SOUTHERN METHODIST UNIVERSITY
MSDS 6371(401)

Kaggle Project

Michael J Wolfe
Sandesh Ojha
Carl Walenciak
Travis Daun

Github Repository

https://github.com/mjwolfe91/SFDS_401_Team3_Kaggle_Project

February 27, 2019

CONTENTS

1. INTRODUCTION

Many factors can impact the sale price of residential real estate. In this analysis we will explore various factors that tend to impact home prices by looking at the neighborhood of North Ames, Edwards and Brookside in Ames, Iowa. We will explore factors like above ground living space, neighborhoods and the impact on sales price. At the end of this analysis, we will quantify the relationship between living area and sales price with respect to these neighborhoods.

After examining the impact of above ground living space on sales price for these neighborhoods, we will provide a complex analysis of factors that can be used to determine sales price for all of Ames. While we will use various methods to determine what factors are important for our predictive model, we will test this model on a blind dataset to show how it performs. The results of this model on the test data will be provided to further build the level of confidence for predicting future home sale prices.

2. DATA DESCRIPTION

The dataset used for this analysis was retrieved from www.kaggle.com/c/house-prices-advanced-regression-techniques.

DATA The data for this evaluation contained 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa. A complete description of these data can be found in Appendix D. The explanatory variables contain both categorical and numeric attributes. Table D.1, provides a high level summary of the variables and variable types contained in this dataset.

PREPROCESSING The variable `LotFrontage` posed a challenge since it was a continuous numerical variable that contained `NA`. This is because the variable was for the linear feet of street connected to property. In many cases (259 of 1460) this was either unknown or unrecorded. Our team made the decision to convert these NA values to 0. We believe this is an acceptable practice since we performed a sensitivity analysis by replacing the NAs with values from 0 to 140 (mean value is 70) with no impact on the linear regression model selection process and this factor was not utilized in any of our final models.

TRAINING Training of the linear regression models will be done utilizing the `training.csv` data obtained from above. A five fold cross validation will be employed for model selection.

TESTING Testing of the linear regression models will be done utilizing the `test.csv` data obtained from above.

RESULTS Datafiles containing the test results of the linear regression models can be found in our Github repository at https://github.com/mjwolfe91/SFDS_401_Team3_Kaggle_Project

3. Analysis Question 1

3.1. Restatement of Problem

REQUEST Century 21 Ames only sells houses in the Noth Ames, Edwards and Brookside neighborhoods and wants an estimate of how the sale price of the house is related to the square footage of the living area of the house. Additionally, Century 21 Ames would like to know if the sale price (and its relationship to square footage) depends on which neighborhood the house is located in. A fit a model will be used to answer this question.

DELIVERABLE Provide the estimate (or estimates if it varies by neighborhood) as well as confidence intervals for any estimate(s) you provide. Provide evidence that the model assumptions are met and that any suspicious observations (outliers / influential observations) have been identified and addressed. Finally, a conclusion that quantifies the relationship between living area and sale price with respect to these three neighborhoods.

3.2. Build and Fit the Model

Listing 1 of Appendix C contains the SAS code used to check the assumptions, clean the data, and run the model to determine the best estimate for sale price based on square footage for the North Ames, Edwards, and Brookside neighborhoods. We removed four data points from the original dataset due to being outliers. First sale prices greater than $300,000 were determined to not be representative of the total population of these three neighborhoods. Second, sale condition was limited to only those that were normal sales. Again we feel that home sales that were not normal sales such as foreclosures, linked properties, and land purchases were not representative of the population of interest.

3.3. Checking Assumptions

3.3.1. Assumptions

LINEARITY The linearity assumption is met by the reviewing the scatter plots associated with the data. Fig B.1 shows a plot of SalePrice vs. GrLivArea and by removing the outliers the linearity assumption is reasonably met.

CONSTANT VARIANCE The residual plot, Fig B.2 resembles somewhat of a random scatter of points around the 0 line, although there is a slight suspicion of non-constant variance judging from the dense cloud around the predicted value of $130,000. Also shown is the Studentized Residual Plot which is very similar to the residual plot, although this plot identifies potential outlying observations.

NORMALITY Based upon the histograms and q-q plots in Fig B.3 there is no evidence to suggest against normality. Additionally the random scatter associated with the residual plots in Fig B.2 also support the normality assumption.

INDEPENDENCE The independence assumption can be assumed to be maintained since these are all unique sales in a free housing market.

3.3.2. Outliers and Influential Points

OUTLIERS / INFLUENTIAL OBSERVATIONS There are distinct outliers that can be seen both in the Cook's D and Outlier and Leverage Diagnostics seen in Fig B.4. By removing the observations that resulted from non-normal sales conditions such as foreclosures and sale prices that were significantly outside the population, the Cook's D and Outlier and Leverage Diagnostics significantly improved. We are confident that the removal of these observations was appropriate since they do not represent the population as a whole. Additionally we are confident that the remaining data does not contain significant outliers or leverage points that need to be addressed further.

The model is a reasonable fit without transformations. The removal of observations not reflective of the population, such as non-normal sale conditions and home sales greater than $300,000, seems appropriate and enables the required model assumptions to be met.

3.3.3. Effect by Neighborhood

MODEL A model was developed to individually look at the neighborhoods of interest to determine if the sale price is impacted by neighborhood. Fig B.5 shows the sales price vs living area by neighborhood. The model shows (Table A.1) that the intercept and slope for Brookside and Edwards do differ from North Ames with statistical significance (p-values: < .0001 and .0077). As such we have determined to choose this model which allows for different intercept and slopes based upon the neighborhood. The resultant model can be seen in Fig B.6.

3.4. Model Metrics

Model metrics can be seen in Table A.2.

3.4.1. Adj R^2

ADJUSTED R^2 obtained for this model is 0.53.

3.4.2. Internal Press

PRESS obtained by this model is $1.96E11$.

3.5. Parameters

3.5.1. Estimates

The parameter estimates can be seen in Table A.1. With these estimates a separate equation can be written for each neighborhood to predict sale price based on living area.

$$\begin{aligned} SalePrice &= 74982 + 54.93(GrLivArea) - 55776(BrkSide) - 30274(Edwards) + \\ &\quad 33.31(GrLivArea)(BrkSide) + 8.24(GrLivArea)(Edwards) \end{aligned}$$

NORTH AMES

$$SalePrice = 74982 + 54.93(GrLivArea)$$

EDWARDS

$$\begin{aligned} SalePrice &= 74982 + 54.93(GrLivArea) - 30274 + 8.24(GrLivArea) \\ &= 44708 + 63.17(GrLivArea) \end{aligned}$$

BROOKSIDE

$$\begin{aligned} SalePrice &= 74982 + 54.93(GrLivArea) - 55776 + 33.31(GrLivArea) \\ &= 19206 + 88.24(GrLivArea) \end{aligned}$$

3.5.2. Interpretation

β_0 The intercept in this model provides an estimate (74982) of the sale price of a home in North Ames (reference neighborhood) with a living area of zero. Of course, this is extrapolation and does not have a clear, practical meaning.

β_1 For each 100 square foot increase in the living area of a home in North Ames, the estimated sale price increases $54.93.

β_2 This is the adjustment of the intercept for a home in Brookside with respect to a home in North Ames. For a living area of zero, the home in Brookside has an estimated sale price of $55,776 less than a home in North Ames.

β_3 This is the adjustment of the intercept for a home in Edwards with respect to a home in North Ames. For a living area of zero, the home in Edwards has an estimated sale price of $30,274 less than a home in North Ames.

β_4 For each 100 square foot increase in the living area of a home in Brookside, the estimated sale price increases $33 from the change with the home in North Ames.

β_5 For each 100 square foot increase in the living area of a home in Edwards, the estimated sale price increases $8 from the change with the home in North Ames.

3.5.3. Confidence Intervals

The confidence intervals for the estimates can be seen in Table A.1.

3.6. Conclusion

The square feet of above ground living space is a statistically significant factor to use to predict the sale price of homes in the North Ames, Edwards, and Brookside neighborhoods of Ames, Iowa. The existing sale prices of homes in these neighborhoods that are sold, under $300,000 and underwent a normal sales condition meet all the assumptions required to generate an appropriate linear regression model. We also examined the differences in predicted prices in these neighborhoods and determined that the sale prices of homes in each neighborhood do differ from each other. Homes in North Ames under 1600 square feet are predicted to sell for a higher price compared to the other two neighborhoods. Out of these three neighborhoods, the price per square foot in Brookside increases at the highest rate. 8 Homes greater

than approximately 1000 square feet in Brookside are predicted to sell for higher prices than comparable homes in Edwards. Homes greater than approximately 1600 square feet in Brookside are predicted to sell for higher prices than comparable homes in North Ames. Homes in the Edwards neighborhood are predicted to sell for the lowest price except for homes that are smaller than 1000 square feet.

4. Analysis Question 2

4.1. Restatement of Problem

Build the most predictive model for sales prices of homes in all of Ames Iowa. This includes all neighborhoods. Your group is limited to only the techniques we have learned in 6371 (no random forests or other methods we have not yet covered). Specifically, you should produce 4 models: one from forward selection, one from backwards elimination, one from stepwise selection, and one that you build custom. The custom model could be one of the three preceding models or one that you build by adding or subtracting variables at your will. Generate an adjusted R2, CV Press and Kaggle Score for each of these models and clearly describe which model you feel is the best in terms of being able to predict future sale prices of homes in Ames, Iowa.

4.2. Model Selection

4.2.1. Forward

$$\begin{aligned} SalePrice &= -1271664 + 648.153(YearBuilt) + 47.049(TotalBsmtSF) \\ &\quad +80.004(GrLivArea) - 671551(RoofMatl\_ClyTile) \\ &\quad +62871(BsmtQual\_Ex) \end{aligned}$$

The model summary can be seen in Tables A.3 and A.4. The SAS code for the forward model can be found in Listing 2 of Appendix C.

Assumptions

The fit diagnostics and residuals by regressors for sale price can be seen in Fig B.7. A scatterplot of the predicted values vs. the sale price using the features obtained by the forward selection model can be seen in Fig B.8.

Linearity Visual tests support assumptions of linearity. There is no evidence to suggest that these data are not linear based on the scatterplot in Fig B.8.

Constant Variance Examining Fig B.7 supports a loose random scatter of points around the origin.

Normality Visual analysis of histograms and q-q plots in Fig B.9 suggests normality of the data. Additionally the random cloud scatter plotted in Fig B.7 supports the normality assumption. Left and right leaning trailing tails in the q-q plot does suggest that outliers on both ends may not meet the normality assumption. We will address this in later analysis.

Independence Century21 Ames indicated all samples are unique sales in a free housing market, supporting assumptions of independence.

OUTLIERS AND INFLUENTIAL POINTS

OUTLIERS / INFLUENTIAL OBSERVATIONS There are distinct outliers that can be seen both in the Cook's D and Outlier and Leverage Diagnostics seen in Fig B.10; however, these outliers do not have high enough leverage to engender corrective action.

RESULTS The model performance metrics can be seen in Table A.5. The predicted results of the test file obtained a Kaggle score of 0.20306 as can be seen in Table A.6.

4.2.2. Backward

$$\begin{aligned} SalePrice &= 120177 + 35.192(BsmtFinSF1) + 85.189(1stFlrSF) \\ &\quad +59.655(2ndFlrSF) - 120853(OverallQual\_3) \\ &\quad -110889(OverallQual\_4) - 101945(OverallQual\_5) \\ &\quad -86980(OverallQual\_6) - 57950(OverallQual\_7) \\ &\quad -615180(RoofMatl\_ClyTile) \end{aligned}$$

The model summary can be seen in Table A.7. The SAS code for the forward model can be found in Listing 3 of Appendix C.

ASSUMPTIONS

The fit diagnostics and residuals by regressors for sale price can be seen in Fig B.11. A scatterplot of the predicted values vs. the sale price using the features obtained by the forward selection model can be seen in Fig B.12.

LINEARITY Visual tests of the scatter plots in Fig B.12 meet assumptions of linearity.

CONSTANT VARIANCE The residual plots in Fig B.11 support a loose random scatter of points around the origin. The Studentized Residual Plot also supports this assumption while highlighting outlying observations. We will address these in later analysis.

NORMALITY Visual analysis of histograms and q-q plots in Fig B.13 suggests normality of the data. Additionally the random cloud scatter plotted in Fig B.11 also support the normality assumption. supports the normality assumption. Left and right leaning trailing tails in the q-q plot does suggest that outliers on both ends may not meet the normality assumption. We will address this in later analysis.

INDEPENDENCE Century21 Ames indicated all samples are unique sales in a free housing market, supporting assumptions of independence.

OUTLIERS AND INFLUENTIAL POINTS

OUTLIERS / INFLUENTIAL OBSERVATIONS There are distinct outliers that can be seen both in the Cook's D and Outlier and Leverage Diagnostics seen in Fig B.14; however, these outliers do not have high enough leverage to engender corrective action.

RESULTS The model performance metrics can be seen in Table A.8. The predicted results of the test file obtained a Kaggle score of 0.23930 as can be seen in Table A.9.

4.2.3. Stepwise

$$\begin{aligned} SalePrice &= -1271664 + 648.153(YearBuilt) + 47.049(TotalBsmtSF) \\ &\quad +80.004(GrLivArea) - 671551(RoofMatl\_ClyTile) \\ &\quad +62871(BsmtQual\_Ex) \end{aligned}$$

The model summary can be seen in Tables A.10 and A.11. The SAS code for the forward model can be found in Listing 4 of Appendix C

ASSUMPTIONS

The fit diagnostics and residuals by regressors for sale price can be seen in Fig B.15. A scatterplot of the predicted values vs the sale price using the features obtained by the forward selection model can be seen in Fig B.16.

LINEARITY Visual tests of the scatter plots in Fig B.16 meet assumptions of linearity.

CONSTANT VARIANCE The residual plots in Fig B.15 support a loose random scatter of points around the origin. The Studentized Residual Plot also supports this assumption while highlighting outlying observations. We will address these in later analysis.

NORMALITY Visual analysis of histograms and q-q plots in Fig B.17 suggests normality of the data. Additionally the random cloud scatter plotted in Fig B.15 also support the normality assumption. supports the normality assumption. Left and right leaning trailing tails in the q-q plot does suggest that outliers on both ends may not meet the normality assumption. We will address this in later analysis

INDEPENDENCE Century21 Ames indicated all samples are unique sales in a free housing market, supporting assumptions of independence.

OUTLIERS AND INFLUENTIAL POINTS

OUTLIERS / INFLUENTIAL OBSERVATIONS There are distinct outliers that can be seen both in the Cook's D and Outlier and Leverage Diagnostics seen in Fig B.18; however, these outliers do not have high enough leverage to engender corrective action.

RESULTS The model performance metrics can be seen in Table A.12. The predicted results of the test file obtained a Kaggle score of 0.20306 as can be seen in Table A.13.

4.2.4. Custom

$$\begin{aligned} SalePrice &= 280051 + (28.48)GrLivArea + (37.54)1stFlrSF + \\ &\quad (38.97)2ndFlrSF - (554.60)Age + (86.1448)RemodFactor + \\ &\quad Neighborhood^* + BldgType^* + SaleCondition^* + \\ &\quad OverallQual^* + OverallCond^* + \\ &\quad HouseStyle^* \end{aligned}$$

(* Estimates for categorical variables left out for ease of reading. See Table A.15 for more detail.)

The model summary can be seen in Tables A.14 and A.15. The SAS code for the custom model can be found in Listing 5 of Appendix C

ASSUMPTIONS

The fit diagnostics and residuals by regressors for sale price can be seen in Fig B.19. A scatterplot of the predicted values vs the sale price using the features obtained by the forward selection model can be seen in Fig B.20.

LINEARITY Visual tests of the scatter plots in Fig B.20 meet assumptions of linearity.

CONSTANT VARIANCE The residual plots in Fig B.19 support a loose random scatter of points around the origin. The Studentized Residual Plot also supports this assumption while highlighting outlying observations. We will address these in later analysis.

