Modeling Assigment #3

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.

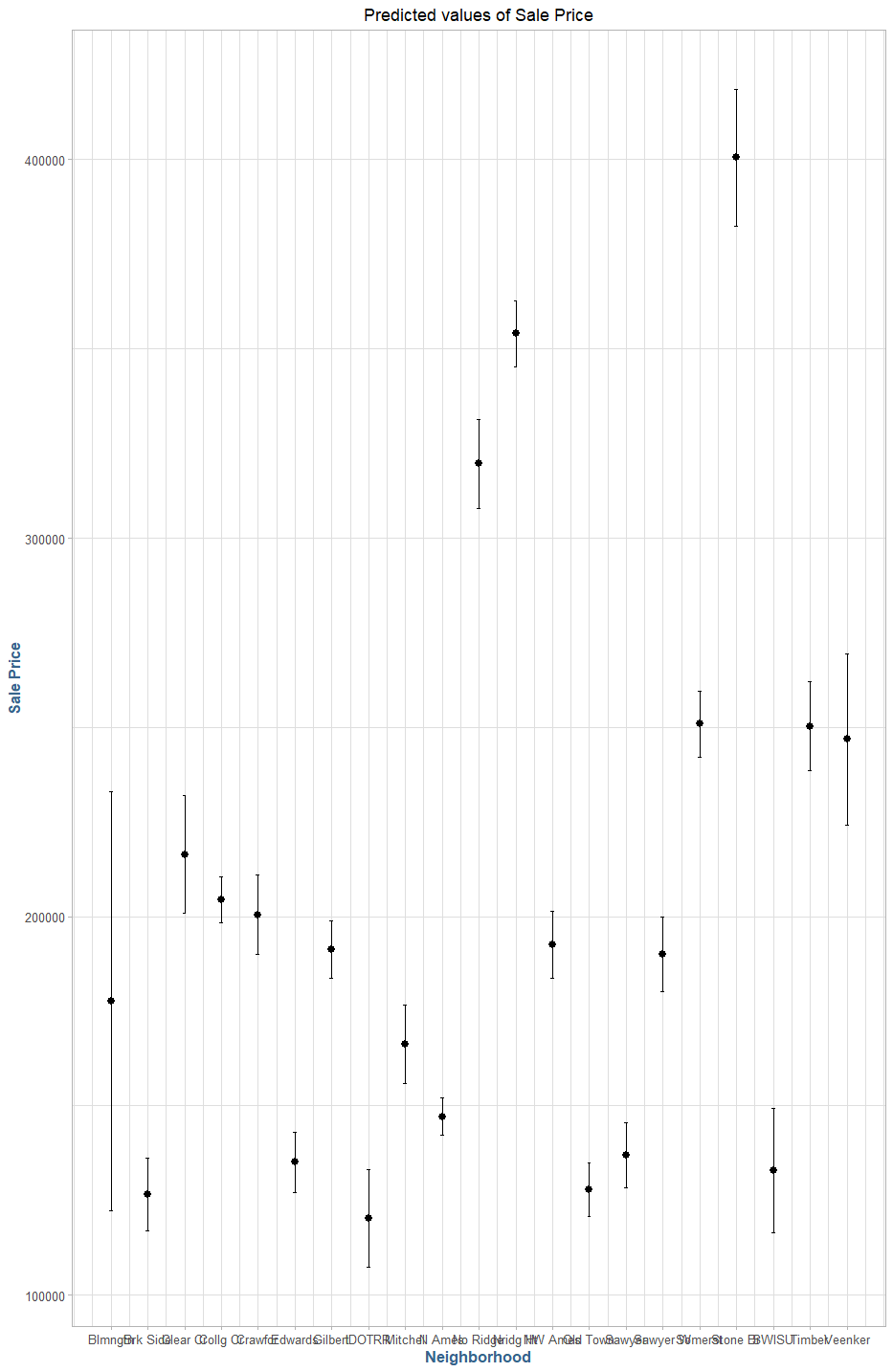
In this report, we will use the Ames dataset which is an alternative to the famous Boston housing data to perform exploratory data analysis through variable derivation, validation, selection and visualization to measure the relevance of these indicators as they pertain to the value of the home in terms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

### preparing the categorical variables

For this part of the lab we will take a systematic approach to examining the relationships between the categorical variables in the data set in relation to the desired response variable. We will look at the subset of 43 columns that contain categorical information and extract the R2, residual standard error from the model fitted to predict the sale price, as well as the mean difference between levels, the number of levels, and the percent of the data that is populated with this attribute.

