Linear Regression Analysis of Ames Housing Dataset
Kristin Henderson
Summer 2024

Introduction

In this analysis, I aim to address two questions regarding home sale prices in Ames, Iowa using various regression techniques. The first question focuses on understanding the relationship between the sale price and the square footage of the living area in three specific neighborhoods. This estimate is intended to provide valuable market-specific insights for Century 21 Ames. The second question seeks to build a predictive model for home sale prices across all of Ames, identifying which combination of home characteristics produces the most robust model and best predictions.

Data Description

The data used in these analyses are from the Ames Housing Dataset (De Cock, 2011). This dataset was obtained from the "House Prices - Advanced Regression Techniques" competition on Kaggle (Montoya, 2016). The training set has 1460 observations with 81 variables, and the testing set has 1459 observations with 80 variables, missing the 'SalePrice' response variable.

Variables of importance for these analyses:

- SalePrice: The property's sale price in dollars, the target variable.
- BsmtFinSF1: Type 1 finished square feet.
- GarageCars: Size of garage in car capacity.
- GrLivArea: Above ground living area square feet.
- LotArea: Lot size in square feet.
- MSSubClass: The building class.
- Neighborhood: Physical locations within Ames city limits.
- OverallCond: Overall condition rating.
- OverallQual: Overall material and finish quality.
- YearBuilt: Original construction date.

Analysis Question 1

Problem

I aim to use linear regression to estimate the relationship between home sale price and the living area in square feet within the North Ames, Edwards and Brookside neighborhoods. I would also like to determine if this relationship depends on the neighborhood.

Build and Fit the Model

I will start by examining the most simple model, $\mu\{\text{Price} \mid \text{Area}\} = \beta_0 + \beta_1\text{Area}$. I will also include neighborhood as an indicator variable, $\mu\{\text{Price} \mid \text{Area, Neighborhood}\} = \beta_0 + \beta_1\text{Area} + \beta_2\text{Brookside} + \beta_3\text{Edwards}$, and examine interaction terms with Area and Neighborhood, $\mu\{\text{Price} \mid \text{Area, Neighborhood}\} = \beta_0 + \beta_1\text{Area} + \beta_2\text{Brookside} + \beta_3\text{Edwards} + \beta_4\text{Area*Brookside} + \beta_5\text{Area*Edwards}$. Additionally, I will explore if any transformations of the data may be appropriate.

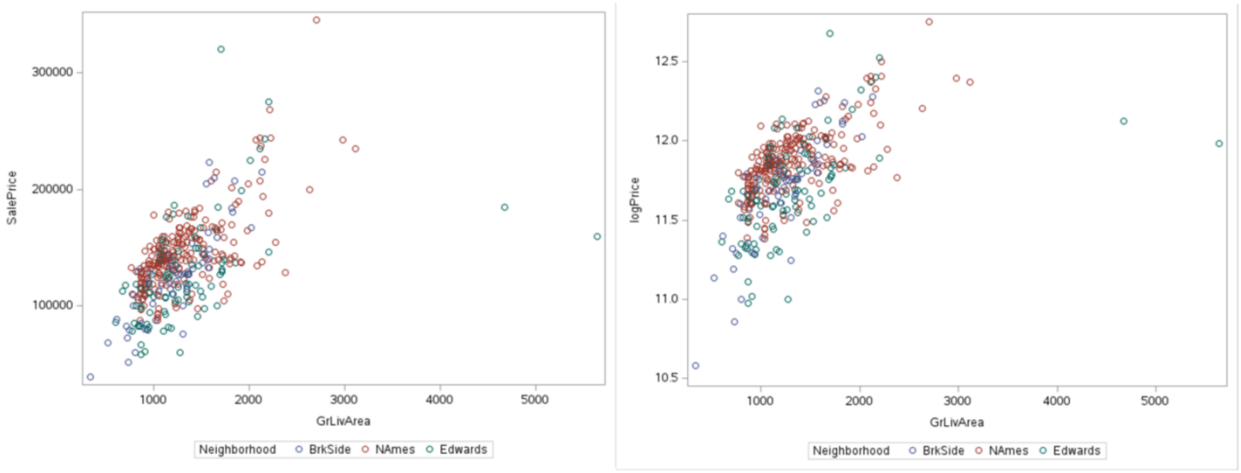


Figure 1: Scatterplot of sale price in dollars and above ground living area in square feet with neighborhoods plotted by color (blue: Brookside, red: North Ames, green: Edwards). Original, untransformed data are on the left, and data with log transformation of sale price are on the right.

