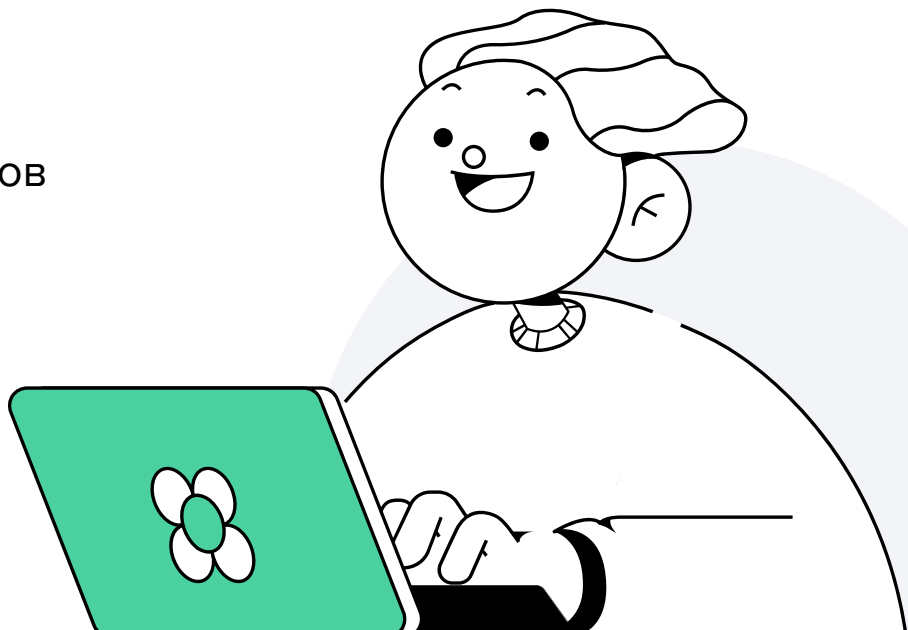


# Презентация дипломного проекта



# Структура презентации

- 1 Описание компании
- 2 Сводка по конкурентам
- 3 Формулирование проблемы и описание данных
- 4 Предобработка данных
- 5 Анализ данных и сегментация клиентов
- 6 Дизайн A/B теста
- 7 Вывод



# Описание компании



1

# Описание компании

Интернет-магазин TENNIS STORE основан в 2017 году и занимает лидирующие позиции в продаже товаров для тенниса.

Головной офис базируется в Калининграде. Близость к Европе помогает поддерживать широкий ассортимент и предлагать конкурентные цены на товар.

Компания сотрудничает с федерацией тенниса калининградской области и является спонсором Турниров в Калининграде, Пскове, Москве и Санкт-Петербурге. Собственная команда участвует в турнирах по теннису разных уровней и категорий.

Маркетинговое УТП: В наличии более 12'000 теннисных товаров из Европы от ведущих мировых брендов: Nike, Wilson, Adidas, Babolat и т.д. Оригинальные ракетки, кроссовки, одежда, мячи с доставкой по всей России



# Сводка по конкурентам



2

# Прямые конкуренты

## SALETENNIS

УТП: Наш магазин предлагает Вам большой выбор товаров, как для новичков, так и для спортсменов – профессионалов.

## TENNIS PRO

УТП: Добро пожаловать в Tennis-Pro – идеальный выбор для всех ваших теннисных пожеланий!

## TENNIS DIRECT

УТП: Интернет-магазин TennisDirect приглашает купить теннисные товары - ракетки, мячи, струны, сумки, кроссовки, а также тренажеры и инвентарь для большого тенниса.

## RUS TENNIS

УТП: У нас Вы найдете большой выбор теннисных ракеток – для детей, юниоров и взрослых, для любителей и профессионалов; теннисные мячи Dunlop и Head, которые отличает превосходные игровые свойства и длительный срок службы

## СПОРТМАСТЕР

УТП: В каталоге интернет-магазина Спортмастер представлен широкий ассортимент женской и мужской экипировки для большого тенниса, а также аксессуаров и спортивного оборудования: мячей и ракеток. Оформите заказ на сайте с доставкой по России и получите бонусы на следующую покупку.

## МАРКЕТПЛЕЙСЫ

Требует дополнительного исследования.



# Посещаемость сайтов

Магазин	Посещаемость	Уникальные посетители	Доля уникальных посетителей	Среднее время сессии	Среднее количество страниц
tennis-store	30304	18100	60%	0:01:39	2,72
tennis-pro	11507	9702	84%	0:12:30	5,28
saletennis	13707	9053	66%	0:02:29	3,64
tennisdirect	3522	3456	98%	0:00:49	9,19
rus-tennis	1017	988	97%	0:02:39	1,01
sportmaster/tennis	31640	24604	78%	0:03:59	4,68
sportmaster	14.7M	7.1M	49%	0:06:01	4,7



# Прочее

Магазин	Email Marketing	Мобильное приложение	Маркетплейсы	Бонусная система
tennis-store	Да	Нет	Нет	Нет
tennis-pro	Да	App Store	Нет	Да
saletennis	Да	Нет	Нет	Нет
tennisdirect	Да	Нет	Нет	Нет
rus-tennis	Да	Нет	Wildberries + Ozon + Я. Маркет	Нет
sportmaster	Да	App Store + Google Play	Wildberries + Ozon + Я. Маркет	Да

