MODELING ASSIGMENT #2

BRANDON MORETZ

INTRODUCTION

To accurately forecast the value of a home, we must find a relevant dataset that contains accurate information of comparable inventory so that we can explore the significant variables of a home which ultimately determine the sale price of the residence. Once we have explored the data set and selected an appropriate sample from the population, our task will be to create both single and multivariate regression models that leverages these key indicators in the data to predict the value of a home given based upon its features. Once we have constructed the models, we will form hypothesis tests at our stated confidence intervals and conduct statistical significance tests upon these models.

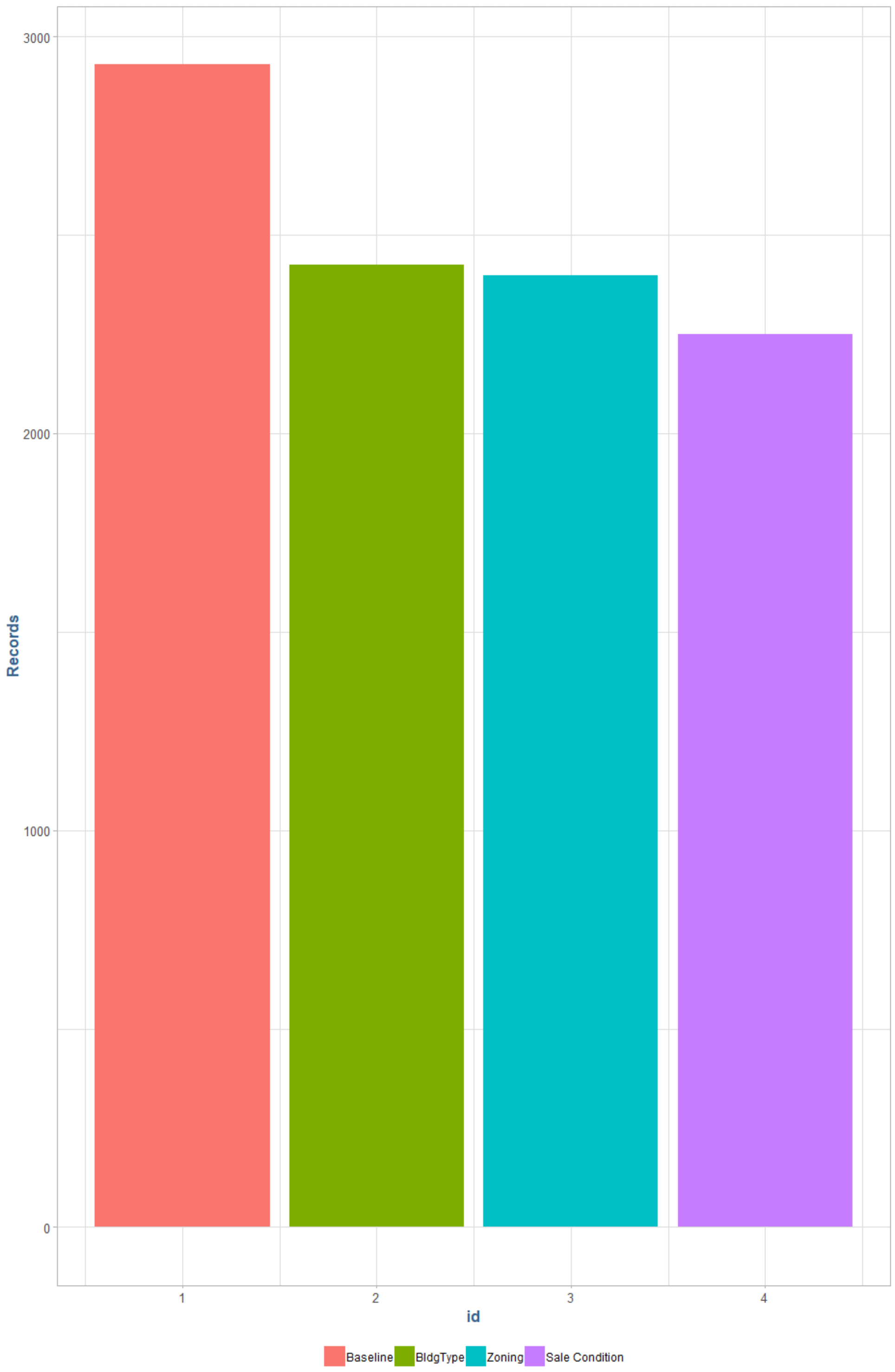
In this report, we will use the Ames dataset which is an alternative to the famous Boston housing data to perform exploratory data analysis through variable derivation, validation, selection and visualization to measure the relevance of these indicators as they pertain to the value of the home in terms of a dollar estimate.

SAMPLE DEFINITION

This data is from the Ames Iowa Assessor's Office and contains characteristics regarding residential properties sold in Ames from 2006 to 2010.

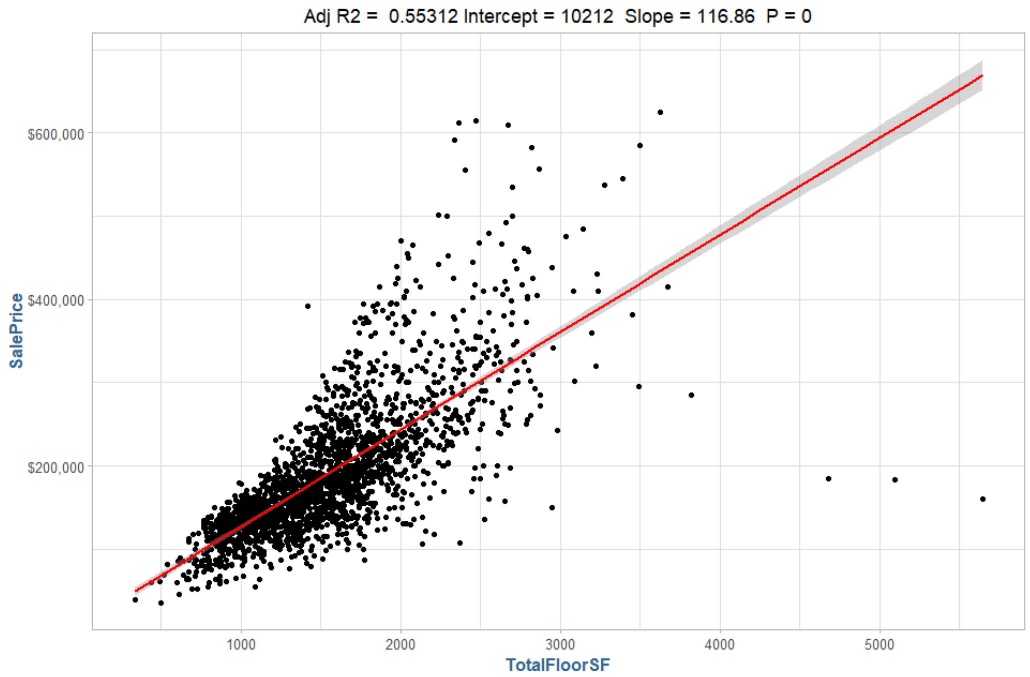
The Ames housing dataset contains approximately three-thousand observations of eighty-two variables collected from the Ames Assessor's Office specifically for assessing value of individual residential properties sold in Ames, Iowa from 2006 to 2010. Given that this data was collected for specifically this purpose, it should be an ideal source of information for our observational study and resulting regression modeling.

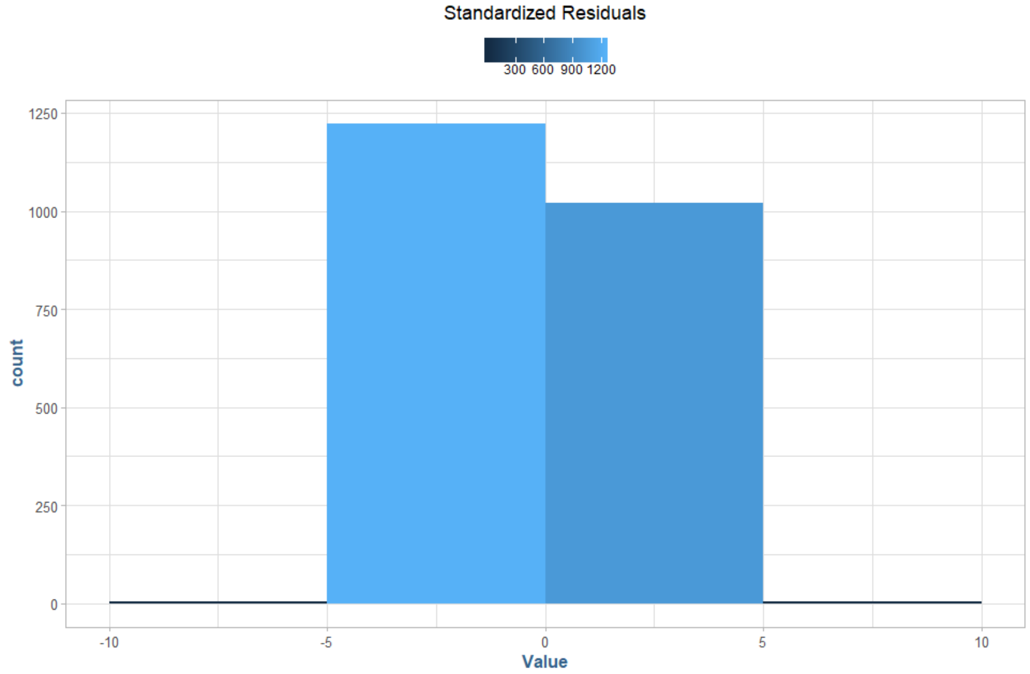
For the sample of homes, we are looking for a "normal" set of homes to build our regression models. For normal in this example, we will choose to only look at single-family style homes (including residential zoning) and non-abnormal sale conditions. We also restrict our analysis to homes with a sale price less than $700,000 as many of our homes meet this criterion. We can see the waterfall of our sample size with each of the preceding chart:



SIMPLE LINEAR REGRESSION MODELS

For the first simple linear regression model, we will look at how the total square footage of the property as an indicator of sale price. We chose this variable because amongst the variables listed in the appendix of variables because it has the highest degree of collinearity in the set of continuous variables.