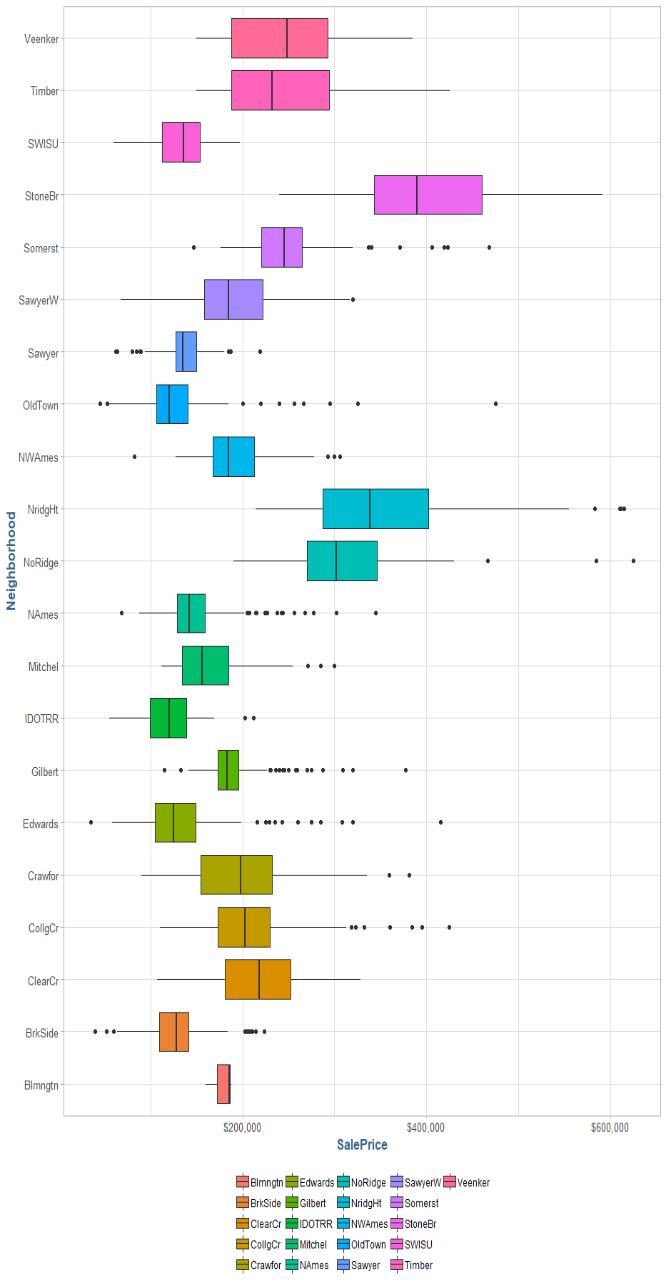
The reason we chose these metrics is due to the variance explained by each category is an indicator of the relative “goodness-of-fit”, and the RSE and mean difference give us a sense of the variance found in each of the levels, where the lower the variance and higher the R2 will give us a good idea of how useful this metric will be in predictive modeling, and a high value for the mean level difference denotes that there will be a greater chance for statistical differences in the levels than if the values were all clustered together. The full results of this exercise can be found in the [appendix](#_Appendix).

The first categorical variable we will explore is the neighborhood variable, as it has both a high R2 (.667), and relatively low residual standard error ($48,633) / high mean level difference ($204,189). In the following chart we can see the predicted sale price by neighborhood category:

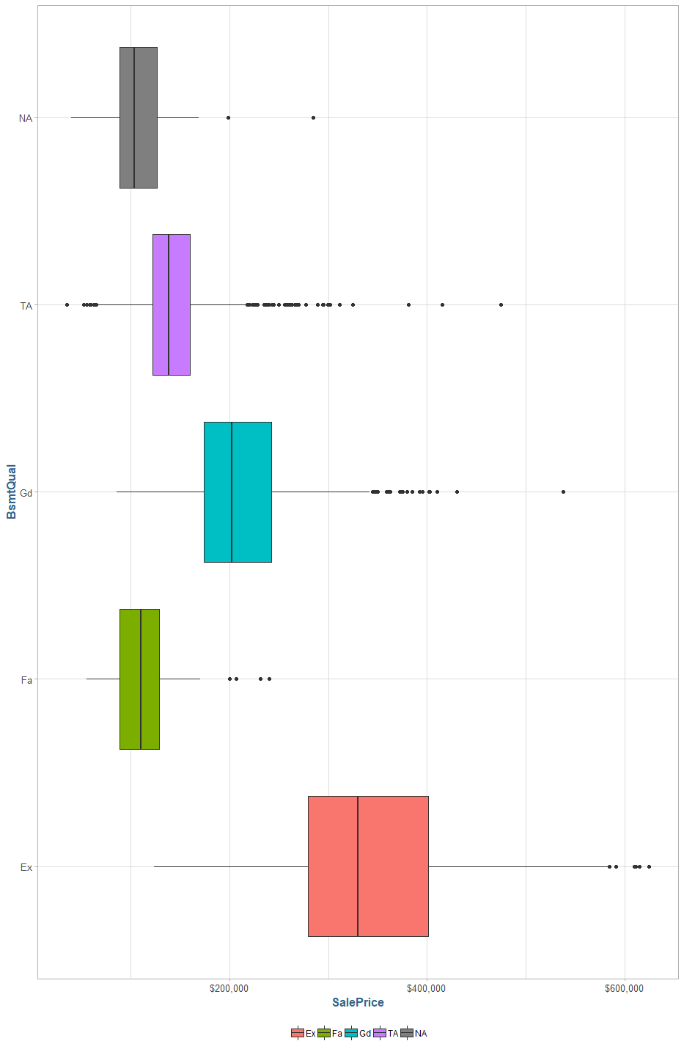


Unfortunately, the model predicted values for each neighborhood has a large interval of values that they could fall into. The following table summarizes each neighborhood by mean sale price:

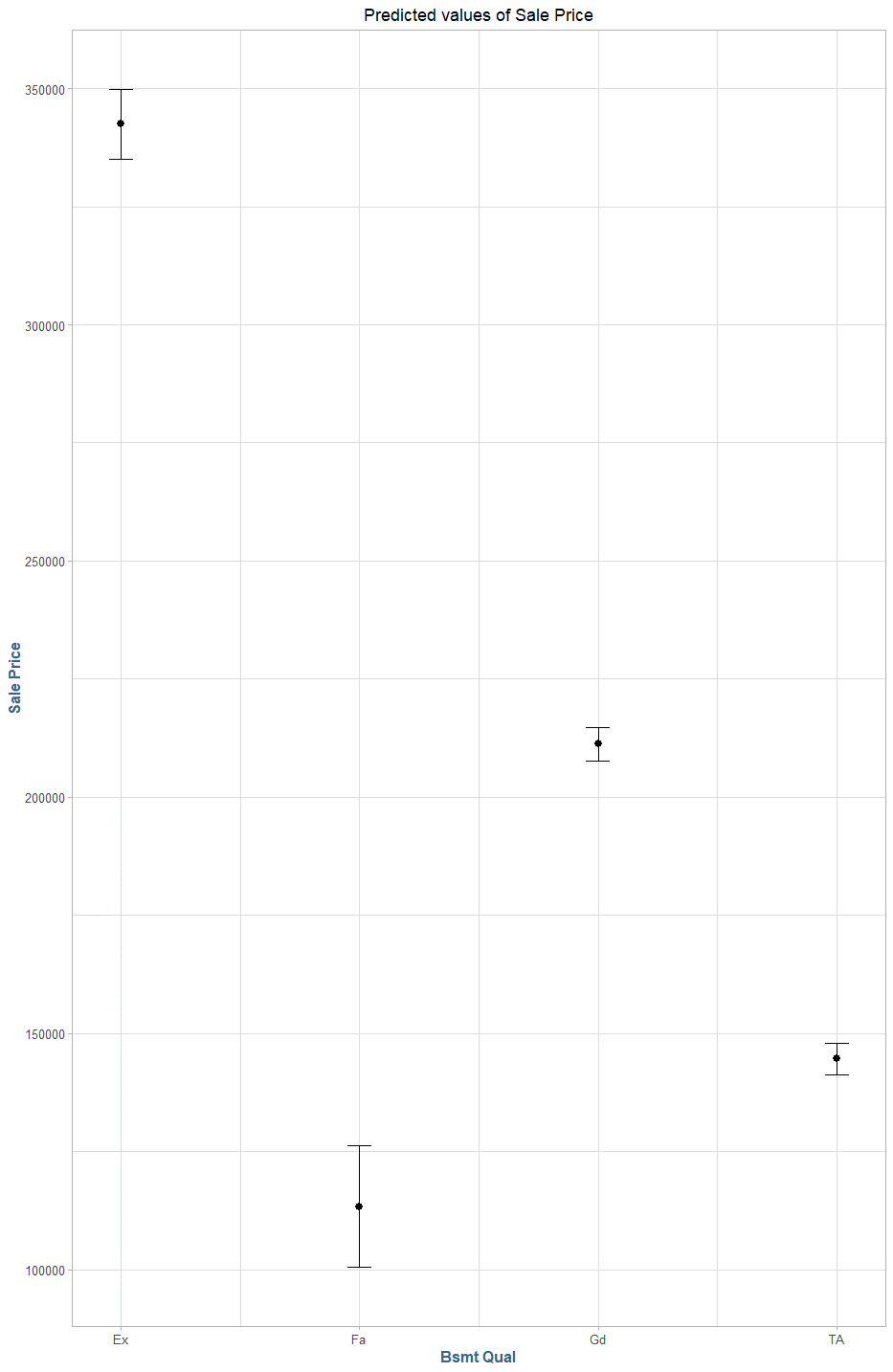


Looking at the modeling data, we can see that the outliers in sale price per neighborhood are vast which would explain the large variances in the model. We also note that of the twenty coefficient terms generated by the model, only five of them had significant p-values below the 5% level.

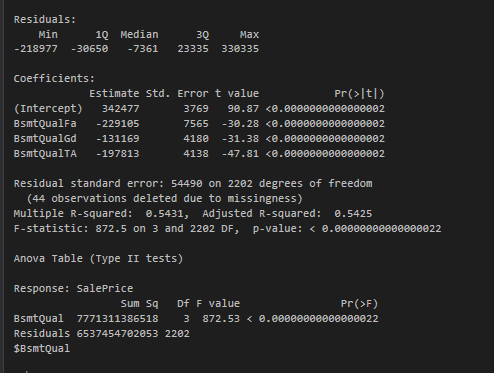
The next categorical variable we will examine is basement quality (**BsmtQual**). For this variable, we note the high percentage of values for our sample (98%), the high R2 and low residual standard error from the corresponding model (.5431 and $54,450, respectively) and the relatively high mean level difference of $203k.



