Assignment #5
Nitin Gaonkar

Introduction:
The purpose of this assignment is to build regression models for the home sale price, in this assignment I will be using dummy coding of categorical Variables and then build regression models, In the later part of the assignment I will be working on validation frameworks and validations.

Results:
PART A: Dummy Coding of Categorical Variables

I have selected fireplaces categorical variable for my analysis. Below are the means for each category:

The MEANS Procedure

Fireplaces=0

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
1422	141195.77	44546.56	13100.00	360000.00

Fireplaces=1

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
1274	213556.00	81459.23	12789.00	625000.00

Fireplaces=2

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
221	242316.16	113124.75	80400.00	755000.00

Fireplaces=3

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
12	255820.83	96637.06	160000.00	462000.00

Fireplaces=4

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
1	260000.00	.	260000.00	260000.00

a. Model saleprice vs fireplaces
Model in equation form:

saleprice= 145729 + 58512 * Fireplaces

Here for each unit increase in Fireplace will increase sale price by $ 58512

Model salesprice vs. Fireplaces

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Number of Observations Read	2930
Number of Observations Used	2930

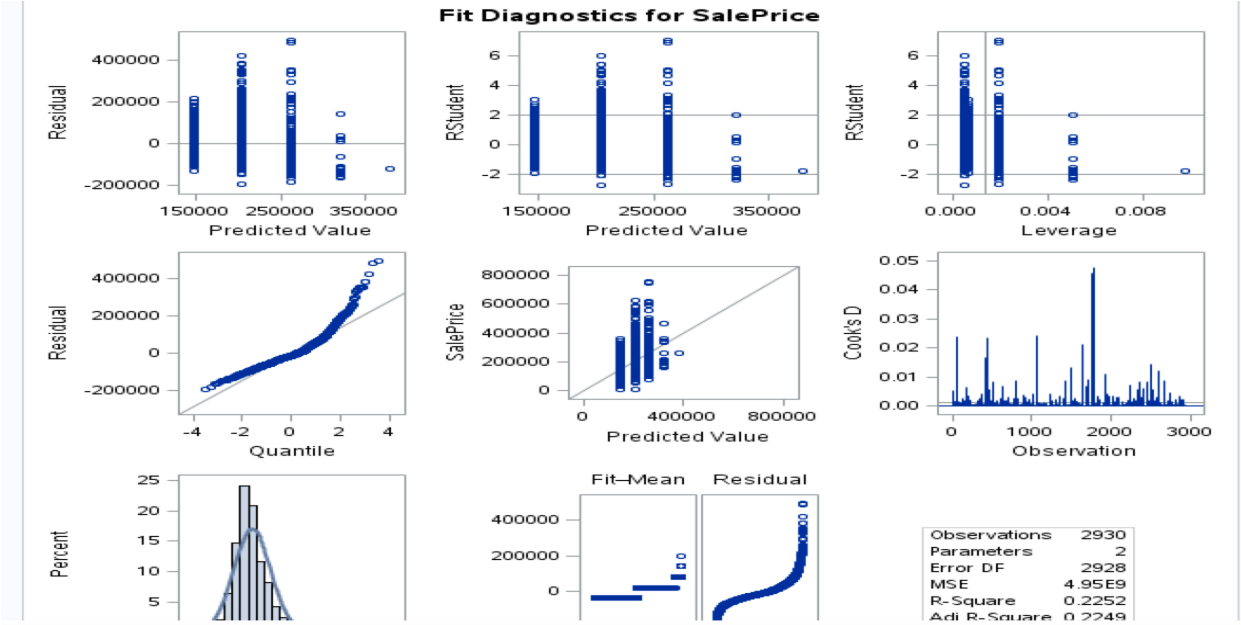
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	4.20966E12	4.20966E12	851.07	<.0001
Error	2928	1.448288E13	4946337816		
Corrected Total	2929	1.869254E13			

Root MSE	70330	R-Square	0.2252
Dependent Mean	180796	Adj R-Sq	0.2249
Coeff Var	38.90030		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	145729	1770.04460	82.33	<.0001
Fireplaces	1	58512	2005.67334	29.17	<.0001



2. Here I have used fireplaces as the categorical variable, I have used the SAS code to create the dummy variables, below is the output of the proc freq:

Fireplaces	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	1422	48.53	1422	48.53
1	1274	43.48	2696	92.01
2	221	7.54	2917	99.56
3	12	0.41	2929	99.97
4	1	0.03	2930	100.00

fireplace_0	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	1508	51.47	1508	51.47
1	1422	48.53	2930	100.00

fireplace_1	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	1656	56.52	1656	56.52
1	1274	43.48	2930	100.00

fireplace_2	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	2709	92.46	2709	92.46
1	221	7.54	2930	100.00

fireplace_3	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	2918	99.59	2918	99.59
1	12	0.41	2930	100.00

Below I ran the multi regression on the dummy variables vs saleprice

Model in equation form:

Saleprice= 260000 -118804 * fireplace_0 – 46444 * fireplace_1 -17684 * fire_place_2 – 4179.1667 fireplace_3

If we don't have a fireplace, then fireplace_0 is 1 rest of the betas are zero so the equation would be

Saleprice= 260000 -118804 * fireplace_0

If we have 1 fireplace:

Saleprice= 260000 - 46444 * fireplace_1 -17684

Respectively for other values of fireplaces.

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Number of Observations Read	2930
Number of Observations Used	2930

