رواد مصر الرقمية

وزارة الاتصالات
وتكنولوجيا المعلومات

Superstore PROJECT

By:

Ahmed Mohamed

Abdallah Adel

INTRODUCTION

This project focuses on transforming a raw superstore dataset into a Clean,Organized, And Analysis - ready structure.

The main objective was to :

1- Fix Data Quality Issues.

2- Normalize the Dataset into a relational Database

3- Analyze Business Performance Using SQL and Python

4- Build Dashboards to Visualize Key insights Across different Tools

DATA ISSUES

When we were cleaning the data . We didn't just looking for nulls . We had to hint for integrity issues . And we found 2 huge issues in product's data .

First : We had "duplicate products" , the same product name like "stables" appearing with two different IDs .

second : we had " corrupt IDS " The same product ID pointing to 2 different names .

Tools Used :

1

2

3

4

Excel

Used for initial data cleaning, organizing datasets, detecting missing values, and performing basic statistical analysis.

Tools Used :

1

2

3

4

SQL

Used to extract, filter, and aggregate data from the database, and to run complex queries for deeper insights.

Tools Used :

1

2

3

4

Python

Used for advanced data cleaning, exploratory data analysis , and applying analytical libraries such as Pandas, NumPy, and Matplotlib.

Tools Used :

1

2

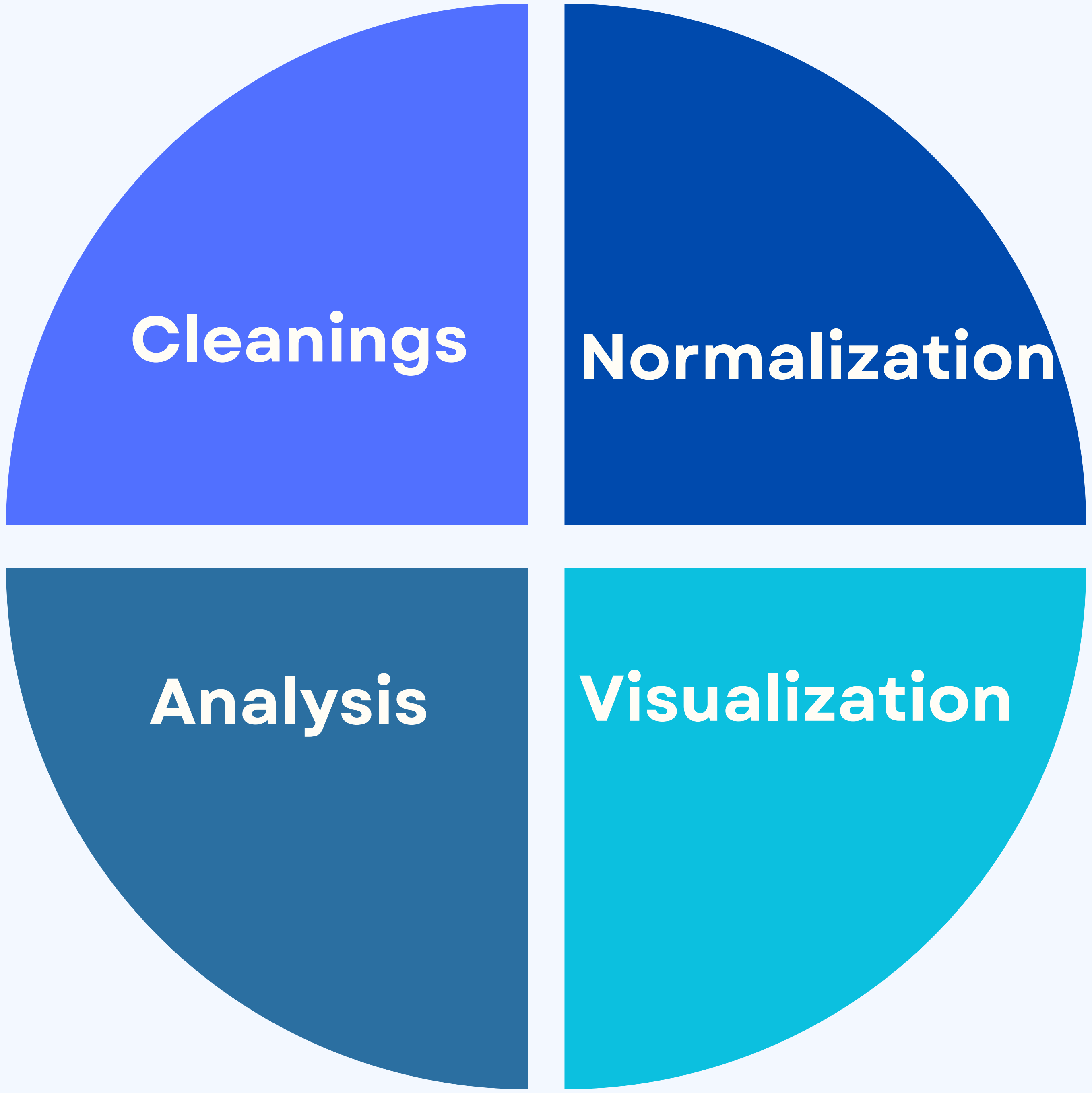
3

4

Tableau

Used to create interactive dashboards and visualizations that clearly present the insights and trends discovered in the analysis.

