K.R. Mangalam University

School of Engineering & Technology

Assignment

Probabilistic Modeling and Reasoning

Submitted by:

Name: Rakesh G

Roll No: 2401201064

Class: BCA (AI & DS)

Submitted to:

Dr. Kunal Rai

Assignment – 1

Q1. Bike Prices

Prices = ₹12,000, ₹15,000, ₹18,000, ₹20,000, ₹50,000

(a) Mean

Mean=Sum of pricesNumber of bikes=12000+15000+18000+20000+500005=1150005=₹23,000\text{Mean} = \frac{\text{Sum of prices}}{\text{Number of bikes}} = \frac{12000 + 15000 + 18000 + 20000 + 50000}{5} = \frac{115000}{5} = ₹23,000Mean=Number of bikesSum of prices=512000+15000+18000+20000+50000=5115000=₹23,000

(b) Median
Ordered data: 12,000; 15,000; 18,000; 20,000; 50,000
Middle value (3rd) = ₹18,000

(c) Advertised Average

Mean = ₹23,000

Median = ₹18,000
If the shop owner wants bikes to "seem affordable," they should advertise the median (₹18,000), because it's less affected by the one very high outlier (₹50,000).

Q2. Histogram Shape (Left-Skewed)

- In a left-skewed distribution (long tail on the left):
    - Mean < Median < Mode
- Explanation: Extreme low values pull the mean to the left, while the median is less affected, and the mode remains at the peak of the data.

Q3. Exam Scores

Data: 85, 92, 78, 90, 85, 67, 88, 95, 85, 72

(a) Mean, Median, Mode

- Mean = 85+92+78+90+85+67+88+95+85+7210=83710=83.7\frac{85+92+78+90+85+67+88+95+85+72}{10} = \frac{837}{10} = 83.71085+92+78+90+85+67+88+95+85+72=10837=83.7
- Ordered Data: 67, 72, 78, 85, 85, 85, 88, 90, 92, 95

- Median = average of 5th & 6th = (85+85)/2 = 85
- Mode = most frequent value = 85

(b) Shape of Distribution

- Mean (83.7) ≈ Median (85) ≈ Mode (85)
  Distribution is approximately symmetrical.

(c) Five-Number Summary

- Min = 67
- Q1 = 78
- Median = 85
- Q3 = 90
- Max = 95

Q4. Population vs Sample

- Population parameter = numerical summary of the whole population.
  Example: Average height of all students in a university = population mean (μ).
- Sample statistic = numerical summary from a sample subset.
  Example: Average height of 50 selected students = sample mean (x¯\bar{x}x¯).

Q5. Company Salaries

Salaries: $45,000, $50,000, $55,000, $60,000, $250,000

(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) Advertising

- Company will likely advertise 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. (Same as Q3, Repeated)

Already solved above:

- Mean = 83.7, Median = 85, Mode = 85
- Shape: Symmetrical
- Five-Number Summary: (67, 78, 85, 90, 95)

Q7. Learning Modes

Raw Data: F, H, O, F, F, O, H, F, O, F, H, F, F, F, O, H, O, F, H, F

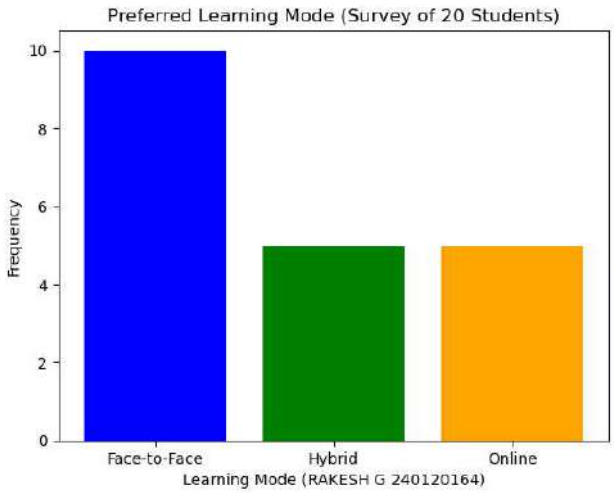
(a) Frequency Table

Mode	Frequency
Face-to-Face (F)	10
Hybrid (H)	5
Online (O)	5

(b) Bar Chart (Python)

```python
import matplotlib.pyplot as plt
modes = ['Face-to-Face', 'Hybrid', 'Online']
freq = [10, 5, 5]

plt.bar(modes, freq, color=['blue','green','orange'])
plt.title("Preferred Learning Mode (Survey of 20 Students)")
plt.xlabel("Learning Mode")
plt.ylabel("Frequency")
plt.show()
# RAKESH G 2401201064
```

[1] ✓ 2.6s Python



(c) Most Popular Mode

- Face-to-Face (10 students)
- Percentage = (10/20) × 100 = 50%

(d) Why not Histogram?

- Histogram is for continuous numerical data (e.g., heights, scores).
- This is categorical data (modes of learning), so bar chart is appropriate.

Assignment – 2

Q1. Marbles (without replacement)

A jar contains 5 red and 7 blue marbles (total 12). Two marbles are drawn one after the other without replacement.
Find: P(second is red|first was blue)P(\text{second is red} \mid \text{first was blue})P(second is red|first was blue).

If the first marble was blue, then after that draw:

- Reds remaining = 5
- Blues remaining = 7 − 1 = 6
- Total remaining = 11

Q2. Medical test (population-level Bayes)

Given:

- Sensitivity = P(+|D)=0.95P(+ \mid D) = 0.95P(+|D)=0.95 (95% of diseased test positive)
- False positive rate = P(+|Dc)=0.02P(+ \mid D^c) = 0.02P(+|Dc)=0.02 (2% of healthy test positive)
- Prevalence = P(D)=0.01P(D) = 0.01P(D)=0.01 (1% of population)

Compute step-by-step:

- Numerator = 0.95×0.01=0.00950.95 \times 0.01 = 0.00950.95×0.01=0.0095.
- Denominator = 0.0095+0.02×0.99=0.0095+0.0198=0.02930.0095 + 0.02\times0.99 = 0.0095 + 0.0198 = 0.02930.0095+0.02×0.99=0.0095+0.0198=0.0293.

