Michael Venit

MSDS 410

Modeling Assignment #2

Introduction

The Ames Housing dataset consists of 2,930 observations representing housing sales in Ames, IA from 2006 to 2010. The data contains 82 features of each property transaction. The variables, including the sale price, lend themselves to the creation of a model with the purpose of predicting sale price accurately. The sample population must be defined for future modeling work as well as performing exploratory analysis on different variables to determine their usefulness as potential predictors of sale price. The following analysis is meant to provide estimates of home values for typical homes in Ames, Iowa.

Defining the Sample Population

There are several sub-populations of housing transactions in the dataset. The objective of this particular analysis is to provide estimates of home values for 'typical' homes in Ames, Iowa, therefore a sub-sample of the original dataset was created. The table below represents a waterfall of filter conditions. These are conditions that have been filtered out of the original dataset in order to create a population that more accurately represents typical home sales.

Drop Condition	Observations Dropped	Remaining Observations
Non-Residential Zoning	29	2901
Multi-Family Homes	502	2399
Non-Normal Sale Conditions	411	1988
Recommended Exclusions	1	1987

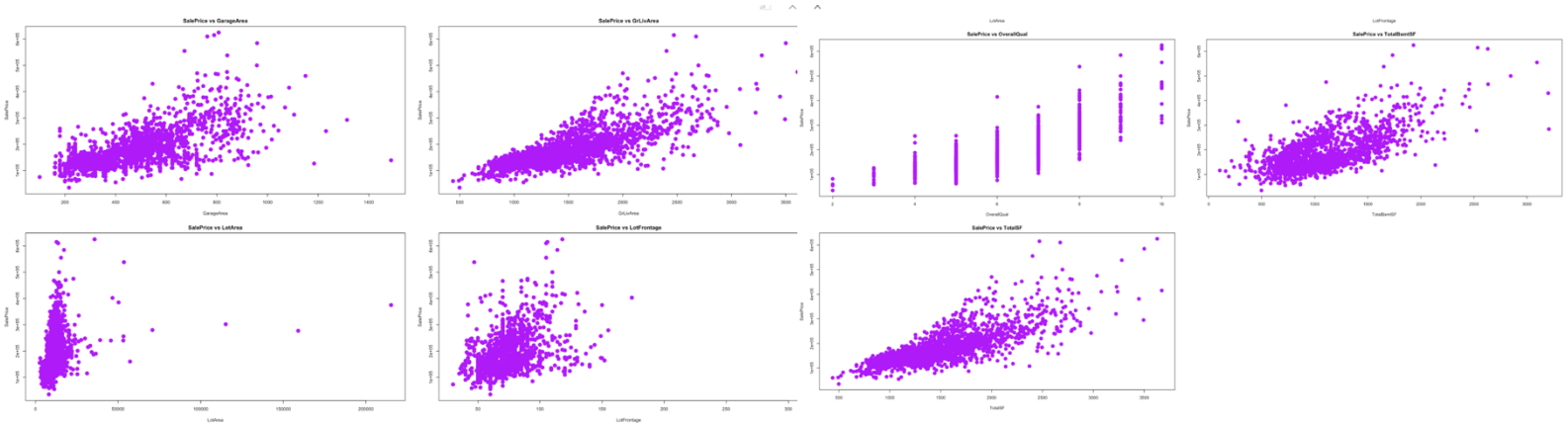
Non-residential zones were removed so the data would only contain residential areas. Additionally, multi-family homes were excluded along with non-normal sale conditions. The documentation associated with the dataset recommended removing properties greater than a certain threshold for above ground livable area (square feet) as they do not represent the population that is the focus of this analysis. The resultant sample for the analysis consists of 1,987 observations. The variables that will be included initially in the analysis are all continuous with the exception of overall quality (OverallQual) which is discrete. From the summary statistics table, a large number of the continuous variables can be observed to have their minimum value as zero. Checking for missing values prior to examining the distributions was also performed to verify this information. Below is a summary table of the variables with their count of missing values. The MasVnrArea variable will be imputed with zeros since it's possible that houses do not have that particular cosmetic feature. The LotFrontage variable will be imputed with the median value. This variable represents the linear feet of street connected to the property, so it would not make sense imputing this value to 0. After imputing the aforementioned missing values, the count of zero values was obtained as a proportion of total observations.

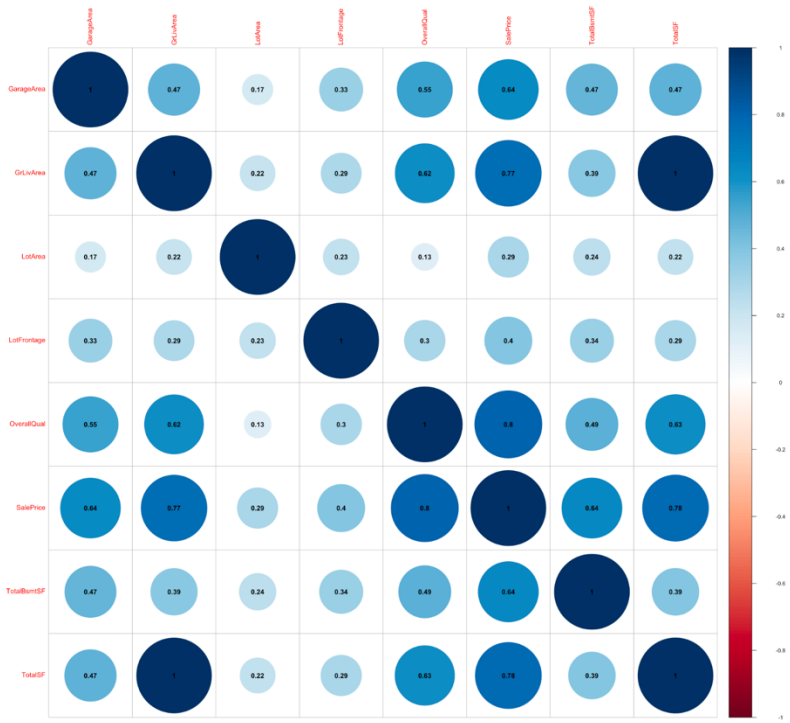
key	Max	Mean	Med	Min	SD
BsmtFinSF1	2288	441.30	390	0	419
BsmtFinSF2	1526	58.25	0	0	181
BsmtUnfSF	2336	534.68	458	0	402
FirstFlrSF	3820	1146.44	1065	334	360
GarageArea	1488	469.16	473	0	201
GrLivArea	3820	1495.67	1450	334	492
logSalePrice	13	12.03	12	10	0
LotArea	215245	10802.17	9750	2500	7772
LotFrontage	313	72.89	70	30	20
LowQualFinSF	1064	4.87	0	0	49
MasVnrArea	1600	94.20	0	0	171
MiscVal	15500	54.33	0	0	529
OpenPorchSF	570	46.36	25	0	65
OverallQual	10	6.01	6	1	1
PoolArea	800	2.16	0	0	35
price_sqft	249	121.42	120	45	27
SalePrice	625000	179600.25	162500	35000	71962
ScreenPorch	576	17.90	0	0	60
SecondFlrSF	1836	344.36	0	0	428
ThreeSsnPorch	508	2.84	0	0	27

