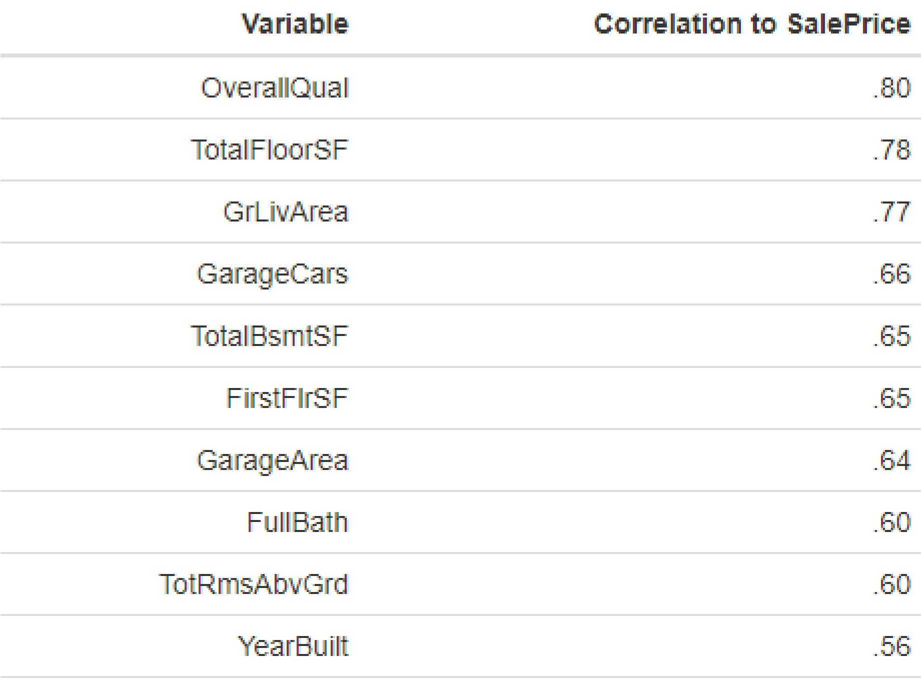
NORTHWESTERN UNIVERSITY | SCHOOL OF PROFESSIONAL STUDIES

Modeling Assignment #2: Building Linear Regression Models
MSDS 410

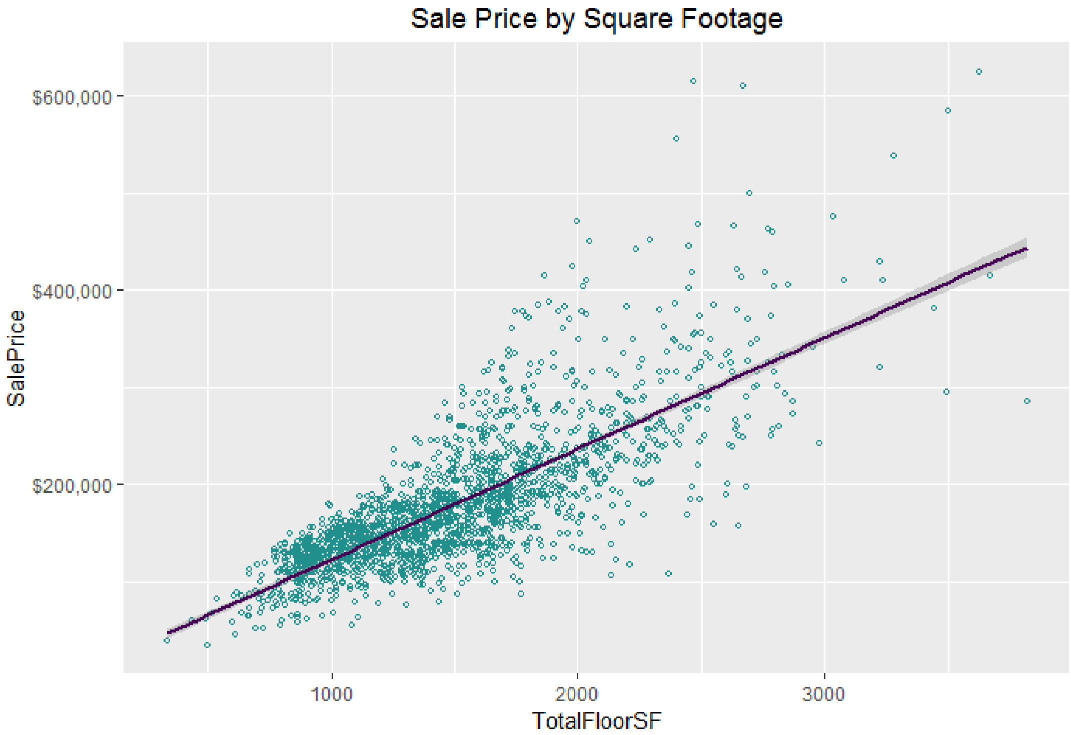
PART A: Simple Linear Regression Models

(1) I have chosen TotalFloorSF as my explanatory variable to predict SalePrice. The following table shows the 10 non-factor variables with the highest Pearson correlation to SalePrice. TotalFloorSF has the highest correlation after OverallQual (which I have ignored due to it being used in section 2 below.)



Variable	Correlation to SalePrice
OverallQual	.80
TotalFloorSF	.78
GrLivArea	.77
GarageCars	.66
TotalBsmtSF	.65
FirstFlrSF	.65
GarageArea	.64
FullBath	.60
TotRmsAbvGrd	.60
YearBuilt	.56

a.



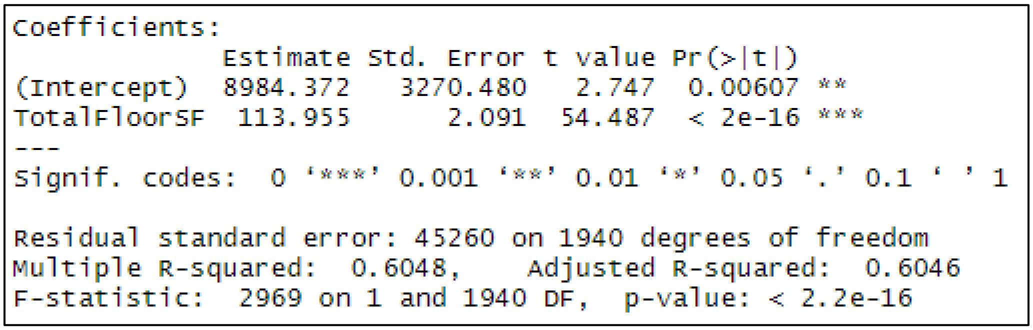
b. Y = 113.955*TotalFloorSF + 8984.372

- 8984.372: Our intercept is the value of a home that has 0 total square feet. This is outside of the realistic parameters of our model and could not be used to estimate an empty plot of land. It simply serves as the y-intercept of the line of best fit.
- 113.955*TotalFloorSF: For every square foot increase in size of the home, the predicted price increases by $113.96.

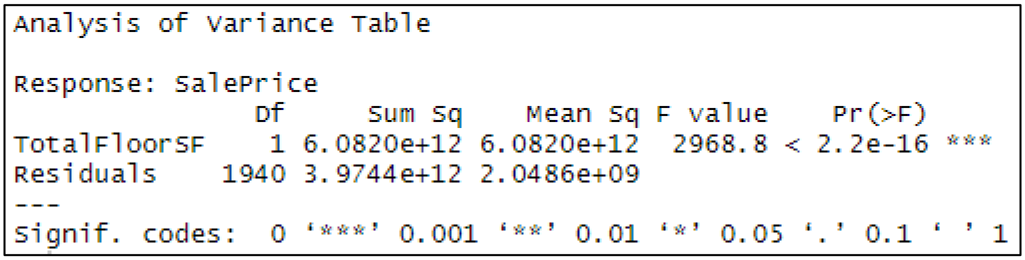
c. R-Squared: 0.605

- This model accounts for 60.5% of the variance in SalePrice.

d.



