

УДК 004.4238Python:004.6pandas
ББК 32.973.2
Х35

Авторы благодарят Софию Хайслер, Джоша Девлина, Александра Анина, Бенджамина Прайка, Теда Петроу, Степана Сокола за предоставленные дополнительные материалы

Хейдт М., Груздев А. В.
Х35 Изучаем pandas / пер. с англ. А. В. Груздева. – М.: ДМК Пресс, 2019. – 682 с.: ил.

ISBN 978-5-97060-670-4

Библиотека pandas – популярный пакет для анализа и обработки данных на языке Python. Он предлагает эффективные, быстрые, высокопроизводительные структуры данных, которые позволяют существенно упростить работу. Данная книга познакомит вас с обширным набором инструментов, предлагаемых библиотекой pandas, – начиная с обзора загрузки данных с удаленных источников, выполнения численного и статистического анализа, индексации, агрегации и заканчивая визуализацией данных и анализом финансовой информации.

Издание предназначено всем разработчикам на языке Python, интересующимся обработкой данных.

УДК 004.4238Python:004.6pandas
ББК 32.973.2

Copyright © Packt Publishing 2017. First published in the English language under the title 'Learning Pandas – Second Edition – (9781787123137)'.

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 978-1-78712-313-7 (англ.)
ISBN 978-5-97060-670-4 (рус.)

Copyright © 2017 Packt Publishing
© Оформление, издание, перевод, ДМК Пресс, 2019

Содержание

Предисловие	15
Глава 1. Библиотека pandas и анализ данных	19
Знакомство с библиотекой pandas	19
Обработка данных, анализ, наука и библиотека pandas	21
Обработка данных.....	22
Анализ данных	22
Наука о данных.....	23
Предназначение библиотеки pandas	23
Процесс анализа данных.....	23
Процесс	24
Взаимосвязь между книгой и процессом анализа данных.....	28
Понятия «данные» и «анализ» в контексте нашего знакомства с библиотекой pandas	29
Типы данных	29
Временные ряды	31
Общие понятия анализа и статистики	31
Другие библиотеки Python, работающие вместе с библиотекой pandas	34
Численные и научные вычисления – NumPy и SciPy	34
Статистический анализ – StatsModels	35
Машинное обучение – scikit-learn	35
РуМС – стохастическое байесовское моделирование	35
Визуализация данных – matplotlib и seaborn.....	36
Выводы	36
Глава 2. Запуск библиотеки pandas	37
Установка Anaconda	37
IPython и Jupyter Notebook.....	39
IPython	39
Jupyter Notebook	40
Знакомство со структурами данных библиотеки pandas – Series и DataFrame	43
Импорт pandas.....	43
Объект Series.....	44
Объект DataFrame.....	48
Загрузка данных из CSV-файла в объект DataFrame.....	52
Визуализация	55
Выводы	56
Глава 3. Представление одномерных данных с помощью объекта Series	57
Настройка библиотеки pandas	58
Создание объекта Series	58
Создание объекта Series с помощью питоновских списков и словарей	58
Создание объекта Series с помощью функций NumPy	60
Создание объекта Series с помощью скалярного значения	61

Свойства <code>.index</code> и <code>.values</code>	61
Размер и форма объекта <code>Series</code>	62
Установка индекса во время создания объекта <code>Series</code>	63
Использование методов <code>.head()</code> , <code>.tail()</code> и <code>.take()</code> для вывода значений	64
Получение значений в объекте <code>Series</code> по метке или позиции	65
Поиск по метке с помощью оператора <code>[]</code> и свойства <code>.ix[]</code>	65
Явный поиск по позиции с помощью свойства <code>.iloc[]</code>	67
Явный поиск по меткам с помощью свойства <code>.loc[]</code>	67
Создание срезов объекта <code>Series</code>	68
Выравнивание данных по меткам индекса	73
Выполнение логического отбора	76
Переиндексация объекта <code>Series</code>	78
Модификация объекта <code>Series</code> на месте	81
Выводы	83

Глава 4. Представление табличных и многомерных данных

с помощью объекта <code>DataFrame</code>	84
Настройка библиотеки <code>pandas</code>	85
Создание объектов <code>DataFrame</code>	85
Создание объекта <code>DataFrame</code> на основе результатов функций <code>NumPy</code>	85
Создание объекта <code>DataFrame</code> с помощью питоновского словаря и объектов <code>Series</code>	87
Создание объекта <code>DataFrame</code> на основе CSV-файла	89
Доступ к данным внутри объекта <code>DataFrame</code>	90
Отбор столбцов в объекте <code>DataFrame</code>	91
Отбор строк в объекте <code>DataFrame</code>	92
Поиск скалярного значения по метке и позиции с помощью <code>.at[]</code> и <code>.iat[]</code>	93
Создание среза датафрейма с помощью оператора <code>[]</code>	94
Логический отбор строк	94
Одновременный отбор строк и столбцов	96
Выводы	96

