Assignment #3:
Oscar Hernandez

Introduction:

The purpose of this report is to provide results from the fitting of regression models that use the Ames Housing data. The housing data set contains 2930 observations spread across 83 variables which are either continuous, discrete, ordinal or nominal. The variables are representative of housing characteristics such as bathroom square footage, size of garage, and exterior quality. The report will summarize all aspects of the model building process. Specifically, the report will cover how the sample population data was determined, results from exploratory data analysis and a discussion of the (multiple) linear regression models that were built. The final goal is to determine which model is most appropriate to predict the value of a "typical" home.

Data:

As mentioned, it is necessary to define a "typical" home. The "typical" home will represent our population of interest and will ultimately become our sample population data. Therefore, we will create "drop conditions" from the original housing data. The "drop conditions" that have been chosen to represent our "typical" home will be placed in a waterfall which is displayed in Figure 1.

Drop Condition	Observations
01: Not Residential	168
02: Not SFR	440
03: Lot Area Greater than 20,000 Square Feet	81
04: Poor House Condition	11
05: Built Pre-1950	533
06: Poor Exterior Condition	152
07: Does not have all utilities	1
08: Home Non-Functional	5
09: Non-Normal Sale	267
10: Sale Price Greater than $400,000	26
99: Eligible Sample	1246

Figure 1: Drop Condition Waterfall

The "drop conditions" listed above represent items that are not representative of a "typical home" in Ames, Iowa. For example, the decision was made to exclude homes priced over $400,000 which resulted in 26 non-eligible observations. Furthermore, it makes sense to remove observations where the property sale took place in a non-residential zone (168 observations) and the property was not a single-family home which is indicated by the "SFR" tag (440 observations). The waterfall created provides us with a sample of 1246 eligible observations from the original 2930 in the housing data.

Prior to moving further along in the model building process, the decision was made to remove observations in which a missing value was present for any variable. The "na.omit" function was applied to our dataframe which reduced the 1246 observations to 973.

Exploratory Data Analysis

As part of the model building process, promising predictor variables must be chosen. An appropriate start would be to check the correlation between a select number of predictor variables (chosen at random) and our response variable, SalePrice. Figure 2 is a table of correlation values between a variety of predictor variables.

Predictor Variable	Correlation
LotArea	0.45
GrLivArea	0.80
GarageArea	0.70
LotFrontage	0.35
EnclosedPorch	-0.06
OpenPorchSF	0.38
PoolArea	0.04
TotRmsAbvGrd	0.64

Figure 2: Table of Correlation Values for SalePrice

For our exploratory data analysis discussion, the two promising predictor variables will be LotArea and GrLivArea. They were chosen based on their correlation with SalePrice (higher than 0.4) and because they are variables that are measured in square footage, which we feel are important for our analysis. First, boxplots were created for each of the predictor variables to check for any outliers, skewness and overall spread of the distribution. Figure 2 displays the boxplots that were created.

From Figure 3, we can see that there are several outliers for both variables which are heavily concentrated on the top half of the plots. The presence of these outliers can be problematic since they may affect the results of our model(s). Furthermore, it would appear that both variables are positively-skewed.

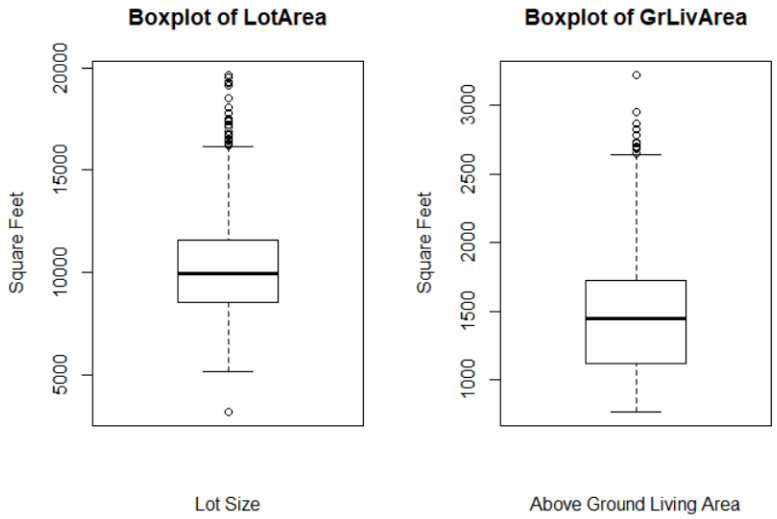

Figure 3: Side by Side Boxplot of LotArea and GrLivArea

Next, it is necessary to create a scatterplot between the predictor variables and the response variable SalePrice. The scatterplots will provide further evidence of a linear relationship between the two variables. Figure 4 displays the scatterplot of SalePrice and LotArea.

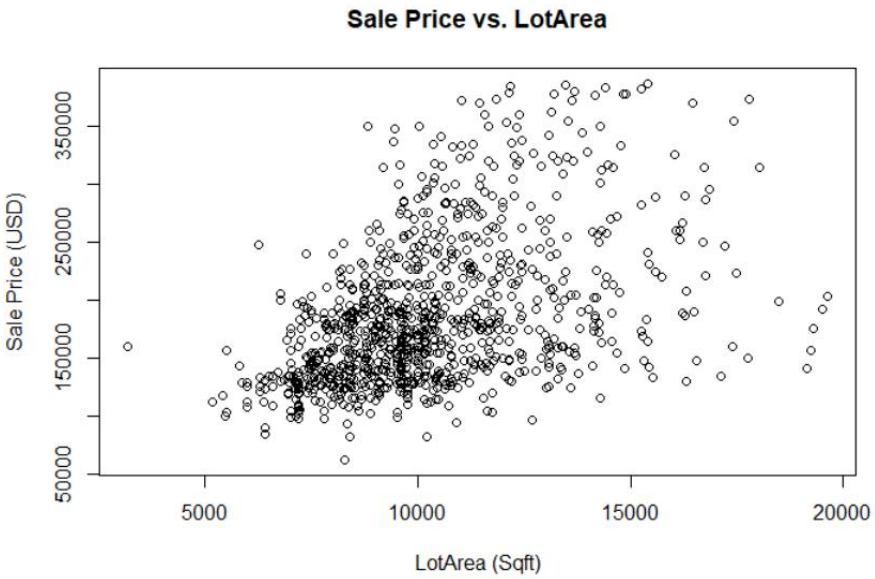

Figure 4: Scatterplot of SalePrice vs. LotArea

From Figure 4, we can see limited evidence of a linear relationship between SalePrice and LotArea. Specifically, the variation in the data becomes greater as LotArea increases past 10,000 square feet. It may be necessary to transform SalePrice to see if that offers better results. Overall, our initial exploration of these two variables leads us to think that LotArea isn't a strong predictor of SalePrice on its own.

Figure 5 is a scatterplot of SalePrice and GrLivArea. From the plot, we can see clear evidence of a linear relationship between the two variables. Specifically, it would appear that there is a strong correlation for observations between 500 and 1700 square feet and $100,000 and $200,000 sale price. Overall, there is some evidence here that leads us to believe that GrLivArea will be an appropriate predictor variable for SalePrice.

