KAGGLE PROJECT (MSDS 6371)

H. H. Nguyen, I. Nwaogu and H. Wang

I. INTRODUCTION

- Kaggle is an online platform for data scientist communities, it is also a place where datasets are explored in order to build models that can serve the community.

- When we ask a home buyer to describe their dream house, and he/she probably will not begin with the height of the basement ceiling or the proximity to an east-west railroad. However, this playground competition's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence.

- In this project [1], we will use 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, to predict the final price of each home by using different methods in order to choose the best model. We will practice feature engineering and regression algorithms to achieve the lowest prediction error.

II. DATA DESCRIPTION

- The Ames Housing dataset was compiled by Dean De Cock [2] describing the sale of individual residential property in Ames, Iowa from 2006 to 2010. The data set contains 2930 observations to give us an opportunity to study a large number of explanatory variables included 23 nominal, 23 ordinal, 14 discrete, and 20 continuous variables that involved in assessing home values. For the analysis, we focus on the relationship between the square footage of the living area of the house and sale price in 3 neighborhoods. For the second analysis, we will consider all neighborhoods, 28 ones, exactly.

III. ANALYSIS OF QUESTION 1

A. Restatement of Problem

- The real estate company "Century 21 Ames" (Ames – Iowa) has commissioned us to analyze the Ames Housing dataset [2] with respect to its business. The company only sells houses in the NAmes, Edwards and BrkSide neighborhoods. Based on its interests related to three specific neighborhoods, we will estimate how square footage relates to the sale and if the sale price (and its relationship to square footage) depends on which neighborhood the house is located in. In order to compete this analysis, we will restrict our model to only focus on these neighborhoods' variables.

B. Build and Fit the Model

- In the first step, by examining the Scatter plot SalePrice vs. GrLIvArea [*see Appendix, Plot 1.3*], it looks like there is a linear relationship between the living area and the sale price. However, there

are some outliers. We need to further check with some tentative models and decide how to deal with these outliers.

We assume here that the observations are independent.

1. We now build a first tentative model:

$$\mu(\text{SalePrice}) = b_0 + b_1(\text{GrLIvArea}). \quad (1)$$

From Plot 1.5 [Appendix], we can see

- Four outliers with studentized residuals larger than ± 2.5 found.
- One outlier with highest Cook's D > 5.
- Adj R-Square = 0.3406.
- Judging from the scatter plot, QQ plot and histogram of the residuals, there is no evidence that the residuals do no follow a normal distribution with constant variance.

Because we only have to deal with 4 outliers from 383 observations, we can remove the outliers and continue with our testing.

2. We rerun the same model (1) without these four outliers:
From Plot 1.10 and Plot 1.12 [*Appendix*], we can see that

- Scatter plots indicate random distributed residuals.
- Cook's D values lower than 0.10.
- Straight line in QQ plot and symmetric shape of histogram indicate the normal distribution.
- Adj R-Square = 0.449.

3. In this step, we will build another tentative model with Neighborhood variables:

$$\mu(\text{SalePrice}) = b_0 + b_1(\text{GrLIvArea}) + b_2(d1)+b_3(d2) + b_4(d1*\text{GrLivArea}) + b_5(d2*\text{GrLivArea}). \quad (2)$$

When we are dealing with categorical variables, d1 = 1 when neighborhood = "NAmes" otherwise d1=0 then d2 =1 when neighborhood = "BrkSide".
From Plot 1.10 and Plot 1.19 [Appendix], we can see that

- Scatter plots indicate random distributed residuals.
- Cook's D values lower than 0.20.
- Straight line in QQ plot and symmetric shape of histogram indicate the normal distribution.
- Adj R-Square = 0.5165.

- We will use this model to fit the data from 3 neighborhoods. We will assume the observations are independent.

C. Analysis

- After the last step, we have decided to choose the model (2) for our analysis. By observing the scatter plots 1.21, 1.22 and 1.23 [Appendix], there is no evidence that a transformation is necessary.

- By using Table 1.16 [Appendix] from out SAS code, we will rewrite the model (2) as follows

$$\mu(\text{SalePrice}) = 37100.4 + 70.2(\text{GrLivArea}) + 43225.3(d1) - 17128.9(d2) - 20.6(d1*\text{GrLivArea}) + 17(d2*\text{GrLivArea}). \quad (3)$$

- When we analyze the residual plots [see Appendix, Plot 1.18], we see
  - Constant SD: The scatter plots of residuals indicate a random distribution around line zero, no pattern found.
  - Normality and zero mean: Both QQ plot and histogram indicate normal distribution of the residual.
  - Identify Any Influential Observations: No significant outliers in residual plots.

- We can simplify the model (3) by separating it by each neighborhood:

- Regression Model for NAmes (d1=1, d2=0):
$$\mu(\text{SalePrice}) = 80325.7 + 49.6*\text{GrLivArea}. \quad (4)$$
- Regression Model for BrkSide (d1=0, d2=1):
$$\mu(\text{SalePrice}) = 19971.5 + 87.2*\text{GrLivArea}. \quad (5)$$
- Regression Model for Edwards (d1=0, d2=0):
$$\mu(\text{SalePrice}) = 37100.4 + 70.2*\text{GrLivArea}. \quad (6)$$

D. Conclusion and Interpretation

- This model is a good fit to the data set of the three neighborhoods, p-value < .0001 for F-test with df(5, 373). 52.2% of the variability of Sale Price can be explained by the living area of the house 'GrLivArea'. Neighborhood Edwards has the highest estimated mean of sale price followed by BrkSide and NAmes.
  - In neighborhood NAmes, every 100 sq. ft living area increase results in an estimated $4,960 increase on sale price, with 95% confidence interval from $1,945 to $7,967.
  - In neighborhood BrkSide, every 100 sq. ft 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. ft living area increase results in an estimated $7,020 increase on sale price, with 95% confidence interval from $5,618 to $8,413, with p-value <0.0001 in t-Test.

- Because it is an observational study, we cannot make any causal inference here. However, there is a correlation between sale prices, square footage and these three neighborhoods although there are many confounding variables in the dataset.

IV. ANALYSIS OF QUESTION 2

A. Restatement of Problem

- In this analysis, we will build a predictive model for sale prices of individual residential property in all neighborhoods in Ames, Iowa.

- Our method is to use multiple linear regression to evaluate and analyze all variables in the dataset in order to get a good model.

- We will use Stepwise, Forward, Backward and Custom process selection in our analysis. We will also compare the parameters of the different models over the adjusted R^2, internal CV Press and Kaggle Score.

B. Pre-processing data

- First, we assume all observations in the dataset are independent.

- We will prepare the data by selecting numerical variables as predictor. By removing variables with NA's (eg. LotFrontage, MasVnrArea, GarageYrBlt) and variables with wrong name (eg. 1stFlrSF, 2ndFlrSF, 3SsnPorch), we have 31 numerical variables as predictors.

- Plot 2.2 [Appendix], there are linear relationship between SalePrice and OverallQual (OverallCond, YearBuilt, YearRemodAdd, respectively). So we would use these variables for our analysis. Similarly, according to the scatter plots 2.3 – 2.6 [Appendix], we sort the following 12 out of 29 numerical variables as candidate predictors for SalePrice: OverallQual, OverallCond, YearBuilt, YearRemodAdd, GrLivArea, FullBath, TotRmsAbvGrd, Fireplaces, GarageCars, GarageArea, WoodDeckSF, OpenPorchSF.

- Base on the SME's Output, we also select 13 categorical variables as candidates: Utilities, LotConfig, Neighborhood, ExterQual, BsmtQual, HeatingQC, KitchenQual, FireplaceQu, GarageQual, PoolQC, Fence, SaleType, SaleCondition.

- We will build model starting from numerical variables, target to use 3-5 numerical variables, then check with each categorical variable to find good fit.

- Finally, we remove 4 outliers of the dataset before building and fitting models.

C. Build models

1. Forward Selection (proc reg): (see Appendix for SAS codes)

Here we run the Forward Selection process through proc reg using the data in the last step (Pre-processing data).

- We have 12 Numeric Predictors: OverallQual, OverallCond, YearBuilt,, YearRemodAdd GrLivArea, FullBath, TotRmsAbvGrd, Fireplaces, GarageCars,, GarageArea WoodDeckSF, OpenPorchSF. Please refer to Table 2.19 and Plot 2.20 [Appendix], the figure shows us

- Top 4 variables with the highest Partial R^2 will be selected in the model (OverallQual, GrLivArea, GarageArea, YearBuilt).
- After 11 steps, Adj R-Square = 0.7996.
- Scatter plots indicate random distributed residuals.
- Cook's D values lower than 0.20.
- A small violation with the normality from QQ plot and histogram.
- Constant variance: We do not see a large concern with variance.
- The scatter plots of residuals which indicate a random distribution around line zero, no pattern found.
- Identify Any Influential Observations: No significant outliers in residual plots.

- Next we will run a Model with Top 4 Numeric Predictors. Looking atTable 2.23, Plot 2.24 and Plot 2.25, we have the similar observations results as in the previous model. Here Adj R-Square = 0.7806 after 4 steps. There is curve relationship between SalePrice and OverallQual then we will try log transformation on OverallQual in the next model.

- With the same model with the top 4 Numeric Predictors, after transforming the data, the Adj R-Square is 0.821. Residual scatter plots look much better, straight line in QQ plot looks linear [see Table 2.26, Plot 2.27, Plot 2.28 – Appendix].

Our regression model is

$$\mu(\text{logSalePrice}) = b_0 + b_1(\text{logOverallQual}) + b_2(\text{GrLivArea}) + b_3(\text{YearBuilt}) + b_4(\text{GarageArea}) \quad (7)$$

with $b_0 = 4.34765$, $b_1 = 0.57644$, $b_2 = 0.00031753$, $b_3 = 0.00033580$ and $b_4 = 0.00305$.

2. Backward Selection (proc reg): (see Appendix for SAS codes)

We will apply the Backward Selection in this model using same data from previous.

- We will use 12 numerical variables as Predictors for this selection type. Please refer to Table 2.29, Plots 2.30-2.32 [Appendix], we have similar observation results as the Forward Selection Method with 12 Numeric Predictors (Cook's D, Constant Variance, Normality...). By running SAS program, we eliminate 2 variables and the Adj R-Square = 0.7993.

Then our regression model is

$$\mu(\text{SalePrice}) = b_0 + b_1(\text{OverallQual}) + b_2(\text{OverallCond}) + b_3(\text{YearBuilt}) + b_4(\text{YearRemodAdd}) + b_5(\text{GrLivArea}) + b_6(\text{FullBath}) + b_7(\text{TotRmsAbvGrd}) + b_8(\text{Fireplaces}) + b_9(\text{GarageArea}) + b_{10}(\text{WoodDeckSF}) \quad (8)$$

with $b_0 = -1368572$, $b_1 = 18928$, $b_2 = 4392.2$, $b_3 = 483.5$, $b_4 = 173.34$, $b_5 = 74.36$, $b_6 = -10289$, $b_7 = -2989.37$, $b_8 = 9977.1$, $b_9 = 56.26$ and $b_{10} = 33.48$.

3. Stepwise Selection (proc reg): (see Appendix for SAS codes)

In this part, we use the Stepwise Selection method through proc reg using the data from the last step (Pre-processing data).

- We build a model using Stepwise method with 12 Numeric Predictors. We have similar observations from the Scatter plots and Histogram of Residuals, about Normality or Cook's D or Constance Variance. After 12 steps, the Adj R-Square is 0.7993 [see Plot 2.36 – Appendix].

- Next, we build a model with Top 4 Numeric Predictors and we do a log transformation on SalePrice and OverallQual. By running SAS program, the analysis shows an Adj R-Square = 0.821. Residual scatter plots looks much better, QQ plot is linear and the Histogram indicates that normality holds. Cook's D values lower than 0.125. [see Table 2.38 and Plot 2.39 – Appendix].

Our regression model is

$$\mu(\text{logSalePrice}) = b_0 + b_1(\text{logOverallQual}) + b_2(\text{GrLivArea}) + b_3(\text{YearBuilt}) + b_4(\text{GarageArea}) \quad (9)$$

with $b_0 = 4.34765$, $b_1 = 0.57644$, $b_2 = 0.00031753$, $b_3 = 0.00033580$ and $b_4 = 0.00305$.

Next, we will customize a model with categorical variables added.

4. Custom Model:

Next, we will customize a model with categorical variables added. (see Appendix for SAS codes)

- Forward Selection and Cross Validation (CV) (proc glmselect):

```
proc glmselect data=train2Ames3NoOutlier;
model logSalePrice = logOverallQual | GrLivArea GarageArea | YearBuilt | OverallCond | YearRemodAdd
 FullBath TotRmsAbvGrd Fireplaces GarageCars WoodDeckSF OpenPorchSF LotConfig_num
 Neighborhood_num ExterQual_num BsmtQual_num HeatingQC_num KitchenQual_num FireplaceQu_num
 GarageQual_num PoolQC_num SaleCondition_num
/selection=Forward(stop=cv) cvmethod=random(5) stats=adjrsq;
output out= results p= predict;
run;
```

- Adj R-Square = 0.8665 after 12 steps

- CV PRESS = 31.55516
- 12 parameters are selected to be used in model
- Kaggle score: 0.15450

- Backward selection with Cross Validation (CV) (proc glmselect):

```
proc glmselect data=train2Ames3NoOutlier;
model logSalePrice = logOverallQual | GrLivArea GarageArea | YearBuilt | OverallCond | YearRemodAdd
 FullBath TotRmsAbvGrd Fireplaces GarageCars WoodDeckSF OpenPorchSF LotConfig_num
 Neighborhood_num ExterQual_num BsmtQual_num HeatingQC_num KitchenQual_num FireplaceQu_num
 GarageQual_num PoolQC_num SaleCondition_num
/selection=Backward(stop=cv) cvmethod=random(5) stats=adjrsq;
output out= results p= predict;
run;
```

- Adj R-Square = 0.8719 after 9 steps
- CV PRESS = 30.55471
- 23 parameters are selected to be used in model
- Kaggle score: 0.15149

- Stepwise selection with Cross Validation (CV) (proc glmselect):

```
proc glmselect data=train2Ames3NoOutlier;
model logSalePrice = logOverallQual | GrLivArea GarageArea | YearBuilt | OverallCond | YearRemodAdd
 FullBath TotRmsAbvGrd Fireplaces GarageCars WoodDeckSF OpenPorchSF LotConfig_num
 Neighborhood_num ExterQual_num BsmtQual_num HeatingQC_num KitchenQual_num FireplaceQu_num
 GarageQual_num PoolQC_num SaleCondition_num
/selection=Stepwise(stop=cv) cvmethod=random(5) stats=adjrsq;
output out= results p= predict;
run;
```

- Adj R-Square = 0.8671 after 13 steps
- CV PRESS = 31.3319
- 14 parameters are selected to be used in model
- Kaggle score: 0.15463

D. Conclusion

After running the three models in the last step (Build models), we upload our files to Kaggle to get our Kaggle score. Finally, we chose our custom model as Backward Selection with Cross Validation.

