**House Price Prediction Using Regression Techniques**

**Authors:** Swetha Manupati, Varshini Sherapur Basavarajappa, Vinisha Recharla Purushotham, Shyam Sundar Theerdhala, Anirudh Tedlapally

**2. Introduction and Problem Statement :**

Buying a house is a significant financial decision, and the housing market can be complex and unpredictable due to factors such as property size, location, and amenities. This project aims to build a model to predict house prices accurately based on these factors using regression techniques. This model will help:

Homebuyers make informed purchasing decisions, Sellers and developers set competitive prices based on market trends, The market by promoting transparency and understanding of the main factors influencing house prices.

**3. Background, Motivation, and Significance :**

The housing market is challenging to navigate, and price transparency is often limited. Homebuyers face the risks of overpaying, while sellers and developers need data-driven tools to price properties competitively. This model uses real data to provide reliable price predictions, thereby aiding all parties involved in housing transactions. Its benefits include:

Supporting informed decision-making for homebuyers and sellers.

Allowing sellers to set competitive prices with confidence.

Enhancing transparency by identifying the most impactful factors affecting house prices.

**4. Research Questions :**

The project addresses the following research questions:

1. What are the primary factors that significantly influence house prices?

2. How accurately can regression techniques predict housing prices?

3. Which regression model Linear, Polynomial, Ridge, or Lasso achieves the best balance of accuracy and generalizability?

**5. Dataset Overview :**

Dataset Name: Housing Price Prediction

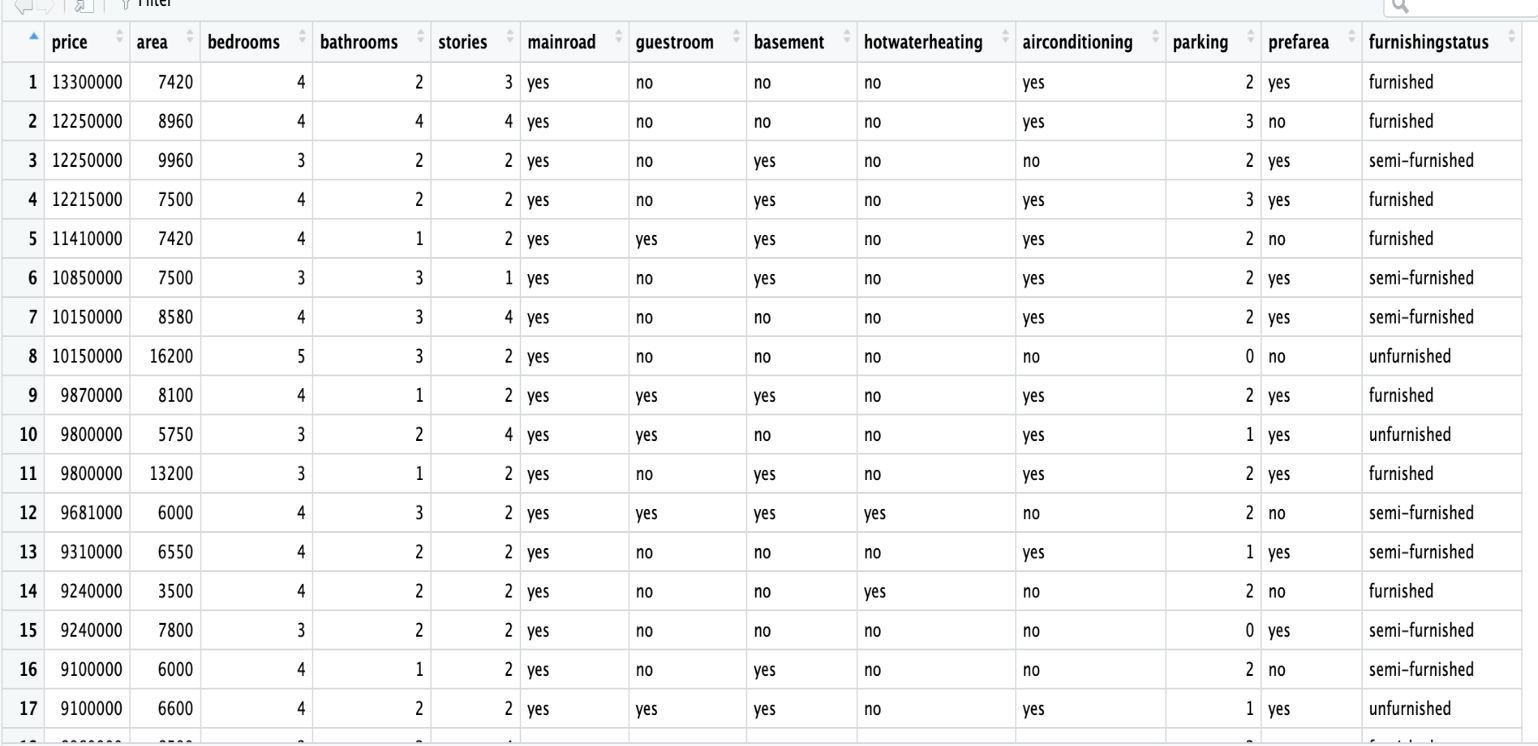
Source: [Housing Price Prediction| Kaggle](https://www.kaggle.com/datasets/harishkumardatalab/housing-price-prediction)

Sample Size: 545 records with 13 variables (1 dependent, 12 independent)

**Target Variable:** Price (House price to predict)

**Independent Variables:**  
**Numerical Variables:** Area, Bedrooms, Bathrooms, Stories, Parking.

**Categorical Variables:** Mainroad, Guestroom, Basement, Hot Water Heating, AirConditioning, Prefarea, Furnishing Status.



