An Analysis of Housing Values in Ames, Iowa

Famous Four

Sachin Chavan, Edward Fry, Gabriel Gonzales, Suresh Nagarajean

Introduction

The Kaggle competition gives machine learning and data science students the opportunity to try their hand against a real world data set. In this case, the data comes from housing values in Ames, Iowa. The goal is to predict housing values based on the contents of that data. In this paper, we will explore the quest to obtain the lowest (best) Kaggle score while using only techniques that have been recently learned, such as regression.

Data Description

The data used for this exercise comes from the Ames Housing dataset, which was put together by Dean de Cock. It consists of 79 explanatory variables and 2921 observations of properties in Ames and has been split into two subsets – a training set with 1461 observations and a test set with 1460. For the first analysis in this paper, we have focused on square footage and three particular neighborhoods where the homes are located. The second analysis expands this to look at potentially any of the explanatory variables. More information on the details can be found in the analyses that follow.

Question 1 Analysis

Problem

Assume that Century 21 Ames (a real estate company) in Ames Iowa has commissioned you to answer a very important question with respect to their business. Century 21 Ames only sells houses in the *NAmes*, *Edwards* and *BrkSide* neighborhoods and would like to simply get an estimate of how the *SalePrice* of the house is related to the square footage of the living area of the house (*GrLIvArea*) and if the *SalesPrice* (and its relationship to square footage) depends on which neighborhood the house is located in. Build and fit a model that will answer this question, keeping in mind that realtors prefer to talk about living area in increments of 100 sq. ft. Provide your client with the estimate (or estimates if it varies by neighborhood) as well as confidence intervals for any estimate(s) you provide. It turns out that Century 21's leadership team has a member that has some statistical background. Therefore, make sure and provide evidence that the model assumptions are met and that any suspicious observations (outliers / influential observations) have been identified and addressed. Finally, of course, provide your client with a well written conclusion

that quantifies the relationship between living area and sale price with respect to these three neighborhoods. Remember that the company is only concerned with the three neighborhoods they sell in.

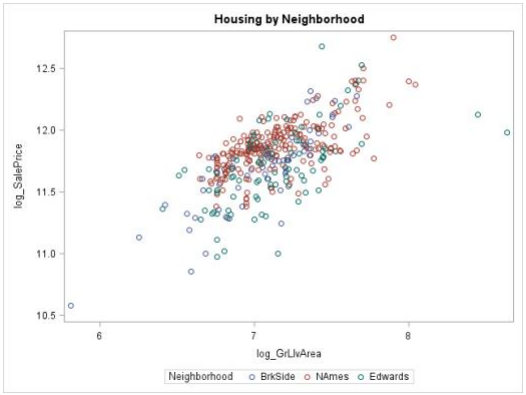
Build and Fit the Model

In order to construct the model, the data must first be cleaned. There were a number of missing values, which had to be removed in order for the valid values to be used in computation. Once this was done, assumptions had to be checked in order to make sure that we were working under conditions whereby the results would be statistically valid. Next, various models were calculated and compared. Finally, through trial and error, various parameter combinations were computed and the resulting error examined in order to find the best combination of parameters and model selection to predict the sale price based on square footage and neighborhood.

Upon first looking at the plot of the data in *Plot 1*, it's apparent that normality and equal spread assumptions are violated. The values clump together in the lower left corner.

Plot 1

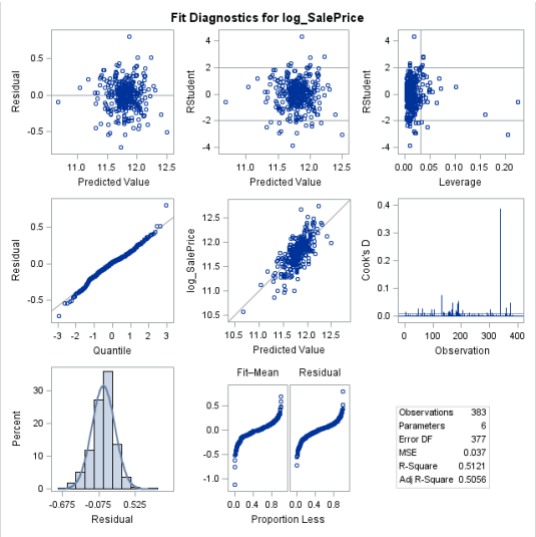
To address this, a log-log transformation was applied, resulting in a much better fit from the scatter plot in *Plot 2*.



