Assignment # 7

May 2021

1)

After reviewing the variables in the Ames Dataset, I will explore the additional variables of overall quality, overall condition, and land slope.

Overall Quality:

This variable was viewed in previous assignments and tested high in correlation to sales price, it is also a seeming natural connection to sales price and will be tested further.

Overall Quality	Min	1st Quartile	Median	Mean	3rd Quartile	Max
1	39,300	39,300	39,300	39,300	39,300	39,300
2	35,000	45,000	59,000	56,200	60,000	82,000
3	58,500	64,500	78,500	84,104	99,800	126,175
4	40,000	89,000	109,900	111,195	128,000	228,500
5	62,383	122,000	134,250	135,806	147,975	284,700
6	76,000	143,000	164,950	166,235	185,188	290,000
7	105,000	180,000	200,000	205,881	229,000	328,000
8	122,000	236,625	263,275	261,200	289,250	334,000
9	239,000	295,500	315,500	307,588	324,500	333,168325
10	325,000	325,000	325,000	325,000	325,000	325,000

Avg. mean = 31.744.44

Overall Condition:

Overall condition sems very similar to overall quality and would be a great variable to test against sales price. Multicollinearity of overall condition and overall quality is something to keep in mind though when viewing results.

Average mean = 28,508

Land Slope:

Over the past assignments, a surprising variable that showed some affect on sales price was land scope. Here, we will explore it further and see if scale of the slope affects the sales price.

Land Slope	Min	1st Quartile	Median	Mean	3rd Quartile	Max
Gentle	35,000	130,000	158,500	169,407	200,000	333,168
Moderate	39,300	131,938	176,250	181,984	235,125	334,000
Severe	130,000	200,000	268,500	238,000	278,500	302,000

Avg. mean = 34,297

The three average means are all very close together in range with the minimum being overall condition at 28k and the max being land slope at 34k. Land Slope proved to be the highest mean and thus the variable to further study when comparing against sales price. The difference in means between each

level supports the logic of dummy coding for each categorical variable as the changes are not equal and a constant value will be left out of the model.

2)

Train: 1213; test: 515; Total after sampling population: 1,728

3)

Deck or Porch Sq Footage – C	Lot Frontage – C	Lot Area - C
Month Sold - D	Total Basement Sq Footage - C	Total Floor Sq Footage - C
Ground Living Area - C	Garage Area - C	Pool Area - C
Bedrooms - D	House Age - D	Total Baths - D
Garage Cars - D	Quality Index - O	Total Square Foot Calc - C

I chose these 15 predictor variables based on our past assignments and correlation. The three model procedures I selected are forward, backward, and stepwise.

Forward:

Predicted Sales Price = Total SqFt + Garage Cars + Quality index - House Age + Total Floor Sqft + Total basement sqft + lot Area + Deck or Porch Sqft + Garage Area + Lot Frontage

Variable	VIF
Garage Cars	4.955833
Garage Area	4.506977
Total Sqft Calc	3.381428
Total Floor Sqft	2.847250
Total Basement Sqft	1.641661
House Age	1.550249
Quality Index	1.216705
Deck or Porch Sqft	1.198968
Lot Area	1.179369
Lot Frontage	1.014972

The vif values are very low and thus none require to be removed because of these low values.

Backward:

Predicted Sales Price = Total SqFt + Garage Cars + Quality index - House Age + Total Floor Sqft + Total basement sqft + lot Area + Deck or Porch Sqft + Garage Area + Lot Frontage

Variable	VIF
Garage Cars	4.955833
Garage Area	4.506977
Total Sqft Calc	3.381428
Total Floor Sqft	2.847250
Total Basement Sqft	1.641661
House Age	1.550249
Quality Index	1.216705
Deck or Porch Sqft	1.198968
Lot Area	1.179369
Lot Frontage	1.014972

The vif values are very low and thus none require to be removed because of these low values.

