K.R. MANGALAM UNIVERSITY

SCHOOL OF ENGINEERING & TECHNOLOGY

ASSIGNMENT

PROBABILISTIC MODELING AND REASONING

Submitted by:

Name: Akash Sharma

Roll No: 2401201108

Course: BCA (AI & DS) Sec: B

Submitted to:

Dr. Kunal Rai

ASSIGNMENT 1

Q1. Bike prices:

(a) **Mean price**

Add all prices: 12000 + 15000 + 18000 + 20000 + 50000 = 119000

Number of bikes = 5

Mean = Total sum / Number of values = 119000 / 5 = ₹23,800

(b) **Median price**

Arrange in ascending order: 12000, 15000, 18000, 20000, 50000

Odd number of values, median is middle one = 3rd value = ₹18,000

(c) **Which average to advertise?**

Mean is heavily affected by the outlier ₹50,000.

Median is not affected much and better represents central tendency.

So, advertising median price ₹18,000 will make bikes seem more affordable.

Q2. Describe the relationship between a histogram's shape and the values of its mean, median, and mode in a left-skewed distribution.

Answer:

In a **left-skewed** distribution, the tail is longer to the left.

Mode > Median > Mean

Mean is pulled toward the tail (left), so less than median.

Q3. Exam scores:

Sol:

(a) **Mean:**

- Sum all scores: 67+72+78+85+90+91+95 = 578
- Number of scores = 7
- Mean = 578 / 7 ≈ 82.57

(b) **Median:**

- Number of values is odd (7), median is middle value = 4th score in sorted list = 85

(c) **Mode:**

- All appear once, so no mode (or no repeating value).

(d) **Distribution shape:**

- Scores roughly symmetrical as mean and median close.

(e) **Five-number summary:**

- Min = 67
- Q1 = Median of lower half = median of {67,72,78} = 72
- Median = 85
- Q3 = Median of upper half = median of {90,91,95} = 91
- Max = 95

Q4. Differentiate between a population parameter and a sample statistic with an example of each.

Answer:

- Population parameter: Measure describing whole population (e.g., mean income of all citizens).
- Sample statistic: Measure computed from a sample subset (e.g., mean income of sampled group).

Q5. Employees:

Sol:

(a) Mean

45000+50000+55000+60000+2500005=4600005=92,000\frac{45000+50000+55000+60000+250000}{5} = \frac{460000}{5} = 92,000545000+50000+55000+60000+250000=5460000=92,000

Median = 55,000 (middle value)

(b) Adver sing

- Company will likely adver se mean = $92,000 (looks higher).
- Employees prefer median = $55,000 (represents typical salary).
  Difference occurs because the mean is inflated by one extreme outlier ($250,000).

Q6. Students data:

Sol:

The solution for this is same as question 3

(a) **Mean:**

- Sum all scores: 67+72+78+85+90+91+95 = 578
- Number of scores = 7
- Mean = 578 / 7 ≈ 82.57

(b) **Median:**

- Number of values is odd (7), median is middle value = 4th score in sorted list = 85

(c) **Mode:**

- All appear once, so no mode (or no repeating value).

(d) **Distribution shape:**

- Scores roughly symmetrical as mean and median close.

