Ames Housing Price Prediction*

Analysis of Structural Housing Features and Sale Prices

Andy Jiang

June 6, 2025

This paper analyzes housing sale prices in Ames, Iowa using multiple linear regression. The dataset is filtered to focus on homes with 2–4 bedrooms, built after 1940, and rated average to excellent in quality. Predictors include log-transformed ground living area, garage area, year built, number of bedrooms, and overall quality. Log transformations address skewness and heteroscedasticity. The final model achieves an R^2 of 0.819, with ground living area and overall quality identified as the most influential predictors. The results provide interpretable insights for investors, buyers, and policymakers.

Table of contents

*Code and data are available at: https://github.com/AndyYanxunJiang/ames-housing-price-prediction.

1 Introduction

The housing market significantly impacts financial stability and investment. The challenge many people face in finding affordable housing highlights the importance of understanding the factors that influence house prices. Given rising real estate costs, identifying factors influencing home values has become both timely and relevant. This study aims to analyze the key factors affecting housing prices in Ames, Iowa, focusing on attributes such as the number of bedrooms (two to four), garage area, ground living area, overall quality (average to excellent), and year built. The research question seeks to determine which housing attributes most significantly influence the sales price of homes in average to excellent quality grade with two to four bedrooms.

There are several peer-reviewed papers that have studied this topic. (Shukla 2024) demonstrated the effectiveness of multiple linear regression in predicting housing prices, listing property size and number of bedrooms as significant predictors, reflecting their intrinsic value and desirability. (Ye 2024) compared different predictive models specifically for Ames housing prices and identified overall quality of the house and living area as the most influential factors. (Han 2023) explored predicting house prices in Ames, Iowa with the same dataset using various advanced regression models and identified key features influencing house prices. They found overall quality of material and finish and year built to be the most significant variables, as well as garage area and ground living area to be significant variables. These studies support the selection of predictor variables and the use of linear regression for this analysis.

Previous research suggests that linear regression is an effective tool for examining the direct relationships between multiple predictors (such as housing attributes) and the response variable (sales price) (GeeksforGeeks 2023). By modeling how factors like size, age, quality, and market trends affect house prices, linear regression quantifies these relationships and provides clear, interpretable coefficients. This makes it a widely used and suitable method for this study, aiming to offer insights into the factors that influence home prices in Ames, Iowa, with a focus on houses in average or above condition with at least two bedrooms.

2 Data Description

The dataset used in this analysis is from the Ames Housing data available through the AmesHousing R package (Kuhn and Johnson 2024). Originally curated by Dean De Cock for use in data science education. The original dataset describes the sale of individual residential properties in Ames, Iowa, from 2006 to 2010 and includes 82 variables related to various attributes of each property.

The response variable for this study is the final sale price of each home. Since sales price is highly right-skewed, we apply log transformation to stabilize the variance and make the

distribution suitable for regression analysis. This transformation is commonly used in price modeling to address issues related to non-normality and heteroscedasticity (Gupta 2024).

Figure 1: Distribution of House Sale Prices.

Prior to log transformation in Figure 1, sale price is heavily right skewed.

After the log transformation in Figure 2, sale price is more normally distributed.

Table 1: Predictors, Measurement Units, Restrictions, and Literature Mentions.

Predictor	Measurement	Restriction	Mentioned
Ground Living Area	Square feet	N/A	Paper 2 Paper 3
Garage Area	Square feet	Garage_Area > 0	Paper 2 Paper 3
Year House Built	Year	Year_Built >= 1940	Paper 3
Bedrooms Above Ground	Count	Two to four bedrooms c(2, 3, 4)	Paper 1

Predictor	Measurement	Restriction	Mentioned
Overall Quality of Material and Finish	Quality Grade		
(10 Levels: Very_Poor to Very_Excellent)	Average to Excellent grade		
c("Average", "Above_Average", "Good", "Very_Good", "Excellent")	Paper 2		
Paper 3			

According to (Shukla 2024), the predictor variables reflect the intrinsic value and desirability of the house, directly tied to the house's sale price. These variables capture essential aspects of a property's value and appeal, which are key determinants of its market price.

Boxplot for ground living area in Figure 3 has a longer upper whisker and outliers extending further from the median, indicating right skewness. We applied log transformation in Figure 4 to deal with the skewness, now the distribution looks more normal.

Garage area in Figure 5 has a similar distribution as ground living area, showing a right skew. We again apply log transformation in Figure 6 to address the same skewness issue. The normality of distribution is drastically improved.

For the year built distribution in Figure 7, due to the discrete and linear nature of time in, there seem to have high and low extremes which can possibly be explained by economic factors. However, the count differences do not necessarily heavily impact residual variances.

Houses generally have 2-4 bedrooms above ground, and as we can see in Figure 8, houses with 3 bedrooms above ground dominate.

From Figure 9, most houses have average to good quality grades, with a rare number of excellent grades.

3 Preliminary Model Results

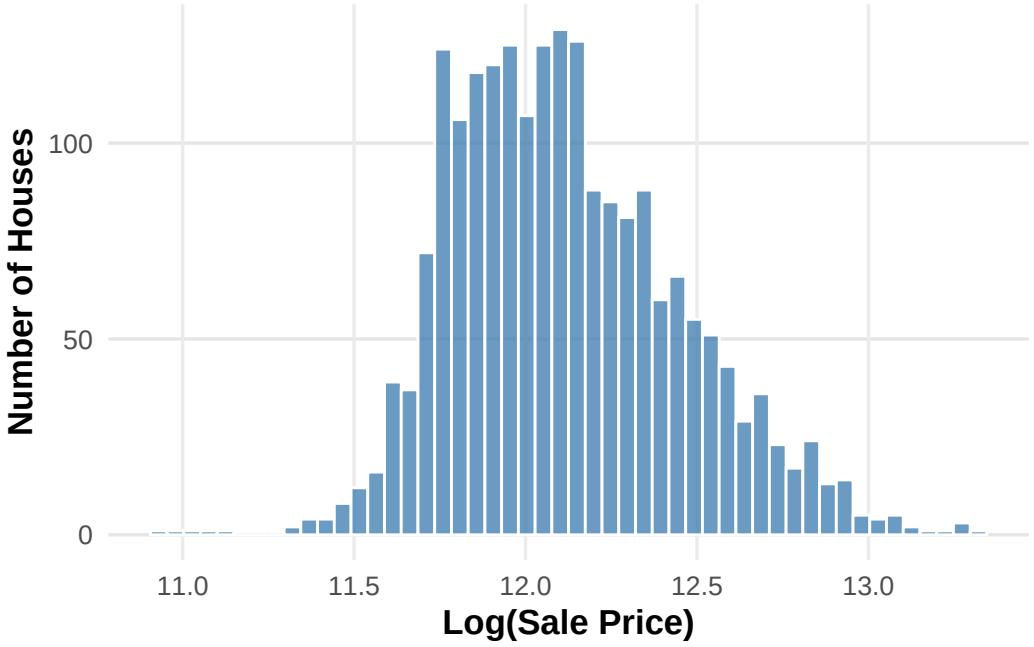

Figure 2: Distribution of Log-Transformed Sale Prices.

