

# Pandas

В ДЕЙСТВИИ

Борис Пасхавер



# **Pandas** **В ДЕЙСТВИИ**

**Борис Пасхавер**



**Санкт-Петербург • Москва • Минск**

**2023**

ББК 32.973.2-018.1  
УДК 004.43  
П19

### Пасхавер Борис

П19 Pandas в действии. — СПб.: Питер, 2023. — 512 с.: ил. — (Серия «Библиотека программиста»).

ISBN 978-5-4461-1941-7

Язык Python помогает упростить анализ данных. Если вы научились пользоваться электронными таблицами, то сможете освоить и pandas! Несмотря на сходство с табличной компоновкой Excel, pandas обладает большей гибкостью и более широкими возможностями. Эта библиотека для Python быстро выполняет операции с миллионами строк и способна взаимодействовать с другими инструментами. Она даст идеальную возможность выйти на новый уровень анализа данных.

**16+** (В соответствии с Федеральным законом от 29 декабря 2010 г. № 436-ФЗ.)

ББК 32.973.2-018.1  
УДК 004.43

Права на издание получены по соглашению с Manning Publications. Все права защищены. Никакая часть данной книги не может быть воспроизведена в какой бы то ни было форме без письменного разрешения владельцев авторских прав.

Информация, содержащаяся в данной книге, получена из источников, рассматриваемых издательством как надежные. Тем не менее, имея в виду возможные человеческие или технические ошибки, издательство не может гарантировать абсолютную точность и полноту приводимых сведений и не несет ответственности за возможные ошибки, связанные с использованием книги. Издательство не несет ответственности за доступность материалов, ссылки на которые вы можете найти в этой книге. На момент подготовки книги к изданию все ссылки на интернет-ресурсы были действующими.

ISBN 978-1617297434 англ.  
ISBN 978-5-4461-1941-7

©2021 by Manning Publications Co. All rights reserved  
© Перевод на русский язык ООО «Прогресс книга», 2022  
© Издание на русском языке, оформление ООО «Прогресс книга», 2022  
© Серия «Библиотека программиста», 2022

# *Краткое содержание*

---

Предисловие . . . . .	17
Благодарности . . . . .	19
О книге. . . . .	22
Об авторе . . . . .	26
Иллюстрация на обложке . . . . .	27
От издательства . . . . .	28

## **Часть I**

### **Основы pandas**

<b>Глава 1.</b> Знакомство с библиотекой pandas . . . . .	30
<b>Глава 2.</b> Объект Series . . . . .	53
<b>Глава 3.</b> Методы класса Series . . . . .	91
<b>Глава 4.</b> Объект DataFrame . . . . .	120
<b>Глава 5.</b> Фильтрация объектов DataFrame . . . . .	160

## **Часть II**

### **Библиотека pandas на практике**

<b>Глава 6.</b> Работа с текстовыми данными . . . . .	198
<b>Глава 7.</b> Мультииндексные объекты DataFrame . . . . .	220
<b>Глава 8.</b> Изменение формы и сводные таблицы . . . . .	260

## **6** Краткое содержание

<b>Глава 9.</b> Объект GroupBy . . . . .	286
<b>Глава 10.</b> Слияние, соединение и конкатенация . . . . .	307
<b>Глава 11.</b> Дата и время . . . . .	332
<b>Глава 12.</b> Импорт и экспорт данных . . . . .	366
<b>Глава 13.</b> Настройка pandas . . . . .	390
<b>Глава 14.</b> Визуализация . . . . .	401

### **Приложения**

<b>Приложение А.</b> Установка и настройка . . . . .	414
<b>Приложение Б.</b> Экспресс-курс по языку Python . . . . .	437
<b>Приложение В.</b> Экспресс-курс по библиотеке NumPy . . . . .	481
<b>Приложение Г.</b> Генерирование фиктивных данных с помощью Faker . . . . .	490
<b>Приложение Д.</b> Регулярные выражения . . . . .	497

