Ames, Iowa Housing Price Analysis

Katherine Nguyen

DS6371 – Statistical Foundations for Data Science

Dallas, TX

Introduction

The Ames Housing Price dataset contains comprehensive data on housing sales in Ames, Iowa, and provides valuable insights into how various factors influence home prices. This dataset includes attributes such as the square footage of living areas, neighborhood types, number of bathrooms, and other structural and location-related features. It is especially useful for analyzing the relationship between home features and their sale prices, as well as for identifying trends in specific neighborhoods. For instance, Century 21 Ames is interested in understanding how the square footage of living areas impacts the sale price within three specific neighborhoods: North Ames, Edwards, and Brookside. A key part of this analysis is to explore whether the relationship between sale price and living area varies across these neighborhoods. Considering the breadth of this data set, building predictive models at varying levels is possible for house prices across all neighborhoods in Ames, IA. By creating and comparing multiple linear regression models—one simple and two multiple regressions using different sets of variables—this analysis identifies the most effective model for predicting home prices in the area. Both analyses require careful evaluation of model assumptions, identification of potential outliers, and a well-supported response to what factors most strongly influence house prices in Ames.

Data Description

The Ames House Price dataset contains 1,460 data points and 81 variables, each representing various features of homes in Ames, Iowa. The dataset includes both continuous and categorical variables, capturing a wide range of attributes such as the size of the living area, the number of bedrooms, the year of construction, the neighborhood, and the condition of the property. Additionally, it includes variables related to the home's exterior, foundation, basement, and other structural details, as well as factors like the number of bathrooms and the presence of a garage. The target variable, SalePrice, represents the sale price of the house and is the primary focus for predictive modeling. Key variables for the multiple linear regression analysis include GrLivArea (above-ground living area in square feet), OverallQual (overall material and finish quality), FullBath (full bathrooms above grade), GarageCars (size of garage in car capacity), KitchenQual (kitchen quality), ExterQual (exterior material quality), and 1stFlrSF (first floor square feet), which are relevant for predicting sales prices of homes in all of Ames, Iowa. Additionally, Neighborhood is a key categorical variable for analysis 1 to examine how location affects price, particularly for areas covered by Century 21 Ames. For a more detailed description of all variables, including their meanings and specific codes for categorical variables, refer to the data description file available on the Kaggle website, which provides comprehensive information on each feature in the dataset. With a mix of numerical, categorical, and ordinal features, this dataset provides a comprehensive foundation for understanding the key factors that influence home prices in Ames.

Analysis Question 1

> **Problem:** Century 21 Ames, a real estate company based in Ames, Iowa, has asked for an analysis to better understand the relationship between the sale price of homes and the square footage of their living areas. The company is particularly interested in homes located in the Names (North Ames), Edwards, and BrkSide (Brookside) neighborhoods, outlined in figure 1. They want to know how the sale price correlates with living area size and whether this relationship varies across these neighborhoods. The task is to build and fit a model that estimates this relationship, providing confidence intervals for any estimates.

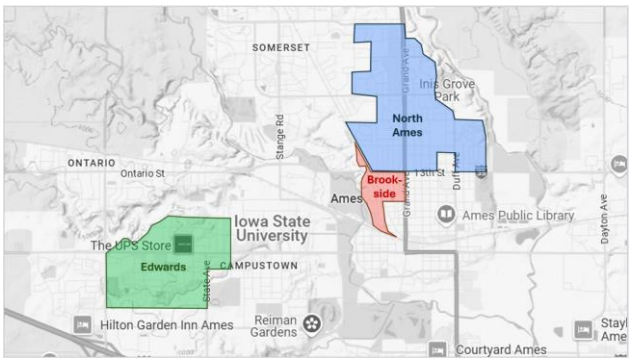


Fig. 1 *Illustration of Century 21 Ames neighborhoods*

Building and Fitting the Model: The analysis provided involves understanding the relationship between house sale price and living area (square footage) across three neighborhoods: NAmes, Edwards, and BrkSide, to assess if this relationship varies by location. The process starts with data filtering and cleaning. The dataset was restricted to houses in the specified neighborhoods, and records with missing values in key variables such as SalePrice and GrLivArea were excluded if present (none were found).

