Ames Housing Data Analysis

Bin Y, Shravan R.

Introduction

In this project, we are going to conduct 2 analyses according the sample data we got for Ames, Iowa. In the first analysis, we will find a model to demonstrate the parameters which will influence the housing price in Ames, Iowa. It provided some ideas on which factors are considerable when the customer wants to buy a house in this area. We conducted another analysis to generate the most predictive model. Then demonstrate the processes of the variable selection using various of the variable selection methods like LASSO, stepwise etc. We selected top 3 models to do the comparison and provided the Adjust R-square, AIC, ASE(Test) and Kaggle Score for these models. In the end, we picked two of the categorical variables to perform a 2-way ANOVA to test the additive and nonadditive model.

Data Description

The datasets are from Ames Housing Dataset which has housing information for the location in Iowa Ames. There are 2 datasets in this analysis. One is the training set which has 1460 observations of house information with 79 explanatory variables which including 36 continuous variables and 43 categorical variables. These records have the real sales price we can used them as training set. Another dataset is the test dataset which has 1459 observations without sale price.

Exploratory Analysis and Data Cleaning

Data cleaning is very important step in the data analysis. So we have done below data cleaning before our analysis question #1 and #2. Please see the SAS codes in Appendix I. The cleaned data are ready to run on both analyses.

1. There are 2 columns in train1.csv has “-1” in it. Replaced to the correct column name: KitchenAbvGr and Functional. In File test.csv, replace NA to -1 for all of the continuous variables.
2. Test.csv dataset has extra 3 columns than tain1.csv: Alley, Fence,PoolQC, MiscFeature. They are dropped.
3. For both files, renamed the header variaalbe name from _1rstFlr to **FirstFlrSF**, _2ndFlr to **SecondFlrSF**, _3SsnPorch to **ThreeSsnPorch**,
4. In both test.csv and train1.csv, there are some -1 values in the data. We treated it as the missing value, replaced them as the median of the variable.
5. In both test.csv and train1.csv, there are some NA values in categorical variables. We replaced them as the mod of the variable.
6. Some continuous variables have 0 value and we need to do transformation on some of the value, so we add 0.1 on the record which has 0 value. In this way it will minimum affect the prediction of the data.
7. For these 79 variables, we ran a summary in R to see the distribution of each variable (please see the result in Appendix II). Based on the data type and the distribution, we grouped the variables as below:

a. **Continuous Variables**: numeric value that has infinite possible value. SalePrice will be the response variable.
b. **Numeric Categorical Variables**: Categorical variables but has numeric value.
c. **Character Categorical Variables:** Categorical variables but has characterize value.
d. **Variables not Used:** these variables have not much meaningful data or totally have no value in test data set. So we excluded them in our analyses.

Continuous Variables (19)	Numeric Categorical Variables(17)	Character Categorical Variables(38)	Variables not Used (5)
SalePrice LotFrontage LotArea MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF _1stFlrSF _2ndFlrSF LowQualFinSF GrLivArea GarageArea WoodDeckSF OpenPorchSF EnclosedPorch _3SsnPorch ScreenPorch PoolArea	OverallQual OverallCond YearBuilt YearRemodAdd Fireplaces BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd Fireplaces GarageYrBlt GarageCars MoSold YrSold	MSSubClass MSZoning LotShape LandContour LandSlope Neighborhood Condition1 Condition2 BldgType HouseStyle RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType ExterQual ExterCond Foundation BsmtQual BsmtExposure BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual GarageCond PavedDrive MiscVal SaleType SaleCondition Street Utilities	Alley PoolQC Fence MiscFeature FireplaceQu

Table 1. Explanatory Variable Analysis

Analysis Question 1:

State the Problem:

We are developing a model that can be used for the typical homebuyers, contractors, and realtors. The goal is for the model to be easily interpreted to provide insight into some of the important factors involved when determining home prices in Ames, Iowa. When considering parameters, ease of measurement and interpretability need to be considered.

Model Selection

Type of Selection: Our goal for the model is ease of interpretation. To accomplish this, we evaluated each variable from the perspective of a homebuyer, realtor, or contractor. We made sure the variable was not too subjective, could be assessed quickly, was a commonly available measurement for the majority of homes on the market, and a common feature assessed by most homebuyers through popular listing sites (Realtor.com, Zillow.com, Refin.com). Once we evaluated each variable we used the Least Absolute Shrinkage and Selection Operator (LASSO) selection process to narrow the variables even further to the most significant. This allows the model to be simple yet valuable to our audience. Please see the SAS code in Appendix III.

After evaluating each variable for its interpretability, we narrowed our initial set of variables to 14 variables. After using the LASSO technique, the final model had 8 explanatory variables. Below is a table of the variables selected for the final model.

Categorical Variable	Continuous Variable
Neighborhood	LogGrLivArea
	YearBuilt
	LogFirstFlrSF
	LogLotArea
	LogTotalBsmtSF
	OpenPorchSF
	LogTotalBath

Fig.1.1 Explanatory Variables Selected

The final model has an $R^2 = 0.840169$.

Checking Assumptions: Before accepting the model, we must verify regression model assumptions are met. The plots below represent the Histogram, Q-Q Plot, and Studentized Residuals vs Predicted Values for residual analysis. Their interpretation will be discussed when we verify the model assumptions.

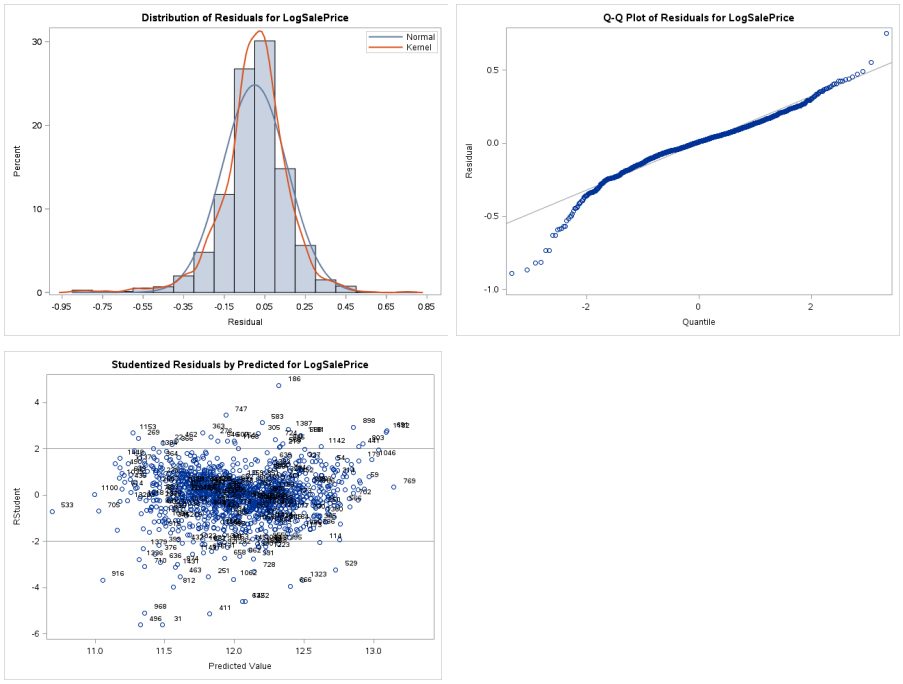


Fig.1.2 Histogram, Q-Q Plot and Studentized Residual Plot

To analyze influential points, we utilized the plots below, looks like observation 496, 524 and 1299 are high influential point. We will remove them from the data set and move on the test. In this model After determining the regression model variables for problem 1, we observed additional highly leveraged observations, but our analysis of those observations did not conclude a need to remove them from the data set.

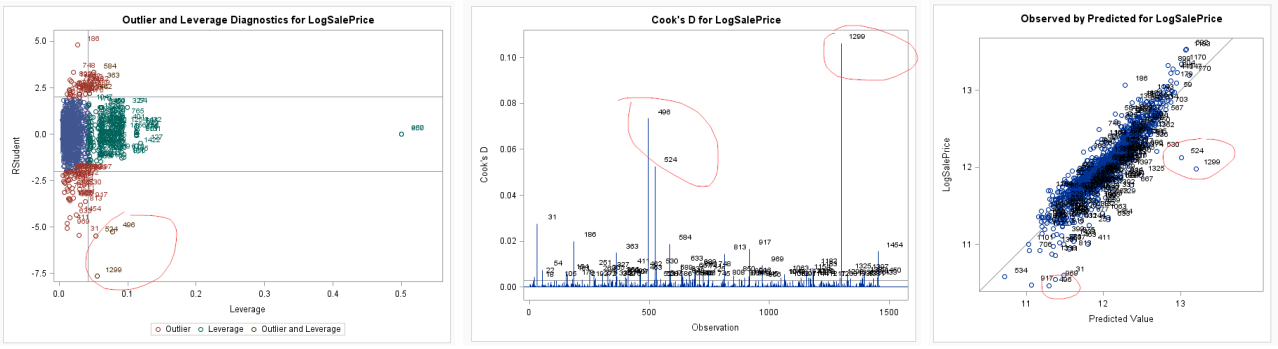


Fig.1.3 Leverage and Cook's D Plots

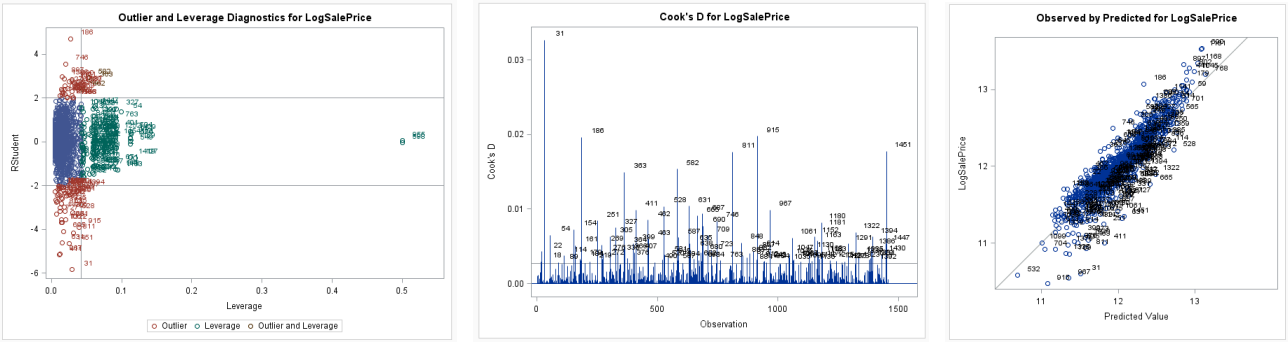


Fig.1.3 Leverage and Cook's D Plots After Removing Influential Points

Looking at the histogram and Q-Q plot we see skewness toward the lower values, but our sample size will allow us to assume normality. The studentized vs predicted residual plot shows little to no evidence against constant variance. Our plot analysis verifies that our model has met the linearity, normality, and constant variance assumption for regression models.

