
House Price Prediction Analysis

Sikandar Shrestha
Dulce X. Cid Sanabria

Data Source link:
<https://www.kaggle.com/datasets/zafarali27/house-price-prediction-dataset/data>

Problem Statement & Goals

Context: Real estate markets are complex economic indicators. Accurate pricing models are essential for buyers, sellers, and economists.

Goal: Predict house price using available structural and categorical features.

Objective: Compare different ML regression models to identify the key predictors and achieve the most accurate price prediction possible within the limitations of the dataset..



Dataset Overview

Source

Kaggle - House Price Prediction Dataset

<https://www.kaggle.com/datasets/zafarali27/house-price-prediction-dataset>
/data

Size

2,000 observations × 10 variables

Target variable

Price

Key predictors

Numeric: Area, YearBuilt, Id

Categorical: Bedrooms, Bathrooms, Floors, Location, Condition, Garage

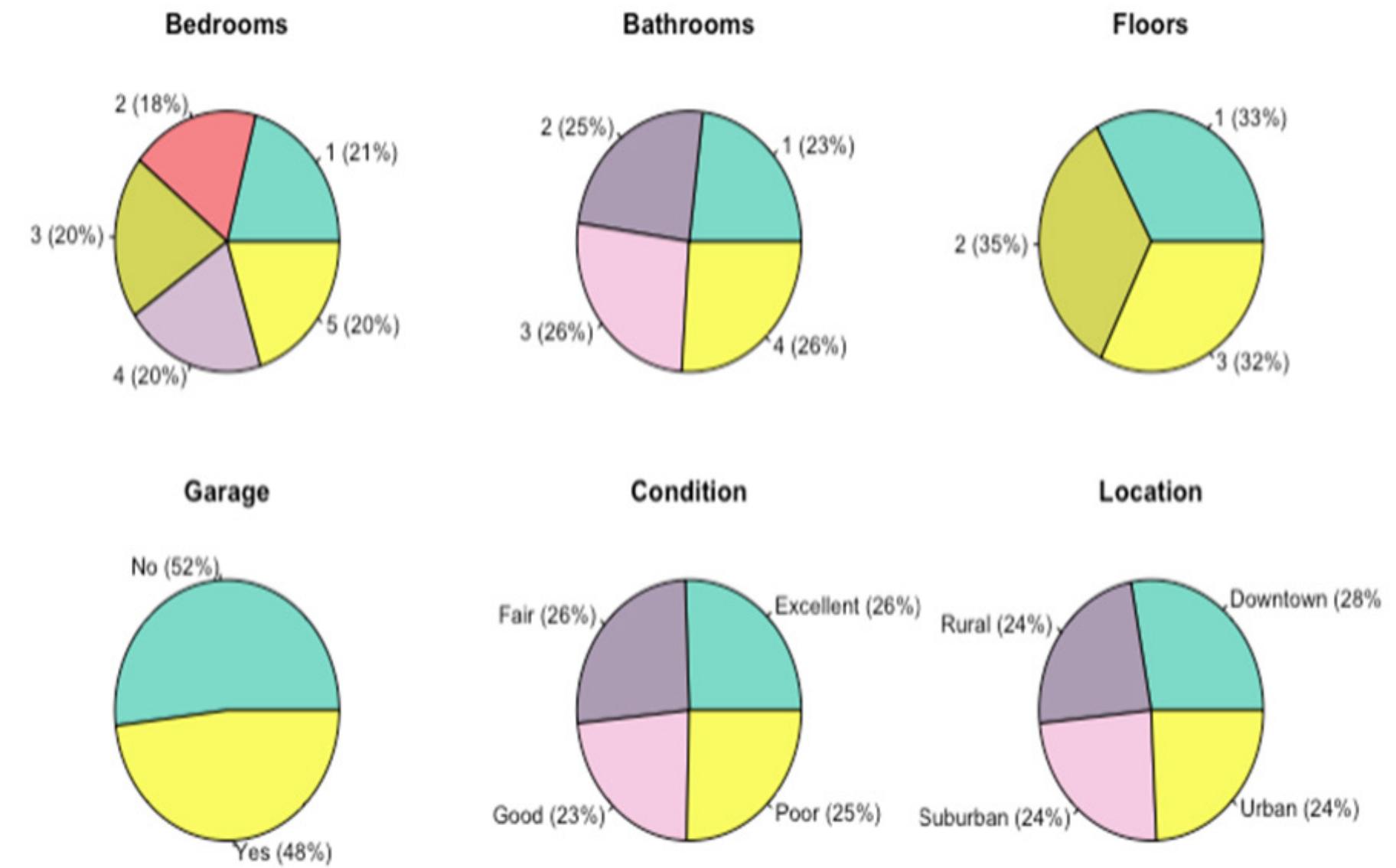
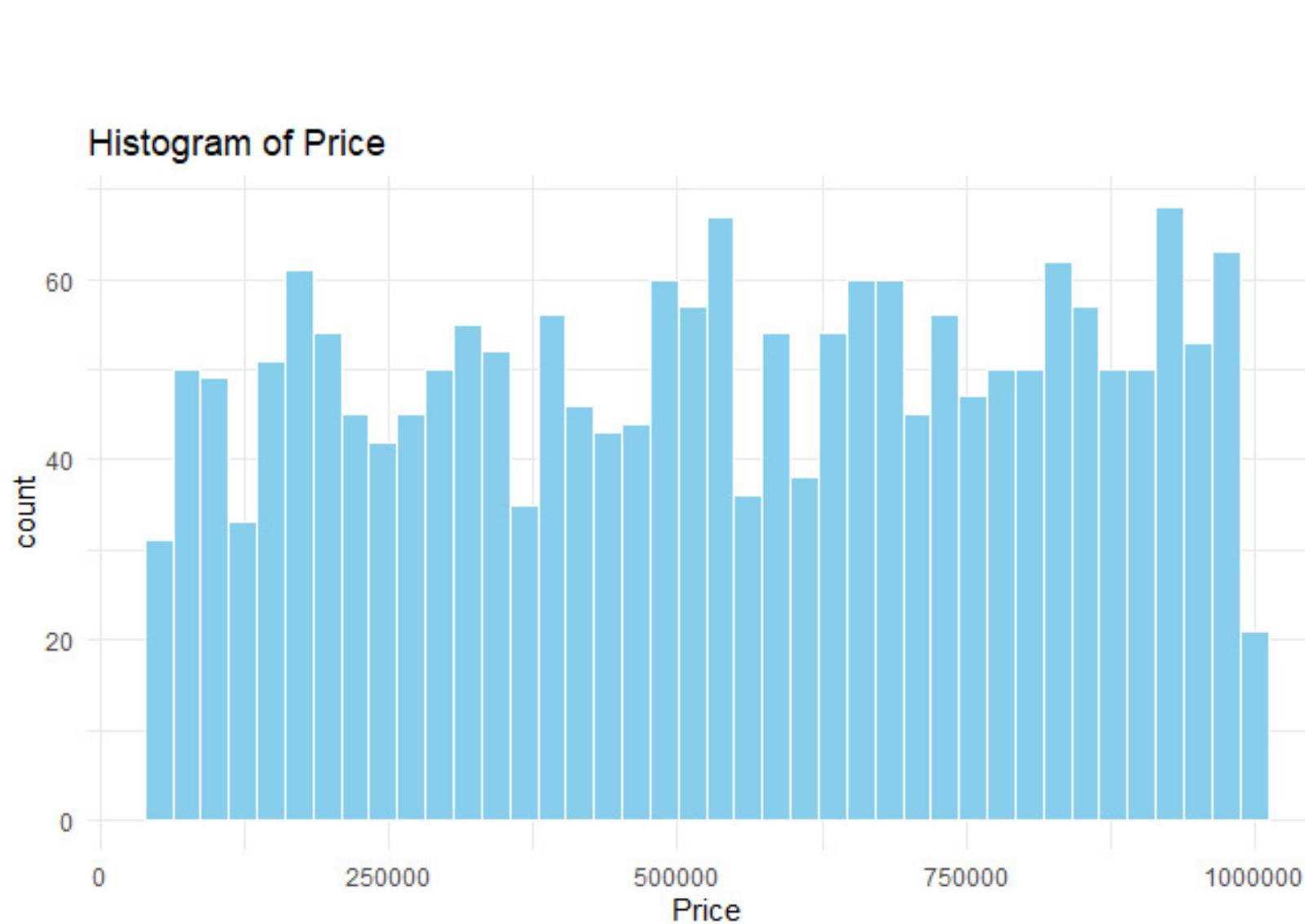
Preprocessing

- No missing values in the dataset
- Skewness analysis revealed symmetric distributions
- Drop ID column (non-informative variable)



Exploratory Data Analysis (EDA)

