

*Посвящаю эту книгу
своему покойному дедушке
Владимиру Замашикову*

Содержание

Об авторе	13
О техническом рецензенте	14
Признательности	15
От издательства	17
Введение	18
Установка языка R и среды RStudio	20
Скачивание и установка языка R.....	20
Скачивание и установка среды RStudio.....	21
Конфигурирование среды RStudio	21
Установка пакетов.....	21
Часть I. РАБОТА С ДАННЫМИ	23
Глава 1. Манипулирование данными	24
1.1. Основы	24
1.1.1. Типы данных	26
1.2. Скачивание данных	28
1.2.1. Пример данных	28
1.3. Простейшие методы управления данными с помощью подпакета dplyr.....	30
1.3.1. Функция select()	30
1.3.2. Функция filter().....	31
1.3.3. Функция arrange()	32
1.3.4. Функция mutate().....	33
1.3.5. Функция case_match().....	34
1.3.6. Функция summarize()	34
1.3.7. Функция group_by()	34
1.3.8. Функция ungroup().....	35
1.3.9. Аргумент .by.....	36
1.3.10. Функция rowwise()	37
1.3.11. Функция count()	38
1.3.12. Функция rename()	38
1.3.13. Функция row_number()	39
1.3.14. Функция skim().....	39
1.4. Обследование даты и времени с помощью подпакета lubridate	40

1.4.1. Функции <code>ymd()</code> , <code>md()</code> , <code>hms()</code> , <code>ymd_hms()</code>	40
1.4.2. Функции <code>year()</code> , <code>month()</code> , <code>day()</code>	41
1.5. Краткий итог.....	41
Глава 2. Упорядоченные данные	42
2.1. Пример.....	44
2.2. Подпакет <code>tidyr</code>	45
2.2.1. Функция <code>pivot_longer()</code>	45
2.2.2. Функция <code>pivot_wider()</code>	46
2.2.3. Функции <code>separate()</code> и <code>unite()</code>	47
2.3. Функции <code>tibble()</code> и <code>tribble()</code>	49
2.4. Пакет <code>janitor</code> : очистка данных.....	50
2.4.1. Функция <code>clean_names()</code>	50
2.4.2. Функция <code>remove_empty()</code>	51
2.4.3. Функция <code>remove_constant()</code>	51
2.4.4. Функции <code>convert_to_date()</code> и <code>convert_to_datetime()</code>	52
2.4.5. Функция <code>row_to_names()</code>	52
2.5. Краткий итог.....	53
Глава 3. Реляционные данные	54
3.1. Типы отношений.....	54
3.1.1. Отношение «один к одному» (1:1).....	54
3.1.2. Отношение «один ко многим» (1:M).....	55
3.1.3. Отношение «многие ко многим» (M:N).....	55
3.2. Понятие ключей.....	56
3.3. Типы соединений.....	56
3.3.1. Внешние соединения.....	57
3.3.2. Фильтрация соединений.....	60
3.4. Визуализация отношений.....	61
3.5. Краткий итог.....	63
Глава 4. Валидация данных	64
4.1. Инспекция данных вручную.....	64
4.2. Улаживание проблем с данными.....	65
4.2.1. Подтверждение истинности своих условий.....	65
4.2.2. Точная валидация с помощью пакета <code>pointblank</code>	67
4.3. Краткий итог.....	69
Глава 5. Восполнение пропущенных данных	70
5.1. Типы пропущенных данных.....	70
5.2. Работа с пропущенными данными.....	71
5.2.1. Явная обработка пропущенных данных с помощью функции <code>complete()</code>	72
5.2.2. Простые методы восполнения.....	74
5.2.3. Наилучший сценарий из наихудших и наихудший сценарий из наилучших.....	79