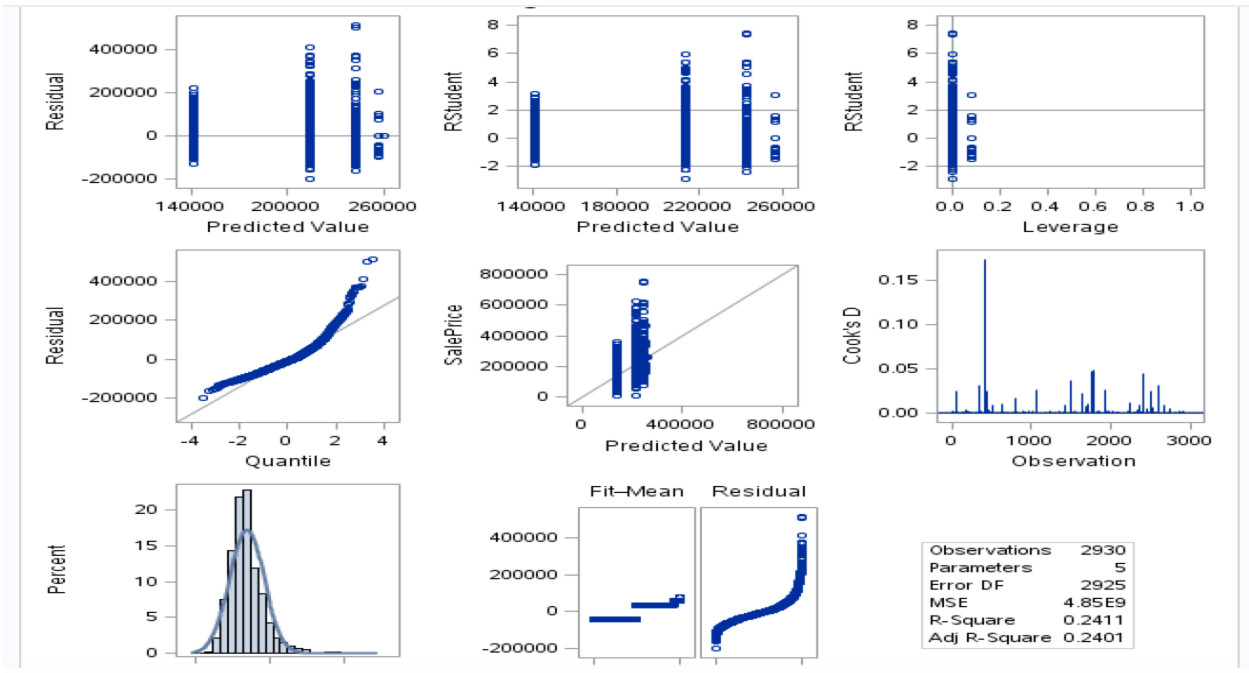
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	4.507472E12	1.126868E12	232.36	<.0001
Error	2925	1.418507E13	4849594929		
Corrected Total	2929	1.869254E13			

Root MSE	69639	R-Square	0.2411
Dependent Mean	180796	Adj R-Sq	0.2401
Coeff Var	38.51800		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	260000	69639	3.73	0.0002
fireplace_0	1	-118804	69664	-1.71	0.0882
fireplace_1	1	-46444	69666	-0.67	0.5050
fireplace_2	1	-17684	69796	-0.25	0.8000
fireplace_3	1	-4179.16667	72483	-0.06	0.9540



	r-squared value	fit-mean/residual	cook's D	qq plot	scatter plot	Histogram
Model1 (saleprice vs fireplaces)	0.2252	Residual is greater than the fit mean	We can observe lot of outliers	points are off the line towards end of plot	Lot of points are away from the line and not distributed in a st line	Lot of points are not covered in the SAS curve
Model 2 (saleprice vs dummy variables	0.2411	residual is greater than the fit mean	Lesser outliers compared to model 1	points are off the line towards end of plot	Not much diff between model 1 and model 2	Slightly better than the model1

After observing the metrics of both the models I don't see a huge difference in both of the models, except for the slight increase in R value, also the increase is because of the more predictor variables, the adjusted R value is increased as well. By looking at these values and the ODS output I think the model 2 is slightly better than the model1.

3. Hypothesis test:

F test:

H0= fireplace_0 = fireplace_1=fireplace_2=fireplace_3=0

HA= one or more co-efficient is non zero.

Alpha value =0.01

F value looks significant and also the p value looks significant we can reject the null hypotheses.

T- Test:

H0= H0= fireplace_0 = fireplace_1=fireplace_2=fireplace_3=0

HA= one or more co-efficient is non zero.

By checking the P values of the co-efficient and looks like they are not significant, we cannot reject the null hypothesis.

4. Here I have used extercond as the categorical variable, I have used the SAS code to create the dummy variables, below is the output mean.

The FREQ Procedure

ExterCond	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Ex	12	0.41	12	0.41
Fa	67	2.29	79	2.70
Gd	299	10.20	378	12.90
Po	3	0.10	381	13.00
TA	2549	87.00	2930	100.00

ExterCond_Ex	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	2918	99.59	2918	99.59
1	12	0.41	2930	100.00

ExterCond_Gd	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	2631	89.80	2631	89.80
1	299	10.20	2930	100.00

ExterCond_TA	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	381	13.00	381	13.00
1	2549	87.00	2930	100.00

ExterCond_Fa	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	2863	97.71	2863	97.71
1	67	2.29	2930	100.00

Below I ran the multi regression on the dummy variables vs saleprice

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Number of Observations Read	2930
Number of Observations Used	2930

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	5.051388E11	1.262847E11	20.31	<.0001
Error	2925	1.81874E13	6217913961		
Corrected Total	2929	1.869254E13			

Root MSE	78854	R-Square	0.0270
Dependent Mean	180796	Adj R-Sq	0.0257
Coeff Var	43.61475		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	82867	45526	1.82	0.0688
ExterCond_Ex	1	134833	50900	2.65	0.0081
ExterCond_Gd	1	82937	45754	1.81	0.0700
ExterCond_TA	1	101554	45553	2.23	0.0259
ExterCond_Fa	1	24697	46534	0.53	0.5957

Model in equation form: saleprice= 82867 + 134833* extercond_ex + 82937 * extercond_gd + 101554 * extercond_TA + 24697 * extercond_Fa

5.
I fitted the model in the below order:
Forward select method
adjusted square method
backward selection method
Stepwise

Mallow's cp
AIC

Below are the continuous and the dummy variables used:
fireplace_0 fireplace_1 fireplace_2 fireplace_3 ExterCond_Ex ExterCond_Gd ExterCond_TA ExterCond_Fa LotFrontage LotArea BsmtUnfSF GrLivArea MasVnrArea TotalBsmtSF FirstFlrSF GarageArea WoodDeckSF

Forward select method: in this selection procedure after analysis of the summary table I could see that the below variables were significant in this model:
BsmtUnfSF, GrLivArea, MasVnrArea ,TotalBsmtSF, FirstFlrSF, GarageArea ,WoodDeckSF, fireplace_2 fireplace_3.