To normalize the target variable, the sale prices were log-transformed, creating a new variable LogSalePrice. This transformation, illustrated in fig. 2, is critical for meeting the normality assumption of linear regression.

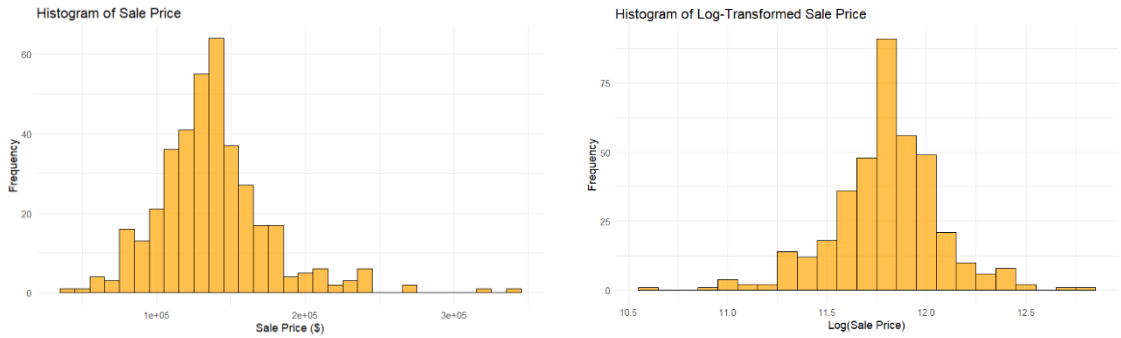


Fig. 2 *Normality comparison between SalePrice and Log(SalePrice)*

Additionally, outliers were identified based on initial scatterplots (fig. 3), particularly focusing on points that might unduly influence the regression results. These outliers, particularly the points above 4000 sqft, were removed to ensure the analysis focused on typical patterns in the data.

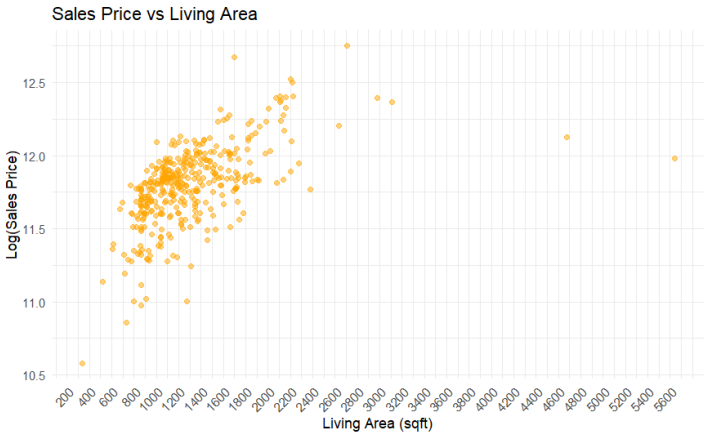


Fig. 3 *Scatter plot of original data before removal of outliers*

Assumptions: The diagnostic checks (Appendix A-1) affirm that all the models' assumptions are adequately met. Residual plots display no discernible patterns, confirming homoscedasticity and independence. Q-Q plots demonstrate that the residuals closely follow a normal distribution, validating the normality assumption. Additionally, histograms of residuals (Appendix A-2) align with a normal distribution curve, further supporting this conclusion. Outlier analysis involved inspecting leverage and influence measures, such as Cook's Distance. There do not appear to be any more extremely influential points that should be removed.

Comparing Competing Models: Separate linear regression models were developed for each neighborhood, with LogSalePrice as the dependent variable and GrLivArea as the explanatory variable. The results in table 1 summarize the performance of each model. The adjusted R^2 values describe the proportion of variance in LogSalePrice explained in each model while accounting for the number of predictors. The Brookside model performs the best with highest adjusted R^2 value, 0.6734, across the models, indicating that the above ground living area explains 67.34% of the variance. The internal CV PRESS (cross validation predicted residual error sum of squares) values reflect prediction errors based on cross-validation, further highlighting the accuracy that the Brookside model predicts with its lowest value. This shows that the model's predictions are closer to actual values for Brookside compared to the other neighborhoods. On the other hand, the combined neighborhoods model has the highest CV PRESS value, suggesting that separating neighborhoods improves predictive performance.

