## **Sales Data Analysis Report**

### **📌 Objective**

The goal of this project was to perform a detailed sales data analysis for a commercial store to uncover patterns, identify trends, and derive actionable insights using historical sales data.

## **🛠 Tools and Libraries Used**

* **Python**: For data manipulation and analysis
* **Pandas**: Data loading, preprocessing, and manipulation
* **Matplotlib & Seaborn**: For visualizing trends and patterns
* **Numpy**: For mathematical operations
* **Scipy**: For statistical analysis
* **Statsmodels**: For time series analysis

## **🗂 Data Overview**

* The dataset contained information on customer demographics, sales revenue, product details, and geographical data.
* The initial dataset had 26 columns and 34,866 rows.

### **Key Columns**

* **Date, Month**: Time-based data for time series analysis
* **Customer Age, Gender**: Demographic information
* **Country, State**: Geographical location data
* **Product Category, Sub Category**: Product information
* **Quantity, Unit Cost, Unit Price, Revenue**: Sales data

## **🧹 Data Cleaning**

### **Steps Taken:**

1. **Handled Missing Values**:
   * One empty row was removed using .drop().
2. **Dropped Irrelevant Columns**:
   * A column without useful data was removed.
3. **Corrected Data Types**:
   * The "Year" column was converted from float to int.
4. **Handled Duplicates**:
   * Verified and confirmed no duplicate rows.
5. **Whitespace and Case Correction**:
   * Cleaned string columns using .str.strip() and .str.lower().

## **🧪 Feature Engineering**

### **New Features Created:**

* **Profit**: Calculated using Revenue - Cost.
* **Revenue per Unit**: Revenue / Quantity.
* **Profit Margin**: (Profit / Revenue) \* 100.
* **Day of the Week & Week of the Month**: Extracted from the "Date" column to detect seasonal patterns.

**Why?** These new features helped in understanding product profitability, customer purchase behavior, and time-based sales trends.

## **📊 Exploratory Data Analysis (EDA)**

### **✅ Visualization Techniques Used:**

* **Boxplots**: For detecting and visualizing outliers.
* **Histograms and KDE Plots**: To observe the distribution of numerical variables.
* **Pie Charts**: For gender distribution and country-wise sales.
* **Bar Charts**: For analyzing top products, states, and age groups.
* **Line Charts**: For visualizing sales trends over time.
* **Correlation Heatmap**: To identify relationships between variables.

## **⚙️ Outlier Detection and Treatment**

* **Boxplots and IQR (Interquartile Range)**: Used to detect outliers.
* **Z-Score**: Applied for more precise outlier detection.
* **Log Transformation**: Applied to normalize skewed data, reduce the effect of extreme outliers, and visualize the data effectively.

**Decision**:

* Outliers representing data errors or extreme anomalies were removed.
* Genuine business outliers (e.g., unusually high sales) were retained for further insights.

## **📅 Time Series Analysis**

* **Seasonal Patterns**: Detected using line charts and decomposition.
* **Monthly and Weekly Trends**: Sales were found to peak in **December** and **June** with a significant drop in **August**.
* **Post-Holiday Dip**: Sales decline in **January** and **February**, indicating a post-holiday effect.

## **🔎 Key Insights**

1. **Geographical Insights**
   * **Germany** and the **U.S.** generated the highest revenue and profit.
   * **California** and **England** were the most profitable states.
2. **Product Analysis**
   * **Mountain Bikes** and **Road Bikes** contributed the most to revenue.
   * Profits were highly dependent on product category, with **Bike Accessories** offering smaller margins.
3. **Customer Demographics**
   * The **26-35** age group made the most purchases.
   * Gender distribution was fairly balanced.
4. **Time-Based Insights**
   * Sales followed a strong seasonal pattern with peaks during summer and winter holidays.
   * A sharp revenue drop in **August** suggested external factors like reduced demand or supply chain issues.

## **📈 Challenges Faced**

* **Data Quality**: Presence of missing and inconsistent values.
* **Outlier Management**: Deciding whether to remove or keep outliers.
* **Business Context**: Limited understanding of real-world factors influencing revenue drop.

**Solution**:

* Cleaned data effectively using Pandas.
* Applied domain-agnostic decisions based on visualization insights.