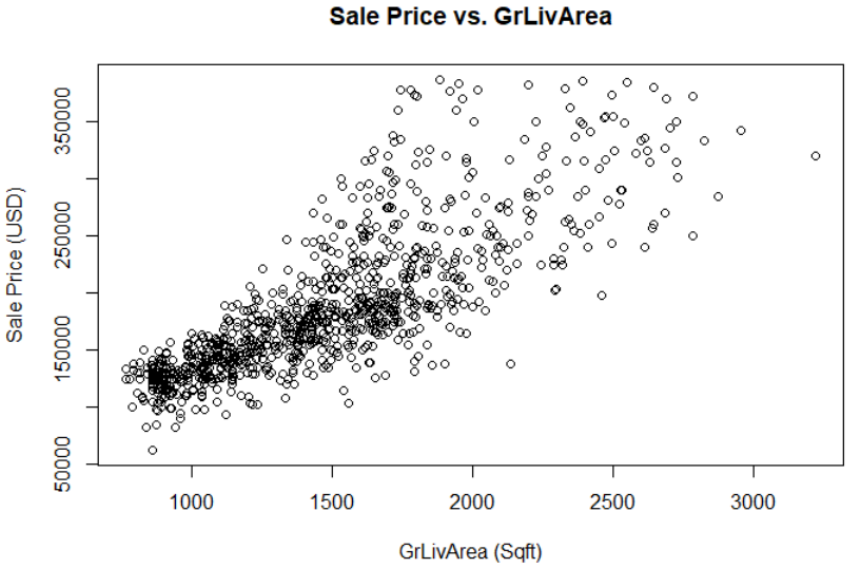

Figure 5: Scatterplot of SalePrice vs. GrLivArea

Simple Linear Regression Models

Now that an exploratory data analysis of our predictor variables is complete, we can fit our simple linear regression models. Using the "lm()" function in R, we will create LotArea_model, which is our first simple linear regression model. The equation for the LotArea_model is y = 68690.06 + 11.5LotArea. This model uses SalePrice as the response variable and LotArea as the predictor variable. Our second linear regression will be called GrLivArea_model which keeps SalePrice as the response

variable and uses GrLivArea as the predictor variable. The equation for GrLivArea_model is y = 19504.58 + 112.92GrLivArea.

Figure 6 displays a side by side comparison of the key outputs from each model. From the table, we can see that each coefficient estimate is statistically significant as indicated by their corresponding p-values. An interesting part of the output is the significant difference in the R^2/Adj. R^2 for each model.

	SalePrice			SalePrice		
	B	*CI*	*p*	*B*	*CI*	*p*
(Intercept)	68690.06	53301.21 – 84078.92	<.001	19504.58	11270.23 – 27738.93	<.001
LotArea	11.50	10.05 – 12.96	<.001			
GrLivArea				112.92	107.60 – 118.23	<.001
Observations		973			973	
R^2 / adj. R^2		.199 / .198			.642 / .641	

Figure 6: Table of Model Outputs

LotArea_model (LotArea Predictor Variable)

The next step in the model building process is to assess the model fit. The assessment will review normality and homoscedasticity using a Q-Q Plot and scatterplot with regards to the residuals from LotArea_model. Figure 7 shows that the residuals from LotArea_model are not necessarily normally distributed given that the data points are not tightly fit on the line.

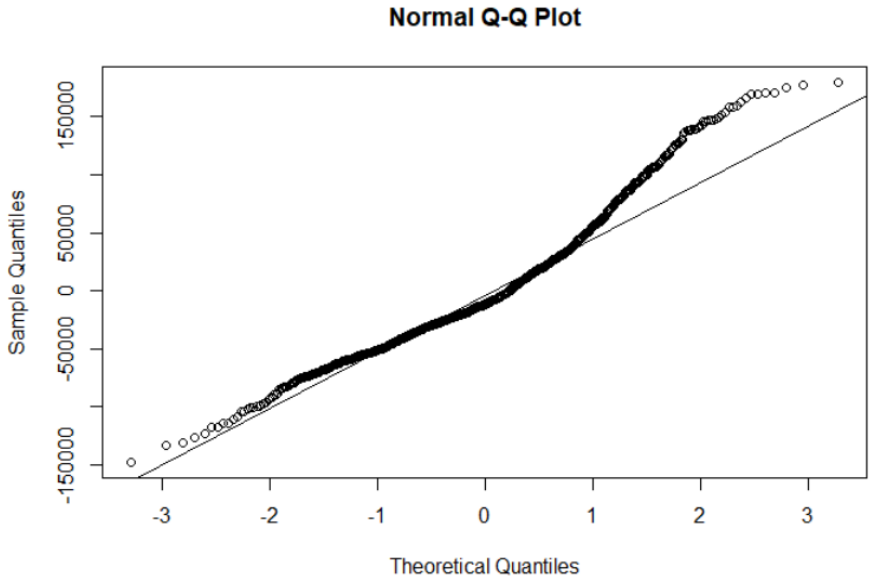

Figure 7: Q-Q Plot LotArea_model Residuals

Next, we must view the relationship between LotArea_model residuals and the LotArea variable. Figure 8 displays the scatterplot that is necessary to determine the level of homoscedasticity present. The plot provides limited evidence to support that the homoscedasticity assumption is present for LotArea_Model. The plots in Figure 8 does not show constant error variance and appears to resemble a cone shape, which is a typical sign of heteroscedasticity. In terms of goodness-of-fit, we have evidence to support that the normality and homoscedasticity assumptions have been violated based on the plots. Furthermore, the R-Squared for LotArea_model is 0.2, which means that the LotArea predictor variable does not explain a large proportion of the variability of SalePrice. Given that our statistical assumptions appear to have been violated it, would not be appropriate to conduct a t-test for our predictor variable. Therefore, it can be stated that this model has a poor goodness-of-fit.

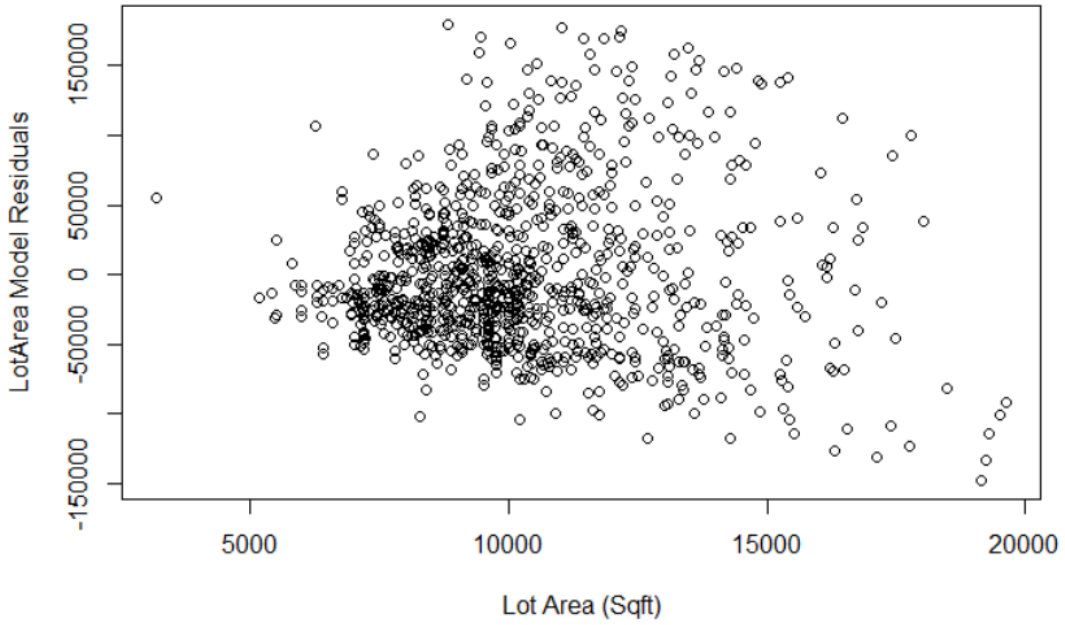

Figure 8: Scatterplot of LotArea_model residuals and LotArea

GrLivArea_model (GrLivArea Predictor Variable)

Figure 9 and 10 were created to explore the same assumptions for the GrLivArea_model. From Figure 9, we can see the data points further along the line deviate in a more extreme fashion than in the LotArea_model. As such, using Figure 9, it can be stated there is evidence to overturn the assumption of normality for the GrLivArea_model.