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	0.4992	0.4992	2050.50	2411.66	<.0001
2	TotalBsmtSF	2	0.1270	0.6263	919.195	821.89	<.0001
3	GarageArea	3	0.0572	0.6835	410.529	437.14	<.0001
4	MasVnrArea	4	0.0171	0.7006	259.653	138.30	<.0001
5	fireplace_0	5	0.0130	0.7137	145.521	109.79	<.0001
6	BsmtUnfSF	6	0.0052	0.7188	101.560	44.23	<.0001
7	WoodDeckSF	7	0.0037	0.7225	70.5091	32.22	<.0001
8	ExterCond_Fa	8	0.0035	0.7260	41.5612	30.54	<.0001
9	fireplace_3	9	0.0025	0.7285	21.4578	22.00	<.0001
10	fireplace_2	10	0.0005	0.7289	19.3924	4.05	0.0442
11	LotFrontage	11	0.0003	0.7293	18.4518	2.93	0.0869
12	LotArea	12	0.0003	0.7295	17.9683	2.48	0.1155
13	ExterCond_Ex	13	0.0002	0.7298	17.8574	2.11	0.1467
14	ExterCond_TA	14	0.0002	0.7300	17.6427	2.21	0.1370
15	ExterCond_Gd	15	0.0005	0.7305	15.5188	4.12	0.0424
16	fireplace_1	16	0.0001	0.7306	16.7131	0.81	0.3695
17	FirstFlrSF	17	0.0001	0.7306	18.0000	0.71	0.3985

Backward selection method
In this procedures we can see that below 12 variables are significant and below 5 variables are not significant and can be eliminated from the model

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-46641	6124.66811	1.093644E11	57.99	<.0001
fireplace_1	17659	2056.24719	1.390827E11	73.75	<.0001
fireplace_2	25685	3978.38701	78603553524	41.68	<.0001
fireplace_3	-49174	14906	20523740019	10.88	0.0010
ExterCond_Ex	52105	14208	25361861526	13.45	0.0003
ExterCond_Gd	28768	6140.69671	41389547239	21.95	<.0001
ExterCond_TA	32763	5567.54658	65302518041	34.63	<.0001
BsmtUnfSF	-12.95797	2.29075	60342149971	32.00	<.0001
GrLivArea	58.38643	2.29452	1.221071E12	647.50	<.0001
MasVnrArea	60.45941	5.75377	2.082214E11	110.41	<.0001
TotalBsmtSF	50.92517	2.69059	6.755717E11	358.24	<.0001
GarageArea	91.55235	5.00155	6.318755E11	335.07	<.0001
WoodDeckSF	44.84070	7.83085	61834356672	32.79	<.0001

Bounds on condition number: 4.5283, 272.82

All variables left in the model are significant at the 0.1000 level.

Summary of Backward Elimination							
Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	fireplace_0	16	0.0000	0.7306	16.2195	0.22	0.6394
2	FirstFlrSF	15	0.0001	0.7305	14.9587	0.74	0.3899
3	ExterCond_Fa	14	0.0001	0.7304	13.8003	0.84	0.3589
4	LotArea	13	0.0003	0.7302	14.1704	2.37	0.1237
5	LotFrontage	12	0.0003	0.7299	14.7858	2.62	0.1060

AIC procedure:

Here we can see that the lowest value for the AIC and BIC is 51718.7820 and 51720.948 respectively

And we can see the variables used by model from the below screen shot.

Number in Model	Adjusted R-Square	R-Square	AIC	BIC	Variables in Model
14	0.7289	0.7304	51718.7820	51720.9840	fireplace_1 fireplace_2 fireplace_3 ExterCond_Ex ExterCond_Gd ExterCond_TA LotFrontage LotArea BsmtUnfSF GrLivArea MasVnrArea TotalBsmtSF GarageArea WoodDeckSF

Adjusted R square value

Here we can see that the highest value for the r square is 0.7306

And we can see the variables used by model from the below screen shot.

17	0.7287	0.7306	fireplace_0 fireplace_1 fireplace_2 fireplace_3 ExterCond_Ex ExterCond_Gd ExterCond_TA ExterCond_Fa LotFrontage LotArea BsmtUnfSF GrLivArea MasVnrArea TotalBsmtSF FirstFlrSF GarageArea WoodDeckSF

Mallow's cp

Here we can see that cp value is 15 and there are 15 parameters in the model, cp=p, below are the variables which best fits the model.

15	15.0270	0.7305	51720.0033	51722.2292	fireplace_1 fireplace_2 fireplace_3 ExterCond_Ex ExterCond_Gd ExterCond_TA LotFrontage LotArea BsmtUnfSF GrLivArea MasVnrArea TotalBsmtSF FirstFlrSF GarageArea WoodDeckSF
15	15.0298	0.7305	51720.0062	51722.2320	fireplace_0 fireplace_1 fireplace_2 ExterCond_Ex ExterCond_Gd ExterCond_TA ExterCond_Fa LotFrontage LotArea BsmtUnfSF GrLivArea MasVnrArea TotalBsmtSF GarageArea WoodDeckSF

Stepwise selection:

Here we can see that the lowest value for the AIC and BIC is 51718.7820 and 51720.948 respectively And we can see the variables used by model from the below screen shot