Predictive Models	Adjusted R^2	CV PRESS	Kaggle Score
Forward	.8676	31.2575	0.15450
Backward	.8719	30.5547	0.15149
Stepwise	.8671	31.3319	0.15463
CUSTOM	.8719	30.5547	0.15149

V. REFERENCES

1. MSDS 6371 – Project Description, SMU, 2019. (https://hnguye01.github.io/6371Description.html)

2. Dean De Cock; Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project. *Journal of Statistics Education.* **19**(3), 2011. (http://jse.amstat.org/v19n3/decock.pdf)

HUY HOANG NGUYEN - Southern Methodist University - *Email address:* hoangnguyen@smu.edu

IKENNA NWAOGU - Southern Methodist University - *Email address:* inwaogu@smu.edu

HAO WANG – Southern Methodist University - *Email address:* wangmichael@smu.edu

VI. APPENDIX

1. SAS codes and Outputs for Analysis of Question 1

- Import the data set train.csv and test.csv:

```
FILENAME REFFILE '/folders/myfolders/MS6371/train.csv';
PROC IMPORT DATAFILE=REFFILE
        DBMS=CSV
        OUT=WORK.TRAIN;
        GETNAMES=YES;
        DATAROW=2;
RUN;

FILENAME REFFILE1 '/folders/myfolders/MS6371/test.csv';
PROC IMPORT DATAFILE=REFFILE1
        DBMS=CSV
        OUT=WORK.TEST;
        GETNAMES=YES;
        DATAROW=2;
RUN;
```

There are 1460 observations and 81 variables in the train.csv and 1459 observations and 80 variables in the test.csv.

- Create a new dataset "trainAmes" with only 3 neighborhoods Names, BrkSide and Edwards that "Century 21 Ames" works on:

```
data trainAmes;
set train;
where neighborhood = 'Names' OR
    neighborhood = 'BrkSide' OR
    neighborhood = 'Edwards';
run;

proc print data=trainAmes;
run;
```

Output (383 observations and 81 variables):

Obs	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neighborhood	Conditi
1	10	190	RL	50	7420	Pave	NA	Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Artery
2	15	20	RL	NA	10920	Pave	NA	IR1	Lvl	AllPub	Corner	Gtl	NAmes	Norm
3	16	45	RM	51	6120	Pave	NA	Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm
4	17	20	RL	NA	11241	Pave	NA	IR1	Lvl	AllPub	CulDSac	Gtl	NAmes	Norm
5	20	20	RL	70	7560	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	NAmes	Norm
6	27	20	RL	60	7200	Pave	NA	Reg	Lvl	AllPub	Corner	Gtl	NAmes	Norm
7	29	20	RL	47	16321	Pave	NA	IR1	Lvl	AllPub	CulDSac	Gtl	NAmes	Norm
8	30	30	RM	60	6324	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	BrkSide	Feedr
9	34	20	RL	70	10552	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	NAmes	Norm
10	38	20	RL	74	8532	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	NAmes	Norm
11	39	20	RL	68	7922	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	NAmes	Norm

(Table 1.1)

- Prepare data – Add columns with transformed data:

```
/* Prepare data: Add columns with transformed data  */
data trainAmes;
set trainAmes;
if neighborhood = “Names” then d1=1; else d1=0;
if neighborhood = “BrkSide” then d2=1; else d2=0;
run;

proc print data = trainAmes;
run;
```

Output (383 observations and 83 variables):

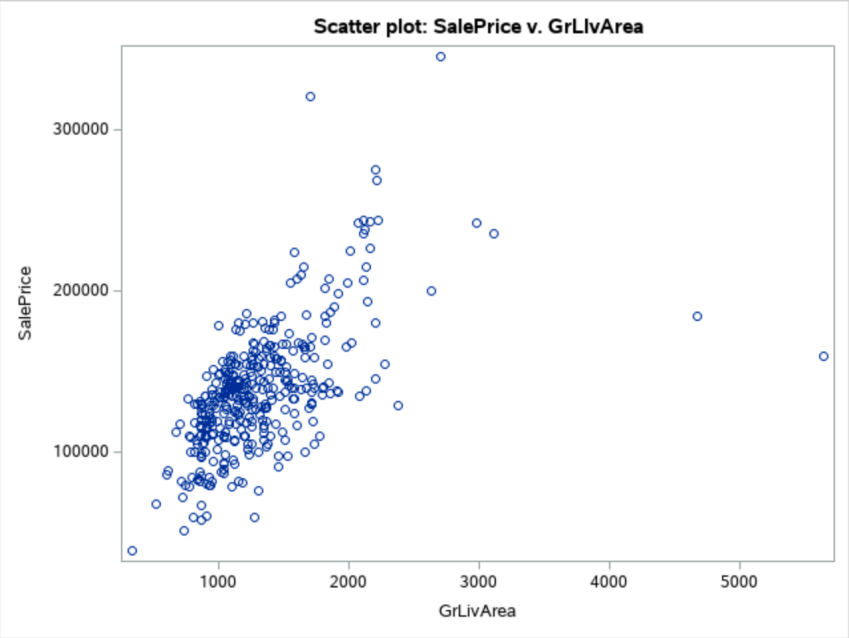
eckSF	OpenPorchSF	EnclosedPorch	_3SsnPorch	ScreenPorch	PoolArea	PoolQC	Fence	MiscFeature	MiscVal	MoSold	YrSold	SaleType	SaleCondition	SalePrice	d1	d2
0	4	0	0	0	0	NA	NA	NA	0	1	2008	WD	Normal	118000	0	1
0	213	176	0	0	0	NA	GdWo	NA	0	5	2008	WD	Normal	157000	1	0
48	112	0	0	0	0	NA	GdPrv	NA	0	7	2007	WD	Normal	132000	0	1
0	0	0	0	0	0	NA	NA	Shed	700	3	2010	WD	Normal	149000	1	0
0	0	0	0	0	0	NA	MnPrv	NA	0	5	2009	COD	Abnorml	139000	1	0
222	32	0	0	0	0	NA	NA	NA	0	5	2010	WD	Normal	134800	1	0
288	258	0	0	0	0	NA	NA	NA	0	12	2006	WD	Normal	207500	1	0
49	0	87	0	0	0	NA	NA	NA	0	5	2008	WD	Normal	68500	0	1
0	38	0	0	0	0	NA	NA	NA	0	4	2010	WD	Normal	165500	1	0
0	0	0	0	0	0	NA	NA	NA	0	10	2009	WD	Normal	153000	1	0
0	52	0	0	0	0	NA	NA	NA	0	1	2010	WD	Abnorml	109000	1	0

(Table 1.2)

- Plot the data:

```
title ‘Scatter plot: SalePrice v. GrLIvArea’;
PROC sgplot DATA=trainAmes;
  scatter x=GrLIvArea y=SalePrice;
run;
```

Output (Scatter plot SalePrice v. GrLIvArea):



(Plot 1.3)

It looks like there is a linear relationship between the living area and the sale price. However, there are some outliers.

- Build the first model:

```
/* Develop a Tentative Model 1 */
proc reg data= trainAmes;
model SalePrice = GrLIvArea / vif clb cli clm;
run;
```

Output:

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Number of Observations Read	383
Number of Observations Used	383

Analysis of Variance

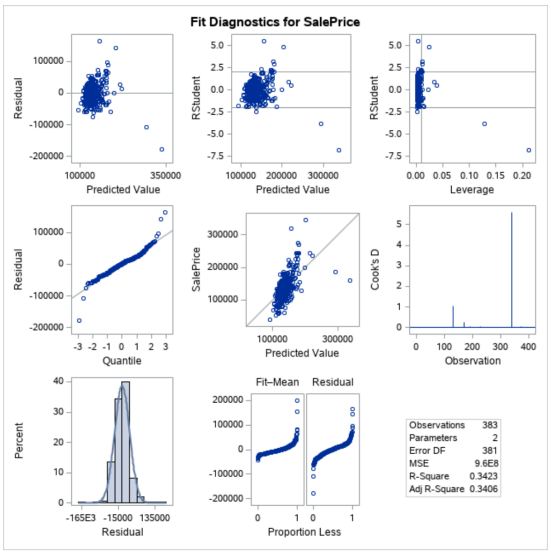
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1.903676E11	1.903676E11	198.29	<.0001
Error	381	3.657846E11	960064442		
Corrected Total	382	5.561521E11			

Root MSE	30985	R-Square	0.3423
Dependent Mean	138063	Adj R-Sq	0.3406
Coeff Var	22.44267		

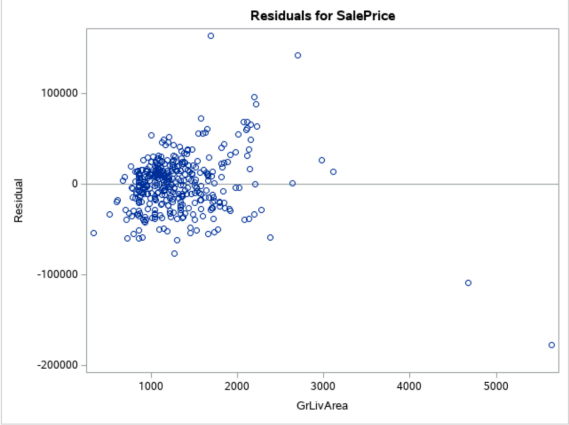
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation	95% Confidence Limits	
Intercept	1	78206	4536.05353	17.24	<.0001	0	69287	87124
GrLivArea	1	45.97896	3.26522	14.08	<.0001	1.00000	39.55885	52.39907

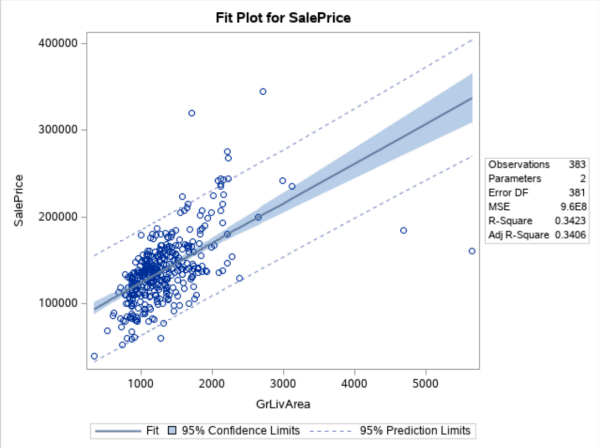
(Table 1.4)



(Plot 1.5)



(Plot 1.6)



(Plot 1.7)

Observations:
Four outliers with studentized residuals larger than ±2.5 found.
One outlier with highest Cook's D (>5)
Straight line in QQ plot and symmetric shape of histogram indicate the normal distribution.
Adj R-Square = 0.3406

- Remove the 4 outliers:

```
/* Remove the 4 outliers due to high residuals */
data trainAmesOutlier;
set train;
where Id = 643 or Id=725 OR Id=1299 or Id=524;

proc print data=trainAmesOutlier;
run;
```

Obs	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neigh
1	524	60	RL	13	40094	Pave	NA	IR1	Bnk	AllPub	Inside	Gtl	Edwa
2	643	80	RL	75	13860	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	NAme
3	725	20	RL	86	13286	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	Edwa
4	1299	60	RL	31	63887	Pave	NA	IR3	Bnk	AllPub	Corner	Gtl	Edwa

(Table 1.8)

```
data trainAmesNoOutlier;
set trainAmes;
where Id ~= 643 AND Id ~= 725 AND Id ~= 1299 AND Id ~= 524;
run;

proc print data=trainAmesNoOutlier;
run;
```

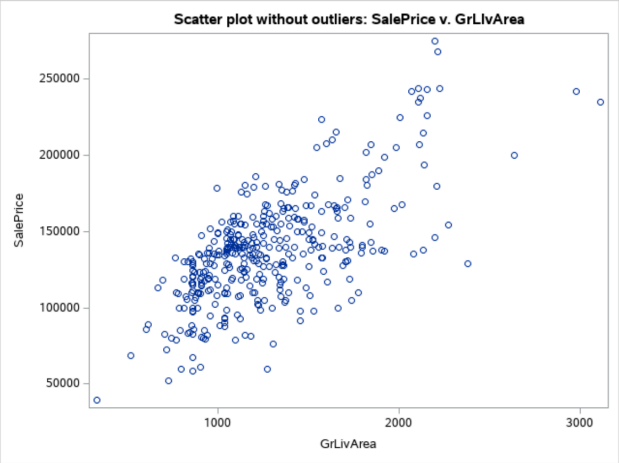
372	1436	20	RL	80	8400	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	NAmes
373	1437	20	RL	60	9000	Pave	NA	Reg	Lvl	AllPub	FR2	Gtl	NAmes
374	1444	30	RL	NA	8854	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	BrkSide
375	1449	50	RL	70	11767	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	Edwards
376	1451	90	RL	60	9000	Pave	NA	Reg	Lvl	AllPub	FR2	Gtl	NAmes
377	1453	180	RM	35	3675	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	Edwards
378	1459	20	RL	68	9717	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	NAmes
379	1460	20	RL	75	9937	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	Edwards

(Table 1.9)

From 383 observations, we now have 279 observations after removing 4 outliers.