Model 1: $\hat{Y} = 10212 + 116.862\beta_1$, $R^2 = 0.5533$

The intercept, or β_0, in this model, $10,212, represents the sale price of a home with no square footage, or an empty lot, which is well outside any meaningful values, therefore this value is likely a simple placeholder than any meaningful value. The coefficient β_1 is the total square footage of the house, and each unit increase in total square footage adds approximately $166.86 to the sale price, which makes sense given that homes are often thought of in terms of price per square foot.

$H_0 : \beta_1 = 0$

$H_a : \beta_1 \neq 0$

$t_1 = \hat{B}_1 / S_{\hat{B}1} = 116.862 / 2.215 =$ **52.7594**

t-test with 99% confidence ($\alpha = 0.01$), threshold: $t_{\alpha/2,\, n-p-2} = t_{0.005,\, 2247} = 2.578$

Reject H_0, since $|t_0| > t_{0.005,\, 2247}$

We can reject the null hypothesis that total square footage in this model is no better than the simple slope. Total square footage has a statistically significant impact on the sale price of a given home. The R^2 here suggests that overall quality can be used to explain approximately 55.33% of the global variance in the sale price.

Sum of Squares due to **Regression** = SS_R = 8,108,616,754,275

Sum of Squared **Error** = SS_E = 6,545,927,067,500

Sum of Squares **Total** = $SS_T = SS_R + SS_E$ = 14,654,543,821,775

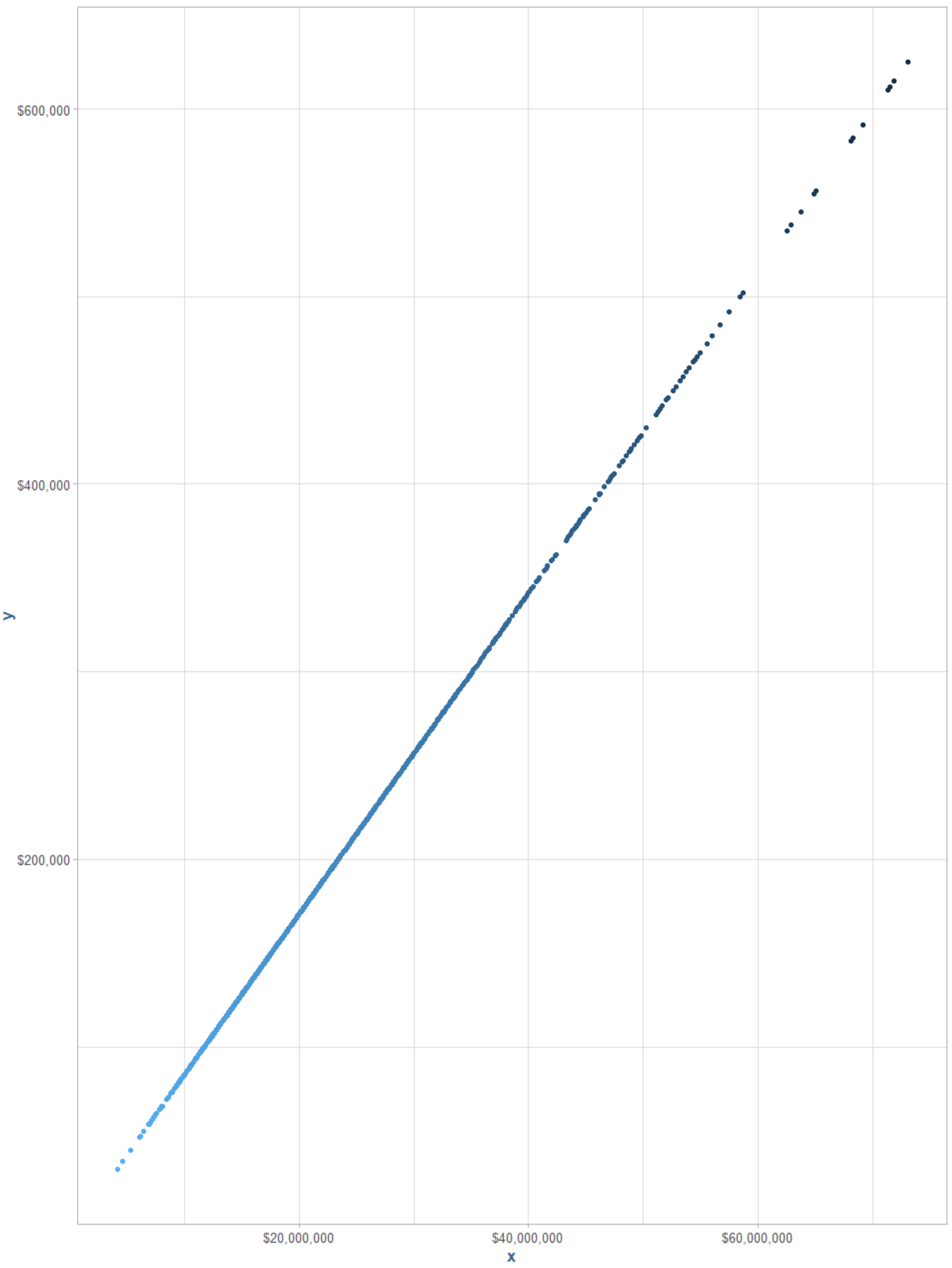
Let,

N = 2250, p = 1

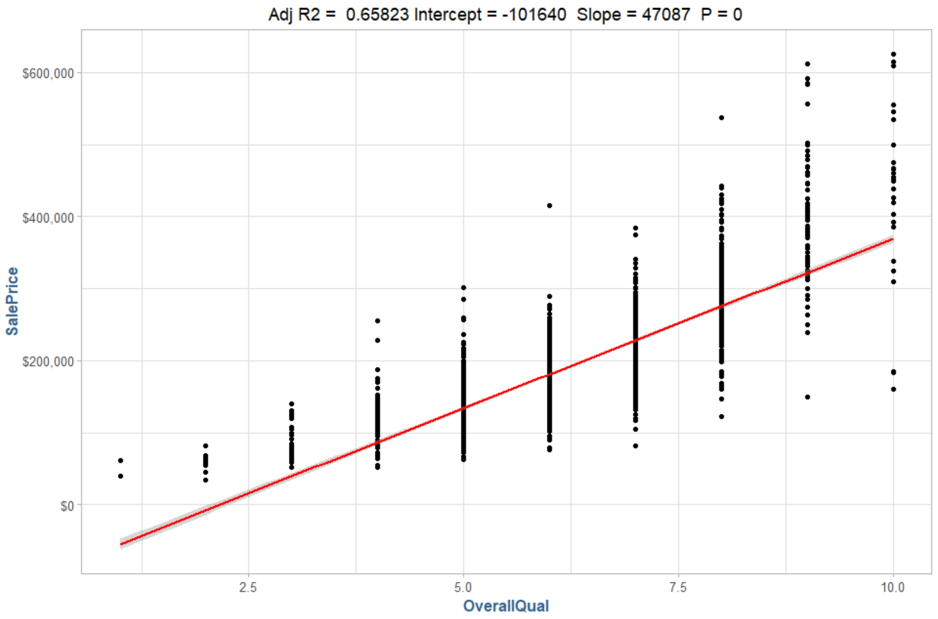
$F = [(SS_T - SS_E) / p] / [SS_E / (n - p - 1)]$ = 8,108,616,754,275 / 2,911,889,265 = **2784.658** on p = 1 and 2248 DF

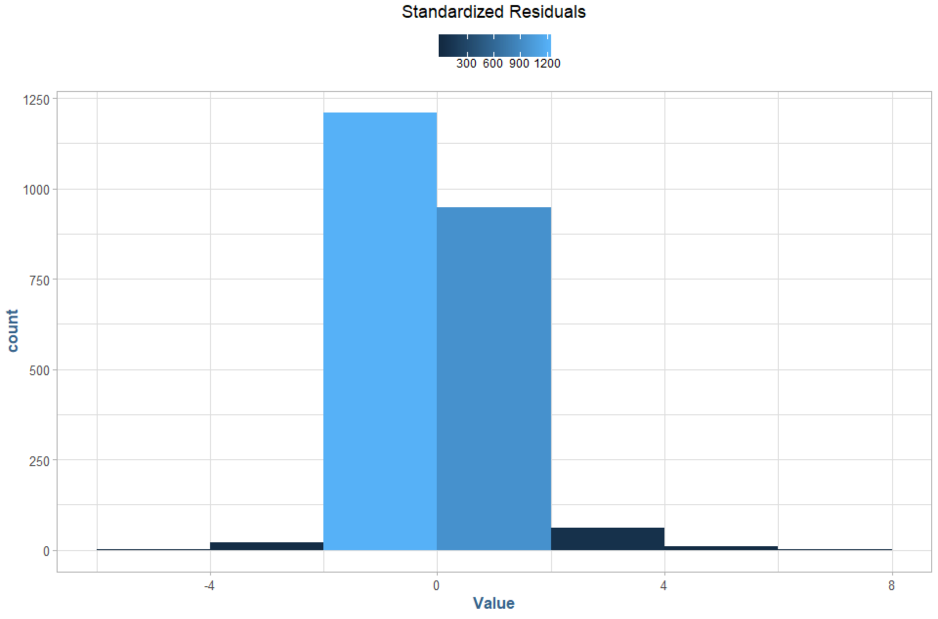
p-value: < 0.0001

There is insufficient evidence (F = 2785, P < 0.001) to conclude that at the slope parameters is not equal to zero (reject the null). This model explains more variance than the intercept alone. A further examination of the standardized residuals and indicate that there are indeed large outliers in the data set given that some of these fall well above/below three standard deviations of the mean, however, there are very view data points that lie in this range with the overwhelming majority falling within the expected range. The following plot shows y prediction against the sale price, with lighter colors indicated less error in the prediction.



