Experiment No.4

Title: Applying and interpreting different plots

Batch:A2 **Roll No.:16010421063** **Experiment No.: 4**

Aim: Applying and interpreting different plots

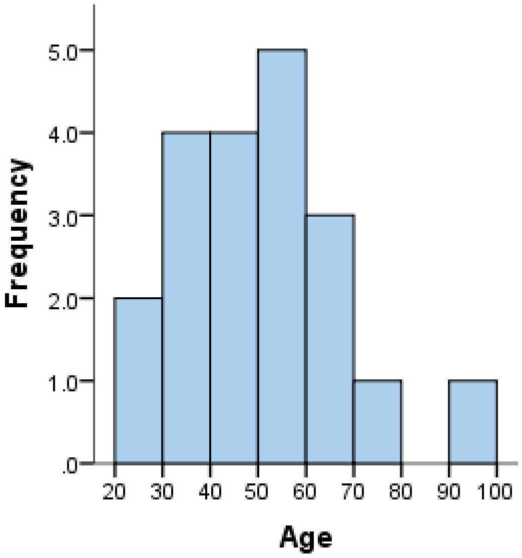
Resources needed: Any programming language/ Rapid Miner, any data source (RDBMS/Excel/CSV)

Theory:

For data preprocessing to be successful, it is essential to have an overall picture of your data. Basic statistical descriptions alone cannot be used to identify properties of the data and highlight which data values should be treated as noise or outliers. The plots such as Box Plot, Q-Q Plot, Histogram and Scatterplots provide various information to the data analyst. Data visualization is very much needed because a visual summary of information makes it easier to identify patterns and trends than looking through thousands of rows. Before applying plots suitability of the attribute should be checked.

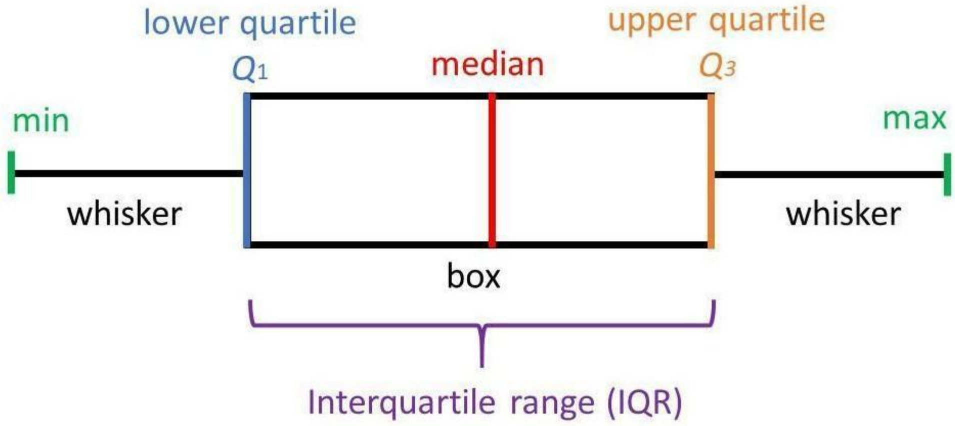
Histogram

Histogram gives accurate representation of the distribution of numeric data. A histogram is a chart that shows frequencies for intervals of values of a continuous variable. It summarize a Univariate Data set. In histogram of a continuous frequency table, x-axis marks class intervals on a suitable scale and y-axis marks frequency of each class interval. The interval of value is known as bin and they all have the same widths. The upper and lower class limits of the new exclusive type classes are known as class boundaries. Histograms also give us much more complete information about our data.



