# **Kaggle Project**

## **House Prices: Advanced Regression Techniques**

Sangrae Cho, Dan Crouthamel, Jenna Ford, Christian Nava

August 7, 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]](#footnote-1)

## **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 sales price of a house is related to the square footage of its living area. Also, the client wants to know if the sales price depends on the house’s neighborhood.

###### Solution Outline

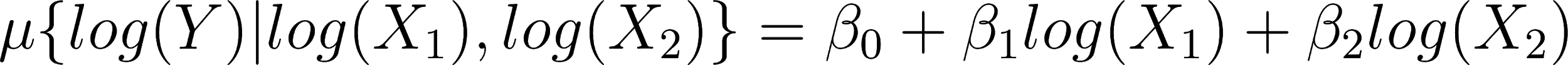
As linear regression techniques have been successful in predicting housing prices, in this project, they will be used to provide an estimate of the relationship between square footage of a house and the square footage of its living area.

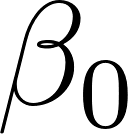
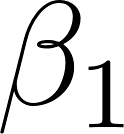
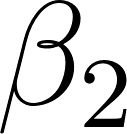
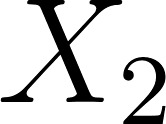
The metric used to evaluate submissions for this Kaggle competition is the Root-Mean-Squared-Error (RMSE) between the logarithm of the predicted value and the logarithm of the observed sales price.

**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 of the data where both the response and explanatory (predictor) variables are logged and is given by the following

[](https://www.codecogs.com/eqnedit.php?latex=%5Cmu%5C%7Blog(Y)%7Clog(X_1)%2Clog(X_2)%5C%7D%20%3D%20%5Cbeta_0%2B%5Cbeta_1log(X_1)%2B%5Cbeta_2log(X_2)%250)

where [](https://www.codecogs.com/eqnedit.php?latex=%5Cbeta_0%250) is the intercept from the linear regression equation, [](https://www.codecogs.com/eqnedit.php?latex=%5Cbeta_1%250) and [](https://www.codecogs.com/eqnedit.php?latex=%5Cbeta_2%250) are the regression coefficients , and [](https://www.codecogs.com/eqnedit.php?latex=X_1%250) and [](https://www.codecogs.com/eqnedit.php?latex=X_2%250) are explanatory variables.

NAmes: 10.6711 + .47302\*LogLiveArea

Edwards: 10.0239 + 0.67333\*LogLiveArea

BrkSide: 9.68755 + 0.81964\*LogLiveArea

**Checking Assumptions**

##### *Linear Relationship*

Assumption is satisfied per the quantile plot (see Appendix).

##### *Normality*

Per the histogram of the residuals 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 1.10.

##### *Independence*

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

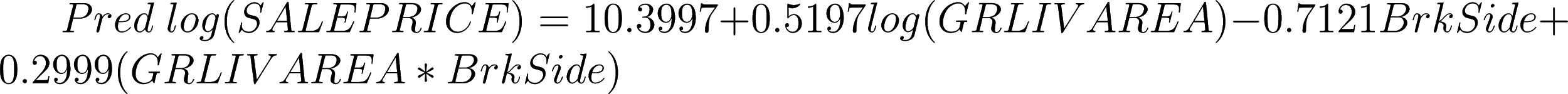
##### *Outliers*

**Comparing Competing Models**

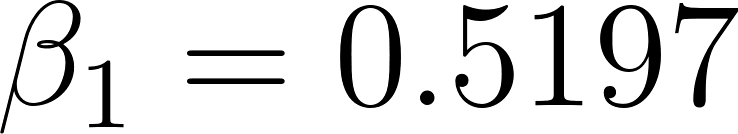
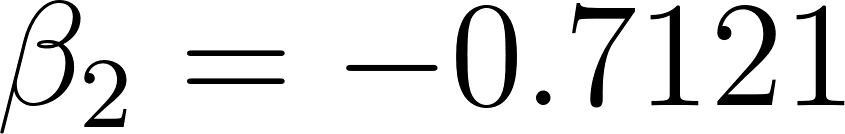
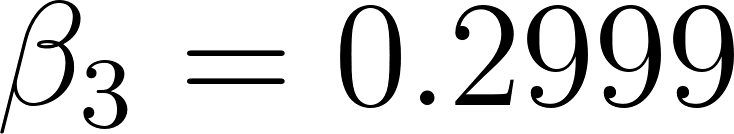
##### *Adj R2*

The adjusted r-squared statistic, or the coefficient of determination, is the percentage of the total variation in the response variable that is accounted for or explained by the explanatory variables. In our case, 50.56% of the variation in log SALEPRICE was explained by

**Parameters**

[****](https://www.codecogs.com/eqnedit.php?latex=Pred%5C%3Blog(SALEPRICE)%20%3D%2010.3997%2B0.5197log(GRLIVAREA)-0.7121BrkSide%2B0.2999(GRLIVAREA*BrkSide)%250)

##### *Estimates*

[**](https://www.codecogs.com/eqnedit.php?latex=%5Cbeta_0%3D10.3397%250), [](https://www.codecogs.com/eqnedit.php?latex=%5Cbeta_1%3D0.5197%250), [](https://www.codecogs.com/eqnedit.php?latex=%5Cbeta_2%3D%20-%200.7121%250), [](https://www.codecogs.com/eqnedit.php?latex=%5Cbeta_3%3D0.2999%250)

##### *Interpretation*

Holding all other variables constant, it is estimated that a 10-fold increase in the square footage of the living area is associated with a ( = 3.3088) increase in the median sales price of a home in Ames, Iowa (p-value < .0001). A 95% confidence interval for the multiplicative increase is

**Conclusion**

## **III. Analysis Question 2**

### **Problem Statement**

## **IV. 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**;



\*EDA AND REGRESSIONS WITH TRANSFORMATIONS;

