Документация

Дипломная работа по теме:

«Анализ продаж e-commerce магазина (поиск инсайтов, составление рекомендаций стейкхолдерам)».

Профессия “Аналитик данных”, DA-114

Букшенко Артур Максимович

г. Москва, 2025г

Оглавление

Введение

Цели проекта:

Проведение комплексного анализа продаж интернет-магазина для выявления ключевых закономерностей, скрытых инсайтов и формирования практических рекомендаций, направленных на увеличение прибыли и оптимизацию бизнес-процессов. Результаты работы предназначены для ключевых стейкхолдеров - генерального директора и директора по маркетингу компании.

Бизнес-контекст:

В условиях высокой конкуренции на рынке электронной коммерции критически важным становится использование данных для:

- Повышения эффективности товарного ассортимента
- Оптимизации маркетинговых активностей
- Улучшения клиентского опыта
- Максимизации прибыли

Основные задачи анализа:

- **Оценка динамики продаж**
  - Анализ географического распределения для выявления наиболее перспективных рынков
  - Исследование сезонных колебаний спроса
- **ABC-анализ товарного ассортимента**
  - Классификация продуктов по степени их вклада в выручку
  - Выявление ключевых товаров, требующих особого внимания
- **RFM-сегментация клиентской базы**
  - Разделение покупателей на группы по степени их ценности для бизнеса
  - Разработка персонализированных маркетинговых стратегий
- **Выявление сезонных трендов**
  - Анализ продаж топ-5 позиций для определения цикличности спроса
  - Подготовка рекомендаций по управлению запасами

Ожидаемый результат:

На основании проведенного анализа будут сформулированы конкретные рекомендации, позволяющие:

- Оптимизировать товарный ассортимент
- Улучшить эффективность маркетинговых кампаний
- Повысить лояльность ключевых клиентов
- Увеличить общую прибыльность бизнеса

Данное исследование предоставит руководству компании надежную аналитическую основу для принятия стратегических решений в области управления продажами и маркетингом.

Блок 1. Описание исходного датасета и типов данных (8 столбцов)

Для исследования был взят датасет “TATA: Online Retail Dataset” со статистикой продаж интернет-магазина.

№	Имя Столбца	Описание	Тип данных
1	**InvoiceNo**	Номер счета-фактуры (Уникальный номер счета-фактуры который идентифицирует каждую транзакцию)	object
2	**StockCode**	Код акций (Уникальный код товара, который используется для идентификации конкретного продукта в системе)	object
3	**Description**	Описание (Описание товара)	object
4	**Quantity**	Количество (Количество единиц товара, купленных в рамках данной транзакции)	int64
5	**InvoiceDate**	Дата счета-фактуры (Дата и время, когда была совершена покупка)	object
6	**UnitPrice**	Цена (Цена за единицу товара)	float64
7	**CustomerID**	Идентификатор клиента (Уникальный идентификатор клиента, который позволяет отслеживать покупки конкретного клиента)	float64
8	**Country**	Страна (Страна, из которой совершен заказ)	object

Блок 2. Подготовка и преобразование данных

В ходе исследования качества данных были сделаны следующие изменения:

2.1 Преобразование

В столбце “**InvoiceDate**”

- Изменен тип данных на “`datetime64[ns]`”
- Декомпозирован на отдельные составляющие:
  - Месяц (`Month`): название месяца
  - Год (`Year`): числовое значение года

В столбце “**Quantity**” - выполнена явная конвертация столбцов в числовые форматы (numeric)

В столбце “**UnitPrice**” - выполнена явная конвертация столбцов в числовые форматы (numeric)

Блок 3. Очистка данных

На этапе предобработки данных выявлено и удалено 5,268 дубликатов по всем столбцам (0.97% от общего объема записей). Это обеспечило корректность последующего анализа без значительной потери данных.

