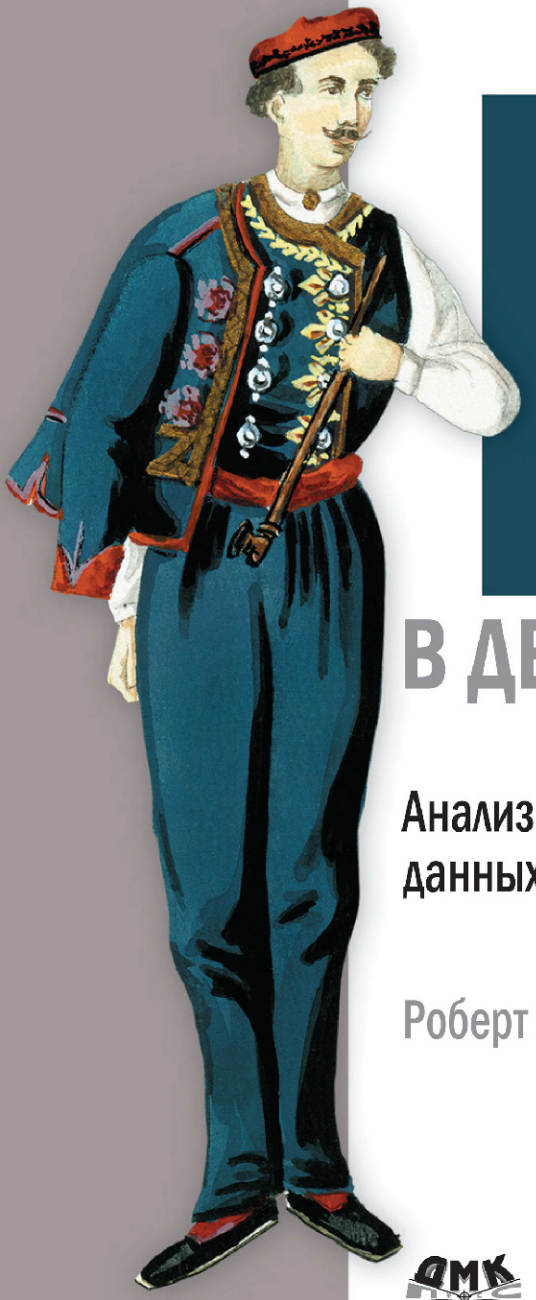

R в действии

Анализ и визуализация данных на языке R

Роберт И.Кабаков

ДМК ПРЕСС ИЗДАТЕЛЬСТВО

MANNING

Роберт И. Кабаков

R в действии

Анализ и визуализация данных в программе R

R in Action

Data analysis and graphics withR

ROBERT I. KABACOFF

MANNING
SHELTER ISLAND

R в действии

Анализ и визуализация данных в программе R

РОБЕРТ И. КАБАКОВ

2-е издание, электронное

ДМК ПРЕСС
издательство

Москва, 2023

УДК 311:004.9R
ББК 60.6с515
К12

Кабаков, Роберт И.

К12 R в действии. Анализ и визуализация данных на языке R / Р. И. Кабаков ; пер. с англ. П. А. Волковой. — 2-е изд., эл. — 1 файл pdf : 590 с. — Москва : ДМК Пресс, 2023. — Систем. требования: Adobe Reader XI либо Adobe Digital Editions 4.5 ; экран 10". — Текст : электронный.

ISBN 978-5-89818-347-9

R — это мощный язык для статистических вычислений и графики, который может справиться поистине с любой задачей в области обработки данных. Он работает во всех важных операционных системах и поддерживает тысячи специализированных модулей и утилит. Все это делает R замечательным средством для извлечения полезной информации из гор сырых данных.

«R в действии» — это руководство по обучению этому языку с особым вниманием к практическим задачам. В данной книге представлены полезные примеры статистической обработки данных и описаны изящные методы работы с запутанными и неполными данными, а также с данными, распределение которых отлично от нормального и с которыми трудно справиться обычными методами. Статистический анализ — это только одна сторона дела. Вы также овладеете обширными графическими возможностями для визуального исследования и представления данных.