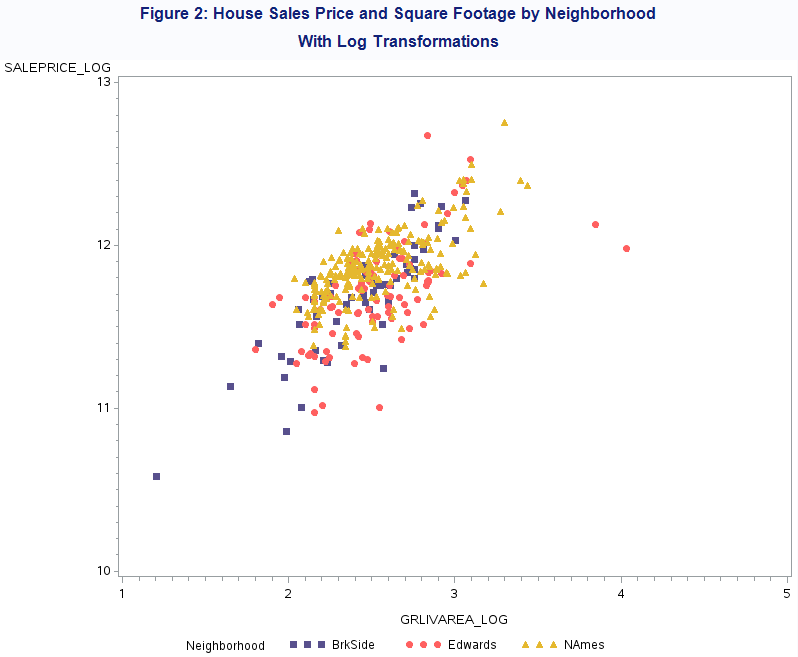
**PROC** **GPLOT** DATA=TRAIN2;

PLOT SALEPRICE\_LOG\*GRLIVAREA\_LOG=NEIGHBORHOOD;

TITLE1 'Figure 2: House Sales Price and Square Footage by Neighborhood';

TITLE2 'With Log Transformations';

**RUN**;



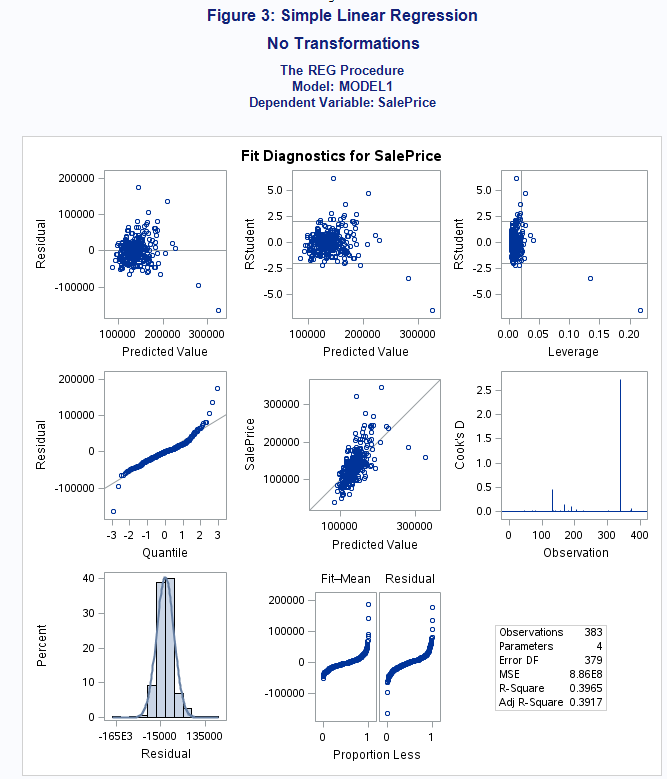
**PROC** **REG** DATA=TRAIN2;

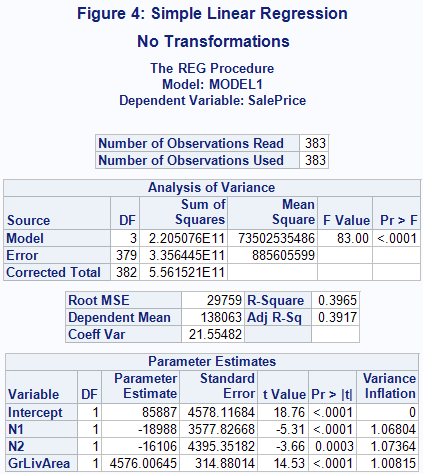
MODEL SALEPRICE = N1 N2 GRLIVAREA /VIF;

TITLE1 'Figure 3: Simple Linear Regression';

TITLE2 'No Transformations';

**RUN**;





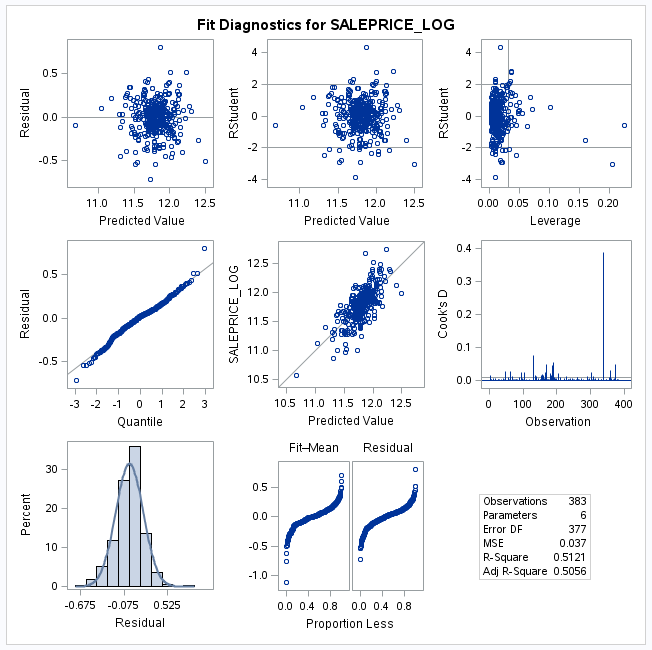
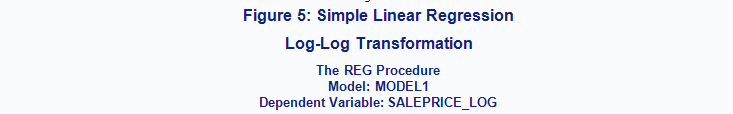
**PROC** **REG** DATA=TRAIN2 plots=all;

