Assignment #2

Regression Model Building

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 2 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, a transformation of the response variable from the sale price to the natural logarithm of the sale price will be conducted. We will then refit the 3 models using log(SalePrice) as the response variable instead of SalePrice and an analysis of goodness-of-fit and a comparison between the models using relevant plots will be conducted, displayed, and discussed to see which transformed model fit the best and if the models improved. Lastly, through this analysis, conditions or situations where a model is not appropriately specified 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

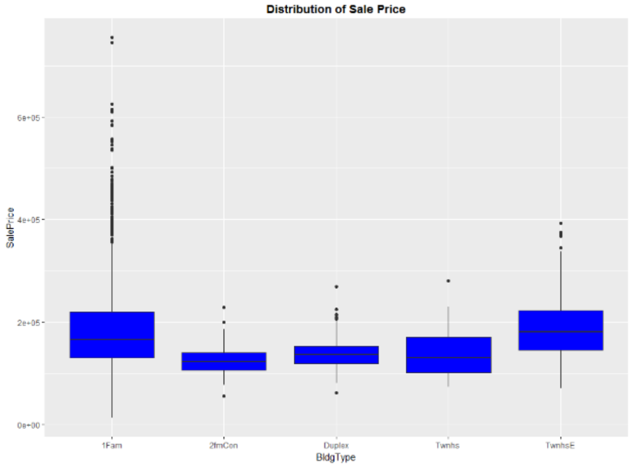


Figure 2: Boxplot of Sale Price & Sale Condition

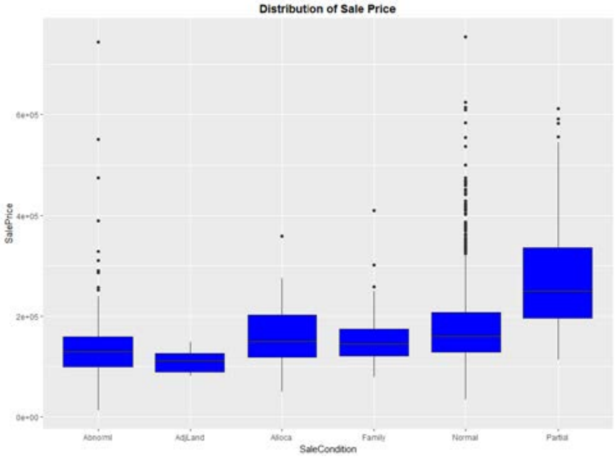
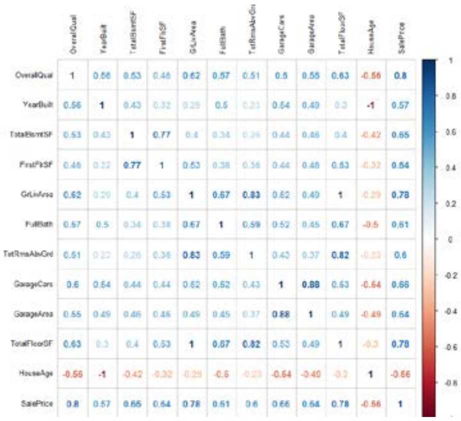