**Variable Types and Distributions:**

Categorical Variables: Yes/No or multiple levels Converted to numerical format for analysis.

Numerical Variables : continuous and discrete values Standardised to enhance model performance.

**6. Methodology**

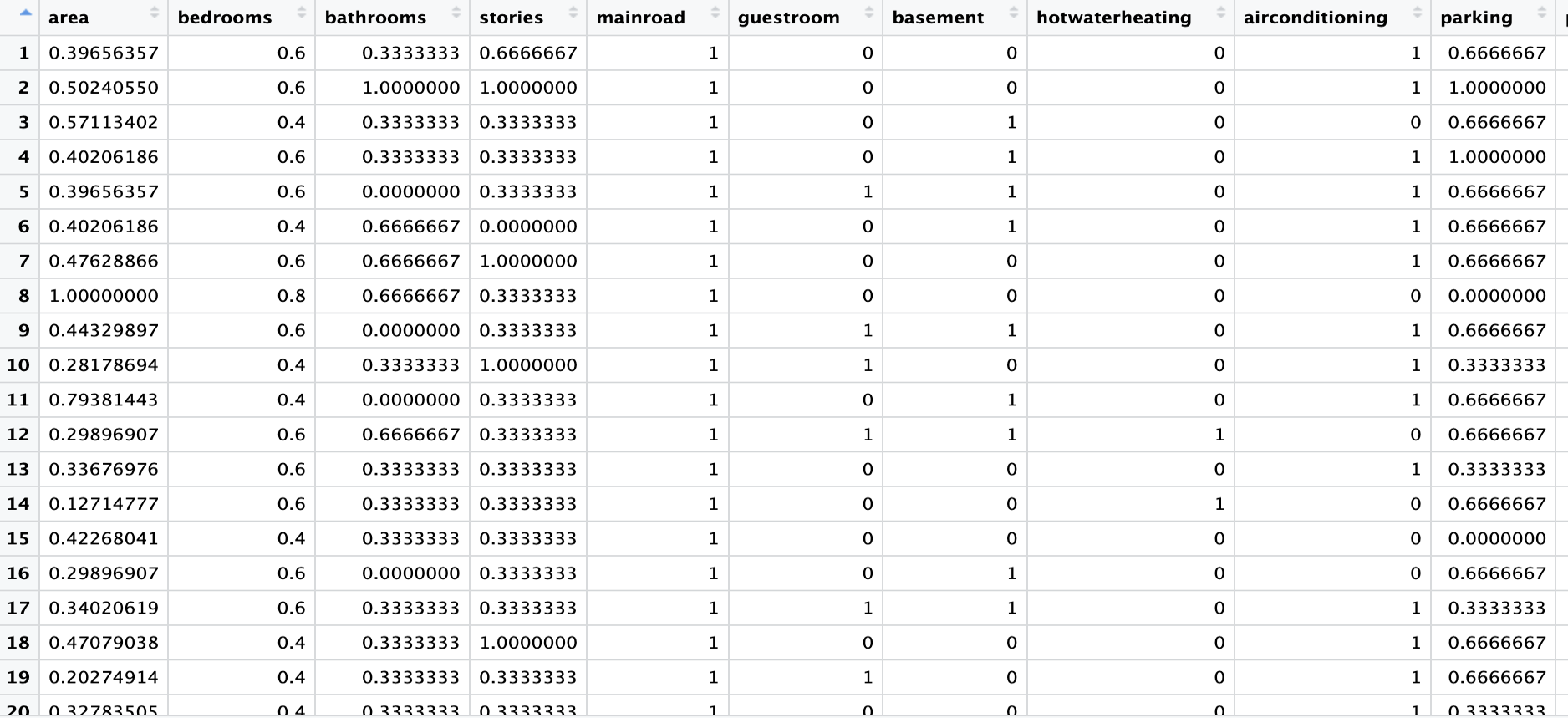
**6.1 Data Preprocessing**

**Data Cleaning:** The dataset was checked for null values, duplicates, and inconsistencies, all of which were not present, ensuring data accuracy.

**Encoding:** Binary categorical variables (e.g: Mainroad, Guestroom) were encoded (Yes = 1, No = 0), and multi-level categories for Furnishing Status were also numerically encoded.

**Standardization and Scaling:** Continuous features like Area and Bedrooms were standardised to maintain consistent scale, which improves the performance of models like Ridge and Lasso that are sensitive to variable scales.