NORMALITY Visual analysis of histograms and q-q plots in Fig B.21 suggests normality of the data. Additionally the random cloud scatter plotted in Fig B.19 also support the normality assumption. supports the normality assumption. Left and right leaning trailing tails in the q-q plot does suggest that outliers on both ends may not meet the normality assumption. We will address this in later analysis

INDEPENDENCE Century21 Ames indicated all samples are unique sales in a free housing market, supporting assumptions of independence.

OUTLIERS AND INFLUENTIAL POINTS

OUTLIERS / INFLUENTIAL OBSERVATIONS There are distinct outliers that can be seen both in the Cook's D and Outlier and Leverage Diagnostics seen in Fig B.22; however, these outliers do not have high enough leverage to engender corrective action.

RESULTS The model performance metrics can be seen in Table A.16. The predicted results of the test file obtained a Kaggle score of 0.14826 as can be seen in Table A.17.

4.3. Comparing Competing Models

Predictive Models	**Adjusted R^2**	**Press**	**Kaggle Score**
Forward	0.7789	2.37E12	0.20306
Backward	0.7846	1.87E12	0.23930
Stepwise	0.7789	2.37E12	0.20306
CUSTOM	0.8456	1.59E12	0.14826

Table 4.1: Analysis Results

4.4. Conclusion

Model selection methods such as the forward, stepwise, and backward methods enabled model performance to be improved by reducing the number of factors to those of statistical significance. The best results came from modifying the results of these selection methods with domain specific knowledge to generate a more accurate and powerful model. Each model selection method had various strengths as can be seen in Table 4.1 but the most accurate results came from a combination of methods combined with educated choices for model parameters.

A. List of Tables

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	74982.03784	B	5925.64914	12.65	<.0001	63324.70074	86639.37493
GrLivArea100	54.93056	B	4.36399	12.59	<.0001	46.34542	63.51570
Neighborhood BrkSide	-55776.16653	B	12012.18226	-4.64	<.0001	-79407.34256	-32144.99051
Neighborhood Edwards	-30273.99848	B	11296.53114	-2.68	0.0077	-52497.29731	-8050.69965
Neighborhood NAmes	0.00000	B	.	.	.	.	.
GrLivArea*Neighborho BrkSide	33.31303	B	9.33444	3.57	0.0004	14.94970	51.67637
GrLivArea*Neighborho Edwards	8.24334	B	8.45713	0.97	0.3304	-8.39409	24.88077
GrLivArea*Neighborho NAmes	0.00000	B	.	.	.	.	.

Table A.1: Results of Neighborhood Impact on Sales Price.

Root MSE	24094
Dependent Mean	137977
R-Square	0.5333
Adj R-Sq	0.5291
AIC	7037.53791
AICC	7037.72196
BIC	6705.63517
C(p)	4.00000
PRESS	1.961216E11
SBC	6718.75845
ASE	573508240

Table A.2: Results of Neighborhood Impact on Sales Price.

The GLMSELECT Procedure

Forward Selection Summary

Step	Effect Entered	Number Effects In	Model R-Square	Adjusted R-Square	AIC	AICC	BIC	CP	SBC	PRESS	ASE	CV PRESS	F Value	Pr > F
0	Intercept	1	0.0000	0.0000	32291.6565	32291.6653	30916.9518	26023.1178	30923.8798	8.53501E12	6216310560	8.5396E12	0.00	1.0000
1	GrLivArea	2	0.5038	0.5034	31332.8678	31332.8853	29956.6544	12224.7250	29970.3144	4.27855E12	3084477950	4.27422E12	1390.02	<.0001
2	YearBuilt	3	0.6454	0.6449	30874.1151	30874.1444	29496.4868	8347.2758	29516.7850	3.0718E12	2204083788	3.0765E12	546.43	<.0001
3	BsmtQual_Ex	4	0.7095	0.7089	30602.6871	30602.7310	29223.6955	6593.2177	29250.5803	2.53391E12	1805567590	2.54596E12	301.72	<.0001
4	RoofMatl_ClyTile	5	0.7375	0.7367	30466.0929	30466.1545	29085.6893	5830.2421	29119.2094	2.27335E12	1631965597	2.535E12	145.31	<.0001
5	TotalBsmtSF	6	0.7797	0.7789*	30227.3690*	30227.4512*	28845.9016*	4674.2217*	28885.7088*	1.91652E12*	1369166863	2.37142E12*	262.00	<.0001

* Optimal Value of Criterion

Table A.3: Forward Selection Model Summary.

Regression of Sale Price Using Forward Selection Results

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Number of Observations Read	2919
Number of Observations Used	1460
Number of Observations with Missing Values	1459

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	7.204143E12	1.440829E12	1045.51	<.0001
Error	1454	2.003768E12	1378107301		
Corrected Total	1459	9.207911E12			

Root MSE	37123	R-Square	0.7824
Dependent Mean	180921	Adj R-Sq	0.7816
Coeff Var	20.51881		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	-1314593	69671	-18.87	<.0001	-1451261	-1177926
YearBuilt	1	671.56397	35.78464	18.77	<.0001	601.36894	741.75901
TotalBsmtSF	1	45.58624	2.82417	16.14	<.0001	40.04635	51.12613
GrLivArea	1	78.17387	2.11050	37.04	<.0001	74.03393	82.31382
RoofMatl_ClyTile	1	-659375	39300	-16.78	<.0001	-736466	-582285
BsmtQual_EX	1	65879	3973.03277	16.58	<.0001	58086	73673

Table A.4: Forward Selection Model 95% Confidence Limits.

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 5).

Effects:	Intercept YearBuilt TotalBsmtSF GrLivArea RoofMatl_ClyTile BsmtQual_Ex

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	5	6.645434E12	1.329087E12	966.48
Error	1365	1.877128E12	1375185179	
Corrected Total	1370	8.522562E12		

Root MSE	37083
Dependent Mean	185182
R-Square	0.7797
Adj R-Sq	0.7789
AIC	30227
AICC	30227
BIC	28846
C(p)	4674.22175
PRESS	1.916518E12
SBC	28886
ASE	1369166863
CV PRESS	2.371423E12

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	-1271664	73606	-17.28
YearBuilt	1	648.152763	37.758936	17.17
TotalBsmtSF	1	47.048875	2.906695	16.19
GrLivArea	1	80.004257	2.189524	36.54
RoofMatl_ClyTile	1	-671551	39444	-17.03
BsmtQual_Ex	1	62871	4009.622196	15.68

Table A.5: Forward Selection Model Performance.

Your most recent submission

Name	Submitted	Wait time	Execution time	Score
results_fw.csv	just now	2 seconds	0 seconds	0.20306

Complete

Jump to your position on the leaderboard ▾

Table A.6: Forward Selection Model Performance - Kaggle.

Regression of Sale Price Using Backward Selection Results

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Number of Observations Read	2919
Number of Observations Used	1460
Number of Observations with Missing Values	1459

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	7.234662E12	8.038513E11	590.69	<.0001
Error	1450	1.973249E12	1360861622		
Corrected Total	1459	9.207911E12			

Root MSE	36890	R-Square	0.7857
Dependent Mean	180921	Adj R-Sq	0.7844
Coeff Var	20.39001		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	114937	5909.53288	19.45	<.0001	103345	126530
BsmtFinSF1	1	36.21019	2.45403	14.76	<.0001	31.39636	41.02402
_1stFlrSF	1	86.75759	3.29659	26.32	<.0001	80.29098	93.22419
_2ndFlrSF	1	59.78139	2.53313	23.60	<.0001	54.81239	64.75039
OQ7	1	-55133	3365.79499	-16.38	<.0001	-61735	-48531
OQ6	1	-85111	3437.23031	-24.76	<.0001	-91853	-78368
OQ5	1	-99775	3584.89690	-27.83	<.0001	-106807	-92743
OQ4	1	-108450	4789.69091	-22.64	<.0001	-117845	-99054
OQ3	1	-121786	8938.50314	-13.62	<.0001	-139320	-104252
RoofMatl_ClyTile	1	-623167	39508	-15.77	<.0001	-700666	-545667

Table A.7: Backward Selection Model 95% Confidence Limits.

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 581).

Effects:	Intercept BsmtFinSF1 _1stFlrSF _2ndFlrSF OverallQual_3 OverallQual_4 OverallQual_5 OverallQual_6 OverallQual_7 RoofMatl_ClyTile

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	9	6.700172E12	7.444635E11	555.98
Error	1361	1.82239E12	1339008268	
Corrected Total	1370	8.522562E12		

Root MSE	36592
Dependent Mean	185182
R-Square	0.7862
Adj R-Sq	0.7848
AIC	30195
AICC	30195
PRESS	1.869959E12
SBC	28874

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	120177	6171.859488	19.47
BsmtFinSF1	1	35.191793	2.494860	14.11
_1stFlrSF	1	85.189346	3.430542	24.83
_2ndFlrSF	1	59.655197	2.597953	22.96
OverallQual_3	1	-120853	10372	-11.65
OverallQual_4	1	-110889	5142.322601	-21.56
OverallQual_5	1	-101945	3689.034840	-27.63
OverallQual_6	1	-86980	3505.752865	-24.81
OverallQual_7	1	-57950	3401.236685	-17.04
RoofMatl_ClyTile	1	-615180	39362	-15.63

Table A.8: Backward Selection Model Performance.

Your most recent submission

Name	Submitted	Wait time	Execution time	Score
results_bw.csv	just now	0 seconds	0 seconds	0.23930

Complete

Jump to your position on the leaderboard ▾

Table A.9: Backward Selection Model Performance - Kaggle.

The GLMSELECT Procedure

Stepwise Selection Summary

Step	Effect Entered	Effect Removed	Number Effects In	Model R-Square	Adjusted R-Square	AIC	AICC	BIC	CP	SBC	PRESS	ASE	CV PRESS	F Value	Pr > F
0	Intercept		1	0.0000	0.0000	32291.6565	32291.6653	30916.9518	26023.1178	30923.8798	8.53501E12	6216310560	8.5396E12	0.00	1.0000
1	GrLivArea		2	0.5038	0.5034	31332.8678	31332.8853	29956.6544	12224.7250	29970.3144	4.27855E12	3084477950	4.27422E12	1390.02	<.0001
2	YearBuilt		3	0.6454	0.6449	30874.1151	30874.1444	29496.4868	8347.2758	29516.7850	3.0718E12	2204083788	3.0765E12	546.43	<.0001
3	BsmtQual_Ex		4	0.7095	0.7089	30602.6871	30602.7310	29223.6955	6593.2177	29250.5803	2.53391E12	1805567590	2.54596E12	301.72	<.0001
4	RoofMatl_ClyTile		5	0.7375	0.7367	30466.0929	30466.1545	29085.6893	5830.2421	29119.2094	2.27335E12	1631965597	2.535E12	145.31	<.0001
5	TotalBsmtSF		6	0.7797	0.7789*	30227.3690*	30227.4512*	28845.9016*	4674.2217*	28885.7088*	1.91652E12*	1369166863	2.37142E12*	262.00	<.0001

* Optimal Value of Criterion

Table A.10: Stepwise Selection Model Summary.

Regression of Sale Price Using Stepwise Selection Results

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Number of Observations Read	2919
Number of Observations Used	1460
Number of Observations with Missing Values	1459

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	7.204143E12	1.440829E12	1045.51	<.0001
Error	1454	2.003768E12	1378107301		
Corrected Total	1459	9.207911E12			

Root MSE	37123	R-Square	0.7824
Dependent Mean	180921	Adj R-Sq	0.7816
Coeff Var	20.51881		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	-1314593	69671	-18.87	<.0001	-1451261	-1177926
YearBuilt	1	671.56397	35.78464	18.77	<.0001	601.36894	741.75901
TotalBsmtSF	1	45.58624	2.82417	16.14	<.0001	40.04635	51.12613
GrLivArea	1	78.17387	2.11050	37.04	<.0001	74.03393	82.31382
RoofMatl_ClyTile	1	-659375	39300	-16.78	<.0001	-736466	-582285
BsmtQual_EX	1	65879	3973.03277	16.58	<.0001	58086	73673

Table A.11: Stepwise Selection Model 95% Confidence Limits.

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 5).

Effects:	Intercept YearBuilt TotalBsmtSF GrLivArea RoofMatl_ClyTile BsmtQual_Ex

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	5	6.645434E12	1.329087E12	966.48
Error	1365	1.877128E12	1375185179	
Corrected Total	1370	8.522562E12		

Root MSE	37083
Dependent Mean	185182
R-Square	0.7797
Adj R-Sq	0.7789
AIC	30227
AICC	30227
BIC	28846
C(p)	4674.22175
PRESS	1.916518E12
SBC	28886
ASE	1369166863
CV PRESS	2.371423E12

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	-1271664	73606	-17.28
YearBuilt	1	648.152763	37.758936	17.17
TotalBsmtSF	1	47.048875	2.906695	16.19
GrLivArea	1	80.004257	2.189524	36.54
RoofMatl_ClyTile	1	-671551	39444	-17.03
BsmtQual_Ex	1	62871	4009.622196	15.68

Table A.12: Stepwise Selection Model Performance.

Your most recent submission

Name	Submitted	Wait time	Execution time	Score
results_sw.csv	just now	0 seconds	0 seconds	0.20306

Complete

Jump to your position on the leaderboard ▾

Table A.13: Stepwise Selection Model Performance - Kaggle.

The GLM Procedure

Class Level Information

Class	Levels	Values
BsmtQual	5	Ex Fa Gd NA TA
OverallQual	10	1 2 3 4 5 6 7 8 9 10
OverallQual	10	1 2 3 4 5 6 7 8 9 10
OverallCond	9	1 2 3 4 5 6 7 8 9
Neighborhood	25	Blmngtn Blueste BrDale BrkSide ClearCr CollgCr Crawfor Edwards Gilbert IDOTRR MeadowV Mitchel NAmes NPkVill NWAmes NoRidge NridgHt OldTown SWISU Sawyer SawyerW Somerst StoneBr Timber Veenker
BldgType	5	1Fam 2fmCon Duplex Twnhs TwnhsE
SaleCondition	6	Abnorml AdjLand Alloca Family Normal Partial
HouseStyle	8	1.5Fin 1.5Unf 1Story 2.5Fin 2.5Unf 2Story SFoyer SLvl

Table A.14: Custom Model Summary.

