

Final Project

Анализ и прогнозирование продаж в розничной торговле

Обзор проекта

В этом проекте вы будете работать с историческими данными о розничных продажах, чтобы понять, какие факторы больше всего влияют на прибыль компании. Данные охватывают клиентов, заказы, товары и детали каждой продажи - это позволит взглянуть на бизнес с разных сторон. Вы будете исследовать, как скидки, количество товаров и категории влияют на общую выручку и прибыль, и какие сегменты клиентов приносят наибольший доход.

Проект охватывает полный цикл обработки данных: от загрузки и нормализации до анализа в SQL, продвинутого анализа и визуализаций в Power BI и построения моделей прогнозирования с помощью Python.

Критерии оценивания

1. SQL - 25 баллов

- Импорт и подготовка данных (5 баллов)
- Выполнение SQL-запросов (15 баллов)
- Объединение таблиц и экспорт в CSV (5 баллов)

2. Power BI - 35 баллов

- Загрузка данных и настройка модели (5 баллов)
- Ответы на бизнес-вопросы, визуализация (20 баллов)
- Функциональность отчёта (фильтры, drill-through) (10 баллов)

3. Python - 25 баллов

- Предобработка данных и EDA (10 баллов)
- Построение и интерпретация модели линейной регрессии (15 баллов)

4. Презентация - 15 баллов

- Качество визуализаций и выводов (5 баллов)
- Структура и логика презентации (5 баллов)
- Ясность в объяснении бизнес-проблемы и инсайтов (5 баллов)

Структура данных

Данные разбиты на 4 логически связанные таблицы:

1. **customers** — информация о клиентах: ID, имя, сегмент, регион, город, почтовый код.
2. **products** — товары: ID, категория, подкатегория, название.
3. **orders** — заказы: ID заказа, дата заказа, дата доставки, способ доставки, ID клиента.
4. **sales** — детализация продаж: ID заказа, ID товара, количество, скидка, прибыль, сумма продажи.

Ваши задачи

1. SQL

Создание таблиц в SQL

- Импортируйте все четыре таблицы в базу данных;
- Проверьте типы данных в таблицах;
- Проверьте все таблицы на наличие лишних колонок и дубликатов.

```
--Check duplicates--
```

```
SELECT
    Order_ID,
    Product_ID,
    COUNT(*) AS duplicate_count
FROM
    `da-nfactorial.student_Zhansaya_Sovetbek.sales`
GROUP BY
    1,2
HAVING
    COUNT(*) > 1;
```

```
--Check Why Duplicates--
```

```
Select * from `da-nfactorial.student_Zhansaya_Sovetbek.sales`
where Order_ID='CA-2016-137043';
```

```
--Remove Duplicates--
```

```
CREATE OR REPLACE TABLE `da-nfactorial.student_Zhansaya_Sovetbek.products` AS
SELECT DISTINCT *
FROM `da-nfactorial.student_Zhansaya_Sovetbek.products`;
```

Сначала я проверила таблицу на наличие дубликатов. Затем, с помощью поиска повторяющихся значений ID, я определила строки, которые полностью совпадают по всем колонкам. Если все поля были идентичны, я считала такие строки дубликатами и удалила их из таблицы.

SQL задачи:

- Выведите топ-5 товаров с наибольшей выручкой;

```
SELECT
    products.Product_Name,
    ROUND(SUM(sales.Sales), 2) AS Total_Revenue
FROM
    `da-nfactorial.student_Zhansaya_Sovetbek.sales` AS sales
JOIN
    `da-nfactorial.student_Zhansaya_Sovetbek.products` AS products
ON
    sales.Product_ID = products.Product_ID
GROUP BY
    1
ORDER BY
    Total_Revenue DESC
LIMIT 5;
```

Row	Product_Name	Total_Revenue
1	HON 5400 Series Task Chairs f...	21870.58
2	Riverside Palais Royal Lawyers Bookcase, Royale Cherry Finish	15610.97
3	Bretford Rectangular Conferenc...	12995.29
4	Global Troy Executive Leather L...	12975.38
5	Sauder Forest Hills Library, Woo...	12921.64

- Рассчитайте средний размер скидки в каждом регионе;

```
SELECT
    customers.Region,
    ROUND(AVG(sales.Discount), 4) AS Avg_Discount
FROM
    `da-nfactorial.student_Zhansaya_Sovetbek.sales` AS sales
JOIN
    `da-nfactorial.student_Zhansaya_Sovetbek.orders` AS orders
```

```

ON sales.Order_ID = orders.Order_ID
JOIN
`da-nfactorial.student_Zhansaya_Sovetbek.customers` AS customers
ON orders.Customer_ID = customers.Customer_ID
GROUP BY
customers.Region
ORDER BY
Avg_Discount DESC;

```

Row	Region	Avg_Discount
1	Central	0.2084
2	East	0.169
3	South	0.1633
4	West	0.1574

- Найдите самых лояльных клиентов за весь период;

```

SELECT
customers.Customer_ID,
customers.Customer_Name,
COUNT(DISTINCT orders.Order_Date) AS Purchase_Days
FROM
`da-nfactorial.student_Zhansaya_Sovetbek.orders` AS orders
JOIN
`da-nfactorial.student_Zhansaya_Sovetbek.customers` AS customers
ON
orders.Customer_ID = customers.Customer_ID
GROUP BY
1,2
ORDER BY
Purchase_Days DESC
LIMIT 5;

```

Row	Customer_ID	Customer_Name	Purchase_Days
1	SV-20365	Seth Vernon	9
2	JE-15745	Joel Eaton	8
3	ZC-21910	Zuschuss Carroll	8
4	LC-16885	Lena Creighton	7
5	LA-16780	Laura Armstrong	7

- Сравните общую прибыль между категориями товаров;

```
SELECT
    products.Sub_Category AS Product_Category,
    ROUND(SUM(sales.Profit), 2) AS Total_Profit
FROM
    `da-nfactorial.student_Zhansaya_Sovetbek.sales` AS sales
JOIN
    `da-nfactorial.student_Zhansaya_Sovetbek.products` AS products
ON
    sales.Product_ID = products.Product_ID
GROUP BY
    1
ORDER BY
    Total_Profit DESC;
```

Row	Product_Category ▾	Total_Profit ▾
1	Chairs	26719.71
2	Furnishings	14569.59
3	Bookcases	-3452.87
4	Tables	-17725.48

