2-е издание

R для Data Science

Импорт, упорядочивание, преобразование, визуализация и моделирование данных

Хэдли Уикхэм
Майн Четинкая-Рэндел
Гаррет Гролемунд

УДК 004.032.2
ББК 32.973.26-018.2
У35

R for Data Science, 2E
Import, Tidy, Transform, Visualize, and Model Data

Garrett Grolemund, Hadley Wickham, Mine Çetinkaya-Rundel

Уикхэм, Хэдли.

У35 R для Data Science. Импорт, упорядочивание, преобразование, визуализация и моделирование данных, 2-е издание / Хэдли Уикхэм, Майн Четинкая-Рэндел, Гаррет Гролемунд: [перевод с английского Е. Белова]. — Алматы : Астана иностранная пресса, 2026. — 704 с. — (O'Reilly. Книги по программированию).

ISBN 978-601-12-6020-6

«R для Data Science. Импорт, упорядочивание, преобразование, визуализация и моделирование данных, 2-е издание» — это практичное введение в работу с данными на языке R, созданное ведущими специалистами сообщества. Книга помогает начинающим дата-сайентистам освоить основные инструменты, необходимые для полноценного анализа данных: от импорта и очистки до визуализации и моделирования.

Авторы последовательно показывают, как превращать разрозненные данные в структурированную информацию, выбирать подходящие инструменты и эффективно представлять результаты анализа. Читатель научится загружать данные из разных источников, строить информативные графики и автоматизировать части анализа с помощью базового программирования на R. Упражнения помогают закрепить навыки и сразу применять их на практике.

УДК 004.032.2
ББК 32.973.26-018.2

ISBN 978-601-12-6020-6

Оглавление

Часть IV
Импорт

Часть V
Программирование

Часть VI
Коммуникация

Введение

Data Science — это увлекательная дисциплина, которая позволяет преобразовывать необработанные первичные данные в догадки, гипотезы и новые знания. Цель настоящего издания — помочь в освоении наиболее важных инструментов языка программирования R, что позволит вам эффективно и самостоятельно заниматься наукой о данных, получая удовольствие от процесса. Прочитав данную книгу, вы получите инструменты для решения широкого спектра задач в области обработки данных с использованием основных возможностей языка R.

Предисловие ко второму изданию

Добро пожаловать во второе издание «R для Data Science. Импорт, упорядочивание, преобразование, визуализация и моделирование данных»! Настоящая книга являет собой серьезную переработку первого издания. Мы удалили материал, который, по нашему мнению, больше не является полезным, добавив информацию, которую следовало бы включить в первое издание. С целью отражения наилучших практик применения языка R в новое издание мы включим обновленные примеры кода и его описание. Мы очень рады приветствовать нового соавтора Майн Четинкая-Рэндел, известного преподавателя в области науки о данных и нашу коллегу в компании Posit (ранее известной как RStudio).

Ниже приводится краткое описание наиболее значительных изменений:

— Первая часть книги переименована в «Весь процесс от А до Я». Цель этого раздела — сформировать у читателя общее представвление об «игре с данными» в целом, прежде чем мы углубимся в детали.
— Вторая часть книги озаглавлена «Визуализация данных». По сравнению с первым изданием в этой части книги инструменты визуализации данных и лучшие практики рассматриваются более подробно. Наилучшим источником подробностей о визуализации по-прежнему является книга ggplot2 (https://ggplot2-book.org/), однако настоящее издание охватывает большее количество ключевых методов работы.
— Третья часть книги теперь называется «Преобразование» и содержит новые главы, посвященные работе с числовыми данными, логическими

векторами и отсутствующими значениями. Ранее данные вопросы уже освещались в главе об обработке данных, однако теперь, чтобы подробно охватить все детали, потребовалось более широкое описание.

