**Assignment #1**

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 dataset, our task will be to create a multivariate regression model that leverages these key indicators in the data to predict the value of a home given based upon its features.

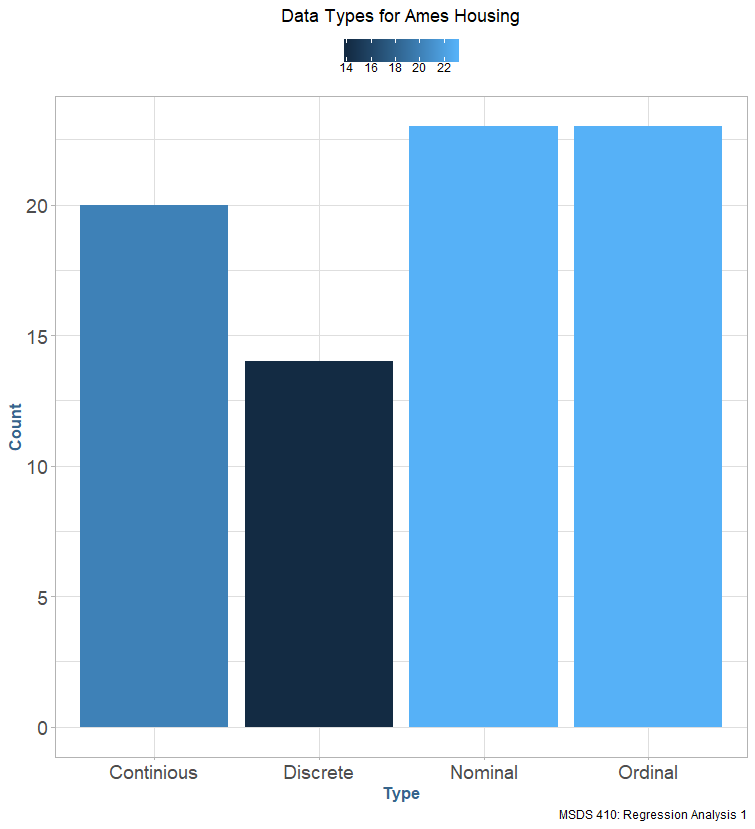
In this report, we will use the Ames dataset which is an alternative to the famous Boston housing data to perform exploratory data analysis through variable derivation, validation, selection and visualization to measure the relevance of these indicators as they pertain to the value of the home in terms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

### Data Survey

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

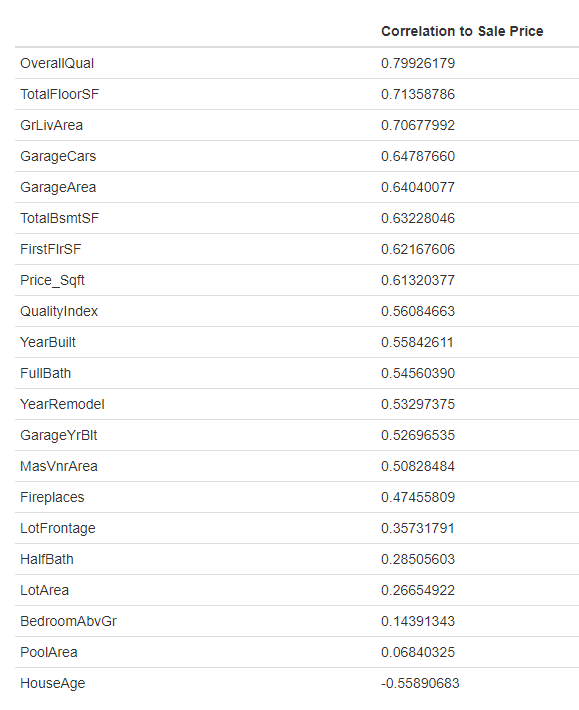
The Ames housing dataset contains approximately three-thousand observations of eighty-two variables collected from the Ames Assessor’s Office specifically for the purpose of assessing value of individual residential properties sold in Ames, Iowa from 2006 to 2010. Given that this data was collected for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

Taking a deeper dive into the eighty-two characteristics of each property, we can classify twenty-three as nominal, twenty-three as ordinal, fourteen discrete and twenty continuous.



In addition to the variables provided in the dataset, we can also derive our own calculated and derived variables for use in our predictive model. Given that our overall goal is to predict the sale price we can calculate the total square footage of the house by combining the square footage of all the floors and break each property down to a price per square foot. This will provide us with a generalized common denominator for each property so that we can assess the impact of features such as house style, neighborhood and quality on a per-square foot basis.

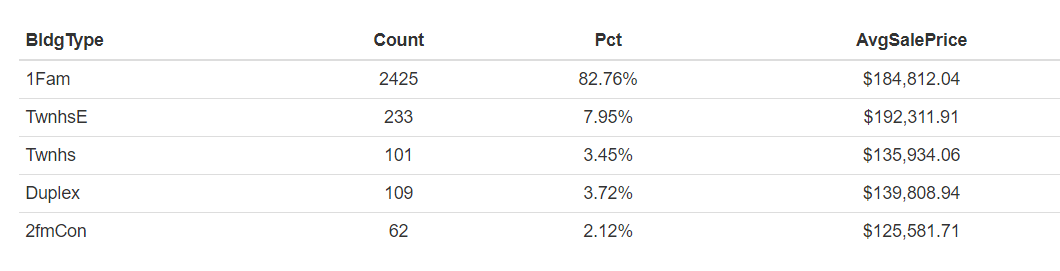
Since we are interested in building a linear model to predict a homes sale price given its features, we should consider which variables are correlated to the sale price. In the following figure we can see the correlation of several variables to the sale price, which we will note here and explore more in-depth in the preceding sections. The following variables were chosen from the total dataset based on initial intuition around housing prices and what drives them. A full accounting of the correlation variables can be found in the [appendix](#Feature_Correlation).



These variables will form the baseline for our model. The simple correlation does not tell us explicitly how useful a given variable will be in building a model, but it is useful to note that intuitively overall quality, total square-footage, and living are have a highly positive correlation, and house age is a strong negative correlation as we would suspect with our simple intuition.

### Sample population

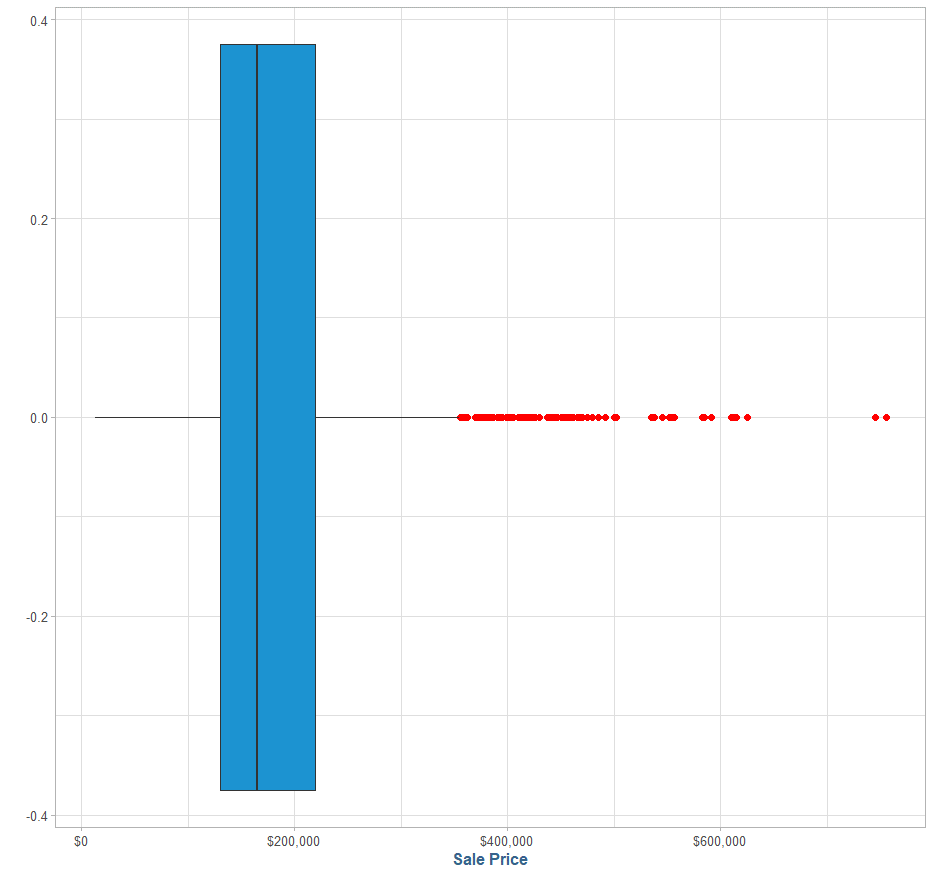
To check the similarity to other property types we can examine the average sale price by building type in the following table:



