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.

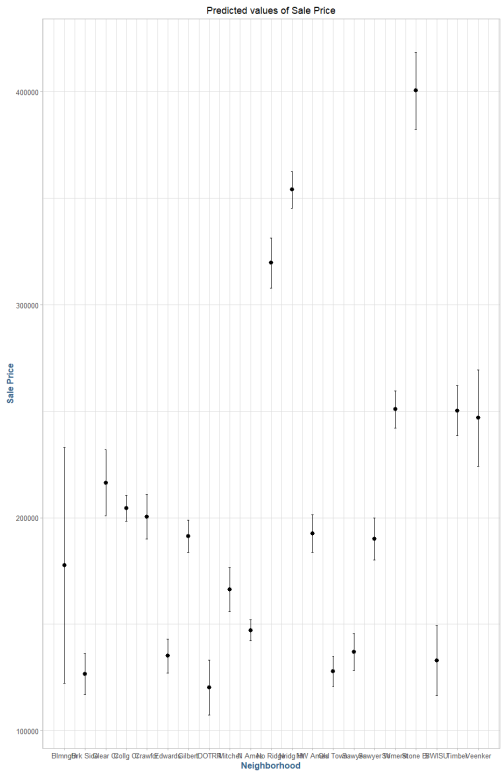
In this report, we will use the Ames dataset which is an alternative to the famous Boston housing data to perform exploratory data analysis through variable derivation, validation, selection and visualization to measure the relevance of these indicators as they pertain to the value of the home in terms 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 R^2, 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.

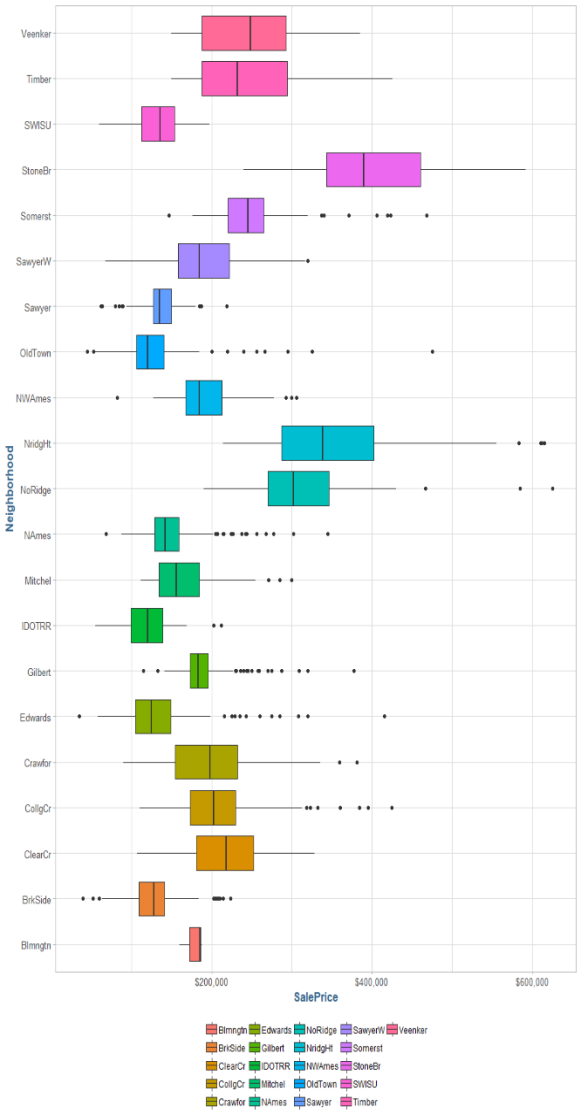
The reason we chose these metrics is due to the variance explained by each category is an indicator of the relative "goodness-of-fit", and the RSE and mean difference give us a sense of the variance found in each of the levels, where the lower the variance and higher the R^2 will give us a good idea 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.

The first categorical variable we will explore is the neighborhood variable, as it has both a high R^2 (.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:



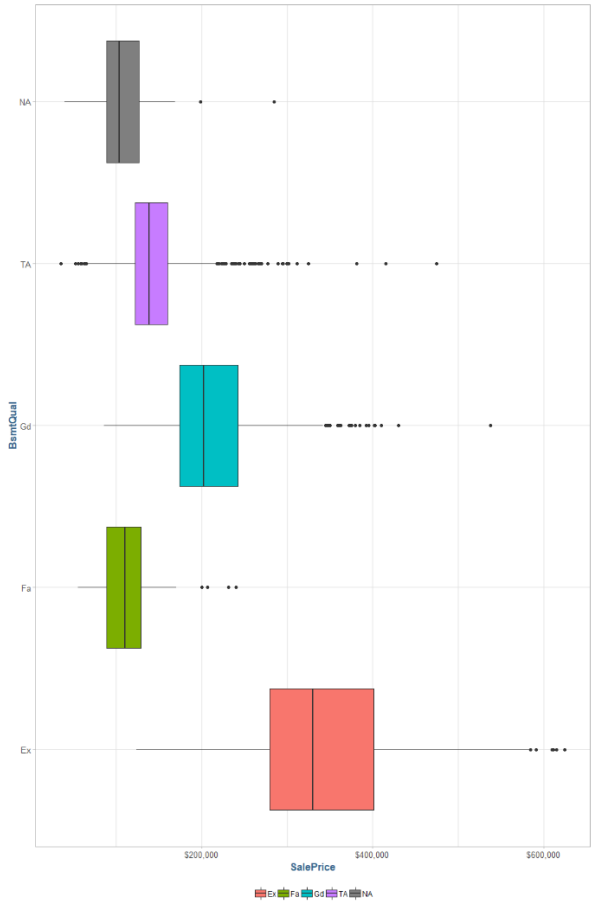
Unfortunately, the model predicted values for each neighborhood has a large interval of values that they could fall into. The following table summarizes each neighborhood by mean sale price:

Neighborhood	MeanPrice
Blmngtn	$177,689
BrkSide	$126,593
ClearCr	$216,417
CollgCr	$204,354
Crawfor	$200,527
Edwards	$135,078
Gilbert	$191,278
IDOTRR	$120,188
Mitchel	$166,295
NAmes	$147,157
NoRidge	$319,616
NridgHt	$353,990
NWAmes	$192,478
OldTown	$127,828
Sawyer	$136,980
SawyerW	$190,008
Somerst	$250,835
StoneBr	$400,546
SWISU	$132,984
Timber	$250,326
Veenker	$246,797

