Noé Flores
Automated Variables Selection and Predictive Modeling

Introduction:
The target for this analysis is to build a model with potential predictive capability towards the Sale Price of homes in our housing data set. Our previous models have been developed through Simple Linear Regression (SLR), and more slightly more complex Multiple Linear Regression Models (MLR) in our analysis. Our pursuit here entails moving forward with model development through more advanced automatic selection processes and adding some categorical variables into the model as well.

Results and Analysis:
We begin by examining our dataset to find a categorical variable with the potential to help us develop our predictive model. In this analysis, we will use the variables we feel are visual to buyers and indicate a level of quality of the homes. One variable we identify as a possible candidate is Exterior Quality, which is used to evaluate the quality of the material on the exterior of the house. The basis for choosing this particular variable is once again, exploratory and speculative in nature. This particular variable is ordinal and can take on one of 4 specific quality level distinctions listed below.

Figure 1: Exterior Quality Variable.

Exter Qual: Variable	Condition
Ex	Excellent
Gd	Good
TA	Typical/Average
Fa	Fair

In order to incorporate this variable into our model, we will need to numerically code each of the conditions in the variables listed above. This will allow us to run a regression analysis against those exterior conditions. The description for those numerical assignments is listed in Figure.2.

Figure 2: SalePrice GrlivArea and Log of SalePrice and GrLivArea.

Exter Qual: Variable	Condition	Numerical code
Ex	Excellent	**1**
Gd	Good	**2**
Ta	Typical/Average	**3**
Fa	Fair	**4**

Now that the variable's specific quality distinctions have been recorded with numerical values, we can run some general statistics to find measures of central tendency such as the mean SalePrice for each exterior quality level. We can see the spread of these results in Figure.3, and they appear to be in line with assumptions related to the overall quality distinction of the homes.

Figure.3 Central tendency results for Saleprice with respect to exterior quality.

Exterior qual=1	Excellent	Variable:Saleprice		
N	Mean	Std Dev	Minimum	Maximum
107	$377,918.62	$106,987.71	$160,000.00	$755,000.00
Exterior qual=2	**Good**	**Variable:Saleprice**		
N	Mean	Std Dev	Minimum	Maximum
989	$230,756.38	$70,411.14	$52,000.00	$745,000.00
Exterior qual=3	**Typical/Average**	**Variable:Saleprice**		
N	Mean	Std Dev	Minimum	Maximum
1799	$143,373.97	$41,503.85	$12,789.00	$415,000.00
Exterior qual=4	**Fair**	**Variable:Saleprice**		
N	Mean	Std Dev	Minimum	Maximum
35	$89,923.74	$38,013.50	$13,100.00	$200,000.00

We can now proceed to fit a simple linear regression model of Exterior Quality as our predictor variable for Sale Price.

$$\text{SalePrice} = \beta_0 + \beta_1 \text{ (Extqual)} + \varepsilon$$

The results of running the regression model on Sale price produced the following parameter results:

Figure.4 Regression model #1SalePrice with Exterior Quality

Variable	Degrees of Freedom	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	430,742	4855.76110	88.71	<.0001
extqual	1	-96,083	1821.85306	-52.74	<.0001

The complete fitted equation for our model is the following:

$$\text{SalePrice} = 430{,}742 + (-96{,}083 * \text{extqual})$$

The predicted value is SalePrice, the model coefficients in the equation indicate to us that if Extqual was 1, our lowest possible quality indicator, the final SalePrice of the house would be $334,659. The mean results we got from the new numerically coded Exterior Quality variable provide us with some clues to asses this regression models fit. With extqual set to one, our result of $334,659 fell short of the mean value of $377,918, with respect to sale price in our dataset. If extqual was 2, our equation result would be $238,576. These results differs from the means in our central measure analysis. This doesn't make sense, and there is some reason behind it. The parameter estimates tell us that our intercept is statistically significant, given that it has a t-value of 88.71, but our coefficient is not significant with a t-value of -52.74. The R-square was also slightly below our desired level of 50% and the examination of the residuals didn't present a lot of scattering across the vertical lines associated with our categorical variable. The real reason behind lack of fit is that extqual is not a continuous variable that can be simply plugged into our equation.

The use of numerical coding was insufficient with respect to our desired goal of incorporating our categorical variable of exterior quality into a regression analysis as a predictor of Sale Price. Dummy coding variables provide another method for utilizing categorical information in regression analysis. We will need to create three dummy coded variables for Extqual holding one over as the basis for interpretation. Generally speaking, with ***k*** categories there will be ***k-1*** coded variables. Each dummy coded variable uses one degree of freedom, so ***k*** groups has ***k-1*** degrees of freedom. Figure.5 provides a visual for the process we described.

Figure.5 Exterior Quality Dummy Coding breakdown

ExtQual Variable	Excellent	Good	Typical/Average
Excellent (ex)	1	0	0
Good (gd)	0	1	0
Typical/Average (ta)	0	0	1
Fair (fa)	0	0	0

The actual dummy coding procedure yields the following results listed in the table below.

Figure.6 Exterior Quality Dummy Variable

ExterQual	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Ex	107	3.65	107	3.65
Fa	35	1.19	142	4.85
Gd	989	33.75	1131	38.6
TA	1799	61.4	2930	100
extqual_ex	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	2823	96.35	2823	96.35
1	107	3.65	2930	100
extqual_gd	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	1941	66.25	1941	66.25
1	989	33.75	2930	100
extqual_ta	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	1131	38.6	1131	38.6
1	1799	61.4	2930	100

We can now proceed to fit a simple linear regression model of Exterior Quality as our predictor variable for Sale Price.

$$\mathbf{Saleprice} = \beta_0 + \beta_1\,(\mathbf{extqual\_ex}) + \beta_2\,(\mathbf{extqual\_gd}) + \beta_3\,(\mathbf{extqual\_ta}) + \varepsilon$$

The results of running the regression model on Saleprice produced the following parameter results:

Figure.7 Regression model #1SalePrice with Exterior Quality

Variable	Degrees of Freedom	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	89,924	9507.89095	9.46	<.0001
extqual_ex	1	287,995	10953	26.29	<.0001
extqual_gd	1	140,833	9674.66698	14.56	<.0001
extqual_ta	1	53,450	9599.93464	5.57	<.0001

The complete fitted equation for our model is the following:

SalePrice = 89,924 + 287,995 * extqual_ex + 140,833 * extqual_gd + 53,450 * extqual_ta

A preliminary overview lets us know that we have statistically significant betas, all with positive t-values and p-values well below our benchmark of .05. We can now fit the model in terms of the individual beta coefficients.