Final model and parameter interpretations

LogSalePrice = LogGrLivArea LogOverall YearBuilt LogFirstFlrSF LogGarageCars LogLotArea LogTotalBsmtSF OpenPorchSF LogTotalBath Neighborhood

Model Test Results: Please see the SAS code in AppendixIII.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	30	193.5256486	6.4508550	249.86	<.0001
Error	1426	36.8157237	0.0258175		
Corrected Total	1456	230.3413723			

R-Square	Coeff Var	Root MSE	LogSalePrice Mean
0.840169	1.336192	0.160678	12.02508

Source	DF	Type I SS	Mean Square	F Value	Pr > F
LogGrLivArea	1	125.0509038	125.0509038	4843.65	<.0001
YearBuilt	1	42.2348244	42.2348244	1635.90	<.0001
LogFirstFlrSF	1	8.0036741	8.0036741	310.01	<.0001
LogLotArea	1	2.9925614	2.9925614	115.91	<.0001
logTotalBsmtSF	1	3.8243219	3.8243219	148.13	<.0001
OpenPorchSF	1	0.7563372	0.7563372	29.30	<.0001
Neighborhood	24	10.6630257	0.4442927	17.21	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
LogGrLivArea	1	18.59103337	18.59103337	720.09	<.0001
YearBuilt	1	5.27670808	5.27670808	204.39	<.0001
LogFirstFlrSF	1	1.52109216	1.52109216	58.92	<.0001
LogLotArea	1	2.55432231	2.55432231	98.94	<.0001
logTotalBsmtSF	1	2.61874777	2.61874777	101.43	<.0001
OpenPorchSF	1	0.68513122	0.68513122	26.54	<.0001
Neighborhood	24	10.66302569	0.44429274	17.21	<.0001

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Lir	
Intercept	-2.456596869	B	0.64145212	-3.83	0.0001	-3.714887922	-1.1983
LogGrLivArea	0.485598535		0.01809600	26.83	<.0001	0.450100903	0.5210
YearBuilt	0.004440286		0.00031059	14.30	<.0001	0.003831025	0.0050
LogFirstFlrSF	0.148855685		0.01939298	7.68	<.0001	0.110813852	0.1868
LogLotArea	0.123592737		0.01242545	9.95	<.0001	0.099218621	0.1479
logTotalBsmtSF	0.023950817		0.00237810	10.07	<.0001	0.019285864	0.0286
OpenPorchSF	0.000372582		0.00007233	5.15	<.0001	0.000230706	0.0005
Neighborhood Blmngtn	-0.065710006	B	0.06450739	-1.02	0.3085	-0.192249570	0.0608
Neighborhood Blueste	-0.046571191	B	0.12627836	-0.37	0.7123	-0.294282476	0.2011
Neighborhood BrDale	-0.191372130	B	0.06768950	-2.83	0.0048	-0.324153821	-0.0585
Neighborhood BrkSide	-0.117351756	B	0.05594556	-2.10	0.0361	-0.227096193	-0.0076
Neighborhood ClearCr	-0.130727280	B	0.05759743	-2.27	0.0234	-0.243712061	-0.0177
Neighborhood CollgCr	-0.136563969	B	0.05062676	-2.70	0.0071	-0.235874890	-0.0372
Neighborhood Crawfor	0.050370451	B	0.05513170	0.91	0.3611	-0.057777495	0.1585
Neighborhood Edwards	-0.259212960	B	0.05225443	-4.96	<.0001	-0.361716760	-0.1567
Neighborhood Gilbert	-0.183352125	B	0.05271697	-3.48	0.0005	-0.286763259	-0.0799
Neighborhood IDOTRR	-0.304461031	B	0.05853602	-5.20	<.0001	-0.419286984	-0.1896
Neighborhood MeadowV	-0.260517850	B	0.06580488	-3.96	<.0001	-0.389602618	-0.1314
Neighborhood Mitchel	-0.233052166	B	0.05375864	-4.34	<.0001	-0.338506664	-0.1275
Neighborhood NAmes	-0.191156609	B	0.05032114	-3.80	0.0002	-0.289868019	-0.0924
Neighborhood NPkVill	-0.095902191	B	0.07426040	-1.29	0.1968	-0.241573550	0.0497
Neighborhood NWAmes	-0.204044795	B	0.05214145	-3.91	<.0001	-0.306326967	-0.1017
Neighborhood NoRidge	0.023496860	B	0.05533372	0.42	0.6712	-0.085047373	0.1320
Neighborhood NridgHt	0.076581473	B	0.05248699	1.46	0.1448	-0.026378524	0.1795
Neighborhood OldTown	-0.186623426	B	0.05472674	-3.41	0.0007	-0.293976980	-0.0792
Neighborhood SWISU	-0.173052233	B	0.06153089	-2.81	0.0050	-0.293753017	-0.0523
Neighborhood Sawyer	-0.227336608	B	0.05248104	-4.33	<.0001	-0.330284938	-0.1243
Neighborhood SawyerW	-0.191353343	B	0.05304294	-3.61	0.0003	-0.295403906	-0.0873
Neighborhood Somerst	-0.056338873	B	0.05269110	-1.07	0.2851	-0.159699270	0.0470

Fig.1.4 Linear Regression Test Results

Parameter Interpretation

1) LogGrLivArea – Paremeter = $(e^{ln(x)})^{0.485598535} = x^{0.485598535}$. An additional 10% increase in square foot of living space above grade (ground) would increase the sale price estimate by $1.1^{0.485598535} = 1.04737$
   - 95% confidence interval $(x^{0.4501}, x^{0.5211})$
2) YearBuilt - SalePrice estimate will change by $(e^{0.004440286})^{x} = 1.00445^{x}$. For every year newer a house is, the SalePrice estimate will increase by $(e^{0.004440286})^{1} = 1.00445$ or, on average, increase by .445%.
   - 95% confidence interval $(1.003838^{x}, 1.00505^{x})$
3) LogFirstFlrSF– Paremeter = $(e^{ln(x)})^{0.148855685} = x^{0.14886}$. An additional 10% increase in square foot of living space on the first floor would increase the sale price estimate by $1.1^{0.14886} = 1.0143$
   - 95% confidence interval $(x^{0.11081}, x^{0.18689})$

4) LogLotArea - Paremeter = $(e^{ln(x)})^{0.123592737} = x^{0.12359}$. An additional 10% increase in square foot of lot area would increase the sale price estimate by $1.1^{0.12359} = 1.0118$
   - 95% confidence interval $(x^{0.09921}, x^{0.14796})$
5) LogTotalBsmtSF - Paremeter = $(e^{ln(x)})^{0.023950817} = x^{0.023950}$. An additional 10% increase in square foot of basement would increase the sale price estimate by $1.1^{0.023950} = 1.00229$
   - 95% confidence interval $(x^{0.019285}, x^{0.02867})$
6) OpenPorchSF - SalePrice estimate will change by $(e^{0.000273582})^{x} = 1.000274^{x}$. For every additional square foot of open porch, the SalePrice estimate will increase by $(e^{0.000273582})^{1} = 1.000274$ or increase Sale Price by 0.027%.
   - 95% confidence interval $(1.0002307^{x}, 1.0005^{x})$
7) Neighborhoods represent various area within Ames Iowa that the house can be located. The parameter for each neighborhood would be used in the $(e^{parameter})$ equation to calculate how it would change the sale price

Conclusion

The model represents a simple yet informative model for prospective homebuyers, realtors, and contractors can use to estimate a home price or estimate the change in price based on additional features relative another house. We've tested the assumptions and conclude that this model can be used to predict sale prices in Ames, Iowa. This model explains about 84% (R^2 = 0.840169) of the variation in home prices in Ames Iowa. Since this is an observational study, the results can only be applied to houses in Ames, Iowa and do not represent a causal relationship between the variables and home sale prices.

Analysis Question 2

State the Problem:

Build the model to efficiently predict the housing price in Ames, Iowa which listed in test.csv dataset (including 1459 houses). In this analysis, we will select reasonable amount of the explanatory variables into the model and use LASSO, Model Averaging, plus Manual method to cross validate the model and provide the reports.

Variable Encoding and Combining

Some of the continuous variables may be more meaningful if we put them together. For example, some of the categorical variables we encoded them then combine. For example, BsmtQual and BsmtCond are related with the basement, we encoded each variable to integer then add them together as a numeric variable TotalBsmtQual. Please see the detailed SAS code in Appendix I:

a. Combined OverallQual and OverallCond as **TotalOverall**.
b. Combined BsmtFullBath and BsmtHalfBath as **TotalBsmtBath**

c. Combined GrLivArea and TotalBsmtSF as TotalSF
d. Combined FullBath, HalfBath, BsmtFullBath, BsmtHalfBath as **TotalBat.** For Half Bath variables we muliplied 0.5 as well.
e. Combined BsmtFinSF1 and BsmtFinSF2 as **TotalBsmtFinSF**
f. Combined YrSold and MoSold as **SoldYearMo**
g. Encoded BsmtQual and BsmtCond and Combined them as **TotalBsmtQual**
h. Encoded ExterQual and ExterCond and Combined them as **TotalExteriorQual**
i. Encoded Exterior1st and Exterior2nd and Combined them as **TotalExterior**
j. Encoded BsmtFinType1 and BsmtFinType2 and Combined them as **TotalBsmtFinType**
k. Encoded GarageQual, GarageFinish, and GarageCond and Combined them as **TotalGarageQual**
l. Encoded LandSlopeand LotShapeand Combined them as **TotalLandQual**
m. Encoded variable BsmtExposure, Functional,KitchenQual
n. Changed variable MSSubClass and MoSold as character variable as it will make more sense.

Data transformation:

When we first look at the histogram of the SalePrice, it is right skewed. We may need some transform on this field. You can see it is normal after we applied the natual log transformation (Fig.2.1).

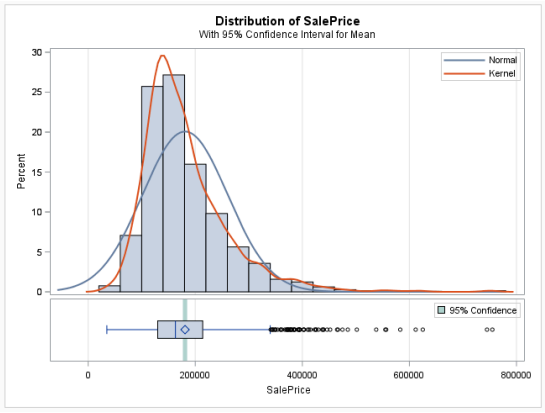
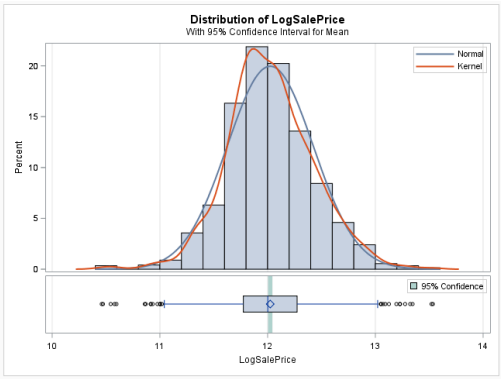
Before Transformation	After Transformation
	

Fig. 2.1. SalePrice Variable Transformation Comparison

We plotted the other continuous variables into a scatter plot matrix, some variables potentially need some data transformations on some variables (Fig.2.2). Based on the results, we selected below variables to do the natural log transformation. The natural log transaction will improve the linear regression and reducing the right skewness in the data. Please see Appendix II for the variable list of the log transformation.

Before Transformation	After Log Transformation

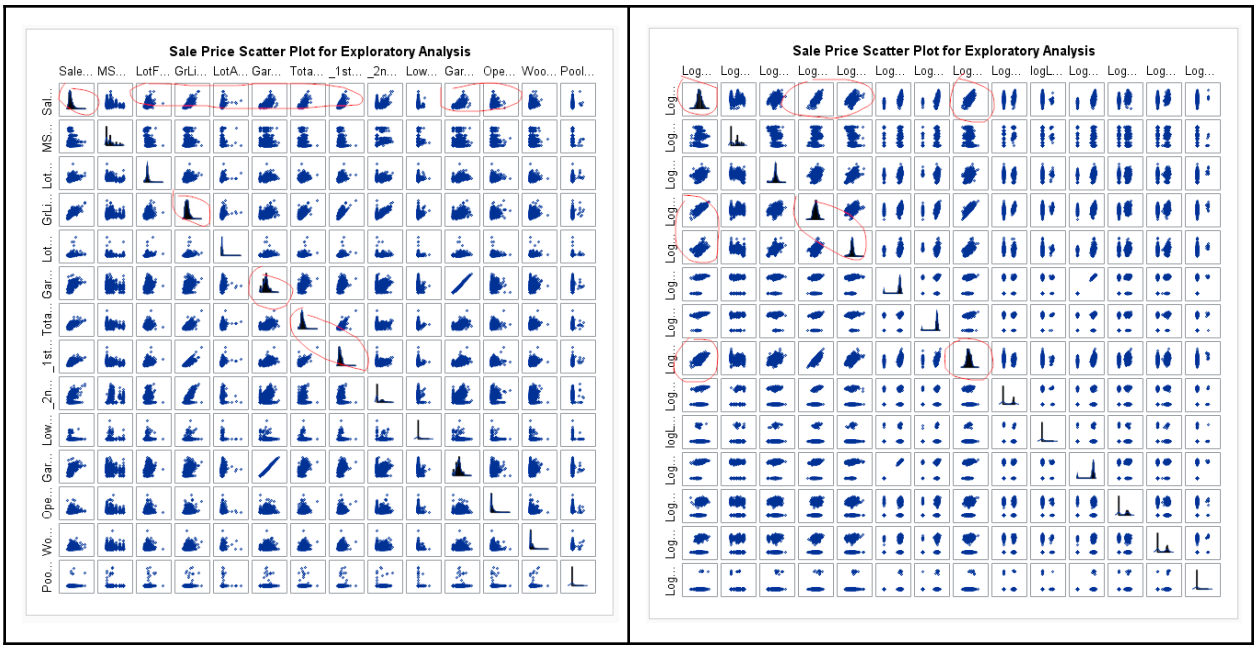


Fig.2.2. Scatter Plot of Continuous Variables Log Transformation Comparison.