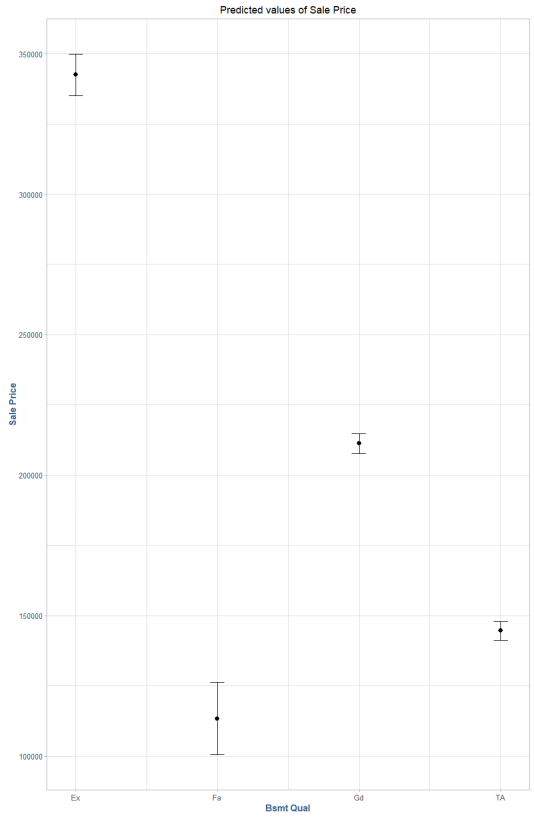


Looking at the modeling data, we can see that the outliers in sale price per neighborhood are vast which would explain the large variances in the model. We also note that of the twenty coefficient terms generated by the model, only five of them had significant p-values below the 5% level.

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 R^2 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:

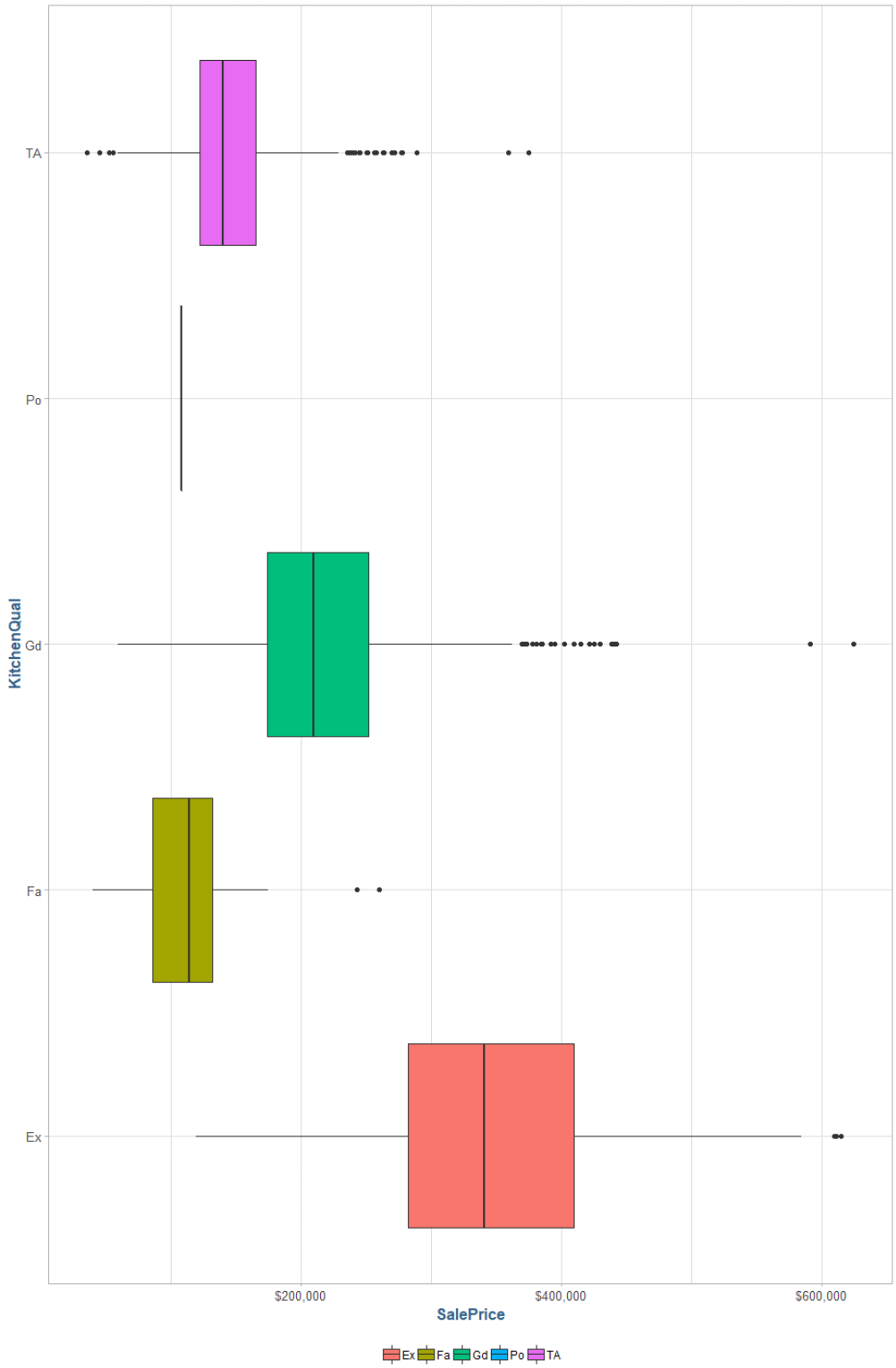


In the above graphic we see that each of the prediction intervals fall within a relatively tight bound. We can also look at the model diagnostics and see that each of the coefficient terms generated by the linear model have statistical significance with low standard errors (in the 3-7-thousand-dollar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

