Клиентская аналитика по данным интернет-магазина детских товаров

Выполнили: Акимов Николай
Медведева Екатерина
Тарджуманян Элен
Шумский Петр

Наша команда

Акимов Николай

Очистка данных, витрина

Шумский Петр

K-means сегментация

Медведева Екатерина

RFM сегментация, оценка кластеризаций, сводные таблицы, профили клиентов

Тарджуманян Элен

DBSCAN сегментация, презентация

Дорожная карта проекта

Подготовка данных

Создание сводных таблиц и витрины покупателей

Сегментация (RFM)

K-means Сегментация

Сегментация DBSCAN

Оценка качества сегментаций, профили

Предложения продавцу

Исследование данных, поиск пропусков и нулей

- **Идентификаторы** - есть пропуски в почте
- **Гео-признаки** - использовать не будем
- Много пропусков в признаках **категории товара** (10,8% и 15,3%)
- С отсутствующей **причиной отмены** 90% строк. Такие заказы оставим
- 20т. и 138т. строк с нулевыми **количествами** товаров
  77т. строк с пустыми значениями **маржи и ценой закупки**

Column	Count Unique	Count Zeros	Count Nans	% of NaNs	dtype
Дата	28887	0	0	0	object
ДатаДоставки	107	0	626	0.1	object
НомерЗаказаНаСайте	178418	0	4	0	object
НовыйСтатус	14	0	0	0	object
СуммаЗаказаНаСайте	14316	0	0	0	object
СуммаДокумента	14747	0	0	0	object
МетодДоставки	7	0	0	0	object
ФормаОплаты	2	0	0	0	object
Регион	508	0	5142	0.7	object
Группа2	13	0	74686	10.8	object
Группа3	93	0	74686	10.8	object
Группа4	411	0	105625	15.3	object
Тип	5	0	0	0	object
Номенклатура	73256	0	0	0	object
ТипТовара	2	0	74690	10.8	object
Отменено	2	0	0	0	object
ПричинаОтмены	31	0	624836	90.4	object
Количество	71	20307	0	0	int64
Цена	16856	0	0	0	object
СуммаСтроки	17527	0	0	0	object
ЦенаЗакупки	24782	0	76913	11.1	object
МесяцДатыЗаказа	2	0	0	0	int64
ГодДатыЗаказа	1	0	0	0	object
ПВЗ_код	4	0	228886	33.1	object
Статус	5	0	0	0	object
Гео	3	0	0	0	object
Маржа	43917	0	76913	11.1	object
СуммаУслуг	151	0	0	0	object
СуммаДоставки	151	0	0	0	object
НомерСтроки	147	0	0	0	int64
КоличествоПроданоКлиенту	69	137763	0	0	int64
ДатаЗаказаНаСайте	81	0	4	0	object
Телефон_new	114448	0	0	0	object
ЭлектроннаяПочта_new	6176	0	13117	1.9	object
Клиент	5168	0	374	0.1	object
ID_SKU	85380	0	0	0	object
ГородМагазина	64	0	0	0	object
МагазинЗаказа	89	0	686556	99.3	object

В датасете 692т. строк

Чистка данных

В качестве ID выбрали номер телефона. Убрали строки с нулевым значением телефона

Оставили:
- Заказы, которые были доставлены до клиента
- Товары, количество которых продано > 0
- Строки, в которых проданы именно товары, а не услуги
- Строки с известной и положительной ценой закупки и маржой

Провели анализ статусов заказа, пропущенных значений в категориях товаров, количественных показателей

Удалили дублирующие по смыслу и ненужные нам столбцы, такие как:
- Временные
- Геоданные
- Дополнительная информация о клиенте и заказе

Анализ статусов заказов

Нам нужны только заказы, которые дошли до покупателя и не были возвращены
Анализ статусов в совокупности даст понимание, какие из них нужно оставить

Количество по полю НомерЗаказаНаСайте	Названия столбцов					
Названия строк	В процессе	Возврат	Доставлен	Не определен	Отменен	Общий итог
В резерве	235	0	0	0	0	235
Возврат	0	1266	0	0	0	1266
Возврат из ПВЗ	0	96901	0	0	0	96901
Доставлен	0	0	251099	0	0	251099
К отгрузке	0	0	222550	0	0	222550
Комплектация Регион	1	0	0	0	0	1
Отменен	0	0	0	0	9325	9325
Отменяется	0	0	0	924	0	924
Отправлен в ПВЗ	7	0	0	0	0	7
Передан в Регион	10	0	0	0	0	10
Принят в ПВЗ	4079	0	0	0	0	4079
Расформирован ПВЗ	0	0	0	0	10808	10808
Скомплектован Регион	2390	0	0	0	0	2390
Частичный возврат	0	24369	0	0	0	24369
Общий итог	**6722**	**122536**	**473649**	**924**	**20133**	**623964**

Рассматриваем только не отмененные заказы!

В столбце [Статус] нам нужен статус Доставлен

Он включает статусы Доставлен и К отгрузке из столбца [НовыйСтатус]

[Новый статус] = Доставлен нам подходит. Проанализируем К отгрузке по другим признакам

Анализ заказов с новым статусом Доставлен

Статус Доставлен у всех заказов, купленных в магазине *
Статус К отгрузке присвоен всем товарам, полученным другими способами получения заказа

Количество по полю НомерЗаказаНаСайте	Названия столбцов		
Названия строк	Доставлен	К отгрузке	Общий итог
DPD	0	5176	5176
Pick point	0	13580	13580
Курьерская	0	153754	153754
Магазины	251099	1	251100
Самовывоз	0	42872	42872
Транспортная компания	0	7167	7167
Общий итог	**251099**	**222550**	**473649**

Оставляем строки с этими двумя статусами.
Они есть у 76% не отмененных заказов.

** Один заказ, купленный в магазине имеет статус К отгрузке. Скорее всего ошибка*

Очистка данных - итог

Было:
691539 строк
38 столбцов

Минус 49 % данных

Стало:
356673 строки
11 столбцов

Остались столбцы: Номер Заказа, Метод доставки, Форма оплаты, Категория товара, Тип товара, Количество, Цена, Сумма строки, Маржа, Дата, Телефон

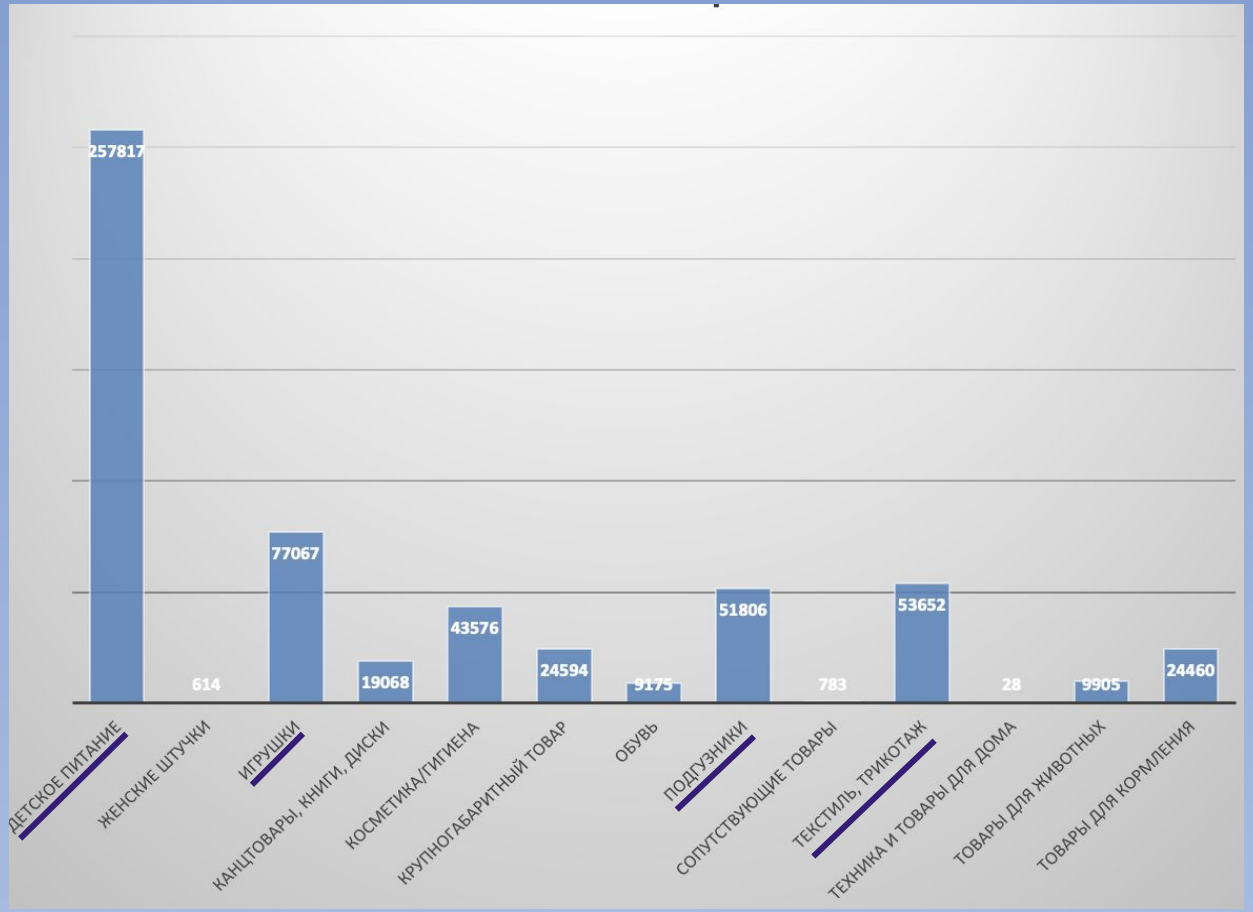
После построения сводных таблиц по товарам такие данные как сумма чека будут пересчитаны. Так же соберем таблицу заказов и по ней витрину по клиентам.

