A Project

Report On

Prediction Using Supervised And Unsupervised Learning

By

Tejaswini Vaidya

B.E

Batch: 2020 – 5129

Center: Thane

Under the Guidance of,

Mr. Mathivanan Balakrishnan

Technical Trainer

EduBridge

(School of coding)

Introduction:

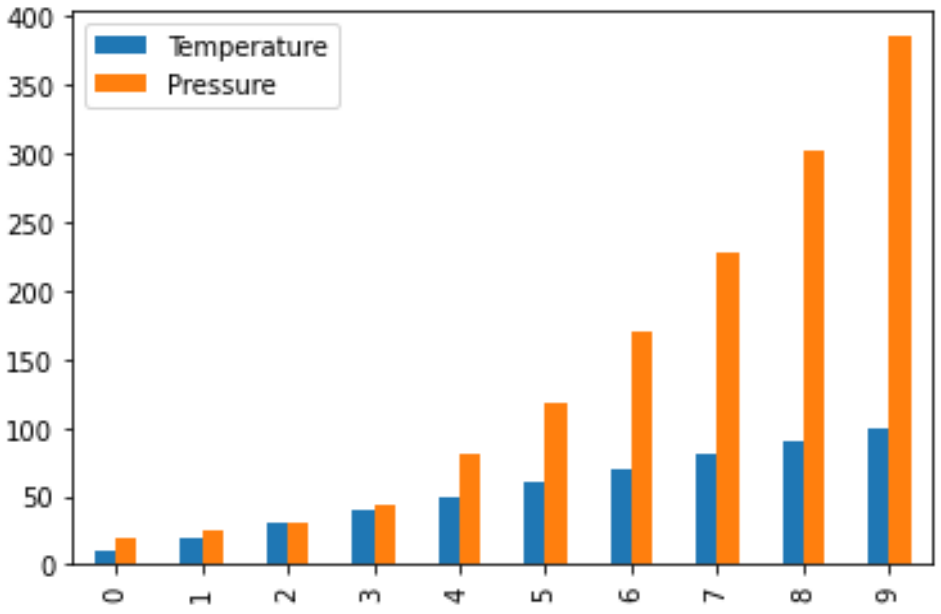
In this project, we are taking different datasets and performing different machine learning algorithms on that dataset. The Machine Learning algorithm we have used on our project is as follows:

1. Linear Regression.
2. Logistic Regression.
3. K-Means Clustering.

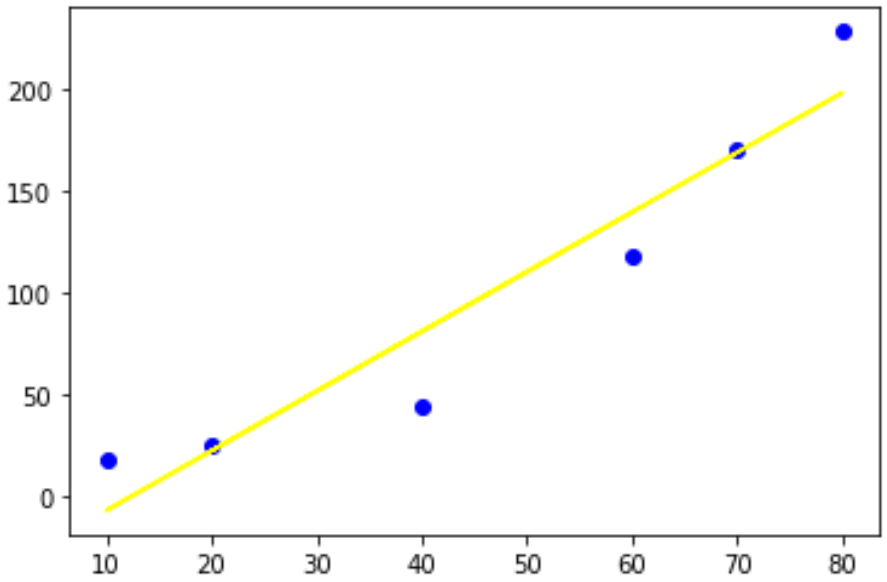
1. **Linear Regression:** Linear Regression is one of the easiest and popular machine learning algorithms. It is a statistical method used for prediction analysis for continuous or numeric variables. Here, we have taken a dataset of pressure which is of two columns and ten rows. Below are the screenshots:

I. Screenshot using Python language:

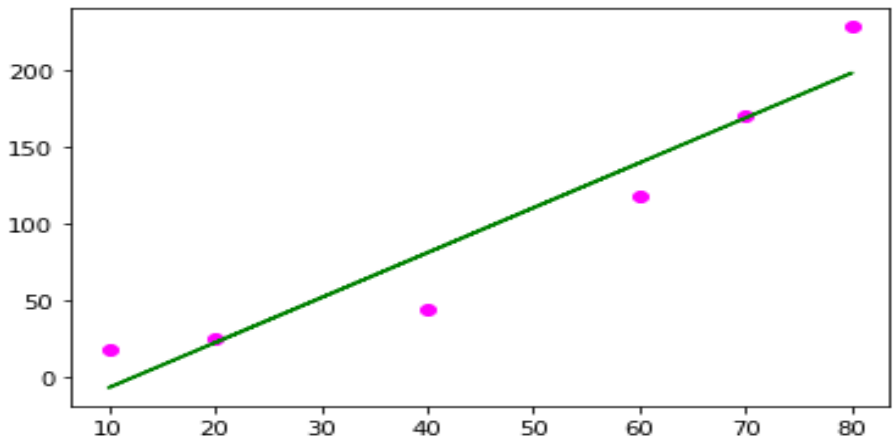
1.Bar graph of the dataset:



2.Visualizing the training data:

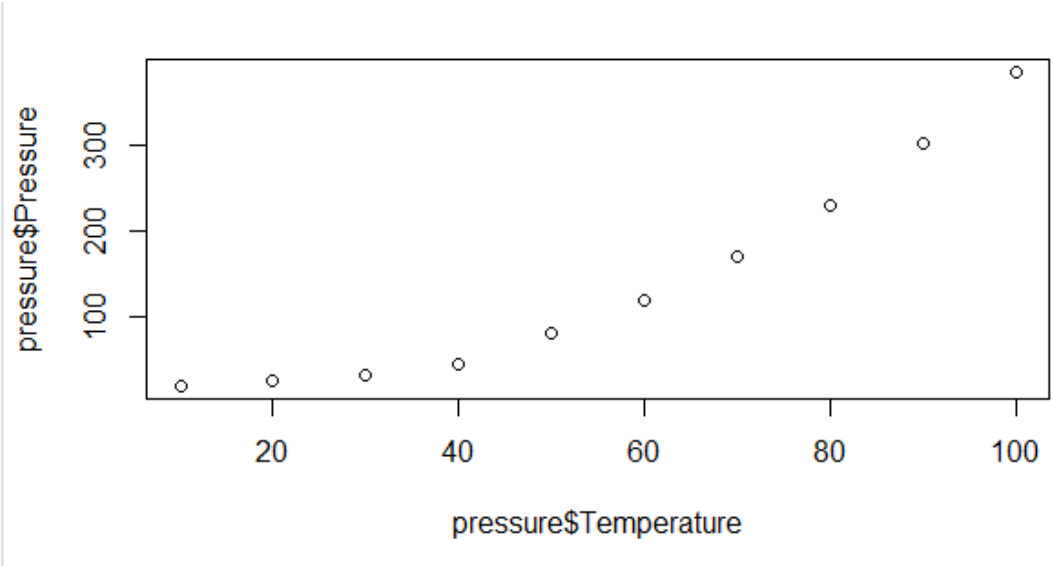


3.Visualising the test data:

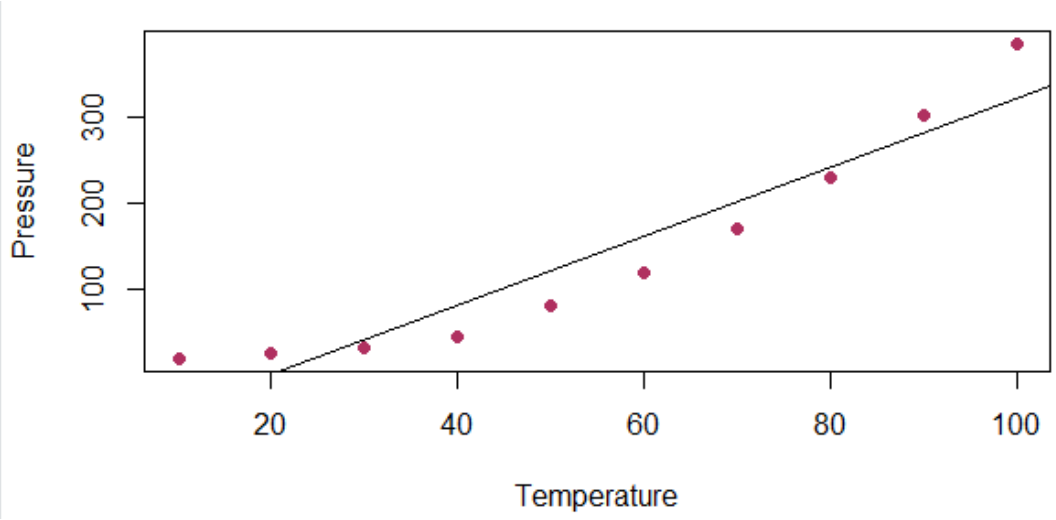


II. Screenshot using R language:

1.Plotting Temperature and Pressure:



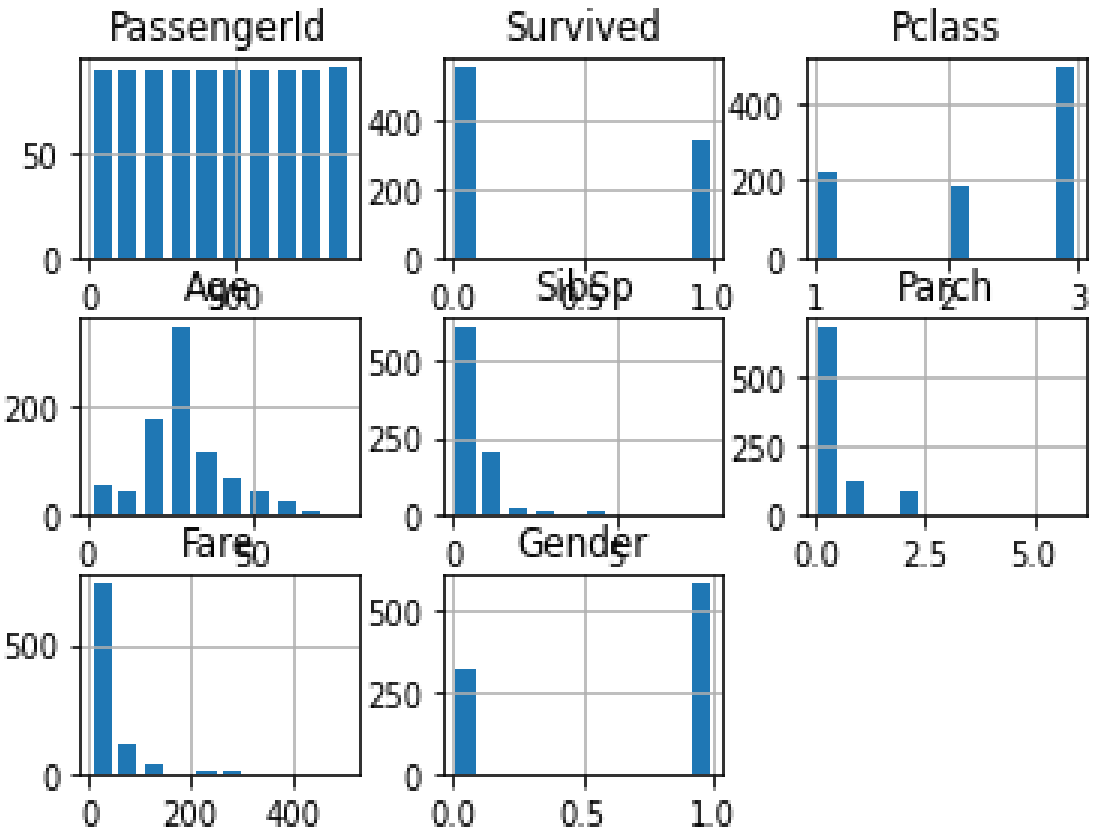
2. Output in R:



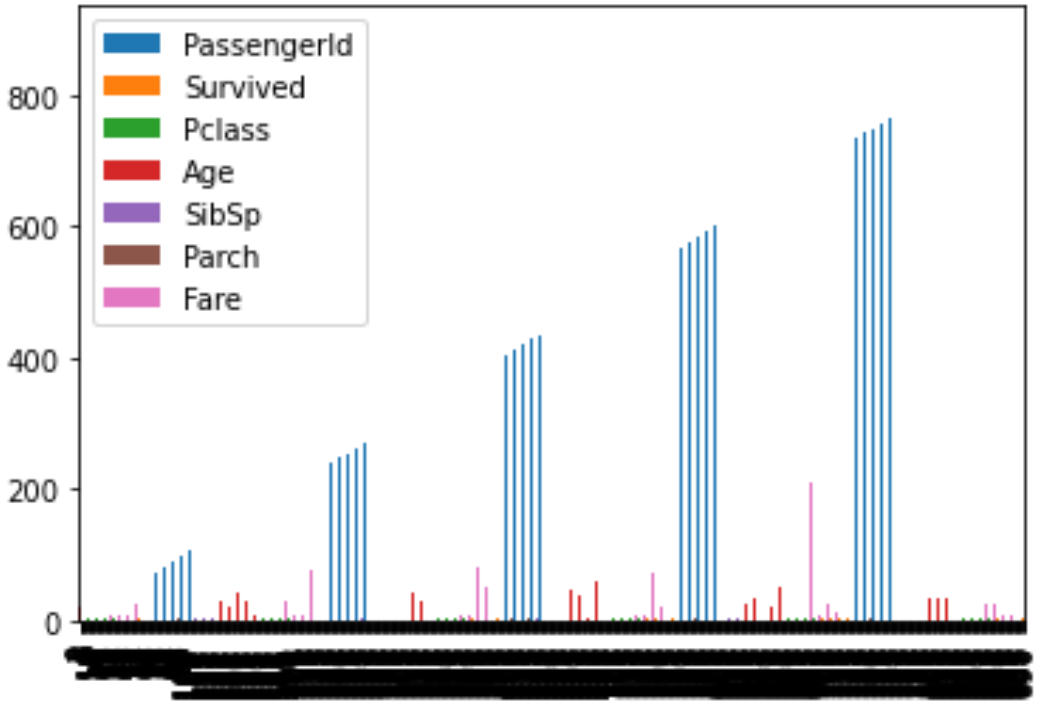
2. **Logistic Regression**: It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1. While performing logistic regression in python we have used the dataset of titanic. Below are the screenshots.

I. **Using python Language:** The dataset used in this algorithm is the titanic dataset. The prediction of Survival of passengers is done.

1.Histogram of the dataset:

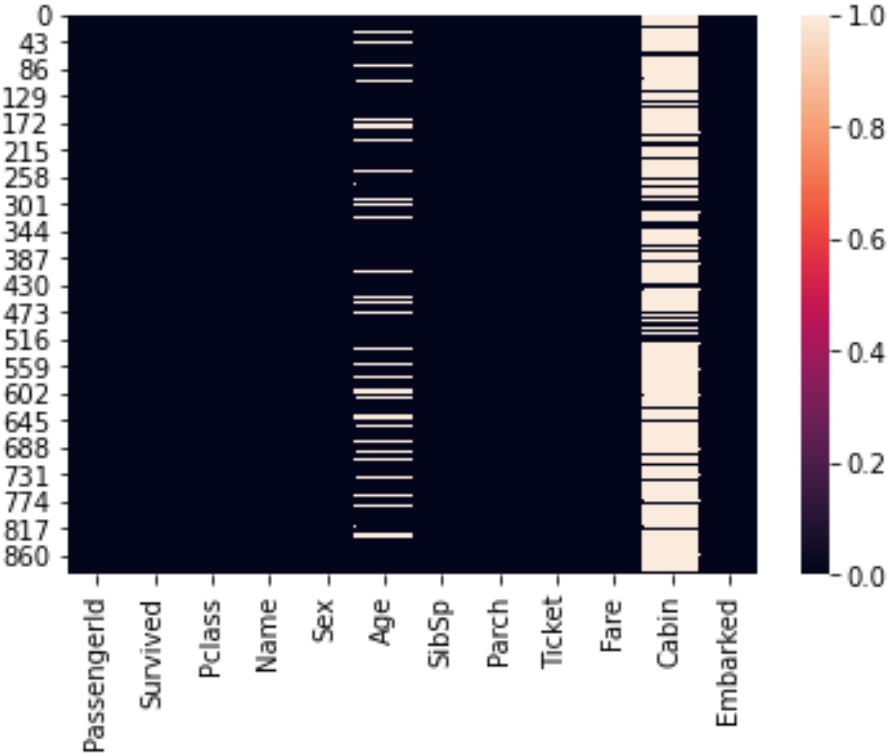


2.Bar plot of the dataset:



2. Heatmap of the dataset:

This graph shows the null values present in the dataset.



3. Confusion Matrix:

From this we come to know the correct prediction and wrong prediction.

```
[[151  24]
 [ 38  82]]
```

4. Classification Report:

From this we will come to know brief about the result of the dataset.

