Дипломная работа по теме:

Исследование данных супермаркета

(поиск инсайтов, составление рекомендаций стейкхолдерам, построение дашборда)

Подготовила: Мукаева Гера
Курс: «Аналитик BI: с нуля до middle»
Код набора: ABU-77

2025 г

Содержание:

Описание проекта

1. Заказчик – директор супермаркета, которому нужно улучшить показатели продаж

2. Задача - провести комплексный анализ данных по заказам с целью увеличения дохода, а именно:
- распределение объемов продаж по штатам, сегментам, категориям, подкатегориям в разрезе годов
- выявление лидирующих категорий и категорий с наиболее низкими показателями
- анализ эффективности каналов доставки
- сегментация клиентов (RFM – анализ)

3. Круг стейкхолдеров:
- директор продаж и маркетинга
- отдел логистики
- финансовый департамент

4. Бизнес-требования:
- на дашборде должны быть отражены продажи по штатам, сегментам, категориям, подкатегориям в разрезе годов
- отсутствие ошибок в данных

5. Гипотезы для проверки:
- более быстрые “Ship mode” (“first class” vs “standard”) увеличивает повторные покупки
- клиенты сегмента “Corporate” генерируют больший средний чек

6. Метрики для оценки:
- объем продаж по штатам, сегментам, категориям и подкатегориям: общий доход (выручка)
- темп роста продаж: процентное изменение объема продаж за период по сравнению с предыдущим
- доля каждой категории в общем объеме продаж
- время доставки: вычисляется как среднее время от оформления заказа до его получения клиентом
- RFM: для отдела маркетинга на основе RFM – метрик формируем группы клиентов для дальнейшего развития стратегии взаимодействия
    - Recency: время с последней покупки (чем меньше значение, тем активнее клиент)
    - Frequency: частота покупок за определенный период (это говорит о лояльности клиента)
    - Monetary: общая сумма покупок, сделанных клиентом (это помогает выделить наиболее ценных клиентов)

Описание данных

Ссылка на данные: Superstore Sales Dataset

Атрибут	Пример данных	Описание
Row ID	1	Номер строки
Order ID	CA-2017-152156	Номер заказа
Order Date	08/11/2017	Дата заказа
Ship Date	11/11/2017	Дата доставки
Ship Mode	Second Class	Тип доставки
Customer ID	CG-12520	Идентификатор клиента
Customer Name	Claire Gute	Имя клиента
Segment	Consumer	Сегмент
Country	United States	Страна
City	Henderson	Город
State	Kentucky	Штат
Postal Code	42420.0	Почтовый индекс
Region	South	Регион
Product ID	FUR-BO-10001798	Идентификатор продукта
Category	Furniture	Категория
Sub-Category	Bookcases	Подкатегория
Product Name	Bush Somerset Collection Bookcase	Название продукта
Sales	261.9600	Цена

Этапы анализа

Преобразование и очистка данных выполнена в visual studio code:
ссылка на ноутбук google colab

- преобразование данных:
  - исключен столбец ‘Postal Code’ и ‘Country’
  - приведение к единому naming convention
  - проверка данных на пустые значения и дубликаты
  - изменение типа данных
- сохранение витрины данных superstore_mart для анализа
- нормализация датафрейма для базы данных для хранения данных
- выгрузка данных в формате csv (так как нет возможности подключения к БД)
- создание ER-диаграммы (концептуальной, логической и физической моделей)
- импорт нормализованных данных и витрины данных в DBeaver

Анализ витрины данных в Tableau Public:

- загрузка витрины данных superstore_mart
- создание вычисляемых полей:

Вычисляемое поле	**Формула**
Рост год от года	(SUM([Sales]) - LOOKUP(SUM([Sales]), -1)) / ABS(LOOKUP(SUM([Sales]), -1))
Доля продаж	SUM([Sales]) / TOTAL(SUM([Sales]))
Количество заказов клиента	{fixed [Customer Id] : count([Order Id])}
Длительность доставки	DATEDIFF('day', [Order Date], [Ship Date])
Повторные покупки	if [Количество заказов клиента] > 1 then 1 else 0 end
Процент клиентов с повторной покупкой	sum([Повторные покупки]) / COUNTD([Customer Id])
Средний чек	sum([Sales]) / COUNTD([Order Id])
Sales для сортировки	if [Направление сортировки] = 'Максимум' then [Sales] else -[Sales] end
Темп роста_выбор даты	if [Темп роста дата] = 'Год' then DATETRUNC('year', [Order Date]) else DATETRUNC('quarter', [Order Date]) end
Топ 10 штатов + прочие	if [Топ 10 штатов по объему продаж] then [State] else 'Прочие' end
Названия штатов для тултипа	if [Топ 10 штатов + прочие] = 'Прочие' then 'Штаты вне топ 10' else 'Штат: ' + [State] end
Топ 10 городов + прочие	if [Топ 10 городов по объему продаж] then [City] else 'Прочие' end
Название городов для тултипа	if [Топ 10 городов + прочие] = 'Прочие' then 'Города вне топ 10' else 'Город: ' + [City] end
Название заголовка для темпа роста	if [Темп роста дата] = 'Год' then 'годам' else 'кварталам' end
Топ 10 подкатегорий + прочие	if [Топ 10 подкатегорий по объему продаж] then [Sub Category] else 'Прочие'

	end
Название подкатегорий для тултипа	if [Топ 10 подкатегорий + прочие] = 'Прочие' then 'Подкатегории вне топ 10' else 'Подкатегория: ' + [Sub Category] end
Recency	DATEDIFF('day', {FIXED [Customer Id]: MAX([Order Date])}, DATE('2019-01-01'))
Frequency (создано для удобства)	{FIXED [Customer Id]: COUNTD([Order Id])}
Monetary	{FIXED [Customer Id]: sum([Sales])}
R-score	if [Recency] <= { FIXED : PERCENTILE([Recency], 0.2)} then 5 elseif [Recency] <= {FIXED : PERCENTILE([Recency], 0.4)} then 4 elseif [Recency] <= { FIXED : PERCENTILE([Recency], 0.6)} then 3 elseif [Recency] <= { FIXED : PERCENTILE([Recency], 0.8)} then 2 else 1 end
F-score	if [Frequency] <= { FIXED : PERCENTILE([Frequency], 0.2)} then 1 elseif [Frequency] <= { FIXED : PERCENTILE([Frequency], 0.4)} then 2 elseif [Frequency] <= { FIXED : PERCENTILE([Frequency], 0.6)} then 3 elseif [Frequency] <= { FIXED : PERCENTILE([Frequency], 0.8)} then 4 else 5 end
M-score	if [Monetary] <= { FIXED : PERCENTILE([Monetary], 0.2)} then 1 elseif [Monetary] <= { FIXED : PERCENTILE([Monetary], 0.4)} then 2 elseif [Monetary] <= { FIXED : PERCENTILE([Monetary], 0.6)} then 3 elseif [Monetary] <= { FIXED : PERCENTILE([Monetary], 0.8)} then 4 else 5 end
RFM-score	int(STR([R-score]) + STR([F-score]) + STR([M-score]))
RFM-group	if [R-score] = 5 and [F-score] >= 4 and [M-score] >= 4 then 'чемпионы' elseif [R-score] >= 4 and [F-score] >= 3 then 'лояльные' elseif [R-score] = 5 and [F-score] = 1 then 'новые клиенты' elseif [R-score] <=2 and [F-score] >=4 then 'уходящие' elseif [R-score] <=2 and [F-score] <= 2 then 'потерянные' else 'прочие' end

- создание сетов:
  - Топ 10 штатов по объему продаж
  - Топ 10 подкатегорий по объему продаж
  - Топ 10 городов по объему продаж
- создание параметров:
  - Темп роста дата
  - Направление сортировки

