В память об отце
Рабби Барри Дов Лернер (1942–2023)

Он научил меня:

- быть чрезвычайно любознательным;
- делиться всем, чему научился;
- верить в людей;
- делать все это с юмором.

Оглавление

Предисловие **9**
Благодарности **11**
Об этой книге **12**
Об авторе **17**
О переводчике **17**
Об изображении на обложке **18**
Глава 1. Объект Series **19**
УПРАЖНЕНИЕ 1. Оценки за ежемесячные тесты 24
УПРАЖНЕНИЕ 2. Масштабирование оценок 37
УПРАЖНЕНИЕ 3. Считаем цифры разряда десятков 42
УПРАЖНЕНИЕ 4. Описательная статистика 52
УПРАЖНЕНИЕ 5. Температура по понедельникам 56
УПРАЖНЕНИЕ 6. Пассажиропоток в такси 60
УПРАЖНЕНИЕ 7. Длинные, средние и короткие поездки в такси 63
Заключение 67
Глава 2. Объект DataFrame **68**
УПРАЖНЕНИЕ 8. Чистый доход 73
УПРАЖНЕНИЕ 9. Налоговое планирование 77
УПРАЖНЕНИЕ 10. Добавление новых товаров 88
УПРАЖНЕНИЕ 11. Лидеры продаж 94
УПРАЖНЕНИЕ 12. Поиск выбросов 97
УПРАЖНЕНИЕ 13. Интерполяция 104
УПРАЖНЕНИЕ 14. Выборочное обновление 108
Заключение 112
Глава 3. Импорт и экспорт **113**
УПРАЖНЕНИЕ 15. Загадочные поездки на такси 117
УПРАЖНЕНИЕ 16. Такси и пандемия 125
УПРАЖНЕНИЕ 17. Установка типов данных для столбцов 134
УПРАЖНЕНИЕ 18. Файл passwd в датафрейм 138
УПРАЖНЕНИЕ 19. Курсы биткоина 142
УПРАЖНЕНИЕ 20. Большие города 148
Заключение 151
Глава 4. Индексы **152**
УПРАЖНЕНИЕ 21. Парковочные талоны 154
УПРАЖНЕНИЕ 22. Оценки за вступительные тесты 167

УПРАЖНЕНИЕ 23. Олимпийские игры ... 172
УПРАЖНЕНИЕ 24. Олимпийские сводные таблицы ... 185
Заключение ... 192

Глава 5. Очистка данных ... 193

УПРАЖНЕНИЕ 25. Очистка данных о парковках ... 197
УПРАЖНЕНИЕ 26. Уход знаменитостей ... 207
УПРАЖНЕНИЕ 27. Титаник и интерполяция ... 214
УПРАЖНЕНИЕ 28. Несогласованные данные ... 220
Заключение ... 227

Глава 6. Группировка, объединение и сортировка ... 228

УПРАЖНЕНИЕ 29. Самые продолжительные поездки на такси ... 232
УПРАЖНЕНИЕ 30. Сравним поездки на такси ... 243
УПРАЖНЕНИЕ 31. Расходы туристов по странам ... 256
Заключение ... 266

Глава 7. Сложная группировка, объединение и сортировка ... 267

УПРАЖНЕНИЕ 32. Температура в разных городах ... 267
УПРАЖНЕНИЕ 33. Оценки за вступительные тесты, часть 2 ... 279
УПРАЖНЕНИЕ 34. Снежные и дождливые города ... 293
УПРАЖНЕНИЕ 35. Вино и туризм... ... 302
Заключение ... 313

Глава 8. Промежуточный проект ... 314

Задача ... 315
Заключение ... 336

Глава 9. Строки ... 337

УПРАЖНЕНИЕ 36. Анализируем Алису ... 343
УПРАЖНЕНИЕ 37. Винные слова ... 350
УПРАЖНЕНИЕ 38. Зарплата программистов ... 360
Заключение ... 373

