Выполнение поставленной задачи было разделено на несколько этапов. Для этого были определены соответствующие функции.

Загрузка и предобработка данных – функция from_parquet(). Были загружены данные из трёх таблиц, которые впоследствии объединяются в одну по столбцам ‘city’ и ‘shop_id’.

В загруженных данных были обнаружены пустые поля в следующих столбцах:

```
shop_id                          0
neighborhood                     0
city                        338384
year_opened                      0
is_on_the_road               39978
is_with_the_well            371080
is_with_additional_services 377919
shop_type                   779399
location                    338384
date                             0
owner                            0
number_of_counters               0
goods_type                       0
total_items_sold                 0
dtype: int64
```

Было решено заполнить отсутствующие данные в столбцах в функции process_empty_fields(). Пропуски в полях ‘is_on_the_road’, ‘is_with_the_well’, ‘is_with_additional_services’, ‘shop_type’, ‘location’, ‘city’ были заменены на строку ‘неизвестно’. Поля, где год открытия ‘year_opened’ = -1 были удалены.

Далее необходимо заменить данные типа object на целочисленные. Это осуществляется с помощью функции replace_string_columns(), которая возвращает преобразованную таблицу и словарь с закодированными объектами.

Для столбца ‘total_items_sold’ проведем min-max нормализацию в функции normalize_column().

Стоит отметить, что из-за больших размеров таблицы (4694733 строк) данные для осуществления кластеризации необходимо разделить на батчи. Кластеризация проводилась в цикле для каждой даты совершения покупки с использованием агломеративной кластеризации, где новые кластеры создаются путем объединения более мелких кластеров и, таким образом, дерево создается от листьев к стволу.

В качестве критерия связи linkage был выбран метод ward, так как он сводит к минимуму дисперсию объединяемых кластеров. Количество

кластеров было выбрано равное 3 из предположения, что магазины могут быть разделены на крупные, средние и маленькие.

После получения кластеров для всех магазинов, осуществляется поиск наиболее часто встречающегося присвоенного кластера в рамках каждого магазина (shop_id).

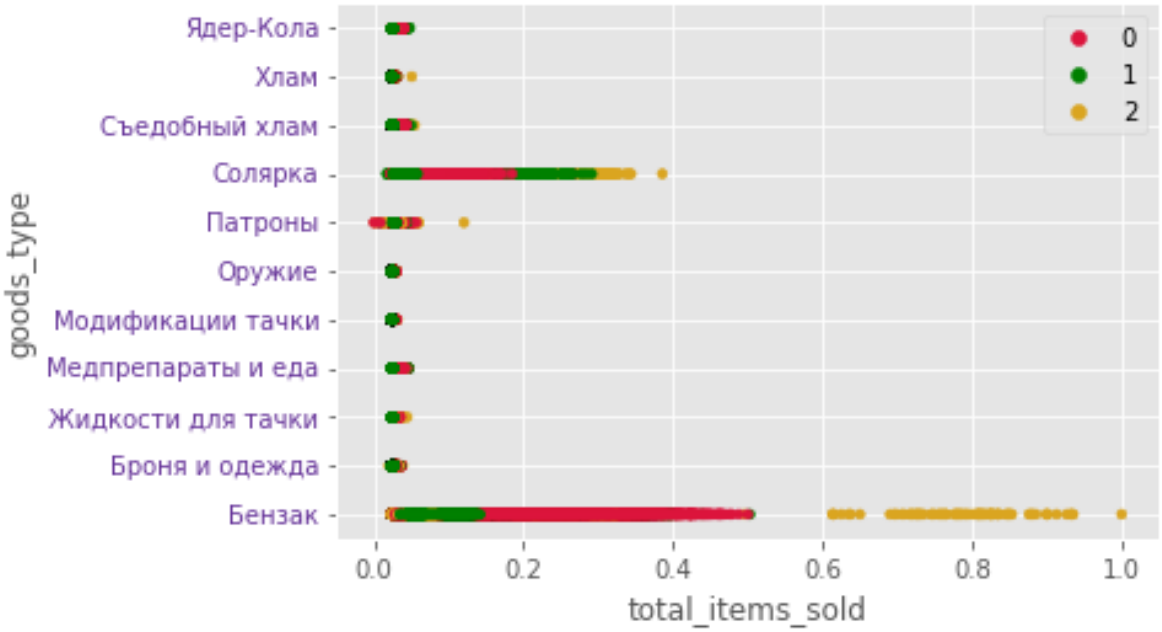
Одним из основных признаков, влияющим на повышение эффективности продаж, целесообразно считать количество проданных товаров – ‘total_items_sold’. Изобразим на графике кластера зависимости числа проданных товаров от различных признаков.

