Prediction of the Sales Price for Houses based on the Ames Housing Dataset

1. Project Description

The objective of this project is to develop a predictive model capable of estimating the sale prices of single-family homes based on various property features. The project is based on the Ames Housing Dataset, which is widely used in machine learning analyses and real estate market research.

The project aims to apply regression techniques to predict the SalePrice — the final sale price of homes. The process includes the following steps: data cleaning, feature preparation, exploratory data analysis, model training, evaluation, and drawing conclusions. The ultimate goal is to create an effective model that can accurately forecast house prices and understand which features have the most significant impact on property valuations in Ames.

2. Dataset Description

The dataset used in this project is the Ames Housing Dataset, a comprehensive collection of residential property records from Ames, Iowa, USA. It is widely recognized in the machine learning community for benchmarking regression models and developing predictive analytics related to real estate pricing.

Key characteristics of the dataset include:

- Number of Instances: 2,930 observations.
- Number of Features: 79 variables describing various attributes of the houses.
- Target Variable: SalePrice, representing the final sales price of each house.

Data Types: The dataset includes both numerical and categorical variables. These cover a broad range of property aspects such as lot size, number of bedrooms, quality of construction, type of exterior material, neighborhood location, and year built.

The richness and diversity of these features make this dataset an ideal candidate for applying regression techniques and exploring feature importance in house price prediction tasks. Additionally, the mix of numerical and categorical data provides opportunities for practicing feature engineering, data transformation, and model optimization.

3. Data Cleaning and Data Preparation

To The data cleaning and preparation process involved several important steps to improve the quality and usability of the dataset before modeling. The following actions were taken:

- **Handling Missing Data:**
  Columns with 40% or more missing values were removed from the dataset. Such columns typically do not provide sufficient information for meaningful analysis and could introduce noise or bias into the model. Removing them helped ensure that the dataset remained reliable and easier to work with.
- **Managing Columns with a High Proportion of Zero Values:**
  Columns with more than 40% zero values were carefully analyzed. Rather than removing all such columns immediately, several strategies were applied to preserve useful information:
  - Porch-related columns were combined into a new feature, Total Porch SF, representing the total area of all porches.

  - The basement finished area columns, BsmtFin SF 1 and BsmtFin SF 2, were summed into a new column, Total BsmtFin SF, capturing the overall finished basement space.
  - The Fireplaces column, indicating the number of fireplaces, was transformed into a binary variable (True if at least one fireplace was present, otherwise False), focusing on the presence rather than the count.

  These transformations reduced sparsity and created more meaningful features for modeling. The remaining columns with a high proportion of zeros, which could not be effectively combined or reinterpreted, were removed from the dataset.

- **Removing Columns with Little Variability:**
  Columns where more than 90% of the entries had the same value were dropped. Such features provided minimal variability and were unlikely to contribute useful predictive information, so their removal helped simplify the dataset.
- **Imputing Remaining Missing Values:**
  For the remaining missing values, different strategies were applied depending on the type of variable:
  - For numerical features, missing values were filled with the mode (most frequent value) to avoid distorting the distribution of data.
  - For categorical features, missing values were replaced with the label 'missing', allowing the model to treat missing data as an informative category without losing any records.

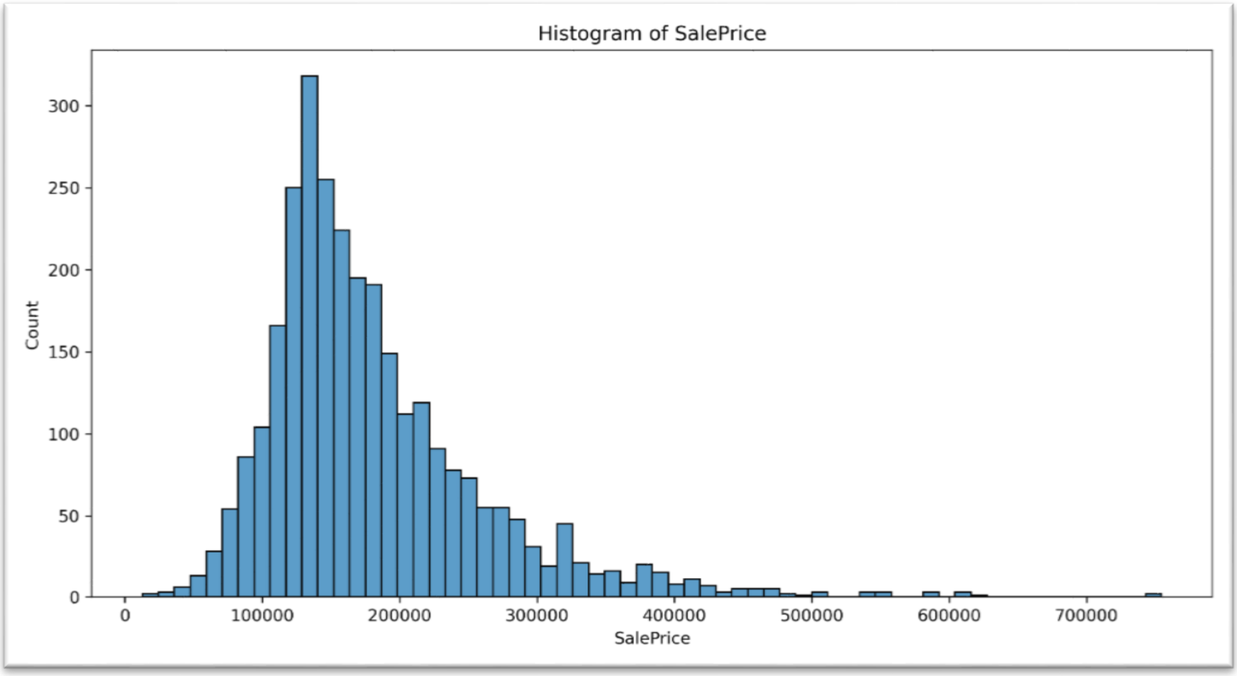
These preprocessing steps were essential to prepare the dataset for feature engineering, model training, and evaluation, ensuring that the predictive modeling process would be as robust and accurate as possible.