We Divided Our Project into 4 Phases



Cleanings

Normalization

Analysis

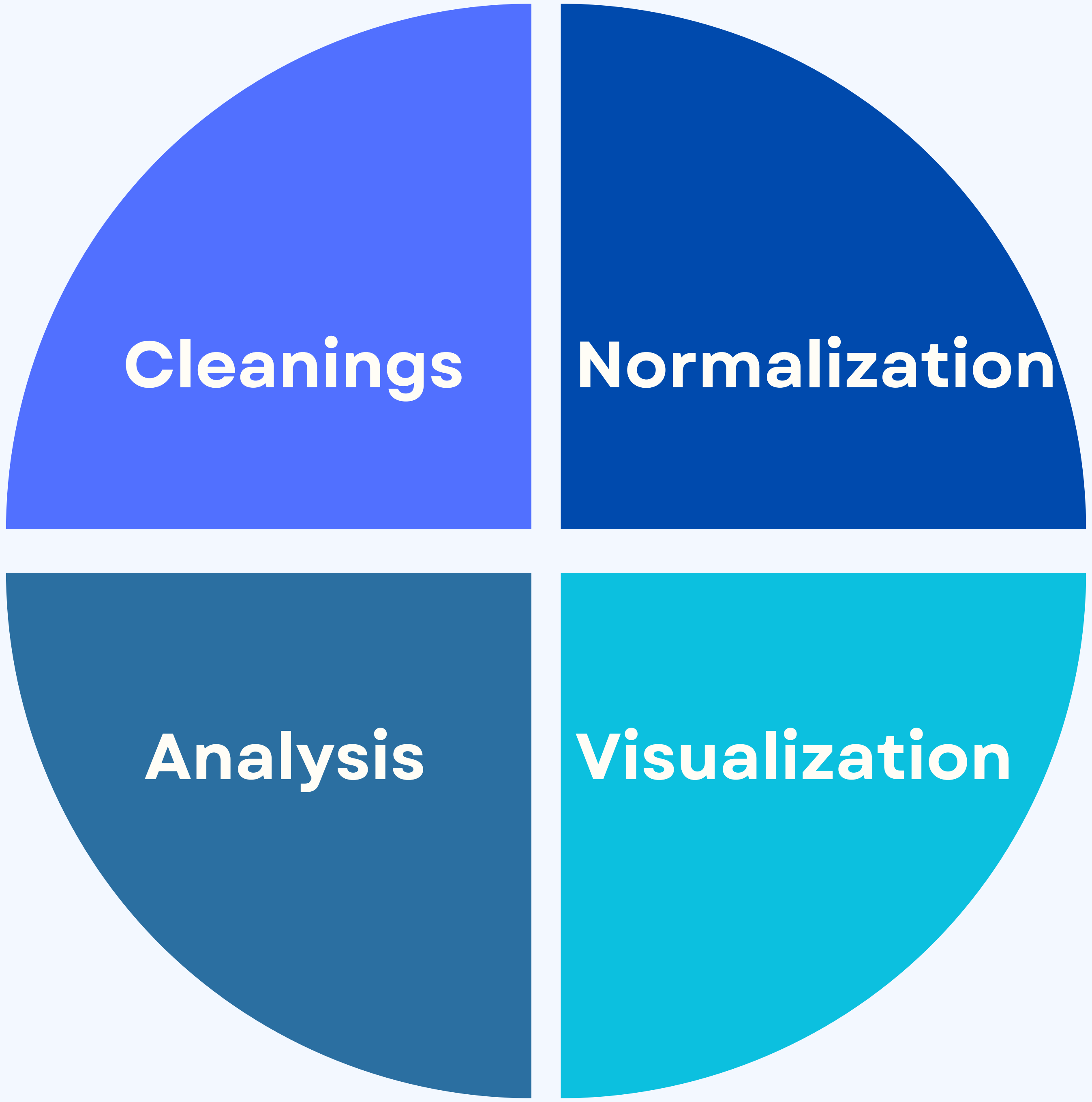
Visualization

Cleanings Steps :

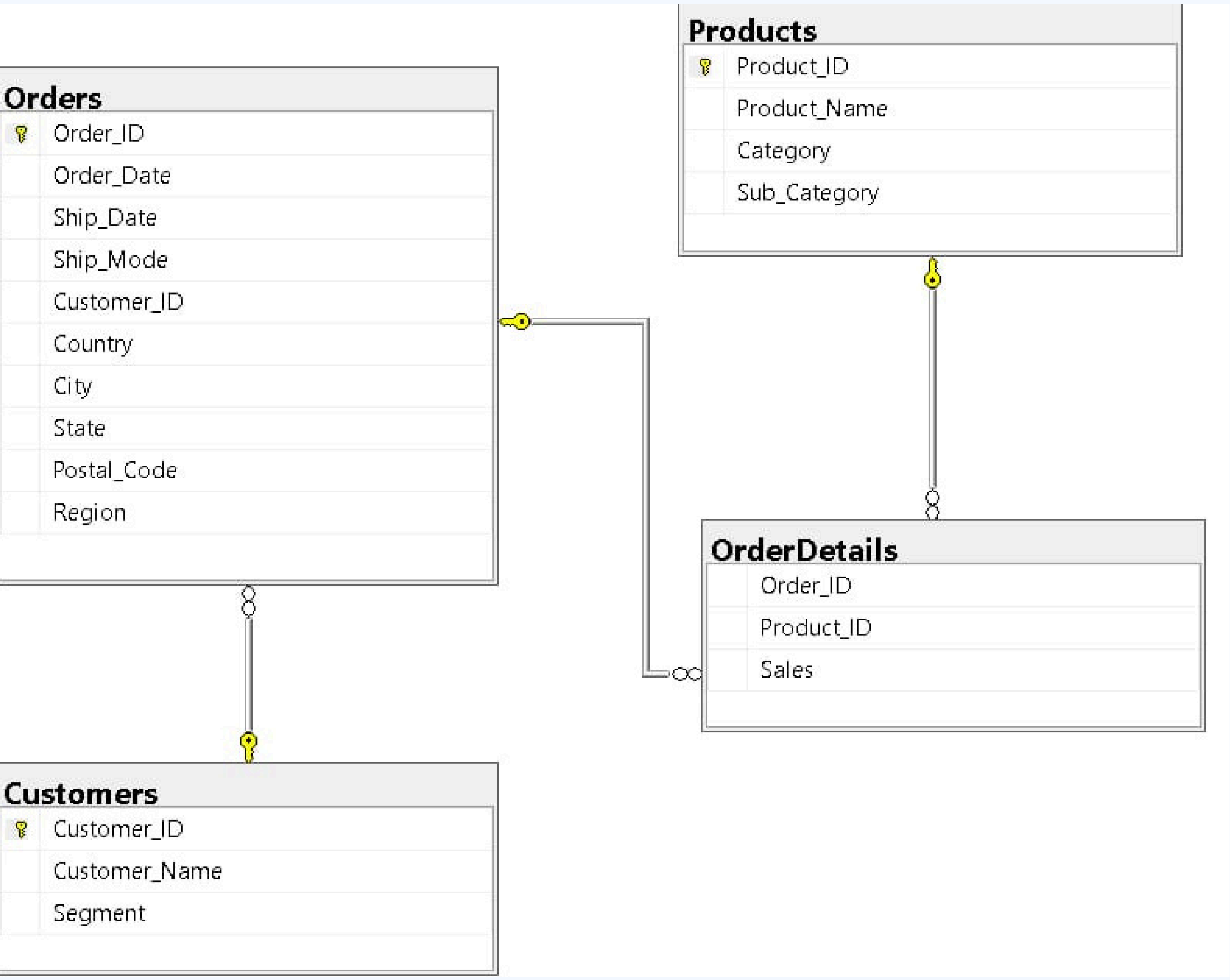
To ensure data accuracy we performed:

- Removal of duplicate rows
- Correction and standrization of product names
- Resolving conflicting product IDs
- Ensuring consistent formatting and data types

Cleanings


Cleanings
Normalization
Analysis
Visualization

Normalization



Products
- Product_ID
- Product_Name
- Category
- Sub_Category

Orders
- Order_ID
- Order_Date
- Ship_Date
- Ship_Mode
- Customer_ID
- Country
- City
- State
- Postal_Code
- Region

OrderDetails
- Order_ID
- Product_ID
- Sales

Customers
- Customer_ID
- Customer_Name
- Segment

We normalized the dataset up to Third Normal Form (3NF) to reduce redundancy and improve data integrity

The final tables:

1- Customers

2-Orders

3-Order Details

4-Products

Each table contains a unique entity, and all relationships are defined clearly through keys

Cleanings

Normalization

Analysis

Visualization

Analysis

In this stage, we used SQL and Python to analyze the structured Superstore dataset after cleaning and normalization.
The analysis focused on understanding order behavior, customer patterns, product performance, and shipping efficiency.

Main Analysis Points:

1. Order Trends

- Analyzed the number of orders per month and per year.
- Identified peak ordering periods and slow seasons.

Analysis

2. Customer Insights

- Studied the distribution of customers across regions and states.
- Identified high-order customers and repeated buyers.

3. Product Performance

- Counted how many times each product was ordered.
- Identified most frequently ordered products vs. low-demand products.

