

УДК 004.438R
ББК 32.973.22
Л76

Дж. Д. Лонг и Пол Титор

Л76 Р. Книга рецептов: Проверенные рецепты для статистики, анализа и визуализации данных / пер. с англ. Д. А. Беликова. – М.: ДМК Пресс, 2020. – 510 с.: ил.

ISBN 978-5-97060-835-7

Язык R – мощный инструмент статистического программирования, десятки тысяч людей ежедневно используют его для проведения серьезного статистического анализа. Но не все задачи, даже простые, удастся быстро решить с его помощью, если не знать определенных тонкостей.

Эта книга предлагает практические советы по решению разнообразных задач с подробным разбором каждой из них. От основных задач автор переходит к вводу и выводу, общей статистике, графике, линейной регрессии – любая значительная работа с R подразумевает знакомство с большинством этих областей или с ними всеми.

Издание пригодится для разработчиков на R с разным уровнем подготовки – от новичков до уверенных пользователей, желающих расширить свой кругозор.

УДК 004.438R
ББК 32.973.22

Authorized Russian translation of the English edition of R Cookbook 2E ISBN 9781492040682
© 2019 J.D. Long and Paul Teetor.

This translation is published and sold by permission of O'Reilly Media, Inc., which owns or controls all rights to publish and sell the same.

Все права защищены. Любая часть этой книги не может быть воспроизведена в какой бы то ни было форме и какими бы то ни было средствами без письменного разрешения владельцев авторских прав.

ISBN 978-1-492-04068-2 (англ.)
ISBN 978-5-97060-835-7 (рус.)

© 2019 J.D. Long and Paul Teetor
© Оформление, издание, перевод, ДМК Пресс, 2020

Оглавление

Предисловие редактора	12
Добро пожаловать в Книгу рецептов R.....	13
Благодарности.....	18

Глава 1. Начало работы и получение справочной информации	19
1.1. Загрузка и установка R.....	20
1.2. Установка RStudio.....	22
1.3. Запуск RStudio	23
1.4. Ввод команд.....	24
1.5. Выход из RStudio	26
1.6. Прерывание R.....	28
1.7. Просмотр прилагаемой документации	28
1.8. Получение справки по функции	30
1.9. Поиск в прилагаемой документации	31
1.10. Получение справки по пакету	33
1.11. Поиск справки в интернете	34
1.12. Поиск соответствующих функций и пакетов	36
1.13. Поиск в списках рассылки	37
1.14. Отправка вопросов в Stack Overflow или в другое место в сообществе	38

Глава 2. Немного основ.....	42
2.1. Вывод на экран.....	42
2.2. Установка переменных	44
2.3. Перечисление переменных	45
2.4. Удаление переменных	46
2.5. Создание вектора	48
2.6. Базовая статистика	49
2.7. Создание последовательностей	51
2.8. Сравнение векторов	53
2.9. Выбор элементов вектора.....	54
2.10. Векторная арифметика.....	57
2.11. Разбираемся с приоритетом оператора	58
2.12. Меньше печатать и больше делать	60
2.13. Создание конвейера вызовов функций.....	61
2.14. Как избежать некоторых распространенных ошибок	64

Глава 3. Навигация по программному обеспечению	69
3.1. Получение и настройка рабочего каталога	69
3.2. Создание нового проекта RStudio.....	70
3.3. Сохранение своего рабочего пространства	72
3.4. Просмотр истории команд	73
3.5. Сохранение результата предыдущей команды	74

3.6. Отображение загруженных пакетов через путь поиска	75
3.7. Просмотр списка установленных пакетов	76
3.8. Доступ к функциям в пакете	77
3.9. Доступ к встроенным наборам данных	78
3.10. Установка пакетов из CRAN	80
3.11. Установка пакета из GitHub	81
3.12. Установка или изменение зеркала CRAN по умолчанию	82
3.13. Запуск сценария	84
3.14. Запуск пакетного сценария	85
3.15. Поиск домашнего каталога R	87
3.16. Настройка запуска R	88
3.17. Использование R и RStudio в облаке	91