УДК 311:004.9R
ББК 60.6с515

Электронное издание на основе печатного издания: R в действии. Анализ и визуализация данных на языке R / Р. И. Кабаков ; пер. с англ. П. А. Волковой. — Москва : ДМК Пресс, 2014. — 588 с. — ISBN 978-5-97060-077-1. — Текст : непосредственный.

ISBN 978-5-89818-347-9

ОТ ПЕРЕВОДЧИКА

По моему глубокому убеждению, на сегодняшний день это лучшая книга, посвященная обработке данных в статистической среде R, для неспециалистов. Я рада, что теперь она стала доступной русскоязычным читателям. Надеюсь, мой перевод не сильно испортил эту книгу. По крайней мере, в некоторых местах она точно стала лучше, потому что я исправила довольно многочисленные и не всегда безобидные опечатки, обнаруженные в исходном издании мною и другими читателями, которые оставили свои отзывы на форуме издательства.

Я благодарю Александра Лободу, который рассказал мне о существовании данной книги, Дмитрия Мовчана, с энтузиазмом воспринявшего мое предложение опубликовать ее перевод, Бориса Демешева за консультации по переводу некоторых статистических терминов и Алексея Шипунова за техническую поддержку. Я особенно признательна Сергею Петрову и Сергею Мастицкому за внимательное прочтение рукописи перевода и конструктивные замечания.

ОГЛАВЛЕНИЕ

ПРЕДИСЛОВИЕ

Что толку в книжке, если в ней нет ни картинок, ни разговоров?

Алиса. «Алиса в Стране чудес»[1]

Оно чудесно и наделено сокровищами, способными удовлетворить всех от мала до велика, но не предназначено для робких духом.

Кью. Сериал «Звездный путь: следующее поколение»

Когда я начал писать эту книгу, я потратил достаточно много времени в поисках хорошего эпиграфа. В итоге я остановился на этих двух. R – это потрясающе гибкие приложение и язык для исследования, визуализации и понимания данных. Я выбрал цитату из «Алисы в Стране чудес», чтобы передать суть современного статистического анализа – интерактивного процесса, состоящего из исследования, визуализации и интерпретации.

Вторая цитата отражает широко распространенное мнение о том, что работе в R сложно научиться. Я надеюсь показать вам, что это не обязательно должно быть так. R – мощная программа с таким большим числом доступных аналитических и графических функций (по последним подсчетам их более 50 000), что она может в одинаковой степени навести ужас и на новичков, и на опытных пользователей. Однако в этом мнимом безумии есть поэзия и логика. Вооружившись руководствами и инструкциями, вы сможете ориентироваться в огромном разнообразии возможностей, выбрав те инструменты, которые нужны для того, чтобы уверенно, эффективно и элегантно выполнить вашу задачу.

Я впервые познакомился с R несколько лет назад, когда хотел получить новую должность консультанта по статистике. Предполагаемый работодатель перед интервью спросил меня, владею ли я R. Следуя обычным советам специалистов по подбору персонала, я немедленно сказал «да» и стал учиться работать в этой программе. Я был опытным статистиком и исследователем с 25 годами опыта

1 Перевод Н. Демуровой.

программирования в SAS и SPSS, свободно владевшим несколькими языками программирования. Чего же тут может быть сложного? Знаменитые последние слова.

По мере того как я пытался выучить язык программирования (как можно быстрее, ведь день собеседования приближался с угрожающей быстротой), я находил или тома, посвященные глубинной структуре языка, или многочисленные трактаты об отдельных продвинутых статистических методах, написанных специалистами в данной области для своих коллег. Встроенная помощь была написана очень лаконично и служила скорее справочником, чем учебным пособием. Каждый раз, когда мне казалось, что я освоил общую логику и возможности R, находилось что-то новое, заставлявшее почувствовать себя невежественным и ничтожным.

