Assignment #5
Crystal Mosley

Introduction:

The purpose of this data examination is to design a model which has *some* predictive explanatory value for the SalePrice of a home, given relative information about observed sales. We will choose some categorical variables that are based on observable features of the houses. It is best to choose observable features as opposed to the categorical variables that are indications of a level of quality due to the potential bias of the quality which is dependent on the person deciding the level of quality.

Results:

Coding of Categorical Variable: HouseStyle

This variable contains eight different categories—

1Story	One story
1.5Fin	One and one-half story: 2nd level finished
1.5Unf	One and one-half story: 2nd level unfinished
2Story	Two story
2.5Fin	Two and one-half story: 2nd level finished
2.5Unf	Two and one-half story: 2nd level unfinished
SFoyer	Split Foyer
SLvl	Split Level

Mean SalePrice of HouseStyle, without categories of each variable: (*note: it utilizes 1Story as the default category*)

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
4	184000.00	60404.19	105000.00	244000.00

Simple linear regression model:

$$SalePrice = \beta_0 + \beta_1 HouseStyle + \varepsilon$$

Parameter Estimations and Model Diagnostics:

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	176091	2413.64078	72.96	<.0001
h_Style	1	1851.04809	751.81358	2.46	0.0139

Model Estimator and Performance:

Source	
Root MSE	79818
R-Square	0.0021
Adj R-Square	0.0017
F Value	6.06

Model Equation*:*

SalePrice = 176091 + 1851.04809 x HouseStyle

Interpretation:

In reviewing this model, we see that the SalePrice is $177,942.05. The *mean* table shows the mean value to be $184,000.00. There's alert for concern since the mean value is off by $6,057.95.

Dummy Coding of Categorical Variable: HouseStyle

We will now discuss the dummy coding of HouseStyle. Since HouseStyle is an eight-way variable, we will be able to model it with seven dummy coded variables by holding over one variable as the basis of interpretation.

Mean SalePrice of HouseStyle, with categories of each variable:

Analysis Variable : SalePrice						
House Style	N Obs	N	Mean	Std Dev	Minimum	Maximum
1.5Fin	314	314	137529.92	47225.67	37900.00	410000.00
1.5Unf	19	19	109663.16	20569.59	64500.00	139400.00
1Story	1481	1481	178699.88	81066.94	12789.00	615000.00
2.5Fin	8	8	220000.00	118211.98	104000.00	475000.00
2.5Unf	24	24	177158.33	76114.76	97500.00	415000.00
2Story	873	873	206990.16	85349.91	40000.00	755000.00
SFoyer	83	83	143472.66	31220.08	70000.00	224500.00
SLvl	128	128	165527.38	34348.13	91000.00	345000.00

First, we use a data procedure to dummy code HouseStyle as an indicator variable. To evaluate progress, we utilize the PROC FREQ of HouseStyle:

House Style	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1.5Fin	314	10.72	314	10.72
1.5Unf	19	0.65	333	11.37
1 Story	1481	50.55	1814	61.91
2.5Fin	8	0.27	1822	62.18
2.5Unf	24	0.82	1846	63.00
2 Story	873	29.80	2719	92.80
SFoyer	83	2.83	2802	95.63
SLvl	128	4.37	2930	100.00

Simple linear regression model: (without h_Style_8 since it'll be the basis of interpretation)

$$SalePrice = \beta_0 + \beta_1 h\_Style\_1 + \beta_2 h\_Style\_2 + \beta_3 h\_Style\_3 + \beta_4 h\_Style\_4 + \beta_5 h\_Style\_5 + \beta_6 h\_Style\_6 + \beta_7 h\_Style\_7 + \varepsilon$$

Parameter Estimations and Model Diagnostics:

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	B	165527	6790.26723	24.38	<.0001
h_Style_1	B	13172	7077.62191	1.86	0.0628
h_Style_2	B	-27997	8056.25449	-3.48	0.0005
h_Style_3	B	-55864	18887	-2.96	0.0031
h_Style_4	B	41463	7271.04417	5.70	<.0001
h_Style_5	B	54473	27997	1.95	0.0518
h_Style_6	B	11631	17088	0.68	0.4962
h_Style_7	B	-22055	10827	-2.04	0.0417
h_Style_8	0	0	.	.	.

Model Estimator and Performance:

Source	
Root MSE	76823
R-Square	0.0774
Adj R-Square	0.0752
F Value	35.04

Model Equation:

$$SalePrice = 165527 + 13172 \times h\_Style\_1 - 27997 \times h\_Style\_2 - 55864 \times h\_Style\_3 + 41463 \times h\_Style\_4 + 54473 \times h\_Style\_5 + 11631 \times h\_Style\_6 - 22055 \times h\_Style\_7$$

Interpretations:

In reviewing this model, if HouseStyle is 1, then the model becomes:

$$SalePrice = 165527 + 13172$$

- Also, if HouseStyle is 1, then the SalePrice in this model is $178,699. Comparing this value to the *mean table*, we see 1Story had a mean of $178,699.88

If HouseStyle is 2, then the model becomes:

$$SalePrice = 165527 - 27997$$

- Also, if HouseStyle is 2, then the SalePrice in this model is $137,530. Comparing this value to the *mean table*, we see 1.5Fin had a mean of $137,529.92

If HouseStyle is 3, then the model becomes:

SalePrice = 165527 - 55864

- Also, if HouseStyle is 3, then the SalePrice in this model is $109,663. Comparing this value to the *mean table*, we see 1.5Unf had a mean of $109,663.16

If HouseStyle is 4, then the model becomes:

SalePrice = 165527 + 41463

- Also, if HouseStyle is 4, then the SalePrice in this model is $206,990. Comparing this value to the *mean table*, we see 2Story had a mean of $206,990.16

If HouseStyle is 5, then the model becomes:

SalePrice = 165527 + 54473

- Also, if HouseStyle is 5, then the SalePrice in this model is $220,000. Comparing this value to the *mean table*, we see 2.5Fin had a mean of $220,000

If HouseStyle is 6, then the model becomes:

SalePrice = 165527 + 11631

- Also, if HouseStyle is 6, then the SalePrice in this model is $177,158. Comparing this value to the *mean table*, we see 2.5Unf had a mean of $177,158.33

If HouseStyle is 7, then the model becomes:

SalePrice = 165527 - 22055

- Also, if HouseStyle is 7, then the SalePrice in this model is $143,472. Comparing this value to the *mean table*, we see SFoyer had a mean of $143,472.66

Lastly, if HouseStyle is 8, then the model becomes:

SalePrice = 165527

- Also, if HouseStyle is 8, then the SalePrice in this model is $165,527. Comparing this value to the *mean table*, we see SLvl had a mean of $165,527.38

Dummy Code Hypothesis Testing:

$H_0 : \beta_{1..7} = 0$ versus $H_1 : \beta_{1..7} \neq 0$

(*same as* $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$)

Equation for f-stat of overall f-test: $F_0 = \frac{\frac{SSR}{k}}{\frac{SSE}{(n-p)}}$

$$F_0 = \frac{\frac{1.4475}{7}}{\frac{1.7245}{(2929-8)}} = 35.0259$$

[Note: the DF of the parameter estimates shows as "B", which indicates bias; therefore, only the ANOVA corrected total of observations were used.]

