ASSIGNMENT DA-1

Roll No: 41205

Problem Statement:

Download the Iris flower dataset or any other dataset into a DataFrame. (eg https://archive.ics.uci.edu/ml/datasets/Iris) 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. (eg. 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 distributions. 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 outlier.

Objective:

1. To under basics of data visualization
2. Understand basic data analysis

Outcome: One will be able to load, analyze and visualize datasets using Python and its libraries.

Pre-requisites:

1. 64-bit Linux OS
2. Programming Languages: Python

Hardware Specification:

1. x86_64 bit
2. 2/4 GB DDR RAM
3. 80 - 500 GB SATA HD
4. 1GB NIDIA TITAN X Graphics Card

Software Specification:

1. Ubuntu 14.04

Theory:

- The Iris dataset was used in R.A. Fisher's classic 1936 paper, The Use of Multiple Measurements in Taxonomic Problems, and can also be found on the UCI Machine Learning Repository.

- It includes three iris species with 50 samples each as well as some properties about each flower. One flower species is linearly separable from the other two, but the other two are not linearly separable from each other.

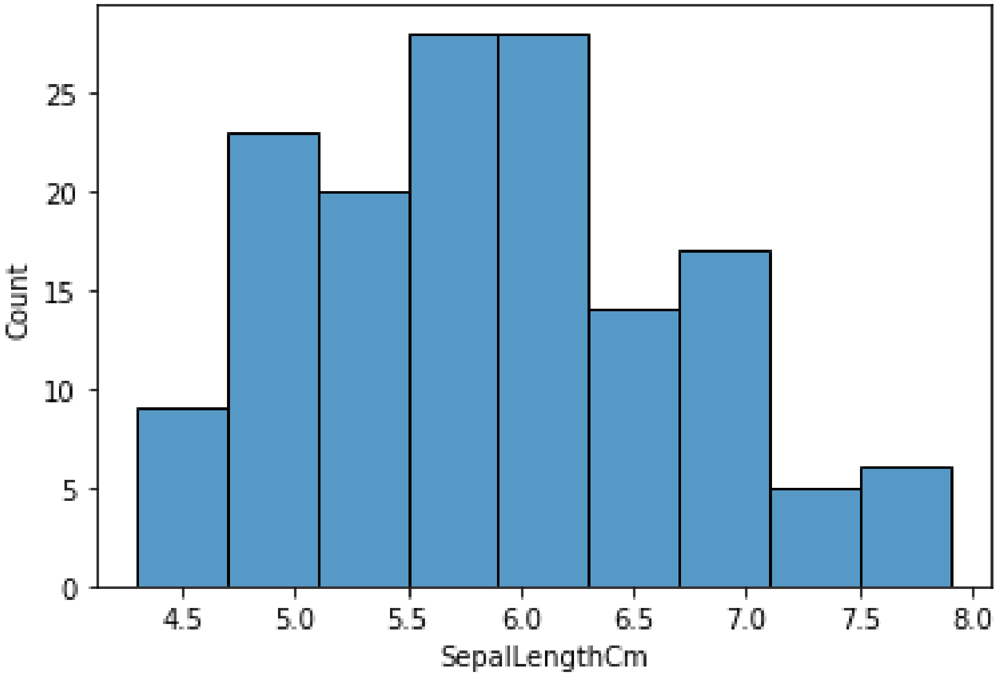
- The columns in this dataset are:
  1. Id
  2. SepalLengthCm
  3. SepalWidthCm
  4. PetalLengthCm
  5. PetalWidthCm
  6. Species