In the preceding table we can clearly see that single-family homes make up most of the data. Given that many homes are of type single-family, and the other building types make up such a small portion of the observed data it might over complicate the model in order to accommodate the special cases for these homes. Therefore, we will exclude any homes not of type single-family from the proceeding analysis in order to meet the objective of predicting a ‘typical’ home price.

### DATA QUALITY CHECKS

We will perform some standard data quality checks against the primary target variable, sale price. No negative values or non-applicable values are reflected in the population. However, there are some cases of relatively extreme outliers in the sale price as we can see in a boxplot for the variable in the following figure.



The values on the far right represent homes with valuations over $700,000 which is far greater than the average value, which is the objective of this analysis, therefore they will be excluded from the rest of this analysis. The boxplot also shows a considerable amount of kurtosis in the distribution of sale prices, so special attention will be given to monitor the impact on the residuals.

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**.

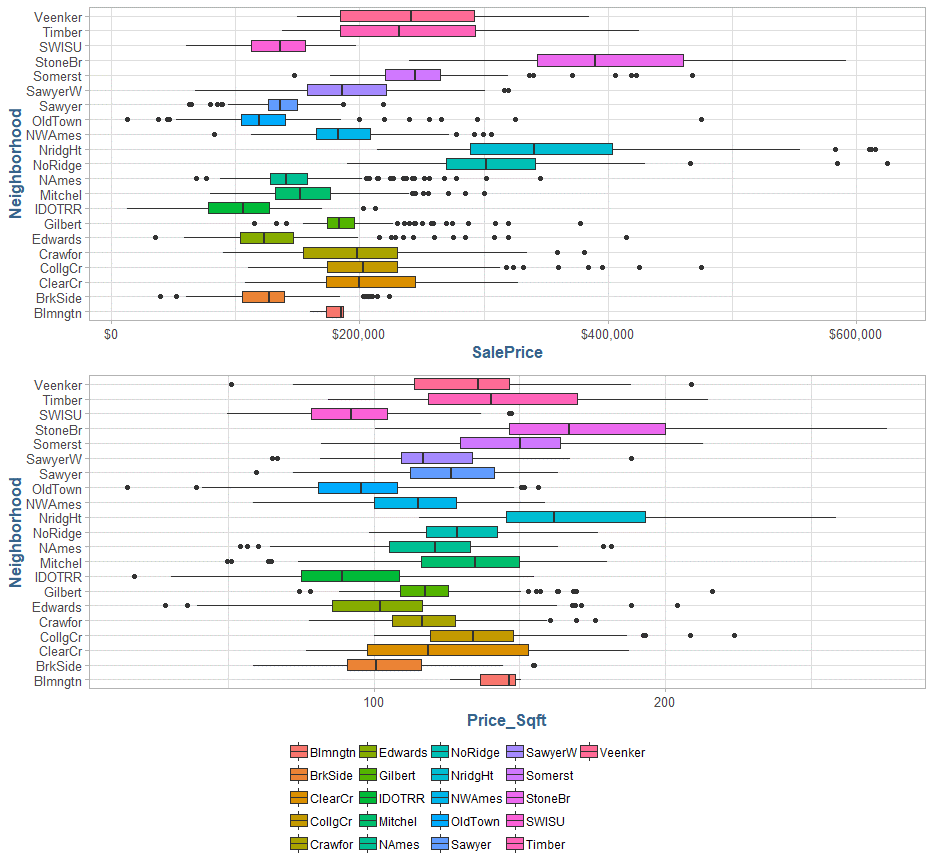
At a high level we can group these features into the following categories: temporal discrete (**YearBuilt**, **YearRemodel, GarageYrBlt**), discrete (**FullBath**, **HalfBath**, **Fireplaces**, **GarageCars, HouseAge**), continuous (**Price\_Sqft, TotalFloorSF, TotalBsmtSF, FirstFloorSF, GrLivArea, MasVnrArea, LotArea, LotFrontage**) and ordinal (**OverallQual, QualityIndex, KitchenQual**).

Further data quality checks on a subset of selected features can be found in the [appendix](#Data_Quality_Check).

### Initial exploratory data analysis

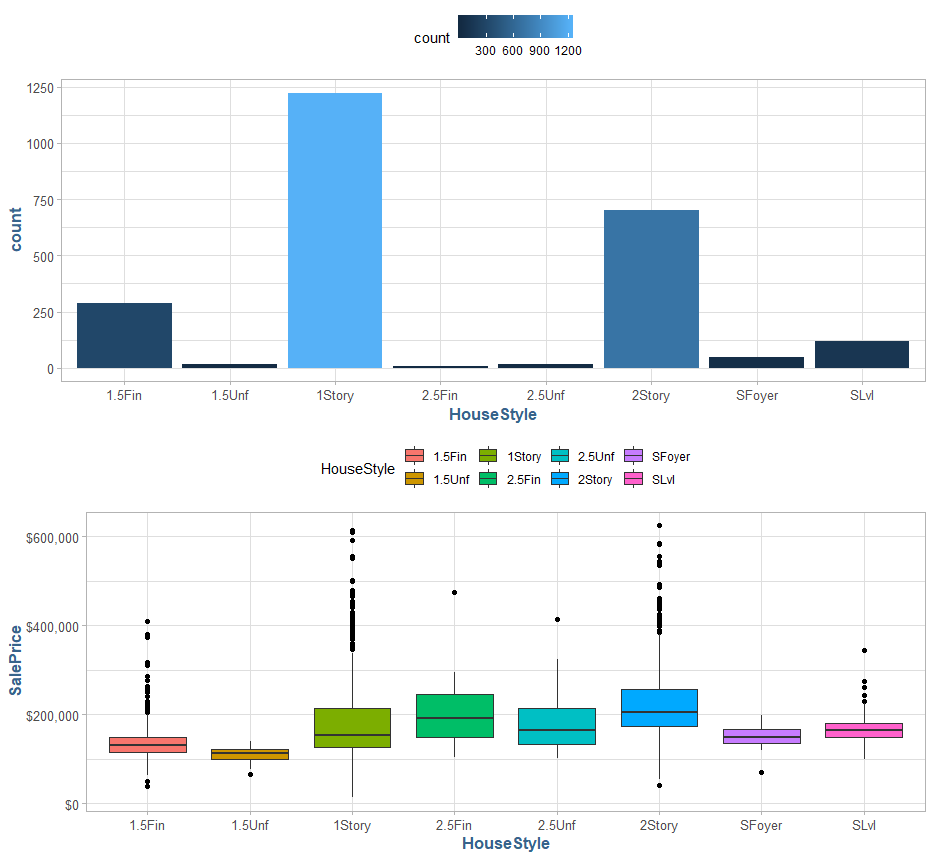
For initial exploratory data analysis, we will choose ten features from the housing dataset that have either a high correlation to sale price (for numeric variables) or have an intuitively strong connection to the sale price (categorical variables such as neighborhood or masonry veneer type).

