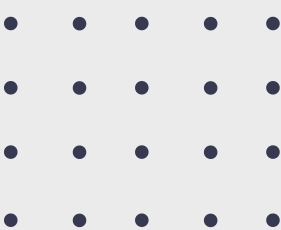


Презентация дипломного проекта

Афонин Артем Викторович

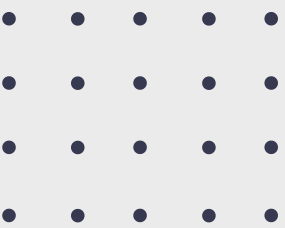
Продуктовая Аналитика





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 А/Б теста
7. Вывод





Описание компании

1

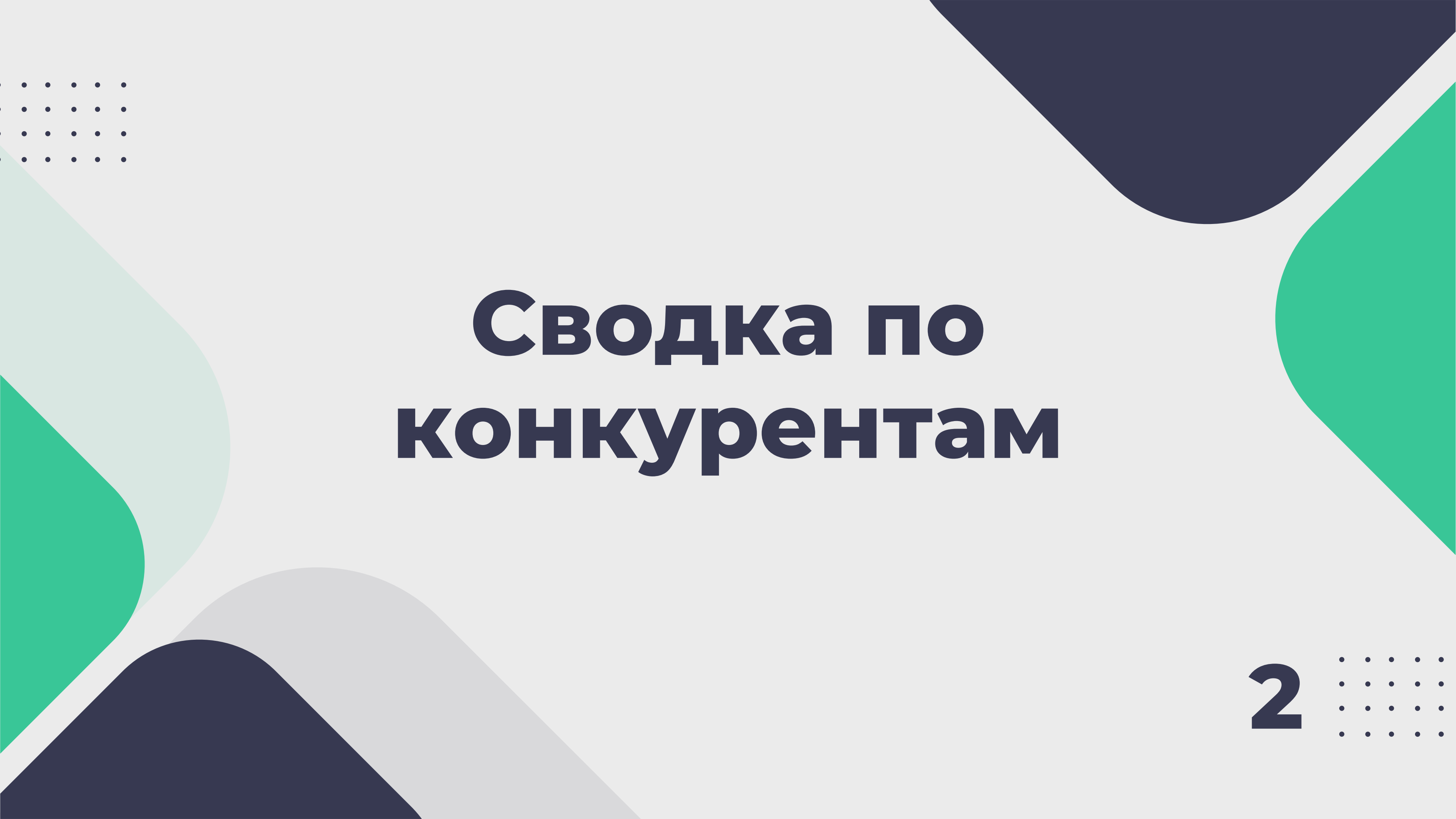




TENNIS STORE — интернет-магазин теннисных товаров, основанный в 2017 году. Занимает лидирующие позиции на российском рынке.

Головной офис расположен в Калининграде, что обеспечивает выгодную логистику из Европы и конкурентные цены. Ассортимент — более 12 000 оригинальных товаров от Nike, Wilson, Babolat, Adidas и других брендов.

Компания не только продает товары, но и сотрудничает с Федерацией тенниса Калининградской области, спонсирует турниры в четырёх городах и продвигает спорт через участие собственной команды.

The background features several large, overlapping abstract shapes in dark blue, light green, and light grey. In the top-left and bottom-right corners, there are decorative patterns of small dark dots arranged in a grid.

Сводка по конкурентам

2 

Прямые конкуренты

Профильный интернет-магазин:

- SALETENNIS
- TENNIS PRO
- TENNIS DIRECT
- RUS TENNIS

Многопрофильный интернет-магазин:

- СПОРТМАСТЕР

Маркетплейсы:

Требуется дополнительное исследование



Посещаемость сайтов

Магазин	Посещаемость	Уникальные посетители	Доля уникальных посетителей	Среднее время сессии	Среднее количество страниц
tennis-store	30 304	18 100	60%	01:39	2.72
tennis-pro	11 507	9 702	84%	12:30	5.28
saletennis	13 707	9 053	66%	02:29	3.64
tennisdirect	3 522	3 456	98%	00:49	9.19
rus-tennis	1 017	988	97%	02:39	1.01
sportmaster/tennis	31 640	24 604	78%	03:59	4.68
sportmaster	14.7M	7.1M	49%	06:01	4.7



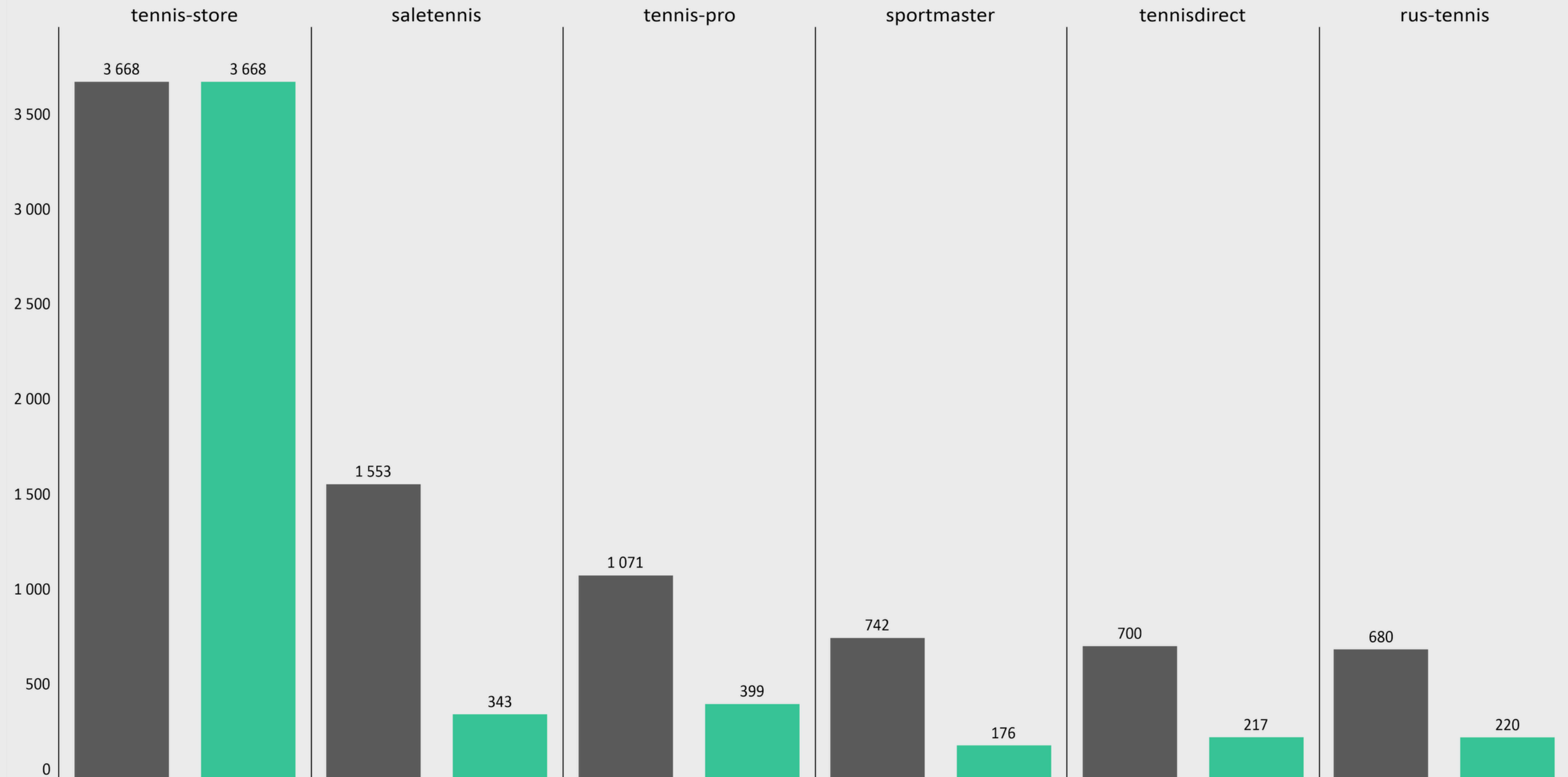
Прочее


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

■ Всего товаров
■ Аналогичный товар



The background features several large, overlapping geometric shapes in shades of dark blue, teal, and light grey. In the top-left corner, there is a 4x5 grid of small dark dots. In the bottom-right corner, there is a large dark blue number '3' followed by a 4x5 grid of small dark dots.