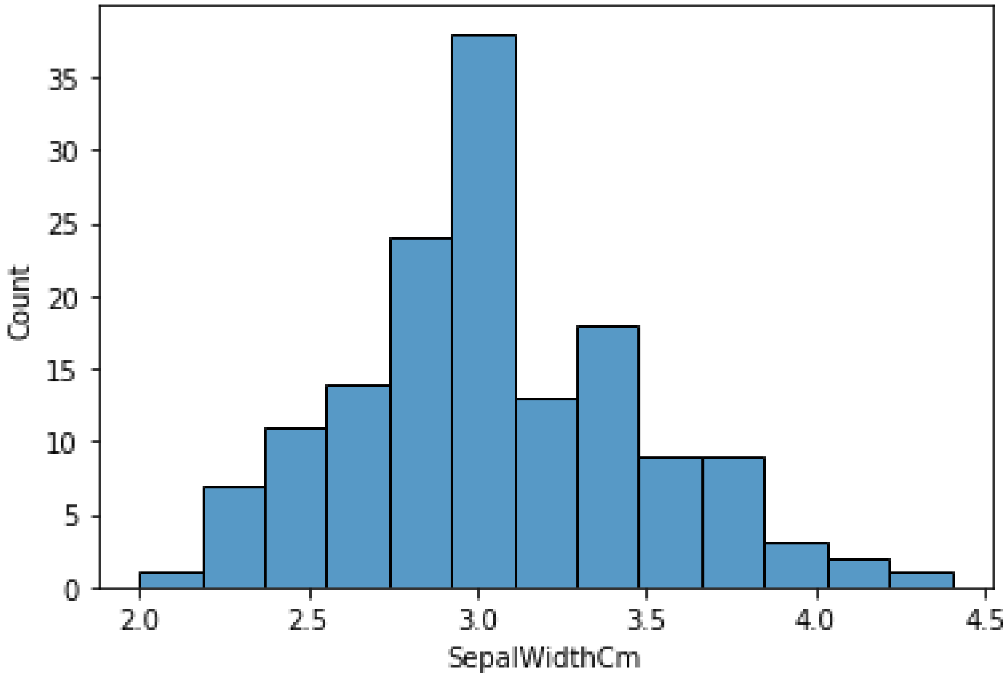
- Data can be analyzed using the pandas library of Python.

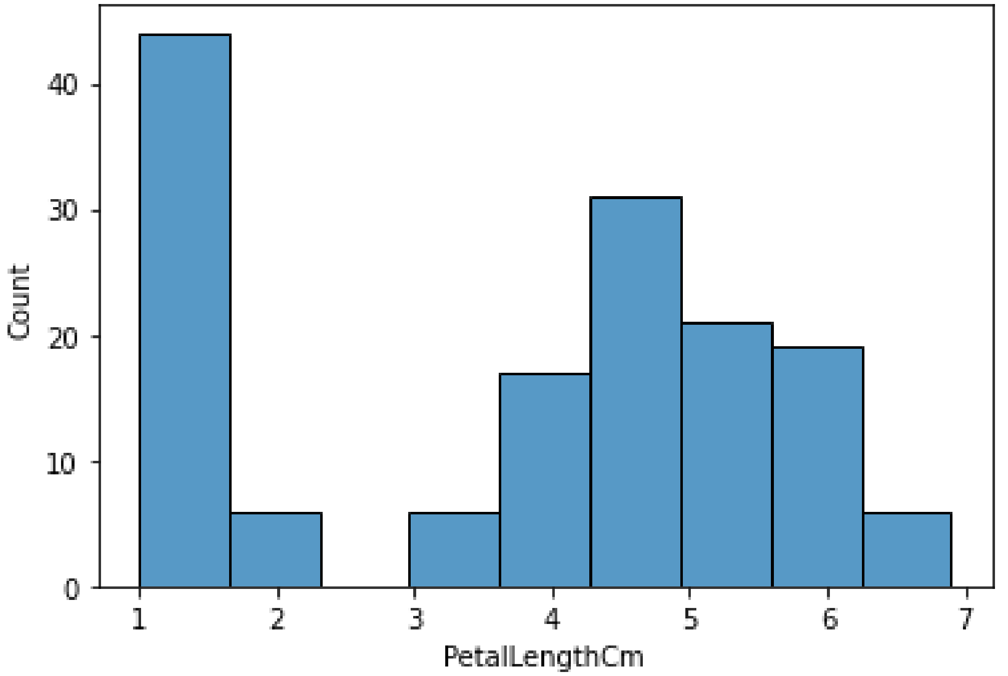
- The data can be brought into dataframes and displayed.