Plot 2

Checking Assumptions

Fit diagnostics are given in the set of plots represented in *Plot 3*.



Plot 3

Residual Plots

An examination of the residual plots for the log-log transformed data shows a more randomized pattern for both the regular and studentized residuals. The residual Q-Q plot shows a nice linear trend, and the residual histogram is very nearly normally distributed.

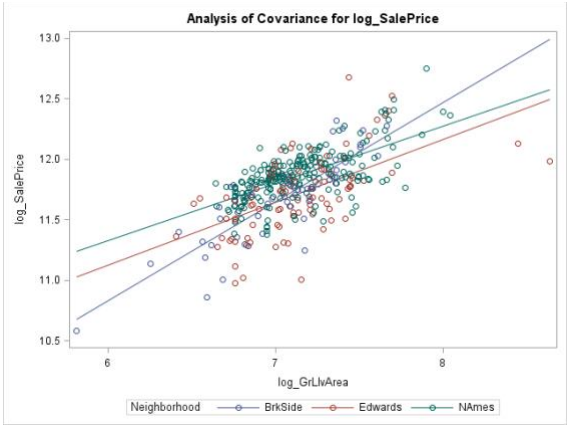
Influential point analysis

Observation 340 stood out immediately on the Cook's D plot, so it is necessary to examine this outlier and see what is going on. Upon further inspection, the data point appears to be valid, so there is no good reason to exclude it from the analysis, and we can proceed with caution.

Assumptions

After adjusting the data with a log-log transformation, the assumptions appear to be met.

- **Linearity** – the prices by living area do appear to be linear, even though each neighborhood has a slightly different regression line as shown in *Plot 4*.



Plot 4

- **Normality** – from the residual plots and histogram in *Plot 3*, the log transformed data appears to be normally distributed.
- **Independence** – each home and neighborhood is evaluated independently. While there may be a correlation between similar neighborhoods, there is no reason to assume that a given home would have a direct dependency on another.
- **Equal spread** – the spreads are visually similar from the plots.

Comparing Competing Models

Adj R^2

The adjusted R-squared is *.5056* for all three model selection methods (forward, backward, and stepwise), which means that approximately 51% of the predicted sales price can be attributed to the neighborhood and living area variables. This is a great level of correlation, but it does answer the question of how much these two variables influence the price, and it also implies that we should continue to examine the parameters to see what other variables may be having a large effect on the price.

Internal CV Press

The CV Press for forward selection was *14.93*, backward selection was *14.87*, and stepwise was *14.65*. Based on this, we chose to use stepwise selection.

Parameters

Estimates

For backward selection, therefore, the parameter estimates worked out to be, as shown in *Table 1*.

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	8.492728	0.324417	26.18
Neighborhood BrkSide	1	-2.579807	0.599881	-4.30
Neighborhood Edwards	1	-0.486220	0.517508	-0.94
Neighborhood NAmes	0	0	.	.
log_GrLiv*Neighborho BrkSide	1	0.819648	0.071629	11.44
log_GrLiv*Neighborho Edwards	1	0.519667	0.056476	9.20
log_GrLiv*Neighborho NAmes	1	0.473024	0.045429	10.41

Table 1

Interpretation

From the data computed in table x, a good regression for this scenario would be:

*μ(log(SalePrice) = 8.49 – 2.58*BrkSide - .47*Edwards + .52*log(GrLivArea)*Edwards + .82*log(GrLivArea)*BrkSide + .47*log(GrLivArea)*BrkSide*

Confidence Intervals

If we include confidence intervals, then the regression equation would be:

*μ(log(SalePrice) = [8.17, 8.81]+ [-3.18, -1.98]*BrkSide + [-.98, .04]*Edwards + [.46, .58]*log(GrLivArea)*Edwards + [.75, .89]*log(GrLivArea)*BrkSide + [.42, .52]*log(GrLivArea)*BrkSide*

Conclusion

Based on this analysis, it is clear that neighborhood and living area play a role (51%) in the eventual sales price. However, it is also clear that other variables must also play a significant role; therefore, additional analysis is recommended. Of course, these conclusions demonstrate only correlation and not causality since the homes were not randomly selected, and the conclusions only apply to the data included in this study.

