**Name:** Shubhankar D. Jakate

**Roll no:** 282019 **PRN:** 22310371

**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
* **Libraries**: Pandas, NumPy

**Key Pandas Functions Used**

1. pd.read\_csv("file.csv") / pd.read\_excel("file.xlsx"): Load data from CSV or Excel files.
2. df.shape: Get the number of rows and columns.
3. df.isnull().sum(): Identify missing values in each column.
4. df.dtypes: Display the data types of each column.
5. (df == 0).sum(): Count 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 a CSV or Excel file into a Pandas DataFrame.
   * Display the dataset's shape and preview the first few rows.
2. **Data Cleaning and Preprocessing**
   * Identify missing values and handle them using imputation or removal.
   * Detect zero values and analyze their significance in the dataset.
   * Verify and modify column data types if necessary.
3. **Data Manipulation and Analysis**
   * Index and select specific columns or rows.
   * Sort data based on relevant attributes.
   * Count unique values and analyze column formats.
   * Convert data types to appropriate formats when needed.

**Advantages of Using Pandas**

1. **Simple and Intuitive**: Provides easy-to-use data structures (Series and DataFrame).
2. **Powerful Data Handling**: Efficiently processes large datasets.
3. **Comprehensive Functionality**: Supports data cleaning, transformation, and analysis.

**Challenges**:

1. Can consume significant memory for very large datasets.
2. Requires careful handling of missing and incorrect values to ensure data accuracy.

**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.