Candidate Models

First, we select a list of possible factors which will affect the housing prices. and run the regression against them to try out which combination of the variables are the best. Then used the LASSO and Stepwise cross validation method to validate each model we selected. Please see below for the model we found. We will explain later on how we select the models.

LASSO Selected Model#1

logSalePrice = RoofMatl MasVnrType Neighborhood BldgType ExterQual_gr BsmtQual_gr LogGrLivArea LogLotArea LogTotalOverall LogTotalBath logTotalBsmtBath GarageArea FirstFlrSF YearRemodAdd YearBuilt logTotalGarageQual logTotalLandQual LogSaleAge

Stepwise Selected Model#2

logSalePrice = Neighborhood BldgType BsmtQual_gr LogGrLivArea LogLotArea LogTotalOverall logTotalBsmtBath GarageArea FirstFlrSF SecondFlrSF YearRemodAdd YearBuilt LogSaleAge

Best Model#3

logSalePrice = RoofMatl MasVnrType Neighborhood BldgType Exterior1st_gr Exterior2nd_gr ExterQual_gr BsmtQual_gr LogGrLivArea LogLotArea LogTotalBsmtSF LogWoodDeckSF LogTotalOverall LogTotalBath LogTotalBsmtBath GarageArea FirstFlrSF SecondFlrSF YearRemodAdd YearBuilt logTotalExteriorQual logTotalGarageQual logTotalBsmtQual logTotalBsmtFinType logTotalLandQual logSaleAge

Model Selection:

Type of selections: we use LASSO and stepwise to determine the variables that can be eliminated. Please see Appendix V for the parameter and results we did using LASSO and stepwise.

Check Assumptions:

Residual Plots: please see below residual plots for the model we selected. Please see Fig. 2.1. from the residual plots, you can see the residuals clustered and evenly distributed around the horizontal 0.0 line. Only have several outliers we will analysis it later.

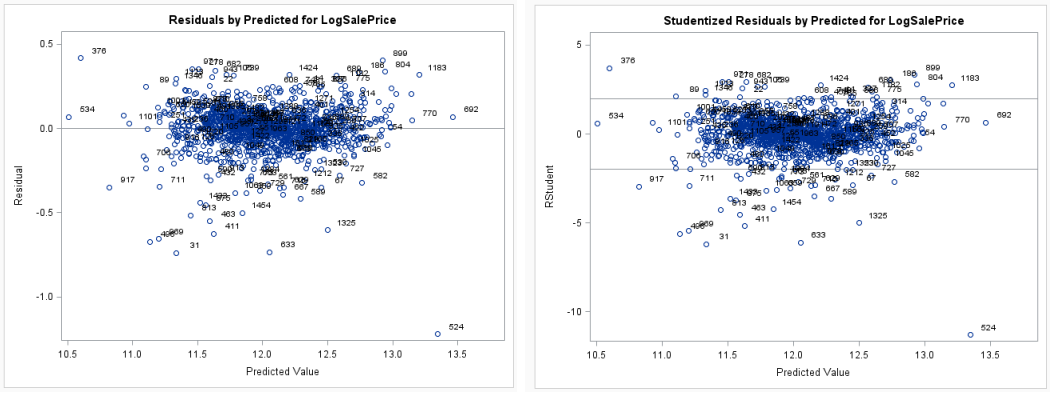


Fig.2.3 Residual Plots for the Model

Influential point analysis (Cook's D and Leverage): Looking as the matrix scatter plot (Fig. 2.4), we do see some influential point with high leverage – low residual on observation 524. We have cleaned up the data point in our model. There are still some highly leveraged data points (id 89) but most of them are between 0 to 0.25 in Cook'D plot. We can ignore them and move on. Please see the Cook'D plots for before and after removing the influence point 524 in Appendix VI.

Linearity: looking at the scatter plot and, it looks like it has the linear relationship between the variables and the sale price. But there are some outliers we may want to have further check.

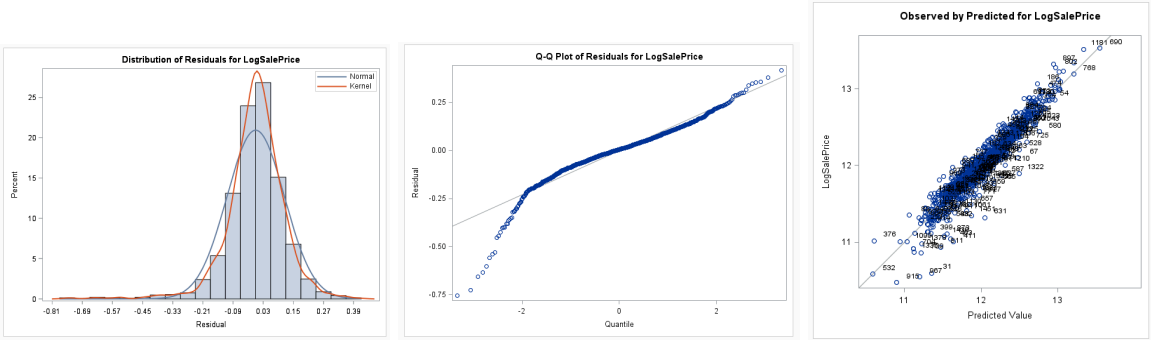


Fig.2.5 glm Result For the Model

Normality: looking at the histogram and QQ plot, it looks like a little bit left skewed but there is not enough evidence to against the normality. We will move on.

Equal SD: looking at the residual vs. predicted value. There is not enough evidence against unequal standard deviations.

Independence: it didn't say how the researcher chose housing data. We assume the all of variables we selected in the model are independence.

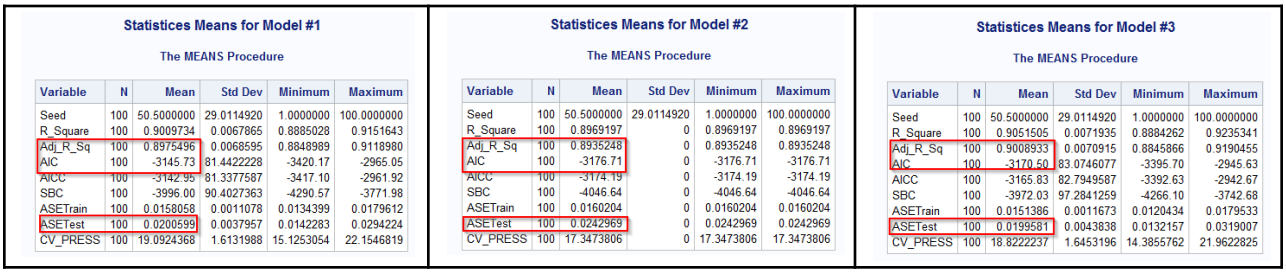
Extra Caution: There is no extra caution after we removed observation 524 from the model.

Based on the assumption checking, the data robust with the multiple linear regression. We are going to move on with the multiple linear regression test.

Comparing Competing Models

How do we know which model is better model? Glmselect will give you Adj_R-Seq, AIC and ASE(Test), but it will depend on the seed as well. So, we use a SAS macro to run

each model 100 times with different seed. Then get the average of Adj_R-sq, AIC, ASE (Test). Please see SAS in Appendix IV. It looks like the Model#3 has better performance.



Statistices Means for Model #1

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
Seed	100	50.5000000	29.0114920	1.0000000	100.0000000
R_Square	100	0.9009734	0.0067865	0.8885028	0.9151643
Adj_R_Sq	100	0.8975496	0.0068595	0.8848989	0.9118980
AIC	100	-3145.73	81.4422228	-3420.17	-2965.05
AICC	100	-3142.95	81.3377587	-3417.10	-2961.92
SBC	100	-3996.00	90.4027363	-4290.57	-3771.98
ASETrain	100	0.0158058	0.0011078	0.0134399	0.0179612
ASETest	100	0.0200599	0.0037957	0.0142283	0.0294224
CV_PRESS	100	19.0924368	1.6131988	15.1253054	22.1546819

Statistices Means for Model #2

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
Seed	100	50.5000000	29.0114920	1.0000000	100.0000000
R_Square	100	0.8969197	0	0.8969197	0.8969197
Adj_R_Sq	100	0.8935248	0	0.8935248	0.8935248
AIC	100	-3176.71	0	-3176.71	-3176.71
AICC	100	-3174.19	0	-3174.19	-3174.19
SBC	100	-4046.64	0	-4046.64	-4046.64
ASETrain	100	0.0160204	0	0.0160204	0.0160204
ASETest	100	0.0242969	0	0.0242969	0.0242969
CV_PRESS	100	17.3473806	0	17.3473806	17.3473806

Statistices Means for Model #3

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
Seed	100	50.5000000	29.0114920	1.0000000	100.0000000
R_Square	100	0.9051505	0.0071935	0.8884262	0.9235341
Adj_R_Sq	100	0.9008933	0.0070915	0.8845866	0.9190455
AIC	100	-3170.50	83.0746077	-3395.70	-2945.63
AICC	100	-3165.83	82.7949587	-3392.63	-2942.67
SBC	100	-3972.03	97.2841259	-4266.10	-3742.68
ASETrain	100	0.0151386	0.0011673	0.0120434	0.0179533
ASETest	100	0.0199581	0.0043838	0.0132157	0.0319007
CV_PRESS	100	18.8222237	1.6453196	14.3855762	21.9622825

Fig.2.6 Model Comparison Results for Selected Model

Please see table 1. for the comparison of the 3 models on the Adjusted R2, AIC and ASE(Test). Model#3 has higher AdjustedR2, lower AIC, lower ASE(Test) and lower Kaggle Score, it is the model we are going to select.

Test Set Models	Adjusted R2	AIC	ASE (Test)	Kaggle Score
Model 1 (LASSO)	0.8975496	-3145.73	0.0200599	0.12823
Model 2 (Stepwise)	0.8935248	-3176.71	0.0242969	0.13017
Model 3 (Best)	0.9008933	-3170.50	0.0199581	0.12541

Table.1 Model Comparison Results

Kaggle highest rank. Team Name “Cool Bin”.