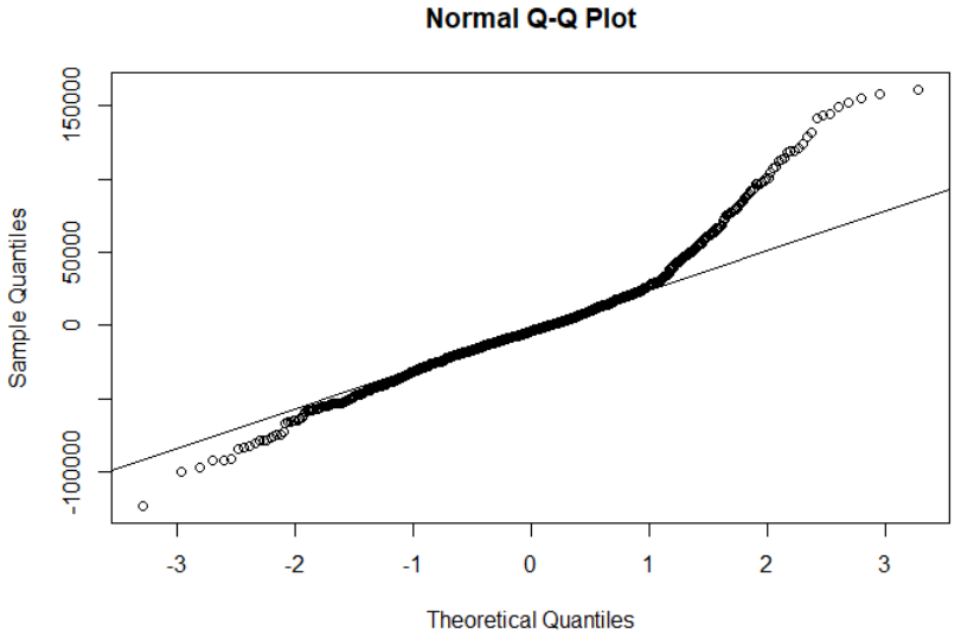

Figure 9: Q-Q Plot GrLivArea_model Residuals

Figure 10 does not offer strong evidence of the homoscedasticity assumption being valid for the GrLivArea_model. The error variance is not constant throughout the plot and does appear to be cone shaped. Moreover, the R-Squared for GrLivArea_model is 0.642 which indicates that more than half of the proportion of variability of SalePrice is explained by the GrLivArea. This a far better result compared to the LivArea variable. Overall, in terms of goodness-of-fit, the GrLivArea_model is more promising than

the LotArea_model due to its higher R-squared but it's important to note that our statistical assumptions have been violated based on the plot analysis.

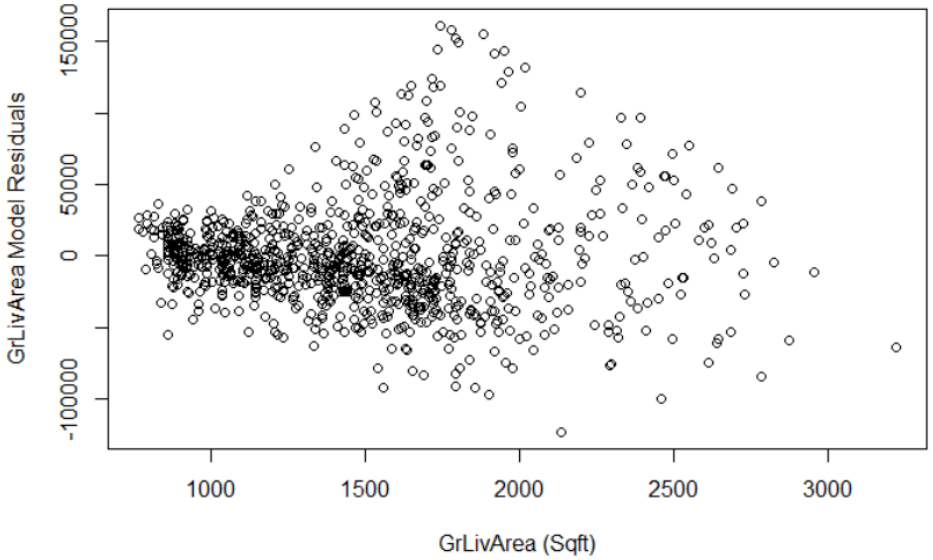

Figure 10: Scatterplot of GrLivArea_model residuals and GrLivArea

Multiple Linear Regression Model

As part of the model building process, it makes sense to build a multiple linear regression model that includes the effect of both LotArea and GrLivArea on the response variable SalesPrice (Note: LotArea and GrLivArea have a correlation coefficient of 0.37). The same programming function was used to create our multiple linear regression model – MLR_model. The equation for this model is y = -12431.42 + 4.4LotArea + 103.95GrLivArea. Figure 11 displays the relevant output for this new model. It would appear that the intercept coefficient is not as statistically significant as prior models based on its p-value but that's not overly concerning. The two predictor variables are statistically significant based on their respective p-values. Furthermore, we see that this model has a higher adjusted R-Squared compared to the GrLivArea_model which was the simple linear regression model that had the highest adjusted R-Squared from our previous section discussion.

	SalePrice		
	B	*CI*	*p*
(Intercept)	-12431.42	-23258.38 – -1604.46	.024
LotArea	4.40	3.38 – 5.41	<.001
GrLivArea	103.95	98.42 – 109.48	<.001
Observations		973	
R^2 / adj. R^2		.667 / .666	

Figure 11: Table of Model Outputs

To assess the goodness-of-fit of the MLR_model, the first step that will be taken is using a Q-Q Plot with its residuals to check for normality assumption. From Figure 12, we can see that the normality assumption appears to be violated for the residuals as indicated by the data points not tightly fitting on the line.

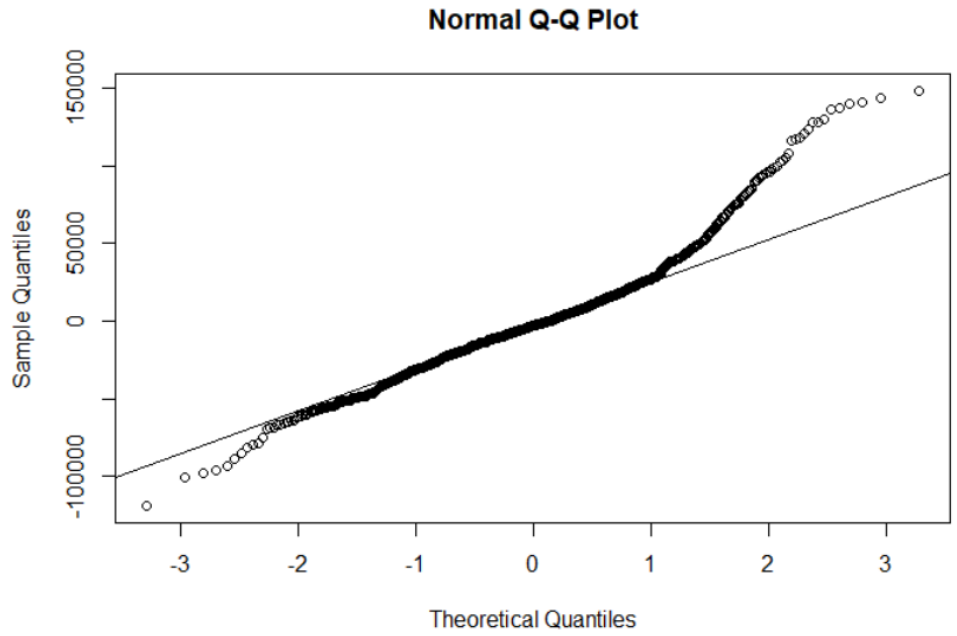

Figure 12: Q-Q Plot MLR_model Residuals

Secondly, Figure 13 is a scatterplot of MLR_model residuals versus MLR_model fitted values. In Figure 13, we are looking to see if we can predict non-zero values for residuals based on the fitted values. Based on the variation in the data points, it would be difficult to predict those non-zero values. This means that the homoscedasticity assumption is present for the MLR_model.

Overall, the MLR_model does fit better than the simple linear regression models from earlier in the report. More predictor variables in a model don't always mean that the model will fit better. However, in this case, we saw more evidence of homoscedastic in our multiple linear regression model which is a desired quality in terms of goodness-of-fit. Another criteria used in assessing fit was adjusted R-Squared. The MLR_model had a higher adjusted R-Squared (0.67) than either one of the previous simple linear regression models. In terms of normality, all the models appear to violate this assumption. Given that normality appears to be violated, it is best not to use a t-statistic/F-statistic to further assess fitness. Assuming normality wasn't violated, all the models created have a statistically significant t-statistic and F-statistic.