Формулирование проблемы и описание данных




Описание проблемы

В настоящий момент эффективность продаж снижается, поскольку отсутствует сегментация клиентов. Что затрудняет разработку маркетинговых стратегий, усложняет процесс продаж и делает взаимодействие с клиентами менее эффективным.

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

Для решения задачи я проанализирую историю продаж за два года.

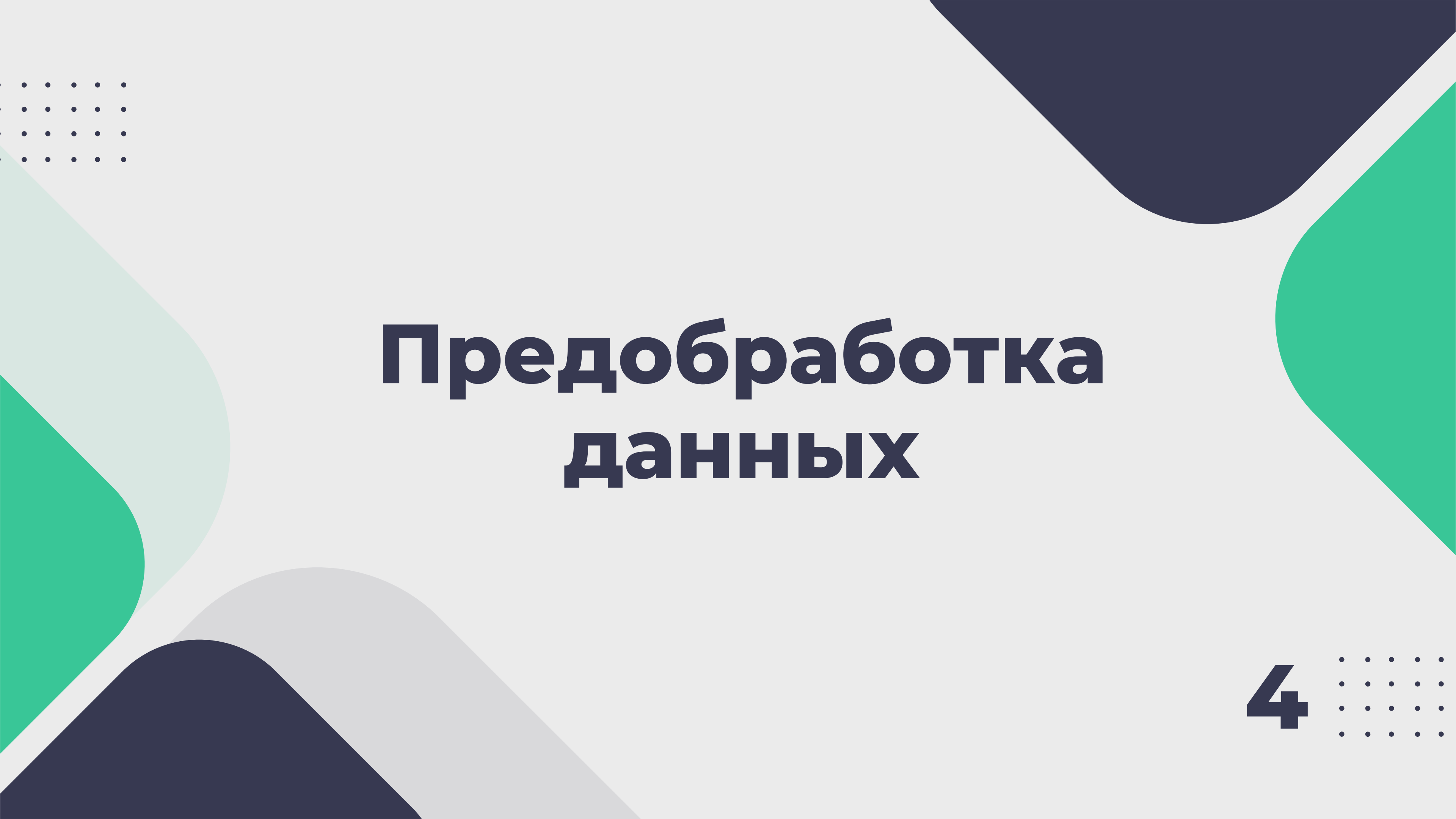


Описание данных

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
0	536365	22752	Теннисная ракетка Yonex New EZONE 105 (275g)	1	2023-12-01 08:26:00	4840	17850.0	Москва и МО
1	536366	85123A	Теннисная ракетка Wilson Ultra Team V4.0	1	2023-12-01 08:26:00	2550	17850.1	Москва и МО
2	536367	71053	Теннисная ракетка Head Geo Speed (MM TRADE)	1	2023-12-01 08:26:00	1142	17850.3	Москва и МО
3	536368	84406B	Теннисная ракетка Wilson Blade 98 (18X20) V8.0...	1	2023-12-01 08:26:00	3998	17850.5	Москва и МО
4	536369	84029G	Теннисная ракетка Wilson Six.One Lite 102	1	2023-12-01 08:26:00	3005	17850.8	Москва и МО
5	536370	84029E	Теннисная ракетка Yonex New EZONE 100L (285g) ...	1	2023-12-01 08:26:00	4993	17851.0	Москва и МО



Предобработка данных

Обработка null значений

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

Количество строк: 974022

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.

