#### MSDS 6371 – KAGGLE PROJECT

# **Introduction**

           Century 21 Ames has commissioned us to answer how the sale price of their houses in the North Ames Edwards and Brookside neighborhoods are related to the square footage of the living area of the house. Additionally, they would like to know if the sales price (and its relationship to square footage) depends on which neighborhood the house is located in. I. answering this we have been asked to build the most predictive model for sales prices of homes in all of Ames, Iowa (This includes all of the neighborhoods).

# **Data Description**

* The data involved includes details of 1460 houses in Ames Iowa Neighborhood. The data is a consolidated with 80 Features of all the houses in this area. There are totally 38 numerical and 44 categorical variables.
* Data has been cleaned all the missing values have been replaced by either the mean of the variable or the meaningful keywords used for analysis.

# **QOI 1 - Analysis Question 1:**

**QOI 1 - Restatement of Problem:**

The data was taken from Kaggle - The House Prices Advanced Regression techniques. Since the analysis was done for Century 21 Ames (a real estate company) we had to filter the data set for houses in below neighborhoods (Names, Edwards and BrkSide)

So here we have the Sales price of 383 houses with details of the living area respectfully.

**QOI 1 - ANALYSIS MODEL 1:**

In [Appendix, Plot 1.1](#Plot) we have analyzed the Scatter plot between Sales Price and the Living Room Area of houses in 3 neighborhoods (NAmes, Edwards and BrkSide). Here we are using a trial method to narrow out the best model eliminating all the outliers and inferential points.

**Model 1: Simple Linear Regression between Living Room Area and the Sales Price of the house.**

**Predicted Sales Price = 0+1Living Room Area**

0 – Intercept of the model otherwise the theoretical price of a house if the Area is zero, which ideally is not possible.

1 – Slope of the model, meaning with every 100 sq ft increase in the living room area the Sales Price increase by 1

**QOI 1 - ASSUMPTIONS:**

Judging by the scatter plot, QQ plot, histogram of the residuals there is sufficient evidence of normality and Linearity. But

the residual plots the points are not scattered. However, there are some outliers with high Leverage and Cooks D which

needs to be removed.

**QOI 1 - EVALUATION MODEL 1**

Adj R2 - 0.3406 (i.e., only 34% of the variation is explained by the explanatory variable which in this case it is the

Living RArea.) Using this model outliers and influential points were identified

* 4 outliers with student Residuals that does not lie within -2.5 and 2.5
* 1 outlier with CooksD greater than 5
* 1 outlier with leverage higher than 0.04

So, we decided to rerun the model after without these outliers. So, after removing the outliers we have received the model.

**QOI 1. ANALYSIS MODEL 2**

Now we removed the outliers with high Cooks D and high residual was removed, after which we have implemented the

Below model

**Predicted Sales Price**

**QOI 1. ASSUMPTIONS MODEL 2**

Judging by the scatter plot, QQ plot, histogram of the residuals there is sufficient evidence of normality and Linearity. But

the residual plots the points are but clustered so here transformation might be required.

**QOI 1. EVALUATION OF THIS MODEL 2**

Adj R2 - 0.4963 (i.e., only 49% of the variation is explained by the explanatory variable which in this case it is the

Living Room Area.). Hence this model is better than the first model.

Table

Description automatically generated with medium confidence

Since we have analyzed the plots and removed the outliers, we are going to analyze the Living room area with respect to

the neighborhood variables.

|  |  |  |
| --- | --- | --- |
|  | Dummy1 | Dummy2 |
| Names | 0 | 0 |
| Edwards | 0 | 1 |
| Brkside | 1 | 0 |

**QOI 1: ANALYSIS MODEL 3**

Now we removed the outliers with high Cooks D and high residual was removed, after which we have implemented the

Below model

**QOI 1. ASSUMPTIONS MODEL 3**

Judging by the scatter plot, QQ plot, histogram of the residuals there is sufficient evidence of normality and Linearity.

Now the residual plots look much random when comparing to the other models.

**QOI 1: EVALUATION OF THIS MODEL 3**

Adj R2 - 0.5122(i.e., only 51% of the variation is explained by the explanatory variable which in this case it is the

Living Room Area.). Upon analyzing all the models, we can assume that this is the final model

Table

Description automatically generated with medium confidence

**Adj. R2 = 0.5171**

**Since the assumptions are all met the Adjusted R^2 is higher compared to all other models we will proceed with Model 3 for QOI 1.**

**Predicted Sales Price**

Y is the final sale price of the house

𝛽0 is the y-intercept, or the theoretical value of houses in the NAmes neighborhood

𝛽1 is square footage area of the house

𝛽2 is the coefficient representing the modifier to Sales price for houses located in the BrkSide  
𝛽3 is the coefficient representing the modifier to Sales price for houses located in the Edwards  
𝛽4 is the coefficient representing area of the house affects if the house is located in BrkSide

𝛽5 is the coefficient representing area of the house affects if the house is located in Edwards

Below is the parameter estimate table for the same.

#### Table Description automatically generated

#### Now we are simplifying

#### Regression Model for NAmes (d1=0, d2=0): μ (Sale Price) = 11.46 + 0.00030\*GrLivArea

#### Regression Model for BrkSide (d1=0, d2=1): μ (Sale Price) = 10.78851+0.0046072\*GrLivArea

#### Regression Model for Edwards (d1=0, d2=0): μ (Sale Price) = 11.23+ 0.0005\*GrLivArea

#### 

**Confidence intervals & Interpretation**

The above model is a good fit to predict the sales price of all the 3 neighborhoods, There is a p-value < .0001 for the Overall F-test proving the significance.