— Четвертая часть книги называется «Импорт». Эта часть включает в себя набор глав, который выходит за рамки чтения простых текстовых файлов и включает в себя работу с электронными таблицами, получение данных из баз данных, работу с большими данными, выстраивание иерархических данных и сбор данных с веб-сайтов.
— Часть «Программирование» перешла из предыдущего издания, претерпев при этом множество изменений. Новая версия данного раздела сфокусирована на наиболее важных нюансах написания и выполнения функций. Программная реализация функций теперь включает подробное описание того, как прописывать функции этапа подготовки данных, например рассматриваются функции коллекции пакетов tidyverse (уделяется внимание проблеме оценки на этапе подготовки), поскольку за последние несколько лет их программная реализация стала проще и приобрела большую важность. Кроме того, новое издание пополнилось новой главой о важных базовых функциях R, которые вы, вероятно, можете встретить в примерах кода на R в реальной практике.
— Раздел «Моделирование» был удален. Мы никогда не располагали достаточными ресурсами, чтобы в рамках одной книги должным и полноценным образом осветить аспект моделирования. Кроме того, на сегодняшний день доступны гораздо лучшие источники. Как правило, мы рекомендуем использовать пакеты tidymodels и прочитать *Tidy Modeling with R* Макса Куна и Джулии Силге (изд-во O'Reilly).
— Часть «Коммуникация» осталась, но была тщательным образом обновлена и теперь включает описание работы с Quarto взамен R Markdown. Данное издание написано в Quarto, и это явно инструмент будущего.

Чему вы научитесь

Наука о данных — обширная область, и вы не сможете освоить все, прочитав всего лишь одну книгу. Наша цель — дать вам базовые навыки работы с наиболее важными инструментами, а также передать достаточно знаний, чтобы при необходимости вы могли находить ресурсы для расширения ваших компетенций. Наша модель этапов типового проекта по анализу данных выглядит примерно так, как показано на рис. I-1.

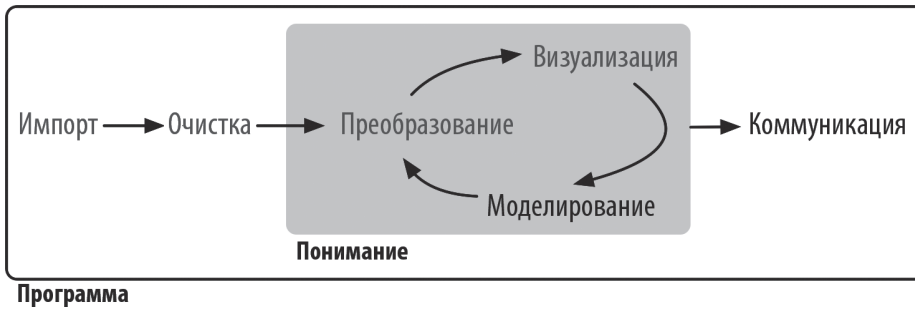

Рис. I-1. *В нашей модели процесса науки о данных вы начинаете с импорта данных и их очистки. Затем вы понимаете свои данные с помощью итерационного цикла преобразования, визуализации и моделирования. Завершается процесс передачей результатов другим людям*

Во-первых, необходимо импортировать данные в среду R. Обычно это означает, что вы берете данные, хранящиеся в файле, базе данных или интерфейсе программирования веб-приложений (API), и загружаете их во фрейм данных в R. Если вы не можете импортировать данные в R, то вы не сможете выполнить их анализ.

После того как вы импортировали данные, рекомендуется привести их в порядок. Упорядочение данных означает их хранение в согласованной форме, которая соответствует семантике набора данных и способу его хранения. Проще говоря, ваши данные должны иметь вид, при котором в каждом столбце указана переменная, а в каждой строке — значение. Подготовка данных очень важна, поскольку последовательная структура позволяет сосредоточить усилия на ответах на вопросы о данных, а не на попытках привести данные к структуре, требуемой той или иной функцией.

Как только ваши данные будут подготовлены, следующим шагом будет их обработка. Обработка включает в себя сужение круга релевантных значений (например, всех людей в одном городе или всех данных за последний год), создание новых переменных, которые являются функциями существующих переменных (например, вычисление скорости на основе расстояния и времени) и вычисление сводных статистических данных (например, подсчет количества или вычисление среднего значения). Вместе процессы подготовки и обработки данных мы называем «разборкой», поскольку работа с данными, полученными в их естественном виде, напоминает борьбу с упорным соперником.

Как только вы подготовите данные с необходимыми переменными, у вас будут два основных механизма генерации знаний: визуализация и моделирование.

