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.

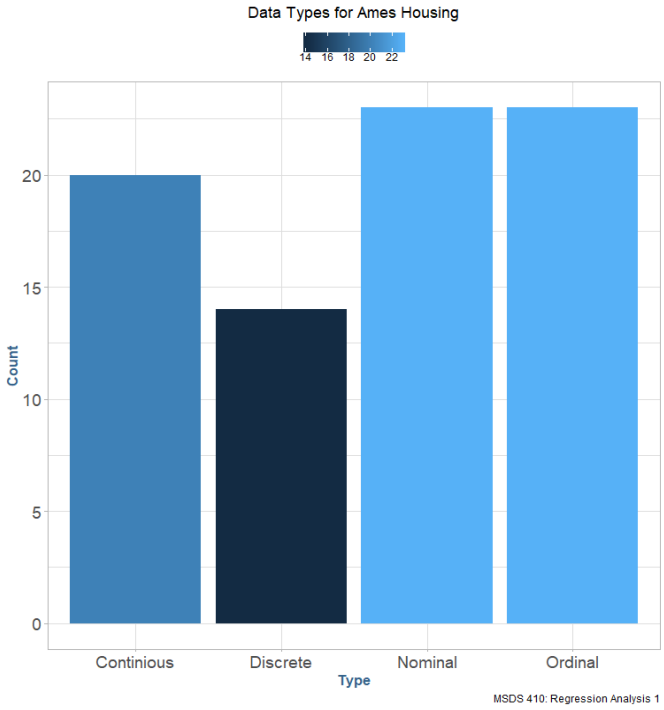
In this report, we will use the Ames dataset which is an alternative to the famous Boston housing data to perform exploratory data analysis through variable derivation, validation, selection and visualization to measure the relevance of these indicators as they pertain to the value of the home in terms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

	Correlation to Sale Price
OverallQual	0.79926179
TotalFloorSF	0.71358786
GrLivArea	0.70677992
GarageCars	0.64787660
GarageArea	0.64040077
TotalBsmtSF	0.63228046
FirstFlrSF	0.62167606
Price_Sqft	0.61320377
QualityIndex	0.56084663
YearBuilt	0.55842611
FullBath	0.54560390
YearRemodel	0.53297375
GarageYrBlt	0.52696535
MasVnrArea	0.50828484
Fireplaces	0.47455809
LotFrontage	0.35731791
HalfBath	0.28505603
LotArea	0.26654922
BedroomAbvGr	0.14391343
PoolArea	0.06840325
HouseAge	-0.55890683

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

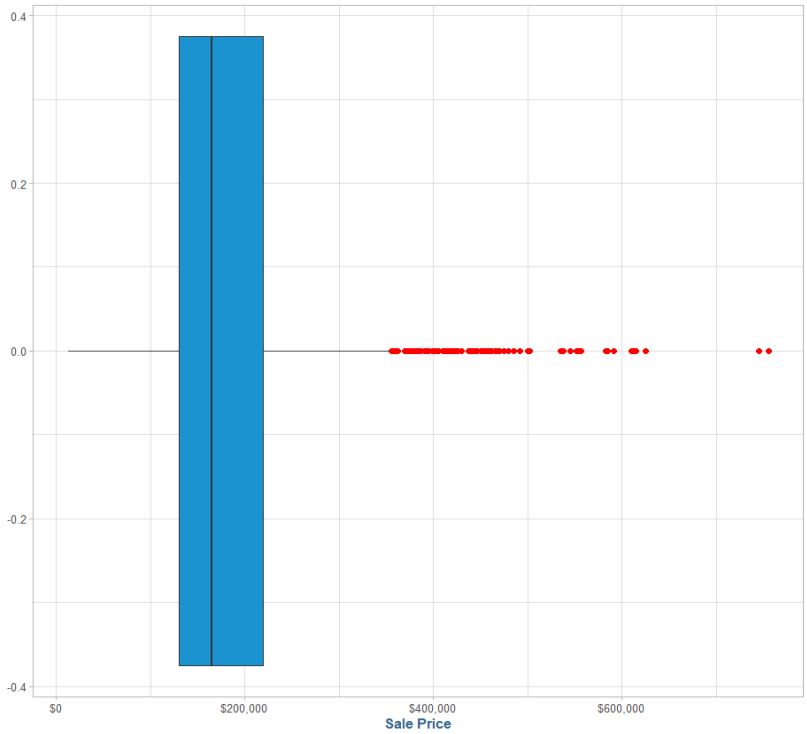
To check the similarity to other property types we can examine the average sale price by building type in the following table:

BldgType	Count	Pct	AvgSalePrice
1Fam	2425	82.76%	$184,812.04
TwnhsE	233	7.95%	$192,311.91
Twnhs	101	3.45%	$135,934.06
Duplex	109	3.72%	$139,808.94
2fmCon	62	2.12%	$125,581.71

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**.

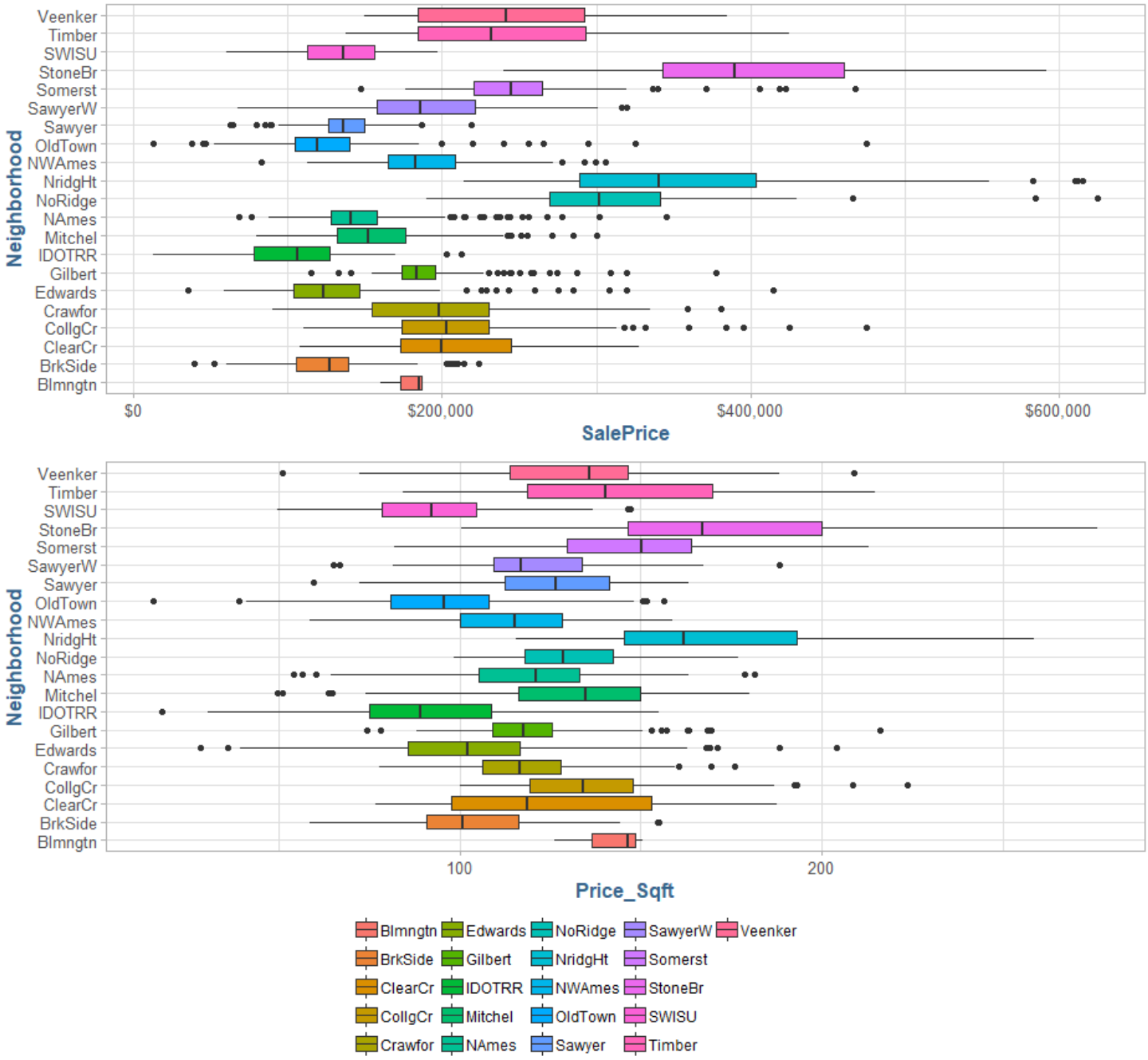
At a high level we can group these features into the following categories: temporal discrete (**YearBuilt**, **YearRemodel, GarageYrBlt**), discrete (**FullBath**, **HalfBath**, **Fireplaces**, **GarageCars, HouseAge**), continuous (**Price_Sqft, TotalFloorSF, TotalBsmtSF, FirstFloorSF, GrLivArea, MasVnrArea,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

Further data quality checks on a subset of selected features can be found in the appendix.