Купленные товары → Проданные заказы → Витрина клиентов

Анализ по количеству купленных товаров

Чаще всего покупали:

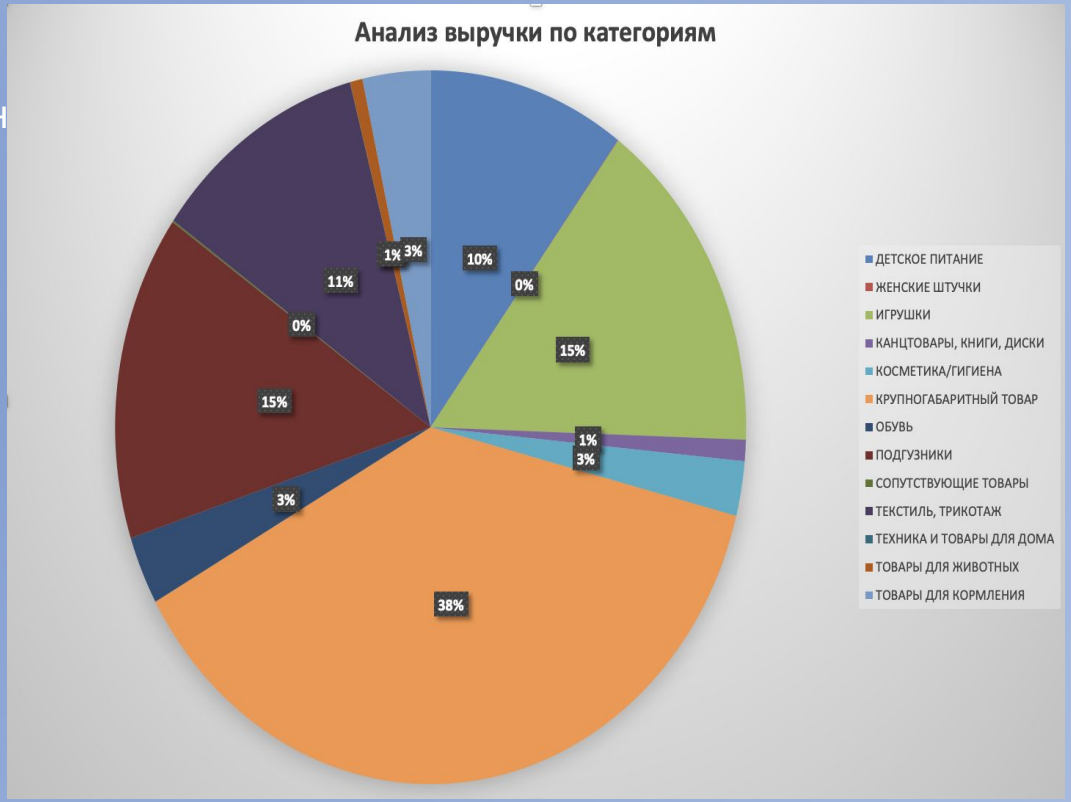
- Детское питание(257 тыс.)
- Игрушки(77 тыс.)
- Текстиль, трикотаж(53 тыс.)
- Подгузники(51 тыс.)



Анализ выручки по категориям

Наибольшую выручку принесли:

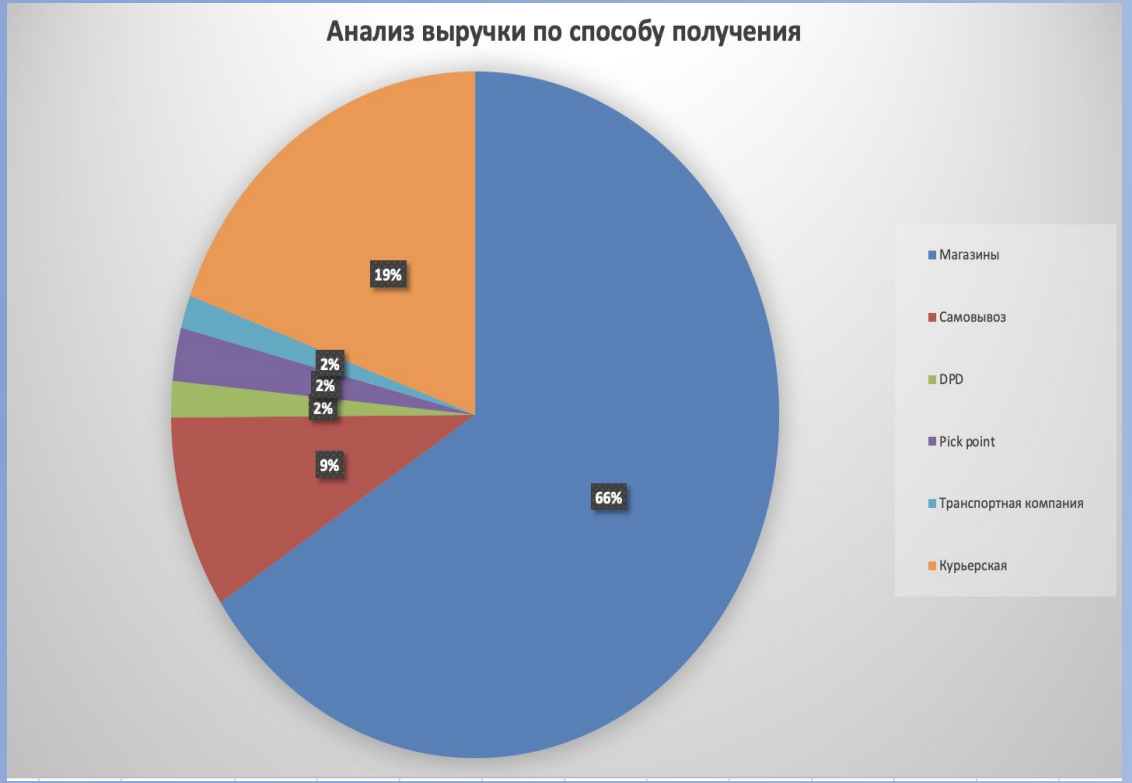
- **Крупногабаритный товар:** 125 млн
- **Игрушки:** 51 млн.
- **Подгузники:** 49 млн.
- **Текстиль трикотаж:** 36 млн.
- **Детское питание:** 33 млн.
- **Сопутствующие товары:** 25 млн.



Анализ выручки по способу получения

Наибольшую выручку принесли:

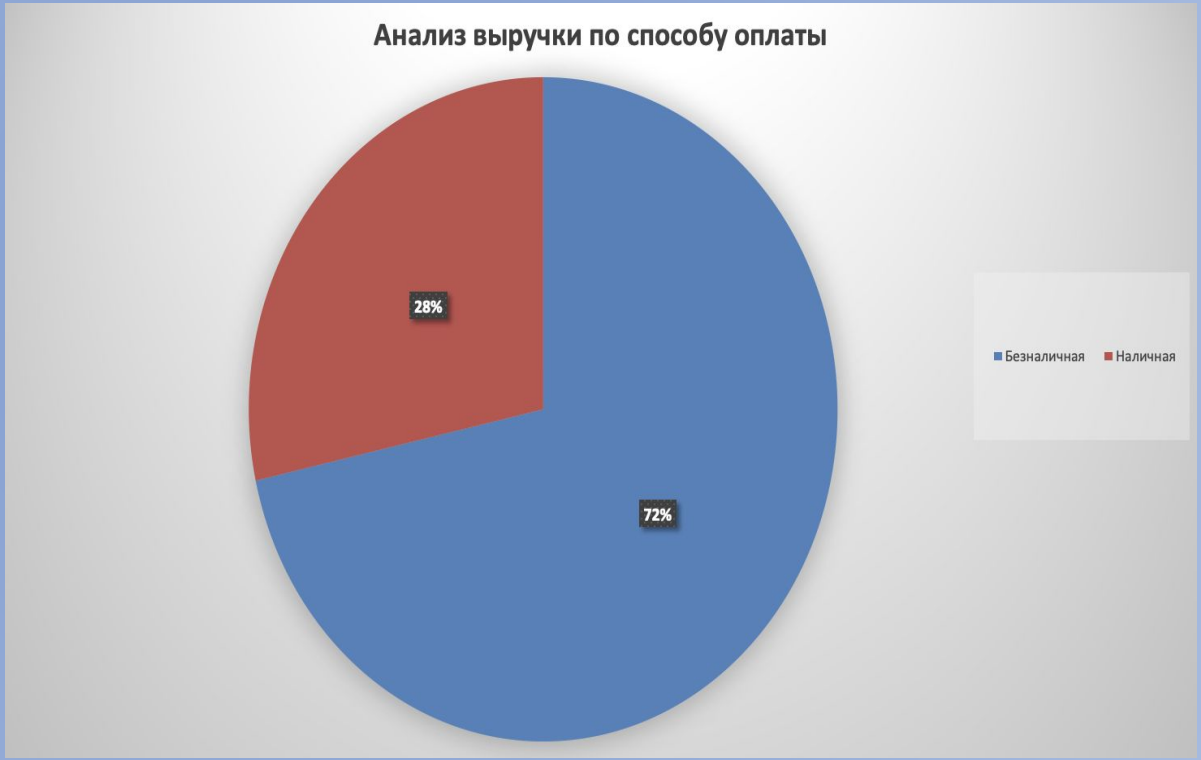
- **Магазины:** 218 млн.
- **Курьерская:** 64 млн.
- **Самовывоз:** 30 млн.



Анализ выручки по способу оплаты

Наибольшую выручку принесли:

- **Безналичная:** 238 млн.
- **Наличная:** 94 млн.



Витрина по клиентам, для сегментации собираем по данным следующих столбцов

	id	КоличествоЗаказов	ПоследнийЗаказ	КоличествоТоваров	Выручка	Маржа	Тип_ИГРУШКИ	Тип_ИНОЕ	Тип_КГТ	Тип_ОДЕЖДА	Тип_ППКП	ДЕТСКОЕ ПИТАНИЕ	ЖЕНСКИЕ ШТУЧКИ	ИГРУШКИ	КАНЦТОВАРЫ, КНИГИ, ДИСКИ	КОСМЕТИКА/ ГИГИЕНА
0	55525753-51545355524977	1	17.03.2017 0:00	1	1199.0	789.00	1	0	0	0	0	0	0	1	0	0
1	55574848-48494948544878	1	21.03.2017 0:00	2	1052.0	367.00	2	0	0	0	0	0	0	2	0	0
2	55574848-48495057545270	1	01.03.2017 0:00	2	360.0	150.97	0	0	0	2	0	0	0	0	0	0
3	55574848-48505056534872	1	17.03.2017 0:00	1	2840.0	1488.85	0	0	1	0	0	0	0	0	0	0
4	55574848-49504849525374	1	17.03.2017 0:00	1	24225.0	5371.00	0	0	1	0	0	0	0	0	0	0

Произведена группировка данных по клиентам(по номеру телефона)

