Modeling Assignment 6: Finalizing the Model – Variable Selection Procedures and Validation

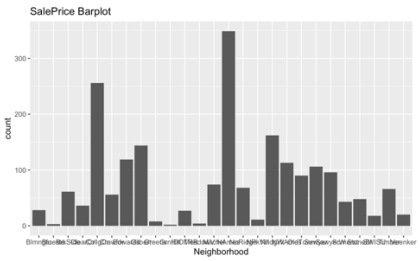
Assignment Overview

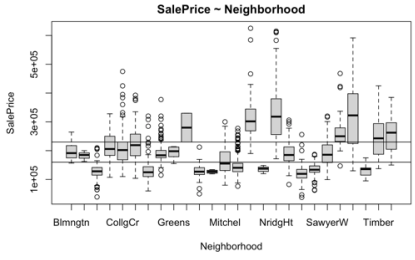
In this assignment we finish our OLS regression model building activities to predict home sale price (SALEPRICE) using the variables in the AMES data set.

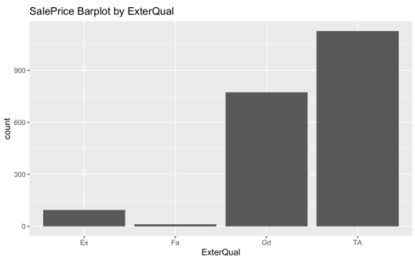
Preparatory Work

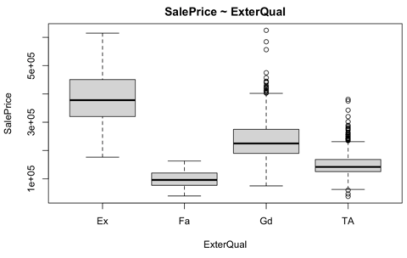
The population of interest remains the same as assignments 1, 3, and 4: *A 1-story building made after 1946, a 2-story building made after 1946, a 1-story PUD made after 1946, or a 1.5 story building. AND in a Residential Low or Medium Density Zoning code. AND less than 4000 sqft of above-ground living area*

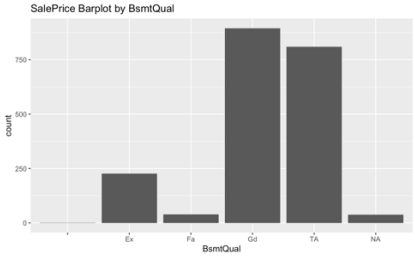
The dummy variables I chose to use in this assignment were Neighborhood, ExterQual, BsmtQual, and KitchenQual. I chose these variables because there is more than one level that contains a significant proportion of the data, as well as visually differing means between levels. The Quality variables are split according to the categories given, but Neighborhood was split into 3 groups: low, medium, and high, according the median SalePrice (see ablines in boxplot).

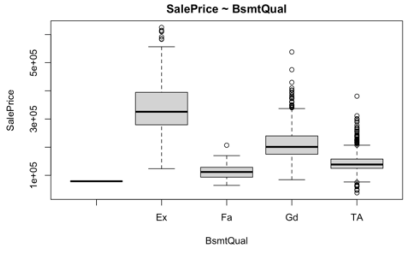


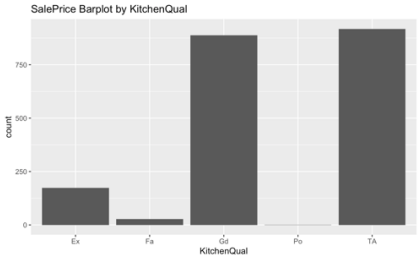


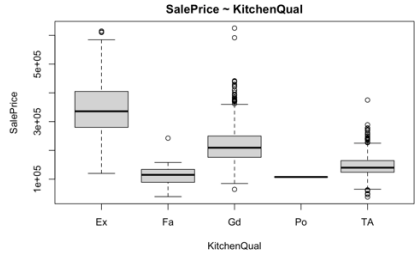












Assignment Tasks

For the tasks in this assignment, the response variable will be: SALEPRICE (Y). The remaining variables will be considered potential explanatory variables (X's).