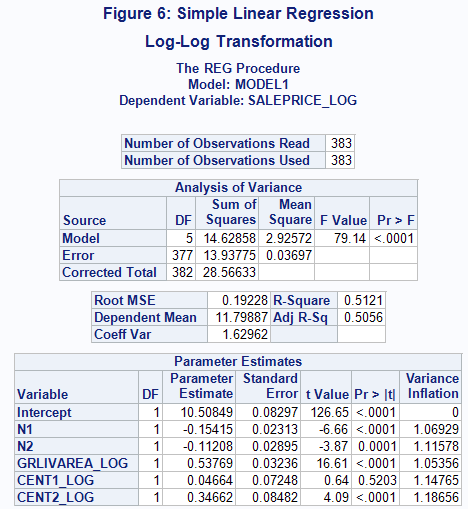
MODEL SALEPRICE\_LOG = N1 N2 GRLIVAREA\_LOG CENT1\_LOG CENT2\_LOG/VIF;

TITLE1 'Figure 5: Simple Linear Regression';

TITLE2 'Log-Log Transformation';

**RUN**;





\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\* 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**;

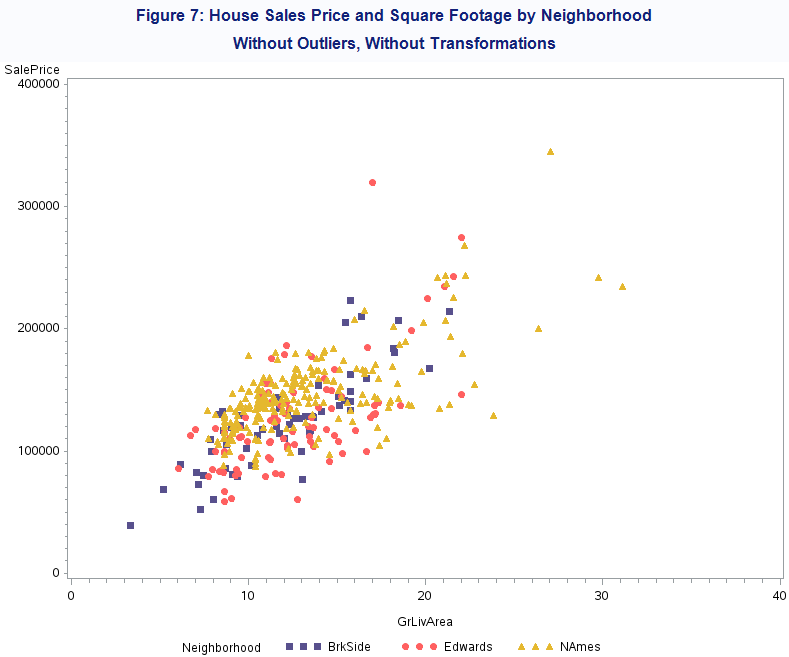
**PROC** **GPLOT** DATA=TRAIN\_NO\_OUTLIERS1;

PLOT SALEPRICE \* GRLIVAREA=NEIGHBORHOOD;

TITLE1 'Figure 7: House Sales Price and Square Footage by Neighborhood';

TITLE2 'Without Outliers, Without Transformations';

**RUN**;



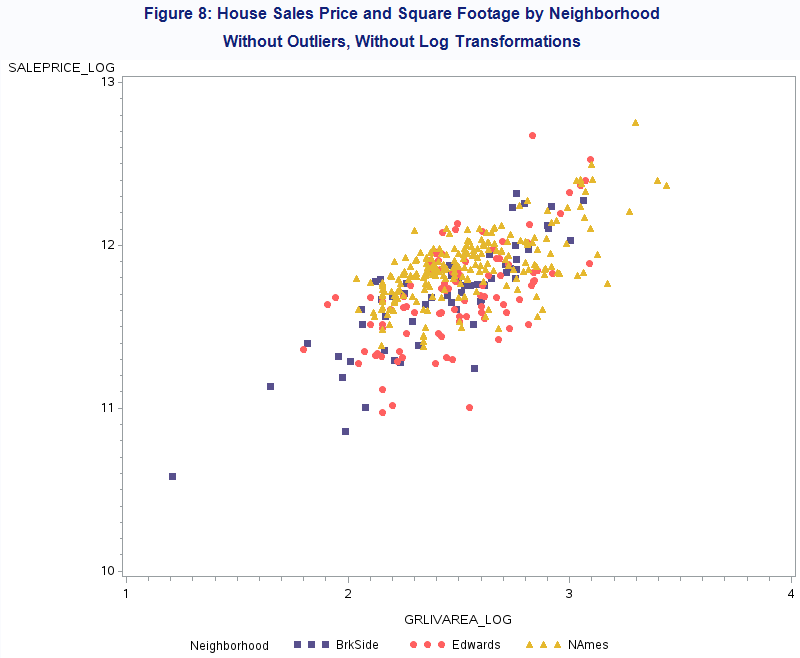
**PROC** **GPLOT** DATA=TRAIN\_NO\_OUTLIERS1;

PLOT SALEPRICE\_LOG \* GRLIVAREA\_LOG=NEIGHBORHOOD;

TITLE1 'Figure 8: House Sales Price and Square Footage by Neighborhood';

TITLE2 'Without Outliers, Without Log Transformations';

