Julia Rodd Assignment 2

Contents

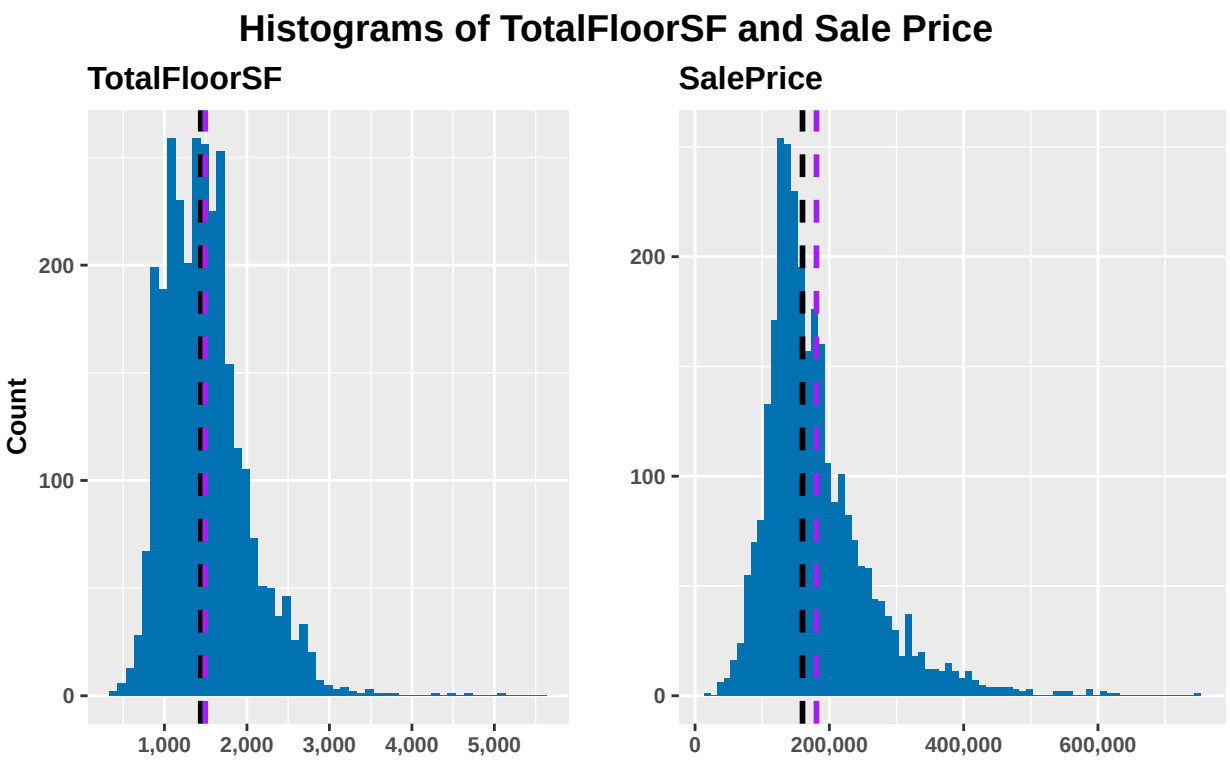
Introduction

This assignment uses housing data from Ames, IA from 2006 - 2010 and expands upon the analyses drawn from Assignment 1. The overall goal of this assignment is to select two variables to predict the sale price of a home and to incorporate these variables into a regression model. Both simple linear regression and multiple linear regression models are created using a subset of the Ames data set. Ultimately, commentary is provided on significance and goodness fit of our various models with conclusions made on model selection and implications. The R ggplot and dplyr packages are the two primary packages used in this assignment.

Results

Section 1: Sample Definition

In order to select the two most promising predictors, it is important to first understand the distribution of our data. This information will help to validate that we are selecting the appropriate homes (observations) to predict SalePrice.

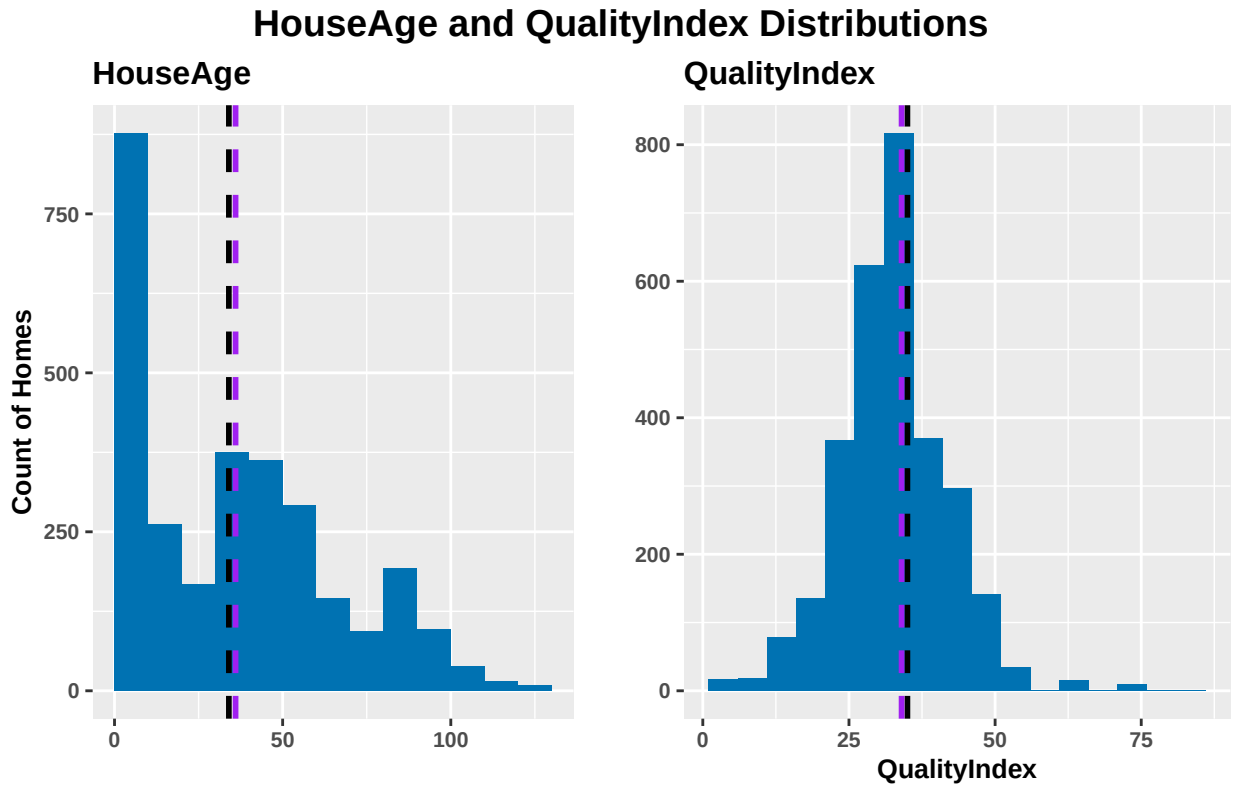


NOTE: Median values are displayed as the dashed black line and Mean values are displayed as the purple dashed line

From the histograms above, we can see that there is a wide range of both TotalFloorSF and SalePrice. There are also a couple homes in the right tails of both distributions that seem to fall outside the 'typical' distribution.

In calculating the IQR for TotalFloorSF, we can use this metric to identify if we have any extreme outliers. Overall, we have 8 extreme outliers (TotalFloorSF >= 3600). Similarly, using the IQR of SalePrice, there are 26 houses that are extreme outliers (SalePrice >= 465500). Only three of these homes with a SalePrice >= 465500 also have a TotalFloorSF >= 3600. We can see that these upper saleprices are pulling the mean (purple line) away from the median (black line).