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	280051.5334	B	16731.55026	16.74	<.0001	247229.8611	312873.2057
GrLivArea	28.4805		22.27568	1.28	0.2013	-15.2169	72.1779
_1stFlrSF	37.5355		22.48390	1.67	0.0953	-6.5703	81.6414
_2ndFlrSF	38.9688		22.01375	1.77	0.0769	-4.2148	82.1523
Age	-554.6041		79.11740	-7.01	<.0001	-709.8059	-399.4024
Neighborhood Blmngtn	-27921.0696	B	12614.12105	-2.21	0.0270	-52665.7311	-3176.4082
Neighborhood Blueste	-30136.4989	B	24455.60793	-1.23	0.2180	-78110.1736	17837.1757
Neighborhood BrDale	-27897.8279	B	13431.79329	-2.08	0.0380	-54246.4872	-1549.1686
Neighborhood BrkSide	-38284.7780	B	10976.10935	-3.49	0.0005	-59816.2116	-16753.3443
Neighborhood ClearCr	-10017.8246	B	11335.13122	-0.88	0.3770	-32253.5383	12217.8892
Neighborhood CollgCr	-32171.4599	B	10022.62574	-3.21	0.0014	-51832.4795	-12510.4403
Neighborhood Crawfor	-8224.8426	B	10789.78513	-0.76	0.4460	-29390.7708	12941.0856
Neighborhood Edwards	-54540.8116	B	10268.02045	-5.31	<.0001	-74683.2131	-34398.4102
Neighborhood Gilbert	-37181.2191	B	10469.27340	-3.55	0.0004	-57718.4111	-16644.0271
Neighborhood IDOTRR	-47511.5755	B	11523.43169	-4.12	<.0001	-70116.6714	-24906.4796
Neighborhood MeadowV	-42217.2180	B	12999.62383	-3.25	0.0012	-67718.1063	-16716.3298
Neighborhood Mitchel	-37274.1687	B	10638.85321	-3.50	0.0005	-58144.0193	-16404.3182
Neighborhood NAmes	-40699.1158	B	9901.17363	-4.11	<.0001	-60121.8872	-21276.3443
Neighborhood NPkVill	-19221.0611	B	14513.00144	-1.32	0.1856	-47690.6870	9248.5648
Neighborhood NWAmes	-40998.6151	B	10288.77702	-3.98	<.0001	-61181.7340	-20815.4963
Neighborhood NoRidge	19161.1306	B	10983.05174	1.74	0.0813	-2383.9216	40706.1829
Neighborhood NridgHt	6011.2647	B	10491.11626	0.57	0.5667	-14568.7757	26591.3050
Neighborhood OldTown	-51172.5973	B	10674.78366	-4.79	<.0001	-72112.9313	-30232.2633
Neighborhood SWISU	-47071.4102	B	12192.57577	-3.86	0.0001	-70989.1417	-23153.6787
Neighborhood Sawyer	-43902.9445	B	10376.43305	-4.23	<.0001	-64258.0150	-23547.8740
Neighborhood SawyerW	-32493.5672	B	10473.18608	-3.10	0.0020	-53038.4346	-11948.6998
Neighborhood Somerst	-19868.6343	B	10384.72232	-1.91	0.0559	-40239.9656	502.6969
Neighborhood StoneBr	20168.8349	B	11608.63619	1.74	0.0825	-2603.4036	42941.0734
Neighborhood Timber	-19026.7762	B	10867.50062	-1.75	0.0802	-40345.1560	2291.6037
Neighborhood Veenker	0.0000	B	.	.	.	.	.
BldgType 1Fam	22937.9043	B	4097.83269	5.60	<.0001	14899.3353	30976.4733
BldgType 2fmCon	20754.2946	B	7308.48497	2.84	0.0046	6417.5060	35091.0832

Table A.15: Custom Model 95% Confidence Limits.

Sum of Residuals	-2.857996E-8
Sum of Squared Residuals	1.3610971E12
Sum of Squared Residuals - Error SS	-0.001220703
PRESS Statistic	1.591527E12
First Order Autocorrelation	0.0370548532
Durbin-Watson D	1.9257432676

Observations	1460
Parameters	63
Error DF	1397
MSE	9.74E8
R-Square	0.8522
Adj R-Square	0.8456

Table A.16: Custom Model Performance.

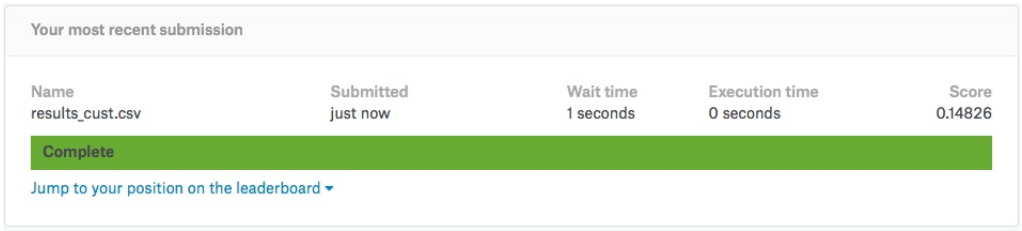

Table A.17: Custom Selection Model Performance - Kaggle.

B. List of Figures

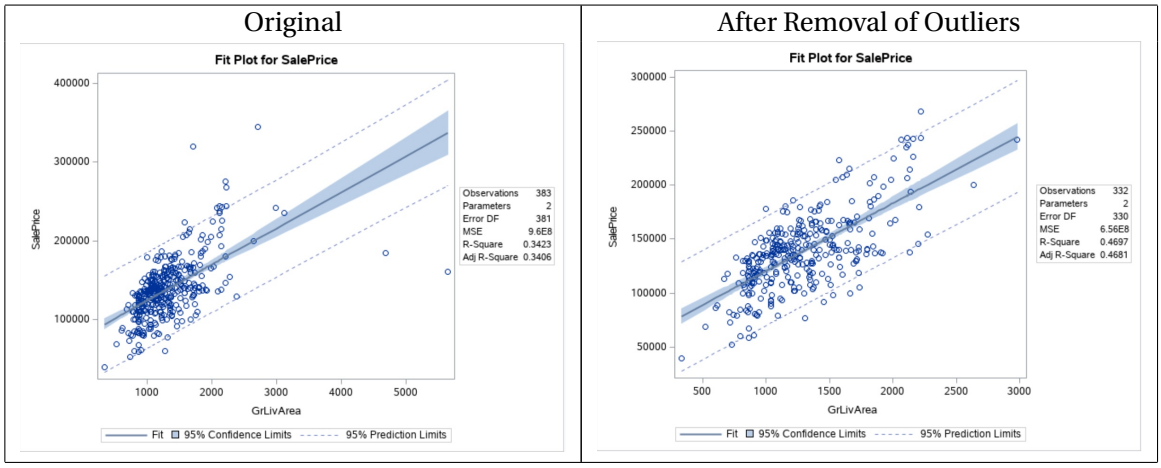

Figure B.1: Simple Linear Regression Models.

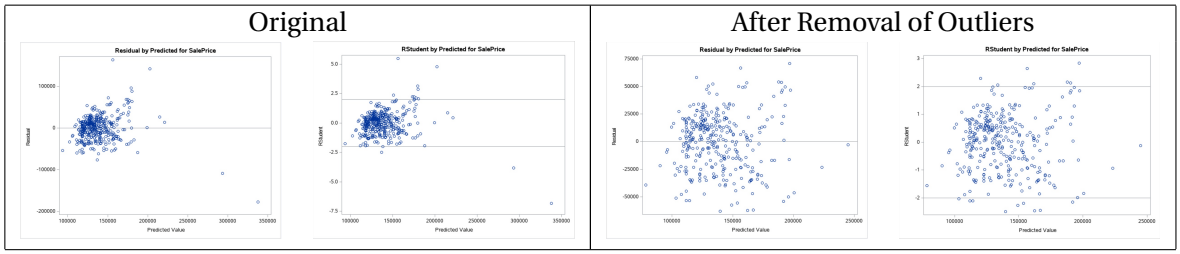

Figure B.2: Residual Plots.

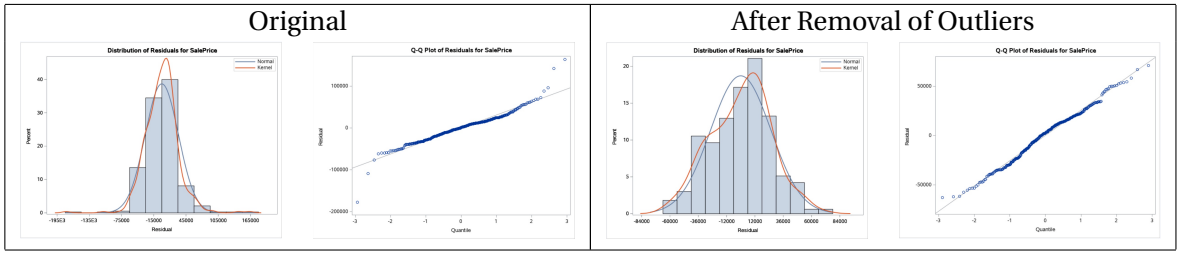

Figure B.3: Normality Plots.

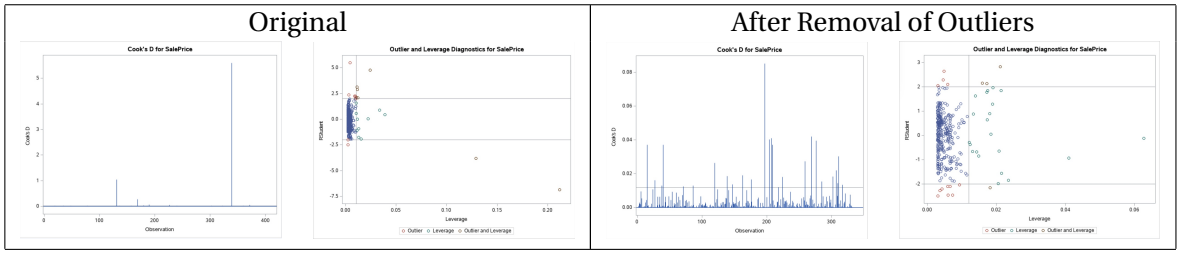

Figure B.4: Influential Point Plots.

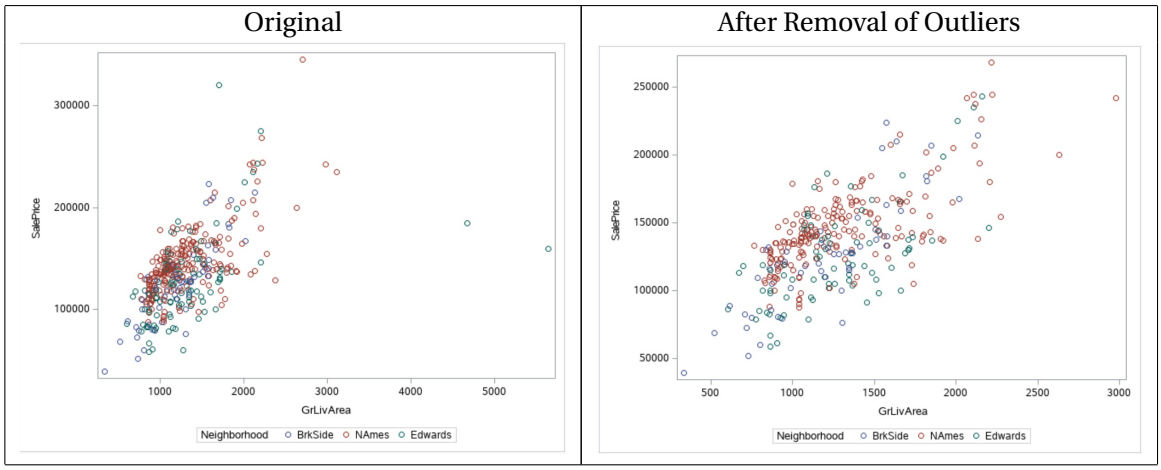

Figure B.5: Scatterplot of Sale Prices vs Living Area by Neighborhood.

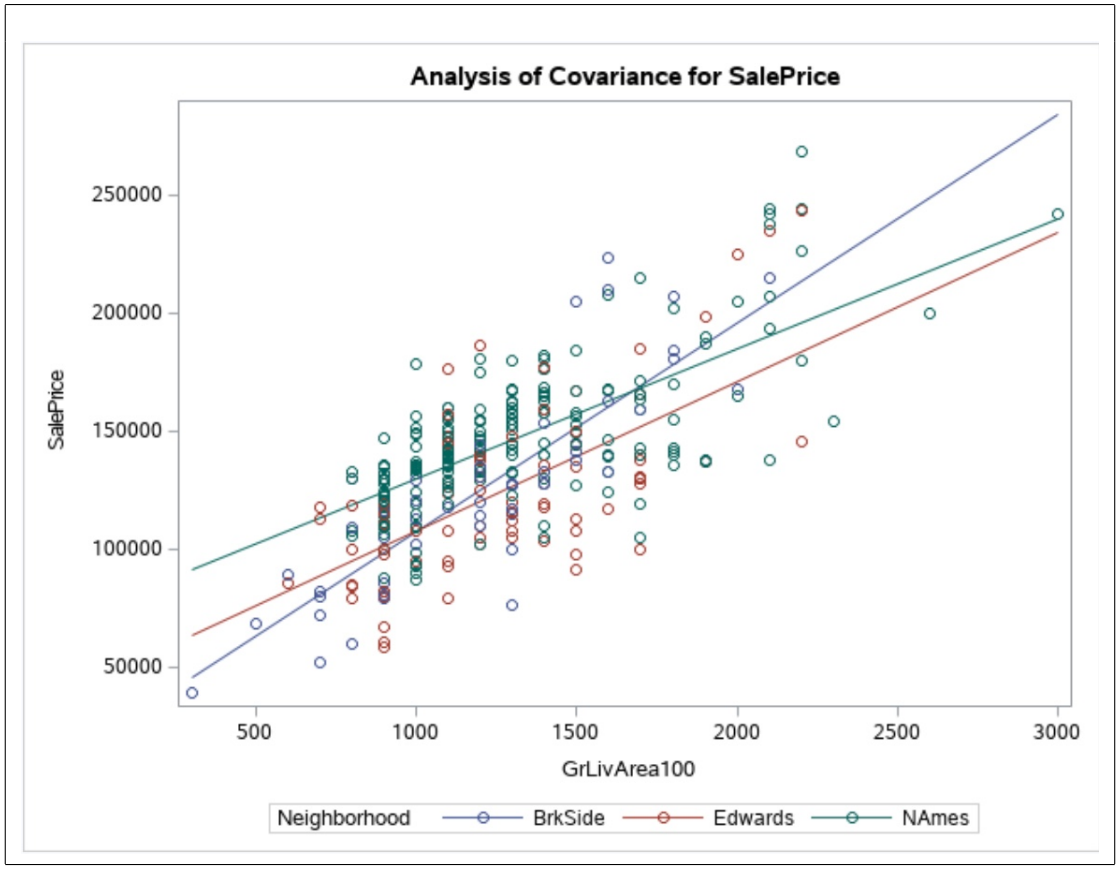

Figure B.6: Linear Regression Model by Neighborhood.

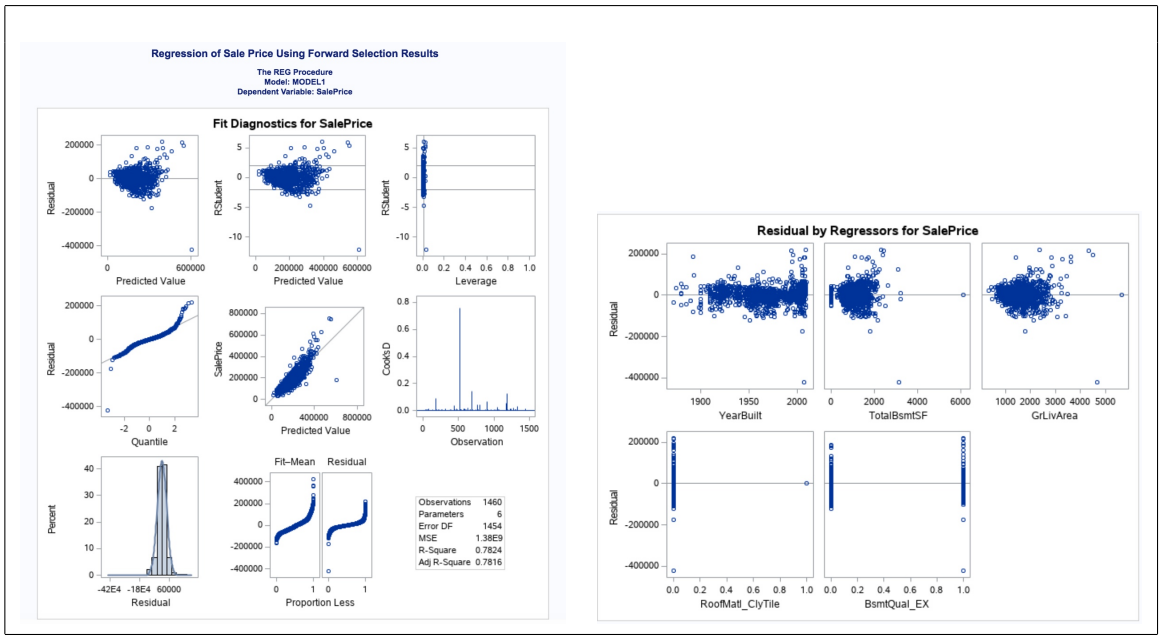

Figure B.7: Plots for Forward Selection Model.