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  8984.372   3270.480   2.747  0.00607 ** 
TotalFloorSF  113.955      2.091  54.487  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 45260 on 1940 degrees of freedom
Multiple R-squared:  0.6048,	Adjusted R-squared:  0.6046 
F-statistic:  2969 on 1 and 1940 DF,  p-value: < 2.2e-16
```

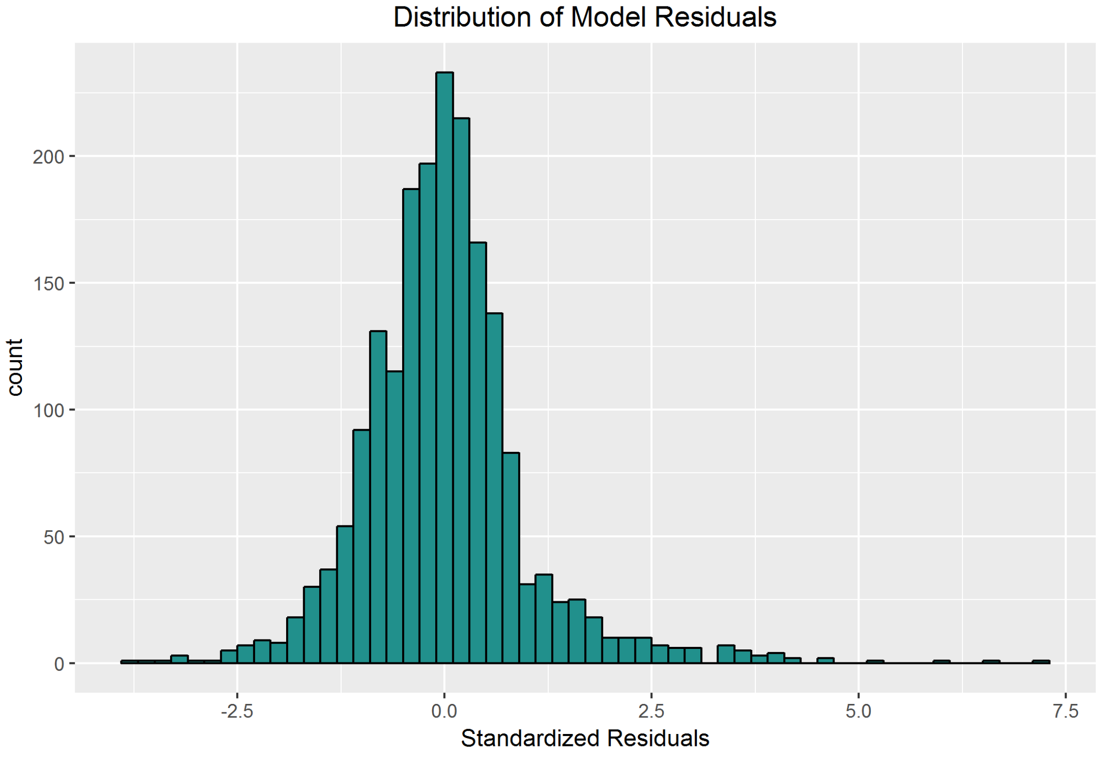


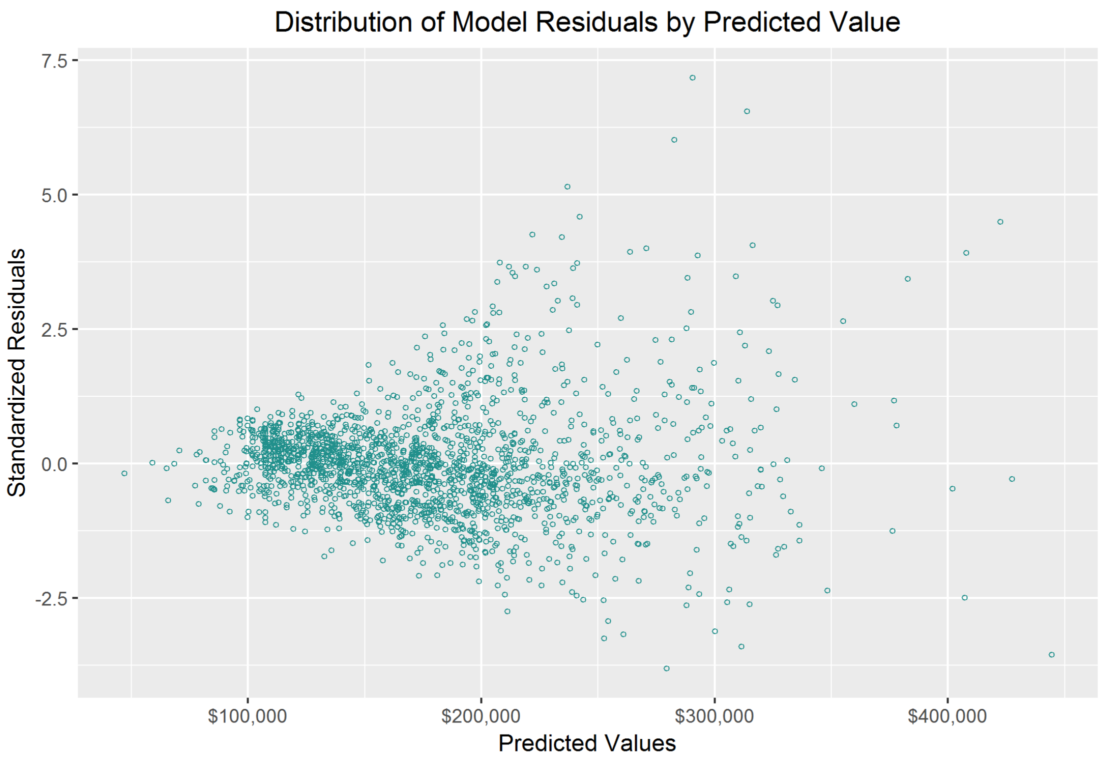
```
Analysis of Variance Table

Response: SalePrice
               Df     Sum Sq    Mean Sq F value    Pr(>F)    
TotalFloorSF    1 6.0820e+12 6.0820e+12  2968.8 < 2.2e-16 ***
Residuals    1940 3.9744e+12 2.0486e+09                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Intercept Hypothesis Test:
  - H_0: $\beta_0 = 0$
  - H_A: β_0 does not equal 0

    - Reject the null hypothesis at the 0.001 level
- TotalFloorSF Hypothesis Test:
    - H_0: $\beta_1 = 0$
    - H_A: β_1 does not equal 0
    - Reject the null hypothesis at the 0 level
- Omnibus Hypothesis Test:
    - H_0: All betas included in model (in our case only one) = 0
    - H_A: At least one beta does not equal 0
    - We reject the null hypothesis and conclude that our model significantly explains the variance of SalePrice.

e. The validity of the hypothesis tests are dependent on the underlying assumptions of Independence, Normality, and Homoscedasticity being well met. To assess this, use the model from part a) to calculate predicted values for each record. Then use the predicted values to compute residuals. Yes, many of the packages automatically give you the predicted and residuals, but you should know how to code and compute these values. Next standardize the residuals but subtracting off the mean and dividing by the standard deviation for each residual (i.e. you will have to obtain those summary statistics first). Check on the underlying assumptions by plotting:

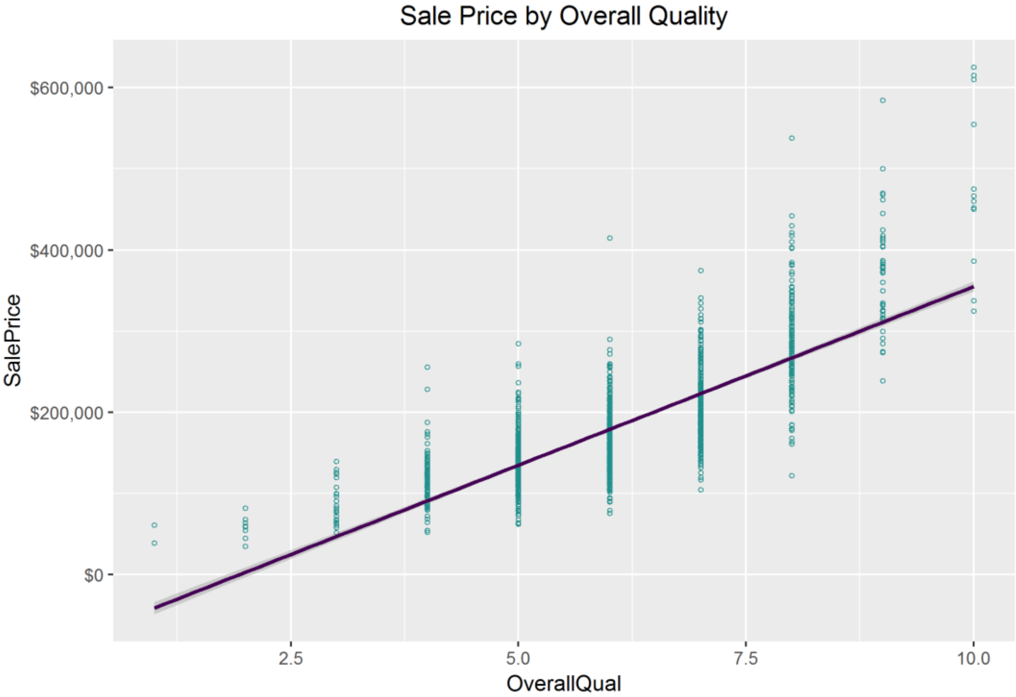




In the scatter plot graph, there is a clear fan shape formed by the residuals, which suggests a violation of the assumption of homoscedasticity. There appear to be some extreme outliers, which can be observed in the far right tail of the histogram. While these outliers could strongly influence our model, even without those points there is still a general fan shape of the standardized residuals as the predicted sale price increases. The issue that this causes is that as our predicted price increases, the observed and expected error of our model increases. The higher predictions are less accurate than the lower predictions.

(2) Let Y = sale price be the dependent or response variable. Use the OVERALL QUALITY variable as the explanatory variable (X) to predict Y. Fit a simple linear regression model using X to predict Y. Call this Model 2. You should:

a.



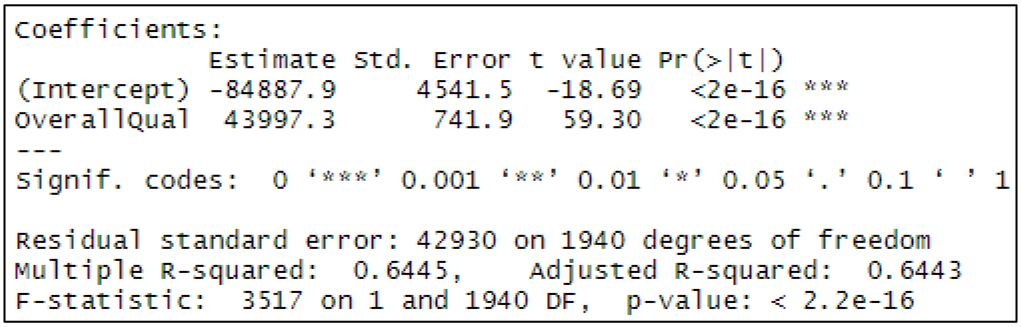
b. Y = 43997.254*OverallQual - 84887.949

- -84887.949: Our intercept is the value of a home that was rated a 0 in Overall Quality. This is outside of the realistic parameters as the lowest quality rating observed is 1. It simply serves as the y-intercept of the line of best fit.
- 43997.254*OverallQual: For every one point increase in Overall Quality, the predicted price increases by $43,997.25.
- OverallQual is an ordinal variable, unlike TotalSqFt form model 1, which was totally continuous. Each level of Overall Quality has sale prices at many different values. For prediction purposes this poses a challenge because this model will assume that for a given Quality score each home will have the exact same sale price.