# Оглавление

---

Предисловие . . . . .	17
Благодарности . . . . .	19
О книге. . . . .	22
Для кого она предназначена . . . . .	22
Как организована эта книга: дорожная карта . . . . .	22
О коде . . . . .	24
Источники информации в Интернете . . . . .	25
Об авторе . . . . .	26
Иллюстрация на обложке . . . . .	27
От издательства . . . . .	28

## **Часть I** **Основы pandas**

<b>Глава 1.</b> Знакомство с библиотекой pandas . . . . .	30
1.1. Данные в XXI веке . . . . .	31
1.2. Знакомство с pandas . . . . .	31
1.2.1. Библиотека pandas по сравнению с визуальными приложениями электронных таблиц . . . . .	34
1.2.2. Pandas по сравнению с конкурентами . . . . .	36

## 8 Оглавление

1.3. Обзор библиотеки <code>pandas</code> . . . . .	38
1.3.1. Импорт набора данных . . . . .	38
1.3.2. Операции над объектами <code>DataFrame</code> . . . . .	41
1.3.3. Подсчет значений в <code>Series</code> . . . . .	44
1.3.4. Фильтрация столбца по одному или нескольким критериям. . . . .	45
1.3.5. Группировка данных. . . . .	48
Резюме . . . . .	51
<b>Глава 2. Объект <code>Series</code></b> . . . . .	<b>53</b>
2.1. Обзор <code>Series</code> . . . . .	54
2.1.1. Классы и экземпляры . . . . .	55
2.1.2. Наполнение объекта <code>Series</code> значениями. . . . .	56
2.1.3. Пользовательские индексы для <code>Series</code> . . . . .	58
2.1.4. Создание объекта <code>Series</code> с пропущенными значениями . . . . .	62
2.2. Создание объектов <code>Series</code> на основе объектов языка <code>Python</code> . . . . .	63
2.3. Атрибуты <code>Series</code> . . . . .	65
2.4. Извлечение первой и последней строк . . . . .	68
2.5. Математические операции . . . . .	70
2.5.1. Статистические операции . . . . .	70
2.5.2. Арифметические операции. . . . .	78
2.5.3. Транслирование. . . . .	81
2.6. Передача объектов <code>Series</code> встроенным функциям языка <code>Python</code> . . . . .	84
2.7. Упражнения . . . . .	86
2.7.1. Задачи . . . . .	86
2.7.2. Решения . . . . .	87
Резюме . . . . .	89
<b>Глава 3. Методы класса <code>Series</code></b> . . . . .	<b>91</b>
3.1. Импорт набора данных с помощью функции <code>read_csv</code> . . . . .	92
3.2. Сортировка объектов <code>Series</code> . . . . .	98
3.2.1. Сортировка значений с помощью метода <code>sort_values</code> . . . . .	98
3.2.2. Сортировка по индексу с помощью метода <code>sort_index</code> . . . . .	101
3.2.3. Получение минимального и максимального значений с помощью методов <code>nsmallest</code> и <code>nlargest</code> . . . . .	102