After applying log transformations to both the living area explanatory and the sale price response variables and comparing plots of the untransformed data and the log transformed data as in Figure 1, the log transformations seems to provide a better linear fit. I will proceed with candidate models: one with log-transformed price and the untransformed area (log-lin), and the other with log transformations of both price and area (log-log). A comparison of these scatterplots is in Supplementary Figure 1 in the appendix.

Checking Assumptions

Residual Plots

The residual plots of the log-lin transformed data (Figure 2) satisfy the assumptions of linear regression better than that of the untransformed data. The residuals of the log-log transformed data in Supplementary Figure 2 also look quite good.

Influential point analysis (Cook's D and Leverage)

One house in the dataset has a large Cook's D (Figure 3). This house has a large square footage but a low sale price, likely due to its partial sale condition. I found four partially completed houses within the neighborhoods of interest, with at least two being influential points. Those two also are the only houses with square footages greater than 3500. Although I could restrict the data range and exclude partially completed houses above 3500 square feet, I decided to exclude all partially completed houses from this analysis, because there are only four. If analyzing all neighborhoods, this strategy might not be appropriate. In the full dataset, there are many more houses listed in partial condition as there are houses with large square footages.

After restricting the analysis and removing these data, the leverage plots look much better, with little evidence of additional influential points, except for one very small and low-priced house.

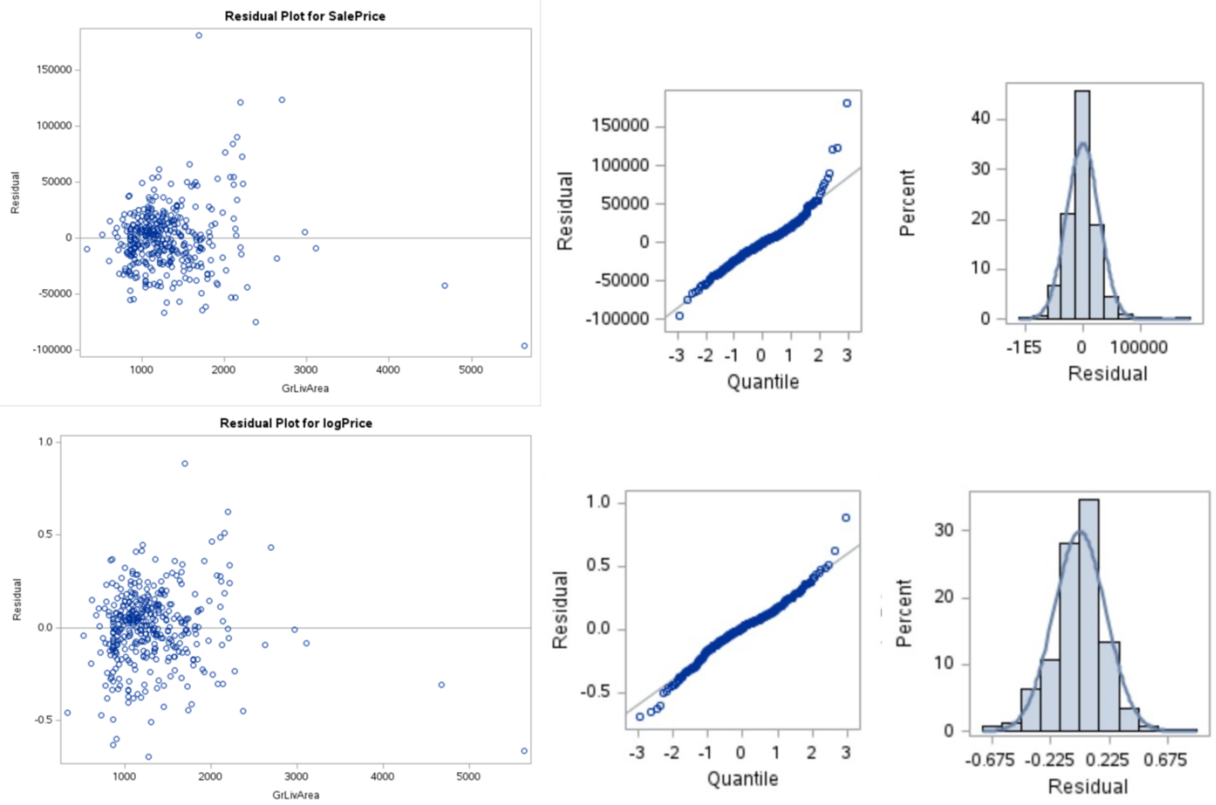


Figure 2: From left to right, residual plots, QQ plots, and histograms of untransformed data (top) and data with log transformation of sale price (bottom).

