Iris Dataset Analysis Report

**EXECUTIVE**:

It is the analysis of the data about all the flowers belonging to three species and find out what are there distinctive characteristics are such as sepal length, sepal width, petal length and petal width.

**INTRODUCTION AND PURPOSES**:

The plots made during the analysis show the most prominent data from the distribution of species on a pie chart to the decision tree that shows the classification of the species based on all the other features in the dataset.

The analysis was conducted by Suven Consultants and Technology Pvt. Ltd. for an online internship corresponding to the course of “Data Analysis Using R”. The analyser of this dataset is Rahul Sar.

**LIMITATIONS**:

This analysis has got certain limitations:

* Accuracy percent of the prediction model for this dataset is less.
* Analysis cannot be automated using R language.

**METHODS**:

* **Verification of data consistency:**

Checking and omitting the NA and blank values.

* **Box Plotting:**

Generic bar plotting of input variables.

* **Conversion of data:**

Converting the categorical data into numeric values.

* **Validation of the statistical analysis:**

Validating the statistical data.

* **Decision Tree:**

Finding correlation between data items and plotting a decision tree for classification.

**SAMPLE**:

The personal who contributed to the analysis are as follows:

* **Niraj Sharma**: Instructor at SCTPL, taught how to use R language for data analysis.
* **Rahul Sar**: Student at SCTPL, taught by Niraj Sharma, Made the full analysis on the Iris Dataset along with this report with the help of Niraj Sharma.

**INSTRUMENTATION**:

The tools used by us are as follows:

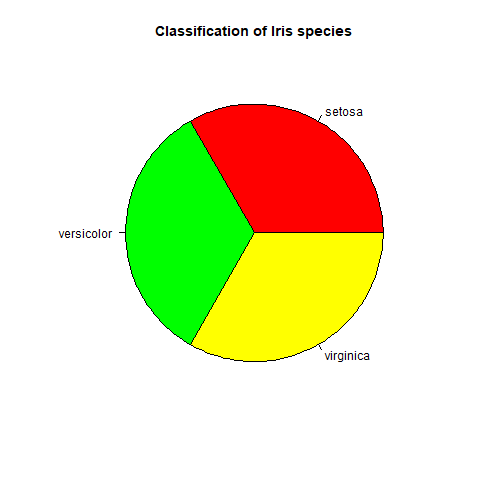
* Rstudio IDE
* R V3.5.1

**RESULTS**:

We did exploratory analysis on the iris dataset and below are our findings:

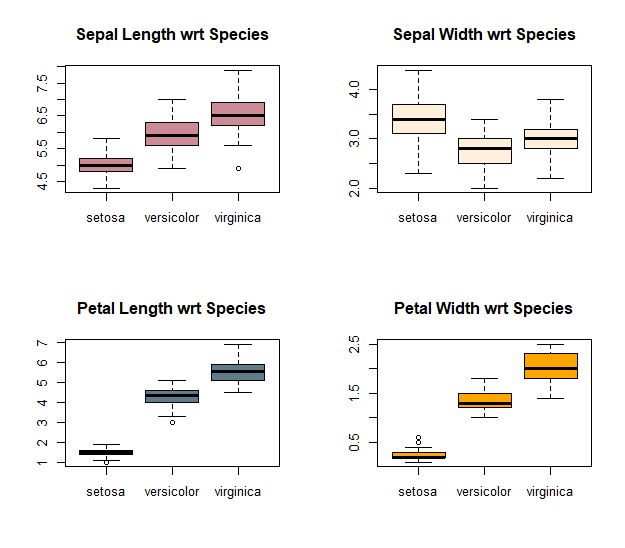
* The dimensions of this dataset is 150 rows and 5 columns.
* The names of features are "sepal\_length", "sepal\_width", "petal\_length", "petal\_width", "species"
* The class of the iris dataset is data.frame.
* The class of species is character, rest other features are of numeric data type.
* There are no missing values in the Iris dataset.
* Sepal length has Min.:4.3, Max.:7.9, Mean:5.843, Median:5.8, 1st Quantile:5.1, 3rd Quantile:6.4.
* Sepal width has Min.:2, Max.:4.4, Mean:3.054, Median:3, 1st Quantile:2.8, 3rd Quantile:3.3.
* Petal length has Min.:1,Max.:6.9,Mean:3.759,Median:4.35,1st Quantile:1.6,3rd Quantile:5.1.
* Petal width has Min.:0.1,Max.:2.5,Mean:1.199,Median:1.3,1st Quantile:0.3,3rd Quantile:1.8.

We plotted a box plot graph according to the species and pie chart as well.



