features.ipynb

Обработка пропущенных данных

Была построена тепловая карта пропущенных данных, и выявлено, что отсутствует:

- 6,7 % данных о городе;
- 0,8 % данных о нахождении на дороге;
- 7,1 % данных о наличии колодца;
- 7,2 % данных о дополнительных сервисах;
- 15,6 % данных о типе магазина.

Так как пропущено в среднем не более 10% данных каждой категории, принято решение заменить пропуски случайными значениями выбранной категории, но **с привязкой к владельцам магазинов.**

Для этого к таблице магазинов была добавлена информация о владельцах, и выявлено, что некоторые магазины имеют более одного владельца.

При заполнении пропусков идея состояла в том, что определенные владельцы имеют магазины в определенных городах, с какими-то определенными характеристиками (например Бомбисты владеют магазинами только в двух городах : Газтаун и Свинцовая Ферма, относящихся, в свою очередь, к одной локации) и было бы ошибочно определить магазины Бомбистов с отсутствующей информации о местонахождении к какой-то другой локации, отличной от Радиоактивной Пустоши.

Также, например, некоторые владельцы могут иметь магазины только определенных типов, а не всех четырех.

Поэтому принято решение заполнять пропуски **случайно выбранными значениями из разброса значений категории для каждого владельца.**

В информации о годе открытия магазина также отсутствовали данные, но так как разброс лет слишком большой даже при группировке данных по владельцам, было принято решение исключить эту фичу из анализа, так как восстановить данных случайными или средними значениями было бы некорректно.

Далее в таблицу с магазинами была добавлена информация о локации, а в таблицу продаж – день недели и номер месяца для каждой продажи.

Шкалирование данных об объеме продаж

Так как сравнивать объемы продаж бензака и, например, хлама некорректно, было решено поочередно отшкалировать (привести распределения к диапазону от 0 до 1) объемы каждого типа товара во всех магазинах и соединить в одну фичу ‘total_items_scaled’.

Генерация фичей

Из отшкалированных объемов продаж были сгенерированы фичи:

- Доля проданных товаров каждой категории в **месячном разрезе**;
- Доля проданных товаров каждой категории в **разрезе дней недели**.

Также была добавлена фича «Среднее количество продавцов/прилавков в разрезе дней недели»
Далее все категориальные фичи были закодированы.