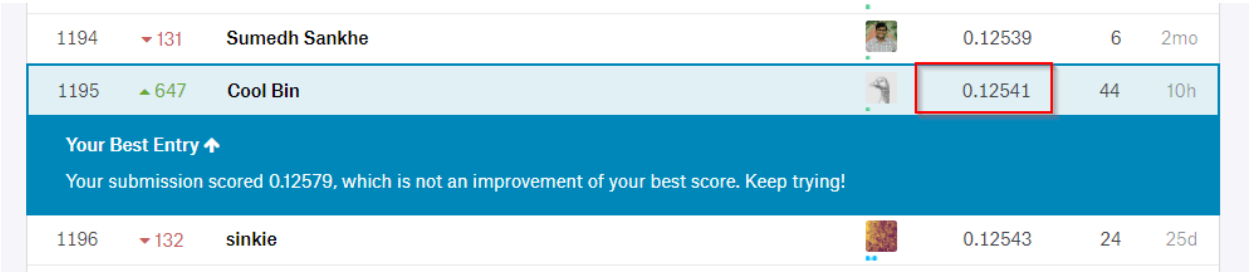


Fig.2.6 Kaggle Score for the Best Model

Conclusion: Based on the analysis of Ames housing data, we do find out the housing price is predictable base on the historical data of the house. There are 90.1% (R^2) of below housing information can describe the sale price of the house in the location of the Ames of state of Iowa. We do have 2 other models selected by LASSO and stepwise. But they are not as good as this custom model. We will provide the parameters included in the model and the accuracy of the model to the customer. Based on inputs from the customer, we can change (add/remove/modify) the parameters to provide more practical significance to the model. The

house cannot be reassigned to another group, so this is observation study. Any correlations made ban only be applied to the Ames, Iowa housing market.

RoofMatl MasVnrType Neighborhood BldgType Exterior1st_gr Exterior2nd_gr ExterQual_gr BsmtQual_gr LogGrLivArea LogLotArea LogTotalBsmtSF LogWoodDeckSF LogTotalOverall LogTotalBath LogTotalBsmtBath GarageArea FirstFlrSF SecondFlrSF YearRemodAdd YearBuilt logTotalExteriorQual logTotalGarageQual logTotalBsmtQual logTotalBsmtFinType logTotalLandQual logSaleAge

Appendix I

Data Import and Cleaning SAS Code:

```
  /*********************************************
** MSDS 6372 Project #1:   Analysis
** 02/11/2018 Create by:Bin Yu
**                                          Shravan Reddy
*********************************************/

/* Import Test dataset*/

FILENAME REFFILE
'C:\Users\yubin\OneDrive\MyWork\SMU\MSDS6372\Unit6Project\test.csv';

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

/* Import Training dataset*/
FILENAME REFFILE
'C:\Users\yubin\OneDrive\MyWork\SMU\MSDS6372\Unit6Project\train1.csv';

PROC IMPORT DATAFILE=REFFILE
      DBMS=CSV Replace
      OUT=train;
      GETNAMES=YES;
RUN;


/*
proc print data = train;
run;
proc print data = test;
run;

*/
/*Drop fields which have not much true data*/
data test;
```

```
      set test;
      Drop Ally;
      Drop Fence;
      Drop MiscFeature;
run;

/**Add . in SalePrice on the Test dataset*/

data test;
set test;
SalePrice = .;
;

/**Merge Train and Test Dataset*/
data train2;
      set train test;
run;


/*
proc means data = train2  n nmiss;
run;
proc print data = train2;
run;
proc sql;
      Select *
      From train2
      where BsmtBath<0;
Run
*/
/* There are some value -1 in the source data could be missing value, replace
it to median of the value in the variable*/

proc sql;
      /*Continouse Variables*/
      update train2
      set LotFrontage=68.00000 /*Median*/
      where LotFrontage=-1;
      update train2
      set MasVnrArea=0.01 /*avoid the log error*/
      where MasVnrArea=-1 | MasVnrArea=0;
      update train2
      set GarageYrBlt=1979.00000 /*Median*/
      where GarageYrBlt=-1;
      update train2
      set BsmtFinSF1=368.50000 /*Median*/
      where BsmtFinSF1=-1;
      update train2
      set BsmtFinSF1=0.01 /*avoid the log error*/
```

```
where BsmtFinSF1=0;
update train2
set BsmtFinSF2=0.00000 /*Median*/
where BsmtFinSF2=-1;
update train2
set BsmtFinSF2=0.01 /*avoid the log error*/
where BsmtFinSF2=0;

update train2
set BsmtUnfSF=467.00000 /*Median*/
where BsmtUnfSF=-1;
update train2
set BsmtUnfSF=0.01 /*avoid the log error*/
where BsmtUnfSF=0;

update train2
set TotalBsmtSF=989.50000 /*Median*/
where TotalBsmtSF=-1;
update train2
set TotalBsmtSF=0.01 /*avoid the log error*/
where TotalBsmtSF=0;

update train2
set BsmtFullBath=0.00000 /*Median*/
where BsmtFullBath=-1;
update train2
set BsmtHalfBath=0.00000 /*Median*/
where BsmtHalfBath=-1;
update train2
set GarageCars=2.00000 /*Median*/
where GarageCars=-1;
update train2
set GarageCars=0.01 /*avoid the log error*/
where GarageCars=0;
update train2
set GarageArea=480.00000 /*Median*/
where GarageArea=-1;
update train2
set GarageArea=0.01 /*avoid the log error*/
where GarageArea=0;

update train2
set SecondFlrSF=0.01 /*avoid the log error*/
where SecondFlrSF=0;
update train2
set LowQualFinSF=0.01 /*avoid the log error*/
where LowQualFinSF=0;

update train2
```

```
        set WoodDeckSF=0.01 /*avoid the log error*/
        where WoodDeckSF=0;
        update train2
        set OpenPorchSF=0.01 /*avoid the log error*/
        where OpenPorchSF=0;
        update train2
        set PoolArea=0.01 /*avoid the log error*/
        where PoolArea=0;
        update train2
        set EnclosedPorch=0.01 /*avoid the log error*/
        where EnclosedPorch=0;
        update train2
        set ThreeSsnPorch=0.01 /*avoid the log error*/
        where ThreeSsnPorch=0;
        update train2
        set MiscVal=0.01 /*avoid the log error*/
        where MiscVal=0;
        update train2
        set ScreenPorch=0.01 /*avoid the log error*/
        where ScreenPorch=0;


Run;
Quit;
/*fix bad data*/
Proc SQL;
        update train2
        set Neighborhood='NAmes' /*Mod*/
        where Neighborhood='-1mes';
run;

/* Add New Variables */
/*get the average of the salesprice by Neighborhood*/
proc sql;
        create table temp as
        select Neighborhood,avg(SalePrice) as Neighborhood_Avg
                from train2
                group by Neighborhood;
run;

Proc Sort data =train2;
by Neighborhood;
run;
/*Merge 2 dataset to get the average saleprice for each record*/
Data train2;
        merge Train2 temp;
        by Neighborhood;
run;
```

```
/*Categorical Variable None Values*/
Proc SQL;
      update train2
      set MSZoning='FV' /*Mod*/
      where MSZoning='NA';
      update train2
      set Exterior1st='Wd Sdng' /*Mod*/
      where Exterior1st='NA';
      update train2
      set Exterior2nd='Wd Sdng' /*Mod*/
      where Exterior2nd='NA';
      update train2
      set MasVnrType='Stone' /*Mod*/
      where MasVnrType='NA'  | MasVnrType='-1';
      update train2
      set BsmtQual='NoA' /*Mod*/
      where BsmtQual='NA' | BsmtQual='-1';
      update train2
      set BsmtCond='NoA' /*Mod*/
      where BsmtCond='NA' | BsmtCond='-1';
      update train2
      set BsmtExposure='NoA' /*Mod*/
      where BsmtExposure='NA' | BsmtExposure='-1';
      update train2
      set BsmtFinType1='NoA' /*Mod*/
      where BsmtFinType1='NA'  | BsmtFinType1='-1';
      update train2
      set BsmtFinType2='NoA' /*Mod*/
      where BsmtFinType2='NA' | BsmtFinType2='-1';
      update train2
      set Electrical='FuseF' /*Mod*/
      where Electrical='NA' | Electrical='-1';
      update train2
      set KitchenQual='Ex' /*Mod*/
      where KitchenQual='NA' | KitchenQual='-1';
      update train2
      set Functional='Mod' /*Mod*/
      where Functional='NA';
      update train2
      set FireplaceQu='NoA' /*Mod*/
      where FireplaceQu='NA' | FireplaceQu='-1';
      update train2
      set GarageType='NoA' /*Mod*/
      where GarageType='NA'  | GarageType='-1';
      update train2
      set GarageFinish='NoA' /*Mod*/
      where GarageFinish='NA'  | GarageFinish='-1';
      update train2
      set GarageQual='NoA' /*Mod*/
```

```
      where GarageQual='NA'  | GarageQual='-1';
      update train2
      set GarageCond='NoA' /*Mod*/
      where GarageCond='NA'  | GarageCond='-1';

      update train2
      set Utilities='NoA' /*Mod*/
      where Utilities='NA';
      update train2
      set SaleType='COD' /*Mod*/
      where SaleType='NA';
      update train2
      set LotConfig='CulDSa' /*Fix length issue*/
      where LotConfig='CulDSac';
run;
/*Encode Categorical Variables */
data train2;
      Set train2;

      If BsmtQual='Ex' then BsmtQual_gr=5;
            else if BsmtQual='Gd' then BsmtQual_gr=4;
            else if BsmtQual='TA' then BsmtQual_gr=3;
            else if BsmtQual='Fa' then BsmtQual_gr=2;
            else if BsmtQual='Po' then BsmtQual_gr=1;
            else BsmtQual_gr=0;
      If BsmtCond='Ex' then BsmtCond_gr=5;
            else if BsmtCond='Gd' then BsmtCond_gr=4;
            else if BsmtCond='TA' then BsmtCond_gr=3;
            else if BsmtCond='Fa' then BsmtCond_gr=2;
            else if BsmtCond='Po' then BsmtCond_gr=1;
            else BsmtCond_gr=0;
      If  BsmtFinType1='GLQ' then BsmtFinType1_gr=6;
            else if BsmtFinType1='ALQ' then BsmtFinType1_gr=5;
            else if BsmtFinType1='BLQ' then BsmtFinType1_gr=4;
            else if BsmtFinType1='Rec' then BsmtFinType1_gr=3;
            else if BsmtFinType1='LWQ' then BsmtFinType1_gr=2;
            else if BsmtFinType1='Unf' then BsmtFinType1_gr=1;
            else BsmtFinType1_gr=0;
      If  BsmtFinType2='GLQ' then BsmtFinType2_gr=6;
            else if BsmtFinType2='ALQ' then BsmtFinType2_gr=5;
            else if BsmtFinType2='BLQ' then BsmtFinType2_gr=4;
            else if BsmtFinType2='Rec' then BsmtFinType2_gr=3;
            else if BsmtFinType2='LWQ' then BsmtFinType2_gr=2;
            else if BsmtFinType2='Unf' then BsmtFinType2_gr=1;
            else BsmtFinType2_gr=0;
      If BsmtExposure='Gd' then BsmtExposure_gr=5;
            else if BsmtExposure='Av' then BsmtExposure_gr=4;
            else if BsmtExposure='Mn' then BsmtExposure_gr=3;
            else BsmtExposure_gr=0;
```

```
If ExterQual='Ex' then ExterQual_gr=5;
      else if ExterQual='Gd' then ExterQual_gr=4;
      else if ExterQual='TA' then ExterQual_gr=3;
      else if ExterQual='Fa' then ExterQual_gr=2;
      else if ExterQual='Po' then ExterQual_gr=1;
      else ExterQual_gr=0;
If ExterCond='Ex' then ExterCond_gr=5;
      else if ExterCond='Gd' then ExterCond_gr=4;
      else if ExterCond='TA' then ExterCond_gr=3;
      else if ExterCond='Fa' then ExterCond_gr=2;
      else if ExterCond='Po' then ExterCond_gr=1;
      else ExterCond_gr=0;
If Exterior1st='VinylSd' then Exterior1st_gr=6;
      else if Exterior1st='MetalSd' then Exterior1st_gr=5;
      else if Exterior1st='HdBoard' then Exterior1st_gr=4;
      else if Exterior1st='Wd Sdng' then Exterior1st_gr=3;
      else if Exterior1st='Plywood' then Exterior1st_gr=2;
      else if Exterior1st='(Other)' then Exterior1st_gr=1;
      else Exterior1st_gr=0;
If Exterior2nd='VinylSd' then Exterior2nd_gr=6;
      else if Exterior2nd='MetalSd' then Exterior2nd_gr=5;
      else if Exterior2nd='HdBoard' then Exterior2nd_gr=4;
      else if Exterior2nd='Wd Sdng' then Exterior2nd_gr=3;
      else if Exterior2nd='Plywood' then Exterior2nd_gr=2;
      else if Exterior2nd='(Other)' then Exterior2nd_gr=1;
      else Exterior2nd_gr=0;
If Functional='Typ' then Functional_gr=8;
      else if Functional='Min1' then Functional_gr=7;
      else if Functional='Min2' then Functional_gr=6;
      else if Functional='Mod' then Functional_gr=5;
      else if Functional='Maj1' then Functional_gr=4;
      else if Functional='Maj2' then Functional_gr=3;
      else if Functional='Sev' then Functional_gr=2;
      else if Functional='Sal' then Functional_gr=1;
      else Functional_gr=0;
If GarageQual='Ex' then GarageQual_gr=5;
      else if GarageQual='Gd' then GarageQual_gr=4;
      else if GarageQual='TA' then GarageQual_gr=3;
      else if GarageQual='Fa' then GarageQual_gr=2;
      else if GarageQual='Po' then GarageQual_gr=1;
      else GarageQual_gr=0;
If GarageCond='Ex' then GarageCond_gr=5;
      else if GarageCond='Gd' then GarageCond_gr=4;
      else if GarageCond='TA' then GarageCond_gr=3;
      else if GarageCond='Fa' then GarageCond_gr=2;
      else if GarageCond='Po' then GarageCond_gr=1;
      else GarageCond_gr=0;
If GarageFinish='Fin' then GarageFinish_gr=3;
      else if GarageFinish='RFn' then GarageFinish_gr=2;
```

```
      else if GarageFinish='Unf' then GarageFinish_gr=1;
      else GarageFinish_gr=0;
If KitchenQual='Ex' then KitchenQual_gr=5;
      else if KitchenQual='Gd' then KitchenQual_gr=4;
      else if KitchenQual='TA' then KitchenQual_gr=3;
      else if KitchenQual='Fa' then KitchenQual_gr=2;
      else if KitchenQual='Po' then KitchenQual_gr=1;
      else KitchenQual_gr=0;
If LandSlope='Gtl' then LandSlope_gr=3;
      else if LandSlope='Mod' then LandSlope_gr=2;
      else if LandSlope='Sev' then LandSlope_gr=1;
      else LandSlope_gr=0;
If LotShape='Reg' then LotShape_gr=4;
      else if LotShape='IR1' then LotShape_gr=3;
      else if LotShape='IR2' then LotShape_gr=2;
      else if LotShape='IR3' then LotShape_gr=1;
      else LotShape_gr=0;
If MSSubClass='20' then MSSubClass_Chr='MSSC20';
      else if MSSubClass='30' then MSSubClass_Chr='MSSC50';
      else if MSSubClass='40' then MSSubClass_Chr='MSSC40';
      else if MSSubClass='45' then MSSubClass_Chr='MSSC45';
      else if MSSubClass='50' then MSSubClass_Chr='MSSC50';
      else if MSSubClass='60' then MSSubClass_Chr='MSSC60';
      else if MSSubClass='70' then MSSubClass_Chr='MSSC70';
      else if MSSubClass='75' then MSSubClass_Chr='MSSC75';
      else if MSSubClass='80' then MSSubClass_Chr='MSSC80';
      else if MSSubClass='85' then MSSubClass_Chr='MSSC85';
      else if MSSubClass='90' then MSSubClass_Chr='MSSC90';
      else if MSSubClass='120' then MSSubClass_Chr='MSSC120';
      else if MSSubClass='150' then MSSubClass_Chr='MSSC150';
      else if MSSubClass='160' then MSSubClass_Chr='MSSC160';
      else if MSSubClass='180' then MSSubClass_Chr='MSSC180';
      else if MSSubClass='190' then MSSubClass_Chr='MSSC190';
      else MSSubClass_Chr=0;
If MoSold='1' then MoSold_Chr='Jan';
      else if MoSold='2' then MoSold_Chr='Feb';
      else if MoSold='3' then MoSold_Chr='Mar';
      else if MoSold='4' then MoSold_Chr='Apr';
      else if MoSold='5' then MoSold_Chr='May';
      else if MoSold='6' then MoSold_Chr='Jun';
      else if MoSold='7' then MoSold_Chr='Jul';
      else if MoSold='8' then MoSold_Chr='Aug';
      else if MoSold='9' then MoSold_Chr='Sep';
      else if MoSold='10' then MoSold_Chr='Oct';
      else if MoSold='11' then MoSold_Chr='Nov';
      else if MoSold='12' then MoSold_Chr='Dec';
      else MoSold_Chr='NoA';
If PavedDrive='Y' then PavedDrive_gr=3;
      else if PavedDrive='P' then PavedDrive_gr=2;
```

