

УДК 004.438Python:004.6

ББК 32.973.22

M15

Уэс Маккинни

M15 Python и анализ данных: Первичная обработка данных с применением pandas, NumPy и Jupiter / пер. с англ. А. А. Слинкина. 3-е изд. – М.: МК Пресс, 2023. – 536 с.: ил.

ISBN 978-5-93700-174-0

Перед вами авторитетный справочник по переформатированию, очистке и обработке наборов данных на Python. Третье издание, переработанное с учетом версий Python 3.10 и pandas 1.4, содержит практические примеры, демонстрирующие эффективное решение широкого круга задач анализа данных. По ходу дела вы узнаете о последних версиях pandas, NumPy и Jupiter.

Книга принадлежит перу Уэса Маккинни, создателя библиотеки pandas, и может служить практическим современным руководством по инструментарию науки о данных на Python. Она идеально подойдет как аналитикам, только начинающим осваивать Python, так и программистам на Python, еще незнакомым с наукой о данных и научными приложениями. Файлы данных и прочие материалы к книге находятся в репозитории на GitHub и на сайте издательства dmkpress.com.

УДК 004.438Python:004.6

ББК 32.973.22

Authorized Russian translation of the English edition of Python for Data Analysis, 2nd edition. ISBN 9781491957660 © 2022 Wesley McKinney.

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 (анг.) 978-1-09810-403-0

ISBN (рус.) 978-5-93700-174-0

Copyright © 2022 Wesley McKinney

© Оформление, издание, перевод, ДМК Пресс, 2023

Оглавление

Об авторе	13
Об иллюстрации на обложке	14
Предисловие от издательства	15
Предисловие	16
Графические выделения	16
О примерах кода	17
Как с нами связаться	17
Благодарности	18
Глава 1. Предварительные сведения.....	22
1.1. О чем эта книга?.....	22
Какого рода данные?	22
1.2. Почему именно Python?	23
Python как клей.....	23
Решение проблемы «двух языков»	24
Недостатки Python.....	24
1.3. Необходимые библиотеки для Python	25
NumPy.....	25
pandas	25
matplotlib.....	27
IPython и Jupyter	27
SciPy.....	28
scikit-learn	28
statsmodels	29
1.4. Установка и настройка.....	29
Miniconda в Windows	30
GNU/Linux.....	30
Miniconda в macOS.....	31
Установка необходимых пакетов	32
Интегрированные среды разработки (IDE)	33
1.5. Сообщество и конференции.....	33
1.6. Структура книги.....	34
Примеры кода	35
Данные для примеров	35
Соглашения об импорте.....	36

Глава 2. Основы языка Python, IPython и Jupyter-блокноты..... 37

2.1. Интерпретатор Python.....	38
2.2. Основы IPython	39
Запуск оболочки IPython.....	39
Запуск Jupyter-блокнота.....	40
Завершение по нажатию клавиши Tab.....	43
Интроспекция.....	45
2.3. Основы языка Python.....	46
Семантика языка	46
Скалярные типы	53
Поток управления.....	61
2.4. Заключение	64

Глава 3. Встроенные структуры данных, функции и файлы..... 65

3.1. Структуры данных и последовательности	65
Кортеж	65
Список	68
Словарь.....	72
Множество.....	76
Встроенные функции последовательностей	78
Списковое, словарное и множественное включения.....	80
3.2. Функции.....	82
Пространства имен, области видимости и локальные функции	83
Возврат нескольких значений	84
Функции являются объектами.....	85
Анонимные (лямбда-) функции	87
Генераторы.....	87
Обработка исключений.....	90
3.3. Файлы и операционная система	92
Байты и Unicode в применении к файлам	96
3.4. Заключение	98

Глава 4. Основы NumPy: массивы и векторные вычисления 99

4.1. NumPy ndarray: объект многомерного массива.....	101
Создание ndarray	102
Тип данных для ndarray	104
Арифметические операции с массивами NumPy.....	107
Индексирование и вырезание	108
Булево индексирование	113
Прихотливое индексирование.....	116
Транспонирование массивов и перестановка осей	117
4.2. Генерирование псевдослучайных чисел.....	119
4.3. Универсальные функции: быстрые поэлементные операции над массивами	120
4.4. Программирование на основе массивов.....	123
Запись логических условий в виде операций с массивами.....	125