(1) The Predictive Modeling Framework

With a 70/30 train/test split we now have two data sets: one for in-sample model development and one for out-of-sample model assessment. Our 70/30 training/test split is the most basic form of cross-validation.

Show a table of observation counts for your train/test data partition in your data section.

Dataset	N	%
Train	1416	70.51793%
Test	592	29.48207%
Total	2008	100%

(2) Model Identification by Automated Variable Selection

Create a pool of candidate predictor variables. This pool of candidate predictor variables needs to have at least 15-20 predictor variables, you can have more.

Include a well-designed list or table of your pool of candidate predictor variables in your report.

Variable	Variable Type	Logic
OverallQual	continuous	High correlation with SalePrice (*Assignment 1*)
GrLivArea	continuous	High correlation with SalePrice (*Assignment 1*)
GarageCars	continuous	High correlation with SalePrice (*Assignment 1*)
GarageArea	continuous	High correlation with SalePrice (*Assignment 1*)
FirstFlrSF	continuous	High correlation with SalePrice (*Assignment 1*)
YearBuilt	continuous	High correlation with SalePrice (*Assignment 1*)
FullBath	continuous	High correlation with SalePrice (*Assignment 1*)
TotRmsAbvGrd	continuous	High correlation with SalePrice (*Assignment 1*)
YearRemodel	continuous	High correlation with SalePrice (*Assignment 1*)
Neighborhood_low	discrete	(*Prep Work*) median SalePrice of Neighborhood <= 160000
Neighborhood_med	discrete	(*Prep Work*) median SalePrice of Neighborhood > 160000 and <= 230000
Neighborhood_hi	discrete	(*Prep Work*) median SalePrice of Neighborhood > 230000
ExterQual_2	discrete	(*Prep Work*) ExterQual=="Fa"
ExterQual_3	discrete	(*Prep Work*) ExterQual=="TA"
ExterQual_4	discrete	(*Prep Work*) ExterQual=="Gd"
ExterQual_5	discrete	(*Prep Work*) ExterQual=="Ex"
BsmtQual_2	discrete	(*Prep Work*) BsmtQual=="Fa"
BsmtQual_3	discrete	(*Prep Work*) BsmtQual=="TA"
BsmtQual_4	discrete	(*Prep Work*) BsmtQual=="Gd"
BsmtQual_5	discrete	(*Prep Work*) BsmtQual=="Ex"
KitchenQual_1	discrete	(*Prep Work*) KitchenQual=="Po"
KitchenQual_2	discrete	(*Prep Work*) KitchenQual=="Fa"
KitchenQual_3	discrete	(*Prep Work*) KitchenQual=="TA"
KitchenQual_4	discrete	(*Prep Work*) KitchenQual=="Gd"
KitchenQual_5	discrete	(*Prep Work*) KitchenQual=="Ex"

Did the different variable selection procedures select the same model or different models? **Display the final estimated models and their VIF values for each of these four models in your report.** *Model Comparison:* Now that we have our final models, we need to compare the in-sample fit and predictive accuracy of our models. **For each of these four models compute the adjusted R-Squared, AIC, BIC, mean squared error, and the mean absolute error for each of these models for the training sample.** Each of these metrics represents some concept of 'fit'. In addition to the values **provide the rank for each model in each metric. If a model is #2 in one metric, then is it #2 in all metrics? Should we expect each metric to give us the same ranking of model 'fit'.**

In-Sample Summary Statistics (best ranking | mid ranking | worst ranking)