Глава 10. Даты и время ... 374

УПРАЖНЕНИЕ 39. Короткие, средние и длинные поездки на такси ... 381
УПРАЖНЕНИЕ 40. Пишем и читаем даты ... 388
УПРАЖНЕНИЕ 41. Цены на нефть ... 397
УПРАЖНЕНИЕ 42. Чаевые за поездки на такси ... 402
Заключение ... 412

Глава 11. Визуализация ... 413

УПРАЖНЕНИЕ 43. Города ... 416
УПРАЖНЕНИЕ 44. Погода в ящиках с усами ... 430
УПРАЖНЕНИЕ 45. Анализируем стоимость поездок на такси с помощью графиков ... 439
УПРАЖНЕНИЕ 46. Машины, нефть и мороженое ... 455

УПРАЖНЕНИЕ 47. Такси и визуализация в Seaborn ... 474
Заключение ... 483

Глава 12. Оптимизация ... 484

УПРАЖНЕНИЕ 48. Категории ... 490
УПРАЖНЕНИЕ 49. Быстрое чтение, быстрая запись ... 497
УПРАЖНЕНИЕ 50. query и eval ... 507
Заключение ... 515

Глава 13. Итоговый проект ... 516

Задача ... 516
Столбцы и их описание ... 519

Заключение ... 548

Предметный указатель ... 549

Предисловие

Когда я только начинал преподавать Python в компаниях по всему миру, я не был удивлен тем, как мои студенты используют этот язык программирования. Обычно они применяли его для написания скриптов взамен менее выразительных Bash-скриптов, создания серверных веб-приложений, разработки автоматизированных тестов и работы с реляционными базами данных.

Спустя какое-то время я с удивлением обнаружил, что они также используют Python для анализа данных. Да, это мощный и легкий в освоении язык, но с быстродействием у него всегда были проблемы. Как же с его помощью можно анализировать данные?

Вскоре я узнал то, что многие уже знали: оказывается, библиотека NumPy способна объединить легкость использования Python с эффективностью C. Я быстро вскочил в этот вагон и уже совсем скоро начал активно применять эту связку в аналитике и обучать тех, кому только предстояло сделать для себя такое открытие. В то же время сам NumPy мне казался чересчур низкоуровневым инструментом для достижения моих целей.

Когда я познакомился с pandas, все встало на свои места. Эта библиотека объединила в себе скорость и эффективность NumPy с богатейшим API, позволяющим легко выполнять задачи, встающие перед аналитиком каждый день. Я люблю сравнивать pandas с машиной, оснащенной автоматической коробкой передач, значительно превосходящей в удобстве автомобиль с ручной коробкой, коим мне представляется NumPy. Pandas позволяет просто и быстро выполнять чтение и запись в самых разных форматах, очищать исходные данные, анализировать и визуализировать их. В общем, он дал мне все, что было нужно. Я был пленен его очарованием.

Спустя десятилетие после моего знакомства с pandas интерес к нему возрос до небес. Сегодня трудно представить себе аналитика данных, не пользующегося этой библиотекой. За это время где я только ни читал свои курсы по pandas – от команд стартапов до государственных учреждений и от небольших хеджевых фондов до компаний, входящих в первую сотню мирового рейтинга.

Pandas подходит к решению задач иначе по сравнению со стандартными библиотеками, входящими в состав Python. Синтаксис один, но структуры данных и принципы работы с ними совершенно разные. При этом библиотека pandas настолько обширна и разнообразна, что в ее хитросплетениях немудрено запутаться. В отличие от базового Python, незримо продвигающего догму «Должен быть только один способ решения задачи…», pandas не исключает множества вариантов и подходов к одной и той же проблеме. В то же время бывает непросто определить, какой из возможных способов окажется наиболее быстрым и простым в эксплуатации, даже (или особенно) если вы обладаете приличным опыт работы с Python.

Именно по этой причине я являюсь большим поклонником практического подхода к обучению. Только практика позволит вам проникнуть во все тайные

комнаты библиотеки pandas и научиться применять обнаруженные приемы для решения своих задач. Вам недостаточно тренироваться на вымышленных наборах данных – если вы хотите действительно хорошо освоить pandas, вам придется, вооружившись этой библиотекой и изрядной долей храбрости, подступаться к обработке и анализу данных из реального мира с характерными для них недостатками в виде пропущенных значений, смешанных форматов и отсутствия четкой структуры.