INITIAL EXPLORATORY DATA ANALYSIS

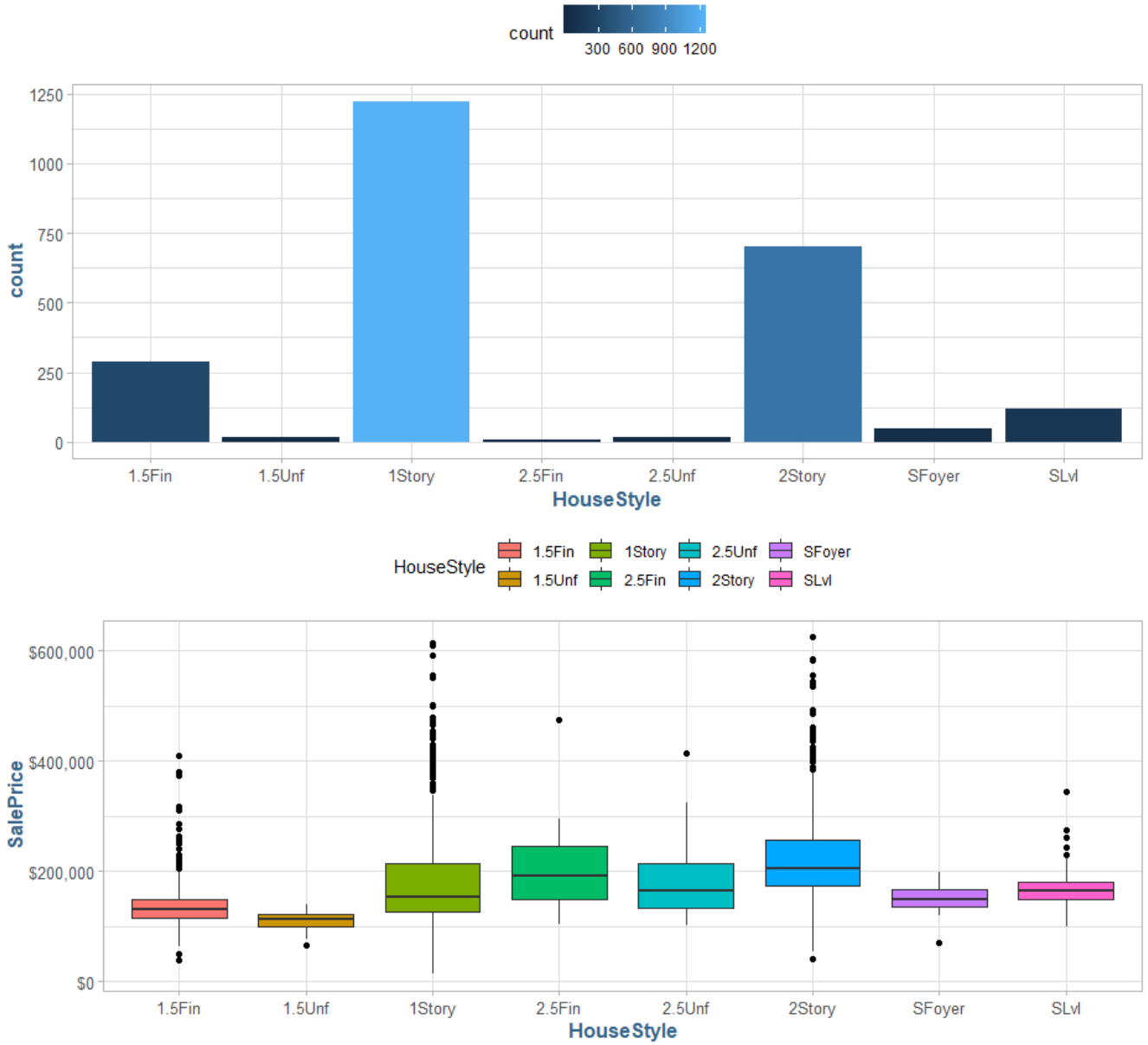
For initial exploratory data analysis, we will choose ten features from the housing dataset that have either a high correlation to sale price (for numeric variables) or have an intuitively strong connection to the sale price (categorical variables such as neighborhood or masonry veneer type).

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:



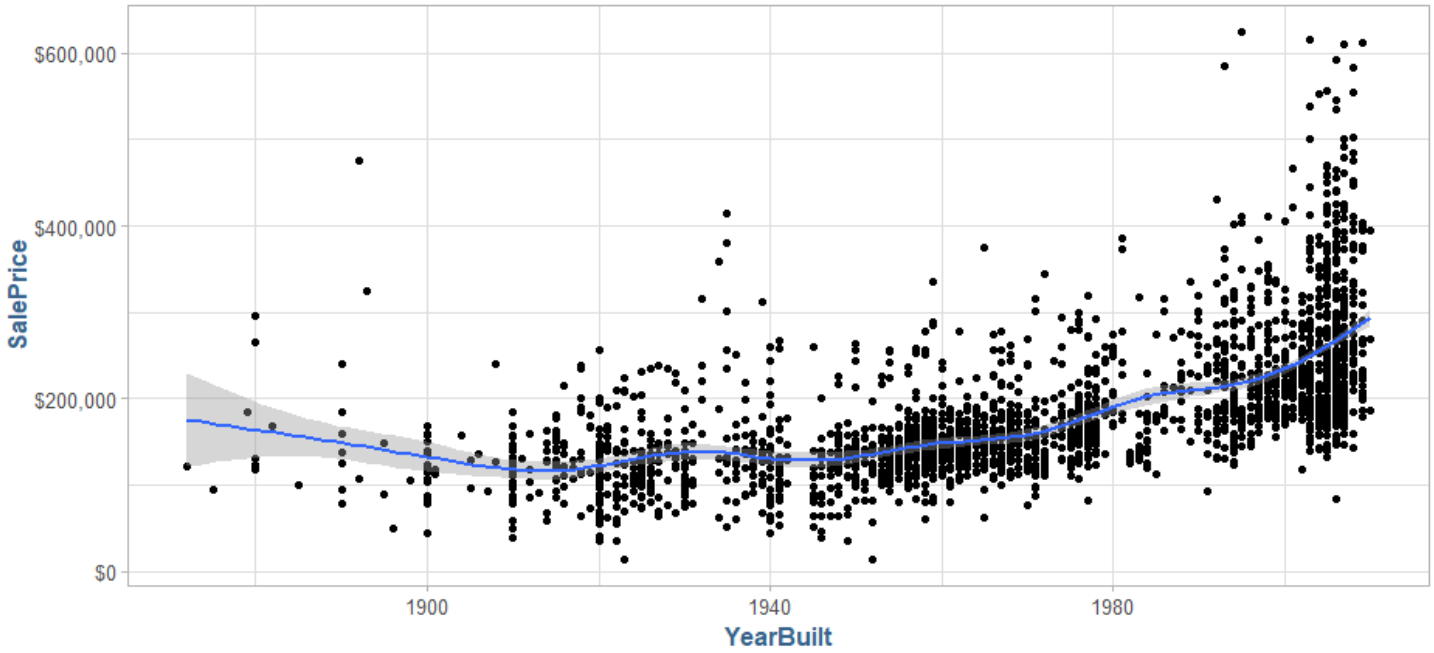
Unfortunately, the distributions are highly scattered and show minimal clustering and therefore will likely lead to low predictive value.

