Introduction

The objective of this assignment is to perform automated variable selection techniques for identifying the “best” regression model for predicting sale price for homes in the Ames, Iowa area. The first phase includes the assessment of which predictor variables, based on common sense and business justification, made sense to include in a predictive model. After conducting some preliminary exploratory data analysis (EDA), it was concluded that the following predictor variable candidates would be considered in the model:

X1-GrLivArea
X2-LotArea
X3-AgeAtSale
X4-TotalBsmtSF
X5-total_baths_calc
X6-TotRmsAbvGrd
X7-highend_ind Neighborhood grouping
X8-midend_ind Neighborhood grouping
X9-good_heating
X10-excl_kitchen kitchen quality
X11-central_air
X12-fireplace_ind
X13-garage_ind
X14-good_basement_ind
X15-concr_foundation
X16-quality_index
X17-brick_exterior
X18-lot_frontage
X19-new_bldg
X20-old_bldg
X21-pos_cond
X22-recent_remodel

Training data set

The data set has been split in test/train data set as shown below. We will test the model accuracy by training the model on the 70% of the data set and validating its accuracy on remaining 30%.

Training and Validation Breakdown

train	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	470	31.89	470	31.89
1	1004	68.11	1474	100.00

Adjusted R-Squared Model (Model_AdjR2)

The results of the adjusted R-squared variable selection method determined that not all predictor variables candidates (X1-X22) should be included in the final model. The adjusted R-squared of the final model was 0.8941.   Table 1 below, you can see the summary output of the variable selection process with the final result in the first row.

Number in Model	Adjusted R-Square	R-Square	Variables in Model
20	0.8941	0.8962	GrLivArea LotArea AgeAtSale TotalBsmtSF total_baths_calc TotRmsAbvGrd highend_ind midend_ind good_heating excl_kitchen fireplace_ind garage_ind good_basement_ind concr_foundation quality_index brick_exterior lot_frontage new_bldg pos_cond recent_remodel
21	0.8941	0.8963	GrLivArea LotArea AgeAtSale TotalBsmtSF total_baths_calc TotRmsAbvGrd highend_ind midend_ind good_heating excl_kitchen central_air fireplace_ind garage_ind good_basement_ind concr_foundation quality_index brick_exterior lot_frontage new_bldg pos_cond recent_remodel
19	0.8940	0.8960	GrLivArea LotArea AgeAtSale TotalBsmtSF total_baths_calc TotRmsAbvGrd highend_ind midend_ind good_heating excl_kitchen fireplace_ind garage_ind good_basement_ind concr_foundation quality_index brick_exterior lot_frontage new_bldg pos_cond
20	0.8940	0.8961	GrLivArea LotArea AgeAtSale TotalBsmtSF total_baths_calc TotRmsAbvGrd highend_ind midend_ind good_heating excl_kitchen central_air fireplace_ind garage_ind good_basement_ind concr_foundation quality_index brick_exterior lot_frontage new_bldg pos_cond
21	0.8940	0.8962	GrLivArea LotArea AgeAtSale TotalBsmtSF total_baths_calc TotRmsAbvGrd highend_ind midend_ind good_heating excl_kitchen fireplace_ind garage_ind good_basement_ind concr_foundation quality_index brick_exterior lot_frontage new_bldg old_bldg pos_cond recent_remodel
22	0.8940	0.8963	GrLivArea LotArea AgeAtSale TotalBsmtSF total_baths_calc TotRmsAbvGrd highend_ind midend_ind good_heating excl_kitchen central_air fireplace_ind garage_ind good_basement_ind concr_foundation quality_index brick_exterior lot_frontage new_bldg old_bldg pos_cond recent_remodel

The adjusted R-squared model suggest to exclude variables central_air(X11) and old_bldg(X20).

Maximum R-Squared Model (Model_MaxR)

Like the adjusted R-squared variable selection method, the results of the maximum R-squared variable selection method determined that not all predictor variable candidates (X1-X22) should be included in the final model. The R-squared of the final model was 0.8963. In Table below, we can see the ANOVA and parameter estimates for the final suggested model from the maximum R-squared variable selection method. The model suggests excluding variables central_air(X11) and old_bldg(X20) as none of them is statistically significant.

Maximum R-Square Improvement: Step 22

Variable old_bldg Entered: R-Square = 0.8963 and C(p) = 23.0000

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	22	4.194377E12	1.906535E11	385.41	<.0001
Error	981	4.852826E11	494681557		
Corrected Total	1003	4.67966E12			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-53934	9640.19669	15483684932	31.30	<.0001
GrLivArea	69.54801	3.31138	2.182108E11	441.11	<.0001
LotArea	1.96205	0.29212	22315866889	45.11	<.0001
AgeAtSale	-219.32033	56.96199	7333533380	14.82	0.0001
TotalBsmtSF	38.28898	2.33774	1.327032E11	268.26	<.0001
total_baths_calc	5132.67410	1289.54560	7836816679	15.84	<.0001
TotRmsAbvGrd	-3580.02397	933.47814	7275942985	14.71	0.0001
highend_ind	35391	3828.79777	42266485117	85.44	<.0001
midend_ind	12984	3049.24304	8968871691	18.13	<.0001
good_heating	3955.14707	1827.02742	2318253160	4.69	0.0306
excl_kitchen	35186	3692.97613	44907956022	90.78	<.0001
central_air	3782.46140	4414.59196	363156308	0.73	0.3918
fireplace_ind	7284.99790	1647.75170	9669433752	19.55	<.0001
garage_ind	7991.88535	5133.16659	1199094722	2.42	0.1198
good_basement_ind	15397	1926.24636	31606635334	63.89	<.0001
concr_foundation	7407.05451	2225.57792	5479377099	11.08	0.0009
quality_index	1349.95149	98.00211	93862295314	189.74	<.0001
brick_exterior	10199	4163.31178	2968885202	6.00	0.0145
lot_frontage	-6946.67150	4483.66691	1187442423	2.40	0.1216
new_bldg	9269.62301	2581.11490	6380213124	12.90	0.0003
old_bldg	76.54691	3854.21978	195123	0.00	0.9842
pos_cond	6411.97868	4269.05417	1115954603	2.26	0.1334
recent_remodel	2454.97983	1910.88693	816491359	1.65	0.1992

Bounds on condition number: 6.052, 948.57

The above model is the best 22-variable model found.

No further improvement in R-Square is possible.