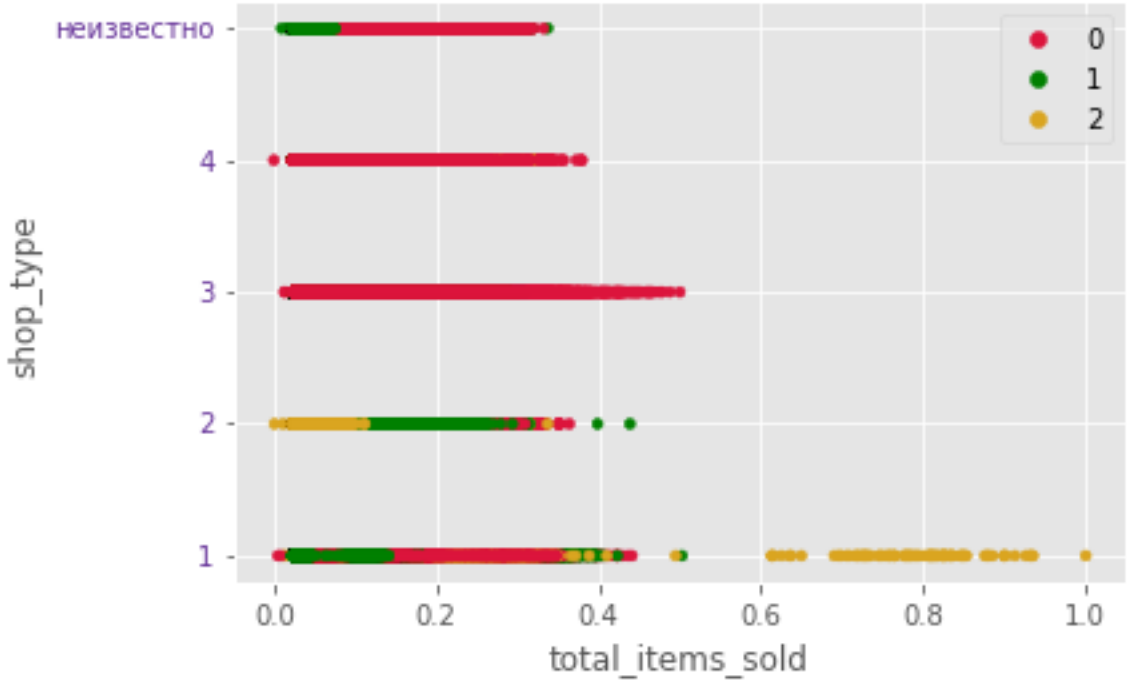
Анализ кластеризованных данных

По построенным ниже зависимостям можно заметить, что Бензак лидирует по количеству проданных товаров, и магазины, в которых он продается в большом количестве, относятся ко 2-му кластеру. Тип наиболее успешных магазинов – 1 (shop_type). Однако, было совершено всего 82 покупки в магазинах 1-го типа, где количество проданных товаров было больше среднего.

Магазины, отнесенные к 1-му кластеру, имеют широкий спектр проданных товаров, но больше всего продается солярка и бензак. Это магазины типов 1, 2.