clustering.ipynb

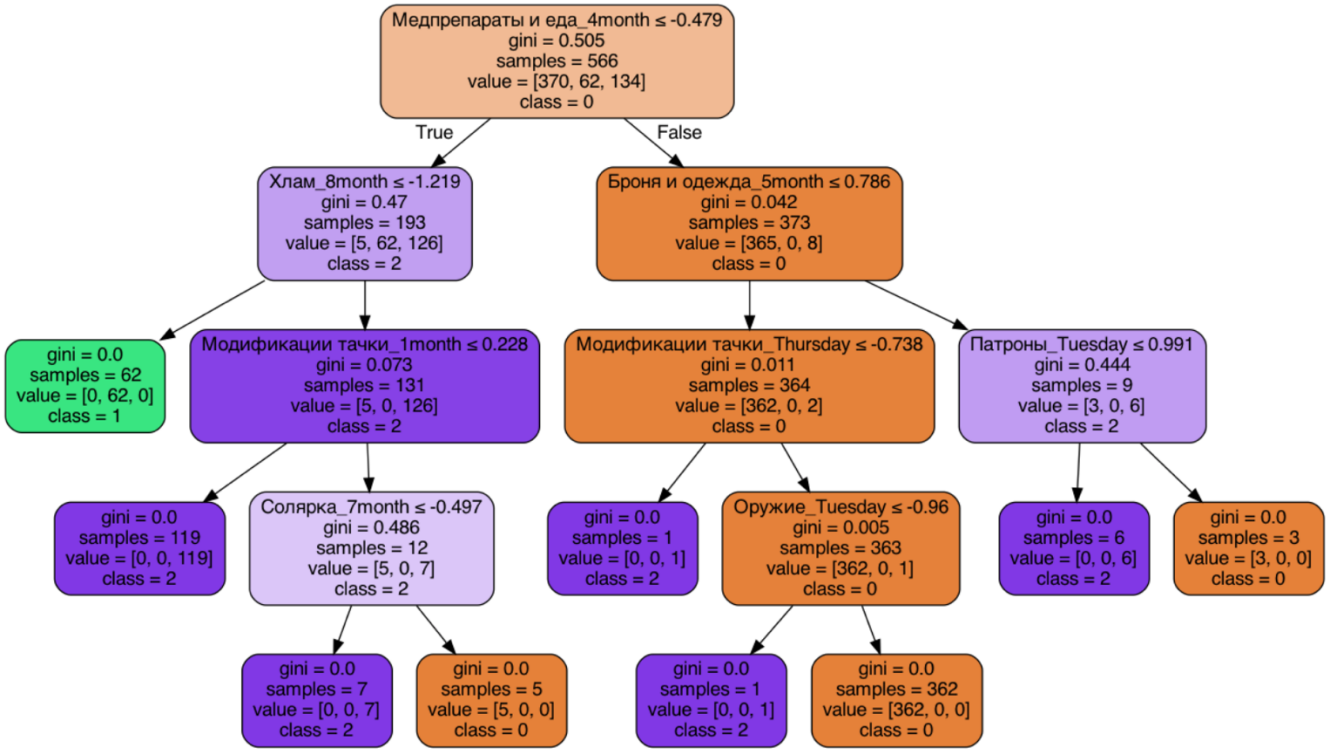
K-means

Так как размерность данных достаточно большая, было решено использовать анализ главных компонент для сокращения размерности.

Были выбраны три компоненты, так как они объясняют практически 70% дисперсии и анализировать данные в трехмерном пространстве достаточно удобно.

Далее был построен график каменистой осыпи, на котором определилось количество кластеров равное трем.

К полученной кластеризации был применен алгоритм дерева решений, для определения влияния различных фичей на распределение по кластерам.

