Изображение выглядит как текст

Автоматически созданное описание

Дисклеймер: в данном кейсе используются замаскированные «живые» данные настоящей задачи, при этом все названия в кейсе подобраны случайно сохраняя смысл данных, все совпадения с реально существующими – случайны.

# **Контекст**

В далеком 2148 году мир переживает последствия кризиса и глобальной войны. Постапокалиптическую пустошь населяют безжалостные воины, но все еще есть место для честных предпринимателей.

Вы работаете в **Компании**, управляющей сетью магазинов, которая торгует различными товарами, пользующимися спросом в данной реальности.

Магазины выглядят примерно так, как представлено на изображении:

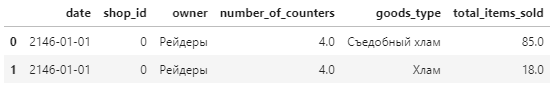


Вам доступны исторические данные о продажах за 2 года и данные о характеристиках магазинов.

# **Исходные данные**

Формат всех источников - .parquet

# Данные по продажам - sales



Изображение выглядит как текст

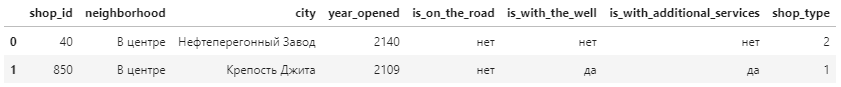
Автоматически созданное описание

* Изображение выглядит как текст

  Автоматически созданное описаниеИзображение выглядит как текст

  Автоматически созданное описаниеshop\_id – уникальный идентификатор магазина
* owner–владелец магазина
* number\_of\_counters– количество работающих прилавков/продавцов
* goods\_type–тип товара
* total\_items\_sold– суммарные продажи в этот день в этом магазине в штуках

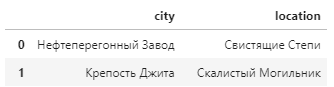
# Данные по характеристикам магазинов - shops



* neighborhood – в какой окрестности находится магазин
* city–в каком городе находится магазин
* year\_opened – в каком году открыт магазин
* is\_on\_the\_road – находится ли магазин прямо у дороги
* is\_with\_the\_well – есть ли у магазина колодец
* is\_with\_additional\_services – есть ли в магазине дополнительные сервисы
* shop\_type – тип магазина

Данный источник является менее точным, чем данные о продажах, так как ведётся и обновляется вручную.

# Данные о разбивке городов по локациям - cities



# **Задача**

Для лучшего управления магазинами, в частности, для более оптимального планирования промо-кампаний и прогнозирования спроса, вам необходимо **разбить магазины на кластеры похожих**. Единственный способ, которым пользовалась компания в прошлом – это разбитие по географическому признаку, то есть по городам. Но вы верите, что прочие характеристики магазинов, а самое главное, профили продаж магазинов, помогут сделать это гораздо точнее.

Вы должны изучить данные, выбрать метрику качества кластеризации, придумать и посчитать информативные признаки (например, доля продаж «патронов» по пятницам) и построить наиболее Изображение выглядит как текст

Автоматически созданное описаниеИзображение выглядит как текст

Автоматически созданное описаниекачественный алгоритм кластеризации, а также описать смысл каждого кластера в понятном для управляющих вашей **Компании** виде.

**Дополнительным** плюсом будет, если вы сможете обосновать, насколько построенная кластеризация лучше, чем разбитие просто по городам.

## Формат предоставления результатов

* Оформленный репозиторий с кодом проекта (в виде .zip файла или ссылки на github/аналог), удовлетворяющий следующим критериям:
  + воспроизводимый
  + содержащий Jupyter Notebook (или несколько) c визуализациями и комментариями к проделанным шагам

Язык решения – python, но если вы им не владеете, допускаются решения так же и на R.

* .tsv (tab separated value) файл с результатами кластеризации, содержащий поля:
  + shop\_id – уникальный идентификатор магазина
  + cluster\_id – идентификатор кластера
* Краткий отчет для **Компании** в любом формате (pdf/ppt/doc и т.п.) о качестве созданной кластеризации и смысловому описанию каждого кластера. Также приветствуются дополнительные комментарии:
  + по имеющимся данным
  + чтобы вы сделали, имея больше ресурсов/времени/данных (каких?)

# **Критерии оценки**

При оценке работ мы будем смотреть на:

* внимательность к данным
* методичность и последовательность размышлений
* полноту и корректность решения
* аккуратность в оформлении кода и его читаемость
* информативность и понятность итогового отчета

# **Изображение выглядит как текст Автоматически созданное описаниеОбратная связь и вопросы ­**

Ваши вопросы по кейсу пишите на платформе teachbase в разделе с кейсом Data Science.

Ответы на основные вопросы по решению кейса будут публиковаться в телеграм канале:   
<https://t.me/joinchat/AAAAAEh4S1F4E03jxyRVTQ>

# **Изображение выглядит как текст Автоматически созданное описаниеПрочие комментарии**

Для чтения .parquet файлов вам понадобиться pyarrow

conda install pyarrow

затем

pandas.read\_parquet(‘data.parquet’)

Изображение выглядит как текст

Автоматически созданное описание