GPN 2020. Data Sciense

Выполнил: Александр Широков

Solution
15.11.2020 г.

Содержание

1 Input Data

1.1 Problem Definition

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.

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

Problem:

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 качественный алгоритм кластеризации, а также описать смысл каждого кластера в понятном для управляющих вашей Компании виде.

С этого момента начается моё слово.

1.2 Table *sales*

Таблица *sales* содержала 5081459 записей по продажам за 2 года с 01.01.2146 по 01.01.2148

	date	shop_id	owner	number_of_counters	goods_type	total_items_sold
0	2146-01-01	0	Рейдеры	4	Съедобный хлам	6.0
1	2146-01-01	0	Рейдеры	4	Хлам	26.0
2	2146-01-01	0	Рейдеры	4	Бензак	10537.0
3	2146-01-01	1	Рейдеры	5	Съедобный хлам	17.0
4	2146-01-01	1	Рейдеры	5	Хлам	9.0

Рис.1 Таблица *sales*

со следующими признаками:

- DATE - дата продажи товара $Y - M - d$
- SHOP_ID - уникальный идентификатор магазина: в интервале $[0, 844]$, INTEGER
- OWNER - владелец магазина, строковый тип, 5 уникальных владельцев:

1. Рейдеры - 3906481 - самая многочисленная группа
2. Воины полураспада - 595022
3. Стервятники - 275076
4. Последователи Апокалипсиса - 169120
5. Бомбисты - 135760

- NUMBER_OF_COUNTERS - количество работающих прилавков/продавцов, INTEGER
- GOODS_TYPE - тип товара, всего 11 видов товара, string
- TOTAL_ITEMS_SOLD - суммарные продажи в этот день в магазине в штуках, FLOAT

В таблице отсутствовали пропущенные значения. Проведём некий разведывательный анализ данной таблицы. Для начала я сгенерировал новые признаки:

- YEAR - год продажи: $[2146, 2147]$
- MONTH - месяц продажи: $[1, \ldots, 12]$
- DAY - день продажи: $[1, 31]$
- DOY - номер дня в году: $[1, 365]$
- DAY_NAME - название дня недели: MONDAY, ..., SATURDAY
- MONTH_NAME - название месяца: JANUARY, ..., DECEMBER
- IS_WEEKEND - является ли день выходным днём: $[0, 1]$

Для чего я вообще начал проводить разведывательный анализ. Идея моя заключалась в следующем: для кластеризации на некоторые группы магазинов хотелось бы найти такие признаки, по которым наши магазины имели какие-то различия в продажах, количестве продавцов и.т.д. Для этого я и начала получать описательные статистики в некоторых разрезах. Вот некоторые результаты, которые удалось получить:

- Среднее количество продаж по месяцам в разрезе SHOP_ID не выявило никаких явных различий - всегда получалось распределение по типу Пуассоновского, но с очень тяжёлыми хвостами. Такая же картина и в разрезе по годам - картина абсолютно идентичная

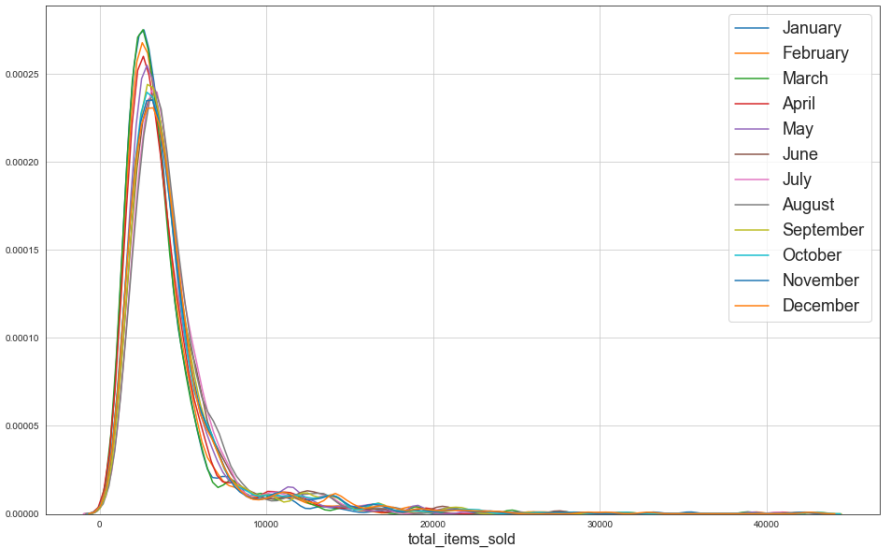

Рис.2 Распределение гистограмма средних продаж по магазинам в разрезе по дням недели и виду товара

- При анализе среднего количества проданных товаров было выяснено, что самыми продаваемыми товарами являются Бензак и Солярка, стягивающие на себя в основном все продажи.