Результат итогового вида обработанного датасета, а также выводы по причинам очистки данных приведены в таблице:

№	Имя столбца	Преобразование данных	% NaN	Очистка данных
1	**InvoiceNo**	object	0.00	Без изменений. Нет пустот. Обнаружены возвраты с отрицательным значением Quantity, где номер транзакции (InvoiceNo) начинается на букву "C". Данные транзакции будут удаляться только для проведения RFM-анализа. Для других видов анализа они сохранены, так как при расчете общих сумм положительные и отрицательные значения компенсируют друг друга. Удалена 1 строчка с большой суммой и отрицательным количеством для того чтобы не портить общую картину анализов.
2	**StockCode**	object	0.00	Без изменений. Нет пустот
3	**Description**	object	*0.27*	Пустоты заменены на ‘unknown’, чтобы не искажать общие показатели, основной ориентир будет на StockCode(столбец в котором

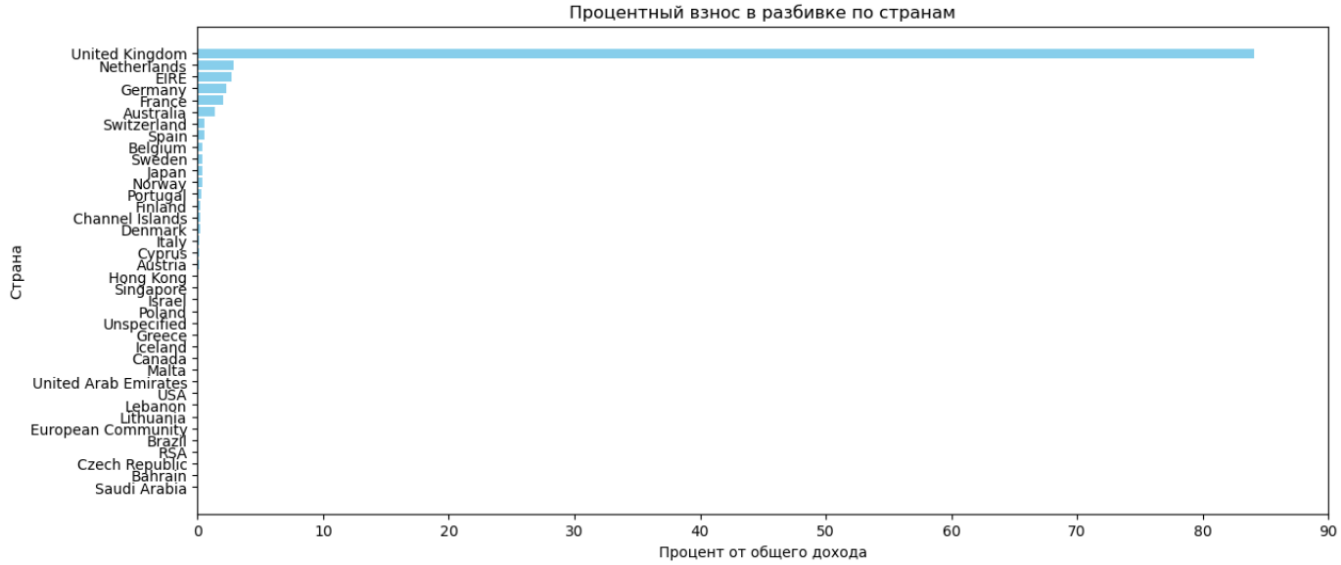
				не было обнаружено пропусков), при необходимости позицию можно определить по StockCode).
4	**Quantity**	int64	0.00	Удалены 1336 строк с отрицательным значением и 0 ценной, которые сложно отнести к какой либо категории товара
5	**InvoiceDate**	datetime64[ns]	0.00	Без изменений. Нет пустот. 2011.12 не полный месяц, данные только до 2011-12-09
6	**UnitPrice**	float64	0.00	Удалены 3 строчки с задолженностями ‘Adjust bad debt’. 2 из них с отрицательным значением, чтобы не искажать общие показатели
7	**CustomerID**	float64	*24.93*	Пустоты заменены на ‘0’, чтобы не искажать общие показатели. Для RFM анализа будет сделана дополнительная очистка по удалению строк с 0 значением
8	**Country**	object	0.00	Без изменений. Нет пустот