Dummy variables h_Style_1, h_Style_5, h_Style_6, and h_Style_7 do not yield statistically significant results; however, h_Style_2, h_Style_3, and h_Style_4, yield statically significant results.

Dummy Coding of Categorical Variable: GarageType

We will now discuss the dummy coding of GarageType. Since GarageType is a seven-way variable, we will be able to model it with seven dummy coded variables by holding over one variable as the basis of interpretation.

Mean SalePrice of GarageType, with categories of each variable:

Analysis Variable : SalePrice						
GarageType	N Obs	N	Mean	Std Dev	Minimum	Maximum
2Types	23	23	154639.13	34973.25	87000.00	235000.00
Attchd	1731	1731	203772.28	79141.09	13100.00	755000.00
Basment	36	36	150473.42	52342.58	55993.00	359100.00
BuiltIn	186	186	249344.15	95003.54	91000.00	582933.00
CarPort	15	15	105566.67	24985.92	71000.00	164900.00
Detchd	782	782	132468.47	39580.63	12789.00	475000.00
NA	157	157	104949.25	34069.81	34900.00	260000.00

Next, we use a data procedure to dummy code GarageType as an indicator variable. To evaluate progress, we utilize the PROC FREQ of GarageType:

GarageType	Frequency	Percent	Cumulative Frequency	Cumulative Percent
2Types	23	0.78	23	0.78
Attchd	1731	59.08	1754	59.86
Basment	36	1.23	1790	61.09
BuiltIn	186	6.35	1976	67.44
CarPort	15	0.51	1991	67.95
Detchd	782	26.69	2773	94.64
NA	157	5.36	2930	100.00

Automated Variable Selection Procedures:

We will utilize the continuous predictor variables from previous assignments that correlated with SalePrice. We will take the top five to incorporate into the automated variable selection strategy: GrLivArea, GarageArea, TotalBsmtSF, FirstFlrSF, and MasVnrArea.

Forward Selection Model: (h_Style)

SalePrice = -11084.84 + 80.4605 x GrLivArea + 89.0710 x GarageArea + 50.5407 x TotalBsmtSF -19.6204 x FirstFlrSF + 57.8657 x MasVnrArea – 29491.34 x h_Style_2 - 12004.24 x h_Style_4 - 62091.40 x h_Style_5 - 38312.13 x h_Style_6 - 3371.79 x h_Style_8

Model Estimator and Performance:

Source	
Root MSE	43538.68
C_p	9.2941
R-Square	0.7034
Adj. R-Square	0.7024
F Value	686.42
AIC	62069.94
BIC	62072.04

Forward Selection Model: (g_Type)

SalePrice = 131.174 + 56.9212 x GrLivArea + 102.990 x GarageArea + 44.3824 x TotalBsmtSF + 56.9500 x MasVnrArea – 65201.25 x g_Type_1 - 16694.73 x g_Type_3 + 20572.73 x g_Type_4 – 39958.15 x g_Type_5 – 24012.41 x g_Type_6

Model Estimator and Performance:

Source	
Root MSE	42210.15
C_p	8.1257
R-Square	0.7212
Adj. R-Square	0.7203
F Value	831.91
AIC	61888.90
BIC	61890.98

Backward Selection Model: (h_Style)

SalePrice = -11654.26 + 79.4499 x GrLivArea + 88.9630 x GarageArea + 50.9243 x TotalBsmtSF -18.6403 x FirstFlrSF + 57.6834 x MasVnrArea – 28623.30 x h_Style_2 - 10797.24 x h_Style_4 – 60270.53 x h_Style_5 - 37085.70 x h_Style_6

Model Estimator and Performance:

Source	
Root MSE	43535.91
C_p	7.9244

R-Square	0.7034
Adj. R-Square	0.7024
F Value	762.72
AIC	62068.58
BIC	62070.66

Backward Selection Model: (g_Type)

SalePrice = 131.174 + 56.9212 x GrLivArea + 102.990 x GarageArea + 44.3824 x TotalBsmtSF + 56.9500 x MasVnrArea – 65201.25 x g_Type_1 - 16694.73 x g_Type_3 + 20572.73 x g_Type_4 – 39958.15 x g_Type_5 – 24012.41 x g_Type_6

Model Estimator and Performance:

Source	
Root MSE	42210.15
C_p	8.1257
R-Square	0.7212
Adj. R-Square	0.7203
F Value	831.91
AIC	61888.90
BIC	61890.98

Stepwise Selection Model: (h_Style)

SalePrice = -11654.26 +79.4499 x GrLivArea + 88.9630 x GarageArea + 50.9243 x TotalBsmtSF - 18.6403 x FirstFlrSF + 57.6834 x MasVnrArea – 28623.30 x h_Style_2 – 10797.24 x h_Style_4 – 60270.53 x h_Style_5 – 37085.70 x h_Style_6

Model Estimator and Performance:

Source	
Root MSE	43535.91
C_p	7.9244
R-Square	0.7034
Adj. R-Square	0.7024
F Value	762.72
AIC	62068.58
BIC	62070.66

Stepwise Selection Model: (g_Type)

SalePrice = 131.174 + 56.9212 x GrLivArea + 102.990 x GarageArea + 44.3824 x TotalBsmtSF + 56.9500 x MasVnrArea – 65201.25 x g_Type_1 - 16694.73 x g_Type_3 + 20572.73 x g_Type_4 – 39958.15 x g_Type_5 – 24012.41 x g_Type_6

Model Estimator and Performance:

Source	
Root MSE	42210.15
C_p	8.1257
R-Square	0.7212
Adj. R-Square	0.7203
F Value	831.91
AIC	61888.90
BIC	61890.98

Adjusted R-Square Selection Model: (h_Style)

SalePrice = -11654.26 + 79.4499 x GrLivArea + 88.9630 GarageArea + 50.9243 x TotalBsmtSF - 18.6403 x FirstFlrSF + 57.6834 x MasVnrArea - 28623.30 x h_Style_2 - 10797.24 x h_Style_4 - 60270.53 x h_Style_5 - 37085.70 x h_Style_6

Model Estimator and Performance:

Source	
Root MSE	43535.91
C_p	7.9244
R-Square	0.7034
Adj. R-Square	0.7024
AIC	62068.58
BIC	62070.66

Adjusted R-Square Selection Model: (g_Type)

SalePrice = 131.17 + 56.9212 x GrLivArea + 102.990 x GarageArea + 44.3824 x TotalBsmtSF + 56.9500 x MasVnrArea – 65201.25 x g_Type_1 - 16694.73 x g_Type_3 + 20572.73 x g_Type_4 - 39958.15 x g_Type_5 – 24012.41 x g_Type_6

Model Estimator and Performance:

Source	
Root MSE	42210.15
C_p	8.1257
R-Square	0.7212

Adj. R-Square	0.7203
AIC	61888.90
BIC	61890.98

Mallow's C_p Selection Model: (h_Style)

SalePrice = -11654.26 + 79.4499 x GrLivArea + 88.9630 x GarageArea + 50.9243 x TotalBsmtSF - 18.6403 x FirstFlrSF + 57.6834 x MasVnrArea – 28623.30 x h_Style_2 - 10797.24 x h_Style_4 - 60270.53 x h_Style_5 - 37085.70 x h_Style_6

Model Estimator and Performance:

Source	
Root MSE	43535.91
C_p	7.9244
R-Square	0.7034
Adj. R-Square	0.7024
AIC	62068.58
BIC	62070.66

Mallow's C_p Selection Model: (g_Type)

SalePrice = 131.17 + 56.9212 x GrLivArea + 102.990 x GarageArea + 44.3824 x TotalBsmtSF + 56.9500 x MasVnrArea – 65201.25 x g_Type_1 - 16694.73 x g_Type_3 + 20572.73 x g_Type_4 - 39958.15 x g_Type_5 – 24012.41 x g_Type_6

Model Estimator and Performance:

Source	
Root MSE	42210.15
C_p	8.1257
R-Square	0.7212
Adj. R-Square	0.7203
AIC	61888.90
BIC	61890.98

AIC Selection Method: The regression procedure within SAS doesn't allow for selection by AIC criterion; therefore, we'll examine the output of C_p selection, and choose the model with the lowest value of AIC.

SalePrice = 131.17 + 56.9212 x GrLivArea + 102.990 x GarageArea + 44.3824 x TotalBsmtSF + 56.9500 x MasVnrArea – 65201.25 x g_Type_1 - 16694.73 x g_Type_3 + 20572.73 x g_Type_4 - 39958.15 x g_Type_5 – 24012.41 x g_Type_6

Model Estimator and Performance:

Source	
Root MSE	42210.15
C_p	8.1257
R-Square	0.7212
Adj. R-Square	0.7203
AIC	61888.90
BIC	61890.98

Comparing Model Performance: (h_Style)

Model	**Cont.**	**Ind.**	**Root MSE**	**C_p**	**R-Squ**	**Adj. R-Squ**	**F Value**	**AIC**	**BIC**
Forward	5	7	43538.68	9.2941	0.7034	0.7024	686.42	62069.94	62072.04
Backward	4	7	43535.91	7.9244	0.7034	0.7024	762.72	62068.58	62070.66
Stepwise	4	7	43535.91	7.9244	0.7034	0.7024	762.72	62068.58	62070.66
Adj. R-Squ	4	8	43535.91	7.9244	0.7034	0.7024	--	62068.58	62070.66
C_p	4	7	43535.91	7.9244	0.7034	0.7024	--	62068.58	62070.66
AIC	4	7	42210.15	8.1257	0.7212	0.7203	--	61888.90	61890.98

Comparing Model Performance: (g_Type)

Model	**Cont.**	**Ind.**	**Root MSE**	**C_p**	**R-Squ**	**Adj. R-Squ**	**F Value**	**AIC**	**BIC**
Forward	5	7	43538.68	9.2941	0.7034	0.7024	686.42	62069.94	62072.04
Backward	4	7	42210.15	8.1257	0.7212	0.7203	831.91	61888.90	61890.98
Stepwise	4	7	42210.15	8.1257	0.7212	0.7203	831.91	61888.90	61890.98
Adj. R-Squ	4	8	42210.15	8.1257	0.7212	0.7203	--	61888.90	61890.98
C_p	4	7	42210.15	8.1257	0.7212	0.7203	--	61888.90	61890.98
AIC	4	7	42210.15	8.1257	0.7212	0.7203	--	61888.90	61890.98

Models which incorporate more parameters become more complex for interpretation. The results for all the model selection methods show that these models all performed well by incorporating almost all of the continuous variables within them. Each variable selection procedure selected different models throughout the automated variable procedure.

All models except for AIC included the FirstFlrSF variable; consequently, the models will need to be refitted due not being statistically significant. Forward selection had the highest C_p value. When comparing model performance for the *g_Type* dummy variables, both C_p and AIC are received the same results because we were wanting to minimize AIC. For both *h_Style* and *g_Type* dummy variables, the C_p values are the highest over the other selections. Backward and Stepwise both show the same results

within each dummy variable table. Next, we'll use the models designed by Backward and Stepwise selection.

Indicator Variable Inclusion:

Refitting the model by incorporating all of the identification variables, resulted in the below.

H_Style Dummy Variable Refitting—

SalePrice = -16842 + 68.5000 x GrLivArea + 88.5224 x GarageArea + 46.5953 x TotalBsmtSF + 57.1992 x MasVnrArea + 662.5789 x h_Style_1 – 21156 x h_Style_2 + 1154.9598 x h_Style_3 + 581.2608 x h_Style_4 – 44405 x h_Style_5 – 25689 x h_Style_6 + 4615.2613 x h_Style_7 + 0 x h_Style_8

Model Estimator and Performance:

Source	
Root MSE	43610
R-Square	0.7026
Adj. R-Square	0.7014
F Value	621.20

G_Type Dummy Variable Refitting—

SalePrice = -470.8594 + 56.96397 x GrLivArea + 102.4136 x GarageArea + 44.3981 x TotalBsmtSF + 56.9820 x MasVnrArea - 64234 x g_Type_1 + 864.0634 x g_Type_2 – 15921 x g_Type_3 + 21401 x g_Type_4 – 39183 x g_Type_5 – 23238 x g_Type_6 + 0 x g_Type_7

Model Estimator and Performance:

Source	
Root MSE	42217
R-Square	0.7212
Adj. R-Square	0.7202
F Value	748.47

The dummy variables *h_Style_8* and *g_Type_7* were set to zero because they make the overall use of each set of indicator variables a linear combination; they are automatically marked as zero's by SAS. Most all the indicator variables were found to be *bias*, according to SAS, and only four are considered to be statistically significant. The complexity and performance of this model are considered to be poor practice.

We'll now run a model that excludes the indicator parameters to simplify the complexity.

SalePrice = -18458 + 63.8058 x GrLivArea + 96.0872 x GarageArea + 49.0212 x TotalBsmtSF + 62.6281 x MasVnrArea

Model Estimator and Performance:

Source	
Root MSE	44130
R-Square	0.6947
Adj. R-Square	0.6943
F Value	1649.63

In reviewing the updated model, we see that all dependent variables are all statistically significant, and the F Value has increased significantly compared to the last model.

Last, we'll run another model and exclude the MasVnrArea variable.

SalePrice = - 29536 + 68.8596 x GrLivArea + 105.0908 x GarageArea + 54.5856 x TotalBsmtSF

Model Estimator and Performance:

Source	
Root MSE	45254
R-Square	0.6795
Adj. R-Square	0.6791
F Value	2066.05

After reviewing the last model, each of the dependent variables are all still statistically significant, and the F Value has increased even more, compared to the last two models.