For the next simple linear regression model, we will look at the variable with the highest degree of linear correlation to the desired response variable of sale price, which is the overall quality (**OverallQual**) discrete variable. Below we can see a simple linear model fitted against the predictor and the resulting fitted statistics of the model as well as resulting residuals:





Model 2: $\hat{Y} = 101641.5 + 47086.6\beta_1$, $R^2 = \mathbf{0.6584}$

Where β_1 is the overall quality of the house

$H_0 : \beta_1 = 0$

$H_a : \beta_1 \neq 0$

$t_1 = \hat{B}_1 / S_{\hat{B}_1} = 47{,}086.6 / 715.4 = \mathbf{65.8186}$

t-test with 99% confidence ($\alpha = 0.01$), threshold: $t_{\alpha/2,\, n-p-2} = t_{0.005,\, 2247} = 2.578$

Reject H_0, since $|t_0| > t_{0.005,\ 2247}$

We can reject the null hypothesis that overall quality in this model is no better than the simple slope. Overall quality (**OverallQual**) of the home has a statistically significant impact on the sale price of a given home. The R^2 here suggests that overall quality can be used to explain approximately 65.84% of the global variance in the sale price.

Sum of Squares due to **Regression** = SS_R = 240.8018

Sum of Squared **Error** = SS_E = 102.964

Sum of Squares **Total** = $SS_T = SS_R + SS_E$ = 343.7658

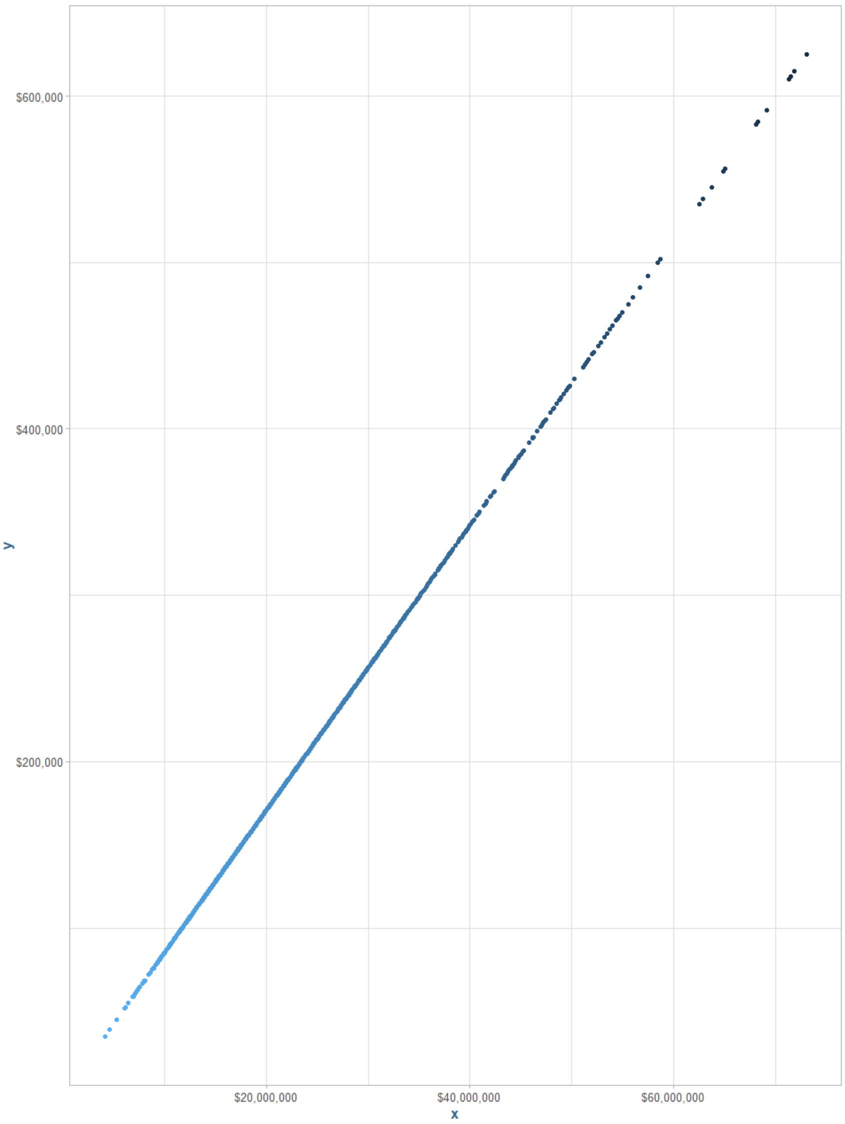
Let,

N = 2250, p = 1

F = $[(SS_T - SS_E) / p] / [SS_E / (n – p – 1)]$ = 9,648,287,918,811 / 2,226,982,163 = **4332.45** on p = 1 and 2248 DF

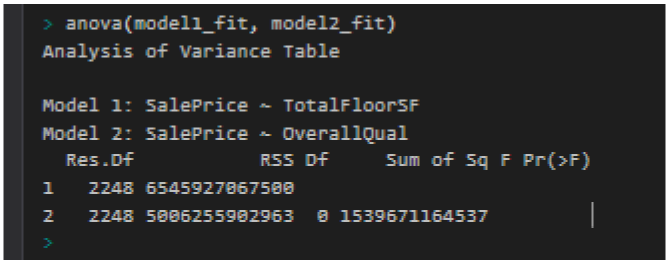
p-value: < 0.0001

There is insufficient evidence (F = 5257, P < 0.001) to conclude that the slope parameters is not equal to zero (reject the null). This model explains more variance than the intercept alone. A further examination of the standardized residuals and indicate that there are indeed large outliers in the data set given that some of them fall well above/below three standard deviations of the mean, however, there are very view data points that lie in this range with the overwhelming majority falling within the expected range. The following plot shows y prediction against the sale price, with lighter colors indicated less error in the prediction.



COMPARISION

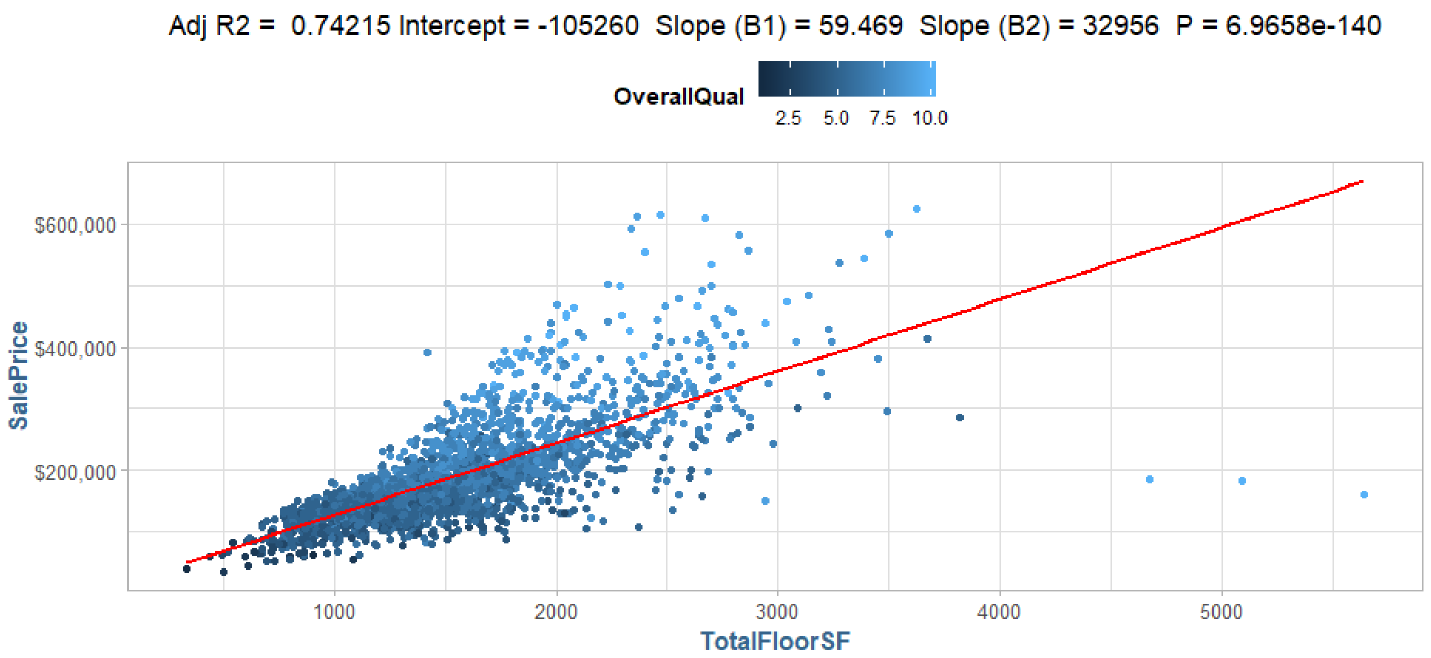
For a comparison of the two models, we can compare the R^2 generated by the each of the resulting models, as well as compare the residual sum of squares generated by each of the linear model fits. For the R^2 value, we can remember that model 1 yielded a value of 0.5533 and model 2 yielded a value of 0.6584. The substantially higher R^2 value in model 2 denotes a relatively higher "goodness" of fit. Additionally, we can look at the two models in an ANOVA setting:

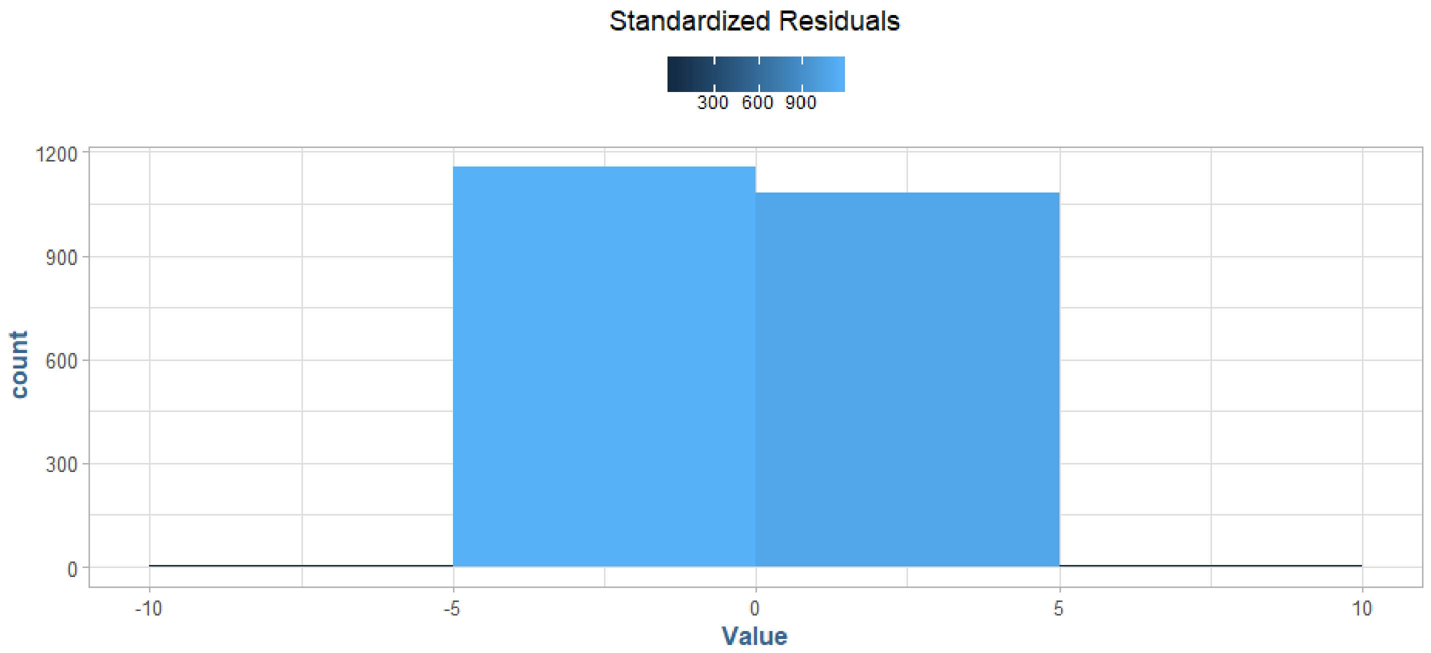


```
> anova(model1_fit, model2_fit)
Analysis of Variance Table

Model 1: SalePrice ~ TotalFloorSF
Model 2: SalePrice ~ OverallQual
  Res.Df           RSS Df     Sum of Sq F Pr(>F)
1   2248 6545927067500
2   2248 5006255902963  0 1539671164537
>
```

Noting that indeed, model 2 has a substantially smaller residual sum of squares, which also indicates it explains more of the variance in the sale price variable than does model 1.

For the next analysis, we will look at a multiple linear regression model with uses both previous explanatory variables in a single model.

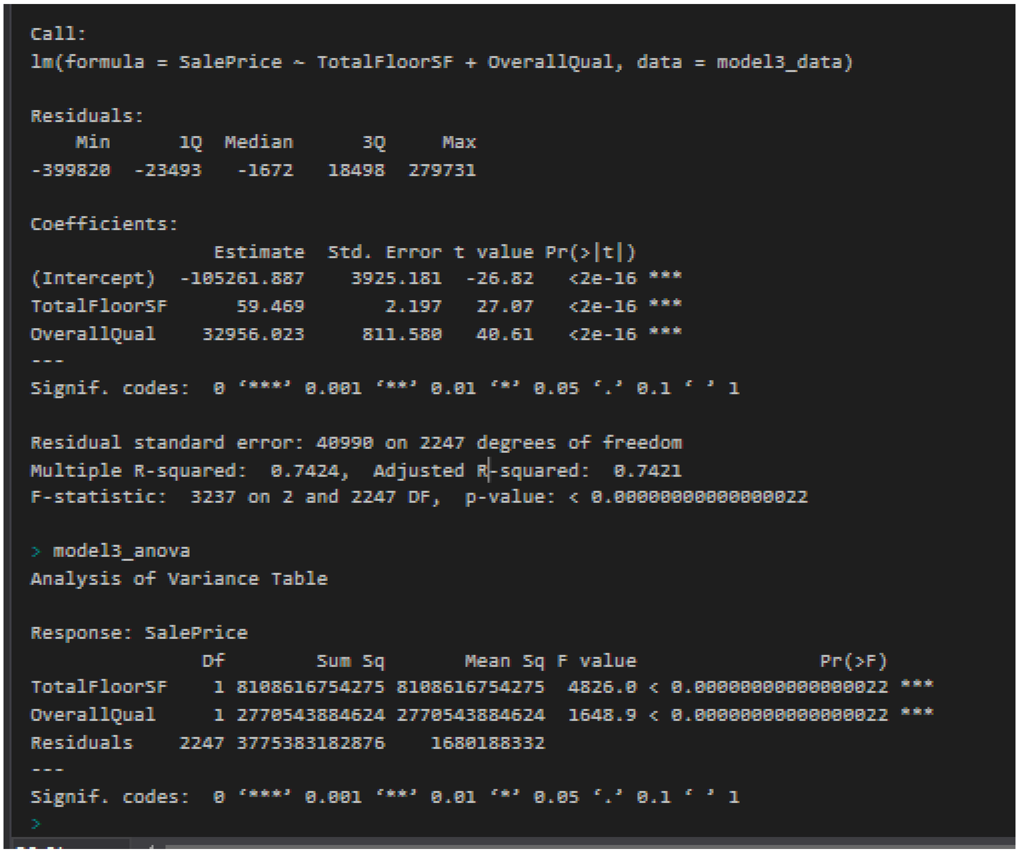




Model 3: $\hat{Y}_{\text{Sale Price}} = -105{,}260 + 59.469\beta_1 + 32{,}956\beta_2$, $R^2 = \mathbf{0.7422}$

In this model we can see there is a negative value for the intercept, meaning that if we were to set both beta one and two coefficients to zero, we would have a home value of negative $105,260, which is clearly irrational in this context. This means that without some degree of value in the coefficients the model is meaningless. For the beta coefficients, we see that each unit of total square footage reflects an additional $59.47 in the sale price, and that for each level of overall quality (on a scale of 1-10), we see an approximately $32, 956 increase in sale price per unit of quality. The R2 in this model indicates that it explains approximately 74.22% of the variance in the sale price using these two indicators, which is better than either of the two previous single variable models we constructed earlier. This reflects a difference of an additional **18.91%** explained variance compared to model 1.

