**Superstore Sales Performance Analysis: A Comprehensive Report**

Noor Tamer, Shahd Ashraf, Ganna Rashad, Menna, Hagar Gamal

Prepared for Data Analytics Course (DEPI)

Supervised by: Eng. Soha

1. **Introduction**

This project analyzes the sales performance of a superstore using Python for data analysis and **Tableau** for visualization. The goal is to uncover key trends in shipping methods, customer segments, regional sales, and product sub-categories to provide actionable insights for improving overall business strategies.

Through careful data cleaning, feature engineering, and a variety of visualizations, we offer a thorough overview of sales performance, highlighting opportunities for improvement and growth.

1. **Data Overview**

The dataset includes several key variables:

• **Sales**: Total sales revenue generated.

• **Shipping Mode**: Different shipping options offered to customers (Standard Class, First Class, Second Class, Same Day).

• **Regions**: Geographical breakdown of sales by East, West, Central, and South regions.

• **Customer Segments**: Customer categories such as Consumer, Corporate, and Home Office.

• **Product Sub-categories**: Items sold, ranging from Phones and Chairs to Accessories and Envelopes.

• **Delivery Time**: Calculated time from order placement to product delivery.

1. **Data Cleaning**

The initial phase of the project was to clean the data to ensure consistency and accuracy. We handled missing values, converted dates, removed irrelevant columns, and eliminated duplicates, setting the foundation for precise analysis.

Here’s the complete code used for data cleaning:

import **pandas** as **pd**

*# load dataset*

df = **pd**.**read\_csv**("Superstore Sales Dataset.csv")

*#impute the missing values with the mode*

df['Postal Code'].**fillna**(df.**mode**, inplace=True)

*#Convert Date Columns to Datetime*

df['Order Date'] = **pd**.**to\_datetime**(df['Order Date’])

df['Ship Date'] = **pd**.**to\_datetime**(df['Ship Date'])

1. **Feature Engineering**

Feature engineering added depth to the analysis by creating new variables. For instance, we calculated **Delivery Time** from the difference between order and shipping dates. This allowed us to measure shipping performance across different modes.

# Create a new column for Delivery Time

df['Delivery Time'] = (df['Ship Date'] **-** df['Order Date']).dt.days

# Group by shipping mode to calculate average delivery time

shipping\_sales = df.**groupby**('Ship Mode')['Sales'].**sum**().**reset\_index**()

# Preview delivery time by shipping mode

**print**(shipping\_sales)

**Insight**: Understanding delivery times across shipping modes revealed that **Same Day** deliveries were the fastest, while **Standard Class** had the longest average delivery time. This insight can inform logistics decisions to improve customer satisfaction.

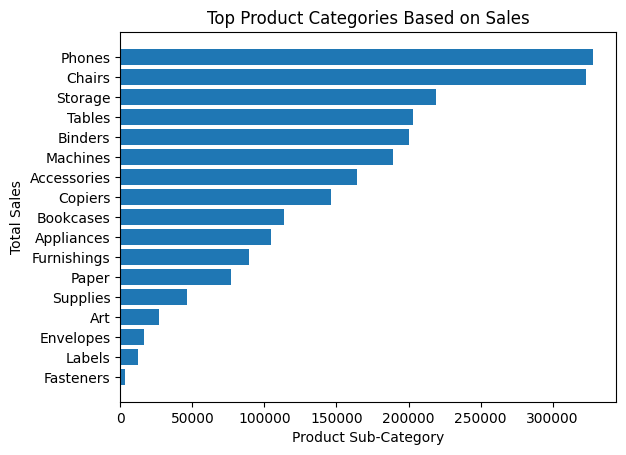
1. **Data Visualization**

Visualization is a critical part of any data analysis, and in this project, we used both **Python** for exploratory visualizations and **Tableau** for building a comprehensive interactive dashboard.

**Exploratory Visualizations with Python:**

1. **Top Product Sub-categories by Sales**

This chart helped identify the top 10 performing product categories by sales, giving insights into where most sales revenue is generated. **Phones**, **Chairs**, and **Storage Supplies** are the top-selling sub-categories, making them critical areas for focus in future business strategies.

