

Я хотел бы поблагодарить моих коллег из BlueGranite. Это большая честь – быть частью команды умных и талантливых людей. Железо точит железо. Также я благодарен окружению, в котором вырос – на западе Мичигана в штате Индиана. Опыт, полученный там, для меня поистине бесценен. Хочу сказать спасибо Патрику Леблану (Patrick Leblanc), Мико Юку (Mico Yuk), Терри Моррису (Terry Morris), доктору Брандейсу Маршаллу (Dr. Brandeis Marshall) и доктору Сайдику Вотсон (Dr. Sydeaka Watson). Вы обладаете качествами, которые я очень ценю, и я постоянно учился на ваших примерах.

Кроме того, хотелось бы выразить благодарность всем моим напарникам по командам, начиная с низших лиг и заканчивая колледжем. Спорт был одним из лучших моих учителей в жизни, и мне выпала большая честь пройти через все это вместе с вами. Я с огромным уважением и почтением отношусь ко многим из вас. Спасибо моим бывшим тренерам, особенно в старшей школе и колледже. Вызовы, которые вы передо мной ставили, закалили мой характер и позволили стать лучше во многих аспектах жизни. И за это я вам очень признателен.

Также я безмерно благодарен моей семье. Вырастить ребенка очень непросто, и я бесконечно ценю помощь, которую мне оказывали мои родственники. Спасибо вам, Тим Адамс младший и маленькая Камилла. Вы обладаете редким сочетанием высокого интеллекта и простоты в общении. Совсем скоро мы поменяемся ролями: вы будете учить, а я – учиться.

Хочу отметить признанием моих братьев в Мичигане и братьев, с которыми познакомился, играя в футбол в Луисвилле. Мы вместе прошли через многое. Что бы ни случилось, мы останемся братьями навсегда. Хотел бы сказать спасибо и моим двоюродным братьям и сестрам. Мы выросли вместе, у нас много общих воспоминаний. Многие из вас поддерживали меня в не самые лучшие времена. Я этого никогда не забуду. Спасибо моим родным братьям и сестрам: Полетте (мир ее праху), Стефани, Тине и Люку, а также моему племяннику Джунбагу, который был мне как брат. Когда мы нуждались друг в друге, мы всегда оказывались рядом. Пусть так будет и дальше. Наконец, я благодарен моим родителям: Лютеру и Эрнестине Уэйд. Я очень ценю вашу любовь, поддержку и жертвенность во имя своих детей. Как сказал бы папа: «Я люблю вас, и это навсегда!»

Последним по порядку, но не по значению, я хотел бы сказать спасибо Всевышнему. Я благодарен за все способности и таланты, которыми Ты меня наделил. Обещаю применить их по назначению, чтобы принести максимальную пользу обществу.

Содержание

От издательства.....	20
Об авторе.....	21
О техническом редакторе.....	22
Благодарности	23
Введение.....	24

Часть I. СОЗДАНИЕ ПОЛЬЗОВАТЕЛЬСКОЙ ВИЗУАЛИЗАЦИИ ПРИ ПОМОЩИ R	41
---	----

Глава 1. Грамматика графиков	42
---	----

Пошаговое создание визуализации в Power BI при помощи R.....	43
Шаг 1. Настройте Power BI	43
Шаг 2. Перенесите визуальный элемент R в рабочую область Power BI.....	43
Шаг 3. Определитесь с набором данных.....	43
Шаг 4. Спроектируйте визуальный элемент в среде разработки R.....	44
Шаг 5. Используйте следующий шаблон для разработки элемента на R	45
Шаг 6. Добавьте скрипту функциональности	45
Рекомендованные шаги по созданию визуального элемента на R при помощи ggplot2.....	46
Шаг 1. Импортируйте нужные для скрипта пакеты	46
Шаг 2. Выполните необходимое преобразование исходных данных	47
Шаг 3. Создайте визуализацию при помощи функции ggplot()	48
Шаг 4. Добавьте нужные геометрии.....	48
Шаг 5. Определите заголовки, подзаголовки и подписи	50
Шаг 6. Приведите в порядок оси.....	50
Шаг 7. Примените тему при необходимости	53
Шаг 8. Используйте функцию theme() для настройки оформления	54
Дополнительный шаг: задайте цвета точек на диаграмме рассеяния	55
Важность оперирования «чистыми» данными.....	58
Популярные геометрии.....	58
Управление эстетиками через шкалы	64
Встроенные в пакет ggplot2 темы.....	65
Использование визуальных элементов R в службе Power BI.....	66
Вспомогательные пакеты ggplot2.....	66
Заключение	67