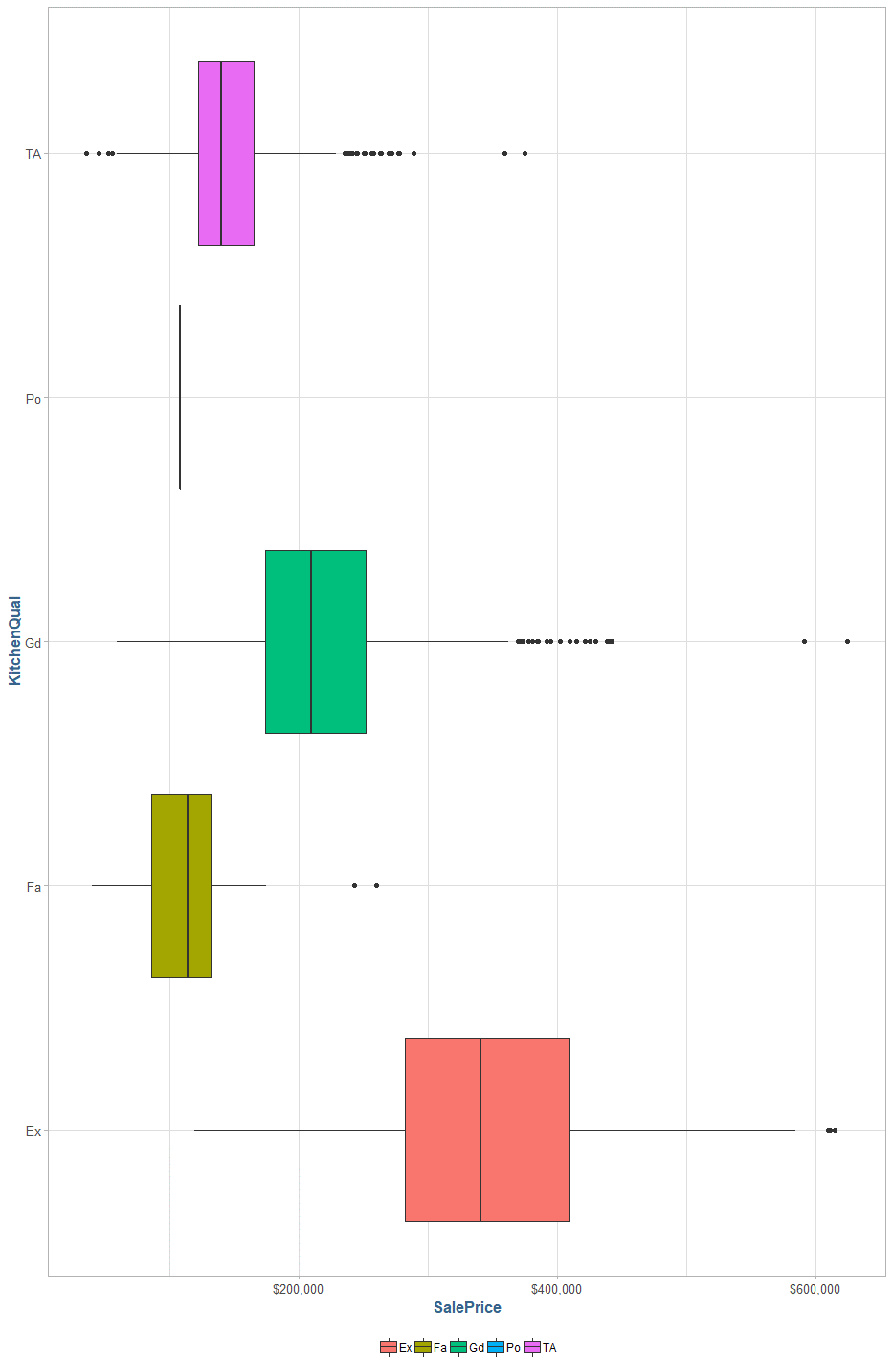
Even though there are some outliers in each of the groups, the predicted values for the sale price based solely upon the basement quality are promising:



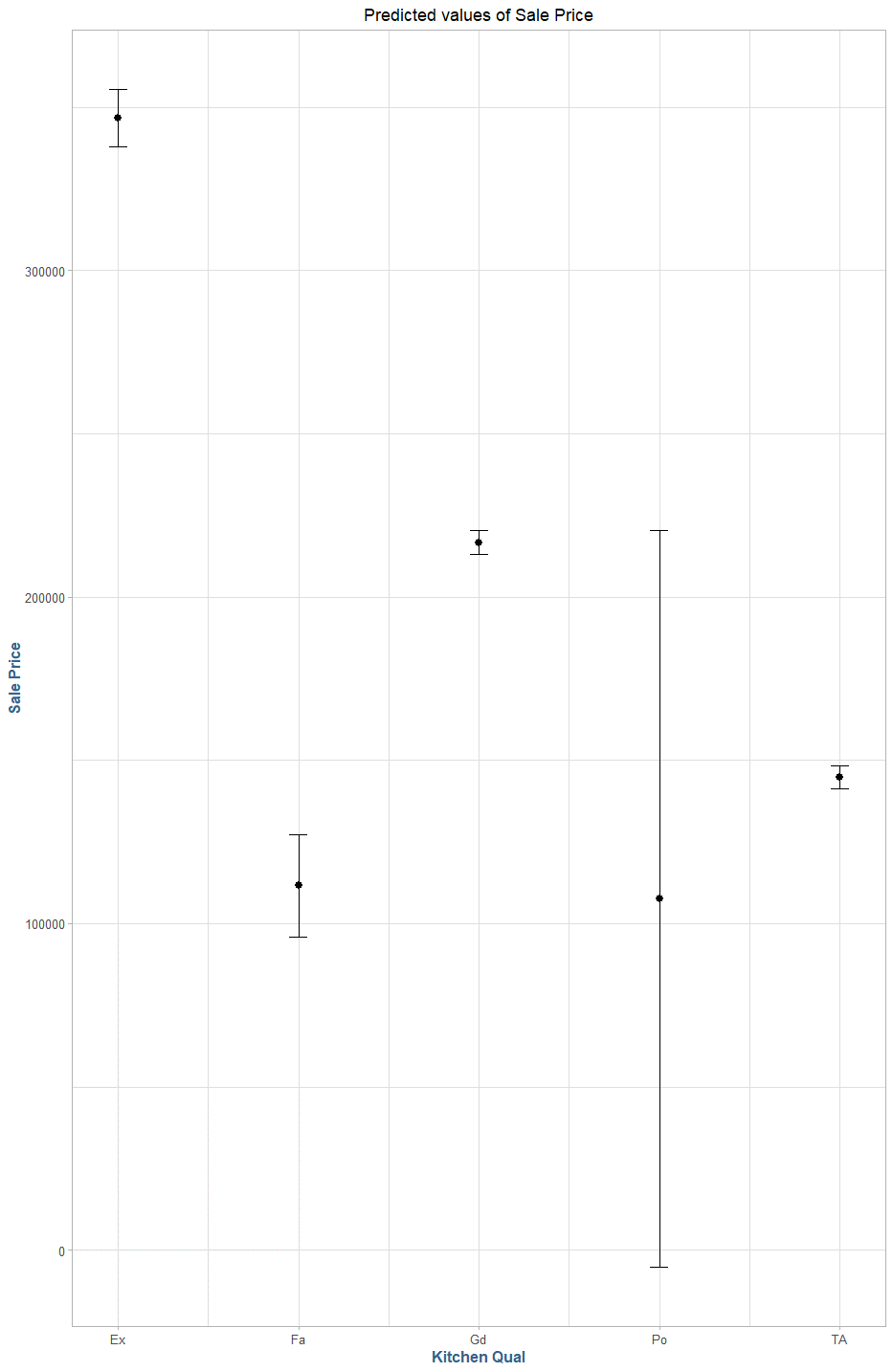
In the above graphic we see that each of the prediction intervals fall within a relatively tight bound. We can also look at the model diagnostics and see that each of the coefficient terms generated by the linear model have statistical significance with low standard errors (in the 3-7-thousand-dollar range):