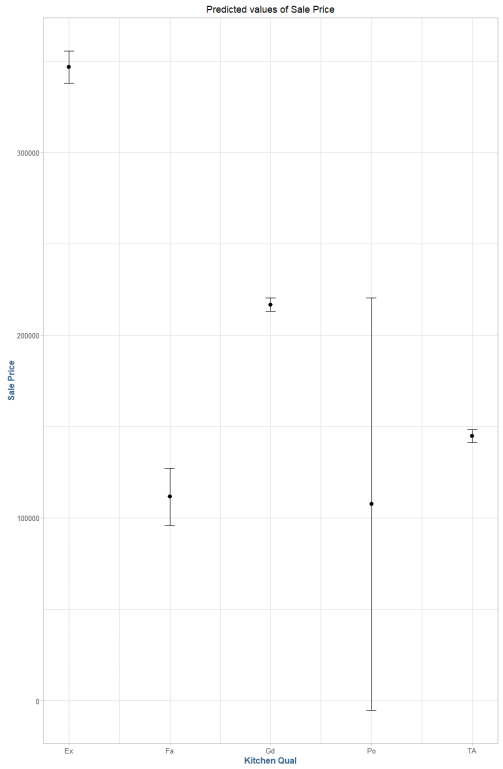
```
Residuals:
    Min      1Q  Median      3Q     Max
-218977  -30650   -7361   23335  330335

Coefficients:
             Estimate Std. Error t value            Pr(>|t|)
(Intercept)   342477       3769   90.87 <0.0000000000000002
BsmtQualFa   -229105       7565  -30.28 <0.0000000000000002
BsmtQualGd   -131169       4180  -31.38 <0.0000000000000002
BsmtQualTA   -197813       4138  -47.81 <0.0000000000000002

Residual standard error: 54490 on 2202 degrees of freedom
  (44 observations deleted due to missingness)
Multiple R-squared:  0.5431,  Adjusted R-squared:  0.5425
F-statistic: 872.5 on 3 and 2202 DF,  p-value: < 0.00000000000000022

Anova Table (Type II tests)

Response: SalePrice
               Sum Sq   Df F value               Pr(>F)
BsmtQual  7771311386518    3  872.53 < 0.00000000000000022
Residuals 6537454702053 2202
$BsmtQual
```

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 R^2 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:

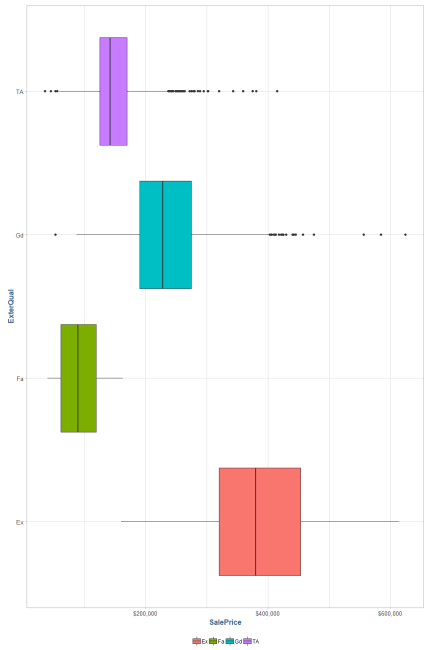


Additionally, like with the basement quality variable, all of the coefficients generated by the linear model appear to have statistical significance:

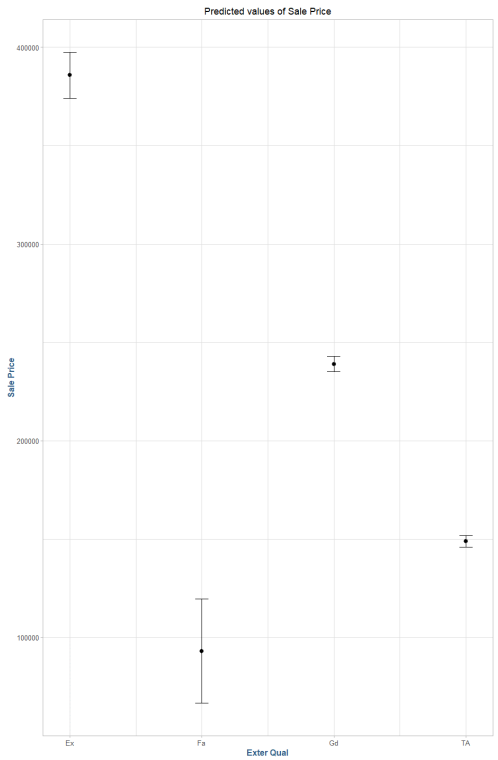
```
Residuals:
    Min      1Q  Median      3Q     Max
-227751  -31834   -5706   27641  408294

Coefficients:
             Estimate Std. Error t value            Pr(>|t|)
(Intercept)    346751       4522  76.673 < 0.0000000000000002
KitchenQualFa  -235079       9174 -25.623 < 0.0000000000000002
KitchenQualGd  -130045       4900 -26.539 < 0.0000000000000002
KitchenQualPo  -239251      57739  -4.144           0.0000354
KitchenQualTA  -201917       4843 -41.694 < 0.0000000000000002

Residual standard error: 57560 on 2245 degrees of freedom
Multiple R-squared:  0.4924,  Adjusted R-squared:  0.4915
F-statistic: 544.5 on 4 and 2245 DF,  p-value: < 0.00000000000000022

Anova Table (Type II tests)

Response: SalePrice
                 Sum Sq   Df F value               Pr(>F)
KitchenQual 7216197403794    4  544.49 < 0.00000000000000022
Residuals   7438346417981 2245
$KitchenQual
```

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 R^2 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:

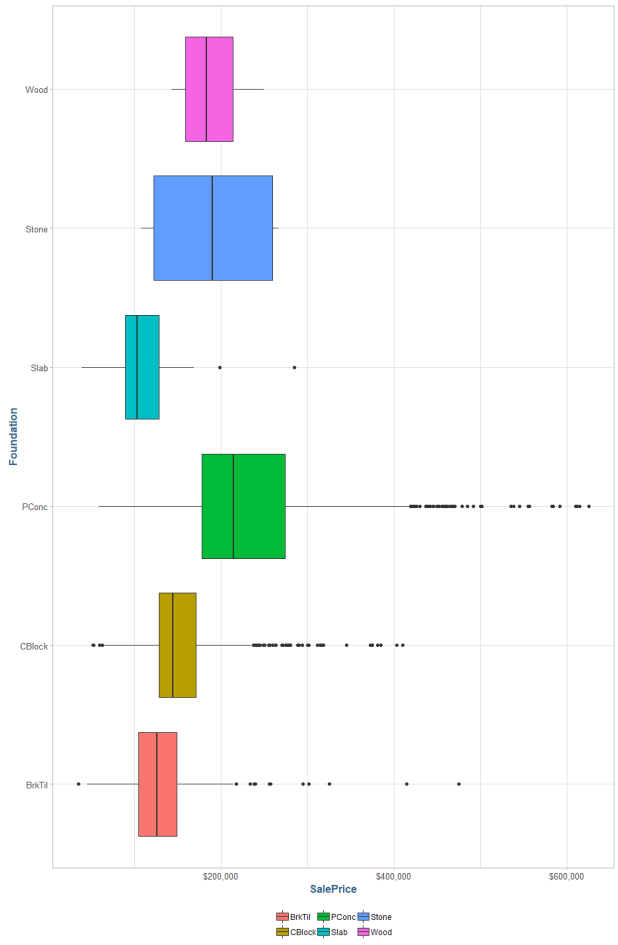


The following model diagnostics also confirm the statistical significance between the different levels of the exterior quality:

```
Residuals:
    Min      1Q  Median      3Q     Max
-225778  -31920   -7820   25452  386043

Coefficients:
            Estimate Std. Error t value            Pr(>|t|)
(Intercept)   385778       5980   64.50 <0.0000000000000002
ExterQualFa  -292716      14792  -19.79 <0.0000000000000002
ExterQualGd  -146820       6310  -23.27 <0.0000000000000002
ExterQualTA  -236958       6166  -38.43 <0.0000000000000002

Residual standard error: 55780 on 2246 degrees of freedom
Multiple R-squared:  0.5231,  Adjusted R-squared:  0.5225
F-statistic: 821.2 on 3 and 2246 DF,  p-value: < 0.00000000000000022

Anova Table (Type II tests)

Response: SalePrice
                Sum Sq   Df F value              Pr(>F)
ExterQual 7665629926936    3  821.16 < 0.00000000000000022
Residuals 6988913894838 2246
$ExterQual
```

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 R^2 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:

```
Residuals:
    Min      1Q  Median      3Q     Max
-175374  -38961  -11962   22576  390626

Coefficients:
                Estimate Std. Error t value             Pr(>|t|)
(Intercept)       132500       4265  31.067 < 0.0000000000000002
FoundationCBlock   22423       4820   4.653          0.00000347
FoundationPConc   101874       4768  21.364 < 0.0000000000000002
FoundationSlab    -18234      13832  -1.318               0.1876
FoundationStone    56813      28237   2.012               0.0443
FoundationWood     57250      34451   1.662               0.0967

Residual standard error: 68370 on 2244 degrees of freedom
Multiple R-squared:  0.2842,   Adjusted R-squared:  0.2826
F-statistic: 178.2 on 5 and 2244 DF,  p-value: < 0.00000000000000022

Anova Table (Type II tests)

Response: SalePrice
                    Sum Sq   Df F value               Pr(>F)
Foundation  4164200647607    5  178.15 < 0.00000000000000022
Residuals  10490343174168 2244
$Foundation
```

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:

Total	Train	Test
2206	1548	658

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

Column	Type	Variable
SalePrice	*Continious*	*Response*
QualityIndex	Ordinal	Quality Index
TotalSqftCalc	Continious	Total Sf (Calcualted)
YearBuilt	Temporal Discrete	Year Built
YearRemodel	Temporal Discrete	Year Remodel
LotArea	Continious	Lot Area
GrLivArea	Continious	Greater Living Area
TotalBath	Discrete	Total Bath
TotalBsmtSF	Continious	Basement Square Footage
HouseAge	Ordinal	House Age
FullBath	Discrete	Full Bath
HalfBath	Discrete	Half Bath
BsmtQual.	Dummy Coded	
BsmtQual.Ex	Dummy Coded	
BsmtQual.Fa	Dummy Coded	Basement Quality
BsmtQual.Gd	Dummy Coded	
BsmtQual.Po	Dummy Coded	
BsmtQual.TA	Dummy Coded	
KitchenQual.Ex	Dummy Coded	
KitchenQual.Fa	Dummy Coded	
KitchenQual.Gd	Dummy Coded	Kitchen Quality
KitchenQual.Po	Dummy Coded	
KitchenQual.TA	Dummy Coded	
ExterQual.Ex	Dummy Coded	
ExterQual.Fa	Dummy Coded	Exterior Quality
ExterQual.Gd	Dummy Coded	
ExterQual.TA	Dummy Coded	
Foundation.BrkTil	Dummy Coded	
Foundation.CBlock	Dummy Coded	
Foundation.PConc	Dummy Coded	Foundation
Foundation.Slab	Dummy Coded	
Foundation.Stone	Dummy Coded	
Foundation.Wood	Dummy Coded	
MasVnrType.	Dummy Coded	
MasVnrType.BrkCmn	Dummy Coded	
MasVnrType.BrkFace	Dummy Coded	Masony Veneer Type
MasVnrType.CBlock	Dummy Coded	
MasVnrType.None	Dummy Coded	
MasVnrType.Stone	Dummy Coded	

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:

FwdColumn	FwdValue	BwdColumn	BwdValue	StepColumn	StepValue	JunkColumn	JunkValue
TotalSqftCalc	4.132553	TotalSqftCalc	4.127039	TotalSqftCalc	4.127051	QualityIndex	35.163851
GrLivArea	3.609015	GrLivArea	3.599269	GrLivArea	3.599347	OverallQual	22.976090
YearBuilt	3.210712	HouseAge	3.169469	HouseAge	3.196701	OverallCond	19.159262
BsmtQual.Gd	2.818976	BsmtQual.Gd	2.818670	BsmtQual.Gd	2.818784	GrLivArea	3.297625
BsmtQual.Ex	2.753506	BsmtQual.Ex	2.755935	BsmtQual.Ex	2.755983	TotalSqftCalc	2.802107
ExterQual.TA	2.547616	Foundation.PConc	2.536628	ExterQual.TA	2.546690	NA	NA
Foundation.PConc	2.539858	KitchenQual.Ex	2.472708	Foundation.PConc	2.537231	NA	NA
KitchenQual.Ex	2.489883	ExterQual.Gd	2.420601	KitchenQual.Ex	2.489628	NA	NA
TotalBsmtSF	2.198202	TotalBsmtSF	2.197763	TotalBsmtSF	2.198527	NA	NA
KitchenQual.Gd	2.102867	ExterQual.Ex	2.106223	KitchenQual.Gd	2.102167	NA	NA
ExterQual.Ex	1.752743	KitchenQual.Gd	2.089835	ExterQual.Ex	1.752702	NA	NA
QualityIndex	1.471487	QualityIndex	1.436333	QualityIndex	1.471881	NA	NA
MasVnrType.None	1.410051	MasVnrType.None	1.408281	MasVnrType.None	1.408481	NA	NA
LotArea	1.204622	LotArea	1.204328	LotArea	1.204648	NA	NA
ExterQual.Fa	1.134770	MasVnrType.	1.035287	ExterQual.Fa	1.135158	NA	NA
MasVnrType.	1.035327	NA	NA	MasVnrType.	1.035306	NA	NA

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.