The AIC (akaike information criterion) value which balances goodness-of-fit with model complexity (lower values indicate better performance), reveals that the North Ames model achieves the best balance of fit and simplicity, with the lowest AIC. Brookside has a higher AIC, suggesting slightly more model complexity, while Edwards performs the worst with a positive AIC, reflecting both poorer fit and greater complexity.

Overall, these metrics demonstrate that modeling neighborhoods separately enhances performance compared to a combined model. The Brookside model explains the most variance and is the most accurate in prediction, though North Ames achieves the best balance between fit and simplicity. Edwards, in contrast, shows weaker predictive strength and fit. These results highlight the dependency of considering individual neighborhoods when modeling the relationship between living area and sale price.

Table 1. *Comparing Model Statistics*

Models	Adjusted R^2	Internal CV PRESS	AIC
All 3 Neighborhoods	0.4119	308771210722.523	-104.94
North Ames	0.4186	132149776632.633	-189.37
Edwards	0.3773	108435928010.778	5.6723
Brookside	0.6734	29704361945.6668	-19.352

Parameters:

The estimated coefficients reveal that for every additional square foot of living area, the percentage increase in sale price (approximated using the log-transformed coefficients) varies by neighborhood. Comparing to the estimate of the model that combines all three models (table 2), one can conclude that the SalePrice does depend on neighborhood.

In general, across all three neighborhoods covered by Century 21 Ames, for a 100 square foot increase in the above ground living area, the median sale price increases by 4.43% with a 95% confidence that the increase is between 3.90% and 4.97%, described in table 2. It is estimated that for an above ground living area of zero, the predicted sale price is $e^{11.23}$, or $75,357.60. However, this is extrapolation, as no above ground living areas of 0 square feet are in the data set. The model can be seen in equation 1 below.

$$\log\left(\widehat{SalePrice}\right)_{Combined} = 11.23 + 0.0004438 \times GrLivArea \qquad \textbf{Eq. 1}$$

In North Ames, for a 100 square foot increase in the above ground living area, the median sale price increases by 3.30%, with a 95% confidence that the increase is between 2.77% and 3.80%. It is estimated that for an above ground living area of zero, the predicted sale price is extrapolated to be $e^{11.44}$, or $92,709.01. The model can be interpreted with equation 2 below.

$$\log(\widehat{SalePrice})_{NAmes} = 11.44 + 0.0003241 \times GrLivArea \qquad \textbf{Eq. 2}$$

In Edwards, for a 100 square foot increase in the above ground living area, the median sale price increases by 5.53% with a 95% confidence that the increase is between 4.09% and 6.99%, described in table 4. It is estimated that for an above ground living area of zero, the predicted sale price is extrapolated to be $e^{11.03}$, or $61,697.58. The model can be seen in equation 3 below.

$$\log(\widehat{SalePrice})_{Edwards} = 11.03 + 0.0005387 \times GrLivArea \qquad \textbf{Eq. 3}$$

In Brookside, for a 100 square foot increase in the above ground living area, the median sale price increases by 7.66% with a 95% confidence that the increase is between 6.22% and 9.13%, described in table 5. It is estimated that for an above ground living area of zero, the predicted sale price is extrapolated to be $e^{10.79}$, or $48,533.04. The model can be seen in equation 4 below.

$$\log(\widehat{SalePrice})_{Brookside} = 10.79 + 0.0007382 \times GrLivArea \qquad \textbf{Eq. 4}$$

Conclusion: This analysis gives sufficient evidence to support that the relationship between sale price and above ground square footage depends on the neighborhood, with all significant parameters determined by their p-values < 0.05. Houses in Brookside exhibit the highest increase in sale price per square foot compared to North Ames and Edwards, reflecting typical market dynamics. This study should allow Century 21 Ames to understand how above ground square footage impacts pricing differently across neighborhoods, aiding in strategic decisions and future market predictions.

