Data Analysis and Preprocessing using Python: A Case Study of the Titanic Dataset

Introduction:

Data analysis uses raw data to find useful information. Data preprocessing is a key step to ensure data quality. This exercise shows data analysis and preprocessing using the Titanic dataset and Python Pandas.

1. Importing Libraries:

To work with data in Python, we use libraries. Pandas helps with table-like data, and NumPy handles number calculations.

2. Locating an Open Source Dataset:

The Titanic dataset from Kaggle (https://www.kaggle.com/c/titanic/data) has passenger information, like survival, class, age, and cabin. It's good for learning data analysis.

3. Loading the Dataset into a Pandas DataFrame:

We use Pandas to load the dataset from a file into a DataFrame, which is like a table. The pd.read_csv() function reads the train.csv file into a DataFrame.

4. Data Preprocessing: Checking for Missing Values and Initial Statistics:

Data preprocessing cleans and organizes data. Real-world data can be messy. We check for missing values and look at the data's basic statistics.

- Once the dataset is loaded, it's crucial to get an initial understanding of its characteristics.

5. Data Formatting and Data Normalization:

We make sure the data is in the right format (like text or numbers).

- Data Formatting (Data Types): Data formatting involves ensuring that data is stored in the correct data type
- Data Normalization: Data normalization is the process of scaling numerical data to a standard range.

Algorithm:

1. Import Libraries: Import pandas and numpy.

2. Load Data: Read train.csv into a Pandas DataFrame.
3. Explore Data: Use head(), isnull().sum(), describe(), info(), shape.
4. Impute 'Age': Fill missing 'Age' based on the mean age per 'Pclass'.
5. Drop 'Cabin': Remove the 'Cabin' column.
6. Drop Remaining NaNs: Remove any rows with remaining missing values.
7. Check Data Types: Display dtypes.

Conclusion:

This exercise covered basic data analysis and preprocessing using Python and Pandas. We explored the Titanic dataset, handled missing values, and prepared the data for modeling. Preprocessing ensures data quality. For deeper analysis, more techniques can be used.