Глава 2. Создание пользовательских визуализаций на R в Power BI при помощи ggplot2	68
Диаграмма с аннотацией	69
Шаг 1. Загрузите исходные данные.....	70
Шаг 2. Создайте срез на основании года на панели фильтров.....	71
Шаг 3. Настройте визуальный элемент R в Power BI.....	71
Шаг 4. Экспортируйте данные в R Studio для дальнейшей разработки.....	72
Шаг 5. Загрузите необходимые пакеты.....	73
Шаг 6. Создайте переменные для проверки данных.....	73
Шаг 7. Выполните проверку данных.....	74
Шаг 8. Добавьте столбцы к набору данных, необходимые для нашей визуализации.....	75
Шаг 9. Создайте переменные для динамических составляющих диаграммы.....	76
Шаг 10. Постройте диаграмму при помощи функции ggplot().....	78
Шаг 11. Добавьте слой со столбчатой диаграммой на визуальный элемент.....	78
Шаг 12. Добавьте текстовый слой на визуальный элемент.....	79
Шаг 13. Измените ось y.....	80
Шаг 14. Преобразуйте вертикальную столбчатую диаграмму в горизонтальную.....	81
Шаг 15. Добавьте на диаграмму динамическую аннотацию.....	81
Шаг 16. Добавьте динамические заголовки и подпись на визуальный элемент.....	83
Шаг 17. Удалите метки с осей x и y.....	84
Шаг 18. Удалите легенду с диаграммы.....	84
Шаг 19. Измените внешний вид диаграммы при помощи темы theme_few().....	85
Шаг 20. Расположите заголовки по центру.....	86
Шаг 21. Перенесите код в Power BI.....	87
Пузырьковая диаграмма	89
Шаг 1. Загрузите исходные данные.....	90
Шаг 2. Загрузите данные в Power BI.....	90
Шаг 3. Создайте срез на основании года.....	90
Шаг 4. Выполните базовую настройку визуального элемента R.....	91
Шаг 5. Экспортируйте данные в R Studio для разработки элемента.....	91
Шаг 6. Загрузите требуемые пакеты.....	91
Шаг 7. Создайте переменные для проверки данных.....	91
Шаг 8. Создайте код для проверки.....	92
Шаг 9. Определите цвета для конференций и дивизионов.....	92
Шаг 10. Динамически определите заголовок диаграммы.....	93
Шаг 11. Создайте набор данных для диаграммы.....	93
Шаг 12. Создайте диаграмму при помощи функции ggplot.....	94
Шаг 13. Добавьте слой для пузырьковой диаграммы при помощи геометрии geom_point.....	94
Шаг 14. Добавьте метки на диаграмму.....	96

Шаг 15. Измените цвет границ и заливок пузырьков на диаграмме	97
Шаг 16. Создайте заголовок диаграммы	98
Шаг 17. Задайте тему.....	98
Шаг 18. Перенесите код в Power BI	98
Визуализация прогнозирования	99
Шаг 1. Загрузите исходные данные	101
Шаг 2. Создайте срез по кватербекам на панели фильтров	102
Шаг 3. Настройте визуальный элемент R в Power BI.....	102
Шаг 4. Экспортируйте данные в R Studio для дальнейшей разработки.....	102
Шаг 5. Загрузите необходимые пакеты	102
Шаг 6. Создайте переменные для проверки данных.....	103
Шаг 7. Выполните проверку данных	103
Шаг 8. Создайте динамический заголовок для визуализации.....	104
Шаг 9. Создайте набор данных, необходимый для составления прогноза	104
Шаг 10. Постройте прогноз.....	105
Шаг 11. Постройте диаграмму.....	105
Шаг 12. Перенесите код в Power BI	106
Линейная диаграмма с затенением	107
Шаг 1. Загрузите исходные данные	109
Шаг 2. Загрузите данные в Power BI	110
Шаг 3. Создайте срезы в отчете	111
Шаг 4. Настройте визуальный элемент R в Power BI.....	111
Шаг 5. Экспортируйте данные в R Studio для дальнейшей разработки.....	112
Шаг 6. Загрузите необходимые пакеты	112
Шаг 7. Создайте переменные для проверки данных	112
Шаг 8. Выполните проверку данных	113
Шаг 9. Создайте новый датафрейм на основании датафрейма dataset.....	113
Шаг 10. Создайте переменные для динамических составляющих диаграммы.....	114
Шаг 11. Создайте наборы данных, необходимые для наложения тени	114
Шаг 12. Создайте наборы данных, необходимые для отрисовки графика.....	118
Шаг 13. Создайте символьный вектор для хранения цветовой схемы затенения	118
Шаг 14. Постройте диаграмму при помощи функции ggplot()	119
Шаг 15. Добавьте слой для создания затенения	119
Шаг 16. Добавьте линейную диаграмму на основании выбора пользователя	120
Шаг 17. Раскрасьте фоновую заливку в соответствии с предопределенной цветовой схемой партий	121
Шаг 18. Отформатируйте ось y в соответствии с выбором пользователя	121
Шаг 19. Добавьте метки на оси x и y.....	121
Шаг 20. Снабдите диаграмму динамическим заголовком и подзаголовком	122
Шаг 21. Измените внешний вид диаграммы в стиле журнала The Economist	122