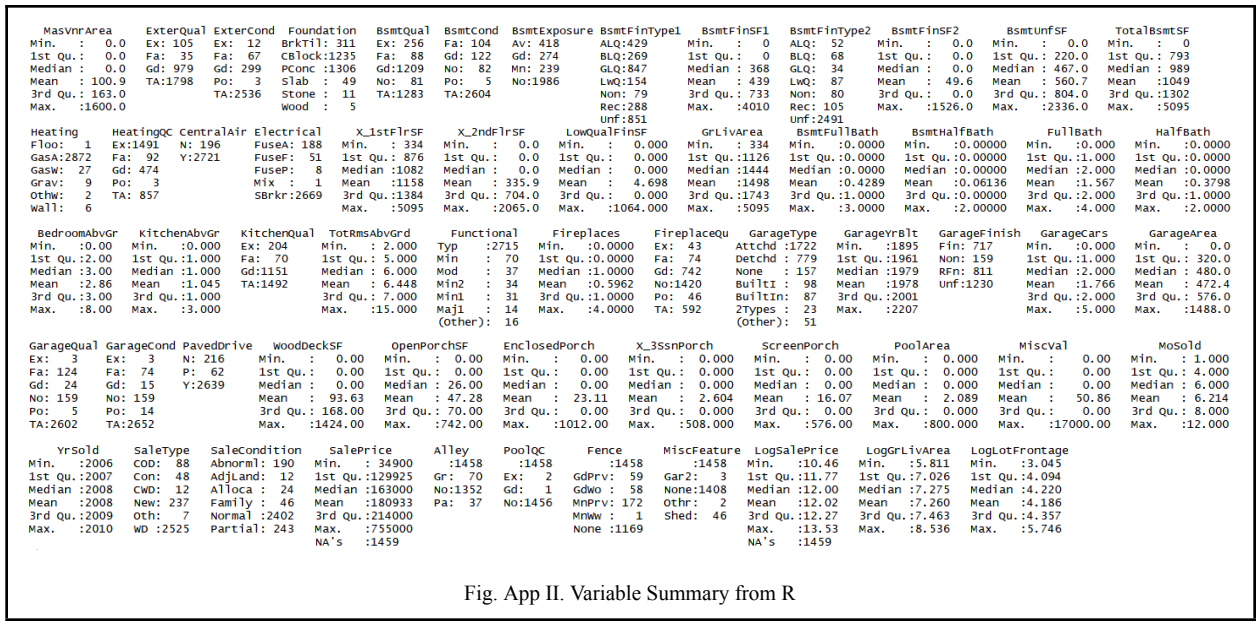
```
            else if PavedDrive='N' then PavedDrive_gr=1;
            else PavedDrive_gr=0;
      If Street='Pave' then Street_gr=3;
            else if Street='Grv1' then Street_gr=2;
            else Street_gr=0;
      If Utilities='AllPub' then Utilities_gr=4;
            else if Utilities='NoSewr' then Utilities_gr=3;
            else if Utilities='NoSeWa' then Utilities_gr=2;
            else if Utilities='ELO' then Utilities_gr=1;
            else Utilities_gr=0;

      If Neighborhood in
('NridgHt','Veenker','Somerst','Timber','StoneBr','NoRidge') then
Neighborhood_gr=3;
            else If Neighborhood in
('Gilbert','NWAmes','Blmngtn','CollgCr','ClearCr','Crawfor') then
Neighborhood_gr=2;
            else If Neighborhood in
('Blueste','SWISU','NAmes','NPkVill','Mitchel','SawyerW') then
Neighborhood_gr=1;
            else Neighborhood_gr=0;
run;
/*Combine as new variables*/

data train2;
      Set train2;
      TotalOverall = OverallQual+OverallCond; /*combine 2 columns*/
      TotalBsmtQual = BsmtQual_gr+BsmtCond_gr; /*combine 2 columns*/
      TotalBsmtFinSF = BsmtFinSF1+BsmtFinSF2; /*combine 2 columns*/
      TotalExteriorQual = ExterQual_gr+ExterCond_gr; /*combine 2 columns*/
      TotalExterior = Exterior1st_gr+Exterior2nd_gr; /*combine 2 columns*/
      TotalGarageQual = GarageQual_gr+GarageCond_gr + GarageFinish_gr;
/*combine 2 columns*/
      TotalBsmtBath = BsmtFullBath+BsmtHalfBath; /*combine 2 columns*/
      TotalBsmtFinType = BsmtFinType1_gr+BsmtFinType2_gr;
      TotalBath = FullBath+0.5*HalfBath + BsmtFullBath+ 0.5*BsmtHalfBath;
/*combine 2 columns*/
      TotalLandQual = LandSlope_gr+LotShape_gr;
      TotalSF     = GrLivArea + TotalBsmtSF;
      SoldYearMo = YrSold * 100 +MoSold;
      SaleAge = YrSold - YearBuilt;
run;
/*Fix log issue update to 0.01 if original value less than 0 to not change
much original values*/
Proc SQL;
      update train2
      set TotalOverall=0.01 /*avoid the log error*/
      where TotalOverall=0;
      update train2
```

```
        set TotalBsmtQual=0.01 /*avoid the log error*/
        where TotalBsmtQual=0;
        update train2
        set TotalBsmtFinSF=0.01 /*avoid the log error*/
        where TotalBsmtFinSF=0;
        update train2
        set TotalExteriorQual=0.01 /*avoid the log error*/
        where TotalExteriorQual=0;
        update train2
        set TotalGarageQual=0.01 /*avoid the log error*/
        where TotalGarageQual=0;
        update train2
        set TotalBsmtBath=0.01 /*avoid the log error*/
        where TotalBsmtBath=0;
        update train2
        set TotalBsmtFinType=0.01 /*avoid the log error*/
        where TotalBsmtFinType=0;
        update train2
        set TotalBath=0.01 /*avoid the log error*/
        where TotalBath=0;
        update train2
        set TotalLandQual=0.01 /*avoid the log error*/
        where TotalLandQual=0;
        update train2
        set TotalSF=0.01 /*avoid the log error*/
        where TotalSF=0;
        update train2
        set SaleAge=0.01 /*avoid the log error*/
        where SaleAge<=0;

Run;
quit;

/*Natual Log Transformation*/
data train2;
 set train2;
 LogSalePrice = log(SalePrice);
 LogGrLivArea = log(GrLivArea);
 LogLotFrontage = log(LotFrontage);
 LogMasVnrArea = log(MasVnrArea);
 LogLotArea = log(LotArea);
 LogGarageArea = log(GarageArea);
 logBsmtUnfSF = log(BsmtUnfSF);

 LogFirstFlrSF = log(FirstFlrSF);
 LogTotRmsAbvGrd = log(TotRmsAbvGrd);
 LogSecondFlrSF = log(SecondFlrSF);
 LogPoolArea = log(PoolArea);
 LogTotalBath = log(TotalBath);
```

```
LogWoodDeckSF = log(WoodDeckSF);
LogOpenPorchSF = log(OpenPorchSF);
LogEnclosedPorch = log(EnclosedPorch);
LogThreeSsnPorch = log(ThreeSsnPorch);
logMiscVal =log(MiscVal);
logScreenPorch = log(ScreenPorch);

LogSaleAge=Log(SaleAge);
LogTotalBsmtFinSF = log(TotalBsmtFinSF);
LogTotalSF = log(TotalSF);
logTotalBsmtSF = log(TotalBsmtSF);
logLowQualFinSF=log(LowQualFinSF);

LogTotalOverall = log(TotalOverall);
logTotalBsmtBath =log(TotalBsmtBath);

logTotalBsmtQual = log(TotalBsmtQual);
logTotalExteriorQual = log(TotalExteriorQual);
logTotalGarageQual = log(TotalGarageQual);
logTotalBsmtFinType = log(TotalBsmtFinType);
logTotalLandQual = log(TotalLandQual);
logNeighborhood_avg = log(Neighborhood_avg);


run;
/*sort the train2 by ID*/
proc sort data=train2;
by ID;
run;
```