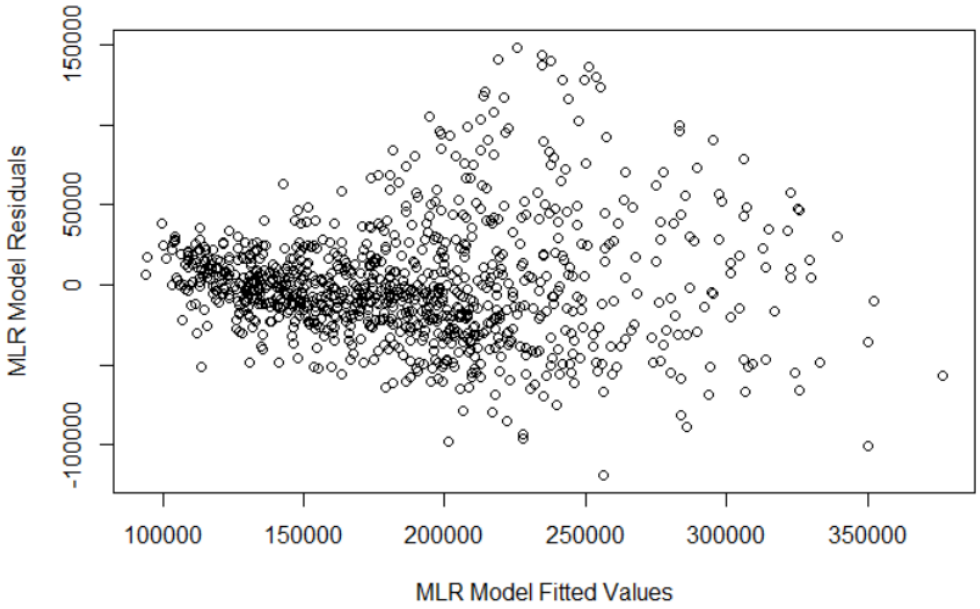

Figure 13: Scatterplot of MLR_model Residuals and Fitted Values

Neighborhood Accuracy

In the Neighborhood Accuracy section of this report, we will explore neighborhood predictive accuracy from a variety of different angles. The residuals by neighborhood from the MLR_model will be discussed. Furthermore, we will explore the relationship between the mean Mean Absolute Error (MAE) and mean SalePrice/SQFT for each neighborhood. Finally, this section will introduce the concept of indicator variables and their impact on MAE.

Residuals by Neighborhood

In this section of the report, we will utilize our multiple linear regression model – MLR_model – to assess neighborhood accuracy. Specifically, we are looking at the relationship between the residuals of MLR_model for each neighborhood to assess which neighborhoods are better fit by our model. As a reminder, residuals measure the difference between observed values and predicted values with regards to a response variable. Therefore, an ideal residual would be 0 which would indicate that there is no difference in the observed value and the predicted value of an ith observation.

Figure 14 is a boxplot of the residuals by neighborhood. Keeping in mind our concept of an ideal residual, to determine if a neighborhood is better fit by the model, we would want to see the neighborhood's residual median close to 0, limited residual outliers and a small Interquartile Range (IQR). Therefore, given the criteria we just described, the boxplot provides evidence to support that the Crawford, Mitchell, Northwest Ames, Old Town, Sawyer and Sawyer West neighborhoods are better fit by the model in comparison to all the neighborhoods.

Furthermore, it can be said that Northridge Heights, Somerset and Stone Brook neighborhoods are consistently underpriced. The reason why this can be said is because the median residual for these neighborhoods is close to $50,000 (the highest compared to all other neighborhoods). Referring to the definition of a residual, this means that our model is underpricing these neighborhoods because there is a large positive gap. On the contrary, using the same logic, we can say that the Clear Creek and Iowa Dot and Rail Road neighborhoods are being overpriced by our model. Both those neighborhoods have the median residuals close to or surpassing -$50,000 – large negative gaps.

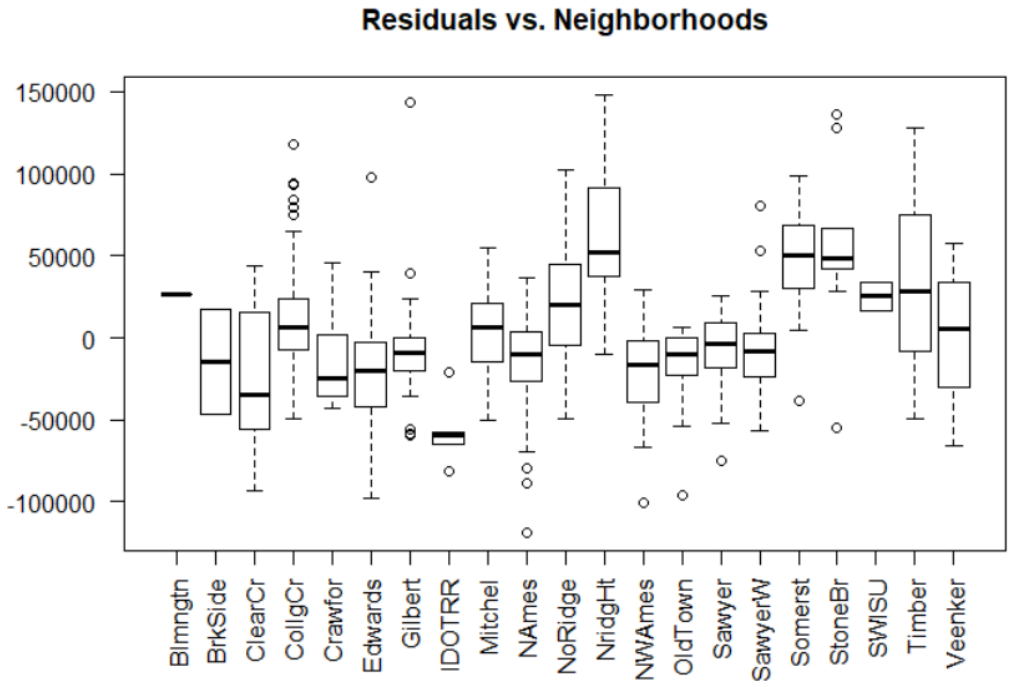

Figure 14: Boxplot Residuals by Neighborhood

Mean Absolute Error and SalePrice/Square Foot

Another part of our neighborhood accuracy section is reviewing the relationship between MAE and the SalePrice/Square Foot by neighborhood. To review this relationship, the mean MAE and mean SalePrice/Square Foot for each neighborhood was calculated (included in the Appendix). The R script

used for these computations is in the Code section of this report. Figure 15 is a scatter plot of the mean MAE and mean SalePrice/Square Foot for each neighborhood. The plot does not show any meaningful relationship between these two quantities. For example, the Old Town neighborhood has a mean SalePrice/SQFT of $100 and a mean MAE of $20,485. The North Ames neighborhood has a mean SalePrice/SQFT of $81 and a mean MAE of $19,870. These two neighborhoods have similar mean MAE but very different mean SalePrice/SQFT. Other similar examples exist in Figure 15 which highlight the conclusion that there isn't any meaningful relationship between these two quantities.

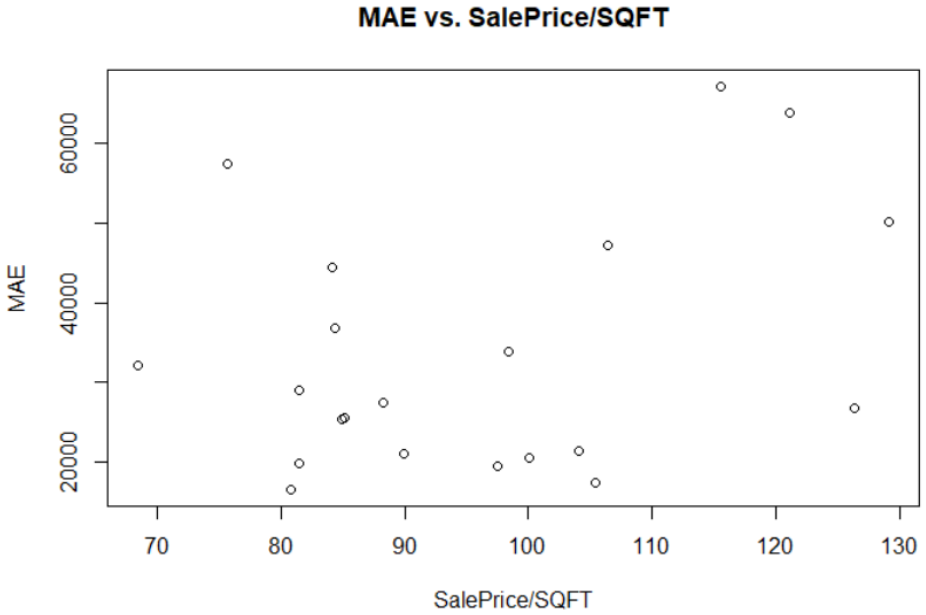

Figure 15: Plot of MAE vs. SalePrice/SQFT

Indicator Variables

The last section of Neighborhood accuracy involves refitting MLR_model with indicator variables. Our indicator variables will be created by grouping neighborhoods based on SalePrice/SQFT. The R script used to create these indicator variables is in the Code section of this report. The four groups are defined by having SalePrice/SQFT in one of the following ranges:

- Greater than or equal to $40 and less than $80
- Greater than or equal to $80 and less than $85
- Greater than or equal to $85 and less than $100
- Greater than or equal to $100

Furthermore, the base category is defined by having SalePrice/SQFT less than $40. The MAE of our original MLR_model is $26,242 and the MAE of our refitted MLR_model is $25,029. Therefore, we can conclude that our refitted model fits better based on the MAE. A lower MAE is desired since it signals that the model has higher predictive accuracy (i.e. less error). Computing the ratios between the two MAEs results in 1.048. This means that the original MLR_model has approximately 5% more error than the refitted model.

SalePrice versus Log SalePrice as the Response

The final section of this report will include the introduction of two new multiple linear regression models, the transformation of the response variable and a discussion of which model fits better.

SalePrice Model

The SalePrice Model was created using four continuous predictor variables – LotArea, GrLivArea, LotFrontage and GarageArea and one discrete predictor variable – TotRmsAbvGrd. The equation for this model is y = -23481.7 + 3.06LotArea + 82.73GrLivArea – 13.99LotFrontage + 128.04GarageArea –

1048.25TotRmsAbvGrd. The R script used to create this model is in the Code section of this report. Figure 16 displays the relevant summary data for the SalePrice Model.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-23481.7	6803.787	-3.451	0.000582
LotArea	3.06	0.475	6.441	1.86E-10
GrLivArea	82.726	4.21	19.649	< 2e-16
LotFrontage	-13.986	69.488	-0.201	0.840523
GarageArea	128.035	6.711	19.077	< 2e-16
TotRmsAbvGrd	-1048.25	1337.017	-0.784	0.433221

Figure 16: Output for SalePrice Model

Furthermore, the R-Squared for the SalePrice Model is 0.76. This summary data can be retrieved based on the R script located in the Code section of this report. We interpret this model by looking at the coefficient estimates. For example, using the LotArea as the predictor variable, we can say that a one unit increase in LotArea will result in $3.06 increase in the SalePrice of a home keeping all the other predictor variables constant. It's important to remember that predictor variables may have different units of measurement so just because a particular predictor variable has a larger coefficient does not mean that it is the best predictor. Also, with regards to interpreting our results, we can state that the LotFrontage and TotRmsAbvGrd predictor variables are not statistically significant based on their respective p-values. This can be interpreted as saying that a change in those predictor variables will not be associated with a change in the response variable. Lastly, using R-Squared we can get an idea of the proportion of variability of SalePrice that is explained by our predictor variables, which in this case is more than 3/4ths. It would be more useful comparing it to another model that has different predictor variables.

Log SalePrice Model

In the Log SalePrice Model the only change that was made was that a log transformation of SalePrice was completed prior to fitting the model. The equation for Log SalePrice Model is y = 11.02 + 1.60E-05 LotArea + 4.18E-04GrLivArea - 2.73E-04LotFrontage + 6.12E-04GarageArea – 7.43TotRmsAbvGrd. Figure 17 displays the relevant summary data for the Log SalePrice Model.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.10E+01	3.27E-02	337.189	< 2e-16
LotArea	1.60E-05	2.28E-06	7.016	4.28E-12
GrLivArea	4.18E-04	2.02E-05	20.678	< 2e-16
LotFrontage	-2.73E-04	3.34E-04	-0.817	0.414
GarageArea	6.12E-04	3.22E-05	18.982	< 2e-16
TotRmsAbvGrd	-7.43E-04	6.42E-03	-0.116	0.908

Figure 17: Output for Log SalePrice Model

The R-Squared for the Log SalePrice Model is 0.78. In cases where the response variable has been log transformed and the predictor variables have not, it can be said that the response variable changes by (100*Coefficient Estimate) percent for a one unit increase in the predictor variable keeping the other predictor variables constant. Therefore, a one unit increase in LotArea would result in a 0.16% increase in SalePrice. Also, the LotFrontage and TotRmsAbvGrd can be interpreted in the same way they were from the original SalePrice model, in which they are not statistically significant. The R-Squared in this model shows that the predictor variables represent more than 3/4ths of the variability of the response variable.

Comparison and Discussion of Model Fits

Since we are building statistical models to predict "typical" home prices, we are concerned with predictive accuracy. Therefore, we are going to compare the Mean Square Error (MSE) and Mean Absolute Error between both models. Prior to doing that, we need to assess goodness-of-fit for each model to ensure that our model did not violate underlying probabilistic assumptions. Specifically, we will be assessing normality and homoscedasticity.

Figure 18 helps us assess the normality assumption for the SalePrice Model. As noted by the divergence towards the upper half of the line, we have some evidence that the normality assumption has been violated.

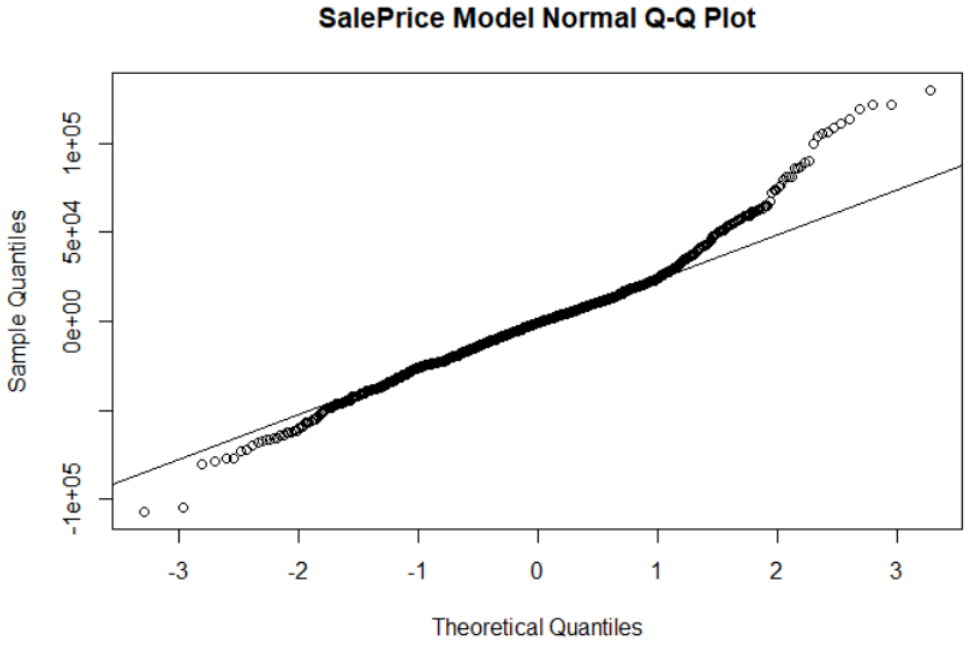

Figure 18: SalePrice Model Q-Q Plot

Furthermore, Figure 19 allows us to assess the Log SalePrice Model for normality. Based on the plot, we can see that there is divergence from the line on both ends of the line which is more pronounced when compared to the Q-Q plot of the SalePrice Model. The normality assumption appears to have been violated here as well based on the Q-Q plot output.

