Nuoya Rezsonya & Alexandra Norman

Housing Price Prediction Project

1. Introduction and data description

We are challenged with tough real estate problem, sale price prediction. Wouldn't it be wonderful if we can have a prediction tool to tell agents and buyers how much they should budget for their dream property? This is age old problem, no one has been able to solve it well. Most transactions are based on agent's experience, not data analysis. In recent years, computing chips and storage technology have advanced more than 10-fold, Data Science has become one of the new era. In this report, we will perform analysis based on Kaggle (https://www.kaggle.com/c/house-prices-advanced-regression-techniques) House Price challenge.

The data set describes the sale of individual residential property in Iowa from 2006 to 2010. There are two data sets presented, both large data sets. The training data set contains final sale price, it is to be used for the modelling section and the test data set only contains explanatory variable. We will predict the sale price of the test data based on the model we have built using training data set. The training data set contains 1460 observations and 80 variables. The test data set contains 1459 observations and 79 variables (no sale price).

2. Preliminary analysis on raw data

To perform multiple linear regression, assumptions behind regression (normality, linearity and constant variance assumptions) must be met. Before we start to do any statistical analysis on the data, we come up with a scatter plot to see if there is any obvious linear relationship between the final sale price and the square footage of the living area. By observing the scatter plot, one can have an overall knowledge of the existance of potential outliers. One can also decide whether transformation is needed. The scatter plot on original data doesn't show the linearity assumption is met. Therefore, transformation is needed. Logarithm transformation is performed on both price and living area. Besides the potential outliers, the data after transformation shows more evidence of linearity.

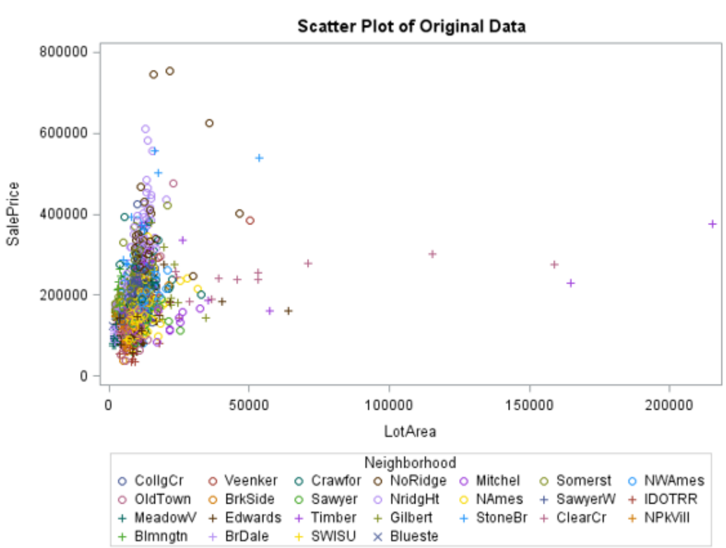


Figure 1: the scatter plot of on original data

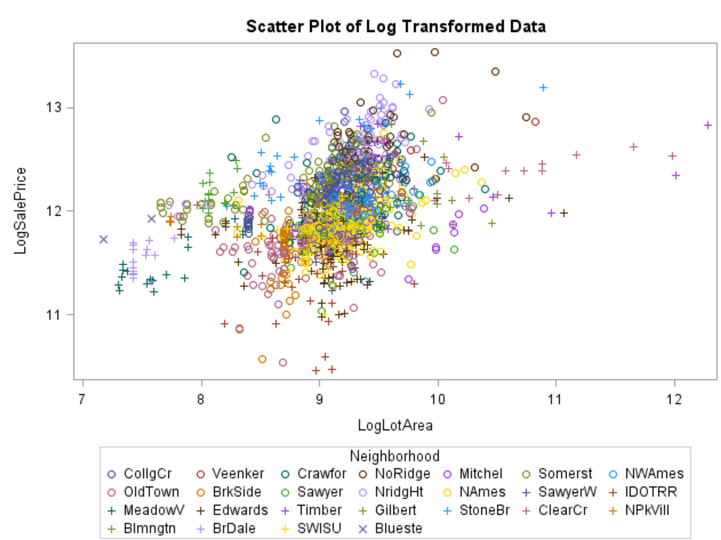


Figure 2: the scatter plot of on log transformed data

3. Analysis One

3.1 Restatement of the Problem

Real estate agents, contractors and prospective buyers are interested in knowing what characteristics of a house drives sales prices in Ames Iowa. We are challenged with creating a model that is easily interpreted that can help real estate agents, contractors and buyers with those insights.

3.2 Outlier removal

Multiple regression of all variables is performed on the transformed data and we check the residual plot to make sure that regression assumptions are met. The residual plot is on the right-hand side (Figure 3: residual plot). There are spikes on Cook's D plot, but the values are small. The normality, linearity and constant SD assumptions are almost met. We choose to keep all the current data points and go ahead perform variable and model selection process.

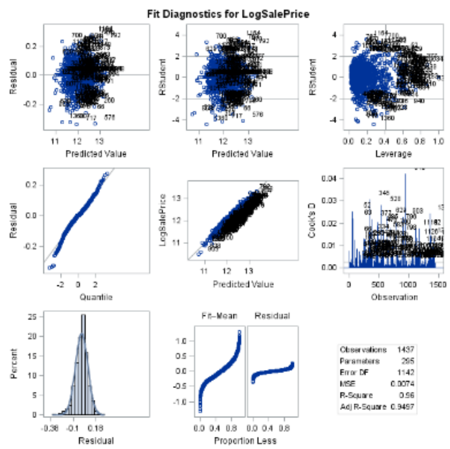


Figure 3: residual plot

3.3 Variable selection

Since the question of interest is what are important factors of a house that could help give insight on what drives sales price in Ames, Iowa and this model should be formed to facilitate the easy interpretation of parameters for use. We use different variable selection techniques with the same criterion first (the details of variable selection are as below), then we picked one categorical variable and two continuous variables from the results to do multiple regression on the log transformed data. We will only talk about the best model we achieved—the LASSO model, but fit statistics from other models will be shown in Appendix:

Selection method	Choose criterion
LASSO	CV PRESS
Forward selection	CV PRESS
Stepwise selection	CV PRESS

3.3 Model construction:

Model equation is: $\hat{\mu}\{\text{LogSalePrice}\} = 9.630643100 + 0.195123495 \times \text{LogLotArea} + \beta_c \times \text{CentralAir} + 0.289104149 \times \text{Bathrooms}$.

After encoding all the categorical variables in this model, we then check the variance inflation factor to make sure there is no multilinearity problem in the model. The result is as shown (Figure 4: VIF check).

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	B	9.63064	0.12279	78.43	<.0001	0
Col2	LogLotArea	1	0.19512	0.01378	14.16	<.0001	1.05927
Col3	CentralAir N	B	-0.36007	0.02868	-12.56	<.0001	1.04250
Col4	CentralAir Y	0	0	.	.	.	.
Col5	Bathrooms	1	0.28910	0.00926	31.24	<.0001	1.10049

Figure 4: VIF check

3.4 Assumption checking by residual plots:

- Normality: Judging from scatter plot and histogram of residuals, the date set looks fit well for normality. The qq plot does show minor deviation, but not strong evidence against normality.
- Linear Trend: The scatter plot (Figure 2: the scatter plot of log transformed data) indicates a strong linear trend between each log(LotArea) and log(SalePrice).
- Equal SD: There is little evidence from the scatter plots of heteroscedasticity.
- Independence: We will assume sale price of houses are independent.
- Influential points check: Cook's D does show spikes, but the Cook's Ds are all small, we will proceed with caution.

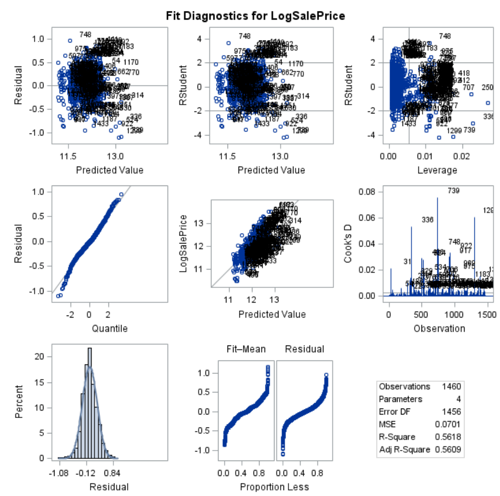


Figure 5: diagnostic plots

3.4 Model and parameter Interpretation

3.4.1 Model equation

In this model, we have one categorical variables: Central Air(Yes/No). The regression result from SAS is as shown on the right-hand side (Figure 6: regression result from SAS).

The model equation is: $\hat{\mu}$\{LogSalePrice\} = 9.630643100+ 0.195123495 ×LogLotArea+ + β_c×CentralAir +0.289104149×Bathrooms.

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	9.630643100	B	0.12278996	78.43	<.0001	9.389778974	9.871507227
LogLotArea	0.195123495		0.01378268	14.16	<.0001	0.168087460	0.222159531
CentralAir N	-0.360069611	B	0.02867552	-12.56	<.0001	-0.416319363	-0.303819858
CentralAir Y	0.000000000	B	.	.	.	.	.
Bathrooms	0.289104149		0.00925551	31.24	<.0001	0.270948598	0.307259700

Figure 6: regression result from SAS

3.4.2 Parameter interpretation

- For houses which have central air (CentralAir_Y, variable CentralAir =0):
  1. Intercept =9.630643100: The predicted median of logged sales price of a house with zero living space regardless of central air and fireplace quality is 9.630643100, which is not practical.
  2. Parameter of Log(LotArea): If variable Bathrooms is fixed, a doubling of the above lot area equates to the multiplicative change of 1.144822 (2^0.195123495). A doubling of the above ground living area equates to an increase of 14.4822% in the estimated medium of the sale price. A 95% confident interval for the parameter is (0.1680,

0.2222). Therefore a 95 % confidence interval for medium increase in house sale price rate after a doubling of the above ground living area is (56.175% to 58.326%).