У этих методов есть взаимодополняющие сильные и слабые стороны, поэтому в реальной ситуации анализ данных потребует множества итераций каждого из методов.

Визуализация — фундаментальный аспект человеческого познания. Эффективная визуализация не только раскрывает неожиданные закономерности в данных, но и стимулирует новые вопросы, подталкивая к переосмыслению исследовательской задачи. Она может указать на некорректную постановку проблемы или необходимость сбора дополнительных данных. Хотя визуализация способна на неожиданные открытия, ее масштабируемость ограничена: для интерпретации результатов требуется участие человека.

Модели служат дополнительным инструментом визуализации. После четкой формулировки исследовательского вопроса модель может обеспечить ответ. Будучи математическими или вычислительными инструментами, модели, как правило, хорошо масштабируются. Даже при отсутствии масштабируемости, увеличение вычислительных мощностей (приобретение дополнительных компьютеров) обычно обходится дешевле, чем расширение исследовательской группы (привлечение большего числа специалистов). Однако каждая модель базируется на определенных предположениях, которые она, по своей природе, не может подвергнуть критике. Следовательно, модель неспособна на принципиально новые, неожиданные открытия.

Финальный этап в науке о данных — коммуникация, критически важная составляющая любого проекта анализа данных. Независимо от эффективности моделей и интерпретации данных посредством визуализации, результаты останутся бесполезными, если не будут эффективно донесены до целевой аудитории.

Все перечисленные инструменты тесно связаны с программированием, которое является универсальным средством, применяемым практически на всех этапах проекта по науке о данных. Необязательно быть высококвалифицированным программистом для достижения успеха в этой области, однако углубление знаний в программировании весьма полезно. Улучшение навыков программирования позволяет автоматизировать рутинные задачи и эффективнее решать новые проблемы.

Эти инструменты будут применяться в каждом проекте по науке о данных, но для большинства проектов их недостаточно. Существует принцип 80/20, который гласит, что вы сможете выполнить примерно 80% работы над проектом, используя инструменты, представленные в этой книге. Однако для завершения оставшихся 20% вам понадобятся дополнительные инструменты. В данной книге мы даем ссылки на источники, с помощью которых вы сможете углубить свои знания и изучить новые инструменты.

Как устроена эта книга

Предыдущая структура описания инструментов обработки данных выстроена в порядке, соответствующем их использованию на этапе анализа. Однако, исходя из нашего опыта, начинать изучение с импорта и очистки данных не является оптимальным подходом. Эти процессы в 80% случаев представляют собой рутинные и достаточно скучные задачи, а в оставшиеся 20% могут вызывать замешательство и разочарование. Это не самое удачное начало для освоения нового предмета. Вместо этого мы сосредоточимся на визуализации и обработке данных, которые уже были успешно импортированы и подготовлены. Такой подход позволит вам сохранять высокий уровень мотивации, поскольку вы сможете увидеть результаты своей работы и понять, что затраченные усилия оправданы.

В каждой главе мы придерживаемся последовательной структуры: начинаем с нескольких вдохновляющих примеров, чтобы представить общую картину, а затем углубляемся в детали темы. Каждый раздел книги включает упражнения, которые позволят вам закрепить полученные знания на практике. Хотя может возникнуть соблазн пропустить выполнение этих упражнений, помните, что ничто так не способствует обучению, как практическая работа с реальными задачами.

Чему вы не научитесь

Наша книга не охватывает некоторые ключевые темы, о которых мы считаем важным упомянуть. Мы придерживаемся принципа: необходимо концентрировать усилия на самых значимых аспектах, чтобы вы могли как можно быстрее приступить к работе. Поэтому в рамках данной книги невозможно рассмотреть весь круг важных вопросов.

Моделирование

Моделирование играет ключевую роль в науке о данных, однако это обширная тема, и, к сожалению, в рамках нашей книги нет возможности рассмотреть ее должным образом. Для более глубокого изучения моделирования мы настоятельно рекомендуем ознакомиться с книгой *Tidy Modeling with R*, написанной нашими коллегами Максом Куном и Джулией Силге (изд-во O'Reilly). В этом издании вы познакомитесь с семейством пакетов tidymodels, которые, как можно предположить, имеют много общего с пакетами tidyverse, используемыми в нашей книге.

