LP1 Assignment DA1

Summary statistics, data visualization and boxplot for the features on the Iris dataset

Date - 9th August, 2020.

Assignment Number - DA1

Title

Summary statistics, data visualization and boxplot for the features on the Iris dataset

Problem Definition

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

Use Python 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 outliers.

Learning Objectives

● Learn to use dataset, dataframes, features of dataset in an application

● Learn to compute summary statistics for the features.

● Learn to use visualization techniques. Learning Outcomes I will be able to compute statistics on the features of the dataset, use histograms and boxplot on the features of the dataset.

Software Packages and Hardware Apparatus Used

● Operating System: 64-bit Ubuntu 18.04

● Programming Language: Python 3

● Jupyter Notebook Environment: Google Colaboratory

● Python Libraries: Numpy, Pandas, MatPlotLib

Related Mathematics

Mathematical Model Let S be the system set:

S = {s; e;X; Y; Fme; Ff;DD;NDD; Fc; Sc}

where Dataset is loaded into the dataframe

s=start state

Iris Dataset

e=end state

Summary statistics for each feature is computed.

X=set of inputs

X = {X1} Where X1 = IRIS Dataset

- 5 Features
  - 4 Numerical Feature
  - 1 Nominal Feature
- Data Count - 150

Y=set of outputs

1) Number of features
2) Types of features
3) Minimum value for each feature in the dataset
4) Maximum value for each feature in the dataset
5) Mean for each feature in the dataset
6) Range for each feature in the dataset
7) Standard deviation for each feature in the dataset
8) Variance for each feature in the dataset
9) Percentiles for each feature in the dataset
10) Histogram for each feature in the dataset
11) Boxplot for each feature in the dataset

Fme is the main function

It calls friend functions

Ff is the set of friend functions

Ff = {f1,f2,f3,f4,f5,f6}

Where,

f1 = function to load dataset into data frame

f2 = function to get number of features

f3 = function to get feature type

f4 = function to get minimum, maximum, mean, range, standard deviation, variance and percentile for each feature

f5 = function to draw histogram for each feature

f6 = function to draw boxplot for each feature

DD= Deterministic Data IRIS dataset

- 5 Features
  - 4 Numerical Feature
  - 1 Nominal Feature
- Data Count - 150

NDD=Non-deterministic data

No non deterministic data

Fc =failure case:

No failure case identified for this application

Concepts related Theory

Data analysis is a process of inspecting, cleansing, transforming, and modelling data with the goal of discovering useful information, informing conclusions, and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, while being used in different business, science, and social science domains.

A data set (or dataset) is a collection of data. Most commonly a data set corresponds to the contents of a single database table, or a single statistical data matrix, where every column of the table represents a particular variable, and each row corresponds to a given member of the data set in question.

Mean, standard deviation, regression, sample size determination and hypothesis testing are the fundamental data analytics methods

Mean

The sum of all the data entries divided by the number of entries.

Range

The difference between the maximum and minimum data entries in the set.

Range = (Max. data entry) – (Min. data entry)

Standard Deviation

The standard deviation measures variability and consistency of the sample or population. In most real-world applications, consistency is a great advantage.

$$\text{Mean} = \frac{\sum fx}{\sum f}$$

$$\text{Standard deviation} = \sqrt{\frac{\sum fx^2}{\sum f} - \left(\frac{\sum fx}{\sum f}\right)^2}$$

Variance

Variance is the average squared deviation from the mean.

Percentile

Let p be any integer between 0 and 100. The pth percentile of the data set is the data value at which p percent of the value in the data set is less than or equal to this value.

Steps for Execution

1. Download Iris Dataset
2. Open Google Colaboratory
3. Upload Iris Dataset to Google Colaboratory
4. Import Python Packages like Numpy, Pandas, MatPlotLib

5. Get features from the Dataset into Pandas Dataframe

6. Give Feature Names to Pandas Dataframe

7. Get Feature Count and Type of Feature

8. Compute Statistics in the Problem Definition

9. Generate Histograms and Boxplots for features of the dataset

Useful Python Functions

Pandas Functions

read_csv - Read Data from Iris Dataset downloaded and uploaded to Google Colab

shape - Get (number of samples,number of features)

iteritems - Iterate through features/columns of the dataset

describe - Get Useful Statistics on the dataset like mean, max value, min value, standard deviation and percentiles

