

O'REILLY®

alist

Изучаем Data Science

Обработка, исследование, визуализация
и моделирование данных с помощью Python



Сэм Лау,
Джозеф Гонсалес,
Дебора Нолан

Оглавление

Отзывы на книгу	13
Предисловие	14
Требуемые базовые знания.....	15
Структура книги.....	15
Принятые условные обозначения.....	16
Примеры применения кода.....	16
Онлайн-обучение O'Reilly.....	17
Наши контакты.....	17
Благодарности.....	18
ЧАСТЬ I. ЖИЗНЕННЫЙ ЦИКЛ DATA SCIENCE	19
Глава 1. Жизненный цикл Data Science	21
Этапы жизненного цикла.....	21
Примеры жизненного цикла.....	24
Краткие выводы.....	25
Глава 2. Постановка вопроса и охват данных	26
Большие данные и новые возможности.....	26
Пример: Google Flu Trends.....	27
Целевая совокупность, фрейм доступных данных, выборка.....	29
Пример: что побуждает к активности участников онлайн-сообщества?.....	31
Пример: кто победит на выборах?.....	31
Пример: связь экологических угроз со здоровьем человека.....	32
Инструменты и протоколы.....	34
Измерение параметров природных явлений.....	34
Пример: определение содержания CO ₂ в воздухе.....	35
Точность.....	36
Виды смещения.....	38
Виды дисперсии.....	40
Краткие выводы.....	41
Глава 3. Структура данных и моделирование	43
Урновая модель.....	44
Схемы выборки.....	46
Выборочное распределение статистики.....	48
Моделирование выборочного распределения.....	49
Моделирование при помощи гипергеометрического распределения.....	50

Пример: моделирование смещения и дисперсии опросов избирателей	52
Урновая модель опросов в Пенсильвании	53
Урновая модель со смещением	55
Проведение более масштабных опросов	56
Пример: моделирование рандомизированного испытания вакцины	58
Охват	58
Урновая модель случайного распределения	60
Пример: измерение качества воздуха	61
Краткие выводы	64
Глава 4. Моделирование при помощи сводной статистики.....	66
Константная модель	66
Минимизация потерь.....	68
Средняя абсолютная ошибка.....	69
Среднеквадратичная ошибка.....	72
Выбор функции потерь	74
Краткие выводы	74
Глава 5. Пример из практики: почему мой автобус всегда опаздывает?.....	76
Постановка вопроса и охват данных.....	76
Первичная обработка данных.....	77
Изучение расписания автобусов	80
Моделирование времени ожидания	83
Краткие выводы	87
ЧАСТЬ II. ТАБЛИЧНЫЕ ДАННЫЕ.....	89
Глава 6. Работа с датафреймами с помощью <i>pandas</i>	91
Подмножество.....	92
Охват данных и постановка вопроса	92
Датафрейм и индекс	93
Срез.....	94
Фильтрация строк.....	98
Пример: когда стало популярным имя Luna?	100
Агрегирование	102
Агрегирование и базовая группировка.....	103
Пример: использование функции <code>.value_counts()</code>	104
Группировка по нескольким столбцам.....	105
Пользовательские агрегатные функции	106
Поворот	109
Соединение	110
Внутренние соединения.....	111
Левые, правые и внешние соединения	113
Пример: популярность категорий имен из статьи <i>New York Times</i>	114
Преобразование	116
Метод <code>apply</code>	117
Пример: популярность имен на букву "L"	118
Цена метода <code>apply</code>	120
Отличие датафрейма от других представлений данных	121
Датафрейм и электронная таблица	121

Датафрейм и матрица.....	121
Датафрейм и отношение.....	122
Краткие выводы.....	123
Глава 7. Работа с отношениями с помощью SQL.....	124
Подмножество.....	124
Основы SQL: <i>SELECT</i> и <i>FROM</i>	125
Что такое отношение?.....	126
Срез.....	126
Фильтрация строк.....	128
Пример: когда стало популярным имя Luna?.....	129
Агрегирование.....	131
Базовая группировка и агрегирование с помощью <i>GROUP BY</i>	131
Группировка по нескольким столбцам.....	132
Другие агрегатные функции.....	133
Соединение.....	134
Внутренние соединения.....	135
Левые и правые соединения.....	136
Пример: популярность категорий имен из статьи <i>NYT</i>	137
Преобразование и обобщенные табличные выражения.....	138
Функции SQL.....	139
Многошаговые запросы с использованием оператора <i>WITH</i>	141
Пример: популярность имен на букву "L".....	142
Краткие выводы.....	142
ЧАСТЬ III. ОСМЫСЛЕНИЕ ДАННЫХ.....	145
Глава 8. Первичная обработка файлов.....	147
Примеры источников данных.....	148
Исследование DAWN.....	148
Безопасность пищевых продуктов в ресторанах Сан-Франциско.....	149
Форматы файлов.....	150
Формат с разделителями.....	150
Формат с фиксированной шириной.....	152
Иерархические форматы.....	153
Свободно форматированный текст.....	153
Кодировка файла.....	154
Размер файла.....	156
Командная оболочка и инструменты командной строки.....	159
Форма и гранулярность таблицы.....	163
Гранулярность данных о проверках и нарушениях в ресторанах.....	164
Форма и гранулярность исследования DAWN.....	166
Краткие выводы.....	168
Глава 9. Первичная обработка датафрейма.....	170
Пример: первичная обработка результатов измерений содержания CO ₂ в обсерватории Мауна-Лоа.....	171
Проверка качества.....	174
Обработка недостающих данных.....	176
Изменение формы таблицы данных.....	177