Distribution analysis

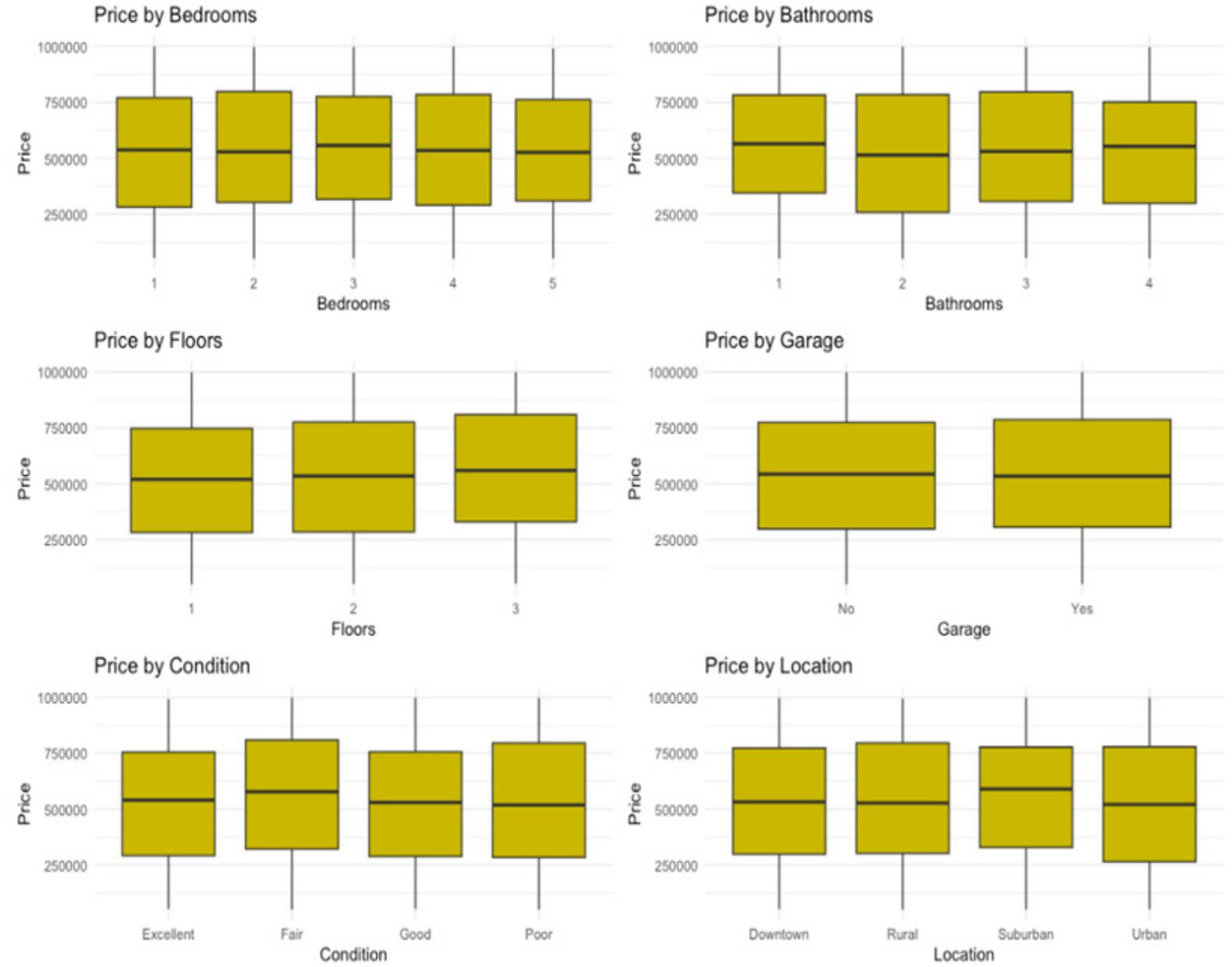


The distribution of home Price is fairly uniform across the full range (from budget-friendly to premium luxury), with a slight linear trend.