(e) **Five-number summary:**

- Min = 67
- Q1 = Median of lower half = median of {67,72,78} = 72
- Median = 85

- Q3 = Median of upper half = median of {90,91,95} = 91
- Max = 95

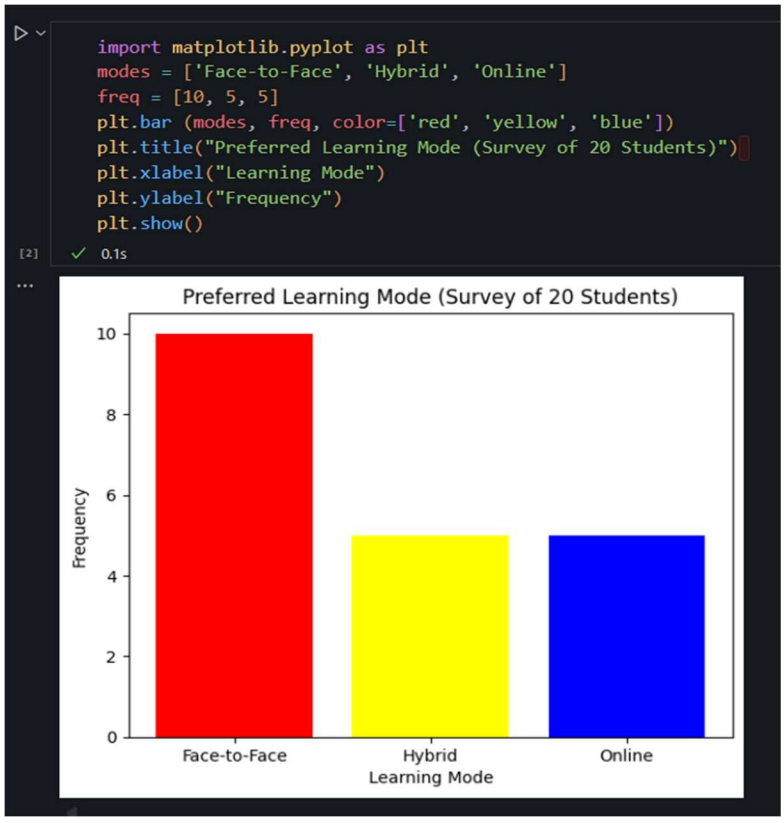
Q7. Modes of transportation:

Sol:

(a) Frequency table:

Mode	**Frequency**
F (In-person)	11
H (Hybrid)	5
O (Online)	4

(b) Bar chart would plot mode categories vs frequencies.



(c) Most popular mode: F (In-Person), 55% (11/20 students)

(d) Histogram not suitable for categorical data; bar chart is appropriate.

ASSIGNMENT 2

Q.1 A jar contains 5 red and 7 blue marbles. Two marbles are drawn randomly one after the other without replacement. What is the probability that the second marble is red, given that the first one was blue?

Sol:

- Total marbles = 12 red + 6 blue = 18
- First marble blue drawn: Remaining marbles = 17
- Remaining red marbles = 12
- Probability = 12/17

Q2. Suppose a laboratory test on a blood sample yields two possible results, positive or negative. According to the industry reports, the blood sample 95% of the people with a particular disease yield positive results. But 2% of the people without the disease also give positive results (false positive). 1% of the total population is infected by the disease. Determine the probability that a person chosen randomly from the population will have disease given that the blood sample of the person tests positive?

Sol:

Given:

- P(Disease) = 0.01 (1%)
- P(No disease) = 0.99
- P(Positive | Disease) = 0.9
- P(Positive | No disease) = 0.05

Calculate P(Disease | Positive):
Using Bayes theorem,

to calculate numerator and denominator:

- Numerator: $0.9 \times 0.01 = 0.009$
- Denominator: $(0.9 \times 0.01) + (0.05 \times 0.99) = 0.009 + 0.0495 = 0.0585$
- P(Disease | Positive) = $0.009 / 0.0585 \approx 0.1538$ or 15.38%

Q3. A patient from the population given above walks into the clinic of a doctor with some symptoms of the disease. After examining the patient and without checking the blood test report, the doctor gives his opinion that there are 30% chances that the patient is suffering from the disease. How should revise his opinion after checking the blood report.

Sol:

Use Bayes:

- Numerator = P(+|D)P(D)=0.95×0.30=0.285P(+\mid D)P(D) = 0.95 \ mes 0.30 = 0.285P(+|D)P(D)=0.95×0.30=0.285.
- Denominator = 0.285+P(+|Dc)P(Dc)=0.285+0.02×0.70=0.285+0.014=0.2990.285 + P(+\mid D^c)P(D^c) = 0.285 + 0.02 \ mes 0.70 = 0.285 + 0.014 = 0.2990.285+P(+|Dc)P(Dc)=0.285+0.02×0.70=0.285+0.014=0.299.

Interpreta on: Because the doctor already assessed a high prior (30%), a posi ve test pushes the posterior probability very high — about 95.2%. That's why priors ma er: the same test result leads to very different posteriors depending on the prior belief.

Q4 and Q5:

Sol:

Given:

- Total students = 50
- Students studying Physics = 30
- Students studying Chemistry = 25
- Students studying both Physics and Chemistry = 15

Step 1: Understand the sets and overlaps.

- The "both" group (15 students) is the intersection of Physics and Chemistry sets.
- Students studying only Physics are those in Physics but not in Chemistry.