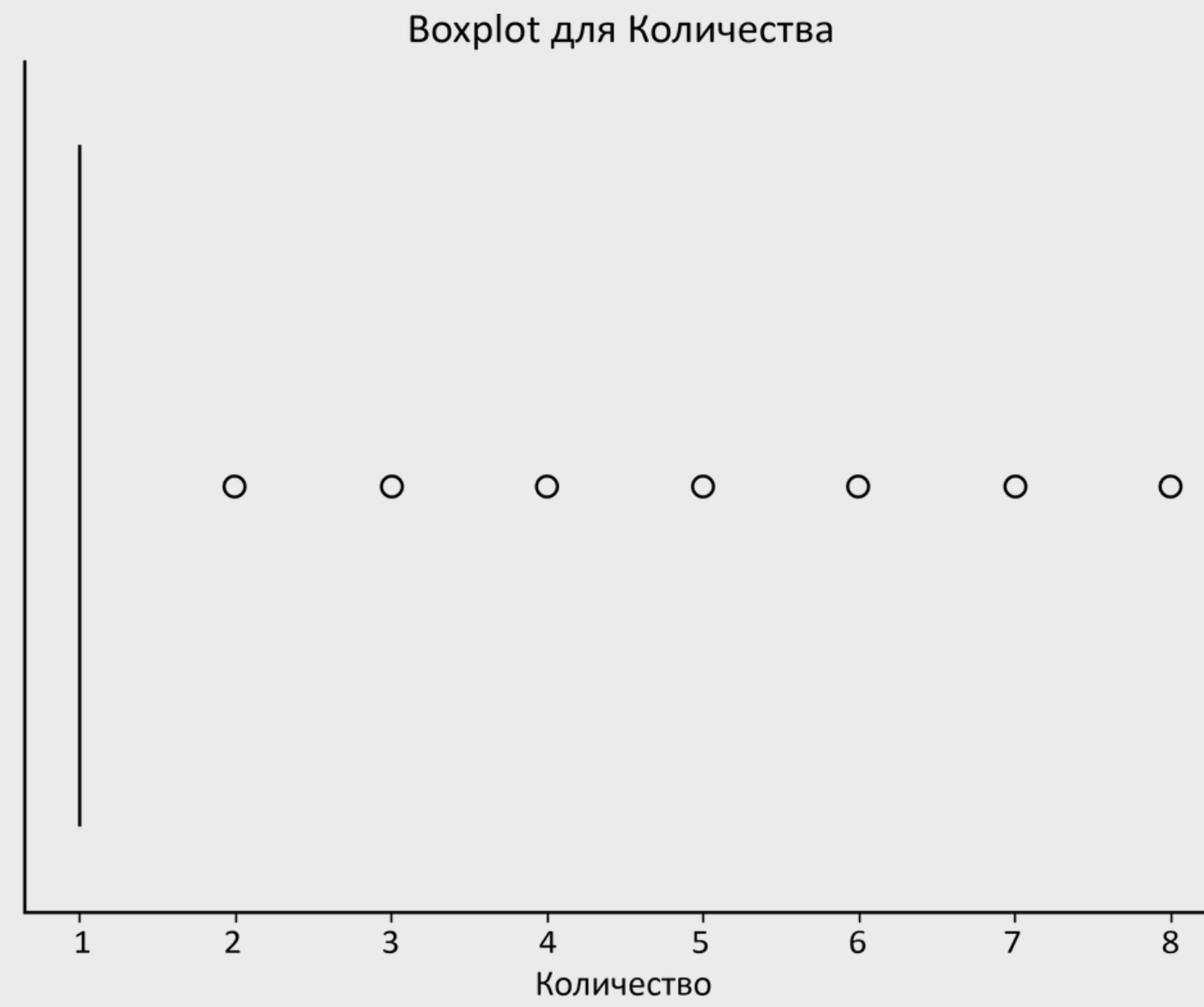
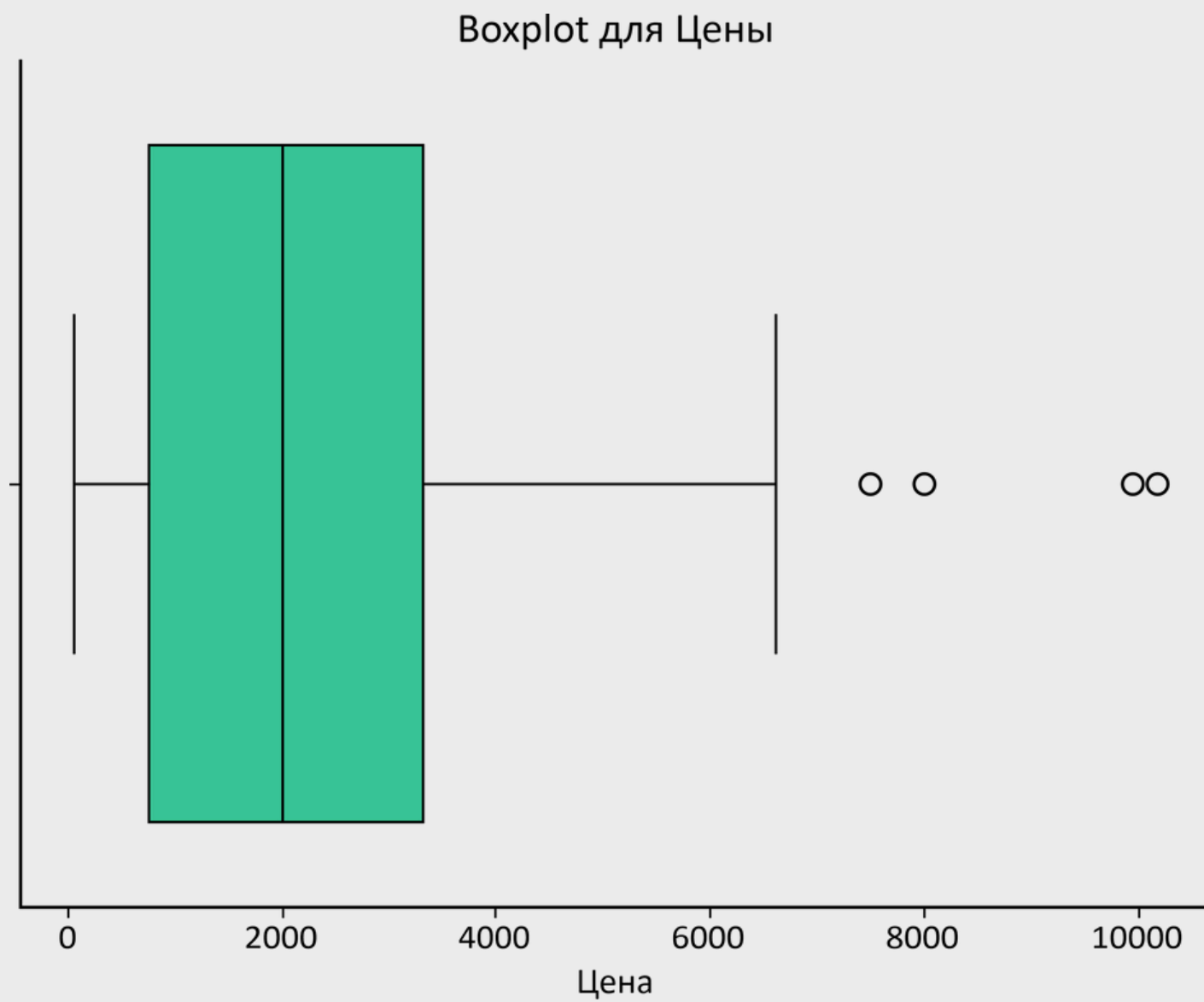


