Краткое оглавление

Оглавление

Предисловие 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 – возможно, ошибку в основном тексте или программном коде, – мы будем очень благодарны, если вы сообщите нам о ней. Сделав это, вы избавите других читателей от недопонимания и поможете нам улучшить последующие издания этой книги.

Если вы найдете какие-либо ошибки в коде, пожалуйста, сообщите о них главному редактору по адресу dmkpress@gmail.com, и мы исправим это в следующих тиражах.

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 dmkpress@gmail.com.

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

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 довольно много времени на выбор хорошего эпиграфа. В итоге я остановился на этих двух. R – это потрясающе гибкая платформа и язык для исследования, визуализации и интерпретации 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 R сложен в изучении. Я надеюсь показать вам, что это не так. R обладает настолько широкими возможностями и предлагает такое огромное число аналитических и графических функций (по последним подсчетам их более 50 000), что в одинаковой степени может вызывать бессознательный страх и у новичков, и у опытных пользователей. Однако в этом кажущемся безумии есть своя логика и поэзия. Вооружившись руководствами и инструкциями, вы сможете сориентироваться в огромном разнообразии возможностей и выбрать те инструменты, которые нужны для эффективного и элегантного решения вашей задачи.

Первое мое знакомство с R состоялось несколько лет назад, когда я подал заявление о приеме на должность консультанта по статистике. На встрече перед собеседованием будущий работодатель

[1] Перевод Н. Демуровой.

спросил меня, владею ли я языком R. Следуя стандартным советам специалистов по подбору персонала, я немедленно сказал «да» и приступил к его изучению. Я был опытным статистиком и исследователем с 25-летним опытом программирования в SAS и SPSS, свободно владел несколькими языками программирования. Что тут может быть сложного? Знаменитые последние слова.

Стремясь выучить этот язык программирования (как можно быстрее, ведь день собеседования приближался с угрожающей быстротой), я находил или тома, посвященные внутренней структуре языка, или многочисленные трактаты об отдельных продвинутых статистических методах, написанных специалистами в данной области для своих коллег. Встроенная справка была слишком лаконичной и служила скорее справочником, чем учебным пособием. Каждый раз, когда мне казалось, что я освоил общую логику и возможности R, обнаруживалось что-то новое, заставлявшее почувствовать себя невежественным и ничтожным.

Взявшись осваивать R, я подошел к процессу с позиции исследователя данных. Я пытался понять, что нужно сделать, чтобы успешно обработать, проанализировать и интерпретировать данные, и выделил следующие важные аспекты:

- доступ к данным (получение данных из разных источников);
- очистка данных (замена или удаление пропущенных значений, преобразование признаков в более удобный для обработки формат);
- аннотирование данных (чтобы можно было вспомнить, что представляет каждый их фрагмент);
- обобщение данных (вычисление описательных статистик, помогающих характеризовать данные);
- визуализация данных (потому что картинка на самом деле стоит тысячи слов);
- моделирование данных (выявление зависимостей и проверка гипотез);
- оформление результатов (подготовка таблиц и диаграмм достаточного для публикации качества).

Затем я постарался понять, как можно 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 рассказываю все, что узнал сам.

Затем, спустя год, Марьян Бейс (Marjan Bace) из издательства Manning позвонила и спросила, не хочу ли я написать книгу про R. К этому времени у меня уже было 50 статей в научных журналах, четыре технических руководства, многочисленные главы в книгах и целая книга по методологии исследований, и что тут может быть сложного? Рискую повториться – знаменитые последние слова.

Первое издание вышло в 2011 году, а второе – в 2015-м. Над третьим изданием я начал работать два с половиной года назад. Описание R всегда было непростой задачей, но за последние несколько лет произошла почти что революция, обусловленная ростом популярности больших данных, широким внедрением программного обеспечения tidyverse (tidyverse.org), быстрой разработкой новых подходов к прогнозной аналитике и машинному обучению, а также появлением новых и более мощных технологий визуализации данных. Я хотел отразить все эти важные изменения в третьем издании.