Chart:  


Code:

import **matplotlib**.**pyplot** as **plt**

import **seaborn** as **sns**

*# Group the data by product sub category versus the sales*

product\_subcategory = df.**groupby**(['Sub-Category'])['Sales'].**sum**().**reset\_index**()

*# Sort the product category in their descending order and identify top product category*

top\_product\_subcategory = product\_subcategory.**sort\_values**(by='Sales', ascending=False)

*# Print the states*

**print**(top\_product\_subcategory.**reset\_index**(drop=True))

top\_product\_subcategory = top\_product\_subcategory.**sort\_values**(by='Sales', ascending=True)

*# Ploting a bar graph*

**plt**.**barh**(top\_product\_subcategory['Sub-Category'], top\_product\_subcategory['Sales'])

*# Labels*

**plt**.**title**('Top Product Categories Based on Sales')

**plt**.**xlabel**('Product Sub-Category')

**plt**.**ylabel**('Total Sales')

**plt**.**xticks**(rotation=0)

**plt**.**show**()

1. **Average Delivery Time by Shipping Mode**

**Same Day** shipping had the fastest delivery times (less than a day), while **Standard Class** had the longest (5 days). This is crucial for improving shipping logistics and offering better customer satisfaction.

Chart:

A graph of a number of blue rectangular bars

Description automatically generated

Code:

import **matplotlib**.**pyplot** as **plt**

import **seaborn** as **sns**

delivery\_time\_by\_mode = df.**groupby**('Ship Mode')['Delivery Time'].**mean**().**reset\_index**()

**plt**.**figure**(figsize=(10, 6))

**sns**.**barplot**(x='Ship Mode', y='Delivery Time', data=delivery\_time\_by\_mode)

**plt**.**title**('Average Delivery Time by Shipping Mode')

**plt**.**xlabel**('Shipping Mode')

**plt**.**ylabel**('Average Delivery Time (days)')

**plt**.**show**()

1. **Sales by Region**

The bar chart visualizes the total sales by state, focusing on the top 10 performing states. This chart is useful to compare how different states contribute to overall sales. **California** leads in total sales, followed by **New York** and **Texas**. States like **Arizona** and **Kentucky** showed lower sales performance.

Chart:  
A graph of sales

Description automatically generated

Code:

import **matplotlib**.**pyplot** as **plt**

import **seaborn** as **sns**

regional\_sales = df.**groupby**(['Region', 'State'])['Sales'].**sum**().**reset\_index**()

**plt**.**figure**(figsize=(12, 8))

**sns**.**barplot**(x='Sales', y='State', hue='Region', data=regional\_sales)

**plt**.**title**('Sales by State and Region')

**plt**.**xlabel**('Total Sales')

**plt**.**ylabel**('State')

**plt**.**show**()

1. **Sales by Customer Segment and Region**

This bar chart compares sales across different customer segments and regions. The **Consumer** segment drives the highest sales, particularly in the **West** region, suggesting that targeted strategies in this area could further boost revenue.

Chart:  
A graph of sales

Description automatically generated

Code:

import **matplotlib**.**pyplot** as **plt**

import **seaborn** as **sns**

segment\_analysis = df.**groupby**(['Segment', 'Region'])['Sales'].**sum**().**reset\_index**()

**plt**.**figure**(figsize=(12, 8))

**sns**.**barplot**(x='Segment', y='Sales', hue='Region', data=segment\_analysis)

**plt**.**title**('Sales by Customer Segment and Region')

**plt**.**xlabel**('Total Sales')

**plt**.**ylabel**('Customer Segment')

**plt**.**show**()

1. **Correlation between Delivery Time and Sales**

The correlation between **Sales** and **Delivery Time** is near zero, indicating no significant relationship between how much is sold and how long it takes to deliver.

Chart:  
**A red and blue squares

Description automatically generated**

Code:

import **matplotlib**.**pyplot** as **plt**

import **seaborn** as **sns**

*# Correlation matrix*

correlation\_matrix = df[['Sales', 'Delivery Time']].**corr**()

**plt**.**figure**(figsize=(8, 6))

**sns**.**heatmap**(correlation\_matrix, annot=True, cmap='coolwarm')

**plt**.**title**('Correlation between Sales and Delivery Time')

**plt**.**show**()

1. **Total Sales by Product Category**

**Technology** leads in sales, followed by **Office Supplies** and **Furniture**. This suggests a strong market for tech-related products, and potential investment in this category could drive more sales.

