Assignment #3:

Valon Tika

Introduction:

The purpose of this assignment is to determine 2 variables to use for initial regression modeling using the Ames Iowa housing dataset. Section 1 will detail the sample data selection along with the drop conditions used to get to the subset of the data. Section 2 will discuss the categorial variables performance as individual predictors to the sales price by the beta. Section 3 will describe the multi-linear regression model that will use 2 continuous variable predictors along with the improvement performance analysis of the model. Section 4 will show the investigation neighborhood activity comparing it to the multi-linear regression that will be created. Section 5 will focus on taking the sales price and comparing it to the logarithmic function of the Sales Price to see changes in how the model interacts with changes to the response variable with at least 4 continuous predictor variables and a randomly chosen discrete variable.

Task:

Section 1: Sample population Definition

The data set that will be used will be defined only within the Ames, Iowa market. There are 2,930 records that were recorded as sold in Ames from 2006 to 2010. There are 82 total variables (23 ordinal, 14 discrete, and 20 continuous variables) stored within this data set.

The problem that this assignment will try to solve is by predicting the housing market within this specific set of data. The first step of this will be to do data sampling by using drop conditions and other sub-setting methods to narrow down our training set of data. The next step will be to define two

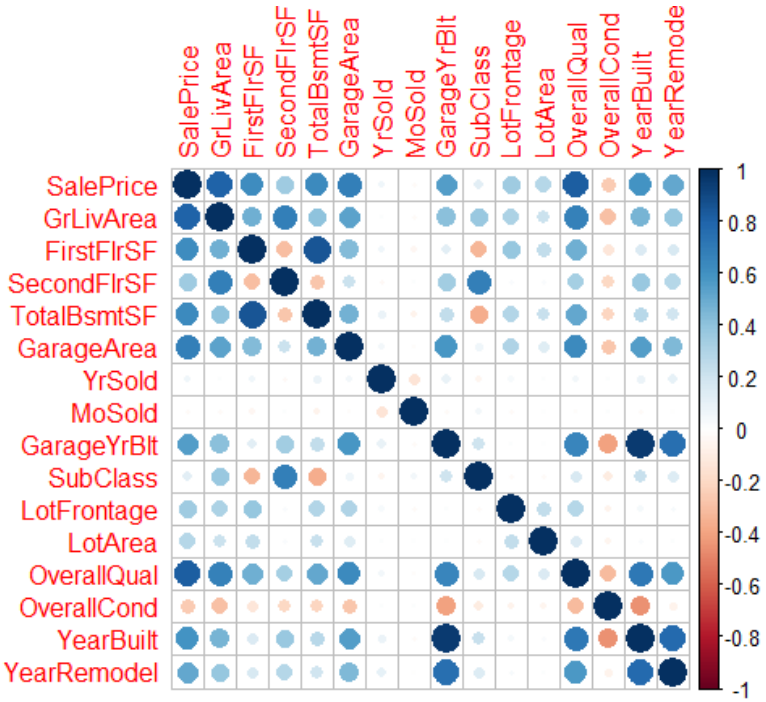
We will continue with the eligible population set from week one where we used single family homes as our subset of data to perform our modeling efforts. This is also considered as our drop conditions when filtering out the records that we don't need for our analysis. To recap, the goal will be to try and figure out which properties have a normal sale condition, paved street, year build past 1950, a basement, and a general living area greater than 800 square feet. We first concluded a set of conditions that labeled the records for what was considered as "drop conditions" for an appropriate housing subset. Here we can determine that we will start with 1,470 homes that are eligible for further analysis.

Condition	Count
Not SFR	505
Non-Normal Sale	423
Street Not Paved	6
Built Pre-1950	489
No Basement	28
LT 800 SqFt	9

Eligible Sample	1,470
Total	**2,930**

To determine the two variables to use for further regression modeling, correlation analysis will need to be done with a subset of variables. Linear regressions assume linear relationship between prediction and predictor variables. To choose the best two features to predict the sales price we will study the correlation of the sales prices with numeric features in the data and select he two for which the correlation is the highest.

After further analysis, 20 variables were selected and a correlation matrix was create to make sense of the numerical features of these variables. Upon reviewing the correlation table, the two variable the sales price had the highest correlation with the garage area and the general living area being strongly correlated relative to the sales price.



Section 2: Simple Linear Regression Modeling

The two variables that were selected (general living area and garage area), which had the highest correlation to the sales price will be used each individually in a simple linear regression against the sales price to see how they perform as predictors.

If we follow the simple linear regression line as ($Y = \beta_1 + \beta_2 X$), then below will be the simple linear regression model of general living area to sales price:

***Sales Price* = β_1 + β_2*General Living Area**

In the equation, "B_1" is the intercept and "B_2" is the slope. To generate our slope and intercept for the equation, the use of the lm() function in R will help generate the results needed to

complete the equation. The results of the simple linear regression are below for general living area:

```
Residuals:
    Min      1Q  Median      3Q     Max
-130930  -24203   -3044   16198  305551


Call:
lm(formula = house$SalePrice ~ house$GrLivArea)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     9059.915   3782.309   2.395   0.0167 *
house$GrLivArea  121.615      2.343  51.898   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 43410 on 1468 degrees of freedom
Multiple R-squared:  0.6472,   Adjusted R-squared:  0.647
F-statistic:  2693 on 1 and 1468 DF,  p-value: < 2.2e-1
```

The explicit equation for general living area would be:

Sales Price = 9059.915 + 121.6 * General Living Area

Where the intercept (B_1) would be 9059.915 and the slope (B_2) would be 121.6. This means that if you move left or right along the X-axis (general living area) by one square foot, the sale price of the house will adjust by ~$121.62 dollars. Below is a graph to display this model better.

