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 Aahan Gupta.

**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.

* **Bar 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.

**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.
* **Aahan Gupta**: Student at SCTPL, taught by Niraj Sharma, Made the full analysis on the Titanic Dataset along with this report with the help of Ashish Gupta and Niraj Sharma.
* **Ashish Gupta**: Father of Aahan Gupta, who helped Aahan Gupta in making this report.

**INSTRUMENTATION**:

The tools used by us are as follows:

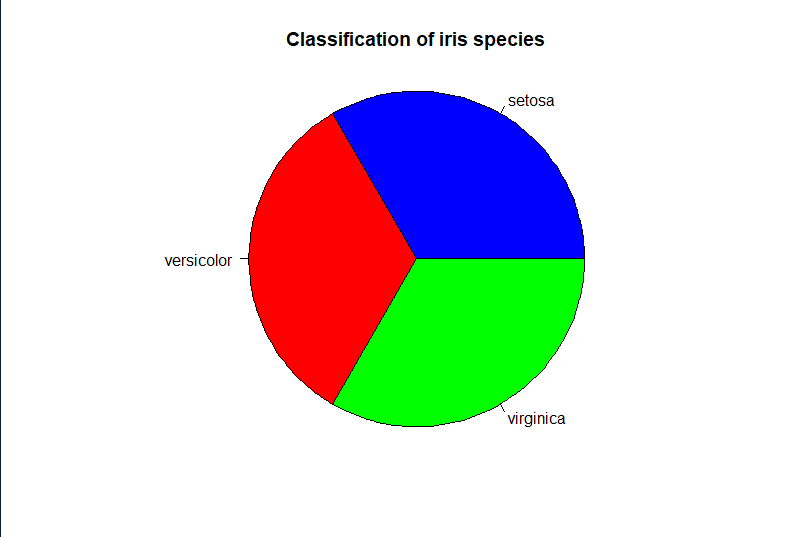
* Rstudio IDE
* R V3.5.1
* Sublime Text Editor

**RESULTS**:

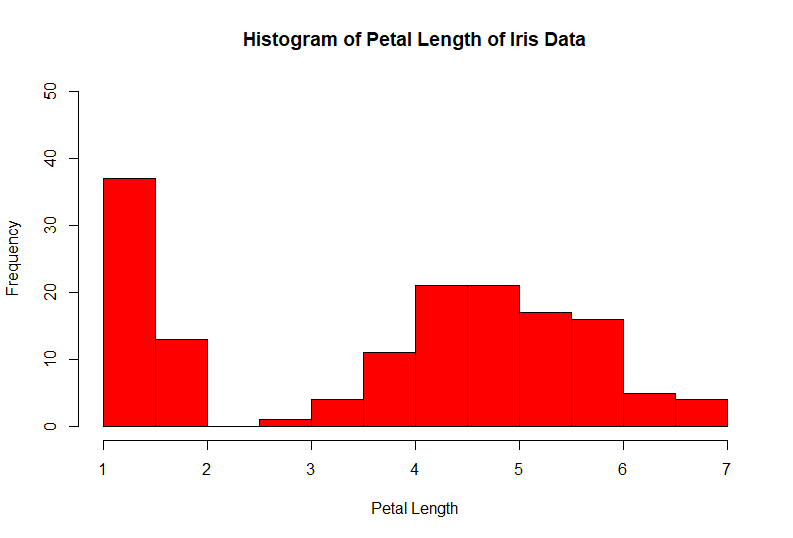
We did exploratory analysis on the iris dataset and below are our findings:

* 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.

