Assignment #3

Data Analysis and Regression

Name: Young, Brent

Predict 410 Section #: 57

Quarter: Summer 2017

Introduction

Context

The dataset that we will be working with is called Ames Housing data (includes 2,930 rows) and is observational data collected by Ames Assessor's Office. The data includes houses sold in Ames, Iowa from 2006 to 2010 with SalePrice as the response variable and 81 predictors (includes nominal, ordinal, discrete, and continuous variables). The final goal is to build a Predictive model (e.g., multiple linear regression) to predict SalePrice of a house using other attributes. In order to accomplish this, an iterative regression process focused on statement of the problem, selection of potentially relevant variables, data collection, model specification, parameter estimation, model adequacy checking, model validation and model use will be conducted within the next five weeks.

Objectives/Purpose

The overall purpose/objective of assignment 3 is to begin building regression models (e.g., simple linear regression models, multiple linear regression models, and regression models for the transformed response log (SalePrice)) for the home sale price by fitting these specific models. First, a waterfall of my drop conditions with counts will be provided to define the sample data/population of interest that we will want to use for the modeling purpose and ensure that the sample data is representative of the population that we want to model. Second, an initial exploratory data analysis/views of the data will be conducted so that we can select two of the most promising predictor variables for predicting SalePrice. Third, the two predictor variables will then be used to fit two simple linear regression models by using diagnostic plots (e.g., residual plots) to assess goodness-of-fit of each model. ANOVA, summary tables, predictive error, and multiple r-squared will also be used to answer the following questions: Is my model significant or not, what is my model and how do I interpret it, and how good is my model. Fourth, we will then combine the two simple linear regression models so that we can begin the creation of the multiple linear regression model. Relevant diagnostic plots will also be conducted to assess the goodness-of-fit of each model, while analyzing the results to see if it fits better than the simple linear regression models (e.g., using r-squared and adjusted r-squared). Fifth, we will make a boxplot of the residuals by neighborhood, so that we can see which neighborhoods are better fit by the model and which ones are consistently overpredicted/underpredicted. Mean MAE and the mean price per square foot for each neighborhood will also be computed so that we can see if there is a relationship between these two quantities. Next, the neighborhoods will then be grouped by price per square foot and dummy variables will be created to be included in the multiple regression model. This will allow us to determine our base category, refit the model with the indicator variables, and compare the MAE of the original multiple regression model with our new multiple regression model to determine which model fits better based on the MAE. Sixth, a transformation of the response variable from the sale price to the natural logarithm of the sale price will be conducted. We will then fit two models using the same set of predictor variables vs. SalePrice and log(SalePrice) so that we can interpret these models, figure out which model fits better, and discuss whether the transformation of log(SalePrice) improved the model fit. We will also consider transformations to the predictors and will fit one more model using any transformations that we find appropriate to improve the model. Lastly, variable transformation and impacts of outlier deletion on the modeling process and the results will be discussed and next steps in the modeling process will be discussed so that we can continue to enhance the model.

Section 1: Sample Definition

Figure 1: Boxplot of Sale Price & Building Style

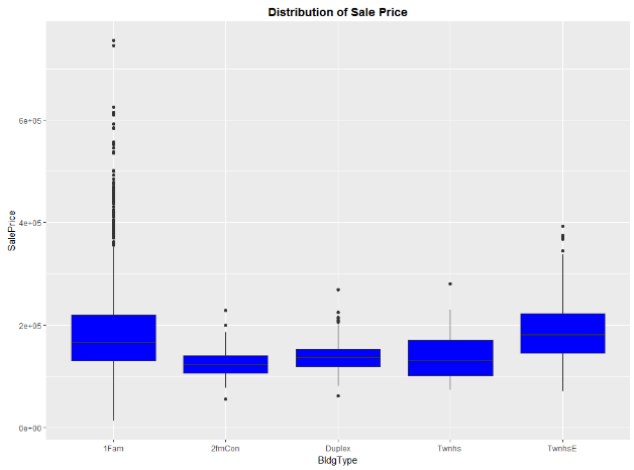


Figure 2: Boxplot of Sale Price & Sale Condition

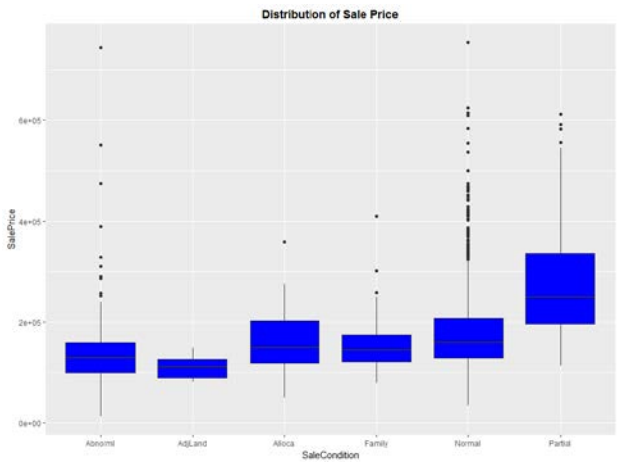


Figure 3: Waterfall of 'Drop Conditions'

```
        1Fam 2fmCon Duplex  Twnhs TwnhsE
Before  2425     62    109    101    233

        Abnorml AdjLand  Alloca  Family  Normal Partial
Before  190       12       24       46    2413     245

        1Fam 2fmCon Duplex  Twnhs TwnhsE
After   2002      0      0      0      0

        Abnorml AdjLand  Alloca  Family  Normal Partial
After     0        0        0        0    2002    0
```

Definition of Sample Data & Observations: Figure 1 shows a boxplot of SalePrice & Bldg Type and Figure 2 shows a boxplot of SalePrice & Sale Condition. When comparing figure 1 & 2, 'single-family' homes and 'normal' sale have similar medians as well as the amount and location of the outliers. As a result, based on this, it makes sense for the sample population/data of interest for 'typical' homes in Ames, Iowa to be 'single-family' homes with 'normal' sales in Ames, Iowa. Figure 3 shows the population of interest ('single family' homes and sale condition 'normal' in Ames, Iowa) after the drop conditions were applied, which comes out to 2002 rows and 81 variables.

Section 2: Simple Linear Regression Models (EDA)

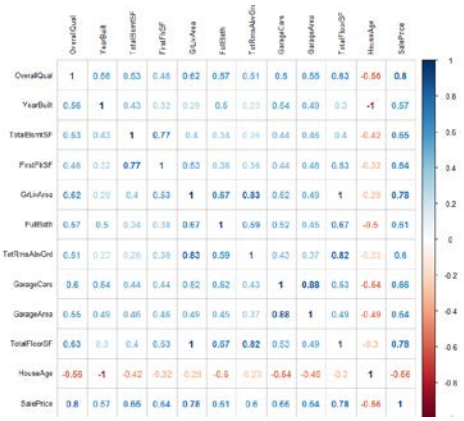


Figure 4: Correlation Matrix of Numeric Variables +/- 0.50

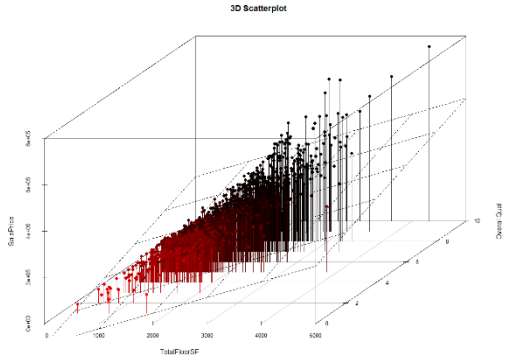


Figure 5: 3D Scatterplot of SalePrice, TotalFloorSF, and OverallQual

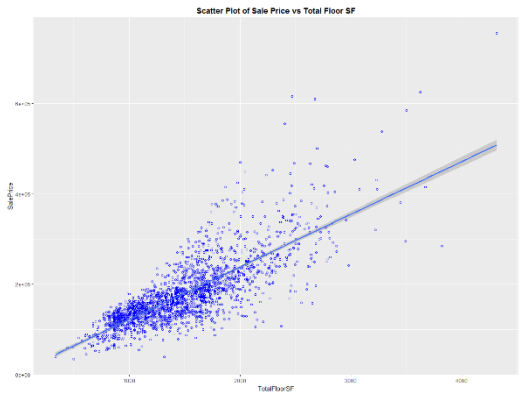


Figure 6: Scatterplot of SalePrice vs. TotalFloorSF

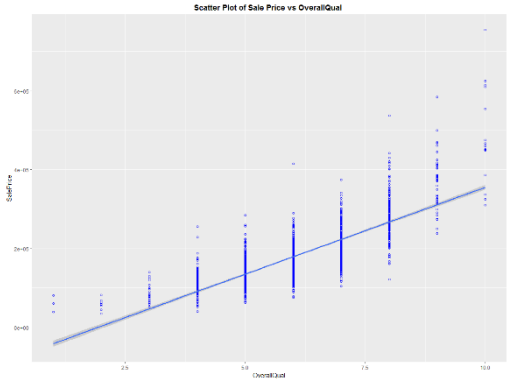


Figure 7: Scatterplot of SalePrice vs. OverallQual

Observations: Figure 4 shows a Correlation Matrix of numeric variables that had correlations beyond at least +0.5 or –0.5. OverallQual (0.8) and TotalFloorSF (0.78) have the strongest positive correlations with SalePrice. Scatterplots of SalePrice vs. TotalFloorSF (figure 6) shows a "funnel" shape and heteroscedasticity, with a positive correlation between TotalFloorSF and SalePrice (as TotalFloorSF increases, SalePrice increases). In terms of the variable OverallQual, the scatterplot (figure 7) shows a positive correlation between OverallQual and SalePrice (as OverallQual increases, SalePrice increases), but does not show a nice linearly correlated relationship. The 3D scatterplot of SalePrice, TotalFloorSF, and OverallQual shows a similar story that we saw in figure 6 and 7, which shows a hyperplane sloping upwards (higher the OverallQual and TotalFloorSF, the higher the price). As a result, since OverallQual (0.8) and TotalFloorSF (0.78) have the strongest positive correlations with SalePrice, we will use these two predictor variables as the most promising for predicting SalePrice. However, it's important to note that we will need to consider a transformation of SalePrice at some point in the model building process. By doing transformation, it will help achieve linearity, homogeneity of variance, and normality/symmetric about the regression equation.

Section 2: Simple Linear Regression Models

Section 2.1: Model #1 (TotalFloorSF)

Figure 8: Analysis of Variance for SalePrice ~ TotalFloorSF

```
Analysis of Variance Table

Response: SalePrice
               Df     Sum Sq    Mean Sq F value    Pr(>F)
TotalFloorSF    1 6.5852e+12 6.5852e+12  3152.3 < 2.2e-16 ***
Residuals    2000 4.1780e+12 2.0890e+09
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

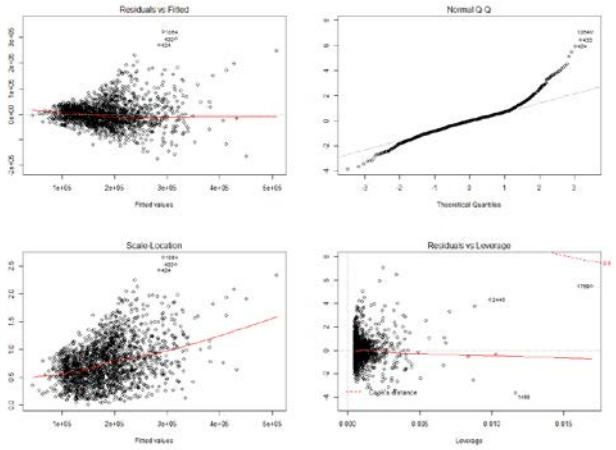
Observations: Figure 8 shows an ANOVA for SalePrice ~ TotalFloorSF. The F-statistic showed 3152 with a p-value = < 2.2e-16. Given that the p-value is very small, we can reject the null hypothesis that all the regr ession coefficients are equal to zero. Therefore, the model has produced statistically significant results to be investigated.

Figure 9: Simple Linear Regression Model SalePrice ~ TotalFloorSF

```
Call:
lm(formula = SalePrice ~ TotalFloorSF, data = subdat)

Residuals:
    Min      1Q  Median      3Q     Max
-174349  -25635   -1347   19989  321751

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5910.614   3250.827   1.818   0.0692 .
TotalFloorSF  116.331      2.072  56.145   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 45710 on 2000 degrees of freedom
Multiple R-squared:  0.6118,   Adjusted R-squared:  0.6116
F-statistic:  3152 on 1 and 2000 DF,  p-value: < 2.2e-16
```

Observations: Figure 9 shows a summary of the Linear Regression Model SalePrice ~ TotalFloorSF. The equation of the regression line is: SalePrice = 5910.61 + 116.33*TotalFloorSF. Therefore, for every additional 1 square-feet, average sales price goes up by $116.33. Since the t-test of TotalFloorSF is statistically significant (p<0.001), we can use this equation. The residual standard error of 45710, shows us that when predicting SalePrice, one standard error = $45,710. The multiple R-squared value of 0.6118 indicates that 61.18% of the variation in SalePrice is explained by the predictor variable TotalFloorSF. Overall, this concludes that the model is “mediocre” for one variable given the multiple R-squared value of 0.6118.

Figure 10: Scatterplots with Residuals & QQ-Plot of Residuals



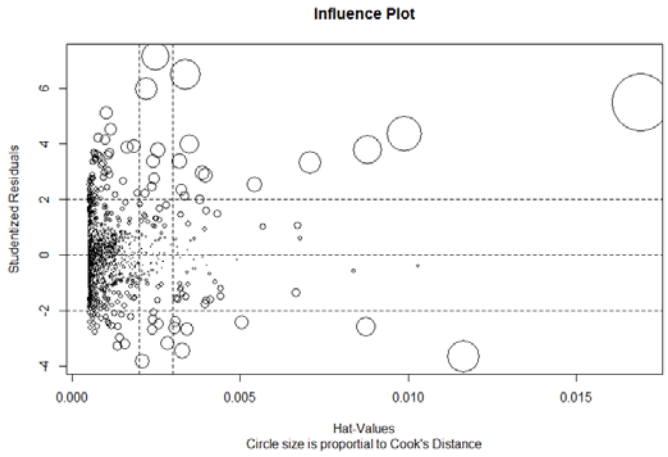
Observations: Figure 10, shows scatterplots with residuals and qq-plots of residuals so that we can check to make sure the model is meeting all the assumptions. The QQ plot reveals that the density distribution is non-normal due to a systematic pattern created by outliers (which is evident in the residuals versus leverage plot). This is present in the plot where some of the data points are progressively departing from the line in the lower left hand corner and upper right hand corner of the plot. This indicates non-normality and shows us that it does not correspond relatively well to a standard normal distribution. The scatterplot of residuals vs. fitted shows us that there are a large amount of data points on the left side of the plot and fewer data points on the right side of the plot. This pattern is an indication of heteroscedasticity (the residual plot "flares-out" in a funnel pattern as x gets larger), which is a violation of the assumption of constant variance for error terms. The points are also "concentrated", as evident in the scale-location plot, which should be "random". By comparison, a healthy normal probability plot of the residuals would be relatively linear and would have a random scatter of data over the range of values for the independent variable. In addition, it is highly desirable for the residuals to conform to a normal distribution with few to no outliers. As a result, given the assumptions of a simple regression analysis and the revelations/results from above, the regression model does not fit the data particularly well and can be improved.

Figure 11: Predictions: Simple Linear Regression Model SalePrice ~ TotalFloorSF

	fit	lwr	upr
1	198555.5	108894.73	288216.3
2	110143.6	20452.96	199834.2
3	160515.1	70854.53	250175.7
4	251370.0	161676.32	341063.6
5	195414.5	105754.54	285074.6
6	192506.3	102846.84	282165.7

Observations: Figure 11 shows us that the predicted value of the first house is $198,555.5. Additionally, the lower and upper confidence bands shows $108,894.73 and $288,216.3, respectively. This means that the 95% confidence band on this predicted value is $108,894.73 and $288,216.3, which is a "wide" band. This means that our model is not that great, which was validated by the multiple R-squared value of 0.6118 and large predicted error of $54,960.

Figure 11B: Influence Plot



Observations: Figure 11B shows an Influence Plot with studentized residuals on the Y-axis and hat values on the x-axis. The plot shows us that there are x and y outliers (e.g., high hat values mean that in the x space those are outliers, they are far away from the centroid in the x-space). Additionally, the plot shows us that some of those outliers are influential points as indicated by the high hat values and large circles and high residuals and large circles as well (Cooks Distance). This means that we have some significant influential points in the x and y outlier direction. This was also confirmed using summary code of "Potentially influential observations of lm(formula = SalePrice ~ TotalFloorSF)" (see appendix) where we saw influential points by looking at Covariance Ratio, Cooks Distance, and Hat Matrix.

Section 2.2: Model #2 (OverallQual)

Figure 12: Analysis of Variance for SalePrice ~ OverallQual