Summary / ANOVA



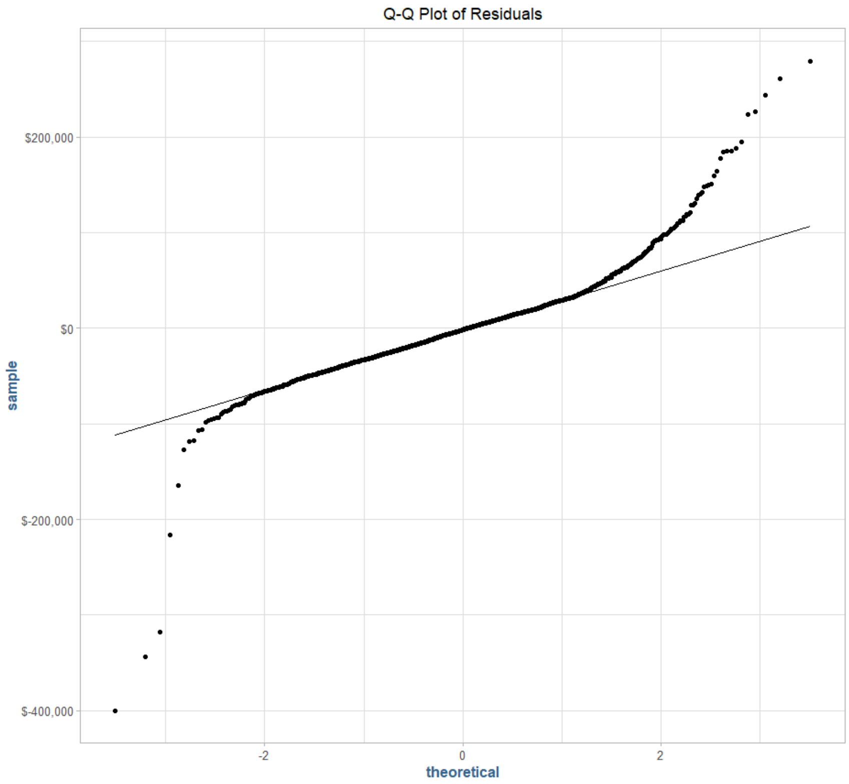
```
Call:
lm(formula = SalePrice ~ TotalFloorSF + OverallQual, data = model3_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-399820  -23493   -1672   18498  279731 

Coefficients:
                Estimate  Std. Error t value Pr(>|t|)    
(Intercept)  -105261.887    3925.181  -26.82   <2e-16 ***
TotalFloorSF      59.469       2.197   27.07   <2e-16 ***
OverallQual    32956.023     811.580   40.61   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 40990 on 2247 degrees of freedom
Multiple R-squared:  0.7424,	Adjusted R-squared:  0.7421 
F-statistic:  3237 on 2 and 2247 DF,  p-value: < 0.00000000000000022

> model3_anova
Analysis of Variance Table

Response: SalePrice
               Df        Sum Sq       Mean Sq F value                Pr(>F)    
TotalFloorSF    1 8108616754275 8108616754275  4826.0 < 0.00000000000000022 ***
OverallQual     1 2770543884624 2770543884624  1648.9 < 0.00000000000000022 ***
Residuals    2247 3775383182876    1680188332                                  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```

Sum of Squares due to **Regression** = SS_R = 8,108,616,754,275 + 2,770,543,884,624
Sum of Squared **Error** = SS_E = 3,775,383,182,876
Sum of Squares **Total** = SS_T = SS_R + SS_E = 14,654,543,821,775

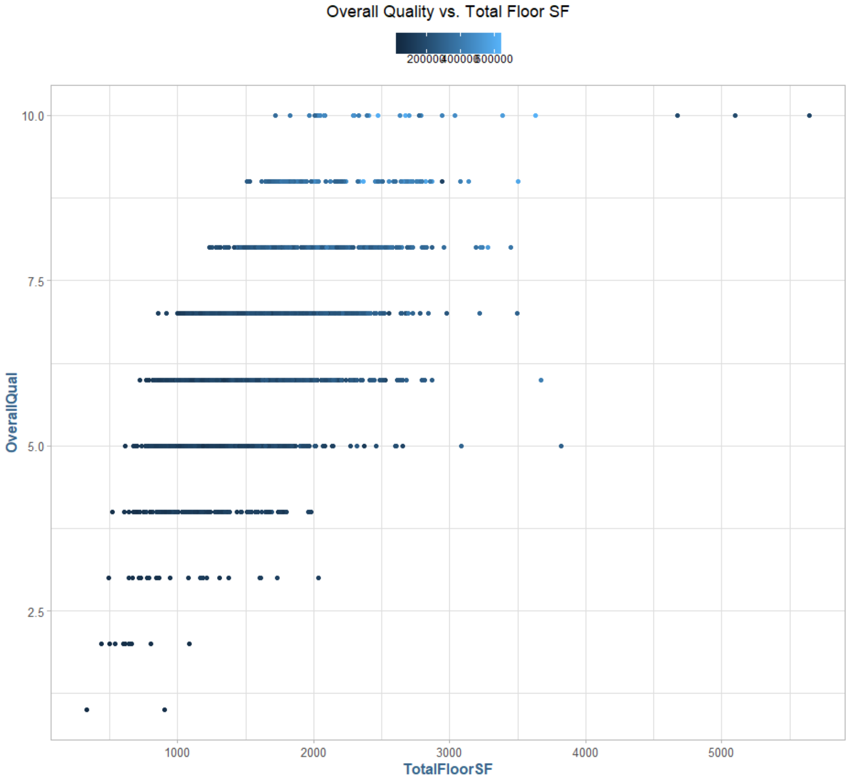
Let,
N = 2250, p = 2

F = [(SS_T - SS_E) / p] / [SS_E / (n – p – 1)] = 5,439,580,319,449 / 1,680,188,332 = **3237.483** on p = 2 and 2247 DF
p-value: < 0.0001

There is insufficient evidence (F = **3237**, P < 0.001) to conclude that at least one of the slope parameters is not equal to zero (reject the null). This model explains more variance than the intercept alone, and in this case the intercept is of no practical use without the additional explained variance provided by the beta coefficients. The standardized residuals here are again normally distributed around mean zero. However, the heavy-tailed distributions indicate that the further ends of the spectrum of sale price will yield less reliable predictions.



In this model, we should keep both coefficients both because they both display statistical significance at the individual t-test level in the model (t < 0.001 for both variables) as well as in the practical sense in that one would not expect the total square footage of a home to be related to the overall quality of the home. You can have high-quality, low-area homes, and low-quality, high-area homes. As we can see in the following example, we get a relatively uniform distribution of homes in each quality bucket, with the sale price varying from low (dark) with lower quality/smaller area, to high (light) with higher quality and more area.

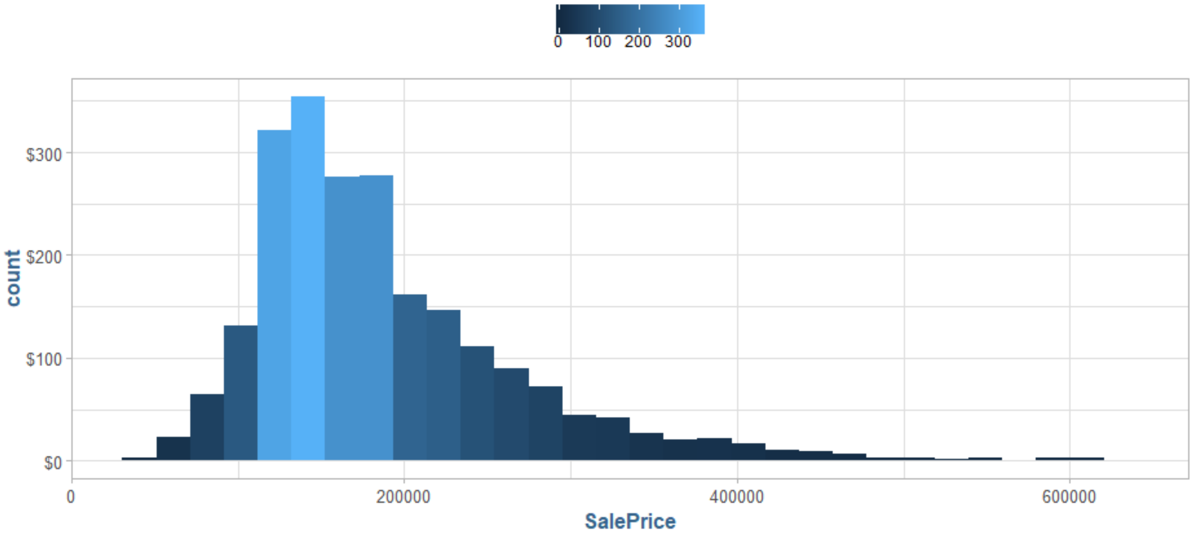


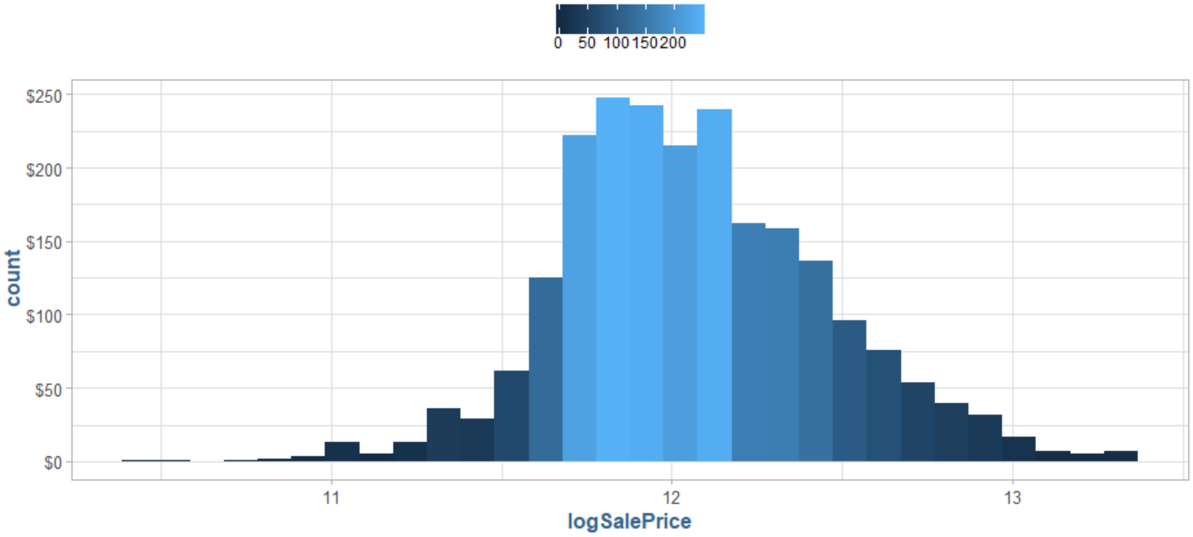
TRANSFORMED RESPONSE

For the next exercise, we will look at each of the preceding models fitted to a logarithmic transformation of the response variable, log sale price. In the following table, we will compare the R^2 of the normal model, the log model, and then look at the delta (log – norm), of the R^2 in the two models.

Model	NormRsq	LogRsq	Delta
1	0.5533	0.5733	0.0200
2	0.6584	0.7005	0.0421
3	0.7424	0.7821	0.0397

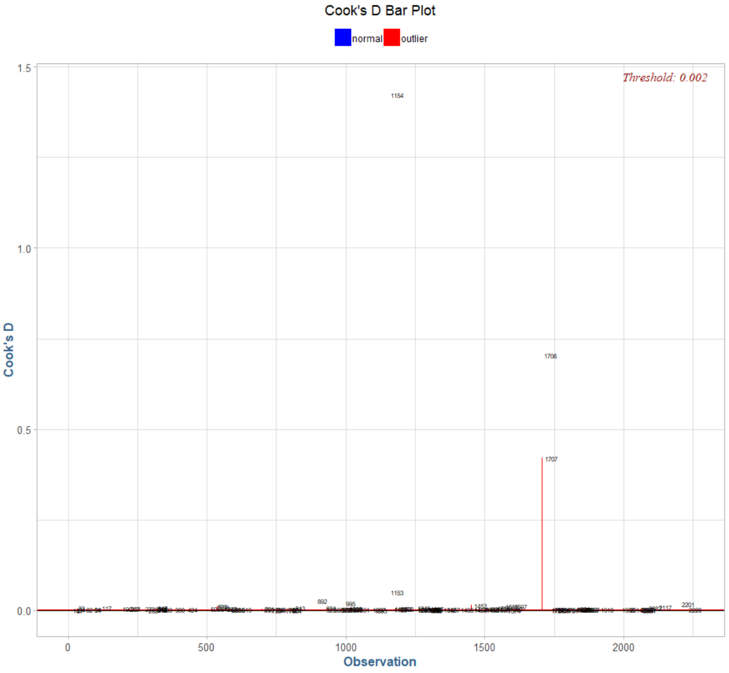
We can see that the log transformation adds additional explained variance to each of the explanatory variables, especially in model 2 and 3 with almost 4%. The reason this particular transformation is useful is that the response variable, sale price, is relatively skewed as we have noted throughout this lab, especially at the tails of the data. This transformation enables us to more safely assume normality in the underlying data, as well as better fit linear models to the transformation of the response. Below we can see the distribution of the normal sale price, and the natural log transformed equivalent:



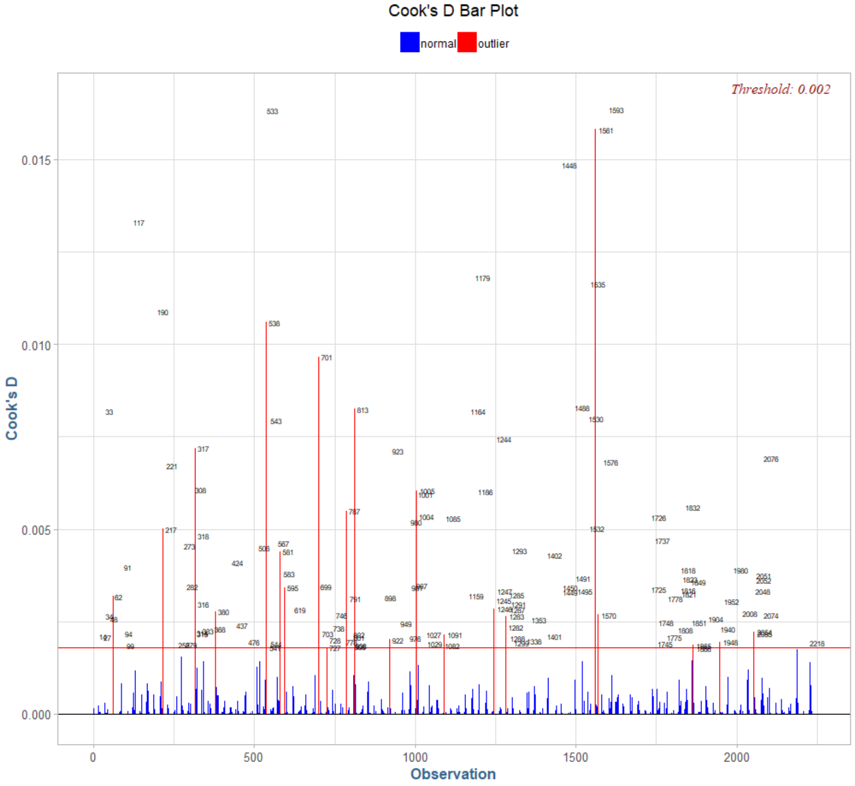


MULTIPLE LINEAR REGRESSION AND INFLUENTAL POINTS

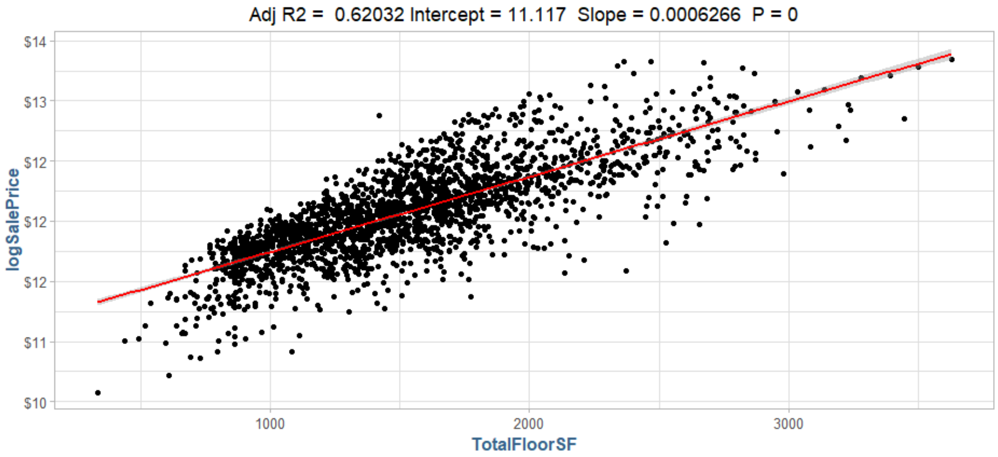
The presence of outliers in the sale price variable have been noted throughout this lab, and in this section, we will attempt to address these influential points with additional techniques. The first technique we will look at is Cook's distance, which attempts to identify observations of heavy influence in the data.

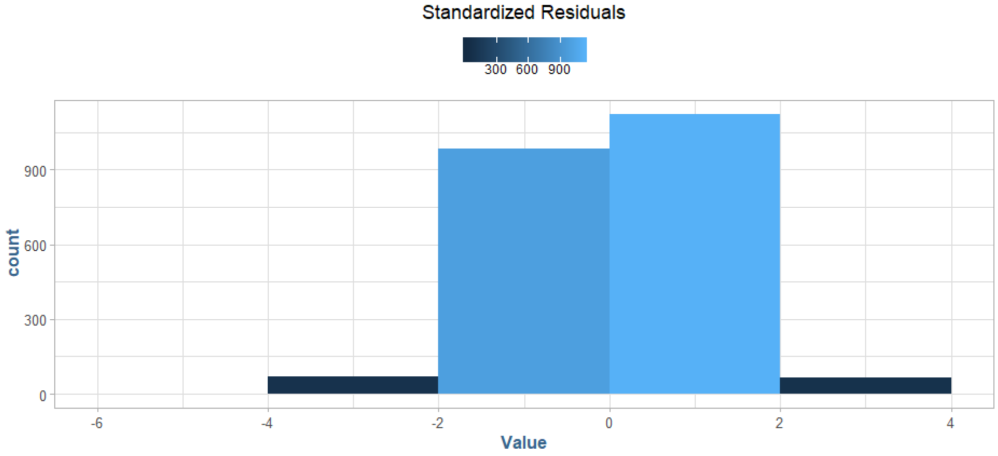


After several iterations of running Cook's distance to delete outlier influential data points (removing a total of 8 rows, 1706, 1707, 1154, 892, 994, 1151, 2195 and 2111) we can note the following change in Cook's distance display:



Now, we note the following updates to the model fit post-outlier removal:





Taking note of the improvement in R² is an additional 5% over the previous model fitted with the outliers present.

I believe techniques that detect outliers that significantly impact the fit of a model and the resulting improved accuracy, explained variance, and then the ultimately the overall "usability" and reliability of the resulting model to be a valuable statistical technique. We should take care to remember our overall objective here, and that is to accurately predict the price of an *average* home given some information about it. If we concern ourselves too much with fitting our model to all data points, outliers or not, we will ultimately overfit the model and it will be of little use to future, unseen data sets.

FINAL MODEL

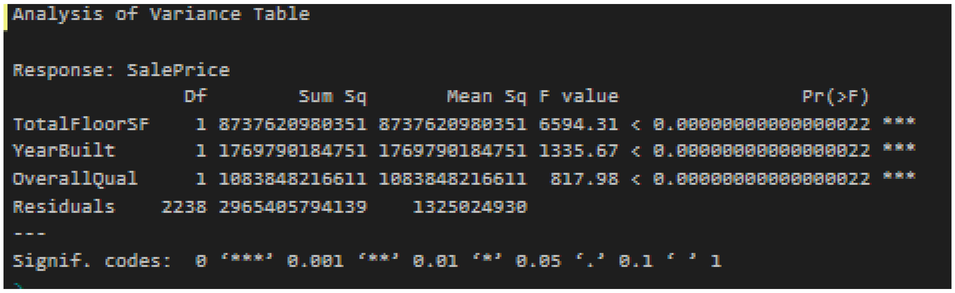
For the final model in this lab, I would like to revisit our feature correlation table which can be found in the appendix to select three variables that have both high linear correlation to the sale price, as well as an intuitively strong presence of correlation with minimal overlap or interaction between the terms. Running a simple comparison of the R² after each of these terms is added to the model yields the following table:

Model	Features	Rsq	Change
1	TotalFloorSF	0.6002488	NA
2	TotalFloorSF + GrLivArea	0.6014527	0.001203857
3	SalePrice ~ TotalFloorSF + GrLivArea + OverallQual	0.7732617	0.171809062
4	SalePrice ~ TotalFloorSF + OverallQual	0.7732263	-0.000035417
5	SalePrice ~ TotalFloorSF + OverallQual + YearBuilt	0.7962854	0.023059046
6	SalePrice ~ TotalFloorSF + OverallQual + MasVnrArea	0.7950077	-0.001277671
7	SalePrice ~ TotalFloorSF + OverallQual + LotArea	0.7863691	-0.008638584

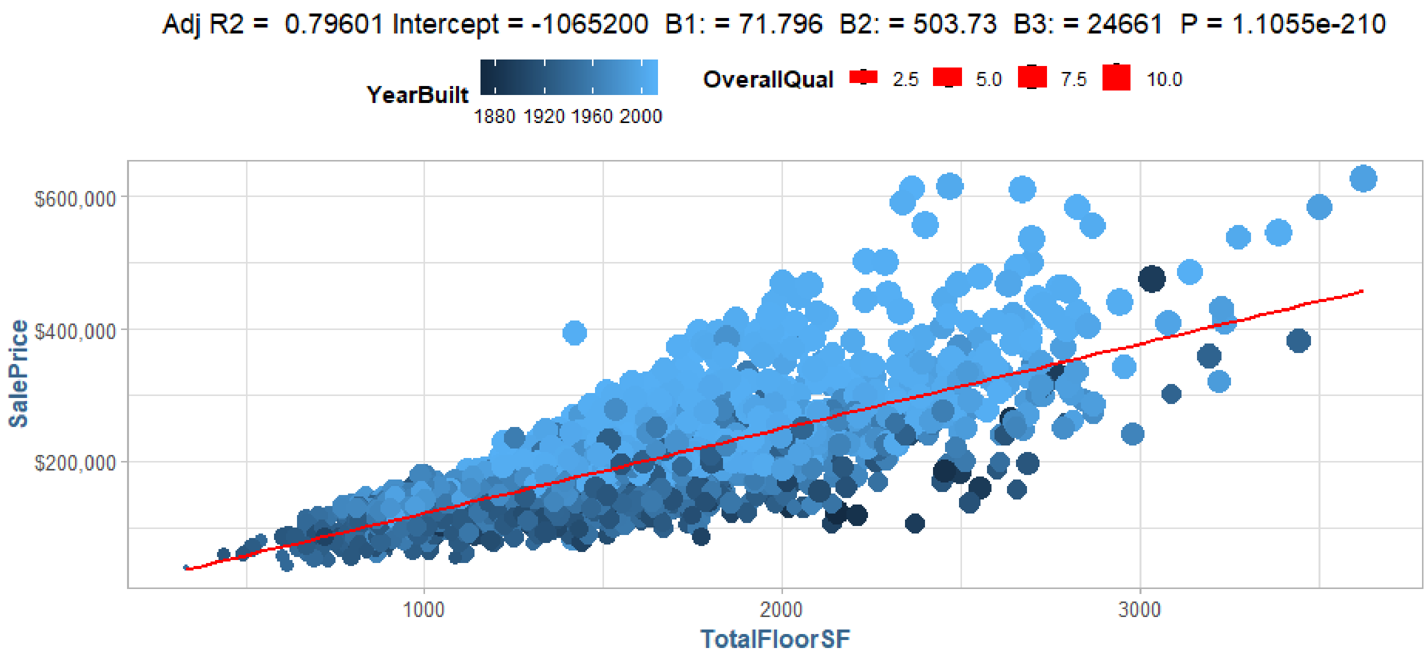
We will pick the total square footage (**TotalFloorSF**), year built (**YearBuilt**) temporal discrete, and the overall quality (**OverallQual**). We will exclude the above ground living area (**GrLivArea**) variable as it does not result in a meaningful impact on the R² value when it is removed, thus preferring as simple a model as possible all other things being equal. For the set of variables chosen for exploration, we will use total square footage (**TotalFloorSF**), overall quality (**OverallQual**) and year built (**YearBuilt**), as these had the most impact to the R2 metric in the previous table.

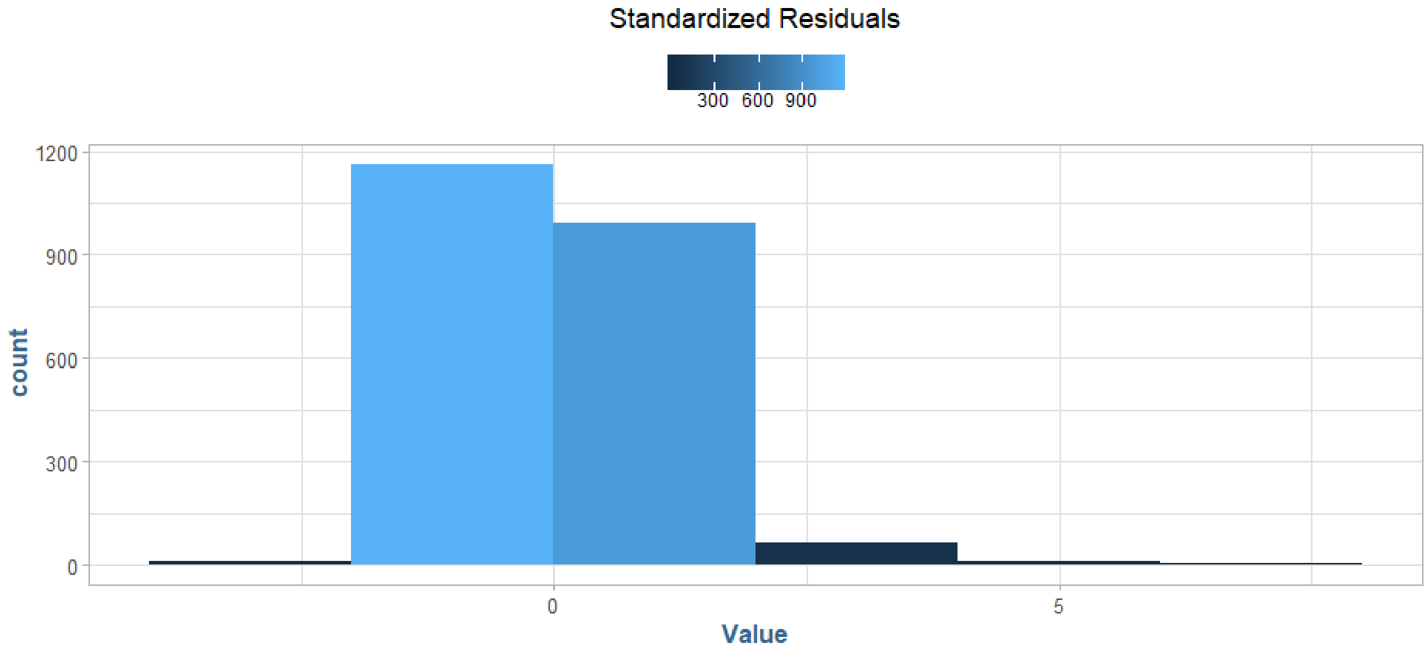
The below diagram uses the natural log transform of the sale price against the total square footage of the house. Along with this core data, we can see the points are colored by oldest (dark) to newest (light) and the size of the points are controlled by the overall quality of the home (where the bigger the higher the quality).

ANOVA