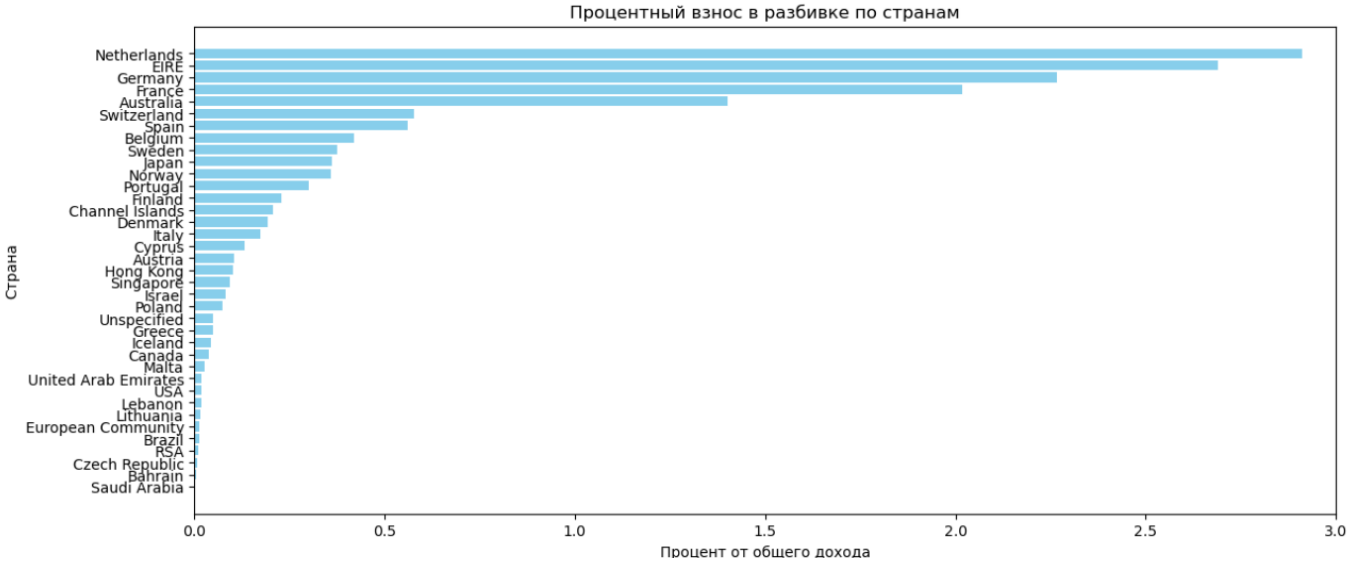
Блок 4. Анализ данных для стейкхолдеров

Целью блока является поиск тенденций и инсайтов для составления рекомендаций стейкхолдерам.

4.1.1 Анализ общих продаж

Динамика продаж за период 2010.12 - 2011.12





Вывод:

Географическое распределение выручки крайне неравномерно:

- Великобритания генерирует 84.05% общего дохода — абсолютный лидер.
- ТОП-5 стран (Великобритания, Нидерланды, Ирландия, Германия, Франция) обеспечивают 93.9% выручки.
- 35 остальных рынков в сумме дают лишь 6.1% дохода.

Неэффективные рынки:

- 25 стран (например, Бахрейн, ОАЭ, США) приносят менее 0.1% выручки каждая.
- Некоторые направления (Саудовская Аравия, Чехия) демонстрируют минимальную активность (<0.01%).

Потенциал роста:

- Страны с 2-3% долей (Германия, Франция, Ирландия) могут стать точками роста при увеличении маркетинговых инвестиций.

Рекомендации для стейкхолдеров:

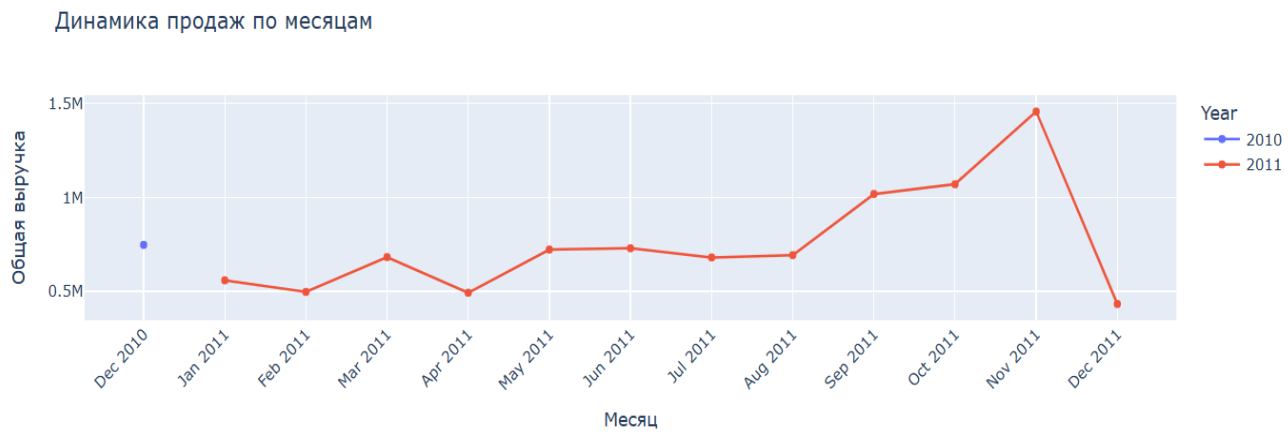
Оптимизация ресурсов:

- Сфокусироваться на ТОП-5 рынках, где **ROI**(окупаемость вложенных инвестиций) максимален.
- Для стран с долей <0.1% провести аудит:
    - Если затраты на логистику/поддержку превышают выручку рассмотреть возможность сокращения присутствия.

Тактические действия:

- Для Великобритании (84% выручки):
    - Внедрить программу лояльности для удержания клиентов.
    - A/B-тестировать повышение среднего чека (например, через кросс-продажи).
- Для перспективных рынков (Нидерланды, Германия, Франция):
    - Увеличить рекламный бюджет на 15-20% с акцентом на локальные тренды.

4.1.2 Анализ продаж с разбивкой по месяцам за период 2010.12 - 2011.12



	Year	Month	Total	Revenue
0	2010	December	746723.610	NaN
5	2011	January	558448.560	-0.252135
4	2011	February	497026.410	-0.109987
8	2011	March	682013.980	0.372189
1	2011	April	492367.841	-0.278068
9	2011	May	722094.100	0.466574
7	2011	June	728947.230	0.009491
6	2011	July	680156.991	-0.066932
2	2011	August	692448.520	0.018072
12	2011	September	1017596.682	0.469563
11	2011	October	1069368.230	0.050876
10	2011	November	1456145.800	0.361688
3	2011	December	432701.060	-0.702845

Вывод:

- Пиковые месяцы: ноябрь (+36,2% к октябрю), сентябрь (+47% к августу) и май (+46,7% к апрелю).
- Сезонные спады: апрель (-27,8% к марту), январь (-25,2% к декабрю 2010).

- **Рекомендации для стейкхолдеров:**

Оптимизация маркетинга:

- Увеличить бюджет в ключевые месяцы:
  - Ноябрь: запустить предновогодние акции (например, "Черная пятница").
  - Сентябрь: продвигать товары для учебы/хобби.
  - Май: предложить "летние стартовые наборы".
- Для спадов (апрель, январь):
  - Ввести программы удержания ("Скидка 15% на следующий заказ").

Логистика и запасы:

- Пиковые периоды:
  - Увеличить складские запасы на 30% для топ-20 товаров.
  - Нанять временный персонал для обработки заказов.

Планирование:

- Разработать "антикризисный" бюджет на январь-апрель с акцентом на retargeting.
- Внедрить ежеквартальные промо-кампании для сглаживания спадов.

4.2.1 ABC анализ по товарам

StockCode	Quantity	Share	Cumsum	abc
22197	56427	0.01	0.01	A
84077	53751	0.01	0.02	A
85099B	47260	0.01	0.03	A
85123A	39067	0.01	0.04	A
84879	36282	0.01	0.04	A
...	...	...	...	...
85044	1	0.00	1.00	C
20950	1	0.00	1.00	C
23843	0	0.00	1.00	C
85047	0	0.00	1.00	C
79323B	0	0.00	1.00	C

3932 rows × 4 columns

Вывод:

Распределил товар по категориям:

- **Категория А** (первые 20% товаров):
  - Вклад в продажи: ~80% от общего объема.

