**Name:** Manish More

**Roll no:** 282010 **PRN:** 22310220

**Class :** SY CSE AI **Batch:** B1

**Assignment 1**

**Statement:**

**Perform the following operations using R/Python on suitable data sets:**

a) read data from different formats (like csv, xls)

b) Find Shape of Data

c) Find Missing Values

d) Find data type of each column

e) Finding out Zero's

f) Indexing and selecting data, sort data,

g) Describe attributes of data, checking data types of each column,

h) counting unique values of data, format of each column, converting variable data type (e.g.

from long to short, vice versa)

**Objective**

1. Enhance skills in data loading, exploration, and preprocessing using Python (Pandas) or R.
2. Understand fundamental techniques for handling missing values and formatting data.
3. Develop proficiency in organizing, indexing, and sorting datasets for better analysis.

**Tools and Resources:**

* Software: Google Colab or Jupyter Notebook
* Libraries: Pandas, NumPy

**Key Pandas Functions Used:**

1. pd.read\_csv("file.csv"): Load data from a CSV file.
2. df.shape: Get the number of rows and columns in the dataset.
3. df.isnull().sum(): Identify missing values for each column.
4. df.dtypes: Display data types of columns.
5. (df == 0).sum(): Count the number of zero values in the dataset.
6. df.sort\_values(by='column\_name'): Sort data based on a specific column.
7. df.describe(): Generate summary statistics for numerical columns.
8. df.nunique(): Count unique values in each column.
9. df['column\_name'] = df['column\_name'].astype(new\_type): Convert data types.

**Methodology**

**1. Data Loading and Exploration:**

* Read data from the file heart.csv using Pandas.
* Display the shape and preview the first few rows of the dataset.

**2. Data Cleaning and Preprocessing:**

* Identified and handled missing values.
* Detected and analyzed zero values in the dataset.
* Verified and modified column data types where necessary.

**3. Data Manipulation and Analysis:**

* Indexed and selected specific rows/columns.
* Sorted data based on a relevant attribute.
* Counted unique values in columns and analyzed data formats.
* Converted data types to ensure proper formats for analysis.

**Advantages of Using Pandas**

1. Simple and Intuitive: Provides flexible data structures (Series and DataFrame).
2. Efficient Data Handling: Enables quick processing of large datasets.
3. Comprehensive Features: Supports extensive data cleaning and transformation operations.

**Challenges**:

1. Managing missing values without losing essential information.
2. Handling datasets with large numbers of zero values affecting computations

**Conclusion**

This assignment provided hands-on experience with essential Pandas functions for data exploration and preprocessing. I have learned how to:

* Load datasets and analyze their structure.
* Handle missing and zero values.
* Sort, filter, and convert data efficiently.