Statistical Studies of IRIS Data

presented by:

Ola Adel

2nd Undergraduate — Faculty of Computer and Information Sciences - Ain Shams University

Statistical Analysis and Applications (SCC234)

Dr. Maryam Al-Berry

Table of Contents

Introduction

The Iris flower data set or Fisher's Iris data set is a multivariate data set introduced by the British statistician and biologist Ronald Fisher in his 1936 paper The use of multiple measurements in taxonomic problems as an example of linear discriminant analysis.[1] It is sometimes called Anderson's Iris data set because Edgar Anderson collected the data to quantify the morphologic variation of Iris flowers of three related species.[2] Two of the three species were collected in the Gasp é Peninsula "all from the same pasture, and picked on the same day and measured at the same time by the same person with the same apparatus".[3]

The data set consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters. Based on the combination of these four features.

The main research question, Is the mean of virginica iris Sepal Lengths is greater than the mean of versicolor iris Sepal Lengths?

I think it′ s very important question, we can develop a linear discriminant model to distinguish the species from each other, and it′ s will make the classification most precise if we want to classify undefined flower by its sepal length, botany scientist also need the answer of this question!

First, we will determine appropriate null and alternative hypotheses, then we will check that the data distribution and finally we will make T test.

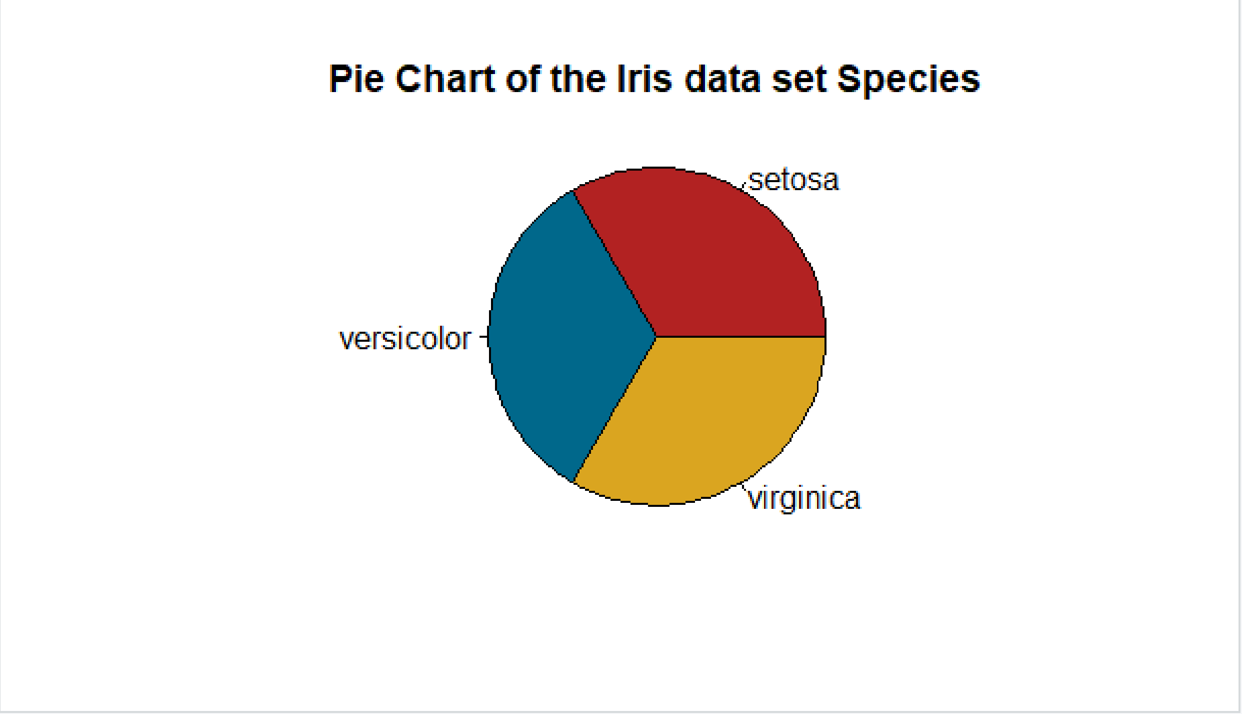
With the inferential tests, we will get the answers of most questions!

Descriptive Statistics

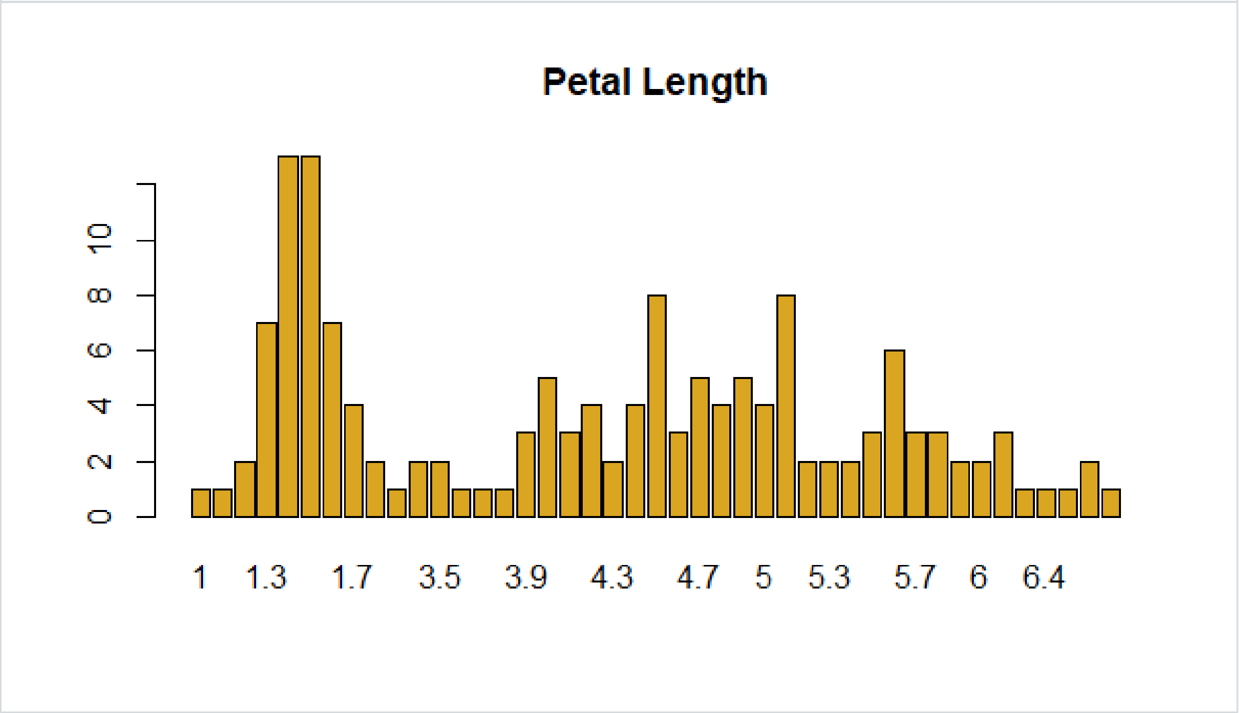
Summary of Data set

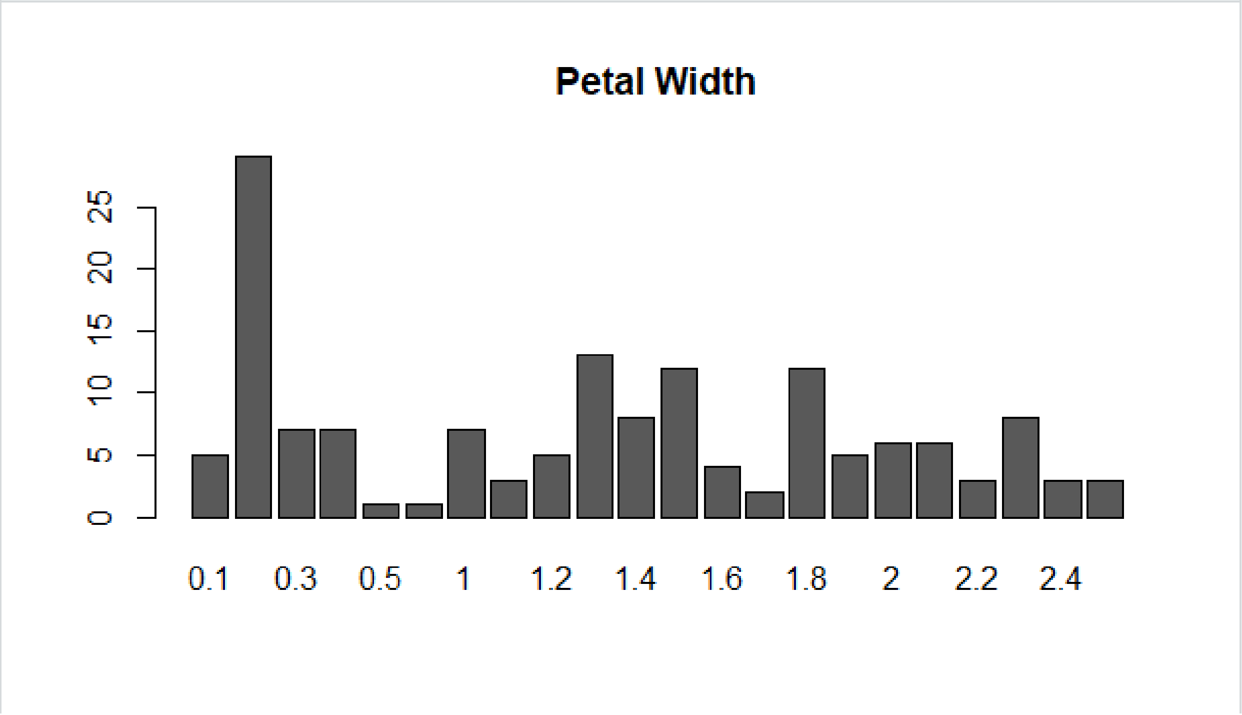
	Sepal Length	Sepal Width	Petal Length	Petal Width
Min	4.300	2.000	1.000	0.100
1st Qu	5.100	2.800	1.600	0.300
Median	5.800	3.000	4.350	1.300
Mean	5.843	3.057	3.758	1.199
3rd Qu	6.400	3.300	5.100	1.800
Max	7.900	4.400	6.900	2.500

