Assignment 1

Name: Yashraj Mohan Nimbalkar

Roll No: 281039 Batch: A2

Statement

In this assignment, we focus on the following tasks:

a) Import data from different file formats.

b) Apply indexing, data selection, and sorting techniques.

c) Analyze data attributes, determine data types, and count distinct values.

d) Modify and reformat columns, and convert data types when needed.

e) Detect and manage missing data efficiently.

Objective

1. Introduce the Pandas library and its powerful tools for handling structured data, including reading files like CSV and Excel.

2. Learn essential data cleaning and transformation methods.

3. Gain hands-on experience in handling and processing real-world datasets to build a solid foundation in data analysis.

Resources Used

• Software: Visual Studio Code

• Library: Pandas

Introduction to Pandas

Pandas is a widely-used, open-source Python library designed for efficient data manipulation and analysis. It offers user-friendly and adaptable data structures that simplify working with organized datasets.

1. Main Data Structures

• Series – A one-dimensional labeled array capable of holding data of any type.
• DataFrame – A two-dimensional, tabular data structure with labeled axes (rows and columns), where each column can hold a different data type.

2. Notable Features

Pandas enables a variety of data operations such as:

• Importing data from multiple file types including CSV, Excel, and SQL databases.
• Filtering, grouping, reshaping, and sorting data seamlessly.
• Performing both basic descriptive analysis and more complex statistical evaluations.

Frequently Used Functions

1. pd.read_csv() – Reads CSV files and stores the content into a DataFrame.
2. head() – Displays the initial rows of the dataset.
3. sort_values() – Arranges data based on specified column values.
4. describe() – Generates statistical summaries for numerical columns.
5. unique() – Identifies and returns unique entries in a specific column.

Methodology

1. Loading and Inspecting the Dataset
   • Dataset Used: A sample dataset (e.g., diabetes data or health metrics) featuring attributes like age, glucose level, BMI, etc.
   • Initial Exploration: Load the dataset with Pandas, examine dimensions, column types, and check for missing or null values.
2. Data Cleaning and Preparation
   • Handling Missing Values: Fill missing entries using techniques like mean, median, or mode imputation; alternatively, drop them.
   • Data Tidying: Remove duplicate entries, correct inconsistent values, and standardize data formats.
3. Feature Engineering
   • Selection: Identify and retain key features relevant to analysis, guided by correlation scores or domain knowledge.
   • Encoding: Transform categorical variables into numerical format using label encoding or one-hot encoding.

Pros of Using Pandas

1. Intuitive and beginner-friendly for data manipulation tasks.
2. Offers robust data structures like DataFrames and Series.
3. Supports a comprehensive suite of tools for data wrangling and analysis.

Cons of Using Pandas

1. Performance may lag with very large datasets due to memory limitations.
2. It is Python-centric and lacks broad integration with other programming languages.

Conclusion

This assignment offered a practical introduction to Pandas for managing and analyzing data with Python. Through hands-on exercises, we practiced tasks like data importing, cleaning, transformation, and summarization. These foundational skills are vital for progressing further into the field of data science and performing efficient data analysis.