3. Parameter of Bathrooms: If variable Log(LotArea) is fixed, a doubling of the above Bathrooms equates to the multiplicative change of 1.221881(2^0.289104149). A doubling of the above ground living area equates to an increase of 22.1881% in the estimated medium of the sale price. A 95% confident interval is (0.2709, 0.3073). Therefore a 95 % confidence interval for medium increase in house sale price rate after a doubling of the above ground living area is (60.33% to 61.87%).

- For houses which have no central air (CentralAir_N , variable CentralAir =1):
  1. Intercept =9.630643100-0.360069611 =9.270573: The predicted median of logged sales price of a house with zero living space regardless of central air and fireplace quality is 9.630643100, which is not practical.
  2. Parameter of Log(LotArea): If variable Bathrooms is fixed, a doubling of the above lot area equates to the multiplicative change of 1.144822 (2^0.195123495). A doubling of the above ground living area equates to an increase of 14.4822% in the estimated medium of the sale price. A 95% confident interval for the parameter is (0.1680, 0.2222). Therefore a 95 % confidence interval for medium increase in house sale price rate after a doubling of the above ground living area is (56.175% to 58.326%).
  3. Parameter of Bathrooms: If variable Log(LotArea) is fixed, a doubling of the above Bathrooms equates to the multiplicative change of 1.221881(2^0.289104149). A doubling of the above ground living area equates to an increase of 22.1881% in the estimated medium of the sale price. A 95% confident interval is (0.2709, 0.3073). Therefore a 95 % confidence interval for medium increase in house sale price rate after a doubling of the above ground living area is (60.33% to 61.87%).

4. Analysis Two

4.1 Restatement of the Problem

We were tasked with the challenge to predict what factors impact the future sales prices on houses in Ames, Iowa. We are to build the most predictive model for sales prices of homes in all of Ames Iowa using only the knowledge we have from our studies so far.

4.2 Model Selection

Since question of interest is how what factors are the most predictive in figuring out the sales price of houses we used different variable selection techniques with the same criterion first (the details of variable selection are as below), then we keep all selected variables from different methods.

Selection method	Choose criterion
LASSO	CV PRESS
Forward selection	CV PRESS
Custom	CV PRESS

To achieve more sophisticated models, we checked for multicollinearity and removed any columns that had a high VIF before doing the variable selection. We dropped the following columns due to multicollinearity: MSSubClass, LandSlope, YearBuilt, Foundation BsmtExposure, GrLivArea, GarageType, GarageYrBlt. For the LASSO and FORWARD selection model we will use the same parameter inputs (see below) and use a partition of 0.5.

Parameters Selected for LASSO and Forward as Inputs:
MSZoning, LotFrontage, Street, LotShape, LandContour, Utilities, LotConfig, Neighborhood, BldgType, HouseStyle, OverallQual, OverallCond, YearRemodAdd, MasVnrType, MasVnrArea, ExterQual, ExterCond, BsmtQual, BsmtCond, BsmtFinType1, BsmtFinType2, TotalBsmtSF, Heating, HeatingQC, CentralAir, Electrical, BedroomAbvGr, KitchenAbvGr, KitchenQual, TotRmsAbvGrd, Functional, Fireplaces, FireplaceQu, GarageFinish, GarageCars, GarageArea, GarageQual, GarageCond, PavedDrive, PoolArea, MiscVal, MoSold, YrSold, SaleType, SaleCondition, Bathrooms, PorchSF, TotalSF, Exterior, Condition, Roof, LogLotArea, Neighborhood*LogLotArea, Neighborhood*OverallCond

LASSO Model

For the LASSO model we used the parameters in the box above, CV as the criterion and did a 50|50 partitioning on the train data to figure out what variables to use on predicting sales prices. This model picked the following variables as factors that help predict the sales price of houses: MSZoning, OverallQual, OverallCond, YearRemodAdd, MasVnrArea, BsmtQual, BsmtFinType1, HeatingQC, CentralAir, Fireplaces, GarageCars, GarageArea, GarageCond, PavedDrive, SaleCondition, Bathrooms, TotalSF, Exterior, LogLotArea

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	104.86385	8.73865	407.48	<.0001
Error	712	15.26932	0.02145		
Corrected Total	724	120.13317			

Root MSE	0.14644
Dependent Mean	12.02676
R-Square	0.8729
Adj R-Sq	0.8708
AIC	-2045.73622
AICC	-2045.14467
BIC	-2786.84821
C(p)	1080.89118
SBC	-2713.11599
ASE (Train)	0.02106
ASE (Test)	0.02151
CV PRESS	12.45352

Cross Validation Details			
	Observations		
Index	Fitted	Left Out	CV PRESS
1	571	154	2.5277
2	592	133	2.0422
3	584	141	2.7715
4	574	151	3.4606
5	579	146	1.6515
Total			12.4535

Figure 7: LASSO model results.

$$\log \text{SalePrice} = \beta_0 + \beta_1\text{MSZoning} + \beta_2\text{OverallQual} + \beta_3\text{OverallCond} + \beta_4\text{YearRemodAdd} + \beta_5\text{MasVnrArea} + \beta_6\text{BsmtQual} + \beta_7\text{BsmtFinType1} + \beta_8\text{HeatingQC} + \beta_9 CentralAir + \beta_{10}\text{Fireplaces} + \beta_{11}\text{GarageCars} + \beta_{12}\text{GarageArea} + \beta_{13}\text{GarageCond} + \beta_{14}\text{PavedDrive} + \beta_{15}\text{SaleCondition} + \beta_{16} Bathrooms + \beta_{17}\text{TotalSF} + \beta_{18} Exterior + \beta_{19} LogLotArea$$

Forward Selection

For the forward selection we used the parameters in the box above, CV as the criterion and did a 50|50 partitioning on the train data to figure out what variables to use on predicting sales prices. This model picked the following variables as factors that help predict the sales price of houses: MSZoning, OverallQual, YearRemodAdd, BsmtQual, Heating, KitchenAbvGr, KitchenQual, Functional, Fireplaces, GarageCars, GarageQual, Bathrooms, TotalSF, LogLotArea, OverallCond*Neighborhood.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	25	113.30860	4.53234	338.44	<.0001
Error	702	9.40111	0.01339		
Corrected Total	727	122.70971			

Root MSE	0.11572
Dependent Mean	12.02710
R-Square	0.9234
Adj R-Sq	0.9207
AIC	-2384.41658
AICC	-2382.25658
BIC	-3125.17094
C(p)	285.69719
PRESS	10.65351
SBC	-2995.06875
ASE (Train)	0.01291
ASE (Test)	0.01542
CV PRESS	10.59053

Cross Validation Details			
	Observations		
Index	Fitted	Left Out	CV PRESS
1	589	139	2.0770
2	576	152	2.3519
3	580	148	2.9495
4	596	132	1.4970
5	571	157	1.7150
Total			10.5905

Figure 8: Forward Selection model results.

$$\log \text{SalePrice} = \beta_0 + \beta_1\text{MSZoning} + \beta_2\text{OverallQual} + \beta_3\text{YearRemodAdd} + \beta_4\text{BsmtQual} + \beta_5\text{Heating} + \beta_6\text{KitchenAbvGr} + \beta_7\text{KitchenQual} + \beta_8\text{Functional} + \beta_9\text{Fireplaces} + \beta_{10}\text{GarageCars} + \beta_{11}\text{GarageQual} + \beta_{12}\text{Bathrooms} + \beta_{13}\text{TotalSF} + \beta_{14}\text{LogLotArea} + \beta_{15}\text{OverallCond} * \text{Neighborhood}$$

Custom Selection

For the custom model we used a combination of variables that were found significant for the forward selection, LASSO and elastic net all with a partition of 50. We ran each one 3 times to see what variables were found multiple times and included them in the custom model. After the variable selection we made sure that all 19 variables listed below were included in the model: MSZoning, OverallQual, YearRemodAdd, BsmtQual, HeatingQC, CentralAir, KitchenQual, GarageCars, GarageArea, Bathrooms, TotalSF, LogLotArea, Fireplaces, OverallCond*Neighborhood, ExterCond, Functional, SaleCondition, SaleType, PorchSF.

Root MSE	0.10887
Dependent Mean	12.02856
R-Square	0.9303
Adj R-Sq	0.9213
AIC	-1890.82347
AICC	-1872.63584
BIC	-2442.53782
C(p)	66.00000
PRESS	8.76952
SBC	-2175.12736
ASE (Train)	0.01048
ASE (Test)	0.02658

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	65	79.61900	1.22491	103.35	<.0001
Error	503	5.96146	0.01185		
Corrected Total	568	85.58046			

$\log \text{SalePrice} = \beta_0 + \beta_1\text{MSZoning} + \beta_2\text{OverallQual} + \beta_3\text{YearRemodAdd} + \beta_4\text{BsmtQual} + \beta_5\text{HeatingQC} + \beta_6\text{CentralAir} + \beta_7\text{KitchenQual} + \beta_8\text{GarageCars} + \beta_9\text{GarageArea} + \beta_{10}\text{Bathrooms} + \beta_{11}\text{TotalSF} + \beta_{12}\text{LogLotArea} + \beta_{13}\text{Fireplaces} + \beta_{14}\text{OverallCond} * \text{Neighborhood} + \beta_{15}\text{ExterCond} + \beta_{16}\text{Functional} + \beta_{17}\text{SaleCondition} + \beta_{18}\text{SaleType} + \beta_{19}\text{PorchSF}$

4.3 Assumption checking by residual plots

4.3.1 LASSO

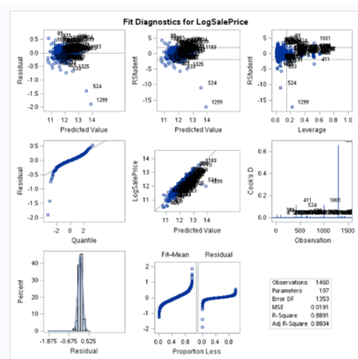
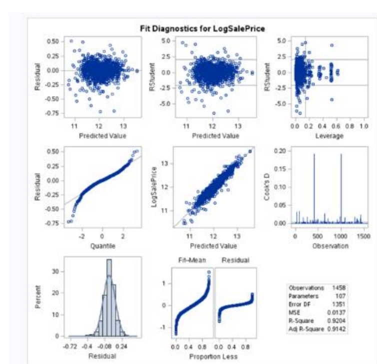


Figure 10: residual plot with outliers

- Normality: Judging from scatter plot and histogram of residuals, the data set looks like it could be slightly left skewed but still looks to fit normality fairly close. The qq plot does show some deviation, but this could be due to the few outliers and not strong evidence against normality. After removing the outliers, the qq plot has less deviation so there is not strong evidence against normality.