A few things should be noted about the results of this regression output. First, the Adjusted R-Squared value is .647, which is somewhat lower than what would be preferred for this analysis. The P-value of general living area is very small (<2e-16), meaning that the model suggests that the predictor could be associated with changes in the response variable with a high confidence interval.

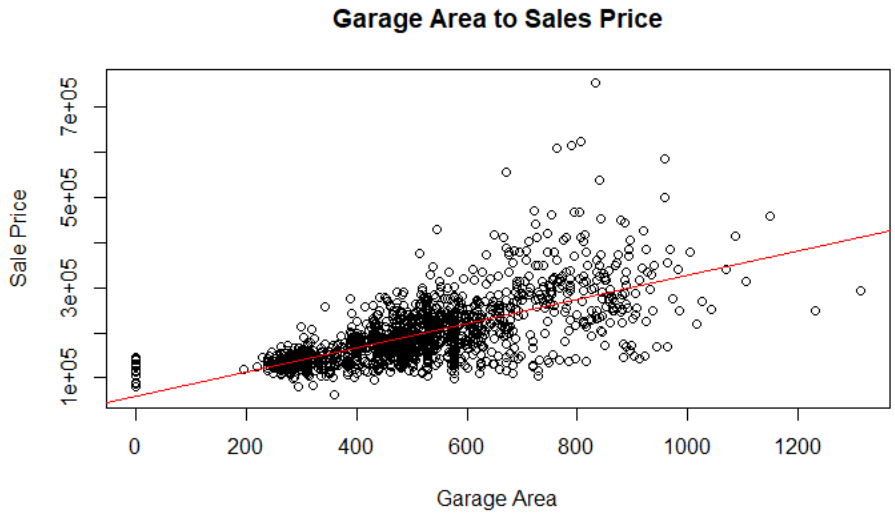
Let's perform the same kind of analysis and perform a simple linear regression and use the garage area as to see how well it can predict the sales price.

```
Call:
lm(formula = house$SalePrice ~ house$GarageArea)

Residuals:
    Min      1Q  Median      3Q     Max
-159218  -29894   -3181   21355  473244

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    58552.399   4391.974   13.33   <2e-16 ***
house$GarageArea   268.274      8.077   33.21   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 55230 on 1468 degrees of freedom
Multiple R-squared:  0.429,    Adjusted R-squared:  0.4286
F-statistic:  1103 on 1 and 1468 DF,  p-value: < 2.2e-16
```

The same simple linear regression line equation ((*Y* = β_1 + $\beta_2 X$), will be populated based on the results of the coefficient and intercept that was generated:

Sales Price = 58552.399 + 268.274 * General Living Area

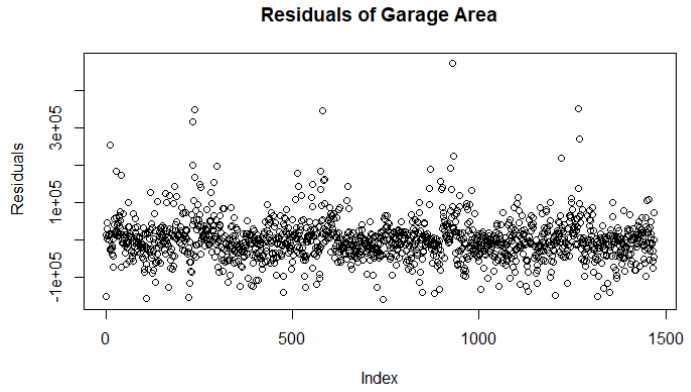
Where the intercept (B_1) would be 58552.399 and the slope (B_2) would be 268.274. This means that if you move left or right along the X-axis (general living area) by one square foot, the sale price of the house will adjust by ~$286.27 dollars. Below is a graph to display the model.

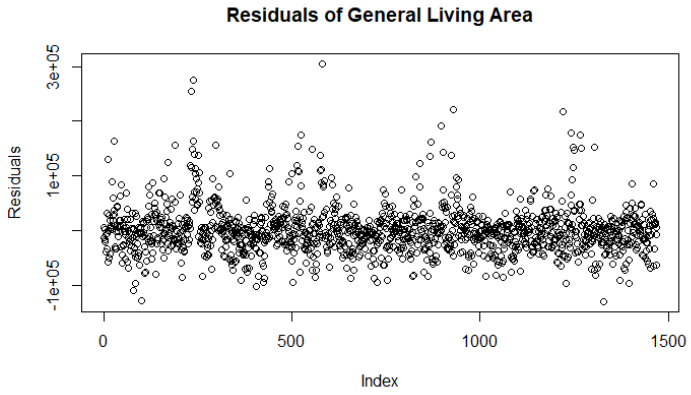


The adjusted R-Square value is much lower at .4286 compared to using general living area as a predictor. The P-value is significantly low, though and meaning that this model also suggests that the predictor could be associated with changes in the response variable.

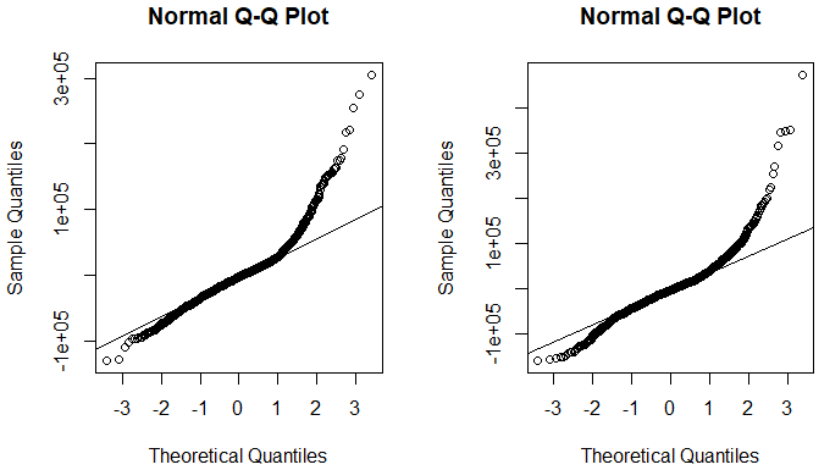
The F-Statistic tests whether the regression coefficient can be zero, or where it means that there is no relation with the predictor variable. Since the P-value is less than .05 then the predictor is important, so the coefficient cannot be 0 for both models

One of the assumptions of a linear regression is that the residuals are normally distributed with zero mean and constant variance. We will look below at scatterplot of both residuals to determine if this is the case.