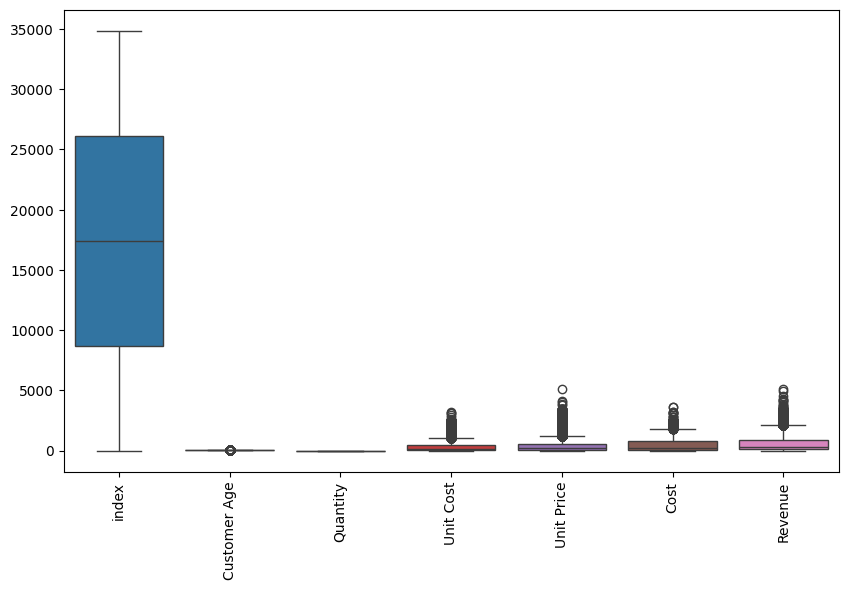
## **🛠 Recommendations**

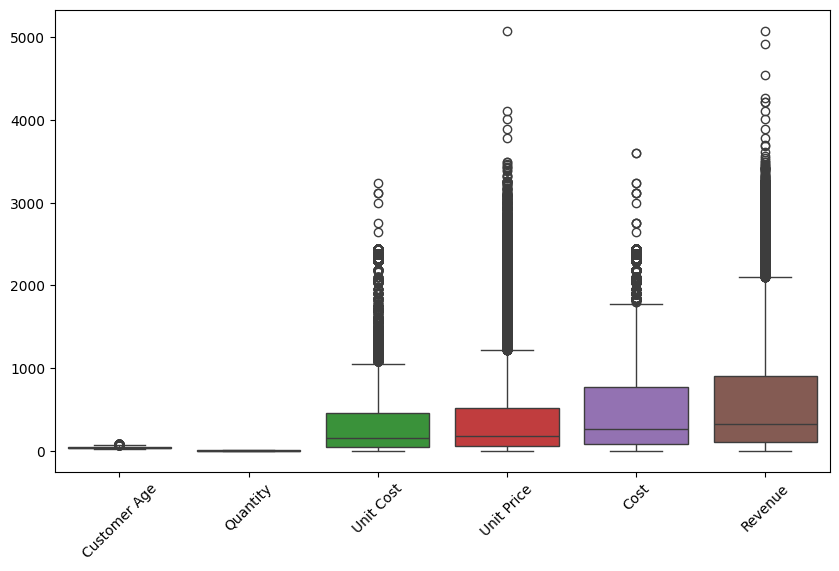
1. **Manage Seasonal Trends**:
   * Plan marketing campaigns before peak months like **June** and **December**.
2. **Investigate Revenue Drops**:
   * Conduct surveys or investigate supply chain issues for **August**.
3. **Focus on Profitable Products**:
   * Prioritize production and marketing of **Mountain Bikes** and **Road Bikes**.
4. **Customer Retention**:
   * Develop loyalty programs targeting the **26-35** age group.

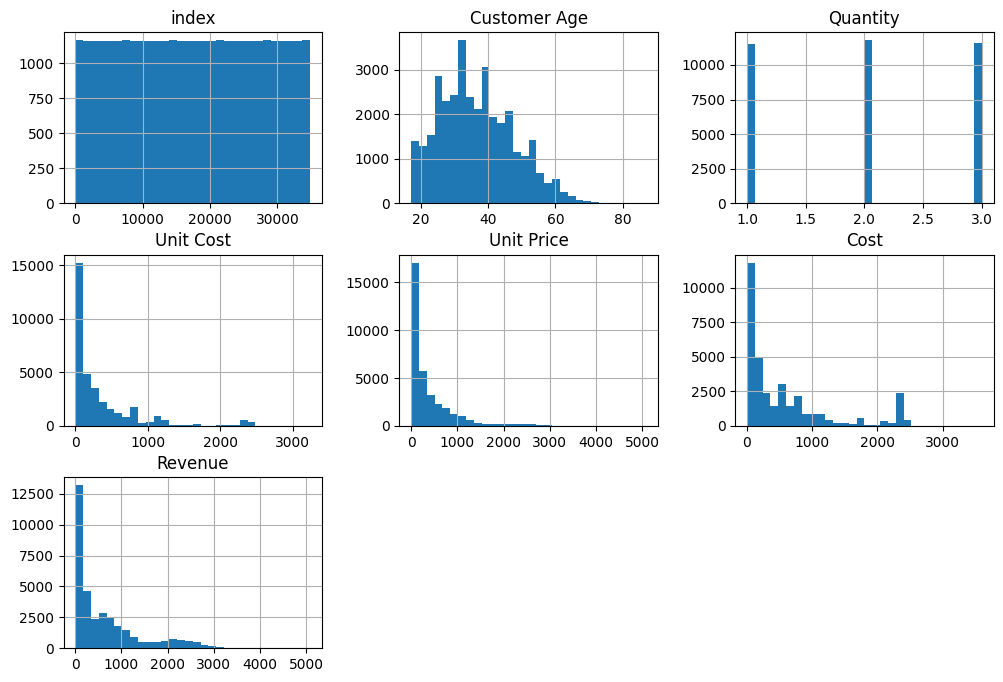
## **🏁 Conclusion**

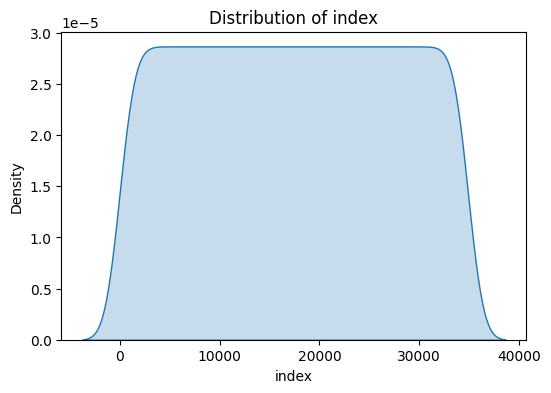
The sales data analysis successfully identified key patterns and insights, providing actionable recommendations. By applying effective data cleaning, visualization, and analysis techniques, we derived meaningful business insights to support strategic decision-making.

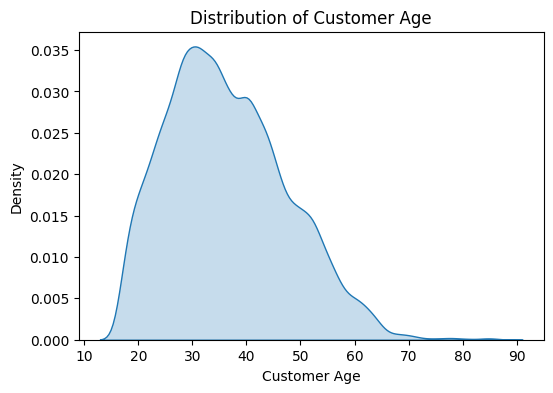
**Graphs and Charts Generated to Draw Conclusion:-**