- Linear Trend: The scatter plots indicate a strong linear trend between selected variables and log(SalePrice).
- Equal SD: There is little evidence from the scatter plots of heteroscedasticity.
- Independence: We will assume the observations are independent.
- Influential points check: Cook's D does show spikes (figure 10) at ID 524 and 1299 we will remove and recheck plots. After removing the influential points, there are spikes on Cook's D plot,

but the values are small. The normality, linearity and constant SD assumptions are almost met. We choose to keep all the current data points.

Figure 11: residual plot with outliers removed

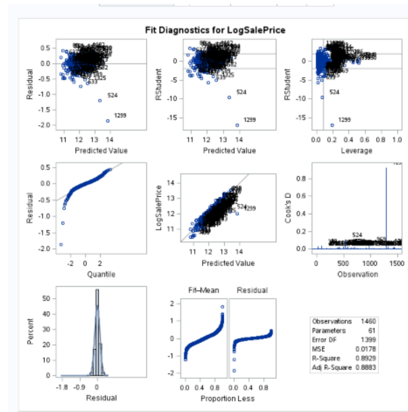


Figure 12: residual plot with outliers

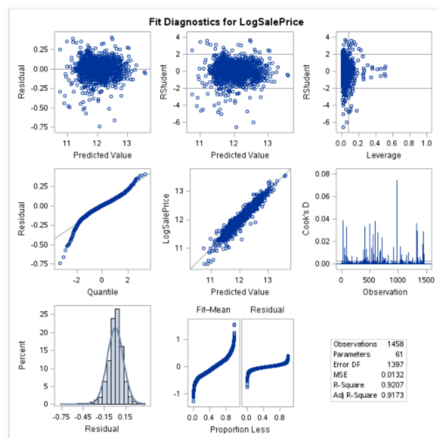


Figure 13: residual plot without outliers

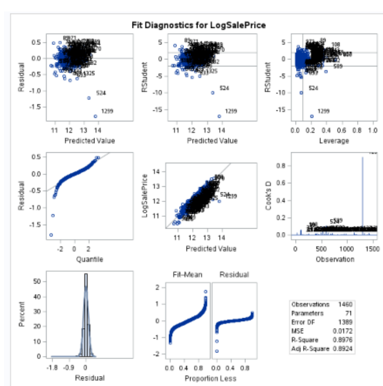


Figure 14: residual plot with outliers

4.3.2 Forward Selection

- Normality: Judging from scatter plot and histogram of residuals, the data set looks like it could be slightly left skewed but overall looks to closely fit normality. The qq plot does show some deviation, but this could be due to the few outliers and not strong evidence against normality. After removing the outliers, the qq plot has less deviation so there is not strong evidence against normality.
- Linear Trend: The scatter plots indicate a strong linear trend between selected variables and log(SalePrice).
- Equal SD: There is little evidence from the scatter plots of heteroscedasticity.
- Independence: We will assume the observations are independent.
- Influential points check: Cook's D does show spikes (figure 12) at ID 524 and 1299 similar to what we saw for the LASSO model, we will remove and recheck plots. After removing the influential points, there are still spikes on Cook's D plot, but the values are small and when attempted to remove saw no change. The normality, linearity and constant SD assumptions are almost met. We choose to keep all the current data points after removing the first two influential points.

4.3.3 Custom Model

- Normality: Judging from scatter plot and histogram of residuals, the data set looks like it could be slightly left skewed but still looks to fit normality besides the outliers. The qq plot does show some deviation, but this could be due to the few outliers and not strong evidence against normality. After removing the outliers, the qq plot has less deviation so there is not strong evidence against normality.

- Linear Trend: The scatter plots indicate a strong linear trend between selected variables and log(SalePrice).
- Equal SD: There is little evidence from the scatter plots of heteroscedasticity.
- Independence: We will assume the observations are independent.
- Influential points check: Cook's D does show spikes (figure 10) at ID 524 and 1299 we will remove and recheck plots. After removing the influential points, there are spikes on Cook's D plot, but the values are small. The normality, linearity and constant SD assumptions are almost met. We choose to keep all the current data points.

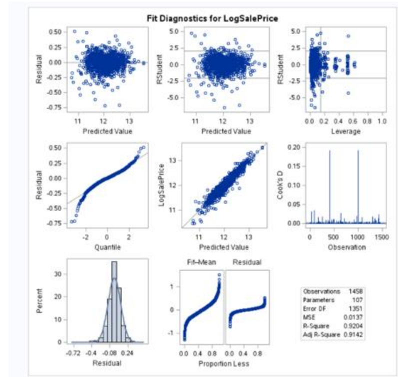


Figure 15: residual plot with outliers removed

4.4 Comparing competing models

Test Set Models	Adjusted R2	AIC	BIC	ASE (Test)	CV Press	Kaggle Score
Forward	0.9207	-2384.1658	-3125.17094	0.01542	10.59053	0.13732
LASSO	0.8708	-2045.73622	-2786.84821	0.02151	12.45352	0.14161
Custom	0.9213	-1890.82347	-2442.53782	0.02658	8.76952	0.17406

4.5 Conclusion:

This is an observation study; no inference can be draw for general housing market price prediction. The result only applies to sale prices on houses in Ames Iowa, however we are able to gleam some insights for the real estate agents, contractors and prospective buyers on what characteristics might be driving sales prices in Ames. Some of the additional insight that we did get after doing a few different models is that there does seem to be at least to influential points (ID = 524 and ID=1299) in this dataset that needed to be removed. From Analysis One we are able to say that LogLotArea, Bathrooms and Central Air can help give insight into the LogSalesPrices of houses in Ames Iowa. Those three variables explain 56.1% of variation in sales price of houses in Ames Iowa. The variables found within the forward selection model were all found to be significantly statistic when trying to predict sales prices of houses in Ames Iowa. Although there are some influential points in the model the custom model seems to be a good fit for this data set to predict sales prices. About r2 = 92.3% of the variation in sales price of houses in Ames Iowa is explained by the variables in the forward model. Leaving 7.7% for the other factors combined.

Extra

The two categorical variables we picked are PavedDrive and CentralAir. PavedDrive is a categorical variable that has three levels while CentralAire has two levels. We tested out both additive and none additive models and provide confidence intervals of different combinations of effects.

Additive model

LogSalePrice=12.17814003 + β_1 ×PavedDrive + β_2×CentralAir

Dependent Variable: LogSalePrice

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	35.9318643	11.9772881	90.82	<.0001
Error	1450	191.2326081	0.1318846		
Corrected Total	1453	227.1644724			

R-Square	Coeff Var	Root MSE	LogSalePrice Mean
0.158176	3.019944	0.363159	12.02536

Source	DF	Type I SS	Mean Square	F Value	Pr > F
PavedDrive	2	20.92592810	10.46296405	79.33	<.0001
CentralAir	1	15.00593622	15.00593622	113.78	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
PavedDrive	2	8.81892676	4.40946338	33.43	<.0001
CentralAir	1	15.00593622	15.00593622	113.78	<.0001

From the overall ANOVA table, we can see that the F value is 90.82 with the corresponding p-value smaller than 0.0001. At a significance level of α = 0.05, we will have to reject the null hypothesis of this overall ANOVA test and conclude that there is house price difference caused by different paved driveway and the central air availability. Based on the type I and type III table, we also find that the effect of PavedDrive and the effect of CentralAir are both statistically significant.

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	12.07814003	B	0.01007471	1198.86	<.0001	12.05837746	12.09790260
PavedDrive N	-0.32315422	B	0.04221513	-7.65	<.0001	-0.40596348	-0.24034497
PavedDrive P	-0.22832838	B	0.06737300	-3.39	0.0007	-0.36048735	-0.09616940
PavedDrive Y	0.00000000	B	.	.	.	.	.
CentralAir N	-0.44229672	B	0.04146476	-10.67	<.0001	-0.52363405	-0.36095939
CentralAir Y	0.00000000	B	.	.	.	.	.

We are using a house that has central air and paved driveway Y as the reference level. From the parameter estimate table, all the parameters are statistically significant. The confidence intervals of the house prices with different combinations of effects are as below.

Central Air =Y, PavedDrive=Y

Non-additive model

LogSalePrice=12.08069948 + β_1 ×PavedDrive + β_2×CentralAir + β_3×(CentralAir×PavedDrive)

Dependent Variable: LogSalePrice

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	36.7424780	7.3484956	55.88	<.0001
Error	1448	190.4219944	0.1315069		
Corrected Total	1453	227.1644724			

R-Square	Coeff Var	Root MSE	LogSalePrice Mean
0.161744	3.015617	0.362639	12.02536

Source	DF	Type I SS	Mean Square	F Value	Pr > F
PavedDrive	2	20.92592810	10.46296405	79.56	<.0001
CentralAir	1	15.00593622	15.00593622	114.11	<.0001
CentralAi*PavedDrive	2	0.81061373	0.40530687	3.08	0.0462

Source	DF	Type III SS	Mean Square	F Value	Pr > F
PavedDrive	2	4.78283582	2.39141791	18.18	<.0001
CentralAir	1	3.64023513	3.64023513	27.68	<.0001
CentralAi*PavedDrive	2	0.81061373	0.40530687	3.08	0.0462

From the overall ANOVA table, we can see that the F value is 55.88 with the corresponding p-value smaller than 0.0001. At a significance level of α = 0.05, we will have to reject the null hypothesis of this overall ANOVA test and conclude that there is house price difference caused by different paved driveway, the central air availability and the interaction of them. Based on the type I and type III table, we also find that the effect of PavedDrive and the effect of CentralAir are both statistically significant while the interaction of them is on the edge of being not statistically significant.