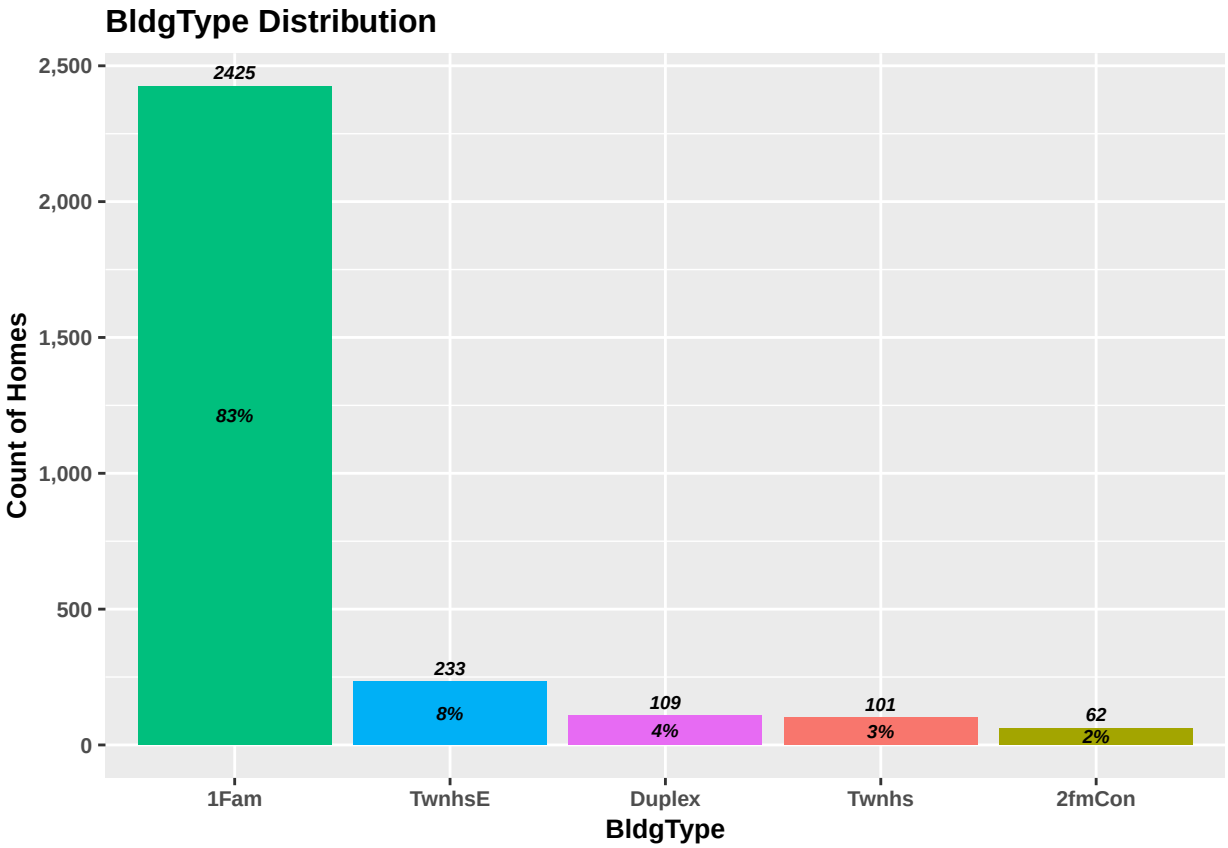
We will keep the outlier information into consideration. Now, we will examine the distribution of homes by HouseAge and QualityIndex.



From the histogram of HouseAge, we can see that the majority of the homes are between 0-10 years of age. Furthermore, we can see that there is a wide range of house ages, but that the mean and median values are close together. There are 0 homes that are considered extreme outliers.

Examining QualityIndex, we see similar trends from other histograms. There is a wide range of quality index, and the mean/median values are very close together and even in the center of the distribution. There are 11 homes that are considered extreme outliers (QualityIndex >= 70).

We continue our analysis by examining the distribution of homes by BldgType.



From the bar chart of BldgType, we can readily see that the majority of the homes in our data set (83%) are single family houses. This observation is a key point and will drive how we form our sample data set, as the characteristics of a home, including their sale price, vary by the type of building.

From this bar chart, we can make two conclusions:

1. BldgType will not be a good predictor in a regression model, since the majority of the data falls into one category, single family home (1Fam); and,
2. A 'typical' Ames house is a single family home

It is unknown how representative this data set is of the building types actually present in Ames. However, we can still move forward with our analysis without this information.

Drop Conditions

Because our primary goal is to create a regression model with SalePrice as our response variable, we need to be cognizant of outliers and the types of homes in our data set.

We will refine our goal as creating a regression model for **typical** Ames houses. We define a typical Ames house as:

1. A single family home (BldgType == "1Fam");
2. Houses with a total square feet < 3,600 (TotalFloorSF < 3600); and,
3. Houses a sale price < 465,500 (SalePrice < 465500)

In defining our data this way, we remove 536 homes from our data set and are left with 2394 homes to model. Even though we are removing a large portion of homes from our data set, this sample data set will be easier to model, as we will not have additional variability by trying to model different building types.

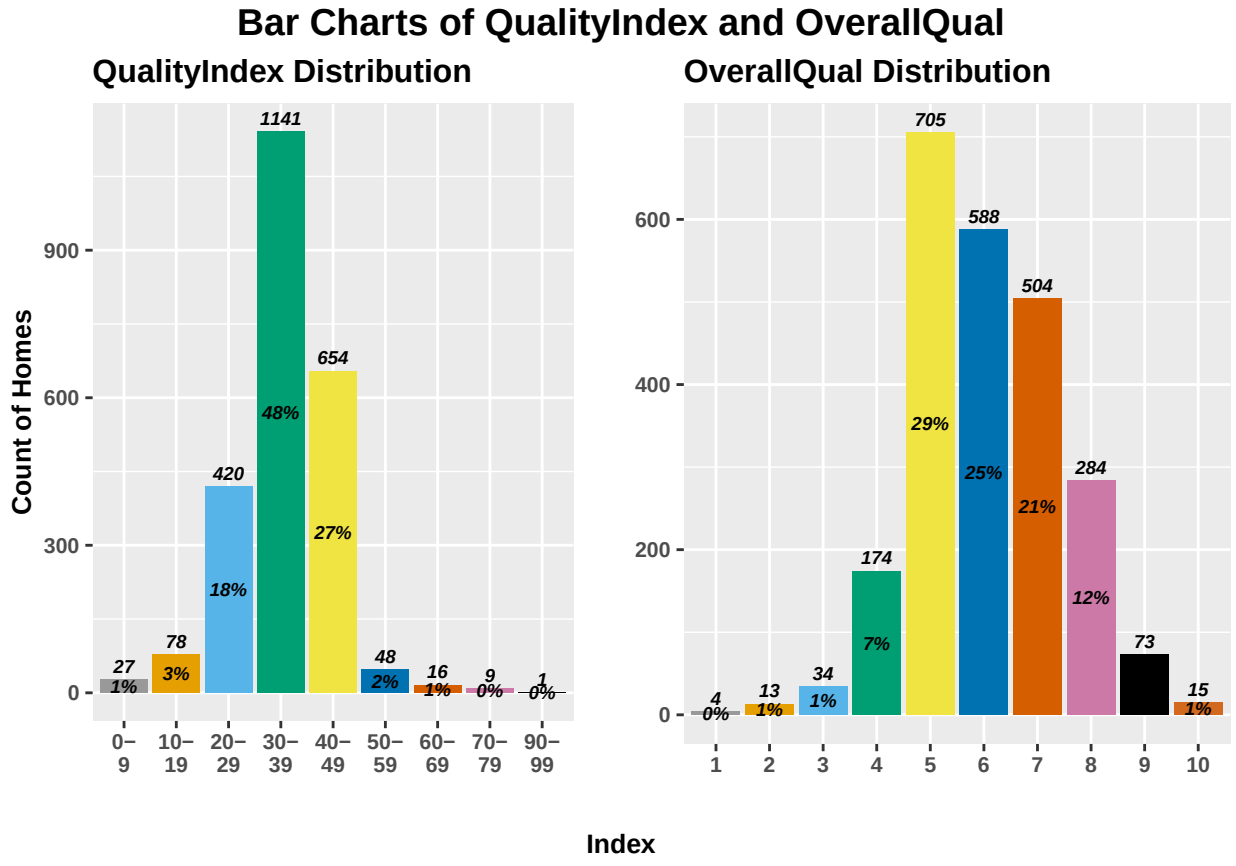
For now, we do not elect to refine our data set by QualityIndex. However, this may be one criterion to revisit to potentially improve model fit.

Section 2: Exploratory Data Analysis

We will now explore our sample data set and identify two promising predictors to include in our regression model.

It is important to note that we will be creating both simple linear regression and multiple linear regression models. Because of this, we will need to focus our exploratory data analysis on continuous predictors. Where possible, we will incorporate some analyses with categorical variables, since analyzing categorical variables will help to inform future considerations.

