Northwestern University
Master of Science in Data Science

By: Ali Gowani

Table of Contents

Overview: *4*

Purpose, Overview and Explanatory Variables: 4

Figure 1: Ames Dataset: Data Types 4

Figure 2: mydata Column (82) Types 5

Task 0: Define the Sample Population – Exploratory Data Analysis 5

Figure 3: mydata_filtered, missing data 6

Figure 4: mydata_filtered, removing missing data 7

PART A: Simple Linear Regression Models *7*

Task 1: 7

Figure 5: Scatter Plot of Sale Price vs Total Floor SF with Regression Line 7

Figure 6: Histogram of residual 9

Figure 7: Scatter Plot of Standardized Residuals vs Predicted Values 9

Task 2: 10

Figure 8: Sale Price vs Quality 10

Figure 9: Histogram of residual 11

Figure 10: Scatter Plot of Standardized Residuals vs Predicted Values 12

Task 3: 12

PART B: Multiple Linear Regression Models *12*

Task 4: 12

Figure 11: Histogram of residual 14

Figure 12: Scatter Plot of Standardized Residuals vs Predicted Values 14

Task 5: 15

Figure 13: Histogram of residual 17

Figure 14: Scatter Plot of Standardized Residuals vs Predicted Values 17

PART C: Multiple Linear Regression Models on Transformed Response Variable *18*

Task 6: 18

Figure 15: Histogram comparison of SalePrice and LogSalePrice 18

Figure 16: R-Squared and Adjusted R-Squared values for SalePrice and LogSalePrice 19

Task 7: 19

PART D: Multiple Linear Regression and Influential Points *20*

Task 8: 20

Figure 17: Visualize various charts for Log Transformation with Model 4 20

Figure 18: Influence Plot 20

Figure 19: Scatter Plot of Residual vs OverallQual 21

PART E: Beginning to Think About a Final Model *21*

Task 9: 21

Conclusion and Reflections:..... 22

Appendix 23

A: Data Quality Check (mydata) 23

B: Histogram distribution of features 33

C: Model 1 (ANNOVA and Summary Tables) 34

D: Model 2 (ANNOVA and Summary Tables)..... 35

E: Model 3 (ANNOVA and Summary Tables) 36

F: Model 4 (ANNOVA and Summary Tables) 37

G: Model 5 (ANNOVA and Summary Tables)..... 38

Overview:

Purpose, Overview and Explanatory Variables:

Let's revisit the Ames data, so we recall how the data looks like at a high level. The Ames dataset is provided by the Ames (Iowa) Assessor's Office, which was used to generate the assessed values for individual residential properties sold in Ames between 2006 and 2010. In total, there are 2,930 observations and 80 features (plus 2 Identifiers). From this, we can gather that there are enough observations for us to model our data in order to predict SalePrice as our response variable. However, we would need to reduce the number of features so that our model is not overfitting and filter our data to ensure we are predicting the SalesPrice for the appropriate "property."

Figure 1: Ames Dataset: Data Types

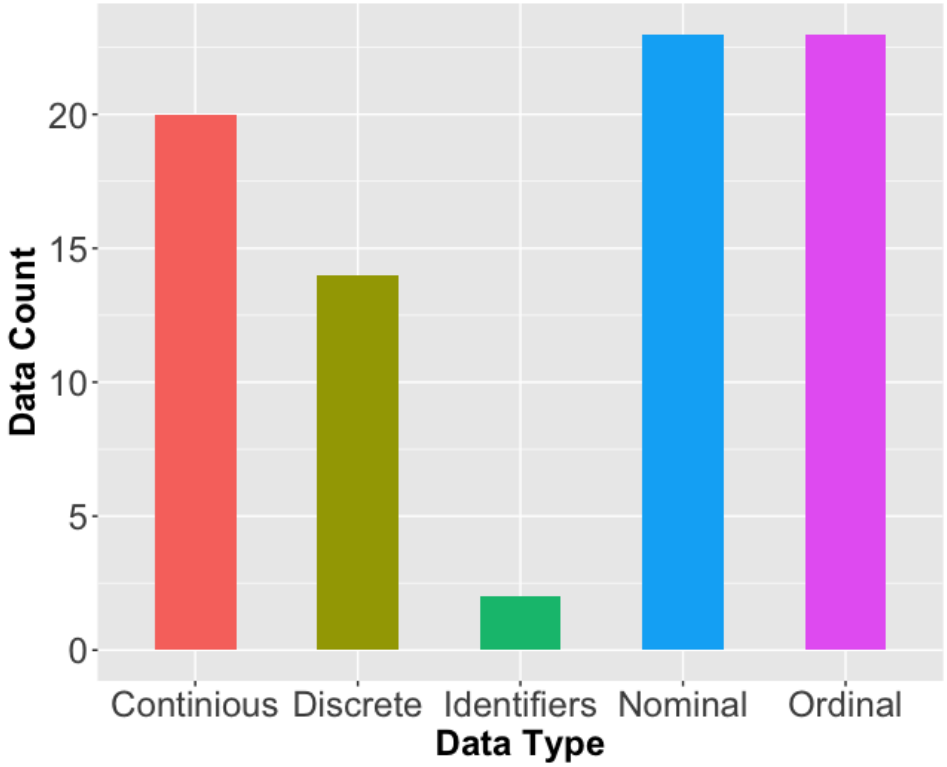
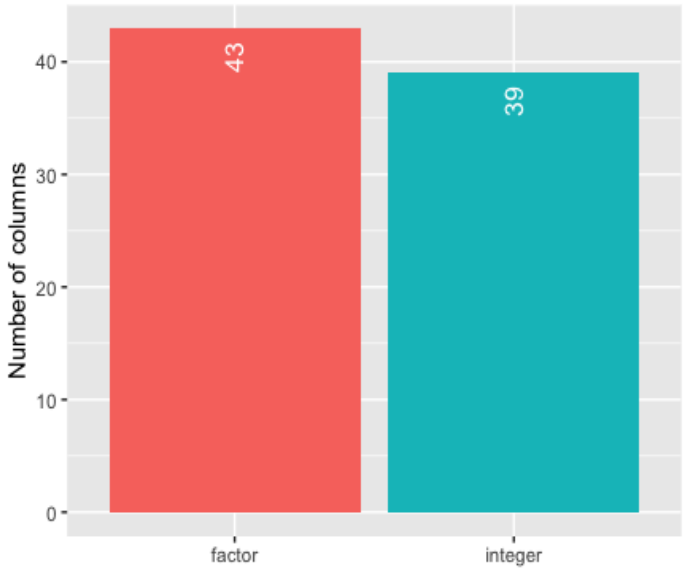


Figure 2: mydata Column (82) Types



The above charts show the various types of data and this is important as it will help us analyze the data. When graphing the data to better understand it, we should be mindful of the different types of data that will be well suited for visualization, for example, between discrete and continuous variables.

Task 0: Define the Sample Population – Exploratory Data Analysis

We can now define the Sample Population for our exercise. In this case, our response variable is SalePrice but SalePrice for what? A feature that can significantly impact SalePrice variable is Building Type (BldgType). For example, it is widely known to use square feet as predictor of SalePrice, assuming, other factors are held equal, but this can be misleading if we are comparing the square feet of a townhouse versus a single-family home. We also want to look at a range for SalePrice to ensure comparison of like types of homes and not too extreme. As before, we have limited our SalePrice of a range between $50,000 and $500,000.

Another component of the Sample Population that I want to define that was not defined in our last exercise for Ames data set, is Zoning. When looking at the various data profiling report, I noticed that even with reducing our sample population to Building Type and a range of Sale Price, we were getting homes that were in commercial zones and agriculture zones. This time I decided to limit the model's exposure to these observations.

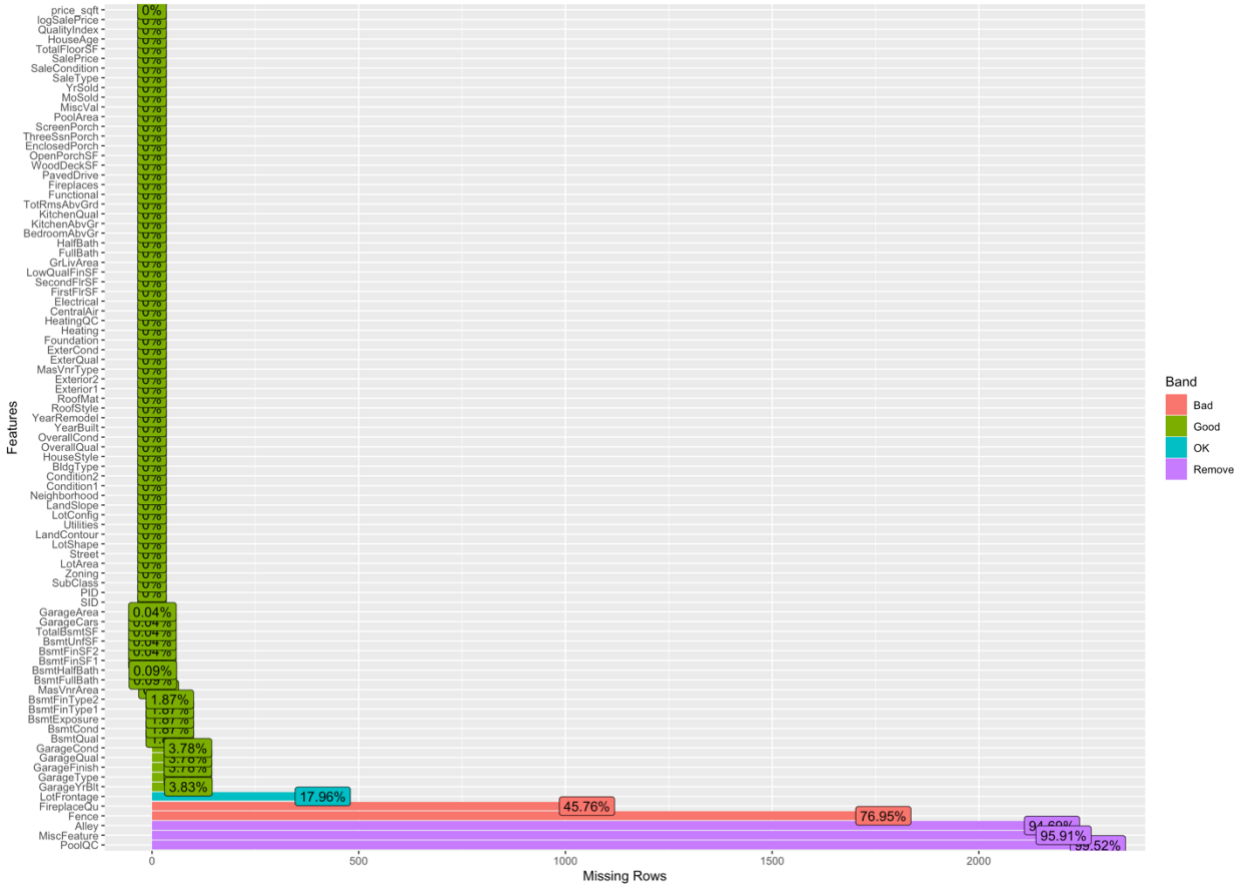
Sample Population (Observations: 2299, Variables: 87):

- Building Type is Single Family (1Fam)
- Sale Price range is between $50,000 and $500,000
- Zoning is Residential High Density (RH), Residential Low Density (RL), Residential Low Density Park (RP) or Residential Medium Density (RM)

We have now defined our sample population. The number of observations has been reduced from 2,930 to 2,299. We can now go deeper into Exploratory Data Analysis (EDA) to see if we can input and clean the data for our sample population.

There are several features, such as, alley, fence, etc. that have significant number of missing data. We should further analyze this to determine whether it makes sense to impute data or whether the feature is even meaningful. If there are more than 5% of missing data from any given feature then it is best to remove the feature as imputing it would insert bias into our dataset.