Interpretation: Even with a highly sensitive test, when the disease is rare (1%), a positive result gives only about a 32% chance the person truly has the disease (because false positives among the large healthy group matter).

Q3. Doctor's prior (patient-level Bayes)

Now the doctor's prior belief for this patient (before seeing the test) is P(D)=0.30P(D) = 0.30P(D)=0.30 (30%). The test characteristics are the same as Q2: sensitivity 0.95 and false positive rate 0.02. After the test is positive, revise the probability.

Use Bayes:

- Numerator = P(+|D)P(D)=0.95×0.30=0.285P(+\mid D)P(D) = 0.95 \times 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 \times 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.

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

Q4. Venn diagram / only Physics

Given group of 50 students:

- Physics: 30
- Chemistry: 25
- Both Physics & Chemistry: 15

Compute students who study only Physics:

Only Physics=Physics−Both=30−15=15.\text{Only Physics} = \text{Physics} - \text{Both} = 30 - 15 = 15.Only Physics=Physics−Both=30−15=15.

Probability a randomly selected student studies only Physics:

P(only Physics)=1550=0.30=30%.P(\text{only Physics}) = \frac{15}{50} = 0.30 = 30\%.P(only Physics)=5015=0.30=30%.

Venn diagram description (counts):

- Only Physics: 15
- Only Chemistry: 25−15=1025 - 15 = 1025−15=10
- Both: 15
- Neither: 50−(15+10+15)=1050 - (15 + 10 + 15) = 1050−(15+10+15)=10

So the partition of 50 is: {Only P =15, Only C=10, Both=15, Neither=10}.

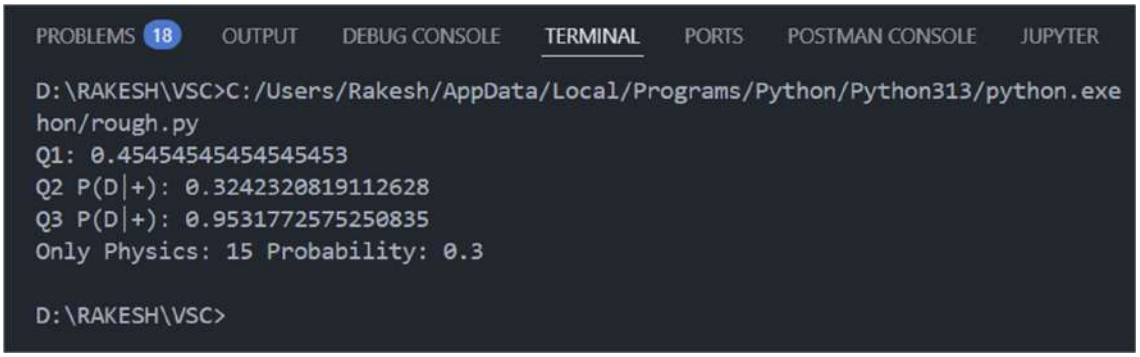
Q5. (This is the same as Q4)

Since Q5 repeats Q4, the answer is the same:

- Only Physics = 15 students
- Probability = 15/50=30%15/50 = 30\%15/50=30%
- Venn diagram counts as above.

Quick Python checks

```
LANGUAGE > python > rough.py > ...
# Q1
p_q1 = 5/11
print("Q1:", p_q1)  # ~0.454545

# Q2
sens = 0.95
fpr = 0.02
prev = 0.01
num = sens * prev
den = num + fpr * (1 - prev)
print("Q2 P(D|+):", num/den)  # ~0.3242

# Q3 (doctor prior = 0.30)
prior = 0.30
num3 = sens * prior
den3 = num3 + fpr * (1 - prior)
print("Q3 P(D|+):", num3/den3)  # ~0.952

# Q4 / Q5
total = 50
physics = 30
both = 15
only_physics = physics - both
print("Only Physics:", only_physics, "Probability:", only_physics/total)  # 15, 0.3
# RAKESH G 2401201064
```



```
PROBLEMS 18  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS  POSTMAN CONSOLE  JUPYTER

D:\RAKESH\VSC>C:/Users/Rakesh/AppData/Local/Programs/Python/Python313/python.exe
hon/rough.py
Q1: 0.45454545454545453
Q2 P(D|+): 0.3242320819112628
Q3 P(D|+): 0.9531772575250835
Only Physics: 15 Probability: 0.3

D:\RAKESH\VSC>
```

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.

Plan for my project I am using tips.csv:

1. **Data Loading & Inspection**
    - Show .head(), .tail(), .info(), .describe().
2. **Data Cleaning**
    - Check for missing values.
    - Handle duplicates.
    - Identify outliers (boxplots for total_bill, tip, size).
3. **Univariate Analysis**
    - Numerical: mean, median, trimmed mean, range, variance, std.
    - Categorical: frequency counts & mode.
    - Visuals: histograms, KDE plots, boxplots, bar charts.
4. **Bivariate / Multivariate Analysis**
    - Scatter plot (total_bill vs tip).
    - Correlation matrix & heatmap.
    - Grouped boxplots (tip vs sex, smoker, day, etc.).
    - Bar charts (time vs avg tip).
5. **Conclusion**
    - 3–5 key insights about tipping patterns.