- Определите, какая доля продаж была совершена со скидкой, от общего объема продаж;

```
SELECT
    ROUND(SUM(CASE WHEN Discount > 0 THEN Sales ELSE 0 END) / SUM(Sales), 4) AS
Discount_Share
FROM
    `da-nfactorial.student_Zhansaya_Sovetbek.sales`;
```

Row	Discount_Share ▾
1	0.6548

- Объедините все четыре таблицы в единую таблицу заказов с деталями и экспортируйте результат в csv.

```
SELECT
    orders.Order_ID,
    orders.Order_Date,
    orders.Ship_Date,
    orders.Ship_Mode,
    customers.Customer_ID,
    customers.Customer_Name,
    customers.Segment,
```

```

customers.Country,
customers.City,
customers.State,
customers.Postal_Code,
customers.Region,
sales.Product_ID,
products.Product_Name,
products.Category,
products.Sub_Category,
sales.Sales,
sales.Quantity,
sales.Discount,
sales.Profit
FROM
`da-nfactorial.student_Zhansaya_Sovetbek.sales` AS sales
JOIN
`da-nfactorial.student_Zhansaya_Sovetbek.orders` AS orders
ON
sales.Order_ID = orders.Order_ID
JOIN
`da-nfactorial.student_Zhansaya_Sovetbek.customers` AS customers
ON
orders.Customer_ID = customers.Customer_ID
JOIN
`da-nfactorial.student_Zhansaya_Sovetbek.products` AS products
ON
sales.Product_ID = products.Product_ID;

```

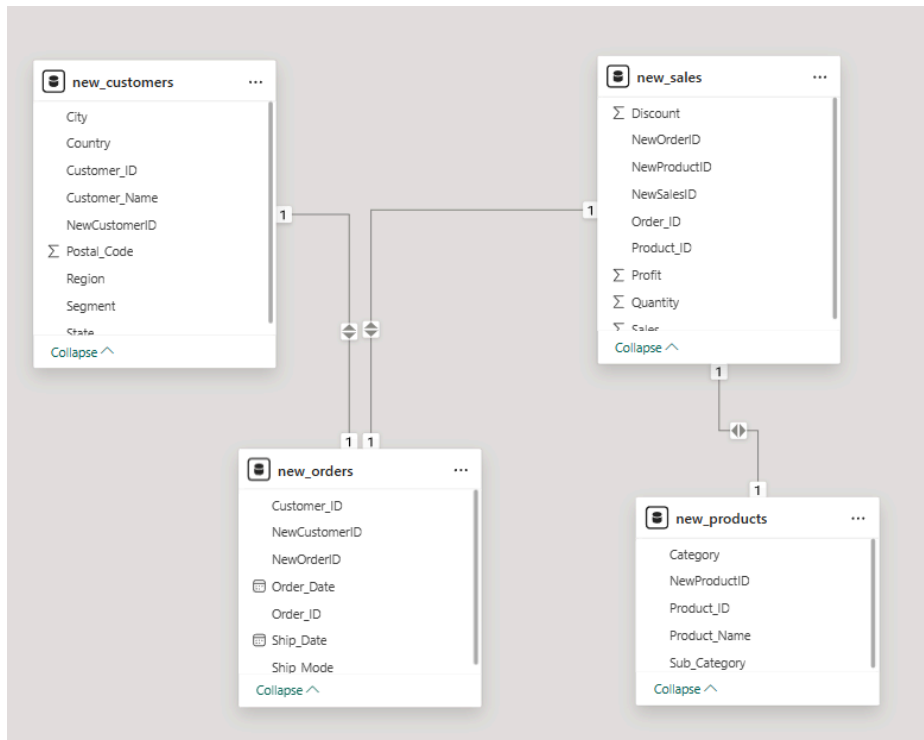
2. PowerBI

Загрузка данных

- Импортируйте 4 таблицы customers, products, orders, sales;
- Проверьте и настройте связи между таблицами;
 Когда я хотела соединить таблицы, то у меня вышла связь many-to-many из-за дубликатов. Из-за этого я создала уникальный ключ в Google Big Query, потому что в Power BI было сложно этого сделать. Моя идея такая если ID повторяется и но там разные данные в других колонках то я добавила дополнительные цифры для этого ID.

'Customer_ID'-1

'Customer_ID'-2



- Проверьте, чтобы типы полей были корректны.

Бизнес-вопросы

Ответьте как минимум на 7 вопросов, используя визуализации и меры в Power BI:

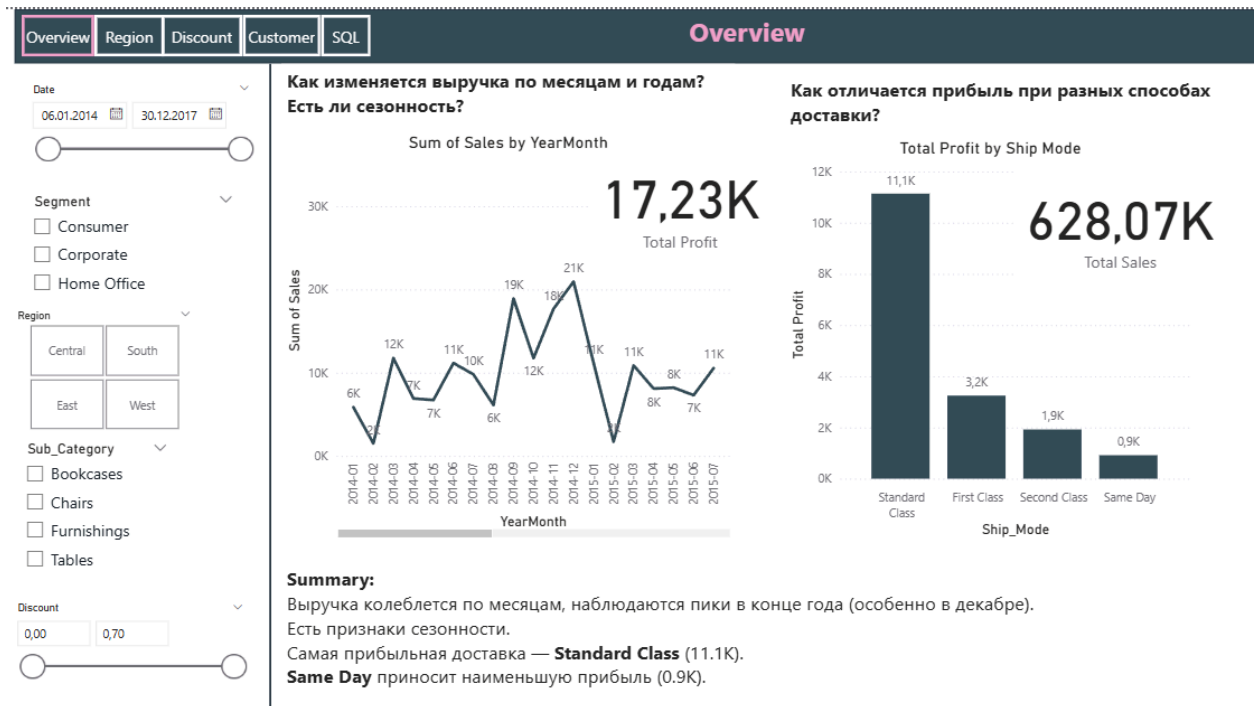
1. Какие категории товаров приносят наибольшую прибыль в разных регионах?
2. Есть ли зависимость между размером скидки и прибылью?
3. Как изменяется выручка по месяцам и годам? Есть ли сезонность?
4. Какие города приносят наибольшую выручку?
5. В каких подкатегориях товаров чаще всего применяются скидки?
6. Есть ли товары, которые часто продаются в убыток? -
7. Как отличается прибыль при разных способах доставки?
8. Что происходит с прибылью, если фильтровать по регионам/категориям/сегментам? -
9. Каков средний размер заказа в зависимости от региона или сегмента клиента?
10. Сколько заказов приходится на одного клиента в среднем?

