Ames Housing Price Prediction*

Analysis of Structural Housing Features and Sale Prices

Andy Jiang

June 11, 2025

This paper investigates the determinants of housing prices in Ames, Iowa using multiple linear regression. Focusing on homes with 2–4 bedrooms, built after 1940, and rated average to excellent in quality, we examine the impact of structural features such as ground living area, garage area, year built, number of bedrooms, and overall quality. Log transformations are applied to address skewness and heteroscedasticity in key variables. The final model explains 81.9% of the variation in sale price, with ground living area and overall quality emerging as the most influential predictors. These findings offer practical insights for homebuyers, investors, 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.

This paper is organized as follows. In Section 2, we describe the dataset, explain the log transformation of the response variable, and explore the distributions of key predictors. Section 3 presents our initial regression model using untransformed predictors, highlights assumption violations through residual analysis, and motivates further model refinement. In Section 4, we describe the process of model selection and improvements, including log-transformations and categorical filtering, and show how these changes address diagnostic issues. Section 5 summarizes the results from the final regression model, interprets each coefficient, evaluates model performance, and compares the findings to existing literature. Finally, in Section 6, we draw key conclusions, offer recommendations for stakeholders, discuss limitations, and suggest directions for future research.

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).

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

Predictor	Measurement	Restriction	Mentioned
Bedrooms Above Ground	Count	Two to four bedrooms c(2, 3, 4)	Paper 1
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

Table 2: P-values of Predictors in Final Model

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

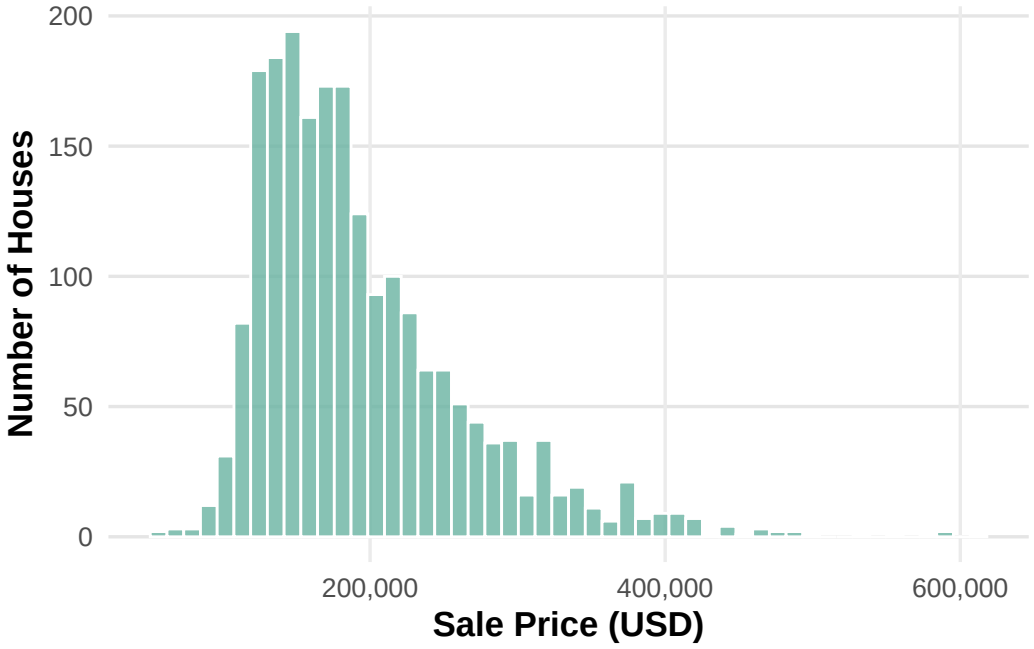

Figure 1: Distribution of House Sale Prices.

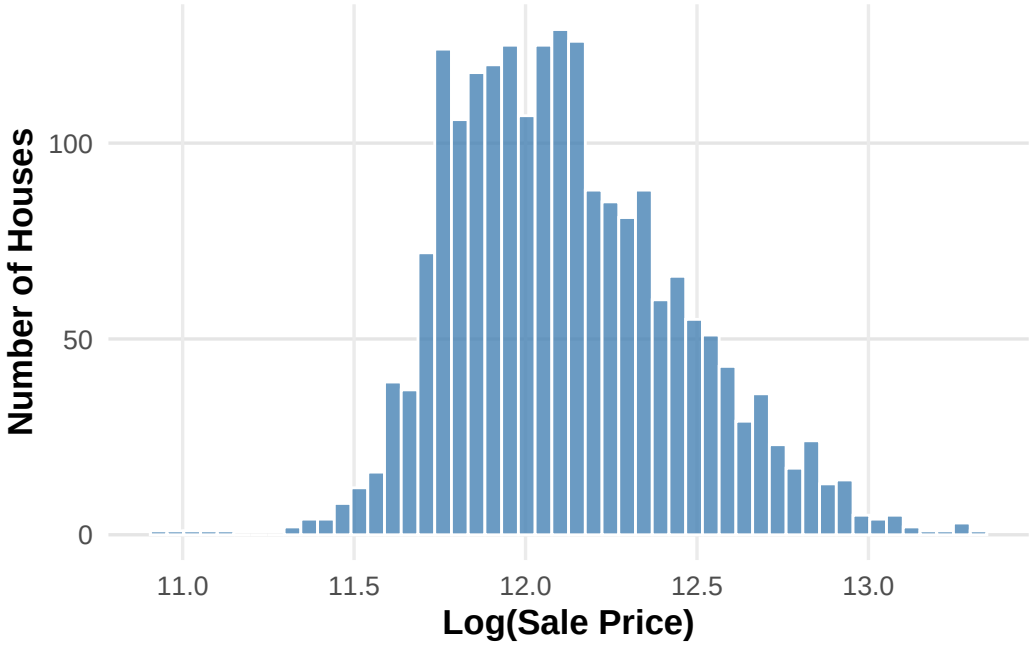

Figure 2: Distribution of Log-Transformed Sale Prices.

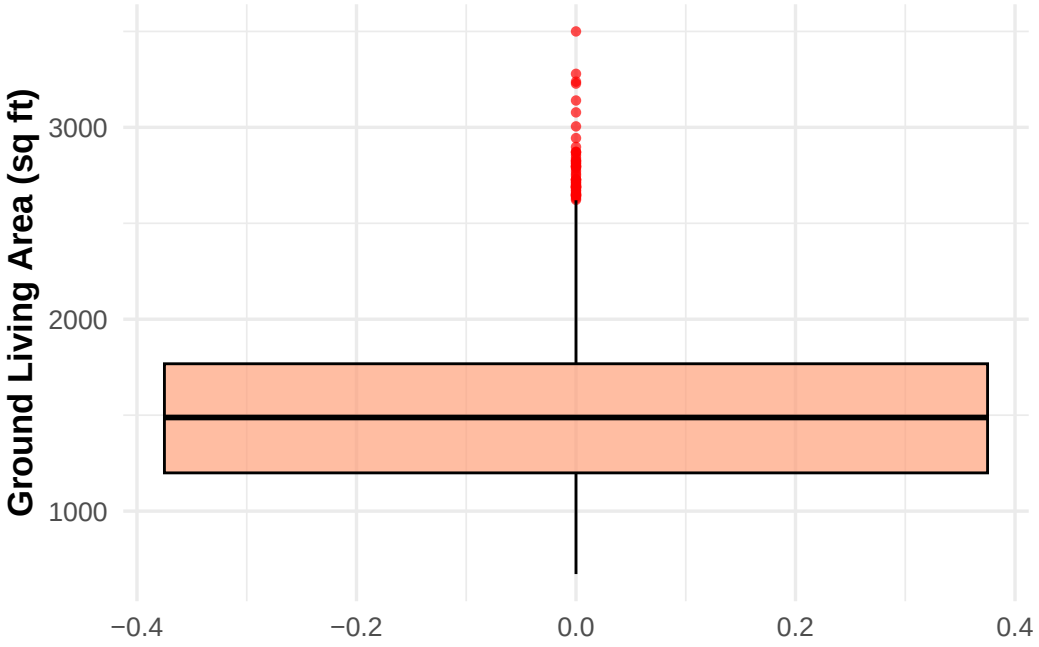

Figure 3: Boxplot of Ground Living Area.

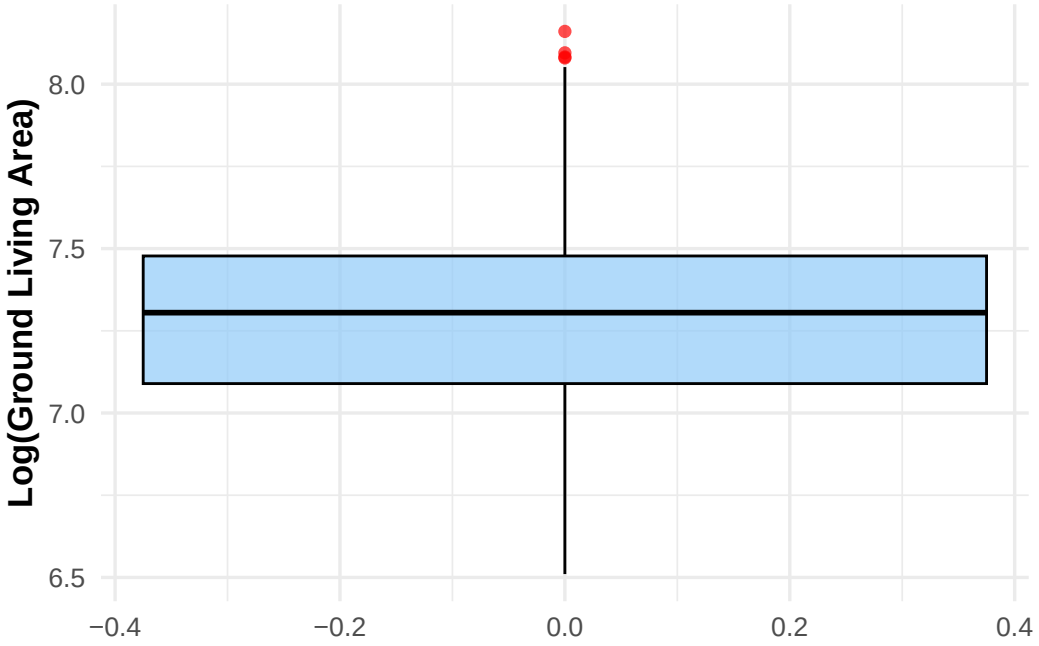

Figure 4: Boxplot of Log-Transformed Ground Living Area.