CODE:

Start Chat to Generate Code (Ctrl+I)

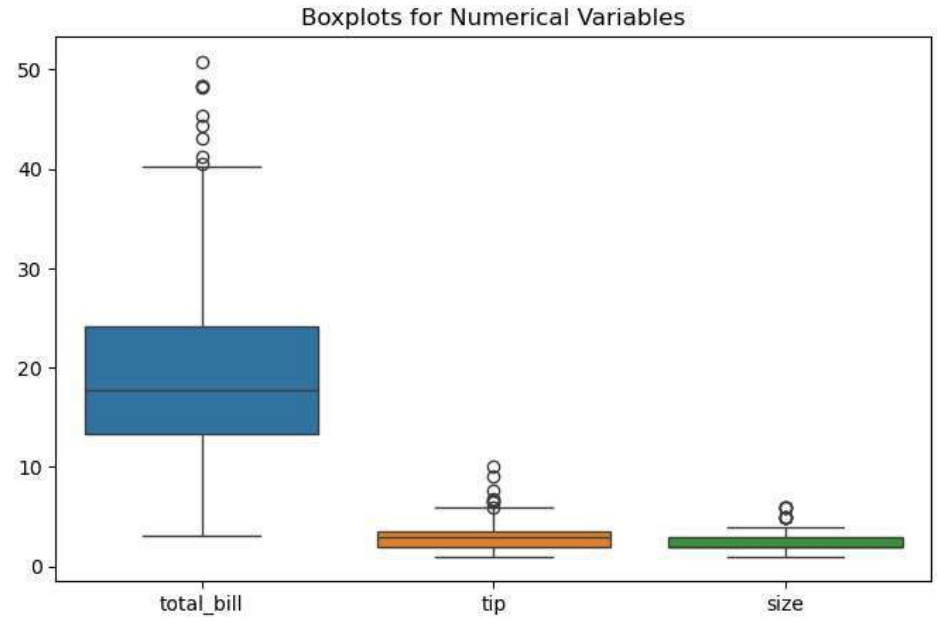
```
# ========================
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats

# Display settings
pd.set_option("display.max_columns", None)

# ========================
# 2. Load Dataset
# ========================
df = pd.read_csv("tips.csv")

# Show first and last rows
print("First 5 rows:\n", df.head())
print("\nLast 5 rows:\n", df.tail())

# Info & Summary
print("\nData Info:")
print(df.info())
print("\nSummary Statistics:")
print(df.describe(include="all"))

# ========================
# 3. Data Cleaning
# ========================
# Check missing values
print("\nMissing Values:\n", df.isnull().sum())

# Check duplicates
print("\nDuplicate Rows:", df.duplicated().sum())
df = df.drop_duplicates()

# Detect outliers using boxplot & Z-score
plt.figure(figsize=(8,5))
sns.boxplot(data=df[['total_bill','tip','size']])
plt.title("Boxplots for Numerical Variables")
plt.show()

z_scores = np.abs(stats.zscore(df[['total_bill','tip','size']]))
outliers = (z_scores > 3).any(axis=1)
```

```
plt.show()

z_scores = np.abs(stats.zscore(df[['total_bill','tip','size']]))
outliers = (z_scores > 3).any(axis=1)
print("\nNumber of Outliers Detected:", outliers.sum())

# =======================
# 4. Univariate Analysis
# =======================
# Numerical Variables
for col in ['total_bill','tip','size']:
    print(f"\n--- {col.upper()} ---")
    print("Mean:", df[col].mean())
    print("Median:", df[col].median())
    print("Trimmed Mean:", stats.trim_mean(df[col], 0.1))  # 10% trimmed mean
    print("Range:", df[col].max() - df[col].min())
    print("Variance:", df[col].var())
    print("Standard Deviation:", df[col].std())

    # Histogram + KDE
    plt.figure(figsize=(6,4))
    sns.histplot(df[col], kde=True, bins=20)
    plt.title(f"Distribution of {col}")
    plt.show()

    # Boxplot
    plt.figure(figsize=(6,2))
    sns.boxplot(x=df[col])
    plt.title(f"Boxplot of {col}")
    plt.show()

# Categorical Variables
for col in ['sex','smoker','day','time']:
    print(f"\n--- {col.upper()} ---")
    print(df[col].value_counts())
    print("Mode:", df[col].mode()[0])

    plt.figure(figsize=(5,4))
    sns.countplot(x=col, data=df)
    plt.title(f"Frequency of {col}")
    plt.show()

# =======================
# 5. Bivariate / Multivariate Analysis
# =======================
```