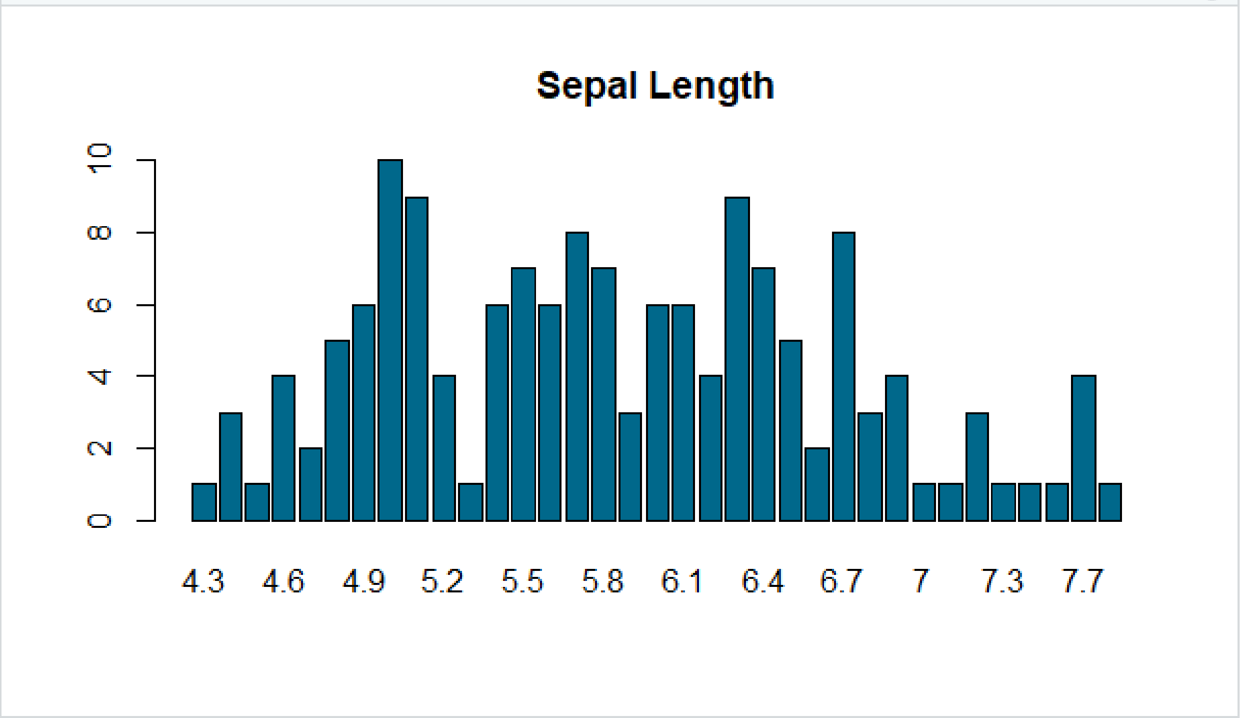
Pie Chart

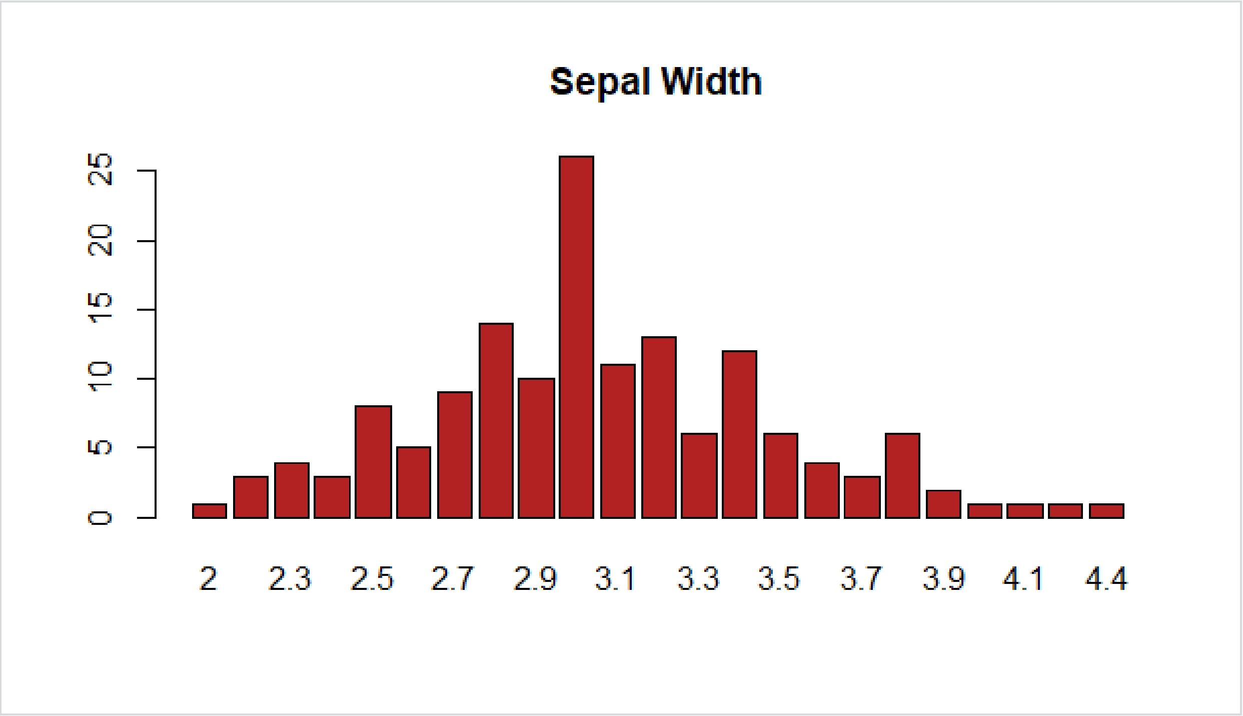

Pie Chart of the Iris data set Species
setosa
versicolor
virginica

Bar Plots

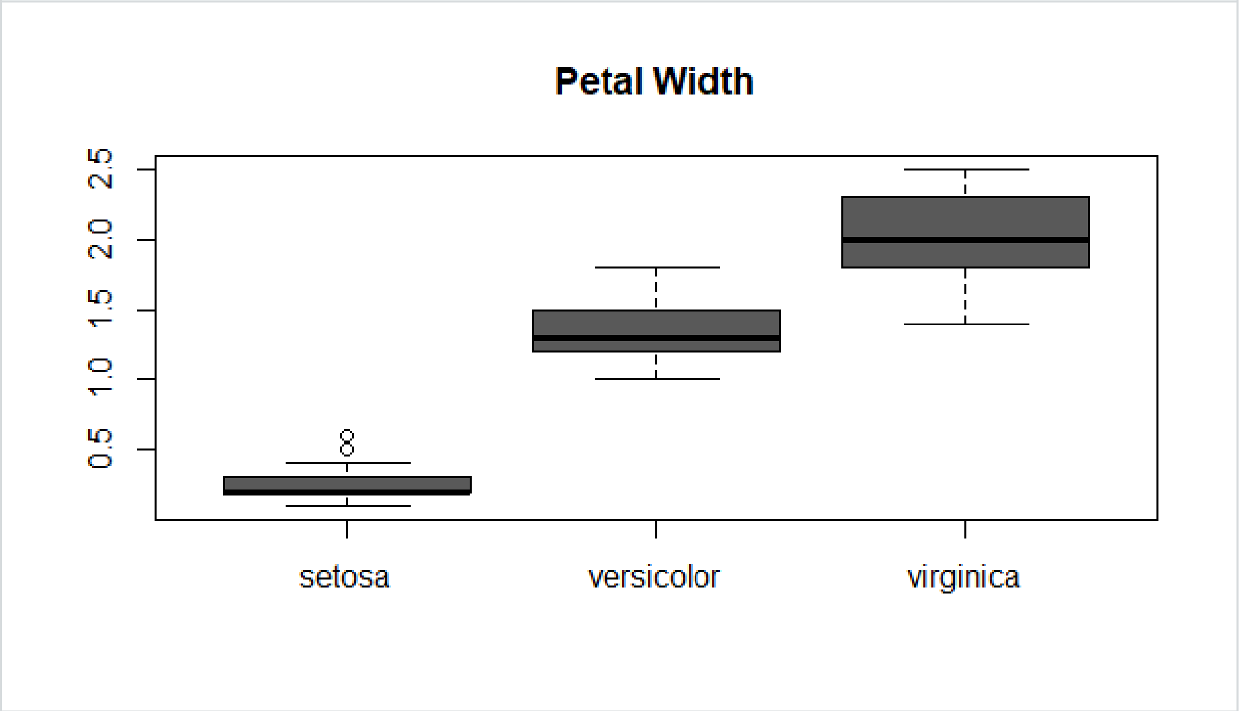

Petal Length
0
2
4
6
8
10
1
1.3
1.7
3.5
3.9
4.3
4.7
5
5.3
5.7
6
6.4

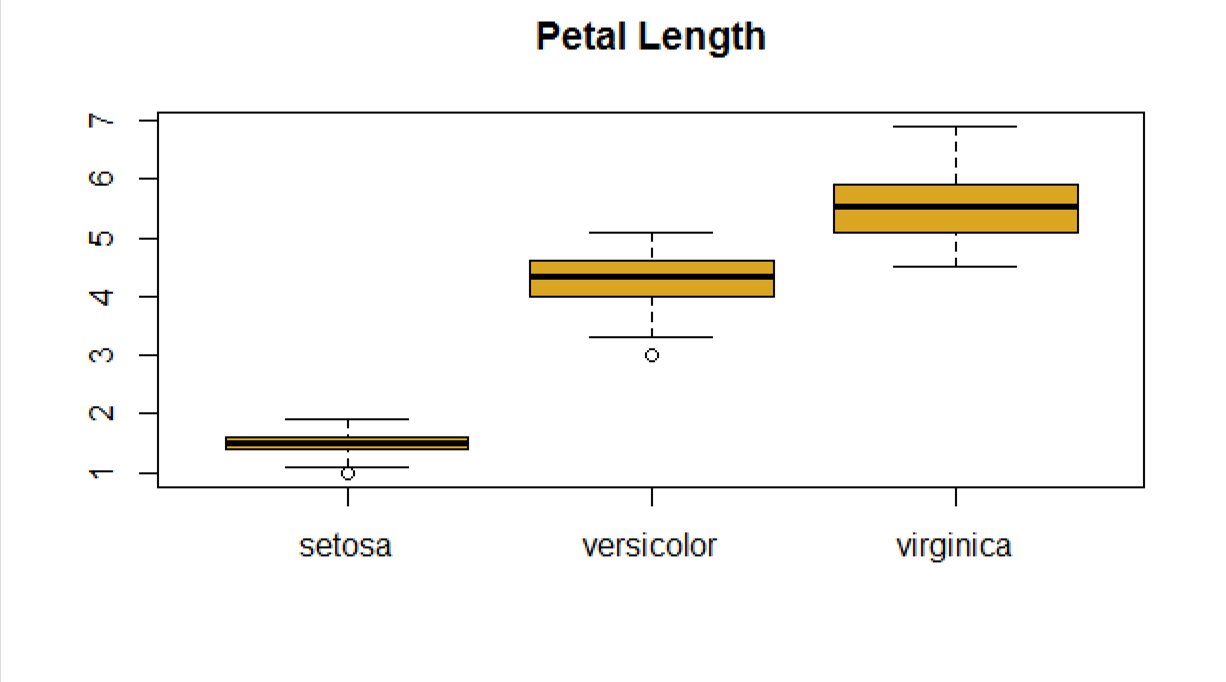

Petal Width
0
5
10
15
20
25
0.1
0.3
0.5
1
1.2
1.4
1.6
1.8
2
2.2
2.4

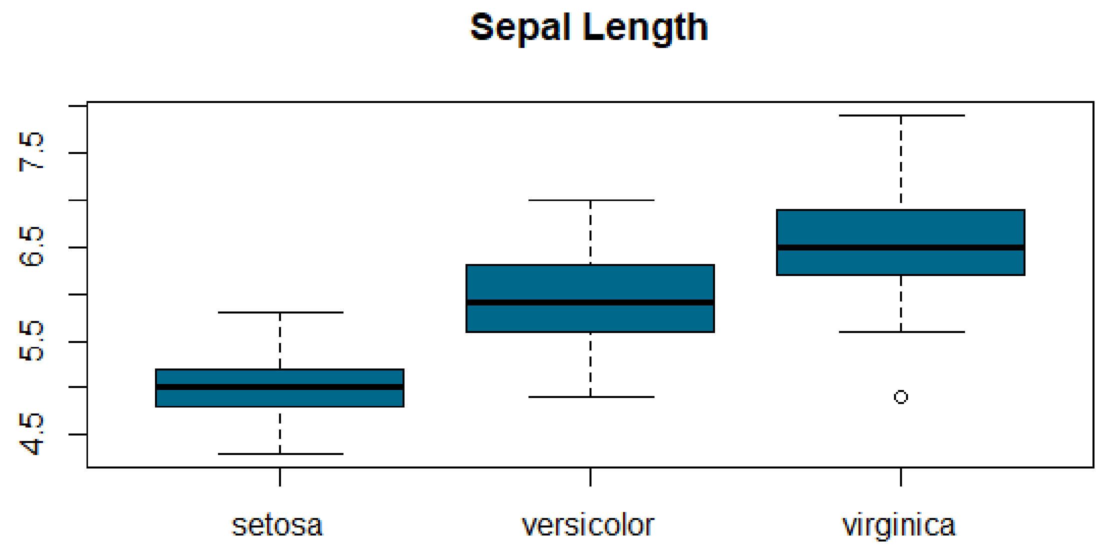

Sepal Length
0
2
4
6
8
10
4.3
4.6
4.9
5.2
5.5
5.8
6.1
6.4
6.7
7
7.3
7.7

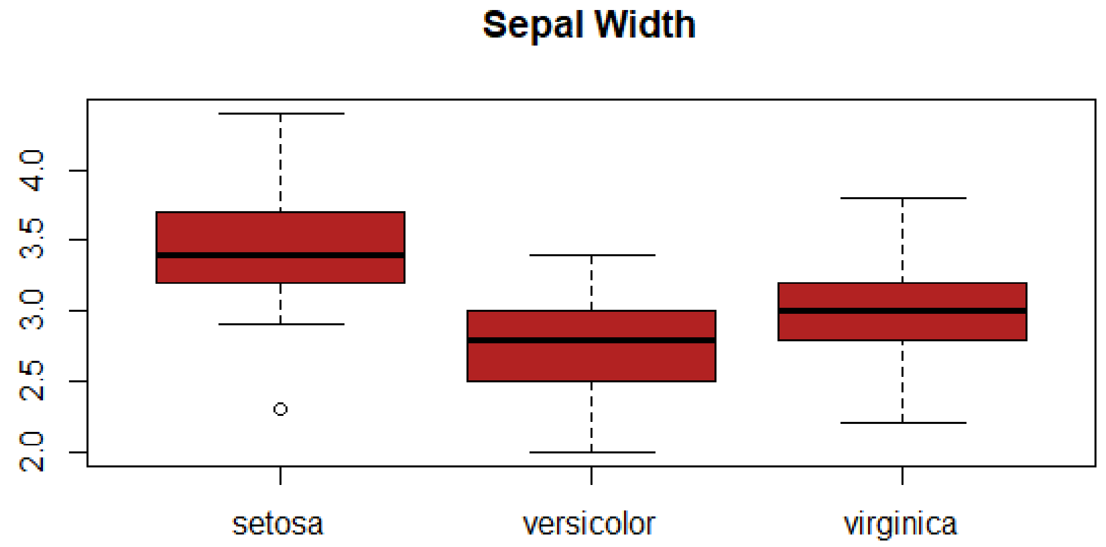