- Plot the scatter plot without outliers:

```
/* Plot the scatter plot without outliers */
title 'Scatter plot without outliers: SalePrice v. GrLIvArea';
PROC sgplot DATA=trainAmesNoOutlier;
  scatter x=GrLIvArea y=SalePrice;
run;
```



(Plot 1.10)

- Build the second model without outliers:

```
/* Develop a Tentative Model 2 without outliers */
proc reg data= trainAmesNoOutlier;
model SalePrice = GrLIvArea / vif clb cli clm;
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Number of Observations Read	379
Number of Observations Used	379

Analysis of Variance

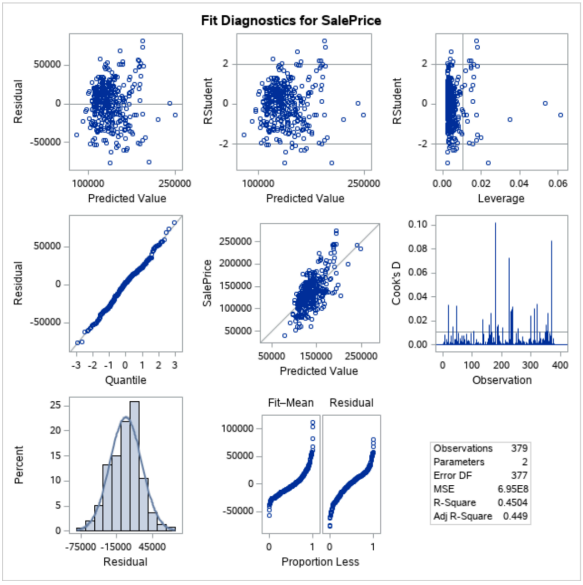
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2.148668E11	2.148668E11	309.00	<.0001
Error	377	2.621477E11	695351977		
Corrected Total	378	4.770145E11			

Root MSE	26370	R-Square	0.4504
Dependent Mean	136855	Adj R-Sq	0.4490
Coeff Var	19.26817		

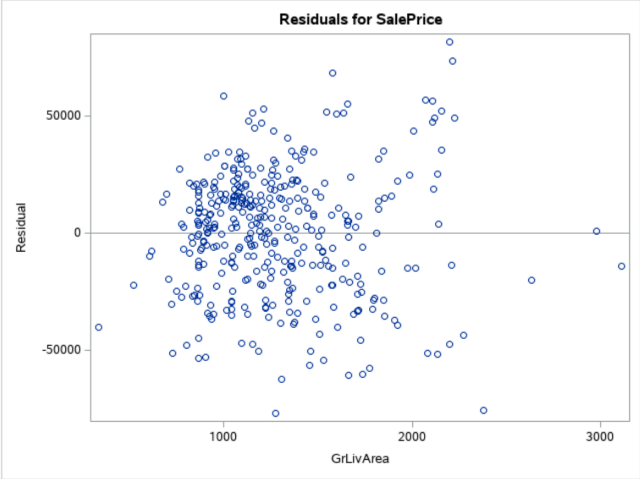
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation	95% Confidence Limits	
Intercept	1	58785	4643.19442	12.66	<.0001	0	49655	67915
GrLivArea	1	61.14848	3.47859	17.58	<.0001	1.00000	54.30861	67.98835

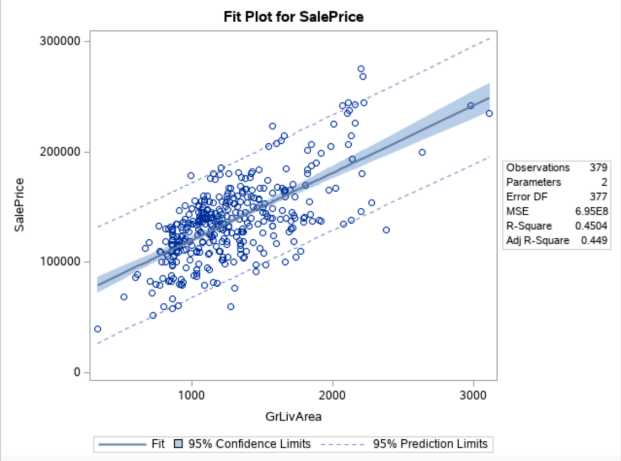
(Table 1.11)



(Plot 1.12)



(Plot 1.13)



(Plot 1.14)

Observations:
Scatter plots indicate random distributed residuals.
Cook's D values lower than 0.10
Straight line in QQ plot and symmetric shape of histogram indicate the normal distribution.
Adj R-Square = 0.449

- Build the third tentative model without outliers:

```
/* Develop a Tentative Model 3 without outliers */
proc glm data= trainAmesNoOutlier plots = all;
/* class neighborhood (ref = "Edwards"); */
model SalePrice = GrLIvArea | d1 | d2 / solution clparm cli;
run;
```

The GLM Procedure

Number of Observations Read	379
Number of Observations Used	379

The GLM Procedure

Dependent Variable: SalePrice

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	249429640074	49885928015	81.76	<.0001
Error	373	227584871181	610147107.72		
Corrected Total	378	477014511255			

R-Square	Coeff Var	Root MSE	SalePrice Mean
0.522897	18.04909	24701.16	136855.4

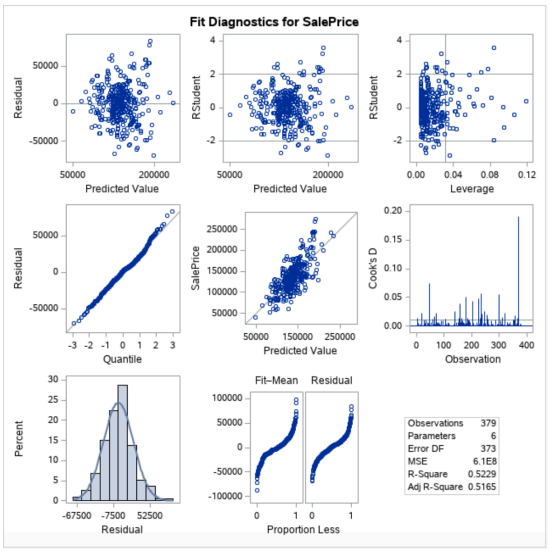
(Table 1.15)

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	37100.42156		9283.71546	4.00	<.0001	18845.44062	55355.40251
GrLivArea	70.15837		7.10756	9.87	<.0001	56.18246	84.13427
d1	43225.29073		10837.81644	3.99	<.0001	21914.41218	64536.16929
GrLivArea*d1	-20.59712		8.20386	-2.51	0.0125	-36.72874	-4.46551
d2	-17128.90777		14154.89029	-1.21	0.2270	-44962.29557	10704.48003
GrLivArea*d2	17.00416		11.05135	1.54	0.1247	-4.72660	38.73493
d1*d2	0.00000	B	.	.	.	.	.
GrLivArea*d1*d2	0.00000	B	.	.	.	.	.

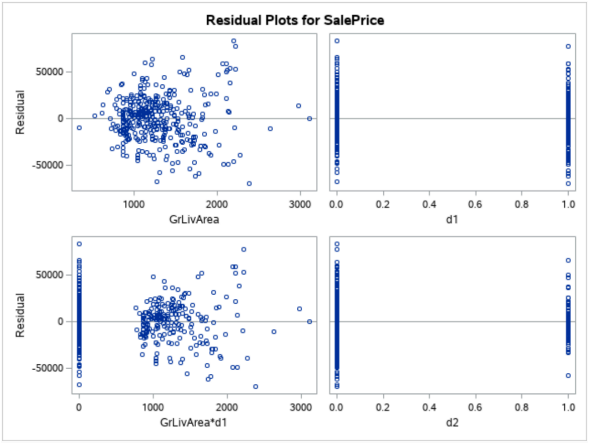
(Table 1.16)

Sum of Residuals	-2.43017E-9
Sum of Squared Residuals	227584871181
Sum of Squared Residuals - Error SS	0.0003662109
PRESS Statistic	237769407660
First Order Autocorrelation	-0.046536661
Durbin-Watson D	2.0908161961

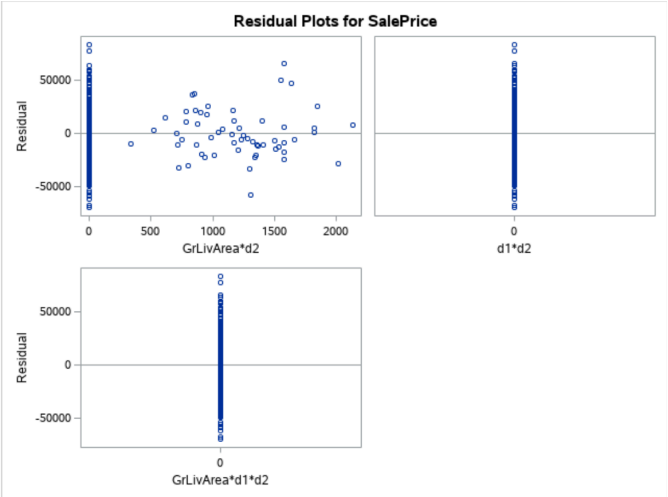
(Table 1.17)



(Plot 1.18)



(Plot 1.19)

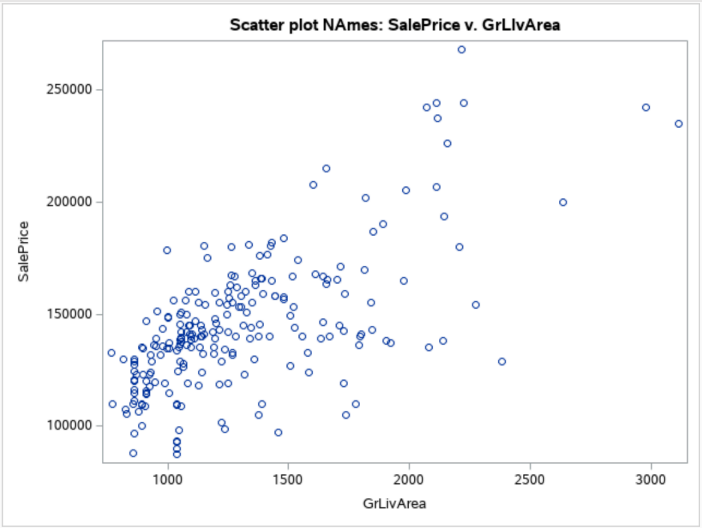


(Plot 1.20)

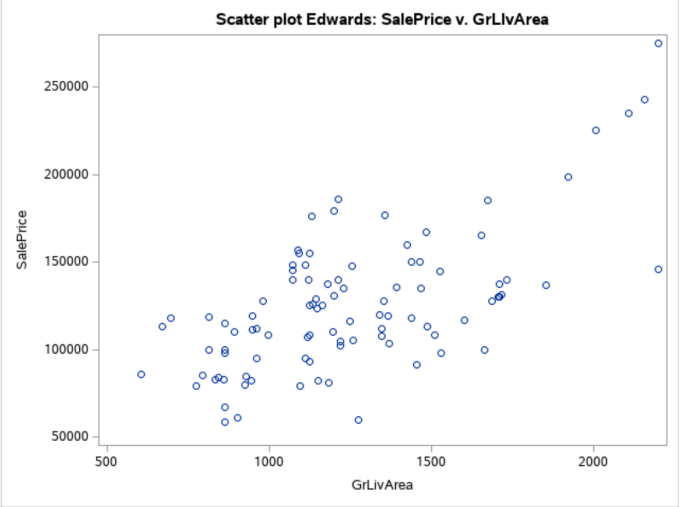
Observations:
Scatter plots indicate random distributed residuals.
Cook's D values lower than 0.20
Straight line in QQ plot and symmetric shape of histogram indicate the normal distribution.
Adj R-Square = 0.5165

- Scatter plots of 3 neighborhoods (Model 3):

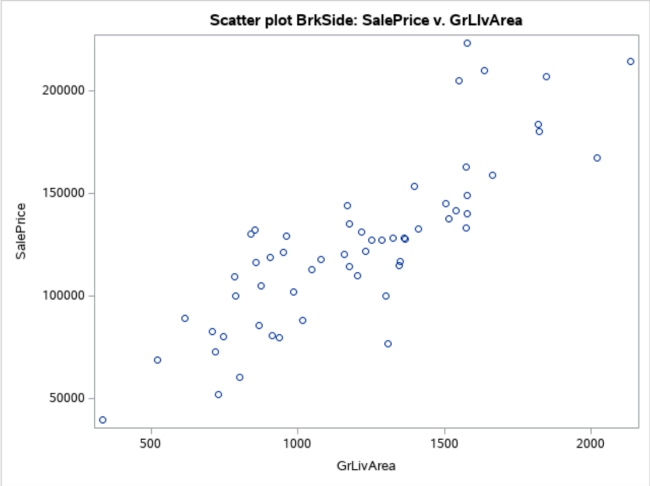
```
/* Use Model 3 to fit data */
/* Scatter plots of three neighborhoods */

title 'Scatter plot NAmes: SalePrice v. GrLIvArea';
PROC sgplot DATA=trainAmesNoOutlier;
where neighborhood = 'NAmes';
  scatter x=GrLIvArea y=SalePrice;
run;

title 'Scatter plot Edwards: SalePrice v. GrLIvArea';
PROC sgplot DATA=trainAmesNoOutlier;
where neighborhood = 'Edwards';
  scatter x=GrLIvArea y=SalePrice;
run;

title 'Scatter plot BrkSide: SalePrice v. GrLIvArea';
PROC sgplot DATA=trainAmesNoOutlier;
where neighborhood = 'BrkSide';
  scatter x=GrLIvArea y=SalePrice;
run;
```



(Plot 1.21)



(Plot 1.22)



(Plot 1.23)

There is no evidence that a transformation is necessary from the scatterplots.

2. SAS codes and Outputs for Analysis of Question 2

- Preparing the data:

