**House Pricing Project**

**1. Data Import and Exploration**

* Loaded the housing dataset using Pandas to begin the analysis.
* Explored the data structure by checking column types, dimensions, and summary statistics.
* Identified missing values and visualized the data distributions to understand patterns and inconsistencies.

**2. Data Cleaning**

* Replaced missing numerical values with their respective column averages to maintain consistency.
* Imputed categorical variables with the most frequently occurring value to handle missing data.
* Removed irrelevant columns and detected outliers based on statistical thresholds and domain knowledge.

**3. Feature Engineering**

* Added new features, such as TotalSquareFootage, to capture meaningful property characteristics.
* Transformed skewed numerical columns using log transformation to reduce the effect of extreme values.
* Encoded categorical columns using one-hot encoding to make them usable by machine learning models.

**4. Data Splitting**

* Divided the dataset into training (80%) and testing (20%) subsets to evaluate model performance on unseen data.

**5. Model Selection**

* Tested multiple regression models, including Linear Regression, Ridge Regression, Random Forest Regressor, and Gradient Boosting Regressor.
* Selected the best model based on metrics such as R-squared and error values.

**6. Model Training**

* Trained the selected model on the training dataset.
* Used GridSearchCV to fine-tune hyperparameters and optimize model performance.

**7. Evaluation Metrics**

* Assessed the model using:
  + Mean Absolute Error (MAE)
  + Mean Squared Error (MSE)
  + Root Mean Squared Error (RMSE)
  + R-squared (R²) score

**8. Model Interpretation**

* Examined feature importance using the Random Forest model to identify key variables influencing housing prices.
* Visualized and interpreted these insights to provide actionable findings.

**9. Prediction and Visualization**

* Generated predictions on the test dataset.
* Created a scatter plot to compare actual prices with predicted prices.
* Analyzed residuals to understand the error distribution of the model.