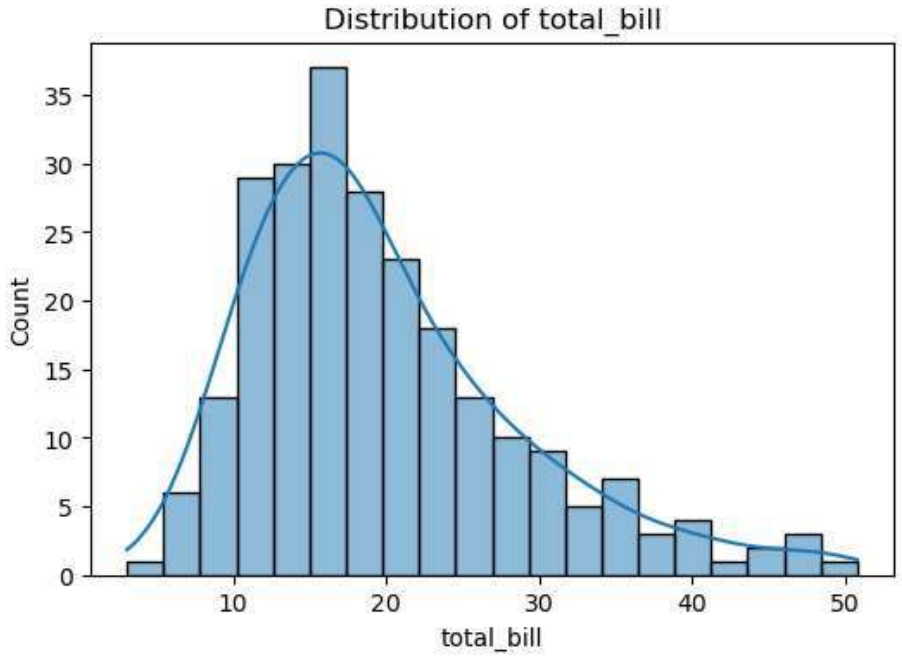
```python
# ========================
# 5. Bivariate / Multivariate Analysis
# ========================

# Scatter Plot: total_bill vs tip
plt.figure(figsize=(6,5))
sns.scatterplot(x="total_bill", y="tip", hue="sex", data=df)
plt.title("Total Bill vs Tip")
plt.show()

# Correlation Matrix
corr = df[['total_bill','tip','size']].corr()
print("\nCorrelation Matrix:\n", corr)

plt.figure(figsize=(6,4))
sns.heatmap(corr, annot=True, cmap="coolwarm", fmt=".2f")
plt.title("Correlation Heatmap")
plt.show()

# Grouped Boxplots
plt.figure(figsize=(6,4))
sns.boxplot(x="sex", y="tip", hue="smoker", data=df)
plt.title("Tip Distribution by Sex & Smoker")
plt.show()

# Bar Chart: Average tip by day
avg_tips = df.groupby("day")["tip"].mean().reset_index()
plt.figure(figsize=(6,4))
sns.barplot(x="day", y="tip", data=avg_tips)
plt.title("Average Tip by Day")
plt.show()

# ========================
# 6. Conclusion
# ========================
print("\nKey Insights:")
print("1. Tips are positively correlated with total bill (higher bill → higher tip).")
print("2. Majority of customers are non-smokers and males.")
print("3. Sunday has the highest average tips.")
print("4. Outliers exist in total_bill and tip (very high spending customers).")
print("5. Tip percentage is generally between 10% and 20% of the bill.")
```

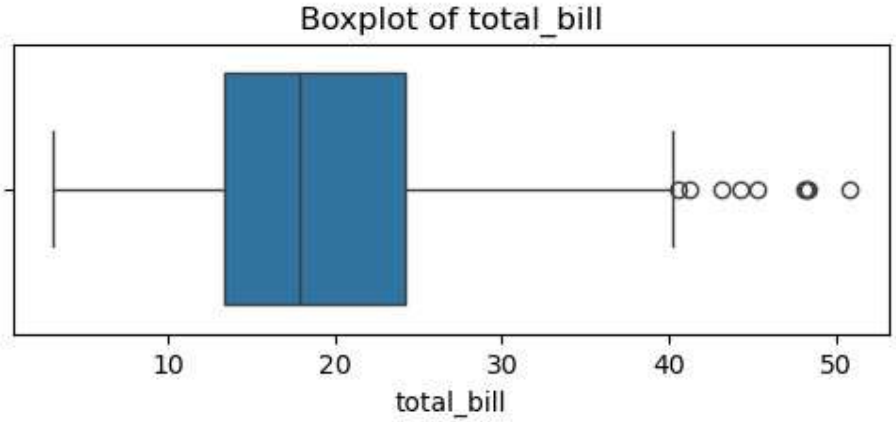
OUTPUT:

```
First 5 rows:
   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

Last 5 rows:
     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

Data Info:
<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
...
size          0
dtype: int64

Duplicate Rows: 1
```