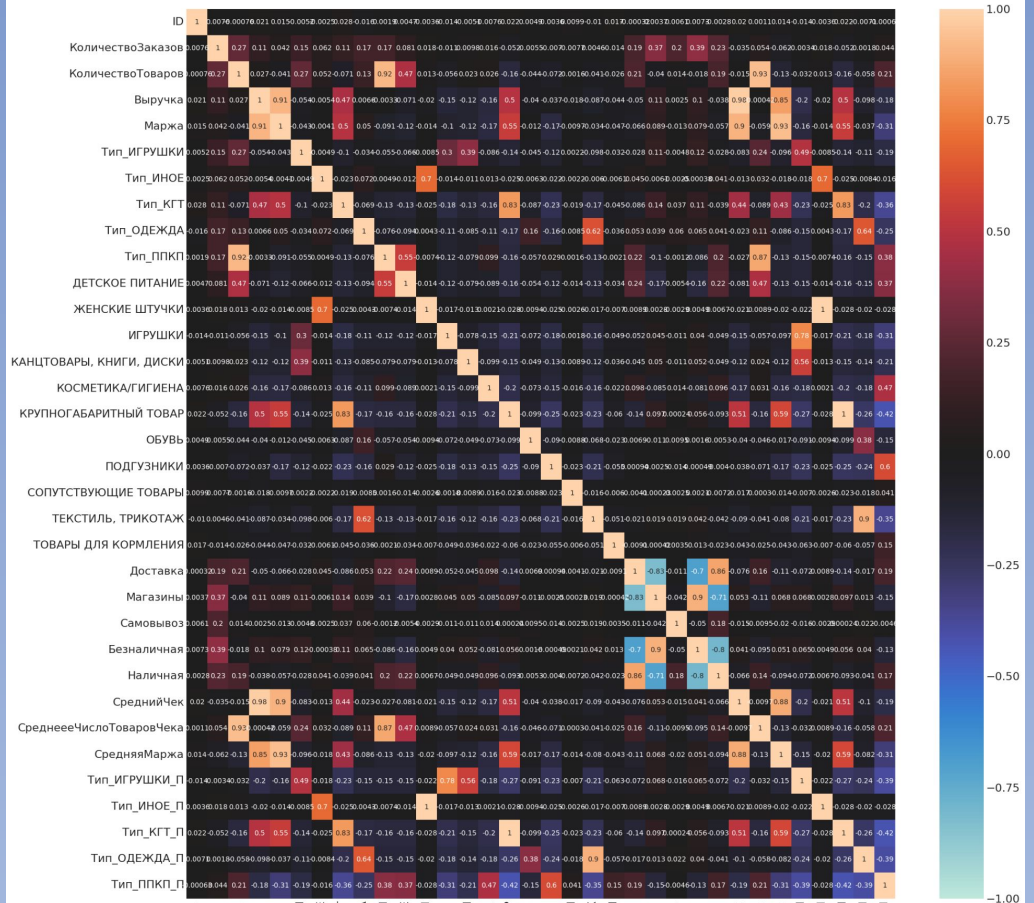
Категориальные переменные закодированы One Hot Encoding

КРУПНОГАБАРИТНЫЙ ТОВАР	ОБУВЬ	ПОДГУЗНИКИ	СОПУТСТВУЮЩИЕ ТОВАРЫ	ТЕКСТИЛЬ, ТРИКОТАЖ	ТОВАРЫ ДЛЯ КОРМЛЕНИЯ	Доставка	Магазины	Самовывоз	Безналичная	Наличная
0	0	0	0	0	0	1	0	0	0	1
0	0	0	0	0	0	0	1	0	1	0
0	0	0	0	2	0	0	1	0	1	0
1	0	0	0	0	0	0	1	0	1	0
1	0	0	0	0	0	0	1	0	1	0

Корреляция данных

Удалим:

- Столбцы, в которых считалось соотношение в долях для типа товара ('Тип_ИГРУШКИ_П','Тип_ИНОЕ_П','Тип_КГТ_П','Тип_ОДЕЖДА_П','Тип_ППКП_П') - коррелируют со столбцами типа товаров
- Средний чек, он сильно коррелирует со средней маржой
- Среднее число товаров в чеке - есть сильная корреляция с количеством товаров



RFM-сегментация

Как давно клиент делал заказ?

Мы берем в качестве данного показателя количество месяцев. Чем меньше времени прошло с момента последней активности, тем больше вероятность того, что клиент повторит действие.

Как часто клиент делал заказ?

Количество заказов, которые делал клиент.
Чем больше заказов, тем большая вероятность того, что клиент повторит его в будущем

Какая общая сумма выручки каждого клиента?

Чем больше денег потрачено, тем большая вероятность того, что клиент сделает заказ

Разделим клиентов на группы для построения коммуникаций с ними

Результаты RFM

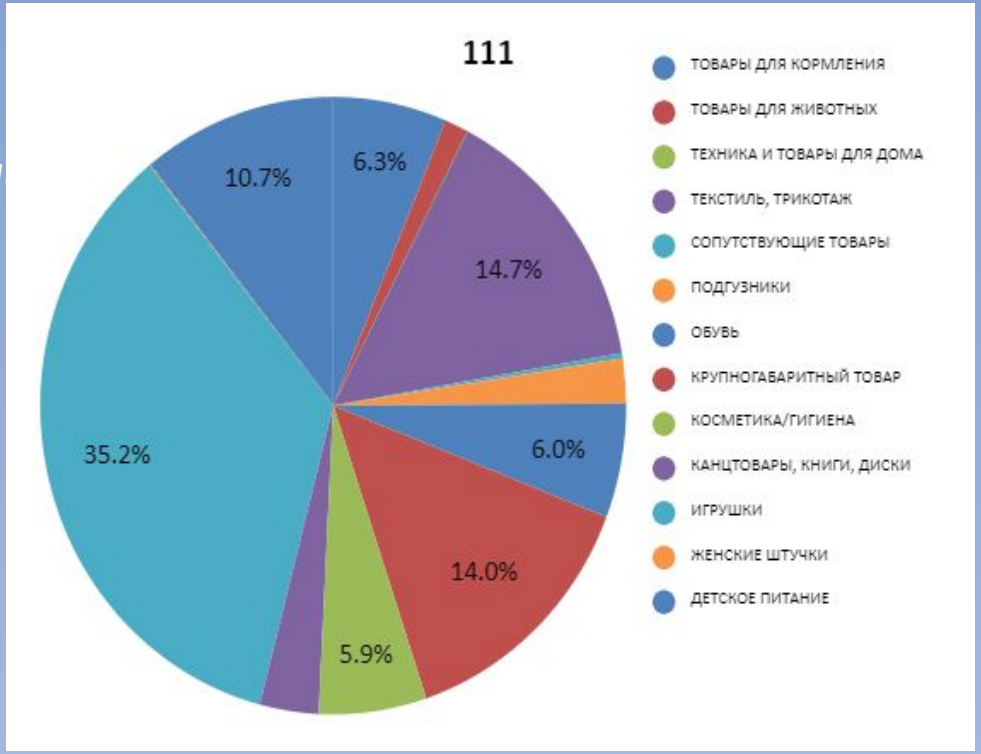
- **111** - сегмент тех, которые сделали последний заказ **в первые полгода**,количество заказов **не превышало 3**, выручка **не превышала 5000**
- **112** - сегмент тех, которые сделали последний заказ **в первые полгода**,количество заказов **не превышало 3**, выручка **превышала 5000**
- **121** - сегмент тех, которые сделали последний заказ **в первые полгода**,количество заказов **превышало 3**, выручка **не превышала 5000**
- **122** - сегмент тех, которые сделали последний заказ **в первые полгода**,количество заказов превышало 3, выручка **превышала 5000**

Результаты RFM

- **211** - сегмент тех, которые сделали последний заказ **во вторые полгода**,количество заказов **не превышало 3**, выручка **не превышала 5000**
- **212** - сегмент тех, которые сделали последний заказ **во вторые полгода**,количество заказов **не превышало 3**, выручка **превышала 5000**
- **221** - сегмент тех, которые сделали последний заказ **во вторые полгода**,количество заказов **превышало 3**, выручка **не превышала 5000**
- **222** - сегмент тех, которые сделали последний заказ **во вторые полгода**,количество заказов **превышало 3**, выручка **превышала 5000**

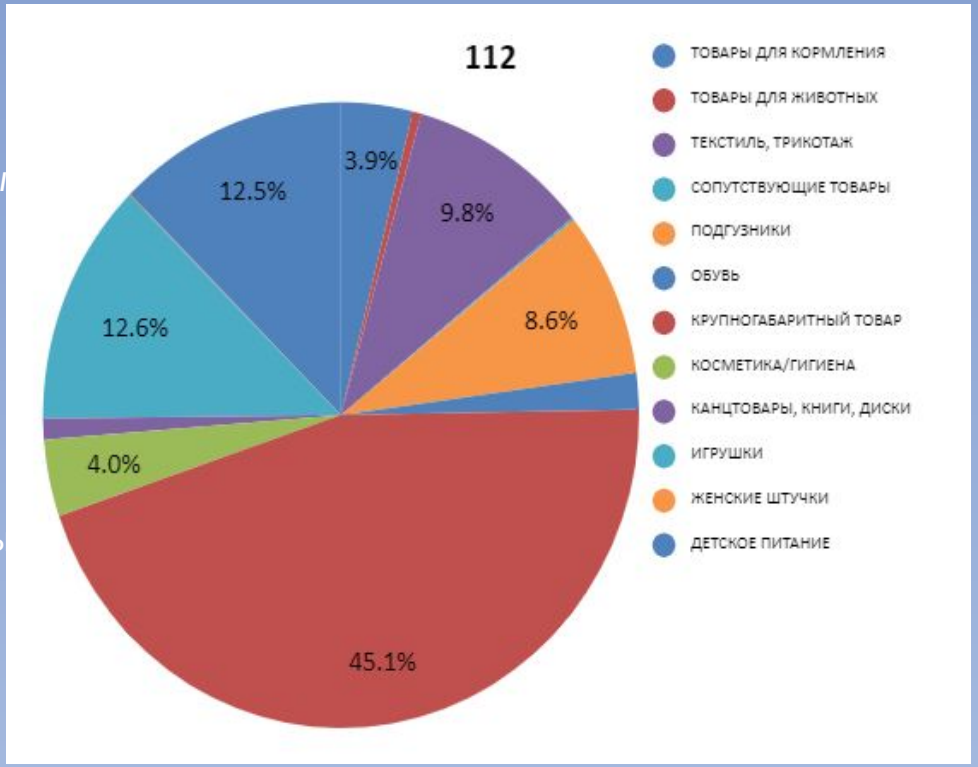
1 группа

111(57% - 46010 клиентов)
Оплата чаще всего производилась безналичны расчетом(72%)
Доставка - магазины(66%), доставка(18%) и самовывоз(14%)
Средний чек - 1833
Кол-во товаров - 4
Кол-во заказов - 1
Самые популярные товары - игрушки(35%), крупногабаритный товар(14%), текстиль, трикотаж(15%)



