Assignment No: -1

Operations on Data

Problem Statement: -

Download insurance dataset from following link.

Insurance Csv (kaggle.com)

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

Objective :

1) This assignment aims to introduce you to the Pandas library and its basic functions. The library provides functionality for reading different file formats such as CSV and Excel.

2) Additionally, it familiarizes users with data cleaning and preprocessing techniques.

3) Enhance our skills in handling data in various formats, improving our proficiency in data analysis and manipulation.

S/W Packages and H/W apparatus used: OS: Windows, Kernel: Python 3, Tools: Google Colab

Libraries and Packages Used: NumPy, Pandas

Theory:

1) Pandas is a **powerful** and **widely-used** open-source Python library for data manipulation and analysis.

2) It provides **easy-to-use** data structures and functions, making it an essential tool for working with structured data.

3) At the core of Pandas are two main data structures: **Series** and **DataFrame.**

4) A **Series i**s a one-dimensional labeled array capable of holding any data type .

5) **DataFrame** is a two-dimensional labeled data structure with columns of potentially different types.

6) These data structures allow users to perform a wide range of operations on data, including loading data from various file formats (such as CSV, Excel, SQL databases), manipulating data (e.g., sorting, filtering, grouping), and performing statistical and analytical tasks.

Methodology:

a) Reading Data from Different Formats:

Utilize libraries like Pandas to read data from CSV, Excel, or other formats into data structures like Data Frames.

b) Indexing, Selecting, and Sorting Data:

Use Pandas indexing methods like loc[] for label-based indexing and sorting functions like sort_values().

c) Describing Attributes of Data and Checking Data Types:

Leverage Pandas' describe() function for summary statistics and dtypes attribute for data type inspection.

d) Counting Unique Values, Checking Format, and Converting Variable Data Type:

Employ Pandas' functions like nunique() for counting unique values, dtypes attribute for format checking, and astype() for data type conversion.

e) Identifying Missing Values and Filling Them:

Use Pandas isnull() and fillna() functions to identify and fill missing values respectively.

Advantages and Applications:

These operations facilitate efficient data manipulation, preprocessing, and analysis, crucial for exploratory data analysis (EDA) and modeling tasks.

They enable researchers, data scientists, and analysts to gain insights, prepare data for modeling, and build predictive models for various domains like finance, healthcare, marketing, etc.

Limitations with Example:

One limitation could be the computational overhead and memory usage, especially with large datasets. For instance, sorting or indexing large datasets may require significant computational resources and time.

Working / Algorithm:

1. Read data from different formats using Pandas.
2. Perform indexing, selection, and sorting operations on the data.
3. Describe attributes and check data types of each column.
4. Count unique values, check formats, and convert variable data types as needed.
5. Identify missing values and fill them using appropriate methods.
6. Iterate through the dataset, applying these operations as necessary.

Conclusion:

The ability to perform diverse data operations using Python's Pandas library is indispensable for efficient data analysis and modeling. By mastering these operations, analysts can effectively preprocess data, uncover insights, and build robust predictive models, thus enabling data-driven decision-making and innovation across various domains.