CS544\_project

**Data Analysis of Iris dataset**

1. **Introduction –**

For the final project we will be analyzing famous (Fisher's or Anderson's) iris data set which gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. We will analyze this data as per Module3. The species are *Iris setosa*, *versicolor*, and *virginica*.

The source of this dataset is **“datasets”** library of R. This data set has 150 cases (rows) and 5 variables (columns) named Sepal.Length, Sepal.Width, Petal.Length, Petal.Width, and Species. Where species is the categorical variable and the other variables are numerical.

1. **Overview of the preprocessing ,import and analysis steps**-

**Preprocessing and importing data**

Before importing the dataset we will check the following things.

* Accuracy of data- if the data entered is correct or not
* Consistency of data- check if similar type of data is kept in all the places or not.
* Completeness: Check if the data is available or not recorded
* Interpretability: The understandability of the data.
* If data comes from trustable source.

If we find some loop in the data quality after checking the above things then we will perform the suitable solution steps and if data passes the above checks then we will further proceed and import data.

To import dataset, we will use the “datasets” library of R software by following the below steps-(Note that we are using the inbuilt data from R , these steps are not in general and will vary if data is taken from some other source.)

* First we will download the respective package in our case it will be “datasets” package using the code “install.packages(datasets)” and then load the respective package using the code “library(package name)”e.g. library(datasets).
* After loading the package we will the load the iris data by the code “data(iris)” and for the glimpse of data structure we used “head(iris)”.

**Analysis-**

In the Analysis We will be performing univariate ,bivariate and multivariate analysis.

We will be performing categorical and numerical analysis since our data consist of these two types of variable. We will be generating the appropriate plots and summaries as per module 3 and will be interpreting the results.

We will be drawing the random samples of data and show the applicability of Central Limit Theorem and explain the various sampling methods.

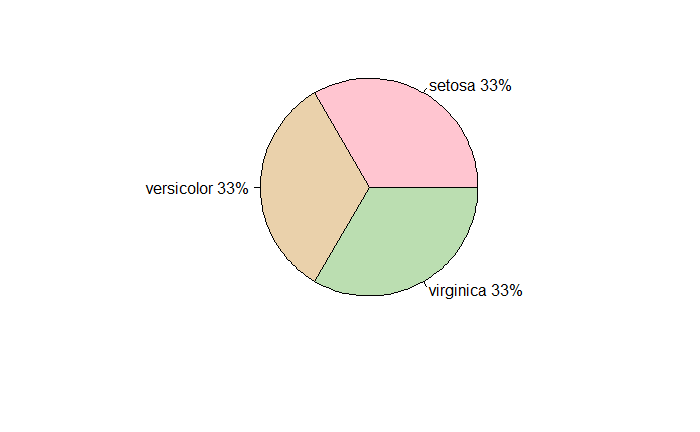
1. **Results-**

**Univariate analysis-**

Here we observed that species has 3 type and each type has been obtained 50 times from total 150 observations. and below is the frequency distribution table of species, which shows that each type of species has probability 0.33 to show up.

