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.

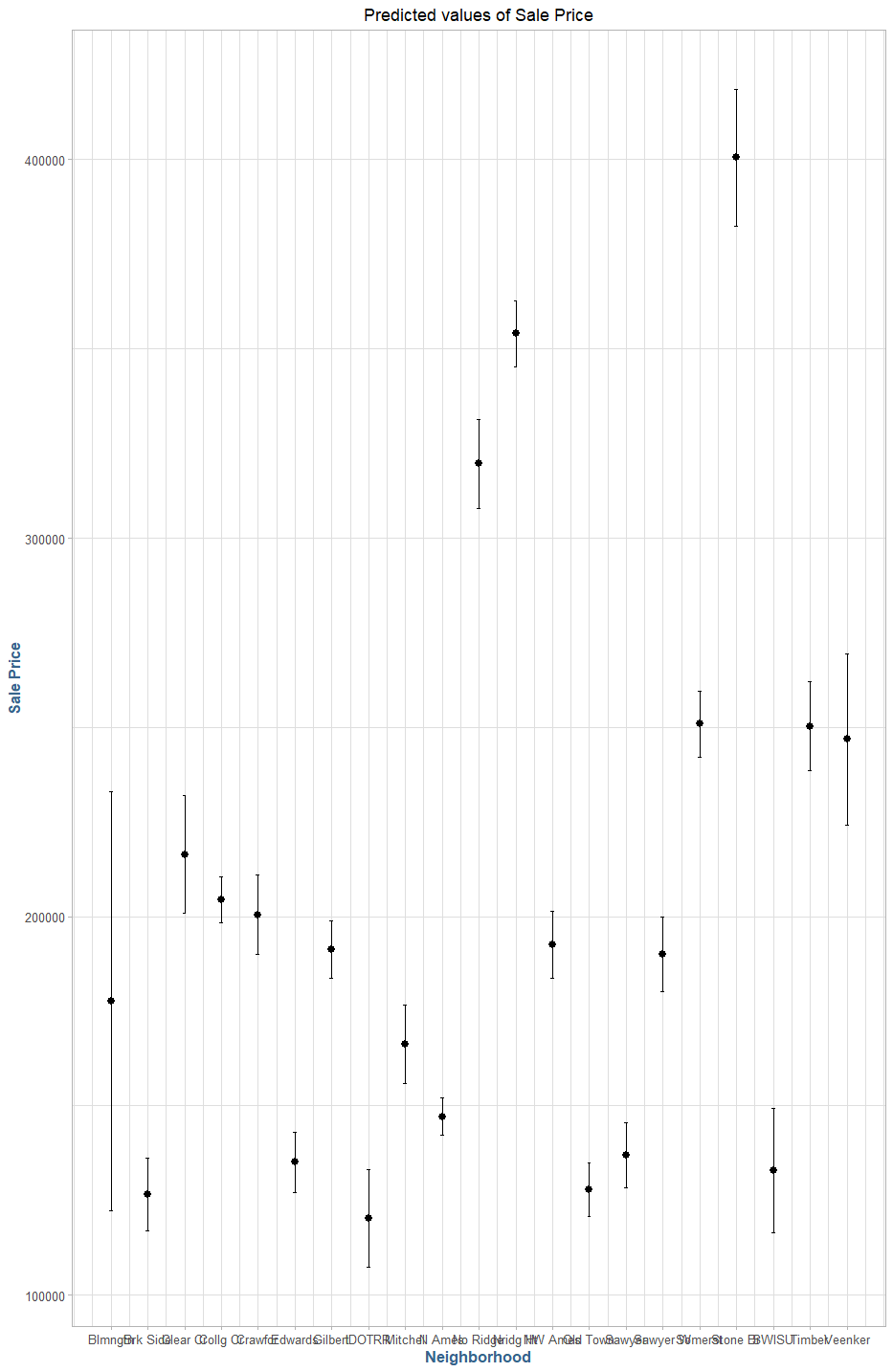
In this report, we will use the Ames dataset which is an alternative to the famous Boston housing data to perform exploratory data analysis through variable derivation, validation, selection and visualization to measure the relevance of these indicators as they pertain to the value of the home in terms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

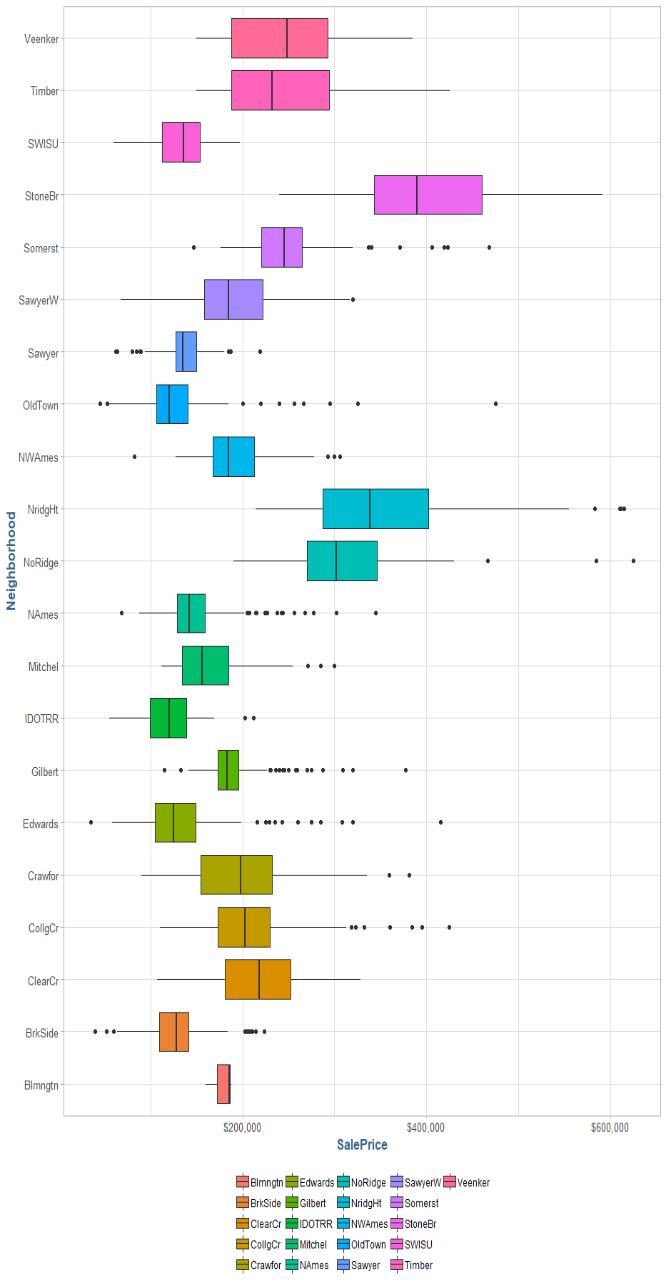
The reason we chose these metrics is due to the variance explained by each category is an indicator of the relative “goodness-of-fit”, and the RSE and mean difference give us a sense of the variance found in each of the levels, where the lower the variance and higher the R2 will give us a good idea of 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:

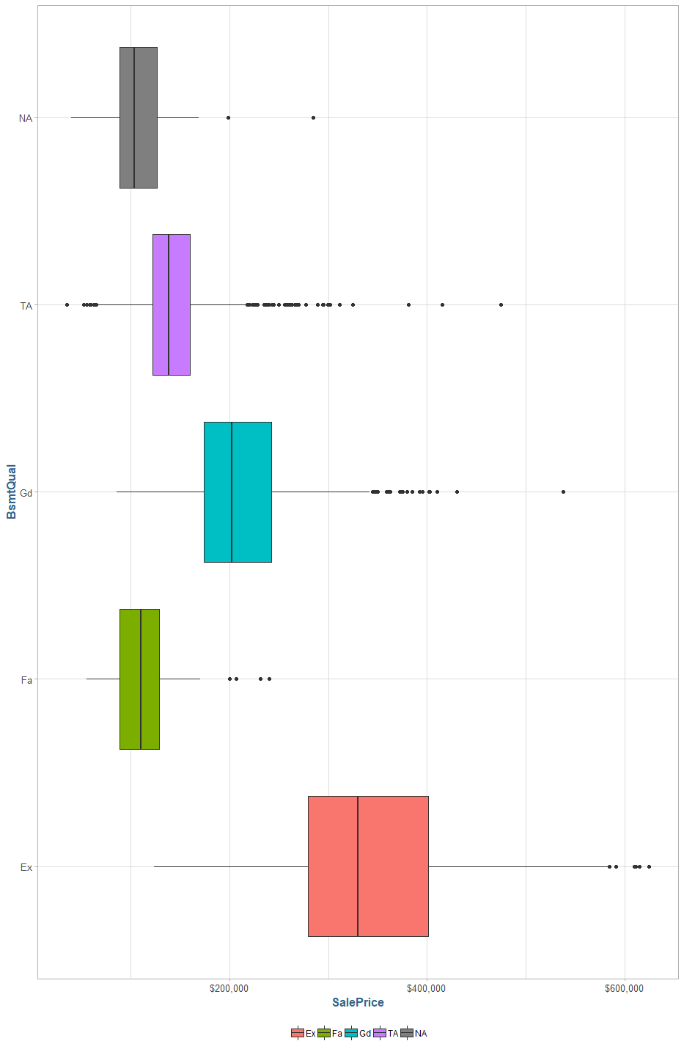


Unfortunately, the model predicted values for each neighborhood has a large interval of values that they could fall into. The following table summarizes each neighborhood by mean sale price:

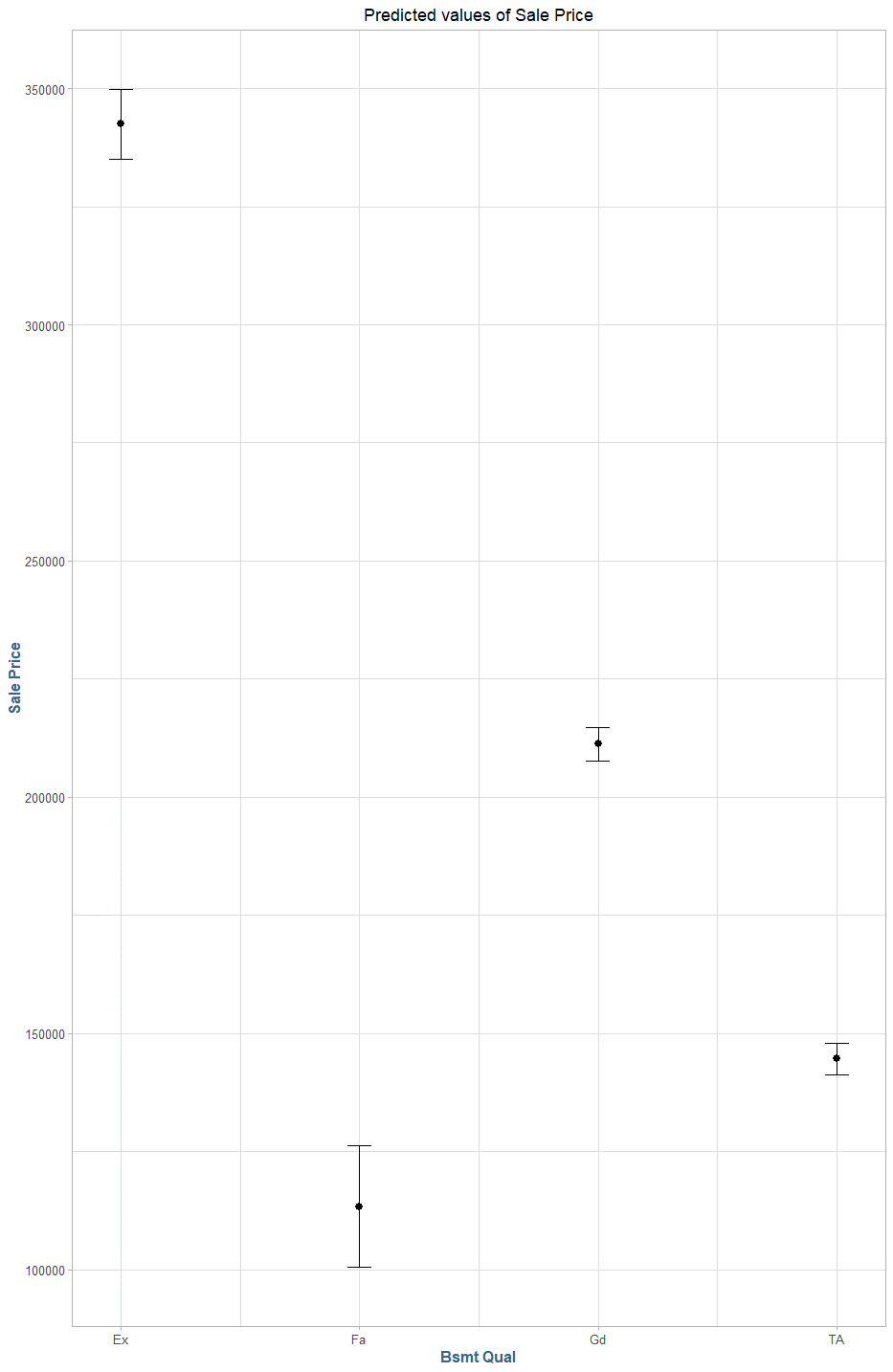


Looking at the modeling data, we can see that the outliers in sale price per neighborhood are vast which would explain the large variances in the model. We also note that of the twenty coefficient terms generated by the model, only five of them had significant p-values below the 5% level.

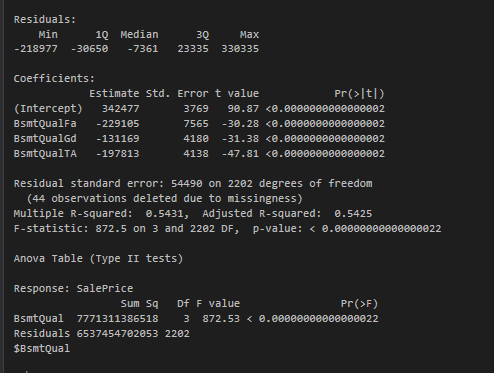
The next categorical variable we will examine is basement quality (**BsmtQual**). For this variable, we note the high percentage of values for our sample (98%), the high R2 and low residual standard error from the corresponding model (.5431 and $54,450, respectively) and the relatively high mean level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



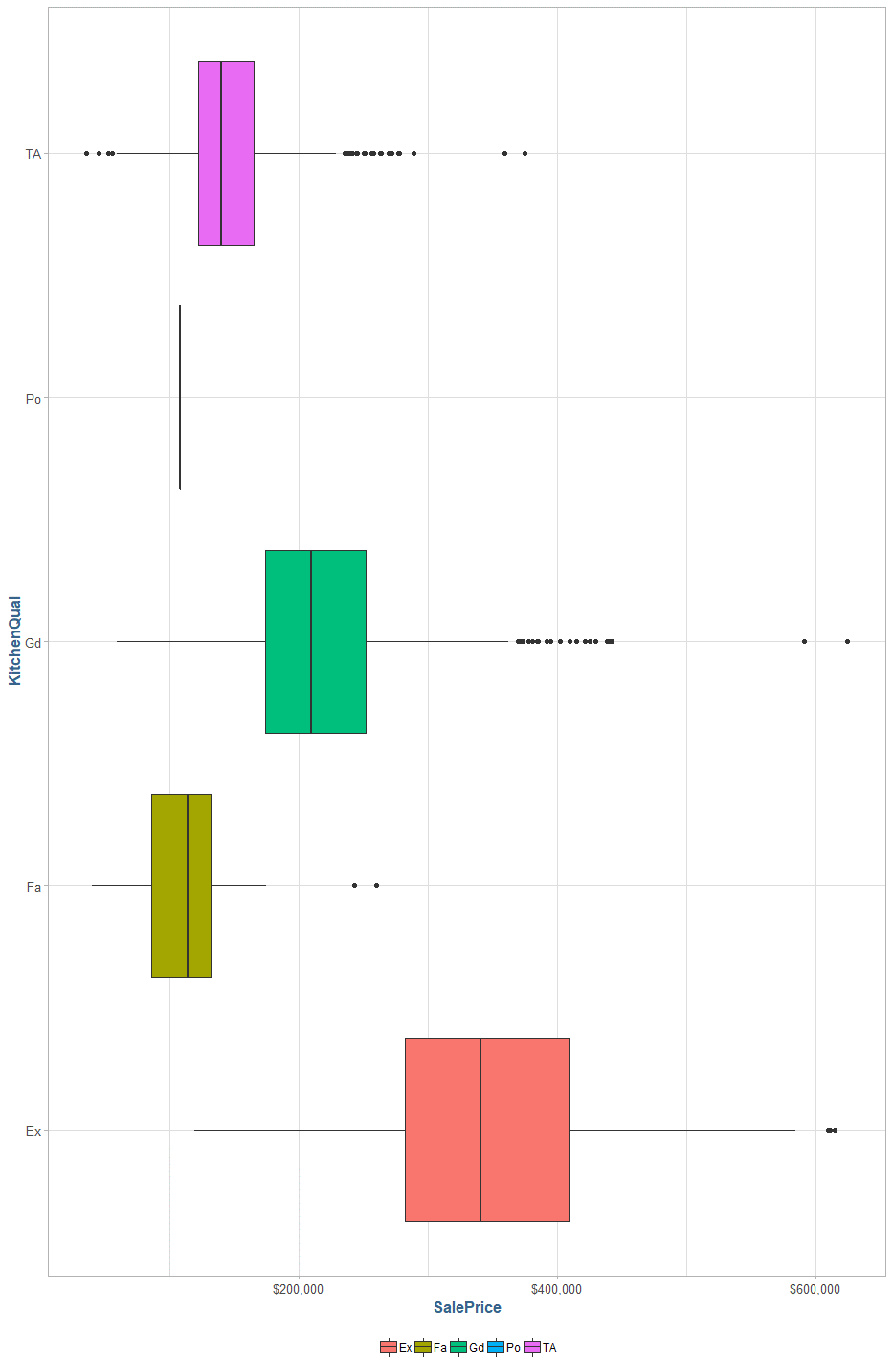
Even though there are some outliers in each of the groups, the predicted values for the sale price based solely upon the basement quality are promising:



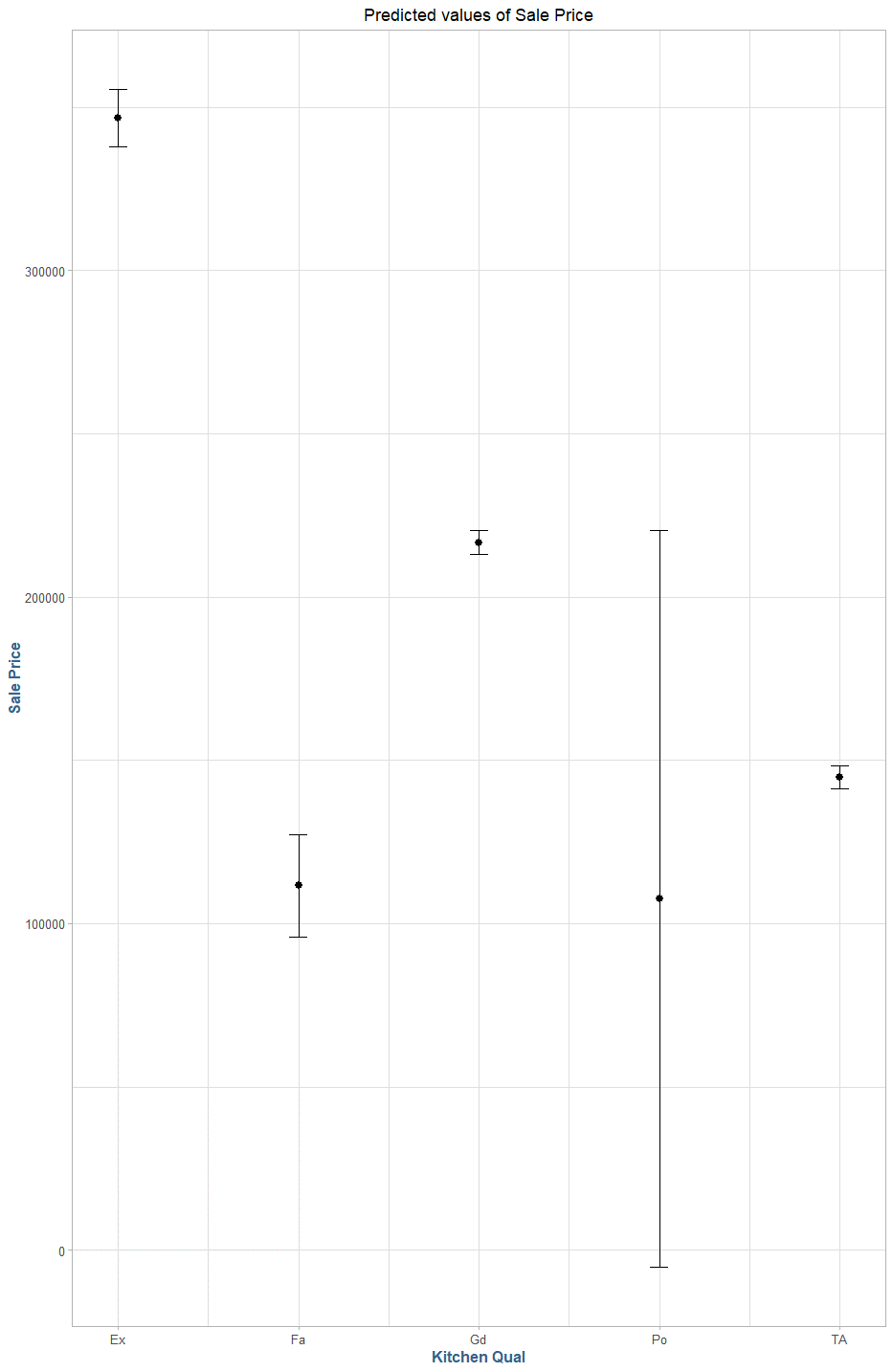
In the above graphic we see that each of the prediction intervals fall within a relatively tight bound. We can also look at the model diagnostics and see that each of the coefficient terms generated by the linear model have statistical significance with low standard errors (in the 3-7-thousand-dollar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



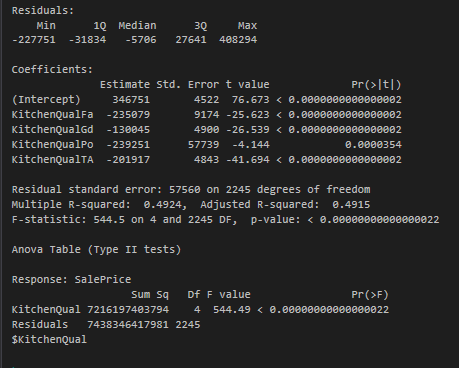
The basement quality (**BsmtQual**) variable will be included for further analysis. Continuing with the theme of quality indicator categorical variables, we will move on to the kitchen quality (**KitchenQual**) variable. The kitchen quality variable has good indicators of predictability from the corresponding 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:



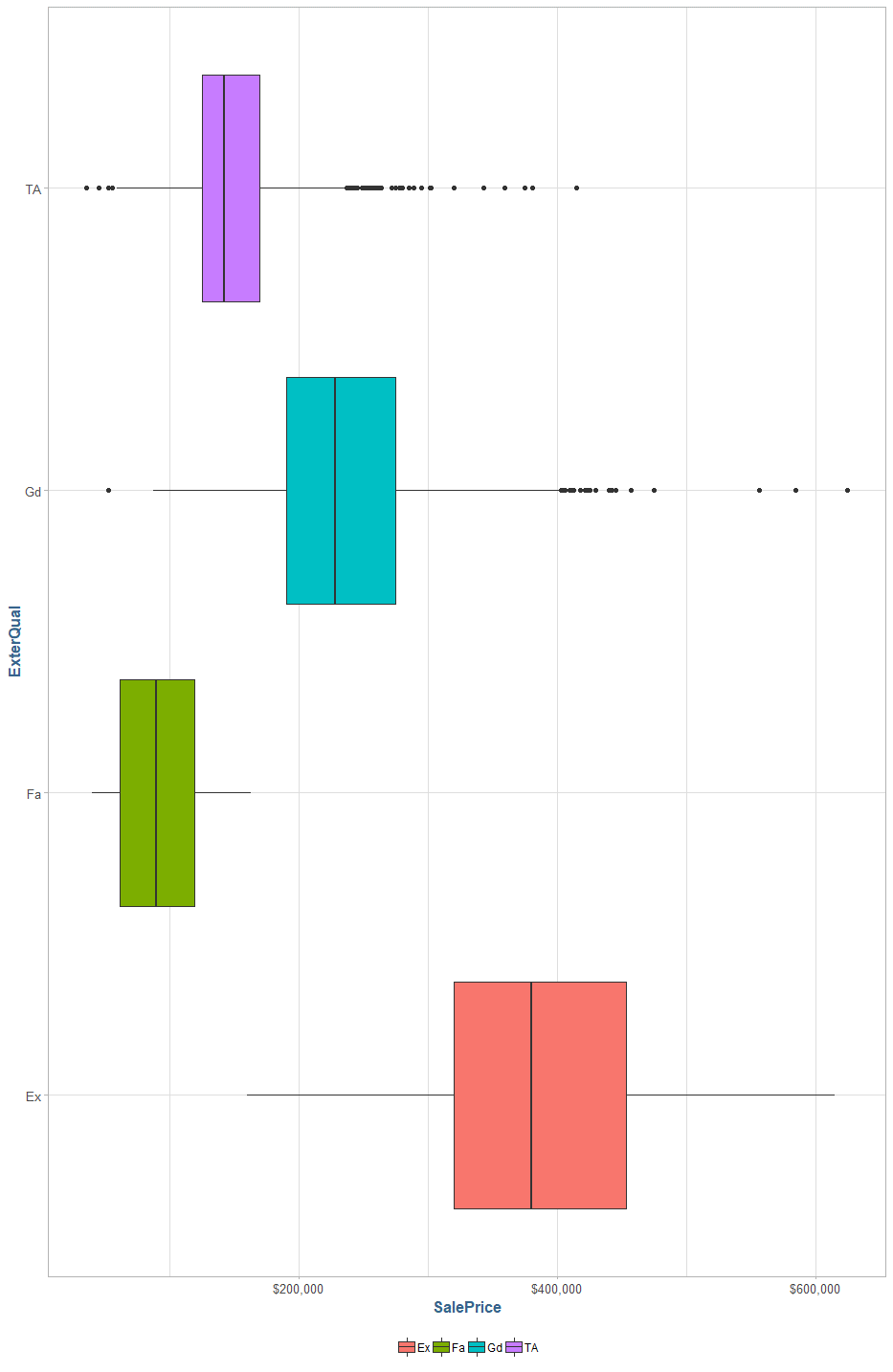
The predictive intervals for sale price based upon kitchen quality are tightly bound intervals, with the exception of the “poor” group, which has a large prediction interval:



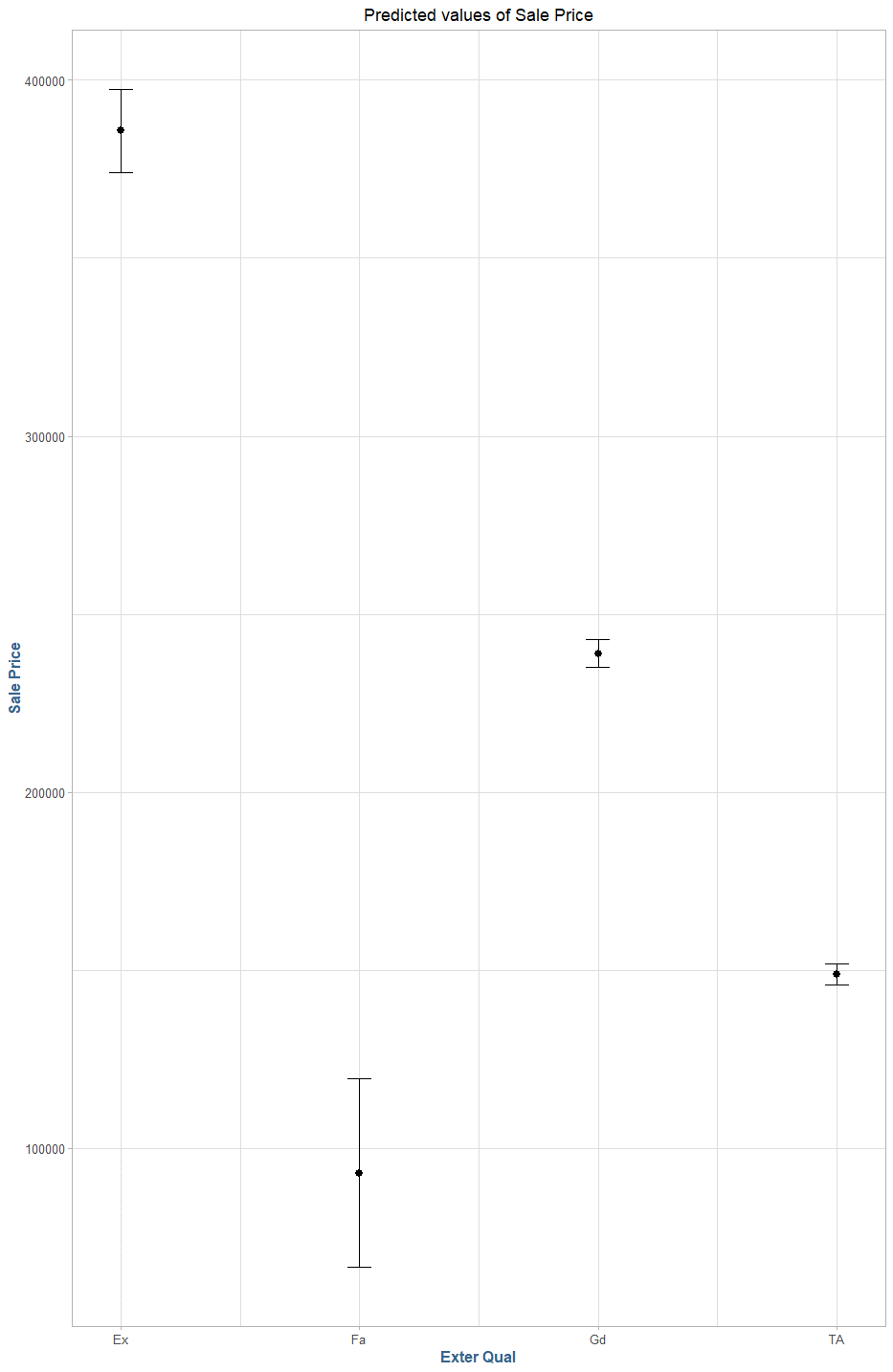
Additionally, like with the basement quality variable, all of the coefficients generated by the linear model appear to have statistical significance:



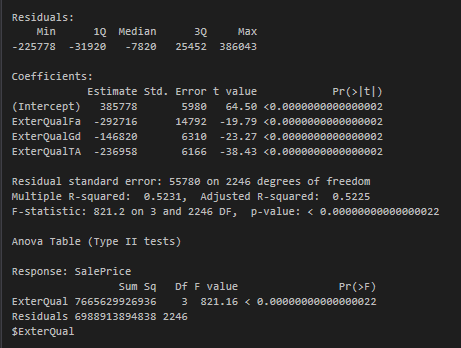
We will keep the kitchen quality variable for further analysis, and we will also examine the last individual quality indicator variable, exterior quality (**ExterQual**). The model generated by external quality variable yields a model with a moderately high R2 and a low residual standard error (.5231 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:



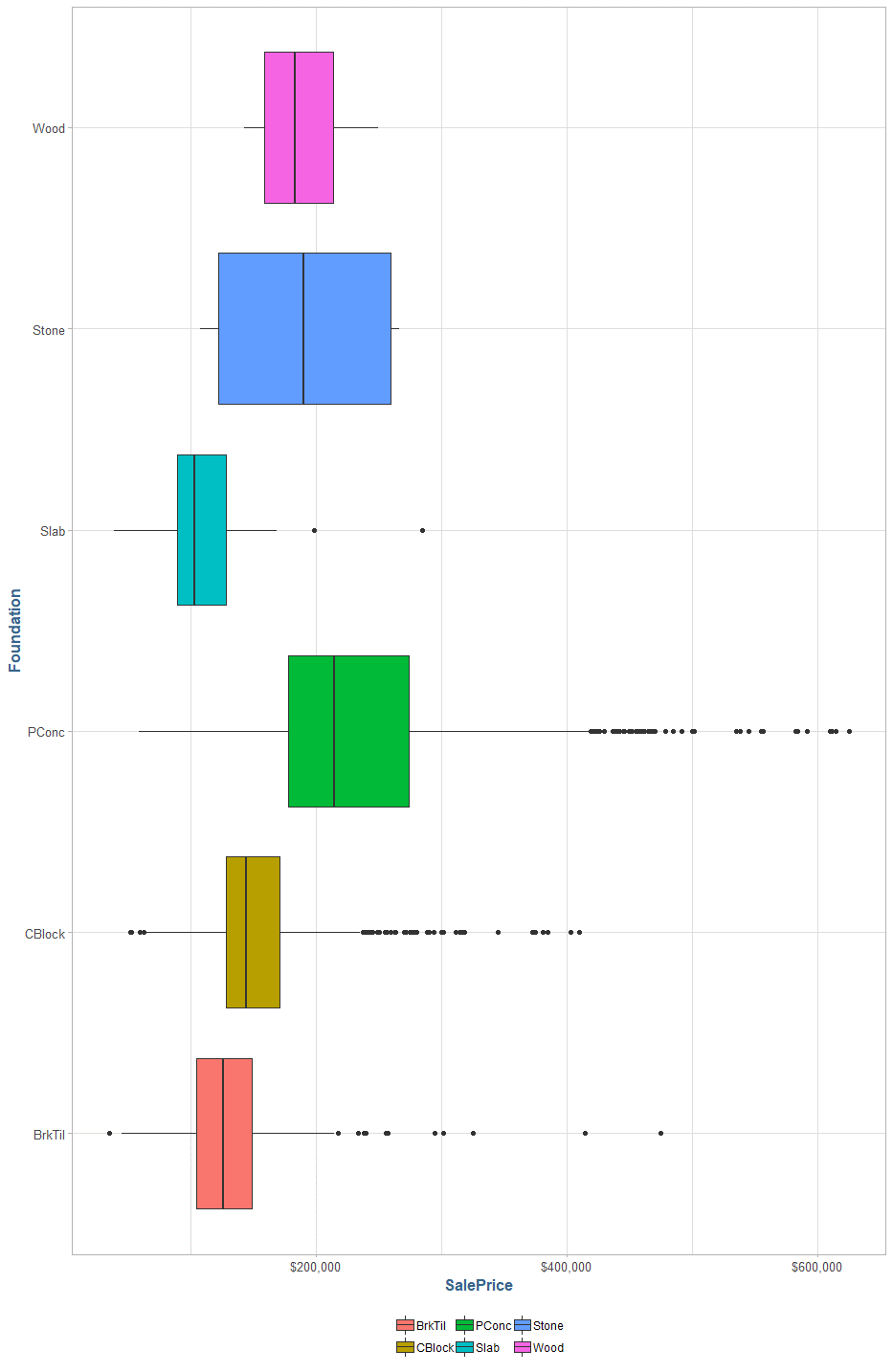
While there are outliers, there does seem to be a significant amount of clustering for the sale price. Additionally, the prediction intervals are small and dispersed as we would hope:



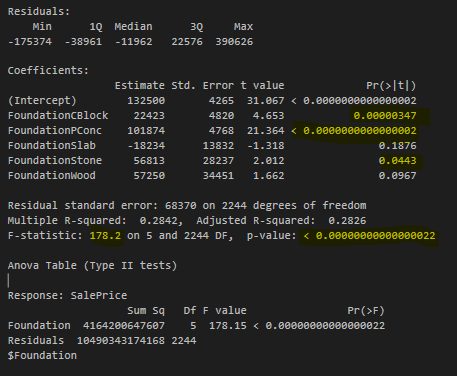
The following model diagnostics also confirm the statistical significance between the different levels of the exterior quality:



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

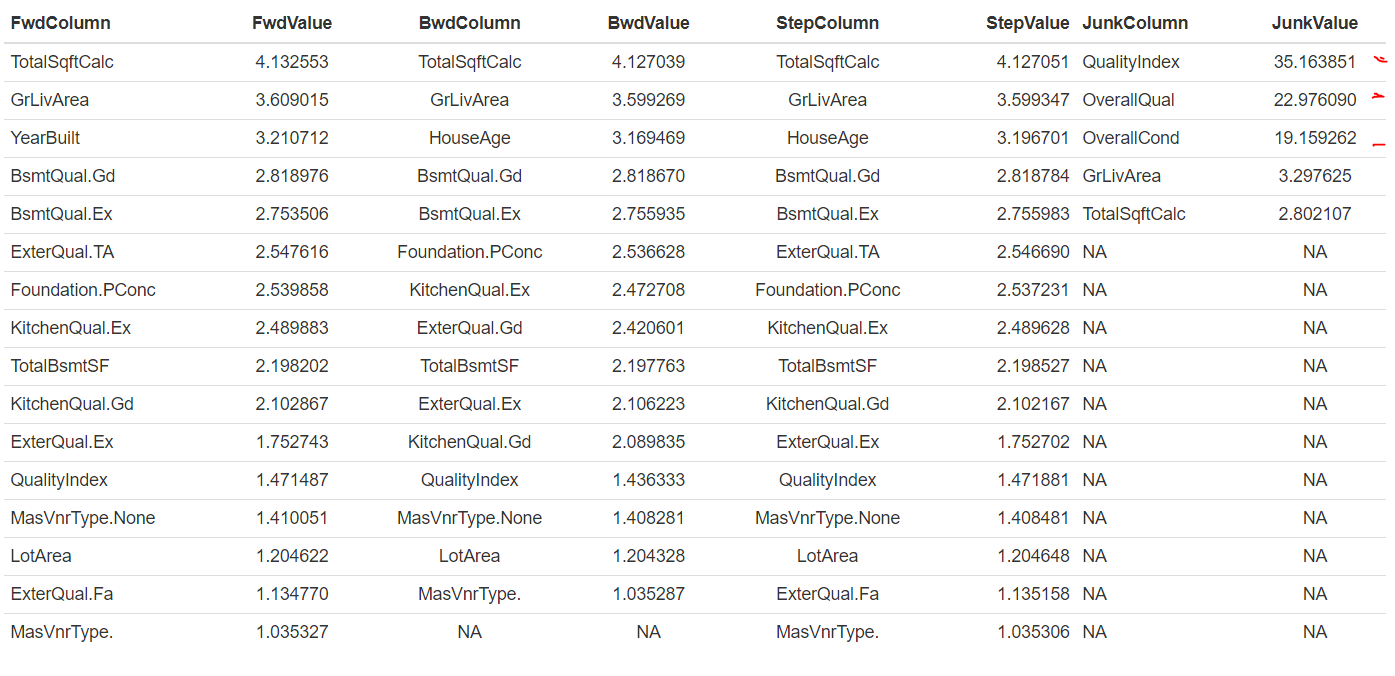
In this section we will split our sample into two parts, a training set and a test set. The purpose of this is to build a model on one set of data, then validate our model on unseen observations in the test set. We will use a ‘standard’ 70/30 split for this exercise, and the number of respective observations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

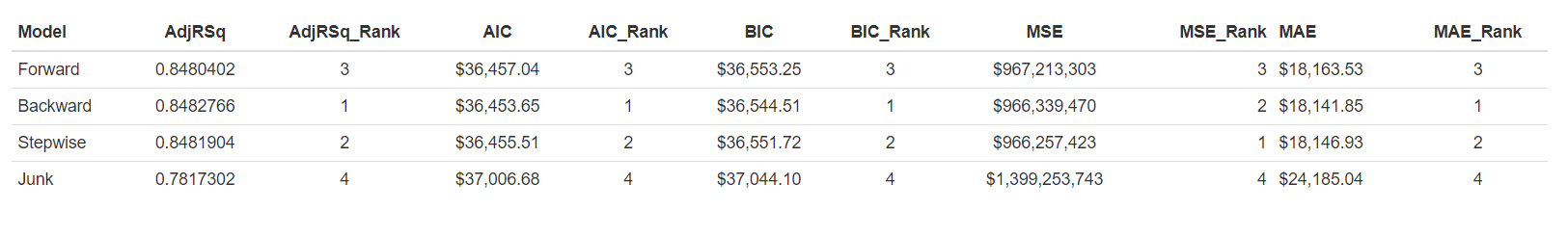
The auto-selection of parameters is not exactly the same for each of the three linear models. In the following figure we can see the columns each process generated, and their corresponding VIF values:



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.

