Assignment No : DA-1

Problem Definition :

Download the Iris flower dataset or any other dataset into a Data frame.

(e.g., [https://archive.ics.uci.edu/ml/datasets/Iris](https://archive.ics.uci.edu/ml/datasets/Iris%20) )

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 the boxplots should be combined into a single plot. Compare distributions and identify outliers.

Objectives :

1. To understand commands in Python/R.
2. To understand data visualization & exploratory data analysis.

Outcomes :

1. Understand data visualization and perform operations for min, max, mean, range, std. deviation, variance, and percentiles.
2. Understand and implement various plots (Histogram, Box Plots).

Hardware & Software Requirements :

Hardware Requirements : Personal Computer (PC)

Software Requirements :

1. 32/64 Bit Linux/Mac/Windows operating system
2. Python 2.X/3.X
3. IDE/Notebook (PyCharm/Google colab/Jupyter Notebook etc.)

Concept Related Theory :

* Feature :

 A feature is an individual measurable property or characteristic of a dataset.

* What is data visualization ?

Data visualization is the graphical representation of information and data. By using [visual elements like charts, graphs, and maps](https://www.tableau.com/data-insights/reference-library/visual-analytics), data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

Plots used in data visualization :

1. Histogram :

A histogram is a graphical representation that organizes a group of data points into user-specified ranges. Similar in appearance to a [bar graph](https://www.investopedia.com/terms/b/bar-graph.asp), the histogram condenses a data series into an easily interpreted visual by taking many data points and grouping them into logical ranges or bins.

* A histogram is a bar graph-like representation of data that buckets a range of outcomes into columns along the x-axis.
* The y-axis represents the number count or percentage of occurrences in the data for each column and can be used to visualize data distributions.

1. Box Plots :

* In [statistics](https://en.wikipedia.org/wiki/Descriptive_statistics), a box plot or boxplot is a method for graphically depicting groups of numerical data through their [quartiles](https://en.wikipedia.org/wiki/Quartile).
* Box plots may also have lines extending from the boxes (*whiskers*) indicating variability outside the upper and lower quartiles, hence the terms box-and-whisker plot and box-and-whisker diagram.

A boxplot is a standardized way of displaying the dataset based on a [five-number summary](https://en.wikipedia.org/wiki/Five-number_summary): the minimum, the maximum, the sample median, and the first and third quartiles.

* [Minimum](https://en.wikipedia.org/wiki/Sample_minimum) (Q0 or 0th [percentile](https://en.wikipedia.org/wiki/Percentile)): the lowest data point excluding any outliers.
* [Maximum](https://en.wikipedia.org/wiki/Sample_maximum) (Q4 or 100th percentile): the largest data point excluding any outliers.
* [Median](https://en.wikipedia.org/wiki/Median) (Q2 or 50th percentile): the middle value of the dataset.
* [First quartile](https://en.wikipedia.org/wiki/First_quartile) (Q1 or 25th percentile): also known as the lower quartile qn(0.25), is the median of the lower half of the dataset.
* [Third quartile](https://en.wikipedia.org/wiki/Third_quartile) (Q3 or 75th percentile): also known as the upper quartile qn(0.75), is the median of the upper half of the dataset.

Libraries/Commands used :

Libraries : pandas, NumPy, Matplotlib, seaborn.

Imports :

* 1. import pandas as pd;
  2. from matplotlib import pyplot as plt;
  3. import seaborn as sns;

Commands :

* 1. Read the csv file : iris = pd.read\_csv('iris.data')
  2. Naming the columns : iris.columns =['sepal\_length', 'sepal\_width', 'petal\_length', 'petal\_width', 'class']
  3. List of features : list(iris.columns)
  4. Type of features : list(iris.dtypes)
  5. Plot Histogram : plt.hist()
  6. Plot Boxplot : sns.boxplot()

Conclusion :

Learnt about Histograms, Boxplots and commands used in Python to visualize and study the data.

Implemented the given problem statement and plotted the Histograms and Boxplot for each feature.