Mallow's Cp Model (Model_MCp)

As with the previous two models, the results of the Mallow's Cp variable selection method also determined that not all predictor variable candidates (X1-X22) should be included in the final model. The Mallow's Cp of the final model was 19.3852. In Table 3 below, we can see some of the output from the Mallow's Cp variable selection process. The best model is shown in the first row of the table.

Number in Model	C(p)	R-Square	Variables in Model
19	19.3852	0.8960	GrLivArea LotArea AgeAtSale TotalBsmtSF total_baths_calc TotRmsAbvGrd highend_ind midend_ind good_heating excl_kitchen fireplace_ind garage_ind good_basement_ind concr_foundation quality_index brick_exterior lot_frontage new_bldg pos_cond
18	19.5649	0.8958	GrLivArea LotArea AgeAtSale TotalBsmtSF total_baths_calc TotRmsAbvGrd highend_ind midend_ind good_heating excl_kitchen fireplace_ind garage_ind good_basement_ind concr_foundation quality_index brick_exterior new_bldg pos_cond
18	19.6215	0.8958	GrLivArea LotArea AgeAtSale TotalBsmtSF total_baths_calc TotRmsAbvGrd highend_ind midend_ind good_heating excl_kitchen fireplace_ind good_basement_ind concr_foundation quality_index brick_exterior lot_frontage new_bldg pos_cond
18	19.6531	0.8958	GrLivArea LotArea AgeAtSale TotalBsmtSF total_baths_calc TotRmsAbvGrd highend_ind midend_ind good_heating excl_kitchen fireplace_ind garage_ind good_basement_ind concr_foundation quality_index brick_exterior lot_frontage new_bldg
20	19.7446	0.8962	GrLivArea LotArea AgeAtSale TotalBsmtSF total_baths_calc TotRmsAbvGrd highend_ind midend_ind good_heating excl_kitchen fireplace_ind garage_ind good_basement_ind concr_foundation quality_index brick_exterior lot_frontage new_bldg pos_cond recent_remodel
17	19.7519	0.8956	GrLivArea LotArea AgeAtSale TotalBsmtSF total_baths_calc TotRmsAbvGrd highend_ind midend_ind good_heating excl_kitchen fireplace_ind good_basement_ind concr_foundation quality_index brick_exterior new_bldg pos_cond
17	19.8901	0.8956	GrLivArea LotArea AgeAtSale TotalBsmtSF total_baths_calc TotRmsAbvGrd highend_ind midend_ind good_heating excl_kitchen fireplace_ind good_basement_ind concr_foundation quality_index brick_exterior lot_frontage new_bldg
17	19.9387	0.8956	GrLivArea LotArea AgeAtSale TotalBsmtSF total_baths_calc TotRmsAbvGrd highend_ind midend_ind good_heating excl_kitchen fireplace_ind garage_ind good_basement_ind concr_foundation quality_index brick_exterior new_bldg

Apart from old_bldg(X20) and central_air(X11) the model suggest to exclude recent_remodel(X22) variable.

Forward Selection Model (Model_F)

Like proceeding variable selection methods, the results of the forward variable selection method also determined that not all predictor variable candidates (X1-X22) should be included in the final model. In table below, we can see the summary of the forward selection method and you'll notice that the p-value from the nested F-tests did not increase until the 10th variable was entered into the model. For this method, we chose a *slentry* value of 0.15 as our threshold for variables to be allowed to enter into the model.

Bounds on condition number: 4.333, 670.15

No other variable met the 0.1500 significance level for entry 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	0.5387	0.5387	3363.84	1170.14	<.0001
2	TotalBsmtSF	2	0.1817	0.7204	1646.54	650.78	<.0001
3	AgeAtSale	3	0.0616	0.7820	1066.27	282.34	<.0001
4	quality_index	4	0.0484	0.8304	610.084	285.35	<.0001
5	excl_kitchen	5	0.0269	0.8574	357.259	188.49	<.0001
6	highend_ind	6	0.0107	0.8681	258.062	80.84	<.0001
7	good_basement_ind	7	0.0100	0.8781	165.040	82.08	<.0001
8	LotArea	8	0.0065	0.8846	105.460	56.14	<.0001
9	fireplace_ind	9	0.0021	0.8868	87.3370	18.67	<.0001
10	concr_foundation	10	0.0014	0.8881	76.5495	12.00	0.0006
11	midend_ind	11	0.0017	0.8898	62.7638	15.02	0.0001
12	total_baths_calc	12	0.0012	0.8910	53.2835	11.03	0.0009
13	new_bldg	13	0.0015	0.8925	41.0800	13.83	0.0002
14	TotRmsAbvGrd	14	0.0015	0.8940	28.4736	14.41	0.0002
15	brick_exterior	15	0.0007	0.8947	24.3117	6.11	0.0136
16	good_heating	16	0.0007	0.8953	20.1251	6.17	0.0132
17	pos_cond	17	0.0003	0.8956	19.7519	2.37	0.1241
18	garage_ind	18	0.0002	0.8958	19.5649	2.19	0.1396
19	lot_frontage	19	0.0002	0.8960	19.3852	2.18	0.1400

The model suggest exclusion of central_air(X11), old_bldg(X20), recent_remodel(X22) variables

Backward Model (Model_B)

Also this time the results of the backward variable selection method determined that not all predictor variable candidates (X1-X22) should be included in the final model. In table below, we can see the summary of the beckward elimination method. For this method, we chose a *slentry* value of 0.15 as our threshold for variables to be allowed to enter into the model.

Backward Elimination: Step 3

Variable recent_remodel Removed: R-Square = 0.8960 and C(p) = 19.3852

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	19	4.193197E12	2.206946E11	446.41	<.0001
Error	984	4.864625E11	494372477		
Corrected Total	1003	4.67966E12			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-49213	8604.19466	16173318918	32.71	<.0001
GrLivArea	69.37933	3.30577	2.17756E11	440.47	<.0001
LotArea	1.96529	0.29142	22484324653	45.48	<.0001
AgeAtSale	-235.30525	48.18330	11790290034	23.85	<.0001
TotalBsmtSF	38.10275	2.33151	1.320365E11	267.08	<.0001
total_baths_calc	5199.85003	1285.76114	8085657317	16.36	<.0001
TotRmsAbvGrd	-3631.66316	928.73408	7559318441	15.29	<.0001
highend_ind	35586	3825.09585	42787872651	86.55	<.0001
midend_ind	12819	2979.75470	9149682815	18.51	<.0001
good_heating	4705.84648	1759.11934	3537848163	7.16	0.0076
excl_kitchen	35274	3679.21696	45440631704	91.92	<.0001
fireplace_ind	7054.77178	1631.23665	9246689699	18.70	<.0001
garage_ind	7631.59505	5101.66229	1106270945	2.24	0.1350
good_basement_ind	15356	1920.78023	31595902404	63.91	<.0001
concr_foundation	7584.67800	2140.86293	6205129619	12.55	0.0004
quality_index	1383.79985	95.28834	1.042609E11	210.90	<.0001
brick_exterior	10375	4156.52181	3080233308	6.23	0.0127
lot_frontage	-6588.78684	4461.34659	1078284276	2.18	0.1400
new_bldg	9952.69529	2487.88429	7911800844	16.00	<.0001
pos_cond	6428.73949	4267.55870	1121881359	2.27	0.1323

Bounds on condition number: 4.333, 670.15

All variables left in the model are significant at the 0.1500 level.