Проверка качества данных.....	178
Качество в плане охвата	179
Качество измерений и регистрируемых значений	179
Качество в плане связанных признаков	180
Проверка качества на пригодность к исследованию.....	181
Выяснение необходимости в исправлении данных	182
Пропущенные значения и записи.....	183
Преобразования и временные метки.....	185
Преобразование временных меток	186
Конвейеризация в преобразованиях	188
Изменение структуры.....	190
Пример: первичная обработка данных о нарушениях правил безопасности в ресторанах	192
Сужение фокуса.....	193
Агрегирование данных о нарушениях.....	194
Извлечение информации из описания нарушений	196
Краткие выводы.....	199
Глава 10. Разведочный анализ данных.....	201
Типы признаков данных	202
Пример: породы собак	204
Преобразование качественных признаков	210
Переразметка категорий.....	210
Сворачивание категорий	211
Преобразование количественных значений в порядковые	212
Роль типов признаков	213
На что обратить внимание в распределении.....	214
Что необходимо выяснить во взаимосвязи	218
Два количественных признака	218
Один качественный и один количественный признак	219
Два качественных признака	221
Сравнения в многомерных системах	223
Руководящие принципы разведочного анализа	226
Пример: цены на жилую недвижимость.....	227
Изучение цен	228
Дальнейшие шаги.....	230
Изучение прочих признаков.....	231
Углубленный анализ взаимосвязей	234
Фиксация местоположения	236
Результаты EDA	238
Краткие выводы.....	239
Глава 11. Визуализация данных.....	240
Выбор масштаба для выяснения структуры.....	240
Заполнение области данных	241
Учет нулевого значения.....	242
Выяснение формы данных с помощью преобразований.....	244
Кренение для расшифровки взаимосвязей.....	246
Выявление взаимосвязей с помощью спрямления	247
Сглаживание и агрегирование данных	249
Методы сглаживания для определения формы данных.....	250

Методы сглаживания для выявления взаимосвязей и тенденций.....	252
Настройка методов сглаживания	254
Сведение распределений к квантилям.....	255
Случаи, когда сглаживание нежелательно	257
Упрощение значимых сравнений.....	259
Подчеркивание важного различия.....	259
Упорядочивание групп	261
Отказ от стекинга	263
Выбор цветовой палитры.....	265
Принципы проведения сравнений на графиках.....	266
Учет особенностей исходных данных при визуализации.....	267
Данные, собранные с течением времени.....	268
Исследования по данным наблюдений.....	269
Неравномерная выборка	271
Географические данные.....	272
Добавление контекста	273
Пример: результаты 100-метрового спринта	273
Создание графиков с помощью <i>plotly</i>	275
Объекты <i>Figure</i> и <i>Trace</i>	276
Изменение макета.....	277
Функции построения графиков	279
Аннотации к изображению.....	281
Другие инструменты визуализации	282
<i>Matplotlib</i>	282
Грамматика графики	282
Краткие выводы.....	283

Глава 12. Тематическое исследование: проверка точности показателей

качества воздуха	285
Постановка вопроса, структура и охват данных.....	286
Поиск близко расположенных датчиков	288
Первичная обработка списка локаций AQS.....	288
Первичная обработка списка локаций PurpleAir	291
Сопоставление датчиков AQS и PurpleAir	292
Первичная обработка и очистка данных датчиков AQS	294
Проверка гранулярности.....	295
Удаление ненужных столбцов	296
Проверка достоверности дат	297
Проверка качества показателей PM2.5	298
Первичная обработка показаний датчиков PurpleAir	299
Проверка гранулярности.....	300
Визуализация временных меток.....	302
Проверка частоты выборки.....	303
Обработка пропущенных значений	305
Разведочный анализ показаний PurpleAir и AQS	306
Создание модели для корректировки показаний PurpleAir	312
Краткие выводы.....	314

