ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ «ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет “Высшая школа бизнеса”

Прикладной анализ данных

Отчет по проекту

Бизнес-дашборд для ритейла

Выполнили:

Васенёва Валерия, БАСБ251

Гаджибегова Кристина, БАСБ252

Лашхия Софья, БАСБ251

Рощина Надежда, БАСБ252

Москва 2025

Ссылки

- общий репозиторий: github
- датасет: kaggle
- обработка датасета: google collab
- дашборд: ссылка в общем репозитории

Введение

Целью данного проекта является разработка бизнес-дашборда для анализа продаж в розничной торговле и формулировка аналитических гипотез на основе данных. Работа выполнена в рамках курса по прикладному анализу данных и направлена на демонстрацию полного цикла аналитического проекта - от первичной обработки данных до визуализации и проверки гипотез.

Основой исследования послужил датасет *Superstore Sales Dataset*, содержащий информацию о продажах товаров в различных штатах США за период 2015–2018 гг. Для решения поставленной задачи использовалась комбинированная методология, включающая статистический анализ и визуальную аналитику.

Процесс работы строился по следующей схеме: Данные - Анализ - Визуализация - Проверка гипотез - Выводы.

На первом этапе был проведен первичный анализ данных и исследовательский анализ (EDA) с использованием Python, что позволило выявить ключевые закономерности и выдвинуть гипотезы (1–6). На втором этапе был разработан интерактивный дашборд в Power BI, который объединил основные бизнес-показатели и стал инструментом визуальной проверки дополнительной гипотезы (7), а также средством дальнейшего анализа.

Описание датасета

В рамках проекта использован датасет *Superstore Sales Dataset*, содержащий детализированные данные о продажах товаров в супермаркете. Датасет включает 9800 строк и 18 столбцов, охватывающих период с 2015 по 2018 год. Каждая строка соответствует отдельному товару в составе заказа, поэтому один заказ может содержать несколько записей.

Данные охватывают продажи в различных штатах США и включают три основные категории товаров - *Furniture*, *Office Supplies* и *Technology*. Ключевые поля датасета содержат сведения о заказах (идентификаторы, даты оформления и доставки, способ доставки), о клиентах (идентификаторы, имена, сегмент, местоположение), а также о

самих товарах (идентификатор, категория, подкатегория, наименование) и финансовые показатели (сумма продажи в долларах).

Основные числовые характеристики датасета следующие:

- Средний чек составляет $230.77 при стандартном отклонении $626.65, что указывает на высокую вариативность покупок.
- Минимальная сумма продажи - $0.44, максимальная - $22,638.48.
- 75% всех продаж не превышают $210.61.
- Всего зафиксировано 4922 уникальных заказа, 793 клиента и 1861 товар.
- Общий объем продаж за весь период - $2,261,536.78.
- Среднее время доставки - 4 дня.

Первичный анализ данных (EDA)

Продажи демонстрируют устойчивый рост на протяжении исследуемого периода. После небольшой стагнации в 2015–2016 годах (снижение на 4,26%) наблюдается резкий рост в 2017 году (+30,6%) и стабильное укрепление результатов в 2018 году (+20,3%). Общий прирост продаж за 4 года составил 50,5%.

Наиболее активным регионом стал Запад (31,4% всех продаж), за ним следуют Восток (29,6%), Центральный (21,8%) и Юг (17,2%). При этом средний чек на Юге оказался самым высоким ($243), несмотря на наименьший объем продаж. Лидирующими штатами являются Калифорния ($446 тыс.), Нью-Йорк ($306 тыс.) и Техас ($168 тыс.), а среди городов выделяются Нью-Йорк, Лос-Анджелес, Сиэтл и Детройт.

Клиентская база стабильно растет: количество уникальных клиентов увеличилось с 589 до 690 за четыре года. На одного клиента приходится в среднем 6 заказов за весь период, то есть около 1,5 заказа в год. Ключевые клиенты демонстрируют разное поведение - от высокой активности до крупных, но редких покупок.

По сегментам бизнеса распределение выглядит следующим образом:

- *Consumer* - 50,8% продаж, средний чек $225;
- *Corporate* - 30,4%, средний чек $233;
- *Home Office* - 18,8%, но наибольший средний чек ($243).

Временной анализ показывает ярко выраженную сезонность. Основные пики продаж приходятся на ноябрь и сентябрь, при этом в ноябре фиксируется максимальный объем ($117 938). В 2018 году почти половина всех продаж (45%) пришлась на последние три месяца года.

Дни недели также демонстрируют различия в активности. Наибольшие продажи приходятся на вторник и субботу, тогда как четверг показал неожиданно низкие результаты. Этот паттерн требует дополнительного анализа.