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	12.08069948	B	0.01012815	1192.78	<.0001	12.06083205	12.10056690
PavedDrive N	-0.35894678	B	0.04993610	-7.19	<.0001	-0.45690162	-0.26099195
PavedDrive P	-0.29144567	B	0.07471301	-3.90	0.0001	-0.43800298	-0.14488836
PavedDrive Y	0.00000000	B	.	.	.	.	.
CentralAir N	-0.50676578	B	0.05083150	-9.97	<.0001	-0.60647705	-0.40705452
CentralAir Y	0.00000000	B	.	.	.	.	.
CentralAi*PavedDrive N N	0.15146163	B	0.09403577	1.61	0.1075	-0.03299929	0.33592254
CentralAi*PavedDrive N P	0.36725831	B	0.17315055	2.12	0.0341	0.02760556	0.70691106
CentralAi*PavedDrive N Y	0.00000000	B	.	.	.	.	.
CentralAi*PavedDrive Y N	0.00000000	B	.	.	.	.	.
CentralAi*PavedDrive Y P	0.00000000	B	.	.	.	.	.
CentralAi*PavedDrive Y Y	0.00000000	B	.	.	.	.	.

We are using a house that has central air and paved driveway Y as the reference level. From the parameter estimate table, except the interaction of no central air and paved driveway N and the interaction of no central air and paved driveway P, the rest of parameters are statistically significant. The confidence intervals of the house prices with different combinations of effects are as below.

Central Air =Y, PavedDrive=Y

1. When a house has central air and with a paved driveway Y, the house price is \$175982.5 ($e^{12.07814003}$) with a 95% confidence interval at ($e^{12.0584}$, $e^{12.0979}$) which is (\$172542.7, \$179494.5). This combination of effect is statistically significant with a p-value < 0.0001.	1. When a house has central air and with a paved driveway Y, the house price is \$176433.5 ($e^{12.08069948}$) with a 95% confidence interval at ($e^{12.0608}$, $e^{12.1006}$) which is (\$172957.3, \$179979.8). This combination of effect is statistically significant with a p-value < 0.0001.
Central Air =Y, PavedDrive=N	Central Air =Y, PavedDrive=N
2. When a house has central air and with a paved driveway N, the house price is \$127387.1 ($e^{12.07814003-0.32315422}$) with a 95% confidence interval at ($e^{(12.0584-0.4060)}=e^{11.6979}$, $e^{(12.0979-0.2403)}=e^{12.0017}$) which is (\$114967, \$141153). This combination of effect is statistically significant with a p-value < 0.0001.	2. When a house has central air and with a paved driveway N, the house price is \$123223.2 ($e^{12.08069948-0.35894678}$) with a 95% confidence interval at ($e^{(12.0608-0.4569)}=e^{11.6039}$,$e^{(12.1006-0.2610)}=e^{11.8396}$) which is (\$109524.1, \$138635). This combination of effect is statistically significant with a p-value < 0.0001.
Central Air =Y, PavedDrive=P	Central Air =Y, PavedDrive=P
3. When a house has central air and with a paved driveway P, the house price is \$140058 ($e^{12.07814003-0.22832838}$) with a 95% confidence interval at ($e^{(12.0584-0.3605)}=e^{11.6979}$,$e^{(12.0979-0.0962)}=e^{12.0017}$) which is (\$120318.8, \$163031.7). This combination of effect is statistically significant with a p-value < 0.0001.	3. When a house has central air and with a paved driveway P, the house price is \$131828.1 ($e^{12.08069948-0.29144567}$) with a 95% confidence interval at ($e^{(12.0608-0.4380)}=e^{11.6228}$, $e^{(12.1006-0.14489)}=e^{11.95571}$) which is (\$111613.8, \$155703.7). This combination of effect is statistically significant with a p-value < 0.0001.
Central Air =N, PavedDrive=Y	Central Air =N, PavedDrive=Y
4. When a house has no central air and with a paved driveway Y, the house price is \$113079.2 ($e^{12.07814003-0.44229672}$) with a 95% confidence interval at ($e^{(12.0584-0.5236)}=e^{11.5348}$, $e^{(12.0979-0.3610)}=e^{11.7369}$) which is (\$102211.6, \$125103.9). This combination of effect is statistically significant with a p-value < 0.0001.	4. When a house has no central air and with a paved driveway Y, the house price is \$106290.8 ($e^{12.08069948-0.50676578}$) with a 95% confidence interval at ($e^{(12.0608-0.60647705)}=e^{11.45432}$,$e^{(12.1006-0.40705452)}=e^{11.69355}$) which is (\$94307.88, \$119796.5). This combination of effect is statistically significant with a p-value < 0.0001.
Central Air =N, PavedDrive=N	Central Air =N, PavedDrive=N
5. When a house has no central air and with a paved driveway N, the house price is \$81853.72 ($e^{12.07814003-0.44229672-0.32315422}$) with a 95% confidence interval at ($e^{(12.0584-0.4060-0.5236)}=e^{11.1288}$, $e^{(12.0979-0.2403-0.3610)}=e^{11.4966}$) which is (\$68104.6, \$98380.71). This combination of effect is statistically significant with a p-value < 0.0001.	5. When a house has no central air and with a paved driveway N, the house price is \$86374.57 ($e^{12.08069948-0.35894678-0.50676578+0.15146163}$) with a 95% confidence interval at ($e^{(12.0608-0.4569-0.60647705-0.03299929)}=e^{10.96442}$,$e^{(12.1006-0.2610-0.40705452+0.33592254)}=e^{11.76847}$) which is (\$57781.27, \$129116.5). This combination of effect is not statistically significant with a p-value ≈ 0.336.
Central Air =N, PavedDrive=P	Central Air =N, PavedDrive=P
6. When a house has no central air and with a paved driveway P, the house price is \$99459.92($e^{12.07814003-0.22832838-0.44229672}$) with a	6. When a house has no central air and with a paved driveway P, the house price is \$114662.3 ($e^{12.08069948-0.29144567-0.50676578+0.36725831}$) with a 95%

95% confidence interval at ($e^{(12.0584-0.3605-0.5236)}$ $=e^{11.1743}$, $e^{(12.0979-0.0962-0.3610)}=e^{11.6407}$) which is (\$71274.93, \$113629.7). This combination of effect is statistically significant with a p-value < 0.0001.	confidence interval at ($e^{(12.0608-0.43800298-0.60647705+0.02760556)}$ $=e^{11.04393}$, $e^{(12.1006-0.14488836-0.40705452+0.70691106)}=e^{12.2557}$) which is (\$62563.04, \$210148.6). This combination of effect is not statistically significant with a p-value ≈ 0.707.
Conclusion	Conclusion
From the overall variance analysis table, we can see that at a significance level of α = 0.05 with a p-value is smaller than 0.0001, we will have to reject the null hypothesis of this overall ANOVA test and conclude that different paved driveway and the central air availability have effect on the house sale price. This is an observation study. No causal inferences can be drawn. The result can only apply to this sample data. From above, one can say that different combinations of type of paved driveway and central air availability can affect the house price.	From the overall ANOVA table, we can see that at significance level of α = 0.05 with a p-value is smaller than 0.0001, we will have to reject the null hypothesis of this overall ANOVA test and conclude that there is house price difference caused by different paved driveway, the central air availability and the interaction of them. However, when we break down to each combination of effect, there is no evidence showing that the interaction of different paved driveway and central air availability has effects on the house prices. This is an observation study. No causal inferences can be drawn. The result can only apply to this sample data. From above, one can say that different combinations of type of paved driveway and central air availability can affect the house price, but not the interaction of them.

