**Name:** Manisha Apchunde
**Roll no:** 281076
**PRN:** 22420244
**Batch:** A3

# Assignment 01

## Aim:

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) This assignment aims to introduce you to the Pandas library and its basic functions. The library provides functionality for reading different file formats such as CSV and Excel.

2) Additionally, it familiarizes users with data cleaning and preprocessing techniques.

3) Enhance our skills in handling data in various formats, improving our proficiency in data analysis and manipulation.

## Resources used:

Programming Language: Python

1. Software used : Visual Studio Code

2. Library used : Pandas

## Introduction To pandas:

1) Pandas is a powerful and widely-used open-source Python library for data manipulation and analysis.

2) It provides easy-to-use data structures and functions, making it an essential tool for working with structured data.
3) At the core of Pandas are two main data structures: Series and DataFrame.

4) A Series is a one-dimensional labelled array capable of holding any data type.

5) DataFrame is a two-dimensional labelled data structure with columns of potentially different types.

6) These data structures allow users to perform a wide range of operations on data, including loading data from various file formats (such as CSV, Excel, SQL databases), manipulating data (e.g., sorting, filtering, grouping), and performing statistical and analytical tasks.

## Some basic functions that we used in program:

1. **pd.read_csv():**

   This function is used to read data from a CSV file into a DataFrame.

2. **head():**

   It is used to display the first few rows of the DataFrame, providing a quick overview of the data.

3. **sort_values():**

   This function sorts the DataFrame by the values of a specified column (in this case, 'Age'), allowing data to be arranged in ascending order.

4. **describe():**

   It generates descriptive statistics for numerical columns in the DataFrame, such as count, mean, standard deviation, minimum, and maximum values.

5. **head():**

   It returns the first n rows of a DataFrame, providing a quick way to preview the structure and content of the dataset.

6. **describe():**

7. This function generates descriptive statistics for numerical columns in the DataFrame, such as count, mean, standard deviation, minimum, and maximum values.

8. **unique():**

This function returns an array of unique values in a column of the DataFrame, useful for identifying distinct categories or groups in categorical data.

## Methodology:

**1.** Data Collection and Exploration:

- **Collect Data**:
  Obtain the heart attack prediction dataset, ensuring it contains relevant features such as age, gender, blood pressure, cholesterol levels, etc.
- **Explore Data**:

  Load the dataset into a pandas DataFrame and explore its structure, including the number of samples, features, data types, and any missing or erroneous values.

**2.** Data preprocessing:

- **Handle Missing Values**:
  Identify and handle missing values appropriately, considering strategies like imputation with mean, median, or mode, or removal of rows or columns with significant missing data.
- **Data Cleaning**:

  Perform data cleaning tasks such as removing duplicates, correcting erroneous entries, and ensuring consistency in data formatting.

**3.** Feature Engineering:

- **Feature Selection**: Select relevant features for heart attack prediction, considering domain knowledge and feature importance techniques like correlation analysis or feature importance scores.
- **Feature Encoding**: Encode categorical variables into numerical format using techniques like one-hot encoding or label encoding to make them suitable for machine learning algorithms.

## Advantages:

1. It is very easy to used library thats why it is also famous library.

2. It provided powerful data structure  like Series and DataFrame.

3. It comes with wide functionality for data manipulation.

## Disadvantages:

1. Pandas may comsume significant memory while working with large dataset.

2. It vary much integrated with python eco-system,which may limit its interoperability with other programming languages or environments.

## Conclusion:

This assignment covered fundamental data exploration techniques using Python. We successfully loaded datasets from different formats, checked their structure, handled missing values, sorted data, and analyzed key attributes. These operations are crucial in preprocessing, ensuring clean and well-structured data for further machine learning tasks