Model	Vars	Max VIF	R^2 adj	AIC	BIC	MSE	MAE
Forward	SalePrice ~ OverallQual + GrLivArea + BsmtQual_5 + FirstFlrSF + GarageArea + ExterQual_5 + YearBuilt + **Neighborhood_hi** + KitchenQual_5 + YearRemodel + FullBath + Neighborhood_low + TotRmsAbvGrd	GrLivArea 4.907968	0.8722	32462.62	32541.15	837064986	19603.58
Backward	SalePrice ~ OverallQual + YearBuilt + YearRemodel + FirstFlrSF + GrLivArea + FullBath + TotRmsAbvGrd + GarageArea + Neighborhood_low + **Neighborhood_med** + ExterQual_5 + BsmtQual_5 + KitchenQual_5	Neighborhood_low 5.056915	0.8722	32462.62	32541.15	837064986	19603.58
Stepwise	SalePrice ~ OverallQual + GrLivArea + BsmtQual_5 + FirstFlrSF + GarageArea + ExterQual_5 + YearBuilt + **Neighborhood_hi** + KitchenQual_5 + YearRemodel + FullBath + Neighborhood_low + TotRmsAbvGrd	GrLivArea 4.907968	0.8722	32462.62	32541.15	837064986	19603.58
Junk	formula = SalePrice ~ LotArea + SecondFlrSF + HalfBath + OpenPorchSF + WoodDeckSF + Fireplaces	HalfBath 1.831630	0.3631	35419.13	35461.18	4225217216	46462.05

5 is considered a concerning VIF, and 10 is considered an unacceptable VIF. With the highest VIF in all model being 5.05, we will revisit the topic of multicollinearity in coefficient interpretations in the last task of this assignment. The Junk model has the lowest VIF and least number of predictors.
All AIC models have the same variables (with Backward having a different reference group for Neighborhood) which yields the same predictive accuracy metrics across all 3 models. Coefficients for all 3 models can be found in the final section of this assignment.

(3) *Predictive Accuracy*

In predictive modeling, we are interested in how well our model performs (predicts) out-of-sample. That is the point of predictive modeling. For each of the four models **compute the Mean Squared Error (MSE) and the Mean Absolute Error (MAE) for the test sample**. **Which model fits the best based on these criteria? Did the model that fit best in-sample predict the best out-of-sample? Should we have a preference for the MSE or the MAE? What does it mean when a model has better predictive accuracy in-sample then it does out-of-sample?**

Out-of-Sample Summary Statistics

Model	MSE	MAE
Forward	758844538	19436.14
Backward	758844538	19436.14
Stepwise	758844538	19436.14
Junk	4608179966	48435.42

As mentioned in the last section, all 3 AIC models have the same predictors which yields the same MSE and MAE, and the Junk model again performs the most poorly. The in-sample errors are lower because, the model is built to minimize error based on the training dataset, and the out-of-same errors are higher due to using a testing dataset. MAE is better understood for interpretation because the scale is the same as the SalePrice, whereas MSE is better used for penalizing larger residuals.

(4) *Operational Validation*

We have validated these models in the statistical sense, but what about the business sense? Do MSE or MAE easily translate to the development of a business policy? Typically, in applications we need to be able to hit defined cut-off points, i.e. we set a policy that we need to be p% accurate. Let's define a variable called PredictionGrade, and consider the predicted value to be 'Grade 1' if it is within ten percent of the actual value, 'Grade 2' if it is not Grade 1 but within fifteen percent of the actual value, Grade 3 if it is not Grade 2 but within twenty-five percent of the actual value, and 'Grade 4' otherwise.

Produce these prediction grades for the in-sample training data and the out-of-sample test data. Note that we want to show these tables in distribution form, not counts. Distribution form is more informative and easier for your reader (and you!) to understand, hence we have normalized the table object.

Prediction Grade Distribution based on Model and Dataset

Model	Data Sample	Grade 1: [0.0.10]	Grade 2: (0.10,0.15]	Grade 3: (0.15,0.25]	Grade 4: (0.25+]
Forward	In-sample	0.6080508	0.1475989	0.1440678	0.1002825
	Out-of-sample	0.58319039	0.20926244	0.13550600	0.07204117
Backward	In-sample	0.6080508	0.1475989	0.1440678	0.1002825
	Out-of-sample	0.58319039	0.20926244	0.13550600	0.07204117
Stepwise	In-sample	0.6080508	0.1475989	0.1440678	0.1002825
	Out-of-sample	0.58319039	0.20926244	0.13550600	0.07204117
Junk	In-sample	0.2923729	0.1137006	0.1998588	0.3940678
	Out-of-sample	0.2668919	0.1317568	0.2111486	0.3902027

How accurate are the models under this definition of predictive accuracy? How do these results compare to our predictive accuracy results? Did the model ranking remain the same?