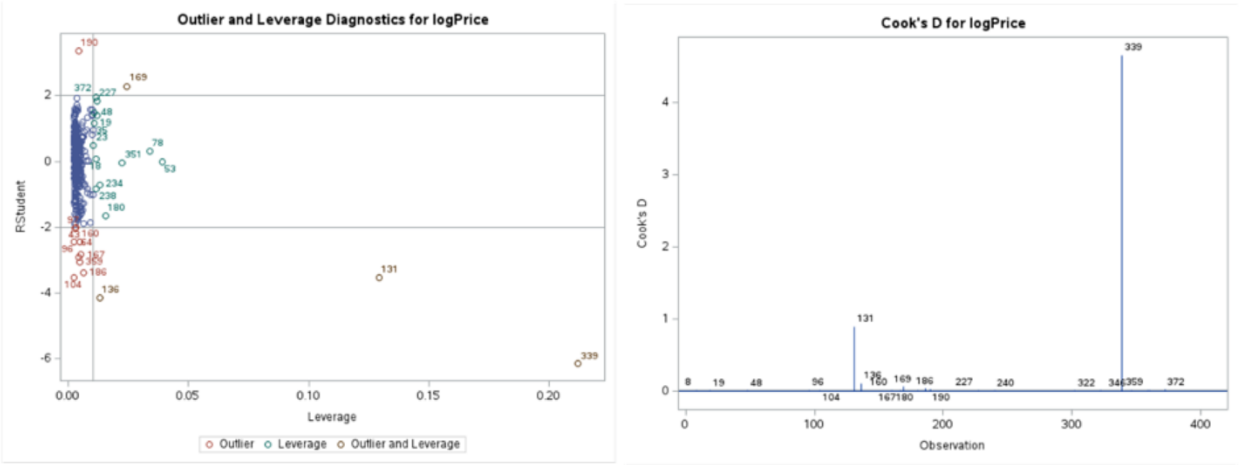


Figure 3: Plots of studentized residuals and leverage (left) and Cook's D (right) for the log-lin transformed data before removing partially completed houses from the analysis.

Address Assumptions

1. Linearity: Examining the scatterplots of the original and log transformed price data in Figure 1, the log-lin transformed data appear to provide a better linear fit.

2. Equal Standard Deviation: The residual plot of the untransformed data in Figure 2 shows some evidence of heteroscedasticity. The log-lin transformation reduces this, showing more random scattering with no evidence against equal standard deviation.
3. Normality: Both the original and log transformed histograms in Figure 2 are consistent with normal distributions. The QQ plot and random scattering of residuals in the log-lin transformed data are more consistent with a normal distribution than those of the untransformed data.
4. Independence: I have no evidence against independence.

Comparing Competing Models: Adj R^2**,** Internal CV PRESS**,** AIC

Effects:	Intercept GrLivArea Neighborhood GrLivArea*Neighborho

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	5	15.12432	3.02486	85.00
Error	373	13.27352	0.03559	
Corrected Total	378	28.39783		

Root MSE	0.18864
Dependent Mean	11.79698
R-Square	0.5326
Adj R-Sq	0.5263
AIC	-877.31911
AICC	-877.01722
PRESS	13.84633
SBC	-1234.69389

Effects:	Intercept logArea Neighborhood logArea*Neighborhood

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	5	15.12337	3.02467	84.99
Error	373	13.27446	0.03559	
Corrected Total	378	28.39783		

Root MSE	0.18865
Dependent Mean	11.79698
R-Square	0.5326
Adj R-Sq	0.5263
AIC	-877.29202
AICC	-876.99013
PRESS	13.80478
SBC	-1234.66680

Effects:	Intercept logArea Neighborhood

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	3	14.43888	4.81296	129.30
Error	375	13.95895	0.03722	
Corrected Total	378	28.39783		

Root MSE	0.19293
Dependent Mean	11.79698
R-Square	0.5085
Adj R-Sq	0.5045
AIC	-862.23634
AICC	-862.07548
PRESS	14.32273
SBC	-1227.48620

Figure 4: Comparison of statistics of competing models. Left: interaction variables and independent slopes when price is log transformed. Center: or when both price and area are log transformed; Right: indicator variables with the same slope when both price and area are log transformed.