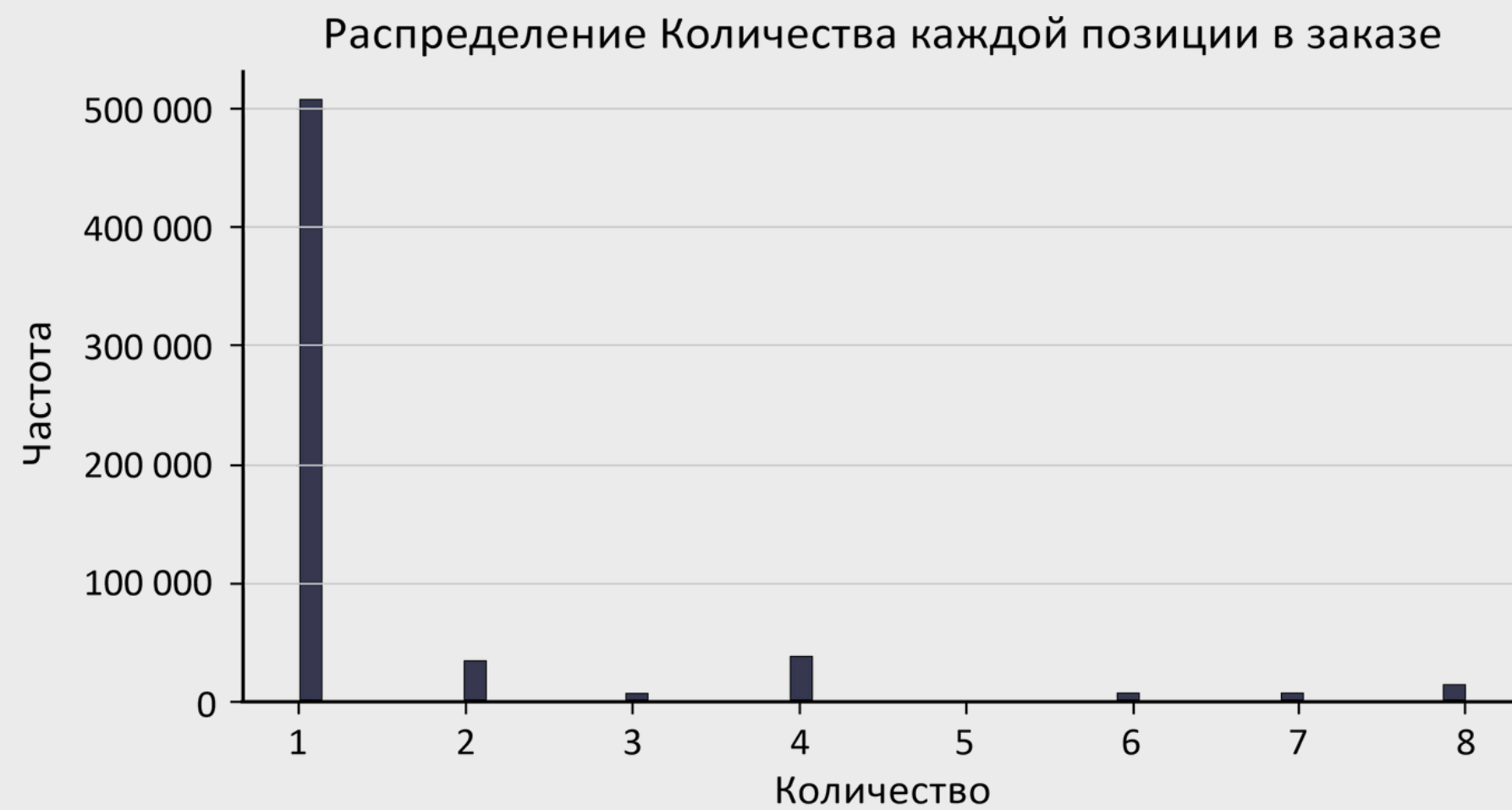
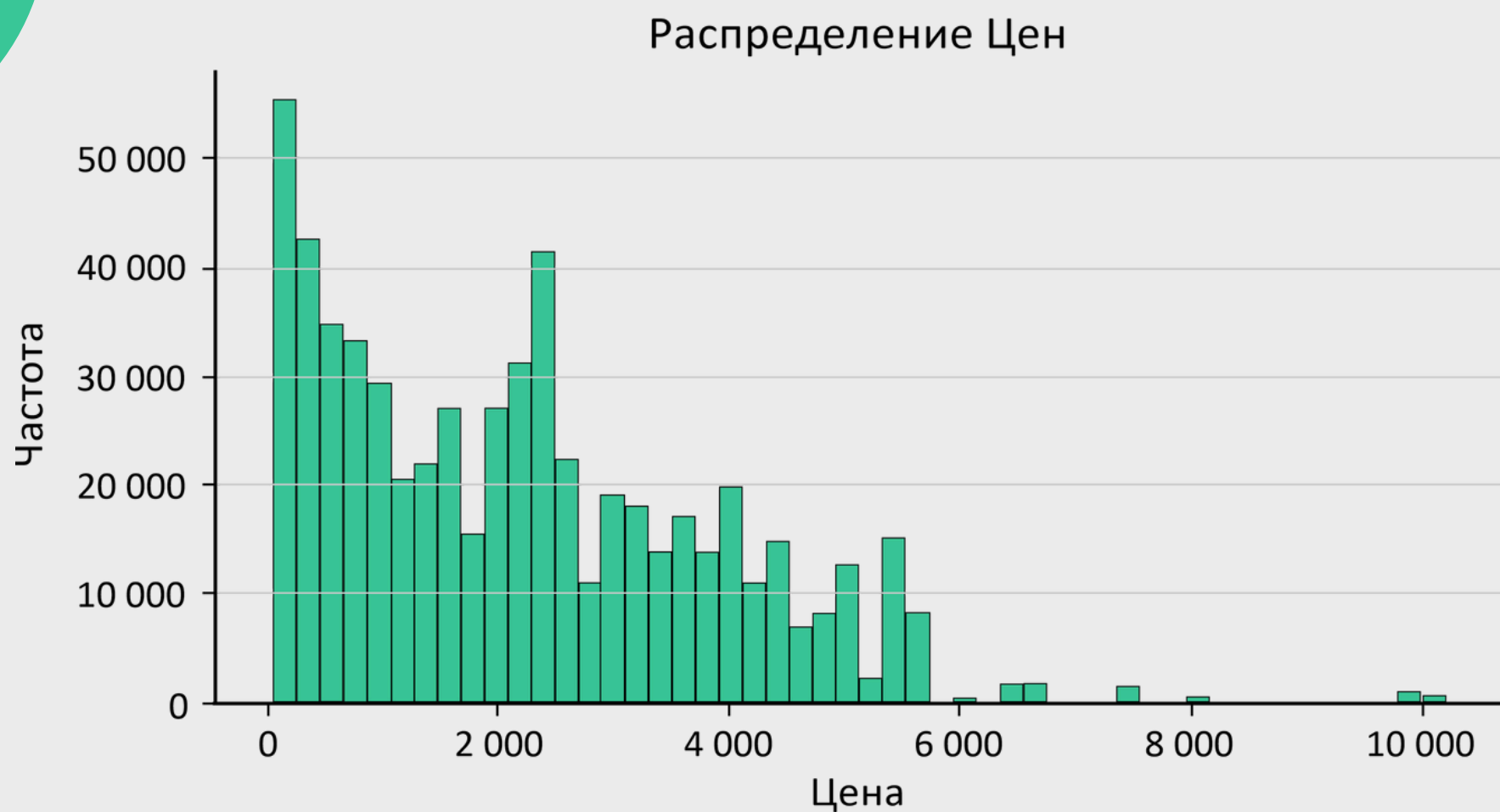
Обработка ошибок

- 1) *'ID заказа'* имеет тип данных object, что может говорить о наличии иных символов, кроме цифр. После проверки были удалены строки, содержащие "С" в id, т.к. это отмененные заказы.
- 2) *'Артикул'* - данные корректны.
- 3) *'Название'* - данные корректны.
- 4) *'Количество'* - никаких отклонений.
- 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.
- 8) *'Регион'* - данные корректны.

Обработка выбросов

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

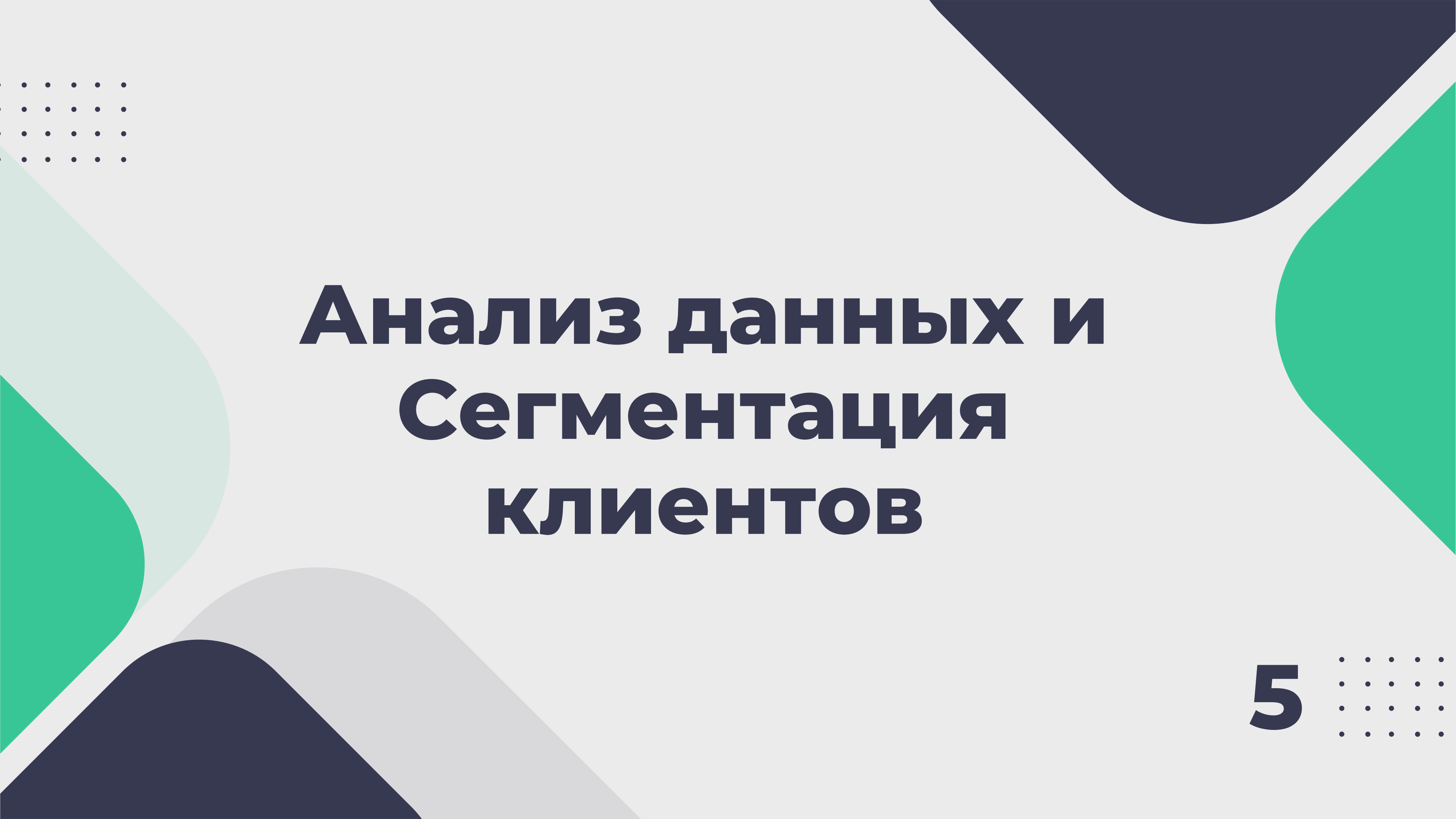




Как видно на графиках, аномальных значений нет.

После обработки данных количество строк сократилось на 38%.
Осталось 603 894 строки.



The background features a light gray base with several large, overlapping abstract shapes in dark blue, teal, and light green. In the top-left and bottom-right corners, there are 4x5 grids of small dark dots.

Анализ данных и 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
- Продажи в мес.
- Средний чек
- Среднее количество товара в заказе
- LifeTime (мес.)
- Месяцев с последней покупки
- Доля заказов в выходные

	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	7	48	97707.86	13	52611.92	683955	1	0.428571
1	12348.0	3	1	2105.00	9	701.67	6315	4	0.000000
2	12349.0	1	59	312099.00	1	312099.00	312099	2	0.000000
3	12352.0	7	38	34823.43	9	27084.89	243764	2	0.428571



Построим матрицу корреляции (метод Пирсона).