По видам доставки распределение следующее:

- *Standard Class* - основной способ (59,3% заказов), средний чек $228.85, среднее время доставки 5,0 дней.
- *Second Class* - 19,9%, средний чек $236.55, доставка за 3,25 дня.
- *First Class* - 15,3%, доставка за 2,18 дня.
- *Same Day* - 5,5%, средний чек $232.75, доставка почти мгновенная (0,04 дня).

Таким образом, бизнес демонстрирует здоровую структуру: массовый сегмент с экономичной доставкой обеспечивает устойчивость, а премиальные классы - скорость и лояльность клиентов.

Средняя корзина клиента включает 2 товара на сумму около $460. Максимальная сумма заказа составила $23,661.23, а наибольшее разнообразие - 14 товаров в одном заказе. Это указывает на преобладание рационального потребления и умеренных покупок.

Наибольший вклад в выручку вносит категория *Technology* ($827,456), за ней следуют *Office Supplies* и *Furniture*. Региональным лидером является Запад ($710,220), а среди сегментов - *Consumer*

Формулировка гипотез

На основе проведённого исследовательского анализа данных были сформулированы ключевые аналитические гипотезы, направленные на выявление закономерностей продаж, особенностей поведения клиентов и взаимодействия товаров. Гипотезы можно разделить на две группы: те, которые проверялись средствами Python в рамках первичного анализа данных (гипотезы 1–6), и гипотезу, проверку которой обеспечивал интерактивный дашборд в Power BI (гипотеза 7).

1. 20% клиентов приносят 80% выручки (принцип Парето).

2. Категория *Technology* характеризуется наибольшей сезонностью.

3. Понедельник является днем с наибольшим средним чеком.

4. Дорогие заказы доставляются быстрее.

5. Клиенты из малых городов более лояльны, чем клиенты из крупных.

6. Товары из категории *Office Supplies* часто покупаются совместно с товарами *Technology*.
7. Пик продаж в году приходится на сентябрь, поскольку школы и офисы совершают сезонные закупки перед новым учебным и деловым годом.

Принцип проверки гипотез:

- Гипотезы 1–6 были проверены через статистический и дескриптивный анализ данных в Python (с применением t-тестов, коэффициентов сезонности и анализа корзин товаров).
- Гипотеза 7 проверялась с помощью интерактивного дашборда в Power BI, что позволило визуально оценить сезонные пики продаж и выявить закономерности в динамике по месяцам.

Таким образом, сформулированные гипотезы служат основой для дальнейшего аналитического исследования и проверки бизнес-идей.

Проверка гипотез с помощью ручного анализа

Для проверки первых шести гипотез был проведен статистический и дескриптивный анализ данных с использованием Python. Каждая гипотеза рассматривалась отдельно с применением соответствующих методов анализа: расчет средних значений, коэффициентов сезонности, анализа корзин товаров и t-тестов для проверки значимости различий.

Гипотеза 1. 20% клиентов приносят 80% выручки

Анализ показал, что 20% клиентов формируют лишь 48,5% выручки. Топ-10 клиентов обеспечивают 6,8% дохода, а топ-50 - 22,6%

Вывод: гипотеза опровергнута - концентрация выручки ниже ожидаемой, клиентская база более равномерная.

Гипотеза 2. Категория Technology имеет высокую сезонность

Коэффициенты сезонности: Furniture - 0,598; Office Supplies - 0,573; Technology - 0,601

Вывод: гипотеза подтверждена - категория *Technology* демонстрирует наибольшие колебания по сезонам.

Гипотеза 3. Понедельник - день самых выгодных покупок

Средние чеки по дням недели варьируются от $218.95 (понедельник) до $264.03 (четверг)

Вывод: гипотеза опровергнута - наиболее выгодные покупки совершаются по четвергам.

Гипотеза 4. Дорогие заказы доставляются быстрее

Среднее время доставки по квартилям стоимости варьируется от 4,0 до 3,9 дней. T-тест показал p-value=0.3252, различие статистически незначимо

Вывод: гипотеза опровергнута - стоимость заказа не влияет на скорость доставки

Гипотеза 5. Клиенты из малых городов более лояльны

Среднее количество заказов на клиента: малые города - 1.00, крупные - 1.02.
T-тест показал значимую разницу ($p<0.001$), но в противоположную сторону

Вывод: гипотеза опровергнута - клиенты из крупных городов демонстрируют немного большую лояльность

Гипотеза 6. Товары Office Supplies часто покупаются вместе с Technology

Наиболее частые комбинации категорий в корзине:

- Furniture + Office Supplies - 944 раз,
- Office Supplies + Technology - 901 раз,
- Furniture + Technology - 473 раза

Вывод: гипотеза частично подтверждена - связь между Office Supplies и Technology действительно сильная, хотя Furniture + Office Supplies встречается чаще.