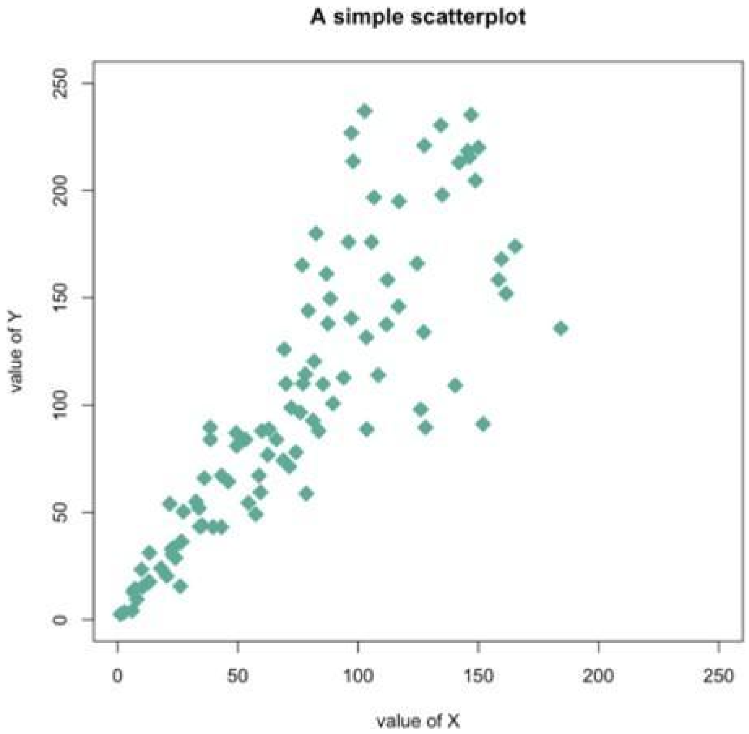
Box plot

Boxplot also known as box-and-whisker plot is a way to show the distribution of values based on the five-number summary: minimum, first quartile, median, third quartile, and maximum. The minimum and the maximum are just the min and max values from the data set. The median is the value that separates the higher half of a data from the lower half. The first quartile is the median of the data values to the left of the median in our ordered values. The third quartile is the median of the data values to the right of the median in our ordered values. Boxplot can also show outliers and IQR(Inter Quartile range) .



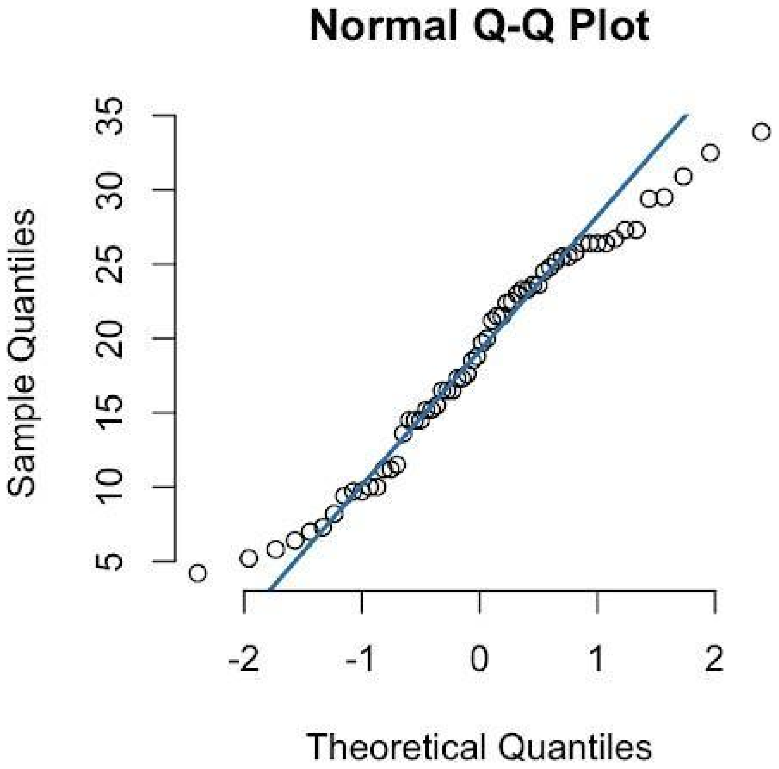
Scatterplots

A scatter plot is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data. A scatter plot can be used either when one continuous variable that is under the control of the experimenter and the other depends on it or when both continuous variables are independent. A scatter plot can suggest various kinds of correlations between variables with a certain confidence interval.



Quantile-Quantile Plot

A Q–Q (quantile-quantile) plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Many distributional aspects can be obtained from a q-q plot like, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot. It helps to assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential.