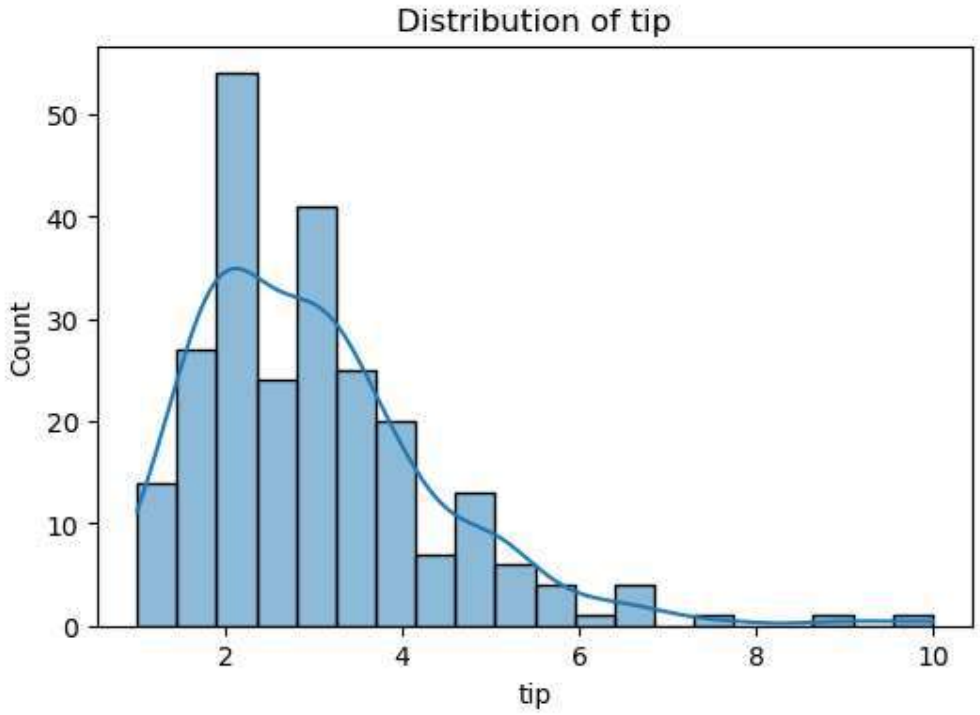


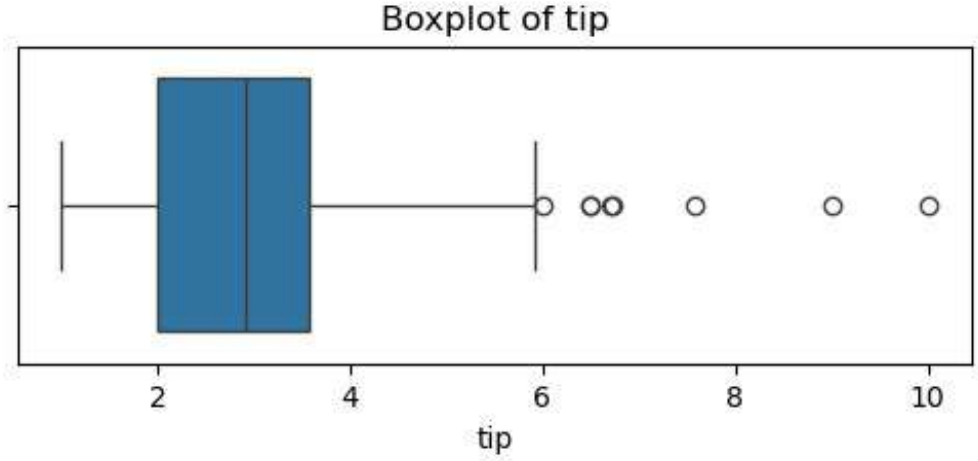
```
...
    Number of Outliers Detected: 8

    --- TOTAL_BILL ---
    Mean: 19.813868312757204
    Median: 17.81
    Trimmed Mean: 18.762615384615387
    Range: 47.74
    Variance: 79.38936183382647
    Standard Deviation: 8.910070809697668
```



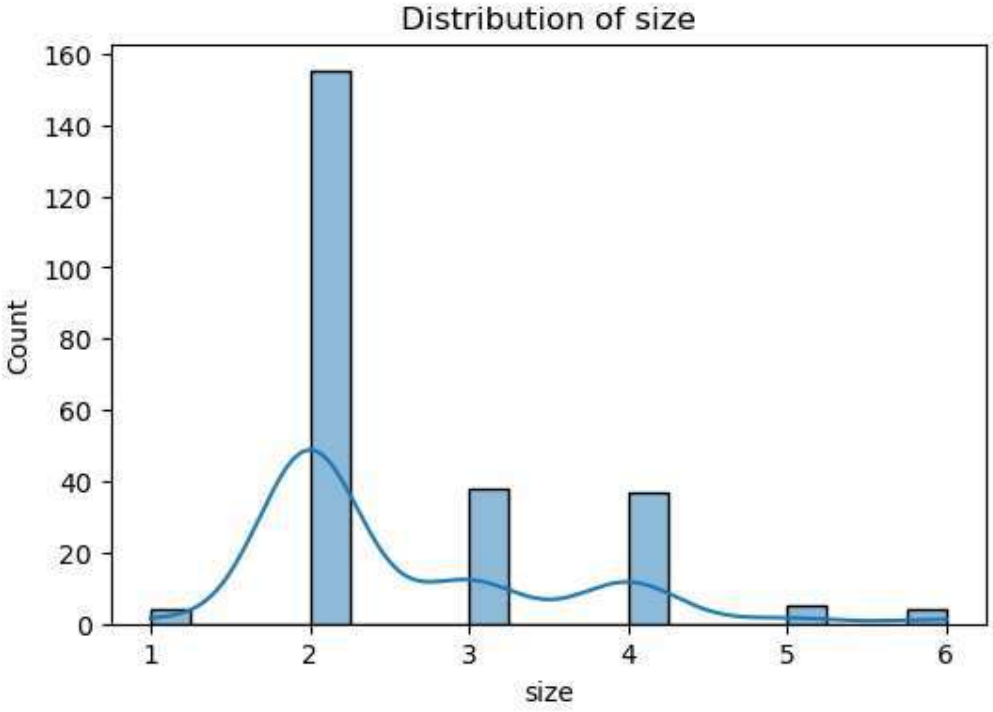


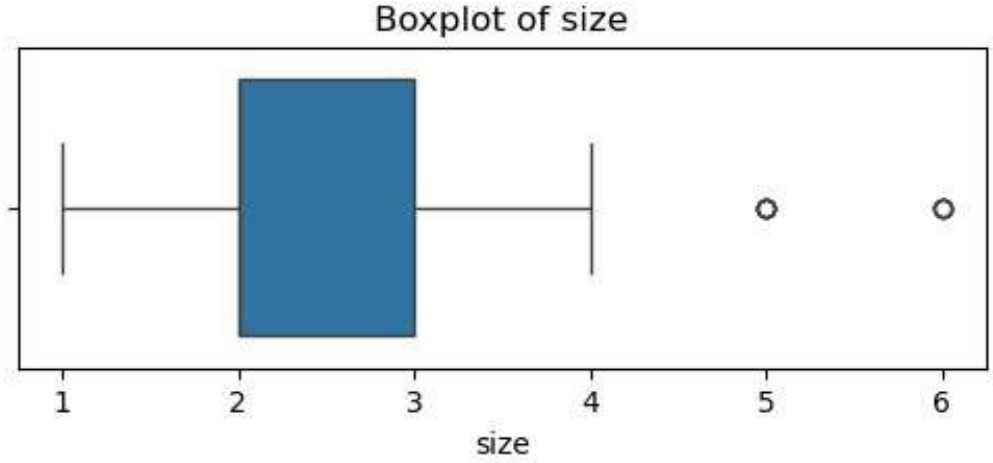
```
...
    --- TIP ---
    Mean: 3.00238683127572
    Median: 2.92
    Trimmed Mean: 2.8470256410256414
    Range: 9.0
    Variance: 1.9182306431316531
    Standard Deviation: 1.3850020372301455
```