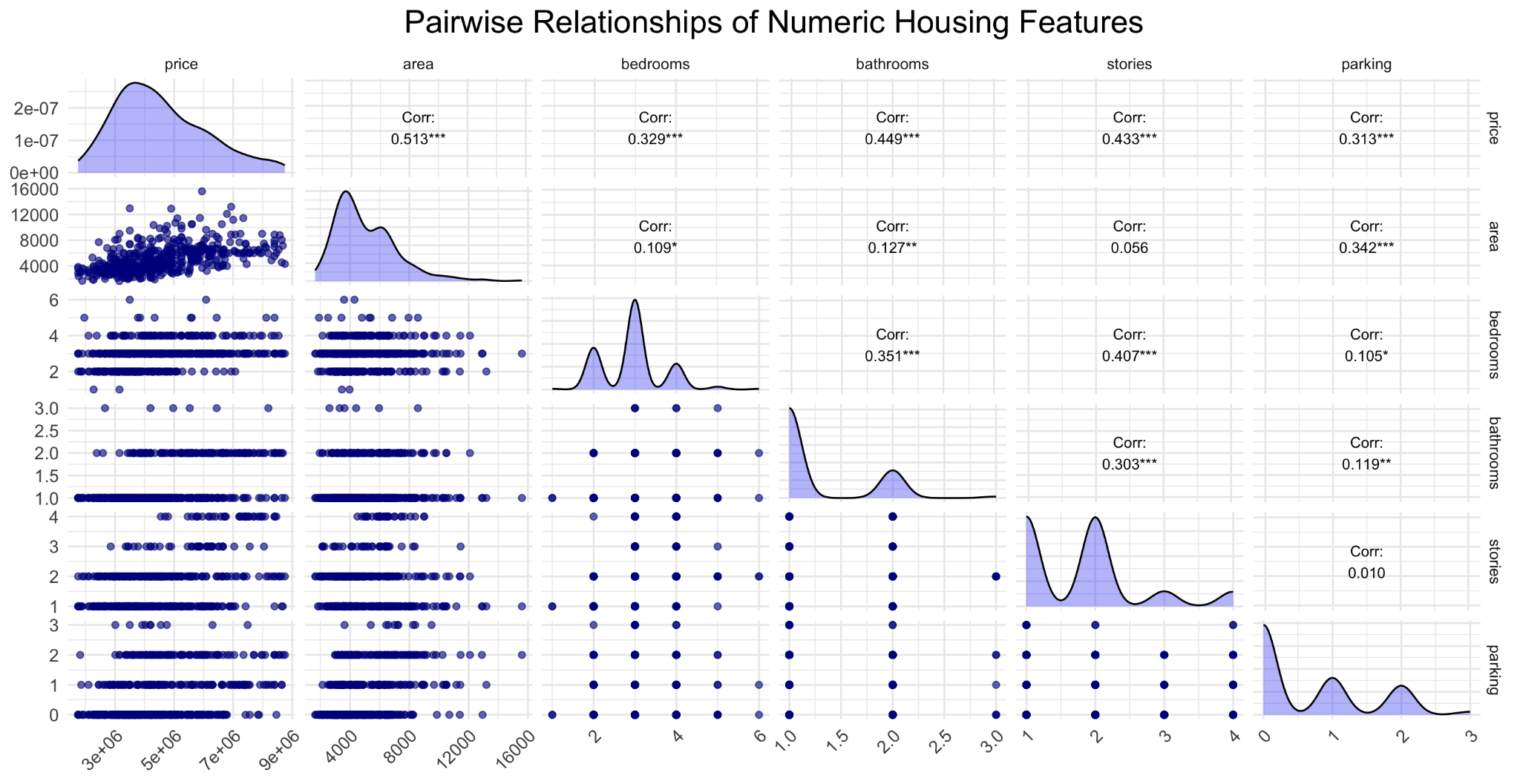
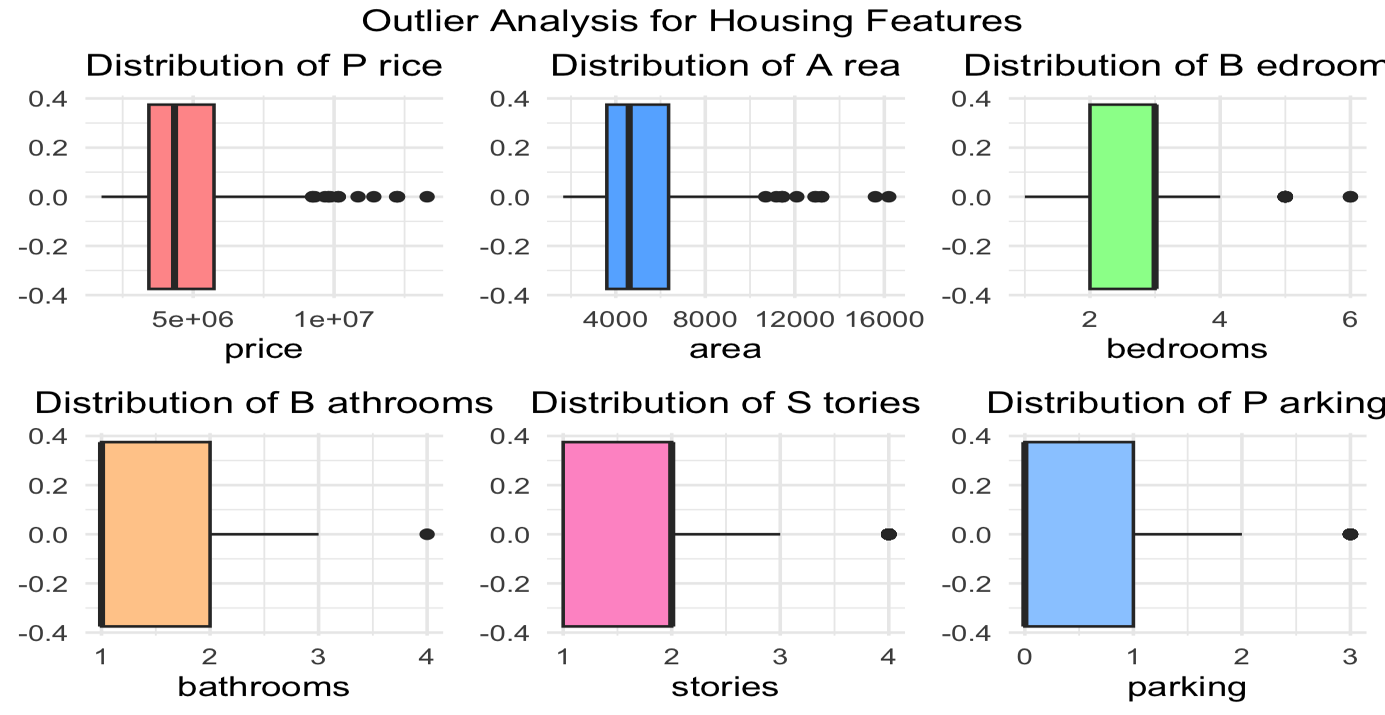
**After Data Encoding and Data Standardization :**



