

*Посвящается моей жене Рейчел
и моим детям Лейле и Элиасу*

СОДЕРЖАНИЕ

От издательства.....	11
Отзывы о книге «R за пределами статистики».....	12
Об авторе	13
Благодарности	14
Введение	15
 Часть I. ВИЗУАЛИЗАЦИИ	 19
Глава 1. Экспресс-курс программирования на языке R	20
Настройка.....	20
Установка R и RStudio	21
Знакомство с интерфейсом RStudio	21
Файлы скриптов R.....	22
Базовый синтаксис R	23
Арифметические операторы	23
Операторы сравнения	24
Функции.....	25
Работа с данными	26
Импорт данных.....	26
Сохранение данных в объектах.....	28
Установка пакетов	29
Проекты RStudio.....	31
Анализ данных с помощью tidyverse	32
Функции tidyverse	32
Конвейер tidyverse.....	33
Комментарии.....	34
Как получить помощь	35
Заключение	36
Дополнительные ресурсы.....	36
 Глава 2. Принципы визуализации данных.....	 37
Визуализация засухи	38

Грамматика графики	41
Работа с ggplot	42
Соотнесение данных с эстетическими свойствами	43
Выбор геометрических объектов	44
Изменение эстетических свойств	48
Установка темы	49
Обновленная визуализация засухи	50
Построение графика по одному региону и году	51
Изменение эстетических свойств	53
Фасетизация графиков	54
Добавление финальных нюансов	57
Полный код визуализации	61
Заключение	64
Дополнительные ресурсы	65
 Глава 3. Пользовательские темы визуализации данных	67
Стилизация графика с помощью пользовательской темы	68
Пример графика	68
Пользовательская тема BBC	70
Компоненты темы BBC	71
Определение функции	71
Текст	73
Легенда	74
Оси	75
Линии сетки	76
Фон	77
Маленькие множественные графики	77
Цвет	79
Заключение	79
Дополнительные ресурсы	80
 Глава 4. Карты и геопространственные данные	81
Краткий курс по геопространственным данным	82
Тип геометрии	83
Размеры	86
Ограничительная рамка	86
Система отсчета координат	87
Столбец geometry	88
Повторное создание карты COVID-19	88
Импорт данных	89
Расчет ежедневных случаев COVID-19	91
Вычисление показателей заболеваемости	92
Добавление геопространственных данных	94
Создание карты	96
Создание ваших собственных карт	100
Импорт необработанных данных	101
Доступ к геопространственным данным с помощью функций R	101
Использование подходящих проекций	103
Обработка геопространственных данных	104

Заключение	105
Дополнительные ресурсы	106
Глава 5. Создание эффективных таблиц	107
Создание фрейма данных	108
Принципы создания таблиц	109
Минимизация беспорядка	109
Разграничение заголовка и основной части	111
Надлежащее выравнивание	112
Выбор корректного уровня точности	115
Осмысленное использование цветовой гаммы	116
Добавление визуализации данных в необходимых случаях	117
Заключение	119
Дополнительные ресурсы	120
Часть II. ОТЧЕТЫ, ПРЕЗЕНТАЦИИ И ВЕБ-САЙТЫ	121
Глава 6. Отчеты R Markdown	122
Создание документа R Markdown	123
Структура документа	124
Метаданные YAML	124
Фрагменты кода R	125
Текст в формате Markdown	127
Встроенный код R	129
Интерактивный запуск фрагментов кода	130
Quarto	132
Заключение	132
Дополнительные ресурсы	132
Глава 7. Параметризованная отчетность	133
Шаблоны отчетов в R Markdown	133
Определение параметров	136
Генерация чисел с помощью параметров	138
Включение параметров в код визуализации	138
Создание скрипта на языке R	140
Связывание документа с помощью кода	140
Создание Tibble с данными о параметрах	140
Лучшие практики	144
Заключение	145
Дополнительные ресурсы	145
Глава 8. Презентации слайд-шоу	146
Почему xaringan?	146
Как работает xaringan	147
Создание нового слайда	148
Настройка размера фигур	149
Инкрементное раскрытие контента	149
Согласование контента с классами контента	150
Добавление фоновых изображений на слайды	152