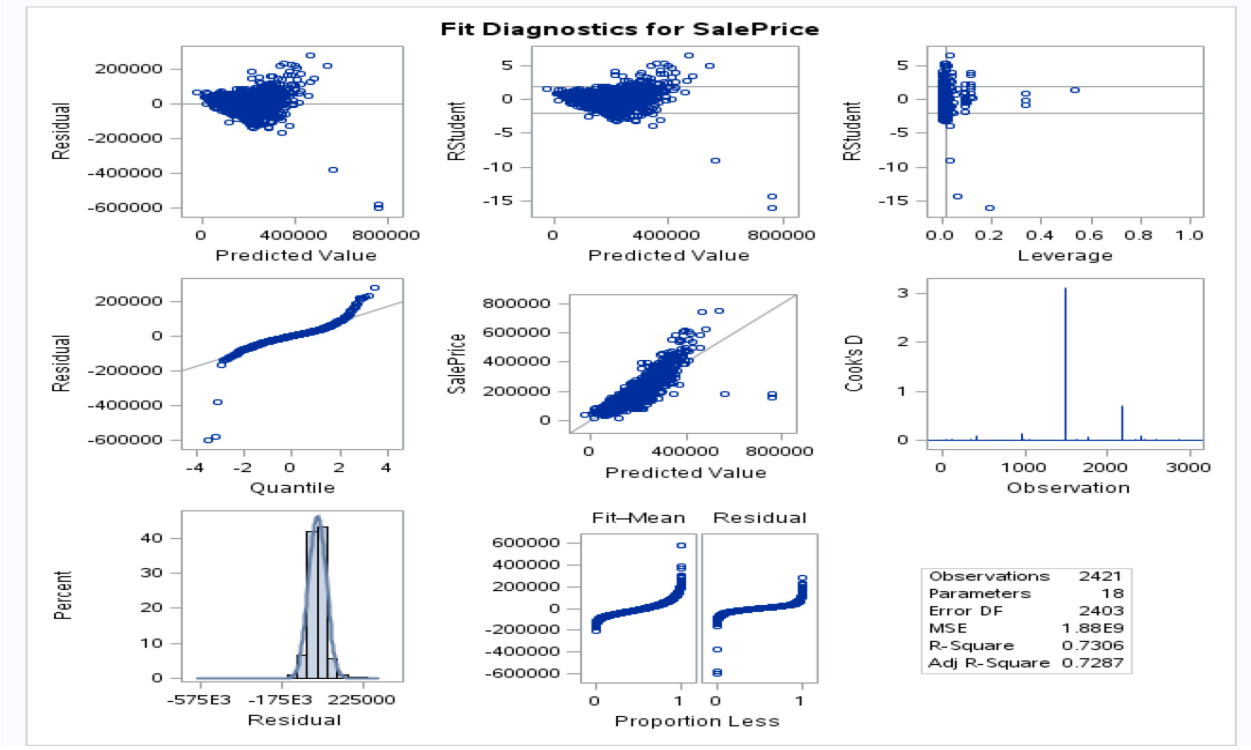
Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	GrLivArea		1	0.4992	0.4992	2050.50	2411.66	<.0001
2	TotalBsmtSF		2	0.1270	0.6263	919.195	821.89	<.0001
3	GarageArea		3	0.0572	0.6835	410.529	437.14	<.0001
4	MasVnrArea		4	0.0171	0.7006	259.653	138.30	<.0001
5	fireplace_0		5	0.0130	0.7137	145.521	109.79	<.0001
6	BsmtUnfSF		6	0.0052	0.7188	101.560	44.23	<.0001
7	WoodDeckSF		7	0.0037	0.7225	70.5091	32.22	<.0001
8	ExterCond_Fa		8	0.0035	0.7260	41.5612	30.54	<.0001
9	fireplace_3		9	0.0025	0.7285	21.4578	22.00	<.0001
10	fireplace_2		10	0.0005	0.7289	19.3924	4.05	0.0442
11	LotFrontage		11	0.0003	0.7293	18.4518	2.93	0.0869
12	LotArea		12	0.0003	0.7295	17.9683	2.48	0.1155
13	ExterCond_Ex		13	0.0002	0.7298	17.8574	2.11	0.1467
14	ExterCond_TA		14	0.0002	0.7300	17.6427	2.21	0.1370
15	ExterCond_Gd		15	0.0005	0.7305	15.5188	4.12	0.0424
16		ExterCond_Fa	14	0.0001	0.7304	14.3394	0.82	0.3650

6. Yes, in few procedures I could see that the dummy coded models variables were selected. I choose the forward selection procedure for my analysis, continuous variables lot Area, lotfrontage, firstflrsf have been removed from the model as they were not statistically significant. I could see that the R squared values decreased a bit since we removed few variables from the model, adjusted r square value did not change much also by observing the ODS output of both model I can see that there is no much difference in the model.

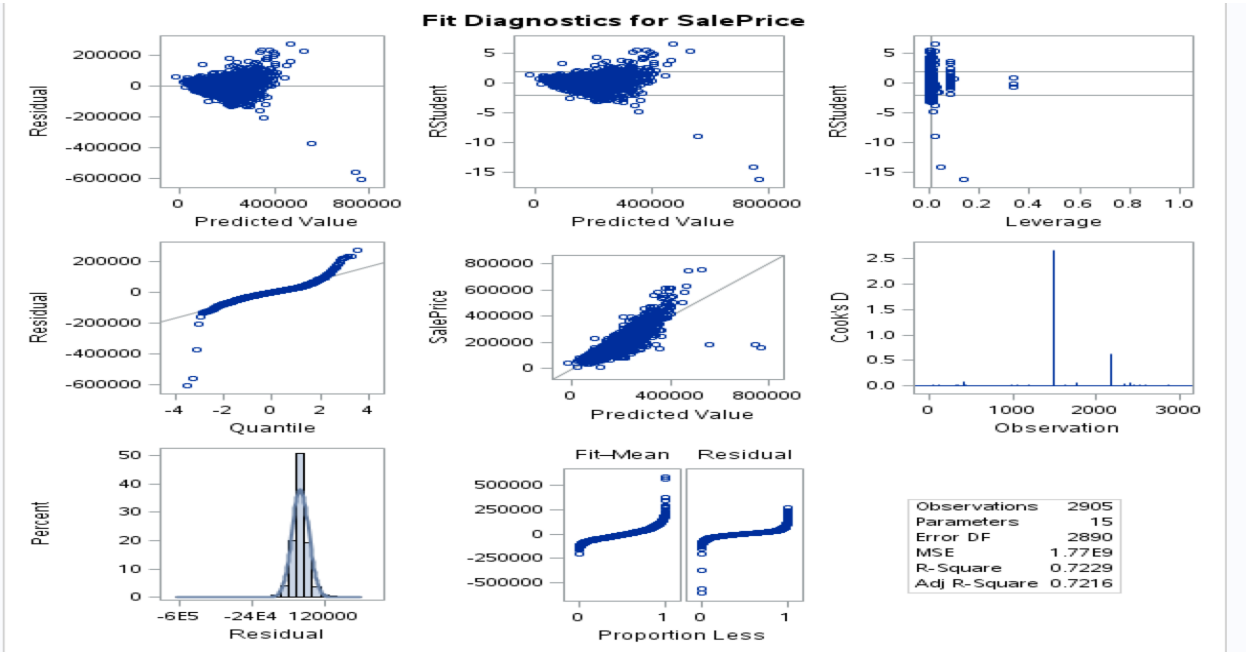
All variables have been entered into the model.