При освоении R я подошел к процессу с точки зрения исследователя, которому нужно обрабатывать данные. Я пытался понять, что нужно сделать, чтобы успешно обработать, проанализировать и понять данные, включая:

- доступ к данным (получение данных из разных источников);
- редактирование данных (замена или удаление пропущенных значений, преобразование признаков в более удобный вид);
- аннотирование данных (чтобы помнить, что представляет собой каждый их фрагмент);
- получение общих сведений о данных (вычисление описательных статистик для того, чтобы охарактеризовать данные);
- визуализация данных (поскольку картинка на самом деле сто́ит тысячи слов);
- моделирование данных (нахождение зависимостей и тестирование гипотез);
- оформление результатов (подготовка таблиц и диаграмм достаточного для публикации качества).

Затем я постарался понять, как я могу использовать R, чтобы выполнить каждую из этих задач. Поскольку я лучше всего учусь, обучая других, со временем я создал сайт (www.statmethods.net), на котором рассказал все, что я узнал.

Затем, около года назад, Марьян Бейс (Marjan Bace), издатель, позвонила и спросила, не хочу ли я написать книгу про R. К этому времени я уже написал 50 статей в научных журналах, четыре технических руководства, многочисленные главы в книгах и целую книгу по методологии исследования, так чего же тут могло быть сложного? Рискую повториться – знаменитые последние слова.

Книгу, которую вы держите в руках, я мечтал иметь много лет назад. Я постарался написать для вас путеводитель по R, который позволит быстро овладеть всей мощью этой замечательной программы с открытым кодом без разочарования и раздражения, которые пришлось испытать мне. Надеюсь, вам понравится.

P.S. Мне предложили ту должность, но я отказался. Однако знакомство с R развернуло мою карьеру в совершенно неожиданном направлении. Жизнь может быть забавной штукой.

БЛАГОДАРНОСТИ

Многие люди приложили значительные усилия, чтобы сделать эту книгу лучше:

- в первую очередь это Марьян Бейс (Marjan Bace), глава издательства Маннинг (Manning), которая предложила мне написать эту книгу;
- Себастьян Стирлинг (Sebastian Stirling), редактор-консультант по аудитории (development editor), который провел многие часы в телефонных беседах со мной, помогая выстроить материал, прояснить основные идеи и в целом сделать текст более интересным. Он также помог мне на многих этапах подготовки книги к изданию;
- Карен Тегмейер (Karen Tegtmeyer), редактор-рецензент (review editor), которая помогла найти рецензентов и координировала процесс рецензирования;
- Мэри Пиргис (Mary Piergies), которая помогала следить за процессом подготовки книги к печати, и ее команду: Лиз Велч (Liz Welch), Сьюзан Харкинс (Susan Harkins) и Рахель Шредер (Rachel Schroeder);
- Пабло Доминик Васелли (Pablo Dominguez Vaselli), корректор, который помог обнаружить ошибки и свежим опытным взглядом проверил программный код;
- рецензенты, которые потратили много времени на внимательное чтение текста, находили опечатки и делали ценные замечания: Крис Вилльямс (Chris Williams), Чарльз Мальпас (Charles Malpas), Анжела Стейплз (Angela Staples), Даниэль Рейс Перейра (Daniel Reis Pereira), Д.Х. Ван Райн (D. H. van Rijn), Кристиан Маркуардт (Christian Marquardt), Амос Фоларин (Amos Folarin), Стюарт Джеффрис (Stuart Jefferys), Дрор Берел (Dror Berel), Патрик Брин (Patrick Breen), Элизабет Островски (Elizabeth Ostrowski), Атеф Оуни (Atef Ouni), Карл Феноллоза (Carles Fenollosa), Рикардо Питробон (Ricardo Pietrobon), Самуэль МакКвилин (Samuel McQuillin),

Ландон Кокс (Landon Cox), Августин Циглер (Austin Ziegler), Рик Вагнер (Rick Wagner), Райн Кокс (Ryan Cox), Сумит Пал (Sumit Pal), Филипп К. Джанер (Philipp K. Janert), Дипак Вохра (Deepak Vohra) и Софи Мормеде (Sophie Mormede);
- многие участники программы раннего доступа издательства Маннинг (Manning Early Access Program, MEAP).