Assumptions to OLS are checking if the residuals are normally distributed with a zero mean and constant variance. It appears they are both following it to a certain extent. Both residuals seem to have zero mean and about constant variance. As the residuals show, there is a symmetrical distribution of points along the scatterplot. They also both show that they're clustered around towards 0 along the y-axis, although there are points below 0, which indicates each model could be lacking a variable to better predict the sales price. Let's look at the quantile plots to get a better look at the distributions of the residuals.



The first model fit of the residuals lie more closely on the normality line then the second model fit since the second model's residuals exhibit more skewness on both tails as opposed to the first model's residuals where the distribution is more right skewed.

Section 3: Multi-Linear Regression Modeling

The next step is to move towards creating a regression model with both predictor variables (general living area and garage area) and use them together to see how they perform in predicting the sales price. A similar approach will be taken when discussing topics such as the R-Squared and see how the goodness of fit the model represents, the P-Value and residuals.

Below is the model output of the multi-linear regression output that was generated:

```
Call:
lm(formula = house$SalePrice ~ house$GrLivArea + house$GarageArea)

Residuals:
    Min      1Q  Median      3Q     Max
-135556  -21964   -2509   15803  292537

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      -20186.081   3637.700  -5.549  3.4e-08 ***
house$GrLivArea      96.115      2.420  39.714  < 2e-16 ***
house$GarageArea    133.391      6.557  20.342  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 38350 on 1467 degrees of freedom
Multiple R-squared:  0.7249,   Adjusted R-squared:  0.7245
F-statistic:  1932 on 2 and 1467 DF,  p-value: < 2.2e-16
```

From the above output, the R-Squared increased to .7249 and the adjusted R-Squared value increased to .7245. This indicates that the model has improved significantly. To interpret the impactfulness of the predictors, we will next look predictor that has a low p-value is likely to be a meaningful addition to your model because changes in the predictor's value are related to changes in the response variable. The P-Value for both variables are below .05 ($P < .05$), which signifies that both predictors together are meaningful to our model. This is an improvement from

the SLR that we performed earlier. Let's create the model for the MLR given 2 observations. The explicit model is shown below.

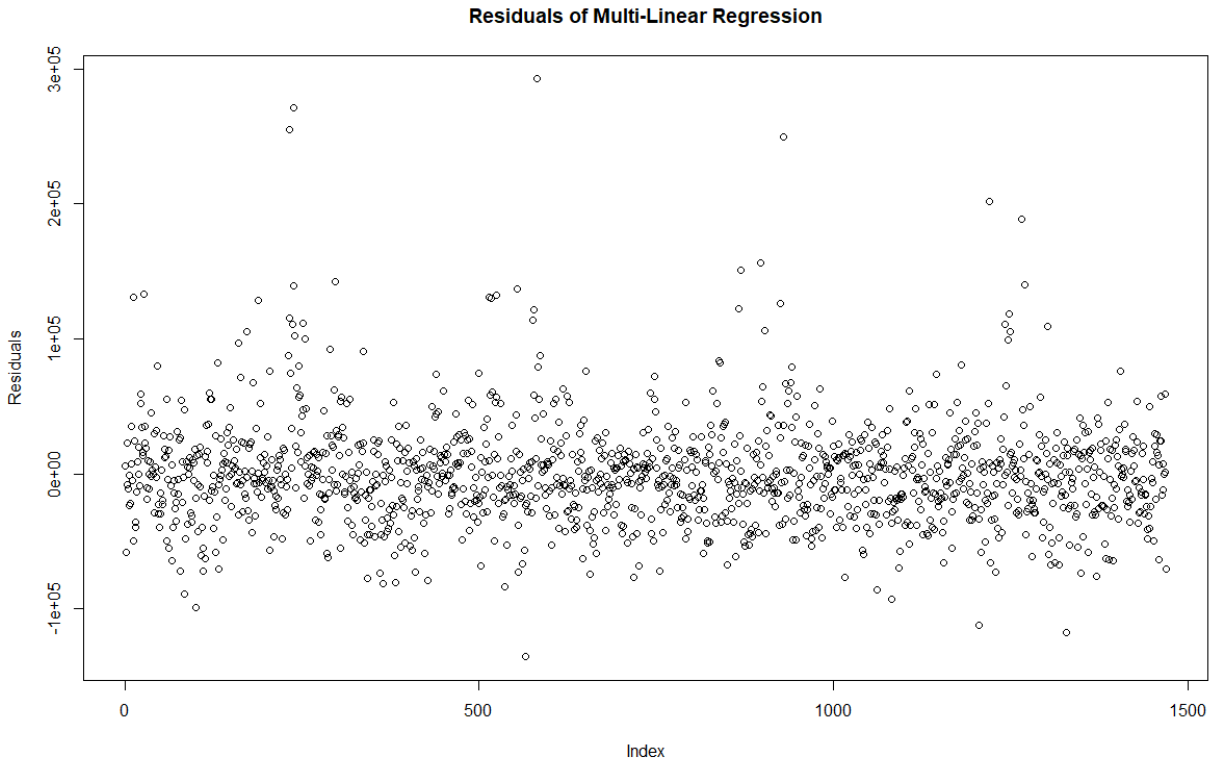
$Y = \beta_1 + \beta_2 X_1 + \beta_2 X_2$