4. Appendix

1. Data cleansing in R

```
#load the libraries
library(ggplot2)
library(olsrr)
library(car)
library(caret)

#read the training data
training <- read.csv('train1.csv', stringsAsFactors=FALSE)
#clean the neighborhood in training first
training$Neighborhood[grep(training$Neighborhood,pattern = "-1mes")] <- "NAmes"
#training$Neighborhood[grep(training$Neighborhood,pattern = "NWAmes")] <- "NAmes"

#correct to character to factor
character_vars <- lapply(training, class) == "character"
training[, character_vars] <- lapply(training[, character_vars], as.factor)

###########################   check  values   ###########################
# check NA values and there is no NA values
colSums(is.na(training))

# this step is to be conservative, na numeric to 0, factor to None
NA_to_zero <- function(x){
```

```
	x[is.na(x)] <- 0
	return(x)
}

training_dropNA <- lapply(training, function(x){
	if(!is.factor(x) & is.numeric(x))
	{return(NA_to_zero(x))}
	else {
		x<-factor(x, exclude=NULL)
		levels(x)[is.na(levels(x))] <- 'None'
		return(x)
	}})

training_dropNA <- as.data.frame(training_dropNA)

# check negative values
# there are columns having negative values: LotFrontage, MasVnrArea, GarageYrBlt
colSums(training_dropNA<0)

NE_to_zero <- function(x){
	x[x<0] <- 0
	return(x)
}

training_dropNE <- lapply(training_dropNA, function(x){
	if(!is.factor(x) & is.numeric(x))
	{return(NE_to_zero(x))}
	else {
		x<-factor(x, exclude=NULL)
		levels(x)[levels(x)<0] <- 'None'
		return(x)
	}})

# check the class of output
class(training_dropNE)
#convert a list to df
training_dropNE <- as.data.frame(training_dropNE)

# check negative values one more time and no more negative values
colSums(training_dropNE < 0)

#check columns with zeros, drop those columns.
colSums(training_dropNE==0)
#drop_column <- c('MasVnrArea','BsmtFinSF1', 'BsmtFinSF2',
'BsmtUnfSF','TotalBsmtSF','X2ndFlrSF','LowQualFinSF',

#'BsmtFullBath','BsmtHalfBath','FullBath','HalfBath','BedroomAbvGr','Kitche.1bvGr','Fireplaces',
```

```
#'GarageCars','GarageArea','WoodDeckSF','OpenPorchSF','EnclosedPorch','X3SsnPorch','Screen
Porch',
                                                         #'PoolArea','MiscVal')
# I dont think we would want to do this since some of those could still be important and my just
not have those attributes

#training_final <- training_dropNE[ , !(names(training_dropNE) %in% drop_column)]

# Add similar variables together like all the bathrooms and the Porch/deck sf to create one
variable with those
training_dropNE$Bathrooms <- (training_dropNE$BsmtFullBath +
0.5*training_dropNE$BsmtHalfBath + training_dropNE$FullBath +
0.5*training_dropNE$HalfBath)
training_dropNE$PorchSF <- (training_dropNE$WoodDeckSF + training_dropNE$OpenPorchSF +
training_dropNE$EnclosedPorch + training_dropNE$X3SsnPorch +
training_dropNE$ScreenPorch)
training_dropNE$TotalSF <- training_dropNE$TotalBsmtSF + training_dropNE$X1stFlrSF +
training_dropNE$X2ndFlrSF + training_dropNE$LowQualFinSF
training_dropNE$Exterior <- paste(training_dropNE$Exterior1st, training_dropNE$Exterior2nd)
training_dropNE$Condition <- paste(training_dropNE$Condition1,
training_dropNE$Condition2)
training_dropNE$Roof <- paste(training_dropNE$RoofStyle, training_dropNE$RoofMatl)

#Add log variables
training_dropNE$LogSalePrice <- log(training_dropNE$SalePrice)
training_dropNE$LogLotArea <- log(training_dropNE$LotArea)

#drop columns that are combined or not needed
drop_column <- c('BsmtFinSF1', 'BsmtFinSF2',
'BsmtUnfSF','X1stFlrSF','X2ndFlrSF','LowQualFinSF',

'BsmtFullBath','BsmtHalfBath','FullBath','HalfBath','WoodDeckSF','OpenPorchSF','EnclosedPorch
',
          'X3SsnPorch','ScreenPorch', 'Exterior1st', 'Exterior2nd', 'Condition1', 'Condition2',
'RoofStyle', 'RoofMatl')

training_final <- training_dropNE[ , !(names(training_dropNE) %in% drop_column)]
colnames(training_final)[which(names(training_final) == "Kitche.1bvGr")] <- "KitchenAbvGr"
colnames(training_final)[which(names(training_final) == "Functio.1l")] <- "Functional"

write.csv(training_final,"train_clean.csv", row.names = FALSE)

######################### whatever we do to the training, we need to do to the test
set#######################################
#read the test data
test <- read.csv('test.csv', stringsAsFactors=FALSE)
```

```
#clean the neighborhood in test first
test$Neighborhood[grep(test$Neighborhood,pattern = "-1mes")] <- "NAmes"
#test$Neighborhood[grep(test$Neighborhood,pattern = "NWAmes")] <- "NAmes"
#correct to character to factor
character_vars_test <- lapply(test, class) == "character"
test[, character_vars_test] <- lapply(test[, character_vars_test], as.factor)

##########################   check  values   ##########################
# check NA values and there is no NA values
colSums(is.na(test))

# this step: na numeric to 0, factor to None
test_dropNA <- lapply(test, function(x){
        if(!is.factor(x) & is.numeric(x))
        {return(NA_to_zero(x))}
        else {
                x<-factor(x, exclude=NULL)
                levels(x)[is.na(levels(x))] <- 'None'
                return(x)
        }})

test_dropNA <- as.data.frame(test_dropNA)
colSums(is.na(test_dropNA)) #no NA anymore

# check negative values
# there are columns having negative values:
# LotFrontage, Neighborhood,  MasVnrType, MasVnrArea, BsmtQual, BsmtCond,
BsmtExposure, BsmtFinType1,
# BsmtFinType2, Electrical, FireplaceQu, GarageType, GarageYrBlt  GarageFinish,
GarageQual,GarageCond
colSums(test_dropNA<0)

test_dropNE <- lapply(test_dropNA, function(x){
        if(!is.factor(x) & is.numeric(x))
        {return(NE_to_zero(x))}
        else {
                x<-factor(x, exclude=NULL)
                levels(x)[levels(x)<0] <- 'None'
                return(x)
        }})

# check the class of output
class(test_dropNE)
#convert a list to df
test_dropNE <- as.data.frame(test_dropNE)

# check negative values one more time and no more negative values
```

```
colSums(test_dropNE<0)

# Add similar variables together like all the bathrooms and the Porch/deck sf to create one variable with those
test_dropNE$Bathrooms <- (test_dropNE$BsmtFullBath + 0.5*test_dropNE$BsmtHalfBath + test_dropNE$FullBath + 0.5*test_dropNE$HalfBath)
test_dropNE$PorchSF <- (test_dropNE$WoodDeckSF + test_dropNE$OpenPorchSF + test_dropNE$EnclosedPorch + test_dropNE$X3SsnPorch + test_dropNE$ScreenPorch)
test_dropNE$TotalSF <- test_dropNE$TotalBsmtSF + test_dropNE$X1stFlrSF + test_dropNE$X2ndFlrSF + test_dropNE$LowQualFinSF
test_dropNE$Exterior <- paste(test_dropNE$Exterior1st, test_dropNE$Exterior2nd)
test_dropNE$Condition <- paste(test_dropNE$Condition1, test_dropNE$Condition2)
test_dropNE$Roof <- paste(test_dropNE$RoofStyle, test_dropNE$RoofMatl)

#Add log variables
test_dropNE$LogSalePrice <- log(test_dropNE$SalePrice)
test_dropNE$LogLotArea <- log(test_dropNE$LotArea)

#drop columns that are combined or not needed
drop_column <- c('BsmtFinSF1', 'BsmtFinSF2',
'BsmtUnfSF','X1stFlrSF','X2ndFlrSF','LowQualFinSF',

'BsmtFullBath','BsmtHalfBath','FullBath','HalfBath','WoodDeckSF','OpenPorchSF','EnclosedPorch
',
              'X3SsnPorch','ScreenPorch', 'Exterior1st', 'Exterior2nd', 'Condition1', 'Condition2',
'RoofStyle', 'RoofMatl')
#check columns with zeros, drop those columns.
test_final <- test_dropNE[ , !(names(test_dropNE) %in% drop_column)]
colnames(test_final)[which(names(test_final) == "Kitche.1bvGr")] <- "KitchenAbvGr"
colnames(test_final)[which(names(test_final) == "Functio.1l")] <- "Functional"
write.csv (test_final,"test_clean.csv", row.names = FALSE)
```