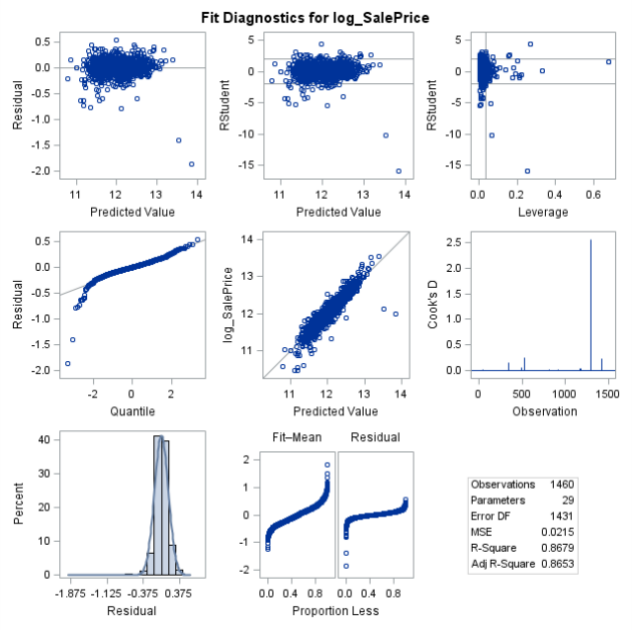
Question 2 Analysis

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 R^2, 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.

Checking Assumptions

As with problem 1, we found it necessary to apply log transformations to many of the variables. Once that was done, the fit diagnostics in *Plot 5* demonstrated that we were ready to proceed.



Plot 5

Residual Plots

An examination of the residual plots for the log-log transformed data shows a more randomized pattern for both the regular and studentized residuals about 0. The residual Q-Q plot shows a

linear trend after a bit of curve at the beginning, and the residual histogram shows a left skew but shows normal distribution otherwise. We can proceed with caution.

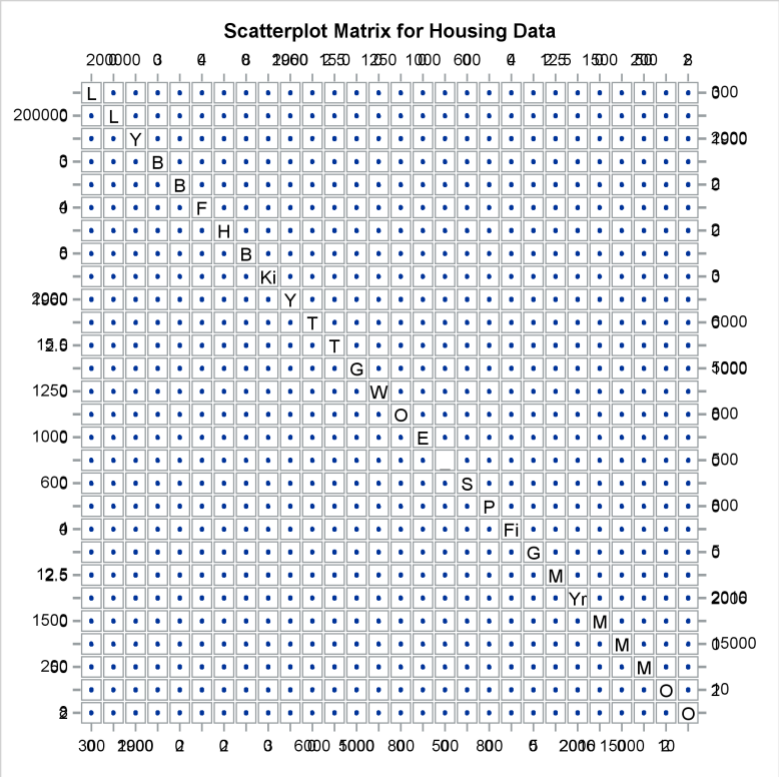
Influential point analysis

Observation 1300 stood out on the Cook's D plot in *Plot 5*, so it is necessary to examine this outlier and see what is going on. Upon further inspection, the data point appears to be valid, so there is no good reason to exclude it from the analysis, and we can proceed with caution.

Assumptions

After adjusting the data with a log-log transformation, the assumptions appear to be met.

- **Linearity** – while it may be difficult to tell due to space constraints, the various plots of the matrix in *Plot 6* appear to be linear when viewed closely.



Plot 6

- **Normality** – from the residual plots and histogram in *Plot 5*, the log transformed data appears to be normally distributed.
- **Independence** – the variables within the same property, like area, presence of a pool or basement, or number of bedrooms, are not related in any discernable way. Further, homes between neighborhoods can be assumed to be independent.
- **Equal spread** – the spreads are visually similar from the plots.