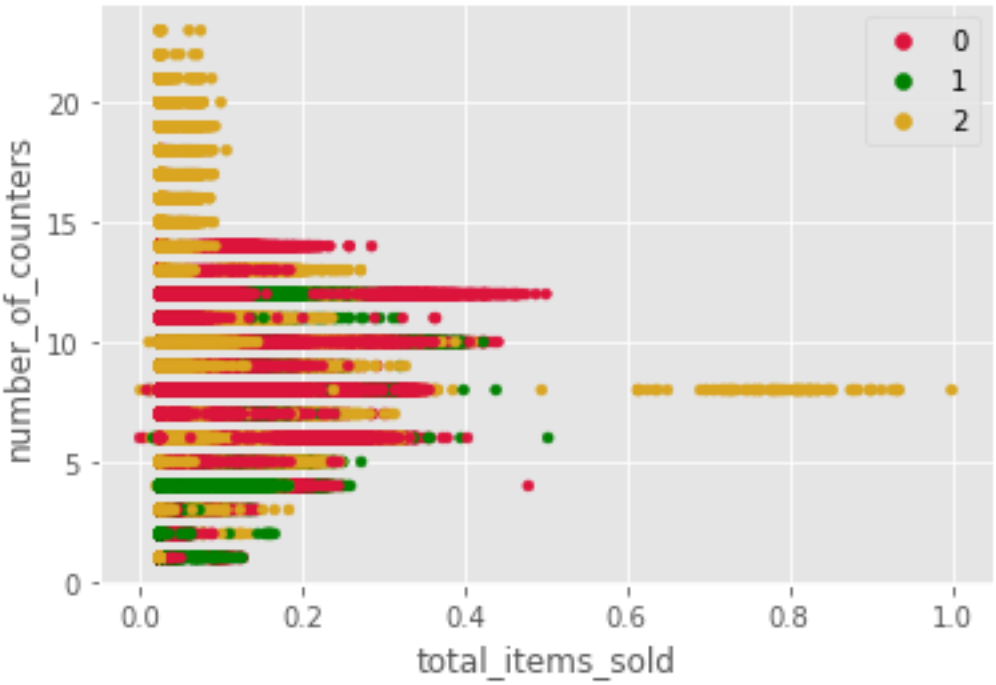
К 0-му кластеру были отнесены магазины типов 1, 3, 4, специализирующиеся на продаже солярки и бензака.





Число кассиров в наиболее успешных магазинах второго кластера равно восьми. Также отсюда можно заметить, что ко второму кластеру были отнесены магазины с числом кассиров более 14, но продажи в этом случае не большие.

В магазинах первого кластера работает мало кассиров, часто это магазины с одним кассиром.