Математические и статистические операции.....	126
Методы булевых массивов	128
Сортировка.....	128
Устранение дубликатов и другие теоретико-множественные операции	130
4.5. Файловый ввод-вывод массивов	130
4.6. Линейная алгебра.....	131
4.7. Пример: случайное блуждание.....	133
Моделирование сразу нескольких случайных блужданий	135
4.8. Заключение	136
Глава 5. Первое знакомство с pandas	137
5.1. Введение в структуры данных pandas	138
Объект Series	138
Объект DataFrame	142
Индексные объекты.....	149
5.2. Базовая функциональность	151
Переиндексация	151
Удаление элементов из оси.....	154
Доступ по индексу, выборка и фильтрация	155
Арифметические операции и выравнивание данных.....	165
Применение функций и отображение	170
Сортировка и ранжирование	172
Индексы по осям с повторяющимися значениями	175
5.3. Редукция и вычисление описательных статистик.....	177
Корреляция и ковариация	180
Уникальные значения, счетчики значений и членство.....	181
5.4. Заключение	185
Глава 6. Чтение и запись данных, форматы файлов.....	186
6.1. Чтение и запись данных в текстовом формате.....	186
Чтение текстовых файлов порциями	193
Вывод данных в текстовом формате.....	195
Обработка данных в других форматах с разделителями.....	196
Данные в формате JSON.....	198
XML и HTML: разбор веб-страниц.....	200
6.2. Двоичные форматы данных.....	203
Формат HDF5.....	205
6.3. Взаимодействие с HTML и Web API	208
6.4. Взаимодействие с базами данных	209
6.5. Заключение	211
Глава 7. Очистка и подготовка данных.....	212
7.1. Обработка отсутствующих данных	212
Фильтрация отсутствующих данных	214
Восполнение отсутствующих данных.....	216

7.2. Преобразование данных	218
Устранение дубликатов	218
Преобразование данных с помощью функции или отображения	220
Замена значений	221
Переименование индексов осей	222
Дискретизация и группировка по интервалам	223
Обнаружение и фильтрация выбросов	226
Перестановки и случайная выборка	227
Вычисление индикаторных переменных	229
7.3. Расширение типов данных	232
7.4. Манипуляции со строками	235
Встроенные методы строковых объектов	235
Регулярные выражения	237
Строковые функции в pandas	240
7.5. Категориальные данные	243
Для чего это нужно	244
Расширенный тип Categorical в pandas	245
Вычисления с объектами Categorical	248
Категориальные методы	250
7.6. Заключение	253

Глава 8. Переформатирование данных:

соединение, комбинирование и изменение формы..... 254

8.1. Иерархическое индексирование	254
Переупорядочение и уровни сортировки	257
Сводная статистика по уровню	258
Индексирование столбцами DataFrame	258
8.2. Комбинирование и слияние наборов данных	260
Слияние объектов DataFrame как в базах данных	260
Соединение по индексу	265
Конкатенация вдоль оси	269
Комбинирование перекрывающихся данных	274
8.3. Изменение формы и поворот	276
Изменение формы с помощью иерархического индексирования	276
Поворот из «длинного» в «широкий» формат	279
Поворот из «широкого» в «длинный» формат	282
8.4. Заключение	284

Глава 9. Построение графиков и визуализация..... 285

9.1. Краткое введение в API библиотеки matplotlib	286
Рисунки и подграфики	287
Цвета, маркеры и стили линий	291
Риски, метки и надписи	292
Аннотации и рисование в подграфике	295
Сохранение графиков в файле	297

Конфигурирование matplotlib	298
9.2. Построение графиков с помощью pandas и seaborn.....	299
Линейные графики.....	299
Столбчатые диаграммы	302
Гистограммы и графики плотности	308
Диаграммы рассеяния.....	310
Фасетные сетки и категориальные данные	313
9.3. Другие средства визуализации для Python	315
9.4. Заключение	316
Глава 10. Агрегирование данных и групповые операции	317
10.1. Как представлять себе групповые операции	318
Обход групп.....	322
Выборка столбца или подмножества столбцов	323
Группировка с помощью словарей и объектов Series	324
Группировка с помощью функций.....	325
Группировка по уровням индекса.....	325
10.2. Агрегирование данных	326
Применение функций, зависящих от столбца, и нескольких функций	328
Возврат агрегированных данных без индексов строк	332
10.3. Метод apply: общий принцип	
разделения–применения–объединения	332
Подавление групповых ключей.....	334
Квантильный и интервальный анализы.....	335
Пример: подстановка зависящих от группы значений	
вместо отсутствующих.....	337
Пример: случайная выборка и перестановка	339
Пример: групповое взвешенное среднее и корреляция.....	341
Пример: групповая линейная регрессия	343
10.4. Групповые преобразования и «развернутая» группировка	343
10.5. Сводные таблицы и перекрестная табуляция	347
Перекрестная табуляция: crosstab.....	350
10.5. Заключение.....	351
Глава 11. Временные ряды	352
11.1. Типы данных и инструменты, относящиеся к дате и времени	353
Преобразование между строкой и datetime	354
11.2. Основы работы с временными рядами.....	356
Индексирование, выборка, подмножества	358
Временные ряды с неуникальными индексами.....	360
11.3. Диапазоны дат, частоты и сдвиг	361
Генерирование диапазонов дат.....	362
Частоты и смещения дат	364
Сдвиг данных (с опережением и с запаздыванием)	366
11.4. Часовые пояса.....	369
Локализация и преобразование	369