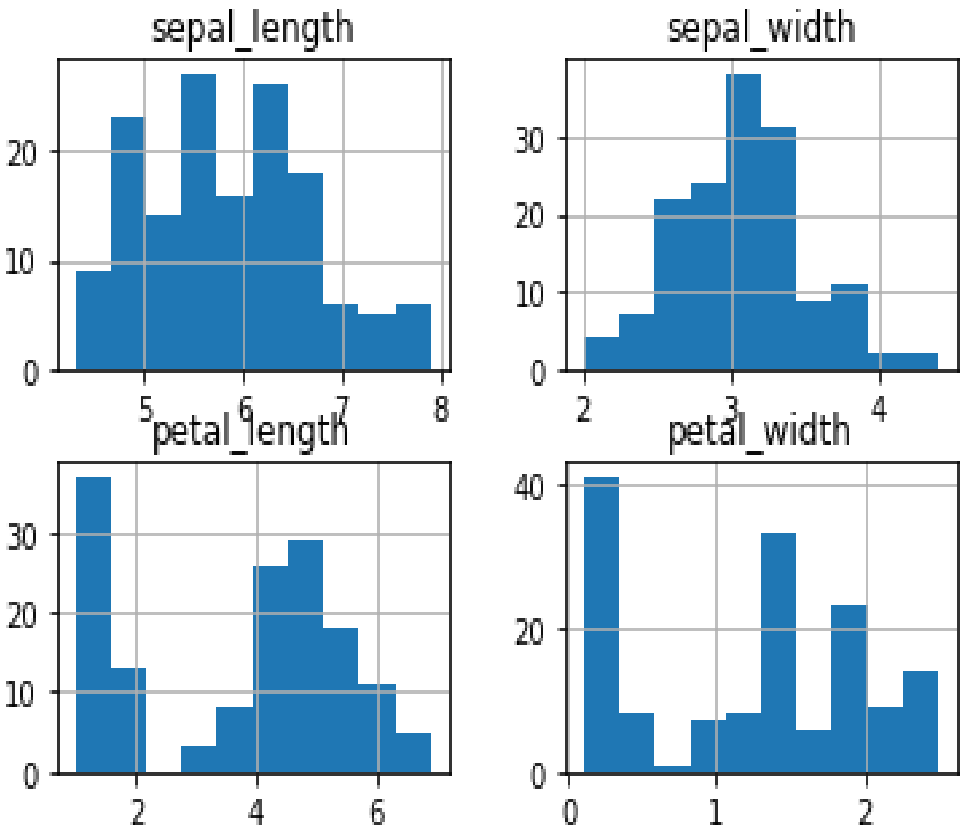
	precision	recall	f1-score	support
0	0.80	0.86	0.83	175
1	0.77	0.68	0.73	120
accuracy			0.79	295
macro avg	0.79	0.77	0.78	295
weighted avg	0.79	0.79	0.79	295

3. k-means clustering: k-means clustering is an unsupervised learning algorithm which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

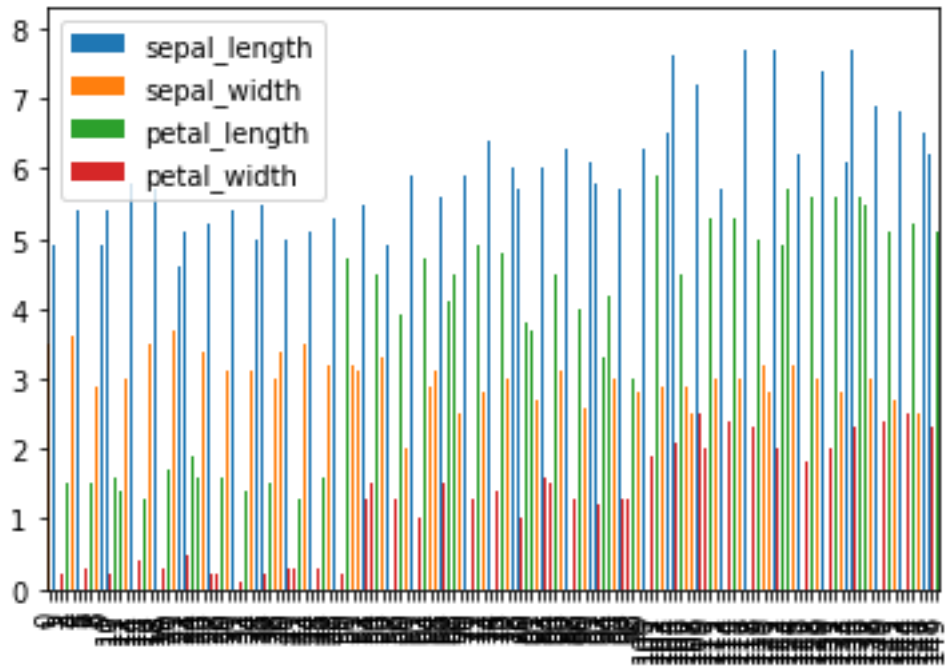
I. Using Python: The dataset used for analysis is the iris dataset. It contains 150 rows and 5 columns.