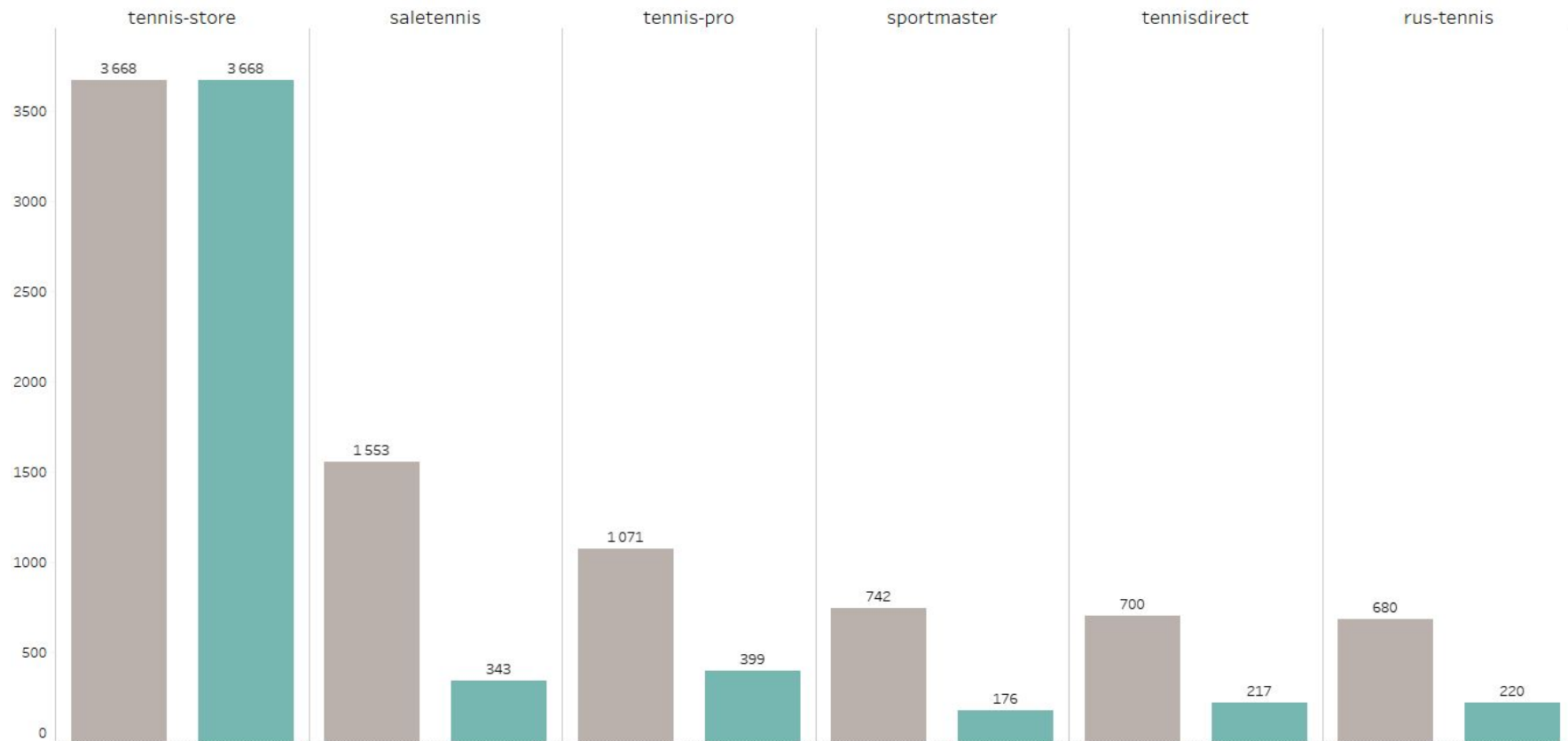




# Сравнение товарной сетки

Название магазина

■ Всего товаров  
■ Аналогичный товар



# Формулирование проблемы и описание данных



3

# Описание проблемы

В настоящий момент эффективность продаж снижена из-за отсутствия сегментации клиентов.

Сегментация важна для компании, т.к. это позволит разрабатывать различные маркетинговые стратегии, правила продаж и взаимодействия с клиентами.

Ключевые цели анализа — выявление кластеров клиентов и описание их поведения.

Для решения задачи будет проанализирована история продаж за два года.



# Описание данных

Файл продаж содержит следующую информацию:

- ID заказа
- Артикул
- Название
- Количество товара
- Дата заказа (с 2023-12-01 по 2024-12-09)
- Цена одной единицы товара
- ID клиента
- Регион доставки

ID заказа	Артикул	Название	Количество	Дата	Цена	ID клиента	Регион
563718	48194	Теннисная ракетка Head Boom MP Alternate 2024 ...	4	2024-08-18 16:18:00	38720	15144.0	Москва и МО
580538	22906	Рюкзак теннисный Under Armour Hustle Sport Bac...	8	2024-12-05 08:38:00	8680	14075.0	Москва и МО
536971	22969	Теннисная ракетка Tecnifibre TF40 315 18x20 + ...	4	2023-12-03 13:40:00	44440	12779.0	Уфа
578808	72351A	Теннисные носки Under Armour HeatGear Tech No ...	2	2024-11-25 13:23:00	2810	16327.0	Москва и МО
580640	22754	Рюкзак теннисный Under Armour Hustle Lite Back...	1	2024-12-05 12:51:00	7100	16910.0	Москва и МО
566562	20996	Теннисные струны Tecnifibre XR3 (12 m)	4	2024-09-13 12:18:00	3570	17365.0	Москва и МО
572233	21086	Теннисная ракетка Prince by Hydrogen Random 280g	8	2024-10-21 14:41:00	39540	14646.0	Екатеринбург



# Предобработка данных



4

# Обработка null значений

Количество строк: 1145002

0	ID заказа	974022	non-null	object
1	Артикул	974022	non-null	object
2	Название	974022	non-null	object
3	Количество	974022	non-null	int64
4	Дата	974022	non-null	object
5	Цена	974022	non-null	int64
6	ID клиента	893078	non-null	float64
7	Регион	974022	non-null	object

Как видно, только столбец 'ID клиента' имеет пропущенные значения. Поскольку нашей задачей является сегментация клиентов и анализ поведения, нам обязательно нужен идентификатор клиента.

Поэтому удалим строки с пропущенными значениями.



# Обработка дублей

Дублями будут считаться строки с идентичными данными в каждом столбце.

Количество дублирующихся строк: 5718.

Дубликатов почти нету. Также удаляем эти строки.



# Обработка ошибок

- 1) 'ID заказа' имеет тип данных object, что может говорить о наличии иных символов, кроме цифр.

C536506	22960	Теннисная ракетка Tecnifibre T-Fight 280 Isofl...	-2	2023-12-01 12:38:00	39540	17897.0	Москва и МО
C536543	22632	Теннисная ракетка Wilson Pro Staff Precision 103	0	2023-12-01 14:30:00	15160	17841.0	Москва и МО
C536543	22355	Теннисные струны Luxilon Natural Gut (12,2 m)	0	2023-12-01 14:30:00	10100	17841.0	Москва и МО
C536548	22244	Сумка теннисная Yonex Pro Racquet Bag 12 pack ...	-1	2023-12-01 14:33:00	34640	12472.0	Казань

В ходе анализа установлено, что значения в 'ID заказа' могут иметь букву 'С' в начале, что говорит об отмене заказа. Отмененные заказы имеют отрицательные или нулевые значения в столбце 'Количество' и не имеют дубликатов в таблице без буквы 'С' в 'ID заказа' и с положительным количеством.

Данные строки не будут использоваться в рамках этого исследования.





# Обработка ошибок

2) 'Артикул' - данные корректны.

3) 'Название' - данные корректны.

4) 'Количество' - никаких отклонений, кроме крайне высоких значений (будет рассмотрено далее).

5) 'Дата' - данные корректны, но необходимо изменить тип данных object.

