

ВТОРОЕ ИЗДАНИЕ

Python и анализ данных

*Первичная обработка данных
с применением pandas, NumPy и IPython*

Уэс Маккини



Москва, 2020

Маккини, У. Python и анализ данных : первичная обработка данных с применением pandas, NumPy и IPython / Уэс Маккини. — 2-е изд. — Москва : ДМК Пресс, 2020. — 539 с. : ил.

УДК 004.438Python

ББК 32

Чит. зал №1 — 2 экз.

Второе издание этой книги дает современное практическое введение в разработку научных приложений на Python, ориентированных на обработку данных. Код переписан под версию Python 3.6, добавлены сведения о последних версиях библиотек pandas, NumPy, IPython и Jupyter.

Описаны те части языка Python и библиотеки для него, которые необходимы для эффективного решения широкого круга аналитических задач: интерактивная оболочка IPython и Jupyter-блокноты, библиотеки NumPy и pandas, библиотека для визуализации данных matplotlib и др.

Издание подойдет как аналитикам, только начинающим осваивать обработку данных, так и опытным программистам на Python, еще не знакомым с научными приложениями.



Содержание

Предисловие	14
Об авторе	20
Об иллюстрации на обложке	21
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	26
matplotlib	27
IPython и Jupyter	27
SciPy	28
scikit-learn	28
statsmodels	29
1.4. Установка и настройка	30
Windows	30
Apple OS X	30
GNU/Linux	31
Установка или обновление Python-пакетов	31

Python 2 и Python 3	32
Интегрированные среды разработки (IDE)	32
1.5. Сообщество и конференции	33
1.6. Структура книги	34
Примеры кода	34
Данные для примеров	35
Соглашения об импорте	35
Жаргон	35
Глава 2. Основы языка Python, IPython и Jupyter-блокноты	36
2.1. Интерпретатор Python	37
2.2. Основы IPython	38
Запуск оболочки IPython	38
Запуск Jupyter-блокнота	39
Завершение по нажатию клавиши Tab	42
Интроспекция	43
Команда %run	45
Исполнение кода из буфера обмена	46
Комбинации клавиш	47
О магических командах	48
Интеграция с matplotlib	50
2.3. Основы языка Python	51
Семантика языка	51
Скалярные типы	59
Поток управления	66
Глава 3. Встроенные структуры данных, функции и файлы	71
3.1. Структуры данных и последовательности	71
Кортеж	71
Список	74
Встроенные функции последовательностей	79
Словарь	81
Множество	85
Списковое, словарное и множественное включения	87
3.2. Функции	89
Пространства имен, области видимости и локальные функции	90
Возврат нескольких значений	91
Функции являются объектами	91
Анонимные (лямбда) функции	93
Каррирование: фиксирование части аргументов	94
Генераторы	94
Обработка исключений	97

3.3.	Файлы и операционная система	100
	Байты и Unicode в применении к файлам	102
3.4.	Заключение	104
Глава 4.	Основы NumPy: массивы и векторные вычисления	105
4.1.	NumPy ndarray: объект многомерного массива	107
	Создание ndarray	108
	Тип данных для ndarray	110
	Арифметические операции с массивами NumPy	113
	Индексирование и вырезание	114
	Булево индексирование	119
	Прихотливое индексирование	121
	Транспонирование массивов и перестановка осей	123
4.2.	Универсальные функции: быстрые поэлементные операции над массивами	125
4.3.	Программирование с применением массивов	127
	Запись логических условий в виде операций с массивами	129
	Математические и статистические операции	131
	Методы булевых массивов	132
	Сортировка	133
	Устранение дубликатов и другие теоретико-множественные операции	134
4.4.	Файловый ввод-вывод массивов	135
4.5.	Линейная алгебра	136
4.6.	Генерация псевдослучайных чисел	138
4.7.	Пример: случайное блуждание	139
	Моделирование сразу нескольких случайных блужданий	141
4.8.	Заключение	142
Глава 5.	Первое знакомство с pandas	143
5.1.	Введение в структуры данных pandas	144
	Объект Series	144
	Объект DataFrame	148
	Индексные объекты	154
5.2.	Базовая функциональность	156
	Переиндексация	156
	Удаление элементов из оси	159
	Доступ по индексу, выборка и фильтрация	161
	Целочисленные индексы	165
	Арифметические операции и выравнивание данных	166
	Применение функций и отображение	172
	Сортировка и ранжирование	174
	Индексы по осям с повторяющимися значениями	177

5.3.	Редукция и вычисление описательных статистик.....	179
	Корреляция и ковариация.....	181
	Уникальные значения, счетчики значений и членство.....	183
5.4.	Заключение	186