Analysis

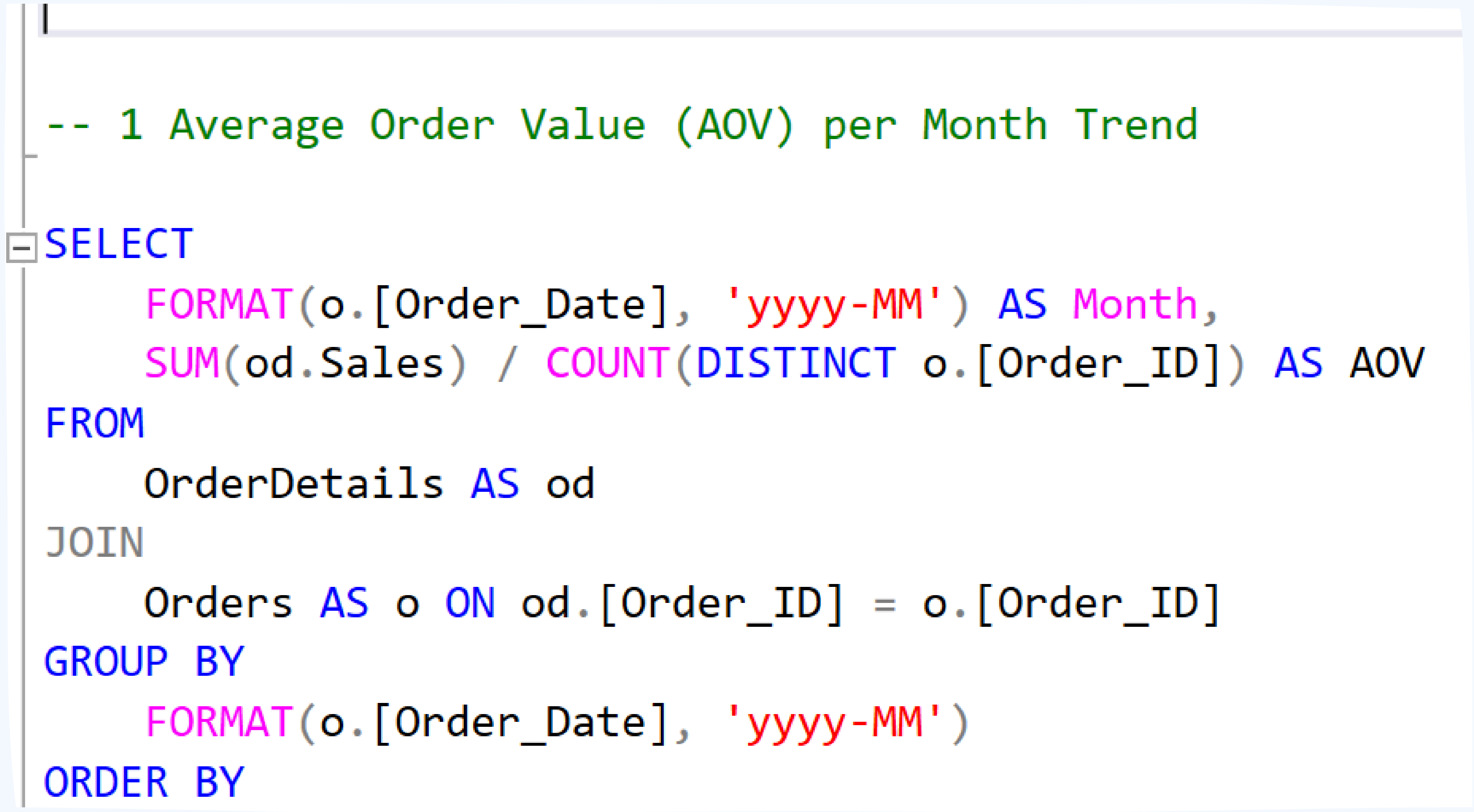
4. Shipping & Delivery Analysis

- Calculated average shipping time for each shipping mode.
- Compared delivery performance across different regions.

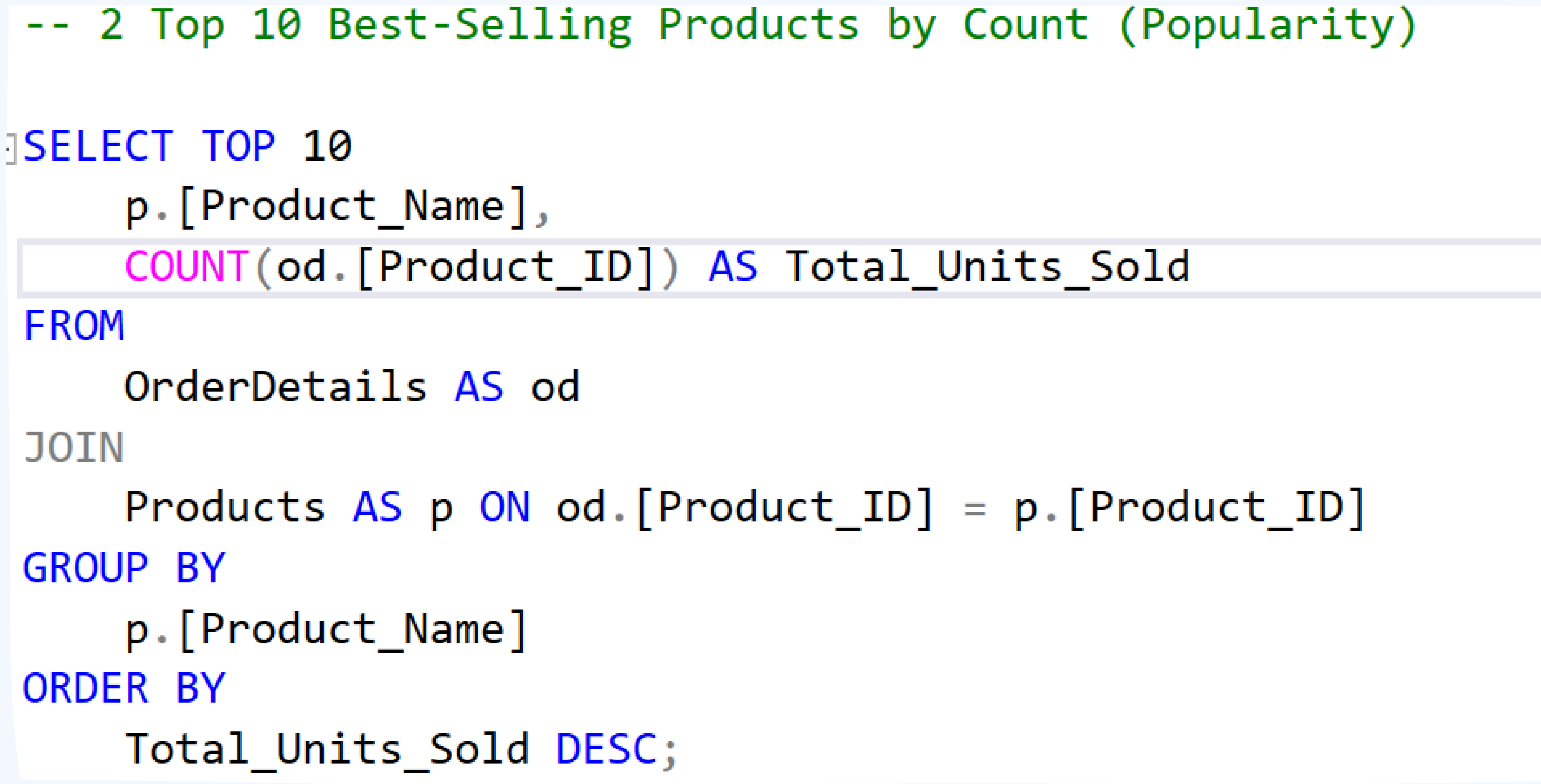
5. Order Distribution by Category & Region

- Analyzed how many orders came from each category.
- Compared category demand across states/regions.