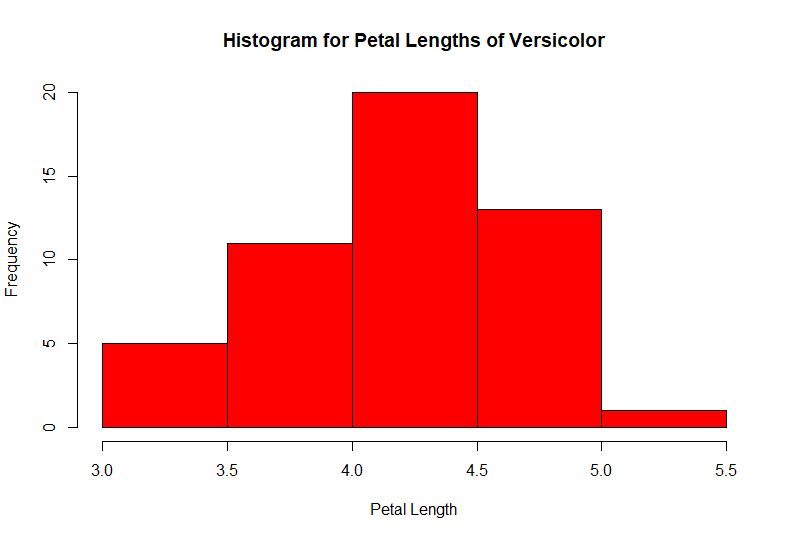


There are equal number of records for each species type

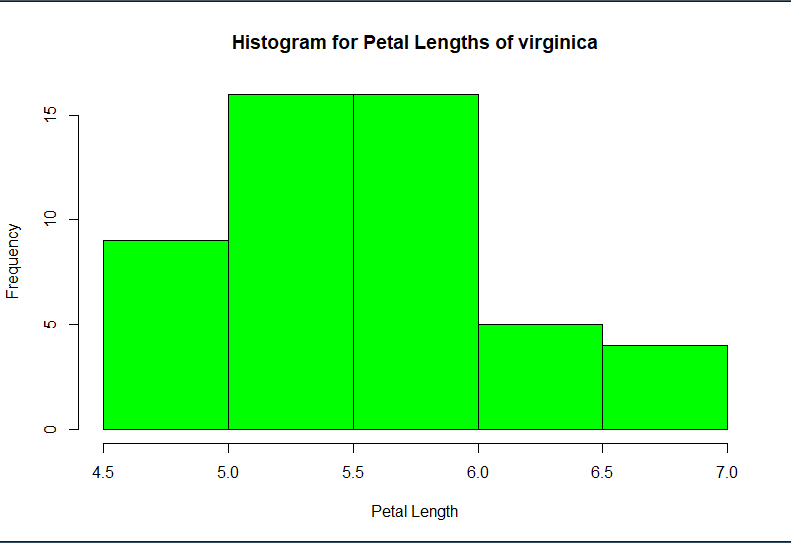


Petal length 1 has the highest frequency.

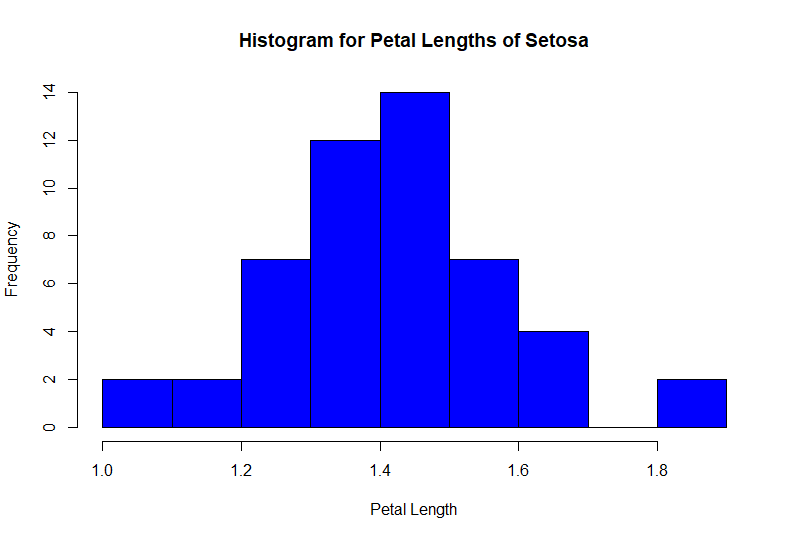
Then we plotted histograms for petal length based on different species.