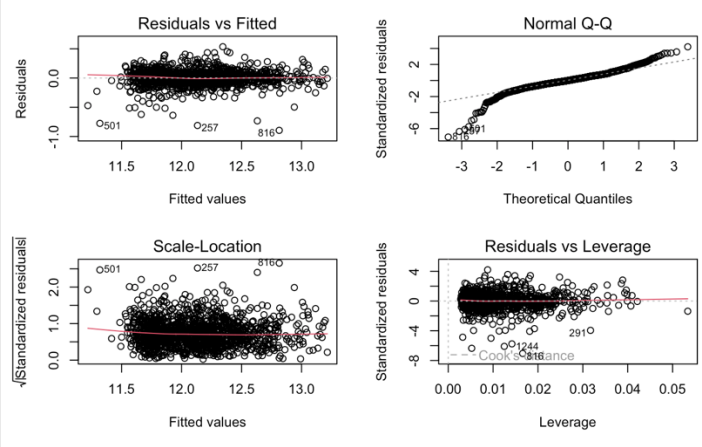
Similar to the previous sections, all AIC models performed the same due to the same predictors. Between in-sample and out-of-sample, the difference was typically less than 5% for each prediction grade. Over 50% of predictions fall under Grade 1 for the AIC models. The Junk model again performed the most poorly having less than 30% of the predictions fall under Grade 1, and the majority of predictions fall in the lowest grade: Grade 4 – over 25% away from the actual value.

Note: The GSEs (Fannie Mae and Freddie Mac) rate an AVM model as 'underwriting quality' if the model is accurate to within ten percent more than fifty percent of the time. Are any of your models 'underwriting quality'?

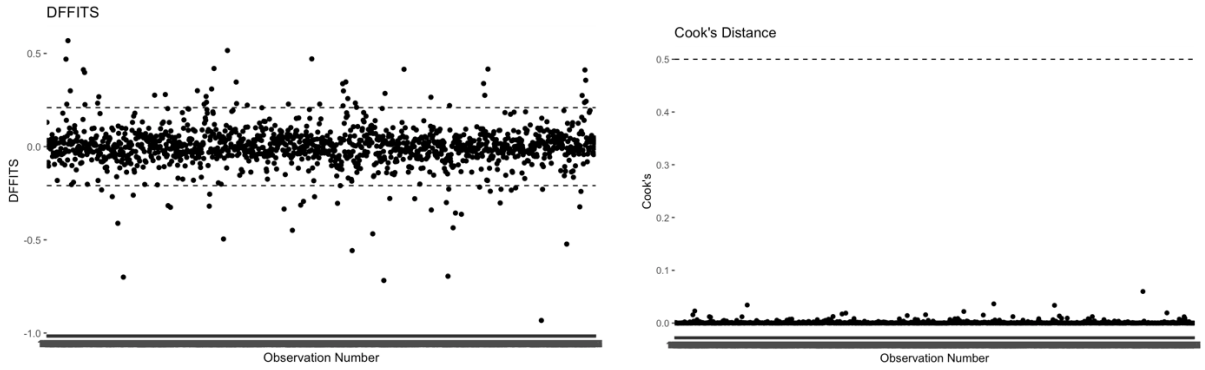
All 3 AIC models are within 10% (Grade 1) more than 50% of the time (58% out-of-sample) which qualifies them as 'underwriting quality.'

6) For which ever model you find to be "Best" after the automated variable selection procedures and all of these comparisons, you will need to re-visit that model and clean it up, as well as conduct residual diagnostics. Frankly, the end of an automated variable selection process is in many ways a starting point. What kinds of things do you want to check for and "clean up"?

We know from past assignments that log(SalePrice) is better for linear regression. We will use the same predictors with this one transformation and check the residual diagnostics.