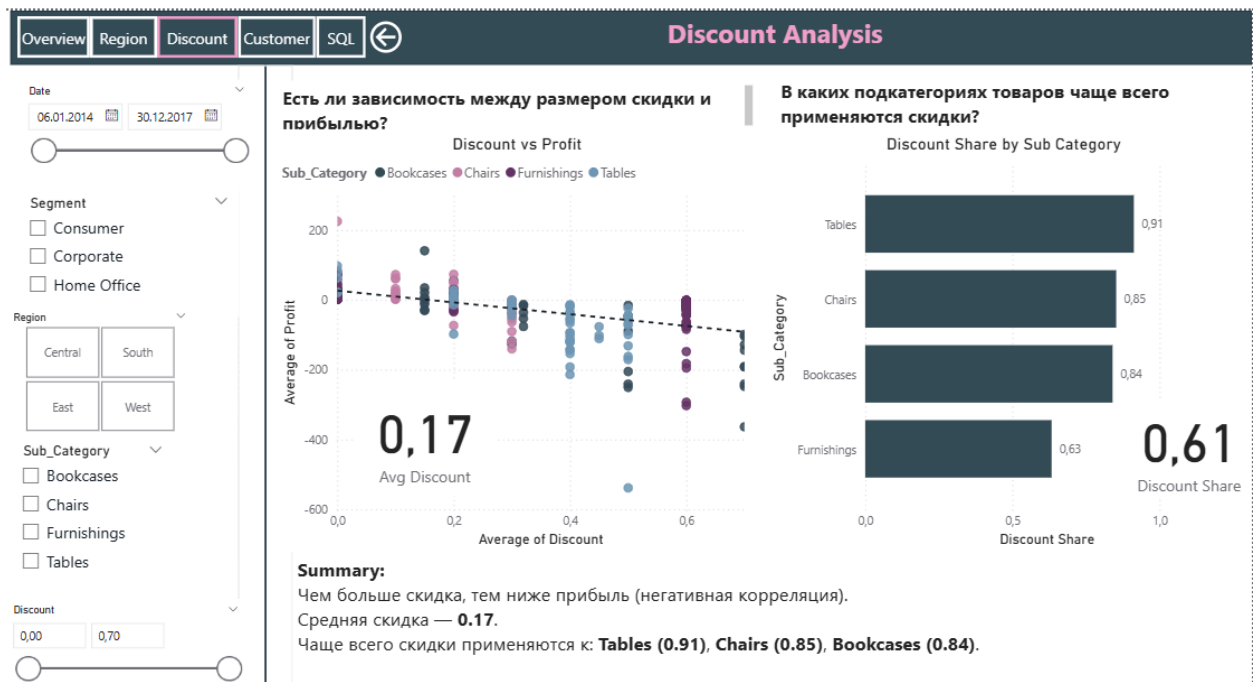
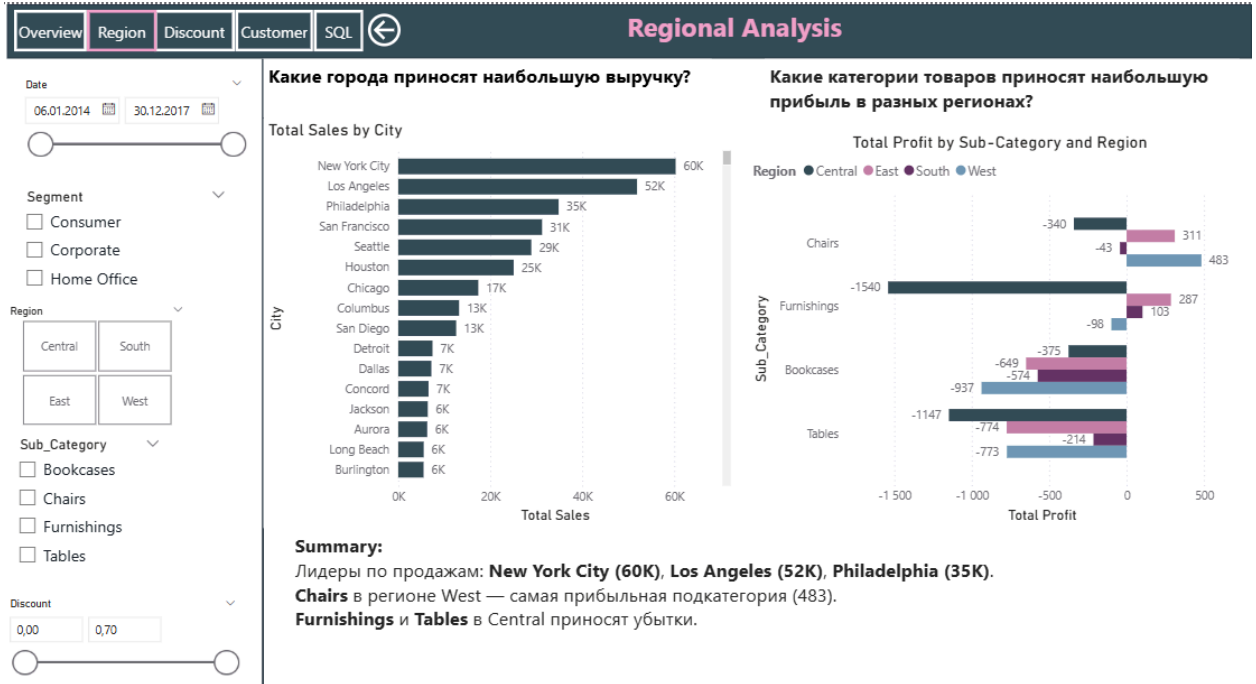
Технические требования