Step 2: Calculate number studying only Physics.

$$\text{Only Physics} = \text{Physics} - \text{Both} = 30 - 15 = 15$$

Step 3: Calculate the probability of only Physics.

$$P(\text{Only Physics}) = \frac{\text{Only Physics}}{\text{Total students}} = \frac{15}{50} = 0.3$$

Interpretation: There's a 30% chance that a randomly selected student studies Physics but not Chemistry.

ASSIGNMENT 3

1. Project Objective:
Perform a complete Exploratory Data Analysis (EDA) on a dataset of your choice. Your goal is to understand the underlying structure of the data, discover patterns and relationships, identify anomalies and outliers, and test your initial hypotheses.

2. Dataset Selection:
You must select a dataset from (link unavailable) Choose a dataset that is rich enough to allow for meaningful analysis. It should have:
- At least 5 variables (columns).
- A mix of numerical and categorical variables is highly recommended.
- A sufficient number of rows (e.g., >100) to make analysis interesting.
Popular beginner-friendly datasets on Kaggle include: Titanic, Iris, House Prices, Netflix Movies and TV Shows, Wine Reviews, or Pokemon Datasets. You are free to choose any that interests you.

3. Technical Requirements:
Your analysis must include the following, implemented in Python:
- Data Loading & Inspection:
  - Load the dataset using pandas.
  - Display the first and last few rows (.head(), .tail()).
  - Check the data types and summary info (.info(), .describe()).
- Data Cleaning (Mandatory):
  - Identify and handle missing values. Explain your method (e.g., removal, imputation).
  - Check for and handle any duplicate entries.
  - Identify outliers using visualizations (e.g., boxplots) and describe how you treated them.
- Univariate Analysis (From Project 1):
  - For numerical variables: Calculate and interpret Mean, Median, Trimmed Mean, Range, Variance, and Standard Deviation.
  - For categorical variables: Calculate frequency counts and modes.
  - Visualize distributions using histograms, KDE plots, and boxplots.
- Bivariate/Multivariate Analysis (From Project 2):
  - Create scatter plots to explore relationships between two numerical variables.
  - Create a correlation matrix and visualize it using a heatmap.
  - Use grouped boxplots or bar charts to explore relationships between categorical and numerical variables.
- Conclusion:
  - Summarize the 3-5 most important insights you discovered from your analysis.

Code:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns


# 1. Load dataset
df = pd.read_csv('C:/Users/Tarun/Downloads/tips.csv')


# 2. Data inspection
print(df.head())      # First 5 rows
print(df.tail())      # Last 5 rows
print(df.info())      # Data types and non-null count
print(df.describe())  # Summary statistics for numerical columns


