Milan Patel

MSDS 6371 Ames Housing Project

February 14, 2022

MSDS 6371 Project Description (Weeks 13 and 14)

Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this Kaggle competition's dataset proves that much more influences price negotiations than the number of bedrooms or the presence of a white-picket fence.

With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this competition challenges you to predict the final price of each home.

- Introduction

This project is based off the Ames Housing dataset on Kaggle. We will first perform an analysis on how the sales price is influenced by the living area square feet in each house in 3 different neighborhoods. While performing an analysis, we will build different models to accurately predict the sales price of the homes with many features.

- Data Description

The data comes from the Ames Housing dataset on Kaggle, which tells us about the sale price of homes in Ames, Iowa. There are 80 different variables, which will all make an impact on the price of a house. We looked at all of the observations in the dataset (1460) and used a variety of these variables to build models to predict the sale price.

-Analysis Question 1:

- **Restatement of Problem**

A real estate company, Century 21 Ames, has hired us to analyze how the living area square feet (GrLivArea) is influential to predict sales price in the 3 neighborhoods (Names, BrkSide, and Edward) they sell homes in.

- **Build and Fit the Model**

Model before log transformation:

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	74676.40154	B	5954.52674	12.54	<.0001	62967.95510	86384.84798
Neighborhood BrkSide	-54704.88774	B	13042.61747	-4.19	<.0001	-80350.71900	-29059.05648
Neighborhood Edwards	-43247.84694	B	11671.23793	-3.71	0.0002	-66197.12068	-20298.57320
Neighborhood NAmes	0.00000	B	.	.	.	.	.
GrLivArea	54.31586	B	4.33457	12.53	<.0001	45.79276	62.83896
GrLivArea*Neighborho BrkSide	32.84667	B	10.16117	3.23	0.0013	12.86665	52.82669
GrLivArea*Neighborho Edwards	21.66057	B	8.79973	2.46	0.0143	4.35757	38.96358
GrLivArea*Neighborho NAmes	0.00000	B	.	.	.	.	.

R-Square	Coeff Var	Root MSE	SalePrice Mean
0.512452	19.45515	26825.22	137882.4

Model after log transformation per 100 square feet:

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	11.44334070	B	0.04202390	272.31	<.0001	11.36070868	11.52597272
Neighborhood BrkSide	-0.65174673	B	0.09204789	-7.08	<.0001	-0.83274144	-0.47075202
Neighborhood Edwards	-0.41785616	B	0.08236942	-5.07	<.0001	-0.57981999	-0.25589233
Neighborhood NAmes	0.00000000	B	.	.	.	.	.
GrLivArea100	0.03241245	B	0.00305911	10.60	<.0001	0.02639730	0.03842761
GrLivArea*Neighborho BrkSide	0.04140983	B	0.00717122	5.77	<.0001	0.02730899	0.05551067
GrLivArea*Neighborho Edwards	0.02145301	B	0.00621039	3.45	0.0006	0.00924146	0.03366455
GrLivArea*Neighborho NAmes	0.00000000	B	.	.	.	.	.

R-Square	Coeff Var	Root MSE	logSalePrice Mean
0.527146	1.604729	0.189318	11.79752

Overall Model: log(SalePrice)=11.4433-0.6517*BrkSide-0.4179*Edwards+(GrLivArea/100)*0.0324+ (GrLivArea/100)*BrkSide*0.0414+ (GrLivArea/100)*Edwards*0.0215

In Neighborhood BrkSide:

log(SalePrice)=11.4433-0.6517*1 +(GrLivArea/100)*0.0324+ (GrLivArea/100)*1 *0.0414= 10.7916+(GrLivArea/100)*0.0738

In Neighborhood Edwards:

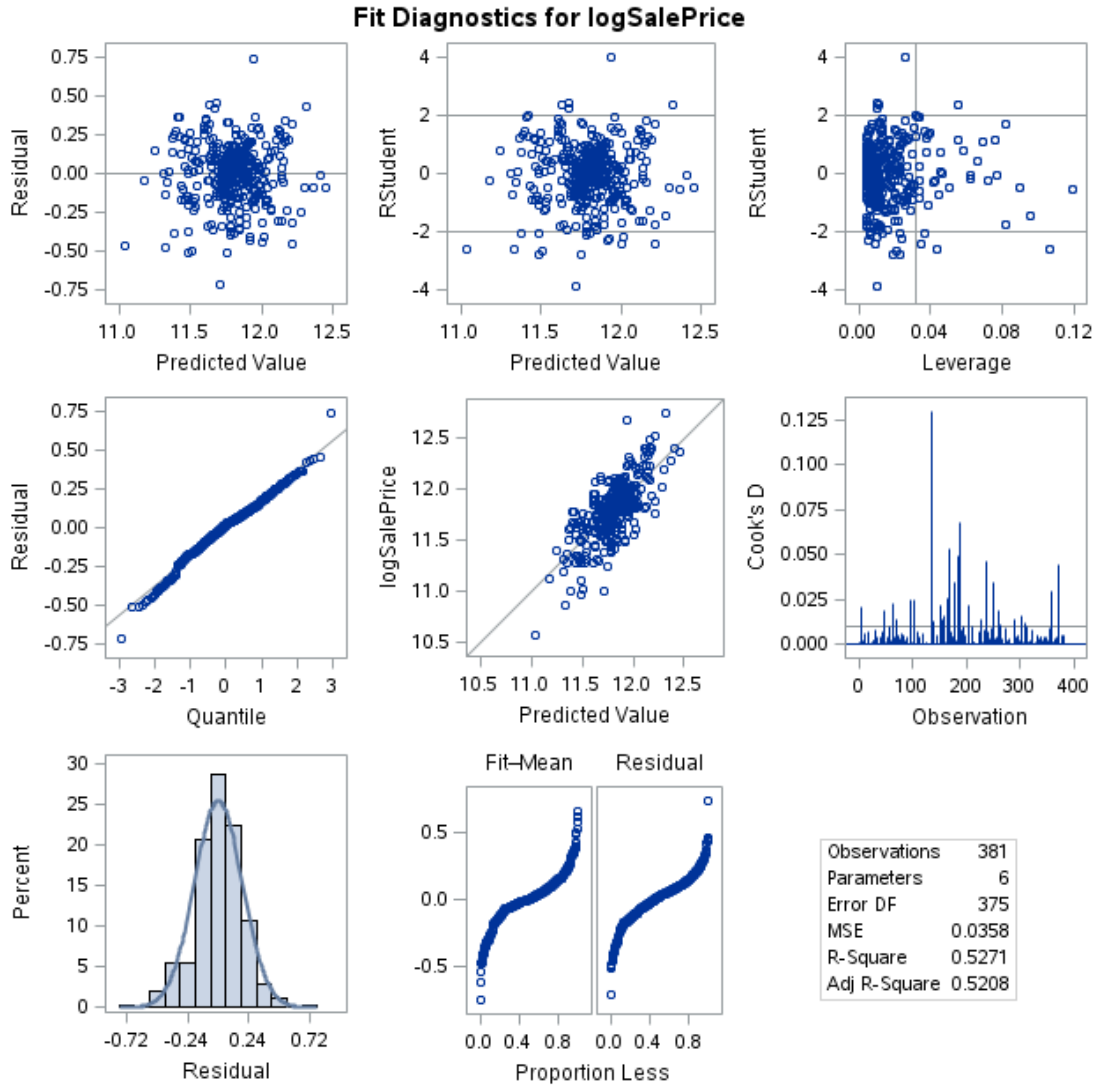
log(SalePrice)=11.4433-0.4179*1+(GrLivArea/100)*0.0324+ (GrLivArea/100)*1*0.0215 = 11.0254+(GrLivArea/100)*0.0539

In Neighborhood NAmes:

log(SalePrice)=11.4433 +(GrLivArea/100)*0.0324

- **Checking Assumptions**

After log transformation by 100 square feet:



Normality: After looking at the histogram, this looks to be normally distributed.

Linearity: After looking at the quantile chart, it looks to be linear.

Equal Standard Deviations: Based on the charts above, we will assume equal standard deviation.

Independence: We will assume our observations are independent.

- **Comparing Competing Models**

  Adj R^2

R-Square for the model before transformation is 0.512 and the R-Square for the model after the transformation is 0.527.