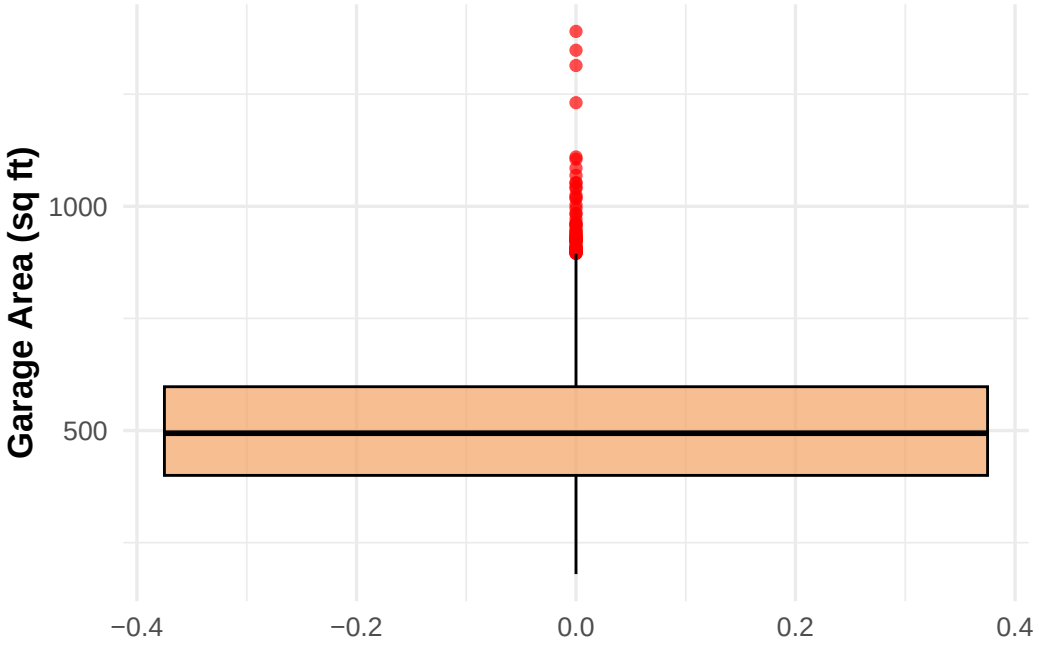

Figure 5: Boxplot of Garage Area.

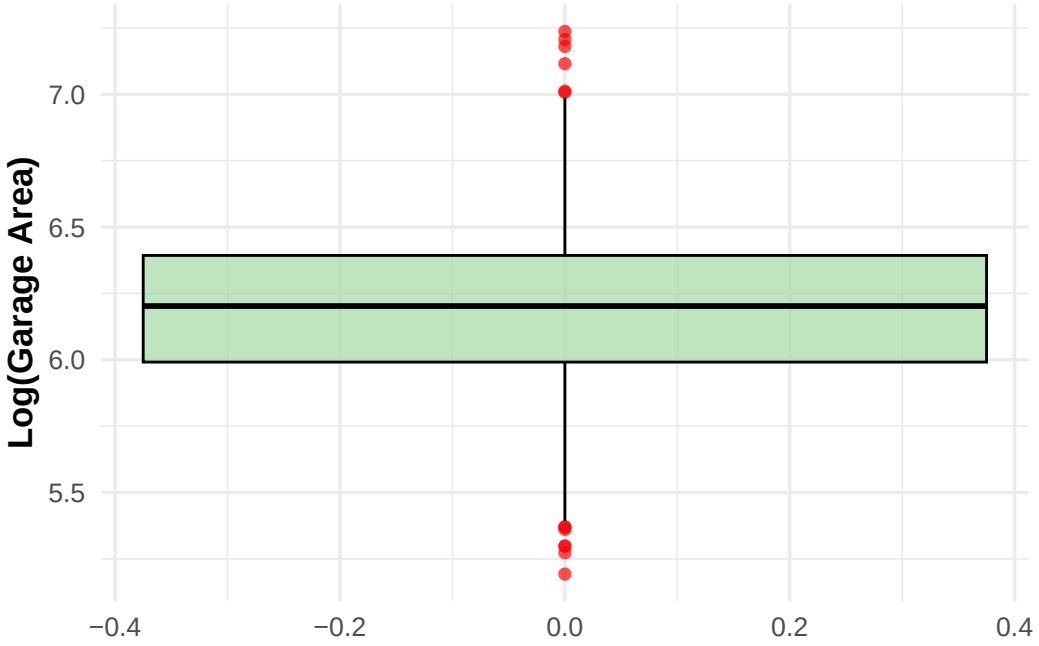

Figure 6: Boxplot of Log-Transformed Garage Area.

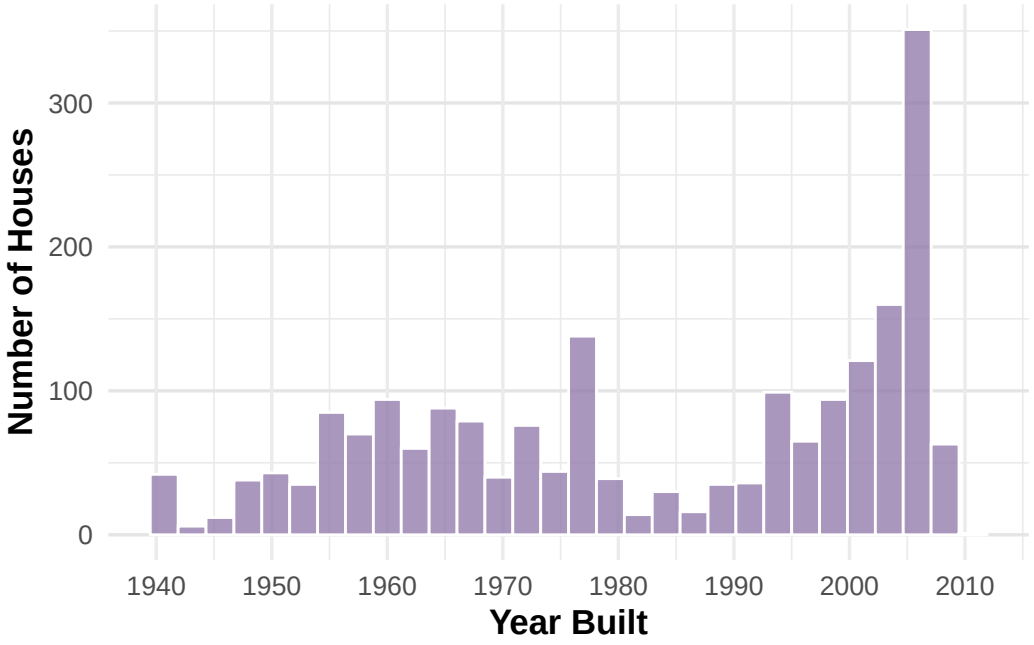

Figure 7: Distribution of House Construction Years.

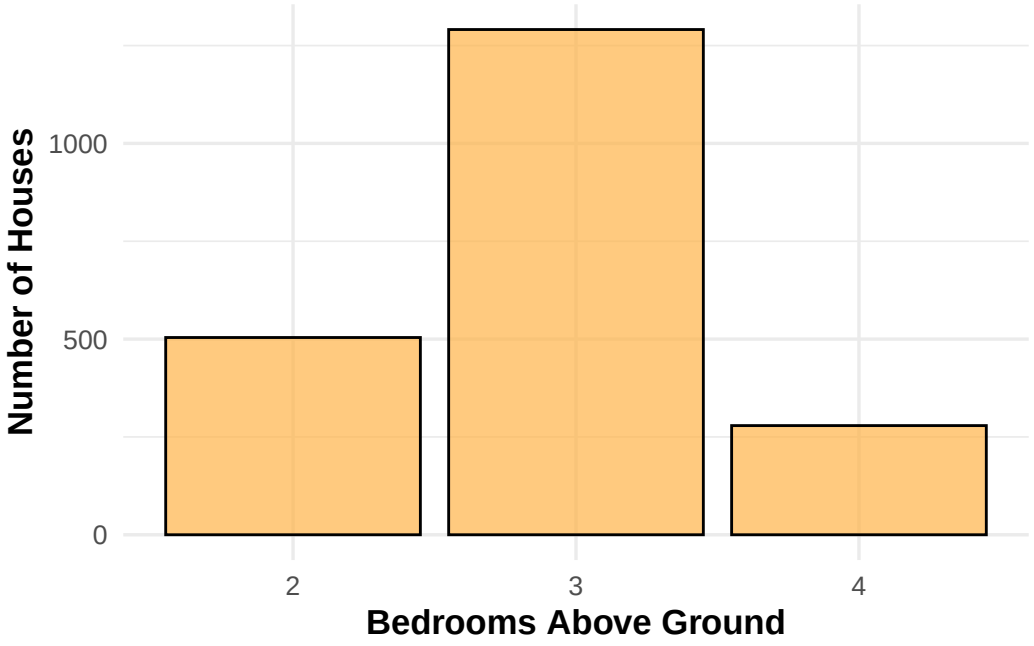

Figure 8: Distribution of Bedrooms Above Ground.