Figure 3: Waterfall of 'Drop Conditions'

```
         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: Exploratory Data Analysis



	OverallQual	YearBuilt	TotalBsmtSF	FirstFlrSF	GrLivArea	FullBath	TotRmsAbvGrd	GarageCars	GarageArea	TotalFloorSF	HouseAge	SalePrice
OverallQual	1	0.56	0.53	0.46	0.62	0.57	0.51	0.6	0.55	0.63	-0.56	0.8
YearBuilt	0.56	1	0.43	0.32	0.29	0.5	0.23	0.54	0.49	0.3	-1	0.57
TotalBsmtSF	0.53	0.43	1	0.77	0.4	0.34	0.26	0.44	0.46	0.4	-0.42	0.65
FirstFlrSF	0.46	0.22	0.77	1	0.53	0.38	0.36	0.44	0.46	0.53	-0.32	0.64
GrLivArea	0.62	0.29	0.4	0.53	1	0.67	0.83	0.52	0.49	1	-0.29	0.78
FullBath	0.57	0.5	0.34	0.38	0.67	1	0.59	0.52	0.45	0.67	-0.5	0.61
TotRmsAbvGrd	0.51	0.23	0.26	0.36	0.83	0.59	1	0.43	0.37	0.82	-0.23	0.6
GarageCars	0.6	0.54	0.44	0.44	0.52	0.52	0.43	1	0.88	0.53	-0.54	0.66
GarageArea	0.55	0.49	0.46	0.46	0.49	0.45	0.37	0.88	1	0.49	-0.49	0.64
TotalFloorSF	0.63	0.3	0.4	0.53	1	0.67	0.82	0.53	0.49	1	-0.3	0.78
HouseAge	-0.56	-1	-0.42	-0.32	-0.29	-0.5	-0.23	-0.54	-0.49	-0.3	1	-0.56
SalePrice	0.8	0.57	0.65	0.64	0.78	0.61	0.6	0.66	0.64	0.78	-0.56	1

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

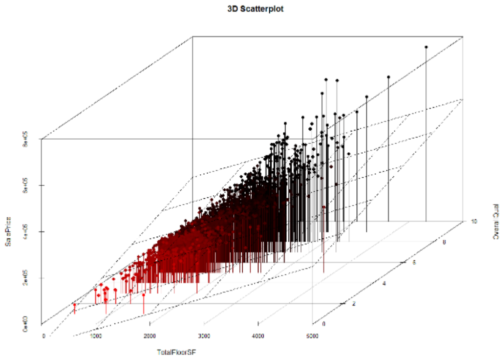


Figure 5: 3D Scatterplot of SalePrice, TotalFloorSF, and OverallQual

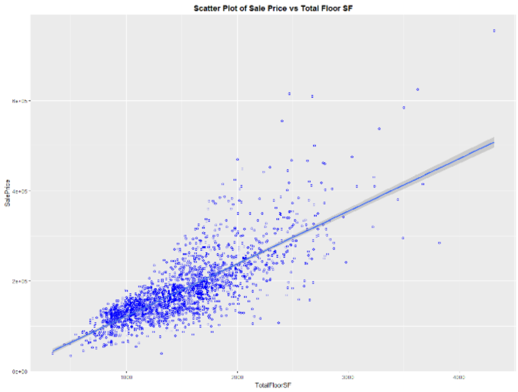


Figure 6: Scatterplot of SalePrice vs. TotalFloorSF

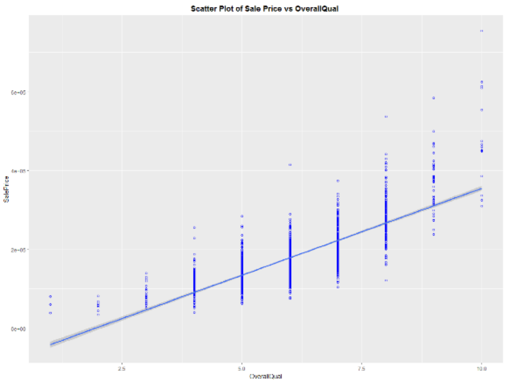


Figure 7: Scatterplot of SalePrice vs. OverallQual

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 3: Simple Linear Regression Models

Section 3.1: 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
```

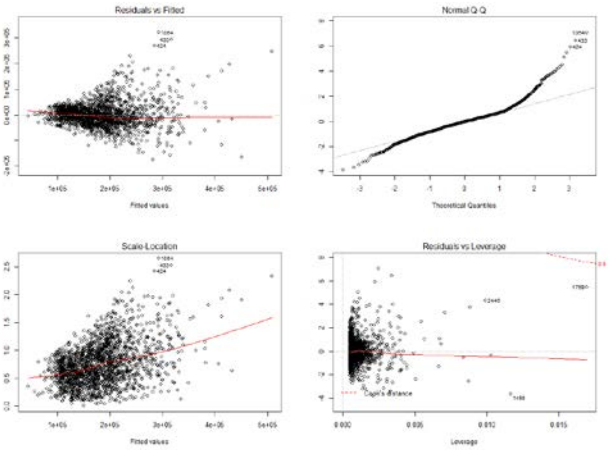
Observations: Figure 8 shows an ANOVA for SalePrice ~ TotalFloorSF. The F-statistic showed 3152 with a p-value = < 2.2e-16. Given that the p-value is very small, we can reject the null hypothesis that all the regression coefficients are equal to zero. Therefore, the model has produced statistically significant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