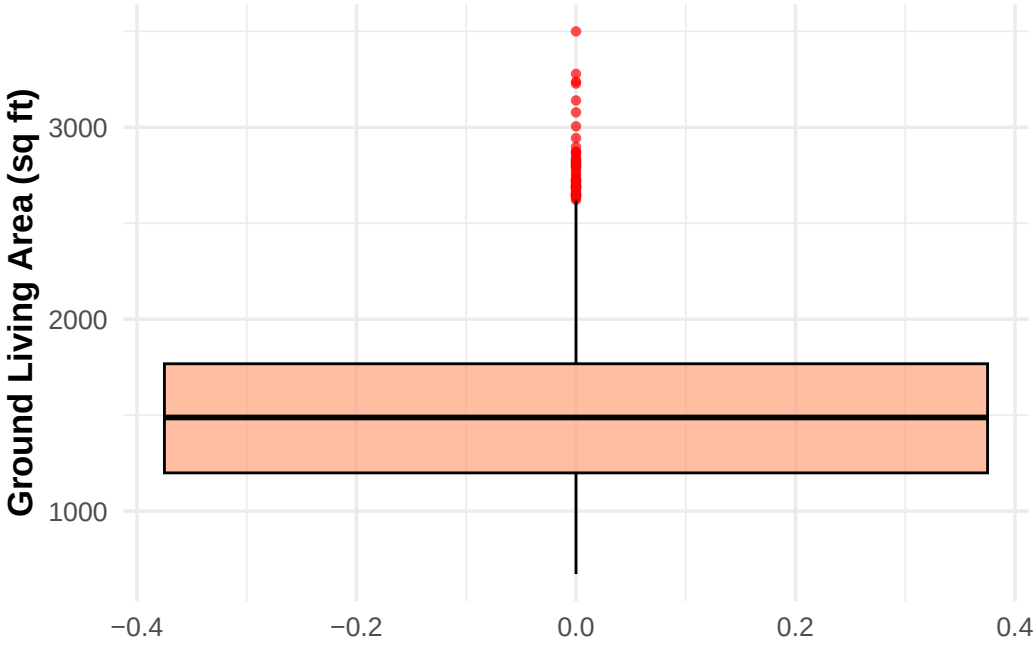

Figure 3: Boxplot of Ground Living Area.

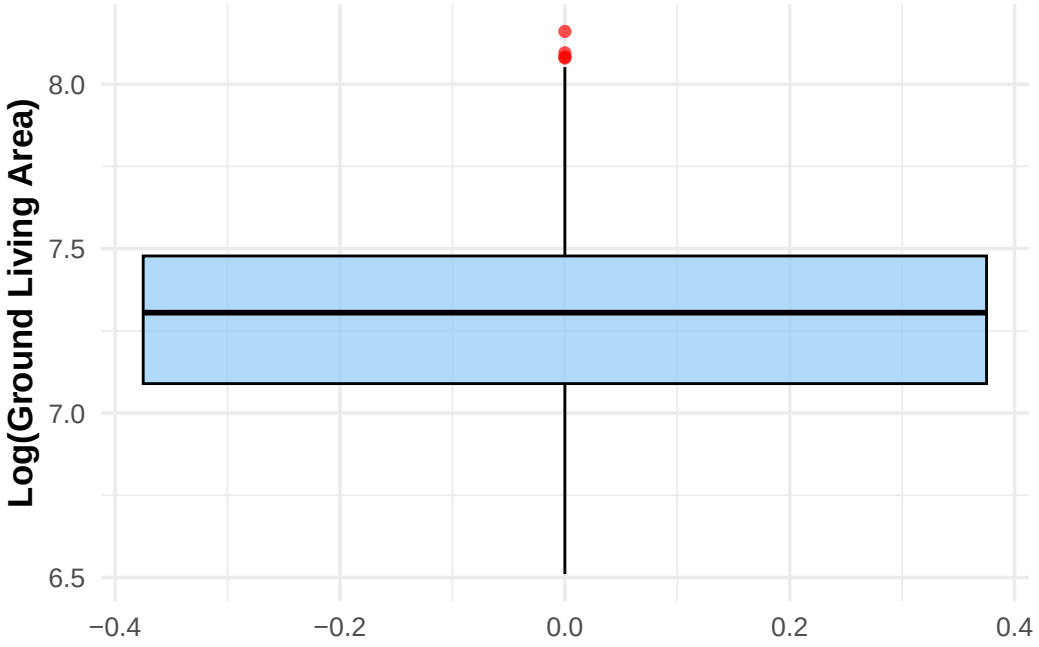

Figure 4: Boxplot of Log-Transformed Ground Living Area.

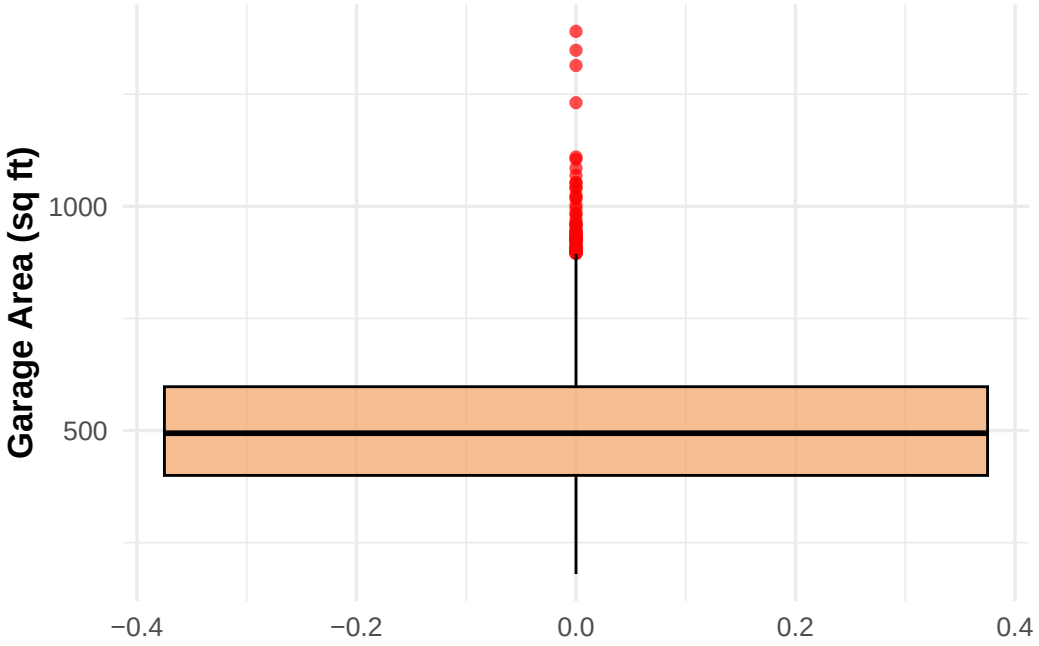

Figure 5: Boxplot of Garage Area.