6) 'Цена' - найдено 510 строк с нулевой ценой. Удаляем эти строки.

7) 'ID клиента' - данные корректны. Выполнено преобразование в целочисленный тип данных, т.к. нет необходимости в типе float.

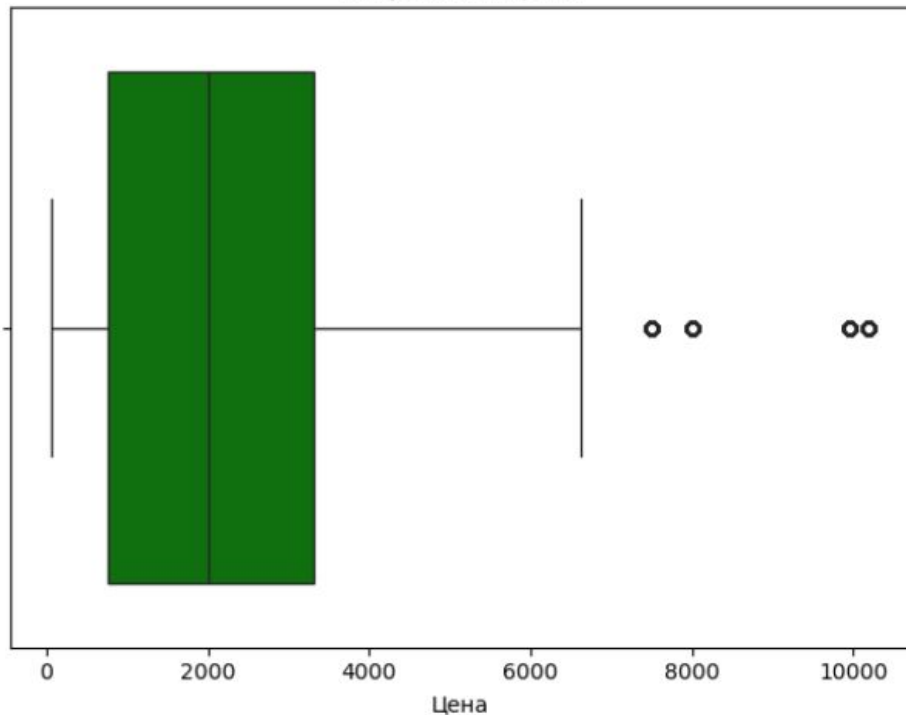
8) 'Регион' - данные корректны



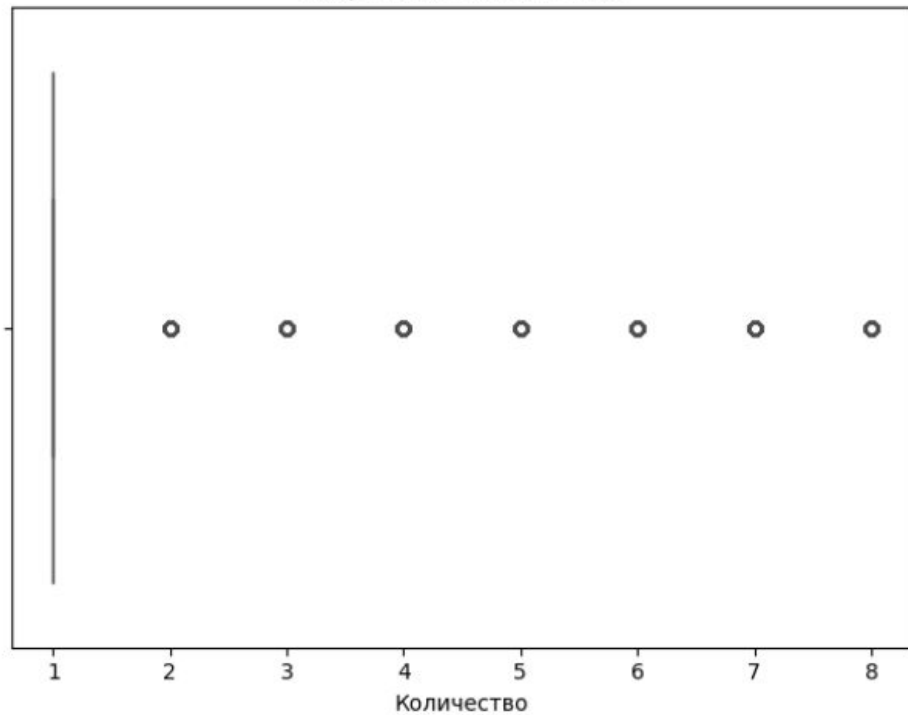
# Обработка выбросов

В столбцах “Цена” и “Количество” могут содержаться выбросы. Посмотрим на графики.

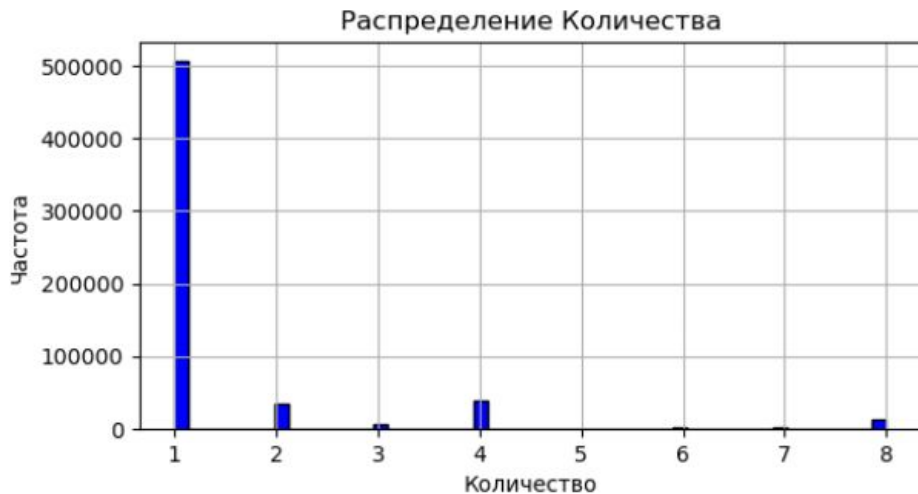
Boxplot для Цены



Boxplot для Количества



# Обработка выбросов



Как видно на графиках, аномальных значений нет.

После обработки данных количество строк сократилось на 38%. Осталось 603894 строки.