Validation Framework—

We will assess the predictive accuracy of the model developed, by using cross-validation and compare/contrast the difference between a statistical model validation, and an application model validation. A univariate distribution and sample will be used to select 70% of the observations from the cleansed dataset.

Adjusted R-Square Selection Model:

Train_Response = - 31889.53 + 74.3801 x GrLivArea + 82.5343 x GarageArea + 56.4940 x TotalBsmtSF + 52.7869 x MasVnrArea - 21291.23 x h_Style_2 - 43727.05 x h_Style_5 - 25151.23 x h_Style_6 + 11109.26 x h_Style_7

Model Estimator and Performance:

Source	
Root MSE	39839.24
Cp	6.0205

R-Square	0.7464
Adj. R-Square	0.7454
AIC	42887.86
BIC	42889.96

Mallow's C_p Selection Model:

Train_Response = - 31889.53 + 74.3801 x GrLivArea + 82.5343 x GarageArea + 56.4940 x TotalBsmtSF + 52.7869 x MasVnrArea - 21291.23 x h_Style_2 - 43727.05 x h_Style_5 - 25151.23 x h_Style_6 + 11109.26 x h_Style_7

Model Estimator and Performance:

Source	
Root MSE	39839.24
C_p	6.0205
R-Square	0.7464
Adj. R-Square	0.7454
AIC	42887.86
BIC	42889.96

AIC Selection Model: The regression procedure within SAS doesn't allow for selection by AIC criterion; therefore, we'll examine the output of C_p selection, and choose the model with the lowest value of AIC.

Train_Response = - 31889.53 + 74.3801 x GrLivArea + 82.5343 x GarageArea + 56.4940 x TotalBsmtSF + 52.7869 x MasVnrArea - 21291.23 x h_Style_2 - 43727.05 x h_Style_5 - 25151.23 x h_Style_6 + 11109.26 x h_Style_7

Model Estimator and Performance:

Source	
Root MSE	39839.24
C_p	6.0205
R-Square	0.7464
Adj. R-Square	0.7454
AIC	42887.86
BIC	42889.96

Forward Selection Model:

Train_Response = - 31075.27 + 75.0555 x GrLivArea + 82.7602 x GarageArea + 58.6846 x TotalBsmtSF - 3.64046 x FirstFlrSF + 52.8141 x MasVnrArea - 21508.66 x h_Style_2 - 43669.53 x h_Style_5 - 25594.87 x h_Style_6 + 11178.89 x h_Style_7

Model Estimator and Performance:

Source	
Root MSE	39841.99
C_p	7.2997
R-Square	0.7465
Adj. R-Square	0.7454
AIC	42889.13
BIC	42891.26

Backward Selection Model:

Train_Response = - 31889.53 + 74.3801 x GrLivArea + 82.5343 GarageArea + 56.4940 x TotalBsmtSF + 52.7869 x MasVnrArea - 21291.23 x h_Style_2 - 43727.05 x h_Style_5 - 25151.23 x h_Style_6 + 11109.26 x h_Style_7

Model Estimator and Performance:

Source	
Root MSE	39839.24
C_p	6.0205
R-Square	0.7464
Adj. R-Square	0.7454
AIC	42887.86
BIC	42889.96

Stepwise Selection Model:

Train_Response = - 31889.53 + 74.3801 x GrLivArea + 82.5343 x GarageArea + 56.4940 x TotalBsmtSF + 52.7869 x MasVnrArea - 21291.23 x h_Style_2 - 43727.05 x h_Style_5 - 25151.23 x h_Style_6 + 11109.26 x h_Style_7

Model Estimator and Performance:

Source	
Root MSE	39839.24
Cp	6.0205
R-Square	0.7464
Adj. R-Square	0.74542
AIC	42887.86
BIC	42889.96

Model Comparisons—

Automatic Variable Selection Model Comparison

Model	Cont.	Ind.	Root MSE	C_p	R-Squ	Adj. R-Squ	F Value	AIC	BIC
Forward	5	7	43538.68	9.2941	0.7034	0.7024	686.42	62069.94	62072.04
Backward	4	7	43535.91	7.9244	0.7034	0.7024	762.72	62068.58	62070.66
Stepwise	4	7	43535.91	7.9244	0.7034	0.7024	762.72	62068.58	62070.66
Adj. R-Squ	4	8	43535.91	7.9244	0.7034	0.7024	--	62068.58	62070.66
C_p	4	7	43535.91	7.9244	0.7034	0.7024	--	62068.58	62070.66
AIC	4	7	42210.15	8.1257	0.7212	0.7203	--	61888.90	61890.98

Training Data Automatic Variable Selection Model Comparison

Model	Cont.	Ind.	Root MSE	C_p	R-Squ	Adj. R-Squ	F Value	AIC	BIC
Forward	5	7	39841.99	7.2997	0.7465	0.7454	659.03	42889.13	42891.26
Backward	4	7	39839.24	6.0205	0.7464	0.7454	741.43	42887.86	42889.96
Stepwise	4	7	39839.24	6.0205	0.7464	0.7454	741.43	42887.86	42889.96
Adj. R-Squ	4	8	39839.24	6.0205	0.7464	0.7454	--	42887.86	42889.96
C_p	4	7	39839.24	6.0205	0.7464	0.7454	--	42887.86	42889.96
AIC	4	7	39839.24	6.0205	0.7464	0.7454	--	42887.86	42889.96

In reviewing the above model comparisons, once again the Backward and Stepwise model selections match each other within their respective tables; however, we notice the R-Square and Adj. R-Square values increase slightly in the training data table, but the other values decreased in the training table.

Next, we'll fit a multiple regression model to predict SalePrice; calling this model "Original". We'll then fit another regression model to predict Train_Response; calling this model "Training".

Original Model vs. Training Model—

1. SalePrice model utilizing the following variables: GrLivArea, GarageArea, TotalBsmtSF, MasVnrARea

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	1.285029E13	3.212574E12	1649.63	<.0001
Error	2900	5.647608E12	1947451081		
Corrected Total	2904	1.84979E13			

Root MSE	44130	R-Square	0.6947
Dependent Mean	180425	Adj R-Sq	0.6943
Coeff Var	24.45884		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-18458	2914.76045	-6.33	<.0001
GrLivArea	1	63.80583	1.96973	32.39	<.0001
GarageArea	1	96.08720	4.68307	20.52	<.0001
TotalBsmtSF	1	49.02124	2.25973	21.69	<.0001
MasVnrArea	1	62.62806	5.23653	11.96	<.0001

u	train	train_response	yhat
0.75040	0	.	197895.84
0.32091	1	105000	152092.48
0.17839	1	172000	168232.27
0.90603	0	.	269764.70
0.35712	1	189900	177287.50

Equations for Original SalePrice Model:

SalePrice = -18458 + 63.8058 x GrLivArea

SalePrice = - 18458 + 96.0872 x GarageArea

SalePrice = -18458 + 49.0212 x TotalBsmtSF

SalePrice = -18458 + 62.6281 x MasVnrArea

2. Train_Response model utilizing the following variables: GrLivArea, GarageArea, TotalBsmtSF, MasVnrARea

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	9.312171E12	2.328043E12	1424.30	<.0001
Error	2019	3.300083E12	1634513856		
Corrected Total	2023	1.261225E13			

Root MSE	40429	R-Square	0.7383
Dependent Mean	180416	Adj R-Sq	0.7378
Coeff Var	22.40881		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-33135	3275.19126	-10.12	<.0001
GrLivArea	1	69.32521	2.17320	31.90	<.0001
GarageArea	1	90.58461	5.16150	17.55	<.0001
TotalBsmtSF	1	58.52987	2.56039	22.86	<.0001
MasVnrArea	1	58.46740	5.72835	10.21	<.0001

u	train	train_response	yhat
0.75040	0	.	199257.15
0.32091	1	105000	146730.82
0.17839	1	172000	171361.60
0.90603	0	.	283924.70
0.35712	1	189900	177773.59

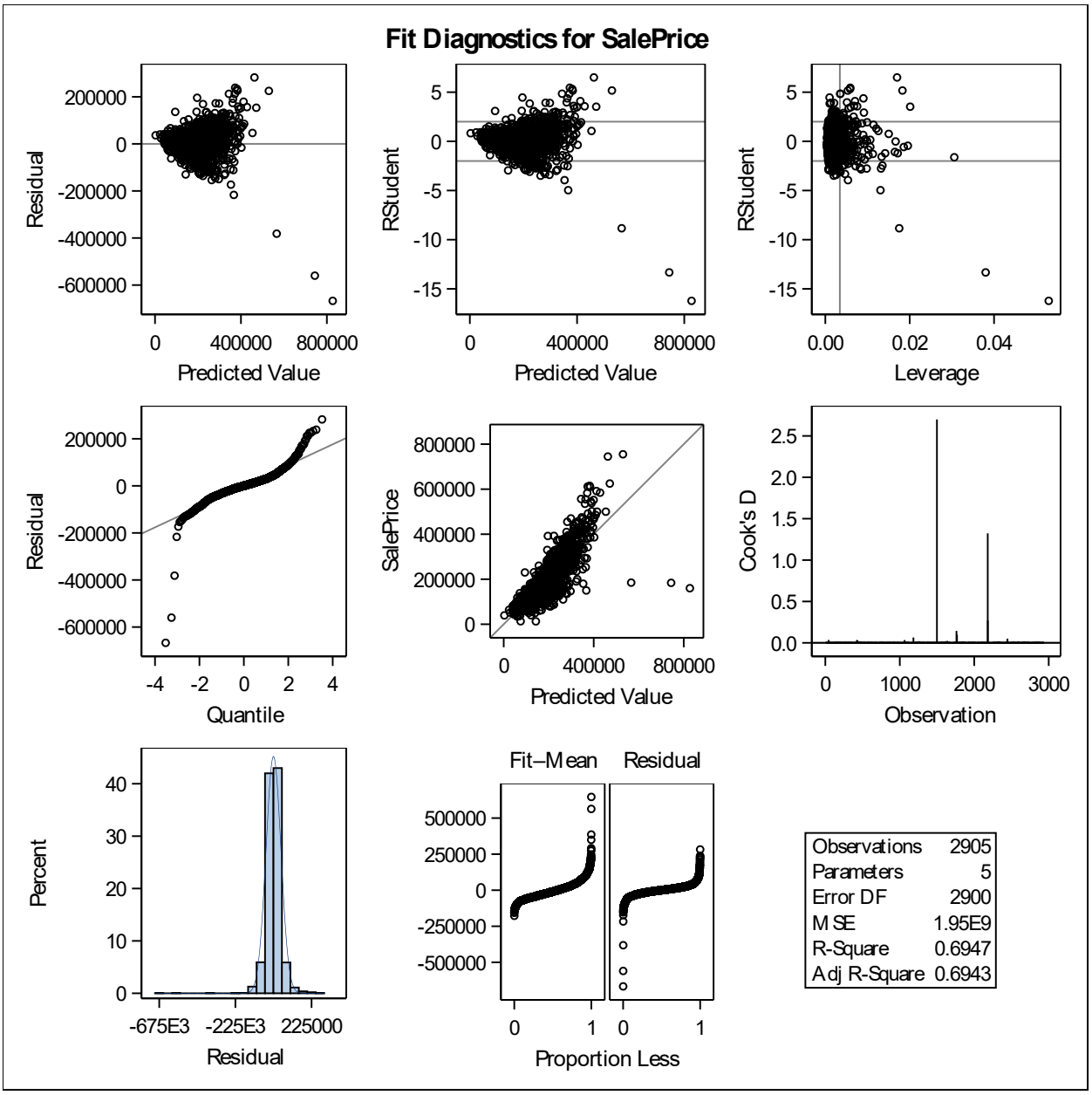
Equations for Training Train_Response Model:

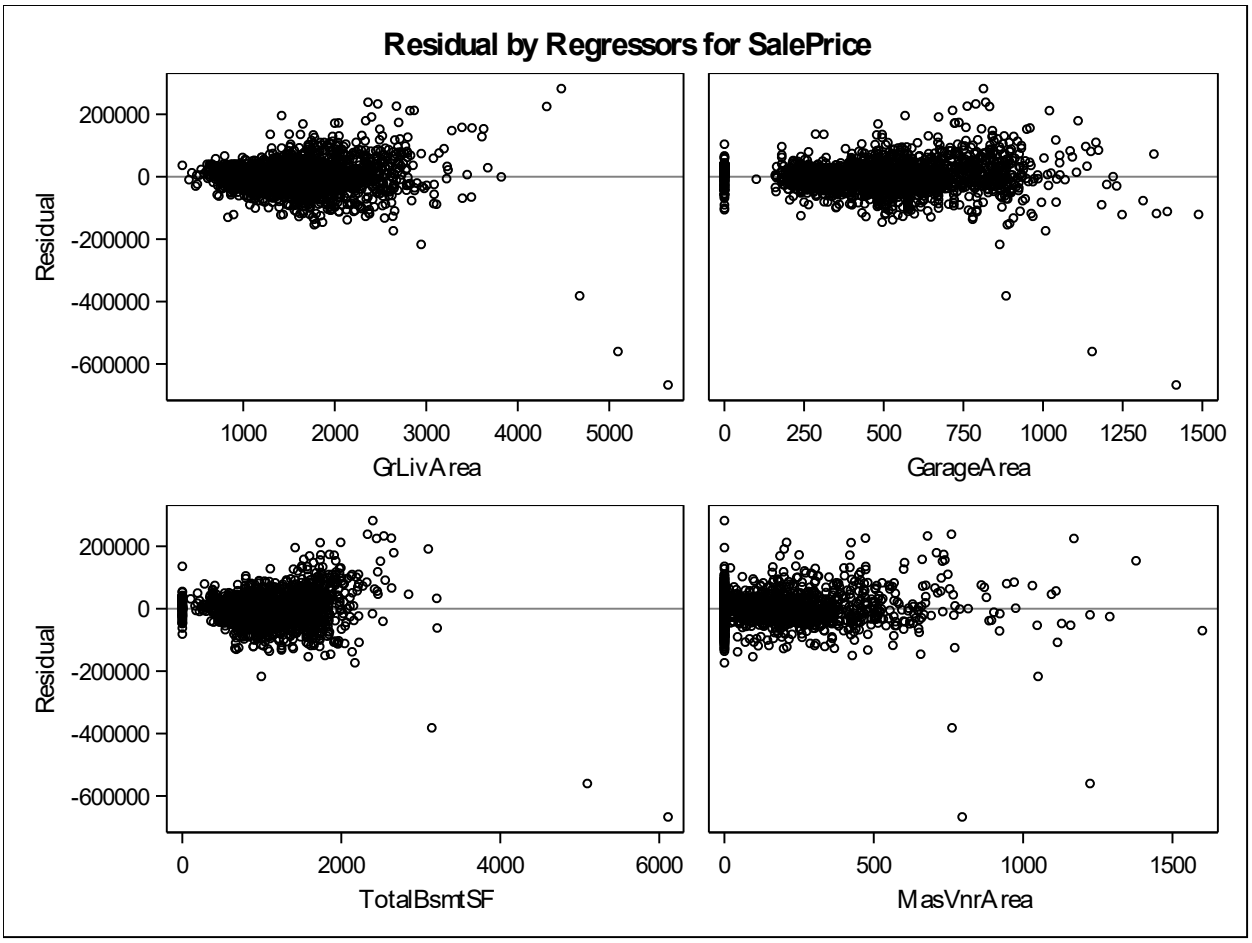
Train_Response = -33135 + 69.3252 x GrLivArea