Каждый из перечисленных людей сделал эту книгу лучше и полнее.

Я также хотел бы поблагодарить многочисленных разработчиков, которые и сделали R такой мощной платформой для анализа данных. Этот список включает не только основную команду разработчиков, но и многочисленных самоотверженных людей, которые создали и поддерживают дополнительные пакеты, значительно расширяющие возможности R. В приложении F перечислены авторы всех пакетов, упомянутых в этой книге. Отдельно я хотел бы упомянуть Джона Фокса (John Fox), Хадли Викхама (Hadley Wickham), Франка Е. Харрела-мл. (Frank E. Harrell), Дипаяна Саркара (Deepayan Sarkar) и Вильяма Ревилла (William Revelle), работами которых я восхищаюсь. Я старался как следует отразить их вклад, а ответственность за все ошибки и искажения, непреднамеренно допущенные в этой книге, лежит исключительно на мне.

На самом деле мне следовало бы начать эту книгу с благодарности моей жене и другу Кэрол Линн (Carol Lynn). Хотя она особенно не интересуется статистикой или программированием, она неоднократно прочитала каждую главу и сделала бесчисленные исправления и предложения. Никаким другим способом нельзя выразить свою любовь к другому человеку лучше, чем прочесть ради него текст по многомерной статистике. Столь же важно, что она переживала многие вечера и выходные, которые я проводил в работе над этой книгой, с тактом, поддержкой и сочувствием. И за что это мне так повезло?

Есть еще два человека, которых я хочу поблагодарить. Один из них – мой отец, любовь которого к науке вдохновляла меня и помогла понять ценность полученных данных. Другой человек – Гари К. Бургер (Gary K. Burger), мой руководитель в магистратуре. Гари заинтересовал меня статистикой и преподаванием, в то время как я собирался стать врачом. Это все он.

ОБ ЭТОЙ КНИГЕ

Если вы выбрали эту книгу, скорее всего, у вас есть какие-то данные, которые нужно собрать в единое целое, преобразовать, исследовать, смоделировать, визуализировать или представить коллегам. Если это так, то R создан для вас! R стал всемирно известным языком программирования для статистического анализа, предсказаний и визуализации данных. В этой программе реализовано множество методов анализа данных, от самых простых до самых сложных и современных.

Эта программа с открытым кодом работает под разными операционными системами, включая Windows, Mac OS X и Linux. Она развивается постоянно, новые методы появляются ежедневно. Кроме того, R поддерживается большим и разнородным сообществом ученых и программистов, которые охотно помогут новичку советами.

Хотя программа R, возможно, больше известна за способность создавать красивые и сложные диаграммы, она может справиться с любой статистической задачей. Базовая версия содержит сотни функций для статистического анализа, управления данными и построения диаграмм. Однако некоторые особенно мощные методы реализованы в дополнительных пакетах, созданных независимыми авторами.

Эта широта возможностей имеет свою цену. Для новичков бывает сложно понять, что такое R и как в ней работать. Даже самые опытные пользователи R с удивлением обнаруживают какие-то возможности, о которых они не подозревали.

«R в действии» представляет собой руководство-путеводитель по R, позволяя в общих чертах ознакомиться с самой программой и ее возможностями. В книге описаны наиболее полезные функции базовой версии и более 90 наиболее часто используемых дополнительных пакетов. На всем протяжении книги акцент делается на практическое применение – на то, чтобы вы, руководствуясь прочитанным, могли проанализировать ваши данные и изложить результаты коллегам. По окончании чтения этой книги вы будете иметь хорошее представление о том, как R работает и где можно получить дополнительную информацию. Вы научитесь применять разнообразные методы для визуали-

зации данных и обретете достаточно умений, чтобы справиться как с простыми, так и со сложными задачами анализа данных.

