ASSIGNMENT NO: 1

Problem Statement -

Perform the following operations using Python on suitable data sets:
a) read data from different formats (like csv, xls)
b) indexing and selecting data, sort 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

S/W Packages and Libraries used:

For the following assignment, the interpreter used was Google Collab and the Libraries used were

- Pandas - It is a powerful data manipulation library in Python, that provides data structures and functions for working with structured data.

- Numpy - NumPy is a fundamental package for scientific computing with Python, providing support for large, multi-dimensional arrays and matrices, along with a collection of mathematical functions to operate on these arrays.

Methodology:

- For Reading Data from Different Formats:
    - Utilize libraries like Pandas to read data from various formats such as CSV, Excel (xls/xlsx), etc.
    - Use of appropriate functions like read_csv() or read_excel() to load the data into Python data structures like DataFrames.
- For Indexing and Selecting Data, Sorting Data:
    - Pandas provides powerful indexing and selection mechanisms using methods like .loc[] and .iloc[].
    - Sorting data can be done using the sort_values() function, specifying the column(s) to sort by.
- For Describing Attributes of Data, Checking Data Types:
    - Use the describe() method to generate descriptive statistics of the data.
    - Check data types of each column using the dtypes attribute of the DataFrame.
- For Counting Unique Values, Formatting Columns, Converting Variable Data Types:
    - Employ functions like value_counts() to count unique values in a column.
    - Convert variable data types using functions like astype() to cast one data type to another.

- Identifying Missing Values and Filling Them:
    - Pandas provides functions like isna() or isnull() to identify missing values.

    - Use methods like fillna() or dropna() to handle missing data by filling them with appropriate values or dropping rows/columns with missing values.

Applications:

- Data Analysis: These operations are fundamental to exploratory data analysis (EDA), a crucial step in any data science or machine learning project.

- Data Cleaning: Handling missing values, formatting data, and converting data types are essential steps in data cleaning pipelines.

Limitations:

- Performance: While Pandas and NumPy are powerful libraries, they may not be optimal for very large datasets due to memory limitations.
- Learning Curve: These libraries have a steep learning curve, especially for beginners, which might hinder the quick implementation of complex operations.

Conclusion:

In conclusion, the assignment showcased the effective use of Python libraries like Pandas and NumPy for diverse data manipulation tasks, from reading various formats to handling missing values. These operations are vital for data analysis, cleaning, and visualization, offering essential insights into structured data. Despite their benefits, it's important to recognize the limitations of these libraries, such as performance constraints with large datasets. Nevertheless, with a solid understanding of their theoretical framework, Pandas and NumPy empower proficient data handling, supporting informed decision-making across domains.