Шаг 22. Перенесите код в Power BI	122
Карта	123
Шаг 1. Загрузите исходные данные	125
Шаг 2. Загрузите данные в Power BI	126
Шаг 3. Создайте срез в отчете на основании выбранного в фильтре штата.....	127
Шаг 4. Настройте визуальный элемент R в Power BI.....	127
Шаг 5. Экпортируйте данные в R Studio для дальнейшей разработки.....	127
Шаг 6. Загрузите необходимые пакеты	127
Шаг 7. Создайте переменные для проверки данных	127
Шаг 8. Выполните проверку данных	128
Шаг 9. Создайте переменные для заголовков диаграммы	128
Шаг 10. Добавьте к набору данных столбец с квинтилем	129
Шаг 11. Создайте символьный вектор для хранения цветовой схемы затенения.....	129
Шаг 12. Постройте диаграмму при помощи функции ggplot()	130
Шаг 13. Добавьте слой с картой.....	130
Шаг 14. Отформатируйте оси <i>x</i> и <i>y</i>	131
Шаг 15. Раскрасьте округа на основании квинтилей.....	131
Шаг 16. Улучшите отображение карты выбранного штата.....	133
Шаг 17. Снабдите диаграмму динамическим заголовком и подзаголовком	133
Шаг 18. Примените тему theme_map()	133
Шаг 19. Перенесите код в Power BI	133
Диаграмма квадрантов	134
Шаг 1. Загрузите исходные данные	137
Шаг 2. Загрузите данные в Power BI	138
Шаг 3. Создайте срезы в отчете по типу игры и четверти матча	138
Шаг 4. Настройте визуальный элемент R в Power BI.....	138
Шаг 5. Экпортируйте данные в R Studio для дальнейшей разработки.....	138
Шаг 6. Загрузите необходимые пакеты	139
Шаг 7. Создайте переменные для проверки данных	139
Шаг 8. Выполните проверку данных	140
Шаг 9. Создайте заголовки диаграммы	140
Шаг 10. Добавьте дополнительные столбцы в набор данных	140
Шаг 11. Постройте диаграмму при помощи функции ggplot()	141
Шаг 12. Используйте геометрию geom_point() для создания диаграммы рассеяния	141
Шаг 13. Добавьте метки игроков для всех квадрантов.....	141
Шаг 14. Добавьте горизонтальные и вертикальные линии, проходящие через центр	142
Шаг 15. Добавьте на диаграмму заголовки квадрантов	143
Шаг 16. Добавьте метки на оси <i>x</i> и <i>y</i>	145
Шаг 17. Снабдите диаграмму динамическими заголовками и подписями	145
Шаг 18. Примените тему theme_tufte.....	145
Шаг 19. Выполните финальную очистку.....	145

Шаг 20. Перенесите код в Power BI	148
Добавление линии регрессии.....	149
Шаг 1. Загрузите исходные данные	150
Шаг 2. Загрузите данные в Power BI	151
Шаг 3. Настройте визуальный элемент R в Power BI.....	151
Шаг 4. Экпортируйте данные в R Studio для дальнейшей разработки.....	151
Шаг 5. Загрузите необходимые пакеты	151
Шаг 6. Создайте переменные для проверки данных.....	151
Шаг 7. Выполните проверку данных	152
Шаг 8. Постройте диаграмму при помощи функции ggplot()	152
Шаг 9. Используйте геометрию geom_point() для создания диаграммы рассеяния	153
Шаг 10. Добавьте на визуализацию слой с линией регрессии	153
Шаг 11. Снабдите диаграмму заголовком и подзаголовком	154
Шаг 12. Примените тему	155
Шаг 13. Выполните финальную очистку.....	155
Шаг 14. Перенесите код в Power BI	155

Часть II. ЗАГРУЗКА ИНФОРМАЦИИ В МОДЕЛЬ ДАННЫХ POWER BI ПРИ ПОМОЩИ R И PYTHON

Глава 3. Чтение файлов CSV.....

Динамическое объединение файлов	158
Пример сценария.....	159
Выбор файлов за скользящий период из 24 месяцев при помощи R	159
Шаг 1. Импортируйте необходимые пакеты для скрипта	159
Шаг 2. Установите рабочую директорию на папку, содержащую наборы данных о продажах	160
Шаг 3. Считайте имена файлов в символьный вектор.....	161
Шаг 4. Создайте вектор дат	162
Шаг 5. Создайте датафрейм, состоящий из двух векторов.....	163
Шаг 6. Получите верхнюю и нижнюю границы желаемого диапазона дат.....	164
Шаг 7. Ограничьте датафрейм только нужными нам месяцами	164
Шаг 8. Создайте датафрейм на основании объединенных файлов	165
Шаг 9. Соберите написанный код и перенесите в редактор скриптов в Power BI.....	166
Выбор файлов за скользящий период из 24 месяцев при помощи Python	168
Шаг 1. Создайте скрипт на Python и загрузите необходимые библиотеки	168
Шаг 2. Установите рабочую директорию на папку Python_Code	168
Шаг 3. Загрузите перечень имен файлов в список	169
Шаг 4. Создайте датафрейм pandas с информацией о файлах для объединения.....	169

Шаг 5. Создайте новый столбец с датой в датафрейме.....	170
Шаг 6. Определите границы нужного нам диапазона дат.....	170
Шаг 7. Ограничьте датафрейм нужным диапазоном	170
Шаг 8. Объедините файлы в единый датафрейм	171
Шаг 9. Перенесите скрипт в Power BI.....	172
Фильтрация строк на основе регулярных выражений	174
Использование регулярных выражений в R	174
Шаг 1. Загрузите необходимые для работы пакеты	175
Шаг 2. Загрузите в R файл с потенциальными избирателями.....	175
Шаг 3. Определите регулярное выражение	175
Шаг 4. Исключите неправильные адреса из набора данных.....	176
Шаг 5. Объедините написанный код в один скрипт и перенесите в редактор скриптов в Power BI.....	176
Использование регулярных выражений в Python	176
Шаг 1. Загрузите необходимые для работы библиотеки	177
Шаг 2. Загрузите в Python файл с избирателями и присвойте его содержимое датафрейму	177
Шаг 3. Определите регулярное выражение	177
Шаг 4. Исключите неправильные адреса из набора данных.....	177
Шаг 5. Объедините написанный код в один скрипт и перенесите в редактор скриптов Python в Power BI.....	177

