Assignment : C1

- Problem Statement : Download the iris flower dataset into a Dataframe using Pyhon /.R and perform following operations :
  - How many features are there and what are there types (eg. numeric, nominal)?
  - Compute and display summary statistics for each feature available in the dataset (eg. Minimum value, maximum value, mean, rang, standard deviation)
  - 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 distribution and identify outliers.

- Objective :
  - To learn the concept and terminologies in data analytics.
  - To learn how to display summary statistics and data visualization.

- Outcomes : we will be able to :-
  - Learn concepts of Data Analytics.
  - Learn the concepts of statistics and data visualization.

- Software/ Hardware Requirements : OS8 windows |ubuntu distribution. Python libraries, Python framework, R studio, Julyter notebook, Anaconda novigator .

- Theory : IRIS Dataset :

The dataset is Multivariate dataset introduced by the British statistician and biocherist Renold Finsher in 1936 .
Dataset consist of 50 sample from each of 3 species of Iris which are sentosa, virginica and versicolor. Four features were measured from each sample : The length and width of the sepal and petal in centimeters . Based on the combination of these four features , Fisher developed a linear discriminant model to distinguish the species from each other. The dataset contains a set of 150 records under five attributes - sepal length, sepal width, petal length, petal width and species.
The data is loaded in python as follows :

```
from sklearn.datasets import load-iris
    data = load-iris ()
    df = pd.DataFrame (data= data ['data'], columns=
                  data ['feature-names'])
```

To Display the features and mere types :

```
print (list (df.columns))
x= df.drop (['target'], axis=1)
x.dtypes.
```

```
y = df['target']
y.Value-Counts()
```

- Summary Statistics :

i] Mean : The avarage value of set of values.

$\bar{x} = \frac{\Sigma x_i}{n}$ x_i = value of attributes, n = total no. of items.

```
df["feature-name"].Mean()
```

ii] Range : The lowest and highest value in dataset.

range = Max - Min.

```
df["feature-name"].min() , df["feature-name"].Max()
```

iii] Standard Deviation : $\sigma = \sqrt{\frac{\Sigma (x_i - \bar{x})^2}{N}}$

iv] Variance : σ^2.

- Data Viscualization :

Histogram : It is suitable for visualization of numeric data over a continuous interval, or a certain time period. The histogram organize large arount of data and provides a visualization quickly using a single Dimension. Histgram is a graphical display of data using bars of different heights. It is similar to a Bar charts but a histogram groups number of ranges. The height of each bar shows how many falls into each range.

```
df.hist.()
plt.Show()
```

2. Box plot : It allows quick graphical examination of one or More datasets . It May seems Primitive than a histogram but they do have some advantages . They are usefull for comparing distribution between groups of data.

Box Plot using Jupyter Notebook:

combined Box plot : x.boxplot()

seperate for each feature : sns. boxplot (x=df ["target"] ,y=df ["feature_name"])

- Test cases :

	Description	Expected output	Result
i]	Feature name	sepal-length, sepal-width, petal-length, petal-width, dtype: object target, dtype: int64	Pass
ii]	Summary statistics	count : 150 (all features) sepal-length : Mean : 5.8 Min : 4.3, Max : 7.9, std: 0.82 sepal-width : Mean : 3.05, Min: 2.0 Max : 4.4, Std : 0.43, Petal-length : Mean : 3.75, Min: 1.0 Max: 6.9, Std : 1.76 Petal width : Mean : 1.19. Min : 0.10, Max : 2.50 , Std : 0.76	Pass

Summary Statistics	target : Mean : 1.0 min : 0.0 , max : 2.0 Std : 0.81	Pass

- Conclusion : Thus we Successfully computed the given operation like Summary Statistics and Data Visualization on IRIS flower Dataset.