Глава 6. Чтение и запись данных, форматы файлов

6.1.	Чтение и запись данных в текстовом формате.....	187
	Чтение текстовых файлов порциями.....	193
	Вывод данных в текстовом формате	195
	Обработка данных в формате с разделителями.....	196
	Данные в формате JSON.....	198
	XML и HTML: разбор веб-страниц.....	200
6.2.	Двоичные форматы данных.....	203
	Формат HDF5	204
	Чтение файлов Microsoft Excel.....	206
6.3.	Взаимодействие с HTML и Web API.....	207
6.4.	Взаимодействие с базами данных.....	209
6.5.	Заключение	210

Глава 7. Очистка и подготовка данных

7.1.	Обработка отсутствующих данных	211
	Фильтрация отсутствующих данных	213
	Восполнение отсутствующих данных.....	215
7.2.	Преобразование данных.....	217
	Устранение дубликатов	217
	Преобразование данных с помощью функции или отображения.....	219
	Замена значений.....	221
	Переименование индексов осей.....	222
	Дискретизация и раскладывание	223
	Обнаружение и фильтрация выбросов	226
	Перестановки и случайная выборка.....	228
	Вычисление индикаторных переменных.....	229
7.3.	Манипуляции со строками	232
	Методы строковых объектов.....	232
	Регулярные выражения	234
	Векторные строковые функции в pandas	237
7.4.	Заключение	240

Глава 8. Переформатирование данных: соединение, комбинирование и изменение формы.....

8.1.	Иерархическое индексирование	241
	Переупорядочение и уровни сортировки	244

Сводная статистика по уровню	245
Индексирование с помощью столбцов DataFrame	246
8.2. Комбинирование и слияние наборов данных	247
Слияние объектов DataFrame как в базах данных	247
Соединение по индексу	252
Конкатенация вдоль оси	256
Комбинирование перекрывающихся данных	261
8.3. Изменение формы и поворот	263
Изменение формы с помощью иерархического индексирования	263
Поворот из «длинного» в «широкий» формат	266
Поворот из «широкого» в «длинный» формат	270
8.4. Заключение	272
Глава 9. Построение графиков и визуализация	273
9.1. Краткое введение в API библиотеки matplotlib	274
Рисунки и подграфики	275
Цвета, маркеры и стили линий	278
Риски, метки и надписи	281
Аннотации и рисование в подграфике	284
Сохранение графиков в файле	286
Конфигурирование matplotlib	288
9.2. Построение графиков с помощью pandas и seaborn	288
Линейные графики	289
Столбчатые диаграммы	291
Гистограммы и графики плотности	296
Диаграммы рассеяния	299
Фасетные сетки и категориальные данные	301
9.3. Другие средства визуализации для Python	303
9.4. Заключение	303
Глава 10. Агрегирование данных и групповые операции	304
10.1. Механизм GroupBy	305
Обход групп	308
Группировка с помощью словарей и объектов Series	311
Группировка с помощью функций	312
Группировка по уровням индекса	313
10.2. Агрегирование данных	313
Применение функций, зависящих от столбца и нескольких функций	315
Возврат агрегированных данных без индексов строк	319
10.3. Метод apply: часть общего принципа разделения-применения-объединения	319
Подавление групповых ключей	322

Квантильный и интервальный анализы.....	322
Пример: подстановка зависящих от группы значений вместо отсутствующих	324
Пример: случайная выборка и перестановка	326
Пример: групповое взвешенное среднее и корреляция	328
Пример: групповая линейная регрессия.....	330
10.4. Сводные таблицы и перекрестное табулирование.....	331
Таблицы сопряженности.....	334
10.5. Заключение	335
Глава 11. Временные ряды	336
11.1. Типы данных и инструменты, относящиеся к дате и времени.....	337
Преобразование между строкой и datetime	338
11.2. Основы работы с временными рядами	341
Индексирование, выборка, подмножества	342
Временные ряды с неуникальными индексами	345
11.3. Диапазоны дат, частоты и сдвиг.....	346
Генерация диапазонов дат.....	347
Частоты и смещения дат.....	349
Сдвиг данных (с опережением и с запаздыванием).....	351
11.4. Часовые пояса.....	354
Локализация и преобразование.....	355
Операции над объектами Timestamp с учетом часового пояса.....	357
Операции между датами из разных часовых поясов	358
11.5. Периоды и арифметика периодов.....	359
Преобразование частоты периода.....	360
Квартальная частота периода	362
Преобразование временных меток в периоды и обратно.....	363
Создание PeriodIndex из массивов.....	365
11.6. Передискретизация и преобразование частоты	367
Понижающая передискретизация.....	369
Повышающая передискретизация и интерполяция.....	371
Передискретизация периодов.....	373
11.7. Скользящие оконные функции.....	374
Экспоненциально взвешенные функции	378
Бинарные скользящие оконные функции	379
Скользящие оконные функции, определенные пользователем	381
11.8. Заключение	382
Глава 12. Дополнительные сведения о библиотеке NumPy	383
12.1. Категориальные данные.....	383
Для чего это нужно	383