Глава 5. Выполнение операций над объектом `DataFrame`

и его содержимым	97
Настройка библиотеки <code>pandas</code>	97
Переименование столбцов	98
Добавление новых столбцов с помощью оператора <code>[]</code> и метода <code>.insert()</code>	99
Добавление столбцов за счет расширения датафрейма	100
Добавление столбцов с помощью конкатенации	101
Переупорядочивание столбцов	102
Замена содержимого столбца	103
Удаление столбцов	103
Присоединение новых строк	105
Конкатенация строк	107
Добавление и замена строк за счет расширения датафрейма	109
Удаление строк с помощью метода <code>.drop()</code>	109
Удаление строк с помощью логического отбора	110
Удаление строк с помощью среза	111
Выводы	111

Глава 6. Индексация данных

Настройка библиотеки <code>pandas</code>	112
--	-----

Важность применения индексов	113
Типы индексов библиотеки pandas	115
Основной тип Index	115
Индексы Int64Index и RangeIndex, в качестве меток используются целые числа	115
Индекс Float64Index, в качестве меток используются числа с плавающей точкой	117
Представление дискретных интервалов с использованием IntervalIndex	117
Категории в качестве индекса – CategoricalIndex	118
Индексация по датам и времени с помощью DatetimeIndex	119
Индексация периодов времени с помощью PeriodIndex	119
Работа с индексами	120
Создание и использование индекса в объекте Series или объекте DataFrame	120
Отбор значений с помощью индекса	121
Преобразование данных в индекс и получение данных из индекса	123
Переиндексация объекта библиотеки pandas	124
Иерархическая индексация	125
Выводы	128

Глава 7. Категориальные данные

Настройка библиотеки pandas	129
Создание категориальных переменных	130
Переименование категорий	135
Добавление категорий	136
Удаление категорий	136
Удаление неиспользуемых категорий	137
Установка категорий	137
Вычисление описательных статистик для категориальной переменной	138
Обработка школьных оценок	138
Выводы	141

Глава 8. Численные и статистические методы

Настройка библиотеки pandas	143
Применение численных методов к объектам библиотеки pandas	143
Выполнение арифметических операций над объектами DataFrame или Series	144
Вычисление количества значений	147
Определение уникальных значений (и их встречаемости)	147
Вычисление минимума и максимума	148
Вычисление <i>n</i> наименьших значений и <i>n</i> наибольших значений	148
Вычисление накопленных значений	149
Выполнение статистических операций с объектами библиотеки pandas	150
Получение итоговых описательных статистик	150
Измерение центральной тенденции: среднее, медиана и мода	151
Вычисление дисперсии и стандартного отклонения	153
Вычисление ковариации и корреляции	154
Дискретизация и квантилизация данных	156
Вычисление ранга значений	160
Вычисление процентного изменения для каждого наблюдения серии	161
Выполнение операций со скользящим окном	161
Создание случайной выборки данных	164
Выводы	165

Глава 9. Загрузка данных

Настройка библиотеки pandas	166
-----------------------------------	-----

8 ❖ Содержание

Работа с CSV-файлами и текстовыми/табличными данными	167
Исследование CSV-файла	167
Чтение CSV-файла в датафрейм	168
Указание индекса столбца при чтении CSV-файла	168
Вывод и спецификация типа данных	169
Указание имен столбцов.....	169
Указание конкретных столбцов для загрузки.....	170
Сохранение датафрейма в CSV-файл	170
Работа с данными, в которых используются разделители полей.....	171
Обработка загрязненных данных, в которых используются разделители полей	172
Чтение и запись данных в формате Excel	174
Чтение и запись JSON-файлов	177
Чтение HTML-файлов из интернета.....	178
Чтение и запись HDF5-файлов	180
Загрузка CSV-файлов из интернета	182
Чтение из базы данных SQL и запись в базу данных SQL.....	182
Загрузка данных с удаленных сервисов	185
Загрузка базы данных по экономической статистике Федерального резервного банка Сент-Луиса	185
Загрузка данных Кеннета Френча	187
Загрузка данных Всемирного банка	188
Выводы	192