There are equal number of records for each species type.

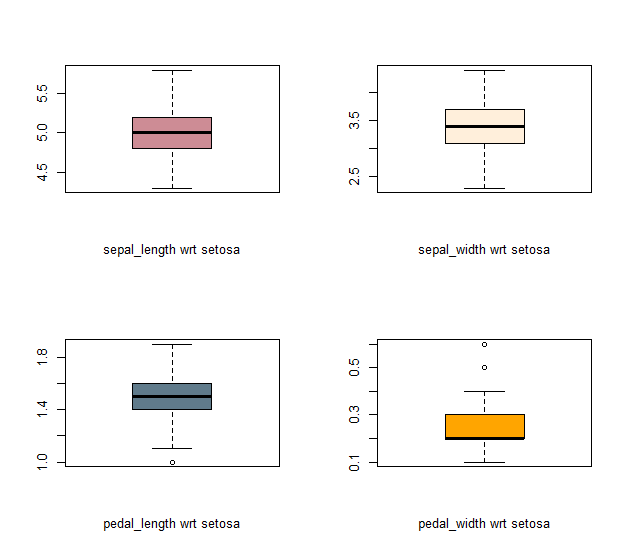
Boxplot for Iris dataset with respect to Species



Virginica has the highest frequency for Sepal Length. Setosa has the highest frequency for Sepal Width. Virginica has the highest frequency for Petal Length. Virginica has the highest frequency for Petal Width.

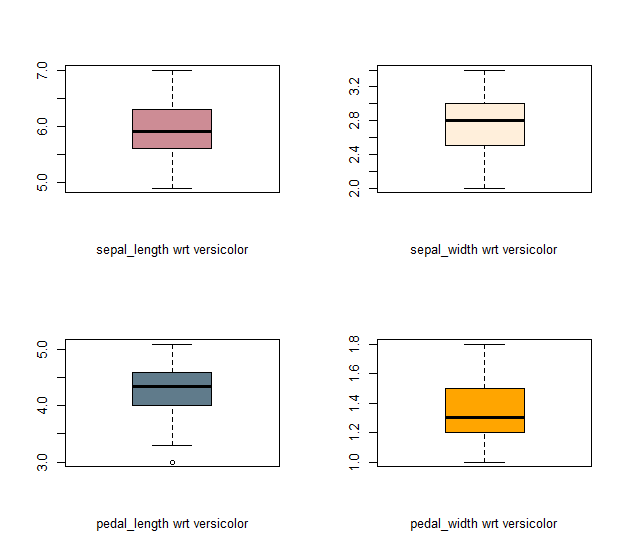
Then we subset of species in different R objects. The species consists of setosa, versicolor and virginica. Then we plotted box plot for each species.

Box plot with respect to Setosa Species.



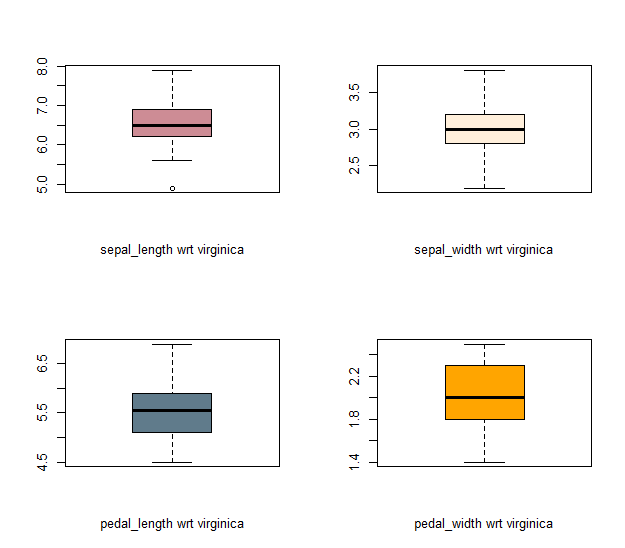
Sepal Length has the highest frequency varying from 4.5 to 5.5.