Exterior Quality level Excellent:

SalePrice = $89,924 + $287,995 = $377,919

The model above is interpreting the sale price with respect to excellent exterior material. If we go back to our table with the mean breakdown for each category and look at the mean value for an Excellent (ex) rating, we see that it is in fact equal to $377,919.

Exterior Quality level Good:

SalePrice = $89,924 + $140,833 = $230,757

The model above is interpreting the sale price with respect to good exterior material. The mean value in our table for a Good (gd) rating is in fact equal to $230,757.

Exterior Quality level Typical/Average:

SalePrice = $89,924 + $53,450 = $143,374

The model above is interpreting the sale price with respect to typical/average exterior material. The mean value in our table for a Typical/Average (ta) rating is equal to $230,757.

Exterior Quality level Fair:

SalePrice = $89,924

The model above is interpreting the sale price with respect to fair exterior material. The mean value in our table for a Fair (fa) rating is equal to $89,924.

The new model utilizing dummy coded variables produced vastly improved results that were consistent with the previous analysis of our data. The predicted model goes through the mean of Y in each category of Exterior Quality.

Hypothesis testing allows us to carry out inferences about our parameters using data from our sample. We can, therefore, formulate the following null and alternative hypothesis that correspond to our dummy coded variable.

Null hypothesis for $\boldsymbol{\beta_1}$ extqual_ex:

(H_0): $\boldsymbol{\beta_1 = 0}$.

Alternative hypothesis for $\boldsymbol{\beta_1}$ extqual_ex:

(H_a): $\boldsymbol{\beta_1 \neq 0}$..

Since $\boldsymbol{\beta_1}$ yielded statistically significant results at a value not equal to zero (287,995), we can reject the null hypothesis and accept the alternative hypothesis for $\boldsymbol{\beta_1}$.

Null hypothesis for $\boldsymbol{\beta_2}$ extqual_gd:

(H_0): $\boldsymbol{\beta_2 = 0}$.

Alternative hypothesis for $\boldsymbol{\beta_2}$ extqual_gd:

(H_a): $\boldsymbol{\beta_2 \neq 0}$..

Since $\boldsymbol{\beta_2}$ yielded statistically significant results at a value not equal to zero (140,833), we can reject the null hypothesis and accept the alternative for $\boldsymbol{\beta_2}$.

Null hypothesis for $\boldsymbol{\beta_3}$ extqual_ta:

(H_0): $\boldsymbol{\beta_3 = 0}$.

Alternative hypothesis for $\boldsymbol{\beta_3}$ extqual_ta:

(H_a): $\boldsymbol{\beta_3 \neq 0}$..

Since $\boldsymbol{\beta_3}$ yielded statistically significant results at a value not equal to zero (53,450), we can reject the null hypothesis and accept the alternative for $\boldsymbol{\beta_3}$.

To further fill out our predictive model, we will dummy code an additional variable from our dataset, Exterior Condition. Exterior Condition (ExterCond) evaluates the present condition of the material on the exterior. Once again, the basis for choosing this particular variable is exploratory and speculative in nature. This particular variable is also ordinal and can take on one of 5 specific quality level distinctions listed below

Figure 8: Exterior Condition Variable.

Exter Cond: Variable	Condition
Ex	Excellent
Gd	Good
TA	Average/Typical
Fa	Fair
Po	Poor

The dummy coding procedure yields the following results listed in the table below.

Figure.9 Exterior Condition dummy Variables

ExterCond	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Ex	12	0.41	12	0.41
Fa	67	2.29	79	2.7
Gd	299	10.2	378	12.9
Po	3	0.1	381	13
TA	2549	87	2930	100
extcond_ex	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
0	2918	99.59	2918	99.59
1	12	0.41	2930	100
extcond_gd	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
0	2631	89.8	2631	89.8
1	299	10.2	2930	100
extcond_ta	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
0	381	13	381	13
1	2549	87	2930	100
extcond_fa	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
0	2863	97.71	2863	97.71
1	67	2.29	2930	100

Up until this point, most of the procedures we have used to build our analysis start with a fixed set of variables that we incorporate into a regression model. What is clear from the information present in our housing dataset is that there are a large number of variables with potential as predictors for sale price. In order to build a more inclusive and robust model, we will incorporate the use of various Automated variable selection procedures. The variables we will include for automatic selection will consist of the following continuous variables and the recently dummy coded categorical variables.

Figure.10 Variables Available for Automated Selection

Variables in Creation Order		
	Variable	**Type**
1	LotFrontage	Num
2	LotArea	Num
3	MasVnrArea	Num
4	BsmtFinSF1	Num
5	BsmtFinSF2	Num
6	BsmtUnfSF	Num
7	TotalBsmtSF	Num
8	FirstFlrSF	Num
9	SecondFlrSF	Num
10	LowQualFinSF	Num
11	GrLivArea	Num
12	GarageArea	Num
13	WoodDeckSF	Num
14	OpenPorchSF	Num
15	EnclosedPorch	Num
16	ThreeSsnPorch	Num
17	ScreenPorch	Num
18	PoolArea	Num
19	MiscVal	Num
20	extqual_ex	Num
21	extqual_gd	Num
22	extqual_ta	Num
23	extcond_ex	Num
24	extcond_gd	Num
25	extcond_ta	Num
26	extcond_fa	Num

Automated Selection Process - Forward Inclusion:
This process starts with a predictor with the highest simple correlation, and on each successive step, adds a new variable which will improve the model. Below is the summary table from that selection process.