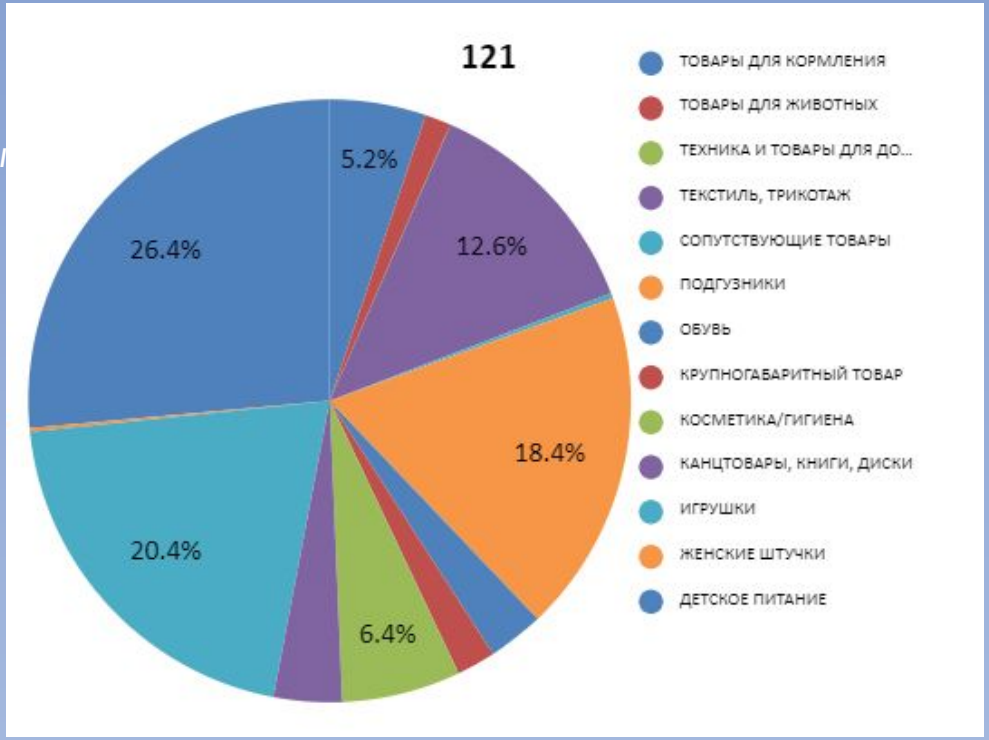
2 группа

112(14% - 11305 клиентов)
Оплата чаще всего производилась безналичны расчетом(70%)
Доставка - магазины(64%), доставка(29%) и самовывоз(7%)
Средний чек - 8922
Кол-во товаров - 10
Кол-во заказов - 1
Самые популярные товары - крупногабаритны товар(45%), игрушки(13%),текстиль,трикотаж (10%)



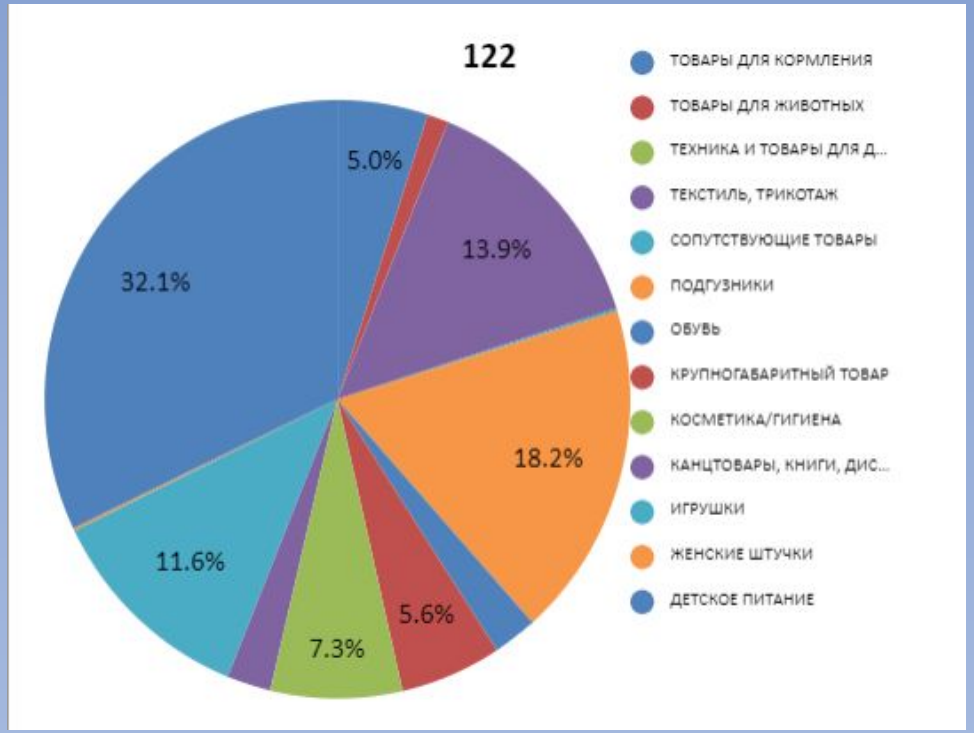
3 группа

121(2% - 1754 клиентов)
Оплата чаще всего производилась безналичным расчетом(59%)
Доставка - магазины(57%), доставка(9%) и самовывоз(34%)
Средний чек - 986
Кол-во товаров - 14
Кол-во заказов - 3
Самые популярные товары - игрушки(20%), детское питание(27%), подгузники(19%)



4 группа

122(5% - 4036 клиентов)
Оплата чаще всего производилась безналичным расчетом(60%)
Доставка - магазины(56%), доставка(21%) и самовывоз(22%)
Средний чек - 3168
Кол-во товаров - 33
Кол-во заказов - 4
Самые популярные товары - детское питание (32%), подгузники(18%), текстиль, трикотаж (14%)



5 группа

211(18% - 14282 клиентов)

Оплата чаще всего производилась безналичны
расчетом(72%)

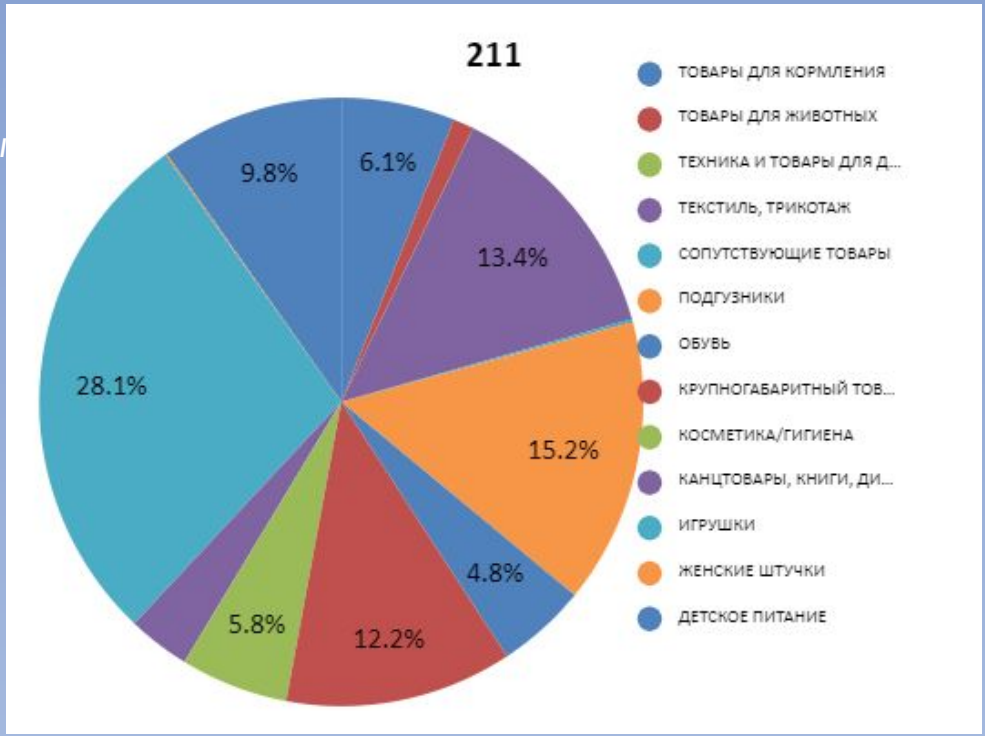
Доставка - магазины(65%), доставка(22%) и самовывоз(13%)

Средний чек - 1867

Кол-во товаров - 4

Кол-во заказов - 1

Самые популярные товары - игрушки(28%), подгузники(15%), текстиль, трикотаж(14%)



6 группа

212(4% - 3217 клиентов)

Оплата чаще всего производилась безналичны расчетом(70%)

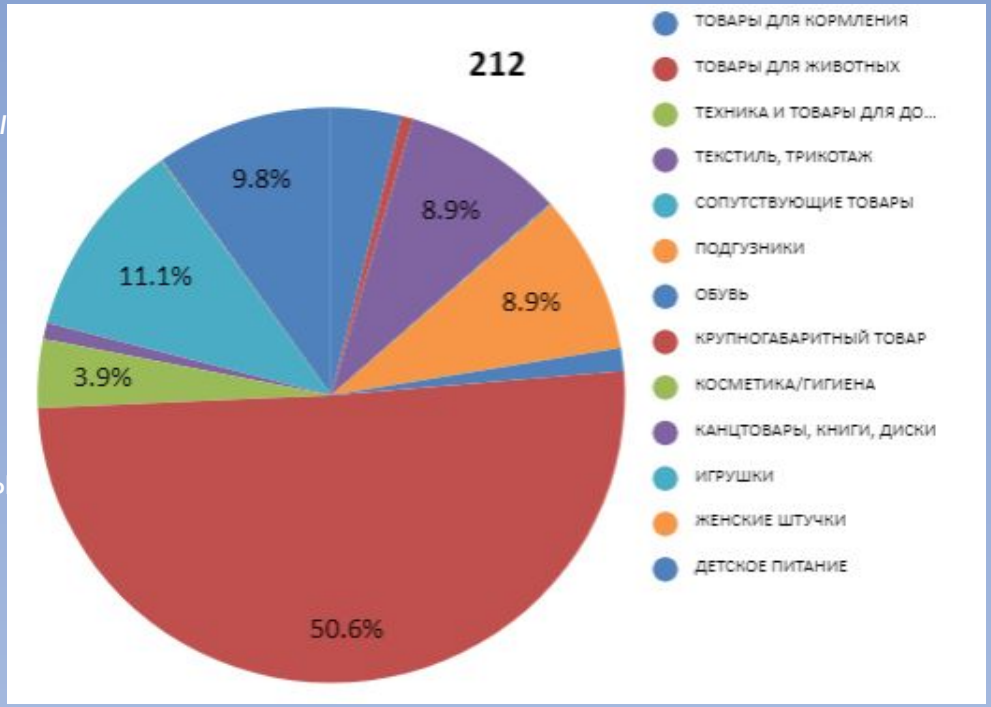
Доставка - магазины(64%), доставка(30%) и самовывоз(7%)

