Lab: Data Preprocessing

1. Load the dataset data.csv into a Pandas DataFrame. Display the first 10 rows of the dataset.
2. What are the data types of each column in the dataset? How can you change the data type of a specific column if needed?
3. Identify any missing values in the dataset. How many missing values are there in each column?
4. Impute the missing values in a numerical column using the mean of that column.
5. For a categorical column with missing values, fill them with the mode (most frequent value) of that column.
6. Drop any rows where more than 2 columns have missing values.
7. Check if there are any duplicate rows in the dataset. How many are there?
8. Remove all duplicate rows from the dataset.
9. Plot a boxplot for a specified numerical column to visually detect outliers.
10. Normalize a numerical column using Min-Max Scaling to bring its values between 0 and1.
11. Standardize another numerical column so that it has mean of 0 and standard deviation of 1.
12. Convert a categorical column with nominal data into numeric format using one-hot encoding.
13. Apply label encoding to another categorical column with ordinal data (e.g., 'low', 'medium', 'high').
14. Calculate the correlation matrix for all numerical columns in the dataset. Identify any pairs of features with a high correlation (e.g., above 0.8). Consider removing one of the correlated features.
15. Suppose you have two datasets: customer_info.csv and transaction_info.csv. Merge them on the CustomerID column. Display the merged DataFrame. (Create two small csv files having same column of CustomerID in both)
16. Create a scatter plot matrix (pairplot) for the numerical columns in the dataset to examine relationships between features. Provide title and legend to the plot.

Application: House Price Prediction data can be downloaded from Kaggle.

Objective: Preprocess a real estate dataset to predict house prices.

Step1: Load a real estate dataset with features like location, size, number of rooms, age of the house, etc.

Step2: Identify missing values in columns such as number of rooms and size. Apply appropriate imputation techniques.

Step3: Create new features such as price per square foot and age at sale.

Step4: Convert categorical variables such as location and property type into numeric values using label encoding.

Step5: Apply Min-Max scaling to features like size, age of the house, and number of rooms.

Step6: Use box plots to identify outliers in the price column.

Step7: Use scatter plots to visualize relationships between key features and the target variable (price). Provide title and legend to the plot.

Deliverable: A well-documented Python script that preprocesses the dataset, along with visualizations and a cleaned dataset ready for regression modeling.