Analysis of Home Sales and Influential Factors

Introduction

A dataset of houses sold in Ames, Iowa between 2006 and 2010 contains 34 quantifiable variables and 46 qualitative variables. Some of these variables can be used to estimate final sales prices, a useful prediction for Real Estate clients to know which housing factors to focus on when presenting houses to potential buyers. One such real estate company, Century 21 is specifically interested in 3 key neighborhoods and wants to measure the relationship between living area and sale price. After answering this question, a predictive model will be built to estimate sales prices for all the homes in the dataset.

Data Description

Dean De Cock is a statistician who was looking to find a better data set than the frequently used but outdated Boston Housing Data Set. While looking for a project for his regression course in graduate school, he heard from a student that Ames City Assessor's office was updating an assessment model that used a dataset of interest. After getting approval from the office and acquiring the dataset, he removed some extraneous variables leaving him with an appropriate dataset now known as, Ames Housing Dataset. The dataset contains 2920 individual residential property sales in Ames, Iowa, sold between 2006 and 2010. For the purposes of this analysis, the data was split into a training and a test set – both with 1460 observations each. The dataset includes 80 variables focusing on "quality and quantity of physical attributes of the properties." Specifically, there are 20 continuous variables, 14 discrete variables, and 46 categorical variables. Dean De Cock's discussion of this dataset is linked in the works cited section.

Analysis Question 1

Since Century 21 only sells houses in North Ames, Edwards, and Brookside, they would like to first get an estimate of how the sale price is related to the square footage of the living area of the house. Second, they would like to know if this relationship is different depending on which neighborhood the house is located in.

Build and Fit the Model

Our first step is to plot the data. We color coded the plot by neighborhood in order to get a good visual representation of the data. As seen in Figure 1.1 (see Appendix A for all Figures), there is visual evidence of an association between Living Area and Sales Price, and the three neighborhoods appear to have similar slopes. As we can see from the green dot outliers, though, the slope of the linear regression for the Edwards neighborhood could be influenced by high leverage outliers. Figure 1.2 estimates the regression slope and confidence intervals for each of the neighborhoods – and as we suspected, the slope for the Edwards neighborhood is different from the slope of the North Ames and Brookside neighborhoods.

Assumptions

As suggested by the visual evidence of the outliers, there may be assumption violations in the raw data of our dataset. We will address the assumptions of linearity, normality, constant variance, and independence.

Linearity

There is visual evidence that a straight line is an adequate model for the linear regression of Living Area and Sale Price. We will address the outliers in the Edwards neighborhood under the Constant Variance assumption discussion. Therefore, we will proceed as if the assumption of linearity is met.

Normality

As illustrated in the variable histograms in Figures 1.3 and 1.4, both the variables Living Area and Sale Price have right-tailed distributions. The QQ Plots, seen in Figures 1.5 and 1.6, for both variables also reflect violations of the normality assumptions with their slightly curved appearances.

One remedy to make the model robust against these violations is to transform the variables. The Living Area variable and the Sale Price variable are good candidates for a log transformation because, in addition to their skewness, they both have a large ratio of minimum and maximum values. The ratio for Living Area is 16.89 (Min: 334, Max: 5642) the ratio for Sale Price is 8.78 (Min: 39300, Max: 345000). The resulting histograms and QQ plots of the log transformed data for both variables can be found in Figures 1.7-1.10. Both the histograms and the QQ plots for the log transformed data show more normal distributions.

Constant Variance

Figures 1.11 and 1.12 show the residual plots for the raw vs the log transformed data. Log transformations successfully addressed violations of the assumptions for constant variance, as evidenced by the cloud of randomly scattered residuals in Figure 1.12. However, we found that there were still some high leverage residuals that were worth examining. As seen in Figure 1.13 and 1.14, observations 130, 131, and 339 have leverages of 0.05 or higher, and are close to the Cook's Distance line. We sorted the observations by Living Area, in descending order, and found that the two largest houses (with over 4000 square feet) had sale prices under $200,000. These were observations 131 and 339, two of our identified high leverage observations. These two mansions are either data invalid at the entry step (perhaps leaving out a zero), or these sales happened under very unusual circumstances, such as seller duress or financier's loss mitigation. In that case, we would not truly be measuring the association of sale price and square footage – we would instead be measuring the extreme effect of the confounding variable. We removed these two observations with confidence that these observations were invalid. Since we only removed 2 observations, the new residual plot looks very similar to Figure 1.12, but we included it as Figure 1.15 for curiosity's sake.

Independence

When discussing the housing market, where neighborhoods may or may not contain builder-grade houses, establishing true independence is a difficult task. Builder-grade houses may violate the independence assumption, as several blocks could be the same 5 models of a home. Also, the same real estate agent could have sold multiple houses in each neighborhood. However, given the volatility of the housing market (especially since our data covers the time period in which the housing market crash of 2008 occurred), two houses next to each other, sold by the same real estate agent, very well could be drastically different sales if one was sold in 2006 and the other sold in 2010. Using a time series model may give us more confidence in the model's robustness to independence assumption violations, but that is beyond the scope of this presentation. We will proceed as if the model is robust to independence assumption violations.

Comparing Competing Models

Adjusted R2

One alternative to the model we chose is to use non-transformed data. One could argue this would be a more practically useful model since the interpretation of log-transformed data is to reference a hypothetical doubling of the explanatory variable. Doubling the square footage of a home is not a real-world practice. That practical utility, however, requires accepting that your parameters may be misleading due to violations of the assumptions. In that sense, using the raw data would be less *statistically* useful than the transformed data. Keeping the possibility of misleading parameters in mind, Figure 1.16 shows a table of how fitting the raw data compares to fitting the transformed data. The table shows that the Adjusted R^2 for the Overall, Edwards, and North Ames estimates are higher for the raw data. This could arguably be a better fitting model – however, these values are still below 0.5. That means that less than 50% of the variance in the Sale Prices for these neighborhoods is accounted for by the variance in the square footage. There are likely other explanatory variables that we should be adding to the model to make it a better fit. On the other hand, for the neighborhood of Brookside, 70% of the variance in the Sales Price is accounted for by the variance in the Living Area Square Footage. That being said, we must also consider that only 58 of the 381 values in the dataset were from the Brookside neighborhood.

Another model to consider would be to fit the model holding Neighborhood constant. This is not a mutually exclusive model, since looking at each neighborhood individually is a step beyond holding the neighborhood constant. Figure 1.17 compares this alternative model, and we do find evidence that when we hold Neighborhood constant, the model using transformed data has a higher Adjusted R^2 overall than when using the raw data. The Adjusted R^2 for the individual neighborhoods does not change when holding Neighborhood constant, of course, since that is what they are already measuring.

Internal CV Press

Another measure of the appropriateness of the model's fit is to run an internal cross validation, using a Leave One Out method, to estimate how well the model predicts each individual observation of the dataset. The comparison of the R^2 scores for the model and its internal cross validation can be found in Figure 1.18 - the difference between the two is less than 1.5%.

Parameters

Estimate Interpretation

We found that the doubling of living area square footage is associated with an estimated $10^{.617}$ (4.14) increase in Sale Price. That is, a doubling of living area square footage is associated with an estimated 314% increase in Sale Price. We found evidence that this estimate varies by neighborhood: for Brookside, the estimated increase is by $10^{.82}$ (6.6), that is, an increase by 560%. For Edwards, the estimated increase is by $10^{.673}$ (4.7), that is, an increase by 370%. For North Ames, the estimated increase is by $10^{.473}$ (2.97), that is, an increase by 197%. A visual representation of these estimates can be seen by observing the regression lines in Figures 1.19 and 1.20.

Confidence Intervals Interpretation

We are 95% confident that in these three neighborhoods, a doubling of living area square footage is associated with an estimated sales price increase between $10^{.547}$ (3.52) and $10^{.687}$ (4.86). That is, between a 252% increase and a 386% increase. For Brookside, the estimated sale price increase is between $10^{.679}$ (4.77) and $10^{.96}$ (9.12). That is, between a 377% increase and an 812% increase. For Edwards, the estimated increase is between $10^{.494}$ (3.12) and $10^{.893}$ (7.82). That is, between a 212% increase and a 682% increase. For North Ames, the estimated increase is between $10^{.4}$ (2.51) and $10^{.546}$ (3.52). That is, an increase between 151% and 252%. A visual representation of these findings can be seen by observing the confidence interval bands on either side of the regression lines in Figures 1.19 and 1.20.

Conclusion

We have evidence that the raw data violates the linear regression model assumptions of normality and constant variance, but we were able to account for these violations by transforming the two variables on a log scale. The conditions for linearity and independence were reasonably met. After fitting a simple linear regression model, we found that the doubling of living area square footage is associated with an estimated 314% increase in Sale Price. We are 95% confident that the increase is between a 252% increase and a 386% increase. We found evidence that there are significant differences between the estimates for each of the 3 neighborhoods. The table in Figure 1.20 summarizes these findings for ease of comparison. These findings do not generalize to the larger Ames Housing dataset, as the three neighborhoods were not a random sample from the dataset. We also cannot make a causal inference since this is observational data. For an interactive demonstration of the regression model for each neighborhood, please see the iShiny App in Appendix B.

Analysis Question 2

Restatement of Problem

For the second section in our analysis, Century 21 would like our team to build the most predictive model for sale prices of homes in all of Ames, Iowa. Unlike the last analysis, this will include all neighborhoods in the dataset. In this analysis, we will be selecting and defending which model we feel predicts sale prices with the most accuracy. We will be including the charts and relevant SAS code in appendices B and D.

Model Selection

For the automatic selection models, we used all the numeric variables and allowed the automatic modeling to select the relevant variables according to its selection process. For the custom model, we used the variables for Lot Area, Year Built, Year Remodel Added, Overall Quality, and Overall Condition. We chose these based on their high Pearson's R score with the Sale Price variable and their p-values. Variables with a high Pearson's R score but p-values above alpha = 0.05 did not fit in the custom model.

Forward

Forward selection begins with no variables in the model. The selection process assesses different criterion for initial selection and begins adding variables based on that criterion. The model will continue

adding variables until it reaches a point where the CV Press Statistic cannot decrease anymore. It will then stop analyzing variables and report the output for the model.

Backward

Backward selection begins with all variables in the model. The selection process assesses different criterion for initial selection and begins subtracting variables based on that criterion. The model will continue subtracting variables until it reaches a point where the CV Press Statistic cannot decrease anymore. It will then stop analyzing variables and report the output for the model.

Stepwise

The stepwise model is a combination of both forward and backward selection. The model will perform both actions to determine which combination of addition or subtraction of variables gives the lowest value in the CV Press Statistic.

Custom

For our Custom model, we have selected variables we feel are the best predictors of the sale pricea We have run our model using the three processes above and determined that out of three runs, the stepwise model gave us the most accurate results.