Глава 4. Чтение данных из Microsoft Excel..... 179

Чтение файлов Excel при помощи R	180
Шаг 1. Импортируйте пакеты tidyverse и readxl.....	180
Шаг 2. Создайте оболочку функции combine_sheets.....	181
Шаг 3. Получите имена листов для объединения из указанной рабочей книги	181
Шаг 4. Преобразуйте символьный вектор, полученный на предыдущем шаге, в именованный символьный вектор.....	181
Шаг 5. Используйте функцию map_dfr() для объединения информации с листов в один датафрейм	182
Шаг 6. Верните датафрейм из функции	183
Шаг 7. Направьте рабочую директорию на папку с файлами Excel.....	183
Шаг 8. Сохраните в переменной excel_file_paths список файлов для обработки.....	184
Шаг 9. Используйте функцию map_dfr() для применения функции combine_sheets() ко всем выбранным файлам	184
Шаг 10. Скопируйте скрипт и вставьте в редактор скриптов R в Power BI через инструмент Получить данные (GetData)	184
Чтение файлов Excel при помощи Python.....	185
Шаг 1. Импортируйте библиотеки os и pandas	186
Шаг 2. Создайте оболочку функции combine_sheets()	186
Шаг 3. Создайте объект Excel на основании пути, переданного в функцию в аргументе excel_file_path	186
Шаг 4. Создайте список имен листов в рабочей книге.....	186

Шаг 5. Используйте метод <code>read_excel()</code> из библиотеки <code>pandas</code> для считывания данных в один датафрейм	187
Шаг 6. Верните датафрейм <code>df</code> из функции <code>combine_sheets</code>	187
Шаг 7. Установите рабочую директорию в папку, в которой находятся файлы Excel.....	187
Шаг 8. Получите список файлов в текущей рабочей директории и присвойте его переменной <code>excel_file_paths</code>	188
Шаг 9. Создайте пустой датафрейм и назовите его <code>combined_workbooks</code> ...	188
Шаг 10. Создайте заготовку для цикла <code>for</code>	188
Шаг 11. Объедините данные со всех листов в один датафрейм при помощи функции <code>combine_sheets()</code>	189
Шаг 12. Добавьте датафрейм <code>combined_workbook</code> к главному датафрейму <code>combined_workbooks</code>	189
Шаг 13. Скопируйте скрипт и вставьте в редактор скриптов Python в Power BI через инструмент Получить данные (GetData)	190
Глава 5. Чтение данных из SQL Server	191
Добавление базы данных AdventureWorksDW_StarSchema к вашему экземпляру SQL Server	191
Чтение данных из SQL Server в Power BI при помощи R	192
Шаг 1. Создайте DSN для подключения к базе данных SQL Server.....	193
Шаг 2. Создайте таблицу лога в SQL Server	196
Шаг 3. Начните написание скрипта на R для загрузки таблицы DimDate	197
Шаг 4. Создайте переменную для хранения имени загружаемой таблицы	197
Шаг 5. Создайте переменную для хранения SQL-выражения.....	197
Шаг 6. Создайте подключение к SQL Server	197
Шаг 7. Извлеките данные из SQL Server и сохраните их в датафрейм.....	198
Шаг 8. Получите текущее время.....	198
Шаг 9. Получите количество прочитанных записей.....	198
Шаг 10. Добавьте в датафрейм запись с информацией для сохранения в лог.....	198
Шаг 11. Сохраните собранную информацию в базе данных	199
Шаг 12. Закройте соединение	200
Шаг 13. Скопируйте написанный скрипт в Power BI	200
Шаг 14. Создайте скрипт для загрузки таблицы DimProduct на базе ReadLog_DimDate.R.....	201
Шаг 15. Создайте скрипт для загрузки таблицы DimPromotion.....	202
Шаг 16. Создайте скрипт для загрузки таблицы DimSalesTerritory на основе ReadLog_DimDate.R.....	202
Шаг 17. Создайте скрипт для загрузки таблицы FactInternetSales на основе ReadLog_DimDate.R.....	203
Чтение данных из SQL Server в Power BI при помощи Python	204
Шаг 1. Создайте DSN для SQL Server	204
Шаг 2. Создайте таблицу для ведения логов в SQL Server.....	204
Шаг 3. Создайте скрипт для загрузки таблицы DimDate.....	205

Шаг 4. Определите переменную для хранения имени таблицы, предназначенной для загрузки в Power BI.....	205
Шаг 5. Создайте подключение к базе данных с помощью библиотеки sqlalchemy	205
Шаг 6. Прочитайте содержимое таблицы DimDate и сохраните его в переменной df_read	206
Шаг 7. Получите текущую дату и время и сохраните в переменной timestamp.....	206
Шаг 8. Посчитайте количество записей в таблице DimDate.....	206
Шаг 9. Добавьте запись в датафрейм с информацией для сохранения логов	207
Шаг 10. Добавьте информацию, добытую на предыдущем шаге, в таблицу логов	207
Шаг 11. Скопируйте скрипт в Power BI	208
Шаг 12. Создайте скрипт для загрузки таблицы DimProduct на основе ReadLog_DimDate.py	208
Шаг 13. Создайте скрипт для загрузки таблицы DimPromotion на основе ReadLog_DimDate.py	209
Шаг 14. Создайте скрипт для загрузки таблицы DimSalesTerritory на основе ReadLog_DimDate.py	210
Шаг 15. Создайте скрипт для загрузки таблицы FactInternetSales на основе ReadLog_DimDate.py	211
Глава 6. Чтение в модель данных Power BI посредством API.....	212
Чтение и загрузка данных в Power BI из API с помощью R.....	212
Шаг 1. Получите персональный ключ API к Census	212
Шаг 2. Загрузите необходимые пакеты R	213
Шаг 3. Определите переменные для возврата из вашего набора данных.....	213
Шаг 4. Создайте символьный вектор, содержащий нужные вам переменные	214
Шаг 5. Сконфигурируйте функцию get_acs	215
Шаг 6. Присвойте переменным (столбцам) осмысленные имена	215
Шаг 7. Скопируйте скрипт в Power BI.....	216
Чтение и загрузка данных в Power BI из API с помощью Python.....	217
Шаг 1. Получите персональный ключ API к Census	217
Шаг 2. Загрузите необходимые библиотеки Python	218
Шаг 3. Определите переменные для возврата из вашего набора данных.....	218
Шаг 4. Создайте список, содержащий нужные вам переменные	219
Шаг 5. Создайте список кортежей с географическими фильтрами для набора данных	220
Шаг 6. Извлеките данные при помощи функции censusdata.download()	220
Шаг 7. Переиндексируйте датафрейм, созданный на шестом шаге.....	221
Шаг 8. Дайте столбцам осмысленные имена	221
Шаг 9. Переименуйте столбцы в датафрейме.....	221
Шаг 10. Скопируйте скрипт в Power BI	222
Заключение	222