```
Analysis of Variance Table

Response: SalePrice
              Df     Sum Sq    Mean Sq F value    Pr(>F)
OverallQual    1 6.9449e+12 6.9449e+12  3637.7 < 2.2e-16 ***
Residuals   2000 3.8183e+12 1.9092e+09
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

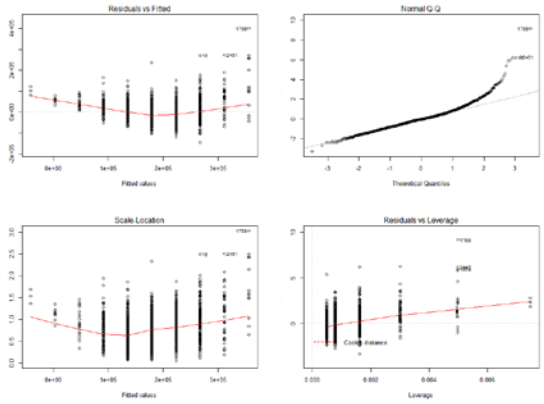
Observations: Figure 12 shows an ANOVA for SalePrice ~ OverallQual. The F-statistic showed 3638 with a p-value = < 2.2e-16. Given that the p-value is very small, we can reject the null hypothesis that all the regr ession coefficients are equal to zero. Therefore, the model has produced statistically significant results to be investigated.

Figure 13: Simple Linear Regression Model SalePrice ~ OverallQual

```
Call:
lm(formula = SalePrice ~ OverallQual, data = subdat)

Residuals:
    Min      1Q  Median      3Q     Max
-145475  -26403   -3650   19400  399411

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -84981.3     4487.5  -18.94   <2e-16 ***
OverallQual  44057.0      730.5   60.31   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 43690 on 2000 degrees of freedom
Multiple R-squared:  0.6452,   Adjusted R-squared:  0.6451
F-statistic:  3638 on 1 and 2000 DF,  p-value: < 2.2e-16
```

Observations: Figure 13 shows a summary of the Linear Regression Model SalePrice ~ OverallQual. The equation of the regression line is: SalePrice = -84981.3 + 44057*OverallQual. Since the t-test of OverallQual is statistically significant (p<0.001), we can use this equation. The residual standard error of 43690, shows us that when predicting SalePrice, one standard error = $43,690. The multiple R-squared value of 0.6452 indicates that 64.52% of the variation in SalePrice is explained by the predictor variable OverallQual. Overall, this concludes that the model is “mediocre” for one variable given the multiple R-squared value of 0.6452.

Figure 14: Scatterplots with Residuals & QQ-Plot of Residuals



Observations:

Figure 14, shows scatterplots with residuals and qq-plots of residuals so that we can check to make sure the model is meeting all the assumptions. The QQ plot reveals that the density distribution is non-normal. This is present in the plot where some of the data points are progressively departing from the line in the upper right hand corner of the plot. This indicates non-normality and shows us that it does not correspond relatively well to a standard normal distribution. The scatterplot of residuals vs. fitted shows us that there is "funnel" pattern. This pattern is an indication of heteroscedasticity (the residual plot "flares-out" as x gets larger), which is a violation of the assumption of constant variance for error terms. We see a similar story in the scale-location plot. By comparison, a healthy normal probability plot of the residuals would be relatively linear and would have a random scatter of data over the range of values for the independent variable. In addition, it is highly desirable for the residuals to conform to a normal distribution with few to no outliers. As a result, given the assumptions of a simple regression analysis and the revelations/results from above, the regression model does not fit the data particularly well and can be improved.

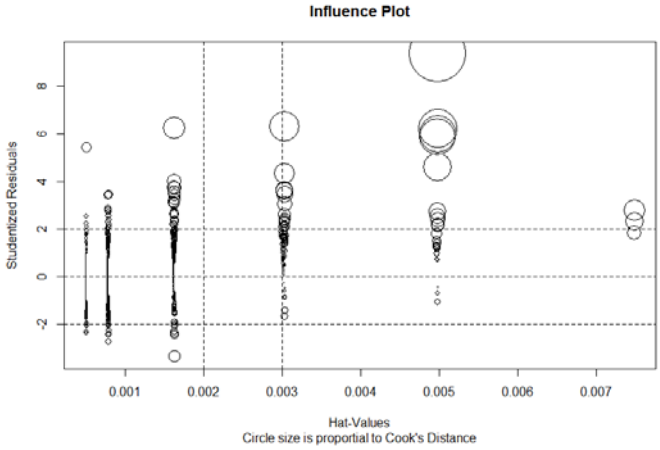
Figure 15: Predictions: Simple Linear Regression Model SalePrice ~ OverallQual

```
       fit       lwr      upr
1 179360.7  93648.68 265072.7
2 135303.7  49579.81 221027.6
3 179360.7  93648.68 265072.7
4 223417.7 137693.60 309141.7
5 135303.7  49579.81 221027.6
6 179360.7  93648.68 265072.7
```

Observations:

Furthermore, figure 15 shows us that the predicted value of the first house is $179,360.7. Additionally, the lower and upper confidence bands shows $93648.68 and $265072.7, respectively. This means that the 95% confidence band on this predicted value is $93648.68 and $265,072.7, which is a "wide" band. This means that our model is not that great, which was validated by the multiple R-squared value of 0.6452 and large predicted error of $43,690.

Figure 15B: Influence Plot



Observations: Figure 15B shows an Influence Plot with studentized residuals on the Y-axis and hat values on the x-axis. The plot shows us that there are x and y outliers (e.g., high hat values mean that in the x space those are outliers, they are far away from the centroid in the x-space). Additionally, the plot shows us that some of those outliers are influential points as indicated by the high hat values and large circles and high residuals and large circles as well (Cooks Distance). This means that we have some significant influential points in the x and y outlier direction. This was also confirmed using summary code of "Potentially influential observations of lm(formula = SalePrice ~ OverallQual)" (see appendix) where we saw influential points by looking at Covariance Ratio, Cooks Distance, and Hat Matrix.

Section 3: Multiple Linear Regression Models – Model #3

Figure 16: Analysis of Variance for SalePrice ~ TotalFloorSF + OverallQual

```
Analysis of Variance Table

Response: SalePrice
               Df     Sum Sq    Mean Sq F value    Pr(>F)
TotalFloorSF    1 6.5852e+12 6.5852e+12  5329.2 < 2.2e-16 ***
OverallQual     1 1.7079e+12 1.7079e+12  1382.1 < 2.2e-16 ***
Residuals    1999 2.4702e+12 1.2357e+09
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Observations: Figure 16 shows an ANOVA for SalePrice ~ TotalFloorSF + OverallQual. The F-statistic on 2 showed 3356 with a p-value = < 2.2e-16. Given that the p-values are very small for both predictor variable s, we can reject the null hypothesis that all the regression coefficients are equal to zero. Therefore, the model has produced statistically significant results to be investigated.

Figure 17: Multiple Linear Regression Model SalePrice ~ TotalFloorSF + OverallQual

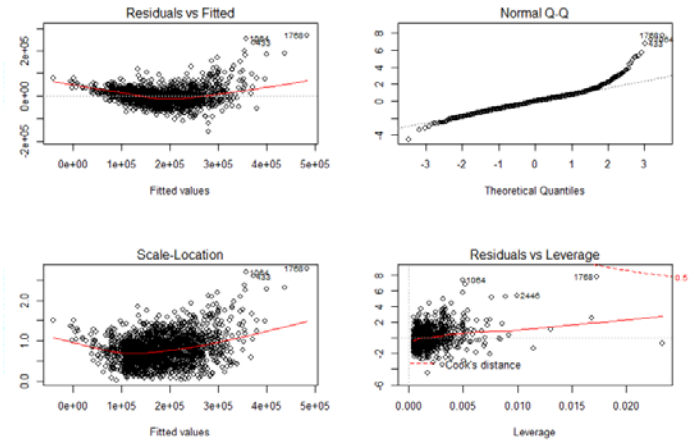
```
Call:
lm(formula = SalePrice ~ TotalFloorSF + OverallQual, data = subdat)

Residuals:
    Min      1Q  Median      3Q     Max
-158800  -21396      74   17646  270800

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -91159.880   3615.068  -25.22   <2e-16 ***
TotalFloorSF     67.956      2.057   33.03   <2e-16 ***
OverallQual   28206.404    758.710   37.18   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 35150 on 1999 degrees of freedom
Multiple R-squared:  0.7705,   Adjusted R-squared:  0.7703
F-statistic:  3356 on 2 and 1999 DF,  p-value: < 2.2e-16
```

Observations: Figure 17 shows a summary of the Multiple Linear Regression Model SalePrice ~ TotalFloorSF + OverallQual. The equation of the regression line is: SalePrice = -91159.88 + 67.956* TotalFloorSF + 28206.40* OverallQual. Since the t-test of both predictor variable are statistically significant (p<0.001), we can use this equation. The residual standard error of 35150, shows us that when predicting SalePrice, one standard error = $35150. The multiple R-squared value of 0.7705 indicates that 77.05% of the variation in SalePrice is explained by the predictor variables TotalFloorSF and OverallQual.

By adding an additional variable, there was a “net gain” since both multiple r-squared and adjusted r-squared increased (if adjusted r-squared decreased, it would be a “net loss”). As a result, this model fits

better than the simple linear regression models since the adjusted r-squared of is 0.7703 for this model is higher compared to model #1 (adjusted r-squared: 0.6116) and model #2 (adjusted r-squared: 0.6451). Adjusted r-squared was used since we are comparing models of different sizes, and as a result, this metric provides a tradeoff between model fit and model complexity; whereas adding more predictor variables will always cause r-squared to increase. Additionally, it's interesting to note that the predicted error is also smaller (predicted error: $35150) than model #1 ($54,960) and model #2 (predicted error: $43,690). This also shows evidence that the multiple linear regression model fits better than the simple linear regression models.

Figure 18: Scatterplots with Residuals & QQ-Plot of Residuals



Observations: Figure 18, shows scatterplots with residuals and qq-plots of residuals so that we can check to make sure the model is meeting all the assumptions. The QQ plot reveals that the density distribution is non-normal. This is present in the plot where some of the data points are progressively departing from the line in the upper right hand corner of the plot. This indicates non-normality and shows us that it does not correspond relatively well to a standard normal distribution. The scatterplot of residuals vs. fitted shows us that there is "bowl shaped" pattern with heteroscedasticity and a few outliers, instead of the "funnel" shaped pattern in model #1 and model #2. By comparison, a healthy normal probability plot of the residuals would be relatively linear and would have a random scatter of data over the range of values for the independent variable. In addition, it is highly desirable for the residuals to conform to a normal distribution with few to no outliers.

Figure 19: Predictions: Multiple Linear Regression Model SalePrice ~ TotalFloorSF + OverallQual

```
       fit       lwr      upr
1 190612.9 121653.16 259572.7
2 110760.3  41778.83 179741.8
3 168391.5  99431.86 237351.1
4 249671.1 180687.30 318655.0
5 160571.7  91589.33 229554.1
6 187079.2 118121.19 256037.3
```

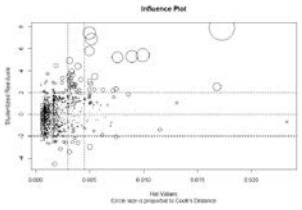
Observations: Figure 19 shows us that the predicted value of the first house is $190,612.9. Additionally, the lower and upper confidence bands shows $121653.16 and $259572.7, respectively. This means that the 95% confidence band on this predicted value is $121653.16 and $$259572.7. This is still a "wide" band, but an improvement (e.g., tighter band) in what we saw in model #1 ($108,894.73 and $288,216.3) and model #2 ($93648.68 and $265,072.7).

Figure 20: Detecting Multicollinearity Using Variance Inflation Factors

```
TotalFloorSF  OverallQual
    1.666774     1.666774
```

Observations: Figure 20 shows us the VIF for the two predictors TotalFloorSF (VIF = 1.666774) and QualityIndex (VIF = 1.666774). A VIF of 1 would mean that no multicolinearity exists at all, while a large VIF number (e.g., 10) would indicate serious multicolinearity issues. As a result, since the VIF for TotalFloorSF and QualityIndex is low, this concludes that we don't have serious multicolinearity issues.

Figure 21: Influence Plot



Observations: Figure 21 shows an Influence Plot with studentized residuals on the Y-axis and hat values on the x-axis. The plot shows us that there are some x and y outliers (e.g., high hat values mean that in the x space those are outliers, they are far away from the centroid in the x-space). Additionally, the plot shows us that some of those outliers are influential points as indicated by the high hat values and large circles and high residuals and large circles as well (Cooks Distance). This means that we have some significant influential points in the x and y outlier direction. This was also confirmed using summary code of "Potentially influential observations of lm(formula = SalePrice ~ TotalFloorSF + OverallQual)" (see appendix) where we saw influential points by looking at Covariance Ratio, Cooks Distance, and Hat Matrix. When comparing the influence plots for model #3 versus models # 1 and 2, there seems to be less influential points in the x-direction and less influential points overall, which is an indication that model #3 fits better.

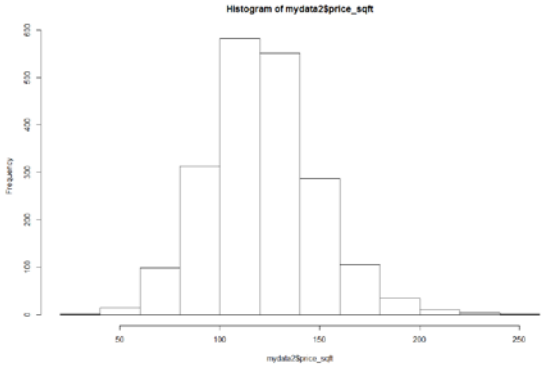
Figure 21: Analysis of Variance Comparing Model #3 vs. Models #1 & 2

```
Analysis of Variance Table

Model 1: SalePrice ~ TotalFloorSF
Model 2: SalePrice ~ TotalFloorSF + OverallQual
  Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
1   2000 4.1780e+12
2   1999 2.4702e+12  1 1.7079e+12 1382.1 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Analysis of Variance Table

Model 1: SalePrice ~ OverallQual
Model 2: SalePrice ~ TotalFloorSF + OverallQual
  Res.Df        RSS Df  Sum of Sq    F    Pr(>F)
1   2000 3.8183e+12
2   1999 2.4702e+12  1 1.3482e+12 1091 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Observations: In conclusion, this model #3 fits better than the simple linear regression models (model #1 and #2) since the adjusted r-squared was higher, predicted error was lower, and there were less influential points than model #1 and #2. Additionally, Figure 22 shows that the F-statistic for model #3 is significantly better than model #1 and #2. However, additional improvement is needed (e.g., transformation of the response variable SalePrice) to handle the problems caused by non-normality, non-linearity, heteroscedasticity or non-constant variance, and outliers/leverage/influential points that we have seen so far in this analysis.