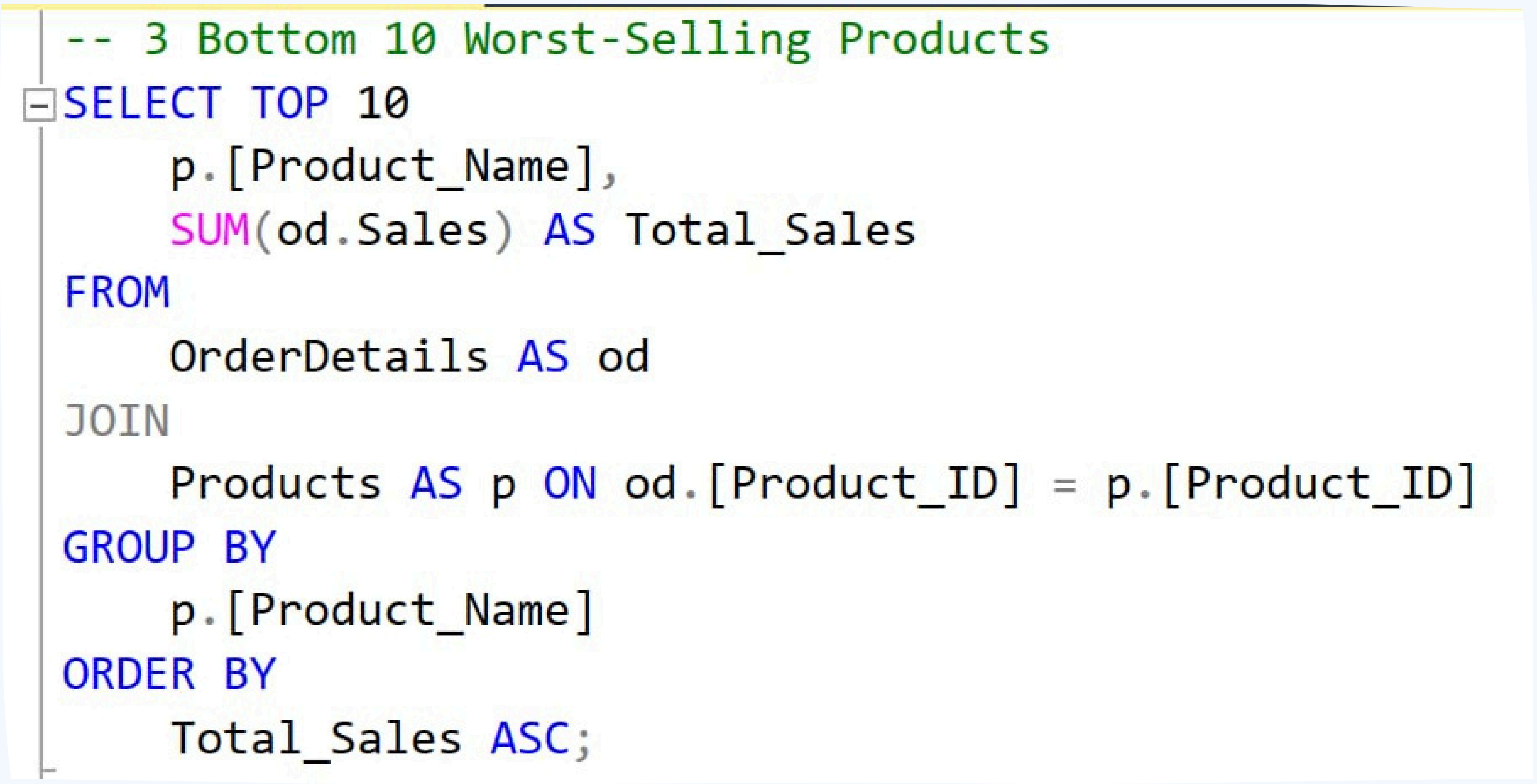
SQL Examples:



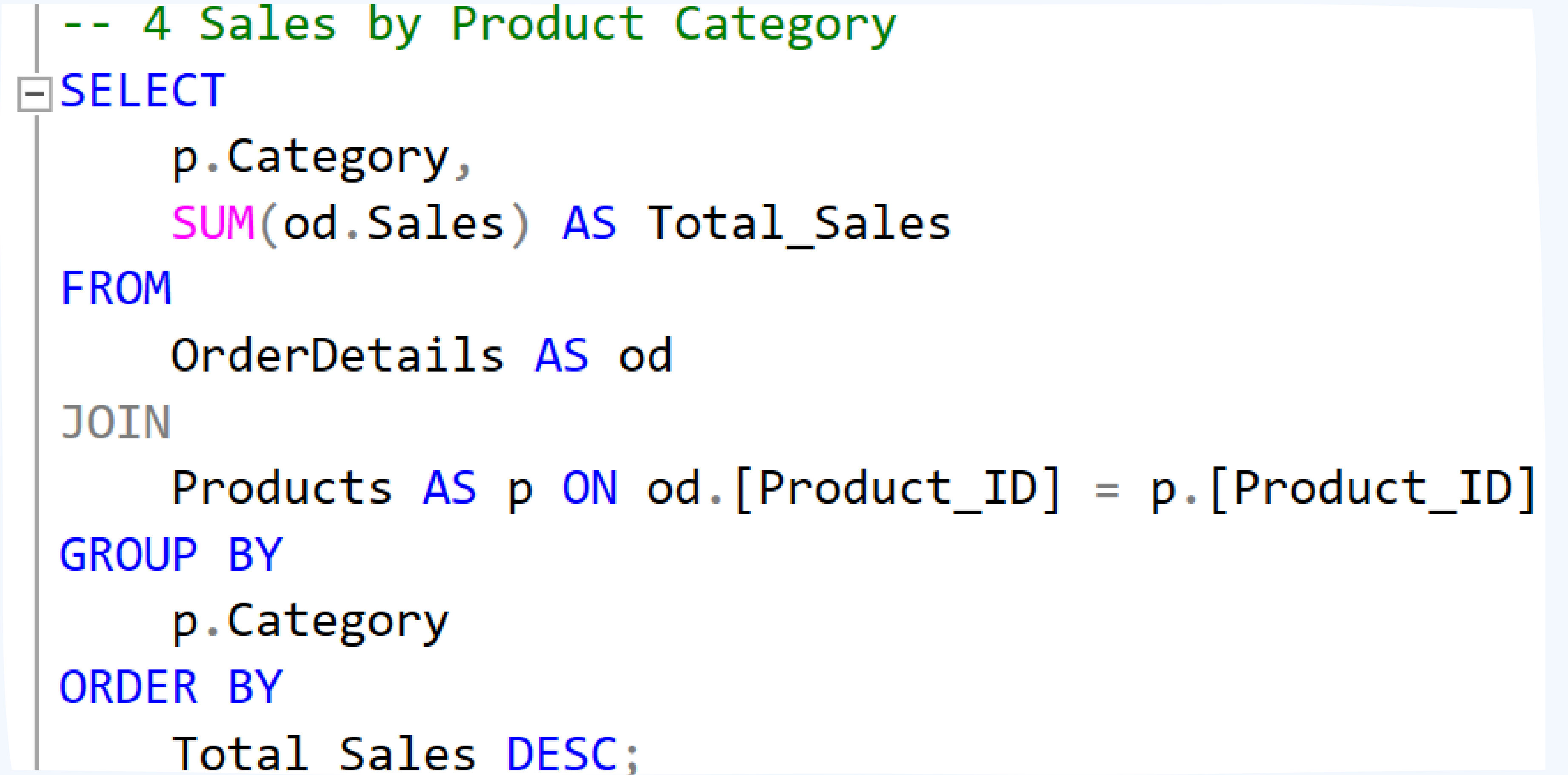
```sql
-- 1 Average Order Value (AOV) per Month Trend

SELECT
    FORMAT(o.[Order_Date], 'yyyy-MM') AS Month,
    SUM(od.Sales) / COUNT(DISTINCT o.[Order_ID]) AS AOV
FROM
    OrderDetails AS od
JOIN
    Orders AS o ON od.[Order_ID] = o.[Order_ID]
GROUP BY
    FORMAT(o.[Order_Date], 'yyyy-MM')
ORDER BY
```



```sql
-- 2 Top 10 Best-Selling Products by Count (Popularity)

SELECT TOP 10
    p.[Product_Name],
    COUNT(od.[Product_ID]) AS Total_Units_Sold
FROM
    OrderDetails AS od
JOIN
    Products AS p ON od.[Product_ID] = p.[Product_ID]
GROUP BY
    p.[Product_Name]
ORDER BY
    Total_Units_Sold DESC;
```