var - Get Variance of features of the data set

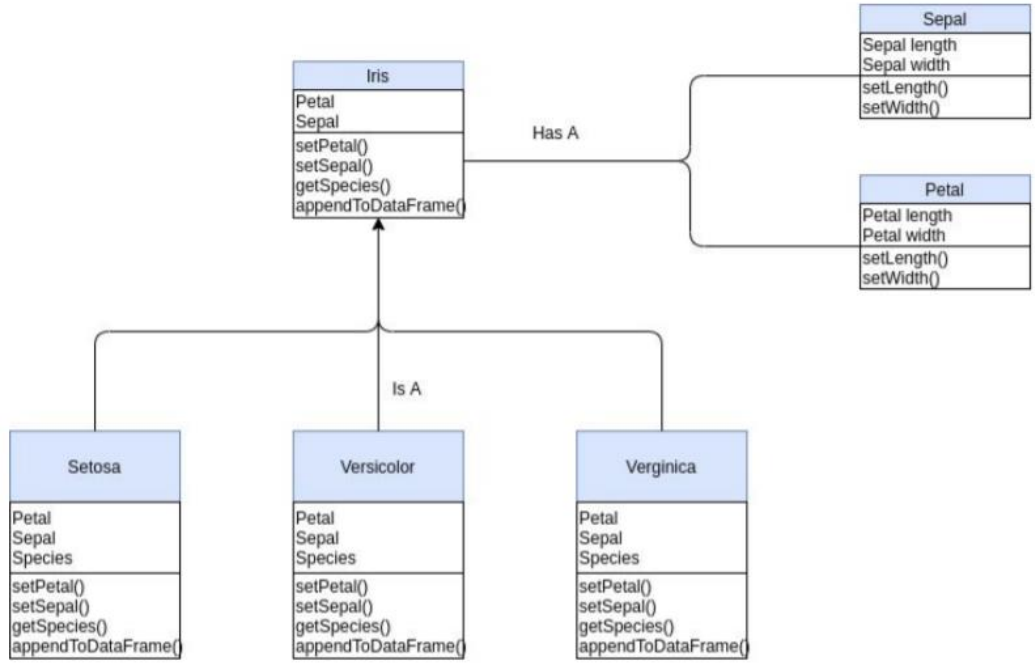
hist - Plot Histogram of features of the dataset

max - Get the max value from all the features

min - Get the min value from all the features

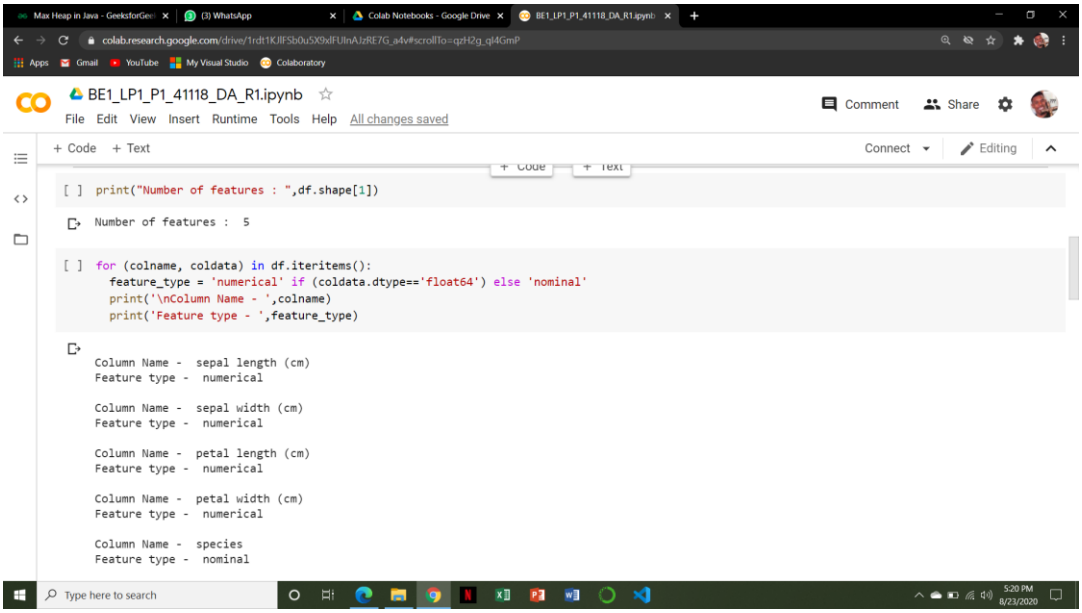
plot - Plot types of graph including box plot