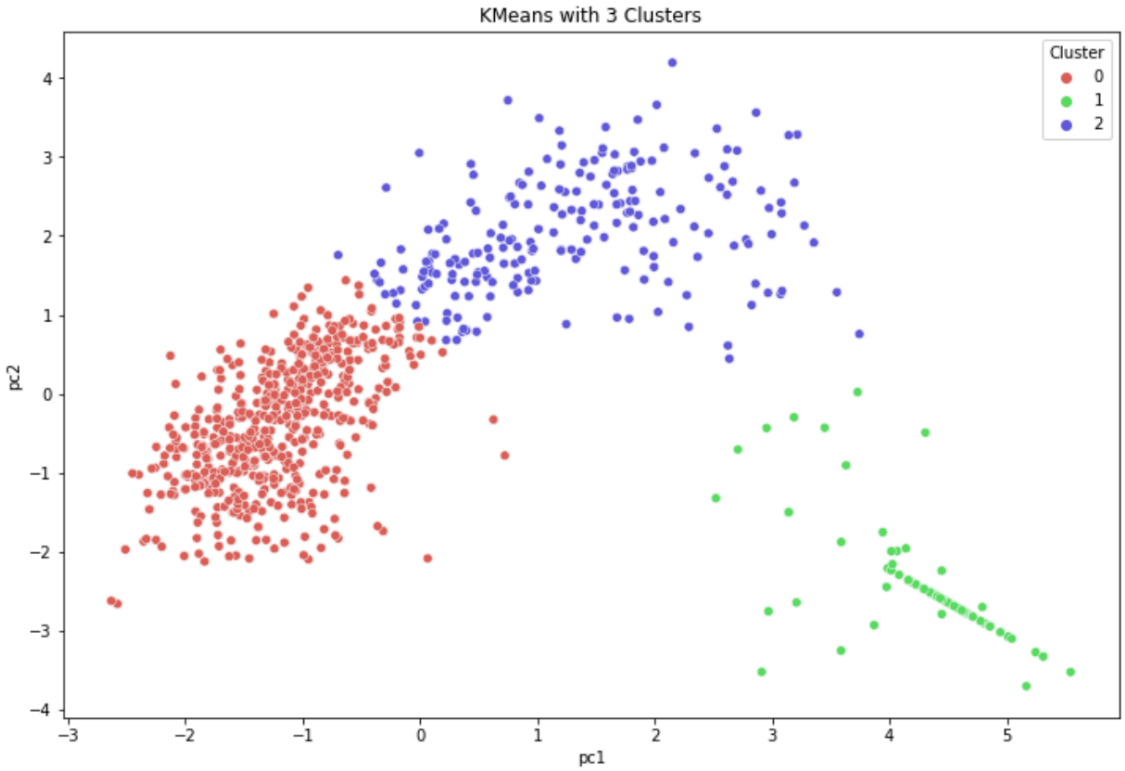


На дереве видно, что решающим главным фактором стала доля продаж медпрепаратов и еды в апреле.
Данные были урезаны только до значимых и был проведен повторный анализ главных компонент, на котором определились две главные компоненты, объясняющие 75% дисперсии.

После повторного применения алгоритма k-means к новым полученным компонентам, было определено три кластера и рассчитаны метрики качества.

Dunn index:	1.4040109210819758
Davies bouldin	0.2438718657321259
Silhouette	0.44543948939225386

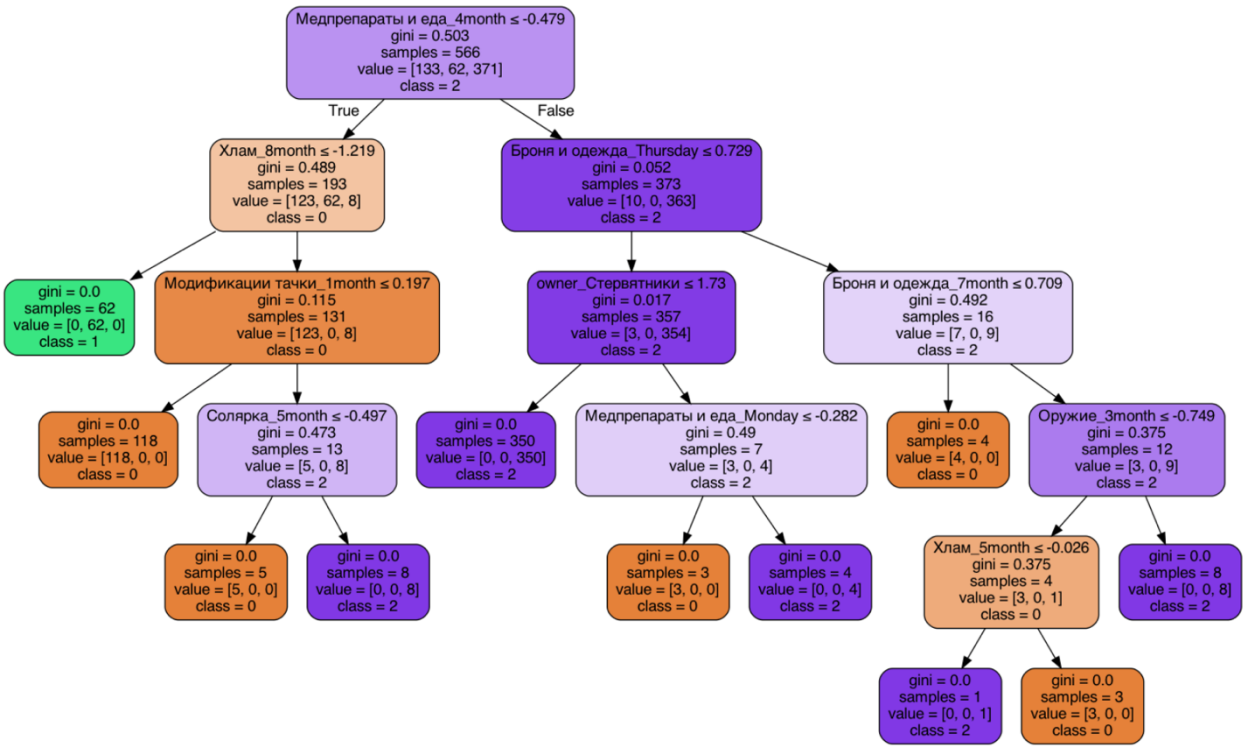
На рисунке ниже представлены кластеры в пространстве двух главных компонент.
Результаты кластеризации находятся в файле k_means.tsv.