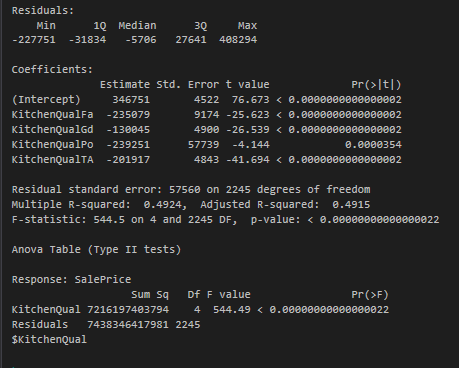
The basement quality (**BsmtQual**) variable will be included for further analysis. Continuing with the theme of quality indicator categorical variables, we will move on to the kitchen quality (**KitchenQual**) variable. The kitchen quality variable has good indicators of predictability from the corresponding linear model, with a moderately high R2 and relatively small residual standard error (.4924 and $57.5k respectively). There is also a relatively high deviation between the means of the levels within the category at $185k. Below we can see the distribution of sale price by levels of kitchen quality:



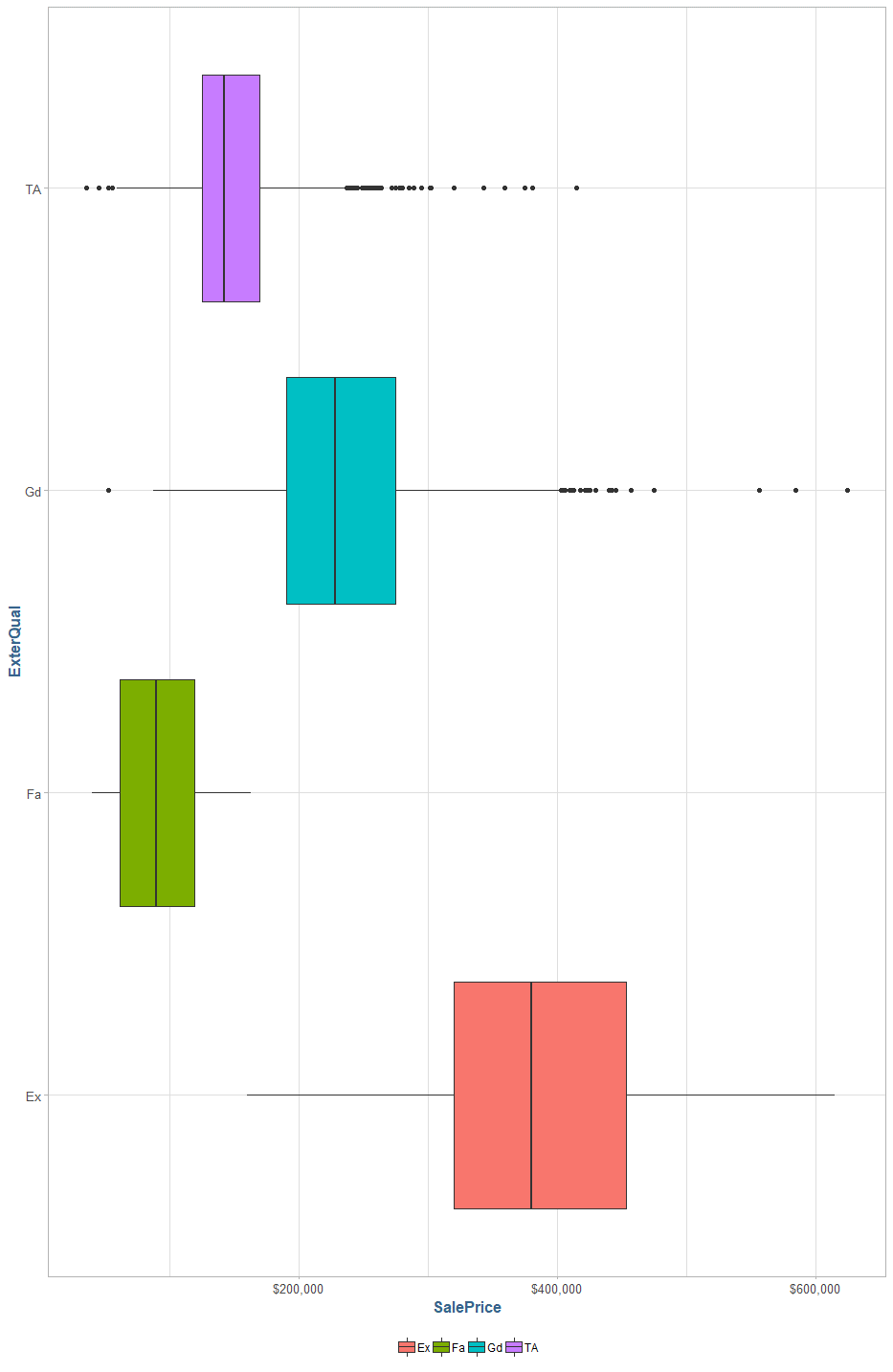
The predictive intervals for sale price based upon kitchen quality are tightly bound intervals, with the exception of the “poor” group, which has a large prediction interval:



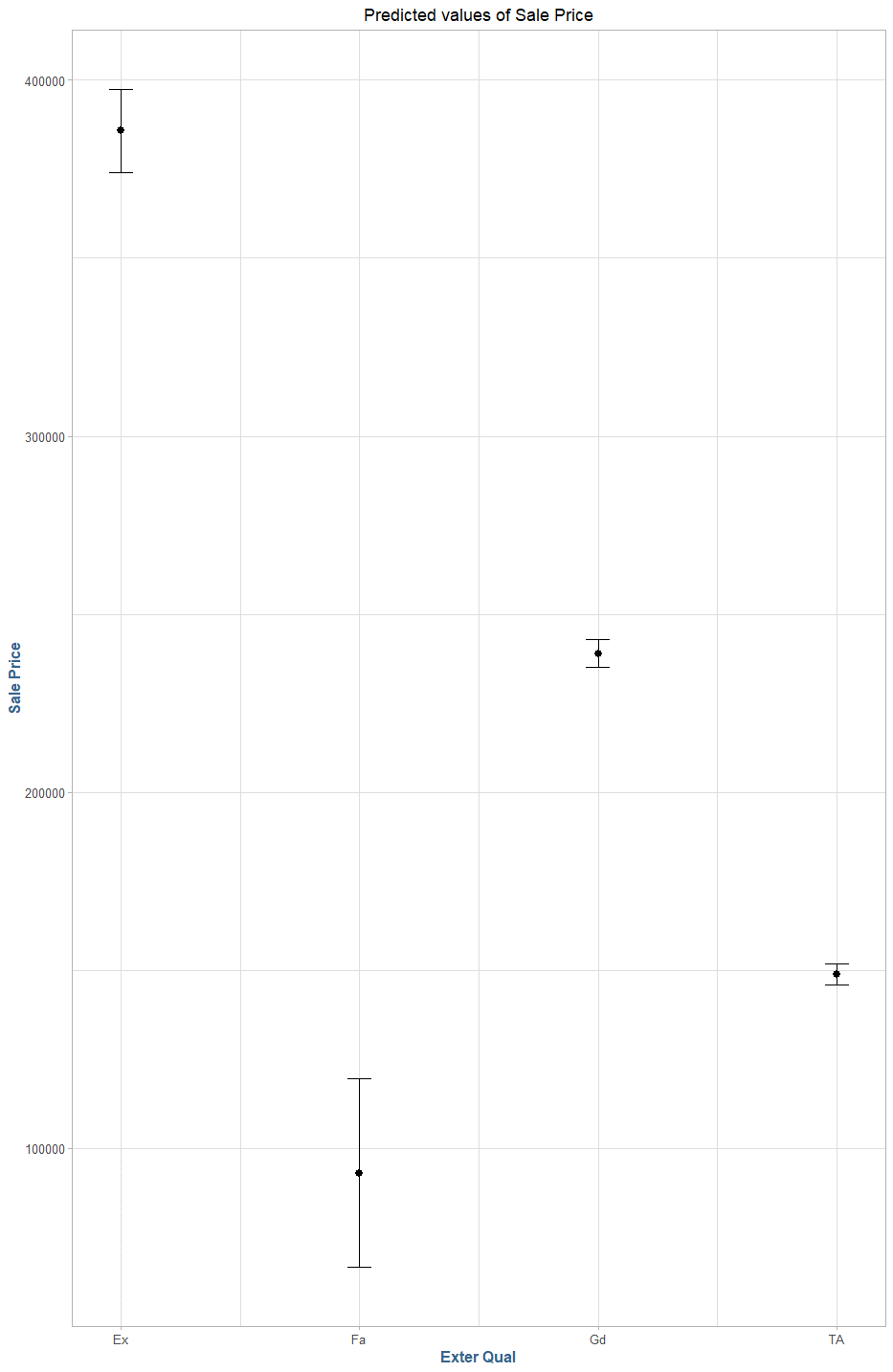
Additionally, like with the basement quality variable, all of the coefficients generated by the linear model appear to have statistical significance:



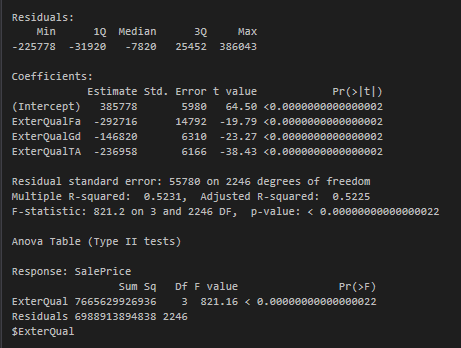
We will keep the kitchen quality variable for further analysis, and we will also examine the last individual quality indicator variable, exterior quality (**ExterQual**). The model generated by external quality variable yields a model with a moderately high R2 and a low residual standard error (.5231 and $55.7k respectively), additionally the mean difference between the levels of the category are relatively high at $216k. Below we can see the distribution of sale price within the various levels of the exterior quality:



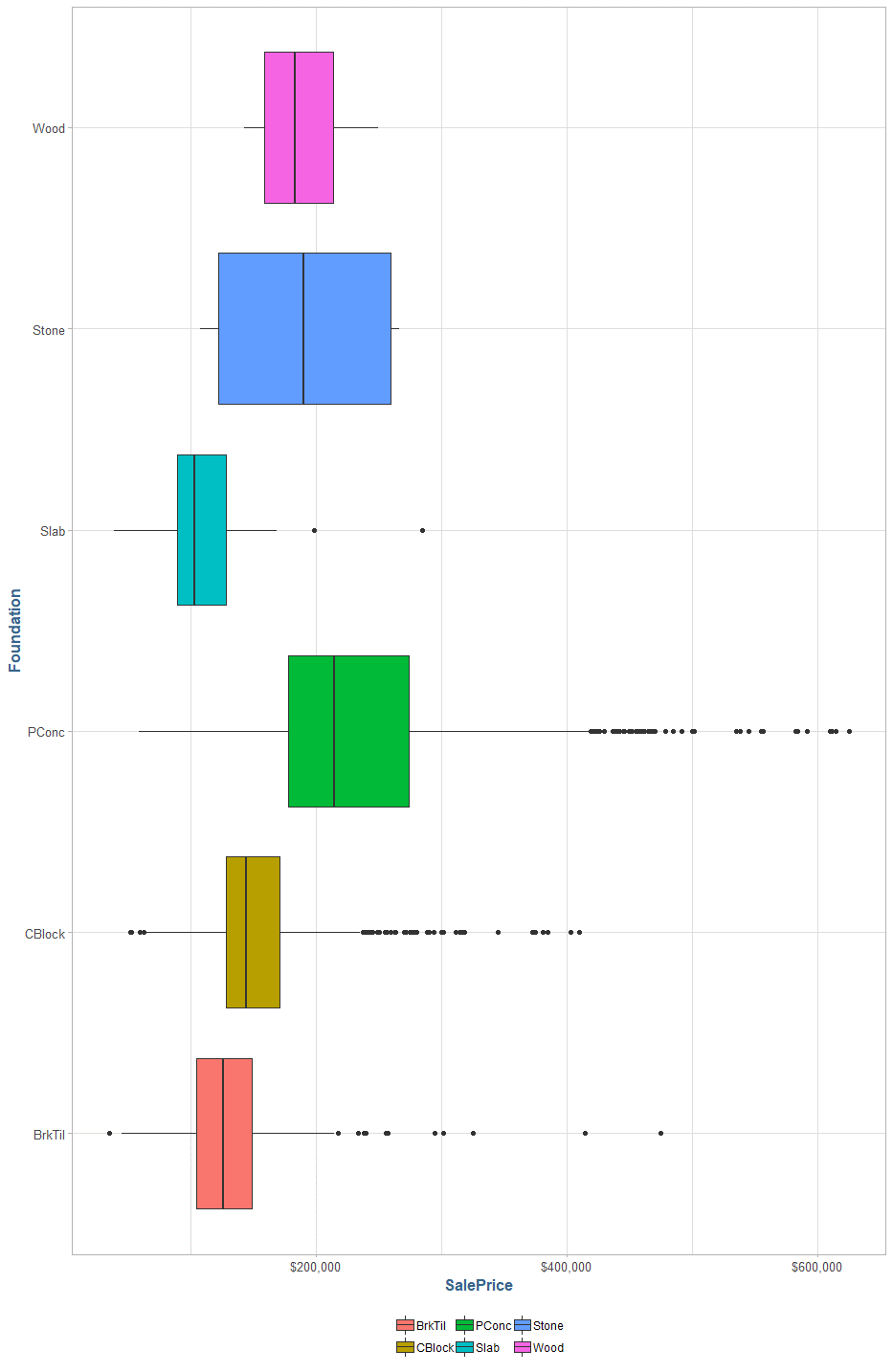
While there are outliers, there does seem to be a significant amount of clustering for the sale price. Additionally, the prediction intervals are small and dispersed as we would hope:



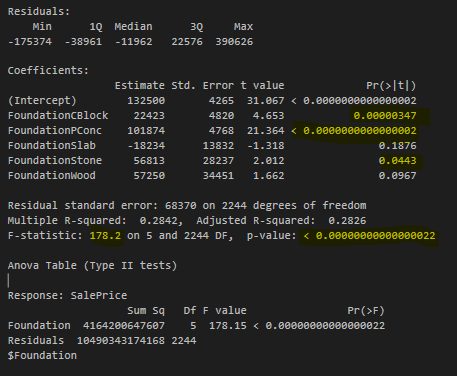
The following model diagnostics also confirm the statistical significance between the different levels of the exterior quality:



The final categorical variable we will look at including in our model is the foundation type (**Foundation**) of the home. The below diagram shows the sale price distributions by foundation type:



We note that this variable produces a slightly lower R2 and a higher residual standard error than the previous variables (.2842 and $68.3k respectively), however, the mean difference is still relatively high for the various levels at $169k and three of the five beta coefficients generated from the model show statistical significance in having an impact on the sale price:



It seems likely that this variable can help account for some of the variance that the quality type variables cannot account for, therefore we will include this variable for further analysis.