3.3. Перезапись объекта Series с помощью параметра inplace . . . . .	104
3.4. Подсчет количества значений с помощью метода value_counts . . . . .	106
3.5. Вызов функции для каждого из значений объекта Series с помощью метода apply . . . . .	112
3.6. Упражнение . . . . .	116
3.6.1. Задача . . . . .	116
3.6.2. Решение. . . . .	117
Резюме . . . . .	119
<b>Глава 4. Объект DataFrame . . . . .</b>	<b>120</b>
4.1. Обзор DataFrame . . . . .	121
4.1.1. Создание объекта DataFrame на основе ассоциативного массива . . . . .	121
4.1.2. Создание объекта DataFrame на основе ndarray библиотеки NumPy . . . . .	123
4.2. Общие черты Series и DataFrame . . . . .	125
4.2.1. Импорт объекта DataFrame с помощью функции read_csv . . . . .	125
4.2.2. Атрибуты Series и DataFrame: сходство и различие . . . . .	127
4.2.3. Общие методы Series и DataFrame . . . . .	130
4.3. Сортировка объекта DataFrame . . . . .	134
4.3.1. Сортировка по одному столбцу . . . . .	134
4.3.2. Сортировка по нескольким столбцам . . . . .	135
4.4. Сортировка по индексу . . . . .	137
4.4.1. Сортировка по индексу строк . . . . .	138
4.4.2. Сортировка по индексу столбцов. . . . .	139
4.5. Задание нового индекса . . . . .	140
4.6. Извлечение столбцов из объектов DataFrame . . . . .	141
4.6.1. Извлечение одного столбца из объекта DataFrame. . . . .	141
4.6.2. Извлечение нескольких столбцов из объекта DataFrame . . . . .	142
4.7. Извлечение строк из объектов DataFrame . . . . .	144
4.7.1. Извлечение строк по метке индекса . . . . .	144
4.7.2. Извлечение строк по позиции индекса . . . . .	146
4.7.3. Извлечение значений из конкретных столбцов . . . . .	149



4.8. Извлечение значений из объектов Series . . . . .	152
4.9. Переименование столбцов и строк . . . . .	152
4.10. Замена индекса . . . . .	154
4.11. Упражнения . . . . .	155
4.11.1. Задачи . . . . .	155
4.11.2. Решения . . . . .	155
Резюме . . . . .	159
<b>Глава 5. Фильтрация объектов DataFrame . . . . .</b>	<b>160</b>
5.1. Оптимизация памяти, используемой набором данных . . . . .	161
5.1.1. Преобразование типов данных с помощью метода astype. . . . .	163
5.2. Фильтрация по одному условию . . . . .	168
5.3. Фильтрация по нескольким условиям . . . . .	173
5.3.1. Условие И . . . . .	173
5.3.2. Условие ИЛИ . . . . .	174
5.3.3. Логическое отрицание (~) . . . . .	175
5.3.4. Методы для работы с булевыми значениями . . . . .	176
5.4. Фильтрация по условию . . . . .	177
5.4.1. Метод isin. . . . .	177
5.4.2. Метод between. . . . .	178
5.4.3. Методы isnull и notnull . . . . .	180
5.4.4. Обработка пустых значений . . . . .	182
5.5. Решение проблемы дубликатов . . . . .	185
5.5.1. Метод duplicated . . . . .	185
5.5.2. Метод drop_duplicates. . . . .	187
5.6. Упражнения . . . . .	191
5.6.1. Задачи . . . . .	191
5.6.2. Решения . . . . .	192
Резюме . . . . .	196

## Часть II

### Библиотека `pandas` на практике

<b>Глава 6. Работа с текстовыми данными</b> . . . . .	198
6.1. Регистр букв и пробелы. . . . .	199
6.2. Срезы строковых значений. . . . .	203
6.3. Срезы строковых значений и замена символов . . . . .	205
6.4. Булевы методы . . . . .	207
6.5. Разбиение строковых значений. . . . .	210
6.6. Упражнение . . . . .	215
6.6.1. Задача . . . . .	215
6.6.2. Решение. . . . .	215
6.7. Примечание относительно регулярных выражений . . . . .	217
Резюме. . . . .	218
 <b>Глава 7. Мультииндексные объекты DataFrame</b> . . . . .	 220
7.1. Объект <code>MultiIndex</code> . . . . .	222
7.2. Объекты <code>DataFrame</code> с мультииндексами . . . . .	226
7.3. Сортировка мультииндексов . . . . .	232
7.4. Выборка данных с помощью мультииндексов. . . . .	236
7.4.1. Извлечение одного или нескольких столбцов . . . . .	237
7.4.2. Извлечение одной или нескольких строк с помощью <code>loc</code> . . . . .	240
7.4.3. Извлечение одной или нескольких строк с помощью <code>iloc</code> . . . . .	246
7.5. Поперечные срезы . . . . .	248
7.6. Операции над индексом. . . . .	249
7.6.1. Замена индекса . . . . .	250
7.6.2. Задание индекса . . . . .	253
7.7. Упражнения . . . . .	255
7.7.1. Задачи. . . . .	255
7.7.2. Решения . . . . .	257
Резюме. . . . .	259