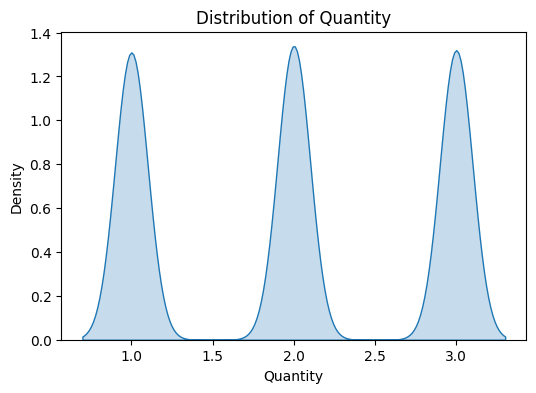
****

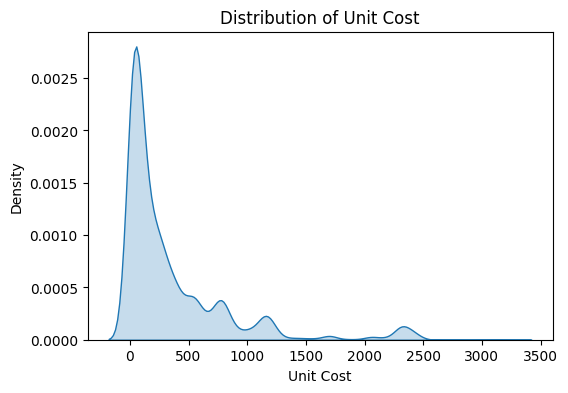
****

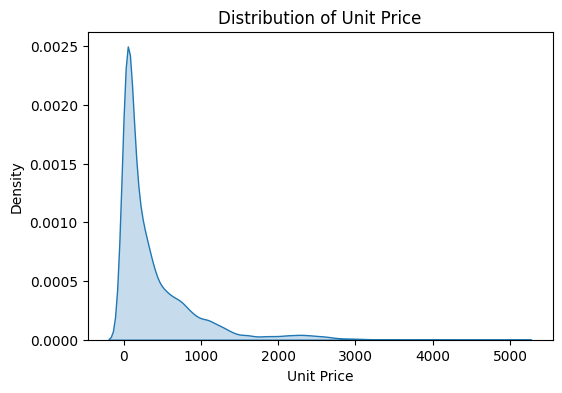
****

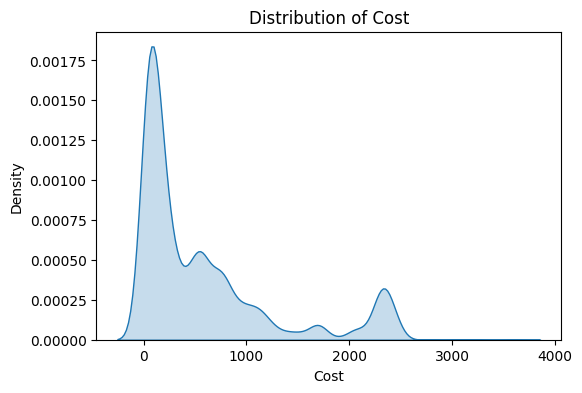
