Iris Dataset Analysis Report

**EXECUTIVE**:

Iris dataset has features – sepal\_length,sepal\_width,petal\_length,petal\_width,species. There are equal number of records for each species type. Petal length 1 has the highest frequency of records.Then we plotted histograms shown below to find out the frequency based on different criteria.Then we plotted histograms for petal length based on different species. There is no strong correlation between any of the two features in this dataset.

**INTRODUCTION**:

The plots made during the analysis show the most prominent data from the distribution data based on the species type on a pie chart and 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 Neha Mirani.

**LIMITATIONS**:

This analysis has got certain limitations:

* Accuracy percentage 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:**

Histogram 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**:

People who contributed to the analysis are as follows:

* **Niraj Sharma**: Instructor at SCTPL, taught us how to use R language for data analysis
* **Neha Mirani**: Student at SCTPL, trained by Niraj Sharma, did the full analysis on the Iris Dataset along with this analysis report

**INSTRUMENTATION**:

The tools used by us are as follows:

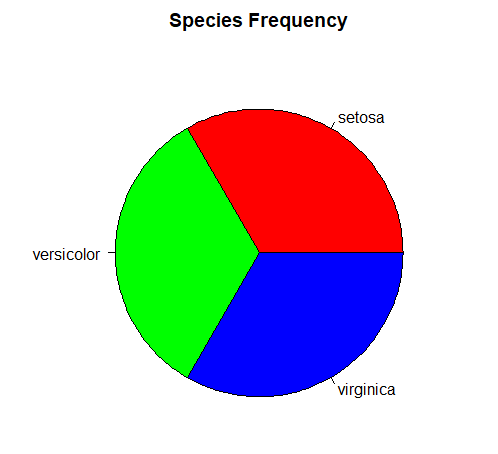
* R Studio IDE
* R v3.5.1

**RESULTS**:

1.)We did exploratory analysis on the Iris dataset and below are our findings:

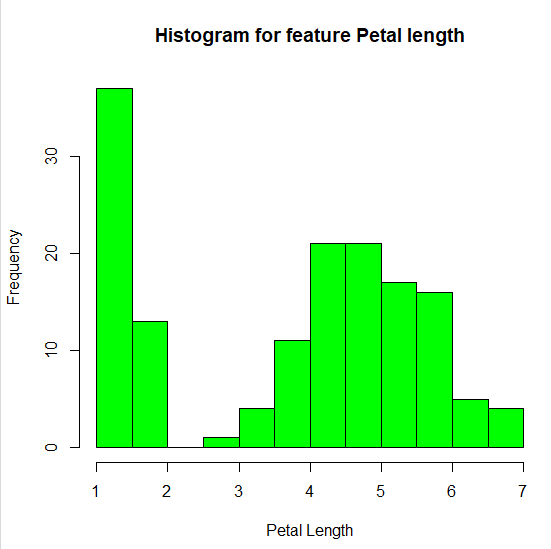
* The dimensions of this dataset is 150 rows and 5 columns
* The names of features are "sepal\_length", "sepal\_width", "petal\_length", "petal\_width", "species"
* The class of the iris dataset is data.frame
* The class of species is character, rest other features are of numeric data type
* There are no missing values in this 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

2.)We plotted a pie chart according to the species as below:



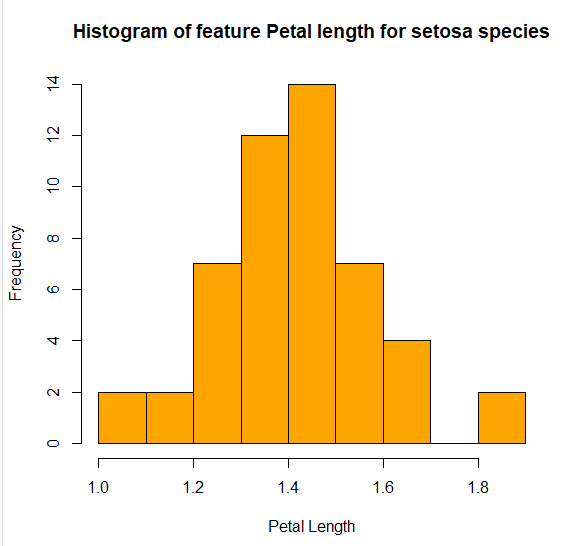
**Finding:** There is equal number of records for each species type.