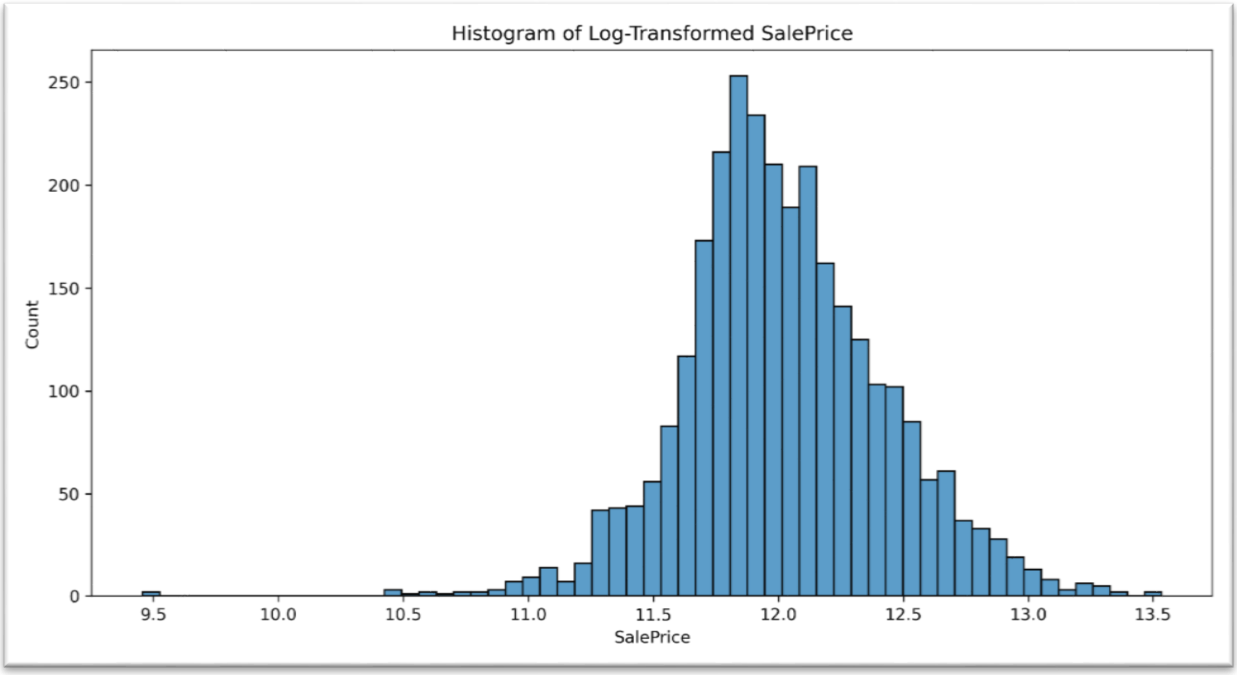
4. Data Exploration and Feature Engineering

During the data mining and feature engineering phase, the focus was on understanding the structure of the dataset and preparing features for modelling in a way that improves predictive performance.