Variable	NA Observations
LotFrontage	384
MasVnrArea	11

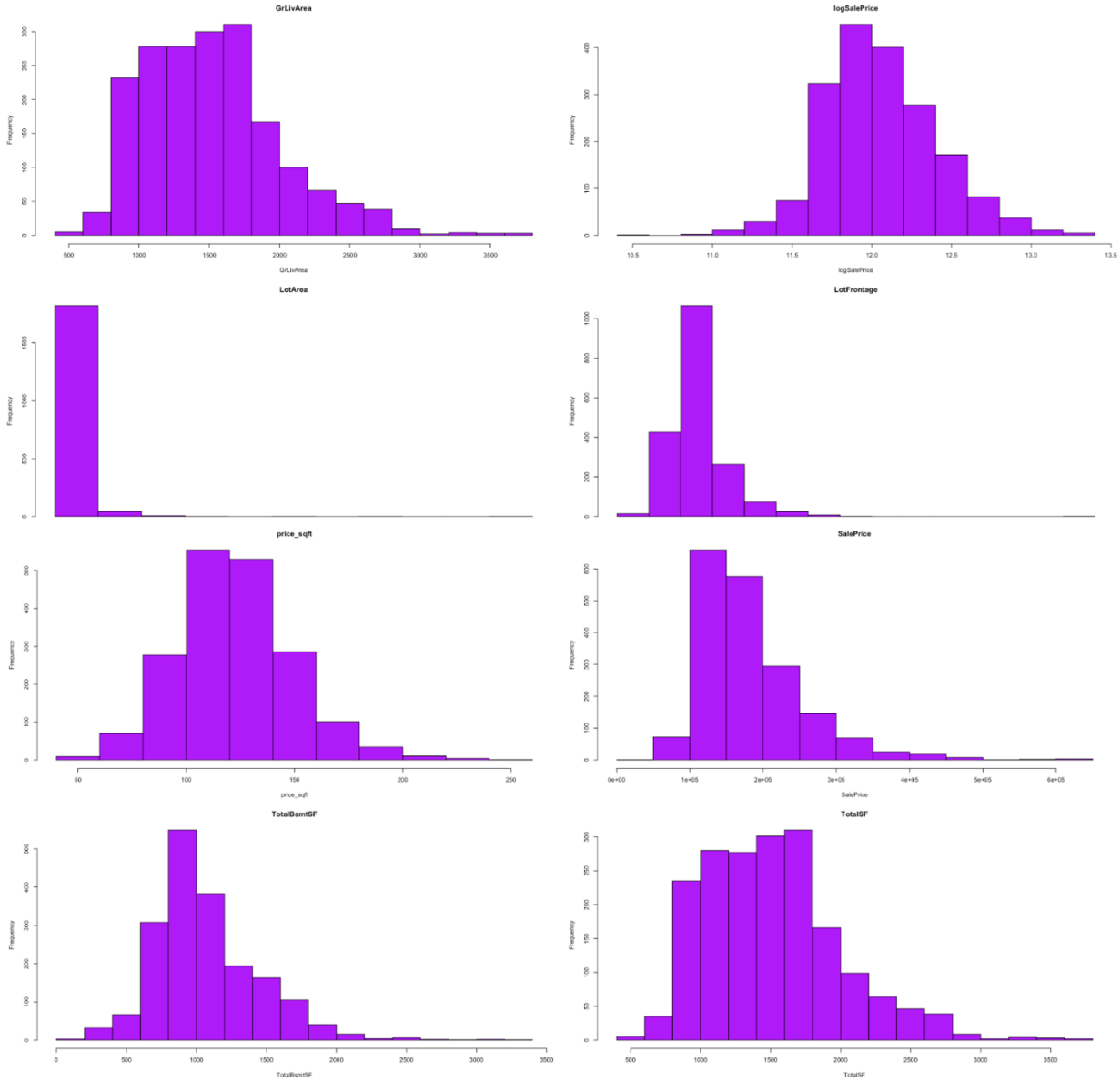
The table reveals that multiple features have a very large proportion of zero valued observations. For this reason, several variables were dropped from the analysis. It was decided that any variables that have > 5% of their observations as zero should be removed. Most of the continuous variables relate to a size measurement. After removal, we are left with the following variables: GarageArea, GrLivArea, logSalePrice, LotArea, LotFrontage, OverallQual, price_sqft, SalePrice, TotalBsmtSF, TotalSF. Since several of these variables have a very small percentage of zero valued observations, those cumulative observations will be dropped from the dataset. Below represents a scatterplot of the remaining dependent variables vs the SalePrice. We can see from the preliminary graphs that there are potentially significant outliers in some of the variables.

Variable	Zero Obs	Zero Obs Percent
BsmtFinSF1	589	30%
BsmtFinSF2	1708	87%
BsmtUnfSF	144	6%
FirstFlrSF	0	0%
GarageArea	71	3%
GrLivArea	0	0%
logSalePrice	0	0%
LotArea	0	0%
LotFrontage	0	0%
LowQualFinSF	1962	99%
MasVnrArea	1242	63%
MiscVal	1902	96%
OpenPorchSF	912	45%
OverallQual	0	0%
PoolArea	1978	99%
price_sqft	0	0%
SalePrice	0	0%
ScreenPorch	1798	90%
SecondFlrSF	1120	57%
ThreeSsnPorch	1962	99%

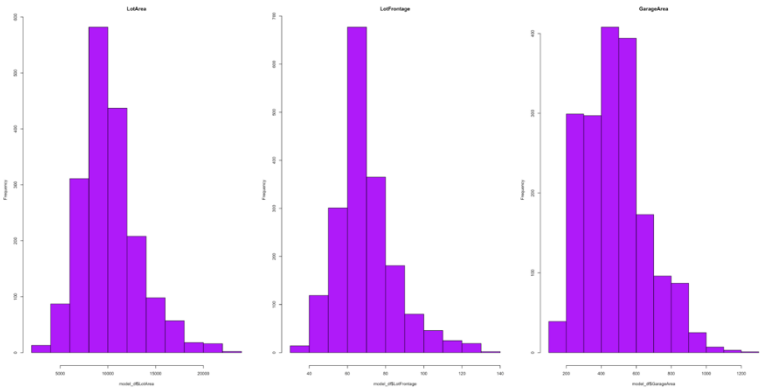




A correlation matrix was developed to determine which features bare a strong linear relationship with SalePrice. The above correlation matrix displays that several of the independent variables have a strong positive linear correlation with the dependent variable. TotalSF and OverallQual both have strong positive correlations. TotalSF and GrLivArea have a perfect positive linear correlation which makes sense as GrLivArea is encapsulated by TotalSF. This indicates that we can most likely drop GrLivArea from the model and it will not have an effect. Below we can see histograms for all of our continuous variables. It is apparent that LotArea and LotFrontage both have a heavy right skew due to outliers. SalePrice also appears to have a slight right skew. logSalePrice appears to be a better response variable.



