Celebal Summer Internship Report

Project Title: House Price Prediction - Data Preprocessing & Feature Engineering

Domain: Data Science | Machine Learning

Tools Used: Python, Pandas, Matplotlib, Seaborn, Scikit-learn

Data Source: https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data

Objective

To perform data preprocessing and feature engineering on housing data in order to prepare it for machine learning models. The goal is to improve data quality, enhance predictive power, and create a reusable dataset for training regression models.

Dataset Overview

- train.csv - 1460 samples (with target 'SalePrice')

- test.csv - 1459 samples (no target)

- 79 features (numeric + categorical)

- Target: SalePrice (house price in USD)

Exploratory Data Analysis (EDA)

1. SalePrice Distribution: Right-skewed. Log-transformed to normalize.

2. Correlation Heatmap: Top 10 features like OverallQual, GrLivArea highly correlated.

3. Missing Values: Visualized and logically imputed.

Data Cleaning

- Filled categorical NAs with 'None'.

- Numeric NAs with median.

- Mode imputation for common categoricals.

- Group-wise imputation for LotFrontage using neighborhood median.

Feature Engineering

- Created TotalSF, TotalBath, HouseAge, RemodelAge, GarageAge.

- Dropped redundant original columns after transformation.

Handling Skew

- Applied log1p() transformation to skewed numeric columns (> 0.75).

Encoding

- One-hot encoded categorical features.

- Used scikit-learn OneHotEncoder with handle_unknown='ignore'.

Outputs

- X_train_preprocessed.csv

- X_test_preprocessed.csv

- y_train.csv

- All plots saved as .png files automatically.

Conclusion

This assignment focused on crafting a clean, informative dataset using preprocessing, logical imputation, outlier treatment, and smart feature engineering. The dataset is now ready for training high-performance regression models.