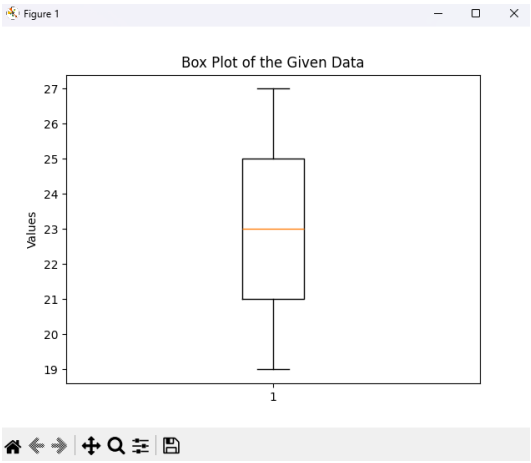
Name: Annie Mahajan
Registration No.: 23BRS1121

Lab – 1
Introduction and Box Plot

Example 1 — Basic Box Plot

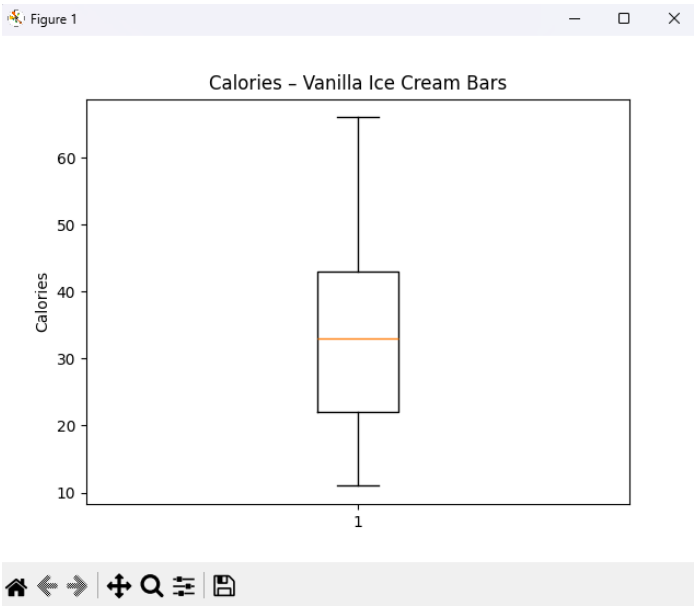
```
import matplotlib.pyplot as plt

data = [20, 19, 21, 22, 23, 24, 25, 27, 26]

plt.boxplot(data)

plt.title("Box Plot of the Given Data")

plt.ylabel("Values")

plt.show()
```



Example 2 — Dispersion of Vanilla Ice Cream Calories

```
import pandas as pd

import matplotlib.pyplot as plt

df = pd.read_csv("Lab1_Vanilla.csv")
```

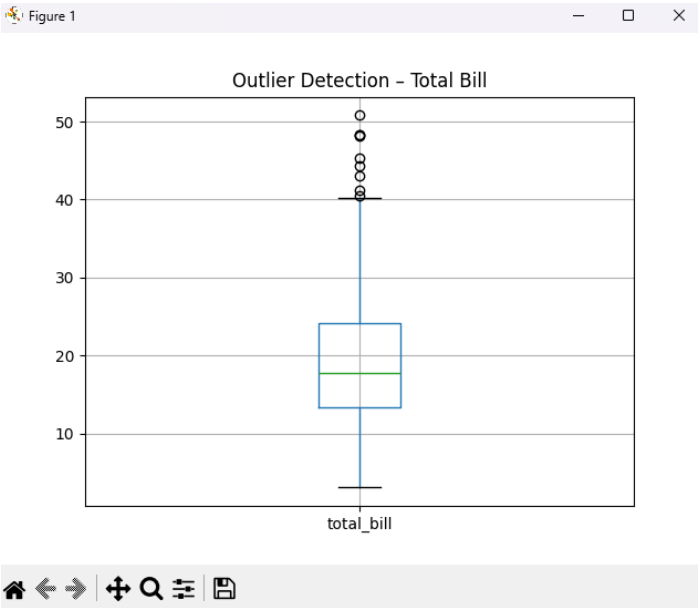
```
print("Mean:", df.iloc[:,0].mean())

print("Median:", df.iloc[:,0].median())

print("Mode:", df.iloc[:,0].mode()[0])

print("\nFive-point summary:\n", df.describe())

Q1 = df.iloc[:,0].quantile(0.25)

Q3 = df.iloc[:,0].quantile(0.75)

IQR = Q3 - Q1

lower = Q1 - 1.5 * IQR

upper = Q3 + 1.5 * IQR

print("\nOutliers:")

print(df[(df.iloc[:,0] < lower) | (df.iloc[:,0] > upper)])

plt.boxplot(df.iloc[:,0])

plt.title("Calories – Vanilla Ice Cream Bars")

plt.ylabel("Calories")

plt.show()
```



```
PS C:\Users\scope1\Downloads\ML> & "C:/Program Files/Python310/python.exe" c:/Users/scope1/Downloads/ML/ex2.py
Mean: 31.88235294117647
Median: 33.0
Mode: 22
Five-Point Summary:
              45
count  17.000000
mean   31.882353
std    13.846851
min    11.000000
25%    22.000000
50%    33.000000
75%    43.000000
max    66.000000

Outliers:
Empty DataFrame
Columns: [45]
Index: []
```