Summary of Backward Elimination							
Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	old_bldg	21	0.0000	0.8963	21.0004	0.00	0.9842
2	central_air	20	0.0001	0.8962	19.7446	0.74	0.3883
3	recent_remodel	19	0.0002	0.8960	19.3852	1.64	0.2003

The model suggest exclusion of central_air(X11), old_bldg(X20), recent_remodel(X22) variables

Stepwise Selection Model (Model_S)

The stepwise selection method was the final option used for model selection. As with previous methods, the stepwise selection method indicated that not all predictor variable candidates should remain in the model. The stepwise variable selection summary is shown below.

Bounds on condition number: 4.333, 670.15

All variables left in the model are significant at the 0.1500 level.

No other variable met the 0.1500 significance level for entry into the model.

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	0.5387	0.5387	3363.84	1170.14	<.0001
2	TotalBsmtSF		2	0.1817	0.7204	1646.54	650.78	<.0001
3	AgeAtSale		3	0.0616	0.7820	1066.27	282.34	<.0001
4	quality_index		4	0.0484	0.8304	610.084	285.35	<.0001
5	excl_kitchen		5	0.0269	0.8574	357.259	188.49	<.0001
6	highend_ind		6	0.0107	0.8681	258.062	80.84	<.0001
7	good_basement_ind		7	0.0100	0.8781	165.040	82.08	<.0001
8	LotArea		8	0.0065	0.8846	105.460	56.14	<.0001
9	fireplace_ind		9	0.0021	0.8868	87.3370	18.67	<.0001
10	concr_foundation		10	0.0014	0.8881	76.5495	12.00	0.0006
11	midend_ind		11	0.0017	0.8898	62.7638	15.02	0.0001
12	total_baths_calc		12	0.0012	0.8910	53.2835	11.03	0.0009
13	new_bldg		13	0.0015	0.8925	41.0800	13.83	0.0002
14	TotRmsAbvGrd		14	0.0015	0.8940	28.4736	14.41	0.0002
15	brick_exterior		15	0.0007	0.8947	24.3117	6.11	0.0136
16	good_heating		16	0.0007	0.8953	20.1251	6.17	0.0132
17	pos_cond		17	0.0003	0.8956	19.7519	2.37	0.1241
18	garage_ind		18	0.0002	0.8958	19.5649	2.19	0.1396
19	lot_frontage		19	0.0002	0.8960	19.3852	2.18	0.1400

At the 0.1500 level variable X20-X22 were excluded from the model.

Model Comparison

There has been slight variation in variable selection between models . Table below shows model fit criteria from the models created using the training sample. Based on the results, we can see that the

		Model_AdjR2	Model_MaxR	Model_MCp	Model_F	Model_B	Model_S
Train	Predictors excluded	X11,X20	X11,X20	X11,X20,X22	X11,X20,X22	X11,X20,X22	X11,X20,X22
Train	Adjusted R2	0.8941	0.8940	0.8940	0.8940	0.8940	0.8940
Train	AIC	20118.9967	20122.23	20118.67	20118.67	20118.67	20118.67
Train	BIC	20121.9478	20125.31	20121.51	20121.51	20121.51	20121.51
Train	Mallow's Cp	19.7446	23	19.3852	19.3852	19.3852	19.3852
Train	MAE	15879.15	15866.46	15907.54	15907.54	15907.54	15907.54
Train	MSE	483716094	483349211	484524420	484524420	484524420	484524420
Test	MAE	15052.86	15016.78	15107.78	15107.78	15107.78	15107.78
Test	MSE	426320079	424752379	430209443	430209443	430209443	430209443

predictive ability of the final model built from the training sample performed very well with the test sample data.

Multicollinearity

Values 10 or above for variance inflation analysis may cause serious problem for data analysis. It seems that there are no immediate risk for multicollinearity even though the value for AgeAtSale is a little bit high.

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-53934	9640.19669	-5.59	<.0001	0
GrLivArea	1	69.54801	3.31138	21.00	<.0001	3.93902
LotArea	1	1.96205	0.29212	6.72	<.0001	1.35207
AgeAtSale	1	-219.32033	56.96199	-3.85	0.0001	6.05198
TotalBsmtSF	1	38.28898	2.33774	16.38	<.0001	1.59373
total_baths_calc	1	5132.67410	1289.54560	3.98	<.0001	2.58350
TotRmsAbvGrd	1	-3580.02397	933.47814	-3.84	0.0001	2.82916
highend_ind	1	35391	3828.79777	9.24	<.0001	1.48772
midend_ind	1	12984	3049.24304	4.26	<.0001	1.95722
good_heating	1	3955.14707	1827.02742	2.16	0.0306	1.67886
excl_kitchen	1	35186	3692.97613	9.53	<.0001	1.43325
central_air	1	3782.46140	4414.59196	0.86	0.3918	1.22050
fireplace_ind	1	7284.99790	1647.75170	4.42	<.0001	1.31963
garage_ind	1	7991.88535	5133.16659	1.56	0.1198	1.09518
good_basement_ind	1	15397	1926.24636	7.99	<.0001	1.59566
concr_foundation	1	7407.05451	2225.57792	3.33	0.0009	2.51224
quality_index	1	1349.95149	98.00211	13.77	<.0001	1.40617
brick_exterior	1	10199	4163.31178	2.45	0.0145	1.05267
lot_frontage	1	-6946.67150	4483.66691	-1.55	0.1216	1.02925
new_bldg	1	9269.62301	2581.11490	3.59	0.0003	1.72775
old_bldg	1	76.54691	3854.21978	0.02	0.9842	2.36132
pos_cond	1	6411.97868	4269.05417	1.50	0.1334	1.07222
recent_remodel	1	2454.97983	1910.88693	1.28	0.1992	1.81775

Pearson Correlation analysis reveals that correlation between AgeAtSale and other variables is within -0.7 and 0.7. The highest value of 0.65 is between AgeAtSale and concr_foundation. The table below shows variables with highest Pearson Correlation.