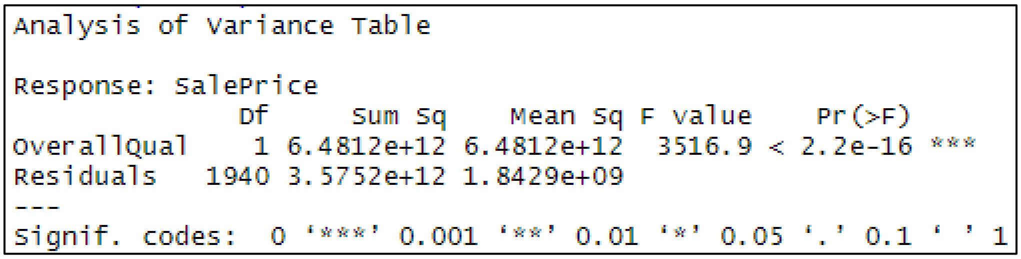
c. R-Squared: 0.644

- This model accounts for 64.4% of the variance in SalePrice.

d.



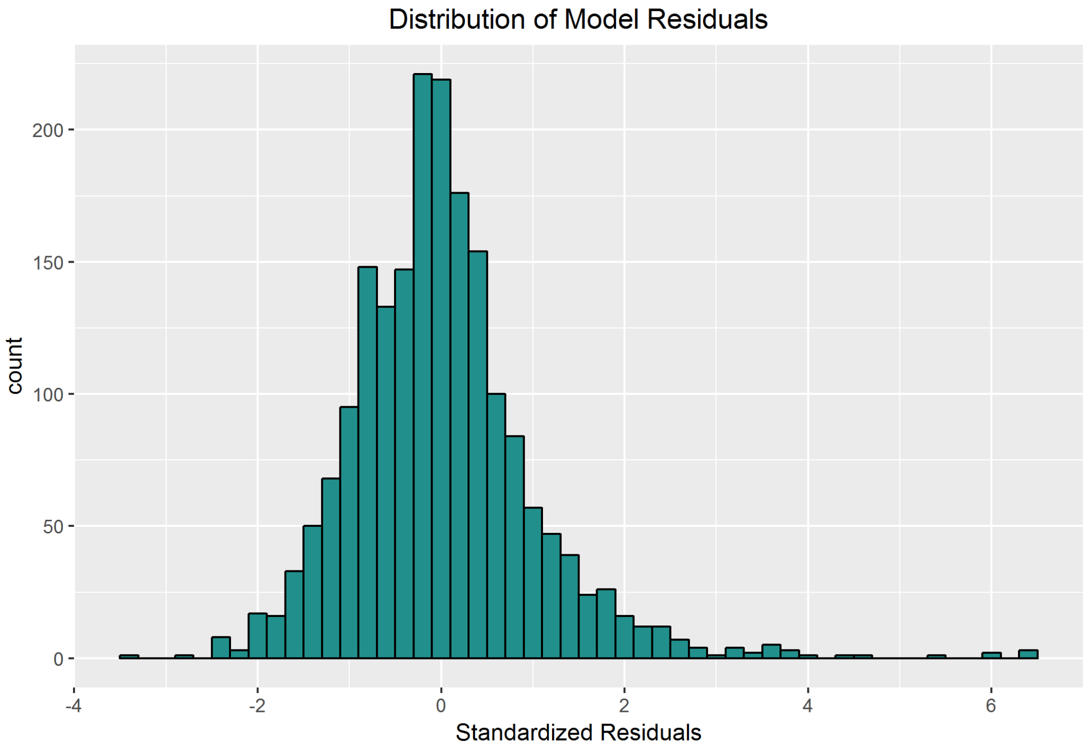
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -84887.9     4541.5  -18.69   <2e-16 ***
OverallQual  43997.3      741.9   59.30   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 42930 on 1940 degrees of freedom
Multiple R-squared:  0.6445,	Adjusted R-squared:  0.6443 
F-statistic:  3517 on 1 and 1940 DF,  p-value: < 2.2e-16
```

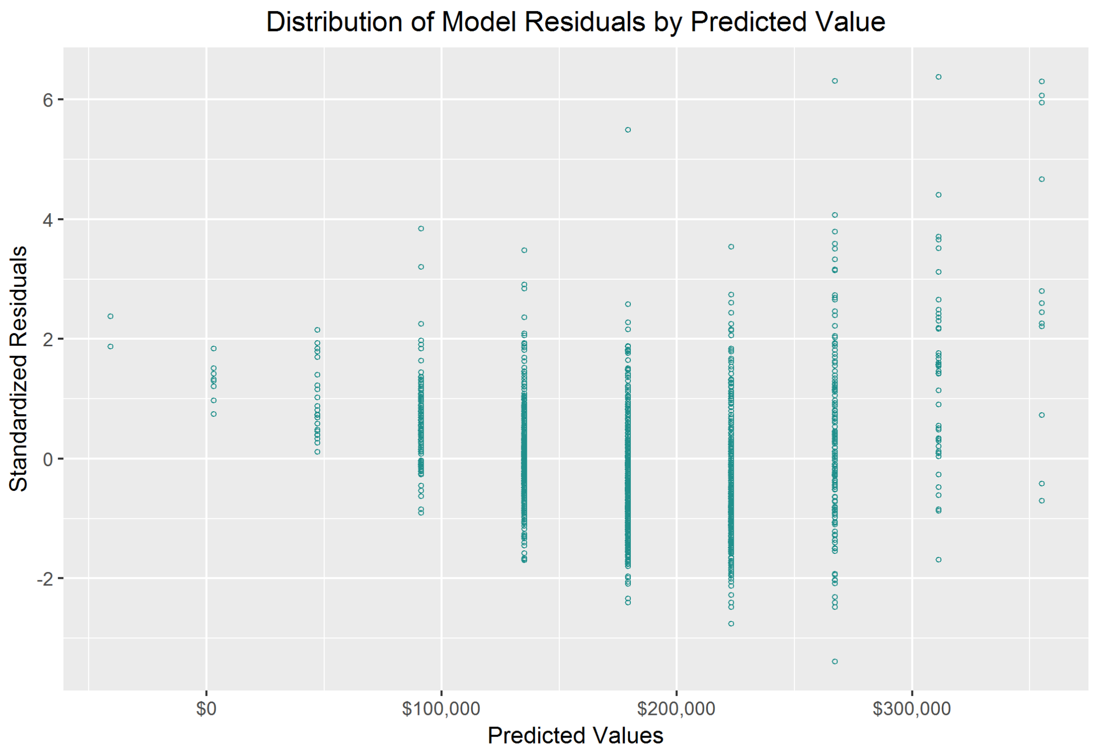


```
Analysis of Variance Table

Response: SalePrice
              Df     Sum Sq    Mean Sq F value    Pr(>F)    
OverallQual    1 6.4812e+12 6.4812e+12  3516.9 < 2.2e-16 ***
Residuals   1940 3.5752e+12 1.8429e+09                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Intercept Hypothesis Test:
  - H_0: $\beta_0 = 0$
  - H_A: β_0 does not equal 0
  - Reject the null hypothesis at the 0 level
- OverallQual Hypothesis Test:
  - H_0: $\beta_1 = 0$
  - H_A: β_1 does not equal 0
  - Reject the null hypothesis at the 0 level
- Omnibus Hypothesis Test:
  - H_0: All betas included in model (in our case only one) = 0
  - H_A: At least one beta does not equal 0
  - We reject the null hypothesis and conclude that our model significantly explains the variance of SalePrice.

e.





The residuals in the scatter plot seem to show a concave fanning pattern over the predicted sales price variable, where they trend more negative as the price increases until trending back positive over the last three quality levels. This shows a heteroscedastic nature for the residuals of this model as well. This histogram of the standardized residuals is right-tailed, which violates the normality of residuals assumption. There appear to be approximately five outlier results with highly positive standardized residuals and one with very negative standardized residual. These points could strongly affect the model.

(3) Model 2 shows a better fit. It has an r-squared value about 4% greater than model 1. The t-statistic on the explanatory variable in model 2 is greater than that of model 1, which suggests a greater statistical significance for that explanatory variable.

PART B: Multiple Linear Regression Models

(4) Fit a multiple regression model that uses 2 continuous explanatory (X) variables to predict Sale Price (Y). These two explanatory(X) variables should be: the explanatory variables from Model 1 and Model 2 above. Call this Model 3. You should:

a. Y = 65.951*TotalFloorSF + 28407.21*OverallQual - 89590.824

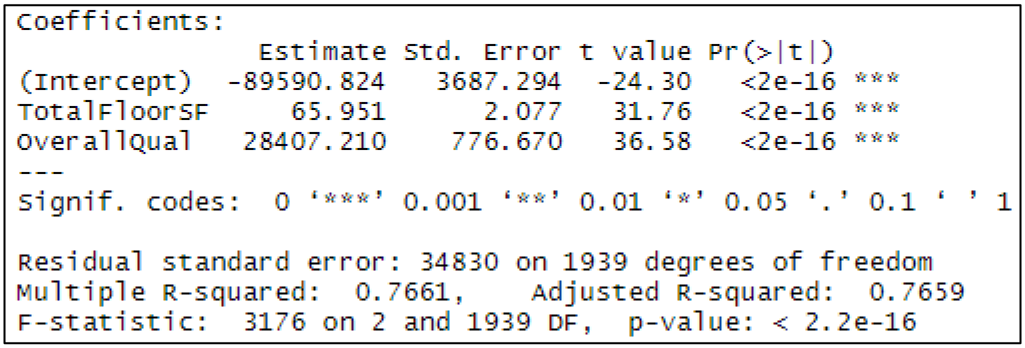
- -89590.824: Our intercept is the value of a home that was rated a 0 in Overall Quality and has zero total square footage. This is outside of the realistic parameters as the lowest quality rating observed is 1 and we do not have any empty plots of land in our sample. It simply serves as the y-intercept of the line of best fit.
- 28407.21*OverallQual: For every one point increase in Overall Quality while holding square footage constant, the predicted price increases by $28,407.21.
- 65.951*TotalFloorSF: For every increase of one square foot while holding the overall quality of the home constant, the predicted sale price increases by $65.95.