Section 4: Neighborhood Accuracy

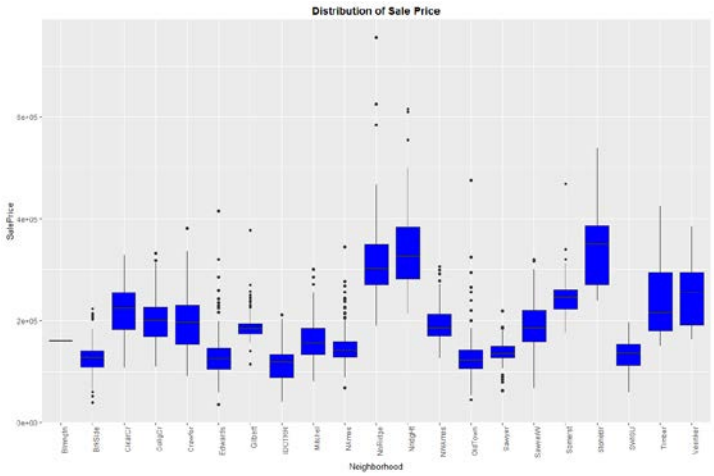
Figure 22: Histogram of Price Per SQFT & Summary Statistics




```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
30.37  103.29  119.82  121.13  137.44  248.99
```

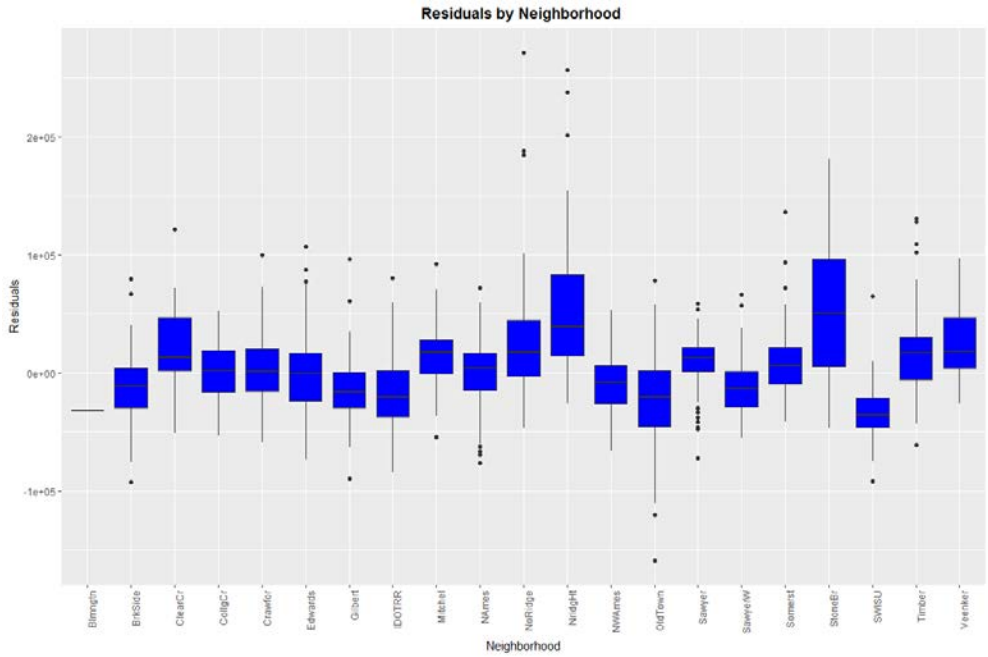
Observations: Figure 22 shows a histogram of price per sqft and summary statistics so that we can use it to help define our clusters. The summary statistics indicate that 25% of houses are $103 or less, 25% are between $103 to $120, 25% are between $120 to $137, and another 25% are more than $137.

Figure 23: Boxplot of SalePrice & Neighborhood



Observations: Figure 23 shows a boxplot of SalePrice &Neighborhood, which allows us to see if SalePrice is correlated with Neighborhood. It appears that neighborhoods such as North Ames, College Creek, and Old Town on average have a low SalePrice, while neighborhoods such as Northridge, Northridge Heights, and Stone Brook have higher SalePrice on average. As a result, this suggests that there is correlation between SalePrice and Neighborhood because the Average SalePrice is different for these different categories. It's also interesting to note that the higher the sale price, the lower the amount of houses sold in that particular neighborhood and vice versa.

Figure 27: Boxplot of Residuals by Neighborhood

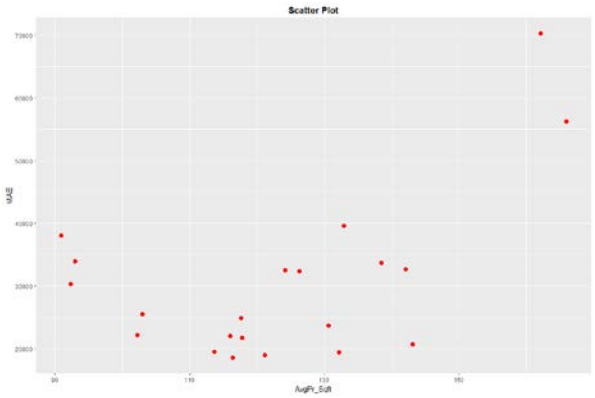


Observations: Figure 27 shows a boxplot of Residuals by Neighborhood, which allows us to see which neighborhoods are better fit by the model. CollgCr, Crawford, Edwards, North Ames, Northwest Ames, and Somerset are better fit by the model. Clear Creek, Northridge Heights, Northridge, and Stone Brook are some neighborhoods that over predicted. Gilbert, Iowa DOT and Rail Road, Old Town, and South & West of Iowa State University are some neighborhoods that are underpredicted.

Figure 28: Mean MAE and the Mean Price per Square Foot

	Neighborhood	MAE	AvgPr_Sqft
1	NridgHt	56277.05	165.95726
2	StoneBr	70276.24	162.117
3	Somerst	20716.89	143.13207
4	Timber	32745.11	142.10616
5	Veenker	33770.68	138.49864
6	NoRidge	39679.42	132.94913
7	CollgCr	19456.93	132.17138
8	Mitchel	23698.78	130.63527
9	Blmngtn	32421.66	126.29937
10	ClearCr	32538.55	124.20393
11	Sawyer	18964.27	121.143
12	SawyerW	21749.51	117.779
13	Crawfor	24909.4	117.66865
14	NAmes	18636.9	116.40979
15	Gilbert	22075.68	116.02615
16	NWAmes	19580.82	113.64843
17	Edwards	25566.94	102.94001
18	BrkSide	22216.23	102.21083
19	OldTown	33939.68	92.9912
20	IDOTRR	30364.13	92.31283
21	SWISU	38096.46	90.91083

Figure 29: Plot of Mean MAE and the Mean Price per Square Foot



Observations: Figure 29 shows a scatterplot of MAE on the y-axis and mean price per square foot on the x-axis. The results show that there is a strong positive correlation between MAE and mean price per square foot (as mean price per square foot increases, MAE increases). This means that as price per square foot increases, the forecasts or predictions are to the eventual outcomes becomes worse or more inaccurate. There will be 3 groups created by sqft: NbhdGrp1 (<=103), NbhdGrp2 (<=120), NbhdGrp3 (<=137), and a baseline category NbhdGrp4 (aka: Other houses).

Figure 31: Analysis of Variance for SalePrice ~ TotalFloorSF + OverallQual + NbhdGrp1 + NbhdGrp2 + NbhdGrp3

```
Analysis of Variance Table

Response: SalePrice
               Df     Sum Sq    Mean Sq  F value    Pr(>F)
TotalFloorSF    1 6.5852e+12 6.5852e+12 11857.25 < 2.2e-16 ***
OverallQual     1 1.7079e+12 1.7079e+12  3075.17 < 2.2e-16 ***
NbhdGrp1        1 6.3280e+11 6.3280e+11  1139.42 < 2.2e-16 ***
NbhdGrp2        1 4.2489e+11 4.2489e+11   765.05 < 2.2e-16 ***
NbhdGrp3        1 3.0393e+11 3.0393e+11   547.26 < 2.2e-16 ***
Residuals    1996 1.1085e+12 5.5537e+08
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 32: Multiple Linear Regression Model SalePrice ~ TotalFloorSF + OverallQual + NbhdGrp1 + NbhdGrp2 + NbhdGrp3

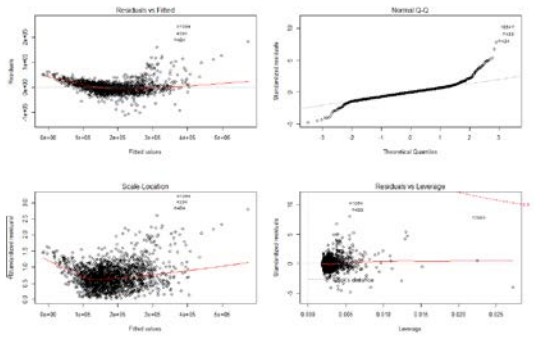
```
Call:
lm(formula = SalePrice ~ TotalFloorSF + OverallQual + NbhdGrp1  +
    NbhdGrp2  + NbhdGrp3 , data = subdat)

Residuals:
    Min      1Q  Median      3Q     Max
-113786  -12165    -998    9457  241527

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    -974.91    3073.42  -0.317    0.751
TotalFloorSF    108.60       1.61  67.434   <2e-16 ***
OverallQual   10620.38     621.44  17.090   <2e-16 ***
NbhdGrp1      -88442.35    1823.49 -48.502   <2e-16 ***
NbhdGrp2      -57380.64    1600.15 -35.860   <2e-16 ***
NbhdGrp3      -36031.23    1540.22 -23.394   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23570 on 1996 degrees of freedom
Multiple R-squared:  0.897,    Adjusted R-squared:  0.8968
F-statistic:  3477 on 5 and 1996 DF,  p-value: < 2.2e-16
```

Observations: Figure 32 shows a summary of the Multiple Linear Regression Model SalePrice ~ TotalFloorSF + OverallQual+ NbhdGrp1 + NbhdGrp2 + NbhdGrp3 . The equation of the regression line is: SalePrice = -974.91 + 108.60* TotalFloorSF + 10620.38* OverallQual -88442.35* NbhdGrp1 -57380.64 * NbhdGrp2 -36031.23*NbhdGrp3 . Since the t-test of all the predictor variables are statistically significant, we can use this equation. The baseline category is NbhdGrp4 (aka: Other houses).

The results suggest that when NbhdGrp1 is compared to the Other houses, NbhdGrp1 homes on average, have a SalePrice of $-88442.35 less and that it is significant. Furthermore, when NbhdGrp2 is compared to the Other houses, NbhdGrp2 homes on average, have a SalePrice of 57380.64 less and that it is significant. Lastly, when NbhdGrp3 is compared to the Other houses, NbhdGrp3 homes on average, have a SalePrice of $36031.23 less and that it is significant. The residual standard error of 23570, shows us

that when predicting SalePrice, one standard error = $23570. The multiple R-squared value of 0.897 indicates that 89.97% of the variation in SalePrice is explained by the predictor variables.

Figure 33: Scatterplots with Residuals & QQ-Plot of Residuals



Observations: Figure 33, shows scatterplots with residuals and qq-plots of residuals so that we can check to make sure the model is meeting all the assumptions. The QQ plot reveals that the density distribution is non-normal. This is present in the plot where some of the data points are progressively departing from the line in the upper right hand corner of the plot. This indicates non-normality and shows us that it does not correspond relatively well to a standard normal distribution. The scatterplot of residuals vs. fitted shows us that there is "funnel shaped" pattern with heteroscedasticity and a few outliers. By comparison, a healthy normal probability plot of the residuals would be relatively linear and would have a random scatter of data over the range of values for the independent variable. In addition, the residual vs. leverage plot shows that there are some influential points on the right side of the graph. It's important to note that it is highly desirable for the residuals to conform to a normal distribution with few to no outliers. This shows that we still have the same issues as we saw in the previous models: non constant variance, non-normality, and influential points. As a result, in order to correct these problems we are going to define a new variable called logSalePrice in the next section, which is the log of SalePrice.

Figure 34: Predictions: MLR Model SalePrice ~ TotalFloorSF + OverallQual+ NbhdGrp1 + NbhdGrp2 + NbhdGrp3 .

	fit	lwr	upr
1	206559.0	160286.04	252831.9
2	92052.6	45767.04	138338.2
3	171046.5	124781.16	217311.9
4	245134.7	198858.03	291411.3
5	171656.9	125375.47	217938.4
6	200911.7	154641.54	247181.9

Observations: Figure 34 shows us that the predicted value of the first house is $206559.0. Additionally, the lower and upper confidence bands shows $160286.04 and $252831.9, respectively. This means that the 95% confidence band on this predicted value is $160286.04 and $252831.9.

Figure 35: MAE of Model #3 vs. New Multiple Regression Model

Model #3

```
> MAE
[1] 25383.46
```

New Multiple Regression Model

```
> MAE2
[1] 15122.9
```

Observations: Figure 35 shows the MAE of Model #3 (25383.46) vs. New Multiple Regression Model (15122.9). Based on the MAE, the new regression model appears to fit better than Model #3, since 15122.9 is less than 25383.46. This makes sense since Adjusted R-squared: 0.8968 was higher than model #3: 0.7703. Additionally, it's interesting to note that the predicted error of $23570 was also lower than model #3: $35150.

Section 5: SalePrice versus Log SalePrice as the Response

Section 5.1: SalePrice Model

Figure 36: Analysis of Variance for SalePrice ~ TotalFloorSF + OverallQual + GarageCars + TotalBsmtSF + NbhdGrp1 + NbhdGrp2 + NbhdGrp3 .

