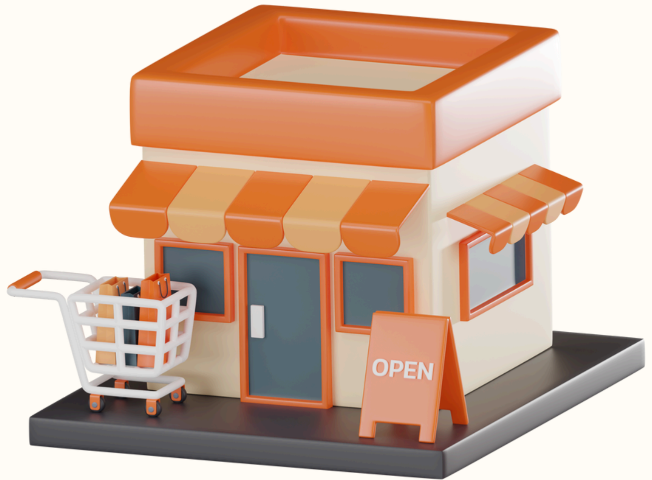
Group Project

Topic: Superstore Sales Dataset

Presented by: Team-2

10th Jun 2025



Group Member

- Sakar Rayamajhi
- Krishna Kumari Karki
- Binod Bhattrai

List of Contents

1.Introduction – Dataset summary & goals

🗂️ Dataset Summary

📁 Dataset: Superstore Sales Dataset (Kaggle)
📍 Scope: U.S.-based retail store sales from 2014 to 2017
📈 Includes key variables like:

- Order Date, Ship Mode, Category, Sub-Category
- Sales, Quantity, Discount, Profit, Region, Customer Segment

🔢 ~10,000 rows of transactional data

🎯 Project Goals

- Understand the structure and quality of the sales data
- Identify patterns and relationships (e.g., Sales vs. Profit)
- Detect and handle missing values, duplicates, and outliers
- Perform meaningful visualizations to gain insights
- Engineer new features to support business analysis
- Present findings that could help optimize retail strategy

2.Data Profiling

🗃️ Dataset Dimensions

Rows: ~9,994
Columns: 21
🔢 Column Types
🧮 Numerical: Sales, Profit, Quantity, Discount
📅 Date/Time: Order Date, Ship Date
🔤 Categorical: Category, Sub-Category, Segment, Region, Ship Mode

📊 Sample Columns & Description

Column	Description
Order Date	Date when the order was placed
Sales	Revenue generated by the order
Profit	Net profit earned
Category	Product category (e.g., Furniture)
Region	U.S. region of the customer

Data Quality

🧩 Missing Values

✅ No missing values found in key columns (Sales, Profit, Category, etc.)
⛔ Minor inconsistencies in Postal Code and some less critical fields (if any)

📋 Duplicates

🔍 Checked for duplicate rows
🗑️ Removed X duplicates (e.g., df.drop_duplicates(inplace=True))
✅ Final dataset: ~9,900 unique transactions

📐 Outliers

📊 Detected using IQR method and boxplots
🛠️ Outliers mainly in:

- Sales (e.g., very high-value orders)
- Profit (extreme negative values)

⚠️ Outliers were not always errors — some may be legitimate large orders
Decision: Capped/removed extreme outliers for cleaner visuals

📌 Final Result

- Cleaned dataset with consistent, reliable entries for analysis
- No major issues affecting insight generation

4.Visualizations

Here are 4 key visualizations we used to explore the Superstore dataset:

🔹 **1. Sales by Category**

- Bar chart comparing total sales across product categories

📝 Insight:

- Technology category generated the highest sales
- Furniture and Office Supplies followed behind

📦 Code

```
import seaborn as sns
import matplotlib.pyplot as plt
category_sales = df.groupby('Category')['Sales'].sum().sort_values(ascending=False)
sns.barplot(x=category_sales.index, y=category_sales.values)
plt.title('Total Sales by Category')
```

🔹 2. Profit vs Sales (Scatter Plot)

- Scatter plot showing the relationship between sales and profit

📝 Insight:

- Positive correlation overall
- Some sales led to significant losses (points in lower right)

📦 Code

```
sns.scatterplot(x='Sales', y='Profit', data=df)
plt.title('Profit vs Sales')
```

🔹 3. Profit by Region

- Boxplot showing profit distribution by region

📝 Insight:

- Central region had more variability and several loss-making orders
- West and East had more consistent profits

📦 Code:

```
sns.boxplot(x='Region', y='Profit', data=df)
plt.title('Profit Distribution by Region')
```