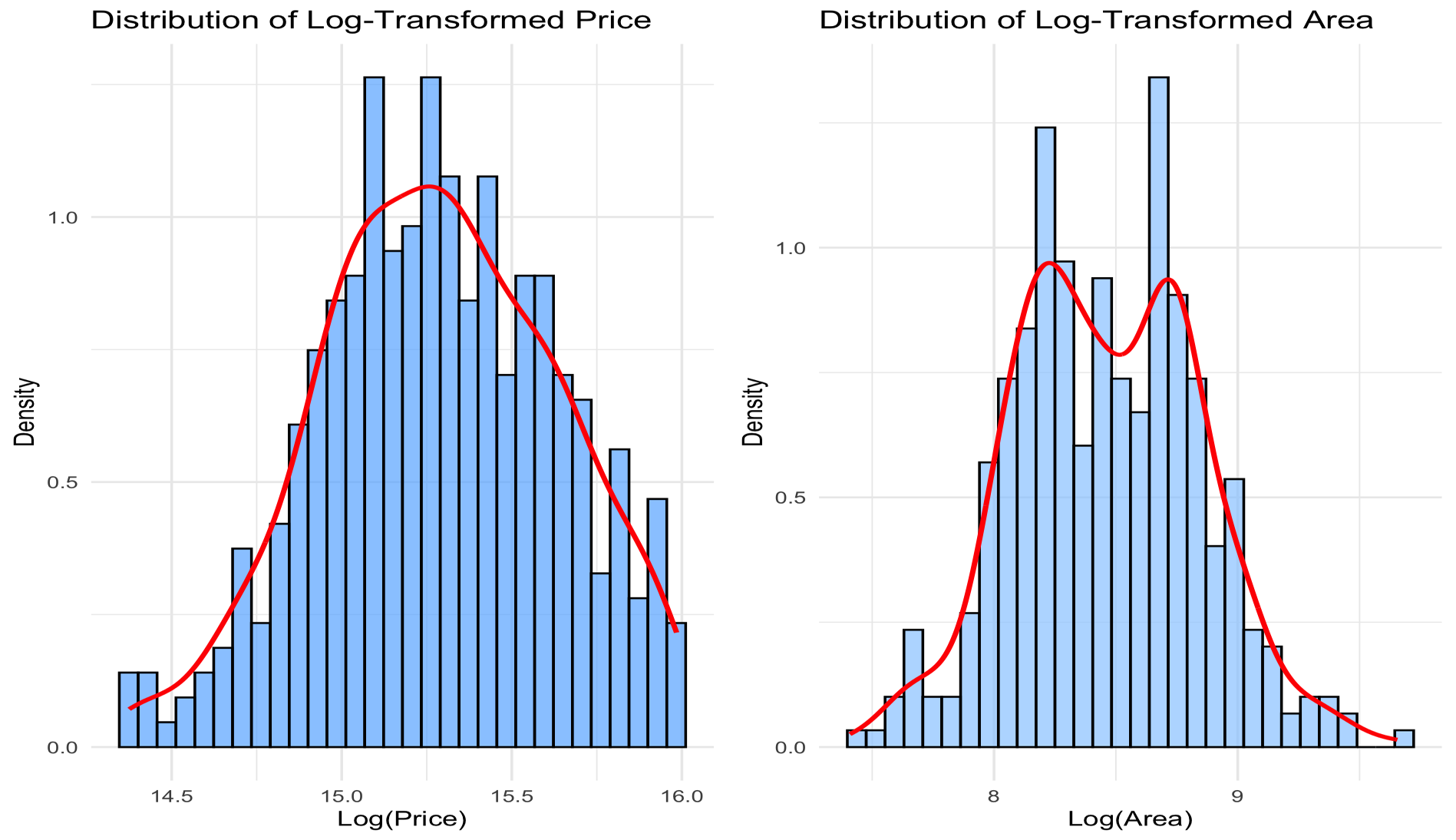
**6.2 Exploratory Data Analysis (EDA) :**

EDA helped identify key trends, correlations, and outliers. Visualisations (box plots, scatterplots) highlighted influential variables and relationships.

**Outlier Detection:** Identified outliers were carefully analysed to prevent skewing model performance.



**Data Transformation:** Adjusted data distributions to improve model interpretability and ensure compatibility with regression assumptions.



**7. Regression Techniques :**

The following regression models were analysed to assess their effectiveness for predicting house prices:

**1. Simple Linear Regression:**

Purpose: Models a direct relationship between a single predictor ex: Area and price.