```
Analysis of Variance Table

Response: SalePrice
               Df     Sum Sq    Mean Sq  F value    Pr(>F)
TotalFloorSF    1 6.5852e+12 6.5852e+12 13156.03 < 2.2e-16 ***
OverallQual     1 1.7079e+12 1.7079e+12  3412.01 < 2.2e-16 ***
GarageCars      1 2.2170e+11 2.2170e+11   442.91 < 2.2e-16 ***
TotalBsmtSF     1 4.8846e+11 4.8846e+11   975.85 < 2.2e-16 ***
NbhdGrp1        1 3.3390e+11 3.3390e+11   667.07 < 2.2e-16 ***
NbhdGrp2        1 2.2547e+11 2.2547e+11   450.44 < 2.2e-16 ***
NbhdGrp3        1 2.0255e+11 2.0255e+11   404.65 < 2.2e-16 ***
Residuals    1994 9.9809e+11 5.0055e+08
```

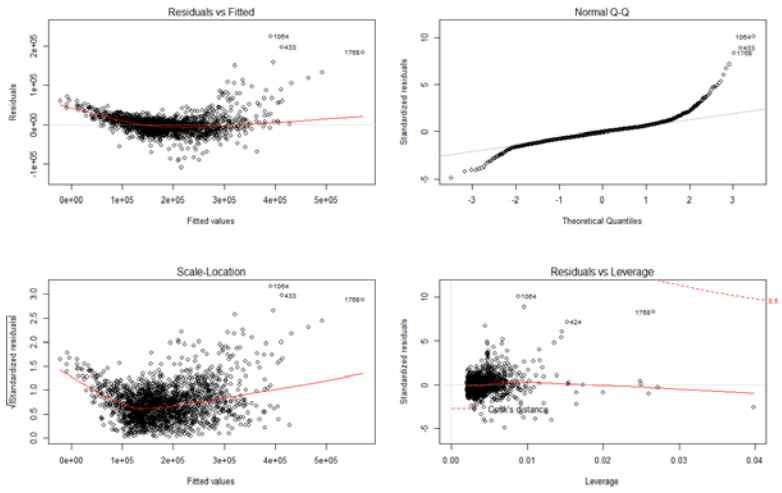
Observations: Figure 36 shows an ANOVA for SalePrice ~ TotalFloorSF + OverallQual + GarageCars + TotalBsmtSF + NbhdGrp1 + NbhdGrp2 + NbhdGrp3. The new continuous variables of GarageCars and TotalBsmtSF were added into this model based on the correlation matrix in figure 4, which showed a moderate positive correlation with SalePrice. A statistically significant result was obtained overall as indicated by the F-statistic which is 2787 with a p-value = < 2.2e-16. This indicates the model has produced statistically significant results to be investigated. The AVOVA tables shows that NbhdGrp1, NbhdGrp2 and NbhdGrp3 all have significant difference when compared to the Others group.

Figure 37: Multiple Linear Regression Model SalePrice ~ TotalFloorSF + OverallQual + GarageCars + TotalBsmtSF + NbhdGrp1 + NbhdGrp2 + NbhdGrp3 .

```
Call:
lm(formula = SalePrice ~ TotalFloorSF + OverallQual + GarageCars +
    TotalBsmtSF + NbhdGrp1  + NbhdGrp2  + NbhdGrp3 , data = subdat)

Residuals:
    Min      1Q  Median      3Q     Max
-109125  -11651    -623    8592  224435

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -15033.060   3068.749  -4.899 1.04e-06 ***
TotalFloorSF     99.189      1.707  58.109  < 2e-16 ***
OverallQual    9111.117    605.366  15.051  < 2e-16 ***
GarageCars     3775.728    949.302   3.977 7.22e-05 ***
TotalBsmtSF      22.944      1.622  14.144  < 2e-16 ***
NbhdGrp1      -75755.962   1962.025 -38.611  < 2e-16 ***
NbhdGrp2      -47887.151   1655.520 -28.926  < 2e-16 ***
NbhdGrp3      -30555.890   1518.994 -20.116  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22370 on 1994 degrees of freedom
Multiple R-squared:  0.9073,   Adjusted R-squared:  0.9069
F-statistic:  2787 on 7 and 1994 DF,  p-value: < 2.2e-16
```

Observations: Figure 37 shows a summary of the Multiple Linear Regression Model SalePrice ~ TotalFloorSF + OverallQual + GarageCars + TotalBsmtSF + NbhdGrp1 + NbhdGrp2 + NbhdGrp3. The equation of the regression line is: SalePrice = -15033.060 + 99.189* TotalFloorSF + 9111.117* OverallQual + 3775.728* GarageCars+ 22.944*TotalBsmtSF-75755.962*NbhdGrp1-47887.151*NbhdGrp2 - 30555.890*NbhdGrp3. The results suggest that when NbhdGrp1 is compared to the Other houses, NbhdGrp1 homes on average, have a SalePrice of $75755.962 less and that it is significant. Furthermore, when NbhdGrp2 is compared to the Other houses, NbhdGrp2 homes on average, have a SalePrice of 47887.151 less and that it is significant. Lastly, when NbhdGrp3 is compared to the Other houses, NbhdGrp3 homes on average, have a SalePrice of $30555.890 less and that it is significant. The residual standard error of 22370, shows us that when predicting SalePrice, one standard error = $22370. The multiple R-squared value of 0.9073 indicates that 90.73% of the variation in SalePrice is explained by the predictor variables.

Figure 38: Scatterplots with Residuals & QQ-Plot of Residuals



Observations: Figure 38, shows scatterplots with residuals and qq-plots of residuals so that we can check to make sure the model is meeting all the assumptions. The QQ plot reveals that the density distribution is non-normal. This is present in the plot where some of the data points are progressively departing from the line in the upper right hand corner of the plot. This indicates non-normality and shows us that it does not correspond relatively well to a standard normal distribution. The scatterplot of residuals vs. fitted shows us that there is "funnel shaped" pattern with heteroscedasticity and a few outliers. By comparison, a healthy normal probability plot of the residuals would be relatively linear and would have a random scatter of data over the range of values for the independent variable. In addition, the residual vs. leverage plot shows that there are some influential points on the right side of the graph. It's important to note that it is highly desirable for the residuals to conform to a normal distribution with few to no outliers. As a result, in order to correct the problems of non-constant variance, non-normality, and influential points we are going to transform SalePrice by creating a new variable called logSalePrice.

Figure 39: Predictions: MLR Model SalePrice ~ TotalFloorSF + OverallQual + GarageCars + TotalBsmtSF + NbhdGrp1 + NbhdGrp2 + NbhdGrp3 .

	fit	lwr	upr
1	205665.80	161735.03	249596.6
2	95521.02	51573.66	139468.4
3	175168.18	131218.11	219118.2
4	266109.60	222070.30	310148.9
5	173057.87	129115.70	217000.0
6	196974.63	153041.79	240907.5

Observations: Figure 39 shows us that the predicted value of the first house is $205665.80. Additionally, the lower and upper confidence bands shows $161735.03 and $249596.6, respectively. This means that the 95% confidence band on this predicted value is $161735.03 and $249596.6.

Figure 39B: Detecting Multicollinearity Using Variance Inflation Factors

```
TotalFloorSF   OverallQual    GarageCars   TotalBsmtSF NbhdGrp1
     1.796233      2.235166      1.716518      1.456676      1.014764
NbhdGrp2   NbhdGrp3
     1.045863      1.019533
```

Observations: Figure 39B shows us the VIF for the all the predictors in the model. A VIF of 1 would mean that no multicolinearity exists at all, while a large VIF number (e.g., 10) would indicate serious multicolinearity issues. As a result, since the VIF for all the predictors above are low, this concludes that we don't have serious multicolinearity issues.

Section 5.2: Log SalePrice Model

Figure 40: Analysis of Variance for L_SalePrice ~ TotalFloorSF + OverallQual + GarageCars + TotalBsmtSF + NbhdGrp1 + NbhdGrp2 + NbhdGrp3 .

```
Analysis of Variance Table

Response: L_SalePrice
               Df  Sum Sq Mean Sq  F value    Pr(>F)
TotalFloorSF    1 172.778 172.778 16777.95 < 2.2e-16 ***
OverallQual     1  52.334  52.334  5081.97 < 2.2e-16 ***
GarageCars      1   7.639   7.639   741.83 < 2.2e-16 ***
TotalBsmtSF     1   9.273   9.273   900.43 < 2.2e-16 ***
NbhdGrp1         1  12.981  12.981  1260.52 < 2.2e-16 ***
NbhdGrp2         1   4.063   4.063   394.52 < 2.2e-16 ***
NbhdGrp3         1   2.627   2.627   255.15 < 2.2e-16 ***
Residuals    1994  20.534   0.010
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

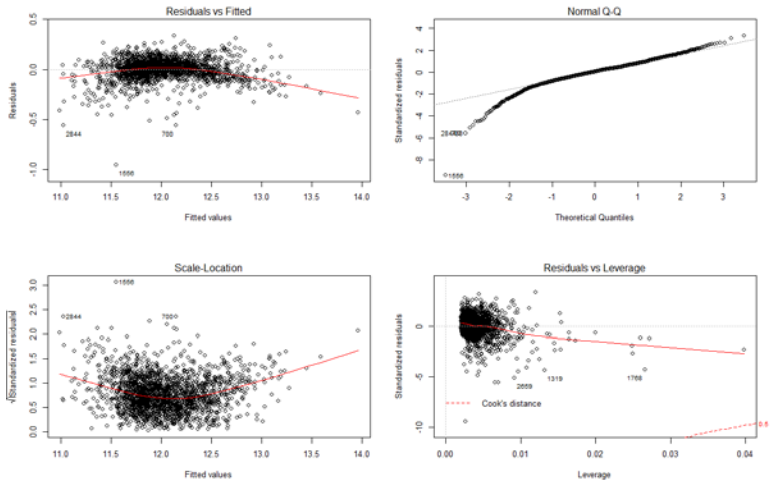
Observations: Figure 40 shows an ANOVA for L_SalePrice ~ TotalFloorSF + OverallQual + GarageCars + TotalBsmtSF + NbhdGrp1 + NbhdGrp2 + NbhdGrp3. A statistically significant result was obtained overall as indicated by the F-statistic which is 3630 with a p-value = < 2.2e-16. This indicates the model has produced statistically significant results to be investigated. The AVOVA tables shows that NbhdGrp1, NbhdGrp2, NbhdGrp3 all have significant difference when compared to the Others group.

Figure 41: Multiple Linear Regression Model L_SalePrice ~ TotalFloorSF + OverallQual + GarageCars + TotalBsmtSF + NbhdGrp1 + NbhdGrp2 + NbhdGrp3 .

```
Call:
lm(formula = L_SalePrice ~ TotalFloorSF + OverallQual + GarageCars +
    TotalBsmtSF + NbhdGrp1  + NbhdGrp2  + NbhdGrp3 , data = subdat)

Residuals:
     Min       1Q   Median       3Q      Max
-0.95383 -0.04997  0.00767  0.05967  0.33814

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.096e+01  1.392e-02 787.615   <2e-16 ***
TotalFloorSF  4.798e-04  7.742e-06  61.965   <2e-16 ***
OverallQual   5.905e-02  2.746e-03  21.504   <2e-16 ***
GarageCars    3.878e-02  4.306e-03   9.007   <2e-16 ***
TotalBsmtSF   9.103e-05  7.358e-06  12.372   <2e-16 ***
NbhdGrp1      -3.784e-01  8.899e-03 -42.518   <2e-16 ***
NbhdGrp2      -1.910e-01  7.509e-03 -25.433   <2e-16 ***
NbhdGrp3      -1.101e-01  6.890e-03 -15.973   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1015 on 1994 degrees of freedom
Multiple R-squared:  0.9272,   Adjusted R-squared:  0.927
F-statistic:  3630 on 7 and 1994 DF,  p-value: < 2.2e-16
```

Observations: Figure 41 shows a summary of the Multiple Linear Regression Model L_SalePrice ~ TotalFloorSF + OverallQual + GarageCars + TotalBsmtSF + NbhdGrp1 + NbhdGrp2 + NbhdGrp3. The equation of the regression line is: L_SalePrice = 1.096e+01+ 4.798e-04* TotalFloorSF + 5.905e-02* OverallQual + 3.878e-02* GarageCars+ 9.103e-05*TotalBsmtSF- 3.784e-01 *NbhdGrp1 - 1.910e-01*NbhdGrp2 -1.101e-01*NbhdGrp3. The results suggest that when NbhdGrp1 is compared to the Other houses, NbhdGrp1 homes on average, have a SalePrice of $38 less and that it is significant. Furthermore, when NbhdGrp2 is compared to the Other houses, NbhdGrp2 homes on average, have a SalePrice of $19 less and that it is significant. Lastly, when NbhdGrp3 is compared to the Other houses, NbhdGrp3 homes on average, have a SalePrice of $10 less and that it is significant. This shows that the interpretation of the log(SalePrice) model is different from the price model because you have to multiply the coefficient by 10 to the 2nd or 4th power in order to get the dollar amount. Additionally, when comparing the residual standard error, it shows 0.1015, which is different than the SalePrice Model. The multiple R-squared value of 0.9272 indicates that 92.72% of the variation in SalePrice is explained by the predictor variables. This multiple r-squared is higher than the Sale Price model, which means that the model fits better after transformation was applied.

Figure 42: Scatterplots with Residuals & QQ-Plot of Residuals



Observations: Figure 42, shows scatterplots with residuals and qq-plots of residuals so that we can check to make sure the model is meeting all the assumptions. After transformation, QQ plot reveals that the density distribution is close to normal since most of the dots are on the line. By comparison, in the SalePrice Model, we saw the assumption of normality being violated when the dots were drastically departing from the line as seen in figure 38, QQ plot. The scatterplot of residuals vs. fitted shows us that the normal probability plot of the residuals appear have a random scatter of data over the range of values for the independent variable, but linearity can be improved. By comparison, in the SalePrice Model, we saw a "bowl" shaped pattern with heteroscedasticity, which is seen in figure 38. As a result, given the assumptions of a regression analysis and the revelations/results from above, the transformed log SalePrice model appears to fit the data better than the SalePrice model, though there still seems to be outliers/influential points, slight non-normality, and non-linearity that needs to be addressed by possible transformations of the predictor variables. In conclusion, given that the r-squared was higher and most of the assumptions improved, the log SalePrice model fits better than the SalePrice model.