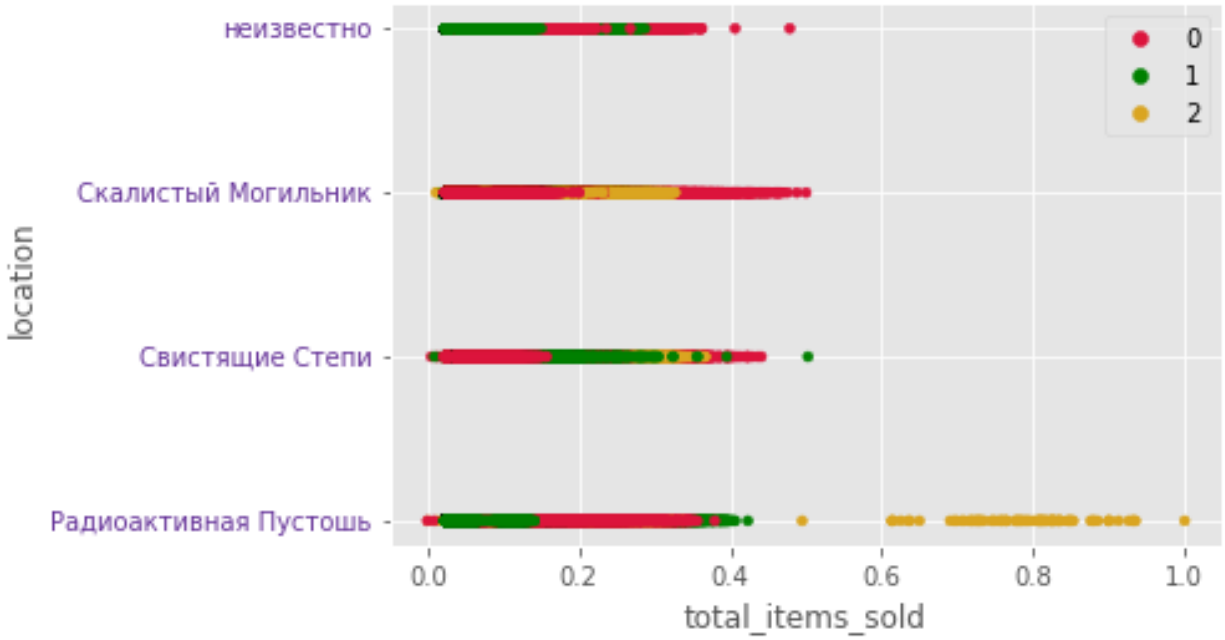


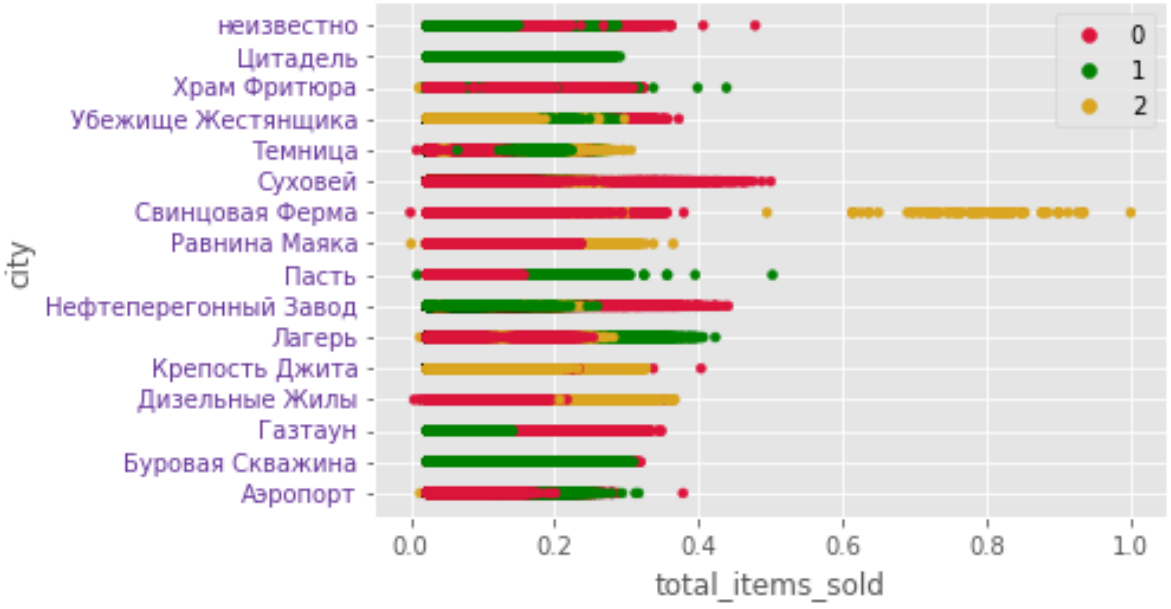
Анализируя рисунки ниже, отметим, что наиболее эффективное по числу проданных товаров расположение магазинов 2-го кластера – в

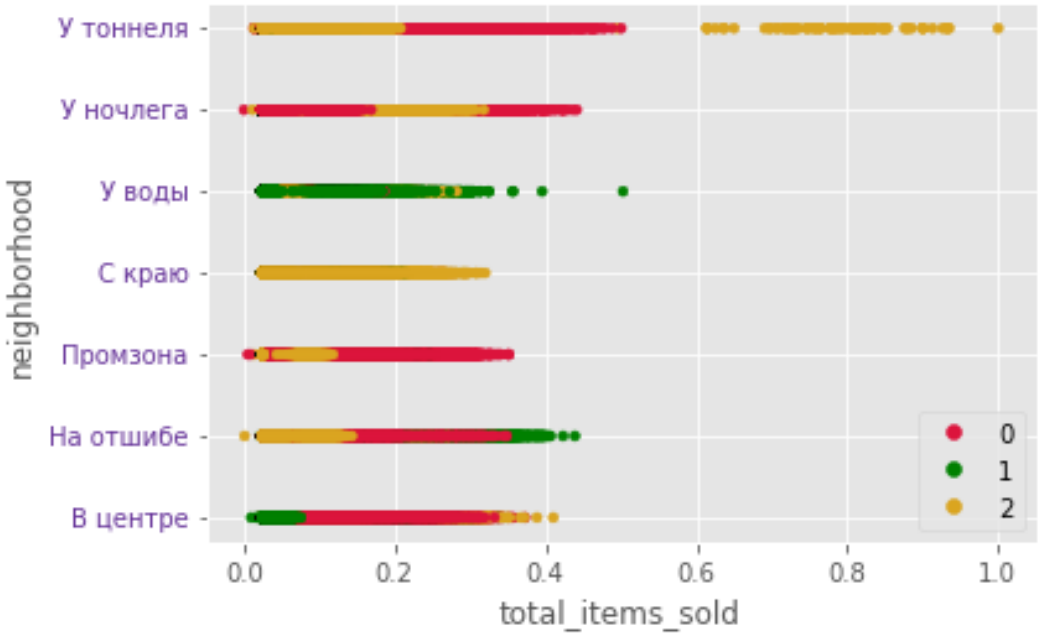
Радиоактивной Пустоши, город Свинцовая Ферма, у тоннеля, не у дороги. Без колодца, но с дополнительными услугами. Также эти магазины встречаются в городе Убежище Жестянщика, но там низкие продажи. В магазинах 2-го кластера с колодцем низкое количество проданных товаров.