[Appendix II](#)

```
> train3 <- read.csv("C:\\Users\\yubin\\OneDrive\\MyWork\\SMU\\MSDS6372\\Unit6Project\\train2.csv")
> summary(train3)
       Id          MSSubClass      MSZoning   LotFrontage       LotArea          Street     LotShape   LandContour  Utilities     LotConfig    LandSlope   Neighborhood    Condition1
 Min.   :   1   Min.   : 20.00   C :  25   Min.   : 21.00   Min.   :  1300   Grvl:  12   IR1: 967   Bnk: 115    AllPub:2914   Corner : 510   Gtl:2776   CollgCr: 267   Norm   :2511
 1st Qu.: 731   1st Qu.: 20.00   FV: 143   1st Qu.: 60.00   1st Qu.:  7476   Pave:2905   IR2:  76   HLS: 120    None   :   2   CulDSa :  82   Mod: 125   OldTown: 239   Feedr  : 163
 Median :1461   Median : 50.00   RH:  26   Median : 68.00   Median :  9452               IR3:  15   Low:  60    NoSeWa:   1   CulDSac:  94   Sev:  16   Names  : 225   RRAn   :  50
 Mean   :1460   Mean   : 57.14   RL:2263   Mean   : 68.98   Mean   : 10139               Reg:1859   Lvl:2622                  FR2    :  85                 NAmes  : 218   Artery :  48
 3rd Qu.:2190   3rd Qu.: 70.00   RM: 460   3rd Qu.: 78.00   3rd Qu.: 11556                                                    FR3    :  14                 Edwards: 192   Arter  :  44
 Max.   :2919   Max.   :190.00            Max.   :313.00   Max.   :215245                                                    Inside :2132                 Somerst: 182   PosN   :  38
                                                                                                                                                          (Other):1594   (Other):  63
   Condition2     BldgType      HouseStyle    OverallQual      OverallCond      YearBuilt     YearRemodAdd  RoofStyle      RoofMatl     Exterior1st   Exterior2nd     MasVnrType
 Norm   :2888   1Fam  :2423   1Story :1471   Min.   : 1.000   Min.   :1.000   Min.   :1872   Min.   :1950   Flat :  20   CompShg:2875   Vinylsd:1025   Vinylsd:1014   BrkCmn :  25
 Feed   :   7   2fmCon:  62   2Story : 870   1st Qu.: 5.000   1st Qu.:5.000   1st Qu.:1953   1st Qu.:1965   Gable:2310   Membran:   1   MetalSd: 450   MetalSd: 447   BrkFace: 879
 Feedr  :   6   Duplex: 109   1.5Fin : 314   Median : 6.000   Median :5.000   Median :1973   Median :1993   Gambr:  22   Metal  :   1   HdBoard: 442   HdBoard: 406   None   :1742
 PosA   :   4   Twnhs :  96   SLvl   : 128   Mean   : 6.086   Mean   :5.565   Mean   :1971   Mean   :1984   Hip  : 549   Roll   :   1   Wd Sdng: 412   Wd Sdng: 392   Stone  : 271
 Artе   :   3   TwnhsE: 227   SFoyer :  83   3rd Qu.: 7.000   3rd Qu.:6.000   3rd Qu.:2001   3rd Qu.:2004   Mansa:  11   Tar&Grv:  23   Plywood: 221   Plywood: 270
 PosN   :   3                 2.5Unf :  24   Max.   :10.000   Max.   :9.000   Max.   :2010   Max.   :2010   Shed :   5   WdShake:   9   CemntBd: 125   CmentBd: 125
 (Other):   6                 (Other):  27                                                                   WdShngl:   7   (Other): 242   (Other): 263
```



```
  MasVnrArea     ExterQual ExterCond  Foundation   BsmtQual  BsmtCond  BsmtExposure BsmtFinType1   BsmtFinSF1    BsmtFinType2   BsmtFinSF2       BsmtUnfSF       TotalBsmtSF
Min.   :   0.0   Ex: 105   Ex:  12   BrkTil: 311   Ex: 256   Fa: 104   Av: 418      ALQ:429       Min.   :   0   ALQ:  52   Min.   :   0.0   Min.   :   0.0   Min.   :   0
1st Qu.:   0.0   Fa:  35   Fa:  67   CBlock:1235   Fa:  88   Gd: 122   Gd: 274      BLQ:269       1st Qu.:   0   BLQ:  68   1st Qu.:   0.0   1st Qu.: 220.0   1st Qu.: 793
Median :   0.0   Gd: 979   Gd: 299   PConc :1306   Gd:1209   No:  82   Mn: 239      GLQ:847       Median : 368   GLQ:  34   Median :   0.0   Median : 467.0   Median : 989
Mean   : 100.9   TA:1798   Po:   3   Slab  :  49   No:  81   Po:   5   No:1986      LwQ:154       Mean   : 439   LwQ:  87   Mean   :  49.6   Mean   : 560.7   Mean   :1049
3rd Qu.: 163.0             TA:2536   Stone :  11   TA:1283   TA:2604                Non: 79       3rd Qu.: 733   Non:  80   3rd Qu.:   0.0   3rd Qu.: 804.0   3rd Qu.:1302
Max.   :1600.0                      Wood  :   5                                      Rec:288       Max.   :4010   Rec: 105   Max.   :1526.0   Max.   :2336.0   Max.   :5095
                                                                                     Unf:851                       Unf:2491

  Heating     HeatingQC CentralAir Electrical     X_1stFlrSF     X_2ndFlrSF     LowQualFinSF       GrLivArea     BsmtFullBath    BsmtHalfBath       FullBath        HalfBath
Floo:   1   Ex:1491   N: 196     FuseA: 188   Min.   : 334   Min.   :   0.0   Min.   :   0.000   Min.   : 334   Min.   :0.0000   Min.   :0.00000   Min.   :0.000   Min.   :0.0000
GasA:2872   Fa:  92   Y:2721     FuseF:  51   1st Qu.: 876   1st Qu.:   0.0   1st Qu.:   0.000   1st Qu.:1126   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:1.000   1st Qu.:0.0000
GasW:  27   Gd: 474              FuseP:   8   Median :1082   Median :   0.0   Median :   0.000   Median :1444   Median :0.0000   Median :0.00000   Median :2.000   Median :0.0000
Grav:   9   Po:   3              Mix  :   1   Mean   :1158   Mean   : 335.9   Mean   :   4.698   Mean   :1498   Mean   :0.4289   Mean   :0.06136   Mean   :1.567   Mean   :0.3798
OthW:   2   TA: 857              SBrkr:2669   3rd Qu.:1384   3rd Qu.: 704.0   3rd Qu.:   0.000   3rd Qu.:1743   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:2.000   3rd Qu.:1.0000
Wall:   6                                    Max.   :5095   Max.   :2065.0   Max.   :1064.000   Max.   :5095   Max.   :3.0000   Max.   :2.00000   Max.   :4.000   Max.   :2.0000

 BedroomAbvGr   KitchenAbvGr   KitchenQual  TotRmsAbvGrd      Functional     Fireplaces     FireplaceQu   GarageType    GarageYrBlt   GarageFinish   GarageCars      GarageArea
Min.   :0.00   Min.   :0.000   Ex: 204     Min.   : 2.000   Typ    :2715   Min.   :0.0000   Ex:  43   Attchd :1722   Min.   :1895   Fin: 717   Min.   :0.000   Min.   :   0.0
1st Qu.:2.00   1st Qu.:1.000   Fa:  70     1st Qu.: 5.000   Min    :  70   1st Qu.:0.0000   Fa:  74   Detchd : 779   1st Qu.:1961   Non: 159   1st Qu.:1.000   1st Qu.: 320.0
Median :3.00   Median :1.000   Gd:1151     Median : 6.000   Mod    :  37   Median :1.0000   Gd: 742   None   : 157   Median :1979   RFn: 811   Median :2.000   Median : 480.0
Mean   :2.86   Mean   :1.045   TA:1492     Mean   : 6.448   Min2   :  34   Mean   :0.5962   No:1420   BuiltI :  98   Mean   :1978   Unf:1230   Mean   :1.766   Mean   : 472.4
3rd Qu.:3.00   3rd Qu.:1.000               3rd Qu.: 7.000   Min1   :  31   3rd Qu.:1.0000   Po:  46   BuiltIn:  87   3rd Qu.:2001              3rd Qu.:2.000   3rd Qu.: 576.0
Max.   :8.00   Max.   :3.000               Max.   :15.000   Maj1   :  14   Max.   :4.0000   TA: 592   2Types :  23   Max.   :2207              Max.   :5.000   Max.   :1488.0
                                                            (Other):  16                             (Other):  51

GarageQual GarageCond PavedDrive   WoodDeckSF      OpenPorchSF     EnclosedPorch     X_3SsnPorch       ScreenPorch       PoolArea          MiscVal            MoSold
Ex:   3    Ex:   3    N: 216     Min.   :   0.00   Min.   :  0.00   Min.   :   0.00   Min.   :  0.000   Min.   :  0.00   Min.   :  0.000   Min.   :    0.00   Min.   : 1.000
Fa: 124    Fa:  74    P:  62     1st Qu.:   0.00   1st Qu.:  0.00   1st Qu.:   0.00   1st Qu.:  0.000   1st Qu.:  0.00   1st Qu.:  0.000   1st Qu.:    0.00   1st Qu.: 4.000
Gd:  24    Gd:  15    Y:2639     Median :   0.00   Median : 26.00   Median :   0.00   Median :  0.000   Median :  0.00   Median :  0.000   Median :    0.00   Median : 6.000
No: 159    No: 159               Mean   :  93.63   Mean   : 47.28   Mean   :  23.11   Mean   :  2.604   Mean   : 16.07   Mean   :  2.089   Mean   :   50.86   Mean   : 6.214
Po:   5    Po:  14               3rd Qu.: 168.00   3rd Qu.: 70.00   3rd Qu.:   0.00   3rd Qu.:  0.000   3rd Qu.:  0.00   3rd Qu.:  0.000   3rd Qu.:    0.00   3rd Qu.: 8.000
TA:2602    TA:2652               Max.   :1424.00   Max.   :742.00   Max.   :1012.00   Max.   :508.000   Max.   :576.00   Max.   :800.000   Max.   :17000.00   Max.   :12.000

    YrSold     SaleType   SaleCondition   SalePrice      Alley     PoolQC      Fence      MiscFeature  LogSalePrice    LogGrLivArea    LogLotFrontage
Min.   :2006   COD:  88   Abnorml: 190   Min.   : 34900   :1458   :1458   :1458   :1458   Min.   :10.46   Min.   :5.811   Min.   :3.045
1st Qu.:2007   Con:  48   AdjLand:  12   1st Qu.:129925   Gr:  70   Ex:   2   GdPrv:  59   Gar2:   3   1st Qu.:11.77   1st Qu.:7.026   1st Qu.:4.094
Median :2008   CWD:  12   Alloca :  24   Median :163000   No:1352   Gd:   1   GdWo :  58   None:1408   Median :12.00   Median :7.275   Median :4.220
Mean   :2008   New: 237   Family :  46   Mean   :180933   Pa:  37   No:1456   MnPrv: 172   Othr:   2   Mean   :12.02   Mean   :7.260   Mean   :4.186
3rd Qu.:2009   Oth:   7   Normal :2402   3rd Qu.:214000             MnWw :   1   Shed:  46   3rd Qu.:12.27   3rd Qu.:7.463   3rd Qu.:4.357
Max.   :2010   WD :2525   Partial: 243   Max.   :755000             None :1169               Max.   :13.53   Max.   :8.536   Max.   :5.746
                                         NA's   :1459                                        NA's   :1459
```

