Modeling Assigment #2

Brandon Moretz

### Introduction

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

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

### Sample definition

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

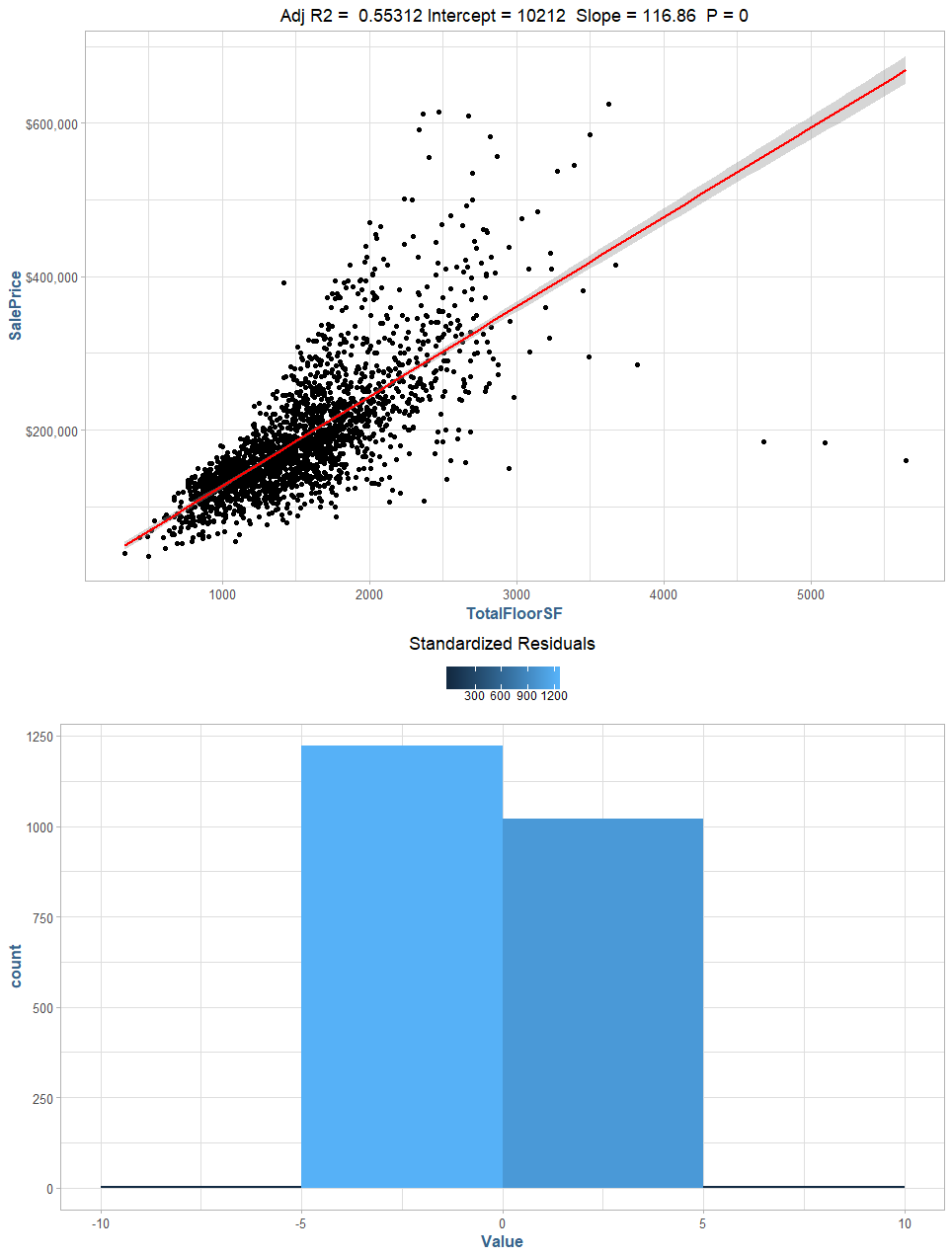
The Ames housing dataset contains approximately three-thousand observations of eighty-two variables collected from the Ames Assessor’s Office specifically for assessing value of individual residential properties sold in Ames, Iowa from 2006 to 2010. Given that this data was collected for specifically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

