Kaggle Project

House Prices: Advanced Regression Techniques

Christian Nava, Jenna Ford, Dan Crouthamel, Sangrae Cho
August 14, 2019

I. Introduction

Data Description

The Ames Housing Dataset is a modern alternative to the popular Boston Housing Dataset. The Ames dataset contains 2,930 observations of individual residential property sales in Ames, Iowa from 2006 to 2010 and contains 80 quantitative and categorical variables. These 80 variables can be further categorized into 20 continuous variables related to area dimensions, 14 discrete variables related to the number and types of rooms in a property (e.g., kitchen, bathroom, bedroom, etc.), and 46 categorical variables (23 nominal and 23 ordinal) describing garages, materials, environmental conditions, and ratings of items within the property.[1] The dataset and its label descriptions can be found on the Kaggle competition data page.

II. Analysis Question 1

Problem Statement

Understanding the factors influencing a home's sale price is a critical business need for a real estate company. The client, Century21 Ames, who sells houses in the North Ames, Edwards, and Brookside neighborhoods wants an estimate that helps them determine how the sale price of a house is related to the square footage of its living area. Also, the client wants to know if the sale price depends on the house's neighborhood (i.e., North Ames, Edwards, or Brookside).

Solution Outline
As linear regression techniques have been successful in predicting housing prices, in this project they will be used to provide an estimate of the relationship between the sale price of a house and the square footage of its living area.

[1]De Cock, D. (2011). Ames, Iowa: Alternative to the Boston housing data as an end of semester regression project. *Journal of Statistics Education*, 19(3).

The metric used to evaluate submissions for this Kaggle competition is the root mean square error (RMSE) between the logarithm of the predicted value and the logarithm of the observed sale price.

Build and Fit the Model

Linear Regression

Linear regression is a method to determine whether one or more predictor variables explain the dependent variable. In this project, we have performed a log-log transformation, $log_e(x)$, of the data where both the response and explanatory (predictor) variables are logged and the regression model is given by the following:

$$\mu\{log(Y) \mid log(X),\ Neighborhood\} = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 log(X) + \beta_4 Cent_1 + \beta_5 Cent_2$$

Where Y is the sale price of the home, X is the square footage of the home's living area (in 100 sq ft increments), β_0 is the intercept from the linear regression equation, β_1, β_2, β_3, β_4 and β_5 are the regression coefficients, D_1 and D_2 are dummy variables for Edwards and Brookside neighborhoods, respectively, and $Cent_1$ and $Cent_2$ are variables representing the interaction between the log of the square footage of the living area and the Edwards and Brookside neighborhoods, respectively. A centering method was used for the $Cent$ interactions.

Check Assumptions

Linear Relationship

Per the residual plot (Figure 5, top left), the residuals appear randomly distributed around 0, which suggests a linear relationship and satisfies the assumption of linearity.

Normality

Per the histogram of the residuals and the QQ plot (Figure 5, middle left and bottom left), the log-transformed data appear normally distributed.

Constant Variance/Multicollinearity

There should be little to no multicollinearity among the predictors. Per the variance inflation factor (VIF), where a value of VIF>10 would indicate the presence of multicollinearity, the assumption is satisfied. VIF values for all variables are below 1.2 (Figure 6).

Independence

We will assume independence. The dataset contains the sales of all homes in Ames, Iowa from 2006 to 2010.

Outliers

The original plots do show some outliers where the square footage is about 4000. Figure 1 shows the scatterplot with all values and no transformations. The Cook's D plot in Figure 3 shows that at least one of these observations over 4000 sq ft has high leverage. All analyses are done removing observations where the house has more than 4000 sq ft.

Comparing Competing Models

Adjusted R^2

The adjusted r-squared statistic is a variation on the coefficient of determination, R^2, that has been adjusted for the number of explanatory variables in the model. With a multivariate regression model, as additional explanatory variables are included in the model, the R^2 value will continue to increase irrespective of the variable's significance. With the adjusted R^2, however, explanatory variables that do not improve the model are penalized, which helps determine if the addition of an explanatory variable improves the model fit. In our model, the R^2 value is 51.21% and the adjusted R^2 value is 50.56%.

Parameters

Estimates

Table 1 shows the parameter estimates for the model.

$$\mu\{log(Y) \mid log(X),\ Neighborhood\} = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 log(X) + \beta_4 Cent_1 + \beta_5 Cent_2$$

Table 1: Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation	95% Confidence Limits	
Intercept	1	10.40966	0.08593	121.13	<.0001	0	10.24068	10.57863
N1	1	-0.14503	0.02293	-6.33	<.0001	1.06949	-0.19012	-0.09995
N2	1	-0.11467	0.02839	-4.04	<.0001	1.10740	-0.17049	-0.05885
GRLIVAREA_LOG	1	0.57731	0.03358	17.19	<.0001	1.04260	0.51129	0.64334
CENT1_LOG	1	0.20031	0.08228	2.43	0.0154	1.10338	0.03854	0.36209
CENT2_LOG	1	0.34662	0.08345	4.15	<.0001	1.16008	0.18254	0.51071

Interpretation of Parameters

β_0 : **Intercept** - The intercept in the model provides an estimate ($e^{10.40966}$ = $33,179) of the sale price of a house in the N. Ames neighborhood (reference neighborhood) with a living area square footage of zero.

β_1 : **N1 -** This is the adjustment of the intercept for a house in the Edwards neighborhood with respect to a house N. Ames neighborhood. For a living area square footage of zero, the house in the Edwards neighborhood has an estimated median sale price of $e^{-0.14503}$ = .86 or 14% less than the house in the N. Ames neighborhood.

β_2 : **N2 -** This is the adjustment of the intercept for a house in the Brookside neighborhood with respect to a house N. Ames neighborhood. For a living area square footage of zero, the house in the Brookside neighborhood has an estimated median sale price of $e^{-0.11467}$ = .89 or 11% less than the house in the N. Ames neighborhood.

β_3 : **GRLIVAREA_LOG -** Each doubling of the living area square footage, with the neighborhood held constant, results in a $2^{0.57731}$ = 1.4921 multiplicative change in the median sale price. This translates to a 49.21% increase in the median sale price.

β_4 : **CENT1_LOG -** Each doubling of the living area square footage for the Edwards neighborhood, with respect to the N. Ames neighborhood, results in a $2^{0.20031}$ = 1.1489 multiplicative change in the median sale price. This translates to a 14.89% increase in the median sale price for the Edwards neighborhood over the N. Ames neighborhood.

β_5 : **CENT2_LOG -** Each doubling of the living area square footage for the Brookside neighborhood, with respect to the N. Ames neighborhood, results in a $2^{0.34662}$ = 1.2716 multiplicative change in the median sale price. This translates to a 27.16% increase in the median sale price for the Brookside neighborhood over the N. Ames neighborhood.

Predictions

Using the model discussed above, we have provided predictions for the 3 neighborhoods at each of their mean living area square footage. These predictions are listed in Table 2.

Table 2: Sales Price Predictions

Neighborhood	Avg. Living Area	Predicted Value	95% Confidence Limits - Mean		95% Confidence Limits - Prediction	
Brookside	1203	$123,797	$117,818	$130,066	$85,0598	$180,160
Edwards	1203	$120,427	$115,983	$125,041	$82,868	$175,027
N. Ames	1203	$139,763	$136,298	$143,315	$96,269	$202,906
Brookside	1310	$132,774	$125,958	$139,944	$91,190	$193,300
Edwards	1310	$127,555	$122,688	$132,628	$87,755	$185,4059
N. Ames	1310	$145,525	$141,917	$149,223	$100,238	$211,272
Brookside	1340	$135,239	$135,239	$142,729	$92,865	$196,929
Edwards	1340	$129,495	$129,495	$134,767	$89,081	$188,245
N. Ames	1340	$147,075	$147,075	$150,874	$101,3064	$213,523

Conclusion

The relationship between the living area and the median sale price of a house with respect to the three neighborhoods of interest can be quantified by the following equations:

$$N.\,Ames:\ \mu\{logSalePrice \mid logLiveArea\} = 10.6711 + 0.47302 * logLiveArea$$

$$Edwards:\ \mu\{logSalePrice \mid logLiveArea\} = 10.0239 + 0.67333 * logLiveArea$$

$$Brookside:\ \mu\{logSalePrice \mid logLiveArea\} = 9.68755 + 0.81964 * logLiveArea$$

Where a doubling in the living area leads to a $2^{0.47302}$ = 1.3888 (39% increase) multiplicative change in the median sale price for the N. Ames neighborhood, a $2^{0.67333}$ = 1.5947 (59% increase) multiplicative change in the median sale price for the Edwards neighborhood, and a $2^{0.81964}$ = 1.7650 (77% increase) multiplicative change in the median sale price for the Brookside neighborhood.

III. Analysis Question 2

Problem Statement

We are seeking to build the best predictive model for sales price from a selection of the following linear regression models: forward selection, backward elimination, stepwise selection, and a custom model of our design. Comparative metrics to determine the best model will include an adjusted R^2 value, CV Press, and a Kaggle Score for each of the four models. Variables with the highest correlation are included in the selection models.

Model Selection

Forward Selection

In a forward selection model we start with just the intercept and no predictors. We subsequently add predictors one at a time using the predictor that improves the fit of the model the most until no new predictor adds a significant improvement to the model.[2]

Our forward selection model achieved an adjusted R^2 of 0.8927, CV press value of 22.0603, and Kaggle score of 0.16250 (Table 2).

[2] SAS/STAT(R) 12.3 User's Guide: High-Performance Procedures, [Forward Selection]

Backward Elimination

The backward elimination model begins with a full model that uses all possible predictors and then successively eliminates the least significant predictors one by one until a stopping condition is met.[3]

For our model, the stopping condition was the predicted residual sum of square with *k*-fold cross validation (CV) for 5 folds. Our backward elimination model achieved an adjusted R^2 of 0.8959, CV press value of 22.6955, and Kaggle score of 0.16193 (Table 3).

Stepwise Selection

The stepwise selection model is a variation on the forward selection model where we start with only the intercept and sequentially add the most significant predictors, however, at each step of adding a predictor, variables that do not significantly improve the fit of the model are eliminated as they would be in a backward elimination model.[4]

Our stepwise selection model achieved an adjusted R^2 of 0.8927, CV press value of 22.0022, and Kaggle score of 0.16193 (Table 3).

Custom Model

For our custom model, we looked at variables that had high correlation and considered variables that we found important when looking for houses such as building type (i.e. single family home), whether the house has central air conditioning/heating and the year the house was built.

Our custom model achieved an adjusted R^2 of 0.9028, CV press value of 23.3070, and Kaggle score of 0.13198 (Table 3). This model outperformed the forward, backward and stepwise selection models.

Check Assumptions for the Custom Model

Linear Relationship

Per the residual plot (Figure 24, upper left graph), the residuals appear randomly distributed around 0, which suggests a linear relationship and satisfies the assumption of linearity.