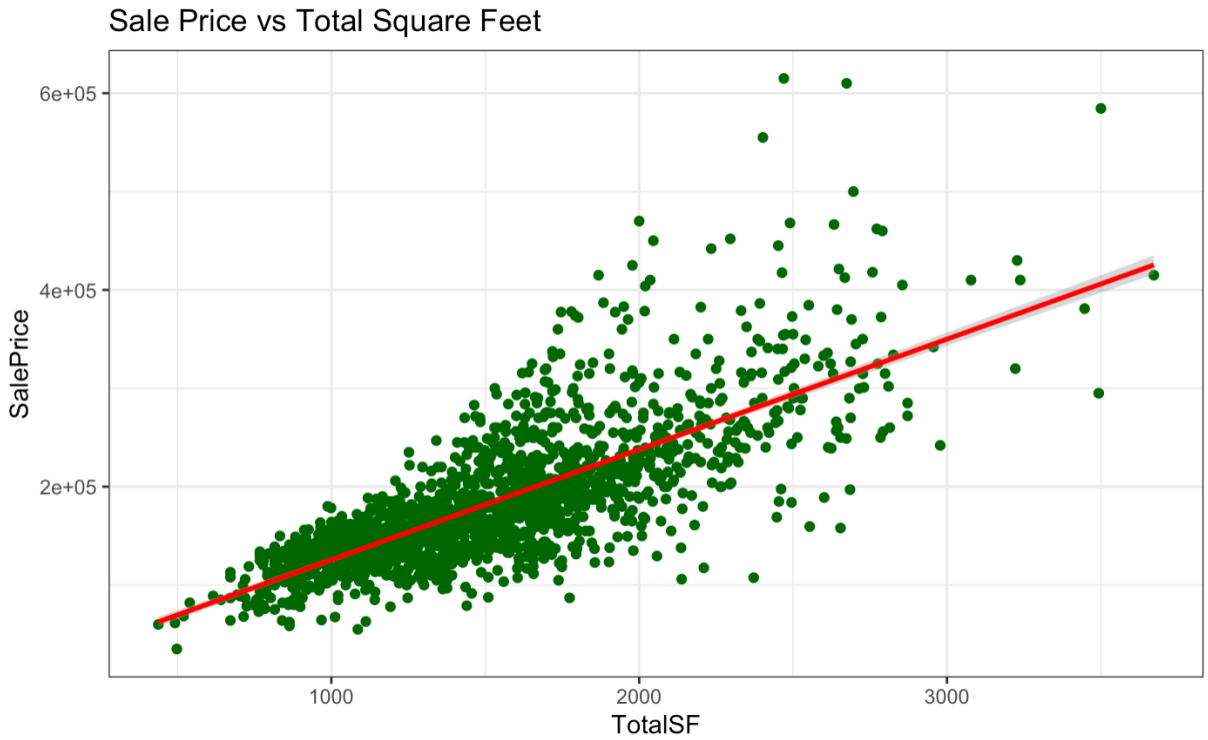
Based on the above histograms and scatterplots, it was decided that extreme outlier values (3 * IQR) from LotArea, LotFrontage, and GarageArea should be removed. Below are the histograms of the three variables after the extreme outlier observations have been removed.



PART A: Simple Linear Regression Models

1. Best Independent Variable

1a). TotalSF was the first variable chosen for modeling due to the fact that this feature demonstrated the highest linear correlation with SalePrice (highest of the continuous variables).



1b). Below is the summary computed from Model 1 along with the regression equation for this model

```
Call:
lm(formula = model_df$SalePrice ~ model_df$TotalSF)

Residuals:
    Min      1Q  Median      3Q     Max
-172070  -24028   -2243   18991  324439

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      13549.890   3421.962    3.96 7.79e-05 ***
model_df$TotalSF   112.150      2.173   51.62  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 44220 on 1827 degrees of freedom
Multiple R-squared:  0.5932,    Adjusted R-squared:  0.593
F-statistic:  2664 on 1 and 1827 DF,  p-value: < 2.2e-16
```

$$\hat{Y} = 13549 + 112.150X$$

The intercept term, 13549, is representative of the SalePrice if the TotalSF were equal to zero. This holds no value contextually. The regression coefficient of 112.150 represents a $112.15 increase in SalePrice for a single unit increase in TotalSF (assuming all else held constant).

1c). The R-squared value for Model 1 is 0.5932. This means that the TotalSF variable accounts for approximately 59% of the variance in SalePrice. Contextually it makes sense that TotalSF is a main driver of price as homes are typically priced at a per/sqft amount. As the size increases in square footage, so does the SalePrice.

1d). Hypothesis tests were conducted on the model coefficient using a two-sided t-test. Prior to stating the hypothesis, the established type I error threshold was set at alpha=0.05. The critical t-value in this t-distribution is calculated with the following formula: $t_{n-2,\,1-\frac{\alpha}{2}} = t_{1829,\,0.975} = 1.9613$ The subsequent formula will be used to calculate the Test Statistic (T) for each of the regression coefficients:

$$T = \frac{(\hat{\beta}_i - \beta_i^{(0)})}{S_{\hat{\beta}_i}}$$

If the absolute value of our coefficient t-statistics greater than 1.9613 we can reject the null hypothesis that the coefficient is equal to zero. The null hypothesis, alternate hypothesis and test statistic for each regression coefficient are stated below.

Intercept:

$$Null = H_0 : \beta_0 = 0$$
$$Alternate = H_a : \beta_0 \neq 0$$
$$T = \frac{(13549.890 - 0)}{3421.962} = 3.96$$

This indicates that we reject the null hypothesis. However, the intercept holds no contextual value in this problem as we cannot have a zero square foot house.

TotalSF:

$$Null = H_0 : \beta_1 = 0$$
$$Alternate = H_a : \beta_1 \neq 0$$
$$T = \frac{112.150 - 0)}{2.173} = 51.62$$

This value indicates that we reject the null hypothesis. TotalSF provides significant information for predicting SalePrice at a type I error rate of 0.05.

Omnibus Overall F-test:

The formula for calculating the F-statistic for overall model is:

$$F = \frac{Mean\ Squared\ Regression}{Mean\ Squared\ Residual} = \frac{(\frac{(SSY-SSE)}{k})}{(\frac{SSE}{(n-k-1)})}$$