The model has an Adj R^2 of 51.2% which means that 51.2% of the variability of Sale Price is explained by the living area of the house.

• In the Names neighborhood, every 100 sq. ft living area increase results in an estimated $1000 increase on sale price, with 95% confidence interval from $1003.255 to $ 9973.45

• In neighborhood BrkSide, every 100 sq living area increase results in an estimated $8,720 increase of sale price, with 95% confidence interval from $0 to $10,788.

• In neighborhood Edwards, every 100 sq living area increase results in an estimated $7,020 increase on sale price, with 95% confidence interval from $5,618 to $8,413.

#### Conclusion

The above analysis corelates to the houses in the 3 neighborhoods Brkside,Edwards and the Names. Prices in the Names neighborhood are the costliest compared to the other neighborhood’s.

# **Analysis Question 2:**

**Restatement of Problem**

We have analyzed the data provided to build a predictive model for sale prices of all the individual residential property in all neighborhoods in Ames, Iowa.

We have analyzed the data using SAS and used multiple linear regression techniques for prediction. For variable selection we have used Stepwise, Forward, Backward process. After which we will compare the parameters of the different models using the stats (adjusted R2, internal CV Press and Kaggle Score) .

**Data Cleaning:**

- For analysis we have started cleaning the data and selecting numerical variables as predictor. We have identified and removed 3 variables with NA’s (eg. Lot Frontage, MasVnrArea, GarageYrBlt). And the Categorical Variables have been converted to numerical equivalent for analysis.

**Model Selection**

**Background**

#### Here in this analysis, we will be building a predictive model to predict the price of the residential properties across all neighborhoods in Iowa. We are going to use various variable selection methods to select only the most significant variables for this analysis. We are going to use Stepwise, Forward, Backward and Custom process selection for our analysis.

#### For this analysis we have done selected the variable based on the VIF and the p value of various variables.

#### 

**Checking Assumptions**

- There is no evidence against Normality, Equal Variances. All the residuals’ plots after the elimination of the outliers that have been identified. However, there is a light curvature in the qq plot because of which the Sale Price was Log transformed.

**Model**

The sale price of the houses was log transformed for all models.

**Model Metrics**

**Custom selection** – Custom model selection with the selected numerical values, for this forward selection we are only using numerical predictors handpicked using the scatterplots (refer Plot 2.8 to Plot 2.10). The below variables are selected based on VIF , Adjusted R^2 values.

Variables Selected for Analysis:

OverallCond ,YearRemodAdd

YearBuilt ,BsmtUnfSF, OpenPorchSF, GarageArea ,TotRmsAbvGrd ,LotArea, WoodDeckSF

Fireplaces ,MasVnrArea ,OverallQual ,GarageCars

GrLivAreaEnclosedPorch FullBath LotConfig\_num\_cat

Neighborhood\_num\_cat ExterQual\_num\_cat BsmtQual\_num\_cat HeatingQC\_num\_cat

KitchenQual\_num\_cat FireplaceQu\_num\_cat

GarageQual\_num\_cat PoolQC\_num\_cat SaleCondition\_num\_cat;

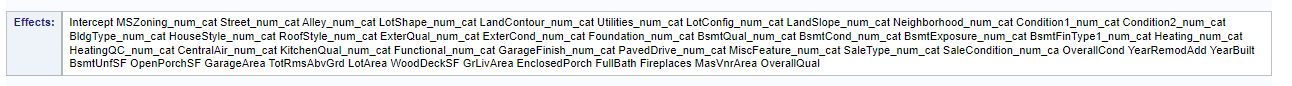
Adj R^2 : 0.8690

CV PRESS : 30.91150

**Backwards selection :** Here we have done backward selection with the below selected variables. Please refer code and plots from Appendix. Below are the features narrowed by this selection process.

Adj R^2 : 0.9012

CV Press : 21.8151



Over all F test shows that the model is significant.

**Stepwise Selection:** Here we have done step wise selection with the below selected variables. Please refer code and plots from appendix. Below are the features narrowed by this selection process.  
Adj R^2 : 0.91

CV Press : 21.79

Over all F test shows that the model is significant.



**Forward selection :** Here we have done forward selection with the below selected variables. Please refer code and plots from Appendix. Below are the features narrowed by this selection process.

Adj R^2 : 0.9218

CV Press : 20.83



Over all F test shows that the model is significant.

|  |  |  |  |
| --- | --- | --- | --- |
|  | CV PRESS | Adj R^2 | Kaggle Score |
| Forward | 20.8376 | 0.9218 | 0.63672 |
| Backward | 21.40984 | 0.9283 | 0.63923 |
| Stepwise | 21.79472 | 0.9121 | 0.13982 |
| Custom | 26.43956 | 0.8898 | 0.1599 |

# Conclusion

# After running the three models we decided to choose the step wise model as our custom model , The custom model and Step wise model have the highest Kaggle scores.

All the Kaggle scores have been placed in GitHub.