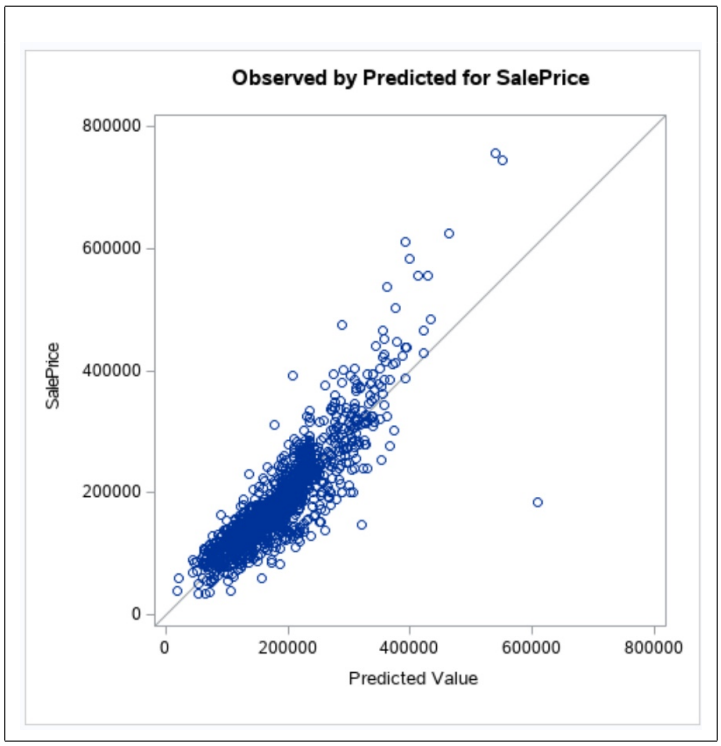

Figure B.8: Scatter Plot Forward Selection Model.

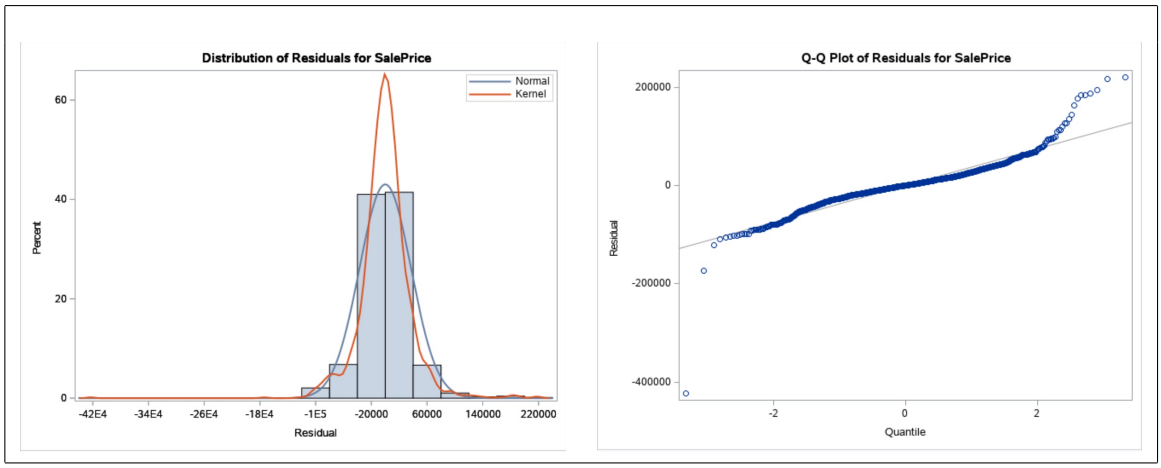

Figure B.9: Normality Plots for Forward Selection Model.

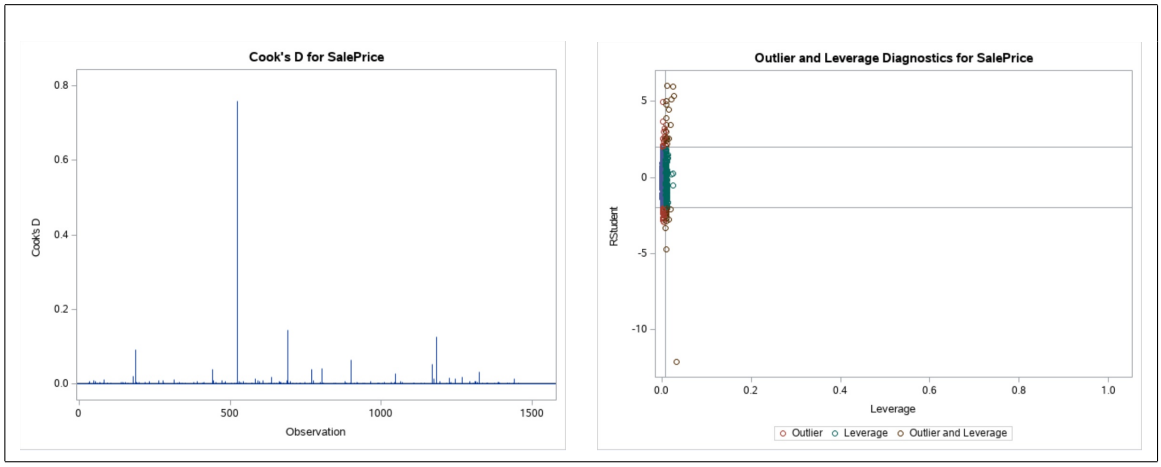

Figure B.10: Influential Point Plots for Forward Selection Model.

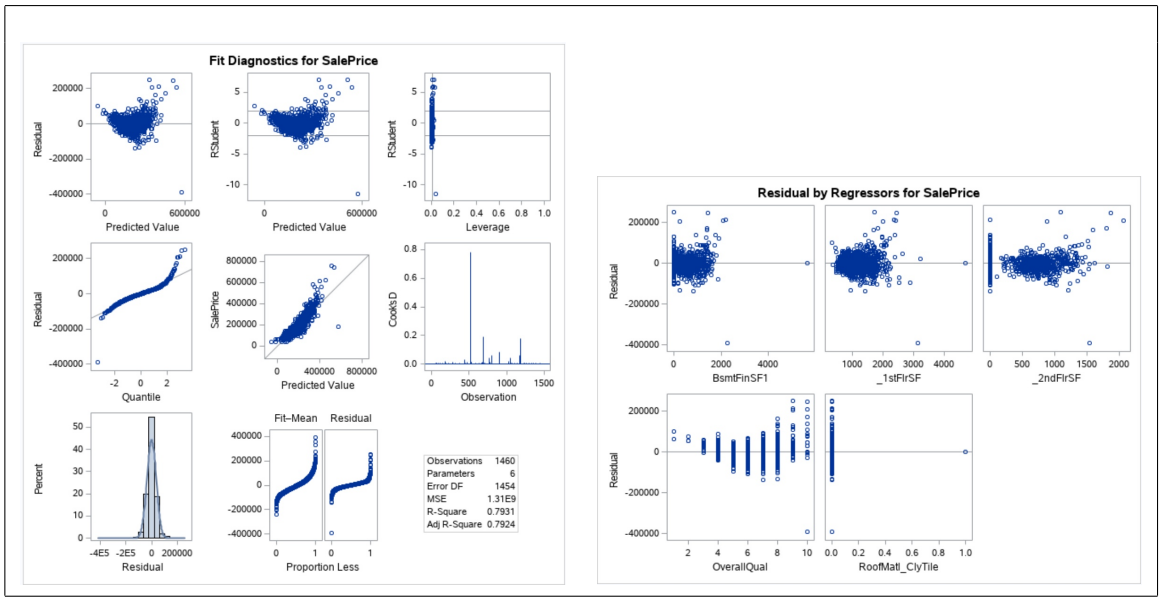

Figure B.11: Plots for Backward Selection Model.

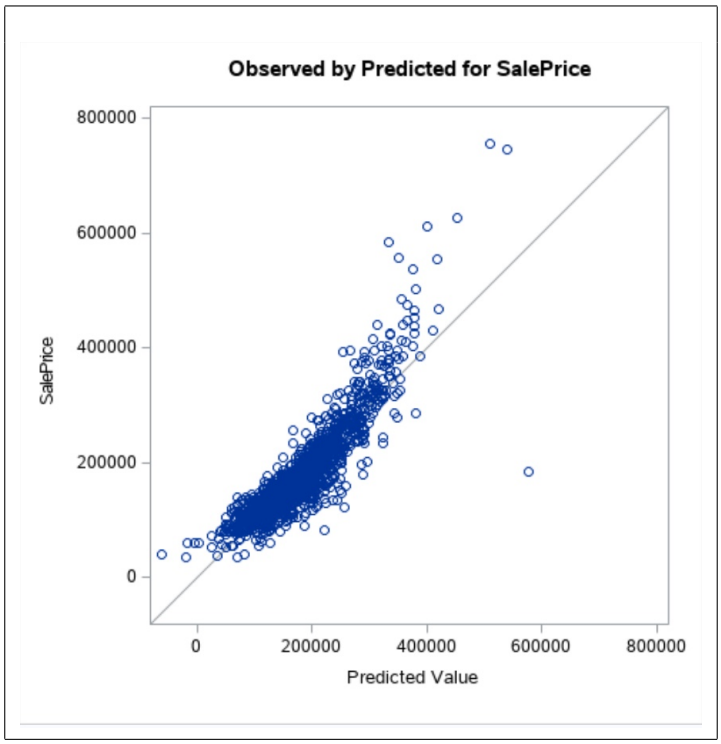

Figure B.12: Scatter Plot Backward Selection Model.

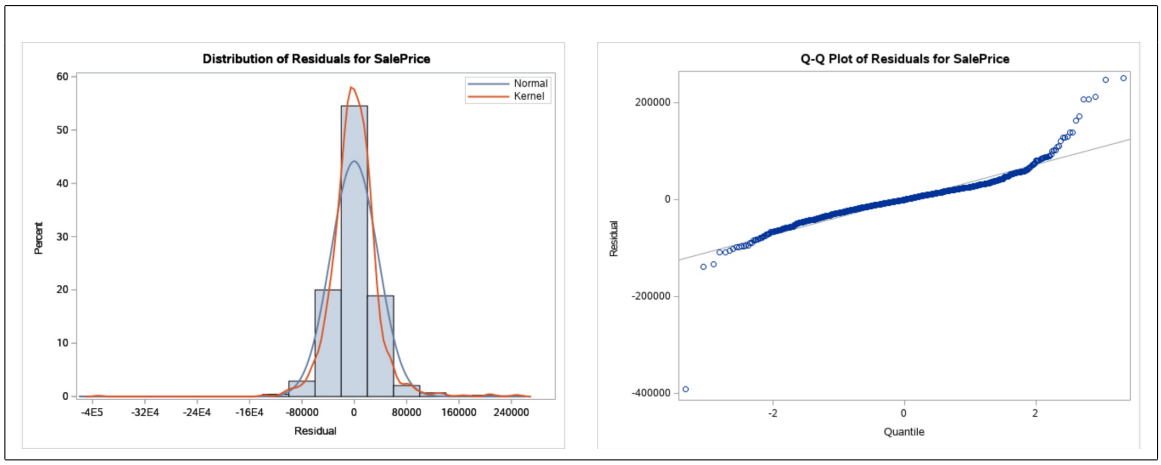

Figure B.13: Normality Plots for Backward Selection Model.

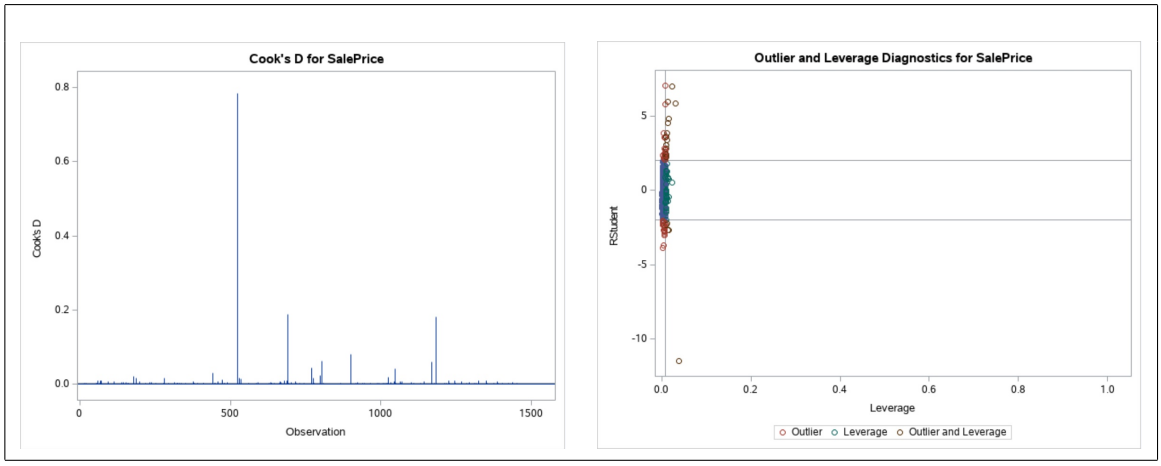

Figure B.14: Influential Point Plots for Backward Selection Model.

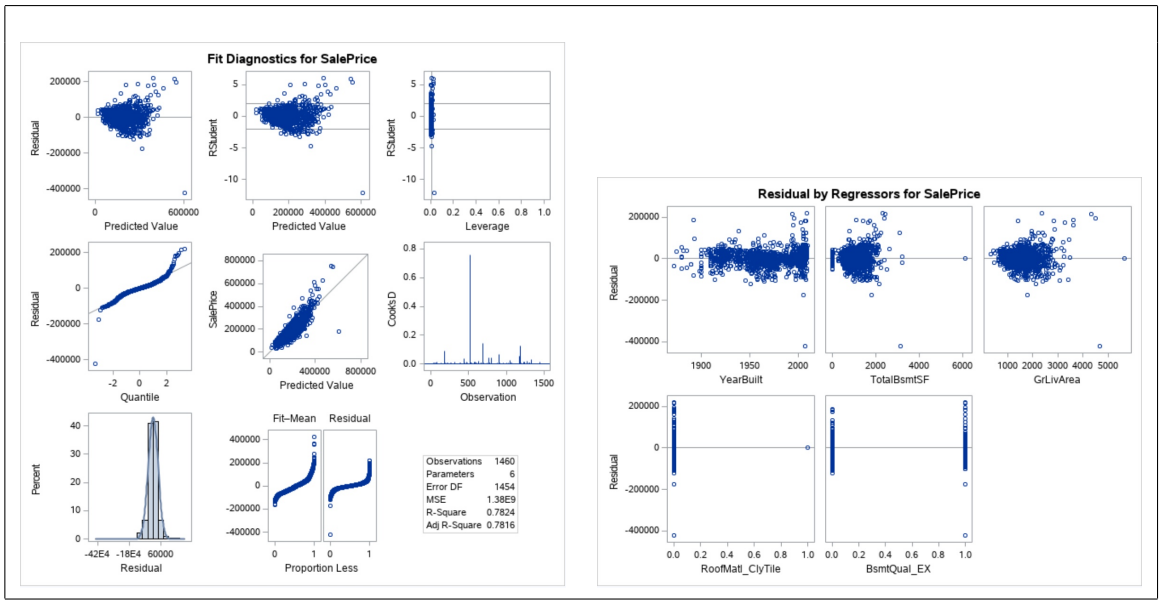

Figure B.15: Plots for Stepwise Selection Model.