Документация

Словарь данных датасета superstore_mart

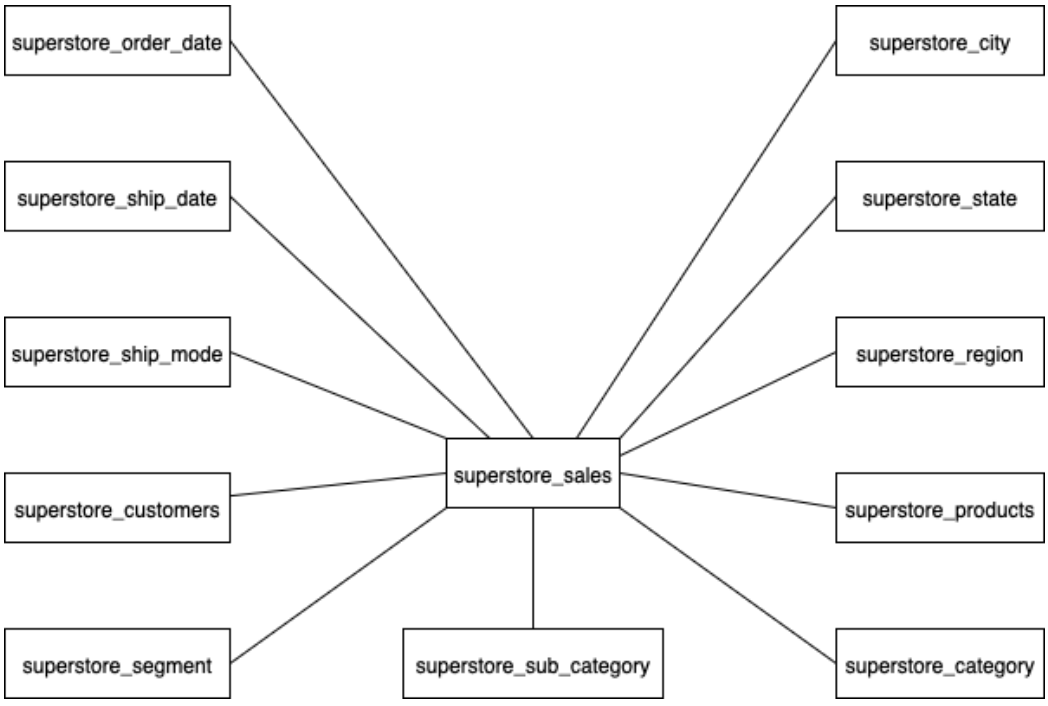
Название таблицы	Название атрибута	Тип данных	Бизнес определение	Пример данных
superstore_mart	row_id	object	Идентификатор строки	1
superstore_mart	order_id	object	Идентификатор заказа	CA-2017-152156
superstore_mart	order_date	datetime64[ns]	Дата заказа	2017-11-08
superstore_mart	ship_date	datetime64[ns]	Дата доставки	2017-11-11
superstore_mart	ship_mode	object	Способ доставки	Second Class
superstore_mart	customer_id	object	Идентификатор клиента	CG-12520
superstore_mart	customer_name	object	Имя клиента	Claire Gute
superstore_mart	segment	object	Сегмент клиента	Consumer
superstore_mart	city	object	Город	Henderson
superstore_mart	state	object	Штат	Kentucky
superstore_mart	region	object	Регион	South
superstore_mart	product_id	object	Идентификатор товара	FUR-BO-10001798
superstore_mart	category	object	Категория	Furniture
superstore_mart	sub_category	object	Подкатегория	Bookcases
superstore_mart	product_name	object	Наименование товара	Bush Somerset Collection Bookcase
superstore_mart	sales	float64	Сумма заказа	261.9600

ER-диаграммы

Сущности	Отношения
superstore_sales: row_id, order_date_id, ship_date_id, ship_mode_id, customer_id, segment_id, city_id, state_id, region_id, category_id, sub_category_id, product_id, sales (9800, 14)	
superstore_order_date: order_date_id, order_date (1230, 2)	superstore_order_date – superstore_sales
superstore_ship_date: ship_date_id, ship_date (1326, 2)	superstore_ship_date - superstore_sales
superstore_ship_mode: ship_mode_id, ship_mode (4, 2)	superstore_ship_mode - superstore_sales