Sepal Width
0
5
10
15
20
25
2
2.3
2.5
2.7
2.9
3.1
3.3
3.5
3.7
3.9
4.1
4.4

Box plots

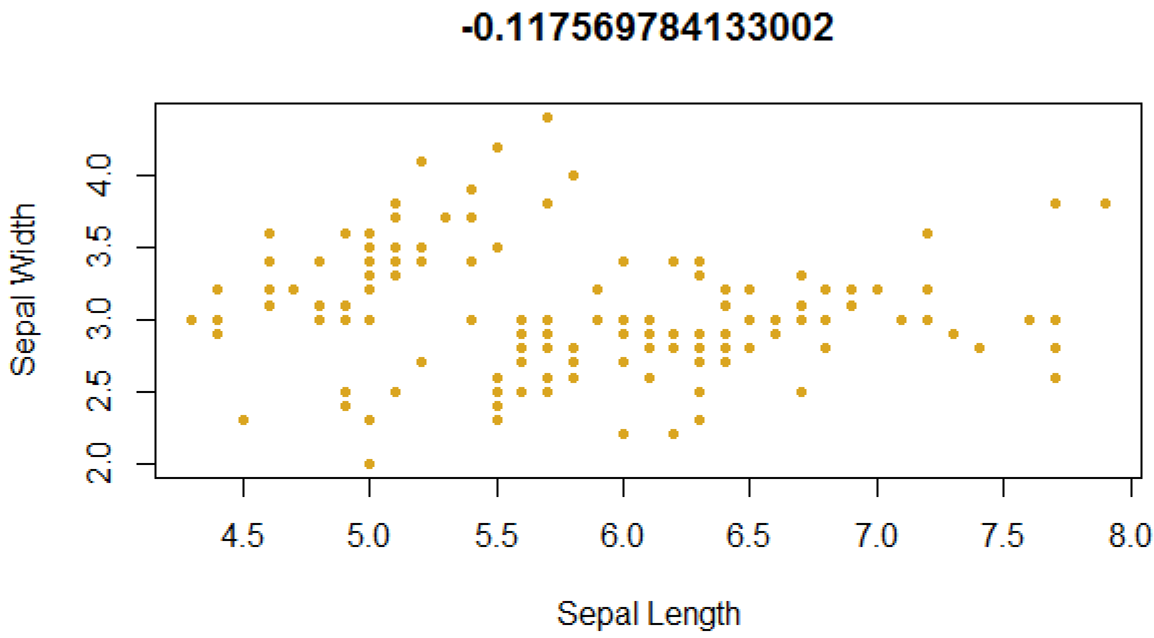

Petal Width
0.5
1.0
1.5
2.0
2.5
setosa
versicolor
virginica

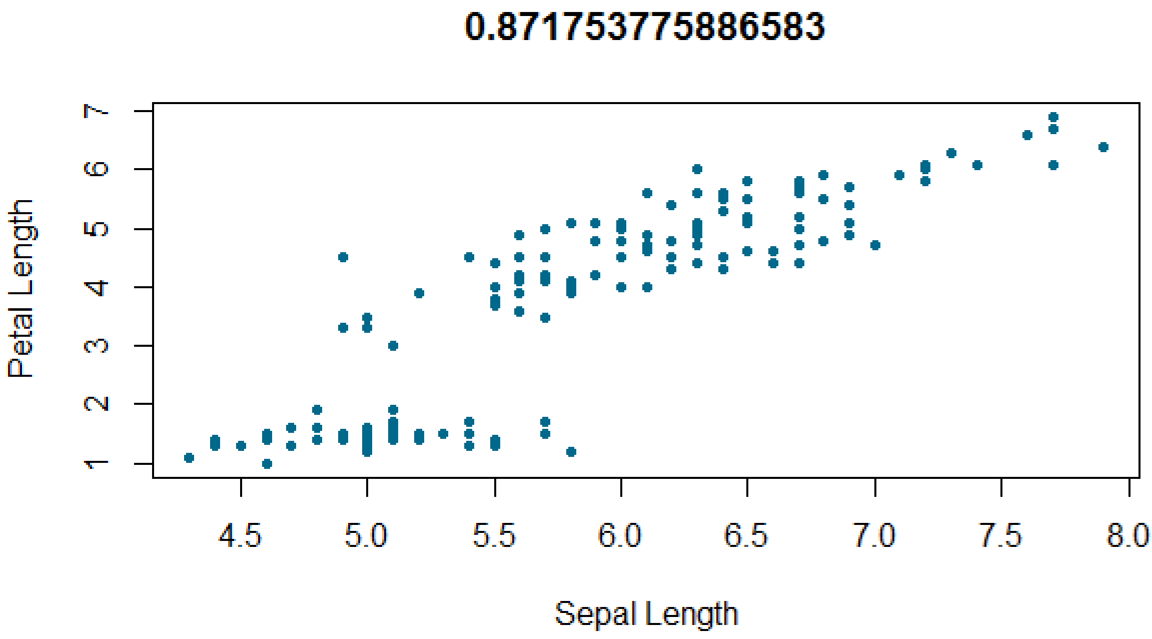

Petal Length
1
2
3
4
5
6
7
setosa
versicolor
virginica

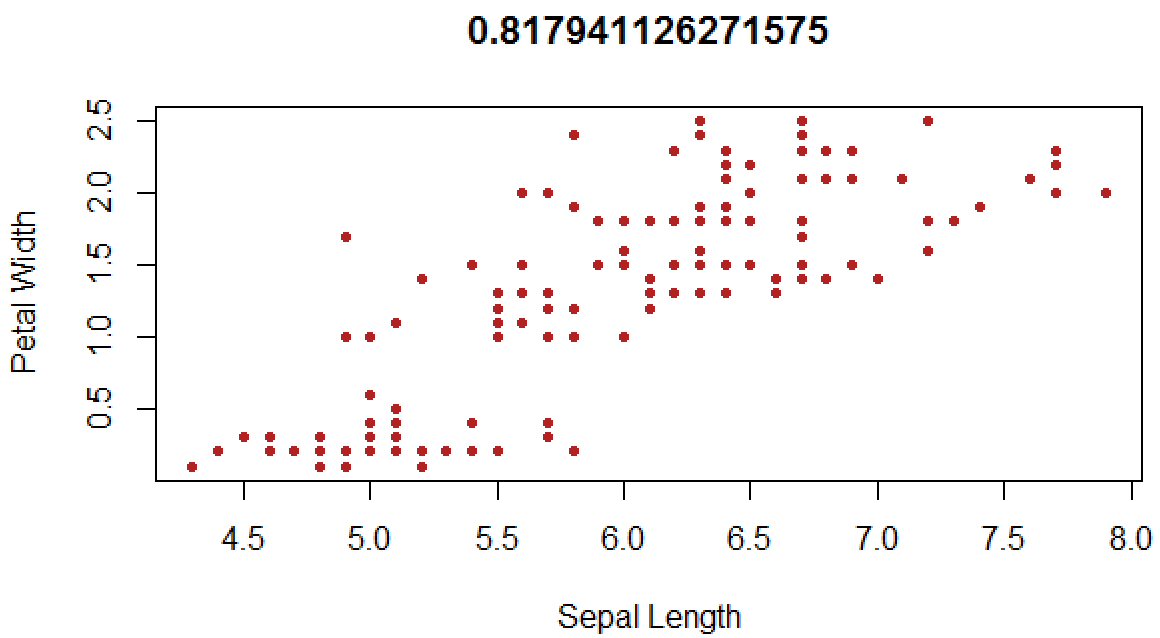

Sepal Length
7.5
6.5
5.5
4.5
setosa
versicolor
virginica

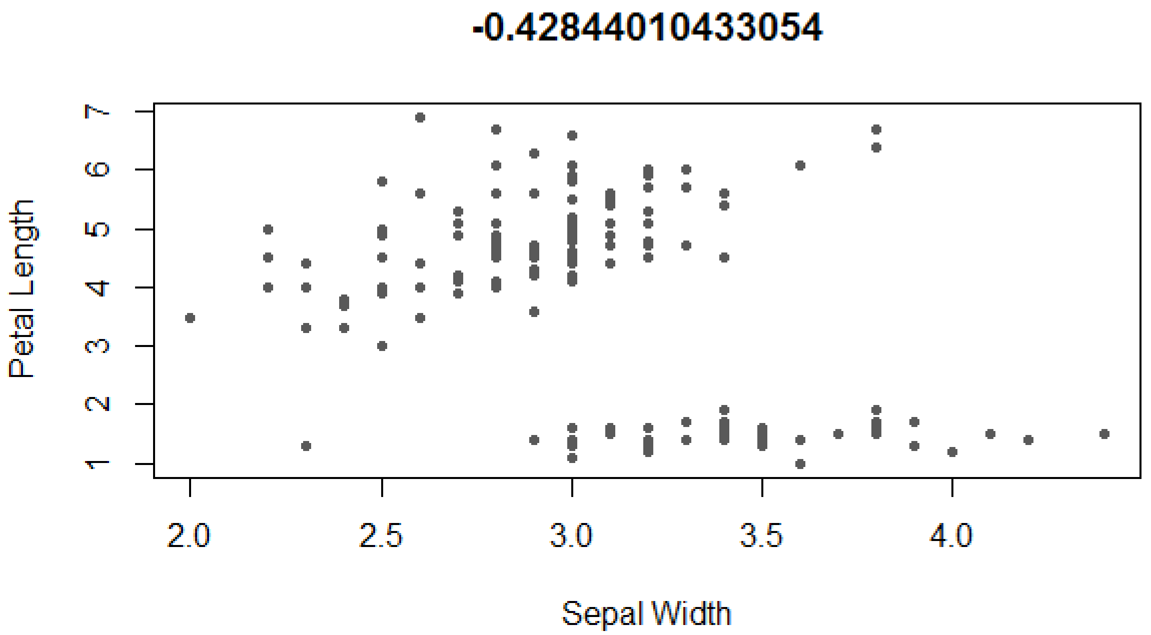

Sepal Width
4.0
3.5
3.0
2.5
2.0
setosa
versicolor
virginica

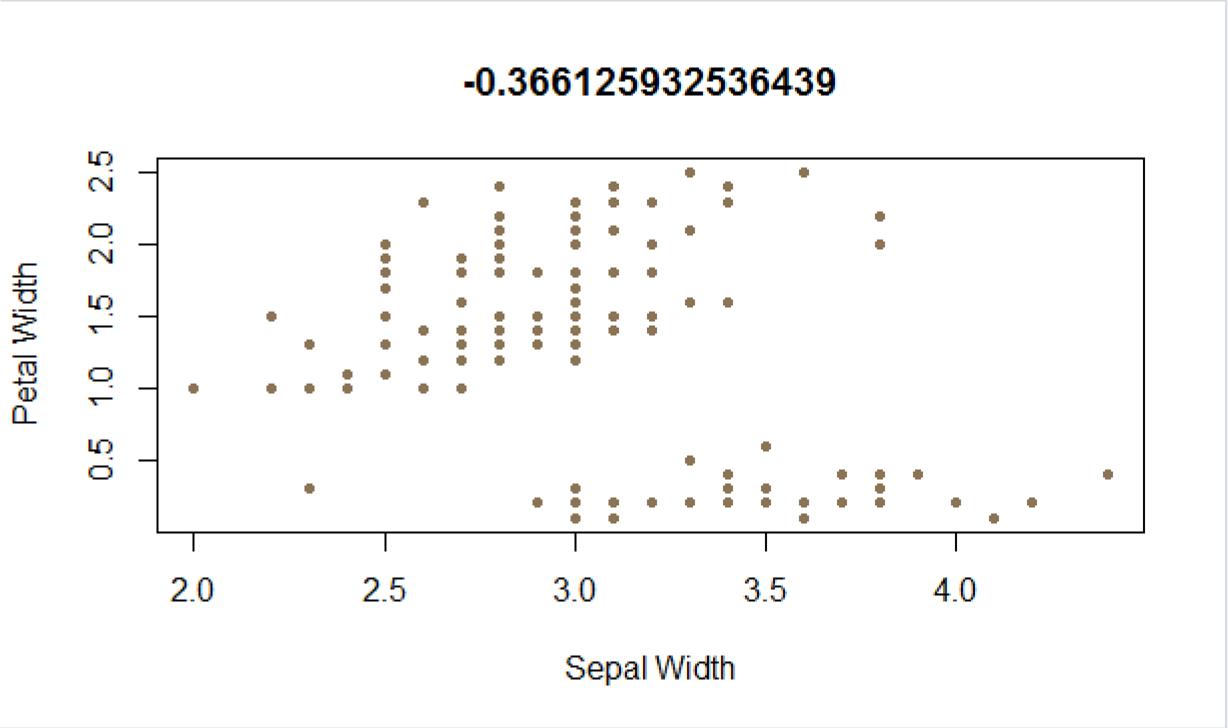