Кому следует прочесть эту книгу

Книга «R в действии» предназначена для любого, кто имеет дело с данными. Опыт в статистическом программировании не требуется. Хотя эта книга доступна и новичкам, в ней содержится достаточно нового и полезного материала, чтобы удовлетворить запросы даже опытных специалистов по R.

Пользователи, не владеющие познаниями в области статистики, которые хотят использовать R для управления данными, их обобщения и представления в графическом виде, смогут легко понять главы 1–6, 11 и 16. Главы 7 и 10 подразумевают, что вы прослушали вводный курс статистики, а главы 8, 9 и 12–15 потребуют более глубоких познаний в этой области. Однако я старался написать каждую главу так, чтобы в ней было что-то интересное и полезное и для новичков, и для опытных статистиков.

Структура книги

Эта книга создана как путеводитель по программе R, с акцентом на методы, которые можно сразу применить для управления данными, их визуализации и осмысления. Книга состоит из 16 глав, сгруппированных в четыре части: «Начало работы», «Базовые методы», «Методы средней сложности» и «Методы повышенной сложности». Дополнительные темы рассмотрены в восьми приложениях.

Глава 1 начинается с обзора программы в целом и характеристик, которые делают ее столь полезной для обработки данных. В главе рассказано, как установить программу и как расширить ее возможности путем установки доступных в Сети дополнительных пакетов. Оставшаяся часть главы посвящена описанию интерфейса программы и рассказу о том, как запускать ее в интерактивном и пакетном режимах.

В главе 2 описаны многие методы импорта данных в программу. Первая половина главы посвящена характеристике типов данных в R и тому, как вводить данные с клавиатуры. Во второй половине главы обсуждаются способы импорта данных из текстовых файлов, вебстраниц, электронных таблиц, других статистических программ и баз данных.

Многие пользователи изначально выбрают R потому, что они хотят создавать диаграммы, так что мы сразу переходим к этой теме в

главе 3. Вам не понадобится долго ждать. Мы обсуждаем, как создавать диаграммы, изменять их и сохранять в разных форматах.

Глава 4 посвящена основам управления данных, включая сортировку, объединение и разбиение наборов данных, а также преобразование, перекодировку и удаление переменных.

Глава 5 основана на главе 4 и содержит описание функций (математических, статистических, текстовых) и управляющих конструкций (циклы, выполнение при условии) для управления данными. Затем мы обсуждаем, как написать вашу собственную функцию в R и как сгруппировать данные различными способами.

В главе 6 рассказано, как создавать наиболее распространенные одномерные диаграммы, такие как столбчатая и круговая диаграммы, диаграмма распределения плотности, диаграмма размахов («ящик с усами») и точечная диаграмма. Все эти диаграммы полезны для изучения характера распределения значений одной переменной.

Глава 7 начинается с описания того, как находить общие характеристики данных, включая использование описательных статистик и сводных таблиц. Затем мы рассматриваем основные способы изучения взаимосвязи между двумя переменными, включая корреляцию, тест Стьюдента, тест хи-квадрат и непараметрические методы.

Глава 8 посвящена применению регрессионных методов для моделирования взаимосвязи между числовой переменной-откликом (outcome variable) и набором из одной или нескольких независимых переменных (predictor variables). Подробно рассмотрены методы подгонки этих моделей, оценки их адекватности и интерпретации их значений.

В главе 9 рассмотрены основные типы планов экспериментов при дисперсионном анализе и его разновидностях. В этой ситуации нас обычно интересует, как комбинации разных типов воздействия или разных условий влияют на числовую переменную-отклик. Также описаны методы оценки адекватности анализа и визуализации результатов.

Детальное описание анализа мощности статистических тестов – предмет главы 10. Она начинается с обсуждения проблемы проверки гипотез; далее описано, как определить объем выборки, необходимый для выявления эффекта заданной величины при заданном уровне достоверности. Это поможет вам повысить вероятность достижения желаемого результата при планировании экспериментов.