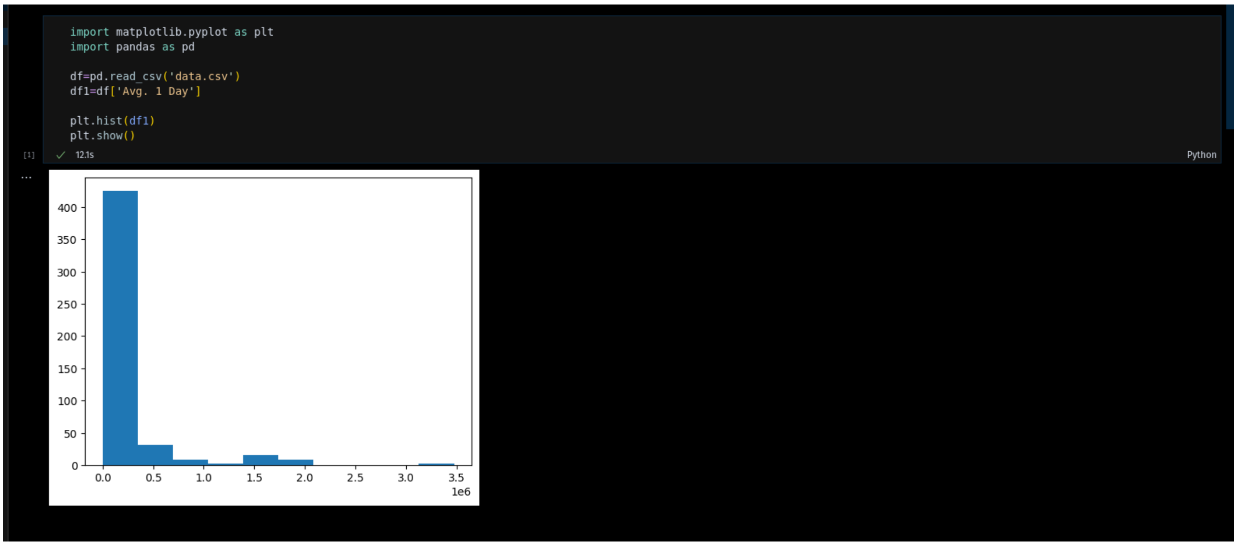
Procedure / Approach /Algorithm / Activity Diagram:

1. Identify the attributes where it will be sensible to apply the below given plots.

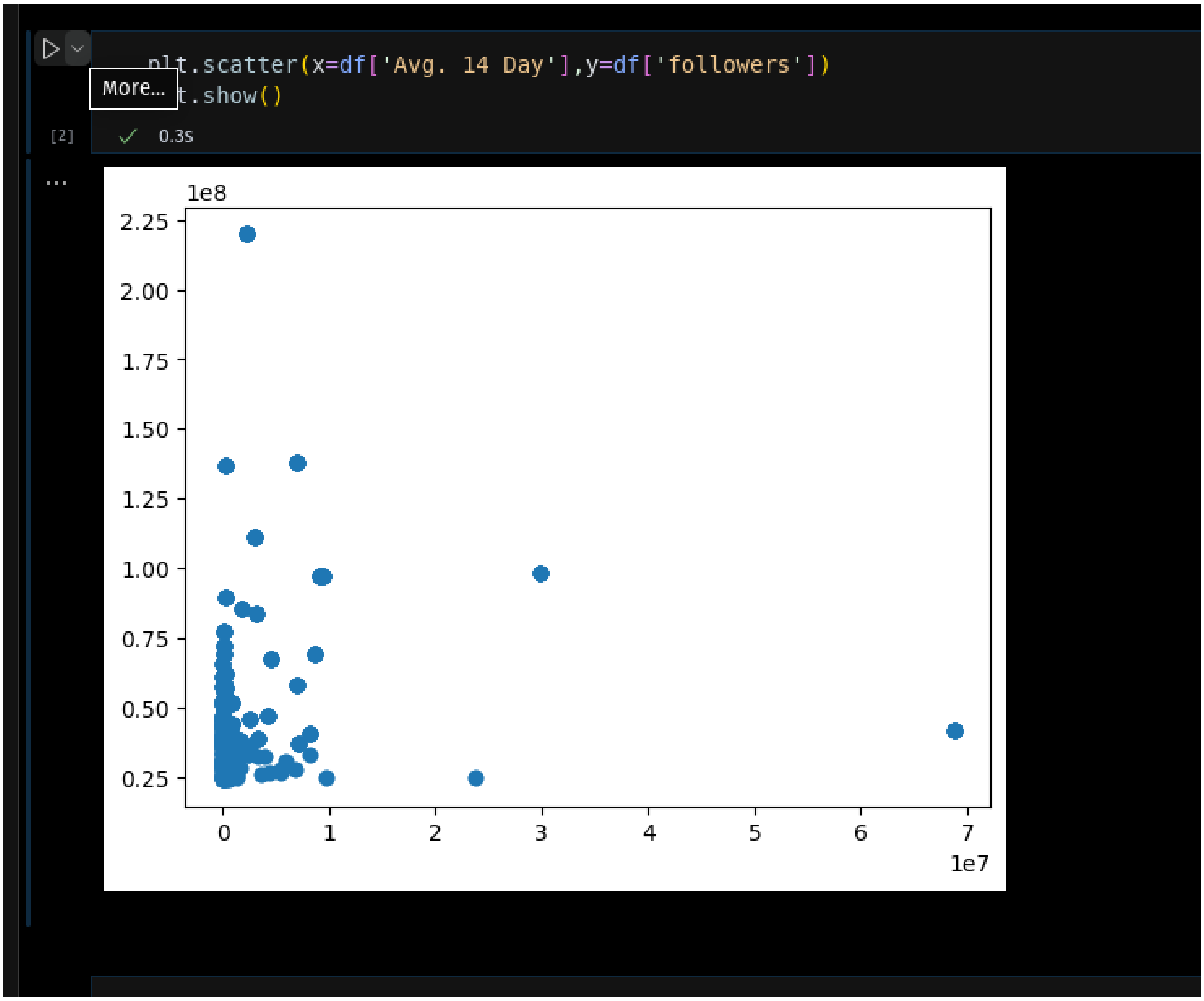
    a. Box Plot
    b. Q Q Plot
    c. Histogram
    d. Scatter Plot

    Apply the above mentioned plots on the identified attributes. Discuss the inferences from these plots in detail.

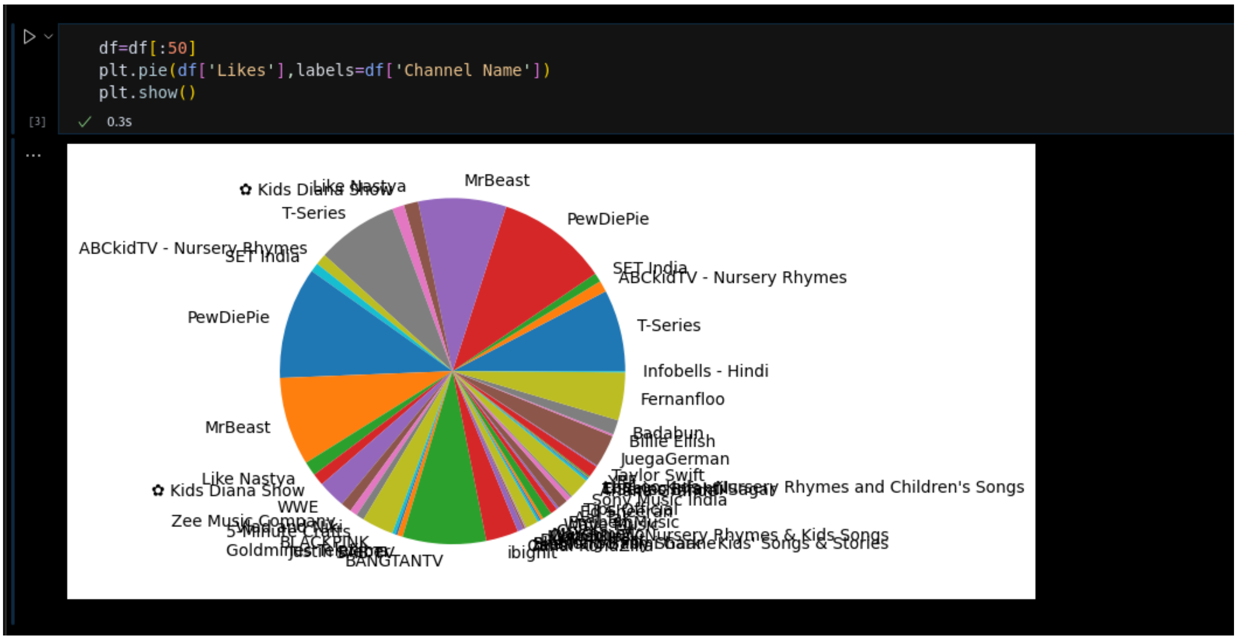
Results: (Program printout with output / Document printout as per the format)



