Michael Venit

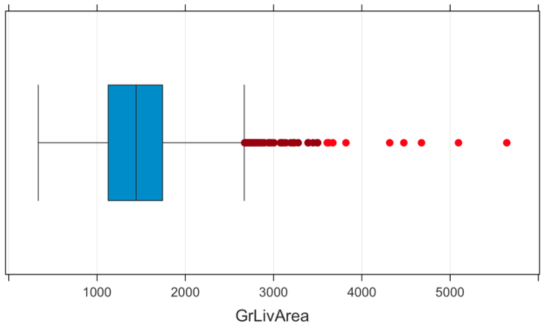
MSDS 410

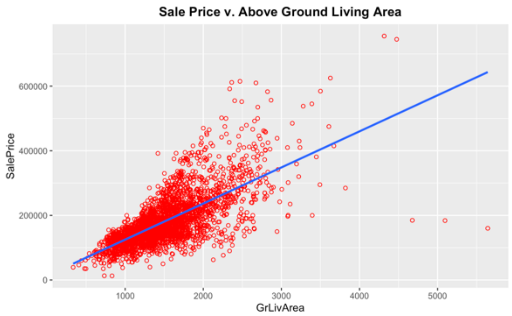
EDA Assignment # 1

Introduction and Sample Definition

The data provided for the purpose of this assignment comes from the Ames, IA Assessor's Office which uses this information to determine individual property values from the years 2006 to 2010. The dataset contains 82 different features, which can be described as nominal, ordinal, discrete and continuous. The goal here is to explore the data set in order to get a better understanding of the information at hand so that it is possible to determine the features that make up an average priced home in Ames, IA. The response variable here is sale price for which predictors will be determined in order to determine which features contribute most to sale price.

As it pertains to the features listed in the data dictionary, I found certain variables to lack less relevance than others as it pertains to assessing the price of a home. Road access, land contour and roof style/slope are just a few examples of features that seem to have little use. On the other hand, features like zoning, qualities, conditions and square footage make sense as predictors for sale price. To begin defining the population of interest I looked at a boxplot of above ground living area square footage as well as a scatter plot of the aforementioned feature with sale price.





From the visualizations I decided to remove any properties > 4,000 sqft to remove the outliers that appear to be atypical to the data. After this point, I filtered certain features prior to performing a quality check on the data:

1. Zoning: removed any commercial, agricultural and industrial zones that would not seem to represent average homes in Ames
2. Building Type: removed all records except single family dethatched homes
3. Sale Condition: set to "Normal"

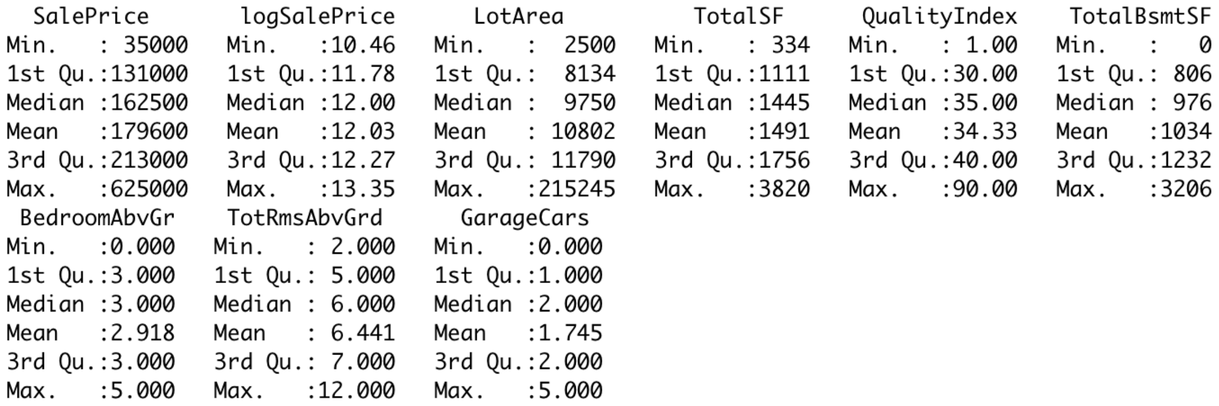
The data set prior to these steps contained 2,931 unique observations, which was then filtered down to have only 1,987. This is a 32% decrease in the size of the data being worked with.

Data Quality Check

The variables mentioned below were chosen as potential predictors for sale price.

Sale Price	Log(Sale Price)	Lot Area	Neighborhood
Condition 1	House Style	Year Built	Total Square Footage
Price per Square Foot	MasVnrArea	Lot Shape	Quality Index
Total Basement Square Footage	Lot Frontage	Bedrooms Above Ground	Ground Living Area
Garage Type	Garage Cars	Garage Year Built	Year Remodel

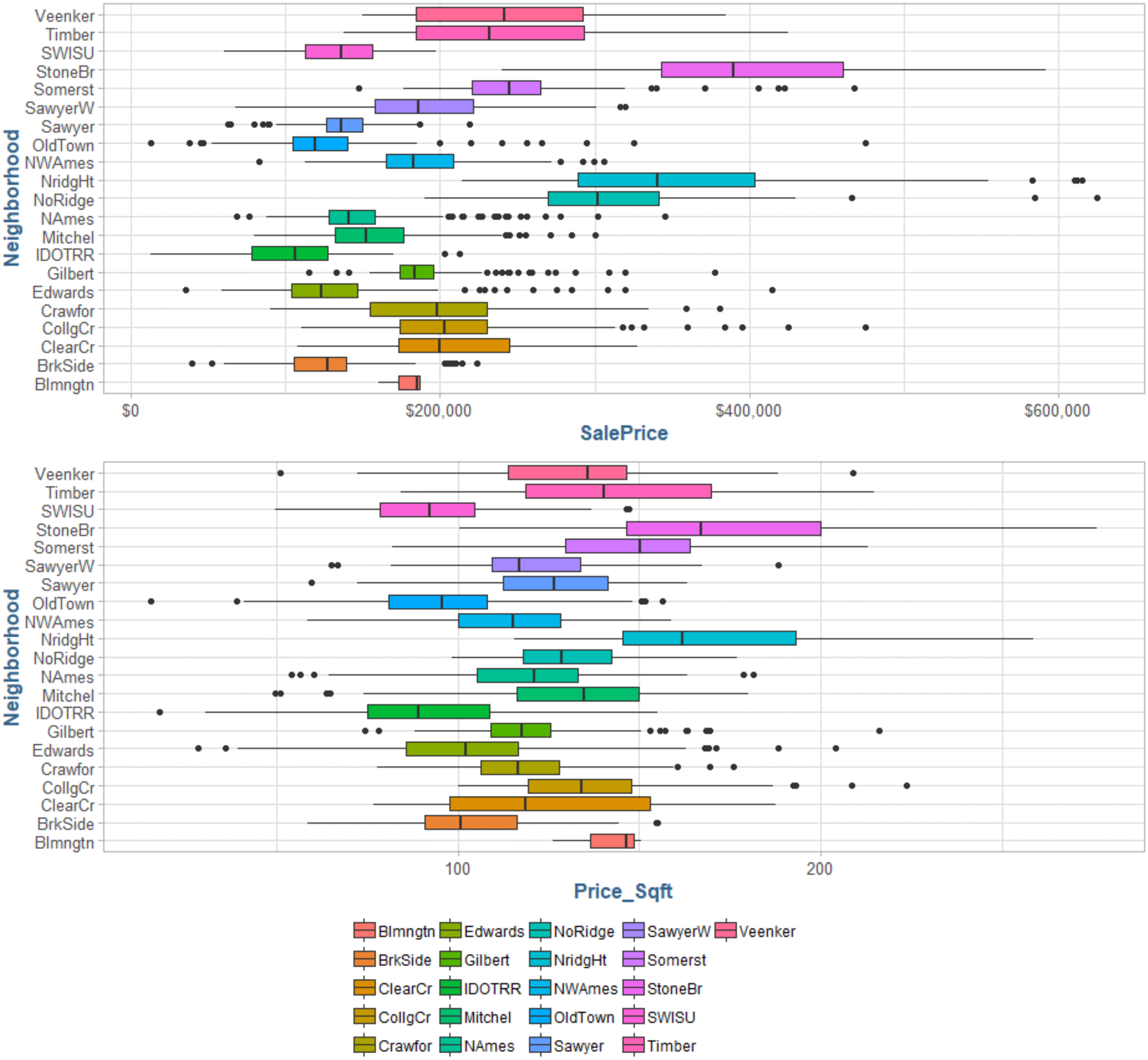
The numeric, continuous variables were explored using the summary() function while the rest were investigated using the table() function. For the predictor variables in this model, we have narrowed down the universe based on intuition and correlation. We can further break down these variables into more granular categorizations by looking at the area of impact and quantifiable measurement we can observe in relation to the desired response variable, sale price. For the selected variables in this analysis, we can see that there are several missing values for the variables GarageYrBlt and LotFrontage, as well as a probable miscoding for the value 2207 in GarageYrBlt. Below you can find a table of summary statistics pertaining to the continuous variables in the dataset.