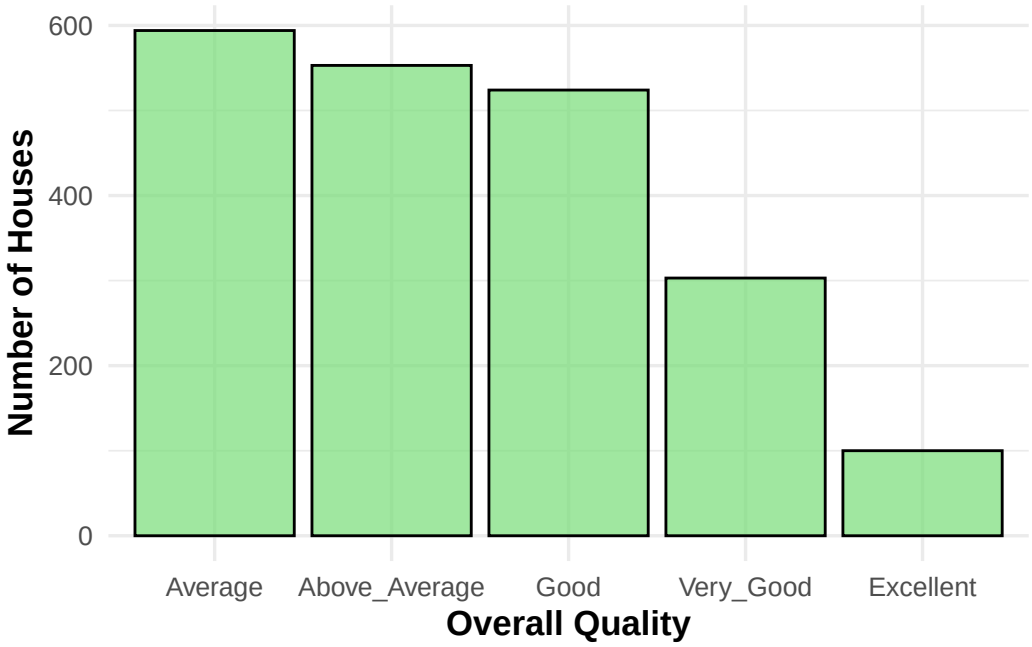

Figure 9: Distribution of Overall Quality Ratings.

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.

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.

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.

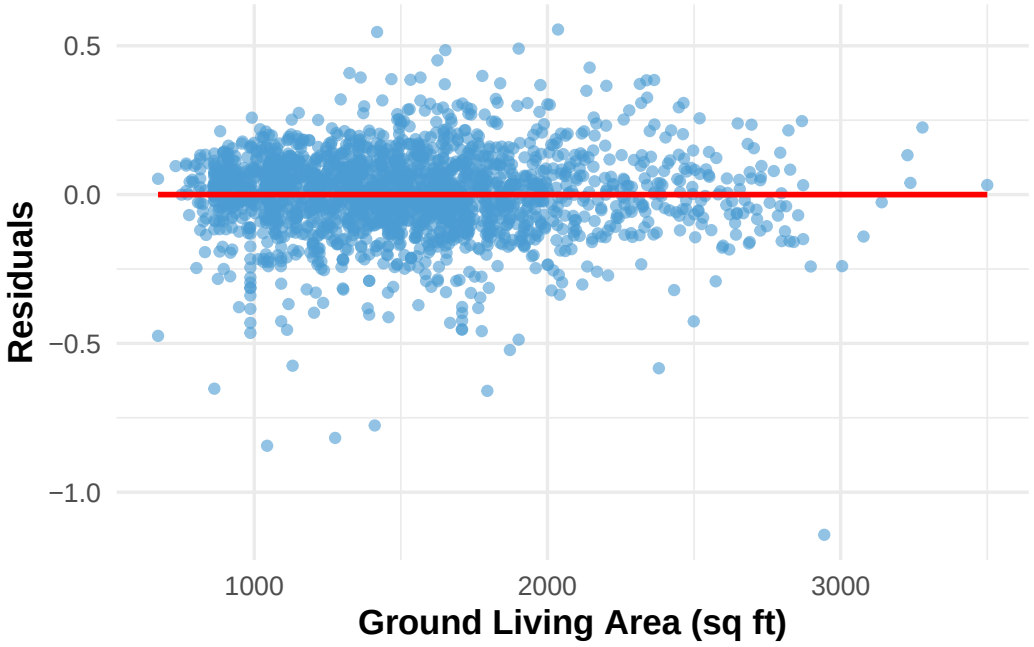

Figure 10: Residuals versus Ground Living Area showing mild left-side clustering.

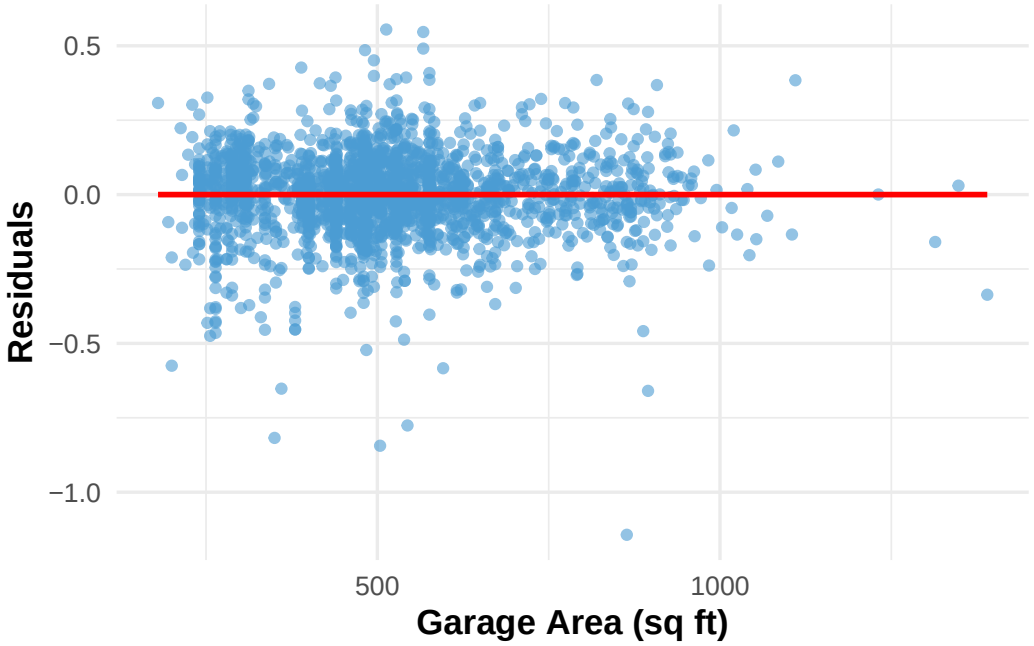

Figure 11: Residuals versus Garage Area showing left-side clustering and heteroscedasticity.

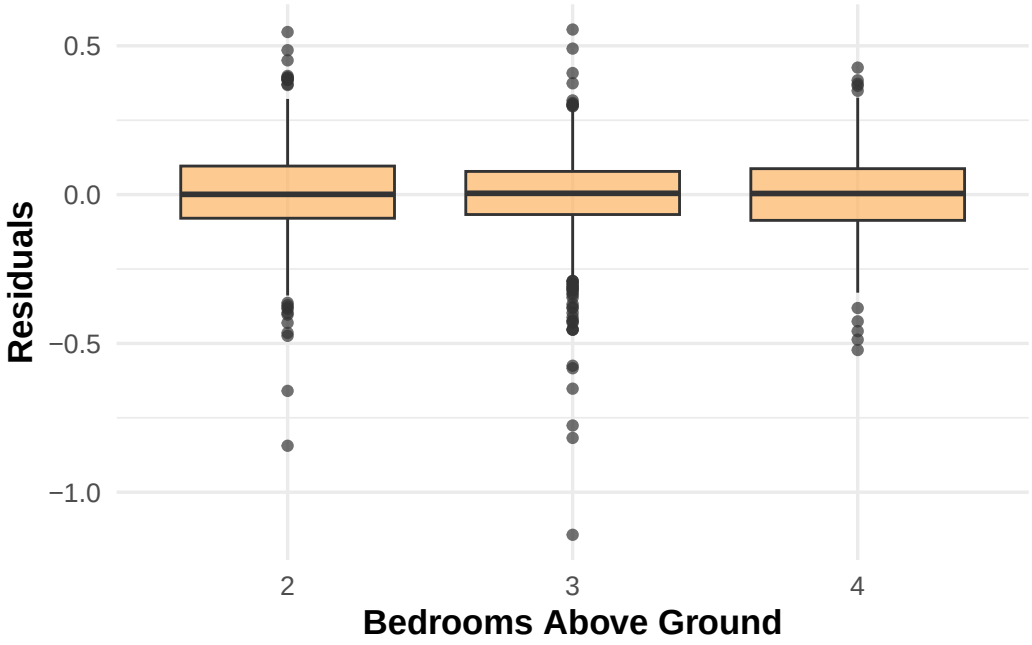

Figure 12: Boxplot of residuals by number of bedrooms above ground.