Normality

Per the histogram of the residuals and the QQ plot (Figure 24, middle left graph), the log-transformed data appear normally distributed. The distribution does appear to have a slight left-skew.

---

[3] Ibid, [Backward Elimination]

[4] Ibid, [Stepwise Selection]

Constant Variance/Multicollinearity

Per the residual plot (Figure 24, upper left graph), the residuals appear randomly distributed around 0 with no evidence of unequal variance. Additionally, an unequal variance is related to a skewed distribution. Per the histogram of the residuals, with a superimposed normal distribution (Figure 24, bottom left), there is no evidence of unequal variance.

Independence

We will assume independence. The dataset contains the sales of all homes in Ames, Iowa from 2006 to 2010.

Outliers

The original plots do show some outliers where the square footage is about 4000. Figure 1 shows the scatterplot with all values and no transformations. The Cook's D plot in Figure 3 shows that at least one of these observations over 4000 sq ft has high leverage. All analyses are done removing observations where the house has more than 4000 sq ft.

Comparing Competing Models

As shown in Table 3, we found that the Custom Model performed the best in terms of Adjusted R^2, CV Press, and Kaggle Score metrics as compared to the Forward Selection, Backward Elimination, and Stepwise Selection models.

Table 3: Metric comparison for all models

Predictive Model	Adjusted R^2	CV Press	Kaggle Score
Forward Selection	0.8927	22.0603	0.16250
Backward Elimination	0.8959	21.6955	0.16193
Stepwise Selection	0.8927	22.0022	0.15571
Custom Model	0.9028	23.3070	0.13198

Conclusion

Our custom model produced the best performance from the models in Table 3. Given the scope of the project and the limitations of the models, there is additional room for improvement that other models can address. Further explorations of models to predict house sale prices include regularized linear models, boosting models (XGBoost, LightGBM), and stacked regression models.

Appendix

Reading in the Data

```
 %MACRO READ_DATA(PATH,NAME,NUM_OBS);
* REPLACE NA WITH MISSING;
DATA _NULL_;
      INFILE "&PATH.&NAME..csv" DSD TRUNCOVER;
      FILE "&PATH.&NAME._missing.csv" DSD;
      LENGTH WORD $200;

      * LOOP THROUGH EACH OF THE 81 COLUMNS AND REPLACE 'NA' WITH .;
      DO I=1 TO &NUM_OBS;
            INPUT WORD @;
            IF I IN (7,31,32,33,34,36,58,59,61,64,65,73,74,75) THEN DO;
                  PUT WORD@;
            END;
            ELSE DO;
                  IF WORD='NA' THEN WORD = .;
                  PUT WORD@;
            END;
      END;

      * OUTPUT THE RECORD TO THE FILE;
      PUT;
RUN;

* IMPORT THE FILE WHERE NAs ARE REPLACED WITH .;
PROC IMPORT DATAFILE="&PATH.&NAME._missing.csv" OUT=&NAME REPLACE;
      GUESSINGROWS=MAX;
RUN;
%MEND;

* CREATE A DATASET WITH THE LIST OF FILES TO READ IN;
DATA LIST;
      * THIS IS WHERE YOU UPDATE WITH THE LOCATION OF YOUR FILES;
      PATH='/data/bnsf/ib/hubops/jford/data_science/kaggle/';
      NAME='train'; NUM_OBS = 81; OUTPUT;
      NAME='test'; NUM_OBS = 80; OUTPUT;
RUN;
```

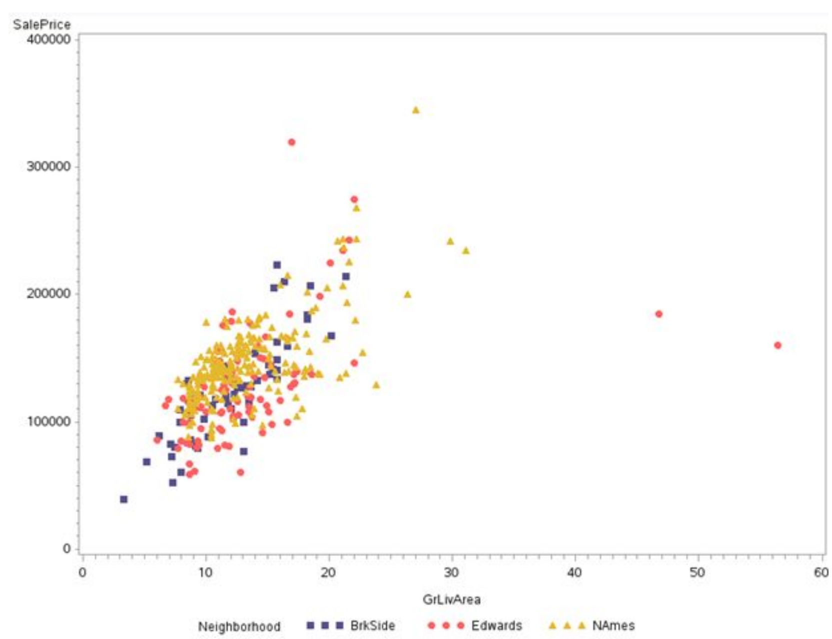
```
* EXECUTE THE MACRO TO LOOP THROUGH THE FILES BEING READ IN;
DATA _NULL_;
      SET LIST;
    CALL EXECUTE('%READ_DATA('||PATH||','||NAME||','||NUM_OBS||')');
RUN;
```

Analysis Question 1

```
* CREATE DUMMY VARIABLES FOR NEIGHBORHOOD AND ONLY KEEP 3
NEIGHBORHOODS;
DATA TRAIN1;
      SET TRAIN;
      IF NEIGHBORHOOD IN ('Edwards','BrkSide','NAmes');
      IF NEIGHBORHOOD = 'Edwards' THEN DO;
            N1=1;
            N2=0;
      END;

      ELSE IF NEIGHBORHOOD = 'BrkSide' THEN DO;
            N1=0;
            N2=1;
      END;

      ELSE DO;
            N1=0;
            N2=0;
      END;

      GRLIVAREA = GRLIVAREA/100;

      SALEPRICE_LOG = LOG(SALEPRICE);
      GRLIVAREA_LOG = LOG(GRLIVAREA);
RUN;

****************************;
* WITH OUTLIERS            ;
****************************;
PROC MEANS DATA=TRAIN1;
RUN;

* CREATE INTERACTION VARIABLES;
DATA TRAIN2;
      SET TRAIN1;
```

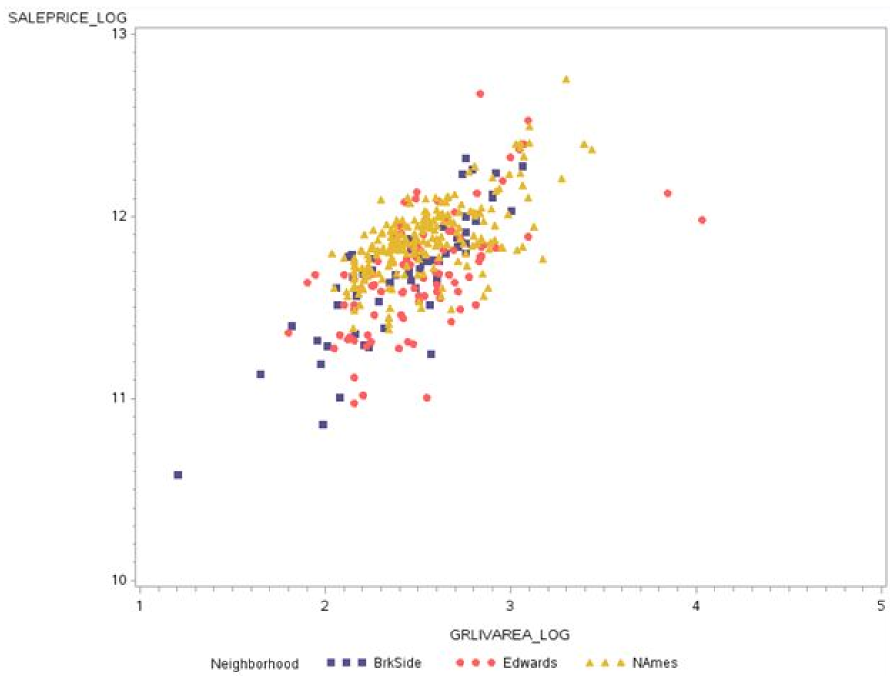
```
      CENT1 = (GRLIVAREA_LOG - 13.0183)*(N1 - 0.2610966);
      CENT2 = (GRLIVAREA_LOG - 13.0183)*(N2 - 0.1514360);

      CENT1_LOG = (GRLIVAREA_LOG - 2.5141431)*(N1 - 0.2610966);
      CENT2_LOG = (GRLIVAREA_LOG - 2.5141431)*(N2 - 0.1514360);

      GRCENT=(GRLIVAREA_LOG - 2.5141431);

RUN;
ODS GRAPHICS ON;

SYMBOL1 V='SQUAREFILLED' C="#58508D" I=NONE;
SYMBOL2 V='DOT' C=ROSE I=NONE;
SYMBOL3 V='TRIANGLEFILLED' C=BIOY I=NONE;

*EDA AND REGRESSIONS WITHOUT TRANSFORMATIONS;
PROC GPLOT DATA=TRAIN2;
PLOT SALEPRICE*GRLIVAREA=NEIGHBORHOOD;
TITLE1 'Figure 1: House Sales Price and Square Footage by
Neighborhood';
TITLE2 'Without Transformations';
RUN;
```

Figure 1: House Sale Price and Square Footage by Neighborhood
'Without Transformations



```
*EDA AND REGRESSIONS WITH TRANSFORMATIONS;
PROC GPLOT DATA=TRAIN2;
PLOT SALEPRICE_LOG*GRLIVAREA_LOG=NEIGHBORHOOD;
TITLE1 'Figure 2: House Sales Price and Square Footage by
Neighborhood';
TITLE2 'With Log Transformations';
RUN;
```

Figure 2: House Sale Price and Square Footage by Neighborhood
With Log Transformations




```
PROC REG DATA=TRAIN2;
	MODEL SALEPRICE = N1 N2 GRLIVAREA /VIF;
	TITLE1 'Figure 3: Simple Linear Regression';
	TITLE2 'No Transformations';
RUN;
```