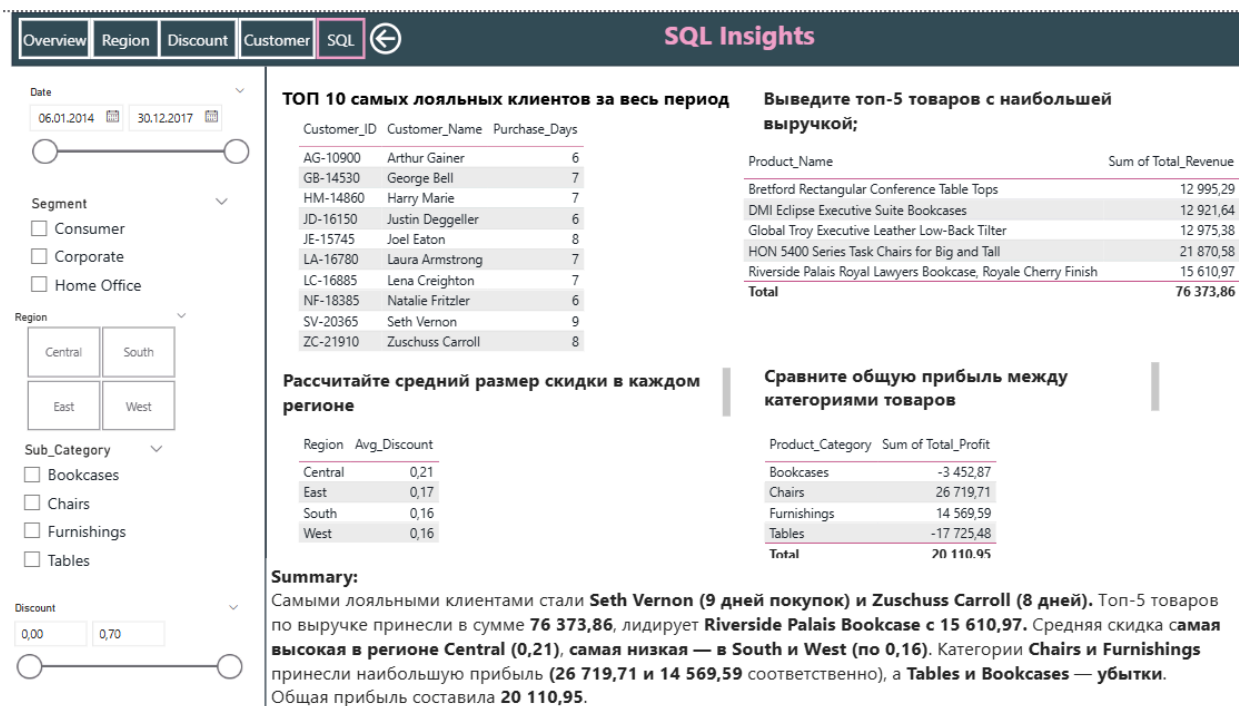
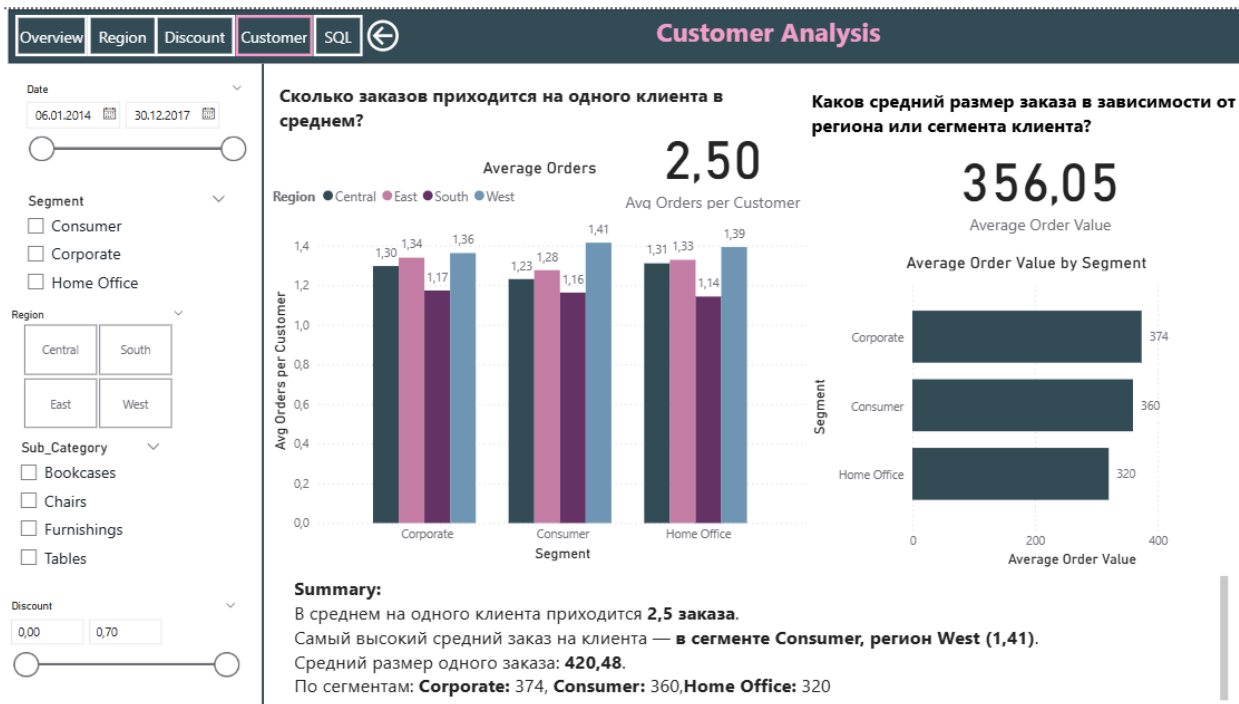
В отчёте обязательно реализуйте следующее:

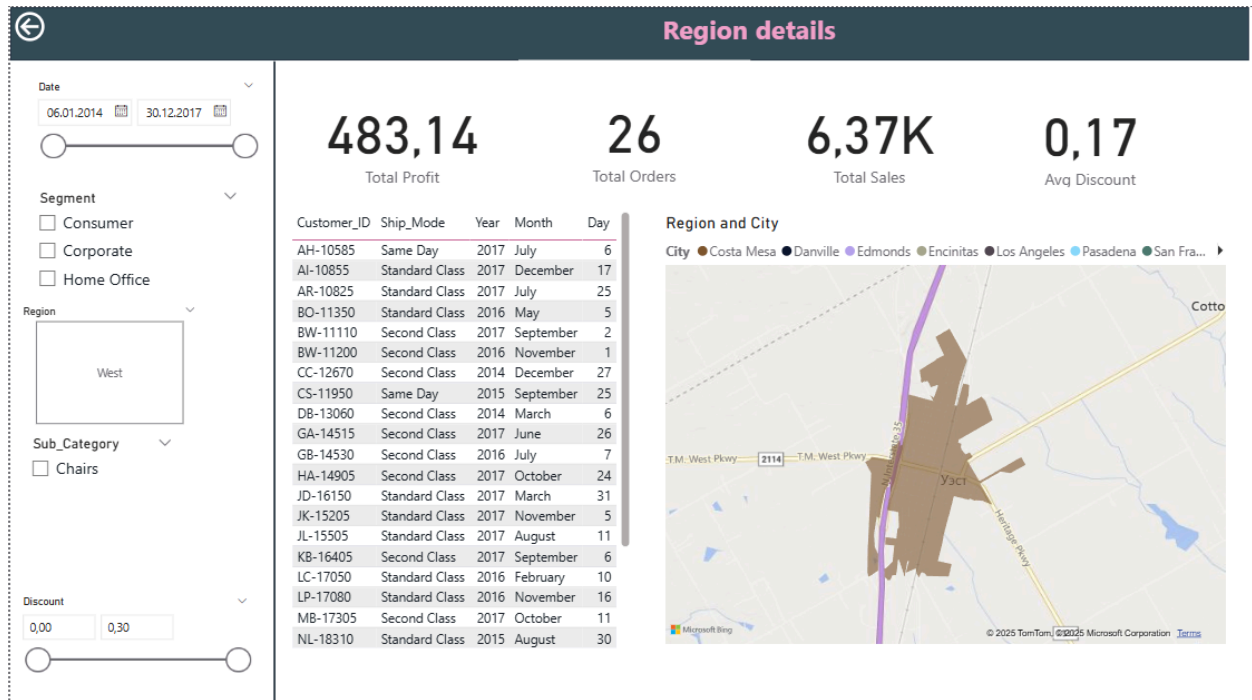
- Фильтрацию отчёта по году, месяцу, сегменту и региону;
- Drill-through страницу с деталями конкретного региона;
- Фильтр, отображающий только заказы со скидками;

- Возможность выбрать категорию товара и отслеживать ее динамику продаж и прибыли во времени;
- Используйте несколько страниц отчета, чтобы логически разделить разные части анализа (например: общие инсайты, региональный анализ, клиентская аналитика и т.д.);
- Отобразите результаты нескольких SQL-запросов в виде отдельных таблиц или визуализаций в отчёте Power BI.
- Обеспечьте удобную навигацию между страницами отчёта (например, через кнопки)









3. Python

- Импортируйте скачанный файл с BigQuery в Python

```
import pandas as pd

df = pd.read_csv('/content/drive/MyDrive/Data Analytics
N!/order_details.csv')
```

- Проверьте типы данных в таблицах

```
df.dtypes
```

Order_ID	object
Order_Date	object
Ship_Date	object
Ship_Mode	object
Customer_ID	object
Customer_Name	object
Segment	object
Country	object
City	object
State	object
Postal_Code	int64
Region	object
Product_ID	object
Product_Name	object
Category	object
Sub_Category	object
Sales	float64
Quantity	int64
Discount	float64
Profit	float64

- Преобразуйте даты (datetime), добавьте столбцы month, year.

```
df['Order_Date'] = pd.to_datetime(df['Order_Date'])

df['Month'] = df['Order_Date'].dt.month

df['Year'] = df['Order_Date'].dt.year
```

Анализ (EDA)

- Постройте графики: распределение продаж, прибыли, скидок.

```
df['Order_Date'] = pd.to_datetime(df['Order_Date'])
mydata = df.groupby('Order_Date')[['Sales', 'Profit',
'Discount']].sum().reset_index()
```

```
#Продажи
```

```
plt.figure(figsize=(14, 4))
sns.lineplot(data=mydata, x='Order_Date', y='Sales', color='blue')
plt.title('Распределение продаж по датам')
plt.xlabel('Дата')
plt.ylabel('Суммарные продажи')
plt.grid(True)
plt.tight_layout()
plt.show()
```



```
#Прибыль
```

```
plt.figure(figsize=(14, 4))
sns.lineplot(data=mydata, x='Order_Date', y='Profit', color='green')
plt.title('Распределение прибыли по датам')
plt.xlabel('Дата')
plt.ylabel('Суммарная прибыль')
plt.axhline(0, color='red', linestyle='--')
plt.grid(True)
plt.tight_layout()
plt.show()
```



```

#Скидки
plt.figure(figsize=(14, 4))
sns.lineplot(data=mydata, x='Order_Date', y='Discount',
color='orange')
plt.title('Распределение скидок по датам')
plt.xlabel('Дата')
plt.ylabel('Суммарная скидка')
plt.grid(True)
plt.tight_layout()
plt.show()