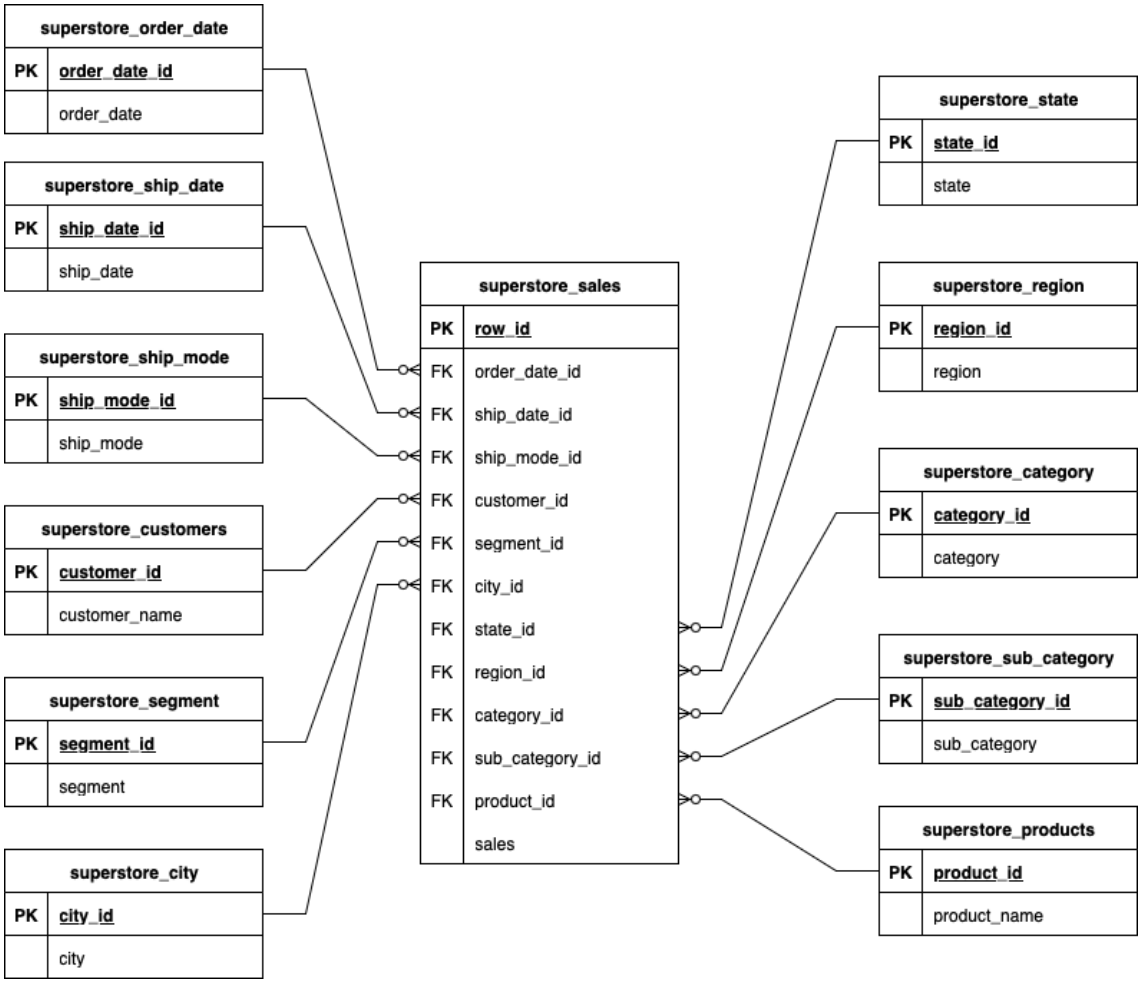
superstore_customers: customer_id, customer_name (793, 2)	superstore_customers – superstore_sales
superstore_segment: segment_id, segment (3, 2)	superstore_segment - superstore_sales
superstore_city: city_id, city (529, 2)	superstore_city - superstore_sales
superstore_state: state_id, state (49, 2)	superstore_state - superstore_sales
superstore_region: region_id, region (4, 2)	superstore_region - superstore_sales
superstore_category: category_id, category (3, 2)	superstore_category - superstore_sales
superstore_sub_category: sub_category_id, sub_category (17, 2)	superstore_sub_category - superstore_sales
superstore_products: product_id, product_name (1893, 2)	superstore_products – superstore_sales