Никаких неожиданных взаимосвязей не наблюдается.

Сегментация клиентов

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

Поскольку K-means использует расстояние для измерения близости точек, необходимо привести данные к одному масштабу.

	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

Также необходимо определить оптимальное число кластеров с помощью метода “Локтя”.

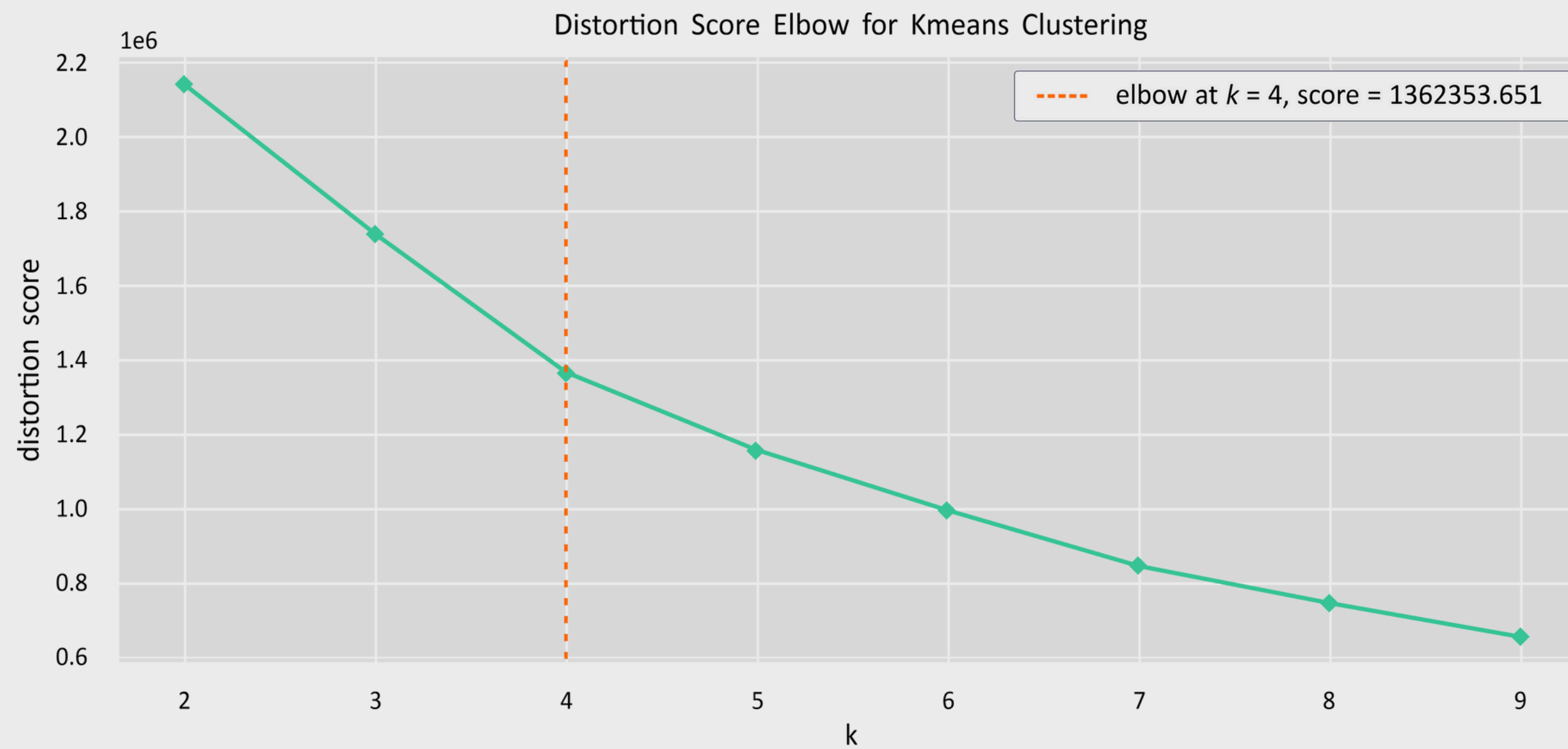


График показывает, что 4 - это оптимальное число кластеров.

Разделим базу клиентов на 4 сегмента и опишем их особенности.



Теперь необходимо оценить качество кластеров, с помощью 3-х показателей.

Silhouette Score (Силуэтный коэффициент, от -1 до 1).

Насколько хорошо каждый объект вписывается в свой кластер по сравнению с другими кластерами. В нашем случае **0.509** — неплохой показатель (чем ближе к 1, тем лучше).

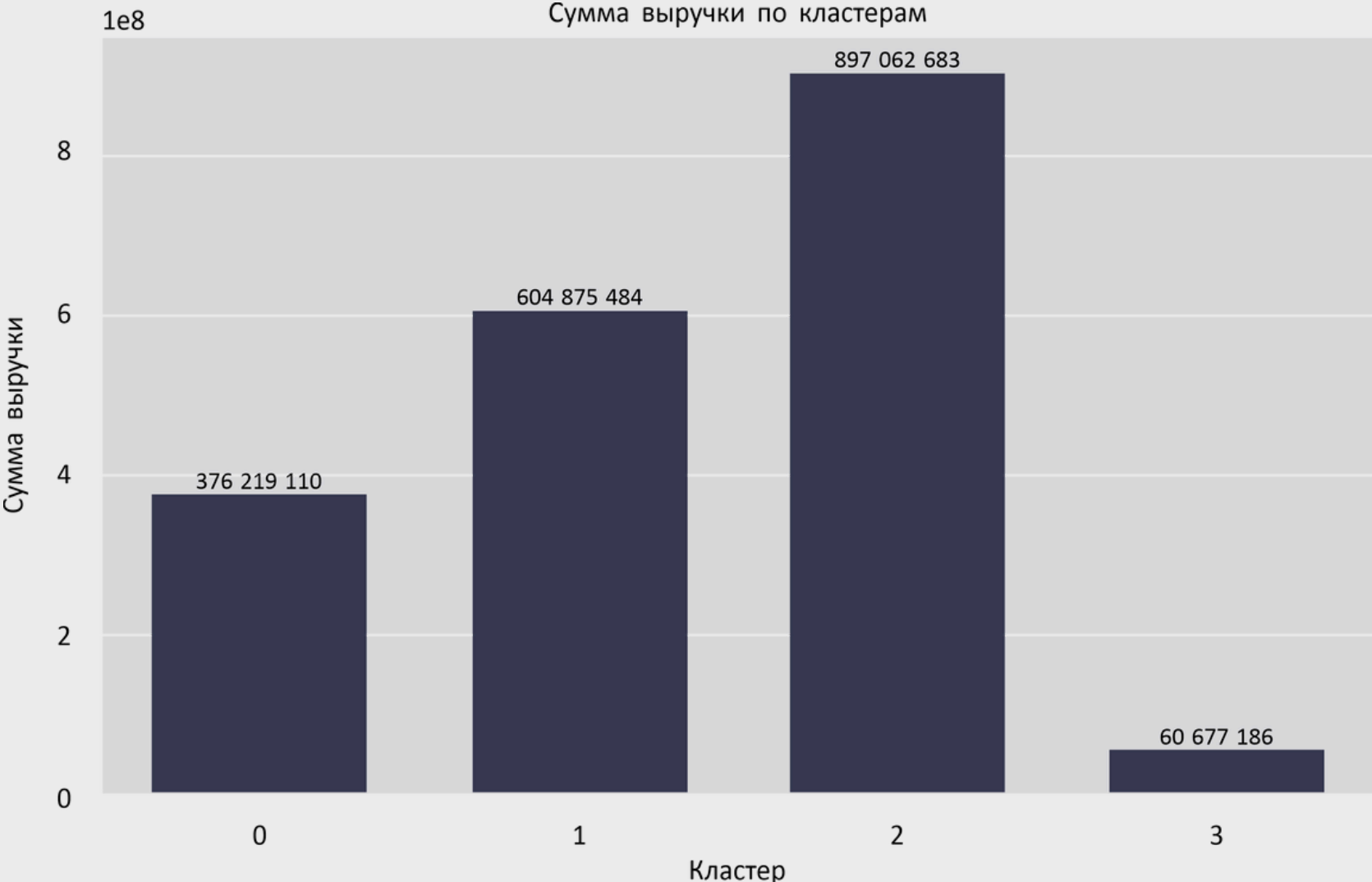
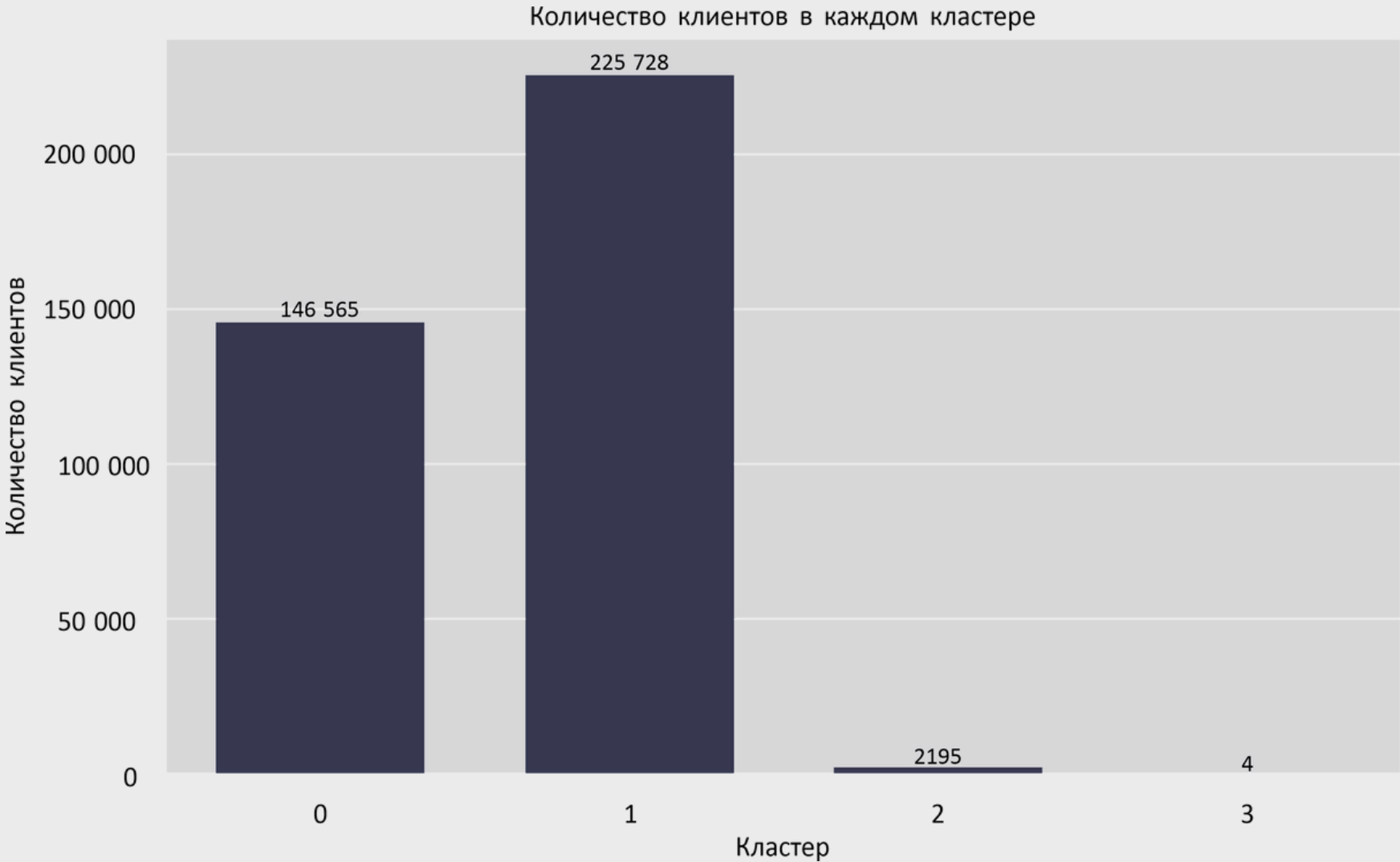
Calinski-Harabasz Score (Индекс Калински-Харабаша, От 0 и выше).