- In this model both β_1 and β_2 are lower than they were in their respective simple linear models.

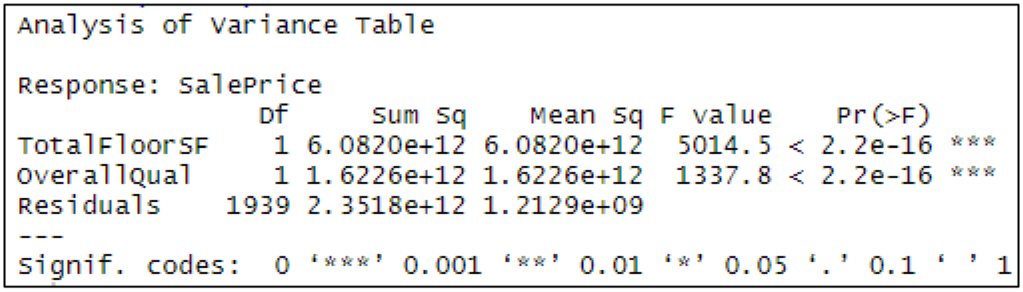
b. R-Squared: 0.776

- This model accounts for 76.6% of the variance in SalePrice.
- This model fits better than either of the simple linear regressions. There is a 0.161 increase in r-squared from model 1 to model 3. That means that our multiple linear regression explains an extra 16% of sale price variance in comparison to the simple linear regression using only total square footage.

c.



```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -89590.824   3687.294  -24.30   <2e-16 ***
TotalFloorSF     65.951      2.077   31.76   <2e-16 ***
OverallQual   28407.210    776.670   36.58   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 34830 on 1939 degrees of freedom
Multiple R-squared:  0.7661,    Adjusted R-squared:  0.7659 
F-statistic:  3176 on 2 and 1939 DF,  p-value: < 2.2e-16
```