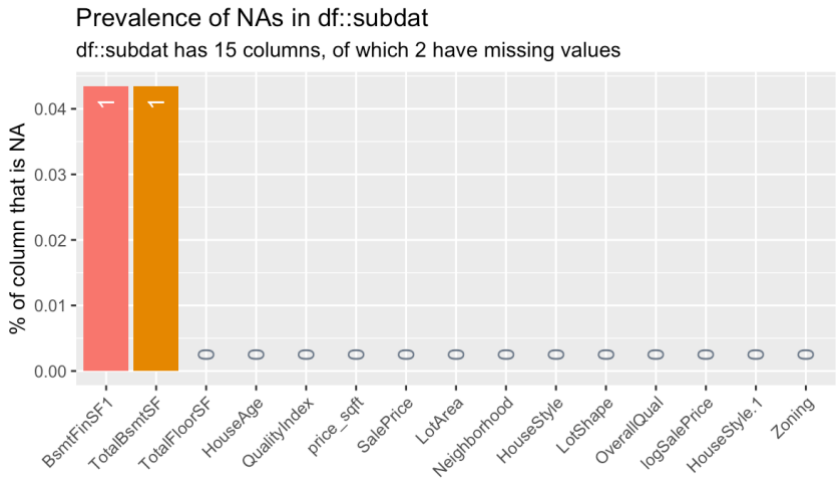
Figure 3: mydata_filtered, missing data



After reviewing the data in Appendix A and Appendix B, I selected the following features: "TotalFloorSF","HouseAge","QualityIndex","price_sqft"," SalePrice", "LotArea", "BsmtFinSF1", "Neighborhood", "HouseStyle", "LotShape", "OverallQual", "logSalePrice", "TotalBsmtSF", "HouseStyle", "Zoning"

I still noticed 1 or 2 observations in TotalBsmtSF and BsmtFinSF1 that had missing values so I removed them accordingly.

Figure 4: mydata_filtered, removing missing data

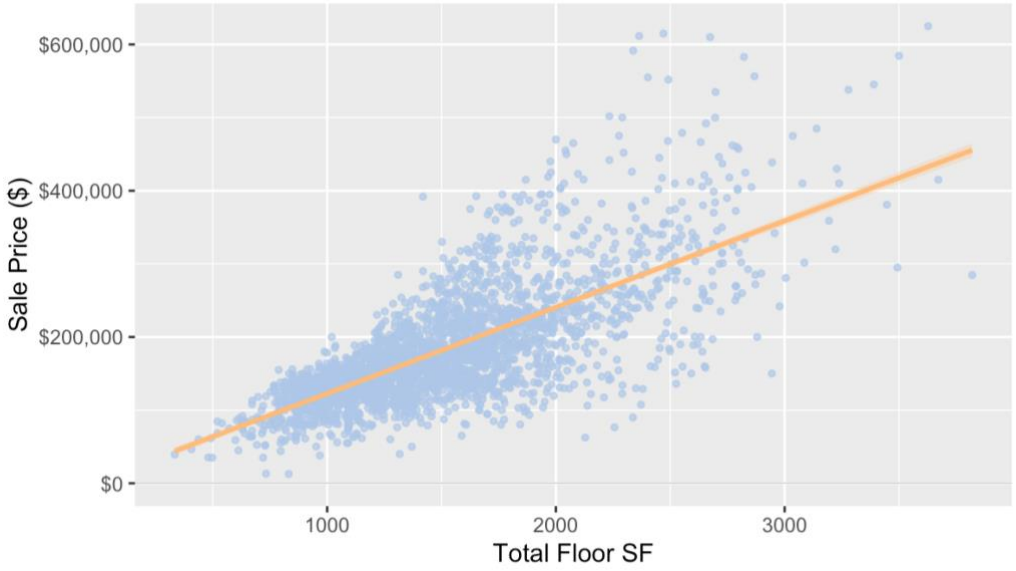


PART A: Simple Linear Regression Models

Task 1:

The best continuous variable that I selected for response variable SalePrice is TotalFloorSF (Total Floor Square Feet). I felt this would provide the most meaningful prediction as we are only able to select one variable.

Figure 5: Scatter Plot of Sale Price vs Total Floor SF with Regression Line



Reference Appendix C for Model 1 Summary and ANOVA Tables. For Model 1, the R-Squared is 0.5282, Y Intercept of 4623.80 and Slope 118.034.

$$\hat{Y} = 4623.8 + 118.034\beta_1$$

$$R^2 = 0.5282$$

Let's define our critical T-statistic to compare our variables:

$$t_{n-1,\left(1-\frac{a}{2}\right)} = t_{2923,0.975} = 1.961$$

T-statistic (t) for the variables is derived by taking the slope of $ß_i$ over the Standard Error (S_i) for $ß_i$.

$$T = \frac{\left(\beta_{\{1\}} - \beta_{\{1\}}^{\{(0)\}}\right)}{S_{(\beta_1)}}$$

1. **TotalFloorSF:**
    a. Null Hypothesis (H_o): $ß_1 = 0$
    b. Alternative Hypothesis (H_a): $ß_1 \neq 0$
    c. T-Statistic:
        i. $T = \frac{(118.034-0)}{2.063} = 57.219$
    d. With an alpha of 0.05 (Type I error) and T-statistic value greater than the critical T-Statistic value, we can reject the Null Hypothesis. This means that the TotalFloorSF variable provides significant information for predicting SalePrice.

We can reject the null hypothesis as the TotalFloorSF has a significant impact on the SalePrice. The R-squared value of 0.5282 can explain ~53% of the variance in SalePrice.

$$F = \frac{(Mean\ Sqrd\ Regression)}{(Mean\ Sqrd\ Residual)} = \frac{\left(\frac{SSY - SSE}{k}\right)}{\left(\frac{SSE}{n-k-1}\right)} = \frac{\left(\frac{18043611023761 - 9535682605311}{1}\right)}{\left(\frac{9535682605311}{2925-1-1}\right)} = 5531$$

The critical F-statistic for Model 1 is:

$$F_{i,n-k-p-1,1-a} = F_{1,2925-1-1,0.95} = 3.84$$

Since the F-statistic for Model 1 is 5531, which is greater than the critical F-statistic for Model 1 at 3.84 and p-value of less than 0.00001 then we can reject the Null Hypothesis. This means that our model contains significant relationship between the explanatory variable (TotalFloorSF) and the response variable of SalePrice.

Figure 6: Histogram of residual

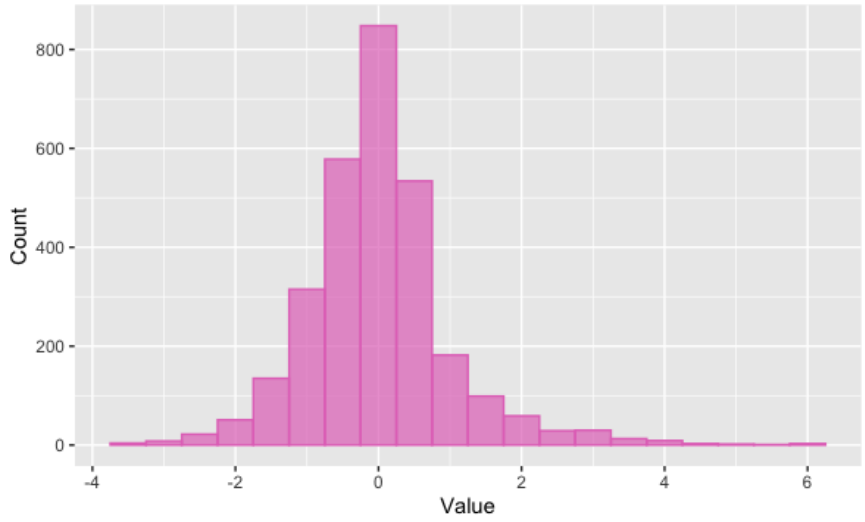
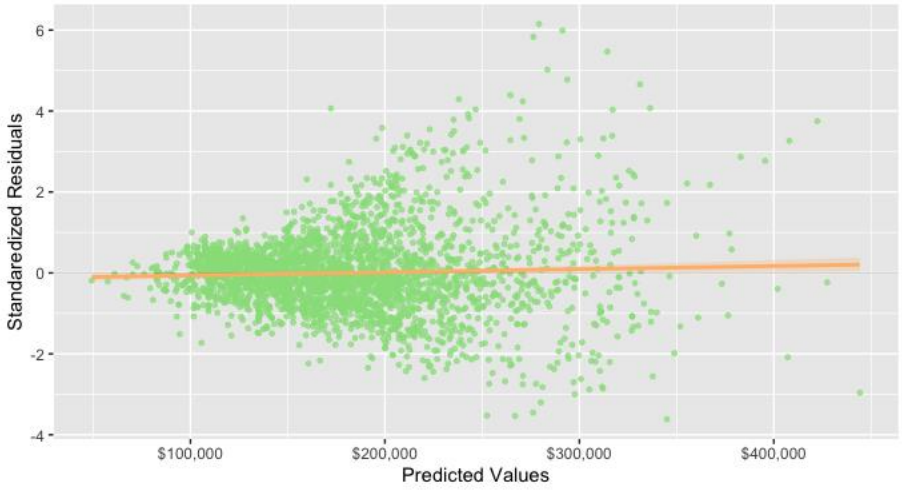


Figure 7: Scatter Plot of Standardized Residuals vs Predicted Values



The histogram above indicates that the distribution is a little skewed to the right. The scatter plot shows that the level of variability of the variable across the range of Predicted Values (Sale Price); this is known as heteroscedasticity. There are few points in the scatter plot that are to the far right and may indicate them as outliers. These could be influential points in our model as they have high residual value.

Task 2:

Figure 8: Sale Price vs Quality

Reference Appendix D for Model 2 Summary and ANOVA tables. For Model 2, the R-Squared is 0.6484, Y Intercept of -94088.40 and Slope 45087.20. The R-Squared value for Model 2 is 0.1202 (~12%) higher than in Model 1.

$$\hat{Y} = -94088.4 + 45087.2\beta_1$$

$$R^2 = 0.6484$$

Let's define our critical T-statistic to compare our variables:

$$t_{n-1,\left(1-\frac{a}{2}\right)} = t_{2923,0.975} = 1.961$$

T-statistic (t) for the variables is derived by taking the slope of ßᵢ over the Standard Error (S_i) for ßᵢ.

$$T = \frac{\left(\beta_{\{1\}} - \beta_{\{1\}}^{\{(0)\}}\right)}{S_{(\beta_1)}}$$

1. **OverallQual:**
    a. Null Hypothesis (H_o): $ß_1 = 0$
    b. Alternative Hypothesis (H_a): $ß_1 \neq 0$
    c. T-Statistic:
        i. $T = \frac{(45087.2-0)}{614.1} = 73.42$
    d. With an alpha of 0.05 (Type I error) and T-statistic value greater than the critical T-Statistic value, we can reject the Null Hypothesis. This means that the OverallQual variable provides significant information for predicting SalePrice.

We can reject the null hypothesis as the OverallQual has a significant impact on the SalePrice. The R-squared value of 0.6484 can explain ~65% of the variance in SalePrice.

$$F = \frac{(Mean\ Sqrd\ Reg.)}{(Mean\ Sqrd\ Residual)} = \frac{\left(\frac{SSY - SSE}{k}\right)}{\left(\frac{SSE}{n-k-1}\right)} = \frac{\left(\frac{18043611023761 - 11701736721987}{1}\right)}{\left(\frac{11701736721987}{2925-1-1}\right)} = 4507$$

The critical F-statistic for Model 1 is:

$$F_{i,n-k-p-1,1-a} = F_{1,2925-1-1,0.95} = 3.84$$

Since the F-statistic for Model 2 is 4507, which is greater than the critical F-statistic for Model 2 at 3.84 and p-value of less than 0.00001 then we can reject the Null Hypothesis. This means that our model contains significant relationship between the explanatory variable (OverallQual) and the response variable of SalePrice.