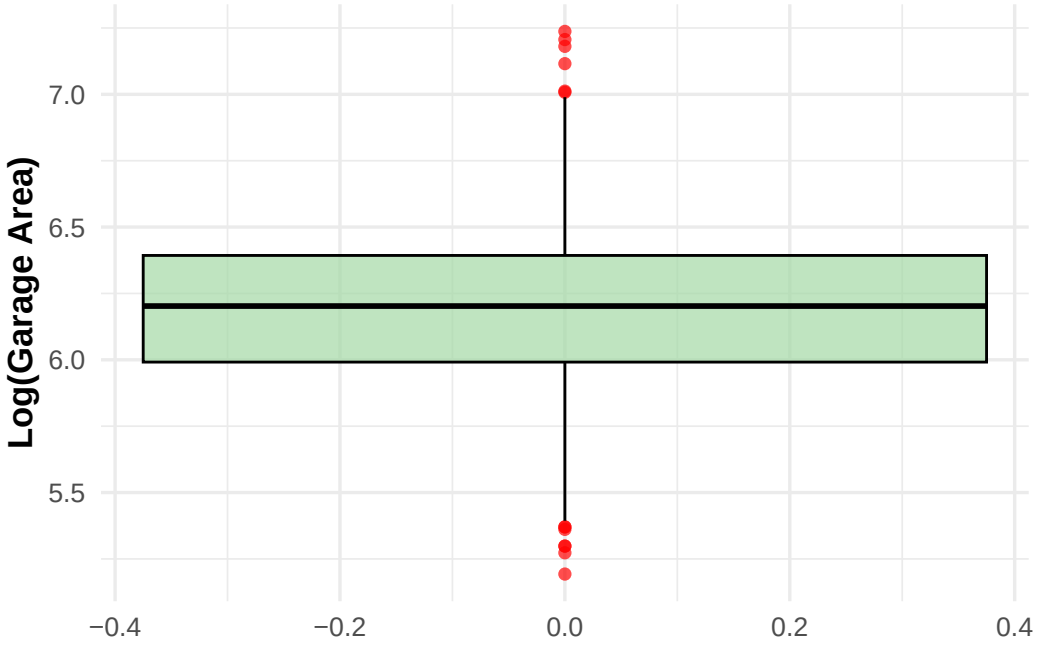

Figure 6: Boxplot of Log-Transformed Garage Area.

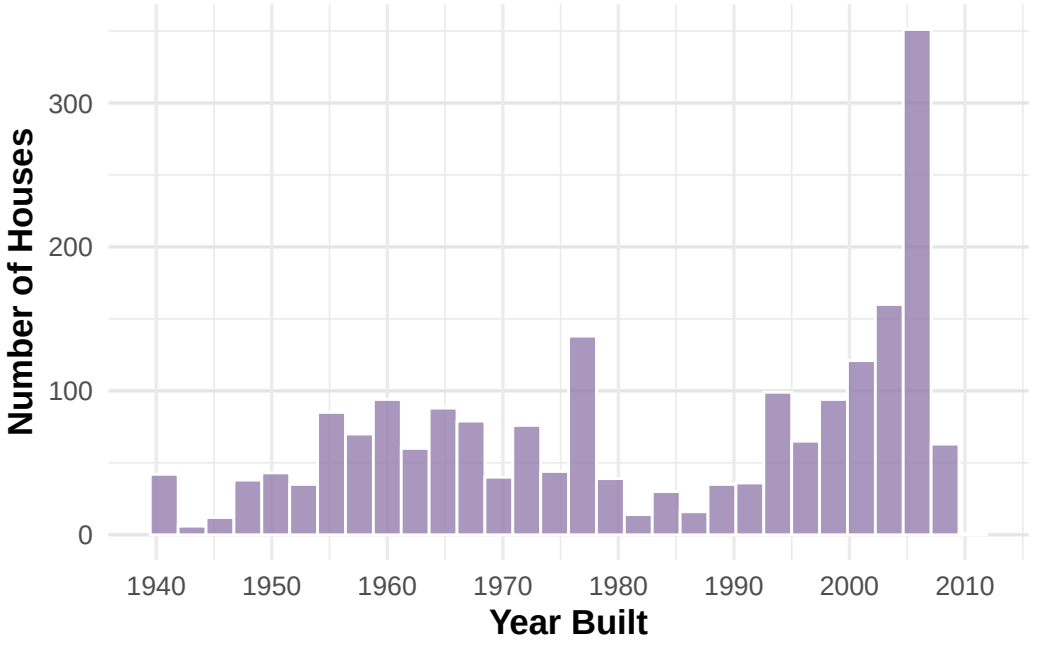

Figure 7: Distribution of House Construction Years.

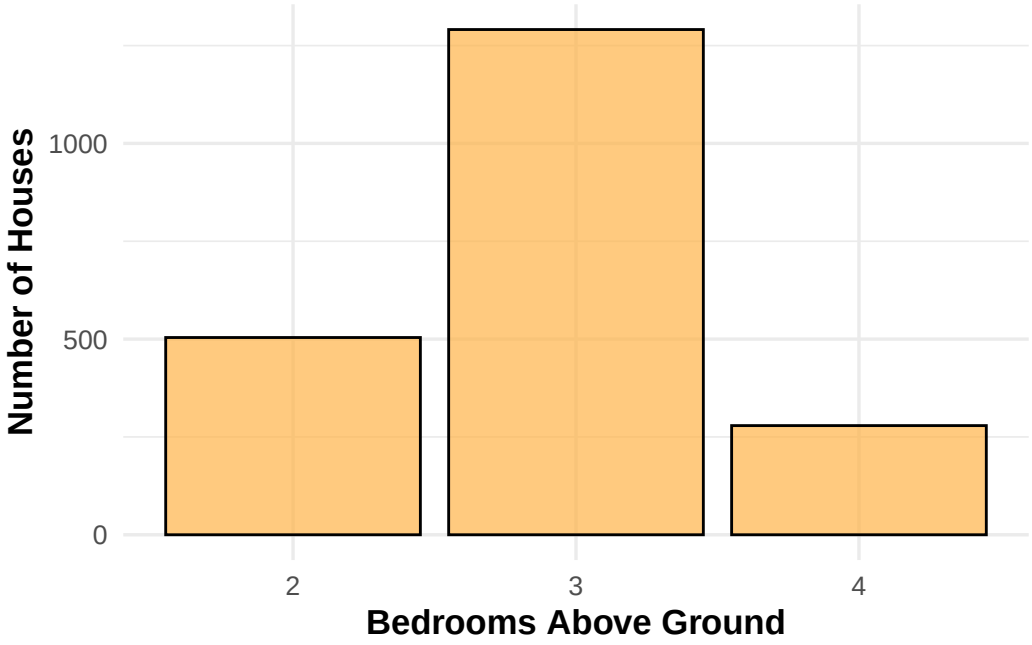

Figure 8: Distribution of Bedrooms Above Ground.