**RUN**;



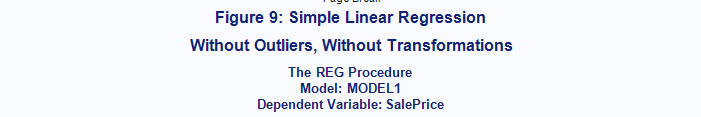
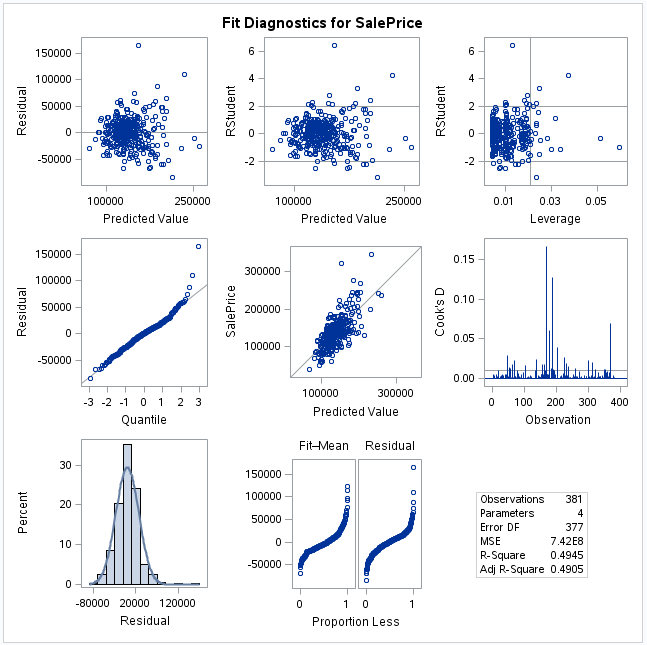
**PROC** **REG** DATA=TRAIN\_NO\_OUTLIERS1 PLOTS=ALL;

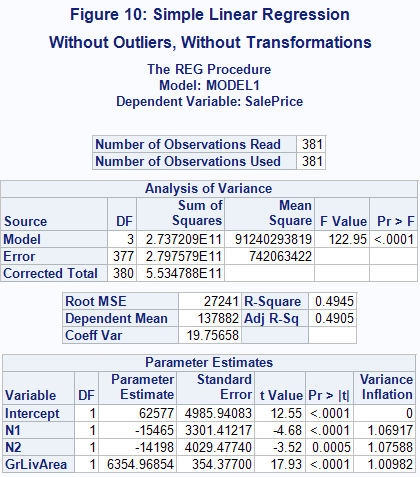
MODEL SALEPRICE = N1 N2 GRLIVAREA /VIF;

TITLE1 'Figure 9: Simple Linear Regression';

TITLE2 'Without Outliers, Without Transformations';

**RUN**;



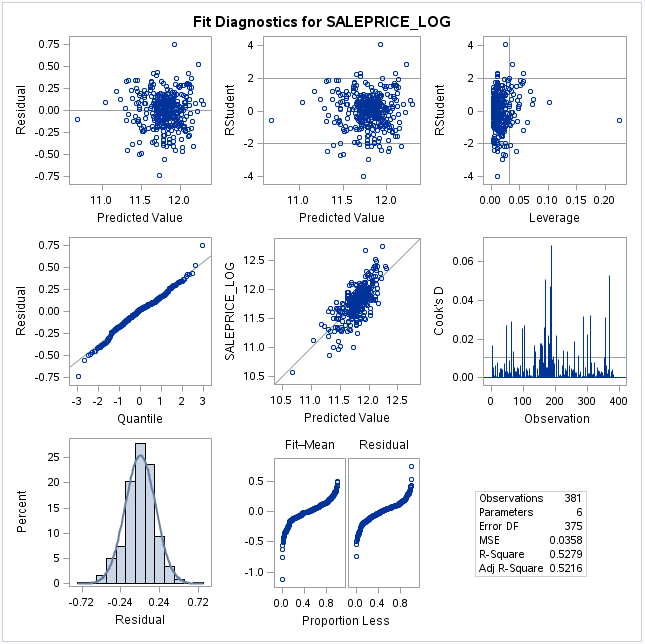
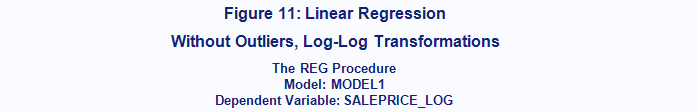
**PROC** **REG** DATA=TRAIN\_NO\_OUTLIERS1 PLOTS=ALL;

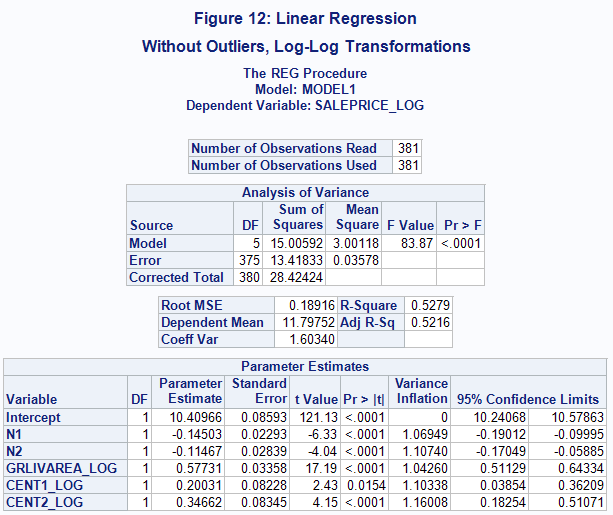
MODEL SALEPRICE\_LOG = N1 N2 GRLIVAREA\_LOG CENT1\_LOG CENT2\_LOG/VIF CLB CLM CLI; \*/