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



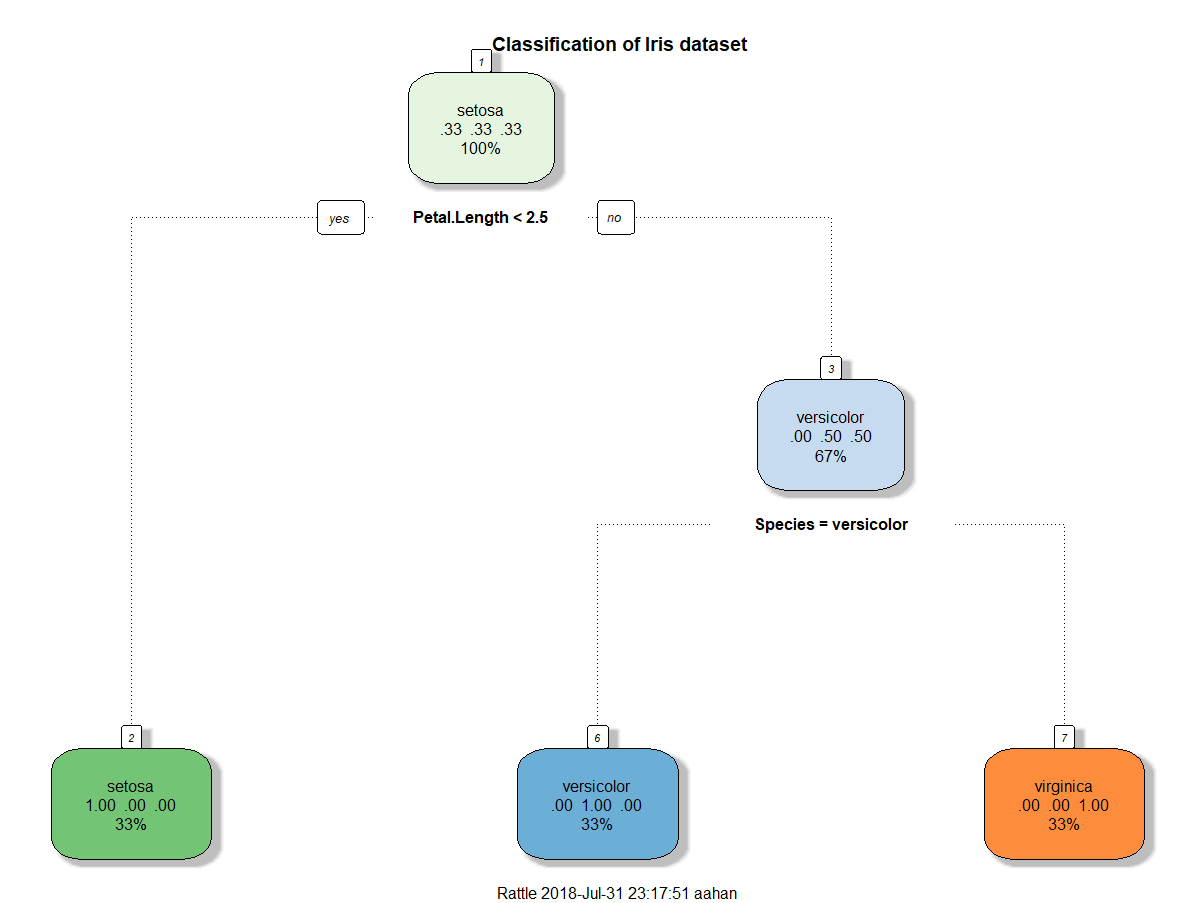
Majority of the virginica flowers have petal length between 5.0 and 6.0.



The majority of the Setosa flowers have petal length between 1.4 and 1.5.

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.

Write the correlations here for the rest features.



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.