Средний чек - 9544

Кол-во товаров - 8

Кол-во заказов - 1

Самые популярные товары - крупногабаритны товар(50%), игрушки(11%), детское питание (10%)



7 группа

221(0,0017% - 139 клиентов)

Оплата чаще всего производилась безналичным расчетом(53%)

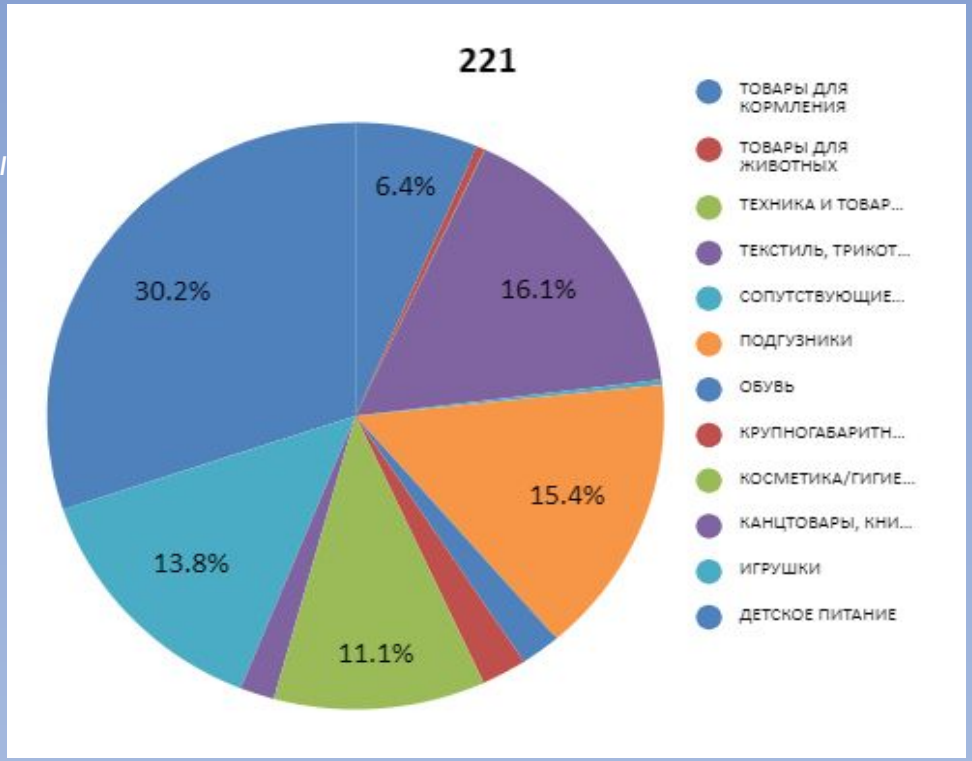
Доставка - магазины(51%), доставка(14%) и самовывоз(35%)

Средний чек - 998

Кол-во товаров - 12

Кол-во заказов - 3

Самые популярные товары - детское питание (30%), текстиль, трикотаж(16%), подгузники (15%)



8 группа

222(0,0026% - 212 клиентов)

Оплата чаще всего производилась безналичны расчетом(61%)

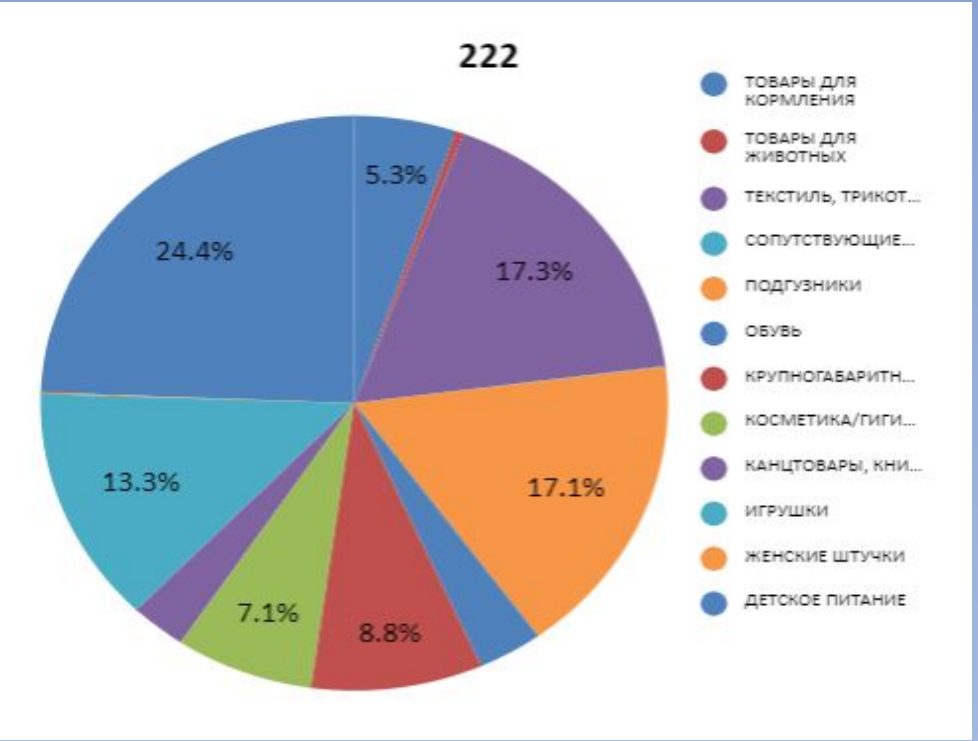
Доставка - магазины(58%), доставка(21%) и самовывоз(22%)

Средний чек - 3263

Кол-во товаров - 22

Кол-во заказов - 3

Самые популярные товары - детское питание (25%), текстиль, трикотаж(17%), подгузники (17%)



Предложения для клиентов на основе анализа

Наиболее популярные товары:

- Игрушки
- Детское питание
- Текстиль
- Трикотаж
- Подгузники
- Крупногабаритный товар

- Стимулировать услуги доставки, так как во всех сегментах преобладает выдача заказа в магазине
- Упростить наличную оплату, так как во всех сегментах наиболее популярен безналичный расчет

Анализ по наибольшей выручке:

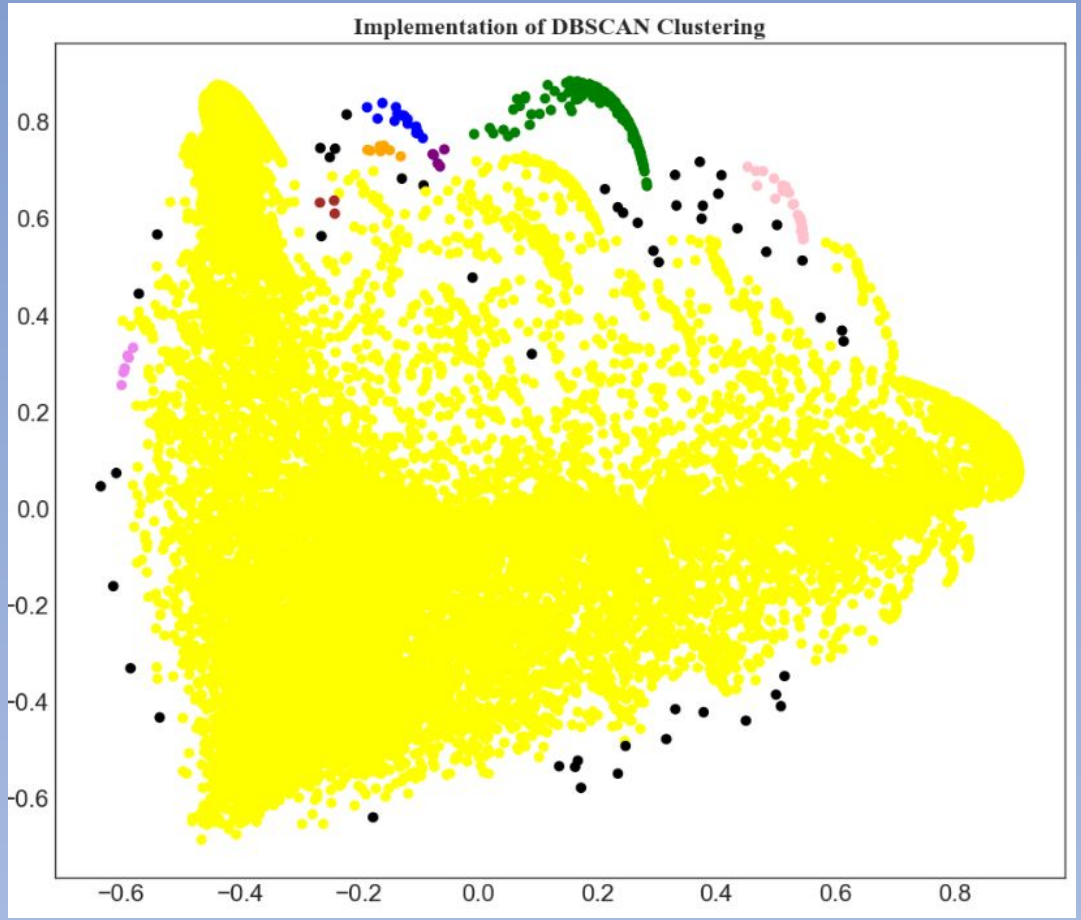
	ДЕТСКОЕ ПИТАН	НСКИЕ ШТУ	ИГРУШКИ	АРЫ, КНИГ	МЕТИКА/ГИГИ	ГАБАРИТНЫЙ	ОБУВЬ	ПОДГУЗНИКИ	ТВУЮЩИЕ	СТИЛЬ, ТРИКО	И ТОВАРЫ	Ы ДЛЯ ЖИВ	Ы ДЛЯ КОРМ	ЛЕНИЯ
111	13081926,6	190477,3	34639799,2	5311666	10993994,4	18741442	6692424	22068717,6	588508,5	18910649,3	13604	1444671	9847234,21	
112	19239947,6	479340,4	24536672,8	4910791	16507812,1	78785261,7	4779349	24198785,9	850146,6	20326953,9	0	1250282	16149674,8	
121	2260605,74	63114,97	2722001,05	806679,8	1431184,87	716574,62	670910,2	2755665,76	114760,5	1970543,86	11947	179016,2	1394512,65	
122	26965368,9	1033140	27077746,6	8941703	22243541,7	21733002	9227867	31380481,8	1531853	24303050,9	37130	2626142	20176048,4	
211	4032126,71	75786	9410366,36	1706035	3609484,58	5366673,73	1865429	7815522,14	133492	5738656,36	20553	409458,7	3062240,82	
212	4049332,29	105222	5455490,09	963669,8	3939978,11	24540808,7	933169,3	6348848,55	197667	5035220,52	21662	394655	4286496,14	
221	198385,11	0	171314,1	21094	158480	53896	52818	221576,11	7573	157537,99	2376	10835	122663	
222	852848,23	6958	1001595,58	213484,8	746308,55	988469,79	337105	1154336,16	9310	904865,07	0	34915	645761,21	

Предложения по сегментам

111	Получите скидку 10% на доставку при приобретении товаров из категории “Игрушки”
112	Приобретите крупногабаритный товар на сумму от 50 тыс. и получите бесплатную доставку
121	Приобретите 3 и более товара из категорий “Подгузники”, “Детское питание” или “Игрушки” и получите подарок на сумму до 1000 рублей на следующий заказ
122	Приобретите товары из категории “Детское питание” и получите скидку 10% на доставку
211	При оформлении 2 заказов получите скидку 5% на следующий
212	Получите скидку 5000 рублей на технику и товары для дома при покупке товаров из категории “Крупногабаритный товар” на сумму от 100 тыс. рублей
221	При заказе товаров из категории “Детское питание” получите скидку 10% на доставку
222	При заказе на сумму от 7000 рублей получите бесплатную доставку в подарок

DBSCAN сегментация

DBSCAN означает пространственную кластеризацию на основе плотности для приложений с шумом. Это – неконтролируемый алгоритм кластеризации, который используется для поиска базовых выборок с высокой плотностью для расширения кластеров.