Часть III. ПРЕОБРАЗОВАНИЕ ДАННЫХ ПРИ ПОМОЩИ R И PYTHON	223
Глава 7. Продвинутое строковое операции и распознавание шаблонов	224
Защита конфиденциальных сведений	225
Защита конфиденциальных сведений в Power BI с помощью R	225
Шаг 1. Импортируйте пакеты tidyverse и stringr.....	225
Шаг 2. Напишите функцию для очистки данных.....	226
Шаг 3. Считайте комментарии в датафрейм	228
Шаг 4. Скройте телефонные номера и номера социального страхования в поле комментария.....	228
Шаг 5. Скопируйте скрипт в Power BI	229
Защита конфиденциальных сведений в Power BI с помощью Python.....	229
Шаг 1. Импортируйте библиотеки pandas, os и re	230
Шаг 2. Напишите функцию mask_text()	230
Шаг 3. Установите рабочую директорию.....	232
Шаг 4. Считайте комментарии в датафрейм	232
Шаг 5. Выполните замену телефонных номеров и номеров социального страхования.....	232
Шаг 6. Скопируйте скрипт в Power BI	232
Подсчет количества слов и предложений в обзорах.....	233
Подсчет количества слов и предложений в обзорах с помощью R	233
Шаг 1. Импортируйте библиотеки tidyverse и stringr.....	233
Шаг 2. Измените рабочую директорию	234
Шаг 3. Считайте информацию из файла	234
Шаг 4. Ограничьте набор данных требуемыми столбцами.....	234
Шаг 5. Добавьте столбцы с количеством слов и предложений	234
Шаг 6. Скопируйте скрипт в Power BI	235
Подсчет количества слов в обзорах с помощью Python.....	235
Шаг 1. Импортируйте библиотеки pandas и os.....	235
Шаг 2. Установите рабочую директорию.....	236
Шаг 3. Считайте информацию из файла	236
Шаг 4. Создайте в датафрейме столбец word_count	236
Шаг 5. Скопируйте скрипт в Power BI	236
Удаление имен неподходящего формата	237
Удаление имен неподходящего формата с помощью R	237
Шаг 1. Импортируйте пакеты tidyverse и stringr.....	237
Шаг 2. Установите рабочую директорию.....	237
Шаг 3. Создайте регулярное выражение с правильным шаблоном имени	238
Шаг 4. Считайте данные в датафрейм	239
Шаг 5. Выполните обновление столбца Name	239
Шаг 6. Скопируйте скрипт в Power BI	239
Удаление имен неподходящего формата с помощью Python.....	239
Шаг 1. Импортируйте библиотеки pandas, re и os	240

Шаг 2. Установите рабочую директорию.....	240
Шаг 3. Считайте данные из файла DimEmployee.csv в датафрейм	240
Шаг 4. Создайте регулярное выражение, соответствующее правильному формату имени	240
Шаг 5. Скомпилируйте регулярное выражение	241
Шаг 6. Напишите функцию для проверки имен на совместимость с шаблоном.....	241
Шаг 7. Примените функцию к столбцу, чтобы избавиться от лишних имен	242
Шаг 8. Скопируйте скрипт в Power BI	242
Определение шаблонов в строках на основании условной логики	243
Поиск шаблонов в строках на основании условной логики с помощью R.....	244
Шаг 1. Импортируйте пакеты tidyverse и stringr.....	244
Шаг 2. Установите рабочую директорию.....	244
Шаг 3. Напишите функцию для поиска изделий	245
Шаг 4. Считайте данные из файла ProductionOrders.csv в датафрейм....	246
Шаг 5. Добавьте в датафрейм df столбец Monitored Products	246
Шаг 6. Скопируйте скрипт в Power BI	246
Поиск шаблонов в строках на основании условной логики с помощью Python	247
Шаг 1. Импортируйте библиотеки pandas, re и os	247
Шаг 2. Установите рабочую директорию.....	247
Шаг 3. Скомпилируйте регулярное выражение	247
Шаг 4. Напишите функцию для распознавания нужных нам деталей....	248
Шаг 5. Считайте данные в датафрейм Pandas	248
Шаг 6. Создайте новый столбец с именем Monitored Products.....	249
Шаг 7. Скопируйте скрипт в Power BI.....	249
Заключение	249
Глава 8. Вычисляемые столбцы с помощью R и Python	250
Создание ключа Google Geocoding API	251
Шаг 1. Зайдите в консоль Google.....	251
Шаг 2. Настройте учетную запись	251
Шаг 3. Добавьте новый проект.....	251
Шаг 4. Активируйте API геокодирования.....	252
Геокодирование адресов с помощью R	254
Геокодирование адресов с помощью Python	256
Вычисление расстояния между точками с помощью пользовательской функции в R	258
Вычисление расстояния между точками с помощью пользовательской функции в Python.....	260
Вычисление расстояния между точками с помощью готовой функции в R...	263
Вычисление расстояния между точками с помощью готовой функции в Python.....	264
Заключение	266