Section 5.3: Comparison and Discussion of Model Fits

Figure 36: Analysis of Variance for SalePrice ~ TotalFloorSF + OverallQual + GarageCars + TotalBsmtSF + NbhdGrp1 + NbhdGrp2 + NbhdGrp3 .

```
Analysis of Variance Table

Response: SalePrice
               Df     Sum Sq    Mean Sq  F value    Pr(>F)
TotalFloorSF    1 6.5852e+12 6.5852e+12 13156.03 < 2.2e-16 ***
OverallQual     1 1.7079e+12 1.7079e+12  3412.01 < 2.2e-16 ***
GarageCars      1 2.2170e+11 2.2170e+11   442.91 < 2.2e-16 ***
TotalBsmtSF     1 4.8846e+11 4.8846e+11   975.85 < 2.2e-16 ***
NbhdGrp1        1 3.3390e+11 3.3390e+11   667.07 < 2.2e-16 ***
NbhdGrp2        1 2.2547e+11 2.2547e+11   450.44 < 2.2e-16 ***
NbhdGrp3        1 2.0255e+11 2.0255e+11   404.65 < 2.2e-16 ***
Residuals    1994 9.9809e+11 5.0055e+08
```

Figure 40: Analysis of Variance for L_SalePrice ~ TotalFloorSF + OverallQual + GarageCars + TotalBsmtSF + NbhdGrp1 + NbhdGrp2 + NbhdGrp3

```
Analysis of Variance Table

Response: L_SalePrice
               Df  Sum Sq Mean Sq  F value    Pr(>F)
TotalFloorSF    1 172.778 172.778 16777.95 < 2.2e-16 ***
OverallQual     1  52.334  52.334  5081.97 < 2.2e-16 ***
GarageCars      1   7.639   7.639   741.83 < 2.2e-16 ***
TotalBsmtSF     1   9.273   9.273   900.43 < 2.2e-16 ***
NbhdGrp1        1  12.981  12.981  1260.52 < 2.2e-16 ***
NbhdGrp2        1   4.063   4.063   394.52 < 2.2e-16 ***
NbhdGrp3        1   2.627   2.627   255.15 < 2.2e-16 ***
Residuals    1994  20.534   0.010
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Comparison/Discussion: Both models were statistically significant as indicated by the F-statistic with a p-value = < 2.2e-16. This indicates that both models produced statistically significant results to be investigated. In Figure 36 & 40, the AVOVA tables shows that NbhdGrp1, NbhdGrp2 and NbhdGrp3 all have significant difference when compared to the Others group.

Figure 37: Multiple Linear Regression Model SalePrice ~ TotalFloorSF + OverallQual + GarageCars + TotalBsmtSF + NbhdGrp1 + NbhdGrp2 + NbhdGrp3 .

```
Call:
lm(formula = SalePrice ~ TotalFloorSF + OverallQual + GarageCars +
    TotalBsmtSF + NbhdGrp1 + NbhdGrp2 + NbhdGrp3, data = subdat)

Residuals:
    Min      1Q  Median      3Q     Max
-109125  -11651    -623    8592  224435

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -15033.060   3068.749  -4.899 1.04e-06 ***
TotalFloorSF     99.189      1.707  58.109  < 2e-16 ***
OverallQual    9111.117    605.366  15.051  < 2e-16 ***
GarageCars     3775.728    949.302   3.977 7.22e-05 ***
TotalBsmtSF      22.944      1.622  14.144  < 2e-16 ***
NbhdGrp1     -75755.962   1962.025 -38.611  < 2e-16 ***
NbhdGrp2     -47887.151   1655.520 -28.926  < 2e-16 ***
NbhdGrp3     -30555.890   1518.994 -20.116  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22370 on 1994 degrees of freedom
Multiple R-squared:  0.9073,   Adjusted R-squared:  0.9069
F-statistic:  2787 on 7 and 1994 DF,  p-value: < 2.2e-16
```

Figure 41: Multiple Linear Regression Model L_SalePrice ~ TotalFloorSF + OverallQual + GarageCars + TotalBsmtSF + NbhdGrp1 + NbhdGrp2 + NbhdGrp3

```
Call:
lm(formula = L_SalePrice ~ TotalFloorSF + OverallQual + GarageCars +
    TotalBsmtSF + NbhdGrp1 + NbhdGrp2 + NbhdGrp3, data = subdat)

Residuals:
     Min       1Q   Median       3Q      Max
-0.95383 -0.04997  0.00767  0.05967  0.33814

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.096e+01  1.392e-02 787.615   <2e-16 ***
TotalFloorSF 4.798e-04  7.742e-06  61.965   <2e-16 ***
OverallQual  5.905e-02  2.746e-03  21.504   <2e-16 ***
GarageCars   3.878e-02  4.306e-03   9.007   <2e-16 ***
TotalBsmtSF  9.103e-05  7.358e-06  12.372   <2e-16 ***
NbhdGrp1    -3.784e-01  8.899e-03 -42.518   <2e-16 ***
NbhdGrp2    -1.910e-01  7.509e-03 -25.433   <2e-16 ***
NbhdGrp3    -1.101e-01  6.890e-03 -15.973   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1015 on 1994 degrees of freedom
Multiple R-squared:  0.9272,   Adjusted R-squared:  0.927
F-statistic:  3630 on 7 and 1994 DF,  p-value: < 2.2e-16
```

Comparison/Discussion: In comparison, the log(SalePrice) model fits better than the SalePrice model. This was evident when comparing their multiple R-squared values of 0.9272 (logSalePrice model) versus Saleprice model of 0.9073. The improvement in the multiple R-squared value for the log(SalePrice) model shows that the transformation of the response to log(SalePrice) improved the model fit.

Figure 38: Scatterplots with Residuals & QQ-Plot of Residuals for SalePrice Model

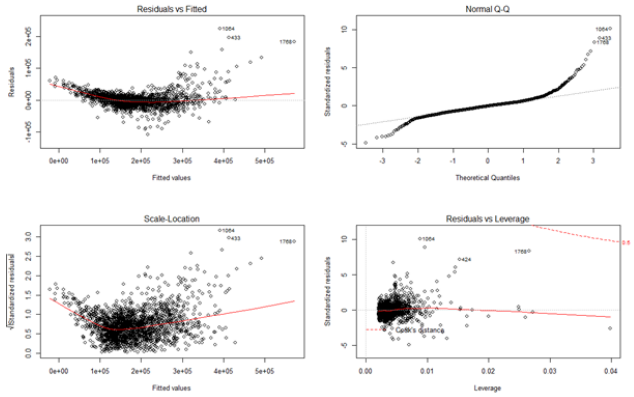
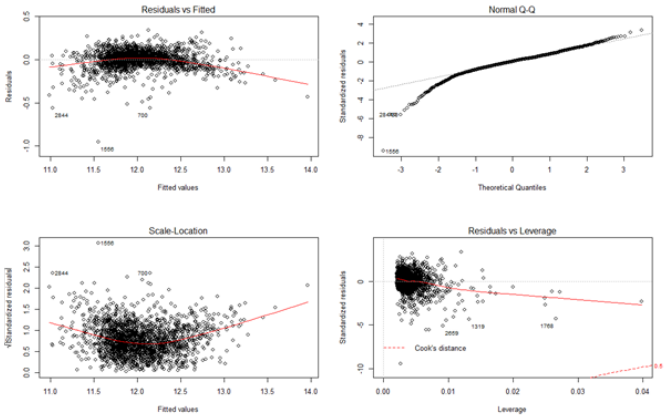


Figure 42: Scatterplots with Residuals & QQ-Plot of Residuals for log(SalePrice) Model



Comparison/Discussion: Figure 42 shows scatterplots with residuals and qq-plots of residuals so that we can check to make sure the model is meeting all the assumptions. After transformation, the QQ plot reveals that the density distribution is close to normal since most of the dots are on the line. By comparison, in the SalePrice Model, we saw the assumption of normality being violated when the dots were drastically departing from the line as seen in figure 38, QQ plot. The scatterplot of residuals vs. fitted shows us that the normal probability plot of the residuals appear to have a random scatter of data over the range of values for the independent variable, but linearity can be improved. By comparison, in the SalePrice Model, we saw a "bowl" shaped pattern with heteroscedasticity, which is seen in figure 38. In conclusion, given that the r-squared was higher and all of the assumptions of normality, linearity, and homoscedasticity improved, the log SalePrice model fits better than the SalePrice model. This also illustrates that the primary reason why we do log or square root transformations is to improve the model assumptions. However, it's important to note that there still seems to be outliers/influential points, slight non-normality, and slight non-linearity that needs to be addressed. As a result, we should consider possible transformations of the predictor variables for the log(SalePrice) model. This will be shown below.

Additional Transformed Model vs. log(SalePrice) Model

Figure 40: Analysis of Variance for L_SalePrice ~ TotalFloorSF + OverallQual + GarageCars + TotalBsmtSF + NbhdGrp1 + NbhdGrp2 + NbhdGrp3 .

```
Analysis of Variance Table

Response: L_SalePrice
               Df  Sum Sq Mean Sq  F value    Pr(>F)
TotalFloorSF    1 172.778 172.778 16777.95 < 2.2e-16 ***
OverallQual     1  52.334  52.334  5081.97 < 2.2e-16 ***
GarageCars      1   7.639   7.639   741.83 < 2.2e-16 ***
TotalBsmtSF     1   9.273   9.273   900.43 < 2.2e-16 ***
NbhdGrp1        1  12.981  12.981  1260.52 < 2.2e-16 ***
NbhdGrp2        1   4.063   4.063   394.52 < 2.2e-16 ***
NbhdGrp3        1   2.627   2.627   255.15 < 2.2e-16 ***
Residuals    1994  20.534   0.010
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 43: Analysis of Variance for L_SalePrice ~ L_TotalFloorSF + OverallQual + GarageCars + TotalBsmtSF + NbhdGrp1 + NbhdGrp2 + NbhdGrp3 .

```
Analysis of Variance Table

Response: L_SalePrice
                Df  Sum Sq Mean Sq  F value    Pr(>F)
L_TotalFloorSF   1 176.697 176.697 22691.99 < 2.2e-16 ***
OverallQual      1  48.024  48.024  6167.39 < 2.2e-16 ***
GarageCars       1   7.625   7.625   979.28 < 2.2e-16 ***
TotalBsmtSF      1   9.349   9.349  1200.67 < 2.2e-16 ***
NbhdGrp1         1  14.566  14.566  1870.62 < 2.2e-16 ***
NbhdGrp2         1   6.141   6.141   788.62 < 2.2e-16 ***
NbhdGrp3         1   4.299   4.299   552.15 < 2.2e-16 ***
Residuals     1994  15.527   0.008
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Comparison/Discussion: Both models were statistically significant as indicated by the F-statistic with a p-value = < 2.2e-16. This indicates that both models produced statistically significant results to be investigated. Figure 40 & 43, the AVOVA tables shows that NbhdGrp1, NbhdGrp2, and NbhdGrp3 all have significant difference when compared to the Others group.

Figure 41: Multiple Linear Regression Model L_SalePrice ~ TotalFloorSF + OverallQual + GarageCars + TotalBsmtSF + NbhdGrp1 + NbhdGrp2 + NbhdGrp3 .

```
Call:
lm(formula = L_SalePrice ~ TotalFloorSF + OverallQual + GarageCars +
    TotalBsmtSF + NbhdGrp1 + NbhdGrp2 + NbhdGrp3, data = subdat)

Residuals:
     Min       1Q   Median       3Q      Max
-0.95383 -0.04997  0.00767  0.05967  0.33814

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.096e+01  1.392e-02 787.615   <2e-16 ***
TotalFloorSF  4.798e-04  7.742e-06  61.965   <2e-16 ***
OverallQual   5.905e-02  2.746e-03  21.504   <2e-16 ***
GarageCars    3.878e-02  4.306e-03   9.007   <2e-16 ***
TotalBsmtSF   9.103e-05  7.358e-06  12.372   <2e-16 ***
NbhdGrp1     -3.784e-01  8.899e-03 -42.518   <2e-16 ***
NbhdGrp2     -1.910e-01  7.509e-03 -25.433   <2e-16 ***
NbhdGrp3     -1.101e-01  6.890e-03 -15.973   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1015 on 1994 degrees of freedom
Multiple R-squared:  0.9272,   Adjusted R-squared:  0.927
F-statistic:  3630 on 7 and 1994 DF,  p-value: < 2.2e-16
```

Figure 44: Multiple Linear Regression Model L_SalePrice ~ L_TotalFloorSF + OverallQual + GarageCars + TotalBsmtSF + NbhdGrp1 + NbhdGrp2 + NbhdGrp3 .

```
Call:
lm(formula = L_SalePrice ~ L_TotalFloorSF + OverallQual + GarageCars +
    TotalBsmtSF + NbhdGrp1 + NbhdGrp2 + NbhdGrp3, data = subdat)

Residuals:
     Min       1Q   Median       3Q      Max
-1.00259 -0.04241  0.00161  0.04975  0.34179