Глава 10. Приведение данных в порядок 193

Настройка библиотеки pandas	193
Что такое приведение данных в порядок?.....	194
Как работать с пропущенными данными	195
Поиск значений NaN в объектах библиотеки pandas	196
Удаление пропущенных данных.....	198
Обработка значений NaN в ходе арифметических операций	201
Заполнение пропущенных данных.....	202
Прямое и обратное заполнение пропущенных значений	203
Заполнение с помощью меток индекса.....	204
Выполнение интерполяции пропущенных значений.....	205
Обработка дублирующихся данных	207
Преобразование данных	210
Сопоставление значений другим значениям	210
Замена значений.....	211
Применение функций для преобразования данных	214
Выводы	218

Глава 11. Объединение, связывание и изменение формы данных 219

Настройка библиотеки pandas	219
Конкатенация данных, расположенных в нескольких объектах.....	220
Понимание семантики конкатенации, принятой по умолчанию	220
Переключение осей выравнивания	224
Определение типа соединения	225
Присоединение вместо конкатенации	226
Игнорирование меток индекса	226
Слияние и соединение данных.....	227
Слияние данных, расположенных в нескольких объектах.....	227

Настройка семантики соединения при выполнении слияния	230
Поворот данных для преобразования значений в индексы и наоборот	233
Состыковка и расстыковка данных	234
Состыковка с помощью неиерархических индексов	234
Расстыковка с помощью иерархических индексов	236
Расплавление данных для преобразования «широкого» формата в «длинный» и наоборот	239
Преимущества использования состыкованных данных	240
Выводы	241

Глава 12. Агрегирование данных

Настройка библиотеки pandas	242
Обзор схемы «разделение–применение–объединение»	243
Данные для примеров	244
Разделение данных	244
Группировка по значениям отдельного столбца	244
Просмотр результатов группировки	245
Группировка по нескольким столбцам	248
Группировка по уровням индекса	249
Применение агрегирующих функций, преобразований и фильтров	251
Применение агрегирующих функций к группам	251
Преобразование групп данных	253
Исключение групп из процедуры агрегирования	258
Выводы	259

Глава 13. Анализ временных рядов

Настройка библиотеки pandas	260
Представление дат, времени и интервалов	261
Объекты datetime, day и time	261
Создание временной метки с помощью объекта Timestamp	263
Использование объекта Timedelta для представления временного интервала	263
Введение во временные ряды	264
Индексация с помощью объекта DatetimeIndex	264
Создание временного ряда с определенной частотой	269
Вычисление новых дат с помощью смещений	271
Представление временных интервалов с помощью смещений дат	271
Привязанные смещения	274
Представление промежутков времени с помощью объектов Period	275
Создание временного интервала с помощью объекта Period	275
Индексация с помощью объекта PeriodIndex	277
Обработка праздников с помощью календарей	279
Нормализация временных меток с помощью часовых поясов	280
Операции с временными рядами	284
Опережение и запаздывание	284
Преобразование частоты временного ряда	287
Увеличение или уменьшение шага дискретизации временного ряда	289
Применение к временному ряду операций на основе скользящего окна	294
Выводы	297

Глава 14. Визуализация

Настройка библиотеки pandas	299
-----------------------------------	-----

10 ❖ Содержание

Основные инструменты визуализации	299
Создание графиков временных рядов	300
Настройка внешнего вида графика временного ряда	302
Виды графиков, часто использующиеся в статистическом анализе данных	314
Демонстрация относительных различий с помощью столбиковых диаграмм	314
Визуализация распределений данных с помощью гистограмм	316
Визуализация распределений категориальных данных с помощью ящичных диаграмм с усами	318
Отображение накопленных итогов с помощью площадных диаграмм	318
Визуализация взаимосвязи между двумя переменными с помощью диаграммы рассеяния	320
Визуализация оценок распределения с помощью графика ядерной оценки плотности	320
Визуализация корреляций между несколькими переменными с помощью матрицы диаграмм рассеяния	321
Отображение взаимосвязей между несколькими переменными с помощью тепловых карт	322
Размещение нескольких графиков на одном рисунке вручную	323
Выводы	325

Приложение 1. Советы по оптимизации вычислений

в библиотеке pandas	326
Базовое итерирование	327
Итерирование с помощью метода <code>.iterrows()</code>	328
Более лучший способ итерирования с помощью метода <code>.apply()</code>	328
Векторизация с помощью объектов Series	329
Векторизация с помощью массивов NumPy	329
Выводы	330