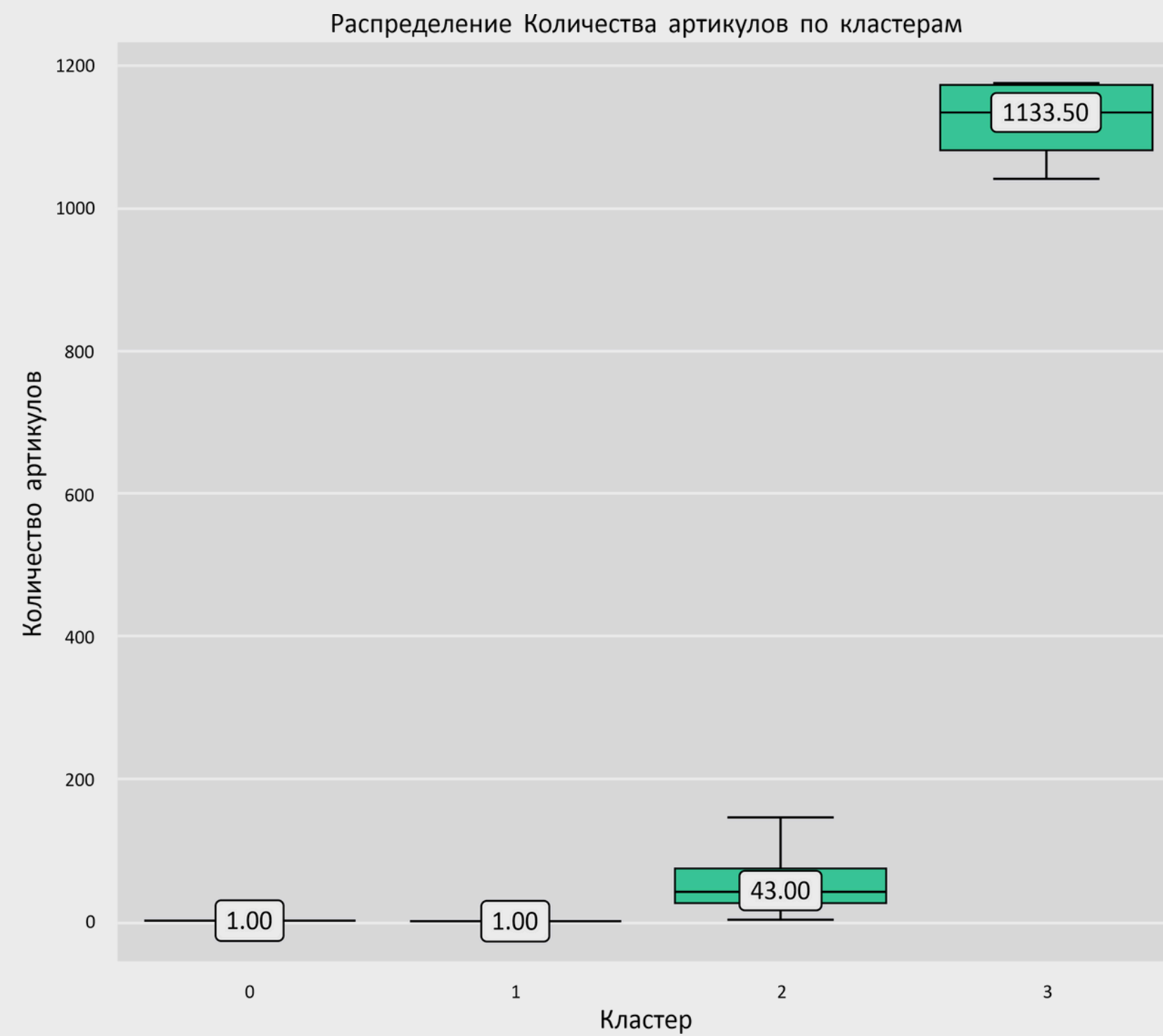
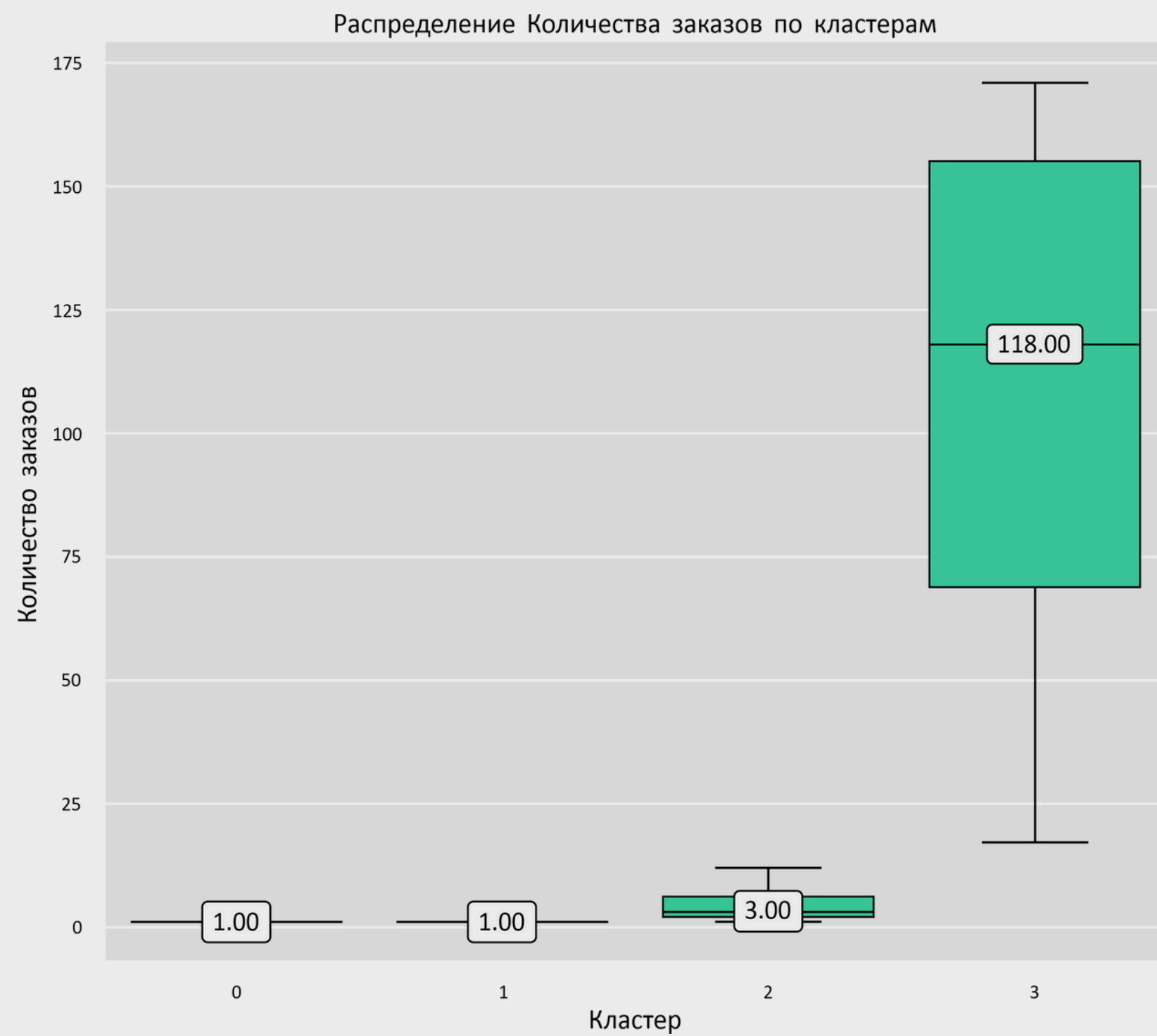
Показывает, насколько кластеры компактны внутри и разделены между собой **149681** — высокий показатель (чем больше, тем лучше).