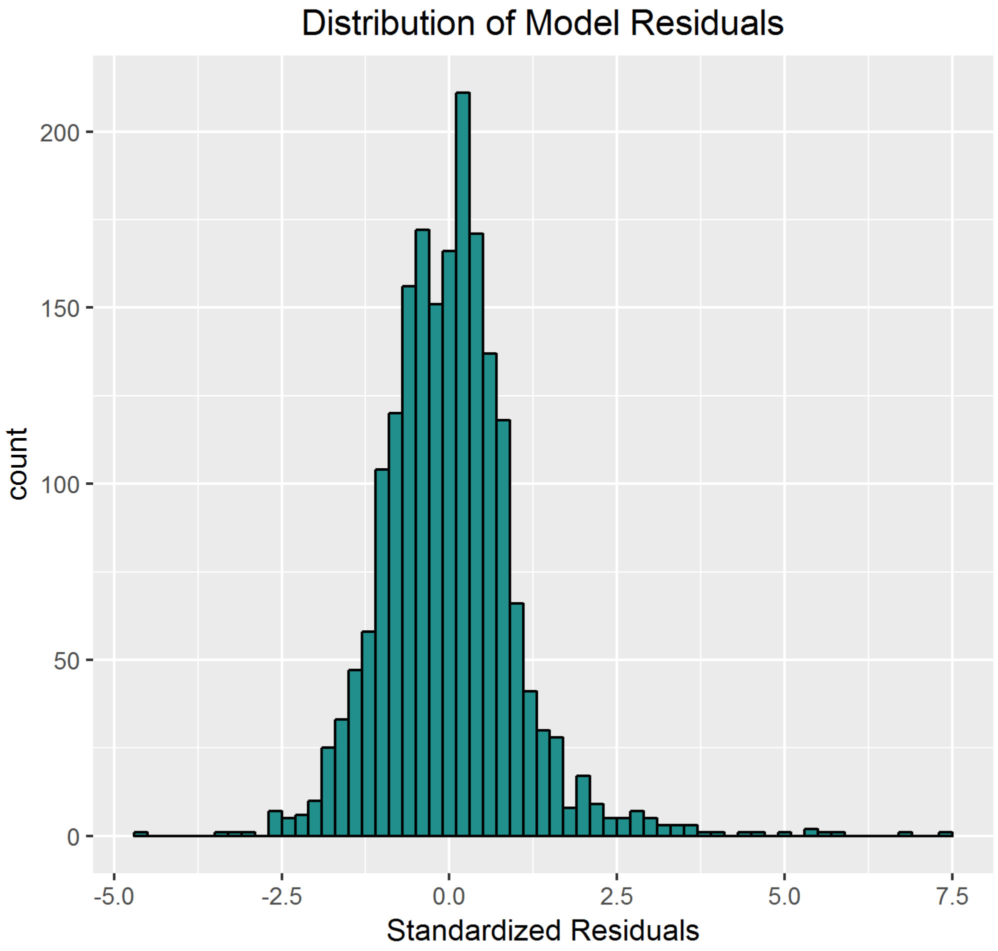


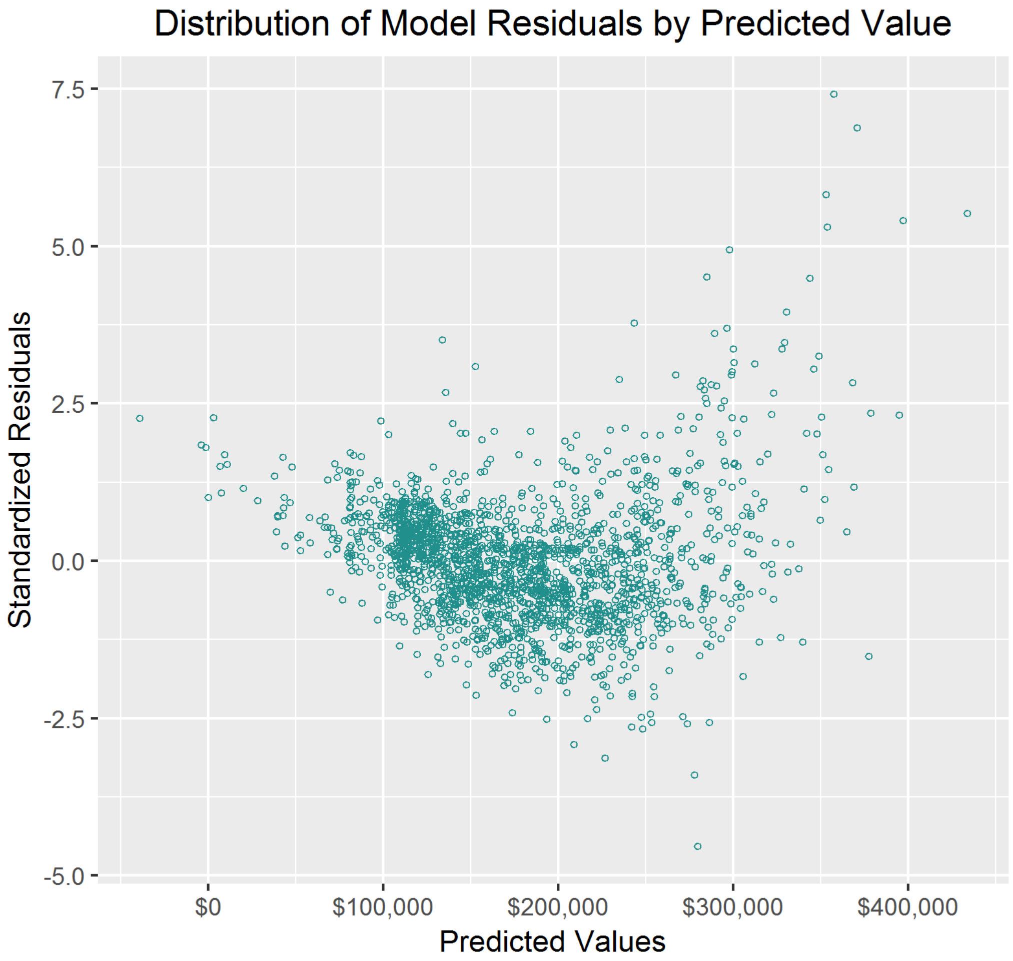
```
Analysis of Variance Table

Response: SalePrice
               Df     Sum Sq    Mean Sq F value    Pr(>F)    
TotalFloorSF    1 6.0820e+12 6.0820e+12  5014.5 < 2.2e-16 ***
OverallQual     1 1.6226e+12 1.6226e+12  1337.8 < 2.2e-16 ***
Residuals    1939 2.3518e+12 1.2129e+09                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

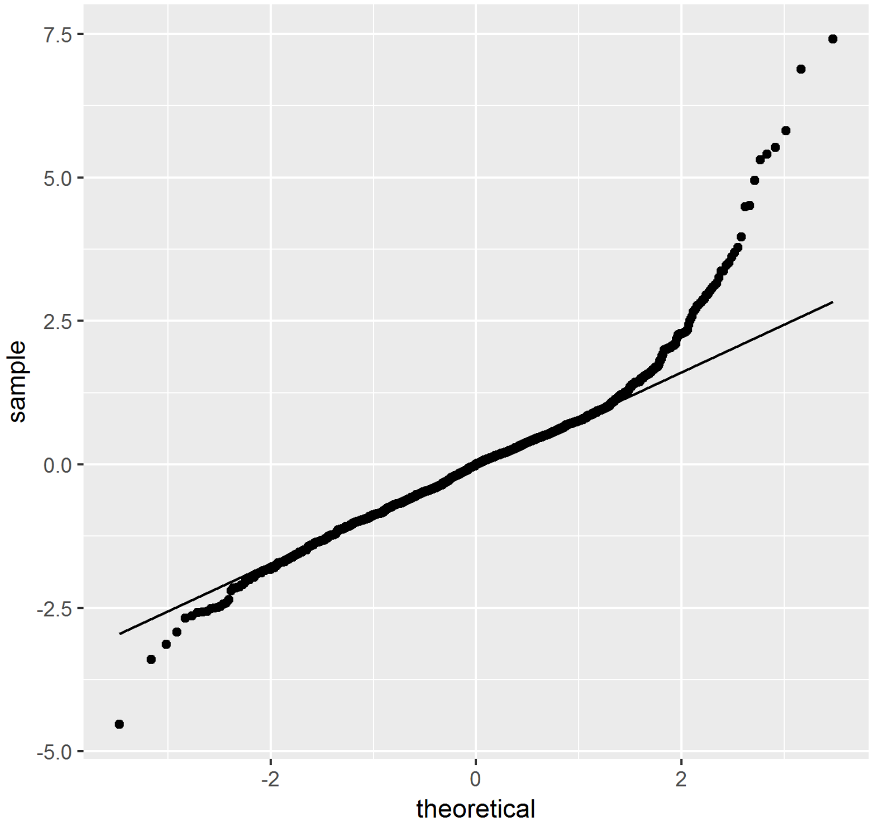
Specify the hypotheses associated with each coefficient of the model and the hypothesis for the omnibus model. Conduct and interpret the hypothesis tests. Intercept Hypothesis Test:

- Intercept Hypothesis Test:
    - H_0: $\beta_0 = 0$
    - H_A: β_0 does not equal 0
    - Reject the null hypothesis at the 0 level
- TotalFloorSF Hypothesis Test:
    - H_0: $\beta_1 = 0$
    - H_A: β_1 does not equal 0
    - Reject the null hypothesis at the 0 level
- OverallQual Hypothesis Test:
    - H_0: $\beta_2 = 0$
    - H_A: β_2 does not equal 0
    - Reject the null hypothesis at the 0 level
- Omnibus Hypothesis Test:
    - H_0: $\beta_1 = \beta_2 = 0$
    - H_A: At least one beta does not equal 0
    - We reject the null hypothesis and conclude that at least one β does not equal zero.

d.


Distribution of Model Residuals
count
200
150
100
50
0
-5.0
-2.5
0.0
2.5
5.0
7.5
Standardized Residuals


Distribution of Model Residuals by Predicted Value
Standardized Residuals
7.5
5.0
2.5
0.0
-2.5
-5.0
$0
$100,000
$200,000
$300,000
$400,000
Predicted Values


sample
7.5
5.0
2.5
0.0
-2.5
-5.0
-2
0
2
theoretical

In the scatter plot for this model, it appears that we can still observe the concavity that we saw in the scatter plot for model 2. The extreme outliers in the histogram are also still present, which can also be seen in the Q-Q plot. There are still some extreme outliers affecting the model at the high end of our predicted sale price, however even with those removed it appears our model would violate the assumption of homoscedasticity.

e. Based on this information, we should retain both variables as predictor variables of Y. All of the current downfalls of model 3 are also more apparent in the simple linear models. Despite the homoscedasticity violations, the multiple linear regression model is a clear improvement over either simple linear model, which is plainly seen in the large increase in R-squared.

(5) Select any other continuous variable you wish. Fit a multiple regression model that uses 3 continuous explanatory (X) variables to predict Sale Price (Y). These three variables should be your variable of choice plus the explanatory variables from Model 3. Call this Model 4. You should:

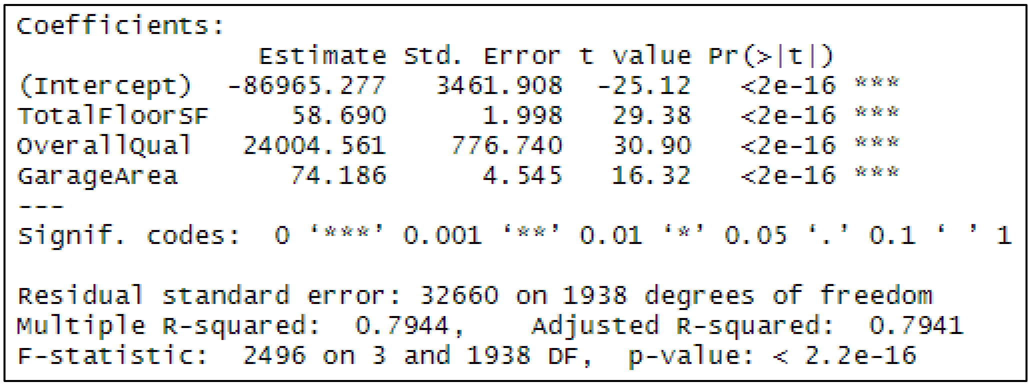
a. Y = 58.69*TotalFloorSF + 24004.561*OverallQual + 74.186*GarageArea - 86965.277

- -86965.277: Our intercept is the value of a home that was rated a 0 in Overall Quality and has zero total square footage and no garage. This is outside of the realistic parameters as the lowest quality rating observed is 1 and we do not have any empty plots of land in our sample. It simply serves as the y-intercept of the line of best fit.
- 74.186*GarageArea: For every one square foot increase in the size of the garage while holding Overall Quality and house square footage constant, we predict an increase in sale price of $74.18.
- 24004.561*OverallQual: For every one point increase in Overall Quality while holding home and garage square footage constant, the predicted price increases by $24,004.56.
- 58.69*TotalFloorSF: For every increase of one square foot in home size while holding the overall quality of the home and garage size constant, the predicted sale price increases by $58.69.
- In this model both β_1 and β_2 are lower than they were in their respective simple linear models and the previous multiple linear model.

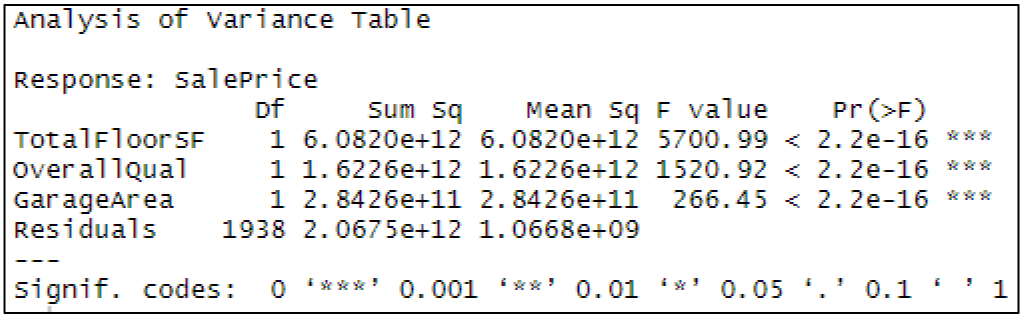
b. R-Squared: 0.794

- This model accounts for 79.4% of the variance in SalePrice.
- This model fits better than either of the simple linear regressions. There is a 0.19 increase in r-squared from model 1 to model 4. That means that our multiple linear regression explains an extra 19% of sale price variance in comparison to the simple linear regression using only total square footage.
- R-squared increase vs model 3: 0.028. The addition of Garage Area accounts for an additional 2.8% of the variation in Sale Price. This variable provides a modest increase over model 3.

c.



```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept) -86965.277   3461.908  -25.12   <2e-16 ***
TotalFloorSF     58.690      1.998   29.38   <2e-16 ***
OverallQual   24004.561    776.740   30.90   <2e-16 ***
GarageArea       74.186      4.545   16.32   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 32660 on 1938 degrees of freedom
Multiple R-squared:  0.7944,    Adjusted R-squared:  0.7941 
F-statistic:  2496 on 3 and 1938 DF,  p-value: < 2.2e-16
```

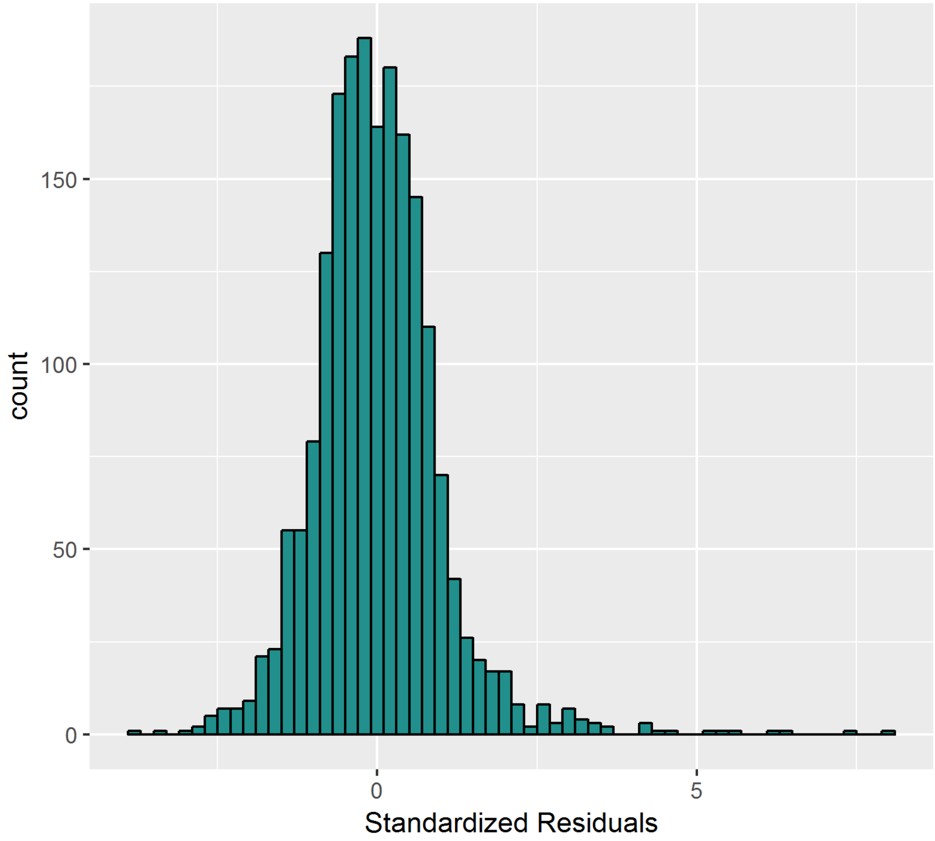