5.2.4. Множественные восполнения	80
5.3. Краткий итог	83
5.3.1. Таблица методов восполнения пропущенных данных.....	84
Часть II. ВОСПРОИЗВОДИМЫЕ НАУЧНЫЕ ИЗЫСКАНИЯ.....	86
Глава 6. Воспроизводимое исследование	87
6.1. Грамотное программирование	88
6.2. Краткий итог	90
Глава 7. Воспроизводимая среда	91
7.1. Пакет <code>renv</code>	92
7.1.1. Рабочий поток.....	92
7.2. Вычислительные среды	93
7.3. Краткий итог	94
Глава 8. Введение в командную оболочку.....	95
8.1. Освоение основных команд	95
8.2. Начало работы с Nano.....	96
Глава 9. Контроль версий с помощью Git и GitHub	98
9.1. Система контроля версий Git и веб-платформа GitHub.....	99
9.2. Основы	100
9.3. Руководство по использованию файла <code>.gitignore</code>	102
9.3.1. Перечисление файлов, которые следует игнорировать.....	103
9.3.2. Файл <code>.gitignore</code> в других программах.....	104
9.4. Краткий итог	104
Глава 10. Стилизация и статический анализ исходного кода.....	105
10.1. Руководство <code>tidyverse</code> по стилю	106
10.1.1. Пробелы и отступы	106
10.1.2. Правила именования	106
10.1.3. Фигурные скобки	107
10.1.4. Комментарии.....	107
10.1.5. Длинные функции	108
10.2. Форматировщик.....	109
10.3. Статический анализатор исходного кода.....	109
10.4. Краткий итог	110
Глава 11. Модульный исходный код	111
11.1. Реиспользование функций	111
11.2. Разбивка исходного кода	112
1.3. Упаковка исходного кода в модули	114
11.4. Краткий итог.....	115

Часть III. ОБЗОР НАУЧНЫХ ПУБЛИКАЦИЙ И НАПИСАНИЕ СТАТЬИ	116
Глава 12. Обзор научных публикаций	117
12.1. Поиск.....	117
12.2. Управление справочными материалами.....	118
12.3. Чтение статей	119
12.4. Ведение заметок.....	120
12.5. Краткий итог.....	121
Глава 13. Написание научной статьи	122
13.1. WYSIWYG.....	122
13.2. Языки разметки.....	123
13.2.1. HTML.....	123
13.2.2. LaTeX.....	123
13.2.3. Упрощенная разметка Markdown	124
13.2.4. YAML	124
13.2.5. Pandoc.....	125
13.3. Quarto.....	125
13.3.1. Ваш первый документ.....	126
13.4. Краткий итог.....	129
Глава 14. Макетирование и ссылки на справочные материалы	130
14.1. Пакет knitr	130
14.2. Блоки div	131
14.3. Диаграммы.....	132
14.4. Цитирование.....	133
14.5. Краткий итог.....	134
Глава 15. Совместная работа и шаблонное форматирование документов	135
15.1. Оптимизация совместной работы с помощью инструмента trackdown	135
15.2. Шаблонное форматирование документов	137
15.3. Краткий итог.....	138
Часть IV. СБОР ДАННЫХ	139
Глава 16. Суммарная ошибка опроса	140
16.1. Репрезентативность – люди, которых вы спрашиваете	143
16.1.1. Взятие выборки	145
16.1.2. От чего зависят ответы	146
16.2. Разработка вопросов	148
16.2.1. Ответы на вопросы.....	148
16.2.2. Рекомендации по эффективной разработке вопросов	150

16.2.3. Влияние порядка следования вопросов	150
16.2.4. В опросе говорится: «Бекон»!.....	151
16.2.5. Типы вопросов	152
16.2.6. Шкалы Лайкерта.....	153
16.2.7. Варианты «не знаю» и «затрудняюсь ответить».....	154
16.2.8. Продолжительность опроса.....	155
16.2.9. Приглашение к опросу.....	156
16.2.10. Итеративное составление вопросов	157
16.3. Инструменты проведения опросов	157
16.3.1. Методы проведения физических опросов	157
16.3.2. Платформы для проведения цифровых опросов.....	158
16.3.3. Платформы для рекрутирования участников	159
16.4. Краткий итог.....	160
Глава 17. Документирование.....	161
17.1. Принципы документирования	161
17.1.1. Иерархическая структура документации	163
17.2. Физическая и электронная документация	165
17.3. Организация данных в электронных таблицах	166
17.3.1. Организация электронных таблиц.....	167
17.3.2. Ввод данных.....	168
17.4. Администрирование	170
17.5. Финансовая отчетность в научных исследованиях.....	171
17.6. Коммуникация	173
17.7. Краткий итог.....	173
Глава 18. Интерфейсы прикладного программирования.....	175
18.1. Основы API.....	175
18.2. Использование API на языке R.....	178
18.3. API системы Qualtrics	180
18.4. Интеграция служб Google со средой языка R.....	182
18.5. API OpenAI.....	184
18.6. Краткий итог.....	186
Часть V. ПРЕЗЕНТАЦИЯ ДАННЫХ.....	187
Глава 19. Основы визуализации данных	188
19.1. Восприятие.....	188
19.1.1. Предвнимательная обработка	189
19.2. Визуальное кодирование	196
19.2.1. Оценивание графиков.....	197
19.3. Краткий итог.....	199
Глава 20. Визуализация данных	200
20.1. Разведывательный и объяснительный типы	200