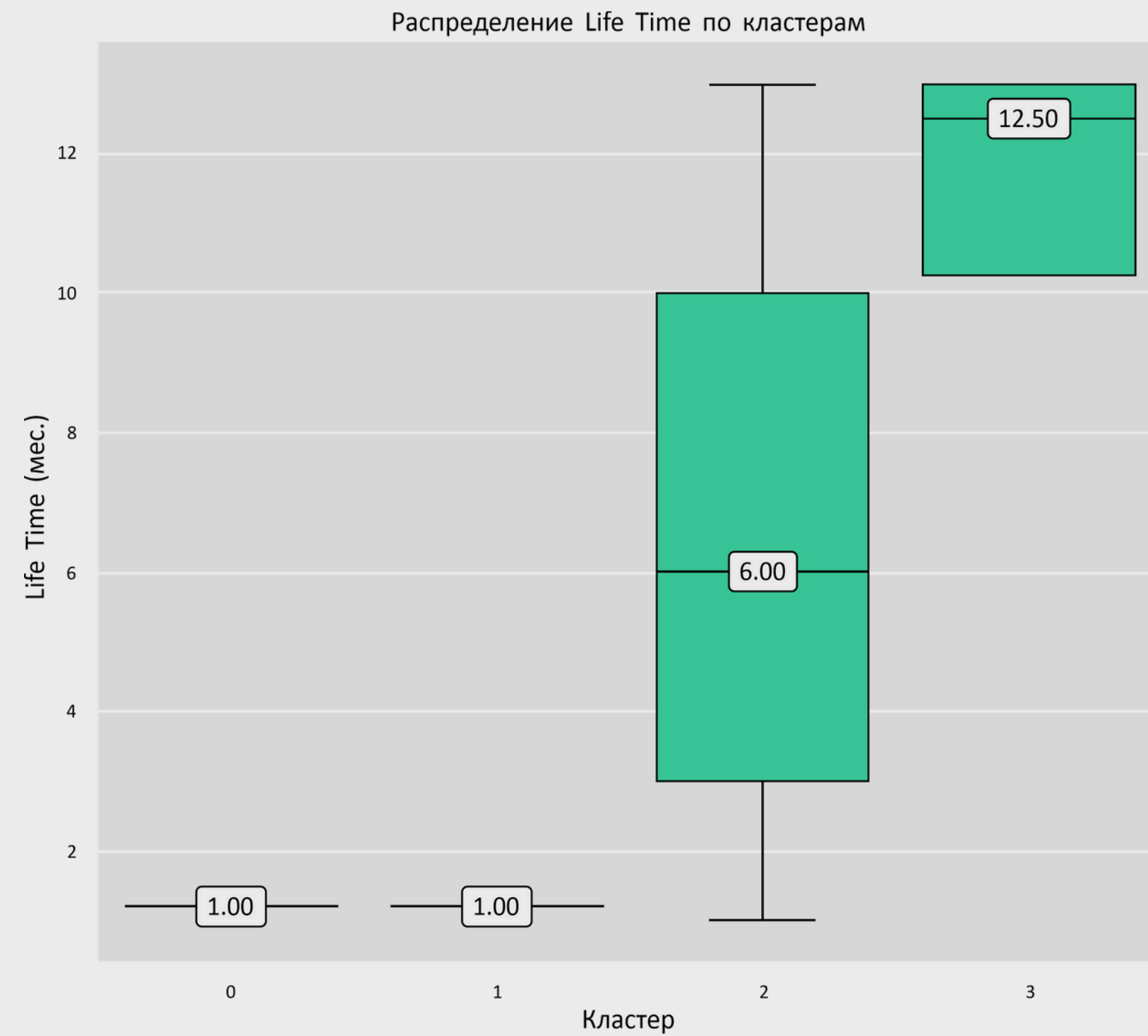
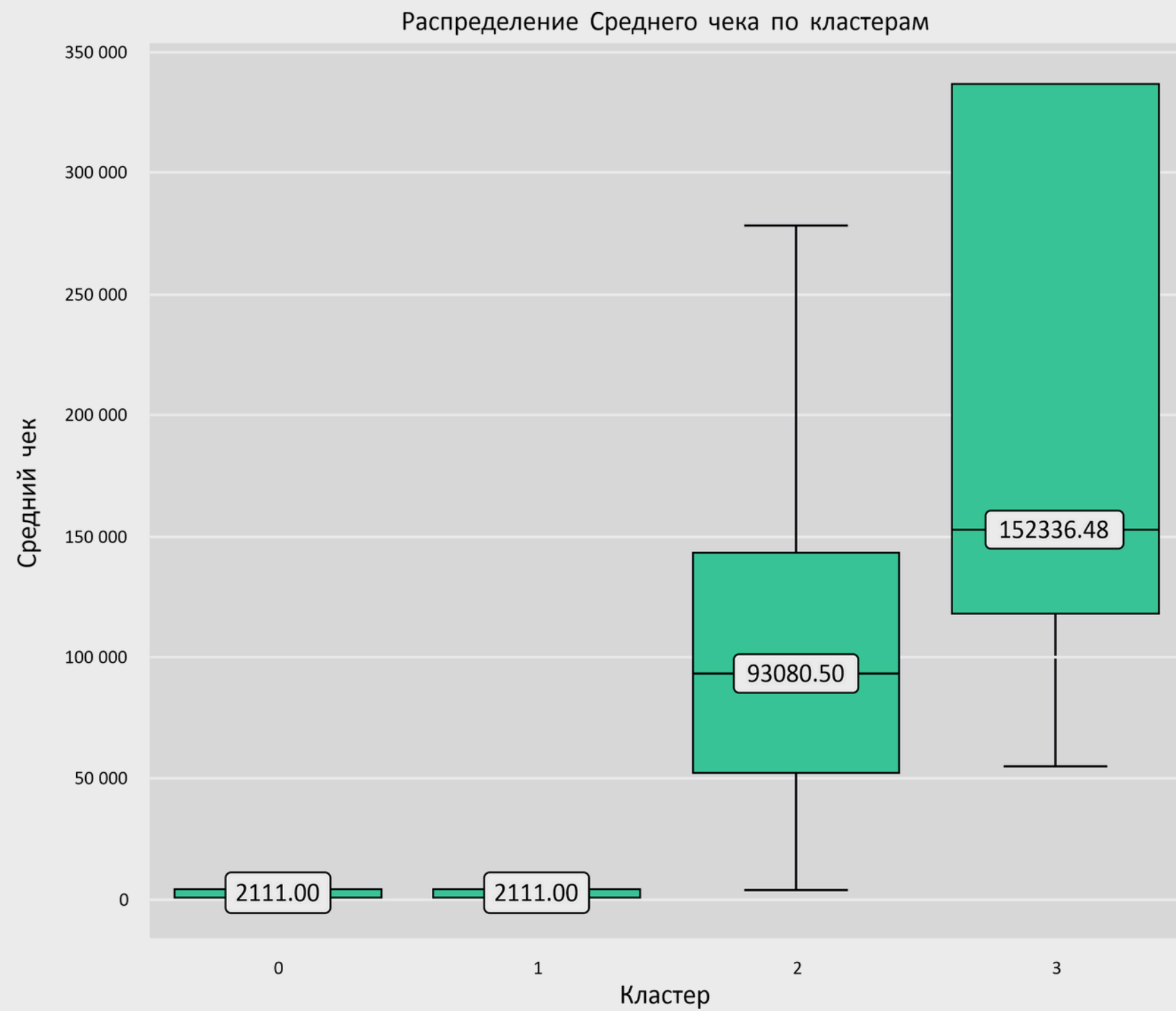
Davies-Bouldin Score (Индекс Дэвиса-Боулдина, от 0 до ∞).