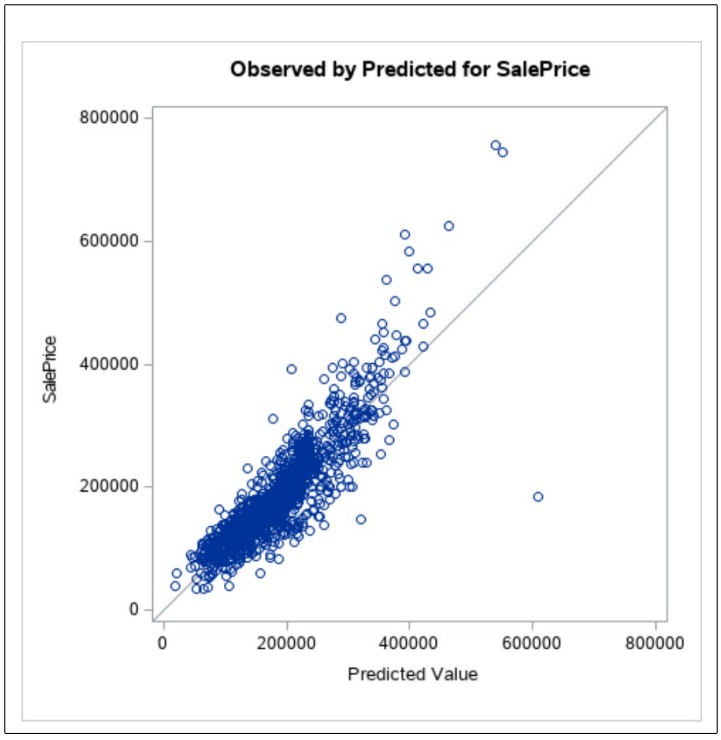

Figure B.16: Scatter Plot Stepwise Selection Model.

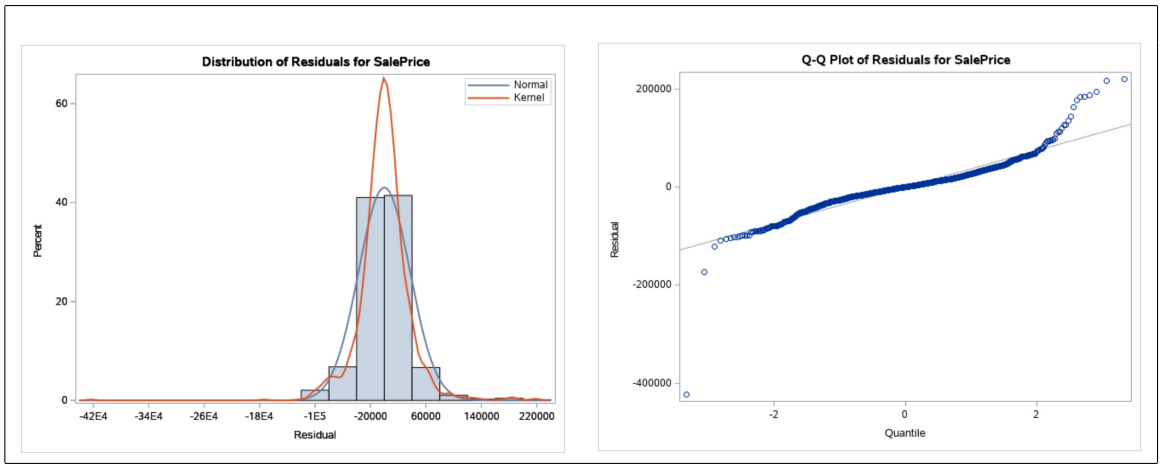

Figure B.17: Normality Plots for Stepwise Selection Model.

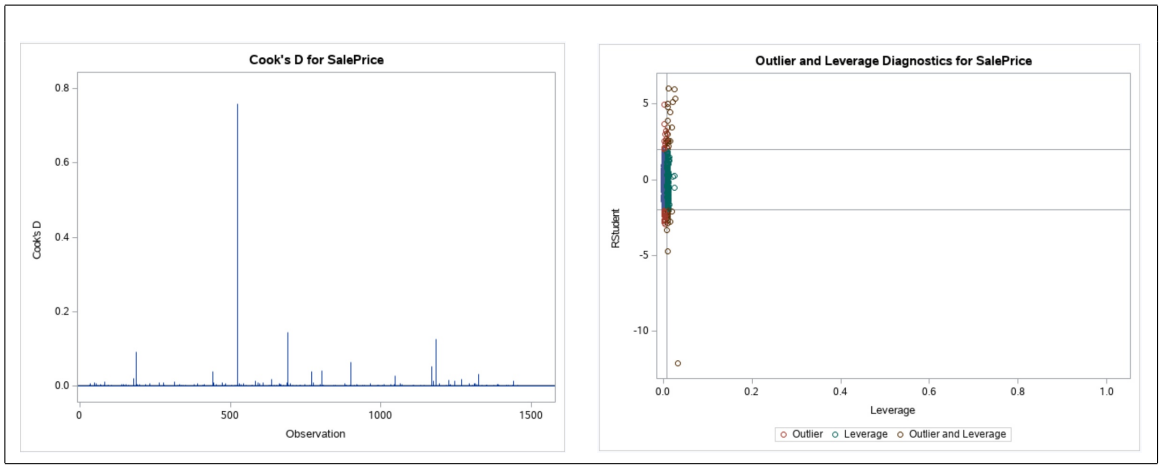

Figure B.18: Influential Point Plots for Stepwise Selection Model.

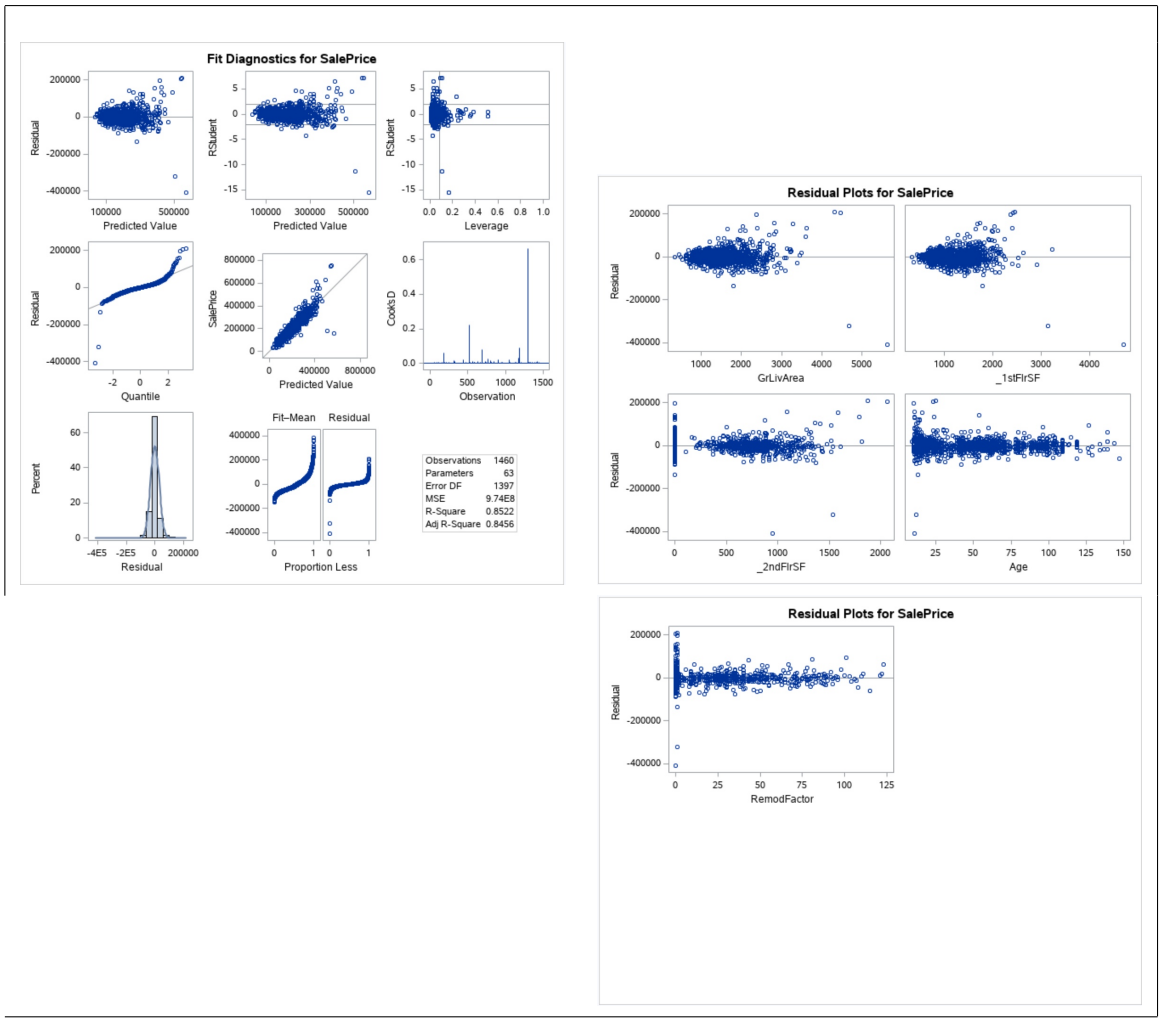

Figure B.19: Plots for Custom Selection Model.

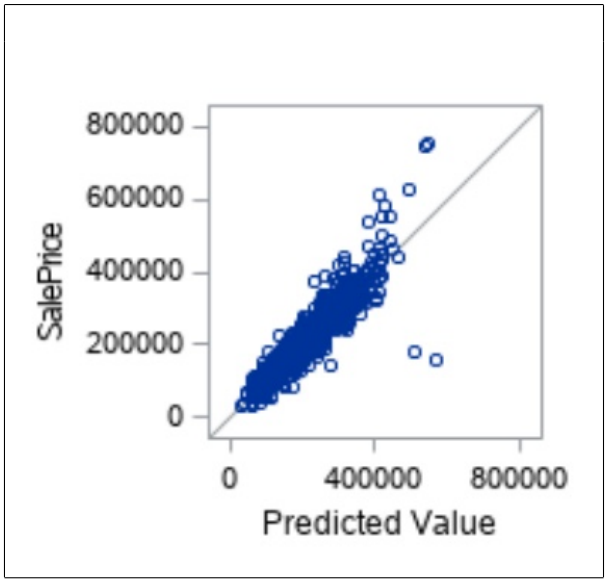

Figure B.20: Scatter Plot Custom Model.

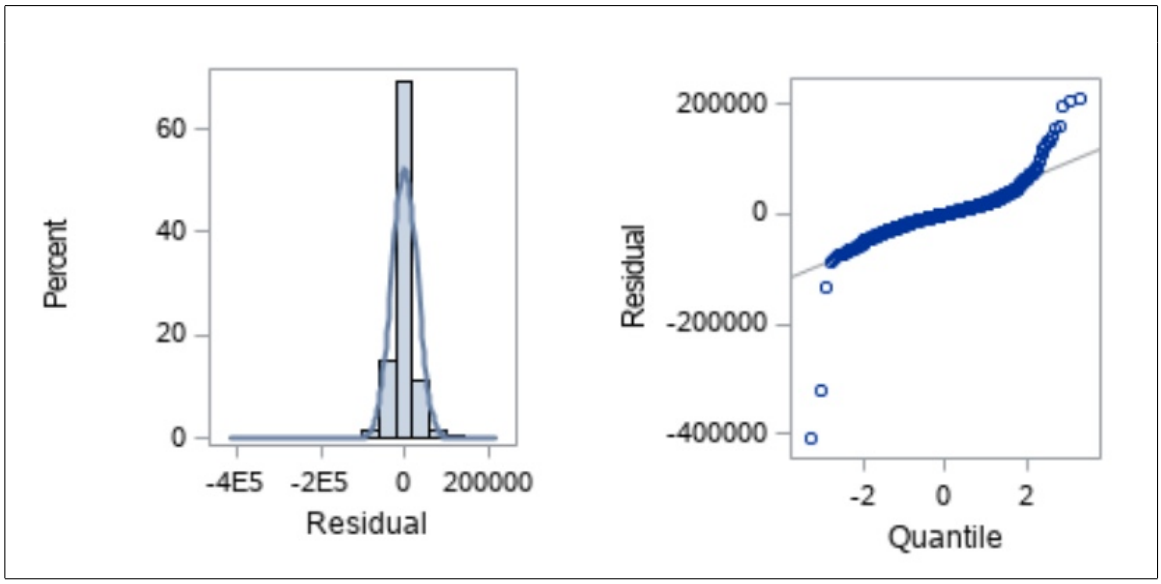

Figure B.21: Normality Plots for Custom Selection Model.

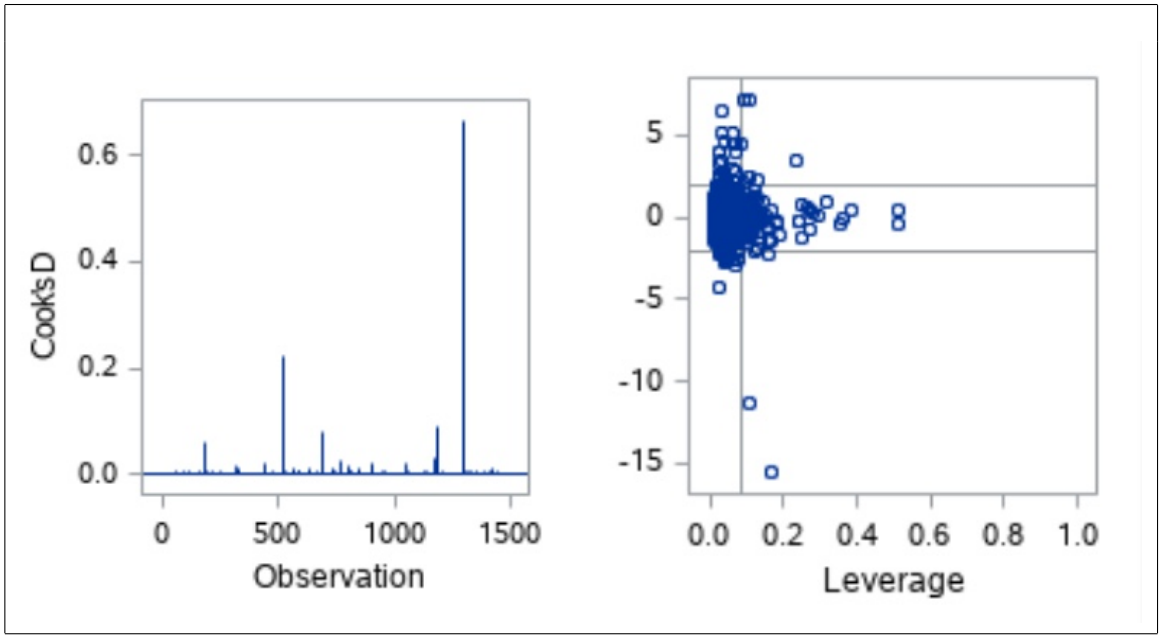

Figure B.22: Influential Point Plots for Custom Selection Model.