- **Parameters**

  Estimates

  Interpretation

  Confidence Intervals

  Starting at sale price of $e^{10.7916}$, a 100 square foot increase of above grade (ground) living area is associated with a $e^{0.0738}$ multiplicative change in median of sale price in the BrkSide neighborhood. We are 95% confident that this increase will be in ($e^{0.027}$, $e^{0.056}$).

  Starting at sale price of $e^{11.0254}$, a 100 square foot increase of above grade (ground) living area is associated with a $e^{0.0539}$ multiplicative change in median of sale price in the Edward neighborhood. We are 95% confident that this increase will be in ($e^{0.009}$, $e^{0.0337}$).

  Starting at sale price of $e^{11.4433}$, a 100 square foot increase of above grade (ground) living area is associated with a $e^{0.0324}$ multiplicative change in median of sale price in the NAmes neighborhood. We are 95% confident that this increase will be in ($e^{0.026}$, $e^{0.038}$).

- **Conclusion**

  There is sufficient evidence to suggest that the model after the transformation is a good fit for the data (p-value < 0.0001). This is an observational study. We cannot make a causal inference. We do not know if this data is randomized from a much bigger population.

- Analysis Question 2

- **Restatement of Problem**

Build a model that can accurately predict the sales price of a home in Ames, Iowa between 2006 and 2010 from all of the variables available to us. We will explore the stepwise, forward, and backward selection models and use our findings to create a better accurate model.

- **Model Selection**

Type of Selection

Stepwise

Using SAS and our selected variables, we created a model using stepwise selection. The stepwise model has selected the following predictors: Neighborhood, BldgType, HouseStyle, GarageType, with the 9 numeric variables we use in the model.

Forward

Using SAS and our selected variables, we created a model using forward selection. The stepwise model has selected the following predictors: Neighborhood, BldgType, HouseStyle, GarageType, with the 9 numeric variables we use in the model.

Backward

Using SAS and our selected variables, we created a model using backward selection. The stepwise model has selected the following predictors: Neighborhood, BldgType, HouseStyle, GarageType, with the 9 numeric variables we use in the model.

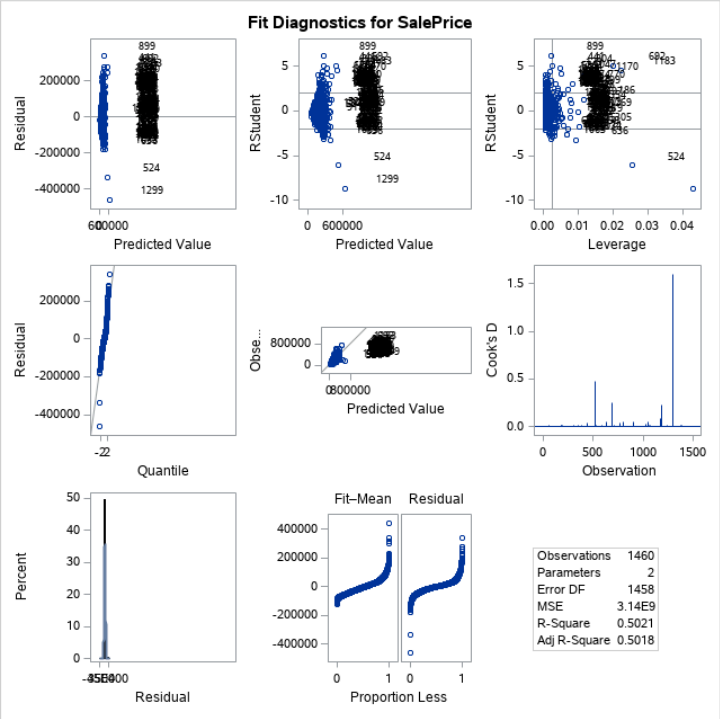
Custom

Using SAS and our selected variables, we created a model using backward selection. The stepwise model has selected the following predictors: Neighborhood, BldgType, HouseStyle, GarageType, with the 9 numeric variables we use in the model. The one numeric variable changed was GrLivArea2 which is taking the GrLivArea and squaring it.