#### **APPENDIX**

VARIABLES SELECTED FOR BACKWARD, FORWARD AND STEP WISE

MSZoning\_num\_cat Street\_num\_cat Alley\_num\_cat LotShape\_num\_cat

LandContour\_num\_cat Utilities\_num\_cat LotConfig\_num\_cat LandSlope\_num\_cat

Neighborhood\_num\_cat Condition1\_num\_cat Condition2\_num\_cat BldgType\_num\_cat

HouseStyle\_num\_cat RoofStyle\_num\_cat Exterior2nd\_num\_cat MasVnrType\_num\_cat

ExterQual\_num\_cat ExterCond\_num\_cat Foundation\_num\_cat BsmtQual\_num\_cat

BsmtCond\_num\_cat BsmtExposure\_num\_cat BsmtFinType1\_num\_cat BsmtFinType2\_num\_cat

Heating\_num\_cat HeatingQC\_num\_cat CentralAir\_num\_cat Electrical\_num\_cat

KitchenQual\_num\_cat Functional\_num\_cat FireplaceQu\_num\_cat GarageType\_num\_cat

GarageFinish\_num\_cat PavedDrive\_num\_cat MiscFeature\_num\_cat SaleType\_num\_cat;

model logPrice = MSZoning\_num\_cat Street\_num\_cat Alley\_num\_cat LotShape\_num\_cat

LandContour\_num\_cat Utilities\_num\_cat LotConfig\_num\_cat LandSlope\_num\_cat

Neighborhood\_num\_cat Condition1\_num\_cat Condition2\_num\_cat BldgType\_num\_cat

HouseStyle\_num\_cat RoofStyle\_num\_cat Exterior2nd\_num\_cat MasVnrType\_num\_cat

ExterQual\_num\_cat ExterCond\_num\_cat Foundation\_num\_cat BsmtQual\_num\_cat

BsmtCond\_num\_cat BsmtExposure\_num\_cat BsmtFinType1\_num\_cat BsmtFinType2\_num\_cat

Heating\_num\_cat HeatingQC\_num\_cat CentralAir\_num\_cat Electrical\_num\_cat

KitchenQual\_num\_cat Functional\_num\_cat FireplaceQu\_num\_cat GarageType\_num\_cat

GarageFinish\_num\_cat PavedDrive\_num\_cat MiscFeature\_num\_cat SaleType\_num\_cat

SaleCondition\_num\_cat OverallCond YearRemodAdd YearBuilt BsmtUnfSF OpenPorchSF

GarageArea TotRmsAbvGrd LotArea WoodDeckSF GrLivArea EnclosedPorch FullBath

Fireplaces MasVnrArea OverallQual

**SAS Code Used for Analysis**

* **Code to import the data set**

FILENAME REFFILE '/home/u58342833/sasuser.v94/train\_cleaned.csv';

PROC IMPORT DATAFILE=REFFILE

DBMS=CSV

OUT=WORK.IMPORT;

GETNAMES=YES;

RUN;

PROC CONTENTS DATA=WORK.IMPORT; RUN;

* **Data cleaned and Filtered for the analysis question 1**

train\_sample1=data.frame(Id=selection$Id,SalePrice=selection$SalePrice,GrLivArea=selection$GrLivArea,Neighborhood=selection$Neighborhood)

target <- c("BrkSide", "NAmes","Edwards")

train\_sample2<-filter(train\_sample1, train\_sample1$Neighborhood %in% target)

train\_sample2<-contrasts(factor(train\_sample2$Neighborhood))

write.csv(train\_sample2,"C:/Users/pearl/OneDrive/Desktop/SMU-FOLDER/Statistics\_Foundations\_2021/Advanced\_Regression/train\_cleaned.csv")

/\* Generated Code (IMPORT) \*/

/\* Source File: test.csv \*/

/\* Source Path: /home/u58342833/sasuser.v94 \*/

/\* Code generated on: 12/7/21, 12:53 AM \*

%web\_drop\_table(WORK.IMPORT1);

FILENAME REFFILE '/home/u58342833/sasuser.v94/test.csv';

OUT=WORK.IMPORT1;

GETNAMES=YES;

RUN;

PROC CONTENTS DATA=WORK.IMPORT1; RUN;

%web\_open\_table(WORK.IMPORT1);

**DATA TABLE 1.1**

Table

Description automatically generated

**Initial Analysis:**

**The relationship between SalePrice and Living Room area is Linear. Sales Price increases as the Greater Living Area increases.**

Chart, scatter chart

Description automatically generated

**Plot 1.1**

title ‘Scatter plot: SalePrice vs Living Room Area’;

PROC sgplot DATA=WORK.IMPORT;

scatter x=GrLIvArea y=SalePrice;

run;

**MODEL 1.1:**

**The relationship between SalePrice and Living Room area.**

proc reg data = WORK.IMPORT plots=all;

model SalePrice = GrLivArea / r vif clb influence;

run;

Table

Description automatically generated

**Plot 1.2 :**

Chart, scatter chart

Description automatically generated

Diagram

Description automatically generated

**Plot 1.3**

**Removal of the 10 outliers which were identified with the below condition**

proc reg data = WORK.IMPORT plots=all;

model SalePrice = GrLivArea / r vif clb influence;

output out=Removal\_Outliers student=studresids cookd=cook h=leverage;

run;

proc print data= Removal\_Outliers;

run;

proc print data=Removal\_Outliers;

where cook>4;

run;

131,339

proc print data=Removal\_Outliers;

where leverage>0.04;

run;

proc print data=Removal\_Outliers;

where studresids<-2.5 OR studresids>2.5 ;

run;

data Removal\_Outliers;

set Removal\_Outliers;

if studresids<-2.5 OR studresids>2.5 then

delete;

run;

data Removal\_Outliers;

set Removal\_Outliers;

if cook>4 then

delete;

run;

data Removal\_Outliers;

set Removal\_Outliers;

if leverage>0.04 then

delete;

run;

proc reg data = Removal\_Outliers plots=all;

model SalePrice = GrLivArea / r vif clb influence;

run;