Figure 9: Histogram of residual

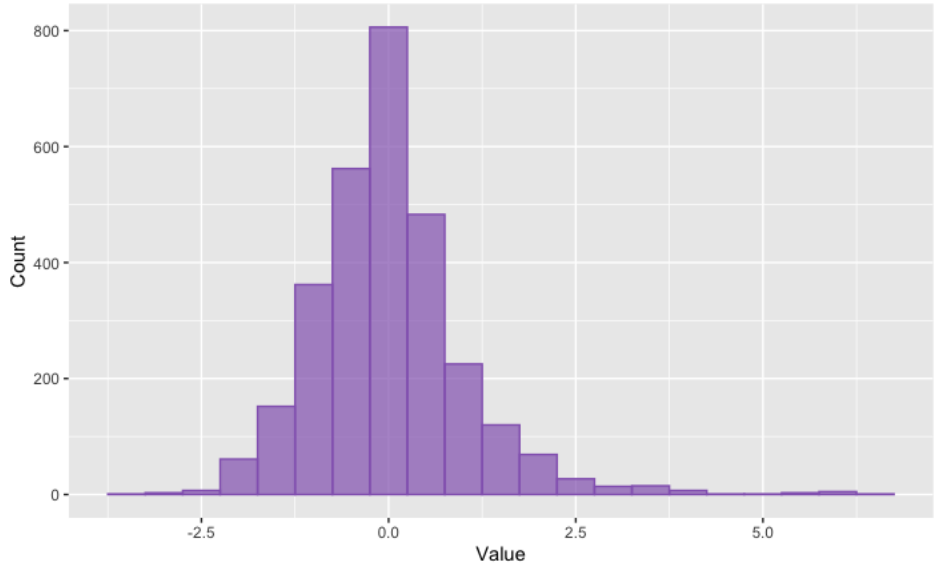
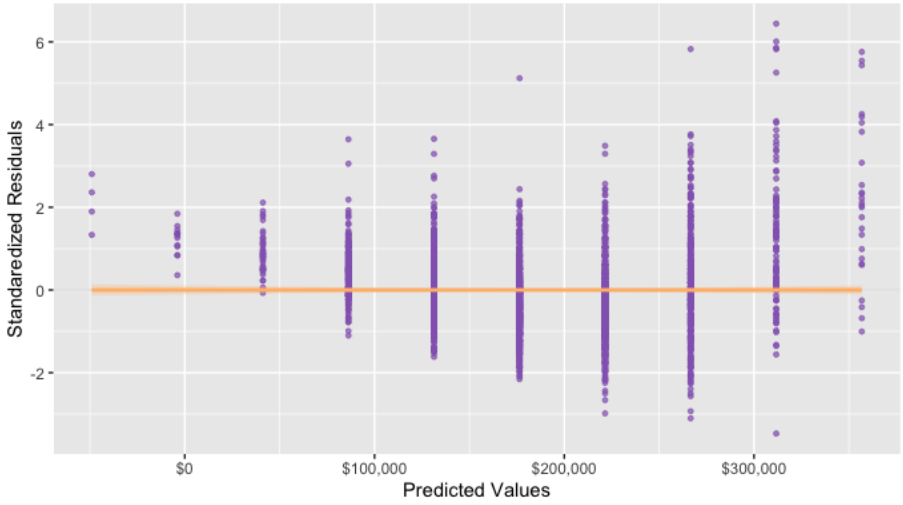


Figure 10: Scatter Plot of Standardized Residuals vs Predicted Values



The histogram above indicates that the distribution is a little skewed to the right. The scatter plot shows that the level of variability of the variable across the range of Predicted Values (Sale Price); this is known as heteroscedasticity. There are few points in the scatter plot that are to the far right and may indicate them as outliers. These could be influential points in our model as they have high residual value.

Task 3:

When comparing these two models, we should look at R-squared values and the F-statistic values to determine which model outperforms the other. Based on these two metrics, it seems that Model 2 outperforms Model 1. However, I do believe that each variable in the model can add a different perspective and a model with perhaps both of these variables and may be others can provide a more robust model.

PART B: Multiple Linear Regression Models

Task 4:

Reference Appendix E for Model 3 Summary and ANOVA tables. For Model 3, the R-Squared is 0.7687, Y Intercept of - 112566.859 and Slope for (OverallQual) β_1 is 32410.371 and for (TotalFloorSF) β_2 is 64.23. The R-Squared value for Model 3 is 0.1203 (~12%) higher than in Model 2.

$$\hat{Y} = -112566.859 + 32410.371\beta_1 + 64.23\beta_2$$

$$R^2 = 0.7687$$

Let's define our critical T-statistic to compare our variables:

$$t_{n-1,\left(1-\frac{a}{2}\right)} = t_{2923,0.975} = 1.961$$

T-statistic (t) for the variables is derived by taking the slope of $ß_i$ over the Standard Error (S_i) for $ß_i$.

$$T = \frac{\left(\beta_{\{1\}} - \beta_{\{1\}}^{\{(0)\}}\right)}{S_{(\beta_1)}}$$

1. **OverallQual:**
    a. Null Hypothesis (H_o): $ß_1 = 0$
    b. Alternative Hypothesis (H_a): $ß_1 \neq 0$
    c. T-Statistic:
        i. $T = \frac{(32410.371-0)}{627.075} = 51.69$
    d. With an alpha of 0.05 (Type I error) and T-statistic value greater than the critical T-Statistic value, we can reject the Null Hypothesis. This means that the OverallQual variable provides significant information for predicting SalePrice.

2. **TotalFloorSF:**
    a. Null Hypothesis (H_o): $ß_2 = 0$
    b. Alternative Hypothesis (H_a): $ß_2 \neq 0$
    c. T-Statistic:
        i. $T = \frac{(64.23-0)}{1.819} = 35.32$
    d. With an alpha of 0.05 (Type I error) and T-statistic value greater than the critical T-Statistic value, we can reject the Null Hypothesis. This means that the TotalFloorSF variable provides significant information for predicting SalePrice.

We can reject the null hypothesis as the OverallQual and TotalFloorSF have a significant impact on the SalePrice. The R-squared value of 0.7536 can explain ~75% of the variance in SalePrice.

The Omnibus Overall F-statistic for Model 3:

a. Null Hypothesis (H_o): $ß_1 = ß_2 = 0$
b. Alternative Hypothesis (H_a): $ß_i \neq 0$ for at least one value of i (e.g.: 1 or 2)

$$F = \frac{(Mean\ Sqrd\ Reg.)}{(Mean\ Sqrd\ Residual)} = \frac{\left(\frac{SSY - SSE}{k}\right)}{\left(\frac{SSE}{n-k-1}\right)} = \frac{\left(\frac{18043611023761 - 13599007521293}{2}\right)}{\left(\frac{13599007521293}{2925-2-1}\right)} = 1937$$

The critical F-statistic for Model 1 is:

$$F_{i,n-k-p-1,1-a} = F_{2,2925-2-1,0.95} = 3.00$$

Since the F-statistic for Model 3 is 1937, which is greater than the critical F-statistic for Model 2 at 3.00 and p-value of less than 0.00001 then we can reject the Null Hypothesis. This means that our model contains significant relationship between the explanatory variables (OverallQual and TotalFloorSF) and the response variable of SalePrice.

Figure 11: Histogram of residual

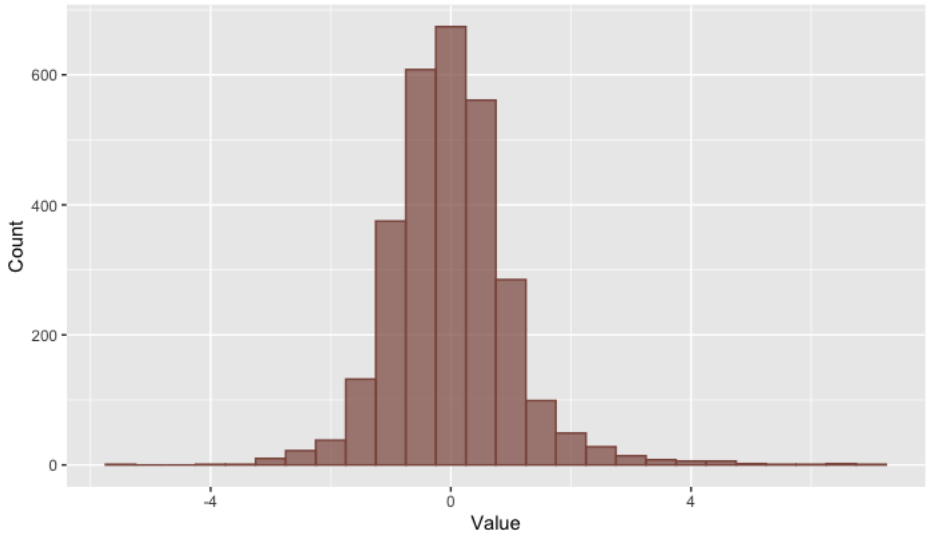
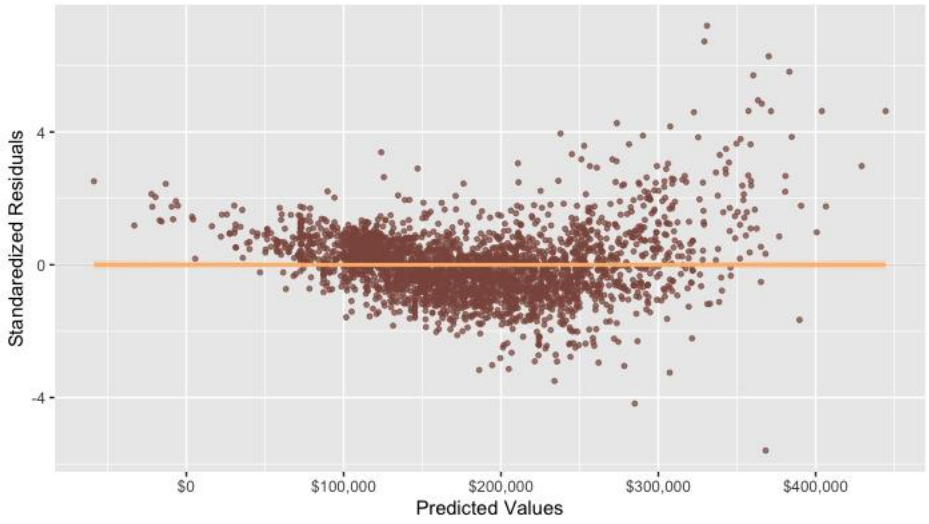


Figure 12: Scatter Plot of Standardized Residuals vs Predicted Values



The histogram above indicates that the distribution is not as skewed to the right as with the previous models and it continues to be normally distributed. The scatter plot shows that the level of variability of the variable across the range of Predicted Values (Sale Price); this is known as heteroscedasticity. It still shows high variability towards the higher predicted sales price values compared to when the predicted sales price values are low.

My suggestion is to keep both coefficients (OverallQual and TotalFloorSF) in this model as they each show a statistical significance for the t-test. In addition, I like the fact that one variable is more focused on quantitative portion of the sales price (e.g.: square feet of the home) and the other on the qualitative ide (e.g.: overall quality of the home). In addition, when the two variables are used in a model, they account for a much larger portion of the variance in our model, when compared to them individually. The R-squared value of them together is much larger than when each are taken separately.

Task 5:

Reference Appendix F for Model 4 Summary and ANOVA tables. For Model 4, the R-Squared is 0.8017, Y Intercept of – 113328.302 and Slope for (OverallQual) β_1 is 25493.956, for (TotalFloorSF) β_2 is 58.279 and for (TotalBsmtSF) β_3 is 49.422. The R-Squared value for Model 4 is 0.033 (~3%) higher than in Model 3.

$$\hat{Y} = -113328.302 + 25493.956\beta_1 + 58.279\beta_2 + 49.422\beta_3$$

$$R^2 = 0.8017$$

Let's define our critical T-statistic to compare our variables:

$$t_{n-1,\left(1-\frac{a}{2}\right)} = t_{2923,0.975} = 1.961$$

T-statistic (t) for the variables is derived by taking the slope of ßᵢ over the Standard Error (S_i) for ßᵢ.

$$T = \frac{\left(\beta_{\{1\}} - \beta_{\{1\}}^{\{(0)\}}\right)}{S_{(\beta_1)}}$$

1. **OverallQual:**
    a. Null Hypothesis (H_o): $ß_1 = 0$
    b. Alternative Hypothesis (H_a): $ß_1 \neq 0$
    c. T-Statistic:
        i. $T = \frac{(25493.956-0)}{619.167} = 41.17$
    d. With an alpha of 0.05 (Type I error) and T-statistic value greater than the critical T-Statistic value, we can reject the Null Hypothesis. This means that the OverallQual variable provides significant information for predicting SalePrice.