### The predictive modeling framework

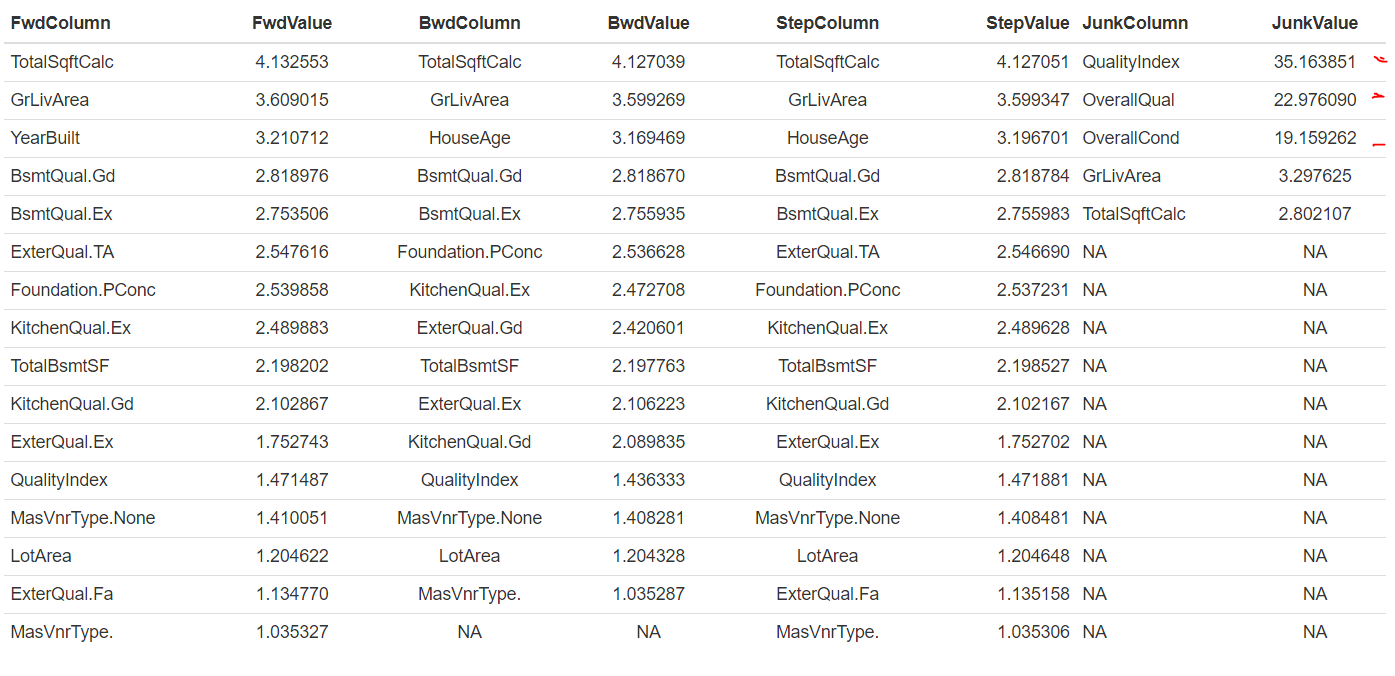
In this section we will split our sample into two parts, a training set and a test set. The purpose of this is to build a model on one set of data, then validate our model on unseen observations in the test set. We will use a ‘standard’ 70/30 split for this exercise, and the number of respective observations can be seen in the table below:



The total number of observations has dipped slightly due to some invalid values in some of the categorical variables of interest which have been scrubbed from the data set. We have narrowed down the universe of possible predictor variables to a subset of nineteen chosen variables with various scales. The preceding table summarizes our variables of choice, and denotes which variables are the result of ‘dummy coding’ where necessary.

The next step in this process is to train our models on the training data. We will train three models, one using all the columns specified in the above variable selection process, one as a simple intercept model to use as a baseline for comparison, and one that uses a simple linear regression so that our step-wise AIC model will be initialized. The final model, name junk, will be created using a multiple linear regression model using quality and square footage variables. The reason we create the junk model is to demonstrate the high degree of collinearity when a model is created using both a derived field and its corresponding underliers (quality index is comprised of overall condition and overall quality).

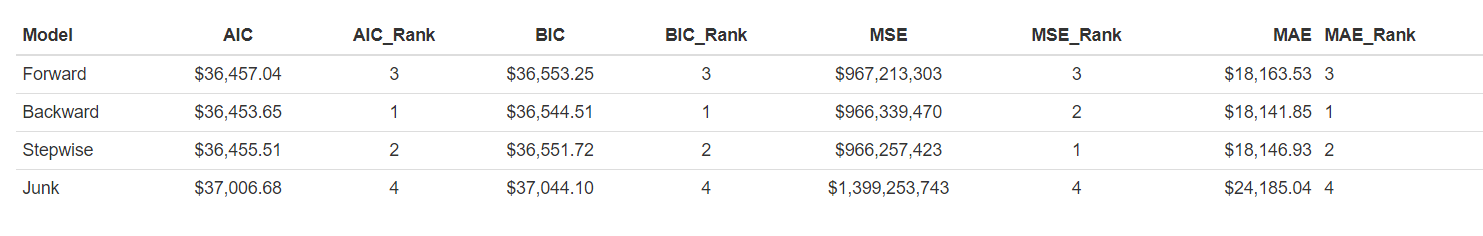
The auto-selection of parameters is not exactly the same for each of the three linear models. In the following figure we can see the columns each process generated, and their corresponding VIF values:



We note the high VIF values on three of the variables in the junk model. As we noted earlier, this is due to the quality index being derived from the other two variables producing a high degree of collinearity. We should be concerned about any column that generates a VIF value over 5 or 10, as there is a high probability of overfitting the model to redundant sets of predictor variables leading to bias in the model.

### model comparison

For the in-sample comparison of the models, we will calculate the Akaike Information Criterion (AIC), Bayesian Information Criterion, Mean Squared Error, and Mean Absolute Error for each of the preceding models and show their relative ranking amongst each of the models.



### Conclusion

### Appendix