Глава 4. Ввод и вывод 93

4.1. Ввод данных с клавиатуры	93
4.2. Вывод меньшего числа цифр (или большего)	94
4.3. Перенаправление вывода в файл	96
4.4. Список файлов	97
4.5. Не удастся открыть файл в Windows – что делать?	99
4.6. Чтение записей фиксированной ширины	100
4.7. Чтение файлов табличных данных	102
4.8. Чтение из файлов CSV	105
4.9. Запись в файлы CSV	107
4.10. Чтение табличных или CSV-данных из интернета	109
4.11. Чтение данных из Excel	110
4.12. Запись таблицы данных в Excel	111
4.13. Чтение данных из файла SAS	113
4.14. Чтение данных из таблиц HTML	115
4.15. Чтение файлов со сложной структурой	117
4.16. Чтение из баз данных MySQL	121
4.17. Доступ к базе данных с помощью dbplyr	124
4.18. Сохранение и транспортировка объектов	125

Глава 5. Структуры данных 128

5.1. Добавление данных в вектор	135
5.2. Вставка данных в вектор	136
5.3. Правило повторного использования	137
5.4. Создание фактора (категориальной переменной)	139
5.5. Объединение нескольких векторов в один вектор и фактор	140
5.6. Создание списка	141
5.7. Выбор элементов по месту в списке	143
5.8. Выбор элементов списка по имени	144
5.9. Создание ассоциативного списка «имя/значение»	146
5.10. Удаление элемента из списка	147
5.11. Преобразование списка в вектор	148
5.12. Удаление элементов NULL из списка	149
5.13. Удаление элементов списка с использованием условия	150
5.14. Инициализация матрицы	152
5.15. Операции с матрицами	153
5.16. Задание описательных имен для строк и столбцов матрицы	154

5.17. Выбор одной строки или столбца из матрицы.....	155
5.18. Инициализация таблицы данных из данных столбца.....	156
5.19. Инициализация таблицы данных из данных строки	157
5.20. Добавление строк в таблицу данных	160
5.21. Выбор столбцов таблицы по их месту в таблице данных.....	162
5.22. Выбор столбцов таблицы данных по имени	165
5.23. Изменение имен столбцов таблицы данных	167
5.24. Удаление значений NA из таблицы данных	168
5.25. Исключение столбцов по имени	169
5.26. Объединение двух таблиц данных.....	170
5.27. Объединение таблиц данных по общему столбцу	171
5.28. Приведение атомарных типов данных.....	173
5.29. Приведение структурированных типов данных	174

Глава 6. Преобразование данных 177

6.1. Применение функции ко всем элементам списка	177
6.2. Применение функции к каждой строке таблицы данных.....	179
6.3. Применение функции к каждой строке матрицы	180
6.4. Применение функции к каждому столбцу	182
6.5. Применение скалярной функции к векторам или спискам	183
6.6. Применение функции к группам данных	186
6.7. Создание нового столбца по условию	187

Глава 7. Строки и даты 189

7.1. Получение длины строки.....	191
7.2. Конкатенация строк	191
7.3. Извлечение подстрок	192
7.4. Разбиение строки по разделителю	193
7.5. Замена подстрок.....	195
7.6. Генерация всех попарных комбинаций строк.....	195
7.7. Получение текущей даты	197
7.8. Преобразование строки в дату	197
7.9. Преобразование даты в строку.....	198
7.10. Преобразование года, месяца и дня в объект Date	199
7.11. Получение даты по юлианскому календарю	200
7.12. Извлечение частей даты	201
7.13. Создание последовательности дат	202

Глава 8. Вероятность 204

8.1. Подсчет количества комбинаций.....	206
8.2. Генерация комбинаций	206
8.3. Генерация случайных чисел.....	207
8.4. Генерация воспроизводимых случайных чисел	209
8.5. Генерация случайной выборки	210
8.6. Генерация случайных последовательностей.....	212
8.7. Случайная перестановка вектора.....	213
8.8. Расчет вероятностей для дискретных распределений	213
8.9. Расчет вероятностей для непрерывных распределений.....	215
8.10. Преобразование вероятностей в квантили	216
8.11. Построение графика функции плотности.....	217