## 12 Оглавление

<b>Глава 8. Изменение формы и сводные таблицы</b> . . . . .	260
8.1. Широкие и узкие данные . . . . .	261
8.2. Создание сводной таблицы из объекта DataFrame . . . . .	263
8.2.1. Метод pivot_table . . . . .	264
8.2.2. Дополнительные возможности для работы со сводными таблицами . . . . .	268
8.3. Перенос уровней индексов с оси столбцов на ось строк и наоборот . . . . .	271
8.4. Расплавление набора данных . . . . .	273
8.5. Развертывание списка значений . . . . .	278
8.6. Упражнения . . . . .	280
8.6.1. Задачи . . . . .	280
8.6.2. Решения . . . . .	281
Резюме . . . . .	285
<b>Глава 9. Объект GroupBy</b> . . . . .	286
9.1. Создание объекта GroupBy с нуля . . . . .	287
9.2. Создание объекта GroupBy из набора данных . . . . .	289
9.3. Атрибуты и методы объекта GroupBy . . . . .	292
9.4. Агрегатные операции . . . . .	296
9.5. Применение собственных операций ко всем группам набора . . . . .	300
9.6. Группировка по нескольким столбцам . . . . .	301
9.7. Упражнения . . . . .	303
9.7.1. Задачи . . . . .	303
9.7.2. Решения . . . . .	304
Резюме . . . . .	306
<b>Глава 10. Слияние, соединение и конкатенация</b> . . . . .	307
10.1. Знакомство с наборами данных . . . . .	309
10.2. Конкатенация наборов данных . . . . .	311
10.3. Отсутствующие значения в объединенных DataFrame . . . . .	314
10.4. Левые соединения . . . . .	316
10.5. Внутренние соединения . . . . .	318
10.6. Внешние соединения . . . . .	320

10.7. Слияние по индексным меткам . . . . .	323
10.8. Упражнения . . . . .	325
10.8.1. Задачи . . . . .	327
10.8.2. Решения . . . . .	327
Резюме . . . . .	330
<b>Глава 11. Дата и время . . . . .</b>	<b>332</b>
11.1. Знакомство с объектом Timestamp . . . . .	333
11.1.1. Как Python работает с датой и временем . . . . .	333
11.1.2. Как pandas работает с датой и временем . . . . .	336
11.2. Хранение нескольких отметок времени в DatetimeIndex . . . . .	339
11.3. Преобразование значений столбцов или индексов в дату и время . . . . .	341
11.4. Использование объекта DatetimeProperties . . . . .	343
11.5. Сложение и вычитание интервалов времени . . . . .	348
11.6. Смещение дат . . . . .	350
11.7. Объект Timedelta . . . . .	353
11.8. Упражнения . . . . .	358
11.8.1. Задачи . . . . .	358
11.8.2. Решения . . . . .	360
Резюме . . . . .	364
<b>Глава 12. Импорт и экспорт данных . . . . .</b>	<b>366</b>
12.1. Чтение и запись файлов JSON . . . . .	367
12.1.1. Загрузка файла JSON в DataFrame . . . . .	369
12.1.2. Экспорт содержимого DataFrame в файл JSON . . . . .	376
12.2. Чтение и запись файлов CSV . . . . .	377
12.3. Чтение книг Excel и запись в них . . . . .	380
12.3.1. Установка библиотек xlrd и openpyxl в среде Anaconda . . . . .	380
12.3.2. Импорт книг Excel . . . . .	381
12.3.3. Экспорт книг Excel . . . . .	384
12.4. Упражнения . . . . .	386
12.4.1. Задачи . . . . .	387
12.4.2. Решения . . . . .	387
Резюме . . . . .	389