Let's substitute our X's for both predictors and take the out our intercepts and slope to finish the explicit model for this regression:

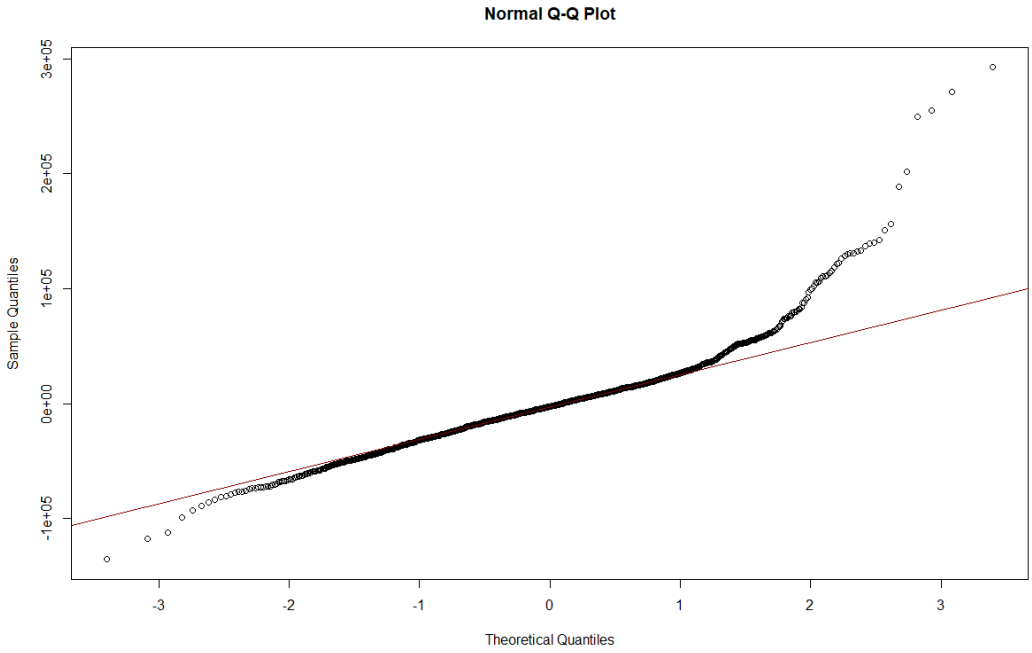
*Y = -20186.081 + 96.115 * General Living Area + 133.391 * Garage Area*

Interpreting this, we conclude that for every unit added/subtracted to general living area, the sales will adjust by 96.115. for every unit added/subtracted to garage area, the sales price will adjust by 133.391. Based on our initial drop conditions, we only are considering properties with general living area square footage greater than 800 square feet, so the model would not theoretically hit a negative sales price.

Again, we will check the assumption of residuals of a linear regression. This states that the residuals are normally distributed with zero mean and constant variance. Let's look at the residuals and how that compares after the new model to get a sense of what is occurring.

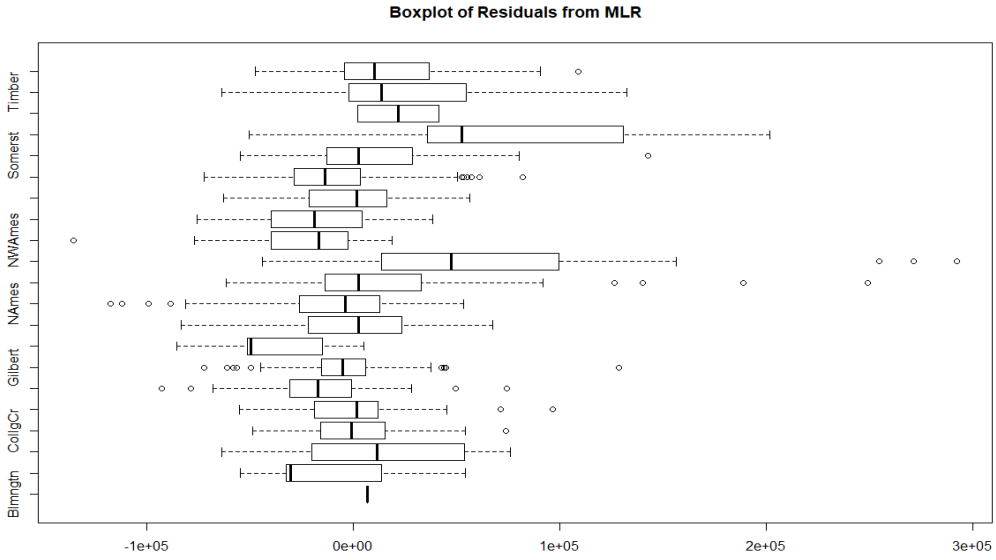


It appears the MLR is following both normal distribution with zero mean and constant variance better than the models independently. The residuals of the MLR show there is a better symmetrical distribution of points along the scatterplot. It also both shows that they're clustered around towards 0 along the y-axis a lot better. For further analysis of normality, let's observe at the normal quantile plot of the residuals.