# 3. Data Cleaning


# Check for missing values
print(df.isnull().sum())


# No missing values detected. If any, handle them here (e.g., imputation or removal)


# Check for duplicates
print(df.duplicated().sum())
df = df.drop_duplicates()


# 4. Identify and treat outliers
```

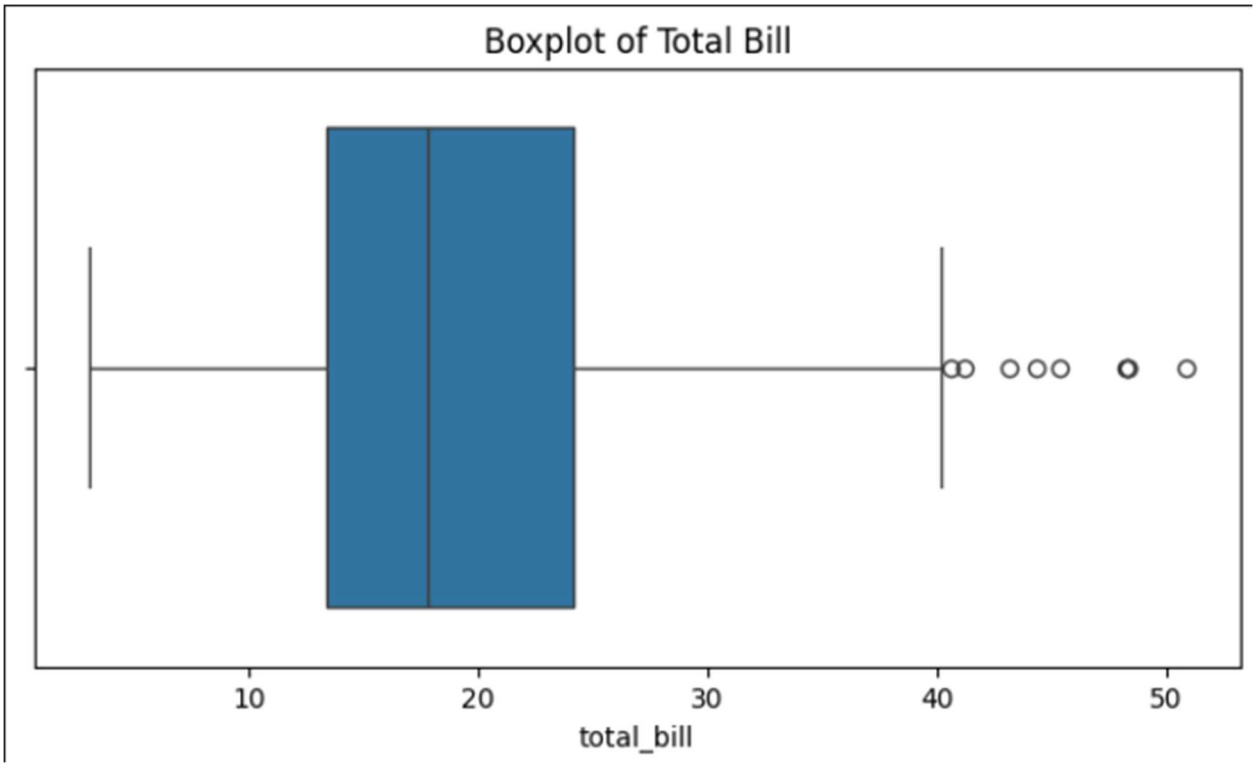
```python
# Example: boxplot for 'total_bill'
plt.figure(figsize=(8,4))
sns.boxplot(x=df['total_bill'])
plt.title('Boxplot of Total Bill')
plt.show()


# Remove outliers based on IQR for 'total_bill'
Q1 = df['total_bill'].quantile(0.25)
Q3 = df['total_bill'].quantile(0.75)
IQR = Q3 - Q1
df_filtered = df[(df['total_bill'] >= (Q1 - 1.5*IQR)) & (df['total_bill'] <= (Q3 + 1.5*IQR))]


# 5. Univariate Analysis


# Numerical variables: mean, median, trimmed mean, range, variance, std deviation for 'total_bill'
print("Mean:", df_filtered['total_bill'].mean())
print("Median:", df_filtered['total_bill'].median())
from scipy.stats import trim_mean
print("Trimmed Mean (10%):", trim_mean(df_filtered['total_bill'], 0.1))
print("Range:", df_filtered['total_bill'].max() - df_filtered['total_bill'].min())
print("Variance:", df_filtered['total_bill'].var())
print("Std deviation:", df_filtered['total_bill'].std())


# Plot histogram, KDE, boxplot for 'total_bill'
plt.figure(figsize=(12,4))
plt.subplot(1,3,1)
sns.histplot(df_filtered['total_bill'], kde=False, bins=20)
plt.title('Histogram of Total Bill')


plt.subplot(1,3,2)
sns.kdeplot(df_filtered['total_bill'])
```

```
plt.title('KDE Plot of Total Bill')


plt.subplot(1,3,3)
sns.boxplot(x=df_filtered['total_bill'])
plt.title('Boxplot of Total Bill')


plt.tight_layout()
plt.show()


# Categorical variables: frequency counts and mode for 'day'
print(df_filtered['day'].value_counts())
print("Mode of Day:", df_filtered['day'].mode()[0])


# Bar chart for 'day'
sns.countplot(x='day', data=df_filtered)
plt.title('Count of tips by Day')
plt.show()


# 6. Bivariate / Multivariate Analysis


# Scatter plot: total_bill vs tip
sns.scatterplot(x='total_bill', y='tip', data=df_filtered)
plt.title('Scatter plot: Total Bill vs Tip')
plt.show()


