Name : Tejas V. Redkar
Roll No : PC-44
PRN: 10322 10937
Batch : C2

DEC Lab Assignment - 1

* Problem Statement:

Data Handling, locate any open source data. Load data into data frame. Perform DataFrame operations. Perform basic statistical operations like mean, median, standard deviation.

* Objectives:

1) To explore various data sources & data repositories.
2) To explore the operations on a dataset file using data frame with basic statistical operations in Python.

* Execution:

Program implementation & output study.

* Conclusion:

Basic operations were performed on the .csv data file using Python.

CLASS:
Date :
Page :

* FAQ's

Q1) State the significance of handling missing values in a dataset.

Ans Handling missing values is crucial for several reasons:

1) Maintaining data integrity: Missing values can introduce several errors & inconsistencies in data analysis. Ignoring them can lead to incorrect or biased results, making it essential to address them properly.
2) Preventing Biased Analysis: If missing values are not handled, the data analysis may become biased towards the observations with complete data, potentially skewing the results.
3) Improving Model Performance: When building machine learning models, missing values can cause issues during training & prediction. Handling them appropriately can result in better model performance & generalization.
4) Meeting Analysis requirement: Some statistical techniques & machine learning algorithms require complete datasets. Addressing missing values is necessary to meet the prerequisites of these methods.
5) Enhancing data visualizations: Missing values can disrupt data visualizations, making it difficult to create meaningful charts & graphs.

6) Facilitating Data Sharing: If the dataset is to be shared, addressing missing values makes it more accessible & usable for other researchers, analysts or stakeholders.

Q2) Explain the central tendency measures with examples.

Ans Central tendency measures are statistical values that represent the center or average of a dataset, providing insights into where the data tends to cluster.
There are three primary measures: mean, median, & mode.

1) Mean (Average):

The mean is the most common measure of central tendency & is calculated by summing all the values in a dataset & dividing by the total number of values.

Fourmula: Mean = (Sum of values) / (Number of values)
Example: Mean of a score: [85, 92, 78, 88, 95]
Mean = (85+92+78+88+95) / 5
= 87.6

2) Median:

It is the middle value when a dataset is sorted

in ascending or descending order. If there is an even number of values, the median is average of the middle - two values.

Example: In above dataset : [85, 92, 78, 88, 95]
when sorted in ascending order - [78, 85, 88, 92, 95]
Median = 88

3) Mode:

The mode is the value that occures most frequently in a dataset. A dataset can have one mode (unimodal) multiple modes or no mode if all values are unique.

Example: In the dataset : [82, 78, 95, 82, 83, 88, 92]

82 is the mode.

Q3) Describe various methods to handle missing values in a dataset.

Ans Handling missing values is an essential step in data preprocessing to ensure accurate & reliable analysis. Some common methods to handle missing values are:

1) Deletion:

Listwise deletion: In this method, entire rows containing missing values.

Pairwise deletion: In this method, missing values are ignored for specific calculations.

when analyzing pair of variables.

2) Mean, median, or mode imputation: Missing values can be replaced using mean (for continuous data), median (for ordinal data) & mode (for categorical data). Imputation provides a simple way of filling missing values.

3) Interpolation: Interpolation methods, such as linear regression or spline interpolation, estimate missing values based on values of adjaunt data points.

Q4) Explain different types of data types.

Ans Data types are useful in computer programming, as they provide specifications as to how data is stored, manipulates & interpreted.

1) Numeric Data Types:
   - ① Integer (int): Represents whole numbers
   - ② Floating point (float): Represents number with decimal point.
   - ③ Double precision (double): Higher precision & larger range of values
   - ④ Long (long): Represents large integers (ex: 1234567890L)

2) Text Data Types:

① String (str or text) : Represents sequence of characters

② Character (char) : Represents a single character.

3) Boolean Data Type :

① Boolean (bool) : Represents binary values, typically True or False.

4) Date & Time Data Types :

① Date : Represents a specific data (formatted)
② Time : Represents specific time (ex: 14:30:00)
③ Datetime : Represents date & time together.

5) Categorical Data Types :

① Enumerations (enums) : Represents a set of predefined named values.
② Ordinal : Represents ordered categorical data
③ Nominal : Represents non-ordered categorical data.

ASN