Книгу, которую вы держите в руках, я мечтал иметь много лет назад. Я постарался написать для вас путеводитель по R, который позволит быстро овладеть всеми возможностями этого уникального продукта с открытым исходным кодом, не испытав разочарований 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 Manning, которая предложила мне написать эту книгу;
- Себастьян Стирлинг (Sebastian Stirling), Дженифер Стоут (Jennifer Stout) и Карен Миллер (Karen Miller), редакторы-консультанты по аудитории (development editor) первого, второго и третьего изданий этой книги. Они провели многие часы в телефонных беседах со мной, помогая организовать материал, прояснить основные идеи и в целом сделать текст более интересным;
- Майк Шепард (Mike Shepard), научный редактор, который помог выявить непонятные места и представил свое экспертное мнение о тестировании кода. Я мог смело положиться на его подробные отзывы и суждения;
- Алекс Драгосавлевич (Aleks Dragosavljević), редактор-рецензент (review editor), помог найти рецензентов и координировал процесс рецензирования;
- Дейдре Хайам, помогавшая следить за процессом подготовки книги к печати, и ее команда: Сюзанна Дж. Фокс (Suzanne G. Fox), мой редактор, и Кэти Теннант (Katie Tennant), корректор;
- рецензенты, которые потратили много времени на внимательное чтение текста, находили опечатки и делали ценные замечания: Ален Ломпо (Alain Lompo), Алессандро Пуциелли (Alessandro Puzielli), Арав Агарвал (Arav Agarwal), Эшли Пол Итли (Ashley Paul Eatly), Клеменс Баадер (Clemens Baader), Дэниел Си Догерти (Daniel C Daugherty), Дэниел Кенни-Юнг (Daniel Kenney-Jung), Эрико Лендзиан (Erico Lendzian),

Джеймс Фронхофер (James Frohnhofer), Жан-Франсуа Морин (Jean-François Morin), Дженис Том (Jenice Tom), Джим Фронхофер (Jim Frohnhofer), Кей Энгельхардт (Kay Engelhardt), Келвин Микс (Kelvin Meeks), Кришна Шреста (Krishna Shrestha), Луис Фелипе Медейро Алвес (Luis Felipe Medeiro Alves), Марио Гизель (Mario Giesel), Мартин Перри (Martin Perry), Ник Дрозд (Nick Drozd), Николь Кенигштейн (Nicole Koenigstein), Роберт Самохил (Robert Samohyl), Тиклу Гангули (Tiklu Ganguly), Том Джеффрис (Tom Jeffries), Ульрих Гогер (Ulrich Gauger), Вишал Сингх (Vishal Singh);

- многие участники программы раннего доступа издательства Manning (Manning Early Access Program, MEAP), купившие книгу до того, как она была закончена, задавали великолепные вопросы, указывали на ошибки и давали ценные подсказки.

Все перечисленные помогли сделать эту книгу лучше и полнее.

Я также хотел бы поблагодарить многочисленных разработчиков, сделавших R такой мощной платформой для анализа данных. Этот список включает не только основную команду разработчиков, но и многочисленных сторонников, создавших и поддерживающих дополнительные пакеты, значительно расширяющие возможности R. В приложении E перечислены авторы всех пакетов, упомянутых в этой книге. Отдельно я хотел бы упомянуть Джона Фокса (John Fox), Хадли Викхама (Hadley Wickham), Франка Е. Харрела младшего (Frank E. Harrell Jr.), 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). Она не особенно интересуется статистикой или программированием, но неоднократно прочитала каждую главу и внесла множество исправлений и предложений. Никаким другим способом нельзя выразить свою любовь к другому человеку лучше, чем прочесть ради него текст по многомерной статистике. Она проявила необычайное терпение в вечера и выходные, которые я проводил в работе над этой книгой, выражая свою поддержку и проявляя такт и сочувствие. 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 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, обобщить, преобразовать, исследовать, смоделировать, визуализировать или представить коллегам. Если это так, то R создан для вас! R стал всемирно известным языком программирования для статистического анализа и визуализации данных. В нем реализовано множество методов анализа данных, от самых простых до самых сложных и современных.

Как проект с открытым кодом он доступен для многих платформ, включая Windows, Mac OS X и Linux. Он постоянно развивается, и ежедневно появляются новые процедуры. Кроме того, R поддерживается большим и многоликим сообществом ученых и программистов, которые охотно помогут новичку советами.

Платформа R больше, пожалуй, 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. Новичкам порой сложно понять, что такое R и как работать с этим языком. Даже самые опытные пользователи R с удивлением обнаруживают какие-то возможности, о которых не подозревали.

Третье издание «R в действии» – это руководство-путеводитель по R, знакомящее с самой платформой и ее возможностями. В книге описаны наиболее полезные функции базовой версии и более 90 наиболее часто используемых дополнительных пакетов. Основной упор в книге делается на практическое применение – на то, чтобы вы, руководствуясь прочитанным, могли проанализировать