Добавление CSS к слайдам	153
Пользовательские CSS	153
Темы	154
Пакет xaringanthemer.....	156
Заключение.....	157
Дополнительные ресурсы.....	157

Глава 9. Веб-сайты

Создание нового проекта distill.....	159
Файлы проекта	160
Документы R Markdown.....	160
Файл _site.yml.....	161
Создание сайта	162
Применение пользовательских CSS.....	163
Работа с контентом веб-сайта	165
Применение макетов distill	169
Добавление интерактивного контента	171
Хостинг веб-сайта	176
Облачный хостинг.....	176
Хостинг GitHub.....	177
Заклучение.....	178
Дополнительные ресурсы.....	178

Глава 10. QUARTO

Создание документа Quarto	179
Сравнение R Markdown и Quarto	181
Формат и исполнение полей YAML	181
Параметры отдельных фрагментов кода.....	182
Тире в именах опций.....	182
Кнопка рендеринга (Render).....	183
Отчеты с параметрами.....	183
Создание презентации.....	186
Инкрементное раскрытие контента	187
Выравнивание контента и добавление фоновых изображений	188
Настройка слайдов с помощью тем и CSS	190
Создание веб-сайтов	192
Разработка веб-сайта.....	192
Установка опций.....	193
Изменение внешнего вида веб-сайта	194
Настройка заголовка и панели навигации	197
Создание более широких макетов.....	197
Размещение вашего веб-сайта на GitHub Pages и Quarto Pub.....	198
Заклучение.....	200
Дополнительные ресурсы.....	200

Часть III. АВТОМАТИЗАЦИЯ И ВЗАИМОДЕЙСТВИЕ

Глава 11. Автоматический доступ к онлайн-данным.....

Импорт данных из Google Sheets с помощью googlesheets4.....	202
---	-----

Подключение к Google.....	203
Чтение данных из таблиц Google.....	203
Работа с данными в R Markdown.....	204
Импорт определенных столбцов	208
Доступ к данным переписи населения с помощью tidycensus	209
Подключение к Бюро переписи населения США с помощью ключа API.....	210
Работа с данными Десятилетней переписи населения США	210
Определение значений переменных переписи населения	211
Использование нескольких переменных переписи населения	212
Анализ данных переписи населения	214
Использование итоговой переменной	215
Визуализация данных Американского общественного опроса	216
Составление диаграмм	217
Создание карт населения с помощью аргумента geometry.....	219
Заключение	222
Дополнительные ресурсы	222
Глава 13. Создание функций и пакетов	223
Создание ваших собственных функций	224
Написание простой функции	224
Добавление аргументов	225
Создание функции для форматирования данных о расе и этнической принадлежности.....	226
Использование ... для передачи аргументов другой функции.....	229
Разработка пакета.....	232
Создание пакета	232
Добавление функций с помощью use_r()	233
Проверка пакета с помощью devtools	233
Добавление взаимозависимых пакетов	235
Корректное обращение к функциям.....	236
Создание документации с помощью Roxugen	237
Добавление лицензии и метаданных.....	239
Написание дополнительных функций.....	240
Инсталляция пакета.....	240
Заключение	241
Дополнительные ресурсы.....	241
В завершение	241
Предметный указатель	243

ОТ ИЗДАТЕЛЬСТВА

Отзывы и пожелания

Мы всегда рады отзывам наших читателей. Расскажите нам, что вы думаете об этой книге – что понравилось или, может быть, не понравилось. Отзывы важны для нас, чтобы выпускать книги, которые будут для вас максимально полезны.

Вы можете написать отзыв на нашем сайте www.dmkpress.com, зайдя на страницу книги и оставив комментарий в разделе «Отзывы и рецензии». Также можно послать письмо главному редактору по адресу dmkpress@gmail.com; при этом укажите название книги в теме письма.

