**Name:** Aayush Avinash Bokde  
**Class:** CSE(AI) B div  
**Roll No.:** 282016  
**PRN No.:** 22310316  
**Subject:** Machine Learning

**Assignment 1 -** Perform the following operations using R/Python on suitable datasets

**Statement**

**Q.** In this assignment, we have to:

Perform the following operations using R/Python on suitable datasets:

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) Find out number of zeroes  
f) Indexing and selecting data, sorting data  
g) Describe attributes of data, checking data types of each column  
h) Count unique values of data, check format of each column, and convert variable data type (e.g., from long to short, and vice versa)

**Objective**

1. To introduce the Pandas library and its basic functionalities for data manipulation.
2. To demonstrate reading and analyzing data in different formats (CSV/XLS).
3. To explore and practice essential data preprocessing and data cleaning techniques.
4. To develop skills in handling structured data for analysis and visualization.

**Resources Used**

* **Software**: Google Colab
* **Library**: Pandas

**Introduction to Pandas**

1. Pandas is a powerful open-source Python library used for data analysis and manipulation.
2. It offers two primary data structures: **Series** (1D) and **DataFrame** (2D tabular data).
3. It supports reading and writing data in various formats like CSV, Excel, SQL databases, JSON, and more.
4. It is especially useful for data wrangling, exploration, and preprocessing before applying machine learning or statistical models.

**Some Basic Functions Used in This Program**

| **Function** | **Description** |
| --- | --- |
| pd.read\_csv() | Reads data from a CSV file |
| shape | Returns the dimensions (rows, columns) of the DataFrame |
| isnull().sum() | Identifies and counts missing values in each column |
| info() | Shows summary information including data types and non-null counts |
| (df == 0).sum() | Finds the number of zeros in each column |
| head() | Displays the first 5 rows of the dataset |
| iloc[] | Enables indexing by position to retrieve specific rows/columns |
| describe() | Generates summary statistics (mean, std, min, etc.) |
| nunique() | Counts unique values in a column |
| dtypes | Returns data types of all columns |

**Methodology**

**1. Data Collection and Loading**

* We used the *Mall\_Customers.csv* dataset, which contains customer information like Age, Gender, Annual Income, and Spending Score.
* It was uploaded and read using pd.read\_csv().

**2. Exploration and Inspection**

* Used shape, head(), info(), and dtypes to inspect the data structure and types.
* Checked for missing values using isnull().sum().

**3. Data Cleaning**

* Searched for zero values that might affect model training or interpretation.
* Retrieved individual rows using iloc[] for validation.

**4. Statistical Summary**

* describe() provided summary statistics to understand the distribution of data.

**5. Feature Analysis**

* Identified how many unique values were present in columns like 'Age'.
* Ensured that the format and data types were appropriate for further processing.

**Advantages of Pandas**

* Easy-to-use syntax with intuitive data structures.
* Fast and efficient data handling, especially with tabular data.
* Supports integration with NumPy, Matplotlib, and Scikit-learn for deeper analysis.
* Excellent for data cleaning, filtering, aggregation, and transformation tasks.

**Disadvantages of Pandas**

* Can consume significant memory when handling very large datasets.
* Mostly tied to Python ecosystem—limited direct use with other programming environments.
* Some operations can be slow for extremely large or real-time datasets.

**Conclusion**

In this assignment, we explored essential data operations using Pandas in Python. We read data from CSV files, checked for missing or zero values, and understood the dataset using descriptive statistics. We practiced indexing, data selection, and converting data types — all of which are crucial preprocessing steps in any data analysis project. This hands-on experience with Pandas will be instrumental in future assignments involving data science, ML models, or statistical exploration.