Figure 3: Simple Linear Regression
No Transformations

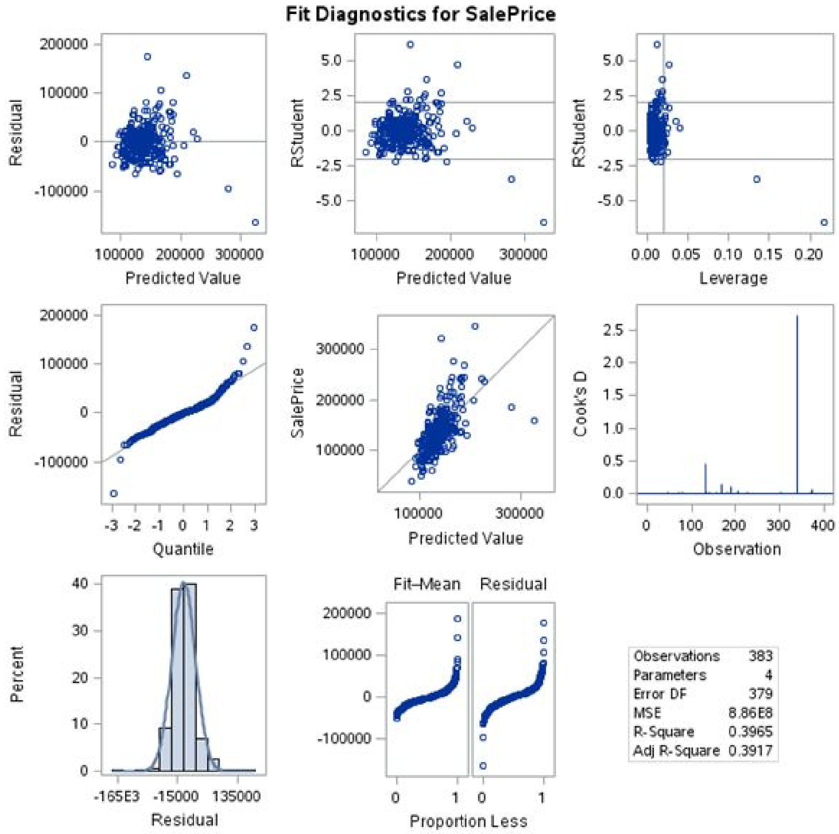


Figure 4: Summary
Simple Linear Regression, No Transformation

Number of Observations Read	383
Number of Observations Used	383

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2.205076E11	73502535486	83.00	<.0001
Error	379	3.356445E11	885605599		
Corrected Total	382	5.561521E11			

Root MSE	29759	R-Square	0.3965
Dependent Mean	138063	Adj R-Sq	0.3917
Coeff Var	21.55482		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	85887	4578.11684	18.76	<.0001	0
N1	1	-18988	3577.82668	-5.31	<.0001	1.06804
N2	1	-16106	4395.35182	-3.66	0.0003	1.07364
GrLivArea	1	4576.00645	314.88014	14.53	<.0001	1.00815

```
PROC REG DATA=TRAIN2 plots=all;
        MODEL SALEPRICE_LOG = N1 N2 GRLIVAREA_LOG CENT1_LOG
    CENT2_LOG/VIF;
    TITLE1 'Figure 5: Simple Linear Regression';
    TITLE2 'Log-Log Transformation';
RUN;
```

Figure 5: Simple Linear Regression
Log-Log Transformation

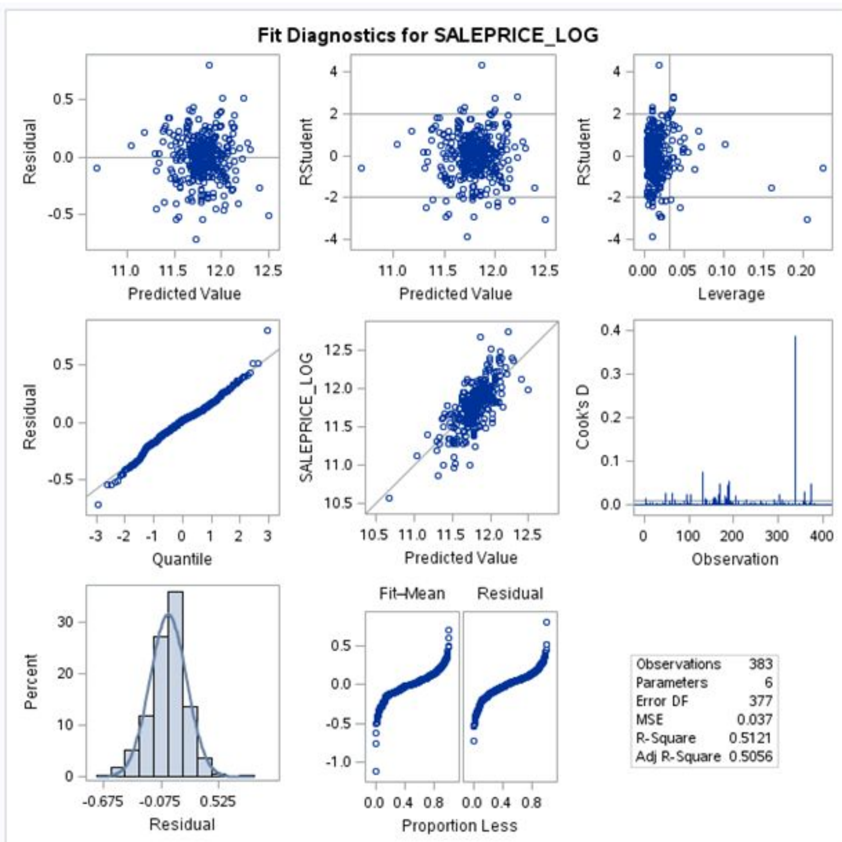


Figure 6: Summary
Simple Linear Regression, Log-Log Transformation

Number of Observations Read	383
Number of Observations Used	383

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	14.62858	2.92572	79.14	<.0001
Error	377	13.93775	0.03697		
Corrected Total	382	28.56633			

Root MSE	0.19228	R-Square	0.5121
Dependent Mean	11.79887	Adj R-Sq	0.5056
Coeff Var	1.62962		

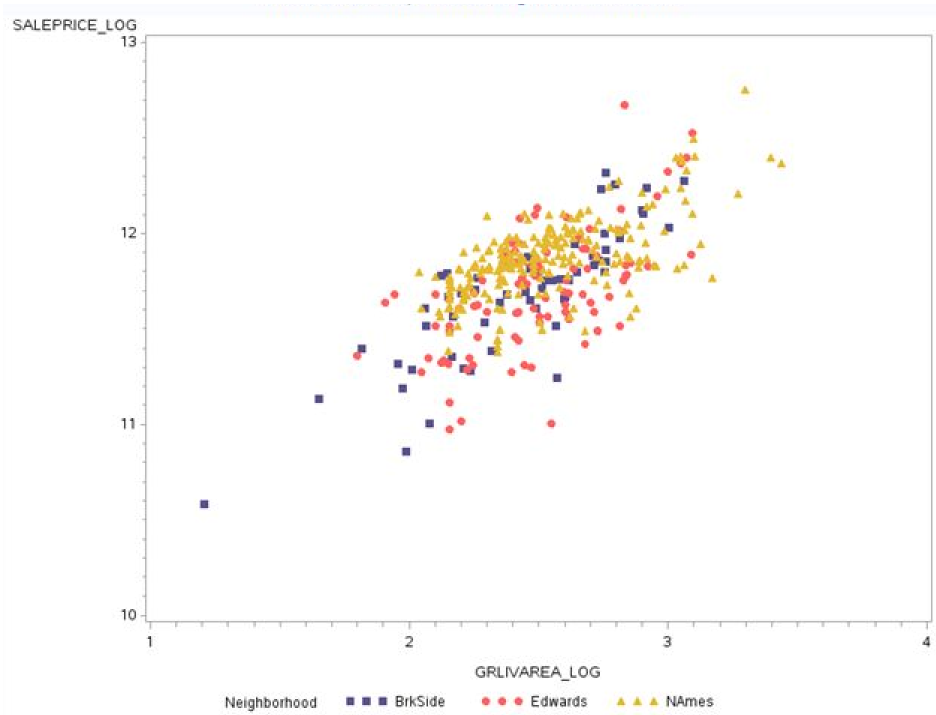
Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	10.50849	0.08297	126.65	<.0001	0
N1	1	-0.15415	0.02313	-6.66	<.0001	1.06929
N2	1	-0.11208	0.02895	-3.87	0.0001	1.11578
GRLIVAREA_LOG	1	0.53769	0.03236	16.61	<.0001	1.05356
CENT1_LOG	1	0.04664	0.07248	0.64	0.5203	1.14765
CENT2_LOG	1	0.34662	0.08482	4.09	<.0001	1.18656

```
****************************;
* WITHOUT OUTLIERS         ;
****************************;
DATA TRAIN_NO_OUTLIERS;
      SET TRAIN1;
      IF GRLIVAREA < 4000/100;
RUN;

PROC MEANS DATA=TRAIN_NO_OUTLIERS;
RUN;

DATA TRAIN_NO_OUTLIERS1;
      SET TRAIN_NO_OUTLIERS;
      CENT1 = (GRLIVAREA_LOG - 12.8158530)*(N1 - 0.2572178);
      CENT2 = (GRLIVAREA_LOG - 12.8158530)*(N2 - 0.1522310);

      CENT1_LOG = (GRLIVAREA_LOG - 2.5066639)*(N1 - 0.2572178);
      CENT2_LOG = (GRLIVAREA_LOG - 2.5066639)*(N2 - 0.1522310);
RUN;
PROC GPLOT DATA=TRAIN_NO_OUTLIERS1;
```

```
      PLOT SALEPRICE * GRLIVAREA=NEIGHBORHOOD;
      TITLE1 'Figure 7: House Sales Price and Square Footage by
      Neighborhood';
      TITLE2 'Without Outliers, Without Transformations';
RUN;
```


```
PROC GPLOT DATA=TRAIN_NO_OUTLIERS1;
      PLOT SALEPRICE_LOG * GRLIVAREA_LOG=NEIGHBORHOOD;
      TITLE1 'Figure 8: House Sales Price and Square Footage by
      Neighborhood';
      TITLE2 'Without Outliers, Without Log Transformations';
RUN;
```

Figure 8: House Sale Price and Square Footage by Neighborhood
Without Outliers, Without Log Transformations




```
PROC REG DATA=TRAIN_NO_OUTLIERS1 PLOTS=ALL;
     MODEL SALEPRICE = N1 N2 GRLIVAREA /VIF;
     TITLE1 'Figure 9: Simple Linear Regression';
     TITLE2 'Without Outliers, Without Transformations';
RUN;
```