Операции над объектами Timestamp с учетом часового пояса	371
Операции над датами из разных часовых поясов	372
11.5. Периоды и арифметика периодов	373
Преобразование частоты периода	374
Квартальная частота периода.....	376
Преобразование временных меток в периоды и обратно.....	377
Создание PeriodIndex из массивов	379
11.6. Передискретизация и преобразование частоты.....	380
Понижающая передискретизация	382
Повышающая передискретизация и интерполяция.....	384
Передискретизация периодов	386
Групповая передискретизация по времени	387
11.7. Скользящие оконные функции	389
Экспоненциально взвешенные функции	392
Бинарные скользящие оконные функции	394
Скользящие оконные функции, определенные пользователем	395
11.8. Заключение.....	396

Глава 12. Введение в библиотеки

моделирования на Python..... 397

12.1. Интерфейс между pandas и кодом модели.....	397
12.2. Описание моделей с помощью Patsy.....	400
Преобразование данных в формулах Patsy	402
Категориальные данные и Patsy.....	404
12.3. Введение в statsmodels	406
Оценивание линейных моделей	407
Оценивание процессов с временными рядами	409
12.4. Введение в scikit-learn	410
12.5. Заключение.....	414

Глава 13. Примеры анализа данных.....415

13.1. Набор данных Bitly с сайта 1.usa.gov	415
Подсчет часовых поясов на чистом Python	416
Подсчет часовых поясов с помощью pandas.....	418
13.2. Набор данных MovieLens 1M	424
Измерение несогласия в оценках.....	428
13.3. Имена, которые давали детям в США за период с 1880 по 2010 год....	432
Анализ тенденций в выборе имен	437
13.4. База данных о продуктах питания министерства сельского хозяйства США.....	446
13.5. База данных Федеральной избирательной комиссии	451
Статистика пожертвований по роду занятий и месту работы	454
Распределение суммы пожертвований по интервалам.....	457
Статистика пожертвований по штатам	459
13.6. Заключение.....	460

Приложение А. Дополнительные сведения

о библиотеке NumPy 461

А.1. Внутреннее устройство объекта ndarray.....	461
Иерархия типов данных в NumPy	462
А.2. Дополнительные манипуляции с массивами.....	463
Изменение формы массива	464
Упорядочение элементов массива в С и в Fortran.....	465
Конкатенация и разбиение массива	466
Эквиваленты прихотливого индексирования: функции take и put	470
А.3. Укладывание	471
Укладывание по другим осям.....	474
Установка элементов массива с помощью укладывания	476
А.4. Дополнительные способы использования универсальных функций.....	477
Методы экземпляра u-функций	477
Написание новых u-функций на Python.....	479
А.5. Структурные массивы и массивы записей	480
Вложенные типы данных и многомерные поля.....	481
Зачем нужны структурные массивы?.....	482
А.6. Еще о сортировке.....	482
Косвенная сортировка: методы argsort и lexsort.....	483
Альтернативные алгоритмы сортировки	485
Частичная сортировка массивов	485
Метод numpy.searchsorted: поиск элементов в отсортированном массиве	486
А.7. Написание быстрых функций для NumPy с помощью Numba	487
Создание пользовательских объектов numpy.ufunc с помощью Numba	489
А.8. Дополнительные сведения о вводе-выводе массивов.....	489
Файлы, отображенные на память.....	489
HDF5 и другие варианты хранения массива.....	491
А.9. Замечания о производительности	491
Важность непрерывной памяти	491

Приложение В. Еще о системе IPython..... 494

В.1. Комбинации клавиш	494
В.2. О магических командах	495
Команда %run	497
Исполнение кода из буфера обмена.....	498
В.3. История команд	499
Поиск в истории команд и повторное выполнение.....	500
Входные и выходные переменные	500
В.4. Взаимодействие с операционной системой	501
Команды оболочки и псевдонимы	502
Система закладок на каталоги	503

В.5. Средства разработки программ.....	504
Интерактивный отладчик.....	504
Хронометраж программы: %time и %timeit	508
Простейшее профилирование: %prun и %run -p.....	510
Построчное профилирование функции.....	512
В.6. Советы по продуктивной разработке кода	
с использованием IPython	514
Перезагрузка зависимостей модуля.....	514
Советы по проектированию программ.....	515
В.7. Дополнительные возможности IPython	516
Профили и конфигурирование.....	516
В.8. Заключение	517
Предметный указатель	518