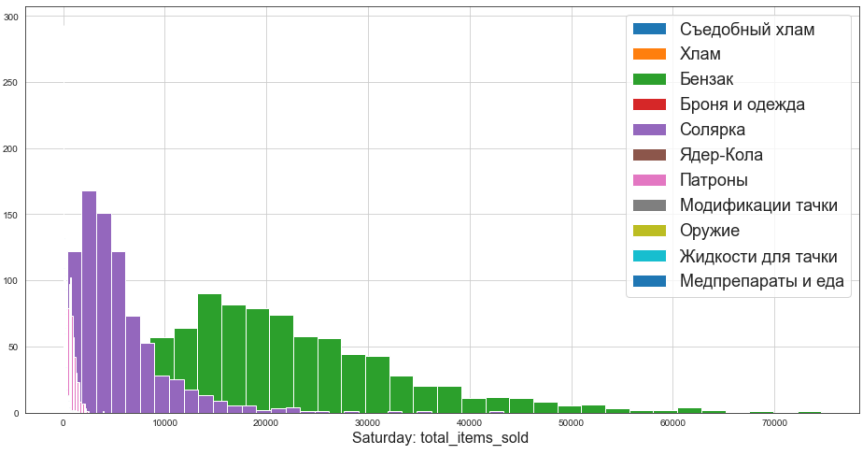

Рис.3 Распределение гистограмма средних продаж по магазинам в разрезе по товарам

Тем не менее вид товара не влияет на распределение продаж в разрезе по дням недели - данная картина наблюдается для каждого дня недели.

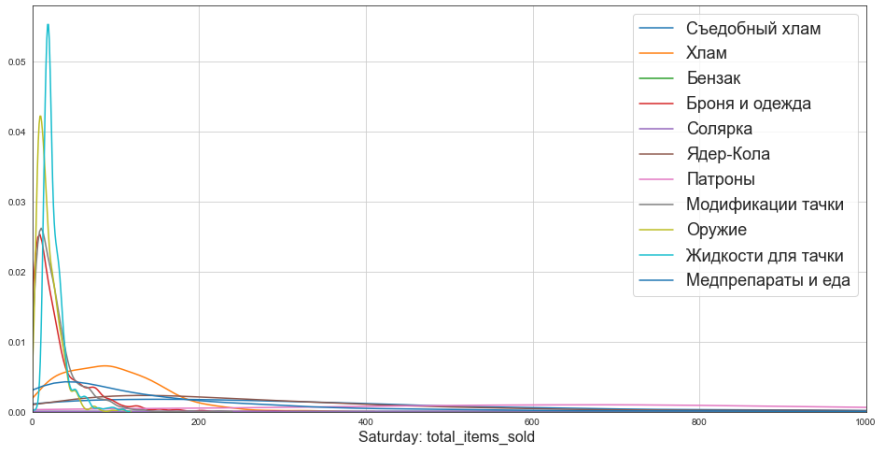

Рис.4 Распределение гистограмма средних продаж по магазинам в разрезе по дням недели (суббота)

Зато если взять общее распределение, то видно, что в пятницу суммарные продажи товаров являются наименьшими, а в среду - наибольшими. Средие суммарные продажи приходятся на понедельник.

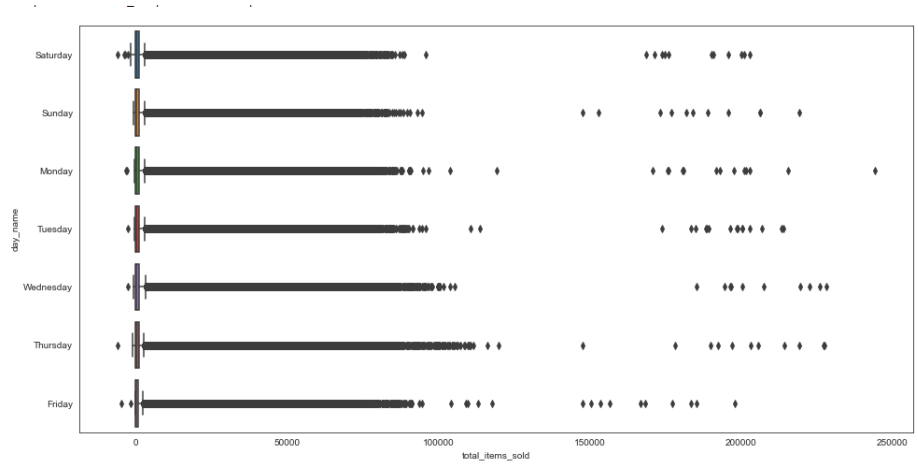

Рис.5 Суммарные продажи в магазинах в разрезе по дням недели

- Проводился так же анализ средних продаж в разрезе по владельцам магазинов - OWNER.