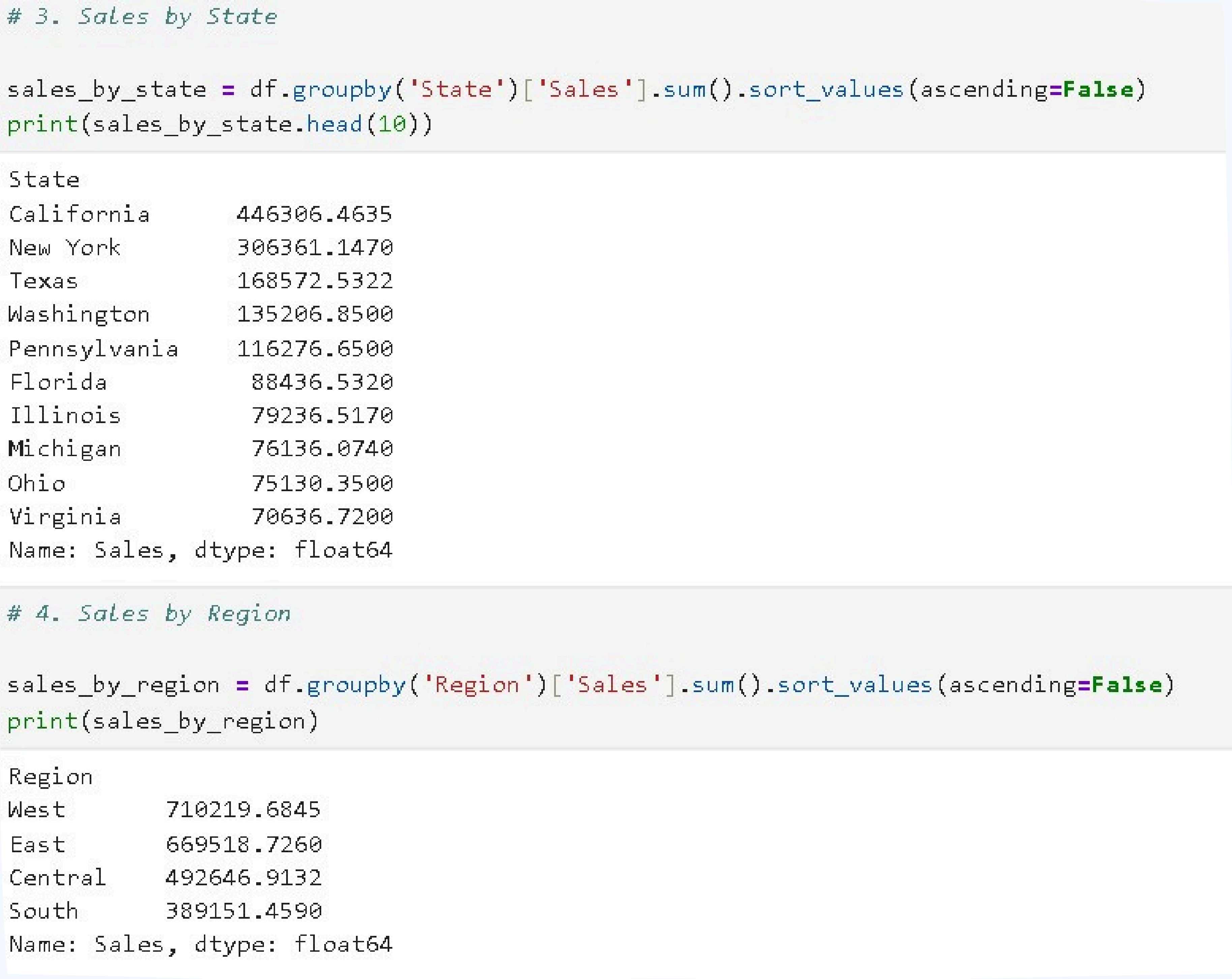


```sql
-- 3 Bottom 10 Worst-Selling Products
SELECT TOP 10
    p.[Product_Name],
    SUM(od.Sales) AS Total_Sales
FROM
    OrderDetails AS od
JOIN
    Products AS p ON od.[Product_ID] = p.[Product_ID]
GROUP BY
    p.[Product_Name]
ORDER BY
    Total_Sales ASC;
```



```sql
-- 4 Sales by Product Category
SELECT
    p.Category,
    SUM(od.Sales) AS Total_Sales
FROM
    OrderDetails AS od
JOIN
    Products AS p ON od.[Product_ID] = p.[Product_ID]
GROUP BY
    p.Category
ORDER BY
    Total_Sales DESC;
```

Python Examples:

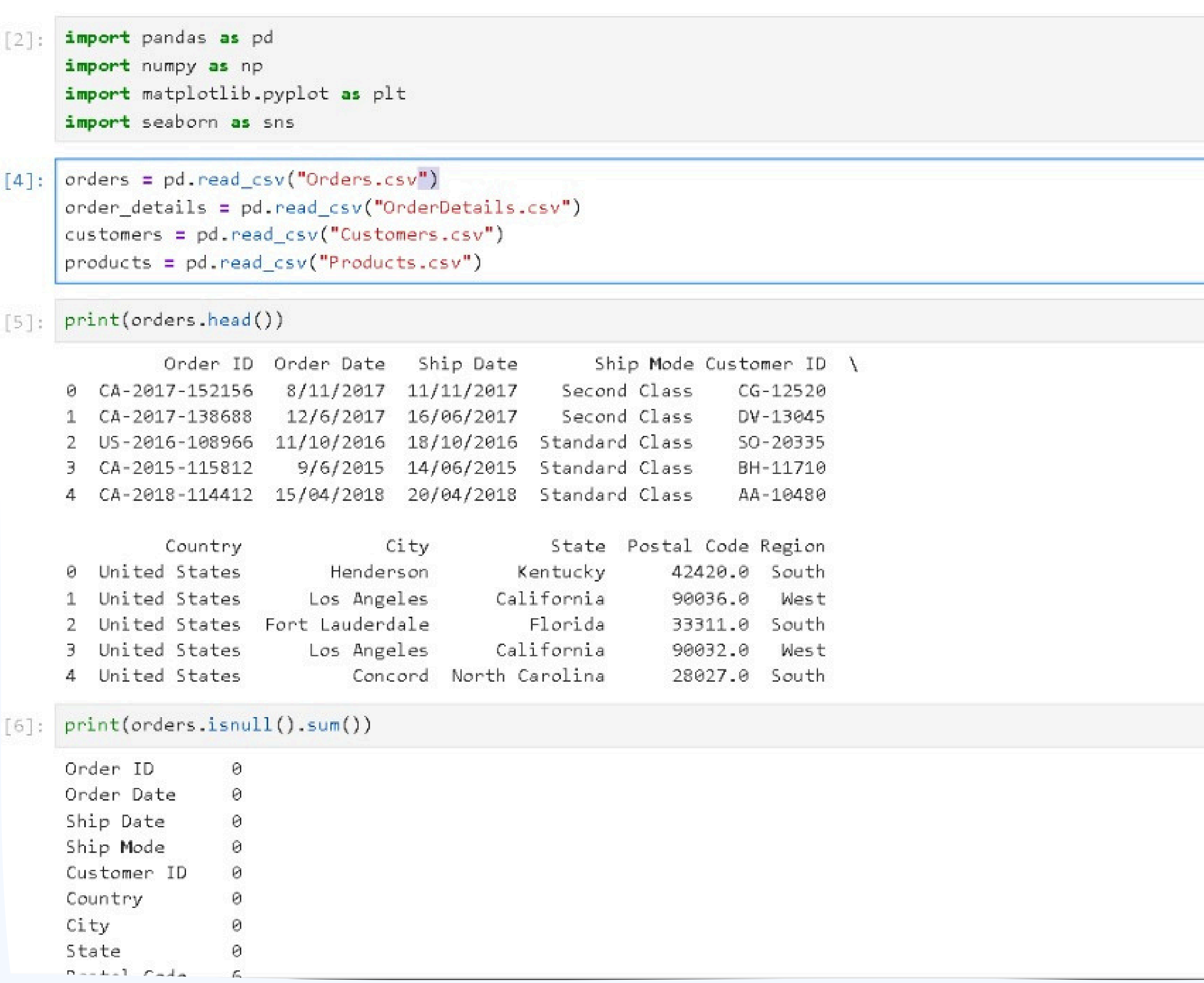


```
# 3. Sales by State

sales_by_state = df.groupby('State')['Sales'].sum().sort_values(ascending=False)
print(sales_by_state.head(10))
```

```
State
California     446306.4635
New York       306361.1470
Texas          168572.5322
Washington     135206.8500
Pennsylvania   116276.6500
Florida         88436.5320
Illinois        79236.5170
Michigan        76136.0740
Ohio            75130.3500
Virginia        70636.7200
Name: Sales, dtype: float64
```

```
# 4. Sales by Region

sales_by_region = df.groupby('Region')['Sales'].sum().sort_values(ascending=False)
print(sales_by_region)
```

```
Region
West       710219.6845
East       669518.7260
Central    492646.9132
South      389151.4590
Name: Sales, dtype: float64
```



```
[2]: import pandas as pd
     import numpy as np
     import matplotlib.pyplot as plt
     import seaborn as sns
```

```
[4]: orders = pd.read_csv("Orders.csv")
     order_details = pd.read_csv("OrderDetails.csv")
     customers = pd.read_csv("Customers.csv")
     products = pd.read_csv("Products.csv")
```

```
[5]: print(orders.head())
```

```
          Order ID  Order Date   Ship Date       Ship Mode Customer ID  \
0  CA-2017-152156   8/11/2017  11/11/2017    Second Class    CG-12520
1  CA-2017-138688   12/6/2017  16/06/2017    Second Class    DV-13045
2  US-2016-108966  11/10/2016  18/10/2016  Standard Class    SO-20335
3  CA-2015-115812    9/6/2015  14/06/2015  Standard Class    BH-11710
4  CA-2018-114412  15/04/2018  20/04/2018  Standard Class    AA-10480

         Country             City           State  Postal Code Region
0  United States        Henderson        Kentucky      42420.0  South
1  United States      Los Angeles      California      90036.0   West
2  United States  Fort Lauderdale         Florida      33311.0  South
3  United States      Los Angeles      California      90032.0   West
4  United States          Concord  North Carolina      28027.0  South
```

```
[6]: print(orders.isnull().sum())
```

```
Order ID       0
Order Date     0
Ship Date      0
Ship Mode      0
Customer ID    0
Country        0
City           0
State          0
Postal Code    6
```

Cleanings

Normalization

Analysis

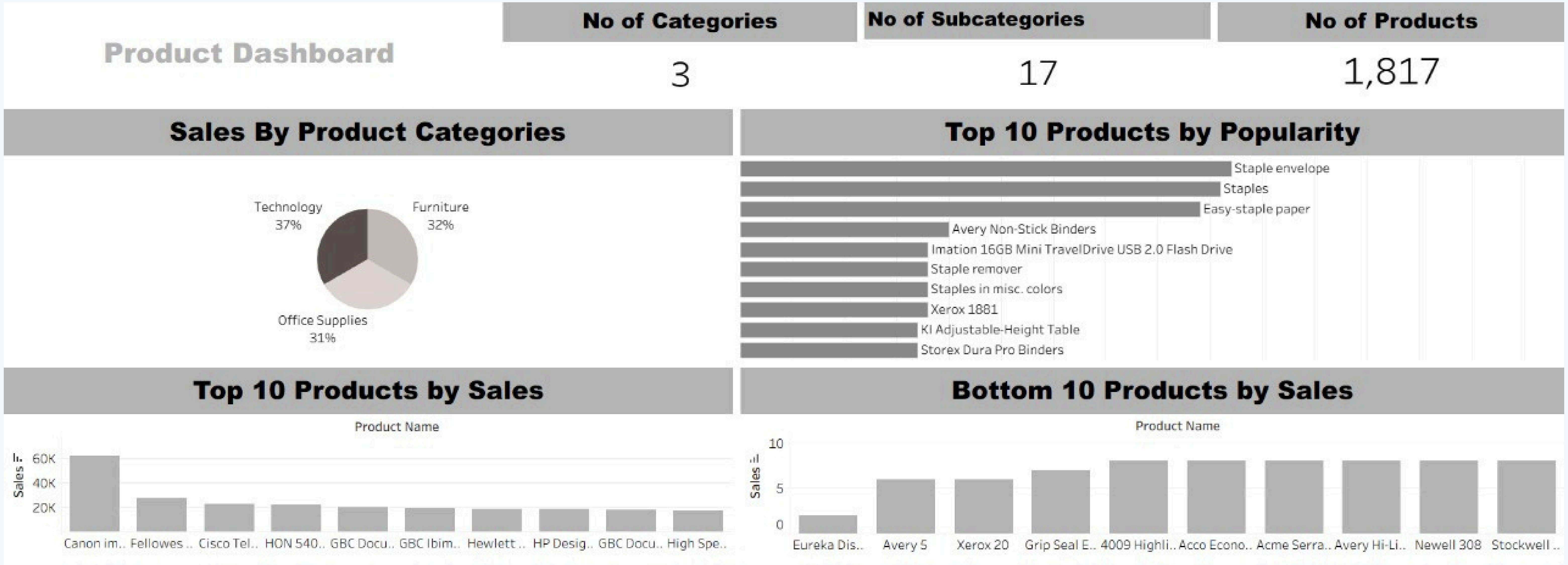
Visualization

Visualization

WE Created 5 interactive Dashboards, each designed to analyze the data from different perspective and support better decision making

Products Dashboard

We begin with the Product Dashboard. This dashboard offers us a comprehensive overview of our product portfolio performance, providing a detailed breakdown of sales percentages by major categories (Technology, Furniture, Office Supplies). It also highlights our top 10 best-selling and most popular products.