## 14 Оглавление

<b>Глава 13. Настройка pandas</b> . . . . .	390
13.1. Получение и изменение параметров настройки pandas . . . . .	391
13.2. Точность . . . . .	396
13.3. Максимальная ширина столбца . . . . .	397
13.4. Порог округления до нуля . . . . .	397
13.5. Параметры контекста . . . . .	398
Резюме . . . . .	400
<b>Глава 14. Визуализация</b> . . . . .	401
14.1. Установка Matplotlib . . . . .	401
14.2. Линейные графики . . . . .	402
14.3. Гистограммы . . . . .	408
14.4. Круговые диаграммы . . . . .	410
Резюме . . . . .	412

## Приложения

<b>Приложение А. Установка и настройка</b> . . . . .	414
А.1. Дистрибутив Anaconda . . . . .	414
А.2. Процесс установки в macOS . . . . .	416
А.2.1. Установка Anaconda в macOS . . . . .	416
А.2.2. Запуск терминала . . . . .	417
А.2.3. Типичные команды, доступные в терминале . . . . .	418
А.3. Процесс установки в Windows . . . . .	419
А.3.1. Установка Anaconda в Windows . . . . .	419
А.3.2. Запуск командной оболочки Anaconda . . . . .	421
А.3.3. Типичные команды, доступные в Anaconda Prompt . . . . .	422
А.4. Создание новых окружений Anaconda . . . . .	424
А.5. Anaconda Navigator . . . . .	429
А.6. Основы Jupyter Notebook . . . . .	432
<b>Приложение Б. Экспресс-курс по языку Python</b> . . . . .	437
Б.1. Простые типы данных . . . . .	438
Б.1.1. Числа . . . . .	439

Б.1.2. Строки . . . . .	439
Б.1.3. Логические значения . . . . .	443
Б.1.4. Объект None . . . . .	443
Б.2. Операторы . . . . .	444
Б.2.1. Математические операторы . . . . .	444
Б.2.2. Операторы проверки на равенство и неравенство . . . . .	446
Б.3. Переменные . . . . .	448
Б.4. Функции . . . . .	449
Б.4.1. Аргументы и возвращаемые значения . . . . .	450
Б.4.2. Пользовательские функции . . . . .	454
Б.5. Модули . . . . .	456
Б.6. Классы и объекты . . . . .	457
Б.7. Атрибуты и методы . . . . .	458
Б.8. Методы строк . . . . .	459
Б.9. Списки . . . . .	463
Б.9.1. Итерации по спискам . . . . .	469
Б.9.2. Генераторы списков . . . . .	470
Б.9.3. Преобразование строки в список и обратно . . . . .	471
Б.10. Кортежи . . . . .	472
Б.11. Словари . . . . .	474
Б.11.1. Итерации по словарям . . . . .	477
Б.12. Множества . . . . .	478
<b>Приложение В. Экспресс-курс по библиотеке NumPy . . . . .</b>	<b>481</b>
В.1. Измерения . . . . .	481
В.2. Объект ndarray . . . . .	483
В.2.1. Создание набора последовательных чисел с помощью метода arange . . . . .	483
В.2.2. Атрибуты объекта ndarray . . . . .	484
В.2.3. Метод reshape . . . . .	485
В.2.4. Функция randint . . . . .	487
В.2.5. Функция randn . . . . .	488
В.3. Объект nan . . . . .	489

<b>Приложение Г. Генерирование фиктивных данных с помощью Faker . . . .</b>	<b>490</b>
Г.1. Установка Faker . . . . .	490
Г.2. Начало работы с Faker . . . . .	491
Г.3. Заполнение набора данных DataFrame фиктивными значениями . . . . .	494
<b>Приложение Д. Регулярные выражения . . . . .</b>	<b>497</b>
Д.1. Введение в модуль re . . . . .	498
Д.2. Метасимволы . . . . .	499
Д.3. Расширенные шаблоны поиска . . . . .	503
Д.4. Регулярные выражения и pandas. . . . .	507