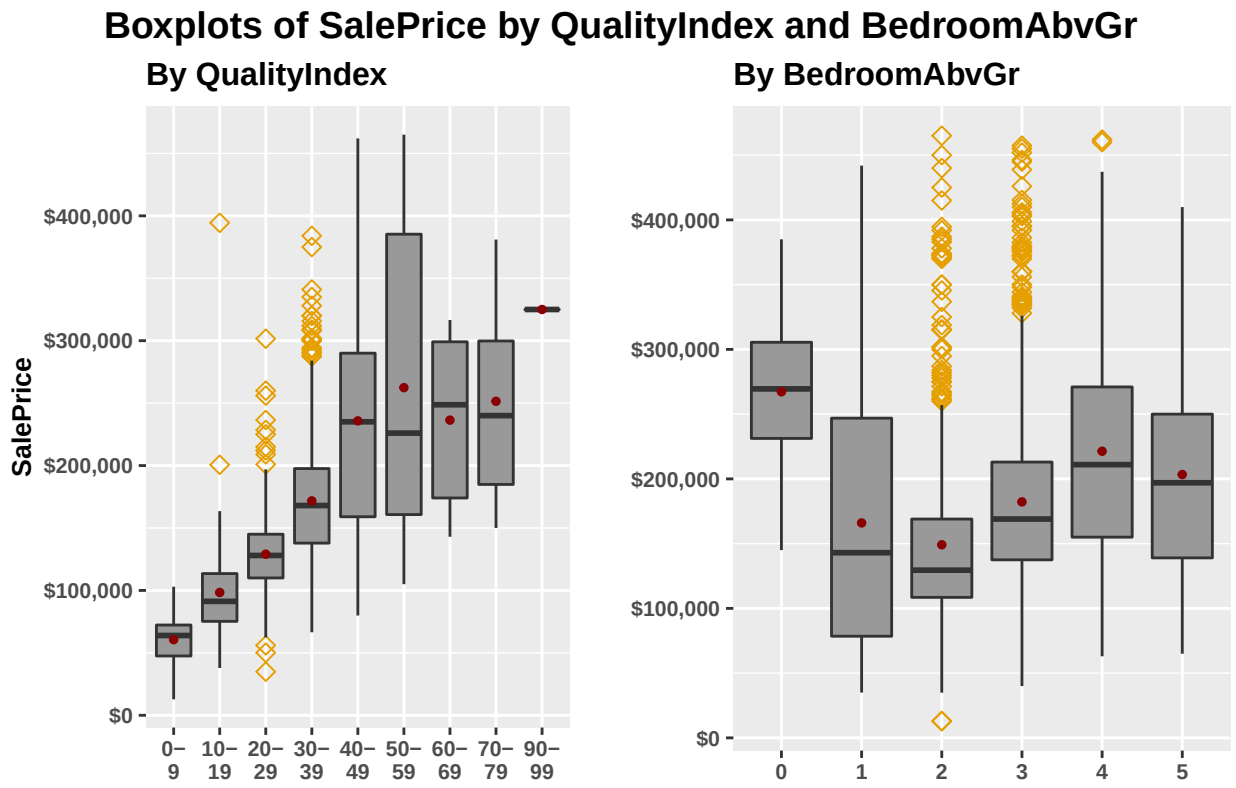
We start our analysis by examining the distribution of two of our quality indices, QualityIndex (OverallQual * OverallCond) and OverallQual side by side. Our hypothesis is that the price of a home varies by its quality, which is why we are spending time exploring these variables.



In comparing these two bar charts, we can see that there is a more even distribution of homes by OverallQual than by QualityIndex. However, we can validate that the relationship between OverallQual and OverallCond is what we would expect it to be. For instance, most homes have a OverallQual in the 5-8 range, and if ther relationship with OverallCond were 1-to-1, then we should have more homes in the 20-49 range. In examining the bar chart of QualityIndex, we do in fact see most homes in this range.

Simple linear regression models require a continuous predictor variable, so we will need to move forward with QualityIndex, as OverallQual is an ordinal categorical variable. We will seek to understand how SalePrice varies by QualityIndex to determine if it is a good potential predictor for our model.

Now, we will examine the relationships between SalePrice and two other variables: QualityIndex (as a factor) and BedroomAbvGr.



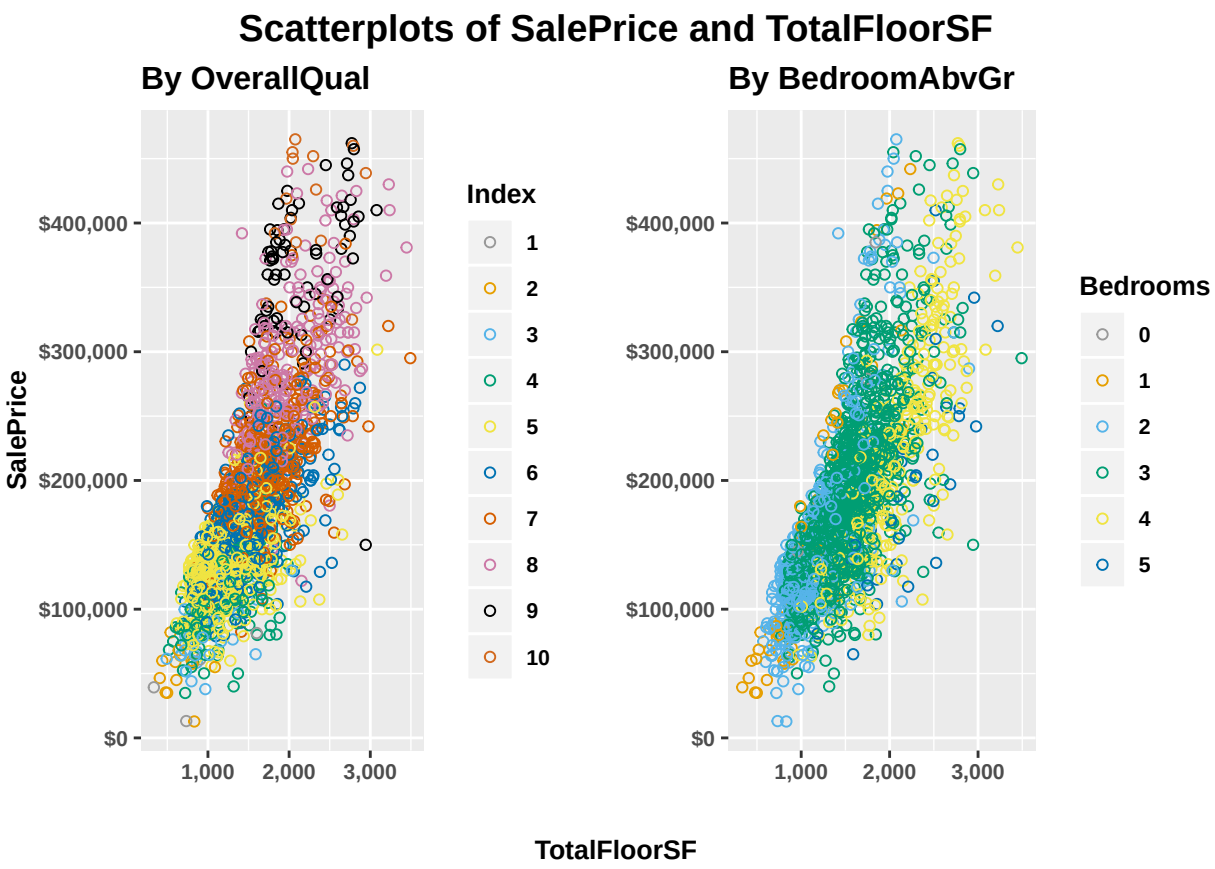
***Mean values are denoted by the darkred dot**

From these boxplots, we can readily see that when examining SalePrice by QualityIndex there appears to be minimal overlap between each of our indices and different median values. While there is variation in SalePrice by BedroomAbvGr, we see differences between 0 bedrooms, 1-2 bedrooms, and 3+ bedrooms. We might consider having less categories if BedroomAbvGr were to be included in our model. Furthermore, surprisingly, homes with 0 bedrooms have higher saleprices. There has to be other characteristics that would contribute to this trend, outside of number of bedrooms. The trends by QualityIndex make logical sense: as quality increases, then so does sale price.

Comparing these two boxplots, QualityIndex seems to be the better potential predictor to include in our regression model because of the differences by each of these quality indices or ratings. It is also important to callout that both variables have outliers in each of their categories.

We continue our analysis by exploring the relationships between SalePrice and the following potential continuous predictors: TotalFloorSF, QualityIndex, and HouseAge. From this analysis, we will choose two predictors. We will also incorporate how these relationships vary by OverallQual and BedroomAbvGr, even though categorical variables will not be included in our model.

We begin by examining scatterplots of TotalFloorSF and SalePrice.