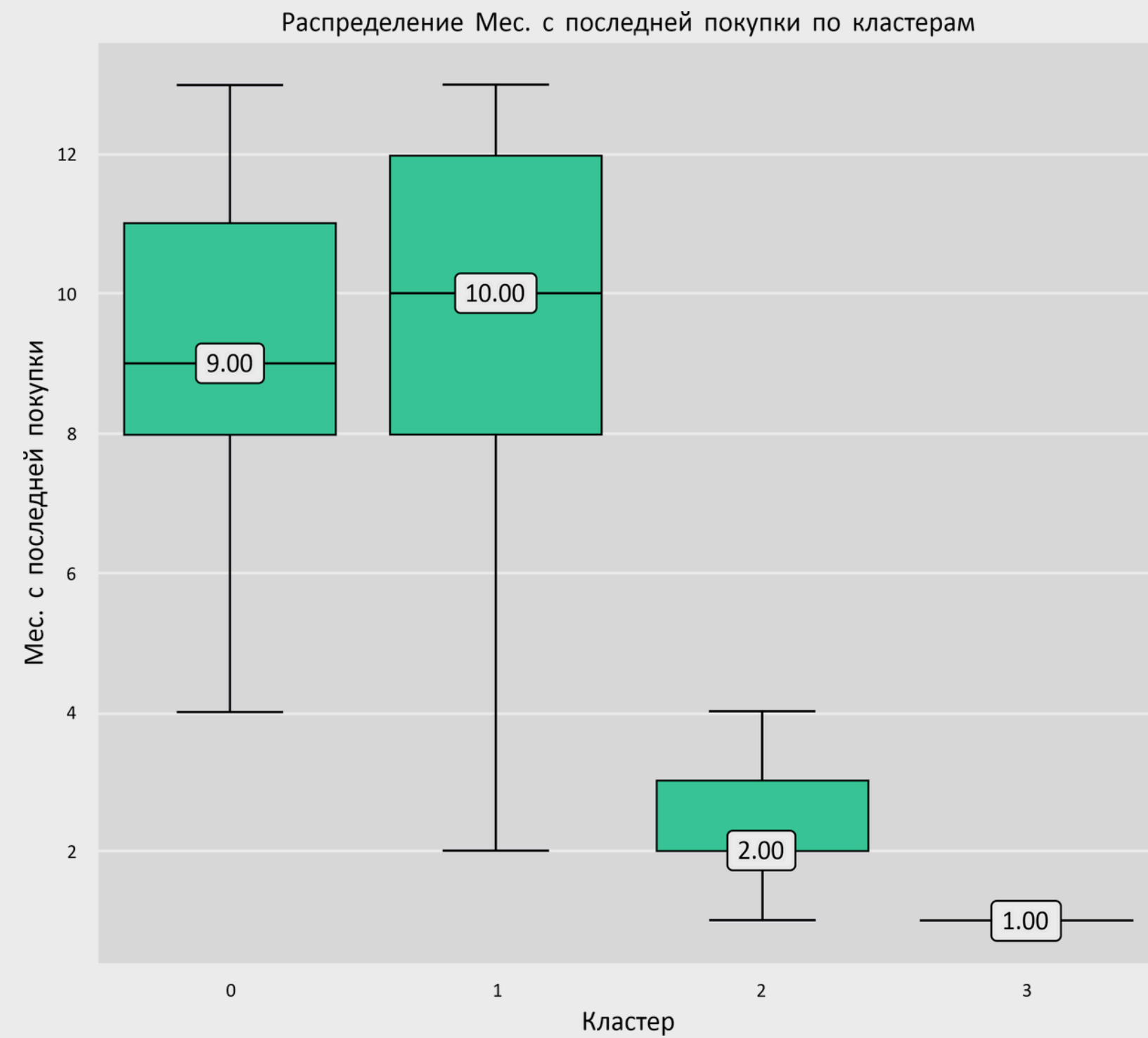
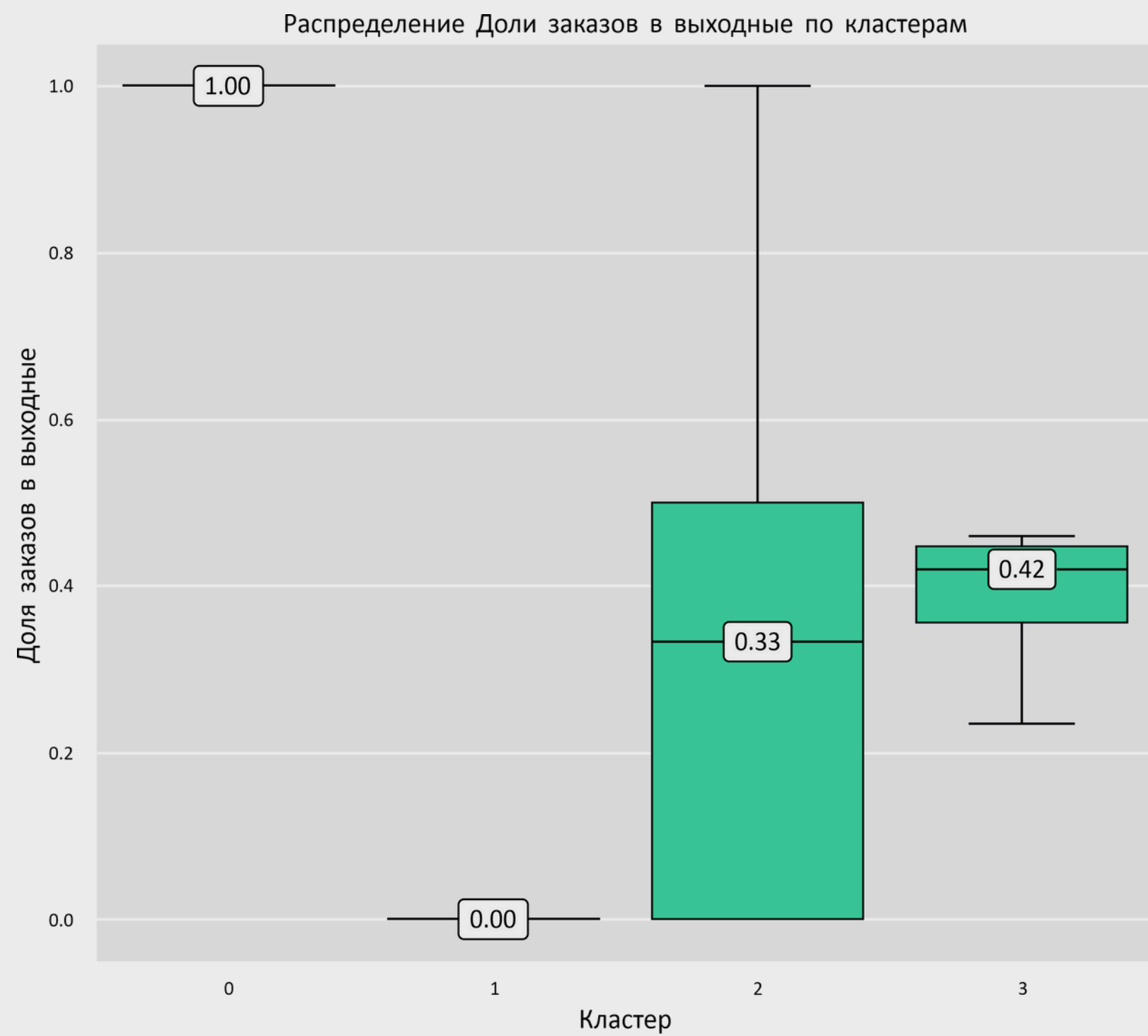
Измеряет насколько кластеры различимы, учитывая их разброс и расстояния между ними. **0.763** — хороший показатель, но кластеры частично пересекаются или имеют размытые границы.

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.

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.



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.

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 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. Если удастся вернуть часть таких клиентов, это может существенно увеличить выручку.

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 А/В теста



1. Определение размера выборки.

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

При расчете по формуле сравнения долей в двух группах получается маленькая выборка.

Поэтому, для повышения точности и получения дополнительной ценной информации, увеличим размер выборки до 10 000 в каждой группе, т.к. затраты на эксперимент минимальны.



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 теста 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.

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.

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 $\alpha = 5\%$ и мощности 80%

ВЫВОД

7






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.

Запланированный 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!