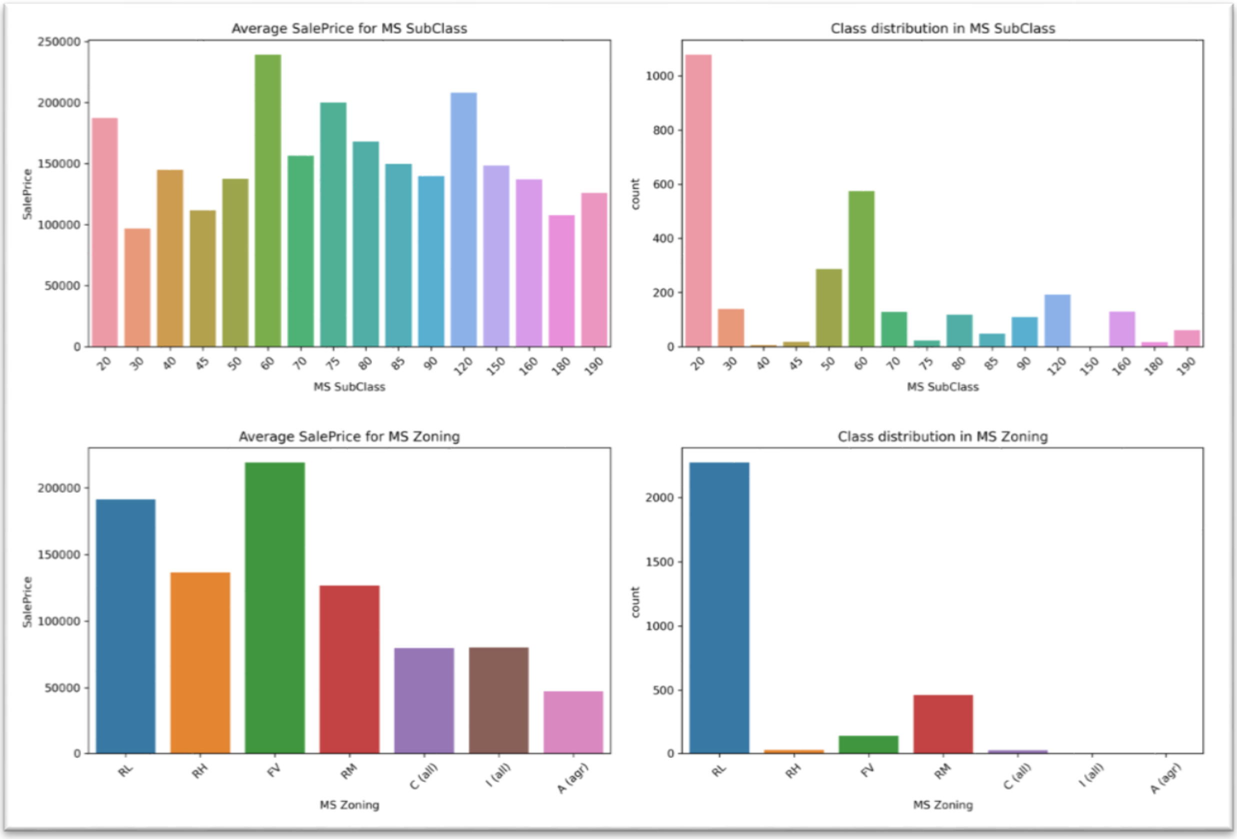
Key actions included:

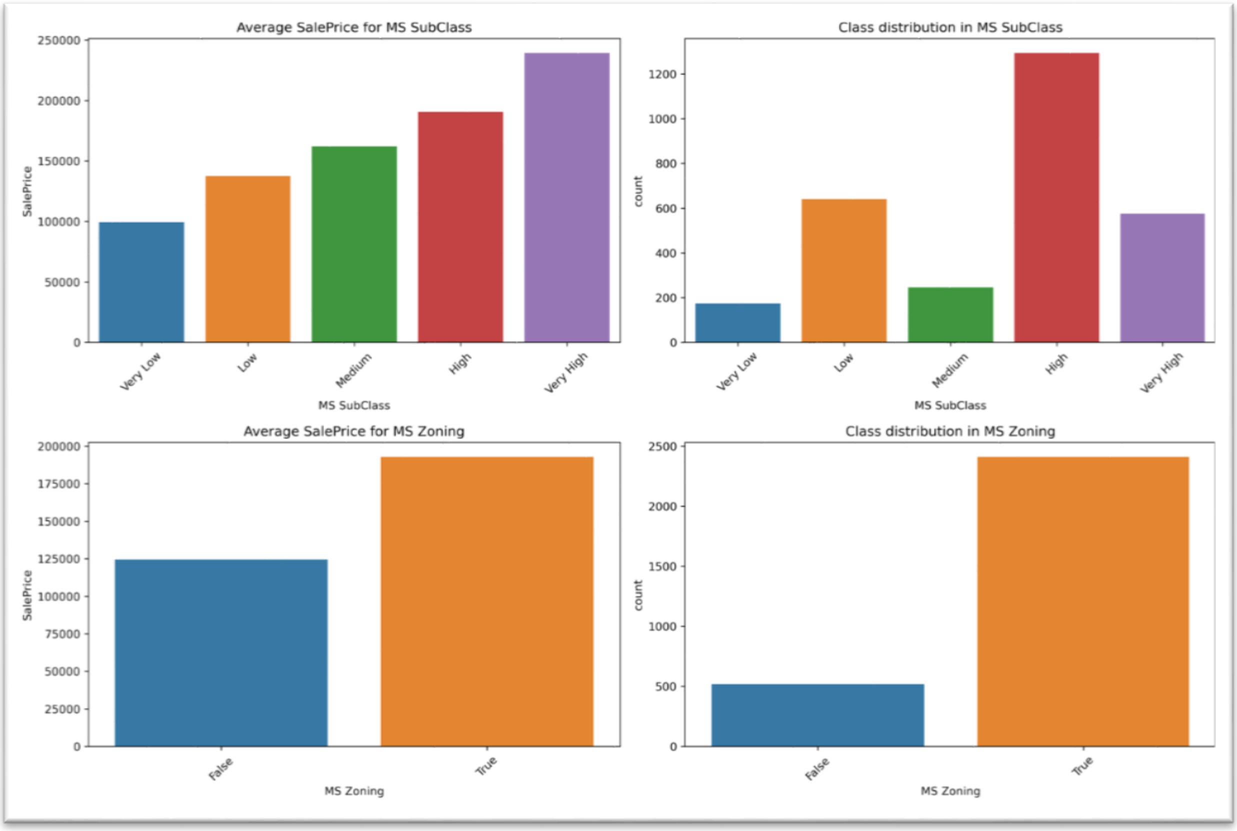
- **Target Variable Analysis:**
  The distribution of the target variable, SalePrice, was analyzed and found to be significantly right-skewed. To normalize its distribution, a logarithmic transformation was applied.





