Name: Kshitij Hundre
Div: D15C
Roll No:18

EXP 2

Aim:

Data Visualization/ Exploratory data Analysis using Matplotlib and Seaborn.

1. Create bar graph, contingency table using any 2 features.
2. Plot Scatter plot, box plot, Heatmap using seaborn.
3. Create histogram and normalized Histogram.
4. Describe what this graph and table indicates.
5. Handle outlier using box plot and Inter quartile range.

Introduction:

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is the first step in your data analysis process developed by “John Tukey” in the 1970s. In statistics, exploratory data analysis is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. By the name itself, we can get to know that it is a step in which we need to explore the data set.

When you are trying to build a machine learning model you need to be pretty sure whether your data is making sense or not. The main aim of exploratory data analysis is to obtain confidence in your data to an extent where you’re ready to engage a machine learning algorithm.

Why do we do EDA?

Exploratory Data Analysis is a crucial step before you jump to machine learning or modeling your data. By doing this you can get to know whether the selected features are good enough to model, are all the features required, are there any correlations based on which we can either go back to the Data Preprocessing step or move on to modeling.

Once EDA is complete and insights are drawn, its feature can be used for supervised and unsupervised machine learning modeling.

In every machine learning workflow, the last step is Reporting or Providing the insights to the Stakeholders and as a Data Scientist you can explain every bit of code but you need to keep in mind the audience. By completing the EDA you will have many plots,heat-maps, frequency distribution, graphs, correlation matrix along with the hypothesis by which any individual can understand what your data is all about and what insights you got from exploring your data set.

Data visualization is very critical to market research where both numerical and categorical data can be visualized, which helps in an increase in the impact of insights and also helps in reducing the risk of analysis paralysis

Advantages of Data visualization:

1. Better Agreement:

In business, for numerous periods, it happens that we need to look at the exhibitions of two components or two situations. A conventional methodology is to experience the massive information of both the circumstances and afterward examine it. This will clearly take a great deal of time.

2. A Superior Method:

It can tackle the difficulty of placing the information of both perspectives into the pictorial structure. This will unquestionably give a superior comprehension of the circumstances. For instance, Google patterns assist us with understanding information identified with top ventures or inquiries in pictorial or graphical structures.

3. Simple Sharing of Data:

With the representation of the information, organizations present another arrangement of correspondence. Rather than sharing the cumbersome information, sharing the visual data will draw in and pass on across the data which is more absorbable.

4. Deals Investigation:

With the assistance of information representation, a salesman can, without much of a stretch, comprehend the business chart of items. With information perception instruments like warmth maps, he will have the option to comprehend the causes that are pushing the business numbers up just as the reasons that are debasing the business numbers. Information representation helps in understanding the patterns and furthermore, different variables like sorts of clients keen on purchasing, rehashing clients, the impact of topography, and so forth.

5. Discovering Relations Between Occasions:

A business is influenced by a lot of elements. Finding a relationship between these elements or occasions encourages chiefs to comprehend the issues identified with their business. For instance, the online business market is anything but another thing today. Each time during certain happy seasons, like Christmas or Thanksgiving, the diagrams of online organizations go up. Along these lines, state if an online organization is doing a normal $1 million business in a specific quarter and the business ascends straightaway, at that point they can rapidly discover the occasions compared to it.