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	0.5006	0.5006	2307.67	2910.08	<.0001
2	TotalBsmtSF	2	0.1258	0.6264	997.355	977.41	<.0001
3	GarageArea	3	0.0532	0.6796	444.475	481.74	<.0001
4	MasVnrArea	4	0.0151	0.6947	289.409	143.04	<.0001
5	fireplace_0	5	0.0123	0.7069	163.555	121.26	<.0001
6	BsmtUnfSF	6	0.0060	0.7129	103.060	60.49	<.0001
7	WoodDeckSF	7	0.0035	0.7164	69.0532	35.26	<.0001
8	ExterCond_Fa	8	0.0034	0.7198	35.8259	34.90	<.0001
9	fireplace_3	9	0.0017	0.7215	19.8594	17.91	<.0001
10	ExterCond_Gd	10	0.0003	0.7218	18.5260	3.32	0.0683
11	ExterCond_Ex	11	0.0003	0.7221	17.4603	3.06	0.0803
12	ExterCond_TA	12	0.0005	0.7226	14.2802	5.18	0.0229
13	fireplace_2	13	0.0002	0.7228	13.8085	2.47	0.1160
14	fireplace_1	14	0.0001	0.7229	15.0000	0.81	0.3687

ODS output of the model before removing the non significant variables.



ODS output after removing the non significant variables.



PART C: Validation Framework

Used the SAS code to split the data 70/30 and for training and test respectively.
Below is the summary table for the forward selection procedure.

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	0.5277	0.5277	1909.29	1880.68	<.0001
2	TotalBsmtSF	2	0.1614	0.6891	684.511	873.04	<.0001
3	GarageArea	3	0.0489	0.7380	314.987	313.52	<.0001
4	BsmtUnfSF	4	0.0148	0.7528	204.578	100.47	<.0001
5	MasVnrArea	5	0.0109	0.7637	123.566	77.58	<.0001
6	fireplace_0	6	0.0066	0.7703	75.1212	48.48	<.0001
7	WoodDeckSF	7	0.0039	0.7743	47.1997	29.24	<.0001
8	ExterCond_Fa	8	0.0035	0.7777	22.8134	26.17	<.0001
9	LotArea	9	0.0008	0.7785	18.9715	5.81	0.0160
10	fireplace_1	10	0.0005	0.7790	17.0197	3.94	0.0474
11	ExterCond_Ex	11	0.0002	0.7792	17.2208	1.79	0.1807
12	LotFrontage	12	0.0002	0.7794	17.7902	1.43	0.2325
13	fireplace_3	13	0.0001	0.7796	18.8204	0.97	0.3256
14	fireplace_2	14	0.0004	0.7799	18.0801	2.74	0.0983
15	FirstFlrSF	15	0.0001	0.7800	19.4690	0.61	0.4350

Here we can see that most of the continuous variables are statistically significant and a R square value of 0.78 for the model. In the ods out put the residual plot look pretty good, in the cook's D there are few outliers,fit mean is greater than the residual which is good, in the histogram the SAS curve covers all the points.

Backward selection:
In this process we can see that the non significant variables have been eliminated, the r square value of the model is 0.7802

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-65060	6644.55898	1.448977E11	95.87	<.0001
fireplace_1	13176	2224.29799	53029090018	35.09	<.0001
fireplace_2	21242	4391.58447	35357986886	23.40	<.0001
fireplace_3	37721	17696	6867380928	4.54	0.0332
ExterCond_Ex	53068	15844	16956134836	11.22	0.0008
ExterCond_Gd	32549	6625.40431	36476323727	24.14	<.0001
ExterCond_TA	33536	6014.66580	46984987288	31.09	<.0001
LotArea	0.30660	0.14625	6642020476	4.39	0.0362
BsmtUnfSF	-17.66666	2.46815	77432764338	51.23	<.0001
GrLivArea	64.27640	2.49993	9.990911E11	661.07	<.0001
MasVnrArea	51.14987	6.01462	1.093028E11	72.32	<.0001
TotalBsmtSF	64.67618	3.06262	6.740038E11	445.97	<.0001
GarageArea	85.58568	5.43958	3.741369E11	247.55	<.0001
WoodDeckSF	45.76841	8.24254	46598165780	30.83	<.0001

Bounds on condition number: 4.5253, 310.01

All variables left in the model are significant at the 0.1000 level.

Summary of Backward Elimination							
Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	FirstFlrSF	16	0.0001	0.7807	16.4678	0.47	0.4941
2	ExterCond_Fa	15	0.0001	0.7805	15.3837	0.92	0.3386
3	fireplace_0	14	0.0002	0.7804	14.5689	1.19	0.2764
4	LotFrontage	13	0.0002	0.7802	13.9717	1.40	0.2364

Here we can see that most of the continuous variables are statistically significant and a R square value of 0.7802 for the model. In the ods out put the residual plot look pretty good, in the cook's D there are lot of outliers, fit mean is greater than the residual which is good, in the histogram the SAS curve covers all the points.

AIC procedure, from the summary table I have picked the best model with lowest value of AIC and BIC

We can see that the model with 15 variables is best model with lowest AIC and BIC.,

In the ods out put the residual plot look good, in the cook's D there are lot of outliers, fit mean is greater than the residual which is good, in the histogram the SAS curve covers all the points.

Number in Model	Adjusted R-Square	R-Square	AIC	BIC	Variables in Model
15	0.7786	0.7805	35629.9202	35632.2387	fireplace_0 fireplace_1 fireplace_2 fireplace_3 ExterCond_Ex ExterCond_Gd ExterCond_TA LotFrontage LotArea BsmtUnfSF GrLivArea MasVnrArea TotalBsmtSF GarageArea WoodDeckSF
16	0.7786	0.7807	35630.9949	35633.3522	fireplace_0 fireplace_1 fireplace_2 fireplace_3 ExterCond_Ex ExterCond_Gd ExterCond_TA ExterCond_Fa LotFrontage LotArea BsmtUnfSF GrLivArea MasVnrArea TotalBsmtSF GarageArea WoodDeckSF

Stepwise procedure:

R square value of this model is 0.7790, almost all the variables are statically significant

In the ods out put the residual plot look good, in the cook's D there are lot of outliers, fit mean is greater than the residual which is good, in the histogram the SAS curve covers all the points.

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	GrLivArea		1	0.5277	0.5277	1909.29	1880.68	<.0001
2	TotalBsmtSF		2	0.1614	0.6891	684.511	873.04	<.0001
3	GarageArea		3	0.0489	0.7380	314.987	313.52	<.0001
4	BsmtUnfSF		4	0.0148	0.7528	204.578	100.47	<.0001
5	MasVnrArea		5	0.0109	0.7637	123.566	77.58	<.0001
6	fireplace_0		6	0.0066	0.7703	75.1212	48.48	<.0001
7	WoodDeckSF		7	0.0039	0.7743	47.1997	29.24	<.0001
8	ExterCond_Fa		8	0.0035	0.7777	22.8134	26.17	<.0001
9	LotArea		9	0.0008	0.7785	18.9715	5.81	0.0160
10	fireplace_1		10	0.0005	0.7790	17.0197	3.94	0.0474

CP procedure:

Here we can see that cp value is 15 and there are 15 parameters in the model, cp=p, below are the variables which best fits the model.