```
Analysis of Variance Table

Response: SalePrice
               Df     Sum Sq    Mean Sq F value    Pr(>F)    
TotalFloorSF    1 6.0820e+12 6.0820e+12 5700.99 < 2.2e-16 ***
OverallQual     1 1.6226e+12 1.6226e+12 1520.92 < 2.2e-16 ***
GarageArea      1 2.8426e+11 2.8426e+11  266.45 < 2.2e-16 ***
Residuals    1938 2.0675e+12 1.0668e+09                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

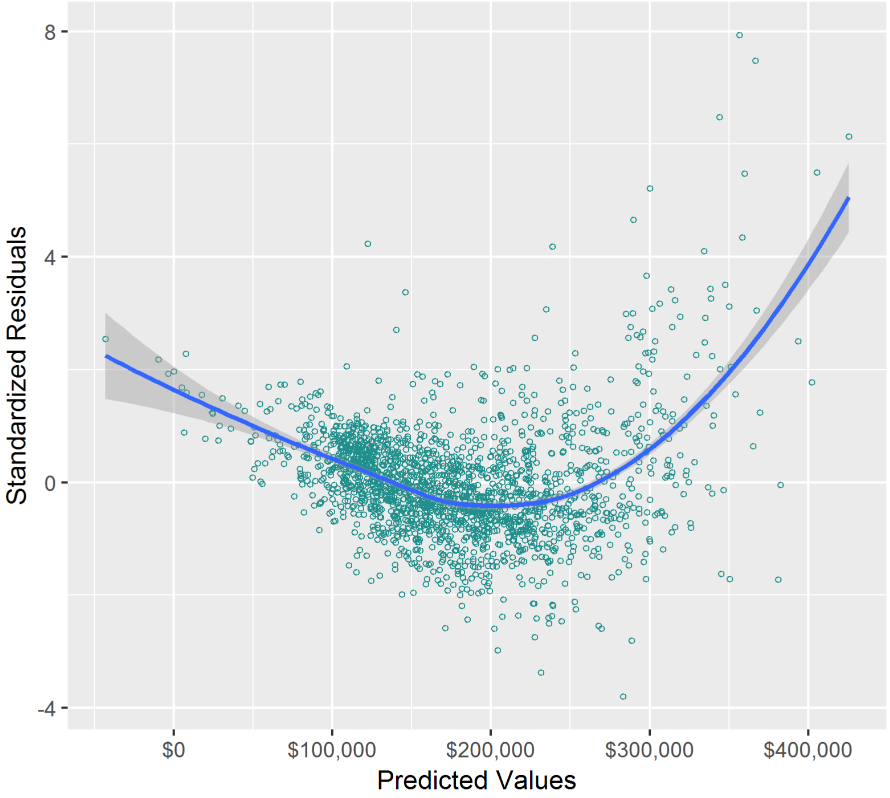
- Intercept Hypothesis Test:
  - H_0: $\beta_0 = 0$
  - H_A: β_0 does not equal 0
  - Reject the null hypothesis at the 0 level
- TotalFloorSF Hypothesis Test:
  - H_0: $\beta_1 = 0$
  - H_A: β_1 does not equal 0
  - Reject the null hypothesis at the 0 level
- OverallQual Hypothesis Test:
  - H_0: $\beta_2 = 0$
  - H_A: β_2 does not equal 0
  - Reject the null hypothesis at the 0 level
- GarageArea Hypothesis Test:
  - H_0: $\beta_3 = 0$
  - H_A: β_3 does not equal 0
  - Reject the null hypothesis at the 0 level
- Omnibus Hypothesis Test:
  - H_0: $\beta_1 = \beta_2 = 0$
  - H_A: At least one beta does not equal 0
  - We reject the null hypothesis and conclude that at least one β does not equal zero.

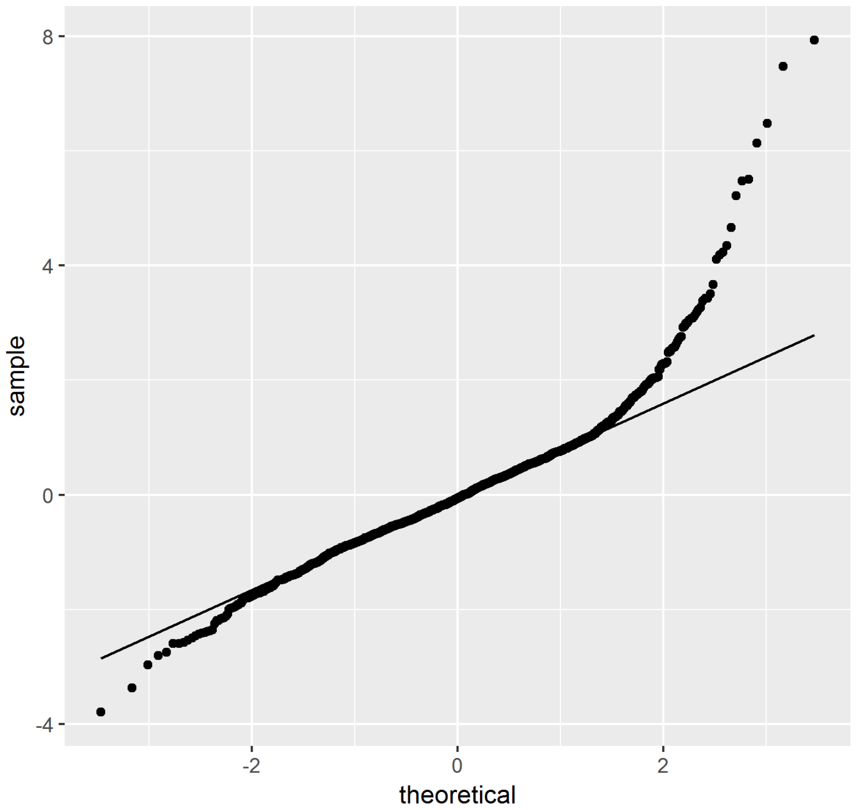
d.

Distribution of Model Residuals



Distribution of Model Residuals by Predicted Value



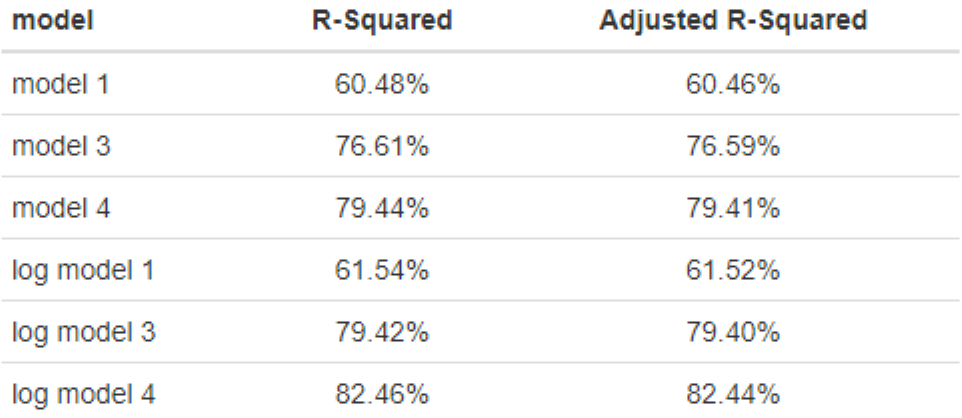


As shown in the LOESS regression line plotted on the scatter plot of standardized residuals, we again see the same trends in the residuals as in previous models. This model violates the assumptions of homoscedasticity in the same manner as previous models. This is also apparent in our QQ plot and histogram. There are outliers that are affecting the model as well.

e. I would argue again in favor of keeping all three variables as predictors. All explanatory variables are highly significant and the inclusion of Garage Area results in a 3% increase in adjusted r-squared over model 3.

PART C: Multiple Linear Regression Models on Transformed Response Variable

(6) The transformed model 4 provides the best fit of all the models we have investigated so far, including showing increases in R-Squared and adjusted R-Squared over the non-transformed models above.