Train_Response = - 33135 + 90.5846 x GarageArea

Train_Response = -33135 + 58.5299 x TotalBsmtSF

Train_Response = -33135 + 58.4674 x MasVnrArea


Fit Diagnostics for SalePrice
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
SalePrice
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 2905
Parameters 5
Error DF 2900
MSE 1.95E9
R-Square 0.6947
Adj R-Square 0.6943



- All variables in both models are statistically significant
- The F Value decreased with the Training model
- The 1Story houses priced at $105,000 had a decrease in the *yhat* for the predicted amount.
- This data shows many outliers fit diagnostics and residuals, as well as heteroscedasticity
- The RMSE, Dep. Mean, and Coeff Var. of the Original model all show a decrease in the Training model, while the R-Square and Adj. R-Square values show an increase in the Training model.

Next, we will assess the Training model fit for the validation data (the 30% sample).

Equation for Mean Squared Error (MSE): SSE/(n-p-1)

- SSE = Error Sum of Squares
    - MSE = 856,051,044.72 – this is very high!

Equation for Mean Absolute Error (MAE): SAE/N

- N = number of non-missing data points; SAE = sum of the absolute errors (or deviations)
    - MAE = 45,914.63

Operational Validation:

MAE calculation

The MEANS Procedure

Analysis Variable : mae				
N	Mean	Std Dev	Minimum	Maximum
2024	27395.71	29258.35	10.0295038	387349.93

u	train	train_response	yhat	mae
0.75040	0	.	200725.18	.
0.32091	1	105000	150914.63	45914.63
0.17839	1	172000	170578.32	1421.68
0.90603	0	.	272879.78	.
0.35712	1	189900	181233.55	8666.45

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	9.414117E12	1.176765E12	741.43	<.0001
Error	2015	3.198138E12	1587165269		
Corrected Total	2023	1.261225E13			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-31890	3339.52338	1.44727E11	91.19	<.0001
GrLivArea	74.38009	2.26375	1.71348E12	1079.59	<.0001
GarageArea	82.53433	5.19069	4.012738E11	252.82	<.0001
TotalBsmtSF	56.49398	2.53992	7.852121E11	494.73	<.0001
MasVnrArea	52.78691	5.69269	1.364708E11	85.98	<.0001
h_Style_2	-21291	3020.05317	78885041250	49.70	<.0001
h_Style_5	-43727	16635	10966384732	6.91	0.0086
h_Style_6	-25151	9572.19144	10957659234	6.90	0.0087
h_Style_7	11109	5534.57070	6394768166	4.03	0.0449

Summary of Forward Selection							
Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	GrLivArea	1	0.5229	0.5229	1765.94	2216.21	<.0001
2	TotalBsmtSF	2	0.1539	0.6768	546.938	962.07	<.0001
3	GarageArea	3	0.0481	0.7248	167.516	352.86	<.0001
4	MasVnrArea	4	0.0135	0.7383	62.3794	104.18	<.0001
5	h_Style_2	5	0.0059	0.7443	17.3748	46.74	<.0001
6	h_Style_6	6	0.0008	0.7451	12.9788	6.38	0.0116
7	h_Style_5	7	0.0008	0.7459	8.2599	6.72	0.0096
8	h_Style_7	8	0.0005	0.7464	6.2363	4.03	0.0449

- All continuous variables are statistically significant, and only one dummy variable shows as statistically significant.
- It seems as though the higher the Partial R-Square, C_p, and F Values are, the more significant the variable becomes

Conclusions:

In regression analysis, we would like our regression model to have significant variables and to produce a high R-squared value. A low *P* value / high R^2 combination indicates that changes in the predictors are related to changes in the response variable, and that our model explains a lot of the response variability. In some cases, it's possible that additional predictors can increase the true explanatory power of the model; however, in other cases, the data contains a fundamentally higher amount of unexplainable variability. On the brighter side, even when R-squared is low, low P values still indicate a real relationship between the significant predictors and the response variable.

From what we have gathered throughout the model building process, it seems that if we add more predictor variables, even the worst correlation to the response variable, the R-Squared results are better. This does not indicate a better *fit* because it also seems like you can put any variable into the multivariate model to get a better explanation of variability in the predictor variable.

This data set presented some challenges. First being that qualitative and categorical variables are much harder to try and use as predictor variables since *bias* goes in much of qualitative data. This data shows the possibility of many outliers that would need to be explored in greater detail. Some of the continuous variables found during the EDA process, showed the potential for mass amounts of null values, which are essentially useless and would need to be extracted from the data set during the model building process.