Упражнения, собранные в этой книге, проистекают из моих курсов и лекций, которые я читаю все последнее десятилетие. Многие из них за эти годы претерпели изменения, которые сделали их только лучше, поскольку в них были учтены все сложности, с которыми могут сталкиваться начинающие разработчики. Моя цель – дать вам возможность попрактиковаться с pandas и приобрести навыки, которые вы сможете успешно применить в своей работе. Подобно тому как каждый учебный полет пилота на авиасимуляторе приближает его к подъему в воздух настоящего самолета с пассажирами, каждое упражнение из этой книги позволит вам чувствовать себя при работе с pandas более уверенно, и вы не будете испытывать проблем при встрече с реальными задачами.

Благодарности

Написанием этой книги я обязан большому количеству людей.

Хотя на обложке книги красуется только мое имя, очень многие в издательстве Manning Publications оказывали мне бесконечную (и терпеливую) поддержку в процессе ее создания. В первую очередь я хотел бы поблагодарить Майка Стивенса (Mike Stephens), вдохновившего меня на написание второй книги (первая была посвящена Python), и Фрэнсис Лефковиц (Frances Lefkowitz), которая знает, где и как нужно надавить, чтобы процесс написания книги пошел легче. Также я благодарен ей за советы, связанные с редактурой. Кроме того, я получил немало дельных советов от технического рецензента Нинослава Черкеза (Ninoslav Cerkez).

Несколько десятков человек выразили желание помогать комментариями к книге в процессе ее написания и редактуры. Их советы помогли мне значительно улучшить книгу и сделать код в упражнениях и описания более выразительным. Я очень благодарен тем, кто купил книгу на стадии предварительного релиза (MEAP) и оставлял свои комментарии в системе liveBook от Manning.

Также хотелось бы сказать спасибо создателям сайта Pandas Tutor (**https://pandastutor.com**) за возможность красиво визуализировать запросы в pandas, подобно тому как это происходит на сайте Python Tutor (**https://pythontutor.com**). В конце большинства упражнений я буду давать ссылку на мое решение на этом сайте. Природа библиотеки pandas и сайта Pandas Tutor вынудила меня работать с ограниченными наборами данных, но визуализация решений от этого не пострадала.

Отдельные слова благодарности хотелось бы выразить в адрес всех рецензентов. Это Ален Куньот (Alain Couniot), Алекс Гарретт (Alex Garrett), Алекс Лукас (Alex Lucas), Александер Коглер (Alexander Kogler), Амилкар де Абро Нетто (Amilcar de Abreu Netto), Кейдж Слагел (Cage Slagel), Дин Лангсам (Dean Langsam), Джордж Маунт (George Mount), Хелен Мари Лабао Баррамеда (Helen Mary Labao Barrameda), Джефф Нойманн (Jeff Neumann), Джефф Смит (Jeff Smith), Хуан Дельгадо (Juan Delgado), Киран Ананта (Kiran Anantha), Микаэл Дотри (Mikael Dautrey), Мики Тебека (Miki Tebeka), Радучу Сергиу Попа (Răducu Sergiu Popa), Садхана Ганапатираджу (Sadhana Ganapathiraju), Салил Аталайе (Salil Athalye), Сатедж Кумар Саху (Satej Kumar Sahu), Срути Шивакумар (Sruti Shivakumar), Стивен Херрера (Steven Herrera) и Сянгбо Мао (Xiangbo Mao) – ваши советы помогли сделать эту книгу лучше.

Наконец, мои самые глубочайшие благодарности семье за их терпение и понимание в отношении моих бесконечных «Минуточку, одну фразу подредактирую и иду…» на протяжении последних трех лет. Спасибо моей жене Шире и троим нашим детишкам: Атаре, Шикме и Амоцу.