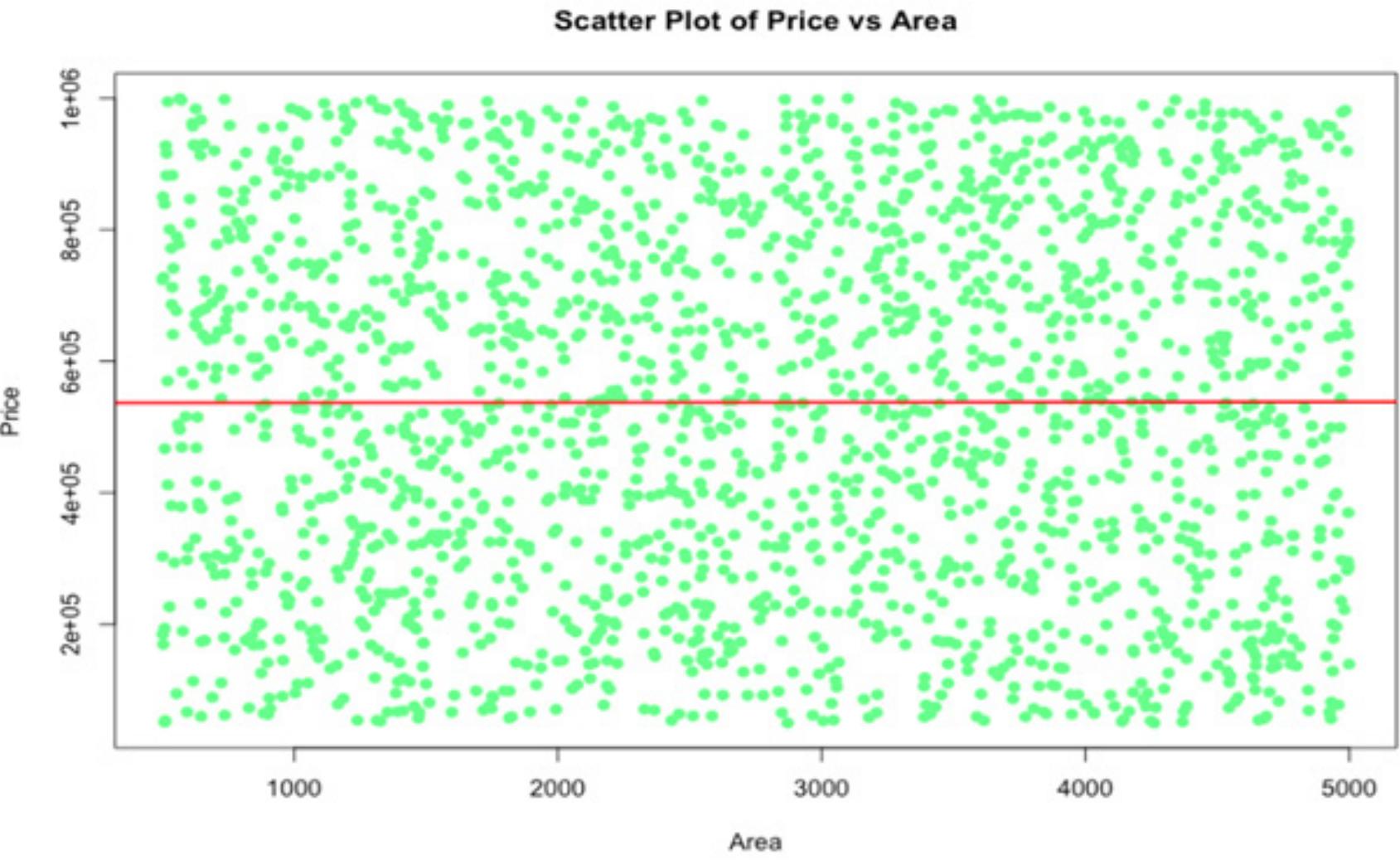
The dataset is well-balanced and representative across the categorical features. There are no extreme imbalances in any category.

Exploratory Data Analysis (EDA)

Bivariate analysis



Price distributions appear consistent across all groups, with similar medians and interquartile ranges, suggesting that these variables have not a visible influence on Price distribution



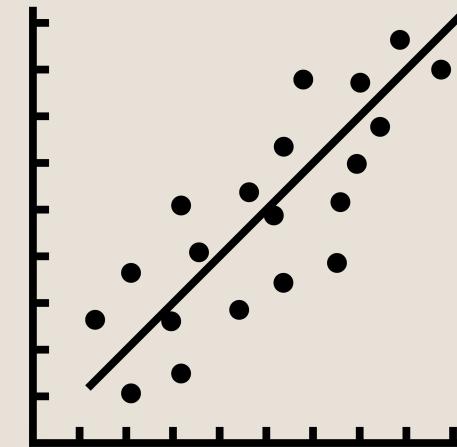
The points are spread uniformly across all area values with no visible trend. The regression line is horizontal, indicating that Area has no visual linear effect on Price.