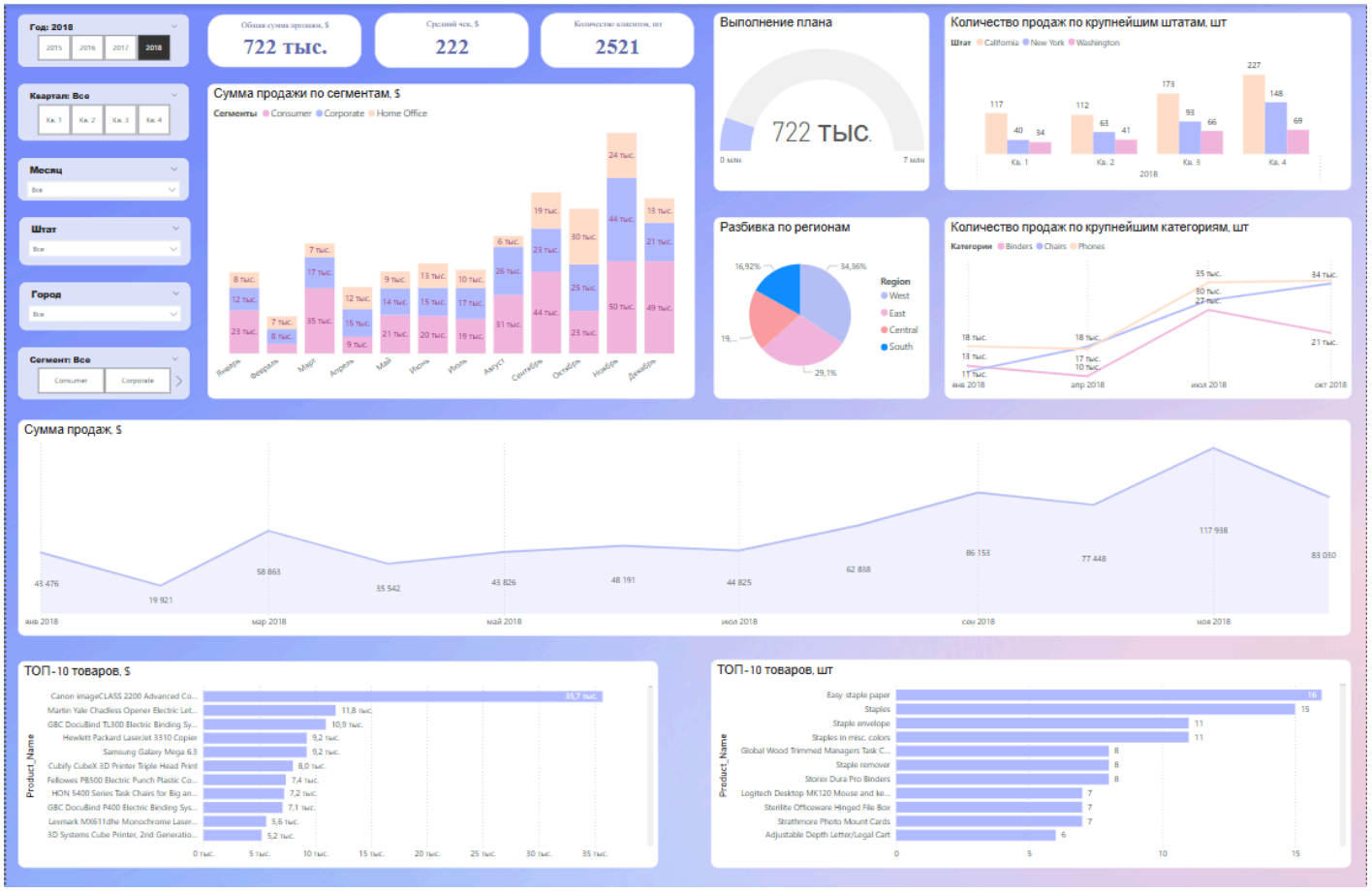
Анализ позволил выявить реальные закономерности в поведении клиентов и структуре продаж, а также открыл направления для более детального изучения сезонных и товарных паттернов на этапе построения дашборда.

Описание дашборда

Интерактивный дашборд разработан в Power BI и предназначен для комплексного анализа продаж по ключевым бизнес-показателям. Он объединяет визуализацию динамики продаж, клиентских характеристик и товарной структуры, а также позволяет гибко фильтровать данные по различным параметрам.

В левой части дашборда расположены панели фильтров, которые позволяют выбрать интересующий период (год, квартал, месяц), а также ограничить анализ по географическим признакам (штат, город) и клиентскому сегменту. Это обеспечивает возможность детализированного анализа и сравнения показателей между различными регионами и категориями клиентов.

