжетную линию в одно изображение. Такой рисунок должен содержать вызов и решение одновременно, что сравнимо с сюжетной линией, начинающейся с введения.

Чтобы привести конкретный пример включения визуализации в историю, я расскажу историю на основе двух графиков. Первый создает проблему, а второй служит решением. Контекст моей истории — рост числа препринтов в области биологических наук (см. также главу 12). Препринты — это черновики, которыми ученые делятся со своими коллегами перед официальным рецензированием и официальной публикацией. Ученые обменивались черновиками своих работ практически с момента возникновения самого понятия научной работы. Однако в начале 1990-х годов, с появлением интернета, физики поняли, что гораздо эффективнее хранить и распространять черновики посредством единого хранилища. Так на свет появился сервер препринтов — веб-сервер, где ученые могут загружать, скачивать и искать черновики научных работ.

Одним из серверов препринтов, используемых до сих пор, является arXiv.org. Довольно быстро сервис начал расширяться, приобретя популярность и в смежных областях, включая математику, астрономию, информатику, статистику, количественные финансы и количественную биологию. В данном случае меня интересуют препринты раздела количественной биологии (q-bio) сайта arXiv.org. В период с 2007 года по конец 2013 года количество заявок ежемесячно росло в геометрической прогрессии, но затем рост внезапно прекратился (рис. 28.1). Очевидно, что в конце 2013 года произошло что-то такое, что радикальным образом повлияло на рост числа заявок по количественной биологии. Что же вызвало столь резкое изменение в количестве подаваемых материалов?

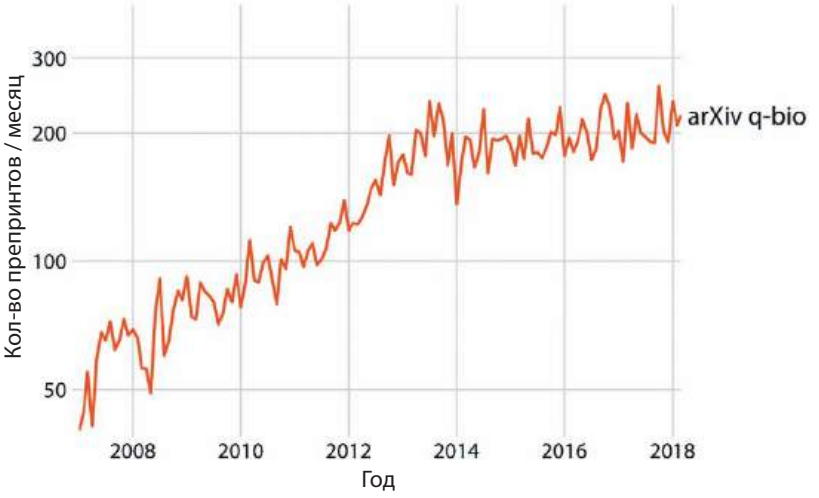


Рис. 28.1. Рост числа ежемесячных заявок в разделе количественной биологии (q-bio) сервера препринтов arXiv.org. В 2014 году наблюдается резкое изменение темпов роста. Если до 2014 года популярность сервиса росла как на дрожжах, то в период с 2014 по 2018 год наблюдается полная стагнация. Обратите внимание, что ось *y* является логарифмической, поэтому линейное увеличение *y* соответствует экспоненциальному росту количества препринтов. Источник: Jordan Anaya, www.prepubmed.org

Итак, моя мысль состоит в том, что конец 2013 года знаменует собой момент, когда количество препринтов в области количественной биологии выросло многократно, и по иронии судьбы это привело к замедлению роста архива q-bio. В ноябре 2013 года лаборатория Cold Spring Harbor (CSHL) Press запустила сервер препринтов в области биологии bioRxiv. CSHL Press — издатель, пользующийся большим уважением среди биологов. Поддержка CSHL Press оказала большую помощь с принятием большого количества препринтов в целом и сервера bioRxiv в частности среди биологов. Те же самые биологи, которые весьма подозрительно относились к arXiv.org, очень тепло приняли bioRxiv. В результате bioRxiv быстро получил признание среди представителей этой науки, чего arXiv в такой степени не довелось испытать ни разу. Фактически вскоре после своего запуска bioRxiv начал показывать быстрый экспоненциальный рост количества препринтов, и замедление роста заявок на arXiv q-bio в точности совпадает с началом экспоненциального роста bioRxiv (рис. 28.2). Скорее всего, большинство специалистов по количественной биологии, которые могли бы отправить препринты в q-bio, решили поместить их в bioRxiv.

Так выглядит моя история о препринтах в области биологии. Я намеренно рассказал ее с помощью двух изображений, хотя второй график (см. рис. 28.2) полностью включает в себя первый (см. рис. 28.1). На мой взгляд, наибольшей

отдачи от этой истории можно добиться, разделив ее на две части, и, если бы мне пришлось с ней выступать, я бы структурировал свой рассказ именно таким образом. Однако рассказать эту историю можно и с помощью одного изображения (см. рис. 28.2). Такая версия может оказаться более подходящей для среды, аудитория которой характеризуется очень небольшим периодом внимания к материалу, например социальные сети.

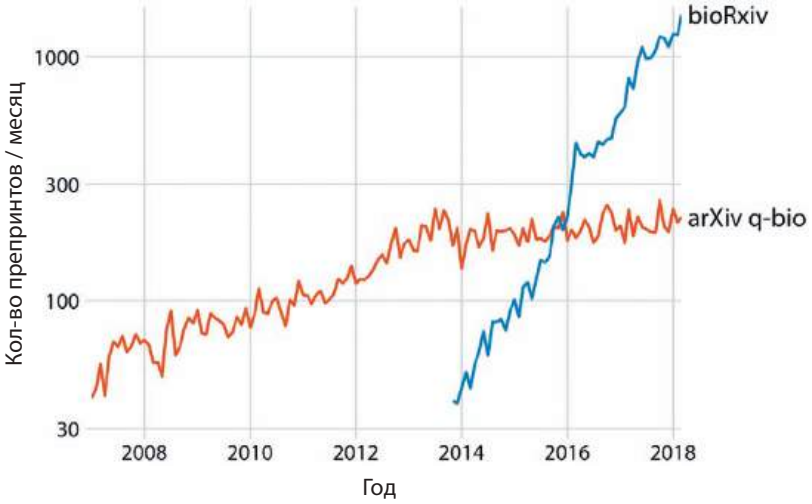


Рис. 28.2. Остановка роста числа заявок в q-bio совпала с появлением сервера bioRxiv. На графике показан рост числа ежемесячных заявок в раздел q-bio универсального сервера препринтов arXiv.org и на выделенный сервер биологических препринтов bioRxiv. Сервер bioRxiv начал функционировать в ноябре 2013 года, и количество отправляемых на него препринтов росло в геометрической прогрессии. По всей видимости, большинство ученых, которые могли бы представить препринты в q-bio, предпочли отправить их на bioRxiv. Источник: Jordan Anaya, http://www.prepubmed.org/

Создавайте визуализации «для генералов»

В оставшейся части этой главы я буду говорить о стратегиях создания отдельных рисунков и их наборов, которые помогут аудитории проникнуться вашей историей и сохранить вовлеченность в рассказ на всем его протяжении. Первое и самое главное правило заключается в том, что вы должны показать аудитории такие графики, которые они точно могут понять. Даже если вы будете безукоризненно следовать всем рекомендациям, которые я дал в этой книге, все равно может быть так, что получившийся у вас график будет неверно понят аудиторией или вызовет у нее недоумение. Если именно это с вами и произошло, вероятно, вы стали жертвой двух распространенных заблуждений: во-первых, что аудитории будет достаточно одного взгляда на график, чтобы

понять посыл всего вашего выступления, и во-вторых, что аудитория может быстро обрабатывать сложные визуализации и на лету схватывать ключевые тенденции и взаимосвязи, которые показаны на диаграммах. Ни одно из этих предположений не соответствует действительности. Нам нужно сделать все возможное, чтобы помочь нашим читателям понять смысл наших визуализаций и увидеть все те закономерности в данных, которые видим мы. Обычно это сводится к формуле «лучше меньше, да лучше». Старайтесь по максимуму упрощать ваши графики. Удалите все элементы, которые не имеют отношения к рассказываемой истории. Должны остаться только наиболее важные моменты. Я называю эту концепцию «рисованием картинок для генералов».

В течение нескольких лет я руководил крупным исследовательским проектом, который финансировала армия США. Руководители программы предупредили меня, чтобы в подаваемые мной ежегодные отчеты я включал минимум графиков, а каждый вошедший в доклад рисунок должен был предельно ясно показывать, в чем именно состоят достижения нашего проекта. Генерал, как мне сказали руководители программы, должен один раз посмотреть на график и сразу понять, как то, что мы делали, улучшило или превзошло предыдущие достижения. Тем не менее, когда мои коллеги, которые были частью этого проекта, присылали мне данные для ежегодного отчета о ходе работ, многие из них не соответствовали этому критерию. Графики, как правило, были слишком сложными, пестрили запутанными техническими терминами или вообще не имели никакого очевидного смысла. Большинство ученых не обучены делать визуализации для генералов.