2. **TotalFloorSF:**
    a. Null Hypothesis (H_o): $ß_2 = 0$
    b. Alternative Hypothesis (H_a): $ß_2 \neq 0$
    c. T-Statistic:
        i. $T = \frac{(58.279-0)}{1.646} = 35.41$

d. With an alpha of 0.05 (Type I error) and T-statistic value greater than the critical T-Statistic value, we can reject the Null Hypothesis. This means that the TotalFloorSF variable provides significant information for predicting SalePrice.

3. **TotalBsmtSF:**
    a. Null Hypothesis (H_o): $ß_3 = 0$
    b. Alternative Hypothesis (H_a): $ß_3 \neq 0$
    c. T-Statistic:
        i. $T = \frac{(49.422-0)}{1.850} = 26.71$
    d. With an alpha of 0.05 (Type I error) and T-statistic value greater than the critical T-Statistic value, we can reject the Null Hypothesis. This means that the TotalBsmtSF variable provides significant information for predicting SalePrice.

We can reject the null hypothesis as the OverallQual, TotalFloorSF and TotalBsmtSF have a significant impact on the SalePrice. The R-squared value of 0.8017 can explain ~80% of the variance in SalePrice.

The Omnibus Overall F-statistic for Model 4:

c. Null Hypothesis (H_o): $ß_1 = ß_2 = ß_3 = 0$
d. Alternative Hypothesis (H_a): $ß_i \neq 0$ for at least one value of i (e.g.: 1, 2 or 3)

$$F = \frac{(Mean\ Sqrd\ Reg.)}{(Mean\ Sqrd\ Residual)} = \frac{\left(\frac{SSY - SSE}{k}\right)}{\left(\frac{SSE}{n-k-1}\right)} = \frac{\left(\frac{4820137546146 - 14460412638438}{3}\right)}{\left(\frac{14460412638438}{2925-3-1}\right)} = 972$$

The critical F-statistic for Model 1 is:

$$F_{i,n-k-p-1,1-a} = F_{3,2925-3-1,0.95} = 2.62$$

Since the F-statistic for Model 3 is 972, which is greater than the critical F-statistic for Model 2 at 2.62 and p-value of less than 0.00001 then we can reject the Null Hypothesis. This means that our model contains significant relationship between the explanatory variables (OverallQual, TotalFloorSF and TotalBsmtSF) and the response variable of SalePrice.

Figure 13: Histogram of residual

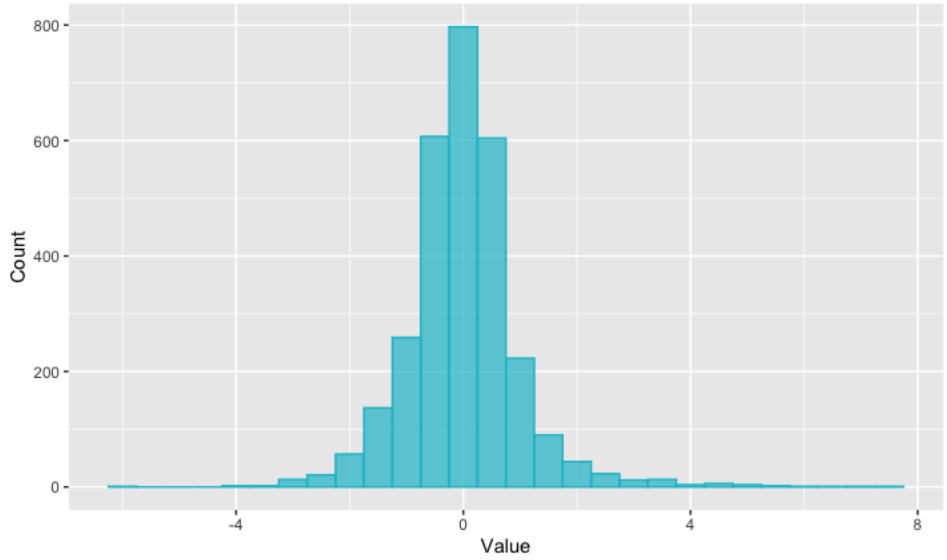
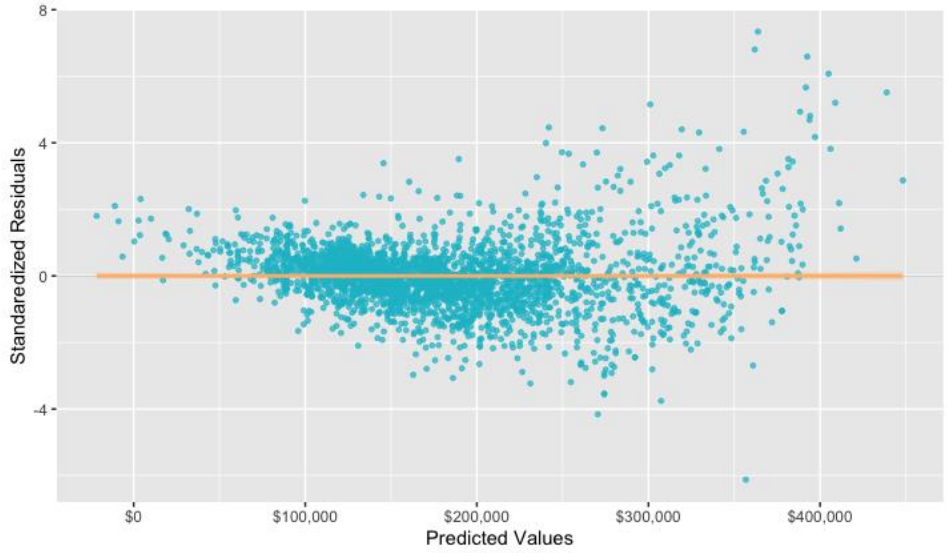


Figure 14: Scatter Plot of Standardized Residuals vs Predicted Values



The histogram above indicates that the distribution is not as skewed to the right as with the previous models and it continues to be normally distributed. The scatter plot shows that the level of variability of the variable across the range of Predicted Values (Sale Price); this is known as heteroscedasticity. It still shows high variability towards the higher predicted sales price values compared to when the predicted sales price values are low.

My suggestion is to keep all coefficients (OverallQual, TotalFloorSF and TotalBsmtSF) in this model as they each show a statistical significance for the t-test. In addition, I like the fact that one variable is more focused on quantitative portion of the sales price (e.g.: square feet of the home) and the other on the qualitative ide (e.g.: overall quality of the home). In addition, when the three variables are used in the

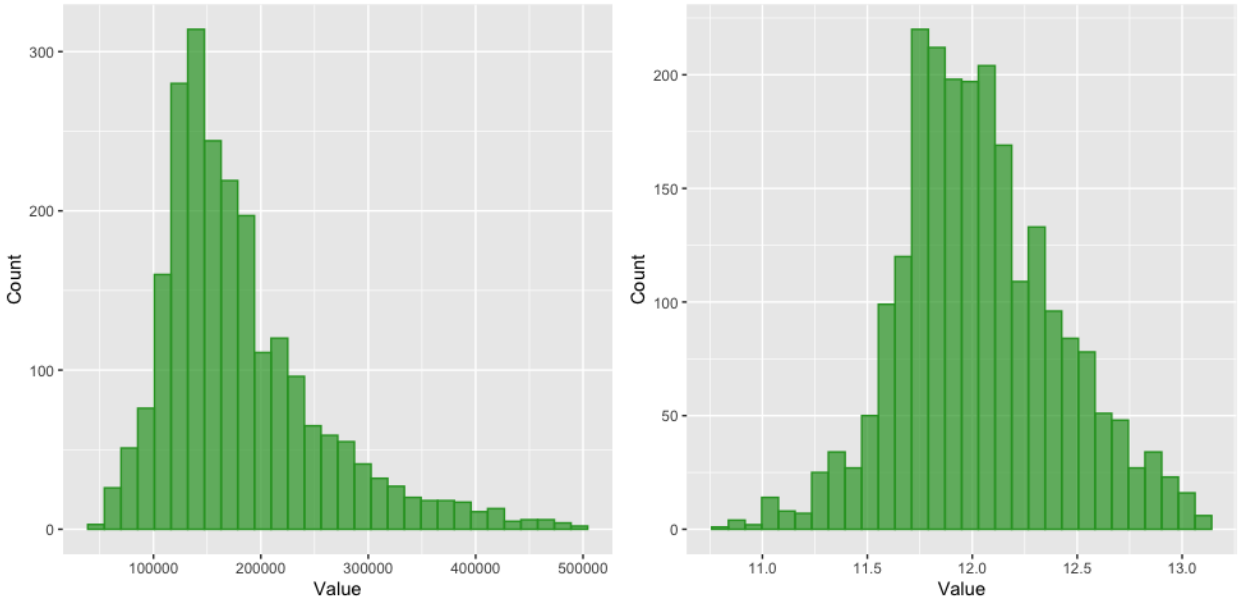
model, they account for a much larger portion of the variance in our model, when compared to them individually. The R-squared value of them together is larger by approximately 3% than when compared to the previous model.

PART C: Multiple Linear Regression Models on Transformed Response Variable

Task 6:

After comparing R-squared and Adjusted R-Squared to the 3 models, it seems that the log transformation does add to the additional explained variance. The transformation allows us to have a more normalized distribution for SalePrice. This is evident in the histograms below. The histogram on the left is of SalePrice and we can see the skewness to the right. However, after the log transformation. Based on this, it is prefered to log transform the SalePrice.

Figure 15: Histogram comparison of SalePrice and LogSalePrice



After the log transformation of Sale Price, I looked at the models and compared the R-Squared and the Adjusted R-Squared values for each of the 3 models against both, the SalePrice and the LogSalePrice variables.

Figure 16: R-Squared and Adjusted R-Squared values for SalePrice and LogSalePrice

Model_Name	Statistic	value
Model 1	Sale Price R^2	0.5283184
Model 3	Sale Price R^2	0.7534214
Model 4	Sale Price R^2	0.8130892
Model 1	Sale Price Adj. R^2	0.7535901
Model 3	Sale Price Adj. R^2	0.8016682
Model 4	Sale Price Adj. R^2	0.5173099
Model 1	Log Sale Price R^2	0.8018718
Model 3	Log Sale Price R^2	0.5174750
Model 4	Log Sale Price R^2	0.7744583
Model 1	Log Sale Price Adj. R^2	0.5281570
Model 3	Log Sale Price Adj. R^2	0.7746125
Model 4	Log Sale Price Adj. R^2	0.8128972

Based on the analysis, it seems that the Adjusted R-squared value is a good indicator of the variance accounted for in our models for SalePrice and LogSalePrice. The Adjusted R-Squared used a penalizing factor for each additional variable added to the model so that variables that do not add importance to the model do not increase its value. Based on this, Model 4 seems to fit the data based.

Task 7:

After analyzing the histograms in Figure 15, it seems that the LN(SalePrice) or the LogSalePrice transformation helps to normalize the SalePrice variable. However, once this has occurred, it is difficult to read and interpret the interaction between the variables and LogSalePrice by just looking at the data. While it may seem that this model has become a “blackbox,” the performance of the model has increased. This is the tradeoff between transforming variables and being able to explain the interaction of the variables to a business executive or a client manager.

We can always transform the LogSalePrice back to SalePrice after the prediction has occurred but this I another step that we would need to take into account, especially, if we want to explain the performance of the model to someone that is not a modeler.