All the statistics in Figure 4 indicate that models using log transformation with independent slopes for each neighborhood perform the best among the candidate models. This is consistent when comparing additional models not shown, which did not account for the neighborhood or include data transformation. The model with log-transformed price but untransformed area performed slightly better than when both variables were log-transformed. I opted for this model due to its easier interpretation. The AIC, BIC and internal CV PRESS are the lowest for this model and are lower than the models with log-log transformations with or without interaction variables. The adjusted R^2 is the same for the equivalent model with log-log transformations and is higher than those without interaction variables.

Parameters

Estimates

General model:

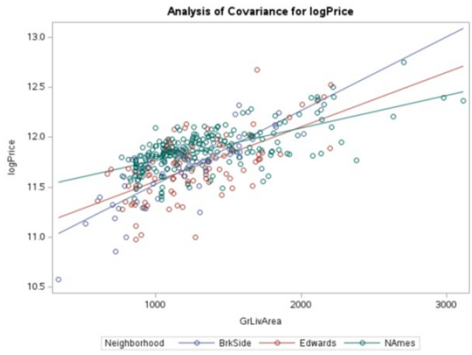
$\mu\{\log(\text{Price}) \mid \text{Area, Neighborhood}\} = 11.4433 + 0.0003*\text{Area} - 0.6517*\text{Brookside} - 0.4334*\text{Edwards} + 0.0004(\text{Area}*\text{Brookside}) + 0.0002(\text{Area}*\text{Edwards})$

Neighborhood specific models:

$\mu\{\log(\text{Price}) \mid \text{Area, Neighborhood = North Ames}\} = 11.0099 + 0.0003*\text{Area}$

$\mu\{\log(\text{Price}) \mid \text{Area, Neighborhood = Edwards}\} = 10.7916 + 0.0005*\text{Area}$

$\mu\{\log(\text{Price}) \mid \text{Area, Neighborhood = Brookside}\} = 11.4433 + 0.0007*\text{Area}$



Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	11.44334070	B	0.04187381	273.28	<.0001	11.36100237	11.52567903
GrLivArea	0.00032412	B	0.00003048	10.63	<.0001	0.00026419	0.00038406
Neighborhood BrkSide	-0.65174673	B	0.09171915	-7.11	<.0001	-0.83209815	-0.47139531
Neighborhood Edwards	-0.43335031	B	0.08240274	-5.26	<.0001	-0.59538246	-0.27131815
Neighborhood NAmes	0.00000000	B	.	.	.	.	.
GrLivArea*Neighborho BrkSide	0.00041410	B	0.00007146	5.80	<.0001	0.00027359	0.00055461
GrLivArea*Neighborho Edwards	0.00022211	B	0.00006199	3.58	0.0004	0.00010022	0.00034401
GrLivArea*Neighborho NAmes	0.00000000	B	.	.	.	.	.

Figure 5: Regression lines and parameter estimate table of chosen model: $\mu\{\log(\text{Price}) \mid \text{Area}, \text{Neighborhood}\} = \beta_0 + \beta_1\text{Area} + \beta_2\text{Brookside} + \beta_3\text{Edwards} + \beta_4\text{Area*Brookside} + \beta_5\text{Area*Edwards}$.

Interpretation

Slopes: There is convincing evidence at the $\alpha = 0.05$ confidence level that for every 100 square foot increase in above ground living area, the mean expected sale price increases in each of these neighborhoods. There is overwhelming evidence that in North Ames, with each additional 100 square feet of living area, the mean expected sale price increases $e^{(0.00032*100)} = 1.0325$ times or by 3.25% (p-value < 0.0001). There is strong evidence that in Edwards, the increase is $e^{(0.00054*100)} = 1.0555$ times or by 5.55% (p-value = 0.0004). There is overwhelming evidence that in Brookside the increase is $e^{(0.00073*100)} = 1.0757$ times or by 7.57% (p-value < 0.0001).