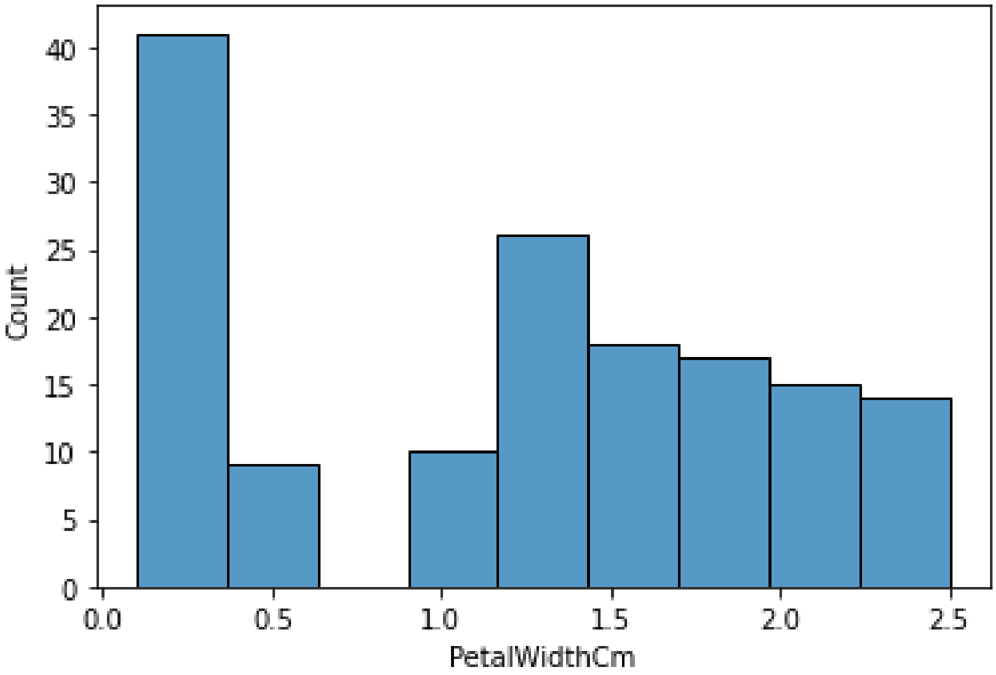
- Visualization can be done using seaborn and matplotlib libraries.

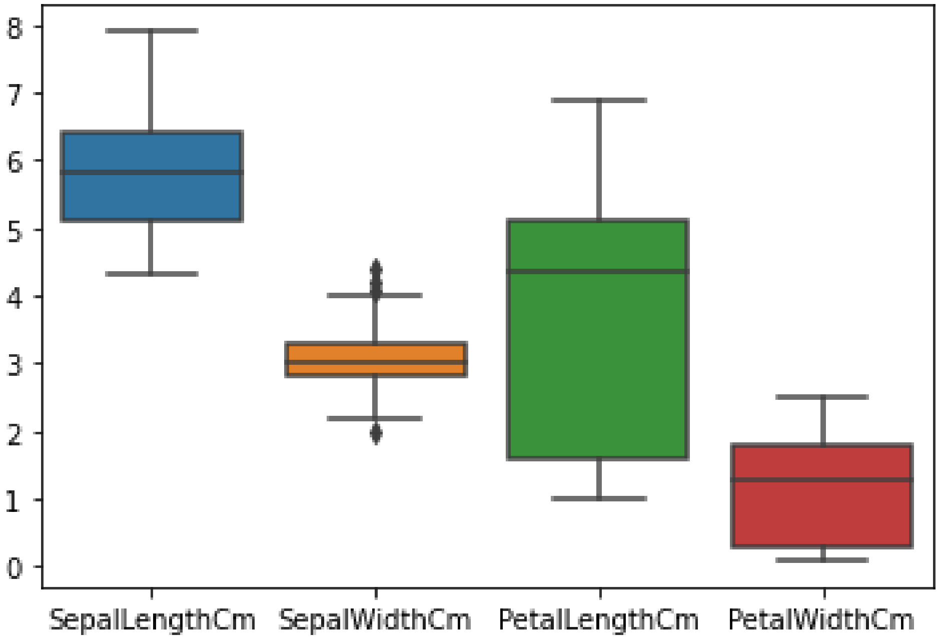
Output:




Count
0
5
10
15
20
25
30
35
2.0
2.5
3.0
3.5
4.0
4.5
SepalWidthCm


Count
0
10
20
30
40
1
2
3
4
5
6
7
PetalLengthCm





Conclusion: Thus, we have analyzed the iris dataset using Python and plotted various charts to visualize the data.