PART D: Multiple Linear Regression and Influential Points

Task 8:

Reference Appendix G for Model 5 Summary and ANOVA tables. After comparing R-squared and Adjusted R-Squared to the 3 models, it seems that the log transformation does add to the additional explained variance. The transformation allows us to have a more normalized

Figure 17: Visualize various charts for Log Transformation with Model 4

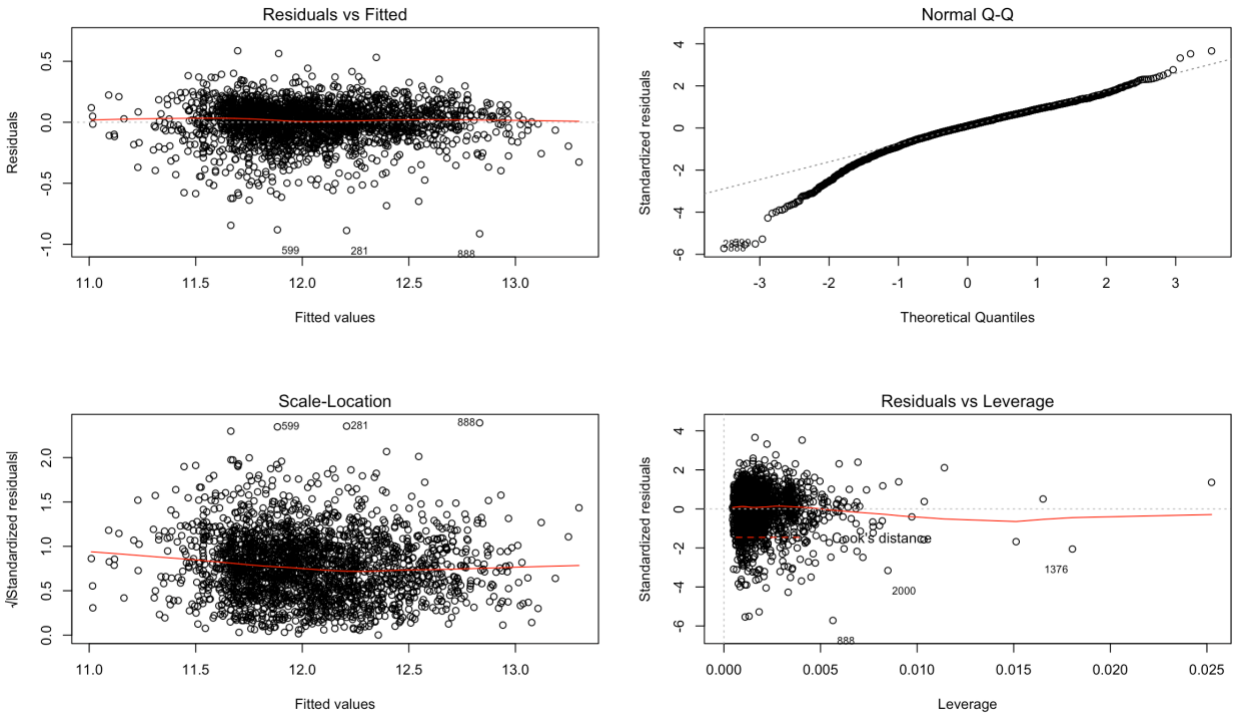
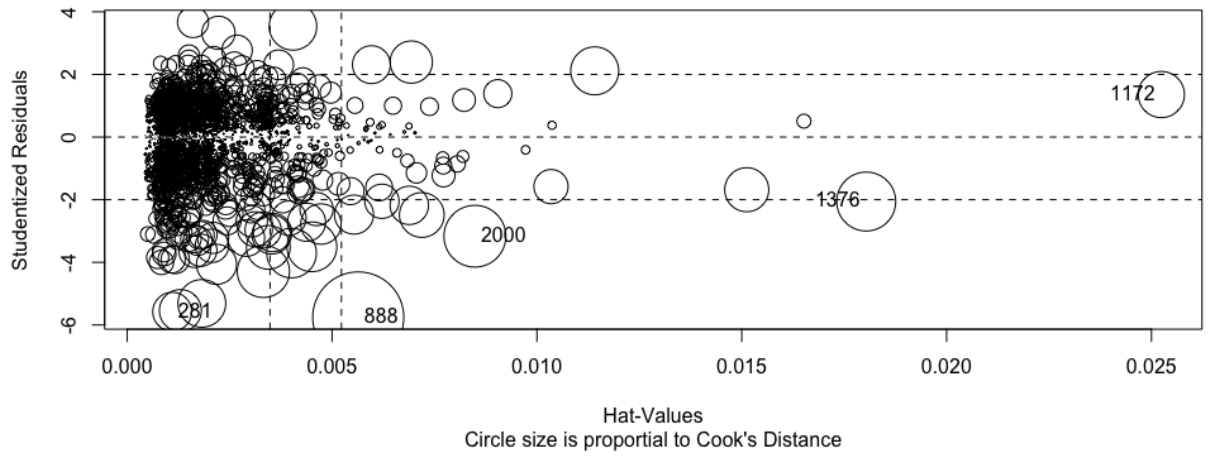
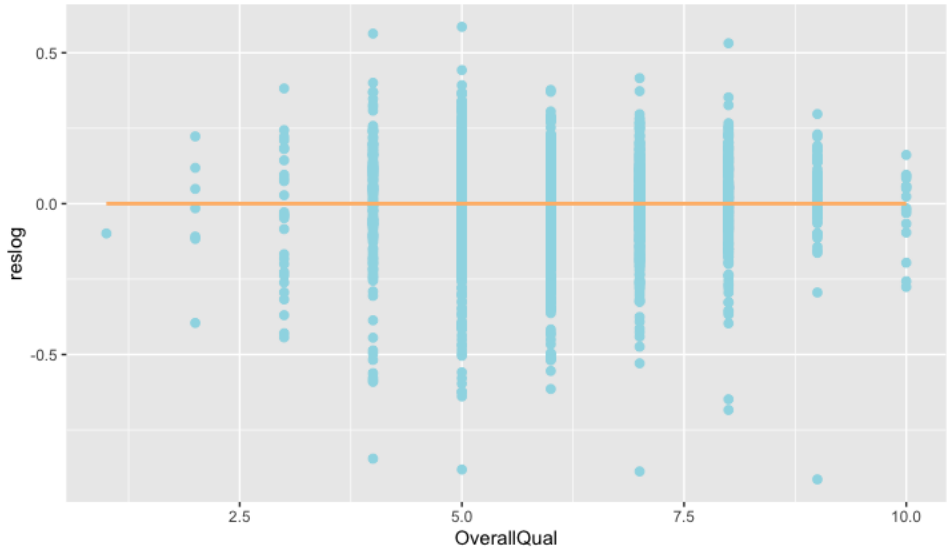


Figure 18: Influence Plot



After reviewing plots above as well as the influential points plot, it seems there are quite a bit of influential points that are impacting the model coefficients. I originally had 2920 observations in Model 4 but after removing the influential points, the observations were reduced to 2291; a difference of 629 observations.

Figure 19: Scatter Plot of Residual vs OverallQual



By removing the influencing points, the model is more accurate in predicting the SalePrice, given the higher R-squared value. However, it is now limited as we removed over 21.5% of the observed data points.

PART E: Beginning to Think About a Final Model

Task 9:

Looking at Model 4, I want to hypothesize that a good mix of explanatory variables would incorporate variables that would explain qualitative features (e.g.: quality of the house) and quantitative features (e.g.: total square feet). My approach would be as follows:

- Identify qualitative and quantitative variables that can provide meaningful information.
- We have approximately 3000 observations so perhaps we can do a grid search or a random grid search to evaluate the relationship across all feature.
- Evaluate outliers and incomplete observations.
- Evaluate several statistics such as DFFITS, Cook's Distance, Leverage, and Influence that can improve the model.

As I think more about how to approach this for our next assignment, I am sure my approach will adjust accordingly.

Conclusion and Reflections:

This was another intense assignment and one of the most challenging assignment that I have faced in the program. It felt like a marathon of tasks to perform; though, I did enjoy doing them as sometimes repetition is key to develop and enhance an approach to modeling.

We create multiple models, conducting t-statistics on each variable and f-statistics on the model. The models were assessed using the R-squared value. In addition, we created plots (e.g.: histograms of residuals, scatter plots comparing residuals and predicted values, etc.), using the predicted values and the actual y (SalesPrice) values and then reviewed the residuals.

I am glad we had the opportunity to learn about new concepts such as DFFITS and influential points. DFFITS shows us how the predicted value changes if certain observations are excluded from our dataset. If we exclude these observations from our model and then refit it then the predicted values will also change. DFFITS statistic is a measure of how the predicted value changes when the observation is removed. It helped me understand that models can always be made better, albeit with according tradeoffs.

I am excited to see how I can continue to fine-tune my approach when it comes to modeling. As I become more comfortable with statistics and evaluating each variable to see how it influences the overall model, I am learning to develop an approach that can be reusable for different business cases. In the next model, I do want to identify my variables for the model based on some cross-validation technique and use other ways of identifying models than correlation or based on our knowledge of the given situation.

Though, I need to set aside more time to go through the assignment as it seems I am not moving or coding fast enough to keep up!

Appendix

A: Data Quality Check (mydata)

Dimensions: 2930 x 87
Duplicates: 0

No	Variable	Stats / Values	Freqs (% of Valid)	Valid	Missing
1	SID [integer]	Mean (sd) : 1465.5 (846) min < med < max: 1 < 1465.5 < 2930 IQR (CV) : 1464.5 (0.6)	2930 distinct values (Integer sequence)	2930 (100%)	0 (0%)
2	PID [integer]	Mean (sd) : 714464497 (188730844.6) min < med < max: 526301100 < 535453620 < 1007100110 IQR (CV) : 378704075 (0.3)	2930 distinct values	2930 (100%)	0 (0%)
3	SubClass [integer]	Mean (sd) : 57.4 (42.6) min < med < max: 20 < 50 < 190 IQR (CV) : 50 (0.7)	16 distinct values	2930 (100%)	0 (0%)
4	Zoning [factor]	1. A (agr) 2. C (all) 3. FV 4. I (all) 5. RH 6. RL 7. RM	2 (0.1%) 25 (0.9%) 139 (4.7%) 2 (0.1%) 27 (0.9%) 2273 (77.6%) 462 (15.8%)	2930 (100%)	0 (0%)
5	LotFrontage [integer]	Mean (sd) : 69.2 (23.4) min < med < max: 21 < 68 < 313 IQR (CV) : 22 (0.3)	128 distinct values	2440 (83.28%)	490 (16.72%)
6	LotArea [integer]	Mean (sd) : 10147.9 (7880) min < med < max: 1300 < 9436.5 < 215245 IQR (CV) : 4115 (0.8)	1960 distinct values	2930 (100%)	0 (0%)
7	Street [factor]	1. Grvl 2. Pave	12 (0.4%) 2918 (99.6%)	2930 (100%)	0 (0%)
8	Alley [factor]	1. Grvl 2. Pave	120 (60.6%) 78 (39.4%)	198 (6.76%)	2732 (93.24%)
9	LotShape [factor]	1. IR1 2. IR2 3. IR3 4. Reg	979 (33.4%) 76 (2.6%) 16 (0.5%) 1859 (63.4%)	2930 (100%)	0 (0%)
10	LandContour [factor]	1. Bnk 2. HLS 3. Low 4. Lvl	117 (4.0%) 120 (4.1%) 60 (2.1%) 2633 (89.9%)	2930 (100%)	0 (0%)