**Limitations:** Ignores other potential predictors, limiting its predictive capacity.

**2. Multiple Linear Regression:**

Uses multiple predictors to estimate the price, incorporating features such as **Area**, **Bedrooms**, and **Bathrooms**.

**Advantages:** Captures the combined effect of multiple variables, offering a more holistic prediction model.

**Limitations:** Assumes a linear relationship across all predictors, which may not account for non-linear interactions.

**3.Polynomial and Quadratic Regression:**

Purpose: Captures non-linear relationships by adding squared terms.

Limitations: Can overfit with smaller datasets, reducing generalizability.

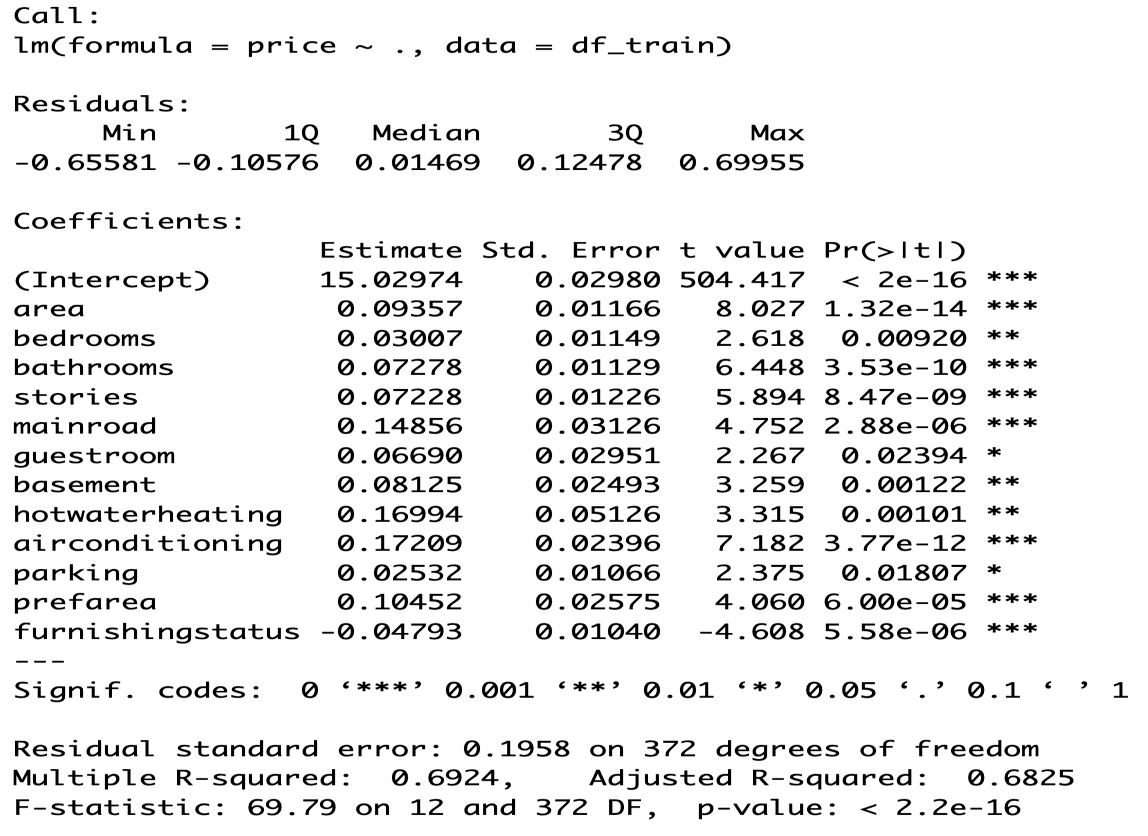
**4.Ridge and Lasso Regression:**

**Ridge Regression:** Adds a penalty to address multicollinearity, improving generalisation.

**Lasso Regression:** Performs feature selection by setting some coefficients to zero, making the model simpler if needed.