Section 3.2: 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
```

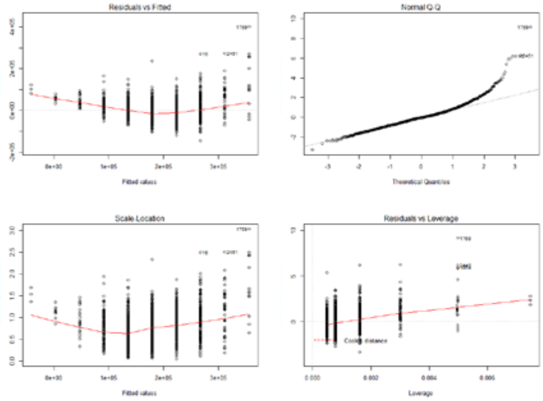
Observations: Figure 12 shows an ANOVA for SalePrice ~ OverallQual. The F-statistic showed 3638 with a p-value = < 2.2e-16. Given that the p-value is very small, we can reject the null hypothesis that all the regression coefficients are equal to zero. Therefore, the model has produced statistically significant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

Section 4: 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.

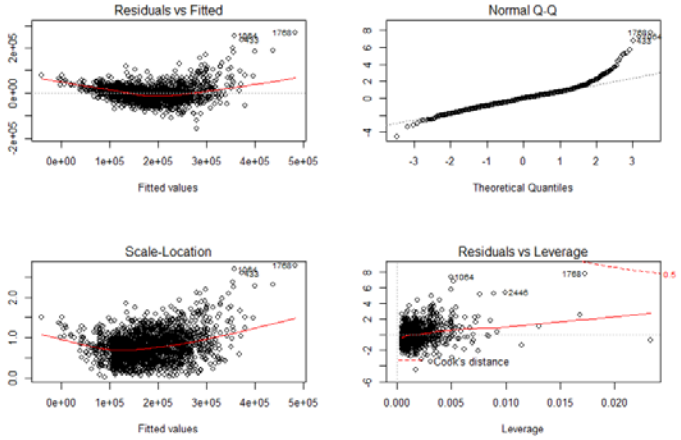
Figure 17: Multiple Linear Regression Model SalePrice ~ TotalFloorSF + OverallQual

```
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).

In conclusion, this model fits better than the simple linear regression models since the adjusted r-squared was higher and predicted error was lower than model #1 and #2. However, additional improvement is needed (e.g., transformation of the response variable SalePrice) to handle the problems cause by non-normality, non-linearity, and heteroscedasticity or non-constant variance that we have seen so far in this analysis.

Section 5: Log SalePrice Response Models

Section 5.1: Model #4 (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
```

Figure 20: Analysis of Variance for log(SalePrice) ~ TotalFloorSF

```
Analysis of Variance Table

Response: L_SalePrice
               Df Sum Sq Mean Sq F value    Pr(>F)
TotalFloorSF    1 172.78 172.778  3157.2 < 2.2e-16
Residuals    2000 109.45   0.055

TotalFloorSF ***
Residuals
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Observations: Figure 20 shows an ANOVA for log(SalePrice) ~ TotalFloorSF. The F-statistic showed 3157 with a p-value = < 2.2e-16. This is similar to what we saw in model #1. Given that the p-value is very small, 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. This was also similar to what we saw in model #1.