Fig. App II. Variable Summary from R

Log Transformation list for the variables:

LogSalePrice = log(SalePrice);
LogGrLivArea = log(GrLivArea);
LogLotFrontage = log(LotFrontage);
LogMasVnrArea = log(MasVnrArea);
LogLotArea = log(LotArea);
LogGarageArea = log(GarageArea);
logBsmtUnfSF = log(BsmtUnfSF);
LogFirstFlrSF = log(FirstFlrSF);
LogTotRmsAbvGrd = log(TotRmsAbvGrd);
LogSecondFlrSF = log(SecondFlrSF);
LogPoolArea = log(PoolArea);
LogTotalBath = log(TotalBath);
LogWoodDeckSF = log(WoodDeckSF);
LogOpenPorchSF = log(OpenPorchSF);
LogEnclosedPorch = log(EnclosedPorch);
LogThreeSsnPorch = log(ThreeSsnPorch);
logMiscVal =log(MiscVal);
logScreenPorch = log(ScreenPorch);
LogSaleAge=Log(SaleAge);
LogTotalBsmtFinSF = log(TotalBsmtFinSF);
LogTotalSF = log(TotalSF);
logTotalBsmtSF = log(TotalBsmtSF);
logLowQualFinSF=log(LowQualFinSF);

For combined categorical variables, we also did some transformation:

LogTotalOverall = log(TotalOverall);

```
logTotalBsmtBath =log(TotalBsmtBath);
logTotalBsmtQual = log(TotalBsmtQual);
logTotalExteriorQual = log(TotalExteriorQual);
logTotalGarageQual = log(TotalGarageQual);
logTotalBsmtFinType = log(TotalBsmtFinType);
logTotalLandQual = log(TotalLandQual);
logNeighborhood_avg = log(Neighborhood_avg);
```

Appendix III

```
/*********************************************
** MSDS 6372 Project #1:   Analysis Question #1
** 02/11/2018 Create by:    Bin Yu
**                                        Shravan Reddy
*********************************************/
/*High Inclurence Points and remove from the dataset*/

/*Evaluated all variables and selected most interpretable and available to
audience

Ran LASSO selection to narrow variables to the most simple model with the most
significant

variables*/

proc glmselect data=train2 plots = all;
      class Neighborhood;
      model LogSalePrice = LogLotFrontage LogLotArea LogTotalBsmtSF
LogFirstFlrSF LogSecondFlrSF
      LogGrLivArea OpenPorchSF LogTotalOverall YearBuilt LogTotalBath
LogTotRmsAbvGrd LogGarageCars
      SoldYearMo Neighborhood
      /selection=LASSO(choose=SBC stop=cv) cvmethod=random(2) hierarchy=single
showpvalues stat=all;
      output out=LassoData p=predict r=resid p = yhat;
      modelaverage tables = (EffectSelectPct(all) ParmEst(All)) alpha = .1;

run;

/*GLM procedure to get estimates and diagnostic plots*/

proc glm data=train2 plots (UNPACK) = DIAGNOSTICS(LABEL);
      class Neighborhood;
      model LogSalePrice = LogGrLivArea YearBuilt LogFirstFlrSF LogLotArea
LogTotalBsmtSF
      OpenPorchSF Neighborhood
      / CLI CLPARM CLM SOLUTION TOLERANCE CLM;
      output out=CustData p=predict PRESS=CVPress;
run;

/*Remove Influential Points*/
data train2;
      set train2;
      if ID ne 1299 & ID ne 524 & ID ne 496;
run;
```

```
/*After Removed Influential Points*/

proc glm data=train2 plots (UNPACK) = DIAGNOSTICS(LABEL);
      class Neighborhood;
      model LogSalePrice = LogGrLivArea YearBuilt LogFirstFlrSF LogLotArea
LogTotalBsmtSF
      OpenPorchSF Neighborhood
      / CLI CLPARM CLM SOLUTION TOLERANCE CLM;
      output out=CustData p=predict PRESS=CVPress;
run;
```

Appendix IV

```
/*********************************************
** MSDS 6372 Project #1:   Analysis Question #2
** 02/11/2018 Create by:    Bin Yu
**                                          Shravan Reddy
*********************************************/

/*Best Model Selected*/
proc glm data = train2 plots(UNPACK) = DIAGNOSTICS(LABEL);
            class RoofMatl MasVnrType BldgType Neighborhood MoSold_Chr
MSSubClass_Chr;
            model logSalePrice = RoofMatl MasVnrType Neighborhood BldgType
Exterior1st_gr Exterior2nd_gr ExterQual_gr BsmtQual_gr
              LogGrLivArea LogLotArea LogTotalBsmtSF LogWoodDeckSF
LogTotalOverall LogTotalBath LogTotalBsmtBath GarageArea FirstFlrSF SecondFlrSF
YearRemodAdd YearBuilt
              logTotalExteriorQual logTotalGarageQual logTotalBsmtQual
logTotalBsmtFinType logTotalLandQual logSaleAge
      / CLI CLPARM CLM SOLUTION TOLERANCE CLM;
run;
Quit;

/*Remove Influential Points*/
data train2;
      set train2;
      if  ID ne 524;
run;

/*After Removeing Influential Points*/
proc glm data = train2 plots(UNPACK) = DIAGNOSTICS(LABEL);
            class RoofMatl MasVnrType BldgType Neighborhood MoSold_Chr
MSSubClass_Chr;
            model logSalePrice = RoofMatl MasVnrType Neighborhood BldgType
Exterior1st_gr Exterior2nd_gr ExterQual_gr BsmtQual_gr
              LogGrLivArea LogLotArea LogTotalBsmtSF LogWoodDeckSF
LogTotalOverall LogTotalBath LogTotalBsmtBath GarageArea FirstFlrSF SecondFlrSF
YearRemodAdd YearBuilt
              logTotalExteriorQual logTotalGarageQual logTotalBsmtQual
logTotalBsmtFinType logTotalLandQual logSaleAge
      / CLI CLPARM CLM SOLUTION TOLERANCE CLM;
run;
Quit;
/*LASSO Model Selection*/
```

```
proc glmselect data = train2 plots (stepaxis = number) = (criterionpanel
ASEPlot) seed =55; /*86*/
            partition fraction (test = .3);
            class RoofMatl MasVnrType BldgType Neighborhood MoSold_Chr
MSSubClass_Chr;
            model logSalePrice = RoofMatl MasVnrType Neighborhood BldgType
Exterior1st_gr Exterior2nd_gr ExterQual_gr BsmtQual_gr
              LogGrLivArea LogLotArea LogTotalBsmtSF LogWoodDeckSF
LogTotalOverall LogTotalBath LogTotalBsmtBath GarageArea FirstFlrSF SecondFlrSF
YearRemodAdd YearBuilt
              logTotalExteriorQual logTotalGarageQual logTotalBsmtQual
logTotalBsmtFinType logTotalLandQual logSaleAge
            / selection = lasso(choose=cv stop =AICC) CVDETAILS;
            /*modelaverage tables = (EffectSelectPct(all) ParmEst(All)) alpha =
.05;*/
run;
Quit;
/*Stepwise Model Selection*/

proc glmselect data = train2 plots (stepaxis = number) = (criterionpanel
ASEPlot) seed =1; /*86*/
            partition fraction (test = .3);
            class RoofMatl MasVnrType BldgType Neighborhood MoSold_Chr
MSSubClass_Chr;
            model logSalePrice = RoofMatl MasVnrType Neighborhood BldgType
Exterior1st_gr Exterior2nd_gr ExterQual_gr BsmtQual_gr
              LogGrLivArea LogLotArea LogTotalBsmtSF LogWoodDeckSF
LogTotalOverall LogTotalBath LogTotalBsmtBath GarageArea FirstFlrSF SecondFlrSF
YearRemodAdd YearBuilt
              logTotalExteriorQual logTotalGarageQual logTotalBsmtQual
logTotalBsmtFinType logTotalLandQual logSaleAge
            / selection = stepwise(choose=cv stop =cv) CVDETAILS;
            /*modelaverage tables = (EffectSelectPct(all) ParmEst(All)) alpha =
.05;*/
run;
Quit;


/*Model Comparison run each model 100 times to get the statistics averages*/

proc sql;
      Drop Table Result_1;
      create table Result_1 (Seed float, R_Square float, Adj_R_Sq float, AIC
float, AICC float, SBC float, ASETrain float, ASETest float, CV_PRESS float);
run;
quit;

%macro AICC_1;
%local i;
%do i=1 %to 100;
      ods output FitStatistics=Output ;
      proc glmselect data = train2 seed =&i; /*55*/
            partition fraction (test = .3);
            class RoofMatl MasVnrType Neighborhood BldgType;
            model logSalePrice =
                        RoofMatl MasVnrType Neighborhood BldgType ExterQual_gr
BsmtQual_gr LogGrLivArea LogLotArea LogTotalOverall
```

```
                        LogTotalBath logTotalBsmtBath GarageArea FirstFlrSF
YearRemodAdd YearBuilt logTotalGarageQual logTotalLandQual LogSaleAge
            / selection = lasso(choose=cv stop =None) CVDETAILS ;

      run;
      Quit;

      proc transpose data = output out = new;
      ID Label1 ;
      var nValue1;
      run;
      Proc SQL;
            insert into Result_1 (Seed, R_Square, Adj_R_Sq, AIC, AICC, SBC,
ASETrain, ASETest, CV_PRESS)
            Select &i, R_Square, Adj_R_Sq, AIC, AICC, SBC, ASE__Train_,
ASE__Test_, CV_PRESS
            From new
      Run;
      Quit;

%end;
%mend AICC_1;
%AICC_1;

/*
Proc sort data = Result;
      by AICC;
run;
proc print data =Result;
run;
*/

/*Macro to get the avarage model statistics*/
proc sql;
      Drop Table Result_2;
      create table Result_2 (Seed float, R_Square float, Adj_R_Sq float, AIC
float, AICC float, SBC float, ASETrain float, ASETest float, CV_PRESS float);
run;
quit;

%macro AICC_2;
%local i;
%do i=1 %to 100;
      ods output FitStatistics=Output ;
      proc glmselect data = train2 seed =&i; /*55*/
            partition fraction (test = .3);
            class BldgType Neighborhood;
            model logSalePrice = Neighborhood BldgType BsmtQual_gr LogGrLivArea
LogLotArea LogTotalOverall logTotalBsmtBath
                  GarageArea FirstFlrSF SecondFlrSF YearRemodAdd YearBuilt
LogSaleAge
            / selection = lasso(choose=cv stop =None) CVDETAILS ;

      run;
      Quit;

      proc transpose data = output out = new;
```

```
      ID Label1 ;
      var nValue1;
      run;
      Proc SQL;
            insert into Result_2 (Seed, R_Square, Adj_R_Sq, AIC, AICC, SBC,
ASETrain, ASETest, CV_PRESS)
            Select &i, R_Square, Adj_R_Sq, AIC, AICC, SBC, ASE__Train_,
ASE__Test_, CV_PRESS
            From new
      Run;
      Quit;

%end;
%mend AICC_2;
%AICC_2;



proc sql;
      Drop Table Result_3;
      create table Result_3 (Seed float, R_Square float, Adj_R_Sq float, AIC
float, AICC float, SBC float, ASETrain float, ASETest float, CV_PRESS float);
run;
quit;

%macro AICC_3;
%local i;
%do i=1 %to 500;
      ods output FitStatistics=Output ;
      proc glmselect data = train2 seed =&i; /*55*/
            partition fraction (test = .3);
          class RoofMatl MasVnrType BldgType Neighborhood MoSold_Chr
MSSubClass_Chr;
            model logSalePrice = RoofMatl MasVnrType Neighborhood BldgType
Exterior1st_gr Exterior2nd_gr ExterQual_gr BsmtQual_gr
              LogGrLivArea LogLotArea LogTotalBsmtSF LogWoodDeckSF
LogTotalOverall LogTotalBath LogTotalBsmtBath GarageArea FirstFlrSF SecondFlrSF
YearRemodAdd YearBuilt
              logTotalExteriorQual logTotalGarageQual logTotalBsmtQual
logTotalBsmtFinType logTotalLandQual logSaleAge

            / selection = lasso(choose=cv stop =None) CVDETAILS ;
      run;
      Quit;

      proc transpose data = output out = new;
      ID Label1 ;
      var nValue1;
      run;
      Proc SQL;
            insert into Result_3 (Seed, R_Square, Adj_R_Sq, AIC, AICC, SBC,
ASETrain, ASETest, CV_PRESS)
            Select &i, R_Square, Adj_R_Sq, AIC, AICC, SBC, ASE__Train_,
ASE__Test_, CV_PRESS
            From new
      Run;
      Quit;
```

```
%end;
%mend AICC_3;
%AICC_3;



/*
Proc sort data = Result;
      by AICC;
run;
proc print data =Result;
run;
*/

proc means data=result_1;
title "Statistices Means for Model #1";
run;

proc means data=result_2;
title "Statistices Means for Model #2";
run;

proc means data=result_3;
title "Statistices Means for Model #3";
run;

title;



proc glm data = train2 plots =all;
      class RoofMatl MasVnrType Neighborhood BldgType;
            model logSalePrice =
                        RoofMatl MasVnrType Neighborhood BldgType ExterQual_gr
BsmtQual_gr LogGrLivArea LogLotArea LogTotalOverall
                        LogTotalBath logTotalBsmtBath GarageArea FirstFlrSF
YearRemodAdd YearBuilt logTotalGarageQual logTotalLandQual LogSaleAge
      / CLI CLPARM CLM SOLUTION TOLERANCE CLM ;
output out = results_Laaso p = Predict PRESS=CVPress;

run;
Quit;


proc glm data = train2 plots =all;
      class BldgType Neighborhood;
            model logSalePrice = Neighborhood BldgType BsmtQual_gr LogGrLivArea
LogLotArea LogTotalOverall logTotalBsmtBath
                  GarageArea FirstFlrSF SecondFlrSF YearRemodAdd YearBuilt
LogSaleAge
      / CLI CLPARM CLM SOLUTION TOLERANCE CLM ;
output out = results_Stepwise p = Predict PRESS=CVPress;

run;
Quit;
```

```
proc glm data = train2 plots =all;
      class RoofMatl MasVnrType BldgType Neighborhood MoSold_Chr
MSSubClass_Chr;
            model logSalePrice = RoofMatl MasVnrType Neighborhood BldgType
Exterior1st_gr Exterior2nd_gr ExterQual_gr BsmtQual_gr
              LogGrLivArea LogLotArea LogTotalBsmtSF LogWoodDeckSF
LogTotalOverall LogTotalBath LogTotalBsmtBath GarageArea FirstFlrSF SecondFlrSF
YearRemodAdd YearBuilt
              logTotalExteriorQual logTotalGarageQual logTotalBsmtQual
logTotalBsmtFinType logTotalLandQual logSaleAge
      / CLI CLPARM CLM SOLUTION TOLERANCE CLM ;
output out = results_Best p = Predict PRESS=CVPress;

run;
Quit;


data results_Laaso2;
set results_Laaso;
if Predict > 0 then logSalePrice = Predict;
if Predict < 0 then logSalePrice = log(180921.2);
SalePrice=exp(logSalePrice);
keep id SalePrice;
where id > 1460;
;

proc export data=results_Laaso2
outfile='C:\Users\yubin\OneDrive\MyWork\SMU\MSDS6372\Unit6Project\submit_Lasso.
csv' dbms=csv Replace;
run;


data results_Stepwise2;
set results_Stepwise;
if Predict > 0 then logSalePrice = Predict;
if Predict < 0 then logSalePrice = log(180921.2);
SalePrice=exp(logSalePrice);
keep id SalePrice;
where id > 1460;
;

proc export data=results_Stepwise2
outfile='C:\Users\yubin\OneDrive\MyWork\SMU\MSDS6372\Unit6Project\submit_Stepwi
se.csv' dbms=csv Replace;
run;

data results_Best2;
set results_Best;
if Predict > 0 then logSalePrice = Predict;
if Predict < 0 then logSalePrice = log(180921.2);
SalePrice=exp(logSalePrice);
keep id SalePrice;
where id > 1460;
;
```

```
proc export data=results_Best2
outfile='C:\Users\yubin\OneDrive\MyWork\SMU\MSDS6372\Unit6Project\submit_Best.c
sv' dbms=csv Replace;
run;
```