The Residual vs Fitted plot has a random linear scatter indicating that a linear relationship between SalePrice and predictors has not been violated and neither has independence of errors. The Normal Q-Q plot approximately follows from the center indicating normality of errors. Scale-Location is also approximately linear indicating homoscedasticity.



DFFITS indicates over 100 influential points and Cook's Distance has no influential points. To avoid modeler bias, we will only remove 10-12 of the most influential points and see how this affects the model in terms of fit. Adjusted R2 goes from 0.8722 to 0.8815, which justifies the removal of these 10-12 influential points for a better fitting model.

- Quantitative variables may have been selected that logically have their coefficients reversed from what it theoretically should be. For example, a final model can have a negative coefficient relating size of home in square feet to price. That wouldn't make sense. Why would that variable be included in the model – or there is something else going on, like multicollinearity that needs to be accounted for. That has to be fixed some way. Always remember, an easy solution for multicollinearity is to simply leave an offending variable out of the model.

Check for multicollinearity from model coefficients' signs

```
Coefficients:
                  Estimate
(Intercept)       3.698e+00
OverallQual       6.282e-02
YearBuilt         1.891e-03
YearRemodel       1.842e-03
FirstFlrSF        1.499e-04
GrLivArea         2.874e-04
FullBath         -2.493e-02
TotRmsAbvGrd     -5.777e-03
GarageArea        2.175e-04
Neighborhood_low -9.134e-02
Neighborhood_med -4.011e-02
ExterQual_5       4.940e-02
BsmtQual_5        5.299e-02
KitchenQual_5     5.367e-02
```

With FullBath, TotRmsAbvGrd

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       4.030e+00  4.368e-01   9.225  < 2e-16 ***
OverallQual       6.281e-02  4.106e-03  15.297  < 2e-16 ***
YearBuilt         1.794e-03  1.996e-04   8.987  < 2e-16 ***
YearRemodel       1.753e-03  2.101e-04   8.342  < 2e-16 ***
FirstFlrSF        1.532e-04  1.018e-05  15.052  < 2e-16 ***
GrLivArea         2.585e-04  8.582e-06  30.114  < 2e-16 ***
GarageArea        2.184e-04  2.017e-05  10.828  < 2e-16 ***
Neighborhood_low -8.749e-02  1.329e-02  -6.582 5.94e-11 ***
Neighborhood_med -4.131e-02  9.872e-03  -4.185 2.98e-05 ***
ExterQual_5       4.688e-02  1.867e-02   2.511 0.012128 *
BsmtQual_5        5.356e-02  1.317e-02   4.069 4.92e-05 ***
KitchenQual_5     5.528e-02  1.487e-02   3.717 0.000207 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1297 on 1947 degrees of freedom
  (38 observations deleted due to missingness)
Multiple R-squared:  0.8816,    Adjusted R-squared:  0.8809
F-statistic:  1318 on 11 and 1947 DF,  p-value: < 2.2e-16
```

Without FullBath, TotRmsAbvGrd

From the model estimates we note that FullBath, TotRmsAbvGrd have negative coefficients. By interpretation, this does not make sense because we would assume that properties with more baths or more rooms above ground should have higher SalePrice. We remove both of these predictors from the model which fixes the interpretation of all coefficients. The highest VIF value of the new model is <10.

- Quantitative variables may be included in the model that are statistically significant, but are not actually predictive. This is mostly an issue when the sample size is large. The issue is large sample size translates into high statistical power. If too much power is present, everything can be statistically significant. Remember, statistical significance does not mean important, it means ruling out chance as the explanation. To guard against an overfit model with too many variables, consider looking at R-squared change. Examine the impact of removing each variable from the final model, one at a time. IF R-squared change for any of the retained variables is too small – why include that variable in the model? It is not contributing to the predictive ability of the model. Think about it! Parsimony is important – simpler models are easier to explain and tend to be better in the long run out of sample! Do you really need a max fit model, or a best explanation model? You are welcome to remove variables from the "final model" until all contribute sufficiently well.

Check change in R2 when removing 1 predictors – avoid overfit

The R2 adj after removing the predictors in the task above goes from 0.8815 to 0.8809 which indicates we did not lose significant prediction power and the fit of the model is still high. We will additionally choose to remove ExterQual_5 (correlated with other Qual vars), BsmtQual_5 (correlated with other Qual vars), Neighborhood_med (correlated with other Neighborhood vars), and YearRemodel (correlated with other Year vars). The R2 adj after removing the predictors in the task above goes from 0.8809 to 0.8734 which indicates we did not lose significant prediction power and the fit of the model is still high.

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       5.936e+00  3.680e-01  16.129  < 2e-16 ***
OverallQual       8.252e-02  3.935e-03  20.974  < 2e-16 ***
YearBuilt         2.513e-03  1.882e-04  13.353  < 2e-16 ***
FirstFlrSF        1.499e-04  1.029e-05  14.562  < 2e-16 ***
GrLivArea         2.559e-04  8.849e-06  28.915  < 2e-16 ***
GarageArea        2.338e-04  2.060e-05  11.350  < 2e-16 ***
Neighborhood_low -5.232e-02  9.740e-03  -5.372 8.71e-08 ***
KitchenQual_5     1.075e-01  1.264e-02   8.505  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.136 on 1989 degrees of freedom
Multiple R-squared:  0.8739,    Adjusted R-squared:  0.8734
F-statistic:  1968 on 7 and 1989 DF,  p-value: < 2.2e-16
```