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
```

Figure 21: Simple Linear Regression Model log(SalePrice) ~ TotalFloorSF

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

Residuals:
     Min       1Q   Median       3Q      Max 
-1.32453 -0.12433  0.01809  0.13809  0.73234 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.114e+01  1.664e-02  669.31   <2e-16
TotalFloorSF 5.959e-04  1.060e-05   56.19   <2e-16

(Intercept)  ***
TotalFloorSF ***
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2339 on 2000 degrees of freedom
Multiple R-squared:  0.6122,   Adjusted R-squared:  0.612 
F-statistic:  3157 on 1 and 2000 DF,  p-value: < 2.2e-16
```

Observations: Figure 21 shows a summary of the Simple Linear Regression Model log(SalePrice) ~ TotalFloorSF. The equation of the regression line is: log(SalePrice) = 1.114e+01 + 5.959e-04* TotalFloorSF. Since the t-test of TotalFloor is statistically significant (p<0.001), we can use this equation. The residual standard error also shows 0.2339, instead of 45710 as seen in model #1. The multiple R-squared value of 0.6122 indicates that 61.22% of the variation in SalePrice is explained by the predictor variable TotalFloorSF. Overall, this multiple r-squared is similar to what we saw in model #1.

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

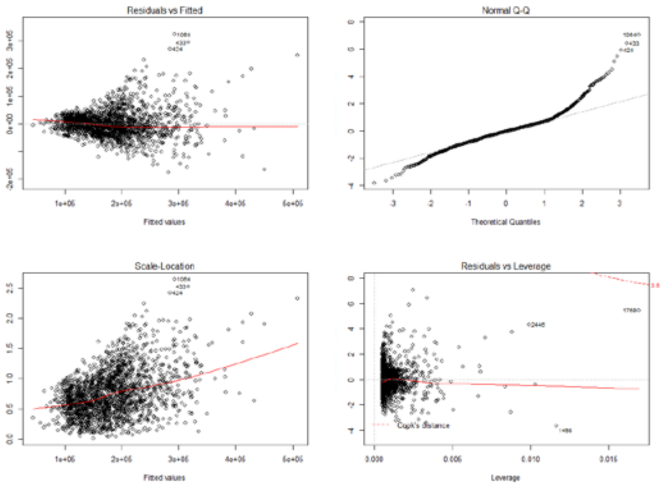


Figure 22: Scatterplots with Residuals & QQ-Plot of Residuals for log(SalePrice)

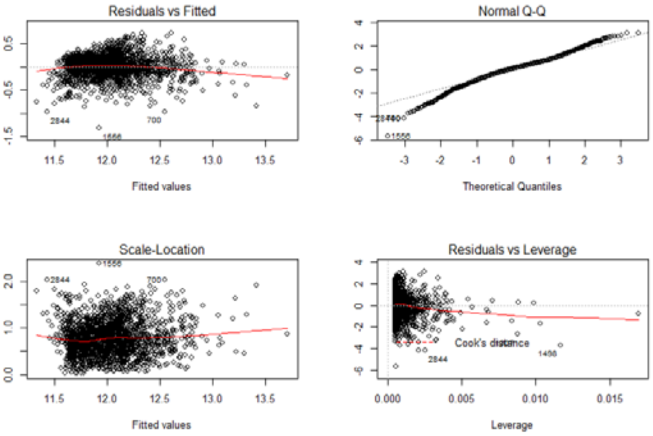


Figure 22, shows scatterplots with residuals and qq-plots of residuals so that we can check to make sure the model is meeting all the assumptions. After transformation, QQ plot reveals that the density distribution is close to normal since most of the dots are on the line, except some dots on the bottom left corner. By comparison, in model #1, we saw the assumption of normality being violated when the dots were departing from the line as seen in figure 10, QQ plot. The scatterplot of residuals vs. fitted shows us that the normal probability plot of the residuals appear to be relatively linear and have a random scatter of data over the range of values for the independent variable. By comparison, in model #1, we saw a “funnel” pattern and heteroscedasticity, which is seen in figure 10. As a result, given the assumptions of a simple regression analysis and the revelations/results from above, the regression model # 5 after transformation appears to fit the data better than model #1, even though r-squared did not improve much and there still seems to be outliers.