Figure.11 Forward Inclusion Automated Selection

Summary of Forward Selection							
Step	Variable Entered	Number of Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	GrLivArea	1	0.4992	0.4992	3747.83	2411.66	<.0001
2	TotalBsmtSF	2	0.127	0.6263	2185.96	821.89	<.0001
3	extqual_ta	3	0.0664	0.6927	1370.67	522.08	<.0001
4	extqual_ex	4	0.0388	0.7315	894.559	349.45	<.0001
5	GarageArea	5	0.0264	0.7579	572.027	262.91	<.0001
6	BsmtUnfSF	6	0.0126	0.7705	418.389	132.98	<.0001
7	extqual_gd	7	0.0091	0.7796	308.43	99.56	<.0001
8	MiscVal	8	0.0083	0.7878	208.841	93.82	<.0001
9	MasVnrArea	9	0.0053	0.7932	145.157	62.2	<.0001
10	WoodDeckSF	10	0.003	0.7962	109.845	35.84	<.0001
11	ScreenPorch	11	0.0026	0.7988	80.2945	30.68	<.0001
12	LotArea	12	0.0019	0.8007	58.9881	22.87	<.0001
13	extcond_fa	13	0.0015	0.8022	42.3738	18.4	<.0001
14	PoolArea	14	0.001	0.8032	31.8871	12.4	0.0004
15	LowQualFinSF	15	0.0008	0.804	24.2104	9.64	0.0019
16	BsmtFinSF1	16	0.0007	0.8046	17.9745	8.23	0.0042
17	EnclosedPorch	17	0.0003	0.8049	16.8404	3.14	0.0767

Automated Selection Process - Backward Elimination:
This process begins with a full model, and on successive steps, deletes the predictor that contributes the least to the model.

Figure.12 Backward Elimination Automated Selection

Summary of Forward Selection								
Step	Variable Entered	Variable Removed	Number of Var In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1		OpenPorchSF	23	0	0.8054	23.0798	0.08	0.7776
2		ThreeSsnPorch	22	0	0.8054	21.1683	0.09	0.7661
3		LotFrontage	21	0	0.8053	19.477	0.31	0.5783
4		LowQualFinSF	20	0	0.8053	17.8072	0.33	0.5654
5	GrLivArea		21	0	0.8053	19.477	0.33	0.5654
6		GrLivArea	20	0	0.8053	17.8072	0.33	0.5654
7		extqual_ta	19	0	0.8053	16.379	0.57	0.4493
8		extcond_fa	18	0.0001	0.8052	15.7272	1.35	0.2454

Automated Selection Process - Stepwise Selection:
This is a combination of forward and backward selection, starting with the predictor having the highest simple correlation

Figure.13 Stepwise Selection Automated Selection

Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number of Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	GrLivArea		1	0.4992	0.4992	3747.83	2411.66	<.0001
2	TotalBsmtSF		2	0.127	0.6263	2185.96	821.89	<.0001
3	extqual_ta		3	0.0664	0.6927	1370.67	522.08	<.0001
4	extqual_ex		4	0.0388	0.7315	894.559	349.45	<.0001
5	GarageArea		5	0.0264	0.7579	572.027	262.91	<.0001
6	BsmtUnfSF		6	0.0126	0.7705	418.389	132.98	<.0001
7	extqual_gd		7	0.0091	0.7796	308.43	99.56	<.0001
8	MiscVal		8	0.0083	0.7878	208.841	93.82	<.0001
9	MasVnrArea		9	0.0053	0.7932	145.157	62.2	<.0001
10	WoodDeckSF		10	0.003	0.7962	109.845	35.84	<.0001
11	ScreenPorch		11	0.0026	0.7988	80.2945	30.68	<.0001
12	LotArea		12	0.0019	0.8007	58.9881	22.87	<.0001
13	extcond_fa		13	0.0015	0.8022	42.3738	18.4	<.0001
14		extqual_ta	12	0.0001	0.802	42.1872	1.79	0.1808
15	PoolArea		13	0.001	0.803	31.8857	12.21	0.0005
16	LowQualFinSF		14	0.0008	0.8039	23.5181	10.33	0.0013
17	BsmtFinSF1		15	0.0007	0.8045	17.2249	8.29	0.004
18	EnclosedPorch		16	0.0003	0.8048	15.9317	3.29	0.0696

Automated Selection Process - Adjusted R-Square:
This is automated selection ranks the models with the best adjusted R-square value

Figure.14 Adj-R-Square Automated Selection

Summary of Adjusted R-Square							
Step	Variable Entered	Number of Variables In	Model Adj-R-Square	Model R-Square	C(p)	AIC	BIC
1	extcond_ex	23	0.8037	0.8054	21.168	50946.435	50948.911
	extcond_gd						
	extcond_ta						
	extcond_fa						
	extqual_ex						
	extqual_gd						
	extqual_ta						
	LotFrontage						
	LotArea						
	MasVnrArea						
	BsmtFinSF1						
	BsmtFinSF2						
	BsmtUnfSF						
	LowQualFinSF						
	GrLivArea						
	GarageArea						
	WoodDeckSF						
	OpenPorchSF						
	EnclosedPorch						
	ThreeSsnPorch						
	ScreenPorch						
	PoolArea						
	MiscVal						

Automated Selection Process - AIC:

This automated process builds the models and ranks them in terms of AIC comparison. Typically, when we compare multiple models, we want to choose the model with the lower AIC value.

Figure.15 AIC Automated Selection

Summary of AIC selection							
Step	**Variable Entered**	**Number of Vars In**	**Model Adj-R-Square**	**Model R-Square**	**C(p)**	**AIC**	**BIC**
1	extcond_ex extcond_gd extcond_ta extcond_fa extqual_ex extqual_gd LotArea MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF LowQualFinSF GrLivArea GarageArea WoodDeckSF EnclosedPorch ScreenPorch PoolArea MiscVal	19	0.8037	0.8053	16.1357	50941.41	50943.81

Automated Selection Process - Mallow's C(p):
This process builds the models and ranks them in terms of Mallow's C(p). Typically, you should look for models where Mallows' Cp is small and relatively close to the number of predictor variables in the model.