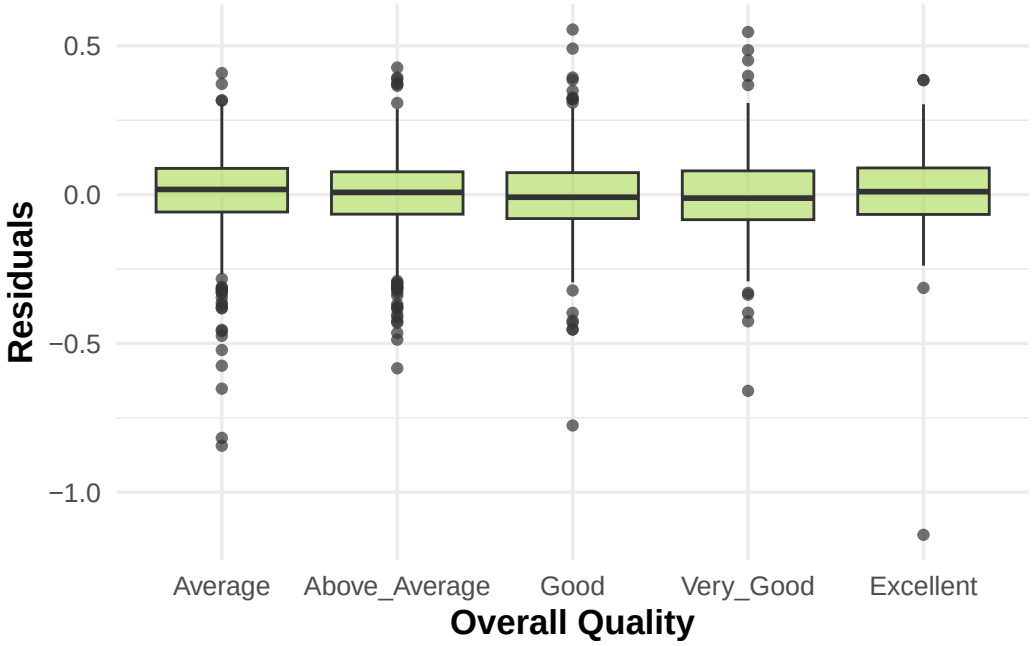

Figure 13: Boxplot of residuals by overall quality rating.

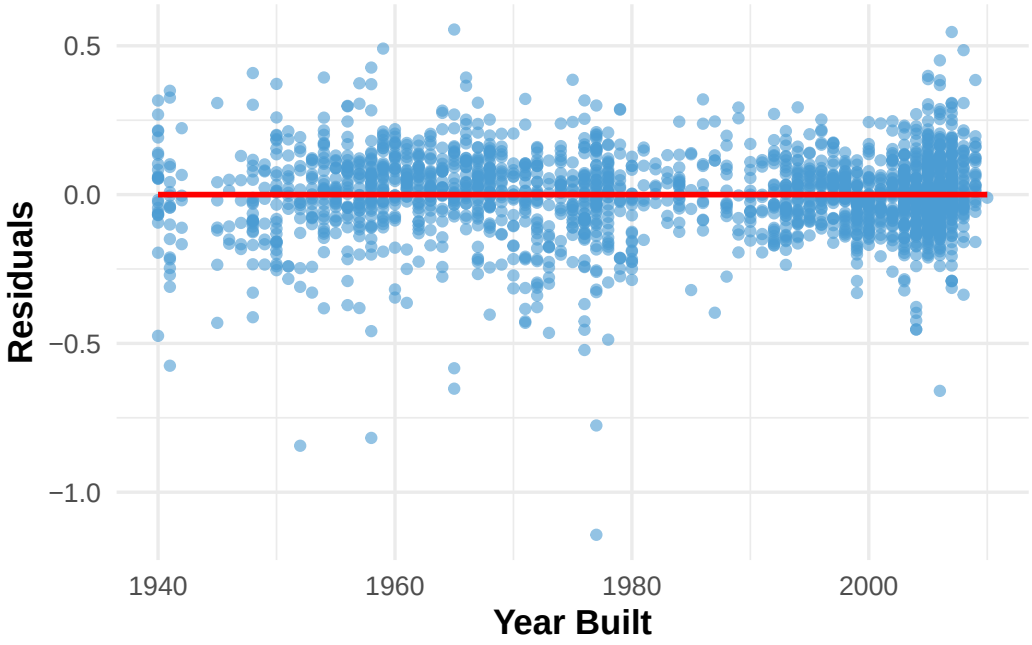

Figure 14: Residuals versus Year Built showing mild structure across time.

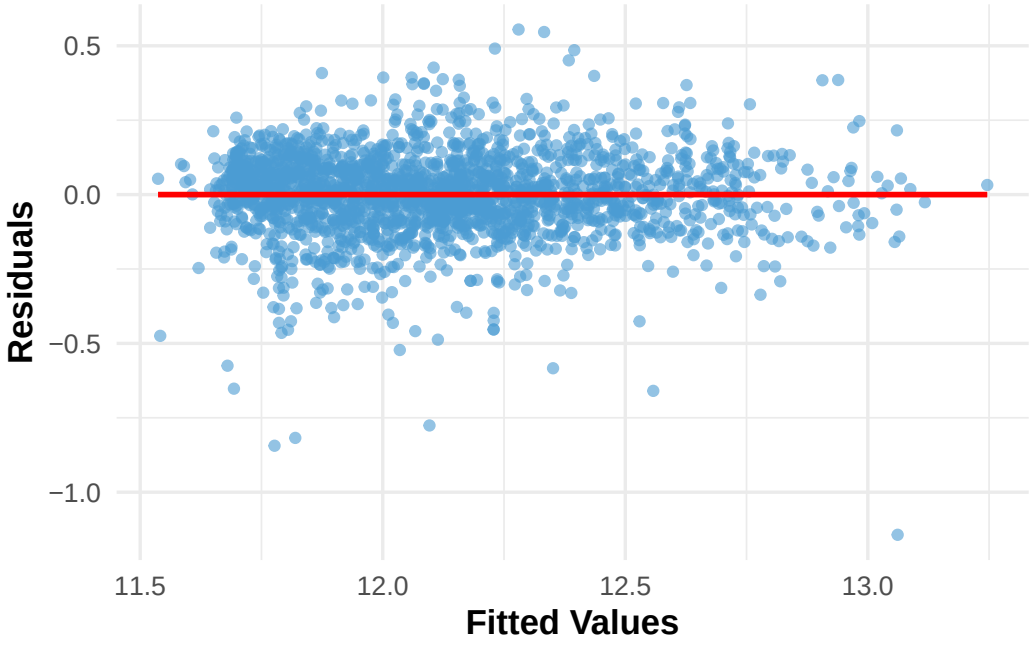

Figure 15: Residuals versus fitted values, assessing model fit and heteroscedasticity.

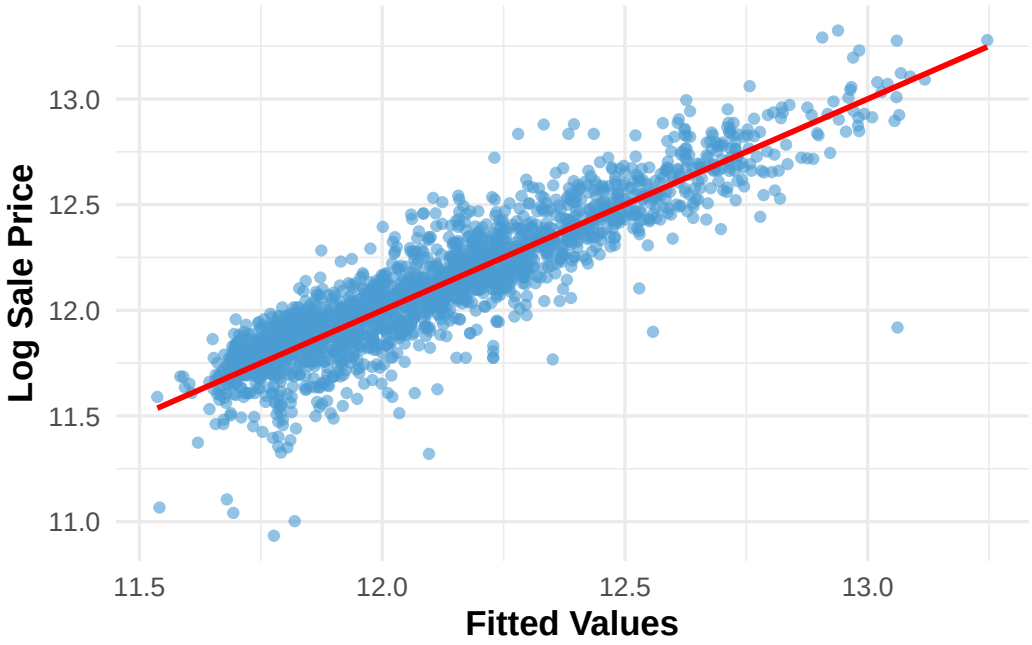

Figure 16: Fitted values versus log sale price, evaluating model prediction alignment.