Об этой книге

В былые времена сбор данных мог быть сопряжен с большими трудностями. Но сегодня, когда датчиками, сенсорами и чипами оборудованы все возможные устройства во всех областях жизнедеятельности человека, эти проблемы окончательно ушли в прошлое. Более того, в наши дни данных в мире собирается столько, что их просто не представляется возможным обработать. Отслеживается буквально все – от сделанных нами шагов на прогулке до эффективности рекламы и температуры в любой точке планеты, если не всей Солнечной системы.

Но вместе с такой активностью мы получили и новую проблему, связанную с обработкой и упорядочиванием всех получаемых данных. Как можно эффективно разобраться во всем многообразии полученных сведений и принимать на их основании решения?

На протяжении многих лет таким средством анализа был Microsoft Excel. И на то были свои причины. Excel представляет собой удобный пакет с графическим интерфейсом, который установлен едва ли не на всех компьютерах в мире. С помощью него вы можете достаточно быстро загрузить данные, очистить их, выполнить нужные вычисления и построить понятные отчеты и даже графики.

Но в последние годы у Excel появился юный и дерзкий конкурент в виде pandas. Изначально эта библиотека предстала перед нами в виде удобной обертки пакета NumPy, сочетающего в себе скорость и эффективность вычислений, присущие языку C, с дружелюбностью Python. Pandas дополнил NumPy новыми полезными методами в области обработки строк и дат со временем, а также позволил визуализировать данные. Кроме того, с помощью Pandas можно удобно читать и писать данные в самых разных форматах, включая онлайн-ресурсы и реляционные базы данных. Все это, помноженное на мощь языка Python, способность обрабатывать гораздо большие массивы данных, по сравнению с Excel, и возможность работать в консольном режиме, без графического интерфейса, уверенно склонило чашу весов в сторону pandas. Я преподавал Python и pandas во многих финансовых учреждениях, в которых аналитиков активно переучивали с Excel на pandas. Кроме того, во многих компаниях из самых разных секторов экономики использование библиотеки pandas утверждено на уровне стандарта.

Но, конечно, аналитики не ограничивались в своей работе одним лишь Excel. Сегодня на pandas переходят многие разработчики из R и Matlab – кто-то по экономическим соображениям, кто-то из-за быстродействия, а кто-то по причине очень развитого сообщества и экосистемы с модулями с открытым исходным кодом на Python, доступными в Python Package Index (*PyPI*).

Проблема с pandas заключается в том, что это огромная библиотека с тысячами методов, которые могут принимать сотни разных параметров. Кроме того, в pandas вы можете одну и ту же задачу решить самыми разными способами, значительно отличающимися в плане быстродействия.

Обучение эффективной работе с pandas – это долгий путь проб и ошибок. Сократить этот путь можно только при помощи интенсивных практических занятий

и решения задач, которые позволят вам лучше понять специфику этой библиотеки, подобно тому как постоянные тренировки в спортзале помогают укрепить мышцы спортсмена.

Именно в этом и состоит основная цель книги, которую вы держите в руках. Решив 50 основных и 150 дополнительных упражнений, которые здесь собраны, вы неожиданно обнаружите в себе способность бегло и уверенно говорить на новом для вас языке pandas. В каждом упражнении вы должны будете загрузить реалистичный набор данных и постараться ответить на поставленные вопросы. В процессе чтения книги вы научитесь применять все наиболее важные методы библиотеки pandas и, что более важно, начнете понимать, когда и какие из них являются наиболее приемлемыми.

Эта книга не учебник по pandas, хотя из нее вы в том числе почерпнете немало теоретических знаний. Вместо этого данная книга призвана помочь вам понять внутреннее устройство pandas и научиться применять эту библиотеку для решения задач в реальном мире.

Пожалуйста, не стоит читать эту книгу от корки до корки, как учебное пособие. Также ошибкой будет прочитать условие задачи, решить для себя, что никакой сложности она для вас не представляет, и проследовать дальше. Многие упражнения включают в себя вопросы, ответы на которые на самом деле более сложны, чем кажутся на первый взгляд. Кроме того, если вы будете просто читать мои решения задач, не пробуя решить их самостоятельно, вы никак не сможете погрузиться в глубины внутреннего устройства pandas. Так что, раз уж вы взялись за эту книгу, не стоит уклоняться от самостоятельного штудирования материала и попыток разобраться в поставленной проблеме собственными силами.