- Variables included in the final model may have coefficients that are essentially zero. These need to be checked to see if those variables are predictive. Use R-squared change to determine if the variable should be retained.

Task above – start with near-0 coefficients

Removing the variable with the coefficient closest to 0 (FirstFlrSF: 0.0001499) changes the R2 adj from 0.8734 to 0.8600 which is incrementally worse than removing the predictors in the tasks above. We will choose to leave this predictor in the final model. The reason these coefficients are so close to zero is because 1.) SalePrice is log transformed so the prediction value is significantly smaller than SalePrice scale, and 2.) continuous variables are in sqft scale so the coefficient is multiplied by a large number.

- A dummy coded variable may have been selected using the automated procedure. If this has happened, you will want to include all but one of the dummy coded variables for the associated categorical variable in the model. It does not matter whether the dummy coded variables are statistically significant or not. The purpose is for interpretation of coefficients and the inclusion of the entire categorical variable in the predictive model. Think of it as variables are used – all or nothing, not just bits and pieces.

Choose whether to include all of a dummy variable levels, or only some

In the sections above, I chose to only include the dummy variables: Neighborhood_low and KitchenQual_5. We tested the inclusions and exclusions of the other dummy variable levels from the model which did not significantly affect R2 adj and the fit of the model, so we removed them to avoid overfitting. This also indicates that there may not be a significant difference between Neighborhood_med and Neighborhood_hi as well as between KitchenQual_1-KitchenQual_4. Those predictors are lumped together as the baseline comparison.

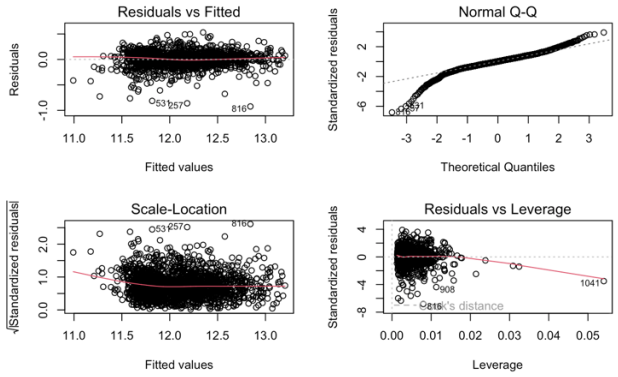
- If you retain one or more categorical variables in your final model, you have an ANCOVA model. This means you are modeling Y with parallel planes. You should then be concerned about unequal slopes for these planes. Of concern is the interaction between the categorical variable and any of the quantitative variables. This is a challenging issue if you have a large number of quantitative explanatory variables in your final model. If interaction between the categorical variable and any of the quantitative variables is a logical possibility, you'll need to test for unequal slopes.