Checking Assumptions

Linearity

There is visual evidence that a straight line is an adequate model for the linear regression for the variables chosen in our automatic models and our custom model with Sale Price. Figure 1.22 displays the scatterplot matrix for the custom model – these are a subset of the scatterplot matrix for the automatic selection model, which would be too large to be legible in this paper. We will address the outliers under the Constant Variance assumption discussion. Therefore, we will proceed as if the assumption of linearity is met.

Normality

Similar to what we found in Analysis Question #1, the raw data violated the normality assumption because it was influenced by the non-normality of the Sale Price variable. Thus, we transformed the Sale Price variable for the custom model as we did in Question #1. We also transformed the other variables in the Custom Model for continuity's sake. As seen in Figure 1.22, there is visual evidence of constant standard deviation with the log-transformed data for each of the custom model variables, and thus for normality. For the automatic selection models, we found that the models were robust to the normality assumption violation because of how many variables were included.

Constant Variance

The Residual plots for non-transformed data look to have clustered effect to them, rather than the cloud we are looking for. In this case, we again chose to view the logged data. After, we can see that the residuals take more of the cloud shape, and we will proceed with the logged data.

There are a few observations in the data with large leverage. As shown in Figure, we can see at least one large outlier with enormous leverage, ID 376 and a couple others with significant leverage, ID 534 and 1299. We are excluding 376, 534, and 1299 from our model. All three observations were sold at significantly reduced sales prices. As we discussed in Analysis Question #1, including these very low sale

price points would be measuring the effect of the extreme confounding variables and are not useful for the purposes of our analysis. There is an indication that these observations have a significant influence on the other sales prices. There are a few other outliers but we could not find an adequate reason to exclude them so they will remain in the model.

Independence

The discussion of the independence assumption violations in Analysis Question #1 applies to the entire dataset. Thus, as explained there, we will proceed as if the independence assumption is met.

Comparing Competing Models

Adj R2 and Internal CV Press

After running the 4 models we put together, the stepwise model on all the numeric variables seems to have the lowest CV press and the highest adjusted R squared. We ran each model three times and then took the averages to find out which would continuously produce the best results, as seen in the table below:

Predictive Models	Adjusted R2 1	CV PRESS 1	Adjusted R2 2	CV PRESS 2	Adjusted R2 3	CV PRESS 3	Avg. R2	Avg. CV PRESS
Forward	0.805	1.85E+12	0.7995	1.82E+12	0.7978	1.85E+12	0.800766667	1.84E+12
Backward	0.8048	1.94E+12	0.806	1.86E+12	0.8055	1.99E+12	0.805433333	1.93E+12
Stepwise	0.8048	1.80E+12	0.7978	1.85E+12	0.7978	1.88E+12	0.800133333	1.84E+12
Custom	0.77	2.24E+12	0.77	2.20E+12	0.77	2.26E+12	0.77	2.23E+12

Kaggle Score

Our best model finished with a Kaggle Score of .28864. This lowest score was from our Custom Model, which used the following variables to predict Sale Price: Lot Area, Year Built, Year Remodeled, Overall Quality, and Overall Condition. The automatic models on the numeric variables had Kaggle scores ranging from .4 to 2.8.

Conclusion

In conclusion, investigating both questions gave us interesting insight into the three neighborhoods in question and the predicted sales price of homes in Ames, Iowa. After performing the first analysis, we found that a doubling of living area square footage is associated with a large increase in sales price regardless of the neighborhood. However, Brookside has the largest increase per doubling of square footage at 560%. During the second analysis, we came up with a model that would predict the sales prices of homes in Ames, Iowa. We used the Forward, Backward, and Stepwise selection model on all the numeric variables and a custom selection of numeric variables. We found the Forward model using the custom selection of numeric variables to be the most accurate model for predicting Sale Price. We then used that model in our Kaggle submission which gave us an RMSE of .28864.

Appendix A: Works Cited

De Cock, Dean. "Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project." *Journal of Statistics Education.* Volume 19, Number 3 (2011), https://jse.amstat.org/v19n3/decock.pdf. Accessed 4/3/2023.

Appendix B: Links

Anthony Burton-Cordova

- https://github.com/Howdoinotgettrolled/StatsProject

Alexandra Thibeaux

- https://github.com/athibeaux/MSDS-DS6371

Rshiny App

- https://athibeaux.shinyapps.io/Ames-Housing-Linear-Regressio/

Appendix C: Figures and Plots

Figure 1.1: Simple Linear Regression Model for Sales Price and Living Area for 3 neighborhoods in Ames, Iowa. Neighborhoods are color-coded and legends are on the right of the graph.

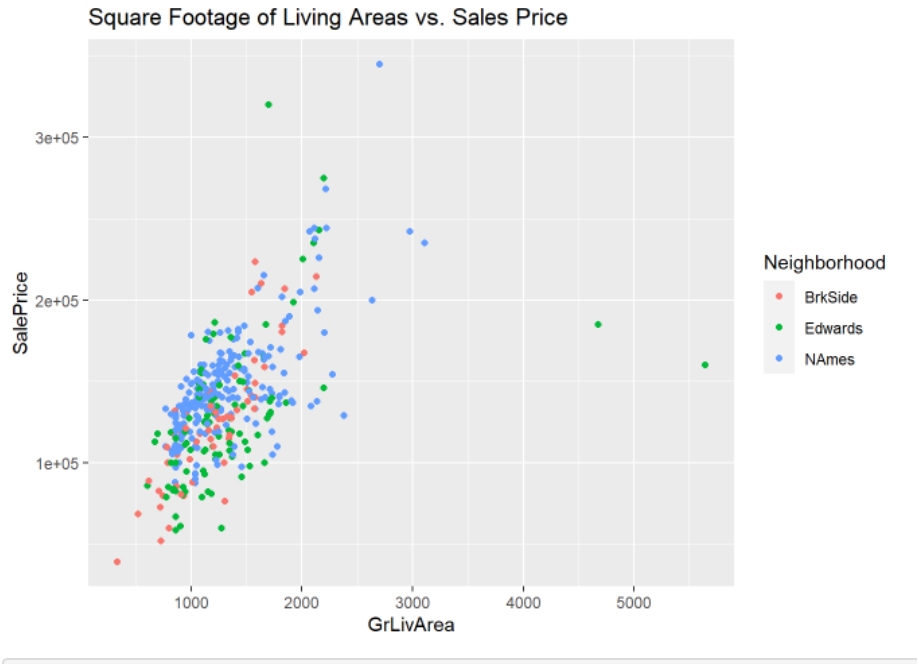


Figure 1.2: Multiple Linear Regression Model for Sales Price and Living Area for 3 neighborhoods in Ames, Iowa. Neighborhoods are color-coded and legends are on the right of the graph. The slope estimates are plotted, as well as the confidence intervals for each slope.