15	15.3837	0.7805	fireplace_0 fireplace_1 fireplace_2 fireplace_3 ExterCond_Ex ExterCond_Gd ExterCond_TA LotFrontage LotArea BsmtUnfSF GrLivArea MasVnrArea TotalBsmtSF GarageArea WoodDeckSF

In the ods out put the residual plot look good, in the cook's D there are lot of outliers, fit mean is greater than the residual which is good, in the histogram the SAS curve covers all the points.

Adjusted R:

16	0.7785	0.7806	fireplace_0 fireplace_1 fireplace_2 fireplace_3 ExterCond_Ex ExterCond_Gd ExterCond_TA LotFrontage LotArea BsmtUnfSF GrLivArea MasVnrArea TotalBsmtSF FirstFlrSF GarageArea WoodDeckSF
15	0.7785	0.7805	fireplace_0 fireplace_1 fireplace_2 fireplace_3 ExterCond_Ex ExterCond_Gd ExterCond_TA LotArea BsmtUnfSF GrLivArea MasVnrArea TotalBsmtSF FirstFlrSF GarageArea WoodDeckSF
13	0.7785	0.7802	fireplace_1 fireplace_2 fireplace_3 ExterCond_Ex ExterCond_Gd ExterCond_TA LotArea BsmtUnfSF GrLivArea MasVnrArea TotalBsmtSF GarageArea WoodDeckSF
17	0.7785	0.7807	fireplace_0 fireplace_1 fireplace_2 fireplace_3 ExterCond_Ex ExterCond_Gd ExterCond_TA ExterCond_Fa LotFrontage LotArea BsmtUnfSF GrLivArea MasVnrArea TotalBsmtSF FirstFlrSF GarageArea WoodDeckSF

0.7807 is the R square value of the model with the 17 parameters.
In the ods out put the residual plot look good, in the cook's D there are lot of outliers, fit mean is greater than the residual which is good, in the histogram the SAS curve covers all the points.

9.Table comparing the metrics from assignment# and assignment 3

	Adjusted R square	Mean squared error	MAE
Model_AdjR2	0.7802	1.51E9	27409.98
Model_Mcp	0.7805	1.51E9	27209.65
Model_f	0.7802	1514496230	27472.15
Model_B	0.7802	1511329281	27270.32
Model_S	0.7790	1516789758	26974.26
Model_Aic	0.7805	1.51E9	27209.65

10. Training sample (forward method)

below are the prediction grading percentage table:

The FREQ Procedure

Prediction_Grade	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Grade 2	347	17.02	347	17.02
Grade 3	823	40.36	1170	57.38
Grade1	869	42.62	2039	100.00

below is the sample observations:

ExterCond_Gd	ExterCond_TA	ExterCond_Fa	u	train	train_response	yhat	Prediction_Grade	pct_diff
0	1	0	0.32091	1	105000	148020.60	Grade 3	0.40972
0	1	0	0.17839	1	172000	183723.35	Grade1	0.06816
0	1	0	0.35712	1	189900	196522.75	Grade1	0.03487
0	1	0	0.22111	1	195500	197229.03	Grade1	0.00884
0	1	0	0.39808	1	191500	160407.33	Grade 3	0.16236
0	1	0	0.12467	1	236500	243740.84	Grade1	0.03062
0	1	0	0.18769	1	189000	189886.95	Grade1	0.00469
1	0	0	0.43607	1	185000	171265.47	Grade1	0.07424
0	1	0	0.26370	1	171500	192531.84	Grade 2	0.12263
0	1	0	0.55486	1	538000	419415.24	Grade 3	0.22042

Sample of the frequency procedure for test data for forward selection method.

The FREQ Procedure

Prediction_Grade	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Grade 2	156	17.51	156	17.51
Grade 3	362	40.63	518	58.14
Grade1	373	41.86	891	100.00

ExterCond_Gd	ExterCond_TA	ExterCond_Fa	u	train	train_response	yhat	Prediction_Grade	pct_diff
0	1	0	0.75040	0	.	216472.43	Grade1	0.00685
0	1	0	0.90603	0	.	285034.31	Grade 3	0.16817
0	1	0	0.78644	0	.	179523.00	Grade 3	0.15914
0	1	0	0.77618	0	.	170464.44	Grade1	0.03090
0	1	0	0.96750	0	.	149188.36	Grade 3	0.17301
0	1	0	0.71393	0	.	208841.33	Grade1	0.01490
0	1	0	0.86134	0	.	307305.99	Grade 3	0.22089
0	1	0	0.86042	0	.	273629.63	Grade 3	0.30300
0	1	0	0.76996	0	.	181633.72	Grade1	0.06843
0	1	0	0.70216	0	.	183143.00	Grade 3	0.15212

Adjrsquare procedure:

Training data:

The FREQ Procedure

Prediction_Grade	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Grade 2	280	13.73	280	13.73
Grade 3	720	35.31	1000	49.04
Grade1	1039	50.96	2039	100.00

test data for adjsqr method:

The FREQ Procedure

Prediction_Grade	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Grade 2	116	13.02	116	13.02
Grade 3	314	35.24	430	48.26
Grade1	461	51.74	891	100.00

Backward procedure (training data):

The FREQ Procedure

Prediction_Grade	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Grade 2	353	17.31	353	17.31
Grade 3	821	40.26	1174	57.58
Grade1	865	42.42	2039	100.00

Backward procedure (test data):

The FREQ Procedure

Prediction_Grade	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Grade 2	154	17.28	154	17.28
Grade 3	362	40.63	516	57.91
Grade1	375	42.09	891	100.00

Cp procedure for training data:

The FREQ Procedure

Prediction_Grade	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Grade 2	337	16.53	337	16.53
Grade 3	838	41.10	1175	57.63
Grade1	864	42.37	2039	100.00

Cp procedure for test data:

The FREQ Procedure

Prediction_Grade	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Grade 2	146	16.39	146	16.39
Grade 3	369	41.41	515	57.80
Grade1	376	42.20	891	100.00

Stepwise method:

The FREQ Procedure

Prediction_Grade	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Grade 2	338	16.58	338	16.58
Grade 3	840	41.20	1178	57.77
Grade1	861	42.23	2039	100.00

