



Python для Data Science для чайников

3-е издание

Джон Пол Мюллер

Лука Массарон

 **БОМБОРА**
ИЗДАТЕЛЬСТВО
Москва

УДК 004.43
ББК 32.973.2-018.2
М98

John Paul Mueller, Luca Massaron
PYTHON FOR DATA SCIENCE FOR DUMMIES
3rd edition

For Dummies® trademark is the exclusive property of Wiley and is used under license. Original English language edition Copyright © 2024 by John Wiley & Sons, Inc. All rights reserved including the right of reproduction in whole in part in any form. This translation published by arrangement with John Wiley & Sons, Inc.

Python для Data Science для чайников, 3-е издание © 2024 by John Wiley & Sons, Inc. Все права, включая право на воспроизведение, защищены полностью или частично в любой форме. Этот перевод опубликован по соглашению с John Wiley & Sons, Inc.

Wiley, the Wiley Publishing Logo, For Dummies, Dummies Man and related trade dress are trademarks or registered trademarks of John Wiley and Sons, Inc. and/or its affiliates in the United States and/or other countries. Used by permission.

Wiley, логотип the Wiley Publishing, For Dummies, «Для чайников», Dummies Man и связанные с ними элементы являются зарегистрированными товарными знаками John Wiley and Sons, Inc. и/или ее филиалами в США и других странах. Используется с разрешения.

Мюллер, Джон Пол.
М98 Python для Data Science для чайников / Джон Пол Мюллер, Лука Массарон ; [перевод с английского К. А. Вантуха]. — Москва : Эксмо, 2026. — 512 с. — (Для чайников).

ISBN 978-5-04-201412-3

Python — один из ведущих языков программирования для анализа данных. В сочетании с методами Data Science он позволяет обрабатывать большие объемы информации, выявлять скрытые закономерности и принимать обоснованные решения. Эта книга представляет собой пошаговое руководство для начинающих, которое поможет освоить работу с данными даже без опыта программирования. Знание основ Python позволит автоматизировать рутинные задачи, откроет путь к практическому применению методов Data Science и поможет повысить профессиональную квалификацию. Вы научитесь работать с данными, хранить и организовывать их с помощью облачных реляционных баз данных, импортировать и очищать информацию, а также применять современные алгоритмы машинного обучения, включая нейронные сети и ансамбли деревьев решений. Издание обновлено с учетом последней версии Python и содержит актуальные примеры и наборы данных.

УДК 004.43
ББК 32.973.2-018.2

ISBN 978-5-04-201412-3

© Вантух К.А., перевод на русский язык, 2026
© Оформление. ООО «Издательство «Эксмо», 2026

ОГЛАВЛЕНИЕ

Введение	19
Часть 1. Первые шаги в data science и языке Python	25
глава 1. Почему Python подходит для data science	27
глава 2. Чудесные свойства языка Python и другие его возможности	41
глава 3. Установка языка Python для data science	55
глава 4. Работа с Google Colab	73
Часть 2. Начало работы с данными	97
глава 5. Работа с Jupyter Notebook	99
глава 6. Работа с реальными данными	110
глава 7. Обработка данных	133
глава 8. Изменение размерности данных	162
глава 9. Применение полученных знаний на практике	174
Часть 3. Визуальное представление информации	189
глава 10. Краткий курс Matplotlib	191
глава 11. Визуализация данных	209
Часть 4. Обработка данных	235
глава 12. Дальнейшие возможности Python	237
глава 13. Исследование анализа данных	261
глава 14. Понижение размерности	291
глава 15. Кластеризация	314
глава 16. Поиск выбросов в данных	333
Часть 5. Обучение на данных	349
глава 17. Знакомство с четырьмя простыми и эффективными алгоритмами	351
глава 18. Перекрестная валидация, выборка и оптимизация	373
глава 19. Увеличение сложности при помощи линейных и нелинейных приемов	399
глава 20. Могущество толпы	442