Section 5.2: Model #5 (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
```

Figure 23: Analysis of Variance for log(SalePrice) ~ OverallQual

```
Analysis of Variance Table

Response: L_SalePrice
              Df  Sum Sq Mean Sq F value    Pr(>F)    
OverallQual    1 193.687 193.687    4375 < 2.2e-16 ***
Residuals   2000  88.542   0.044                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Observations: Figure 23 shows an ANOVA for log(SalePrice) ~ OverallQual. The F-statistic showed 4375 with a p-value = < 2.2e-16. Given that the p-value is very small, 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. This was also similar to what we saw in model #2.

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
```

Figure 24: Simple Linear Regression Model log(SalePrice) ~ OverallQual

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

Residuals:
     Min       1Q   Median       3Q      Max 
-0.96291 -0.12656  0.00577  0.12551  0.91116 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 10.628883   0.021609  491.87   <2e-16 ***
OverallQual  0.232665   0.003518   66.14   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2104 on 2000 degrees of freedom
Multiple R-squared:  0.6863,	Adjusted R-squared:  0.6861 
F-statistic:  4375 on 1 and 2000 DF,  p-value: < 2.2e-16
```

Observations: Figure 24 shows a summary of the Simple Linear Regression Model log(SalePrice) ~ OverallQual. The equation of the regression line is: log(SalePrice) = 10.628883 + 0.232665* OverallQual. Since the t-test of TotalFloor is statistically significant (p<0.001), we can use this equation. The residual standard error also shows 0.2104, instead of 43690 as seen in model #2. Interestingly, the standard error of 0.2104 is smaller than model #4. The multiple R-squared value of 0.6863 indicates that 68.63% of the variation in SalePrice is explained by the predictor variable TotalFloorSF. This multiple r-squared is higher than the original model #2, which means that the model fits better after transformation was applied.

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

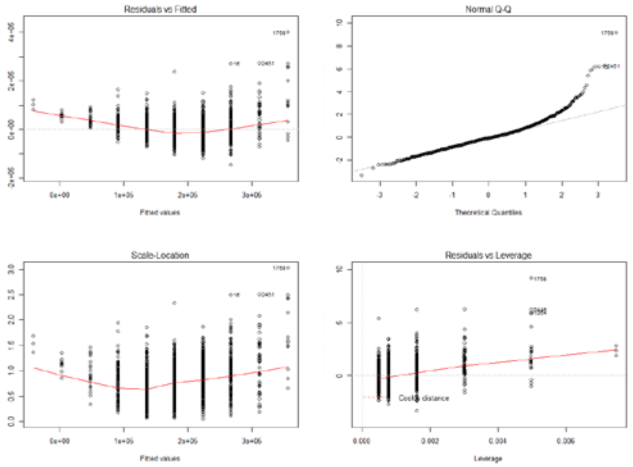
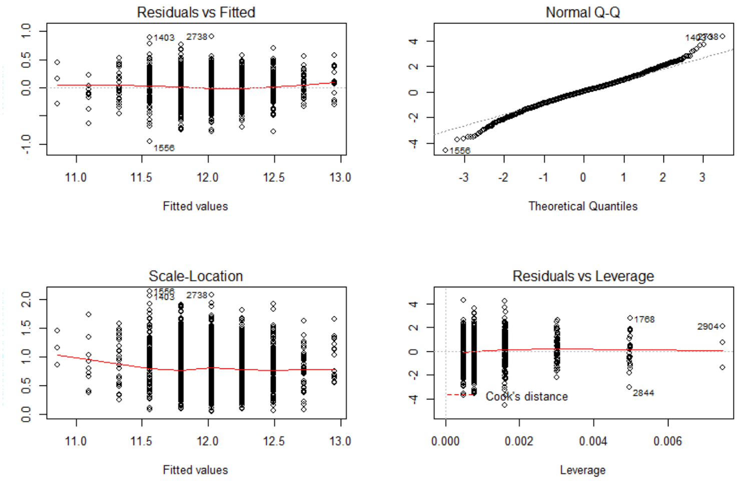