The following is the ANOVA table from Model 1. These values will be used to calculate the overall F-statistic.

```
Analysis of Variance Table

Response: model_df$SalePrice
                  Df     Sum Sq    Mean Sq F value    Pr(>F)
model_df$TotalSF    1 5.2109e+12 5.2109e+12  2664.3 < 2.2e-16 ***
Residuals        1827 3.5732e+12 1.9558e+09
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The hypotheses can be seen below:

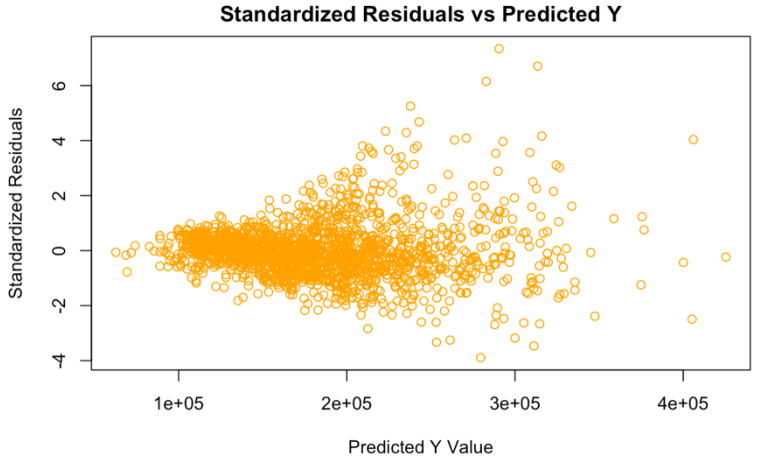
$Null = H_0 : \beta_1 = 0$
$Alternate = H_a : \beta_1 \neq 0$

The F-statistic was calculated in R and is equal to 2664. The critical F-value is calculated with using the formula below:

$$F_{k,\, n-k-1,\, 1-\alpha} = F_{1,\, 1829-1-1,\, 1-0.05} = 3.8466$$

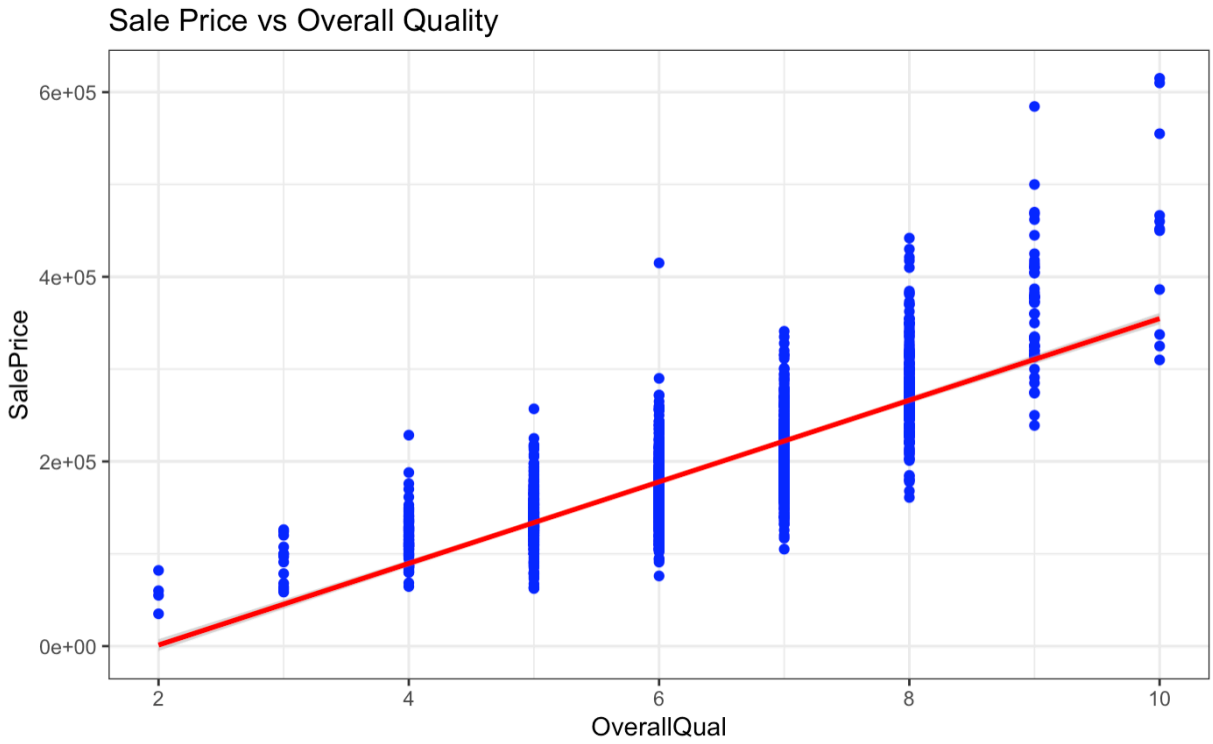
The F-statistic of 2664 is much greater than the critical F-value of 3.8466, allowing for the rejection of the null hypothesis. This indicates that there is a significant relationship between the independent variable and the response variable.

1e). The scatterplot below indicates that the error variance increases as Y-hat increases, which demonstrates a sign of heteroscedasticity. It is expected that the error variance to be constant as Y-hat increases. The histogram indicates a slight right skew, but generally follows a normal distribution. There appears to be 3 points at the top of the scatterplot that could be potential outliers or influential points as they have the highest residual values.



2. Overall Quality

2a).



2b). Below is the summary computed from Model 2. The regression coefficient in this equation can be interpreted as a 1 unit increase in OverallQual results in an increase of $44,196.6 to the SalePrice. OverallQual is an ordinal feature, therefore this is a difficult interpretation. There are only 10 values that the score can be and each score has a range of sale prices. An ordinal value would indicate that there were only 10 different sale prices that would ultimately be predicted, producing significant prediction error.

```
Call:
lm(formula = model_df$SalePrice ~ model_df$OverallQual)

Residuals:
    Min      1Q  Median      3Q     Max
-117133  -25043   -2633   19260  273974

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         -87243.3     4649.0  -18.77   <2e-16 ***
model_df$OverallQual  44196.6      747.3   59.15   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 40610 on 1827 degrees of freedom
Multiple R-squared:  0.6569,    Adjusted R-squared:  0.6567
F-statistic:  3498 on 1 and 1827 DF,  p-value: < 2.2e-16
```

$$\hat{Y} = -87243.3 + 44196.6X$$

2c). The R-squared value of Model 2 is 0.6569, meaning that approximately 66% percent of the variation in SalePrice is accounted for by OverallQual.

2d). The intercept coefficient will be ignored going forward as it carries no value in the context of this problem. The null hypothesis, alternate hypothesis and t-value can be seen below.

OverallQual:

$$Null = H_0 : \beta_1 = 0$$
$$Alternate = H_a : \beta_1 \neq 0$$
$$T = \frac{44196.6 - 0)}{747.3} = 59.15$$

The t-value indicates that we should reject the null hypothesis as OverallQual provides significant information in predicting SalePrice at a type I error rate of 0.05.

Omnibus Overall F-test :

The following is the ANOVA table from Model 2. The values will be used to calculate the overall F-statistic.

```
Analysis of Variance Table

Response: model_df$SalePrice
                       Df     Sum Sq    Mean Sq F value    Pr(>F)
model_df$OverallQual    1 5.7704e+12 5.7704e+12  3498.2 < 2.2e-16
Residuals            1827 3.0137e+12 1.6495e+09
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The hypotheses are the following:

$$Null = H_0 : \beta_1 = 0$$
$$Alternate = H_a : \beta_1 \neq 0$$

The F-statistic was generated using the following code to produce a F-statistic of 3498:

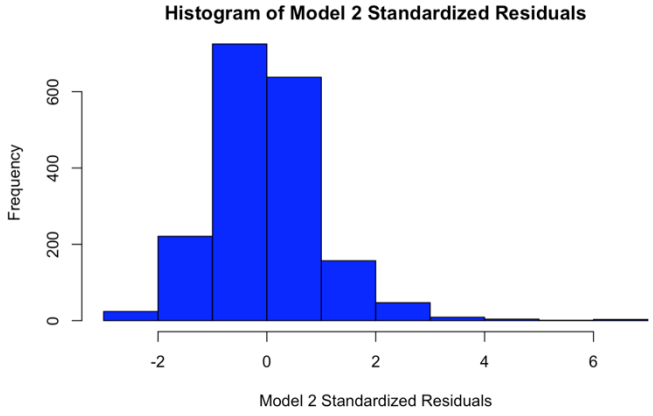
```r
# Write ANOVA results to dataframe
anova_2 <- anova(model_2)
# Extract SSY & SSE
SSY <- sum(anova_2$`Sum Sq`)
SSE <- anova_2$`Sum Sq`[2]
# Define k & n
k <- 1
n <- nrow(model_df)
# Calculate numerator and denominator
numerator_f <- (SSY - SSE) / k
denominator_f <-  SSE / (n - k - 1)
# Calculate F
F <- round(numerator_f / denominator_f)
```