Создавая визуализацию, не рассчитывайте на то, что аудитория с лету разберется в сложном графике.

Услышав эту историю, кто-то может решить, что генералы не очень умны или просто не разбираются в науке. Я думаю, что это не так. Генералы просто очень заняты. У них нет возможности тратить полчаса на то, чтобы расшифровать загадочное изображение. Когда они дают ученым миллионы долларов из карманов налогоплательщиков для проведения фундаментальных исследований, самое малое, чего они могут ожидать взамен, — это горстка предельно четких демонстраций того, что было сделано что-то стоящее и интересное. Кроме того, этот рассказ не следует понимать как рассказ о военном финансировании, в частности. Генерал — это метафора любого человека, до которого вы хотите что-то донести с помощью своей визуализации: научного рецензента для вашей статьи или заявки на грант, редактора газеты, вашего руководителя или начальника вашего руководителя в компании, где вы работаете. Если вы хотите, чтобы ваша история оказала нужное

воздействие на зрителя, вы должны сделать графики, которые подходят для ваших генералов.

По иронии судьбы первое, что встанет на пути создания визуализации для генералов, — это легкость, с которой современное программное обеспечение позволяет нам создавать сложные визуализации данных. Почти безграничные возможности таких программ провоцируют нас поместить на график как можно больший объем информации. Это и вправду так, я вижу явную тенденцию в сторону увеличения сложности в мире визуализации данных. Да, эти визуализации могут выглядеть очень впечатляюще, но вот помогут ли они нам рассказать историю? Рассмотрим рис. 28.3, на котором показаны задержки прибытия для всех рейсов, которые вылетали из Нью-Йорка в 2013 году. Думаю, вам наверняка понадобится некоторое количество времени, чтобы разобраться в этом графике.

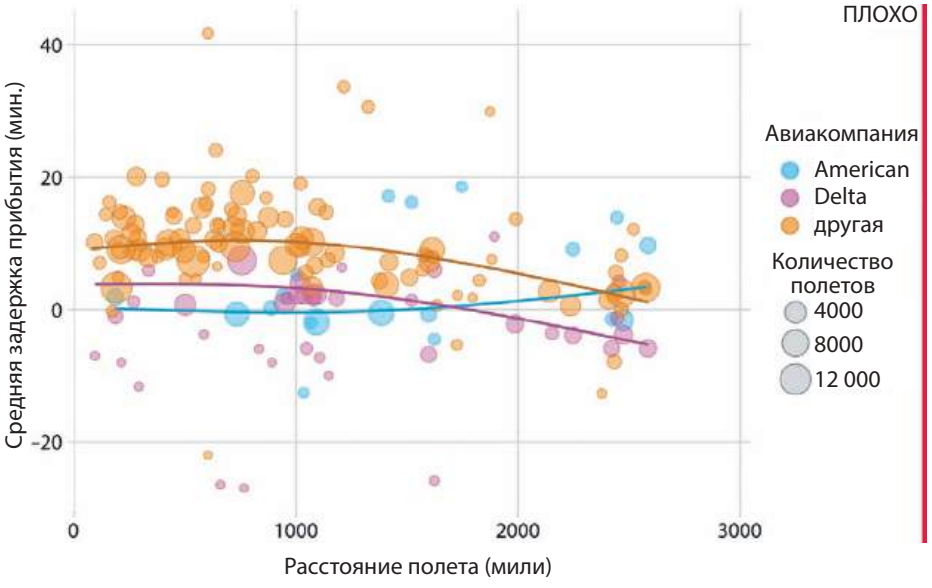


Рис. 28.3. Средняя задержка прибытия самолетов в зависимости от дальности полета до Нью-Йорка. Каждая точка представляет один пункт назначения, а размер точки обозначает количество рейсов из одного из трех основных аэропортов Нью-Йорка (Ньюарк, Аэропорт им. Джона Кеннеди или Ла-Гуардия) в этот пункт назначения в 2013 году. Отрицательные задержки означают, что рейс прибыл раньше заявленного времени. Сплошные линии отображают тенденции между задержкой прибытия и расстоянием. Delta имеет стабильно более низкие задержки прибытия по сравнению с другими авиакомпаниями, независимо от дальности полета. American имеет одну из самых низких задержек на коротких расстояниях и при этом одну из самых высоких задержек на более длинных расстояниях. Эта визуализация относится к категории «плохих», потому что она слишком сложная. Большинство читателей сочтут ее запутанной и не смогут на интуитивном уровне понять, о чем идет речь на этом графике. Источник: US Dept. of Transportation, Bureau of Transportation Statistics

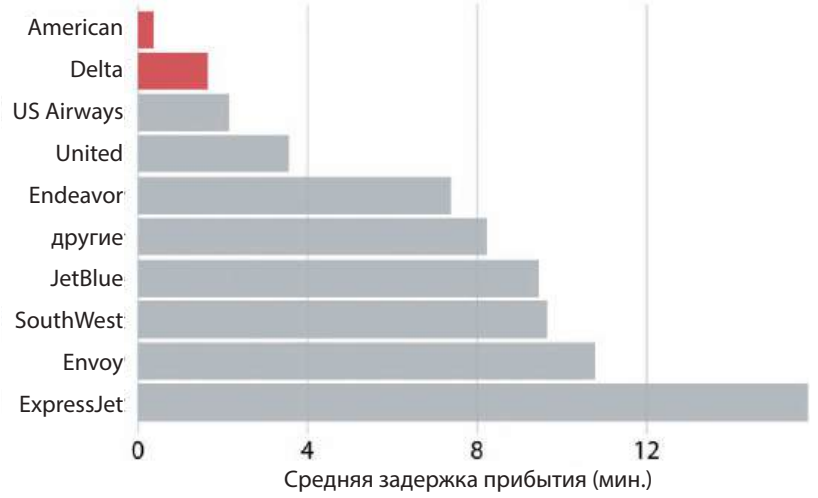


Рис. 28.4. Средняя задержка прибытия рейсов из Нью-Йорка в 2013 году в разрезе авиакомпаний. Авиакомпании American и Delta имеют самые низкие средние задержки прибытия среди всех авиакомпаний, выполняющих рейсы из Нью-Йорка. Источник: US Dept. of Transportation, Bureau of Transportation Statistics

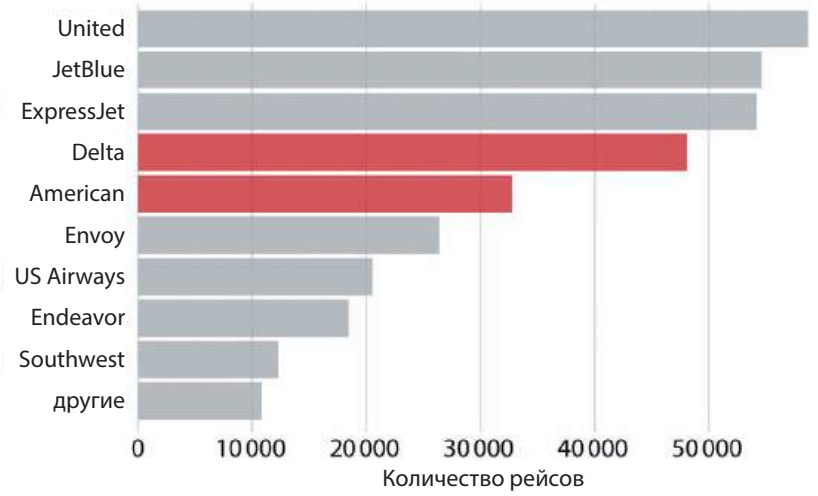


Рис. 28.5. Количество рейсов из Нью-Йорка в 2013 году в разрезе авиакомпаний. Delta и American занимают четвертое и пятое место соответственно по количеству рейсов из Нью-Йорка. Источник: US Dept. of Transportation, Bureau of Transportation Statistics

Я думаю, что наиболее важным посылом рис. 28.3 является то, что American и Delta имеют самые короткие задержки прибытия. Однако эту мысль можно было бы гораздо проще передать простой гистограммой (рис. 28.4). Таким образом, рис. 28.4 является наилучшим решением для визуализации истории о задержках прибытия самолетов авиакомпаний, даже если создание

этого графика не потребует от вас применения навыков визуализации данных. Если же вы задаетесь вопросом, связаны ли как-то эти небольшие задержки с тем, что самолеты этих авиакомпаний не так часто летают из района Нью-Йорк Сити, вы можете представить вторую гистограмму, которая подчеркнет, что American и Delta являются основными перевозчиками в данном регионе (рис. 28.5). Обе эти гистограммы отбрасывают переменную расстояния, показанную на рис. 28.3. Тем лучше. Нам не нужно визуализировать измерения данных, которые не связаны напрямую с нашей историей, даже если они у нас есть и даже если мы можем создать отображающую их визуализацию. Простое и понятное лучше, чем сложное и запутанное. Если вы пытаетесь показать слишком много данных одновременно, вы рискуете не показать вообще ничего.