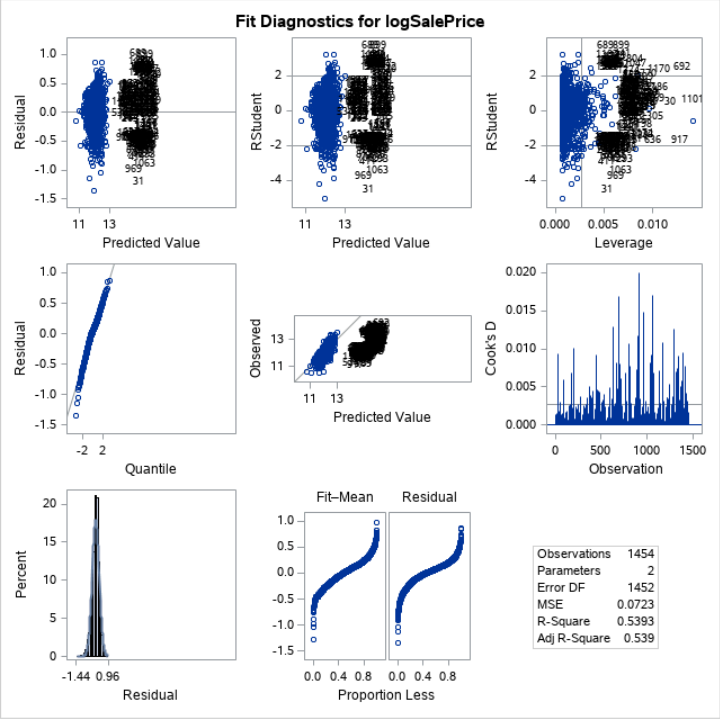
Predictive Models	Adjusted R2	CV PRESS
Forward	.8983	25.3808
Backward	.8984	24.9259
Stepwise	.8913	26.2129
CUSTOM	.8945	27.0655

- **Checking Assumptions**

Before log transformation:



After log transformation:



Normality: After looking at the histogram, it looks to be normally distributed.
Linearity: By looking at the quantile chart, it looks to be linear.
Equal Standard Deviation: Based on the charts above, we will assume equal standard deviations.
Independence: we will assume each observation is independent.

- **Comparing Competing Models**

Predictive Models	Adjusted R2	CV PRESS
Forward	.8983	25.3808
Backward	.8984	24.9259
Stepwise	.8913	26.2129
CUSTOM	.8945	27.0655

With the forward selection model, the adjusted R2 is 0.8983 and the cross-validation press is 25.3808. With the backward selection model, the adjusted R2 is 0.8984 and the cross-validation press is 24.9259. With the stepwise selection model, the adjusted R2 is 0.8913 and the cross-validation press is 26.2129. With the custom selection model, the adjusted R2 is 0.8945 and the cross-validation press is 27.0655.

- **Conclusion: A short summary of the analysis.**

After running and comparing the four models we created, the custom model gives us the best adjusted R2 (0.8945). We choose custom model selection to be our best model.