Intercepts: There is overwhelming evidence that for houses with zero square feet of above-ground living area, the mean expected sale price in North Ames is $e^{11.4433} = \$93{,}274$ (p-value < 0.0001. In Edwards, the mean expected sale price is 35.2% ($1 - e^{-0.4334} = 1 - 0.648$) less than homes in North Ames or $e^{11.0099} = \$60{,}470$. In Brookside, the mean expected sale price is 47.9% ($1 - e^{-0.6517} = 1 - 0.521$) less than homes in North Ames or $e^{10.7916} = \$48{,}611$ (p-value < 0.0001. As houses typically have more than zero square feet of living area, this estimate should be interpreted cautiously and in the context of other predictors.

Confidence Intervals

Slopes: I am 95% confident at the $\alpha = 0.05$ confidence level that for every 100 square foot increase in above-ground living area, the mean expected sale price in North Ames increases by between 2.63% and 3.87%, in Edwards by between 4.29% and 6.82%, and in Brookside by between 6.07% and 9.09%.

Intercepts: I am 95% confident that for houses with zero square feet of living area, the mean expected sale price in North Ames is between $85,905 and $101,286. In Edwards, it is between $51,426 and $71,111 (22.8% to 44.9% lower than North Ames), and in Brookside, it is between $40,587 and $58,215 (37.6% to 56.5% lower than North Ames). Again, interpreting the sale price for houses with zero square feet of living area should be done cautiously.

Conclusion

In the preceding analysis, I examined the relationship between home sale price and living area within the North Ames, Edwards, and Brookside neighborhoods. The results indicated that this relationship varies by neighborhood. For every 100 square foot increase in above-ground living area, the mean expected sale price increases by 3.25% in North Ames, 5.55% in Edwards, and

7.57% in Brookside. Influential observations were addressed, and model assumptions were verified. A multiple linear regression model with a log transformation of the sale price and interaction terms provided the best fit, enabling generation of estimates and confidence intervals for the relationship between living area and sale price in each neighborhood within the Century 21 Ames market.

R Shiny: Price vs Living Area Chart

https://kdhenderson.shinyapps.io/RegressionOfHousingPricesOnSquareFootage/

Analysis Question 2

Problem

I would like to use linear regression to build a model to predict sale price in all of Ames, Iowa.

Candidate Models

- SLR: $\mu\{\log(\text{Price}) \mid \text{OverallQual}\} = \beta_0 + \beta_1\text{OverallQual}$
- MLR 1: $\mu\{\log(\text{Price}) \mid \text{OverallQual}, \log(\text{GrLivArea}), \text{Neighborhood}, \text{MSSubClass}\} = \beta_0 + \beta_1\text{OverallQual} + \beta_2\log(\text{GrLivArea}) + \beta_*\text{Neighborhood} + \beta_*\text{MSSubClass}$ (* where Neighborhood has 25 categories and MSSubClass has 16 categories)
- MLR 2: $\mu\{\log(\text{Price}) \mid \text{OverallQual}, \log(\text{GrLivArea}), \text{OverallCond}, \text{GarageCars}, \text{BsmtFinSF1}, \text{YearBuilt}, \log(\text{LotArea}), \text{Neighborhood}\} = \beta_0 + \beta_1\text{OverallQual} + \beta_2\log(\text{GrLivArea}) + \beta_3\text{OverallCond} + \beta_4\text{GarageCars} + \beta_5\text{BsmtFinSF1} + \beta_6\text{YearBuilt} + \beta_7\log(\text{LotArea}) + \beta_*\text{Neighborhood}$ (* where Neighborhood has 25 categories)
- MLR 4: $\mu\{\log(\text{Price}) \mid \text{OverallQual}, \log(\text{GrLivArea}), \text{OverallCond}, \text{GarageCars}, \text{BsmtFinSF1}, \text{YearBuilt}, \log(\text{LotArea}), \text{Neighborhood}, \text{MSSubClass}\} = \beta_0 + \beta_1\text{OverallQual} + \beta_2\log(\text{GrLivArea}) + \beta_3\text{OverallCond} + \beta_4\text{GarageCars} + \beta_5\text{BsmtFinSF1} + \beta_6\text{YearBuilt} + \beta_7\log(\text{LotArea}) + \beta_*\text{Neighborhood} + \beta_*\text{MSSubClass}$ (* where Neighborhood has 25 categories and MSSubClass has 16 categories)

Checking Assumptions

Influential point analysis (Cook's D and Leverage)

The studentized residuals and leverage and Cook's D plots for the SLR and MLR1 models (left half of Figure 6) do not show evidence of any influential points. However, MLR2 (center right of Figure 6) has points of concern. There was one observation of basement square footage (BsmtFinSF1) above 5000. I decided to restrict the range of the analysis to houses below this threshold. The residuals for MLR4 benefitted from removing this point. The remaining influential point seemed to result from a basement with over 2000 square feet. I considered 4000 as a cutoff, as only two houses fell above 4000. Additionally, only ten houses in the whole dataset were above 2000. However, the house above 4000 and several of the houses above 2000 were in the test set, and I did not want to eliminate them.