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



### Simple linear regression models

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



*Model 1:* Ŷ = 10212 + 116.862β1 + ε, R2 = **0.5533**

Where β1 is the total square footage of the house

H0 : β1 = 0

Ha : β1 ≠ 0

t1 = B̂1 / SB̂1 = 116.862 / 2.215 = **52.7594**

t-test with 99% confidence (α = 0.01), threshold: tα/2, n – p – 2 = t0.005, 2247 = 2.578

**Reject** H0, since |t0| > t0.005, 2247

We can reject the null hypothesis that total square footage in this model is no better than the simple slope. Total square footage has a statistically significant impact on the sale price of a given home. The R2 here suggests that overall quality can be used to explain approximately 55.33% of the global variance in the sale price.

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

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

Sum of Squares **Total** = SST = SSR + SSE = 14,654,543,821,775

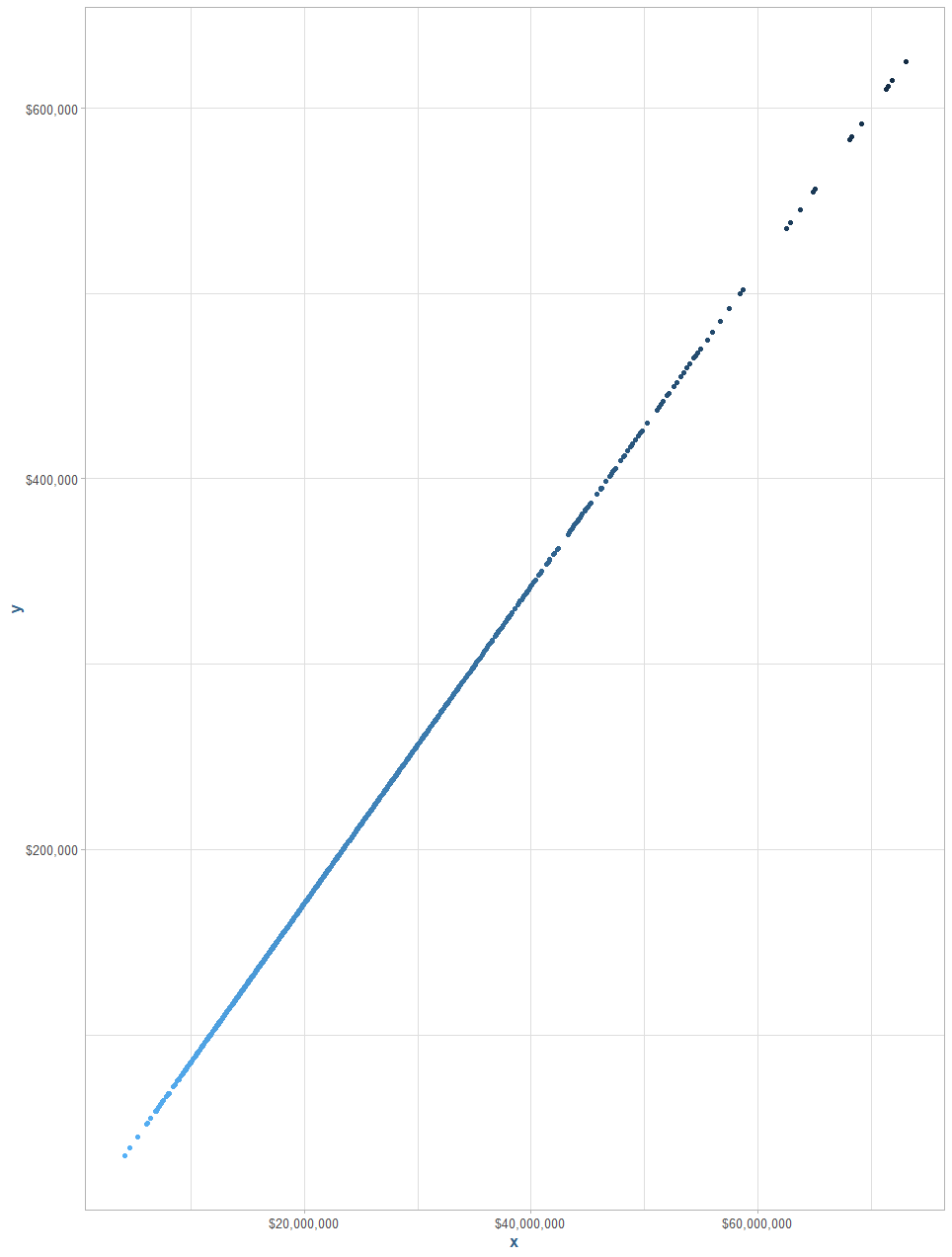
Let,

N = 2250, p = 1

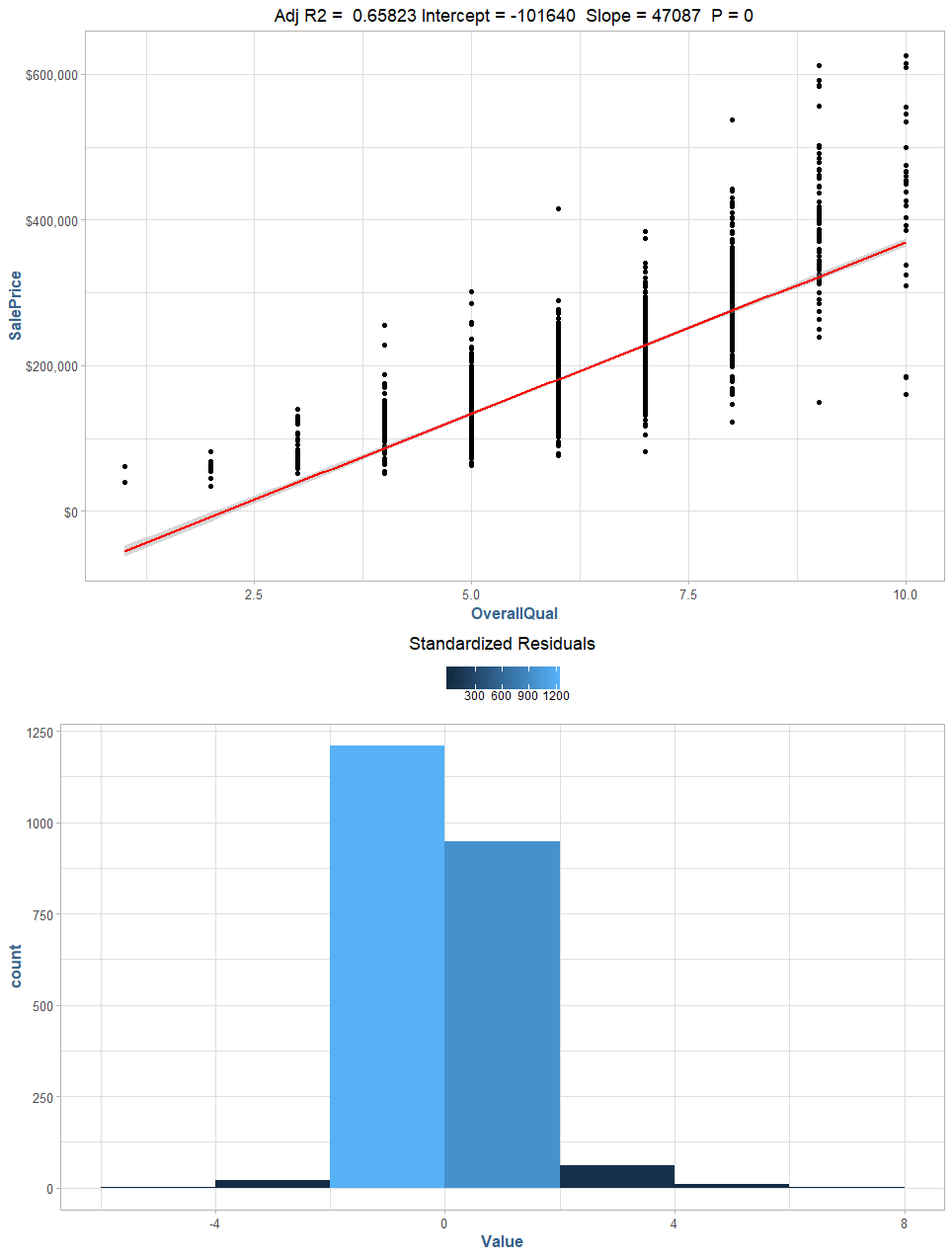
F = [(SST - SSE) / p] / [SSE / (n – p – 1)] = 8,108,616,754,275 / 2,911,889,265 = **2784.658** on p = 1 and 2248 DF

