Assignment No:

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 there 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 the dataset. All of the box plots should be combined into a single plot. Compare distributions and find outlines

Learning Objectives:

- To learn the concepts and terminologies in datasets
- Learn how to summanize and plot charts.

Learning Outcomes:

- To learn the concepts and terminologies in data analysis.
- To learn how to display summary statistics and charts for each feature.

Requirements:

OS: Windows 10 / Fedora 20
Python (Scipy libraries)
Google colbab.

Theory:

Iris flower data set:

The dataset is a multi variate dataset introduced by Rohald Fisher in 1936.

It consists of 50 samples from each of 3 species of Iris, which are Sentosa, Virginia, and vesicolor.

Four features measured from each sample are length and width of sepals and petals in mm

Summary Statistics:

1. Mean: Identifies the average value of set of values.

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

where x_i = value of its attribute
n : total no. of attributes

2. Range: It measures the variability of a dataset in terms of distance between highest and lowest values.

range = max - min.

3. Standard Deviation: It also measures the variability of data set.

$$\sigma = \sqrt{\sum_{i=1}^{n} \frac{(x_1 - \bar{x})^2}{n}}$$

4. Variance: Measures how for the data is spread out.

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

Data Visualization:

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

Histogram:

- A vertical bar chart is used to draw a histogram which represents the distribution of a set of data over a continous interval or certain time period and relationships of a single variable over set of classes.

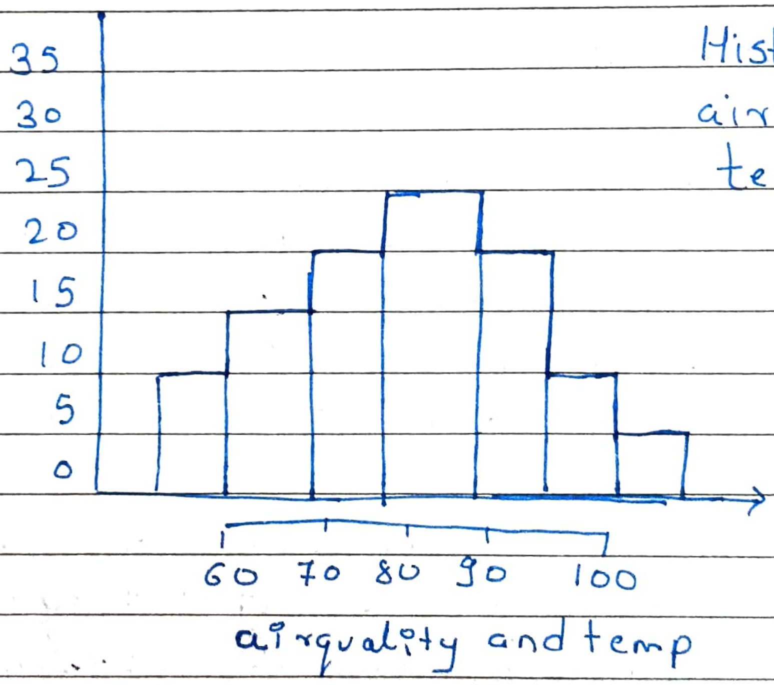
- While representating the tabulated data into histogram, the tabulated frequency at every interval / bin / instance is represented by every bar in a histogram and the total area of a histogram is equal to the number of data

- The one of the most commonly used graphical presentation of dat is histogram.

- Histogram organizes and displays the table data in user-friendly format.

- Histogram is used to graphically represent the huge amount of area/measurements/ dimensions contained by table.

- That means the histogram constructed to visualize the data will make that data easy to understand by representing the number.



Histogram of air quality and temp.

Box plots:

A box plots or box or whisker plot is a graphical summary of a distributions.

- The box in the middle indicates hinges (close first and third quantites) and median.

- The lines show the largest and smallest observations that falls within the distance.

- A box plot can often give a good idea of the data distribution and is often more useful to compare distributions side by side as it is more compact than a histogram.

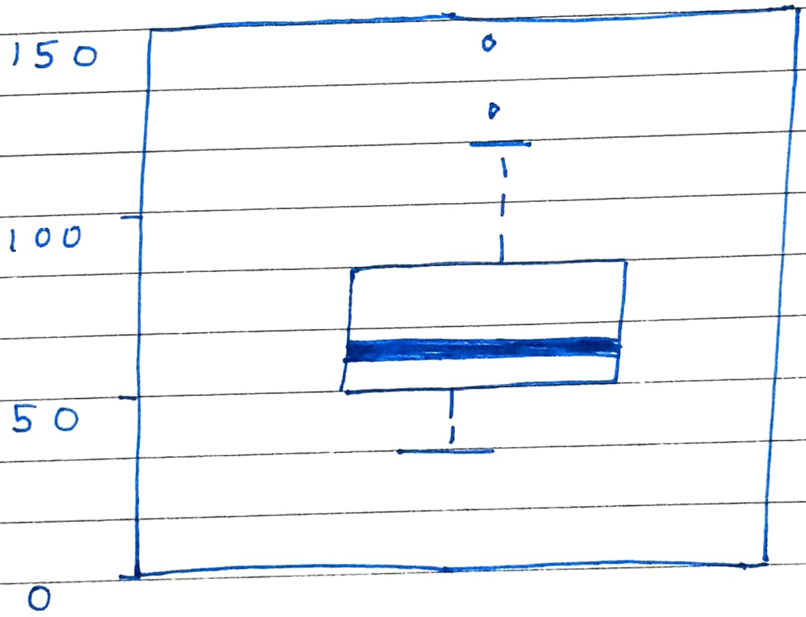


Fig: 1

- Thus, use of blox plot function to calculate quick summaries for all the variables in our set by default.

Conclusion:

Thus we studied about the concepts of data analysis and also visualized the Iris data set using histograms and box plots.