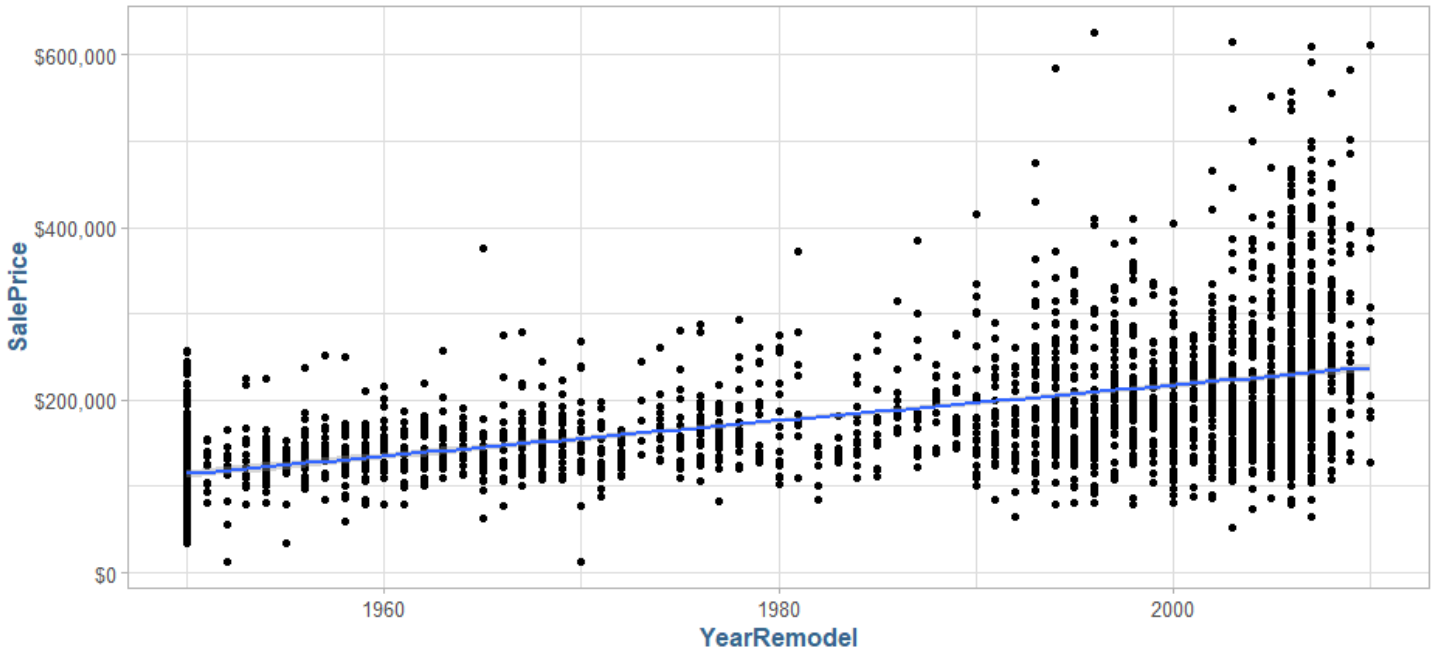
Another categorical variable that has potential for predictability in sale price intuitively is the housing style, as variations in home type should have similar pricing points. The following figure shows the distribution of housing style, and how sale price is distributed amongst the different housing 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.

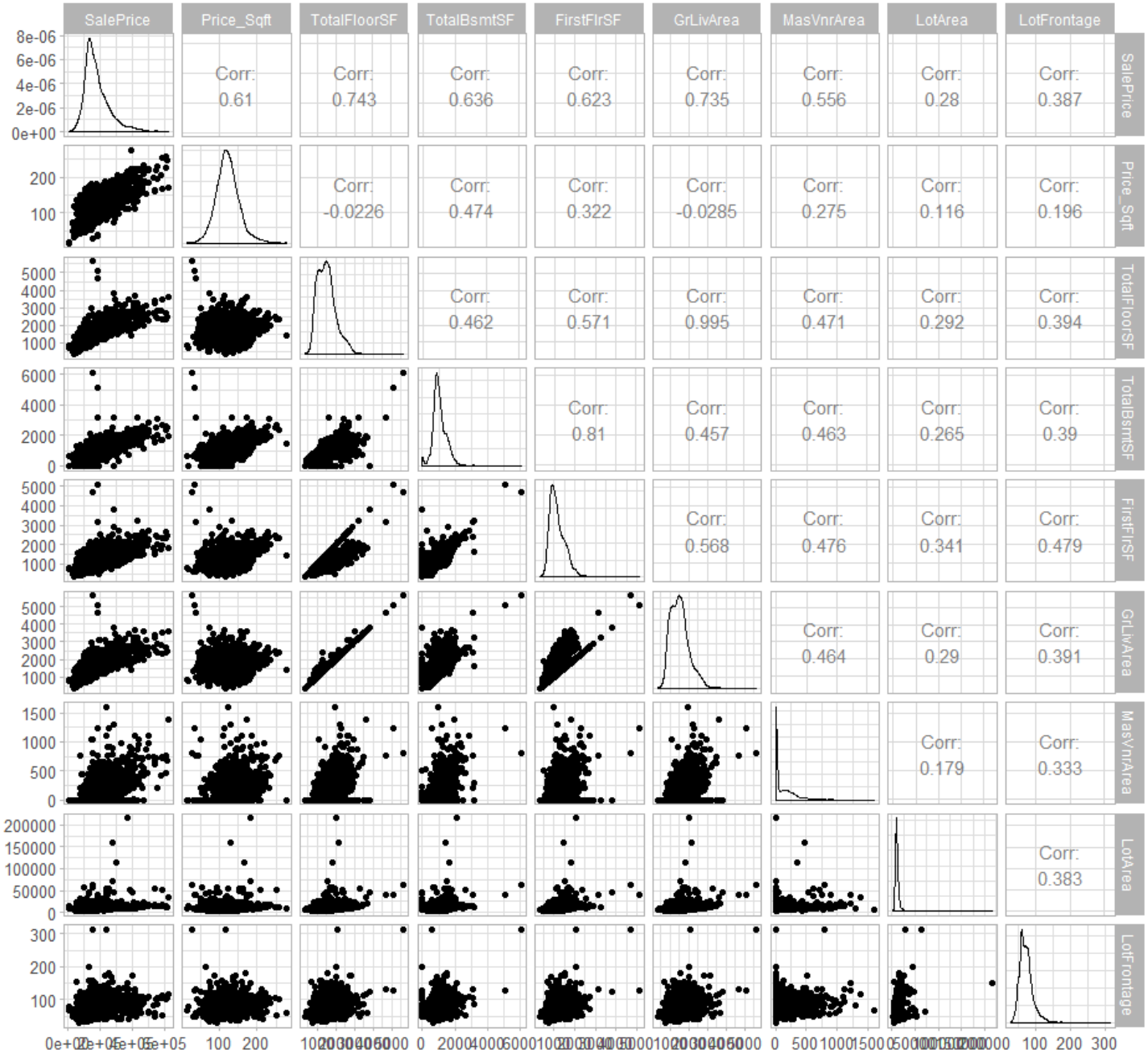
Another set of variables we should explore are the discrete variables, such as year built (**YearBuilt**) and year remodeled (**YearRemodel**). For these variables, we can look at a scatterplot of the year versus the sale price to see if there is a relationship:





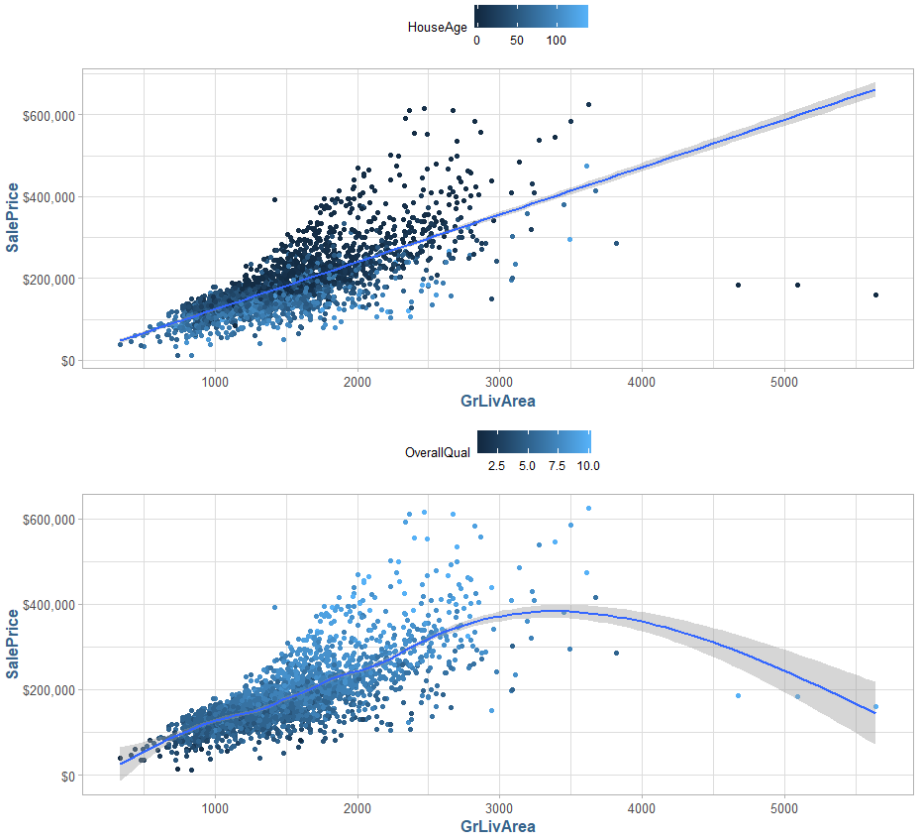
In the above top figure, we can see the year a house was built with a smooth fit against its sale price. The year built has a strong relationship to sale price if the home was built post 1940. In general, the year a home was remodeled shows a clear linear relationship to its sale price. These two variables 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.