```



```

# Группировка по дате
daily = df.groupby('Order_Date')[['Sales', 'Profit',
'Discount']].sum().reset_index()
# Применим скользящее среднее (на 7 дней)
daily['Sales_smooth'] = daily['Sales'].rolling(window=7).mean()
daily['Profit_smooth'] = daily['Profit'].rolling(window=7).mean()
daily['Discount_smooth'] =
daily['Discount'].rolling(window=7).mean()
# Построим график
plt.figure(figsize=(15, 6))
sns.lineplot(x='Order_Date', y='Sales_smooth', data=daily,
label='Продажи', color='blue')
sns.lineplot(x='Order_Date', y='Profit_smooth', data=daily,
label='Прибыль', color='green')
sns.lineplot(x='Order_Date', y='Discount_smooth', data=daily,
label='Скидка', color='orange')

plt.title('Тренды продаж, прибыли и скидок ')
plt.xlabel('Дата')
plt.ylabel('Суммарные значения ')

```

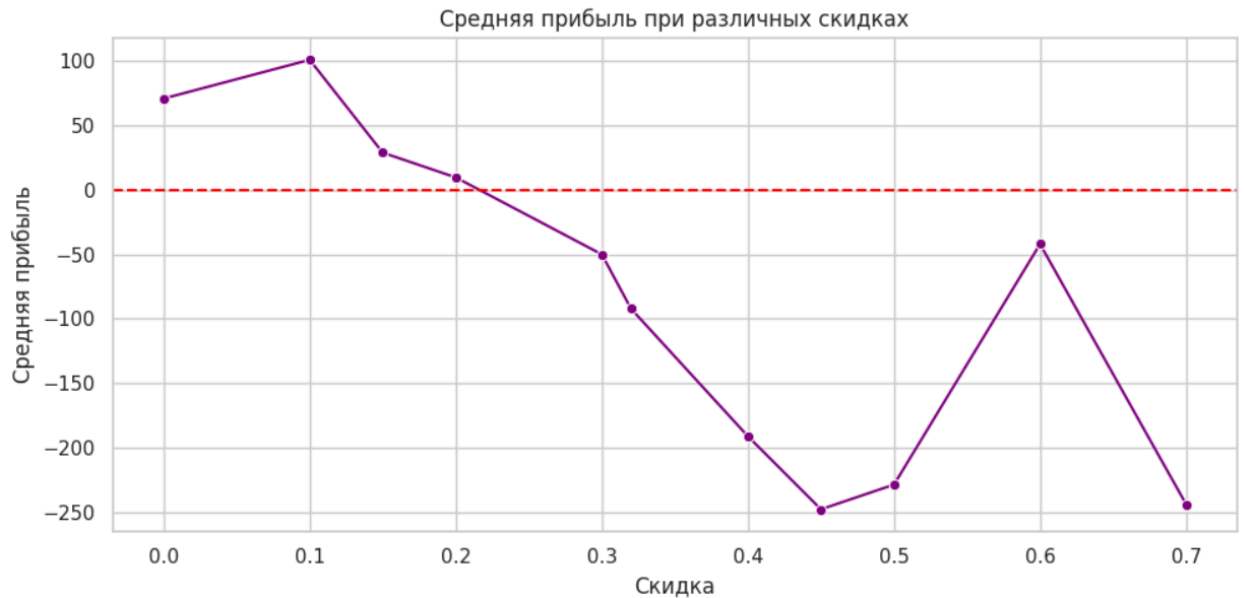
```
plt.legend()
plt.grid(True)
plt.tight_layout()
plt.show()
```



- Исследуйте зависимости между скидкой и прибылью.

```
discount_profit =
df.groupby('Discount')['Profit'].mean().reset_index()

plt.figure(figsize=(10, 5))
sns.lineplot(data=discount_profit, x='Discount', y='Profit',
marker='o', color='purple')
plt.title('Средняя прибыль при различных скидках')
plt.xlabel('Скидка')
plt.ylabel('Средняя прибыль')
plt.axhline(0, linestyle='--', color='red')
plt.grid(True)
plt.tight_layout()
plt.show()
```



```
corr = df[['Discount', 'Profit']].corr()
print("Корреляция между скидкой и прибылью:")
corr
```

Корреляция между скидкой и прибылью:

	Discount	Profit
Discount	1.000000	-0.478304
Profit	-0.478304	1.000000

Linear Regression Model

- Постройте модель Linear Regression для предсказания колонки Profit на основе Sales, Discount, Quantity

```
import pandas as pd
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error, r2_score
# Выбираем признаки
features = ['Sales', 'Discount', 'Quantity']
target = 'Profit'
# Удалим пропущенные значения
df_model = df[features + [target]].dropna()
```



```

# Признаки и целевая переменная
X = df_model[features]
y = df_model[target]
model = LinearRegression()
model.fit(X, y)

# Коэффициенты модели
coefficients = pd.DataFrame({
    'Признак': features,
    'Коэффициент': model.coef_
})
print(coefficients)

```

- Какие признаки влияют на прибыль?

	Признак	Коэффициент
0	Sales	0.057770
1	Discount	-338.482361
2	Quantity	0.230205

- Что означает коэффициент при Discount?
Если скидка становится больше, прибыль сильно падает.
Например, при увеличении скидки всего на чуть-чуть, прибыль уменьшается на **десятки или сотни**.
Это значит, что **скидки — плохо влияют на доход компании**. Чем больше скидка, тем меньше заработок.

- Добавьте новую колонку profit_pred в вашу таблицу и сохраните ее в csv формате(forecast.csv)

```

df['profit_pred'] = model.predict(df[features])
df.to_csv('/content/drive/MyDrive/Data Analytics N!/forecast.csv',
index=False)

```

- Я хотела посмотреть на мой график, как справилась моя модель регрессии с предсказанием.