RShiny App: Price vs. Living Area Chart

To view scatterplots of the home's sale price against the above ground square footage for North Ames, Edwards, Brookside, and other neighborhoods through an interactive graphical user interface, visit the following link: https://knguyen-ds.shinyapps.io/HousingAmesIA/

Analysis Question 2

Problem: The next task was to build the most predictive model for forecasting the sales prices of homes in Ames, Iowa. Multiple regression models were developed and compared with varying factors considered and selected based on their correlation to SalePrice. To evaluate the models, certain metrics including the adjusted R^2, internal CV PRESS, and Kaggle Score (found with the Root-Mean-Squared-Error (RMSE) between the logarithm of the predicted value and the logarithm of the observed sales price) were evaluated for each model. After evaluating the models, the goal was to analyze and identify which model is best to predict future home sale prices in Ames, Iowa.

Candidate Models: Separate linear regression models were developed to predict sale price, with LogSalePrice as the dependent variable and various combinations of the other factors as the independent variables.

The training and testing datasets were cleaned and transformed by combining GrLivArea and TotalBsmtSF to represent the total square footage of a house, established as a new variable, AllSQFT. Then, some missing values, NA, were replaced with "None" or the statistical mean of their respective columns, depending on their data descriptions. The non-numeric ordinal variables, such as ExterQual, ExterCond, etc., were changed to a numeric scale. The explanatory variables with remaining missing values, such as Electrical, were then removed from the analysis. Lastly, the explanatory variables were ranked based on their correlation with SalePrice. The top variables were OverallQual, AllSQFT, GarageCars, ExterQual, GarageArea, and KitchenQual.

Simple Linear Regression with 1 Explanatory Variable: The initial model generated was a simple linear regression model with OverallQual as the explanatory variable due to its high correlation with SalePrice. With this model, the sale price may be modeled as below in equation 5:

$$\log\left(\widehat{SalePrice}\right) = 10.5836 + 0.2357 \times OverallQual \qquad \textbf{Eq. 5}$$

For each increase in OverallQual rank, indicating an improvement in quality, the estimated median sale price will increase by $e^{0.2357}$, or 26.58%, with a 95% confidence that the increase falls between 25.00% and 28.16%. For an OverallQual of 0, the predicted sale price is extrapolated to be $e^{10.58}$, or approximately $39,340.

Multiple Linear Regression with 2 Explanatory Variables: The multiple linear regression model utilized GrLivArea and FullBath as the explanatory variables. With this model, the predicted sale price can be found using equation 6:

$$\log\left(\widehat{SalePrice}\right) = 10.8200 + 0.000647 \times GrLivArea + 0.3227 \times FullBath - 0.0001105 \times (GrLivArea \times FullBath) \qquad \textbf{Eq. 6}$$

For a 100 square foot increase in the above-ground living area, the estimated median sale price increases by approximately 6.48%, with a 95% confidence interval for the increase

between 5.79% and 7.17%. For each additional full bathroom, the median sale price increases by approximately 38.04%, with a 95% confidence interval for the increase between 29.24% and 46.84%. The interaction term between GrLivArea and FullBath has a negative correlation, indicating that at higher levels of living area, the incremental value of additional full bathrooms diminishes. When GrLivArea and FullBath are both zero, the extrapolated predicted sale price is $e^{10.8200}$, or approximately $50,067.

Multiple Linear Regression with 6 Explanatory Variables: The more complex multiple linear regression model used the following explanatory variables: OverallQual, AllSQFT, GarageCars, KitchenQual, ExterQual, and 1stFlrSF, due to their high correlation with SalePrice. With this model, the sale price may be predicted using equation 7:

$$\begin{aligned}\log(\widehat{SalePrice}) = {} & 17.970 - 1.744 \times OverallQual - 0.002322 \times AllSQFT \\ & - 1.466 \times GarageCars - 3.049 \times KitchenQual \\ & - 3.676 \times ExterQual - 0.01375 \times X1stFlrSF \\ & + 0.0004815 \times (OverallQual \times AllSQFT) \\ & + 0.6192 \times (OverallQual \times GarageCars) \\ & + 0.0007574 \times (AllSQFT \times GarageCars) \\ & + 0.6105 \times (OverallQual \times KitchenQual) \\ & + 2.545 \times (GarageCars \times KitchenQual) \\ & + 0.8289 \times (OverallQual \times ExterQual)\end{aligned} \qquad \textbf{Eq. 7}$$