Stepwise:

Predicted Sales Price = Total SqFt + Garage Cars + Quality index - House Age + Total Floor Sqft + Total basement sqft + lot Area + Deck or Porch Sqft + Garage Area + Lot Frontage

Variable	VIF
Garage Cars	4.955833
Garage Area	4.506977
Total Sqft Calc	3.381428
Total Floor Sqft	2.847250
Total Basement Sqft	1.641661
House Age	1.550249
Quality Index	1.216705
Deck or Porch Sqft	1.198968
Lot Area	1.179369
Lot Frontage	1.014972

The vif values are very low and thus none require to be removed because of these low values.

Adjusted r-squared:

Model	Adjusted R-Squared	Rank
Forward	0.8926	**1**
Backward	0.8926	**1**
Stepwise	0.8926	**1**
Junk	0.819	**4**

AIC:

Model	AIC	Rank
Forward	27251.61	**1**
Backward	27251.61	**1**
Stepwise	27251.61	**1**
Junk	27875.36	**4**

BIC:

Model	BIC	Rank
Forward	27312.82	**1**
Backward	27312.82	**1**
Stepwise	27312.82	**1**
Junk	27911.07	**4**

MSE:

Model	MSE	Rank
Forward	328033973	**1**
Backward	328033973	**1**
Stepwise	328033973	**1**
Junk	553124960	**4**

MAE:

Model	MAE	Rank
Forward	13646.06	**1**
Backward	13646.06	**1**
Stepwise	13646.06	**1**
Junk	17753.52	**4**

All three models produced the same or similar results, and all had the same model fit ranking of #1.

4)

MSE:

Model	MSE	Rank
Forward	323541773	**1**
Backward	323541773	**1**
Stepwise	328033973	**3**

MAE:

Model	MAE	Rank
Stepwise	13646.06	**1**
Forward	13753.05	**2**
Backward	13753.05	**2**

The stepwise model has different MSE and MAE scores from the test sample but I hade worse mse scores than the forward and backward models, but did have better mae values. To decide between the two for a better-out-of-sample fit, it is important to note that better predictive accuracy in-sample would likely mean our model is more accurate with the current data, and if it has a higher out-of-sample accuracy. The forecast values will have a higher accuracy/ This is the reason we are creating the different models and suing comparative metrics to find the best fitting model.

5)

G1 = within 10% of actual value

G2 = within 15% of actual value and outside scope of G1

G3 = within 25% actual value and outside scope of G2

G4 = anything outside the scope of G1-3

In Sample:

Model	**Grade 1**	**Grade 2**	**Grade 3**	**Grade 4**
Forward	69.25%	15.91%	11.05%	3.79%
Backward	69.25%	15.91%	11.05%	3.79%
Stepwise	69.25%	15.91%	11.05%	3.79%

Out-of-Sample:

Model	**Grade 1**	**Grade 2**	**Grade 3**	**Grade 4**
Forward	67.77%	19.03%	10.29%	2.91%
Backward	67.77%	19.03%	10.29%	2.91%
Stepwise	67.77%	19.03%	10.29%	2.91%

The in sample and out of sample results produced the same values for all three of the forward, backward, and stepwise models with the out of sample results closely resembling the training data results. Lastly, I noted that if the model is accurate within Grade 1 more than 50% of the time, each of our models are of good quality.

6)

Step 4 showed that the best models are the forward and backward models with the highest rankings for fit and their mse values, so I will use the forward model moving forward to re-visit, clean up, and conduct residual diagnostics. We must further understand our models outputs, standardized residuals, and distributions.

Multicollinearity:

Variable	Coefficient	Relationship
Garage Cars	6623.4193	+
Garage Area	11.4371	+
Total Sqft Calc	13.9408	+
Total Floor Sqft	13.9408	+
Total Basement Sqft	24.7731	+
House Age	-537.5446	-
Quality Index	1458.3783	+
Deck or Porch Sqft	14.1838	+
Lot Area	0.8957	+
Lot Frontage	26.9667	+

House age appears to be the only coefficient with a negative relationship to our target variable of sales price which is in line with how the housing market works.

