CREATING GRAPHS & TABLES

Big Data Algorithms and Statistic 01
BDM 2053

Group ‘D’ Team Members

AJAY KUMAR
C0942384

HAZEL PORTIA ELAINE SANTOS
C0915982

SNEH PATHAK
C0938327

POOJA SHRESTHA
C0931754

SHRIYA UPADHYAY
C0938089

Table of Contents

Q1: Frequency Table

Output Screenshot

```
Frequency Table for Attribute 1
        Bin Range  Frequency
0  (1.098, 1.725]         13
1   (1.725, 2.35]         11
2    (2.975, 3.6]         14
3   (2.35, 2.975]         14

Frequency Table for Attribute 2
        Bin Range  Frequency
0   (2.825, 3.55]         12
1  (2.097, 2.825]         16
2   (3.55, 4.275]         14
3    (4.275, 5.0]         10
```

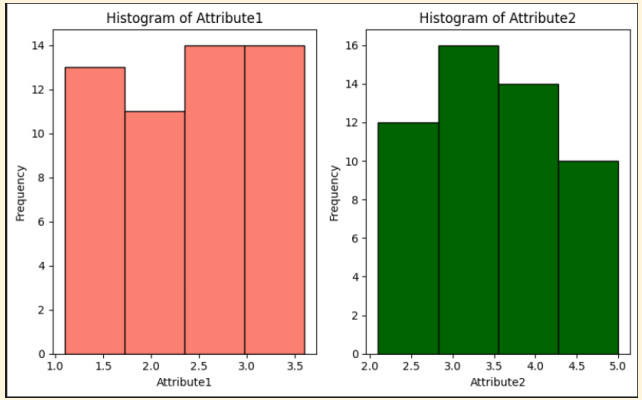
ATTRIBUTE 1: The data is distributed relatively evenly across the bins. The first two bins contain slightly fewer data points (13 and 11, respectively), while the last two bins have slightly more (14 each). This suggests that the values in Attribute 1 are somewhat concentrated in the upper half of the range.

ATTRIBUTE 2: The distribution shows more variation, with the first bin having the highest count (16 data points), and the remaining bins showing a more balanced distribution. This indicates that the values in Attribute 2 are more concentrated in the lower ranges, with fewer values in the higher bins.

Overall, the frequency tables provide insight into how the values for each attribute are spread across the defined bins. If some bins have significantly higher frequencies, it means a large portion of the data falls within those ranges. In contrast, more equal frequencies suggest a uniform data spread across the attribute's range.

Q2: Histogram

Output Screenshot



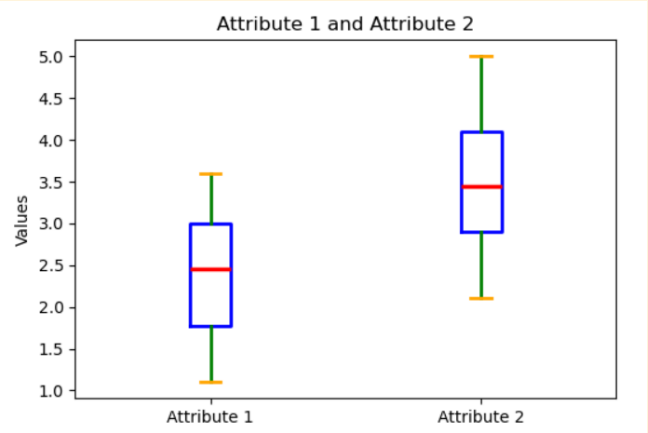
The histograms for both attributes shows how their values are distributed across the data range.

From Attribute1 histogram, the data appears evenly distributed across the range, yet there is a concentration of values in the higher bins.bin 3 & 4 have equal frequency.

Attribute2 histogram also shows a generally even distribution, with more values clustered towards the middle, it looks like a right skewed data distribution.

Q3: BOXPLOT

Output Screenshot



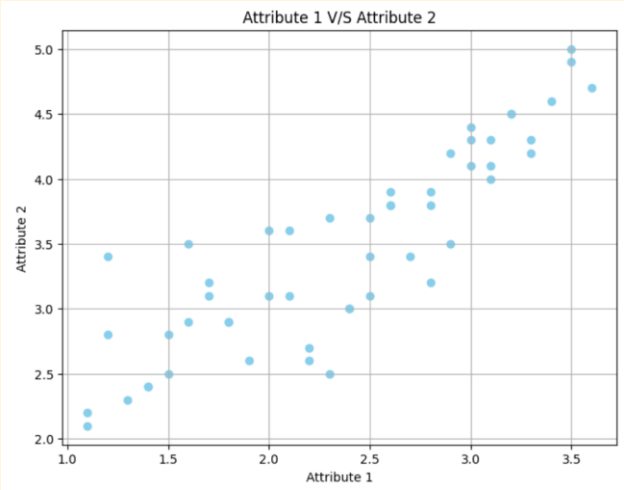
Attribute 1 has a minimum value of 1.1 and a maximum of 3.6, the range of it is 2.5 and the interquartile result is 1.225. Attribute 2 has a minimum value of 2.1 and a maximum value of 5.0. The range of it is 2.9 with an interquartile result of 1.2

Attribute 2 returns the highest value and the highest minimum compared to Attribute 1. Attribute 2 also has a higher median at 2.9 which means it has a larger spread of results.The interquartile results is not very significant, as the difference is only .0025.

If the analysis of the data leans towards the min results, then Attribute 1 is the better option. However, if the result of the data favors the higher results, then the best option is Attribute 2 as it is more spread and it has the maximum figure in the data set.