The estimate of OverallQual as a standalone variable is -1.744, suggesting that, in isolation, a value increase in OverallQual would lead to a decrease in SalePrice by $e^{-1.744} - 1$, or about 81.09%. However, this variable interacts with other variables which affect its true effect on SalePrice. For instance with the interaction between OverallQual and AllSQFT, for a 100 square foot increase, the effect of OverallQual on sale price increases by $e^{0.0007574} - 1$, or 4.92%. This means that the more square footage a home has, the more significant OverallQual becomes in determining its price. The same process can be done with all the significant interaction terms to determine the effect percentages. For AllSQFT as a standalone variable, for a 100 square foot increase in total square footage, the median sale price decreases by 20.61%, with a 95% confidence interval for the decrease between 19.51% and 21.71%. For GarageCars, each additional car capacity in the garage decreases the median sale price by 76.96%, with a 95% confidence interval for the decrease between 63.72% and 90.50%.For KitchenQual, for each rank increase in kitchen quality, the median sale price decreases 95.41%, with a 95% confidence interval for the decrease between 85.98% and 105.08%.For each rank increase in ExterQual, the median sale price decreases by 97.44%, with a 95% confidence interval for the decrease between 88.67% and 106.23%. For a 100 square foot increase in 1stFlrSF, the median sale price decreases by 1.36%, with a 95% confidence interval for the decrease between 0.75% and 1.97%. When OverallQual, AllSQFT, GarageCars, KitchenQual, ExterQual, and 1stFlrSF are all zero, the predicted sale price is extrapolated to approximately $58,965,232.

Assumptions: The resulting plots (Appendix B-1) generated confirm the models' assumptions are supported. Random scattering in the residual plots confirm homoscedasticity and independence. The residuals in the Q-Q plots lie closely to the diagonal line, validating the normality assumption. Additionally, histograms of residuals (Appendix B-2) align with a normal distribution curve, further supporting this conclusion. Outlier analysis involved inspecting leverage and influence measures, such as Cook's Distance. There do not appear to be any more extremely influential points that should be removed.

Comparing Competing Models: Table 2 compares the performance of the three models for predicting home sale prices in Ames, Iowa, using various evaluation metrics: Adjusted R^2, Internal CV PRESS, Kaggle Score (RMSE), and AIC. The Simple Linear Regression model, with only OverallQual as a predictor, has an adjusted R^2 of 0.6686, indicating that about 67% of the variability in sale prices is explained by this single variable. However, it shows a high CV PRESS and high Kaggle Score (RMSE), suggesting that while it performs reasonably, it does not generalize as effectively as the other models. The Multiple Linear Regression (GrLivArea, Full Bath) model, which incorporates GrLivArea and FullBath, has a lower adjusted R^2 of 0.5432, indicating that it explains only 54% of the variability. This model also performs worse in terms of CV PRESS and RMSE, highlighting its reduced predictive power compared to the first model. In contrast, the Multiple Linear Regression (OverallQual, AllSQFT, GarageCars, KitchenQual, ExterQual, 1stFlrSF) model demonstrates the best performance across all metrics, with an adjusted R^2 of 0.8494, explaining 85% of the variability in sale prices. It has the lowest CV PRESS, the best Kaggle Score (RMSE) of 0.16431, and the lowest AIC of -1278.4, indicating it is the most efficient and accurate model for predicting home prices. Thus, the third model, incorporating multiple predictors and interaction terms, provides the most robust and reliable predictions.