Большие данные

В нашей книге мы делаем акцент на работе с небольшими массивами данных, которые хранятся в памяти. Это идеальная отправная точка, поскольку опыт работы с малыми данными необходим для успешного освоения работы с большими данными. Инструменты, которые вы изучите, отлично подходят для обработки сотен мегабайт данных и с должной осторожностью могут использоваться и для работы с несколькими гигабайтами данных. Мы также рассмотрим, как извлекать данные из баз данных и файлов parquet, которые обычно применяются для хранения больших объемов данных. Вы не всегда сможете работать с полным набором данных, но это не проблема: зачастую для ответа на интересующий вас вопрос достаточно всего лишь подмножества или выборки из данных.

Если вы часто сталкиваетесь с работой с большими данными (в диапазоне от 10 до 100 Гб), мы рекомендуем обратить внимание на пакет `data.table`. Мы не рассматриваем его в этой книге, так как он использует другой интерфейс, отличный от tidyverse, и требует знакомства с другими концепциями и методами. Тем не менее его эффективность в работе с данными значительно выше, а прирост производительности оправдывает временные затраты на его изучение, особенно если вы работаете с крупными массивами данных.

Python, Julia и другие

В нашей книге мы не рассматриваем Python, Julia или какие-либо другие языки программирования, которые могут быть полезны в науке о данных. Это не означает, что мы считаем эти инструменты неэффективными. На самом деле многие команды по обработке данных используют комбинацию различных языков как R, так и Python. Однако мы убеждены, что лучше всего осваивать один инструмент за раз, и R является отличным выбором для начала.

Предварительные условия

Пытаясь сделать книгу максимально полезной, мы сделали несколько предположений о ваших знаниях. В целом вы должны обладать базовыми навыками работы с числами, и было бы неплохо, если бы у вас был некоторый опыт программирования. Если вы ранее никогда не программировали, рекомендуем вам ознакомиться с книгой Гаррета Гролемунда *Hands-On Programming with R*; издательство O'Reilly, которая может служить ценным дополнением к материалам данной книги.

Для того чтобы воспроизвести код, представленный в данной книге, вам понадобятся четыре компонента: R, RStudio, набор пакетов R под названием tidyverse и несколько дополнительных пакетов. Пакеты являются основными строительными блоками воспроизводимого кода в R. Они содержат повторно используемые функции, документацию, объясняющую, как их использовать, и примеры данных для работы. Установив и загрузив эти компоненты, вы сможете эффективно использовать материал, представленный в книге.

R

Чтобы загрузить R, перейдите в CRAN (Comprehensive R Archive Network; *https://cloud.r-project.org/*) — ресурс, который представляет собой обширную сеть архивов R. Новая основная версия R выпускается раз в год, а также публикуются два-три второстепенных обновления в год. Рекомендуется регулярно обновлять R, хотя процесс обновления может быть немного сложным, особенно при переходе на основные версии, когда требуется переустановка всех пакетов. Тем не менее не следует откладывать во избежание дальнейших трудностей. Для работы с материалами этой книги мы рекомендуем использовать R версии 4.2.0 или более поздних версий.

RStudio

RStudio — это интегрированная среда разработки (IDE) для программирования на R, которую можно загрузить с официальной страницы загрузки RStudio (*https://posit.co/download/rstudio-desktop/*). Обычно RStudio обновляется несколько раз в год и будет автоматически уведомлять вас о выходе новой версии, поэтому нет необходимости регулярно проверять обновления вручную. Рекомендуется обновляться, чтобы использовать все новые и улучшенные функции. Для работы с материалами этой книги убедитесь, что у вас установлена, по крайней мере, версия RStudio 2022.02.0.

Когда вы запустите RStudio (рис. I-2), вы увидите две ключевые области интерфейса: панель консоли и панель вывода. На данный момент все, что вам нужно, — это ввести код R на панели консоли и нажать Enter, чтобы запустить его. Вы узнаете больше в ходе изучения материалов нашей книги[1].