Address Assumptions

1. Linearity: Log transformation of SalePrice, GrLivArea and LotArea resulted in a better linear fit than the untransformed data (Supplementary figure 3).
2. Equal Standard deviation: The residual plots in Figure 6 are randomly scattered with no strong evidence against equal standard deviation.

3. Normality: Prior to log transformation the distributions looked very right-skewed. After log transformation of the above variables, the histograms, QQ plots and random scattering of residuals in Figure 6 are reasonably consistent with a normal distribution. There is some mild left skewness, but I feel comfortable that the sample size will mitigate this.
4. Independence: There is no evidence against independence.

Residual Plots

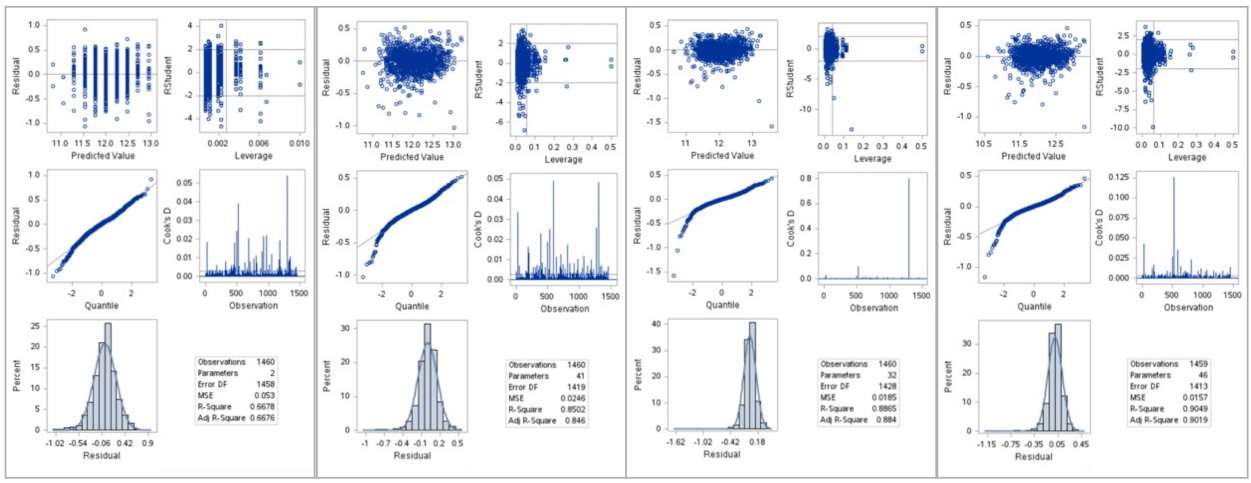


Figure 6: Residual plots for candidate models SLR, MLR1, MLR2, and MLR4 from left to right after log transformation of selected variables. Residuals for MLR4 were plotted after restricting analysis to basement square footages under 5000.

Comparing Competing Models: Adj R^2**,** Internal CV PRESS, AIC, Kaggle Score

Predictive Models	Adjusted R^2	CV PRESS	AIC	Kaggle Score
Simple Linear Regression	.67	76.7	-2838	.229
Multiple Linear Regression 1 (MLR1)	.85	36.5	-3952	.162
MLR2	.90	25.0	-4499	.139
MLR4	.90	24.1	-4558	.136

