Lab 4: Scatter Plots – Bivariate Analysis

Prelab Questions

1.Define bivariate analysis and describe its importance.

It examines the relationship between two variables to understand patterns, trends, or associations. It is important because it helps identify correlations, dependencies, and potential causal relationships.

2.What does the slope of a regression line in a scatter plot indicate?

The slope indicates the rate of change between the dependent and independent variables. A positive slope shows a direct relationship, while a negative slope shows an inverse relationship.

3.How do you identify a strong positive or negative correlation from a scatter plot?

A strong positive correlation appears as points closely aligned along an upward sloping line, whereas a strong negative correlation aligns points along a downward sloping line.

4.What is the purpose of adding a trend line to a scatter plot?

A trend line summarizes the relationship between variables, making patterns clearer and aiding in prediction or analysis.

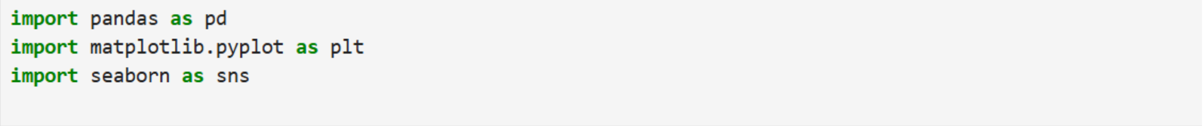
5.Why is it important to consider outliers in scatter plot visualizations?

Outliers can skew the results, obscure true patterns, and affect the accuracy of the analysis. Identifying them ensures better data interpretation.

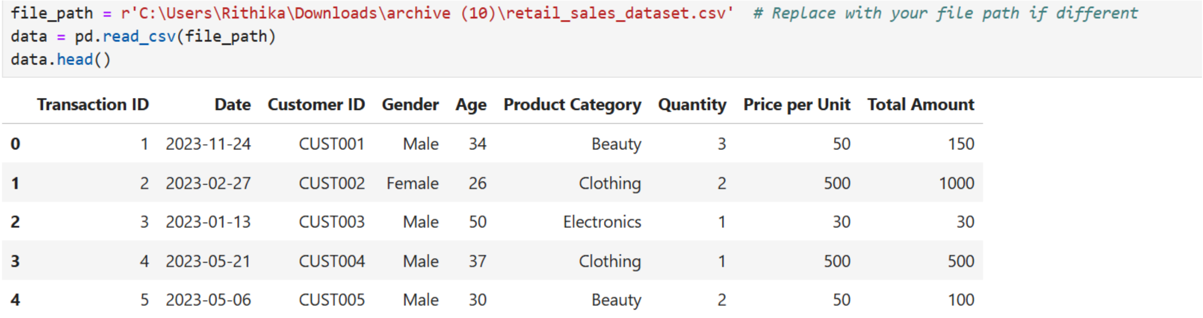
In-Lab Details

Objective:
Analyze relationships between two variables using scatter plots and trend lines.

