**Exploratory Data Analysis & Sales Performance Analysis Report**

**Project 1: General Exploratory Data Analysis (EDA)**

**1. Introduction**

The objective of this project is to perform an in-depth exploratory data analysis (EDA) on the given dataset (Superstore Sales Data). The analysis aims to identify trends, patterns, anomalies, and relationships among various features to derive meaningful business insights.

**2. Data Cleaning & Preparation**

* **Missing Values:** Checked and handled appropriately.
* **Duplicate Removal:** Ensured the dataset does not contain redundant entries.
* **Outlier Detection:** Used the IQR method to remove extreme values in Sales and Profit.
* **Data Type Verification:** Converted categorical and date columns into appropriate formats.

**3. Statistical Analysis**

* **Descriptive Statistics:**
  + Mean, median, standard deviation, and variance were calculated for Sales, Profit, and Discount.
  + Distribution of numerical variables was explored to identify skewness.
* **Correlation Analysis:**
  + A heatmap was used to examine relationships between numerical features.

**4. Data Visualization & Insights**

* **Sales Distribution:** Histogram showed a right-skewed distribution with a majority of sales in the lower range.
* **Profit Boxplot:** Revealed the presence of outliers and variation in profit margins.
* **Correlation Heatmap:** Highlighted a positive correlation between Sales and Profit, and a negative correlation between Discount and Profit.

**5. Conclusion**

The analysis provided key insights into the dataset, including sales distribution, profit trends, and correlations. The findings suggest that discounts negatively impact profits, and sales trends need further exploration for business strategy optimization.

**Project 2: Sales Performance Analysis**

**1. Introduction**

The objective of this project is to analyze sales performance trends, study relationships between key variables, and build a predictive model to forecast sales based on profit and discount values.

**2. Data Cleaning & Preparation**

* **Missing Values:** Checked and handled.
* **Duplicate Entries:** Removed.
* **Date Conversion:** Converted Order.Date into a datetime format for time series analysis.

**3. Exploratory Data Analysis (EDA)**

* **Sales Trends Over Time:**
  + A time series plot showed seasonal trends in sales.
* **Profit vs. Discount:**
  + A scatter plot revealed that higher discounts generally led to lower profits.
* **Sales by Region & Category:**
  + Bar charts highlighted top-performing regions and product categories.

**4. Predictive Modeling**

* **Model Used:** Linear Regression.
* **Features:** Profit, Discount.
* **Target Variable:** Sales.
* **Performance Evaluation:**
  + **R² Score:** Indicates how well the model explains variance in sales.
  + **Mean Squared Error (MSE):** Measures prediction accuracy.

**5. Key Findings & Business Insights**

* **Sales Trends:** Sales fluctuate seasonally, indicating potential peak periods.
* **Discount Impact:** Higher discounts tend to reduce profits.
* **Top Regions & Categories:** Certain regions and product categories contribute significantly to total sales.

**6. Conclusion & Recommendations**

* **Strategic Pricing:** Reduce excessive discounts to improve profitability.
* **Targeted Marketing:** Focus on high-performing regions and categories.
* **Stock Planning:** Predict demand during peak seasons based on historical sales trends.

**Final Deliverables**

Cleaned dataset with missing values handled and outliers removed.  
Statistical summary and correlation analysis.  
Data visualizations (histograms, boxplots, heatmaps, time series, scatter plots, bar charts).  
Linear Regression model for sales prediction with performance evaluation.  
Business recommendations based on insights.