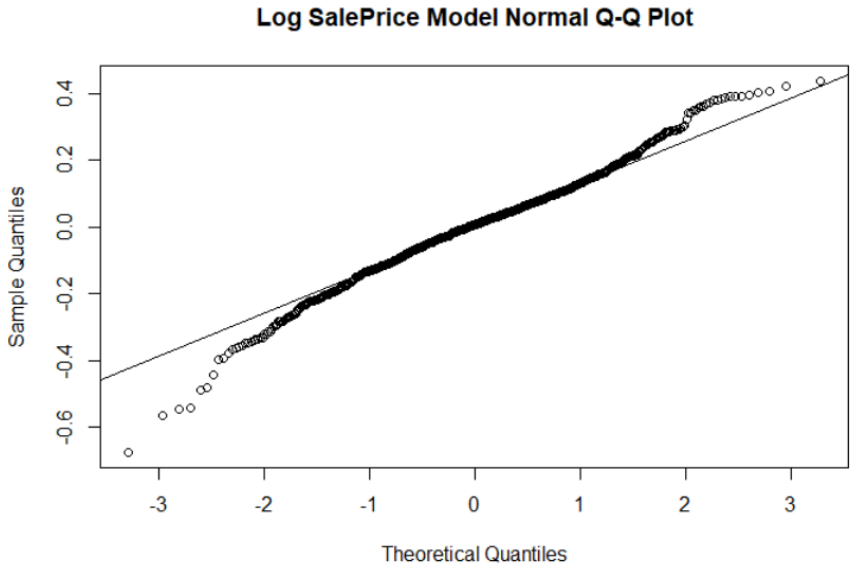

Figure 19: Log SalePrice Model Q-Q Plot

As noted earlier, GrLivArea was a statistically significant predictor variable. Therefore, to check for homoscedasticity we must plot the model residuals versus a predictor variable. This process can be repeated for all continuous predictor variables in the model. We will also plot the model residuals versus the fitted values. Figure 20 displays SalePrice residuals and GrLivArea. There is evidence to support homoscedasticity from this plot. Typically, if the error variances are not constant, a cone shape plot is usually present which is not the case here. It's important to note for GLivArea values between 1750 and 2000, there is less constant variation in the residuals.

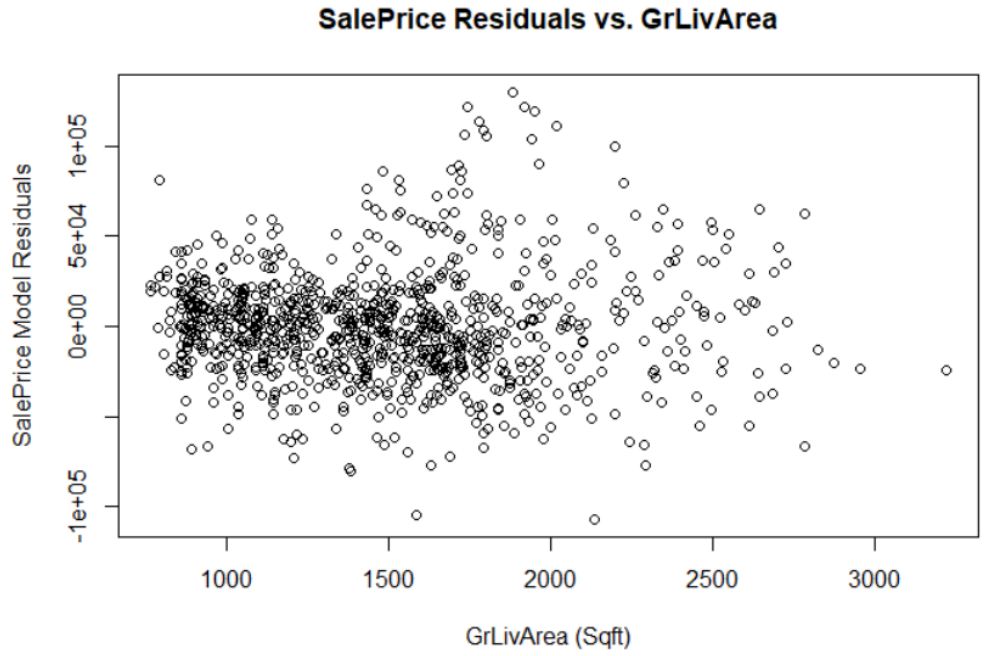

Figure 20: Plot of SalePrice Residuals versus GrLivArea

Also, to check for homoscedasticity in the SalePrice Model, we can plot the residuals versus the fitted values. Figure 21 displays this plot and it appears to show signs of heteroscedasticity based on the shape of the plot.

Figure 22: Plot of SalePrice Residuals versus SalePrice Fitted Values

Figure 23 shows the residuals versus GrLivArea for the Log SalePrice Model. Based on the shape of the plot, it does show more signs of homoscedasticity. The plot does not resemble a cone shape.

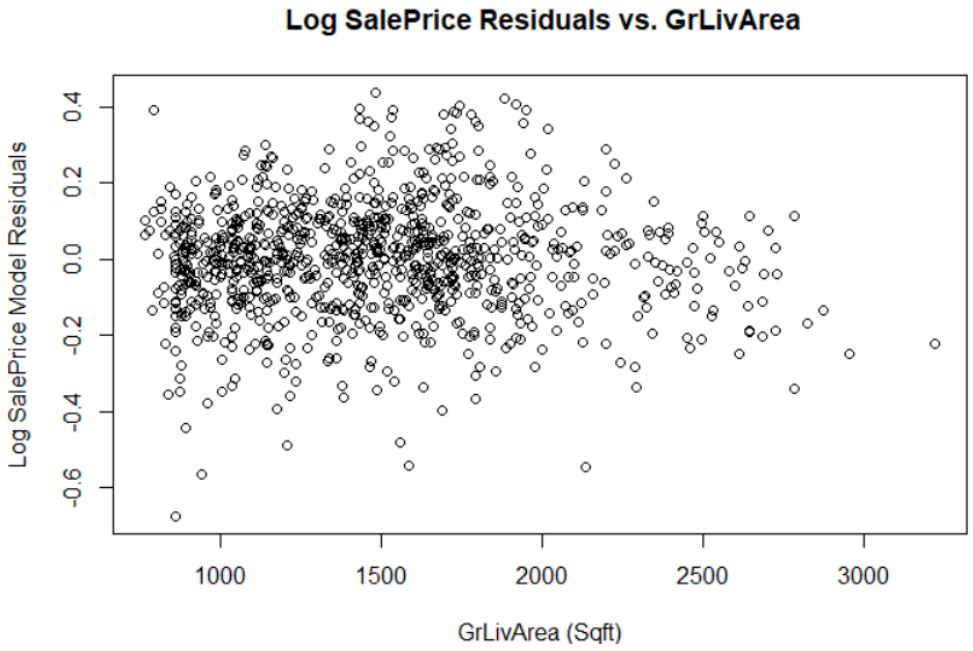

Figure 23: Plot of Log SalePrice Residuals versus GrLivArea

Figure 24 is a plot of the residuals versus the fitted values for the Log SalePrice model. This plot also shows signs of homoscedasticity as it doesn't resemble a cone shape. The error variances show some consistency across the plot but does have minor spots where the error variances are not consistent.