Model Selection

Type of Selection

- **Stepwise** – R^2 = *.87*, CV Press = *32.83*. Stepwise returned these variables: *MSZoning Neighborhood RoofMatl Exterior1st BsmtQual BsmtExposure BsmtFinType1 Heating HeatingQC CentralAir Functional GarageFinish GarageQual GarageCond Fence SaleCondition LotArea YearBuilt BsmtFullBath YearRemodAdd GrLivArea WoodDeckSF EnclosedPorch ScreenPorch Fireplaces GarageCars YrSold MSSubClass OverallQual OverallCond*
- **Forward** – R^2 = *.93*, CV Press = *32.26*. Forward selection resulted in the primary variables of *Neighborhood, YearBuilt, BsmtFullBath, GrLivArea, GarageCars, OverallQual, and OverallCond*
- **Backward** – R^2 = *.82*, CV Press = *27.17*. These variables all show influence with this model: *LotFrontage LotArea YearBuilt BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr YearRemodAdd TotalBsmtSF TotRmsAbvGrd GrLivArea WoodDeckSF OpenPorchSF EnclosedPorch _3SsnPorch ScreenPorch PoolArea Fireplaces GarageCars MoSold YrSold MasVnrArea MiscVal MSSubClass OverallQual OverallCond MSZoning Street Alley LotShape LandContour LandSlope Neighborhood BldgType HouseStyle RoofStyle RoofMatl Exterior1st MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive PoolQC Fence SaleType SaleCondition*
- **Custom** – R^2 = *.91*. For a custom criteria, these variables were examined: *Neighborhood BldgType Heating CentralAir SaleCondition LotArea YearBuilt BsmtFullBath YearRemodAdd TotalBsmtSF GrLivArea WoodDeckSF ScreenPorch PoolArea Fireplaces GarageCars OverallCond OverallQual*

Comparing Competing Models

Adj R^2

The adjusted R-squared varied slightly, with forward selection being the highest at *.93*, which means that approximately 93% of the predicted sales price can be attributed to the variables selected.

Internal CV Press

However, the CV Press for backward selection was the best at *27.17*. Based on this, we initially chose to use backward selection for the model.

Kaggle Score

As shown in *Figure 1*, the custom selection model resulted in the best Kaggle score - *.13305* - of all the models we tried. Since he custom model actually resulted in the best Kaggle score, we ultimately went with the custom model.

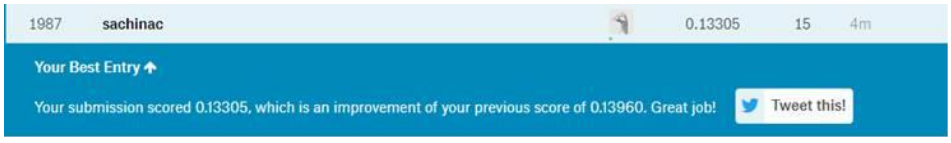


Figure 1

Conclusion

Based on our analysis, the custom model yielded the best (lowest) Kaggle score. As a result, that model is the one that we officially submitted.

Appendix A – Analysis 1 SAS Code

```
/* Load the data */
title 'House Pricing';
PROC IMPORT OUT= housing_ds
        DATAFILE= "..\Data\train.csv"
        DBMS=CSV REPLACE;
        GETNAMES=YES;
        DATAROW=2;
RUN;

/* Filter and transform the data */
data h_transform_ds;
set housing_ds;
  where Neighborhood='NAmes' or Neighborhood='Edwards' or Neighborhood='BrkSide';
log_SalePrice = log(SalePrice);
log_GrLIvArea = log(GrLIvArea);

run;

proc print data=h_transform_ds;
run;

proc corr data=housing_ds plots=all;
run;

/* Set plot symbols */
SYMBOL1 V=plus C=black I=none;
SYMBOL2 V=star C=red I=none;
SYMBOL3 V=circle C=blue I=none;
TITLE 'Housing by Neighborhood';

/* Plot the 3 neighborhoods of interest */
proc sgplot data = h_transform_ds;
scatter x = log_GrLIvArea y = log_SalePrice / group = Neighborhood;
run;

/* Check assumptions */
title 'Housing by Neighborhood';
proc glm data=h_transform_ds plots=all;
class Neighborhood;
model log_SalePrice=Neighborhood|log_GrLIvArea/solution clparm;
run;

title 'Train the Model - Forward Elimination';
proc glmselect data=h_transform_ds plots=all;
class Neighborhood;
model log_SalePrice = Neighborhood|log_GrLIvArea /selection=Forward(stop=CV) cvmethod=random(5) cvdetails=cvpress stats=all;
output out=results_b p=Predict;
run;

title 'Train the Model - Backward Elimination';
proc glmselect data=h_transform_ds plots=all;
class Neighborhood;
model log_SalePrice = Neighborhood|log_GrLIvArea /selection=Backward(stop=CV) cvmethod=random(5) cvdetails=cvpress stats=all;
output out=results_b p=Predict;
run;

title 'Train the Model - Stepwise';
proc glmselect data=h_transform_ds plots=all;
class Neighborhood;
model log_SalePrice = Neighborhood|log_GrLIvArea/selection=stepwise select=cv slentry = .15 sls=.15 cvdetails=cvpress;
output out=results_s p=Predict;
run;
quit;
```