Table 2. *Comparing Model Statistics*

Models	Adjusted R^2	Internal CV PRESS	Kaggle Score (RMSE)	AIC
Simple Linear Regression	0.6686	2521860547973.95	0.22896	-202.82
Multiple Linear Regression (GrLivArea, Full Bath)	0.5432	3890339503310.41	0.28750	261.68
Multiple Linear Regression (OverallQual, AllSQFT, GarageCars, KitchenQual, ExterQual, 1stFlrSF)	0.8494	931032752785.55	0.16431	-1278.4

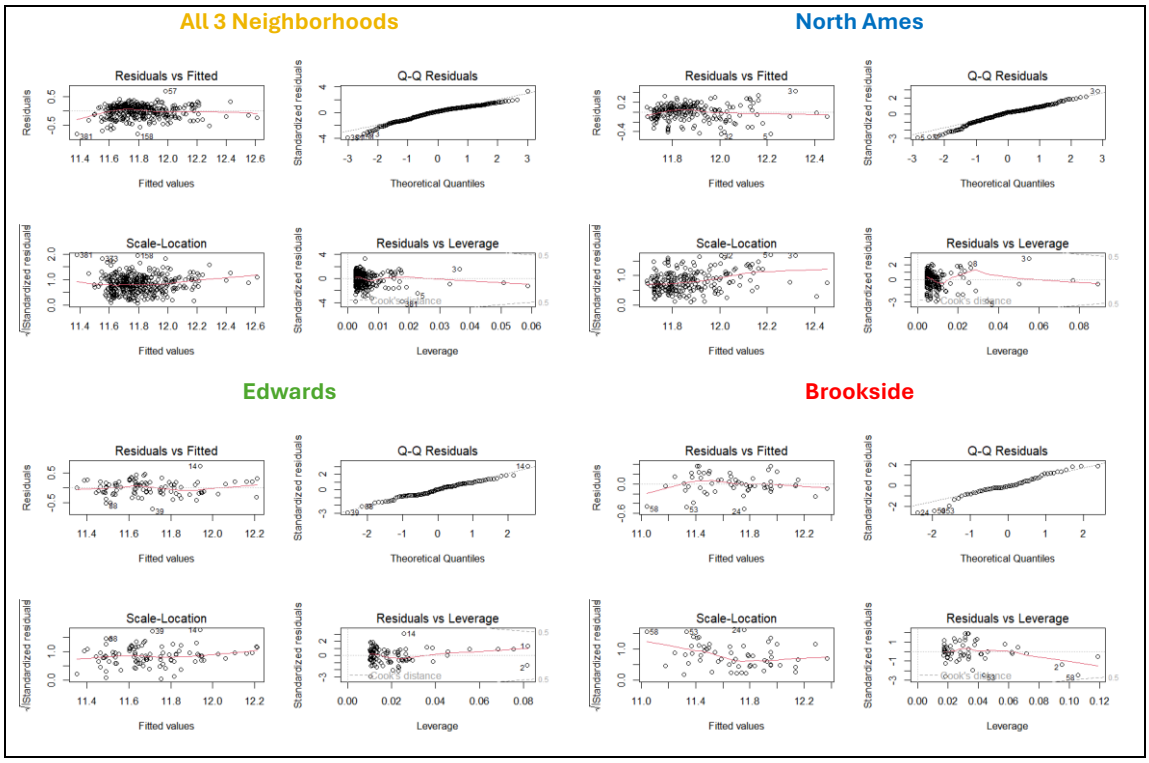
Conclusion: In conclusion, to predict home sale prices in Ames, Iowa, multiple regression models were developed with varying complexity to identify the most predictive model. The first model, a simple linear regression using OverallQual as the sole predictor, explained 67% of the variance in sale prices but had high CV PRESS and RMSE values, indicating limited generalizability. The second model, a multiple linear regression with GrLivArea and FullBath as predictors, explained 54% of the variance and performed worse in terms of CV PRESS and RMSE, showing reduced predictive power. The third, more complex model, incorporated six key predictors—OverallQual, AllSQFT (a combination of GrLivArea and TotalBsmtSF), GarageCars, KitchenQual, ExterQual, and 1stFlrSF—along with interaction terms to capture the combined effects of variables. This model demonstrated the best performance, explaining 85% of the variability in sale prices, and it had the lowest PRESS, best Kaggle Score (RMSE = 0.16431), and lowest AIC, indicating superior predictive accuracy. The model's assumptions were validated through residual analysis, and no influential outliers were identified. Ultimately, the third model provided the most reliable and efficient predictions for future home sale prices in Ames, Iowa, capturing both main effects and interactions between variables.