Choose to test ANCOVA unequal slopes

The only interaction I would like to test is between KitchenQual and OverallQual as these are logically related. Adding the interaction term to the model changes the R2 adj from 0.8734 to 0.8734 which does not improve the model. The T-tests of KitchenQual and the interaction term become statistically insignificant. We will choose not to include the interaction term in the final model.

Once you've cleaned up your model and have come to a TRUE FINAL model, you will want to conduct the **typical goodness of fit and model diagnostic analysis**. Hopefully, all will go well and you won't have any issues. If this is the case:

Residual diagnostics, assumptions, ~~transformations~~ goodness of fit



The Residual vs Fitted plot has a random linear scatter indicating that a linear relationship between SalePrice and predictors has not been violated and neither has independence of errors. The Normal Q-Q plot approximately follows from the center indicating normality of errors. Scale-Location is also approximately linear indicating homoscedasticity. We will accept this as our final model. Coefficients and summary values in the screenshot above.

final_model <- lm(formula = log(SalePrice) ~ OverallQual + YearBuilt + FirstFlrSF + GrLivArea + GarageArea + Neighborhood_low + KitchenQual_5, data=subset_no_outliers)

7) ***Conclusion***

What are the challenges presented by the data?

This data proves to be challenging in many different ways.

From the beginning of the analysis, the data cleaning is a massive undertaking. We must take caution when filtering out data with waterfall statements whether that's due to defining our population of interest or whether there are outliers that do not assimilate to the patterns of the majority of the data. Whatever it is, any data cleaning decisions we make in the beginning are crucial to the types of analyses we can perform and the way we interpret our outcomes. I chose to make small tweaks throughout the assignments on the housing dataset and appropriately changed my interpretations based on feedback.

Data transformations can be tricky as well. The assumptions of linear regression are difficult to satisfy, and the transformations of variables will not always satisfy the assumptions. It can take many tries of trial and error to figure out the appropriate transformation of a single variable. In past assignments we learned that a log transformation of SalePrice helped with the summary plots and satisfying linear regression assumptions. We also learned in previous assignments the need for transforming discrete data into dummy coded variables for analysis as well as creating interaction terms between discrete and continuous variables in combined models.

Lastly, the data became challenging when it came to avoiding multicollinearity. Sometimes it's easy to avoid it by just using logic like OverallCond and OverallQual. But in other cases, we need to check in many different ways: correlations, VIF, coefficient signs. All of these tests are used throughout the analysis. For example the VIF values may be low like in the model for task #6, but the interpretations of the coefficients just did not make sense because of the sign, which indicated multicollinearity with a similar predictor. Model selection based on collinear interpretation was truly a challenge.

What are your recommendations for improving predictive accuracy?

I believe that like above there are many different steps along the way that help improve predictive accuracy:Model selection is important. Having a large pool of predictors helps with improving accuracy as seen in task #2 where we allowed the automated model to add and remove predictors from the large pool we provided.

Like above, variable transformation, including interaction terms, and outlier removal may also improve model accuracy. All of these things (variable selection, variable transformation, interaction terms, and outlier removal) need to be tweaked over many iterations in a methodical process in order to yield high predictive accuracy.

What do you think of the notion of parsimony: simpler models might be preferable over complicated models? Do we really need a max fit model or is a simpler but more interpretable model better?

Take caution when balancing the level of predictive accuracy with overfitting. Modeler bias and overfitting a training model may yield high predictive accuracy for the training data, but it may perform poorly on the testing dataset due to overfitting. Or even if the data does do fine on both the in-sample and out-of sample data, we must take caution on generalizing any outcomes based on the ways we manipulated the data. For example, removing a significant amount of outliers may yield highly accurate results with the subsetted data, but the interpretations may not be true for the population of interest that we want to generalize to in the future.

Overall, it is best to balance predictive power with the restraint of avoiding overfitting models. In this assignment, I chose to drop a lot of variables because they did not contribute to the goodness of fit metric R2 adj relatively as much as other predictors. A simple model is also easier to interpret to see what effects each predictor has on SalePrice.