Часть 6. Часть про число десять	467
глава 21. Десять ресурсов	469
глава 22. Десять задач	476
Указатель	487

СОДЕРЖАНИЕ

ВВЕДЕНИЕ	19
О книге	20
Дурацкие предположения	21
Значки, используемые в книге	22
За рамками книги	23
Что дальше (или Как пользоваться книгой)	24

ЧАСТЬ 1. ПЕРВЫЕ ШАГИ В DATA SCIENCE И ЯЗЫКЕ PYTHON	25
---	----

ГЛАВА 1 Почему Python подходит для data science	27
Представление о Python как языке	28
Сферы применения Python как языка общего назначения ..	28
Python как интерпретируемый язык	30
Python как компилируемый язык	30
Определение для Data Science	31
Возникновение data science	31
Основные навыки специалиста по data science	32
Связь data science, больших данных и искусственного интеллекта	33
Создание конвейера data science	34
Представление о роли языка Python в data science	35
Новое положение специалистов по data science	36
Работа с универсальным, простым и эффективным языком ..	36
Быстрый переход к использованию языка Python	38
Загрузка данных	38
Обучение модели	39
Просмотр результатов	40

ГЛАВА 2 Чудесные свойства языка Python и другие его возможности	41
Работа с Python	42
Вклад языка Python в data science	42
Первое представление о языке	44
Необходимость отступов	44
Работа с Jupyter Notebook и Google Colab	45
Быстрое прототипирование и экспериментирование	45

Соображения относительно скорости выполнения	47
Возможность визуализации	48
Применение экосистемы Python в data science.	50
Научные расчеты: SciPy.	50
Фундаментальные операции для научных расчетов: NumPy.	51
Анализ данных: pandas.	51
Реализация машинного обучения: Scikit-learn	52
Переход к глубокому обучению: TensorFlow и Keras	52
Эффективный анализ: XGBoost	53
Построение графиков: Matplotlib.	54
Создание графов: NetworkX	54
глава 3. Установка языка Python для data science	55
Работа с Anaconda.	56
Работа с Jupyter Notebook.	56
Доступ к Anaconda Prompt.	57
Установка Anaconda под Windows	58
Установка Anaconda на Linux	62
Установка Anaconda на Mac OS X.	63
Загрузка датасетов и примеров кода.	65
Использование Jupyter Notebook	65
Задание кодового репозитория.	66
Создание нового блокнота.	67
Добавление содержимого в блокнот.	68
Экспорт блокнота	69
Общее представление о наборах данных, используемых в данной книге	71
глава 4. Работа с Google Colab	73
Определение Google Colab.	74
Общее представление о том, как работает Colab	74
Об отличиях онлайн-кодирования	75
Поддержка локальной среды выполнения	77
Работа с блокнотами.	78
Создание блокнота	78
Доступ к существующему блокноту	79
Сохранение блокнотов.	81
Скачивание блокнотов.	84
Выполнение ежедневных задач	85
Создание ячеек кода.	85
Создание текстовых ячеек	88

Создание специальных ячеек	89
Редактирование ячеек	89
Перемещение ячеек	90
Использование аппаратного ускорения	90
Выполнение кода	91
Просмотр блокнота	92
Просмотр оглавления	92
Получение информации о блокноте	93
Контроль исполнения кода	93
Совместное использование блокнота	94
Получение справочной информации	95
 ЧАСТЬ 2. НАЧАЛО РАБОТЫ С ДАННЫМИ	 97
 ГЛАВА 5. Работа с Jupyter Notebook	 99
Использование Jupyter Notebook	100
Работа со стилями	100
Получение справки Python	102
Использование магических функций	103
Знакомство с объектом	104
Перезапуск ядра	106
Возврат к точке восстановления	106
Интеграция мультимедиа и графики	107
Внедрение графиков и других изображений	107
Загрузка примеров с сайтов онлайн	108
Получение онлайн-графики и мультимедиа	108
 ГЛАВА 6. Работа с реальными данными	 110
Загрузка, потоковая загрузка и ограниченная выборка	111
Загрузка в память небольших объемов данных	112
Потоковая загрузка данных в память	113
Создание вариантов изображения	114
Построение выборки данных разными способами	115
Доступ к данным в структурированном плоском файле	117
Чтение из текстового файла	118
Чтение формата CSV с разделением запятыми	119
Чтение файлов Excel и Microsoft Office вообще	121
Пересылка данных в форме неструктурированного файла	122
Работа с данными из реляционных баз	126
Взаимодействие с данными из баз NoSQL	127
Доступ к данным в сети	128