# Correlation matrix and heatmap
corr = df_filtered.corr(numeric_only=True)  # Added numeric_only=True to avoid error
sns.heatmap(corr, annot=True, cmap='coolwarm')
plt.title('Correlation Matrix Heatmap')
```

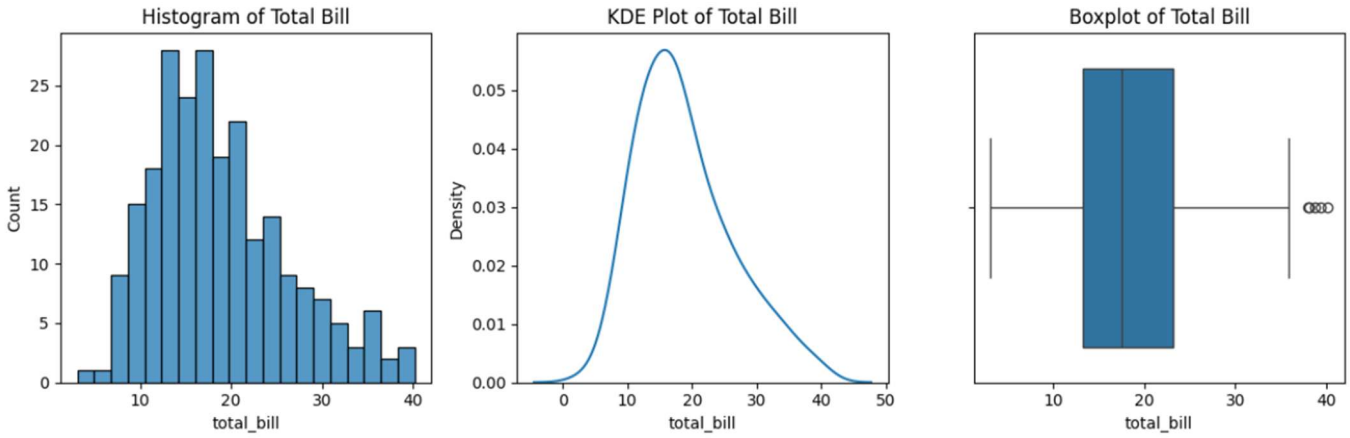
```
# Correlation matrix and heatmap
corr = df_filtered.corr(numeric_only=True)  # Added numeric_only=True to avoid error
sns.heatmap(corr, annot=True, cmap='coolwarm')
plt.title('Correlation Matrix Heatmap')
plt.show()