Q4: Scatter Plot

Output Screenshot



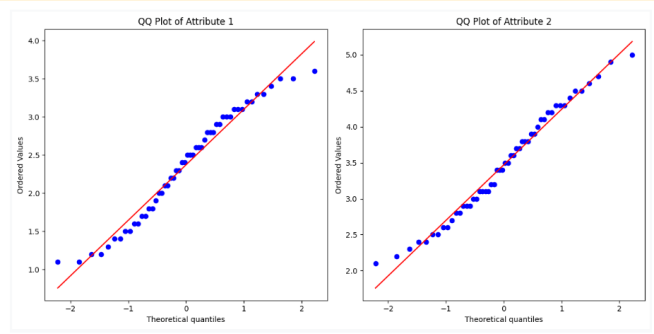
Positive Correlation:
•There is a clear upward trend in the data points, suggesting a **positive correlation** between Attribute 1 and Attribute 2.

Strength of Relationship:
•The points are fairly clustered along a trend, indicating a strong linear relationship.

The scatter plot demonstrates a positive linear relationship between **Attribute 1** and **Attribute 2**. This indicates that as one variable increases, the other tends to increase as well. The strength of the correlation is moderate to strong based on the visual clustering of the points.

Q5: QQ Plot

Output Screenshot



Attribute 1: The points deviate from the straight line, particularly at the tails, indicating that the data for Attribute 1 does not follow a normal distribution. The data shows some skewness or heavy tails.

Attribute 2: The Q-Q plot for Attribute 2 also shows deviations from the straight line, especially at the extremes, suggesting that it does not follow a normal distribution either

However, the central points are closer to the line than Attribute 1, indicating a closer fit to normality but still with deviations.

APPENDIX:

Code for Q1(Frequency):

1. Frequency table for each attribute with 4 bins. Comment on the results.

```
import pandas as pd
import numpy as np
# Loading the dataset
df = pd.read_excel("C:\\Users\\Amrit Raj\\Downloads\\Assignement1_dataset.xlsx")
# Creating 4 bins for each attribute
bins = 4
# Frequency table for Attribute1
attribute1_freq, attribute1_bins = np.histogram(df['Attribute1'], bins=bins)
attribute1_freq_table = pd.DataFrame({
   'Bin Range': pd.cut(df['Attribute1'], bins=bins).unique(),
   'Frequency': attribute1_freq
})
# Frequency table for Attribute2
attribute2_freq, attribute2_bins = np.histogram(df['Attribute2'], bins=bins)
attribute2_freq_table = pd.DataFrame({
   'Bin Range': pd.cut(df['Attribute2'], bins=bins).unique(),
   'Frequency': attribute2_freq
})
print("Frequency Table for Attribute 1")
print(attribute1_freq_table)
print("\nFrequency Table for Attribute 2")
print(attribute2_freq_table)
```

Code for Q2(Histogram):

2. Create a historgram using the same bins for each attribute. Comment on the results.

```
import matplotlib.pyplot as plt

bins = 4

# Histogram for both attributes using above bins

fig, axs = plt.subplots(1, 2, figsize=(8,5))

# Histogram for Attribute1

axs[0].hist(data['Attribute1'],color='salmon', bins=bins, edgecolor='black')

axs[0].set_title('Histogram of Attribute1')

axs[0].set_xlabel('Attribute1')

axs[0].set_ylabel('Frequency')

# Histogram for Attribute2

axs[1].hist(data['Attribute2'],color='darkgreen', bins=bins, edgecolor='black')

axs[1].set_title('Histogram of Attribute2')

axs[1].set_xlabel('Attribute2')

axs[1].set_ylabel('Frequency')

plt.tight_layout()

plt.show()
```

Code for Q3 (Box Plot):

3. Create a boxplot for both attributes. Comment on the results.

```
plt.figure(figsize=(6, 4))

plt.boxplot([df['Attribute1'], df['Attribute2']], labels=['Attribute 1', 'Attribute 2'],
```

```
    boxprops=boxprops, medianprops=medianprops, whiskerprops=whiskerprops,
capprops=capprops)

plt.title('Attribute 1 and Attribute 2')

plt.ylabel('Values')

plt.show()
```

Code for Q4 (Scatter Plot):

4. Create a scatterplot. Comment on the results.

```
#Importing the required Libraries

import pandas as pd

import matplotlib.pyplot as plt

#Importing the dataset

df = pd.read_excel("Assignement 1 dataaset.xlsx")


# Plotting a scatter plot between 'Attribute 1' and 'Attribute 2'

plt.figure(figsize=(8,6))

plt.scatter(df['Attribute1'], df['Attribute2'], color='skyblue')

plt.title('Attribute 1 V/S Attribute 2')

plt.xlabel('Attribute 1')

plt.ylabel('Attribute 2')

plt.grid(True)

plt.show()
```

Code for Q5:

5. Create a QQ plot for each attribute. Are either attributes normally distrubted? Why or why not?

```
import matplotlib.pyplot as plt

import scipy.stats as stats

plt.figure(figsize=(12, 6))


# QQ plot for Attribute 1

plt.subplot(1, 2, 1)

stats.probplot(df['Attribute1'], dist="norm", plot=plt)

plt.title('QQ Plot of Attribute 1')


# QQ plot for Attribute 2

plt.subplot(1, 2, 2)

stats.probplot(df['Attribute2'], dist="norm", plot=plt)

plt.title('QQ Plot of Attribute 2')

plt.tight_layout()

plt.show()
```