глава 7. Обработка данных	133
Жонглирование NumPy и pandas	134
Когда пользоваться NumPy.	135
Когда пользоваться pandas.	135
Проверка данных	136
Выяснение содержания данных	137
Удаление дубликатов	138
Создание карты и плана данных.	139
Работа с категориальными переменными	142
Создание категориальных переменных	143
Переименование уровней	145
Комбинирование уровней	145
Обработка дат, встречающихся в данных	147
Форматирование значений даты и времени	147
Выбор наиболее подходящего формата времени	148
Обработка отсутствия данных	149
Поиск отсутствующих данных	149
Представление отсутствующих данных	150
Импутация отсутствующих данных.	151
Срез и сегментация данных	153
Срез строк	153
Срез столбцов.	154
Сегментация данных	155
Конкатенация и трансформация	155
Добавление новых записей и переменных.	156
Удаление данных	158
Сортировка и перемешивание.	158
Агрегирование данных на любом уровне.	160
 глава 8. Изменение размерности данных	 162
Токенизация данных при помощи модели «мешок слов»	163
Общее представление о модели «мешок слов»	163
Упорядочение элементов текста при помощи n-грамм	166
Реализация преобразований TF-IDF	168
Работа с графами.	170
Общее представление о матрице смежности.	171
Работа с основами NetworkX.	171

глава 9. Применение полученных знаний на практике	174
Рассмотрение задачи и данных в контексте	176
Оценка задачи data science	176
Исследование решений	179
Формулировка гипотезы	180
Подготовка данных	181
Искусство конструирования признаков	181
Определение для конструирования признаков	181
Объединение переменных	182
Общее представление о биннинге и дискретизации	183
Использование индикаторных переменных	184
Преобразование распределений	184
Операции над массивами	185
Использование векторизации	186
Простые арифметические операции над векторами и матрицами	186
Умножение матрицы на вектор	187
Умножение матриц	188

часть 4. ВИЗУАЛЬНОЕ ПРЕДСТАВЛЕНИЕ ИНФОРМАЦИИ 189

глава 10. Краткий курс Matplotlib	191
Начинаем с графика	192
Определение графика	193
Создание нескольких кривых и графиков	194
Сохранение результатов на диске	195
Настройка осей, делений и сетки	195
Настройка осей	196
Форматирование осей	196
Добавление сетки	198
Настройка внешнего вида линии	199
Работа со стилями линий	200
Использование цветов	201
Добавление маркеров	202
Использование меток, аннотаций и легенды	205
Добавление меток	205
Аннотирование графика	206
Добавление легенды	207