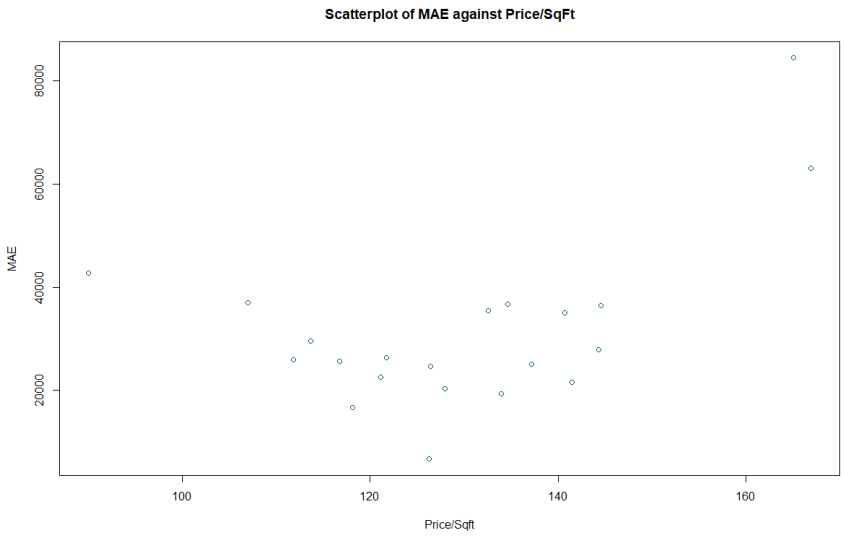
Section 4: Neighborhood Activity

In this section, we will want to take a categorial variable to better manipulate the modeling efforts to see if the model can be better represented using a different kind of variable in our analysis. Neighborhood activity has been chosen as the categorial variable. We will take our current results from the MLR to benchmark against this category. Let's first take the residuals from the MLR that was generate and create a boxplot, by neighborhood, to interpret the analysis.



The residuals of the MLR is the observed value, or the value from fit. Thus, if the residuals are greater than 0, then that means the property is underpriced by our model. If the residuals are less than 0, then that means the property is overpriced by our model. It looks like a lot of neighborhoods from Somerst and North Ames are under priced by the analysis of this box plot. Overall, the model seems pretty close to 0, though, with minimal outliers.

Let's observe the mean absolute error (MAE) and compare that to the price per square foot for each neighborhood. Let's plot that and observe how the measured value of square footage compares to the actual square footage. This can help determine if the model is measuring accurately per neighborhood by square foot, which provides a more detailed analysis of the housing market.



The graph above shows the impact of the 21 neighborhoods and the price per square foot within each neighborhood along their MAE. The MAE is higher for either low price/sqft or high price/sqft. For the average range it is almost the same.

We will now create a new categorial variable to account for the change in the price per square foot and create 6 groups (0-5). Group 1 is for prices less than $100. Group 2 is for prices between $100 and $120. Group 3 is for prices between $120 and $140. Group 4 is for prices between $140 and $160. Group 5 is for groups greater than $160. Group 0 is for properties with $0. We will use the new categorial variable as a third variable in the regression to see how the output will look.

```
Call:
lm(formula = house$SalePrice ~ house$GrLivArea + house$GarageArea +
    house$indi)

Residuals:
   Min     1Q Median     3Q    Max
```

```
-76446  -8774     84   8632 208585

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      -22892.696  18996.092  -1.205  0.22835
house$GrLivArea     129.702      1.342  96.663  < 2e-16 ***
house$GarageArea     13.345      3.700   3.607  0.00032 ***
house$indi1      -52440.619  18964.912  -2.765  0.00576 **
house$indi2      -14567.065  18923.122  -0.770  0.44154
house$indi3       15811.660  18928.384   0.835  0.40366
house$indi4       40773.057  18956.721   2.151  0.03165 *
house$indi5       98414.749  18996.761   5.181 2.52e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18900 on 1462 degrees of freedom
Multiple R-squared:  0.9334,   Adjusted R-squared:  0.9331
F-statistic:  2928 on 7 and 1462 DF,  p-value: < 2.2e-16
```

The R-Squared value is .9331, which is the most fitted model so far within the housing analysis. The P-Value is well below .05, which indicates that we can reject the null hypothesis. We also can conclude that the F-Statistic cannot be tested to see whether the regression coefficients can be 0 since the P-Values are less than .05.

Section 5: Model Comparison of Y versus log(Y)

Next, let’s use the same set of predictor variables, but change the response variable of sales price. 4 random continuous predictor variables were chosen. Those variables are lot frontage, lot area, total basement square foot, and general living area. Along those 4 variables, one random discrete variable was choses, which was the year build. We will plot the residuals to compare the two response variables to get a sense of the observations.

After the log transformation the variance in the error term reduces this transformation. This may be useful when there is heteroskedasticity in the model. Our model is not the case, but it can usually help. Next, we will observe at the regression outputs of the different response variables for further analysis.

Sales Price

```
l:
lm(formula = house$SalePrice ~ house$LotFrontage + house$LotArea +
    house$YearBuilt + house$TotalBsmtSF + house$GrLivArea)

Residuals:
    Min      1Q  Median      3Q     Max
-146468  -19303   -1152   16483  219309

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.987e+06  1.101e+05 -18.042  < 2e-16 ***
house$LotFrontage  2.664e+02  5.639e+01   4.725 2.59e-06 ***
house$LotArea      8.511e-01  1.378e-01   6.175 9.19e-10 ***
house$YearBuilt    9.844e+02  5.619e+01  17.520  < 2e-16 ***
house$TotalBsmtSF  6.429e+01  2.883e+00  22.299  < 2e-16 ***
house$GrLivArea    8.448e+01  2.493e+00  33.882  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 32560 on 1129 degrees of freedom
  (335 observations deleted due to missingness)
Multiple R-squared:  0.8235,   Adjusted R-squared:  0.8227
F-statistic:  1054 on 5 and 1129 DF,  p-value: < 2.2e-16
Call:
lm(formula = house$SalePrice ~ house$LotFrontage + house$LotArea +
    house$YearBuilt + house$TotalBsmtSF + house$GrLivArea)

Residuals:
    Min      1Q  Median      3Q     Max
-146468  -19303   -1152   16483  219309
```