Небольшая доля товаров приносит основную прибыль. Это «ключевые» позиции, требующие особого контроля (оптимизация запасов, защита от дефицита).

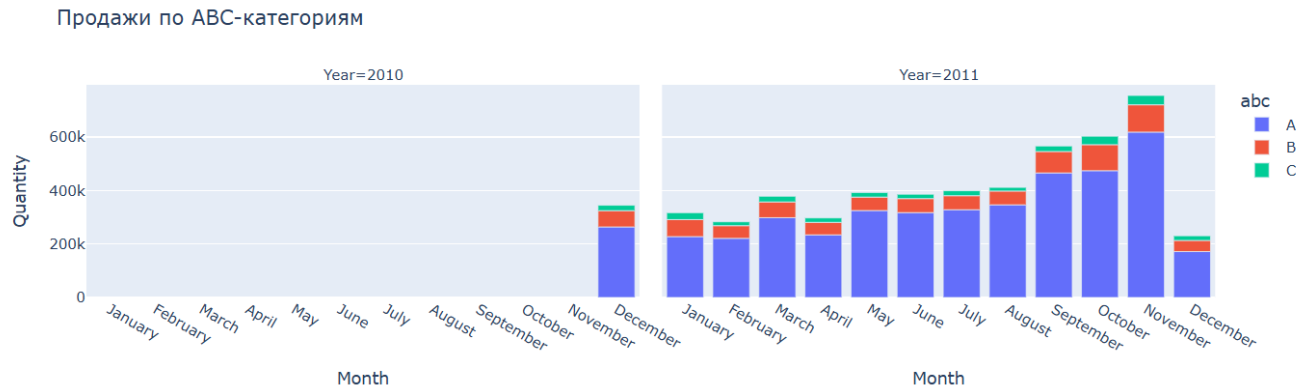
- **Категория B** (следующие 15% товаров):
    - Вклад в продажи: ~15% от общего объема.

Товары со средним спросом. Возможно, имеют потенциал для перевода в категорию A при правильном продвижении.

- **Категория C** (оставшиеся 65% товаров):
    - Вклад в продажи: ~5% от общего объема.

«Хвост» ассортимента. Многие товары могут быть неэффективны. Требуют анализа на исключение или оптимизацию.

4.2.2 Продажи по ABC категориям за период 2010.12 - 2011.12



Вывод:

Прослеживаются сезонные тренды:

- Ноябрь — абсолютный пик для всех категорий (рост на 30-50% к среднему).
- Апрель и январь — периоды снижения спроса для всех категорий.

Рекомендации для стейкхолдеров:

Управление ассортиментом

Для категории A

- Ввести динамическое управление запасами:
    - Автоматический заказ при достижении порогового уровня (например, 20% от месячного объема).

    - Увеличить страховой запас на 25% для ТОП-10 товаров перед ноябрем.
- Оптимизировать логистические цепочки для ключевых позиций.

Для категории B

- Запустить программу лояльности:
    - Предлагать товары B-категории как дополнение к A-товарам (например, «Вместе дешевле»).
    - Выделить 5-7 перспективных позиций для перевода в категорию A через точечное продвижение.

Для категории C

- Провести аудит ассортимента:
    - Вывести из ассортимента товары с нулевыми продажами
    - Перевести 20% низко продаваемых позиций на dropshipping.
- Создать наборы (например, «3 товара категории C по цене 2»).