# Grouped boxplot: total_bill by day
sns.boxplot(x='day', y='total_bill', data=df_filtered)
plt.title('Boxplot: Total Bill by Day')
plt.show()
```

Output:

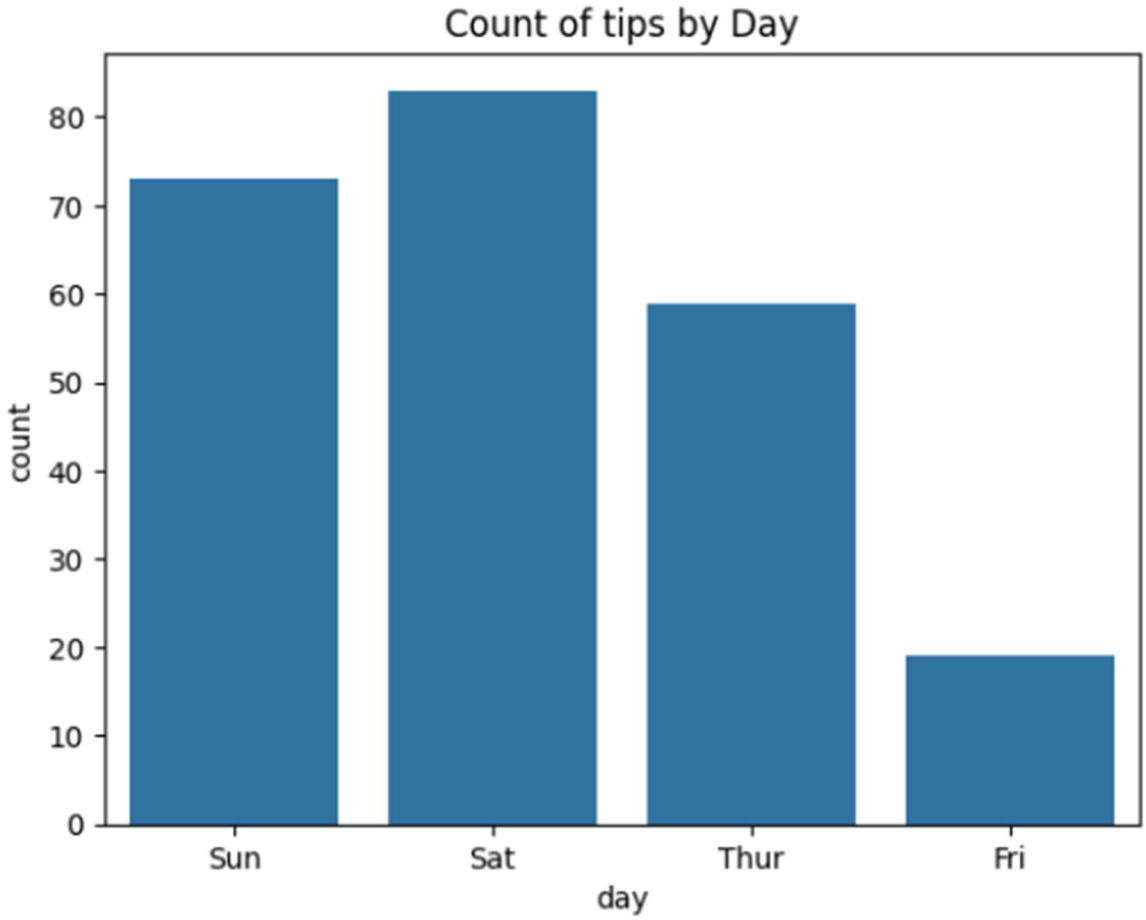
```
   total_bill   tip     sex smoker  day    time  size
0       16.99  1.01  Female     No  Sun  Dinner     2
1       10.34  1.66    Male     No  Sun  Dinner     3
2       21.01  3.50    Male     No  Sun  Dinner     3
3       23.68  3.31    Male     No  Sun  Dinner     2
4       24.59  3.61  Female     No  Sun  Dinner     4
     total_bill   tip     sex smoker   day    time  size
239       29.03  5.92    Male     No   Sat  Dinner     3
240       27.18  2.00  Female    Yes   Sat  Dinner     2
241       22.67  2.00    Male    Yes   Sat  Dinner     2