Ways to improve predictive accuracy would be to:

1. Add more data – It allows the "data to tell for itself," instead of relying on assumptions and weak correlations.
2. Clean up missing values/data points, and outliers – Missing values and outlier values in the data reduces the accuracy of a model or leads to a biased model; consequently, this leads to inaccurate predictions.
3. Cross validation and Transformations

Code:

```
libname mydata '/scs/wtm926/' access=readonly;

data temp1;
 set mydata.ames_housing_data;

data part1;
 set temp1;
  keep SalePrice GrLivArea GarageArea TotalBsmtSF FirstFlrSF MasVnrArea HouseStyle h_Style;
        if HouseStyle='1Story' then h_Style=1;
        if HouseStyle='1.5Fin' then h_Style=2;
        if HouseStyle='1.5Unf' then h_Style=3;
        if HouseStyle='2Story' then h_Style=4;
        if HouseStyle='2.5Fin' then h_Style=5;
        if HouseStyle='2.5Unf' then h_Style=6;
        if HouseStyle='SFoyer' then h_Style=7;
        if HouseStyle='SLvl' then h_Style=8;

proc reg data=part1;
        model SalePrice = h_Style;
run;

proc means data=part1;
 class HouseStyle;
 var SalePrice;
run;

proc freq data=part1;
        tables HouseStyle h_Style;
run;

data part2;
        set temp1;
                keep SalePrice GrLivArea GarageArea TotalBsmtSF FirstFlrSF MasVnrArea HouseStyle
h_Style_1 h_Style_2 h_Style_3 h_Style_4 h_Style_5 h_Style_6 h_Style_7 h_Style_8;
                        if HouseStyle in ('1Story' '1.5Fin' '1.5Unf' '2Story' '2.5Fin' '2.5Unf' 'SFoyer' 'SLvl')
then do;
                                h_Style_1= (HouseStyle eq '1Story');
                                h_Style_2= (HouseStyle eq '1.5Fin');
                                h_Style_3= (HouseStyle eq '1.5Unf');
                                h_Style_4= (HouseStyle eq '2Story');
                                h_Style_5= (HouseStyle eq '2.5Fin');
                                h_Style_6= (HouseStyle eq '2.5Unf');
                                h_Style_7= (HouseStyle eq 'SFoyer');
                                h_Style_8= (HouseStyle eq 'SLvl');
                        end;
```

```
Proc freq data=part2;
	tables HouseStyle h_Style_1 h_Style_2 h_Style_3 h_Style_4 h_Style_5 h_Style_6 h_Style_7
h_Style_8;
run;

proc reg data=part2;
	model SalePrice = h_Style_1 h_Style_2 h_Style_3 h_Style_4 h_Style_5 h_Style_6 h_Style_7
h_Style_8;
run;

data part3;
  set temp1;
   keep SalePrice GrLivArea GarageArea TotalBsmtSF FirstFlrSF MasVnrArea GarageType g_Type;
	if GarageType='2Types' then g_Type=1;
	if GarageType='Attchd' then g_Type=2;
	if GarageType='Basment' then g_Type=3;
	if GarageType='BuiltIn' then g_Type=4;
	if GarageType='CarPort' then g_Type=5;
	if GarageType='Detchd' then g_Type=6;
	if GarageType='NA' then g_Type=7;

proc reg data=part3;
	model SalePrice = g_Type;
run;

proc means data=part3;
  class GarageType;
  var SalePrice;
run;

proc freq data=part3;
	tables GarageType g_Type;
run;

data part4;
	set temp1;
		keep SalePrice GrLivArea GarageArea TotalBsmtSF FirstFlrSF MasVnrArea GarageType
g_Type_1 g_Type_2 g_Type_3 g_Type_4 g_Type_5 g_Type_6 g_Type_7;
				if GarageType in ('2Types' 'Attchd' 'Basment' 'BuiltIn' 'CarPort' 'Detchd' 'NA')
then do;
				g_Type_1 = (GarageType eq '2Types');
				g_Type_2 = (GarageType eq 'Attchd');
				g_Type_3 = (GarageType eq 'Basment');
				g_Type_4 = (GarageType eq 'BuiltIn');
				g_Type_5 = (GarageType eq 'CarPort');
				g_Type_6 = (GarageType eq 'Detchd');
				g_Type_7 = (GarageType eq 'NA');
```

```
                end;

Proc freq data=part4;
        tables GarageType g_Type_1 g_Type_2 g_Type_3 g_Type_4 g_Type_5 g_Type_6 g_Type_7;
run;

proc reg data=part4;
        model SalePrice = g_Type_1 g_Type_2 g_Type_3 g_Type_4 g_Type_5 g_Type_6 g_Type_7;
run;

proc reg data=part2 outest=reg_FW;
        model saleprice = GrLivArea GarageArea TotalBsmtSF FirstFlrSF MasVnrArea h_Style_1
h_Style_2 h_Style_3 h_Style_4 h_Style_5 h_Style_6 h_Style_7 h_Style_8/
selection=forward adjrsq aic bic cp best=5;
proc print data=reg_FW;
run;

proc reg data=part4 outest=reg_FW2;
        model saleprice = GrLivArea GarageArea TotalBsmtSF FirstFlrSF MasVnrArea g_Type_1 g_Type_2
g_Type_3 g_Type_4 g_Type_5 g_Type_6 g_Type_7/
selection=forward adjrsq aic bic cp best=5;
proc print data=reg_FW2;
run;

proc reg data=part2 outest=reg_BW;
        model saleprice = GrLivArea GarageArea TotalBsmtSF FirstFlrSF MasVnrArea h_Style_1
h_Style_2 h_Style_3 h_Style_4 h_Style_5 h_Style_6 h_Style_7 h_Style_8/
selection=backward adjrsq aic bic cp best=5;
proc print data=reg_BW;
run;

proc reg data=part4 outest=reg_BW2;
        model saleprice = GrLivArea GarageArea TotalBsmtSF FirstFlrSF MasVnrArea g_Type_1 g_Type_2
g_Type_3 g_Type_4 g_Type_5 g_Type_6 g_Type_7/
selection=backward adjrsq aic bic cp best=5;
proc print data=reg_BW2;
run;

proc reg data=part2 outest=reg_SW;
        model saleprice = GrLivArea GarageArea TotalBsmtSF FirstFlrSF MasVnrArea h_Style_1
h_Style_2 h_Style_3 h_Style_4 h_Style_5 h_Style_6 h_Style_7 h_Style_8/
selection=stepwise adjrsq aic bic cp best=5;
proc print data=reg_SW;
run;

proc reg data=part4 outest=reg_SW2;
        model saleprice = GrLivArea GarageArea TotalBsmtSF FirstFlrSF MasVnrArea g_Type_1 g_Type_2
g_Type_3 g_Type_4 g_Type_5 g_Type_6 g_Type_7/
```

```
selection=stepwise adjrsq aic bic cp best=5;
proc print data=reg_SW2;
run;

proc reg data=part2 outest=reg_ADJ;
	model saleprice = GrLivArea GarageArea TotalBsmtSF FirstFlrSF MasVnrArea h_Style_1
h_Style_2 h_Style_3 h_Style_4 h_Style_5 h_Style_6 h_Style_7 h_Style_8/
selection=adjrsq aic bic cp best=5;