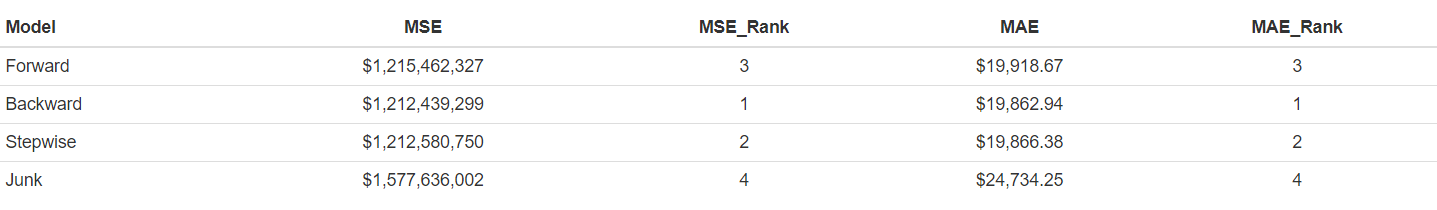


The preceding table shows various measures of fit for the models and the relative ranking for generated value amongst the collection. Both mean squared error and mean absolute error are widely used metrics that measure the average magnitude of a set of errors for a model. The main difference between them is mean absolute error, as the name implies, is agnostic to the direction of the error. The adjusted R2, Akaike Information Criterion and Bayesian Information Criterion are all measures for assessing model fit, although they report distinctly different meanings. Adjusted R2 measures how well the model fits the observed data and penalizes for unnecessary variables in the model and reports a number between 0 and 1, with 1 being 100%, on how much variance is explained by the model and can be interpreted as a percentage. The AIC and BIC scores are closely related in that they report an estimate of how well the model will predict new data, and an estimate for how much information is lost in a given model, and they also have a penalization for unnecessary predictors included in the model.

All of these metrics are important measures for the overall quality of the model, and we should give consideration to each of them, however as we are building a predictive model the AIC and BIC measures should be given more weight. In the model presented here one model happens to be ranked first in all criteria, the backward model is ranked first in all categories for the training data set.

### PRedictive accuracy

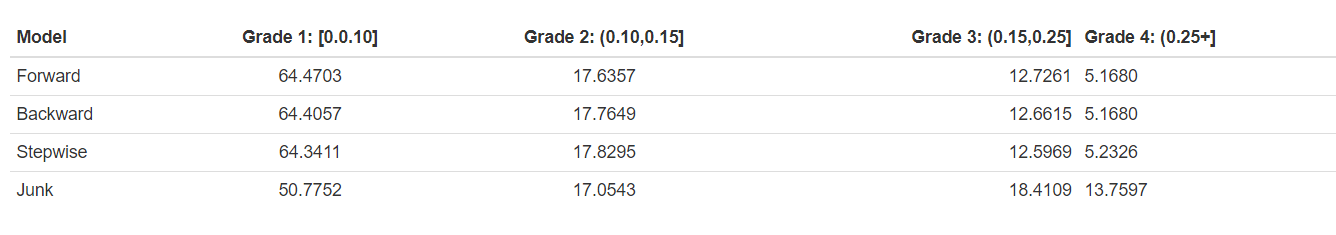
Now we will test each of the models on out of sample data, that is data that the models have not seen before.



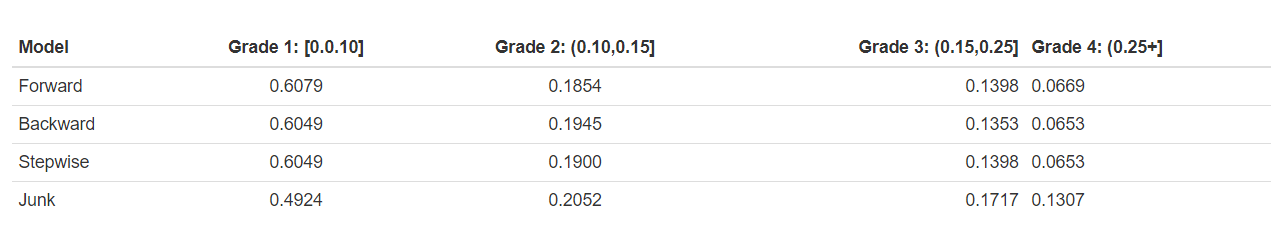
The preceding table summarizes the mean squared error and mean absolute error for each of the models on the test set of data which has not been previously seen by the models. We note again that the ‘backward’ generated model had the top performance in the in-sample test set as well as the out-sample which we can see above.

### operational validation

In a statistical sense all of the metrics above are valid for our evaluation, however, they do not translate easily to the business. Although, we should note here that as far as interpretability reporting the mean absolute error as the average prediction error is much more explainable than the mean squared error. For an even more interpretable evaluation of the model accuracy, we can look at the distribution of predictions grades for each model. The prediction grade is determined by the percent difference of the model predicted value vs the actual value, bucketed into 4 groups: 0-10, 10-15, 15-25 and anything over 25. These grades are for the in-sample, or training data, for each of the respective models.



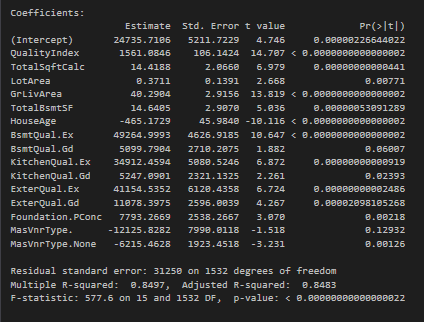
We can see the same metric for the out-of-sample, or test data, grades below:



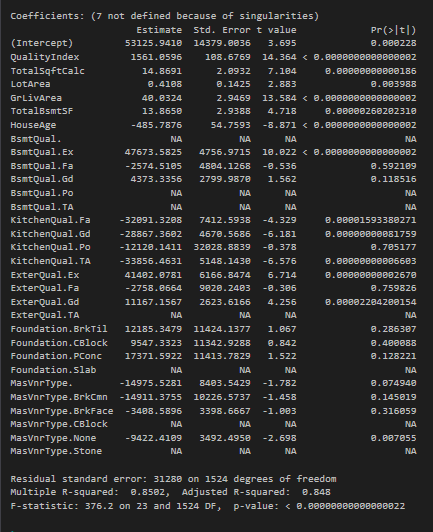
We see that the prediction accuracy for a grade ‘1’ which is a prediction that falls within 10% of the actual home value gets reduced across the board. Interestingly, the ‘Junk’ or baseline model saw the lease decrease in ‘1’ grades, although it is still the worst performing model across the board. Interesting, we note that for this metric of grading the relative accuracy of the predictions the Forward version of the model performed the best in both in-sample and out-sample, beating the previously unanimous Backward model by thirty basis points.