Figure 25: Scatterplots with Residuals & QQ-Plot of Residuals for log(SalePrice)



Observations: Figure 25, 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 still somewhat non-normal since some of the dots are still departing from line. However, this is an improvement to the drastic departing of the dots in the upper left hand corner in figure 14 (model #2, QQ plot). The scatterplot of residuals vs. fitted shows us that the normal probability plot of the residuals appear to be relatively linear and have a random scatter of data over the range of values for the independent variable. By comparison, in model #2, we saw a “funnel” pattern and heteroscedasticity, which is seen in figure 14. As a result, given the assumptions of a simple regression analysis and the revelations/results from above, the regression model # 5 after transformation appears to fit the data better than model #2, since r-squared increased and the assumptions improved.

Section 5.3: Model #6 (SalePrice ~ TotalFloorSF + OverallQual

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
```

Figure 26: Analysis of Variance for log(SalePrice) ~ TotalFloorSF + OverallQual

```
Analysis of Variance Table

Response: L_SalePrice
               Df  Sum Sq Mean Sq F value    Pr(>F)
TotalFloorSF    1 172.778 172.778  6047.0 < 2.2e-16 ***
OverallQual     1  52.334  52.334  1831.6 < 2.2e-16 ***
Residuals    1999  57.117   0.029
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Observations: Figure 26 shows an ANOVA for log(SalePrice) ~ ~ TotalFloorSF + OverallQual. The F-statistic on 2 showed 3939 with a p-value = < 2.2e-16. Given that the p-values are very small, 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. This was also similar to what we saw in model #3.

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
```

Figure 27: Multiple Linear Regression Model log(SalePrice) ~ TotalFloorSF + OverallQual

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

Residuals:
     Min       1Q   Median       3Q      Max 
-1.05907 -0.09260  0.01093  0.11050  0.68371 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.060e+01  1.738e-02  609.72   <2e-16 ***
TotalFloorSF 3.281e-04  9.893e-06   33.16   <2e-16 ***
OverallQual  1.561e-01  3.648e-03   42.80   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.169 on 1999 degrees of freedom
Multiple R-squared:  0.7976,   Adjusted R-squared:  0.7974 
F-statistic:  3939 on 2 and 1999 DF,  p-value: < 2.2e-16
```

Observations: Figure 27 shows a summary of the Multple Linear Regression Model log(SalePrice) ~ TotalFloorSF + OverallQual. The equation of the regression line is: log(SalePrice) = 1.060e+01 + 3.281e-04*TotalFloorSF + 1.561e-01*OverallQual. Since the t-test of TotalFloor & OverallQual are statistically significant ($p<0.001$), we can use this equation. The residual standard error also shows 0.169, instead of 35150 as seen in model #3. Interestingly, the standard error of 0.169 is smaller than both model #4 and model #5. The multiple R-squared value of 0.7976 indicates that 79.76% of the variation in SalePrice is explained by the predictor variable TotalFloorSF. This multiple r-squared is higher than the original model #3, which means that the model fits better after transformation was applied. Additionally, this model also seems to fit better than the all the other transformed models since the adjusted r-squared for this model is 0.7974 compared to model #4 (adjusted r-squared: 0.612) and model #5 (adjusted r-squared: 0.6861). This also shows evidence that the multiple linear regression model after transformation fits the best compared to the other transformed models.