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.844e+00  6.483e-02  90.147  < 2e-16 ***
L_TotalFloorSF 8.333e-01  1.102e-02  75.637  < 2e-16 ***
OverallQual    3.966e-02  2.491e-03  15.920  < 2e-16 ***
GarageCars     2.413e-02  3.787e-03   6.373 2.29e-10 ***
TotalBsmtSF    6.485e-05  6.470e-06  10.023  < 2e-16 ***
NbhdGrp1      -4.558e-01  8.212e-03 -55.506  < 2e-16 ***
NbhdGrp2      -2.494e-01  6.824e-03 -36.546  < 2e-16 ***
NbhdGrp3      -1.431e-01  6.092e-03 -23.498  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08824 on 1994 degrees of freedom
Multiple R-squared:  0.945,    Adjusted R-squared:  0.9448
F-statistic:  4893 on 7 and 1994 DF,  p-value: < 2.2e-16
```

Comparison/Discussion: After TotalFloorSF was transformed to L_TotalFloorSF, it appears that the multiple R-squared improved from 0.9272 to 0.945. Residual error in figure 44 is also lower than figure 41.

Figure 42: Scatterplots with Residuals & QQ-Plot of Residuals for log(SalePrice) Model

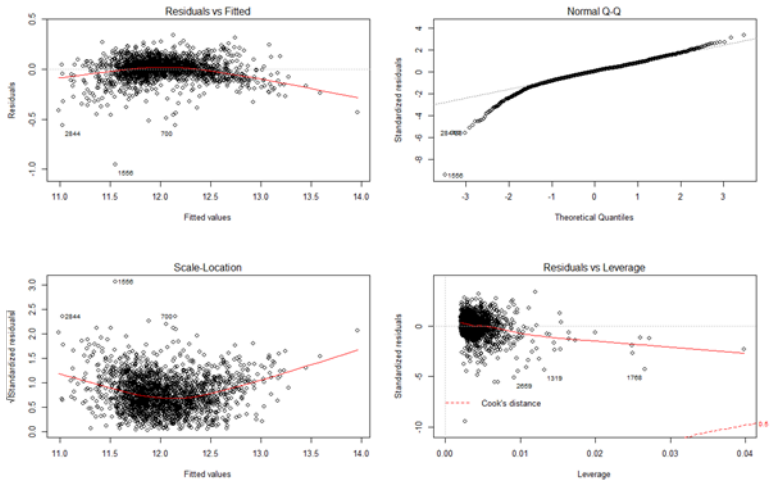
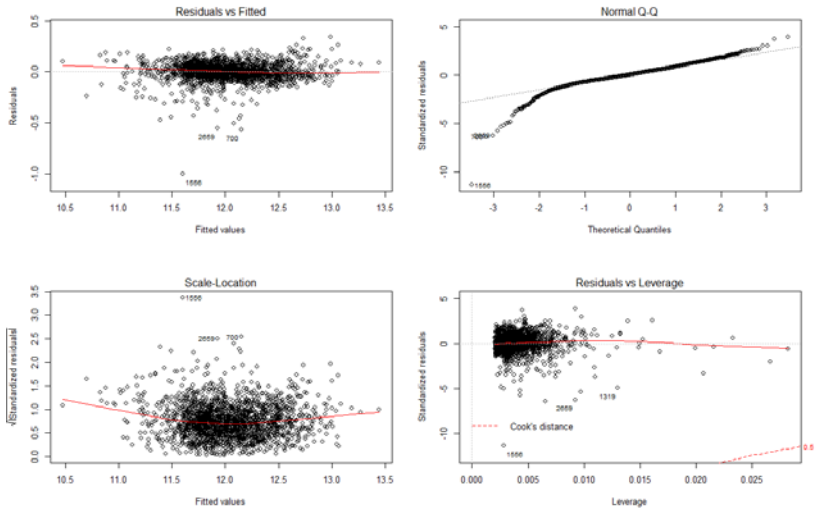


Figure 45: Scatterplots with Residuals & QQ-Plot of Residuals for log(SalePrice) and log(L_TotalFloorSF) Model



Comparison/Discussion: After TotalFloorSF was transformed to L_TotalFloorSF, it appears that the the QQ plot has somewhat improved since the dots have moved closer to the line, indicating that the density distribution has gotten closer to normal compared to Figure 42. Additionally, the scatterplot of residuals vs. fitted shows us that the normal probability plot of the residuals appear to be more linear and have more of random scatter of data over the range of values for the independent variable compared to Figure 42. In conclusion, given that the assumptions of normality, linearity, and homoscedasticity improved, the new model fits better than the L_SalePrice model. However, it's important to note that there still seems to be outliers/influential points and slight non-normality that needs to be addressed. As a result, we should consider additional transformations of the predictor variables, outlier deletion/down weighing them in the future.

Section 6: Summary/Conclusions

In section 1, we defined the sample population/data of interest for 'typical' homes in Ames, Iowa to be 'single-family' homes with 'normal' sales in Ames, Iowa using drop conditions and boxplots. In section 2, a correlation matrix and scatterplots were created and after an analysis, OverallQual (0.8) and TotalFloorSF (0.78) were chosen as the two predictor variables with the most promise for predicting SalePrice due to their strong positive correlations. A simple linear regression models of SalePrice ~ TotalFloorSF and SalePrice ~ OverallQual was then created and relevant diagnostic plots (e.g., ANOVA, summary tables, predictor error, and multiple-r-squared) were created to assess goodness-of-fit of each model. The results showed that both models were significant, but mediocre r-squared values for SalePrice ~ TotalFloorSF (r-squared: 0.6118) and SalePrice ~ OverallQual (r-squared: 0.6452) were shown and large predicted error and "wide" bands existed. Additionally, both models did not fit the data really well because the assumptions were violated: non-normality, non-linearity, heteroscedasticity or non-constant variance, and outliers/influential points were evident.

In section 3, the predictor variables of TotalFloorSF and OverallQual were combined to create a multiple linear regression model: SalePrice ~ TotalFloorSF+OverallQual. Relevant diagnostic plots were conducted to assess the goodness-of-fit of the model. The results showed that the model was significant and that the model fits better than the simple linear regression models since the adjusted r-squared for this model is 0.7703 compared to model #1 (adjusted r-squared: 0.6116) and model #2 (adjusted r-squared: 0.6451). Additionally, predicted error was smaller and tighter than model #1 and model #2. Furthermore, the VIF showed that there were no multicolinearity issues. However, although this model fit better than model #1 and #2, the assumptions were still violated: non-normality, non-linearity, heteroscedasticity or non-constant variance, and outliers/influential points were evident.

In section 4, a boxplot of the residuals by neighborhood was created. The results showed that CollgCr, Crawford, Edwards, North Ames, Northwest Ames, and Somerset are better fit by the model. Clear Creek, Northridge Heights, Northridge, and Stone Brook are some neighborhoods that over predicted. Gilbert, Iowa DOT and Rail Road, Old Town, and South & West of Iowa State University are some that are underpredicted.

Additionally, the Plot of Mean MAE and the Mean Price per Square Foot showed that there is a strong positive correlation between MAE and mean price per square foot (as mean price per square foot increases, MAE increases). This means that as price per square foot increases, the forecasts or predictions are to the eventual outcomes becomes worse or more inaccurate. Next, the neighborhoods were grouped by price per square foot and dummy variables were created to be included in the multiple regression model. The results showed that the MAE of New Multiple Regression Model (15122.9) appeared to fit better than Model #3 (25151.48) since 15122.9 is less than 25383.46. This makes sense since Adjusted R-squared: 0.8968 was higher than model #3: 0.7703. Additionally, it's interesting to note that the predicted error of $23570 was also lower than model #3: $35150.

In section 5, a transformation of the response variable from the sale price to the natural logarithm of the sale price was conducted. Two models were created: L_SalePrice and SalePrice ~ TotalFloorSF + OverallQual + GarageCars +TotalBsmtSF + NbhdGrp1 + NbhdGrp2 + NbhdGrp3.

The results showed that the log(SalePrice) model fits better than the SalePrice model. This was evident when comparing their multiple R-squared values of 0.9272 (logSalePrice model) versus Saleprice model of

0.9073. The improvement in the multiple R-squared value for the log(SalePrice) model shows that the transformation of the response to log(SalePrice) improved the model fit. Additionally, most of the assumptions such as normality and homoscedasticity improved, which illustrates that the primary reason why we do log or square root transformations is to improve the model assumptions. However, it's important to note that there were still outliers/influential points, slight non-normality, and non-linearity that needed to be addressed. As a result, we then considered possible transformations of the predictor variables for the log(SalePrice) model.

In the end, we created another model: L_SalePrice ~ L_TotalFloorSF + OverallQual + GarageCars + TotalBsmtSF + NbhdGrp1 + NbhdGrp2 + NbhdGrp3 . In this model, we transformed TotalFloorSF to L_TotalFloorSF. The results showed that the r-squared improved in comparison to log(SalePrice) model. Addtionally, the assumptions of normality, linearity, and homoscedasticity all improved. However, it's important to note that there still seems to be outliers/influential points and slight non-normality that needs to be addressed. As a result, we should consider additional transformations of the predictor variables, outlier deletion/down weighing them in the future.

In conclusion, variable transformation and outlier deletion impacts the modeling process and the results because it helps improve model fit and the model itself by making the model assumptions truer than before. For instance, a healthy normal probability plot of the residuals would be relatively linear and would have a random scatter of data over the range of values for the independent variable. It's also important and highly desirable for the residuals to conform to a normal distribution with few to no outliers. As a result, conducting variable transformation and outlier deletion improves the modeling assumptions of normality, linearity, and homoscedasticity. For instance, if the results of the diagnostic plots show heteroskedasticity (e.g., the residual plot "flares-out" in a funnel pattern as x gets larger, which is a violation of the assumption of constant variance for error terms), non-normality (e.g., some data points in the QQ plot progressively depart from the line), outliers, influential points etc., variable transformation and outlier deletion can help improve the model assumptions. In the end, these analytical activities are a benefit. However, we should not throw out outliers or downweigh them automatically. Instead, we need to examine them carefully and decide corrective actions.

Lastly, in terms of the next steps we need to continue with "model adequacy checking". This includes addressing the outliers/influential points that we detected and possibly conduct more variable transformations to improve the slight non-normality in our last model. It may also be beneficial to include more continuous and categorical predictor variables as well in the future. After all these are conducted, we can then move to model validation to see if the model is going to satisfy business needs/requirements. This entails determining if the model will be behave or function at it was intended in the operating environment. For instance, we can assess the predictive accuracy of our model using data splitting (aka cross validation), which divides the data into two parts: estimation data (training data, 70%) and prediction data (test data, 30%). If the model validation process is satisfied, we can then move on to model use. However, if any of these assumptions are not satisfied, we will need to go back to model specification, parameter estimation, and continue the process.

APPENDIX

```
Potentially influential observations of
        lm(formula = SalePrice ~ TotalFloorSF, data = subdat) :

     dfb.1_ dfb.TFSF dffit   cov.r   cook.d hat
16   -0.23   0.27     0.28_*  1.00    0.04   0.01_*
38    0.00   0.02     0.05    1.00_*  0.00   0.00
47   -0.18   0.22     0.23_*  0.99_*  0.03   0.00_*
60   -0.02   0.03     0.03    1.00_*  0.00   0.00_*
63   -0.01   0.02     0.02    1.00_*  0.00   0.00_*
66   -0.04   0.05     0.05    1.01_*  0.00   0.01_*
128   0.07  -0.09    -0.10_*  1.00    0.01   0.00
161   0.14  -0.16    -0.17_*  1.00    0.01   0.01_*
246   0.02  -0.01     0.05    1.00_*  0.00   0.00
254   0.09  -0.11    -0.11_*  1.01_*  0.01   0.01_*
292   0.00  -0.02    -0.05    1.00_*  0.00   0.00
293   0.09  -0.12    -0.13_*  1.00_*  0.01   0.00
297  -0.03   0.04     0.07    1.00_*  0.00   0.00
303   0.01  -0.01     0.01    1.00_*  0.00   0.00
322  -0.05   0.07     0.10_*  0.99_*  0.00   0.00
344  -0.02   0.03     0.03    1.00_*  0.00   0.00_*
348  -0.01   0.04     0.09    0.99_*  0.00   0.00
377   0.07  -0.08    -0.09    1.00    0.00   0.00_*
380   0.02  -0.02    -0.02    1.00_*  0.00   0.00
421  -0.01   0.03     0.07    0.99_*  0.00   0.00
422  -0.14   0.17     0.18_*  1.00_*  0.02   0.00_*
424  -0.19   0.25     0.28_*  0.97_*  0.04   0.00
425   0.00   0.02     0.07    0.99_*  0.00   0.00
430  -0.06   0.09     0.12_*  0.99_*  0.01   0.00
432  -0.11   0.14     0.17_*  0.99_*  0.01   0.00
433  -0.29   0.35     0.38_*  0.96_*  0.07   0.00_*
435  -0.01   0.04     0.08    0.99_*  0.00   0.00
439  -0.01   0.03     0.07    0.99_*  0.00   0.00
441   0.01   0.00     0.06    1.00_*  0.00   0.00
442  -0.02   0.03     0.06    1.00_*  0.00   0.00
448  -0.07   0.09     0.10_*  1.00    0.01   0.00
449  -0.12   0.15     0.17_*  0.99_*  0.01   0.00
450   0.00   0.00     0.01    1.00_*  0.00   0.00
458  -0.06   0.09     0.10_*  1.00_*  0.01   0.00
496  -0.08   0.09     0.10_*  1.00    0.01   0.00_*
497  -0.01   0.01     0.01    1.00_*  0.00   0.00_*
499  -0.02   0.03     0.03    1.00_*  0.00   0.00_*
502   0.05  -0.06    -0.07    1.00    0.00   0.00_*
505  -0.08   0.11     0.12_*  1.00    0.01   0.00
514  -0.14   0.17     0.19_*  0.99_*  0.02   0.00
575   0.00   0.02     0.06    1.00_*  0.00   0.00
586   0.05  -0.06    -0.07    1.00    0.00   0.00_*
630   0.01  -0.03    -0.06    1.00_*  0.00   0.00
663  -0.02   0.01    -0.02    1.00_*  0.00   0.00
700   0.12  -0.15    -0.18_*  0.99_*  0.02   0.00
716   0.07  -0.10    -0.12_*  0.99_*  0.01   0.00
717   0.04  -0.05    -0.06    1.00_*  0.00   0.00_*
807   0.04  -0.06    -0.08    1.00_*  0.00   0.00
821   0.00   0.01     0.05    1.00_*  0.00   0.00
822  -0.01   0.03     0.07    0.99_*  0.00   0.00
880   0.00   0.01     0.05    1.00_*  0.00   0.00
```

```
892   0.00   0.02     0.06    1.00_*  0.00   0.00
908   0.00   0.00     0.00    1.00_*  0.00   0.00
910   0.12  -0.14    -0.16_*  1.00    0.01   0.00_*
944   0.00   0.00     0.00    1.00_*  0.00   0.00
957  -0.05   0.07     0.10_*  0.99_*  0.00   0.00
958   0.00   0.02     0.06    0.99_*  0.00   0.00
960  -0.06   0.09     0.12_*  0.99_*  0.01   0.00
961   0.00   0.02     0.06    0.99_*  0.00   0.00
969  -0.06   0.09     0.13_*  0.98_*  0.01   0.00
1012 -0.02   0.03     0.03    1.00_*  0.00   0.00
1013 -0.04   0.06     0.10_*  0.99_*  0.01   0.00
1023  0.08  -0.09    -0.10_*  1.00_*  0.00   0.00_*
1028  0.05  -0.06    -0.06    1.00_*  0.00   0.00_*
1057 -0.10   0.12     0.14_*  1.00_*  0.01   0.00
1058 -0.04   0.07     0.10_*  0.99_*  0.01   0.00
1060 -0.14   0.17     0.18_*  1.00_*  0.02   0.00_*
1061 -0.07   0.09     0.10_*  1.00    0.00   0.00
1064 -0.25   0.32     0.36_*  0.95_*  0.06   0.00
1065 -0.09   0.11     0.12_*  1.00    0.01   0.00_*
1068 -0.10   0.12     0.13_*  1.00    0.01   0.00_*
1069 -0.05   0.06     0.07    1.00_*  0.00   0.00_*
1100  0.00   0.00     0.00    1.00_*  0.00   0.00_*
1103 -0.04   0.06     0.08    1.00_*  0.00   0.00
1105  0.01  -0.01    -0.01    1.00_*  0.00   0.00
1158  0.03  -0.03    -0.04    1.00_*  0.00   0.00_*
1159 -0.05   0.05     0.06    1.00_*  0.00   0.00_*
1192  0.06  -0.08    -0.09    1.00    0.00   0.00_*
1200  0.02  -0.03    -0.03    1.00_*  0.00   0.00_*
1252  0.01  -0.03    -0.06    1.00_*  0.00   0.00
1289  0.08  -0.10    -0.10_*  1.00    0.01   0.00_*
1307  0.21  -0.24    -0.24_*  1.00_*  0.03   0.01_*
1314  0.04  -0.07    -0.09    1.00_*  0.00   0.00
1319  0.06  -0.09    -0.11_*  0.99_*  0.01   0.00
1321  0.00   0.00    -0.01    1.00_*  0.00   0.00_*
1323  0.09  -0.11    -0.13_*  1.00    0.01   0.00
1347  0.02  -0.04    -0.07    1.00_*  0.00   0.00
1350  0.00   0.00     0.00    1.00_*  0.00   0.00
1420  0.08  -0.10    -0.10_*  1.00    0.01   0.00_*
1498  0.34  -0.39    -0.40_*  1.00    0.08   0.01_*
1522  0.11  -0.13    -0.14_*  1.00    0.01   0.00_*
1538  0.04  -0.05    -0.05    1.01_*  0.00   0.01_*
1556 -0.04   0.02    -0.06    0.99_*  0.00   0.00
1574  0.01   0.01     0.05    1.00_*  0.00   0.00
1588 -0.04   0.06     0.09    0.99_*  0.00   0.00
1611  0.02  -0.04    -0.07    1.00_*  0.00   0.00
1636 -0.04   0.06     0.10_*  0.99_*  0.00   0.00
1642 -0.10   0.13     0.16_*  0.99_*  0.01   0.00
1694 -0.08   0.11     0.15_*  0.98_*  0.01   0.00
1698 -0.06   0.08     0.08    1.00    0.00   0.00_*
1701 -0.10   0.12     0.12_*  1.00    0.01   0.00_*
1704 -0.02   0.02     0.03    1.00_*  0.00   0.00
1707 -0.02   0.05     0.09    0.99_*  0.00   0.00
1708  0.01  -0.01    -0.01    1.01_*  0.00   0.00_*
1711  0.00   0.00     0.00    1.00_*  0.00   0.00
1762 -0.07   0.09     0.10_*  1.00    0.00   0.00
1764 -0.14   0.18     0.19_*  0.99_*  0.02   0.00_*
1765 -0.06   0.07     0.08    1.01_*  0.00   0.01_*
```