PYTHON SCRIPT

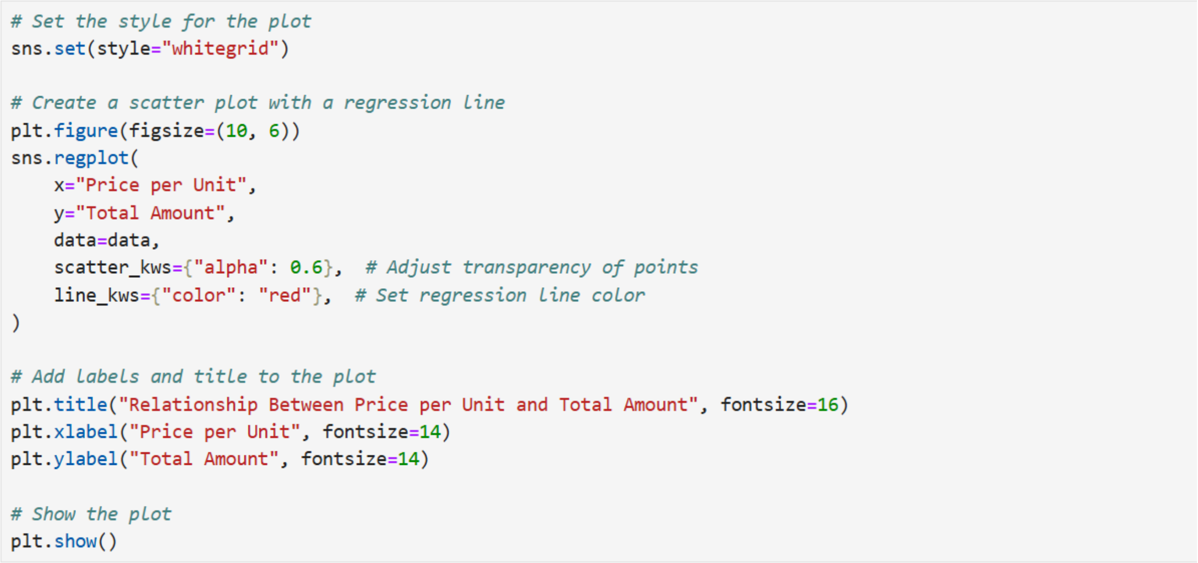


```python
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```



```python
file_path = r'C:\Users\Rithika\Downloads\archive (10)\retail_sales_dataset.csv'  # Replace with your file path if different
data = pd.read_csv(file_path)
data.head()
```

	Transaction ID	Date	Customer ID	Gender	Age	Product Category	Quantity	Price per Unit	Total Amount
0	1	2023-11-24	CUST001	Male	34	Beauty	3	50	150
1	2	2023-02-27	CUST002	Female	26	Clothing	2	500	1000
2	3	2023-01-13	CUST003	Male	50	Electronics	1	30	30
3	4	2023-05-21	CUST004	Male	37	Clothing	1	500	500
4	5	2023-05-06	CUST005	Male	30	Beauty	2	50	100



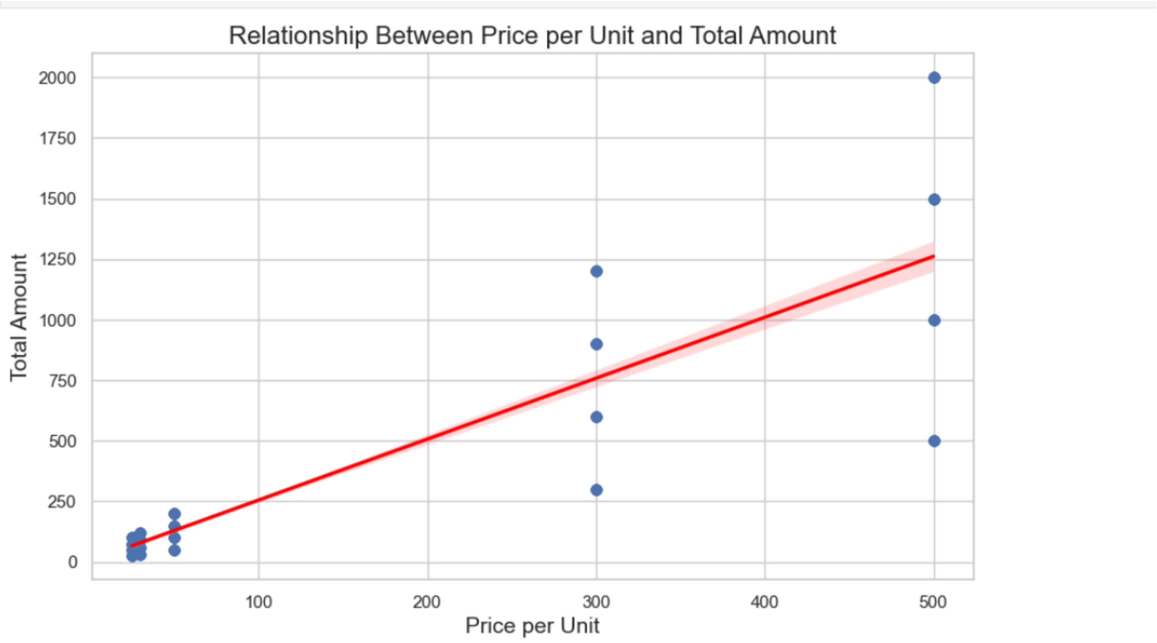
```
# Set the style for the plot
sns.set(style="whitegrid")