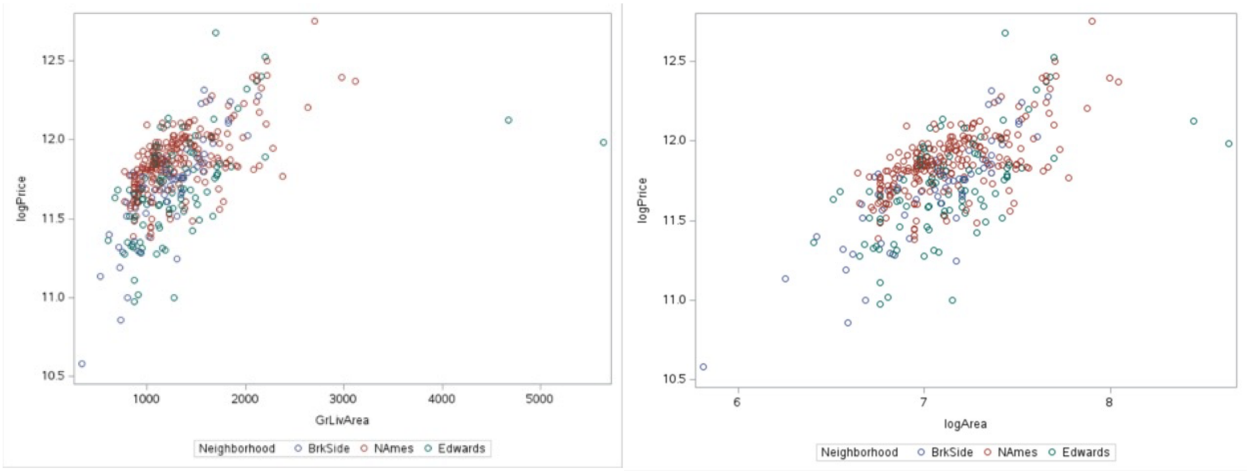
Conclusion

In Analysis 2, I developed several predictive models for home sale prices in Ames, Iowa. Among the candidate models, the Multiple Linear Regression Model 4 (MLR4) performed the best, with a high adjusted R^2 of 0.90, and the lowest CV PRESS, AIC and Kaggle scores. This model included nine predictor variables, two treated categorically and the rest numerically, and employed log transformations of several variables, including the response variable SalePrice. Influential observations and model assumptions were addressed. With one fewer variable (MSSubClass), MLR2 performed almost as well.

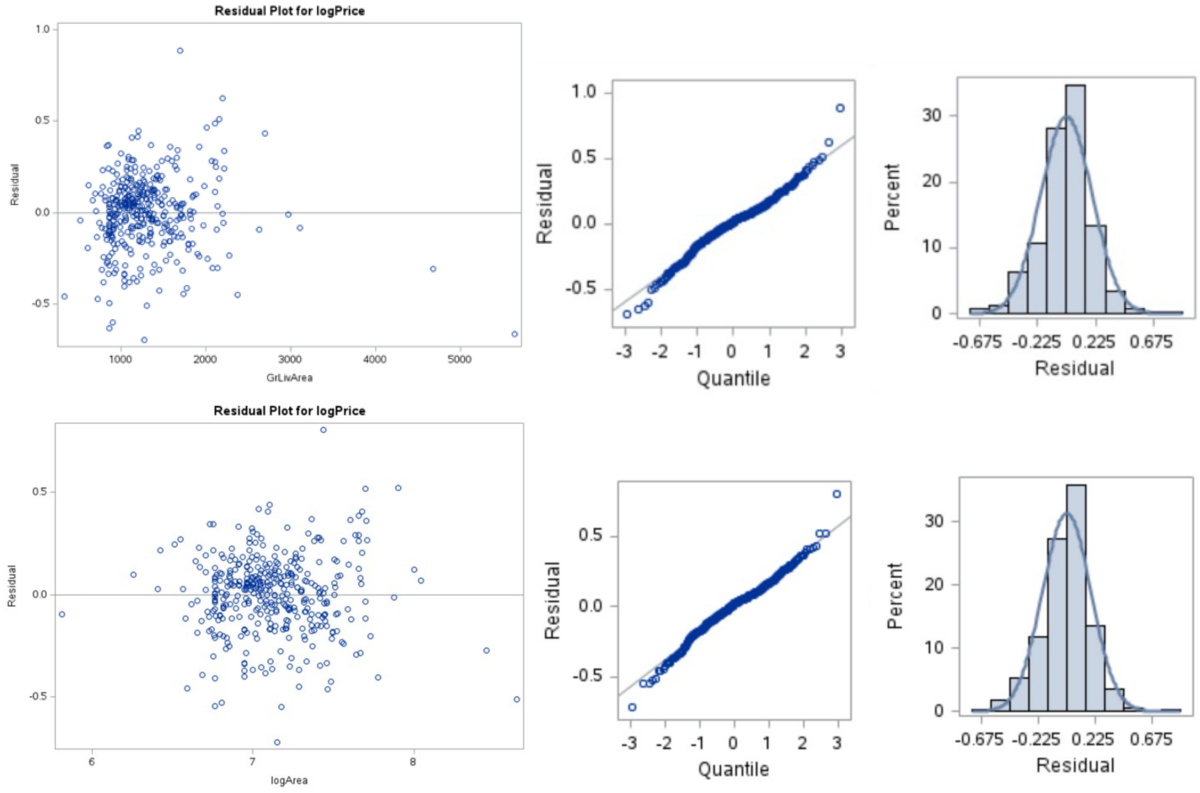
Future improvements could include treating ranked numeric variables as categorical, grouping categories, exploring log transformations on additional variables, and imputing with simpler model predictions rather than median values. Additionally, predicting home prices is highly dependent on the economy and mortgage interest rates, so incorporating these data would benefit future models. MLR4 performed very well and is the best of these candidate models to predict home sale prices in Ames.