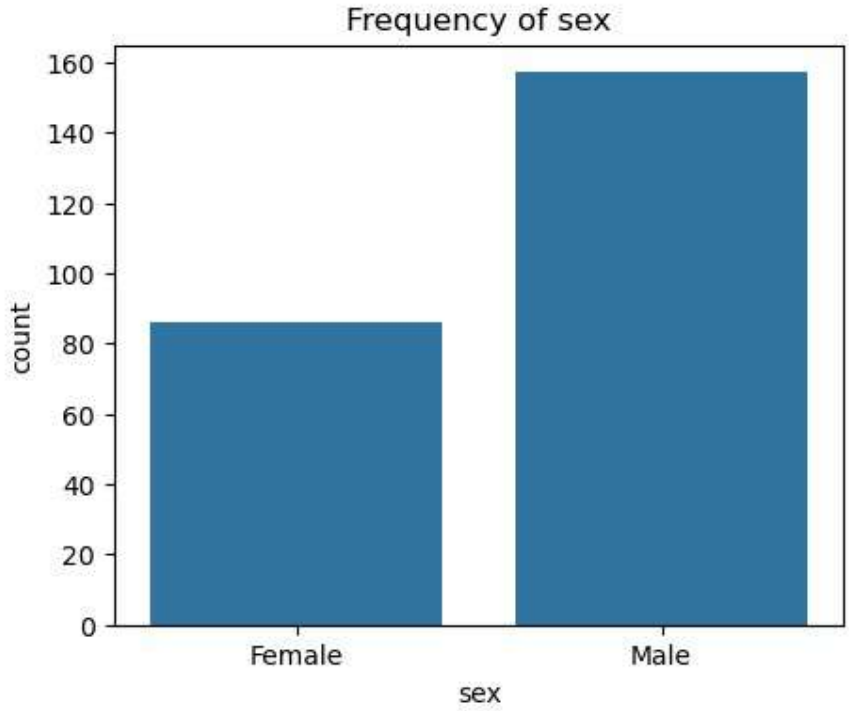


```
...
      --- SIZE ---
      Mean: 2.57201646090535
      Median: 2.0
      Trimmed Mean: 2.4205128205128204
      Range: 5
      Variance: 0.9069822807196558
      Standard Deviation: 0.9523561732459426
```