# Анализ данных

## Сегментация клиентов



5

# Анализ данных

Для построения кластеров, основываясь на тех данных, что у нас есть, создадим дополнительные столбцы:

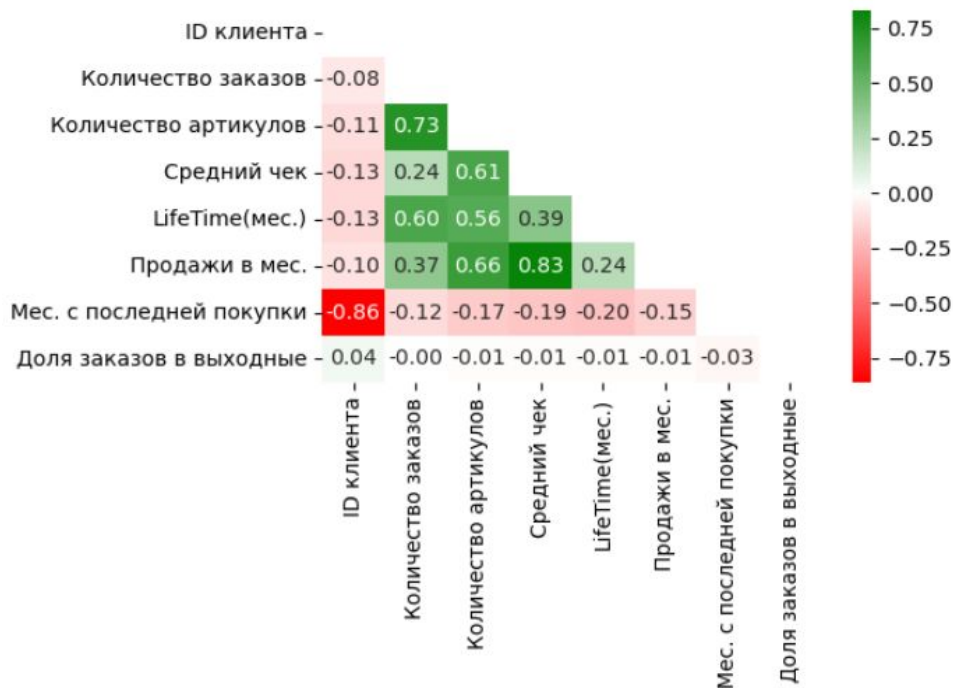
- Сумма заказов (руб.)
- Количество заказов
- Количество артикулов
- Всего товаров
- Средний чек
- Среднее количество товара в заказе
- LifeTime(мес.)
- Месяцев с последней покупки
- Доля заказов в выходные

	ID клиента	Количество заказов	Количество артикулов	Сумма заказов (руб.)	Всего товаров	Средний чек	Среднее количество товара в заказе	LifeTime(мес.)	Мес. с последней покупки	Продажи в мес.
0	12347.0	7	48	683955	288	97707.86	41.14	13	1	52611.92
1	12348.0	3	1	6315	3	2105.00	1.00	9	4	701.67
2	12349.0	1	59	312099	153	312099.00	153.00	1	2	312099.00
3	12350.0	1	1	2105	1	2105.00	1.00	1	11	2105.00
4	12352.0	7	38	243764	101	34823.43	14.43	9	2	27084.89



# Анализ данных

Как видно на тепловой карте, никаких неожиданных взаимосвязей не наблюдается.



Матрица корреляции (метод Пирсона)



# Сегментация клиентов

Для решения этой задачи я воспользуюсь одним из популярных и простых алгоритмов машинного обучения без учителя “K-means” кластеризацией.

K-means группирует данные в заранее заданное количество кластеров ( $k$ ), минимизируя сумму квадратов расстояний между точками данных и центроидами их кластеров. На каждом шаге алгоритм назначает точки ближайшим центроидам и пересчитывает центроиды как среднее значение точек в кластере, повторяя процесс до стабилизации.

Поскольку K-means использует расстояние для измерения близости точек, необходимо привести данные к одному масштабу, т.к. Если признаки имеют разные масштабы (например, возраст в диапазоне 0–100 и доход в диапазоне 0–1 000 000), то признаки с большими значениями могут доминировать при вычислении расстояний.



# Сегментация клиентов

Поскольку K-means использует расстояние для измерения близости точек, необходимо привести данные к одному масштабу, т.к. Если признаки имеют разные масштабы (например, возраст в диапазоне 0–100 и доход в диапазоне 0–1 000 000), то признаки с большими значениями могут доминировать при вычислении расстояний.

	ID клиента	Количество заказов	Количество артикулов	Средний чек	LifeTime(мес.)	Продажи в мес.	Сумма заказов (руб.)	Мес. с последней покупки	Доля заказов в выходные
0	12347.0	8.853291	6.555754	8.084328	21.370968	3.455710	9.790136	-3.750819	0.070594
1	12348.0	2.915823	-0.066924	-0.088285	14.219255	-0.167148	0.016410	-2.427007	-0.809112
2	12349.0	-0.052911	8.105743	26.411559	-0.084170	21.565519	4.426789	-3.309548	-0.809112
3	12350.0	-0.052911	-0.066924	-0.088285	-0.084170	-0.069208	-0.044311	0.661888	-0.809112
4	12352.0	8.853291	5.146674	2.708651	14.219255	1.674158	3.441181	-3.309548	0.070594





# Сегментация клиентов

Также необходимо определить оптимальное число кластеров. Для этого воспользуюсь методом “Локтя”. Он основан на анализе метрики искажения (distortion) или суммы квадратов ошибок (SSE, Sum of Squared Errors). Эта метрика измеряет, насколько далеко точки данных находятся от центроидов их кластеров.