Если вы являетесь экспертом в какой-либо области и заинтересованы в написании новой книги, заполните форму на нашем сайте по адресу http://dmkpress.com/authors/publish_book/ или напишите в издательство по адресу dmkpress@gmail.com.

Список опечаток

Хотя мы приняли все возможные меры для того, чтобы обеспечить высокое качество наших текстов, ошибки все равно случаются. Если вы найдете ошибку в одной из наших книг, мы будем очень благодарны, если вы сообщите о ней главному редактору по адресу dmkpress@gmail.com. Сделав это, вы избавите других читателей от недопонимания и поможете нам улучшить последующие издания этой книги.

Нарушение авторских прав

Пиратство в интернете по-прежнему остается насущной проблемой. Издательство «ДМК Пресс» очень серьезно относится к вопросам защиты авторских прав и лицензирования. Если вы столкнетесь в интернете с незаконной публикацией какой-либо из наших книг, пожалуйста, пришлите нам ссылку на интернет-ресурс, чтобы мы могли применить санкции.

Ссылку на подозрительные материалы можно прислать по адресу электронной почты dmkpress@gmail.com.

Мы высоко ценим любую помощь по защите наших авторов, благодаря которой мы можем предоставлять вам качественные материалы.

ОТЗЫВЫ О КНИГЕ «R ЗА ПРЕДЕЛАМИ СТАТИСТИКИ»

«...Фантастический и бесценный ресурс для всех, кто работает или планирует работать с данными, вне зависимости от уровня их подготовки. Настоятельно рекомендую».

Габриэла де Кейрос, директор департамента ИИ, Microsoft

«Долгожданная книга в экосистеме R».

Оскар Баруффа, автор «Большой книги по R»

«Отличный ресурс для всех изучающих язык R...».

Кара Томпсон, консультант по визуализации данных

«Дэвид представил наиболее эффективные способы использования R для визуализации данных и карт, автоматизации, составления отчетов и разработки веб-сайтов».

Том Мок, менеджер по продуктам, Posit

«Источник захватывающих примеров, актуальных фрагментов кода и полезных объяснений в удобном для восприятия виде...»

*Седрик Шерер,
специалист по визуализации данных и информационной графике*

«Наглядная демонстрация лаконичной эффективности R для быстрого решения нестатистических задач».

*Боб Рудис, вице-президент по науке о данных,
исследованиям безопасности и разработке средств обнаружения,
Greynoise Intelligence*

ОБ АВТОРЕ

Дэвид Кис (David Keyes) является основателем компании R for the Rest of Us (<https://rfortherestofus.com>), где он занимается разработкой курсов, проведением корпоративных тренингов и взаимодействием с различными организациями с целью эффективного использования возможностей языка R. Под его руководством команда консультантов создала множество проектов с использованием R. Обладая качественным профессиональным образованием, он самостоятельно освоил R и теперь помогает использовать этот мощный инструмент всем, кто не может назвать себя пользователем R.

О техническом редакторе

Рита Джордано (Rita Giordano) занимается визуализацией данных и является независимым консультантом-преподавателем LinkedIn в Великобритании. Будучи по специальности физиком, она также имеет докторскую степень по статистике в области структурной биологии и обладает обширным опытом в сфере исследований и работы с данными.

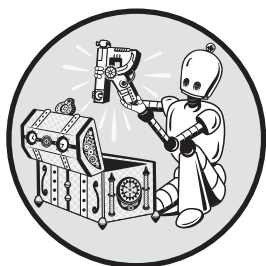
БЛАГОДАРНОСТИ

Эта книга является наглядным свидетельством взаимодействия многочисленных членов сообщества R, которые щедро делятся своими знаниями и вдохновляют других. Я называю себя самоучкой, но на самом деле мое обучение основано на опыте сообщества. На страницах этой книги рассказывается о нескольких пользователях R, от которых я многому научился; многие другие остались неупомянутыми. Спасибо всем, кто работал над развитием R, делился своими знаниями о R и превратил сообщество R в гостеприимное место.