Pearson Correlation Coefficients, N = 1474 Prob > |r| under H0: Rho=0					
	AgeAtSale	total_baths_calc	midend_ind	concr_foundation	old_bldg
AgeAtSale	1.00000	-0.60325 <.0001	-0.60669 <.0001	-0.65950 <.0001	0.62522 <.0001
total_baths_calc	-0.60325 <.0001	1.00000	0.33514 <.0001	0.47357 <.0001	-0.29533 <.0001
midend_ind	-0.60669 <.0001	0.33514 <.0001	1.00000	0.20778 <.0001	-0.59951 <.0001
concr_foundation	-0.65950 <.0001	0.47357 <.0001	0.20778 <.0001	1.00000	-0.18892 <.0001
old_bldg	0.62522 <.0001	-0.29533 <.0001	-0.59951 <.0001	-0.18892 <.0001	1.00000

Operational validation

To assess the operational accuracy of the final model, we placed the predictive scores, absolute value for each observation's actual vs predicted value for the response variable SalePrice, into three categories: Grade 1 (within 10% of the observed value), Grade 2 (between 10-15% of the observed value), Grade 3 (everything else). Below the results for each one of the models.

Model_AdjR2 Prediction Grade

Prediction_Grade	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Grade 1	324	68.94	324	68.94
Grade 2	79	16.81	403	85.74
Grade 3	67	14.26	470	100.00

Model_MaxR Prediction Grade

Prediction_Grade	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Grade 1	328	69.79	328	69.79
Grade 2	74	15.74	402	85.53
Grade 3	68	14.47	470	100.00

Model_MCp Prediction Grade

Prediction_Grade	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Grade 1	327	69.57	327	69.57
Grade 2	74	15.74	401	85.32
Grade 3	69	14.68	470	100.00

Model_F Prediction Grade

Prediction_Grade	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Grade 1	327	69.57	327	69.57
Grade 2	74	15.74	401	85.32
Grade 3	69	14.68	470	100.00

Model_B Prediction Grade

Prediction_Grade	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Grade 1	327	69.57	327	69.57
Grade 2	74	15.74	401	85.32
Grade 3	69	14.68	470	100.00

Model_S Prediction Grade

Prediction_Grade	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Grade 1	327	69.57	327	69.57
Grade 2	74	15.74	401	85.32
Grade 3	69	14.68	470	100.00

Based on the results of the prediction scores above, we can see that 85% of the predicted values were within 15% of the observed value.

Conclusion

In order to see effects of the automated variable selection I had to choose quite a few variables. The model might be hard to interpret for the business users, although it performs quite well. It has been proven that simpler models repeatedly run on larger data sets perform better in the long run.