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.

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.

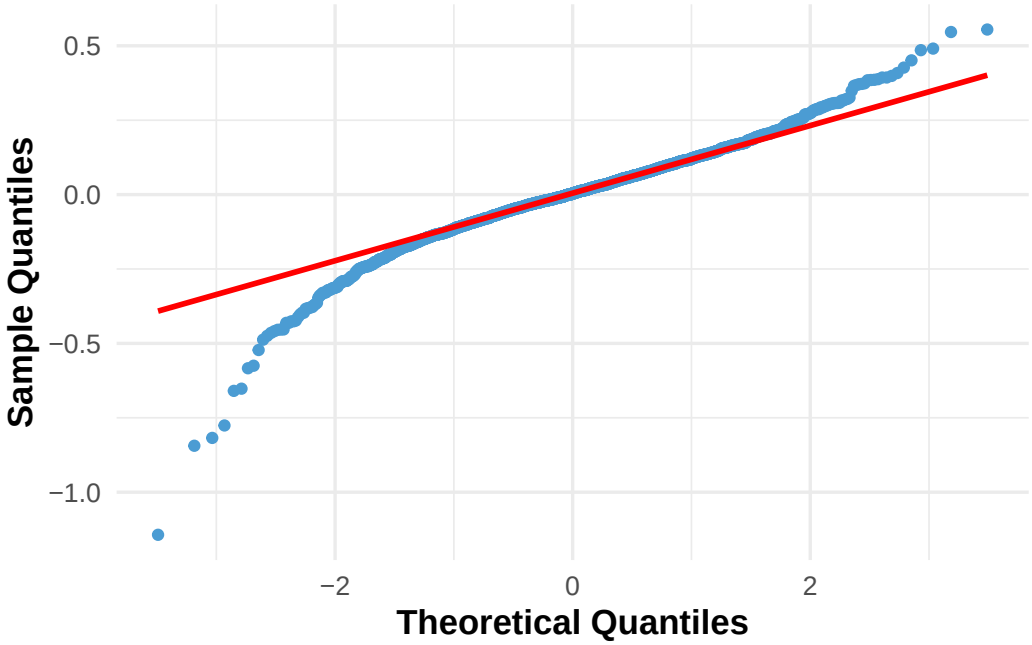

Figure 17: Q-Q plot of residuals to assess normality.

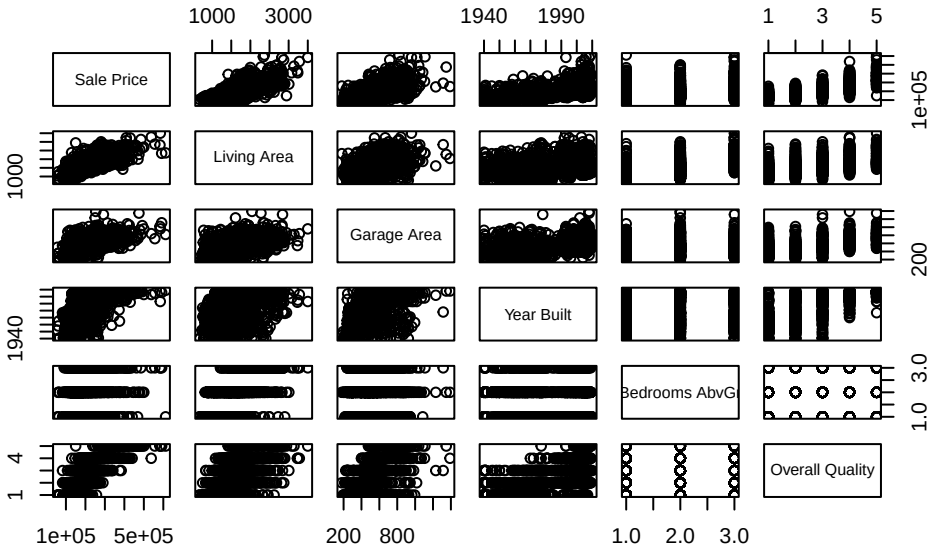

Figure 18: Scatterplot matrix of key predictors and response variable.

4.1 Transformations on Predictor Variables

Both ground living area in Figure 19 and garage area in Figure 20 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.

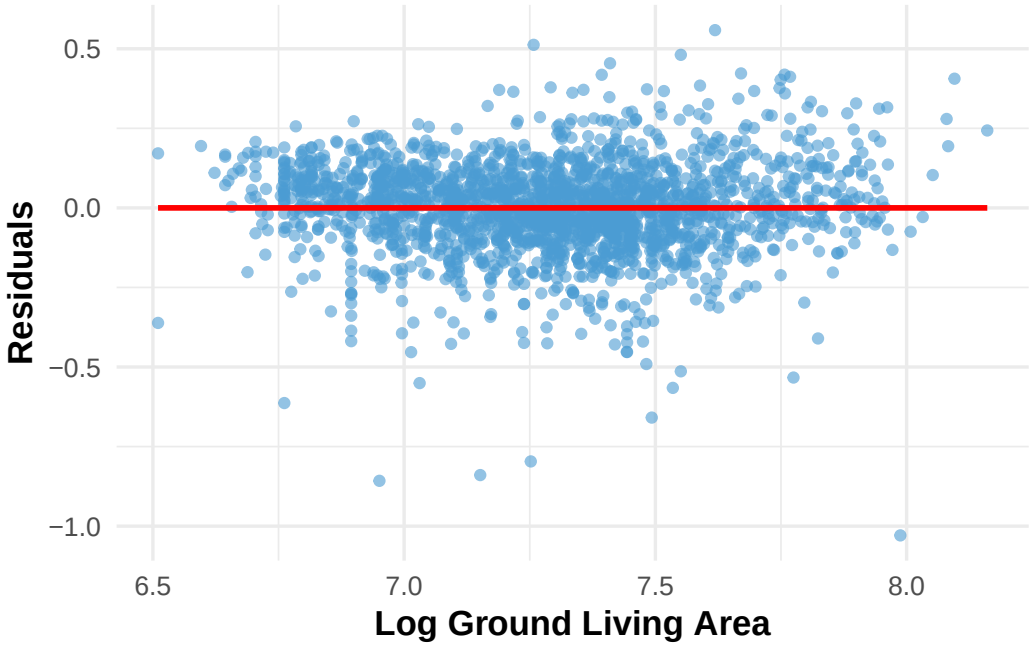

Figure 19: Residuals plotted against log-transformed ground living area to assess linearity and homoscedasticity.

4.2 Categorical Variables and Restrictions

And we continue to limit the number of bedrooms to two, three, and four in Figure 21, and only include houses with rating from average to excellent in Figure 22. The restrictions not only preserved the stability of residuals, but also improved the skewness in the bedroom levels in the updated model with transformed predictors.

4.3 Outliers

With the same data filtering to restrict houses from 1940 or later to remove older houses which likely are outliers in the real world due to depreciation. The residual spread continues to appear

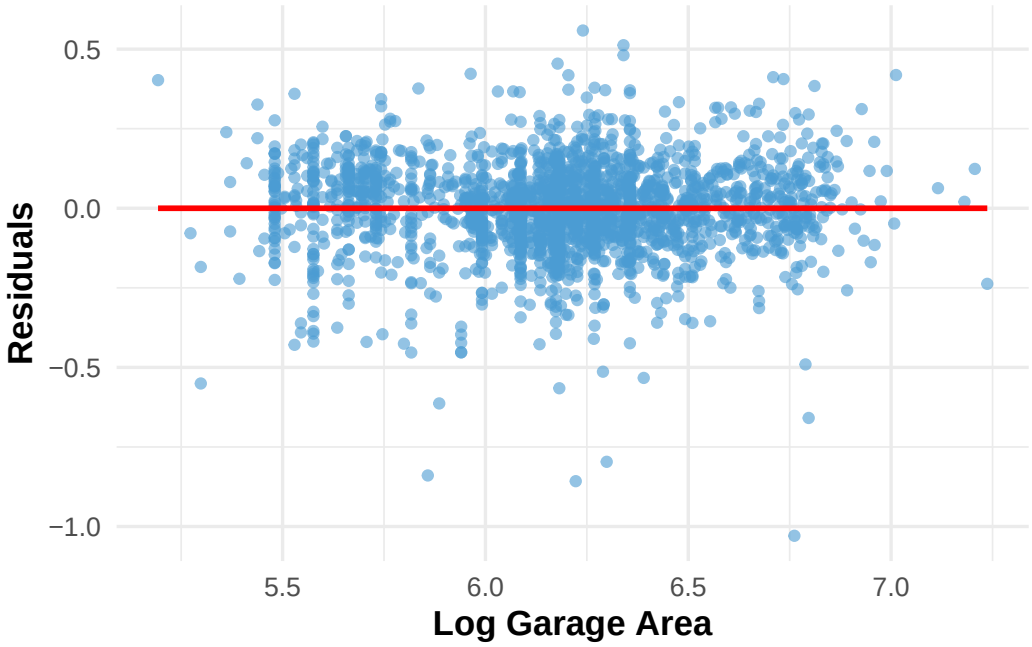

Figure 20: Residuals plotted against log-transformed garage area, examining model fit and variance consistency.

random year by year, showing no visible pattern (Figure 23). However, since this variable is discrete, plus we treat it as a continuous variable since time is an increasing measure, the gaps between some years can be explained by economic factors outside the scope of regression analysis. Which is not a concern for violation of assumptions.