В наше время также сложно удержаться от советов не скармливать мои упражнения ChatGPT с дальнейшим просмотром предлагаемых решений. Мало того, что эти решения зачастую будут неправильными, они также не позволят вам самим пройти полноценный путь обучения, как известно, состоящий из ошибок и работы над ними.

Для кого предназначена эта книга

Если вы прошли курс по pandas, но по-прежнему часто обращаетесь к Stack Overflow или Google за решением той или иной задачи, эта книга для вас. Это не учебник в привычном понимании этого слова, а пособие по освоению внутреннего устройства pandas путем решения практических примеров.

Во многих курсах, посвященных pandas, не делается акцент на необходимости хорошо знать Python перед изучением этой библиотеки. Лично я глубоко убежден в том, что такие знания просто необходимы, и в процессе чтения этой книги вы не раз в этом удостоверитесь. В то же время вам не нужно быть настоящим экспертом в области Python. Мне кажется, вам будет достаточно глубокого понимания основных типов данных, циклов, функций и генераторов списков, а также навыка установки модулей с помощью инструкции `pip`. Кроме того, вам может пригодиться понимание анонимных функций в Python (`lambda`), но и это совсем не обязательно.

Структура книги

Эта книга насчитывает 13 глав, в каждой из которых мы сосредоточимся на отдельном аспекте библиотеки pandas. В упражнениях будут активно использоваться техники из предыдущих упражнений, а иногда и из следующих. К примеру, со строками (глава 9) и датами (глава 10) мы поработаем и в первых главах книги. Названия глав можно воспринимать лишь как обобщение того, с чем вам придется столкнуться при их чтении, а не как строгие правила.

Названия и описания глав книги приведены ниже.

- **Глава 1. Объект Series**. В этой главе вы узнаете, что из себя представляют объекты Series и как можно извлекать из них нужные вам данные.
- **Глава 2. Объект DataFrame**. В данной главе мы поговорим о создании датафреймов и извлечении из них требуемых значений.
- **Глава 3. Импорт и экспорт**. Эта глава будет посвящена чтению и записи данных в различные форматы, включая CSV и JSON.
- **Глава 4. Индексы**. В этой главе мы поговорим о техниках установки и извлечения обычных и множественных индексов в pandas.
- **Глава 5. Очистка данных**. В этой главе мы научимся приводить в порядок беспорядочные данные. В числе прочего мы узнаем, как определять наличие дубликатов, обрабатывать пропущенные значения в данных и удалять ненужные или некорректные данные.
- **Глава 6. Группировка, объединение и сортировка**. Здесь мы обсудим саму суть функционала pandas, заключающуюся в группировании данных, объединении нескольких датафреймов и их сортировке по индексам или значениям. Это настолько важные темы, что мы выделили для них сразу две главы.
- **Глава 7. Сложная группировка, объединение и сортировка**. В этой главе мы продолжим обсуждение ключевых методов библиотеки pandas и выведем их понимание на новый качественный уровень.
- **Глава 8. Промежуточный проект**. В этой главе мы реализуем большой проект на основе данных исследования о разработчиках Python.
- **Глава 9. Строки**. В этой главе мы поговорим о работе с текстовыми данными в библиотеке pandas.
- **Глава 10. Даты и время**. Эта глава будет посвящена взаимодействию со значениями, представляющими дату и время.
- **Глава 11. Визуализация**. Здесь мы будем визуализировать наши данные при помощи API pandas и модуля Seaborn.
- **Глава 12. Оптимизация**. В этой главе мы поговорим об оптимизации в отношении быстродействия и использования памяти при обработке данных.
- **Глава 13. Итоговый проект**. В заключительной главе книги мы реализуем итоговый большой проект на основе данных об американских колледжах и университетах.