глава 11. Визуализация данных	209
Выбор графика в соответствии с задачей	210
Создание сравнений при помощи столбчатых диаграмм	210
Показ распределений при помощи гистограмм	212
Изображение групп при помощи коробчатых диаграмм	214
Выявление закономерностей при помощи диаграмм рассеивания	215
Создание сложных диаграмм рассеивания	217
Отображение групп	217
Отображение корреляции	218
Отображение временных последовательностей	220
Представление времени на осях	220
Изображение трендов относительно времени	222
Отображение географических данных	224
Использование виртуальной среды в Notebook	225
Отображение географических данных при помощи Cartopy	227
Отказ от устаревших библиотек: пакет Basemap	229
Отображение графов	230
Создание неориентированных графов	230
Создание ориентированных графов	232
 ЧАСТЬ 4. ОБРАБОТКА ДАННЫХ	 235
 глава 12. Дальнейшие возможности Python	 237
Работаем с Scikit-learn	238
Знакомство с классами в Scikit-learn	238
Определение задач data science	239
Использование преобразующих функций	243
Соединение оценочных функций в конвейер	245
Преобразование целевых переменных	246
Объединение признаков	246
Обработка неоднородных данных	247
Замер времени и скорость выполнения	250
Замеры времени при помощи timeit	250
Работа с профайлером памяти	253
Параллельные вычисления на нескольких процессорных ядрах	256
Метод перехода к параллельным вычислениям на нескольких ядрах	257
Демонстрация параллельных вычислений	258

глава 13. Исследование анализа данных	261
Методология EDA	262
Определение описательной статистики для числовых данных	264
Измерение центральной тенденции	265
Измерение дисперсии и размаха	266
Работа с процентилями	267
Определение степени нормальности распределения	268
Подсчеты для категориальных данных	270
Общее представление о частотах	271
Создание таблиц сопряженности	272
Создание прикладных графиков для разведочного анализа	273
Исследование при помощи ящичных диаграмм	274
Выполнение t-тестов после ящичных диаграмм	276
Наблюдение в параллельных координатах	277
Построение графиков распределений	278
Построение диаграмм рассеяния	279
Общее представление о корреляции	281
Общее представление о ковариации и корреляции	282
Использование непараметрической корреляции	284
Знакомство с критерием хи-квадрат для таблиц	285
Работа с V Крамера	286
Модификация распределений данных	287
Использование различных статистических распределений	287
Создание стандартизации по Z-показателю	288
Преобразование других важных распределений	288
глава 14. Понижение размерности	291
Общее представление об SVD-разложении	292
Поиск путей сокращения размерности	293
SVD-разложение в качестве средства измерить невидимое	296
Применение факторного анализа и МГК	297
Представление о психометрической модели	297
Поиск скрытых факторов	298
Использование компонентов вместо факторов	299
Сокращение размерности	300
Сжатие информации при помощи t-SNE	301
Общее представление о некоторых приложениях	303
Распознавание лиц при помощи МГК	304
Извлечение тем при помощи НМФ	307
Рекомендательная система для кинофильмов	310

глава 15. Кластеризация	314
Кластеризация методом К-средних	315
Общее представление об алгоритмах, основанных на центроидах	317
Создание примера с данными, содержащими изображение	319
Поиск оптимальных решений	320
Кластеризация больших данных	323
Выполнение иерархической кластеризации	325
Использование метода иерархической кластеризации	327
Визуализация решений, полученных методами агрегатной кластеризации	328
Идентификация новых групп при помощи DBSCAN	329
глава 16. Поиск выбросов в данных	333
Общее представление о поиске выбросов	334
Поиск дополнительных источников проблем	335
Общее представление об аномалиях и новизне	336
Знакомство с простым однофакторным методом	338
Применение гауссова распределения	340
Обработка выбросов	341
Общее представление о многофакторном подходе	343
Применение анализа главных компонент	343
Поиск выбросов при помощи кластерного анализа	345
Автоматизация поиска с помощью леса изоляции (Isolation Forest)	347
часть 5. ОБУЧЕНИЕ НА ДАННЫХ	349
глава 17. Знакомство с четырьмя простыми и эффективными алгоритмами	351
Угадай число: линейная регрессия	352
Общее представление о семействе линейных моделей	352
Использование большего числа переменных	354
Общее представление об ограничениях и проблемах	356
Переход к логистической регрессии	358
Применение логистической регрессии	359
Случай более двух классов	361
Смотрим на вещи проще и опускаемся до наивного Байесова алгоритма	362
Почему «наивный Байес» не так уж наивен	364
Предсказание классификации текста	366