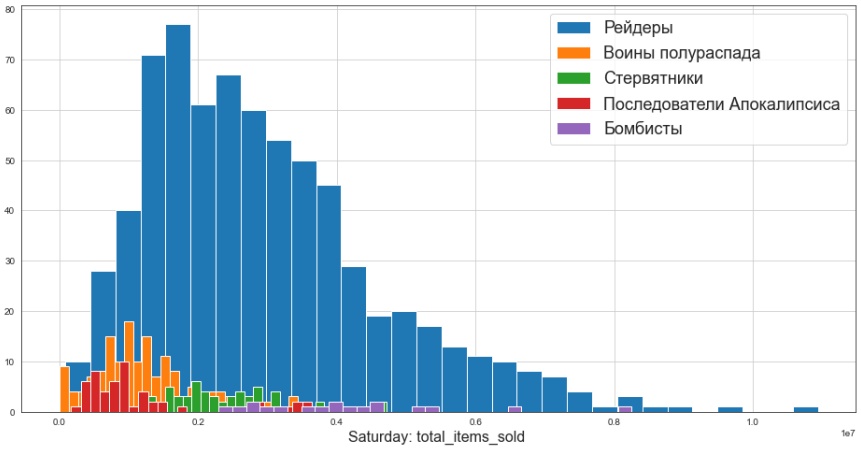

Рис.6 Суммарные продажи в магазинах в разрезе по владельцам магазинов

Видно, что наибольшие продажи на себя стягивают "Рейдеры"(синие), затем "Воины полураспада"(оранжевые) и затем идёт группа из самых маленьких продаж. На самом деле, уже на этом были мысли по поводу кластеризации: если есть какая-то гистограмма плотности распределений, то кластеризацию вполне можно проводить на основании данной плотности, наблюдая за количество горбов в гистограмме. Да и из картинки напрашивается распределение магазинов на "продающих много "продающих средне"и продающих "очень мало". Но ведь неоходимо понять, из-за чего магазины получают преимущество над другими магизинами.

Немного отвлечемся и посмотрим на разницу в распределениях для продаж товара "Бензак"и "Броня и одежда"в разрезе по владельцам:

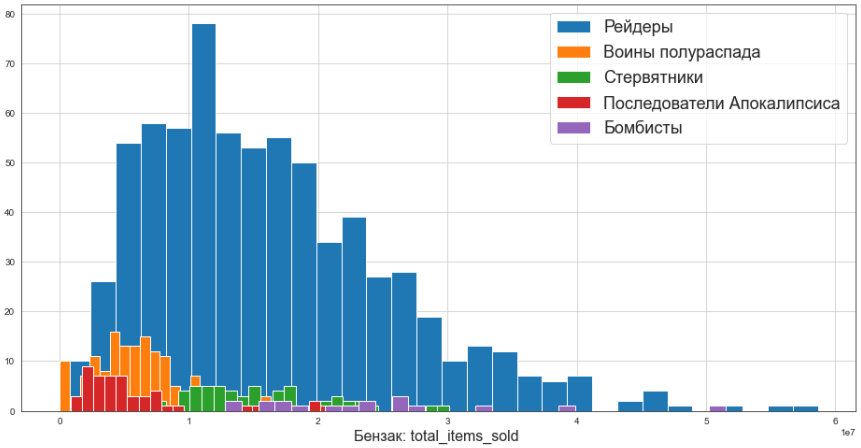

Рис. 7 Продажа товара "Бензак нет явного перекоса в нулевой столбец на гистограмме

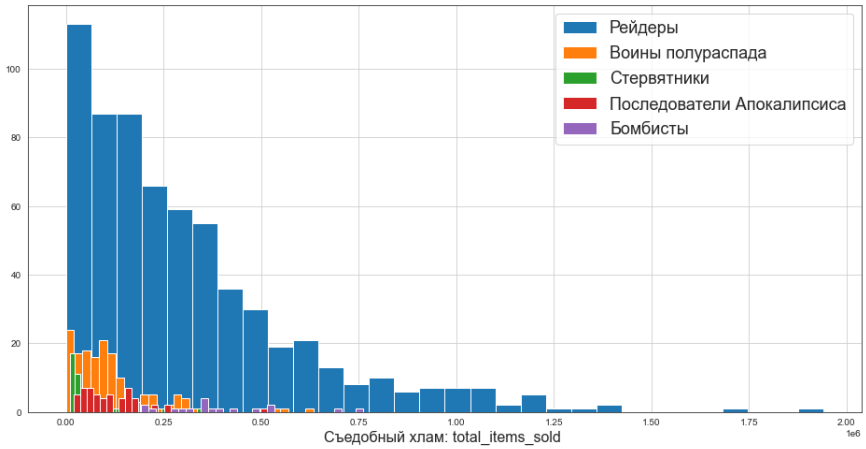

Рис. 8 Видно, что есть нулевой столбик явно выделяющийся на гистограммах - так выглядят все гистограммы для непродающихся товаров

- Далее я посмотрел на количество продавцов и поинтересовался: менялось ли количество продавцов координально с течением времени. Это вопрос даст нам ответ про стабильность ситуации в нашем апокалиптическом обществе.

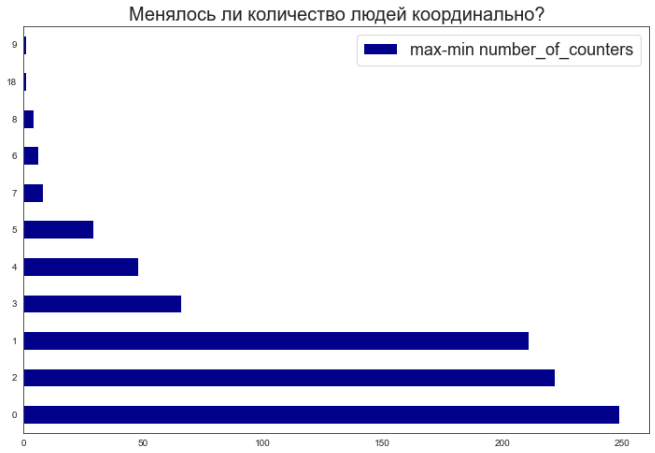

Рис. 9 Разница между максимальным значением продавцов и минимальным в течение времени

Видно, что количество продавцов практически не менялось (± 2). Было только несколько случаев, когда кто-то из владельцев "психанул"и добавил 18 новых продавцов. Видимо, хотел улучшить продажи. В качестве вывода можно сказать, что по данном фактора кластеризация не будет происходить.