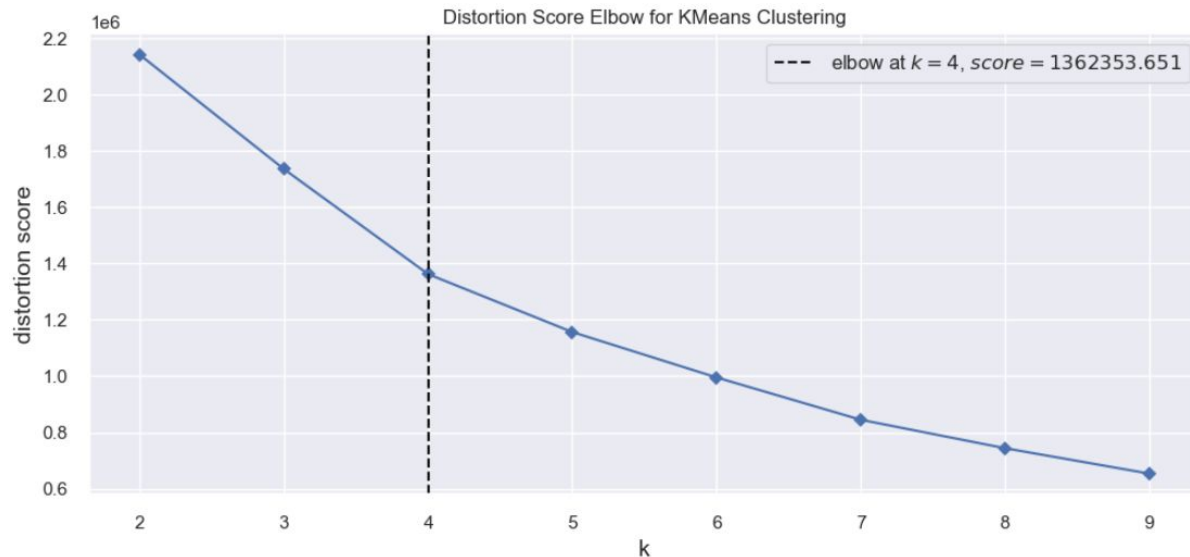


График показывает, что 4 - это оптимальное число кластеров. Разделим базу клиентов на 4 сегмента и опишем их особенности.



# Сегментация клиентов

Теперь необходимо оценить качество кластеров.

**Silhouette Score** (Силуэтный коэффициент, от -1 до 1) измеряет компактность кластеров и их разделенность (насколько объекты внутри одного кластера ближе друг к другу, чем к другим кластерам).

В нашем случае **0.509** — неплохой показатель, но внутри кластеров есть некоторая размытость.

**Calinski-Harabasz Score** (Индекс Калински-Харабаша) показывает, насколько кластеры компактны и разделены.

**149681** — высокий показатель, что говорит о хорошей компактности и разделении кластеров.



# Сегментация клиентов

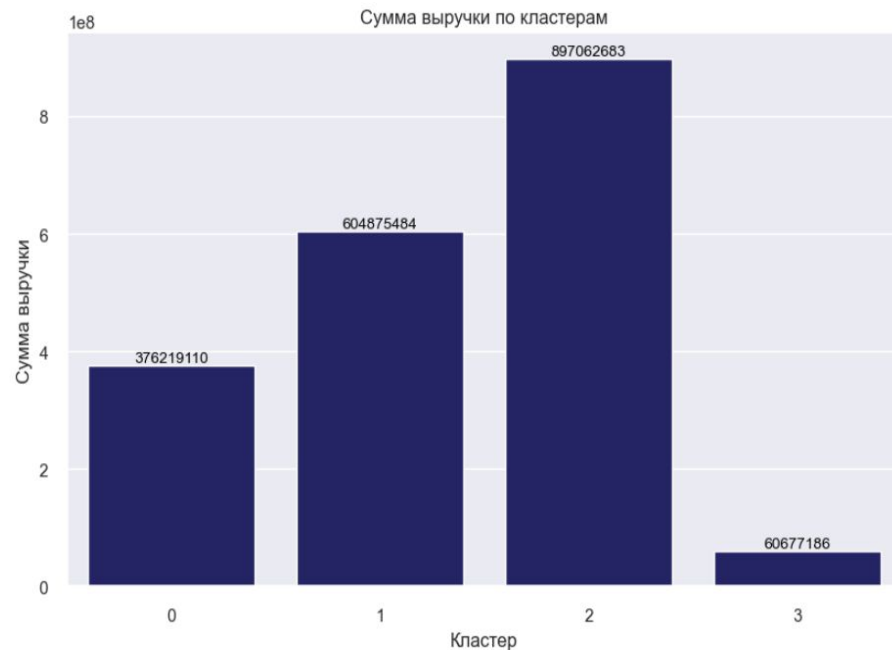
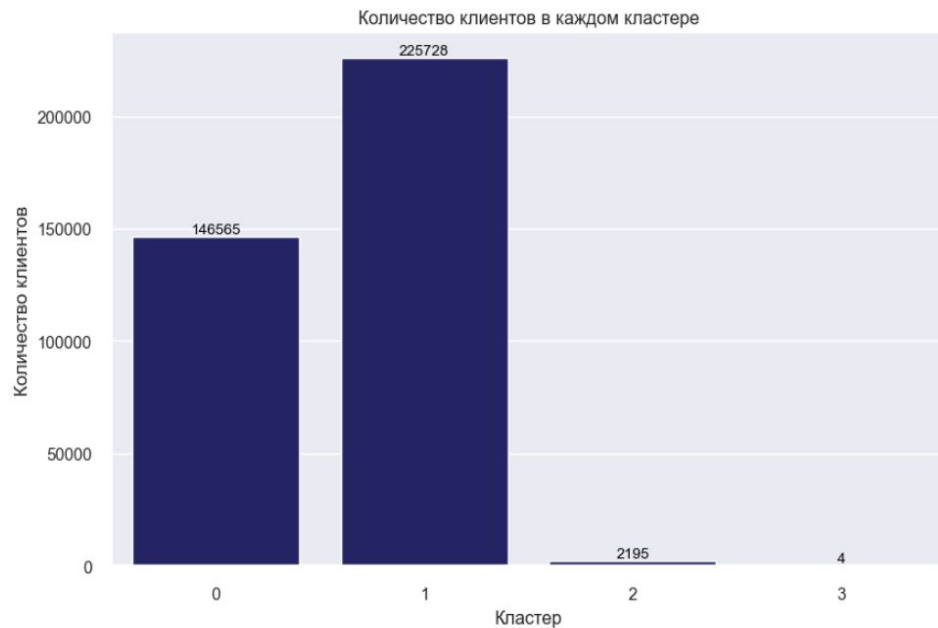
Davies-Bouldin Score (Индекс Дэвиса-Боулдина, от 0 до  $\infty$ ) измеряет насколько кластеры различимы, учитывая их разброс и расстояния между ними.

0.763 — хороший показатель, указывающий на приемлемое качество кластеризации, но кластеры частично пересекаются или имеют размытые границы.