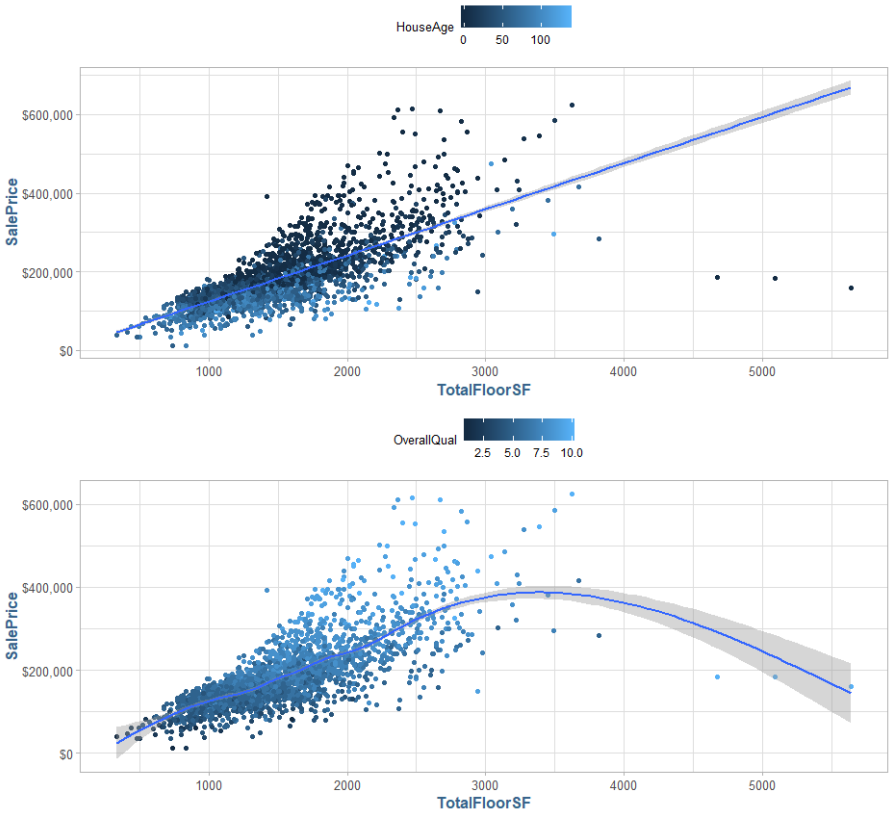


During our initial analysis of the variables contained in the dataset, we noted the high correlation of sale price to the continuous variables above ground living area (**GrLivArea**) and total square footage of the home (**TotalFloorSF**). To explore this relationship a bit deeper, we will have a look 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.



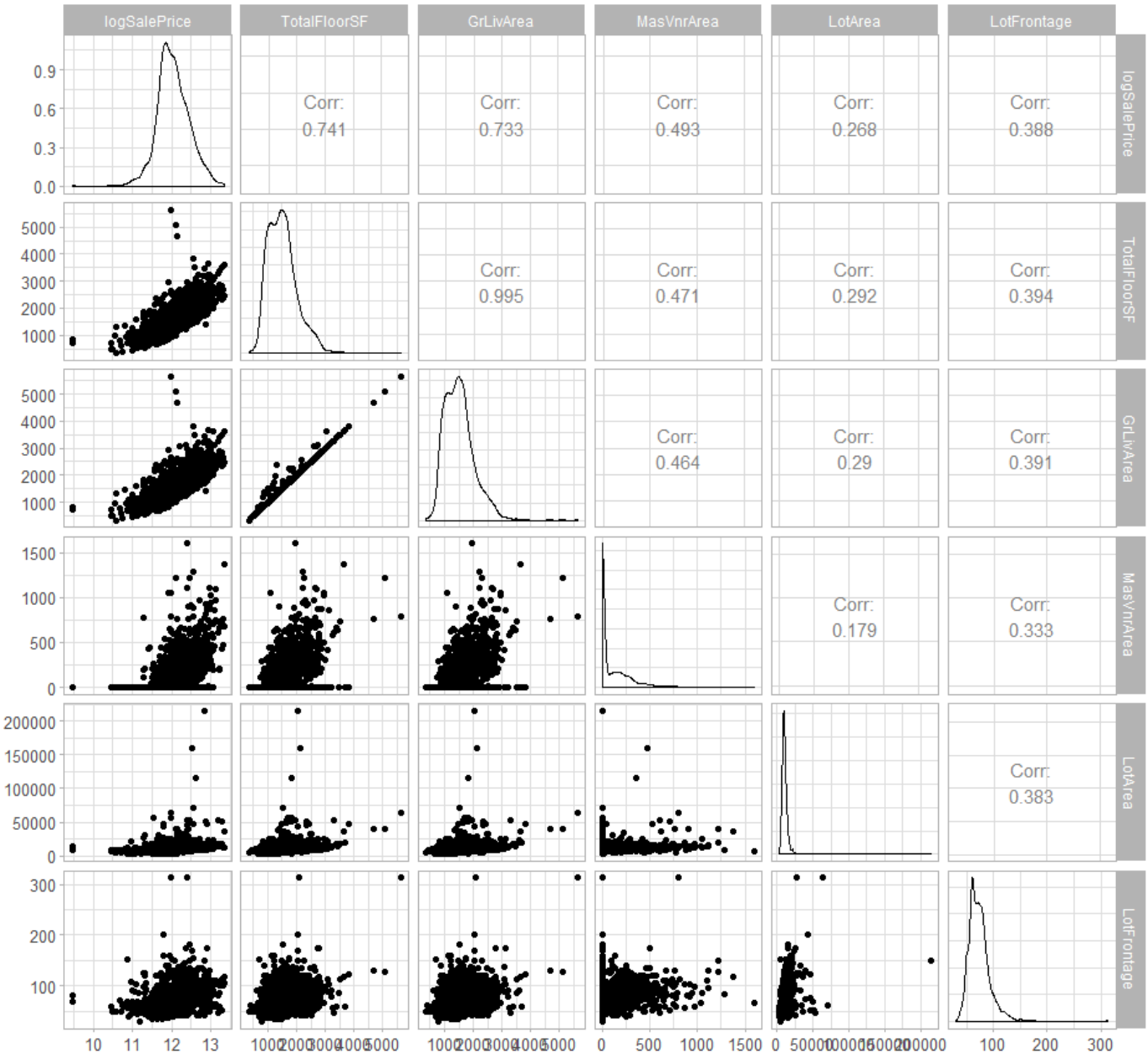
We can similarly look at sale price as a function of total square footage in the following figure:



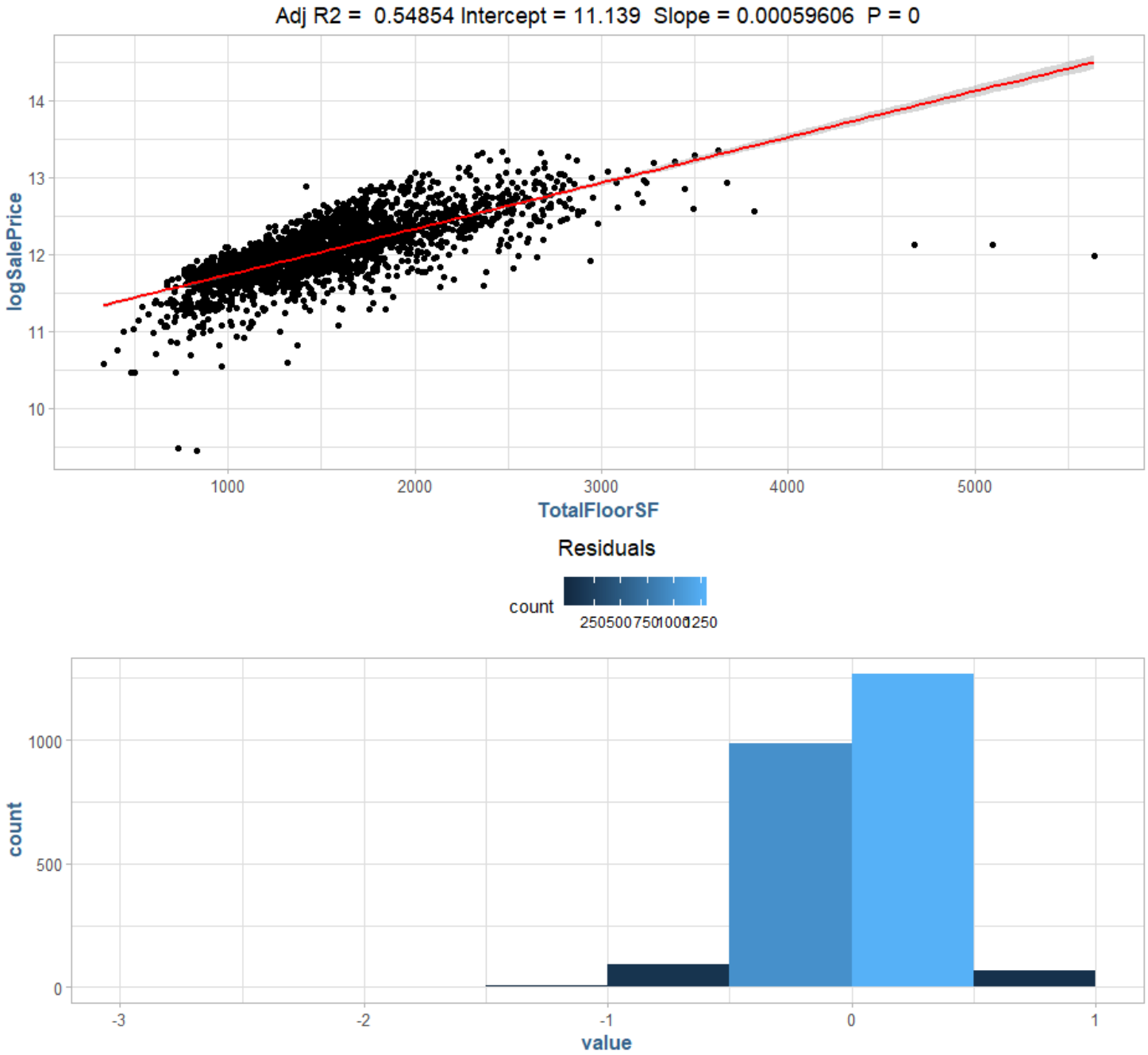
We should observe that in the top chart in both figures we see a strong negative correlation to the houses age depicted by the light to dark color scheme, and a strong positive correlation to the quality index denoted by the inverted dark to light color scheme in the bottom figure.

