Assignment 1

1 Title

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

2 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.

3 Problem Statement

Download the Iris flower dataset or any other dataset into a DataFrame.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 his- togram.
- 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.

4 Outcome

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

5 Software and Hardware requirements

1. Operating System : 64-bit Linux or its derivative
2. Programming Language: Python/R

8 Theory- Concept in brief

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 vari- ety 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 con- tents of a single database table, or a single statistical data matrix, where every column of the ta- ble represents a particular variable, and each row corresponds to a given member of the data set in question.

8.1 Important terms

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

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.

8.2 Mean

: After we find the sum the mean is easily calculated as

$$Mean = \frac{sum}{N} \tag{1}$$

8.3 Variance and Standard Deviation

The formula for variance is:

$$VAR = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 \tag{2}$$

The formula for standard deviation is basically just square root of variance:

$$STD\_DEV = \sqrt{VAR} \tag{3}$$

9 Algorithm design

x: iris dataset

How many features are there and what are their types:

```
x.dtypes
```

Compute and display summary statistics for each feature:

```
x.describe
```

Create a histogram for each feature:

```
plt.hist(x['feature'],bins=15)
plt.show()
```

Create a combined boxplot for each feature in the dataset:

```
x.boxplot()
```

10 Test cases

```
x.dtypes—-¿
sepal length (cm) float64
sepal width (cm) float64
petal length (cm) float64
petal width (cm) float64
dtype: object
```

11 Conclusion

We have successfully conducted the data visualization of the iris dataset and performed various operations on the dataset.