****

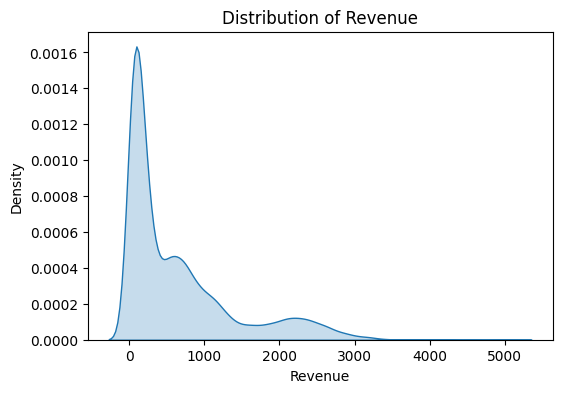
****

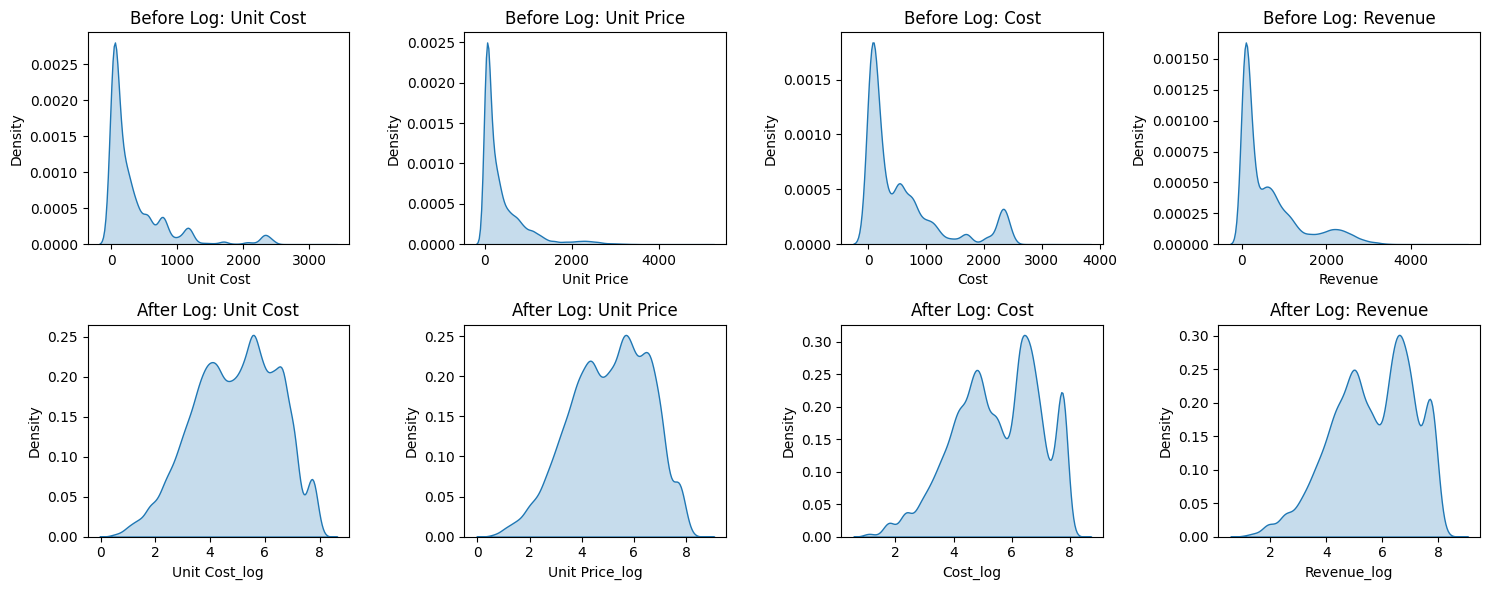
****

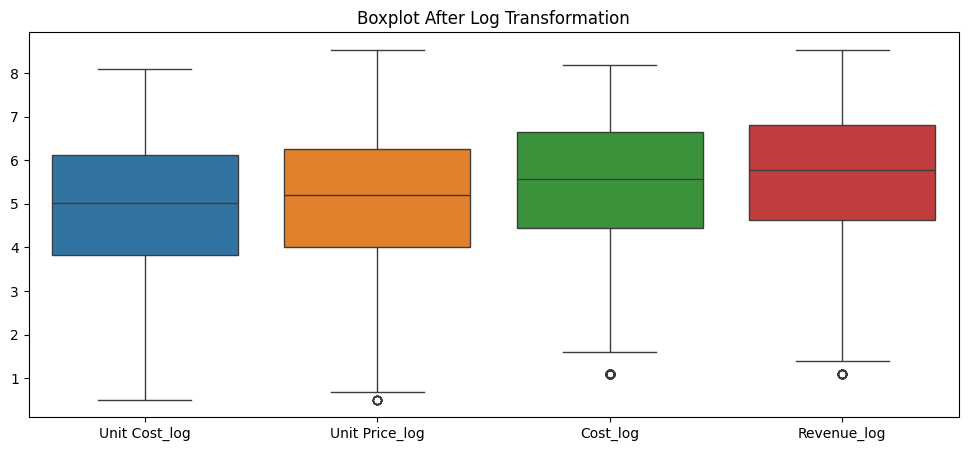
****

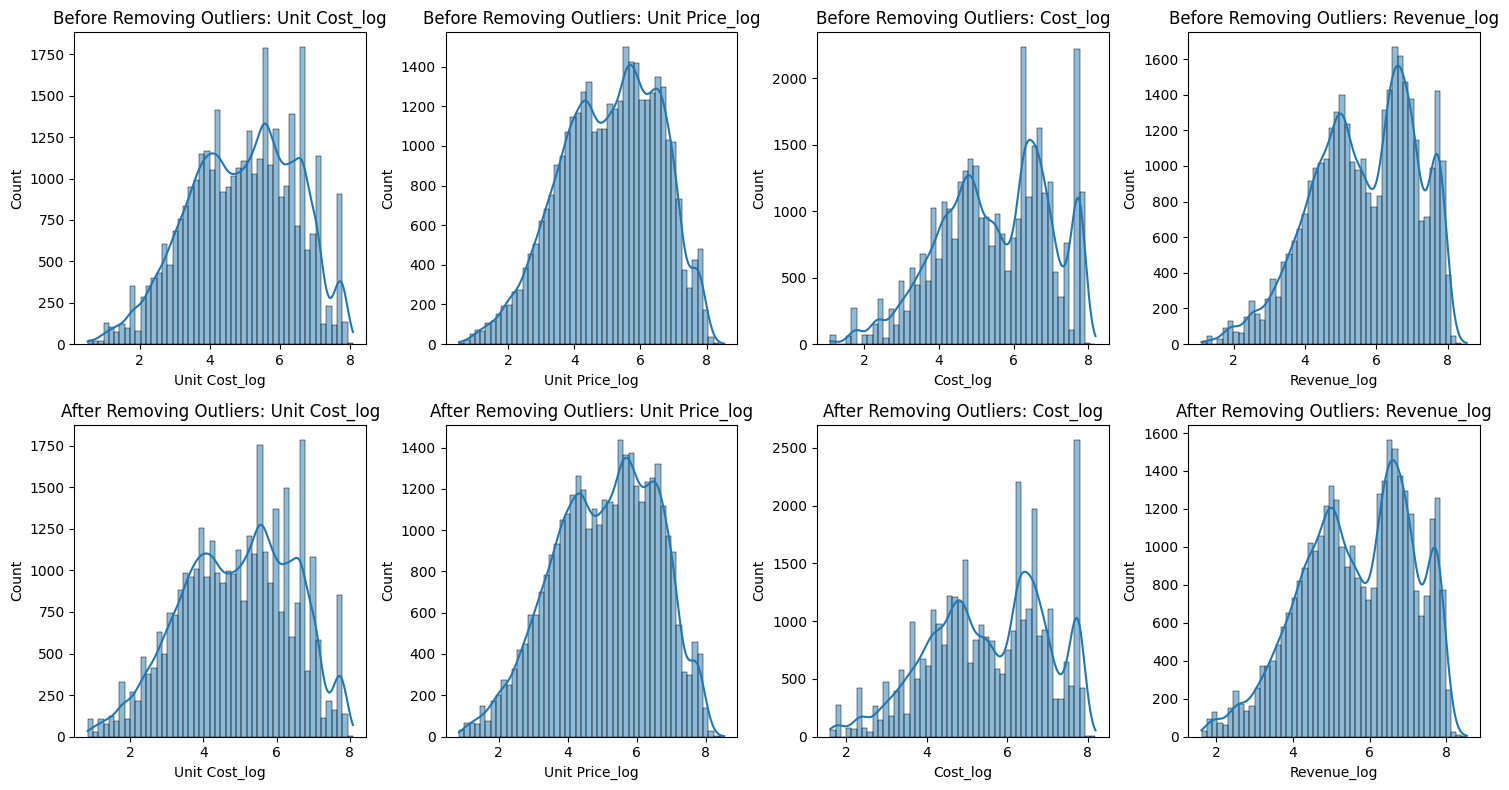