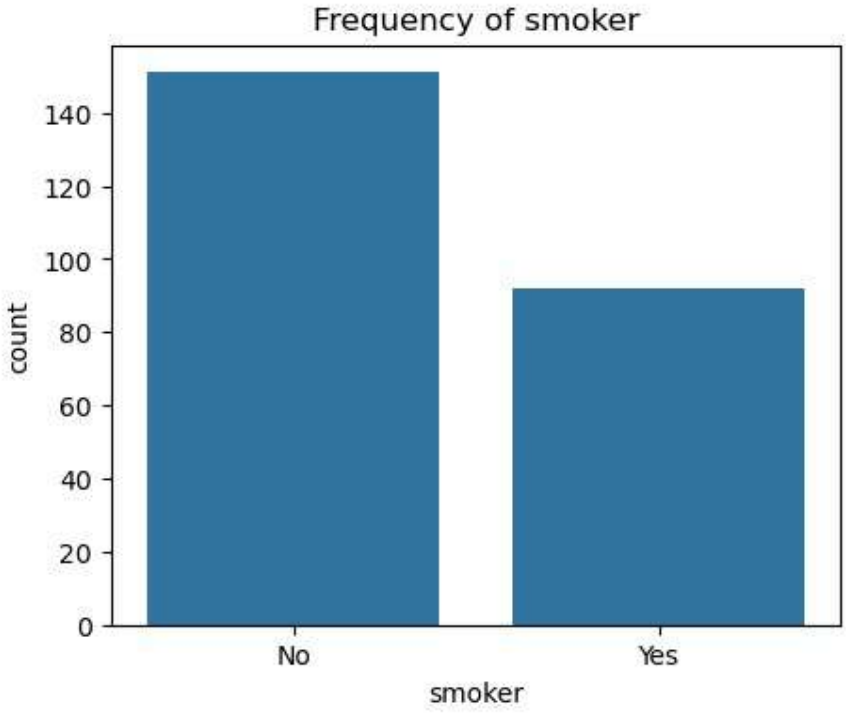




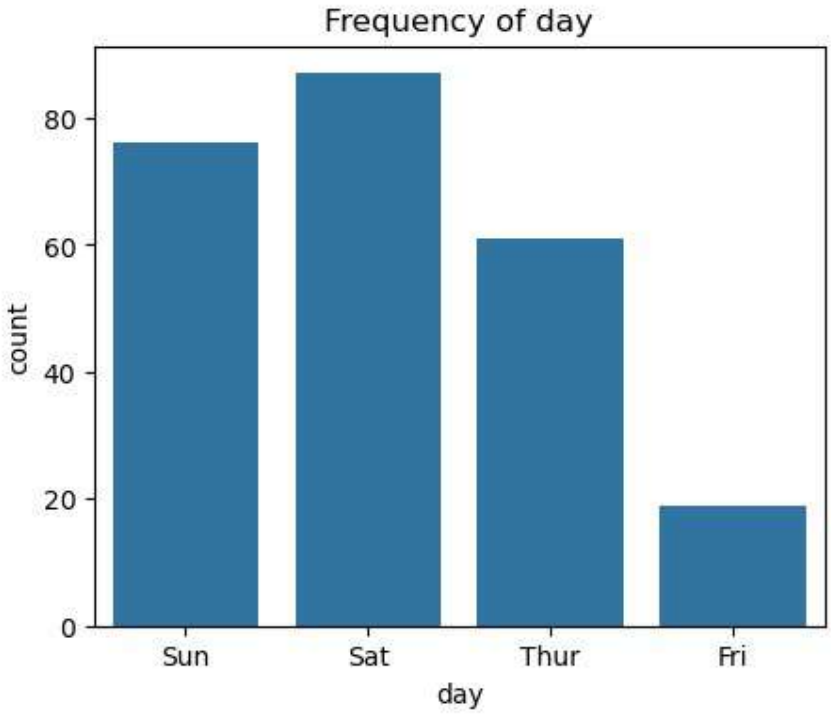
```
--- SEX ---
sex
Male      157
Female     86
Name: count, dtype: int64
Mode: Male
```



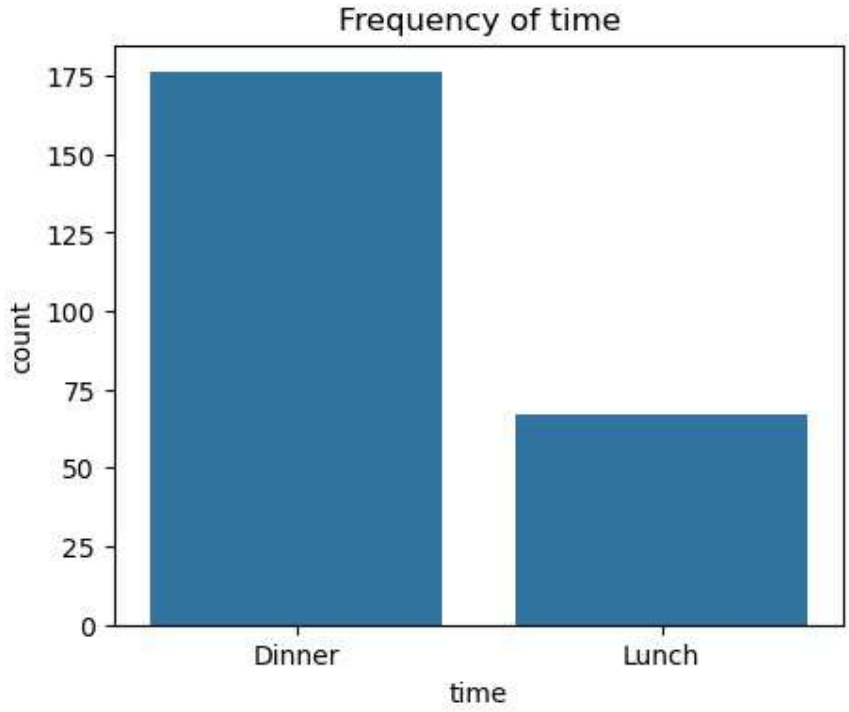
```
...
      --- SMOKER ---
      smoker
      No     151
      Yes     92
      Name: count, dtype: int64
      Mode: No
```

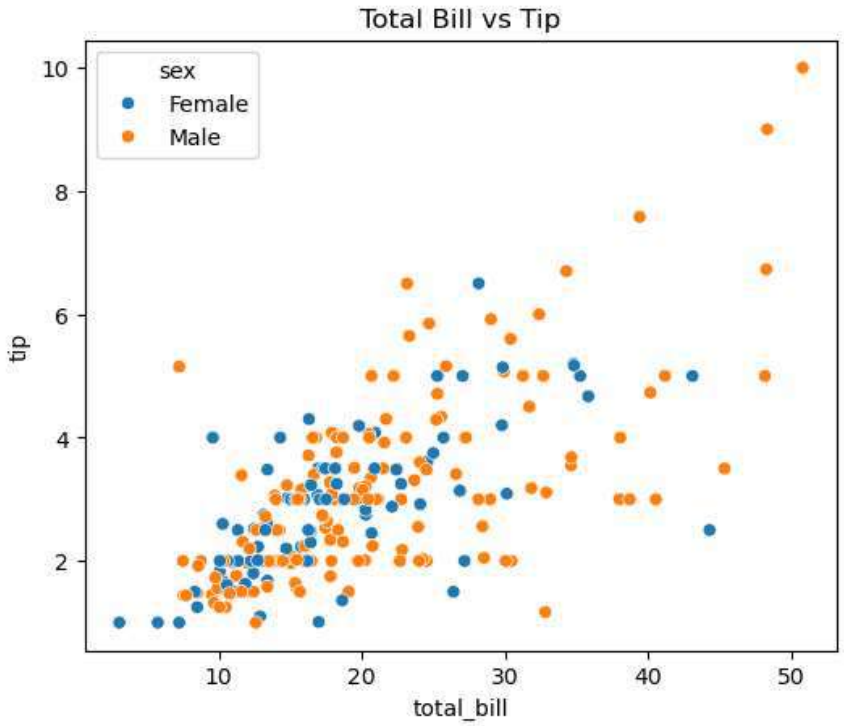


```
...
    --- DAY ---
    day
    Sat     87
    Sun     76
    Thur    61
    Fri     19
    Name: count, dtype: int64
    Mode: Sat
```