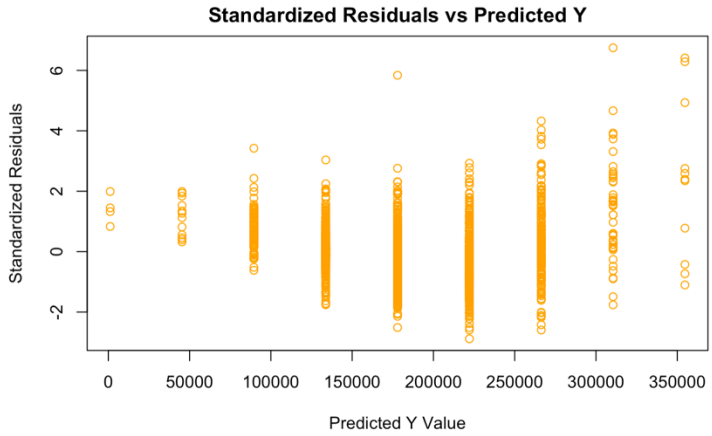
The critical F value is calculated with the following:

$$F_{k,\, n-k-1,\, 1-\alpha} = F_{1,\, 1829-1-1,\, 1-0.05} = 3.8466$$

The F-statistic of 3498 is much greater than the critical F-value of 3.8466, allowing us to reject the null hypothesis. There appears to be a significant relationship between the independent variable and the response variable.

2e). The scatterplot indicates that the error variance increases as Y-hat increases, which signals heteroscedasticity. It is expected that the error variance to be constant as Y-hat increases. The histogram indicates a slight right skew, but generally follows a normal distribution. There appears to be several points at the top of the scatterplot that could be potential outliers or influential points as they have the highest residual values.





3. Based on the R-squared and F-Statistic values, it would appear that Model 2 outperforms Model 1 from an accuracy standpoint. Model 2 accounts for a bigger proportion of variance in SalePrice than in Model 1. This result does not seem intuitive, considering OverallQual is an ordinal variable. This means that there will only ever be 10 different predictions for the SalePrice

of a home. OverallQual could prove to be more impactful when combined with another variable like TotalSF.

PART B: Multiple Linear Regression Models

4. 2 continuous explanatory variables.

4a). It was decided that OverallQual and TotalSF be the explanatory variables used for multiple linear regression.

```
Call:
lm(formula = model_df$SalePrice ~ model_df$TotalSF + model_df$OverallQual)

Residuals:
    Min      1Q  Median      3Q     Max
-115219  -20491     166   17508  257299

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          -91406.752   3818.013  -23.94   <2e-16 ***
model_df$TotalSF         62.729      2.107   29.77   <2e-16 ***
model_df$OverallQual  29416.781    789.012   37.28   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33330 on 1826 degrees of freedom
Multiple R-squared:  0.769,     Adjusted R-squared:  0.7688
F-statistic:  3040 on 2 and 1826 DF,  p-value: < 2.2e-16
```

$$\hat{Y} = -91406.752 + 62.729X_1 + 29416.781X_2$$

Again, the intercept is meaningless in the context of this problem. Beta 1 can be interpreted as a 1 unit increase in TotalSF and will result in an increase of $62.73 in SalePrice if all other variables are held constant. Beta 2 can be interpreted as a 1 unit increase in the OverallQual score will result in an increase of $29,416.78 in SalePrice if all other variables are held constant. Both of the coefficients have decreased in the combined model respective to their individual models.

4b). The multiple R-squared value for Model 3 is 0.769, demonstrating that this model accounts for more variance in the response variable than that of model 1. The difference between the two values is 0.1758, signifying that Model 3 accounts for an additional 17.58% of variance in SalePrice. Since the R-squared value increases with each additional independent variable, the adjusted R-squared value is a more appropriate comparison as it applies a penalty factor to unimportant independent variables being added to the model. The adjusted R-squared value is the same as the R-squared for Model 3, indicating that the increase of the proportion of variance accounted for is valid.

4c). Below, the null hypothesis, alternate hypothesis and t-value can be found for each explanatory variable.

TotalSF:

$$Null = H_0 : \beta_1 = 0$$
$$Alternate = H_a : \beta_1 \neq 0$$
$$T = \frac{(62.729 - 0)}{2.107} = 29.77$$

This value indicates that we reject the null hypothesis. TotalSF provides significant information for predicting SalePrice at a type I error rate of 0.05.

OverallQual:

$$Null = H_0 : \beta_2 = 0$$
$$Alternate = H_a : \beta_2 \neq 0$$
$$T = \frac{(29416.781 - 0)}{789.012} = 37.28$$

The t-value indicates that we should reject the null hypothesis as OverallQual provides significant information in predicting SalePrice at a type I error rate of 0.05.

Omnibus Overall F-test:

The formula for calculating the F-statistic for overall model is:

$$F = \frac{Mean\ Squared\ Regression}{Mean\ Squared\ Residual} = \frac{(\frac{(SSY-SSE)}{k})}{(\frac{SSE}{(n-k-1)})}$$

The following is the ANOVA table from Model 3. The values will be used to calculate the overall F-statistic.

```
Analysis of Variance Table

Response: model_df$SalePrice
                     Df     Sum Sq    Mean Sq F value    Pr(>F)
model_df$TotalSF      1 5.2109e+12 5.2109e+12  4689.9 < 2.2e-16 ***
model_df$OverallQual  1 1.5444e+12 1.5444e+12  1390.0 < 2.2e-16 ***
Residuals          1826 2.0288e+12 1.1111e+09
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

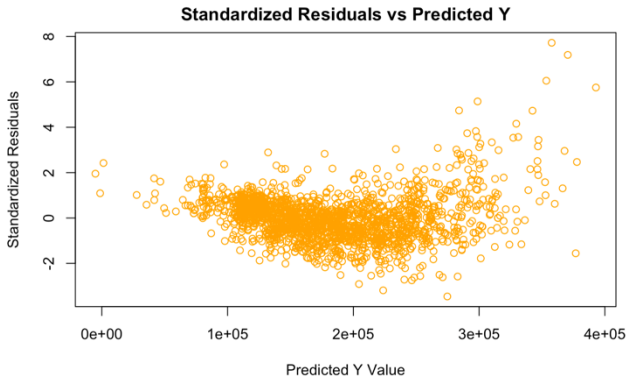
The hypotheses read as follows:

$Null = H_0 : \beta_1 = \beta_2 = 0$
$Alternate = H_a : \beta_1 \neq 0\ or\ \beta_2 \neq 0$

From the original fitting of the model, we can see that the F-statistic was calculated to be 3040, while the critical F-value was generated via the following formula: $F_{k,\,n-k-1,\,1-\alpha} = F_{2,\,1829-2-1,\,1-0.05} = 3.0007$ The F-statistic of 3040 is much greater than the critical F-value of 3.0007, which allows for the rejection of the null hypothesis. There appears to be a significant relationship between the independent variables and the response variable. However, this does not indicate which variables are meaningful.

4d). The scatterplot indicates that the error variance increases as Y-hat increases. This is a sign of heteroscedasticity. We expect the error variance to be constant as Y-hat increases. The histogram indicates a slight right skew, but generally follows a normal distribution and there appears to be

several points at the top of the scatterplot that could be potential outliers or influential points as they have the highest residual values.



4e). Both variables in the model, as when they are combined, they account for a much larger proportion of variance in the response variable than a single feature model.

5. Use of 3 explanatory variables

5a). It was decided that OverallQual, TotalSF and TotalBsmtSF be the explanatory variables used for multiple linear regression in the prediction of SalePrice.