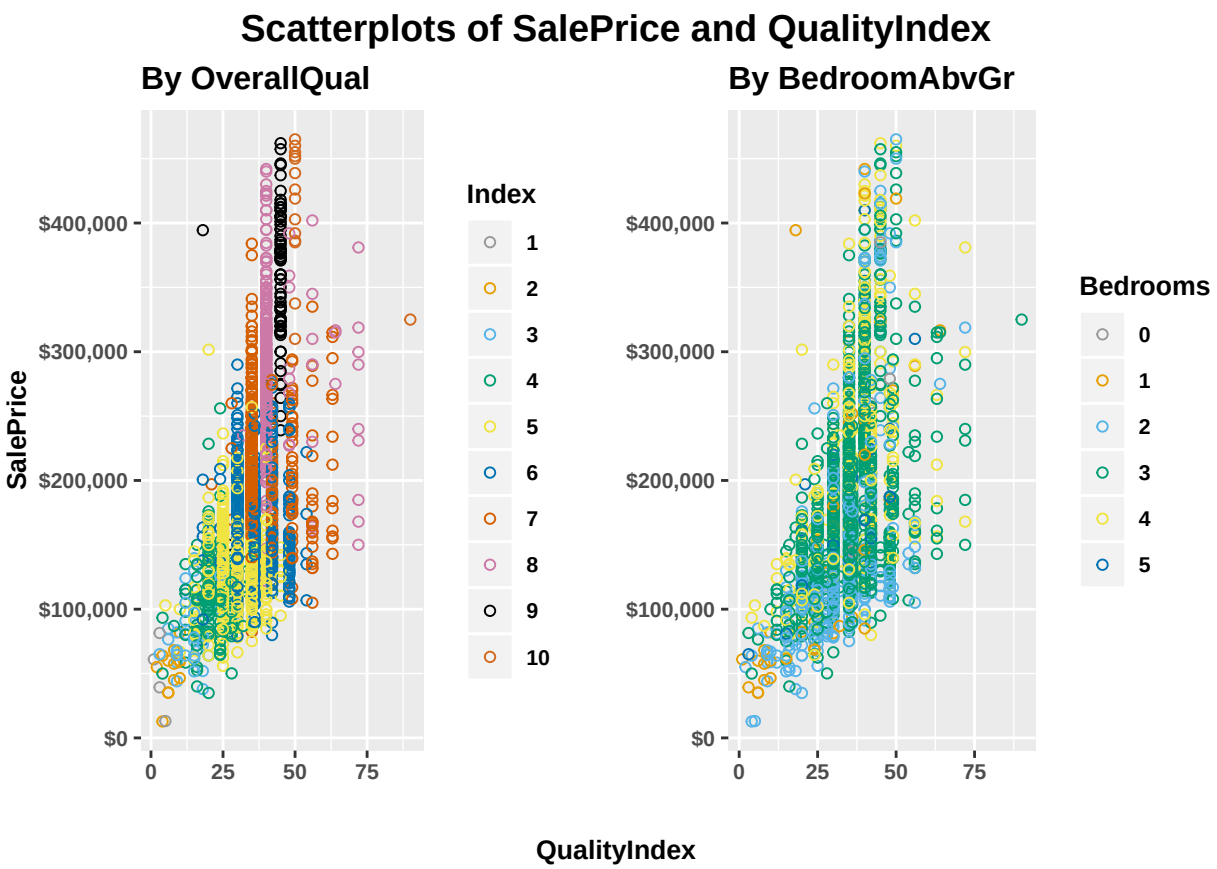
There are several key points we can make from these two scatterplots.

First, we can better understand the relationship between TotalFloorSF and SalePrice. First, we can examine that there is a positive linear relationship with TotalFloorSF and SalePrice. In fact, the correlation between TotalFloorSF and SalePrice is 76%, which is very strong. We note that the removal of our extreme outliers has improved this relationship by removing the variability present in the upper ranges of both of these variables, resulting in a higher correlation. The addition of the regression lines to our scatterplots helps to see this pattern.

Although there is a strong positive linear relationship, we gain much more context from viewing these scatterplots. For one, we can see that this relationship follows a wedge-shaped pattern, also known as heteroscedasticity. We can see that there is increased variabiility in the upper ranges of both SalePrice and TotalFloorSF. This trend means that we will need to transform SalePrice, as heteroscedasticity is a violation of a linear regression model. Additionally, the presence of heteroscedasticity means that TotalFloorSF on its own does not fully describe all of the variability in SalePrice. We would not have obtained this information if we were to solely utilize correlation in determining the 'best' potential predictor.

Second, we can make some observations regarding the interactions with our two categorical variables, OverallQual and BedroomAbvGr. We see clear layering of TotalFloorSF and SalePrice by OverallQual. We still see differences in number of bedrooms, as 4- and 5- bedroom homes are on the right side of the plot.

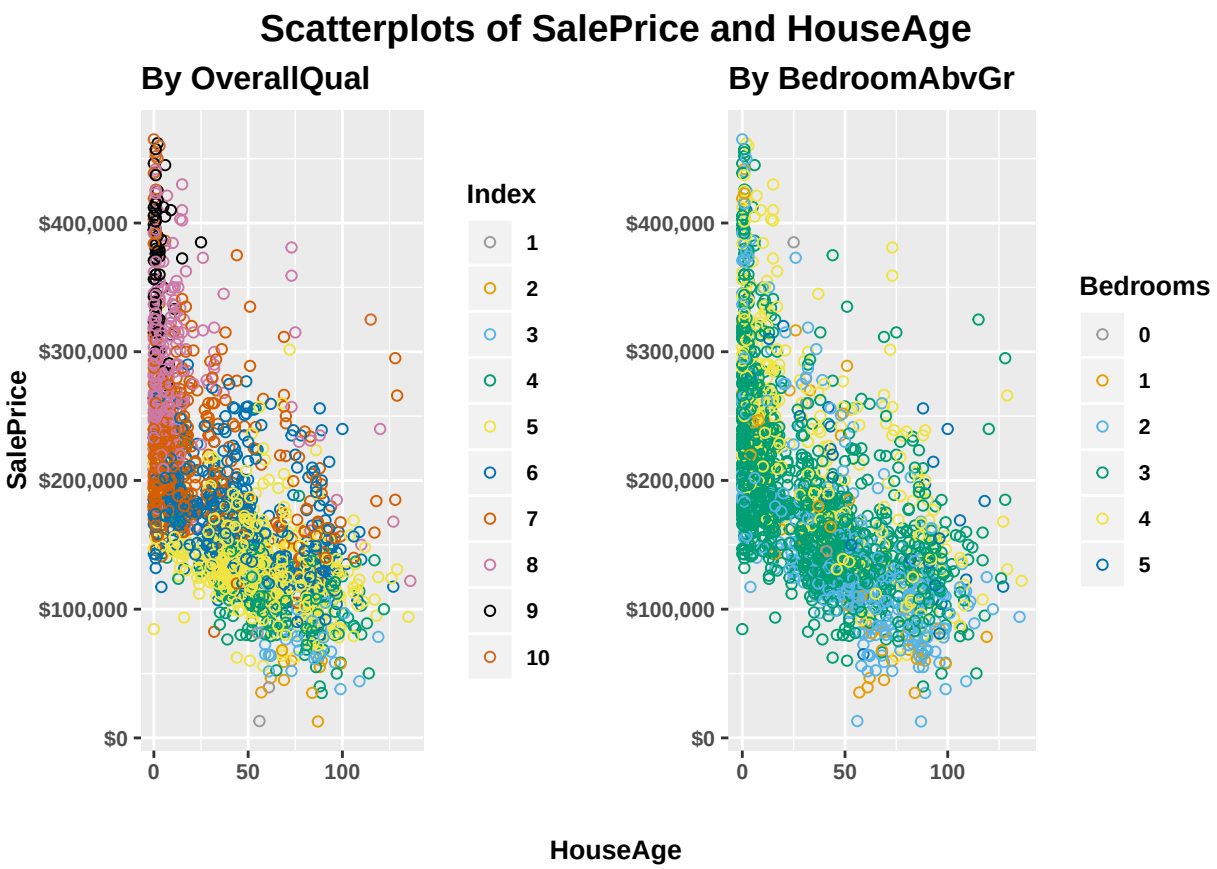
We repeat the exercise above by examining scatterplots of SalePrice and QualityIndex.



We can see that the relationship between SalePrice and QualityIndex is a positive one. For one, the correlation between these two variables is 54%, which is just okay. We can examine that part of why this correlation is just okay is there appears to be increased variability in the 25-50 quality index range. The addition of the regression line helps to demonstrate that the line does not go through the middle of the data cloud.

Futhermore, we can state that we once again see layering by OverallQual. The relationship between QualityIndex and SalePrice does not seem captured by number of bedrooms, as the coloring does not follow a distinct pattern.

Finally, we analyze scatterplots between SalePrice and HouseAge.



From these scatterplots, we can see that HouseAge and SalePrice have a negative relationship. That is, as HouseAge increases then SalePrice decreases. In fact, the correlation between these two variables is -62%. The correlation is slightly improved since we reduced the scope of our data set.