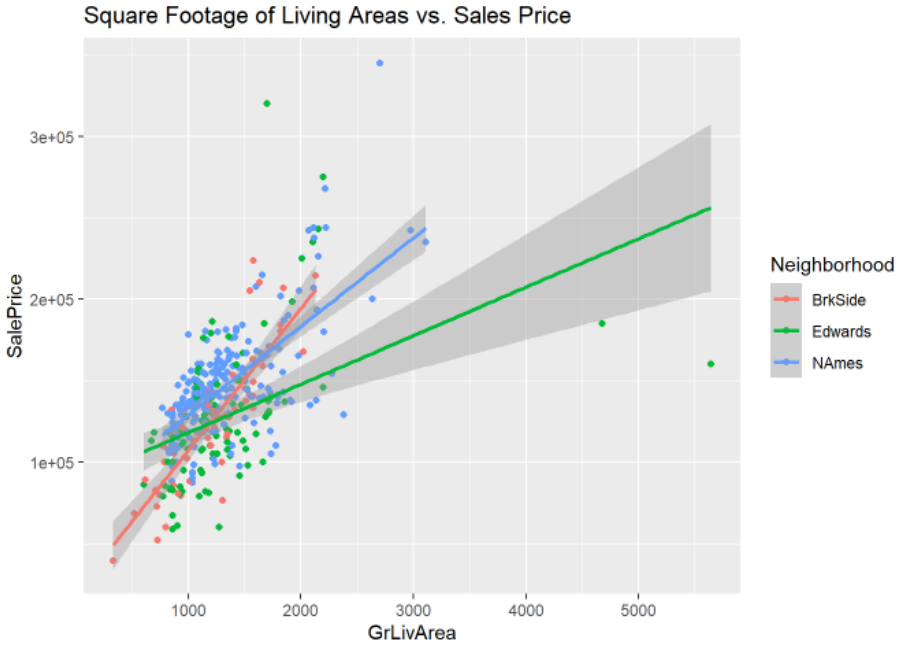


Figure 1.3: Distribution of Living Area Variable - Evidence of Right tailed-ness

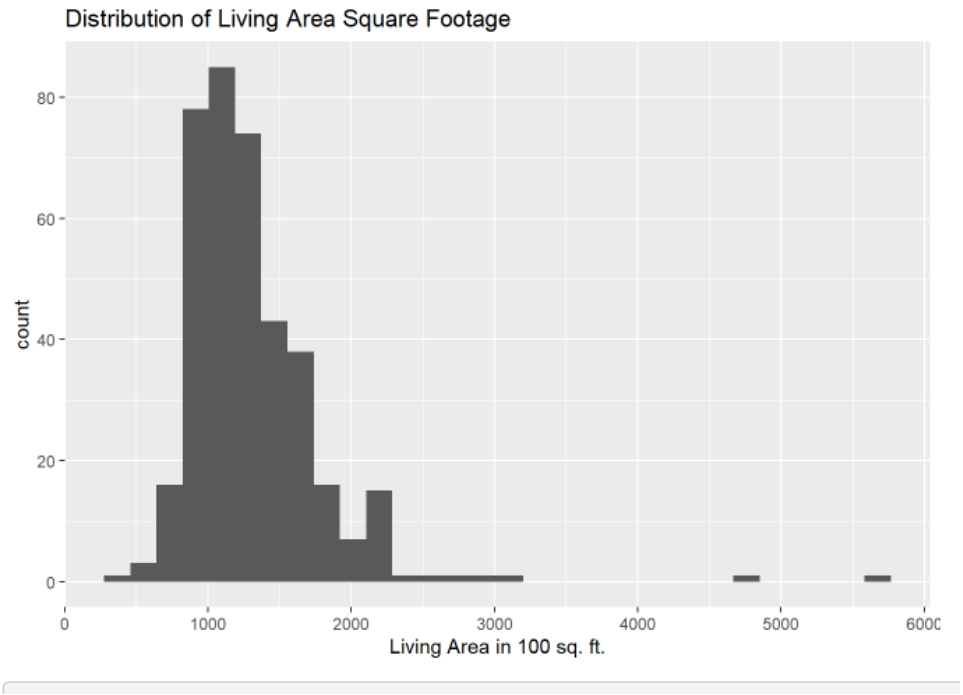


Figure 1.4: Distribution of Sale Price Variable - Evidence of Right tailed-ness

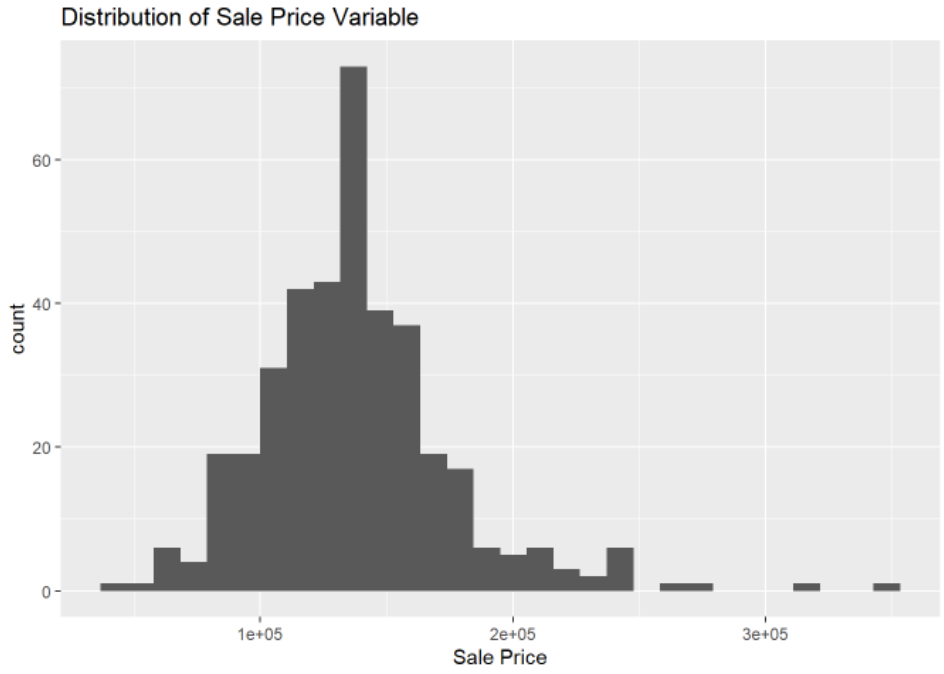


Figure 1.5: QQ Plot of Living Area Variable, slight upward curve visible

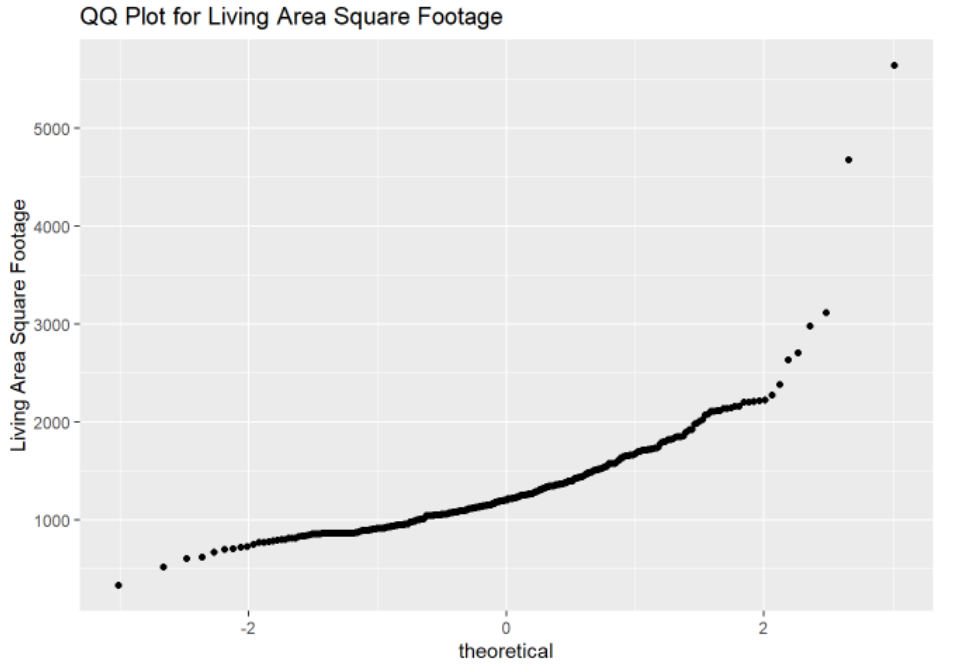


Figure 1.6: QQ Plot of Sale Price variable, slight upward curve visible

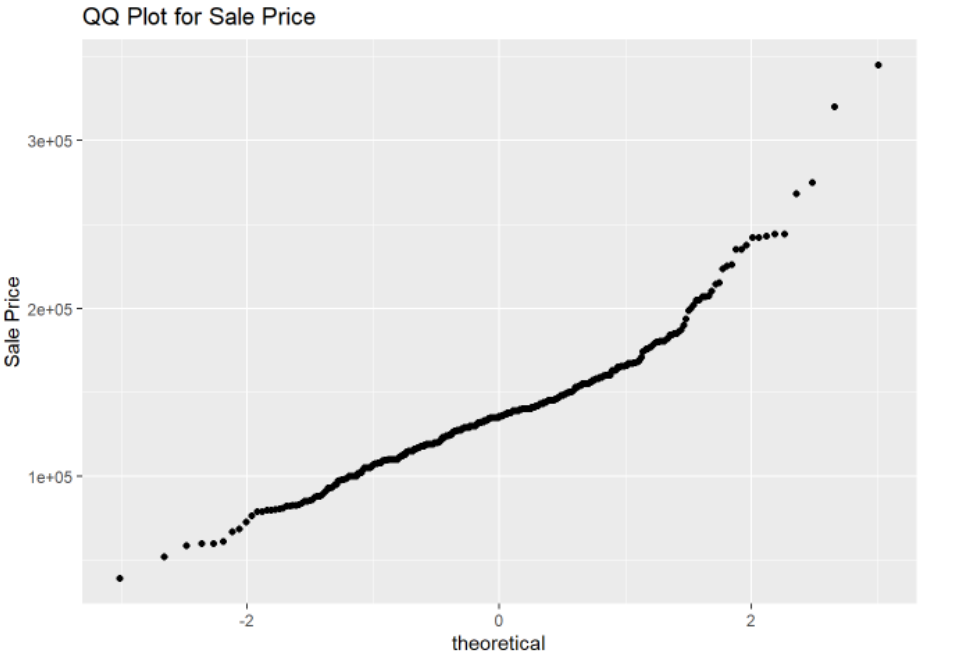


Figure 1.7: Distribution of log-transformed Living Area variable, evidence of normal distribution, robust to right tailed-ness

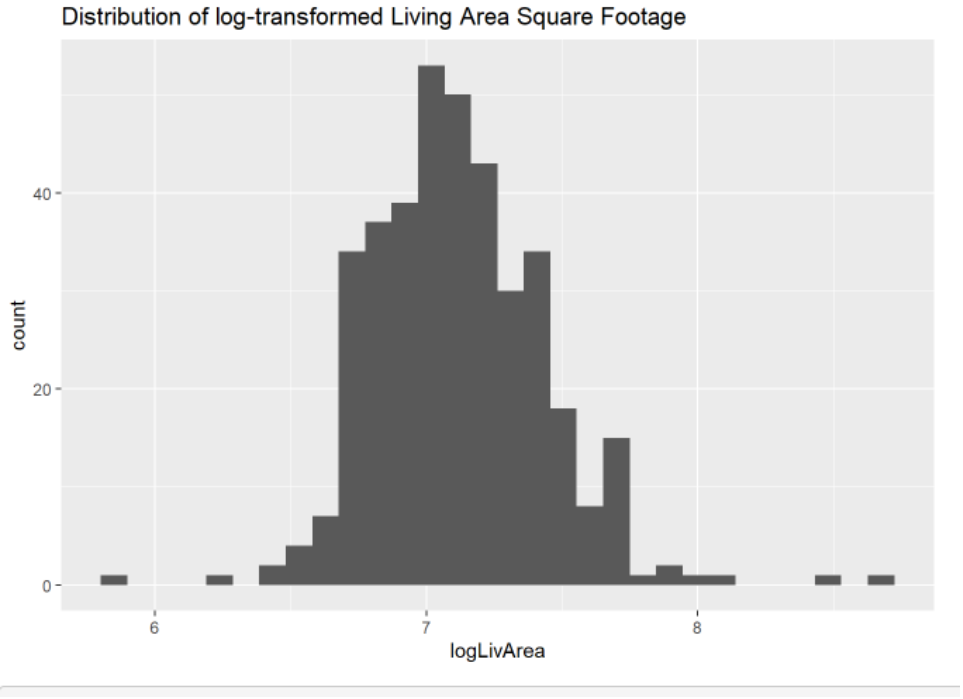


Figure 1.8: Distribution of log-transformed Sale Price variable, evidence of normal distribution, robust to right tailed-ness

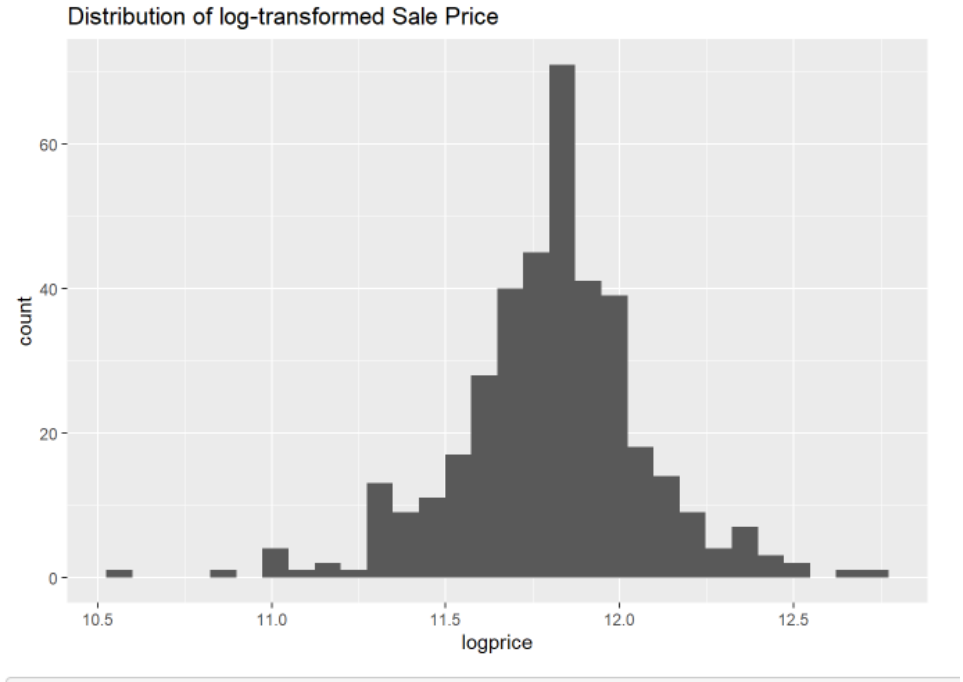


Figure 1.9: QQ Plot for log-transformed Living Area variable, sample distribution more in line with theoretical distribution