```
Call:
lm(formula = model_df$SalePrice ~ model_df$TotalSF + model_df$OverallQual +
    model_df$TotalBsmtSF)

Residuals:
    Min      1Q  Median      3Q     Max
-113376  -17088    -166   15954  208034

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -1.015e+05  3.293e+03  -30.81   <2e-16 ***
model_df$TotalSF     5.731e+01  1.817e+00   31.55   <2e-16 ***
model_df$OverallQual 2.300e+04  7.201e+02   31.94   <2e-16 ***
model_df$TotalBsmtSF 5.398e+01  2.094e+00   25.78   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 28550 on 1825 degrees of freedom
Multiple R-squared:  0.8307,    Adjusted R-squared:  0.8304
F-statistic:  2984 on 3 and 1825 DF,  p-value: < 2.2e-16
```

$$\hat{Y} = -101455.5643 + 57.3082X_1 + 23001.7556X_2 + 53.9852X_3$$

Beta 1 can be interpreted as a 1 unit increase in TotalSF will result in an increase of $57.31 in SalePrice if all other variables are held constant. Beta 2 can be interpreted as a 1 unit increase in the OverallQual score will result in an increase of $23,001.76 in SalePrice if all other variables are held constant. Beta 3 can be interpreted as a 1 unit increase in the TotalBmstSF will result in an increase of $53.99 in SalePrice if all other variables are held constant. Both of the coefficients from Model 3 have decreased in Model 4 with the addition of TotalBsmtSF.

5b). The multiple R-squared value for Model 4 is 0.8307 as it accounts for more variance in the response variable than Model 3. The difference between the two values is 0.0617, signifying that Model 4 is accounting for an additional 6.2% of variance in SalePrice. Since the R-squared value increases with each additional independent variable, the adjusted R-squared value is a more appropriate comparison as it applies a penalty factor to unimportant independent variables being added to the model. The adjusted R-squared value is the same as the R-squared for Model 3, indicating that the increase of the proportion of variance accounted for is valid.

5c). Below, the null hypothesis, alternate hypothesis and t-value can be found for each explanatory variable.

TotalSF:

$$Null = H_0 : \beta_1 = 0$$
$$Alternate = H_a : \beta_1 \neq 0$$
$$T = \frac{(57.3082 - 0)}{1.817} = 31.55$$

This value indicates that we reject the null hypothesis. TotalSF provides significant information for predicting SalePrice at a type I error rate of 0.05.

OverallQual:

$Null = H_0 : \beta_2 = 0$

$Alternate = H_a : \beta_2 \neq 0$

$$T = \frac{(23001.7556 - 0)}{720.1} = 31.94$$

The t-value indicates that we should reject the null hypothesis as OverallQual provides significant information in predicting SalePrice at a type I error rate of 0.05.

TotalBsmtSF:

$Null = H_0 : \beta_3 = 0$

$Alternate = H_a : \beta_3 \neq 0$

$$T = \frac{(53.9852 - 0)}{2.094} = 25.78$$

The t-value indicates that we should reject the null hypothesis as TotalBsmtSF provides significant information in predicting SalePrice at a type I error rate of 0.05.

Omnibus Overall F-test:

The formula for calculating the F-statistic for overall model is:

$$F = \frac{Mean\ Squared\ Regression}{Mean\ Squared\ Residual} = \frac{(\frac{(SSY - SSE)}{k})}{(\frac{SSE}{(n-k-1)})}$$

The following is the ANOVA table from Model 4. The values will be used to calculate the overall F-statistic.

```
Analysis of Variance Table

Response: model_df$SalePrice
                      Df     Sum Sq    Mean Sq F value    Pr(>F)
model_df$TotalSF       1 5.2109e+12 5.2109e+12 6393.94 < 2.2e-16 ***
model_df$OverallQual   1 1.5444e+12 1.5444e+12 1895.07 < 2.2e-16 ***
model_df$TotalBsmtSF   1 5.4150e+11 5.4150e+11  664.44 < 2.2e-16 ***
Residuals           1825 1.4873e+12 8.1497e+08
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

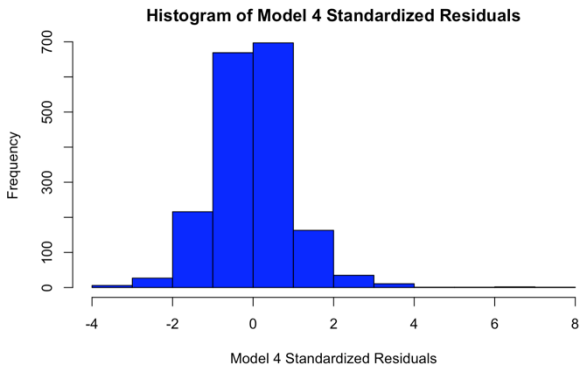
The hypotheses are as follows:

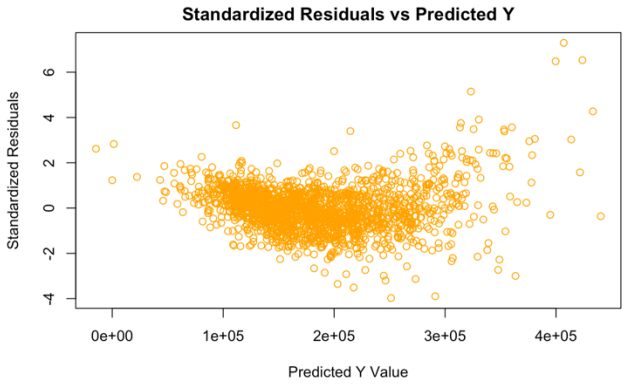
$Null = H_0 : \beta_1 = \beta_2 = \beta_3 = 0$

$Alternate = H_a : \beta_1 \neq 0\ or\ \beta_2 \neq 0\ or\ \beta_3 \neq 0$

From the original fitting of the model, we can see that the F-statistic was calculated to be 2984, while the critical F-value was generated via the following formula: $F_{k,\,n-k-1,\,1-\alpha} = F_{3,\,1829-3-1,\,1-0.05} = 2.6098$ The F-statistic of 2984 is much greater than the critical value of 2.6098, which allows us to reject the null hypothesis. This indicates that there is a significant relationship between the independent variables and the response variable, though this does not indicate which variables are meaningful.