- **Feature Reduction and Grouping:**
  In categorical variables, all of categories were significantly reduced. Similar or rarely occurring classes were grouped together to simplify the feature space and avoid overfitting.

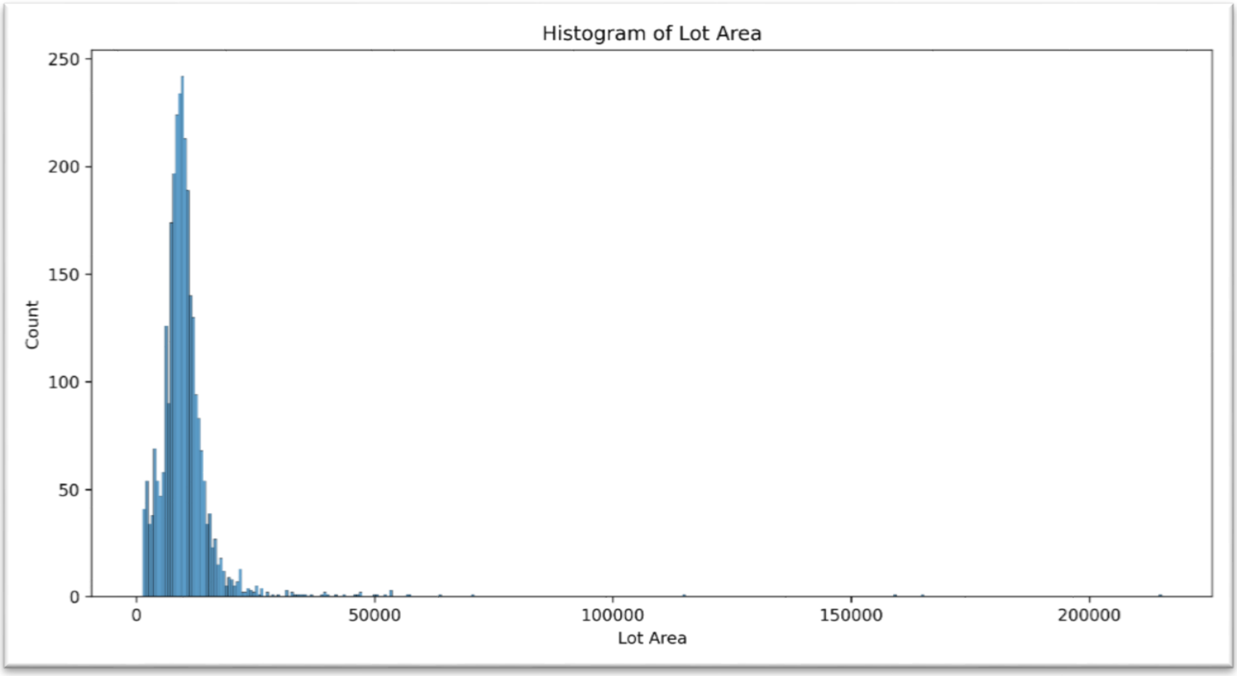


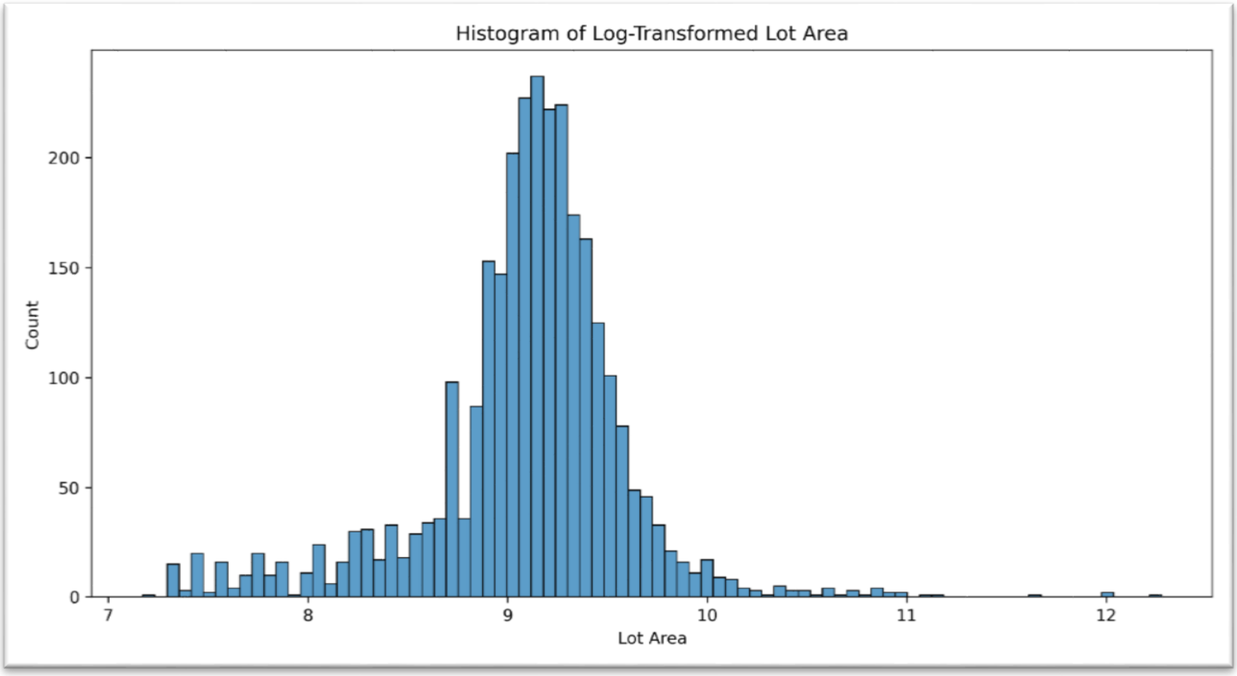


- **Feature Transformation for Numerical Variables:**
  Several numerical features representing years: YearBuilt, YearRemodAdd,