Часть IV. МАШИННОЕ ОБУЧЕНИЕ И ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ В POWER BI ПРИ ПОМОЩИ R И PYTHON267

Глава 9. Применение методов машинного обучения и искусственного интеллекта к моделям данных Power BI..... 268

Применение алгоритмов машинного обучения к набору данных перед загрузкой в модель Power BI.....	269
Прогнозирование цен на недвижимость с помощью R.....	270
Шаг 1. Пусть аналитик данных сохранит для вас модель.....	270
Шаг 2. Загрузите пакет tidyverse.....	270
Шаг 3. Загрузите объект модели и набор данных для оценки.....	271
Шаг 4. Ограничьте датафрейм столбцами, необходимыми для вашей модели.....	271
Шаг 5. Примените модель машинного обучения к своему набору данных для составления прогноза цен на недвижимость.....	271
Шаг 6. Добавьте прогноз к исходному набору данных.....	272
Шаг 7. Скопируйте скрипт в Power BI.....	272
Прогнозирование цен на недвижимость с помощью Python.....	272
Шаг 1. Пусть аналитик данных сохранит для вас модель.....	273
Шаг 2. Загрузите необходимые библиотеки.....	273
Шаг 3. Загрузите объект модели и актуальный набор данных.....	273
Шаг 4. Извлеките нужную информацию из датафрейма.....	274
Шаг 5. Примените модель к подготовленному набору данных для расчета прогноза.....	274
Шаг 6. Добавьте прогнозные данные к исходному набору данных.....	275
Шаг 7. Скопируйте скрипт в Power BI.....	275
Использование готовых моделей ИИ для расширения функционала моделей данных в Power BI.....	276
Настройка Cognitive Services в Azure.....	277
Виртуальная машина для анализа данных (Data Science Virtual Machine – DSVM).....	277
Анализ тональности текста в Microsoft Cognitive Services при помощи Python.....	278
Шаг 1. Загрузите набор данных с отзывами Yelp с сайта Kaggle.....	279
Шаг 2. Загрузите нужные библиотеки, модули и функции для скрипта.....	279
Шаг 3. Инициализируйте переменные для работы скрипта.....	280
Шаг 4. Считайте фрагмент файла с отзывами в датафрейм.....	280
Шаг 5. Преобразуйте датафрейм к формату, приемлемому для службы Microsoft Cognitive Services.....	281
Шаг 6. Оцените отзывы посетителей при помощи метода sentiment().....	281
Шаг 7. Создайте датафрейм, содержащий оценки отзывов.....	282
Шаг 8. Скопируйте скрипт в Power BI.....	282
Применение сторонних моделей машинного обучения к моделям данных Power BI.....	283

Конфигурирование средства анализа настроения текста в IBM Watson.....	284
Шаг 1. Заведите аккаунт в IBM Cloud.....	284
Шаг 2. Выполните вход в IBM Cloud.....	284
Шаг 3. Перейдите на страницу Tone Analyzer.....	284
Шаг 4. Настройте службу Tone Analyzer.....	284
Шаг 5. Получите ключ API.....	285
Написание скрипта на Python для анализа настроения текста в IBM Watson.....	286
Шаг 1. Импортируйте необходимые библиотеки и модули.....	286
Шаг 2. Создайте экземпляр класса IAMAuthenticator.....	286
Шаг 3. Создайте экземпляр класса ToneAnalyzerV3.....	286
Шаг 4. Установите ссылку на службу для созданного объекта.....	287
Шаг 5. Создайте датафрейм с исходными данными для анализа.....	287
Шаг 6. Создайте основу для датафрейма с оценочными данными.....	287
Шаг 7. Определите циклическую конструкцию для отправки документов в службу IBM Watson.....	288
Шаг 8. Отформатируйте и оцените настроение текста в документе.....	288
Шаг 9. Извлеките результат анализа текста и инициализируйте переменные исходными значениями.....	289
Шаг 10. Пройдите по тонам и присвойте их значения соответствующим переменным.....	290
Шаг 11. Создайте датафрейм на основе списка listReturnedUtterance.....	291
Шаг 12. Объедините датафреймы dfReturnedUtterance и dfDocuments.....	291
Шаг 13. Скопируйте скрипт в Power BI.....	292

Глава 10. Создание моделей анализа данных и скриптов для обработки информации.....