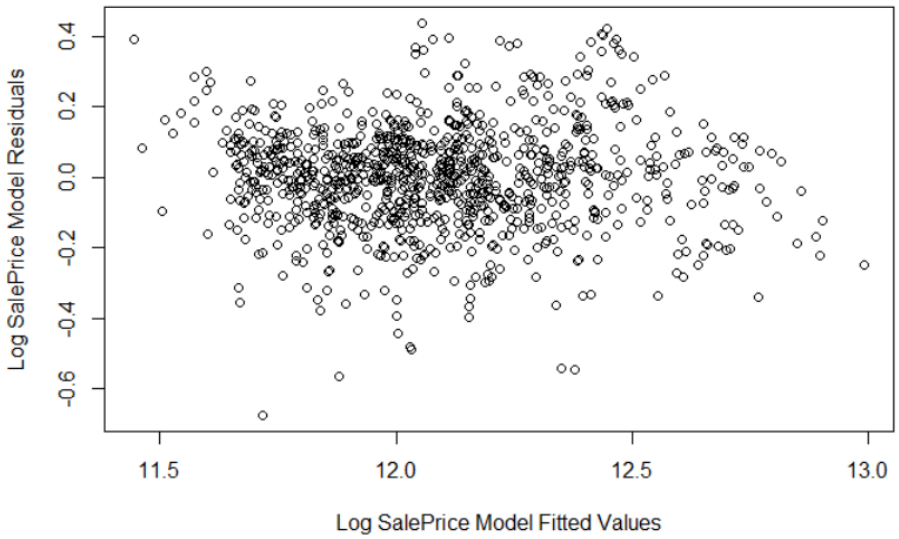

Figure 24: Plot of Log SalePrice Residuals versus Log SalePrice Fitted Values

In our discussion of goodness-of-fit, we must also consider the F-statistic and R-Squared for each model. This was computed using R script that is in the Code section of the report. The SalePrice Model F-statistic is 611.9 and is statistically significant based on its p-value. The same can be said for the Log SalePrice Model which as an F-statistic of 671.4. Therefore, regarding an Overall F-test for a regression effect, both models would reject the null hypothesis. In other words, there is enough statistical evidence to support the claim that at least one predictor variable (in either model) has an estimated coefficient that is statistically different from zero which means that each model provides some predictive information on the response variable. If we would have failed to reject the null, then we would consider discarding our model(s). Moreover, the SalePrice Model has an R-Squared of 0.76 and the Log SalePrice

Model has an R-Squared of 0.78. Based on R-Squared output, we can say that the Log SalePrice model does a slightly better job at explaining the variation in SalePrice since it has a higher R-Squared value.

The last component of this section will compare the MSE and MAE of both models. The ratios for each quantity will be defined by SalePrice Model/Log SalePrice Model. The MSE ratio is 1.03 and the MAE ratio 1.06. Based on these ratios, we can say that there is more error in the SalePrice Model. Also, it can be said that the Log SalePrice has more predictive accuracy since its MSE and MAE is lower.

Based on the evidence from the plots displayed earlier, taking into account R-Squared and also comparing the MSE/MAE ratios, we can say that the Log SalePrice model fits better. The transformation of SalePrice did improve the model fit. A log transformation can improve a model fit when a variable is negatively or positive skewed. The log transformation can pull the outlying data closer to the bulk of the data with the intent of making the variable have a normal distribution. I don't think it is necessary to consider transformations of the predictor variables. I would say that to improve the model fit, it would be more prudent to consider different predictor variables than the ones used to build these models.

Conclusion:

Overall, between LotArea and GrLivArea, it would appear that GrLivArea is a better predictor of SalePrice. The evidence outlined in this report gives reason as to why that is the case. Simply thinking about what the variables represent, it makes sense to think that the square footage of above ground living area (i.e. space where people can actually live in) would be better at predicting the value of a "typical" home as opposed to knowing the square footage of the lot where the home resides. In terms of answering our statistical problem, using both predictor variables together in a multiple linear regression model is the best approach. Lastly, in the final section of this report, we looked at the impact that log transformation has on model fit. Two different multiple linear regression models were created and through multiple pieces of evidence, it was concluded that transforming our response variable, SalePrice, resulted in a better fit model. Therefore, as we continue creating new models to predict "typical" home prices that include different or more predictor variables than used in this report, we need to consider transforming SalePrice in those new models to identify which is the "better" model.

Appendix:

Neighborhood	Mean MAE
Blmngtn	26739.92
BrkSide	32152.68
ClearCr	44475.86
CollgCr	21345.38
Crawfor	27421.83
Edwards	29045.38
Gilbert	17386.79
IDOTRR	57346.3
Mitchel	21138.44
NAmes	19870.7
NoRidge	33881.76
NridgHt	63902.69
NWAmes	25550.46
OldTown	20485.67
Sawyer	16572.07
SawyerW	19413.4
Somerst	50095.24
StoneBr	67139.23
SWISU	25302.14
Timber	47281.58
Veenker	36751.7

Neighborhood	Mean SalePrice/SFT
Blmngtn	126.29937
BrkSide	68.46142
ClearCr	84.13797
CollgCr	104.0274
Crawfor	88.24469
Edwards	81.44519
Gilbert	105.42103
IDOTRR	75.73055
Mitchel	89.94133
NAmes	81.49837
NoRidge	98.36637
NridgHt	121.05186
NWAmes	85.15438
OldTown	100.11628
Sawyer	80.76462
SawyerW	97.48759
Somerst	129.14811
StoneBr	115.5854
SWISU	84.93171
Timber	106.39821
Veenker	84.35933