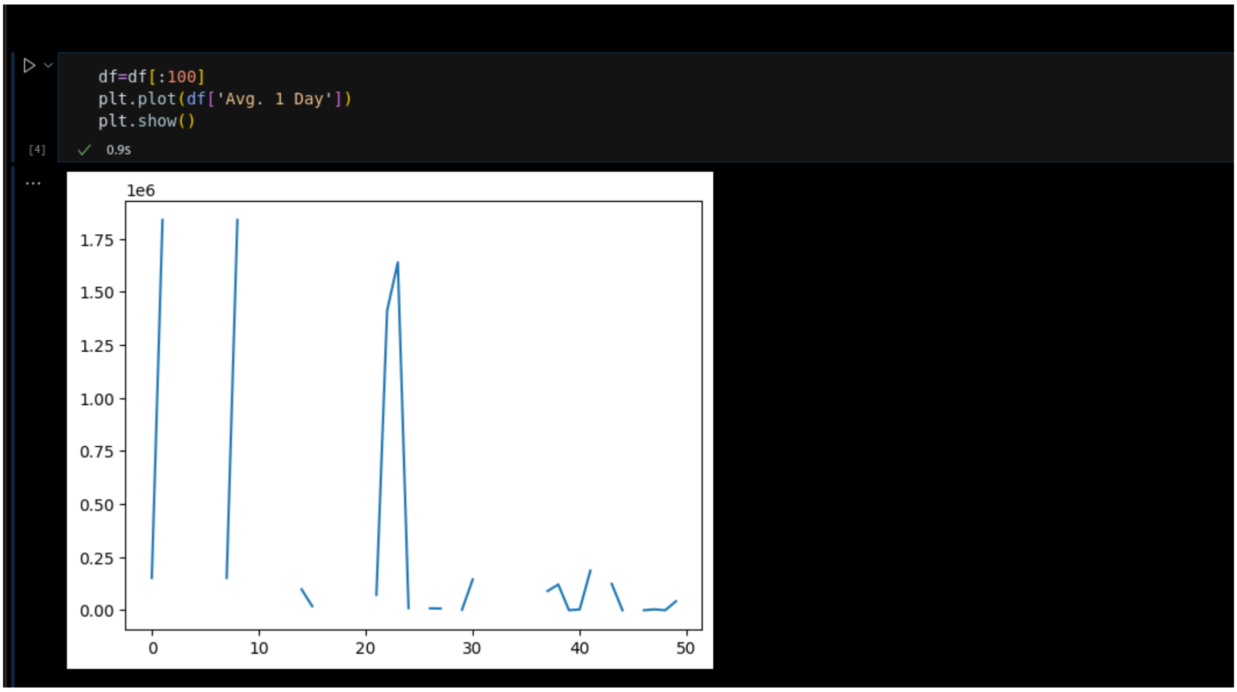
```
import matplotlib.pyplot as plt
import pandas as pd

df=pd.read_csv('data.csv')
df1=df['Avg. 1 Day']

plt.hist(df1)
plt.show()
```