Appendix

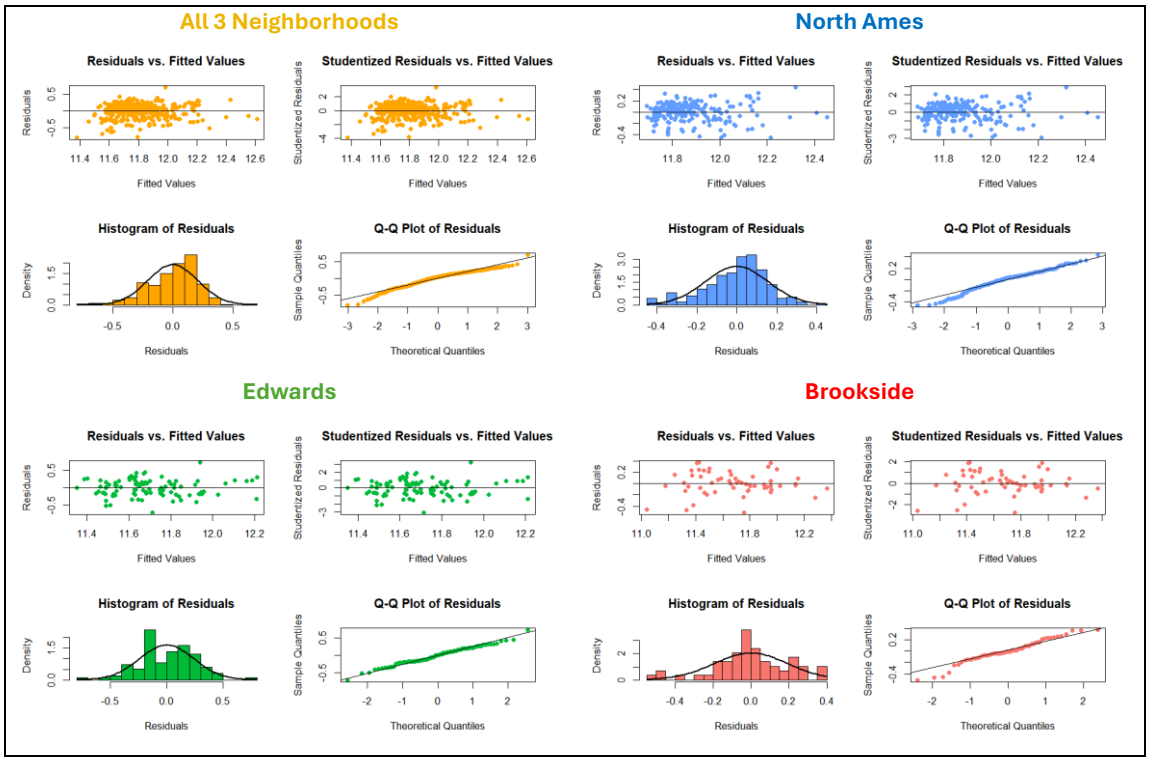
A. Analysis Question 1

Full script/code can be found by following this link: https://github.com/knguyen-ds/AmesHousingProj/blob/main/AnalysisQuestion1/Pt1Code.docx