Flowchart Design

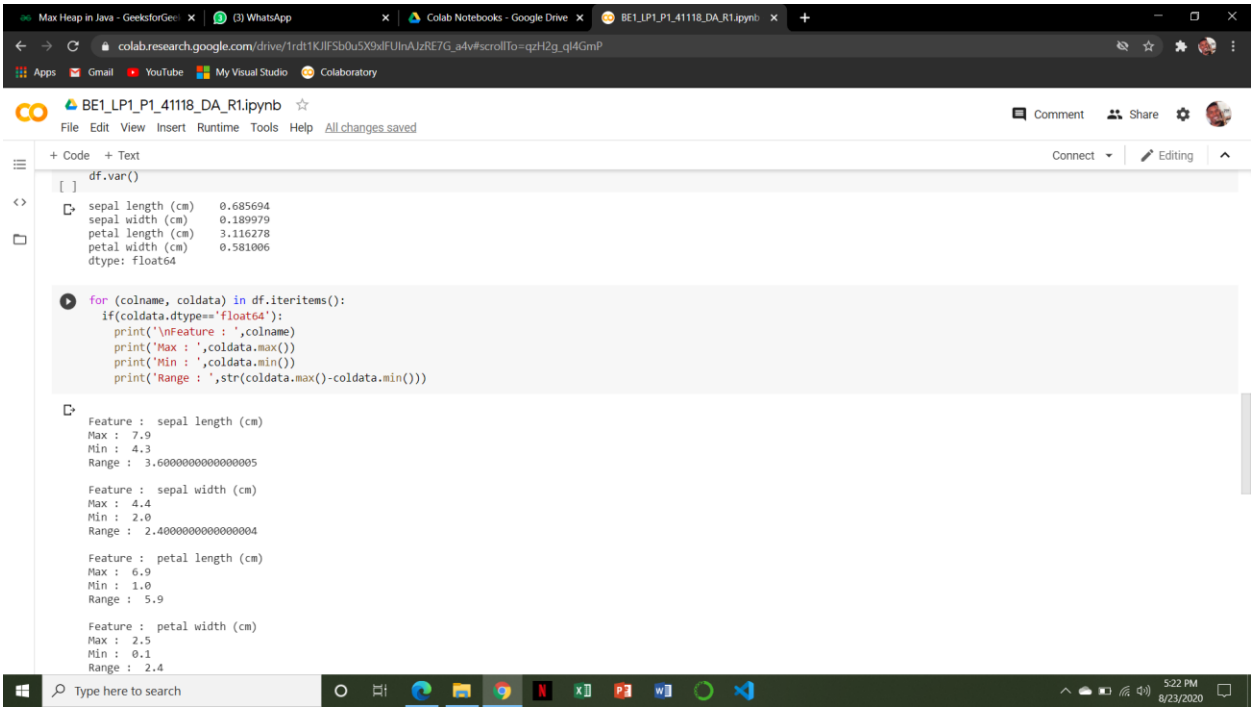


Output Screenshots –

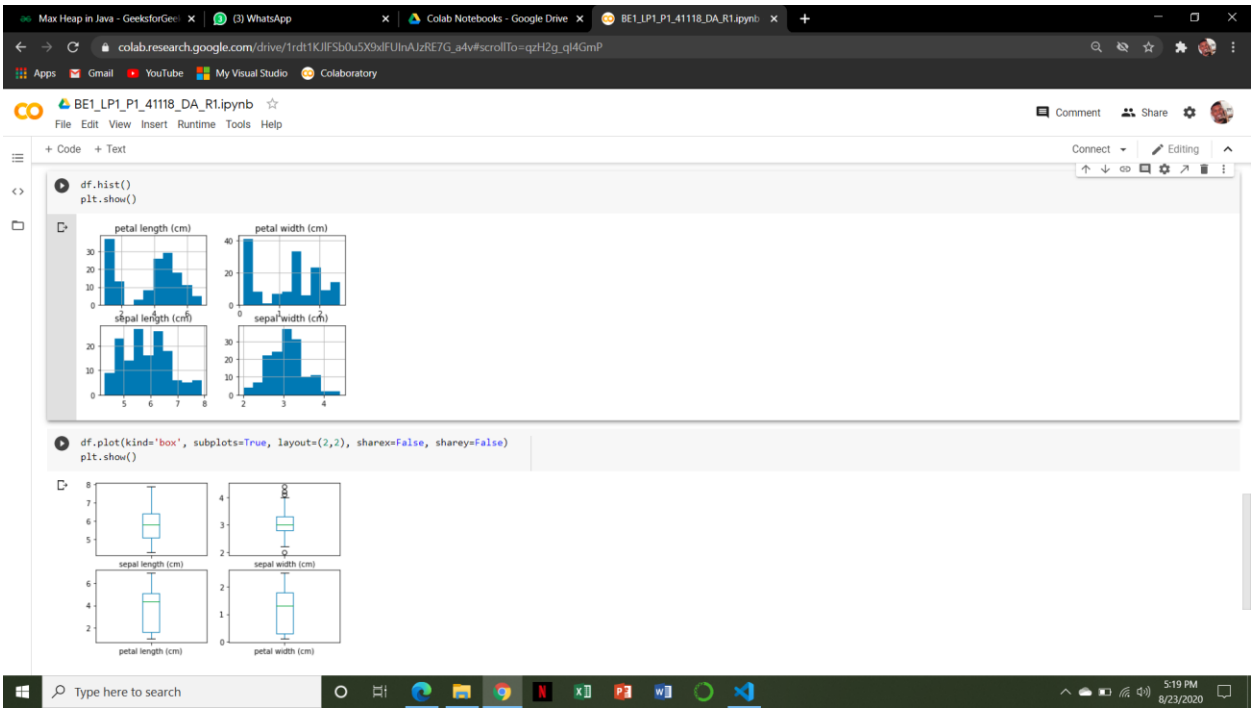
Number of features, Feature Types and Statistics



Variance and Range



Histograms and Boxplots



Notebook Link –

https://colab.research.google.com/drive/1rdt1KJlFSb0u5X9xlFUInAJzRE7G_a4v?usp=sharing

Conclusion –

We have successfully computed statistics on the features of the Iris dataset, and used histograms and boxplot on the features of the dataset.