Roll no. 41310

Page No.:
Date:

* Assignment DA1 *

* **Title:-**

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

* **Problem statements:-**

Download the iris flower dataset into Dataframe. Use Python/R for perform following:

① How many features are there and what are their types (e.g. numeric, nominal etc.)

② Compute and display summary statistics for each feature available in dataset (eg min value, max value, mean, range, variance, standard deviation and percentiles.

③ Data visualization:- Create a histogram for each feature in the dataset to illustrate feature distributions. Plot each histogram

④ Create a boxplot from each feature in dataset. All of the boxplots should be combined single plot. Compare distributions and identify outliers.

* **Learning Objective:-**

① Learn to use dataset, dataframes, features of dataset in an application.

② Learn to compute summary statistics for the features.

③ Learn to use visualization technique.

* **Learning Outcome:-**
Students will be able to compute the statistics on the features of dataset.
Use histograms and boxplot on the features of 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 dataframe
s = start state
e = end state i.e. summary statistics of 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 type.
② summary statistics of each feature
③ Data visualization (histogram, boxplot)

Fme is set of main functions.
Fme = { f1, f2, f3 }.
where,
f_1 = function to load dataset into dataframe
f2 :- function to get number of features
f3 :- function to draw histogram for each feature
f4 :- function to get feature type
f5 :- function to draw boxplot for each feature

DD:- Deterministic Data
NDD:- Non-Deterministic data.
FC:- failure Case
No failure Case identified for application.

Theory:-

Data analysis is a procces of inspecting, cleansing transforming and modelling data with the goal of discovering the useful information, informing Conclusion, supporting and decision making. Data Analysis has multiple approaches, encompassing diverse techniques under the variety of names, while used in different business, science and technology & social science domains.

A dataset is Collection of data. Most Commonly dataset Correspond to the Contents of single database table, on single statistical data matrix where each Column of table represent Perticular Variable.

Mean, standard deviation, variance size, min and max are the fundamentels of data analytics process.

$$* \text{ Mean} = \frac{\text{sum of data enteries.}}{\text{No. of data enteries.}}$$

Population mean: $\mu = \frac{\Sigma x}{n}$

Sample mean: $\bar{x} = \frac{\Sigma x}{n}$,

* **Range**:- Difference between max. and mi
data entriesin set.

Range = (Max. data entey) - (Min data entey)

* **Standard deviation**:-
It measures Variabilit and Consistency of the sample on population in most real word applicatio Consistency is great advantege

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

Sample S.D. = $s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$

* **Variance**:-
The average squared deviation from mean is Variance.

* **Percentile**:-
Let p be any integer between 0 to 100. The p^{th} percentile of dataset is data value at which p percent of the value in dataset are less than or equal to this value.

Python Commonds/Algorithm Steps:-

① Importing numpy for numerical Calculation if required or pandas can be used. Import pandas for data access and manipulation. Import seaborn for plotting boxplot.

② Reading CSV Content in dataset named dataframe.

=> dataset = pd.read-csv(r, "iris.csv")

③ Printing shape i.e. Rows & Columns.

=> print (dataset.shape)

④ printing Columns and their datatypes.

=> dataset.dtypes.

⑤ To check Column is numeric or nominal.

=> is_numeric-dtype(dataset['Column'])

⑥ Describe mean, median, min, max, percentile

=> print (dataset.describe())

⑦ printing histograms.

=> dataset['Column'].plot.hist()

⑧ Printing boxplots:-

=> sns.boxplot(data=dataset)

⑨ for Compairing outliers.

=> sns.distplot(dataset['Column'], label="...", color='blue')

Test Cases:-

	Expected Output	Actual Output	Result.
①	Import dataset	Iris. CSV imported in dataframe	Pass.
②	Printing no. of Rows & Column	using shape in python. No. of rows & Column Printed (150,5)	Pass.

Expected output	Actual output	Result.
③ Peinting Columns & datatypes.	5 Columns printed with datatype	Pass
④ Check Column is numeric or not	All 4 Column for histogram are numeric	Pass.
⑤ Peinting histogram	4 histogram for Sepal length, sepal width, petal length, peted width printed	Pass.
⑥ Peinting boxplot	Box plot with all 4 Column peinted showing min, max & percentile	Pass
⑦ Compairing outlier	outliers are Compared by using distplot	Pass.

Conclusion:- Analysis of iris flower dataset is performed and shown using visualization technique such as histogram and boxplot.