Figure.16 Mallow's C(p) Automated Selection

Summary of AIC selection							
Step	Variable Entered	Number of Vars In	Model Adj-R-Square	Model R-Square	C(p)	AIC	BIC
1	extcond_fa extqual_ex extqual_gd LotArea MasVnrArea BsmtFinSF1 BsmtFinSF2 TotalBsmtSF LowQualFinSF GrLivArea GarageArea WoodDeckSF EnclosedPorch ScreenPorch PoolArea MiscVal	16	0.8035	0.8048	15.9317	50941.2584	50943.5139

Automated Selection Process - overview:

Figure.17 Automated selection model Comparison

Selection Method	Number of Vars In	Categorical Var	Contin Var	Adj-R-Square	Model R-Square	C(p)	AIC	BIC
Forward Selection	17	4	13	0.8035	0.8049	16.8404	50942.16	50944.45
Backward	18	5	13	0.8037	0.8052	15.7272	50941.02	50943.37
Stepwise	16	3	13	0.8035	0.8048	15.9317	50941.26	50943.51
Adj-R-Square	23	7	16	0.8037	0.8054	21.168	50946.43	50948.91
AIC	19	6	13	0.8037	0.8053	16.1357	50941.41	50943.81
Mallow's C(p)	16	3	13	0.8035	0.8048	15.9317	50941.26	50943.51

The automated selection process yielded different results across each selection process. This is the portion of an analysis where the data scientist earns their keep. The commonality among all of these automated selection methods is the end goal. Each of these processes is used to build and refine a model that optimizes the goodness-of-fit, conforms to assumptions, and provides predictive capability. All of the methods we used are valid in that sense. However, as we can observe from the breakdown in the table and the individual results of the selection process, no two models were the same, and they each used some combination of continuous and categorical variables to achieve their ideal and best fitting model.

This brings us back to our statement of a data scientist earning their keep. Adjusted R-Square, R-Square, Mallow's C(p), AIC, and BIC are all valid tools for determining model fit and potential. There is very little difference in the Adjusted R-square and R-square values across the different selection processes, and the same can be said of AIC and BIC. What stands out then is Mallow's C(p). As we previously mentioned, we want Mallow's C(p) to be relatively small and close to the number of predictor variables in the model. Two models from the Automated selection process fit that desired description best, Stepwise selection, and the model that selected parameters based on the Mallow's C(p) value. Their only differentiation was one continuous variable. The choice therefore comes down to selecting one of those two models, and in the end, we will side with the Stepwise Selected model objectively basing our decision on the process of building the model through stepwise selection.

We will now consider the base model to be from our automated stepwise selection. We will use this model and make some adjustments for interpretive purposes. One thing we will be doing is adding all relevant dummy coded categorical variables, and removing any statistically insignificant continuous variables. The table below has a breakdown of the variables included from the stepwise selection process and their p-values

Figure.18 Stepwise Selection variables

Number of Vars In	Variable Entered	Var Type	Pr > F
1	GrLivArea	Cont	<.0001
2	TotalBsmtSF	Cont	<.0001
4	extqual_ex	Cat	<.0001
5	GarageArea	Cont	<.0001
6	BsmtUnfSF	Cont	<.0001
7	extqual_gd	Cat	<.0001
8	MiscVal	Cont	<.0001
9	MasVnrArea	Cont	<.0001
10	WoodDeckSF	Cont	<.0001
11	ScreenPorch	Cont	<.0001
12	LotArea	Cont	<.0001
13	extcond_fa	Cat	<.0001
13	PoolArea	Cont	0.0005
14	LowQualFinSF	Cont	0.0013
15	BsmtFinSF1	Cont	0.004
16	EnclosedPorch	Cont	0.0696

From the table, we see that we have 16 total variables consisting of 13 continuous variables and 3 categorical variables. The categorical variables come from external quality and external condition, this means that we will add all coded dummy variables for each of those categories. EnclosedPorch is our lone continuous variable with a statistically insignificant p-value. We will remove that variable from our adjusted model.

Figure.19 Base Model Adjustments

Variable Entered	Variable Removed	Var Type
extqual_ta		Cat
extcond_ex		Cat
extcond_gd		Cat
extcond_ta		Cat
	EnclosedPorch	Cont

The results of running the adjusted regression model on Saleprice produced the following parameter results:

Figure.20 Adjusted base model Parameter results

Variable	Degrees of Freedom	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-46904	21466	-2.19	0.029
extcond_ex	1	51419	23477	2.19	0.0286
extcond_gd	1	46378	21105	2.2	0.0281
extcond_ta	1	48493	21018	2.31	0.0211
extcond_fa	1	25798	21346	1.21	0.2269
extqual_ex	1	122121	8048.78364	15.17	<.0001
extqual_gd	1	50413	6954.39476	7.25	<.0001
extqual_ta	1	7699.98599	6739.54987	1.14	0.2533
GrLivArea	1	57.09394	1.73062	32.99	<.0001
TotalBsmtSF	1	28.60456	4.28965	6.67	<.0001
GarageArea	1	52.80673	4.03276	13.09	<.0001
BsmtUnfSF	1	-7.6854	4.09822	-1.88	0.0609
MiscVal	1	-10.33855	1.1922	-8.67	<.0001
MasVnrArea	1	31.62968	4.44223	7.12	<.0001
WoodDeckSF	1	40.27801	5.70364	7.06	<.0001
ScreenPorch	1	66.0905	12.12383	5.45	<.0001
LotArea	1	0.43201	0.0916	4.72	<.0001
PoolArea	1	-52.09911	19.11962	-2.72	0.0065
LowQualFinSF	1	-53.1198	14.73991	-3.6	0.0003
BsmtFinSF1	1	13.06077	4.22052	3.09	0.002

The complete fitted equation for our model is the following:

SalePrice = -46,904 + 51419 * extcond_ex + 46,378 * extcond_gd + 48,493 * extcond_ta + 25,798 * extcond_fa + 122,121 * extqual_ex + 50,413 * extqual_gd + 7,699.99 * extqual_ta + 57.09 * GrLivArea + 28.60 * TotalBsmtSF **+ 52.80 * GarageArea** + (- **7.69 *** BsmtUnfSF) + (- **10.34 *** MiscVal) + **31.63 * MasVnrArea** + **40.28 * WoodDeckSF** + **66.09 * ScreebPorch**+ **.43 * LotArea** + (- **52.10 * PoolArea) +** (**-53.12 * LowQualFinSF)** + **13.06 * BsmtFinSF1**

A preliminary overview of this new model doesn't look like a good fit. We have several variables that are now considered to be statistically non-significant, and we also have negative t-values present.

The goodness-of-fit comparison between our base model, which was selected through the automated stepwise process, and our newly fitted model are below.