Vif:

Variable	VIF
Garage Cars	4.955833
Garage Area	4.506977
Total Sqft Calc	3.381428
Total Floor Sqft	2.847250
Total Basement Sqft	1.641661
House Age	1.550249
Quality Index	1.216705
Deck or Porch Sqft	1.198968
Lot Area	1.179369
Lot Frontage	1.014972

The vif values don't indicate multicollinearity high enough to leave any of the variables out of the final model.

R-Squared values:

Due to large sample size we are going to remove variables one by one to analyze the changes in r-squared – keeping those that impact the predictive ability.

Variable Removal	R-Squared	Change in R-Squared
Full Model	0.893	
Lot Frontage	0.8925	0.0005
Garage Area	0.8919	0.0006
Deck or Porch SF	0.8905	0.0014
Garage Cars	0.8819	0.0086
Lot Area	0.8759	0.006
Total SF Calc	0.8664	0.0095
Total Basement SF	0.8196	0.0468
Total Floor SF	0.5685	0.2511
Quality Index	0.3501	0.2184
House Age	0	0.3501

Lot frontage, garage area, deck or porch sf, garage cars, lot area all have values less than 1 percentage, we will remove these variables as the don't contribute to the impact on predictive ability. Keeping these variables has a total r-squared value of .8759 with only .0171 less than the full model after removing the five variables.

Coefficients with change in r-squared:

Variable	Coefficient	Change in R-Squared
Total SF Calc	15.202	0.0095
Total Basement SF	29.949	0.0468
Total Floor SF	50.119	0.2511
Quality Index	1556.851	0.2184
House Age	-630.491	0.3501

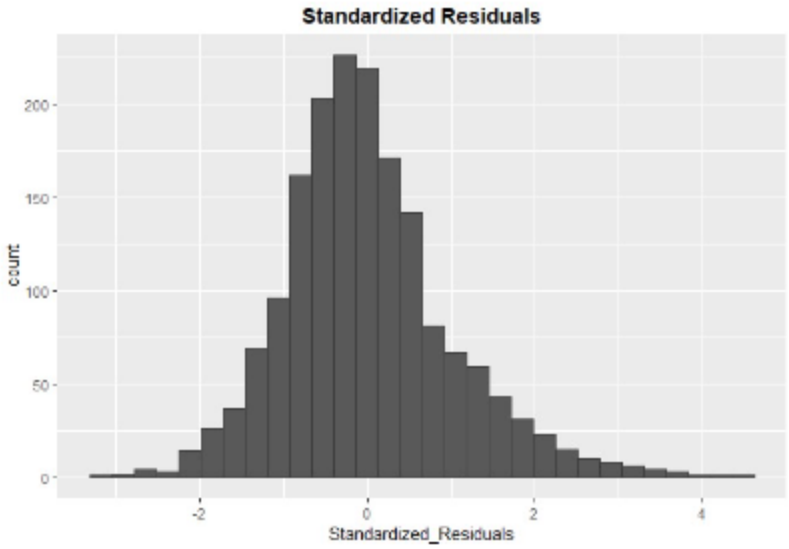
All of the variables had high r-squared change values and thus should be kept as they impact the predicted value.

