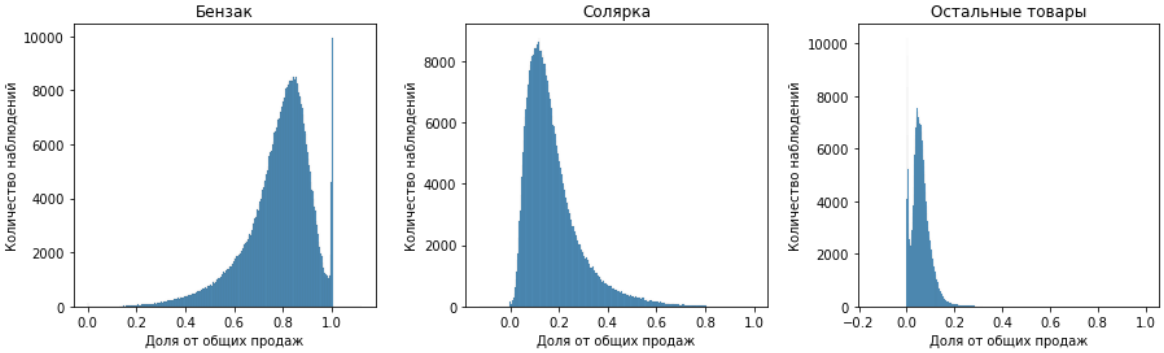
Отчет о кластеризации магазинов

Имеющиеся данные

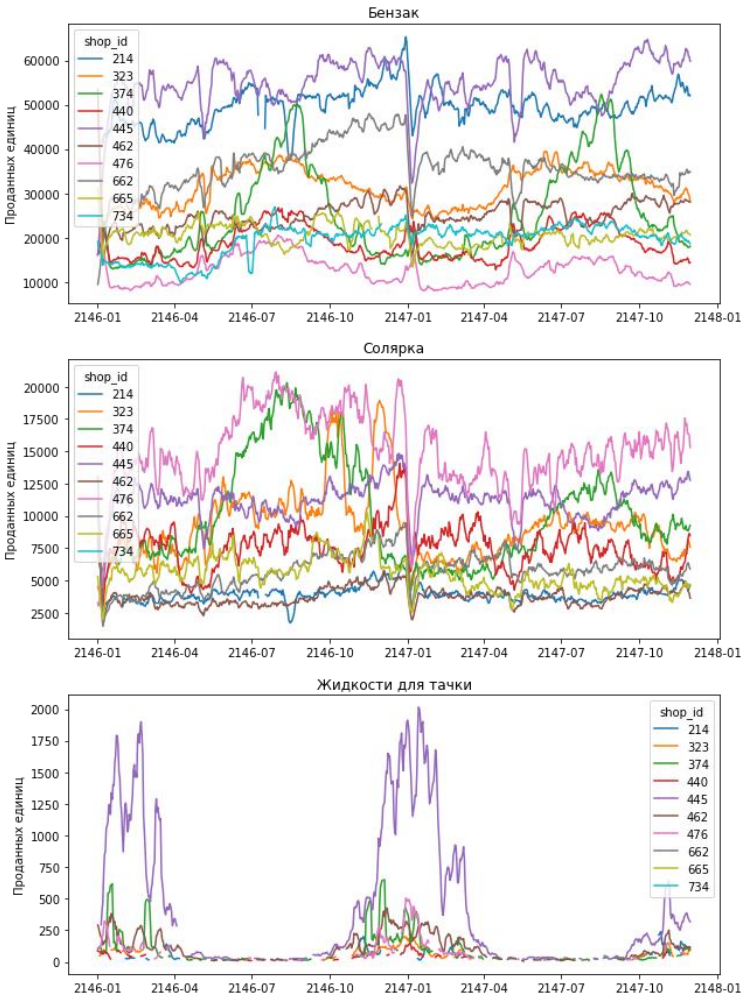
- Данные о продажах различных товаров за 2 года
- Мета информация о магазинах

Обзор имеющихся данных

Большую долю продаж магазинов составляют бензак и солярка.