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 policies, therefore we would expect a strong correlation, or clustering of home prices and similar price per square footage in the same neighborhood. We can see the distributions of sale price and price / sqft in the following figure:



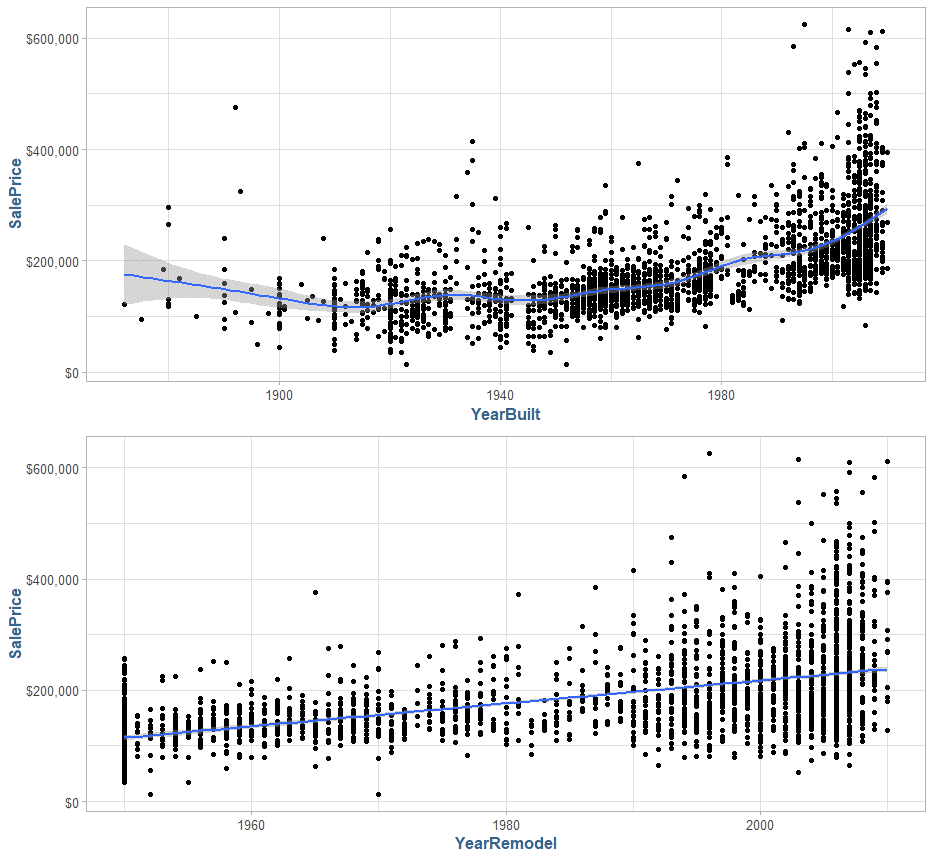
Unfortunately, the distributions are highly scattered and show minimal clustering and therefore will likely lead to low predictive value.

Another categorical variable that has potential for predictability 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:



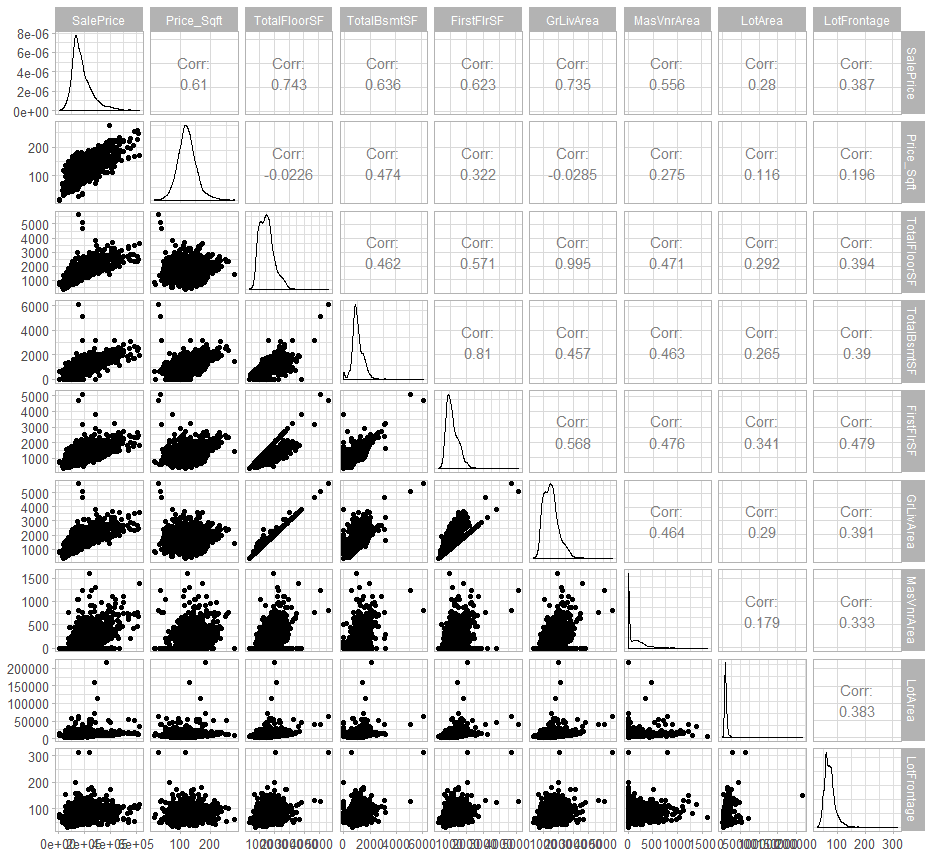
As we see in the above figure, by far the most common housing styles are traditional one-story and two-story homes. These homes account for almost 80% of the total sample, which is not surprising given that we filtered out only single-family homes in the preceding section. However, looking at the distribution for the sale price amongst the different housing types, it does not appear we will be able to rely on the housing style as an indicator for the sale price as 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 (**YearBuilt**) and year remodeled (**YearRemodel**). For these variables, we can look at a scatterplot of the year versus the sale price to see if there is a relationship:



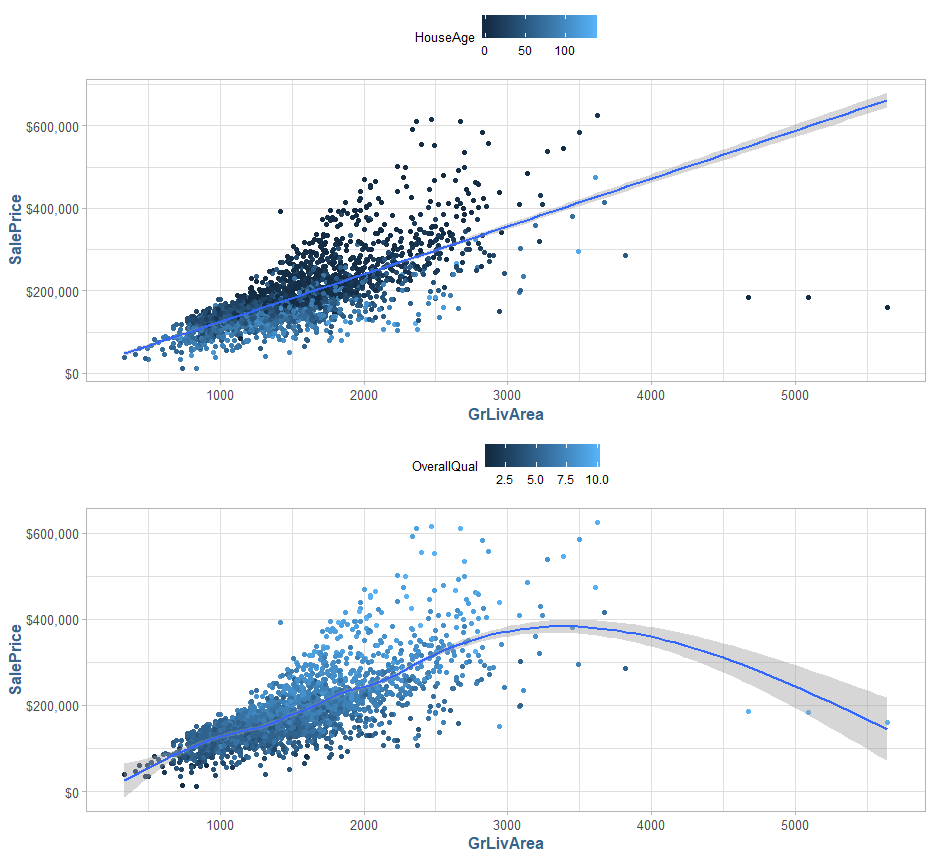
In the above top figure, we can see the year a house was built with a smooth fit against its sale price. The year built has a strong relationship to sale price if the home was built post 1940. In general, the year a home was remodeled shows a clear linear relationship to its sale price. These two variables show promise for predictability in the sale price of a home and should be considered when we are in the modeling phase.