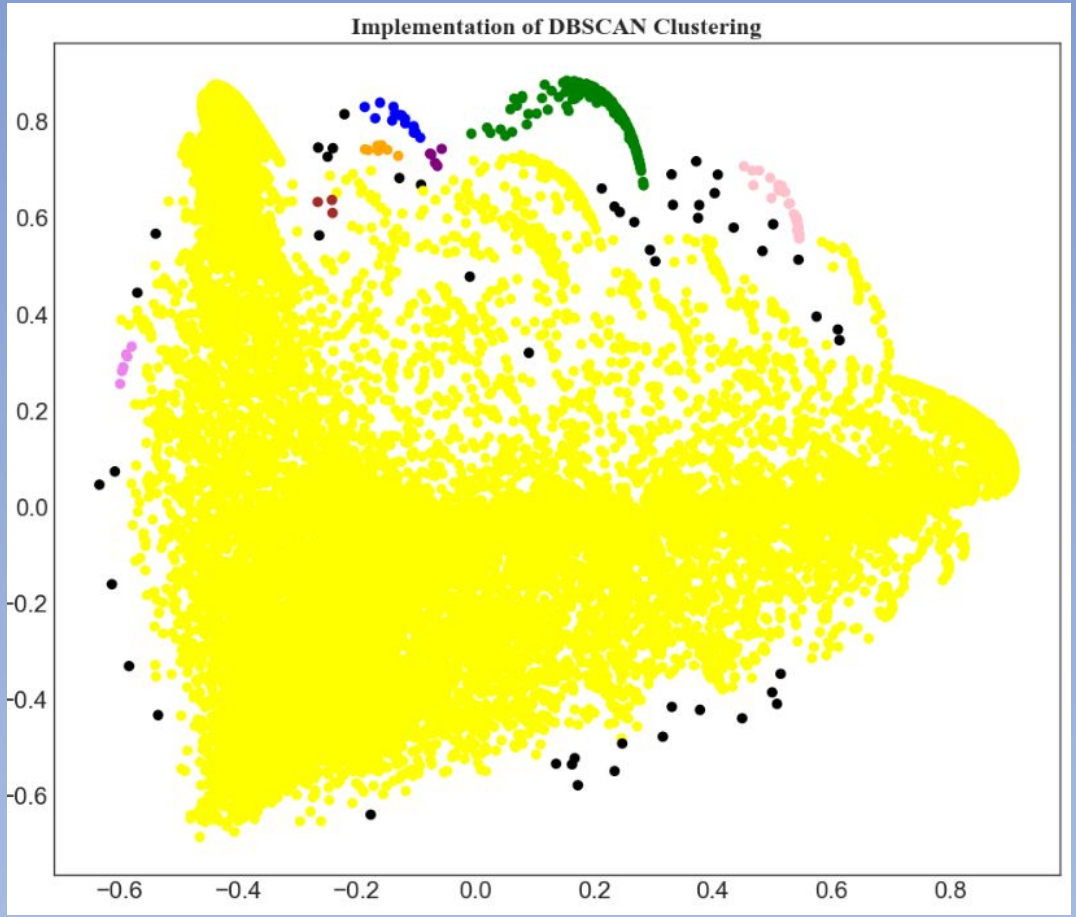


DBSCAN сегментация

DBSCAN имеет 2 параметра : радиус и кол-во точек, которые он может охватить. Мы взяли следующие параметры:

eps=0.036, min_samples=4

Итог:
DBSCAN не подходит для нашей задачи. Он не используется, если кол-во параметров большое. Поэтому у нас 99% точек вошли в один кластер.



K-means сегментация

Метод кластеризации, который разбивает множество элементов векторного пространства на заранее известное число кластеров k.
Алгоритм стремится минимизировать среднеквадратичное отклонение на точках каждого кластера.
На каждой итерации перевычисляется центр масс для каждого кластера, полученного на предыдущем шаге, и векторы разбиваются на кластеры снова в соответствии с тем, какой из новых центров ближе.
Алгоритм заканчивается, когда на очередной итерации не происходит изменение кластеров.

Минусы:

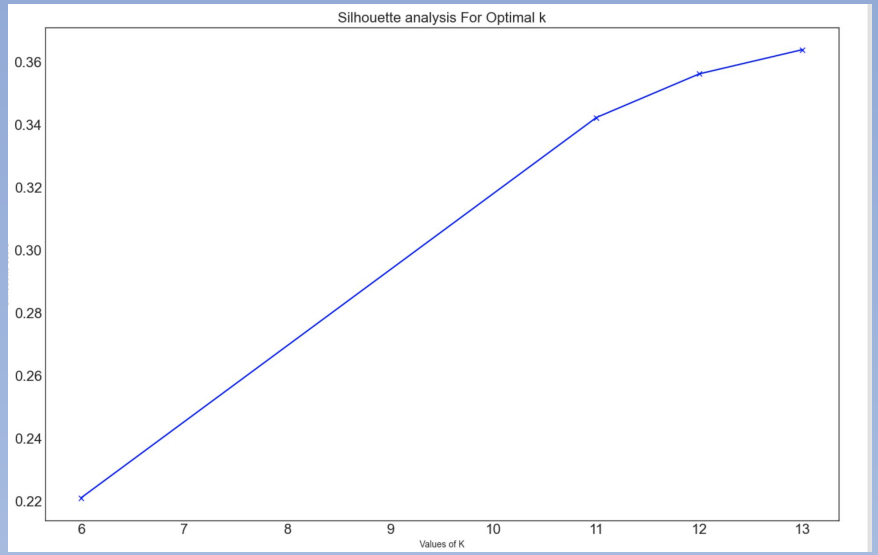
- Нужно заранее знать число кластеров.
- Алгоритм чувствителен к выбору начальных центров кластеров.

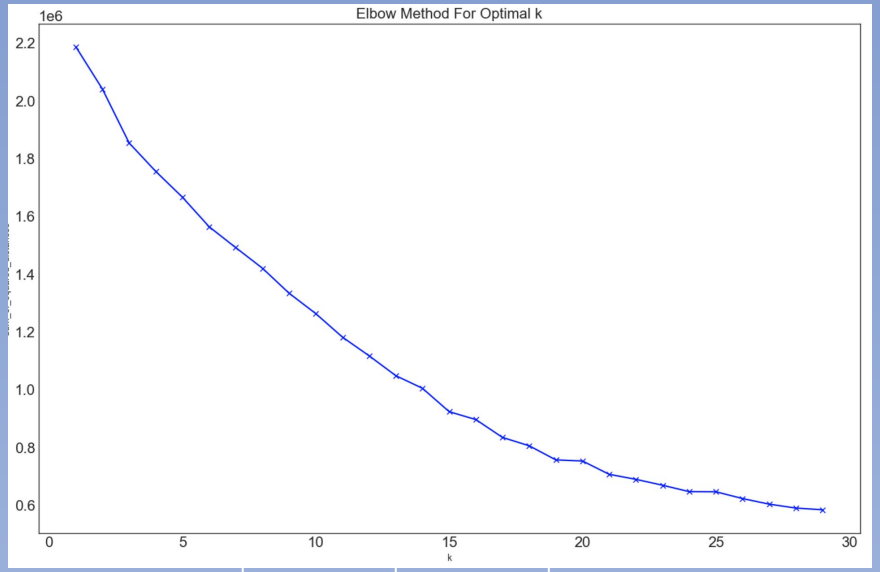
Плюсы:

- Скорость реализации
- Масштабируемость до огромных наборов данных
- Простота реализации

Выбор количества кластеров для модели K-means

- **Метод “локтя”** показал, что что лучшее количество кластеров - **17**.
- **Метод “силуэта”** показал, что лучшее количество кластеров - **13**.
- **Метод “Калински-Харабаш”** показал, что лучшее количество кластеров - **13.**