Глава 9. Общая статистика	222
9.1. Получение сводки данных	224
9.2. Расчет относительных частот	226
9.3. Представление факторов в виде таблицы и создание таблиц сопряженности	227
9.4. Проверка категориальных переменных на независимость	228
9.5. Расчет квантилей (и квартилей) набора данных	229
9.6. Инвертирование квантиля	230
9.7. Преобразование данных в z-оценки	230
9.8. Проверка среднего значения выборки (t-критерий)	231
9.9. Формирование доверительного интервала для среднего значения	233
9.10. Формирование доверительного интервала для медианы	234
9.11. Тестирование доли выборки	235
9.12. Формирование доверительного интервала для доли	236
9.13. Проверка на нормальность	237
9.14. Тест последовательностей	238
9.15. Сравнение средних значений двух выборок	239
9.16. Непараметрическое сравнение местоположений двух выборок	241
9.17. Проверка значимости корреляции	242
9.18. Проверка групп на предмет наличия равных пропорций	244
9.19. Парные сравнения между средними значениями групп	245
9.20. Проверка двух выборок, чтобы определить, принадлежат ли они одному закону распределения	246

Глава 10. Графики	248
10.1. Создание точечной диаграммы	252
10.2. Добавление заголовка и меток	253
10.3. Добавление (или удаление) координатной сетки	254
10.4. Применение темы к графику ggplot	258
10.5. Создание точечной диаграммы	261
10.6. Добавление (или удаление) условных обозначений	263
10.7. Построение регрессионной линии точечной диаграммы	267
10.8. Построение точечных диаграмм	270
10.9. Создание по одной точечной диаграмме	272
10.10. Создание гистограммы	274
10.11. Добавление доверительных интервалов в гистограмму	276
10.12. Раскраска гистограммы	279
10.13. Построение линии из точек x и y	281
10.14. Изменение типа, ширины или цвета линии	282
10.15. Построение нескольких наборов данных	285
10.16. Добавление вертикальных или горизонтальных линий	286
10.17. Создание диаграммы размаха	288
10.18. Создание диаграммы размаха для каждого уровня фактора	290
10.19. Создание гистограммы	292
10.20. Добавление оценки плотности к гистограмме	293
10.21. Создание графика квантиль-квантиль	295
10.22. Создание других графиков квантиль-квантиль	297
10.23. Построение переменной в нескольких цветах	300
10.24. График функции	302
10.25. Отображение нескольких графиков на одной странице	304

Глава 11. Линейная регрессия и дисперсионный анализ	309
11.1. Простая линейная регрессия	311
11.2. Множественная линейная регрессия	313
11.3. Получение регрессионной статистики	314
11.4. Общая информация о регрессии	318
11.5. Линейная регрессия без свободного члена	321
11.6. Регрессия только тех переменных, которые сильно коррелируют с вашей зависимой переменной	322
11.7. Линейная регрессия с эффектами взаимодействия	325
11.8. Выбор наиболее подходящих переменных регрессии	327
11.9. Регрессия для подмножества данных	331
11.10. Использование выражения в формуле регрессии	333
11.11. Полиномиальная регрессия	334
11.12. Регрессия на преобразованных данных	335
11.13. Поиск наиболее подходящего степенного преобразования (тест Бокса–Кокса)	337
11.14. Формирование доверительных интервалов для коэффициентов регрессии	342
11.15. Построение невязок регрессии	342
11.16. Диагностика линейной регрессии	344
11.17. Обнаружение влиятельных наблюдений	348
11.18. Тестирование невязок на наличие автокорреляции (критерий Дарбина–Уотсона)	349
11.19. Предсказываем новые значения	351
11.20. Формирование интервалов предсказаний	352
11.21. Однофакторный дисперсионный анализ	352
11.22. Создание диаграммы взаимодействия	354
11.23. Находим различия между средними значениями групп	356
11.24. Устойчивый дисперсионный анализ (критерий Краскела–Уоллиса)	358
11.25. Сравнение моделей с использованием функции anova	360