GarageYrBlt were transformed into "age" features by calculating the difference between the maximum year in the dataset and the feature year. This transformation helped capture the aging effect, which is often more meaningful in real estate valuation than absolute years.

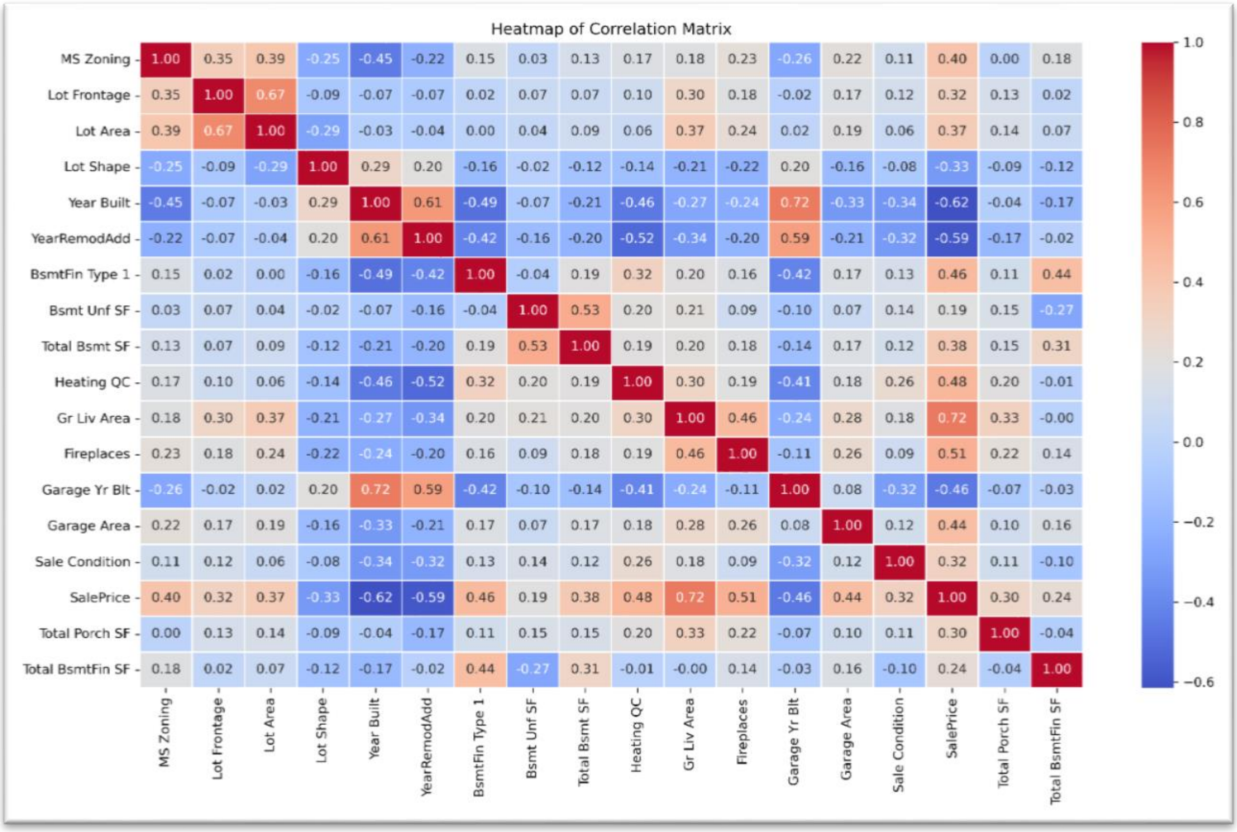
- **Logarithmic Transformations for Selected Features:**
  In addition to SalePrice, other highly skewed numerical variables were log-transformed to reduce skewness and stabilize variance, facilitating better model learning.