```
Analysis of Variance Table

Response: SalePrice
               Df        Sum Sq       Mean Sq F value                Pr(>F)    
TotalFloorSF    1 8737620980351 8737620980351 6594.31 < 0.00000000000000022 ***
YearBuilt       1 1769790184751 1769790184751 1335.67 < 0.00000000000000022 ***
OverallQual     1 1083848216611 1083848216611  817.98 < 0.00000000000000022 ***
Residuals    2238 2965405794139    1325024930                                  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

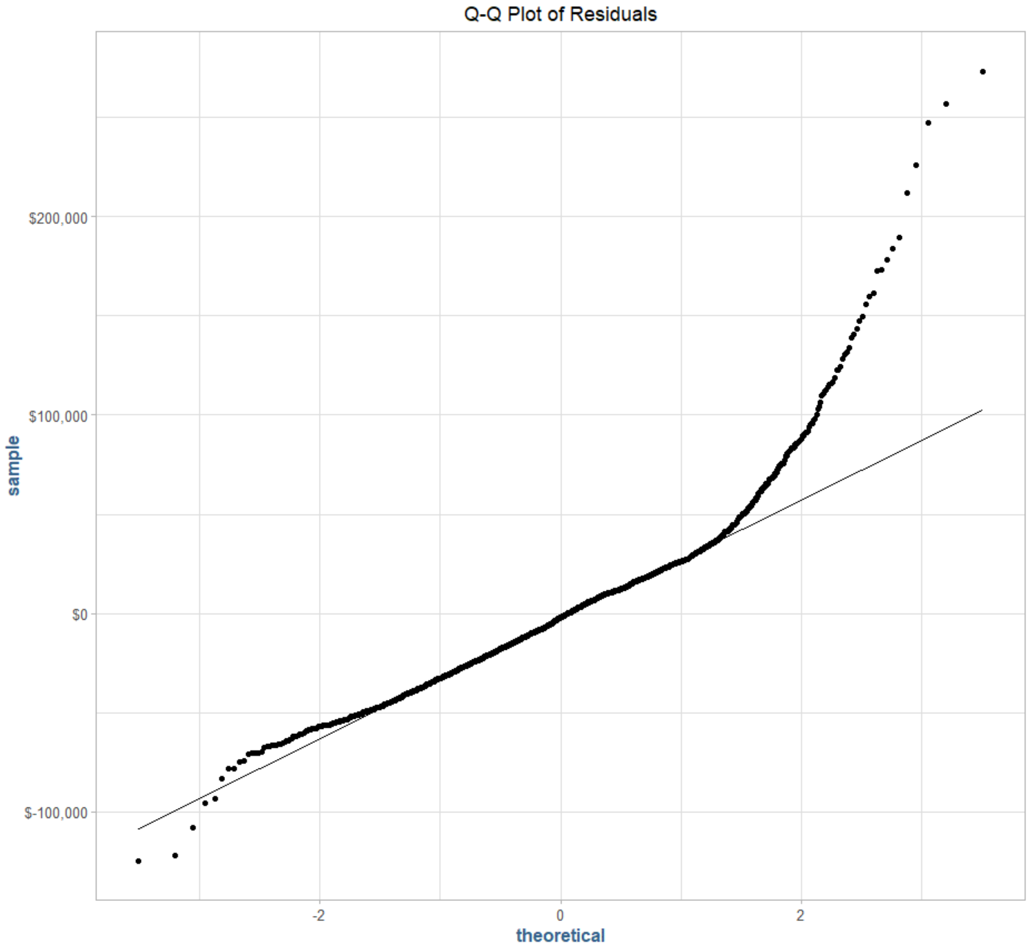




Model 5: $\hat{Y}_{\text{Sale Price}} = -1{,}065{,}200 + 71.976\beta_1 + 503.73\beta_2 + 24661\beta_3$, $R^2 = \mathbf{0.7960}$

Here we see the intercept is -$1,065,200, with such a low value it is safe to say that the intercept in this case is useless without the use of the coefficients in the model. As for the beta coefficients, we see that total square footage adds approximately $71.98 per square foot, year built adds $503.73 per year built after 1880, and overall quality adds approximately $24,661 per unit increase.

The resulting Q-Q plot of the residuals to look for normality given we removed substantially influencing outliers from the data set:



While not perfect, is much improved over the original. We also note the high percentage of variance explained in the data set at 79.60%, which is superior to anything we have seen thus far in our analysis.

CONCLUSION

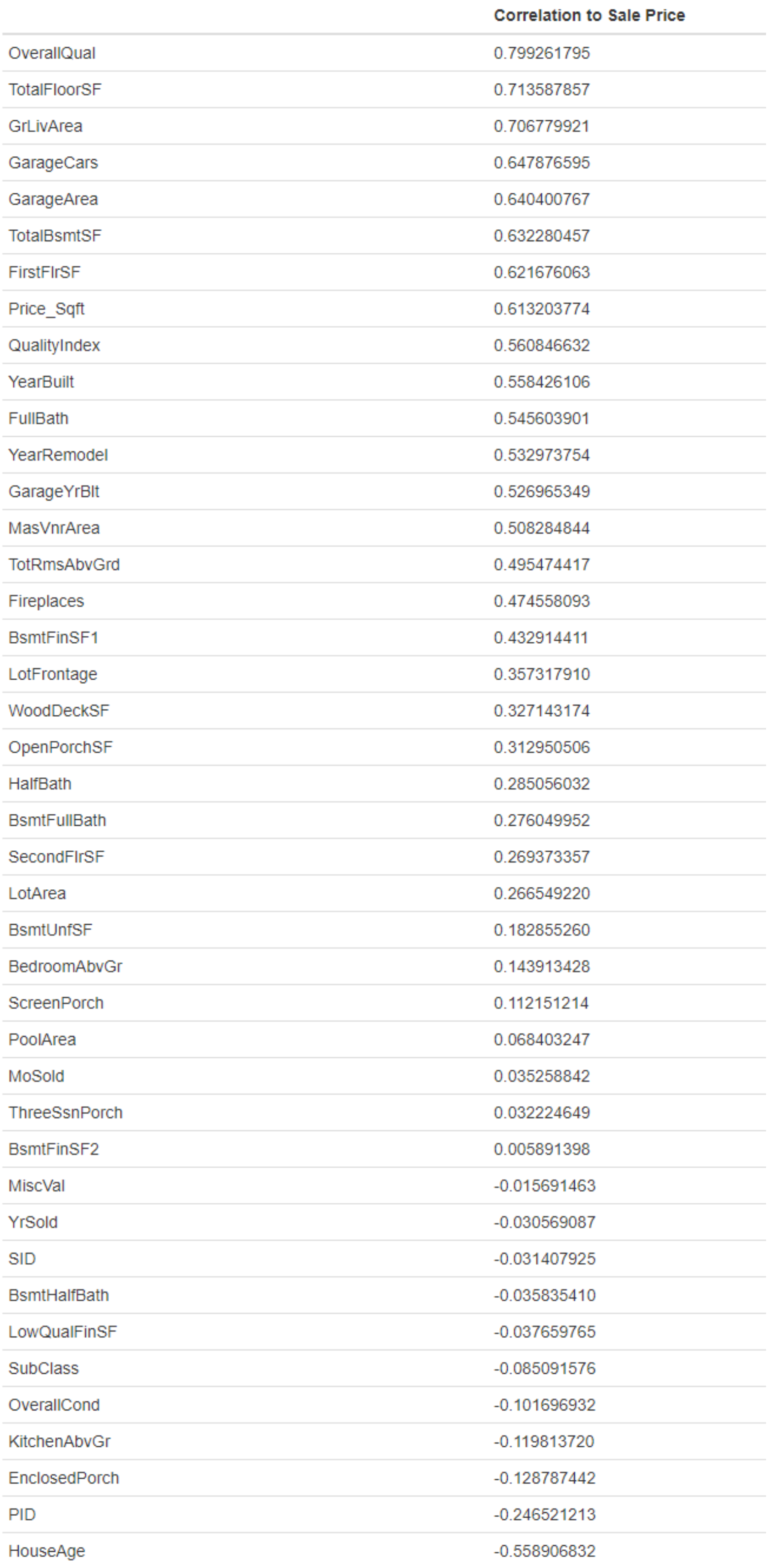
In this lab we explored several variations of regression models in order to accurately predict the sale price of a home. We first further refined our sample from the original definition we constructed in the first modeling lab, refining our data further by building type, zoning and sale condition so that out model will be more reflective of the properties we are trying to account for.