ЧАСТЬ IV. ДРУГИЕ ИСТОЧНИКИ ДАННЫХ.....	317
Глава 13. Операции с текстом.....	319
Примеры текстов и заданий.....	319
Преобразование текста в стандартный формат	319
Извлечение фрагмента текста для создания признака	320
Преобразование текста в признаки	320
Анализ текста.....	321
Манипуляции со строками.....	322
Преобразование текста в стандартный формат с помощью строковых методов Python	322
Строковые методы в <i>pandas</i>	323
Извлечение фрагментов текста с помощью разделения строк.....	324
Регулярные выражения	325
Конкатенация литералов.....	325
Классы символов.....	326
Символ подстановки.....	327
Инвертированные классы символов.....	327
Сокращения классов символов	327
Анкеры и границы	327
Исключение метасимволов	328
Квантификаторы.....	328
Создание признаков при помощи чередования и группировки	330
Справочные таблицы	331
Анализ текста.....	333
Краткие выводы	338
Глава 14. Обмен данными	339
Формат NetCDF	339
Формат JSON	344
HTTP	348
REST	352
XML, HTML, XPath	356
Пример: веб-скрапинг результатов забегов из Википедии.....	359
XPath.....	361
Пример: доступ к курсам валют ЕЦБ	363
Краткие выводы	366
ЧАСТЬ V. ЛИНЕЙНОЕ МОДЕЛИРОВАНИЕ	369
Глава 15. Линейные модели	371
Простая линейная модель	372
Пример: простая линейная модель оценки качества воздуха.....	375
Интерпретация линейных моделей.....	377
Оценка качества подгонки.....	378
Подгонка простой линейной модели	379
Модель множественной линейной регрессии	381
Подбор параметров модели множественной линейной регрессии	386
Пример: где находится страна возмозможностей?.....	389
Объяснение восходящей мобильности на основе времени в пути на работу	391
Связь восходящей мобильности с использованием нескольких переменных	393

Конструирование признаков для числовых измерений	397
Конструирование признаков для категориальных измерений	401
Краткие выводы	408
Глава 16. Выбор модели	410
Переобучение	411
Пример: энергопотребление	411
Метод <i>train_test_split</i>	416
Перекрестная проверка	420
Регуляризация	425
Смещение и дисперсия модели	426
Краткие выводы	429
Глава 17. Теория логического вывода и прогнозирования	431
Распределения: популяционное, эмпирическое, выборочное	431
Принципы проверки гипотез	433
Пример: ранговый критерий сравнения продуктивности соавторов Википедии	435
Пример: проверка эффективности вакцины с помощью пропорций	439
Бутстрап-процедура построения выводов	441
Доверительный интервал	446
Интервал прогнозирования	449
Пример: прогнозирование опоздания автобуса	449
Пример: прогнозирование размера краба	450
Пример: прогнозирование прироста краба	451
Вероятность выводов и прогнозов	454
Формализация теории для статистик средних рангов	454
Общие свойства случайных величин	457
Вероятность в основе интервалов и тестирования	459
Вероятность в основе выбора модели	462
Краткие выводы	464
Глава 18. Тематическое исследование: как взвесить осла	466
Постановка вопроса и охват данных	466
Первичная обработка и преобразование данных	467
Разведочный анализ данных	472
Моделирование веса осла	475
Функция потерь при назначении анестетиков	475
Построение простой линейной модели	476
Подгонка множественной линейной модели	478
Добавление в модель качественных признаков	479
Оценка модели	482
Краткие выводы	484
ЧАСТЬ VI. КЛАССИФИКАЦИЯ	487
Глава 19. Классификация	489
Пример: поваленные ураганом деревья	489
Моделирование и классификация	492
Константная модель	492
Исследование взаимосвязи между буреломом и размером деревьев	493

Моделирование долей (и вероятностей).....	495
Логистическая модель.....	496
Логарифм отношения шансов	497
Применение логистической кривой.....	498
Функция потерь для логистической модели	499
От вероятностей к классификации.....	502
Матрица несоответствий	504
Точность и полнота	505
Краткие выводы.....	508
Глава 20. Численная оптимизация	509
Основы градиентного спуска	510
Минимизация потери Хубера.....	512
Выпуклые и дифференцируемые функции потерь.....	514
Варианты градиентного спуска.....	515
Стохастический градиентный спуск.....	516
Мини-пакетный градиентный спуск.....	517
Метод Ньютона	517
Краткие выводы.....	518
Глава 21. Тематическое исследование: распознавание фейковых новостей	520
Уточнение вопроса и выяснение охвата данных	521
Получение и "выпас" данных	522
Разведочный анализ данных.....	526
Разведочный анализ источников публикаций	527
Разведочный анализ даты публикации.....	529
Разведочный анализ слов в статье	530
Моделирование	532
Однословная модель	532
Многословная модель	534
Прогнозирование с помощью преобразования TF-IDF	536
Краткие выводы.....	539
Приложение 1. Дополнительный материал	541
Приложение 2. Источники данных.....	548
Предметный указатель.....	553
Об авторах.....	558
Об изображении на обложке.....	559