242       17.82  1.75    Male     No   Sat  Dinner     2
243       18.78  3.00  Female     No  Thur  Dinner     2
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 244 entries, 0 to 243
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   total_bill  244 non-null    float64
 1   tip         244 non-null    float64
 2   sex         244 non-null    object
 3   smoker      244 non-null    object
 4   day         244 non-null    object
 5   time        244 non-null    object
 6   size        244 non-null    int64
dtypes: float64(2), int64(1), object(4)
...
time          0
size          0
dtype: int64
1
```

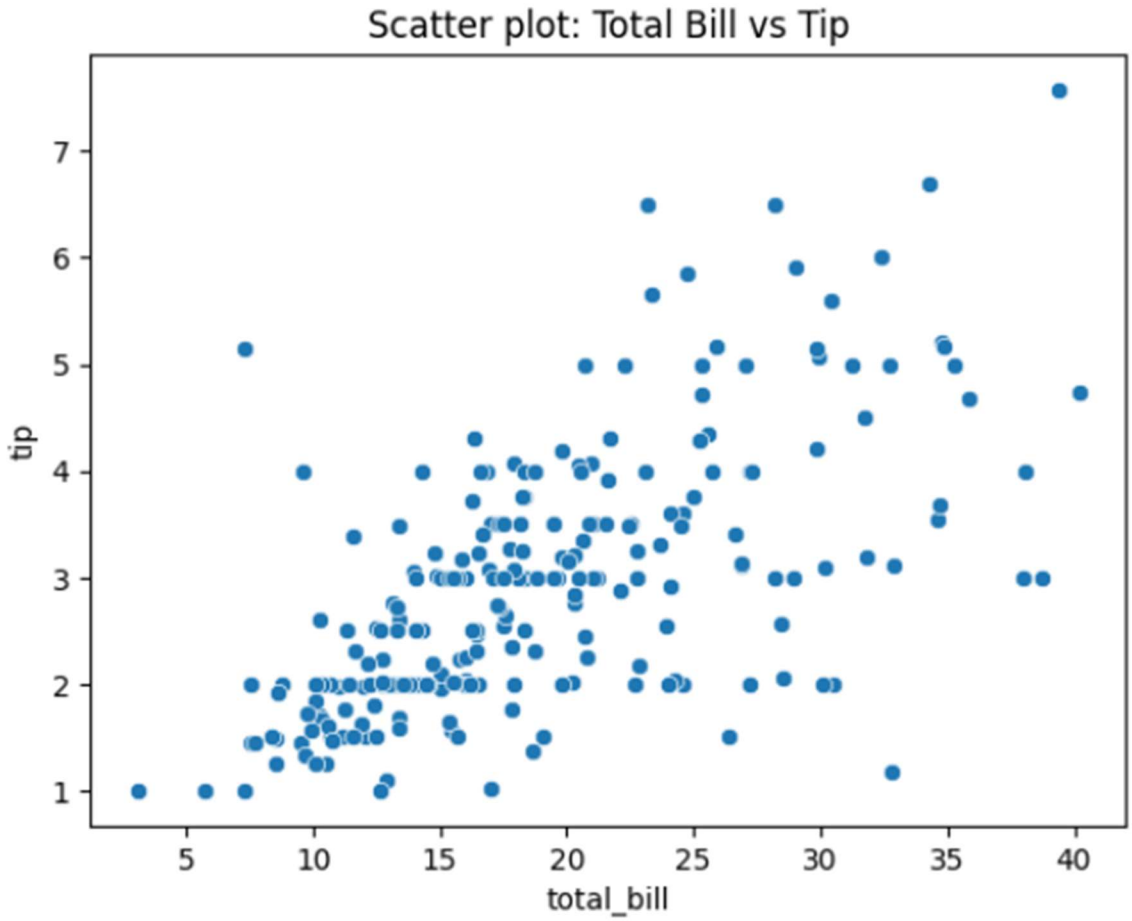


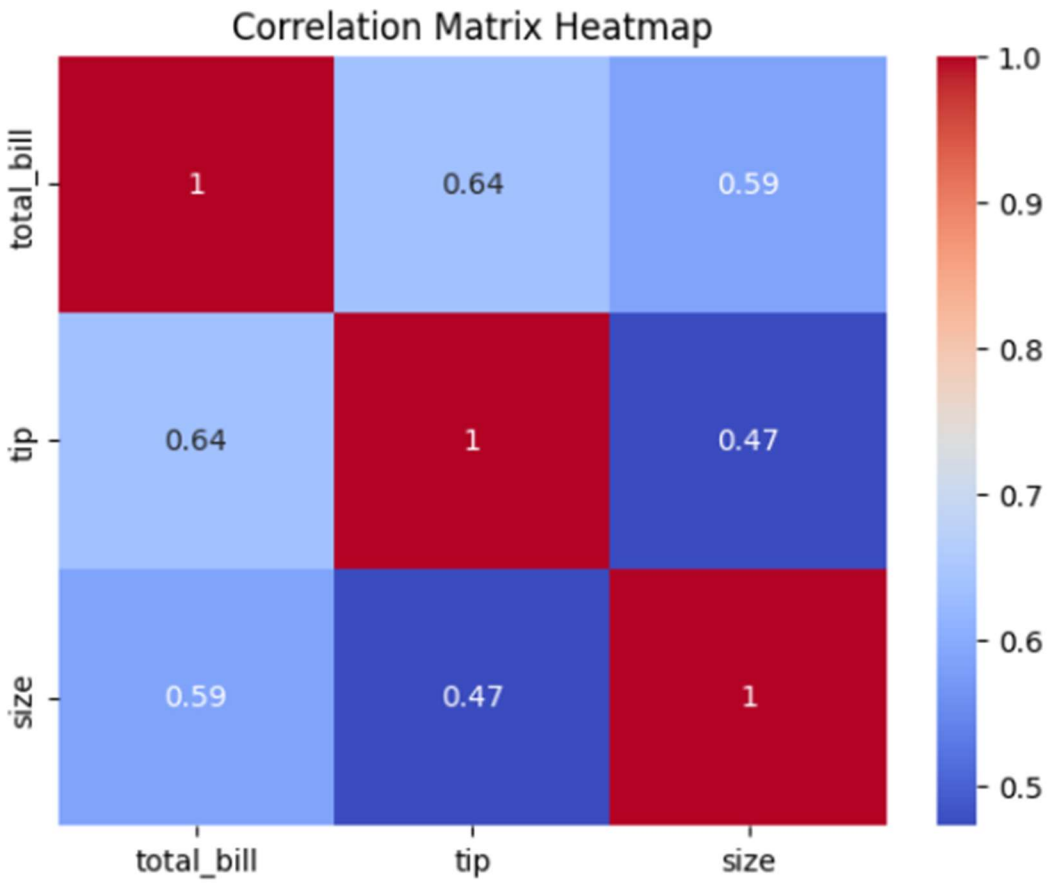