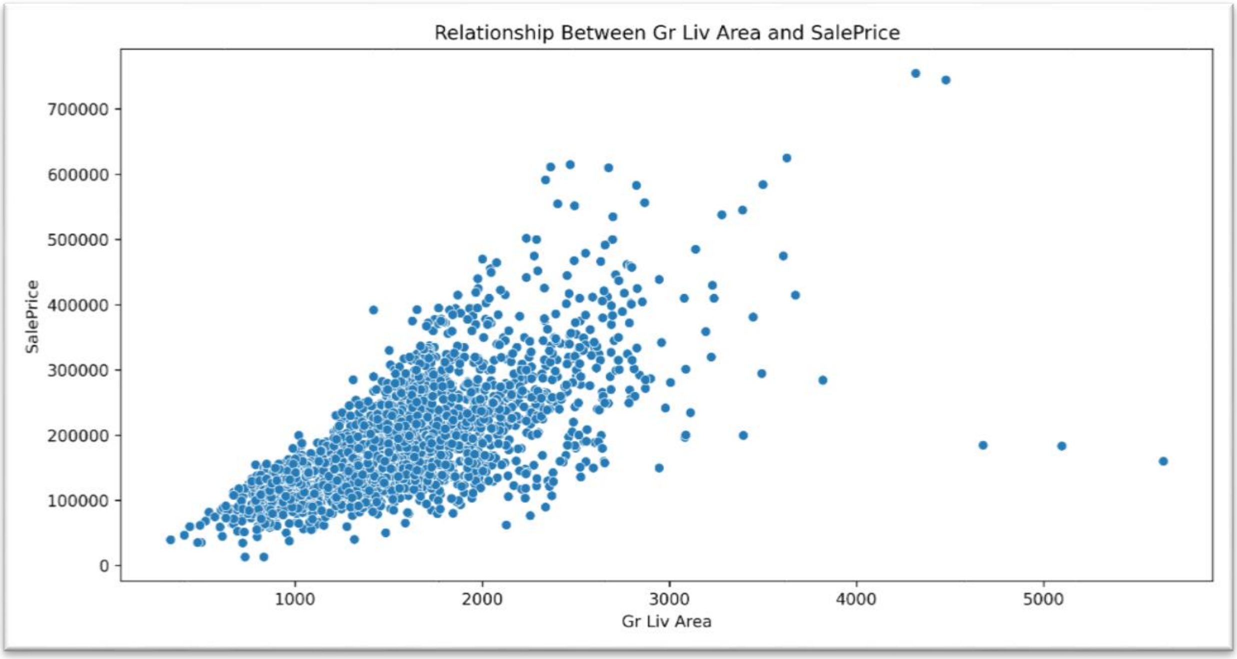


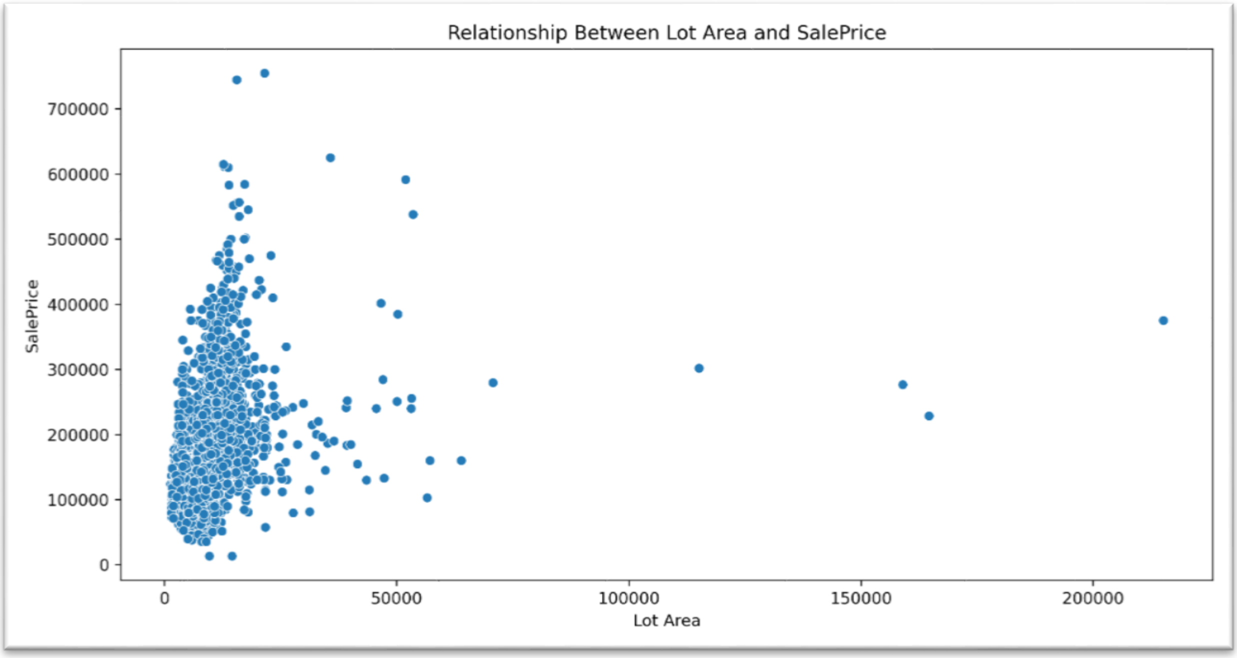
- **Correlation Analysis:**
  A correlation heatmap was created to visualize the relationships between all columns

in the dataset and to examine how strongly each feature is correlated with SalePrice. The heatmap made it easy to identify which variables have the strongest correlations with SalePrice, highlighting the key features that influence house prices. The strongest positive correlations were observed with variables such as OverallQual, GrLivArea, GarageCars, and TotalBsmtSF.



- **Relationship with SalePrice:**
  Key numerical features showing strong correlations with SalePrice were further explored using scatter plots to visualize the nature of their relationships.





Through these targeted exploration and feature engineering steps, we enhanced the dataset's structure and increased its suitability for regression modeling, while keeping the number of features manageable and meaningful.

5. Model Training