Постепенный переход к сложным визуализациям

И все же в некоторых случаях нам и правда не обойтись без довольно сложных графиков, содержащих большое количество информации. В подобных сценариях мы можем помочь нашему зрителю, показав ему сначала упрощенную версию рисунка и только потом окончательный вариант графика со всеми нюансами. Я настоятельно рекомендую использовать подобный подход и для случая презентаций. Никогда не переходите сразу к самому сложному, начните с легкоусвояемого подмножества информации.

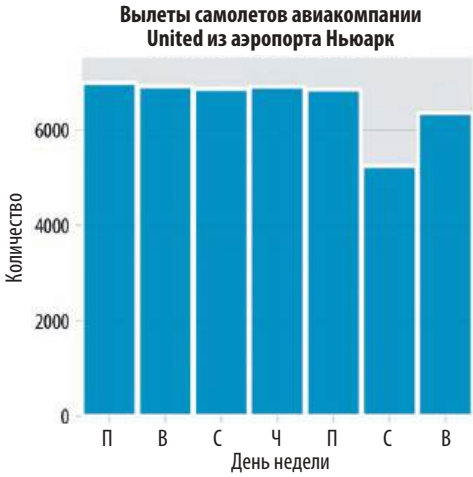


Рис. 28.6. Количество вылетающих самолетов авиакомпании United из аэропорта Ньюарк в 2013 году в разбивке по дням недели. В большинстве рабочих дней количество вылетов примерно одинаковое, но в выходные дни вылетов становится меньше. Источник: US Dept. of Transportation, Bureau of Transportation Statistics

Рис. 28.7. Вылеты авиарейсов из района Нью-Йорк Сити в 2013 году с разбивкой по авиакомпаниям, аэропортам и дням недели. Рейсы United Airlines и ExpressJet составляют большинство вылетов из аэропорта Ньюарк (EWR); JetBlue, Delta, American и Endeavour составляют большую часть вылетов из аэропорта имени Джона Кеннеди; а Delta, American, Envoy и US Airways составляют большую часть вылетов из Ла-Гуардия (LGA). У большинства авиакомпаний, но не у всех, в выходные дни вылетов меньше, чем в будни. Источник: US Dept. of Transportation, Bureau of Transportation Statistics

Эта рекомендация тем актуальнее, если в качестве финального графика будет использоваться малая панельная визуализация (см. главу 20), на которой показана сетка небольших графиков, имеющих схожую структуру. Понять полную сетку будет намного проще, если ранее аудитория видела один из ее элементов. Например, на рис. 28.6 показано общее количество вылетов United Airlines из аэропорта Ньюарк (EWR) в 2013 году с разбивкой по дням недели. Как только мы присмотримся к этому графику и разберемся в нем, нам будет гораздо проще воспринять подобного рода информацию уже для десяти авиакомпаний и трех аэропортов (рис. 28.7).

Визуализации должны быть запоминающимися

Преимущество простых и понятных графиков, таких как обычные столбчатые диаграммы, перед сложными заключается в том, что они не содержат отвлекающих факторов, легко читаются и позволяют зрителям сосредоточиться на наиболее важных моментах, которые вы и хотите донести. Однако у простоты тоже есть свой недостаток: слишком простые графики могут выглядеть слишком общими. У них нет никаких особенностей, которые бы выделяли их и которыми бы они запоминались аудитории. Если бы я быстро пролистал перед вами 10 гистограмм, вам было бы трудно различить их, а потом вспомнить, что же на них было показано. Например, если вы бегло взглянете на рис. 28.8, вы можете заметить визуальное сходство с рис. 28.5, приведенным ранее в этой главе. Тем не менее между этими двумя графиками нет ничего общего, кроме того, что они являются столбчатыми диаграммами. На рис. 28.5 показано количество рейсов из Нью-Йорка в разбивке по авиакомпаниям, а на рис. 28.8 показаны наиболее популярные домашние животные в домохозяйствах США. Ни на одном из этих графиков нет какого-либо элемента, который помогал бы читателю интуитивно понять, какой теме посвящен данный график, и поэтому эти рисунки вряд ли оставят какой-то след в памяти аудитории.

Исследования человеческого восприятия показывают, что люди склонны запоминать более визуально сложные и уникальные рисунки [Bateman et al., 2010; Borgo et al., 2012]. Однако визуальная уникальность и сложность не только влияют на запоминаемость, но и иногда препятствуют способности человека быстро получать информацию, а также затрудняют идентификацию небольших различий в значениях. Одна из крайностей — это прекрасно запоминающиеся, но при этом крайне запутанные графики. Даже если подобного рода рисунок выглядит как произведение искусства, это еще не делает его хорошей визуализацией данных. Другая крайность — это изображения, которые предельно понятны, но вместе с тем непримечательны и скучны,

и которые в итоге могут не оказать того эффекта, на который мы рассчитывали. В общем, нам необходимо отыскать баланс между двумя крайностями, чтобы наши визуализации были и запоминающимися, и понятными. (Однако целевая аудитория также имеет значение. Если график предназначен для научно-технической публикации, нас, как правило, меньше беспокоит запоминаемость диаграмм, чем если бы наш материал предназначался для печати в массовом печатном издании или блоге.)

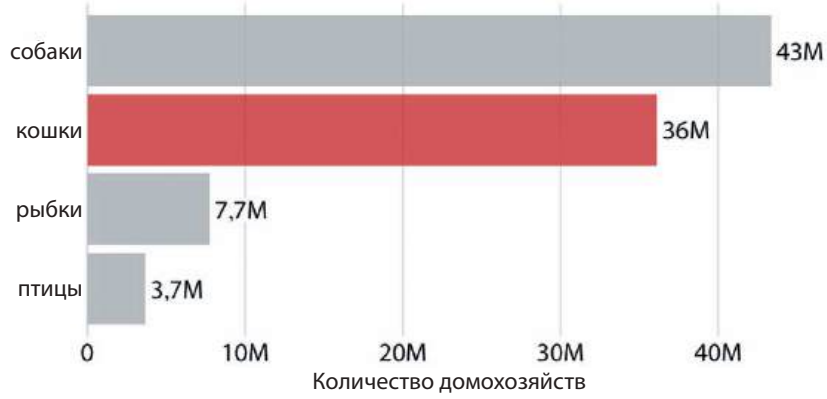


Рис. 28.8. Количество домашних хозяйств, имеющих одного или нескольких наиболее популярных домашних питомцев: собак, кошек, рыбок или птиц. Эта столбчатая диаграмма совершенно понятная, но абсолютно незапоминающаяся. Столбец «кошки» был выделен исключительно для создания визуального сходства с рис. 28.5. Источник: 2012 US Pet Ownership & Demographics Sourcebook, American Veterinary Medical Association

Мы можем сделать рисунок более запоминающимся, добавив в него визуальные элементы, которые отражают особенности данных, такие как рисунки или пиктограммы вещей или объектов, о которых идет речь. Один из распространенных подходов заключается в том, чтобы показывать значения данных в виде повторяющихся изображений так, чтобы каждая копия изображения соответствовала определенному количеству представленной переменной. Например, мы можем заменить столбцы с рис. 28.8 на повторяющиеся изображения собаки, кошки, рыбы и птицы, нарисованные в таком масштабе, чтобы каждое полное изображение животного соответствовало 5 миллионам домашних хозяйств (рис. 28.9). Таким образом, визуально рис. 28.9 все еще функционирует как гистограмма, однако теперь мы сделали его чуть более визуально сложным, что, в свою очередь, делает график более запоминающимся, и показали данные с использованием изображений, которые непосредственно отражают их значение. Достаточно беглого взгляда, чтобы отметить и запомнить, что собак и кошек на диаграмме намного больше, чем рыбок или птиц. Важно отметить, что в данном случае целью использования изображений

является представление данных, а не украшение диаграммы или создание подписей к осям. Психологические эксперименты показали, что последние варианты скорее отвлекают, чем помогают [Haroz, Kosara, Franconeri, 2015].