[1] Если вы хотите получить полное представление обо всех функциях RStudio, ознакомьтесь с Руководством пользователя RStudio.

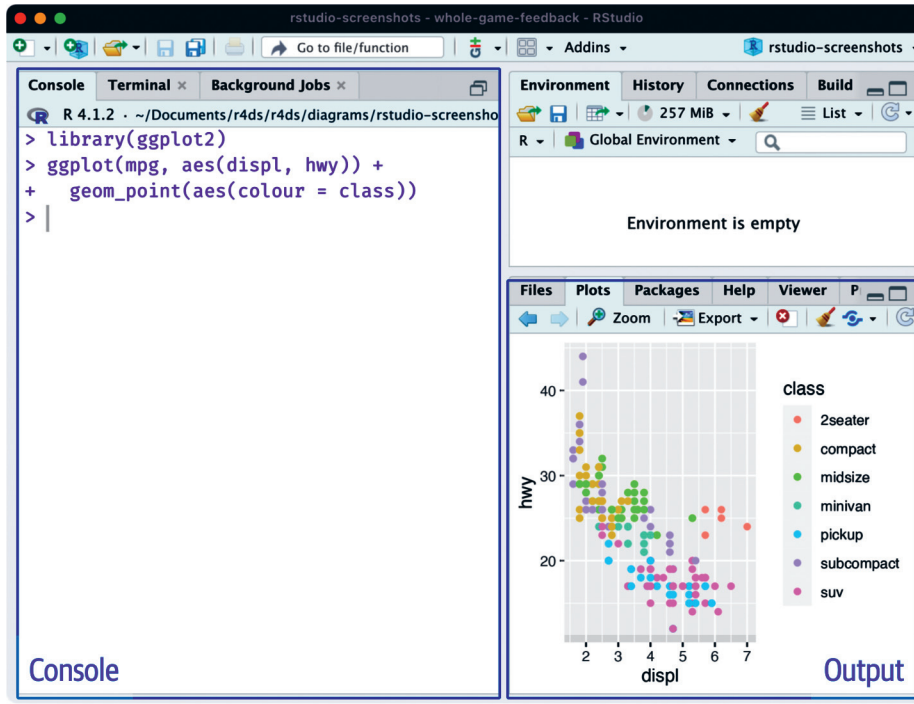

Рис. I-2. *Среда разработки RStudio имеет две ключевые области: вводите код R в консоли слева и просматривайте графики в панели вывода справа*

The Tidyverse

Вам потребуется установить несколько пакетов для R. Пакет R представляет собой совокупность функций, данных и документации, которые расширяют возможности базового R. Использование пакетов является ключевым аспектом эффективной работы с R. Большинство пакетов, которые вы будете изучать в этой книге, входят в состав так называемого набора пакетов, известного как tidyverse. Все пакеты в tidyverse реализуют единую философию программирования данных, что позволяет им легко взаимодействовать друг с другом.

Вы можете установить полный пакет tidyverse с помощью следующей команды:

```
install.packages("tidyverse")
```

На вашем компьютере введите эту строку кода в консоли и нажмите Enter, чтобы выполнить команду. R загрузит необходимые пакеты из CRAN и установит их на ваш компьютер.

Учтите, что вы не сможете использовать функции, объекты или справочные материалы пакета, пока он не будет загружен. После успешной установки пакета вы можете активировать его с помощью функции `library()` (*https://rdrr.io/r/base/library.html*), как показано ниже:

```
library(tidyverse)
#> --- Attaching core tidyverse packages --- tidyverse 2.0.0 ----
#> ✓ dplyr     1.1.0.9000  ✓ readr     2.1.4
#> ✓ forcats   1.0.0       ✓ stringr   1.5.0
#> ✓ ggplot2   3.4.1       ✓ tibble    3.1.8
#> ✓ lubridate 1.9.2       ✓ tidyr     1.3.0
#> ✓ purrr     1.0.1
#> --- Conflicts ------------------ tidyverse_conflicts() ----
#> ✖ dplyr::filter() masks stats::filter()
#> ✖ dplyr::lag()    masks stats::lag()
#> ℹ Use the conflicted package (<http://conflicted.r-lib.org/>)
#>   to force all conflicts to become errors
```

Это означает, что при загрузке tidyverse активируются девять основных пакетов: dplyr, forcats, ggplot2, lubridate, purrr, readr, stringr, tibble и tidyr. Эти пакеты образуют ядро tidyverse и будут использоваться практически в каждом вашем анализе данных.

