**Name:** Rutuja Tikhole

**Roll no:** 281004

**Batch:** A1

**Assignment 1**

**Statement:**

Q. In this assignment we have to do

a) Read data from different formats

b) Indexing and selecting, sorting data

c) Describe attributes of data, checking data types of each column

d) Counting unique values of data, format of each column, converting variable data type (e.g.from long to short, vice versa)

e) Identifying missing values and fill in the missing values.

**Objective:**

1. This assignment introduces the Pandas library and its core functionalities, focusing on reading different file formats such as CSV and Excel.
2. It also covers data cleaning and preprocessing techniques to improve data quality.
3. The goal is to enhance our ability to handle, manipulate, and analyze structured data effectively.

**Resources Used:**

1. Software: Jupyter Notebook
2. Library: Pandas

**Introduction to Pandas:**

1. Pandas is a robust, open-source Python library for manipulating and analyzing data.
2. It is a vital tool for managing structured data since it offers user-friendly data structures and operations.
3. Two essential data structures form the foundation of the library: a. Series: A labeled array with one dimension that can hold different kinds of data. b. DataFrame: A labeled data structure with two dimensions that can store many kinds of data in its columns.
4. Pandas facilitates a wide range of data operations, including cleaning, filtering, sorting, grouping, statistical analysis, and loading data from several file formats (CSV, Excel, and SQL databases).

**Key Functions Used in the Program:**

1. pd.read\_csv(): Loads data from a CSV file into a DataFrame.
2. shape: Returns the number of rows and columns in the dataset.
3. isnull().sum(): Identifies missing values within the dataset.
4. dtypes: Displays the data type of each column.
5. (df == 0).sum(): Counts the number of zero values in each column.
6. sort\_values(): Sorts the DataFrame based on a specified column in ascending or descending order.
7. describe(): Generates descriptive statistics such as count, mean, standard deviation, minimum, and maximum values for numerical columns.
8. unique(): Returns an array of unique values in a given column, useful for identifying distinct categories in categorical data.

**Methodology:**

1. Data Collection and Exploration:
   1. Data Acquisition: Obtain a relevant dataset containing essential features.
   2. Data Exploration: Load the dataset into a Pandas DataFrame and inspect its structure, including the number of rows, columns, data types, missing values, and inconsistencies.
2. Data Preprocessing:
   1. Handling Missing Values: Identify and address missing values through imputation or removal.
   2. Data Cleaning: Remove duplicates, rectify incorrect entries, and standardize formatting to ensure data consistency.
3. Feature Engineering:
   1. Feature Selection: Identify the most relevant features using domain knowledge and statistical techniques.
   2. Feature Encoding: Convert categorical data into numerical format using techniques like one-hot encoding or label encoding to facilitate analysis.

**Advantages of Pandas:**

1. User-friendly syntax, making it accessible for beginners and professionals alike.
2. Robust data structures like Series and DataFrame for efficient data handling.
3. Extensive functionality for data manipulation, allowing operations such as filtering, merging, and reshaping data.
4. Compatibility with other Python libraries such as NumPy and Matplotlib for advanced analysis and visualization.

**Disadvantages of Pandas:**

1. Handling very large datasets can lead to high memory consumption, impacting performance.
2. While highly integrated with Python, its interoperability with other programming languages is limited compared to some other data-processing tools.

**Conclusion:**

To sum up, this assignment gave me a basic understanding of the Pandas package, which is an essential tool for Python data analysis and manipulation. We looked at key features for effectively reading, organizing, and processing data. We improved our ability to handle structured data, clean and preprocess information, and conduct statistical analysis by using Pandas in real-world situations. Further developments in data science and machine learning applications are facilitated by these ideas.