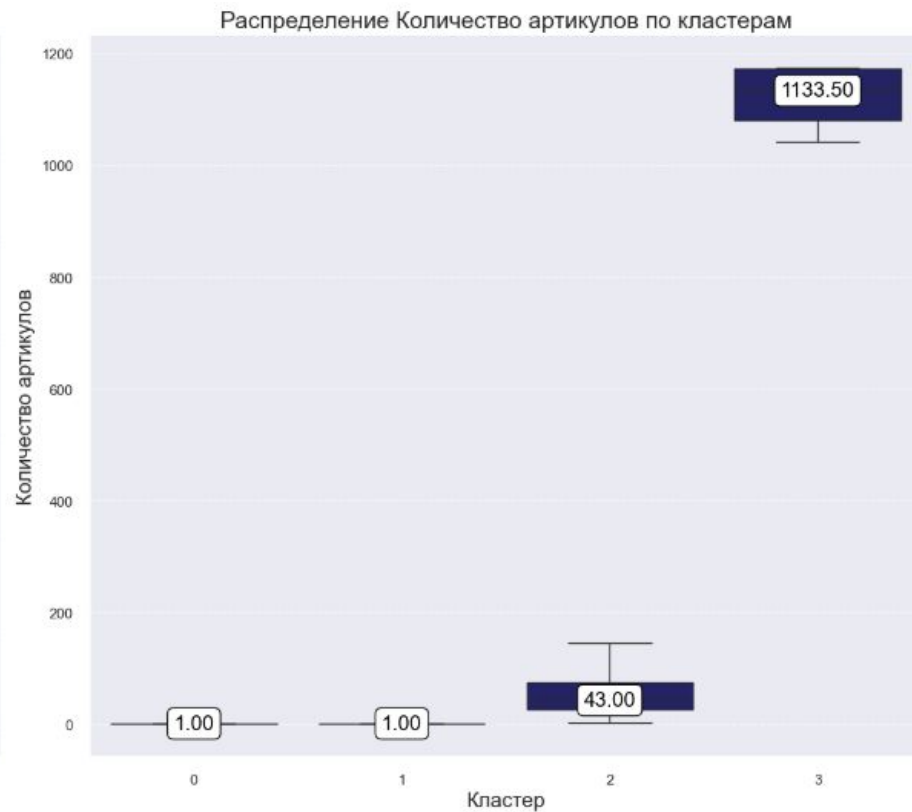
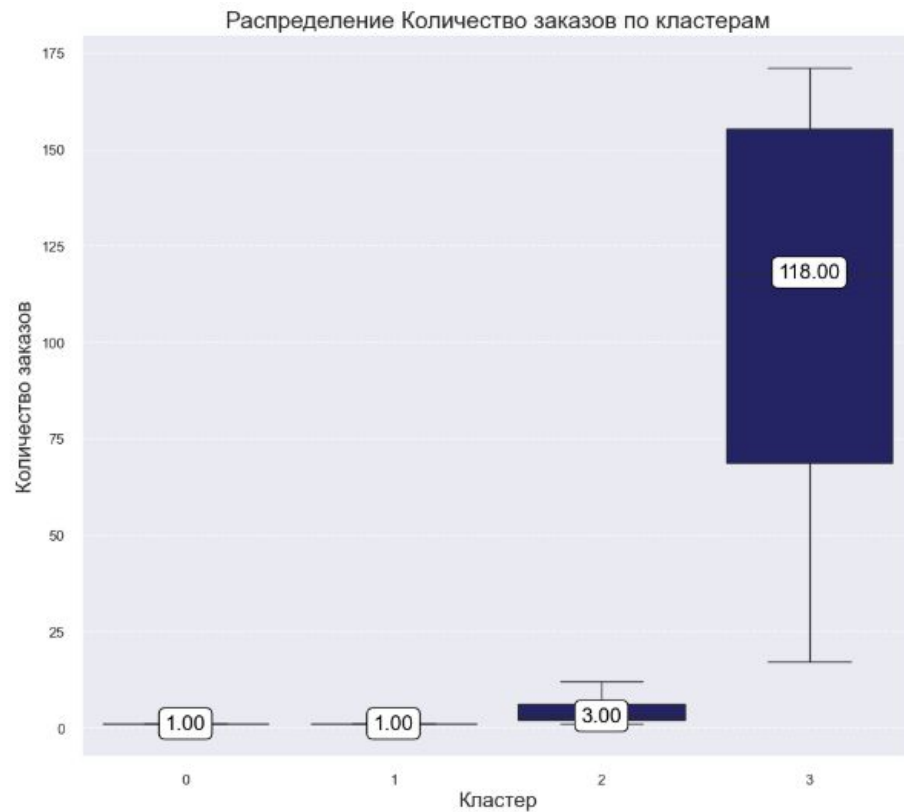


# Описание кластеров

Рассмотрим характеристики каждого кластера.