Маркетинг и продвижение

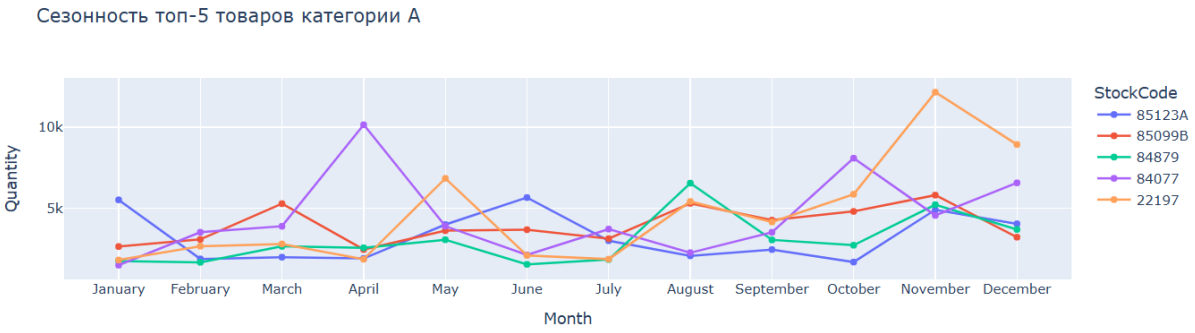
- Пиковые месяцы (ноябрь, сентябрь):
    - Увеличить рекламный бюджет на 30% для категории A.
    - Запустить предзаказы на сезонные хиты за 2 месяца до пика.
- Сезонные спады (январь, апрель):
    - Предлагать скидки 10-15% на товары категории B.
    - Провести A/B-тесты новых маркетинговых стратегий.

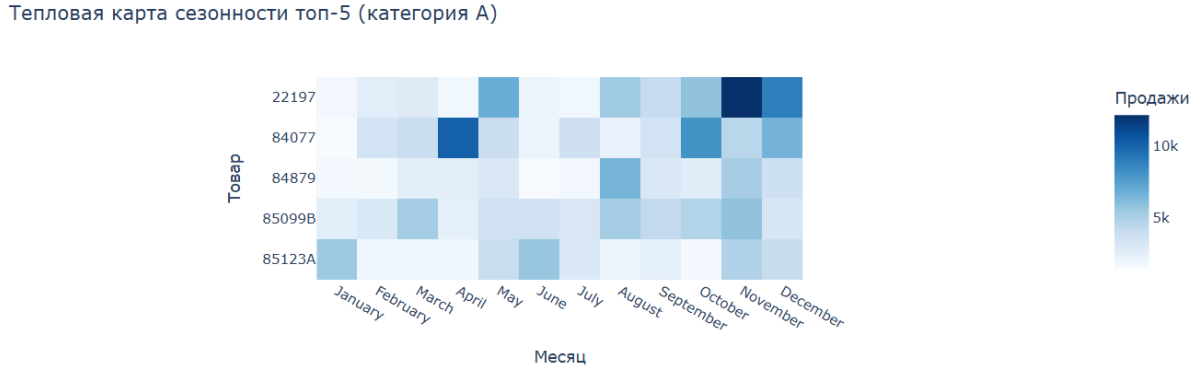
4.3.1 Топ 5 товаров категории A

StockCode	Quantity	Share	Cumsum	abc
22197	56427	0.01	0.01	A
84077	53751	0.01	0.02	A
85099B	47260	0.01	0.03	A
85123A	39067	0.01	0.04	A
84879	36282	0.01	0.04	A

Вывод:
Получившийся топ показывает абсолютных лидеров продаж.

4.3.2 Сезонность топ-5 товаров категории A





Вывод:

Ярко выраженная сезонность

- Пиковые месяцы: ноябрь (макс. продажи 22197), декабрь (22197), май (22197).
- Спад: январь-апрель (минимальные значения для большинства позиций).

Особенности товаров

- Стабильные хиты: 22197, 84077, 85099B — высокие продажи круглый год.
- Сезонные всплески:
  - 84879 — пик в августе

Распределение по месяцам

- Ноябрь-декабрь: 40-50% годовых продаж для большинства ТОП-товаров.

- Летние месяцы: умеренный спрос с локальными пиками.

Рекомендации для стейкхолдеров:

- Сфокусируйтесь на управлении 5 ключевыми товарами
- Автоматизируйте процессы пополнения запасов для 22197 и 84077.
- Используйте пиковые месяцы для максимизации прибыли через таргетированный маркетинг.