Глава 11 – это продолжение главы 5. В ней рассказано, как создать диаграммы для визуализации связей между двумя и более переменными. Обсуждаются разные типы двух- и трехмерных диаграмм

рассеяния, матриц диаграмм рассеяния, графиков, коррелограмм и мозаичных диаграмм.

В главе 12 представлены аналитические методы, которые хорошо работают, когда данные происходят из неизвестных или смешанных типов распределения, когда размеры выборок малы, когда выбросы представляют собой проблему или когда разработка статистического теста на основании наблюдаемого распределения слишком сложна. Это метод повторной выборки (resampling) и бутстреп-анализ (bootstrapping) – подходы, требующие большого объема вычислений и легко реализуемые в R.

Глава 13 описывает, как применять регрессионный анализ, рассмотренный в главе 8, к данным с распределением, отличным от нормального. Глава начинается с описания обобщенных линейных моделей. Затем более подробно рассматриваются случаи, когда нужно предсказать переменную-отклик, представленную либо категориальными (логистическая регрессия), либо счетными данными (пуассоновская регрессия).

Одна из сложностей, связанных с многомерными данными, – это проблема снижения их размерности. В главе 14 описаны методы, с помощью которых большое число коррелирующих друг с другом переменных преобразуется в меньший набор независимых переменных (анализ главных компонент), а также методы обнаружения скрытой структуры в имеющемся наборе переменных (факторный анализ). Детально разобраны многочисленные этапы этих типов анализа.

В соответствии с нашим намерением описать актуальные методы анализа данных глава 15 посвящена современным подходам к решению распространенной проблемы пропущенных значений в данных. В R реализованы разнообразные изящные подходы к анализу неполных в силу разных причин данных. Здесь описаны лучшие из этих методов, вместе с разъяснениями, когда сто́ит применять каждый из них, а каких лучше избегать.

Глава 16 завершает обсуждение диаграмм рассмотрением некоторых наиболее сложных и полезных методов визуализации данных. Рассмотрена визуализация очень сложных данных с использованием панельной (или категоризированной) графики, даны основные сведения о новом пакете ggplot2, также кратко описаны способы работы с диаграммами в режиме реального времени.

В послесловии перечислены многие из лучших сайтов, которые следует посетить, чтобы научиться работать в R, влиться в сообщество пользователей R, получить ответы на возникшие вопросы и отсле-

живать изменения в этом стремительно развивающемся программном продукте.

И последнее, но не менее важное: восемь приложений (от A до H) содержат дополнительные сведения по таким полезным темам, как пользовательский интерфейс, настройка и обновление программы, экспорт данных, получение результатов высокого полиграфического качества, использование R для матричной алгебры (по образцу MATLAB) и работа с большими объемами данных.

Примеры

Для того чтобы сделать книгу настолько широко применимой, насколько возможно, я выбрал примеры из разных областей знаний, включая психологию, социологию, медицину, биологию, бизнес и технические науки. Ни один из примеров не требует специальных знаний в соответствующей области.

Наборы данных, используемые в этих примерах, были выбраны потому, что они позволяют формулировать интересные вопросы и имеют небольшой размер. Это позволяет сосредоточиться на рассматриваемом методе и быстро понять происходящее. Когда учишься новым методам, меньше – значит лучше.

Наборы данных либо поставляются с базовой версией R, либо доступны в составе дополнительных пакетов, которые можно скачать из Интернета. Программный код для каждого примера размещен на сайте http://www.manning.com/RinAction. Для получения максимальной отдачи от этой книги я рекомендую выполнять примеры по ходу их прочтения.

В заключение нужно вспомнить известную сентенцию, которая гласит, что если спросить двух статистиков, как анализировать определенный набор данных, получишь три разных ответа. Можно понимать этот афоризм по-разному, ведь каждый ответ приблизит вас к пониманию данных. Я не утверждаю, что предлагаемый мной тот или иной способ анализа данных – лучший или единственный путь к решению конкретной задачи. Я предлагаю вам применить разные подходы к данным, используя знания, приобретенные во время чтения книги, и посмотреть, что вы сможете узнать. R – интерактивная программа, и лучший способ чему-то научиться в ней – это экспериментировать.

