Superstore Sales Data: Exploratory Data Analysis

(Report)

1. Data Loading and Initial Exploration:

- The dataset was loaded into a pandas DataFrame.
- Initially it was inspected that dataset has 10,014 rows and 21 columns
- Further it was observed that ‘Ship Mode’ and ‘Quantity’ columns had null values
- Some statistical information regarding numerical columns was inspected
- Duplicates were found on the basis of unique Order ID’s i.e only 5009 were unique and many records were duplicated.
- It was observed that ‘Sales Price’,’Profit’ and ‘Discount’ has right skewed distributions.

2. Handling Duplicates:

- Duplicates records were identified on basis of Order ID’s and removed
- 5005 rows were affected while performing this operation

3. Date Handling:

- Order Date and Ship Date columns were converted to **datetime format** for consistency.
- Extracted year from Order ID and compared with Order Date
- **16 mismatched rows** were corrected by aligning Order Date year with Order ID year.
- Checked for and confirmed there were **no cases** where Ship Date was earlier than Order Date.

4. Imputation of Missing Values:

- Calculated Days to Ship = Ship Date - Order Date.
- If Days to Ship <= 6 → Ship Mode set to **"Same Day”**.
- If Days to Ship > 6 and Days to Ship <= 28 → set to “First Class”.
- If Days to Ship > 28 and Days to Ship <= 34 → set to “Second Class”.
- If Days to Ship > 34 → set to “Standard Class".
- For missing Quantity, median was selected as imputation strategy based on data distribution.

5. Data Masking and String Handling:

- Dropped the Customer Name column to protect **PII**.
- Created a new Customer Name Masked column with initials only.

- Converted Postal Code to **string format** and padded to ensure 5-character formatting.

6. Data Type Conversion:

- Converted Quantity to int, and Sales Price to float, ensuring correct data types for calculations.

7. Handling Inconsistent Categorical Data:

- Cleaned State column by replacing abbreviations (e.g., "CA") with full names (e.g., "California").

8. Feature Engineering:

- Created new columns:
- Original Price = Sales Price / (1 - Discount)
- Total Sales = Sales Price * Quantity
- Total Profit = Profit * Quantity
- Discount Price = Original Price * Discount
- Total Discount = Discount Price * Quantity
- Shipping Urgency based on Days to Ship:
- 0 days → “Immediate"
- 1–3 days → “Urgent"
- 3 days → "Standard"

- Calculated Days Since Last Order.
- Created a customer-level aggregation dataset (Total Sales, Quantity, Discount) and merged it with the original dataset.

9. Outlier Detection and Handling

- Created a function remove_outliers(df, col) using the **3 × IQR** method:
- Removes only **extreme values** to retain valid data while mitigating noise.
- Applied the function on:
- Sales Price
- Profit

10. Customer Segmentation and Analysis

- Computed Customer Sales Quintile and Customer Profit Quintile using **pd.qcut( )** method in pandas library.
- Created a **cross-tabulation** to explore the relationship between sales and profitability segments using **pd.crosstab( )** method.

11. Final Analysis and Visualisations:

Sales and Profit Analysis

- Top 10 Profitable Products: Products like Canon ImageCLASS Printer, Logitech Keyboard, and Apple AirPort Time Capsule emerged as top profit-generators.
- Top 10 Loss-Making Products: Items such as Bush Somerset Bookcase, Hon 5400 Chair, and large office furniture units consistently incurred losses.
- The scatter plot with regression line revealed a positive correlation between Sales and Profit, but with noticeable outliers showing some high-sales items still led to losses.
- A joint plot helped visualise dense clusters of products generating moderate profits.

Customer Insights

- Created a Heatmap to evaluate relationship between customers' sales and profit quintiles
- A strong linear pattern was observed between high sales and high profit but sometimes moderate sales also lead to high profits
- Created a pivot table to understand how different product categories perform across customer segments
- The most profitable combinations are [(Consumer - Office Supplies), (Corporate - Office Supplies),(Home Office - Office Supplies)] and the least profit making combinations are [(consumer - Technology), (Corporate - Technology), (Home Office - Technology)].

Shipping and Delivery Analysis

- Plotted a Pie chart to visualise the distribution of orders by Shipping Urgency
- Plotted violin chart to explore the distribution of Profit across different Days to Ship categories
- The violin plot suggested that profit distribution does not depend much on shipping days as it is almost same for the all three categories
- Created multiple bar charts to visualise and compare the profitability of different shipping modes
- On basis of Average profit Second Class was observed to be most profitable amongst all and Standard Class contributed the maximum profit amongst all
- Profitability was also analysed by categorising further on the basis of segments. To analyse profitable classes amongst all segments

- Pivot table was created to find out the most preferred and profitable Ship Mode amongst all regions, which came out to be “Standard Class” for all regions

Regional Sales and Profitability

- Created Bar Charts to analyse and visualise total sales and profit region wise
- West was observed to have maximum sales as well as profit
- Further Total Sales and Profit was analysed sub-divided on basis of Shipping mode.
- “Standard Class” was observed to be most profitable in all regions
- State-wise pivot table analysis highlighted the most and least profitable states.
- Correlation analysis (after encoding states) showed regional profitability trends.

Pricing and Discount Strategy

- Discount vs. Profitability analysed via scatter plot + trend line.
- Original Price vs. Discounted Price compared using line plots across product categories.

Temporal Trends

- Time series plots tracked Sales and Profit over months/years, revealing seasonal spikes.
- Line/bar charts visualised Order Frequency by Month.
- Year-over-year growth chart showed trends in sales and profit over time.