4.4 Fitted Values

For the fitted value residual plot in Figure 24, after the log transformations on garage area and ground living area, the spread now displays no visible pattern and seems randomly scattered.

For the response versus fitted plot in Figure 25, the points now are more randomly scattered along the diagonal, as a result of the log transformation on garage area and ground living area, which reduced the previous deviation.

4.5 Normality and Predictor Relationship

After log-transforming the response, the right-skewed tail is reduced (Figure 26). A left skew still persists after transforming the two predictors suggesting deviations from normal, even though the residual plots do not show it. This suggests that the model slightly underestimates

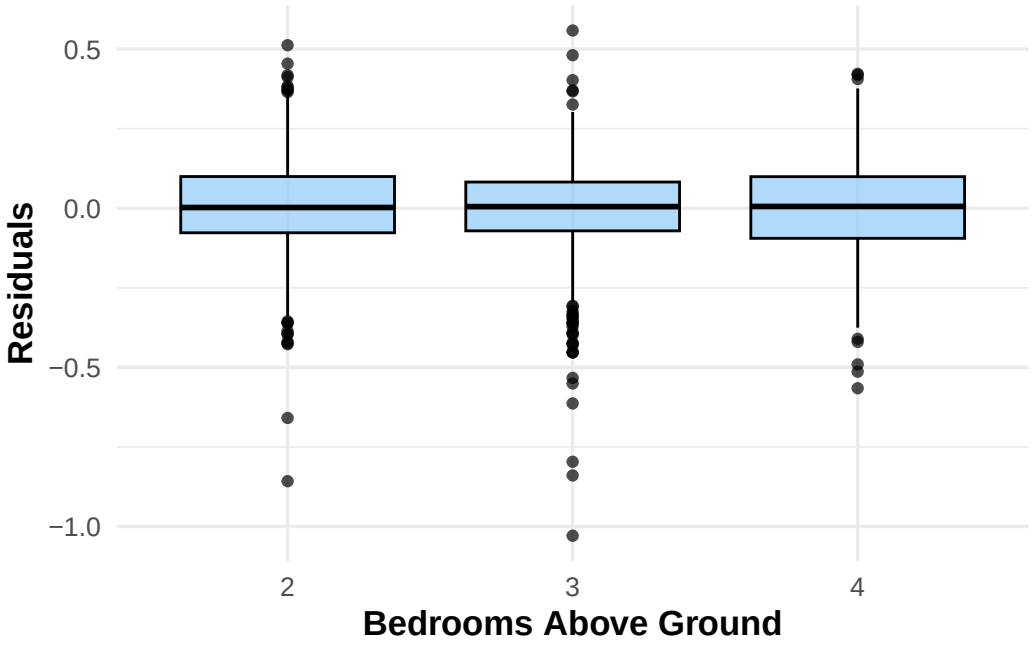

Figure 21: Boxplot of residuals from the transformed model by number of bedrooms above ground.

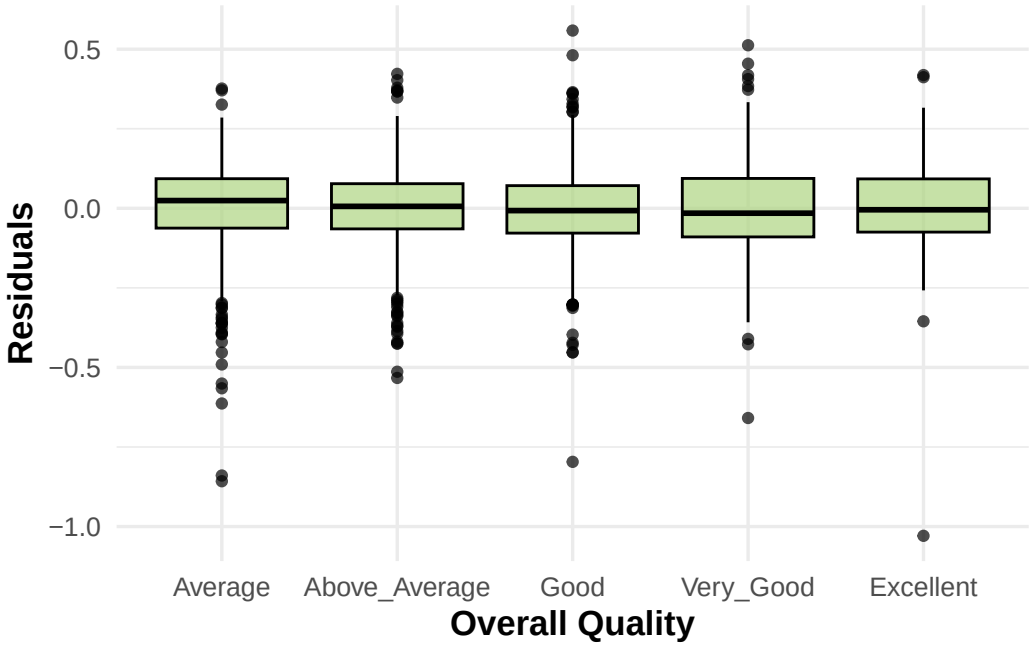

Figure 22: Boxplot of residuals from the transformed model by overall quality rating.

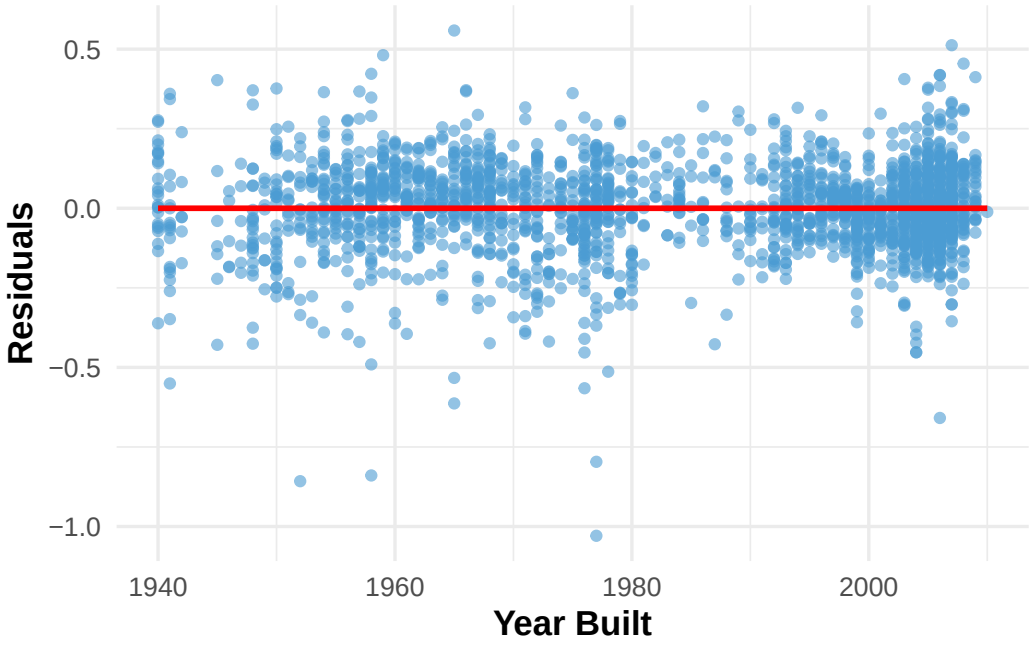

Figure 23: Scatterplot of residuals from the transformed model versus year built, showing no strong temporal structure.

the prices of the lower valued houses. Likely from unmeasured factors like investors perception, which can disproportionate the lower end houses. From the QQ-plot we observed a left tail deviation that could not be addressed even after we performed the transformations which could be a potential limitation to our study. Despite this left tail deviation, the overall distribution is still reasonably normal, with no visible violation from residual patterns.

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
Year Built	0.421	0.545	1.000

In the pairwise plot (Figure 27), the relationship between year built and other variables appears stepped or uneven. This is expected, as year built is a discrete variable treated as continuous, which naturally leads to gaps between data points. Other pairwise comparisons showed weak linear trends or randomness, suggesting that relationships are approximately linear or lack strong patterns. We also examined the correlation matrix (Table 3) for numeric variables and found no high correlations, indicating that multicollinearity is unlikely. Based on these

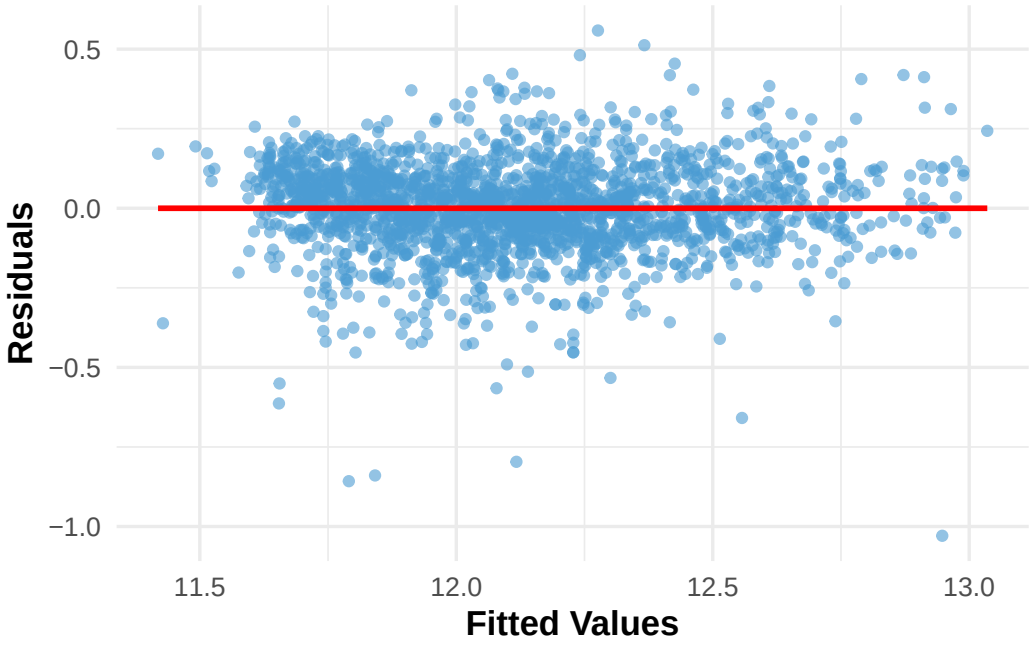

Figure 24: Plot of residuals versus fitted values from the transformed model. A random scatter pattern supports homoscedasticity.

findings, we retained all selected predictors. After applying necessary transformations and restrictions, none of the variables violated model assumptions, resulting in a cleaner and more interpretable model.