Я также хотел бы поблагодарить команду R for the Rest of Us. Много нового о возможностях R я почерпнул именно через непосредственное сотрудничество с талантливыми пользователями R.

Наконец, я хотел бы поблагодарить всех, кто помогал мне в написании этой книги. Технический рецензент Рита Джордано помогла мне удостовериться, что все работает, и предложила отличные идеи для улучшения. Мой редактор Фрэнсис Со внесла фантастический вклад в работу над этой книгой. Биллу Поллоку и всей команде издательства No Starch: спасибо, что отнеслись с пониманием ко мне и моей странной идее написать книгу о нестатистическом использовании инструмента, созданного для статистики.

ВВЕДЕНИЕ



В начале 2020 года, когда весь мир боролся с распространением вируса COVID-19, одной стране удалось добиться успеха тогда, когда другие не смогли. Речь о Новой Зеландии. Причин, по которым Новая Зеландия смогла справиться с COVID-19, достаточно много. Одной из них стал язык программирования R (да, это правда).

Этот скромный инструмент для анализа данных помог Новой Зеландии в борьбе с COVID-19, потому что позволил команде Министерства здравоохранения генерировать ежедневные отчеты о случаях заболевания по всей стране. На основании информации из этих отчетов власти смогли разработать политику, благодаря которой страна в значительной степени избежала распространения COVID-19. Однако команда специалистов была небольшой, и поэтому ежедневная подготовка отчетов с помощью такого инструмента, как Excel, была бы невозможна. Как рассказывал мне руководитель проекта Крис Нокс (Chris Knox), «попытаться повторить сделанное нами в системе “наведи и кликни мышкой” попросту невозможно».

Вместо этого нескольким сотрудникам удалось написать код на языке R, и этот код можно было запускать ежедневно для создания обновленных отчетов. В этих отчетах не было сложной статистики; это были в буквальном смысле подсчеты случаев COVID-19. Их ценность заключалась во всем остальном, что может делать R: анализе и визуализации данных, создании отчетов и автоматизации рабочих процессов.

В этой книге рассматривается множество вариантов использования R для взаимодействия и автоматизации задач. Из нее вы узнаете:

- как создавать визуализации данных, карты и таблицы профессионального качества;
- как заместить громоздкий мультиинструментальный рабочий процесс для создания отчетов с помощью R Markdown;
- как использовать параметризованные отчеты для создания нескольких отчетов одновременно;

- как создавать презентации со слайд-шоу и веб-сайты с помощью R Markdown;
- как автоматизировать процесс импорта онлайн-данных из Google Sheets и Бюро переписи населения США;
- как создать собственные функции для автоматизации многократно повторяющихся задач;
- как собрать свои функции в пакет, которым можно поделиться с другими.

И самое главное – вы сможете делать все это без проведения какого-либо статистического анализа, более сложного, чем вычисление средних величин.

Разве R не предназначен исключительно для статистического анализа?

Многие считают, что R – это всего лишь инструмент для углубленного статистического анализа. Однако этот язык подходит не только для работы с числовыми значениями. В конце концов, каждому пользователю R каким-либо образом приходится иллюстрировать свои выводы и сообщать о своих результатах, будь то визуализация данных, отчеты, веб-сайты или презентации. Кроме того, чем больше вы используете R, тем чаще вам захочется автоматизировать задачи, которые обычно выполняются вручную.

Будучи квалифицированным антропологом без профильного образования в области статистики, я часто стеснялся использовать R для своих задач визуализации и коммуникации. Но дело в том, что R отлично подходит для этих целей. Пакет `ggplot2` – это инструмент, который выбирают многие ведущие дизайнеры в области информационных технологий. Пользователи во всем мире активно используют возможности R для автоматизации отчетности и повышения эффективности своей работы. Язык R не только заменяет многие инструменты, но также позволяет выполнять более эффективно другие ваши повседневные задания, такие как создание отчетов и таблиц.