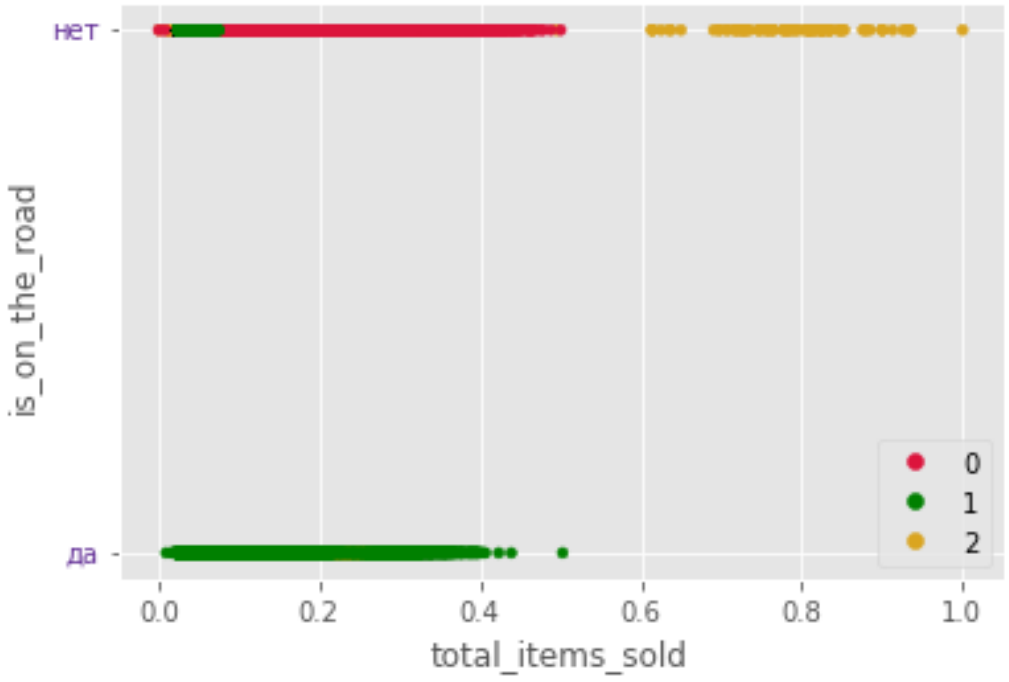
Что касается 1-го кластера, то магазины этого кластера чаще всего встречается в Свистящих Степях и в Радиоактивной Пустоши, во многих городах у воды, у дороги. Без колодца и без дополнительных услуг.