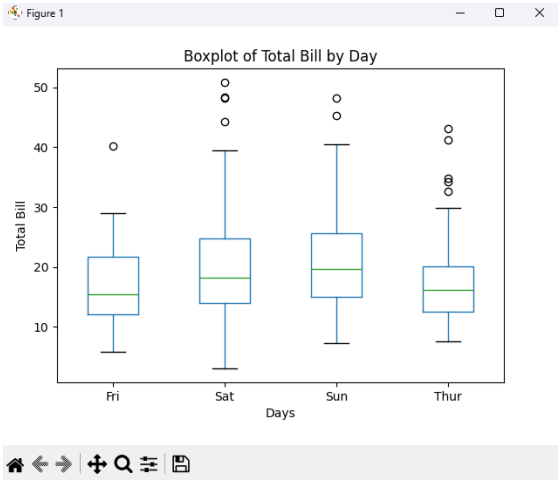
Example 3 — Outlier Detection in Tips Dataset

```
import pandas as pd

import matplotlib.pyplot as plt

df = pd.read_csv("Lab1_Tips.csv")

df.boxplot(column=['total_bill'])

plt.title("Outlier Detection – Total Bill")

plt.show()
```



Example 4 — Boxplot by Day (Grid Off)

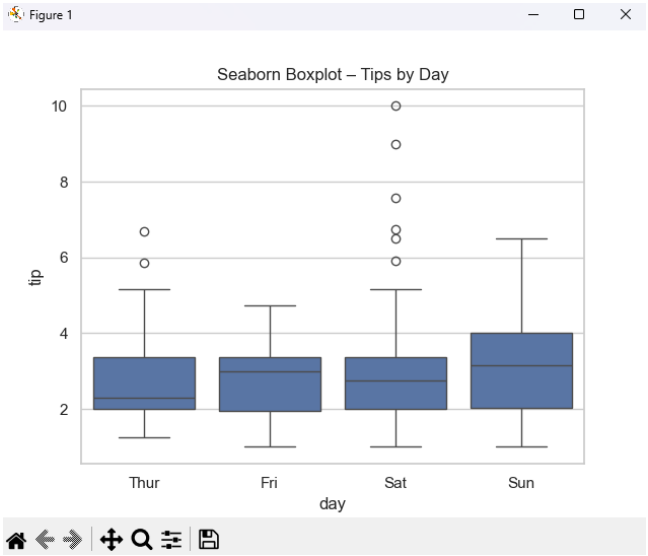
```
import pandas as pd

import matplotlib.pyplot as plt

df = pd.read_csv("Lab1_Tips.csv")

df.boxplot(by='day', column=['total_bill'], grid=False)

plt.ylabel("Total Bill")

plt.xlabel("Day")

plt.title("Boxplot of Total Bill by Day")

plt.suptitle("")

plt.show()
```



Example 5 — Seaborn Boxplot

```
import seaborn as sns

import matplotlib.pyplot as plt

sns.set(style='whitegrid')

tips = sns.load_dataset("tips")

sns.boxplot(x='day', y='tip', data=tips)
```

```
plt.title("Seaborn Boxplot – Tips by Day")

plt.show()
```

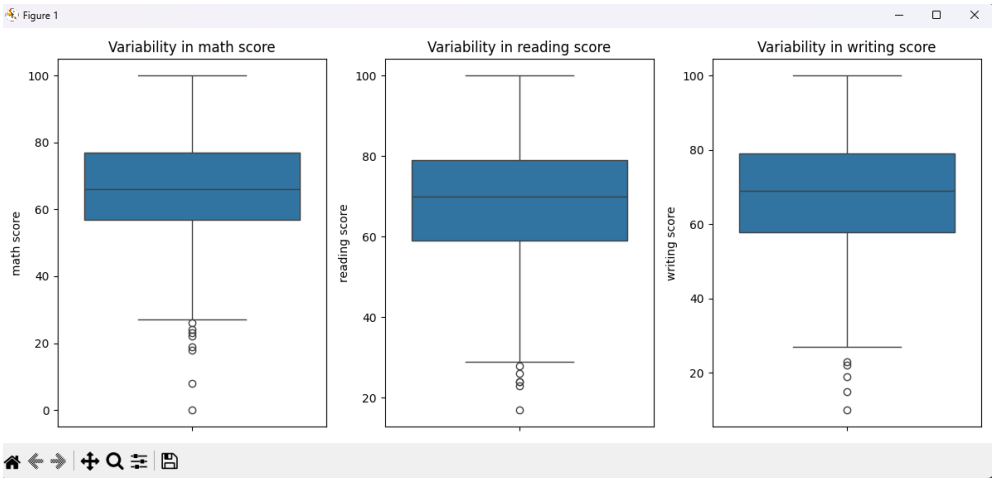


Example 6 — Variability in Student Performance Dataset

```
import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

df = pd.read_csv("Lab1_StudentsPerformance.csv")

numeric_cols = ["math score", "reading score", "writing score"]

print("Summary Statistics:")

print(df[numeric_cols].describe())

plt.figure(figsize=(12, 4))

for i, col in enumerate(numeric_cols):

    plt.subplot(1, 3, i+1)

    sns.boxplot(y=df[col])

    plt.title("Variability in " + col)
```