Agglomerative

При построении иерархической агломеративной кластеризации метод главных компонент не был использован, так как ухудшал результаты работы алгоритма.

Было определено три кластера и построено дерево решений для выявления значимых фичей.



Видно, что доля продаж медпрепаратов и еды в апреле также, как и в алгоритме к-средних является главным решающим фактором, но есть и различия.

Далее кластеризация была построена заново по важным фичам. Метрики качества полученной кластеризации:

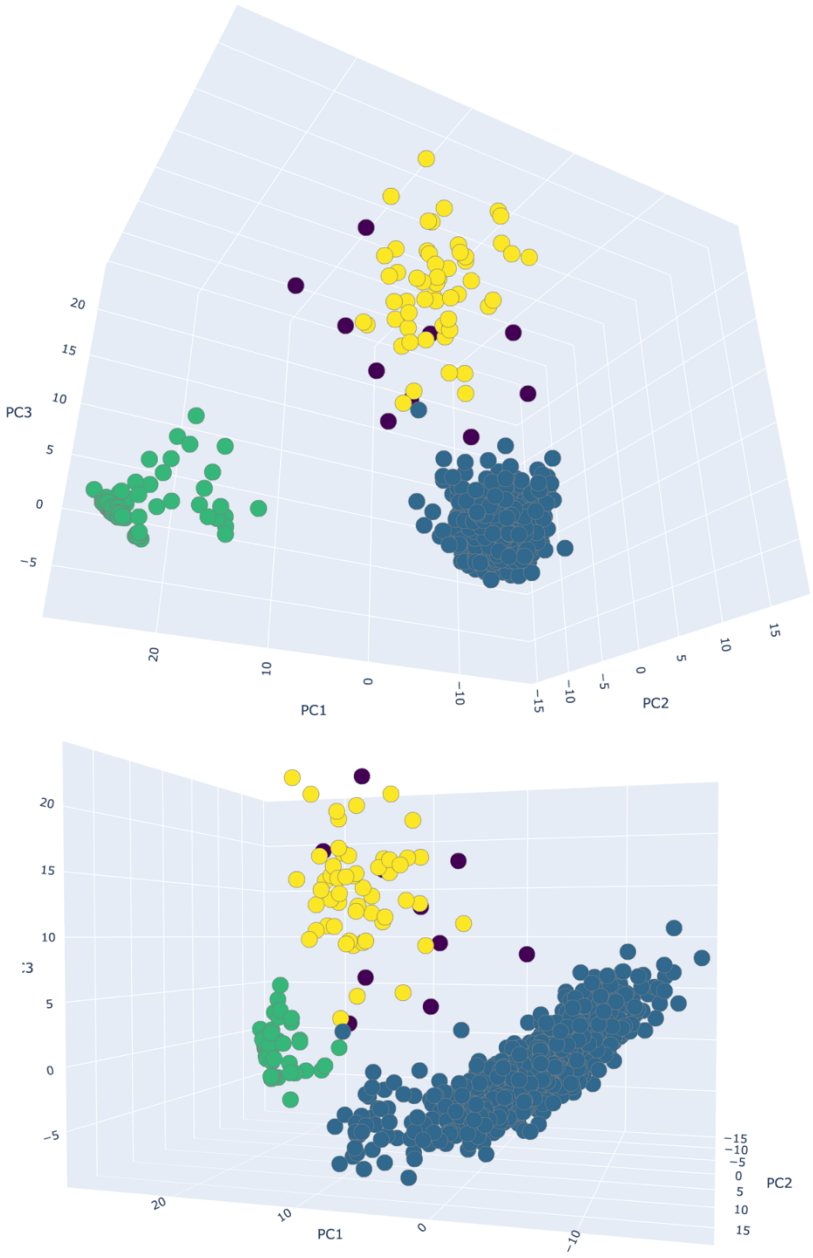
Dunn index:	0.805171220027242
Davies bouldin	0.4074887799236224
Silhouette	0.4540787920141201

