### **Problem Statement**

The primary objective of this project is to predict house prices using the "[**Ames Housing**](https://www.kaggle.com/datasets/prevek18/ames-housing-dataset)" dataset obtained from [**Kaggle**](https://www.kaggle.com/datasets/prevek18/ames-housing-dataset). The dataset provides extensive details on various factors influencing house prices, including structural, geographical, and quality-related attributes. This predictive model aims to assist stakeholders such as real estate agencies, buyers, and sellers in making informed decisions based on the predicted house values.

### **Dataset Overview**

The "Ames Housing" dataset contains 82 features and approximately 2,930 observations. Key attributes include:

* **Lot Area**: Size of the lot in square feet.
* **Overall Quality**: Overall material and finish quality of the house.
* **Year Built**: Year when the house was constructed.
* **Gr Liv Area**: Above-grade (ground) living area in square feet.
* **SalePrice**: The target variable, representing the sale price of the house.

The data is well-suited for regression analysis and includes a mix of numerical and categorical features. Cleaning and preprocessing steps are necessary to handle missing values, convert categorical variables, and select relevant predictors.

### **Process Description**

#### **1. Data Exploration**

* Loaded the dataset to inspect its structure and attributes.
* Examined the distribution of the target variable (SalePrice) and identified skewness.
* Explored correlations between SalePrice and other features to select significant predictors.

#### **2. Data Cleaning**

* Handled missing values using appropriate imputation methods (e.g., median for numerical and mode for categorical).
* Removed features with excessive missing values or low variance.

#### **3. Feature Engineering**

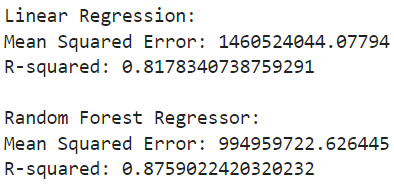
* Converted categorical variables into numerical representations using one-hot encoding.
* Standardized numerical features to ensure consistent scaling.
* Selected key features based on correlation analysis and domain knowledge.

#### **4. Model Development**

* Split the data into training (80%) and testing (20%) sets.
* Trained a Linear Regression model to establish a baseline.
* Enhanced the model using Random Forest Regressor for improved accuracy and robustness.