Test data for stepwise method to get the accuracy

The FREQ Procedure

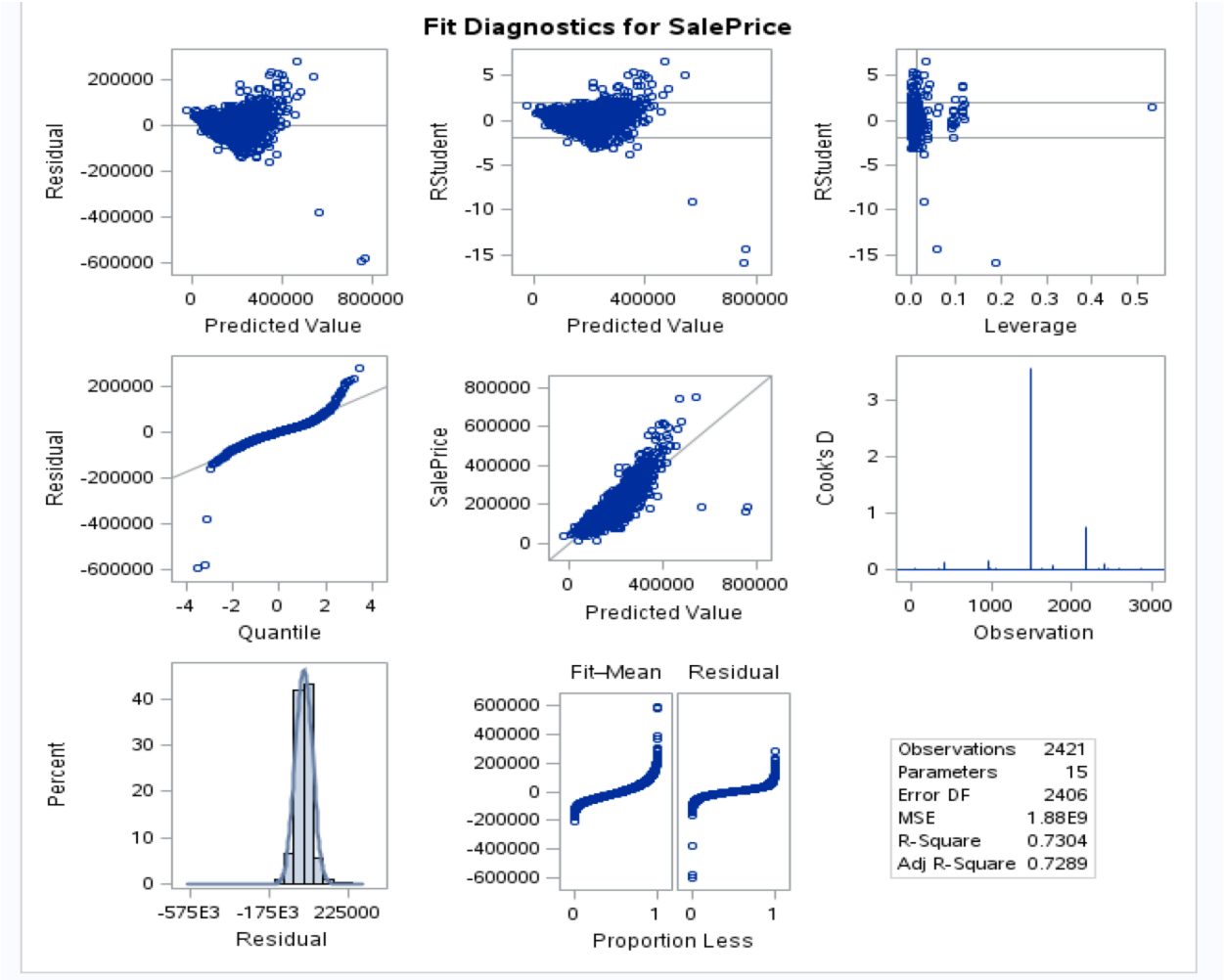
Prediction_Grade	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Grade 2	144	16.16	144	16.16
Grade 3	368	41.30	512	57.46
Grade1	379	42.54	891	100.00

After comparing all the models, I think the model with the ADJRSQ is the best fitted model, it has an accuracy of 51.74

After revisiting the issues for this model below were the variables which we used for the best fit:

fireplace_0 fireplace_1 fireplace_2 fireplace_3 ExterCond_Ex ExterCond_Gd ExterCond_TA ExterCond_Fa LotFrontage LotArea BsmtUnfSF GrLivArea MasVnrArea TotalBsmtSF FirstFlrSF GarageArea WoodDeckSF

And below is the ODS output of the model :



Conclusion:

Overall in this assignment I have tried multi regression by dummy coding and using the various procedures and the all the models were validated by the validation frame work.
In this data set there was few important predictor variables missing and also there were not enough observations to train the model and run the validation on test set. I think one way of improving the accuracy is by adding varied data that covers most of the characteristics of data.