- Распределение продаж по будням и на выходных:

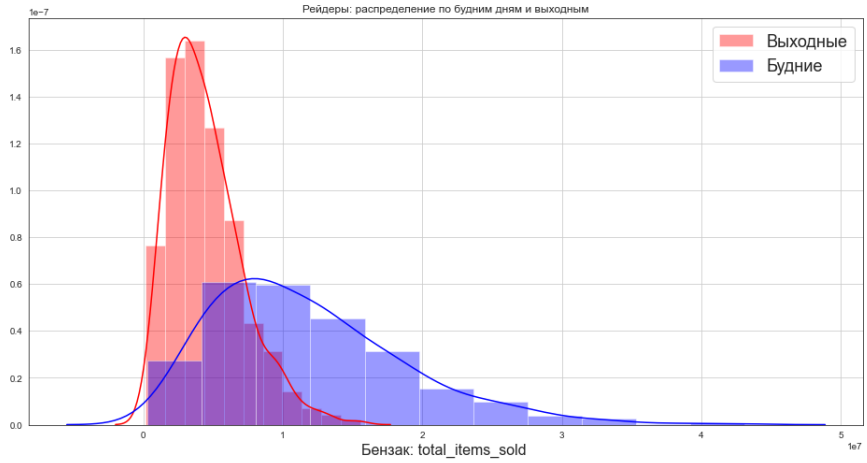

Рис. 10 Суммарные продажи по будням и на выходных

По выходным суммарное значение выручки больше, чем по будням, но как мы видим - с очень маленькими хвостами, а по будням - среднее значение выручки меньше, но хвосты тяжелые, поэтому логично предположить, что плотность распределения средних значений будет примерно одинаковая.

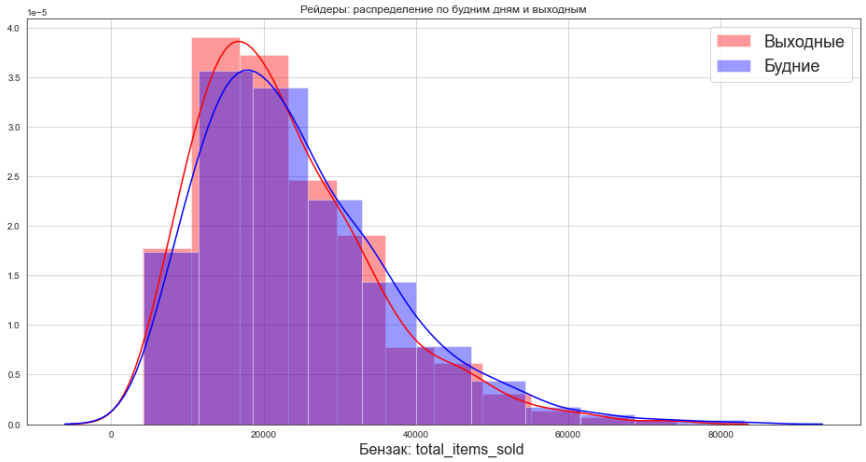

Рис. 11 Средние продажи по будням и на выходных. Как и было предположено

Различий в гистограммах в разрезе по дням недели и товарам не были выявлены.

Выводы из таблицы *sales*: были выявлены различия в распределениях продаж товара "Бензак"и "Солярка"в сравнении с другими товарами, было выявлено, что по пятницам товары продаются меньше всего на недели, а в среду - больше всего, а так же магазины подразделяются на три категории в разрезе по владельцам магазинов: большие, средние и очень маленькие продажи. Перейдем к рассмотрению таблицы *cities*.

1.3 Table *cities*

В таблице *cities* представлена информация о принадлежности города определенному местоположению - LOCATION. Всего региона 3: Скалистый Могильник, Свистящие Степи и Радиоактивная Пустошьи города распределены равномерно по данным регионам: $6, 4, 4$ соответственно. Более информации нет, поэтому перейдем к таблице *shops*.

1.4 Table *shops*

Данная таблица содержит информацию по характеристикам магазинов и не зря сказано в описании, что данная таблица является менее точной, так как ведётся вручную - уже с самого начала я ожидал трудности с заполнением пропущенных данных.

Таблица *shops* содержала 845 записей с характеристиками магазинов со следующими признаками:

- SHOP_ID - PK, уникальный идентификатор магазина: $[0, \ldots, 844]$
- NEIGHBORHOOD - в какой окрестности находится магазин, категориальный признак, имеется 7 уникальных значений, без пропущенных: В центре, Промзона, У ночлега, У воды, На отшибе, У тоннеля< С краю
- CITY - в каком городе открыт магазин, категориальный признак.