****

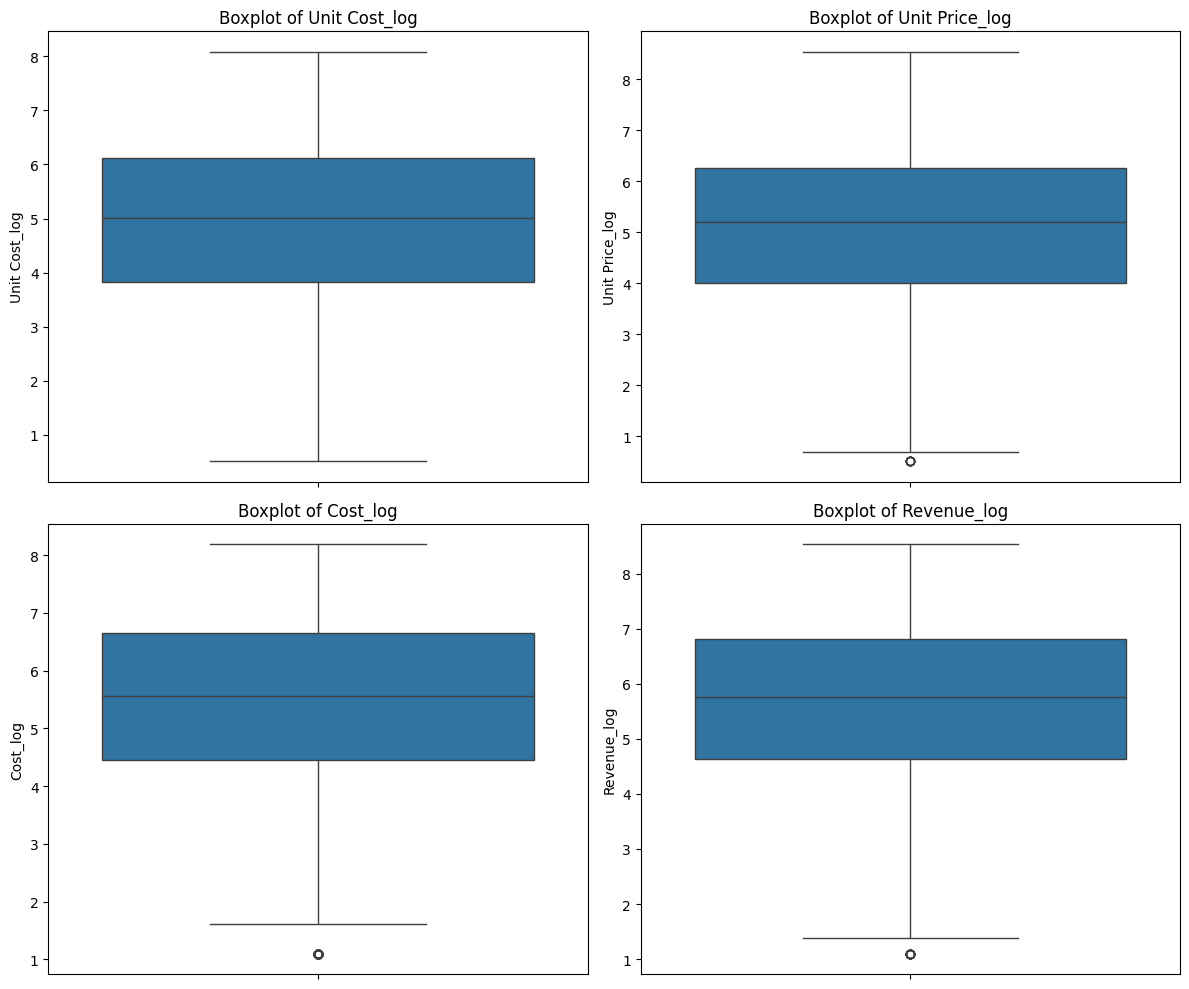
****

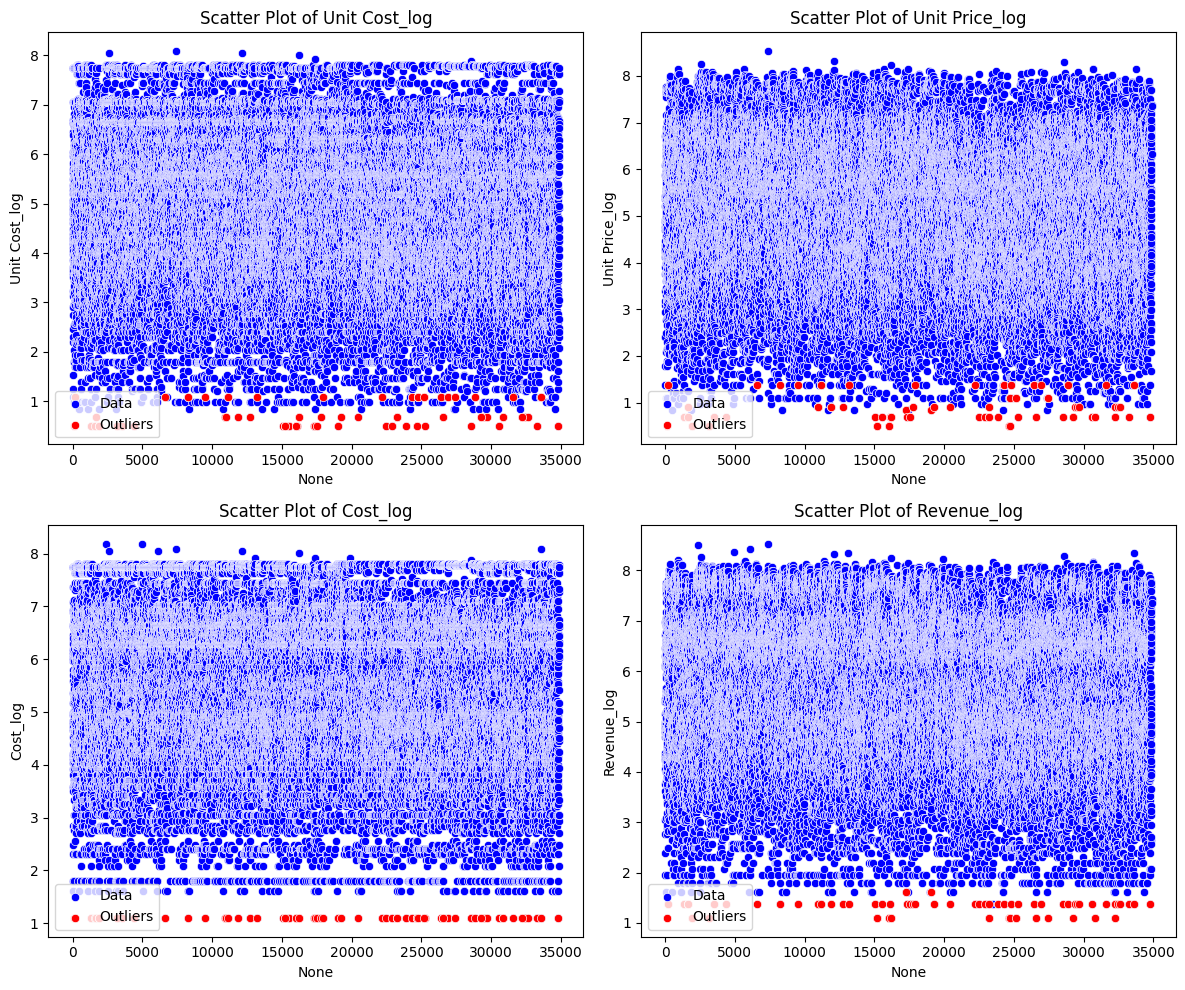
****

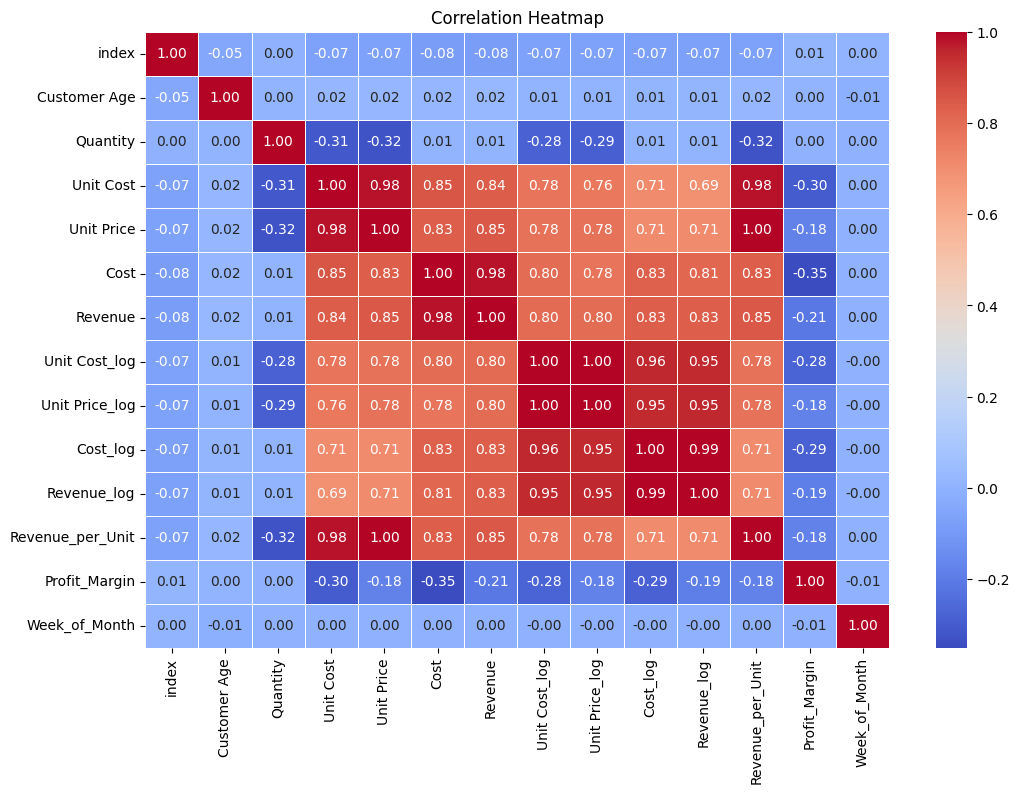
****

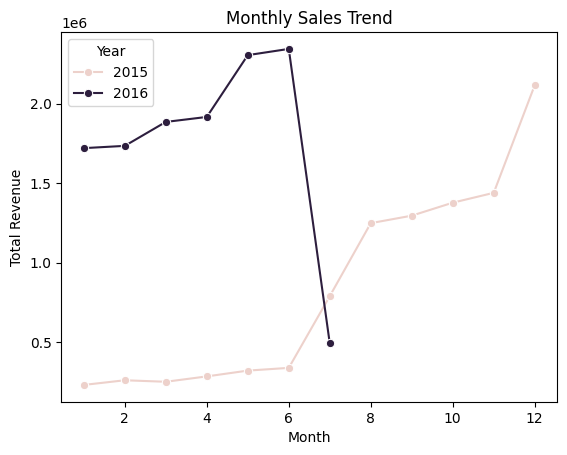