Code:
Paste your code in at the end.

```
libname mydata "/scs/wtm926/" access=readonly;
proc datasets library=mydata;
```

```
data my_assign;
set mydata.ames_housing_data;

#####assignment 5######
data part1;
set my_assign;
keep Fireplaces saleprice;
run;

proc sort data=part1;
by Fireplaces;
run;


proc means data=part1;
by Fireplaces;
var saleprice;
run;

proc sort data=my_assign;
by Fireplaces;
run;
data part3;
set my_assign;
keep saleprice fireplaces fireplace_0 fireplace_1 fireplace_2 fireplace_3;
if fireplaces in ('0','1','2','3','4') then do;
fireplace_0 = (fireplaces eq '0');
fireplace_1 = (fireplaces eq '1');
fireplace_2 = (fireplaces eq '2');
fireplace_3 = (fireplaces eq '3');
end;

proc freq data =part3;
tables fireplaces fireplace_0 fireplace_1 fireplace_2 fireplace_3;
run;

5.
data combined;
set my_assign;
set part2;
set part3;
run;
proc reg data=combined;
model saleprice =  fireplace_0 fireplace_1 fireplace_2 fireplace_3 ExterCond_Ex
ExterCond_Gd ExterCond_TA ExterCond_Fa
```

```
            LotFrontage LotArea BsmtUnfSF GrLivArea MasVnrArea TotalBsmtSF
FirstFlrSF GarageArea WoodDeckSF /
             selection=forward;
            run;

backward selection:
data combined;
set my_assign;
set part2;
set part3;
run;
proc reg data=combined;
model saleprice =  fireplace_0 fireplace_1 fireplace_2 fireplace_3 ExterCond_Ex
ExterCond_Gd ExterCond_TA ExterCond_Fa
            LotFrontage LotArea BsmtUnfSF GrLivArea MasVnrArea TotalBsmtSF
FirstFlrSF GarageArea WoodDeckSF /
             selection=backward;
            run;

data combined;
set my_assign;
set part2;
set part3;
run;
proc reg data=combined;
model saleprice =  fireplace_0 fireplace_1 fireplace_2 fireplace_3 ExterCond_Ex
ExterCond_Gd ExterCond_TA ExterCond_Fa
            LotFrontage LotArea BsmtUnfSF GrLivArea MasVnrArea TotalBsmtSF
FirstFlrSF GarageArea WoodDeckSF /
             selection=adjrsq AIC BIC;
            run;

data combined;
set my_assign;
set part2;
set part3;
run;
proc reg data=combined;
model saleprice =  fireplace_0 fireplace_1 fireplace_2 fireplace_3 ExterCond_Ex
ExterCond_Gd ExterCond_TA ExterCond_Fa
            LotFrontage LotArea BsmtUnfSF GrLivArea MasVnrArea TotalBsmtSF
FirstFlrSF GarageArea WoodDeckSF /
             selection=cp aic bic;
            run;

data combined;
```

```
set my_assign;
set part2;
set part3;
run;
proc reg data=combined;
model saleprice =  fireplace_0 fireplace_1 fireplace_2 fireplace_3 ExterCond_Ex
ExterCond_Gd ExterCond_TA ExterCond_Fa
            LotFrontage LotArea BsmtUnfSF GrLivArea MasVnrArea TotalBsmtSF
FirstFlrSF GarageArea WoodDeckSF /
             selection=cp aic bic;
            run;
```

valdation framework:

```
data temp;  set mydata.ames_housing_data;  * 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=.;  run;
```

forward selection:

```
data temp;
set combined;
* generate a uniform(0,1) random variable with seed set to 123;
u = uniform(123);
*proc print data=temp (obs=10);
*run;
if (u < 0.70) then train = 1;
else train = 0;
if (train=1) then train_response=SalePrice;
else train_response=.;
run;
proc reg data=temp;
model train_response =  fireplace_0 fireplace_1 fireplace_2 fireplace_3 ExterCond_Ex
ExterCond_Gd ExterCond_TA ExterCond_Fa
 LotFrontage LotArea BsmtUnfSF GrLivArea MasVnrArea TotalBsmtSF FirstFlrSF GarageArea
WoodDeckSF /
 selection=forward;
            run;

proc reg data=temp;
model train_response =  fireplace_0 fireplace_1 fireplace_2 fireplace_3 ExterCond_Ex
ExterCond_Gd ExterCond_TA ExterCond_Fa
             LotFrontage LotArea BsmtUnfSF GrLivArea MasVnrArea TotalBsmtSF FirstFlrSF
GarageArea WoodDeckSF /
             selection=adjrsq AIC BIC;
            run;
```

Adjusted R:

```
data temp;
set combined;
* generate a uniform(0,1) random variable with seed set to 123;
u = uniform(123);
*proc print data=temp (obs=10);
*run;
if (u < 0.70) then train = 1;
else train = 0;
if (train=1) then train_response=SalePrice;
else train_response=.;
run;
*proc print data=temp (obs=10);
*run;
proc reg data=temp;
model train_response =  fireplace_0 fireplace_1 fireplace_2 fireplace_3 ExterCond_Ex
ExterCond_Gd ExterCond_TA ExterCond_Fa
            LotFrontage LotArea BsmtUnfSF GrLivArea MasVnrArea TotalBsmtSF FirstFlrSF
GarageArea WoodDeckSF /
             selection=adjrsq;
            run;
```

9.

```
proc reg data=temp;
model train_response = fireplace_0 fireplace_1 fireplace_2 fireplace_3 ExterCond_Ex
ExterCond_Fa
            LotFrontage LotArea BsmtUnfSF GrLivArea MasVnrArea TotalBsmtSF
FirstFlrSF GarageArea WoodDeckSF /
            selection=backward;
            output out=part9 predicted=yhat;
            run;

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

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

proc reg data=temp;
model train_response = fireplace_1 fireplace_2 fireplace_3 ExterCond_Ex ExterCond_Fa
```

```
            LotFrontage LotArea BsmtUnfSF GrLivArea MasVnrArea TotalBsmtSF
FirstFlrSF GarageArea WoodDeckSF /
            selection=backward;
            output out=part9 predicted=yhat;
            run;

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

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


proc reg data=temp;
model train_response= fireplace_0 fireplace_1 fireplace_2 fireplace_3 ExterCond_Ex
ExterCond_Gd ExterCond_TA ExterCond_Fa
            BsmtUnfSF GrLivArea MasVnrArea TotalBsmtSF GarageArea WoodDeckSF /
            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='Grade1';
 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;


for test data:

proc reg data=temp;
model train_response= fireplace_0 fireplace_1 fireplace_2 fireplace_3 ExterCond_Ex
ExterCond_Gd ExterCond_TA ExterCond_Fa
            BsmtUnfSF GrLivArea MasVnrArea TotalBsmtSF GarageArea WoodDeckSF /
            selection=forward;
            output out =part10 predicted=yhat;

          title ' ';

        proc print data=part10 (obs=10);
run;

Data part10b;
set part10;


if train=1 then delete;

 Length Prediction_Grade $7.;

  pct_diff= abs((yhat-saleprice)/ saleprice);

if pct_diff LE .10 then Prediction_Grade='Grade1';
 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;



proc reg data=temp;
```

```
model train_response= fireplace_0 fireplace_2 fireplace_3 ExterCond_Ex ExterCond_Gd
ExterCond_TA ExterCond_Fa
            LotFrontage LotArea BsmtUnfSF GrLivArea MasVnrArea TotalBsmtSF
FirstFlrSF GarageArea WoodDeckSF /
             selection=stepwise;
            output out =part10 predicted=yhat;

          title ' ';

        proc print data=part10 (obs=10);
run;

Data part10b;
set part10;


if train=1 then delete;

 Length Prediction_Grade $7.;

  pct_diff= abs((yhat-saleprice)/ saleprice);

if pct_diff LE .10 then Prediction_Grade='Grade1';
 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;
```