It is quite clear how automated variable selection can be valuable when designing a predictive model. That should be combined with domain knowledge of stakeholders in order to achieve the best results.

```
libname mydata                    '/libname mydata                  '/scs/crb519/PREDICT_410/SAS_Data/'
access=readonly;


ods graphics on;

ods noproctitle;

title 'Assignment 5';

*data set A method b for defining variables without YearBuilt funkar utan yb 10 variabler

* create water drop condition to establish sample;




data building;

                    set mydata.ames_housing_data;

                    format drop_condition $40.;

                    if (BldgType ne '1Fam') then drop_condition='01: Not a Single Family';

                    else if (Zoning not in ('RH','RL','RM','FV')) then drop_condition='02: Non-Residential
Zoning';

                    else if (GrLivArea > 4000 and SalePrice < 200000) then drop_condition='03: Large Area
and Low Sale Price';

                    else if (FullBath < 1) then drop_condition='04: No Bath';

                    else if (LotArea > 20000) then drop_condition='05: Lot Area Over 20,000sqf';

                    else if (SalePrice > 700000) then drop_condition='06: Sale Price Over $700,000';

                    else if (SalePrice < 15000) then drop_condition='07: Sale Price Below $15,000';

                    else if (GrLivArea < 1000) then drop_condition='08: Living Area Below 1000sqf';

                    else if (GrLivArea > 3000) then drop_condition='09: Living Area Above 3000sqf';

                    else if (AgeAtSale > 100) then drop_condition='10: Age Over 100';

                    else if (Functional ne 'Typ')  then drop_condition='11: Non-Typcial Function';

                    else if (SaleCondition ne 'Normal') then drop_condition='12: Non-Normal Sale';

                    else drop_condition='99: Sample Population';

                    AgeAtSale = YrSold - YearBuilt;
```

```
                    logSalePrice = log(SalePrice);
run;


proc freq data=building;
 tables drop_condition;
 title 'Sample Waterfall';
run;



data train;
   set building
     (where = (drop_condition = '99: Sample Population'));

   u = uniform(123);
   if (u < 0.70) then train = 1; else train = 0;
   if (train=1) then train_response=SalePrice; else train_response=.;

   * create derived  variables;
                    total_SF = max(GrLivArea,0) + max(TotalBsmtSF,0);
                    total_baths = max(FullBath,0) + max(BsmtFullBath,0);
                    total_halfbaths = max(HalfBath,0) + max(BsmtHalfBath,0);
                    total_baths_calc = total_baths + total_halfbaths;


                    *Kitchen categorical variables;
```

```
if (KitchenQual = 'Ex') then excl_kitchen = 1; else excl_kitchen = 0;

*Fireplace;
if (Fireplaces>0) then fireplace_ind=1; else fireplace_ind=0;

*Exterior;
if (ExterQual in ('Ex', 'Gd')) then good_exterior=1; else good_exterior=0;
*Neighborhood variables;


if (Neighborhood in ('StoneBr','NridgHt','Greens','GrnHill')) then
  highend_ind = 1;
else
  highend_ind = 0;
if (Neighborhood in ('BrDale','IDOTRR','MeadowV','OldTown','SWISU')) then
  midend_ind = 0;
else
  midend_ind = 1;

*Foundation;
if (Foundation = 'PConc') then concr_foundation = 1; else concr_foundation = 0;

quality_index = OverallCond*OverallQual;

if(CentralAir='Y') then central_air=1; else central_air=0;

if (GarageCars>0) then garage_ind=1; else garage_ind=0;
```

```
        if (BsmtQual in ('Ex', 'Gd')) or (BmstCond in ('Ex', 'Gd'))
        then good_basement_ind=1;
        else good_basement_ind=0;


        if (BsmtFinType1 = 'GLQ') then good_basement_ind=1; else good_basement_ind=0;
        if (HeatingQC = 'Ex') then good_heating = 1; else good_heating = 0;


        if (Exterior1 in ('BrkComm','BrkFace')) or (Exterior2 in ('BrkComm','BrkFace'))
          then brick_exterior=1; else brick_exterior=0;
        if (LotConfig in ('FR2','FR3')) then lot_frontage=1; else lot_frontage=0;
        if (AgeAtSale < 5) then new_bldg = 1; else new_bldg = 0;
        if (AgeAtSale > 85) then old_bldg = 1; else old_bldg = 0;
        if (Condition1 in ('PosA','PosN') or Condition2 in ('PosA','PosN')) then pos_cond = 1;
else pos_cond = 0;
        YearsSinceRemodel =  YrSold - YearRemodel;
        if (YearsSinceRemodel >= 0 and YearsSinceRemodel <= 10) then recent_remodel = 1;
else recent_remodel = 0;
run;



* Model_AdjR2: use automatic variable selection with adjusted R squre ;
proc reg data = train plots =  diagnostics(unpack) outest = Model_AdjR2_est;
  Model_AdjR2: model train_response =
    GrLivArea LotArea AgeAtSale TotalBsmtSF total_baths_calc
    TotRmsAbvGrd highend_ind midend_ind good_heating excl_kitchen
        central_air fireplace_ind garage_ind good_basement_ind concr_foundation
        quality_index  brick_exterior lot_frontage new_bldg old_bldg pos_cond
recent_remodel
    / selection=adjrsq best=1 aic bic cp vif;
```

```
  output out=Model_AdjR2_out (keep=train SalePrice yhat) predicted=yhat;
  title 'Model_AdjR2';
run;



 * Model_MaxR: use automatic variable selection with MaxR ;
proc reg data = train plots =  diagnostics(unpack) outest = Model_MaxR_est;
  Model_MaxR: model train_response =
    GrLivArea LotArea AgeAtSale TotalBsmtSF total_baths_calc
    TotRmsAbvGrd highend_ind midend_ind good_heating excl_kitchen
                     central_air fireplace_ind garage_ind good_basement_ind concr_foundation
                     quality_index  brick_exterior lot_frontage new_bldg old_bldg pos_cond
recent_remodel
    / best=1 selection=maxr aic bic cp vif;
  output out=Model_MaxR_out (keep=train SalePrice yhat) predicted=yhat;
  title 'Model_MaxR';
run;

*------------------------------------------------------------------------------------------
* Modle_MCp
*------------------------------------------------------------------------------------------;
* Model_MCp: use automatic variable selection with Minimal Mallow's Cp ;
proc reg data = train plots =  diagnostics(unpack) outest = Model_MCp_est;
  Model_MCp: model train_response =
    GrLivArea LotArea AgeAtSale TotalBsmtSF total_baths_calc
    TotRmsAbvGrd highend_ind midend_ind good_heating excl_kitchen
                     central_air fireplace_ind garage_ind good_basement_ind concr_foundation
```

```
                    quality_index  brick_exterior lot_frontage new_bldg old_bldg pos_cond
recent_remodel

   / best=1 selection=cp best=1 aic bic cp vif;

  output out=Model_MCp_out (keep=train SalePrice yhat) predicted=yhat;

  title 'Model_MCp';

run;




* Model_F: use automatic variable selection with forward selection ;