Глава 12. Полезные хитрости	362
12.1. Просмотр данных	362
12.2. Вывод на экран результата присваивания	364
12.3. Суммирование строк и столбцов	365
12.4. Вывод данных в столбцах	366
12.5. Объединение данных	367
12.6. Поиск положения определенного значения	368
12.7. Выбор каждого n -го элемента вектора	368
12.8. Поиск минимумов или максимумов	369
12.9. Генерация всех комбинаций нескольких переменных	371
12.10. Преобразование таблицы данных	372
12.11. Сортировка таблицы данных	373
12.12. Удаление атрибутов из переменной	374
12.13. Раскрываем структуру объекта	375
12.14. Определяем время выполнения кода	378
12.15. Избавляемся от предупреждений и сообщений об ошибках	379
12.16. Извлечение аргументов функции из списка	380
12.17. Определение собственных бинарных операторов	382
12.18. Избавляемся от сообщения о запуске	383
12.19. Получение и настройка переменных среды	384

12.20. Разбиение кода на секции	385
12.21. Локальная параллелизация выполнения кода	386
12.22. Удалённая параллелизация выполнения кода	388

Глава 13. За пределами основных цифр и статистики 392

13.1. Минимизация или максимизация однопараметрической функции	392
13.2. Минимизация или максимизация многопараметрической функции	393
13.4. Метод главных компонент	396
13.5. Простая ортогональная регрессия	397
13.6. Поиск кластеров в данных	399
13.7. Прогнозирование бинарной переменной (логистическая регрессия)	402
13.8. Бутстрэппинг	404
13.9. Факторный анализ	406

Глава 14. Анализ временных рядов 411

14.1. Представление данных временного ряда	412
14.2. Построение данных временных рядов	415
14.3. Извлечение самых старых или самых последних наблюдений	417
14.4. Выбор элементов из временного ряда	419
14.5. Объединение нескольких временных рядов	421
14.6. Заполнение временного ряда	423
14.7. Смещение временного ряда	425
14.8. Вычисление последовательных различий	427
14.9. Выполнение расчетов по временным рядам	428
14.10. Вычисление скользящей средней	429
14.11. Применение функции по календарному периоду	431
14.12. Применение функции <code>rollapply</code>	433
14.13. Построение функции автокорреляции	434
14.14. Тестирование временного ряда на наличие автокорреляций	436
14.15. Построение функции частичной автокорреляции	437
14.16. Поиск корреляций с временным лагом между двумя временными рядами	439
14.17. Удаление тренда из временного ряда	440
14.18. Подгонка модели ARIMA	443
14.19. Удаление незначимых коэффициентов ARIMA	446
14.20. Выполнение диагностики для модели ARIMA	448
14.21. Прогнозирование по модели ARIMA	450
14.22. Построение прогноза	451
14.23. Тестирование на наличие возвращения к среднему	452
14.24. Сглаживание временного ряда	455

Глава 15. Простое программирование 457

15.1. Выбор из двух альтернатив: <code>if/else</code>	458
15.2. Итерация с помощью цикла	460
15.3. Определение функции	461
15.4. Создание локальной переменной	462
15.5. Выбор из нескольких альтернатив: функция <code>switch</code>	463
15.6. Определение значений по умолчанию для параметров функции	464
15.7. Подать сигнал с помощью сообщения об ошибке	465

15.8. Защита от ошибок	466
15.9. Создание анонимной функции	467
15.10. Создание коллекции многократно используемых функций	468
15.11. Автоматическое форматирование кода	469

Глава 16. R Markdown и публикации 471

16.1. Создание нового документа	472
16.2. Добавление заголовка, автора или даты	474
16.3. Форматирование текста документа	476
16.4. Вставка заголовков документов	476
16.5. Вставка списка	477
16.6. Вывод результатов из кода R	478
16.7. Контролируем, какой код и результаты отображаются	480
16.8. Вставка графика	481
16.9. Вставка таблицы	484
16.10. Вставка таблицы данных	485
16.11. Вставка математических уравнений	488
16.12. Генерация вывода HTML	488
16.13. Генерация вывода в формате PDF	490
16.14. Генерация вывода в формате Microsoft Word	492
16.15. Генерация выходных данных презентации	498
16.16. Создание параметризованного отчета	500
16.17. Организация рабочего процесса в R Markdown	503

Об авторах 506

Колофон 506

Предметный указатель 507

От редакции 508