  На данном этапе хочется остановиться, потому что именно в этом признаке скрывается **причина неудчаной кластеризации по городам**. Если просмотреть на частоту встречаемости городов, то увидим следующее:

Рис. 12 Встречаемость определённых городов в таблице *sales*

Видно, что в практические 7% случаев неизвестно (None), в каком городе находится магазин. В этом и есть причина, по которому кластеризация по городам не является оптимальным - во-первых, у нас много городов и большое количество кластеров.. весьма неинтерпретируемо, во-вторых у нас весьма велика вероятность (она сравнима с вероятностями других городов) того, что какой-то магазин мы просто не сможем обнаружить в этом столбце и примерно 7 столбцов наблюдений мы просто выкидываем. А вдруг данный SHOP_ID принадлежит "Рейдерам"и мы потеряем огромное количество прибыли из рассмотрения (такое же возможно). Поэтому, кластеризация по городам - не очень хорошая идея, постараемся придумать получше, получается. Да и что делать с пропущенными значениями - не особо понятно. Было принято решение не включать данный категориальный признак в общее рассмотрение, раз кластеризация не будет проводиться по нему.

- YEAR_OPENED - в каком году был открыт магазин. В 63 случаях (второе место по встречаемости) неизвестно было в каком году был открыт магазин, поэтому было предпринято следующее преобразование данного признака: отсутствующие значение заменим на *медиану* по столбцу, а дальше каждое значение заменим на разность $2147 - X$, где X - значение года открытия. Таким образом мы переходим к некоторой описательной статистики открытия города.

- IS_ON_THE_ROAD - находится ли магазин прямо у дороги. Распределение значений следующее: нет - 614, да - 224 и в 7 случаях неизвестно. Данный категориальный признак можно прокодировать с помощью OneHotEncoding, а пропущенные значения заменить на самое встречаемое значение - 0.

- IS_WITH_THE_WELL - есть ли у магазина колодец. У данного признака есть явный перекос в сторону "нет не имеет, поэтому было принято решение не применять кодирование - был бы очень сильный дисбаланс.

- IS_WITH_ADDITIONAL_SERVICES - есть ли в магазине дополнительные сервисы. Данный признак является сбалансированным, по половине наблюдений за "нет"и "да".

- SHOP_TYPE - тип магазина, всего 4 типа. И на самом деле я изначально ставил очень много надежд на этот признак и вот в чём дело. У данного признака есть много пропущенных значений - примерно пятая часть. И если бы по данному признаку можно было бы кластеризировать магазины, то задача переформулировалась в задачу обучения с учителем - есть 4 класса и необходимо предсказать для отсутсвующих классов принадлежность определённому классу. Эта задача была бы намного приятнее. Были предприняты попытки генерирования признаков и на основании алгоритма K-ближайших соседей собственной

реализации (из прошлых времен) предсказывать классы - по виду магазина. Но, к сожалению, данный признак не показал различий в распределениях продаж товаров, продаж по владельцам магазинов и.т.д. Но переход к задаче обучения с учителем на основании какого-то признака - та идея, которую следует иметь ввиду и я бы в эту сторону хорошенько подумал.

Из некоторых выводов: мы заполнили пропущенные значения< отбросили несбалансированные признаки и предложил несколько идей для кластеризации. Так же было сделано наблюдение, что если в одном из последних 6 признаков отсутствовало значение, то, в принципе информация о данном магазине отсутстовала и данные можно удалить из таблицы для упрощения кластеризации.

Осталось рассмотреть общую MERGE таблицу.

1.5 Table Merge: *sales* / *shops* / *cities*

Соединим сначала таблицы *shops* и *sales* по столбцу CITY, а затем полученную таблицу соединим по столбцу SHOP_ID с таблицей *sales*.

Проведём опять анализ данной таблицы.

- Для начала хотелось бы понять, какое местоположение является наиболее выгодным: для этого будем использовать данные о суммарных продажах по магазинам в разрезе по столбцам OWNER и NEIGHBORHOOD. Для того, чтобы получить информацию о наилучшем местоположении, я провёл следующий алгоритм: брал магазины с владельцем OWNER и сопоставлял самому большому значению выручки в разрезе по NEIGHBORHOOD минимальный ранг и.т.д. В итоге так сделал для всех owner и посчитал сумму рангов. В итоге получились следующие результаты:

  1. В ЦЕНТРЕ - наиболее выгодное местоположение, сумма рангов - 0.
  2. У ТОННЕЛЯ, С КРАЮ - делят 2/3 места, сумма рангов - 9
  3. ПРОМЗОНА - суммар рангов - 13
  4. У НОЧЛЕГА - суммар рангов - 16
  5. НА ОТШИБЕ - суммар рангов - 20
  6. У ВОДЫ - суммар рангов - 21

  Мы нашли признак, по которому распределения разные выручки. Возьмём на вооружение..