```
Mean: 18.82346153846154
Median: 17.465
Trimmed Mean (10%): 18.195478723404257
Range: 37.1
Variance: 55.42335406074612
Std deviation: 7.44468629700044
```

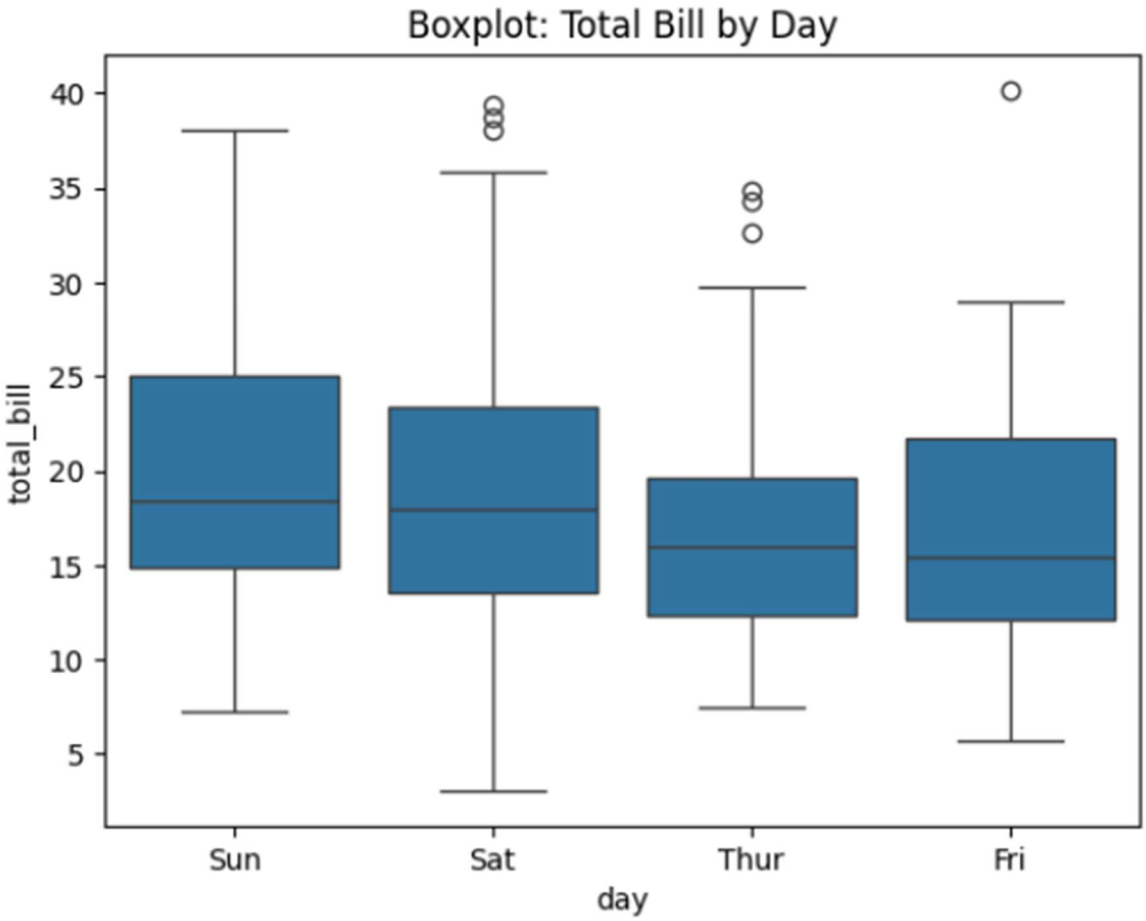


```
day
Sat     83
Sun     73
Thur    59
Fri     19
Name: count, dtype: int64
Mode of Day: Sat
```




Scatter plot: Total Bill vs Tip
tip
total_bill
1
2
3
4
5
6
7
5
10
15
20
25
30
35
40


Correlation Matrix Heatmap
total_bill
tip
size
1
0.64
0.59
0.64
1
0.47
0.59
0.47
1
total_bill
tip
size
1.0
0.9
0.8
0.7
0.6
0.5


Boxplot: Total Bill by Day
total_bill
40
35
30
25
20
15
10
5
Sun
Sat
Thur
Fri
day