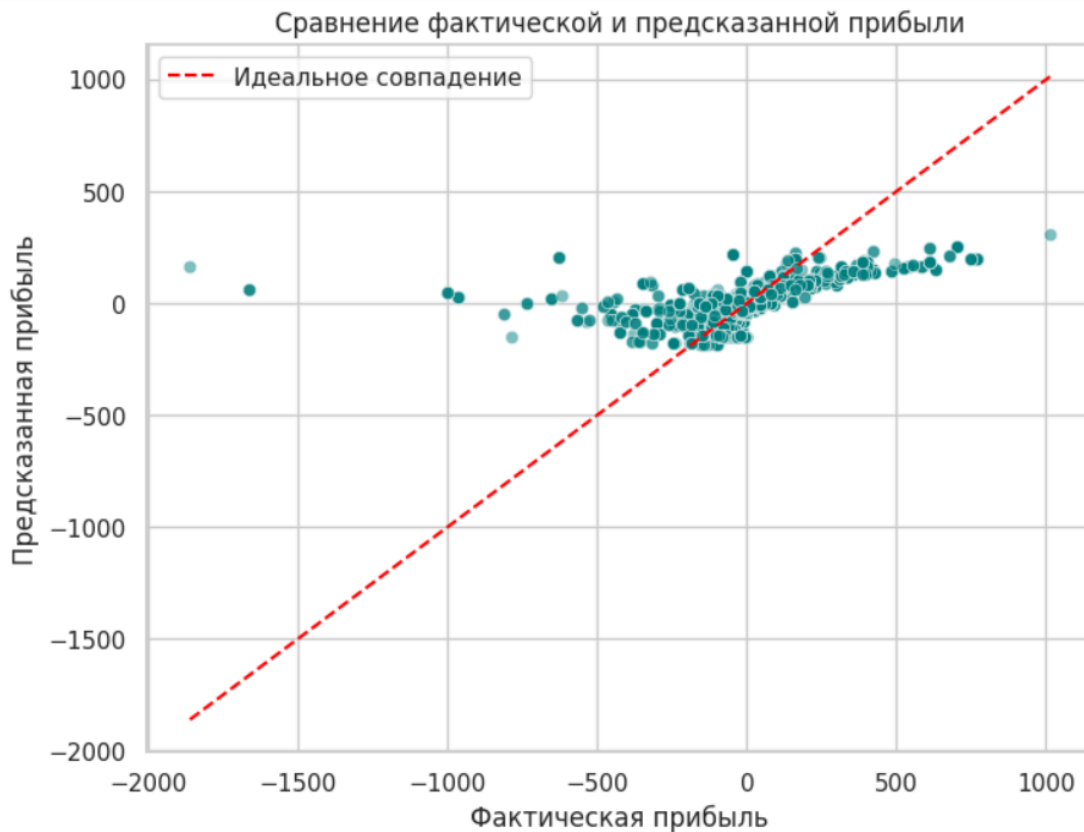
```

import matplotlib.pyplot as plt
import seaborn as sns

plt.figure(figsize=(8, 6))
sns.scatterplot(x=df['Profit'], y=df['profit_pred'], alpha=0.5,
color='teal')
plt.plot([df['Profit'].min(), df['Profit'].max()],
         [df['Profit'].min(), df['Profit'].max()],
         color='red', linestyle='--', label='Идеальное совпадение')

```

```
plt.xlabel('Фактическая прибыль')
plt.ylabel('Предсказанная прибыль')
plt.title('Сравнение фактической и предсказанной прибыли')
plt.legend()
plt.grid(True)
plt.show()
```



В целом неплохо, но я здесь вижу выбросы и что все скоплены в одной точке. Из-за этого я переделала свою модель линейной регрессии.

Удалила выбросы и добавила полиномиальные признаки, затем использовала RandomForest для линейной регрессии.

```
# Удалим выбросы по правилу IQR
Q1 = df['Profit'].quantile(0.25)
Q3 = df['Profit'].quantile(0.75)
IQR = Q3 - Q1
```

```

# Фильтруем только адекватные значения
df_clean = df[(df['Profit'] >= Q1 - 1.5 * IQR) & (df['Profit'] <= Q3 + 1.5
* IQR)]

from sklearn.preprocessing import PolynomialFeatures
X = df_clean[['Sales', 'Discount', 'Quantity']]
y = df_clean['Profit']
poly = PolynomialFeatures(degree=2, include_bias=False)
X_poly = poly.fit_transform(X)

from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error, r2_score
import numpy as np

# Делим данные
X_train, X_test, y_train, y_test = train_test_split(X_poly, y,
test_size=0.2, random_state=42)

# Обучаем модель
model = RandomForestRegressor(random_state=42)
model.fit(X_train, y_train)

# Предсказания
y_pred = model.predict(X_test)

# Оценки качества модели
mse = mean_squared_error(y_test, y_pred)
rmse = np.sqrt(mse)
r2 = r2_score(y_test, y_pred)

# Выводим результаты
print("MSE:", mse)
print("RMSE:", rmse)
print("R2:", r2)