### Revision

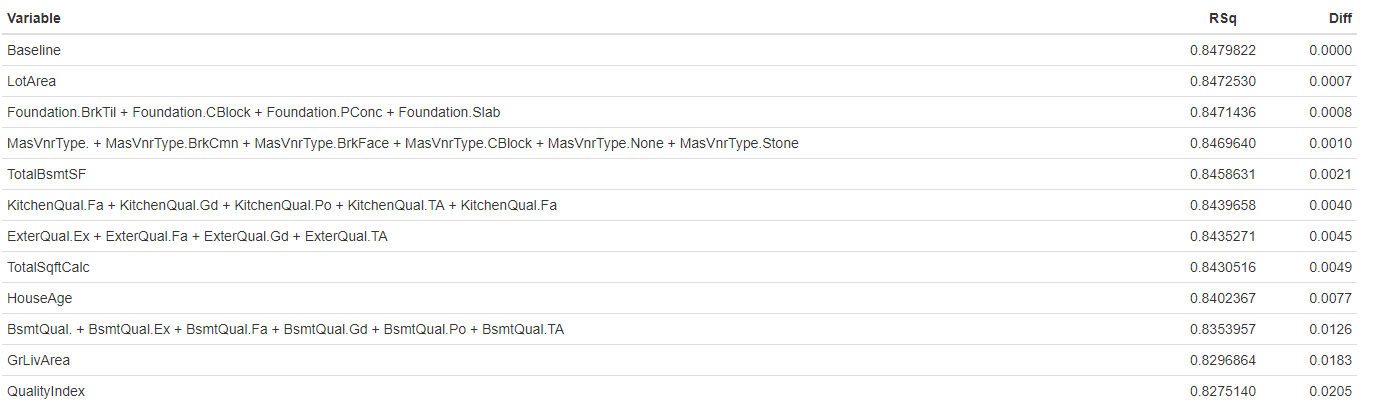
At this point we are going to pick the ‘Backward’ generated model and perform a deep dive on the model parameters and regression diagnostics. First, we want to examine each of the coefficients generated by the backward parameter selection technique. The auto selection technique selected fifteen variables from our data set to predict the price of a home, which we can see here:



The variable selection technique picked some of the columns from the dummy coded variables such as basement quality, kitchen quality and masonry veneer type, however, not all the categories were included. We will include all the columns and then re-evaluate. The baseline version of our ‘final’ model will include the following coefficient terms:

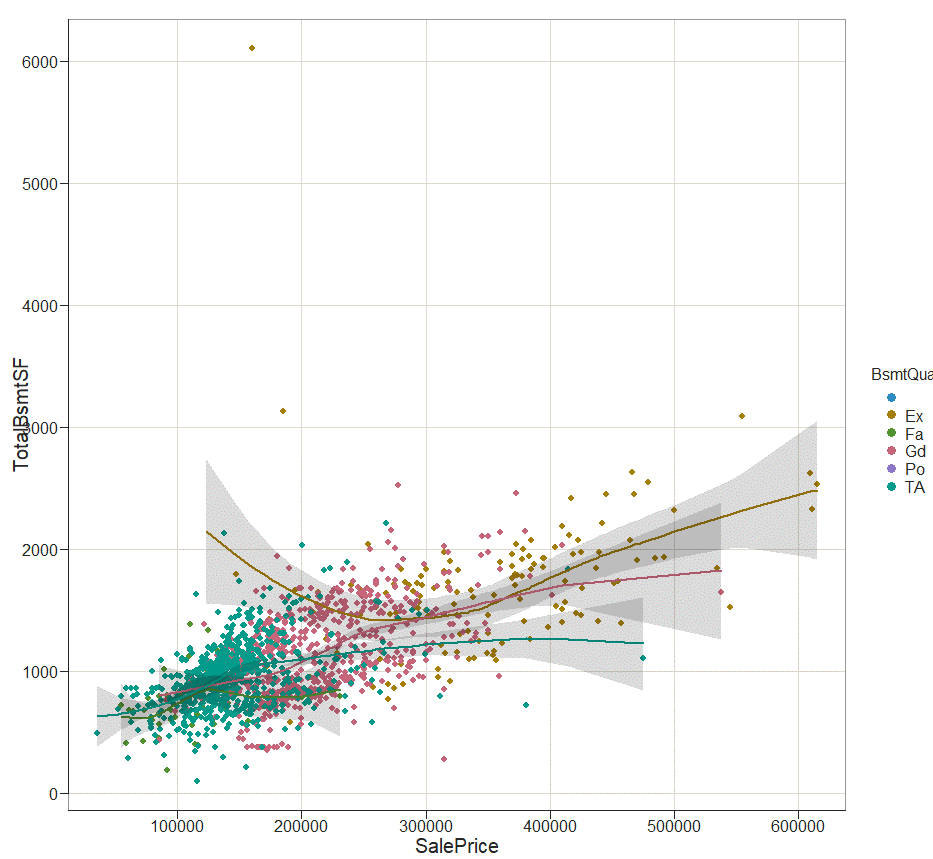


We will now go through and examine the change in R2 by removing the above terms that have the highest probability of being extraneous variables. The following table summarizes the change in R2 by removing the term in variable column:



In the above table we note a few variables that have negligible impact on the R2 of the model, which are listed in order of impact from smallest to largest. For the final model we will remove the lot area, foundation type and masonry veneer type from the model due to their close to zero impact on the R2 score and the additional increased probability of overfitting due to the increased model complexity.

Since there could be an interaction between the size of the basement (**TotalBsmtSF**) and the basement quality (**BsmtQual**), we will test for interaction with the unequal slopes method. Below, we can see the interaction plot of the basement size and the basement quality:



We will model these terms separately from the final model to test for interaction between these two variables separately. The full model is defined as follows:

Ŷ = 94,710.92 + 52.56β1 + 155,167.34β2 + 49,038.22β3 + 21,946.79β4 – 0.97β5 – 92.9β6 + 29.41β7

For this model we can interpret the intercept of $94,710 as a placeholder value as it is the mean sale price for homes is approximately $200,000 in the dataset, it is too far below a reasonable value to be interpreted when the basement size is zero. Then for each 1 unit increase in basement square footage we increase the sale price by approximately $52 per square foot, and $155,168 if it has excellent quality, $48,945 if it is fair quality and $21,976 if it has good quality.

This compares to the reduced model,

Ŷ = 81,477.56 + 66.41β1 + 143,701.97β2 - 12,709.5β3 +52,849.7β3

Where again the intercept can be interpreted as a placeholder value due to its small size relative to the data set, and here we would add $66.41 per square foot of basement size regardless of quality, and add $143,701 if its excellent quality, subtract $12,709 if its fair quality and add $52,849 if its good quality.

F = (1,630,694,755,062 – 1,603,752,680,650) / 3 ) / (1,603,752,680,650 / 644)

= **3.6399**

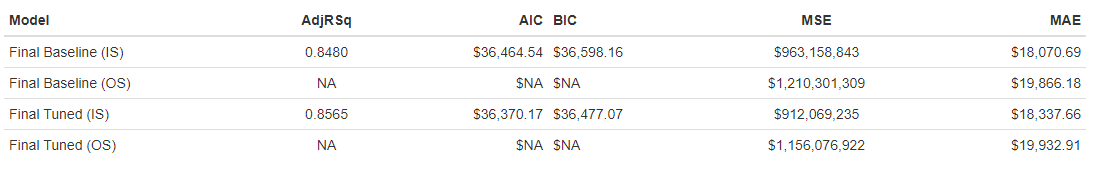
And at a 90% confidence level (F3,544 = **2.6187**), we can reject the null hypothesis that the more complex model here is the better fit, and we will include these interaction terms in the final model.

The final model can be defined as:

Ŷ = 59,158.39 + 1,493.27β1 + 17.36β2 + 14.07β3 + 41.96β3 - 536β4 + 116,464.2β5 + 18536.05β6 – 16,983.75β7 – 31,227.05β8 – 29,798β9 – 4,493.25β10 – 34,782.47β11 + 55,582.55β12 – 4,783.73β13 + 10,060β14 -40.7β15 –27.1β16 + 21.72β17

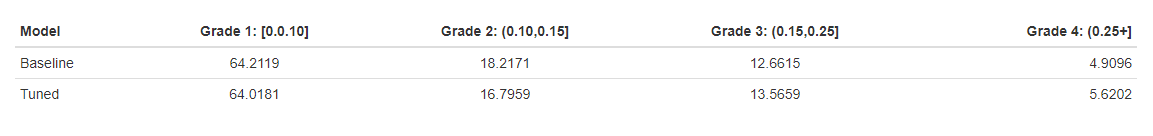
Which we can interpret as the intercept of $59,158 must be a placeholder value in the model since it is well below any reasonable value of a given home. And for each unit of quality index (on a scale of 1-10), we can add $1,493 and for each square foot of the home we increase the value by $17.36. For the basement we can add $14.07 per square foot, and $116,432 if its excellent quality, $18,508 if its fair quality, and subtract $16,962 if its good quality. We can also add approximately $42 per square foot of above ground living area, subtract $536 for every year older the home is. For the kitchen we can subtract $31,227 if its fair quality, $29.798 if its good quality, $34,782 if it’s typical and $4,4493 if it’s poor quality.