```
*** Analysis 2  ***/
/** Multiple Regression Model for SalePrice Prediction **/
/* Prepare data */
data trainAmes2;
set train;
/* Select numerical variables as predictors */
/* Removing variables with NA's: LotFrontage MasVnrArea GarageYrBlt */
/* Removing variables with wrong name: 1stFlrSF 2ndFlrSF 3SsnPorch */
```

```
keep Id MSSubClass LotArea OverallQual OverallCond YearBuilt YearRemodAdd BsmtFinSF1
BsmtFinSF2 BsmtUnfSF
 LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr
KitchenAbvGr TotRmsAbvGrd
 Fireplaces GarageCars GarageArea WoodDeckSF OpenPorchSF EnclosedPorch ScreenPorch
PoolArea MiscVal MoSold
 YrSold SalePrice;
run;
```

Then we have selected 31 numerical variables as predictors. Here we have removed variables with NA's (LotFrontage, MasVnrArea, GarageYrBlt) and variables with wrong name (1stFlrSF, 2ndFlrSF, 3SsnPorch).

We have 1460 observations and 31 variables.

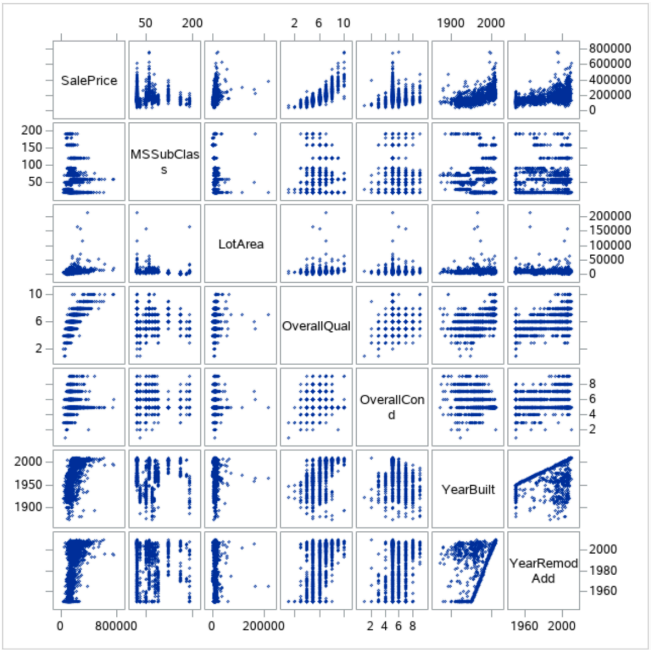
```
proc print data=trainAmes2;
run;
```

1453	1453	180	3675	5	5	2005	2005	547	0	0	0	1072
1454	1454	20	17217	5	5	2006	2006	0	0	1140	0	1140
1455	1455	20	7500	7	5	2004	2005	410	0	811	0	1221
1456	1456	60	7917	6	5	1999	2000	0	0	953	0	1647
1457	1457	20	13175	6	6	1978	1988	790	163	589	0	2073
1458	1458	70	9042	7	9	1941	2006	275	0	877	0	2340
1459	1459	20	9717	5	6	1950	1996	49	1029	0	0	1078
1460	1460	20	9937	5	6	1965	1965	830	290	136	0	1256

(Table 2.1)

- Variable selections with Scatter plots:

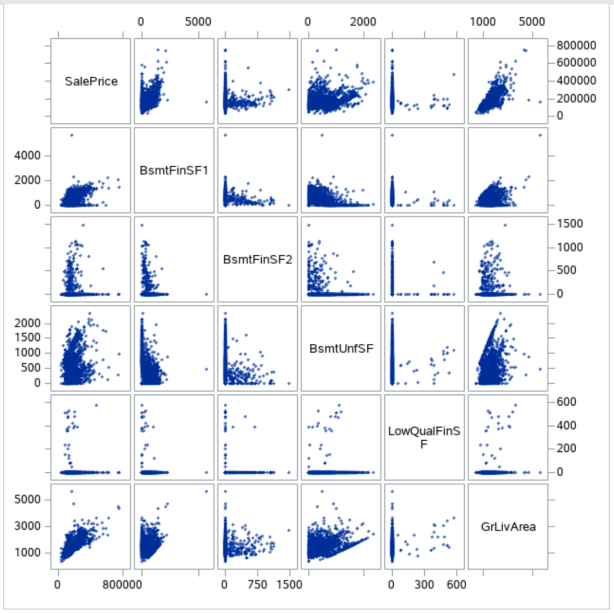
```
/* 1: Variable selection with scatter plots */
PROC sgscatter DATA=trainAmes2;
matrix SalePrice MSSubClass LotArea OverallQual OverallCond YearBuilt YearRemodAdd;
run;
```



(Plot 2.2)

Linear relationship: OverallQual, OverallCond, YearBuilt, YearRemodAdd
Nonlinear relationship: NSSubClass, LotArea
Keep 4 variables with linear relationship

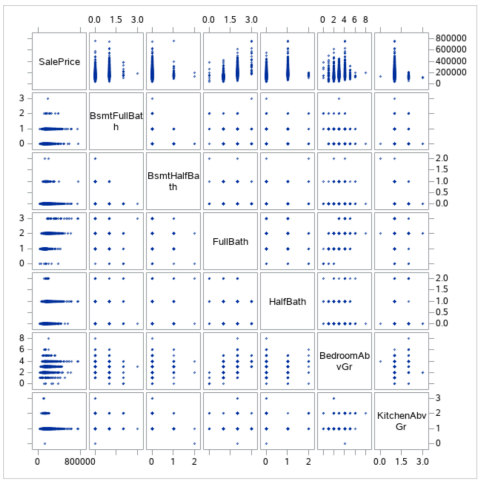
```
PROC sgscatter DATA=trainAmes2;
matrix SalePrice BsmtFinSF1 BsmtFinSF2 BsmtUnfSF LowQualFinSF GrLivArea;
run;
```



(Plot 2.3)

Linear relationship: GrLivArea
Nonlinear relationship: BsmtFinSF1, BsmFinFS2, BsmUnfSF, LowQualFinSF
Keep 1 variable with linear relationship

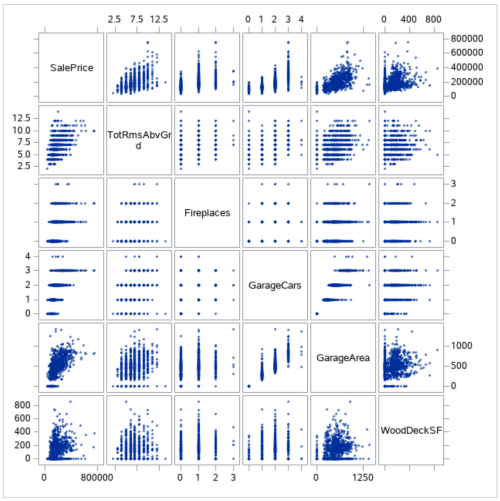
```
PROC sgscatter DATA=trainAmes2;
matrix SalePrice BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr;
run;
```



(Plot 2.4)

Linear relationship: FullBath
Nonlinear relationship: BsmtFullBath, BsmtHalfBath, HalfBath, BedroomAbvGr, KitchenAbvGr
Keep 1 variable with linear relationship

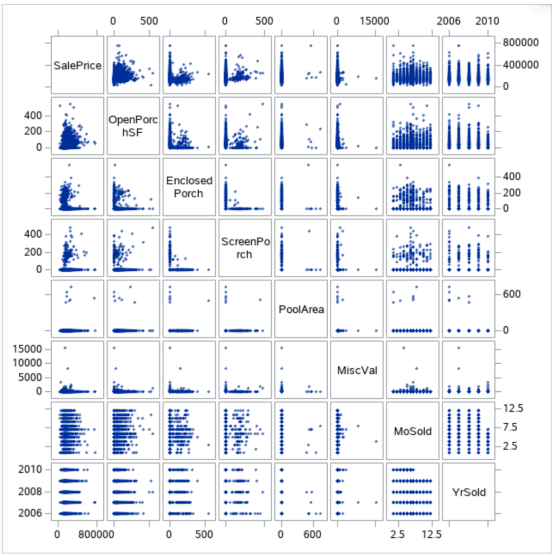
```
PROC sgscatter DATA=trainAmes2;
matrix SalePrice TotRmsAbvGrd Fireplaces GarageCars GarageArea WoodDeckSF;
run;
```



(Plot 2.5)

Linear relationship: TotRmsAbvGrd, Fireplaces, GarageCars, GarageArea, WoodDeckSF
Nonlinear relationship: None
Keep 5 variable with linear relationship

```
PROC sgscatter DATA=trainAmes2;
matrix SalePrice OpenPorchSF EnclosedPorch ScreenPorch PoolArea MiscVal MoSold YrSold;
run;
```



(Plot 2.6)

Linear relationship: OpenPorchSF
Nonlinear relationship: EnclosedPorch, ScreenPorch, PoolArea, MiscVal, MoSold, YrSold
Keep 1 variable with linear relationship

/** 12 numerical variables are selected as candidate predictors through checking scatter plots **/
/** OverallQual, OverallCond, YearBuilt, YearRemodAdd, GrLivArea, FullBath, TotRmsAbvGrd, Fireplaces, GarageCars, GarageArea, WoodDeckSF, OpenPorchSF **/
/** 13 categorical variables are selected as candidate predictors base on SME's judgement **/
/** Utilities, LotConfig, Neighborhood, ExterQual, BsmtQual, HeatingQC, KitchenQual, FireplaceQu, GarageQual, PoolQC, Fence, SaleType, SaleCondition **/

- Remove outliers:

```
/*** Remove outliers ***/
data trainAmes2NoOutlier;
set trainAmes2;
where Id ~= 643 AND Id ~= 725 AND Id ~= 1299 AND Id ~= 524;
logSalePrice = log(SalePrice);
```

```
logOverallQual = log(OverallQual);
run;
proc print data=trainAmes2NoOutlier;
run;
```

1449	1453	180	3675	5	5	2005	2005	547	0	0	0	1072	1	0	1	0	2	1
1450	1454	20	17217	5	5	2006	2006	0	0	1140	0	1140	0	0	1	0	3	1
1451	1455	20	7500	7	5	2004	2005	410	0	811	0	1221	1	0	2	0	2	1
1452	1456	60	7917	6	5	1999	2000	0	0	953	0	1647	0	0	2	1	3	1
1453	1457	20	13175	6	6	1978	1988	790	163	589	0	2073	1	0	2	0	3	1
1454	1458	70	9042	7	9	1941	2006	275	0	877	0	2340	0	0	2	0	4	1
1455	1459	20	9717	5	6	1950	1996	49	1029	0	0	1078	1	0	1	0	2	1
1456	1460	20	9937	5	6	1965	1965	830	290	136	0	1256	1	0	1	1	3	1

(Table 2.7)

Then our dataset has 1456 observations (train.csv).

- Forward Method: Build model1 with 12 numerical predictors:

```
/** Forward Method  **/
/* 2: Build model1 with 12 numerical predictors */
proc reg data=trainames2Nooutlier;
model SalePrice = OverallQual OverallCond YearBuilt YearRemodAdd GrLivArea FullBath
TotRmsAbvGrd
 Fireplaces GarageCars GarageArea WoodDeckSF OpenPorchSF / selection=forward slentry=0.1
slstay=0.1 adjrsq;
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Number of Observations Read	1456
Number of Observations Used	1456

Forward Selection: Step 1

Variable OverallQual Entered: R-Square = 0.6320 and C(p) = 1217.382

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	5.790255E12	5.790255E12	2497.57	<.0001
Error	1454	3.370883E12	2318351446		
Corrected Total	1455	9.161138E12			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-98853	5734.81184	6.888402E11	297.13	<.0001
OverallQual	45903	918.49896	5.790255E12	2497.57	<.0001

Bounds on condition number: 1, 1

(Table 2.8)

Forward Selection: Step 2

Variable GrLivArea Entered: R-Square = 0.7405 and C(p) = 432.4422

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	6.783999E12	3.391999E12	2073.32	<.0001
Error	1453	2.377139E12	1636021578		
Corrected Total	1455	9.161138E12			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-110665	4841.31005	8.548452E11	522.51	<.0001
OverallQual	32040	954.83966	1.842108E12	1125.97	<.0001
GrLivArea	63.75882	2.58701	9.937437E11	607.41	<.0001

Bounds on condition number: 1.5314, 6.1257

(Table 2.9)

Forward Selection: Step 3

Variable GarageArea Entered: R-Square = 0.7676 and C(p) = 238.0537

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	7.031997E12	2.343999E12	1598.53	<.0001
Error	1452	2.129141E12	1466350526		
Corrected Total	1455	9.161138E12			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-105614	4599.82587	7.730302E11	527.18	<.0001
OverallQual	26911	986.27045	1.091684E12	744.49	<.0001
GrLivArea	57.52888	2.49560	7.79221E11	531.40	<.0001
GarageArea	75.44182	5.80105	2.479984E11	169.13	<.0001

Bounds on condition number: 1.823, 14.751

(Table 2.10)

Forward Selection: Step 4

Variable YearBuilt Entered: R-Square = 0.7812 and C(p) = 141.2523

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	7.156763E12	1.789191E12	1295.22	<.0001
Error	1451	2.004375E12	1381374861		
Corrected Total	1455	9.161138E12			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-864007	79925	1.614302E11	116.86	<.0001
OverallQual	21867	1094.54018	5.51348E11	399.13	<.0001
YearBuilt	399.18902	42.00357	1.24766E11	90.32	<.0001
GrLivArea	64.01796	2.51660	8.938923E11	647.10	<.0001
GarageArea	59.41614	5.87754	1.411656E11	102.19	<.0001

Bounds on condition number: 2.3833, 29.734

(Table 2.11)