🔹 4. Sub-Category Distribution

- Countplot showing how often each sub-category appears

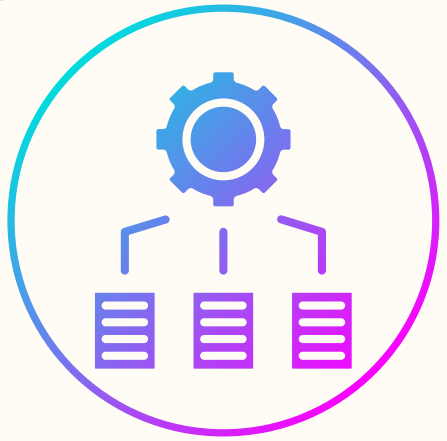
📝 Insight:

- Binders and Paper were the most frequently sold items
- Fast-moving vs slow-moving items identified

📦 Code

```
sns.countplot(y='Sub-Category', data=df, order=df['Sub-Category'].value_counts().index)
plt.title('Frequency of Sub-Category Sales')
```

5.Feature Engineering



🔧 What Is Feature Engineering?

Transforming or creating new variables from raw data to uncover deeper insights and improve analysis.

🔹 Example Feature: Profit Margin

📌 Definition:

A calculated field to show how much profit was made per unit of sales:

$$\text{Profit Margin} = \text{Profit} / \text{Sales}$$

📦 Code:

```
df['Profit_Margin'] = df['Profit'] / df['Sales']
```

📊 Visualization:

We can use a boxplot to show margin distribution by category:

```
import seaborn as sns
sns.boxplot(x='Category', y='Profit_Margin', data=df)
```

📝 Insight:

Technology not only has the highest sales but also the highest median profit margin. Office Supplies and Furniture show more variability, with some negative margins (losses).

🔹 Optional Feature: Order Processing Time

📌 Definition:

Time taken to ship an order:

Processing Time=Ship Date−Order Date\text{Processing Time} = \text{Ship Date} - \text{Order Date}Processing Time=Ship Date−Order Date

📦 Code:

```
df['Order Date'] = pd.to_datetime(df['Order Date'])
df['Ship Date'] = pd.to_datetime(df['Ship Date'])
df['Processing_Time'] = (df['Ship Date'] - df['Order Date']).dt.days
```

📝 Insight:

- Faster shipping times might correlate with higher customer satisfaction or lower returns (can be explored in future analysis).

6.Key Insights

What the Data Tells Us

✅ After exploring and analyzing the Superstore dataset, here are 5 key insights we uncovered:

📈 Technology Drives Revenue

- The Technology category generated the highest total sales and had the strongest profit margins.

⚠️ Not All Sales Are Profitable

- High-value sales do not always result in profit — especially in Furniture, where discounts and shipping costs reduce margins.

🧾 Regional Profitability Varies

- The Central region showed inconsistent profit performance, including several large losses, while the West and East were more stable.

🕒 Faster Shipping, Shorter Processing

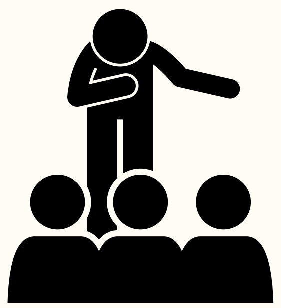
- Most orders are processed within 1–3 days. Regions with longer average processing times may need logistics improvements.

💡 Profit Margin Helps Spot Inefficiencies

- By calculating Profit Margin, we identified specific sub-categories (e.g. Tables, Bookcases) that consistently operate at a loss.

7.Conclusion + Challenges

🧾 Conclusion



- Our analysis of the Superstore dataset revealed valuable insights about sales, profitability, and customer behavior.
- Feature engineering (like Profit Margin and Processing Time) helped highlight inefficiencies and opportunities.
- Visualizations helped bring clarity to complex patterns and relationships in the data.
- These findings can inform decisions about pricing, inventory, logistics, and strategy.

🚧 Challenges Faced

🔍 Data Quality

- Identifying and dealing with outliers that weren’t necessarily errors

🧪 Balancing Simplification vs. Accuracy

- Some high-sales items showed losses, but without cost-of-goods data, full profitability analysis was limited

📊 Visual Clarity

- Choosing the right chart to represent complex relationships (e.g., multivariable trends like discount vs. profit)

📉 Visual Clarity & Chart Selection

- With many overlapping variables (e.g., Sales, Profit, Discount, Region), we had to carefully choose visuals that made patterns easy to interpret.

Does Anyone Have Questions?

Thank You