Model	AdjRSq	AdjRSq_Rank	AIC	AIC_Rank	BIC	BIC_Rank	MSE	MSE_Rank	MAE	MAE_Rank
Forward	0.8480402	3	$36,457.04	3	$36,553.25	3	$967,213,303	3	$18,163.53	3
Backward	0.8482766	1	$36,453.65	1	$36,544.51	1	$966,339,470	2	$18,141.85	1
Stepwise	0.8481904	2	$36,455.51	2	$36,551.72	2	$966,257,423	1	$18,146.93	2
Junk	0.7817302	4	$37,006.68	4	$37,044.10	4	$1,399,253,743	4	$24,185.04	4

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 R^2, Akaike Information Criterion and Bayesian Information Criterion are all measures for assessing model fit, although they report distinctly different meanings. Adjusted R^2 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.

Model	MSE	MSE_Rank	MAE	MAE_Rank
Forward	$1,215,462,327	3	$19,918.67	3
Backward	$1,212,439,299	1	$19,862.94	1
Stepwise	$1,212,580,750	2	$19,866.38	2
Junk	$1,577,636,002	4	$24,734.25	4

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.

Model	Grade 1: [0.0.10]	Grade 2: (0.10,0.15]	Grade 3: (0.15,0.25]	Grade 4: (0.25+]
Forward	64.4703	17.6357	12.7261	5.1680
Backward	64.4057	17.7649	12.6615	5.1680
Stepwise	64.3411	17.8295	12.5969	5.2326
Junk	50.7752	17.0543	18.4109	13.7597

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

Model	Grade 1: [0.0.10]	Grade 2: (0.10,0.15]	Grade 3: (0.15,0.25]	Grade 4: (0.25+]
Forward	0.6079	0.1854	0.1398	0.0669
Backward	0.6049	0.1945	0.1353	0.0653
Stepwise	0.6049	0.1900	0.1398	0.0653
Junk	0.4924	0.2052	0.1717	0.1307

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:

```
Coefficients:
                  Estimate  Std. Error t value            Pr(>|t|)
(Intercept)      24735.7106   5211.7229   4.746     0.00000226644022
QualityIndex      1561.0846    106.1424  14.707 < 0.0000000000000002
TotalSqftCalc       14.4188      2.0660   6.979     0.00000000000441
LotArea              0.3711      0.1391   2.668              0.00771
GrLivArea           40.2904      2.9156  13.819 < 0.0000000000000002
TotalBsmtSF         14.6405      2.9070   5.036     0.00000053091289
HouseAge          -465.1729     45.9840 -10.116 < 0.0000000000000002
BsmtQual.Ex      49264.9993   4626.9185  10.647 < 0.0000000000000002
BsmtQual.Gd       5099.7904   2710.2075   1.882              0.06007
KitchenQual.Ex   34912.4594   5080.5246   6.872     0.00000000000919
KitchenQual.Gd    5247.0901   2321.1325   2.261              0.02393
ExterQual.Ex     41154.5352   6120.4358   6.724     0.00000000002486
ExterQual.Gd     11078.3975   2596.0039   4.267     0.00002098105268
Foundation.PConc  7793.2669   2538.2667   3.070              0.00218
MasVnrType.      -12125.8282   7990.0118  -1.518              0.12932
MasVnrType.None   -6215.4628   1923.4518  -3.231              0.00126

Residual standard error: 31250 on 1532 degrees of freedom
Multiple R-squared:  0.8497,  Adjusted R-squared:  0.8483
F-statistic: 577.6 on 15 and 1532 DF,  p-value: < 0.00000000000000022
```

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:

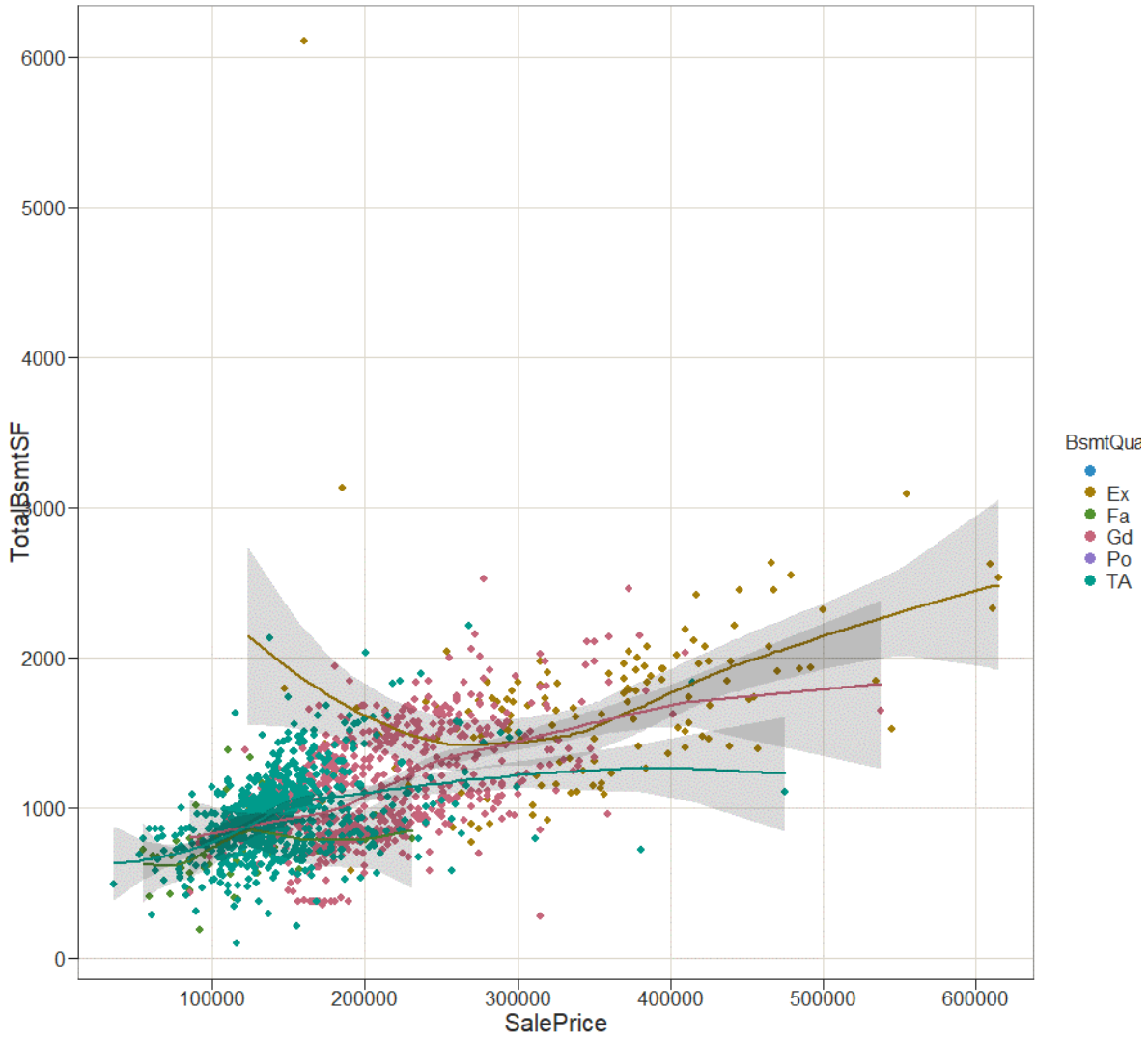
```
Coefficients: (7 not defined because of singularities)
                    Estimate  Std. Error t value             Pr(>|t|)
(Intercept)        53125.9410  14379.0036   3.695             0.000228
QualityIndex        1561.0596    108.6769  14.364 < 0.0000000000000002
TotalSqftCalc         14.8691      2.0932   7.104     0.00000000000186
LotArea                0.4108      0.1425   2.883             0.003988
GrLivArea             40.0324      2.9469  13.584 < 0.0000000000000002
TotalBsmtSF           13.8650      2.9388   4.718     0.00000260202310
HouseAge            -485.7876     54.7593  -8.871 < 0.0000000000000002
BsmtQual.                 NA          NA      NA                   NA
BsmtQual.Ex         47673.5825   4756.9715  10.022 < 0.0000000000000002
BsmtQual.Fa         -2574.5105   4804.1268  -0.536             0.592109
BsmtQual.Gd          4373.3356   2799.9870   1.562             0.118516
BsmtQual.Po                NA          NA      NA                   NA
BsmtQual.TA                NA          NA      NA                   NA
KitchenQual.Fa     -32091.3208   7412.5938  -4.329     0.00001593380271
KitchenQual.Gd     -28867.3602   4670.5686  -6.181     0.00000000081759
KitchenQual.Po     -12120.1411  32028.8839  -0.378             0.705177
KitchenQual.TA     -33856.4631   5148.1430  -6.576     0.00000000006603
ExterQual.Ex        41402.0781   6166.8474   6.714     0.00000000002670
ExterQual.Fa        -2758.0664   9020.2403  -0.306             0.759826
ExterQual.Gd        11167.1567   2623.6166   4.256     0.00002204200154
ExterQual.TA               NA          NA      NA                   NA
Foundation.BrkTil   12185.3479  11424.1377   1.067             0.286307
Foundation.CBlock    9547.3323  11342.9288   0.842             0.400088
Foundation.PConc    17371.5922  11413.7829   1.522             0.128221
Foundation.Slab            NA          NA      NA                   NA
MasVnrType.        -14975.5281   8403.5429  -1.782             0.074940
MasVnrType.BrkCmn  -14911.3755  10226.5737  -1.458             0.145019
MasVnrType.BrkFace  -3408.5896   3398.6667  -1.003             0.316059
MasVnrType.CBlock          NA          NA      NA                   NA
MasVnrType.None     -9422.4109   3492.4950  -2.698             0.007055
MasVnrType.Stone           NA          NA      NA                   NA

Residual standard error: 31280 on 1524 degrees of freedom
Multiple R-squared:  0.8502,  Adjusted R-squared:  0.848
F-statistic: 376.2 on 23 and 1524 DF,  p-value: < 0.00000000000000022
```

We will now go through and examine the change in R^2 by removing the above terms that have the highest probability of being extraneous variables. The following table summarizes the change in R^2 by removing the term in variable column:

Variable	RSq	Diff
Baseline	0.8479822	0.0000
LotArea	0.8472530	0.0007
Foundation.BrkTil + Foundation.CBlock + Foundation.PConc + Foundation.Slab	0.8471436	0.0008
MasVnrType. + MasVnrType.BrkCmn + MasVnrType.BrkFace + MasVnrType.CBlock + MasVnrType.None + MasVnrType.Stone	0.8469640	0.0010
TotalBsmtSF	0.8458631	0.0021
KitchenQual.Fa + KitchenQual.Gd + KitchenQual.Po + KitchenQual.TA + KitchenQual.Fa	0.8439658	0.0040
ExterQual.Ex + ExterQual.Fa + ExterQual.Gd + ExterQual.TA	0.8435271	0.0045
TotalSqftCalc	0.8430516	0.0049
HouseAge	0.8402367	0.0077
BsmtQual. + BsmtQual.Ex + BsmtQual.Fa + BsmtQual.Gd + BsmtQual.Po + BsmtQual.TA	0.8353957	0.0126
GrLivArea	0.8296864	0.0183
QualityIndex	0.8275140	0.0205

In the above table we note a few variables that have negligible impact on the R^2 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 R^2 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:

$$\hat{Y} = 94{,}710.92 + 52.56\beta_1 + 155{,}167.34\beta_2 + 49{,}038.22\beta_3 + 21{,}946.79\beta_4 - 0.97\beta_5 - 92.9\beta_6 + 29.41\beta_7$$

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,

$$\hat{Y} = 81{,}477.56 + 66.41\beta_1 + 143{,}701.97\beta_2 - 12{,}709.5\beta_3 + 52{,}849.7\beta_3$$

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.