3.) Next we plotted a histogram to find out the frequency for the Petal Length values as below:

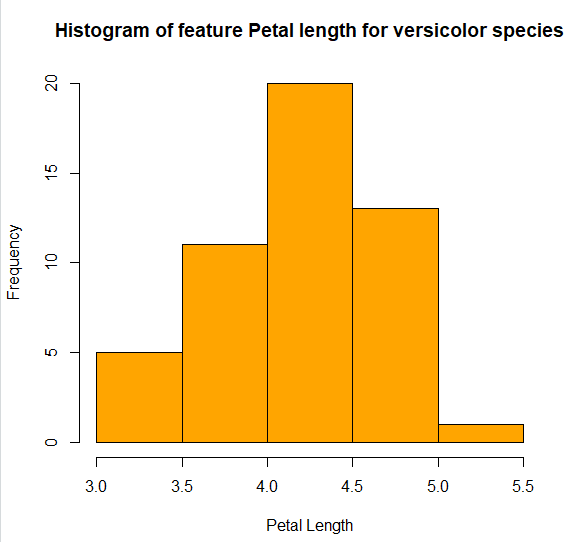


**Finding:** Petal Length 1 has the highest frequency.

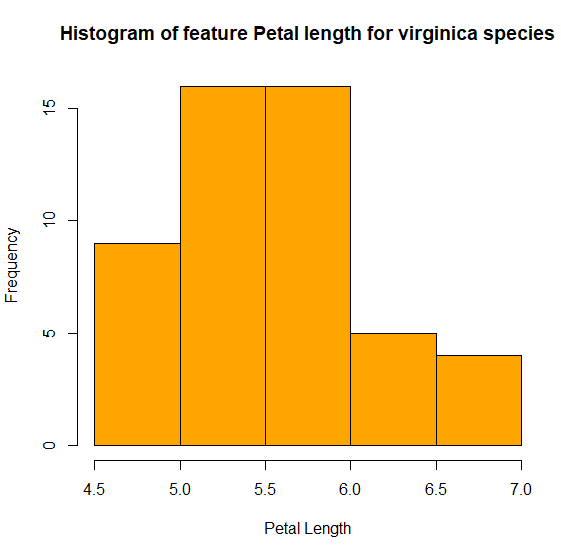
4) Next we plotted a histogram for the Petal Length feature based on each different species as below:



**Finding:** The majority of the Setosa flowers have petal length between 1.4 and 1.5



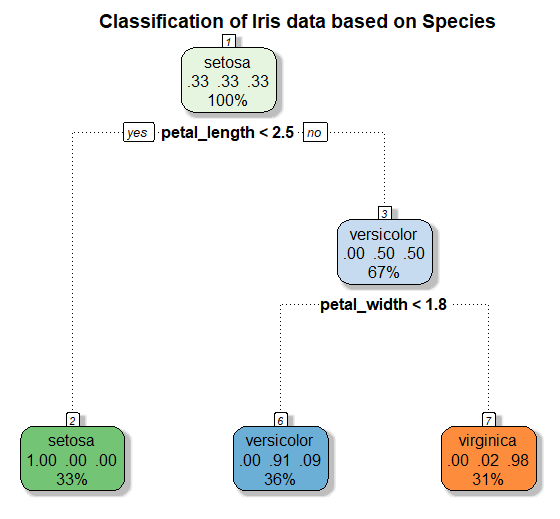
**Finding:** The majority of the Versicolor flowers have petal length between 4.0 and 4.5



**Finding:** Majority of the Virginica flowers have petal length between 5.0 and 6.0

5) We found the correlation between randomly picked two features from the dataset - Sepal Length and Sepal Width and the value derived is -0.1093692 which indicates that there is weak downhill (negative) relationship between these two features

6) We then plotted a decision tree based on Iris species which 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 real-life recommendation may not be possible.

**SUMMARY**:

We performed exploratory analysis on Iris dataset. Then we plotted different graphs like boxplot, histogram and scatterplot to find out the frequency of records in this dataset based on different criteria.

We found out correlation between random attributes to find out how strong their values are correlated with each other.

**REFERENCES**:

* <https://www.analyticsvidhya.com/>
* <https://www.r-bloggers.com/>
* <https://www.tutorialspoint.com/index.htm>
* Notes provided by SCTPL