Name: Snehal Yelwande
Roll No: 281063
Batch: A3

Assignment 1

Question:

Perform the following operations using Python on a suitable dataset:
a) Read data from different formats (like CSV, XLS).
b) Find the shape of the data.
c) Check for missing values.
d) Identify the data type of each column.
e) Find the number of zero values.
f) Perform indexing, selecting, and sorting data.
g) Describe dataset attributes and check data types.
h) Count unique values in columns and convert variable data types.

Objective:

1. Learn the basics of the Pandas library and its functions.
2. Understand data cleaning and preprocessing techniques.
3. Improve data handling skills for better analysis and manipulation.

Resources Used:

- **Software:** Visual Studio Code
- **Library:** Pandas

Introduction to Pandas:

Pandas is a widely used Python library for data manipulation and analysis. It provides powerful data structures like:

1. **Series:** A one-dimensional labelled array.
2. **Data Frame:** A two-dimensional labelled table with multiple columns.

These structures help in loading, organizing, analyzing, and visualizing data efficiently.

Basic Functions Used in the Program:

1. pd.read_csv(): Reads a CSV file into a Data Frame.
2. head(): Displays the first few rows of the dataset.
3. sort_values(): Sorts data based on a specific column.
4. describe(): Provides summary statistics of numerical columns.
5. info(): Shows data types, memory usage, and non-null counts.

Methodology:

1. **Data Collection and Exploration:**
   - Load the dataset into Pandas.
   - Check for missing values, incorrect data, and overall structure.
2. **Data Preprocessing:**
   - Handle missing values using techniques like mean/median imputation or row removal.
   - Clean the data by removing duplicates and fixing errors.
3. **Feature Engineering:**
   - Select relevant features for analysis.
   - Convert categorical variables into numerical values.

Advantages of Pandas:

- Easy to use and learn.
- Provides powerful data manipulation tools.
- Handles large datasets efficiently.

Disadvantages of Pandas:

- Can be memory-intensive for very large datasets.
- Works mainly within the Python ecosystem, limiting cross-language compatibility.

Conclusion:

This assignment introduced the Pandas library and its basic functions for handling and analyzing data. We explored how to read different file formats, clean missing values, and manipulate data efficiently. These skills will help in future data analysis projects and improve our understanding of real-world data handling.