**Name:** Shubh Dahane  
**Roll No:** 282015

**PRN No :**22310312  
**Batch:** B1  
**Assignment 1**

**Statement:**

**Q.** 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 Zeros  
f) Indexing and selecting data, sorting data  
g) Describe attributes of data, checking data types  
h) Counting unique values, data format of each column, converting variable data type

**Objective:**

1. To get familiar with the **Pandas** library and its core functionalities.
2. To understand basic **data cleaning**, **data exploration**, and **preprocessing** tasks.
3. To build proficiency in handling structured data for analysis.

**Resources Used:**

* **Software Used:** Jupyter Notebook
* **Library Used:** Pandas

**Introduction to Pandas:**

1. **Pandas** is a popular open-source Python library for data manipulation and analysis.
2. It offers two primary data structures:
   * **Series** – One-dimensional labeled array.
   * **DataFrame** – Two-dimensional labeled data structure.
3. It allows operations such as:
   * Reading CSV/Excel files.
   * Handling missing or null values.
   * Statistical analysis and transformation.
   * Indexing, slicing, and sorting data.

**Basic Functions Used:**

| **Function** | **Purpose** |
| --- | --- |
| pd.read\_csv() | Reads data from a CSV file |
| df.shape | Returns the number of rows and columns |
| df.isnull().sum() | Checks for missing values |
| df.dtypes or df.info() | Displays data types of each column |
| (df == 0).sum() | Finds zero entries in the dataset |
| df.iloc[], df.loc[], df[] | Indexing and data selection |
| df.sort\_values() | Sorts the data based on a column |
| df.describe() | Descriptive statistics |
| df.nunique() / value\_counts() | Finds unique values |
| astype() | Converts column data type |
| fillna(), interpolate() | Handles missing values |

**Methodology:**

**1. Data Collection and Exploration:**

* Used a heart attack prediction dataset stored as a CSV file.
* Loaded into a Pandas DataFrame and explored initial rows using df.head().
* Checked the **shape**, **data types**, **missing values**, and **unique values**.

**2. Data Preprocessing:**

* Checked for and handled missing values in the Ca column using interpolation.
* Converted Ca column to int64 after filling nulls.
* Detected zero entries to understand skewness or placeholders in data.

**3. Data Manipulation:**

* Indexed and selected specific rows/columns using both .iloc and .loc.
* Sorted dataset based on cholesterol level (Chol) in ascending order.
* Displayed basic statistics using .describe().

**Advantages:**

1. Easy to use and very powerful for data analysis.
2. Supports complex operations on large datasets with few lines of code.
3. Great integration with other data science libraries like NumPy, Matplotlib, etc.

**Disadvantages:**

1. Can consume a lot of memory with large datasets.
2. Primarily designed for Python, which limits integration with other languages.

**Conclusion:**

This assignment provided hands-on experience with the **Pandas** library—an essential tool for data manipulation in Python. It covered a wide range of operations including reading data, handling missing values, sorting, indexing, and data transformation. These skills are foundational for any data science or machine learning pipeline, and mastering them will lead to better and more efficient data analysis in future projects.

A graph showing age distribution

AI-generated content may be incorrect.

A graph of heart disease

AI-generated content may be incorrect.

A graph showing a number of people

AI-generated content may be incorrect.

A graph showing different colored boxes

AI-generated content may be incorrect.