C. Source Code (SAS) for Analysis

C.1. Analysis 1

Listing 1: Analysis 1 SAS Code (Analysis1.sas).

```
*-------------------------------------------------------------*
| Import train.csv                                            |
| Set REFFILE for train.csv                                   |
*-------------------------------------------------------------*;

* FILENAME REFFILE '/home/mwolfe0/train.csv';
FILENAME REFFILE
'/folders/myfolders/MSDS6371/GroupProject/Datasets/train.csv';

PROC IMPORT DATAFILE=REFFILE DBMS=CSV REPLACE OUT=TRAIN;
        GETNAMES=YES;
RUN;

*-------------------------------------------------------------*
| Subset the data to only include homes sold in the           |
| neighborhoods of interest - NAmes, BrkSide, and Edwards|
| Round the gross living area to the nearest 100 SF           |
| Keep only the variables of Neighboorhood, GrLivArea,        |
| and SalePrice in the dataset                                |
*-------------------------------------------------------------*;

DATA HOMES1;
SET TRAIN (KEEP=Neighborhood GrLivArea SalePrice);
IF Neighborhood EQ "NAmes" |
   Neighborhood EQ "BrkSide" |
   Neighborhood EQ "Edwards";
GrLivArea100 = ROUND(GrLivArea, 100); /*FLOOR(GrLivArea);*/
RUN;

*-------------------------------------------------------------*
| Descriptive statistics on the HOMES1 dataset for            |
| GrLivArea and SalePrice                                     |
*-------------------------------------------------------------*;

PROC UNIVARIATE DATA=HOMES1;
        CLASS Neighborhood;
        VAR GrLivArea SalePrice;
RUN;

*-------------------------------------------------------------*
| Scatter plot of sale prices in the three neighborhoods |
| vs Gross Living Area                                        |
*-------------------------------------------------------------*;

PROC SGPLOT DATA=HOMES1;
```

```
46          SCATTER X=GrLivArea Y=SalePrice/GROUP=Neighborhood;
47  RUN;
48
49  *-------------------------------------------------------------*
50  | Regression model of homes in the three neighborhoods       |
51  | combined for Sale Price based on Gross Living Area         |
52  | to check assumptions on the data in these three            |
53  | neighborhoods                                              |
54  *-------------------------------------------------------------*;
55
56  PROC REG DATA=HOMES1 PLOTS=ALL;
57          MODEL SalePrice=GrLivArea / CLB;
58          RUN;
59
60  *-------------------------------------------------------------*
61  | Regression model of homes in the three neighborhoods       |
62  | using an equal slope model                                 |
63  *-------------------------------------------------------------*;
64
65  PROC GLM DATA=HOMES1;
66          CLASS Neighborhood;
67          MODEL SalePrice=GrLivArea Neighborhood / CLPARM;
68          RUN;
69
70  *-------------------------------------------------------------*
71  | Regression model of homes in the three neighborhoods       |
72  | using an equal intercept model (slopes differ)             |
73  *-------------------------------------------------------------*;
74
75  PROC GLM DATA=HOMES1;
76          CLASS Neighborhood;
77          MODEL SalePrice=GrLivArea*Neighborhood / CLPARM;
78          RUN;
79
80  *-------------------------------------------------------------*
81  | Regression model of homes in the three neighborhoods       |
82  | using a model that allows slopes and intercepts to         |
83  | vary                                                       |
84  *-------------------------------------------------------------*;
85
86  PROC GLM DATA=HOMES1;
87          CLASS Neighborhood;
88          MODEL SalePrice=Neighborhood GrLivArea*Neighborhood / CLPARM;
89          RUN;
90
91  *-------------------------------------------------------------*
92  | Remove Outliers:                                           |
93  | SaleCondition is not normal (confounding effect on         |
94  | prices)                                                    |
95  | SalePrice is greater than 300,000 since they are not       |
```

```
96  | representative of the overall population in these      |
97  | three neighborhoods.                                   |
98  | Keep only the variables of Neighboorhood, GrLivArea,   |
99  | and SalePrice in the dataset                           |
100 *--------------------------------------------------------*;
101
102 DATA HOMES2;
103 SET TRAIN (KEEP=Neighborhood GrLivArea SalePrice SaleCondition);
104 IF Neighborhood EQ "NAmes" |
105    Neighborhood EQ "BrkSide" |
106    Neighborhood EQ "Edwards";
107 IF SalePrice LT 300000;
108 IF SaleCondition EQ "Normal";
109 GrLivArea100 = ROUND(GrLivArea, 100);
110 RUN;
111
112 *--------------------------------------------------------*
113 | Descriptive statistics on the HOMES1 dataset for       |
114 | GrLivArea and SalePrice                                |
115 *--------------------------------------------------------*;
116
117 PROC UNIVARIATE DATA=HOMES2;
118         CLASS Neighborhood;
119         VAR GrLivArea SalePrice;
120 RUN;
121
122 *--------------------------------------------------------*
123 | Scatter plot of sale prices in the three neighborhoods |
124 | vs Gross Living Area                                   |
125 *--------------------------------------------------------*;
126
127 PROC SGPLOT DATA=HOMES2;
128         SCATTER X=GrLivArea Y=SalePrice;
129         REG X=GrLivArea Y=SalePrice;
130 RUN;
131
132 PROC SGPLOT DATA=HOMES2;
133         SCATTER X=GrLivArea Y=SalePrice/GROUP=Neighborhood;
134 RUN;
135
136 *--------------------------------------------------------*
137 | Regression model of homes in the three neighborhoods   |
138 | using a model that allows slopes and intercepts to     |
139 | vary                                                   |
140 | Output 95% confidence limit for parameter estimates    |
141 *--------------------------------------------------------*;
142
143 PROC REG DATA=HOMES2 PLOTS=ALL;
144         MODEL SalePrice=GrLivArea / CLB;
145         RUN;
```

```
146
147 *-------------------------------------------------------------*
148 | Regression model of homes in the three neighborhoods        |
149 | using an equal slope model                                  |
150 *-------------------------------------------------------------*;
151
152 PROC GLM DATA=HOMES2;
153         CLASS Neighborhood;
154         MODEL SalePrice=GrLivArea Neighborhood / CLPARM;
155         RUN;
156
157 *-------------------------------------------------------------*
158 | Regression model of homes in the three neighborhoods        |
159 | using an equal intercept model (slopes differ)              |
160 *-------------------------------------------------------------*;
161
162 PROC GLM DATA=HOMES2;
163         CLASS Neighborhood;
164         MODEL SalePrice=GrLivArea*Neighborhood / CLPARM;
165         RUN;
166
167 *-------------------------------------------------------------*
168 | Regression model of homes in the three neighborhoods        |
169 | using a model that allows slopes and intercepts to          |
170 | vary                                                        |
171 *-------------------------------------------------------------*;
172
173 PROC GLM DATA=HOMES2;
174         CLASS Neighborhood;
175         MODEL SalePrice=Neighborhood GrLivArea*Neighborhood / CLPARM;
176         RUN;
177
178
179 *-------------------------------------------------------------*
180 | Alternate Method with interaction terms                     |
181 |                                                             |
182 | Keep only the variables of Neighboorhood, GrLivArea,        |
183 | and SalePrice in the dataset                                |
184 | d1 = NAmes, d2 = BrkSide, Control = Edwards                 |
185 *-------------------------------------------------------------*;
186
187 DATA HOMES3;
188 SET TRAIN (KEEP=Neighborhood GrLivArea SalePrice SaleCondition);
189 IF Neighborhood EQ "NAmes" |
190    Neighborhood EQ "BrkSide" |
191    Neighborhood EQ "Edwards";
192 IF SalePrice LT 300000;
193 IF SaleCondition EQ "Normal";
194 GrLivArea100 = ROUND(GrLivArea, 100);
195   IF Neighborhood = 'NAmes' THEN d1 = 1; ELSE d1=0;
```

```
196    IF Neighborhood = 'BrkSide' THEN d2 = 1; ELSE d2=0;
197                  int1 = d1*GrLivArea100; int2 = d2*GrLivArea100;
198 RUN;
199
200 *-------------------------------------------------------------*
201 | Plots to check assumptions                                  |
202 | d1 = NAmes, d2 = BrkSide, Control = Edwards                 |
203 *-------------------------------------------------------------*;
204
205 PROC SGPLOT DATA=HOMES3;
206 HISTOGRAM GrLivArea100;
207 DENSITY GrLivArea100/TYPE=NORMAL;
208 TITLE "Histogram of Gross Living Area in NAmes, BrkSide, and Edwards";
209 RUN;
210
211 PROC SGPLOT DATA=HOMES3;
212 SCATTER X=GrLivArea100 Y=SalePrice;
213 TITLE "Gross Living Area vs Sale Price in NAmes, BrkSide, and Edwards";
214 RUN;
215
216 PROC REG DATA=HOMES3 PLOT=ALL;
217 model SalePrice = GrLivArea100/CLB;
218 RUN;
219
220 *-------------------------------------------------------------*
221 | Run regression model with interaction terms using dummy|
222 | variables                                                   |
223 | d1 = NAmes, d2 = BrkSide, Control = Edwards                 |
224 | Output 95% confidence limit for parameter estimates         |
225 *-------------------------------------------------------------*;
226 PROC REG DATA=HOMES3;
227          model SalePrice = GrLivArea100 d1 d2 int1 int2/VIF CLB;
228          title
229          'Regression of Sale Price on Gross Living Area
230          with Interaction Terms';
231          RUN;
232
233 *-------------------------------------------------------------*
234 | center the interaction terms based on the means of          |
235 | GrLivArea100 and d1 and d2 to correct for the               |
236 | inflated VIF                                                |
237 *-------------------------------------------------------------*;
238
239 PROC MEANS DATA=HOMES3;
240 var GrLivArea100 d1 d2;
241 run;
242
243 DATA center;
244 set HOMES3;
245 cent1 = (GrLivArea100 - 1280.72)*(d1-0.593);
```

```
246 cent2 = (GrLivArea100 - 1280.72)*(d2-0.164);
247 RUN;
248
249 PROC REG DATA=center PLOTS=ALL;
250 model SalePrice = GrLivArea100 d1 d2 cent1 cent2/VIF CLB;
251 title
252         'Regression of Sale Price on Gross Living Area
253         with Interaction Terms';
254 RUN;
255
256 PROC GLM DATA=HOMES3 PLOT=ALL;
257 CLASS Neighborhood;
258 model SalePrice=GrLivArea100|Neighborhood/solution CLPARM;
259 RUN;
260
261 PROC GLMSELECT DATA=HOMES3;
262 CLASS Neighborhood;
263 MODEL SalePrice = GrLivArea Neighborhood
264 / cvmethod=random(5) stats=all;
265 RUN;
```

C.2. Analysis 2

C.2.1. Forward Selection

Listing 2: Forward Selection SAS Code (Forward.sas).

```
*-------------------------------------------------------------*
| Import train.csv                                            |
| Import test.csv                                             |
| Set REFFILE for train.csv                                   |
| Set REFFILE2 for test.csv                                   |
*-------------------------------------------------------------*;

*FILENAME REFFILE '/home/mwolfe0/train.csv';
*FILENAME REFFILE2 '/home/mwolfe0/test.csv';
FILENAME REFFILE
'/folders/myfolders/MSDS6371/GroupProject/Datasets/train.csv';
FILENAME REFFILE2
'/folders/myfolders/MSDS6371/GroupProject/Datasets/test.csv';

PROC IMPORT DATAFILE=REFFILE DBMS=CSV REPLACE OUT=TRAIN;
        GETNAMES=YES;
RUN;

PROC IMPORT DATAFILE=REFFILE2 DBMS=CSV REPLACE OUT=TEST;
        GETNAMES=YES;
RUN;

*-------------------------------------------------------------*
| Combine train and test into one datafile HOMES              |
*-------------------------------------------------------------*;

DATA HOMES;
        SET TRAIN TEST;
        IF LotFrontage EQ "NA" THEN LotFrontage = 0;
        LotFront = input(LotFrontage, 8.);
        drop LotFrontage;
        RENAME LotFront=LotFrontage;
RUN;

*-------------------------------------------------------------*
| Code for forward selection                                  |
| Set seed to a constant for model comparison                 |
| Class variable input with split option to allow             |
|       classification variable to be able to enter or        |
|       leave the model independently                         |
| Stop=CV specifies the model will stop when the              |
|       predicted residual sum of square is reached with      |
|       k-fold cross validation                               |
| CVMethod specifies how subsets ar formed for                |
|       cross validation                                      |
| OUTPUT Dataset to RESULTS with the predicted variable       |
```

```
47 |          based on the final model                              |
48 *----------------------------------------------------------------*;
49
50 PROC GLMSELECT DATA=HOMES SEED=71669132;
51          CLASS MSSubClass MSZoning Street Alley LotShape LandContour
52                      Utilities LotConfig LandSlope Neighborhood Condition1
53                      Condition2 BldgType HouseStyle OverallQual OverallCond
54                      RoofMatl Exterior1st Exterior2nd MasVnrArea ExterQual
55                      ExterCond Foundation BsmtQual BsmtExposure BsmtFinType1
56                      BsmtFinType2 Heating HeatingQC CentralAir Electrical
57                      KitchenQual Functional FireplaceQu GarageType
58                      GarageFinish GarageQual GarageCond PavedDrive PoolQC
59                      Fence MiscFeature SaleType SaleCondition RoofStyle
60                      BsmtCond MasVnrType
61                      / split;
62          MODEL SalePrice= LotArea YearBuilt YearRemodAdd BsmtFinSF1
63                      BsmtFinSF2 BsmtUnfSF TotalBsmtSF _1stFlrSF _2ndFlrSF
64                      LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath
65                      FullBath HalfBath BedroomAbvGr KitchenAbvGr
66                      TotRmsAbvGrd Fireplaces GarageYrBlt GarageCars
67                      GarageArea WoodDeckSF OpenPorchSF EnclosedPorch
68                      _3SsnPorch ScreenPorch PoolArea MiscVal MoSold YrSold
69                      MSSubClass MSZoning Street Alley LotShape LandContour
70                      Utilities LotConfig LandSlope Neighborhood Condition1
71                      Condition2 BldgType HouseStyle OverallQual OverallCond
72                      RoofMatl Exterior1st Exterior2nd MasVnrArea ExterQual
73                      ExterCond Foundation BsmtQual BsmtExposure BsmtFinType1
74                      BsmtFinType2 Heating HeatingQC CentralAir Electrical
75                      KitchenQual Functional FireplaceQu GarageType
76                      GarageFinish GarageQual GarageCond PavedDrive PoolQC
77                      Fence MiscFeature SaleType SaleCondition LotFrontage
78                      RoofStyle BsmtCond MasVnrType
79          / selection =forward(stop=CV) cvmethod=random(5) stats=all;
80          OUTPUT OUT=RESULTS P=PREDICT;
81 RUN;
82
83 *----------------------------------------------------------------*
84 | Create a datafile RESULTS_FW of predicted values for          |
85 | SalePrice for house id greater than 1460 which                |
86 | is where the Kaggle test set data begins.                     |
87 *----------------------------------------------------------------*;
88
89 DATA RESULTS_FW;
90          SET RESULTS;
91
92          IF PREDICT > 0 THEN
93                  SalePrice=Predict;
94
95          IF PREDICT < 0 THEN
96                  SalePrice=10000;
```

```
        KEEP id SalePrice;
        WHERE id > 1460;
RUN;

*-------------------------------------------------------------*
| Export a datafile for predicted values for                  |
| SalePrice for house id greater than 1460 which              |
| is where the Kaggle test set data begins.                   |
*-------------------------------------------------------------*;

*FILENAME REFFILE3 '/home/mwolfe0/results_fw.csv';
FILENAME REFFILE3
'/folders/myfolders/MSDS6371/GroupProject/Datasets/results_fw.csv';

PROC EXPORT DATA=RESULTS_FW FILE=REFFILE3 DBMS=CSV REPLACE;
RUN;

*-------------------------------------------------------------*
| Check model assumptions.                                    |
*-------------------------------------------------------------*;

DATA HOMES2;
SET HOMES (KEEP=YearBuilt TotalBsmtSF GrLivArea RoofMatl BsmtQual
                SalePrice);
        IF RoofMatl EQ "ClyTile" THEN RoofMatl_ClyTile=1;
        ELSE RoofMatl_ClyTile=0;
        IF BsmtQual = 'Ex' THEN BsmtQual_EX = 1;
        ELSE BsmtQual_EX=0;
RUN;

PROC REG DATA=HOMES2 PLOT=ALL;
        model SalePrice = YearBuilt TotalBsmtSF GrLivArea
        RoofMatl_ClyTile BsmtQual_EX/CLB;
        title
        'Regression of Sale Price Using Forward Selection Results';
        RUN;

PROC GLM DATA=HOMES2 PLOT=ALL;
CLASS RoofMatl_ BsmtQual;
MODEL SalePrice = YearBuilt TotalBsmtSF GrLivArea|
        RoofMatl_ClyTile BsmtQual_EX/solution CLPARM;
RUN;
```

