Date:-

Assignment No 01

Title:-

Download the Iris flower dataset or any other dataset into a DataFrame.
Use Python/R and perform following:-

- How many features are there and what are their types (e.g. numeric, nominal)?
- Compute and display summary statistics for each feature available in the dataset, (e.g. minimum value, maximum value, mean, range, standard deviation, variance and percentiles.)
- Data visualization - Create a histogram for each feature in the dataset to illustrate the feature distribution. Plot each histogram
- Create a boxplot for each feature in the dataset. All of the boxplots should be combined into a single plot. Compare distributions and identify outliers.

Problem Statement:-

Implement a dataset into a dataframe.
Implement the following operations.

1) Display data set details.
2) Calculate min, max, mean, range, standard deviation, variance.
3) Create boxplot and histogram.

Software and Hardware requirements:-

- 64-bit OS Linux.
- Python
- Google Colab.

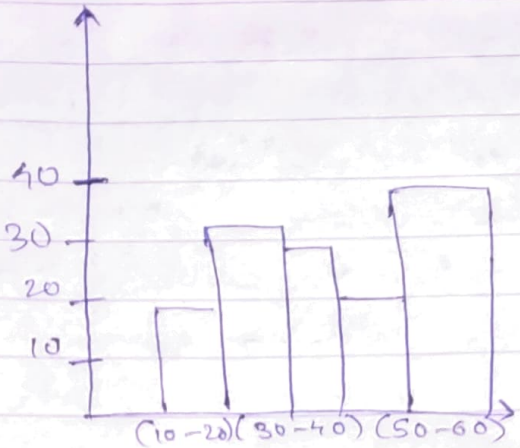
Learning Objectives:-

To learn the concept of how to display statistics for each feature Available in the dataset.

Theory:

Data Visualization:-

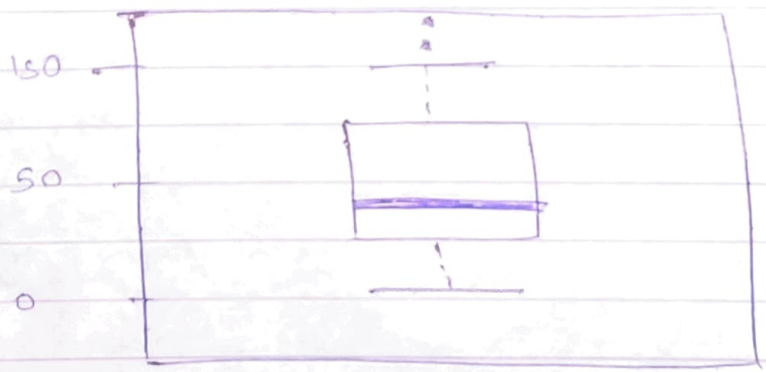
1) Histogram: Vertical bar chart is used to draw a histogram: which represents the distribution of a set of data over a continuous interval or certain time period and relationships of a single variable over set of classes.
While representing the tabulated data into an histogram, the tabulated frequency at every interval / bin / instance is represented by every bar in a histogram. And the total area of a histogram is equal to the number of data. it is one of the most commonly used graphical representation of data.



Histogram example

2) Boxplot :- a graphical summary of distributions

- The box in the middle indicates hinges (close to the first and third quantities) and median.
- The lines show largest and smallest observations that falls within the distance.
- A boxplot can often more useful to compare distributions side by side as it more compact than a histogram.



boxplot example.

- Boxplots are used to calculate quick summaries for all the variables in our set by default.

About the Iris dataset:

The IRIS dataset is a multivariate dataset introduced by British statistician, eugencist and biologists Ronald Fisher in a paper he wrote in 1936

The dataset contains 50 samples from each of the three species of Iris (Iris setosa, Iris virginica & Iris versicolor). Four features were measured for each sample: the length & width of the sepals and petals in centimeters.

Fisher developed a linear discriminant model to distinguish the species from each other.

This dataset is thus very useful for statistical classification techniques in machine learning as well as good starter dataset.

Conclusion:

The python commands for basic statiscal techniques were understand and data visualization was performed on the results.