Assignment: 1

Problem Statement:

Perform the following operations using R/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

Software Library Package:

Python with pandas.

1. Theory:

The Pandas library in Python is a versatile tool for data manipulation and analysis. It offers intuitive data structures like DataFrames and Series, enabling easy organization and manipulation of data. With Pandas, tasks such as cleaning, filtering, grouping, and aggregation become streamlined. Its integration with other libraries like NumPy and Matplotlib enhances its capabilities for data analysis and visualization. Popular for its simplicity and extensive documentation, Pandas is essential for data scientists, analysts, and developers working with datasets of any size.

1.1 Methodology:

- Read data from various formats such as CSV using `pd.read_csv()` function.
- Perform indexing and selecting data using DataFrame indexing methods.
- Sort data using `sort_values()` function.

- Describe attributes of data using `describe()` function.

- Check data types of each column using `dtypes`.

- Count unique values of data using `nunique()` function.

- Convert variable data types using `astype()` function.

- Identify missing values using `isnull()` function.

- Fill in the missing values using `fillna()` function.

1.2 Advantages and Applications:

- Python's pandas library provides a comprehensive toolset for data manipulation and analysis.

- Suitable for handling various data formats and performing data preprocessing tasks.

- Widely used in data science, machine learning, and analytics projects.

- Offers flexibility and efficiency in data manipulation tasks.

1.3 Limitations:

- Performance may degrade with very large datasets.

- Requires familiarity with Python programming and pandas library.

- Handling complex data transformations may require additional coding and understanding of pandas functionalities.

2. Working/Algorithm:

- Read data from CSV file using `pd.read_csv()`.

- Index and select data using DataFrame indexing methods.

- Sort data using `sort_values()` function.

- Describe attributes of data using `describe()` function.

- Check data types using `dtypes`.

- Convert data types using `astype()` function.

- Identify missing values using `isnull()` function.

- Fill missing values using `fillna()` function.

3. Conclusion:

- Python's pandas library offers a powerful and flexible platform for data manipulation and analysis tasks.

- By leveraging pandas functionalities, tasks such as reading data from different formats, indexing, sorting, describing attributes, checking data types, counting unique values, converting data types, and handling missing values can be efficiently performed.

- These operations are fundamental for data preprocessing in various data science and analytics projects, facilitating further analysis and modeling.