2. **Analysis 1 code:**

```
/*read data to sas*/
proc import datafile="H:\MSDS6372 stats2\projects\pj1\train_clean.csv"
     dbms=dlm out=train replace;
     delimiter=',';
     getnames=yes;
run;

ods graphics on /  width=10in height=10in;
/*check linearity assumption of MLR*/
proc sgplot data=train;
title 'Scatter Plot of Original Data';
scatter x=LotArea y=SalePrice/group=Neighborhood; run;

proc sgplot data=train;
title 'Scatter Plot of Log Transformed Data';
scatter x = LogLotArea y= LogSalePrice/group=Neighborhood; run;
ods graphics off;

proc import datafile="H:\MSDS6372 stats2\projects\pj1\test_clean.csv"
     dbms=dlm out=test replace;
```

```
     delimiter=',';
     getnames=yes;
run;

data test2;
set test;
SalePrice = .;
;
data train2;
set train test2;
run;
proc means data=train2 NMISS N; run;

/*************************************************************************/
/*using log price and log area, perform stepwise with cv*/
proc glmselect data = train plots= (aseplot);
class MSZoning Street LotShape LandContour Utilities LotConfig LandSlope
Neighborhood BldgType HouseStyle MasVnrType
        ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure
      BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical
        KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual
      GarageCond  PavedDrive SaleType     SaleCondition Exterior Condition
Roof;

model LogSalePrice=MSSubClass MSZoning LotFrontage LogLotArea Street LotShape
LandContour Utilities
                        LotConfig LandSlope Neighborhood BldgType HouseStyle
OverallQual OverallCond YearBuilt
                        YearRemodAdd MasVnrType MasVnrArea ExterQual
ExterCond Foundation BsmtQual BsmtCond BsmtExposure
                        BsmtFinType1 BsmtFinType2 TotalBsmtSF Heating
HeatingQC CentralAir Electrical GrLivArea BedroomAbvGr
                        KitchenAbvGr KitchenQual TotRmsAbvGrd Functional
Fireplaces FireplaceQu GarageType GarageYrBlt GarageFinish
                        GarageCars GarageArea GarageQual GarageCond
PavedDrive PoolArea MiscVal MoSold YrSold SaleType SaleCondition Bathrooms
                        PorchSF TotalSF Exterior Condition Roof
Neighborhood*LogLotArea Neighborhood*OverallCond /selection =
stepwise(choose=cv) SHOWPVALS stats=all;
run;
/*to fit*/
ods graphics on;
proc glm data=train2 PLOTS=DIAGNOSTICS(label);
class MSZoning Street LotShape LandContour Utilities LotConfig LandSlope
Neighborhood BldgType HouseStyle MasVnrType
        ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure
      BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical
        KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual
      GarageCond  PavedDrive SaleType     SaleCondition Exterior Condition
Roof;

model LogSalePrice= CentralAir GarageCars Bathrooms/cli clm solution CLPARM;
output out=logstep p=Predict_LogSalePrice;
run;quit;
ods graphics off;
proc glmmod data=logstep outdesign=GLMDesignstep outparm=GLMParmalexstep;
class MSZoning Street LotShape LandContour Utilities LotConfig LandSlope
Neighborhood BldgType HouseStyle MasVnrType
        ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure
      BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical
```

```
        KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual
      GarageCond PavedDrive SaleType SaleCondition Exterior Condition Roof;

model LogSalePrice= CentralAir GarageCars Bathrooms;
run;
proc print data=GLMDesignstep; run;
proc print data=GLMParmalexstep; run;
proc reg data=GLMDesignstep;
   DummyVars: model LogSalePrice = COL2-COL5/VIF; /* dummy variables except
intercept */
   ods select ParameterEstimates;
quit;
/*to get the test set done*/
proc sql;
create table log_step_final as
select *, mean(Predict_LogSalePrice)as MeanPredict from logstep
group by Neighborhood;quit;
data log_step_final1;
set log_step_final;
if Predict_LogSalePrice > 0 then SalePrice = exp(Predict_LogSalePrice);
else SalePrice = exp(MeanPredict);
keep id SalePrice;
where id > 1460;
run;quit;
proc means data=log_step_final1 NMISS N; run;
Proc export data=log_step_final1
outfile='H:\MSDS6372 stats2\projects\pj1\logstep1.csv'
DBMS=CSV Replace;
run;
/***************************************************************************/
/*using log price and log area, perform forward with cv*/
proc glmselect data = train plots= (aseplot);
class MSZoning Street LotShape LandContour Utilities LotConfig LandSlope
Neighborhood BldgType HouseStyle MasVnrType
        ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure
      BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical
        KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual
      GarageCond  PavedDrive SaleType     SaleCondition Exterior Condition
Roof;

model LogSalePrice=MSSubClass MSZoning LotFrontage LogLotArea Street LotShape
LandContour Utilities
                        LotConfig LandSlope Neighborhood BldgType HouseStyle
OverallQual OverallCond YearBuilt
                        YearRemodAdd MasVnrType MasVnrArea ExterQual
ExterCond Foundation BsmtQual BsmtCond BsmtExposure
                        BsmtFinType1 BsmtFinType2 TotalBsmtSF Heating
HeatingQC CentralAir Electrical GrLivArea BedroomAbvGr
                        KitchenAbvGr KitchenQual TotRmsAbvGrd Functional
Fireplaces FireplaceQu GarageType GarageYrBlt GarageFinish
                        GarageCars GarageArea GarageQual GarageCond
PavedDrive PoolArea MiscVal MoSold YrSold SaleType SaleCondition Bathrooms
                        PorchSF TotalSF Exterior Condition Roof
Neighborhood*LogLotArea Neighborhood*OverallCond /selection =
forward(choose=cv) SHOWPVALS stats=all;
run;
ods graphics on;
proc glm data=train2 PLOTS=DIAGNOSTICS(label);
class MSZoning Street LotShape LandContour Utilities LotConfig LandSlope
Neighborhood BldgType HouseStyle MasVnrType
```

```
        ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure
      BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical
        KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual
      GarageCond  PavedDrive SaleType     SaleCondition Exterior Condition
Roof;

model LogSalePrice=MSZoning GarageCars Bathrooms/cli clm solution CLPARM;
output out=logforward p=Predict_LogSalePrice;
run;quit;
proc glmmod data=logforward outdesign=GLMDesignfor outparm=GLMParmfor;
class MSZoning Street LotShape LandContour Utilities LotConfig LandSlope
Neighborhood BldgType HouseStyle MasVnrType
        ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure
      BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical
        KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual
      GarageCond  PavedDrive SaleType     SaleCondition Exterior Condition
Roof;

model LogSalePrice=MSZoning GarageCars Bathrooms;
run;
proc print data=GLMDesignfor; run;
proc print data=GLMParmfor; run;
proc reg data=GLMDesignfor;
   DummyVars: model LogSalePrice = COL2-COL8/VIF; /* dummy variables except
intercept */
   ods select ParameterEstimates;
quit;
ods graphics off;
/*to get the test set done*/
proc sql;
create table log_forward_final as
select *, mean(Predict_LogSalePrice)as MeanPredict from logforward
group by Neighborhood;quit;
data log_forward_final1;
set log_forward_final;
if Predict_LogSalePrice > 0 then SalePrice = exp(Predict_LogSalePrice);
else SalePrice = exp(MeanPredict);
keep id SalePrice;
where id > 1460;
run;quit;
proc means data=log_forward_final1 NMISS N; run;
Proc export data=log_forward_final1
outfile='H:\MSDS6372 stats2\projects\pj1\logforward1.csv'
DBMS=CSV Replace;
run;
proc glmselect data = train plots(stepaxis = number) = (criterionpanel
ASEPlot);

class MSZoning Street LotShape LandContour Utilities LotConfig LandSlope
Neighborhood BldgType HouseStyle MasVnrType YearBuilt YearRemodAdd
  ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1
BsmtFinType2 Heating HeatingQC CentralAir Electrical
  KitchenQual FireplaceQu GarageType GarageFinish GarageQual GarageCond
PavedDrive SaleType SaleCondition Exterior Condition Roof;

model LogSalePrice= LotFrontage LogLotArea Street Utilities LotConfig
Neighborhood BldgType HouseStyle OverallQual OverallCond YearBuilt
YearRemodAdd ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure
BsmtFinType1 BsmtFinType2 TotalBsmtSF
Heating HeatingQC CentralAir Electrical GrLivArea BedroomAbvGr KitchenQual
TotRmsAbvGrd
```

Nuoya Rezsonya & Alexandra Norman

```
Fireplaces FireplaceQu GarageType GarageYrBlt GarageFinish GarageCars
GarageArea GarageQual GarageCond PavedDrive
PoolArea MiscVal MoSold YrSold SaleType SaleCondition Bathrooms PorchSF
TotalSF Exterior Condition Roof
Neighborhood*LogLotArea Neighborhood*OverallCond /selection = LASSO
(choose=CV) SHOWPVALS stats=all;
run;
/*to fit*/

ods graphics on;
proc glm data=train2 PLOTS=DIAGNOSTICS(label);
class MSZoning Street LotShape LandContour Utilities LotConfig LandSlope
Neighborhood BldgType HouseStyle MasVnrType
        ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure
      BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical
        KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual
      GarageCond PavedDrive SaleType SaleCondition Exterior Condition Roof;

model LogSalePrice= LogLotArea CentralAir Bathrooms /cli clm solution CLPARM;
output out=logalex p=Predict_LogSalePrice;
run;quit;
ods graphics off;
proc glmmod data=logalex outdesign=GLMDesignalex outparm=GLMParmalex;
class MSZoning Street LotShape LandContour Utilities LotConfig LandSlope
Neighborhood BldgType HouseStyle MasVnrType
        ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure
      BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical
        KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual
      GarageCond PavedDrive SaleType SaleCondition Exterior Condition Roof;

model LogSalePrice= LogLotArea CentralAir Bathrooms;
run;
proc print data=GLMDesignalex; run;
proc print data=GLMParmalex; run;
proc reg data=GLMDesignalex;
   DummyVars: model LogSalePrice = COL2-COL5/VIF; /* dummy variables except
intercept */
   ods select ParameterEstimates;
quit;
/*to get the test set done*/
proc sql;
create table log_alex_final as
select *, mean(Predict_LogSalePrice)as MeanPredict from logalex
group by Neighborhood;quit;
data log_alex_final1;
set log_alex_final;
if Predict_LogSalePrice > 0 then SalePrice = exp(Predict_LogSalePrice);
else SalePrice = exp(MeanPredict);
keep id SalePrice;
where id > 1460;
run;quit;
proc means data=log_alex_final1 NMISS N; run;
Proc export data=log_alex_final1
outfile='H:\MSDS6372 stats2\projects\pj1\logalex1.csv'
DBMS=CSV Replace;
run;
```

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	B	11.14742	0.02106	529.22	<.0001	0
Col2	CentralAir N	B	-0.26445	0.02623	-10.08	<.0001	1.06938
Col3	CentralAir Y	0	0	.	.	.	.
Col4	GarageCars	1	0.23227	0.00969	23.97	<.0001	1.33903
Col5	Bathrooms	1	0.21867	0.00915	23.89	<.0001	1.31953

Stepwise selection parameter results

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	11.14742460	B	0.02106380	529.22	<.0001
CentralAir N	-0.26445242	B	0.02623231	-10.08	<.0001
CentralAir Y	0.00000000	B	.	.	.
GarageCars	0.23226669		0.00969139	23.97	<.0001
Bathrooms	0.21866749		0.00915407	23.89	<.0001

Stepwise selection VIF check

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	B	11.02084	0.02226	495.15	<.0001	0
Col2	MSZoning C (a	B	-0.45808	0.07677	-5.97	<.0001	1.04192
Col3	MSZoning FV	B	0.17009	0.03466	4.91	<.0001	1.32818
Col4	MSZoning RH	B	0.08332	0.06148	1.36	0.1755	1.06467
Col5	MSZoning RL	B	0.15738	0.01834	8.58	<.0001	1.45866
Col6	MSZoning RM	0	0	.	.	.	.
Col7	GarageCars	1	0.23822	0.00958	24.87	<.0001	1.33044
Col8	Bathrooms	1	0.20484	0.00930	22.03	<.0001	1.38556

Forward selection parameter results

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	11.02083872	B	0.02225751	495.15	<.0001
MSZoning C (a	-0.45807829	B	0.07677408	-5.97	<.0001
MSZoning FV	0.17009192	B	0.03466304	4.91	<.0001
MSZoning RH	0.08332379	B	0.06148159	1.36	0.1755
MSZoning RL	0.15737762	B	0.01834178	8.58	<.0001
MSZoning RM	0.00000000	B	.	.	.
GarageCars	0.23822130		0.00957794	24.87	<.0001
Bathrooms	0.20484240		0.00930037	22.03	<.0001