- **Appendix**

Well commented SAS Code for Analysis 1 and 2

Analysis 1 Appendix

```
/*importing analysis 1 train file that was filtered to the 3 neighborhoods*/
%web_drop_table(WORK.IMPORT);


FILENAME REFFILE '/home/u59649446/sasuser.v94/MSDS 6371 Project/analysis1.csv';

PROC IMPORT DATAFILE=REFFILE
	DBMS=CSV
	OUT=WORK.IMPORT;
	GETNAMES=YES;
RUN;

PROC CONTENTS DATA=WORK.IMPORT; RUN;


%web_open_table(WORK.IMPORT);


/*importing analysis 1 test file*/
%web_drop_table(WORK.IMPORT3);


FILENAME REFFILE '/home/u59649446/sasuser.v94/MSDS 6371
Project/analysis1test.csv';

PROC IMPORT DATAFILE=REFFILE
	DBMS=CSV
	OUT=WORK.IMPORT3;
	GETNAMES=YES;
RUN;

PROC CONTENTS DATA=WORK.IMPORT3; RUN;


%web_open_table(WORK.IMPORT3);

/* identify dataset */
data train;
set work.import(rename = (GarageYrBlt = GarageYrBlt_num));
GarageYrBlt = input(GarageYrBlt, 8.);
```

```
drop GarageYrBlt_num;
run;

proc print data=train;
run;

/*checking for outliers */
proc reg data=train plots(label) = (Cooksd);
id id;
model SalePrice = GrLivArea;
run;

/* Remove observations 524 and 1299 due to it being outliers */
data train2;
set train;
if id = 524 then delete;
if id = 1299 then delete;
run;

/*checking for more outliers */
proc reg data=train2 plots(label) = (Cooksd);
id id;
model SalePrice = GrLivArea;
run;

proc glm data=train2 plots=all;
class Neighborhood;
model SalePrice = Neighborhood | GrLivArea/solution clparm;
run;

/*Apply log transformation because of outliers*/
data logtrain;
set train2;
logGrLivArea = log(GrLivArea);
logSalePrice = log(SalePrice);
run;

proc print data=logtrain;
run;

proc glm data=logtrain plots=all;
class Neighborhood;
model logSalePrice = Neighborhood | logGrLivArea/solution clparm;
run;

/*living area is in increments of 100 sq. ft. */
data train3;
```

```
set train2;
logSalePrice = log(SalePrice);
GrLivAreaBy100 = GrLivArea/100;
run;

proc print data=train3;
run;

proc glm data=train3 plots=all;
class Neighborhood;
model logSalePrice = Neighborhood | GrLivAreaBy100 /solution clparm;
run;

/*exploratory data analysis*/
proc univariate data=logtrain;
var SalePrice logSalePrice;
histogram SalePrice logSalePrice;
run;

proc univariate data=logtrain;
var GrLivArea logGrLivArea;
histogram GrLivArea logGrLivArea;
run;

proc sgplot data = train2;
scatter x = GrLivArea y = SalePrice;
run;

proc sgplot data=train3;
scatter x = GrLivAreaBy100 y = logSalePrice;
run;

/*Adding SalePrice column to Test*/
data test;
set work.import3;
SalePrice = .;
run;

/*Combining train and test*/
data train4;
set train test;
run;

/*Creating the results*/
proc glm data = train4 plots=all;
class Neighborhood;
model SalePrice = Neighborhood GrLivArea / cli solution;
```

```
output out = results p = predict;
run;

/*Predicting the SalePrice*/
data results2;
set results;
if predict > 0 then SalePrice = predict;
if predict < 0 then SalePrice = 100000;
keep id SalePrice;
where id > 1460;
run;

proc means data=results2;
var SalePrice;
run;

proc print data=results2;
run;
```