Figure 9: Simple Linear Regression
Without Outliers, Without Transformations

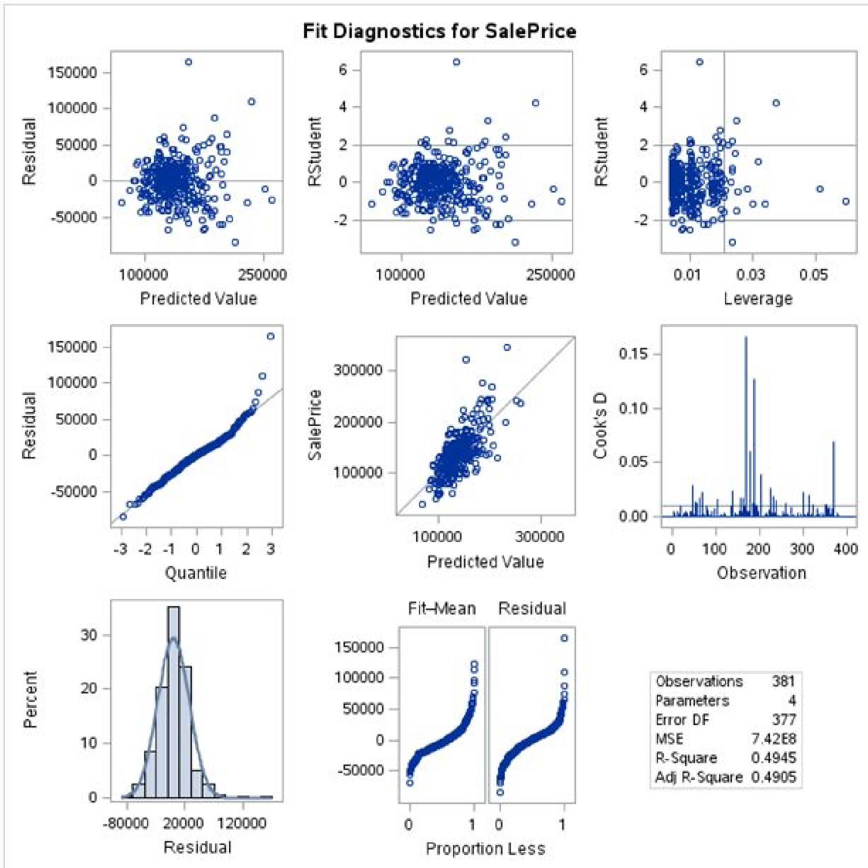


Figure 10: Summary
Simple Linear Regression, Untransformed Data Without Outliers

Number of Observations Read	381
Number of Observations Used	381

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2.737209E11	91240293819	122.95	<.0001
Error	377	2.797579E11	742063422		
Corrected Total	380	5.534788E11			

Root MSE	27241	R-Square	0.4945
Dependent Mean	137882	Adj R-Sq	0.4905
Coeff Var	19.75658		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	62577	4985.94083	12.55	<.0001	0
N1	1	-15465	3301.41217	-4.68	<.0001	1.06917
N2	1	-14198	4029.47740	-3.52	0.0005	1.07588
GrLivArea	1	6354.96854	354.37700	17.93	<.0001	1.00982

```
PROC REG DATA=TRAIN_NO_OUTLIERS1 PLOTS=ALL;
          MODEL SALEPRICE_LOG = N1 N2 GRLIVAREA_LOG CENT1_LOG
     CENT2_LOG/VIF CLB CLM CLI; */
     TITLE1 'Figure 11: Linear Regression';
     TITLE2 'Without Outliers, Log-Log Transformations';
RUN;
```

Figure 11: Linear Regression
Without Outliers, Log-Log Transformations

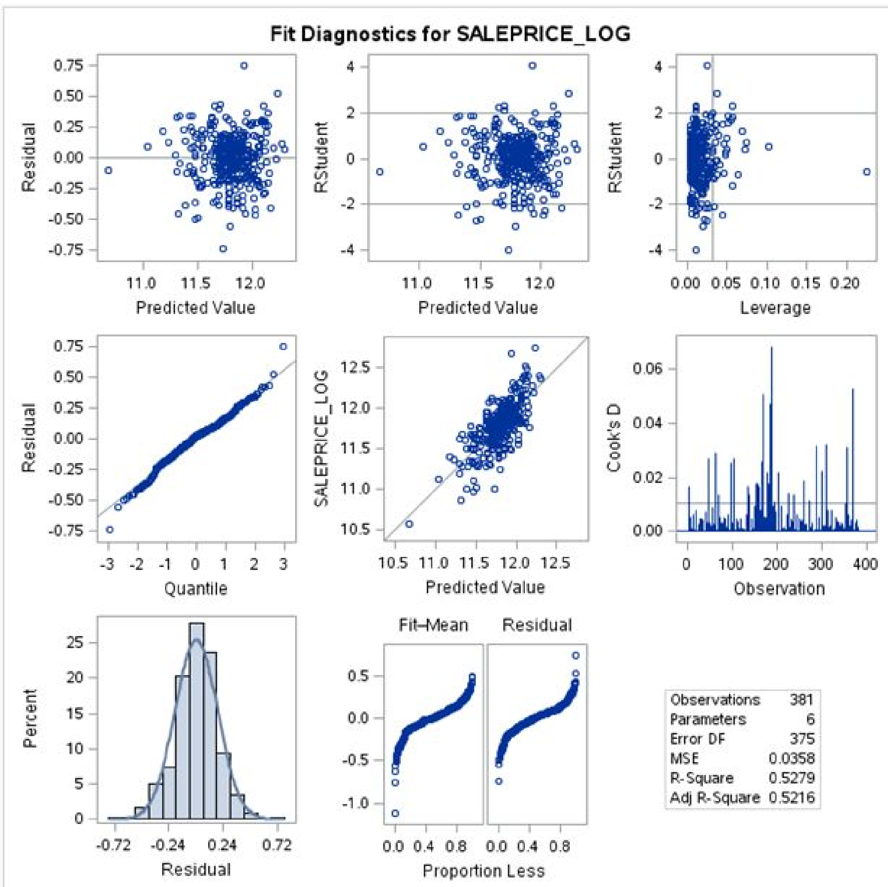


Figure 12: Summary
Simple Linear Regression, Log-transformed Data Without Outliers

Number of Observations Read	381
Number of Observations Used	381

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	15.00592	3.00118	83.87	<.0001
Error	375	13.41833	0.03578		
Corrected Total	380	28.42424			

Root MSE	0.18916	R-Square	0.5279
Dependent Mean	11.79752	Adj R-Sq	0.5216
Coeff Var	1.60340		

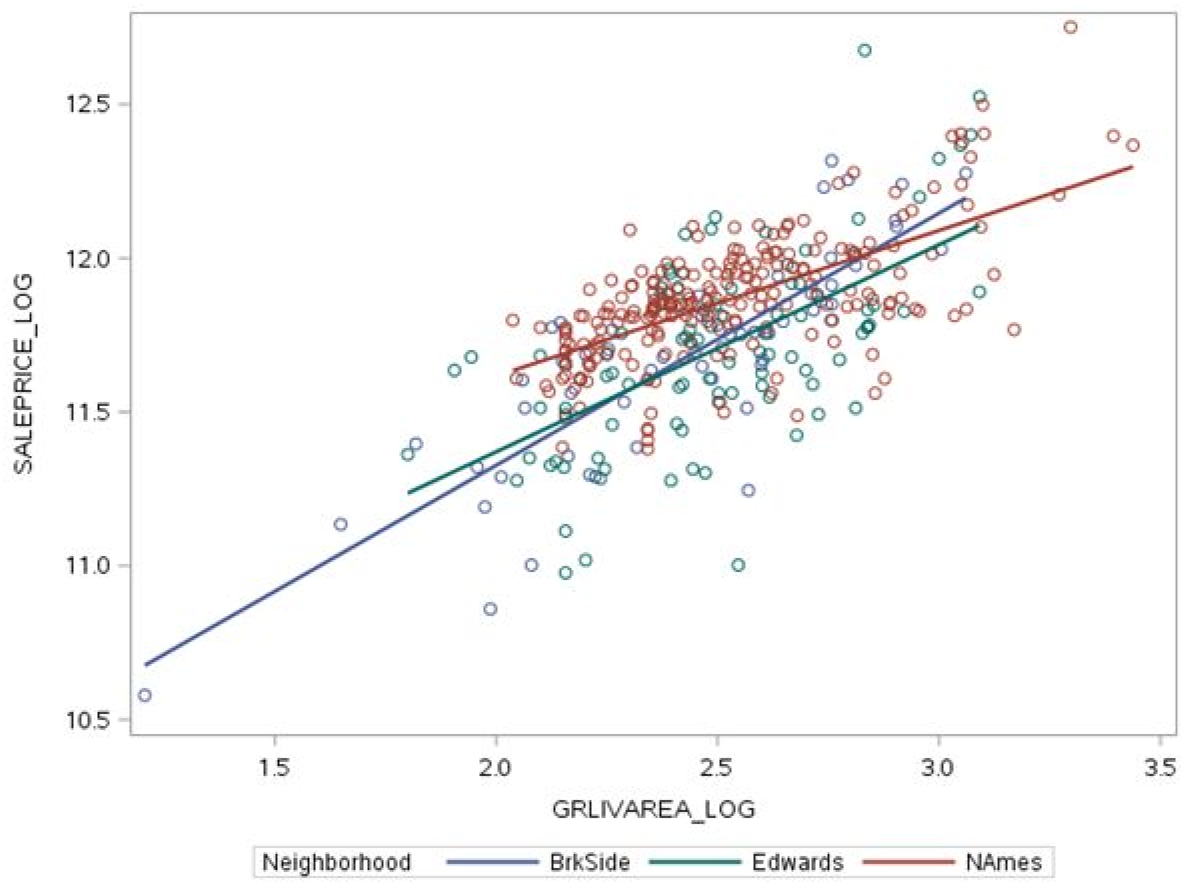
Parameter Estimates								
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation	95% Confidence Limits	
Intercept	1	10.40966	0.08593	121.13	<.0001	0	10.24068	10.57863
N1	1	-0.14503	0.02293	-6.33	<.0001	1.06949	-0.19012	-0.09995
N2	1	-0.11467	0.02839	-4.04	<.0001	1.10740	-0.17049	-0.05885
GRLIVAREA_LOG	1	0.57731	0.03358	17.19	<.0001	1.04260	0.51129	0.64334
CENT1_LOG	1	0.20031	0.08228	2.43	0.0154	1.10338	0.03854	0.36209
CENT2_LOG	1	0.34662	0.08345	4.15	<.0001	1.16008	0.18254	0.51071

Figure 13: Output Statistics
Simple Linear Regression, Log-transformed Data Without Outliers