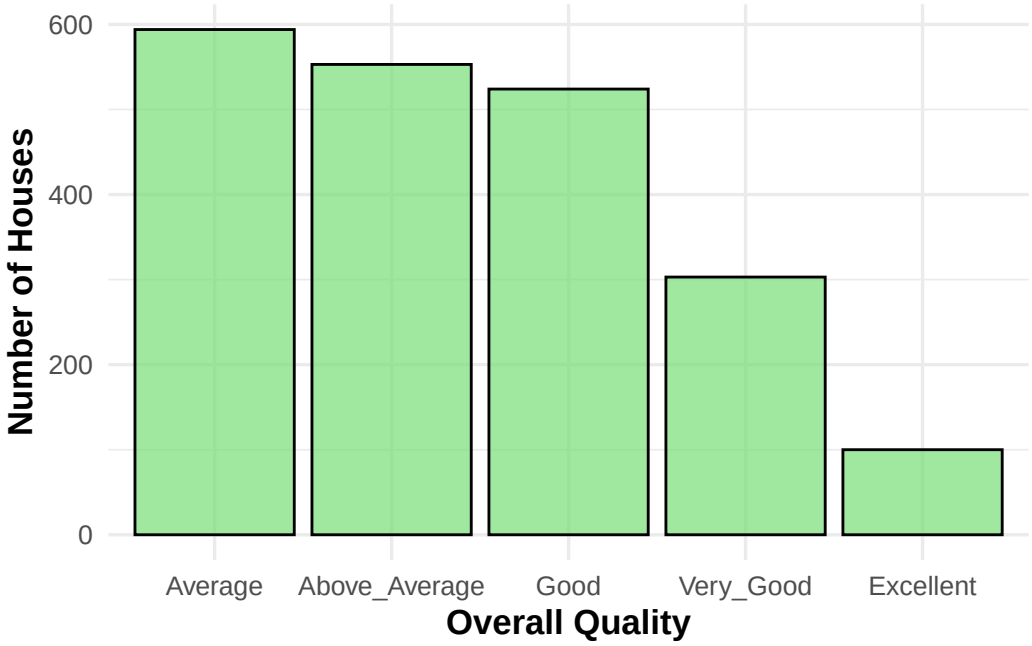

Figure 9: Distribution of Overall Quality Ratings.

Table 2: P-values of Predictors in Final Model

Predictor	p.value
3 bedrooms above ground	0.0984
all other predictors	<0.05

Our preliminary model has a log transformed response and the original predictors without transformation. Most predictors are statistically significant, except houses with 3 bedrooms above ground Table 2.

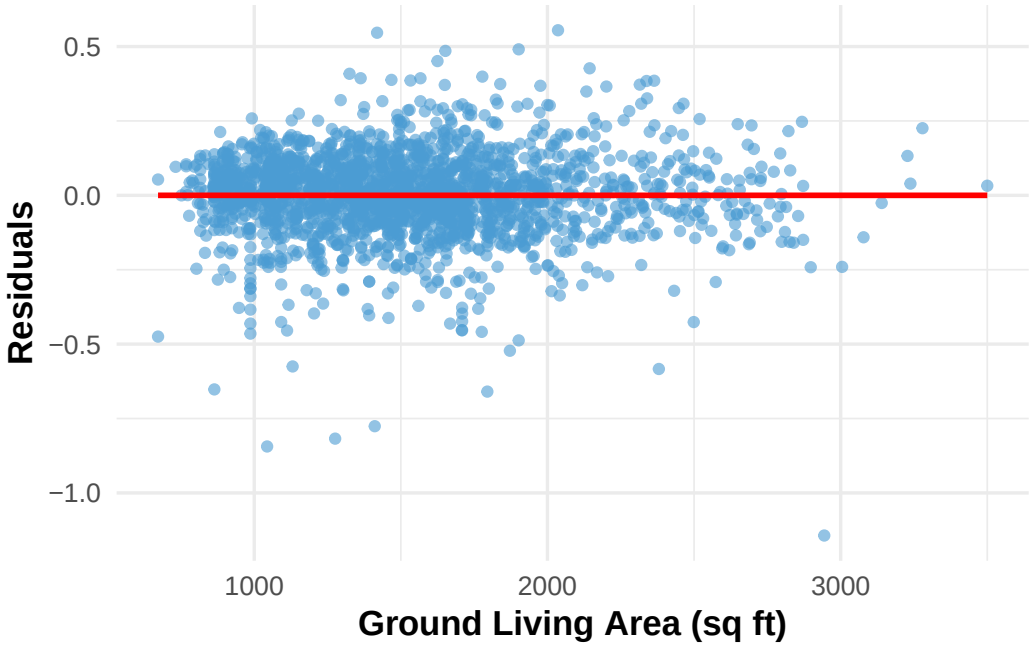

Figure 10: Residuals versus Ground Living Area showing mild left-side clustering.

In the ground living area residual plot Figure 10, residuals appear mostly random with slight clustering toward the left, indicating potential heteroscedasticity. For the garage area residual plot Figure 11, the residuals also cluster more on the left, also suggesting potential correlated errors or heteroscedasticity. These patterns suggested that homoscedasticity assumptions were not fully satisfied.

For the residual plots for both the bedrooms above ground Figure 12 and overall quality Figure 13. By restricting the number of bedrooms to between two and four, and limiting quality grades to average and above. We achieve a more consistent interquartile range across predictor levels. While this does not ensure elimination of potential violations in all relationships, it reduces the extreme deviation visibly, improving stability.