Принятые обозначения

В книге использованы следующие типографские обозначения:

- `моноширинный шрифт` использован для программного кода, который нужно вводить именно так, как указано в книге;

- моноширинный шрифт также использован внутри основного текста для обозначения фрагментов кода или ранее упомянутых объектов;
- *курсив* внутри программного кода – это указатель места заполнения. Его следует заменять подходящим текстом или значениями, соответствующими вашей задаче. Например, *путь_к_моему_файлу* должен быть заменен указанием пути к реальному файлу на вашем компьютере;
- R – это интерактивный язык, который информирует пользователя о готовности принять команду приглашением (> по умолчанию). Многие фрагменты программного кода в книге скопированы из интерактивных сессий. Если вы видите строки кода, которые начинаются с >, не набирайте этот символ приглашения к вводу команды;
- пояснения к программному коду приведены в виде внутритекстовых комментариев. В дополнение к этому некоторые пояснения обозначены нумерованными кружками, такими как ❶, которые отсылают к объяснению ниже по тексту;
- для того чтобы сэкономить место или сделать текст более понятным, мы иногда добавляли в вывод результатов интерактивных сессий дополнительные пробелы или удаляли текст, который напрямую не относился к обсуждаемой теме.

Об авторе

Доктор наук Роберт Кабаков – вице-президент по исследовательской работе (Vice President of Research) в Группе исследований менеджмента (Management Research Group – MRG), международной фирме, специализирующейся на организационном развитии и консалтинге. У него за спиной более 20 лет опыта в сфере исследовательских и статистических консультаций в областях заботы о здоровье, финансовых операций, производства, бихевиоризма, управления и академической науки. Прежде чем присоединиться к MRG, Р. Кабаков был профессором психологии в юго-восточном университете Нова (Nova Southeastern University) во Флориде, где он преподавал количественные методы и статистическое программирование в магистратуре. В последние два года он поддерживает сайт Quick-R – учебное пособие по R.

ОБ ИЛЛЮСТРАЦИИ НА ОБЛОЖКЕ

На обложке книги «R в действии» изображен «Мужчина из Задара». Эта иллюстрация позаимствована из альбома, посвященного хорватским национальным костюмам середины XIX века. Альбом составлен Николой Арсеновичем (Nikola Arsenovic) и опубликован в 2003 году музеем этнографии, расположенным в хорватском городе Сплит. Иллюстрация получена благодаря любезной помощи библиотекаря музея. Этот музей расположен в римской части средневекового центра города: недалеко от развалин дворца императора Диоклетиана, датированных примерно 304 годом нашей эры. Альбом состоит из красочных изображений людей из разных регионов Хорватии с описанием их костюмов и образа жизни.

Задар – это древнеримский город на севере далматского побережья Хорватии. Ему более 2000 лет, в течение сотен лет он был крупным портом по пути из Константинополя на запад. Расположенный на полуострове в окружении небольших островков Адриатического моря, этот живописный город служит популярной туристической достопримечательностью, привлекая своими архитектурными сокровищами – развалинами римских времен, рвами и старыми каменными стенами. Персонаж с обложки облачен в синие шерстяные брюки и белую льняную рубашку, поверх которой он надел синий жилет и куртку, отделанные характерной для этого района красочной вышивкой. Красные шерстяные пояс и шляпа дополняют костюм.

Принятая манера одеваться и стиль жизни заметно изменились за последние 200 лет. Региональные различия, столь заметные в прошлом, сильно размылись. Сейчас по внешнему виду сложно различить жителей разных континентов, не говоря уже о разных деревнях или городах, расположенных всего лишь на расстоянии нескольких километров друг от друга. Возможно, разнообразие культур преобразовалось в разнообразие личной жизни – и уж, конечно, в более разнообразную и динамичную технологичную жизнь.

Издательство Маннинг отмечает изобретательность и инициативность компьютерных технологий обложками книг, на которых представлено разнообразие региональных культур два столетия назад. Эти культуры оживают в нашей памяти благодаря иллюстрациям из старых книг и коллекциям, таким как эта.