Output Statistics								
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	11.7	11.6356	0.0251	11.5862	11.6851	11.2604	12.0109	0.0428
2	12.0	11.8669	0.0126	11.8421	11.8917	11.4942	12.2397	0.0971
3	11.8	11.4455	0.0320	11.3826	11.5084	11.0683	11.8227	0.3451
4	11.9	11.7621	0.0161	11.7305	11.7938	11.3889	12.1354	0.1496
5	11.8	11.8983	0.0129	11.8729	11.9238	11.5255	12.2712	-0.0561
6	11.8	11.7104	0.0195	11.6720	11.7488	11.3365	12.0843	0.1011
7	12.2	11.9826	0.0166	11.9499	12.0153	11.6092	12.3560	0.2603
8	11.1	11.0389	0.0604	10.9200	11.1577	10.6484	11.4293	0.0957
9	12.0	12.0113	0.0185	11.9749	12.0476	11.6375	12.3850	0.005468
10	11.9	11.8833	0.0127	11.8583	11.9082	11.5105	12.2561	0.0549
11	11.6	11.7865	0.0148	11.7574	11.8155	11.4134	12.1596	-0.1874
12	11.3	11.6696	0.0195	11.6314	11.7079	11.2957	12.0435	-0.3551
13	12.0	11.8930	0.0128	11.8678	11.9182	11.5202	12.2658	0.0899
14	11.9	11.8264	0.0132	11.8004	11.8523	11.4535	12.1992	0.0301
15	11.6	11.7077	0.0250	11.6587	11.7568	11.3326	12.0829	-0.0594
16	11.8	11.9057	0.0131	11.8799	11.9315	11.5329	12.2785	-0.1304
17	12.1	11.9278	0.0138	11.9006	11.9550	11.5548	12.3007	0.1757
18	12.1	12.1347	0.0282	12.0793	12.1901	11.7587	12.5108	-0.0340
19	12.4	12.1381	0.0285	12.0822	12.1941	11.7620	12.5143	0.2668
20	11.9	11.7993	0.0142	11.7714	11.8272	11.4263	12.1723	0.0845

```
PROC SGPLOT DATA=TRAIN_NO_OUTLIERS1 ;
     TITLE "Sales Price by Living Area and Neighborhood";
     REG Y=SALEPRICE_LOG X=GRLIVAREA_LOG / GROUP=NEIGHBORHOOD ;
     TITLE1 'Figure 14: Linear Regression';
     TITLE2 'Without Outliers, Log-Log Transformations';
RUN;
```

Figure 14: Linear Regression
Without Outliers, Log-Log Transformations

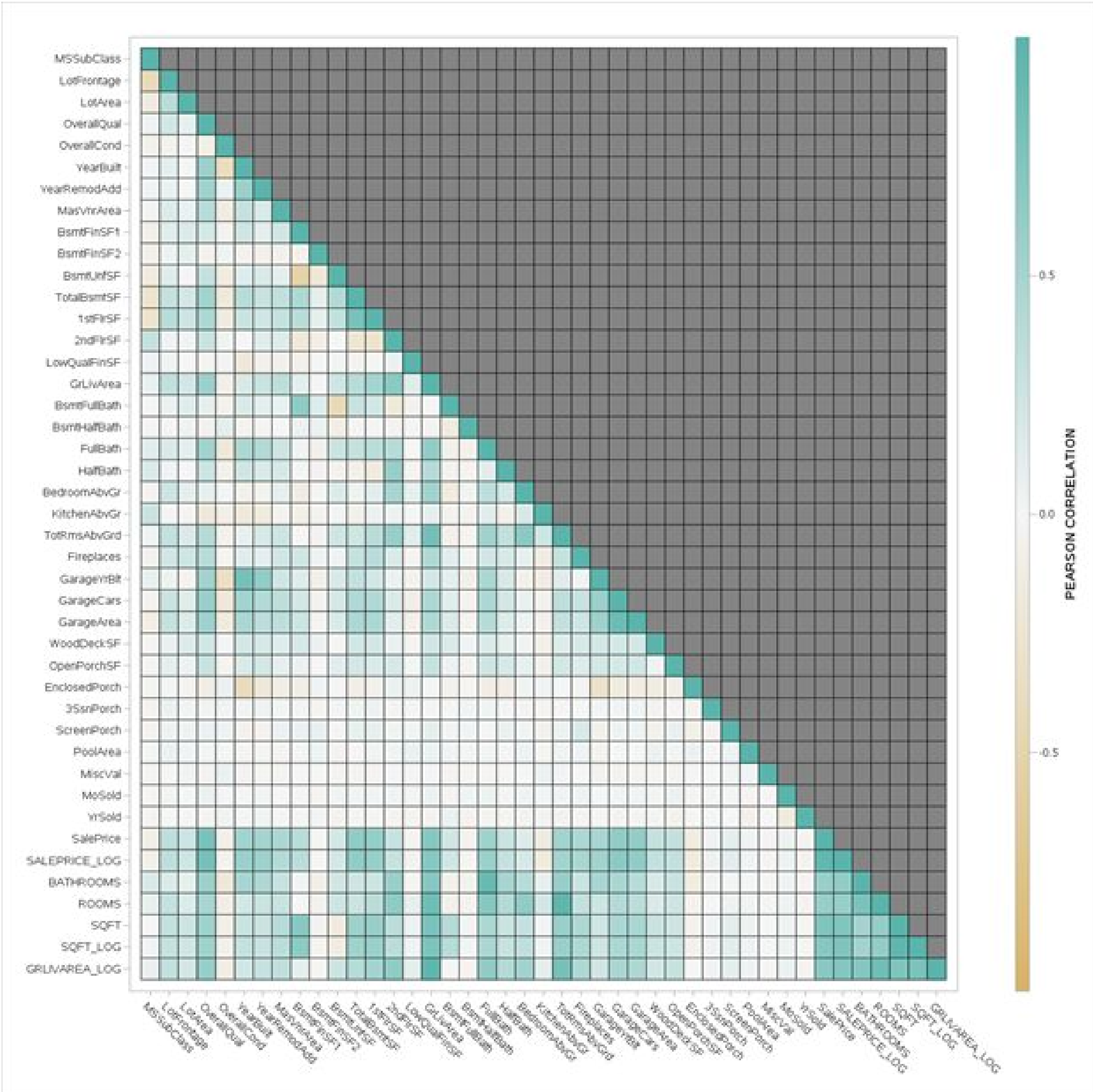


Analysis Question 2

```
* DATA PREP;
DATA TRAIN2;
     SET TRAIN;

     SALEPRICE_LOG = LOG(SALEPRICE);

     IF GRLIVAREA < 4000;

     BATHROOMS = .5*HALFBATH + FULLBATH;
     ROOMS = BATHROOMS + TOTRMSABVGRD;

     SQFT = (BSMTFINSF1 + GRLIVAREA)/100;
     SQFT_LOG = LOG(SQFT);

     GRLIVAREA = GRLIVAREA/100;
     GRLIVAREA_LOG = LOG(GRLIVAREA);
RUN;

* CORRELATION MATRIX;
* CODE FROM
HTTPS://BLOGS.SAS.COM/CONTENT/SASDUMMY/2013/06/12/CORRELATIONS-MATRIX
-HEATMAP-WITH-SAS/;

%MACRO PREPCORRDATA(IN=,OUT=);
  /* RUN CORR MATRIX FOR INPUT DATA, ALL NUMERIC VARS */
  PROC CORR DATA=&IN. NOPRINT
    PEARSON
    OUTP=WORK._TMPCORR
    VARDEF=DF
  ;
  RUN;

  /* PREP DATA FOR HEAT MAP */
DATA &OUT.;
  KEEP X Y R;
  SET WORK._TMPCORR(WHERE=(_TYPE_="CORR"));
  ARRAY V{*} _NUMERIC_;
  X = _NAME_;
  DO I = DIM(V) TO 1 BY -1;
    Y = VNAME(V(I));
    R = V(I);
    /* CREATES A LOWER TRIANGULAR MATRIX */
```

```
    IF (I<_N_) THEN
      R=.;
    OUTPUT;
  END;
RUN;

PROC DATASETS LIB=WORK NOLIST NOWARN;
  DELETE _TMPCORR;
QUIT;
%MEND;

ODS PATH WORK.MYSTORE(UPDATE) SASHELP.TMPLMST(READ);

PROC TEMPLATE;
  DEFINE STATGRAPH CORRHEATMAP;
   DYNAMIC _TITLE;
    BEGINGRAPH;
      ENTRYTITLE _TITLE;
      RANGEATTRMAP NAME='MAP';
      RANGE -1 - 1 / RANGECOLORMODEL=(CXD8B365 CXF5F5F5 CX5AB4AC);
      ENDRANGEATTRMAP;
      RANGEATTRVAR VAR=R ATTRVAR=R ATTRMAP='MAP';
      LAYOUT OVERLAY /
     XAXISOPTS=(DISPLAY=(LINE TICKS TICKVALUES))
     YAXISOPTS=(DISPLAY=(LINE TICKS TICKVALUES));
     HEATMAPPARM X = X Y = Y COLORRESPONSE = R /
          XBINAXIS=FALSE YBINAXIS=FALSE
          NAME = "HEATMAP" DISPLAY=ALL;
     CONTINUOUSLEGEND "HEATMAP" /
                    ORIENT = VERTICAL LOCATION = OUTSIDE
                    TITLE="PEARSON CORRELATION";
      ENDLAYOUT;
    ENDGRAPH;
  END;
RUN;

ODS GRAPHICS /HEIGHT=2400 WIDTH=2400 IMAGEMAP;

%PREPCORRDATA(IN=TRAIN2(DROP=ID) ,OUT=CORR_MATRIX);
PROC SGRENDER DATA=CORR_MATRIX TEMPLATE=CORRHEATMAP;
   DYNAMIC _TITLE_="CORR MATRIX";
RUN;
```

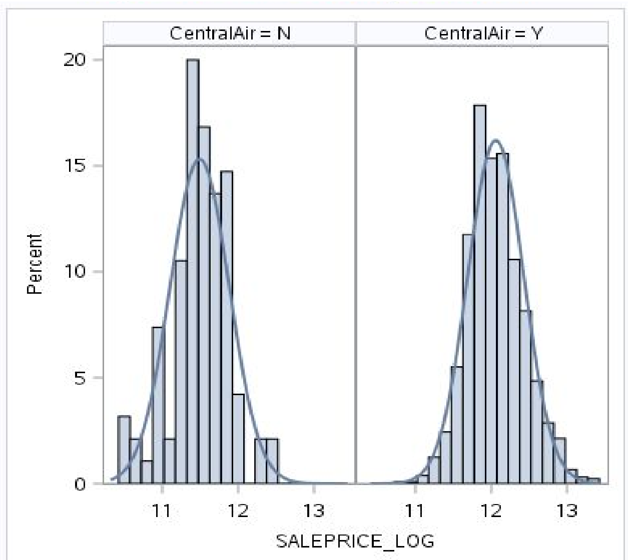
Figure 15: Correlation Matrix




```
 * EDA FOR CATEGORICAL VARIABLES;
ODS GRAPHICS /HEIGHT=600 WIDTH=600 IMAGEMAP;

PROC SGPANEL DATA=TRAIN2 NOAUTOLEGEND;
     TITLE "Figure 16: Sales Price by Central Air";
```