Рис. 28.9. Количество домохозяйств, имеющих одного или нескольких наиболее популярных домашних питомцев, показанных в виде диаграммы изотипа. Каждое полное изображение животного обозначает 5 миллионов домашних хозяйств, в которых живут соответствующие домашние животные. Источник: 2012 US Pet Ownership & Demographics Sourcebook, American Veterinary Medical Association

Такие визуализации, как рис. 28.9, часто называют диаграммами изотипа. Слово isotype является акронимом названия International System Of TYpographic Picture Education и, строго говоря, относится к упрощенным подобным логотипам пиктограммам, которые представляют объекты, животных, растения или людей [Haroz, Kosara и Franconeri, 2015]. Тем не менее я думаю, что имеет смысл использовать термин «диаграмма изотипа» более широко, чтобы применять его к любому типу визуализации, где для указания величины значения используются копии одного и того же изображения. Все-таки префикс «изо-» означает «один и тот же», а «тип» может означать определенный вид, класс или группу.

Будьте последовательны, но не повторяйтесь

Когда в главе 19 мы обсуждали составные визуализации, я упомянул, что очень важно использовать согласованный визуальный язык для всех частей большого изображения. То же самое верно и для всех остальных визуализаций, не только составных. Если мы создадим три графика, которые являются частью одной большой истории, то их визуальная составляющая должна быть подобрана таким образом, чтобы они выглядели как единое целое. Использование единого визуального языка не означает, однако, что все должно выглядеть

абсолютно одинаково. Напротив, важно, чтобы визуализации, описывающие различные анализы, были визуально различимыми, чтобы ваша аудитория могла легко распознать, где заканчивается один анализ и начинается другой. Лучшим способом этого достичь является использование разных подходов к визуализации разных частей общей истории. Если столбчатая диаграмма у вас уже была, задействуйте диаграмму рассеяния, коробчатую или линейную диаграмму. В противном случае различные анализы рискуют смешаться в сознании аудитории, и зрителям будет трудно отличить одну часть истории от другой. Например, если мы изменим рис. 20.8 из раздела «Составные визуализации» на с. 256 таким образом, чтобы на нем были нарисованы исключительно столбчатые диаграммы, мы получим значительно более запутанный график (рис. 28.10).

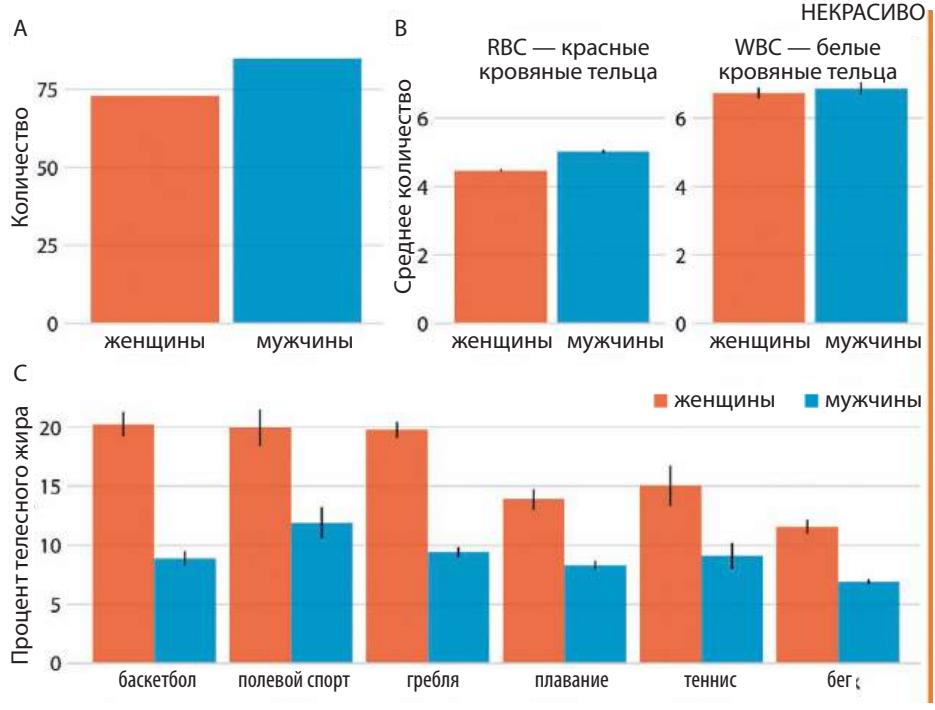


Рис. 28.10. Физиология и телосложение спортсменов мужского и женского пола. Планки погрешности показывают стандартную ошибку среднего. На графике слишком много повторений. Он показывает те же данные, что и рис. 20.8, и использует согласованный визуальный язык, но при этом на всех его панелях присутствует один и тот же тип визуализации (столбчатые диаграммы). Из-за этого читателю будет сложнее понять, что части A, B и C демонстрируют абсолютно разные выводы. Источник: [Telford and Cunningham, 1991]

При создании презентации или доклада старайтесь использовать различные типы графиков для каждого нового анализа.

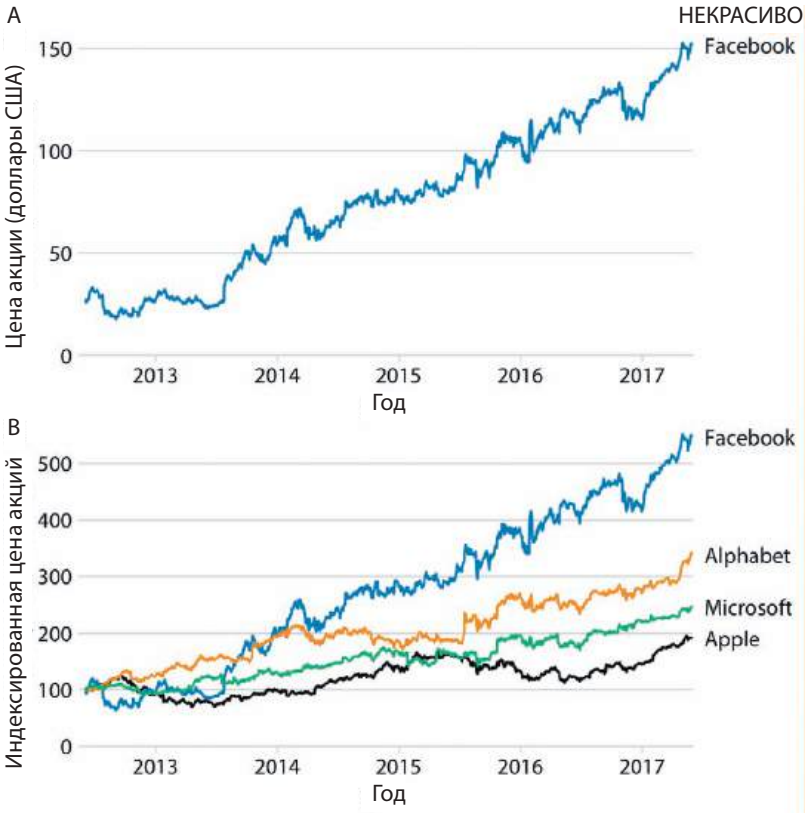


Рис. 28.11. Рост цены акций Facebook в течение пяти лет по сравнению с другими технологическими компаниями. A. Цена акций Facebook выросла с 25 долларов за акцию в середине 2012 года до 150 долларов за акцию в середине 2017 года. B. Цены других крупных технологических компаний выросли не настолько за тот же период времени. На 1 июня 2012 года цены были проиндексированы до значения 100, чтобы их можно было легко сравнивать. Эта визуализация относится к категории «некрасивых», потому что части A и B повторяют друг друга. Источник: Yahoo! Finance

Наборы повторяющихся визуализаций часто появляются при показе длинных историй из многих частей, где каждая часть основана на однотипных исходных данных. В таком случае неизбежно возникает соблазн использовать для каждой части один и тот же тип визуализации. Однако в совокупности эти рисунки вряд ли смогут привлечь внимание аудитории. Давайте в качестве примера рассмотрим историю о цене акций Facebook, состоящую из двух частей: 1) цена акций Facebook быстро выросла в период с 2012 по 2017 год; 2) рост цен Facebook опередил рост цен других крупных технологических компаний. Возможно, вы захотите визуализировать эти два утверждения с помощью двух графиков, которые показывают цену акций с течением времени, как показано на рис. 28.11. Обратите внимание, что у графика на рис. 28.11A

есть четкое предназначение и он должен оставаться неизменным, в то же время рис. 28.11B является повторяющимся и заслоняет главную мысль визуализации. Нас не особо интересует точное изменение во времени цен акций Alphabet, Apple или Microsoft; мы просто хотим подчеркнуть, что акции каждой из этих компаний меньше выросли в цене, чем акции Facebook.

Лично я бы оставил часть A как есть, а часть B заменил столбчатой диаграммой, показывающей процентное увеличение цены акций (рис. 28.12). Теперь у нас есть два разных графика, каждый из которых уникален и которые хорошо сочетаются друг с другом. Часть A позволяет читателю ознакомиться с необработанными исходными данными, а часть B подчеркивает величину эффекта, отказываясь при этом от показа любой второстепенной информации.