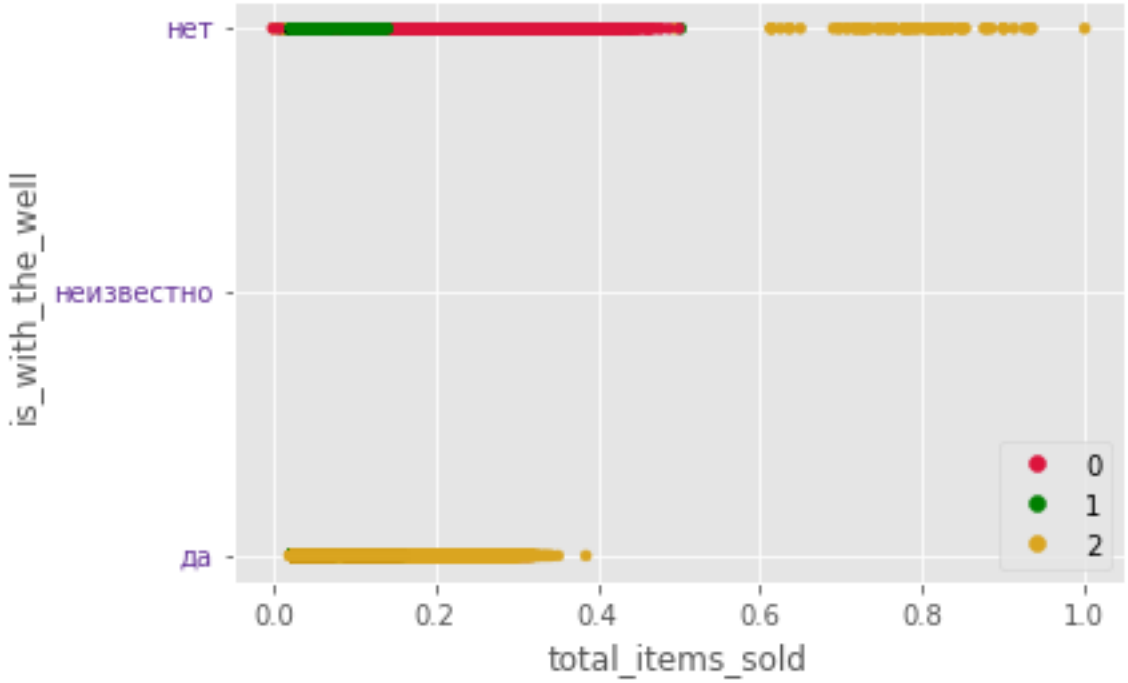
Колодец в магазинах 0-го кластера отсутствует. В остальном эти магазины располагаются повсеместно, кроме Цитадели.