Категориальные типы в pandas.....	385
Вычисления с категориальными значениями.....	388
Категориальные методы.....	390
12.2. Дополнительные способы использования GroupBy.....	393
Групповые преобразования и GroupBy с «развертыванием»	393
Групповая передискретизация по времени	397
12.3. Сцепление методов.....	399
Метод pipe.....	400
12.4. Заключение	401
Глава 13. Введение в библиотеки моделирования на Python.....	402
13.1. Интерфейс между pandas и кодом модели.....	402
13.2. Описание моделей с помощью Patsy.....	405
Преобразование данных в формулах Patsy.....	408
Категориальные данные и Patsy.....	410
13.3. Введение в statsmodels	412
Оценивание линейных моделей	413
Оценивание процессов с временными рядами.....	416
13.4. Введение в scikit-learn.....	417
13.5. Продолжение своего образования.....	420
Глава 14. Примеры анализа данных.....	422
14.1. 1.usa.gov data from Bitly.....	422
Подсчет часовых поясов на чистом Python.....	423
Подсчет часовых поясов с помощью pandas.....	425
14.2. Набор данных MovieLens 1M	432
Измерение несогласия в оценках.....	437
14.3. Имена, которые давали детям в США за период с 1880 по 2010 год.....	439
Анализ тенденций в выборе имен	444
14.4. База данных о продуктах питания министерства сельского хозяйства США.....	453
14.5. База данных федеральной избирательной комиссии	459
Статистика пожертвований по роду занятий и месту работы.....	462
Распределение суммы пожертвований по интервалам.....	465
Статистика пожертвований по штатам	467
14.6. Заключение	468
Приложение А. Дополнительные сведения о библиотеке NumPy.....	469
A.1. Внутреннее устройство объекта ndarray.....	469
Иерархия типов данных в NumPy.....	470
A.2. Дополнительные манипуляции с массивами.....	471

Изменение формы массива.....	472
Упорядочение элементов массива в C и в Fortran.....	474
Конкатенация и разбиение массива.....	474
Повторение элементов: функции <code>tile</code> и <code>repeat</code>	477
Эквиваленты прихотливого индексирования: функции <code>take</code> и <code>put</code>	479
A.3. Укладывание.....	480
Укладывание по другим осям.....	482
Установка элементов массива с помощью укладывания.....	484
A.4. Дополнительные способы использования универсальных функций.....	485
Методы экземпляра <code>u-функций</code>	485
Написание новых <code>u-функций</code> на Python.....	488
A.5. Структурные массивы.....	489
Вложенные типы данных и многомерные поля.....	489
Зачем нужны структурные массивы?.....	490
A.6. Еще о сортировке.....	491
Косвенная сортировка: методы <code>argsort</code> и <code>lexsort</code>	492
Альтернативные алгоритмы сортировки.....	493
Частичная сортировка массивов.....	494
Метод <code>numpy.searchsorted</code> : поиск элементов в отсортированном массиве.....	495
A.7. Написание быстрых функций для NumPy с помощью Numba.....	496
Создание пользовательских объектов <code>numpy.ufunc</code> с помощью Numba.....	498
A.8. Дополнительные сведения о вводе-выводе массивов.....	498
Файлы, спроецированные на память.....	498
HDF5 и другие варианты хранения массива.....	500
A.9. Замечания о производительности.....	500
Важность непрерывной памяти.....	500
Приложение В. Еще о системе IPython.....	503
V.1. История команд.....	503
Поиск в истории команд и повторное выполнение.....	503
Входные и выходные переменные.....	504
V.2. Взаимодействие с операционной системой.....	505
Команды оболочки и псевдонимы.....	506
Система закладок на каталоги.....	507
V.3. Средства разработки программ.....	507
Интерактивный отладчик.....	507
Хронометраж программы: <code>%time</code> и <code>%timeit</code>	512
Простейшее профилирование: <code>%prun</code> и <code>%run -p</code>	514
Построчное профилирование функции.....	516
V.4. Советы по продуктивной разработке кода с использованием IPython.....	518
Перезагрузка зависимостей модуля.....	518

Советы по проектированию программ	519
В.5. Дополнительные возможности IPython	521
Делайте классы дружественными к IPython	521
Профили и конфигурирование	521
В.6. Заключение	523
Предметный указатель.....	524