11	Utilities [factor]	1. AllPub 2. NoSeWa 3. NoSewr	2927 (99.9%) 1 (0.0%) 2 (0.1%)	2930 (100%)	0 (0%)
12	LotConfig [factor]	1. Corner 2. CulDSac 3. FR2 4. FR3 5. Inside	511 (17.4%) 180 (6.1%) 85 (2.9%) 14 (0.5%) 2140 (73.0%)	2930 (100%)	0 (0%)
13	LandSlope [factor]	1. Gtl 2. Mod 3. Sev	2789 (95.2%) 125 (4.3%) 16 (0.5%)	2930 (100%)	0 (0%)
14	Neighborhood [factor]	1. Blmngtn 2. Blueste 3. BrDale 4. BrkSide 5. ClearCr 6. CollgCr 7. Crawfor 8. Edwards 9. Gilbert 10. Greens [18 others]	28 (1.0%) 10 (0.3%) 30 (1.0%) 108 (3.7%) 44 (1.5%) 267 (9.1%) 103 (3.5%) 194 (6.6%) 165 (5.6%) 8 (0.3%) 1973 (67.3%)	2930 (100%)	0 (0%)
15	Condition1 [factor]	1. Artery 2. Feedr 3. Norm 4. PosA 5. PosN 6. RRAe 7. RRAn 8. RRNe 9. RRNn	92 (3.1%) 164 (5.6%) 2522 (86.1%) 20 (0.7%) 39 (1.3%) 28 (1.0%) 50 (1.7%) 6 (0.2%) 9 (0.3%)	2930 (100%)	0 (0%)
16	Condition2 [factor]	1. Artery 2. Feedr 3. Norm 4. PosA 5. PosN 6. RRAe 7. RRAn 8. RRNn	5 (0.2%) 13 (0.4%) 2900 (99.0%) 4 (0.1%) 4 (0.1%) 1 (0.0%) 1 (0.0%) 2 (0.1%)	2930 (100%)	0 (0%)
17	BldgType [factor]	1. 1Fam 2. 2fmCon 3. Duplex 4. Twnhs 5. TwnhsE	2425 (82.8%) 62 (2.1%) 109 (3.7%) 101 (3.5%) 233 (8.0%)	2930 (100%)	0 (0%)
18	HouseStyle [factor]	1. 1.5Fin 2. 1.5Unf 3. 1Story 4. 2.5Fin 5. 2.5Unf 6. 2Story 7. SFoyer 8. SLvl	314 (10.7%) 19 (0.7%) 1481 (50.5%)	2930 (100%)	0 (0%)

			8 (0.3%) 24 (0.8%) 873 (29.8%) 83 (2.8%) 128 (4.4%)		
19	OverallQual [integer]	Mean (sd) : 6.1 (1.4) min < med < max: 1 < 6 < 10 IQR (CV) : 2 (0.2)	1 : 4 (0.1%) 2 : 13 (0.4%) 3 : 40 (1.4%) 4 : 226 (7.7%) 5 : 825 (28.2%) 6 : 732 (25.0%) 7 : 602 (20.5%) 8 : 350 (11.9%) 9 : 107 (3.6%) 10 : 31 (1.1%)	2930 (100%)	0 (0%)
20	OverallCond [integer]	Mean (sd) : 5.6 (1.1) min < med < max: 1 < 5 < 9 IQR (CV) : 1 (0.2)	1 : 7 (0.2%) 2 : 10 (0.3%) 3 : 50 (1.7%) 4 : 101 (3.5%) 5 : 1654 (56.5%) 6 : 533 (18.2%) 7 : 390 (13.3%) 8 : 144 (4.9%) 9 : 41 (1.4%)	2930 (100%)	0 (0%)
21	YearBuilt [integer]	Mean (sd) : 1971.4 (30.2) min < med < max: 1872 < 1973 < 2010 IQR (CV) : 47 (0)	118 distinct values	2930 (100%)	0 (0%)
22	YearRemodel [integer]	Mean (sd) : 1984.3 (20.9) min < med < max: 1950 < 1993 < 2010 IQR (CV) : 39 (0)	61 distinct values	2930 (100%)	0 (0%)
23	RoofStyle [factor]	1. Flat 2. Gable 3. Gambrel 4. Hip 5. Mansard 6. Shed	20 (0.7%) 2321 (79.2%) 22 (0.8%) 551 (18.8%) 11 (0.4%) 5 (0.2%)	2930 (100%)	0 (0%)
24	RoofMat [factor]	1. ClyTile 2. CompShg 3. Membran 4. Metal 5. Roll 6. Tar&Grv 7. WdShake 8. WdShngl	1 (0.0%) 2887 (98.5%) 1 (0.0%) 1 (0.0%) 1 (0.0%) 23 (0.8%) 9 (0.3%) 7 (0.2%)	2930 (100%)	0 (0%)

25	Exterior1 [factor]	1. AsbShng 2. AsphShn 3. BrkComm 4. BrkFace 5. CBlock 6. CemntBd 7. HdBoard 8. ImStucc 9. MetalSd 10. Plywood [6 others]	44 (1.5%) 2 (0.1%) 6 (0.2%) 88 (3.0%) 2 (0.1%) 126 (4.3%) 442 (15.1%) 1 (0.0%) 450 (15.4%) 221 (7.5%) 1548 (52.8%)	2930 (100%)	0 (0%)
26	Exterior2 [factor]	1. AsbShng 2. AsphShn 3. Brk Cmn 4. BrkFace 5. CBlock 6. CmentBd 7. HdBoard 8. ImStucc 9. MetalSd 10. Other [7 others]	38 (1.3%) 4 (0.1%) 22 (0.8%) 47 (1.6%) 3 (0.1%) 126 (4.3%) 406 (13.9%) 15 (0.5%) 447 (15.3%) 1 (0.0%) 1821 (62.2%)	2930 (100%)	0 (0%)
27	MasVnrType [factor]	1. (Empty string) 2. BrkCmn 3. BrkFace 4. CBlock 5. None 6. Stone	23 (0.8%) 25 (0.9%) 880 (30.0%) 1 (0.0%) 1752 (59.8%) 249 (8.5%)	2930 (100%)	0 (0%)
28	MasVnrArea [integer]	Mean (sd) : 101.9 (179.1) min < med < max: 0 < 0 < 1600 IQR (CV) : 164 (1.8)	445 distinct values	2907 (99.22%)	23 (0.78%)
29	ExterQual [factor]	1. Ex 2. Fa 3. Gd 4. TA	107 (3.6%) 35 (1.2%) 989 (33.8%) 1799 (61.4%)	2930 (100%)	0 (0%)
30	ExterCond [factor]	1. Ex 2. Fa 3. Gd 4. Po 5. TA	12 (0.4%) 67 (2.3%) 299 (10.2%) 3 (0.1%) 2549 (87.0%)	2930 (100%)	0 (0%)
31	Foundation [factor]	1. BrkTil 2. CBlock 3. PConc 4. Slab 5. Stone 6. Wood	311 (10.6%) 1244 (42.5%) 1310 (44.7%) 49 (1.7%) 11 (0.4%) 5 (0.2%)	2930 (100%)	0 (0%)

32	BsmtQual [factor]	1. (Empty string) 2. Ex 3. Fa 4. Gd 5. Po 6. TA	1 (0.0%) 258 (9.0%) 88 (3.1%) 1219 (42.8%) 2 (0.1%) 1283 (45.0%)	2851 (97.3%)	79 (2.7%)
33	BsmtCond [factor]	1. (Empty string) 2. Ex 3. Fa 4. Gd 5. Po 6. TA	1 (0.0%) 3 (0.1%) 104 (3.6%) 122 (4.3%) 5 (0.2%) 2616 (91.8%)	2851 (97.3%)	79 (2.7%)
34	BsmtExposure [factor]	1. (Empty string) 2. Av 3. Gd 4. Mn 5. No	4 (0.1%) 418 (14.7%) 284 (10.0%) 239 (8.4%) 1906 (66.8%)	2851 (97.3%)	79 (2.7%)
35	BsmtFinType1 [factor]	1. (Empty string) 2. ALQ 3. BLQ 4. GLQ 5. LwQ 6. Rec 7. Unf	1 (0.0%) 429 (15.0%) 269 (9.4%) 859 (30.1%) 154 (5.4%) 288 (10.1%) 851 (29.8%)	2851 (97.3%)	79 (2.7%)
36	BsmtFinSF1 [integer]	Mean (sd) : 442.6 (455.6) min < med < max: 0 < 370 < 5644 IQR (CV) : 734 (1)	995 distinct values	2929 (99.97%)	1 (0.03%)
37	BsmtFinType2 [factor]	1. (Empty string) 2. ALQ 3. BLQ 4. GLQ 5. LwQ 6. Rec 7. Unf	2 (0.1%) 53 (1.9%) 68 (2.4%) 34 (1.2%) 89 (3.1%) 106 (3.7%) 2499 (87.6%)	2851 (97.3%)	79 (2.7%)
38	BsmtFinSF2 [integer]	Mean (sd) : 49.7 (169.2) min < med < max: 0 < 0 < 1526 IQR (CV) : 0 (3.4)	274 distinct values	2929 (99.97%)	1 (0.03%)
39	BsmtUnfSF [integer]	Mean (sd) : 559.3 (439.5) min < med < max: 0 < 466 < 2336 IQR (CV) : 583 (0.8)	1137 distinct values	2929 (99.97%)	1 (0.03%)
40	TotalBsmtSF [integer]	Mean (sd) : 1051.6 (440.6) min < med < max: 0 < 990 < 6110 IQR (CV) : 509 (0.4)	1058 distinct values	2929 (99.97%)	1 (0.03%)
41	Heating [factor]	1. Floor 2. GasA 3. GasW 4. Grav 5. OthW 6. Wall	1 (0.0%) 2885 (98.5%) 27 (0.9%) 9 (0.3%)	2930 (100%)	0 (0%)

			2 (0.1%) 6 (0.2%)		
42	HeatingQC [factor]	1. Ex 2. Fa 3. Gd 4. Po 5. TA	1495 (51.0%) 92 (3.1%) 476 (16.2%) 3 (0.1%) 864 (29.5%)	2930 (100%)	0 (0%)
43	CentralAir [factor]	1. N 2. Y	196 (6.7%) 2734 (93.3%)	2930 (100%)	0 (0%)
44	Electrical [factor]	1. (Empty string) 2. FuseA 3. FuseF 4. FuseP 5. Mix 6. SBrkr	1 (0.0%) 188 (6.4%) 50 (1.7%) 8 (0.3%) 1 (0.0%) 2682 (91.5%)	2930 (100%)	0 (0%)
45	FirstFlrSF [integer]	Mean (sd) : 1159.6 (391.9) min < med < max: 334 < 1084 < 5095 IQR (CV) : 507.8 (0.3)	1083 distinct values	2930 (100%)	0 (0%)
46	SecondFlrSF [integer]	Mean (sd) : 335.5 (428.4) min < med < max: 0 < 0 < 2065 IQR (CV) : 703.8 (1.3)	635 distinct values	2930 (100%)	0 (0%)
47	LowQualFinSF [integer]	Mean (sd) : 4.7 (46.3) min < med < max: 0 < 0 < 1064 IQR (CV) : 0 (9.9)	36 distinct values	2930 (100%)	0 (0%)
48	GrLivArea [integer]	Mean (sd) : 1499.7 (505.5) min < med < max: 334 < 1442 < 5642 IQR (CV) : 616.8 (0.3)	1292 distinct values	2930 (100%)	0 (0%)
49	BsmtFullBath [integer]	Mean (sd) : 0.4 (0.5) min < med < max: 0 < 0 < 3 IQR (CV) : 1 (1.2)	0 : 1707 (58.3%) 1 : 1181 (40.3%) 2 : 38 (1.3%) 3 : 2 (0.1%)	2928 (99.93%)	2 (0.07%)
50	BsmtHalfBath [integer]	Mean (sd) : 0.1 (0.2) min < med < max: 0 < 0 < 2 IQR (CV) : 0 (4)	0 : 2753 (94.0%) 1 : 171 (5.8%) 2 : 4 (0.1%)	2928 (99.93%)	2 (0.07%)
51	FullBath [integer]	Mean (sd) : 1.6 (0.6) min < med < max: 0 < 2 < 4 IQR (CV) : 1 (0.4)	0 : 12 (0.4%) 1 : 1318 (45.0%) 2 : 1532 (52.3%) 3 : 64 (2.2%) 4 : 4 (0.1%)	2930 (100%)	0 (0%)
52	HalfBath [integer]	Mean (sd) : 0.4 (0.5) min < med < max: 0 < 0 < 2 IQR (CV) : 1 (1.3)	0 : 1843 (62.9%) 1 : 1062 (36.2%) 2 : 25 (0.9%)	2930 (100%)	0 (0%)
53	BedroomAbvGr [integer]	Mean (sd) : 2.9 (0.8) min < med < max: 0 < 3 < 8 IQR (CV) : 1 (0.3)	0 : 8 (0.3%) 1 : 112 (3.8%) 2 : 743 (25.4%)	2930 (100%)	0 (0%)