Analysis 2 Appendix

```
%web_drop_table(projectData);


FILENAME REFFILE '/home/u59649446/sasuser.v94/MSDS6371 Analysis 2
Project/projectData.xlsx';

PROC IMPORT DATAFILE=REFFILE
	DBMS=XLSX
	OUT=projectData;
	GETNAMES=YES;
RUN;

PROC CONTENTS DATA=projectData; RUN;


%web_open_table(projectData);

/*Cleaning train dataset*/
data train;
set projectData;
   if lotFrontage = "NA" then lotFrontage = "0";
   if GarageYrBlt = "NA" then GarageYrBlt = YearBuilt;
   if MasVnrArea = "NA" then MasVnrArea = "0";
   LotFrontage1 = input(LotFrontage, 8.);
   GarageYrBlt1 = input(GarageYrBlt, 8.);
   MasVnrArea1 = input(MasVnrArea, 8.);
   drop lotFrontage GarageYrBlt MasVnrArea;
   rename lotFrontage1 = lotFrontage GarageYrBlt1 = GarageYrBlt MasVnrArea1 =
MasVnrArea;
run;

data train; set train; run;


data view(keep=lotFrontage GarageYrBlt MasVnrArea); set train; run;
proc print data = view; run;
proc contents data = view; run;

/*Checking for outliers*/
proc reg data=train plots(label) = (Cooksd);
id id;
model SalePrice = GrLivArea;
run;
```

```
/*removing the outliers */
data train2;
set train;
if id = 1299 then delete;
if id = 524 then delete;
if id = 1183 then delete;
if id = 347 then delete;
if id = 496 then delete;
if id = 1424 then delete;
run;

/* log transformations */
data train3;
set train2;
logSalePrice = log(SalePrice+1);
logLotArea = log(LotArea+1);
logLotFrontage = log(LotFrontage+1);
logMasVnrArea = log(MasVnrArea+1);
logBsmtFinSF1 = log(BsmtFinSF1+1);
logBsmtFinSF2 = log(BsmtFinSF2+1);
logBsmtFullBath = log(BsmtFullBath+1);
logBsmtHalfBath = log(BsmtHalfBath+1);
logTotalBsmtSF = log(TotalBsmtSF+1);
logBsmtUnfSF = log(BsmtUnfSF+1);
log_1stFlrSF = log(_1stFlrSF+1);
log_2ndFlrSF = log(_2ndFlrSF+1);
logGarageArea = log(GarageArea+1);
logGarageYrBlt = log(GarageYrBlt+1);
logGarageCars = log(GarageCars+1);
logWoodDeckSF = log(WoodDeckSF+1);
logOpenPorchSF = log(OpenPorchSF+1);
logEnclosedPorch = log(EnclosedPorch+1);
log_3SsnPorch = log(_3SsnPorch+1);
logScreenPorch = log(ScreenPorch+1);
logPoolArea = log(PoolArea+1);
logGrLivArea = log(GrLivArea+1);
logBedroomAbvGr = log(BedroomAbvGr+1);
logLowQualFinSF = log(LowQualFinSF+1);
logKitchenAbvGr = log(KitchenAbvGr+1);
logTotRmsAbvGrd = log(TotRmsAbvGrd+1);
logFirePlaces = log(Fireplaces+1);
logFullBath = log(FullBath+1);
logHalfBath = log(HalfBath+1);
logMiscVal = log(MiscVal+1);
logYearBuilt = log(YearBuilt+1);
logYearRemodAdd = log(YearRemodAdd+1);
```

```
logOverallQual = log(OverallQual+1);
logOverallCond = log(OverallCond+1);
logMSSubClass = log(MSSubClass+1);

/* accounting for curvature where most evident based on heuristics */
GrLivArea2 = GrLivArea**2;
GarageArea2 = GarageArea**2;
_1stFlrSF2 = _1stFlrSF**2;
TotalBsmtSF2 = TotalBsmtSF**2;
_2ndFlrSF2 = _2ndFlrSF**2;
WoodDeckSF2 = WoodDeckSF**2;
run;

proc reg data=train3 plots(label) = (Cooksd);
id id;
model logSalePrice = logGrLivArea;
run;

proc glmselect data=train3 plots=all;
Class Neighborhood BldgType HouseStyle GarageType;
model logSalePrice = logLotFrontage logLotArea logOverallQual logOverallCond
logYearBuilt logYearRemodAdd logMasVnrArea
logTotalBsmtSF logGrLivArea | Neighborhood BldgType HouseStyle
GarageType/selection= backward(select = cv choose = cv stop = cv) CVDETAILS;
run;


proc glmselect data=train3 plots=all;
Class Neighborhood BldgType HouseStyle GarageType;
model logSalePrice = logLotFrontage logLotArea logOverallQual logOverallCond
logYearBuilt logYearRemodAdd logMasVnrArea
logTotalBsmtSF logGrLivArea | Neighborhood BldgType HouseStyle
GarageType/selection= forward(select = cv choose = cv stop = cv) CVDETAILS;
run;


proc glmselect data=train3 plots=all;
Class Neighborhood BldgType HouseStyle GarageType;