Methodology

The dataset was split into 70% for training and 30% for testing.

Linear Models

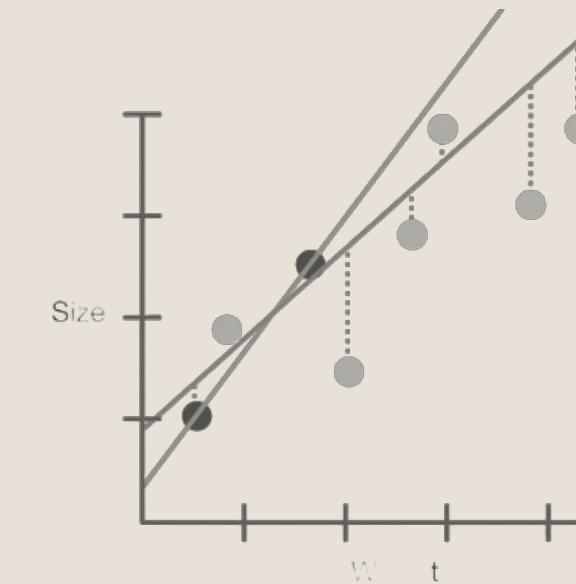
- Linear Regression
- Log-Linear Regression



Evaluation Metrics

Regularized Regression

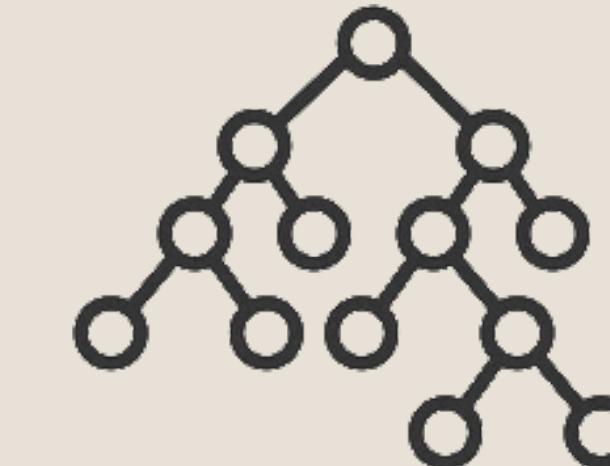
- Ridge Regression
- LASSO Regression
- Elastic Net Regression



RMSE & MAE – lower values indicate better predictive accuracy
 R^2 – higher values indicate stronger model fit

Ensemble Methods

- Random Forest
- Gradient Boosting



Model Comparison

Model <chr>	RMSE <dbl>	Rsquared <dbl>	MAE <dbl>
Elastic Net	277342.4	1.516719e-03	239762.3
LASSO	277343.7	1.519818e-03	239761.1
Ridge	277528.1	6.966896e-04	239967.2
Gradient Boosting	277533.5	2.078314e-03	239977.4
Linear	278576.5	2.588715e-03	239451.0
Random Forest	284466.1	1.579524e-05	243657.5
Log-Linear	292810.6	5.104606e-03	249686.0
Log-Linear + Interactions	294868.4	3.926175e-03	251405.9

Despite generally weak overall performance, **Elastic Net** was selected as the best-performing model because:

- It achieved the **lowest RMSE and MAE**, while also addressing multicollinearity and enabling automatic feature selection.
- Its structure remains close to a linear model, improving interpretability compared to more complex approaches.
- Although improvements over LASSO and Ridge are modest, it provides a balanced compromise and handles correlated predictors more effectively

However, **low R²** values and relatively high errors indicate limited overall predictive performance for this dataset. Best model still performs poorly

Prioritize Data Improvement:

Given the low model performance, future efforts should focus on acquiring richer and more representative data, particularly property characteristics and market conditions.

Reevaluate Dataset Validity:

The low R^2 values suggest that the current dataset may be noisy, incomplete, or insufficiently representative of real housing market dynamics.

Reconsider Problem Formulation:

Alternative approaches, such as predicting price ranges, may be more suitable given the limitations of the current data.

Communicate Model Limitations:

Clearly communicate to stakeholders that, given the current data and model performance, the model has limited value for accurate price prediction.

Recommendations / Next Steps



Conclusion

Elastic Net was chosen primarily for its stability and robustness rather than its predictive accuracy.

Overall, the findings indicate that the dataset is unsuitable for robust house price prediction.

Meaningful practical application would require more informative and representative variables, or a reassessment of the data generation process, to ensure realistic and reliable relationships between housing prices and explanatory features.