В верхней части дашборда представлены ключевые метрики: общая сумма продаж, средний чек, количество клиентов и прогресс по выполнению плана. В качестве цели была выбрана сумма продаж в 7 миллионов долларов за весь период продаж. Эти показатели дают общее представление о текущем состоянии бизнеса и позволяют быстро оценить эффективность продаж.



Центральная область визуализации включает круговую диаграмму распределения продаж по регионам и столбчатую диаграмму динамики продаж по клиентским

сегментам в разрезе месяцев. Эти графики позволяют выявлять региональные различия и сезонные тренды в поведении разных категорий покупателей.

Правая часть дашборда содержит две ключевые визуализации: столбчатую диаграмму количества продаж по крупнейшим штатам и линейный график динамики продаж по основным товарным категориям. Они помогают определить, какие регионы и категории товаров вносят наибольший вклад в выручку.

Нижняя часть дашборда отведена под временной анализ - здесь расположен основной график динамики продаж по периодам (годам, кварталам или месяцам в зависимости от выбора фильтров). В самом низу размещены таблицы с топ-10 товаров по объему продаж и по количеству проданных единиц, что позволяет оперативно оценивать ассортимент и эффективность конкретных позиций.

Использование дашборда для проверки гипотез

Интерактивный дашборд в Power BI позволяет не только визуализировать общие показатели продаж, но и служит инструментом быстрой проверки аналитических гипотез, сформулированных в ходе исследования. Благодаря гибкой системе фильтров можно оперативно анализировать данные в различных разрезах - по времени, регионам, сегментам клиентов и категориям товаров.

Гипотезы о сезонности, региональных различиях и поведении клиентов можно проверять визуально через дашборд.

- Сезонные колебания продаж категории *Technology* подтверждаются линейным графиком динамики продаж по месяцам.
- Региональные и сегментные различия можно проанализировать через круговую диаграмму распределения по регионам и столбчатую диаграмму продаж по сегментам.
- Средний чек и общие суммы продаж по временным периодам, сегментам товаров или клиентов можно анализировать с помощью верхней панели с ключевыми метриками.
- Графики в нижней части дашборда позволяют визуально оценивать наиболее прибыльные категории товаров и общие тенденции продаж по времени. Раздел с топ-10 товаров помогает верифицировать гипотезы о наиболее популярных категориях и сочетаниях продуктов.

Гипотеза 7. Пик продаж в году приходится на сентябрь

Для проверки гипотезы о сезонном пике продаж был проведен анализ месячной динамики продаж за все четыре года с помощью дашборда.
Результаты показали, что наибольшие объемы продаж действительно приходятся на сентябрь (подготовка к новому учебному году) и ноябрь (подготовка к новому календарному году). При этом сентябрь занимает уверенное второе место по продажам и характеризуется стабильным ростом по сравнению с летними месяцами.

Вывод: гипотеза частично подтверждена - сентябрь действительно является одним из пиковых месяцев продаж, однако наибольший пик наблюдается в ноябре, что связано с ростом потребительской активности в предновогодний период.

Заключение

В рамках проекта была выполнена полная аналитическая работа по продажам супермаркета с использованием Python и Power BI. Проведенный исследовательский анализ данных (EDA) позволил выявить ключевые закономерности в поведении клиентов, структуре продаж, сезонности и региональных различиях.

Были сформулированы и проверены семь гипотез:

- Первичные гипотезы (1–6), проверенные средствами Python, позволили уточнить реальную концентрацию выручки среди клиентов, выявить особенности сезонности категорий товаров, оценить влияние стоимости заказа на скорость доставки, а также проверить закономерности покупок по дням недели и взаимосвязь товаров в корзине. Из шести гипотез две подтвердились частично или полностью (2 и 6), остальные опровергнуты.
- Дополнительная гипотеза (7) была проверена с помощью интерактивного дашборда. Анализ показал, что сентябрь является одним из пиковых месяцев продаж, однако наибольший пик приходится на ноябрь.

Разработанный дашборд объединяет ключевые метрики, визуализации и фильтры, что позволяет интерактивно исследовать данные, подтверждать или опровергать гипотезы и поддерживать принятие управленческих решений. Он обеспечивает наглядность, детализацию и гибкость анализа, объединяя результаты статистического исследования и визуальные инструменты.

Таким образом, проект продемонстрировал полный цикл аналитической работы: от первичного анализа данных и формулировки гипотез до визуализации и проверки результатов. Полученные выводы позволяют бизнесу лучше понимать структуру продаж, поведение клиентов и сезонные тренды, а дашборд служит эффективным инструментом для дальнейшего мониторинга и анализа.