The below table summarizes the baseline metrics for the baseline and tuned models, both in and out of sample:

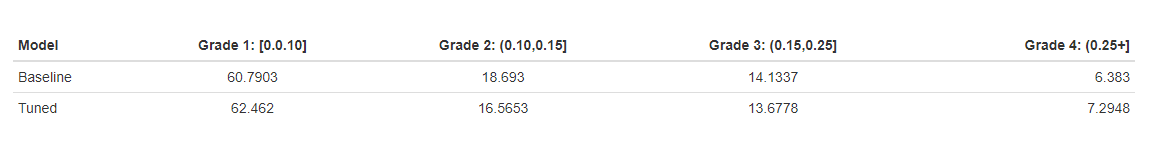


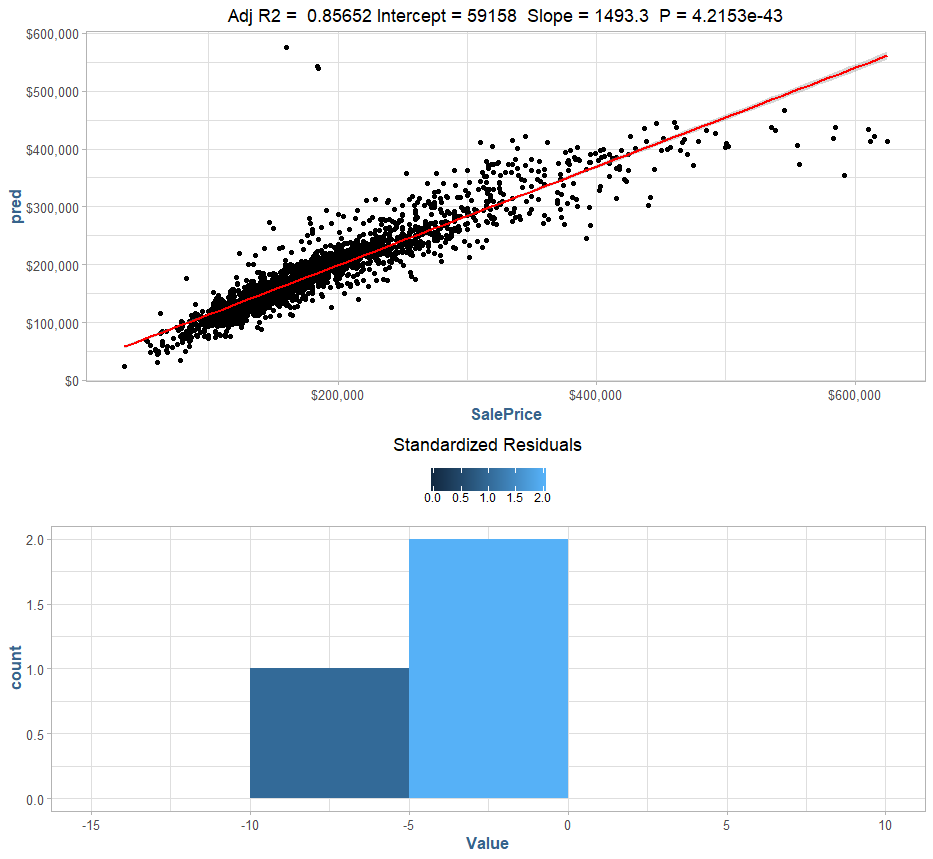
We can see that our tuning helped improve the explained variance of the data, as well as lower the mean squared error for both in and out of sample data. The tuned model has a slightly higher mean absolute error; however, both the AIC and BIC scores are lower in the tuned version. Now that we have seen some statistical improvements in the model, let’s revisit our business case with the grading of the individualized prediction scores:

For the in-sample data, we note the following results which is essentially no change from the baseline version to the tuned version:

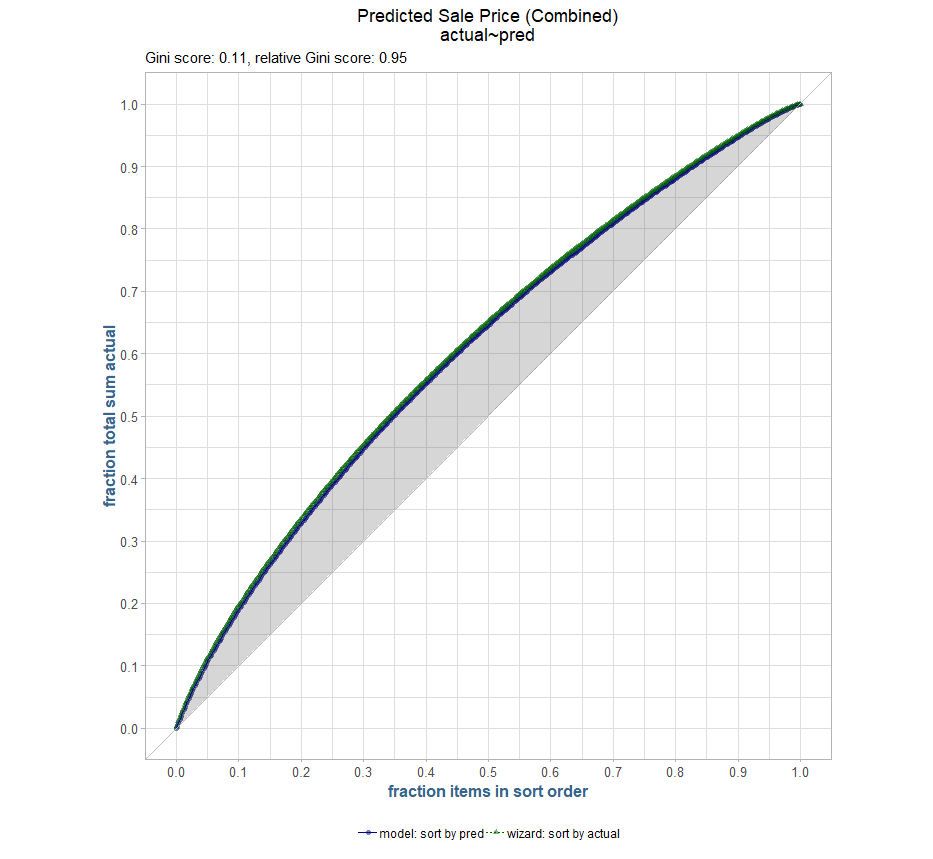


However, in data the model hasn’t seen before we notice immediately a 2% increase in the Grade 1 category, which we are trying to get as high as possible.



Now, let’s finally look our tuned model on the full data set: 

In the top chart we see the predicted sale price vs the actual sale price, and we have normally distributed residuals as we would expect. Additionally, we can see that compared to a ‘perfect’ model denoted by the green line in the below chart, our model, the blue line, comes close to the ideal model. We also note a .95 Gini score that ranks a models predictive accuracy:



This is the best model we have produced thus far, in both the statistical sense and the business sense.

### Conclusion

In this lab we began by looking closer at some categorical variables we have previously excluded from our analysis due to lack of proper technique for handling them. We explored some categorical variables that intuitively should have some relevance on the sale price of a given home, of which we picked a few for further analysis. Next, we partitioned our data set with a standard 70/30 split to develop a training / testing predictive modeling framework. Using this framework, we ran 3 separate auto variable selection modeling techniques, and then developed a baseline ‘junk’ model for comparative purposes. We then explored many different statistical measures of ‘good-ness of fit’ including the familiar R2, the Akaike Information Criterion (AIC), the Bayesian Information Criterion, and the standard mean squared error and mean absolute error. From these statistical measures, we then look at an operational validation procedure for a business case of model validation which measures the percentage of error in the predictive accuracy and cut them into buckets. We evaluated the models using both frameworks, then selected the one we felt fit both cases the best holistically.

After we generated a ‘best’ model, we then looked for categorical variables that had missing baseline cases in the model and added them where appropriate. From that exhaustive model, we looked at the change in R2 from the final model terms and removed ones that seemed to add little value in favor of a simpler model. The final step in the model building process was testing for interactions between our categorical variables and our continuous variables, which we found an interaction that helped boost the predictive accuracy of our model.

After our we finished tuning our model, we then re-examined both version of the final model, pre-and-post the tuning process and found that there was both improvement in the statistical sense and the business case sense, which validated our work. We then examined the final model on the full data set, validated the residual distributions and looked at prediction curves for the model. We concluded this lab with the most robust and accurate model of the housing data thus far.

### Appendix