Code:

```
library(sjPlot)
library(sjlabelled)
#Read in our Ames Housing Data
ames_df <- read.csv('ames_housing_data.csv', header = TRUE, stringsAsFactors = FALSE)

ames_df$dropCondition <- ifelse(!ames_df$Zoning %in% c('RH','RL','RP','RM'), '01: Not Residential',
 ifelse(ames_df$BldgType!='1Fam','02: Not SFR',
 ifelse(ames_df$LotArea>20000,'03: Lot Area Greater than 20,000 Square Feet',
 ifelse(ames_df$OverallCond<3,'04: Poor House Condition',
 ifelse(ames_df$YearBuilt<1950,'05: Built Pre-1950',
 ifelse(!ames_df$ExterCond %in% c('Ex', 'GD', 'TA', 'Fa'),'06: Poor Exterior Condition',
 ifelse(ames_df$Utilities!='AllPub','07: Does not have all utilities',
 ifelse(!ames_df$Functional %in% c('Typ', 'Min1', 'Min2', 'Mod'),'08: Home Non Functional',
 ifelse(ames_df$SaleCondition!='Normal','09: Non-Normal Sale',
 ifelse(ames_df$SalePrice>400000, '10: Sale Price Greater than $400,000',
 '99: Eligible Sample'))))))))))


#Create a table summarizing Drop Conditions
table(ames_df$dropCondition)

#Save table as 'waterfall'
waterfall <- table(ames_df$dropCondition)

#Display the table as a column matrix
as.matrix(waterfall, 11,1)

# Eliminate all observations that are not part of the eligible sample population
eligible_population <- subset(ames_df,dropCondition=='99: Eligible Sample')

# Check that all remaining observations are eligible
table(eligible_population$dropCondition)

# Create a new DataFrame that only includes variables we like
keep.vars <- c('SID','PID','LotFrontage','LotArea','LotConfig','Neighborhood',
          'HouseStyle','OverallQual','OverallCond','YearBuilt','YearRemodel','Exterior1',
          'BsmtFinSF1','BsmtFinSF2','CentralAir','GrLivArea','BsmtFullBath','BsmtHalfBath',
          'FullBath','HalfBath','BedroomAbvGr','TotRmsAbvGrd','Fireplaces','GarageCars',
          'GarageArea','WoodDeckSF','OpenPorchSF','EnclosedPorch','ThreeSsnPorch',
          'ScreenPorch','PoolArea','MoSold','YrSold','SaleCondition','SalePrice')

#Subset eligible_population DataFrame with keep.vars vector
skinny_ames_df <- eligible_population[, keep.vars]

#Delete observations with missing values
skinny_ames_df <- na.omit(skinny_ames_df)

#Table of correlation for EDA
cor(skinny_ames_df$LotArea, skinny_ames_df$SalePrice)
cor(skinny_ames_df$GrLivArea, skinny_ames_df$SalePrice)
cor(skinny_ames_df$GarageArea, skinny_ames_df$SalePrice)
```

```
cor(skinny_ames_df$LotFrontage, skinny_ames_df$SalePrice)
cor(skinny_ames_df$EnclosedPorch, skinny_ames_df$SalePrice)
cor(skinny_ames_df$OpenPorchSF, skinny_ames_df$SalePrice)
cor(skinny_ames_df$PoolArea, skinny_ames_df$SalePrice)
cor(skinny_ames_df$TotRmsAbvGrd, skinny_ames_df$SalePrice)

#Complete EDA on promising predictor variables in skinny_ames_df
#First, use boxplots to inspect distribution of two most promising variables
par(mfrow=c(1,2))
boxplot(skinny_ames_df$LotArea, main = 'Boxplot of LotArea', xlab= 'Lot Size', ylab='Square Feet')
boxplot(skinny_ames_df$GrLivArea, main = 'Boxplot of GrLivArea', xlab='Above Ground Living Area',
ylab = 'Square Feet')
par(mfrow=c(1,1))

#Next create scatterplot of SalePrice versus each promising predictor variable
plot(skinny_ames_df$LotArea, skinny_ames_df$SalePrice, ylab='Sale Price (USD)', main='Sale Price vs.
LotArea', xlab='LotArea (Sqft)')
plot(skinny_ames_df$GrLivArea, skinny_ames_df$SalePrice, ylab='Sale Price (USD)', main='Sale Price vs.
GrLivArea', xlab='GrLivArea (Sqft)')

#Fit a linear regression model using LotArea variable
LotArea_model <- lm(SalePrice ~ LotArea, data=skinny_ames_df)

#Fit a linear regression model using GrLivArea
GrLivArea_model <- lm(SalePrice ~ GrLivArea, data=skinny_ames_df)

#Display summary of LotArea_model
sjt.lm(LotArea_model)

#Display summary of GrLivArea_model
sjt.lm(GrLivArea_model)

#Display summary of both models
sjt.lm(LotArea_model, GrLivArea_model)

#Now, we need to produce some diagnostic plots to assess the goodness-of-fit of each model
#We will check the assumptions of normality and homoscedasticity with:
# 1. QQ Plot
# 2. Scatterplot of residuals versus predictor variable

#GOF plots for LotArea_model
#Plot for normaility
qqnorm(LotArea_model$residuals)
qqline(LotArea_model$residuals)
#Plot for homoscedasticity
plot(skinny_ames_df$LotArea,LotArea_model$residuals, xlab='Lot Area (Sqft)', ylab='LotArea Model
Residuals')

#GOF plots for GrLivArea_model
#Plot for normaility
qqnorm(GrLivArea_model$residuals)
qqline(GrLivArea_model$residuals)
```

```
#Plot for homoscedasticity
plot(skinny_ames_df$GrLivArea,GrLivArea_model$residuals, xlab='GrLivArea (Sqft)', ylab='GrLivArea
Model Residuals')

#Now we will built a multiple linear regrssion model using the same promising predictor variables
#Let's first look at the correlation between our predictor variables
cor(skinny_ames_df$LotArea, skinny_ames_df$GrLivArea)

#Fit the multiple linear regression model with both predictor variables
MLR_model <- lm(SalePrice ~ LotArea+GrLivArea, data=skinny_ames_df)

#GOF plots for MLR_model
#Plot for normaility
qqnorm(MLR_model$residuals)
qqline(MLR_model$residuals)
#Plot for homoscedasticity
plot(skinny_ames_df$LotArea,MLR_model$residuals, xlab='Lot Area (Sqft)', ylab='MLR Model Residuals')
plot(skinny_ames_df$GrLivArea,MLR_model$residuals, xlab='GrLivArea (Sqft)', ylab='MLR Model
Residuals')
plot(MLR_model$fitted.values,MLR_model$residuals, xlab='MLR Model Fitted Values', ylab='MLR Model
Residuals')

#Neighborhood Accuracy
boxplot(MLR_model$residuals~skinny_ames_df$Neighborhood, las=2, main='Residuals vs.
Neighborhoods')

#Add Total Square Foot variable
skinny_ames_df$TotalSqftCal<-
skinny_ames_df$BsmtFinSF1+skinny_ames_df$BsmtFinSF2+skinny_ames_df$GrLivArea
#Add Sale Price / Total Square Foot variable
skinny_ames_df$SalePricePerSquareFoot<-skinny_ames_df$SalePrice/skinny_ames_df$TotalSqftCal

#Compute the mean MAE for each neighborhood
meanMAE_MLR_model <- aggregate(abs(MLR_model$residuals),
by=list(Neighborhood=skinny_ames_df$Neighborhood), FUN=mean)
meanMAE_MLR_model

#Compute the mean Sale Price / Total Square Foot for each neighborhood
meanSPperSQFT_MLR_model <- aggregate(skinny_ames_df$SalePricePerSquareFoot,
by=list(Neighborhood=skinny_ames_df$Neighborhood), FUN=mean)
meanSPperSQFT_MLR_model

#Plot the mean MAE and the mean Sale Price / Total Square Foot for each neighborhood
plot(meanSPperSQFT_MLR_model[,2],meanMAE_MLR_model[,2], main = 'MAE vs. SalePrice/SQFT', xlab
= 'SalePrice/SQFT', ylab='MAE')

#Group the neighborhoods by Sale Price / Total Square Foot
#This will be done by using indicator variables
#There will be four groups - the base group will be neighborhoods that have a Sale Price / Total Square
Foot less than $80
skinny_ames_df$SPSF40_80 <- ifelse(skinny_ames_df$SalePricePerSquareFoot>=40 &
skinny_ames_df$SalePricePerSquareFoot<80,1,0)
```

```
skinny_ames_df$SPSF80_85 <- ifelse(skinny_ames_df$SalePricePerSquareFoot>=80 &
skinny_ames_df$SalePricePerSquareFoot<85,1,0)
skinny_ames_df$SPSF85_100 <- ifelse(skinny_ames_df$SalePricePerSquareFoot>=85 &
skinny_ames_df$SalePricePerSquareFoot<100,1,0)
skinny_ames_df$SPSF100 <- ifelse(skinny_ames_df$SalePricePerSquareFoot>=100,1,0)

#Refit MLR_model with the indicator variables
MLR_model_refit <- lm(SalePrice ~ LotArea+GrLivArea+SPSF40_80+SPSF80_85+SPSF85_100+SPSF100,
data=skinny_ames_df)

#Commpute MAE for MLR_model and MLR_model_refit
MAE_MLR_model <-mean(abs(MLR_model$residuals))
MAE_MLR_model_refit <-mean(abs(MLR_model_refit$residuals))

#We are going to fit two models using four continuous predictor variables and one discrete variable
#One model will have SalePrice as the response variable and the second model will have log(SalePrice)
as the response variable
Q5_model <- lm(SalePrice ~ LotArea+GrLivArea+LotFrontage+GarageArea+TotRmsAbvGrd,
data=skinny_ames_df)
Q5_logmodel <- lm(log(SalePrice) ~ LotArea+GrLivArea+LotFrontage+GarageArea+TotRmsAbvGrd,
data=skinny_ames_df)

#Summary of both models
summary(Q5_model)
summary(Q5_logmodel)

#Check Q5_model for normality and homoscedasticity
qqnorm(Q5_model$residuals,main = 'SalePrice Model Normal Q-Q Plot')
qqline(Q5_model$residuals)

plot(skinny_ames_df$GrLivArea,Q5_model$residuals, xlab='GrLivArea (Sqft)', ylab='SalePrice Model
Residuals', main ='SalePrice Residuals vs. GrLivArea')
plot(Q5_model$fitted.values, Q5_model$residuals, xlab='SalePrice Model Fitted Values', ylab='SalePrice
Model Residuals')

#Check Q5_logmodel for normality and homoscedasticity
qqnorm(Q5_logmodel$residuals,main = 'Log SalePrice Model Normal Q-Q Plot')
qqline(Q5_logmodel$residuals)

plot(skinny_ames_df$GrLivArea,Q5_logmodel$residuals, xlab='GrLivArea (Sqft)', ylab='Log SalePrice
Model Residuals', main =' Log SalePrice Residuals vs. GrLivArea')
plot(Q5_logmodel$fitted.values, Q5_logmodel$residuals, xlab='Log SalePrice Model Fitted Values',
ylab='Log SalePrice Model Residuals')

#Calculate MSE and MSA for each model
MSE_Q5Model <- mean(Q5_model$residuals^2)
MAE_Q5Model <- mean(abs(Q5_model$residuals))

MSE_Q5LogModel <- mean((skinny_ames_df$SalePrice-exp(Q5_logmodel$fitted.values))^2)
MAE_Q5LogModel <- mean(abs(skinny_ames_df$SalePrice-exp(Q5_logmodel$fitted.values)))

MSE_Q5Model/MSE_Q5LogModel
MAE_Q5Model/MAE_Q5LogModel
```