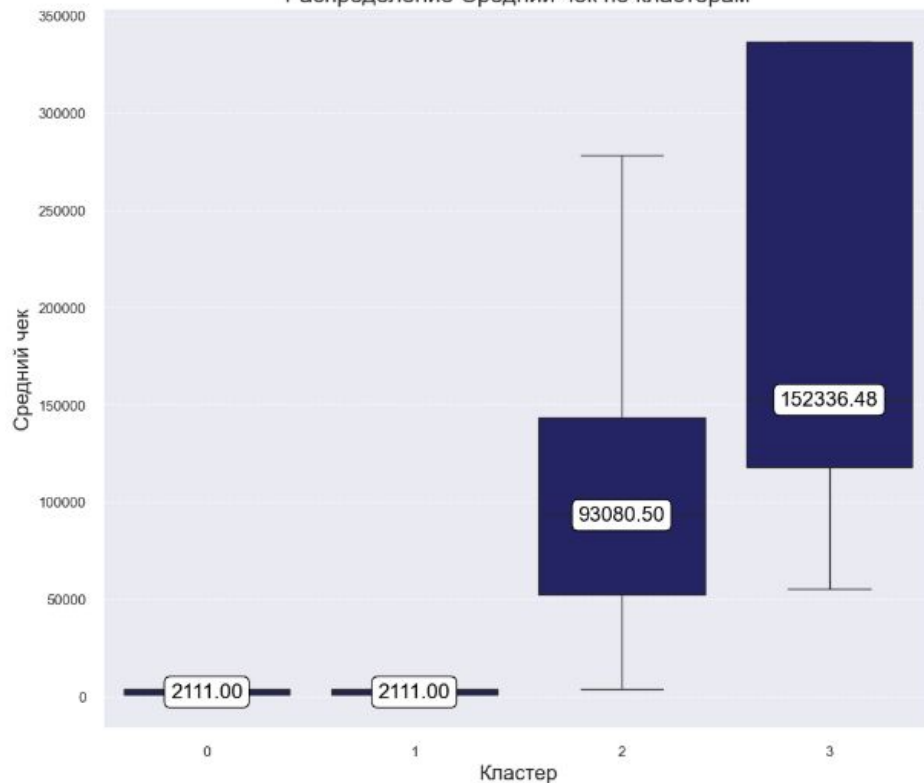


# Описание кластеров

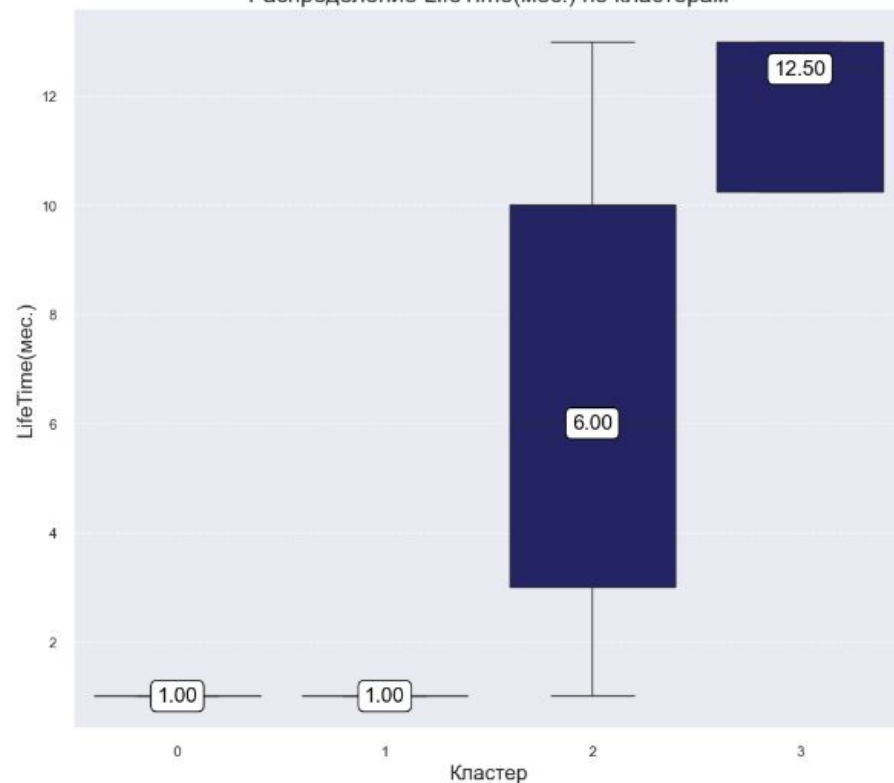


# Описание кластеров

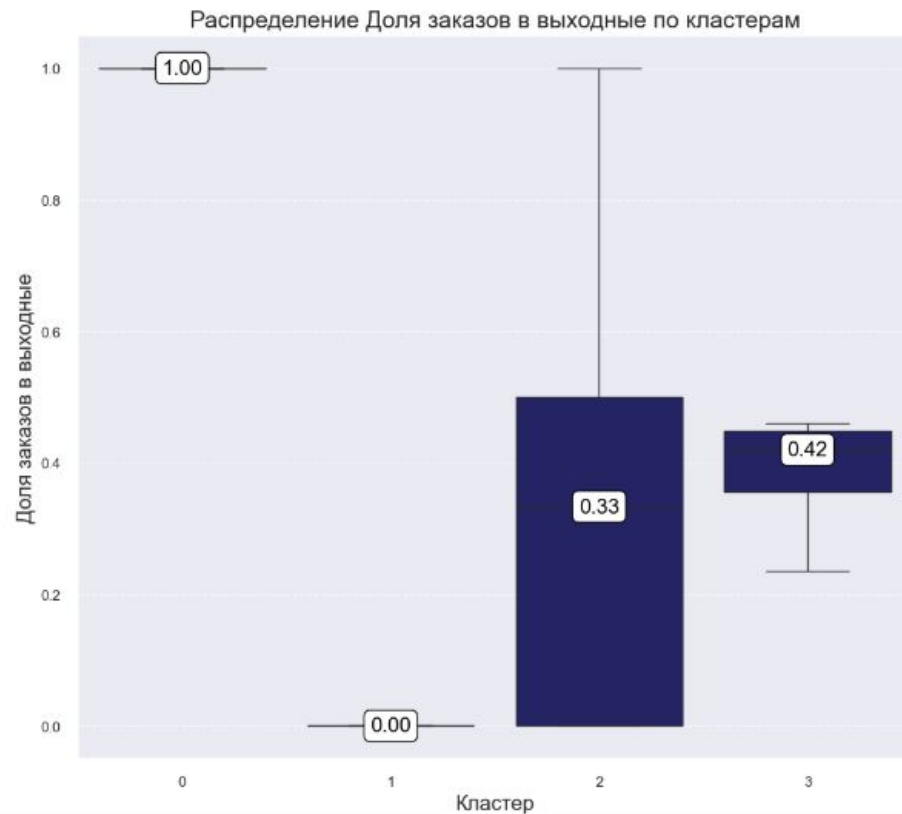
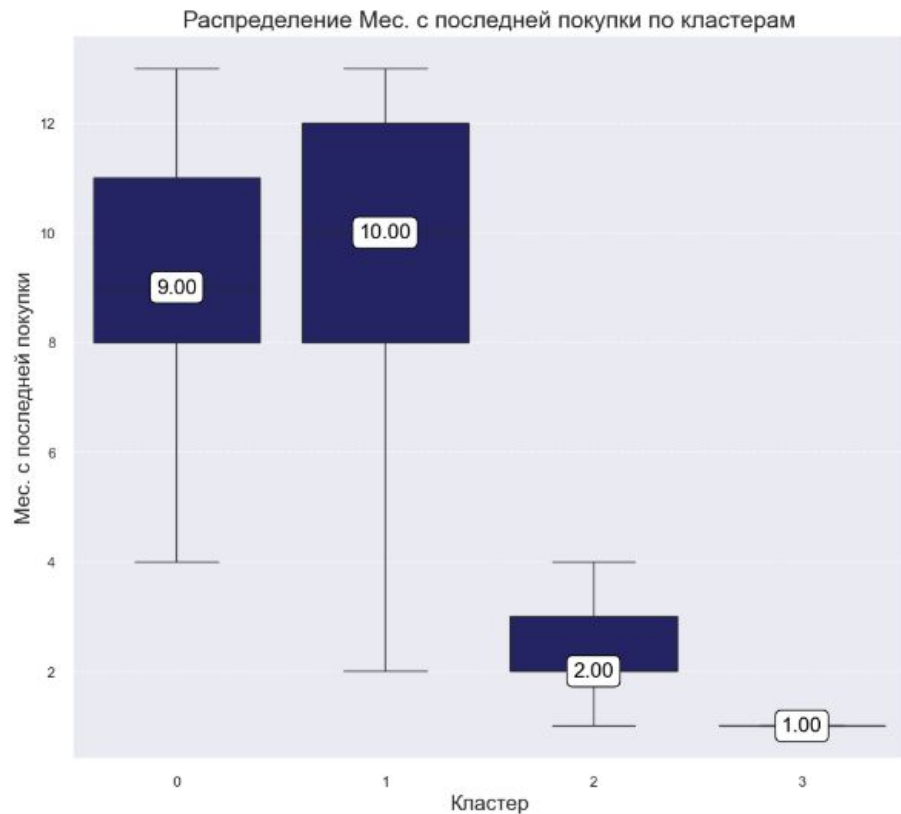
Распределение Средний чек по кластерам



Распределение LifeTime(мес.) по кластерам



# Описание кластеров



# Описание кластеров

## Кластер 0

Разовые покупатели. 39.14% клиентов относится к этому кластеру, вклад в выручку 19.40%. Это клиенты, которые совершают всего один заказ с небольшим средним чеком (2111 руб). Их Life Time составляет 1 месяц - это минимальное значение, по факту это разовая покупка. Давность покупки от 4 до 13 месяцев, медиана 9 месяцев. Совершают покупки только в выходные дни.