Appendix V

LASSO: used stop at AICC parameter and seed =55. Looking at Fig.2.3, it stopped at step 27.

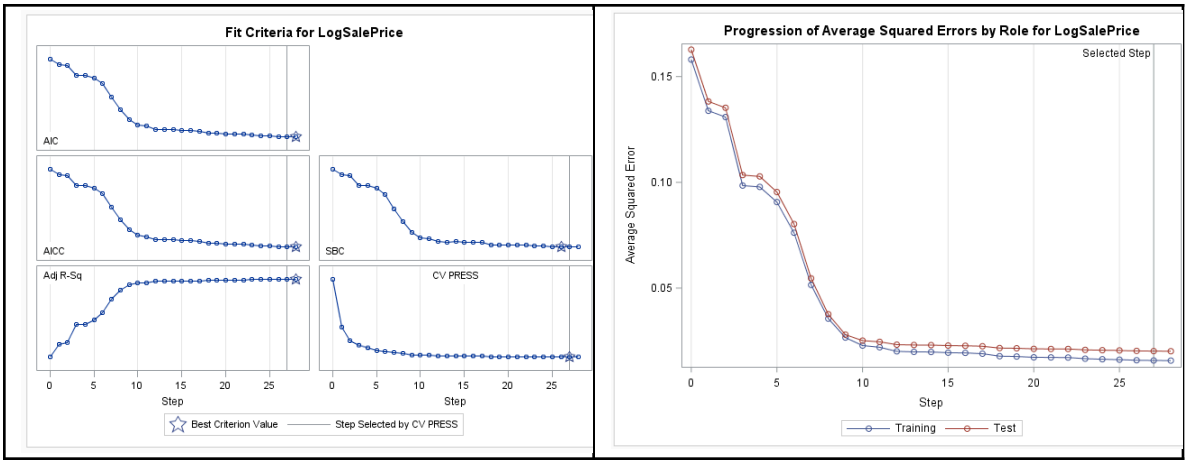


Fig.2.3 LASSO Model Selection Results.

Stepwise: Used stop at cv parameter and seed=1. Looking at Fig.2.4, it stopped at step 13.

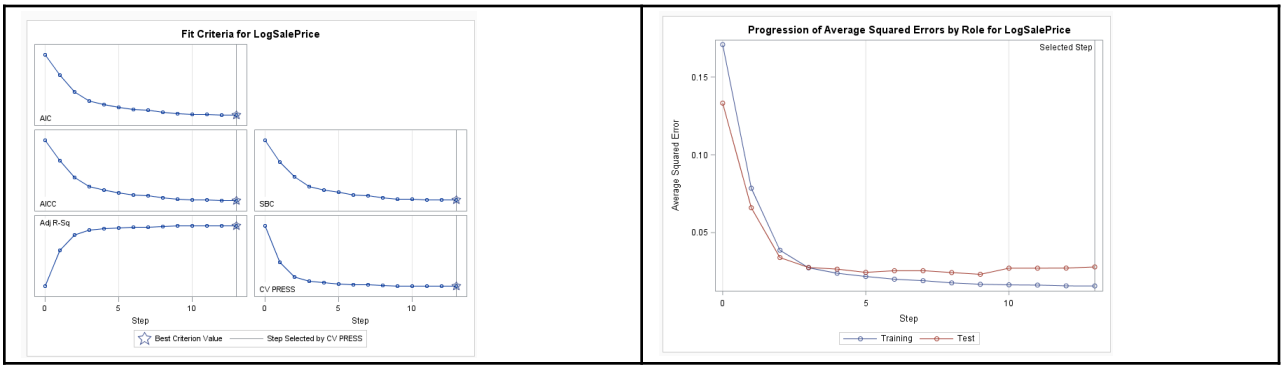


Fig.2.4 Stepwise Model Selection Results

Appendix VI

From the Cook's D and residual plot, you can see observation 524 is the influential point. We may want to remove it from the model.

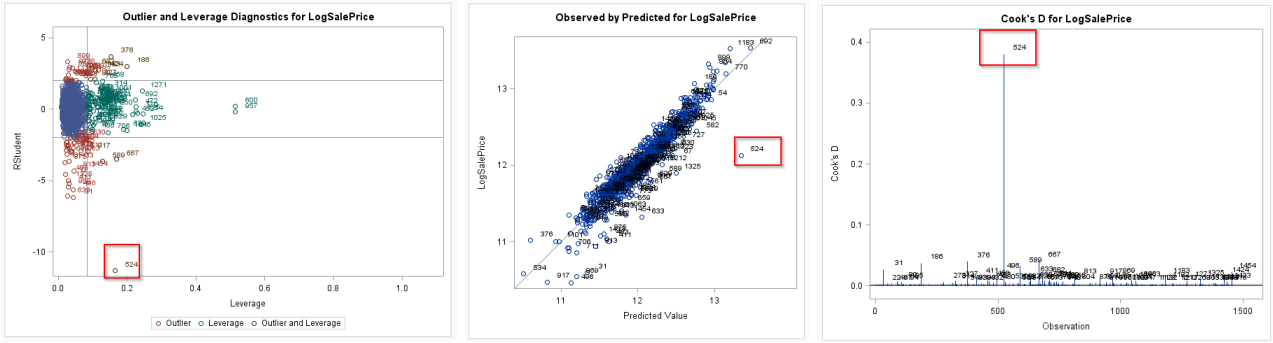


Fig.2.4 Cook's D and Influential Point Analysis

Here is what looks like after removed 524 observations.

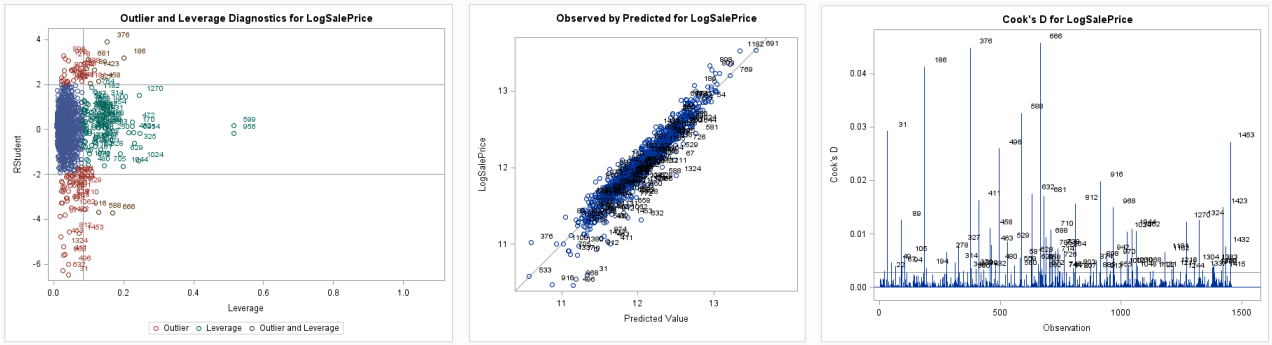


Fig.2.5 Cook's D and Influential Point Analysis After Removing Influential Point