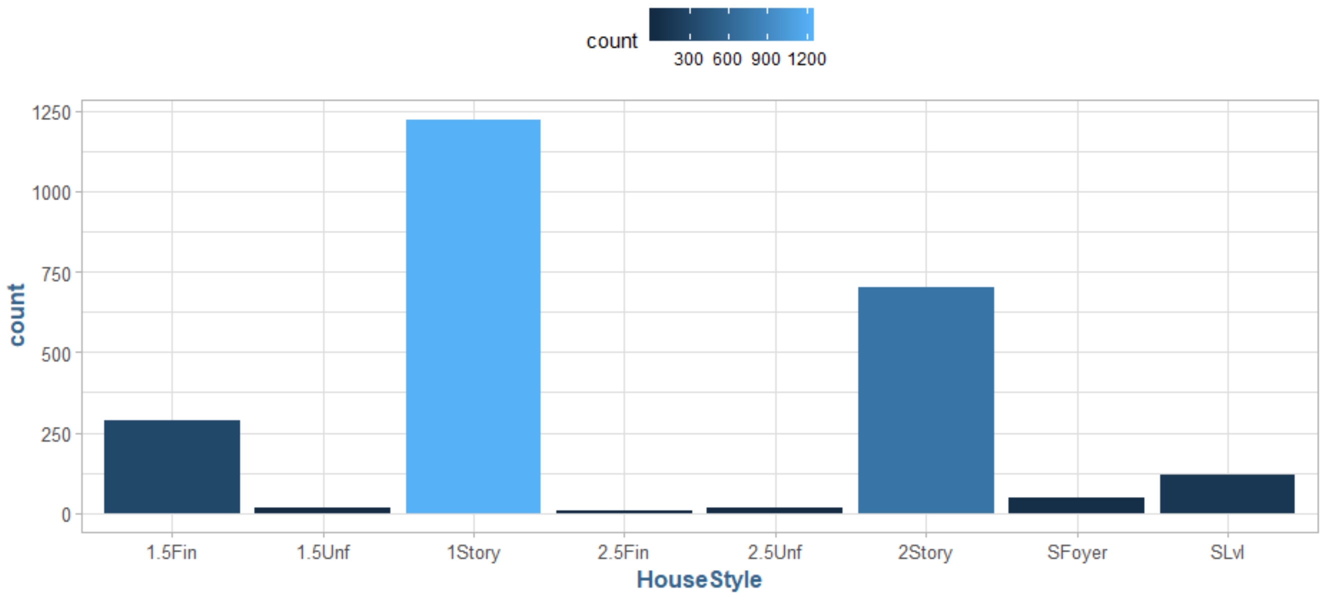
```
   SalePrice        logSalePrice       LotArea          TotalSF        QualityIndex     TotalBsmtSF
 Min.   : 35000   Min.   :10.46   Min.   :  2500   Min.   : 334   Min.   : 1.00   Min.   :   0
 1st Qu.:131000   1st Qu.:11.78   1st Qu.:  8134   1st Qu.:1111   1st Qu.:30.00   1st Qu.: 806
 Median :162500   Median :12.00   Median :  9750   Median :1445   Median :35.00   Median : 976
 Mean   :179600   Mean   :12.03   Mean   : 10802   Mean   :1491   Mean   :34.33   Mean   :1034
 3rd Qu.:213000   3rd Qu.:12.27   3rd Qu.: 11790   3rd Qu.:1756   3rd Qu.:40.00   3rd Qu.:1232
 Max.   :625000   Max.   :13.35   Max.   :215245   Max.   :3820   Max.   :90.00   Max.   :3206
  BedroomAbvGr    TotRmsAbvGrd      GarageCars
 Min.   :0.000   Min.   : 2.000   Min.   :0.000
 1st Qu.:3.000   1st Qu.: 5.000   1st Qu.:1.000
 Median :3.000   Median : 6.000   Median :2.000
 Mean   :2.918   Mean   : 6.441   Mean   :1.745
 3rd Qu.:3.000   3rd Qu.: 7.000   3rd Qu.:2.000
 Max.   :5.000   Max.   :12.000   Max.   :5.000
```

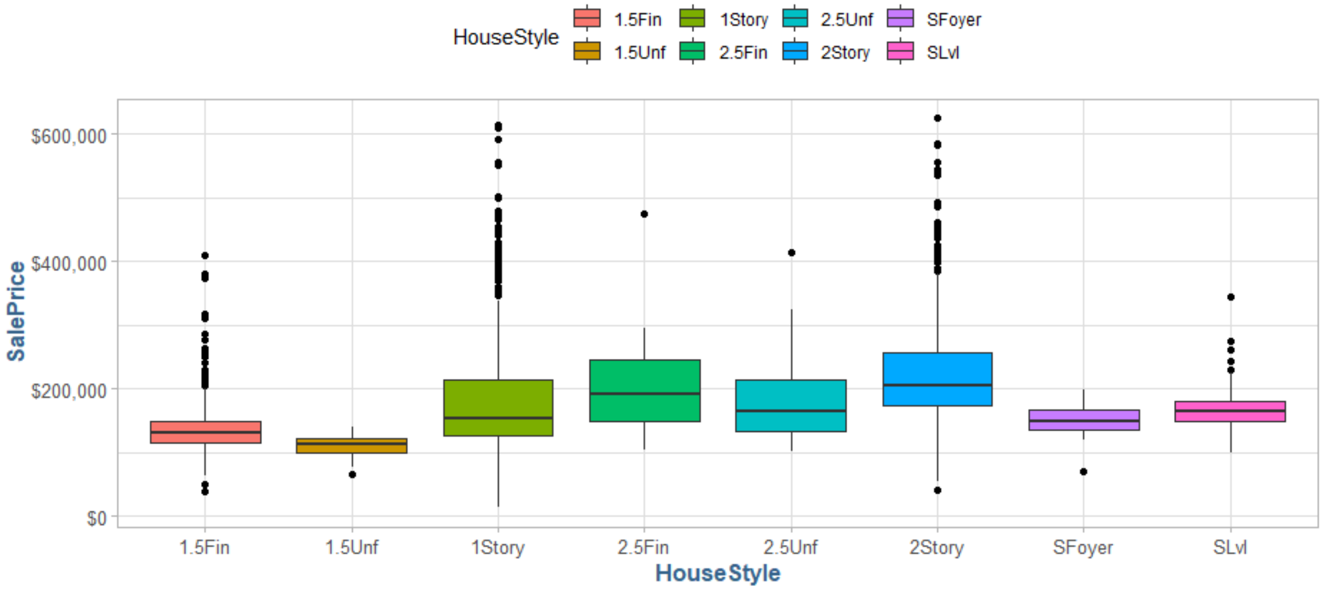
Initial Exploratory Analysis

For initial exploratory data analysis, we will choose ten features from the housing dataset that have either a high correlation to sale price (for numeric variables) or have an intuitively strong connection to the sale price (categorical variables such as neighborhood). Homes in specific neighborhoods are often built using the same set of pre-approved floor plans and layouts that have similar specifications due to homeowner's association, therefore we would expect a strong correlation with neighborhood. We can see the distributions of sale price and price per square foot in the following figure according to neighborhood:



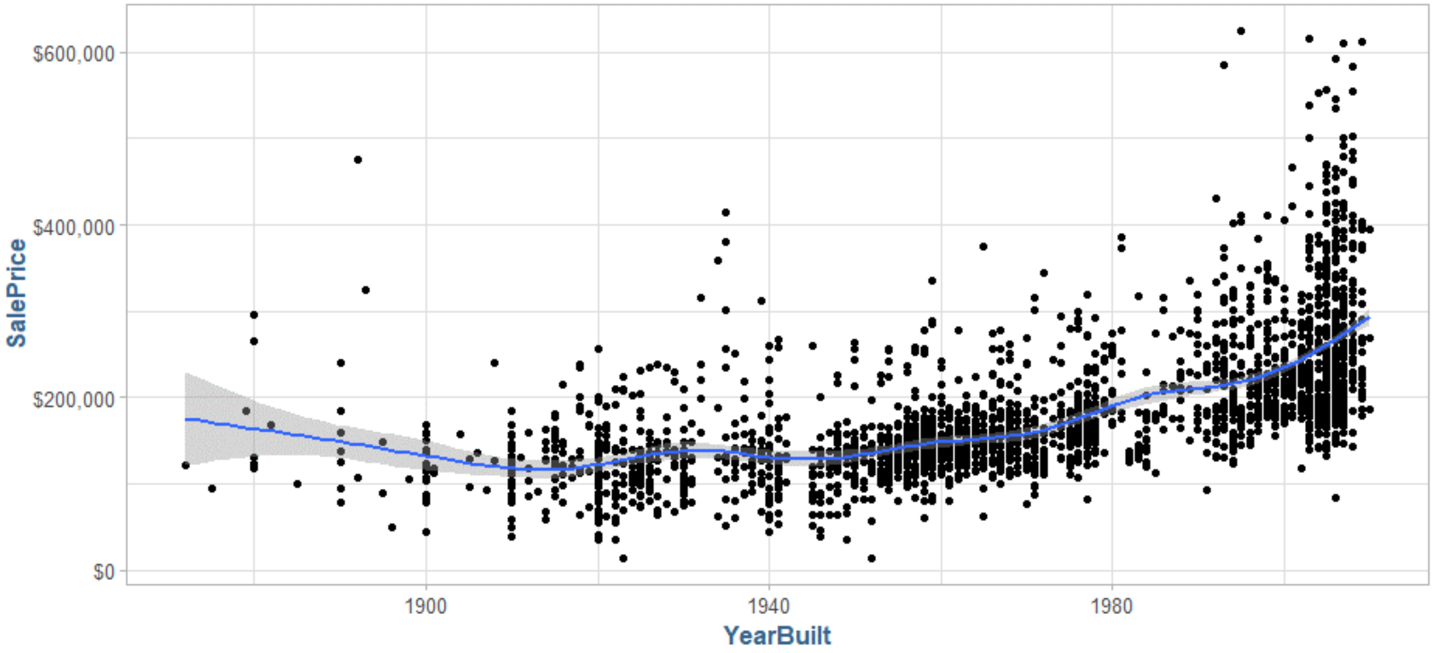
The distributions seem to be highly scattered and show minimal clustering and therefore will likely lead to low predictive value. Another categorical variable that has potential for a quality prediction in sale price intuitively is the housing style, as variations in home type should have similar pricing points. The following figure shows the distribution of housing style, and how sale