model logSalePrice = logLotFrontage logLotArea logOverallQual logOverallCond
logYearBuilt logYearRemodAdd logMasVnrArea
logTotalBsmtSF logGrLivArea | Neighborhood BldgType HouseStyle
GarageType/selection= stepwise(select = cv choose = cv stop = cv) CVDETAILS;
run;


/*importing test dataset*/
%web_drop_table(test);
```

```
	FILENAME REFFILE '/home/u59649446/sasuser.v94/MSDS 6371
Project/analysis2test.xlsx';

	PROC IMPORT DATAFILE=REFFILE
		DBMS=XLSX
		OUT=test;
		GETNAMES=YES;
	RUN;

	PROC CONTENTS DATA=test; RUN;


	%web_open_table(test);

	/*Cleaning test dataset*/
	data test2;
	set test;
	   if lotFrontage = "NA" then lotFrontage = "0";
	   if GarageYrBlt = "NA" then GarageYrBlt = YearBuilt;
	   if MasVnrArea = "NA" then MasVnrArea = "0";
	   LotFrontage1 = input(LotFrontage, 8.);
	   GarageYrBlt1 = input(GarageYrBlt, 8.);
	   MasVnrArea1 = input(MasVnrArea, 8.);
	   drop lotFrontage GarageYrBlt MasVnrArea;
	   rename lotFrontage1 = lotFrontage GarageYrBlt1 = GarageYrBlt MasVnrArea1 =
MasVnrArea;
	run;

	data test2; set test2; run;


	data view(keep=lotFrontage GarageYrBlt MasVnrArea); set test2; run;
	proc print data = view; run;
	proc contents data = view; run;

	/* log transformations */
	data test3;
	set test2;
	logSalePrice = log(SalePrice+1);
	logLotArea = log(LotArea+1);
	logLotFrontage = log(LotFrontage+1);
	logMasVnrArea = log(MasVnrArea+1);
	logBsmtFinSF1 = log(BsmtFinSF1+1);
	logBsmtFinSF2 = log(BsmtFinSF2+1);
	logBsmtFullBath = log(BsmtFullBath+1);
```

```
logBsmtHalfBath = log(BsmtHalfBath+1);
logTotalBsmtSF = log(TotalBsmtSF+1);
logBsmtUnfSF = log(BsmtUnfSF+1);
log_1stFlrSF = log(_1stFlrSF+1);
log_2ndFlrSF = log(_2ndFlrSF+1);
logGarageArea = log(GarageArea+1);
logGarageYrBlt = log(GarageYrBlt+1);
logGarageCars = log(GarageCars+1);
logWoodDeckSF = log(WoodDeckSF+1);
logOpenPorchSF = log(OpenPorchSF+1);
logEnclosedPorch = log(EnclosedPorch+1);
log_3SsnPorch = log(_3SsnPorch+1);
logScreenPorch = log(ScreenPorch+1);
logPoolArea = log(PoolArea+1);
logGrLivArea = log(GrLivArea+1);
logBedroomAbvGr = log(BedroomAbvGr+1);
logLowQualFinSF = log(LowQualFinSF+1);
logKitchenAbvGr = log(KitchenAbvGr+1);
logTotRmsAbvGrd = log(TotRmsAbvGrd+1);
logFirePlaces = log(Fireplaces+1);
logFullBath = log(FullBath+1);
logHalfBath = log(HalfBath+1);
logMiscVal = log(MiscVal+1);
logYearBuilt = log(YearBuilt+1);
logYearRemodAdd = log(YearRemodAdd+1);
logOverallQual = log(OverallQual+1);
logOverallCond = log(OverallCond+1);
logMSSubClass = log(MSSubClass+1);

/* accounting for curvature where most evident based on heuristics */
GrLivArea2 = GrLivArea**2;
GarageArea2 = GarageArea**2;
_1stFlrSF2 = _1stFlrSF**2;
TotalBsmtSF2 = TotalBsmtSF**2;
_2ndFlrSF2 = _2ndFlrSF**2;
WoodDeckSF2 = WoodDeckSF**2;
run;

/* Convert alphanumeric variables to numeric and Add SalePrice with blank values */
data test4;
set test3;
BsmtFinSF1One = input(BsmtFinSF1, 8.);
BsmtFinSF1Two = input(BsmtFinSF2, 8.);
BsmtUnfSF1 = input(BsmtUnfSF, 8.);
TotalBsmtSF1 = input(TotalBsmtSF, 8.);
BsmtFullBath1 = input(BsmtFullBath, 8.);
BsmtHalfBath1 = input(BsmtHalfBath, 8.);
```

```
        GarageCars1 = input(GarageCars, 8.);
        GarageArea1 = input(GarageArea, 8.);
        drop BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF BsmtFullBath BsmtHalfBath
GarageCars GarageArea;
        rename BsmtFinSF1One = BsmtFinSF1 BsmtFinSF1Two = BsmtFinSF2 BsmtUnfSF1 =
BsmtUnfSF TotalBsmtSF1 = TotalBsmtSF BsmtFullBath1 = BsmtFullBath BsmtHalfBath1 =
BsmtHalfBath
        GarageCars1 = GarageCars GarageArea1 = GarageArea;
        SalePrice=.;
        run;

        /* Combine train and Test */
        data combine;
        set train3 test4;
        run;