Chart:

A graph of blue rectangular shapes

Description automatically generated

Code:

import **matplotlib**.**pyplot** as **plt**

import **seaborn** as **sns**

category\_sales = df.**groupby**('Category')['Sales'].**sum**().**reset\_index**()

**plt**.**figure**(figsize=(10, 6))

**sns**.**barplot**(x='Category', y='Sales', data=category\_sales)

**plt**.**title**('Total Sales by Product Category')

**plt**.**xlabel**('Total Sales')

**plt**.**ylabel**('Product Category')

**plt**.**show**()

1. **Total Sales by Year**

Sales show a steady upward trend, with a notable peak in 2017 and further growth in 2018. This indicates the store has been consistently growing, making it essential to sustain this growth trajectory.

Chart:  
**A graph with a line and a dotted line

Description automatically generated**

import **matplotlib**.**pyplot** as **plt**

import **seaborn** as **sns**

*# Convert the "Order Date" column to datetime format*

df['Order Date'] = **pd**.**to\_datetime**(df['Order Date'], dayfirst=True)

*# Group the data by years and calculate the total sales amount for each year*

yearly\_sales = df.**groupby**(df['Order Date'].dt.year)['Sales'].**sum**()

yearly\_sales = yearly\_sales.**reset\_index**()

yearly\_sales = yearly\_sales.**rename**(columns={'Order Date': 'Year', 'Sales':'Total Sales'})

*# yearly\_sales =*

*# Print the total sales for each year*

**print**(yearly\_sales)

**plt**.**plot**(yearly\_sales['Year'], yearly\_sales['Total Sales'], marker='o', linestyle='-')

**plt**.**xlabel**('Year')

**plt**.**ylabel**('Total Sales')

**plt**.**title**('Total Sales by Year')

*# Display the plot*

**plt**.**tight\_layout**()

**plt**.**show**()

1. **Total Sales in 2018 by Quarter**

The **fourth quarter** shows the highest sales, with a significant spike in December, likely due to holiday promotions and increased consumer spending. The first quarter starts off slower, indicating a seasonal effect on sales performance.

Chart:  
**A graph with a dotted line

Description automatically generated**

import **matplotlib**.**pyplot** as **plt**

import **seaborn** as **sns**

df['Order Date'] = **pd**.**to\_datetime**(df['Order Date'], dayfirst=True)

*# Filter the data for the year 2018*

year\_sales = df[df['Order Date'].dt.year **==** 2018]

*# Calculate the quarterly sales for 2018*

quarterly\_sales = year\_sales.**resample**('Q', on='Order Date')['Sales'].**sum**()

quarterly\_sales = quarterly\_sales.**reset\_index**()

quarterly\_sales = quarterly\_sales.**rename**(columns={'Order Date': 'Quarter', 'Sales':'Total Sales'})

**print**("Quarterly Sales for 2018:")

**print**(quarterly\_sales)

*# Create a line graph for total sales by year*

**plt**.**plot**(quarterly\_sales['Quarter'], quarterly\_sales['Total Sales'], marker='o', linestyle='--')

**plt**.**xlabel**('Year')

**plt**.**ylabel**('Total Sales')

**plt**.**title**('Total Sales in 2018 quarters')

*# Display the plot*

**plt**.**tight\_layout**()

**plt**.**xticks**(rotation=75)

**plt**.**show**()

**Tableau Visualizations:**

We created a **comprehensive dashboard** in Tableau that highlights:

• **Total Sales by Shipping Mode** (with Standard Class dominating).

• **Correlation Between Sales and Delivery Time** (minimal correlation, showing that delivery time doesn’t heavily influence sales).

