Title: Day 6 - Kaggle Competition Data Analysis and Outlier Handling

Introduction:

On Day 6 of the internship, the team engaged in a Kaggle competition, with team members assigned to specific roles, including a data provider, data analyzer, documenter, and note-taker. The competition involved working with the provided dataset to analyze and prepare the data for further modeling.

Tasks and Actions:

Data Provider:

The data provider's role was to supply the team with the necessary dataset for the Kaggle competition. Two datasets were loaded using Pandas: 'train.csv' and 'test.csv'. The 'train.csv' dataset was specifically used for analysis and model development.

Data Analyzer:

The data analyzer's role included exploring and analyzing the dataset. The following tasks were performed:

Column Renaming: The column 'KitchenQual' was renamed to 'QualKitchen' for clarity and consistency.

Box Plots for Outliers: Visual representations of outliers were created using box plots for attributes 'GarageArea' and 'WoodDeckSF'. This provided insights into the distribution and presence of potential outliers.

Handling Outliers: Outliers were detected and handled using the Z-score method. Z-scores were calculated for attributes 'Age' and 'WeekOfMonthClaimed,' and outliers were identified based on a specified Z-score threshold. The resulting dataset with outliers removed was stored in 'data_no_outliers' (not mentioned in code).

Categorical to Numerical Conversion: Categorical data transformation was performed by using LabelEncoder. The 'GarageCars' attribute was encoded to convert it into numerical form.

Correlation Analysis: Correlation between numerical attributes and the target variable 'SalePrice' was analyzed. This helped in understanding the relationship between variables.

Missing Values Analysis: The number of missing values in each column was counted to identify the presence of null values.

Data Types Identification: Data types of all columns were identified to distinguish between numerical and categorical attributes.

Handling Missing Values: Missing values in the 'LotFrontage' column were filled by imputing the mean value.

Documenter:

The documenter was responsible for documenting the tasks and actions taken by the team throughout the Kaggle competition. This documentation provides a clear record of the work performed, including data analysis, outlier handling, and data preprocessing.

Note Taker:

The note-taker's role was to take detailed notes on the tasks performed, including specific data transformations, methods used for handling outliers, and the results obtained during the analysis. These notes were valuable for tracking the team's progress during the competition.

Conclusion:

Day 6 was dedicated to participating in a Kaggle competition, involving data analysis and preparation. The team analyzed the dataset, identified and handled outliers, transformed categorical data into numerical form, and conducted correlation analysis. The documentation and detailed notes ensured that the team's work was well-documented and transparent.