20.2. Грамматика графики	202
20.2.1. Графические интерфейсы для пакета ggplot2	208
20.3. Интерактивные графики.....	210
20.3.1. HTML-виджеты.....	211
20.4. Краткий итог.....	213

Глава 22. Подходящий график для работы

21.1. Сравнение категорий	214
21.1.1. Леденцовый график	215
21.1.2. Пулевой график.....	215
21.2. Распределение	216
21.2.1. График плотности.....	217
21.2.2. Многоугольник частот	218
21.2.3. Коробчатый график	219
21.2.4. Скрипичный график.....	219
21.2.5. Ульевый график.....	220
21.2.6. График дождевых облаков.....	221
21.2.7. Графики по краям	222
21.3. Пропорции	223
21.3.1. Многослойный столбчатый график	223
21.3.2. Круговой график	224
21.3.3. Вафельный график	226
21.3.4. Древовидные карты	226
21.4. Корреляция	227
21.4.1. Диаграмма рассеяния.....	227
21.4.2. Коррелограммы.....	227
21.5. Изменение во временной динамике.....	229
21.5.1. Линейный график.....	229
21.5.2. Водопадный график	229
21.6. Краткий итог.....	230

Глава 22. Цветовые данные.....

22.1. Какие цвета выбрать	231
22.1.1. Взаимодополняющая гармония с положительной/отрицательной коннотацией	232
22.1.2. Почти взаимодополняющая гармония для выделения двух рядов, один из которых находится в центре внимания	233
22.1.3. Аналогичная/триадическая гармония для выделения трех рядов	234
22.1.4. Выделение одного ряда на фоне двух соотнесенных рядов	235
22.1.5. Схема на основе аналогичной взаимодополняющей гармонии для одного главного ряда и трех вторичных ему рядов	236
22.1.6. Схема на основе двойной взаимодополняющей гармонии для двух пар, одна из которых является доминирующей	237
22.1.7. Прямоугольная или квадратная схема на основе взаимодополняющей гармонии для четырех рядов с одинаковым акцентом	237

22.1.8. Схема последовательных цветов.....	238
22.1.9. Схемы с расходящимися цветами.....	239
22.1.10. Предварительно созданные шкалы	239
22.2. Цветовые системы	240
22.2.1. Внимание: цветовые карты могут увеличивать риск смерти!	244
22.2.2. Итак, что же следует использовать?.....	245
22.3. Краткий итог.....	246
Глава 23. Создание таблиц	247
23.1. Таблицы gt.....	247
23.1.1. Подготовка данных.....	248
23.2. Таблицы DT	255
23.3. Краткий итог	257
Эпилог.....	258
Справочные материалы.....	259
Предметный указатель.....	263

Об авторе



Никита Ткаченко возглавляет консалтинговое агентство Evalyn, специализирующееся на аудите обслуживания клиентов на основе искусственного интеллекта (ИИ) и технологических решениях по анализу данных. Он помогает организациям любого размера использовать ИИ и данные, чтобы оптимизировать процессы принятия решений, систематизировать операции и улучшать качество обслуживания покупателей. Обладая солидным опы-

том в области изысканий и аналитики, Никита также преподает курсы по инструментам исследования, руководит студентами и проводит академические исследования в Университете Сан-Франциско.

О техническом рецензенте



Сиджо Валаяккад Маникандан – опытный специалист в области искусственного интеллекта и науки о данных. Обладает обширным опытом управления крупномасштабными проектами в области науки о данных для корпораций из списка Fortune 500. Сиджо получил степень магистра наук в области деловой аналитики в знаменитом Техасском университете в Остине и степень бакалавра инженерных наук в Институте технологии и науки Бирлы в Пилани, Индия.