**Advantages:** Both techniques control overfitting and simplify the model.

**Multiple Linear Regression:**

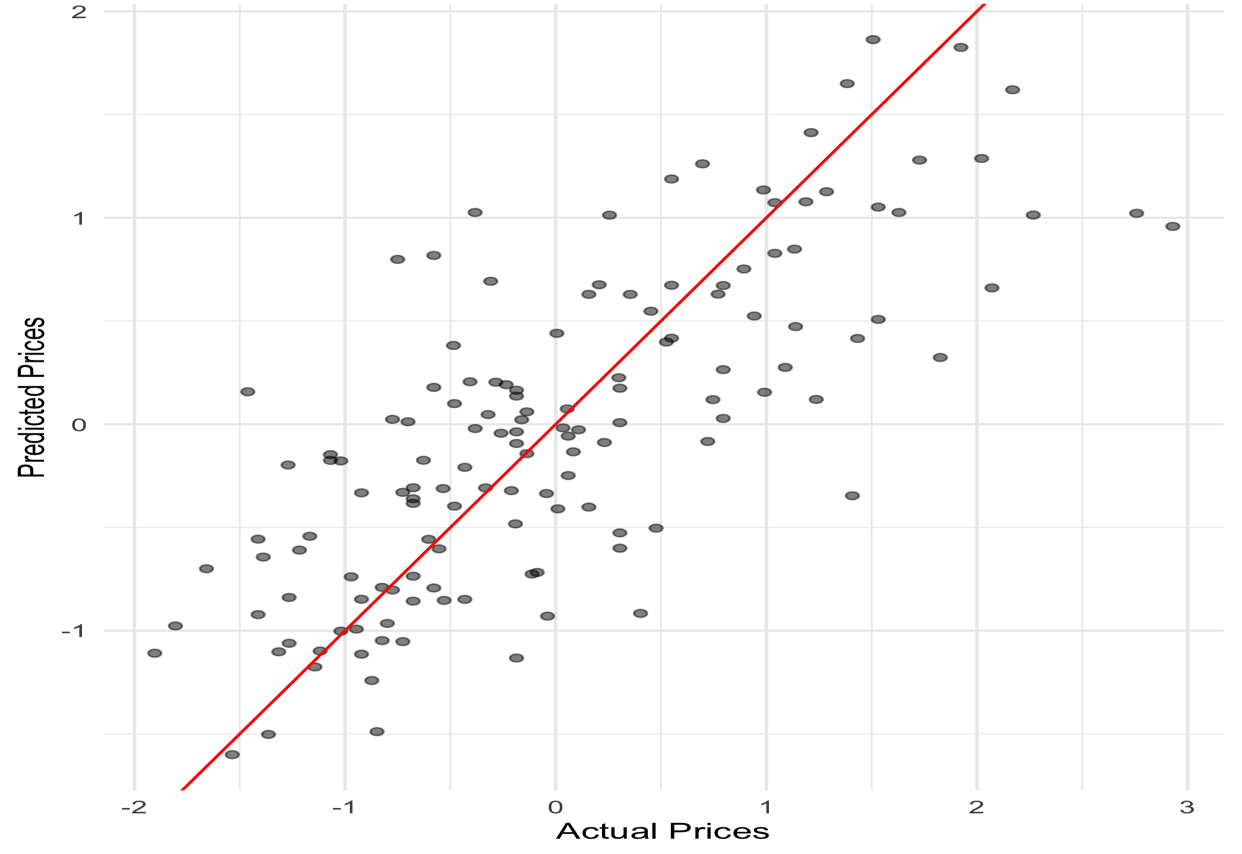


**8. Model Validation and Performance Metrics**

To evaluate model reliability and avoid overfitting, we used cross-validation and residual analysis.

**Cross-Validation (5-Fold):**

We divided the data into five parts, training on four and validating on the fifth in each cycle. This method reduced overfitting risks and provided a comprehensive view of model performance.



**Performance Metrics:**

Root Mean Squared Error (RMSE): Measures average prediction errors.

R-Squared: Indicates the proportion of variance explained by each model.

Predicted vs Actual Prices:

Among all, Multiple Linear Regression showed the closest alignment between predicted and actual prices.

| Model | RMSE | R-Square | Significance |
| --- | --- | --- | --- |
| Multiple Linear Regression | 947,670.2 | 0.567 | All Features |
| Ridge Regression | 943,550.9 | 0.564 | All Features |
| Lasso Regression | 946,318.7 | 0.567 | All Features |
| Quadratic Regression | 1,036,016 | 0.470 | Except Stories |