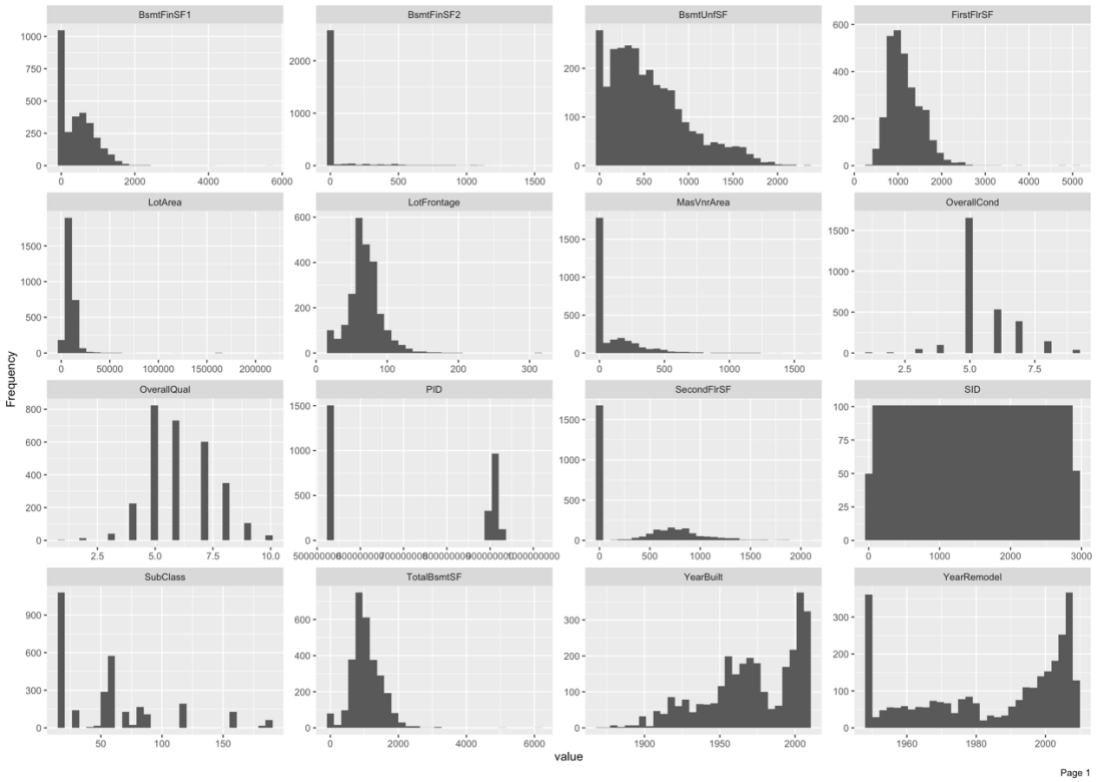
			3 : 1597 (54.5%) 4 : 400 (13.7%) 5 : 48 (1.6%) 6 : 21 (0.7%) 8 : 1 (0.0%)		
54	KitchenAbvGr [integer]	Mean (sd) : 1 (0.2) min < med < max: 0 < 1 < 3 IQR (CV) : 0 (0.2)	0 : 3 (0.1%) 1 : 2796 (95.4%) 2 : 129 (4.4%) 3 : 2 (0.1%)	2930 (100%)	0 (0%)
55	KitchenQual [factor]	1. Ex 2. Fa 3. Gd 4. Po 5. TA	205 (7.0%) 70 (2.4%) 1160 (39.6%) 1 (0.0%) 1494 (51.0%)	2930 (100%)	0 (0%)
56	TotRmsAbvGrd [integer]	Mean (sd) : 6.4 (1.6) min < med < max: 2 < 6 < 15 IQR (CV) : 2 (0.2)	14 distinct values	2930 (100%)	0 (0%)
57	Functional [factor]	1. Maj1 2. Maj2 3. Min1 4. Min2 5. Mod 6. Sal 7. Sev 8. Typ	19 (0.7%) 9 (0.3%) 65 (2.2%) 70 (2.4%) 35 (1.2%) 2 (0.1%) 2 (0.1%) 2728 (93.1%)	2930 (100%)	0 (0%)
58	Fireplaces [integer]	Mean (sd) : 0.6 (0.6) min < med < max: 0 < 1 < 4 IQR (CV) : 1 (1.1)	0 : 1422 (48.5%) 1 : 1274 (43.5%) 2 : 221 (7.5%) 3 : 12 (0.4%) 4 : 1 (0.0%)	2930 (100%)	0 (0%)
59	FireplaceQu [factor]	1. Ex 2. Fa 3. Gd 4. Po 5. TA	43 (2.9%) 75 (5.0%) 744 (49.3%) 46 (3.0%) 600 (39.8%)	1508 (51.47%)	1422 (48.53%)
60	GarageType [factor]	1. 2Types 2. Attchd 3. Basment 4. BuiltIn 5. CarPort 6. Detchd	23 (0.8%) 1731 (62.4%) 36 (1.3%) 186 (6.7%) 15 (0.5%) 782 (28.2%)	2773 (94.64%)	157 (5.36%)
61	GarageYrBlt [integer]	Mean (sd) : 1978.1 (25.5) min < med < max: 1895 < 1979 < 2207 IQR (CV) : 42 (0)	103 distinct values	2771 (94.57%)	159 (5.43%)
62	GarageFinish [factor]	1. (Empty string) 2. Fin 3. RFn 4. Unf	2 (0.1%) 728 (26.2%)	2773 (94.64%)	157 (5.36%)

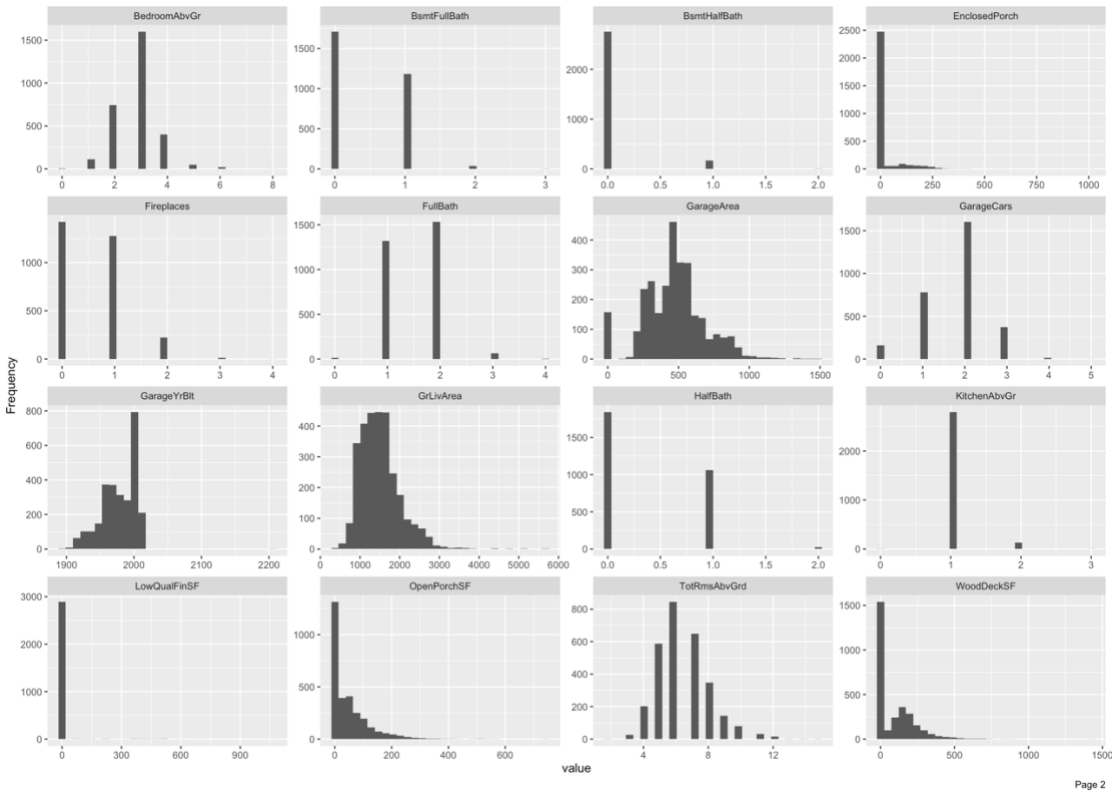
			812 (29.3%) 1231 (44.4%)		
63	GarageCars [integer]	Mean (sd) : 1.8 (0.8) min < med < max: 0 < 2 < 5 IQR (CV) : 1 (0.4)	0 : 157 (5.4%) 1 : 778 (26.6%) 2 : 1603 (54.7%) 3 : 374 (12.8%) 4 : 16 (0.5%) 5 : 1 (0.0%)	2929 (99.97%)	1 (0.03%)
64	GarageArea [integer]	Mean (sd) : 472.8 (215) min < med < max: 0 < 480 < 1488 IQR (CV) : 256 (0.5)	603 distinct values	2929 (99.97%)	1 (0.03%)
65	GarageQual [factor]	1. (Empty string) 2. Ex 3. Fa 4. Gd 5. Po 6. TA	1 (0.0%) 3 (0.1%) 124 (4.5%) 24 (0.9%) 5 (0.2%) 2615 (94.3%)	2772 (94.61%)	158 (5.39%)
66	GarageCond [factor]	1. (Empty string) 2. Ex 3. Fa 4. Gd 5. Po 6. TA	1 (0.0%) 3 (0.1%) 74 (2.7%) 15 (0.5%) 14 (0.5%) 2665 (96.1%)	2772 (94.61%)	158 (5.39%)
67	PavedDrive [factor]	1. N 2. P 3. Y	216 (7.4%) 62 (2.1%) 2652 (90.5%)	2930 (100%)	0 (0%)
68	WoodDeckSF [integer]	Mean (sd) : 93.8 (126.4) min < med < max: 0 < 0 < 1424 IQR (CV) : 168 (1.3)	380 distinct values	2930 (100%)	0 (0%)
69	OpenPorchSF [integer]	Mean (sd) : 47.5 (67.5) min < med < max: 0 < 27 < 742 IQR (CV) : 70 (1.4)	252 distinct values	2930 (100%)	0 (0%)
70	EnclosedPorch [integer]	Mean (sd) : 23 (64.1) min < med < max: 0 < 0 < 1012 IQR (CV) : 0 (2.8)	183 distinct values	2930 (100%)	0 (0%)
71	ThreeSsnPorch [integer]	Mean (sd) : 2.6 (25.1) min < med < max: 0 < 0 < 508 IQR (CV) : 0 (9.7)	31 distinct values	2930 (100%)	0 (0%)
72	ScreenPorch [integer]	Mean (sd) : 16 (56.1) min < med < max: 0 < 0 < 576 IQR (CV) : 0 (3.5)	121 distinct values	2930 (100%)	0 (0%)
73	PoolArea [integer]	Mean (sd) : 2.2 (35.6) min < med < max: 0 < 0 < 800 IQR (CV) : 0 (15.9)	14 distinct values	2930 (100%)	0 (0%)
74	PoolQC [factor]	1. Ex 2. Fa 3. Gd 4. TA	4 (30.8%) 2 (15.4%) 4 (30.8%)	13 (0.44%)	2917 (99.56%)