4.4.1 RFM анализ, разработка системы меток

	CustomerID	Recency	Frequency	Monetary	R	F	M	rfm_score	RFM_Label	rfm_segment
0	12346	325	1	77183.60	2	1	5	8	2 1 5	Hibernating
1	12347	366	182	4310.00	1	5	5	11	1 5 5	Can't loose them
2	12348	357	31	1797.24	1	3	4	8	1 3 4	At risk
3	12349	18	73	1757.55	5	4	4	13	5 4 4	Campions
4	12350	309	17	334.40	2	2	2	6	2 2 2	Hibernating
...	...	...	...	...	...	...	...	...	...	...
4334	18280	277	10	180.60	3	1	1	5	3 1 1	About to sleep
4335	18281	180	7	80.82	4	1	1	6	4 1 1	Promising
4336	18282	125	12	178.05	4	1	1	6	4 1 1	Promising
4337	18283	336	721	2045.53	2	5	4	11	2 5 4	Can't loose them
4338	18287	201	70	1837.28	4	4	4	12	4 4 4	Loyal customers

4339 rows × 10 columns

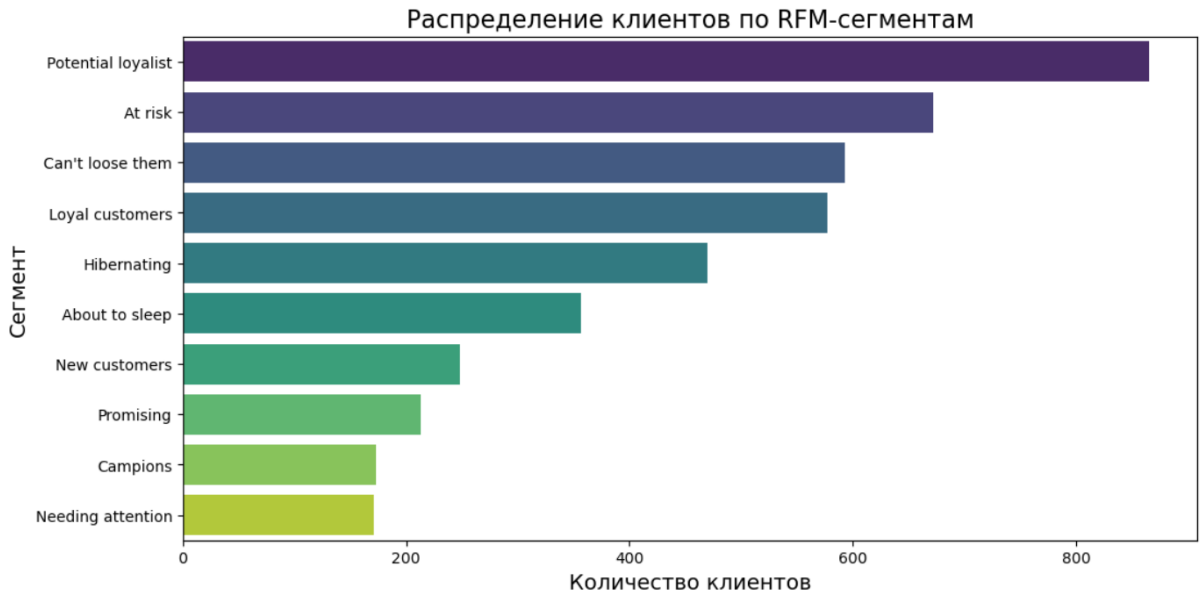
Вывод:

Разработана система меток которая позволила разбить клиентов на 10 ключевых групп:

- **Champions**
  - Клиенты с максимальными показателями по всем трем параметрам (R, F, M). Совершают покупки регулярно, недавно и на большие суммы.
- **Loyal Customers**
  - Высокие Frequency и Monetary, но Recency может быть чуть ниже, чем у Champions. Лояльны, но менее активны в последнее время.
- **Potential Loyalist**
  - Недавние покупатели (высокий R), совершившие 2+ покупки (средний F) и потратившие выше среднего (M). Перспектива стать Champions.
- **New Customers**

    - Высокий Recency (купили недавно), но низкие Frequency и Monetary (пока только одна небольшая покупка).
- **Promising**
    - Недавние (R), но пока слабая вовлеченность (низкие F и M). Аналог New Customers, но с чуть большей историей.
- **Needing Attention**
    - Средние значения по всем параметрам. Не выделяются, но могут уйти в "спящие".
- **About To Sleep**
    - Низкие Frequency и Monetary, но покупка была относительно недавно (средний R). На грани перехода в "спящие".
- **At Risk**
    - Высокие F и M в прошлом, но давно не покупали (низкий R). Ценные, но теряющие интерес.
- **Can't Loose Them**
    - Исторически VIP-клиенты (высокие F и M), но длительное отсутствие (низкий R). Критически важны для бизнеса.
- **Hibernating**
    - Низкие R, F, M. Покупали давно, редко и мало. Часто "мертвый" сегмент.