6. Investigating Openings and Patterns:

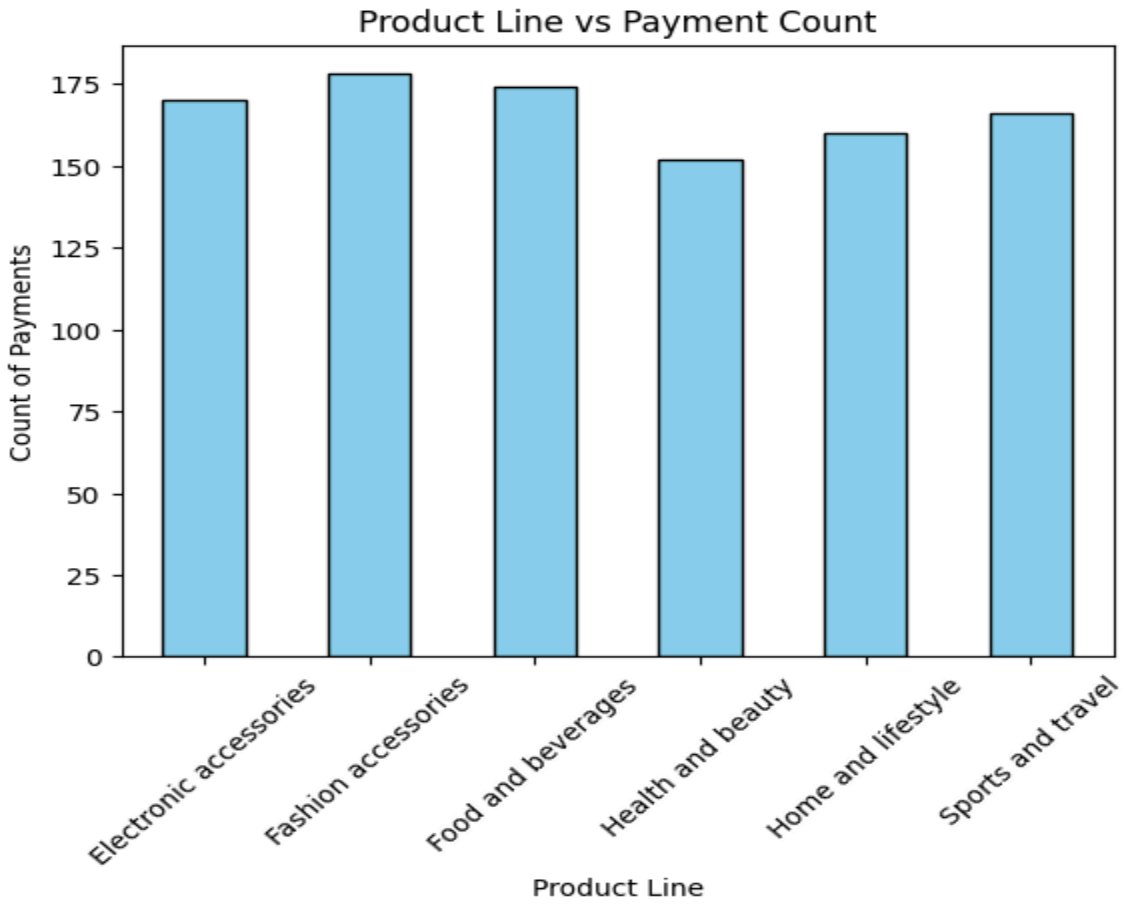
With the huge loads of information present, business chiefs can discover the profundity of information in regard to the patterns and openings around them. Utilizing information representation, the specialists can discover examples of the conduct of their clients, subsequently preparing for them to investigate patterns and open doors for business.

1) Bar Graph (Product Line vs Gender_male)

Inference:

- This bar graph shows the distribution of sales for different product lines based on payment count..
- If the bar height for one product line is significantly different between different payment count, it suggests a payment-based sales pattern.
- For instance, if *Health and Beauty* has a higher payment bar, this may indicate a preference trend.

```
[ ] import matplotlib.pyplot as plt


    # Bar plot for product line and payment method
    df.groupby('Product line')['Payment'].count().plot(kind='bar', color='skyblue', edgecolor='black')
    plt.title('Product Line vs Payment Count')
    plt.xlabel('Product Line')
    plt.ylabel('Count of Payments')
    plt.xticks(rotation=45)
    plt.show()
```



2) Contingency Table (Product Line and Payment Method)

What: A table that shows the frequency distribution of variables.
Why: Helps analyze relationships between categorical variables.

Inference:

- The table provides a frequency distribution of product line purchases across different payment methods.
- For example, if *Cash* is more frequently used for *Food and Beverages*, it might suggest customer spending habits at that department.

```
contingency_table = pd.crosstab(df['Product line'], df['Payment'])
print(contingency_table)
```

```
Payment                 Cash  Credit card  Ewallet
Product line
Electronic accessories    71           46       53
Fashion accessories       57           56       65
Food and beverages        57           61       56
Health and beauty         49           50       53
Home and lifestyle        51           45       64
Sports and travel         59           53       54
```

3) Scatter Plot

- **What:** A graph of points that shows the relationship between two variables.
- **Why:** Useful to identify patterns, correlations, or clusters in data.

  The scatter plot reveals the relationship between the unit price and gender which paid that price.

  A positive trend might indicate that higher unit prices lead to larger total sales, possibly due to bulk purchases.

  `hue='Gender_male'`: Points are colored based on the boolean male indicator.

  Displays the count of transactions for each product line categorized by gender (where `1` indicates Male and `0` indicates Female).