To identify the most effective regression algorithm for predicting house prices, five models were trained and evaluated:

- Linear Regression,
- Lasso Regression,
- Random Forest Regressor,
- XGBoost Regressor,

- K-Nearest Neighbors Regressor.

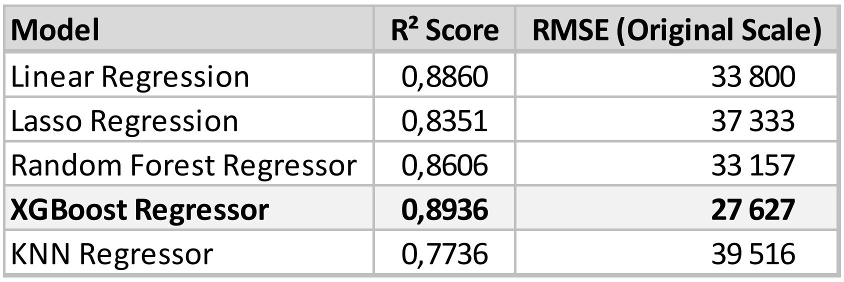
All models were trained using a consistent pipeline:

- The dataset was first one-hot encoded to handle categorical variables.
- The data was split into a training set (80%) and a test set (20%).
- A grid search with 5-fold cross-validation was used for hyperparameter tuning, optimizing for the R² score.

Hyperparameter tuning grids were defined for each model, and the best parameters were selected based on performance. The GridSearchCV routine was applied in a uniform way to allow a fair comparison across all models.

6. Model Evaluation

The performance of each model was evaluated on the test set using both R² score and Root Mean Squared Error (RMSE) on the original (exponentiated) scale of SalePrice, as log transformation had been applied earlier.