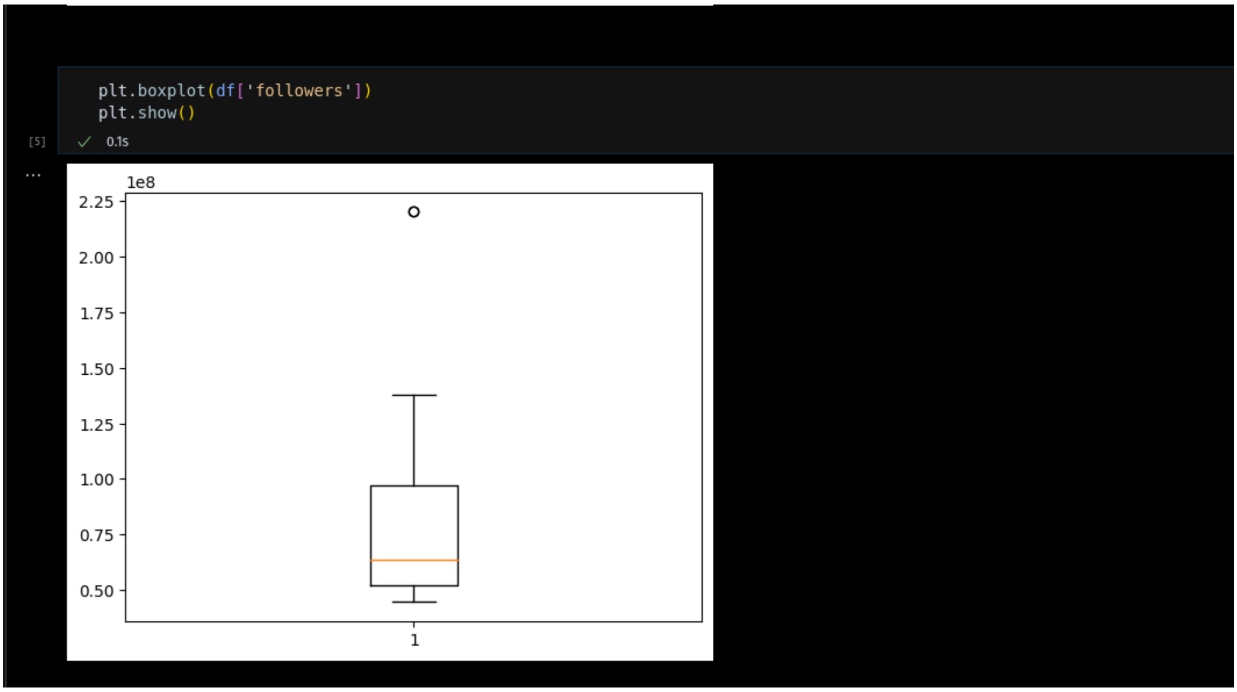
```
plt.scatter(x=df['Avg. 14 Day'],y=df['followers'])
t.show()
```



```
df=df[:50]
plt.pie(df['Likes'],labels=df['Channel Name'])
plt.show()
```



```
df=df[:100]
plt.plot(df['Avg. 1 Day'])
plt.show()
```



```
plt.boxplot(df['followers'])
plt.show()
```

Questions:

1. Why is it important to measure the dispersion in the dataset?

Ans- While measures of central tendency are used to estimate "normal" values of a dataset, measures of dispersion are important for describing the spread of the data, or its variation around a central value. Two distinct samples may have the same mean or median, but completely different levels of variability, or vice versa.

2. Discuss the other purposes/advantages of the plots used in this experiment.

Ans- Advantages of Box Plot

- Graphically display a variable's location and spread at a glance
- Provide some indication of the data symmetry and skewness
- Unlike many other methods of data display, box plots show outliers

Advantages of Histogram

- The main advantage of a histogram are its simplicity and versatility.
- It can be used in many different situation to offer an insightful look at frequency distribution

Advantages of Scatter Plot

- Show a Relationship and a trend in the data relationship
- Show all data points including minimum and maximum and outliers
- Can highlight correlations. Retains the exact data values and

Advantages of Pie Chart

- It represents data visually as a fractional part of a whole, which can be an effective communication tool for the even uninformed audience.
- The need for readers to examine or measure underlying numbers themselves can be removed by using this chart.
- To emphasize points you want to make, you can manipulate pieces of data in the pie chart.

Outcomes:

CO4:Comprehend various data visualization techniques

Conclusion: (Conclusion to be based on the objectives and outcomes achieved)

Successfully understood and implemented various graphs/plots

Grade: AA / AB / BB / BC / CC / CD /DD

Signature of faculty in-charge with date

References:

Books/ Journals/ Websites:

1. Han, Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann 3nd Edition