**Assignment 1**

**Name: Yash Kishor Maske**

**Roll No: 282007**

**PRN: 22310125**

**Class: SY CSE AI**

**Batch: B1**

**Statement**

Perform the following operations using R/Python on appropriate datasets:  
a) Load data from different formats (e.g., CSV, Excel).  
b) Determine the shape of the dataset.  
c) Identify missing values.  
d) Check the data type of each column.  
e) Count the number of zeros in the dataset.  
f) Perform indexing, selection, and sorting of data.  
g) Describe dataset attributes and verify column data types.  
h) Count unique values, check column formats, and convert variable data types (e.g., long to short and vice versa).

**Objective**

1. Develop skills in data loading, exploration, and preprocessing using Python (Pandas) or R.
2. Understand fundamental techniques for handling missing values and formatting data.
3. Gain proficiency in organizing, indexing, and sorting datasets for effective 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 datasets from CSV or Excel files.
2. df.shape – Retrieve the number of rows and columns.
3. df.isnull().sum() – Detect missing values in each column.
4. df.dtypes – Display column data types.
5. (df == 0).sum() – Count zero values in the dataset.
6. df.sort\_values(by='column\_name') – Sort data by a specified 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 column data types.

**Methodology**

**1. Data Loading and Exploration**

* Import data from CSV or Excel files into a Pandas DataFrame.
* Display the dataset's shape and preview the first few rows.

**2. Data Cleaning and Preprocessing**

* Identify and handle missing values through imputation or removal.
* Detect and analyze the presence of zero values in the dataset.
* Verify and modify column data types if required.

**3. Data Manipulation and Analysis**

* Select specific rows or columns for analysis.
* Sort data based on relevant attributes.
* Count unique values and analyze column formats.
* Convert data types to appropriate formats when necessary.

**Advantages of Using Pandas**

1. **User-Friendly** – Offers easy-to-use data structures (Series and DataFrame).
2. **Efficient Processing** – Handles large datasets effectively.
3. **Versatile** – Supports data cleaning, transformation, and analysis with built-in functions.

**Challenges**

1. **Memory Usage** – Large datasets may consume significant memory.
2. **Data Accuracy** – Careful handling of missing and incorrect values is required.

**Conclusion**

This assignment provided practical experience with Pandas for data exploration and preprocessing. Key learnings include:

* Loading datasets and analyzing their structure.
* Handling missing and zero values.
* Sorting, filtering, and converting data efficiently.