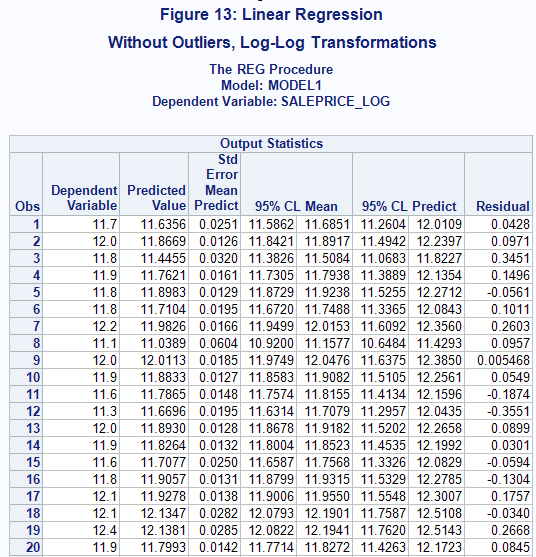
TITLE1 'Figure 11: Linear Regression';

TITLE2 'Without Outliers, Log-Log Transformations';

**RUN**;







**PROC** **SGPLOT** DATA=TRAIN\_NO\_OUTLIERS1 ;

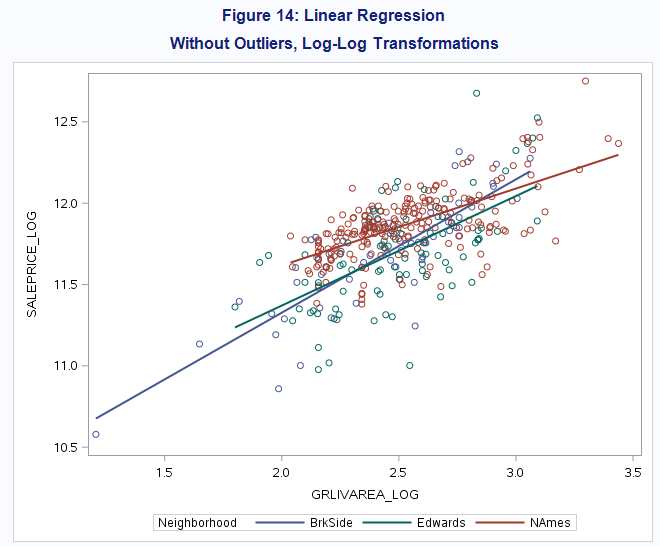
TITLE "Sales Price by Living Area and Neighborhood";

REG Y=SALEPRICE\_LOG X=GRLIVAREA\_LOG / GROUP=NEIGHBORHOOD ;

TITLE1 'Figure 14: Linear Regression';

TITLE2 'Without Outliers, Log-Log Transformations';

**RUN**;



**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**;

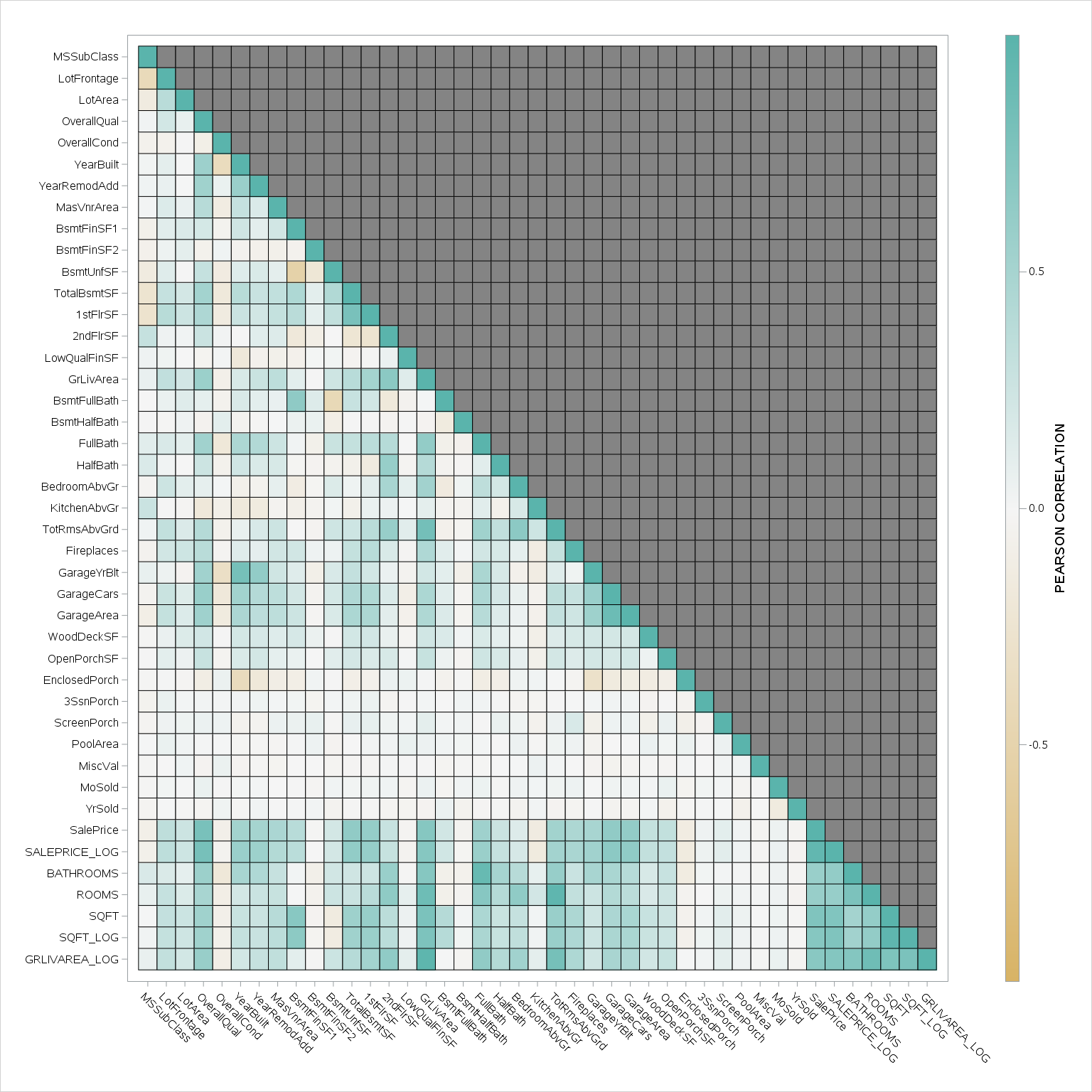
ODS GRAPHICS /HEIGHT=**2400** WIDTH=**2400** IMAGEMAP;