p-value: < 0.0001

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



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



Model 2: Ŷ = 101641.5 + 47086.6β1 + ε, R2 = **0.6584**

Where β1 is the overall quality of the house

H0 : β1 = 0

Ha : β1 ≠ 0

t1 = B̂1 / SB̂1 = 47,086.6 / 715.4 = **65.8186**

t-test with 99% confidence (α = 0.01), threshold: tα/2, n – p – 2 = t0.005, 2247 = 2.578

**Reject** H0, since |t0| > t0.005, 2247

We can reject the null hypothesis that overall quality in this model is no better than the simple slope. Overall quality (**OverallQual**) of the home has a statistically significant impact on the sale price of a given home. The R2 here suggests that overall quality can be used to explain approximately 65.84% of the global variance in the sale price.

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

Sum of Squared **Error** = SSE = 102.964

Sum of Squares **Total** = SST = SSR + SSE = 343.7658

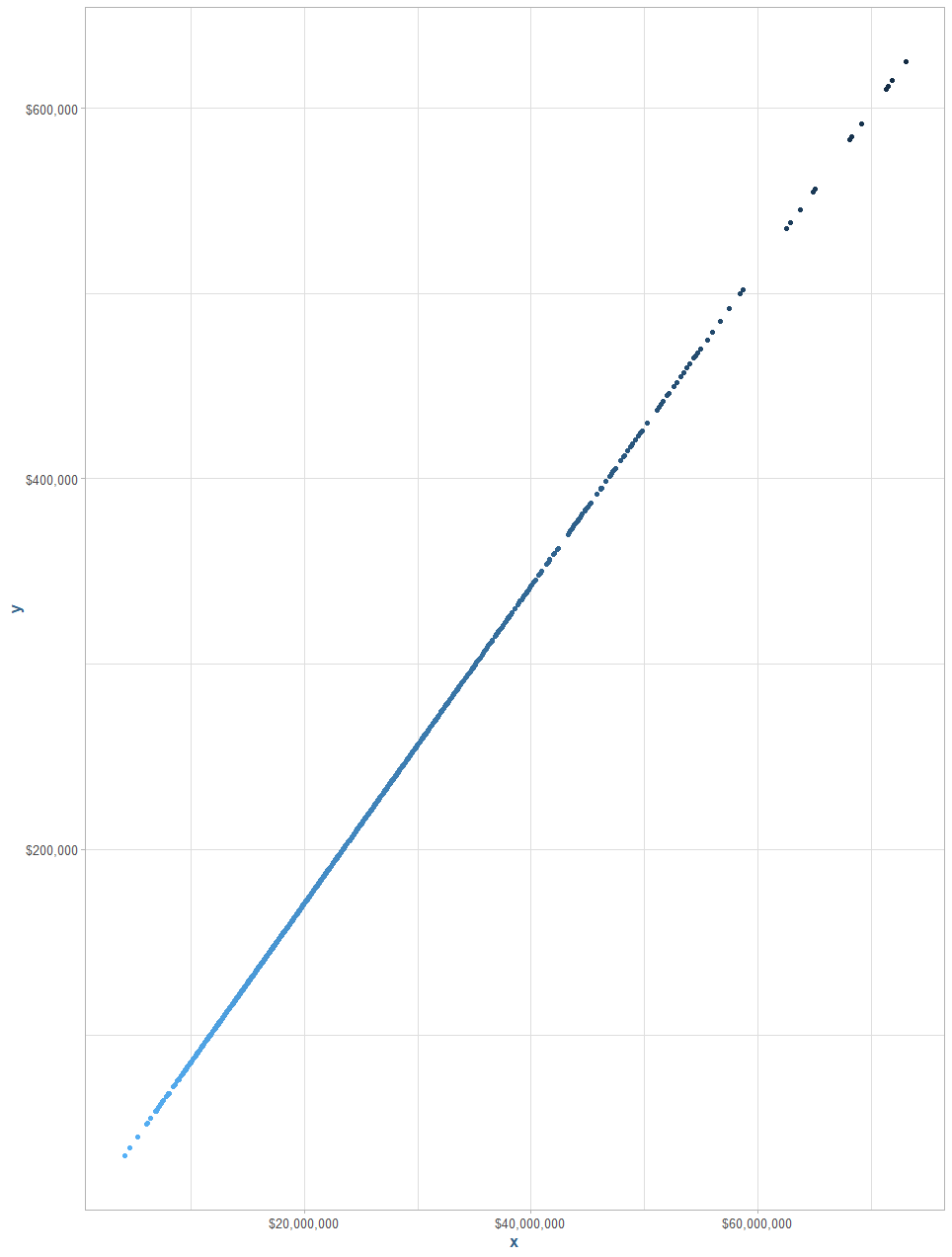
Let,

N = 2250, p = 1

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

p-value: < 0.0001

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

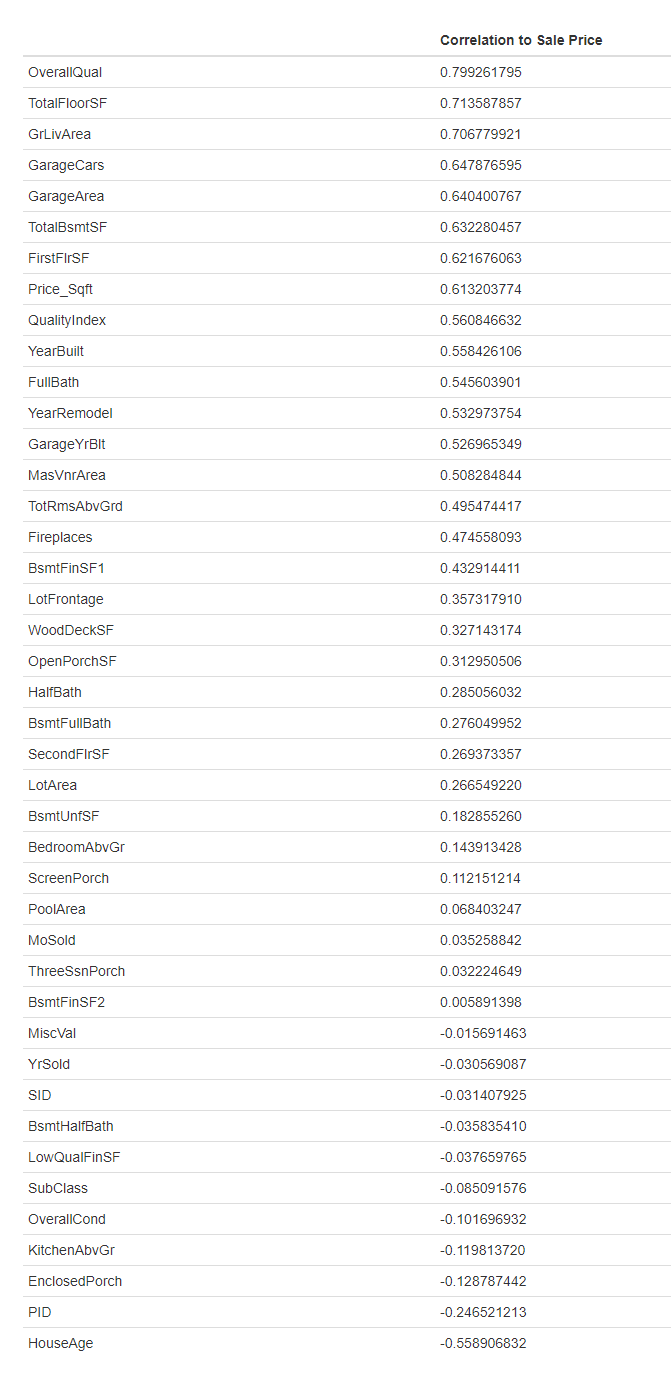


### Research

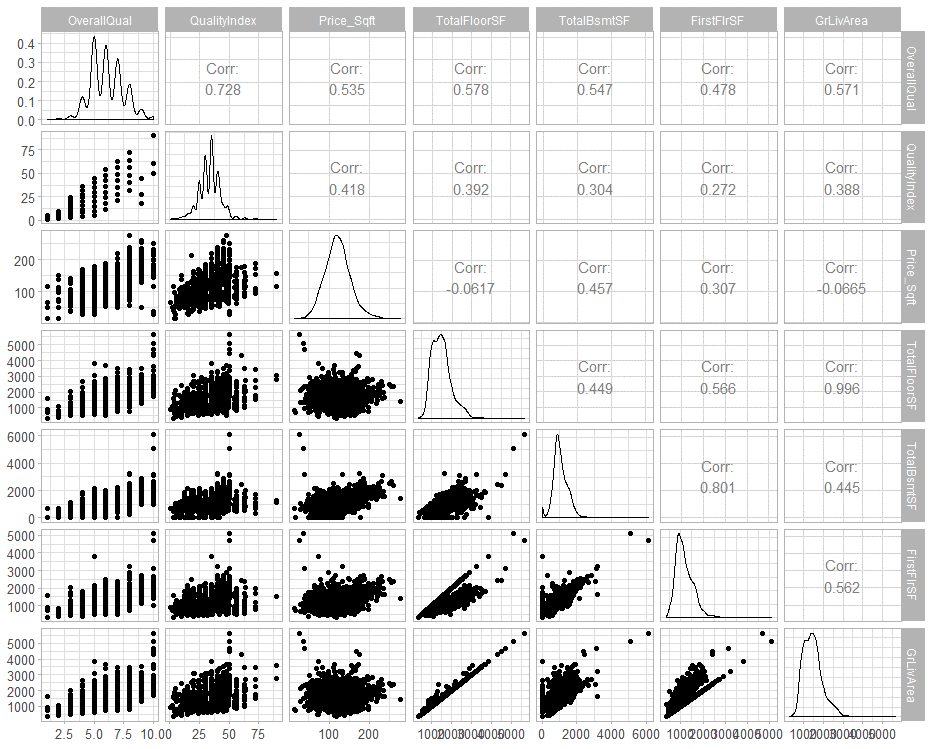
### Conclusion

### Appendex

#### Feature correlation



High Impact



Continuous