proc reg data = train plots =  diagnostics(unpack) outest = Model_F_est;

  Model_F: model train_response =

    GrLivArea LotArea AgeAtSale TotalBsmtSF total_baths_calc

    TotRmsAbvGrd highend_ind midend_ind good_heating excl_kitchen

                    central_air fireplace_ind garage_ind good_basement_ind concr_foundation

                    quality_index  brick_exterior lot_frontage new_bldg old_bldg pos_cond
recent_remodel

   / best=1 selection=forward slentry=0.15 aic bic cp vif;

  output out=Model_F_out (keep=train SalePrice yhat) predicted=yhat;

  title 'Model_F';

run;


* Model_B: use automatic variable selection with backward selection ;

proc reg data = train plots =  diagnostics(unpack) outest = Model_B_est;

  Model_B: model train_response =

    GrLivArea LotArea AgeAtSale TotalBsmtSF total_baths_calc

    TotRmsAbvGrd highend_ind midend_ind good_heating excl_kitchen

                    central_air fireplace_ind garage_ind good_basement_ind concr_foundation
```

```
                    quality_index  brick_exterior lot_frontage new_bldg old_bldg pos_cond
recent_remodel
   / selection=backward slstay=0.15 best=1 aic bic cp vif;
  output out=Model_B_out (keep=train SalePrice yhat) predicted=yhat;
  title 'Model_B';
run;



* Model_S: use automatic variable selection with stepwise selection ;
proc reg data = train plots =  diagnostics(unpack) outest = Model_S_est;
  Model_S: model train_response =
   GrLivArea LotArea AgeAtSale TotalBsmtSF total_baths_calc
   TotRmsAbvGrd highend_ind midend_ind good_heating excl_kitchen
                    central_air fireplace_ind garage_ind good_basement_ind concr_foundation
                    quality_index  brick_exterior lot_frontage new_bldg old_bldg pos_cond
recent_remodel
   / best=1 selection=stepwise slentry=0.15 slstay=0.15 aic bic cp vif;
  output out=Model_S_out (keep=train SalePrice yhat) predicted=yhat;
  title 'Model_S';
run;

* create macros for printing the estimators and calculating MSE and MAE;
%macro print_est(indata,mytitle);
proc print data=&indata.;
title &mytitle.;
run; quit;
%mend;
```

```
%macro data_validation(indata,outdata);
data &outdata.;
  set &indata.;
  err = SalePrice - yhat;
  abserr = abs(SalePrice - yhat);
  sqerr  = (SalePrice - yhat) ** 2;
  if (abserr <= SalePrice * 0.1) then
    Prediction_Grade = 'Grade 1';
  else if (abserr <= SalePrice * 0.15) then
    Prediction_Grade = 'Grade 2';
  else
    Prediction_Grade = 'Grade 3';
run; quit;
%mend;

%macro print_grade(indata,mytitle);
proc freq data=&indata.;
  tables Prediction_Grade;
  where train=0;
  title &mytitle.;
run; quit;
%mend;

%macro print_mse_mae(indata,mytitle);
proc means data=&indata. mean min max median sum;
  var abserr sqerr;
  class train;
```

```
  title &mytitle;
run; quit;
%mend;



* print model estimates for all six models ;
%print_est(indata=%str(Model_AdjR2_est),mytitle=%str('Model_AdjR2 estimator'));
%print_est(indata=%str(Model_MaxR_est),mytitle=%str('Model_MaxR estimator'));
%print_est(indata=%str(Model_MCp_est),mytitle=%str('Model_MCp estimator'));
%print_est(indata=%str(Model_F_est),mytitle=%str('Model_F estimator'));
%print_est(indata=%str(Model_B_est),mytitle=%str('Model_B estimator'));
%print_est(indata=%str(Model_S_est),mytitle=%str('Model_S estimator'));

* run model in-sample and out-of-sample validation and print results ;
%data_validation(indata=%str(Model_AdjR2_out),outdata=%str(Model_AdjR2_validation));
%print_mse_mae(indata=%str(Model_AdjR2_validation),mytitle=%str('Model_AdjR2 MSE MAE'));
%data_validation(indata=%str(Model_MaxR_out),outdata=%str(Model_MaxR_validation));
%print_mse_mae(indata=%str(Model_MaxR_validation),mytitle=%str('Model_MaxR MSE MAE'));
%data_validation(indata=%str(Model_MCp_out),outdata=%str(Model_MCp_validation));
%print_mse_mae(indata=%str(Model_MCp_validation),mytitle=%str('Model_MCp MSE MAE'));
%data_validation(indata=%str(Model_F_out),outdata=%str(Model_F_validation));
%print_mse_mae(indata=%str(Model_F_validation),mytitle=%str('Model_F MSE MAE'));
%data_validation(indata=%str(Model_B_out),outdata=%str(Model_B_validation));
%print_mse_mae(indata=%str(Model_B_validation),mytitle=%str('Model_B MSE MAE'));
%data_validation(indata=%str(Model_S_out),outdata=%str(Model_S_validation));
%print_mse_mae(indata=%str(Model_S_validation),mytitle=%str('Model_S MSE MAE'));
```

```
* print model operational grades ;

%print_grade(indata=%str(Model_AdjR2_validation),mytitle=%str('Model_AdjR2 Prediction Grade'));

%print_grade(indata=%str(Model_MaxR_validation),mytitle=%str('Model_MaxR Prediction Grade'));

%print_grade(indata=%str(Model_MCp_validation),mytitle=%str('Model_MCp Prediction Grade'));

%print_grade(indata=%str(Model_F_validation),mytitle=%str('Model_F Prediction Grade'));

%print_grade(indata=%str(Model_B_validation),mytitle=%str('Model_B Prediction Grade'));

%print_grade(indata=%str(Model_S_validation),mytitle=%str('Model_S Prediction Grade'));



* run a final operationa validation of the model produced by F/S/B;

proc reg data = train plots =  diagnostics(unpack) outest = Model_Final_est;

  Model_Final: model train_response =

                              GrLivArea      LotArea        AgeAtSale      TotalBsmtSF
            total_baths_calc

                              TotRmsAbvGrd                  highend_ind    midend_ind     good_heating
            excl_kitchen

                              fireplace_ind  good_basement_ind             concr_foundation
            quality_index

                              brick_exterior new_bldg;

  output out=Model_Final_out (keep=train SalePrice YrSold yhat) predicted=yhat;

  title 'Model_Final';

run;


%data_validation(indata=%str(Model_Final_out),outdata=%str(Model_Final_validation));




*Multicollinearity;
```

```
proc reg data=train;

model train_response =

   GrLivArea LotArea AgeAtSale TotalBsmtSF total_baths_calc

   TotRmsAbvGrd highend_ind midend_ind good_heating excl_kitchen

                        central_air fireplace_ind garage_ind good_basement_ind concr_foundation

                        quality_index  brick_exterior lot_frontage new_bldg old_bldg pos_cond
recent_remodel / corrb vif collin;



                        ods graphics on;

title 'Correlation matrix ';

proc corr data=train plots=scatter(alpha=.20 .30);

  var  GrLivArea LotArea AgeAtSale TotalBsmtSF total_baths_calc

   TotRmsAbvGrd highend_ind midend_ind good_heating excl_kitchen

                        central_air fireplace_ind garage_ind good_basement_ind concr_foundation

                        quality_index  brick_exterior lot_frontage new_bldg old_bldg pos_cond
recent_remodel ;

 run;

ods graphics off;



                        ods graphics on;