%***PREPCORRDATA***(IN=TRAIN2(DROP=ID) ,OUT=CORR\_MATRIX);

**PROC** **SGRENDER** DATA=CORR\_MATRIX TEMPLATE=CORRHEATMAP;

DYNAMIC \_TITLE\_="CORR MATRIX";

**RUN**;



\* EDA FOR CATEGORICAL VARIABLES;

ODS GRAPHICS /HEIGHT=**600** WIDTH=**600** IMAGEMAP;

**PROC** **SGPANEL** DATA=TRAIN2 NOAUTOLEGEND;

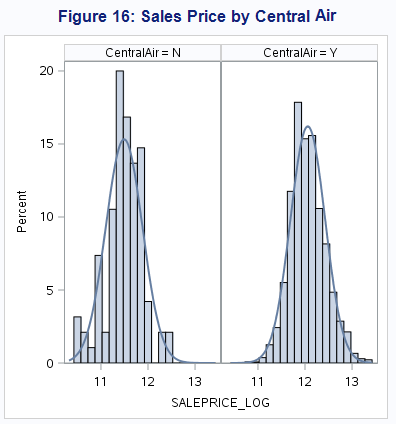
TITLE "Figure 16: Sales Price by Central Air";

PANELBY CENTRALAIR;

HISTOGRAM SALEPRICE\_LOG;

DENSITY SALEPRICE\_LOG;

**RUN**;



**PROC** **SGPANEL** DATA=TRAIN2 NOAUTOLEGEND;

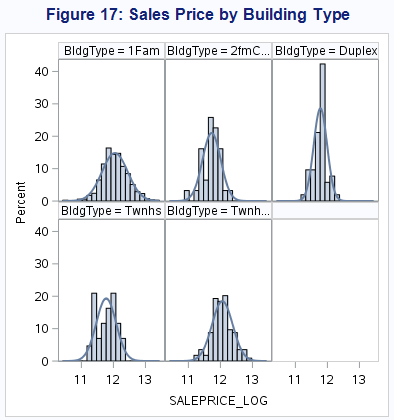
TITLE "Figure 17: Sales Price by Building Type";

PANELBY BldgType;

HISTOGRAM SALEPRICE\_LOG;

DENSITY SALEPRICE\_LOG;

**RUN**;



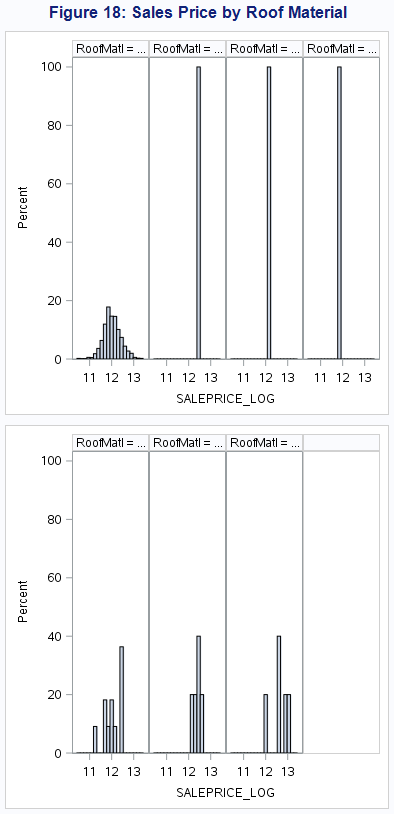
**PROC** **SGPANEL** DATA=TRAIN2(WHERE=(MISSING(ROOFMATL)=**0**)) NOAUTOLEGEND;

TITLE "Figure 18: Sales Price by Roof Material";

PANELBY ROOFMATL /COLUMNS=**4**;

HISTOGRAM SALEPRICE\_LOG;

**RUN**;



ODS GRAPHICS /HEIGHT=**600** WIDTH=**2400** IMAGEMAP;