INITIAL EXPLORATORY DATA ANALYSIS FOR MODELING

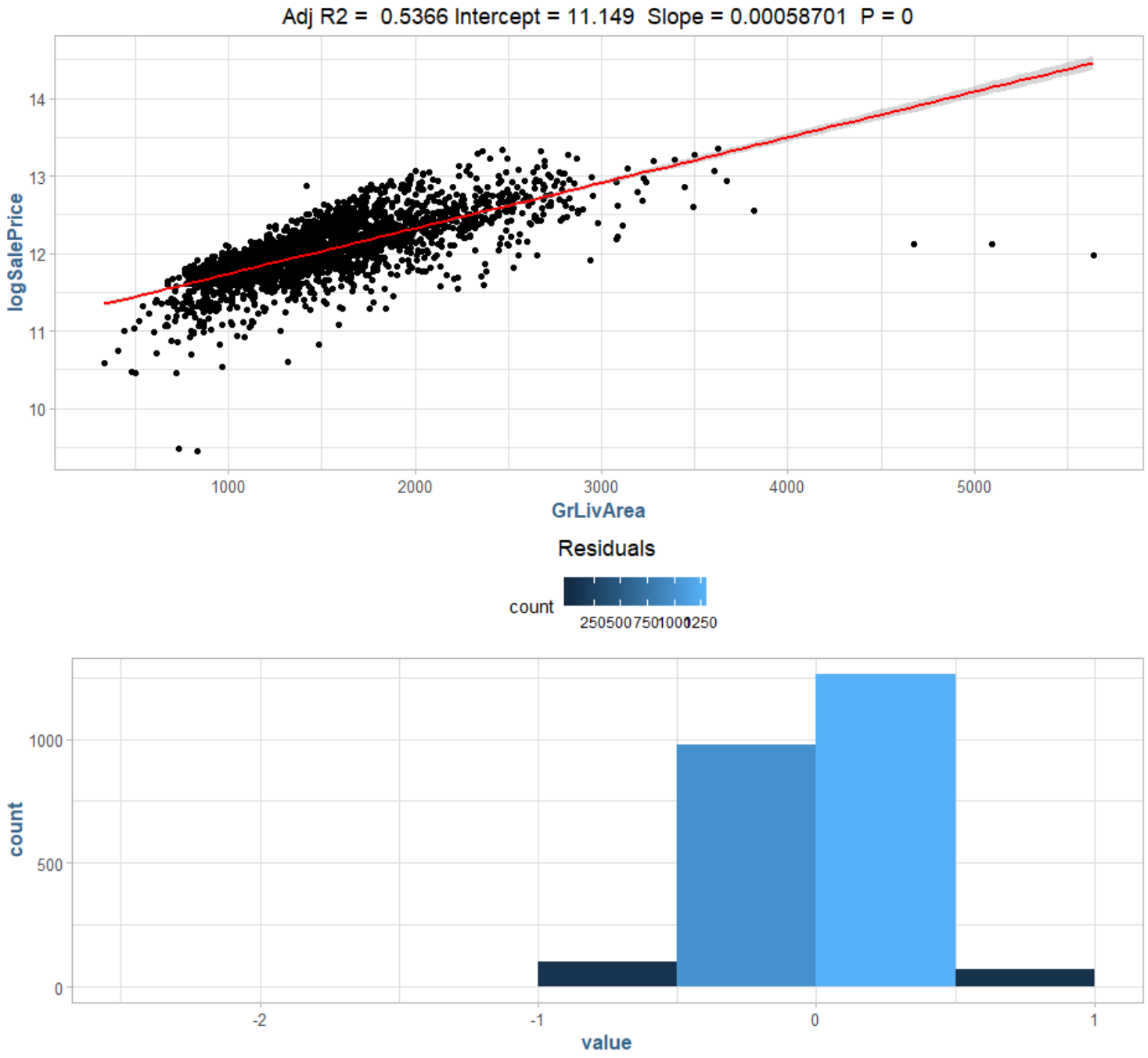
For the modeling phase of this analysis, we will use the log transformed sale price response variable. The features that exhibited the most predictive ability during the exploratory data analysis are shown in a scatterplot matrix against the response variable in the figure below:



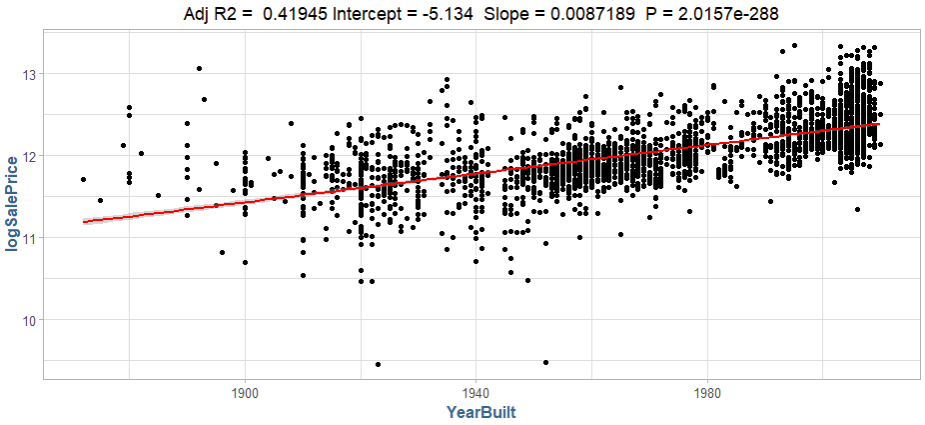
The first model we will work with is a simple linear model based on log sale price as a function of total square footage, which can be seen in the following figure:

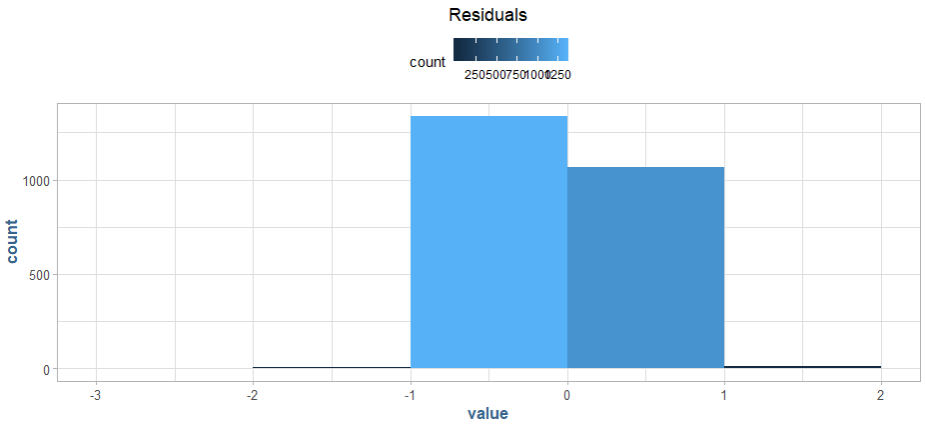


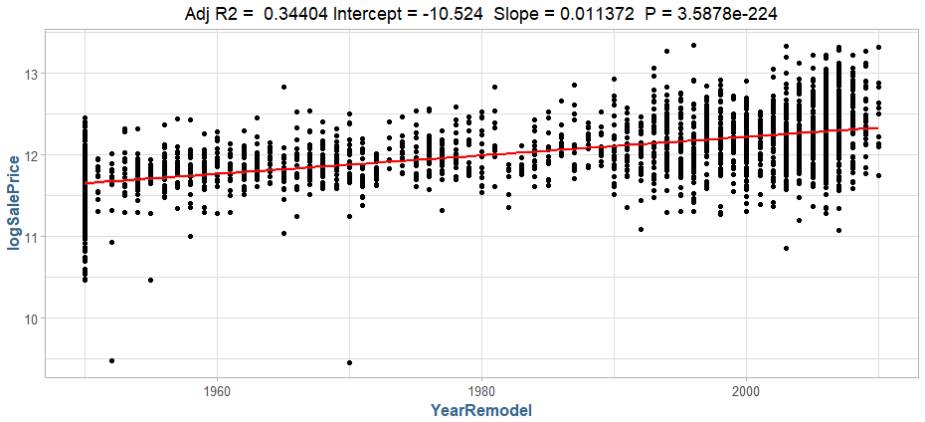
We note that the simple linear model fitted on the total square footage works for a segment of the sample, however, exhibits a great deal of error in many cases. Next, we will look at a model fitted to the above ground living area (**GrLivArea**) in the following figure:

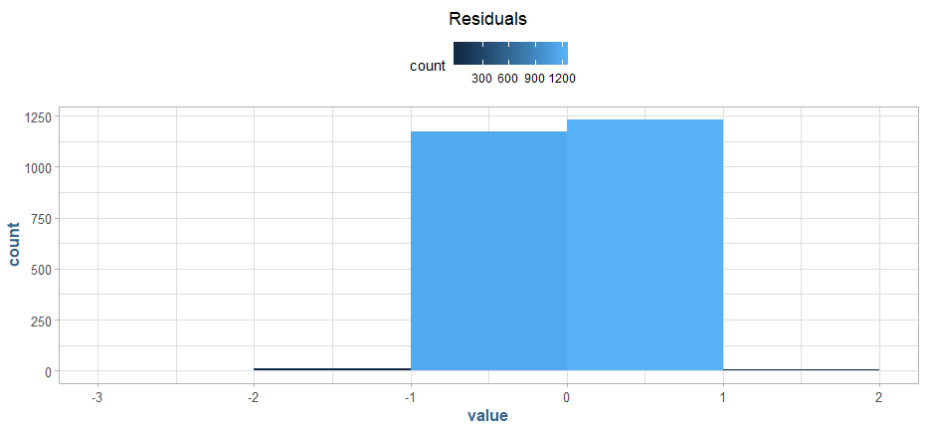


This model performs similarly to the total square footage model we looked at previously, although it does perform slightly better in some cases. The next two models are based upon the more linear relationships we discovered in our exploratory phase, the year the house was built (**YearBuilt**) and the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

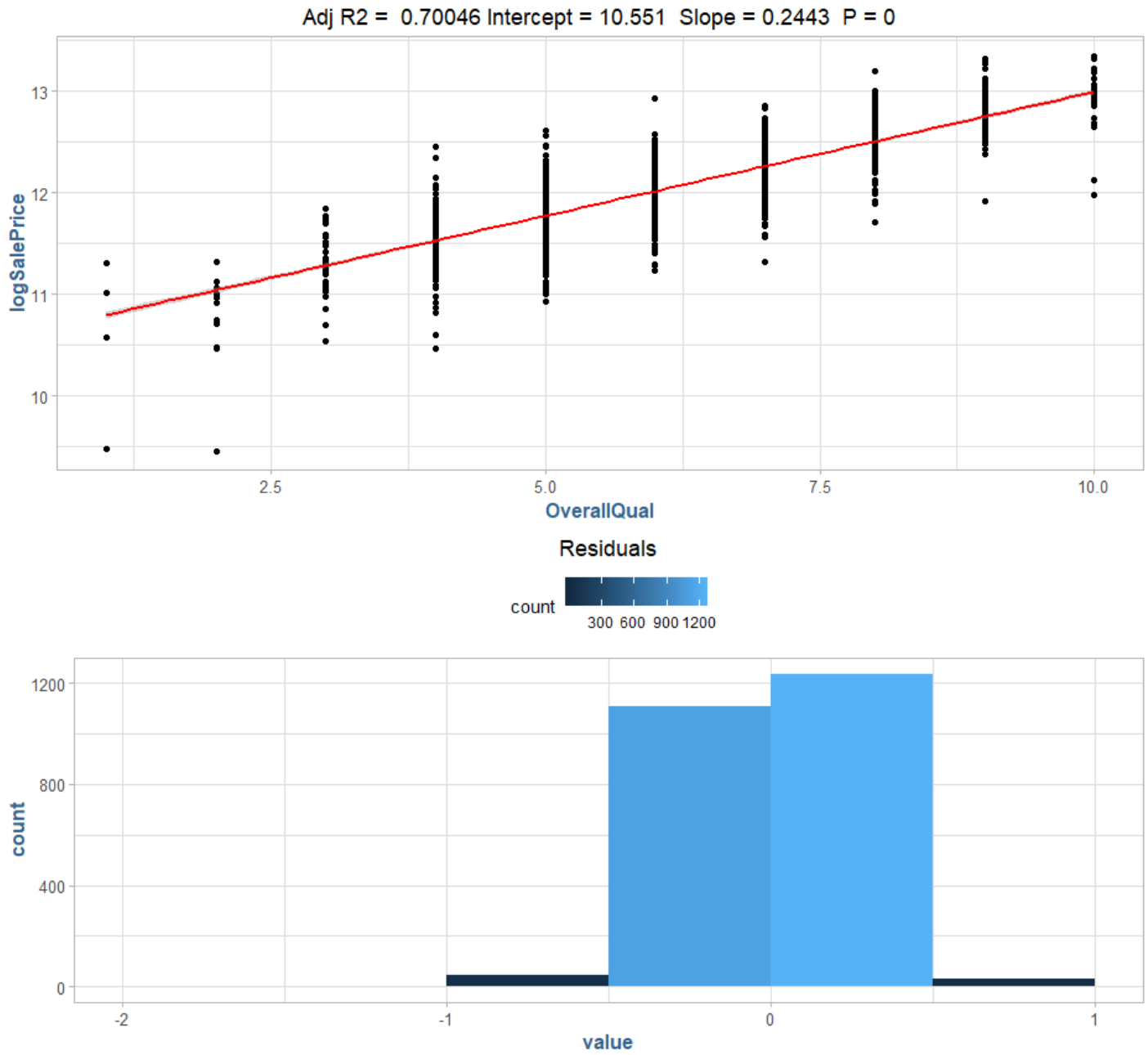

Adj R2 = 0.41945 Intercept = -5.134 Slope = 0.0087189 P = 2.0157e-288
logSalePrice
13
12
11
10
1900
1940
1980
YearBuilt


Residuals
count
250500750100012250
count
1000
500
0
-3
-2
-1
0
1
2
value

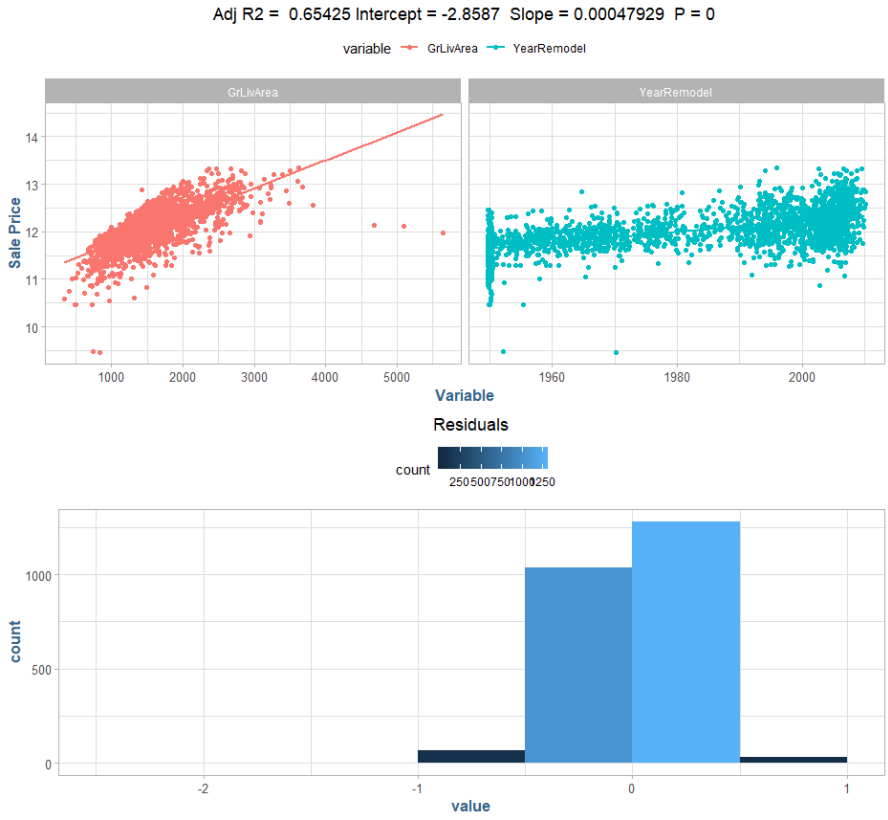

Adj R2 = 0.34404 Intercept = -10.524 Slope = 0.011372 P = 3.5878e-224
logSalePrice
13
12
11
10
1960
1980
2000
YearRemodel


Residuals
count
300 600 900 1200
count
1250
1000
750
500
250
0
-3
-2
-1
0
1
2
value