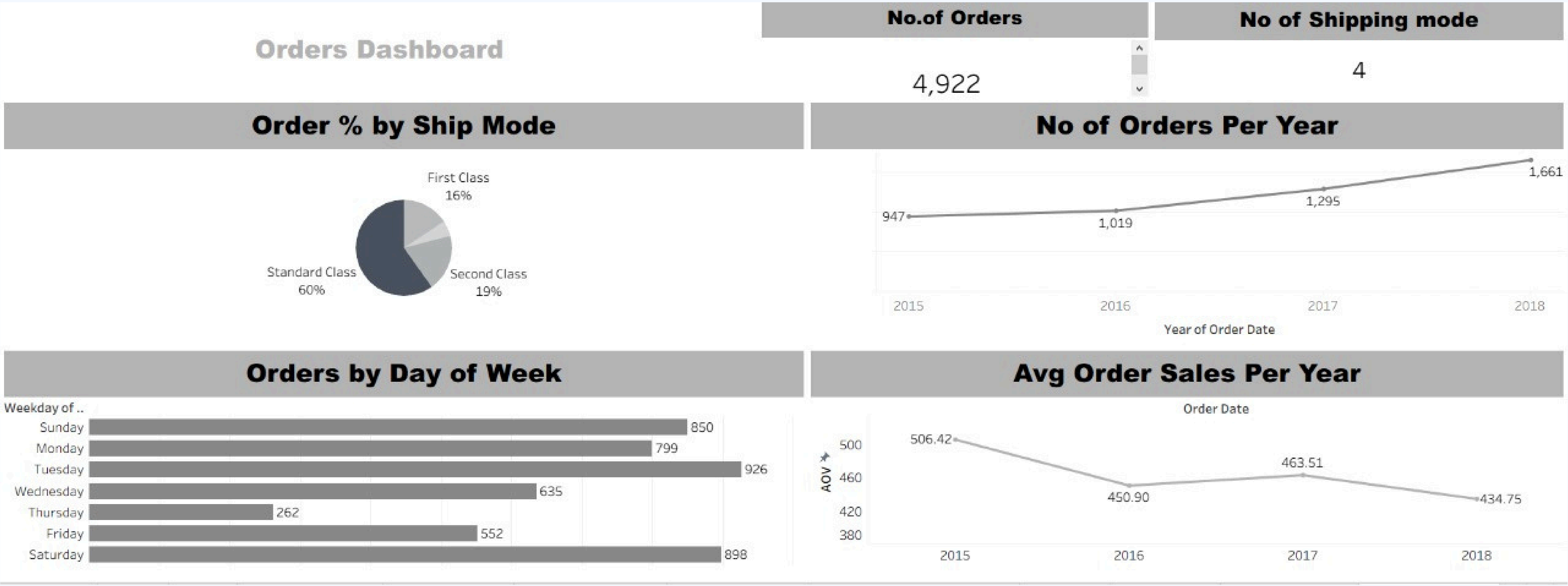


Product Dashboard

No of Categories	No of Subcategories	No of Products
3	17	1,817

Orders Dashboard

we conclude with the Orders Dashboard, which is essential for understanding purchasing patterns and business volume. This board shows us how orders are distributed across the four ship modes, illustrates the order growth trend over the years, and identifies the days of the week with the highest and lowest order rates, supporting better operational and inventory planning



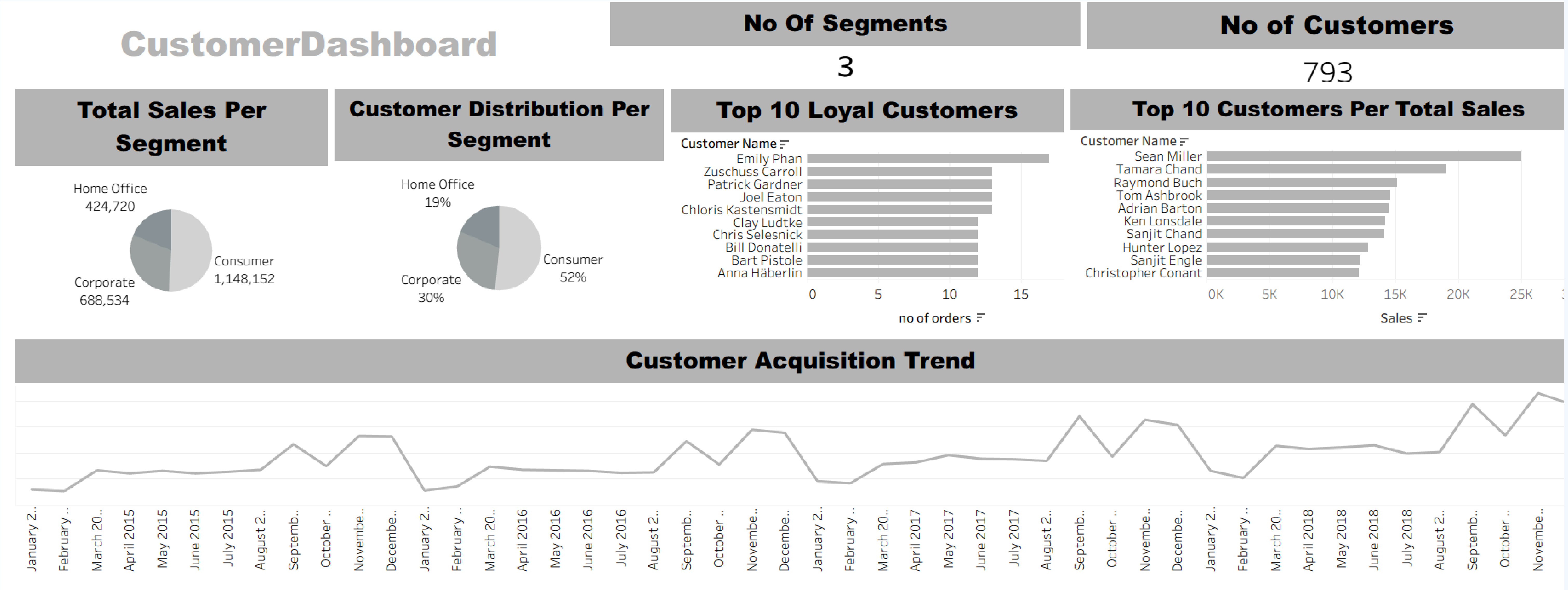
Orders Dashboard

No.of Orders: 4,922

No of Shipping mode: 4

Customers Dashboard

The Customer Dashboard gives us a deeper understanding of our valuable customer base. This view displays the customer distribution across different segments (Corporate, Home Office, Consumer) and identifies our top customers based on total sales, in addition to our most loyal customers in terms of the number of orders