Для достижения исключительных результатов он сочетает свои академические знания и практический опыт.

Сиджо внес значительный вклад в развитие науки о данных и научных исследований благодаря не только своим профессиональным и академическим достижениям, но и своей преданности сообществу. Он входит в состав нескольких авторитетных организаций, таких как Американская статистическая ассоциация, проводит независимые исследования, а также активно рецензирует научные и профессиональные монографии, был наставником молодых исследователей данных и руководил молодыми стартапами. Более того, его опыт позволил ему войти в состав жюри нескольких престижных премий, включая Webby Awards.

Будучи целеустремленным исследователем и идейным лидером, Сиджо продолжает оказывать глубокое влияние на сферу науки о данных, вдохновляя новое поколение исследователей данных. Его вклад продвинул междисциплинарную область науки о данных и научных исследований вперед и помог определить будущее этой быстро развивающейся отрасли, что сделало его бесценным активом для сообщества и провидцем в своей сфере.

Признательности

Посвящаю эту книгу своему покойному дедушке Владимиру Замашикову.

Выражаю свою глубочайшую благодарность Томасу Вайнанди, который убедил меня превратить черновой вариант в опубликованную книгу и направлял меня на протяжении всего процесса. Без него эта книга была бы невозможна.

Я глубоко признателен Парсу Рахими за его неизменную поддержку, тщательный анализ всех моих глав, правок и идей, а также за предоставленные художественные работы с элементами фантастики.

Также выражаю благодарность Алессандре Кассар и Майклу Джонасу за их поддержку на протяжении всего моего академического пути. Особая благодарность Питеру Лоренцену, Керу Гиббсу и Джесси Анттила-Хьюзу за их неоценимое наставничество и поддержку.

Я признателен Шивани Шукле, Брюсу Вайдику, Мише Цыгману, Марио Лиму, Стиву Треттелу, Мехмету Эмре, Эндрю Хоббсу, Арману Хачияну, Робизону Хубулашвили и Конраду Пошу за рецензирование глав и предоставленную поддержку. Кроме того, я благодарен преподавательскому составу, библиотекарям, сотрудникам писательского центра и Центра деловых исследований и инноваций (CBSI) Университета Сан-Франциско за их наставничество.

Гордон Гетти заслуживает благодарности за то, что вдохновил меня более уверенно писать о своих навыках.

В процессе публикации мне помогал коллектив издательства APress, в особенности Шауль Элсон, редактор-заказчик, Лаура Берендсон, редактор-разработчик, и Гриффин Винклер, ассистент редактора. Мой технический рецензент Сиджо Маникандан предоставил комментарии, которые значительно обогатили книгу дополнительными примерами и пояснениями. Выражаю особую благодарность всем сотрудникам издательства APress, благодаря чьим усилиям эта книга стала реальностью.

Выражаю глубокую признательность Джону Четвинду, Джейку Консгроуву, Тимофею Лопухову и Александру Рому за их поддержку, любопытство и доверие к моей работе.

Должен поблагодарить свою семью, включая мою маму Екатерину Ткаченко, чья вера в меня вдохновила меня на запуск этого проекта, моего отца Антона Ткаченко, приемную мать Дарью Ткаченко, а также бабушку и дедушку Валентину и Александра Ткаченко и Раису Замашикову.

Особая благодарность Анастасии Терновской за ее непоколебимую веру в меня.

Я очень признателен Марии Аксутенко за создание иллюстрации «кошка-исследователь».

Особая благодарность также коллективу Posit за их неоценимый вклад и поддержку сообщества.

Наконец, выражаю искреннюю признательность сообществу разработчиков открытого исходного кода и всем тем, кто щедро делился своими знаниями в интернете. Вы сделали это реальным, и я рад возможности принять участие в обсуждении. Надеюсь, что эта книга окажется полезной на вашем пути к овладению методами научных изысканий.

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.

Введение

Эта книга родилась из моих разочарований и опыта в сфере высшего образования и профессиональной деятельности. Она была написана на основе заметок и материалов весеннего курса 2023 года по разработке опросов под влиянием восторженных откликов и содержательных вопросов моих студентов.