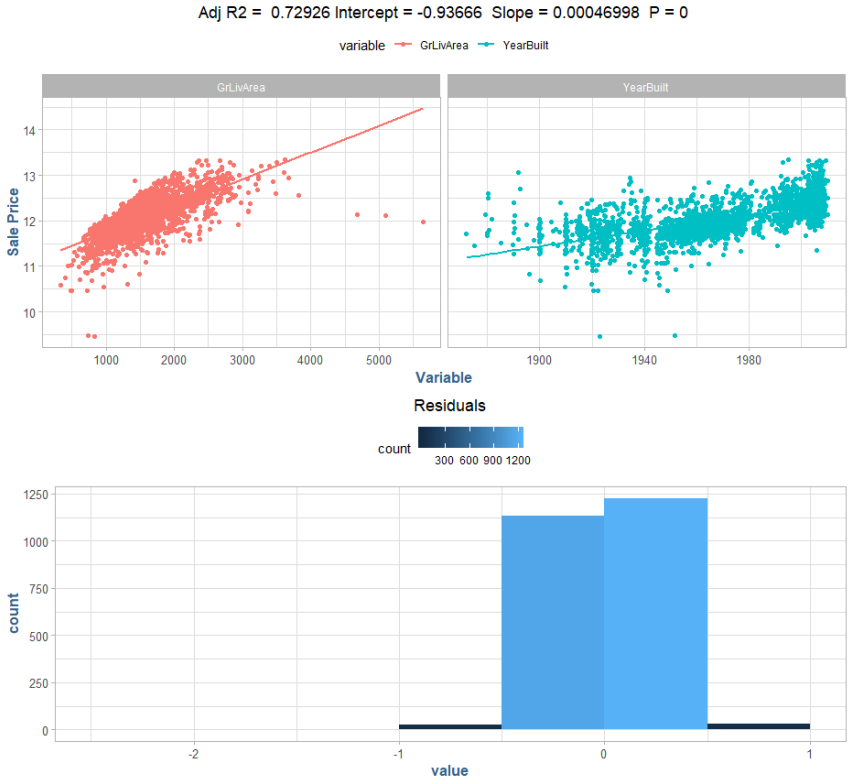
These two models perform substantially better based on the temporal variables than the previous models based on the continuous variables relating to housing area. The final single variable linear model we will look at is one built using the highest correlation variable we saw during out initial exploration,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



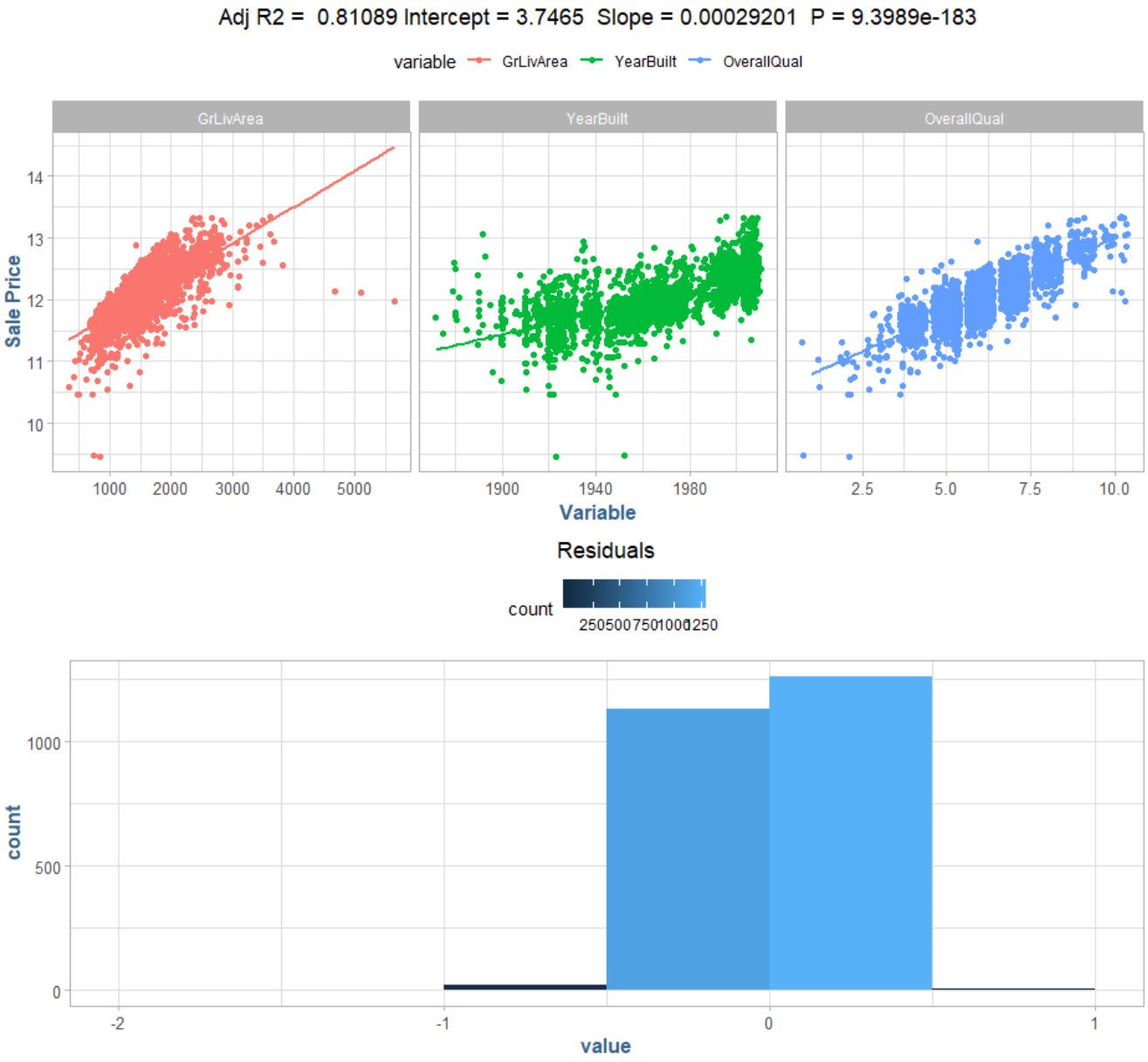
The results of the single linear regression are overall less than promising; however, we can use multiple variables to help accommodate for the unexplained bias using a standard linear model. In the following figure we will look at two promising variables from the previous section, above ground living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

FEATURE CORRELATION

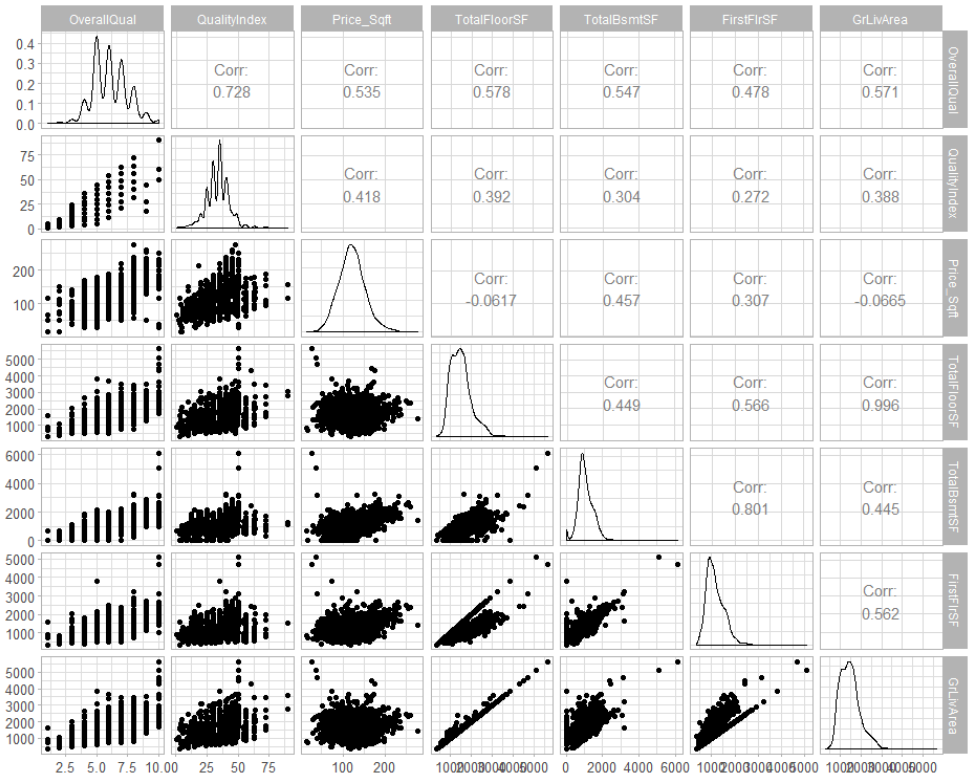
	Correlation to Sale Price
OverallQual	0.799261795
TotalFloorSF	0.713587857
GrLivArea	0.706779921
GarageCars	0.647876595
GarageArea	0.640400767
TotalBsmtSF	0.632280457
FirstFlrSF	0.621676063
Price_Sqft	0.613203774
QualityIndex	0.560846632
YearBuilt	0.558426106
FullBath	0.545603901
YearRemodel	0.532973754
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

	value
1	OverallQual
2	QualityIndex
3	Price_Sqft
4	TotalFloorSF
5	TotalBsmtSF
6	FirstFlrSF
7	GrLivArea
8	GarageCars
9	GarageArea
10	GarageYrBlt
11	FullBath
12	HalfBath
13	MasVnrArea
14	Fireplaces
15	KitchenQual
16	HouseAge
17	YearRemodel
18	YearBuilt
19	LotArea
20	LotFrontage

