Week 3. Lab 2: DataPreprocessing

Duration: Due date by 3pm

Objective:

Apply foundational data preprocessing techniques on the "house_prices.csv" dataset, focusing on detecting and addressing inconsistencies, invalid entries, duplicate records, data range issues, format inconsistencies, and missing values.

Instructions for Submission:

1. **Python Script (**.ipynb**):**
    - Ensure your script includes comments to explain the purpose and functionality of each code block.
2. **PDF Conversion:**
    - After completing the lab in a Python script, convert this file to a PDF that includes all code and output.
    - Ensure that the PDF is well-formatted, readable, and includes your name and student ID at the top.

Questions:

Project 1: Data Cleaning and Preprocessing

Objective: Clean the dataset by addressing missing values, outliers, and duplicate records.

Tasks:

1. **Identify and Remove Duplicates:**
    - Write code to detect and remove any duplicate records in the dataset.
2. **Handling Missing Values:**
    - Assess which columns have missing data and the proportion of missing data.
    - Apply different strategies to handle missing values:
        - For numerical variables with missing data, replace missing values using the mean or median.

        - For categorical variables, replace missing values with the mode or a specific category like "Unknown".
3. **Outlier Detection and Handling:**
    - Use the Z-Score method to identify outliers in numerical variables. Any data points that are more than 3 standard deviations from the mean should be considered outliers.
    - Apply Winsorization to cap extreme outliers at a specified percentile.

Project 2: Data Normalization and Standardization

Objective: Normalize and standardize numerical features to prepare for further analysis.

Tasks:

1. **Standardizing Features:**
    - Standardize the 'Size(sqft)' and 'Price' columns using z-score standardization so that they have a mean of 0 and a standard deviation of 1.
2. **Mean Normalization:**
    - Apply mean normalization to the 'Bedrooms' and 'Bathrooms' columns to scale the features to have a mean of 0.
3. **Min-Max Scaling:**
    - Scale the 'Year Built' column using Min-Max scaling to transform the data to a range of 0 to 1.

Project 3: Encoding Categorical Variables

Objective: Encode categorical variables using different strategies.

Tasks:

1. **Label Encoding:**
    - Apply label encoding to the 'Location' column where each unique location is assigned a unique integer.
2. **One-Hot Encoding:**
    - Perform one-hot encoding on the 'House_Type' column to create binary variables for each category.

Project 4: Multivariate Analysis

Objective: Conduct multivariate analysis to explore relationships between different features.

Tasks:

1. **Correlation Analysis:**
    - Generate a heatmap of correlations between all numerical features.
    - Discuss any significant correlations observed and potential reasons behind these relationships.
2. **Pairplot Analysis:**
    - Create pairplots to visualize pairwise relationships between the 'Size(sqft)', 'Bedrooms', 'Bathrooms', and 'Price'. Identify any patterns or clusters.

These projects will allow you to apply the concepts they have learned in class practically, reinforcing their understanding and preparing them for real-world data analysis tasks.

Deliverables:

- Upload your .ipynb and .pdf files to the designated Canvas submission area before the deadline.
- Ensure that both files are named appropriately (e.g., LastName.Week2-Lab1) and include necessary comments and documentation.