        /*Backward cross validation*/
        proc glmselect data=combine plots=all;
        Class Neighborhood BldgType HouseStyle GarageType;
        model logSalePrice = logLotFrontage logLotArea logOverallQual logOverallCond
logYearBuilt logYearRemodAdd logMasVnrArea
        logTotalBsmtSF logGrLivArea | Neighborhood BldgType HouseStyle
GarageType/selection= backward(select = cv choose = cv stop = cv) CVDETAILS;
        output out = resultsbackward p = predict;
        run;

        /*Forward cross validation*/
        proc glmselect data=combine plots=all;
        Class Neighborhood BldgType HouseStyle GarageType;
        model logSalePrice = logLotFrontage logLotArea logOverallQual logOverallCond
logYearBuilt logYearRemodAdd logMasVnrArea
        logTotalBsmtSF logGrLivArea | Neighborhood BldgType HouseStyle
GarageType/selection= forward(select = cv choose = cv stop = cv) CVDETAILS;
        output out = resultsforward p = predict;
        run;

        /*Stepwise cross validation*/
        proc glmselect data=combine plots=all;
        Class Neighborhood BldgType HouseStyle GarageType;
        model logSalePrice = logLotFrontage logLotArea logOverallQual logOverallCond
logYearBuilt logYearRemodAdd logMasVnrArea
        logTotalBsmtSF logGrLivArea | Neighborhood BldgType HouseStyle
GarageType/selection= stepwise(select = cv choose = cv stop = cv) CVDETAILS;
        output out = resultsstepwise p = predict;
        run;

        /*Custom cross validation with GrLivArea2*/
```

```
proc glmselect data=combine plots=all;
Class Neighborhood BldgType HouseStyle GarageType;
model logSalePrice = logLotFrontage logLotArea logOverallQual logOverallCond
logYearBuilt logYearRemodAdd logMasVnrArea
logTotalBsmtSF GrLivArea2 | Neighborhood BldgType HouseStyle
GarageType/selection= backward(select = cv choose = cv stop = cv) CVDETAILS;
output out = resultscustom p = predict;
run;

proc sgscatter data = train3;
matrix logSalePrice GrLivArea /Group= LotConfig;
run;

proc means data= combine;
var SalePrice;
run;

proc univariate data=resultscustom;
var logSalePrice Predict;
run;


/*Predicting the SalePrice for each CV type*/
data results;
        set resultsbackward;
        Predict2 = exp(Predict);
        if Predict2 > 0 then SalePrice = Predict2;
        if Predict2 < 0 then SalePrice = 180600.98;
        keep id SalePrice;
        where id > 1460;
run;

data results2;
        set resultsforward;
        Predict2 = exp(Predict);
        if Predict2 > 0 then SalePrice = Predict2;
        if Predict2 < 0 then SalePrice = 180600.98;
        keep id SalePrice;
        where id > 1460;
run;

data results3;
        set resultsstepwise;
        Predict2 = exp(Predict);
        if Predict2 > 0 then SalePrice = Predict2;
        if Predict2 < 0 then SalePrice = 180600.98;
        keep id SalePrice;
```

```
		where id > 1460;
	run;

	data results4;
		set resultscustom;
		Predict2 = exp(Predict);
		if Predict2 > 0 then SalePrice = Predict2;
		if Predict2 < 0 then SalePrice = 180600.98;
		keep id SalePrice;
		where id > 1460;
	run;


	/*Exporting the final results*/
	PROC EXPORT
	DATA=results
	DBMS= xlsx
	LABEL
	OUTFILE='/home/u59649446/sasuser.v94/MSDS6371 Analysis 2
Project/finalbackward.xlsx'
	REPLACE;
	run;

	PROC EXPORT
	DATA=results2
	DBMS= xlsx
	LABEL
	OUTFILE='/home/u59649446/sasuser.v94/MSDS6371 Analysis 2
Project/finalforward.xlsx'
	REPLACE;
	run;

	PROC EXPORT
	DATA=results3
	DBMS= xlsx
	LABEL
	OUTFILE='/home/u59649446/sasuser.v94/MSDS6371 Analysis 2
Project/finalstepwise.xlsx'
	REPLACE;
	run;

	PROC EXPORT
	DATA=results4
	DBMS= xlsx
	LABEL
	OUTFILE='/home/u59649446/sasuser.v94/MSDS6371 Analysis 2
Project/finalcustom.xlsx'
```

```
REPLACE;
run;
```