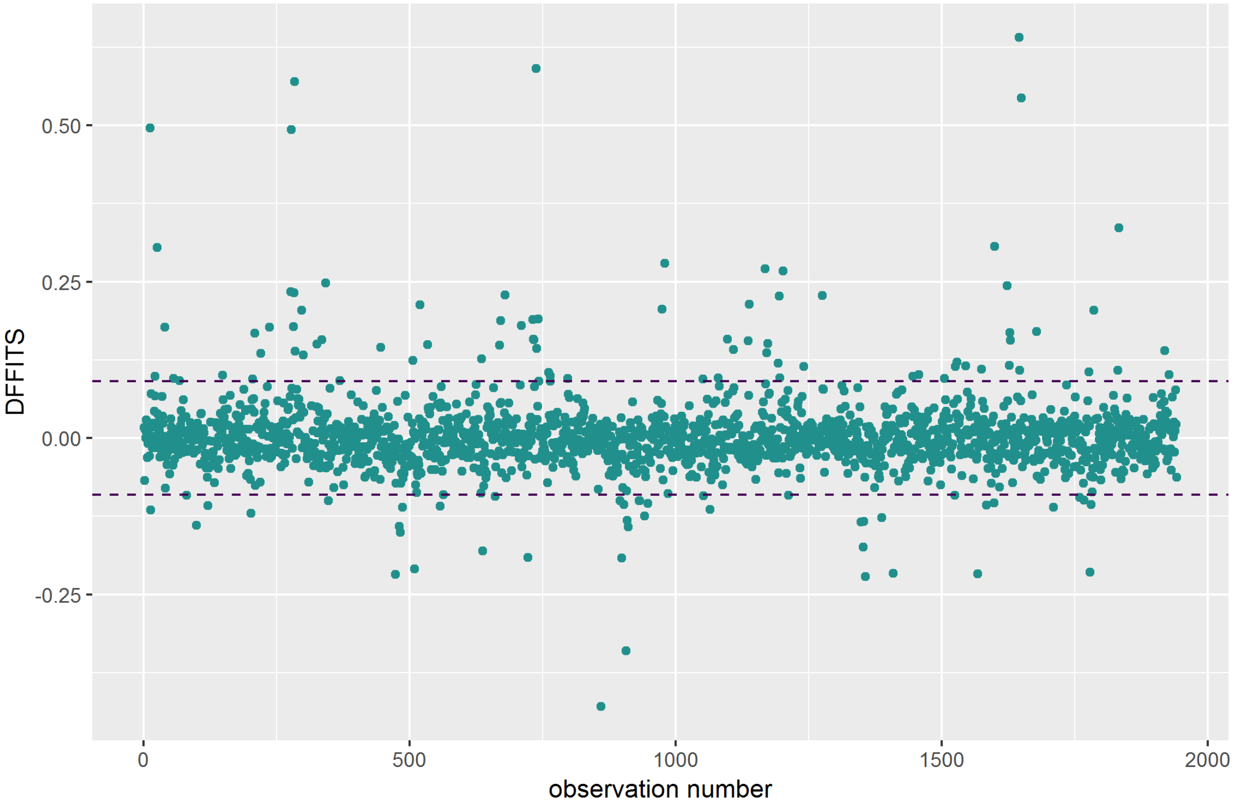


model	R-Squared	Adjusted R-Squared
model 1	60.48%	60.46%
model 3	76.61%	76.59%
model 4	79.44%	79.41%
log model 1	61.54%	61.52%
log model 3	79.42%	79.40%
log model 4	82.46%	82.44%

(7) The output of our log transformed model can be interpreted as e^Y = Sale Price. This is different from the non-transformed model because those did not require any further transformation of the output in order to predict the sale price. The increase in predictive power of the transformed model justifies the transformation of the target variable. Any audience with a sufficiently technical background typically does not need to know the finer details in how a prediction is made. They are typically only concerned with how accurate the model ends up being. For any non-technical person who is interested in learning about the inner workings of the model, it is imperative that a data scientist is able to distill the ideas behind the methodology down in a way that is easily understood to a non-technical audience.

PART D: Multiple Linear Regression and Influential Points

(8) DFFITS threshold: 2*sqrt((p+1)/(n-p-1)) = 2*sqrt((3+1)/(1942-3-1)) = 0.09086



As seen in the chart above, we have many influential points based on DFFITS values, 125 total. After the removal of these 125 influential points, model 4 improves substantially. The new R-Squared value is 0.8388, which represents a 4% improvement over the original model 4.

However I do not believe that this improvement justifies the removal of the influential data points. The 125 influential points represents ~6% of the total sample. Removing them simply because their inclusion does not work well with our model feels disingenuous and misleading, especially considering we removed outlier observations before beginning the analysis. Removal after fitting the models post-hoc is simply choosing the data to make the model more convenient to use.

PART E: Beginning to Think About a Final Model

(9) Use your approach to identify a good multiple regression model to predict SALEPRICE(Y) from the set of continuous explanatory variables available to you in the AMES dataset. For this task you need to:

a. My approach at developing a final model is a mix of data observation and domain knowledge. The first variable that I added in was House Age. This makes intuitive sense since newer houses tend to be more expensive and need less maintenance. This variable was also negatively correlated with sale price at -0.56. In order to capture the additional square footage that is not captured by the TotalFloorSF variable, I first added Gross Livable Area. However this variable was not a significant predictor and did little to add to R-Squared so it was removed. Total Basement Sq Feet proved to be a much better predictor and helped capture the additional increase in home size.

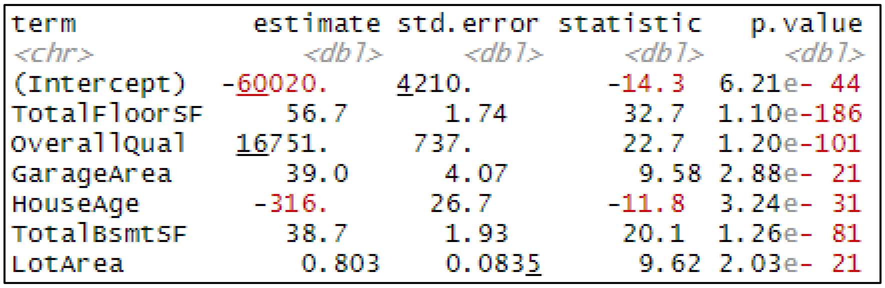
   The next variable I tested out was the number of full bathrooms. While the variable was significant, it had a counterintuitive beta which suggested that adding bathrooms decreased a home's value. It also did little to add to the R-Squared value of the model. I suspect that the signal of this variable is captured in other variables since larger homes tend to have more bathrooms. Therefore I decided to proceed without the Full Bathrooms variable.

   The final variable added into the model was Lot Area. The rationale for this variable is that a larger lot could be more valuable since it provides the opportunity to add an addition or just enjoy a large yard. This variable did prove to be significant and increased the R-Squared value so it was included.

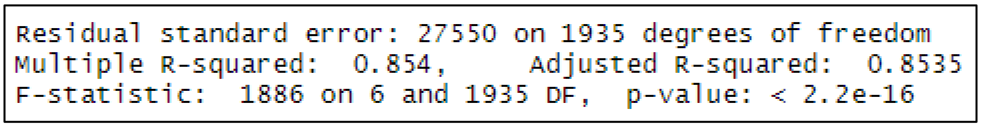
b. Y = 56.7*TotalFloorSF + 16751*OverallQual + 39.0*GarageArea – 316*HouseAge + 38.7*TotalBsmtSF + 0.803*LotArea – 60020

   - -60020: Our intercept is the value of a home that was rated a 0 in Overall Quality, has zero total square footage, no garage, no basement, an age of 0, and a 0 square foot lot. This is obviously outside of the realistic parameters. It simply serves as the y-intercept of the line of best fit.
   - 0.803*LotArea: Every increase of one square foot of lot size increases our price prediction by 0.80, when holding all other variables constant.
   - 38.7*TotalBsmtSF: For every one square foot increase in home size, while holding the other variables constant, predicted price increases $38.70.
   - -316*HouseAge: For every year older the house is, the predicted price drops $316.
   - 39*GarageArea: For every one square foot increase in the size of the garage while holding all other variables constant, we predict an increase in sale price of $39.
   - 16751*OverallQual: For every one point increase in Overall Quality while holding all other variables constant, the predicted price increases by $24,004.56.
   - 56.7*TotalFloorSF: For every increase of one square foot in home size while holding all other variables constant, the predicted sale price increases by $56.70.
   - In this model β_1, β_2, and β_3 are lower than they were in their respective simple linear models and the previous multiple linear models.

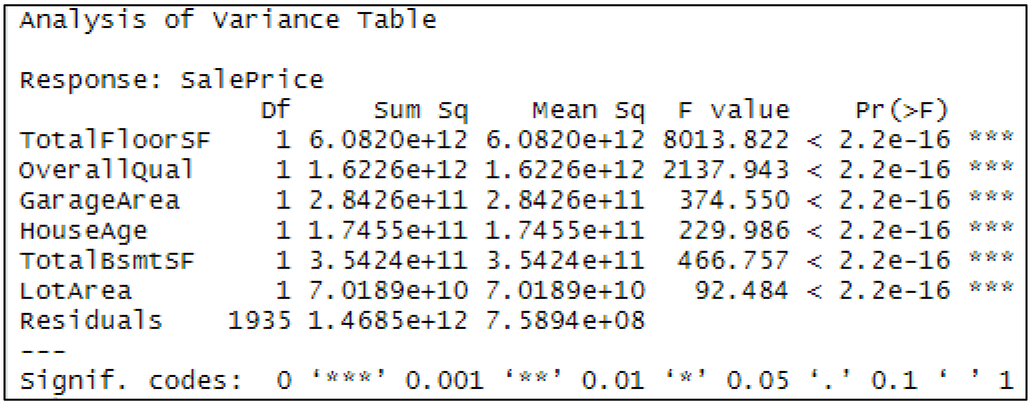
c. Provided the tidyverse coefficient table in order to make beta estimates more legible.



```
term           estimate std.error statistic   p.value
<chr>             <dbl>     <dbl>     <dbl>     <dbl>
(Intercept)  -60020.     4210.        -14.3  6.21e- 44
TotalFloorSF     56.7       1.74       32.7  1.10e-186
OverallQual   16751.      737.         22.7  1.20e-101
GarageArea       39.0       4.07        9.58 2.88e- 21
HouseAge       -316.       26.7       -11.8  3.24e- 31
TotalBsmtSF      38.7       1.93       20.1  1.26e- 81
LotArea           0.803     0.0835      9.62 2.03e- 21
```