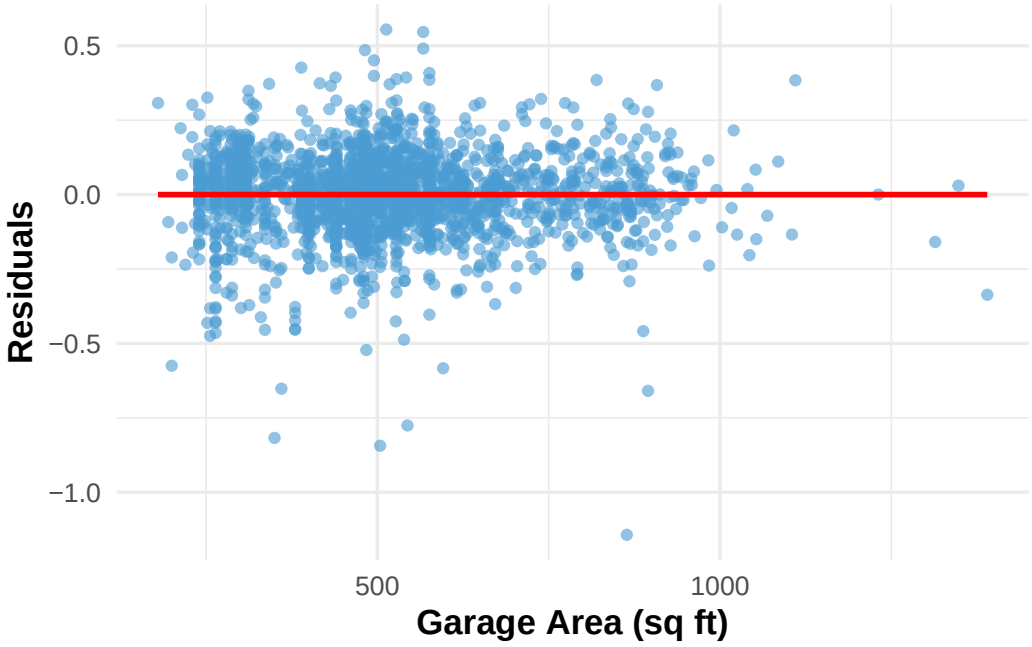

Figure 11: Residuals versus Garage Area showing left-side clustering and heteroscedasticity.

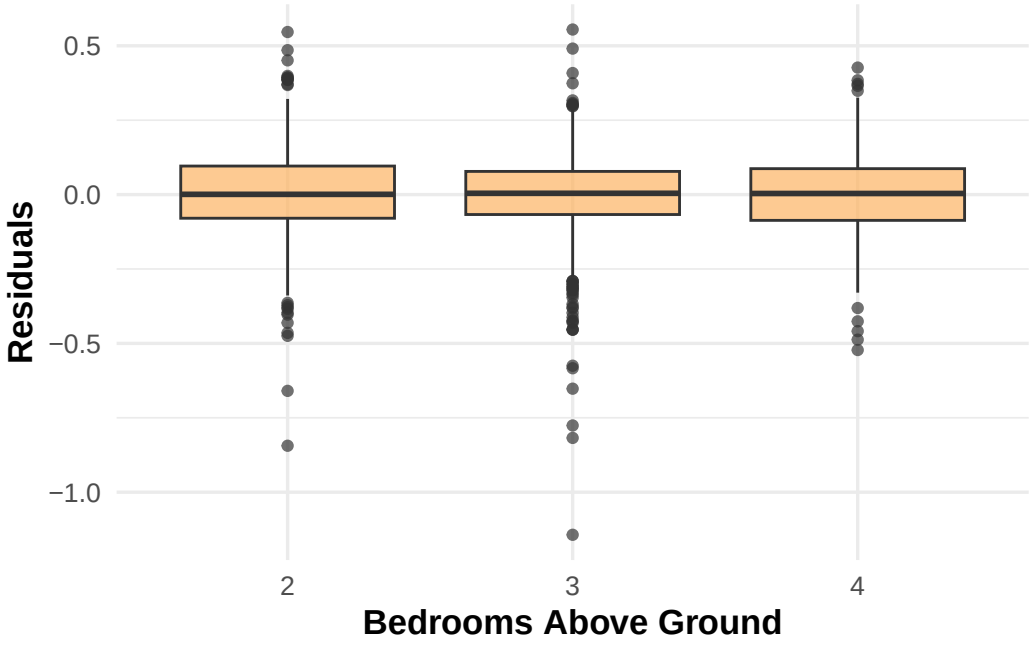

Figure 12: Boxplot of residuals by number of bedrooms above ground.

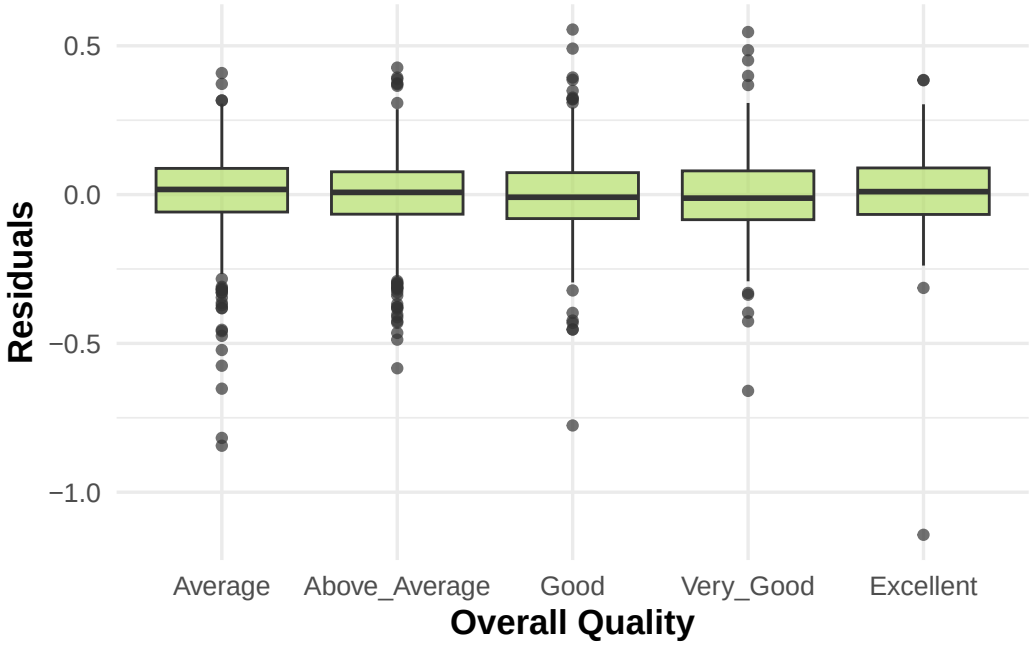

Figure 13: Boxplot of residuals by overall quality rating.