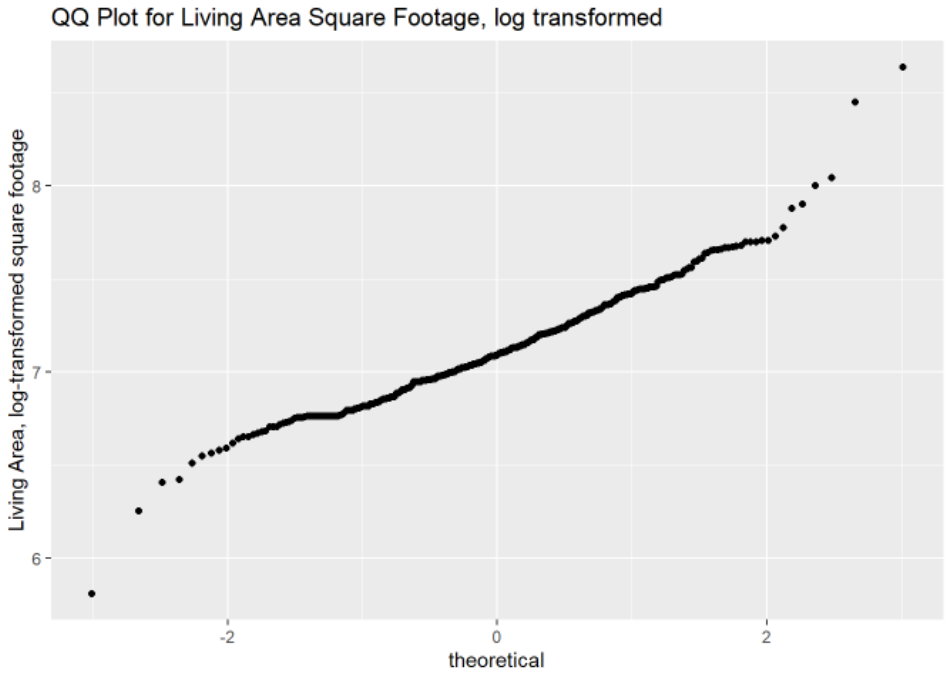


Figure 1.10: QQ Plot for log-transformed Sale Price variable, sample distribution more in line with theoretical distribution

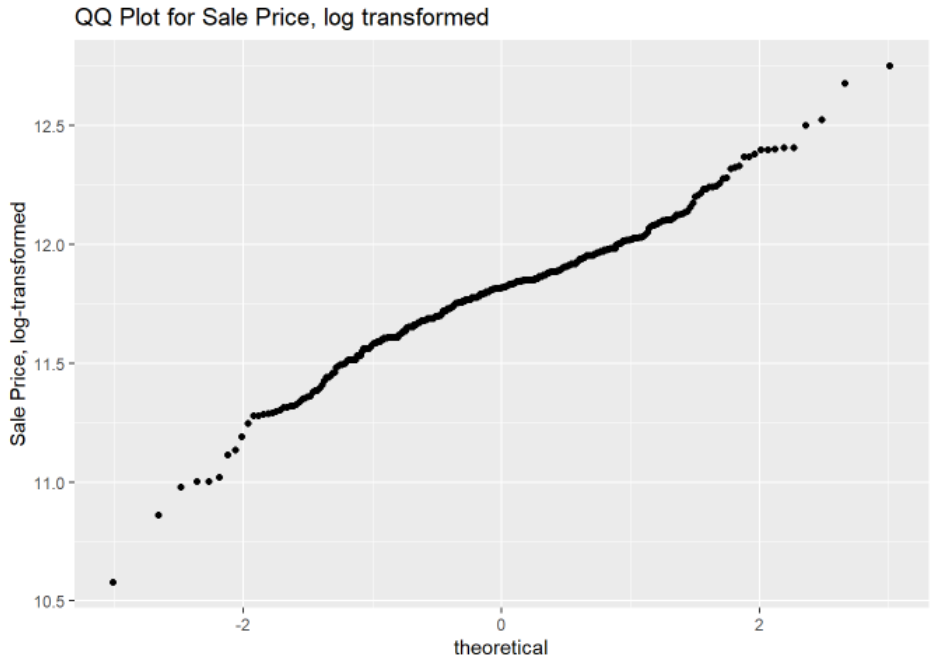


Figure 1.11: Residuals vs Fitted values plot for raw data, showing a cluster effect.

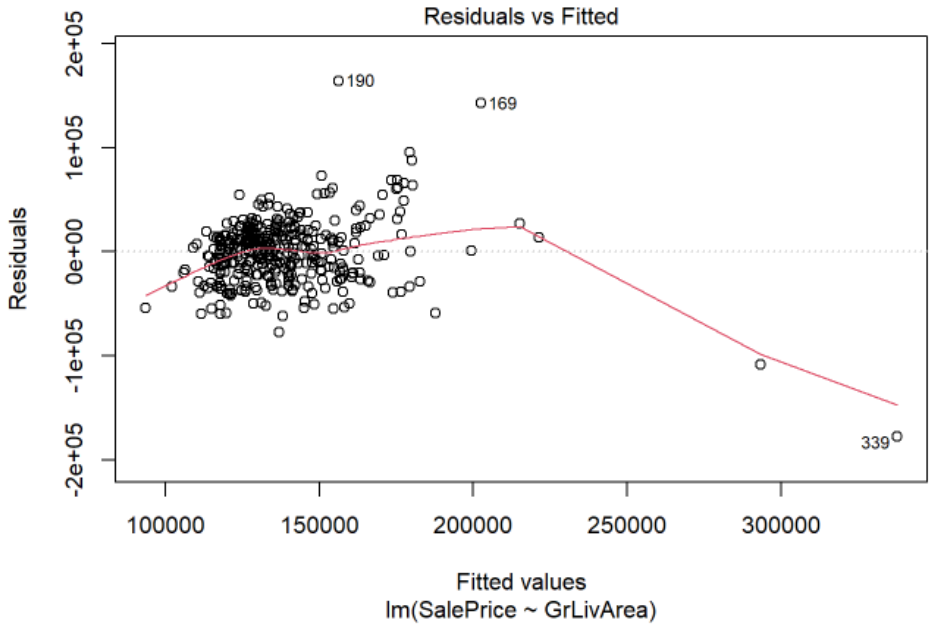


Figure 1.12: Residuals vs. Fitted values plot for log-transformed data, with more randomly scattered residuals.

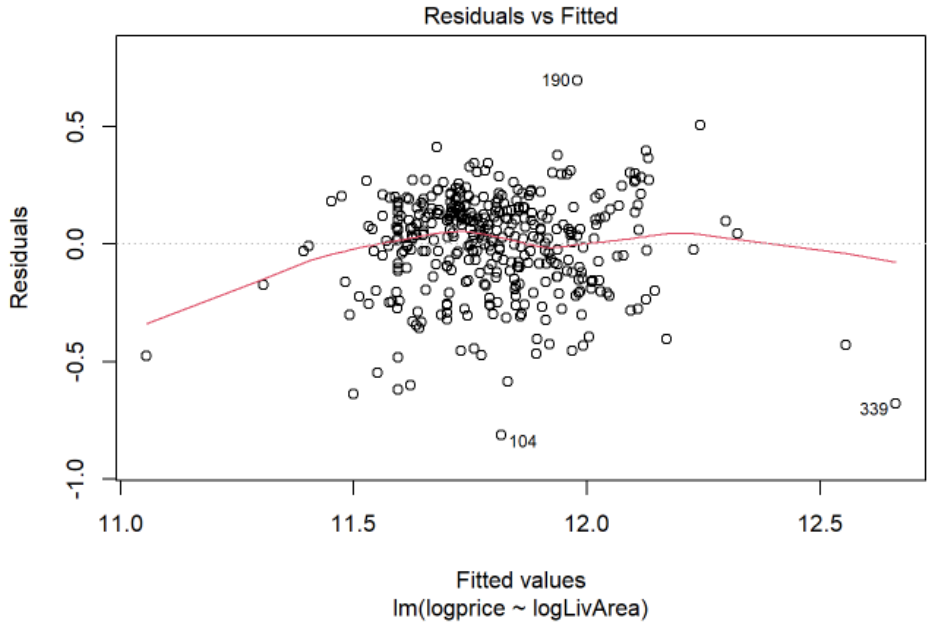


Figure 1.13: Residual Plot for Residuals vs Leverage of raw data, showing two observations that exceed Cook's Distance

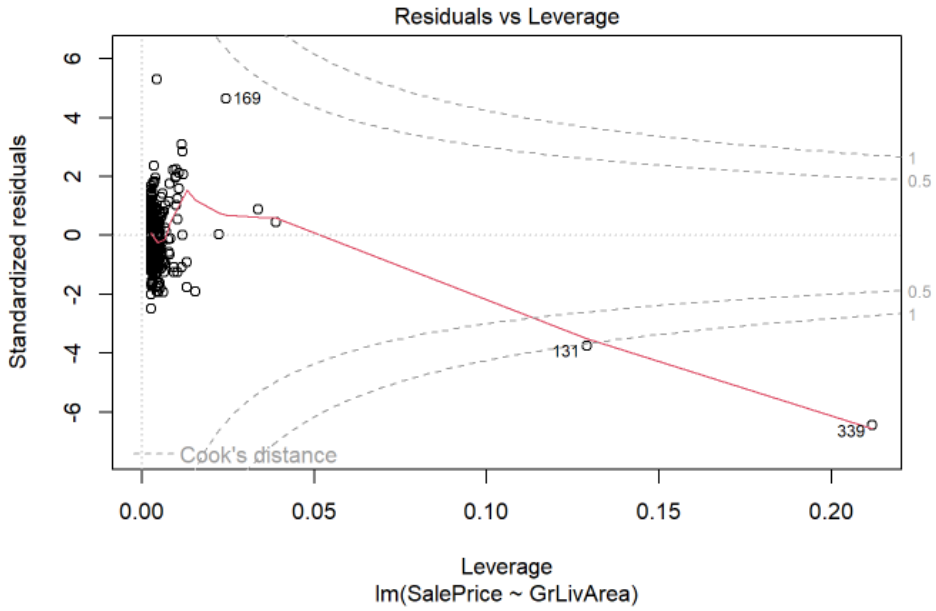


Figure 1.14: Residual plot of residuals vs. leverage for log-transformed data, showing three points approaching Cook's Distance