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

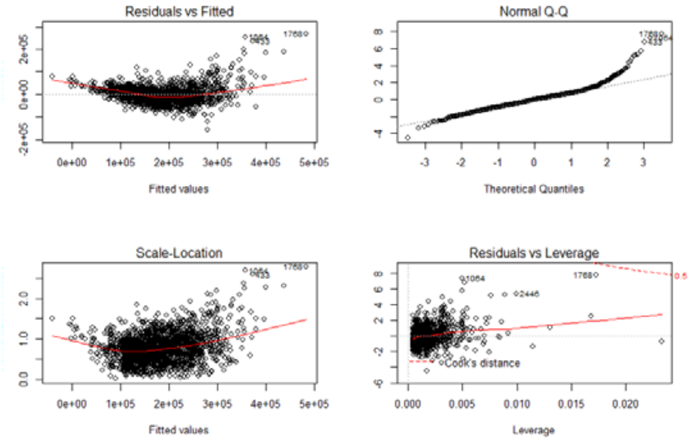
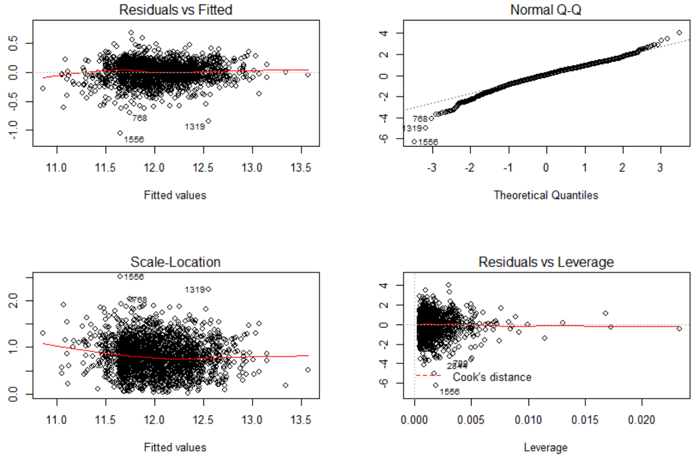


Figure 28: Scatterplots with Residuals & QQ-Plot of Residuals for log(SalePrice)



Observations: Figure 25, 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 model #3, we saw the assumption of normality being violated when the dots were drastically departing from the line as seen in figure 18, QQ plot. The scatterplot of residuals vs. fitted shows us that the normal probability plot of the residuals appear to be relatively linear and have a random scatter of data over the range of values for the independent variable. By comparison, in model #3, we saw a "bowl" shaped pattern with heteroscedasticity, which is seen in figure 18. As a result, given the assumptions of a simple regression analysis and the revelations/results from above, the multiple linear regression model # 6 after transformation appears to fit the data better than model #3, though there still seems to be outliers. In conclusion, given that the adjusted r-squared was higher and the standard error was smaller than the transformed simple linear regression models #4 and #5, and all of the by assumptions of normality, linearity, and homoscedasticity were satisfied, model #6 fits the best compared to the transformed models #4 and #5.

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. In section 3, simple linear regression models of SalePrice ~ TotalFloorSF and SalePrice ~ OverallQual were 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, and heteroscedasticity or non-constant variance were evident.

In section 4, 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. However, although this model fit better than model #1 and #2, the assumptions were still violated: non-normality, non-linearity, and heteroscedasticity or non-constant variance were evident.

As a result, in order to address this issue, a transformation of the response variable from the sale price to the natural logarithm of the sale price was conducted. We then refit the 3 models using log(SalePrice) as the response variable instead of SalePrice and an analysis of goodness-of-fit and a comparison between the models using relevant plots were conducted, displayed, and discussed to see which transformed model fit the best and if the models improved. Overall, all the transformed models were significant. Additionally, the transformed model r-squared improved and assumptions were all improved and/or met compared to their original models and therefore fit better. Lastly, given that the adjusted r-squared was higher and the standard error was smaller than the transformed simple linear regression models #4 and #5, and all of the by assumptions of normality, linearity, and homoscedasticity were satisfied, model #6 fit the best compared to model #4 and #5.

In conclusion, a model is not appropriately specified when the following exist: the model is not significant, r-squared is low, predicted error is large, and “wide” predicted error bands exists. Additionally, a model is not appropriately specified when assumptions are violated: non-normality, non-linearity, and heteroscedasticity or non-constant variance are evident in residual vs. fitted plots and QQ plot. Lastly, in terms of the next steps we need to continue with “model adequacy checking”. This includes testing diagnostics, addressing outliers through detection, addressing multicollinearity, and PCA transformation. After all these are conducted, we can then move to model validation to see if the model is going to satisfy business needs/requirements and ultimately 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.