Forward selection VIF check

3. Analysis 2 Code:

```
/*read train data to sas*/
proc import datafile="H:\MSDS6372 stats2\projects\pj1\train_clean.csv"
     dbms=dlm out=train replace;
     delimiter=',';
     getnames=yes;
run;

ods graphics on /  width=10in height=10in;
/*check linearity assumption of MLR*/
proc sgplot data=train;
scatter x=LotArea y=SalePrice/group=Neighborhood; run;
/*log scatter is better*/
proc sgplot data=train;
scatter x = LogLotArea y= LogSalePrice/group=Neighborhood; run;
ods graphics off;
/*encode all the categorical variables and remove columns with high VIF*/
proc glmmod data=train outdesign=GLMDesign outparm=GLMParm;
class MSZoning Street LotShape LandContour Utilities LotConfig LandSlope
Neighborhood BldgType HouseStyle MasVnrType
        ExterQual ExterCond Foundation BsmtQual BsmtCond BsmtExposure
      BsmtFinType1 BsmtFinType2 Heating HeatingQC CentralAir Electrical
        KitchenQual Functional FireplaceQu GarageType GarageFinish GarageQual
      GarageCond  PavedDrive SaleType     SaleCondition Exterior Condition
Roof;
model LogSalePrice=MSSubClass MSZoning LotFrontage LogLotArea Street LotShape
LandContour Utilities
                        LotConfig LandSlope Neighborhood BldgType HouseStyle
OverallQual OverallCond YearBuilt
                        YearRemodAdd MasVnrType MasVnrArea ExterQual
ExterCond Foundation BsmtQual BsmtCond BsmtExposure
                        BsmtFinType1 BsmtFinType2 TotalBsmtSF Heating
HeatingQC CentralAir Electrical GrLivArea BedroomAbvGr
```

```
                        KitchenAbvGr KitchenQual TotRmsAbvGrd Functional
Fireplaces FireplaceQu GarageType GarageYrBlt GarageFinish
                        GarageCars GarageArea GarageQual GarageCond
PavedDrive PoolArea MiscVal MoSold YrSold SaleType SaleCondition Bathrooms
                        PorchSF TotalSF Exterior Condition Roof;
run;
proc print data=GLMDesign; run;
proc print data=GLMParm; run;
proc reg data=GLMDesign;
   DummyVars: model LogSalePrice = COL2-COL315/VIF; /* dummy variables except
intercept */
   ods select ParameterEstimates;
quit;
/*after checking the result, there are variables that have large
VIF,MSSubClass LandSlope YearBuilt Foundation BsmtExposure GrLivArea
GarageType GarageYrBlt */
/*delete them*/
data train_new(drop = MSSubClass LandSlope YearBuilt Foundation BsmtExposure
GrLivArea GarageType GarageYrBlt);
set train;
run;
Proc export data=train_new
outfile='H:\MSDS6372 stats2\projects\pj1\train_new.csv'
DBMS=CSV Replace;
run;
proc import datafile="H:\MSDS6372 stats2\projects\pj1\test_clean.csv"
     dbms=dlm out=test replace;
     delimiter=',';
     getnames=yes;
run;
data test(drop = MSSubClass LandSlope YearBuilt Foundation BsmtExposure
GrLivArea GarageType GarageYrBlt);
set test;run;
proc print data=test;
run;
data test2;
set test;
SalePrice = .;
;
data train2;
set train_new test2;
run;
/*LASSO MODEL VARIABLE SELECTION*/
ods graphics on;
proc glmselect data = train_new plots= (aseplot criteria);partition
fraction(test = .5);
class MSZoning Street LotShape LandContour Utilities LotConfig Neighborhood
BldgType HouseStyle MasVnrType
        ExterQual ExterCond BsmtQual BsmtCond BsmtFinType1 BsmtFinType2
Heating HeatingQC CentralAir Electrical
        KitchenQual Functional FireplaceQu GarageFinish GarageQual GarageCond
PavedDrive SaleType     SaleCondition Exterior Condition Roof;

model LogSalePrice=MSZoning   LotFrontage Street LotShape LandContour
Utilities LotConfig Neighborhood BldgType HouseStyle OverallQual
                              OverallCond YearRemodAdd MasVnrType MasVnrArea
ExterQual ExterCond BsmtQual BsmtCond BsmtFinType1
                              BsmtFinType2 TotalBsmtSF Heating HeatingQC
CentralAir Electrical BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd
                              Functional Fireplaces FireplaceQu GarageFinish
GarageCars GarageArea GarageQual GarageCond PavedDrive PoolArea   MiscVal
```

```
                              MoSold YrSold SaleType SaleCondition Bathrooms
PorchSF TotalSF Exterior Condition Roof LogLotArea Neighborhood*LogLotArea
Neighborhood*OverallCond/selection = lasso (choose=cv stop=cv) cvdetails=all
SHOWPVALS stats=all;
run;
ods graphics off;
/*to fit*/
ods graphics on;
proc glm data=train2 PLOTS=DIAGNOSTICS(label);
class MSZoning Street LotShape LandContour Utilities LotConfig Neighborhood
BldgType HouseStyle MasVnrType
        ExterQual ExterCond BsmtQual BsmtCond BsmtFinType1 BsmtFinType2
Heating HeatingQC CentralAir Electrical
        KitchenQual Functional FireplaceQu GarageFinish GarageQual GarageCond
PavedDrive SaleType     SaleCondition Exterior Condition Roof;

model LogSalePrice= MSZoning OverallQual OverallCond YearRemodAdd MasVnrArea
                              BsmtQual BsmtFinType1 HeatingQC CentralAir
                              Fireplaces GarageCars GarageArea GarageCond
PavedDrive SaleCondition Bathrooms TotalSF Exterior LogLotArea/cli clm
solution;
output out=loglasso p=Predict_LogSalePrice;
run;quit;
ods graphics off;

/*removing outliers*/
data final2;
set train2;
if Id eq 1299 then delete;
if Id eq 524 then delete;
run;
ods graphics on;
proc glm data=final2 PLOTS=DIAGNOSTICS(label);
class MSZoning Street LotShape LandContour Utilities LotConfig Neighborhood
BldgType HouseStyle MasVnrType
        ExterQual ExterCond BsmtQual BsmtCond BsmtFinType1 BsmtFinType2
Heating HeatingQC CentralAir Electrical
        KitchenQual Functional FireplaceQu GarageFinish GarageQual GarageCond
PavedDrive SaleType     SaleCondition Exterior Condition Roof;

model LogSalePrice= MSZoning OverallQual OverallCond YearRemodAdd MasVnrArea
                              BsmtQual BsmtFinType1 HeatingQC CentralAir
                              Fireplaces GarageCars GarageArea GarageCond
PavedDrive SaleCondition Bathrooms TotalSF Exterior LogLotArea/cli clm
solution;
output out=loglasso p=Predict_LogSalePrice;
run;quit;
ods graphics off;

/*to get the test set done*/
proc sql;
create table log_lasso_final2 as
select *, mean(Predict_LogSalePrice)as MeanPredict from loglasso
group by Neighborhood;quit;
data log_lasso_final_q2;
set log_lasso_final2;
if Predict_LogSalePrice > 0 then SalePrice = exp(Predict_LogSalePrice);
else SalePrice = exp(MeanPredict);
keep id SalePrice;
where id > 1460;
run;quit;
```

```
proc means data=log_lasso_final_q2 NMISS N; run;
Proc export data=log_lasso_final_q2
outfile='H:\MSDS6372 stats2\projects\pj1\loglasso2.csv'
DBMS=CSV Replace;
run; /*kaggle 0.14161*/


/*FORWARD SELECTION MODEL VARIABLE SELECTION*/
ods graphics on;
proc glmselect data = train_new plots= (aseplot criteria);partition
fraction(test = .5);
class MSZoning Street LotShape LandContour Utilities LotConfig Neighborhood
BldgType HouseStyle MasVnrType
        ExterQual ExterCond BsmtQual BsmtCond BsmtFinType1 BsmtFinType2
Heating HeatingQC CentralAir Electrical
        KitchenQual Functional FireplaceQu GarageFinish GarageQual GarageCond
PavedDrive SaleType     SaleCondition Exterior Condition Roof;

model LogSalePrice=MSZoning   LotFrontage Street LotShape LandContour
Utilities LotConfig Neighborhood BldgType HouseStyle OverallQual
                              OverallCond YearRemodAdd MasVnrType MasVnrArea
ExterQual ExterCond BsmtQual BsmtCond BsmtFinType1
                              BsmtFinType2 TotalBsmtSF Heating HeatingQC
CentralAir Electrical BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd
                              Functional Fireplaces FireplaceQu GarageFinish
GarageCars GarageArea GarageQual GarageCond PavedDrive PoolArea   MiscVal
                              MoSold YrSold SaleType SaleCondition Bathrooms
PorchSF TotalSF Exterior Condition Roof LogLotArea Neighborhood*LogLotArea
Neighborhood*OverallCond/selection = stepwise(choose=cv stop=cv)
cvdetails=all SHOWPVALS stats=all;
run;
ods graphics off;
/*to fit*/
ods graphics on;
proc glm data=train2 PLOTS=DIAGNOSTICS(label);
class MSZoning Street LotShape LandContour Utilities LotConfig Neighborhood
BldgType HouseStyle MasVnrType
        ExterQual ExterCond BsmtQual BsmtCond BsmtFinType1 BsmtFinType2
Heating HeatingQC CentralAir Electrical
        KitchenQual Functional FireplaceQu GarageFinish GarageQual GarageCond
PavedDrive SaleType     SaleCondition Exterior Condition Roof;

model LogSalePrice= MSZoning OverallQual YearRemodAdd BsmtQual Heating
KitchenAbvGr KitchenQual Functional
                              Fireplaces GarageCars GarageQual Bathrooms
TotalSF LogLotArea OverallCond*Neighborhood/cli clm solution;
output out=logforward p=Predict_LogSalePrice;
run;quit;
ods graphics off;
/*removing outliers*/
data final2;
set train2;
if Id eq 1299 then delete;
if Id eq 524 then delete;
run;
ods graphics on;
proc glm data=final2 PLOTS=DIAGNOSTICS(label);
class MSZoning Street LotShape LandContour Utilities LotConfig Neighborhood
BldgType HouseStyle MasVnrType
        ExterQual ExterCond BsmtQual BsmtCond BsmtFinType1 BsmtFinType2
Heating HeatingQC CentralAir Electrical
```

```
        KitchenQual Functional FireplaceQu GarageFinish GarageQual GarageCond
PavedDrive SaleType     SaleCondition Exterior Condition Roof;

model LogSalePrice= MSZoning OverallQual YearRemodAdd BsmtQual Heating
KitchenAbvGr KitchenQual Functional
                              Fireplaces GarageCars GarageQual Bathrooms
TotalSF LogLotArea OverallCond*Neighborhood/cli clm solution;
output out=logforward p=Predict_LogSalePrice;
run;quit;
ods graphics off;

/*to get the test set done*/
proc sql;
create table log_forward_final2 as
select *, mean(Predict_LogSalePrice)as MeanPredict from logforward
group by Neighborhood;quit;

data log_forward_final_q2;
set log_forward_final2;
if Predict_LogSalePrice > 0 then SalePrice = exp(Predict_LogSalePrice);
else SalePrice = exp(MeanPredict);
keep id SalePrice;
where id > 1460;
run;quit;
proc means data=log_forward_final_q2 NMISS N; run;
Proc export data=log_forward_final_q2
outfile='H:\MSDS6372 stats2\projects\pj1\logforward2.csv'
DBMS=CSV Replace;
run;/*kaggle 0.13732*/

/*CUSTOM MODEL VARIABLE SELECTION*/
```

