TITLE	**Summary statistics,data visualization and boxplot for the features on the Iris dataset or any other dataset.**
PROBLEM STATEMENT / DEFINITION	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. (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 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.
OBJECTIVE	• Learn to use dataset, dataframes, features of dataset in an application • Learn to compute summary statistics for the features. • Learn to use visualization techniques.
S/W PACKAGES AND HARDWARE APPARATUS USED	1. Operating System : 64-bit Open source Linux or its derivative 2. Programming Languages: PYTHON/R
FERENCES	• Mark Gardner, “Beginning R: The Statistical Programming Language”, Wrox Publication, ISBN: 978-1-118-16430-3 • David Dietrich, Barry Hiller, “Data Science and Big Data Analytics”, EMC education services, Wiley publications, 2012, ISBN0-07-120413-X • Luis Torgo, “Data Mining with R, Learning with Case Studies”, CRC Press, Talay and Francis Group, ISBN9781482234893
STEPS	Refer to theory, algorithm, test input, test output
INSTRUCTIONS FOR WRITING JOURNAL	1. Date 2. Assignment no. 3. Problem definition 4. Learning objective 5. Learning outcome 6. Related Mathematics 7. Concepts related Theory 8. Test cases 9. Program code with proper documentation.

	10. Output of program. 11. Conclusion and applications (the verification and testing of outcomes)

Assignment No. DA1

- **Aim:**

Summary statistics, data visualization and boxplot for the features on the Iris dataset or any other dataset.

- **Problem Statement / Definition:**
  - 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 outliers.

- **Prerequisites**

  Database management system, Python/R programming

- **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 Outcome:**

Students will be able to compute statistics on the features of the dataset, use histograms and boxplot on the features of the dataset.

- **Related Mathematics**

Mathematical Model

Let S be the system set:
S = {s; e;X; Y; Fme;DD;NDD; Fc; Sc} where Dataset is loaded into the dataframe
s=start state
e=end state i.e. Summary statistics for each feature is computed.
X=set of inputs
X = {X1}
where
X1 = IRIS or any other dataset
where ,
Y=set of outputs
1) Number of features and their types.
2) Summary statistics of the each feature (minimum value, maximum value, mean, range, standard deviation, variance and percentiles)
3) Data Visualization- histogram for each feature in the dataset , boxplot for each feature in the dataset
Fme is the set of main functions
Fme = {f1,f2,f3}
where
f1 = function to load dataset into dataframe
f2 = function to to get number of features
f3 = function to get feature type
f3 = function to get minimum,maximum,mean,range,standard deviation,variance and percentile for each feature
f4 = function to draw histogram for each feature
f5 = function to draw boxplot for each feature
DD= Deterministic Data
IRIS dataset
NDD=Non-deterministic data
No non deterministic data
Fc =failure case:
No failure case identified for this application

- **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.

$$\text{Population Mean: } \mu = \frac{\Sigma x}{N}$$

$$\text{Sample Mean: } \overline{x} = \frac{\Sigma x}{n}$$

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 measure variability and consistency of the sample or population. In most real-world applications, consistency is a great advantage. In statistical data analysis, less variation is often better.

$$\text{Population Standard Deviation} = \sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$$

$$\text{Sample Standard Deviation } = s = \sqrt{\frac{\sum(x-\overline{x})^2}{n-1}}$$

Variance: The average squared deviation from the mean is also known as the variance.

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

- How to calculate percentiles: Use the following steps for calculating percentiles for small data sets.
- Step 1: Sort the data in ascending order (from smallest to largest)

$\left(\frac{p}{100}\right)n,$ • Step Step 3: 2: Calculate ith = the 100 where p is the percentile and n is the sample size.

Step 3: If i is an integer the pth percentile is the mean of the data values in position i and i+1.If i is not an integer then round up to the next integer and use the value in this position.

R commands:

- R command to load dataset from an URL.

```
url<- "https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data"
filename<-"./iris.csv"
        download.file(url=url, destfile = filename, method ="curl")
```

- To get number of rows in the dataset:

  ```
  nrow(dataset)
  ```

- To get number of features in the dataset:

  ```
  ncol(dataset)
  ```

- To get minimam in the column: min(dataset$column_name)
- To get maximam in the column: max(data$column_name)
- To get mean in the column:colMeans(x=dataset, na.rm = TRUE)
- To get range in the column: range(as.data.frame(dataset[,col], drop=false))
- To get standard deviation and variance in the dataset:

  ```
  apply(dataset, 2, sd)
  apply(dataset,2,var)
  ```

- **Test data:**

  Iris data from https://archive.ics.uci.edu/ml/datasets/Iris dataset.