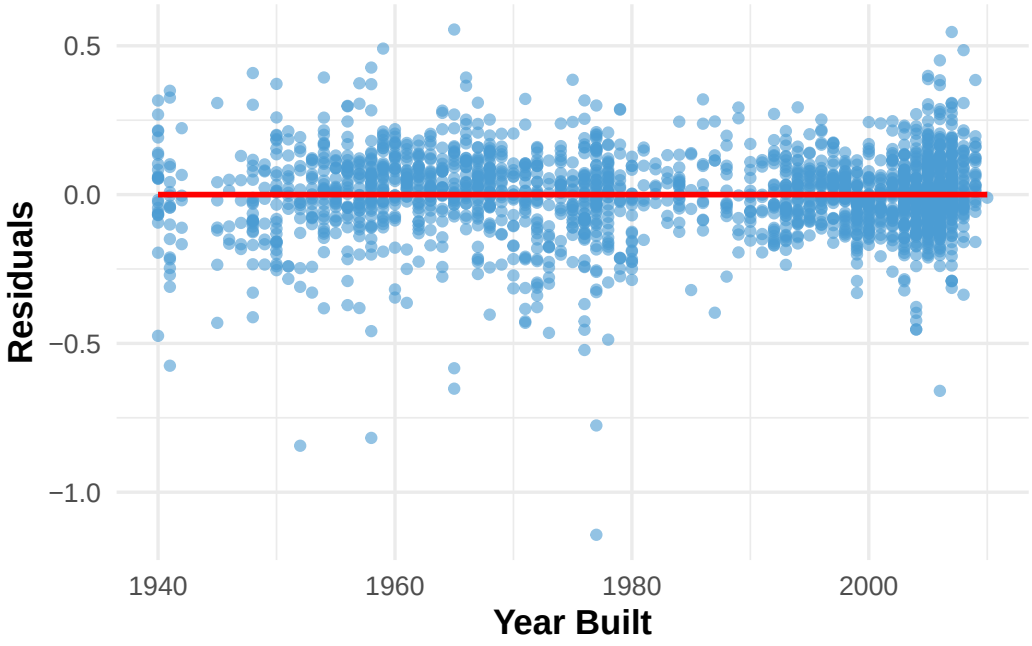

Figure 14: Residuals versus Year Built showing mild structure across time.

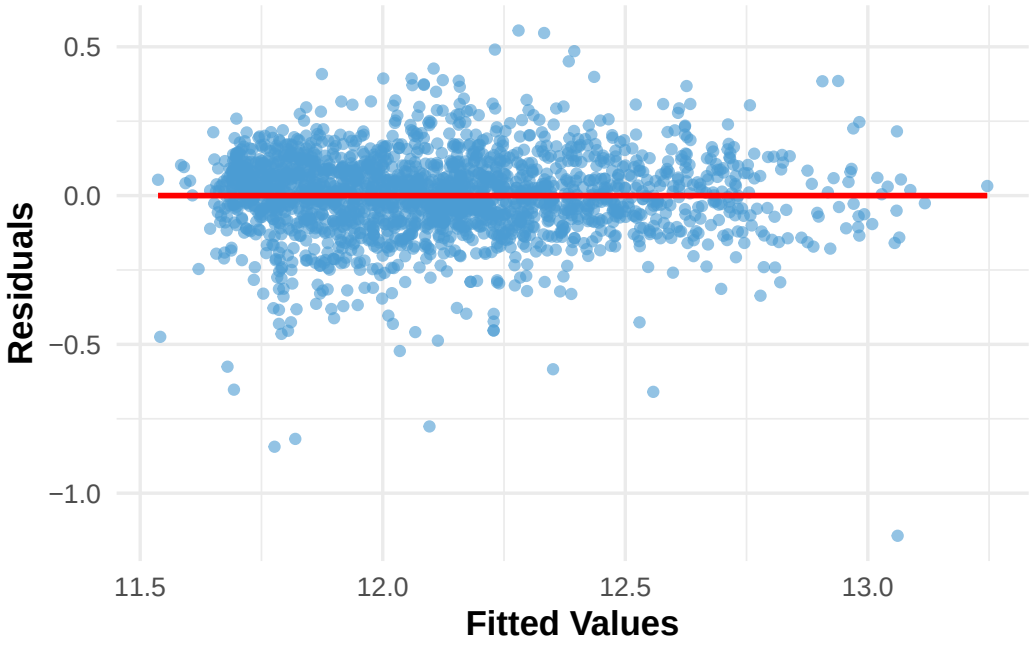

Figure 15: Residuals versus fitted values, assessing model fit and heteroscedasticity.

For the year built residual plot Figure 14, although discrete, it is best treated as a continuous variable since time is an increasing measure. We restricted the data to post-1940 to exclude outliers. The residual spread appears random year by year, showing no visible pattern. For the fitted value residual plot Figure 15, similar to the ground living area plot, slight clustering on the left which also suggests potential heteroscedasticity.

For the response versus fitted plot Figure 16, the points scatter mostly random along the diagonal, though the left side slightly heavier, consistent with previous residual plot observations, indicating some deviations from randomness.

After log-transforming the response Figure 17, the right-skewed tail is significantly reduced but still shows some deviations, indicating skewness still persists. A strong left skew suggests the residuals deviates from normal, potentially violating normality assumption. In the pairwise plot Figure 18, the relationship between ground living area and garage area shows both a linear trend and scattered cluster, suggesting multiple underlying relationships violating linearity. Other variable relationships appear either random or linear, largely aligning with the assumption.