proc print data=reg_ADJ;
run;

proc reg data=part4 outest=reg_ADJ2;
	model saleprice = GrLivArea GarageArea TotalBsmtSF FirstFlrSF MasVnrArea g_Type_1 g_Type_2
g_Type_3 g_Type_4 g_Type_5 g_Type_6 g_Type_7/
selection=adjrsq aic bic cp best=5;
proc print data=reg_ADJ2;
run;

proc reg data=part2 outest=reg_CP;
	model saleprice = GrLivArea GarageArea TotalBsmtSF FirstFlrSF MasVnrArea h_Style_1
h_Style_2 h_Style_3 h_Style_4 h_Style_5 h_Style_6 h_Style_7 h_Style_8/
selection=cp adjrsq aic bic cp best=5;
proc print data=reg_CP;
run;

proc reg data=part4 outest=reg_CP2;
	model saleprice = GrLivArea GarageArea TotalBsmtSF FirstFlrSF MasVnrArea g_Type_1 g_Type_2
g_Type_3 g_Type_4 g_Type_5 g_Type_6 g_Type_7/
selection=cp adjrsq aic bic cp best=5;
proc print data=reg_CP2;
run;

proc reg data=part2 outest=reg_AIC;
	model saleprice = GrLivArea GarageArea TotalBsmtSF FirstFlrSF MasVnrArea h_Style_1
h_Style_2 h_Style_3 h_Style_4 h_Style_5 h_Style_6 h_Style_7 h_Style_8/
selection=aic adjrsq aic bic cp best=5;
proc print data=reg_AIC;
run;

proc reg data=part4 outest=reg_AIC2;
	model saleprice = GrLivArea GarageArea TotalBsmtSF FirstFlrSF MasVnrArea g_Type_1 g_Type_2
g_Type_3 g_Type_4 g_Type_5 g_Type_6 g_Type_7/
selection=aic adjrsq aic bic cp best=5;
proc print data=reg_AIC2;
run;

ods graphics on;
```

```
proc reg data=part2;
  model SalePrice = GrLivArea GarageArea TotalBsmtSF MasVnrArea h_Style_1 h_Style_2 h_Style_3
h_Style_4 h_Style_5 h_Style_6 h_Style_7 h_Style_8;

proc reg data=part2;
  model SalePrice = GrLivArea GarageArea TotalBsmtSF MasVnrArea;

proc reg data=part2;
  model SalePrice = GrLivArea GarageArea TotalBsmtSF;
run;

proc reg data=part4;
  model SalePrice = GrLivArea GarageArea TotalBsmtSF MasVnrArea g_Type_1 g_Type_2 g_Type_3
g_Type_4 g_Type_5 g_Type_6 g_Type_7;

proc reg data=part4;
  model SalePrice = GrLivArea GarageArea TotalBsmtSF MasVnrArea;

proc reg data=part4;
  model SalePrice = GrLivArea GarageArea TotalBsmtSF;
run;

ods graphics off;

data part8;
        set part2;

                u = uniform(123);

                if (u < 0.70) then train = 1;
                else train = 0;

                if (train=1) then train_response=SalePrice;
                else train_response=.;
        run;


proc reg data=part8 outest=ADJ_train;
        model train_response = GrLivArea GarageArea TotalBsmtSF FirstFlrSF MasVnrArea h_Style_1
h_Style_2 h_Style_3 h_Style_4 h_Style_5 h_Style_6 h_Style_7 h_Style_8/
selection=adjrsq aic bic cp best=5;

proc print data=ADJ_train;
run;

proc reg data=part8 outest=cp_train;
        model train_response = GrLivArea GarageArea TotalBsmtSF FirstFlrSF MasVnrArea h_Style_1
h_Style_2 h_Style_3 h_Style_4 h_Style_5 h_Style_6 h_Style_7 h_Style_8/
```

```
selection=cp adjrsq aic bic cp best=5;

proc print data=cp_train;
run;

proc reg data=part8 outest=aic_train;
        model train_response = GrLivArea GarageArea TotalBsmtSF FirstFlrSF MasVnrArea h_Style_1
h_Style_2 h_Style_3 h_Style_4 h_Style_5 h_Style_6 h_Style_7 h_Style_8/
selection=aic adjrsq aic bic cp best=5;

proc print data=aic_train;
run;

proc reg data=part8 outest=forward_train;
        model train_response = GrLivArea GarageArea TotalBsmtSF FirstFlrSF MasVnrArea h_Style_1
h_Style_2 h_Style_3 h_Style_4 h_Style_5 h_Style_6 h_Style_7 h_Style_8/
selection=forward adjrsq aic bic cp best=5;

proc print data=forward_train;
run;

proc reg data=part8 outest=backward_train;
        model train_response = GrLivArea GarageArea TotalBsmtSF FirstFlrSF MasVnrArea h_Style_1
h_Style_2 h_Style_3 h_Style_4 h_Style_5 h_Style_6 h_Style_7 h_Style_8/
selection=backward adjrsq aic bic cp best=5;

proc print data=backward_train;
run;

proc reg data=part8 outest=stepwise_train;
        model train_response = GrLivArea GarageArea TotalBsmtSF FirstFlrSF MasVnrArea h_Style_1
h_Style_2 h_Style_3 h_Style_4 h_Style_5 h_Style_6 h_Style_7 h_Style_8/
selection=stepwise adjrsq aic bic cp best=5;

proc print data=stepwise_train;
run;

/* proc reg data=part8;
        model train_response = GrLivArea GarageArea TotalBsmtSF MasVnrArea/
        selection=forward;
output out=part9 predicted=yhat;

proc print data=part9(obs=5);
run;

proc reg data=part8;
        model SalePrice = GrLivArea GarageArea TotalBsmtSF MasVnrArea/
        selection=forward;
```

```
output out=part9 predicted=yhat;

proc print data=part9(obs=5);
run; */

proc reg data=part8;
	model SalePrice = GrLivArea GarageArea TotalBsmtSF MasVnrArea;
	ouput out=non_dummy predicted=yhat;

proc print data=non_dummy(obs=5);
run;

proc reg data=part8;
	model train_response= GrLivArea GarageArea TotalBsmtSF MasVnrArea;
	ouput out=non_dummy predicted=yhat;

proc print data=non_dummy(obs=5);
run;

proc reg data=part8;
	model SalePrice = GrLivArea GarageArea TotalBsmtSF MasVnrArea h_Style_1 h_Style_2
h_Style_3 h_Style_4 h_Style_5 h_Style_6 h_Style_7 h_Style_8;
	ouput out=dummy_v predicted=yhat;

proc print data=dummy_v(obs=5);
run;

proc reg data=part8;
	model train_response= GrLivArea GarageArea TotalBsmtSF MasVnrArea h_Style_1 h_Style_2
h_Style_3 h_Style_4 h_Style_5 h_Style_6 h_Style_7 h_Style_8;
	ouput out=dummy_v predicted=yhat;

proc print data=dummy_v(obs=5);
run;

data part9b;
	set part9;
		mae = abs(yhat - train_response);

proc print data=part9b(obs=5);

proc means data=part9b;
	var mae;
	title 'MAE calculation';
run;

proc reg data=part8;
```

```
	model train_response = GrLivArea GarageArea TotalBsmtSF MasVnrArea h_Style_1 h_Style_2
h_Style_3 h_Style_4 h_Style_5 h_Style_6 h_Style_7 h_Style_8/
	selection=forward;
	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;
```