We also constructed simple linear regression models based upon the highest colinear continuous variable, total square footage, interpreted the model and conducted a t-test for statistical significance, which there was evidence for. We examined the plot of predicted y-values against the actual y values and examined these residuals from the model in further detail. We conducted a similar exercise for the overall quality variable, and then compared the two models by their R^2 values. We concluded that the second model not only had a higher R^2, but also demonstrated lower residual sum of squares, noting that it explained more of the variance in the data than did model 1.

Next, we constructed a multiple linear regression model from both indicator variables and performed the omnibus F-test on the model, resulting in statistical significance in at least one of the coefficient terms, while each had an individually strong t-test value as well. We examined the distribution of the residuals, noting that there does appear to be some heavily influencing outliers in this data set. To attempt resolution on some of these values, we transformed the response variable into its natural log representation, which transformed the response into a clearly normal distribution, and iterated over several runs of Cook's distance in order to remove the heaviest impacting variables. The final fit on the post-cleanup data set yielded a substantial improvement.

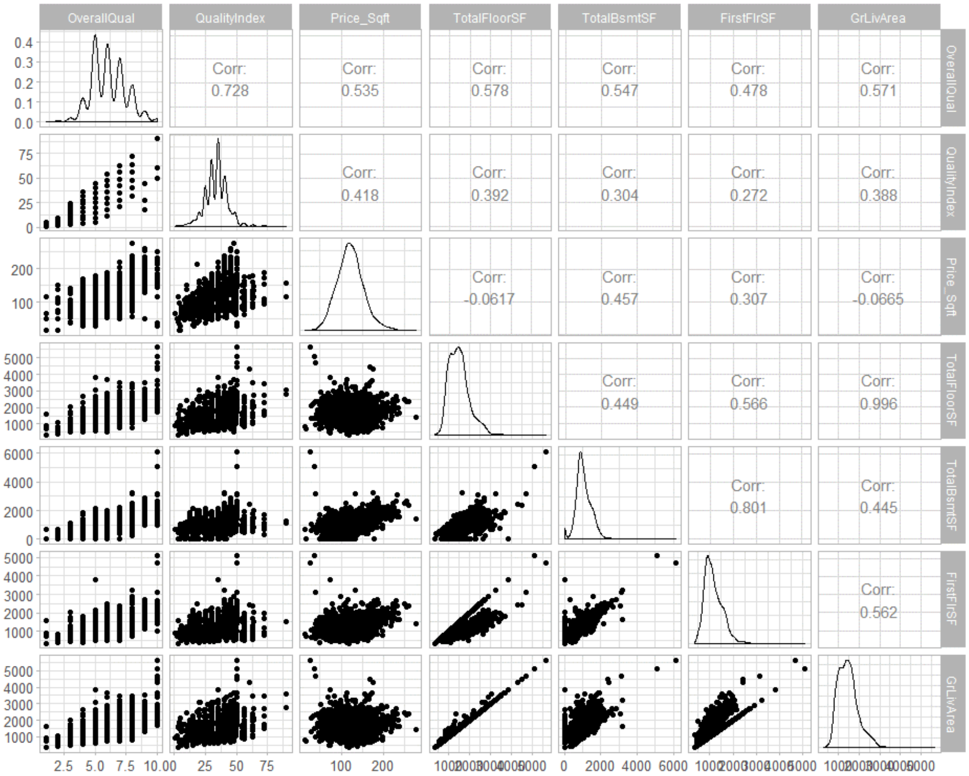
Finally, we started to construct our final model for predicting home values using a process of elimination by looking at changes in the R^2 metric amongst several models with varying terms. We concluded with a model that has both a high R^2, statistically and intuitively impacting terms, and a dense, normal distribution of errors in the predictions. We noted the improved form to normally of the residuals given the removal of the outliers in prior work and generated our best performing model to date. Our next step should be to evaluate the performance of this model using a cross-validation technique.

FEATURE CORRELATION

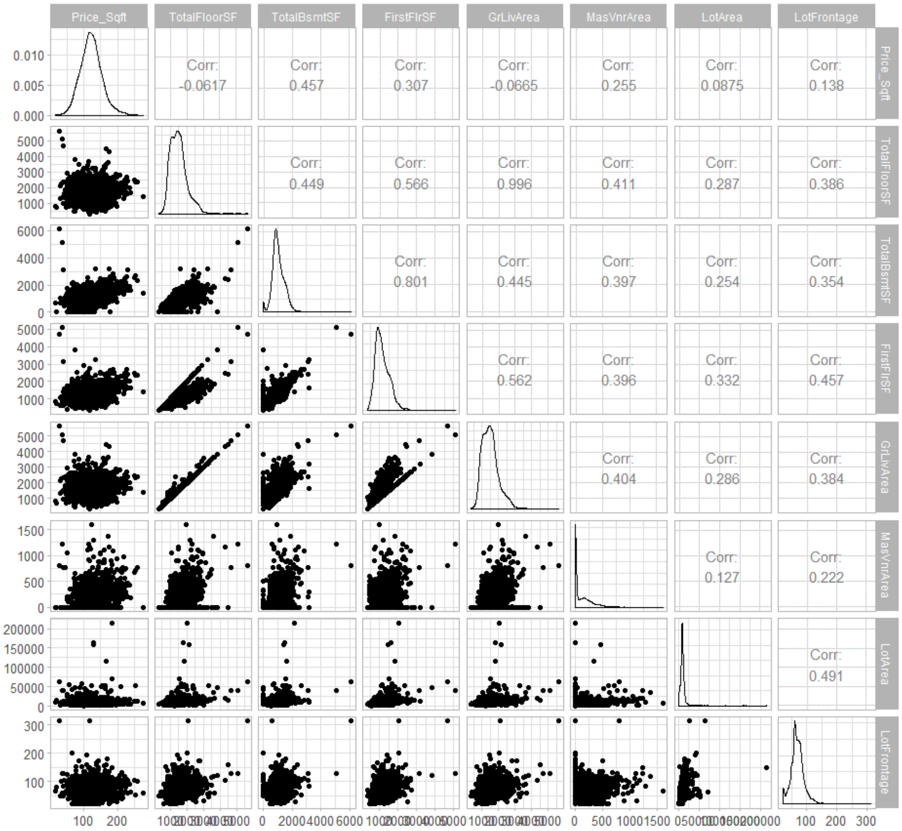


	Correlation to Sale Price
OverallQual	0.799261795
TotalFloorSF	0.713587857
GrLivArea	0.706779921
GarageCars	0.647876595
GarageArea	0.640400767
TotalBsmtSF	0.632280457
FirstFlrSF	0.621676063
Price_Sqft	0.613203774
QualityIndex	0.560846632
YearBuilt	0.558426106
FullBath	0.545603901
YearRemodel	0.532973754
GarageYrBlt	0.526965349
MasVnrArea	0.508284844
TotRmsAbvGrd	0.495474417
Fireplaces	0.474558093
BsmtFinSF1	0.432914411
LotFrontage	0.357317910
WoodDeckSF	0.327143174
OpenPorchSF	0.312950506
HalfBath	0.285056032
BsmtFullBath	0.276049952
SecondFlrSF	0.269373357
LotArea	0.266549220
BsmtUnfSF	0.182855260
BedroomAbvGr	0.143913428
ScreenPorch	0.112151214
PoolArea	0.068403247
MoSold	0.035258842
ThreeSsnPorch	0.032224649
BsmtFinSF2	0.005891398
MiscVal	-0.015691463
YrSold	-0.030569087
SID	-0.031407925
BsmtHalfBath	-0.035835410
LowQualFinSF	-0.037659765
SubClass	-0.085091576
OverallCond	-0.101696932
KitchenAbvGr	-0.119813720
EnclosedPorch	-0.128787442
PID	-0.246521213
HouseAge	-0.558906832

High Impact



Continuous


Price_Sqft
TotalFloorSF
TotalBsmtSF
FirstFlrSF
GrLivArea
MasVnrArea
LotArea
LotFrontage
Corr: -0.0617
Corr: 0.457
Corr: 0.307
Corr: -0.0665
Corr: 0.255
Corr: 0.0875
Corr: 0.138
Corr: 0.449
Corr: 0.566
Corr: 0.996
Corr: 0.411
Corr: 0.287
Corr: 0.386
Corr: 0.801
Corr: 0.445
Corr: 0.397
Corr: 0.254
Corr: 0.354
Corr: 0.562
Corr: 0.396
Corr: 0.332
Corr: 0.457
Corr: 0.404
Corr: 0.286
Corr: 0.384
Corr: 0.127
Corr: 0.222
Corr: 0.491