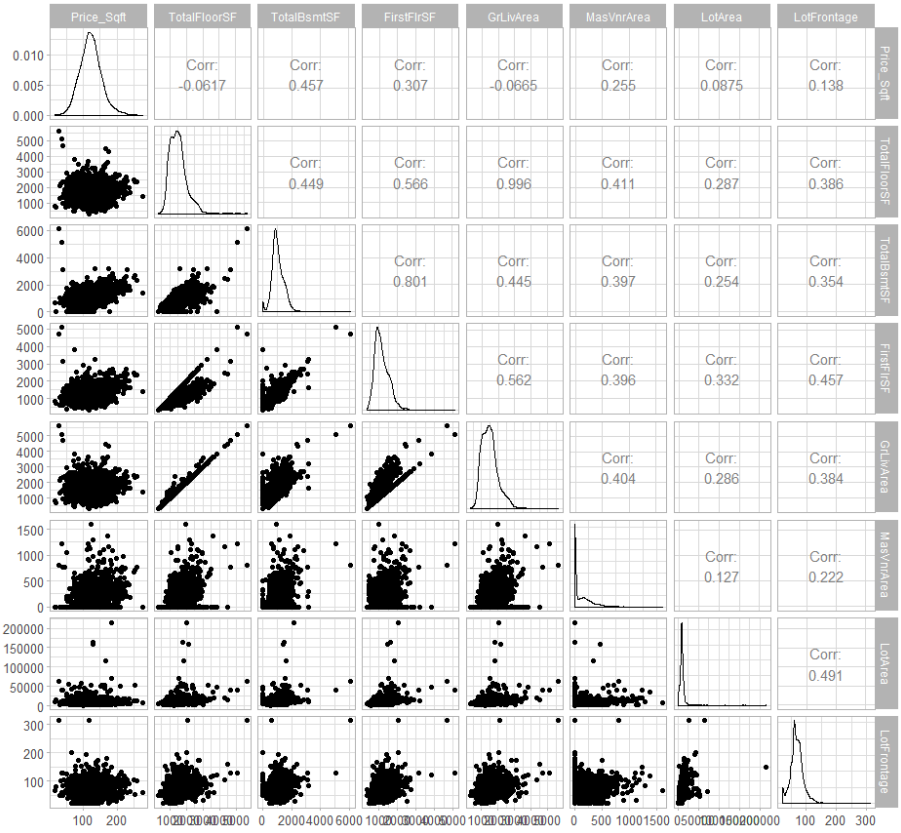
Skim

```
Skim summary statistics
 n obs: 2930
 n variables: 20

-- Variable type:factor ---------------------------------------------------------------------------
    variable missing complete    n n_unique                          top_counts ordered
 KitchenQual       0     2930 2930        5 TA: 1494, Gd: 1160, Ex: 205, Fa: 70   FALSE

-- Variable type:integer --------------------------------------------------------------------------
     variable missing complete    n      mean       sd   p0     p25    p50      p75   p100     hist
   Fireplaces       0     2930 2930      0.6     0.65    0    0       1        1        4 ▇▁▁▁▁▁▁▁
   FirstFlrSF       0     2930 2930  1159.56   391.89  334  876.25 1084     1384     5095 ▆▇▂▁▁▁▁▁
     FullBath       0     2930 2930     1.57     0.55    0    1       2        2        4 ▃▇▁▇▁▁▁▁
   GarageArea       1     2929 2930   472.82   215.05    0  320     480      576     1488 ▂▇▅▁▁▁▁▁
   GarageCars       1     2929 2930     1.77     0.76    0    1       2        2        5 ▂▇▂▁▁▁▁▁
  GarageYrBlt     159     2771 2930  1978.13    25.53 1895 1960    1979     2002     2207 ▂▅▇▃▂▁▁▁
    GrLivArea       0     2930 2930  1499.69   505.51  334 1126    1442     1742.75  5642 ▆▇▃▁▁▁▁▁
     HalfBath       0     2930 2930     0.38     0.5     0    0       0        1        2 ▇▁▅▁▁▁▁▁
     HouseAge       0     2930 2930    36.43    30.29   -1    7      34       54      136 ▇▆▃▂▁▁▁▁
      LotArea       0     2930 2930 10147.92 7880.02 1300 7440.25 9436.5 11555.25 215245 ▇▁▁▁▁▁▁▁
  LotFrontage     490     2440 2930    69.22    23.37   21   58      68       80      313 ▂▇▁▁▁▁▁▁
   MasVnrArea      23     2907 2930    101.9   179.11    0    0       0      164     1600 ▇▁▁▁▁▁▁▁
  OverallQual       0     2930 2930     6.09     1.41    1    5       6        7       10 ▁▃▇▅▂▁▁▁
 QualityIndex       0     2930 2930    33.76     9.18    1   30      35       40       90 ▁▃▇▂▁▁▁▁
  TotalBsmtSF       1     2929 2930  1051.61   440.62    0  793     990     1302     6110 ▃▇▂▁▁▁▁▁
 TotalFloorSF       0     2930 2930  1495.01   503.13  334 1120    1440     1740     5642 ▆▇▃▁▁▁▁▁
    YearBuilt       0     2930 2930  1971.36    30.25 1872 1954    1973     2001     2010 ▁▁▁▂▃▅▆▇
  YearRemodel       0     2930 2930  1984.27    20.86 1950 1965    1993     2004     2010 ▃▁▁▁▂▃▅▇

-- Variable type:numeric --------------------------------------------------------------------------
   variable missing complete    n  mean    sd    p0    p25    p50    p75   p100     hist
 Price_Sqft       0     2930 2930 121.6 31.89 15.37 100.57 120.43 140.01 276.25 ▁▃▇▃▁
>
```

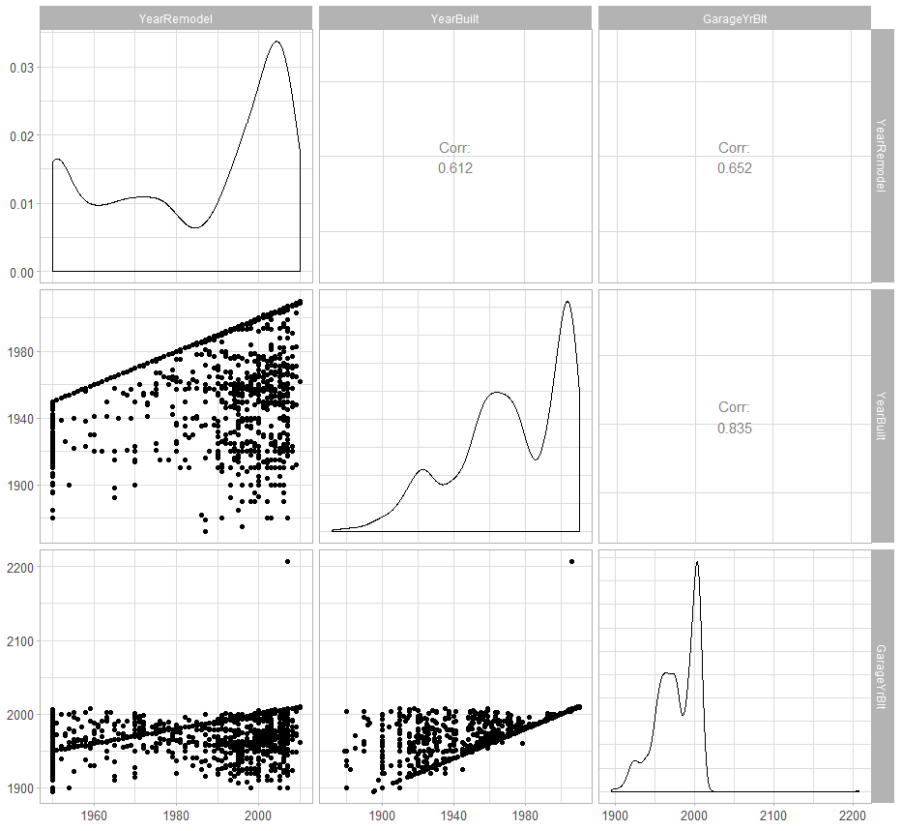
High Impact



Continuous



Temporal


YearRemodel
YearBuilt
GarageYrBlt
Corr:
0.612
Corr:
0.652
Corr:
0.835
0.03
0.02
0.01
0.00
1980
1940
1900
2200
2100
2000
1900
1960
1980
2000
1900
1940
1980
1900
2000
2100
2200