Концептуальная модель:

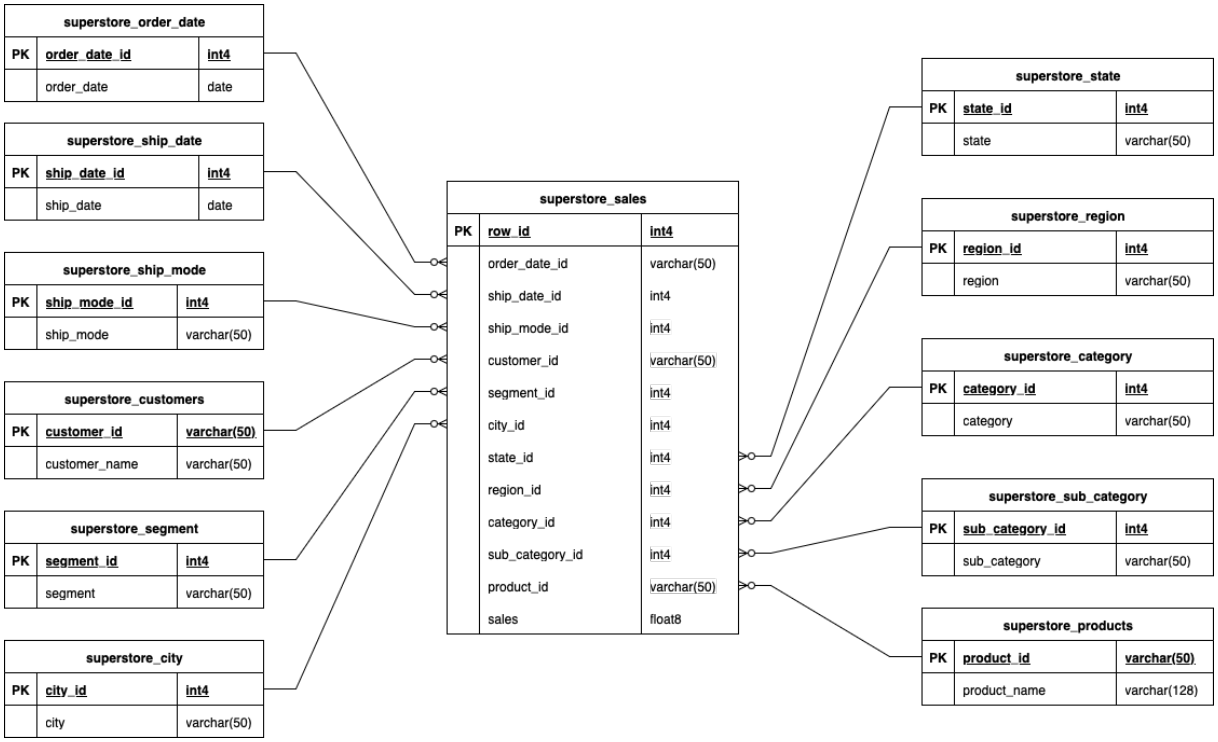


Логическая модель:

связь один ко многим



Физическая модель:

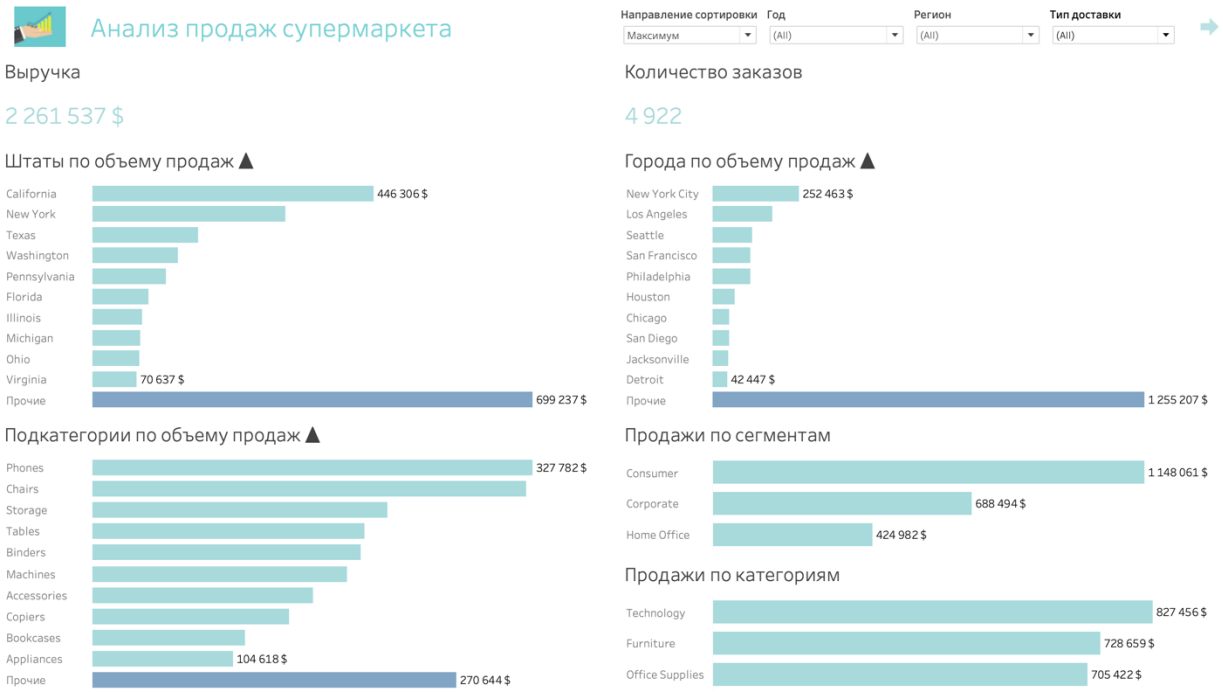


Результаты

Создано три дашборда, в которых есть следующие элементы анализа:

1. **Анализ продаж супермаркета**

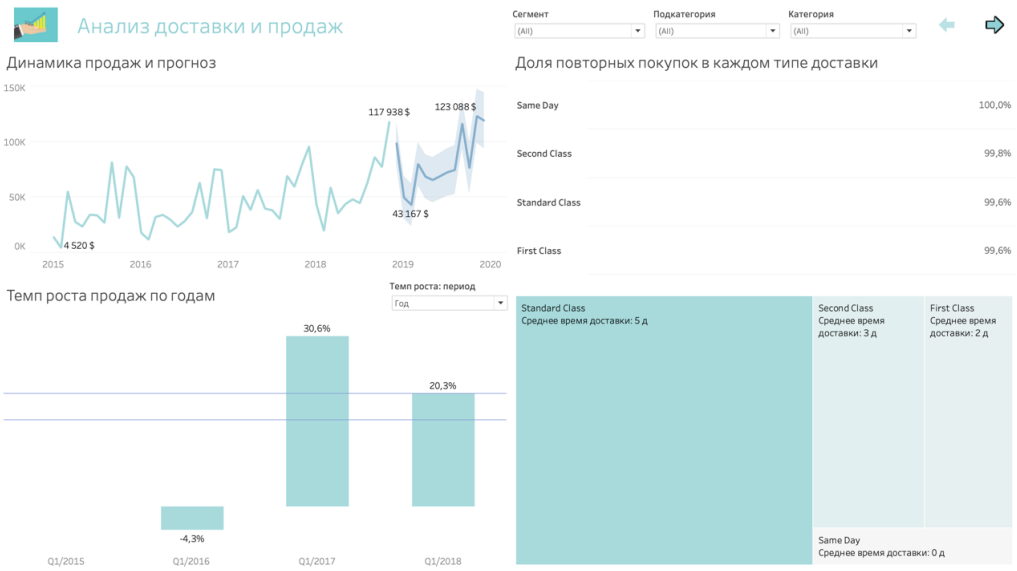
    - ключевые показатели:
        - выручка
        - количество заказов
    - диаграммы:
        - штаты по объему продаж
        - города по объему продаж
        - подкатегории по объему продаж
        - продажи по сегментам
        - продажи по категориям
    - фильтры:
        - по направлению сортировки
        - по годам
        - по регионам
        - по типу доставки
    - тултипы:
        - объем продаж
        - доля продаж
        - средний чек
    - кнопка вперед – переход на дашборд «Анализ по типу доставки и динамика продаж»