Прогнозирование цен на недвижимость в Power BI с помощью R со службой SSMLS.....	294
Написание скрипта на языке R для добавления модели в SQL Server.....	295
Шаг 1. Загрузите необходимые пакеты.....	296
Шаг 2. Загрузите модель R в вашу сессию.....	296
Шаг 3. Подключитесь к базе данных.....	296
Шаг 4. Определите переменные модели.....	296
Шаг 5. Напишите выражение на T-SQL для добавления модели в базу данных.....	297
Шаг 6. Добавьте код, необходимый для запуска выражения T-SQL из R.....	297
Шаг 7. Сохраните скрипт.....	298
Использование SSMLS совместно с R для оценки данных.....	298
Шаг 1. Запустите SQL Server Management Studio.....	298
Шаг 2. Создайте подключение к серверу, который хотите использовать.....	298
Шаг 3. Добавьте базу данных BostonHousingInfo на ваш сервер.....	298
Шаг 4. Добавьте модель в базу данных.....	299
Шаг 5. Создайте в базе данных хранимую процедуру для прогноза.....	299

Шаг 6. Извлеките данные с прогнозами из SQL Server в Power BI.....	301
Прогнозирование цен на недвижимость в Power BI с помощью Python со службой SSMLS.....	303
Написание скрипта на языке Python для добавления модели в SQL Server.....	303
Шаг 1. Подберите версии библиотек.....	303
Шаг 2. Создайте окружение conda.....	304
Шаг 3. Напишите код для загрузки модели в SQL Server.....	305
Использование SSMLS совместно с Python для оценки данных.....	307
Шаг 1. Запустите SQL Server Management Studio.....	307
Шаг 2. Создайте подключение к серверу, который хотите использовать.....	307
Шаг 3. Добавьте базу данных BostonHousingInfo на ваш сервер.....	307
Шаг 4. Добавьте модель в базу данных.....	307
Шаг 5. Создайте в базе данных хранимую процедуру для прогноза.....	308
Шаг 6. Извлеките данные с прогнозами из SQL Server в Power BI.....	310
Анализ тональности текста в Power BI с помощью R со службой SSMLS.....	311
Добавление готовых моделей R в SSMLS с помощью PowerShell.....	311
Шаг 1. Проверьте, установлены ли предварительно обученные модели.....	311
Шаг 2. Откройте PowerShell от имени администратора.....	312
Шаг 3. Загрузите скрипт PowerShell.....	312
Шаг 4. Запустите загруженный скрипт в PowerShell.....	312
Решение проблем.....	312
Использование готовой модели R в SSMLS для анализа тональности текста в Power BI.....	312
Шаг 1. Определите хранимую процедуру.....	313
Шаг 2. Определите переменные.....	313
Шаг 3. Инициализируйте переменную @Query.....	313
Шаг 4. Инициализируйте переменную @RScript.....	314
Шаг 5. Сконфигурируйте процедуру sp_execute_external_script.....	315
Шаг 6. Определите выходные данные.....	316
Шаг 7. Создайте хранимую процедуру в базе данных.....	316
Шаг 8. Вызовите процедуру из Power BI.....	317
Анализ тональности текста в Power BI с помощью Python со службой SSMLS.....	318
Добавление готовых моделей Python в SSMLS.....	319
Шаг 1. Проверьте, установлены ли предварительно обученные модели.....	319
Шаг 2. Откройте PowerShell от имени администратора.....	319
Шаг 3. Загрузите скрипт PowerShell.....	319
Шаг 4. Запустите загруженный скрипт в PowerShell.....	319
Решение проблем.....	320
Использование готовой модели Python в SSMLS для анализа тональности текста в Power BI.....	320
Шаг 1. Определите хранимую процедуру.....	320
Шаг 2. Определите переменные.....	321

Шаг 3. Инициализируйте переменную @Query	321
Шаг 4. Инициализируйте переменную @PythonScript.....	321
Шаг 5. Сконфигурируйте процедуру sp_execute_external_script	322
Шаг 6. Определите выходные данные	322
Шаг 7. Создайте хранимую процедуру в базе данных.....	322
Шаг 8. Вызовите процедуру из Power BI.....	323
Вычисление расстояния между точками в Power BI с помощью R со службой SSMLS.....	324
Шаг 1. Убедитесь, что в SSMLS загружен пакет dplyr.....	325
Шаг 2. Запустите SSMS и подключитесь к SQL Server	325
Шаг 3. Добавьте базу данных CalculateDistance на ваш сервер.....	325
Шаг 4. Создайте хранимую процедуру для расчета расстояний.....	326
Шаг 5. Вызовите процедуру из Power BI	328
Вычисление расстояния между точками в Power BI с помощью Python со службой SSMLS.....	329
Шаг 1. Запустите SSMS и подключитесь к SQL Server	330
Шаг 2. Добавьте базу данных CalculateDistance на ваш сервер.....	330
Шаг 3. Создайте хранимую процедуру для расчета расстояний.....	331
Шаг 4. Вызовите процедуру из Power BI	333
Предметный указатель.....	335

От издательства

Отзывы и пожелания

Мы всегда рады отзывам наших читателей. Расскажите нам, что вы думаете об этой книге – что понравилось или, может быть, не понравилось. Отзывы важны для нас, чтобы выпускать книги, которые будут для вас максимально полезны.

Вы можете написать отзыв на нашем сайте www.dmkpress.com, зайдя на страницу книги и оставив комментарий в разделе «Отзывы и рецензии». Также можно послать письмо главному редактору по адресу dmkpress@gmail.com; при этом укажите название книги в теме письма.

Если вы являетесь экспертом в какой-либо области и заинтересованы в написании новой книги, заполните форму на нашем сайте по адресу http://dmkpress.com/authors/publish_book/ или напишите в издательство по адресу dmkpress@gmail.com.

Скачивание исходного кода примеров

Скачать файлы с дополнительной информацией для книг издательства «ДМК Пресс» можно на сайте www.dmkpress.com на странице с описанием соответствующей книги.