Figure.20 Base Model and Newly fitted Model Comparison

Model	Number of Vars	Categorical Var	Continues Var	Adj-R-Square	Model R-Square	C(p)	AIC	BIC
Stepwise Base	16	3	14	0.8035	0.8048	15.9317	50941.26	50943.51
Base Adjusted	19	7	15	0.7951	0.7965	20	60994.55	60996.83

The breakdown of goodness-of-fit solidifies what we saw in the initial analysis of parameter estimates. While the R-Square and Adjusted R-Square remains relatively high in the adjusted model, there is a drop-off present as well. The most telling information comes from Mallow's C(p) and AIC. The value of Mallow's C(p) is 20 in the adjusted model, versus 19 parameter variable. The C(p) for our base model is only slightly off when compared to the variables used to model Sale Price. Also, if we look at both models, we see that the AIC value is lower in our base model. Both of these metrics gives us confidence in saying that the Stepwise selected model is a better fit than our adjusted model.

Model Validation Framework:

To better assess the validity and accuracy of our model, we will now move into the use of cross-validation. To accomplish this goal, we will be creating a training and test split of our dataset on a 70-30 line. This will allow us to run our model using 70% of the dataset and examine the model's overall accuracy on the remaining 30% of the data. The use of cross-validation is an important component of regression analysis because it functions as our primary way of measuring the predictive performance for our models.

To begin, we will compare our previously used automated stepwise selection model using the full dataset and run the same automated selection method and model components on the training data.

Figure.21 Original Model and Train fitted Model Comparison

Original Source	DF	Sum of Squares	Mean Square	F-Value	Pr>F
Model	16	1.35304E+13	8.4565E+11	619.53	<.0001
Error	2404	3.28144E+12	1364992620		
Corrected Total	2420	1.68118E+13			

Train Data Source	DF	Sum of Squares	Mean Square	F-Value	Pr>F
Model	15	8.55336E+12	5.70224E+11	421.7	<.0001
Error	1668	2.25546E+12	1352192878		
Corrected Total	1683	1.08088E+13			

The models both exhibit statistical significance with high F-Values and statistically significant P-Values as well.

The model components in each case are also very similar. We have almost the same components in each model, with the exception of the variable BsmtUnSf1 missing from the model run against the training data. It's worth noting that this variable also presented the least amount of statistical significance. The fact that there is a good amount of overlap is promising for our original model.

Figure.22 Original Model and Train fitted Model Components Comparison

Original Variable	Degrees of Freedom	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	8053.72851	2886.664	10625070074	7.78	0.0053
extcond_fa	-23703	4871.39469	32317070724	23.68	<.0001
extqual_ex	116721	4613.32135	8.74E+11	640.14	<.0001
extqual_gd	44667	1957.40001	7.11E+11	520.72	<.0001
LotArea	0.67819	0.12937	37514404775	27.48	<.0001
MasVnrArea	36.91232	5.0724	72284548842	52.96	<.0001
BsmtFinSF1	13.94814	4.94538	10858328026	7.95	0.0048
BsmtUnfSF	-7.38777	4.77879	3262281509	2.39	0.1222
TotalBsmtSF	26.67017	5.02873	38394135525	28.13	<.0001
LowQualFinSF	-48.1582	15.85819	12588201752	9.22	0.0024
GrLivArea	56.29836	1.99128	1.09E+12	799.33	<.0001
GarageArea	54.1106	4.44264	2.02E+11	148.35	<.0001
WoodDeckSF	43.87018	6.72619	58067241520	42.54	<.0001
EnclosedPorch	-22.08343	12.16652	4497082309	3.29	0.0696
ScreenPorch	73.30442	13.59309	39696681993	29.08	<.0001
PoolArea	-67.91598	21.42119	13721035871	10.05	0.0015
MiscVal	-16.40993	1.51694	1.60E+11	117.02	<.0001

Train Data Variable	Degrees of Freedom	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	14699	3412.52922	25086158372	18.55	<.0001
extcond_fa	-22755	5712.13743	21457673147	15.87	<.0001
extqual_ex	117513	5658.41505	5.83E+11	431.31	<.0001
extqual_gd	46706	2323.63664	5.46E+11	404.03	<.0001
LotArea	0.54362	0.21398	8727450377	6.45	0.0112
MasVnrArea	34.11503	5.96735	44194331854	32.68	<.0001
BsmtFinSF1	20.35603	2.4059	96799144916	71.59	<.0001
TotalBsmtSF	17.09958	2.82695	49473740517	36.59	<.0001
LowQualFinSF	-70.93991	18.58989	19690941762	14.56	0.0001
GrLivArea	55.63866	2.37272	7.44E+11	549.87	<.0001
GarageArea	52.30836	5.31885	1.31E+11	96.72	<.0001
WoodDeckSF	33.60257	8.1896	22764506494	16.84	<.0001
EnclosedPorch	-35.69432	15.17672	7479644493	5.53	0.0188
ScreenPorch	67.21581	16.41717	22666518695	16.76	<.0001
PoolArea	-71.97082	23.27064	12934055202	9.57	0.002
MiscVal	-16.18131	1.57194	1.43E+11	105.96	<.0001

The overall fit between the two models is also fairly similar between the them, but there are a few things worth mentioning.

Figure.23 Goodness-of-fit Original Model and Train fitted Model Comparison

Model	Number of Vars In	Cat Var	Continuous Var	Adj-R-Square	Model R-Square	C(p)	AIC	BIC
Stepwise Original	16	3	13	0.8035	0.8048	15.9317	50941.26	50943.51
Training Model	15	3	12	0.7895	0.7913	17.5265	35422.01	35424.29

Mallow's C(p) is slightly better in our original model but the AIC measurement is leaning towards the training model.