Appendix B – Analysis 2 SAS Code

```
/*

Following fields are numeric fields but few records contains NA. This code replaces NA by 0

Fields edited prior to processing

LotFrontage GarageYrBlt MasVnrArea GarageCars BsmtFullBath BsmtHalfBath BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF

*/

data house_train_ds;
infile '..\Data\train.csv' dsd truncover;
file '..\Data\train_sas.csv' dsd;
length word $200;
do i=1 to 81;
input word @;
if(word='NA' and (i=4 or i=60 or i=27 or i=62 or i=63 or i=48 or i=49 or i=35 or i=37 or i=38 or i=39)) then word=0;
put word @;
end;
put;
run;

data house_test_ds;
infile '..\Data\test.csv' dsd truncover;
file '..\Data\test_sas.csv' dsd;
length word $200;
do i=1 to 80;
input word @;
if(word='NA' and (i=4 or i=60 or i=27 or i=62 or i=63 or i=48 or i=49 or i=35 or i=37 or i=38 or i=39)) then word=0;
put word @;
end;
put;
run;

title 'House Pricing Training Dataset';
PROC IMPORT OUT= house_train_sas_ds
        DATAFILE= "..\Data\train_sas.csv"
        DBMS=CSV REPLACE;
        GETNAMES=YES;
        DATAROW=2;
RUN;
PROC IMPORT OUT= house_test_sas_ds
        DATAFILE= "..\Data\test_sas.csv"
        DBMS=CSV REPLACE;
        GETNAMES=YES;
        DATAROW=2;
RUN;

/* Log transform and clean the data */
data house_train_log_ds;
set house_train_sas_ds;

log_SalePrice = log(SalePrice);
IF MiscVal    > 0   then log_MiscVal      = log(MiscVal);    else log_MiscVal=0;
if GarageYrBlt = 0   then GarageYrBlt  = YearBuilt;
if LotFrontage = 0   then LotFrontage  = mean(LotFrontage);
if MasVnrArea  = 0   then MasVnrArea   = mean(MasVnrArea);
run;

/* Log transform and clean the data */
data house_test_log_ds;
set house_test_sas_ds;
```

```
log_SalePrice = log(SalePrice);
IF MiscVal    > 0   then log_MiscVal      = log(MiscVal);    else log_MiscVal=0;
if GarageYrBlt = 0   then GarageYrBlt  = YearBuilt;
if LotFrontage = 0   then LotFrontage  = mean(LotFrontage);
if MasVnrArea  = 0   then MasVnrArea   = mean(MasVnrArea);
run;

/* Combine both train and test data sets */
data housing_ds;
 set house_train_log_ds house_test_log_ds;
run;


/*
  Removed following variables from the model
  as correlation matrix showed that these are highly correlated so they are redundant

  Correlated variables removed
  1stFlrSF 2ndFlrSF BsmtFinSF1 BsmtFinSF2 BsmtUnfSF GarageYrBlt GarageArea TotRmsAbvGrd

  removed because scatterplot doesn't look good.
  _3SsnPorch TotalBsmtSF

  log transformation applied on below variables.
  These are the variables that represents Areas in Square feet.
  log_3SsnPorch log_EnclosedPorch log_GrLivArea log_LotArea log_MasVnrArea log_OpenPorchSF log_TotalBsmtSF log_WoodDeckSF


  Summary
  ===================================
  Total Indepdendent Variables - 82
  Total Used for analysis - 72
  Quantitative Variables  - 26
  Qualitative Variables   - 46

*/

%macro quantitative_vars;
LotFrontage
LotArea
YearBuilt
BsmtFullBath
BsmtHalfBath
FullBath
HalfBath
BedroomAbvGr
KitchenAbvGr
YearRemodAdd
TotalBsmtSF
TotRmsAbvGrd
GrLivArea
WoodDeckSF
OpenPorchSF
EnclosedPorch
_3SsnPorch
ScreenPorch
PoolArea
Fireplaces
GarageCars
MoSold
YrSold
MasVnrArea
MiscVal
MSSubClass
OverallQual
OverallCond