Это наиболее массовый, но наименее лояльный сегмент. Клиенты делают одну покупку и не возвращаются. Требуется работа по повышению удержания.





# Описание кластеров

## Кластер 1

Этот кластер копирует характеристики класса 0, за исключением того, что все заказы сделаны в выходные дни.

60.28% клиентской базы, 31.20% от выручки.

В дальнейшей работе необходимо объединить с кластером 0.



# Описание кластеров

## Кластер 2

Клиенты с высокой покупательной способностью, активные и малочисленные.

Хотя этот сегмент крайне мал (0.59% клиентской базы), приносит 46.27% всей выручки. Клиенты совершают в среднем по 3 покупки, но с очень высоким средним чеком (93 080 руб) и большим количеством артикулов (43). Их Life Time 1-13 месяцев, медиана — 6 месяцев, а время с последней покупки составляет от 1 до 7 месяцев (медиана — 2 месяца). Заказы чаще совершаются в будние дни.

Это лояльные клиенты с высокой ценностью. Их необходимо удерживать, например с помощью программ лояльности и персональных предложений.



# Описание кластеров

## Кластер 3

Максимально лояльные и высокодоходные клиенты. По всем признакам b2b сегмент. Кластер крайне мал ( $>0.1\%$ ) и приносит 3.13% всей выручки. В среднем клиенты совершают большое количество заказов (118) с высоким средним чеком (152 336 руб). Также у них самый высокий Life Time (от 10 до 13 месяцев, медиана — 12,5 месяца). Их интересует большое количество разнообразных товаров (в среднем 1133 артикулов). Время с последней покупки — всего 1 месяц. Около половины заказов совершаются в будние дни.

Это наиболее ценный сегмент с высокой частотой заказов и большим вкладом в выручку на 1 покупателя. Их необходимо поддерживать индивидуальным подходом, например, персональными менеджерами, специальной ценовой программой, эксклюзивными предложениями и др. Однако этот сегмент очень мал. Возможно из-за заточенности бизнеса на работу с b2c, а не b2b.



# Дизайн А/Б теста



6

# Гипотеза

Как видно из результатов сравнения кластеров, у компании значительная доля клиентов совершает только одну покупку и больше не возвращается (кластеры 0 и 1). Если удастся вернуть часть таких клиентов, это может существенно увеличить выручку. Чтобы их поведение стало больше похоже на поведение клиентов из кластера 2.

**Гипотеза:** Если предложить клиентам из кластера "0" скидку 10% на второй заказ, то в течение 2-х недель доля повторных покупок среди них увеличится на 8%, так как это привлечет их внимание и скидка снизит барьер для повторной покупки.



# План

## 1. Определение размера выборки.

Расчет по формуле:

$$n = \frac{(Z_{\alpha/2} \cdot \sqrt{2p(1-p)} + Z_{\beta} \cdot \sqrt{p_1(1-p_1) + p_2(1-p_2)})^2}{(p_2 - p_1)^2}$$

где:

- $n$  — минимальный размер выборки на одну группу (контрольную или тестовую).
- $p_1$  — текущая конверсия (1.3%).
- $p_2$  — ожидаемая конверсия после изменения (15%).
- $p = \frac{p_1 + p_2}{2}$  — средняя конверсия.
- $Z_{\alpha/2}$  — критическое значение нормального распределения для уровня значимости  $\alpha$  (1.96 для 5%).
- $Z_{\beta}$  — критическое значение нормального распределения для статистической мощности  $1 - \beta$  (обычно 0.84 для 80% мощности).

При:

- уровень значимости  $\alpha$  - 5%
- Статистическая мощность  $(1-\beta)$  — 80%
- Размер выборки = 124

Для повышение точности увеличим размер выборки до 20000, т.к. затраты на эксперимент минимальны.



# План

## 2. Разделение на группы.

Группа А (контрольная группа): Клиенты, которые не получают никакого специального предложения.

Группа В (тестовая группа): Клиенты, которые получают предложение скидки.

Распределение по группам будет случайным образом поровну. Но после, необходимо проверить группы на однородность по среднему чеку и давности покупки с помощью t-теста или тест Манна-Уитни.

## 3. Формат предложения.

Отправляем клиентам из группы В персонализированное предложение по e-mail, с предложением скидки 10% на второй заказ с ограничением по времени в 2 недели.



# План

## 4. Метрики для оценки.

Первичная метрика: Доля клиентов, совершивших повторную покупку.

Вторичные метрики:

- средний чек повторных покупок
- общая выручка от повторных покупок
- время между первой и второй покупкой

## 5. Продолжительность эксперимента.

Эксперимент длится 2 недели.





# План

## 6. Анализ результатов.

Если разница в доле повторных покупок между группами составит 8% или более, гипотеза подтверждается. Если нет — гипотеза опровергается.

Далее проверим статистическую значимость различий с помощью статистических тестов. Подойдет z-тест для пропорций, т.к. данные бинарны (купил/не купил).

Если  $|Z| > 1.96$ , различия статистически значимы на уровне  $\alpha = 0.05$



# Вывод

7

# Вывод

Компания имеет хороший потенциал для развития продаж за счет дифференцированного подхода к клиентам.

В ходе разделения клиентской базы на кластеры была выявлена одна из проблем: 99% клиентов совершают всего одну покупку и не возвращаются (низкий Retention). Несмотря на их огромную численность, они формируют лишь 51% выручки.

Спланированный А/Б тест – лишь первый шаг к глубокой персонализации работы с клиентами.

В дальнейшем можно:

- провести серию CustDev для большего понимания потребностей сегментов
- провести анализ потребительских корзин для системы рекомендаций покупок
- улучшить показатели рекламы и многое другое.



**Спасибо за внимание!**