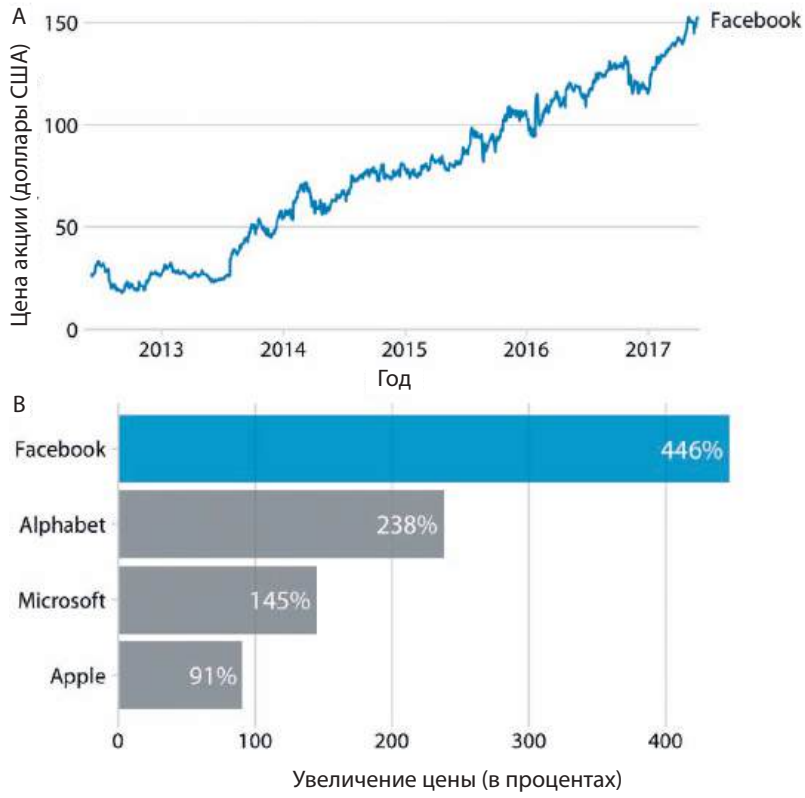


Рис. 28.12. Рост цены акций Facebook в течение пяти лет и сравнение их с акциями других технологических компаний. A. Цена акций Facebook выросла с примерно 25 долларов за акцию в середине 2012 года до 150 долларов за акцию в середине 2017 года, увеличившись почти на 450%. B. Цены акций других крупных технологических компаний не выросли так же сильно за тот же период времени. Рост их стоимости варьировался от 90 до почти 240%. Источник: Yahoo! Finance

Рис. 28.12 иллюстрирует тот общий принцип, которому я следую при подготовке визуализаций для рассказа: я начинаю с графика, максимально приближенного к отображению необработанных данных, а на последующих диаграммах я постепенно увеличиваю количество производных величин. Производные величины (такие как процентное увеличение, средние значения, коэффициенты подогнанных моделей и т. д.) полезны для подчеркивания основных тенденций в больших и сложных наборах данных. Поскольку эти данные выведены из исходных, они менее интуитивно понятны. Если мы покажем их прежде исходной информации, наша аудитория, скорее всего, попросту не сможет понять, какие именно мы сделали вычисления. С другой стороны, если мы попытаемся показать все тенденции на необработанных данных, нам в конечном итоге понадобится слишком много графиков и/или мы будем повторять сами себя.

Итак, существует ли какое-то магическое число или формула, чтобы понять, сколько графиков нам потребуется для рассказа своей истории? Увы, нет. Все зависит от того, где именно будет показана ваша визуализация. Для короткого поста в блоге или твита одного графика более чем достаточно. Для научных работ я рекомендую использовать от трех до шести визуализаций. Если ваш доклад содержит более шести изображений, то некоторые из них, возможно, стоит перенести в раздел приложений или дополнительных материалов. Документировать все собранные нами доказательства — хорошее решение, но мы не должны изнурять нашу аудиторию, заставляя зрителей интерпретировать чрезмерное количество похожих друг на друга рисунков. Однако в некоторых контекстах большее количество графиков может оказаться вполне подходящим вариантом. В таком случае нам, скорее всего, придется рассказывать несколько историй или одну большую, но с ответвлениями по ходу сюжета. Например, если меня просят выступить с часовой научной презентацией, я, скорее всего, постараюсь рассказать три разные истории. Аналогично в книге или диссертации наверняка будет больше одной истории, но при этом, как правило, по одной истории на главу или раздел. В этих сценариях каждая отдельная сюжетная линия или подсюжет должны быть представлены максимум тремя — шестью визуализациями. На протяжении всей этой книги я следую этому принципу на уровне разделов в главах, в чем вы можете убедиться, если перечитаете их. Каждый раздел самодостаточен и обычно содержит не более шести изображений.

Аннотированный список литературы

Ни одна книга не может охватить все, что нужно знать по данной теме. Я рекомендую вам ознакомиться и с другими текстами по визуализации данных, чтобы углубить свое понимание и улучшить технические навыки в создании визуализаций. Далее приведен небольшой список книг, которые я считаю интересными, заставляющими задуматься или полезными. Книги, перечисленные в первом разделе, наиболее схожи по тематике с настоящей книгой и могут содержать дополнительные или альтернативные точки зрения по темам, которые я затронул. Книги, перечисленные в разделе «Книги по программированию», посвящены важной теме создания визуализаций с использованием различных подходов к программированию и доступных программных библиотек. Во всех остальных разделах перечислены книги, которые расширят ваши знания о визуализации данных и улучшат ваши навыки в коммуницировании с помощью визуальных элементов и данных.

Размышления о данных и их визуализации

В следующих книгах рассматриваются мыслительные процессы и принятие решений, необходимые для преобразования данных в визуализации. Данные материалы являются текстами вводного уровня и посвящены принципам выбора визуализаций, а также рассказывают о подводных камнях, которых следует остерегаться.

Alberto Cairo. The Truthful Art. New Riders, 2016.

Отличное всестороннее введение в визуализацию данных, в частности для журналистов. В книге рассматриваются многие важные концепции визуализации данных, например принципы визуализации распределения, тенденции, неопределенность и карты. Во многих главах эта книга служит введением в основные принципы статистики, объясняя такие понятия, как совокупность, выборка и доверительный уровень.

Stephen Few. Show Me the Numbers. Analytics Press, 2012.

Книга о визуализации данных для профессионалов в области бизнеса. По охвату и целевой аудитории она схожа со следующей книгой, но содержит больше материала и более подробно освещает многие темы. Однако

в сравнении с трудом Коул Нуссбаумер Кнафлик (см. далее) эта книга не так хорошо написана и тщательно подготовлена.

Cole Nussbaumer Knaflic. Storytelling with Data. John Wiley & Sons, 2015.

Хорошо написанная и скрупулезно подготовленная книга о том, как превращать данные в графики. Основная аудитория книги — люди, создающие деловую графику. В своей области данная книга является отличным справочником, однако она не охватывает многие темы, которые важны для ученых, такие как визуализация распределений, тенденций или неопределенности.

Книги по программированию

Все перечисленные ниже книги рассказывают о визуализации данных с точки зрения создания кода.

Kieran Healy. Data Visualization: A Practical Introduction. Princeton University Press, 2018.

Введение в использование ggplot2 для визуализации данных. Рекомендуется в качестве продолжения после книги Уикема и Гроулмунда «Язык R в задачах науки о данных» (упомянута далее в этом списке).

Scott Murray. Interactive Data Visualization for the Web: An Introduction to Designing with D3. 2nd ed. O'Reilly Media, 2017.

Введение в создание интерактивных онлайн-визуализаций с помощью D3 с использованием HTML, CSS, JavaScript и SVG.

Jake VanderPlas. Python Data Science Handbook: Essential Tools for Working with Data. O'Reilly Media, 2016.

Введение в использование языка программирования Python для анализа данных. Содержит обширный материал по визуализации данных с использованием Python Matplotlib и Seaborn.

Хэдли Уикем, Гарретт Гроулмунд. Язык R в задачах науки о данных. Импорт, подготовка, обработка, визуализация и моделирование данных. М.: Вильямс, 2018.

Всестороннее введение в использование языка программирования R для анализа данных. Содержит несколько глав по использованию ggplot2 для визуализации данных.

Тексты по статистике

Вводные тексты по статистике обычно содержат материал по визуализации данных, охватывающий такие темы, как диаграммы рассеяния, гистограммы,

«ящики с усами» и линейные графики. Список подобных текстов довольно велик. Здесь я упомянул лишь некоторые из них, на которые стоит обратить внимание.

David M. Diez, Christopher D. Barr, Mine Çetinkaya-Rundel. OpenIntro Statistics. 3rd ed. OpenIntro, Inc., 2015.

