LP I

Assignment C1

Data Analysis

Title :- Iris Data Analysis Date of completion :- 5.11.20

Problem Statement :- Download the Iris Flower Dataset or any other dataset into a DataFrame. Use Python/R and Perform following

1) How many features are there & what are their types?
2) Compute & display summary statistics for each feature available in the dataset.
3) Data Visualization :- Create a histogram for each feature in the dataset to illustrate the feature distributions. Plot each histogram.
4) Create a boxplot for each feature in the dataset. All of the boxplots should be combined into a single plot. Compare distribution & identify outliers

Learning Objective :- 1) Understand dataframes and its features
2) Analyse Iris dataset.

Learning Outcome :- Students will be able to
1) Analyze different datasets.

Software/Hardware Requirement :- OS (Linux), Python, ~~gotat~~ Iris Dataset.

Page:
Date: / /

Theory:-

Libraries used:-

1) pandas
2) Numpy
3) matplotlib
4) .

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 ie summary statistics for each feature is captured

X = set of inputs

Y = set of outputs

DD → Deterministic Data

NDD → Non deterministic Data

fc → Failure case

1. Data set is collection of data.
2. Data analysis is a process of inspecting, cleansing, transforming, & modelling data with the goal of discovering useful information, informat^n conclusion & decision making.
3. Mean, Standard, Variance, regression, hypothesis are the fundamental data analytics methods.

Page:
Date: / /

Mean

Sum of data entities divided by no. of entities.

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

Sample Mean $\bar{x} = \frac{\Sigma x}{n}$

Standard deviation:-

measure variability and consistency of the sample or population.

$$\sigma = \sqrt{\frac{\Sigma (x-\mu)^2}{N}}$$

Variance:-

averaged squared deviation from the mean.

Dataset Used :- Iris dataset.

- describe()

gives all the parameters like mean, std. deviation, variance

- hist()

creates histogram

- boxplot()

plot a box plot

Page :
Date : / /

Test cases

Input	Actual Output	Expected O/p	Remark
describe for column 1.	count 150.0 mean 5.84 std V. 0.82 min 4.30 max 7.90	count 150.0 mean 5.84 std 0.82 min 4.30 max 7.9	Passed
plot histogram	plotted	plotted	Passed
Boxplot	plotted	plotted	Passed

Conclusion:- Thus I analyzed iris dataset successfully.