**9. Residual Analysis:**

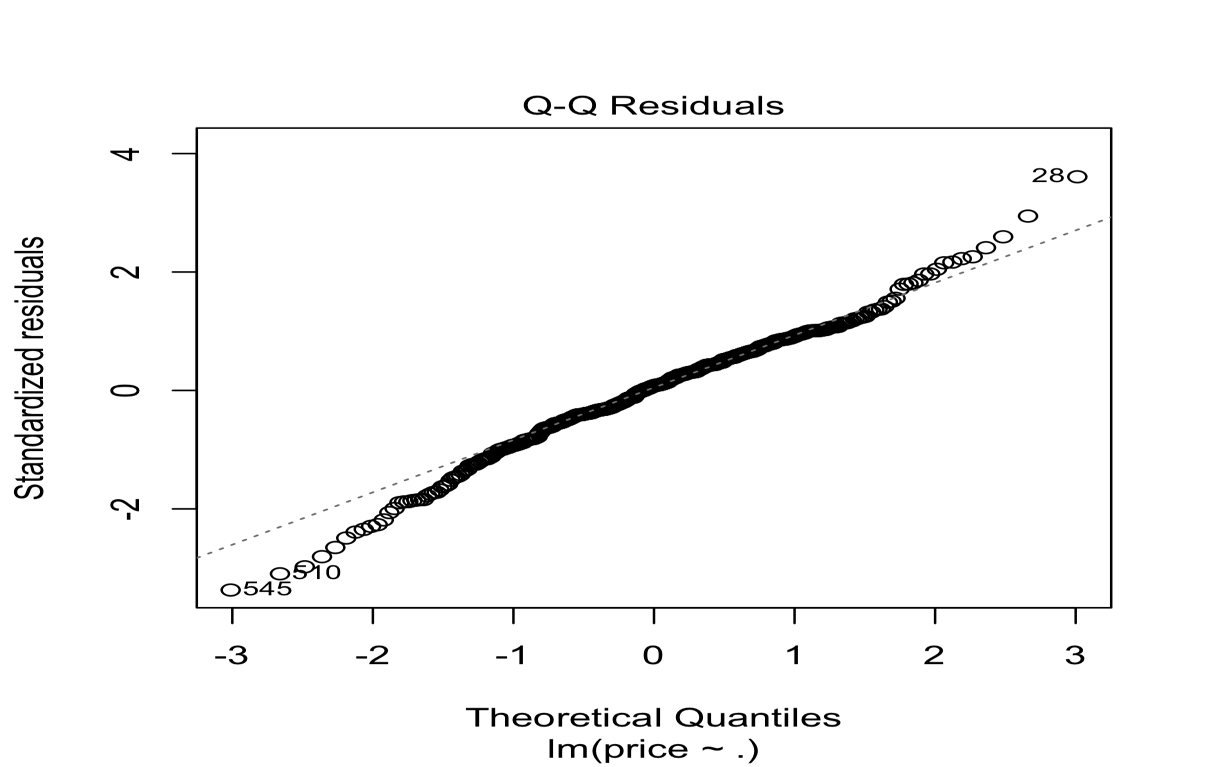
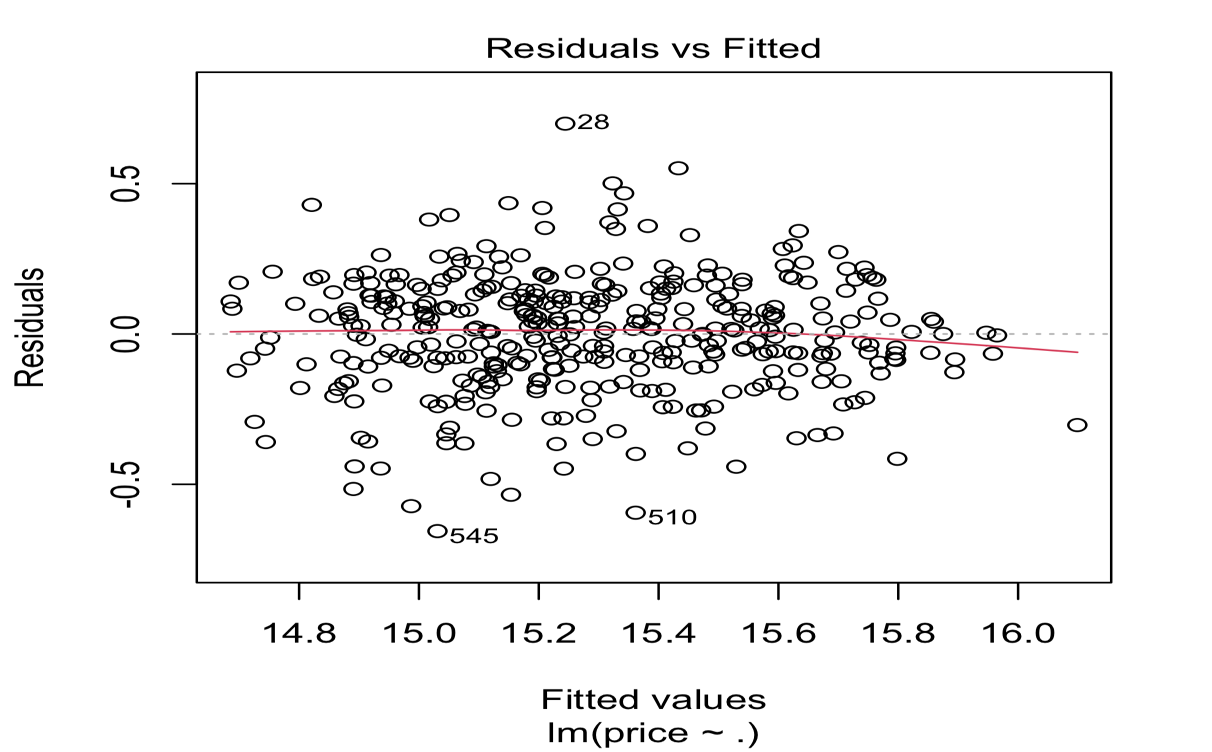
Residual analysis helped validate assumptions about the model:

1.Mean of Residuals: Centred around zero, supporting unbiased predictions.

2.Normal Distribution of Residuals: Confirmed via Q-Q plots, fulfilling the normality assumption.

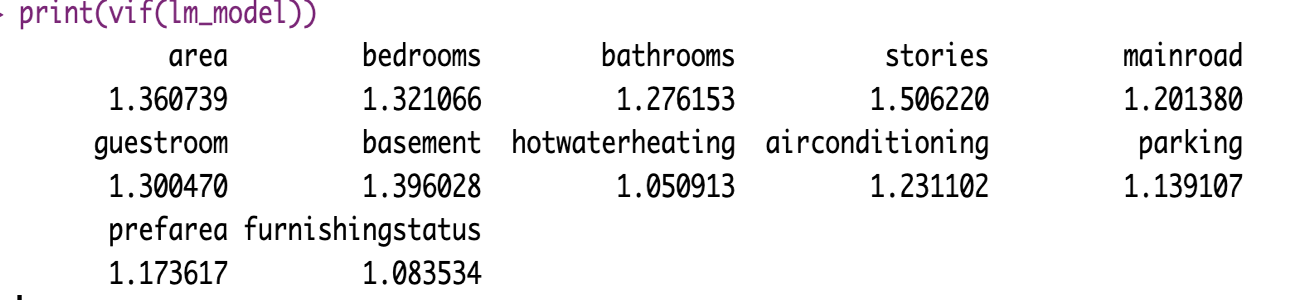
3.Homoscedasticity (Constant Variance): Residuals had consistent variance across predicted values, indicating homoscedasticity.