Учебник-введение в статистику с открытым исходным кодом. Книга находится в свободном доступе, то же самое касается файлов LaTeX и листингов R, с помощью которых были созданы сама книга и рисунки к ней.

Susan Holmes, Wolfgang Huber. Modern Statistics for Modern Biology. Cambridge University Press, 2018.

Руководство по статистике, в котором особое внимание уделено вычислительным инструментам, необходимым в современной биологии. Книга находится в свободном доступе, а все примеры из нее снабжены листингами на языке программирования R.

Chester Ismay, Albert Y. Kim. Modern Dive: An Introduction to Statistical and Data Sciences via R. moderndive.com.

Вводный учебник, существующий только в онлайн-формате, посвящен основам статистики и науки о данных. Книга охватывает как теоретические концепции, так и практические подходы с использованием языка программирования R.

Исторические тексты

Книги этого раздела представляют интерес прежде всего по историческим причинам. В свое время это были очень популярные и полезные издания, но сейчас аналогичный материал можно найти и в других книгах, а также в более современном формате.

William S. Cleveland. The Elements of Graphing Data. 2nd ed. Hobart Press, 1994.

Одна из первых книг об информационном дизайне, написанная для статистиков. Книга содержит множество примеров диаграмм рассеяния, линейных графиков, гистограмм и ящичковых диаграмм, а также обсуждает их в контексте анализа данных и статистического моделирования. Данная книга популяризирует точечный график Кливленда.

William S. Cleveland. Visualizing Data. Hobart Press, 1993.

Сопутствующая The Elements of Graphing Data книга того же автора. Данный материал больше сосредоточен на математической стороне вопроса и не затрагивает тему человеческого восприятия.

Edward R. Tufte. Envisioning Information. Graphics Press, 1990.

Эта книга популяризировала концепцию малого множителя.

Edward R. Tufte. The Visual Display of Quantitative Information. 2nd ed. Graphics Press, 2001.

Эта книга, впервые опубликованная в 1983 году, оказала большое влияние на область визуализации данных. Автор ввел в профессиональный (и не только) обиход такие понятия, как визуальный шум, соотношение данных и чернил и спарклайны. В книге также был показан первый график слоупграф (но без названия). При этом небольшая часть рекомендаций, приведенных в этой книге, не прошли проверку временем, например, аргументация в пользу предельно минималистичного дизайна графика.

Книги по смежной тематике

Ниже приведен список изданий, тесно связанных с темами визуализации данных и эффективной коммуникации.

Joshua Schimel. Writing Science. Oxford University Press, 2011.

Данная книга в увлекательной форме учит, как писать тексты на научные и технические темы с помощью историй. Несмотря на то что эта книга не о визуализации данных, она будет очень полезна всем тем, кто пишет технические статьи.

Jonathan Schwabish. Better Presentations. Columbia University Press, 2016.

Краткое и информативное руководство по проведению презентаций. Обязательно к прочтению всем, кто регулярно использует слайды для выступлений или презентаций.

Maureen C. Stone. A Field Guide to Digital Color. A K Peters, 2003.

Подробное руководство о том, как цвета захватываются, обрабатываются и воспроизводятся на компьютере.

Colin Ware. Information Visualization. 3rd ed. Morgan Kaufmann, 2012.

Книга посвящена принципам визуализации, в ней рассматриваются такие темы, как работа зрительной системы человека и восприятие различных графических моделей. Книга охватывает множество различных сценариев визуализации, включая пользовательские интерфейсы и виртуальные миры, но при этом сравнительно меньше внимания уделяется визуализации данных в виде двухмерных иллюстраций.

Технические примечания

Данная книга создана с помощью R Markdown и пакетов `bookdown`, `rmarkdown` и `knitr`. Рисунки сделаны с помощью библиотеки ggplot2 с использованием нескольких дополнительных пакетов, таких как `cowplot`, `geofacet`, `ggforce`, `ggmap`, `ggrepel`, `ggridges`, `hexbin`, `patchwork`, `sf`, `statebins`, `tidybayes` и `treemapify`. Манипуляции с цветом осуществлялись с помощью пакетов `colorpace` и `colorblindr`. Чтобы скомпилировать примеры из всех частей этой книги, используйте актуальные версии указанных пакетов.

Исходный код книги доступен по адресу **github.com/clauswilke/dataviz**. Для книги также требуется вспомогательный пакет R `dviz.supp`, код которого доступен по адресу **github.com/clauswilke/dviz.supp**.

В процессе работы над книгой использовалась следующая среда:

```
## R version 3.5.0 (2018-04-23)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS Sierra 10.12.6
##
## Matrix products: default
## BLAS: /Library/Frameworks/ ... /libRblas.0.dylib
## LAPACK: /Library/Frameworks/ ... /libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/ ... /C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats graphics grDevices utils datasets methods base
##
## other attached packages:
##  [1] nycflights13_1.0.0 gapminder_0.3.0 RColorBrewer_1.1-2
##  [4] gganimate_1.0.0.9000 ungeviz_0.1.0 emmeans_1.3.1
##  [7] mgcv_1.8-24          nlme_3.1-137   broom_0.5.1
## [10] tidybayes_1.0.3      maps_3.3.0     statebins_2.0.0
## [13] sf_0.7-1             maptools_0.9-4 sp_1.3-1
## [16] rgeos_0.3-28         ggspatial_1.0.3 geofacet_0.1.9
## [19] plot3D_1.1.1         magick_1.9     hexbin_1.27.2
## [22] treemapify_2.5.0     gridExtra_2.3  ggmap_2.7.904
## [25] ggthemes_4.0.1       ggridges_0.5.1 ggrepel_0.8.0
## [28] ggforce_0.1.1        patchwork_0.0.1 lubridate_1.7.4
```

```
## [31] forcats_0.3.0       stringr_1.3.1   purrr_0.2.5
## [34] readr_1.1.1         tidyr_0.8.2     tibble_1.4.2
## [37] tidyverse_1.2.1     dviz.supp_0.1.0 dplyr_0.8.0.9000
## [40] colorblindr_0.1.0   ggplot2_3.1.0   colorspace_1.4-0
## [43] cowplot_0.9.99
##
## loaded via a namespace (and not attached):
##  [1] rjson_0.2.20             deldir_0.1-15
##  [3] class_7.3-14             rprojroot_1.3-2
##  [5] estimability_1.3         ggstance_0.3.1
##  [7] rstudioapi_0.7           farver_1.0.0.9999
##  [9] ggfittext_0.6.0          svUnit_0.7-12
## [11] mvtnorm_1.0-8            xml2_1.2.0
## [13] knitr_1.20               polyclip_1.9-1
## [15] jsonlite_1.5             png_0.1-7
## [17] compiler_3.5.0           httr_1.3.1
## [19] backports_1.1.2          assertthat_0.2.0
## [21] Matrix_1.2-14            lazyeval_0.2.1
## [23] cli_1.0.1.9000           tweenr_1.0.1
## [25] prettyunits_1.0.2        htmltools_0.3.6
## [27] tools_3.5.0              misc3d_0.8-4
## [29] coda_0.19-2              gtable_0.2.0
## [31] glue_1.3.0               Rcpp_1.0.0
## [33] cellranger_1.1.0         imguR_1.0.3
## [35] xfun_0.3                 strapgod_0.0.0.9000
## [37] rvest_0.3.2              MASS_7.3-50
## [39] scales_1.0.0             hms_0.4.2
## [41] yaml_2.2.0               stringi_1.2.4
## [43] e1071_1.7-0              spData_0.2.9.4
## [45] RgoogleMaps_1.4.3        rlang_0.3.0.1
## [47] pkgconfig_2.0.2          bitops_1.0-6
## [49] geogrid_0.1.1            evaluate_0.11
## [51] lattice_0.20-35          tidyselect_0.2.5
## [53] plyr_1.8.4               magrittr_1.5
## [55] bookdown_0.7             R6_2.3.0
## [57] generics_0.0.2           DBI_1.0.0
## [59] pillar_1.3.0             haven_1.1.2
## [61] foreign_0.8-71           withr_2.1.2.9000
## [63] units_0.6-1              modelr_0.1.2
## [65] crayon_1.3.4             arrayhelpers_1.0-20160527
## [67] rmarkdown_1.10           progress_1.2.0.9000
## [69] jpeg_0.1-8               rnaturalearth_0.1.0
## [71] grid_3.5.0               readxl_1.1.0
## [73] digest_0.6.18            classInt_0.2-3
## [75] xtable_1.8-3             munsell_0.5.0
## [77] concaveman_1.0.0
```

Примечания

Bateman, S., R. Mandryk, C. Gutwin, A. Genest, D. McDine, and C. Brooks. 2010. "Useful Junk? The Effects of Visual Embellishment on Comprehension and Memorability of Charts." *ACM Conference on Human Factors in Computing Systems*, 2573–2582. *https://dl.acm.org/citation.cfm?doid=1753326.1753716*.