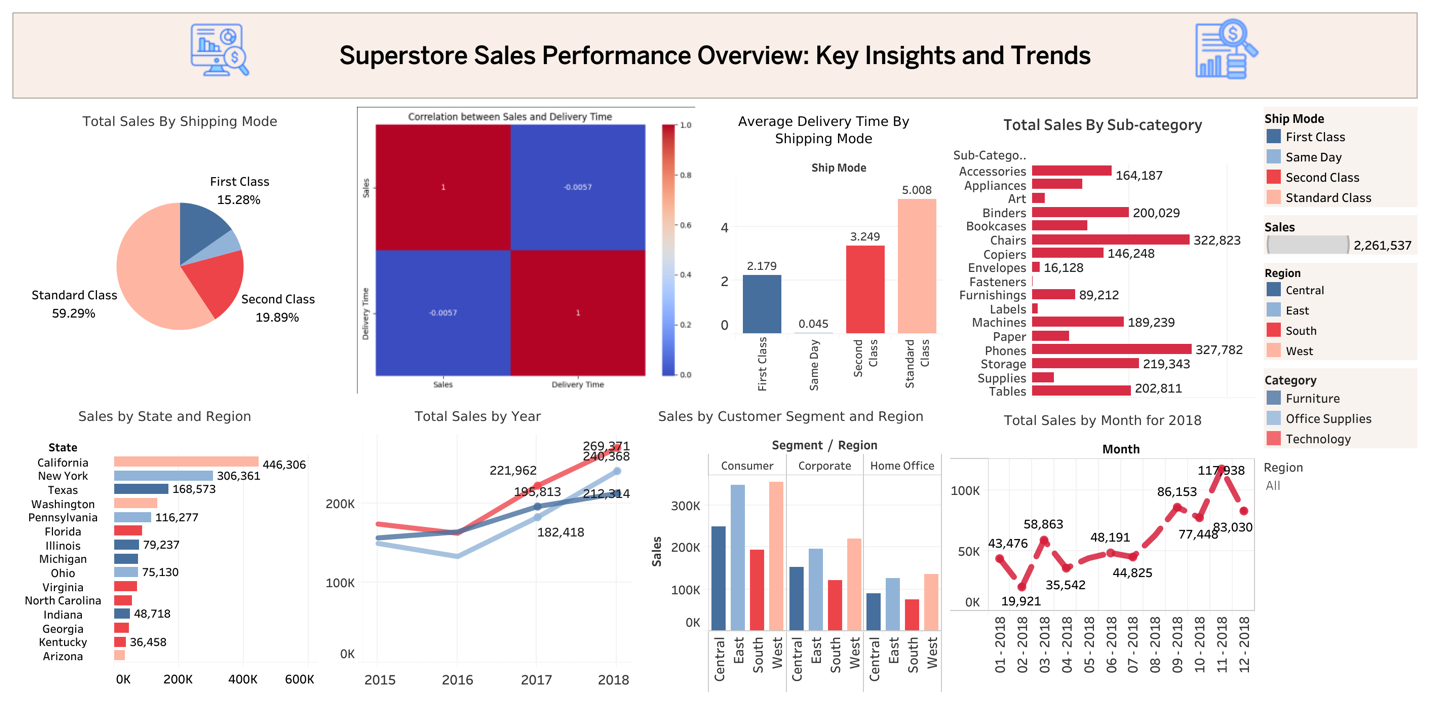
• **Sales by Region and State** (with the West and California leading in sales).

• **Top Sales by Product Sub-categories** (Phones, Chairs, and Storage Supplies).

• **Monthly and Yearly Sales Trends** (steady increase in sales, peaking in December).

**Why Tableau?**

Tableau was chosen for its powerful interactive capabilities, making it easier for stakeholders to explore the data and gain insights quickly. It allows us to drill down into specific regions, product categories, and customer segments for a more granular understanding.



Link for interactive dashboard: <https://public.tableau.com/views/DEPI_GRAD_PROJECT/Dashboard2?:language=en-GB&publish=yes&:sid=&:redirect=auth&:display_count=n&:origin=viz_share_link>

1. **Key Findings**

From our comprehensive analysis:

• **Shipping Mode**: Standard Class accounts for 59.29% of sales, but Same Day shipping offers significantly faster delivery times.

• **Sales by Region**: The West region, particularly California, leads in sales. States like Kentucky and Arizona lag behind, suggesting room for improvement.

• **Product Performance**: Phones, Chairs, and Storage Supplies are the top-selling sub-categories, contributing significantly to overall sales.

• **Customer Segments**: The Consumer segment is the most profitable, with the West region being a strong market for all segments.

1. **Conclusion and Recommendations**

• **Optimize Shipping Options**: Promote Same Day and First Class shipping modes for time-sensitive products, as these offer better delivery times with potential to attract more customers.

• **Expand Marketing in Lower-Performing States**: Target states with lower sales (e.g., Kentucky and Arizona) with localized marketing strategies to increase performance in those regions.

• **Product Prioritization**: Focus inventory and marketing efforts on top-performing product categories like Phones and Chairs to capitalize on high demand.

• **Holiday Campaigns**: December sales show a clear spike, so leveraging holiday promotions and targeted campaigns could further boost end-of-year revenue.