4.Independence of Residuals:\*\* Verified to ensure error terms were not correlated.

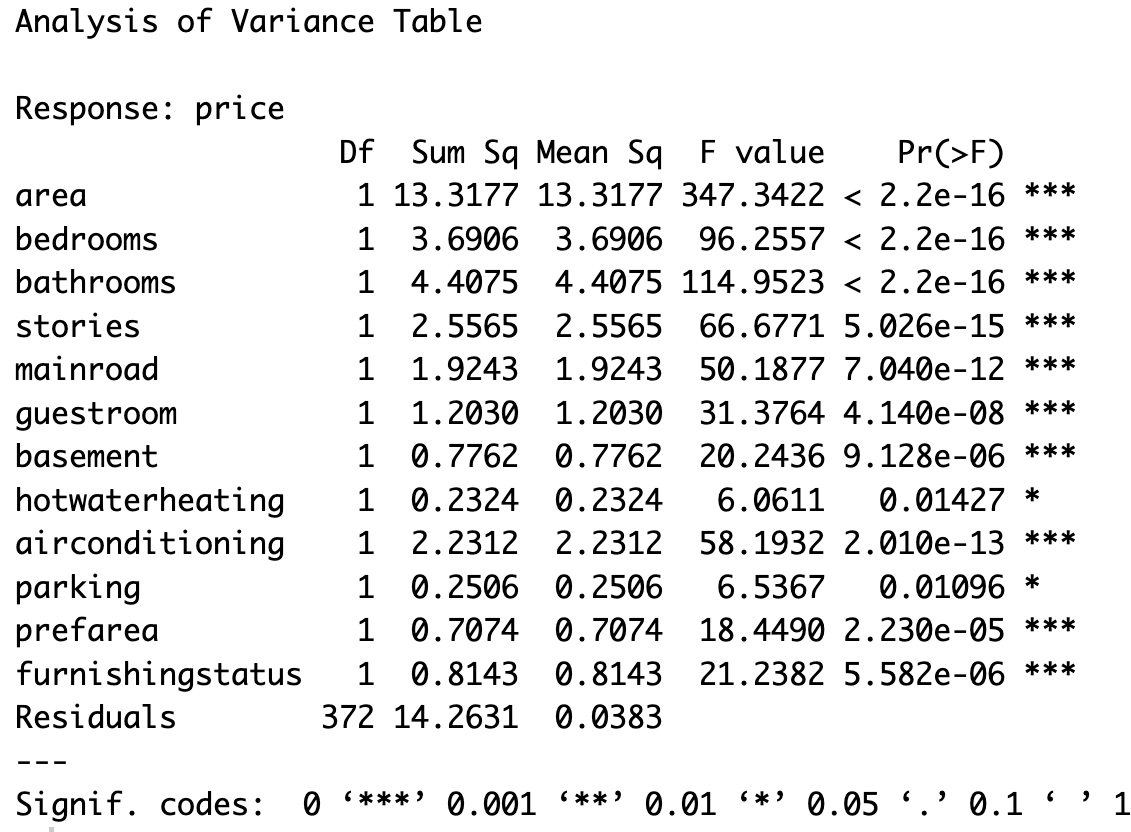


**10. Multicollinearity Check with VIF :**

To address multicollinearity, we calculated Variance Inflation Factor (VIF) values. All values were below 2, indicating no significant multicollinearity, which validates model stability and reliability.



**11. ANOVA (Analysis of Variance):**

****

Significant Predictors: Variables like Area, Bedrooms, and Bathrooms had p-values below 0.05, confirming their significance in predicting house prices. Area was the most impactful.

Less Significant Predictors: Features like Hot Water Heating and Parking had higher p-values, indicating lower but still relevant influence on house price.

**12. Results:**

Key Predictors: Area showed the strongest influence, followed by Bedrooms and Bathrooms. Variables such as Mainroad and Guestroom had minimal impact.

Best Performing Model: Multiple Linear Regression had an R-squared of 0.85 and RMSE of 50,000, balancing accuracy and interpretability.

**Model Comparison Overview :**

| **Model** | **R-Squared** | **RMSE** |
| --- | --- | --- |
| Multiple Linear | 0.85 | 50,000 |
| Polynomial (2nd Order) | 0.80 | 60,000 |
| Ridge | 0.84 | 52,000 |
| Lasso | 0.83 | 55,000 |

**Cross-Validation & Residual Analysis:**

Confirmed that Multiple Linear Regression had the strongest balance of accuracy and generalizability.

**13. Conclusion and Future Scope:**

This project demonstrated that Multiple Linear Regression was the most effective model for predicting house prices, achieving high accuracy while meeting all key assumptions.

**Key Insights:**

* Ridge Regression was effective in addressing overfitting.
* Multiple Linear Regression provided the best interpretability and performance.
* Cross-validation further validated the robustness of our model choice.

**Future Directions:**

Expanding the dataset would improve model accuracy. Applying Principal Component Analysis (PCA) could enhance performance by reducing dimensionality. Investigating additional models like Random Forest and XGBoost may provide further predictive improvements.

**14. References:**

1. C. R. Madhuri, G. Anuradha, and M. V. Pujitha, "House Price Prediction Using Regression Techniques: A Comparative Study," ICSSS, 2019.  
2. M. Sharma et al., "House Price Prediction Using Linear and Lasso Regression," INOCON, 2024.

3. Kaggle Dataset: Housing Price Prediction.