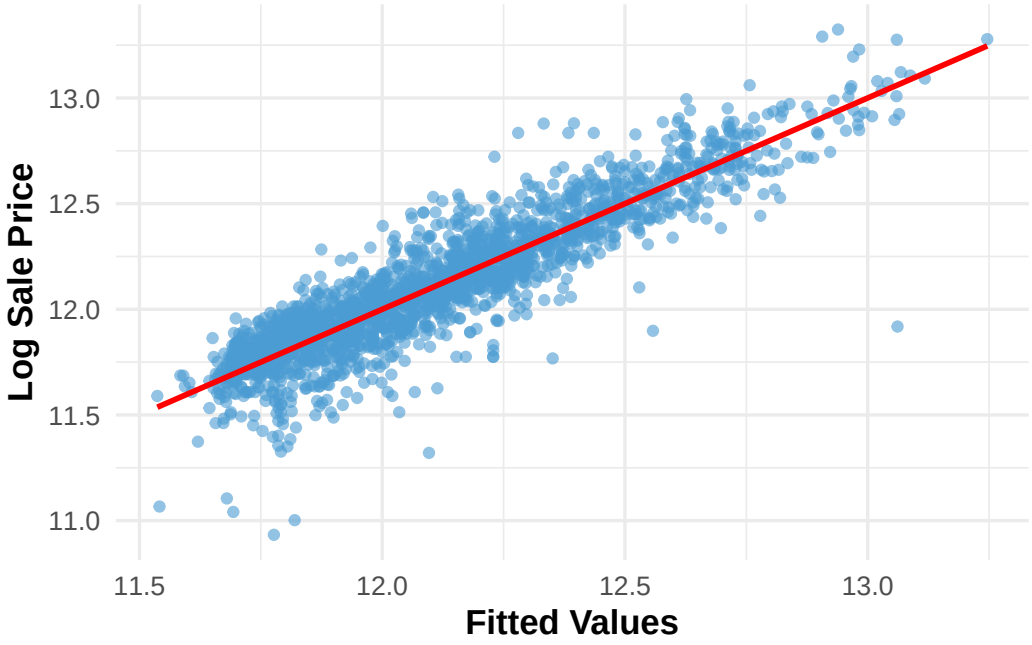

Figure 16: Fitted values versus log sale price, evaluating model prediction alignment.

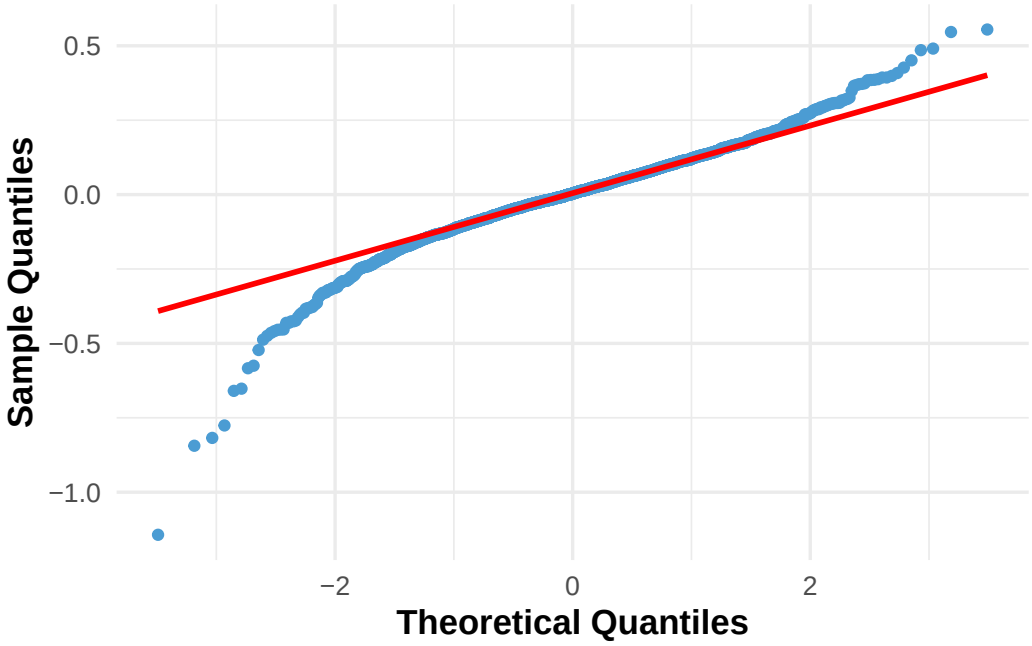

Figure 17: Q-Q plot of residuals to assess normality.

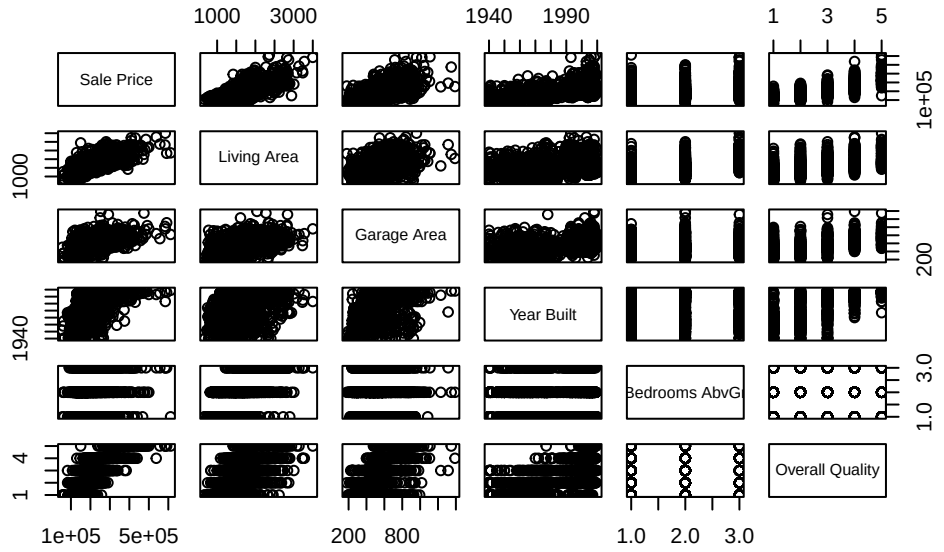

Figure 18: Scatterplot matrix of key predictors and response variable.

4 Model Selection

Based on the diagnostic issues from Section 3, several improvements were made to build a more solid model for predicting house prices, including transformation, data restrictions, and data interpretation.

4.1 Transformations on Predictor Variables

Both ground living area and garage area are log transformed to address right skewed distributions and reduce heteroscedasticity observed from preliminary analysis. The transformations showed that the residuals are more evenly scattered, drastically eliminating previous violations. With the garage area filtered to greater than zero, we avoid undefined values from log transformation.

Table 3: Correlation Matrix of Numerical Predictors

	Log.Ground.Living.Area	Log.Garage.Area	Year.Built
Log Ground Living Area	1.000	0.518	0.421
Log Garage Area	0.518	1.000	0.545

	Log.Ground.Living.Area	Log.Garage.Area	Year.Built
Year Built	0.421	0.545	1.000

5 Final Model Inference and Results {#sec-final}z

Table 4: Final Model Coefficients with 95% CI and p-values

Predictor	Estimate	Std..Error	X95..CI	p.value
(Intercept)	2.6813	0.4588	[1.7816, 3.5810]	<0.0001
Log Ground Living	0.5325	0.0178	[0.4976, 0.5674]	<0.0001
Log Garage Area	0.1480	0.0122	[0.1240, 0.1719]	<0.0001
Year Built	0.0023	0.0002	[0.0018, 0.0028]	<0.0001
Bedrooms Above Ground: 3	-0.0156	0.0080	[-0.0313, 0.0001]	0.0515
Bedrooms Above Ground: 4	-0.0721	0.0134	[-0.0984, -0.0459]	<0.0001
Overall Quality: Above Average	0.0322	0.0094	[0.0138, 0.0507]	0.0006
Overall Quality: Good	0.1131	0.0122	[0.0892, 0.1370]	<0.0001
Overall Quality: Very Good	0.2689	0.0151	[0.2392, 0.2986]	<0.0001
Overall Quality: Excellent	0.4833	0.0204	[0.4432, 0.5233]	<0.0001