4.4.2 Распределение клиентов по RFM- сегментам



Вывод:

Доминирующие группы

- **Potential loyalist** (865 клиентов) – перспективные клиенты с высоким потенциалом лояльности

- **At risk** (672 клиента) – ценные клиенты, начинающие отдаляться
- **Can't loose them** (593 клиента) – VIP-клиенты с высокой частотой покупок, но давно не совершавшие заказов

Наиболее ценные сегменты

- **Champions** (173 клиента)
    - Совершают покупки регулярно и недавно
    - Генерируют максимальный доход (высокий Monetary)
- **Loyal customers** (577 клиентов)
    - Постоянные покупатели со стабильной частотой заказов
    - Средний чек выше среднего

Проблемные зоны:

- **Hibernating** (470 клиентов) – "спящие" клиенты с низкой активностью
- **About to sleep** (357 клиентов) – клиенты на грани ухода
- **At risk** (672 клиента) – требуют немедленного внимания для удержания

Рекомендации для стейкхолдеров:

Для топ-сегментов (Champions, Loyal customers)

- **Программы лояльности**
    - Ввести статусную систему с эксклюзивными бонусами
    - Персональные предложения на основе истории покупок
- **Маркетинг**
    - Early access к новинкам
    - Приглашения на закрытые мероприятия/презентации

Для перспективных сегментов (Potential loyalist, Promising)

- **Стимулирование повторных покупок**
    - Купон на скидку для второго заказа
    - Программа "Приведи друга" с усиленными бонусами
- **Персонализация**
    - Рекомендации на основе первого заказа
    - Таргетированные email-рассылки

Для проблемных сегментов (At risk, About to sleep, Hibernating)

- **Реактивационные кампании**

    - Персональные письма с эксклюзивным предложением
    - Push-уведомления о новых поступлениях в интересующих категориях
- **Специальные условия**
    - Бесплатная доставка для следующего заказа
    - Подарок к следующей покупке

Для VIP-клиентов (Can't loose them):

- **Персональный менеджер**
    - Отдельный канал поддержки
    - Индивидуальные условия сотрудничества
- **Эксклюзивные предложения**
    - Доступ к limited edition товарам
    - Сервис персонального шоппера

Итоги проекта и заключение

Проведенный комплексный анализ данных позволил выявить ключевые закономерности в продажах, клиентском поведении и эффективности товарного ассортимента. На основании исследования сформулированы конкретные рекомендации для оптимизации бизнес-процессов и увеличения прибыли.

Ключевые результаты:

- **Географический анализ**
    - 84% выручки генерирует Великобритания
    - 25 стран обеспечивают менее 0.1% дохода каждая
- **Сезонность**
    - Пики продаж в ноябре (+36%), сентябре (+47%) и мае (+47%)
    - Спад в январе (-25%) и апреле (-28%)
- **ABC-анализ**
    - 20% товаров (категория A) дают 80% выручки
- **RFM-анализ**
    - Выделено 10 клиентских сегментов
    - 593 VIP-клиента ("Can't loose them") требуют особого внимания
    - 865 перспективных клиентов ("Potential loyalist")

Оценка выполнения поставленных задач:

- **Оценка динамики продаж**
    - Построены графики динамики, выявлены сезонные тренды
- **Анализ географического распределения**

    - Определены ключевые рынки, даны рекомендации по оптимизации
- **ABC-анализ товарного ассортимента**
    - Товары классифицированы, выявлены проблемные позиции
- **RFM-сегментация клиентской базы**
    - Разработана система сегментации с рекомендациями
- **Выявление сезонных трендов**
    - Определены пиковые и спадовые периоды

Заключение

Данное исследование предоставляет аналитическую основу для принятия управленческих решений. Следующий шаг — тестирование рекомендаций на практике с последующим контролем метрик. Реализация предложенных мер позволит компании увеличить прибыльность и укрепить позиции на рынке.