Clusters indicate common pricing patterns, while isolated points suggest anomalies or rare pricing behavior.

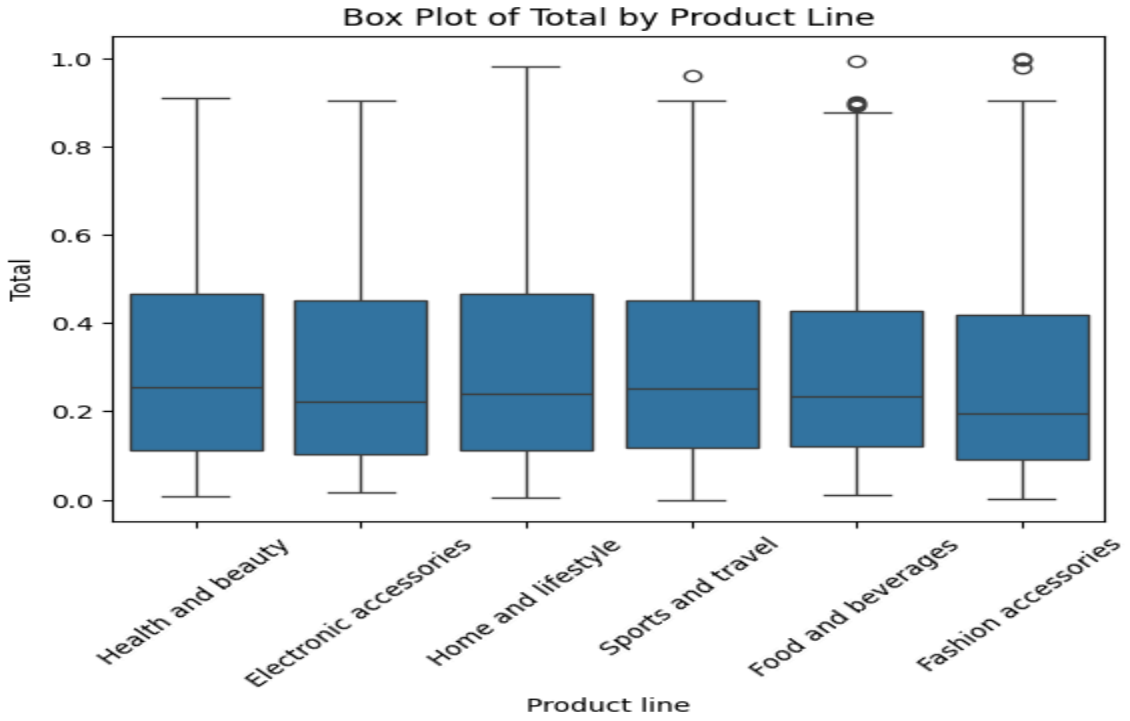
```
import seaborn as sns

sns.scatterplot(data=df, x='Unit price', y='Total', hue='Gender_Male')
plt.title('Unit Price vs Total with Gender (Male=1)')
plt.show()
```

4) Inference: Box Plot of Total by Product Line

- **Spread of Total Sales:**
  The box plot shows the distribution of total sales for each product line. The height of the boxes represents the range of typical sales amounts.
- **Median Sales:**
  The central line inside each box represents the **median sales value** for that product line. Comparing these medians indicates which product line generally brings in higher sales.

- **Interquartile Range (IQR):**
  The box's length (from Q1 to Q3) shows the middle 50% of the sales values. Wider boxes indicate more variability in sales for that product line.
- **Outliers:**
  Data points outside the whiskers are outliers, suggesting unusually high or low sales values. These might be special sales events or data inconsistencies.
- **Example Observation:**
  - If the *Electronic Accessories* product line shows several outliers on the higher side, it may indicate a few exceptionally large transactions.
  - If *Fashion Accessories* has a narrow IQR, it suggests consistent sales amounts without significant variability.

```
#box plot
sns.boxplot(data=df, x='Product line', y='Total')
plt.title('Box Plot of Total by Product Line')
plt.xticks(rotation=45)
plt.show()
```