Boxplot with respect to Versicolor Species

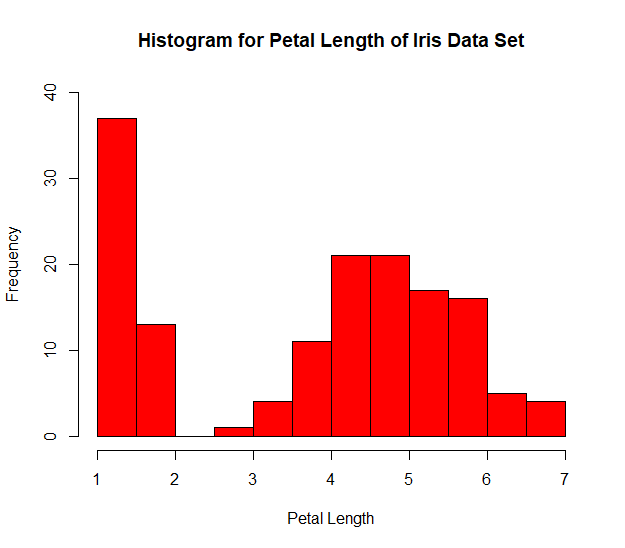


Sepal Length has the highest frequency ranging from 5 to 7.

Boxplot with respect to Virginica Species

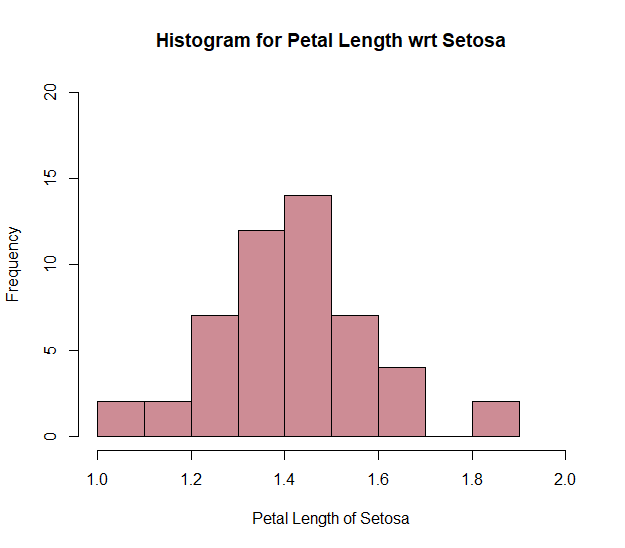


Sepal Length has the highest frequency ranging from 5 to 8.

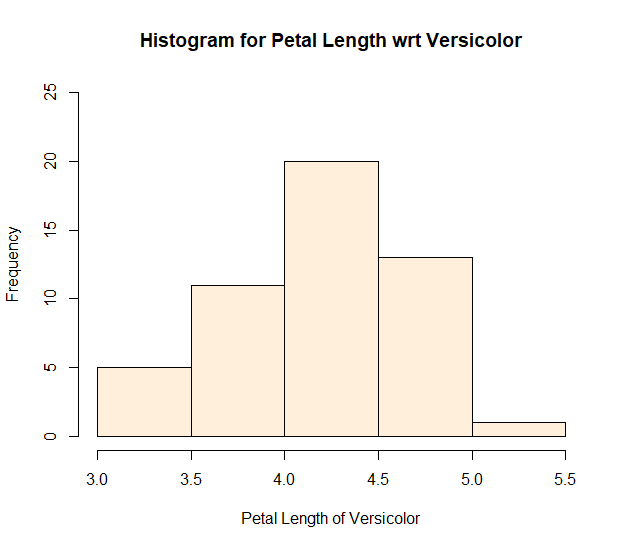


Petal length 1 has the highest frequency.

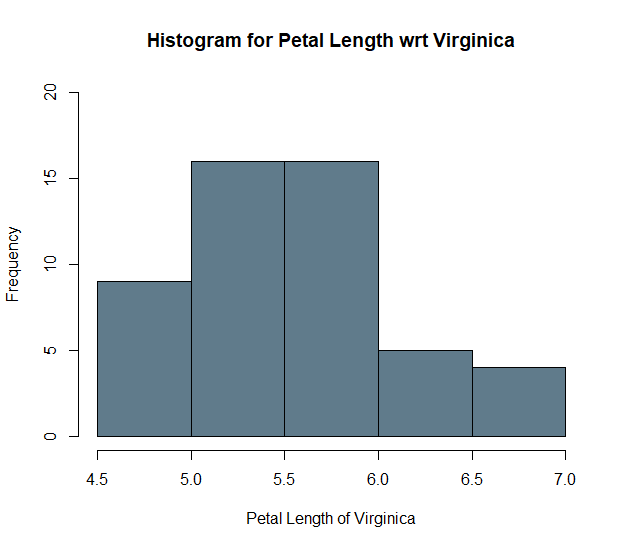
Then we plotted histograms for petal length based on different species.



Majority of the Setosa flowers have Petal Length between 1.4 and 1.5.



Majority of the versicolor flowers have petal length between 4.0 and 4.5.