$$F = (1{,}630{,}694{,}755{,}062 - 1{,}603{,}752{,}680{,}650) / 3) / (1{,}603{,}752{,}680{,}650 / 644)$$

$$= \mathbf{3.6399}$$

And at a 90% confidence level ($F_{3,544}$ = **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:

$$\hat{Y} = 59{,}158.39 + 1{,}493.27\beta_1 + 17.36\beta_2 + 14.07\beta_3 + 41.96\beta_3 - 536\beta_4 + 116{,}464.2\beta_5 + 18536.05\beta_6 - 16{,}983.75\beta_7 - 31{,}227.05\beta_8 - 29{,}798\beta_9 - 4{,}493.25\beta_{10} - 34{,}782.47\beta_{11} + 55{,}582.55\beta_{12} - 4{,}783.73\beta_{13} + 10{,}060\beta_{14} - 40.7\beta_{15} - 27.1\beta_{16} + 21.72\beta_{17}$$

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:

Model	AdjRSq	AIC	BIC	MSE	MAE
Final Baseline (IS)	0.8480	$36,464.54	$36,598.16	$963,158,843	$18,070.69
Final Baseline (OS)	NA	$NA	$NA	$1,210,301,309	$19,866.18
Final Tuned (IS)	0.8565	$36,370.17	$36,477.07	$912,069,235	$18,337.66
Final Tuned (OS)	NA	$NA	$NA	$1,156,076,922	$19,932.91

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:

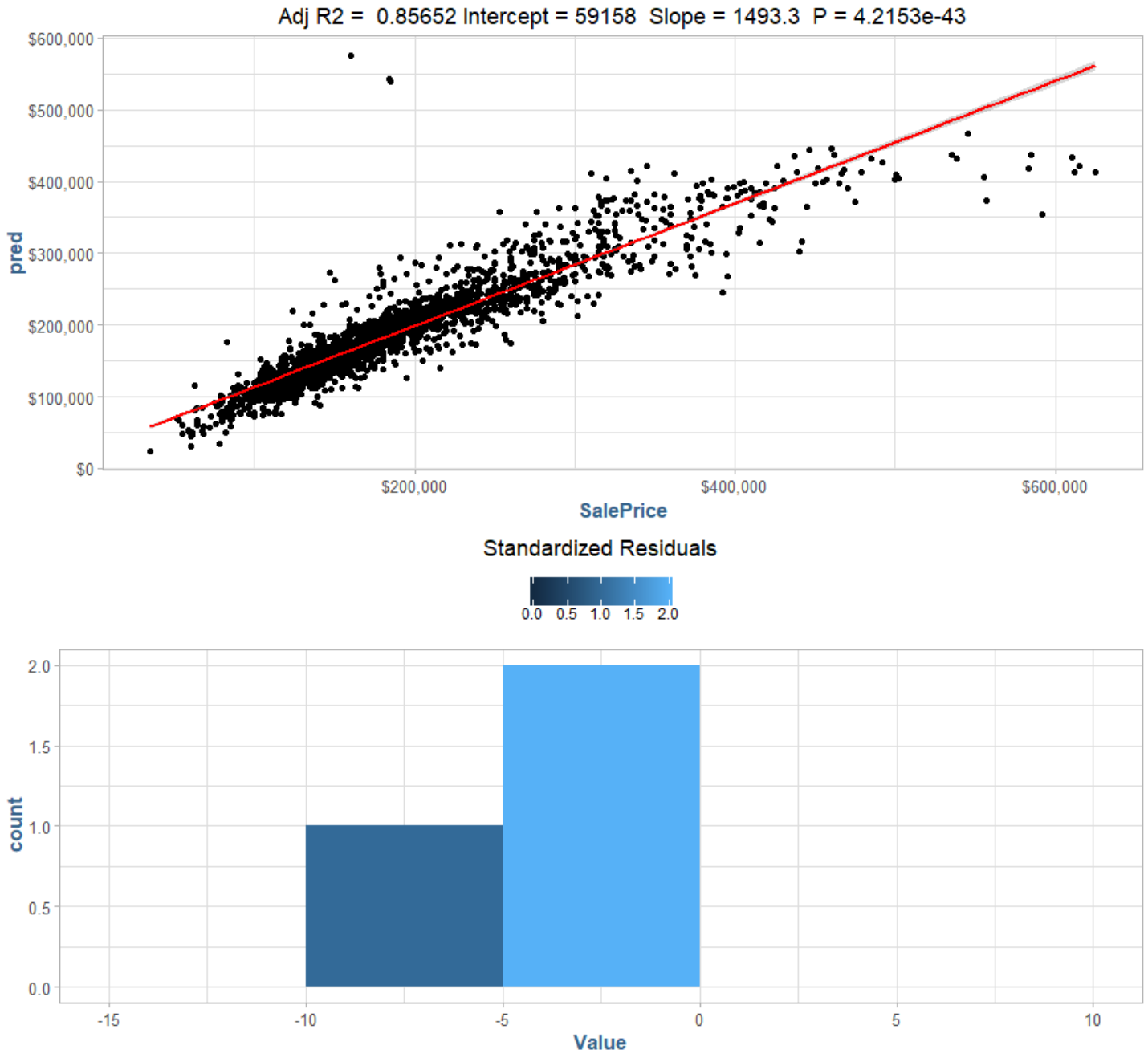
For the in-sample data, we note the following results which is essentially no change from the baseline version to the tuned version:

Model	Grade 1: [0.0.10]	Grade 2: (0.10,0.15]	Grade 3: (0.15,0.25]	Grade 4: (0.25+]
Baseline	64.2119	18.2171	12.6615	4.9096
Tuned	64.0181	16.7959	13.5659	5.6202