- Далее я попытался проверить теорию о кластеризации по SHOP_ID, но пришло разочарование:

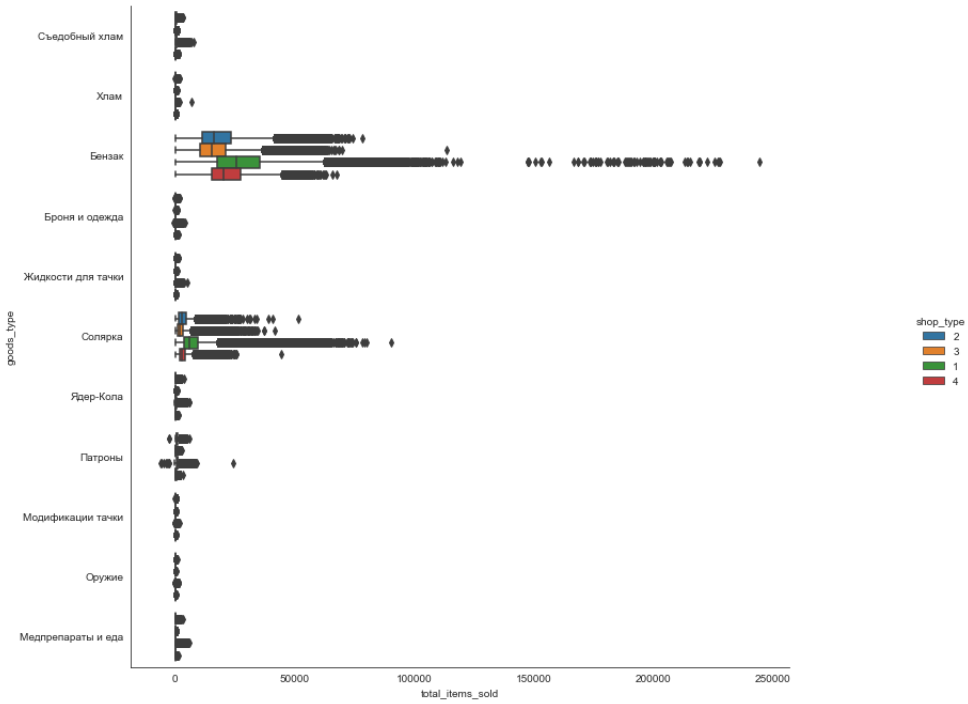

Рис. 13 Разрез по SHOP_ID в зависимости от продажи товаров

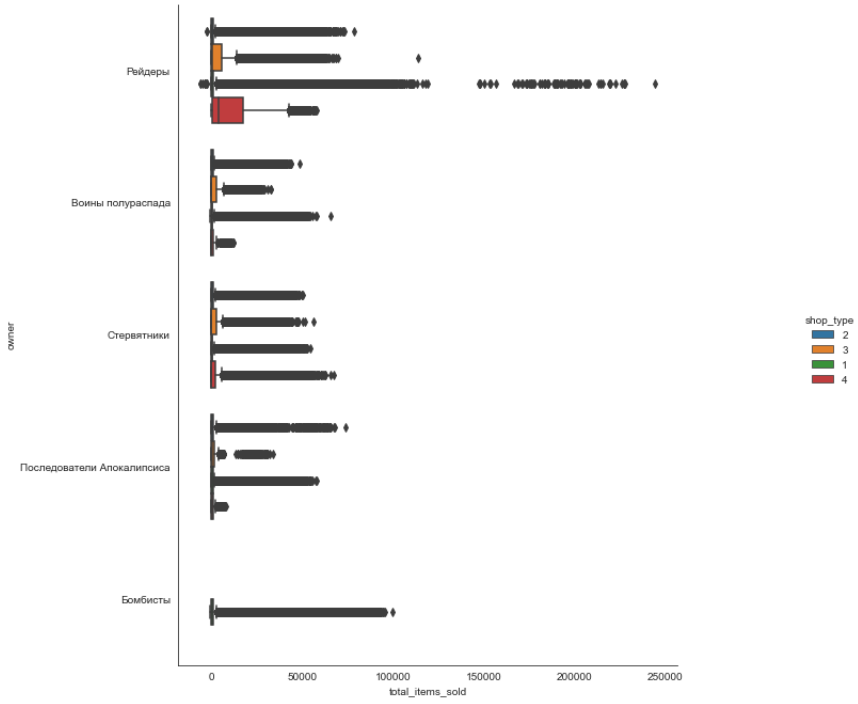

Рис. 14 Разрез по SHOP_ID в зависимости от владельца

Видно, что распределения одинаковы в разрезах по товарам одинаковое, а вот второй рисунок.. Он оставляет надежду на кластеризацию

по данному признаку, но я просто побоялся, если честно, его интерпретировать, потому что по идее можно здесь разбить на 3 класса, но бомбисты уж очень явно различаются... В общем идею с применением обучения с учителем по предсказанию пропущенных значений в SHOP_ID я отложил и не притронулся больше.