Упражнения составляют основную часть глав этой книги. При этом каждое упражнение будет разбито на следующие секции:

- **Упражнение**: условие задачи для обдумывания способа ее решения;
- **Подробный разбор**: детальное описание задачи и способа ее решения;
- **Решение**: код решения и (в большинстве случаев) ссылка на код на сайте Pandas Tutor, чтобы вы могли его запустить. Код решения вместе с проверочными кодами доступны также в сопроводительных материалах на странице книги на сайте издательства и в репозитории на GitHub по адресу **https://github.com/reuven/pandas-workout**;
- **Дополнительные упражнения**: три вспомогательных упражнения на ту же тему, которые помогут вам лучше понять обсуждаемый предмет. Детального описания решений этих упражнений вы в книге не найдете, но сами решения[1] будут представлены.

Код решений в книге

Эта книга содержит большое количество кода на языке Python. В отличие от большинства книг код в этой книге стоит воспринимать как руководство к действию по написанию собственного кода, а не просто как полезное чтиво. При наличии у вас достаточного опыта написания кода на Python с использованием библиотеки pandas вы вполне можете написать и более оптимальный код в сравнении с тем, который приведен в книге. В этом случае вы можете написать мне по адресу, приведенному в последнем абзаце книги, и мы вместе поучимся и порадуемся.

Помимо этой книги, код решений всех упражнений, включая дополнительные, можно найти в следующих местах:

- в сопроводительных материалах на странице книги на сайте издательства и в репозитории GitHub по адресу **https://github.com/reuven/pandas-workout**. Код организован по главам и номерам упражнений, чтобы вам удобно было загрузить нужное решение и запустить его на своем компьютере;
- на сайте Pandas Tutor по адресу **https://pandastutor.com**, представляющем великолепное место для изучения всех тонкостей библиотеки pandas. Работая с этим сайтом, вы можете ввести практически любой свой код и увидеть, как на самом деле он работает, с демонстрацией и анимацией всех выполняемых преобразований. В подавляющем большинстве упражнений из этой книги есть ссылки на сайт Pandas Tutor, чтобы вы могли легко и быстро перейти на нужную страницу и запустить пример. Обратите внимание, что в этих примерах обычно будут использоваться небольшие наборы данных.

Код в этой книге будет перемежаться пояснениями и комментариями, а для лучшей читаемости мы будем выделять код `моноширинным шрифтом`.

Внешне в книге фрагменты кода могут отличаться от того, как они представлены на сайте. Ограничения книжного формата вынудили нас вставлять переносы строк и другие элементы форматирования. Кроме того, если пояснения того или

[1] В переводном издании. – *Прим. перев.*

иного фрагмента кода даются в отдельном абзаце, в самом коде могут отсутствовать соответствующие комментарии. На сайте код представлен с полным набором комментариев.

Я надеюсь, что сочетание кода на странице книги, пояснений, ссылки на сайт Pandas Tutor и кода для скачивания поможет вам лучше понять все происходящее в упражнениях и применить полученные знания в своих рабочих сценариях.

Требования к программному/аппаратному обеспечению

Первое и главное, что вы должны установить для плодотворного чтения этой книги, – это, конечно, Python и библиотека pandas. Загрузить и установить Python легче всего по адресу **https://www.python.org**. Я рекомендую установить последнюю доступную версию. Также существуют и другие способы установки Python, включая Windows Store или Homebrew на Mac. Фрагменты кода из этой книги должны успешно работать с любой версией Python начиная с 3.9. На финальном прогоне кода я использовал версию 3.12.

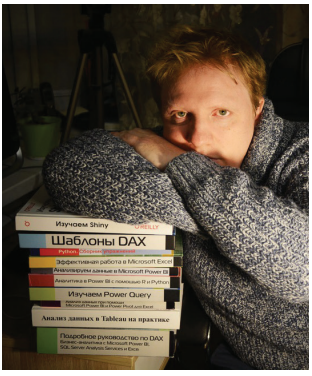
Также вам понадобится библиотека pandas. Я использовал версию 2.1.4, но весь код должен нормально работать со всеми версиями 2.1.x. Загрузить и установить библиотеку можно, воспользовавшись командой `pip install pandas` в терминале.