For the continuous variables in the dataset, we can look at a scatterplot matrix to explore the relationships between them.

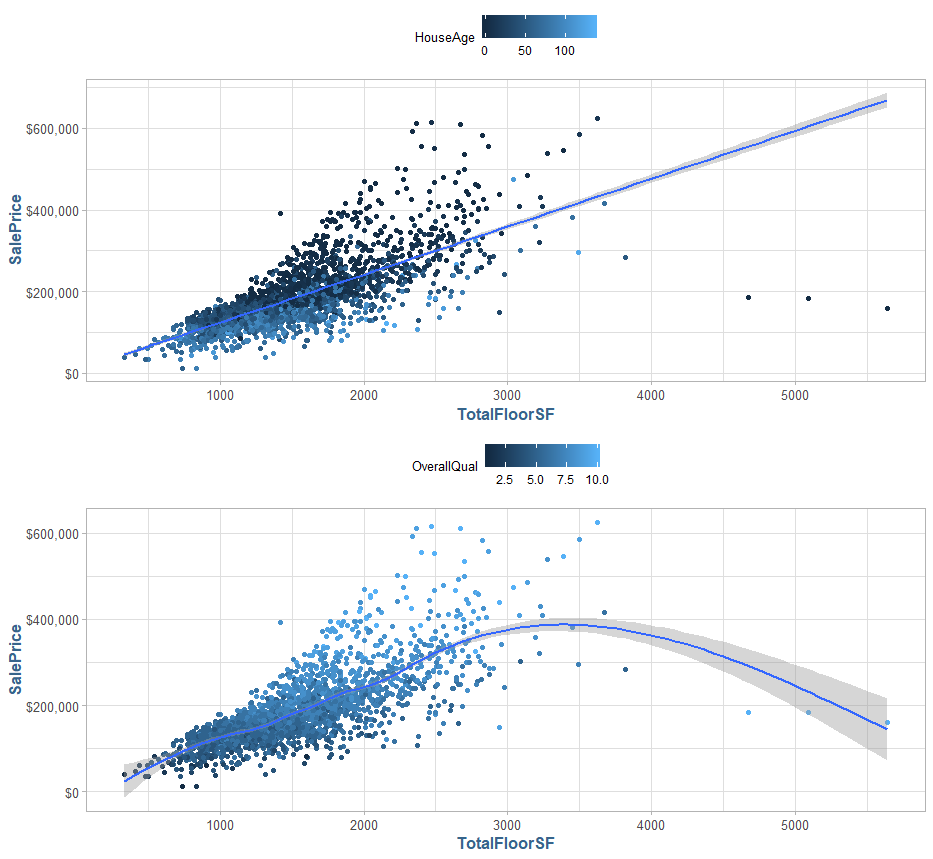


During our initial analysis of the variables contained in the dataset, we noted the high correlation of sale price to the continuous variables above ground living area (**GrLivArea**) and total square footage of the home (**TotalFloorSF**). To explore this relationship a bit deeper, we will have a look at the relationship between them. We will also note that the color scheme used in the relationship reflects the house age and quality features.

The following figure shows sale price as a function of above ground living area, with the linear model showing the overall positive trend and high correlation, as well as a more flexible (LOSSES) model that shows the potential for overfitting due to some outliers in the dataset.



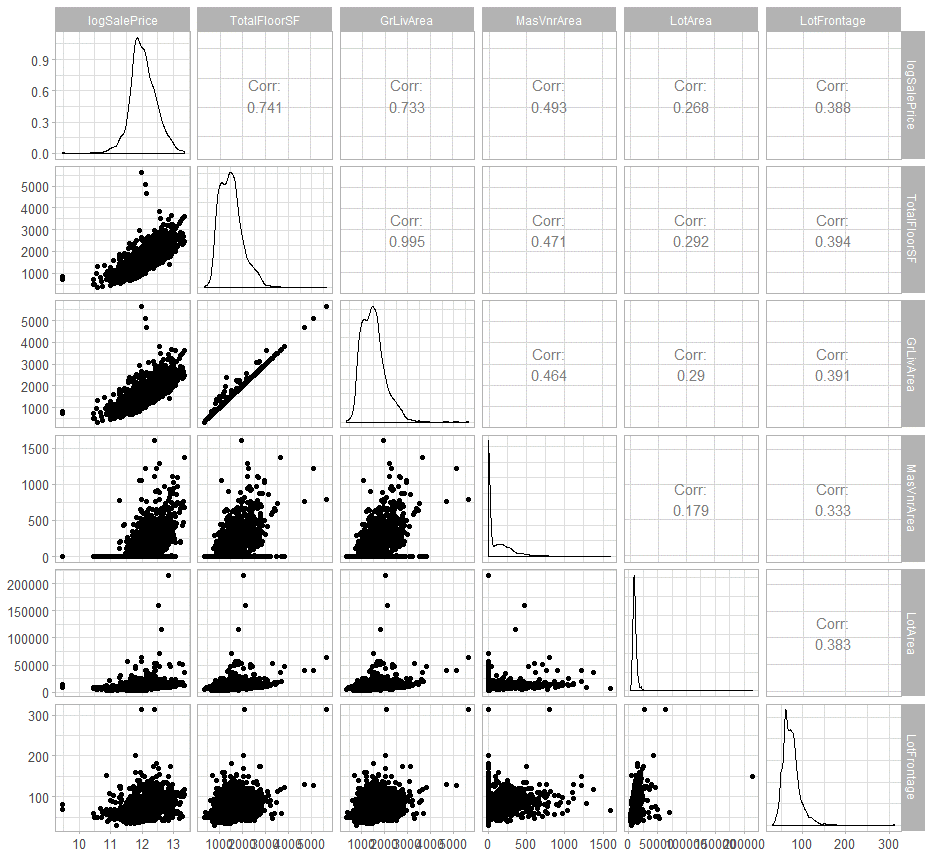
We can similarly look at sale price as a function of total square footage in the following figure:



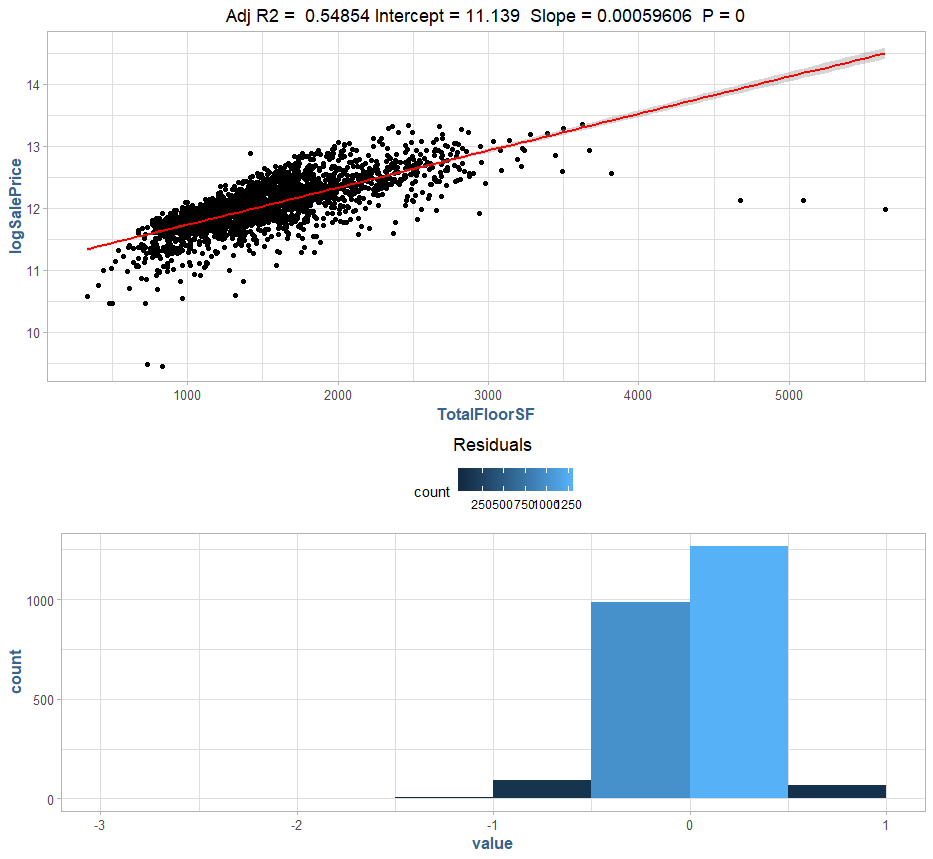
We should observe that in the top chart in both figures we see a strong negative correlation to the houses age depicted by the light to dark color scheme, and a strong positive correlation to the quality index denoted by the inverted dark to light color scheme in the bottom figure.

### Initial exploratory data analysis for modeling

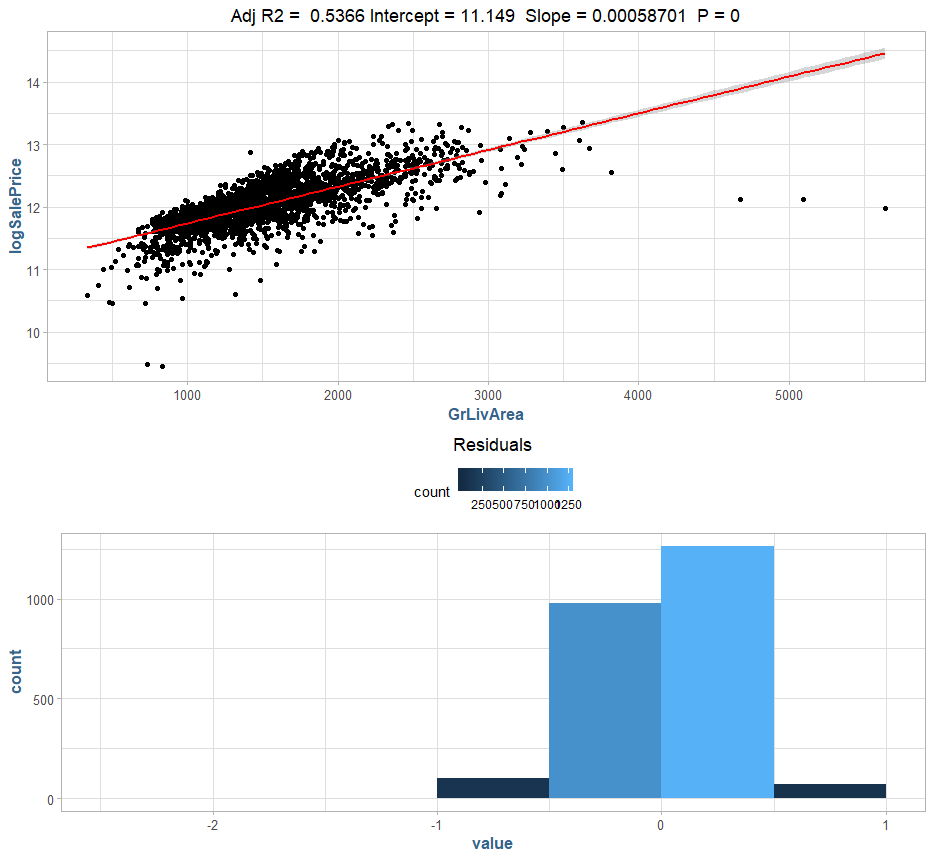
For the modeling phase of this analysis, we will use the log transformed sale price response variable. The features that exhibited the most predictive ability during the exploratory data analysis are shown in a scatterplot matrix against the response variable in the figure below:



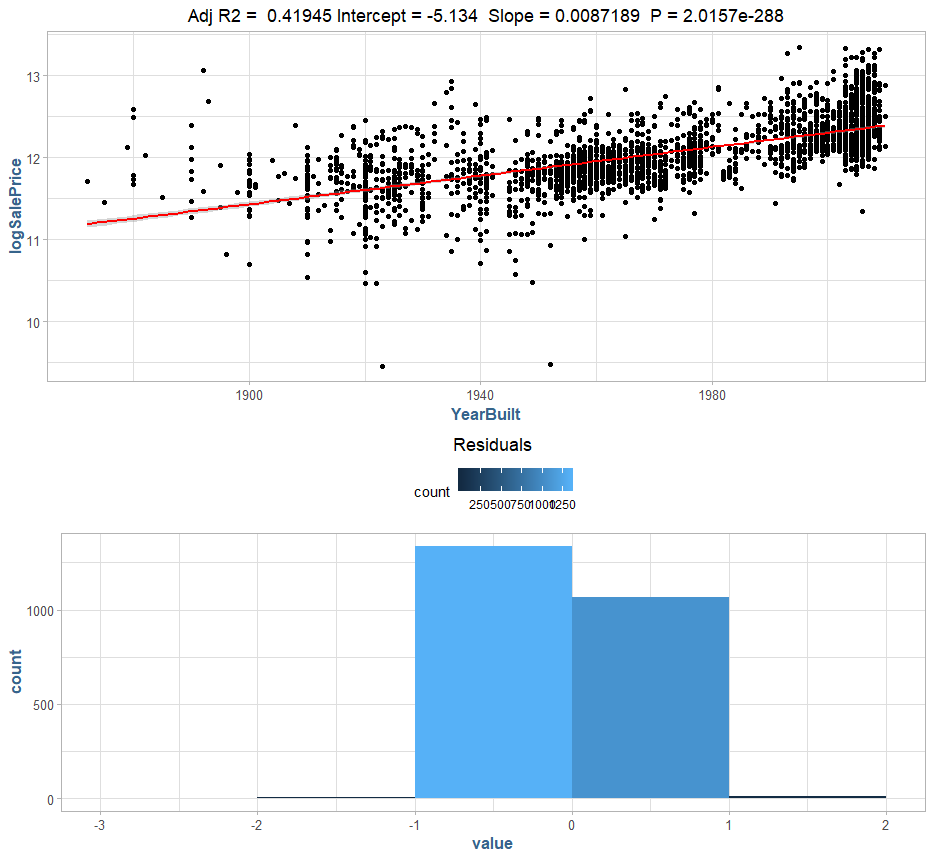
The first model we will work with is a simple linear model based on log sale price as a function of total square footage, which can be seen in the following figure:

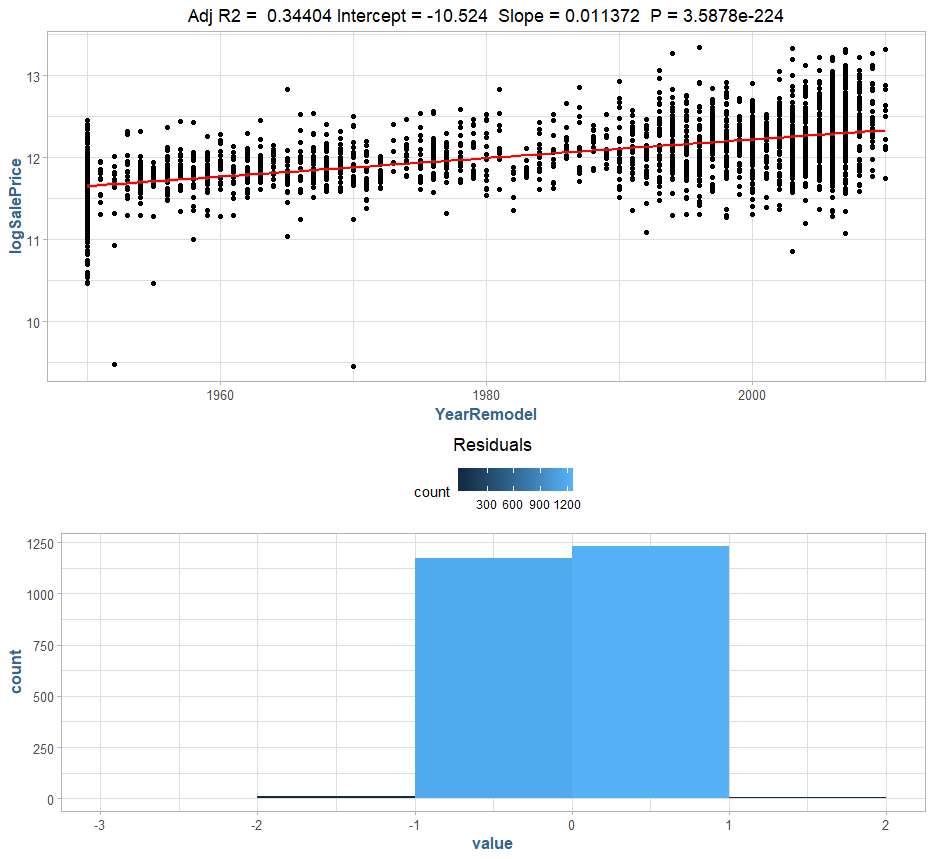


We note that the simple linear model fitted on the total square footage works for a segment of the sample, however, exhibits a great deal of error in many cases. Next, we will look at a model fitted to the above ground living area (**GrLivArea**) in the following figure:

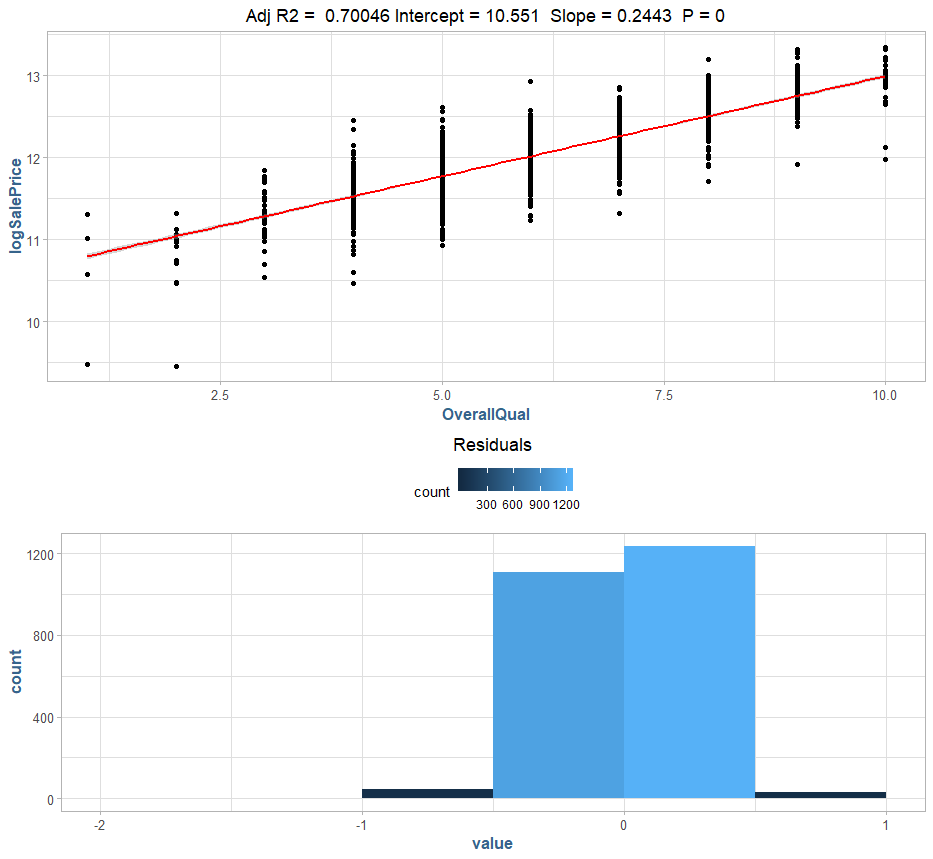


This model performs similarly to the total square footage model we looked at previously, although it does perform slightly better in some cases. The next two models are based upon the more linear relationships we discovered in our exploratory phase, the year the house was built (**YearBuilt**) and the year it was remodeled (**YearRemodel**), which we will see in the following two figures:

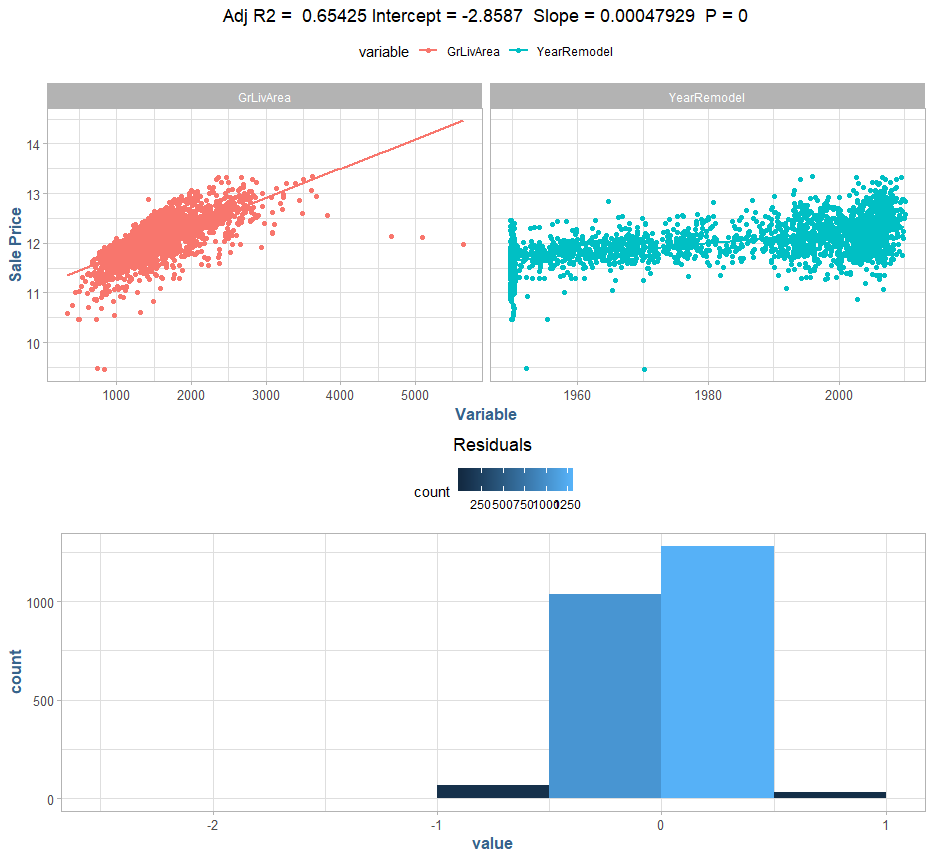




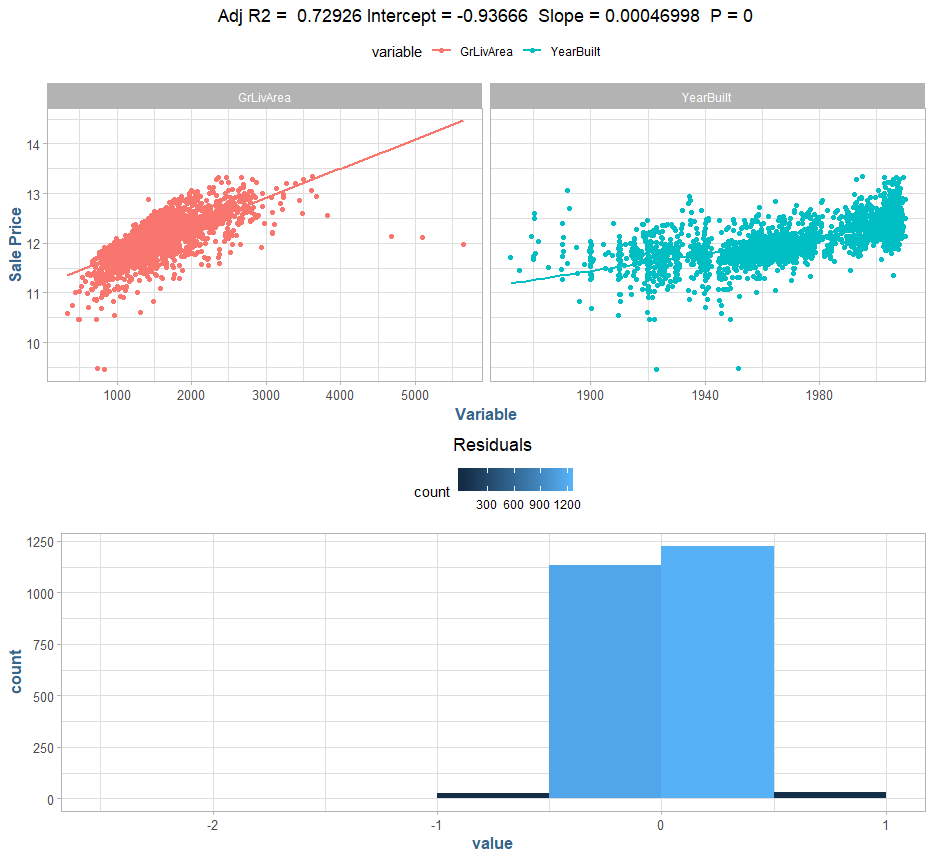
These two models perform substantially better based on the temporal variables than the previous models based on the continuous variables relating to housing area. The final single variable linear model we will look at is one built using the highest correlation variable we saw during out initial exploration, overall quality:



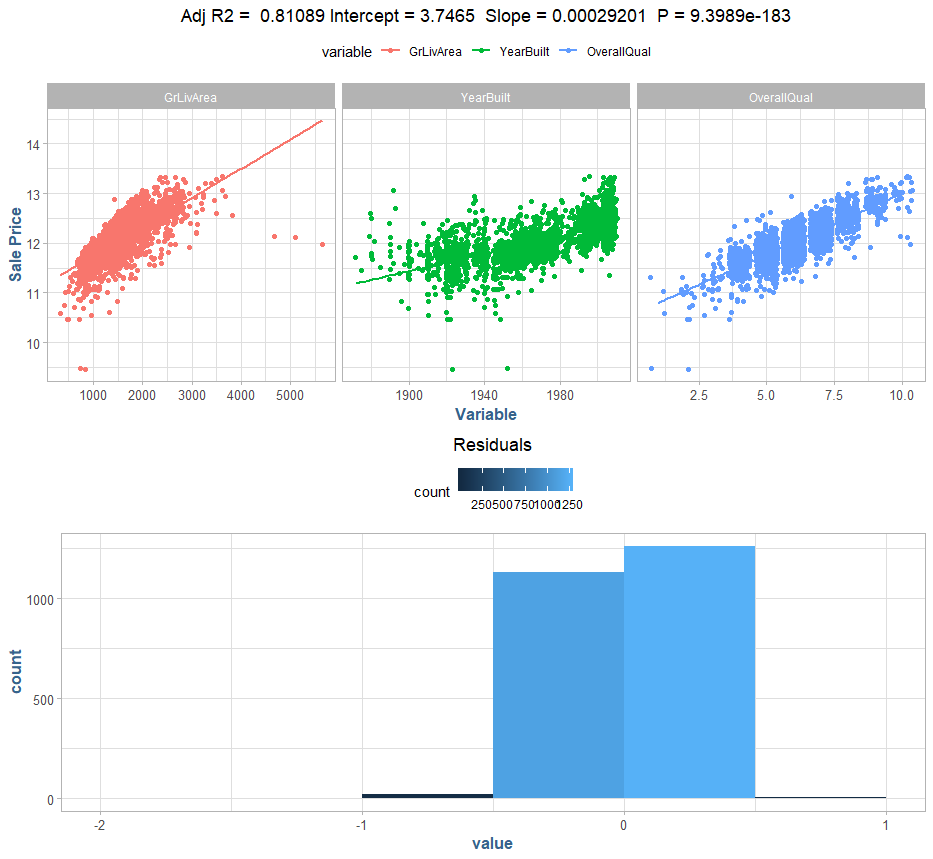
The results of the single linear regression are overall less than promising; however, we can use multiple variables to help accommodate for the unexplained bias using a standard linear model. In the following figure we will look at two promising variables from the previous section, above ground living area (**GrLivArea**) and year remodel (**YearRemodel**):



This model doesn’t perform as well as the single variable model using overall quality index (**OverallQual**). Expanding this model to use the more linear variable year built (**YearBuilt**), we can see a slight increase in performance:



The final model we will examine in this lab is a multiple linear regression model fitted to use the top three variables in each class, namely above ground living area (**GrLivArea**) for continuous, year built (**YearBuilt**) temporal discrete, and overall quality (**OverallQual**) ordinal variable. The result is depicted in the following figure:

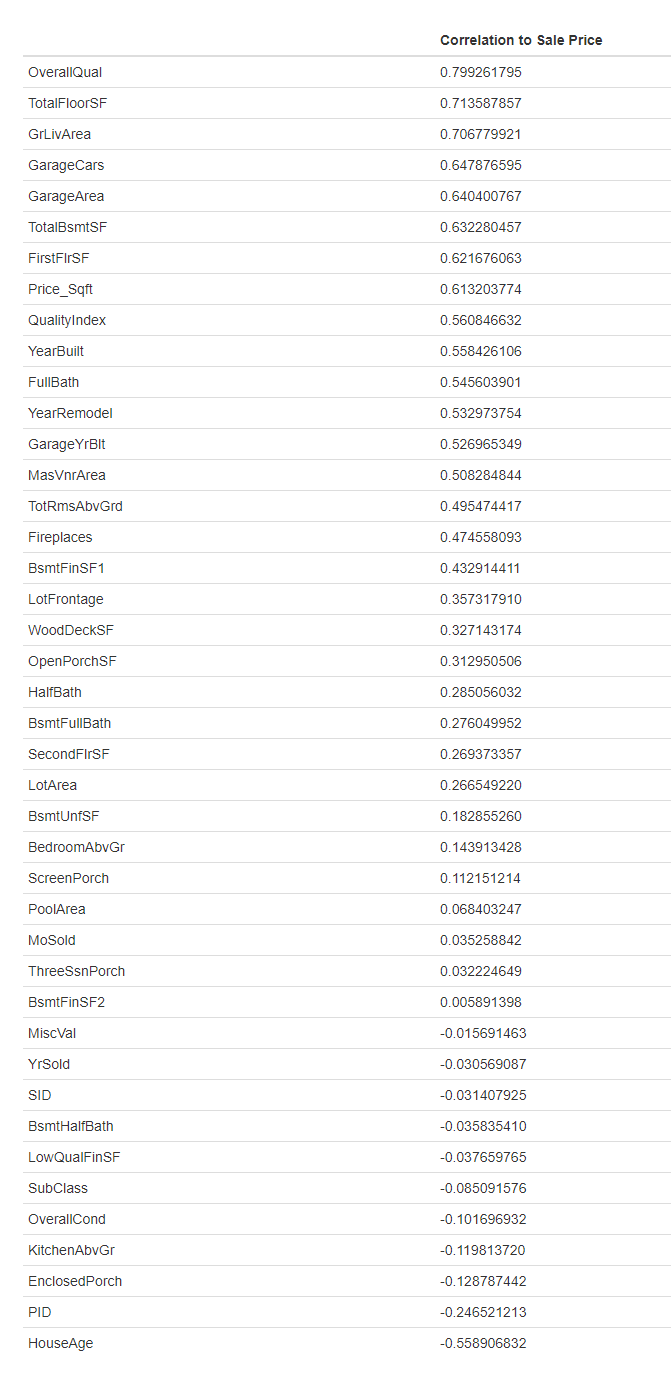


### Conclusion

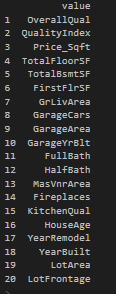
In this lab we have performed a data survey, defined a sample population, executed a data quality check on selected features, basic exploratory data analysis and initial model formulation based upon our findings. The data suggests there are indeed fundamental relationships between the housing explanatory variables and the desired response, the sale price. Amongst the sample population we defined, single-family residences with a sale price less than $600,000, we see a semi-colinear relationship from the overall square footage and above ground living area, a stronger colinear relationships to the temporal discrete variables’ year built and year remodel and as well as a strong colinear relationship to the overall quality ordinal variable. 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.

### Appendex

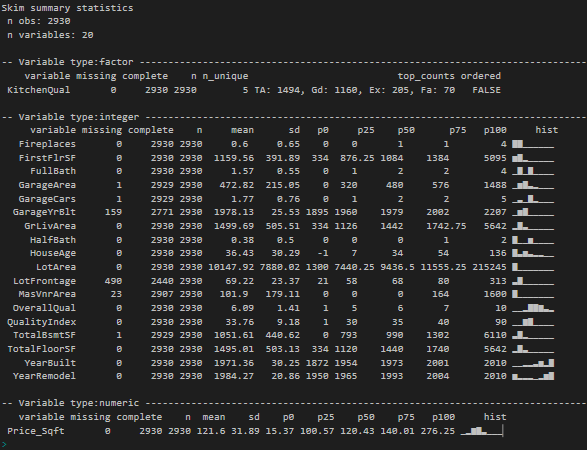
#### Feature correlation



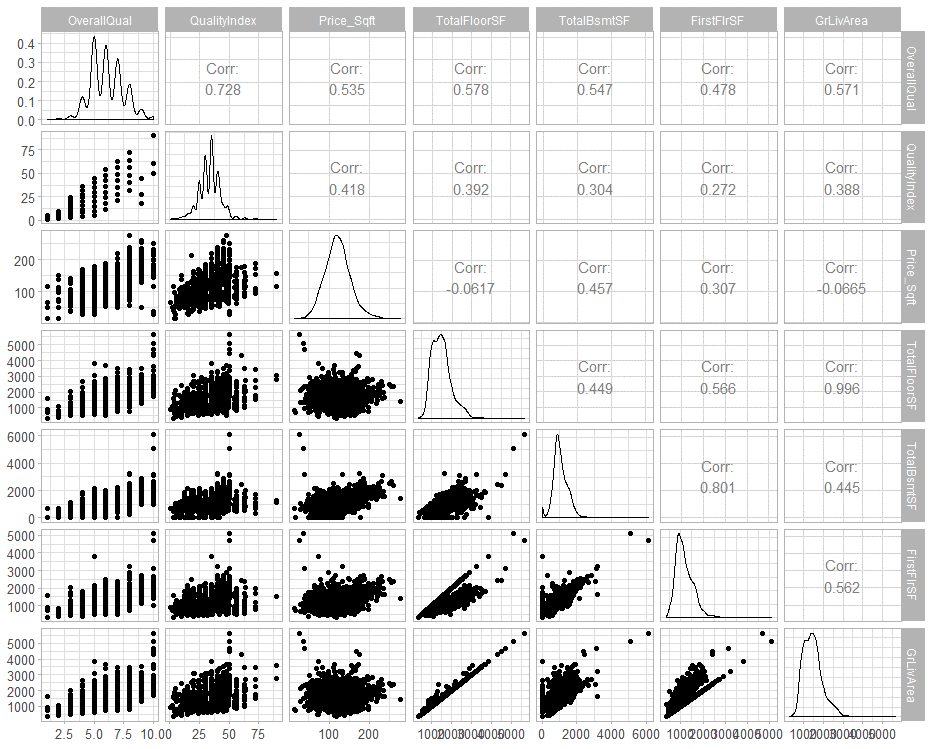
#### Data Quality Check



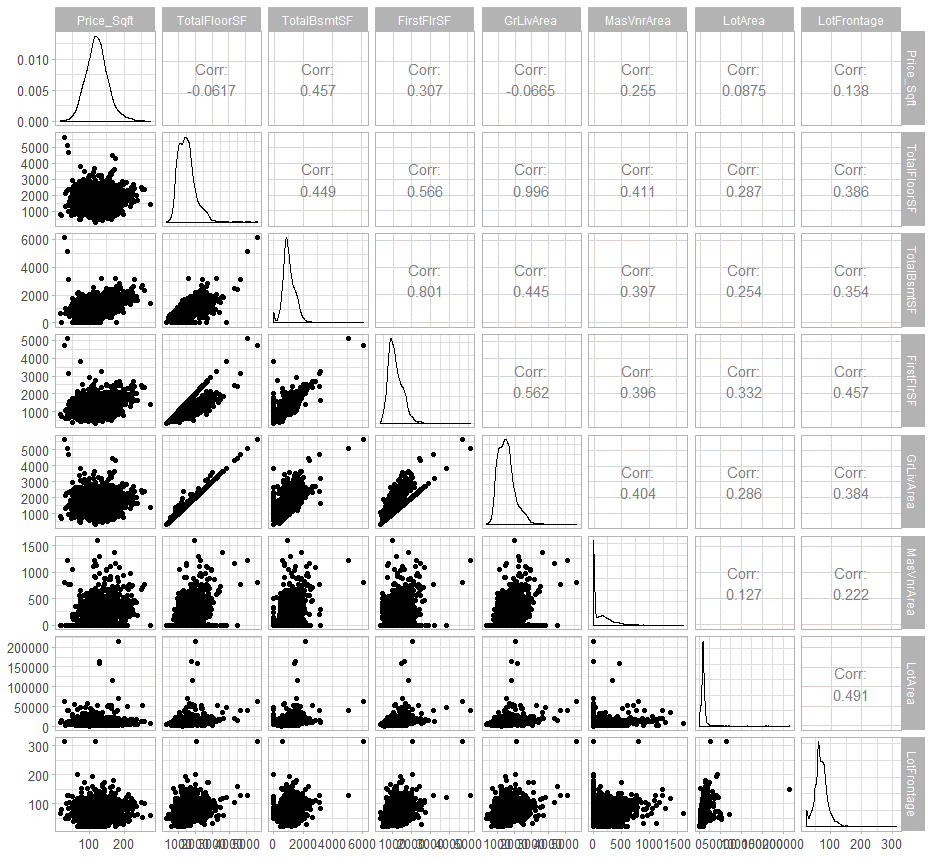
Skim



High Impact



Continuous



Temporal