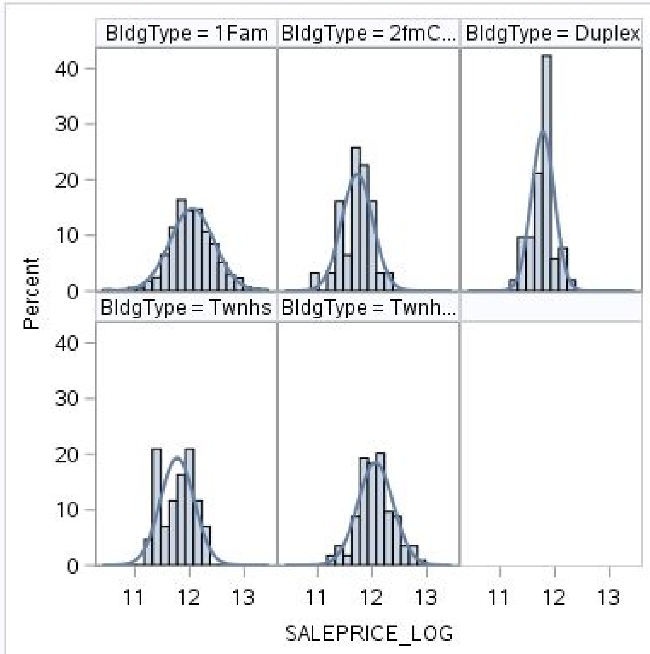
```
      PANELBY CENTRALAIR;
      HISTOGRAM SALEPRICE_LOG;
      DENSITY SALEPRICE_LOG;
RUN;
```

Figure 16: Sale Price by Central Air



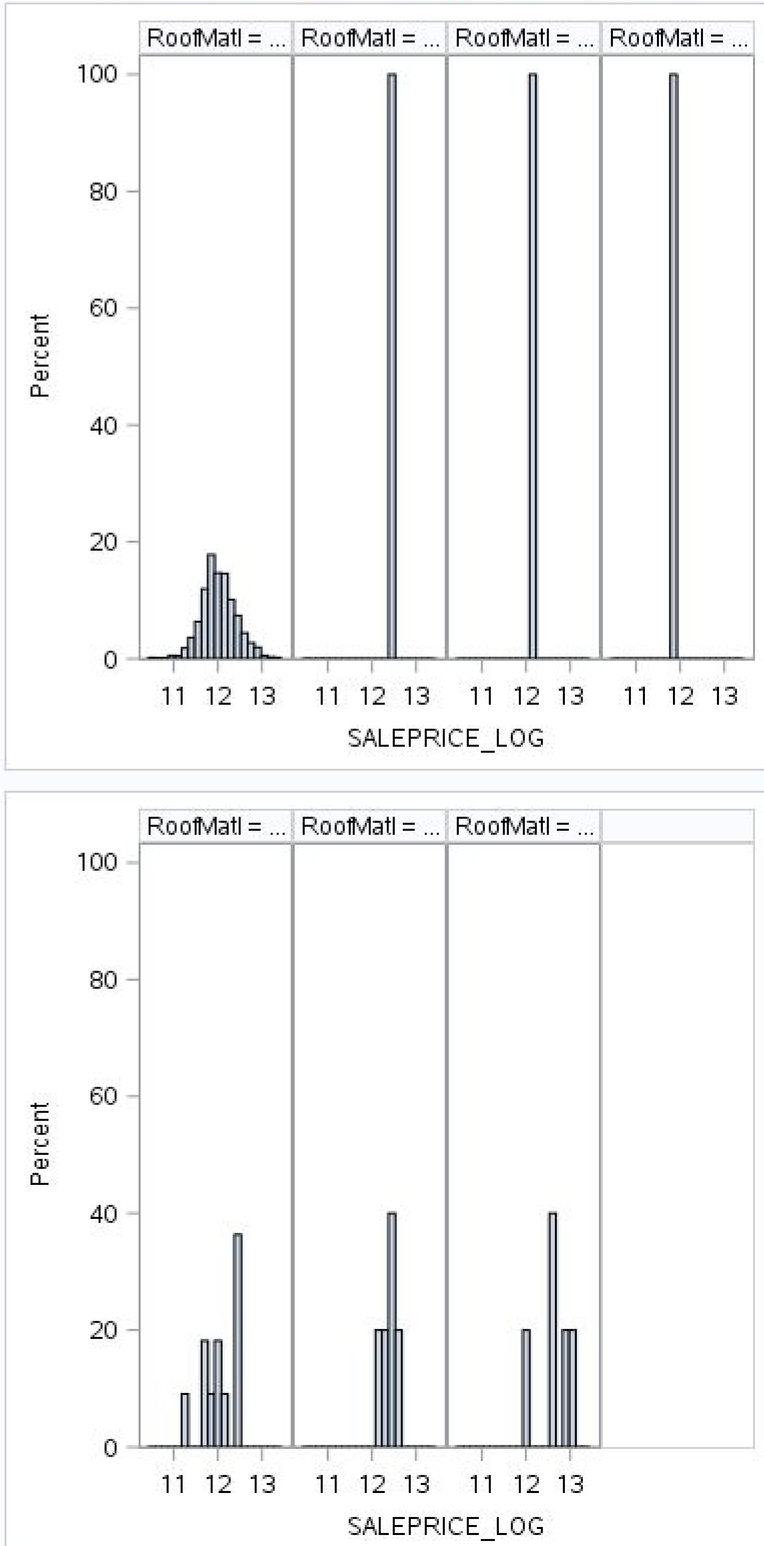

```
PROC SGPANEL DATA=TRAIN2 NOAUTOLEGEND;
      TITLE "Figure 17: Sales Price by Building Type";
      PANELBY BldgType;
      HISTOGRAM SALEPRICE_LOG;
      DENSITY SALEPRICE_LOG;
RUN;
```

Figure 17: Sale Price by Building Type

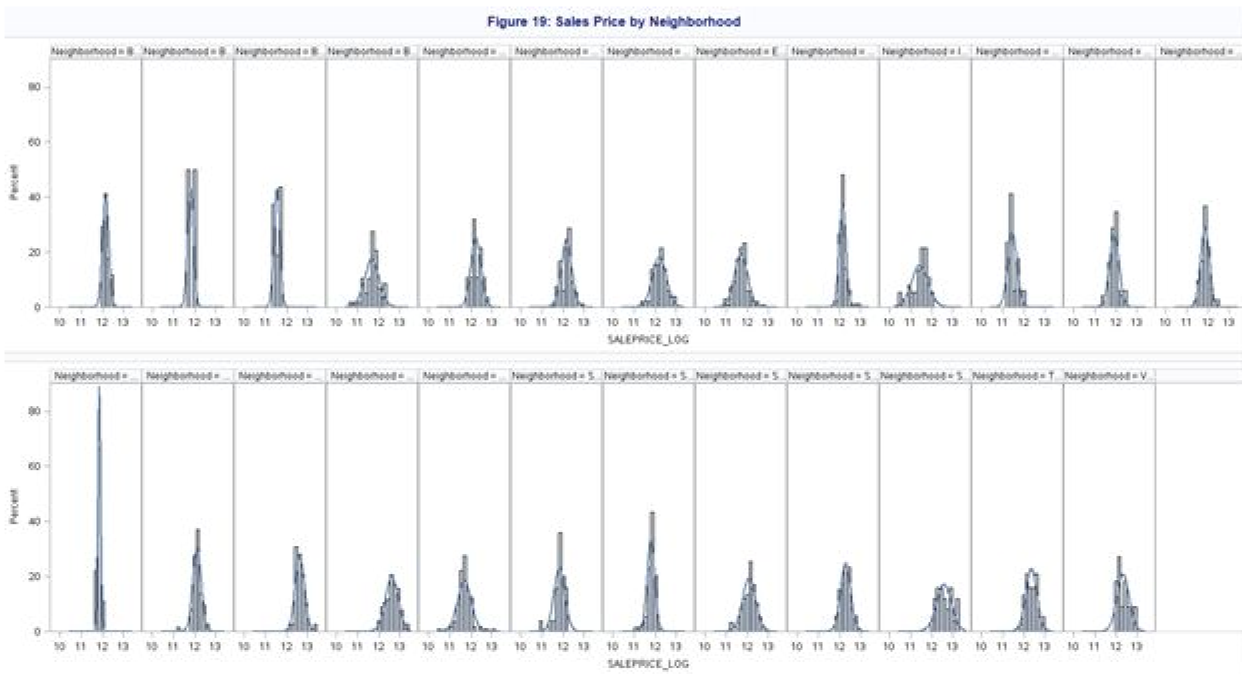



```
PROC SGPANEL DATA=TRAIN2(WHERE=(MISSING(ROOFMATL)=0)) NOAUTOLEGEND;
      TITLE "Figure 18: Sales Price by Roof Material";
      PANELBY ROOFMATL /COLUMNS=4;
      HISTOGRAM SALEPRICE_LOG;
RUN;
```

Figure 18: Sale Price by Roof Material



```
ODS GRAPHICS /HEIGHT=600 WIDTH=2400 IMAGEMAP;
PROC SGPANEL DATA=TRAIN2 NOAUTOLEGEND;
     TITLE "Figure 19: Sales Price by Neighborhood";
     PANELBY NEIGHBORHOOD / columns=13;
     HISTOGRAM SALEPRICE_LOG;
     DENSITY SALEPRICE_LOG;
RUN;
```



Note: It is difficult to read Figure 19. The point of the figure is to show how different the distributions of SALEPRICE_LOG look for each neighborhood. This is a significant variable to include in the model.

```
PROC GLMSELECT DATA=TRAIN2;
          CLASS EXTERQUAL BSMTQUAL KITCHENQUAL GARAGEFINISH
     GARAGETYPE HEATINGQC BSMTEXPOSURE LOTSHAPE GARAGECOND
     CENTRALAIR FOUNDATION NEIGHBORHOOD;
          MODEL SALEPRICE_LOG =  LOTAREA WOODDECKSF OPENPORCHSF
     FIREPLACES MASVNRAREA  GARAGEYRBLT YEARBUILT ROOMS GARAGEAREA
     GARAGECARS OVERALLQUAL SQFT_LOG EXTERQUAL BSMTQUAL KITCHENQUAL
     GARAGEFINISH GARAGETYPE HEATINGQC BSMTEXPOSURE LOTSHAPE
     GARAGECOND CENTRALAIR FOUNDATION NEIGHBORHOOD
       / SELECTION = FORWARD(STOP=CV) CVMETHOD=RANDOM(5)
STATS=ADJRSQ;
     TITLE 'Figure 20: Forward Selection Model';
RUN;
```

Figure 20: Summary
Forward Selection Model

Forward Selection Summary						
Step	Effect Entered	Number Effects In	Number Parms In	Adjusted R-Square	SBC	CV PRESS
0	Intercept	1	1	0.0000	-2664.9883	193.8435
1	OverallQual	2	2	0.6641	-4149.9911	65.2888
2	SQFT_LOG	3	3	0.8002	-4853.7698	38.7845
3	Neighborhood	4	27	0.8428	-5033.0914	31.8471
4	GarageArea	5	28	0.8535	-5122.7393	29.6788
5	KitchenQual	6	31	0.8622	-5188.0498	27.9850
6	LotArea	7	32	0.8679	-5239.5168	26.7793
7	CentralAir	8	33	0.8732	-5289.5237	25.6880
8	ROOMS	9	34	0.8783	-5338.9957	24.7700
9	OpenPorchSF	10	35	0.8814	-5368.5859	24.1556
10	YearBuilt	11	36	0.8840	-5392.2700	23.6960
11	WoodDeckSF	12	37	0.8861	-5411.8592	23.3086
12	Fireplaces	13	38	0.8881	-5429.3126	22.9509
13	HeatingQC	14	42	0.8903	-5431.9604	22.5515
14	BsmtExposure	15	46	0.8927	-5436.8589*	22.0446
15	GarageCars	16	47	0.8928	-5432.6911	22.0157
16	GarageType	17	52	0.8950*	-5429.5735	21.8002*
* Optimal Value of Criterion						

Selection stopped at a local minimum of the cross validation PRESS.