- Теперь проанализируем, есть ли различия при наличии дороги

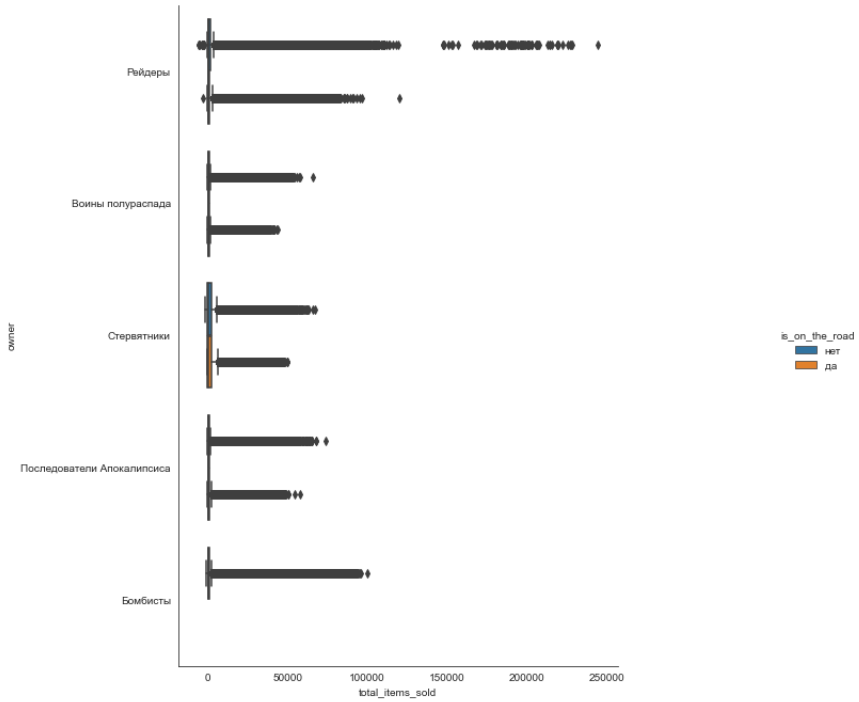

Рис. 15 Видим, что лучше продаются магазины не находящиеся рядом с дорогой.

- Проанализировав признак IS_WITH_THE_WELL было выяснено, что распределения не зависят от наличия или отсутствия данного признака - удалим из рассмотрения.

- IS_WITH_ADDITIONAL_SERVICES

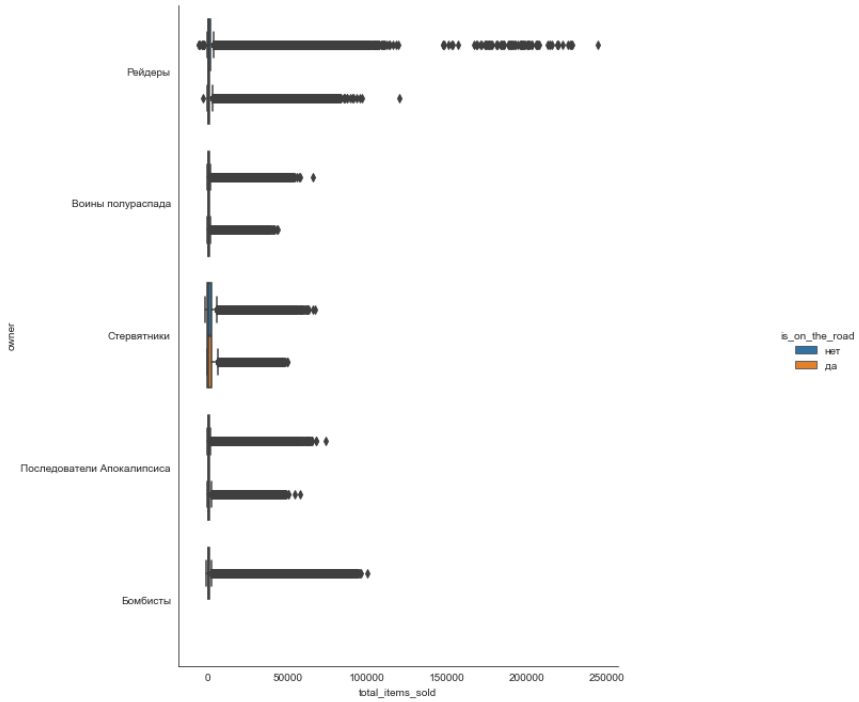

Рис. 16 Видим, что у "Стервятников"продажи больше при наличии дополнительных сервисов

Некоторые выводы: мы подходим к важному этапу: построение гипотезы кластеризации. Действуя по изначально намеченному плану, мы искали признаки, на основании которых распределения тех или иных статистик различались. Перейдем к построению гипотезы кластеризации:

1.6 Hypothesis of Clusterisation

На данном этапе появилася **гипотеза кластеризации**: будем кластеризировать магазины по выгодности местоположения, откуда следует принадлежность определённому кластеру: маленькой, средней или большой компании, а так же, чем выгоднее местоположение, тем больше выручка с основных товаров, которые продаются в магазинах: Солярка, Бензак.

Это гипотеза звучит логично: разделим наши магазины на несколько групп по выгодности местоположения (у реки или в центре - разница есть), а из выгодности местоположения делается вывод о том, какая компания могла бы выкупить выгодное местоположение - очевидно, что богатая. Из выгодности местоположения вытекает увеличиние выручки.

Перейдем к генерации признаков и кластеризации.