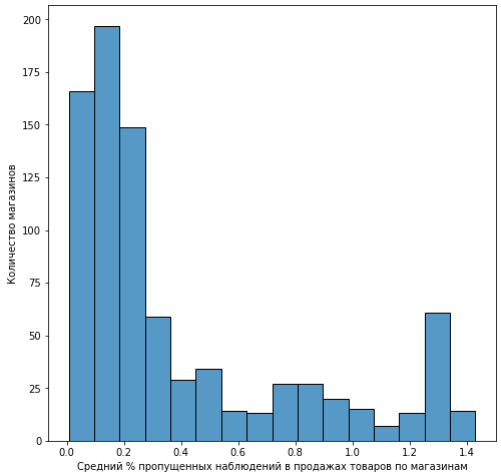
Нетопливные товары будем считать сопутствующими.



Продажи незначительно растут во времени, при этом колеблются вокруг примерно одного значения. В продажах многих товаров ярко выражена сезонность (например, в продажа жидкостей для тачек). Также в некоторых магазинах наблюдается резкое падение/увеличение спроса после начала 2147 года.

Однако, мне интересны не абсолютные значения продаж, а соотношение продаж различных товаров. Поэтому, несмотря на незначительные изменения спроса из года в год, будем использовать при кластеризации магазинов весь данный промежуток.

Также в данных о продажах есть много пропущенных дней, что накладывает ограничения на применимость моей модели кластеризации.



Задача кластеризации:

Нахождение в данных инсайдов, которые помогут в управлении магазинами.

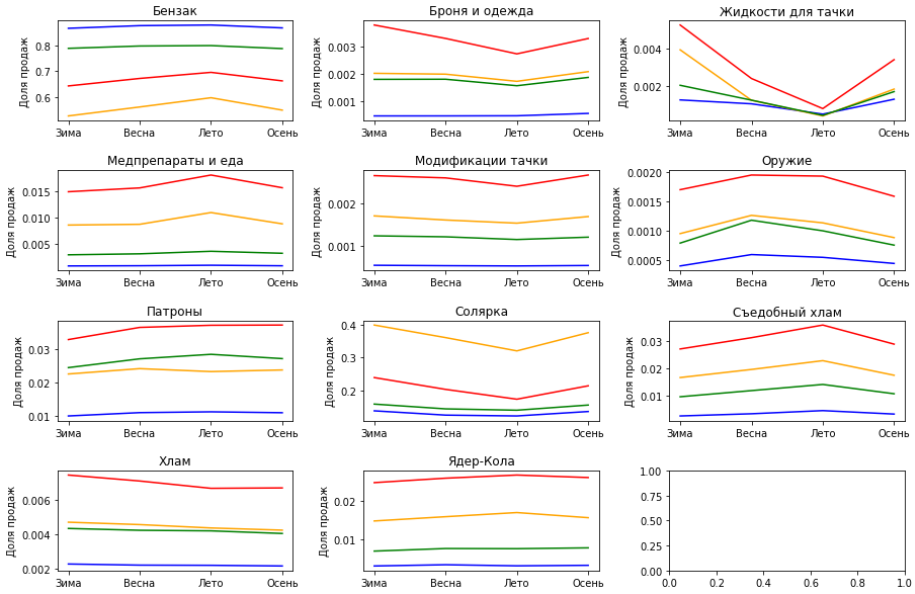
Кластеризация магазинов:

На мой взгляд **самое важное для бизнеса** отличие магазинов заключается в их **профилях продаж**. Именно понимание того, какой профиль продаж существует в конкретном магазине, поможет принимать наиболее важные решения в их управлении. Возможно, 1 магазин посещают для покупки съедобных товаров, а другие для покупки топлива.

Поэтому я провел кластеризацию следующим образом:

1) Данные о продажах поделены на **4 сезона**: зима, весна, лето, осень.
   Это сделано для того, чтобы можно было легко проинтерпретировать полученные кластеры, но при этом получить ценную информацию об изменении профиля продаж магазина в течение года.
2) Посчитаны **доли продаж** каждого товара в кажом сезоне.
   Таким образом я игнорирую информацию об абсолютных продажах и сосредатачиваюсь на относительных. Это поможет найти похожие по профилям продаж магазины вне зависимости от объема продаж в них.
3) Данные отксалированы и с помощью алгоритма k-means разбиты на **4 кластера**
   Было выбрано именно 4 кластера, т.к. при большем количестве сильно снижается качество кластеризаации и затрудняется интерпретация, а при меньшем количестве кластеров не получается извлечь из полученных кластеров полезной для бизнеса информации.