Screenshot are as follows:

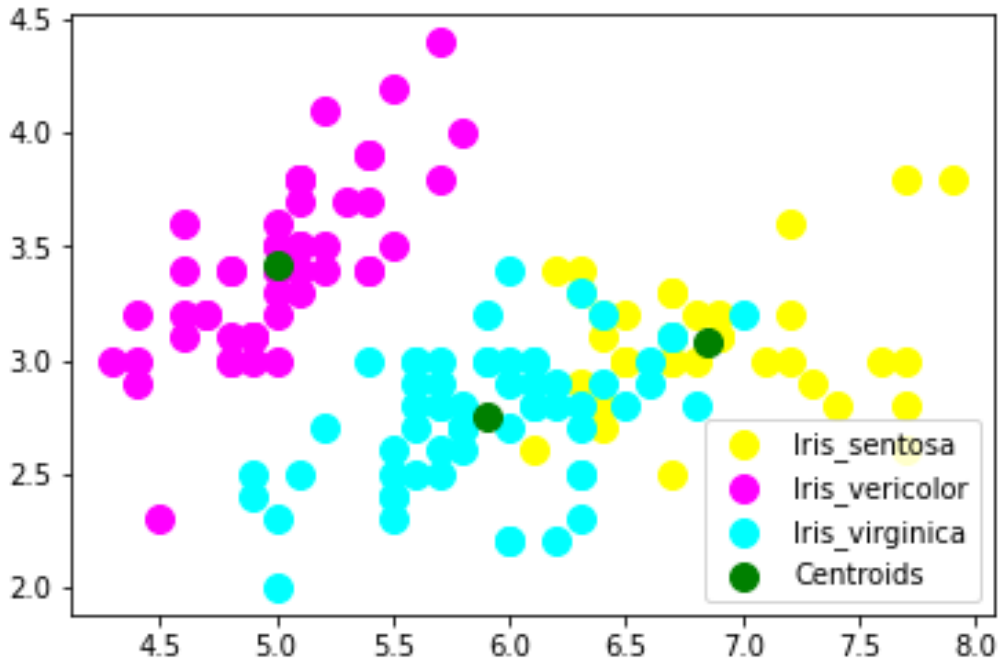
1. **Histogram of the dataset:**



2. **Bar chart of the dataset:**



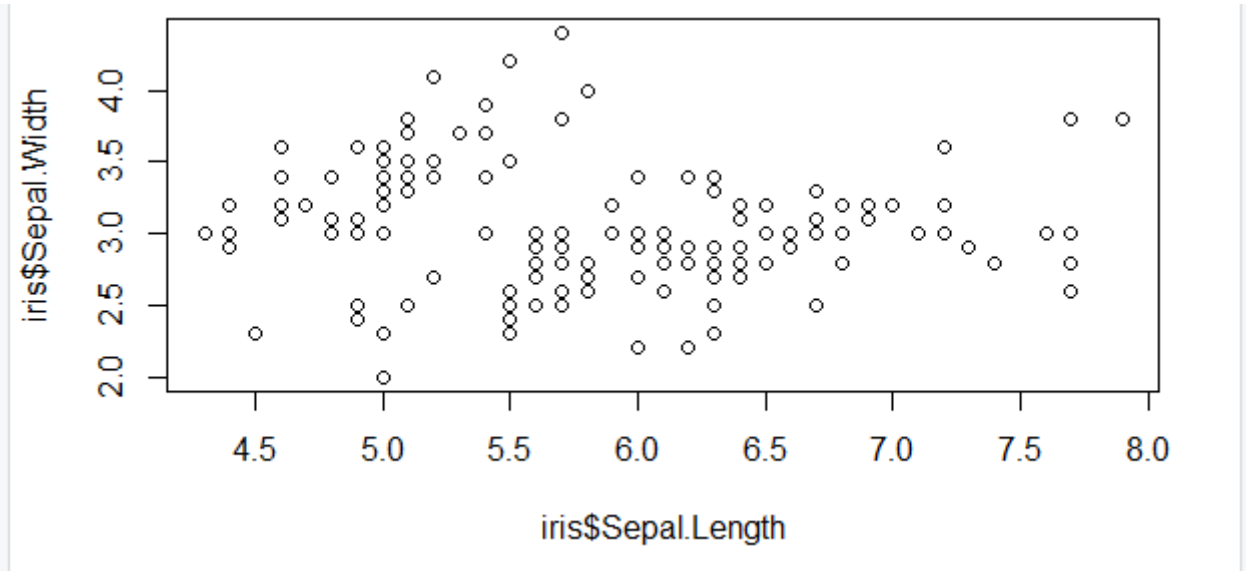
3. Output of KMeans:



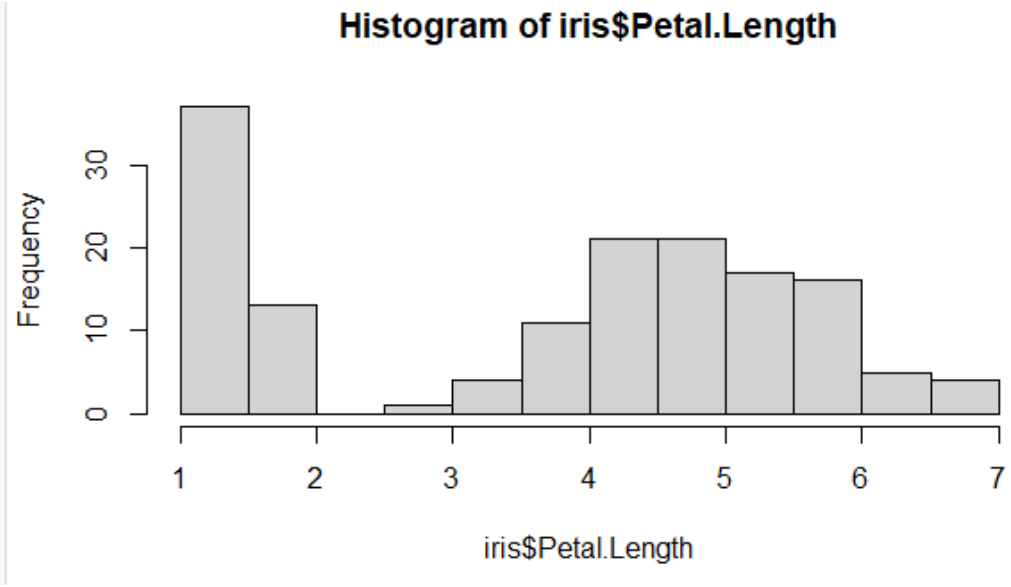
I. Analysis in R:

We have used the same dataset named iris.

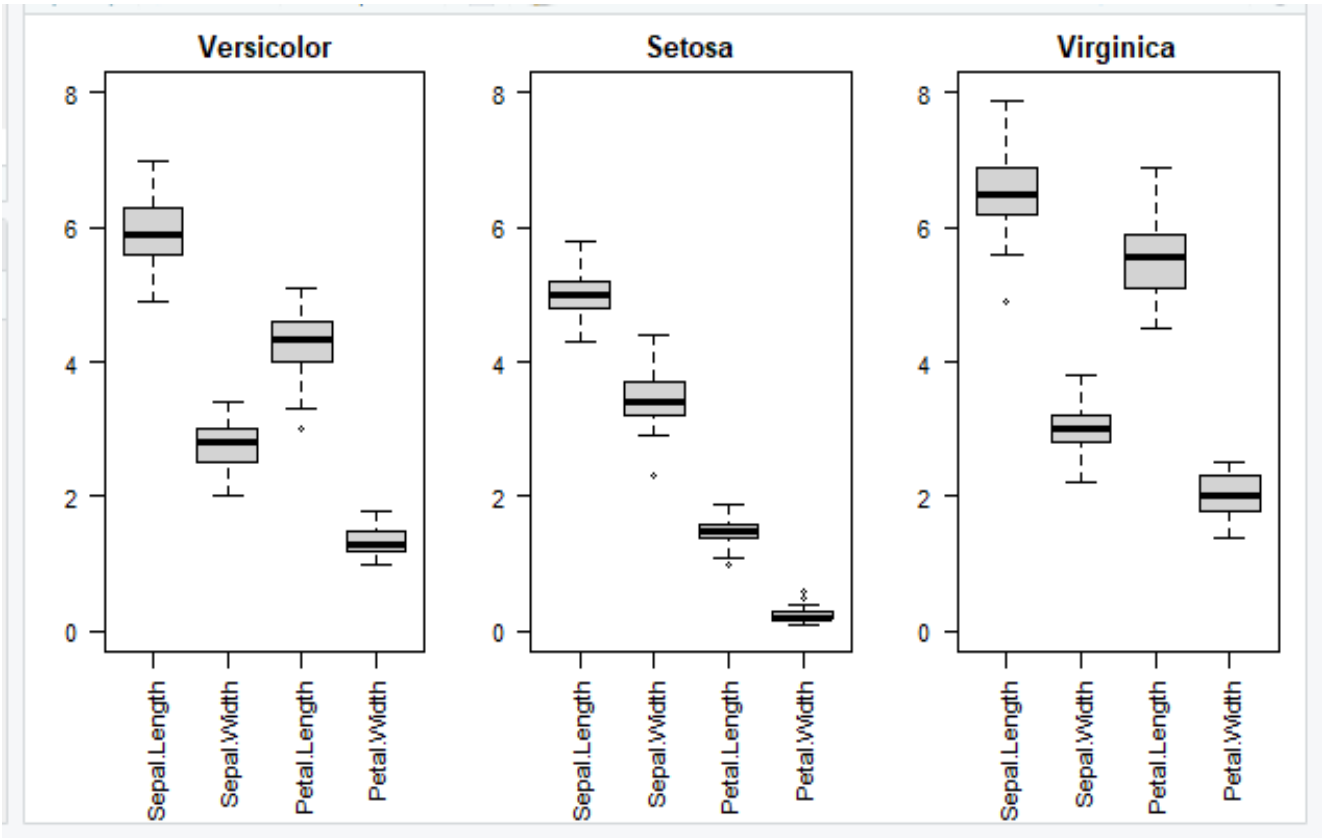
1. Plotting Sepal.length and Sepal.Width:



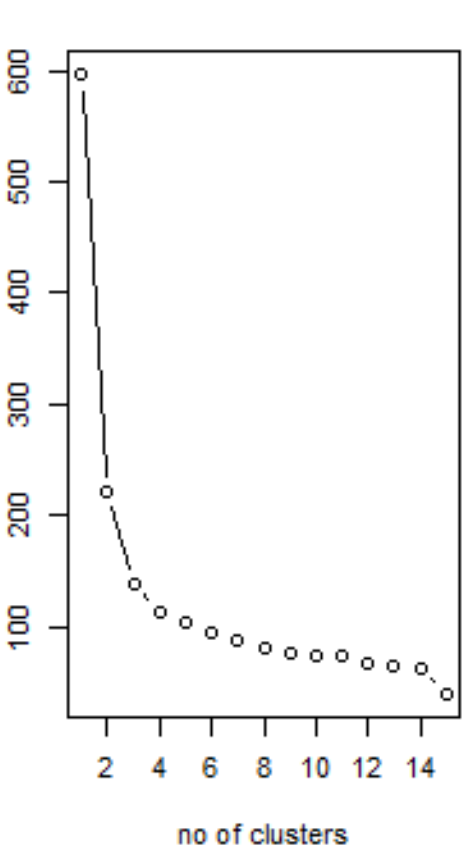
2. **Histogram of Petal.length:**

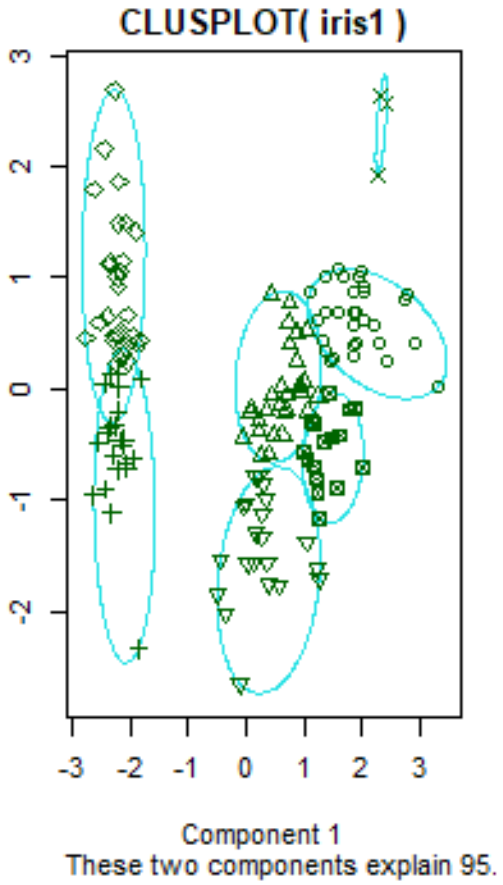


3. **Boxplot graph:**



4. Output of KMeans:





Software Requirements:

- **Software Used:** Spyder 4.1.5, RStudio Version 1.4.1106.
- **Back end**: MongoDB 4.4.4.
- **Operating System:** Windows 10, 64-bit operating system.

Thank you