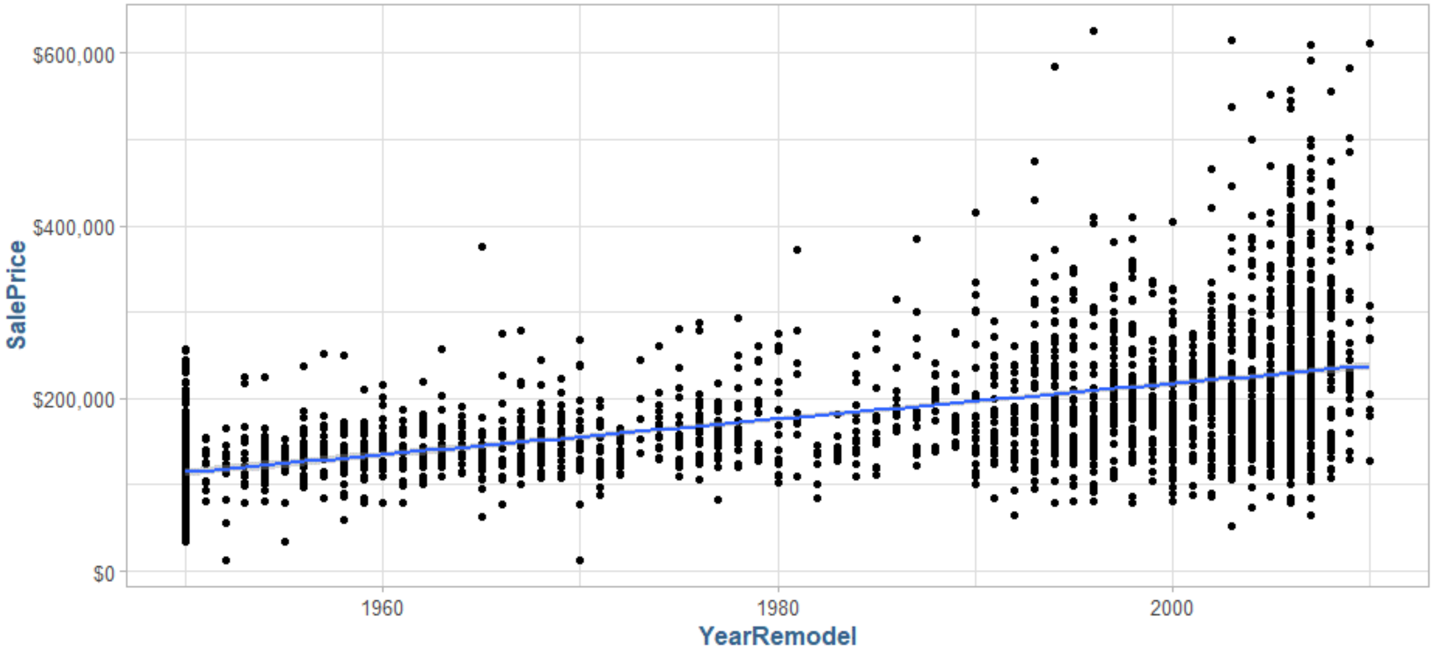




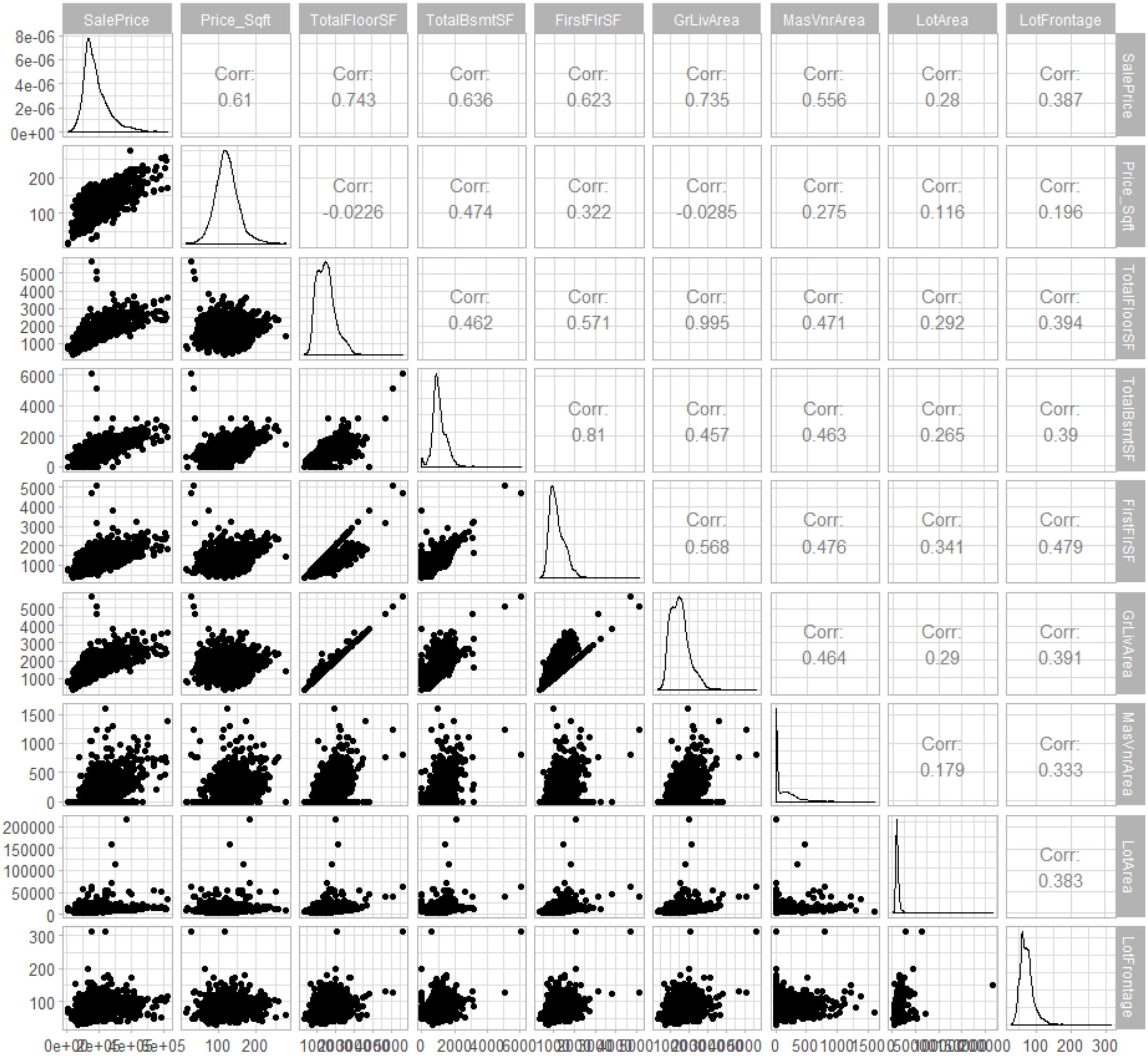
price is distributed amongst the different housing styles:

As can be seen in the above figures, by far the most common housing styles are one-story and two-story homes, accounting for near 80% of the total sample. However, looking at the distribution for the sale price amongst the different housing types, it does not appear we will be able to rely on the housing style as a predictor due to the amount of outliers for one-story and two-story homes would overcomplicate the model to fit these values. Another set of variables we should explore are the discrete variables, such as year built and year remodeled. For these variables, we can look at a scatterplot of the year versus the sale price to observe whether or not there is a relationship:

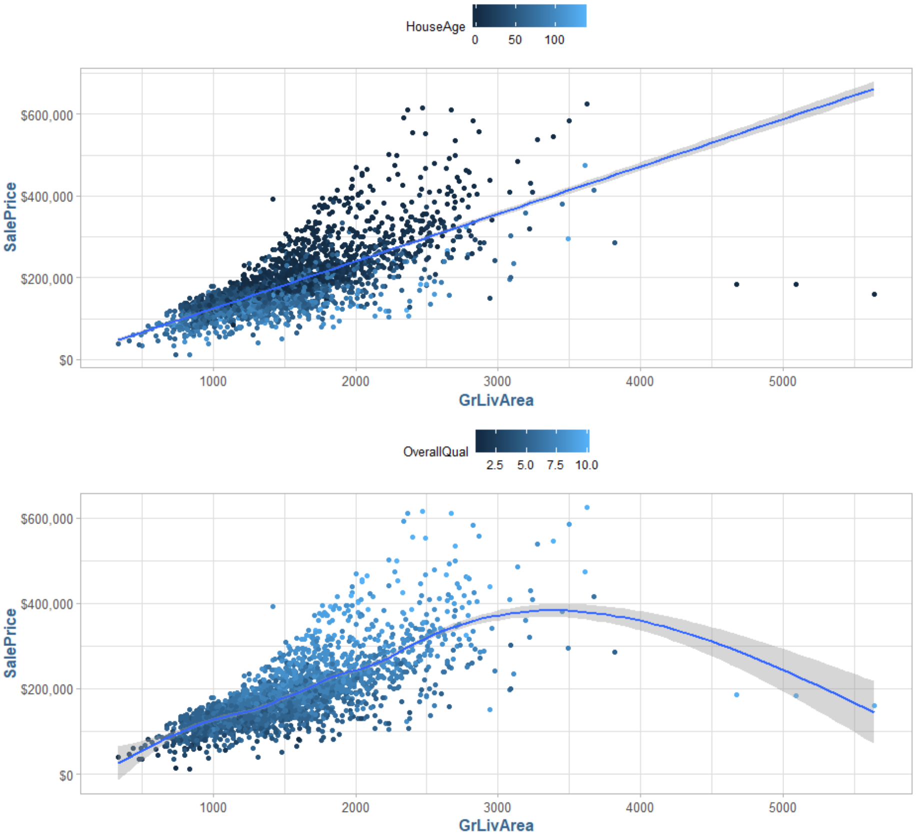




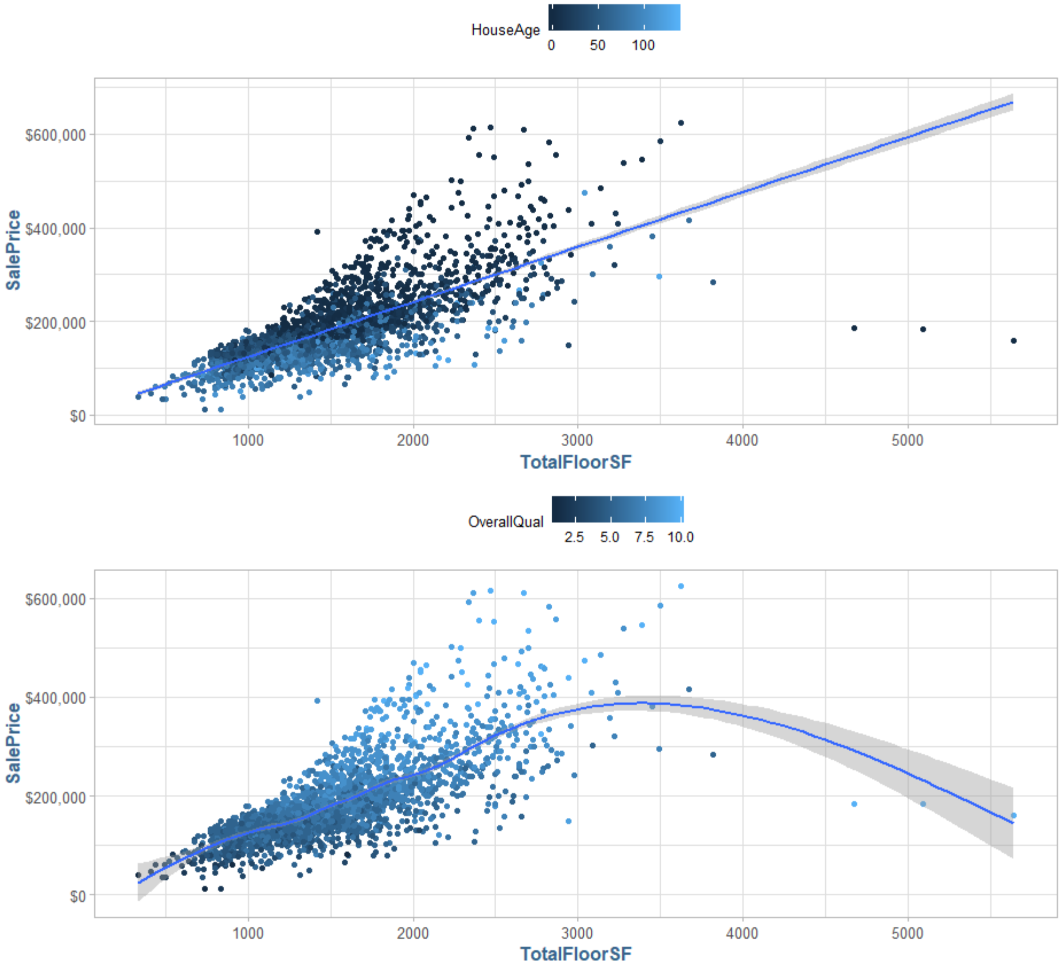
In the above top figure, we observe the year a house was built with a smooth fit against its sale price. YearBuilt has a strong relationship to sale price if the home was built post 1940. In general, the year a home was remodeled shows a clear linear relationship to its sale price. Hence, these two variables would be suitable as predictors of sale price.



For the continuous variables in the dataset, we can look at a scatterplot matrix to explore the relationships between them, which can be seen above. During our initial analysis of the variables contained in the dataset, we noted the high correlation of sale price to the continuous variables above ground living area and total square footage of the home. It would be interesting to look at the relationship between these two variables. The following figure visualizes sale price as a function of above ground living area. The linear model showing the overall positive trend and high correlation, as well as a more flexible (LOSSES) model that demonstrates that there is potential for overfitting as can be seen by the outliers in the dataset.

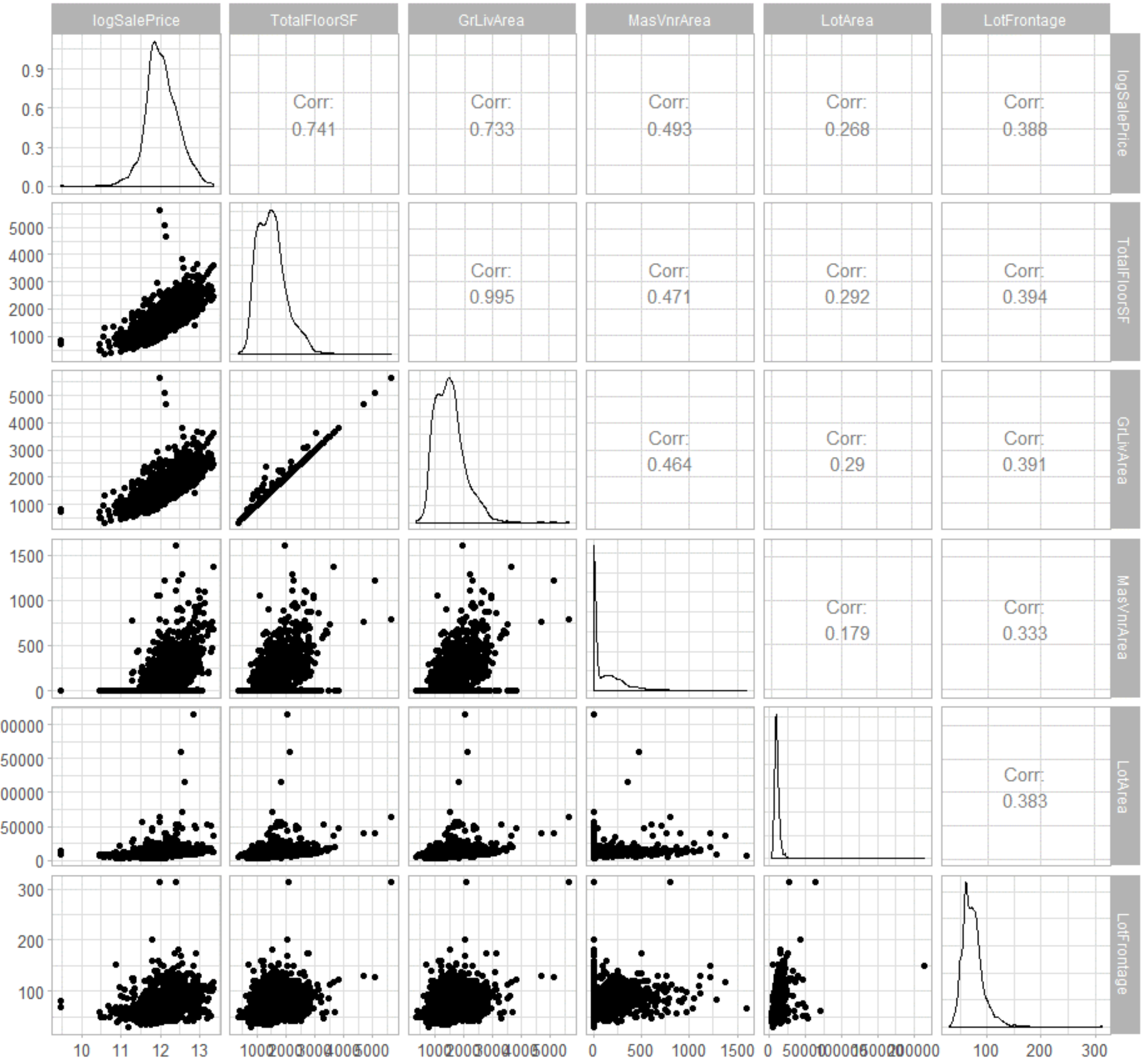


We can similarly look at sale price as a function of total square footage in the following figure. In the top chart in both figures we see a strong negative correlation to the houses age depicted by the light to dark color scheme, and a strong positive correlation to the quality index denoted by the inverted dark to light color scheme in the bottom figure.