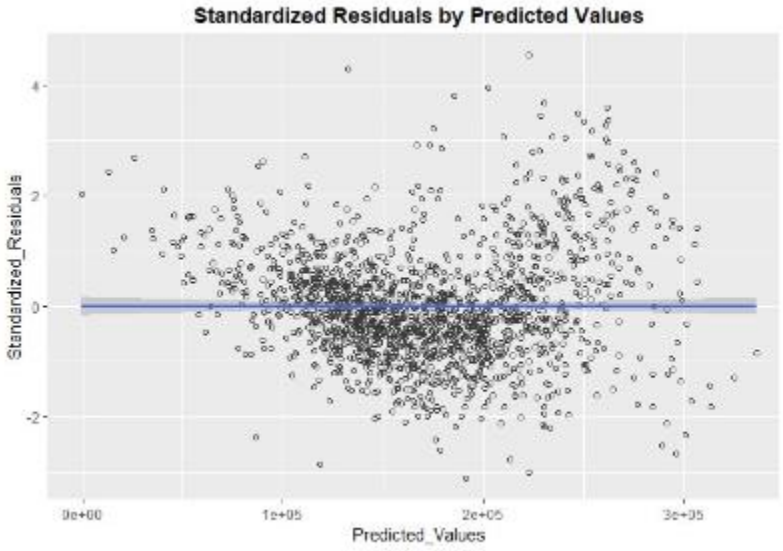
Dummy coded variables:

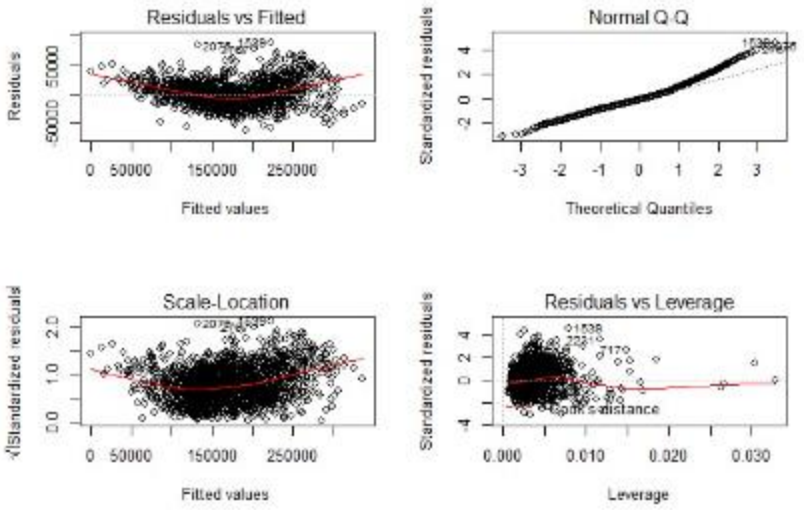
I did not include any dummy coded variables as they were not needed in the final model, however, if there had been they would have been included for the categorical variables.

Unequal Slopes:

Quality index is the only categorical model we have chosen for the final model and it is a combination of overall quality and condition of home. The interaction between this variable and the four others do not appear to be a logical possibility, so there is no need to test.


Standardized Residuals
count
200
150
100
50
0
-2
0
2
4
Standardized_Residuals


Standardized Residuals by Predicted Values
Standardized_Residuals
4
2
0
-2
0e+00
1e+05
2e+05
3e+05
Predicted_Values



There are quite a few outliers in the theoretical quantiles and standardized residuals, but none of our plots show any indication that the assumptions of our model are false. The standardized residuals are normally distributed and there appears to be an equal distribution above and below the trend line in predicted values compared to actual values.

7)

I liked that we went back to the Ames data set for this assignment and have been using it to build on most of our assignments. It has been good to be able to delve deeper into the dataset and build upon our previous models and underlying assumptions about the data. In this assignment, I was concerned by the amount of outliers that are present and wonder if more fitting would be required to get a more accurate model. This would likely need more training data, and since we only have a snapshot int time for Ames, would probably require more years worth of data to be able to view time lapses or changes over time. Independent factors like economy and gdp would also have to be taken into consideration for the outcomes. The max fitting model and the simple model seem both to be good models but would need to be applied in the right situations. The max fitting model seems to be more powerful but the simpler model more intuitive, it would also depend on who the target audience is.