

Devanshu Surana

PC-23, Panel C, Batch C1

1032210755

## DEC Lab Assignment 1

Problem statement: Data Handling, Locate any open source data Load data into data frame. Perform Data frame operations, Perform basic statistical operations like mean, median, standard deviation.

### Objectives:

- 1) To explore various Data Sources and Data Repositories
- 2) To explore the operation on a dataset file using data frame with basic statistical operations in python.

### Methodology

- 1) Identify and study various data sources.

#### A-1) Public Data Sources.

- Open data Portal
- Publicly Accessible websites
- Research publications

#### Private Data Sources

- Corporate Data
- Market Research firm
- Subscription services

## Government Data Sources

- Government databases
- Census data
- Geospatial data
- Weather and environmental data

It is important to note that access to certain government and private data sources may be restricted or subject to privacy regulations and legal requirements.

2) To study a dataset with various operation using Python command.

- Load Dataset
- Read CSV file
- Display:

Head: Displays first five values of dataset.

Tail: Displays last five values of dataset

Describe: Gets summary statistics for numeric cols.

Summary: Depending upon your or specific analysis you can create a custom summary of your dataset using Pandas.

## Data Handling.

### Remove duplicates

By, 'drop\_duplicates()' keeps the first occurrence of each duplicated row and removed subsequent duplicates.



Identify and display missing values.

- Using the 'isna()' or 'isnull()' both return a dataset frame of the same shape as the original. where each element is 'True' if the corresponding element in original Dataframe is missing (NaN or None) and 'false' otherwise.

Statistical basic operations to handle missing values.

- for ex: rows with missing values : we can use the command dropna.

df.dropna(inplace = True).

Execution

Program implementation and output study

Conclusion :

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

FAQ's

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

→ Some significant reasons are :

- 1) Preservation of data Integrity
- 2) Avoidance of Bias.
- 3) Improved model ~~Prefer~~ Performance
- 4) Enhanced Data Visualisation

5) Reduced Noise in Analysis.

Q2) Explain Central tendency measures with examples.  
→ 1) Mean: Avg of all values divided by the no. of values.

Ex: Consider scores of test = {85, 92, 88, 78, 95}  
$$\text{mean} = \frac{85 + 92 + 88 + 78 + 95}{5} = 88.6$$

mean is 88.6

2) Median: Middle value of a dataset when its arranged in ascending or descending order.

for dataset: 78, 85, 88, 92, 95 median is 88.

3) Mode: Mode is most frequently occurred element in a dataset

for ex: In dataset (1, 1, 2, 2, 2, 3, 4, 4, 4, 4)  
mode = 4

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

→ Mean/Median/Mode imputation: Replace null/missing values with median/mode/mean.

K-nearest Neighbours (KNN) Imputation: Impute missing values by considering the values of their K-nearest neighbors in dataset.



- Create a New Category: for categorised data, you can create a new category (for eg: "Unknown" or "others").

- Machine-learning based Imputation: Train machine learning models (eg: decision trees, random forests, or neural networks) to predict missing values on other features in dataset.

Q4) Explain different datatype in Python.

→ Numeric:

int (integer): Represent whole numbers for ex: '+5', '-3'

float: Represent decimal point numbers for ex: '3.14', '-0.5'

Complex: Represent complex with a real and imaginary part for ex:  $3+2j$ .

Text

str (string): represents text or sequence of characters enclosed in single or double quotes. for ex: "Hello World".

Boolean Data Type

bool (Boolean): Represents either True / False.

Sequence Types

List: Represents an ordered collection of items that can be of different data types.

Tuple : Represent an ordered collection of items, similar to lists, but tuples are immutable, meaning their elements cannot be changed once defined.

range + Represents an immutable sequence of numbers and is often used for looping.

✓

✓