Camlin

Assignment

- 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 the foll:-

(1) How many features & their type?

(2) Compare & display summary statistics for each feature available in dataset (eg: min, max, mean, std-dev, variance ...)

(3) Data visualization - create a histogram for each feature in data set 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 distributions & find outliers.

- Objective:

- To learn the concept & terminologies in data analytics

- To learn how to display summary statistics & charts for each feature.

- Outcome:

We will be able to

- learn the concepts in data analytics
- learn how to summarize & plot charts.

- Requirements:
  - OS: Windows 10 / Ubuntu (64bit)
  - Python (SciPy libraries) / Rstudio = R lib.

- Theory:

A) Iris flower dataset:-
- The dataset is a multivalue dataset introduced by British satisfac-tion & biochemist Ronald Fisher 1936
- Dataset consists of 50 samples from each of 3 species of Iris which are Sentosa virginica & versicolor
- Four features measured from each sample are length & width of sepals & 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 & 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

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

(.) Applications:

1.) Histogram:

- It is suitable for visualizing distribution of numeric data over a continuous interval or certain time period.

- It organizes large amnt of data and provides visualization quality 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 adv.
- They take up space & are particularly useful for comparing distributions between several groups of data

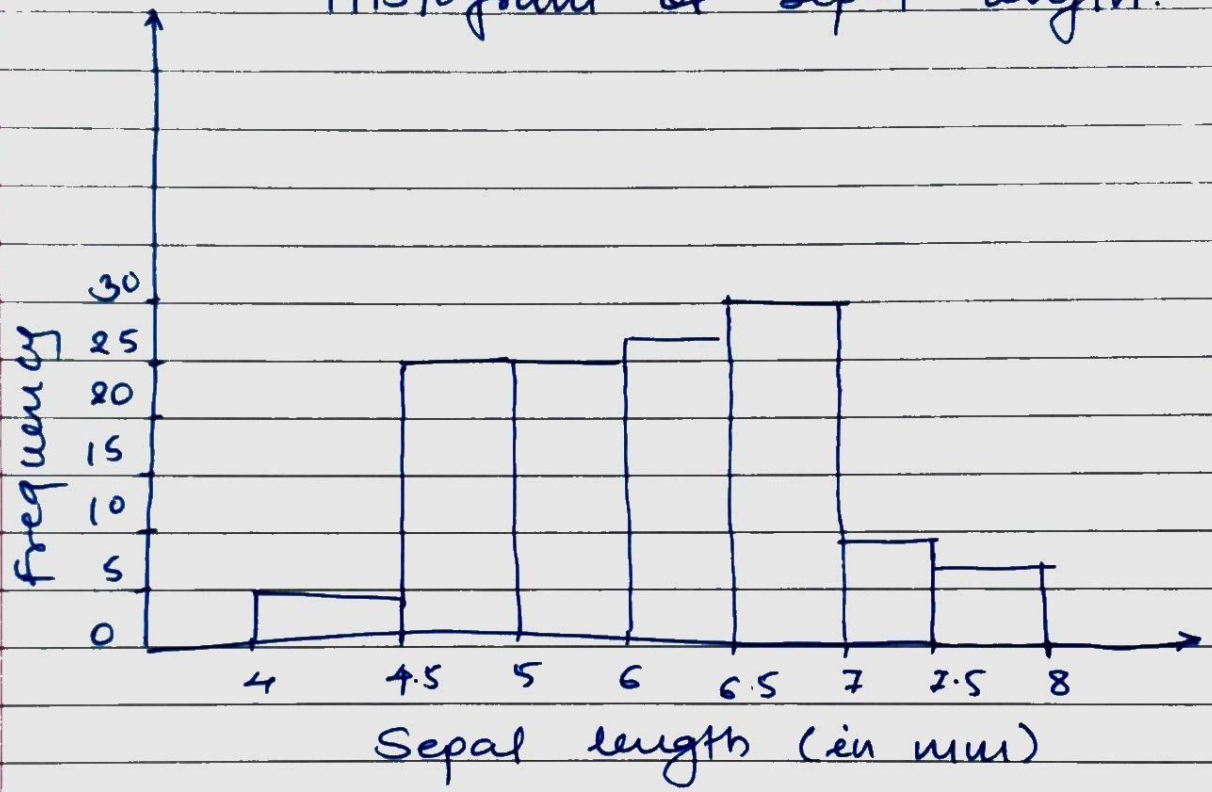
3.) Data visualization:-
- It quickly creates insightful dat visuals.
- They allow anyone to organize and present information quickly.

- Test cases

Histogram of sepal length.

where
input → column of sepal length
and
output → mean = 5.843 mm.

Histogram of sepal length.



- Conclusion:

Thus, we studied about concepts in data analytics and the dataset We also presented the data in charts and box plots.