Часть I.

НАЧАЛО РАБОТЫ

Испытайте R в действии! R – одно из наиболее популярных современных программных средств для анализа данных и их визуализации. Это бесплатная программа с открытым кодом, предназначенная для операционных систем Windows, Mac OS X и Linux. Благодаря этой книге вы приобретете навыки, необходимые для овладения этой многофункциональной программой и ее эффективного использования для обработки ваших собственных данных.

Книга разделена на четыре части. Первая часть посвящена установке программы в ее базовой версии, знакомству с интерфейсом, импорту данных и преобразованию их в удобный для дальнейшего анализа вид.

Глава 1 познакомит вас с программной средой R. Эта глава начинается с обзора программы R и ее особенностей, которые делают ее столь мощным программным средством для современного анализа данных. После краткого объяснения того, как скачать и установить программу, следует описание пользовательского интерфейса на ряде простых примеров. Затем вы научитесь тому, как увеличить функциональность базовой версии при помощи расширений системы команд (так называемых дополнительных пакетов), которые можно скачать бесплатно с сетевых хранилищ. В конце главы размещены примеры, которые позволят применить ваши новые умения.

Как только вы познакомились с интерфейсом R, возникает следующая задача – загрузить ваши данные в программу. В современном богатом информацией мире данные могут поступать из разных источников и в разных форматах. В главе 2 описано множество методов, которые можно использовать для импорта данных в R. Первая половина главы посвящена описанию форматов, в которых R хранит данные, и

рассказу о том, как вводить данные вручную. Во второй части обсуждаются методы импорта данных из текстовых файлов, веб-страниц, электронных таблиц, других статистических программ и баз данных.

Исходя из последовательности действий при обработке данных, возможно, следующим пунктом имело бы смысл обсудить управление данными и устранение ошибок в них. Однако многие пользователи, впервые познакомившиеся с R, больше интересуются ее мощными графическими возможностями. Чтобы не игнорировать этот интерес и не заставлять вас ждать, в главе 3 мы немедленно переходим к графическому представлению данных. В этой главе обсуждается, как создавать диаграммы, изменять их параметры и сохранять диаграммы в разных форматах. Рассказано, как на диаграммах выбирать цвета, типы символов и линий, шрифты, делать заголовки, надписи и списки условных обозначений. В заключение описано, как объединить несколько диаграмм на одном изображении.

После того как вам представилась возможность испытать графические возможности R, настало время вернуться к анализу данных. Они редко сразу поступают в готовом к использованию виде. Часто бывает необходимо потратить значительное количество времени, комбинируя данные из различных источников, устраняя ошибки (неправильно закодированные, несоответствующие, отсутствующие данные) и создавая новые (комбинированные, трансформированные, перекодированные) переменные, прежде чем вы сможете перейти к решению интересующих вас задач. В главе 4 описаны все основные способы управления данными в R, включая сортировку, слияние и разделение наборов данных, а также трансформацию, перекодировку и удаление переменных.

Глава 5 основана на материале, изложенном в главе 4. Здесь рассказано, как использовать числовые (арифметические, тригонометрические и статистические) и текстовые функции (разбиение строк, объединение и замену) в управлении данными. Для иллюстрации многих из описанных функций в этом разделе использованы многочисленные примеры. Далее разобраны управляющие конструкции (циклы, исполняемые при определенных условиях команды). После прочтения этого раздела вы научитесь создавать собственные функции в R. Это позволит расширить возможности R, объединив многие команды в одну легко настраиваемую функцию. В заключение обсуждаются мощные методы реорганизации и группировки данных, которые часто бывают полезными при подготовке данных к дальнейшему анализу.

После прочтения главы 1 вы подробно познакомитесь с программированием в среде R. Вы приобретете навыки, необходимые для ввода данных и получения их из внешних источников, а также для устранения ошибок в данных. Кроме того, вы получите опыт создания, изменения параметров и сохранения различных типов диаграмм.