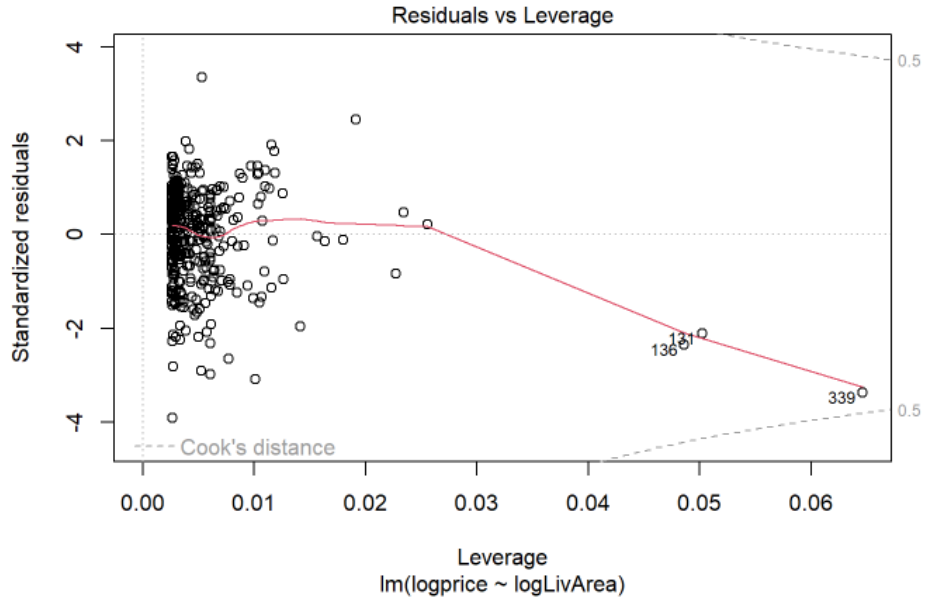


Figure 1.15: Residual plot after removing 2 invalid data points, showing a random scattering of residuals.

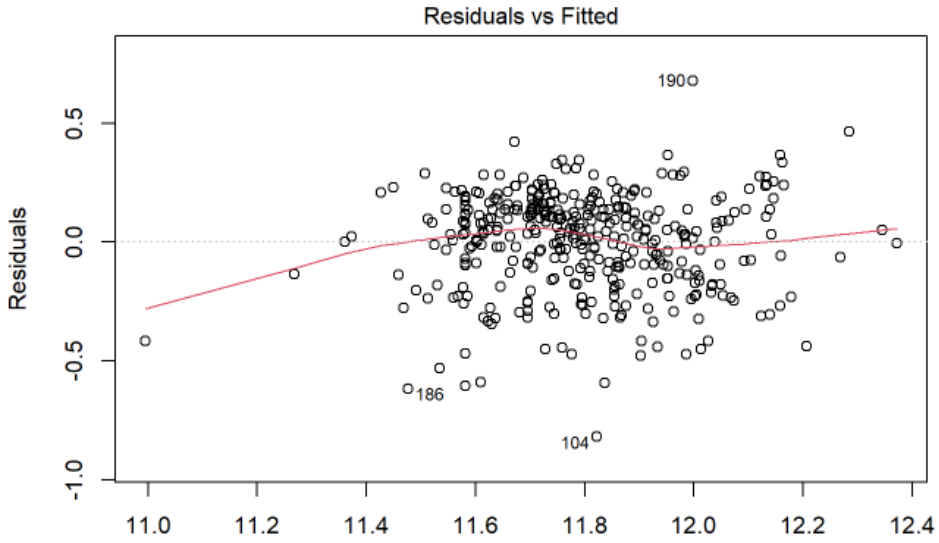


Figure 1.16

Measure: Adjusted R^2	Overall	Brookside	Edwards	North Ames
Raw Data	0.4559	0.6921	0.386	0.4587
Transformed Data	0.4415	0.7029	0.3598	0.417

Figure 1.17

Measure: Adjusted R^2	Overall	Brookside	Edwards	N. Ames
Raw Data, holding Neighborhood constant	0.4905	0.6921	0.386	0.4587
Transformed Data, holding Neighborhood constant	0.5002	0.7029	0.3598	0.417

Figure 1.18

Measure: R^2	Model R^2	Internal CV	Difference
Raw Data	0.4573	0.4486	0.0087
Transformed Data	0.443	0.4361	0.0069
Raw Data, holding Neighborhood constant	0.4945	0.4802	0.0143
Transformed Data, holding Neighborhood constant	0.5041	0.4914	0.0127

Figure 1.20: Scatterplot displaying Regression line with confidence intervals for the explanatory variable of Living Area Square Footage (transformed on the log scale) and the response variable of Sale Price (transformed on the log scale)

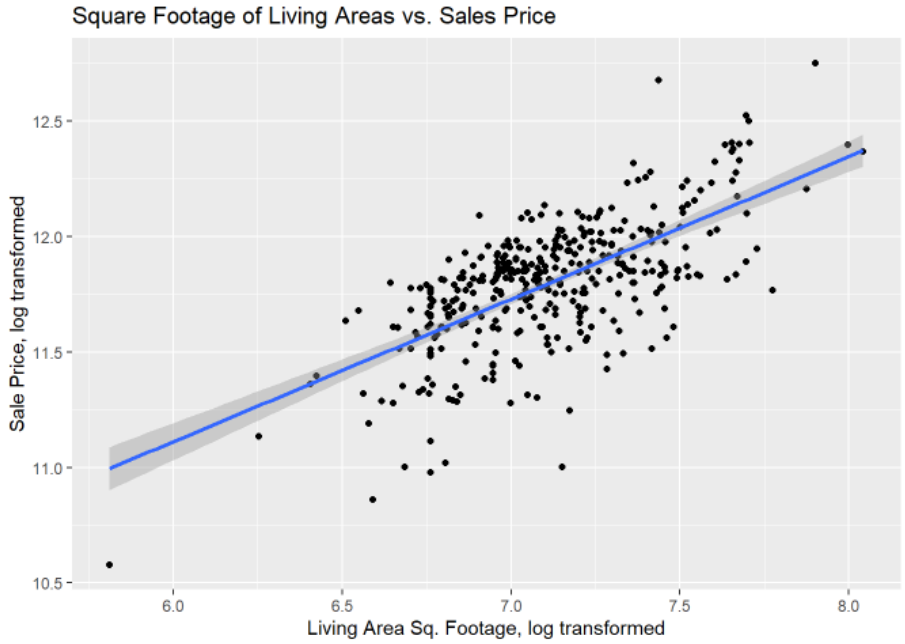


Figure 1.21: Scatterplot displaying regression lines with confidence intervals for the neighborhoods of Brookside, Edwards, and North Ames using Living Area Square Footage as the explanatory variable (transformed on the log scale) and the response variable as Sale Price (transformed on the log scale).

Figure 1.21

Model:	Estimate	% Increase	Lower CI	% Increase	Upper CI	% Increase
Overall	$10^{.617}$ (4.14)	314%	$10^{.547}$ (3.52)	252%	$10^{.687}$ (4.86)	386%
Brookside	$10^{.82}$ (6.6)	560%	$10^{.679}$ (4.77)	377%	$10^{.96}$ (9.12)	812%
Edwards	$10^{.673}$ (4.7)	370%	$10^{.494}$ (3.12)	212%	$10^{.893}$ (7.82)	682%
North Ames	$10^{.473}$ (2.97)	197%	$10^{.4}$ (2.51)	151%	$10^{.546}$ (3.52)	252%

Figure 1.22: Scatterplot matrix of the log-transformed data for the custom model

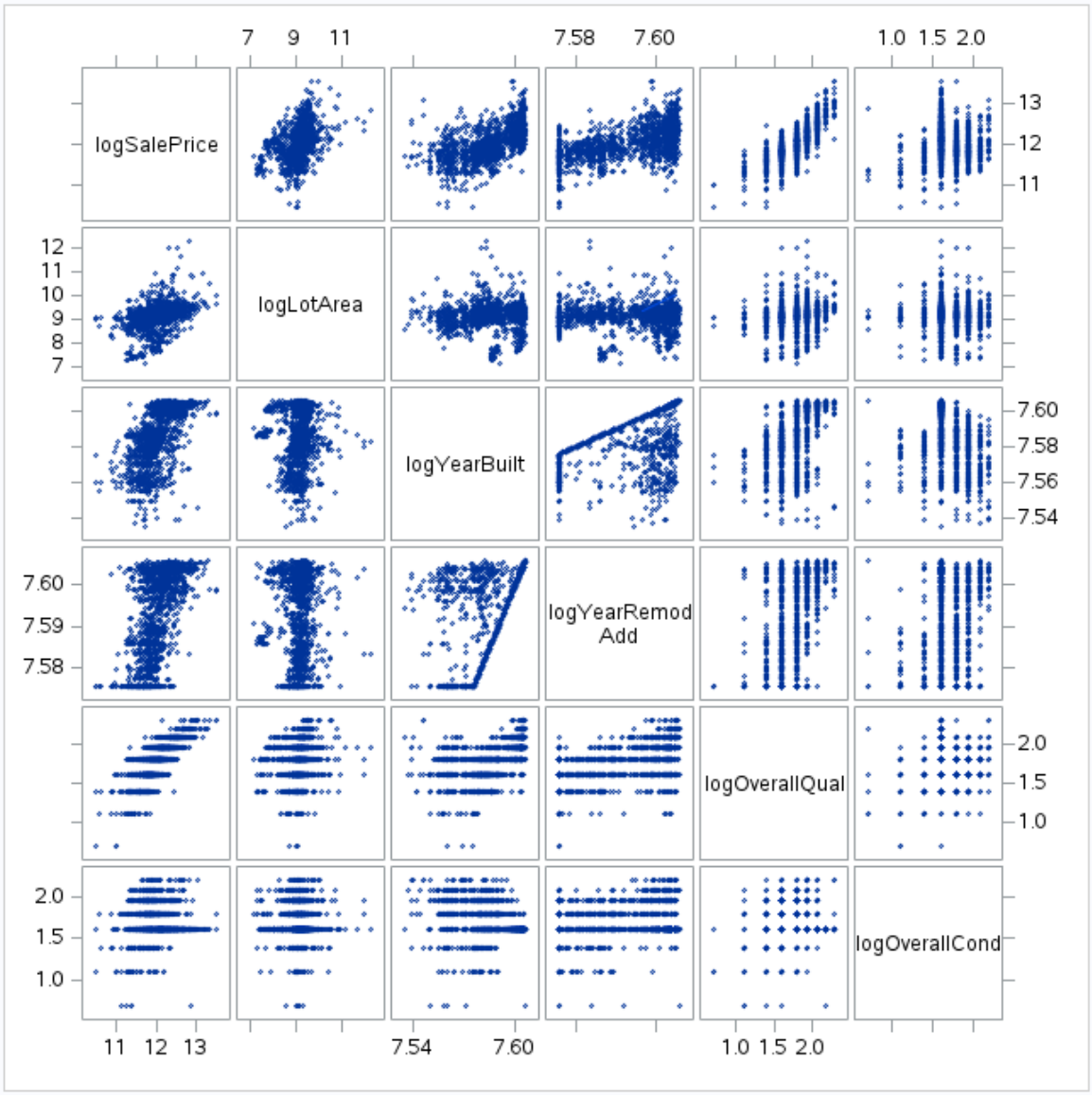


Figure 1.23: Correlation statistics for the Custom model

Pearson Correlation Coefficients, N = 1457 Prob > |r| under H0: Rho=0						
	logSalePrice	logLotArea	logYearBuilt	logYearRemodAdd	logOverallQual	logOverallCond
logSalePrice	1.00000	0.40253 <.0001	0.58451 <.0001	0.56365 <.0001	0.80003 <.0001	-0.01367 0.6022
logLotArea	0.40253 <.0001	1.00000	0.01879 0.4735	0.02394 0.3611	0.15971 <.0001	-0.00499 0.8492
logYearBuilt	0.58451 <.0001	0.01879 0.4735	1.00000	0.58841 <.0001	0.56241 <.0001	-0.32416 <.0001
logYearRemodAdd	0.56365 <.0001	0.02394 0.3611	0.58841 <.0001	1.00000	0.54092 <.0001	0.08103 0.0020
logOverallQual	0.80003 <.0001	0.15971 <.0001	0.56241 <.0001	0.54092 <.0001	1.00000	-0.03439 0.1896
logOverallCond	-0.01367 0.6022	-0.00499 0.8492	-0.32416 <.0001	0.08103 0.0020	-0.03439 0.1896	1.00000