Chart, scatter chart

Description automatically generated

**Plot 1.4 – Residual Plot**

Diagram, engineering drawing

Description automatically generated with medium confidence

**Plot 1.5 – Model 2**

**After removing the outliers with high cooks distance and Leverage, We implemented the second model. Now this model**

**Predicted Sales Price = 0+1Living Room Area+2Living Room Area\*Dummy1+3Living Room Area\*Dummy2+**

**4\*Dummy2+5\*Dummy2**

**Plot 1.6**

Diagram

Description automatically generated

**Plot 1.7**

Chart, scatter chart

Description automatically generated

**Plot 1.8**

Chart, scatter chart

Description automatically generated

**Plot 1.9**

Diagram

Description automatically generated with medium confidence

**Plot 1.10**

Chart, scatter chart

Description automatically generated

**Plot 1.11**

Chart, scatter chart

Description automatically generated

**CODE 1.12**

proc print data=WORK.IMPORT;

run;

data WORK.IMPORT;

set WORK.IMPORT;

if ID='1299' then delete;

if ID='524' then delete;

if ID='643' then delete;

run;

proc glm data= WORK.IMPORT plots=all;

class Neighborhood;

model SalePrice= GrLivArea | Neighborhood/solution clm;

run;

data WORK.IMPORT;

set WORK.IMPORT;

lSalePrice=log(SalePrice);

lGrLivArea=log(GrLivArea);

run;

proc glm data= WORK.IMPORT plots=all;

class Neighborhood;

model lSalePrice= GrLivArea | Neighborhood/solution clm;

run;

proc glm data= WORK.IMPORT plots= all;

class Neighborhood;

model lSalePrice= GrLivArea | Neighborhood/solution clm clparm; ;

run;

**ANALYSIS 2 : CODE SCATTER PLOT TO SELECT NUMERICAL VARIABLES**

* **Based on the below code and the scatter plots we can say that Sales price has a linear relationship with following variables.**
* **Total there are 38 numerical predictors provided in the data set**

**/\*Id MSSubClass LotFrontage LotArea OverallQual OverallCond YearBuilt YearRemodAdd MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF X1stFlrSF X2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd Fireplaces GarageYrBlt GarageCars GarageArea WoodDeckSF OpenPorchSF EnclosedPorch X3SsnPorch ScreenPorch PoolArea MiscVal MoSold YrSold SalePrice \***

**/\*\* 17 numerical variables are selected as candidate predictors through checking scatter plots \*\*/ /\*\* ,** GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd Fireplaces GarageYrBlt GarageCars GarageArea WoodDeckSF OpenPorchSF EnclosedPorch X3SsnPorch ScreenPorch PoolArea MiscVal MoSold YrSold SalePrice **\*\*/**

/\*\*\* Code \*\*/

/\* 1: Variable selection with scatter plots \*/

/\* Linear Relationship with OverallQual SalePrice GarageCars GarageArea WoodDeckSF\*/

/\* Non Linear Relationship with SalePrice MSSubClass LotArea \*/

PROC sgscatter DATA=WORK.IMPORT2;

matrix SalePrice MSSubClass LotArea OverallQual ;

run;

/\* Linear Relationship with OverallQual\*/

/\* Linear Relationship with OverallQual\*/

/\* Non Linear Relationship with SalePrice MSSubClass LotArea \*/

PROC sgscatter DATA=WORK.IMPORT2;

matrix SalePrice OverallCond YearBuilt YearRemodAdd;

run;

/\* Non Linear Relationship with BsmtFullBath BsmtHalfBath \*/

PROC sgscatter DATA=WORK.IMPORT2;

matrix SalePrice BsmtFullBath BsmtHalfBath ;

run;

/\* Non Linear Relationship with BedroomAbvGr \*/

PROC sgscatter DATA=WORK.IMPORT2;

matrix SalePrice FullBath HalfBath BedroomAbvGr KitchenAbvGr;

run;

PROC sgscatter DATA=WORK.IMPORT2;

matrix SalePrice TotRmsAbvGrd Fireplaces

run;

PROC sgscatter DATA=WORK.IMPORT2;

matrix SalePrice GarageCars GarageArea WoodDeckSF;

run;

**Plot 2.1:**

A picture containing diagram

Description automatically generated

**Plot 2.2:**

Word

Description automatically generated with medium confidence

**Plot 2.3:**

Calendar

Description automatically generated

**Plot 2.4**

Calendar

Description automatically generated

**Plot 2.5**

Graphical user interface, diagram, application, Word

Description automatically generated

**Plot 2.6**

Calendar

Description automatically generated

**Plot 2.7:**

A picture containing diagram

Description automatically generated

**Now we totally identified the the variables to be used so we remove the outliers that we identified**

set union\_all;

/\* this had an exceptionally high sales price \*/

if id = 826 then delete;