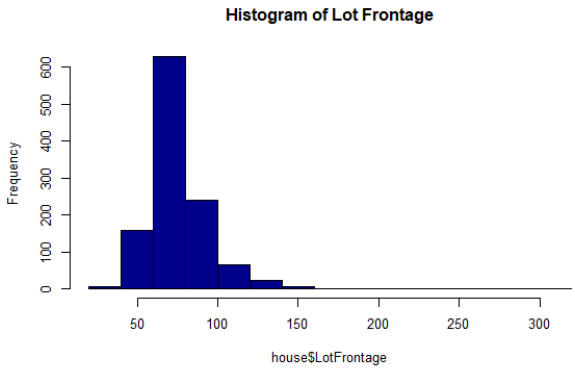
```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.987e+06  1.101e+05 -18.042  < 2e-16 ***
house$LotFrontage  2.664e+02  5.639e+01   4.725 2.59e-06 ***
house$LotArea      8.511e-01  1.378e-01   6.175 9.19e-10 ***
house$YearBuilt    9.844e+02  5.619e+01  17.520  < 2e-16 ***
house$TotalBsmtSF  6.429e+01  2.883e+00  22.299  < 2e-16 ***
house$GrLivArea    8.448e+01  2.493e+00  33.882  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 32560 on 1129 degrees of freedom
  (335 observations deleted due to missingness)
Multiple R-squared:  0.8235,   Adjusted R-squared:  0.8227
F-statistic:  1054 on 5 and 1129 DF,  p-value: < 2.2e-16
```

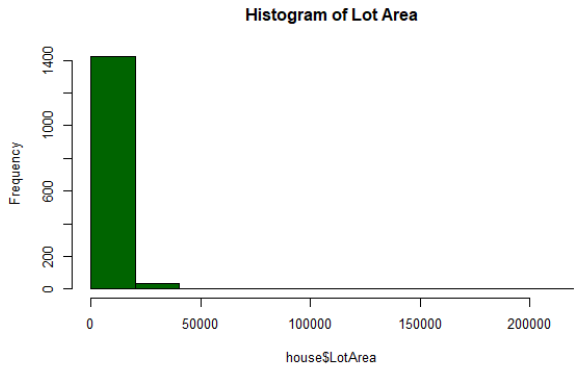
Log of Sales Price

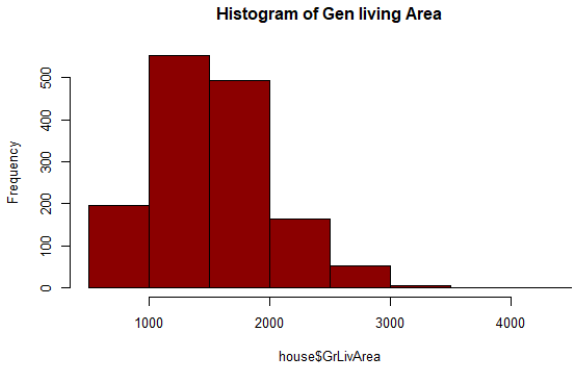
```
Call:
lm(formula = log(house$SalePrice) ~ house$LotFrontage + house$LotArea +
    house$YearBuilt + house$TotalBsmtSF + house$GrLivArea)

Residuals:
     Min       1Q   Median       3Q      Max
-0.64228 -0.06829  0.00013  0.07174  0.39944

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -4.064e-01  4.081e-01  -0.996     0.32
house$LotFrontage  1.331e-03  2.090e-04   6.371 2.74e-10 ***
house$LotArea      3.451e-06  5.108e-07   6.756 2.27e-11 ***
house$YearBuilt    5.830e-03  2.082e-04  27.994  < 2e-16 ***
house$TotalBsmtSF  2.349e-04  1.069e-05  21.985  < 2e-16 ***
house$GrLivArea    3.736e-04  9.241e-06  40.430  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1207 on 1129 degrees of freedom
  (335 observations deleted due to missingness)
Multiple R-squared:  0.8741,   Adjusted R-squared:  0.8736
F-statistic:  1568 on 5 and 1129 DF,  p-value: < 2.2e-16
```

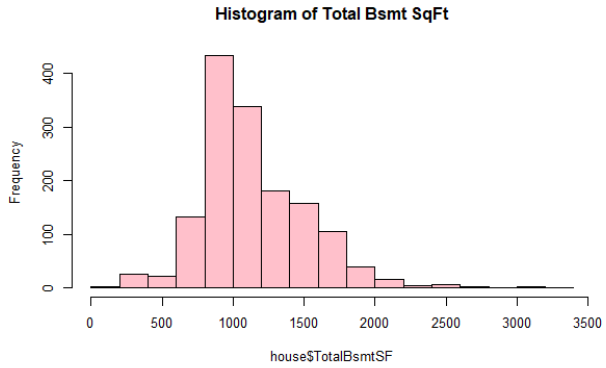
At first glance, the R-Squared would indicate that the model with the log of sales price fits better for predicting than the first model by having an adjusted R-Squared of .8736 compared to the first one which had an adjusted R-Squared value of .8235. The P-Value of both models sit below .05, which signifies that both models hold somewhat strong. The F-Statistic cannot be tested to see whether the regression coefficients can be 0 since the P-Values are less than .05.

Let’s observe the predictor variables to see if we can make some transformations to better the performance of the model for this analysis. Let’s plot the 4 continuous variables to see if we notice anything that we can work on.