Для решения упражнений из этой книги вам совсем не обязательно устанавливать графическую среду разработки (IDE) для Python, но с ней вам будет удобнее. Две наиболее популярные среды разработки – это PyCharm (от JetBrains) и Visual Studio Code (от Microsoft). Лично я большой поклонник Jupyter Notebook, который можно установить с помощью команды `pip install jupyter`.

Об авторе

Реувен Лернер (Reuven M. Lerner) – инструктор по Python и pandas, преподающий онлайн и офлайн как для сотрудников крупных компаний, так и в частном порядке. Реувен также выпускает еженедельную рассылку о Python под названием «Better Developers» и рассылку «Bamboo Weekly» с задачами по pandas. Реувен обладает степенью бакалавра Массачусетского технологического института (MIT) в области компьютерных наук и степенью доктора в области педагогических наук Северо-Западного университета (Northwestern). Автор книги «Python Workout», вышедшей в издательстве Manning в 2020 году.

О переводчике



Александр Гинько, обладающий богатым опытом работы в сфере ИТ и более десяти лет посвятивший переводам книг и статей на самые разные темы, в последние годы специализируется на переводе книг в области бизнес-аналитики и программирования для издательства «ДМК Пресс» по направлениям Python, SQL, Power BI, DAX, Excel, Power Query, Tableau, R... На данный момент в активе Александра уже порядка 25 книг, включая одну авторскую, и он продолжает плодотворно работать над переводом новых книг.

Возможно, вам также будут интересны книги «Сверхбыстрый Python» (**https://dmkpress.com/catalog/computer/programming/python/978-5-93700-226-6**) и «Введение в статистическое обучение с примерами на Python» (**https://dmkpress.com/catalog/computer/statistics/978-5-93700-217-4**) в переводе Александра.

Помимо перевода книг, Александр ведет свой канал в Telegram (**https://t.me/alexanderginko_books**), на котором вы можете из первых уст получить ответы на все интересующие вас вопросы об уже переведенных книгах, находящихся в работе и запланированных на будущее. Также на канале можно найти промокоды на все книги Александра для покупки книг на сайте издательства «ДМК Пресс» с большими скидками. Купить книги Александра и следить за переводом новых книг в режиме реального времени можно также с помощью его бота в Telegram по адресу **https://t.me/alexanderginko_books_bot**.

Об изображении на обложке

Картина на обложке книги носит название «Женщина с Тунгуски, Северная Сибирь» (*Femme Tongouse* или *Woman of Tunguska, Northern Siberia*) и принадлежит коллекции художника Жака Грассе де Сэйнт-Совера (Jacques Grasset de Saint-Sauveur). Впервые картина была показана в 1788 году. Все изображения тщательно прорисованы и раскрашены вручную.

В те времена очень легко было по одежде определить местожительство, род занятий и статус человека. Издательство Manning традиционно оформляет обложки книг по компьютерной тематике шедеврами мирового искусства, отдавая дань богатому разнообразию региональных культур прошлых веков.

Глава 1

Объект Series

Если у вас есть опыт работы с библиотекой *pandas*, вы знаете, что с ее помощью нам обычно приходится взаимодействовать с двумерными данными в виде таблиц со столбцами и строками, называемых *датафреймами* (data frame). В то же время каждый столбец представляет собой одномерную структуру, именуемую *Series*[2], что видно на рис. 1.1. Таким образом, вы можете представлять себе датафрейм как коллекцию объектов Series.