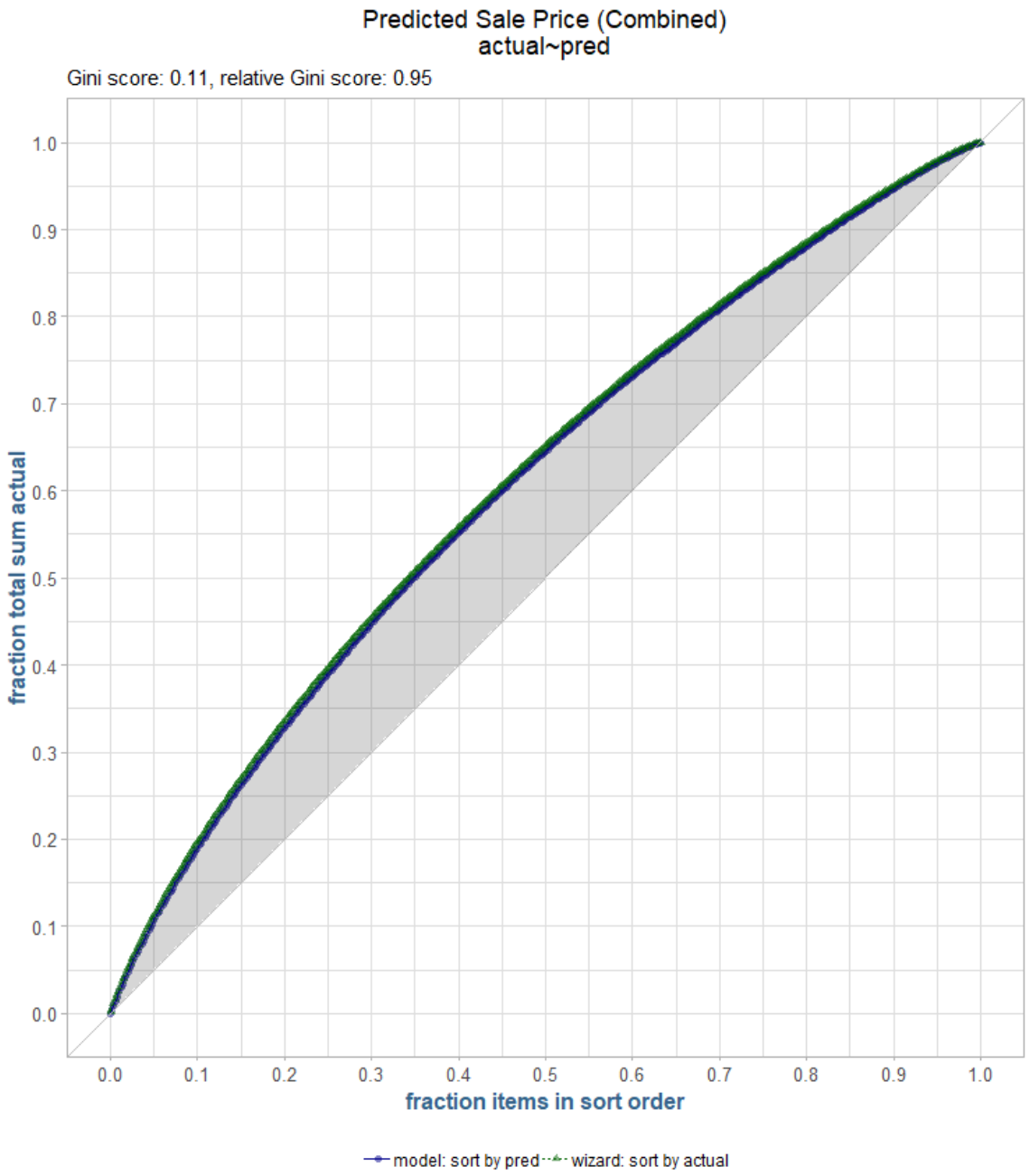
However, in data the model hasn't seen before we notice immediately a 2% increase in the Grade 1 category, which we are trying to get as high as possible.

Model	Grade 1: [0.0.10]	Grade 2: (0.10,0.15]	Grade 3: (0.15,0.25]	Grade 4: (0.25+]
Baseline	60.7903	18.693	14.1337	6.383
Tuned	62.462	16.5653	13.6778	7.2948

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 R^2, 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 a 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 of these 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 R^2 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

Column	RSq	RSE	MeanDiff	Levels	PctPopulated
PoolQC	65.15	$79,538.98	$266,873	4	0.5
Neighborhood	63.70	$48,633.44	$204,189	21	100.0
BsmtQual	54.31	$54,450.27	$202,956	4	98.0
ExterQual	52.31	$55,745.54	$216,654	4	100.0
KitchenQual	49.24	$57,510.01	$185,493	5	100.0
Foundation	28.42	$68,296.76	$169,188	6	100.0
Alley	27.69	$39,208.31	$156,261	2	5.3
GarageFinish	26.73	$68,824.93	$187,457	4	96.6
BsmtFinType1	23.89	$70,277.41	$175,368	6	98.0
Exterior1	23.84	$70,445.41	$183,215	14	100.0
MasVnrType	23.57	$70,570.17	$209,064	5	100.0
Exterior2	23.28	$70,704.92	$189,070	17	100.0
GarageType	21.44	$71,267.45	$173,876	6	96.6
HeatingQC	20.04	$72,180.01	$149,365	5	100.0
BsmtExposure	19.63	$72,216.64	$215,797	5	98.0
SaleType	14.54	$74,623.48	$166,685	10	100.0
SaleCondition	13.58	$75,038.94	$187,758	5	100.0
HouseStyle	11.70	$75,851.57	$177,362	8	100.0
FireplaceQu	11.26	$82,226.60	$219,843	5	55.4
Zoning	9.81	$76,660.14	$164,995	4	100.0
Fence	8.55	$48,923.75	$152,808	4	21.7
MiscFeature	8.28	$41,680.05	$183,451	5	4.0
PavedDrive	7.68	$77,560.84	$147,877	3	100.0
RoofStyle	7.01	$77,839.20	$195,070	6	100.0
LotShape	7.00	$77,846.70	$213,092	4	100.0
Electrical	5.86	$78,322.18	$133,237	5	100.0
CentralAir	5.40	$78,512.19	$149,095	2	100.0
Condition1	5.25	$78,576.43	$183,271	9	100.0
GarageQual	4.29	$78,663.08	$172,748	6	96.6
LandContour	3.99	$79,095.89	$202,014	4	100.0
GarageCond	3.18	$79,118.97	$146,069	6	96.6
Condition2	2.69	$79,629.15	$222,347	6	100.0
ExterCond	2.34	$79,770.61	$156,242	5	100.0
BsmtCond	2.30	$79,622.69	$177,295	5	98.0
LotConfig	1.94	$79,933.51	$201,504	5	100.0
Functional	1.60	$80,072.98	$153,704	6	100.0
BsmtFinType2	1.46	$79,966.55	$197,817	7	98.0
RoofMat	0.81	$80,395.98	$217,837	7	100.0
Heating	0.66	$80,455.76	$122,328	6	100.0
LandSlope	0.28	$80,608.51	$202,053	3	100.0
Street	0.04	$80,706.32	$166,652	2	100.0
Utilities	0.04	$80,707.09	$151,680	2	100.0