Name: Rohit Mahaveer Bhabire.

Roll. No :281031 Batch: A2

Assignment No. 1

Statement:

Perform the following operations using R/Python on a suitable dataset:

a) Read data from different formats (like CSV, XLS)
b) Find the shape of the data
c) Find missing values
d) Find the data type of each column
e) Identify zero values
f) Indexing and selecting data, sorting data
g) Describe attributes of data, checking data types of each column
h) Count unique values, identify format of each column, convert variable data types (e.g., from long to short, and vice versa)

Objectives of This Assignment:

1. To introduce the **Pandas** library and its core functionalities.
2. To become familiar with **data cleaning** and **preprocessing** techniques.
3. To enhance skills in managing and manipulating **data from various file formats**.
4. To develop a foundation for advanced **data analysis and visualization** tasks.

Resources Used:

- **Software:** Google Colab
- **Language:** Python
- **Library:** Pandas

Introduction to Pandas:

1. **Pandas** is a widely-used open-source Python library for **data manipulation and analysis**.
2. It provides easy-to-use **data structures** and **data analysis tools** for working with structured data.
3. The core components of Pandas:
    - **Series:** A one-dimensional labeled array.
    - **DataFrame:** A two-dimensional labeled structure with columns of potentially different types.
4. With Pandas, users can:
    - Load data from CSV, Excel, or databases
    - Filter, sort, and group data
    - Perform statistical, cleaning, and transformation operations

---

Basic Functions Used:

1. **pd.read_csv()** – Reads a CSV file into a DataFrame.
2. **head()** – Displays the first few rows of the DataFrame.
3. **sort_values()** – Sorts the DataFrame based on a column (e.g., 'Age').

4. **describe()** – Provides summary statistics for numerical columns.
5. **info()** – Gives info about data types, null values, and memory usage.

Methodology:

1. **Data Collection:**
   Imported data using Pandas from a .csv file.
2. **Data Exploration & Cleaning:**
   - Checked shape, data types, and missing values
   - Identified zero values and unique entries
   - Sorted the data for better understanding
   - Converted data types as needed for analysis
3. **Data Analysis:**
   Used descriptive statistics and data exploration functions to analyze patterns and structure in the dataset.

Advantages:

- **User-Friendly**: Easy to use and intuitive for beginners.
- **Powerful Structures**: Series and DataFrame simplify data manipulation.
- **Flexible and Fast**: Supports a wide range of operations for structured data.
- **Integrated with Python**: Works seamlessly with NumPy, Matplotlib, and other libraries.

Disadvantages:

- **Memory Usage**: Can be heavy on memory with large datasets.
- **Python-Centric**: Limited interoperability with non-Python environments.

Conclusion:

This assignment served as an excellent introduction to the **Pandas library**, providing practical experience in **data handling and preprocessing**. We covered essential operations including reading data, exploring its structure, handling missing or zero values, and converting data types. These foundational tasks form the basis of any data analysis pipeline and are critical for building more complex data science and machine learning projects. Mastering Pandas will be invaluable for future assignments and real-world data analytics scenarios.