Becker, R. A., W. S. Cleveland, and M.-J. Shyu. 1996. "The Visual Design and Control of Trellis Display." *Journal of Computational and Graphical Statistics* 5: 123–155.

Bergstrom, C. T., and J. West. 2016. "The Principle of Proportional Ink." *http://callingbullshit.org/tools/tools_proportional_ink.html*.

Borgo, R., A. Abdul-Rahman, F. Mohamed, P. W. Grant, I. Reppa, and L. Floridi. 2012. "An Empirical Study on Using Visual Embellishments in Visualization." *IEEE Transactions on Visualization and Computer Graphics* 18: 2759–2768. *https://ieeexplore.ieee.org/document/6327282/*.

Brewer, Cynthia A. 2017. "ColorBrewer 2.0. Color Advice for Cartography." *http://www.ColorBrewer.org*.

Carr, D. B., R. J. Littlefield, W. L. Nicholson, and J. S. Littlefield. 1987. "Scatterplot Matrix Techniques for Large N." *Journal of the American Statistical Association* 82: 424–436.

Clauset, A., C. R. Shalizi, and M. E. J. Newman. 2009. "Power-Law Distributions in Empirical Data." *SIAM Review* 51: 661–703.

Cleveland, R. B., W. S. Cleveland, J. E. McRae, and I. Terpenning. 1990. "STL: A Seasonal-Trend Decomposition Procedure Based on Loess." *Journal of Official Statistics* 6: 3–73.

Cleveland, W. S. 1979. "Robust Locally Weighted Regression and Smoothing Scatterplots." *Journal of the American Statistical Association* 74: 829–836.

–––. 1993. *Visualizing Data*. Summit, New Jersey: Hobart Press.

Dua, D., and E. Karra Taniskidou. 2017. "UCI Machine Learning Repository." University of California, Irvine, School of Information; Computer Sciences. *https://archive.ics.uci.edu/ml*.

Fisher, R. A. 1936. "The Use of Multiple Measurements in Taxonomic Problems." *Annals of Eugenics* 7: 179–188. doi:10.1111/j.1469–1809.1936.tb02137.x.

Haroz, S., R. Kosara, and S. L. Franconeri. 2015. "ISOTYPE Visualization: Working Memory, Performance, and Engagement with Pictographs." *ACM Conference on Human Factors in Computing Systems*, 1191–1200. *https://dl.acm.org/citation.cfm?doid=2702123.2702275*.

–––. 2016. "The Connected Scatterplot for Presenting Paired Time Series." *IEEE Transactions on Visualization and Computer Graphics* 22: 2174–2186. *https://ieeexplore.ieee.org/document/7332976*.

Hullman, J., P. Resnick, and E. Adar. 2015. "Hypothetical Outcome Plots Outperform Error Bars and Violin Plots for Inferences About Reliability of Variable Ordering." *PLOS ONE* 10: e0142444. *https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0142444*.

Kale, A., F. Nguyen, M. Kay, and J. Hullman. 2018. "Hypothetical Outcome Plots Help Untrained Observers Judge Trends in Ambiguous Data." *IEEE Transactions on Visualization and Computer Graphics* 25: 892–905. *https://ieeexplore.ieee.org/document/8440816*.

Kay, M., T. Kola, J. Hullman, and S. Munson. 2016. "When (Ish) Is My Bus? User-Centered Visualizations of Uncertainty in Everyday, Mobile Predictive Systems." CHI Conference on Human Factors in Computing Systems, 5092–5103. *https://dl.acm.org/citation.cfm?doid=2858036.2858558*.

Marcos, M. L., and J. Echave. 2015. "Too Packed to Change: Side-Chain Packing and Site-Specific Substitution Rates in Protein Evolution." *PeerJ* 3: e911.

McDonald, Ian. 2017. "DW-NOMINATE Using ggjoy." *http://rpubs.com/ianrmcdonald/293304*.

Molyneaux, L., S. K. Gilliam, and L. C. Florant. 1947. "Differences in Virginia Death Rates by Color, Sex, Age, and Rural or Urban Residence." *American Sociological Review* 12: 525–535.

Okabe, M., and K. Ito. 2008. "Color Universal Design (CUD): How to Make Figures and Presentations That Are Friendly to Colorblind People."

Paff, M. L., B. R. Jack, B. L. Smith, J. J. Bull, and C. O. Wilke. 2018. "Combinatorial Approaches to Viral Attenuation." bioRxiv, 29918. *https://www.biorxiv.org/content/10.1101/299180v1*.

Schimel, J. 2011. *Writing Science: How to Write Papers That Get Cited and Proposals That Get Funded*. Oxford: Oxford University Press.

Sidiropoulos, N., S. H. Sohi, T. L. Pedersen, B. T. Porse, O. Winther, N. Rapin, and F. O. Bagger. 2018. "SinaPlot: An Enhanced Chart for Simple and Truthful Representation of Single Observations over Multiple Classes." *Journal of*

Computational and Graphical Statistics 27: 673–676. *https://www.tandfonline.com/doi/abs/10.1080/10618600.2017.1366914?journalCode=ucgs20*.

Stone, M., D. Albers Szafir, and V. Setlur. 2014. "An Engineering Model for Color Difference as a Function of Size." 22nd Color and Imaging Conference, 253–258.

Telford, R. D., and R. B. Cunningham. 1991. "Sex, Sport, and Body-Size Dependency of Hematology in Highly Trained Athletes." *Medicine and Science in Sports and Exercise* 23: 788–794.

The Economist online. 2011. "Corrosive Corruption." *https://www.economist.com/graphic-detail/2011/12/02/corrosive-corruption*.

Tufte, E. R. 1990. *Envisioning Information*. Cheshire, Connecticut: Graphics Press.

–––. 2001. *The Visual Display of Quantitative Information*. 2nd ed. Cheshire, Connecticut: Graphics Press.

Wehrwein, A. 2017. "It Brings Me ggjoy." *https://austinwehrwein.com/data-visualization/it-brings-me-ggjoy/*.

Wickham, H. 2016. *ggplot2: Elegant Graphics for Data Analysis*. 2nd ed. New York: Springer.

Wikipedia, User:Schutz. 2007. "File:Piecharts.svg." *https://en.wikipedia.org/wiki/File:Piecharts.svg*.

Yates, F. 1935. "Complex Experiments." *Supplement to the Journal of the Royal Statistical Society* 2: 181–247. *https://www.jstor.org/stable/2983638?origin=crossref&seq=1#page_scan_tab_contents*.

Предметный указатель

Г

Д

Е

П

Р

С

Об авторе

Клаус О. Уилке — профессор интегративной биологии Техасского университета в Остине. Он имеет докторскую степень в области теоретической физики Рурского университета в Бохуме, Германия. Клаус является автором или соавтором более 170 научных публикаций, охватывающих темы вычислительной биологии, математического моделирования, биоинформатики, эволюционной биологии, биохимии белков, вирусологии и статистики. Он является автором нескольких популярных пакетов R, используемых для визуализации данных, таких как `cowplot` и `ggridges`, а также участвует в разработке пакета ggplot2.

Об изображении на обложке

Животное на обложке книги — это попугай розелла (*Platycercus icterotis*), небольшой вид попугаев на юго-западе Австралии. Название *icterotis* происходит от древнегреческого слова, которое означает «желтое ухо», и связано с желтыми пятнами на щеках птицы. Розелла действительно очень красочная: голова и шея птицы имеют красный цвет, спинка — полосатая зеленая, черная и красная, на крыльях синие перья, а хвост — сине-зеленый. Длина птиц этого вида в среднем составляет около 25 см.

Этот попугай водится в лесах, сельскохозяйственных угодьях и парках. Обычно он питается травой, семенами и фруктами, но в период размножения ему требуется больше белка, поэтому в это время он также поедает личинок насекомых. Птицы кормятся на земле, собираясь группами по 20 особей в местах, где много пищи. Спаривающиеся пары гнездятся в дуплах деревьев (часто предпочитают эвкалипты) и откладывают от двух до семи яиц в каждом выводке.

На западе розеллы часто живут в вольерах, период жизни в неволе составляет более 15 лет.

Иллюстрация на обложке сделана Карен Монтгомери и основана на черно-белой гравюре из книги Шоу «Зоология».