Stop Details				
Candidate For	Effect	Candidate CV PRESS		Compare CV PRESS
Entry	GarageYrBlt	21.8429	>	21.8002

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	45	173.46847	3.85485	253.48
Error	1321	20.08928	0.01521	
Corrected Total	1366	193.55776		

Root MSE	0.12332
Dependent Mean	12.05158
R-Square	0.8962
Adj R-Sq	0.8927
AIC	-4307.99612
AICC	-4304.57535
SBC	-5436.85892
CV PRESS	22.06034

```
PROC GLMSELECT DATA=TRAIN2;
     CLASS EXTERQUAL BSMTQUAL KITCHENQUAL GARAGEFINISH GARAGETYPE
     HEATINGQC BSMTEXPOSURE LOTSHAPE GARAGECOND CENTRALAIR
     FOUNDATION NEIGHBORHOOD;
     MODEL SALEPRICE_LOG =  LOTAREA WOODDECKSF OPENPORCHSF
     FIREPLACES MASVNRAREA  GARAGEYRBLT YEARBUILT ROOMS GARAGEAREA
     GARAGECARS OVERALLQUAL SQFT_LOG EXTERQUAL BSMTQUAL KITCHENQUAL
     GARAGEFINISH GARAGETYPE HEATINGQC BSMTEXPOSURE LOTSHAPE
     GARAGECOND CENTRALAIR FOUNDATION NEIGHBORHOOD
     / SELECTION = BACKWARD(STOP=CV) CVMETHOD=RANDOM(5) STATS=ADJRSQ;
     TITLE 'Figure 21: Backward Selection Model';
RUN;
```

Figure 21: Summary
Backward Elimination Model

Step	Effect Removed	Number Effects In	Number Parms In	Adjusted R-Square	SBC	CV PRESS
0		25	75	0.8968*	-5311.0668	21.9172
1	Foundation	24	70	0.8966	-5339.9249	21.8334
2	GarageCond	23	66	0.8964	-5362.1820	21.8092
3	BsmtQual	22	62	0.8958	-5377.7975	21.7674
4	GarageFinish	21	60	0.8959	-5391.5433*	21.6955*

* Optimal Value of Criterion

Selection stopped at a local minimum of the cross validation PRESS.

Stop Details

Candidate For	Effect	Candidate CV PRESS		Compare CV PRESS
Removal	LotShape	21.7832	>	21.6955

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	59	174.27156	2.95376	200.17
Error	1307	19.28619	0.01476	
Corrected Total	1366	193.55776		

Root MSE	0.12147
Dependent Mean	12.05158
R-Square	0.9004
Adj R-Sq	0.8959
AIC	-4335.76569
AICC	-4329.96952
SBC	-5391.54326
CV PRESS	21.69550

```
PROC GLMSELECT DATA=TRAIN2;
      CLASS EXTERQUAL BSMTQUAL KITCHENQUAL GARAGEFINISH
    GARAGETYPE HEATINGQC BSMTEXPOSURE LOTSHAPE GARAGECOND
    CENTRALAIR FOUNDATION NEIGHBORHOOD;
      MODEL SALEPRICE_LOG =  LOTAREA WOODDECKSF OPENPORCHSF
    FIREPLACES MASVNRAREA  GARAGEYRBLT YEARBUILT ROOMS GARAGEAREA
    GARAGECARS OVERALLQUAL SQFT_LOG EXTERQUAL BSMTQUAL KITCHENQUAL
    GARAGEFINISH GARAGETYPE HEATINGQC BSMTEXPOSURE LOTSHAPE
    GARAGECOND CENTRALAIR FOUNDATION NEIGHBORHOOD
    / SELECTION = STEPWISE(STOP=CV) CVMETHOD=RANDOM(5) STATS=ADJRSQ;
    TITLE 'Figure 22: Stepwise Selection Model';
RUN;
```

Figure 22: Summary
Stepwise Selection Model

Stepwise Selection Summary							
Step	Effect Entered	Effect Removed	Number Effects In	Number Parms In	Adjusted R-Square	SBC	CV PRESS
0	Intercept		1	1	0.0000	-2664.9883	194.2986
1	OverallQual		2	2	0.6641	-4149.9911	65.2742
2	SQFT_LOG		3	3	0.8002	-4853.7698	38.7690
3	Neighborhood		4	27	0.8428	-5033.0914	31.5664
4	GarageArea		5	28	0.8535	-5122.7393	29.4708
5	KitchenQual		6	31	0.8622	-5188.0498	27.6232
6	LotArea		7	32	0.8679	-5239.5168	26.5829
7	CentralAir		8	33	0.8732	-5289.5237	25.7274
8	ROOMS		9	34	0.8783	-5338.9957	24.7889
9	OpenPorchSF		10	35	0.8814	-5368.5859	24.1254
10	YearBuilt		11	36	0.8840	-5392.2700	23.7349
11	WoodDeckSF		12	37	0.8861	-5411.8592	23.2176
12	Fireplaces		13	38	0.8881	-5429.3126	22.8090
13	HeatingQC		14	42	0.8903	-5431.9604	22.4108
14	BsmtExposure		15	46	0.8927*	-5436.8589*	22.0022*
* Optimal Value of Criterion							

Selection stopped as adding or dropping any effect does not improve the selection criterion.

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	45	173.46847	3.85485	253.48
Error	1321	20.08928	0.01521	
Corrected Total	1366	193.55776		

Root MSE	0.12332
Dependent Mean	12.05158
R-Square	0.8962
Adj R-Sq	0.8927
AIC	-4307.99612
AICC	-4304.57535
SBC	-5436.85892
CV PRESS	22.00219

```
* CUSTOM MODEL;
PROC GLM DATA=TRAIN2 PLOTS=ALL;
      CLASS NEIGHBORHOOD BLDGTYPE ROOFMATL CENTRALAIR;
            MODEL SALEPRICE_LOG = OVERALLQUAL OVERALLCOND YEARBUILT
      ROOFMATL BSMTFINSF1 TOTALBSMTSF GRLIVAREA_LOG CENTRALAIR
      NEIGHBORHOOD | BLDGTYPE / SOLUTION CLPARM ;
RUN;

PROC GLMSELECT DATA=TRAIN2 PLOTS=ALL;
      CLASS NEIGHBORHOOD BLDGTYPE ROOFMATL CENTRALAIR;
            MODEL SALEPRICE_LOG = OVERALLQUAL OVERALLCOND YEARBUILT
      ROOFMATL BSMTFINSF1 TOTALBSMTSF GRLIVAREA_LOG CENTRALAIR
      NEIGHBORHOOD | BLDGTYPE / SELECTION=NONE STATS=PRESS;
RUN;
```

Figure 23a: Summary Custom Model

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	76	207.2343962	2.7267684	178.85	<.0001
Error	1379	21.0249043	0.0152465		
Corrected Total	1455	228.2593005			

R-Square	Coeff Var	Root MSE	SALEPRICE_LOG Mean
0.907890	1.027094	0.123477	12.02194

Source	DF	Type I SS	Mean Square	F Value	Pr > F
OverallQual	1	153.1972213	153.1972213	10048.0	<.0001
OverallCond	1	0.3228806	0.3228806	21.18	<.0001
YearBuilt	1	7.0987914	7.0987914	465.60	<.0001
RoofMatl	6	0.5707126	0.0951188	6.24	<.0001
BsmtFinSF1	1	7.8770132	7.8770132	516.64	<.0001
TotalBsmtSF	1	6.8543041	6.8543041	449.57	<.0001
GRLIVAREA_LOG	1	23.6090418	23.6090418	1548.49	<.0001
CentralAir	1	0.2263898	0.2263898	14.85	0.0001
Neighborhood	24	4.9291890	0.2053829	13.47	<.0001
BldgType	4	1.3259670	0.3314917	21.74	<.0001
Neighborhoo*BldgType	35	1.2228853	0.0349396	2.29	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
OverallQual	1	3.17565599	3.17565599	208.29	<.0001
OverallCond	1	4.01760013	4.01760013	263.51	<.0001
YearBuilt	1	2.04432422	2.04432422	134.08	<.0001
RoofMatl	6	0.27144371	0.04524062	2.97	0.0070
BsmtFinSF1	1	2.06604098	2.06604098	135.51	<.0001
TotalBsmtSF	1	1.62255449	1.62255449	106.42	<.0001
GRLIVAREA_LOG	1	16.96831466	16.96831466	1112.93	<.0001
CentralAir	1	0.21012999	0.21012999	13.78	0.0002
Neighborhood	24	2.21429782	0.09226241	6.05	<.0001
BldgType	4	0.29798939	0.07449735	4.89	0.0006
Neighborhoo*BldgType	35	1.22288532	0.03493958	2.29	<.0001

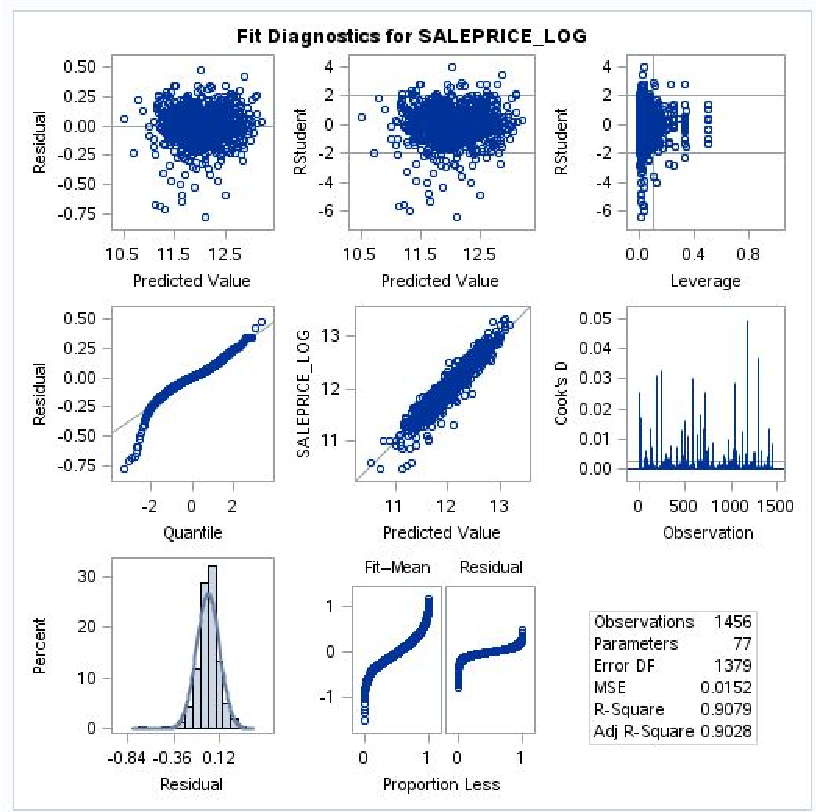
Root MSE	0.12348
Dependent Mean	12.02194
R-Square	0.9079
Adj R-Sq	0.9028
AIC	-4558.15028
AICC	-4549.20039
PRESS	23.30700
SBC	-5609.32476