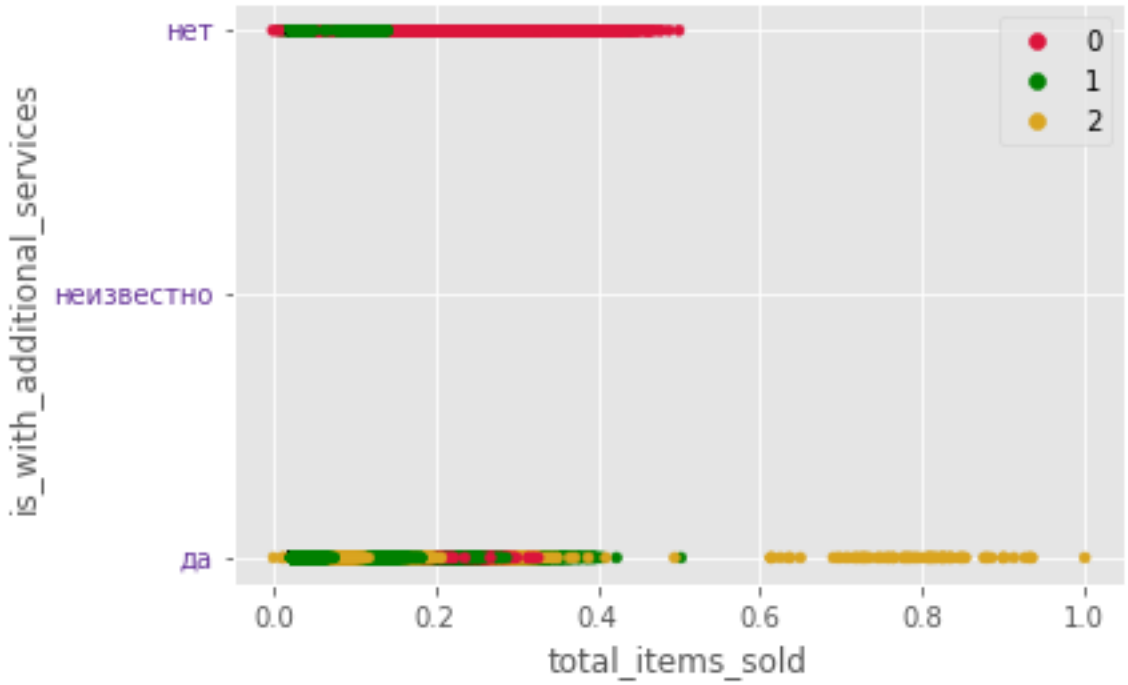




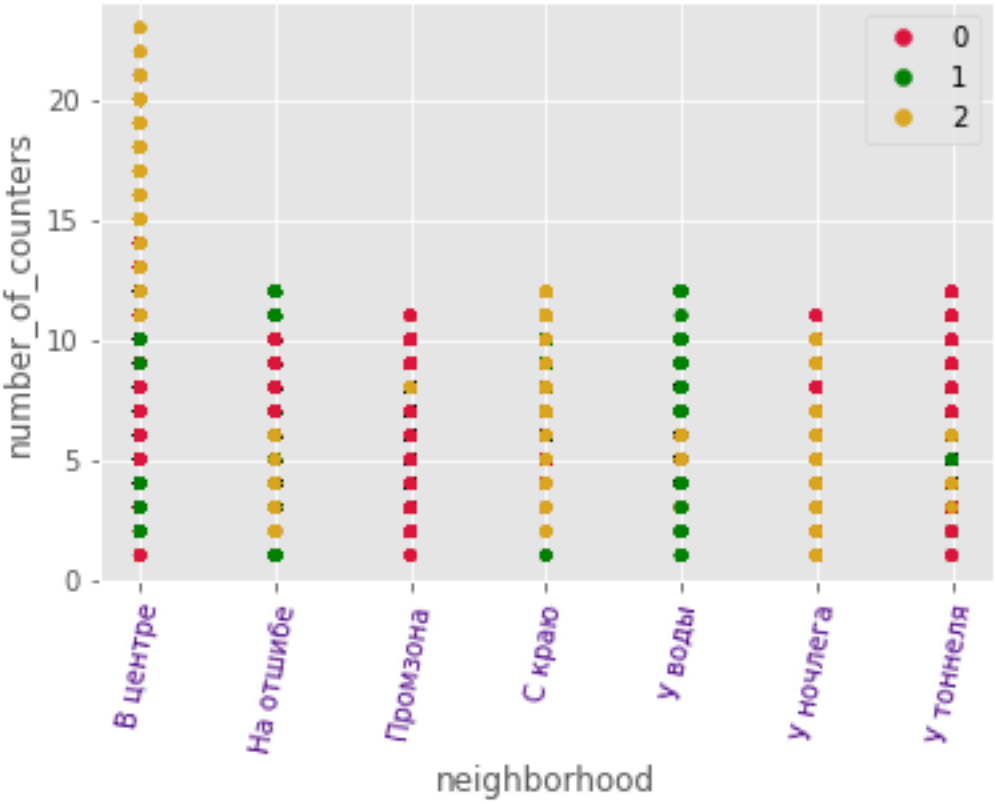

У тоннеля
У ночлега
У воды
С краю
Промзона
На отшибе
В центре
neighborhood
0.0
0.2
0.4
0.6
0.8
1.0
total_items_sold
0
1
2