Figure 1.24: Cook's D chart for identifying outliers and their leverage

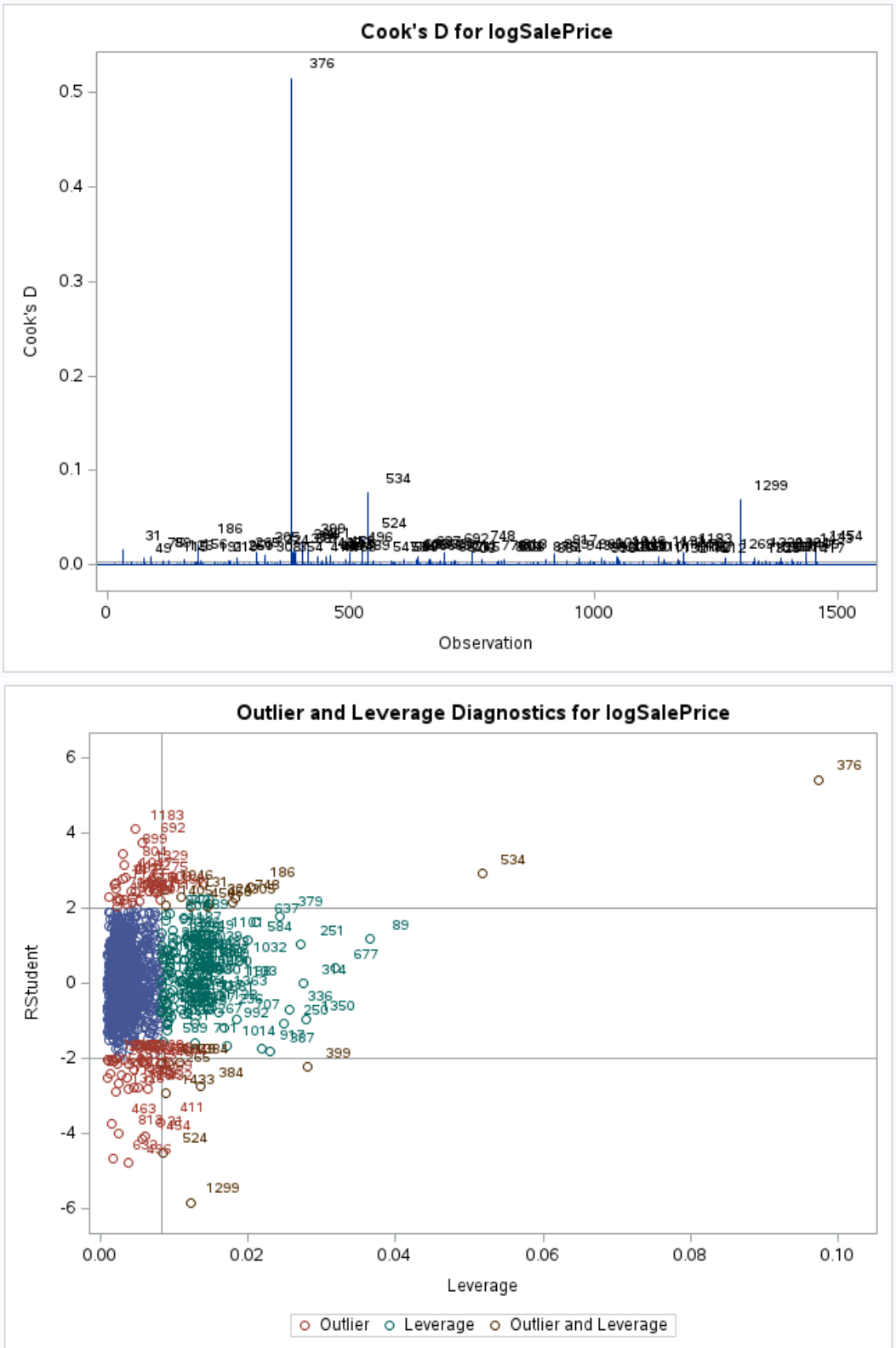
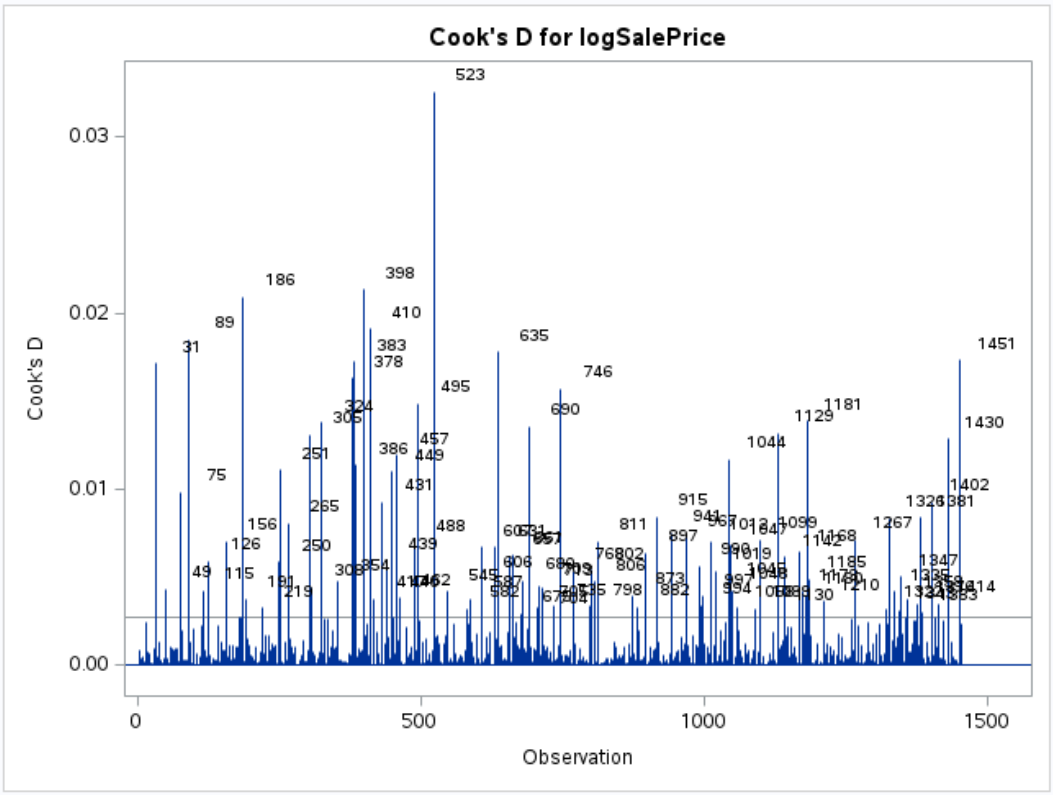
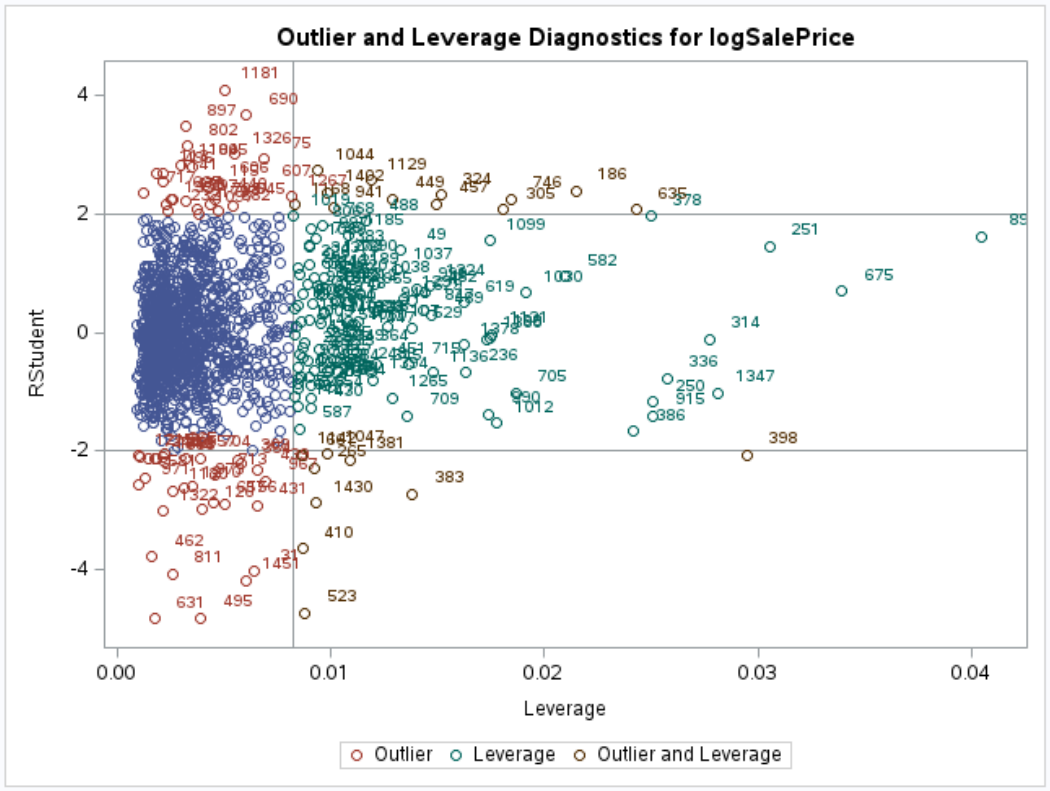