Appendix

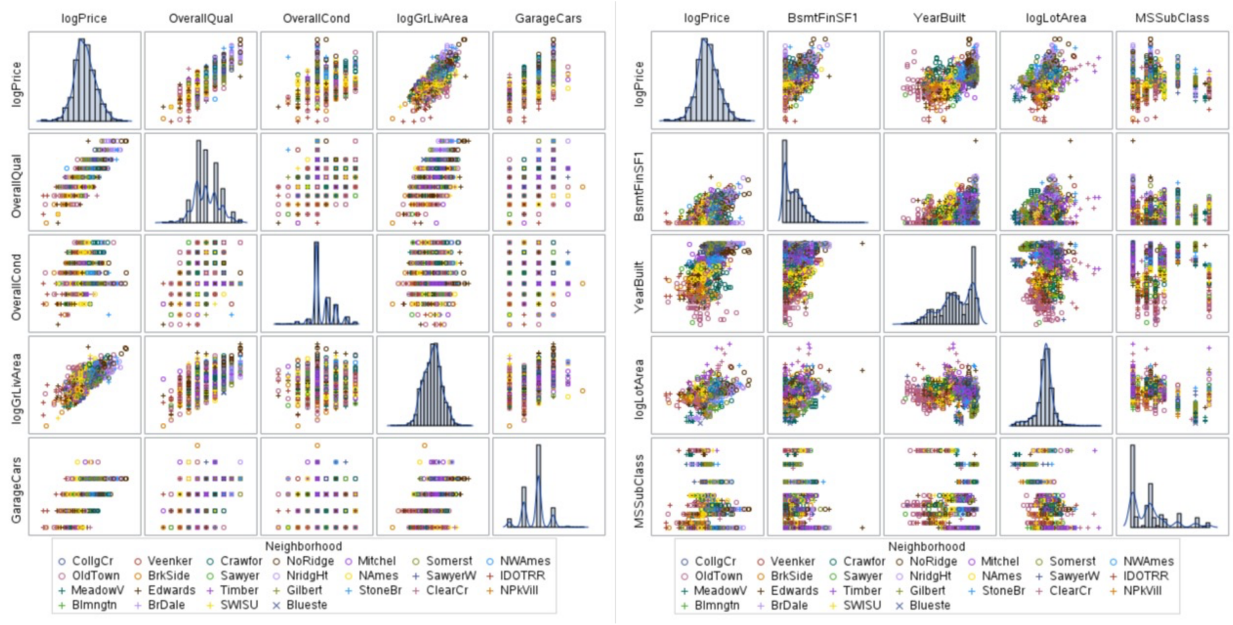
Supplementary Figures for Analysis 1



Supp. Fig. 1: Scatterplot of sale price in dollars and above ground living area in square feet with neighborhoods plotted by color (blue: Brookside, red: North Ames, green: Edwards). Data with log transformation of only sale price are on the left, and data with log transformed price and area are on the right.



Supp. Fig. 2: From left to right, residual plots, QQ plots, and histograms of data with log transformation of sale price (top) and log transformation of sale price and area (bottom).



Supp. Fig. 3: Scatterplots and histograms of variables, some after log transformation, used in candidate models with neighborhoods plotted by color/symbol.

Citations

De Cock, D. 2011. “Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project.” *Journal of Statistics Education* 19 (3).

Montoya A., DataCanary. 2016. House Prices - Advanced Regression Techniques. Kaggle. https://kaggle.com/competitions/house-prices-advanced-regression-techniques