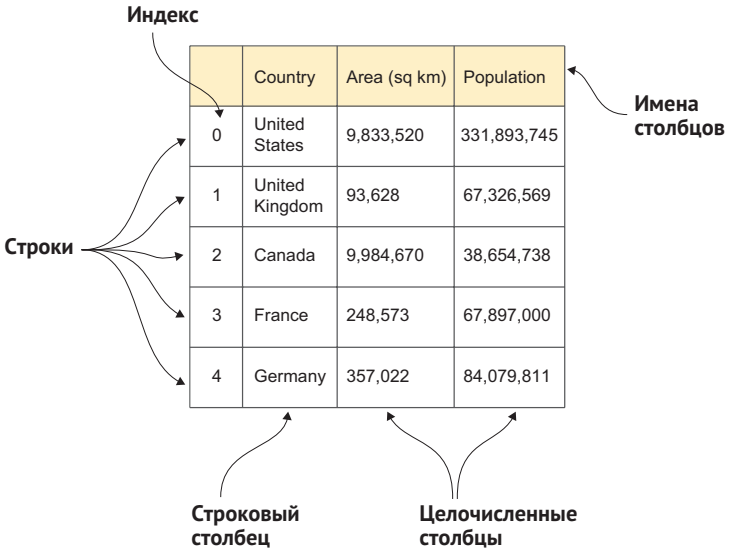


Рис. 1.1. Каждый столбец в датафрейме представляет собой объект Series

Это бывает очень полезно, поскольку вскоре вы узнаете, что большинство методов, применимых к объектам Series, могут быть использованы и с датафреймами с той лишь разницей, что вместо единственного значения они будут возвращать значения для всех столбцов в датафрейме. К примеру, если применить к объекту Series метод `mean`, он вернет среднее значение по столбцу (см. рис. 1.2). Но если вызвать его применительно к датафрейму, pandas под капотом опросит все входящие в датафрейм столбцы на предмет среднего значения и вернет получен-

[2] Мы будем использовать оригинальное название объекта Series, поскольку общепринятого аналога в русском языке не существует. – *Прим. перев.*

ные результаты совокупно в виде нового объекта Series, к которому впоследствии также можно применить разные методы.

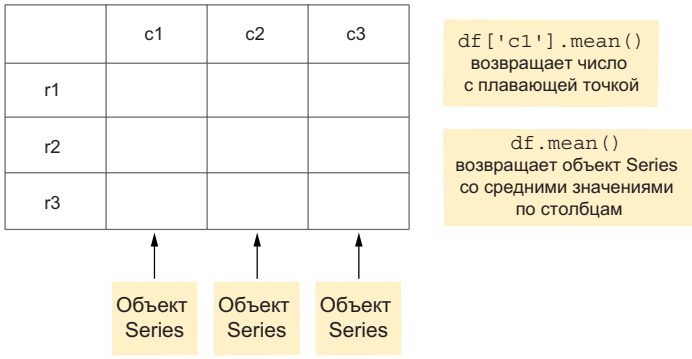


Рис. 1.2. Вызов методов, характерных для объектов Series, таких как `mean`, применительно к датафреймам обычно приводит к получению результата для всех столбцов

Глубокое понимание внутреннего устройства объектов Series поможет вам овладеть библиотекой pandas в полной мере. К примеру, с использованием *булева индекса* (boolean index), также называемого *индексом-маской* (mask index), мы можем легко извлекать строки и столбцы из датафрейма (если вы не знакомы с булевыми индексами, см. врезку «Отбор при помощи булевых значений» далее в этой главе).

Соглашения об именованиях, используемые в этой книге

На протяжении этой книги мы будем часто использовать одни и те же имена для переменных:

- переменной `s` мы будем обозначать объект Series;
- переменная `df` будет ссылаться на датафрейм;
- `pd` представляет собой алиас, или псевдоним, библиотеки pandas, загруженной следующим образом: `import pandas as pd`.

Хотя я являюсь горячим поклонником длинных описательных имен переменных в своих рабочих проектах, в процессе преподавания pandas я предпочитаю ограничиваться короткими именами `s` и `df`. Это бывает очень удобно, особенно с учетом того, что в основном мы будем одновременно использовать один датафрейм или объект Series. В тех редких случаях, когда в моих примерах будет присутствовать более одного датафрейма или Series, я буду добавлять к именам переменных `s` и `df` префиксы или порядковые номера.

Мне также нравится обращаться к классам `Series` и `DataFrame` без использования префикса `pd`. С этой целью я обычно импортирую эти классы из библиотеки pandas явно, как показано ниже:

```
from pandas import Series, DataFrame
```