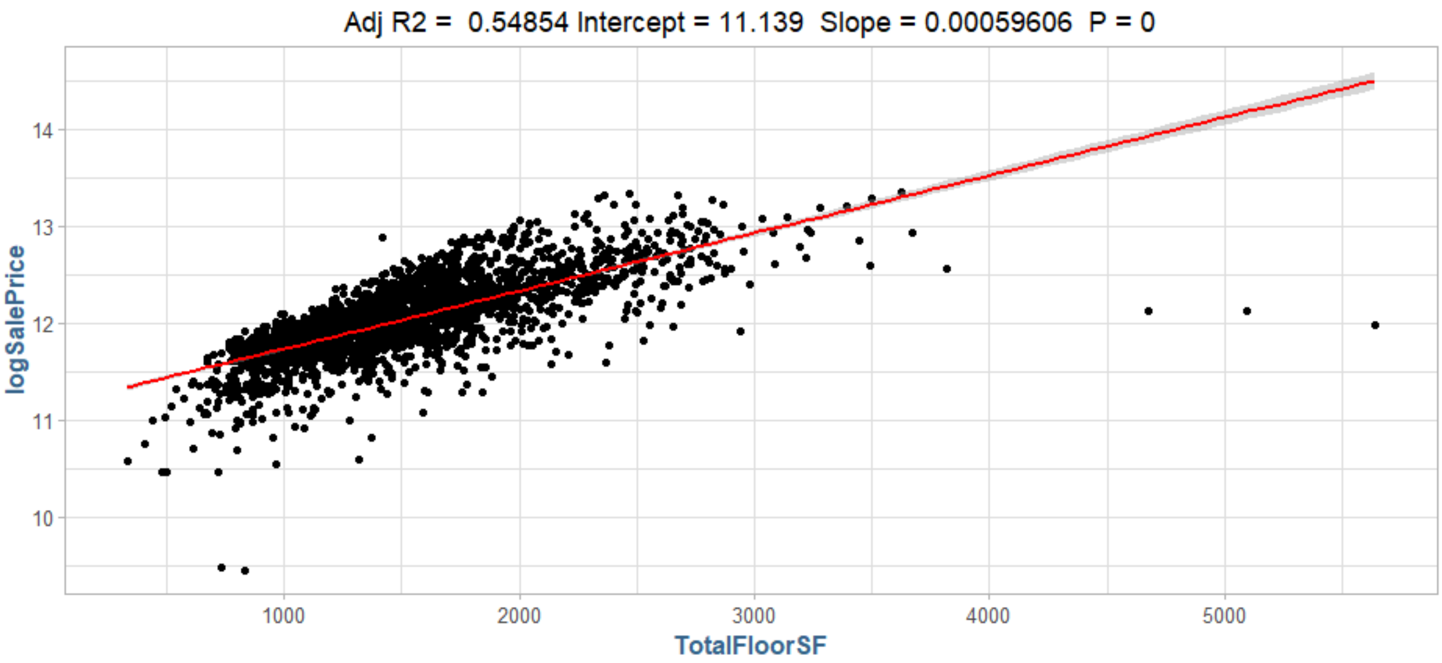


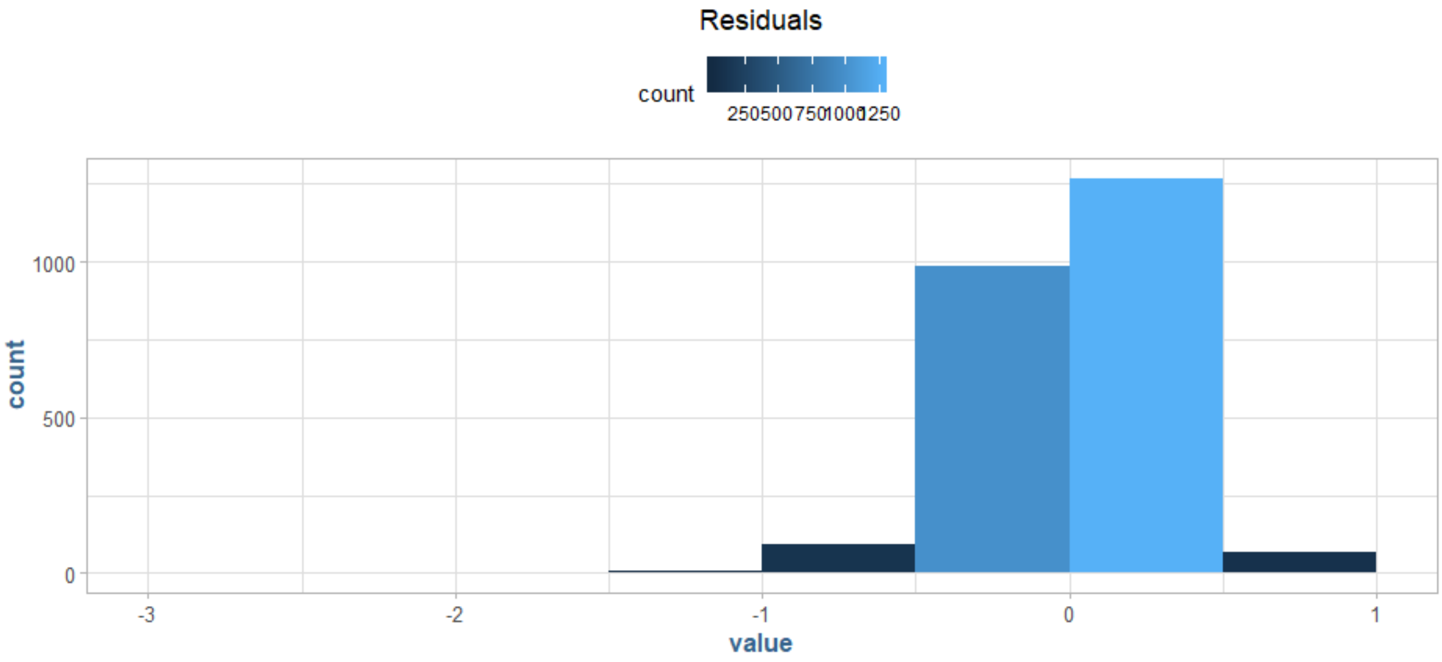
Exploratory Data Analysis for Modeling

For the modeling phase of this analysis, we will use the log transformed sale price response variable. The features that exhibited the most predictive ability during the exploratory data analysis are shown in a scatterplot matrix below:

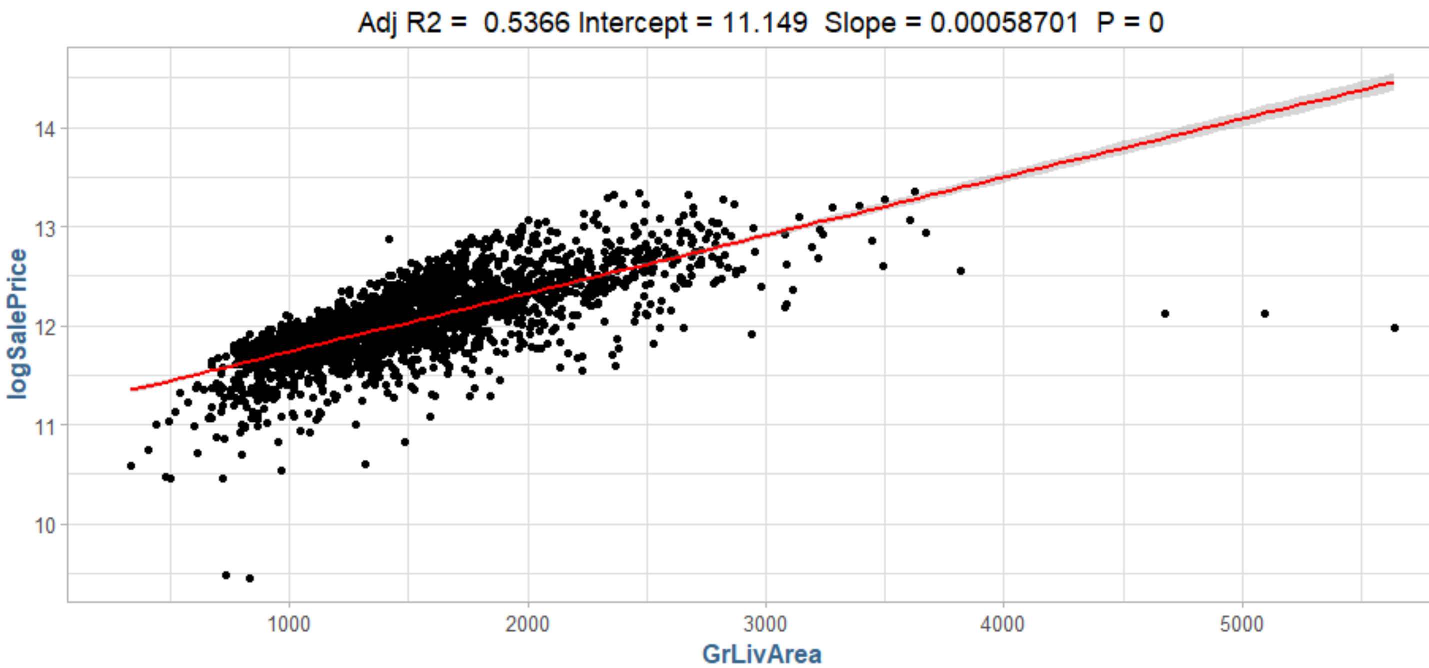


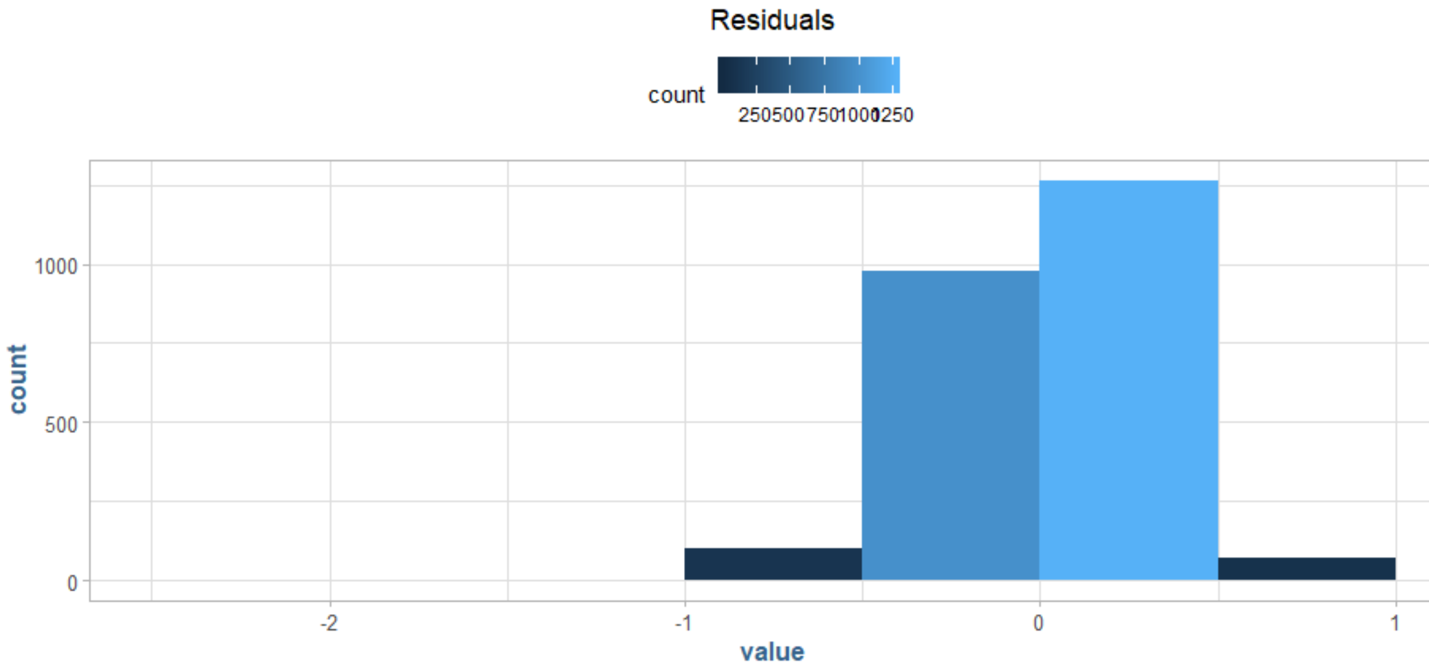
The first model is a simple linear model based on log sale price as a function of total square footage, which can be seen in the following figure below:



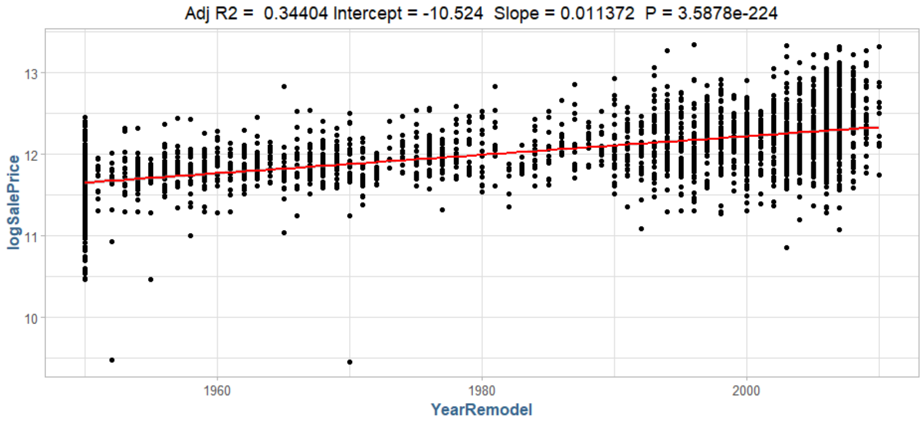


The simple linear model fitted on the total square footage exhibits a great deal of error in many cases, as can be seen above. The adjusted r-squared also indicates that 54.8% of the variation of logSalePrice is explained by TotalFloorSF. The next model below was fitted to the above ground living area:

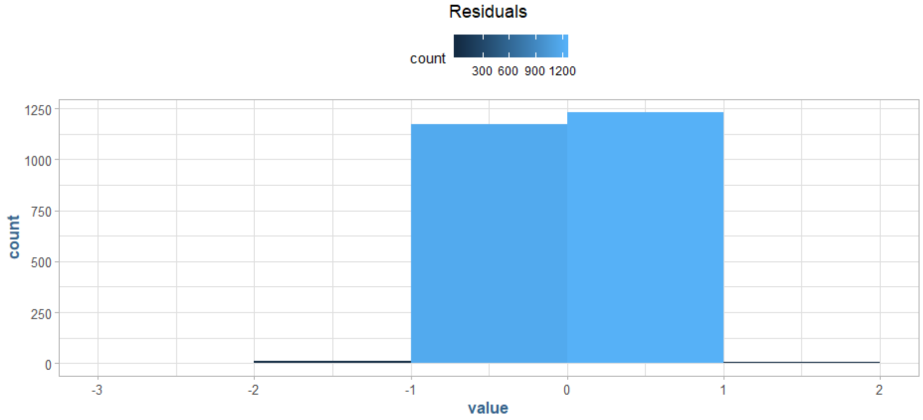




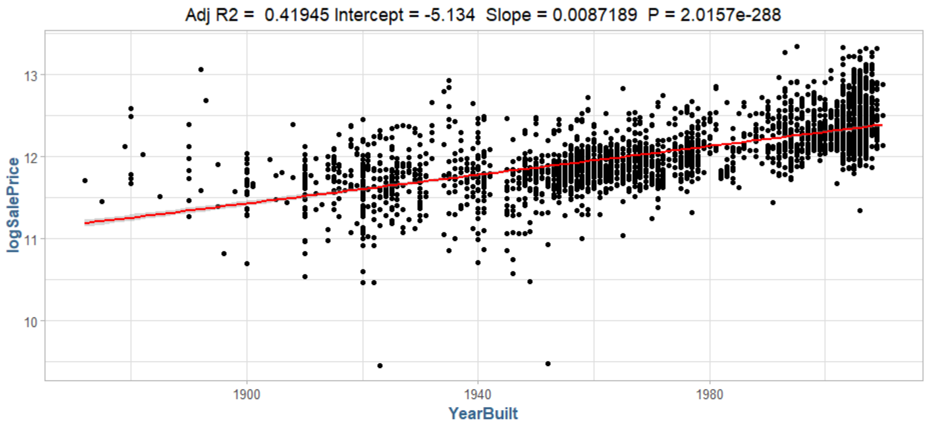
This model performs similarly to the total square footage model we looked at previously, although it does perform slightly better in some cases. The adjusted r-squared also indicates that 53.6% of the variation of logSalePrice is explained by GrLivArea. The next two models are based upon the more linear relationships we discovered in our exploratory phase. The year the house was built and the year it was remodeled, which can be seen in the two figures below:

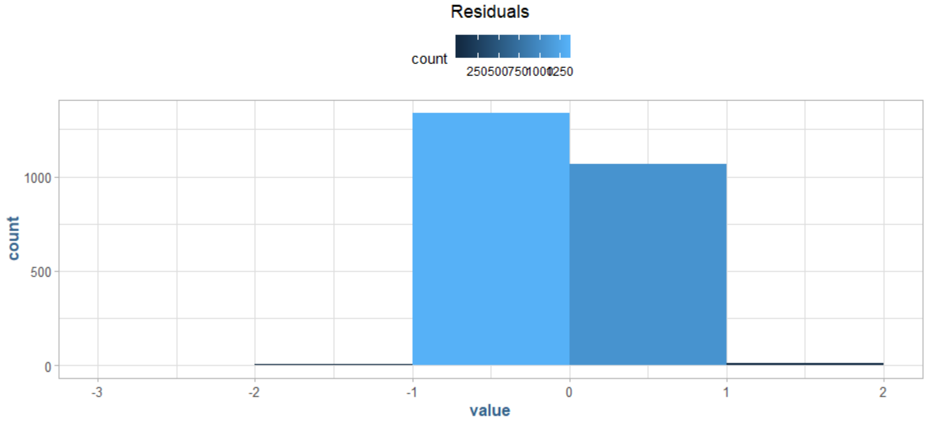


Adj R2 = 0.34404 Intercept = -10.524 Slope = 0.011372 P = 3.5878e-224

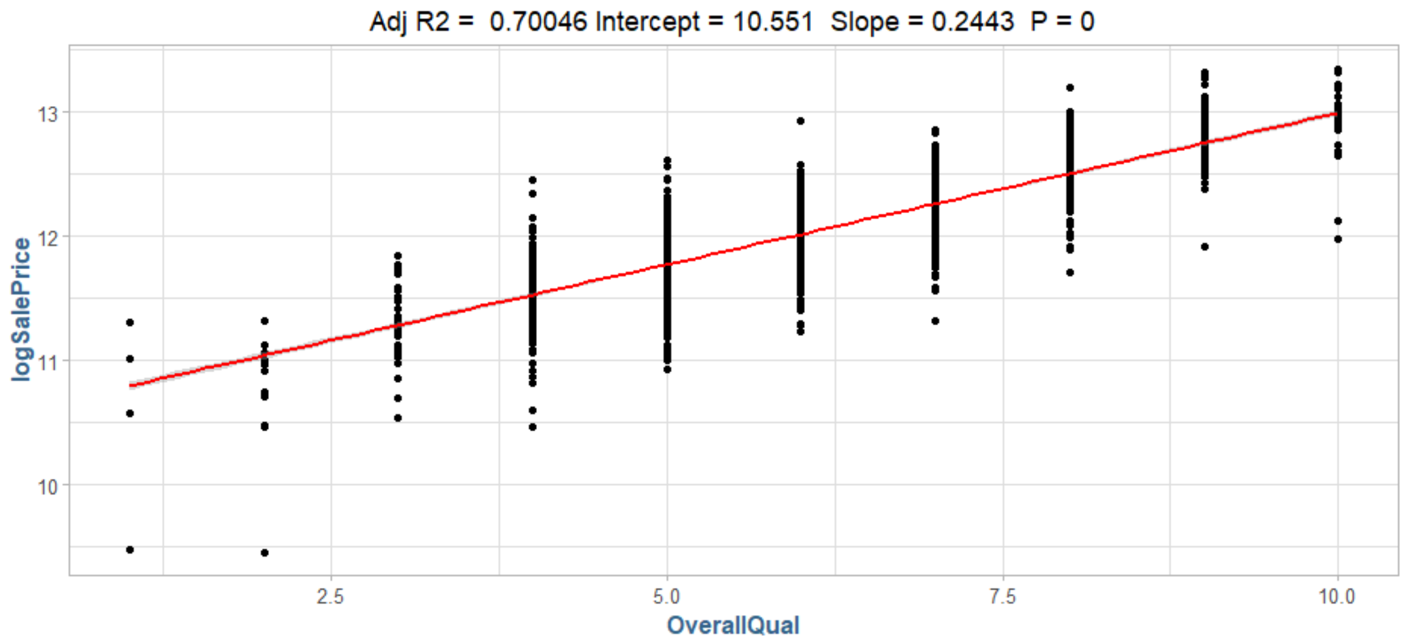


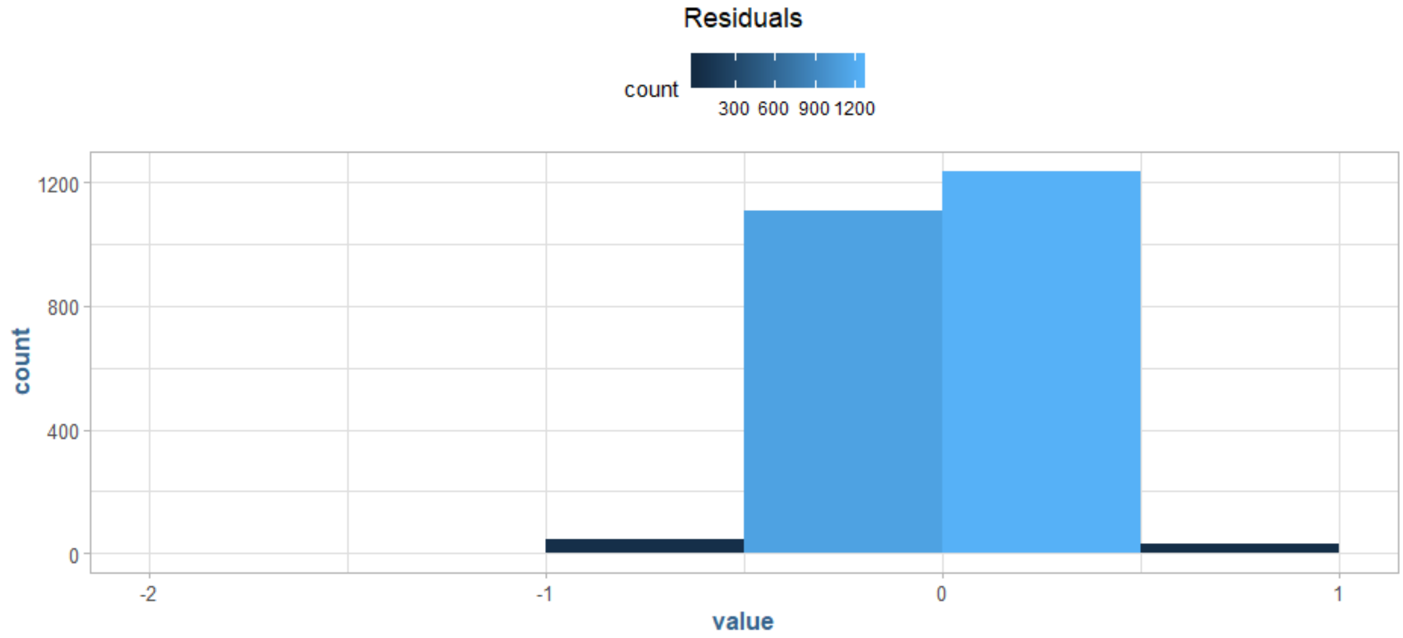
Residuals



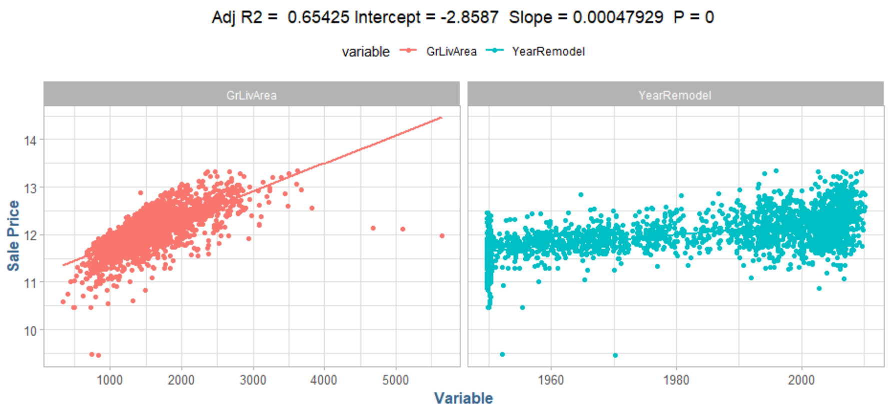


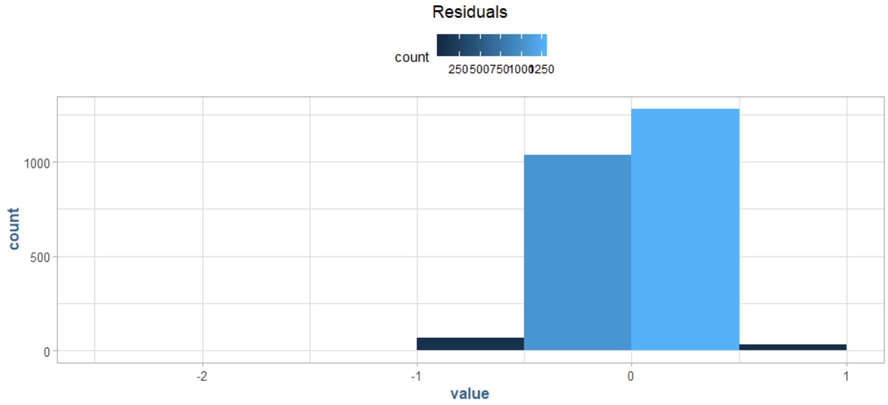
These two models performed far better based on the temporal variables than the previous models which used continuous variables relating to housing area. The final single variable linear model, overall quality index, is the one that generated the highest correlation we saw during the initial exploration phase:



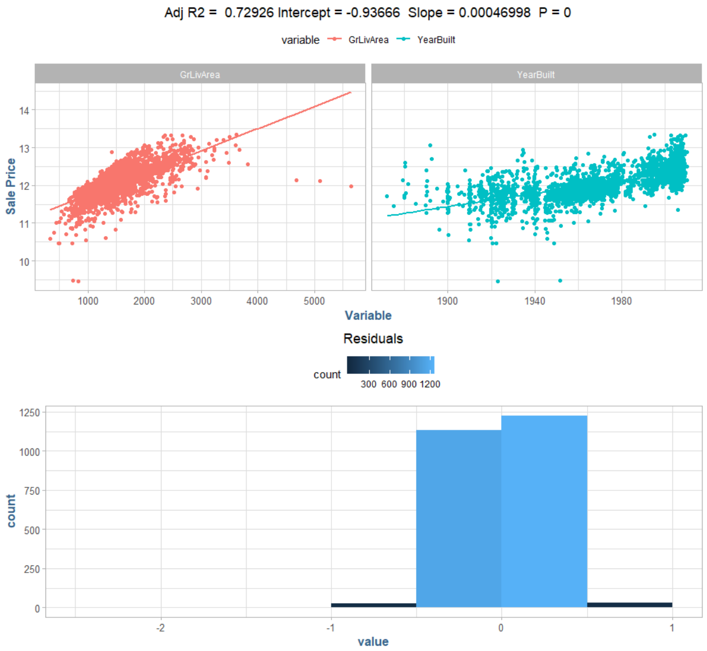


The results of the single linear regression are underwhelming; however, we can use multiple variables to help accommodate for unexplained bias using a standard linear model. In the following figure we will look at two variables that showed potential from the previous section, those being above ground living area and year remodel:

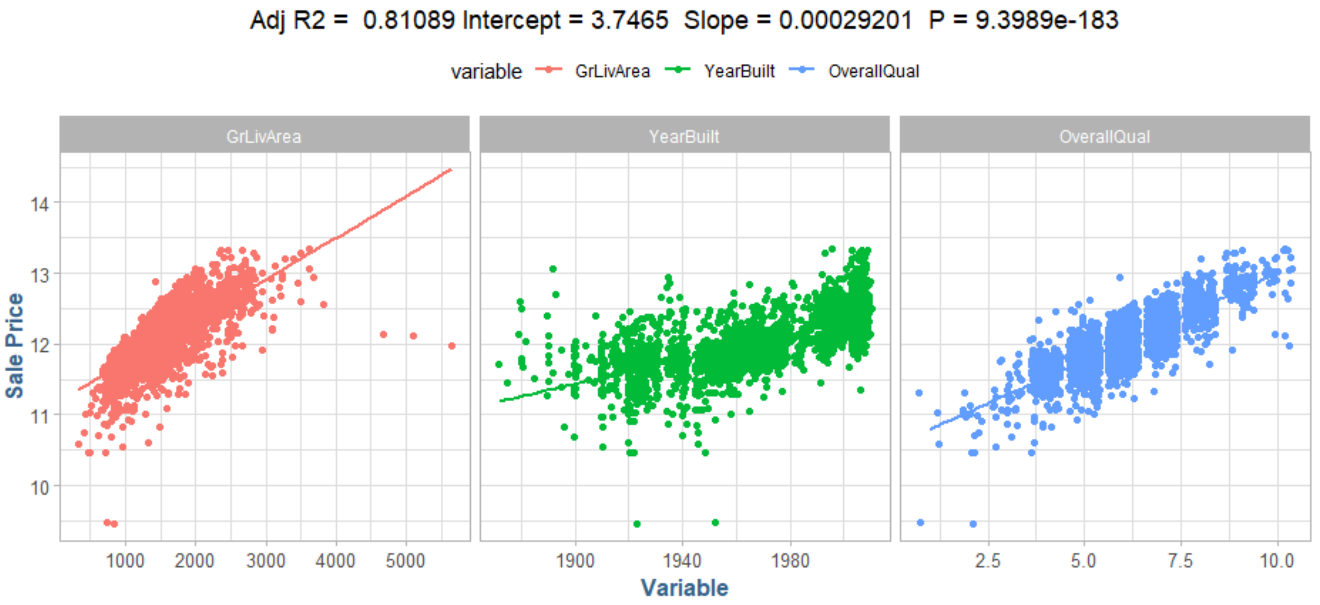




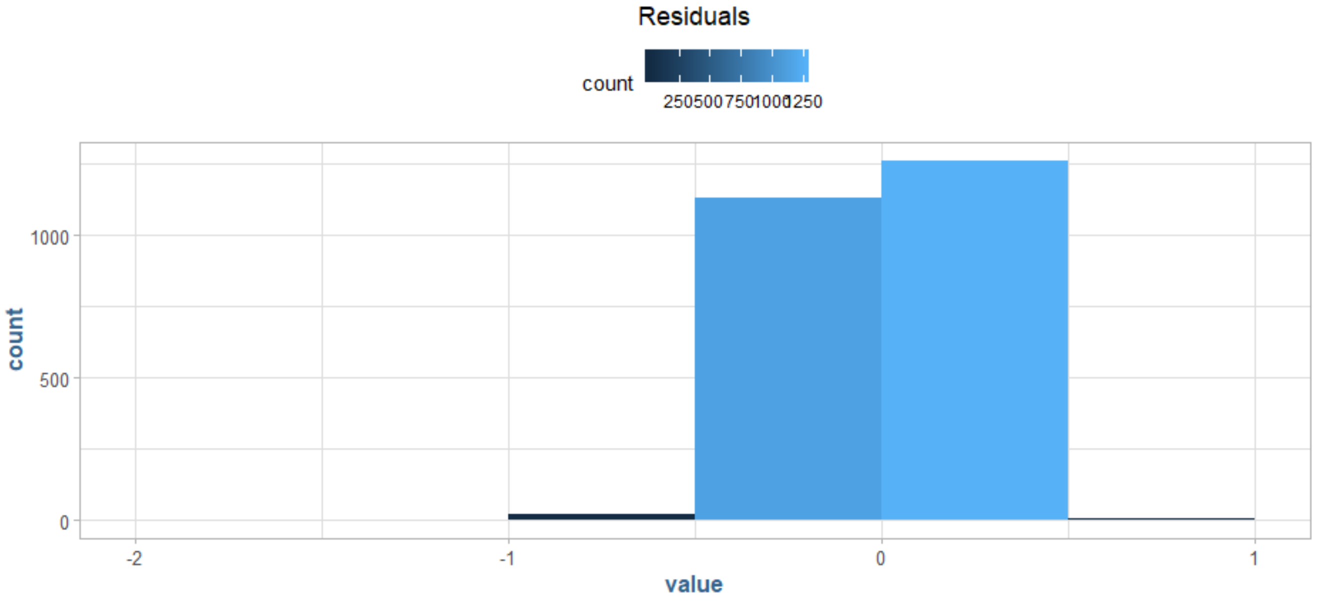
This model does not perform as well the model that solely used the overall quality index The adjusted r-squared indicates that 65.4% of the variance in sale price can be explained by GrLivArea and YrReModel. Expanding this model to use year built, we can see a slight increase in performance:



The final model that will be built will be a multiple linear regression model fitted to use the top three variables that have been solid predictors of sale price. Above ground living area, year built, and overall quality were the variables observed to have the most prediction power as it pertains to price and the adjusted r-squared value indicates that 81.1% of the variance in sale price can be attributed to the aforementioned variables. The result is depicted in the following figure:



Adj R2 = 0.81089 Intercept = 3.7465 Slope = 0.00029201 P = 9.3989e-183



Residuals

Conclusion

In this lab I was able to perform a data survey, define a sample population, executed a data quality check on selected features, conduct exploratory data analysis and generate initial model formulation. The data suggests there are essential relationships between the housing explanatory variables and sale price. From the defined sample population, we see a semi-colinear relationship from the overall square footage and above ground living area, a stronger colinear relationships to the year built and year remodel and as well as a strong colinear relationship to the overall quality. Using a combination of these variables, we can explain a significant portion of the variance in the sample population. For further analysis, we should explore non-linear models in order to generate a more accurate fit to the data at hand.