```
1766 -0.01   0.02     0.02    1.00_*  0.00   0.00
1768 -0.63   0.71     0.72_*  0.99_*  0.26   0.02_*
1772  0.01  -0.01    -0.01    1.00_*  0.00   0.00
1773 -0.07   0.08     0.09    1.01_*  0.00   0.01_*
1775 -0.02   0.03     0.03    1.00_*  0.00   0.00
1779  0.02   0.00     0.05    1.00_*  0.00   0.00
1791  0.00   0.02     0.06    1.00_*  0.00   0.00
1805  0.02  -0.02    -0.03    1.00_*  0.00   0.00
1828  0.00   0.00     0.00    1.00_*  0.00   0.00
1833  0.03  -0.04    -0.04    1.00_*  0.00   0.00_*
1834  0.06  -0.08    -0.08    1.00_*  0.00   0.00_*
1902 -0.01   0.01    -0.01    1.00_*  0.00   0.00_*
1996  0.08  -0.10    -0.13_*  0.99_*  0.01   0.00
1998  0.08  -0.10    -0.11_*  1.00    0.01   0.00
2002  0.12  -0.15    -0.17_*  0.99_*  0.01   0.00
2046  0.10  -0.12    -0.13_*  1.00    0.01   0.00_*
2066  0.15  -0.18    -0.20_*  0.99_*  0.02   0.00_*
2084  0.01  -0.01     0.01    1.00_*  0.00   0.00
2093  0.02  -0.03    -0.03    1.00_*  0.00   0.00_*
2151  0.02  -0.02    -0.03    1.00_*  0.00   0.00
2153  0.01  -0.01    -0.02    1.00_*  0.00   0.00
2214  0.07  -0.08    -0.09    1.00    0.00   0.00_*
2219  0.01  -0.01    -0.01    1.00_*  0.00   0.00_*
2231  0.01  -0.01    -0.01    1.00_*  0.00   0.00_*
2273  0.01  -0.02    -0.02    1.00_*  0.00   0.00
2277  0.01  -0.01    -0.01    1.00_*  0.00   0.00
2279  0.04  -0.06    -0.08    1.00_*  0.00   0.00
2323  0.05  -0.06    -0.06    1.00_*  0.00   0.00_*
2342 -0.08   0.12     0.16_*  0.98_*  0.01   0.00
2351  0.09  -0.10    -0.11_*  1.00    0.01   0.00_*
2385 -0.04   0.07     0.12_*  0.98_*  0.01   0.00
2386 -0.02   0.04     0.06    1.00_*  0.00   0.00
2388 -0.01   0.02     0.02    1.00_*  0.00   0.00
2396 -0.04   0.06     0.09    0.99_*  0.00   0.00
2399 -0.02   0.05     0.09    0.99_*  0.00   0.00
2400 -0.02   0.05     0.09    0.99_*  0.00   0.00
2446 -0.37   0.42     0.44_*  0.99_*  0.09   0.01_*
2447 -0.08   0.09     0.10_*  1.00_*  0.00   0.00_*
2450 -0.03   0.03     0.04    1.00_*  0.00   0.00_*
2451 -0.30   0.35     0.36_*  1.00_*  0.06   0.01_*
2452 -0.02   0.02     0.03    1.00_*  0.00   0.00
2453  0.01  -0.01    -0.01    1.00_*  0.00   0.00
2454  0.02  -0.02    -0.02    1.00_*  0.00   0.00_*
2499  0.05  -0.06    -0.07    1.00    0.00   0.00_*
2501  0.00   0.00     0.00    1.01_*  0.00   0.00_*
2523 -0.03   0.06     0.10_*  0.99_*  0.00   0.00
2528  0.06  -0.07    -0.08    1.00    0.00   0.00_*
2562  0.05  -0.07    -0.09    1.00_*  0.00   0.00
2618  0.01  -0.01    -0.01    1.00_*  0.00   0.00
2654  0.00   0.00     0.00    1.00_*  0.00   0.00
2659  0.01  -0.04    -0.07    0.99_*  0.00   0.00
2667 -0.15   0.18     0.19_*  1.00    0.02   0.01_*
2684  0.01  -0.01     0.01    1.00_*  0.00   0.00
2738  0.03  -0.04    -0.04    1.01_*  0.00   0.01_*
2844 -0.03   0.03    -0.03    1.00_*  0.00   0.00
2892 -0.03   0.04     0.04    1.00_*  0.00   0.00
2904 -0.01  -0.01    -0.06    1.00_*  0.00   0.00
```

```
Potentially influential observations of
        lm(formula = SalePrice ~ OverallQual, data = subdat) :

     dfb.1_ dfb.OvrQ dffit   cov.r   cook.d hat
16   -0.17   0.21     0.25_*  0.96_*  0.03   0.00
17    0.07  -0.08    -0.10_*  1.00    0.00   0.00
42    0.04  -0.04    -0.05    1.00_*  0.00   0.00_*
47   -0.19   0.22     0.24_*  0.99_*  0.03   0.00_*
60   -0.02   0.02     0.03    1.00_*  0.00   0.00_*
66   -0.09   0.11     0.13_*  0.99_*  0.01   0.00
131   0.04  -0.03     0.04    1.00_*  0.00   0.00_*
175   0.07  -0.07     0.08    1.00    0.00   0.00_*
229  -0.03   0.04     0.07    1.00_*  0.00   0.00
235   0.09  -0.09     0.10_*  1.00    0.00   0.00_*
254  -0.03   0.04     0.06    1.00_*  0.00   0.00
274   0.05  -0.04     0.05    1.00_*  0.00   0.00_*
288   0.01  -0.01     0.01    1.00_*  0.00   0.00_*
297  -0.03   0.04     0.07    1.00_*  0.00   0.00
303   0.13  -0.12     0.13_*  1.00    0.01   0.00_*
322  -0.07   0.08     0.08    1.00    0.00   0.00_*
348  -0.07   0.08     0.08    1.00    0.00   0.00_*
421   0.03  -0.03    -0.03    1.01_*  0.00   0.00_*
422  -0.15   0.17     0.19_*  0.99_*  0.02   0.00_*
424  -0.28   0.31     0.33_*  0.99_*  0.05   0.00_*
425  -0.01   0.02     0.02    1.00_*  0.00   0.00_*
429  -0.02   0.03     0.03    1.00_*  0.00   0.00_*
430  -0.09   0.11     0.12_*  1.00    0.01   0.00_*
432  -0.13   0.15     0.16_*  1.00    0.01   0.00_*
433  -0.36   0.39     0.42_*  0.97_*  0.09   0.00_*
435  -0.05   0.06     0.06    1.00    0.00   0.00_*
439  -0.02   0.03     0.03    1.00_*  0.00   0.00_*
441   0.01  -0.01    -0.01    1.00_*  0.00   0.00_*
442  -0.01   0.01     0.02    1.00_*  0.00   0.00_*
443  -0.04   0.04     0.05    1.00_*  0.00   0.00_*
448  -0.04   0.05     0.05    1.01_*  0.00   0.00_*
449  -0.14   0.15     0.17_*  0.99_*  0.01   0.00_*
458  -0.07   0.09     0.11_*  1.00_*  0.01   0.00
496  -0.09   0.10     0.13_*  0.99_*  0.01   0.00
505  -0.09   0.10     0.12_*  0.99_*  0.01   0.00
514  -0.16   0.18     0.20_*  0.99_*  0.02   0.00_*
521   0.06  -0.07    -0.08    1.00    0.00   0.00_*
524   0.06  -0.07    -0.07    1.00_*  0.00   0.00_*
554  -0.01   0.00    -0.05    1.00_*  0.00   0.00
655   0.07  -0.06     0.07    1.00    0.00   0.00_*
663   0.09  -0.09     0.09    1.00_*  0.00   0.00_*
666   0.04  -0.04     0.04    1.00_*  0.00   0.00_*
709   0.07  -0.06     0.07    1.01_*  0.00   0.00_*
743   0.09  -0.09     0.09    1.00_*  0.00   0.00_*
766   0.20  -0.20     0.20_*  1.00_*  0.02   0.01_*
781   0.10  -0.09     0.10_*  1.00_*  0.00   0.00_*
787   0.11  -0.11     0.12_*  1.00    0.01   0.00_*
822  -0.02   0.02     0.03    1.00_*  0.00   0.00_*
826   0.07  -0.08    -0.09    1.00    0.00   0.00_*
867   0.00   0.00     0.00    1.00_*  0.00   0.00_*
892  -0.01   0.01     0.01    1.00_*  0.00   0.00_*
899   0.04  -0.04     0.04    1.00_*  0.00   0.00_*
```

```
908   0.09  -0.09     0.09    1.00_*  0.00   0.00_*
946   0.09  -0.08     0.09    1.00_*  0.00   0.00_*
957  -0.04   0.06     0.10_*  0.99_*  0.00   0.00
958   0.00   0.00     0.01    1.00_*  0.00   0.00_*
960  -0.10   0.11     0.12_*  1.00    0.01   0.00_*
961  -0.01   0.01     0.01    1.00_*  0.00   0.00_*
968   0.00   0.00     0.00    1.00_*  0.00   0.00_*
969  -0.12   0.13     0.14_*  1.00    0.01   0.00_*
1012 -0.01   0.02     0.02    1.00_*  0.00   0.00_*
1013 -0.08   0.09     0.10_*  1.00    0.00   0.00_*
1057 -0.10   0.12     0.14_*  0.99_*  0.01   0.00
1058 -0.07   0.08     0.09    1.00    0.00   0.00_*
1060 -0.15   0.16     0.17_*  1.00    0.01   0.00_*
1061 -0.07   0.08     0.09    1.00    0.00   0.00_*
1064 -0.36   0.40     0.42_*  0.97_*  0.09   0.00_*
1065 -0.10   0.12     0.13_*  1.00    0.01   0.00_*
1068 -0.10   0.12     0.14_*  0.99_*  0.01   0.00
1102  0.02  -0.02    -0.03    1.00_*  0.00   0.00_*
1107 -0.03   0.05     0.08    0.99_*  0.00   0.00
1158  0.01   0.00     0.06    1.00_*  0.00   0.00
1159 -0.06   0.07     0.08    1.00    0.00   0.00_*
1220  0.03  -0.03     0.03    1.00_*  0.00   0.00_*
1221  0.01  -0.01     0.01    1.00_*  0.00   0.00_*
1296  0.03  -0.04    -0.07    1.00_*  0.00   0.00
1319  0.09  -0.11    -0.13_*  0.99_*  0.01   0.00
1320  0.03  -0.04    -0.07    1.00_*  0.00   0.00
1321  0.04  -0.05    -0.05    1.01_*  0.00   0.00_*
1322  0.08  -0.08     0.08    1.00_*  0.00   0.00_*
1358  0.07  -0.08    -0.10_*  1.00_*  0.00   0.00
1403  0.14  -0.13     0.15_*  0.99_*  0.01   0.00
1405  0.03  -0.04    -0.06    1.00_*  0.00   0.00
1408  0.06  -0.05     0.08    0.99_*  0.00   0.00
1417 -0.01   0.00    -0.05    1.00_*  0.00   0.00
1461  0.04  -0.04    -0.05    1.00_*  0.00   0.00_*
1498  0.07  -0.06     0.10_*  0.99_*  0.00   0.00
1515  0.02  -0.02     0.02    1.00_*  0.00   0.00_*
1538 -0.07   0.09     0.10_*  1.00_*  0.01   0.00
1541  0.05  -0.04     0.06    1.00_*  0.00   0.00
1558  0.04  -0.04     0.04    1.00_*  0.00   0.00_*
1570  0.12  -0.11     0.13_*  0.99_*  0.01   0.00
1636 -0.07   0.08     0.08    1.00    0.00   0.00_*
1642 -0.11   0.13     0.16_*  0.99_*  0.01   0.00
1694 -0.13   0.15     0.15_*  1.00    0.01   0.00_*
1698 -0.07   0.08     0.09    1.00    0.00   0.00_*
1701 -0.11   0.12     0.13_*  1.00    0.01   0.00_*
1707 -0.06   0.07     0.08    1.00    0.00   0.00_*
1762 -0.07   0.09     0.11_*  1.00_*  0.01   0.00
1764 -0.15   0.17     0.18_*  1.00    0.02   0.00_*
1765 -0.10   0.11     0.12_*  1.00    0.01   0.00_*
1768 -0.57   0.63     0.66_*  0.92_*  0.21   0.00_*
1773 -0.10   0.13     0.15_*  0.99_*  0.01   0.00
1778 -0.01   0.01     0.01    1.00_*  0.00   0.00_*
1861 -0.07   0.08     0.10_*  1.00_*  0.00   0.00
1902  0.16  -0.15     0.16_*  1.01_*  0.01   0.01_*
1903  0.02  -0.02     0.02    1.00_*  0.00   0.00_*
1904  0.02  -0.02     0.02    1.00_*  0.00   0.00_*
1983  0.09  -0.08     0.09    1.00    0.00   0.00_*
```