Figure 1.25: Cook's D chart for identifying outliers and their leverage





Appendix D: R Code for Analysis Question #1

Ames_Housing_Data

Burton-Cordova & Thibeaux

2023-04-04

Load Data

```
#Load Test Data
 test = read.csv('https://github.com/athibeaux/MSDS-
DDS/raw/main/Project/test.csv', header = TRUE, fill = TRUE)

#Load Train Data
 train = read.csv('https://github.com/athibeaux/MSDS-
DDS/raw/main/Project/train.csv', header = TRUE, fill = NA)

#Summary of Train Data
 summary(train)

#Select Relevant Columns and Neighborhoods for Analysis Question 1
 C21 = train %>% select(GrLivArea,Neighborhood,SalePrice) %>%
filter(Neighborhood == "NAmes" | Neighborhood == "Edwards" | Neighborhood ==
"BrkSide")

C21$Neighborhood <- as.factor(C21$Neighborhood)
 summary(C21)

#Check and Remove NA's
 sum(is.na(C21$GrLIvArea))
```

Addressing Assumptions

Linearity

```
# Without Lines
 C21 %>% ggplot(aes(GrLivArea, SalePrice, color = Neighborhood)) +
geom_point() +
   xlab("Living Area in 100 Sq. Feet") + ylab("Sale Price") +
   ggtitle("Square Footage of Living Areas vs. Sales Price")

# With Lines for each Neighborhood
 C21 %>% ggplot(aes(GrLivArea, SalePrice, color = Neighborhood)) +
geom_point() +
   geom_smooth(method = "lm") +
```

```
  xlab("Living Area in 100 Sq. Feet") + ylab("Sale Price") +
   ggtitle("Square Footage of Living Areas vs. Sales Price")
```

Normality

```
ggpairs(C21)

# Histogram for Sale Price
 C21 %>% ggplot() + geom_histogram(aes(x = SalePrice)) +
  ggtitle("Distribution of Sale Price Variable") + xlab("Sale Price")

# Histogram for Living Area Square Footage
 C21 %>% ggplot() + geom_histogram(aes(x = GrLivArea)) +
   ggtitle("Distribution of Living Area Square Footage") +
   xlab("Living Area in 100 sq. ft.")

# QQ Plot for Living Area Square Footage
 C21 %>% ggplot() + geom_qq(aes(sample = GrLivArea)) +
  ggtitle("QQ Plot for Living Area Square Footage") + ylab("Living Area
Square Footage")

# QQ Plot for Sale Price
 C21 %>% ggplot() + geom_qq(aes(sample = SalePrice)) +
  ggtitle("QQ Plot for Sale Price") + ylab("Sale Price")

# Log Transformation on GrLivArea
 C21$logLivArea = log(C21$GrLivArea)

# Histogram for Log-Transformed Living Area Square Footage
 C21 %>% ggplot() + geom_histogram(aes(x = logLivArea)) +
   xlab("Living Area, log-transformed") +
   ggtitle("Distribution of log-transformed Living Area Square Footage")

# QQ Plot for Log-Transformed Living Area Square Footage
 C21 %>% ggplot() + geom_qq(aes(sample = logLivArea)) +
  ggtitle("QQ Plot for Living Area Square Footage, log transformed") +
  ylab("Living Area, log-transformed square footage")

# Log Transformation on Sale Price
 C21$logprice = log(C21$SalePrice)

# Histogram for Log-Transformed Sale Price
 C21 %>% ggplot() + geom_histogram(aes(x = logprice)) +
   xlab("Sale Price, log-transformed")
   ggtitle("Distribution of log-transformed Sale Price")

# QQ Plot for Log-Transformed Living Area Square Footage
```

```
 C21 %>% ggplot() + geom_qq(aes(sample = logprice)) +
  ggtitle("QQ Plot for Sale Price, log transformed") +
  ylab("Sale Price, log-transformed")
```

Equal Variance

```
# Non Transformed Data
 rd <- lm(SalePrice ~ GrLivArea, data = C21)
 g = rd$residuals
 m<-mean(g)
 std<-sqrt(var(g))
 hist(g, density=20, breaks=20, prob=TRUE, col="red",
      xlab="Residuals",
     main="Residual Histogram with Normal Distribution")
 curve(dnorm(x, mean=m, sd=std),
      col="darkblue", lwd=2, add=TRUE, yaxt="n")

# Display residual plots for non-transformed data
 plot(rd)

# Notice cluster effect of residuals!

# Log-Log Model

log_rd <- lm(logprice ~ logLivArea, data = C21)
 log_g = log_rd$residuals
 log_m<-mean(log_g)
 log_std<-sqrt(var(log_g))
 hist(log_g, density=20, breaks=20, prob=TRUE, col="red",
      xlab="Residuals",
     main="Residual Histogram with Normal Distribution")
 curve(dnorm(x, mean=log_m, sd=log_std),
      col="darkblue", lwd=2, add=TRUE, yaxt="n")

# Display residual plots for log-transformed data
 plot(log_rd)
```

We have a few high leverage residuals. Let's take a closer look.

```
# Examine the largest values in the explanatory variable
 SortBySqFt <- C21[order(C21$GrLivArea, decreasing = TRUE),]
 head(SortBySqFt)

# Wow! Two houses with >4000 square feet were sold for under $200k?? Either
that's a mistake (like someone left out a zero) or I need the number of that
real estate agent IMMEDIATELY so he can get me that kind of price on a
mansion! If these two outliers are an valid observation, we may be looking at
a short sale or foreclosure. Since these kinds of sales are not what Century
```

```
21 wants to measure, we feel confident in removing them from our dataset.

# Save complete dataset as a separate object in case we need it later
 C21_full <- C21

# Find rows with outlier data
 which(C21[ , 1] > 4000)

# Those are the same datapoints highlighted in our Residuals vs Leverage
plot, so let's remove them.
 C21 <- C21[-c(131,339),]
```

Now that we've removed the invalid data points, let's look again at the residuals.

```
# Log-Log Model

log_rd <- lm(logprice ~ logLivArea, data = C21)
 log_g = log_rd$residuals
 log_m<-mean(log_g)
 log_std<-sqrt(var(log_g))
 hist(log_g, density=20, breaks=20, prob=TRUE, col="red",
      xlab="Residuals",
     main="Residual Histogram with Normal Distribution")
 curve(dnorm(x, mean=log_m, sd=log_std),
      col="darkblue", lwd=2, add=TRUE, yaxt="n")

# Display new residual plots
 plot(log_rd)
```

Now let's take another look at the fits

```
# Parameter Estimates
 # Overall
 fit = lm(logprice ~ logLivArea, data = C21)
 summary(fit)
 confint(fit)

# Without Lines
 C21 %>% ggplot(aes(logLivArea, logprice, color = Neighborhood)) +
   geom_point() +
   ylab("Sale Price, log transformed") +
  xlab("Living Area Sq. Footage, log transformed") +
  ggtitle("Square Footage of Living Areas vs. Sales Price")

# By Neighborhood
 fit_hoods = C21 %>% group_by(Neighborhood) %>% do(model = lm(logprice ~
logLivArea, data = .))
```

```
# Brookside
 summary(fit_hoods[[2]][[1]])
 confint(fit_hoods[[2]][[1]])

# Edwards
 summary(fit_hoods[[2]][[2]])
 confint(fit_hoods[[2]][[2]])

# North Ames
 summary(fit_hoods[[2]][[3]])
 confint(fit_hoods[[2]][[3]])

# With Regression Line
 C21 %>% ggplot(aes(logLivArea, logprice)) +
  geom_point() + geom_smooth(method = "lm") +
   ylab("Sale Price, log transformed") +
  xlab("Living Area Sq. Footage, log transformed") +
  ggtitle("Square Footage of Living Areas vs. Sales Price")

# With Lines for each Neighborhood
 C21 %>% ggplot(aes(logLivArea, logprice, color = Neighborhood)) +
  geom_point() + geom_smooth(method = "lm") +
   ylab("Sale Price, log transformed") +
  xlab("Living Area Sq. Footage, log transformed") +
  ggtitle("Square Footage of Living Areas vs. Sales Price")
```

Compare competing models

Fit on non-transformed data

```
# Fit on non-transformed data
 # Parameter Estimates
 # Overall
 fit = lm(SalePrice ~ GrLivArea, data = C21)
 summary(fit)
 confint(fit)

# Without Lines
 C21 %>% ggplot(aes(GrLivArea, SalePrice, color = Neighborhood)) +
   geom_point() +
   ylab("Sale Price") +
  xlab("Living Area Sq. Footage") +
  ggtitle("Square Footage of Living Areas vs. Sales Price")

# By Neighborhood
 fit_hoods = C21 %>% group_by(Neighborhood) %>% do(model = lm(SalePrice ~
```

```
GrLivArea, data = .))

# Brookside
 summary(fit_hoods[[2]][[1]])
 confint(fit_hoods[[2]][[1]])

# Edwards
 summary(fit_hoods[[2]][[2]])
 confint(fit_hoods[[2]][[2]])

# North Ames
 summary(fit_hoods[[2]][[3]])
 confint(fit_hoods[[2]][[3]])

# With Lines for each Neighborhood
 C21 %>% ggplot(aes(GrLivArea, SalePrice, color = Neighborhood)) +
  geom_point() + geom_smooth(method = "lm") +
   ylab("Sale Price") +
  xlab("Living Area Sq. Footage") +
  ggtitle("Square Footage of Living Areas vs. Sales Price")
```

Hold Neighborhood Constant

```
# Transformed Data
 fit = lm(logprice ~ logLivArea + Neighborhood, data = C21)
 summary(fit)
 confint(fit)

# Raw Data
 fit = lm(SalePrice ~ GrLivArea + Neighborhood, data = C21)
 summary(fit)
 confint(fit)
```

Internal CV as a measure for competing models

```
# LOOCV method from library(caret), transformed data

train(logprice ~ logLivArea, method = "lm", data = C21, trControl =
trainControl(method = "LOOCV"))

# LOOCV method from library(caret), raw data

train(SalePrice ~ GrLivArea, method = "lm", data = C21, trControl =
trainControl(method = "LOOCV"))

# LOOCV method from library(caret), transformed data, holding Neighborhood
constant
```

```
train(logprice ~ logLivArea + Neighborhood, method = "lm", data = C21,
trControl = trainControl(method = "LOOCV"))

# LOOCV method from library(caret), raw data, holding Neighborhood constant

train(SalePrice ~ GrLivArea + Neighborhood, method = "lm", data = C21,
trControl = trainControl(method = "LOOCV"))
```