Приложение 2. Улучшение производительности pandas

(из официального пособия по библиотеке pandas)	331
Написание расширений на языке C для pandas	331
«Чистый» Python	331
Обычный Cython	333
Использование библиотеки Numba	333
Jit	334
Vectorize	334
Вычисление выражений с помощью функции <code>eval()</code>	335
Поддерживаемый синтаксис	336
Примеры использования функции <code>eval()</code>	336
Метод <code>DataFrame.eval()</code>	337

Приложение 3. Используем pandas для больших данных

Работаем с данными бейсбольных игр	341
Внутреннее представление датафрейма	343
Подтипы	344
Оптимизация числовых столбцов с помощью понижающего преобразования	345
Сравнение способов хранения числовых и строковых значений	347
Оптимизация типов object с помощью типа category	349
Задаем типы во время считывания данных	353
Выводы	355

Приложение 4. Пример предварительной подготовки данных в pandas (конкурсная задача Tinkoff Data Science Challenge)	356
Считывание CSV-файла в объект DataFrame	357
Преобразование типов переменных	382
Переименование категорий переменных	384
Обработка редких категорий	385
Разбиение набора данных на обучающую и контрольную	390
Импутация пропусков	393
Конструирование новых признаков	398
Создание переменной, у которой значения основаны на значениях исходной переменной	399
Создание бинарной переменной на основе значений количественных переменных	401
Создание переменной, у которой каждое значение – среднее значение количественной переменной, взятое по уровню категориальной переменной	402
Возведение в квадрат	403
Дамми-кодирование (One-hot Encoding)	405
Кодирование контрастами (Effect Coding)	407
Присвоение категориям в лексикографическом порядке целочисленных значений, начиная с 0 (Label Encoding)	407
Создание переменной, у которой каждое значение – частота наблюдений в категории переменных (Frequency Encoding)	409
Кодирование вероятностями зависимой переменной (Likelihood Encoding)	410
Кодировка средним значением зависимой переменной, сглаженным через сигмоидальную функцию	412
Кодировка средним значением зависимой переменной, сглаженным через параметр регуляризации	415
Кодировка простым средним значением зависимой переменной по схеме leave-one-out	415
Кодировка простым средним значением зависимой переменной по схеме K-fold	416
Кодировка средним значением зависимой переменной, сглаженным через сигмоидальную функцию, по схеме K-fold	416
Присвоение категориям в зависимости от порядка их появления целочисленных значений, начиная с 1 (Ordinal Encoding)	421
Бинарное кодирование (Binary Encoding)	422
Создание переменных-взаимодействий	422
Категоризация (биннинг) количественной переменной	423
Дамми-кодирование и подготовка массивов для обучения и проверки	429
Выбор метрики качества	431
Построение моделей случайного леса, градиентного бустинга и логистической регрессии	447
Математический аппарат логистической регрессии	488
Отдельная предварительная подготовка данных для логистической регрессии	494
Построение логистической регрессии в библиотеке H2O	538

Приложение 5. Пример предварительной подготовки данных в pandas (конкурсная задача предсказания отклика ОТП Банка)	563
Этап I. Построение модели на обучающей выборке – части исторической выборки и ее проверка на контрольной выборке – части исторической выборки	566
I.1. Считывание CSV-файла, содержащего исторические данные, в объект DataFrame	566