Молодые специалисты в области исследования данных, как правило, на своих занятиях знакомятся с моделями, экспериментами и теориями и часто возвращаются к этим знаниям позже. Однако для проведения высококачественных изысканий и анализа требуется более глубокое понимание инструментов и того, «как» это делается, а не только того, «что» делается и «почему». Эти знания часто выходят за пределы того, чему учат в рамках стандартной учебной программы. В данной книге я стремлюсь преодолеть этот пробел, помогая вам перейти от знания того, что вы хотите делать, к пониманию того, как это делать. В свои главы я вложил сотни часов разочарования, так что вам не придется проходить этот путь самостоятельно.

Эта книга не является исчерпывающим руководством; если вы ищете именно его, то, возможно, вам лучше поискать в другом месте. Данная же книга может служить картой, очерчивающей необходимые инструменты и темы для вашего исследовательского путешествия. Ее цель состоит в том, чтобы развить вашу интуицию и подсказать места, где найти более подробную информацию. Главы намеренно сделаны краткими, а материал излагается по существу. Они призваны раскрывать и просвещать, а не утомлять. Вы узнаете об эффективном управлении данными, проведении воспроизводимых исследований, составлении обзора научных публикаций и методах написания текстов, а также об эффективной визуализации данных.

Эта книга, изначально написанная под впечатлением от моего обучения в аспирантуре по экономике, представляет ценность для всех дисциплин. В ней содержатся важные сведения для всех, кто работает с данными, от гуманитарных наук до анализа данных и естественных наук. Независимо от того, оттачиваете ли вы свой опыт в области анализа данных или являетесь в ней новичком, настоящая книга обещает предложить вам нечто ценное.

Приведенные в книге примеры в основном написаны на языке R, что предполагает базовое понимание языка, но это не обязательно. Некоторые главы, в особенности те, которые посвящены теории, вообще не требуют знаний

в области программирования. Материал оказался полезен широкому кругу читателей, включая веб-разработчиков, математиков, аналитиков данных и экономистов. Книга организована таким образом, чтобы быть всеохватывающей и давать информацию независимо от вашего уровня владения программированием или профессиональной подготовки.

Ее структура позволяет выбирать гибкие пути чтения; можно просматривать главы последовательно, чтобы усваивать материал систематически либо переходить непосредственно к наиболее актуальным для вас темам.

Установка языка R и среды RStudio

Добро пожаловать в захватывающий мир анализа данных на языке R, мастерски выполненного специально для статистического анализа и визуализации данных. Удобный синтаксис и воспроизводимость данного языка делают его идеальным выбором как для новичков, так и для профессионалов. Однако прежде чем приступить к разведывательному анализу данных и моделированию, важно провести различие между языком программирования R и интегрированной средой разработки (IDE) RStudio, которая расширяет функциональные средства языка R.

Скачивание и установка языка R

Язык R поддерживается и распространяется через всеобъемлющую сеть архивов R (Comprehensive R Archive Network, аббр. CRAN), которая обеспечивает доступ к самой свежей версии и ресурсам.

Для пользователей macOS:

1. Перейдите на веб-сайт CRAN¹.
2. Кликните по **Download R for macOS** (Скачать R для macOS).
3. Выберите подходящую версию:
 - для Apple Silicon (например, M1, M2) надо скачать версию, в названии которой указано «-arm64» (например, R-4.2.2-arm64.pkg);
 - для компьютеров Mac на базе Intel надо выбрать версию без «-arm64» (например, R-4.2.2.pkg).
4. Следуйте инструкциям мастера установки. Обычно достаточно стандартных настроек.

Для пользователей Windows:

1. Перейдите на веб-сайт CRAN.
2. Выберите **Download R for Windows** (Скачать R для Windows).
3. Выберите **base** (базовая), а затем первую ссылку в верхней части страницы (например, Download R-4.2.2 for Windows).
4. Установщик проведет вас по всему процессу. В целях более эффективной установки придерживайтесь стандартных настроек.

¹ См. <https://cran.r-project.org/>.

Конец ознакомительного фрагмента.

Приобрести книгу можно

в интернет-магазине

«Электронный универс»

e-Univers.ru