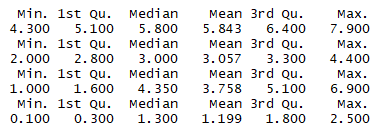


The below pie chart is showing the frequency distribution of the species



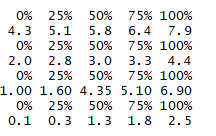
**Pie chart of Species**

* Averages of Sepal.Length, Sepal.Width, Petal.Length and Petal.Width are 5.843333, 3.057333, 3.758 and 1.199333 respectively.
* Medians of Sepal.Length, Sepal.Width, Petal.Length and Petal.Width are 5.8, 3, 4.35 and 1.3 respectively.
* Ranges of Sepal.Length, Sepal.Width, Petal.Length and Petal.Width are (4.3, 7.9),(2.0 ,4.4),(1.0 ,6.9) and (0.1 ,2.5) respectively.
* Differences in the ranges of Sepal.Length, Sepal.Width, Petal.Length and Petal.Width are 3.6, 2.4, 5.9 and 2.4 respectively.
* Variances of Sepal.Length, Sepal.Width, Petal.Length and Petal.Width are 0.6856935, 0.1899794, 3.116278 and 0.5810063 respectively.
* Standard deviations of Sepal.Length, Sepal.Width, Petal.Length and Petal.Width are 0.8280661, 0.4358663, 1.765298 and 0.7622377.
* Summaries of Sepal.Length, Sepal.Width, Petal.Length and Petal.Width are below respectively-



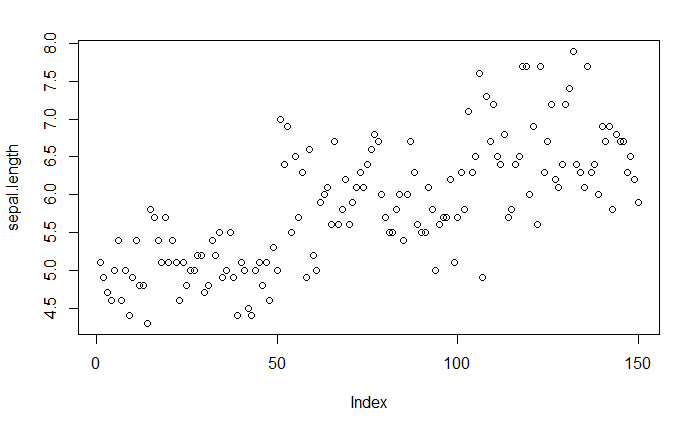
**Descriptive Summary of Sepal.Length, Sepal.Width, Petal.Length and Petal.Width**

* Quantiles of Sepal.Length, Sepal.Width, Petal.Length and Petal.Width are below respectively-



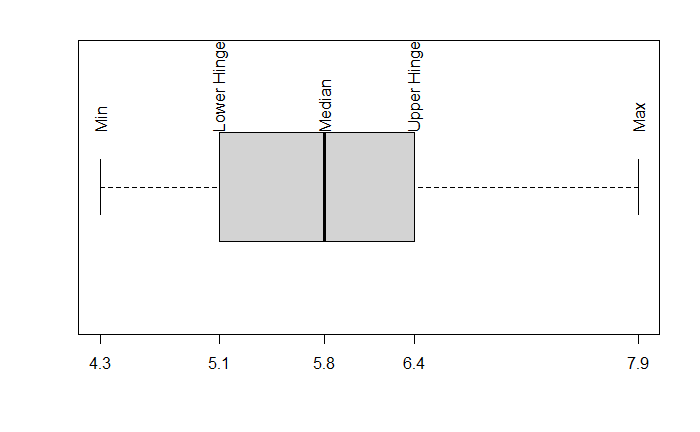
**Quartiles of Sepal.Length, Sepal.Width, Petal.Length and Petal.Width**

* Inter Quartile ranges of Sepal.Length, Sepal.Width, Petal.Length and Petal.Width are 1.3, 0.5, 3.5 and 1.5 respectively.
* From the below scatter plot of the Sepal. Length, we noticed that the observations are too scattered



**Scatter Plot of Sepal length**

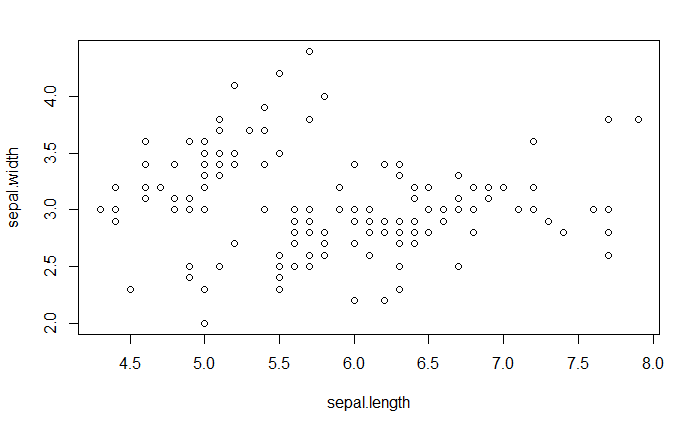
* And from the boxplot we observed there is no outlier present.