To assess the models fit for validation we can use the residuals to calculate the Mean Squared Error (MSE) and the Mean Absolute Error (MAE) for the testing sample, and compare that with the training sample. MAE is the mean of the absolute errors. The absolute error is the absolute value of the difference between the forecasted value and the actual value. The MSE is an average of the squares of the difference between the actual observations and those predicted. The formulas are below:

$$\mathbf{MSE} = \frac{\boldsymbol{SSE}}{\boldsymbol{n-P}}$$

$$\mathbf{MAE} = \frac{\boldsymbol{SAE}}{\boldsymbol{n}}$$

Figure.24 Mean Absolute Error for Training Model vs. Test Model

Training Model		
N	**Mean Absolute Error**	**Mean Square Error**
2020	23268.58	1352192878
Testing Model		
N	**Mean Absolute Error**	**Mean Square Error**
737	23028.4	1316534661

The MAE is the simplest way for us to determine the forecast accuracy of our model. Typically, what we want to see is similar values in our training and testing model. If you look at the table above in Figure.24, you see that our MAE is very close. There is and should be a slight difference in the magnitude of the testing model. Overall, it appears that this model is a good fit, and presents evidence that it could serve for predictive purpose.

Operational Validation:
The final check of our model's predictive ability will come from validating our model in a functional sense. To do this, we will set up different benchmarks for the success of our training model when running against the testing model. The predicted values will be the following:

Figure.25 Predictive Grading Scale for Training Model vs. Test Model

Predicted Value Grade	Parameter Buckets
Grade 1	Less than 10%
Grade 2	Less than 15%
Grade 3	Greater than 15%

These markers will allow us a grade for measuring the accuracy of our Original and Training models. The final results are below in Figure.26.

Figure.25 Predictive Grading Scale for Original and Training Model

Original Model Prediction Grade Level	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Grade 1	1482	50.58	1482	50.58
Grade 2	500	17.06	1982	67.65
Grade 3	948	32.35	2930	100
Training Model Prediction Grade Level	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Grade 1	1038	50.91	1038	50.91
Grade 2	337	16.53	1375	67.44
Grade 3	664	32.56	2039	100

It's a good sign that both models come in with similar results. The question for us is to interpret what those results mean. What these numbers are telling us in both cases is that 51% of the observations fell within 10%. An additional 17% were located within 15% of the residuals giving us a cumulative rate of 67%. This validates some of what we saw earlier. The model does appear to have predictive ability. The statistical measurements and comparisons back up this assessment, and this final operational validation backs it up. It isn't the best model, but it definitely has potential.

Conclusion:
After going through and analyzing the data with various modeling techniques, it is still clear that there some challenges when working with a dataset such as this. The mix of continues and categorical variables in this particular data set leave open possibilities of further exploration and model fitting. The challenge, however, lies in those possibilities. Categorical variables need to be addressed with dummy coding to assess their true predictive capabilities. In short, there are possibilities and difficulties still available to us to work on.

With that said, it is my opinion that while there are countless possibilities for building robust models, I find it easier to visualize one that is simple and with fewer parameters. The variables need to be statistically significant of course, but we can rely on other measures besides R-Square to assess the model's initial predictive ability and run it through a validation process for confirmation of those initial results. I believe dummy coding our variables for inclusion in our regression analysis gives us an avenue towards finding the right model. The ease and speed associated with building models in SAS with Automated variable selection processes make it possible to streamline the potential models. This ease and flexibility also reduce the need for subjective inclusion of variables into the modeling process.

Through statistical inference from our sample, we can expand this model to the broader population data. It isn't perfect by any means, but if we have a large sample such as the AMES housing data, and we can build statistically significant models from this data, the inclusion of the population should yield predictive results as well. To me, it always feels like the best method starts with proper statistical inference on the sample. This will always allow us to build future models up in a concise manner.