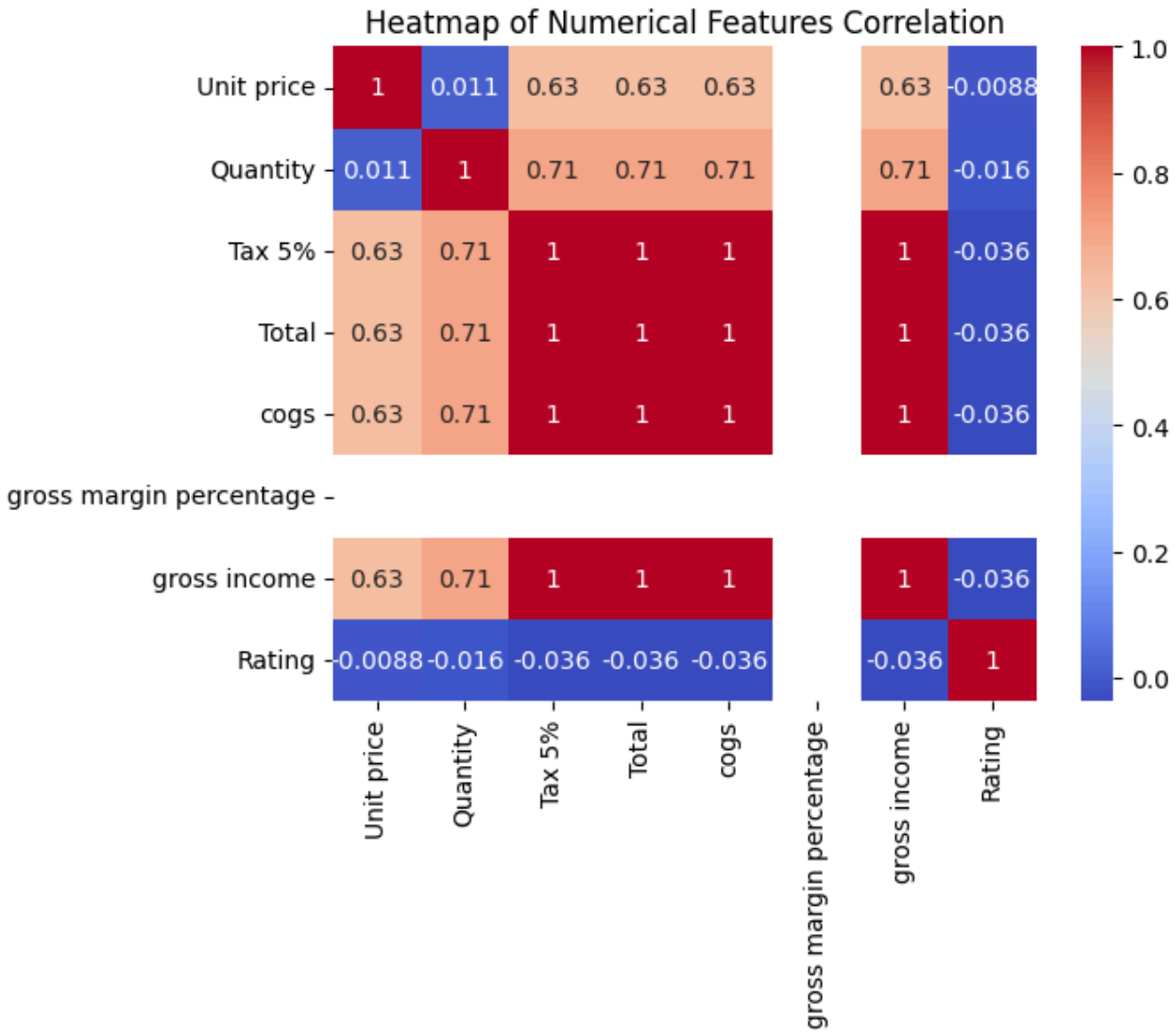
5) Heatmap of Numerical Features Correlation

- **Purpose:**
  This heatmap visually represents the correlation between the numerical features of the dataset. The values range between **-1 to 1**, where:
  - **1:** Strong positive correlation (as one feature increases, the other increases)
  - **0:** No correlation
  - **-1:** Strong negative correlation (as one feature increases, the other decreases)

Key Observations:

- **Total vs Quantity (High Positive Correlation)**
  - A high positive correlation (close to 1) suggests that the total sales amount increases as the number of purchased items (Quantity) increases. This is expected in sales data.
- **Gross Income vs Total (Strong Positive Correlation)**
  - This indicates that a higher total amount is strongly associated with higher gross income. This is intuitive as gross income is often derived from total sales.
- **Weak Correlations:**
  - Some features, like *Unit Price* and *Quantity*, may show weak or no correlation, suggesting that the number of items purchased doesn't necessarily depend on unit prices.
- **No Negative Correlations:**
  - Since this is a sales dataset, most numerical features are likely positively related.

```
#heatmap
numeric_df = df.select_dtypes(include=['float64', 'int64'])

# Generate the heatmap for numerical features
import seaborn as sns
import matplotlib.pyplot as plt

sns.heatmap(numeric_df.corr(), annot=True, cmap='coolwarm')
plt.title('Heatmap of Numerical Features Correlation')
plt.show()
```