5d). The scatterplot indicates that the error variance increases as Y-hat increases, which is indicative of heteroscedasticity. We expect the error variance to be constant as Y-hat increases, while the histogram displays a slight right skew and generally follows a normal distribution. There also appears to be several points at the top of the scatterplot that could be potential outliers or influential points as they have the highest residual values.





5e). I would retain all three variables in the model as they account for a much larger proportion of variance in the response variable than Model 3. The adjusted R-squared increased at the same magnitude as the R-squared value, which is representative that the model was not penalized for an unimportant independent variable.

PART C: Multiple Linear Regression Models on Transformed Response Variable

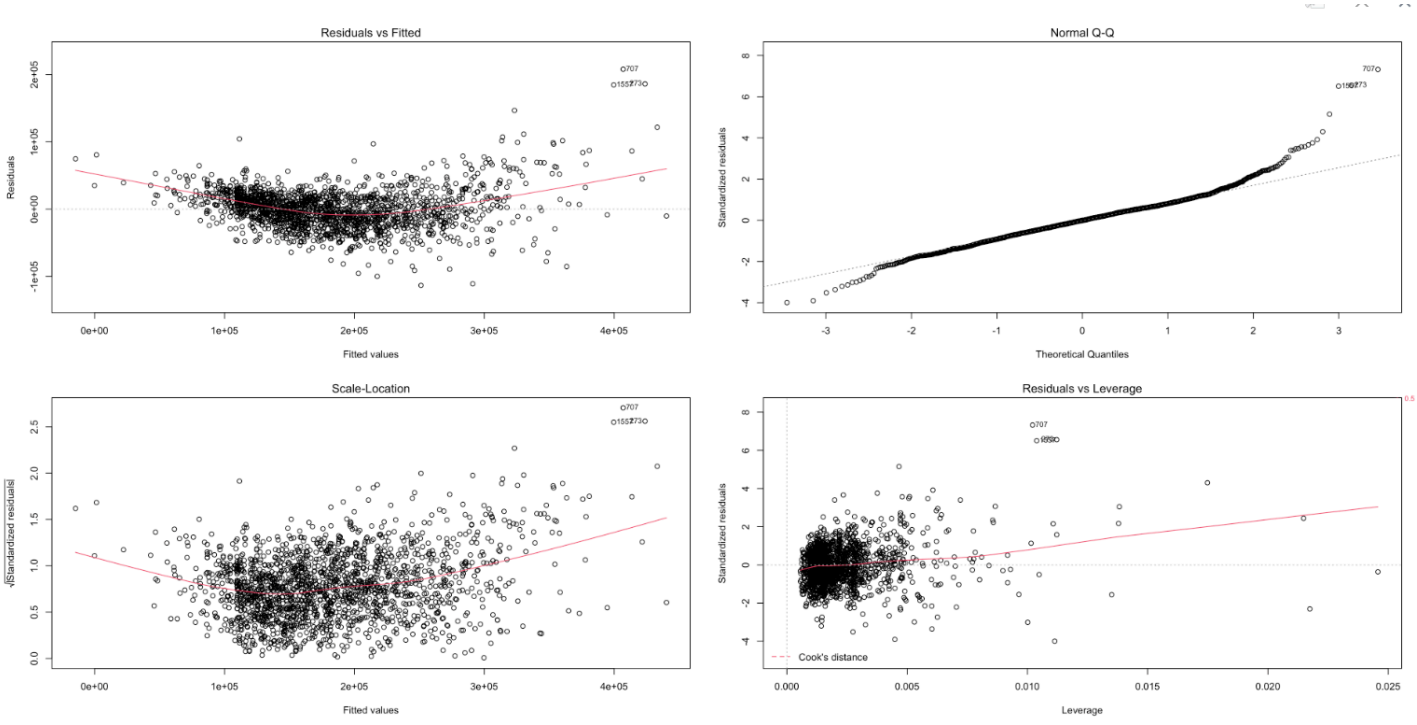
6. Adjusted R-squared values were observed, alongside the RMSE for each model as it pertains to fitting models with logSalePrice. The adjusted R-squared value is a good measure of the proportion of variance accounted for in the response variable as it utilizes a penalty factor so that unimportant independent variables don't increase the value. The measure RMSE is the square root of the mean squared error. Model 4 fits the data the best utilizing both of these measures as can be seen in the table below. Model 4 possesses the highest adjusted R-squared while displaying the most minimal RMSE.

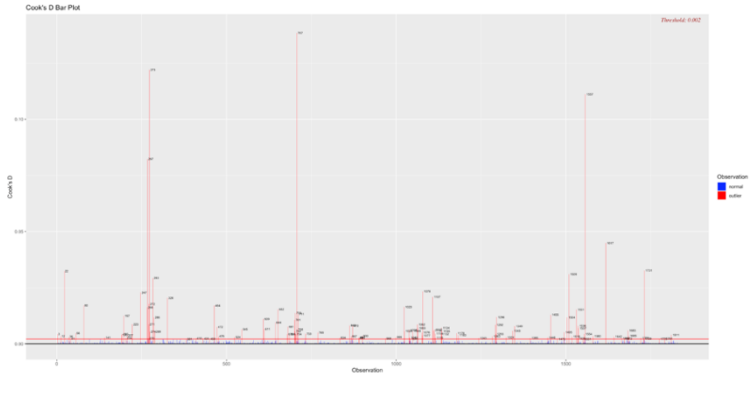
Model	Adjusted R-Squared	RMSE
1	0.6101282	0.2165033
3	0.7934511	0.1575419
4	0.8346686	0.1409105

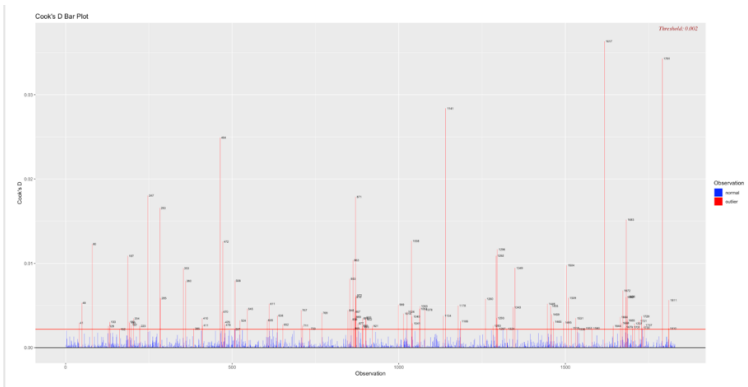
7. It becomes more difficult to explain the model coefficients once variables are transformed. In this particular case, there is only one feature transformation occurring. Hence, it seems worthwhile to sacrifice how understandable the model is versus increased model performance. The predicted logSalePrice can be transformed back to SalePrice prior to present findings to an executive team for the sake of simplicity, as it is more easily understandable in explaining technical concepts to non-technical leaders. If the presentation of the model results were contextualized with the problem statement, then there would be a much better chance of getting a point across, in regard to the data, even if the technical concepts are not fully understood.