**Task 1 results**

Category City Country Customer.ID Customer.Name \

0 Office Supplies Los Angeles United States LS-172304 Lycoris Saunders

1 Office Supplies Los Angeles United States MV-174854 Mark Van Huff

2 Office Supplies Los Angeles United States CS-121304 Chad Sievert

3 Office Supplies Los Angeles United States CS-121304 Chad Sievert

4 Office Supplies Los Angeles United States AP-109154 Arthur Prichep

Discount Market 记录数 Order.Date Order.ID ... Sales \

0 0.0 US 1 2011-01-07 00:00:00.000 CA-2011-130813 ... 19

1 0.0 US 1 2011-01-21 00:00:00.000 CA-2011-148614 ... 19

2 0.0 US 1 2011-08-05 00:00:00.000 CA-2011-118962 ... 21

3 0.0 US 1 2011-08-05 00:00:00.000 CA-2011-118962 ... 111

4 0.0 US 1 2011-09-29 00:00:00.000 CA-2011-146969 ... 6

Segment Ship.Date Ship.Mode Shipping.Cost \

0 Consumer 2011-01-09 00:00:00.000 Second Class 4.37

1 Consumer 2011-01-26 00:00:00.000 Standard Class 0.94

2 Consumer 2011-08-09 00:00:00.000 Standard Class 1.81

3 Consumer 2011-08-09 00:00:00.000 Standard Class 4.59

4 Consumer 2011-10-03 00:00:00.000 Standard Class 1.32

State Sub.Category Year Market2 weeknum

0 California Paper 2011 North America 2

1 California Paper 2011 North America 4

2 California Paper 2011 North America 32

3 California Paper 2011 North America 32

4 California Paper 2011 North America 40

[5 rows x 27 columns]

Missing Values:

Category 0

City 0

Country 0

Customer.ID 0

Customer.Name 0

Discount 0

Market 0

记录数 0

Order.Date 0

Order.ID 0

Order.Priority 0

Product.ID 0

Product.Name 0

Profit 0

Quantity 0

Region 0

Row.ID 0

Sales 0

Segment 0

Ship.Date 0

Ship.Mode 0

Shipping.Cost 0

State 0

Sub.Category 0

Year 0

Market2 0

weeknum 0

dtype: int64

Statistical Summary:

Sales Profit Discount

count 40298.000000 40298.000000 40298.000000

mean 97.511440 12.675087 0.130881

std 107.685222 25.008335 0.203628

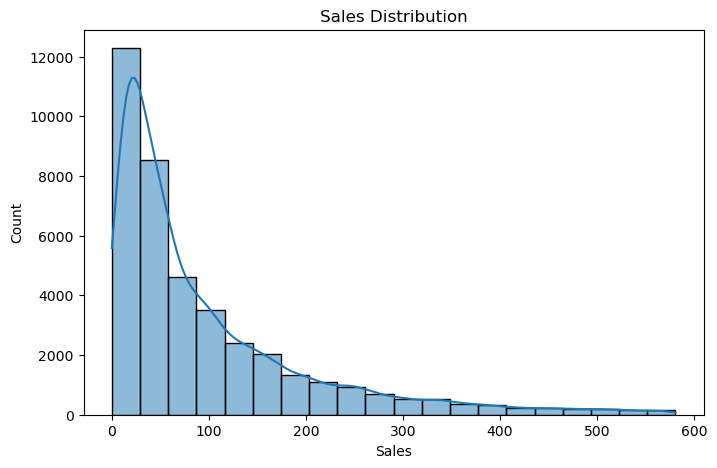
min 0.000000 -55.125000 0.000000

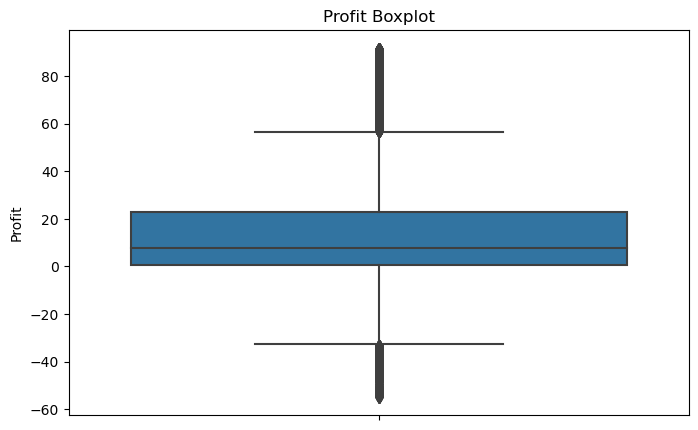
25% 24.000000 0.666000 0.000000

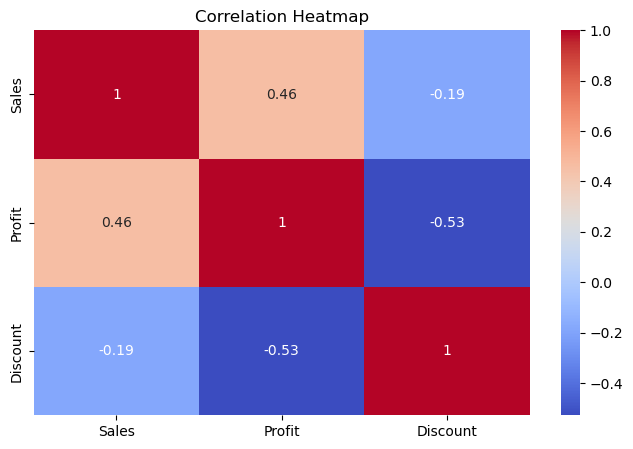
50% 56.000000 7.740000 0.000000

75% 131.000000 22.966000 0.200000

max 581.000000 91.995400 0.800000







**Task 2 results**

Category City Country Customer.ID Customer.Name \

0 Office Supplies Los Angeles United States LS-172304 Lycoris Saunders

1 Office Supplies Los Angeles United States MV-174854 Mark Van Huff

2 Office Supplies Los Angeles United States CS-121304 Chad Sievert

3 Office Supplies Los Angeles United States CS-121304 Chad Sievert

4 Office Supplies Los Angeles United States AP-109154 Arthur Prichep

Discount Market 记录数 Order.Date Order.ID ... Sales \

0 0.0 US 1 2011-01-07 00:00:00.000 CA-2011-130813 ... 19

1 0.0 US 1 2011-01-21 00:00:00.000 CA-2011-148614 ... 19

2 0.0 US 1 2011-08-05 00:00:00.000 CA-2011-118962 ... 21

3 0.0 US 1 2011-08-05 00:00:00.000 CA-2011-118962 ... 111

4 0.0 US 1 2011-09-29 00:00:00.000 CA-2011-146969 ... 6

Segment Ship.Date Ship.Mode Shipping.Cost \

0 Consumer 2011-01-09 00:00:00.000 Second Class 4.37

1 Consumer 2011-01-26 00:00:00.000 Standard Class 0.94

2 Consumer 2011-08-09 00:00:00.000 Standard Class 1.81

3 Consumer 2011-08-09 00:00:00.000 Standard Class 4.59

4 Consumer 2011-10-03 00:00:00.000 Standard Class 1.32

State Sub.Category Year Market2 weeknum

0 California Paper 2011 North America 2

1 California Paper 2011 North America 4

2 California Paper 2011 North America 32

3 California Paper 2011 North America 32

4 California Paper 2011 North America 40

[5 rows x 27 columns]

Missing Values:

Category 0

City 0

Country 0

Customer.ID 0

Customer.Name 0

Discount 0

Market 0

记录数 0

Order.Date 0

Order.ID 0

Order.Priority 0

Product.ID 0

Product.Name 0

Profit 0

Quantity 0

Region 0

Row.ID 0

Sales 0

Segment 0

Ship.Date 0

Ship.Mode 0

Shipping.Cost 0

State 0

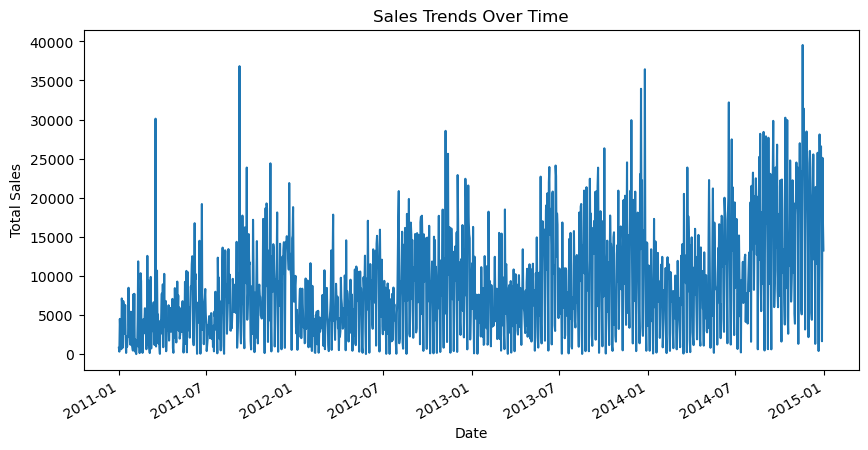
Sub.Category 0

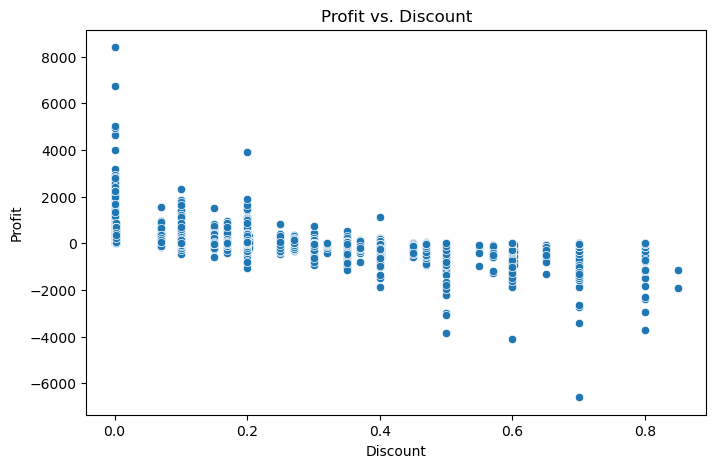
Year 0

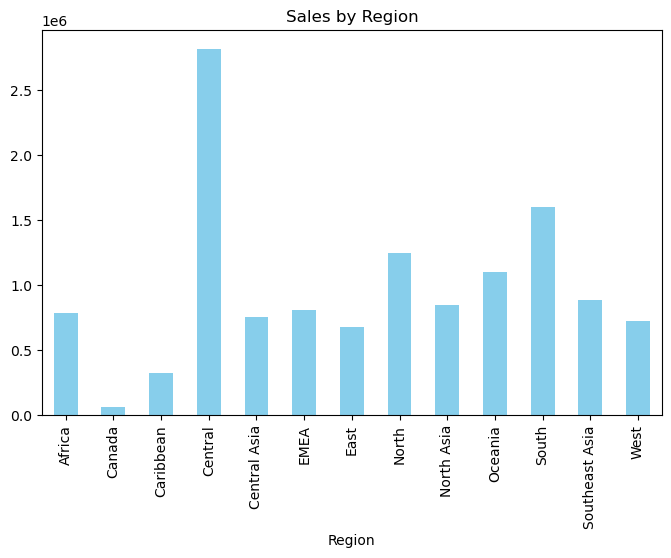
Market2 0

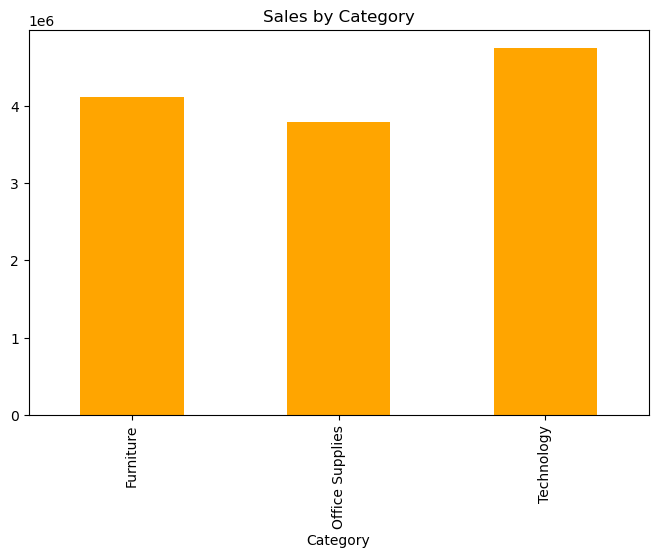
weeknum 0

dtype: int64









Model Performance:

R2 Score: 0.11428823787896558

Mean Squared Error: 241642.6163023337