Name : Vikas Mane
Roll No: B1921152
PRN: 72000291F

Assignment1

Title: Download the Iris flower dataset or any other dataset into a Data Frame. (e.g., 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 andpercentiles
- Data Visualization-Create a histogram for each feature in the dataset to illustrate the feature distributions. Plot eachhistogram.
- Create a boxplot for each feature in the dataset. All of the boxplots should be combined into a single plot. Compare distributions and identifyoutliers.

Aim: Implement a dataset into a data frame. Implement the following operations:
1. Display data setdetails.
2. Calculate min, max, mean, range, standard deviation, variance.
3. Create histograph using histfunction.
4. Create boxplot using boxplotfunction.

System Requirements: Python, Data Visualization libraries, Jupyter Notebook.

Theory:

Data analytics (DA): Data analytics (DA) is the *process of examining data sets in order to draw conclusions about the information they contain, increasingly with the aid of specialized systems and software.* Data analytics technologies and techniques are widely used in commercial industries to enable organizations to make more-informed business decisions and by scientists and researchers to verify or disprove scientific models, theories and hypotheses.

Data Science: Dealing with unstructured and structured data, Data Science is a field that comprises of everything that related to *data cleansing, preparation, and analysis.*

Data Science is the combination of statistics, mathematics, programming, problem-solving, capturing data in ingenious ways, the ability to look at things differently, and the activity of cleansing, preparing and aligning the data.

In simple terms, it is the umbrella of techniques used when trying to extract insights and information from data.

Big Data: *Big Data refers to humongous volumes of data that cannot be processed effectively with the traditional applications that exist.* The processing of Big Data begins with the raw data that isn't aggregated and is most often impossible to store in the memory of a single computer.

A buzzword that is used to describe immense volumes of data, both unstructured and structured, Big Data inundates a business on a day-to-day basis. Big Data is something that can be used to analyze insights which can lead to better decisions and strategic business moves.

The definition of Big Data, given by Gartner is, “Big data is high-volume, and high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing

Data Analytics: Data Analytics the science of examining raw data with the purpose of drawing conclusions about that information.

Data Analytics involves applying an algorithmic or mechanical process to derive insights. For example, running through a number of data sets to look for meaningful correlations between each other.

It is used in a number of industries to allow the organizations and companies to make better decisions as well as verify and disprove existing theories or models.

The focus of Data Analytics lies in inference, which is the process of deriving conclusions that are solely based on what the researcher already knows.

Applications of Data Analysis:

- Healthcare: The main challenge for hospitals with cost pressures tightens is to treat as many patients as they can efficiently, keeping in mind the improvement of the quality of care. Instrument and machine data is being used increasingly to track as well as optimize patient flow, treatment, and equipment used in the hospitals. It is estimated that there will be a 1% efficiency gain that could yield more than $63 billion in the global healthcaresavings.
- Travel: Data analytics is able to optimize the buying experience through the mobile/ weblog and the social media data analysis. Travel sights can gain insights into the customer’s desires and preferences. Products can be up-sold by correlating the current sales to the subsequent browsing increase browse-to-buy conversions via customized packages and offers. Personalized travel recommendations can also be delivered by data analytics based on social mediadata.
- Gaming: Data Analytics helps in collecting data to optimize and spend within as well as across games. Game companies gain insight into the dislikes, the relationships, and the likes of theusers.
- Energy Management: Most firms are using data analytics for energy management, including smart-grid management, energy optimization, energy distribution, and building automation in utility companies. The application here is centered on the controlling and monitoring of network devices, dispatch crews, and manage service outages. Utilities are given the ability to integrate millions of data points in the network performance and lets the engineers use the analytics to monitor thenetwork.

Objective: To learn the concept of how to display summary statistics for each feature available in the dataset

Outcome: After completion of this assignment, it is expected to implement various statistics.

Input: Structured Dataset: Iris Dataset

Output:

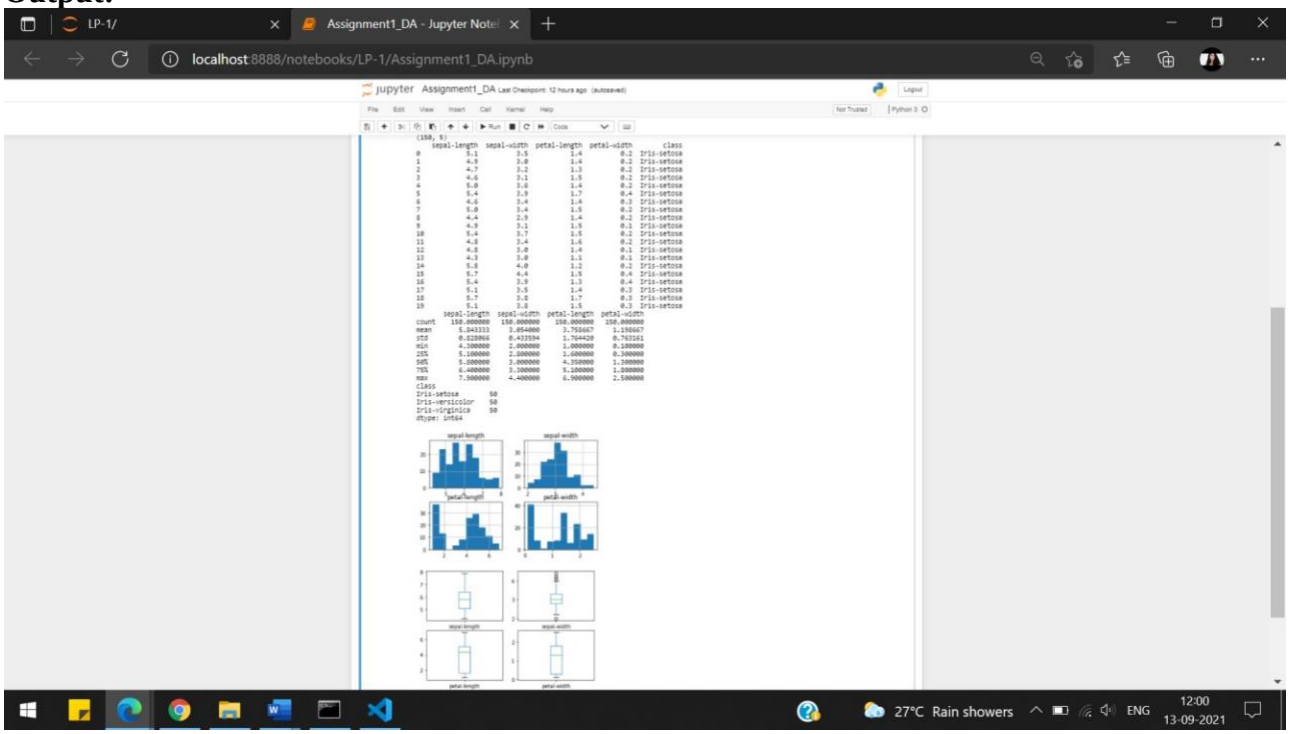
1. DatasetDetails.
2. Min, Max, Mean, Variance value and Percentiles of probabilities
3 Histogram using HistFunction.
4. Boxplot using Boxplot Function.

Result / Conclusion: We learned to analyse data distribution, and data statistics for given dataset.

Program:

```
# Load libraries
import pandas
import matplotlib.pyplot as plt

# Load dataset
url = "https://archive.ics.uci.edu/ml/machine-learning-
databases/iris/iris.data"
names = ['sepal-length', 'sepal-width', 'petal-length', 'petal-width',
'class']
dataset = pandas.read_csv(url, names=names)

# shape
print(dataset.shape)

# head
print(dataset.head(20))

# descriptions
print(dataset.describe())

# class distribution
print(dataset.groupby('class').size())


# histograms
dataset.hist()
plt.show()

# box and whisker plots
dataset.plot(kind='box', subplots=True, layout=(2,2), sharex=False,
sharey=False)
plt.show()
```

Output:


LP-1/
Assignment1_DA - Jupyter Notel
localhost:8888/notebooks/LP-1/Assignment1_DA.ipynb
jupyter Assignment1_DA
(150, 5)
sepal-length sepal-width petal-length petal-width class
0 5.1 3.5 1.4 0.2 Iris-setosa
1 4.9 3.0 1.4 0.2 Iris-setosa
2 4.7 3.2 1.3 0.2 Iris-setosa
3 4.6 3.1 1.5 0.2 Iris-setosa
4 5.0 3.6 1.4 0.2 Iris-setosa
5 5.4 3.9 1.7 0.4 Iris-setosa
6 4.6 3.4 1.4 0.3 Iris-setosa
7 5.0 3.4 1.5 0.2 Iris-setosa
8 4.4 2.9 1.4 0.2 Iris-setosa
9 4.9 3.1 1.5 0.1 Iris-setosa
10 5.4 3.7 1.5 0.2 Iris-setosa
11 4.8 3.4 1.6 0.2 Iris-setosa
12 4.8 3.0 1.4 0.1 Iris-setosa
13 4.3 3.0 1.1 0.1 Iris-setosa
14 5.8 4.0 1.2 0.2 Iris-setosa
15 5.7 4.4 1.5 0.4 Iris-setosa
16 5.4 3.9 1.3 0.4 Iris-setosa
17 5.1 3.5 1.4 0.3 Iris-setosa
18 5.7 3.8 1.7 0.3 Iris-setosa
19 5.1 3.8 1.5 0.3 Iris-setosa
sepal-length sepal-width petal-length petal-width
count 150.000000 150.000000 150.000000 150.000000
mean 5.843333 3.054000 3.758667 1.198667
std 0.828066 0.433594 1.764420 0.763161
min 4.300000 2.000000 1.000000 0.100000
25% 5.100000 2.800000 1.600000 0.300000
50% 5.800000 3.000000 4.350000 1.300000
75% 6.400000 3.300000 5.100000 1.800000
max 7.900000 4.400000 6.900000 2.500000
class
Iris-setosa 50
Iris-versicolor 50
Iris-virginica 50
dtype: int64
sepal-length
sepal-width
petal-length
petal-width
27°C Rain showers
ENG
12:00
13-09-2021