6) Histogram

Inference: Customer Rating Distribution Histogram

1. **Rating Spread:**
   The histogram shows how customer ratings are distributed across different ranges, with the bins dividing ratings from low to high.
2. **Most Common Ratings:**
   If there's a peak near higher ratings (like 8-10), it indicates customer satisfaction, whereas peaks at lower ratings suggest dissatisfaction trends.

3. **Skewness of Ratings:**If the distribution leans towards higher ratings, it suggests overall positive feedback from customers; if it's more balanced, opinions are mixed

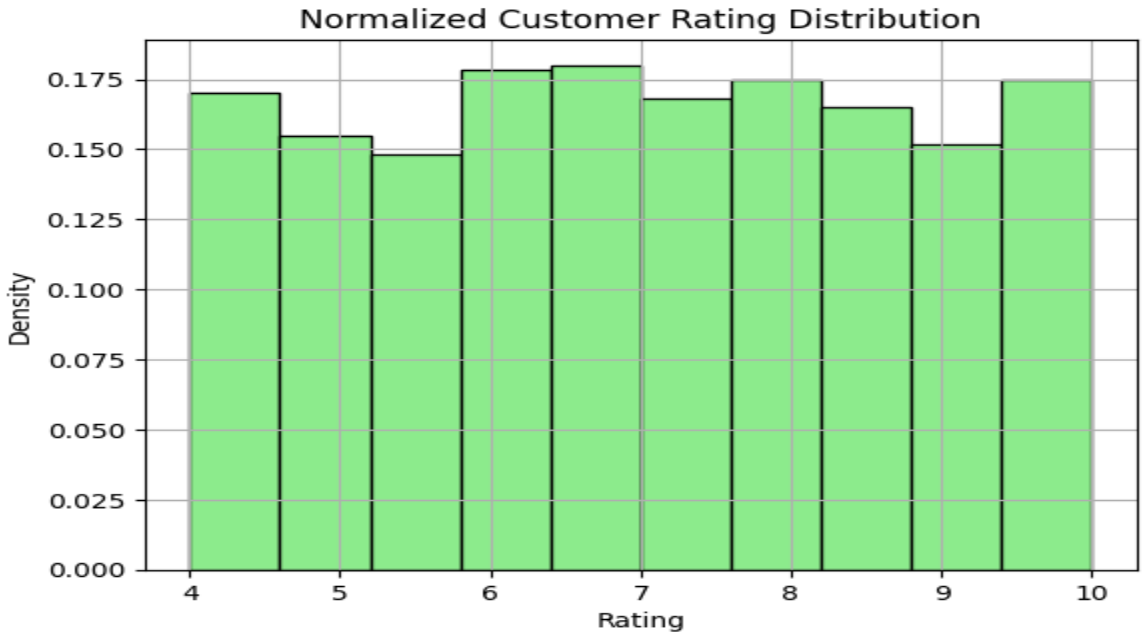
```
#histogram
df['Rating'].hist(bins=10, color='lightblue', edgecolor='black')
plt.title('Customer Rating Distribution')
plt.xlabel('Rating')
plt.ylabel('Frequency')
plt.show()
```

7) Normalized Histogram

Inference: Normalized Customer Rating Distribution Histogram

1. Rating Spread:
   Similar to the regular histogram, the normalized histogram shows the spread of customer ratings across different ranges, with the bins dividing ratings from low to high.
2. Most Common Ratings:
   Peaks in the density indicate the most frequent customer rating ranges. Higher density near higher ratings suggests frequent positive feedback.
3. Probability Distribution:
   Since the histogram is normalized, the y-axis represents probability density rather than frequency, helping visualize the likelihood of different rating ranges.

```
#normalized histogram      both of our histograms will be the same as we have performed normalization when we clean the data
df['Rating'].hist(bins=10, density=True, color='lightgreen', edgecolor='black')
plt.title('Normalized Customer Rating Distribution')
plt.xlabel('Rating')
plt.ylabel('Density')
plt.show()
```