****

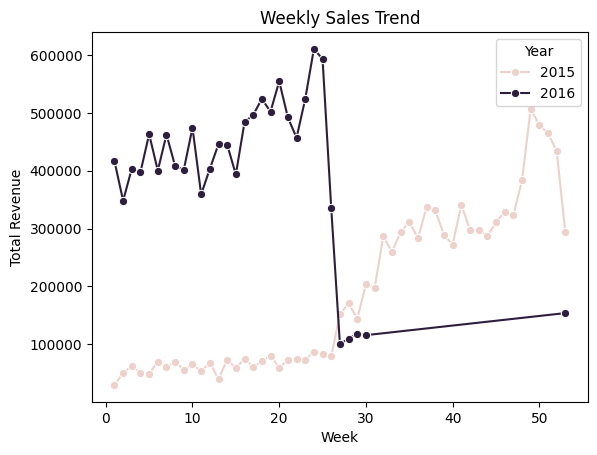
****

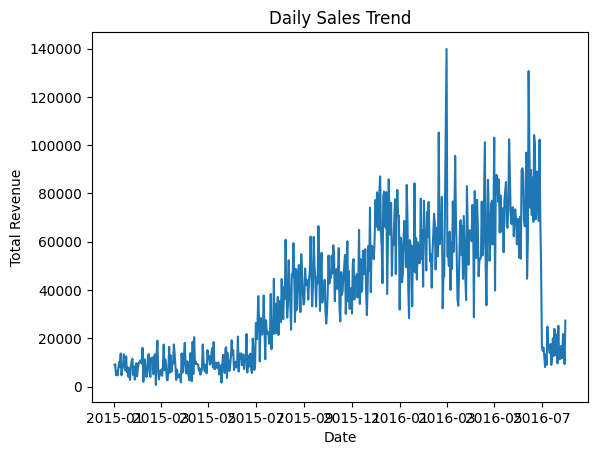
****

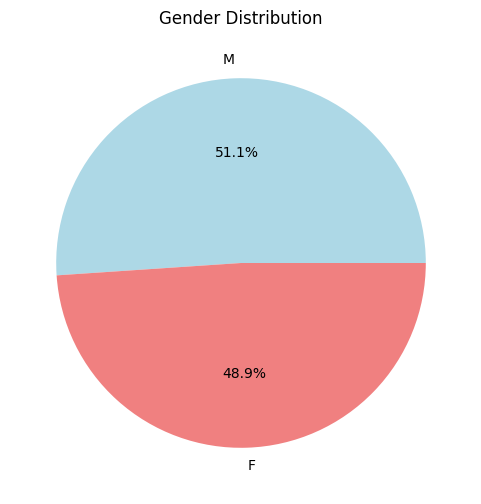
****

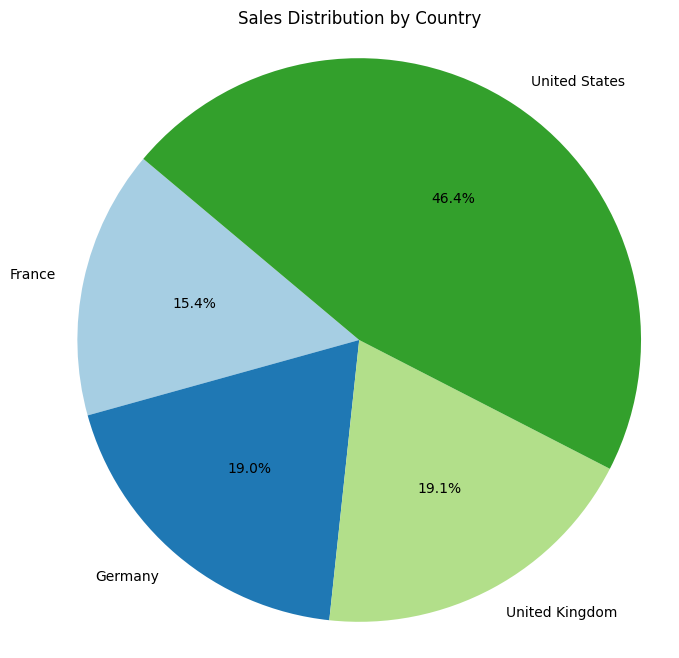