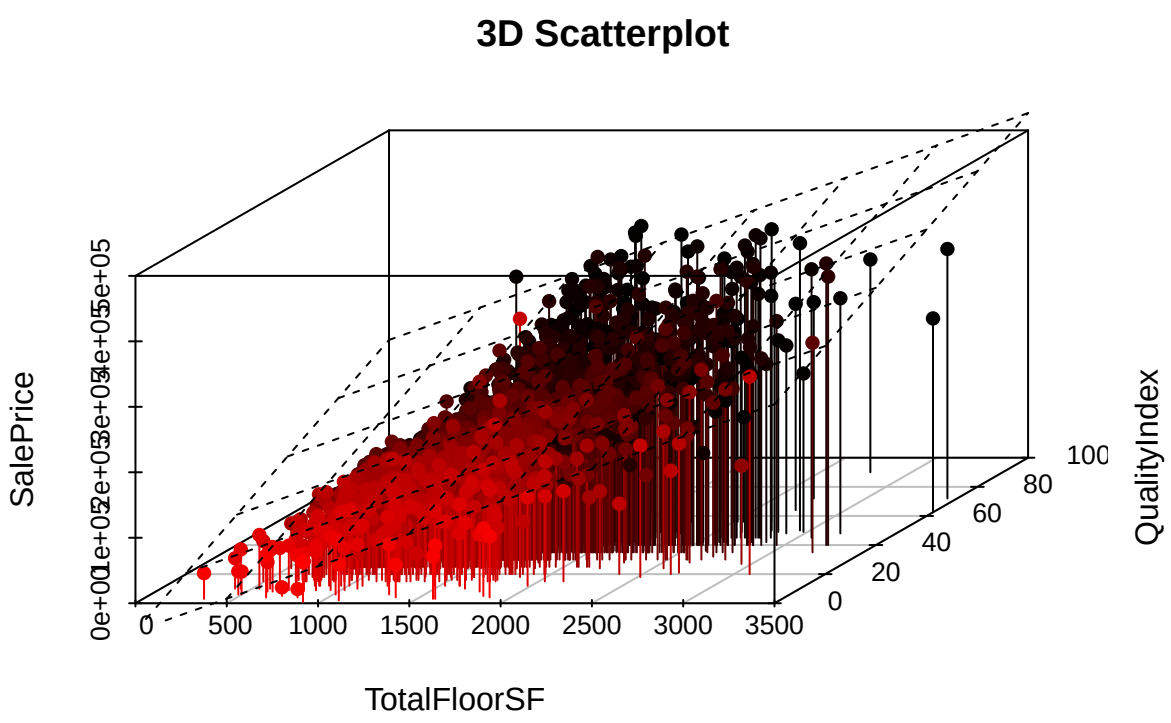
Moreover, we continue to see the trend of clear layering by OverallQual with some patterns by BedroomAbvGr. The plot on the right shows a higher concentration of blues under the regression line and a higher concetration of pinks and blacks above the regression line.

Selection of Two Predictors

From all of our exploratory work, we will move forward with TotalFloorSF and QualityIndex as our two predictors.

There are several reasons why we select QualityIndex over HouseAge. First, from the boxplot of OverallQual (as a factor) and SalePrice, we saw that SalePrice did vary by OverallQual as the ranges across indices showed distinct ranges of SalePrice. Additionally, seeing the trends by OverallQual in the scatterplots above really showed that SalePrice and TotalFloorSF do vary by quality. From Assignment 1 we saw a stronger linear relationship between TotalFloorSF and QualityIndex. Assignment 1 showed essentially no linear relationship between TotalFloorSF and HouseAge. Even though HouseAge has a slighter higher correlation with SalePrice, we are looking for relationships that make logical sense.

We can further validate that these relationships make sense by examining a 3D scatterplot of SalePrice, TotalFloorSF, and QualityIndex.



We can see from the 3D scatterplot that higher sale prices are associated with larger square feets and a higher quality index. We are now ready to move forward with our regression models.

Section 3: Simple Linear Regression Model

In this section, we will create two simple linear regression models from our two predictors of TotalFloorSF and QualityIndex. Commentary is provided on the significance and goodness of fit of each model.

Model 1: TotalFloorSF

We start with a linear regression model of TotalFloorSF as the predictor.

First, we answer the question 'Is my model significant?' From the ANOVA results below, we can see that our p-value is low, which implies that our model is significant. However, we need to keep exploring further to assess how good this model actually is, as significance does not always imply that we have a 'good' model.

```
##                Df    Sum Sq   Mean Sq F value Pr(>F)
## TotalFloorSF    1 7.471e+12 7.471e+12    3296 <2e-16 ***
## Residuals    2392 5.422e+12 2.267e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the output of the coefficients below, we are able to write our model as: y = 6923.1 + 116.5x, where x = the number of total square feet.

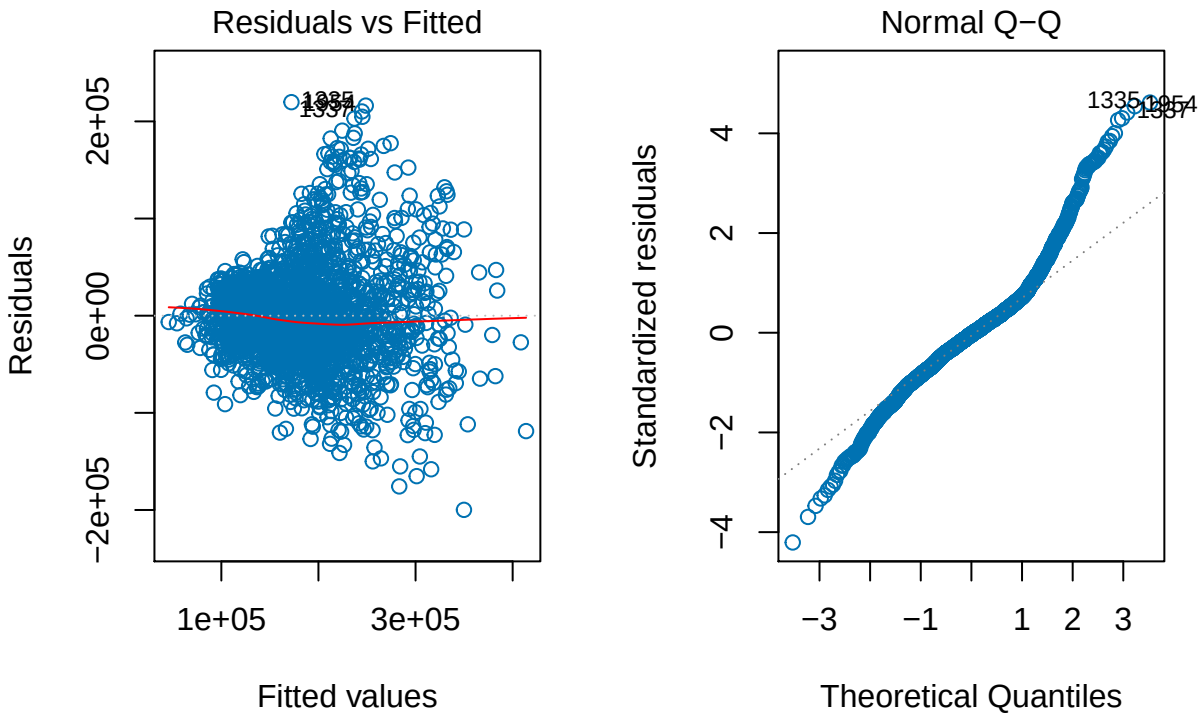
We interpret our model as for each additional 1 square foot, the average sale price of a home increases by 116.5 dollars.

```
##                Estimate  Std. Error   t value   Pr(>|t|)
## (Intercept)  6923.0978 3179.645641  2.177317 0.02955434
## TotalFloorSF  116.4864    2.028989 57.411041 0.00000000
```

We can also examine the residual standard error and R squared value to understand goodness of fit.

The residual standard error is 47608 dollars. This means that the average error for one standard deviation is 47608, which is a pretty large range. The R squared value is 58%, which means that 58% of SalePrice variation is explained by TotalFloorSF. Since we want our R squared value to be close to one, this result is okay but possibly could be better.

We now examine two key residual plots to determine if our assumptions of constant variance and normality hold true for this simple linear regression model.



Right away, we can see that our residuals have a pattern and are not randomly scattered above and below the line. Therefore, the assumption of constant variance is not met. Moreover, the residuals do not follow the normal QQ-line in the tails so we also do not meet the assumption of normality.

Overall, the simple linear regression model with TotalFloorSF does not meet two key assumptions for modeling: constant variance and normality, despite being a significant predictor for SalePrice. The R squared value and residual error are just okay as well, resulting in wider than desired confidence intervals for predicted SalePrice values.

Model 2: QualityIndex

We now turn our attention to a simple linear regression model with QualityIndex as the predictor. We will examine the significance and fit of this model in predicting SalePrice.

Starting with significance, from the ANOVA results below, we can see that our p-value is low, which tells us that our model is significant.