Scatter Plots and correlation values



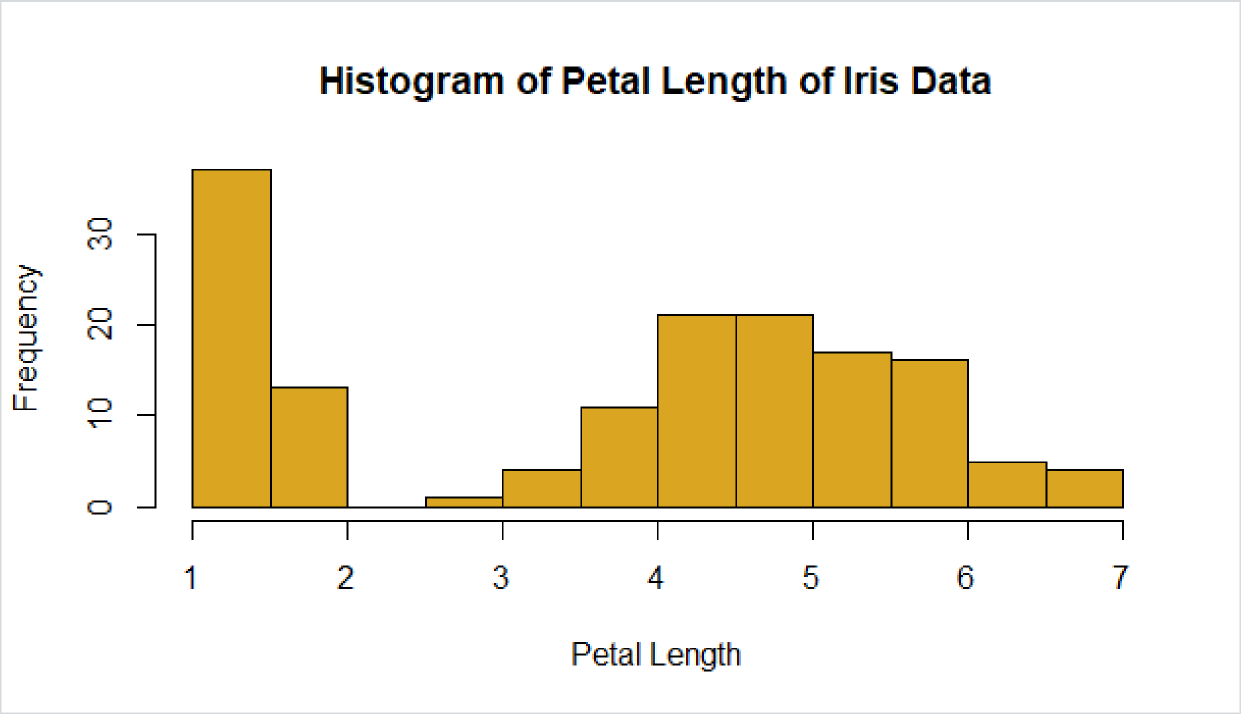


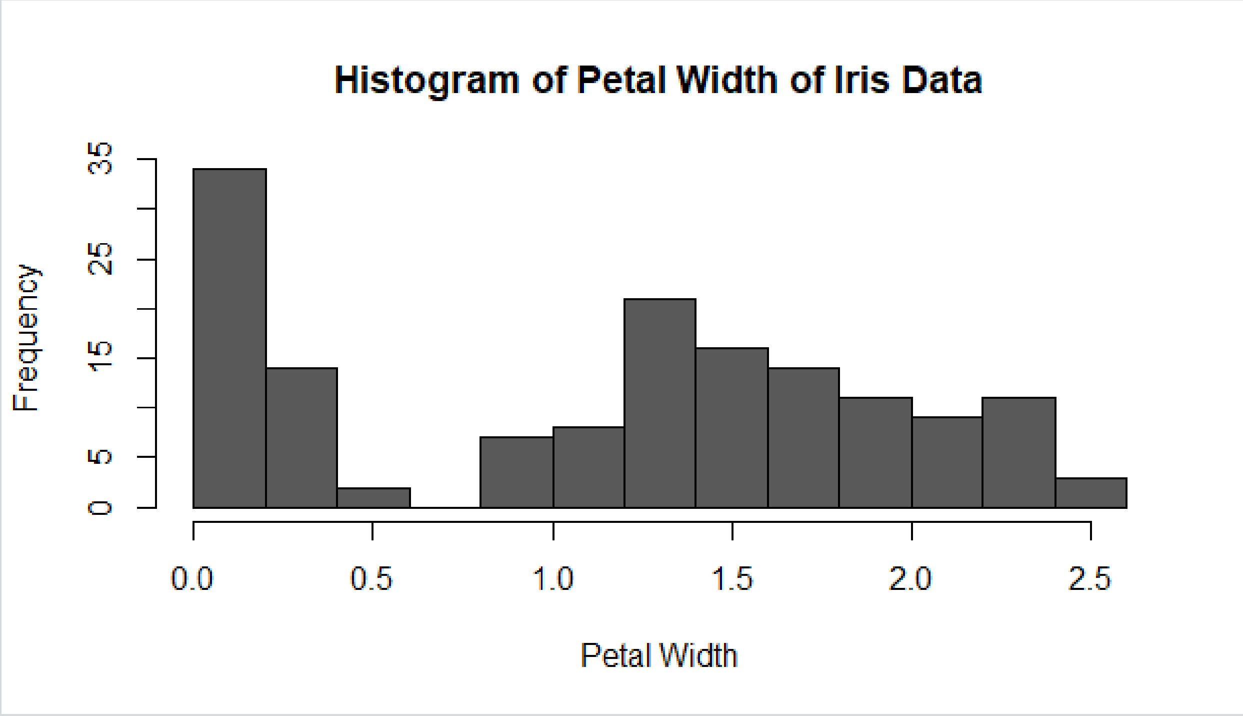

0.817941126271575
Petal Width
0.5
1.0
1.5
2.0
2.5
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
Sepal Length

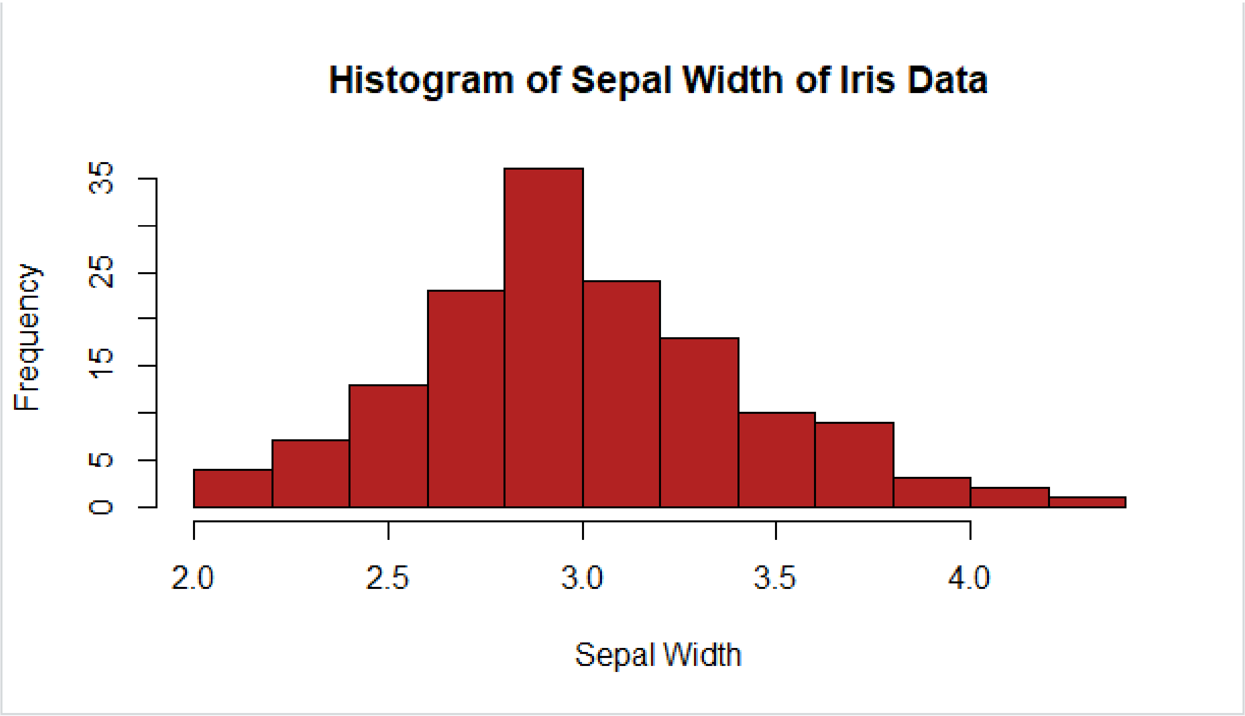

-0.42844010433054
Petal Length
1
2
3
4
5
6
7
2.0
2.5
3.0
3.5
4.0
Sepal Width

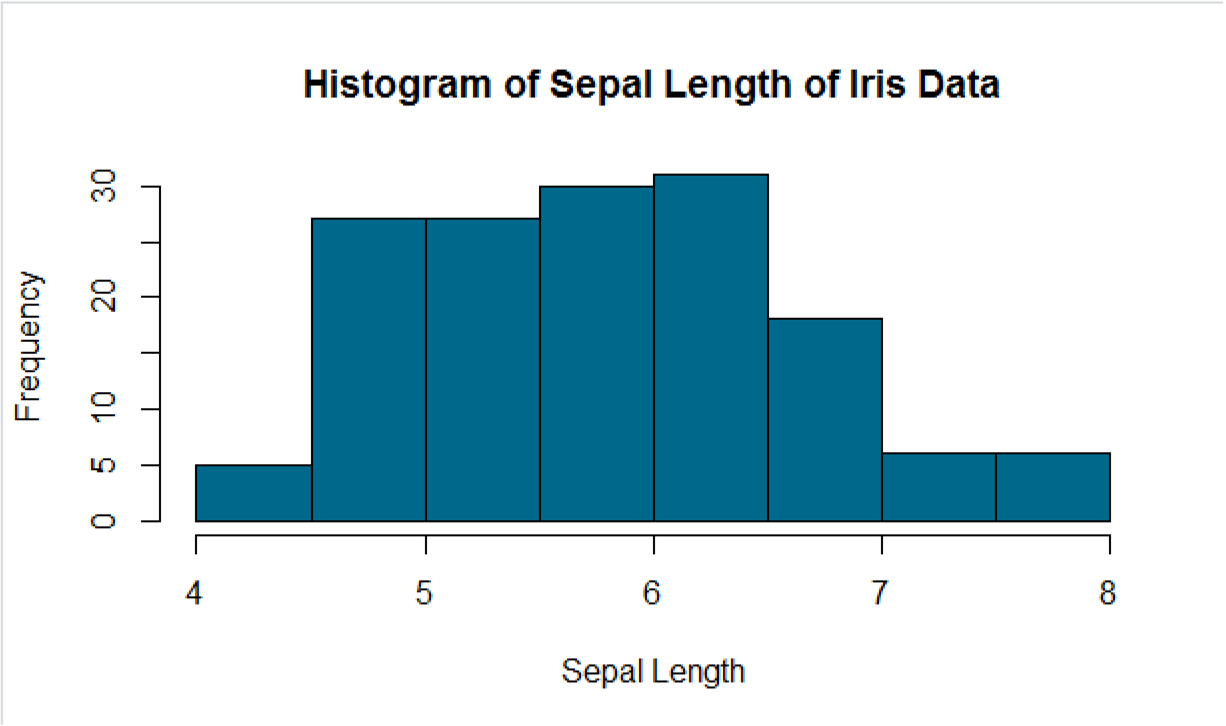

-0.366125932536439
Petal Width
0.5
1.0
1.5
2.0
2.5
2.0
2.5
3.0
3.5
4.0
Sepal Width

Histograms


Histogram of Petal Length of Iris Data
Frequency
0