/\* this is an exceptionally large home \*/

if id = 524 then delete;

run;

title "Cleaning Data For Regression - All the categorical values have been made numerical for analysis ";

data WORK.IMPORT;

set WORK.IMPORT;

if MSZoning="C" then MSZoning\_num\_cat = 0;

else if MSZoning="FV" then MSZoning\_num\_cat = 1;

else if MSZoning="RH" then MSZoning\_num\_cat = 2;

else if MSZoning="RL" then MSZoning\_num\_cat = 3;

else MSZoning\_num\_cat = 4;

if LotFrontage="NA" then LotFrontage\_num\_cat = 0;

else LotFrontage\_num\_cat = LotFrontage;

if Street = "Pave" then Street\_num\_cat = 0;

else Street\_num\_cat = 1;

if Alley = "Pa" then Alley\_num\_cat = 1;

else if Alley = "Gr" then Alley\_num\_cat = 2;

else Alley\_num\_cat = 0;

if LotShape="IR1" then LotShape\_num\_cat = 3;

else if LotShape = "IR2" then LotShape\_num\_cat = 2;

else if LotShape = "IR3" then LotShape\_num\_cat = 1;

else LotShape\_num\_cat = 0;

if LandContour = "Bnk" then LandContour\_num\_cat = 0;

else if LandContour = "HLS" then LandContour\_num\_cat = 1;

else if LandContour = "Low" then LandContour\_num\_cat = 2;

else LandContour\_num\_cat = 3;

if Utilities = "AllPub" then Utilities\_num\_cat = 1;

else Utilities\_num\_cat = 0;

if LotConfig = "Corner" then LotConfig\_num\_cat = 0;

else if LotConfig = "CulDSac" then LotConfig\_num\_cat = 1;

else if LotConfig = "FR2" then LotConfig\_num\_cat = 2;

else if LotConfig = "FR3" then LotConfig\_num\_cat = 3;

else LotConfig\_num\_cat = 4;

if LandSlope= "Gtl" then LandSlope\_num\_cat = 0;

else if LandSlope= "Mod" then LandSlope\_num\_cat = 1;

else LandSlope\_num\_cat = 2;

else if Condition1="PosN" then Condition1\_num\_cat = 4;

else if Condition1="RRAe" then Condition1\_num\_cat = 5;

else if Condition1="RRAn" then Condition1\_num\_cat = 6;

else if Condition1="RRNe" then Condition1\_num\_cat = 7;

else Condition1\_num\_cat = 8;

if Condition2="Norm" then Condition2\_num\_cat = 0;

else if Condition2="Artery" then Condition2\_num\_cat = 1;

else if Condition2="Feedr" then Condition2\_num\_cat = 2;

else if Condition2="PosA" then Condition2\_num\_cat = 3;

else if Condition2="PosN" then Condition2\_num\_cat = 4;

else if Condition2="RRAe" then Condition2\_num\_cat = 5;

else if Condition2="RRAn" then Condition2\_num\_cat = 6;

else if Condition2="RRNe" then Condition2\_num\_cat = 7;

else Condition2\_num\_cat = 8;

if BldgType="1Fam" then BldgType\_num\_cat = 0;

else if BldgType="2fmCon" then BldgType\_num\_cat = 1;

else if BldgType="Duplex" then BldgType\_num\_cat = 2;

else if BldgType="Twnhs" then BldgType\_num\_cat = 3;

else BldgType\_num\_cat = 4;

if HouseStyle="1.5Fin" then HouseStyle\_num\_cat = 0;

else if HouseStyle="1.5Unf" then HouseStyle\_num\_cat = 1;

else if HouseStyle="1Story" then HouseStyle\_num\_cat = 2;

else if HouseStyle="2.5Fin" then HouseStyle\_num\_cat = 3;

else if HouseStyle="2.5Unf" then HouseStyle\_num\_cat = 4;

else if HouseStyle="2Story" then HouseStyle\_num\_cat = 5;

else if HouseStyle="SFoyer" then HouseStyle\_num\_cat = 6;

else HouseStyle\_num\_cat = 7;

if RoofStyle="Flat" then RoofStyle\_num\_cat = 0;

else if RoofStyle="Gable" then RoofStyle\_num\_cat = 1;

else if RoofStyle="Gambrel" then RoofStyle\_num\_cat = 2;

else if RoofStyle="Hip" then RoofStyle\_num\_cat = 3;

else if RoofStyle="Mansard" then RoofStyle\_num\_cat = 4;

else RoofStyle\_num\_cat = 5;

if RoofMatl="CompShg" then RoofMatl\_num\_cat = 1;

else if RoofMatl="Metal" then RoofMatl\_num\_cat = 2;

else if RoofMatl="WdShngl" then RoofMatl\_num\_cat = 3;

else if RoofMatl="WdShake" then RoofMatl\_num\_cat = 4;

else if RoofMatl="ClyTile" then RoofMatl\_num\_cat = 5;

else if RoofMatl="Roll" then RoofMatl\_num\_cat = 6;

else if RoofMatl="Membran" then RoofMatl\_num\_cat = 7;

else RoofMatl\_num\_cat = 0;

**CODE 2.2**

proc print data=WORK.IMPORT1;

run;

proc print data=WORK.IMPORT3;

run;

data WORK.IMPORT3;

set WORK.IMPORT3;

SalePrice=.;

run;

data train2;

set WORK.IMPORT1 WORK.IMPORT3;

run;

data train2;

set train2;

if ID='1299' then delete;

if ID='524' then delete;

if ID='643' then delete;

run;

