Page No. Date

Assignment NO-3.

Title - Descriptive Stastics - Measures of central Tendency and variability.

problem statement - perform the following operations on any open source dataset (e.g. data.csv)
1) provide summary statistics (mean, median, ~~mode~~ minimum, maximum standared deviation) for a dataset (age, income etc)

2) write a python program to display some basic statistical details like. percentile, mean, standared deviation etc.

objective - To analyze and demonastrate knowledge of statistical data analysis techniques for decision making

PREREQUISITE -
1) Basic of python programming.
2) Concept of stastistic mean, median, minimum, maximum, standared deviation.

THEORY -
The data are summarized in some, but not all ways. we close descriptive that are either most often reported, or most often covered in introductory courses. There as follows:

- Central Tendency

Mean
median
mode

- Dispersion

Standared deviation (Std. deviation)
minimum
maximum

- Central Tendency

The mean median and mode are the three measures of central Tendancy. Mean is the arithmetic average of data set.

The mean

$$\bar{x} = \frac{\Sigma x}{N}.$$

Theory -

1. Introduction to big data.

Big data means really a big data, it is collection of Large datasets that cannot be processed using traditional Computing techniques Big data is not merelize a data, rather it has become complete subject which involve various tools, techniques & frame work.. Big data involve the data produce by different devices & applications.

2. Introduction to dataset.

A dataset is a collection of records, Similar to table rows, but the columns can contain not only strings or numbers but also nested data structures such as lists, map & other records.

3. python libraries for data science.

a. Numpy

one of the most fundamental packages in python, Numpy is general purpose array processing package. It provides high performance multidimensional array object.

⊙ what can you do with Numpy?

1. Basic array operations: add, Multiply, slice, flatten, reshape, index arrays.
2. work with datetime or linear algebra
3. Basic slicing & advanced indexing in numpy python.

b. Pandas :

pandas is an open source python packages that provides high performance, easy to use data structure and data analysis.

⊙ what can you do with Pandas?

1> Indexing, manipulating, renaming, sorting, merging data frame.

③ scikit learn :.

Introduced to the world as google summer of code project, scikit learn is a robust machine Learning library For python.

⊙ what can you do with scikit learn ?

1. classification : spam detection, image recognition.

2. clustering - Drug responce , stack price .

3. Regression - customer Segmentation, grouping. experiment outcomes.

4. Description of dataset - The Iris dataset was used in R.A Fishers classic 1936 paper. The use of multiple measurements in taxonomic problems & can also be Found on the UCI machine Learning repository One Flower species is linearly separable from the other two, but the other two are not linearly separable from each other.

5. panda dataframe functions for Load datasets.

1. The dataset is downloads from UCI repository
2. Now read CSV File as dataframe in python from path where you saved the same the Iris dataset is stored in .csv format.

6. Panda data Frame functions for data preprocessing

① cheaking of missing values in dataset.

is null () → cheak missing values or null values

is na () → count of missing values of column & row wise count.

7. panda functions for data formatting & normalization

The Transforming data stage is about convertion the dataset into Format that can be analyzed or modelled effectively.

8. panda functions for handling categorical variables:

- categorial variables have values that describe a quality or 'characterstic' of data unit, life what type or which category.

• label encoding- label encoding refers to converting the labels into a numeric form so as to convert them into the machine-readable form.

Height	Height
Tall	0
medium	1
Short	2

Algorithm.

Step 1: Import pandas & sklearn library for preprocessing

Step 2: Load the iris dataset in dataframe object df

Step 3: observe the unique values for the species column.

Step 4: Define label encoder object knows how to understand word labels

step 5: Encode labels in column 'species'

step 6: observe the unique values for the species column.

Conclusion -

In this way we have explored the function of the python library for data preprocessing data wrangling techniques & how to handle missing values on iris dataset.