- взаимодействие:
  - все элементы в диаграммах могут быть фильтрами для других диаграмм на листе
  - все фильтры на листе влияют на все диаграммы на листе
  - фильтр направление сортировки
    - максимум – сортировка от большего к меньшему значению суммы продаж
    - минимум – сортировка от меньшего к большему значению суммы продаж

2. **Анализ по типу доставки и динамика продаж**

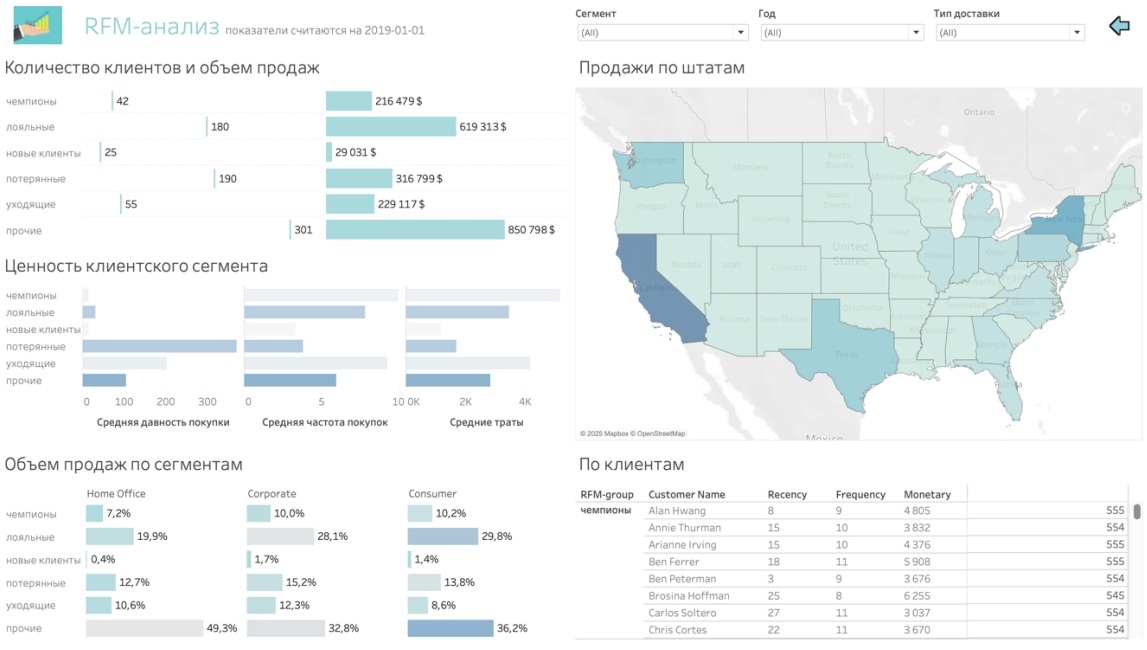
- диаграммы:
  - динамика продаж и прогноз
  - доля повторных покупок в каждом типе доставки
  - темп роста продаж по годам и кварталам
  - количество и среднее время доставки по типу доставки
- фильтры:
  - по сегменту
  - по подкатегориям
  - по категориям
- тултипы:
  - индикатор прогноза
  - дата
  - объем продаж
  - средний чек
  - тип доставки
  - среднее время доставки
  - количество доставок
  - доля продаж
- кнопка темп роста: период – меняет период диаграммы «темп роста продаж по годам и кварталам» на год или квартал
- кнопка назад – возврат к дашборду «Анализ продаж супермаркета»
- кнопка вперед – переход к дашборду «RFM-анализ»



- взаимодействие:
  - диаграмма «динамика продаж и прогноз» может быть фильтром для диаграммы «количество и среднее время доставки по типу доставки»
  - диаграмма «доля повторных покупок в каждом типе доставки» и диаграмма «количество и среднее время доставки по типу доставки» может быть фильтром для всех диаграмм на листе
  - все фильтры на листе влияют на все диаграммы на листе