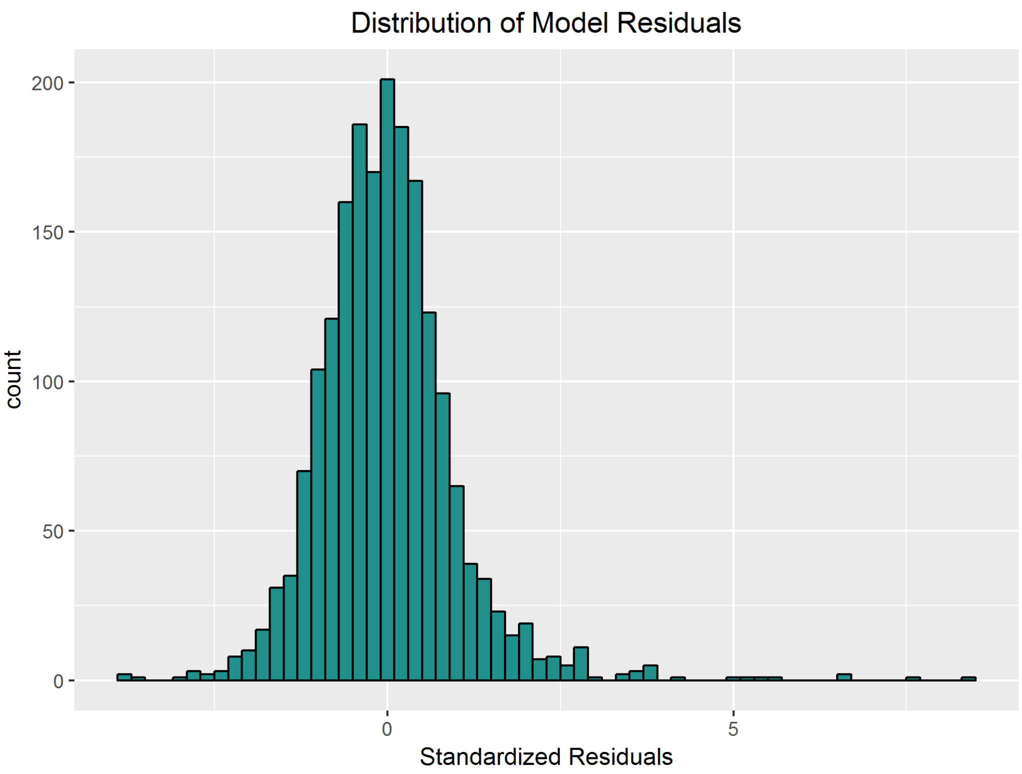
```
Residual standard error: 27550 on 1935 degrees of freedom
Multiple R-squared:  0.854,     Adjusted R-squared:  0.8535
F-statistic:  1886 on 6 and 1935 DF,  p-value: < 2.2e-16
```



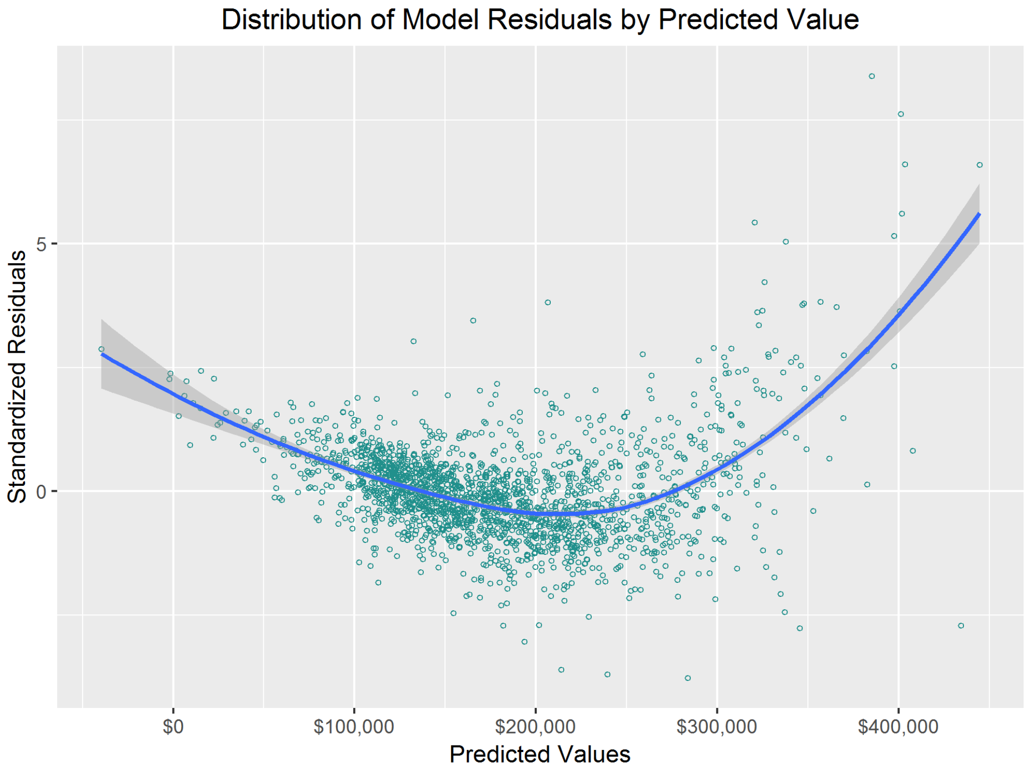
```
Analysis of Variance Table

Response: SalePrice
               Df     Sum Sq    Mean Sq  F value    Pr(>F)
TotalFloorSF    1 6.0820e+12 6.0820e+12 8013.822 < 2.2e-16 ***
OverallQual     1 1.6226e+12 1.6226e+12 2137.943 < 2.2e-16 ***
GarageArea      1 2.8426e+11 2.8426e+11  374.550 < 2.2e-16 ***
HouseAge        1 1.7455e+11 1.7455e+11  229.986 < 2.2e-16 ***
TotalBsmtSF     1 3.5424e+11 3.5424e+11  466.757 < 2.2e-16 ***
LotArea         1 7.0189e+10 7.0189e+10   92.484 < 2.2e-16 ***
Residuals    1935 1.4685e+12 7.5894e+08
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

d. R-Squared: 0.854

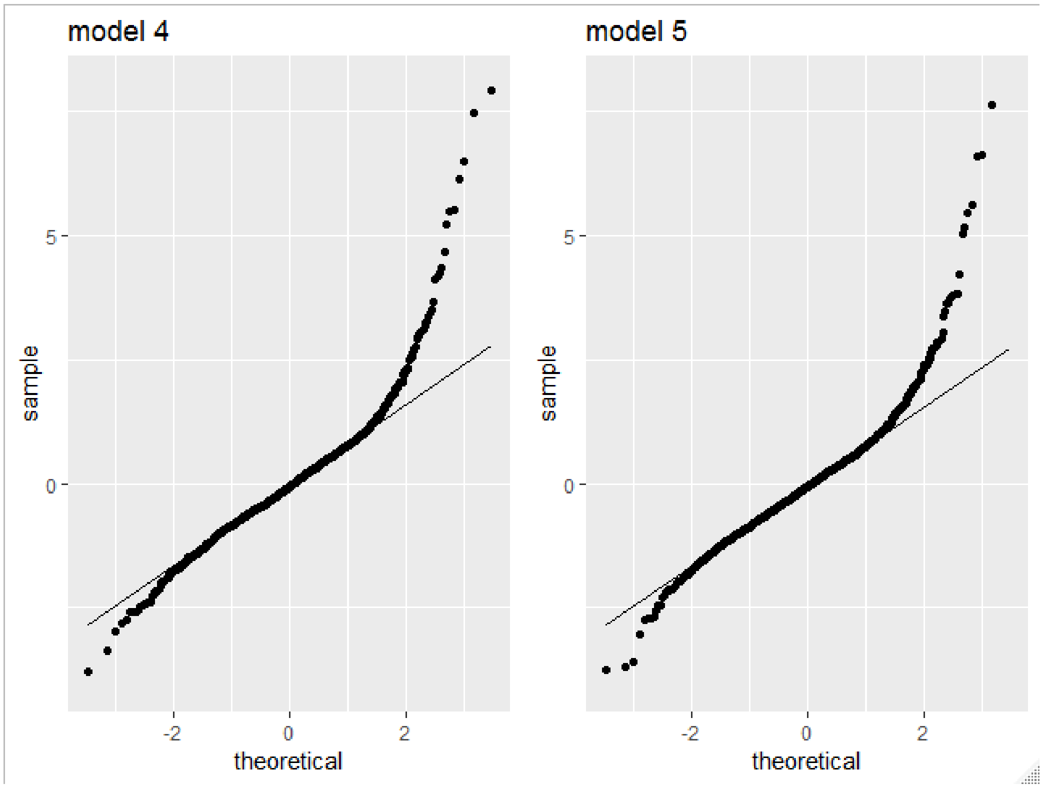
e.



Distribution of Model Residuals



Distribution of Model Residuals by Predicted Value



As with our previous models, we observe the same normality and homoscedasticity violations of the residuals. However, when viewing the Q-Q plot in comparison to model 4, it does appear as though model 5 represents an improvement over previous models. The residuals at the positive end are closer to the Q-Q line than in model 4.

CONCLUSION / REFLECTION

In this assignment we have examined many possible models for the Ames Housing data, ranging from simple to complex. Fitting a simple linear model using total home square feet provided a good baseline model to improve upon. After adding other predictors to our model, we examined variable transformation and outlier deletion. The transformation of our target variable provided a useful way to make the model more accurate without distorting the results, since the output could just be converted back to a final sale price. While this transformation made our model more accurate, it does add some complication when explaining the methodology to non-technical stakeholders. However, the ability to communicate both methodology and outcomes is an extremely important quality in a data scientist and model performance should not suffer due to inability to communicate.

Outlier removal made our model more accurate, but should be used extremely sparingly in order to avoid cherry picking the data to fit the model. Given the number of outliers, removing them from our model suggests that it may not help predict out-of-sample home prices.

When selecting a final model, hypothesis testing played a role in deciding on the variables to include. While they shouldn’t be the only deciding factor on if a variable is included it does provide a good starting point for investigation. Insignificant variables can be disregarded fairly quickly, but significant variables should be examined more in depth to see if they provide value in the fit of the model.

The next step in my modeling process would be investigating the categorical variables in the data set. There is potential to refine the model further using factors. Once a final model is decided upon, it should undergo some validation by testing the predictions on out-of-sample home sales. The validation step will provide valuable feedback on our model's performance.