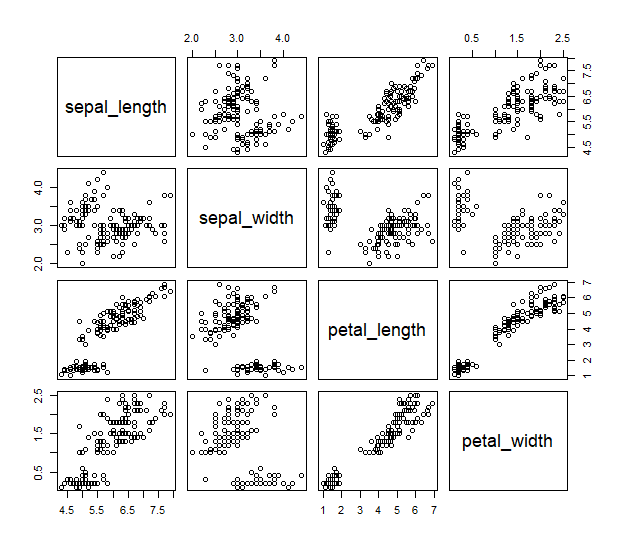


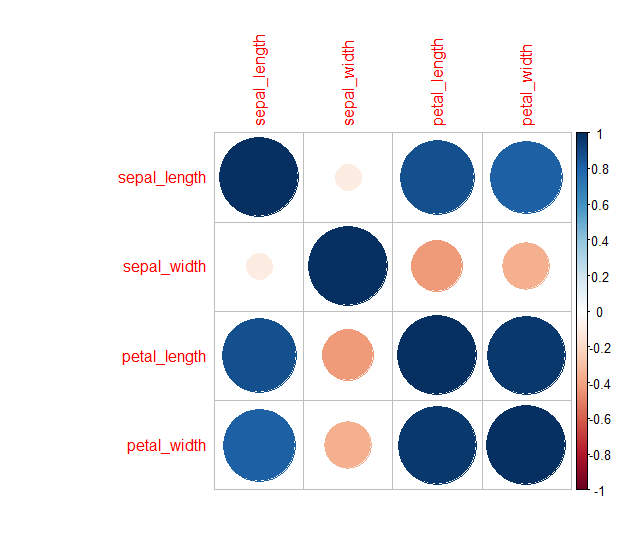
Majority of the versicolor flowers have petal length between 5.0 and 6.0.

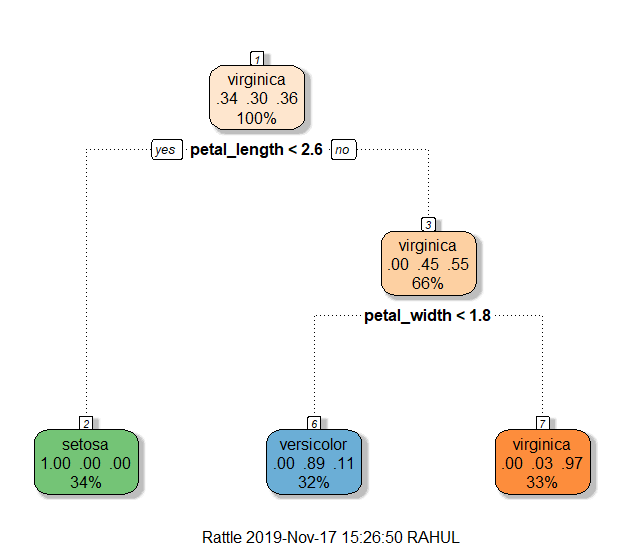
We found out correlation between by randomly picking up 2 features from the dataset that is Sepal Length and Sepal Width and the value derived is -0.1093692 which indicates that there is weak downhill(negative) relationship between these 2 features.

Here is the correlations here for the rest features.

|  |
| --- |
| sepal\_length sepal\_width petal\_length petal\_width  sepal\_length 1.0000000 -0.1093692 0.8717542 0.8179536  sepal\_width -0.1093692 1.0000000 -0.4205161 -0.3565441  petal\_length 0.8717542 -0.4205161 1.0000000 0.9627571  petal\_width 0.8179536 -0.3565441 0.9627571 1.0000000 |
|  |
| |  | | --- | |  | |







We then plotted a decision tree based on iris species which showed us that the setosa flowers have a petal length of less than 2.5. This decision tree helped us to classify the data by plotting each species as a different node based on the values of all the other features.

**RECOMMENDATIONS**:

This is just a practice analysis and no real-life recommendations may not be possible.

**REFERENCES**:

<https://www.analyticsvidhya.com/>

<https://www.r-bloggers.com/>

<https://www.tutorialspoint.com/index.htm>

Notes provided by SCTPL.