It looks as though lot area can be manipulated due to the skewness and high kurtosis of the histogram. Using the logarithmic function of lot area may make this model output somewhat better. We will next perform a MLR on the logarithmic function of sales price, but use the logarithmic function of lot area to see how that changes the model.

```
Call:
lm(formula = log(house$SalePrice) ~ house$LotFrontage + log(house$LotArea) +
    house$YearBuilt + house$TotalBsmtSF + house$GrLivArea)

Residuals:
     Min       1Q   Median       3Q      Max
-0.63937 -0.06994  0.00372  0.07367  0.36891

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.596e+00  4.271e-01  -3.738 0.000195 ***
house$LotFrontage  1.020e-03  2.105e-04   4.845 1.45e-06 ***
log(house$LotArea) 1.323e-01  1.405e-02   9.418  < 2e-16 ***
house$YearBuilt    5.861e-03  2.044e-04  28.670  < 2e-16 ***
house$TotalBsmtSF  2.258e-04  1.059e-05  21.325  < 2e-16 ***
house$GrLivArea    3.588e-04  9.320e-06  38.500  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1185 on 1129 degrees of freedom
  (335 observations deleted due to missingness)
Multiple R-squared:  0.8786,   Adjusted R-squared:  0.878
F-statistic:  1634 on 5 and 1129 DF,  p-value: < 2.2e-16
```

From the output above, the R-Squared value increased slightly to .878, compared to the MLR of the log of sales price with just the lot area of .8736. This shows improvements to the P-Values along the intercept and predictor values as well. The F-Statistic cannot be used to see if the regression coefficients can be 0 since the P-Values are less than .05.

Conclusion

Initial exploratory data analysis should be performed when dealing with the Ames housing dataset before further analysis. Outliers, such as houses greater than 4,000 square feet and houses under 800 square feet will help better the models' fit and will be easier to work with when wrangling with the dataset. Although, this is not as practical to perform these techniques in real world situations, for the purposes of performing the basics of linear regression modeling was optimal.

In conclusion, the analysis of the sales price was very hard to predict with single variables individually, as they provided low results when performing the simple linear regression models individually. This is to be expected, since predicting house prices could have many factors that change the price. By combining the variables, the analysis strengthened when the multilinear regression was performed. MAE showed higher for prices that either had a low price per square foot or high price per square foot. For the average ranges it was relatively the same. Creating categorial variables by neighborhood was helpful in exploring analysis of the MLR while using general living area and garage area as the other two predictors, but still needs further analysis with predictors and adjustments to the response variable.

Model comparisons between the sales price and logarithmic function of the sales price indicates better model fit when the logarithmic function was made to the response variable. Adjustments to the predictor variables also showed better fits for the model, more specifically the lot area. The same 5 variables used to analyze the response variables were used.

The goal was to try and predict housing prices using a subset of the data and using goodness of fit metrics along with adjustments to the data during the modeling efforts to see if the basic OLS metrics were met when analyzing the various regressions that were created. There needs to be an understanding of the parameters and potential concerns of the models when assessing whether to make these models make predictions.