8) Handle outlier using box plot

Inference: Box Plot for Total Sales

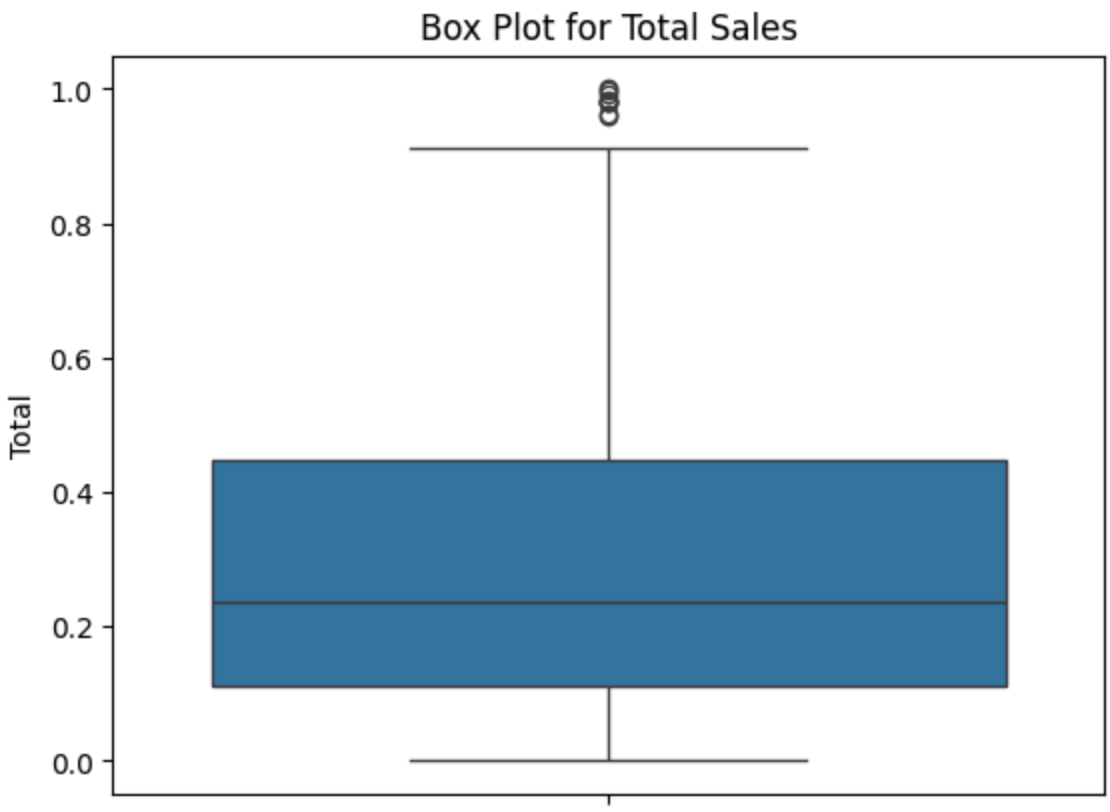
1. Identifying Outliers:
    - Any data points outside the whiskers of the box plot are considered outliers. These points represent unusually high total sales amounts.
2. Sales Variability:
    - The spread of the box shows the range of typical sales values, while the whiskers indicate the overall variability.
3. Business Insight:
    - Outliers may indicate rare high-value transactions or potential data entry errors that require investigation.
    - Understanding these outliers can help identify key trends, such as promotional events leading to significant sales.

Outliers detected in *Total* or *Gross Income* columns suggest extreme sales figures, possibly due to special promotions or data entry errors.
Handling these outliers ensures more accurate statistical analysis.

```
# Handle Outliers Using Box Plot and IQR

#Box Plot to Visualize Outliers
sns.boxplot(data=df, y='Total')
plt.title('Box Plot for Total Sales')
plt.show()
```

```
#Handle Outliers with IQR
Q1 = df['Total'].quantile(0.25)
Q3 = df['Total'].quantile(0.75)
IQR = Q3 - Q1

# Define bounds
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

# Filter out outliers
cleaned_df = df[(df['Total'] >= lower_bound) & (df['Total'] <= upper_bound)]
print(f"Rows before outlier removal: {len(df)}")
print(f"Rows after outlier removal: {len(cleaned_df)}")
```

```
Rows before outlier removal: 1000
Rows after outlier removal: 991
```



Conclusion:

Hence we learned about exploratory data analysis and various types of statistical measures of data along with correlation. We also learnt about visualization and applied these concepts with hands-on experience on our chosen dataset.