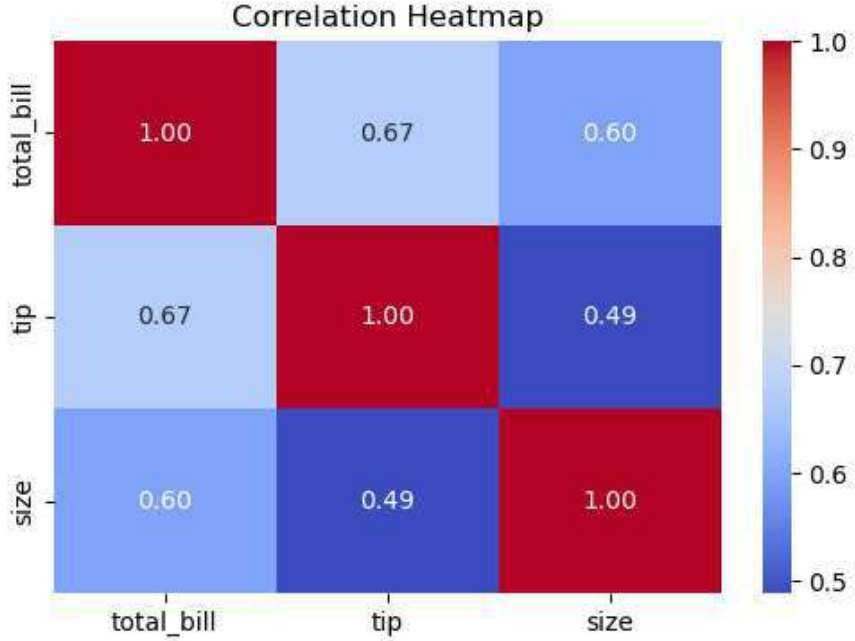


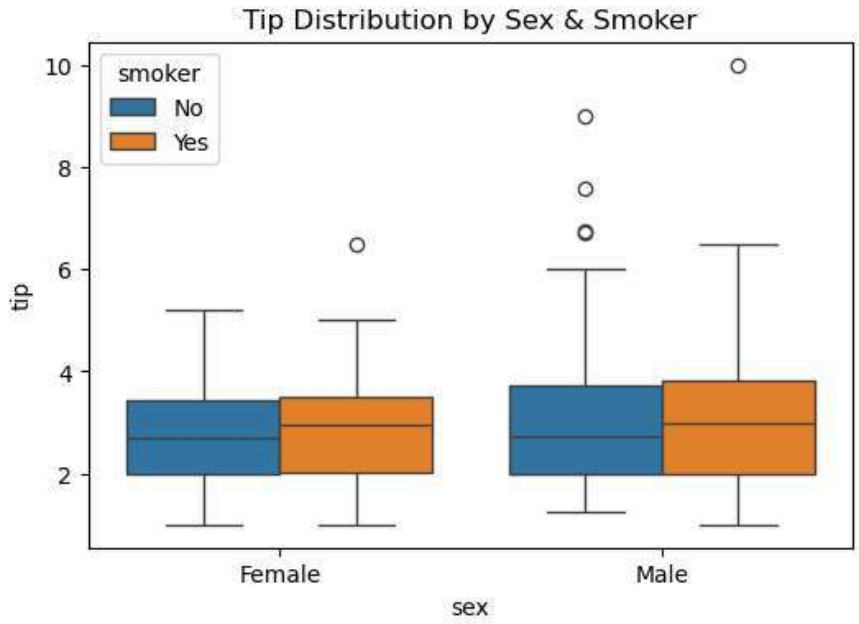
```
...
      --- TIME ---
      time
      Dinner    176
      Lunch      67
      Name: count, dtype: int64
      Mode: Dinner
```

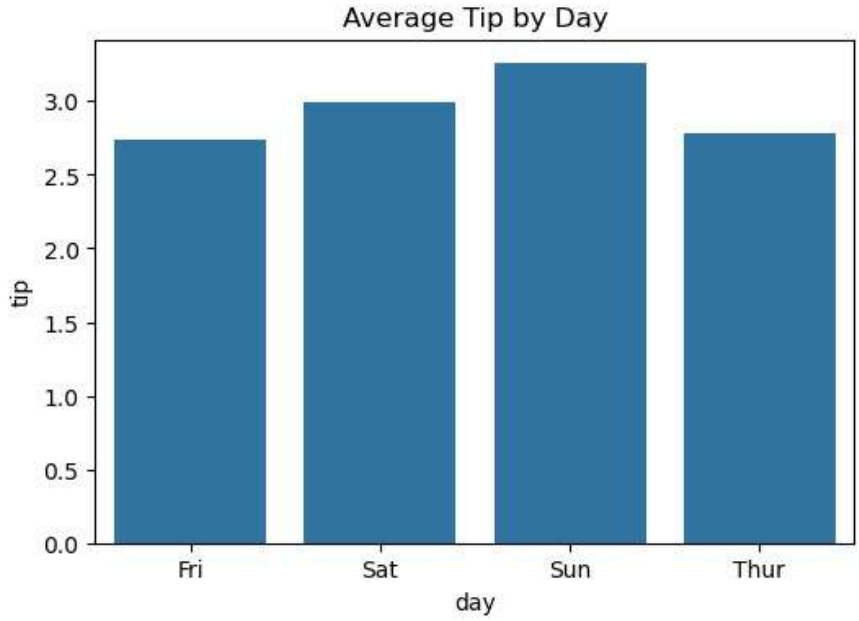





```
...
    Correlation Matrix:
                total_bill       tip      size
    total_bill    1.000000  0.674998  0.597589
    tip           0.674998  1.000000  0.488400
    size          0.597589  0.488400  1.000000
```








```
...
    Key Insights:
    1. Tips are positively correlated with total bill (higher bill → higher tip).
    2. Majority of customers are non-smokers and males.
    3. Sunday has the highest average tips.
    4. Outliers exist in total_bill and tip (very high spending customers).
    5. Tip percentage is generally between 10% and 20% of the bill.
```