I.2. Преобразование типов переменных.....	567
I.3. Импутация пропусков, не использующая результаты математических вычислений (импутация, которую можно выполнять до/после разбиения на обучение/контроль)	570
I.4. Обработка редких категорий	572
I.5. Конструирование новых признаков, не использующее результаты математических вычислений (которое можно выполнять до/после разбиения на обучение/контроль)	575
I.6. Разбиение на обучающую и контрольную выборки.....	577
I.7. Импутация пропусков, использующая статистики – результаты математических вычислений (ее нужно выполнять после разбиения на обучение и контроль).....	577
I.8. Поиск преобразований переменных, максимизирующих нормальность распределения (дается в сокращенном виде).....	578
I.9. Биннинг как один из способов конструирования новых признаков, использующий результаты математических вычислений (нужно выполнять только после разбиения на обучение и контроль).....	582
I.10. Выполнение преобразований, исходя из информации гистограмм распределения и графиков квантиль-квантиль	587
I.11. Конструирование новых признаков	587
I.12. Стандартизация.....	589
I.13. Дамми-кодирование	589
I.14. Подготовка массивов признаков и массивов меток зависимой переменной	590
I.15. Построение логистической регрессии с помощью класса LogisticRegression библиотеки scikit-learn	590
I.16. Настройка гиперпараметров логистической регрессии с помощью класса GridSearchCV	591
I.17. Отбор признаков для логистической регрессии с помощью случайного леса (класса RFE).....	592
I.18. Отбор признаков для логистической регрессии с помощью BorutaPy.....	594
I.19. Проблема дисбаланса классов	597
Этап II. Построение модели на всей исторической выборке и применение к новым данным	605
II.1. Считывание CSV-файла, содержащего исторические данные, в объект DataFrame.....	605
II.2. Предварительная обработка исторических данных.....	605
II.3. Обучение модели логистической регрессии на всех исторических данных	611
II.4. Считывание CSV-файла, содержащего новые данные, в объект DataFrame	611
II.5. Предварительная обработка новых данных	612
II.6. Применение модели логистической регрессии, построенной на всех исторических данных, к новым данным	612
Приложение 6. Работа с датами и строками	615
Работа с датами.....	615
Работа со строками	618
Изменение регистра строк	618
Изменение строкового значения	620
Определение пола клиента по отчеству	620
Удаление лишних символов из строк	624
Удаление повторяющихся строк	627
Извлечение нужных символов из строк	628

Приложение 7. Работа с предупреждением

SettingWithCopyWarning в библиотеке pandas	631
Что представляет из себя предупреждение SettingWithCopyWarning?	632
Присваивание по цепочке (chained assignment)	633
Скрытая цепочка	635
Советы и рекомендации по работе с предупреждением SettingWithCopyWarning	637
Отключение предупреждения	637
Однотипные и многотипные объекты	638
Ложные срабатывания	639
Подробнее о присваивании по цепочке	641
Ложные пропуски	643
И вновь о скрытой цепочке	644

Приложение 8. От Pandas к Scikit-Learn – новый подход

к управлению рабочими процессами	647
Новый уровень интеграции Scikit-Learn с Pandas	647
Краткое резюме и цели статьи	647
Знакомство с классом ColumnTransformer и обновленным классом OneHotEncoder	648
Задача предсказания цен на недвижимость с Kaggle	649
Исследуем данные	649
Удаление зависимой переменной из обучающего набора	649
Кодировка отдельного столбца со строковыми значениями	650
Scikit-Learn – только двумерные данные	650
Импортируем класс, создаем экземпляр класса – модель, обучаем модель –	
трехэтапный процесс работы с моделью	650
У нас NumPy-массив. Где имена столбцов?	651
Проверка корректности первой строки данных	652
Используем метод .inverse_transform()	
для автоматизации данной операции	652
Применение преобразования к тестовому набору	653
Проблема № 1 – новые категории в тестовом наборе	654
Ошибка: Unknown Category	654
Проблема № 2 – пропущенные значения в тестовом наборе	655
Проблема № 3 – пропущенные значения в обучающем наборе	655
Необходимость импутации пропущенных значений	656
Больше о методе .fit_transform()	657
Применение нескольких преобразований к тестовому набору	657
Применение конвейера	658
Почему для тестового набора мы вызываем только метод .transform()?	659
Выполнение преобразований для нескольких столбцов со строковыми значениями	659
Обращение к отдельным этапам конвейера	659
Использование нового ColumnTransformer для отбора столбцов	660
Передаем конвейер в ColumnTransformer	660
Передаем весь объект DataFrame в ColumnTransformer	661
Извлечение названий признаков	661
Преобразование количественных переменных	662
Работа со всеми количественными признаками	662
Передаем конвейер с преобразованиями для категориальных признаков и конвейер	
с преобразованиями для количественных признаков в ColumnTransformer	663
Машинное обучение	664
Перекрестная проверка	665

14 ❖ Содержание

Отбор наилучших значений гиперпараметров с помощью решетчатого поиска	665
Представление результатов решетчатого поиска в виде датафрейма pandas	666
Создание пользовательского трансформера, выполняющего основные преобразования	666
Редкие категории	666
Написание пользовательского класса	666
Применение класса BasicTransformer	669
Использование BasicTransformer в конвейере	669
Биннинг и преобразование количественных переменных с помощью нового класса KBinsDscritizer	670
Отдельная обработка всех столбцов с годами с помощью ColumnTransformer	671
Применение RobustScaler и FunctionTransformer	673
Предметный указатель	677