Научно-популярное издание

БИБЛИОТЕКА ЦИФРОВОЙ ТРАНСФОРМАЦИИ

Уилке Клаус

ОСНОВЫ ВИЗУАЛИЗАЦИИ ДАННЫХ
ПОСОБИЕ ПО ЭФФЕКТИВНОЙ И УБЕДИТЕЛЬНОЙ ПОДАЧЕ ИНФОРМАЦИИ

Главный редактор *Р. Фасхутдинов*
Руководитель направления *В. Обручев*
Ответственный редактор *Л. Салихова*
Продюсер *Н. Витько*
Научный редактор *А. Бояршинов*
Литературный редактор *Ю. Медведева*
Младший редактор *П. Смирнов*
Художественный редактор *Е. Пуговкина*
Компьютерная верстка *Э. Брегис*
Корректоры *Н. Болотина, Л. Крымова, Ю. Киреева*

Страна происхождения: Российская Федерация
Шығарылған елі: Ресей Федерациясы

ООО «Издательство «Эксмо»
123308, Россия, г. Москва, ул. Зорге, д. 1, стр. 1, эт. 20, каб. 2013. Тел.: 8 (495) 411-68-86.
Home page: www.eksmo.ru E-mail: info@eksmo.ru
Өндіруші: «Издательство «Эксмо» ЖШҚ
123308, Ресей, Мәскеу қаласы, Зорге көшесі, 1-үй, 1-құрылыс, 20 қабат, 2013-каб.
Тел.: 8 (495) 411-68-86. Home page: www.eksmo.ru E-mail: info@eksmo.ru.
Тауар белгісі: «Эксмо»
Интернет-магазин : www.book24.ru
Интернет-магазин : www.book24.kz
Интернет-дүкен : www.book24.kz

Импортёр в Республику Казахстан ТОО «РДЦ-Алматы».
Қазақстан Республикасына импорттаушы «РДЦ-Алматы» ЖШС.

Дистрибьютор и представитель по приему претензий на продукцию в Республике Казахстан: ТОО «РДЦ-Алматы»
Дистрибьютор және Қазақстан Республикасында өнімге шағымдар қабылдау жөніндегі өкіл: «РДЦ-Алматы» ЖШС.
Алматы қ., Домбровский көш., 3 «а», литер Б, офис 1.
Тел.: 8 (727) 251-59-90/91/92. E-mail: RDC-Almaty@eksmo.kz

Сведения о подтверждении соответствия издания согласно законодательству РФ о техническом регулировании можно получить на сайте Издательства «Эксмо»: www.eksmo.ru/certification

Техникалық реттеу туралы РФ заңнамасына сай басылымның сәйкестігін растау туралы мәліметтерді мына адрес бойынша алуға болады: http://eksmo.ru/certification/

Произведено в Российской Федерации
Ресей Федерациясында өндірілген

Сертификаттауға жатпайды

Дата изготовления / Подписано в печать 19.12.2023. Формат 70x100$^{1}/_{16}$.
Гарнитура «Newton». Печать офсетная. Усл. печ. л. 28,52.
Тираж экз. Заказ

12+

Москва. ООО «Торговый Дом «Эксмо»
Адрес: 123308, г. Москва, ул. Зорге, д.1, строение 1.
Телефон: +7 (495) 411-50-74. **E-mail:** reception@eksmo-sale.ru

По вопросам приобретения книг «Эксмо» зарубежными оптовыми покупателями обращаться в отдел зарубежных продаж ТД «Эксмо»
E-mail: **international@eksmo-sale.ru**

International Sales: International wholesale customers should contact Foreign Sales Department of Trading House «Eksmo» for their orders.
international@eksmo-sale.ru

По вопросам заказа книг корпоративным клиентам, в том числе в специальном оформлении, обращаться по тел.: +7 (495) 411-68-59, доб. 2151.
E-mail: **borodkin.da@eksmo.ru**

Оптовая торговля бумажно-беловыми и канцелярскими товарами для школы и офиса «Канц-Эксмо»:
Компания «Канц-Эксмо»: 142702, Московская обл., Ленинский р-н, г. Видное-2, Белокаменное ш., д. 1, а/я 5. Тел./факс: +7 (495) 745-28-87 (многоканальный).
e-mail: **kanc@eksmo-sale.ru**, сайт: www.**kanc-eksmo.ru**

Филиал «Торгового Дома «Эксмо» в Нижнем Новгороде
Адрес: 603094, г. Нижний Новгород, улица Карпинского, д. 29, бизнес-парк «Грин Плаза»
Телефон: +7 (831) 216-15-91 (92, 93, 94). **E-mail**: reception@eksmonn.ru

Филиал ООО «Издательство «Эксмо» в г. Санкт-Петербурге
Адрес: 192029, г. Санкт-Петербург, пр. Обуховской обороны, д. 84, лит. «Е»
Телефон: +7 (812) 365-46-03 / 04. **E-mail**: server@szko.ru

Филиал ООО «Издательство «Эксмо» в г. Екатеринбурге
Адрес: 620024, г. Екатеринбург, ул. Новинская, д. 2щ
Телефон: +7 (343) 272-72-01 (02/03/04/05/06/08)

Филиал ООО «Издательство «Эксмо» в г. Самаре
Адрес: 443052, г. Самара, пр-т Кирова, д. 75/1, лит. «Е»
Телефон: +7 (846) 207-55-50. **E-mail**: RDC-samara@mail.ru

Филиал ООО «Издательство «Эксмо» в г. Ростове-на-Дону
Адрес: 344023, г. Ростов-на-Дону, ул. Страны Советов, 44А
Телефон: +7(863) 303-62-10. **E-mail**: info@rnd.eksmo.ru

Филиал ООО «Издательство «Эксмо» в г. Новосибирске
Адрес: 630015, г. Новосибирск, Комбинатский пер., д. 3
Телефон: +7(383) 289-91-42. E-mail: eksmo-nsk@yandex.ru

Обособленное подразделение в г. Хабаровске
Фактический адрес: 680000, г. Хабаровск, ул. Фрунзе, 22, оф. 703
Почтовый адрес: 680020, г. Хабаровск, А/Я 1006
Телефон: (4212) 910-120, 910-211. **E-mail**: eksmo-khv@mail.ru

Республика Беларусь: ООО «ЭКСМО АСТ Си энд Си»
Центр оптово-розничных продаж Cash&Carry в г. Минске
Адрес: 220014, Республика Беларусь, г. Минск, проспект Жукова, 44, пом. 1-17, ТЦ «Outleto»
Телефон: +375 17 251-40-23; +375 44 581-81-92
Режим работы: с 10.00 до 22.00. **E-mail:** exmoast@yandex.by

Казахстан: «РДЦ Алматы»
Адрес: 050039, г. Алматы, ул. Домбровского, 3А
Телефон: +7 (727) 251-58-12, 251-59-90 (91,92,99). E-mail: RDC-Almaty@eksmo.kz

Полный ассортимент продукции ООО «Издательство «Эксмо» можно приобрести в книжных магазинах «Читай-город» и заказать в интернет-магазине: www.chitai-gorod.ru.
Телефон единой справочной службы: 8 (800) 444-8-444. Звонок по России бесплатный.

Интернет-магазин ООО «Издательство «Эксмо»
www.eksmo.ru
Розничная продажа книг с доставкой по всему миру.
Тел.: +7 (495) 745-89-14. E-mail: **imarket@eksmo-sale.ru**

БОМБОРА ИЗДАТЕЛЬСТВО

ISBN 978-5-04-106457-0

БОМБОРА – лидер на рынке полезных и вдохновляющих книг. Мы любим книги и создаем их, чтобы вы могли творить, открывать мир, пробовать новое, расти. Быть счастливыми. Быть на волне.

bombora.ru bomborabooks bombora

O'REILLY®

Основы визуализации данных

з визуализации информации сейчас не ходится ни один бизнес. Итоги продаж, сячные отчеты, презентации новых ектов – все это примеры того, что важ- и нужно правильно визуализировать.

м нагляднее графики, тем весомее мотрятся» ваши аргументы. Изображе- я должны быть ясными, привлекатель- ми и убедительными.

помощью этого руководства основам визуализации анных вы сможете:

- обучиться искусству создания графиков с нуля
- освоить различные форматы визуализации для своих презентаций
- собрать собственный инструментарий для инфографики
- создавать дизайнерские концепции и креативные решения для достижения своих целей
- иллюстрировать свою позицию наглядными примерами
- уделять внимание мелким деталям, зачастую упускаемым другими людьми

Материал книги выстроен в ческой последовательности, о большинство глав можно читат самостоятельный текст – вам не тельно штудировать книгу от кор корки. Не стесняйтесь переходи части к части, чтобы выбрать наи интересный для вас раздел ил который посвящен тому типу дан над которым вы сейчас работает

Жаль, что эта книга не выш 20 лет назад, когда я впер начал заниматься визуализац данных. В ней прекрас систематизированы и подо практические знания автора.

Андрей Бояршин
Director of Production, General Arc

Клаус Уилке – профессор интегратив биологии Техасского университета в О автор более 170 научных публикаций охватывающих темы вычислительной биологии, математического моделиро биоинформатики, эволюционной биол биохимии белков, вирусологии и стат Он также выступает создателем неско популярных пакетов R, используемых визуализации данных.

ISBN 978-5-04-106457-0