Forward Selection: Step 5

Variable Fireplaces Entered: R-Square = 0.7876 and C(p) = 96.9407

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	7.215245E12	1.443049E12	1075.30	<.0001
Error	1450	1.945893E12	1341995095		
Corrected Total	1455	9.161138E12			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-879877	78814	1.672591E11	124.63	<.0001
OverallQual	20720	1092.72981	4.825044E11	359.54	<.0001
YearBuilt	410.98482	41.43907	1.320026E11	98.36	<.0001
GrLivArea	59.32257	2.58044	7.092573E11	528.51	<.0001
Fireplaces	11308	1712.90904	58482034763	43.58	<.0001
GarageArea	58.99972	5.79350	1.391774E11	103.71	<.0001

Bounds on condition number: 2.4451, 44.699

(Table 2.12)

Forward Selection: Step 6

Variable OverallCond Entered: R-Square = 0.7933 and C(p) = 57.7443

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	7.267268E12	1.211211E12	926.70	<.0001
Error	1449	1.89387E12	1307018777		
Corrected Total	1455	9.161138E12			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-1127954	87155	2.189195E11	167.50	<.0001
OverallQual	19336	1100.49240	4.03481E11	308.70	<.0001
OverallCond	5923.02856	938.83307	52022679795	39.80	<.0001
YearBuilt	522.86812	44.57511	1.798379E11	137.59	<.0001
GrLivArea	61.31303	2.56606	7.461982E11	570.92	<.0001
Fireplaces	11256	1690.45953	57951216503	44.34	<.0001
GarageArea	58.93112	5.71752	1.388534E11	106.24	<.0001

Bounds on condition number: 2.5463, 63.632

(Table 2.13)

Forward Selection: Step 7

Variable WoodDeckSF Entered: R-Square = 0.7962 and C(p) = 38.6056

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	7.293962E12	1.041995E12	808.07	<.0001
Error	1448	1.867176E12	1289486490		
Corrected Total	1455	9.161138E12			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-1064121	87698	1.89856E11	147.23	<.0001
OverallQual	19496	1093.65151	4.097594E11	317.77	<.0001
OverallCond	5528.82156	936.53147	44940426814	34.85	<.0001
YearBuilt	491.06052	44.82366	1.547648E11	120.02	<.0001
GrLivArea	59.78079	2.57094	6.971971E11	540.68	<.0001
Fireplaces	10598	1685.31147	50989157246	39.54	<.0001
GarageArea	57.67015	5.68580	1.326588E11	102.88	<.0001
WoodDeckSF	36.23641	7.96432	26693770910	20.70	<.0001

Bounds on condition number: 2.549, 82.86

(Table 2.14)

Forward Selection: Step 8

Variable FullBath Entered: R-Square = 0.7985 and C(p) = 23.6116

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	7.315422E12	9.144277E11	716.89	<.0001
Error	1447	1.845716E12	1275546966		
Corrected Total	1455	9.161138E12			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-1187225	92241	2.113054E11	165.66	<.0001
OverallQual	19730	1089.22482	4.185179E11	328.11	<.0001
OverallCond	5435.38618	931.73420	43408341047	34.03	<.0001
YearBuilt	556.64373	47.36133	1.76199E11	138.14	<.0001
GrLivArea	66.49620	3.03624	6.118112E11	479.65	<.0001
FullBath	-10280	2506.23666	21459977547	16.82	<.0001
Fireplaces	9810.79968	1687.12000	43133363515	33.82	<.0001
GarageArea	56.49681	5.66221	1.269907E11	99.56	<.0001
WoodDeckSF	35.00489	7.92684	24874448890	19.50	<.0001

Bounds on condition number: 2.7056, 120.71

(Table 2.15)

Forward Selection: Step 9

Variable TotRmsAbvGrd Entered: R-Square = 0.7997 and C(p) = 17.3192

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	7.325893E12	8.139881E11	641.35	<.0001
Error	1446	1.835245E12	1269187343		
Corrected Total	1455	9.161138E12			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-1158995	92535	1.991044E11	156.88	<.0001
OverallQual	19487	1089.79027	4.05822E11	319.75	<.0001
OverallCond	5417.77465	929.42880	43125619853	33.98	<.0001
YearBuilt	546.40603	47.37737	1.688165E11	133.01	<.0001
GrLivArea	74.82031	4.19179	4.043583E11	318.60	<.0001
FullBath	-9437.28884	2517.13267	17840540796	14.06	0.0002
TotRmsAbvGrd	-3051.52377	1062.36538	10471561072	8.25	0.0041
Fireplaces	9427.37514	1688.19461	39578769815	31.18	<.0001
GarageArea	56.12708	5.64955	1.252689E11	98.70	<.0001
WoodDeckSF	34.08015	7.91361	23538558261	18.55	<.0001

(Table 2.16)

Forward Selection: Step 10

Variable YearRemodAdd Entered: R-Square = 0.8007 and C(p) = 12.1962

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	7.334888E12	7.334888E11	580.36	<.0001
Error	1445	1.82625E12	1263840837		
Corrected Total	1455	9.161138E12			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-1368572	121235	1.610533E11	127.43	<.0001
OverallQual	18928	1107.51602	3.691427E11	292.08	<.0001
OverallCond	4392.23911	1003.97855	24188863442	19.14	<.0001
YearBuilt	483.46133	52.83797	1.058092E11	83.72	<.0001
YearRemodAdd	173.34216	64.97593	8994889292	7.12	0.0077
GrLivArea	74.36409	4.18644	3.987757E11	315.53	<.0001
FullBath	-10289	2532.05331	20870322480	16.51	<.0001
TotRmsAbvGrd	-2989.37406	1060.38132	10044510277	7.95	0.0049
Fireplaces	9977.10851	1697.19101	43675745824	34.56	<.0001
GarageArea	56.26402	5.63787	1.258706E11	99.59	<.0001
WoodDeckSF	33.47555	7.90017	22692092869	17.95	<.0001

Bounds on condition number: 5.1914, 240.4

(Table 2.17)

Forward Selection: Step 11

Variable OpenPorchSF Entered: R-Square = 0.8011 and C(p) = 11.1511

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	7.338733E12	6.671576E11	528.63	<.0001
Error	1444	1.822405E12	1262053058		
Corrected Total	1455	9.161138E12			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-1351611	121539	1.56082E11	123.67	<.0001
OverallQual	18813	1108.69800	3.633715E11	287.92	<.0001
OverallCond	4381.79706	1003.28604	24073131691	19.07	<.0001
YearBuilt	482.31557	52.80467	1.05292E11	83.43	<.0001
YearRemodAdd	166.23996	65.05731	8240551152	6.53	0.0107
GrLivArea	73.53960	4.21006	3.850733E11	305.12	<.0001
FullBath	-10370	2530.68765	21193213681	16.79	<.0001
TotRmsAbvGrd	-2924.21376	1060.28840	9599482588	7.61	0.0059
Fireplaces	9937.40120	1696.14274	43320999085	34.33	<.0001
GarageArea	55.88612	5.63804	1.240023E11	98.25	<.0001
WoodDeckSF	34.27243	7.90777	23706051751	18.78	<.0001
OpenPorchSF	26.59865	15.23800	3845393792	3.05	0.0811

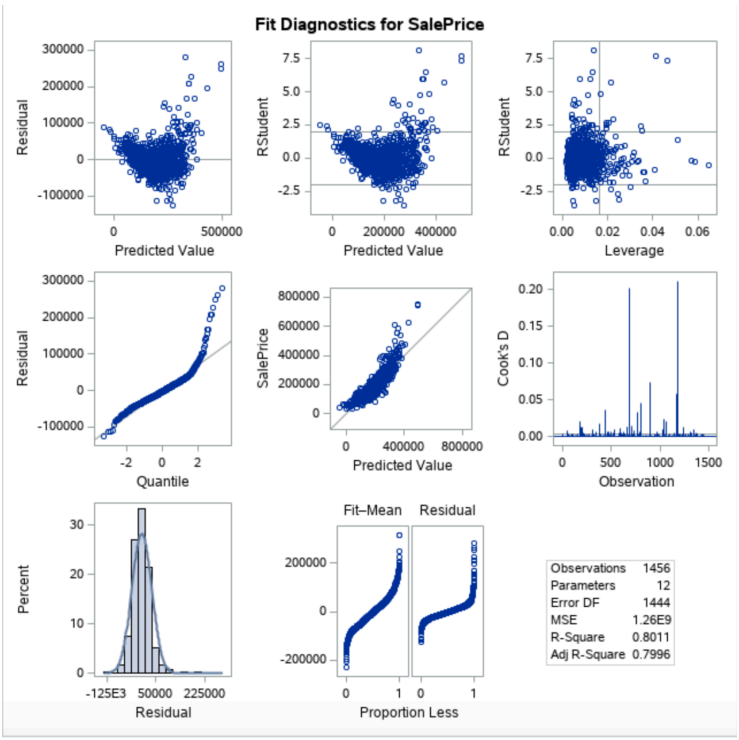
Bounds on condition number: 5.2576, 278.06

No other variable met the 0.1000 significance level for entry into the model.

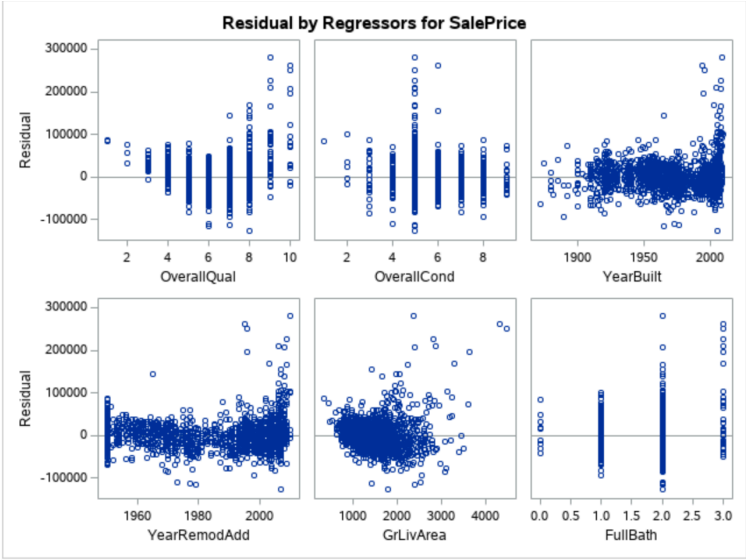
(Table 2.18)

Summary of Forward Selection							
Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	OverallQual	1	0.6320	0.6320	1217.38	2497.57	<.0001
2	GrLivArea	2	0.1085	0.7405	432.442	607.41	<.0001
3	GarageArea	3	0.0271	0.7676	238.054	169.13	<.0001
4	YearBuilt	4	0.0136	0.7812	141.252	90.32	<.0001
5	Fireplaces	5	0.0064	0.7876	96.9407	43.58	<.0001
6	OverallCond	6	0.0057	0.7933	57.7443	39.80	<.0001
7	WoodDeckSF	7	0.0029	0.7962	38.6056	20.70	<.0001
8	FullBath	8	0.0023	0.7985	23.6116	16.82	<.0001
9	TotRmsAbvGrd	9	0.0011	0.7997	17.3192	8.25	0.0041
10	YearRemodAdd	10	0.0010	0.8007	12.1962	7.12	0.0077
11	OpenPorchSF	11	0.0004	0.8011	11.1511	3.05	0.0811

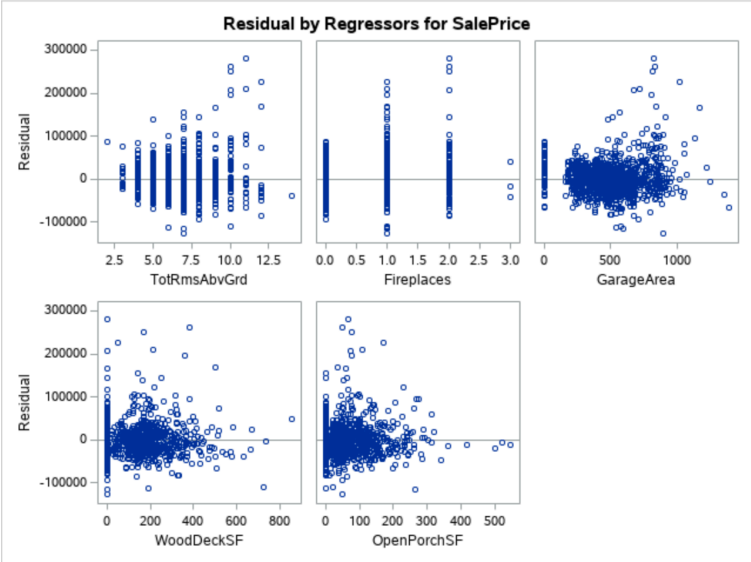
(Table 2.19)



(Plot 2.20)



(Plot 2.21)



(Plot 2.22)

The top 4 variables with the highest Partial R^2 will be selected in the model.
Adj R-Square = 0.7996 after 11 steps

- Build model2 with 4 numerical predictors with the highest partial R-square:

```
/* 2: Build model2 with 4 numerical predictors with the highest partial R-square */
proc reg data=trainames2Nooutlier;
model SalePrice = OverallQual GrLivArea GarageArea YearBuilt / selection=forward slentry=0.1
slstay=0.1 adjrsq;
run;
```

Forward Selection: Step 4

Variable YearBuilt Entered: R-Square = 0.7812 and C(p) = 5.0000

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	7.156763E12	1.789191E12	1295.22	<.0001
Error	1451	2.004375E12	1381374861		
Corrected Total	1455	9.161138E12			