2 Clustersisation

2.1 Feature Engineering

Создадим матрицу признаков, на основании которых мы будем делать кластеризацию. Идея следующая сгенерируем признаки таким образом, чтобы по тем признакам из первой части моего отчёта, где наблюдались различия в распределениях, мы взяли самые плохо продаваемые категории, средние и наиболее хорошо продаваемые. Из этого метода сгенерируем следующую матрицу признаков, состоящую из:

- - среднее количество TOTAL_ITEMS_SOLD по продуктам "Бензак "Оружие "Хлам"и "Броня и Одежда будем придерживаться данной стратегии - выбирать наилучшее по продажам среднее и наихудшее. После каждого добавления заполняем нулевые значения минус максимальным значеним в "dataframe чтобы обозначить различие между возникающими классами более явно.
- среднее количество по дню DAT_NAME недели: Пятница, как наименьшее, Thursday - как наибольшее и MONDAY - как нечто среднее.
- количество продаж по наличию или отсутствию дополнительных сервисов
- среднее количество продаж по владельцам "Рейдеры"и "Бомбисты"

2.2 Use Metric: *Silhouette*

Будем использовать метрику SILHOUETTE для оценки качества кластеризации.

Для одного элемента x она считается так:

$$S(x) = \frac{b(x) - a(x)}{\max\left(a(x), b(x)\right)}$$

где

- $a(x) =$ среднее расстояние от x до точек внутри того же кластера.
- $b(x) =$ среднее расстояние от x до точек внутри ближайшего кластера.

Сама метрика равна среднему значению $S(x)$ от каждого элемента.

Видно, что $-1 \leqslant S(x) \leqslant 1$, причем чем больше $b(x)$ относительно $a(x)$, тем метрика ближе к 1. Чем метрика больше - тем лучше кластеризация.

2.3 Use Method: Agglomerative Clustering

Будем использовать алгоритм AGGLOMERATIVE CLUSTERING. Интуиция у алгоритма простая:

1. Начинаем с того, что высыпаем на каждую точку свой кластер

2. Сортируем попарные расстояния между центрами кластеров по возрастанию

3. Берём пару ближайших кластеров, склеиваем их в один и пересчитываем центр кластера

4. Повторяем п. 2 и 3 до тех пор, пока все данные не склеятся в один кластер

Применим алгоритм AgglomerativeClustering к нашим признакам.

2.4 Clusterisation: Result

	silhouette_score	value_counts
2	0.866026	{1: 649, 0: 196}
3	0.88943	{0: 649, 2: 178, 1: 18}
4	0.689217	{1: 459, 3: 190, 2: 178, 0: 18}

Рис. 17 Результаты кластеризации AgglomerativeClustering по метрике Silhouette

По результатам кластеризации можно сделать вывод, что при $k = 3$ у нас наибольшее значение метрики, но самое главное - этот результат весьма предсказуем - мы разделили наше множество на 3 магазина, причём второй и первый класс, как видно из результатов, весьма похожи друг на друга, а это именно то, о чём мы говорили, когда предполагали, что кластеризовать можно по большим, средним и маленьким компаниям (страница 5). Так же выяснено, что полученные кластеры получились, если принять во внимание местоположение магазина, то есть выручка зависит от выгодности местоположения магазина.

Характеристики полученных кластеров:

- Класс №0 (649) - наиболее представимый класс, большинство из магазинов принадлежат двум самым богатым продавцам, так же практически все данные магазины имеют наибольшую среднюю выручку и магазины располагаются в большинстве своём 'В центре' либо 'У тоннеля', не имеет различий в том, есть ли дополнительные сервисы или нет

- Класс №1 (178) - очень сильно проигрывает классу №0, имеет намного меньшую среднюю выручку, расположены в в остальных участках neighborhood, но имеют различия в выручке при наличии дополнительных сервисов в большую сторону

- Класс №2 (18) - во многом очень схожий с классом №1, имеет схожую выручку, но местоположение - у Воды (в большинстве своём) и выручка не зависит от наличия дополнительных сервисов

Результаты кластеризации были загружены в файл **submission.tsv**

3 Conclusion

Построенный алгоритм качественно проводит кластеризацию на данных объектах и имеет понятную интерпретацию, но по сути мы кластеризуем наши данные на основании количества продаж и (слава богу) сумели найти признак различия выгодности местоположений, что, однако, является не причиной, а следствием основного нашего признака. Поэтому, я бы попросил информацию о количестве людей, находящихся в магазине каждый день, о популяции городов (например), добавил бы некоторого разнообразия в данные, хотя и на доступных нам данных мы смогли вытянуть очень много информации, как мне кажется.

Мне понравилось заниматься данной задачей и мне бы хотелось продолжить ею заниматься, потому что я, конечно, чувствую, что загадка этой задачи не раскрыта и наполовину, но я не ограничился одни лишь предложением данной кластеризации: в отчёте я предлагал несколько идей, в направлении которых можно двигаться в дальнейшем.

В SOLUTION.IPYNB находится оформленный JUPYTER NOTEBOOK с выводами и реализованной кластеризацией.

Спасибо за задание! Мне было интересно решать и, я надеюсь, что на данном этапе мы не остановимся.

С Уважением, Александр Широков