Be

SELECTING VARIABLES BASED ON THE VIF, R^2 and SIGNIFICANCE P VALUE

proc reg data=Train2;

model SalePrice =

MSZoning\_num\_cat

Street\_num\_catAlley\_num\_cat

LotShape\_num\_cat

LandContour\_num\_cat

Utilities\_num\_cat

LotConfig\_num\_cat

LandSlope\_num\_cat

Neighborhood\_num\_cat

Condition1\_num\_cat

Condition2\_num\_cat

BldgType\_num\_cat

HouseStyle\_num\_cat

RoofStyle\_num\_cat

RoofMatl\_num\_cat

Exterior1st\_num\_cat

Exterior2nd\_num\_cat

MasVnrType\_num\_cat

ExterQual\_num\_cat

ExterCond\_num\_cat

Foundation\_num\_cat

BsmtQual\_num\_cat

BsmtCond\_num\_cat

BsmtExposure\_num\_cat

BsmtFinType1\_num\_cat

BsmtFinType2\_num\_cat

Heating\_num\_cat

HeatingQC\_num\_cat

CentralAir\_num\_cat

Electrical\_num\_catKitchenQual\_num\_catFunctional\_num\_catFireplaceQu\_num\_catGarageType\_num\_catGarageFinish\_num\_cat GarageQual\_num\_cat

GarageCond\_num\_catPavedDrive\_num\_catPoolQC\_num\_catFence\_num\_cat

MiscFeature\_num\_catSaleType\_num\_cat

SaleCondition\_num\_cat

OverallCond

YearRemodAdd

YearBuilt

BsmtUnfSF

OpenPorchSF

GarageArea

TotRmsAbvGrd LotArea

WoodDeckSF

GrLivArea

EnclosedPorch FullBath Fireplaces

MasVnrArea

OverallQual

GarageCars

/ stb VIF ss1 ss2 selection=stepwise slentry=0.1 slstay=0.1;

OUTPUT OUT = reg\_House\_price PREDICTED=PRCDT RESIDUAL = h\_Res

L95M = h\_l95m U95M = h\_u95m L95 =h\_l95 U95 = h\_u95

rstudent = h\_rstudent h = lev cookd = Cookd dffits = dffit

STDP = h\_spredicted STDR = h\_s\_residual STUDENT = h\_student;

quit;

Table

Description automatically generated

**Plot 2.8**

Diagram, engineering drawing

Description automatically generated

**CUSTOM SELECTION :**

proc glmselect data=train2;

class LotConfig\_num\_cat

Neighborhood\_num\_cat ExterQual\_num\_cat BsmtQual\_num\_cat HeatingQC\_num\_cat

KitchenQual\_num\_cat FireplaceQu\_num\_cat

GarageQual\_num\_cat PoolQC\_num\_cat SaleCondition\_num\_cat;

model logPrice = GarageCars

OverallCond

YearRemodAdd

YearBuilt

BsmtUnfSF

OpenPorchSF

GarageArea

TotRmsAbvGrd

LotArea

WoodDeckSF

GrLivArea

EnclosedPorch

FullBath

Fireplaces

MasVnrArea

OverallQual | Neighborhood\_num\_cat /selection=stepwise(stop=cv) cvmethod=random(5) stats=adjrsq;

output out= results p= predict;

run;

data results;

set results;

if predict>0 then SalePrice=exp(predict);

if predict<0 then SalePrice=10000;

keep id SalePrice;

where id>1460;

proc export

data=results

dbms=csv

outfile="/home/u58342833/sasuser.v94/Exam/submission\_result.csv"

replace;

run;

Graphical user interface, table

Description automatically generated

**Plot 2,15**

**COCODE**

**BACKWARD ELIMINATION:**

proc glmselect data=train2;

class MSZoning\_num\_cat Street\_num\_cat Alley\_num\_cat LotShape\_num\_cat

LandContour\_num\_cat Utilities\_num\_cat LotConfig\_num\_cat LandSlope\_num\_cat

Neighborhood\_num\_cat Condition1\_num\_cat Condition2\_num\_cat BldgType\_num\_cat

HouseStyle\_num\_cat RoofStyle\_num\_cat Exterior2nd\_num\_cat MasVnrType\_num\_cat

ExterQual\_num\_cat ExterCond\_num\_cat Foundation\_num\_cat BsmtQual\_num\_cat

BsmtCond\_num\_cat BsmtExposure\_num\_cat BsmtFinType1\_num\_cat BsmtFinType2\_num\_cat

Heating\_num\_cat HeatingQC\_num\_cat CentralAir\_num\_cat Electrical\_num\_cat

KitchenQual\_num\_cat Functional\_num\_cat FireplaceQu\_num\_cat GarageType\_num\_cat

GarageFinish\_num\_cat PavedDrive\_num\_cat MiscFeature\_num\_cat SaleType\_num\_cat;

model logPrice = MSZoning\_num\_cat Street\_num\_cat Alley\_num\_cat LotShape\_num\_cat

LandContour\_num\_cat Utilities\_num\_cat LotConfig\_num\_cat LandSlope\_num\_cat

Neighborhood\_num\_cat Condition1\_num\_cat Condition2\_num\_cat BldgType\_num\_cat

HouseStyle\_num\_cat RoofStyle\_num\_cat Exterior2nd\_num\_cat MasVnrType\_num\_cat

ExterQual\_num\_cat ExterCond\_num\_cat Foundation\_num\_cat BsmtQual\_num\_cat

BsmtCond\_num\_cat BsmtExposure\_num\_cat BsmtFinType1\_num\_cat BsmtFinType2\_num\_cat