Appendix E: SAS Code for Analysis Question #2

```
/* Generated Code (IMPORT) */
/* Source File: AmesTrainData.csv */
/* Source Path: /home/u63075710/DS6371/Final Project */
/* Code generated on: 4/7/23, 11:03 AM */

*Train Data;
%web_drop_table(WORK.IMPORT);



FILENAME Data5 '/home/u63075710/DS6371/Final Project/AmesTrainData.csv';

PROC IMPORT DATAFILE=Data5
        DBMS=CSV
        OUT=WORK.Data5;
        GETNAMES=YES;
RUN;

PROC CONTENTS DATA=WORK.Data5; RUN;



%web_open_table(WORK.Data5);

*Test Data;
%web_drop_table(WORK.IMPORT);



FILENAME Data6 '/home/u63075710/DS6371/Final Project/AmesTestData.csv';
```

```
PROC IMPORT DATAFILE=Data6
        DBMS=CSV
        OUT=WORK.Data6;
        GETNAMES=YES;
RUN;

PROC CONTENTS DATA=WORK.Data6; RUN;



%web_open_table(WORK.Data6);

*Create the SalePrice Column;
Data Data6;
set Data6;
SalePrice = .;
run;

*Combine Data from Test and Train;
data Train1;
set data232 Data6;
run;

/* Adding Log Data*/
data Data5;
set Data5;
logSalePrice = log(SalePrice);
logLotArea = log(LotArea);
logYearBuilt = log(YearBuilt);
logYearRemodAdd = log(YearRemodAdd);
```

```
logOverallQual = log(OverallQual);
logOverallCond = log(OverallCond);
;

*Delete Step from Train Data;
data data232;
set Data5;
if _n_ = 376 then delete;
if _n_ = 534 then delete;
if _n_ = 1299 then delete;
run;

proc print data = data232;
run;

*We can view the data, but it takes a few min to load;
proc print Data = Data5;
run;

*Look at the correlation from selected variables;
proc corr data = Data1 plots = scatter;
var SalePrice LotArea YearBuilt YearRemodAdd OverallQual OverallCond;
run;

*Look at the correlation from selected and logged variables;
proc corr data = Data1 plots = scatter;
var SalePrice logLotArea logYearBuilt logYearRemodAdd logOverallQual logOverallCond;
run;
```

```
*Look at the correlation from selected and logged variables (Highest Pearson);

proc corr data = Data232 plots = scatter;

var logSalePrice logLotArea logYearBuilt logYearRemodAdd logOverallQual logOverallCond;

run;


*Look at the correlation from selected and logged variables (Highest Pearson);

proc sgscatter data = Data232;

matrix logSalePrice logLotArea logYearBuilt logYearRemodAdd logOverallQual logOverallCond;

run;


*Look at the correlation from Numeric variables;

proc corr data = Data5 plots = scatter;

var SalePrice TotRmsAbvGrd Fireplaces GarageYrBlt GarageCars GarageArea WoodDeckSF OpenPorchSF
EnclosedPorch ScreenPorch

PoolArea MiscVal MoSold YrSold MSSubClass LotArea OverallQual OverallCond YearBuilt YearRemodAdd
MasVnrArea BsmtFinSF1 BsmtFinSF2

BsmtUnfSF TotalBsmtSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath
BedroomAbvGr KitchenAbvGr;

run;


*Fit and Residualsfrom selected variables;

proc reg data = Data1;

model SalePrice = LotArea YearBuilt YearRemodAdd OverallQual OverallCond / stb clb VIF scorr1 scorr2;

run;


*Fit and Residualsfrom from selected and logged variables;

proc reg data = data232 plots(only label) =(CooksD RStudentByLeverage);

model logSalePrice = logLotArea logYearBuilt logYearRemodAdd logOverallQual logOverallCond / stb clb
VIF scorr1 scorr2;

run;
```

```
*Fit and Residualsfrom from selected and logged variables;

proc reg data = Data5 plots(only label) =(CooksD RStudentByLeverage);

model logSalePrice = logLotArea logYearBuilt logYearRemodAdd logOverallQual logOverallCond / stb clb
VIF scorr1 scorr2;

run;


*Fit and Residualsfrom from Numeric variables;

proc reg data = Data1;

model SalePrice = TotRmsAbvGrd Fireplaces GarageYrBlt GarageCars GarageArea WoodDeckSF
OpenPorchSF EnclosedPorch ScreenPorch

PoolArea MiscVal MoSold YrSold MSSubClass LotArea OverallQual OverallCond YearBuilt YearRemodAdd
MasVnrArea BsmtFinSF1 BsmtFinSF2

BsmtUnfSF TotalBsmtSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath
BedroomAbvGr KitchenAbvGr / stb clb VIF scorr1 scorr2;

run;


/* Variable Selection with Sig Level (Without Class function and Catagorical Variables) */


*Forward Selection;

proc glmselect data = Train1 plots= all;


model SalePrice = TotRmsAbvGrd Fireplaces GarageYrBlt GarageCars GarageArea WoodDeckSF
OpenPorchSF EnclosedPorch ScreenPorch

PoolArea MiscVal MoSold YrSold MSSubClass LotArea OverallQual OverallCond YearBuilt YearRemodAdd
MasVnrArea BsmtFinSF1 BsmtFinSF2

BsmtUnfSF TotalBsmtSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath
BedroomAbvGr KitchenAbvGr/

selection = Forward(stop = CV SLE = .2) stats = adjrsq;

output out= Results p= Predict;

run;
```

```
*Remove negative predictions;
data Results2;
set results;
if Predict > 0 then SalePrice = Predict;
if Predict < 0 then SalePrice = 10000;
keep id SalePrice;
where id > 1460;

proc means data= Results2;
var SalePrice;
run;

proc export data= results2
        outfile= "/home/u63075710/DS6371/Final Project/Results2.csv"
        dbms=csv;
run;



*Backward Selection;
proc glmselect data = Data232;

model SalePrice = TotRmsAbvGrd Fireplaces GarageYrBlt GarageCars GarageArea WoodDeckSF
OpenPorchSF EnclosedPorch ScreenPorch
PoolArea MiscVal MoSold YrSold MSSubClass LotArea OverallQual OverallCond YearBuilt YearRemodAdd
MasVnrArea BsmtFinSF1 BsmtFinSF2
BsmtUnfSF TotalBsmtSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath
BedroomAbvGr KitchenAbvGr/
selection = Backward(stop  = CV SLS = .2) stats = adjrsq;
score data = Data232 out = MYOUTPUT;
```

```
run;

*Stepwise Selection;
proc glmselect data = Data232;

model SalePrice = TotRmsAbvGrd Fireplaces GarageYrBlt GarageCars GarageArea WoodDeckSF
OpenPorchSF EnclosedPorch ScreenPorch
PoolArea MiscVal MoSold YrSold MSSubClass LotArea OverallQual OverallCond YearBuilt YearRemodAdd
MasVnrArea BsmtFinSF1 BsmtFinSF2
BsmtUnfSF TotalBsmtSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath
BedroomAbvGr KitchenAbvGr/
selection = stepwise (stop = CV) stats = adjrsq;
score data = Data232 out = MYOUTPUT;
run;

/* Variable Selection with Sig Level (With chosen Logged Variables) */

*Forward Selection;
proc glmselect data = Data232;
model logSalePrice = logLotArea logYearBuilt logYearRemodAdd logOverallQual logOverallCond/
selection = Forward(stop = CV SLE = .2) stats = adjrsq;
run;

*Backward Selection;
proc glmselect data = Data232;
model logSalePrice = logLotArea logYearBuilt logYearRemodAdd logOverallQual logOverallCond/
selection = Backward(stop  = CV SLS = .2) stats = adjrsq;
run;

*Stepwise Selection;
```

```
proc glmselect data = Data232;
model logSalePrice = logLotArea logYearBuilt logYearRemodAdd logOverallQual logOverallCond/
selection = stepwise (stop = CV) stats = adjrsq;
score data = Data232 out = MYOUTPUT;
run;

/* Stepwise Selection with Sig Level (Custom) */
proc glmselect data = Data232;
model SalePrice = TotRmsAbvGrd GarageCars GarageArea OverallQual YearBuilt YearRemodAdd
TotalBsmtSF GrLivArea FullBath/
selection = stepwise (stop = CV) stats = adjrsq;
run;

/* Variable Selection with Sig Level (With Class function and Catagorical Variables) */
*Forward Selection;
proc glmselect data = Data1;
class MSZoning Street Alley LotShape LandContour Utilities LotConfig LandSlope Neighborhood
Condition1 Condition2 BldgType HouseStyle
RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond
Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1
BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu
GarageType GarageFinish GarageQual GarageCond PavedDrive
PoolQC Fence MiscFeature SaleType SaleCondition;
model SalePrice = TotRmsAbvGrd Fireplaces GarageYrBlt GarageCars GarageArea WoodDeckSF
OpenPorchSF EnclosedPorch ScreenPorch
PoolArea MiscVal MoSold YrSold MSSubClass LotArea OverallQual OverallCond YearBuilt YearRemodAdd
MasVnrArea BsmtFinSF1 BsmtFinSF2
BsmtUnfSF TotalBsmtSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath
BedroomAbvGr KitchenAbvGr
MSZoning Street Alley LotShape LandContour Utilities/
selection = Forward(stop = CV SLE = .2) stats = adjrsq;
```

```
run;


*Backward Selection (Hightes Avg R2 & CV PRESS);

proc glmselect data = Data1;

class MSZoning Street Alley LotShape LandContour Utilities LotConfig LandSlope Neighborhood
Condition1 Condition2 BldgType HouseStyle

RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond
Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1

BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu
GarageType GarageFinish GarageQual GarageCond PavedDrive

PoolQC Fence MiscFeature SaleType SaleCondition;;

model SalePrice = TotRmsAbvGrd Fireplaces GarageYrBlt GarageCars GarageArea WoodDeckSF
OpenPorchSF EnclosedPorch ScreenPorch

PoolArea MiscVal MoSold YrSold MSSubClass LotArea OverallQual OverallCond YearBuilt YearRemodAdd
MasVnrArea BsmtFinSF1 BsmtFinSF2

BsmtUnfSF TotalBsmtSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath
BedroomAbvGr KitchenAbvGr

MSZoning Street Alley LotShape LandContour Utilities/

selection = Backward(stop  = CV SLS = .2) stats = adjrsq;

run;


*Stepwise Selection;

proc glmselect data = Data1;

class MSZoning Street Alley LotShape LandContour Utilities LotConfig LandSlope Neighborhood
Condition1 Condition2 BldgType HouseStyle

RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond
Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1

BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu
GarageType GarageFinish GarageQual GarageCond PavedDrive

PoolQC Fence MiscFeature SaleType SaleCondition;;

model SalePrice = TotRmsAbvGrd Fireplaces GarageYrBlt GarageCars GarageArea WoodDeckSF
OpenPorchSF EnclosedPorch ScreenPorch
```

PoolArea MiscVal MoSold YrSold MSSubClass LotArea OverallQual OverallCond YearBuilt YearRemodAdd MasVnrArea BsmtFinSF1 BsmtFinSF2

BsmtUnfSF TotalBsmtSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr

MSZoning Street Alley LotShape LandContour Utilities/

selection = stepwise (stop = CV) stats = adjrsq;

run;