Figure 23b: Summary Custom Model

Parameter		Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept		3.923520375	B	0.56029422	7.00	<.0001	2.824399185	5.022641564
OverallQual		0.064942062		0.00449981	14.43	<.0001	0.056114848	0.073769276
OverallCond		0.057774011		0.00355905	16.23	<.0001	0.050792274	0.064755748
YearBuilt		0.003208109		0.00027705	11.58	<.0001	0.002664623	0.003751595
RoofMatl	CompShg	-0.205230139	B	0.05714789	-3.59	0.0003	-0.317336346	-0.093123931
RoofMatl	Membran	0.005313558	B	0.13746554	0.04	0.9692	-0.264350639	0.274977754
RoofMatl	Metal	-0.021769180	B	0.13763727	-0.16	0.8744	-0.291770255	0.248231895
RoofMatl	Roll	-0.245072442	B	0.13956431	-1.76	0.0793	-0.518853756	0.028708872
RoofMatl	Tar&Grv	-0.186090711	B	0.06815668	-2.73	0.0064	-0.319792701	-0.052388722
RoofMatl	WdShake	-0.236646400	B	0.07952386	-2.98	0.0030	-0.392647214	-0.080645585
RoofMatl	WdShngl	0.000000000	B	.	.	.	.	.
BsmtFinSF1		0.000108086		0.00000929	11.64	<.0001	0.000089872	0.000126301
TotalBsmtSF		0.000118458		0.00001148	10.32	<.0001	0.000095932	0.000140983
GRLIVAREA_LOG		0.492503503		0.01476301	33.36	<.0001	0.463543121	0.521463885
CentralAir	N	-0.060481041	B	0.01629146	-3.71	0.0002	-0.092439775	-0.028522308
CentralAir	Y	0.000000000	B	.	.	.	.	.
Neighborhood	Blmngtn	-0.216061192	B	0.07811600	-2.77	0.0058	-0.369300243	-0.062822142
Neighborhood	Blueste	-0.411451065	B	0.14305250	-2.88	0.0041	-0.692075114	-0.130827015
Neighborhood	BrDale	-0.411839514	B	0.09518495	-4.33	<.0001	-0.598562478	-0.225116549
Neighborhood	BrkSide	-0.066750591	B	0.15951208	-0.42	0.6757	-0.379663173	0.246161992

Note: Figure 23b only includes a partial screenshot of the parameter estimates table.

Figure 24: Residual Plots
Custom Model



Code for Kaggle Submission

```
* KAGGLE SUBMISSIONS;
DATA TEST2;
      SET TEST;

      SALEPRICE_LOG = LOG(SALEPRICE);

      BATHROOMS = .5*HALFBATH + FULLBATH;
      ROOMS = BATHROOMS + TOTRMSABVGRD;

      SQFT = (BSMTFINSF1 + GRLIVAREA)/100;
      SQFT_LOG = LOG(SQFT);

      GRLIVAREA = GRLIVAREA/100;
      GRLIVAREA_LOG = LOG(GRLIVAREA);
RUN;

* COMBINE TRAIN AND TEST DATASETS;
DATA KAGGLE;
```

```
     SET TRAIN2 TEST2;
RUN;

* CALCULATE MEAN BY NEIGHBORHOOD FOR ANY MISSING PREDICTIONS;
PROC SQL;
     CREATE TABLE MEAN_PRICE AS
     SELECT NEIGHBORHOOD
          ,AVG(SALEPRICE)
     FROM TRAIN2
     GROUP BY 1
     ;
QUIT;

* FORWARD SELECTION MODEL;
PROC GLMSELECT DATA=KAGGLE;
          CLASS EXTERQUAL BSMTQUAL KITCHENQUAL GARAGEFINISH
     GARAGETYPE HEATINGQC BSMTEXPOSURE LOTSHAPE GARAGECOND
     CENTRALAIR FOUNDATION NEIGHBORHOOD;
          MODEL SALEPRICE_LOG =  LOTAREA WOODDECKSF OPENPORCHSF
     FIREPLACES MASVNRAREA  GARAGEYRBLT YEARBUILT ROOMS GARAGEAREA
     GARAGECARS OVERALLQUAL SQFT_LOG EXTERQUAL BSMTQUAL KITCHENQUAL
     GARAGEFINISH GARAGETYPE HEATINGQC BSMTEXPOSURE LOTSHAPE
     GARAGECOND CENTRALAIR FOUNDATION NEIGHBORHOOD
       / SELECTION = FORWARD(STOP=CV) CVMETHOD=RANDOM(5)
STATS=ADJRSQ;
     OUTPUT OUT=RESULTS_FORWARD P=PREDICT;
RUN;

* BACKWARD SELECTION MODEL;
PROC GLMSELECT DATA=KAGGLE;
          CLASS EXTERQUAL BSMTQUAL KITCHENQUAL GARAGEFINISH
     GARAGETYPE HEATINGQC BSMTEXPOSURE LOTSHAPE GARAGECOND
     CENTRALAIR FOUNDATION NEIGHBORHOOD;
          MODEL SALEPRICE_LOG =  LOTAREA WOODDECKSF OPENPORCHSF
     FIREPLACES MASVNRAREA  GARAGEYRBLT YEARBUILT ROOMS GARAGEAREA
     GARAGECARS OVERALLQUAL SQFT_LOG EXTERQUAL BSMTQUAL KITCHENQUAL
     GARAGEFINISH GARAGETYPE HEATINGQC BSMTEXPOSURE LOTSHAPE
     GARAGECOND CENTRALAIR FOUNDATION NEIGHBORHOOD
            / SELECTION = BACKWARD(STOP=CV) CVMETHOD=RANDOM(5)
     STATS=ADJRSQ;
     OUTPUT OUT=RESULTS_BACKWARD P=PREDICT;
RUN;

* STEPWISE SELECTION MODEL;
PROC GLMSELECT DATA=KAGGLE;
```

```
     CLASS EXTERQUAL BSMTQUAL KITCHENQUAL GARAGEFINISH
  GARAGETYPE HEATINGQC BSMTEXPOSURE LOTSHAPE GARAGECOND
  CENTRALAIR FOUNDATION NEIGHBORHOOD;
     MODEL SALEPRICE_LOG =  LOTAREA WOODDECKSF OPENPORCHSF
  FIREPLACES MASVNRAREA  GARAGEYRBLT YEARBUILT ROOMS GARAGEAREA
  GARAGECARS OVERALLQUAL SQFT_LOG EXTERQUAL BSMTQUAL KITCHENQUAL
  GARAGEFINISH GARAGETYPE HEATINGQC BSMTEXPOSURE LOTSHAPE
  GARAGECOND CENTRALAIR FOUNDATION NEIGHBORHOOD
        / SELECTION = STEPWISE(STOP=CV) CVMETHOD=RANDOM(5)
  STATS=ADJRSQ;
  OUTPUT OUT=RESULTS_STEPWISE P=PREDICT;
RUN;

* CUSTOM MODEL;
PROC GLMSELECT DATA=kaggle PLOTS=ALL;
  CLASS NEIGHBORHOOD BLDGTYPE ROOFMATL CENTRALAIR;
     MODEL SALEPRICE_LOG = OVERALLQUAL OVERALLCOND YEARBUILT
  ROOFMATL BSMTFINSF1 TOTALBSMTSF GRLIVAREA_LOG CENTRALAIR
  NEIGHBORHOOD | BLDGTYPE / SELECTION=NONE CVMETHOD=RANDOM(5)
  stats=press;
  OUTPUT OUT=RESULTS_CUSTOM P=PREDICT;
RUN;

%MACRO FILE_SUBMISSION(FILE);
DATA RESULTS2;
SET &FILE;
IF ID > 1460;
SALEPRICE = EXP(PREDICT);

* REPLACE ANY MISSING PREDICTIONS WITH THE MEAN SALES PRICE FOR THE
NEIGHBORHOOD;
if missing(predict) = 1 then do;
if neighborhood = "Blmngtn" then saleprice= 194870.8824;
else if neighborhood = "Blueste" then saleprice=137500;
else if neighborhood = "BrDale" then saleprice=  104493.75;
else if neighborhood = "BrkSide" then saleprice=124834.0517;
else if neighborhood = "ClearCr" then saleprice=212565.4286;
else if neighborhood = "CollgCr" then saleprice=197965.7733;
else if neighborhood = "Crawfor" then saleprice=210624.7255;
else if neighborhood = "Edwards" then saleprice=127318.5714;
else if neighborhood = "Gilbert" then saleprice=192854.5063;
else if neighborhood = "IDOTRR" then saleprice=  100123.7838;
else if neighborhood = "MeadowV" then saleprice=98576.47059;
else if neighborhood = "Mitchel" then saleprice=156270.1225;
else if neighborhood = "NAmes" then saleprice=   145847.08;
else if neighborhood = "NPkVill" then saleprice=142694.4444;
```

```
else if neighborhood = "NWAmes" then saleprice=  189050.0685;
else if neighborhood = "NoRidge" then saleprice=314028.4103;
else if neighborhood = "NridgHt" then saleprice=316270.6234;
else if neighborhood = "OldTown" then saleprice=128225.3009;
else if neighborhood = "SWISU" then saleprice=   142591.36;
else if neighborhood = "Sawyer" then saleprice=  136793.1351;
else if neighborhood = "SawyerW" then saleprice=186555.7966;
else if neighborhood = "Somerst" then saleprice=225379.8372;
else if neighborhood = "StoneBr" then saleprice=310499;
else if neighborhood = "Timber" then saleprice=  242247.4474;
else if neighborhood = "Veenker" then saleprice=238772.7273;
end;

KEEP ID SALEPRICE;
RUN;

PROC EXPORT DATA=RESULTS2
FILE="/data/bnsf/ib/hubops/jford/data_science/kaggle/&FILE..csv"
replace;
RUN;
%MEND;

DATA LIST;
     LENGTH FILE $16.;
     FILE="RESULTS_FORWARD"; OUTPUT;
     FILE="RESULTS_BACKWARD"; OUTPUT;
     FILE="RESULTS_STEPWISE"; OUTPUT;
     FILE="RESULTS_CUSTOM"; OUTPUT;
RUN;

DATA _NULL_;
     SET LIST;
     CALL EXECUTE('%FILE_SUBMISSION('||FILE||')');
RUN;
```