Model	R² Score	RMSE (Original Scale)
Linear Regression	0,8860	33 800
Lasso Regression	0,8351	37 333
Random Forest Regressor	0,8606	33 157
XGBoost Regressor	**0,8936**	**27 627**
KNN Regressor	0,7736	39 516

XGBoost outperformed all other models, achieving the highest R² and the lowest RMSE. This indicates strong predictive performance and robustness to overfitting.

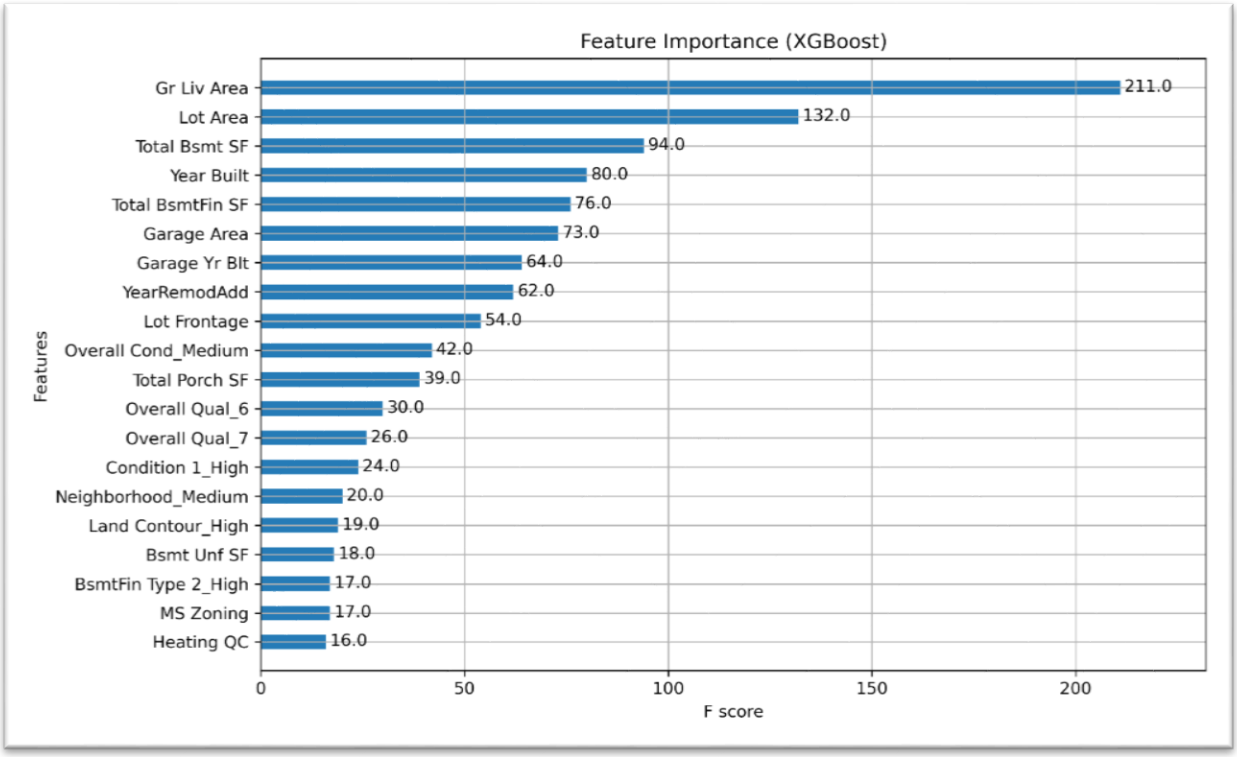
In contrast, KNN and Lasso showed relatively weaker performance, particularly in handling complex relationships within the data.

7. Conclusions

This project demonstrates the value of comprehensive data preprocessing and advanced regression techniques in accurately predicting housing prices. Through careful cleaning, feature engineering, and model selection, a robust predictive pipeline was established.

Key findings include:

- Data transformations such as log-scaling of target variables and feature simplification (e.g., reducing category cardinality) significantly improved model performance.
- XGBoost proved to be the most effective model, balancing both accuracy and generalization.
- Variables such as GrLivArea and Lot Area emerged as the most influential predictors, as shown in the chart below.


Feature Importance (XGBoost)
Features
Gr Liv Area
Lot Area
Total Bsmt SF
Year Built
Total BsmtFin SF
Garage Area
Garage Yr Blt
YearRemodAdd
Lot Frontage
Overall Cond_Medium
Total Porch SF
Overall Qual_6
Overall Qual_7
Condition 1_High
Neighborhood_Medium
Land Contour_High
Bsmt Unf SF
BsmtFin Type 2_High
MS Zoning
Heating QC
211.0
132.0
94.0
80.0
76.0
73.0
64.0
62.0
54.0
42.0
39.0
30.0
26.0
24.0
20.0
19.0
18.0
17.0
17.0
16.0
0
50
100
150
200
F score