нет
да
is_on_the_road
0.0
0.2
0.4
0.6
0.8
1.0
total_items_sold
0
1
2

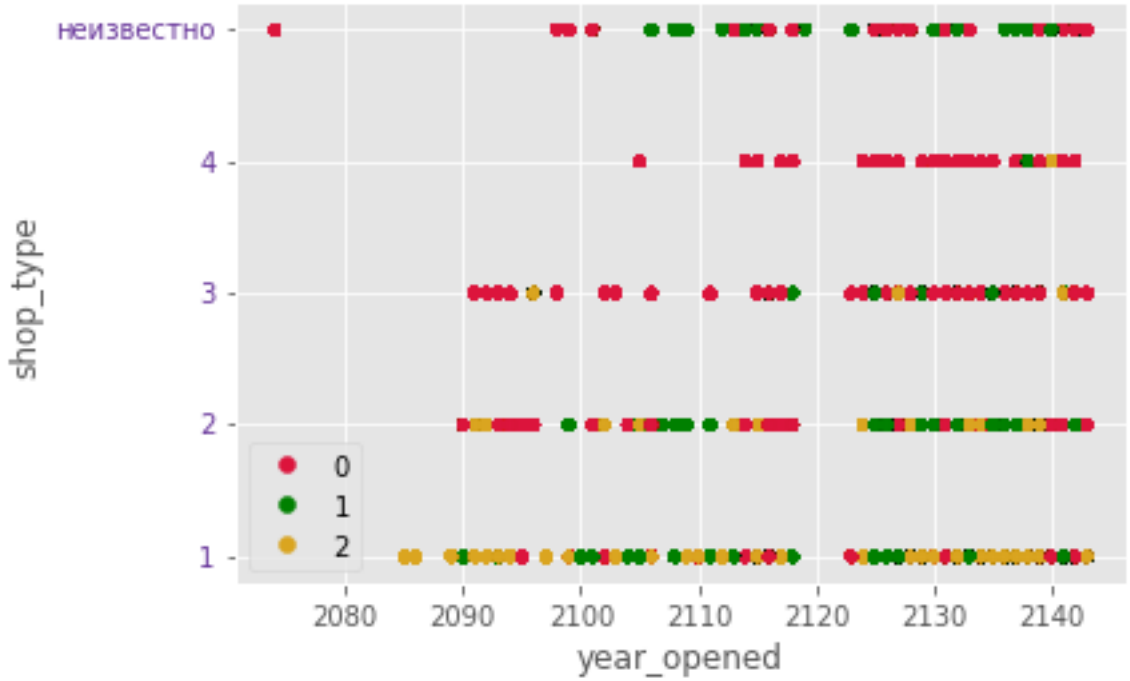




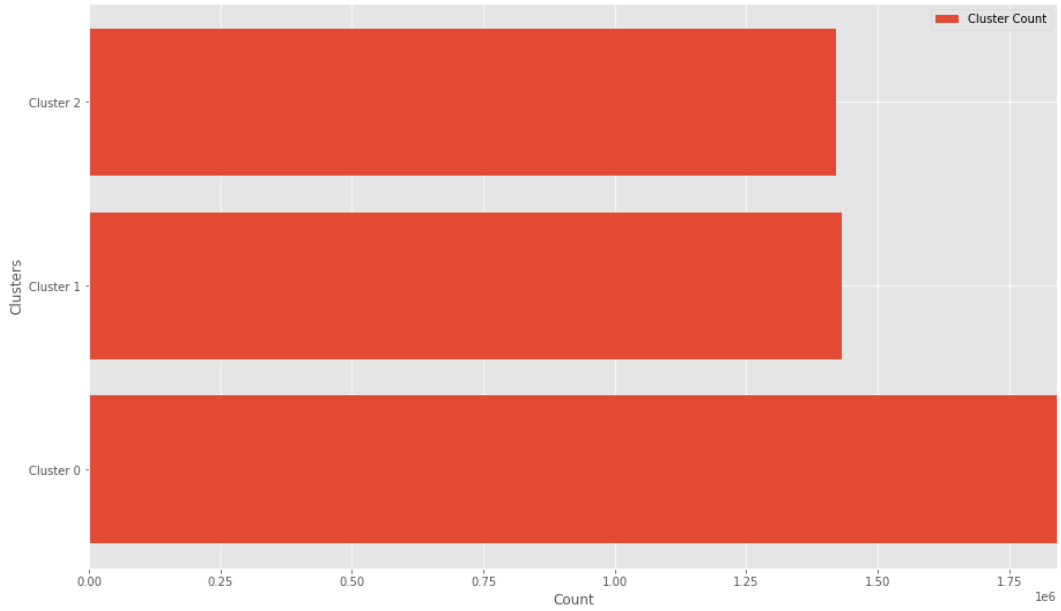
Также можно заметить, что число кассиров в магазинах 1-го кластера самое большое в центре.



В распределении по годам открытия выявить зависимости тяжело.



На диаграмме изображено количество магазинов, отнесенных к каждому кластеру. Можно заметить, что 0-й кластер превосходит все остальные.



Итог

Магазины 1-го кластера представляют собой небольшие магазины с малым числом кассиров и специализируются на продаже солярки.

Магазины 2-го кластера также продают продукты нефтепереработки (бензак и солярку), имеют большой объем продаж. Однако, часто большое число кассиров не является необходимым.

Можно предположить, что магазины, отнесенные к 1-му кластеру, могли бы повысить продажи, добавив дополнительные услуги, как это сделано в магазинах 2-го кластера. Также можно повысить эффективность, расположив магазины, продающие бензак в районе «у тоннеля».

Магазины 0-го кластера объединяют в себе несколько типов магазинов, имеют широкий ассортимент, невысокие продажи бензака. Это самый часто встречаемый кластер.

Финальная таблица с присвоенными кластерами записана в файл results_table.tsv. Так как данные с отсутствующими годами открытия были удалены из набора данных, в финальной таблице их тоже нет.

Результаты работы кластеризации могли бы быть улучшены при обработке всего набора данных целиком, не разделяя данные по датам. Также для оценки не хватает данных о стоимости товаров.