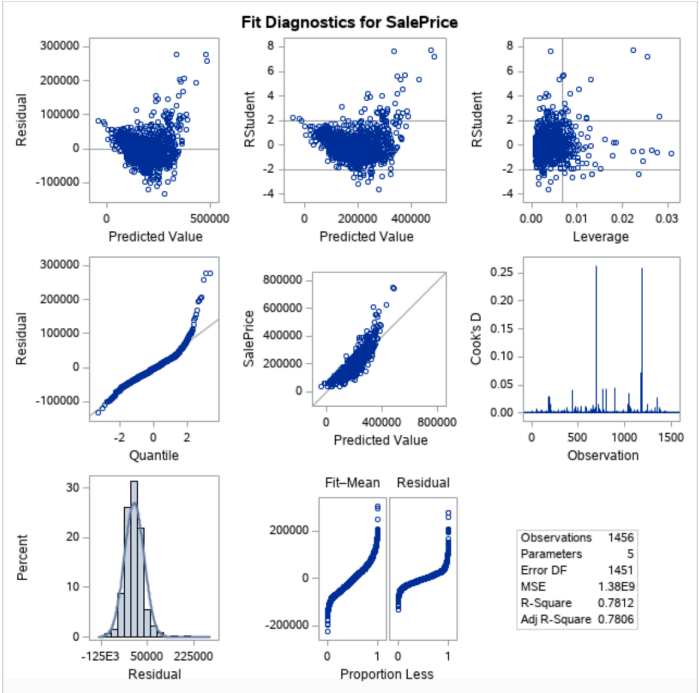
Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-864007	79925	1.614302E11	116.86	<.0001
OverallQual	21867	1094.54018	5.51348E11	399.13	<.0001
GrLivArea	64.01796	2.51660	8.938923E11	647.10	<.0001
GarageArea	59.41614	5.87754	1.411656E11	102.19	<.0001
YearBuilt	399.18902	42.00357	1.24766E11	90.32	<.0001

Bounds on condition number: 2.3833, 29.734

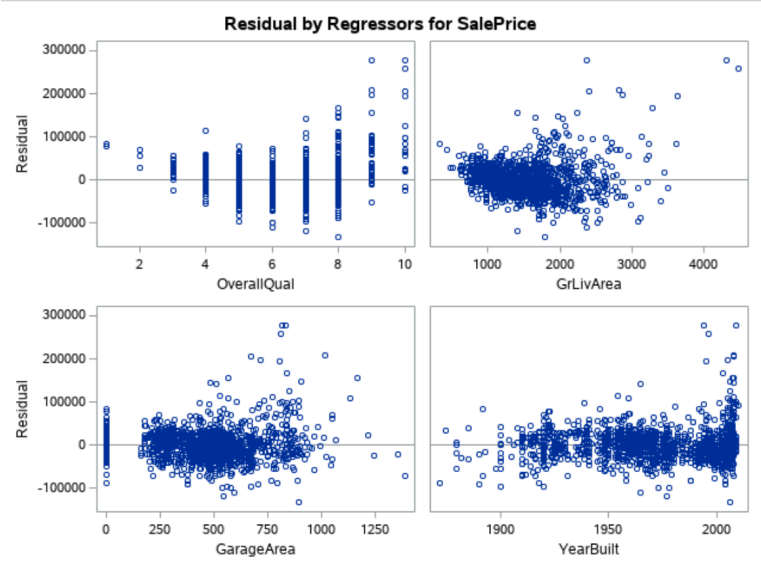
All variables have been entered into the model.

Summary of Forward Selection							
Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	OverallQual	1	0.6320	0.6320	988.238	2497.57	<.0001
2	GrLivArea	2	0.1085	0.7405	270.850	607.41	<.0001
3	GarageArea	3	0.0271	0.7676	93.3202	169.13	<.0001
4	YearBuilt	4	0.0136	0.7812	5.0000	90.32	<.0001

(Table 2.23)



(Plot 2.24)



(Plot 2.25)

Adj R-Square = 0.7806 after 4 steps
Looks like there is curve relationship between SalePrice and OverallQual

```
/* 2: Build model3 with log transformation */
proc reg data=trainames2Nooutlier;
model logSalePrice = logOverallQual GrLivArea GarageArea YearBuilt / selection=forward
slentry=0.1 slstay=0.1 adjrsq;
run;
```

Forward Selection: Step 4

Variable GarageArea Entered: R-Square = 0.8215 and C(p) = 5.0000

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	190.43980	47.60995	1668.92	<.0001
Error	1451	41.39317	0.02853		
Corrected Total	1455	231.83297			

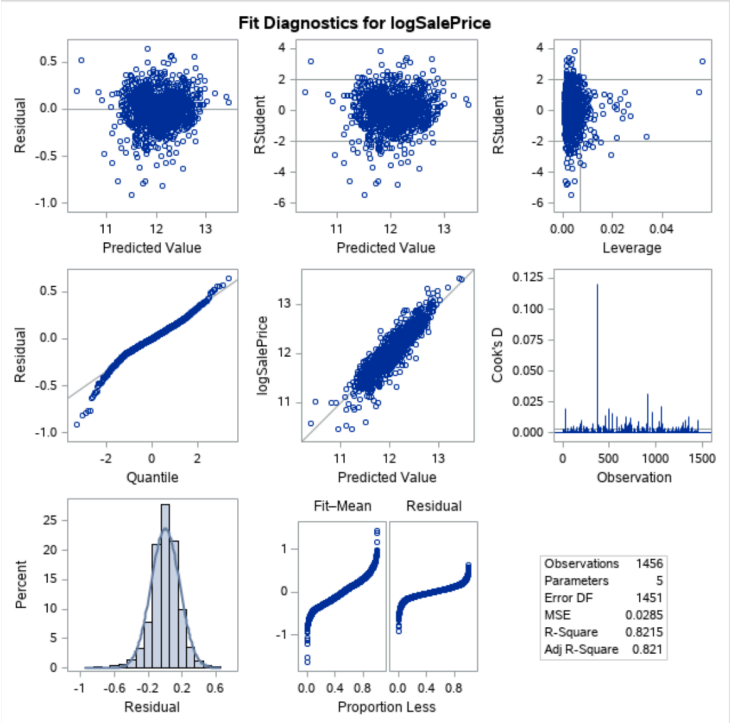
Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	4.34765	0.35084	4.38069	153.56	<.0001
logOverallQual	0.57644	0.02697	13.03661	456.99	<.0001
GrLivArea	0.00031753	0.00001117	23.05313	808.11	<.0001
GarageArea	0.00033580	0.00002655	4.56209	159.92	<.0001
YearBuilt	0.00305	0.00018829	7.48480	262.37	<.0001

Bounds on condition number: 2.1457, 28.208

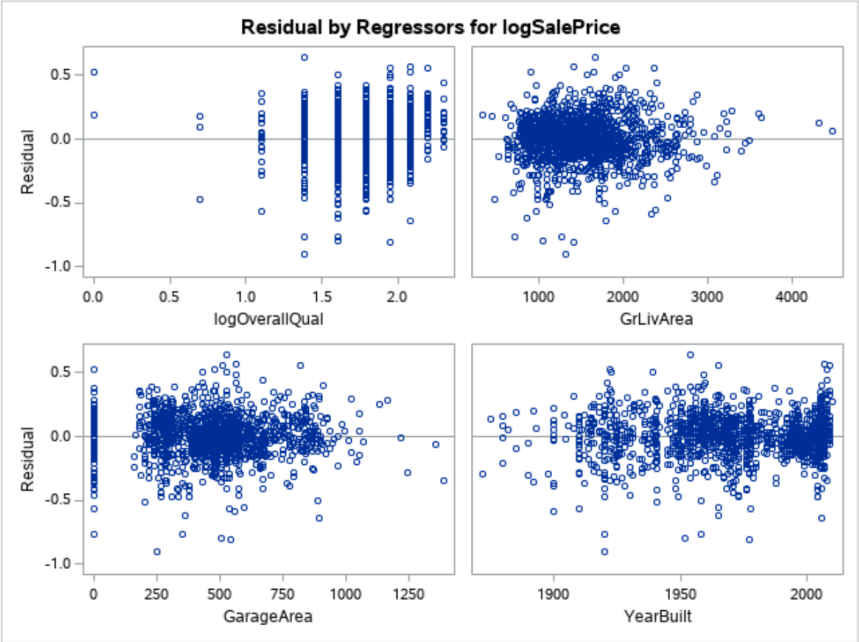
All variables have been entered into the model.

Summary of Forward Selection							
Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	logOverallQual	1	0.6332	0.6332	1528.89	2509.99	<.0001
2	GrLivArea	2	0.1131	0.7463	611.633	647.87	<.0001
3	YearBuilt	3	0.0555	0.8018	162.920	406.25	<.0001
4	GarageArea	4	0.0197	0.8215	5.0000	159.92	<.0001

(Table 2.26)



(Plot 2.27)



(Plot 2.28)

Log transformed variables of OverallQuall and SalePrice are used in model.
Adj R-Square = 0.821 after 4 steps
Residual scatter plots looks much better, straight line in QQ plot.

- Backward Method: Build model4 with 12 numerical predictors

```
/** Backward Method  **/
/* 2: Build model4 with 12 numerical predictors */
proc reg data=trainames2Nooutlier;
model SalePrice = OverallQual OverallCond YearBuilt YearRemodAdd GrLivArea FullBath
TotRmsAbvGrd
 Fireplaces GarageCars GarageArea WoodDeckSF OpenPorchSF / selection=backward slentry=0.1
slstay=0.05 adjrsq aic bic cp partialr2 press sp vif;
run;
```

Summary of Backward Elimination							
Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	GarageCars	11	0.0000	0.8011	11.1511	0.15	0.6975
2	OpenPorchSF	10	0.0004	0.8007	12.1962	3.05	0.0811

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

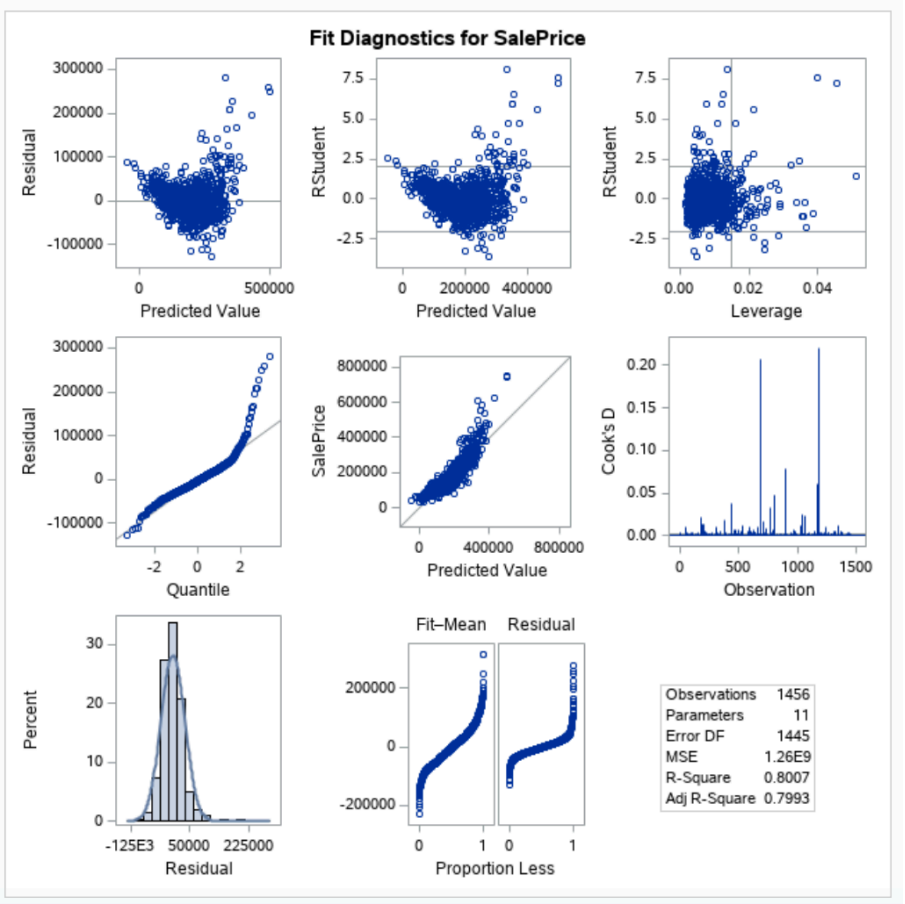
Number of Observations Read	1456
Number of Observations Used	1456

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	7.334888E12	7.334888E11	580.36	<.0001
Error	1445	1.82625E12	1263840837		
Corrected Total	1455	9.161138E12			

Root MSE	35551	R-Square	0.8007
Dependent Mean	180725	Adj R-Sq	0.7993
Coeff Var	19.67110		

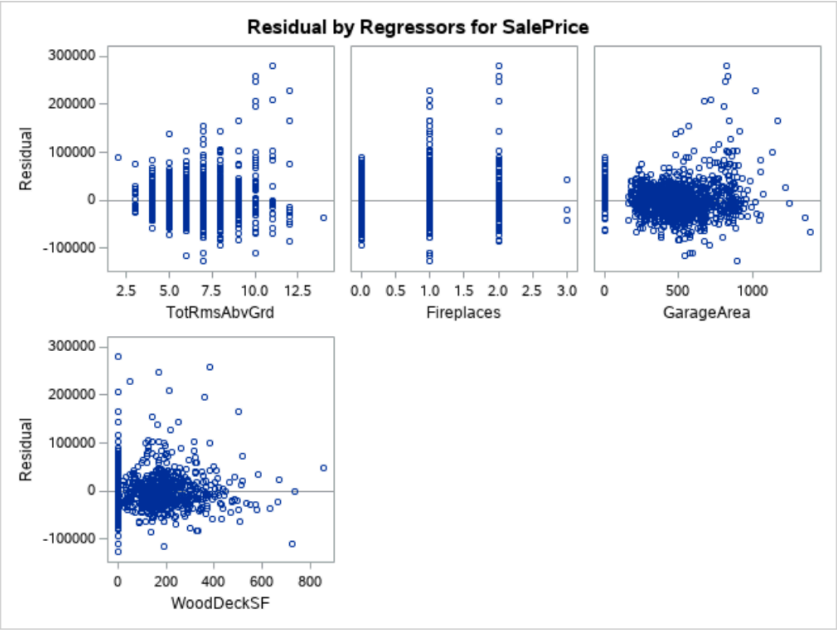
Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Squared Semi-partial Corr Type I	Variance Inflation
Intercept	1	-1368572	121235	-11.29	<.0001	.	0
OverallQual	1	18928	1107.51602	17.09	<.0001	0.63205	2.66704
OverallCond	1	4392.23911	1003.97855	4.37	<.0001	0.00004817	1.43853
YearBuilt	1	483.46133	52.83797	9.15	<.0001	0.00799	2.93140
YearRemodAdd	1	173.34216	64.97593	2.67	0.0077	0.00225	2.07165
GrLivArea	1	74.36409	4.18644	17.76	<.0001	0.12945	5.19141
FullBath	1	-10289	2532.05331	-4.06	<.0001	0.00470	2.23301
TotRmsAbvGrd	1	-2989.37406	1060.38132	-2.82	0.0049	0.00198	3.37712
Fireplaces	1	9977.10851	1697.19101	5.88	<.0001	0.00540	1.35525
GarageArea	1	56.26402	5.63787	9.98	<.0001	0.01431	1.64831
WoodDeckSF	1	33.47555	7.90017	4.24	<.0001	0.00248	1.12635