Listing 3: Backward Selection SAS Code (Backward.sas).

```
*-------------------------------------------------------------*
| Import train.csv                                            |
| Import test.csv                                             |
| Set REFFILE for train.csv                                   |
| Set REFFILE2 for test.csv                                   |
*-------------------------------------------------------------*;

*FILENAME REFFILE '/home/mwolfe0/train.csv';
*FILENAME REFFILE2 '/home/mwolfe0/test.csv';
FILENAME REFFILE
'/folders/myfolders/MSDS6371/GroupProject/Datasets/train.csv';
FILENAME REFFILE2
'/folders/myfolders/MSDS6371/GroupProject/Datasets/test.csv';

PROC IMPORT DATAFILE=REFFILE DBMS=CSV REPLACE OUT=TRAIN;
        GETNAMES=YES;
RUN;

PROC IMPORT DATAFILE=REFFILE2 DBMS=CSV REPLACE OUT=TEST;
        GETNAMES=YES;
RUN;

*-------------------------------------------------------------*
| Combine train and test into one datafile HOMES              |
*-------------------------------------------------------------*;

DATA HOMES;
        SET TRAIN TEST;
        IF LotFrontage EQ "NA" THEN LotFrontage = 0;
        LotFront = input(LotFrontage, 8.);
        drop LotFrontage;
        RENAME LotFront=LotFrontage;
RUN;

*-------------------------------------------------------------*
| Code for backward selection                                 |
| Set seed to a constant for model comparison                 |
| Class variable input with split option to allow             |
|       classification variable to be able to enter or        |
|       leave the model independently                         |
| Stop=10 specifies the model will stop selection at the      |
|       first step for which the selected model has 10        |
|       effects                                               |
| CVMethod specifies how subsets ar formed for                |
|       cross validation                                      |
| OUTPUT Dataset to RESULTS with the predicted variable       |
|        based on the final model                             |
```

```
*-------------------------------------------------------------*;

PROC GLMSELECT DATA=HOMES SEED=71669132;
        CLASS MSSubClass MSZoning Street Alley LotShape LandContour
                Utilities LotConfig LandSlope Neighborhood Condition1
                Condition2 BldgType HouseStyle OverallQual OverallCond
                RoofMatl Exterior1st Exterior2nd MasVnrArea ExterQual
                ExterCond Foundation BsmtQual BsmtExposure BsmtFinType1
                BsmtFinType2 Heating HeatingQC CentralAir Electrical
                KitchenQual Functional FireplaceQu GarageType
                GarageFinish GarageQual GarageCond PavedDrive PoolQC
                Fence MiscFeature SaleType SaleCondition RoofStyle
                BsmtCond MasVnrType
                / split;
        MODEL SalePrice= LotArea YearBuilt YearRemodAdd BsmtFinSF1
                BsmtFinSF2 BsmtUnfSF TotalBsmtSF _1stFlrSF _2ndFlrSF
                LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath
                FullBath HalfBath BedroomAbvGr KitchenAbvGr
                TotRmsAbvGrd Fireplaces GarageYrBlt GarageCars
                GarageArea WoodDeckSF OpenPorchSF EnclosedPorch
                _3SsnPorch ScreenPorch PoolArea MiscVal MoSold YrSold
                MSSubClass MSZoning Street Alley LotShape LandContour
                Utilities LotConfig LandSlope Neighborhood Condition1
                Condition2 BldgType HouseStyle OverallQual OverallCond
                RoofMatl Exterior1st Exterior2nd MasVnrArea ExterQual
                ExterCond Foundation BsmtQual BsmtExposure BsmtFinType1
                BsmtFinType2 Heating HeatingQC CentralAir Electrical
                KitchenQual Functional FireplaceQu GarageType
                GarageFinish GarageQual GarageCond PavedDrive PoolQC
                Fence MiscFeature SaleType SaleCondition LotFrontage
                RoofStyle BsmtCond MasVnrType
        / selection =backward(stop=10) cvmethod=random(5)
                        stats=ADJRSQ stats=PRESS;
        OUTPUT OUT=RESULTS P=PREDICT;
RUN;

*-------------------------------------------------------------*
| Create a datafile RESULTS_BW of predicted values for      |
| SalePrice for house id greater than 1460 which            |
| is where the Kaggle test set data begins.                 |
*-------------------------------------------------------------*;

DATA RESULTS_BW;
        SET RESULTS;

        IF PREDICT > 0 THEN
                SalePrice=Predict;

        IF PREDICT < 0 THEN
                SalePrice=10000;
```

```
        KEEP id SalePrice;
        WHERE id > 1460;
RUN;

*-------------------------------------------------------------*
| Export a datafile for predicted values for                  |
| SalePrice for house id greater than 1460 which              |
| is where the Kaggle test set data begins.                   |
*-------------------------------------------------------------*;

*FILENAME REFFILE3 '/home/mwolfe0/results_bw.csv';
FILENAME REFFILE3
'/folders/myfolders/MSDS6371/GroupProject/Datasets/results_bw.csv';

PROC EXPORT DATA=RESULTS_BW FILE=REFFILE3 DBMS=CSV REPLACE;
RUN;

*-------------------------------------------------------------*
| Check model assumptions.                                    |
*-------------------------------------------------------------*;


DATA HOMES2;
SET HOMES (KEEP=BsmtFinSF1 _1stFlrSF _2ndFlrSF OverallQual RoofMatl
                SalePrice);
        IF RoofMatl EQ "ClyTile" THEN RoofMatl_ClyTile=1;
        ELSE RoofMatl_ClyTile=0;
        IF OverallQual EQ 7 THEN OQ7=1; ELSE OQ7=0;
        IF OverallQual EQ 6 THEN OQ6=1; ELSE OQ6=0;
        IF OverallQual EQ 5 THEN OQ5=1; ELSE OQ5=0;
        IF OverallQual EQ 4 THEN OQ4=1; ELSE OQ4=0;
        IF OverallQual EQ 3 THEN OQ3=1; ELSE OQ3=0;
RUN;

PROC REG DATA=HOMES2 PLOT=ALL;
        model SalePrice = BsmtFinSF1 _1stFlrSF _2ndFlrSF
        OQ7 OQ6 OQ5 OQ4 OQ3 RoofMatl_ClyTile/CLB;
        title
        'Regression of Sale Price Using Backward Selection Results';
        RUN;

PROC GLM DATA=HOMES2 PLOT=ALL;
CLASS OverallQual RoofMatl;
MODEL SalePrice = BsmtFinSF1 _1stFlrSF _2ndFlrSF OverallQual|
        RoofMatl_ClyTile/solution CLPARM;
RUN;
```

Listing 4: Stepwise Selection SAS Code (Stepwise.sas).

```
*-------------------------------------------------------------*
| Import train.csv                                            |
| Import test.csv                                             |
| Set REFFILE for train.csv                                   |
| Set REFFILE2 for test.csv                                   |
*-------------------------------------------------------------*;

*FILENAME REFFILE '/home/mwolfe0/train.csv';
*FILENAME REFFILE2 '/home/mwolfe0/test.csv';
FILENAME REFFILE
'/folders/myfolders/MSDS6371/GroupProject/Datasets/train.csv';
FILENAME REFFILE2
'/folders/myfolders/MSDS6371/GroupProject/Datasets/test.csv';

PROC IMPORT DATAFILE=REFFILE DBMS=CSV REPLACE OUT=TRAIN;
        GETNAMES=YES;
RUN;

PROC IMPORT DATAFILE=REFFILE2 DBMS=CSV REPLACE OUT=TEST;
        GETNAMES=YES;
RUN;

*-------------------------------------------------------------*
| Combine train and test into one datafile HOMES              |
*-------------------------------------------------------------*;

DATA HOMES;
        SET TRAIN TEST;
        IF LotFrontage EQ "NA" THEN LotFrontage = 0;
        LotFront = input(LotFrontage, 8.);
        drop LotFrontage;
        RENAME LotFront=LotFrontage;
RUN;

*-------------------------------------------------------------*
| Code for stepwise selection                                 |
| Set seed to a constant for model comparison                 |
| Class variable input with split option to allow             |
|       classification variable to be able to enter or        |
|       leave the model independently                         |
| Stop=CV specifies the model will stop when the              |
|       predicted residual sum of square is reached with      |
|       k-fold cross validation                               |
| CVMethod specifies how subsets ar formed for                |
|       cross validation                                      |
| OUTPUT Dataset to RESULTS with the predicted variable       |
|        based on the final model                             |
```

```
*-------------------------------------------------------------*;

PROC GLMSELECT DATA=HOMES SEED=71669132;
        CLASS MSSubClass MSZoning Street Alley LotShape LandContour
                Utilities LotConfig LandSlope Neighborhood Condition1
                Condition2 BldgType HouseStyle OverallQual OverallCond
                RoofMatl Exterior1st Exterior2nd MasVnrArea ExterQual
                ExterCond Foundation BsmtQual BsmtExposure BsmtFinType1
                BsmtFinType2 Heating HeatingQC CentralAir Electrical
                KitchenQual Functional FireplaceQu GarageType
                GarageFinish GarageQual GarageCond PavedDrive PoolQC
                Fence MiscFeature SaleType SaleCondition RoofStyle
                BsmtCond MasVnrType
                / split;
        MODEL SalePrice= LotArea YearBuilt YearRemodAdd BsmtFinSF1
                BsmtFinSF2 BsmtUnfSF TotalBsmtSF _1stFlrSF _2ndFlrSF
                LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath
                FullBath HalfBath BedroomAbvGr KitchenAbvGr
                TotRmsAbvGrd Fireplaces GarageYrBlt GarageCars
                GarageArea WoodDeckSF OpenPorchSF EnclosedPorch
                _3SsnPorch ScreenPorch PoolArea MiscVal MoSold YrSold
                MSSubClass MSZoning Street Alley LotShape LandContour
                Utilities LotConfig LandSlope Neighborhood Condition1
                Condition2 BldgType HouseStyle OverallQual OverallCond
                RoofMatl Exterior1st Exterior2nd MasVnrArea ExterQual
                ExterCond Foundation BsmtQual BsmtExposure BsmtFinType1
                BsmtFinType2 Heating HeatingQC CentralAir Electrical
                KitchenQual Functional FireplaceQu GarageType
                GarageFinish GarageQual GarageCond PavedDrive PoolQC
                Fence MiscFeature SaleType SaleCondition LotFrontage
                RoofStyle BsmtCond MasVnrType
        / selection=stepwise(stop=CV) cvmethod=random(5) stats=all;
        OUTPUT OUT=RESULTS P=PREDICT;
RUN;

*-------------------------------------------------------------*
| Create a datafile RESULTS_SW of predicted values for      |
| SalePrice for house id greater than 1460 which            |
| is where the Kaggle test set data begins.                 |
*-------------------------------------------------------------*;

DATA RESULTS_SW;
        SET RESULTS;

        IF PREDICT > 0 THEN
                SalePrice=Predict;

        IF PREDICT < 0 THEN
                SalePrice=10000;
        KEEP id SalePrice;
```

```
        WHERE id > 1460;
RUN;

*-------------------------------------------------------------*
| Export a datafile for predicted values for                  |
| SalePrice for house id greater than 1460 which              |
| is where the Kaggle test set data begins.                   |
*-------------------------------------------------------------*;

*FILENAME REFFILE3 '/home/mwolfe0/results_sw.csv';
FILENAME REFFILE3
'/folders/myfolders/MSDS6371/GroupProject/Datasets/results_sw.csv';

PROC EXPORT DATA=RESULTS_SW FILE=REFFILE3 DBMS=CSV REPLACE;
RUN;

*-------------------------------------------------------------*
| Check model assumptions.                                    |
*-------------------------------------------------------------*;

DATA HOMES2;
SET HOMES (KEEP=YearBuilt TotalBsmtSF GrLivArea RoofMatl BsmtQual
                SalePrice);
        IF RoofMatl EQ "ClyTile" THEN RoofMatl_ClyTile=1;
        ELSE RoofMatl_ClyTile=0;
        IF BsmtQual = 'Ex' THEN BsmtQual_EX = 1;
        ELSE BsmtQual_EX=0;
RUN;

PROC REG DATA=HOMES2 PLOT=ALL;
        model SalePrice = YearBuilt TotalBsmtSF GrLivArea
        RoofMatl_ClyTile BsmtQual_EX/CLB;
        title
        'Regression of Sale Price Using Stepwise Selection Results';
        RUN;

PROC GLM DATA=HOMES2 PLOT=ALL;
CLASS RoofMatl_ BsmtQual;
MODEL SalePrice = YearBuilt TotalBsmtSF GrLivArea|
        RoofMatl_ClyTile BsmtQual_EX/solution CLPARM;
RUN;
```

C.2.4. Custom Model

Listing 5: Custom Model SAS Code (Carls_Custom.sas).

```
*-------------------------------------------------------------*
| Import train.csv                                            |
| Import test.csv                                             |
| Set REFFILE for train.csv                                   |
| Set REFFILE2 for test.csv                                   |
*-------------------------------------------------------------*;
 FILENAME REFFILE
  '/folders/myfolders/MSDS6371/GroupProject/Datasets/train.csv';

/* FILENAME REFFILE */
/* "C:\Users\cwale\OneDrive\Desktop\SMU\Winter18\StatsFoundation\
Case_Study\Data\train.csv"; */

/*FILENAME REFFILE2 */
/* "C:\Users\cwale\OneDrive\Desktop\SMU\Winter18\StatsFoundation\
Case_Study\Data\test.csv" */

FILENAME REFFILE2
'/folders/myfolders/MSDS6371/GroupProject/Datasets/test.csv';


PROC IMPORT OUT= WORK.train
            DATAFILE= REFFILE
            DBMS=CSV REPLACE;
     GETNAMES=YES;
     DATAROW=2;
RUN;

PROC IMPORT OUT= WORK.test
            DATAFILE= REFFILE2
            DBMS=CSV REPLACE;
     GETNAMES=YES;
     DATAROW=2;
RUN;


/*PROC IMPORT DATAFILE=REFFILE DBMS=CSV REPLACE OUT=TRAIN;*/
/*      GETNAMES=YES;*/
/*RUN;*/
/**/
/*PROC IMPORT DATAFILE=REFFILE2 DBMS=CSV REPLACE OUT=TEST;*/
/*      GETNAMES=YES;*/
/*RUN;*/

DATA HOMES;
        SET TRAIN TEST;
        Age=2019-YearBuilt;
```

```
        RemodFactor = YearRemodAdd - YearBuilt;
        Impression=OverallQual + OverallCond/2;
RUN;

*-------------------------------------------------------------*
| Code for custom GLM model                                   |
| NOTE: SAS online "_1st" will not run must change to         |
|       first same for _2nd                                   |
| Code works as is in SAS University Edition                  |
*-------------------------------------------------------------*;
ods graphics on;
PROC GLM DATA=HOMES PLOTS=ALL;
        CLASS BsmtQual OverallQual OverallQual OverallCond
        Neighborhood BldgType SaleCondition HouseStyle;
        MODEL SalePrice = GrLivArea _1stFlrSF _2ndFlrSF Age
        Neighborhood BldgType SaleCondition RemodFactor
        OverallQual OverallCond HouseStyle/ p clparm
                TOLERANCE SOLUTION;
        OUTPUT OUT=RESULTS P=PREDICT;
        RUN;
ods graphics off;
/* .148 - GrLivArea _1stFlrSF _2ndFlrSF Age Neighborhood
BldgType SaleCondition LotArea RemodFactor OverallQual
OverallCond HouseStyle */


DATA RESULTS_CUST;
        SET RESULTS;

        IF PREDICT > 0 THEN
                SalePrice=Predict;

        IF PREDICT < 0 THEN
                SalePrice=10000;
        KEEP id SalePrice;
        WHERE id > 1460;
RUN;

FILENAME REFFILE3
'/folders/myfolders/MSDS6371/GroupProject/Datasets/results_cust.csv';
/* FILENAME REFFILE3
C:\Users\cwale\OneDrive\Desktop\results_cust.csv */

PROC EXPORT DATA=RESULTS_CUST FILE=REFFILE3 DBMS=CSV
                REPLACE;
RUN;
```