Список опечаток

Хотя мы приняли все возможные меры для того, чтобы обеспечить высокое качество наших текстов, ошибки все равно случаются. Если вы найдете ошибку в одной из наших книг, мы будем очень благодарны, если вы сообщите о ней главному редактору по адресу dmkpress@gmail.com. Сделав это, вы избавите других читателей от недопонимания и поможете нам улучшить последующие издания этой книги.

Нарушение авторских прав

Пиратство в интернете по-прежнему остается насущной проблемой. Издательства «ДМК Пресс» и Maker Media очень серьезно относятся к вопросам защиты авторских прав и лицензирования. Если вы столкнетесь в интернете с незаконной публикацией какой-либо из наших книг, пожалуйста, пришлите нам ссылку на интернет-ресурс, чтобы мы могли применить санкции.

Ссылку на подозрительные материалы можно прислать по адресу электронной почты dmkpress@gmail.com.

Мы высоко ценим любую помощь по защите наших авторов, благодаря которой мы можем предоставлять вам качественные материалы.

Об авторе



Райан Уэйд (Ryan Wade) является профессиональным аналитиком данных с более чем 20-летним стажем. Своему образованию и опыту работы Райан обязан тем, что приобрел целостное понимание аналитических процессов как с технической, так и с организационной точки зрения. Является обладателем сертификата MCSE в области бизнес-аналитики и Microsoft R и профессионально программирует на R, Python, DAX, T-SQL, M и VBA применительно к локальным и облачным решениям в аналитике на *платформе данных Microsoft (Microsoft Data Platform)*.

Райан принимает активное участие в открытых мероприятиях по R и Python, а также выступает на конференциях SQLSaturdays, TDWI, BDPA и PASS Summit с лекциями по анализу данных. Разработал полноценный онлайн-курс для ExcelTv, в рамках которого демонстрирует применение языка R в Power BI для проведения углубленного анализа данных и их визуализации.

О техническом редакторе

Аадития Марутхи (Aaditya Maruthi) является ведущим специалистом в области разработки баз данных в крупной компании. Обладает более чем 10-летним стажем работы с различными СУБД, включая Microsoft SQL Server и Oracle. По большей части работал с технологиями от Microsoft, такими как SSAS, SSRS, SSIS и Power BI. Аадития также является обладателем сертификата Certified AWS Solutions Architect Associate.

Благодарности

Хочу поблагодарить технических редакторов Майка Хаффера (Mike Huffer) и Аадитию Марутхи (Aaditya Maruthi) за качественную обратную связь, которая помогла сделать эту книгу лучше. Также хотел бы сказать спасибо Джонатану Геннику (Jonathan Gennick) и Джилл Бальцано (Jill Balzano). Признателен вам за ваше терпение и помощь. Без ваших напутствий мне было бы очень непросто завершить начатое. Кроме того, вы постоянно держали руку на пульсе проекта. Изначально я хотел включить в книгу слишком много всего, и в этом случае ее написание заняло бы невероятно много времени. Вы помогли мне понять, что действительно важно, и это позволило завершить работу в приемлемые сроки.

Введение

Microsoft Power BI является одним из наиболее популярных инструментов в области бизнес-аналитики. За последние годы этот программный комплекс опередил своих прямых конкурентов QlikView и Tableau и прочно занял лидирующее положение на рынке. Одним из главных преимуществ Power BI является то, что он представляет собой нечто гораздо большее, чем просто инструмент визуализации данных. Вот лишь несколько явных преимуществ Power BI:

- встроенный язык запросов *DAX*, позволяющий крайне эффективно извлекать информацию из модели данных с применением сложной бизнес-логики;
- интегрированный инструмент подготовки и преобразования данных *Power Query*, при помощи которого можно легко извлекать и трансформировать исходную информацию в вид, пригодный для анализа;
- движок *Vertipaq*, позволяющий хранить данные в оптимальном для формирования отчетов виде и быстро и эффективно обрабатывающий сложные вычисления;
- заранее подготовленные пакеты интерактивных элементов визуализации, помогающие представлять данные в понятной и четкой форме.

Глядя на этот список преимуществ, вы вполне можете задаться вопросом, зачем же столь мощному инструменту понадобилась помощь языков программирования R и Python. Ответ прост – чтобы заполнить области, в которых встроенные средства недостаточно хороши. Вот лишь несколько примеров применения этих языков программирования в рамках Power BI:

- создание пользовательских элементов визуализации без особых усилий;
- реализация интеллектуальной обработки данных, методов машинного обучения и искусственного интеллекта без необходимости приобретать дорогостоящую подписку на *Power BI Premium*;
- применение продвинутых методов обработки текстовой информации с использованием техник, недоступных в *Power Query* и *DAX*;
- взаимодействие со службами *Microsoft Cognitive Services* без необходимости приобретать подписку на *Power BI Premium*;
- взаимодействие со сторонними интерфейсами API с целью эффективного обогащения моделей данных Power BI;
- и многое другое...

В данной книге мы подробно расскажем о том, как использовать на практике языки программирования R и Python для обеспечения всей перечисленной выше функциональности в Power BI. Язык R идеально подходит для Power BI по причине того, что он был создан специально для анализа данных. Уже долгие годы аналитики активно используют R для преобразования и визуализации информации. Так что то немногое, на что не способна программная среда Power BI, может быть с лихвой компенсировано при помощи языка R.

Что касается Python, то этот язык программирования приобрел чрезвычайную популярность в области анализа данных в последнее десятилетие. Одним

Конец ознакомительного фрагмента.

Приобрести книгу можно

в интернет-магазине

«Электронный универс»

e-Univers.ru