Heating\_num\_cat HeatingQC\_num\_cat CentralAir\_num\_cat Electrical\_num\_cat

KitchenQual\_num\_cat Functional\_num\_cat FireplaceQu\_num\_cat GarageType\_num\_cat

GarageFinish\_num\_cat PavedDrive\_num\_cat MiscFeature\_num\_cat SaleType\_num\_cat

SaleCondition\_num\_cat OverallCond YearRemodAdd YearBuilt BsmtUnfSF OpenPorchSF

GarageArea TotRmsAbvGrd LotArea WoodDeckSF GrLivArea EnclosedPorch FullBath

Fireplaces MasVnrArea OverallQual / selection=backward slentry=0.1 slstay=0.1;

output out=backward\_result p=Predict4;

run;

/\*\* To predict the result of the Backward elimination \*/

data backward\_result;

set backward\_result;

if Predict4>0 then SalePrice=exp(Predict4);

if Predict4<0 then SalePrice=10000;

keep id SalePrice;

where id>1460;

/\*\* Export to csv\*\*/

proc export

data=backward\_result

dbms=csv

outfile="/home/u58342833/sasuser.v94/Exam/submission\_backward\_result.csv"

replace;

run;

Table

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

**FORWARD ELIMINATION:**

proc glmselect data=train2;

class MSZoning\_num\_cat Street\_num\_cat Alley\_num\_cat LotShape\_num\_cat

LandContour\_num\_cat Utilities\_num\_cat LotConfig\_num\_cat LandSlope\_num\_cat

Neighborhood\_num\_cat Condition1\_num\_cat Condition2\_num\_cat BldgType\_num\_cat

HouseStyle\_num\_cat RoofStyle\_num\_cat Exterior2nd\_num\_cat MasVnrType\_num\_cat

ExterQual\_num\_cat ExterCond\_num\_cat Foundation\_num\_cat BsmtQual\_num\_cat

BsmtCond\_num\_cat BsmtExposure\_num\_cat BsmtFinType1\_num\_cat BsmtFinType2\_num\_cat

Heating\_num\_cat HeatingQC\_num\_cat CentralAir\_num\_cat Electrical\_num\_cat

KitchenQual\_num\_cat Functional\_num\_cat FireplaceQu\_num\_cat GarageType\_num\_cat

GarageFinish\_num\_cat PavedDrive\_num\_cat MiscFeature\_num\_cat SaleType\_num\_cat;

model logPrice = MSZoning\_num\_cat Street\_num\_cat Alley\_num\_cat LotShape\_num\_cat

LandContour\_num\_cat Utilities\_num\_cat LotConfig\_num\_cat LandSlope\_num\_cat

Neighborhood\_num\_cat Condition1\_num\_cat Condition2\_num\_cat BldgType\_num\_cat

HouseStyle\_num\_cat RoofStyle\_num\_cat Exterior2nd\_num\_cat MasVnrType\_num\_cat

ExterQual\_num\_cat ExterCond\_num\_cat Foundation\_num\_cat BsmtQual\_num\_cat

BsmtCond\_num\_cat BsmtExposure\_num\_cat BsmtFinType1\_num\_cat BsmtFinType2\_num\_cat

Heating\_num\_cat HeatingQC\_num\_cat CentralAir\_num\_cat Electrical\_num\_cat

KitchenQual\_num\_cat Functional\_num\_cat FireplaceQu\_num\_cat GarageType\_num\_cat

GarageFinish\_num\_cat PavedDrive\_num\_cat MiscFeature\_num\_cat SaleType\_num\_cat

SaleCondition\_num\_cat OverallCond YearRemodAdd YearBuilt BsmtUnfSF OpenPorchSF

GarageArea TotRmsAbvGrd LotArea WoodDeckSF GrLivArea EnclosedPorch FullBath

Fireplaces MasVnrArea OverallQual / selection=backward slentry=0.1 slstay=0.1;

output out=backward\_result p=Predict4;

run;

/\*\* To predict the result of the Backward elimination \*/

data forward\_result;

set forward\_result;

if Predict4>0 then SalePrice=exp(Predict4);

if Predict4<0 then SalePrice=10000;

keep id SalePrice;

where id>1460;

/\*\* Export to csv\*\*/

proc export

data= rward\_result;

dbms=csv

outfile="/home/u58342833/sasuser.v94/Exam/submission\_backward\_result.csv"

replace;

run;

Table

Description automatically generated

Graphical user interface, table

Description automatically generated

/\*\* STEP WISE ELIMINATION\*\*/

proc glmselect data=train2;

class MSZoning\_num\_cat Street\_num\_cat Alley\_num\_cat LotShape\_num\_cat

LandContour\_num\_cat Utilities\_num\_cat LotConfig\_num\_cat LandSlope\_num\_cat

Neighborhood\_num\_cat Condition1\_num\_cat Condition2\_num\_cat BldgType\_num\_cat

HouseStyle\_num\_cat RoofStyle\_num\_cat Exterior2nd\_num\_cat MasVnrType\_num\_cat

ExterQual\_num\_cat ExterCond\_num\_cat Foundation\_num\_cat BsmtQual\_num\_cat

BsmtCond\_num\_cat BsmtExposure\_num\_cat BsmtFinType1\_num\_cat BsmtFinType2\_num\_cat

Heating\_num\_cat HeatingQC\_num\_cat CentralAir\_num\_cat Electrical\_num\_cat

KitchenQual\_num\_cat Functional\_num\_cat FireplaceQu\_num\_cat GarageType\_num\_cat