PART D: Multiple Linear Regression Models and Influential Points

8. As can be seen in the scatter plots below with regard to Cooks Distance, there appear to be 0 points that surpass the threshold for Cooks Distance, while there are many outliers (146) that surpass the leverage threshold. Following the removal of certain points with the highest leverage, the model was then refitted, and it generated a slightly higher R-squared, jumping from .8346 to .8347. However, it should be noted that since SalePrice is the response variable for this model, the model should not be trimmed of its outliers by too much as it will be susceptible to overfitting and not respond well to unseen data.







PART E: Beginning to Think About a Final Model

9a). For the final mode, additional continuous variables from the dataset will be added. Until this point in the modeling process, the variables have all been related to the interior of the home or the overall quality score. I'm going to include some external measurements (i.e. GarageArea, LotArea, LotFrontage) to see how they impact the model, as home buyers are concerned with the property size and exposure to the street.

9b). GarageArea has a coefficient of 52.12 meaning that a 1 unit increase in GarageArea will result in an increase of 52.12 in SalePrice if all other variables held constant. LotArea has a coefficient of 1.667 which equates to a $1.67 increase in SalePrice for a one unit increase in LotArea. LotFrontage has a coefficient of 160.1015 which equates to a 160.11 increase in SalePrice for a one unit increase in LotFrontage.

```
Call:
lm(formula = model_df$SalePrice ~ model_df$TotalSF + model_df$OverallQual +
    model_df$TotalBsmtSF + model_df$GarageArea + model_df$LotArea +
    model_df$LotFrontage)

Residuals:
    Min      1Q  Median      3Q     Max
-114089  -15271    -166   14701  216389

Coefficients:
                       Estimate  Std. Error t value             Pr(>|t|)
(Intercept)          -116919.824    3827.598 -30.547 < 0.0000000000000002
model_df$TotalSF          50.106       1.772  28.270 < 0.0000000000000002
model_df$OverallQual   20536.711     709.228  28.956 < 0.0000000000000002
model_df$TotalBsmtSF      42.247       2.111  20.010 < 0.0000000000000002
model_df$GarageArea       52.119       4.509  11.559 < 0.0000000000000002
model_df$LotArea           1.667       0.243   6.863     0.00000000000918
model_df$LotFrontage     160.101      48.381   3.309             0.000954

Residual standard error: 26760 on 1822 degrees of freedom
Multiple R-squared:  0.8514,    Adjusted R-squared:  0.8509
F-statistic:  1740 on 6 and 1822 DF,  p-value: < 0.00000000000000022
```

9c). The formula for calculating the F-statistic for overall model is (F-statistic = 1740):

$$F = \frac{Mean\ Squared\ Regression}{Mean\ Squared\ Residual} = \frac{(\frac{(SSY-SSE)}{k})}{(\frac{SSE}{(n-k-1)})}$$

The following is the ANOVA table from Model 5. The values will be used to calculate the overall F-statistic.

```
Analysis of Variance Table

Response: model_df$SalePrice
                       Df        Sum Sq       Mean Sq  F value                Pr(>F)
model_df$TotalSF        1 5210864359940 5210864359940 7274.579 < 0.00000000000000022
model_df$OverallQual    1 1544419797500 1544419797500 2156.073 < 0.00000000000000022
model_df$TotalBsmtSF    1  541500622924  541500622924  755.957 < 0.00000000000000022
model_df$GarageArea     1  127312809406  127312809406  177.734 < 0.00000000000000022
model_df$LotArea        1   47042592444   47042592444   65.673 0.0000000000000009654
model_df$LotFrontage    1    7844035756    7844035756   10.951               0.000954
Residuals            1822 1305119525486     716311485
```

The null and alternate hypotheses are seen below with the critical F-value:

$$Null = H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$$

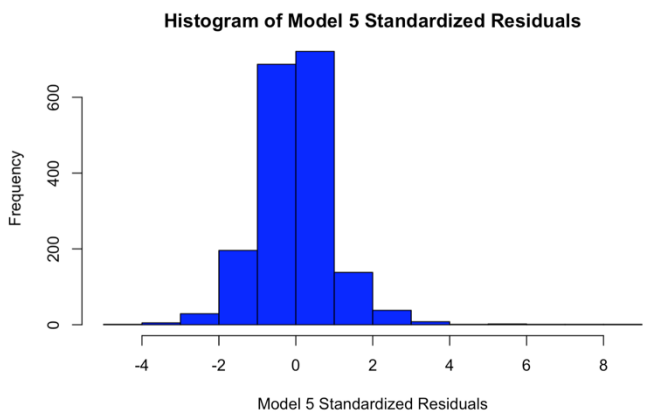
$$Alternate = H_a : \beta_i \neq 0\ for\ at\ least\ one\ i$$

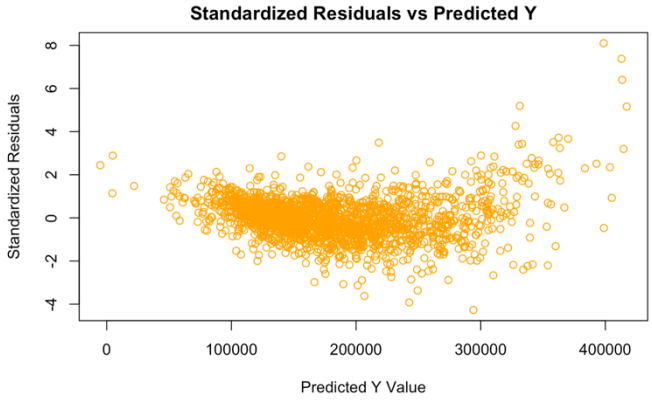
$$F_{k,\, n-k-1,\, 1-\alpha} = F_{6,\, 1829-6-1,\, 1-0.05} = 2.1036$$

The F-statistic of 1740 is much greater than the critical value of 2.1036, allowing us to reject the null hypothesis. This indicates that there is a significant relationship between the independent variables and the response variable. However, this does not indicate which variables are meaningful.

9d). Below demonstrates the normality of the residuals. Over the 5 models we see an increase in R-squared values increases from model 1-5 (0.593, 0.656, 0.768, 0.830, 0.850). Although the models are far from perfect, we are already seeing an increase in accuracy of each model with

added features. Each p-value in Model 5 has a p-value less that 0.05 suggesting statistical significance.





9e). The scatterplot above indicates that the error variance increases as Y-hat increases, which is a sign of heteroscedasticity. There also appears to be a downward trend initially in the residuals.

CONCLUSION

Variable transformation and outlier detection/removal can seemingly have a dramatic effect on the modeling process. If observations are obtained in a rigorous fashion and are determined to be valid data points, then it becomes slightly bias to remove them. The difficulty really comes from future observations that may contain the same attribute values as the influential or outlier points. This will lead to a poor prediction on these observations and can ultimately be beneficial to remove these values as their influence can dramatically change the prediction values for the vast majority of "normal" observations. Statistical tests can still be trusted to a certain extent, but common sense needs to be employed when interpreting them. The intercept, for example, is largely meaningless in the context of this problem and analyzing the test of significance won't change that fact. As it pertains to next steps, I would include examining interactions between

variables and checking for problems with collinearity. I noticed in the correlation matrix that some of the independent variables have strong linear correlations. These are features that would need to be explored and/or accounted for to a further extent.