На кого рассчитана эта книга

Использование R может преобразить вашу работу вне зависимости от уровня вашей подготовки. Эта книга адресована вам, если вы либо уже используете R и стремитесь изучить его возможности для визуализации и коммуникации, либо еще незнакомы с R и пытаетесь понять, подходит ли вам этот язык. Я написал книгу «R для всего остального» таким образом, чтобы она была понятна даже тем, кто никогда не писал ни строчки кода на R. Но даже если вы писали полноценные программы на этом языке, книга позволит вам узнать множество новых приемов, которые пригодятся в дальнейшем.

Язык R – отличный инструмент для любого, кто работает с данными. Возможно, вы занимаетесь исследованиями и рассматриваете новые способы поделиться результатами своей работы. Или вы журналист, который ищет более эффективные способы анализа открытых данных. Или же вы аналитик дан-

ных, уставший от необходимости использовать дорогостоящие проприетарные инструменты. В любом случае, если вам приходится работать с данными, то R будет для вас очень полезен.

Краткое содержание книги

Каждая глава этой книги посвящена одному из возможных способов использования языка R и содержит примеры реальных проектов на R с практическим применением рассматриваемых методик. Я погружаюсь в код проекта и разбираю программы на отдельные фрагменты, для того чтобы помочь вам разобраться с принципами их работы, а также предлагаю способы выйти за рамки приведенного примера. Книга состоит из трех отдельных частей, описанных ниже.

В части I вы познакомитесь с методами использования R для визуализации данных.

- **Глава 1 «Краткий курс программирования на языке R».** Знакомство со средой программирования RStudio и основами синтаксиса языка R, которые понадобятся вам для понимания остальной части книги.
- **Глава 2 «Принципы визуализации данных».** Подробно рассматривается создание визуализации для Scientific American, описывающей ситуацию с засухами в США. Кроме того, в этой главе представлен пакет `ggplot2` для визуализации данных и рассмотрены важные принципы создания высококачественных графиков.
- **Глава 3 «Пользовательские темы для визуализации данных».** Описание процесса создания журналистами BBC пользовательской темы для пакета визуализации данных `ggplot2`. По мере изучения созданного ими пакета вы сможете освоить процесс создания собственной темы.
- **Глава 4 «Карты и геопространственные данные».** Процесс создания карт на языке R с использованием простых данных об объектах. Вы научитесь писать код для создания карт, находить геопространственные данные, выбирать подходящие проекции и применять принципы визуализации данных для повышения эффективности вашей карты.
- **Глава 5 «Проектирование эффективных таблиц».** Вы узнаете, как использовать пакет `gt` для создания высококачественных таблиц с помощью R. Под руководством знатока таблиц Тома Мока вы научитесь принципам дизайна, обеспечивающим эффективное представление данных в таблицах.

Часть II посвящена использованию R Markdown для эффективной коммуникации. Вы узнаете об использовании визуализаций из части I в отчетах, слайд-шоу и статических веб-сайтах, полностью созданных на основе кода R.

- **Глава 6 «Отчеты в R Markdown».** Знакомство с R Markdown – инструментом для создания профессиональных отчетов на языке R. В этой главе рассматривается структура документа в R Markdown, рассказывается о способах использования встроеного кода для автоматического обновления текста отчета при изменении значений данных, а также описываются разнообразные возможности экспорта.