```
##                Df    Sum Sq   Mean Sq F value Pr(>F)
## QualityIndex    1 3.815e+12 3.815e+12    1005 <2e-16 ***
## Residuals    2392 9.077e+12 3.795e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the output of the coefficients below, we are able to write our model as: y = 33644 + 4327.5x, where x = the number of quality index.

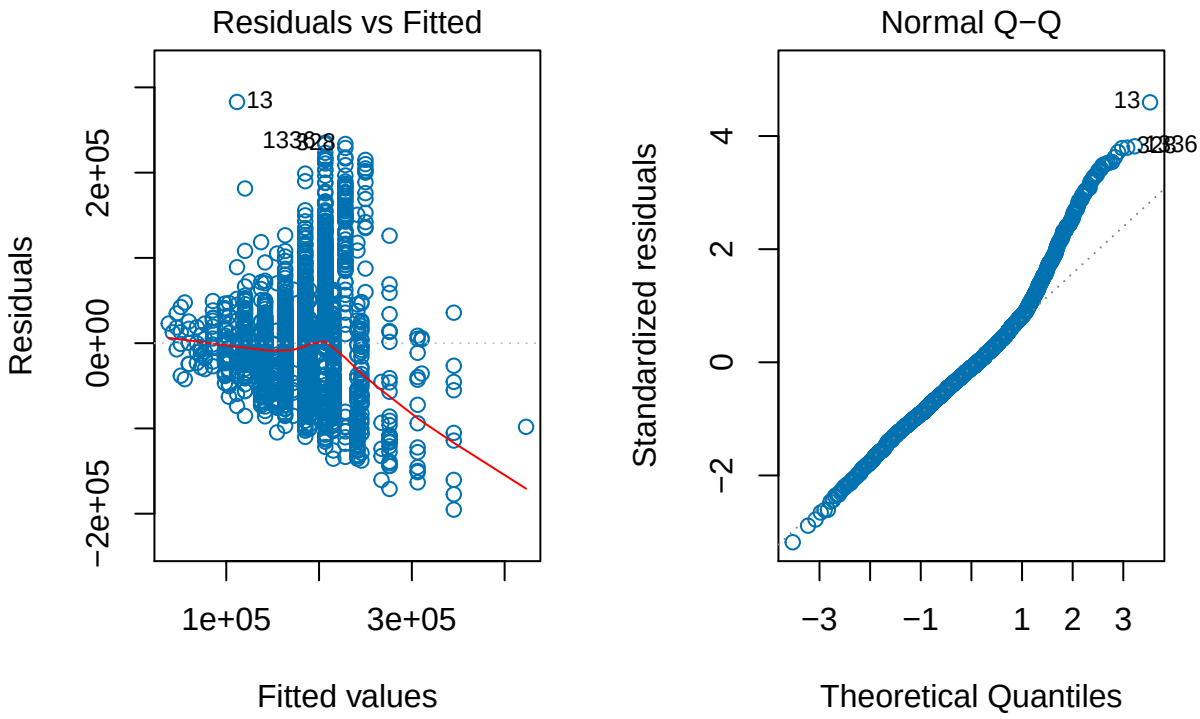
We interpret our model as for each additional increase in quality index, the average sale price of a home increases by 4327.5 dollars.

```
##               Estimate Std. Error   t value      Pr(>|t|)
## (Intercept)  33643.972  4806.2075  7.000108  3.309244e-12
## QualityIndex  4327.462   136.4831 31.706936 1.714607e-184
```

Next, we examine the residual standard error and R squared value to see how good our model actually is.

The residual standard error is 61602 dollars. This means that the average error for one standard deviation is 61602, which is a large range. The R squared value is 30%, which means that 30% of SalePrice variation is explained by QualityIndex, which is a very low amount of variation. This result is not desirable.

We continue our goodness of fit assessment by analyzing two key residual plots to determine if our assumptions of constant variance and normality hold true for this simple linear regression model.



Similar to our first simple linear regression model, we can see that our residuals do not meet the assumptions of constant variance and normality. The residuals are not randomly distributed, since they follow a wedge-shaped pattern and demonstrate banding. Especially on the right tail, the residuals depart from the QQ-line.

Overall, our simple linear regression model with QualityIndex as a predictor does not meet our assumptions of constant variance and normality. More importantly, even though this model is significant, it has poor goodness of fit metrics, making this model a poor choice, on its own, for predicting SalePrice.

Section 4: Multiple Linear Regression Model (Model 3)

We will now examine the significance and goodness and fit of a multiple linear regression model, which includes TotalFloorSF and QualityIndex as predictors.

From the ANOVA results below, we can see that our p-value is low, which implies that our model is significant.

```
##                Df    Sum Sq   Mean Sq F value Pr(>F)
## TotalFloorSF    1 7.471e+12 7.471e+12  3898.8 <2e-16 ***
## QualityIndex    1 8.401e+11 8.401e+11   438.4 <2e-16 ***
## Residuals    2391 4.581e+12 1.916e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the output of the coefficients below, we are able to write our model as: y = -42603.2 + 99x1 + 2224.8x2, where x1 = the number of total square feet and x2 = the number of quality index.

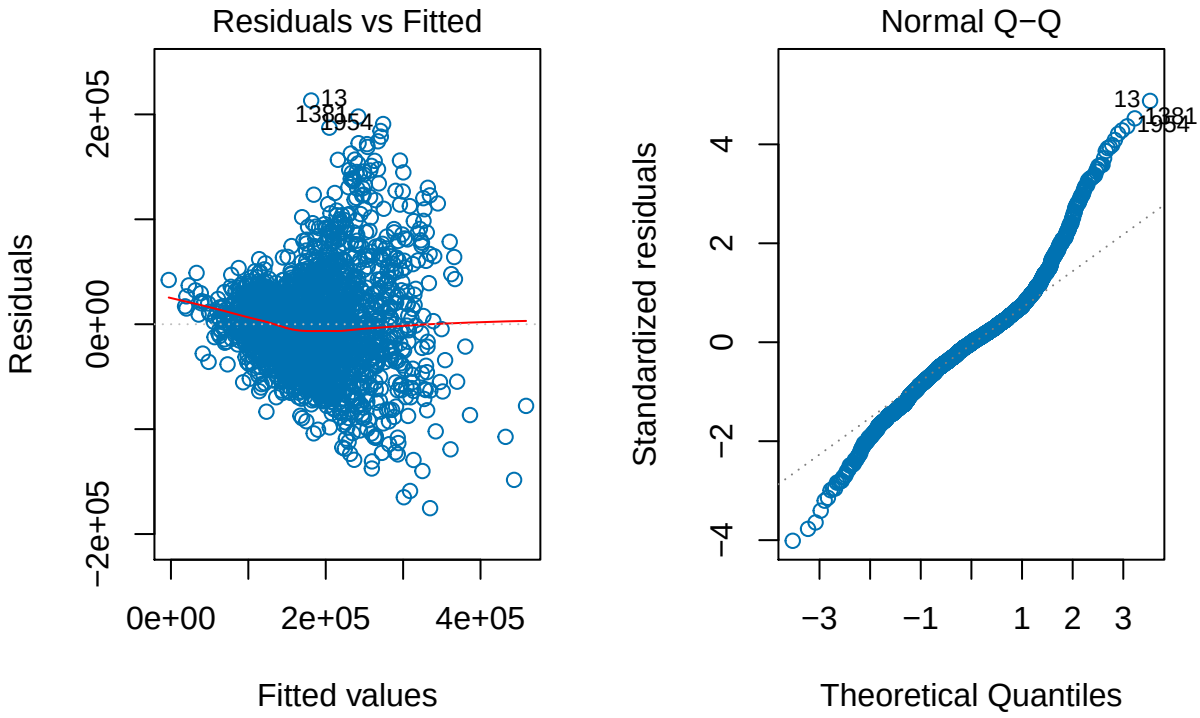
We interpret our model as for each additional 1 square foot, the average sale price of a home increases by 99 dollars when quality index is held constant. Similarly, for each additional increase in quality index, the average sale price of a home increases by 2224.8 dollars when total square feet is held constant.