(Table 2.29)



(Plot 2.30)

(Plot 2.31)



(Plot 2.32)

- Stepwise Method - Build model5 with 12 numerical predictors:

```
/** Stepwise Method  **/
/* 2: Build model5 with 12 numerical predictors */
proc reg data=trainames2Nooutlier;
model SalePrice = OverallQual OverallCond YearBuilt YearRemodAdd GrLivArea FullBath
TotRmsAbvGrd
 Fireplaces GarageCars GarageArea WoodDeckSF OpenPorchSF / selection=stepwise slentry=0.1
slstay=0.05 cp adjrsq aic bic partialr2 press sp vif;
run;
```

Stepwise Selection: Step 12

Variable OpenPorchSF Removed: R-Square = 0.8007 and C(p) = 12.1962

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	7.334888E12	7.334888E11	580.36	<.0001
Error	1445	1.82625E12	1263840837		
Corrected Total	1455	9.161138E12			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-1368572	121235	1.610533E11	127.43	<.0001
OverallQual	18928	1107.51602	3.691427E11	292.08	<.0001
OverallCond	4392.23911	1003.97855	24188863442	19.14	<.0001
YearBuilt	483.46133	52.83797	1.058092E11	83.72	<.0001
YearRemodAdd	173.34216	64.97593	8994889292	7.12	0.0077
GrLivArea	74.36409	4.18644	3.987757E11	315.53	<.0001
FullBath	-10289	2532.05331	20870322480	16.51	<.0001
TotRmsAbvGrd	-2989.37406	1060.38132	10044510277	7.95	0.0049
Fireplaces	9977.10851	1697.19101	43675745824	34.56	<.0001
GarageArea	56.26402	5.63787	1.258706E11	99.59	<.0001
WoodDeckSF	33.47555	7.90017	22692092869	17.95	<.0001

(Table 2.33)

Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	OverallQual		1	0.6320	0.6320	1217.38	2497.57	<.0001
2	GrLivArea		2	0.1085	0.7405	432.442	607.41	<.0001
3	GarageArea		3	0.0271	0.7676	238.054	169.13	<.0001
4	YearBuilt		4	0.0136	0.7812	141.252	90.32	<.0001
5	Fireplaces		5	0.0064	0.7876	96.9407	43.58	<.0001
6	OverallCond		6	0.0057	0.7933	57.7443	39.80	<.0001
7	WoodDeckSF		7	0.0029	0.7962	38.6056	20.70	<.0001
8	FullBath		8	0.0023	0.7985	23.6116	16.82	<.0001
9	TotRmsAbvGrd		9	0.0011	0.7997	17.3192	8.25	0.0041
10	YearRemodAdd		10	0.0010	0.8007	12.1962	7.12	0.0077
11	OpenPorchSF		11	0.0004	0.8011	11.1511	3.05	0.0811
12		OpenPorchSF	10	0.0004	0.8007	12.1962	3.05	0.0811

(Table 2.34)

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Number of Observations Read	1456
Number of Observations Used	1456

Analysis of Variance

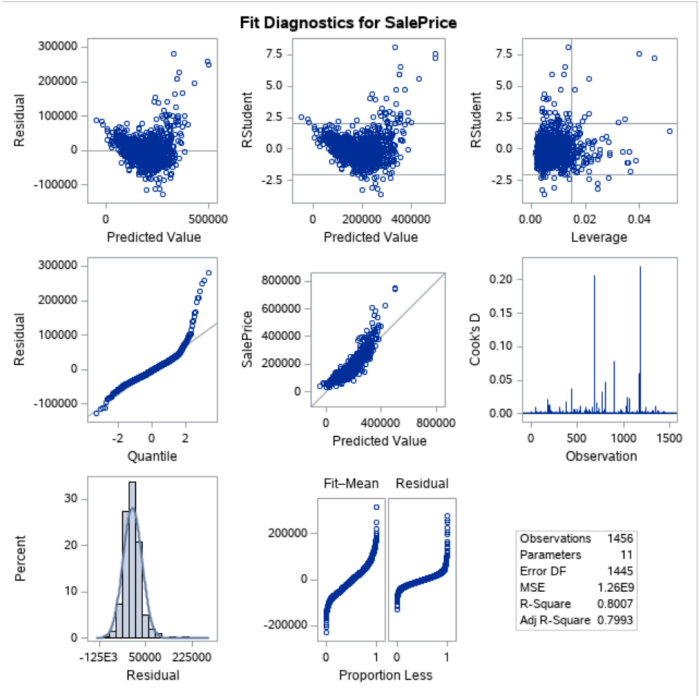
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	7.334888E12	7.334888E11	580.36	<.0001
Error	1445	1.82625E12	1263840837		
Corrected Total	1455	9.161138E12			

Root MSE	35551	R-Square	0.8007
Dependent Mean	180725	Adj R-Sq	0.7993
Coeff Var	19.67110		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Squared Semi-partial Corr Type I	Variance Inflation
Intercept	1	-1368572	121235	-11.29	<.0001	.	0
OverallQual	1	18928	1107.51602	17.09	<.0001	0.63205	2.66704
OverallCond	1	4392.23911	1003.97855	4.37	<.0001	0.00004817	1.43853
YearBuilt	1	483.46133	52.83797	9.15	<.0001	0.00799	2.93140
YearRemodAdd	1	173.34216	64.97593	2.67	0.0077	0.00225	2.07165
GrLivArea	1	74.36409	4.18644	17.76	<.0001	0.12945	5.19141
FullBath	1	-10289	2532.05331	-4.06	<.0001	0.00470	2.23301
TotRmsAbvGrd	1	-2989.37406	1060.38132	-2.82	0.0049	0.00198	3.37712
Fireplaces	1	9977.10851	1697.19101	5.88	<.0001	0.00540	1.35525
GarageArea	1	56.26402	5.63787	9.98	<.0001	0.01431	1.64831
WoodDeckSF	1	33.47555	7.90017	4.24	<.0001	0.00248	1.12635

(Table 2.35)



(Plot 2.36)

Adj R-Square = 0.7993 after 12 steps.

The top 4 variables with the highest Partial R^2 will be selected in the model.

```
/* 2: Build model6 with log transformation */
proc reg data=trainames2Nooutlier;
model logSalePrice = logOverallQual GrLivArea GarageArea YearBuilt / selection=stepwise
slentry=0.1 slstay=0.05 cp adjrsq aic bic partialr2 press sp vif;
run;
```

Stepwise Selection: Step 4

Variable GarageArea Entered: R-Square = 0.8215 and C(p) = 5.0000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	190.43980	47.60995	1668.92	<.0001
Error	1451	41.39317	0.02853		
Corrected Total	1455	231.83297			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	4.34765	0.35084	4.38069	153.56	<.0001
logOverallQual	0.57644	0.02697	13.03661	456.99	<.0001
GrLivArea	0.00031753	0.00001117	23.05313	808.11	<.0001
GarageArea	0.00033580	0.00002655	4.56209	159.92	<.0001
YearBuilt	0.00305	0.00018829	7.48480	262.37	<.0001

Bounds on condition number: 2.1457, 28.208

All variables left in the model are significant at the 0.0500 level.

All variables have been entered into the model.

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	logOverallQual		1	0.6332	0.6332	1528.89	2509.99	<.0001
2	GrLivArea		2	0.1131	0.7463	611.633	647.87	<.0001
3	YearBuilt		3	0.0555	0.8018	162.920	406.25	<.0001
4	GarageArea		4	0.0197	0.8215	5.0000	159.92	<.0001

(Table 2.37)

The REG Procedure
Model: MODEL1
Dependent Variable: logSalePrice

Number of Observations Read	1456
Number of Observations Used	1456

Analysis of Variance

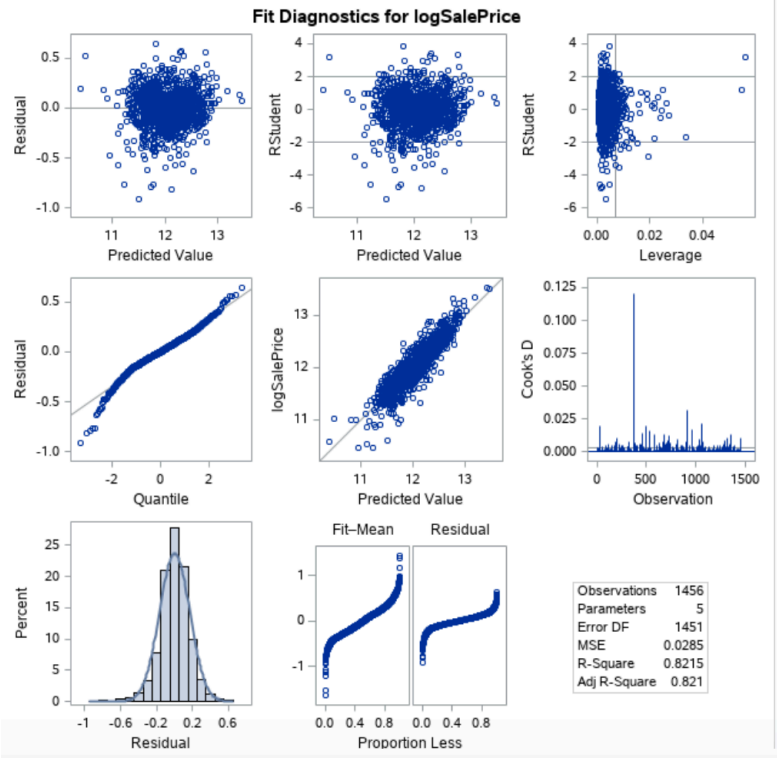
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	190.43980	47.60995	1668.92	<.0001
Error	1451	41.39317	0.02853		
Corrected Total	1455	231.83297			

Root MSE	0.16890	R-Square	0.8215
Dependent Mean	12.02306	Adj R-Sq	0.8210
Coeff Var	1.40480		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Squared Semi-partial Corr Type I	Variance Inflation
Intercept	1	4.34765	0.35084	12.39	<.0001	.	0
logOverallQual	1	0.57644	0.02697	21.38	<.0001	0.63320	2.14570
GrLivArea	1	0.00031753	0.00001117	28.43	<.0001	0.11312	1.63733
GarageArea	1	0.00033580	0.00002655	12.65	<.0001	0.04285	1.61992
YearBuilt	1	0.00305	0.00018829	16.20	<.0001	0.03229	1.64912

(Table 2.38)



(Plot 2.39)