1. Plots to provide evidence of the appropriateness of the model and verify assumptions



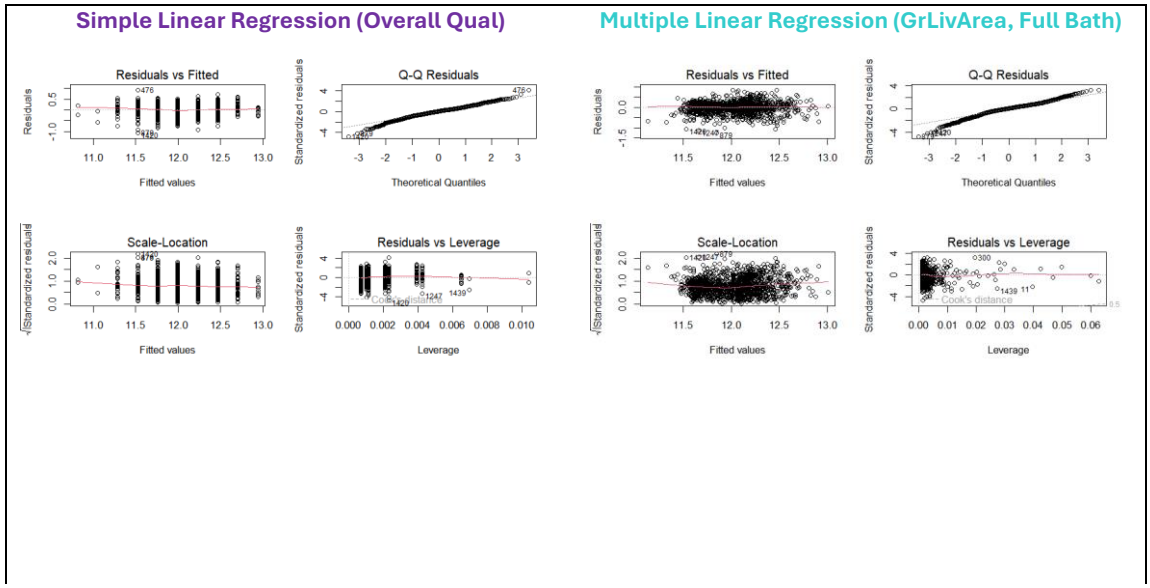
2. Additional plots to verify model assumptions



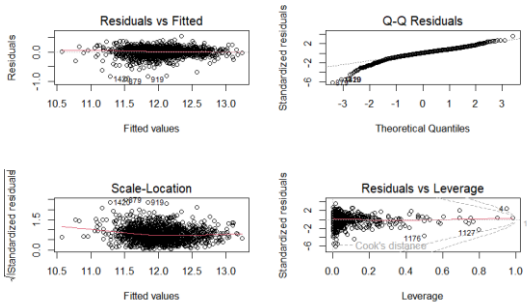
B. Analysis Question 2

Full script/code can be found by following this link: https://github.com/knguyen-ds/AmesHousingProj/blob/main/AnalysisQuestion2/Pt2Code.docx

1. Plots to provide evidence of the appropriateness of the model and verify assumptions

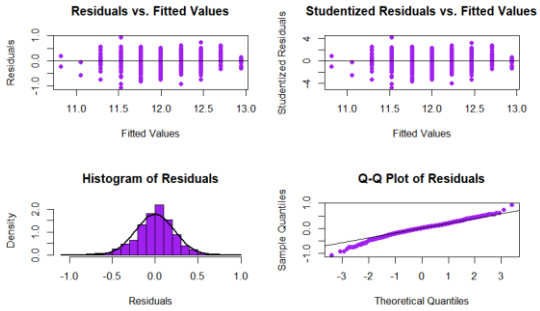


Multiple Linear Regression
(OverallQual, AllSQFT, GarageCars, KitchenQual, ExterQual, 1stFlrSF)

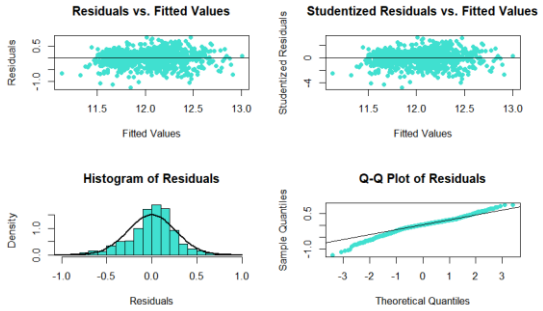


2. Additional plots to verify model assumptions

Simple Linear Regression (Overall Qual)



Multiple Linear Regression (GrLivArea, Full Bath)



Multiple Linear Regression
(OverallQual, AllSQFT, GarageCars, KitchenQual, ExterQual, 1stFlrSF)