```
##                   Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept)  -42603.17311 3760.555086 -11.32896 5.107204e-29
## TotalFloorSF     99.00242    2.043913  48.43769 0.000000e+00
## QualityIndex   2224.83987  106.255063  20.93867 1.594694e-89
```

We can also examine the residual standard error and R squared value to assess the goodness of fit of this multiple linear regression model.

The residual standard error is 43774 dollars. This means that the average error for one standard deviation is 43774, which is a much narrower range than each of our simple linear regression models. The R squared value is 64%, which means that 64% of SalePrice variation is explained by TotalFloorSF and QualityIndex. This R squared value is fairly good, as we want our R squared value close to 1.

We now examine two key residual plots to determine if our assumptions of constant variance and normality hold true for this model.



We see that our multiple linear regression model does not meet the assumptions of constant variance and normality. The plot of residuals vs fitted values follows a wedge-shaped pattern. The QQ-plot shows the residuals depart from the QQ-line in the tails, especially the right tail.

In summary, the multiple linear regression model is significant and improves on goodness of fit metrics compared to the simple linear regression models. However, there are still concerns as this model violates the core assumptions of linear regression.

Section 5: LogSalePrice Response Models

In this section, we re-evaluate our three models using logSalePrice instead of SalePrice as our response variable. We perform this step to consider this transformed variable, as based on our EDA work we saw

patterns of heteroscedasticity.

It is important to callout that the log transformation of SalePrice improves its skewness. Before the transformation, it was 1.13 and after the log transformation, it is -0.28. Given the packages used in this assignment, we want the skewness and kurtosis values to be 0. However, the kurtosis value is not improved. It was 1.38 before the transformation and is 1.67 after the log transformation. This means that we are dealing with a heavy tailed distribution, which provides some context to the results that will be shown below.

For each model shown, we provide commentary on model significance and goodness of fit. We will also draw comparisons to the equivalent models where SalePrice was the response variable.

Model 4: Simple Linear Regression - TotalFloorSF

We start with examining model significance. From the ANOVA results below, we can see that our p-value is low, which tells us that our model is significant.

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## TotalFloorSF    1  215.2  215.23    3103 <2e-16 ***
## Residuals    2392  165.9    0.07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the output of the coefficients below, we are able to write our model as: y = 11.094 + 0.001x, where x = the number of total square feet.

We interpret our model as each additional 1 square foot results in a 0.063% percentage change in the sale price of a home.

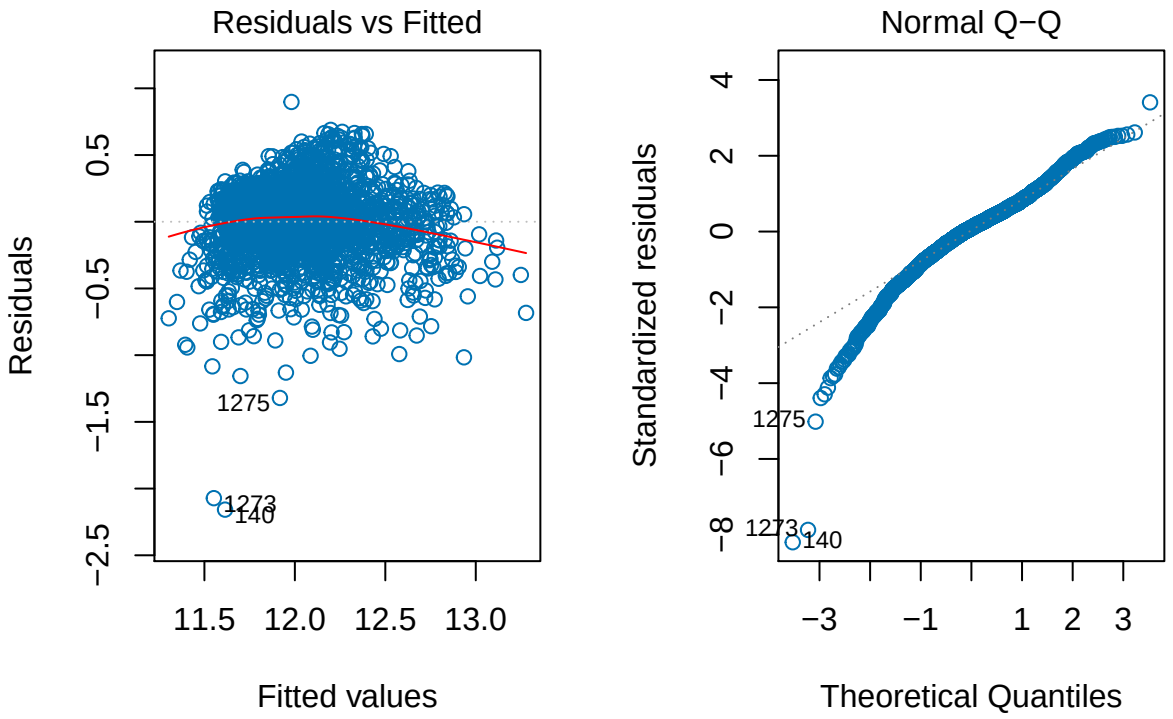
```
##                  Estimate   Std. Error   t value Pr(>|t|)
## (Intercept)  1.109396e+01 1.758987e-02 630.70165        0
## TotalFloorSF 6.252486e-04 1.122441e-05  55.70435        0
```

We can also examine the residual standard error and R squared value to evaluate goodness of fit.

The residual standard error is 0. This means that the average error for one standard deviation is 0, which does not reveal a lot of meaningful information as the log transformation of SalePrice is resulting in this narrower range. The R squared value is 56%, which means that 56% of logSalePrice variation is explained by TotalFloorSF.

This result is equivalent to Model 1. If we compare the adjusted R squared values, Model 1 with SalePrice has an adjusted R squared value of 58% and Model 4 with logSalePrice has an adjusted R squared value of 58%.

We now examine two key residual plots to determine if our assumptions of constant variance and normality hold true for this simple linear regression model.



We can see improvement in the assumptions of constant variance and normality with logSalePrice. The plot of residuals vs fitted values is more random and the residuals more closely follow the QQ-line. The residuals do however depart from the QQ-line in the lower tail now.

Although not perfect, Model 4 with logSalePrice as the response variable and TotalFloorSF has a much better model fit than Model 1.

Model 5: Simple Linear Regression - QualityIndex

From the ANOVA results below, we can see that our p-value is low, which tells us that our model is significant. Next, we interpret this model.

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## QualityIndex    1  131.7   131.7    1263 <2e-16 ***
## Residuals    2392  249.4     0.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the output of the coefficients below, we are able to write our model as: y = 11.163 + 0.025x, where x = the number of quality index.

We interpret our model as each additional increase in quality index results in a 2.543% percentage change in the sale price of a home.

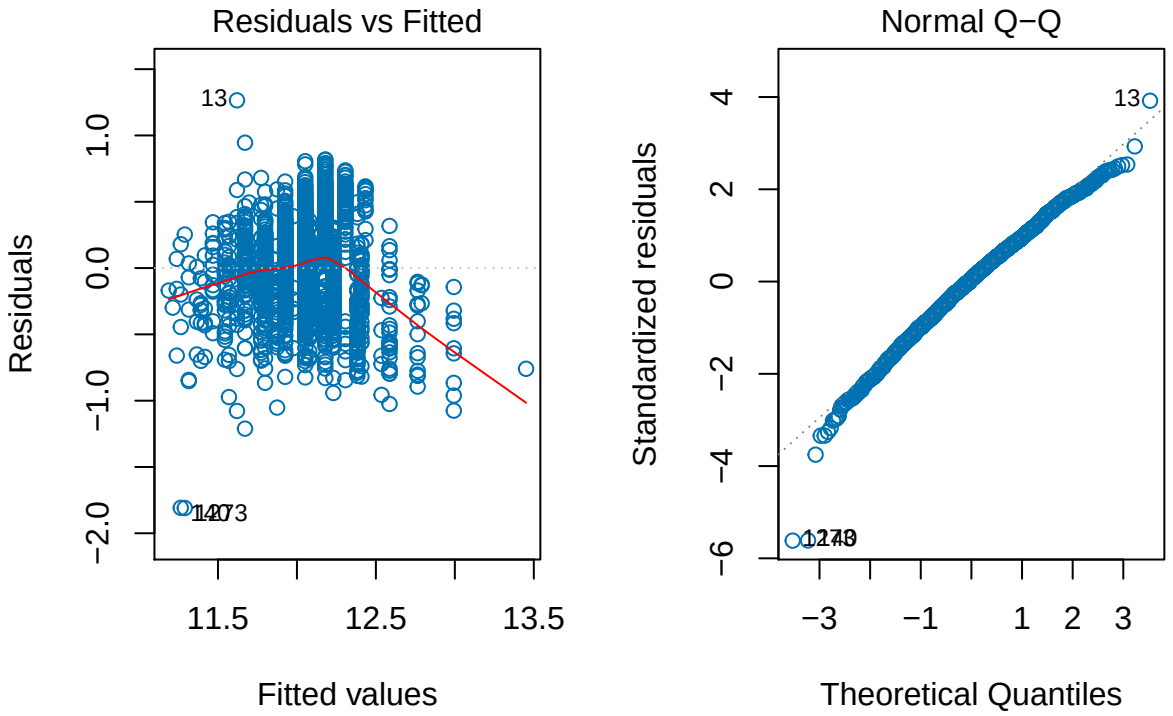
```
##                  Estimate   Std. Error  t value      Pr(>|t|)
## (Intercept)  11.16265880 0.0251949369 443.0517  0.000000e+00
## QualityIndex  0.02542684 0.0007154672  35.5388 1.686939e-222
```

We can also examine the residual standard error and R squared value for goodness of fit.

The residual standard error is 0. This means that the average error for one standard deviation is 0, which once again has been impacted by the log transformation of SalePrice. The R squared value is 35%, which means that 35% of logSalePrice variation is explained by QualityIndex. Since we want our R squared value to be close to one, this result is okay but possibly could be better.

This result is slightly better than Model 2. If we compare the adjusted R squared values, Model 2 with SalePrice has an adjusted R squared value of 30% and Model 4 with logSalePrice has an adjusted R squared value of 35%. However, both R squared values are subpar.

We now examine two key residual plots to determine if our assumptions of constant variance and normality hold true for this simple linear regression model.



Similar to Model 4, the assumptions of constant variance and normality are improved using logSalePrice. However, we can see some outlier values that are potentially driving depatures from our normal lines. There is variation in the residual vs fitted values plot in the lower values and a departure of the residuals from the QQ-line in both tails.

In short, Model 5 has a low R squared value but using logSalePrice improves upon this model meeting linear model assumptions.

Model 6: Multiple Linear Regression - TotalFloorSF and QualityIndex

Now we will combine our two predictors to form a multiple linear regression model.

We once again begin with determining model significance. We can see that our p-value is low, which implies that our model is significant.

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## TotalFloorSF    1 215.23  215.23  3963.7 <2e-16 ***
## QualityIndex    1  36.08   36.08   664.5 <2e-16 ***
## Residuals    2391 129.83    0.05
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the output of the coefficients below, we are able to write our model as: y = 10.769 + 0.001x1 + 0.015x2, where x1 = the number of total square feet and x2 = the number of quality index.

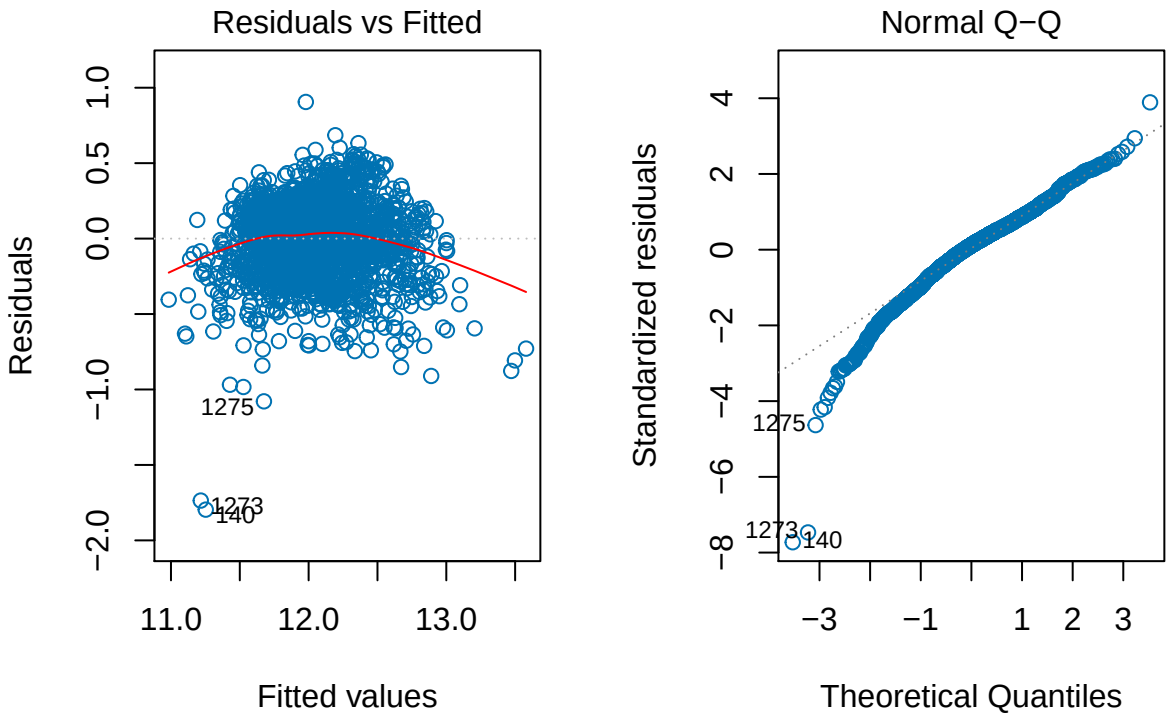
We interpret our model as each additional 1 square foot results in a 0.051% percentage change in the sale price of a home when quality index is held constant. This slight difference from the coefficient above is due to rounding and multiplication. Similarly, each additional increase in quality index results in a 1.458% percentage change in the sale price of a home when total square feet is held constant.