6 Discussion and Conclusion

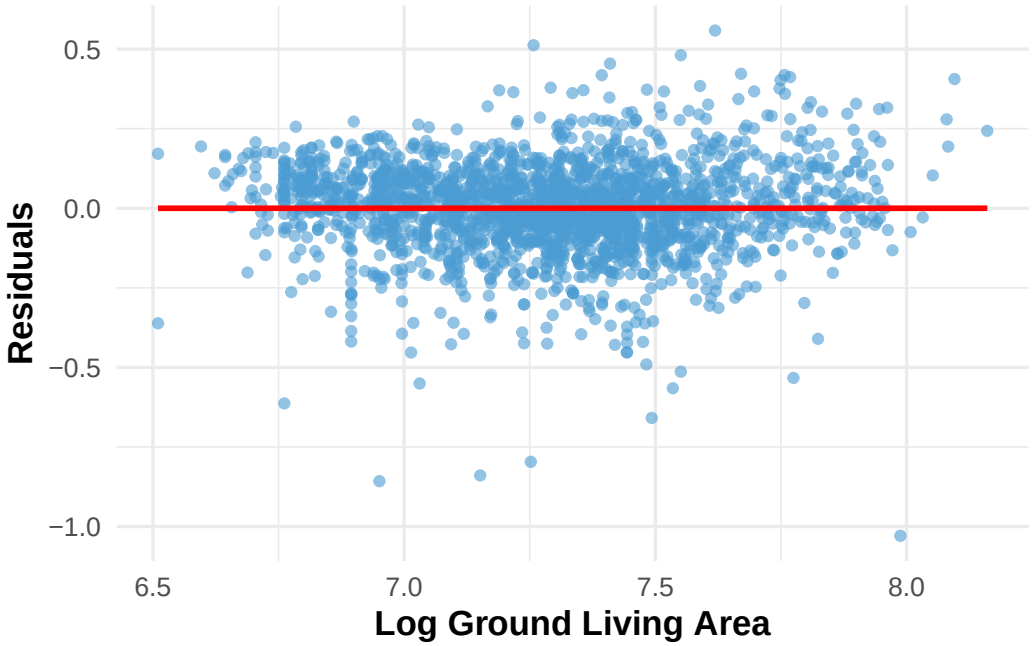

Figure 19: Residuals plotted against log-transformed ground living area to assess linearity and homoscedasticity.

References

GeeksforGeeks. 2023. "Multiple Linear Regression Using r to Predict Housing Prices." https://www.geeksforgeeks.org/multiple-linear-regression-using-r-to-predict-housing-prices/.

Gupta, Shubham. 2024. "Building a California Housing Price Prediction Model Using Gradient Boosting and Feature Selection: A Comprehensive Guide." NGAIF. https://www.ngaif.org/2024/04/building-a-california-housing-price-prediction-model-using-gradient-boosting-and-feature-selection.html.

Han, Yueting. 2023. "Price Prediction of Ames Housing Through Advanced Regression Techniques." *BCP Business & Management* 38. https://www.researchgate.net/publication/369437029_Price_Prediction_of_Ames_Housing_Through_Advanced_Regression_Techniques/fulltext/641b583b66f8522c38c770c2/Price-Prediction-of-Ames-Housing-Through-Advanced-Regression-Techniques.pdf.

Kuhn, Max, and Kjell Johnson. 2024. "AmesHousing: The Ames Iowa Housing Data." https://cran.r-project.org/web/packages/AmesHousing/AmesHousing.pdf.

Shukla, Shivani. 2024. "Predicting Housing Prices Using Multiple Linear Regression: A Comprehensive Analysis." *JETIR* 11 (3). https://www.jetir.org/papers/JETIR2403593.pdf.

Ye, Qiongwei. 2024. "House Price Prediction Using Machine Learning for Ames, Iowa." In *Proceedings of the 4th International Conference on Signal Processing and Machine Learning*.

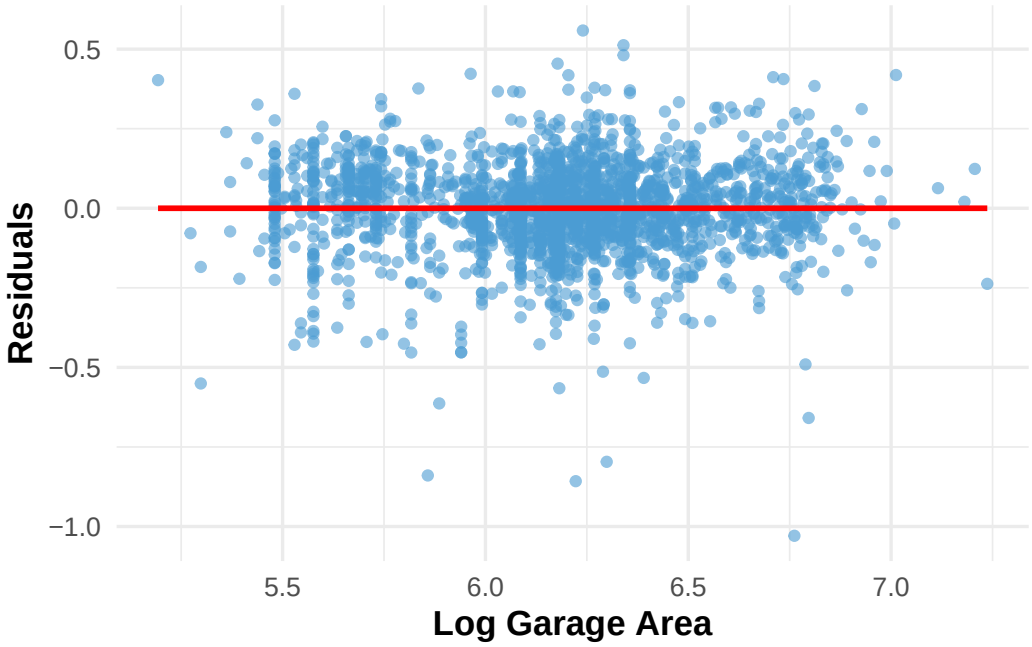

Figure 20: Residuals plotted against log-transformed garage area, examining model fit and variance consistency.

https://www.ewadirect.com/proceedings/ace/article/view/11040/pdf.