3. RFM-анализ

- диаграммы
  - количество клиентов и объем продаж
  - ценность клиентского сегмента
  - объем продаж по сегментам
  - продажи по штатам
  - RFM-анализ по клиентам
- фильтры:
  - по годам
  - по типу доставки
  - по сегменту
- тултипы:
  - RFM-группа
  - количество клиентов
  - объем продаж
  - средняя давность покупки
  - средняя частота покупок
  - средние траты
  - количество клиентов
  - сегмент
  - доля продаж
- кнопка назад – возврат к дашборду «Анализ по типу доставки и динамика продаж»



    - взаимодействие:
        - диаграмма «количество клиентов и объем продаж» и диаграмма «продажи по штатам» может быть фильтром для всех диаграмм
        - все фильтры на листе влияют на все диаграммы на листе

Выводы

- Общая выручка за период анализа – более 2 млн. долларов
- Общее количество заказов за период анализа – около 5 тысяч
- Лидеры по объему продаж и количеству заказов:
    - штат – Калифорния (446 тысяч долларов, 1 тысяча заказов)
    - город – Нью-Йорк (252 тысячи долларов, 439 заказов)
    - сегмент – обычный потребитель (consumer)
    - категория – технологии (technology)
    - подкатегория – телефоны (phones)
- Имеющие наиболее низкие показатели по объему продаж и количеству заказов:
    - штат – Северная Дакота
    - город – Миссури
    - сегмент – домашний офис (home office)
    - категория – офисные принадлежности (office supplies)
    - подкатегория – крепежи (fasteners)
- Средний чек сегмента «Corporate» - 467,77 $, гипотеза неверная
- Заметна положительная динамика роста продаж
- Клиенты чаще всего пользуются стандартной доставкой, которая занимает в среднем 5 дней
- Доля повторных покупок в каждом типе доставок практически одинаковая – гипотеза неверная
- Заметен резкий скачок объема продаж в 2017 году (на 30,6%)
- Заметно снижение объема продаж в первом квартале каждого года
- RFM-анализ (исключая группу прочие клиенты):
    - по объему продаж лидирует группа лояльных
    - по количеству клиентов лидирует группа потерянных
    - количество новых клиентов год от года уменьшается

Рекомендации

- Предполагаем, что маркетинговые усилия по удержанию активности и повышению лояльности в «лидерах» имеют высокий потенциал, поэтому:
    - есть смысл запускать программы лояльности и персональные предложения для этих рынков
    - а также рассмотреть возможность увеличения складов
    - и оптимизировать логистику в лидирующих штатах
- Для наиболее прибыльных категорий и подкатегорий:
    - стоит развивать ассортимент
    - добавить эксклюзивные предложения
    - а также развивать кросс-продажи сопутствующих товаров
- Для наименее прибыльных категорий и подкатегорий:
    - следует проанализировать причины (сезонность, конкуренция, цена)
    - и оптимизировать ассортимент

- Значительный рост продаж в 2017 году:
    - изучить причины скачка продаж
    - повторить эффективные акции и промо
- Среднее время стандартной доставки (5 дней) – можно улучшить клиентский опыт:
    - проанализировать предложения конкурентов
    - оптимизировать логистику
    - запустить спец акции по ускоренной доставке в пиковые периоды
- Повторные покупки не зависят от типа доставки:
    - имеет смысл проработать другие триггеры повторных продаж
- Снижение продаж в первые кварталы могут быть связаны с сезонностью/праздниками:
    - проанализировать причины спада
    - запустить спец акции в «низкий» сезон
- RFM-анализ:
    - нужно разработать методы возвращения потерянных клиентов (рассылки, спецпредложения)
    - сокращение клиентской базы: нужно усилить рекламу для новых клиентов, и разработать реферальные программы

Для роста продаж и повышения лояльности нужно сфокусироваться на удержании и развитии лидирующих рынков и прибыльных категорий, оптимизировать ассортимент и логистику, а также внедрить программу лояльности и персональные предложения. Следует проанализировать причины роста и спада, адаптировать спецакции к сезонности.