Доли средних продаж рахличных товаров по кластерам:



Описание полученных кластеров:

Синий кластер (249 магазинов):

1. Самая большая доля продаж бензака.

2. Незначительная сезонность во всех категориях. Лишь доля продаж жидкостей для тачек резко падает летом (как и во всех остальных кластерах)

Красный кластер (90 магазинов):

1. Преобладают продажи товаров, не относящихся к категории "Топливо" (Бензак, солярка)

2. Яркая сезонность долей продаж почти всех товаров. Доля продаж брони, одежды и модификаций для тачки резко проседают летом, в то время как доли продаж съедобного хлама и медпрепаратов с едой сильно растут летом

Желтый кластер (96 магазинов):

1. Самая высокая доля продажи солярки среди всех кластаеров. Летом доля продаж солярки падает.

2. Летом растет доля продаж бензака

Зеленый кластер (410 магазинов):

1. Большая доля продаж бензака, но при этом меньше, чем в синем кластере. Из этого следуют большие доли продаж остальных товаров.

Интерпретация полученных кластеров

Скорее всего, магазины синего кластера посещают люди на легковых автомобилях (т.к. высока доля продаж бензака), которые не имеют нужды в каких-либо товарах, кроме бензака.

Магазины красного кластера посещают люди на легковых автомобилях, сильно заинтересованные в нетопливных товарах. Возможно, это происходит из-за того, что через эти магазины проезжают люди, которым нужны эти товары, но при этом они либо не имеют возможности заехать в другие магазины, либо решают сэкономить время, купив топливо и другие товары в одном магазине.

Магазины желтого кластера посещают водители грузовых автомобилей (высока доля продаж солярки). При этом летом доля таких водителей снижается и эти магазины начинает посещать большее количество людей на легковых автомобилях.

Магазины зеленого кластера посещают примерно такие же люди, но при этом они больше интересуются остальными товарами

Преимущества данной кластеризации перед кластеризацией на основе географических признаков

Делая кластеризацию на основе географических данных о магазинах, мы получим кластеры магазинов, которые находятся в похожих регионах.

Но у данной кластеризации есть серьезный недостаток: ограниченность данных о географическом положении магазина. Например, 2 магазина могут находиться около дорого, однако по одной дороге ездят одни люди, а по другой – другие. Это будет влиять на то, какие товары и в каких соотношениях будут покупать в данных магазинах.

При этом мой вариант кластеризации, основанный на информации о продажах различных товаров, напрямую группирует магазины исходя их профилей их продаж, что позволяет понять, какие люди приходят в магазины, какие у них потребности и как они изменяются в течение года.

Ограничения применения данной кластеризации:

1. Плохо работает на тех магазинах, где мало данных о продажах товаров.
2. Не учитывает микроколебания профилей продаж (месячные, недельные)

Возможные пути развития исследования:

1. Кластеризация на основе данных о чеках:

Данные о чеках помогут больше понять о том, какие покупатели приходят в наши магазины. Например, можно проанализировать, какие товары чаще покупают друг с другом в разных магазинах и на основе данной информации предлагать посетителям купить в добавок еще какие-то товары.

Также, имея данные о чеках, можно попробовать закодировать продажы в магазинах за определнный период и провести кластеризаци на этих данных. Например, с помощью алгоритма word2vec построить векторное представление каждого чека и, проагрегировав полученные векторы для каждого магазина за определенный период, получить векторное представление магазинов в определенный период, на основе которого можно провести кластеризацию. Как мне кажется, такая кластеризация поможет выявить более тонкие различия между магазинами.

2. Кластеризация на основе рядов продаж:

Можно попробовать кластеризировать магазины на основе корреляции продаж различных товаров между магазинами, например, с помощью алгоритма k-medoids. Однако, этот требует **более четкой постановки бизнес-задачи** и имеет ряд ограничений. Во-первых, необходимо понимать, продажи каких товаров для магазинов важнее и исходя из этого агрегиоровать корреляции между различными товарами. Во-вторых, при таком подходе не получится учесть товары, о продажах которых пропущено много информации.

Выполнил Авилов Илья