title 'Correlation matrix ';

proc corr data=train plots=scatter(alpha=.20 .30);

  var  AgeAtSale  total_baths_calc

    midend_ind

                        concr_foundation
```

```
                    old_bldg  ;
 run;
ods graphics off;scs/crb519/PREDICT_410/SAS_Data/' access=readonly;


ods graphics on;
ods noproctitle;
title 'Assignment 5';
*data set A method b for defining variables without YearBuilt funkar utan yb 10 variabler
* create water drop condition to establish sample;




data building;
                    set mydata.ames_housing_data;
                    format drop_condition $40.;
                    if (BldgType ne '1Fam') then drop_condition='01: Not a Single Family';
                    else if (Zoning not in ('RH','RL','RM','FV')) then drop_condition='02: Non-Residential
Zoning';
                    else if (GrLivArea > 4000 and SalePrice < 200000) then drop_condition='03: Large Area
and Low Sale Price';
                    else if (FullBath < 1) then drop_condition='04: No Bath';
                    else if (LotArea > 20000) then drop_condition='05: Lot Area Over 20,000sqf';
                    else if (SalePrice > 700000) then drop_condition='06: Sale Price Over $700,000';
                    else if (SalePrice < 15000) then drop_condition='07: Sale Price Below $15,000';
                    else if (GrLivArea < 1000) then drop_condition='08: Living Area Below 1000sqf';
                    else if (GrLivArea > 3000) then drop_condition='09: Living Area Above 3000sqf';
                    else if (AgeAtSale > 100) then drop_condition='10: Age Over 100';
                    else if (Functional ne 'Typ')  then drop_condition='11: Non-Typcial Function';
                    else if (SaleCondition ne 'Normal') then drop_condition='12: Non-Normal Sale';
```

```
                        else drop_condition='99: Sample Population';
                        AgeAtSale = YrSold - YearBuilt;
                        logSalePrice = log(SalePrice);
run;


proc freq data=building;
 tables drop_condition;
 title 'Sample Waterfall';
run;



data train;
    set building
      (where = (drop_condition = '99: Sample Population'));

    u = uniform(123);
    if (u < 0.70) then train = 1; else train = 0;
    if (train=1) then train_response=SalePrice; else train_response=.;

    * create derived  variables;
                        total_SF = max(GrLivArea,0) + max(TotalBsmtSF,0);
                        total_baths = max(FullBath,0) + max(BsmtFullBath,0);
                        total_halfbaths = max(HalfBath,0) + max(BsmtHalfBath,0);
                        total_baths_calc = total_baths + total_halfbaths;
```

```
*Kitchen categorical variables;
if (KitchenQual = 'Ex') then excl_kitchen = 1; else excl_kitchen = 0;

*Fireplace;
if (Fireplaces>0) then fireplace_ind=1; else fireplace_ind=0;

*Exterior;
if (ExterQual in ('Ex', 'Gd')) then good_exterior=1; else good_exterior=0;
*Neighborhood variables;

if (Neighborhood in ('StoneBr','NridgHt','Greens','GrnHill')) then
  highend_ind = 1;
else
  highend_ind = 0;
if (Neighborhood in ('BrDale','IDOTRR','MeadowV','OldTown','SWISU')) then
  midend_ind = 0;
else
  midend_ind = 1;

*Foundation;
if (Foundation = 'PConc') then concr_foundation = 1; else concr_foundation = 0;

quality_index = OverallCond*OverallQual;

if(CentralAir='Y') then central_air=1; else central_air=0;
```

```
		if (GarageCars>0) then garage_ind=1; else garage_ind=0;

		if (BsmtQual in ('Ex', 'Gd')) or (BmstCond in ('Ex', 'Gd'))

		then good_basement_ind=1;

		else good_basement_ind=0;


		if (BsmtFinType1 = 'GLQ') then good_basement_ind=1; else good_basement_ind=0;

		if (HeatingQC = 'Ex') then good_heating = 1; else good_heating = 0;


		if (Exterior1 in ('BrkComm','BrkFace')) or (Exterior2 in ('BrkComm','BrkFace'))

		  then brick_exterior=1; else brick_exterior=0;

		if (LotConfig in ('FR2','FR3')) then lot_frontage=1; else lot_frontage=0;

		if (AgeAtSale < 5) then new_bldg = 1; else new_bldg = 0;

		if (AgeAtSale > 85) then old_bldg = 1; else old_bldg = 0;

		if (Condition1 in ('PosA','PosN') or Condition2 in ('PosA','PosN')) then pos_cond = 1;
else pos_cond = 0;

		YearsSinceRemodel =  YrSold - YearRemodel;

		if (YearsSinceRemodel >= 0 and YearsSinceRemodel <= 10) then recent_remodel = 1;
else recent_remodel = 0;

run;



* Model_AdjR2: use automatic variable selection with adjusted R squre ;

proc reg data = train plots =  diagnostics(unpack) outest = Model_AdjR2_est;

  Model_AdjR2: model train_response =

    GrLivArea LotArea AgeAtSale TotalBsmtSF total_baths_calc

    TotRmsAbvGrd highend_ind midend_ind good_heating excl_kitchen

		central_air fireplace_ind garage_ind good_basement_ind concr_foundation
```

```
                        quality_index  brick_exterior lot_frontage new_bldg old_bldg pos_cond
recent_remodel

    / selection=adjrsq best=1 aic bic cp vif;

  output out=Model_AdjR2_out (keep=train SalePrice yhat) predicted=yhat;

  title 'Model_AdjR2';

run;




 * Model_MaxR: use automatic variable selection with MaxR ;

proc reg data = train plots =  diagnostics(unpack) outest = Model_MaxR_est;

  Model_MaxR: model train_response =

    GrLivArea LotArea AgeAtSale TotalBsmtSF total_baths_calc

    TotRmsAbvGrd highend_ind midend_ind good_heating excl_kitchen

                        central_air fireplace_ind garage_ind good_basement_ind concr_foundation

                        quality_index  brick_exterior lot_frontage new_bldg old_bldg pos_cond
recent_remodel

    / best=1 selection=maxr aic bic cp vif;

  output out=Model_MaxR_out (keep=train SalePrice yhat) predicted=yhat;

  title 'Model_MaxR';

run;


*-------------------------------------------------------------------------------------------

* Modle_MCp

*-------------------------------------------------------------------------------------------;

* Model_MCp: use automatic variable selection with Minimal Mallow's Cp ;

proc reg data = train plots =  diagnostics(unpack) outest = Model_MCp_est;

  Model_MCp: model train_response =
```

```
   GrLivArea LotArea AgeAtSale TotalBsmtSF total_baths_calc
   TotRmsAbvGrd highend_ind midend_ind good_heating excl_kitchen
                  central_air fireplace_ind garage_ind good_basement_ind concr_foundation
                  quality_index  brick_exterior lot_frontage new_bldg old_bldg pos_cond
recent_remodel
   / best=1 selection=cp best=1 aic bic cp vif;
  output out=Model_MCp_out (keep=train SalePrice yhat) predicted=yhat;
  title 'Model_MCp';
run;




* Model_F: use automatic variable selection with forward selection ;
proc reg data = train plots =  diagnostics(unpack) outest = Model_F_est;
  Model_F: model train_response =