10
20
30
1
2
3
4
5
6
7
Petal Length


Histogram of Petal Width of Iris Data
Frequency
0
5
15
25
35
0.0
0.5
1.0
1.5
2.0
2.5
Petal Width


Histogram of Sepal Width of Iris Data
Frequency
0
5
15
25
35
2.0
2.5
3.0
3.5
4.0
Sepal Width


Histogram of Sepal Length of Iris Data
Frequency
0
5
10
20
30
4
5
6
7
8
Sepal Length

Correlation and Covariance

Correlation between variables as 4*4 matrix

	Sepal Length	Sepal Width	Petal Length	Petal Width
Sepal Length	1.000000	-0.1175698	0.8717538	0.8179411
Sepal Width	-0.1175698	1.0000000	-0.4284401	-0.3661259
Petal Length	0.8717538	-0.4284401	1.0000000	0.9628654
Petal Width	0.8179411	-0.3661259	0.9628654	1.0000000

Covariance between variables as 4*4 matrix

	Sepal Length	Sepal Width	Petal Length	Petal Width
Sepal Length	0.6856935	-0.0424340	1.2743154	0.5162707
Sepal Width	-0.0424340	0.1899794	-0.3296564	-0.1216394
Petal Length	1.2743154	-0.3296564	3.1162779	1.2956094
Petal Width	0.5162707	-0.1216394	1.2956094	0.5810063

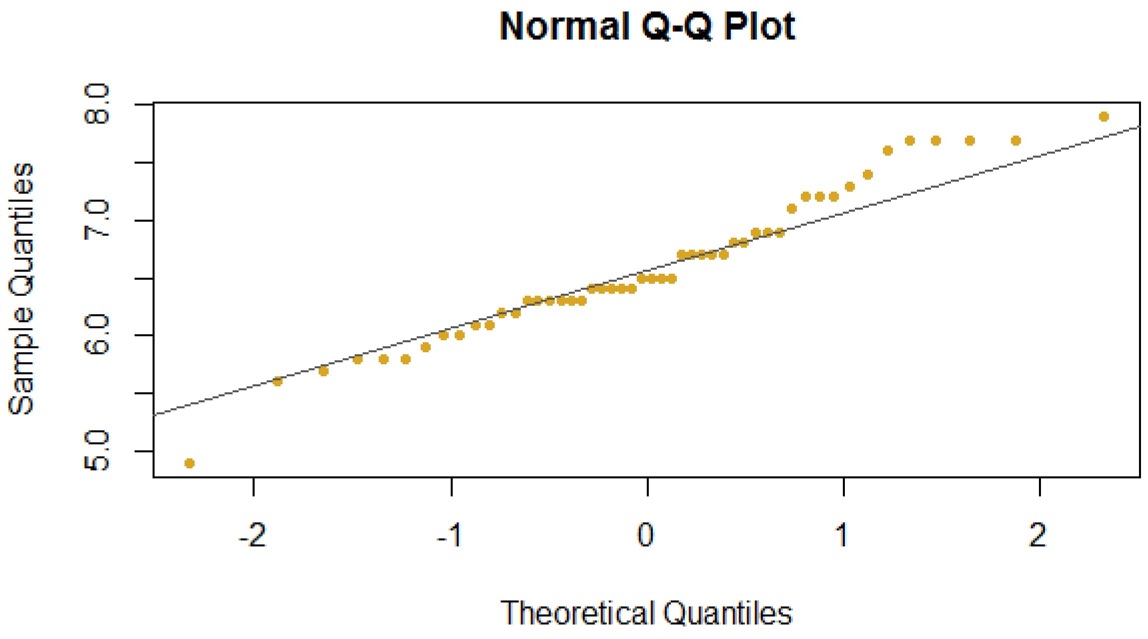
Inferential Statistics

- We will test the following hypothesis using α =0.05.