D. Summary of Data

Attribute Type		Description	Features
Categorical (Qualitative)	Nominal	Only provide enough information to distinguish one object from another	MSSubClass, MSZoning, Street, Alley, LotShape, LandContour, Utilities, LotConfig, LandSlope, Neighborhood, Condition1, Condition2, BldgType, HouseStyle, RoofMatl, Exterior1st, Exterior2nd, MasVnrType, Foundation, BsmtExposure, Heating CentralAir, Electrical, GarageType, GarageFinish PavedDrive, MiscFeature, SaleType, SaleCondition RoofStyle **(30 variables)**
	Ordinal	Provide enough information to order objects	OverallQual, OverallCond, ExterQual, ExterCond, BsmtQual, BsmtCond, BsmtFinType1, BsmtFinType2, HeatingQC, KitchenQual, Functional, FireplaceQu, GarageQual, GarageCond, PoolQC, Fence **(16 variables)**
Numeric (Quantitative)	Interval	Interval attributes difference between values are meaningful	YearBuilt, YearRemodAdd, GarageYrBlt, MoSold, YrSold **(5 variables)**
	Ratio	Differences and ratios are meaningful	LotFrontage, LotArea, BsmtFinSF1, BsmtFinSF2, BsmtUnfSF, TotalBsmtSF, 1stFlrSF, 2ndFlrSF, LowQualFinSF, GrLivArea, BsmtFullBath, MasVnrArea, BsmtHalfBath, FullBath, HalfBath, Bedroom, Kitchen, TotRmsAbvGrd, Fireplaces, GarageCars, GarageArea, WoodDeckSF, OpenPorchSF, EnclosedPorch, 3SsnPorch, ScreenPorch, PoolArea, MiscVal **(28 variables)**

Table D.1: Summary of Dataset

Classification Variables	
MSSubClass	Identifies the type of dwelling involved in the sale
	20 1-STORY 1946 & NEWER ALL STYLES
	30 1-STORY 1945 & OLDER
	40 1-STORY W/FINISHED ATTIC ALL AGES
	45 1-1/2 STORY - UNFINISHED ALL AGES
	50 1-1/2 STORY FINISHED ALL AGES
	60 2-STORY 1946 & NEWER
	70 2-STORY 1945 & OLDER
	75 2-1/2 STORY ALL AGES
	80 SPLIT OR MULTI-LEVEL
	85 SPLIT FOYER
	90 DUPLEX - ALL STYLES AND AGES
	120 1-STORY PUD (Planned Unit Development) - 1946 & NEWER
	150 1-1/2 STORY PUD - ALL AGES
	160 2-STORY PUD - 1946 & NEWER
	180 PUD - MULTILEVEL - INCL SPLIT LEV/FOYER
	190 2 FAMILY CONVERSION - ALL STYLES AND AGES
MSZoning	Identifies the general zoning classification of the sale
	A Agriculture
	C Commercial
	FV Floating Village Residential
	I Industrial
	RH Residential High Density
	RL Residential Low Density
	RP Residential Low Density Park
	RM Residential Medium Density
LotFrontage	Linear feet of street connected to property
LotArea	Lot size in square feet
Street	Type of road access to property
	Grvl Gravel
	Pave Paved
Alley	Type of alley access to property
	Grvl Gravel
	Pave Paved
	NA No alley access
LotShape	General shape of property
	Reg Regular
	IR1 Slightly Irregular
	IR2 Moderately Irregular
	IR3 Irregular
LandContour	General shape of property
	Lvl Near Flat/Level
	Bnk Banked - Quick and significant rise from street grade to building
	HLS Hillside - Significant slope from side to side
	Low Depression

Classification Variables		
Utilities	Identifies the type of dwelling involved in the sale	
	AllPub	All public Utilities (E,G,W,& S)
	NoSewr	Electricity, Gas, and Water (Septic Tank)
	NoSeWa	Electricity and Gas Only
	ELO	Electricity only
LotConfig	Lot configuration	
	Inside	Inside lot
	Corner	Corner lot
	CulDSac	Cul-de-sac
	FR2	Frontage on 2 sides of property
	FR3	Frontage on 3 sides of property
LandSlope	Slope of property	
	Gtl	Gentle slope
	Mod	Moderate slope
	Sev	Severe slope
Neighborhood	Physical locations within Ames city limits	
	Blmgtn	Bloomington Heights
	Blueste	Bluestem
	BrDale	Briardale
	BrkSide	Brookside
	ClearCr	Clear Creek
	CollgCr	College Creek
	Crawfor	Crawford
	Edwards	Edwards
	Gilbert	Gilbert
	IDOTRR	Iowa DOT and Rail Road
	MeadowV	Meadow Village
	Mitchel	Mitchell
	NAmes	North Ames
	NoRidge	Northridge
	NPkVill	Northpark Villa
	NridgHt	Northridge Heights
	NWAmes	Northwest Ames
	OldTown	Old Town
	SWISU	South & West of Iowa State University
	Sawyer	Sawyer
	SawyerW	Sawyer West
	Somerst	Somerset
	StoneBr	Stone Brook
	Timber	Timberland
	Veenker	Veenker
Condition1	Proximity to various conditions	
	Artery	Adjacent to arterial street
	Feedr	Adjacent to feeder street
	Norm	Normal
	RRNn	Within 200' of North-South Railroad
	RRAn	Adjacent to North-South Railroad
	PosN	Near positive off-site feature–park, greenbelt, etc.
	PosA	Adjacent to postive off-site feature
	RRNe	Within 200' of East-West Railroad
	RRAe	Adjacent to East-West Railroad

Classification Variables		
Condition2	Proximity to various conditions (if more than one is present)	
	Artery	Adjacent to arterial street
	Feedr	Adjacent to feeder street
	Norm	Normal
	RRNn	Within 200' of North-South Railroad
	RRAn	Adjacent to North-South Railroad
	PosN	Near positive off-site feature–park, greenbelt, etc.
	PosA	Adjacent to postive off-site feature
	RRNe	Within 200' of East-West Railroad
	RRAe	Adjacent to East-West Railroad
BldgType	Type of dwelling	
	1Fam	Adjacent to arterial street
	2FmCon	Adjacent to feeder street
	Duplx	Normal
	TwnhsE	Normal
	TwnhsI	Normal
HouseStyle	Style of dwelling	
	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
OverallQual	Rates the overall material and finish of the house	
	10	Very Excellent
	9	Excellent
	8	Very Good
	7	Good
	6	Above Average
	5	Average
	4	Below Average
	3	Fair
	2	Poor
	1	Very Poor
OverallQual	Rates the overall material and finish of the house	
	10	Very Excellent
	9	Excellent
	8	Very Good
	7	Good
	6	Above Average
	5	Average
	4	Below Average
	3	Fair
	2	Poor
	1	Very Poor

Classification Variables		
YearBuilt	Original construction date	
YearRemodAdd	Remodel date (same as construction date if no remodeling or additions)	
RoofStyle	Type of roof	
	Flat	Flat
	Gable	Gable
	Gambrel	Gabrel (Barn)
	Hip	Hip
	Mansard	Mansard
	Shed	Shed
RoofMatl	Roof material	
	ClyTile	Clay or Tile
	CompShg	Standard (Composite) Shingle
	Membran	Membrane
	Metal	Metal
	Roll	Roll
	Tar&Grv	Gravel & Tar
	WdShake	Wood Shakes
	WdShngl	Wood Shingles
Exterior1st	Exterior covering on house	
	AsbShng	Asbestos Shingles
	AsphShn	Asphalt Shingles
	BrkComm	Brick Common
	BrkFace	Brick Face
	CBlock	Cinder Block
	CemntBd	Cement Board
	HdBoard	Hard Board
	ImStucc	Imitation Stucco
	MetalSd	Metal Siding
	Other	Other
	Plywood	Plywood
	PreCast	PreCast
	Stone	Stone
	Stucco	Stucco
	VinylSd	Vinyl Siding
	Wd Sdng	Wood Siding
	WdShing	Wood Shingles

Classification Variables		
Exterior2nd	Exterior covering on house (if more than one material)	
	AsbShng	Asbestos Shingles
	AsphShn	Asphalt Shingles
	BrkComm	Brick Common
	BrkFace	Brick Face
	CBlock	Cinder Block
	CemntBd	Cement Board
	HdBoard	Hard Board
	ImStucc	Imitation Stucco
	MetalSd	Metal Siding
	Other	Other
	Plywood	Plywood
	PreCast	PreCast
	Stone	Stone
	Stucco	Stucco
	VinylSd	Vinyl Siding
	Wd Sdng	Wood Siding
	WdShing	Wood Shingles
MasVnrType	Masonry veneer type	
	BrkCmn	Brick Common
	BrkFace	Brick Face
	CBlock	Cinder Block
	None	None
	Stone	Stone
MasVnrArea	Masonry veneer area in square feet	
ExterQual	Evaluates the quality of the material on the exterior	
	Ex	Excellent
	Gd	Good
	TA	Average/Typical
	Fa	Fair
	Po	Poor
ExterCond	Evaluates the present condition of the material on the exterior	
	Ex	Excellent
	Gd	Good
	TA	Average/Typical
	Fa	Fair
	Po	Poor

Foundation	Type of foundation	
	BrkTil	Brick & Tile
	CBlock	Cinder Block
	PConc	Poured Concrete
	Slab	Slab
	Stone	Stone
	Wood	Wood
BsmtQual	Evaluates the height of the basement	
	Ex	Excellent (100+ inches)
	Gd	Good (90-99 inches)
	TA	Typical (80-89 inches)
	Fa	Fair (70-79 inches)
	Po	Poor (<70 inches)
	NA	No Basement
BsmtCond	Evaluates the general condition of the basement	
	Ex	Excellent
	Gd	Good
	TA	Typical - slight dampness allowed
	Fa	Fair - dampness or some cracking or settling
	Po	Poor - Severe cracking, settling, or wetness
	NA	No Basement
BsmtExposure	Refers to walkout or garden level walls	
	Gd	Good Exposure
	Av	Average Exposure
	Av	(split levels or foyers typically score average or above)
	Mn	Mimimum Exposure
	No	No Exposure
	NA	No Basement
BsmtFinType1	Rating of basement finished area	
	GLQ	Good Living Quarters
	ALQ	Average Living Quarters
	BLQ	Below Average Living Quarters
	Rec	Average Rec Room
	LwQ	Low Quality
	Unf	Unfinshed
	NA	No Basement
BsmtFinSF1	Type 1 finished square feet	
BsmtFinType2	Rating of basement finished area (if multiple types)	
	GLQ	Good Living Quarters
	ALQ	Average Living Quarters
	BLQ	Below Average Living Quarters
	Rec	Average Rec Room
	LwQ	Low Quality
	Unf	Unfinshed
	NA	No Basement
BsmtFinSF2	Type 2 finished square feet	
BsmtUnfSF	Unfinished square feet of basement area	
TotalBsmtSF	Total square feet of basement area	

Heating	Type of heating	
	Floor	Floor Furnace
	GasA	Gas forced warm air furnace
	GasW	Gas hot water or steam heat
	Grav	Gravity furnace
	OthW	Hot water or steam heat other than gas
	Wall	Wall furnace
HeatingQC	Heating quality and condition	
	Ex	Excellent
	Gd	Good
	TA	Average/Typical
	Fa	Fair
	Po	Poor
CentralAir	Central air conditioning	
	N	No
	Y	Yes
Electrical	Electrical system	
	SBrkr	Standard Circuit Breakers & Romex
	FuseA	Fuse Box over 60 AMP and all Romex wiring (Average)
	FuseF	60 AMP Fuse Box and mostly Romex wiring (Fair)
	FuseP	AMP Fuse Box and mostly knob & tube wiring (poor)
	Mix	Mixed
1stFlrSF	First Floor square feet	
2ndFlrSF	Second Floor square feet	
LowQualFinSF	Low quality finished square feet (all floors)	
GrLivArea	Above grade (ground) living area square feet	
BsmtFullBath	Basement full bathrooms	
BsmtHalfBath	Basement half bathrooms	
FullBath	Full bathrooms above grade	
HalfBath	Half baths above grade	
Bedroom	Bedrooms above grade (does NOT include basement bedrooms)	
Kitchen	Kitchens above grade	
KitchenQual	Electrical system	
	Ex	Excellent
	Gd	Good
	TA	Average/Typical
	Fa	Fair
	Po	Poor
TotRmsAbvGrd	Total rooms above grade (does not include bathrooms)	
Functional	Home functionality (Assume typical unless deductions are warranted)	
	Typ	Typical Functionality
	Min1	Minor Deductions 1
	Min2	Minor Deductions 2
	Mod	Moderate Deductions
	Maj1	Major Deductions 1
	Maj2	Major Deductions 2
	Sev	Severely Damaged
	Sal	Salvage only
Fireplaces	Number of fireplaces	

Classification Variables		
FireplaceQu	Electrical system	
	Ex	Excellent - Exceptional Masonry Fireplace
	Gd	Good - Masonry Fireplace in main level
	TA	Average - Prefabricated Fireplace in main living area or Masonry Fireplace in basement
	Fa	Fair - Prefabricated Fireplace in basement
	Po	Poor - Ben Franklin Stove
	NA	No Fireplace
GarageType	Garage location	
	2Types	More than one type of garage
	Attchd	Attached to home
	Basment	Basement Garage
	BuiltIn	Built-In (Garage part of house - typically has room above garage)
	CarPort	Car Port
	Detchd	Detached from home
	NA	No Garage
GarageYrBlt	Year garage was built	
GarageFinish	Interior finish of the garage	
	Fin	Finished
	RFn	Rough Finished
	Unf	Unfinished
	NA	No Garage
GarageCars	Size of garage in car capacity	
GarageArea	Size of garage in square feet	
GarageQual	Garage quality	
	Ex	Excellent
	Gd	Good
	TA	Typical/Average
	Fa	Fair
	Po	Poor
	NA	No Garage
PavedDrive	Paved driveway	
	Y	Paved
	P	Partial Pavement
	N	Dirt/Gravel
WoodDeckSF	Wood deck area in square feet	
OpenPorchSF	Open porch area in square feet	
EnclosedPorch	Enclosed porch area in square feet	
3SsnPorch	Three season porch area in square feet	
ScreenPorch	Screen porch area in square feet	
PoolArea	Pool area in square feet	
PoolQC	Garage quality	
	Ex	Excellent
	Gd	Good
	TA	Typical/Average
	Fa	Fair
	NA	No Pool

Classification Variables		
Fence	Fence quality	
	GdPrv	Good Privacy
	MnPrv	Minimum Privacy
	GdWo	Good Wood
	MnWw	Minimum Wood/Wire
	NA	No Fence
MiscFeature	Miscellaneous feature not covered in other categories	
	Elev	Elevator
	Gar2	2nd Garage (if not described in garage section)
	Othr	Other
	Shed	Shed (over 100 SF)
	TenC	Tennis Court
	NA	None
MiscVal	$Value of miscellaneous feature	
MoSold	Month Sold (MM)	
YrSold	Year Sold (YYYY)	
SaleType	Type of sale	
	WD	Warranty Deed - Conventional
	CWD	Warranty Deed - Cash
	VWD	Warranty Deed - VA Loan
	New	Home just constructed and sold
	COD	Court Officer Deed/Estate
	Con	Contract 15% Down payment regular terms
	ConLw	Contract Low Down payment and low interest
	ConLI	Contract Low Interest
	ConLD	Contract Low Down
	Oth	Other
SaleCondition	Condition of sale	
	Normal	Normal Sale
	Abnorml	Abnormal Sale - trade, foreclosure, short sale
	AdjLand	Adjoining Land Purchase
	Alloca	Allocation - two linked properties with separate deeds, typically condo with a garage unit
	Family	Sale between family members
	Partial	Home was not completed when last assessed (associated with New Homes)