```

```
%mend quantitative_vars;

/* categorical (44) */
%macro categorical_vars;
MSZoning
Street
Alley
LotShape
LandContour
LandSlope
Neighborhood
BldgType
HouseStyle
RoofStyle
RoofMatl
Exterior1st
Exterior2nd
MasVnrType
ExterQual
ExterCond
Foundation
BsmtQual
BsmtCond
BsmtExposure
BsmtFinType1
BsmtFinType2
Heating
HeatingQC
CentralAir
Electrical
KitchenQual
Functional
FireplaceQu
GarageType
GarageFinish
GarageQual
GarageCond
PavedDrive
PoolQC
Fence
SaleType
SaleCondition
%mend categorical_vars;

/* Check assumptions */
proc sgscatter data = housing_ds;
title "Scatterplot Matrix for Housing Data";
matrix %categorical_vars %quantitative_vars / group = Neighborhood;
run;

title 'Train the Model - Forward Selection';
proc glmselect data=housing_ds plots=all;
class %categorical_vars;
model log_SalePrice = %categorical_vars %quantitative_vars /selection=Forward(stop=CV) cvmethod=random(5) cvdetails=cvpress stats=all;
output out=results_f p=Predict;
run;
quit;

data final_results_f;
set results_f;
if Predict > log(10000) then SalePrice=exp(Predict);
if Predict < log(10000) then SalePrice=10000;
keep id SalePrice;
where id > 1460;
run;
```

```
proc print data=final_results_f;
run;

proc univariate data=final_results_f;
var SalePrice;
histogram SalePrice;
run;

proc univariate data=house_train_sas_ds;
var SalePrice;
histogram SalePrice;
run;

proc means data=house_train_sas_ds;
var SalePrice;
run;

PROC EXPORT DATA= WORK.FINAL_RESULTS_F
        OUTFILE= "..\Data\msds_submit_forward_003_20190813.csv"
        DBMS=CSV REPLACE;
        PUTNAMES=YES;
RUN;

title 'Train the Model - Backward Elimination';
proc glmselect data=housing_ds plots=all;
class %categorical_vars;
model log_SalePrice = %quantitative_vars %categorical_vars /selection=Backward(stop=CV) cvmethod=random(5) cvdetails=cvpress stats=all;
output out=results_b p=Predict;
run;
quit;

data final_results_b;
set results_b;
if Predict > log(10000) then SalePrice=exp(Predict);
if Predict < log(10000) then SalePrice=10000;
keep id SalePrice;
where id > 1460;
run;

proc print data=final_results_b;
run;

title 'Train the Model - Stepwise';
proc glmselect data=housing_ds plots=all;
class %categorical_vars;
model log_SalePrice = %categorical_vars %quantitative_vars/selection=stepwise select=cv slentry = .15 sls=.15 cvdetails=cvpress;
output out=results_s p=Predict;
run;
quit;

data final_results_s;
set results_s;
if Predict > log(10000) then SalePrice=exp(Predict);
if Predict < log(10000) then SalePrice=10000;
keep id SalePrice;
where id > 1460;
run;


title 'Train the Model - Custom';
proc glm data=housing_ds plots=all;
Class  Neighborhood BldgType Heating CentralAir SaleCondition;

model log_SalePrice = Neighborhood BldgType Heating CentralAir SaleCondition
    LotArea YearBuilt BsmtFullBath YearRemodAdd TotalBsmtSF GrLivArea WoodDeckSF  ScreenPorch
    PoolArea Fireplaces GarageCars OverallCond OverallQual /solution clparm;
```

```
output out=results_custom p=Predict;
run;
quit;

data final_results_custom;
set results_custom;
if Predict > log(10000) then SalePrice=exp(Predict);
if Predict < log(10000) then SalePrice=10000;
keep id SalePrice;
where id > 1460;
run;
```