**PROC** **SGPANEL** DATA=TRAIN2 NOAUTOLEGEND;

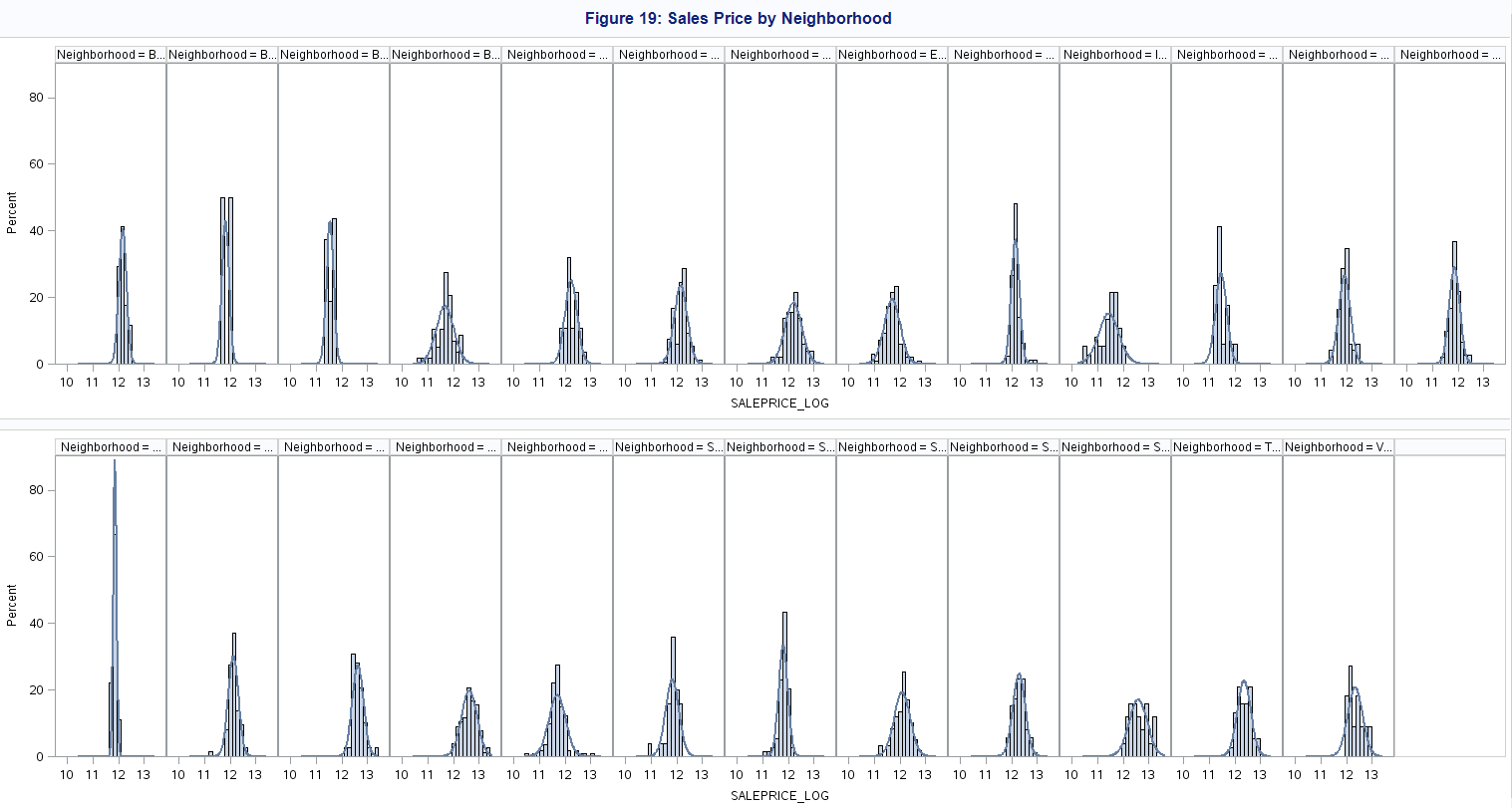
TITLE "Figure 19: Sales Price by Neighborhood";

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;

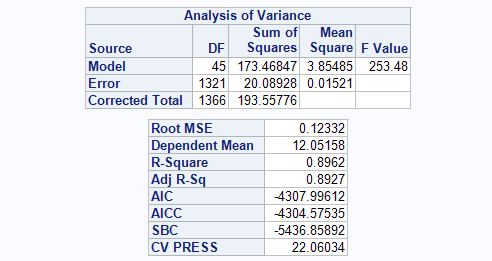
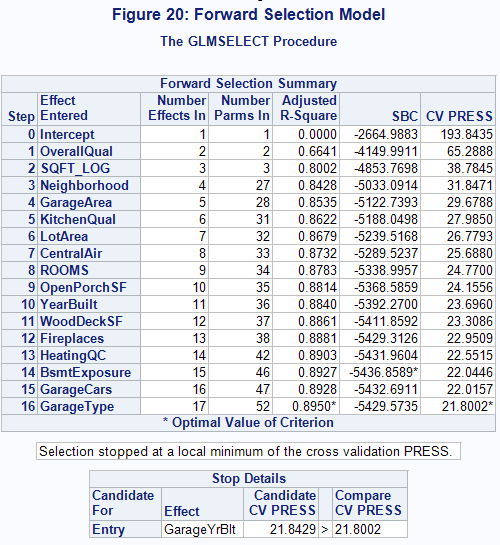
CLASS EXTERQUAL BSMTQUAL KITCHENQUAL GARAGEFINISH GARAGETYPE HEATINGQC BSMTEXPOSURE LOTSHAPE GARAGECOND CENTRALAIR FOUNDATION NEIGHBORHOOD;

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**;



**PROC** **GLMSELECT** DATA=TRAIN2;

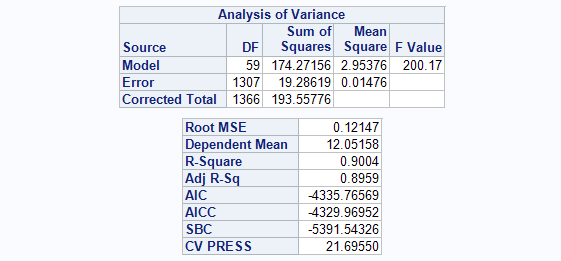
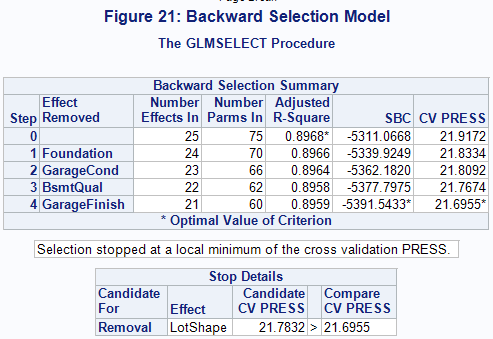
CLASS EXTERQUAL BSMTQUAL KITCHENQUAL GARAGEFINISH GARAGETYPE HEATINGQC BSMTEXPOSURE LOTSHAPE GARAGECOND CENTRALAIR FOUNDATION NEIGHBORHOOD;

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**;