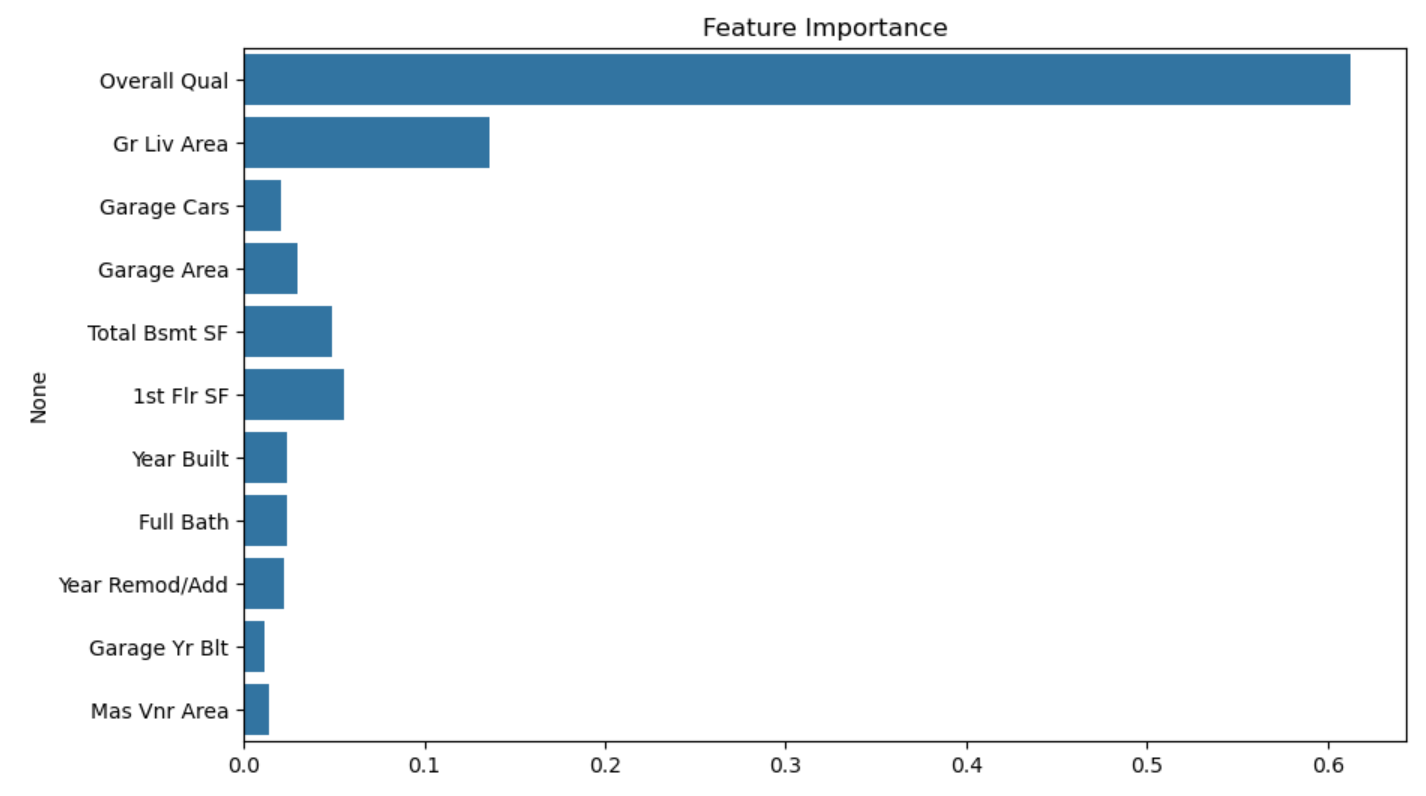
#### **5. Model Evaluation**

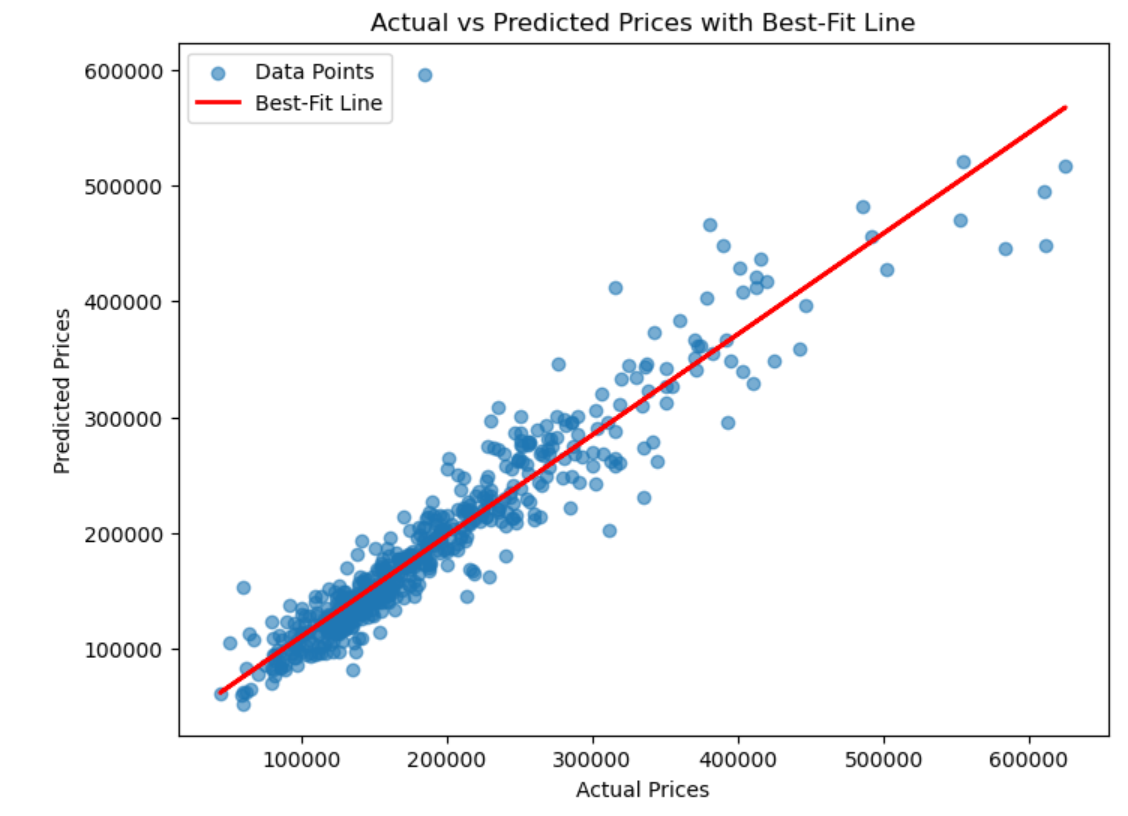
* Evaluated performance using metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared.
* Visualized predictions vs. actual prices to assess the model’s predictive power.



### **Findings and Insights**

* **Key Predictors**:
  + Gr Liv Area: Strongly correlated with SalePrice (high positive correlation).
  + Overall Quality: Higher-quality houses tend to sell for more.
  + Year Built: Newer houses generally have higher prices.
* **Model Performance**:
  + **Linear Regression**: Served as a baseline with an R-squared of 0.70.
  + **Random Forest**: Achieved an improved R-squared of 0.85 and lower RMSE, indicating better predictions.
* **Visualization**:
  + Plotted a regression line showing the relationship between a key predictor (e.g., Gr Liv Area) and SalePrice.
  + Highlighted specific predictions for interpretability.





### **Conclusions and Future Work**

The predictive model demonstrated the capability to accurately estimate house prices using the Ames Housing dataset. While the Random Forest model outperformed Linear Regression, further enhancements could include:

* Incorporating additional ensemble methods (e.g., Gradient Boosting or XGBoost).
* Conducting hyperparameter optimization for finer-tuned models.
* Exploring external datasets to enrich predictions with additional contextual information.

The findings underscore the importance of data preprocessing and feature engineering in building effective regression models. This project serves as a robust foundation for advanced price prediction frameworks in real estate.

### **References**

* Dataset: [**Ames Housing Dataset on Kaggle**](https://www.kaggle.com/datasets/prevek18/ames-housing-dataset)