Date __/__/____

Assignment-C1

* **Title:-** Analysis on Iris Flower dataset.

* **Problem statement:-** Download the iris flower dataset or any other dataset into a dataframe. Use Python / R and perform following:

1) How many features are these and what are their types?
2) Compare and display summary statistics for each feature available in dataset. (E.g.: Min, Max, mean, std dev, variance, percentile).
3) Data visualization - Create a Histogram for each feature in the dataset to illustrate feature distribution.
4) Create a box plot for each feature in dataset. All of the box plots should be combined into a single plot.
   Compare distribution & Find outliers.

* **Objectives:-**
  - To learn the concept and terminologies in data analytics.
  - To learn how to display summary statistics and charts for each feature.

* **Outcomes:-**

  Students will be able to -
  - learn concepts in data analytics.
  - learn how to summarize plot charts.

* **S/W & H/W Requirements:**
  - OS: Window 10/ Ubuntu.
  - Python (scify libraries) / R studio with R libraries.

Date __/__/____

Saathi Notebooks ②

* Theory:-

A) IRIS Flower dataset:-

- The dataset is a multiset variate dataset introduced by the British statistician and biochemist Ronald Fisher in 1936.
- Dataset consists of 50 samples from each of 3 species of Iris, which are sentosa, virginca and versicolor.
- Four features measured from each sample are length and width of sepals and petals in mm.

B) Summary statistics.

1) mean: It identifies the average value of set of values.

$$\bar{x} = \frac{\sum_{i}^{n} x_i}{n}$$ where x_i = Value of attribute.
n = Total no. of items.

2) Range :- It shows the mathematical model between the lowest and highest values in the dataset. It measures the variability of dataset.
Range = max - min.

3) Standard Deviation :- It measures the variability of dataset like range. The smaller standard deviation indicates less variability.

$$\sigma = \sqrt{\frac{\sum_{i}^{n} (x_i - \bar{x})^2}{n}}$$

4) Variance : It measures how far the data is spread out:

Page No.

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}.$$

c) Applications:-

1) Histogram:-

- It is suitable for visualizing distribution of numeric data over a continuous interval or a certain time period.
- The histogram organises large amount of data, and provides a visualization quickly, using a single dimension.

2) Box Plot:-

- It allows quick graphical examination of one or more datasets
- It may seem primitive than a histogram but they do have some advantages.
- They take up space and are particularly useful for comparing distributions between several groups of data.

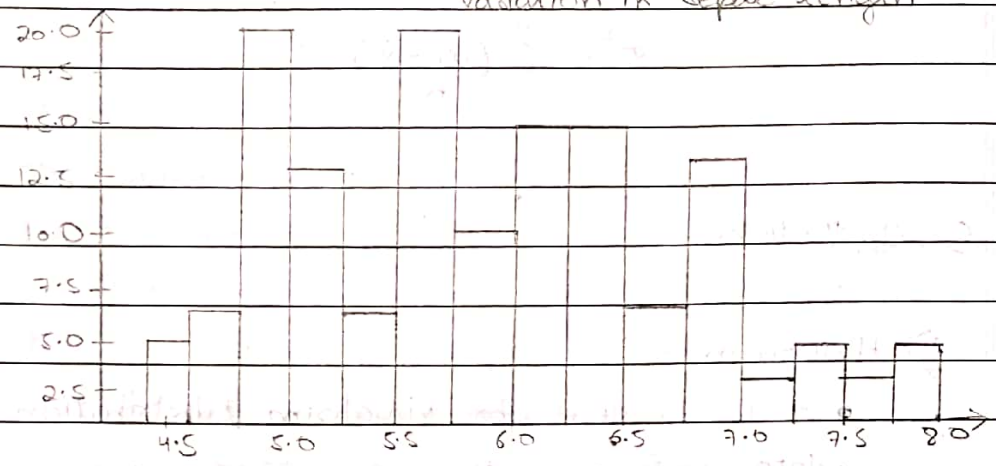
3) Data Visualization:

- It quickly creates insightful data visuals.
- They allow anyone to organize and present information quickly.

* Test-Case:-

Input	Output
Column of Sepal length	mean = 5.843.

Variation in sepal length.



* Conclusion:- We studied about concepts in data analytics, and the data-set. We also presented the data in charts and box plots.