Run this three times

```
proc glmselect data = train_new plots= (aseplot criteria);partition
fraction(test = .5);
class MSZoning Street LotShape LandContour Utilities LotConfig Neighborhood
BldgType HouseStyle MasVnrType
        ExterQual ExterCond BsmtQual BsmtCond BsmtFinType1 BsmtFinType2
Heating HeatingQC CentralAir Electrical
        KitchenQual Functional FireplaceQu GarageFinish GarageQual GarageCond
PavedDrive SaleType     SaleCondition Exterior Condition Roof;

model LogSalePrice= MSZoning LotFrontage Street LotShape LandContour
Utilities LotConfig Neighborhood BldgType HouseStyle OverallQual
                              OverallCond YearRemodAdd MasVnrType MasVnrArea
ExterQual ExterCond BsmtQual BsmtCond BsmtFinType1
                              BsmtFinType2 TotalBsmtSF Heating HeatingQC
CentralAir Electrical BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd
                              Functional Fireplaces FireplaceQu GarageFinish
GarageCars GarageArea GarageQual GarageCond PavedDrive PoolArea   MiscVal
                              MoSold YrSold SaleType SaleCondition Bathrooms
PorchSF TotalSF Exterior Condition Roof LogLotArea Neighborhood*LogLotArea
Neighborhood*OverallCond/selection = forward (choose=cv stop=cv)
cvdetails=all SHOWPVALS stats=all;
run;
/*Run this 3 times*/
proc glmselect data = train_new plots= (aseplot criteria);partition
fraction(test = .5);
class MSZoning Street LotShape LandContour Utilities LotConfig Neighborhood
BldgType HouseStyle MasVnrType
        ExterQual ExterCond BsmtQual BsmtCond BsmtFinType1 BsmtFinType2
Heating HeatingQC CentralAir Electrical
```

```
        KitchenQual Functional FireplaceQu GarageFinish GarageQual GarageCond
PavedDrive SaleType     SaleCondition Exterior Condition Roof;

model LogSalePrice= MSZoning LotFrontage Street LotShape LandContour
Utilities LotConfig Neighborhood BldgType HouseStyle OverallQual
                              OverallCond YearRemodAdd MasVnrType MasVnrArea
ExterQual ExterCond BsmtQual BsmtCond BsmtFinType1
                              BsmtFinType2 TotalBsmtSF Heating HeatingQC
CentralAir Electrical BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd
                              Functional Fireplaces FireplaceQu GarageFinish
GarageCars GarageArea GarageQual GarageCond PavedDrive PoolArea   MiscVal
                              MoSold YrSold SaleType SaleCondition Bathrooms
PorchSF TotalSF Exterior Condition Roof LogLotArea Neighborhood*LogLotArea
Neighborhood*OverallCond/selection = elasticnet (choose=cv stop=cv)
cvdetails=all SHOWPVALS stats=all;
run;
/*Run this 3 times*/
proc glmselect data = train_new plots= (aseplot criteria);partition
fraction(test = .5);
class MSZoning Street LotShape LandContour Utilities LotConfig Neighborhood
BldgType HouseStyle MasVnrType
        ExterQual ExterCond BsmtQual BsmtCond BsmtFinType1 BsmtFinType2
Heating HeatingQC CentralAir Electrical
        KitchenQual Functional FireplaceQu GarageFinish GarageQual GarageCond
PavedDrive SaleType     SaleCondition Exterior Condition Roof;

model LogSalePrice= MSZoning LotFrontage Street LotShape LandContour
Utilities LotConfig Neighborhood BldgType HouseStyle OverallQual
                              OverallCond YearRemodAdd MasVnrType MasVnrArea
ExterQual ExterCond BsmtQual BsmtCond BsmtFinType1
                              BsmtFinType2 TotalBsmtSF Heating HeatingQC
CentralAir Electrical BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd
                              Functional Fireplaces FireplaceQu GarageFinish
GarageCars GarageArea GarageQual GarageCond PavedDrive PoolArea   MiscVal
                              MoSold YrSold SaleType SaleCondition Bathrooms
PorchSF TotalSF Exterior Condition Roof LogLotArea Neighborhood*LogLotArea
Neighborhood*OverallCond/selection = lasso (choose=cv stop=cv) cvdetails=all
SHOWPVALS stats=all;
run;

proc glmselect data = train_new plots= (aseplot criteria);partition
fraction(test = .6);
class MSZoning Street LotShape LandContour Utilities LotConfig Neighborhood
BldgType HouseStyle MasVnrType
        ExterQual ExterCond BsmtQual BsmtCond BsmtFinType1 BsmtFinType2
Heating HeatingQC CentralAir Electrical
        KitchenQual Functional FireplaceQu GarageFinish GarageQual GarageCond
PavedDrive SaleType     SaleCondition Exterior Condition Roof;
model logSalePrice= MSZoning OverallQual YearRemodAdd BsmtQual HeatingQC
CentralAir KitchenQual GarageCars
                              GarageArea Bathrooms TotalSF LogLotArea
Fireplaces OverallCond*Neighborhood ExterCond Functional
                              SaleCondition SaleType PorchSF/ selection =
none CVDETAILS stats=all;
output out = resultscustomv2 p = logPredict;
run;

proc glm data = final PLOTS=DIAGNOSTICS(label);
class MSZoning Street LotShape LandContour Utilities LotConfig Neighborhood
BldgType HouseStyle MasVnrType
```

```
        ExterQual ExterCond BsmtQual BsmtCond BsmtFinType1 BsmtFinType2
Heating HeatingQC CentralAir Electrical
        KitchenQual Functional FireplaceQu GarageFinish GarageQual GarageCond
PavedDrive SaleType     SaleCondition Exterior Condition Roof;

model LogSalePrice= MSZoning OverallQual YearRemodAdd BsmtQual HeatingQC
CentralAir KitchenQual GarageCars
                              GarageArea Bathrooms TotalSF LogLotArea
Fireplaces OverallCond*Neighborhood ExterCond Functional
                              SaleCondition SaleType PorchSF;
output out = resultscustomv2 p = logPredict;
run;

data final2;
set final;
if Id eq 1299 then delete;
if Id eq 524 then delete;
run;

proc glm data = final2 PLOTS=DIAGNOSTICS(label);
class MSZoning Street LotShape LandContour Utilities LotConfig Neighborhood
BldgType HouseStyle MasVnrType
        ExterQual ExterCond BsmtQual BsmtCond BsmtFinType1 BsmtFinType2
Heating HeatingQC CentralAir Electrical
        KitchenQual Functional FireplaceQu GarageFinish GarageQual GarageCond
PavedDrive SaleType     SaleCondition Exterior Condition Roof;

model LogSalePrice= MSZoning OverallQual YearRemodAdd BsmtQual HeatingQC
CentralAir KitchenQual GarageCars
                              GarageArea Bathrooms TotalSF LogLotArea
Fireplaces OverallCond*Neighborhood ExterCond Functional
                              SaleCondition SaleType PorchSF;
output out = resultscustomv3 p = logPredict;
run;

data resultscustom2;
      set resultscustomv3;
      SalePrice = exp(logSalePrice);
      Predict = exp(logPredict);
      run;

proc sql;
create table resultscustom3 as
select *, mean(SalePrice) as MeanSalePricebyNeigh
from resultscustom2
group by Neighborhood;
quit;

data resultscustomfinalv2;
      set resultscustom3;
      if Predict > 0 then SalePrice = Predict;
      if Predict < 0 then SalePrice = MeanSalePricebyNeigh;
      keep id SalePrice;
      where id > 1460;
      ;
```

Extra Credit:

Appendix:

```
/*read train data to sas*/
proc import datafile="H:\MSDS6372 stats2\projects\pj1\train_new.csv"
     dbms=dlm out=train_new replace;
```

```
     delimiter=',';
     getnames=yes;
run;
/*import the test*/
proc import datafile="H:\MSDS6372 stats2\projects\pj1\test_clean.csv"
     dbms=dlm out=test replace;
     delimiter=',';
     getnames=yes;
run;

data test2;
set test;
SalePrice = .;
;
data train2;
set train_new test2;
run;

ods graphics on;
/*two categorical variables are: PavedDrive and centralAir*/
/*the following glm is the additive model*/
proc glm data=train2 PLOTS=(DIAGNOSTICS RESIDUALS);
class MSZoning Street LotShape LandContour Utilities LotConfig Neighborhood
BldgType HouseStyle MasVnrType
        ExterQual ExterCond BsmtQual BsmtCond BsmtFinType1 BsmtFinType2
Heating HeatingQC CentralAir Electrical
        KitchenQual Functional FireplaceQu GarageFinish GarageQual GarageCond
PavedDrive SaleType     SaleCondition Exterior Condition Roof;

model LogSalePrice=PavedDrive CentralAir/CLI CLM SOLUTION CLPARM;
lsmeans CentralAir / pdiff tdiff adjust=bon;
run;quit;
ods graphics off;
/*********************************************************************/
ods graphics on;
/*two categorical variables are: PavedDrive and centralAir*/
/*the following glm is the non-additive model*/
proc glm data=train2 PLOTS=(DIAGNOSTICS RESIDUALS);
class MSZoning Street LotShape LandContour Utilities LotConfig Neighborhood
BldgType HouseStyle MasVnrType
        ExterQual ExterCond BsmtQual BsmtCond BsmtFinType1 BsmtFinType2
Heating HeatingQC CentralAir Electrical
        KitchenQual Functional FireplaceQu GarageFinish GarageQual GarageCond
PavedDrive SaleType     SaleCondition Exterior Condition Roof;

model LogSalePrice=PavedDrive CentralAir PavedDrive*CentralAir /CLI CLM
SOLUTION CLPARM;
lsmeans CentralAir / pdiff tdiff adjust=bon;
run;quit;
ods graphics off;
```