Результаты кластеризации находятся в файле agglomerative.tsv.

DBSCAN

Алгоритм DBSCAN требует определения значения эпсилон и минимального количества точек. Для определения эпсилон был использован метод ближайших соседей и метод локтя. Но полученный по графику эпсилон равный двум не давал хороших результатов при работе алгоритма и был заменен на эпсилон равный пяти.

После работы алгоритма было выявлено три кластера и 10 значений, относящихся к шуму.



Для оценки качества была использована только метрика Силуэт, так как из трех выбранных две остальные не являются информативными для алгоритма DBSCAN.

Silhouette	0.559

Варьирование параметров эпсилон и минимального количества точек не дало лучшего результата, чем представлен выше.
Результаты кластеризации находятся в файле agglomerative.tsv.

Пересечения кластеров

k_means VS agglo	k_means VS DBSCAN	agglo VS DBSCAN
Размеры кластеров: 544 , 539 Пересечения: 525	Размеры кластеров: 544 , 692 Пересечения: 543	Размеры кластеров: 539 , 692 Пересечения: 539
Размеры кластеров: 94 , 89 Пересечения: 89	Размеры кластеров: 94 , 89 Пересечения: 89	Размеры кластеров: 89 , 89 Пересечения: 89
Размеры кластеров: 207 , 217 Пересечения: 193	Размеры кластеров: 207 , 54 Пересечения: 53	Размеры кластеров: 217 , 54 Пересечения: 54

Полученные при помощи различных алгоритмов кластеры являются сильно схожими, поэтому будет дана общая для всех алгоритмов интерпретация кластеров.

Общая интерпретация кластеров

В скобках указаны характеристики, относящиеся только к какому-то определенному алгоритму.

Номер кластера	Описание
0 – (K-means, DBSCAN) **2** – (Agglomerative)	В магазинах этого кластера наблюдается повышенный спрос на: • Броню и одежду (DBSCAN: больше в субботу/воскресенье); • Жидкости для тачки с ноября по март; • Медпрепараты; • Модификации тачки; • Оружие; • Патроны; • Съедобный хлам; • Хлам; • Ядер Колу. В таких магазинах есть дополнительные сервисы и колодец, а также больше всего рабочих прилавков. (k-means, agglomerative: Тип магазина в основном «1»)
1	В магазинах этого кластера самый большой спрос на бензак и солярку, а на остальные категории товаров спрос минимальный. В таких магазинах меньше всего прилавков/продавцов. (k-means, agglomerative: В основном находятся не на дороге)
2 –(K-means, DBSCAN) **0** – (Agglomerative)	В магазинах этого кластера наблюдается повышенный спрос на: • Броню и одежду (DBSCAN: больше в пятницу); • Жидкости для тачки с ноября по март (DBSCAN: с октября по декабрь); • Оружие с мая по ноябрь; • Патроны; • Съедобный хлам; • Хлам; • Ядер Колу Такие магазины в основном находятся в Свистящих степях не на дороге.

Рекомендации

- Проведение кластеризации с помощью Карт Кохонена;
- Проведение кластеризации по каждому месяцу отдельно (либо по сезонности), далее суммирование номеров кластеров и разделение по сегментам;
- Варьирование параметра random_state (возможно улучшение качества кластеризации);
- Добавление (по возможности) информации о времени покупки товара.