Appendix SAS Code

```
/*Read file into SAS and identify it as mydata*/
libname mydata '/scs/wtm926/' access=readonly;

Data temp1;
set mydata.ames_housing_data;

/***Pre-Check data contents***/
proc contents data=temp1 order=varnum;
run;

/***Pre-Freq tables for Categorical variables***/
proc freq data=temp1;
  tables ExterQual ExterCond ;
run;

/***Pre-Mean data for categorical variable ExterQual***/
proc sort data=temp1;
  by ExterQual;
proc means data=temp1;
  by ExterQual;
  var saleprice;
run;

/***Pre-Means data for categorical variable ExterCond***/
proc sort data=temp1;
  by ExterCond;
proc means data=temp1;
  by ExterCond;
  var saleprice;
run;

/******* Part A *******/

/***1. Select New Character variable and numerically code***/
Data Part1;
set temp1;
Keep saleprice extercond exterqual extqual Lotfrontage LotArea MasVnrArea BsmtFinSF1
BsmtFinSF2
   BsmtUnfSF TotalBsmtSF FirstFlrSF SecondFlrSF LowQualFinSF GrLivArea GarageArea
WoodDeckSF OpenPorchSF
   EnclosedPorch ThreeSsnPorch ScreenPorch PoolArea MiscVal;
if exterqual='Ex' then extqual=1;
if exterqual='Gd' then extqual=2;
if exterqual='TA' then extqual=3;
if exterqual='Fa' then extqual=4;
```

```
proc contents data=Part1 order=varnum;
run;

/***Use proc sort and Proc means to find mean of Y(Saleprice)***/
proc sort data=Part1;
  by extqual;
proc means data=Part1;
  by extqual;
  var saleprice;
run;

/***2.Fit a simple linear regression using var part 1***/
proc reg data=part1;
   model saleprice = extqual;
run;

proc freq data=part1;
tables exterqual extqual;
run;

/***3.Set up dummy variable for Categorical variable from part 1***/
Data Part3;
   set part1;
keep saleprice extercond exterqual extqual_ex extqual_gd extqual_ta extqual Lotfrontage LotArea
MasVnrArea
   BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF FirstFlrSF SecondFlrSF LowQualFinSF
GrLivArea GarageArea
   WoodDeckSF OpenPorchSF EnclosedPorch ThreeSsnPorch ScreenPorch PoolArea MiscVal;
   if exterqual in ('Ex' 'Gd' 'TA' 'Fa') then do;
   extqual_ex = (exterqual eq 'Ex');
   extqual_gd = (exterqual eq 'Gd');
   extqual_ta = (exterqual eq 'TA');
end;

Proc freq data=part3;
tables exterqual extqual_ex extqual_gd extqual_ta;
run;

/***3.B Fit a regression model for dummy variable***/
proc reg data=part3;
model saleprice = extqual_ex extqual_gd extqual_ta;
run;

/***5. Select a new Character variable and dummy code***/
Data Part5;
```

```
   set part3;
keep saleprice extercond extcond_ex extcond_gd extcond_ta extcond_fa exterqual extqual_ex
extqual_gd extqual_ta
   extqual Lotfrontage LotArea MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF
FirstFlrSF SecondFlrSF
   LowQualFinSF GrLivArea GarageArea WoodDeckSF OpenPorchSF EnclosedPorch
ThreeSsnPorch ScreenPorch PoolArea
   MiscVal;
if extercond in ('Ex' 'Gd' 'TA' 'Fa' 'Po') then do;
   extcond_ex = (extercond eq 'Ex');
   extcond_gd = (extercond eq 'Gd');
   extcond_ta = (extercond eq 'TA');
   extcond_fa = (extercond eq 'Fa');
end;

Proc freq data=part5;
tables extercond extcond_ex extcond_gd extcond_ta extcond_fa;
run;

/***Check results in table***/
Proc freq data=Part5;
   tables extercond extcond_ex extcond_gd extcond_ta extcond_fa;
run;

/*******PartB*******/

/***6.Use all continuous and dummy variables in automated selection process***/
Data Part6;
   set part5;
keep saleprice extcond_ex extcond_gd extcond_ta extcond_fa extqual_ex extqual_gd extqual_ta
   Lotfrontage LotArea MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF
FirstFlrSF SecondFlrSF
   LowQualFinSF GrLivArea GarageArea WoodDeckSF OpenPorchSF EnclosedPorch
ThreeSsnPorch ScreenPorch PoolArea
   MiscVal;


proc contents data=Part6 order=varnum;
run;

/*******(A-1)Forward Selection*******/
proc reg data=part6;
   model saleprice = extcond_ex extcond_gd extcond_ta extcond_fa extqual_ex extqual_gd
extqual_ta
   Lotfrontage LotArea MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF
FirstFlrSF SecondFlrSF
```

```
   LowQualFinSF GrLivArea GarageArea WoodDeckSF OpenPorchSF EnclosedPorch
ThreeSsnPorch ScreenPorch PoolArea
   MiscVal/
selection=forward;

/*******(A-2) Backward Selection *******/
proc reg data=part6;
   model saleprice = extcond_ex extcond_gd extcond_ta extcond_fa extqual_ex extqual_gd
extqual_ta
   Lotfrontage LotArea MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF
FirstFlrSF SecondFlrSF
   LowQualFinSF GrLivArea GarageArea WoodDeckSF OpenPorchSF EnclosedPorch
ThreeSsnPorch ScreenPorch PoolArea
   MiscVal/
selection=backward;

/*******(A-3) Stepwise Selection *******/
proc reg data=part6;
   model saleprice = extcond_ex extcond_gd extcond_ta extcond_fa extqual_ex extqual_gd
extqual_ta
   Lotfrontage LotArea MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF
FirstFlrSF SecondFlrSF
   LowQualFinSF GrLivArea GarageArea WoodDeckSF OpenPorchSF EnclosedPorch
ThreeSsnPorch ScreenPorch PoolArea
   MiscVal/
selection=Stepwise;

/*******(A-4)Adjusted R-Square + AIC *******/
proc reg data=part6 outest=rsqest;
   model saleprice = extcond_ex extcond_gd extcond_ta extcond_fa extqual_ex extqual_gd
extqual_ta
   Lotfrontage LotArea MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF
FirstFlrSF SecondFlrSF
   LowQualFinSF GrLivArea GarageArea WoodDeckSF OpenPorchSF EnclosedPorch
ThreeSsnPorch ScreenPorch PoolArea
   MiscVal/
selection=adjrsq aic bic cp;

proc print data=rsqest;
proc sort data=rsqest; by _rsq_;
proc print data=rsqest;
/***run;**/

proc print data=rsqest;
proc sort data=rsqest; by _AIC_;
proc print data=rsqest;
```

```
/*******(A-5) Mallow's Cp *******/
proc reg data=part6 outest=rsqest;
   model saleprice = extcond_ex extcond_gd extcond_ta extcond_fa extqual_ex extqual_gd
extqual_ta
   Lotfrontage LotArea MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF
FirstFlrSF SecondFlrSF
   LowQualFinSF GrLivArea GarageArea WoodDeckSF OpenPorchSF EnclosedPorch
ThreeSsnPorch ScreenPorch PoolArea
   MiscVal/
selection=cp;

proc print data=rsqest;
proc sort data=rsqest; by _CP_;
proc print data=rsqest;

/*** 7. Select one of six models as base and refit***/
/******* Base model Stepwise selection*******/
proc reg data=part6;
   model saleprice = extcond_ex extcond_gd extcond_ta extcond_fa extqual_ex extqual_gd
extqual_ta
   Lotfrontage LotArea MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF
FirstFlrSF SecondFlrSF
   LowQualFinSF GrLivArea GarageArea WoodDeckSF OpenPorchSF EnclosedPorch
ThreeSsnPorch ScreenPorch PoolArea
   MiscVal/
selection=Stepwise;

/***Base adjusted model***/
proc reg data=part6;
   model saleprice = extcond_ex extcond_gd extcond_ta extcond_fa extqual_ex extqual_gd
extqual_ta
   GrLivArea TotalBsmtSF GarageArea BsmtUnfSF MiscVal MasVnrArea WoodDeckSF
ScreenPorch LotArea PoolArea
   LowQualFinSF BsmtFinSF1;
run;

proc reg data=part6;
   model saleprice = extcond_ex extcond_gd extcond_ta extcond_fa extqual_ex extqual_gd
extqual_ta
   GrLivArea TotalBsmtSF GarageArea BsmtUnfSF MiscVal MasVnrArea WoodDeckSF
ScreenPorch LotArea PoolArea
   LowQualFinSF BsmtFinSF1/
   selection = rsquare adjrsq aic bic cp;
run;
```

```
**** 8. Creat a train test split  ****;
Data Part8;
set part6;

* generate a uniform(0,1) random variable with seed set to 123;
u = uniform(123);
  if (u < 0.70) then train = 1;
  else train = 0;
  if (train=1) then train_response=SalePrice;
  else train_response=.;
    if (u > 0.70) then test = 1;
  else test = 0;
  if (test=1) then test_response=SalePrice;
  else test_response=.;
run;

/***original Model***/
proc reg data=part8;
   model saleprice = extcond_ex extcond_gd extcond_ta extcond_fa extqual_ex extqual_gd
extqual_ta
   Lotfrontage LotArea MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF
FirstFlrSF SecondFlrSF
   LowQualFinSF GrLivArea GarageArea WoodDeckSF OpenPorchSF EnclosedPorch
ThreeSsnPorch ScreenPorch PoolArea
   MiscVal/
selection=Stepwise;

/***Train Model***/
proc reg data=part8;
model train_response = extcond_ex extcond_gd extcond_ta extcond_fa extqual_ex extqual_gd
extqual_ta
   Lotfrontage LotArea MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF
FirstFlrSF SecondFlrSF
   LowQualFinSF GrLivArea GarageArea WoodDeckSF OpenPorchSF EnclosedPorch
ThreeSsnPorch ScreenPorch PoolArea
   MiscVal/
selection=stepwise;

/***Check AIC BIC for training model ***/
proc reg data=part8;
model train_response = extcond_ex extcond_gd extcond_ta extcond_fa extqual_ex extqual_gd
extqual_ta
   Lotfrontage LotArea MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF
FirstFlrSF SecondFlrSF
   LowQualFinSF GrLivArea GarageArea WoodDeckSF OpenPorchSF EnclosedPorch
ThreeSsnPorch ScreenPorch PoolArea
   MiscVal/
```

```
  selection = rsquare adjrsq aic bic cp;
run;

/***test model***/
proc reg data=part8;
model test_response = extcond_ex extcond_gd extcond_ta extcond_fa extqual_ex extqual_gd
extqual_ta
   Lotfrontage LotArea MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF
FirstFlrSF SecondFlrSF
   LowQualFinSF GrLivArea GarageArea WoodDeckSF OpenPorchSF EnclosedPorch
ThreeSsnPorch ScreenPorch PoolArea
   MiscVal/
selection=stepwise;

**** 9. C ****;
proc reg data=part8;
model train_response = extcond_ex extcond_gd extcond_ta extcond_fa extqual_ex extqual_gd
extqual_ta
   Lotfrontage LotArea MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF
FirstFlrSF SecondFlrSF
   LowQualFinSF GrLivArea GarageArea WoodDeckSF OpenPorchSF EnclosedPorch
ThreeSsnPorch ScreenPorch PoolArea
   MiscVal/
selection=stepwise;
output out=part9 predicted=yhat;

proc print data=part9(obs=5);
Data Part9b;
set part9;
mae = abs(yhat - train_response);

proc print data=part9b(obs=5);
proc means data=part9b;
var mae;
title 'MAE calculation Training';

/***9.C test data***/
proc reg data=part8;
model test_response = extcond_ex extcond_gd extcond_ta extcond_fa extqual_ex extqual_gd
extqual_ta
   Lotfrontage LotArea MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF
FirstFlrSF SecondFlrSF
   LowQualFinSF GrLivArea GarageArea WoodDeckSF OpenPorchSF EnclosedPorch
ThreeSsnPorch ScreenPorch PoolArea
   MiscVal/
selection=stepwise;
output out=part9c predicted=yhat;
```

```
proc print data=part9c(obs=5);
Data Part9d;
set part9c;
mae = abs(yhat - test_response);

proc print data=part9d(obs=5);
proc means data=part9d;
var mae;
title 'MAE calculation Testing';

**** Part 10 ****;
proc reg data=part8;
model train_response = extcond_ex extcond_gd extcond_ta extcond_fa extqual_ex extqual_gd
extqual_ta
   Lotfrontage LotArea MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF
FirstFlrSF SecondFlrSF
   LowQualFinSF GrLivArea GarageArea WoodDeckSF OpenPorchSF EnclosedPorch
ThreeSsnPorch ScreenPorch PoolArea
   MiscVal/
selection=stepwise;
output out=part10 predicted=yhat;
title ' ';
proc print data=part10 (obs=10);
run;

/*****/
Data Part10b;
set part10;

if train_response =. then delete;

Length Prediction_Grade $7.;
pct_diff = abs((yhat - train_response) / train_response);
if pct_diff LE .10 then Prediction_Grade = 'Grade 1';
else if pct_diff GT .10 and pct_diff LE .15 then Prediction_Grade= 'Grade 2';
else Prediction_Grade = 'Grade 3';

proc print data=part10b (obs=10);
run;

proc freq data=part10b;
tables prediction_grade;
run;

/***10 part 2 ***/
proc reg data=part8;
```

```
   model saleprice = extcond_ex extcond_gd extcond_ta extcond_fa extqual_ex extqual_gd
extqual_ta
   Lotfrontage LotArea MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF
FirstFlrSF SecondFlrSF
   LowQualFinSF GrLivArea GarageArea WoodDeckSF OpenPorchSF EnclosedPorch
ThreeSsnPorch ScreenPorch PoolArea
   MiscVal/
selection=Stepwise;
output out=part11 predicted=yhat;
title ' ';
proc print data=part11 (obs=10);
run;

/*****/
Data Part11b;
set part11;

if saleprice =. then delete;

Length Prediction_Grade $7.;
pct_diff = abs((yhat - saleprice) / saleprice);
if pct_diff LE .10 then Prediction_Grade = 'Grade 1';
else if pct_diff GT .10 and pct_diff LE .15 then Prediction_Grade= 'Grade 2';
else Prediction_Grade = 'Grade 3';

proc print data=part11b (obs=10);
run;

proc freq data=part11b;
tables prediction_grade;
run;

quit;
```

References

(1) Black, K. (2008). *Business statistics: For contemporary decision making*. Hoboken, NJ: Wiley.

(2) Montgomery, D. C., Peck, E. A., Vinning, G. G., (2012). *Introduction to Linear Regression Analysis* Hoboken, NJ: Wiley.

(3) Wilcox, R. R. (2009). *Basic statistics: Understanding conventional methods and modern insights*.Oxford: Oxford University Press.

(4) Cody, R. (2011). *SAS: Statistics by Example*. Carey, NC: SAS Institute Inc.

(5) Evaluating forecast Accuracy. https://www.otexts.org/fpp/2/5 (accessed February 4, 2017)