**Boxplot of Sepal Length**

**Bivariate Analysis-**

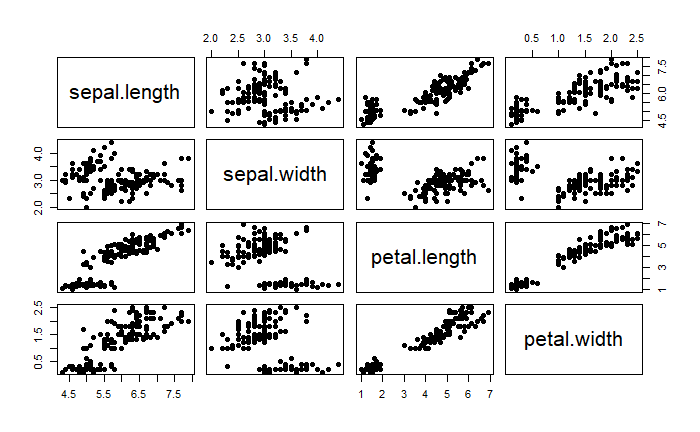
* From the above scatter plot between Sepal length and sepal width, we observed there is no visible pattern.



**Scatter plot Sepal length vs Sepal width**

**Multivariate Analysis-**

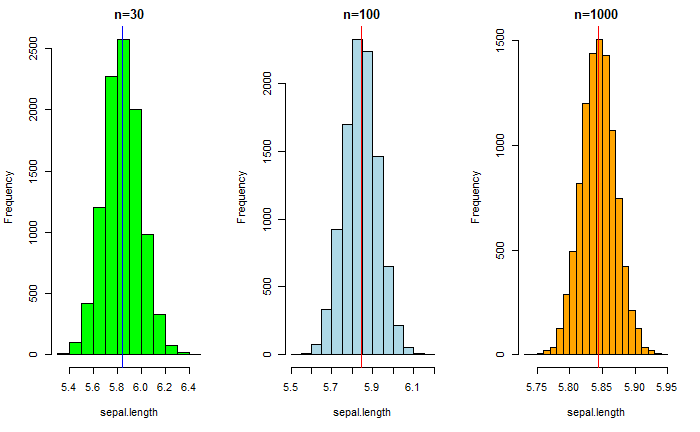
From the below pair plot of variables, we observed that sepal length and petal length, sepal length and petal width has increasing linear relation. Also petal length and petal width also has linear increasing relation.



**Pair Plots Between Sepal length, Sepal width, Petal length, Petal width**

**Applicability of Central limit theorem and various sampling methods-**

To show the applicability of central limit theorem, we considered many samples of sepal length with different sample sizes i.e. 30,100 and 1000 and plotted the histogram of means of these samples for each sizes respectively. And we can see the applicability of CLT in the above histograms.



**Histograms of Sepal length with different sample sizes to show applicability of CLT**

* We used simple random sampling with and without replacement, stratified sampling and systematic sampling on the data and generated the random samples. And observed when we used sampling with replacement, the two sample values are independent. In sampling without replacement, the two sample values weren’t independent. Samples from Stratified random sampling are a more precise since it's a better representation of the overall population.
* But in our case we used sepal length variable to compare all the method and took the average for all the sampling methods we used and for systematic sampling method with equal probability average sepal length was the closest to the population average sepal length.
* The best sampling method can vary based on the different situations and the kind of problem.