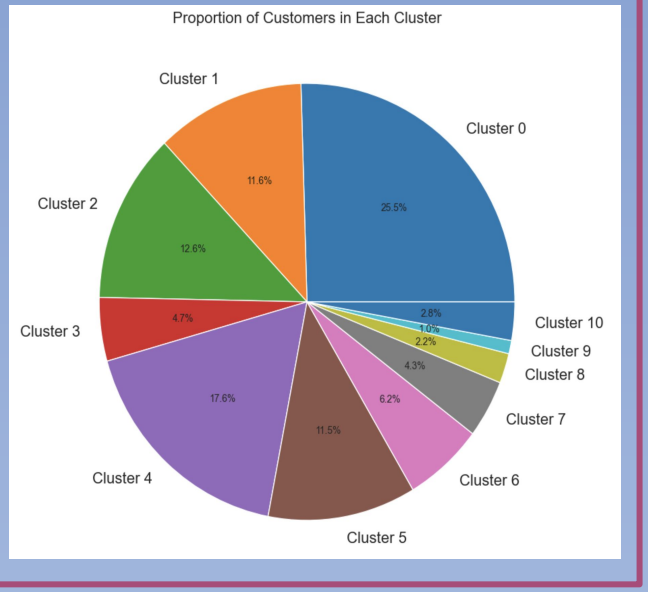




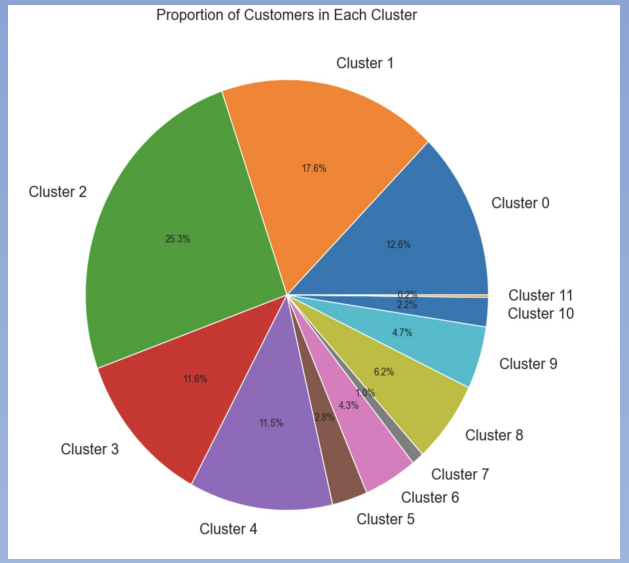

```python
#метод калински показал,что кол-во кластеров должно быть равно 13
import numpy as np
from sklearn.cluster import KMeans
from sklearn.metrics import calinski_harabasz_score

# Define the range of k values to evaluate
k_values = range(3, 14)  # Start from k=2 as it requires a minimum of tw

# Calculate the Calinski-Harabasz index for each k value
ch_scores = []
for k in k_values:
    model = KMeans(n_clusters=k, random_state=42)
    labels = model.fit_predict(features_standardized)
    ch_score = calinski_harabasz_score(features_standardized, labels)
    ch_scores.append(ch_score)

# Find the optimal k value with the highest Calinski-Harabasz index
optimal_k = k_values[np.argmax(ch_scores)]

# Print the optimal k value
print("Optimal k:", optimal_k)
```

```
Optimal k: 13
```

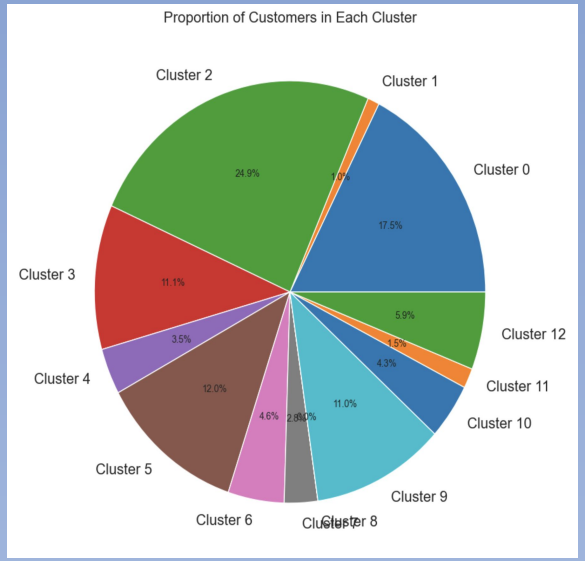
Распределение для 11,12,13 кластеров



11 кластеров



12 кластеров



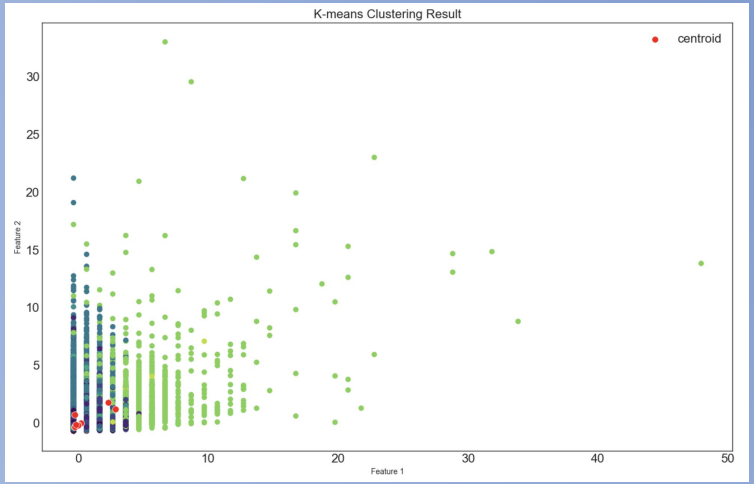
13 кластеров

Вывод: более сбалансированной по количеству клиентов является сегментация для 11 кластеров.

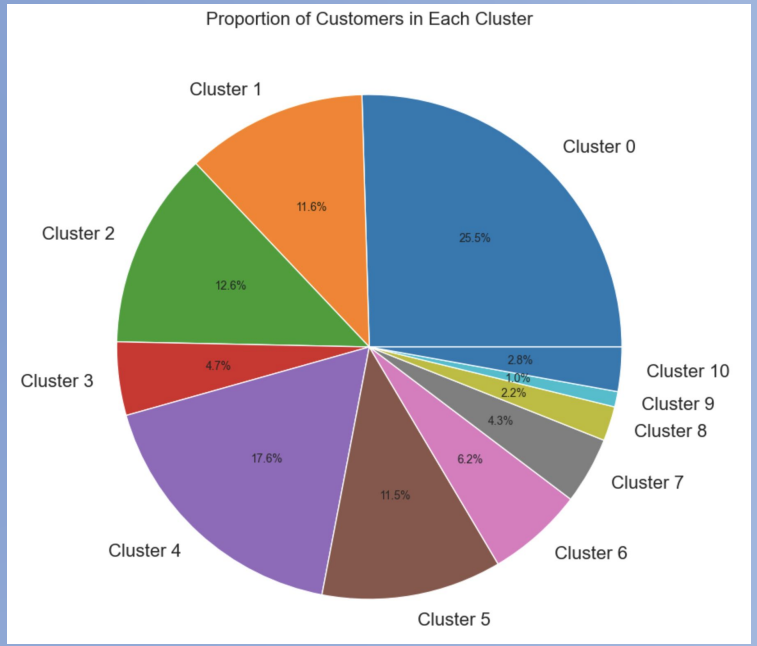
Сегментация K-means

Модель для 11 кластеров дает сбалансированное разделение сегментов.

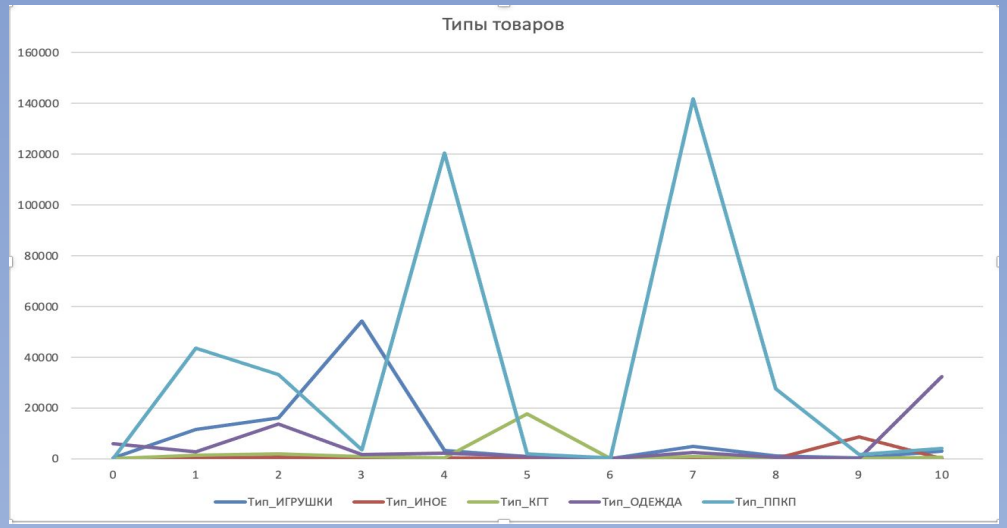
График распределения



Распределение клиентов на 11 сегментов



Анализ сегментов K-means по типу товаров



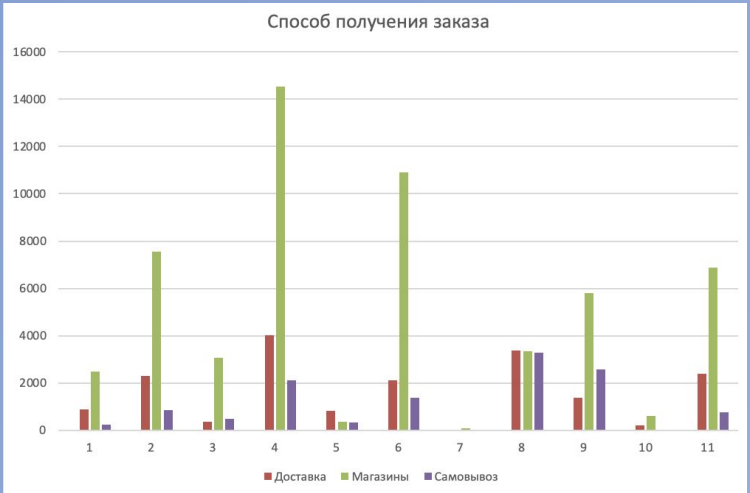
Видим, что наибольшей популярностью пользуются:

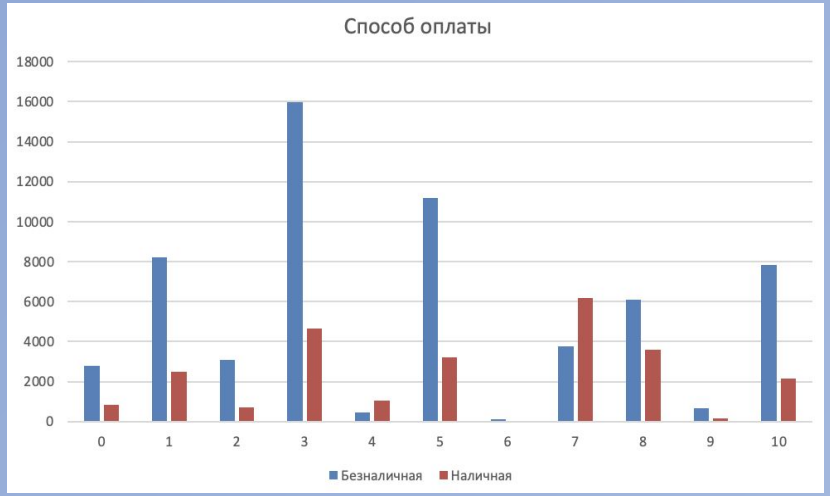
- Тип_ППКП(1,2,4,6,7,8 сегменты)
- Тип_ОДЕЖДА(0,10 сегменты)
- Тип_Игрушки(3 сегмент)

Количество товаров, покупаемых сегментами:

	0	1	2	3	4	5	6	7	8	9	10
Тип_ИГРУЛ	336	11560	16070	54208	3314	996	82	5002	1205	240	3122
Тип_ИНОЕ	8	500	592	131	242	22	0	379	80	8511	82
Тип_КГТ	91	1302	2057	993	456	17715	5	834	424	46	671
Тип_ОДЕЖ	5942	2771	13713	1800	2109	851	22	2415	615	162	32427
Тип_ППКП	170	43719	33169	3650	120418	1883	261	141760	27630	1697	4085

Анализ сегментов K-means по способу оплаты и получения



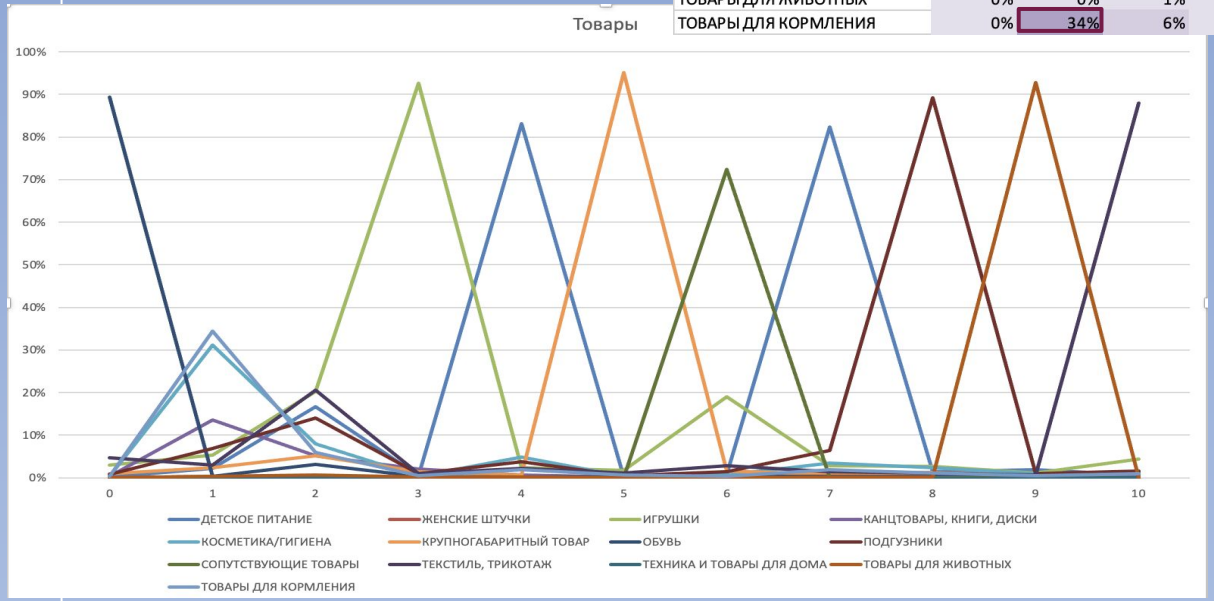


У всех сегментов преобладает безналичный расчет и получения из магазинов.

Анализ сегментов K-means по категориям

Кластеры

	0	1	2	3	4	5	6	7	8	9	10
ДЕТСКОЕ ПИТАНИЕ	0%	2%	17%	0%	83%	0%	1%	82%	1%	2%	1%
ЖЕНСКИЕ ШТУЧКИ	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
ИГРУШКИ	3%	5%	20%	93%	2%	2%	19%	3%	3%	1%	4%
КАНЦТОВАРЫ, КНИГИ, ДИСКИ	0%	14%	5%	2%	1%	0%	0%	0%	0%	0%	1%
КОСМЕТИКА/ГИГИЕНА	0%	31%	8%	0%	5%	0%	1%	3%	2%	1%	1%
КРУПНОГАБАРИТНЫЙ ТОВАР	1%	2%	5%	1%	1%	95%	2%	1%	1%	0%	1%
ОБУВЬ	89%	0%	3%	0%	0%	0%	0%	0%	0%	0%	1%
ПОДГУЗНИКИ	1%	7%	14%	1%	4%	1%	1%	6%	89%	1%	2%
СОПУТСТВУЮЩИЕ ТОВАРЫ	0%	0%	0%	0%	0%	0%	72%	0%	0%	0%	0%
ТЕКСТИЛЬ, ТРИКОТАЖ	5%	3%	20%	1%	2%	1%	3%	1%	1%	1%	88%
ТЕХНИКА И ТОВАРЫ ДЛЯ ДОМА	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
ТОВАРЫ ДЛЯ ЖИВОТНЫХ	0%	0%	1%	0%	0%	0%	0%	0%	0%	93%	0%
ТОВАРЫ ДЛЯ КОРМЛЕНИЯ	0%	34%	6%	0%	2%	1%	0%	2%	1%	0%	1%



Анализ сегментов K-means по среднему чеку и количеству товаров

Максимальный средний чек и среднее количество товаров имеют клиенты 5-ого сегмента. Они чаще всего покупают игрушки.

Выводы по сегментам

1 - 3513 клиентов
Покупают **одежду(обувь)**
Средний чек - **2218**
2 - 10289 клиентов
Покупают **товары для кормления и косметику**
Средний чек - **2476**
3 - 3079 клиентов
Покупают **текстиль и игрушки**
Средний чек - **10495**
4 - 20193 клиентов
Покупают **игрушки**
Средний чек - **2191**
5 - 1212 клиента
Покупают **детское питание**
Средний чек - **13919**

6 - 14176 клиентов
Покупают **крупногабаритные товар**
Средний чек - **7810**
7 - 142 клиента
Покупают **сопутствующие товары**
Средний чек - **2077**
8 - 8931 клиентов
Покупают **детское питание**
Средний чек - **3974**
9 - 9001 клиент
Покупают **подгузники**
Средний чек - **3032**
10 - 786 клиентов
Покупают **товары для животных**
Средний чек - **2844**
11 - 9633 клиента
Покупают **текстиль, трикотаж**
Средний чек - **3109**

Оценка качества кластеризации

1. Коэффициент силуэта находится в диапазоне от -1 до 1, где более высокое значение указывает на лучшее качество кластеризации.

Как мы можем видеть, наш коэффициент **ниже 0,2**, алгоритм DBSCAN применять **не стоит**.

```
The Silhouette Coefficient is: -0.15405393758625138
```

Для K-means коэффициент силуэта 0,36 **считается умеренным**.

```
The Silhouette Coefficient is: 0.36047039094404254
```

2.Индекс Боулдина — это мера качества кластеризации, где более низкое значение указывает на лучшую кластеризацию.
Получившийся индекс можно считать хорошим.

```
The Bouldin Index is: 0.2480105130618571
```

3.Индекс Калински-Харабаша — это мера качества кластеризации, где более высокий индекс указывает на лучшие результаты кластеризации.
Наш индекс можно считать хорошим.

```
The Calinski-Harabasz Index is: 6787.6773735903525
```

Связь между RFM и K-means

- **1 кластер давно(111)** или **недавно(211)** сделал заказ **на маленькую сумму**, чаще всего покупали товары из категории **“Обувь”**
- **2 кластер давно(111)** или **недавно(211)** сделали заказ **на маленькую сумму,** чаще всего покупали товары из категорий **“Товары для кормления” и “Косметика, гигиена”**
- **3 кластер давно** делали заказ **на большую сумму,** чаще всего покупали **“Текстиль, трикотаж” и “Игрушки”**
- **4 кластер давно(111) или недавно(211)** сделали заказ **на маленькую сумму,** чаще всего покупали товары из категории **“Игрушки”**
- **5 кластер давно** делали заказ **на большую сумму,** преобладают товары из категории **“Детское питание”**

Кластеры K-means

Сегменты RFM	0	1	2	3	4	5	6	7	8	9	10
111	72%	65%	1%	72%	4%	38%	73%	52%	60%	66%	64%
112	6%	9%	2%	6%	27%	38%	8%	15%	9%	6%	11%
121	1%	1%	23%	1%	6%	0%	1%	5%	3%	3%	1%
122	1%	1%	70%	0%	54%	1%	0%	7%	3%	5%	1%
211	20%	22%	0%	20%	2%	12%	17%	17%	22%	18%	20%
212	1%	2%	0%	2%	6%	12%	1%	4%	3%	3%	3%
221	0%	0%	1%	0%	0%	0%	0%	1%	0%	0%	0%
222	0%	0%	3%	0%	1%	0%	0%	0%	0%	0%	0%

Связь между RFM и K-means

- **6 кластер давно** делал заказ **на большую и маленькую суммы** в категории **“Крупногабаритный товар”**
- **7 кластер давно** делали заказ **на маленькую сумму** преимущественно из категории **“Сопутствующие товары”**
- **8 кластер давно(111) или недавно(211)** делали заказ **на маленькую сумму** из категории **“Детское питание”**
- **9 кластер давно(111) или недавно(211)** делали заказ **на маленькую сумму из категории “Подгузники”**
- **10 кластер давно(111) или недавно(211)** делали заказ **на маленькую сумму** из категории **“Товары для животных”**
- **11 кластер давно(111) или недавно(211)** делали заказ **на маленькую сумму** из категории **“Текстиль, трикотаж”**

Кластеры K-means

Сегменты RFM	0	1	2	3	4	5	6	7	8	9	10
111	72%	65%	1%	72%	4%	38%	73%	52%	60%	66%	64%
112	6%	9%	2%	6%	27%	38%	8%	15%	9%	6%	11%
121	1%	1%	23%	1%	6%	0%	1%	5%	3%	3%	1%
122	1%	1%	70%	0%	54%	1%	0%	7%	3%	5%	1%
211	20%	22%	0%	20%	2%	12%	17%	17%	22%	18%	20%
212	1%	2%	0%	2%	6%	12%	1%	4%	3%	3%	3%
221	0%	0%	1%	0%	0%	0%	0%	1%	0%	0%	0%
222	0%	0%	3%	0%	1%	0%	0%	0%	0%	0%	0%

Итог

- **Кластерам 1, 2, 4, 7, 8, 9, 10,** которые преимущественно относятся **к 111 и 211**, необходимо напомнить о компании, предложить акцию: “При покупке от 3-ех товаров на сумму более 5000 рублей, получите в подарок скидку 10% на последующий заказ”.
- **Кластерам 3 и 5,** которые преимущественно относятся **к 122**, стоит напомнить о компании, предложить скидку на покупку товаров из категорий “Текстиль, трикотаж” и “Детское питание”.
- **Кластерам 6,** который в равной мере относится **к 111 и 112,** стоит напомнить о компании, предложить скидку 5% на товары из категории “Крупногабаритный товар” при покупке на сумму от 50 тыс. рублей.

Спасибо за внимание!

24 июня 2023 года