CustomerDashboard

No Of Segments

3

No of Customers

793

Total Sales Per Segment

Customer Distribution Per Segment

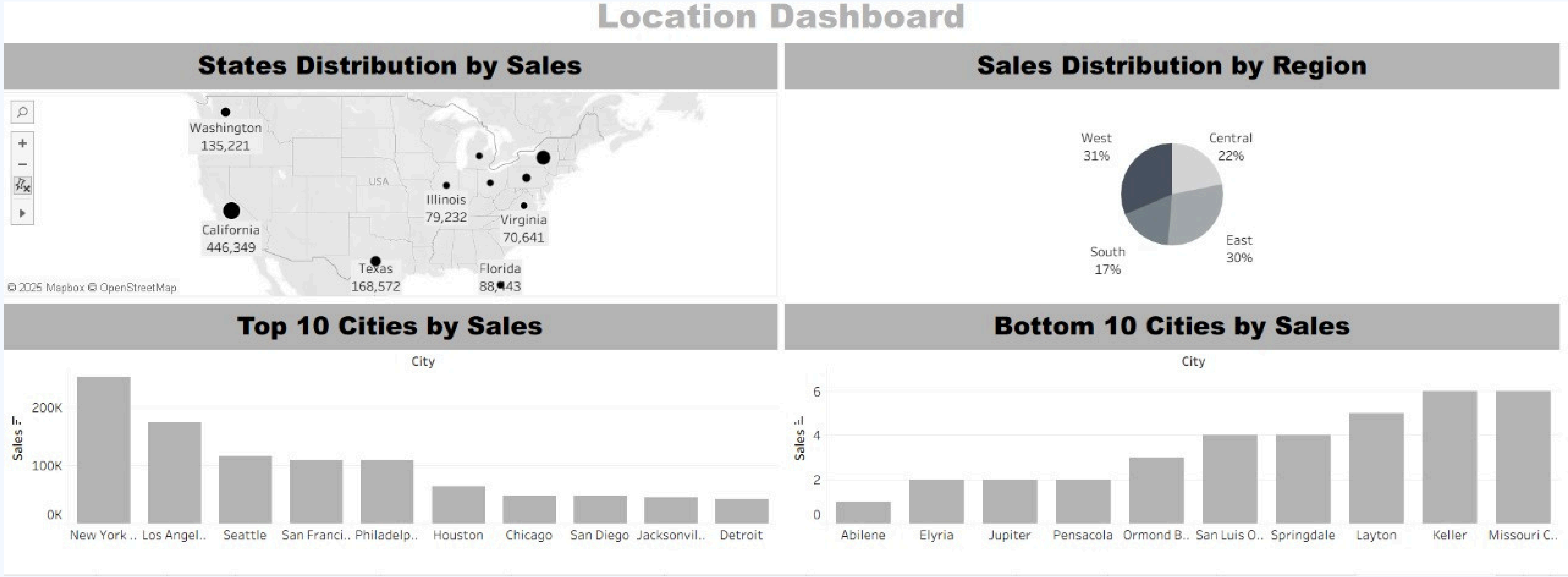
Top 10 Loyal Customers

Top 10 Customers Per Total Sales

Customer Acquisition Trend

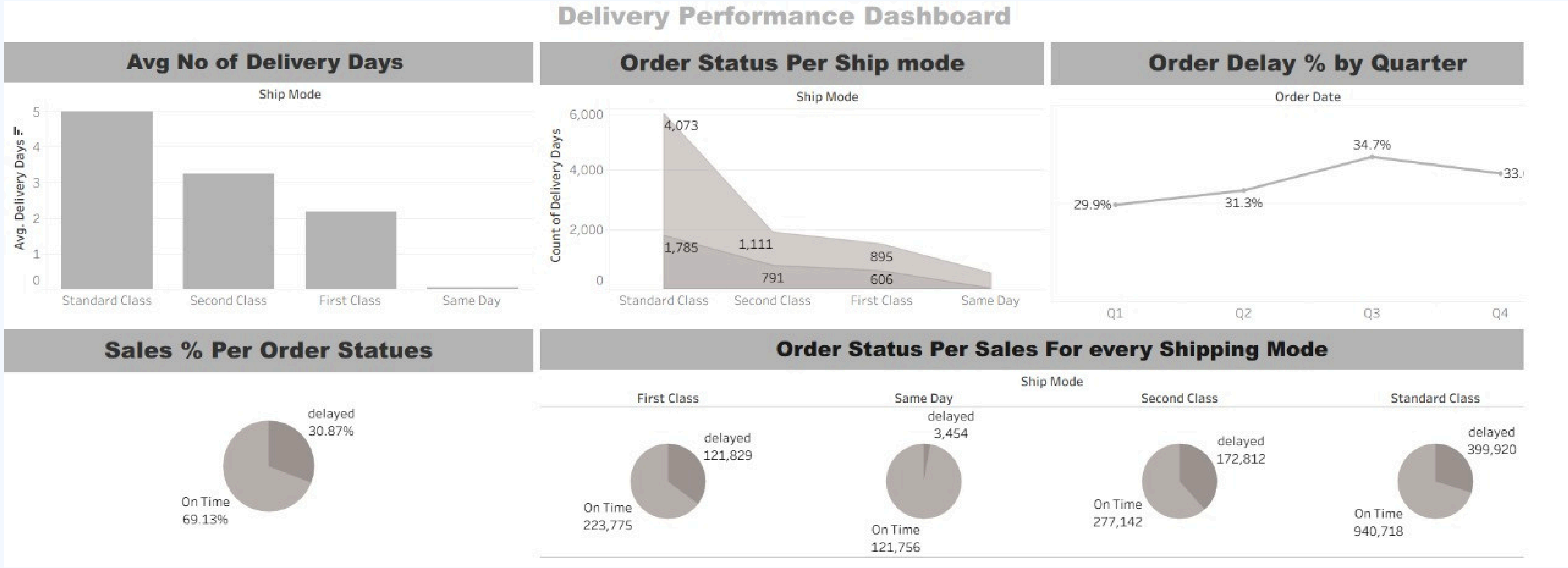
Location Dashboard

the Location Dashboard highlights the geographical dimension of our business. This board shows us where our sales are generated, with a detailed analysis of sales distribution by major regions, states, and the top and bottom 10 cities by sales volume. This analysis helps us guide our expansion and marketing efforts more effectively across different geographies.



Delivery Performance Dashboard

Finally ,we move to the Delivery Performance Dashboard. It focuses on key shipping metrics, such as the average delivery days per ship mode, the overall percentage of delayed versus on-time orders, and the quarterly trend of these delays, enabling us to optimize our supply chains



نص فقرتك

IMPORTANT INSIGHTS

The overall Delayed Order Rate is 30.87%, peaking at 34.7% in Quarter 3.

60% of total orders are shipped via Standard Class, and the West region generates the highest sales at 31%.

The Consumer segment is the largest customer group (52%), yet the highest individual sales are driven by Sean Miller.

CONCLUSION

To finish, our dashboards show that our business is strong and growing. The number of orders increases every year. Our biggest customers are in the Consumer group, and the Technology items sell the most.

However, we need to fix two main problems:

Too many orders are delayed (over 30%).

We depend too much on one shipping type (Standard Class at 60%) and sales from the West area.

We must work on these two points to grow better and faster.

RECOMMENDATIONS

Here are three simple actions we should take:

Fix Delivery Problems:Action: We must quickly lower the delay rate (30.87%).

Reason: We need to find out why orders are slow, especially in Q3 (34.7% delay) and for the most common shipping method, Standard Class.

Grow in New Areas:Action: We need to spend money to sell more in the areas that are currently weak.

Reason: The West area is our best seller, but we need sales to be more even across all regions.

Keep Selling Technology and Keep Top Customers:Action: Keep pushing sales of Technology items.

Reason: Technology is our top sales category (37%). We should also give special offers to our biggest spenders (like Sean Miller) to keep them happy and loyal.

THANK YOU