   GrLivArea LotArea AgeAtSale TotalBsmtSF total_baths_calc
   TotRmsAbvGrd highend_ind midend_ind good_heating excl_kitchen
                  central_air fireplace_ind garage_ind good_basement_ind concr_foundation
                  quality_index  brick_exterior lot_frontage new_bldg old_bldg pos_cond
recent_remodel
   / best=1 selection=forward slentry=0.15 aic bic cp vif;
  output out=Model_F_out (keep=train SalePrice yhat) predicted=yhat;
  title 'Model_F';
run;

* Model_B: use automatic variable selection with backward selection ;
proc reg data = train plots =  diagnostics(unpack) outest = Model_B_est;
  Model_B: model train_response =
```

```
   GrLivArea LotArea AgeAtSale TotalBsmtSF total_baths_calc
   TotRmsAbvGrd highend_ind midend_ind good_heating excl_kitchen
                       central_air fireplace_ind garage_ind good_basement_ind concr_foundation
                       quality_index  brick_exterior lot_frontage new_bldg old_bldg pos_cond
recent_remodel
   / selection=backward slstay=0.15 best=1 aic bic cp vif;
  output out=Model_B_out (keep=train SalePrice yhat) predicted=yhat;
  title 'Model_B';
run;



* Model_S: use automatic variable selection with stepwise selection ;
proc reg data = train plots =  diagnostics(unpack) outest = Model_S_est;
  Model_S: model train_response =
   GrLivArea LotArea AgeAtSale TotalBsmtSF total_baths_calc
   TotRmsAbvGrd highend_ind midend_ind good_heating excl_kitchen
                       central_air fireplace_ind garage_ind good_basement_ind concr_foundation
                       quality_index  brick_exterior lot_frontage new_bldg old_bldg pos_cond
recent_remodel
   / best=1 selection=stepwise slentry=0.15 slstay=0.15 aic bic cp vif;
  output out=Model_S_out (keep=train SalePrice yhat) predicted=yhat;
  title 'Model_S';
run;

* create macros for printing the estimators and calculating MSE and MAE;
%macro print_est(indata,mytitle);
proc print data=&indata.;
title &mytitle.;
```

```
run; quit;
%mend;

%macro data_validation(indata,outdata);
data &outdata.;
  set &indata.;
  err = SalePrice - yhat;
  abserr = abs(SalePrice - yhat);
  sqerr  = (SalePrice - yhat) ** 2;
  if (abserr <= SalePrice * 0.1) then
    Prediction_Grade = 'Grade 1';
  else if (abserr <= SalePrice * 0.15) then
    Prediction_Grade = 'Grade 2';
  else
    Prediction_Grade = 'Grade 3';
run; quit;
%mend;

%macro print_grade(indata,mytitle);
proc freq data=&indata.;
  tables Prediction_Grade;
  where train=0;
  title &mytitle.;
run; quit;
%mend;

%macro print_mse_mae(indata,mytitle);
```

```
proc means data=&indata. mean min max median sum;
  var abserr sqerr;
  class train;
  title &mytitle;
run; quit;
%mend;


* print model estimates for all six models ;
%print_est(indata=%str(Model_AdjR2_est),mytitle=%str('Model_AdjR2 estimator'));
%print_est(indata=%str(Model_MaxR_est),mytitle=%str('Model_MaxR estimator'));
%print_est(indata=%str(Model_MCp_est),mytitle=%str('Model_MCp estimator'));
%print_est(indata=%str(Model_F_est),mytitle=%str('Model_F estimator'));
%print_est(indata=%str(Model_B_est),mytitle=%str('Model_B estimator'));
%print_est(indata=%str(Model_S_est),mytitle=%str('Model_S estimator'));

* run model in-sample and out-of-sample validation and print results ;
%data_validation(indata=%str(Model_AdjR2_out),outdata=%str(Model_AdjR2_validation));
%print_mse_mae(indata=%str(Model_AdjR2_validation),mytitle=%str('Model_AdjR2 MSE MAE'));
%data_validation(indata=%str(Model_MaxR_out),outdata=%str(Model_MaxR_validation));
%print_mse_mae(indata=%str(Model_MaxR_validation),mytitle=%str('Model_MaxR MSE MAE'));
%data_validation(indata=%str(Model_MCp_out),outdata=%str(Model_MCp_validation));
%print_mse_mae(indata=%str(Model_MCp_validation),mytitle=%str('Model_MCp MSE MAE'));
%data_validation(indata=%str(Model_F_out),outdata=%str(Model_F_validation));
%print_mse_mae(indata=%str(Model_F_validation),mytitle=%str('Model_F MSE MAE'));
%data_validation(indata=%str(Model_B_out),outdata=%str(Model_B_validation));
%print_mse_mae(indata=%str(Model_B_validation),mytitle=%str('Model_B MSE MAE'));
```

```
%data_validation(indata=%str(Model_S_out),outdata=%str(Model_S_validation));
%print_mse_mae(indata=%str(Model_S_validation),mytitle=%str('Model_S MSE MAE'));

* print model operational grades ;
%print_grade(indata=%str(Model_AdjR2_validation),mytitle=%str('Model_AdjR2 Prediction Grade'));
%print_grade(indata=%str(Model_MaxR_validation),mytitle=%str('Model_MaxR Prediction Grade'));
%print_grade(indata=%str(Model_MCp_validation),mytitle=%str('Model_MCp Prediction Grade'));
%print_grade(indata=%str(Model_F_validation),mytitle=%str('Model_F Prediction Grade'));
%print_grade(indata=%str(Model_B_validation),mytitle=%str('Model_B Prediction Grade'));
%print_grade(indata=%str(Model_S_validation),mytitle=%str('Model_S Prediction Grade'));



* run a final operationa validation of the model produced by F/S/B;
proc reg data = train plots =  diagnostics(unpack) outest = Model_Final_est;
  Model_Final: model train_response =
                    GrLivArea   LotArea     AgeAtSale   TotalBsmtSF
        total_baths_calc
                    TotRmsAbvGrd            highend_ind midend_ind  good_heating
        excl_kitchen
                    fireplace_ind good_basement_ind     concr_foundation
        quality_index
                    brick_exterior new_bldg;
  output out=Model_Final_out (keep=train SalePrice YrSold yhat) predicted=yhat;
  title 'Model_Final';
run;

%data_validation(indata=%str(Model_Final_out),outdata=%str(Model_Final_validation));
```

```
*Multicollinearity;

proc reg data=train;
model train_response =
   GrLivArea LotArea AgeAtSale TotalBsmtSF total_baths_calc
   TotRmsAbvGrd highend_ind midend_ind good_heating excl_kitchen
                    central_air fireplace_ind garage_ind good_basement_ind concr_foundation
                    quality_index  brick_exterior lot_frontage new_bldg old_bldg pos_cond
recent_remodel / corrb vif collin;


                    ods graphics on;
title 'Correlation matrix ';
proc corr data=train plots=scatter(alpha=.20 .30);
  var  GrLivArea LotArea AgeAtSale TotalBsmtSF total_baths_calc
   TotRmsAbvGrd highend_ind midend_ind good_heating excl_kitchen
                    central_air fireplace_ind garage_ind good_basement_ind concr_foundation
                    quality_index  brick_exterior lot_frontage new_bldg old_bldg pos_cond
recent_remodel ;
 run;
ods graphics off;


                    ods graphics on;
title 'Correlation matrix ';
proc corr data=train plots=scatter(alpha=.20 .30);
```

```
   var  AgeAtSale  total_baths_calc
     midend_ind
                   concr_foundation
                   old_bldg  ;
 run;
ods graphics off;
```