```
2116  0.01   0.00     0.05    1.00_*  0.00   0.00
2155  0.03  -0.03    -0.03    1.00_*  0.00   0.00_*
2188  0.04  -0.04     0.04    1.00_*  0.00   0.00_*
2242  0.10  -0.10     0.11_*  1.00_*  0.01   0.00_*
2279  0.10  -0.10     0.10_*  1.00    0.01   0.00_*
2342 -0.16   0.18     0.20_*  0.99_*  0.02   0.00_*
2379  0.00   0.00     0.00    1.00_*  0.00   0.00_*
2385 -0.10   0.12     0.13_*  1.00    0.01   0.00_*
2386 -0.01   0.02     0.02    1.00_*  0.00   0.00_*
2396 -0.05   0.06     0.06    1.00    0.00   0.00_*
2399 -0.07   0.08     0.08    1.00    0.00   0.00_*
2400 -0.06   0.07     0.08    1.00    0.00   0.00_*
2446 -0.38   0.42     0.44_*  0.97_*  0.10   0.00_*
2447 -0.09   0.11     0.12_*  1.00    0.01   0.00_*
2451 -0.28   0.32     0.35_*  0.97_*  0.06   0.00_*
2523 -0.07   0.08     0.09    1.00    0.00   0.00_*
2596  0.06  -0.06     0.07    1.00    0.00   0.00_*
2599  0.03  -0.02     0.03    1.00_*  0.00   0.00_*
2633  0.03  -0.04    -0.07    1.00_*  0.00   0.00
2643  0.03  -0.05    -0.08    0.99_*  0.00   0.00
2651  0.06  -0.06     0.06    1.00    0.00   0.00_*
2654  0.02  -0.02     0.02    1.00_*  0.00   0.00_*
2656  0.10  -0.09     0.10_*  1.00    0.00   0.00_*
2657  0.02  -0.02     0.02    1.00_*  0.00   0.00_*
2667 -0.17   0.18     0.19_*  1.00    0.02   0.00_*
2670  0.02  -0.02     0.02    1.00_*  0.00   0.00_*
2696  0.05  -0.05     0.06    1.00_*  0.00   0.00_*
2738  0.03   0.00     0.12_*  0.97_*  0.01   0.00
2844  0.05  -0.05     0.05    1.01_*  0.00   0.00_*
2870  0.06  -0.05     0.08    0.99_*  0.00   0.00
2883  0.04  -0.04     0.04    1.00_*  0.00   0.00_*
2892 -0.03   0.04     0.07    1.00_*  0.00   0.00
2904  0.24  -0.24     0.24_*  1.00    0.03   0.01_*


Potentially influential observations of
        lm(formula = SalePrice ~ TotalFloorSF + OverallQual, data = subdat) :

     dfb.1_ dfb.TFSF dfb.OvrQ dffit   cov.r   cook.d hat
16   -0.17   0.40    -0.12     0.45_*  0.97_*  0.07   0.01_*
17    0.07   0.03    -0.09    -0.11    0.99_*  0.00   0.00
42    0.00   0.00    -0.01    -0.01    1.01_*  0.00   0.00
47   -0.20   0.13     0.09     0.28_*  0.98_*  0.03   0.00
60    0.01   0.00     0.00    -0.01    1.01_*  0.00   0.00
66   -0.05   0.12    -0.03     0.14_*  1.01_*  0.01   0.01_*
72    0.01   0.00     0.00     0.06    0.99_*  0.00   0.00
112   0.00  -0.01     0.00    -0.01    1.01_*  0.00   0.00
128  -0.01  -0.04     0.03    -0.04    1.01_*  0.00   0.01_*
131  -0.01   0.00     0.01    -0.01    1.01_*  0.00   0.00
161   0.03  -0.14     0.06    -0.15_*  1.00    0.01   0.01_*
175   0.01   0.01    -0.01     0.01    1.01_*  0.00   0.01_*
202   0.03   0.00    -0.03    -0.07    0.99_*  0.00   0.00
203  -0.01  -0.01     0.01    -0.01    1.01_*  0.00   0.00_*
235   0.03   0.03    -0.04     0.04    1.01_*  0.00   0.01_*
254   0.00  -0.01     0.01    -0.01    1.01_*  0.00   0.01_*
293   0.00  -0.12     0.08    -0.13_*  1.00    0.01   0.00
297  -0.03   0.03     0.02     0.08    0.99_*  0.00   0.00
```

```
303   0.16   0.00    -0.12     0.16_*  1.00    0.01   0.00_*
322  -0.10  -0.02     0.10     0.13_*  1.00    0.01   0.00
348  -0.12  -0.07     0.15     0.17_*  0.99_*  0.01   0.00
421  -0.05  -0.04     0.07     0.07    1.01_*  0.00   0.01_*
422  -0.15   0.11     0.06     0.21_*  0.99_*  0.01   0.00
424  -0.35  -0.01     0.30     0.41_*  0.96_*  0.05   0.00_*
430  -0.13  -0.03     0.13     0.17_*  0.99_*  0.01   0.00
432  -0.18  -0.02     0.17     0.21_*  0.99_*  0.02   0.01_*
433  -0.42   0.10     0.29     0.49_*  0.94_*  0.08   0.01_*
435  -0.10  -0.06     0.13     0.14_*  1.00    0.01   0.00
441  -0.04  -0.04     0.06     0.06    1.00_*  0.00   0.00_*
448  -0.06   0.00     0.05     0.07    1.01_*  0.00   0.00_*
449  -0.15   0.05     0.10     0.19_*  0.99_*  0.01   0.00
458  -0.08   0.04     0.05     0.12_*  0.99_*  0.00   0.00
496  -0.07   0.11    -0.01     0.14_*  1.00    0.01   0.00
497  -0.01   0.01     0.00     0.02    1.00_*  0.00   0.00
505  -0.08   0.08     0.02     0.14_*  0.99_*  0.01   0.00
514  -0.18   0.07     0.11     0.22_*  0.98_*  0.02   0.00
521   0.02   0.02    -0.03    -0.03    1.01_*  0.00   0.00_*
524   0.03   0.01    -0.03    -0.04    1.01_*  0.00   0.01_*
586   0.00   0.00     0.00     0.00    1.01_*  0.00   0.01_*
610  -0.01  -0.03     0.02    -0.03    1.00_*  0.00   0.00
663   0.11   0.00    -0.08     0.11    1.00    0.00   0.00_*
700  -0.05  -0.19     0.16    -0.21_*  0.99_*  0.01   0.01_*
709   0.08   0.00    -0.06     0.08    1.00_*  0.00   0.00_*
716  -0.05  -0.13     0.12    -0.15_*  1.00_*  0.01   0.00
717   0.01  -0.03     0.01    -0.03    1.00_*  0.00   0.00
720   0.07   0.03    -0.08    -0.10    0.99_*  0.00   0.00
743   0.08   0.02    -0.07     0.08    1.00_*  0.00   0.01_*
751   0.03  -0.01    -0.02    -0.06    1.00_*  0.00   0.00
766   0.15   0.06    -0.15     0.17_*  1.01_*  0.01   0.01_*
781   0.11   0.01    -0.09     0.11    1.00    0.00   0.00_*
782   0.02   0.02    -0.03     0.03    1.00_*  0.00   0.00
787   0.04   0.04    -0.06     0.06    1.01_*  0.00   0.01_*
829   0.00   0.00     0.00     0.00    1.00_*  0.00   0.00
867  -0.01   0.00     0.00     0.01    1.00_*  0.00   0.00
908   0.13  -0.01    -0.09     0.13_*  1.00    0.01   0.01_*
910   0.04  -0.15     0.06    -0.16_*  1.00_*  0.01   0.00
912  -0.02  -0.03     0.04    -0.04    1.01_*  0.00   0.00_*
946   0.07   0.02    -0.07     0.08    1.01_*  0.00   0.01_*
957  -0.05   0.07     0.01     0.12_*  0.98_*  0.01   0.00
960  -0.14  -0.03     0.14     0.17_*  0.99_*  0.01   0.00
966   0.00   0.00     0.00     0.00    1.00_*  0.00   0.00
969  -0.16  -0.05     0.17     0.21_*  0.98_*  0.01   0.00
1012  0.01   0.00    -0.01    -0.01    1.00_*  0.00   0.00
1013 -0.12  -0.05     0.14     0.16_*  0.99_*  0.01   0.00
1023  0.00   0.00     0.00     0.00    1.01_*  0.00   0.01_*
1028  0.00  -0.01     0.01    -0.01    1.01_*  0.00   0.01_*
1057 -0.10   0.10     0.02     0.17_*  0.99_*  0.01   0.00
1058 -0.11  -0.04     0.12     0.14_*  1.00_*  0.01   0.00
1060 -0.14   0.05     0.09     0.17_*  1.00    0.01   0.01_*
1064 -0.45   0.02     0.37     0.52_*  0.93_*  0.09   0.00_*
1065 -0.09   0.05     0.04     0.12_*  1.00    0.00   0.00
1068 -0.09   0.12     0.00     0.17_*  0.99_*  0.01   0.00
1100  0.00   0.00     0.00     0.00    1.00_*  0.00   0.00
1158  0.00   0.06    -0.04     0.06    1.01_*  0.00   0.01_*
1159 -0.03   0.02     0.01     0.04    1.01_*  0.00   0.00
```

```
1192  0.00  -0.03     0.02    -0.03    1.01_*  0.00   0.00_*
1200  0.02  -0.02     0.00    -0.03    1.00_*  0.00   0.00
1289  0.00  -0.03     0.02    -0.03    1.01_*  0.00   0.01_*
1307  0.02  -0.15     0.07    -0.15_*  1.01_*  0.01   0.01_*
1314 -0.01  -0.08     0.05    -0.10    0.99_*  0.00   0.00
1319  0.13  -0.05    -0.09    -0.19_*  0.97_*  0.01   0.00
1321  0.10  -0.03    -0.06    -0.11    1.00    0.00   0.01_*
1322  0.03   0.01    -0.03     0.04    1.01_*  0.00   0.01_*
1323  0.04  -0.12     0.04    -0.14_*  0.99_*  0.01   0.00
1346  0.03  -0.04    -0.01    -0.09    0.99_*  0.00   0.00
1358  0.07   0.04    -0.09    -0.11    0.99_*  0.00   0.00
1366 -0.01   0.01    -0.01    -0.05    0.99_*  0.00   0.00
1403  0.12   0.13    -0.17     0.19_*  0.99_*  0.01   0.00
1407  0.00   0.00     0.00     0.00    1.00_*  0.00   0.00
1408  0.06   0.05    -0.07     0.10    0.99_*  0.00   0.00
1420  0.00  -0.02     0.01    -0.02    1.01_*  0.00   0.01_*
1498 -0.01  -0.11     0.08    -0.11    1.02_*  0.00   0.02_*
1522 -0.02  -0.08     0.07    -0.09    1.01_*  0.00   0.01_*
1538 -0.01   0.03    -0.01     0.03    1.01_*  0.00   0.01_*
1570  0.09   0.10    -0.13     0.15_*  0.99_*  0.01   0.00
1588 -0.08   0.00     0.07     0.12_*  0.99_*  0.00   0.00
1636 -0.10  -0.04     0.12     0.14_*  1.00    0.01   0.00
1642 -0.13   0.07     0.07     0.19_*  0.97_*  0.01   0.00
1694 -0.20  -0.08     0.23     0.25_*  0.99_*  0.02   0.01_*
1701 -0.09   0.06     0.03     0.12_*  1.00    0.01   0.00
1707 -0.11  -0.06     0.13     0.15_*  1.00    0.01   0.00
1708 -0.01   0.01     0.00     0.01    1.01_*  0.00   0.01_*
1762 -0.06   0.08     0.01     0.12_*  1.00    0.00   0.00
1764 -0.17   0.03     0.12     0.20_*  1.00_*  0.01   0.01_*
1765 -0.05   0.06     0.01     0.08    1.01_*  0.00   0.01_*
1768 -0.53   0.88    -0.14     1.05_*  0.93_*  0.35   0.02_*
1773 -0.07   0.16    -0.05     0.18_*  1.00    0.01   0.01_*
1805  0.00   0.03    -0.02     0.04    1.00_*  0.00   0.00
1833  0.02  -0.03     0.00    -0.04    1.00_*  0.00   0.00
1865  0.00   0.00     0.00     0.00    1.00_*  0.00   0.00
1902  0.20   0.00    -0.15     0.20_*  1.00    0.01   0.01_*
1987 -0.01  -0.01     0.01    -0.01    1.01_*  0.00   0.00
1996 -0.01  -0.13     0.08    -0.15_*  0.99_*  0.01   0.00
1998  0.03  -0.11     0.04    -0.13_*  0.99_*  0.01   0.00
2002  0.05  -0.17     0.06    -0.19_*  0.99_*  0.01   0.00
2046 -0.01  -0.06     0.05    -0.06    1.01_*  0.00   0.01_*
2066 -0.03  -0.17     0.13    -0.18_*  1.00    0.01   0.01_*
2075  0.05   0.01    -0.03     0.06    1.00_*  0.00   0.00
2078  0.00   0.02    -0.01     0.02    1.00_*  0.00   0.00
2093  0.00   0.02    -0.01     0.02    1.01_*  0.00   0.00
2155 -0.01  -0.01     0.01     0.02    1.01_*  0.00   0.00
2214  0.00  -0.03     0.02    -0.03    1.01_*  0.00   0.00_*
2219  0.00  -0.01     0.00    -0.01    1.01_*  0.00   0.00
2231  0.00  -0.01     0.00    -0.01    1.01_*  0.00   0.00
2242  0.12   0.01    -0.09     0.12_*  1.00    0.00   0.00_*
2279  0.00   0.00     0.00     0.00    1.01_*  0.00   0.01_*
2342 -0.21  -0.05     0.23     0.28_*  0.97_*  0.03   0.00
2351  0.02  -0.08     0.03    -0.09    1.00    0.00   0.00_*
2385 -0.15  -0.07     0.18     0.21_*  0.99_*  0.01   0.00
2399 -0.12  -0.06     0.15     0.16_*  0.99_*  0.01   0.00
2400 -0.11  -0.06     0.14     0.15_*  1.00_*  0.01   0.00
2446 -0.35   0.38     0.04     0.54_*  0.97_*  0.10   0.01_*
```

```
2447 -0.06   0.05     0.02     0.09    1.00    0.00   0.00_*
2451 -0.26   0.41    -0.05     0.50_*  0.97_*  0.08   0.01_*
2454  0.01  -0.01     0.00    -0.02    1.01_*  0.00   0.00
2501 -0.01   0.01     0.00     0.01    1.01_*  0.00   0.00
2523 -0.12  -0.06     0.14     0.16_*  0.99_*  0.01   0.00
2528  0.00  -0.02     0.01    -0.02    1.01_*  0.00   0.01_*
2651  0.01   0.01    -0.02     0.02    1.01_*  0.00   0.00
2661  0.03  -0.06     0.00    -0.10    0.99_*  0.00   0.00
2663  0.03   0.00    -0.03    -0.06    1.00_*  0.00   0.00
2667 -0.14   0.08     0.07     0.18_*  1.00    0.01   0.01_*
 [ reached getOption("max.print") -- omitted 7 rows ]
```