Pune District Education Association's
College Of Engineering
Manjari (Bk.), Hadapsar, Pune-412307.
Accredited by NAAC

set

ASSIGNMENT No 3

Title : Descriptive statistics - Measure of central Tendency & variability.

Aim : Descriptive statistics - Measure of central tendency & variability.

Perform the following operations on any open source dataset. (eg. data, csv).

1. Provide summary statistics for a dataset with numeric variables grouped by one of the quantitative variable. For example, if your categorial variable is age, groups & quantitative variable in income, then provide summary statistic of income grouped by the age groups. Create a list that contains a numeric value for each response to the categorial variable.
2. Write a python program to display some basic statistical details like percentile, mean, standard deviation etc. of the species of 'Iris-setosa', 'Iris-versicolor' & 'Iris-versicolor' of Iris.csv dataset.

Provide the codes with o/p & explain everything that you do in this step.

→ P.T.O.

Pune District Education Association's
College Of Engineering
Manjari (Bk.), Hadapsar, Pune-412307.
Accredited by NAAC

set

Objectives : students should be able to perfoem the statistical operations using python on any open sourcedataset.

Requirement : 1. Basic of python progeamming.
2. Concept of Data preprocessing, Data formatting, Data Normalization & cleaning.

Theory : · Introduction :

Descriptive statistics is the building block of data science. Advanced analytics is often incomplete without analyzing descriptive statistics of the key metrics. In simple terms, descriptive statistics can be defined as the measures of central tendency & these measures can be broken down further into the measures of central tendency & the measures of dispersion.

Measures of central tendency include mean, median & mode, while the measures of variability include standard deviation, variance & the interquartile range. In this guide, you will learn how to compute these measures of descriptive statistics & use them to interpret the data.

→ P.T.O.

We will covet the topics given below :-
1. Mean 2. Median 3. Mode.
4. Standard deviation 5. Variance 6. Interquartile range.
7. Skewness.
We will begin by loading the dataset to be used in this guide.

Data: In this guide, we will be using fictitious data of loan applications containing 600 observations of 10 variables, as described below:

1. Marital_status: whether the applicant is married ("Yes") or not ("No").
2. Dependents: Number of dependents of the applicant.
3. Is_graduate: whether the applicant is graduate. ("Yes") or ("Not").
4. Income: Annual Income of the applicant (in USD).
5. loan_amount: loan amount (in USD) for which the application was submitted.
6. Term_Months: Tenure of the loan (in months).
7. Credit_score: whether the applicants credit score was good ("satisfactory") or Not ("Not_satisfactory").
8. Age: The applicants age (in years).
9. Sex: whether the applicant is female or male.
10. approval_status: whether the loan application was approved ("Yes") or not ("No").

→ P.T.O.

set

CSV file / Dataset - Iris_Dataset

· Required libraries

```
import pandas as pd
import numpy as np.
import matplotlib.pyplot as plt.
import seaborn as sns
from sklearn.datasets import load_Iris.
```

· Functions used -

· df = pd.read_csv("Iris_data_set.csv")

· df.head()

· df.shape()

· df.info()

· df.describe()

· df.isnull().sum()

· plt.show()

PDEA
Pune District Education Association

→ P.T.O.

- mean = grouped_df.mean()
- median = grouped_df.median()
- min = grouped_df.min()
- max = grouped_df.max()
- std = grouped_df.std()
- df.skew()

Do all operations on each column. Also Draw boxplot for each column.

Conclusion : In this guide, you have learned about the fundamentals of the most widely used descriptive statistics & their calculations with python. we covered the following topics in this guide.

1. Mean 2. Median 3. Mode.
4. Standard Deviation 5. Variance 6. Interquartile range
7. Skewness.

It is important to understand the usage of these statistics & which one to use, depending on the problem statement & the data.

→ P.T.O

Pune District Education Association's
College Of Engineering
Manjari (Bk.), Hadapsar, Pune-412307.
Accredited by NAAC

1set

To learn more about data preparation & building machine models using python's 'scikit-learn' library, please refer to the following guides-
1. scikit machine learning.
2. Ensemble modeling with scikit-learn.

1. Which terms include under Measures of central tendency?
→ Mean, Median, Mode.

2. Which terms include under Measures of variability?
→ Standard deviation, variance & interquartile range.

3. Describe Iris dataset.
→ Iris is a collection of instruments, materials, stimuli, data & data coding & analysis tools used for research into languages, signed language learning, etc.
It contains four features (length & width of sepals & petals) of 50 samples of three species of Iris (Iris setosa, Iris virginica & Iris versicolor). Iris dataset contains five columns such as Petal length, Petal width, Sepal length, Sepal Width & species Type. & rows being the samples.

4. Explain Mean
→ Mean represents the arithmetic average of the data.

→ P.T.O.

Pune District Education Association's
College Of Engineering
Manjari (Bk.), Hadapsar, Pune-412307.
Accredited by NAAC

ISE

5. Describe Measures of Central Tendency.
→ It describe the center of the data & often represented by mean, median, mode.

6. Explain Median
→ Median represents the 50th percentile of the middle value of the data, that separates the distribution into two halves.

7. Explain Mode.
→ It represents the most frequent value of a variable in the data.

8. Explain Standard Deviation.
→ It is a measure that is used to quantify the amount of variation of a set of data values from its mean. A low standard deviation for a variable indicates that the data points tend to be close to its mean & vice versa.

9. Explain Variance.
→ It is square of the standard deviation & the covariance of the random variable with itself.

→ P.T.O.

Pune District Education Association's
College Of Engineering
Manjari (Bk.), Hadapsar, Pune-412307.
Accredited by NAAC

Set

10. Explain Interquartile Range (IQR).

→ Measure of statistical dispersion. It is calculated as the difference between upper quartile (75th percentile) & the lower quartile (25th percentile). IQR is very important measure for identifying outliers & could be visualize using boxplot.

12. Explain term outliers.

→ Outliers are always given wrong direction for your expected results. Outliers always talk about extremities. Too small or too large.

13. Explain skewness.

→ It is used to measure of symmetry or lack of symmetry. The skewness value can be positive, negative or undefined. In a perfectly symmetrical distribution, the mean, median & mode will all have the same value.