```

MSE: 49.3723466007307

RMSE: 7.0265458513220205

R2: 0.9530435670313302

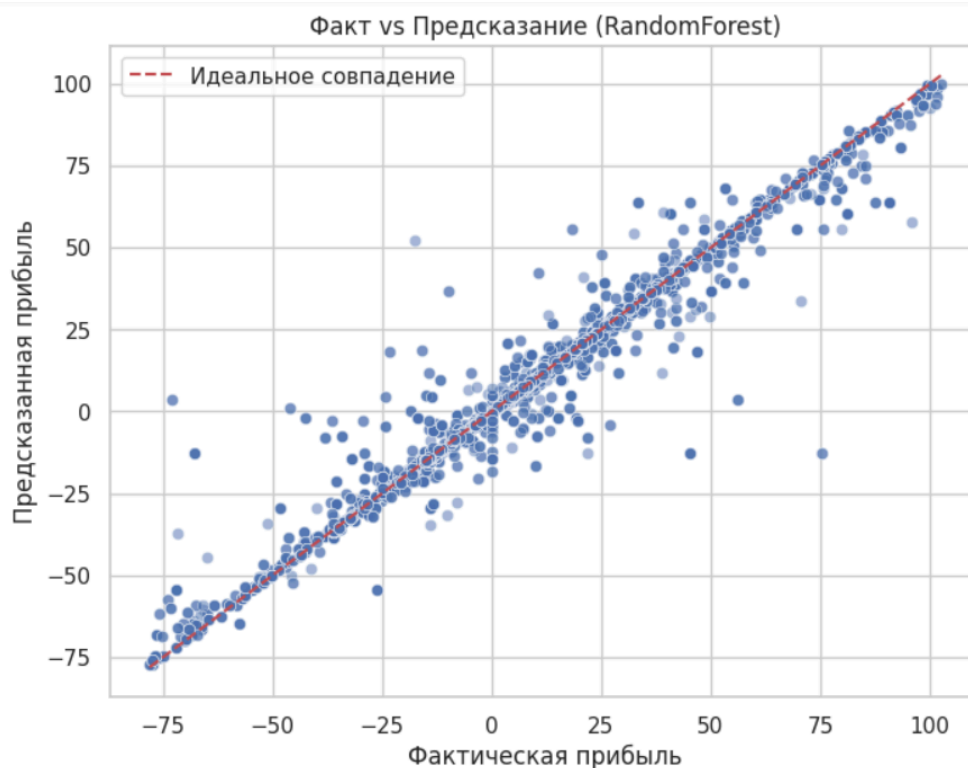
```

df_clean['profit_pred'] = model.predict(X_poly)

# Сохраняем
df_clean.to_csv('/content/drive/MyDrive/Data Analytics
N!/forecast_improved.csv', index=False)

plt.figure(figsize=(8,6))
sns.scatterplot(x=y, y=model.predict(X_poly), alpha=0.5)
plt.plot([y.min(), y.max()], [y.min(), y.max()], 'r--', label='Идеальное
совпадение')
plt.xlabel('Фактическая прибыль')
plt.ylabel('Предсказанная прибыль')
plt.title('Факт vs Предсказание (RandomForest)')
plt.legend()
plt.grid(True)
plt.show()

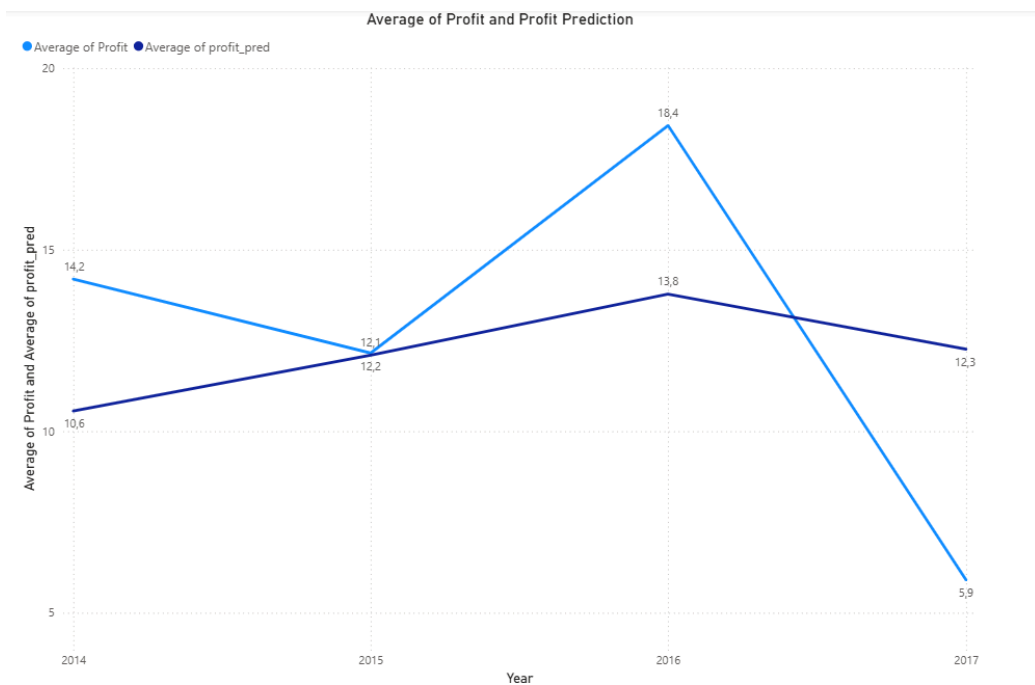
```



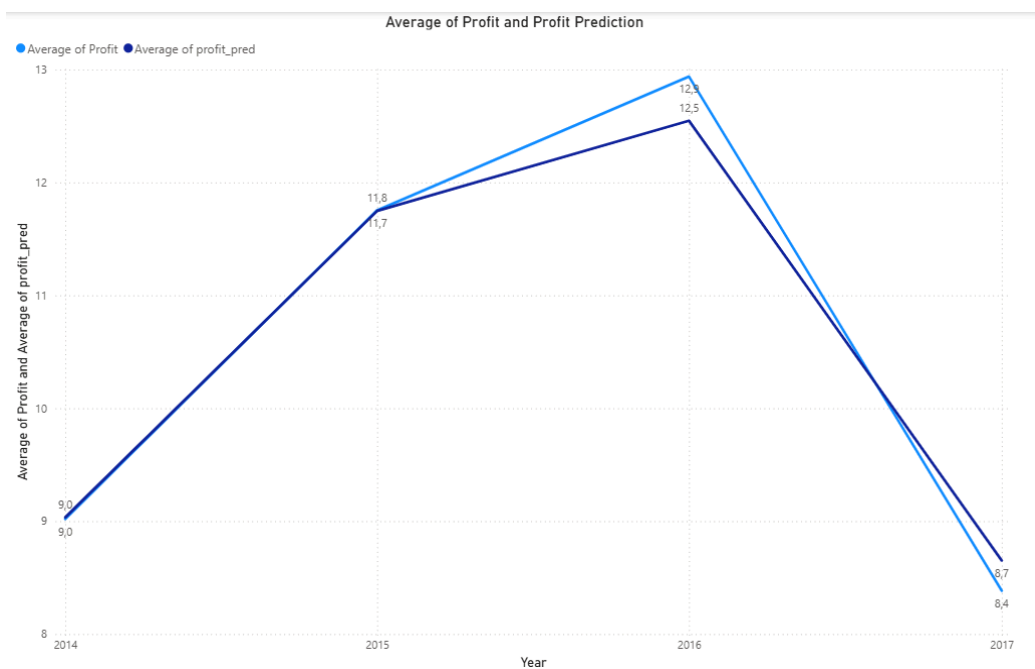
4. Power BI (продолжение)

Вернитесь в свой отчет в Power BI и добавьте визуализацию прогноза продаж (данные из forecast.csv) в сравнение с реальными значениями прибыли.

Первая модель Линейной регрессии:



Вторая модель Линейной регрессии с Random Forest:



Как можно увидеть на графиках вторая модель справилась лучше с предсказаниями.

Summary

Продажи показывают сезонность с пиками в декабре, прибыль снижалась после 2016 года. Самая прибыльная доставка — Standard Class, а Same Day — наименее выгодна. Лидеры по продажам — New York, Los Angeles и Philadelphia. В West Chairs приносят максимум прибыли, а в Central Tables и Furnishings — убытки. Чем выше скидка, тем ниже прибыль, особенно в категории Tables (скидки 91%). Средний клиент делает 2,5 заказа, самый лояльный — Seth Vernon (9 покупок). Топ-5 товаров дали 76К выручки, лидер — Riverside Palais Bookcase. Для роста прибыли стоит сократить скидки, пересмотреть стратегию в Central и развивать доставку Standard Class.