GarageFinish\_num\_cat PavedDrive\_num\_cat MiscFeature\_num\_cat SaleType\_num\_cat;

model logPrice = MSZoning\_num\_cat Street\_num\_cat Alley\_num\_cat LotShape\_num\_cat

LandContour\_num\_cat Utilities\_num\_cat LotConfig\_num\_cat LandSlope\_num\_cat

Neighborhood\_num\_cat Condition1\_num\_cat Condition2\_num\_cat BldgType\_num\_cat

HouseStyle\_num\_cat RoofStyle\_num\_cat Exterior2nd\_num\_cat MasVnrType\_num\_cat

ExterQual\_num\_cat ExterCond\_num\_cat Foundation\_num\_cat BsmtQual\_num\_cat

BsmtCond\_num\_cat BsmtExposure\_num\_cat BsmtFinType1\_num\_cat BsmtFinType2\_num\_cat

Heating\_num\_cat HeatingQC\_num\_cat CentralAir\_num\_cat Electrical\_num\_cat

KitchenQual\_num\_cat Functional\_num\_cat FireplaceQu\_num\_cat GarageType\_num\_cat

GarageFinish\_num\_cat PavedDrive\_num\_cat MiscFeature\_num\_cat SaleType\_num\_cat

SaleCondition\_num\_cat OverallCond YearRemodAdd YearBuilt BsmtUnfSF OpenPorchSF

GarageArea TotRmsAbvGrd LotArea WoodDeckSF GrLivArea EnclosedPorch FullBath

Fireplaces MasVnrArea OverallQual / selection=stepwise (stop=cv) slentry=0.1 slstay=0.1;

output out=stepwise\_result p=Predict4;

run;

/\*\* To predict the result of the Backward elimination \*/

data forward\_result;

set forward\_result;

if Predict4>0 then SalePrice=exp(Predict4);

if Predict4<0 then SalePrice=10000;

keep id SalePrice;

where id>1460;

/\*\* Export to csv\*\*/

proc export

data= rward\_result;

dbms=csv

outfile="/home/u58342833/sasuser.v94/Exam/submission\_backward\_result.csv"

replace;

run;

Graphical user interface, application, table

Description automatically generated

Table

Description automatically generated

Graphical user interface, table

Description automatically generated

BACKWARD \_ELIMINATION

proc glmselect data=train2;   
class MSZoning\_num\_cat Street\_num\_cat Alley\_num\_cat LotShape\_num\_cat   
LandContour\_num\_cat Utilities\_num\_cat LotConfig\_num\_cat LandSlope\_num\_cat   
Neighborhood\_num\_cat Condition1\_num\_cat Condition2\_num\_cat BldgType\_num\_cat   
HouseStyle\_num\_cat RoofStyle\_num\_cat Exterior2nd\_num\_cat MasVnrType\_num\_cat   
ExterQual\_num\_cat ExterCond\_num\_cat Foundation\_num\_cat BsmtQual\_num\_cat   
BsmtCond\_num\_cat BsmtExposure\_num\_cat BsmtFinType1\_num\_cat BsmtFinType2\_num\_cat   
Heating\_num\_cat HeatingQC\_num\_cat CentralAir\_num\_cat Electrical\_num\_cat   
KitchenQual\_num\_cat Functional\_num\_cat FireplaceQu\_num\_cat GarageType\_num\_cat   
GarageFinish\_num\_cat PavedDrive\_num\_cat MiscFeature\_num\_cat SaleType\_num\_cat;  
model logPrice = MSZoning\_num\_cat Street\_num\_cat Alley\_num\_cat LotShape\_num\_cat   
LandContour\_num\_cat Utilities\_num\_cat LotConfig\_num\_cat LandSlope\_num\_cat   
Neighborhood\_num\_cat Condition1\_num\_cat Condition2\_num\_cat BldgType\_num\_cat   
HouseStyle\_num\_cat RoofStyle\_num\_cat Exterior2nd\_num\_cat MasVnrType\_num\_cat   
ExterQual\_num\_cat ExterCond\_num\_cat Foundation\_num\_cat BsmtQual\_num\_cat   
BsmtCond\_num\_cat BsmtExposure\_num\_cat BsmtFinType1\_num\_cat BsmtFinType2\_num\_cat   
Heating\_num\_cat HeatingQC\_num\_cat CentralAir\_num\_cat Electrical\_num\_cat   
KitchenQual\_num\_cat Functional\_num\_cat FireplaceQu\_num\_cat GarageType\_num\_cat   
GarageFinish\_num\_cat PavedDrive\_num\_cat MiscFeature\_num\_cat SaleType\_num\_cat   
SaleCondition\_num\_cat OverallCond YearRemodAdd YearBuilt BsmtUnfSF OpenPorchSF   
GarageArea TotRmsAbvGrd LotArea WoodDeckSF GrLivArea EnclosedPorch FullBath   
Fireplaces MasVnrArea OverallQual / selection=backward (stop=cv) slentry=0.1 slstay=0.1;   
output out=backward\_result p=Predict4;  
run;

/\*\* To predict the result of the Backward elimination \*/

data forward\_result;

set forward\_result;

if Predict4>0 then SalePrice=exp(Predict4);

if Predict4<0 then SalePrice=10000;

keep id SalePrice;

where id>1460;

Table

Description automatically generated

Graphical user interface, application, table

Description automatically generated

Table

Description automatically generated