Обратите внимание, что пакеты в tidyverse регулярно обновляются. Чтобы узнать, доступны ли новые версии пакетов, вы можете запустить функцию `tidyverse_update()`.

Другие пакеты

Существует также множество других пакетов, которые не являются частью tidyverse. Эти пакеты могут решать специфические задачи в различных областях или быть разработаны на основе других принципов. Это не делает их хуже или лучше; они просто отличаются по своему функционалу и подходу. Другими словами, вне tidyverse существует не просто хаотичная масса пакетов, а целые экосистемы, состоящие из взаимосвязанных инструментов. По мере работы над новыми проектами в области науки о данных с использованием R вы будете открывать для себя новые пакеты и новые способы анализа данных.

В этой книге мы будем использовать ряд пакетов, не входящих в состав tidyverse. Например, мы применим указанные ниже пакеты, поскольку они предоставляют интересные датасеты для практических задач в процессе изучения R:

```
install.packages(c("arrow", "babynames", "curl", "duckdb",
"gapminder", "ggrepel", "ggridges", "ggthemes", "hexbin",
"janitor", "Lahman", "leaflet", "maps", "nycflights13", "openxlsx",
"palmerpenguins", "repurrrsive", "tidymodels", "writexl"))
```

Мы будем использовать несколько других пакетов в рамках конкретных примеров и задач. Вам не нужно устанавливать эти пакеты прямо сейчас, но имейте в виду, что, если вы увидите ошибку, связанную с отсутствием какого-либо из этих пакетов, вам потребуется установить его:

```
library(ggrepel)
#> Error in library(ggrepel): there is no package called 'ggrepel'
```

Это означает, что для установки пакета вам необходимо выполнить команду `install.packages("ggrepel")`.

Запуск кода R

В предыдущем разделе было показано несколько примеров запуска кода R. Код в тексте книги выглядит следующим образом:

```
1 + 2
#> [1] 3
```

Если вы запустите этот же код в своей локальной консоли, он будет выглядеть так:

```
> 1 + 2
[1] 3
```

Существуют два основных отличия. Во-первых, в консоли появляется знак > при запуске фрагмента кода, который не отображается в книге. Во-вторых, результат выполнения в книге помечен как комментарий с помощью #>, тогда как в консоли он появляется сразу после введенного кода. Эти различия позволяют вам легко копировать код из электронной версии книги и вставлять его в консоль.

На протяжении всей книги мы придерживаемся последовательного набора соглашений для обозначения кода:

— Функции в тексте представлены кодовым шрифтом и сопровождаются круглыми скобками, например, `sum()` или `mean()`.
— Другие объекты R, такие как данные или аргументы функций, выделяются кодовым шрифтом без круглых скобок, например, `flights` или `x`.
— В некоторых случаях, чтобы было ясно, из какого пакета получен объект, мы указываем имя пакета, за которым следуют два двоеточия. Например, `dplyr::mutate()` или `nycflights13::flights`. Это также является корректным кодом на языке R.

Прочие условные обозначения

В данной книге применяются следующие типографские обозначения:

Курсив
: Используется для обозначения URL-адресов и адресов электронной почты.

`Моноширинный шрифт`
: Применяется в списках программ и абзацах для обозначения элементов программного кода, таких как имена переменных, функции, базы данных, типы данных, переменные среды, операторы, ключевые слова и имена файлов.

`Полужирным моноширинным шрифтом`
: Обозначает команды или текст, которые пользователь должен вводить точно и без изменений.

`Моноширинным курсивом`
: Указывает на текст, который необходимо заменить значениями, предоставленными пользователем, или значениями, зависящими от контекста.

Этот элемент обозначает общее примечание, которое может быть полезным для понимания материала или предоставления дополнительной информации.

Данный элемент сигнализирует о предупреждении или предостережении, которое важно учитывать, чтобы избежать возможных ошибок или проблем.