****

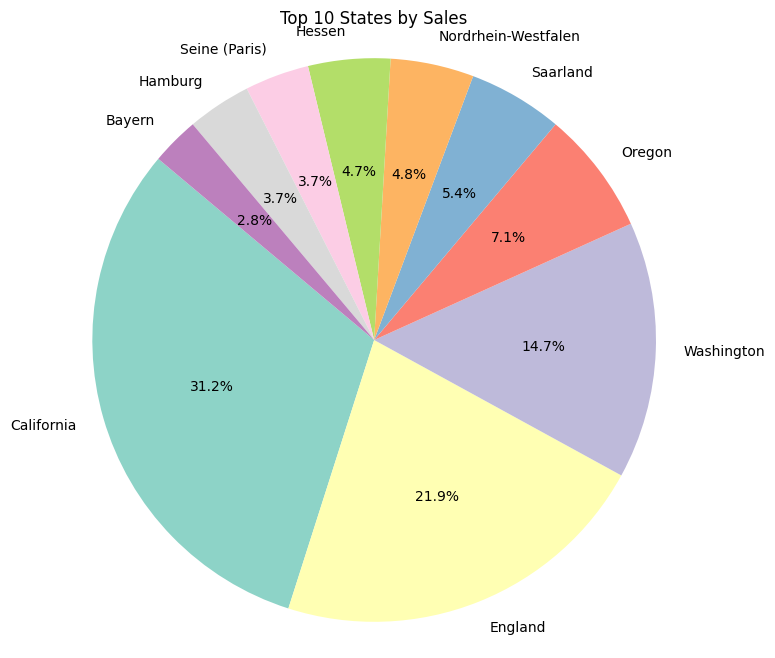
****

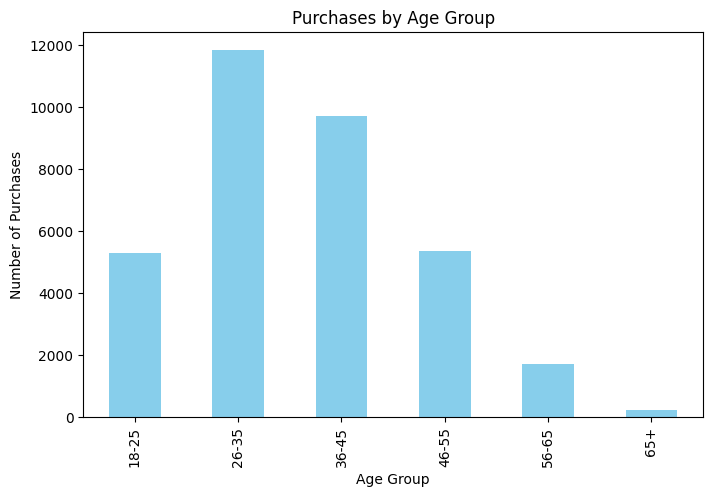
****

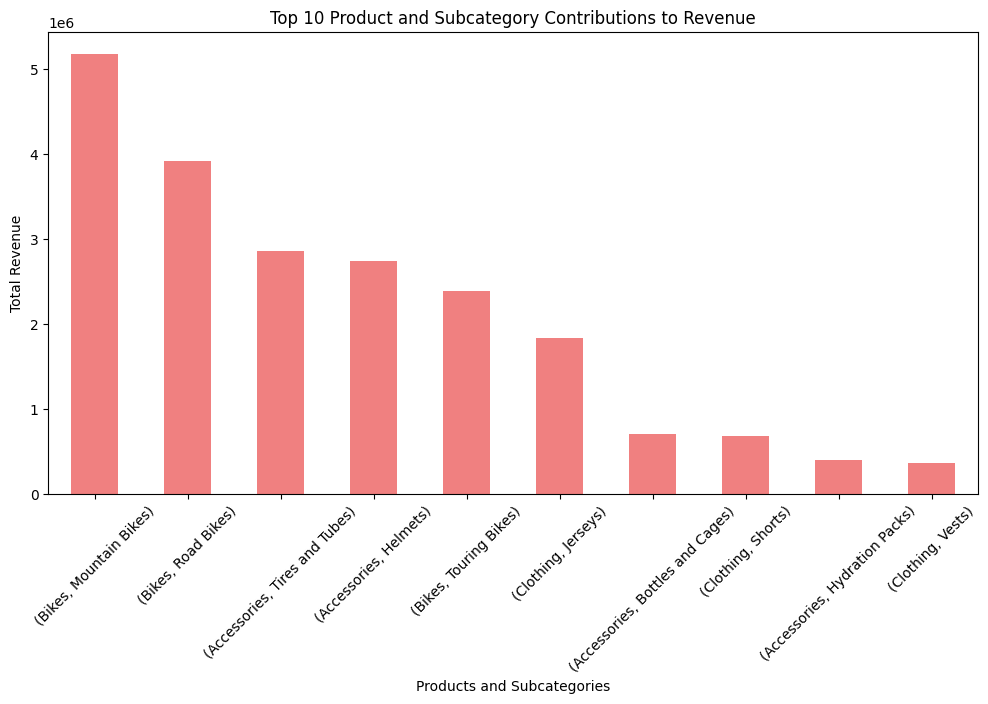
****

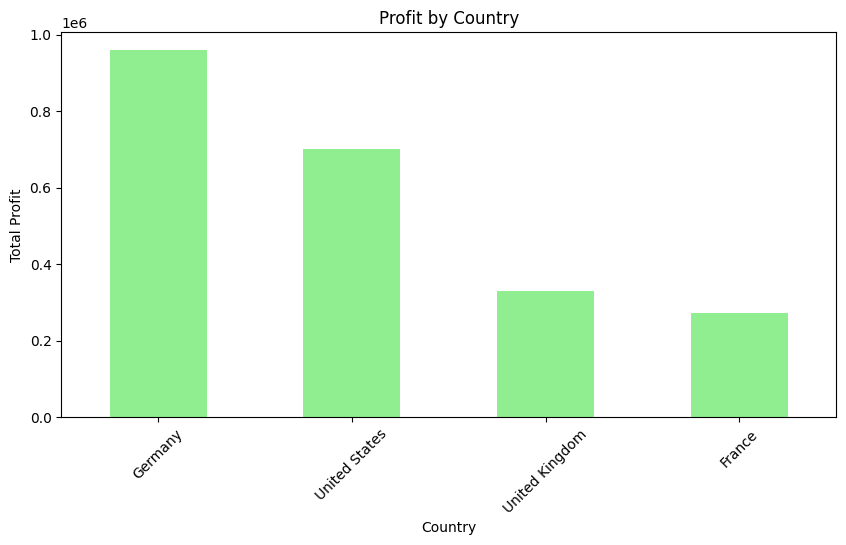