- Ho: The mean of virginica iris Sepal Lengths equals the mean of versicolor iris Sepal Lengths
- Ha: The mean of virginica iris Sepal Lengths is greater than the mean of versicolor iris Sepal Lengths.

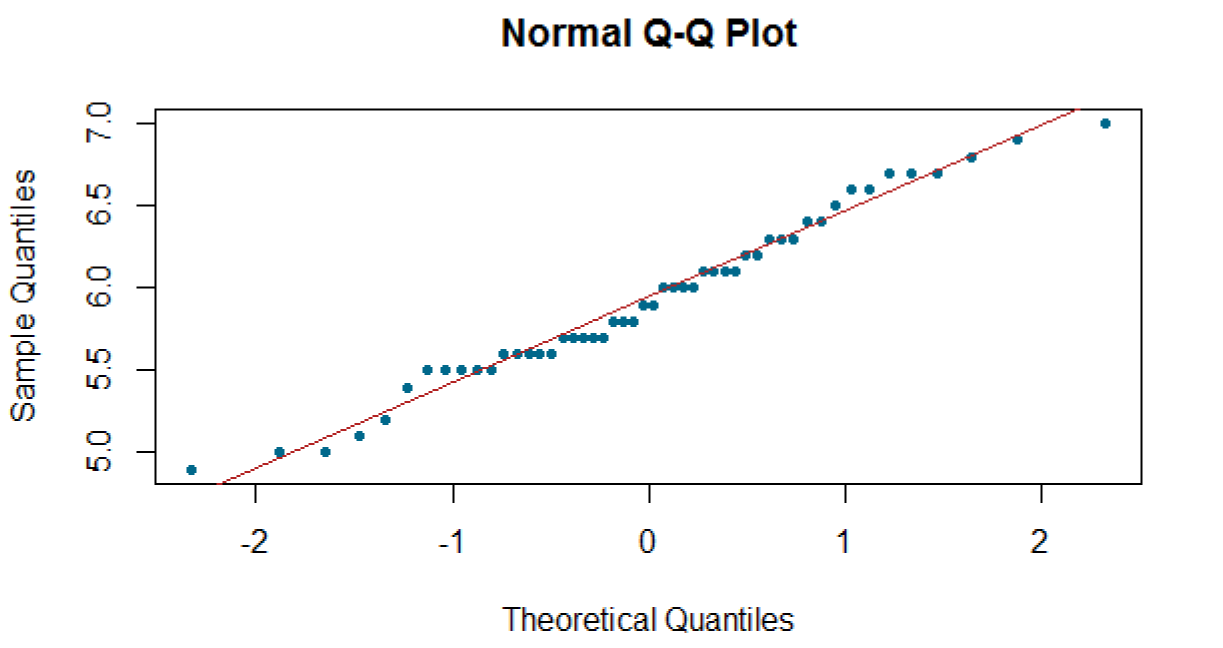
- check that the data is normally distributed

1. Viriginica



- data of Viriginica are normally distributed.

2. Versicolor



- data of Versicolor are normally distributed
- t-test results

  - mean of x = 6.588.

  - mean of y = 5.936.

- (effect size = 6.588 - 5.936 = 0.652)

Conclusion

Our test statistic is 5.6292 with 94.025 degrees of freedom. The P-value is 9.3×10–89.3×10–8, which is less than our α of 0.05. We therefore reject the null hypothesis and conclude that there is evidence that virginica sepal lengths are larger than versicolor sepal lengths. The effect size is 0.652, meaning on average virginica sepal lengths are 0.652 cm longer than versicolor. We have a 95% confidence interval of [0.46, Inf], meaning there is a 95% chance that the true difference in means falls in this range.

Classification using K-Mean Clustering

First, we must preprocess the dataset then apply k-means clustering algorithm and finally verify results of clustering.

1. Preprocess the dataset

Since clustering is a type of Unsupervised Learning, we would not require Class Label(output) during execution of our algorithm. We will, therefore, remove Class Attribute “Species” and store it in another variable. We would then normalize the attributes between 0 and 1 using our own function.

- Class without ”Species” Attribute.

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1          5.1         3.5          1.4         0.2
## 2          4.9         3.0          1.4         0.2
## 3          4.7         3.2          1.3         0.2
## 4          4.6         3.1          1.5         0.2
## 5          5.0         3.6          1.4         0.2
## 6          5.4         3.9          1.7         0.4
```

- Class after normalization.

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1   0.22222222   0.6250000   0.06779661  0.04166667
## 2   0.16666667   0.4166667   0.06779661  0.04166667
## 3   0.11111111   0.5000000   0.05084746  0.04166667
## 4   0.08333333   0.4583333   0.08474576  0.04166667
## 5   0.19444444   0.6666667   0.06779661  0.04166667
## 6   0.30555556   0.7916667   0.11864407  0.12500000
```

2. Apply k-means clustering algorithm

- No. of records in each cluster

```
39 50 61
```

- value of cluster center data point value (3 centers for k=3)

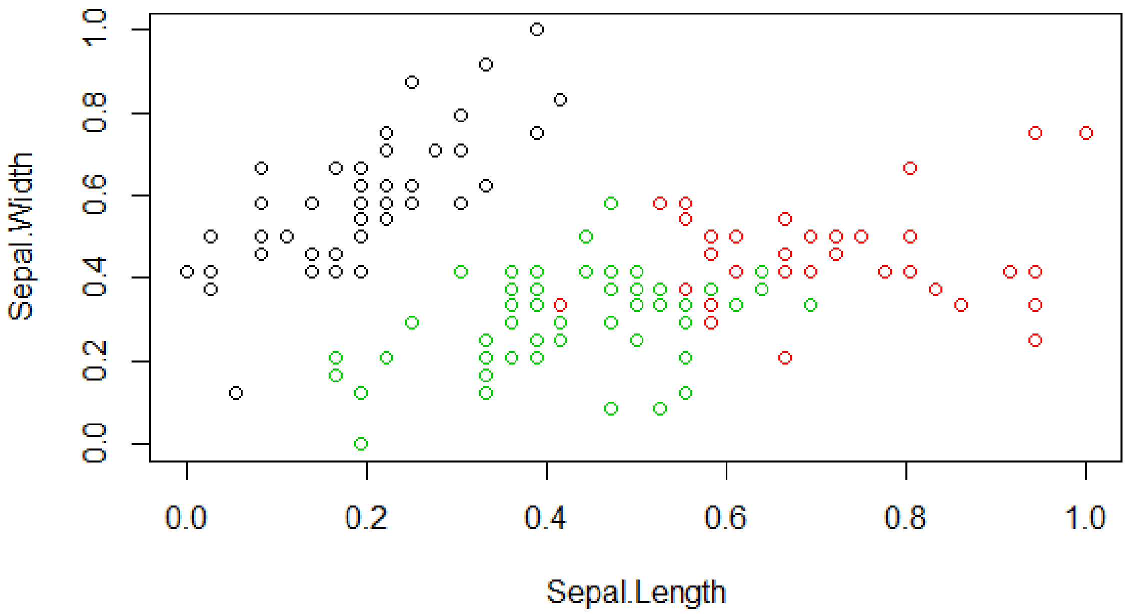
```
  Sepal.Length Sepal.Width Petal.Length Petal.Width
1    0.4412568   0.3073770   0.57571548  0.54918033
2    0.7072650   0.4508547   0.79704476  0.82478632
3    0.1961111   0.5950000   0.07830508  0.06083333
```