plt.tight_layout()

plt.show()




```
PS C:\Users\scope1\Downloads\ML> & "C:/Program Files/Python310/python.exe" c:/Users/scope1/Downloads/ML/ex6.py
First 5 rows:
   gender race/ethnicity parental level of education         lunch test preparation course  math score  reading score  writing score
0  female        group B           bachelor's degree      standard                    none          72             72             74
1  female        group C                some college      standard               completed          69             90             88
2  female        group B             master's degree      standard                    none          90             95             93
3    male        group A          associate's degree  free/reduced                    none          47             57             44
4    male        group C                some college      standard                    none          76             78             75

Summary Statistics:
       math score  reading score  writing score
count  1000.00000    1000.000000    1000.000000
mean     66.08900      69.169000      68.054000
std      15.16308      14.600192      15.195657
min       0.00000      17.000000      10.000000
25%      57.00000      59.000000      57.750000
50%      66.00000      70.000000      69.000000
75%      77.00000      79.000000      79.000000
max     100.00000     100.000000     100.000000
```

Exercise: Apply Variability to tips.csv

import pandas as pd

import seaborn as sns

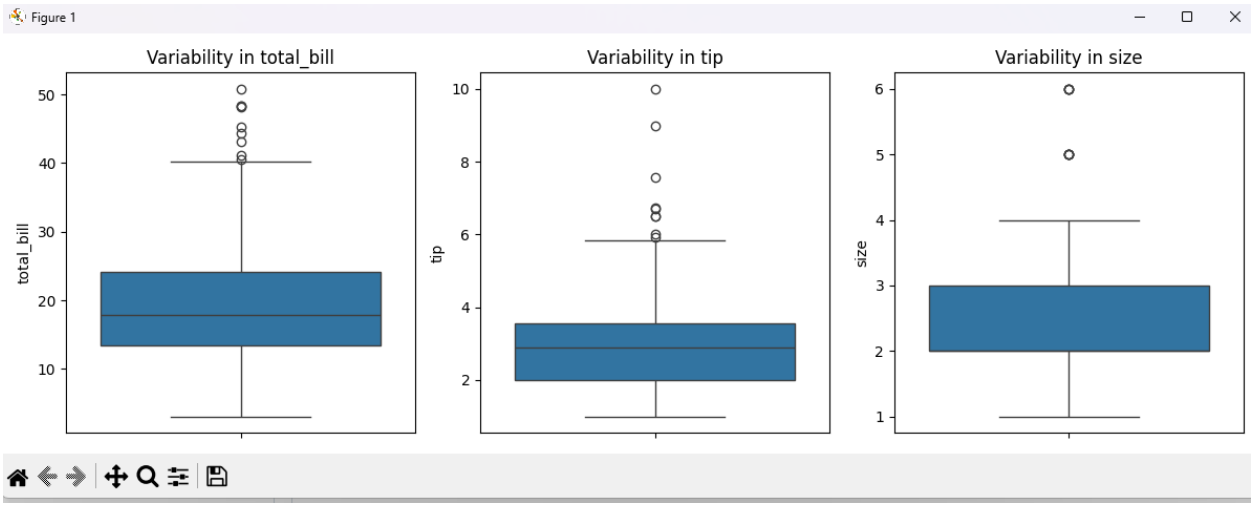
import matplotlib.pyplot as plt

df = pd.read_csv("Lab1_Tips.csv")

numeric_cols = ["total_bill", "tip", "size"]

print("Summary Statistics:")

```
print(df[numeric_cols].describe())

plt.figure(figsize=(12, 4))

for i, col in enumerate(numeric_cols):
    plt.subplot(1, 3, i+1)
    sns.boxplot(y=df[col])
    plt.title("Variability in " + col)

plt.tight_layout()

plt.show()
```




```
PS C:\Users\scope1\Downloads\ML> & "C:/Program Files/Python310/python.exe" c:/Users/scope1/Downloads/ML/Q1.py
Summary Statistics:
       total_bill         tip        size
count  244.000000  244.000000  244.000000
mean    19.785943    2.998279    2.569672
std      8.902412    1.383638    0.951100
min      3.070000    1.000000    1.000000
25%     13.347500    2.000000    2.000000
50%     17.795000    2.900000    2.000000
75%     24.127500    3.562500    3.000000
max     50.810000   10.000000    6.000000
```