5 Final Model Inference and Results

5.1 Model Summary

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

Predictor	Estimate	Std..Error	X95..CI	p.value
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

After testing multiple specifications and addressing potential assumption violations, our final model includes the predictors in Table 4: log ground living area, log garage area, year built, bedrooms above ground, and overall quality. The response variable is the log-transformed sale price which we use to stabilize variance and address skewness from the start.

5.2 Coefficients Interpretation

5.2.1 Intercept

The intercept (2.6813) represents the predicted log sale price of a house that has the base level of each predictor. This corresponds to a house with 2 bedrooms, average quality rating, and zero values for the continuous predictors (i.e., log ground living area, log garage area, and year built). The intercept itself may not be meaningful on its own, but it serves as the base value from which other effects are added.

5.2.2 Log Ground Living Area

The coefficient for log of ground living area is 0.5325. This means that a 1% increase in the ground living area is related with about that amount (0.5325%) increase in sale price, holding all other variables constant. Note that in our case the variable is log transformed, the interpretation changes from absolute nominal change to percent change. This is one of the most important predictors in our model, suggesting the idea that the overall size of a house is a critical factor in determining the price value.

5.2.3 Log Garage Area

The coefficient for log of garage area is 0.1480, which again means that a 1% increase in garage area increases the sale price by 0.1480%. While the effect is smaller than the ground living area, it is still positive meaning it is still statistically significant. This implies that investors/buyers still value garage area, likely due to the practicality like storage of extra things not put in house, or space to install protections for cars.

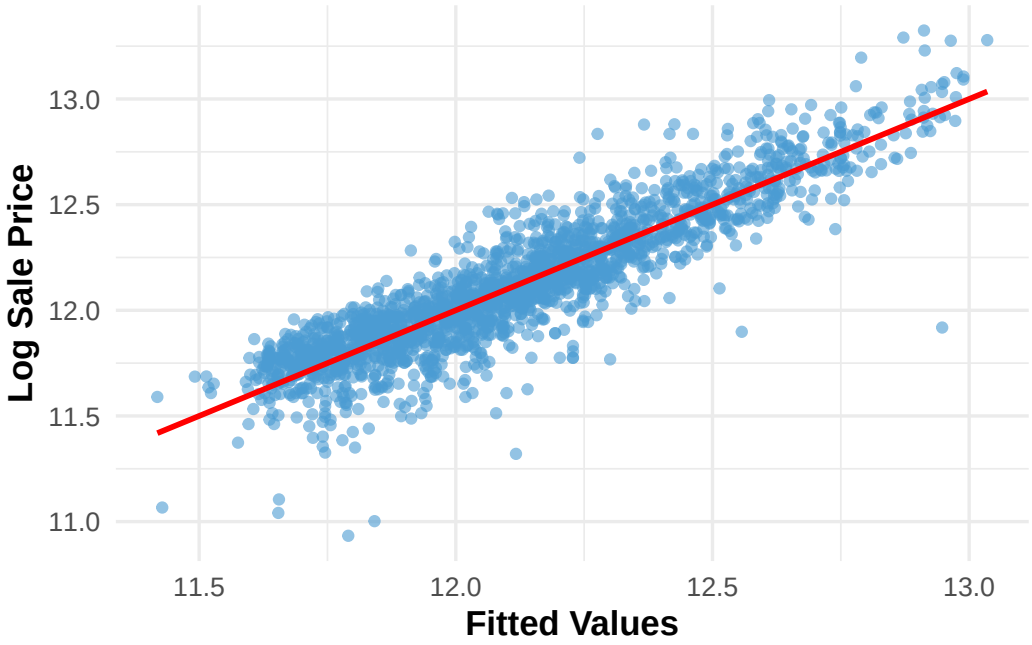

Figure 25: Plot of fitted values versus log-transformed sale prices. The close alignment with the red trend line indicates good model fit.

5.2.4 Year Built

The estimate of 0.0023 for year built suggests that on average each additional year of construction adds 0.23% to the log price of the house. This reflects that people prefer newer houses which are usually built with more modern materials, design and inside features. This coefficient is small but precise, with a very narrow confidence interval.

5.2.5 Bedrooms Above Ground

These coefficients are interpreted relative to the base of two bedrooms above ground. The effect of having three bedrooms is slightly negative (-0.0156), and four bedrooms an even larger negative effect (-0.0721), both compared to two bedroom houses. While the coefficient for three bedrooms is above the 0.05 threshold by a tiny bit, it is considered not significant enough, but the result for four bedrooms is clearly significant. These results suggest that within our filtered sample of average to excellent quality houses, more bedrooms does not necessarily always mean more value. It could reflect trade-offs like smaller living spaces per room for example.

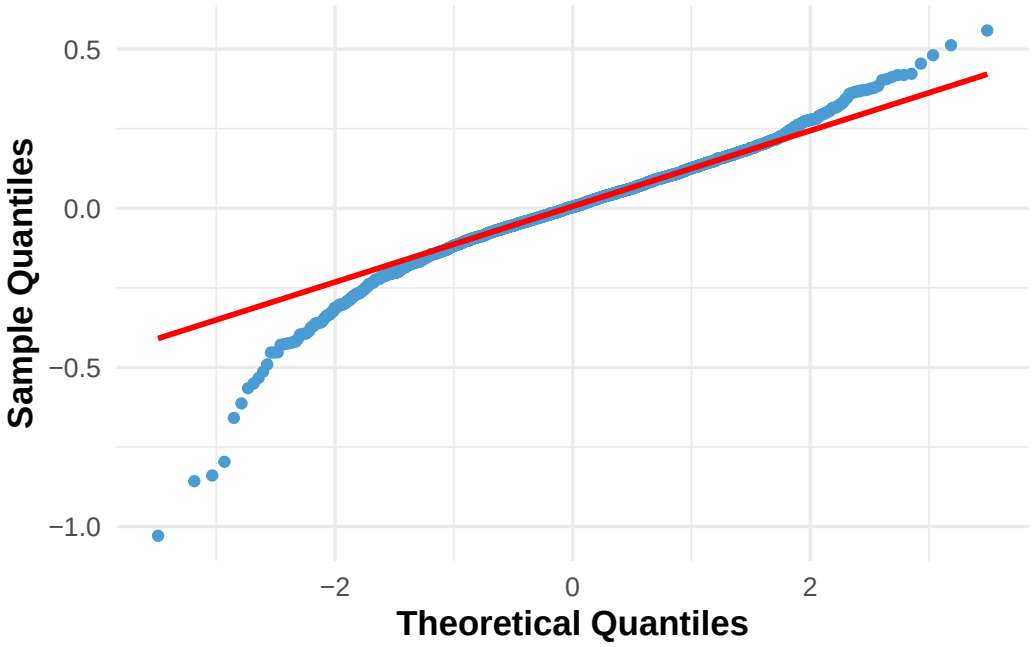

Figure 26: QQ plot of residuals from the final model. The points lie close to the red line, supporting normality.

5.2.6 Overall Quality

Average quality as the base level:

- "Above Average" houses sell for 3.22% more
- "Good" houses sell for 11.31% more
- "Very Good" houses sell for 26.89% more
- "Excellent" houses sell for 48.33% more

The increasing pattern absolutely reflects the importance of quality of material and craftsmanship in houses. The results show a strong and consistent upward trend in sale price as quality improves, and all levels of quality are statistically significant with approximately consistent narrowness of confidence intervals through all levels.