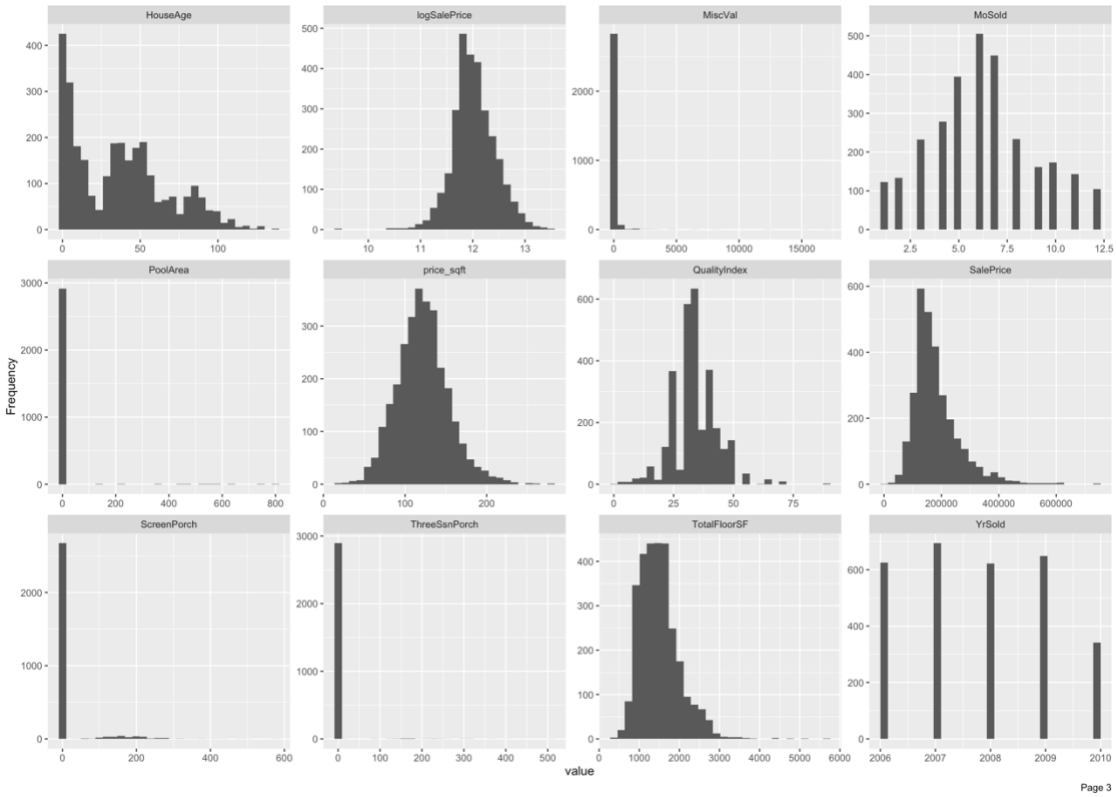
			3 (23.1%)		
75	Fence [factor]	1. GdPrv 2. GdWo 3. MnPrv 4. MnWw	118 (20.6%) 112 (19.6%) 330 (57.7%) 12 (2.1%)	572 (19.52%)	2358 (80.48%)
76	MiscFeature [factor]	1. Elev 2. Gar2 3. Othr 4. Shed 5. TenC	1 (0.9%) 5 (4.7%) 4 (3.8%) 95 (89.6%) 1 (0.9%)	106 (3.62%)	2824 (96.38%)
77	MiscVal [integer]	Mean (sd) : 50.6 (566.3) min < med < max: 0 < 0 < 17000 IQR (CV) : 0 (11.2)	38 distinct values	2930 (100%)	0 (0%)
78	MoSold [integer]	Mean (sd) : 6.2 (2.7) min < med < max: 1 < 6 < 12 IQR (CV) : 4 (0.4)	12 distinct values	2930 (100%)	0 (0%)
79	YrSold [integer]	Mean (sd) : 2007.8 (1.3) min < med < max: 2006 < 2008 < 2010 IQR (CV) : 2 (0)	2006 : 625 (21.3%) 2007 : 694 (23.7%) 2008 : 622 (21.2%) 2009 : 648 (22.1%) 2010 : 341 (11.6%)	2930 (100%)	0 (0%)
80	SaleType [factor]	1. COD 2. Con 3. ConLD 4. ConLI 5. ConLw 6. CWD 7. New 8. Oth 9. VWD 10. WD .	87 (3.0%) 5 (0.2%) 26 (0.9%) 9 (0.3%) 8 (0.3%) 12 (0.4%) 239 (8.2%) 7 (0.2%) 1 (0.0%) 2536 (86.6%)	2930 (100%)	0 (0%)
81	SaleCondition [factor]	1. Abnorml 2. AdjLand 3. Alloca 4. Family 5. Normal 6. Partial	190 (6.5%) 12 (0.4%) 24 (0.8%) 46 (1.6%) 2413 (82.3%) 245 (8.4%)	2930 (100%)	0 (0%)
82	SalePrice [integer]	Mean (sd) : 180796.1 (79886.7) min < med < max: 12789 < 160000 < 755000 IQR (CV) : 84000 (0.4)	1032 distinct values	2930 (100%)	0 (0%)
83	TotalFloorSF [integer]	Mean (sd) : 1495 (503.1) min < med < max: 334 < 1440 < 5642 IQR (CV) : 620 (0.3)	1289 distinct values	2930 (100%)	0 (0%)

84	HouseAge [integer]	Mean (sd) : 36.4 (30.3) min < med < max: -1 < 34 < 136 IQR (CV) : 47 (0.8)	128 distinct values	2930 (100%)	0 (0%)
85	QualityIndex [integer]	Mean (sd) : 33.8 (9.2) min < med < max: 1 < 35 < 90 IQR (CV) : 10 (0.3)	36 distinct values	2930 (100%)	0 (0%)
86	logSalePrice [numeric]	Mean (sd) : 12 (0.4) min < med < max: 9.5 < 12 < 13.5 IQR (CV) : 0.5 (0)	1032 distinct values	2930 (100%)	0 (0%)
87	price_sqft [numeric]	Mean (sd) : 121.6 (31.9) min < med < max: 15.4 < 120.4 < 276.3 IQR (CV) : 39.4 (0.3)	2841 distinct values	2930 (100%)	0 (0%)

B: Histogram distribution of features







C: Model 1 (ANNOVA and Summary Tables)

```
Call:
lm(formula = subdat2$SalePrice ~ subdat2$TotalFloorSF)

Residuals:
    Min      1Q  Median      3Q     Max
-202117  -30105   -1431   23229  328000

Coefficients:
                    Estimate Std. Error t value            Pr(>|t|)
(Intercept)         4623.800   3230.158   1.431               0.152
subdat2$TotalFloorSF  118.034      2.063  57.219 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 53960 on 2923 degrees of freedom
Multiple R-squared:  0.5283,    Adjusted R-squared:  0.5282
F-statistic:  3274 on 1 and 2923 DF,  p-value: < 0.00000000000000022
```

```
Analysis of Variance Table

Response: subdat2$SalePrice
                      Df        Sum Sq       Mean Sq F value
subdat2$TotalFloorSF    1 9532770925496 9532770925496    3274
Residuals            2923 8510840098265    2911679815
                                   Pr(>F)
subdat2$TotalFloorSF < 0.00000000000000022 ***
Residuals
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

D: Model 2 (ANNOVA and Summary Tables)

```
Call:
lm(formula = SalePrice ~ OverallQual, data = subdat3)

Residuals:
    Min      1Q  Median      3Q     Max
-161696  -29022   -2347   21472  299961

Coefficients:
            Estimate Std. Error t value            Pr(>|t|)
(Intercept) -94088.4     3836.7  -24.52 <0.0000000000000002 ***
OverallQual  45087.2      614.1   73.42 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 46590 on 2923 degrees of freedom
Multiple R-squared:  0.6484,    Adjusted R-squared:  0.6483
F-statistic:  5391 on 1 and 2923 DF,  p-value: < 0.00000000000000022
```

```
Analysis of Variance Table

Response: SalePrice
              Df         Sum Sq        Mean Sq F value
OverallQual    1 11699566333793 11699566333793  5390.5
Residuals   2923  6344044689968     2170388194
                           Pr(>F)
OverallQual < 0.00000000000000022 ***
Residuals
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

E: Model 3 (ANNOVA and Summary Tables)

```
Call:
lm(formula = SalePrice ~ OverallQual + TotalFloorSF, data = subdat4)

Residuals:
    Min      1Q  Median      3Q     Max
-218220  -24099   -1091   20445  280691

Coefficients:
               Estimate  Std. Error t value            Pr(>|t|)
(Intercept)  -112566.859    3254.802  -34.59 <0.0000000000000002 ***
OverallQual    32410.371     627.075   51.69 <0.0000000000000002 ***
TotalFloorSF      64.230       1.819   35.32 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 39010 on 2922 degrees of freedom
Multiple R-squared:  0.7536,    Adjusted R-squared:  0.7534
F-statistic:  4468 on 2 and 2922 DF,  p-value: < 0.00000000000000022
```

```
Analysis of Variance Table

Response: SalePrice
               Df         Sum Sq        Mean Sq F value
OverallQual     1 11699566333793 11699566333793  7689.0
TotalFloorSF    1  1897919584110  1897919584110  1247.3
Residuals    2922  4446125105858     1521603390
                          Pr(>F)
OverallQual  < 0.00000000000000022 ***
TotalFloorSF < 0.00000000000000022 ***
Residuals
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

F: Model 4 (ANNOVA and Summary Tables)

```
Call:
lm(formula = SalePrice ~ OverallQual + TotalFloorSF + TotalBsmtSF,
    data = subdat5)

Residuals:
    Min      1Q  Median      3Q     Max
-186816  -19821    -251   18006  242615

Coefficients:
               Estimate  Std. Error t value            Pr(>|t|)
(Intercept)  -113328.302    2920.351  -38.81 <0.0000000000000002 ***
OverallQual    25493.956     619.167   41.17 <0.0000000000000002 ***
TotalFloorSF      58.279       1.646   35.41 <0.0000000000000002 ***
TotalBsmtSF       49.422       1.850   26.71 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 34980 on 2920 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.8019,    Adjusted R-squared:  0.8017
F-statistic:  3939 on 3 and 2920 DF,  p-value: < 0.00000000000000022
```

```
Analysis of Variance Table

Response: SalePrice
              Df         Sum Sq        Mean Sq F value
OverallQual    1 11689331279084 11689331279084 9553.23
TotalFloorSF   1  1897885979293  1897885979293 1551.07
TotalBsmtSF    1   873195380061   873195380061  713.63
Residuals   2920  3572910560635     1223599507
                           Pr(>F)
OverallQual  < 0.00000000000000022 ***
TotalFloorSF < 0.00000000000000022 ***
TotalBsmtSF  < 0.00000000000000022 ***
Residuals
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

G: Model 5 (ANNOVA and Summary Tables)

```
Call:
lm(formula = logSalePrice ~ TotalFloorSF + OverallQual + TotalBsmtSF,
    data = subdat7)

Residuals:
     Min       1Q   Median       3Q      Max
-0.91364 -0.07870  0.01669  0.10302  0.58602

Coefficients:
                Estimate   Std. Error t value            Pr(>|t|)
(Intercept)  10.560756253  0.015491976  681.69 <0.0000000000000002 ***
TotalFloorSF  0.000297471  0.000008877   33.51 <0.0000000000000002 ***
OverallQual   0.129468109  0.003509884   36.89 <0.0000000000000002 ***
TotalBsmtSF   0.000231694  0.000009946   23.30 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1602 on 2291 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.8239,    Adjusted R-squared:  0.8236
F-statistic:  3572 on 3 and 2291 DF,  p-value: < 0.00000000000000022
```

```
Analysis of Variance Table

Response: logSalePrice
              Df  Sum Sq Mean Sq F value                Pr(>F)
TotalFloorSF    1 195.720 195.720 7624.61 < 0.00000000000000022 ***
OverallQual     1  65.398  65.398 2547.71 < 0.00000000000000022 ***
TotalBsmtSF     1  13.931  13.931  542.71 < 0.00000000000000022 ***
Residuals    2291  58.809   0.026
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```