```
##                   Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept)  1.076937e+01 2.001896e-02 537.95871  0.00000e+00
## TotalFloorSF 5.106605e-04 1.088057e-05  46.93322  0.00000e+00
## QualityIndex 1.458139e-02 5.656387e-04  25.77864 1.63224e-129
```

Next, we examine the residual standard error and R squared value to assess goodness of fit.

The residual standard error is 0. This means that the average error for one standard deviation is 0, which is does not reveal too much information as the transformation of SalePrice is impacting this range. The R squared value is 66%, which means that 66% of logSalePrice variation is explained by TotalFloorSF and QualityIndex. Since we want our R squared value to be close to one, this result is good.

This result is slightly better than Model 3. If we compare the adjusted R squared values, Model 3 with SalePrice has an adjusted R squared value of 64% and Model 6 with logSalePrice has an adjusted R squared value of 66%. Both adjusted R squared values indicate that the models are good, but examining the residual plots to determine if our assumptions of constant variance and normality hold true will be key to see which model actually performs better.



Similar to the trends above, we can see that the residual plots are much improved using logSalePrice. The residuals vs fitted values have a more constant variance, despite showing some variability in the left of the plot. Additionally, the residuals follow the normal QQ-line except in the lower left tail.

All in all, we can conclude that this model has a better fit than Model 4.

Summary and Conclusions

In this assignment, we utilized a subset of the Ames dataset to explore and identify two promising predictor variables to utilize in various linear regression models.

From our exploratory data analysis work, we defined typical Ames houses as single family homes that have a total square feet < 3600 and a sale price < 465500. While this definition resulted in the removal of 536 homes from our data set, it did help to also remove inheret variability by building type. Our exploration led us to focus in on TotalFloorSF and QualityIndex as our two predictors for modeling.

Overall, we saw that multiple linear regression models outperformed our simple linear regression models. Furthermore, we saw that models with logSalePrice instead of SalePrice had better conformity to the linear regression assumptions of constant variance and normality.

Even though we were able to successfully create a model (Model 6) that explained a large portion of variability in logSalePrice and largely met linear regression assumptions, there are still several issues with this model that are beyond the scope of this assignment.

First, no conclusions can be made on if Model 6 is truly the 'best' model. No comparisons were made across models as only comparisons were made on if the predictors were in the model were each significant. Because this level of hypothesis testing was not performed, conclusions cannot be drawn on the best predictors of

SalePrice. In addition, Model 6 only had a R squared value 66%. Ideally, we would like to find a model that has a higher R squared value.

Second, in examining the residual plots, there are some concerns. Although we saw that the logSalePrice model residuals conformed more closely to the assumptions of constant variance and normality, there appear to be some outliers or influential points that are preventing these models from fully meeting these assumptions. These outliers would have to be understood better and possibly removed before additional conclusions could be drawn. We also saw that logSalePrice kurtosis value alluded to a heavy tailed distribution but the skewness value was close to 0. This result is concerning, especially when trying to model sale price behavior.

Furthermore, the multiple linear regression models do not include any categorical variables. The Ames data set is primarily comprised of categorical variables. From Assignment 2, we saw some trends by OverallQual and BedroomAbvGr and in assignment 1 we saw that SalePrice varied by Neighborhood. Not including any categorical variables in a regression model or at least comparing models with and without categorical variables is a needed step. Without this level of analysis, we are not incorporating any contextual information that could potentially improve the fit of our model.

All in all, this assignment makes some good progress on determining potential predictors of SalePrice. However, it falls short as there are outliers that still need to be addressed and further model comparisons that need to be made before the appropriate model can be selected to predict sale price.