- **Глава 7 «Параметризованная отчетность».** Описание одного из преимуществ использования R Markdown – возможности одновременного создания нескольких отчетов с помощью так называемой параметризованной отчетности. Вы увидите, как сотрудники Городского института использовали R для создания финансовых сводок по всем 50 штатам США. В процессе вы познакомитесь с особенностями применения параметризованных отчетов и научитесь их использовать.
- **Глава 8 «Презентации в виде слайд-шоу».** Описание возможностей использования R Markdown для создания слайдов с помощью пакета `xaringan`. Вы получите представление о создании собственных презентаций, настройке контента для размещения на слайдах и добавлении эффектов в слайд-шоу.
- **Глава 9 «Веб-сайты».** В этой главе описывается процесс создания собственного веб-сайта с помощью R Markdown и пакета `distill`. На примере веб-сайта о показателях COVID-19 в округе Вестчестер, штат Нью-Йорк, вы сможете ознакомиться с методами создания страниц на сайте, добавления интерактивности с помощью пакетов R и развертывания веб-сайта различными способами.
- **Глава 10 «Quarto».** Описание возможностей использования Quarto, версии R Markdown нового поколения. Вы узнаете об особенностях применения Quarto для всех проектов, в которых вы ранее использовали R Markdown (отчеты, отчеты с параметрами, слайд-шоу, презентации и веб-сайты).

Часть III посвящена возможностям использования R для автоматизации рабочего процесса и обмена данными с другими пользователями.

- **Глава 11 «Автоматический доступ к онлайн-данным».** Изучение двух пакетов R для автоматического импорта данных из Интернета: `google-sheets4` для работы с Google Sheets и `tidycensus` для работы с данными Бюро переписи населения США. Вы познакомитесь с принципами работы пакетов и научитесь использовать их для автоматизации процесса доступа к данным.
- **Глава 12 «Создание функций и пакетов».** В этой главе описывается порядок создания собственных функций и пакетов, а также их совместного использования с другими пользователями, что является одним из главных преимуществ R. Объединение ваших пользовательских функций в пакет может дать возможность другим пользователям R оптимизировать свою работу, о чем вы узнаете на примере пакетов группы разработчиков R, созданных для исследователей Центра онкологии Моффитта.

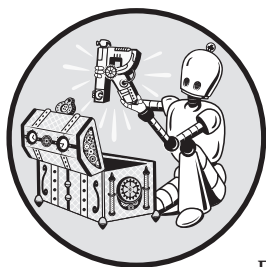
После прочтения этой книги вы сможете использовать R для решения широкого круга нестатистических задач. Вы научитесь эффективно визуализировать данные и представлять свои выводы с помощью карт и таблиц. Вы сможете интегрировать свои результаты в отчеты с помощью R Markdown, а также создавать наглядные презентации и веб-сайты. И вы узнаете о способах автоматизации многих рутинных задач с помощью пакетов, созданных другими или разработанных вами самостоятельно. Так давайте же приступим!

ЧАСТЬ I

ВИЗУАЛИЗАЦИИ

1

ЭКСПРЕСС-КУРС ПРОГРАММИРОВАНИЯ НА ЯЗЫКЕ R



У R сложилась заслуженная репутация достаточно сложного в изучении языка, особенно для тех, кто обращается к нему без предварительного опыта программирования. Эта глава призвана помочь каждому, кто прежде никогда не сталкивался с R. Вы сможете настроить среду программирования R с помощью RStudio и научитесь использовать функции, объекты, пакеты и проекты для работы с данными. Вы также познакомитесь с пакетом *tidyverse*, который содержит основные функции анализа и обработки данных для примеров в этой книге.

Эта глава не является полноценным вводным курсом в программирование на R; скорее, здесь описаны те основы, которые потребуются вам для работы с остальными частями книги. Если у вас уже есть опыт работы с R, можете сразу перейти к главе 2.

Настройка

Для эффективного использования R вам понадобятся два программных продукта. Во-первых, собственно R, который предоставляет базовые вычислительные инструменты для работы с языком. Во-вторых, интегрированная среда разработки (Integrated Development Environment, IDE), такая как RStudio. Именно эта платформа для кодирования упрощает работу с языком R. Лучше всего понять взаимосвязь между R и RStudio можно с помощью аналогии из

Конец ознакомительного фрагмента.

Приобрести книгу можно

в интернет-магазине

«Электронный универс»

e-Univers.ru