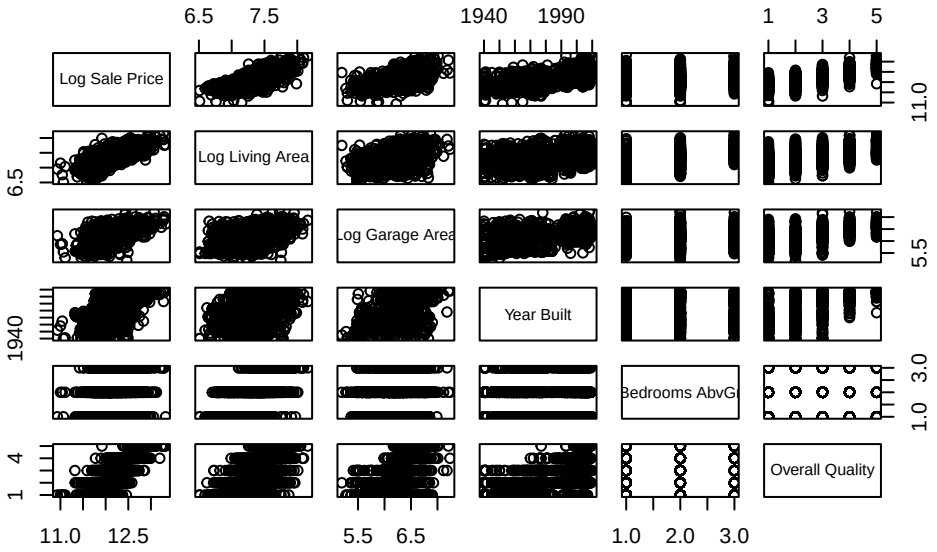

Figure 27: Pairwise scatterplots for predictors and log sale price. Clear linear relationships are observed, especially with log-transformed variables.

5.3 Comparison to Literature

The results of our model align well with findings from previous studies. Both (Ye 2024) and (Han 2023) say that ground living area, garage area, and overall quality of houses were the most influential predictors of house prices in the Ames dataset. Our final model results show this: ground living area had the highest coefficient among continuous predictors, and overall quality shows the largest effect sizes across the model. Han also reported that newer houses have higher prices, which aligns with our positive and significant coefficient of house built.

The only potential meaningful difference from the literature is the negative relationship between bedroom count and price. (Shukla 2024) found that more bedrooms were related with higher prices, likely because bedroom count can be an indirect measure for size of house. However, since our model includes ground living area, the added value of additional bedrooms may be less direct, hence or even negative if more bedrooms means smaller rooms or less free space. This difference highlights the importance of considering multiple dimensions of size and layout when modelling house value.

Additionally, studies from literature used machine learning models which are more complex and can potentially sacrifice some interpretability for predictors compared to our linear model, that offers straightforward interpretations of how each variable contributes to house price, which is useful for real world policy making and buyer decisions.

5.4 Model Performance

Table 5: Model performance statistics: R-squared, Adjusted R-squared, and AIC.

R-squared	0.8190
Adjusted R-squared	0.8182
AIC	-2184.5000

Our final model achieves strong statistical performance (Table 5). The R-squared value is 0.819, meaning that approximately 81.9% of the variation in log sale price is explained by the predictors. This is a strong level of explanatory power especially in housing/real estate modelling where prices are influenced by many complex and often unobservable factors.

The Adjusted R-squared is 0.8182 which is very close to the R-squared. This indicates that our model does not have overfitting problems despite having many predictors. Each included variable adds meaningful explanatory power without increasing the complexity of the model unnecessarily.

The Akaike Information Criterion (AIC) is -2184.5, which is a strong indicator of model fit. Lower values of AIC suggest better trade-offs between fit and complexity. The high negative value suggests the model performs very well on the log transformed response. If we kept the predictors untransformed with skewed distribution like in the preliminary results, the AIC would likely be a lot higher.

In addition, previous visual inspection of residual plots supports the model's performance. Residuals are randomly scattered in plots vs. fitted values, and no obvious non-linear patterns which suggest heteroscedasticity are present. While the Q-Q plot has some left tail deviation as explained in the previous section, suggesting possible underestimation of prices for lower value houses, the overall distribution is again reasonably normal and does not have major concern.

Overall, our final model shows that our chosen predictors significantly affect house prices. The log transformed variables allow for percentage-based interpretation which gives practical insights into how each attribute contributes to price. The model performance is strong as shown by the statistics, and the results are mostly consistent with literature findings.

6 Discussion and Conclusion

Our study explores the relationship between house attributes and the sale prices of houses in Ames, Iowa. Using our carefully filtered dataset and a multiple linear regression model, we focused on how factors like ground living area, garage area, number of bedrooms, year built, and overall quality influence house prices. All predictors were chosen based on the literature we

found, exploratory data analysis, and real world knowledge, with transformations and filtering applied as needed to meet linear regression assumptions.

6.1 Main Conclusions

Our findings from our final model support several main conclusions. First, ground living area comes out as the most influential continuous predictor. A 1% increase in ground living area is related to about a 0.53% increase in house price, holding other variables constant. This aligns with real world economic expectations and confirms that investors/buyers value interior living space very highly.

Second, overall quality is strongly positively related to price. Compared to houses with average ratings, those with excellent quality ratings have nearly a 48.3% premium on price. This clear rise observed across the quality levels suggests that buyers are highly sensitive to the materials, finish, and overall craftsmanship of houses.

Third, garage area also contributes positively to price, though its effect is less strong. A 1% increase in garage area corresponds to about a 0.15% increase in house price. While garage areas are not as prioritized for valuation as living space or quality, they still play an important role especially in real world contexts where car ownership is the norm.

Also, year built also showed a significant but small effect, each additional year corresponds to about a 0.23% increase in house price. This supports the general belief that newer houses are more desirable due to more modern finishes, better layouts, and more energy efficiency. However, the small size of this coefficient suggests that age alone does not dictate price, it must be considered along with quality and size as well.

The most surprising finding was the negative relationship between bedroom count and price. In our filtered sample of houses with two to four bedrooms, having three or four bedrooms is associated with slightly lower prices than having just two. This result is counterintuitive and appears to contradict common housing assumptions and beliefs. Again, It may be explained by trade-offs in design, houses with more bedrooms may have smaller individual rooms or less free open space, reducing the value. Or, it may show buyer preferences in this particular market in Iowa.

6.2 Supporting Evidence

Our findings are mostly consistent with prior research. (Han 2023) and (Ye 2024), who both analyzed the same dataset using more complex models, says that ground living area, garage area, overall quality, and year built as the most important variables in predicting house prices. Our model confirmed this conclusion with a linear model, showing that even simple models can offer high explanatory power and valuable conclusions when the model is correctly specified.

The upward trend seen in overall quality, from average to excellent, supports findings from (Ye 2024) and (Shukla 2024), who emphasized the importance of qualitative attributes in driving buyer behavior. In a real world housing market, buyers are willing to pay high premiums for houses that are seen to require less maintenance or improvements and that offer more luxurious and durable finishes.

As from earlier, our result about bedrooms deviates from the trend found by (Shukla 2024), where bedroom count had a clear positive impact on price. We believe this is due to our model already controlling the overall size through ground living area, which makes additional bedrooms redundant or maybe even detrimental in value. This highlights the importance of understanding how multiple predictors interact in linear regression.

6.3 Recommendations

These findings can be used to help with decision-making by house sellers, buyers, and policy-makers.

For sellers, investing in quality improvements may offer a higher return on investment than simply adding rooms or expanding garage space. Upgrades to materials, finishes, or even design may increase a home's market value more effectively than increasing bedroom count or building other extensions. Similarly, emphasizing free living space, rather than maximizing total area may align better with investor preferences.

For buyers, the results suggest that buyers often pay premiums for subjective things like perceived quality, not just objective attributes like size. This shows the importance of evaluating a house fully rather than simply by number of bedrooms or total area.

For policymakers, the findings may justify incentives for focusing on quality of construction of houses rather than density alone. Encouraging development of wellbuilt midsize houses rather than maximizing floor capacity may better align with market preferences and produce more sustainable long-term housing value.

6.4 Improvements

While our model performed well, achieving an R-squared of 0.819, and Adjusted R-squared of 0.8182, and an AIC of -2184.5, it is not without limitations.

First, we rely on structural features and do not include location based factors, which are known to heavily influence house prices. Attributes like neighborhood for example are missing from the model. Including geographic information or neighborhood level variables can definitely improve the model's explanatory power and capture important omitted variables.

Second, some variables we removed from filtering may still play a meaningful role in price. While we excluded many variable levels in favour of the most desired ones, a more flexible

model might still capture the meaningful effect of them, especially in areas that are less urban.

Third, the left tail deviation in the Q-Q plot suggests the model underestimates prices for the lower priced houses. This could be due to exclusion of lower quality houses from our dataset, omitted variables affecting undervalued houses, and nonlinear effects that are not captured by linear models.

Although the residual plots suggest our model assumptions are met, using more modeling techniques could explore whether nonlinear patterns or interactions between predictors can yield better accuracy in predicting.

Additionally, our analysis uses data from 2006 to 2010. Housing market has changed quite a lot since then, so the exact coefficients may not generalize to present day valuations. However, the relative importance of variables like size and quality should still remain consistent over time.

Finally, using a multilinear regression model for interpretability and transparency is our main goal within the course scope. While machine learning models used in the literature may outperform, they can lack interpretability and require more complex techniques. The trade-off is justified in our case because our main goal was to not only forecast but also understand housing prices.

6.5 Closing Summary

In conclusion, our analysis highlights the relative importance of different house attributes in predicting house prices in Ames, Iowa. Ground living area and overall quality are the strongest drivers of price, while garage area and year built also play important roles. The number of bedrooms unexpectedly has a weak or even negative relationship with price when controlling for size and quality. These findings are supported by peer-reviewed literature and offer useful insights in the real world.

Our model has both accuracy and interpretability balanced, explaining over 80% of the variance in sale prices of houses while maintaining clear and meaningful coefficient interpretations. With additional location based data and more modelling techniques, future studies can build from this baseline to better understand housing markets both in Ames and the entire world.

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