- We will create a custom model with categorical variables added:

```
/** Custom Model with Categorical Variable added  **/
/* create new data set with categorical variables added */
data trainAmes3;
set train;
/* Removing variables with NA's: LotFrontage MasVnrArea GarageYrBlt */
/* Removing variables with wrong name: 1stFlrSF 2ndFlrSF 3SsnPorch */
keep Id MSSubClass LotArea OverallQual OverallCond YearBuilt YearRemodAdd BsmtFinSF1
BsmtFinSF2 BsmtUnfSF
 LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath HalfBath BedroomAbvGr
KitchenAbvGr TotRmsAbvGrd
 Fireplaces GarageCars GarageArea WoodDeckSF OpenPorchSF EnclosedPorch ScreenPorch
PoolArea MiscVal MoSold
 YrSold Utilities LotConfig Neighborhood ExterQual BsmtQual HeatingQC KitchenQual
FireplaceQu GarageQual
 PoolQC Fence SaleType SaleCondition SalePrice;
run;
data trainAmes3NoOutlier;
set trainAmes3;
where Id ~= 643 AND Id ~= 725 AND Id ~= 1299 AND Id ~= 524;
logSalePrice = log(SalePrice);
```

```
logOverallQual = log(OverallQual);
logGrLivArea = log(GrLivArea);
/** 10 Categorical => Numeric **/
/** LotConfig, Neighborhood, ExterQual, BsmtQual, HeatingQC, KitchenQual,
FireplaceQu, GarageQual, PoolQC, SaleCondition **/

/*** LotConfig => LotConfig_num ***/
if LotConfig="Inside" then LotConfig_num =5;
  else if LotConfig="Corner" then LotConfig_num =4;
  else if LotConfig="CulDSac" then LotConfig_num =3;
  else if LotConfig="FR2" then LotConfig_num =2;
  else if LotConfig="FR3" then LotConfig_num =1;

/*** Neighborhood => Neighborhood_num ***/
if Neighborhood="Blmngtn" then Neighborhood_num =25;
  else if Neighborhood="Blueste" then Neighborhood_num =24;
  else if Neighborhood="BrDale" then Neighborhood_num =23;
  else if Neighborhood="BrkSide" then Neighborhood_num =22;
  else if Neighborhood="ClearCr" then Neighborhood_num =21;
  else if Neighborhood="CollgCr" then Neighborhood_num =20;
  else if Neighborhood="Crawfor" then Neighborhood_num =19;
  else if Neighborhood="Edwards" then Neighborhood_num =18;
  else if Neighborhood="Gilbert" then Neighborhood_num =17;
  else if Neighborhood="IDOTRR" then Neighborhood_num =16;
  else if Neighborhood="MeadowV" then Neighborhood_num =15;
  else if Neighborhood="Mitchel" then Neighborhood_num =14;
  else if Neighborhood="Names" then Neighborhood_num =13;
  else if Neighborhood="NoRidge" then Neighborhood_num =12;
  else if Neighborhood="NPkVill" then Neighborhood_num =11;
  else if Neighborhood="NridgHt" then Neighborhood_num =10;
  else if Neighborhood="NWAmes" then Neighborhood_num =9;
  else if Neighborhood="OldTown" then Neighborhood_num =8;
  else if Neighborhood="SWISU" then Neighborhood_num =7;
  else if Neighborhood="Sawyer" then Neighborhood_num =6;
  else if Neighborhood="SawyerW" then Neighborhood_num =5;
  else if Neighborhood="Somerst" then Neighborhood_num =4;
  else if Neighborhood="StoneBr" then Neighborhood_num =3;
  else if Neighborhood="Timber" then Neighborhood_num =2;
  else if Neighborhood="Veenker" then Neighborhood_num =1;

/*** ExterQual => ExterQual_num ***/
if ExterQual="Ex" then ExterQual_num =5;
  else if ExterQual="Gd" then ExterQual_num =4;
  else if ExterQual="TA" then ExterQual_num =3;
```

```
  else if ExterQual="Fa" then ExterQual_num =2;
  else if ExterQual="Po" then ExterQual_num =1;

/*** BsmtQual => BsmtQual_num ***/
if BsmtQual="Ex" then BsmtQual_num =5;
  else if BsmtQual="Gd" then BsmtQual_num =4;
  else if BsmtQual="TA" then BsmtQual_num =3;
  else if BsmtQual="Fa" then BsmtQual_num =2;
  else if BsmtQual="Po" then BsmtQual_num =1;
  else if BsmtQual="NA" then BsmtQual_num =0;

/*** HeatingQC => HeatingQC_num ***/
if HeatingQC="Ex" then HeatingQC_num =5;
  else if HeatingQC="Gd" then HeatingQC_num =4;
  else if HeatingQC="TA" then HeatingQC_num =3;
  else if HeatingQC="Fa" then HeatingQC_num =2;
  else if HeatingQC="Po" then HeatingQC_num =1;

/*** KitchenQual => KitchenQual_num ***/
if KitchenQual="Ex" then KitchenQual_num =5;
  else if KitchenQual="Gd" then KitchenQual_num =4;
  else if KitchenQual="TA" then KitchenQual_num =3;
  else if KitchenQual="Fa" then KitchenQual_num =2;
  else if KitchenQual="Po" then KitchenQual_num =1;

/*** FireplaceQu => FireplaceQu_num ***/
if FireplaceQu="Ex" then FireplaceQu_num =5;
  else if FireplaceQu="Gd" then FireplaceQu_num =4;
  else if FireplaceQu="TA" then FireplaceQu_num =3;
  else if FireplaceQu="Fa" then FireplaceQu_num =2;
  else if FireplaceQu="Po" then FireplaceQu_num =1;
  else if FireplaceQu="NA" then FireplaceQu_num =0;

if GarageQual="Ex" then GarageQual_num =5;
  else if GarageQual="Gd" then GarageQual_num =4;
  else if GarageQual="TA" then GarageQual_num =3;
  else if GarageQual="Fa" then GarageQual_num =2;
  else if GarageQual="Po" then GarageQual_num =1;
  else if GarageQual="NA" then GarageQual_num =0;

if PoolQC="Ex" then PoolQC_num =5;
  else if PoolQC="Gd" then PoolQC_num =4;
  else if PoolQC="TA" then PoolQC_num =3;
  else if PoolQC="Fa" then PoolQC_num =2;
```

```
  else if PoolQC="NA" then PoolQC_num =1;

if SaleCondition="Normal" then SaleCondition_num =6;
  else if SaleCondition="Abnorml" then SaleCondition_num =5;
  else if SaleCondition="AdjLand" then SaleCondition_num =4;
  else if SaleCondition="Alloca" then SaleCondition_num =3;
  else if SaleCondition="Family" then SaleCondition_num =2;
  else if SaleCondition="Partial" then SaleCondition_num =1;
run;

proc contents data=trainAmes3NoOutlier; run;
proc print data=trainAmes3NoOutlier; run;
```

- Forward selection with Cross Validation (CV):

```
/**  class ExterQual; **/
/** Forward selection with Cross Validation (CV) **/
proc glmselect data=trainAmes3NoOutlier;
model logSalePrice = logOverallQual GrLivArea GarageArea YearBuilt OverallCond
YearRemodAdd
 FullBath TotRmsAbvGrd Fireplaces GarageCars WoodDeckSF OpenPorchSF LotConfig_num
 Neighborhood_num ExterQual_num BsmtQual_num HeatingQC_num KitchenQual_num
FireplaceQu_num
 GarageQual_num PoolQC_num SaleCondition_num
/selection=Forward(stop=cv) cvmethod=random(5) stats=adjrsq;
run;
```

- Backward selection with Cross Validation (CV):

```
/** Backward selection with Cross Validation (CV) **/
proc glmselect data=trainAmes3NoOutlier;
model logSalePrice = logOverallQual GrLivArea GarageArea YearBuilt OverallCond
YearRemodAdd
 FullBath TotRmsAbvGrd Fireplaces GarageCars WoodDeckSF OpenPorchSF LotConfig_num
 Neighborhood_num ExterQual_num BsmtQual_num HeatingQC_num KitchenQual_num
FireplaceQu_num
 GarageQual_num PoolQC_num SaleCondition_num
/selection=Backward(stop=cv) cvmethod=random(5) stats=adjrsq;
run;
```

- Stepwise selection with Cross Validation (CV):

```
/** Stepwise selection with Cross Validation (CV) **/
proc glmselect data=trainAmes3NoOutlier;
model logSalePrice = logOverallQual GrLivArea GarageArea YearBuilt OverallCond
YearRemodAdd
 FullBath TotRmsAbvGrd Fireplaces GarageCars WoodDeckSF OpenPorchSF LotConfig_num
 Neighborhood_num ExterQual_num BsmtQual_num HeatingQC_num KitchenQual_num
FireplaceQu_num
 GarageQual_num PoolQC_num SaleCondition_num
/selection=Stepwise(stop=cv) cvmethod=random(5) stats=adjrsq;
run;
```

- Kaggle upload:

```
/** Prepare data for kaggle uploading **/
data test;
set test;
if GarageArea='' then GarageArea=0;
if GarageCars='' then GarageCars=0;
if KitchenQual='NA' then KitchenQual='Po';
SalePrice = .;
run;

data train2;
set train test;
run;

data train2Ames3;
set train2;
/* Removing variables with NA's: LotFrontage MasVnrArea GarageYrBlt */
/* Removing variables with wrong name: 1stFlrSF 2ndFlrSF 3SsnPorch */
keep Id OverallQual GrLivArea GarageArea YearBuilt OverallCond YearRemodAdd FullBath
 TotRmsAbvGrd Fireplaces GarageCars WoodDeckSF OpenPorchSF LotConfig Neighborhood
ExterQual
 BsmtQual HeatingQC KitchenQual FireplaceQu GarageQual PoolQC SaleCondition SalePrice;
run;

data train2Ames3NoOutlier;
set train2Ames3;
where Id ~= 643 AND Id ~= 725 AND Id ~= 1299 AND Id ~= 524;
logSalePrice = log(SalePrice);
logOverallQual = log(OverallQual);
logGrLivArea = log(GrLivArea);
/** 10 Categorical => Numeric **/
/** LotConfig, Neighborhood, ExterQual, BsmtQual, HeatingQC, KitchenQual,
```

```
FireplaceQu, GarageQual, PoolQC, SaleCondition **/

/*** LotConfig => LotConfig_num ***/
if LotConfig="Inside" then LotConfig_num =5;
  else if LotConfig="Corner" then LotConfig_num =4;
  else if LotConfig="CulDSac" then LotConfig_num =3;
  else if LotConfig="FR2" then LotConfig_num =2;
  else if LotConfig="FR3" then LotConfig_num =1;

/*** Neighborhood => Neighborhood_num ***/
if Neighborhood="Blmngtn" then Neighborhood_num =25;
  else if Neighborhood="Blueste" then Neighborhood_num =24;
  else if Neighborhood="BrDale" then Neighborhood_num =23;
  else if Neighborhood="BrkSide" then Neighborhood_num =22;
  else if Neighborhood="ClearCr" then Neighborhood_num =21;
  else if Neighborhood="CollgCr" then Neighborhood_num =20;
  else if Neighborhood="Crawfor" then Neighborhood_num =19;
  else if Neighborhood="Edwards" then Neighborhood_num =18;
  else if Neighborhood="Gilbert" then Neighborhood_num =17;
  else if Neighborhood="IDOTRR" then Neighborhood_num =16;
  else if Neighborhood="MeadowV" then Neighborhood_num =15;
  else if Neighborhood="Mitchel" then Neighborhood_num =14;
  else if Neighborhood="NAmes" then Neighborhood_num =13;
  else if Neighborhood="NoRidge" then Neighborhood_num =12;
  else if Neighborhood="NPkVill" then Neighborhood_num =11;
  else if Neighborhood="NridgHt" then Neighborhood_num =10;
  else if Neighborhood="NWAmes" then Neighborhood_num =9;
  else if Neighborhood="OldTown" then Neighborhood_num =8;
  else if Neighborhood="SWISU" then Neighborhood_num =7;
  else if Neighborhood="Sawyer" then Neighborhood_num =6;
  else if Neighborhood="SawyerW" then Neighborhood_num =5;
  else if Neighborhood="Somerst" then Neighborhood_num =4;
  else if Neighborhood="StoneBr" then Neighborhood_num =3;
  else if Neighborhood="Timber" then Neighborhood_num =2;
  else if Neighborhood="Veenker" then Neighborhood_num =1;

/*** ExterQual => ExterQual_num ***/
if ExterQual="Ex" then ExterQual_num =5;
  else if ExterQual="Gd" then ExterQual_num =4;
  else if ExterQual="TA" then ExterQual_num =3;
  else if ExterQual="Fa" then ExterQual_num =2;
  else if ExterQual="Po" then ExterQual_num =1;

/*** BsmtQual => BsmtQual_num ***/
```

```
if BsmtQual="Ex" then BsmtQual_num =6;
  else if BsmtQual="Gd" then BsmtQual_num =5;
  else if BsmtQual="TA" then BsmtQual_num =4;
  else if BsmtQual="Fa" then BsmtQual_num =3;
  else if BsmtQual="Po" then BsmtQual_num =2;
  else if BsmtQual="NA" then BsmtQual_num =1;

/*** HeatingQC => HeatingQC_num ***/
if HeatingQC="Ex" then HeatingQC_num =5;
  else if HeatingQC="Gd" then HeatingQC_num =4;
  else if HeatingQC="TA" then HeatingQC_num =3;
  else if HeatingQC="Fa" then HeatingQC_num =2;
  else if HeatingQC="Po" then HeatingQC_num =1;

/*** KitchenQual => KitchenQual_num ***/
if KitchenQual="Ex" then KitchenQual_num =5;
  else if KitchenQual="Gd" then KitchenQual_num =4;
  else if KitchenQual="TA" then KitchenQual_num =3;
  else if KitchenQual="Fa" then KitchenQual_num =2;
  else if KitchenQual="Po" then KitchenQual_num =1;

/*** FireplaceQu => FireplaceQu_num ***/
if FireplaceQu="Ex" then FireplaceQu_num =6;
  else if FireplaceQu="Gd" then FireplaceQu_num =5;
  else if FireplaceQu="TA" then FireplaceQu_num =4;
  else if FireplaceQu="Fa" then FireplaceQu_num =3;
  else if FireplaceQu="Po" then FireplaceQu_num =2;
  else if FireplaceQu="NA" then FireplaceQu_num =1;

if GarageQual="Ex" then GarageQual_num =6;
  else if GarageQual="Gd" then GarageQual_num =5;
  else if GarageQual="TA" then GarageQual_num =4;
  else if GarageQual="Fa" then GarageQual_num =3;
  else if GarageQual="Po" then GarageQual_num =2;
  else if GarageQual="NA" then GarageQual_num =1;

if PoolQC="Ex" then PoolQC_num =5;
  else if PoolQC="Gd" then PoolQC_num =4;
  else if PoolQC="TA" then PoolQC_num =3;
  else if PoolQC="Fa" then PoolQC_num =2;
  else if PoolQC="NA" then PoolQC_num =1;

if SaleCondition="Normal" then SaleCondition_num =6;
  else if SaleCondition="Abnorml" then SaleCondition_num =5;
```

```
  else if SaleCondition="AdjLand" then SaleCondition_num =4;
  else if SaleCondition="Alloca" then SaleCondition_num =3;
  else if SaleCondition="Family" then SaleCondition_num =2;
  else if SaleCondition="Partial" then SaleCondition_num =1;
run;

/* Kaggle score: 0.15115, position# 3266 */
proc glmselect data=train2Ames3NoOutlier;
model logSalePrice = logOverallQual | GrLivArea GarageArea | YearBuilt | OverallCond |
YearRemodAdd
 FullBath TotRmsAbvGrd Fireplaces GarageCars WoodDeckSF OpenPorchSF LotConfig_num
 Neighborhood_num ExterQual_num BsmtQual_num HeatingQC_num KitchenQual_num
FireplaceQu_num
 GarageQual_num PoolQC_num SaleCondition_num
/selection=Backward(stop=cv) cvmethod=random(5) stats=adjrsq;
output out= results p= predict;
run;

proc print data=train2Ames3NoOutlier;
where Id=1461;
run;
data results2;
set results;
where id > 1460;
logSalePrice = predict;
SalePrice = 2.718281828459045235360287471352 7**logSalePrice;
keep id logSalePrice SalePrice;

proc print data=results2;
run;

proc means data=results2;
var SalePrice;
run;
```