**PROC** **GLMSELECT** DATA=TRAIN2;

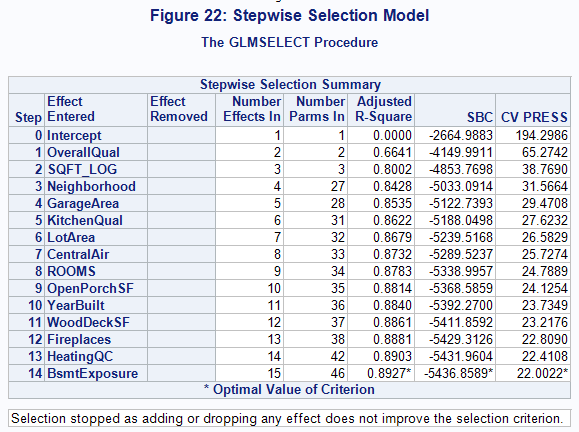
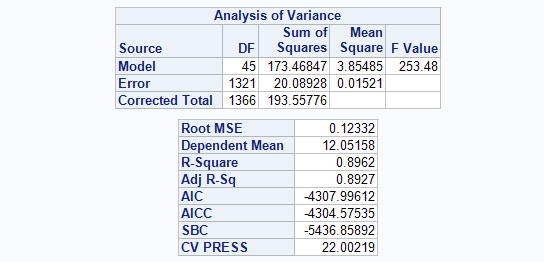
CLASS EXTERQUAL BSMTQUAL KITCHENQUAL GARAGEFINISH GARAGETYPE HEATINGQC BSMTEXPOSURE LOTSHAPE GARAGECOND CENTRALAIR FOUNDATION NEIGHBORHOOD;

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**;

\* 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 ;

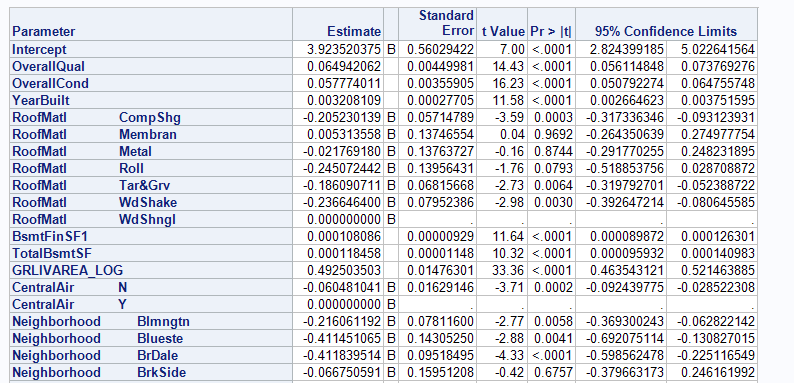
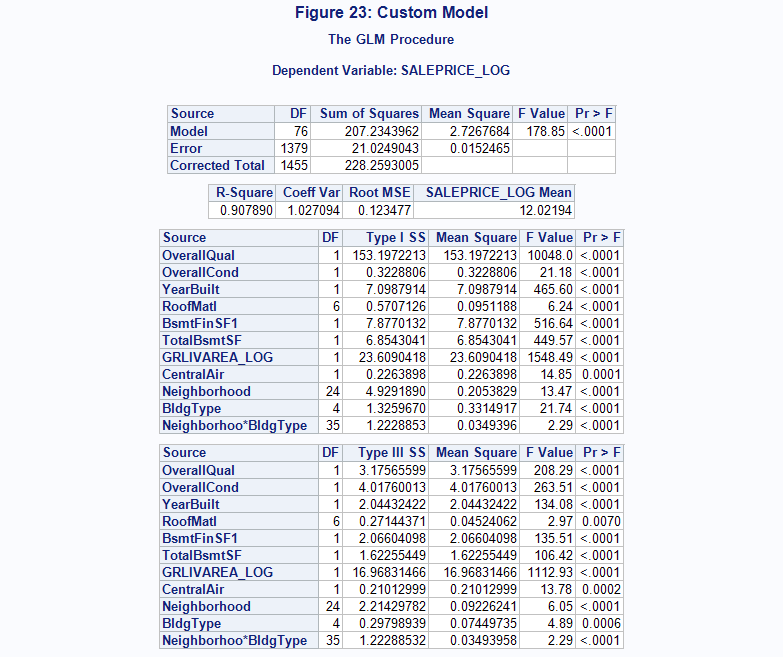
**RUN**;

**PROC** **GLMSELECT** DATA=TRAIN2 PLOTS=ALL;

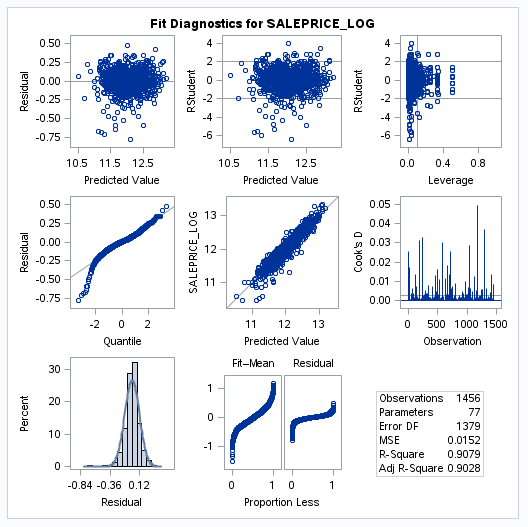
CLASS NEIGHBORHOOD BLDGTYPE ROOFMATL CENTRALAIR;

MODEL SALEPRICE\_LOG = OVERALLQUAL OVERALLCOND YEARBUILT ROOFMATL BSMTFINSF1 TOTALBSMTSF GRLIVAREA\_LOG CENTRALAIR NEIGHBORHOOD | BLDGTYPE / SELECTION=NONE;

**RUN**;



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



**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**;

1. http://jse.amstat.org/v19n3/decock.pdf [↑](#footnote-ref-1)