- cluster vector showing the cluster where each record falls

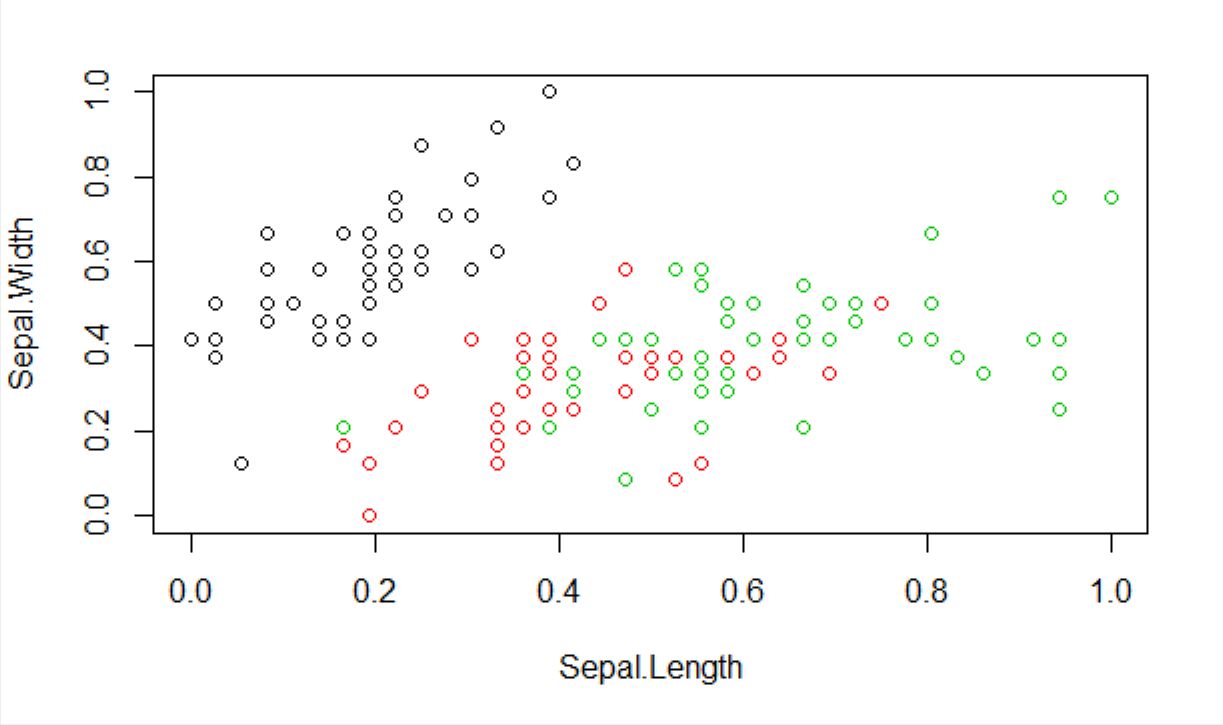
```
[1] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[36] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[71] 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 2 2 2
[106] 2 1 2 2 2 2 2 2 1 2 2 2 2 2 1 2 1 2 1 2 2 1 1 2 2 2 2 2 1 1 2 2 2 1 2
[141] 2 2 1 2 2 2 1 2 2 1
```

3. Verify results of clustering

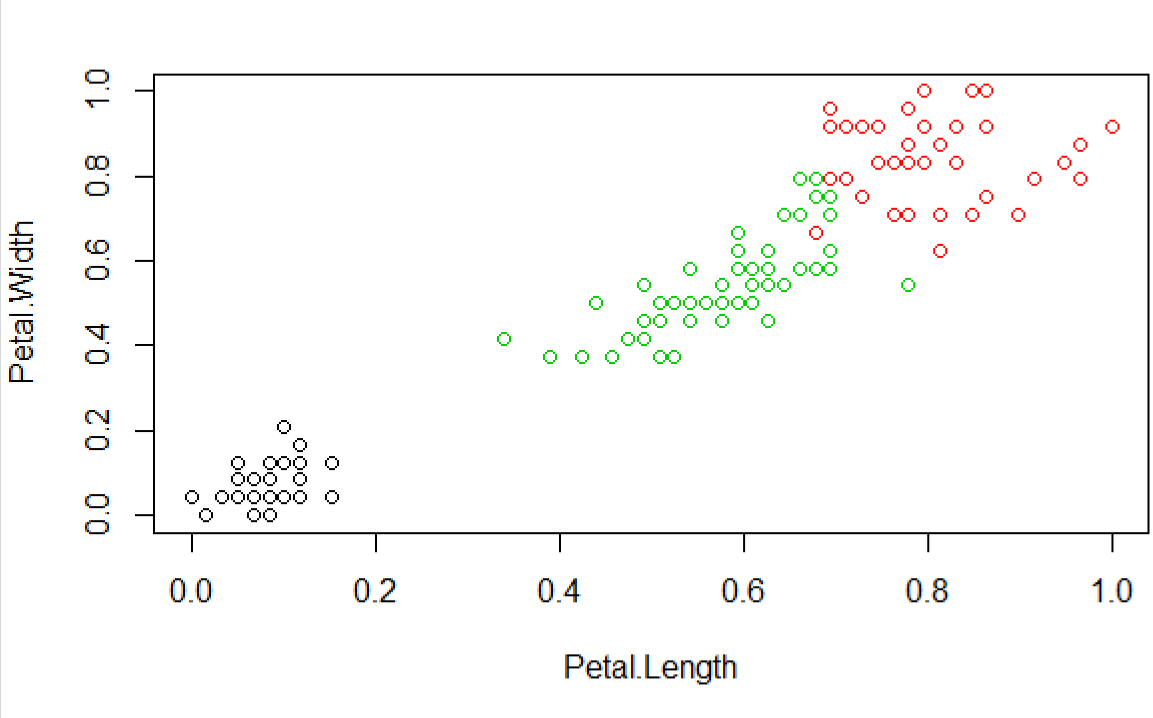
distribution in clusters between sepal length data points and sepal width data points.



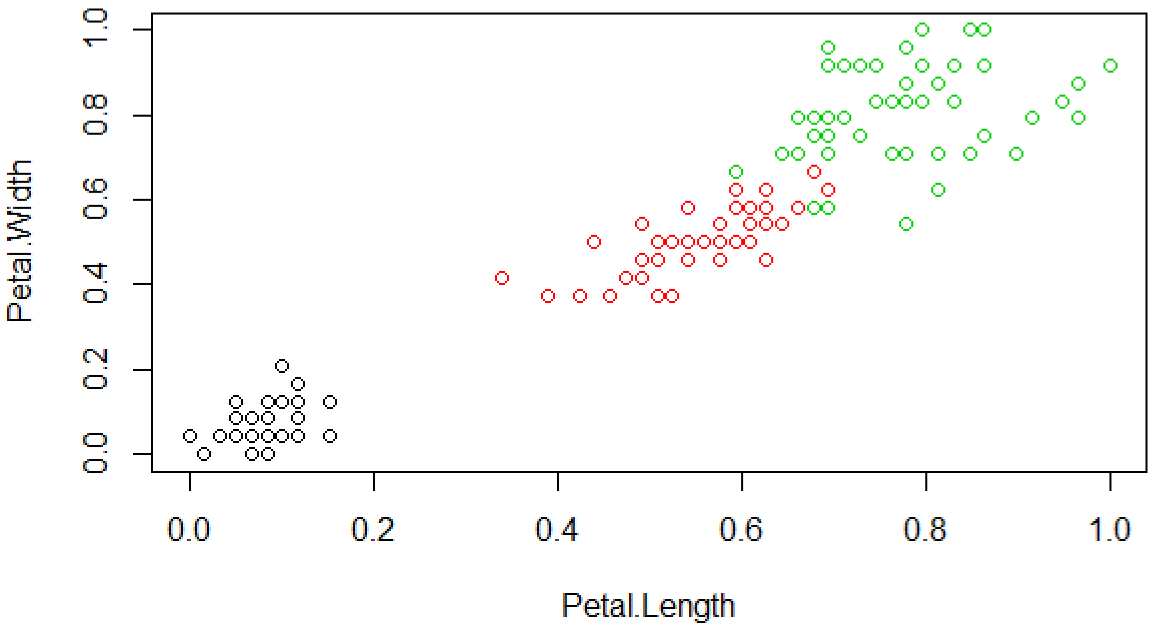
distribution originally by "class" attribute between sepal length data points and sepal width data points.



distribution in clusters between petal length data points and petal width data points.



distribution originally by "class" attribute between petal length data points and petal width data points.



- Result of table shows that Cluster 1 corresponds to Virginica, Cluster 2 corresponds to Versicolor and Cluster 3 to Setosa.

```
    setosa versicolor virginica
 1     0      47       14
 2     0       3       36
 3    50       0        0
```

Total number of correctly classified instances are: 36 + 47 + 50= 133

Total number of incorrectly classified instances are: 3 + 14= 17

Accuracy = 133/(133+17) = 0.88 , our model has achieved 88% accuracy.

Summary

We can conclude that there is evidence that virginica sepal lengths are larger than versicolor sepal lengths. The effect size is 0.652, meaning on average virginica sepal lengths are 0.652 cm longer than versicolor.

Also, there is a strong positive correlation (between sepal length, petal length) and (between sepal length, petal width) and perfect positive correlation between petal width and petal length.

Finally, to improve the classification accuracy further, we may try different values of “k”. In some cases, it is also beneficial to change the algorithm in case k-means is unable to yield good results.