Appendix (R Code)

```
library("tidyverse", lib.loc="~/R/win-library/3.5")
library("ggplot2", lib.loc="~/R/win-library/3.5")
library("stargazer", lib.loc="~/R/win-library/3.5")
library("PerformanceAnalytics", lib.loc="~/R/win-library/3.5")
library(corrplot, lib.loc="~/R/win-library/3.5")

ames.df =
read.csv("c:/Users/vtika/Desktop/R/msds_410/ames_housing_data.csv", header
= TRUE,stringsAsFactors = FALSE)

ames.df = data.frame(ames.df)

str(ames.df)

########################################
###determining our sample set of data###
########################################

##Narrowing down to single family homes and removing houses where
GrllivArea is less than
##800 and greater than 4000

ames.df$dropCondition <- ifelse(ames.df$BldgType!='1Fam','01: Not SFR',
  ifelse(ames.df$SaleCondition!='Normal','02: Non-Normal Sale',
  ifelse(ames.df$Street!='Pave','03: Street Not Paved',
  ifelse(ames.df$YearBuilt <1950,'04: Built Pre-1950',
  ifelse(ames.df$TotalBsmtSF <1,'05: No Basement',
  ifelse(ames.df$GrLivArea <800,'06: LT 800 SqFt',
  '99: Eligible Sample')
  )))))

##follow up with a table to look at the eligible population

table(ames.df$dropCondition)

##Making a matrix to export to create a pic for the report

waterfall <- table(ames.df$dropCondition);

waterfall.matrix <- as.matrix(waterfall,4,2)


out.path <- "c:/Users/vtika/Desktop/R/msds_410/"
file.name <- 'summarstatistics.html';

stargazer(ames.df, type=c('html'),out=paste(out.path,file.name,sep=''),
          title=c('Table XX: Summary Statistics for Boston Housing'),
          align=TRUE, digits=2, digits.extra=2, initial.zero=TRUE,
median=TRUE)

# Eliminate all observations that are not part of the eligible sample
population;
elig.pop <- subset(ames.df,dropCondition=='99: Eligible Sample');
```

```
#Check that all remaining observations are eligible;
table(elig.pop$dropCondition)

##Select appopriate fields from eligible population
elig.pop <- elig.pop %>%
  select(SalePrice,SaleCondition,SaleType,
         GrLivArea,FirstFlrSF,SecondFlrSF,
         TotalBsmtSF,GarageArea,YrSold,
         MoSold,GarageType,GarageYrBlt,
         GarageArea,GarageCond,SubClass,
         Zoning,LotFrontage,LotArea,
         LotShape,Utilities,Neighborhood,
         BldgType,HouseStyle,OverallQual,
         OverallCond,YearBuilt,YearRemodel)


##create the elig.po as house dataset

house <- elig.pop
#View(house)

##### Question 2 #####

##perform EDA on two variables as potential predictor variables

##creating variables as well to determine a few things
library(corrplot)
# Correlation only makes sense for numerical features.
co <- cor(house[sapply(house, function(x) is.integer(x))], use =
"complete.obs")
corrplot(co)

# looking at corrplot sale price seem to has highest correlation with
GrLivArea and GarageArea

# We take these two variables to fit a regression model

fit_1 <- lm(house$SalePrice ~ house$GrLivArea)
fit_2 <- lm(house$SalePrice ~ house$GarageArea)

summary(fit_1)
# Adjusted R-2 squared value of 0.639
summary(fit_2)
# Adjusted R-2 squared value of 0.4348
# First model is a better fit

plot(house$GrLivArea, house$SalePrice, xlab="General Living Area", ylab =
"Sale Price", main = "General Living Area to Sales Price")
abline(fit_1, col="red")
```

```
plot(house$GarageArea, house$SalePrice, xlab="Garage Area", ylab="Sale
Price", main = "Garage Area to Sales Price")
abline(fit_2, col="red")

# Looking at residual plots for both fits

plot(fit_1$residuals, ylab = "Residuals", main = "Residuals of General
Living Area")
plot(fit_2$residuals, ylab= " Residuals", main = "Residuals of Garage
Area")

par(mfrow=c(2,2))
plot(fit_1)

par(mfrow=c(2,2), main = "Residuals versus Fitted for Garage Area")
plot(fit_2)

# Both residual seem to have zero mean and about constant variance.
# We have a closer look at the quantile plots

par(mfrow=c(1,2))
qqnorm(fit_1$residuals)
qqline(fit_1$residuals)
qqnorm(fit_2$residuals)
qqline(fit_2$residuals)

# In the first fit the residuals lie more closely on the line so it is
more closer to a normal distribution.
# Coefficient estimates, p-values and t-values etc are listed in the
summary tables
summary(fit_1)
summary(fit_2)

##### #####

##### Question 3 #####

fit_m <- lm(house$SalePrice~house$GrLivArea+house$GarageArea)
summary(fit_m)
# The AdjustedR squared value has increased to 0.7215 which is greater
than the individual models
# Thus our model has improved.
# But this may not always be the case
summary(fit_m$residuals)

plot(fit_m$residuals, ylab = "Residuals", main = "Residuals of Multi-
Linear Regression")

par(mfrow=c(1,1))
plot(fit_m$residuals, ylab="Residuals")
qqnorm(fit_m$residuals)
qqline(fit_m$residuals, col = "darkred")

# The residuals fit the normal more closely
```

```
##### #####

##### Question 4 #####

# Making new columns for residuals
house$resid <- fit_m$residuals
table(house$Neighborhood)
boxplot(house$resid ~ house$Neighborhood, horizontal=T, main = "Boxplot of
Residuals from MLR")

# Residual is observed value - value from fit
# Thus is rediuals >> 0 means the property is underpriced by our model
# If the residuals is << 0 means the property is overpriced by our model

house$X <- house$SalePrice/house$GrLivArea
y <- house %>%
  group_by(Neighborhood) %>%
  summarize(Y = mean(abs(resid)))
x <- house %>%
  group_by(Neighborhood) %>%
  summarize(x=mean(X))
plot(x$x,y$Y, xlab = "Price/Sqft", ylab = "MAE", main = "Scatterplot of
MAE against Price/SqFt", col = "darkgreen")
# The MAE is higher for either low price/sqft or high price/sqft
# For the average range it is almost the same.

# Creating new categorical variable to account for change in price/sqft
house$indi <- 0
house$indi[house$X < 100] <- 1
house$indi[house$X > 100 & house$X < 120] <- 2
house$indi[house$X > 120 & house$X < 140] <- 3
house$indi[house$X > 140 & house$X < 160] <- 4
house$indi[house$X > 160] <- 5
unique(house$indi)

house$indi <- as.factor(house$indi)

fit_new <- lm(house$SalePrice~house$GrLivArea+house$GarageArea+house$indi)
summary(fit_new)
# Adjusted R squared goes to 0.932!

# Comparing MAEs
mean(abs(fit_m$residuals))
mean(abs(fit_new$residuals)) # new model has lower MAE

##### #####

##### Question 5 #####

# We chose any four continous and any 1 discrete variable
fit1 <-
lm(house$SalePrice~house$LotFrontage+house$LotArea+house$YearBuilt+house$T
otalBsmtSF+house$GrLivArea)
```

```
fit2 <-
lm(log(house$SalePrice)~house$LotFrontage+house$LotArea+house$YearBuilt+ho
use$TotalBsmtSF+house$GrLivArea)

par(mfrow=c(1,1))
plot(fit1$residuals, main = "Residuals of Sales Price MLR")
par(mfrow=c(1,1))
plot(fit2$residuals,main = "Residuals of Log(Sales Price) MLR")
# After the log transformation the variance in the error term reduces
# This transformation might be useful when there is heteroskedasticity in
the model
# In our case it is not the case, but ususally it can help


summary(fit1)
summary(fit2) # R squared value of second model is better

par(mfrow=c(2,2))
hist(house$LotFrontage, col = "darkblue", main = "Histogram of Lot
Frontage")
hist(house$LotArea, col = "darkgreen", main = "Histogram of Lot Area")
hist(house$GrLivArea, col = "darkred", main = "Histogram of Gen living
Area")
hist(house$TotalBsmtSF, col = "pink", main = "Histogram of Total Bsmt
SqFt")

# Lets Transform LotArea

fit3 <-
lm(log(house$SalePrice)~house$LotFrontage+log(house$LotArea)+house$YearBui
lt+house$TotalBsmtSF+house$GrLivArea)
summary(fit3)
# R square value improved slightly!
```