# Create a scatter plot with a regression line
plt.figure(figsize=(10, 6))
sns.regplot(
    x="Price per Unit",
    y="Total Amount",
    data=data,
    scatter_kws={"alpha": 0.6},  # Adjust transparency of points
    line_kws={"color": "red"},  # Set regression line color
)

# Add labels and title to the plot
plt.title("Relationship Between Price per Unit and Total Amount", fontsize=16)
plt.xlabel("Price per Unit", fontsize=14)
plt.ylabel("Total Amount", fontsize=14)

# Show the plot
plt.show()
```

Resources:

- Python (Jupyter Notebook).
- Libraries: Matplotlib, Seaborn.
- Dataset: retail_data.csv with columns for marketing expenses and sales.

Expected Output:

- A scatter plot showing the relationship between Price per Unit and Total Amount.
- A regression line indicating the trend of the relationship.



Postlab Questions

1. How does adding a regression line enhance scatter plot interpretation?

    - **Trend Identification**: A regression line makes it easier to see the overall trend between variables.
    - **Relationship Strength**: The slope indicates the strength and direction (positive or negative) of the relationship.
    - **Outlier Detection**: Points far from the line highlight deviations.
    - **Prediction**: The line enables estimating values for dependent variables.

2. What are the limitations of scatter plots in identifying non-linear relationships?

    - **Hidden Patterns**: Non-linear relationships may not be easily discernible, especially in noisy data.
    - **Dense Clustering**: Overlapping points in clusters can obscure trends.
    - **Ambiguity**: Scatter plots cannot directly quantify non-linear relationships without additional modeling.
    - **Scale Dependency**: Scaling issues may hide or distort patterns.

3. Discuss the significance of correlation coefficients in scatter plots.

    - **Quantification**: Provides a numerical measure of the relationship between variables, ranging from -1 to +1.
    - **Direction**: Indicates whether the relationship is positive (both variables increase) or negative (one decreases as the other increases).
    - **Validation**: Confirms or contrasts visual patterns observed in the plot.
    - **Comparison**: Useful for comparing relationships in different datasets.

4. Suggest improvements to scatter plots when dealing with large datasets.

    - **Transparency**: Reduce marker opacity to reveal overlapping points.
    - **Aggregation**: Use heatmaps or bubble plots to show data density.
    - **Interactive Plots**: Allow zooming, filtering, and exploration with tools like Plotly.
    - **Faceting**: Split data into smaller, category-specific plots.
    - **Color and Size**: Add dimensions using color intensity or marker size.

5. How can categorical variables be incorporated into scatter plots?

    - **Color Coding**: Assign different colors to categories for easy differentiation.
    - **Marker Shapes**: Use varying marker shapes for distinct categories.
    - **Faceted Plots**: Create separate scatter plots for each category.
    - **Convex Hulls**: Highlight groups with category-specific boundaries.
    - **Interactive Legends**: Enable toggling of categories for better focus and clarity.

DONE BY
Durga S – 22BCS026
Nandhini Priya K K -22BCS073
Nithiyaa T – 22BCS082
Priya Dharshini P – 22BCS094
Rithika S – 22BCS103
Sandhiya G – 22BCS109
Santhiya P – 22BCS111