****

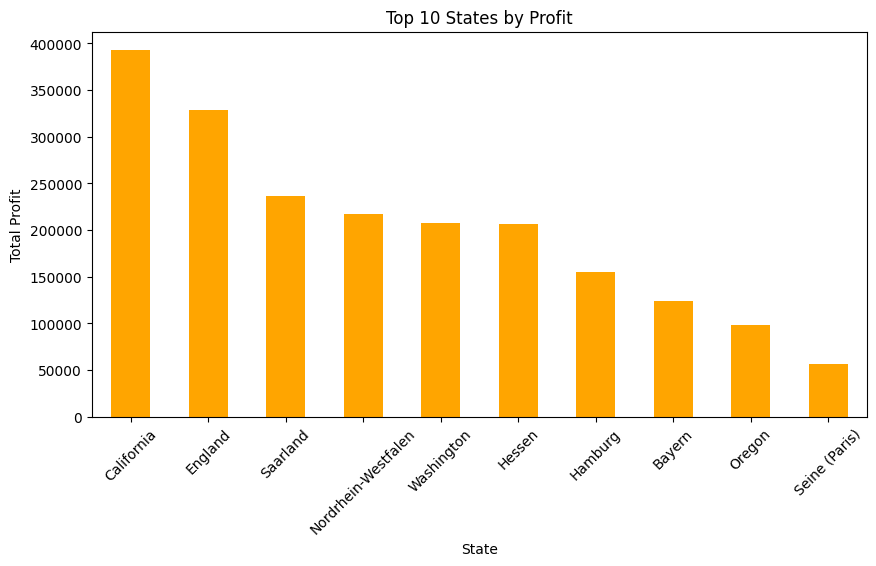
****

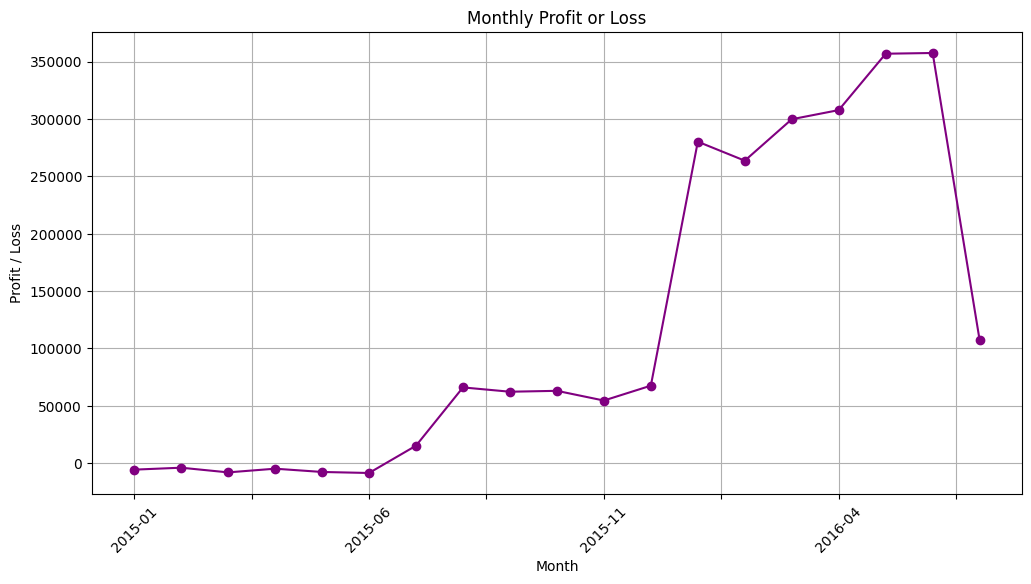
****

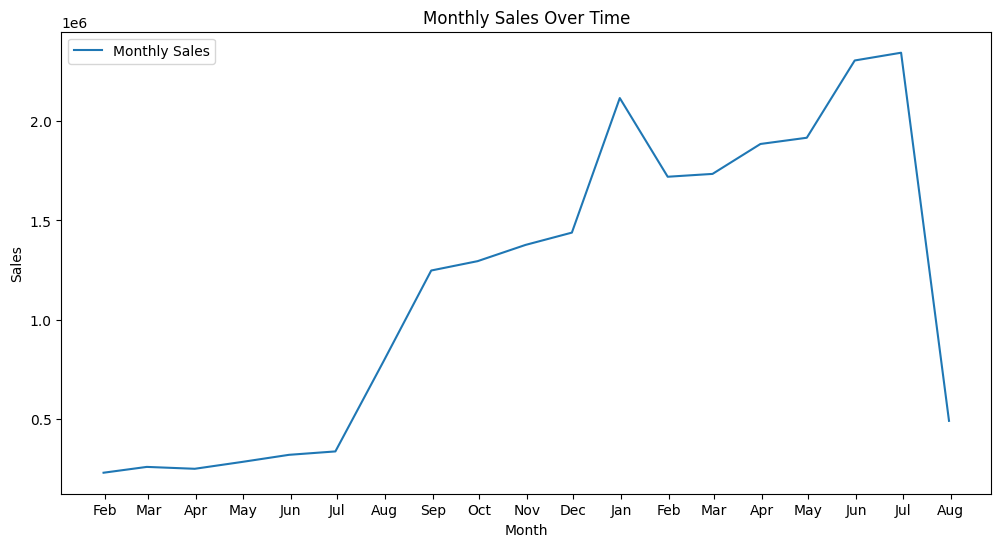
****

****

****

****

****

****