Ленивое обучение на ближайших соседях	368
Прогнозирование после того, как соседи найдены	369
Как разумно выбрать параметр k	371
глава 18. Перекрестная валидация, выборка и оптимизация	373
О проблеме обучения модели	374
Общее представление о предвзятости и дисперсии	376
Тактика выбора моделей	377
Разделение набора данных на учебное и тестовое множество	380
Перекрестная валидация	384
Использование перекрестной валидации на k фолдах	385
Стратификация выборок в сложных данных	386
Профессионально выбираем переменные	389
Выбор при помощи однофакторных метрик	389
Прямой и обратный отбор	391
Накачка гиперпараметров	392
Реализация дискретного поиска	393
Пробуем случайный поиск	397
глава 19. Увеличение сложности при помощи линейных и нелинейных приемов	399
Использование нелинейных преобразований	400
Преобразование переменных	401
Создание связей между переменными	402
Регуляризация линейных моделей	405
Применение регрессии L_2	407
Применение регрессии L_1	408
Необычное применение регуляризации	408
Комбинирование L_1 и L_2 : Elasticnet	409
Едим большие данные по кусочкам	410
Как определить, когда данных слишком много?	410
Реализация стохастического градиентного спуска	410
Общее представление о методе опорных векторов	414
Опора на вычислительный метод	415
Фиксирование многих новых параметров	417
Классификация при помощи SVC	419
Перейти к нелинейности просто	425
Вычисление регрессии при помощи SVR	427
Создание стохастического решения при помощи SVM	430
Играем с нейронными сетями	435
Общее представление о нейронных сетях	435
Классификация и регрессия при помощи нейронов	437

ГЛАВА 20. Могущество толпы	442
Начинаем с простого дерева решений	443
Общее представление о дереве решений	443
Создание деревьев классификации	447
Создание деревьев регрессии	449
Пропадаем в случайном лесу	451
Делаем машинное обучение доступным для масс	452
Работа с классификатором случайного леса	454
Работа с регрессором случайного леса	456
Оптимизация случайного леса	457
Бустинг предсказаний	459
Много слабых моделей выигрывают	459
Классификация методом градиентного бустинга	460
Регрессия методом градиентного бустинга	461
Использование гиперпараметров градиентного бустинга	462
Использование XGBoost (экстремального градиентного бустинга)	463
 ЧАСТЬ 6. ЧАСТЬ ПРО ЧИСЛО ДЕСЯТЬ	 467
ГЛАВА 21. Десять ресурсов	469
Узнаем новости на Reddit	470
Задаем хороший старт на KDNuggets	470
Ищем бесплатные учебные ресурсы на Quora	471
Заряжаемся идеями на Oracle AI & Data Science Blog	471
Получаем доступ к огромному списку ресурсов на Data Science Central	472
Узнаем о новых методологиях для новичков на Data Science 101	472
Получаем наиболее авторитетные источники на Udacity	473
Получаем помощь в решении сложных задач на Conductrics	473
Узнаем факты о науке о данных с открытым кодом на Springboard	474
Находим ресурсы для разработчиков с Джонатаном Боуэром	475
 ГЛАВА 22. Десять задач	 476
Удаление информации, позволяющей идентифицировать личность	477
Создание безопасной среды обработки данных	478
Работа с задачей многих источников данных	479
Оттачивание приемов борьбы с переобучением	480

Находим дорогу через набор данных MovieLens	480
Поиск верного источника данных.	481
Работа с рукописной информацией.	482
Работа с изображениями.	484
Выяснение происхождения данных	485
Работа с очень большим графом	486
УКАЗАТЕЛЬ	487
Об авторах.	505
Посвящение Луки	506
Посвящение Джона.	506
Благодарности Луки.	506
Благодарности Джона	506