Kaggle Project (MSDS 6371)

David Camacho & Stephanie Duarte

Introduction:

We have been asked by Century 21 Ames (a real estate company) in Ames Iowa to get an estimate of the sale price of a house based on the square footage of the living area and to see the sales price (and relationship to square footage) depending on which neighborhood the house is located in for the NAmes, Edwards, and BrkSide neighborhoods. Therefore Century 21 would like for us to build the most predictive model for sales prices of homes in all of Ames, Iowa. This includes all neighborhoods.

Data Description

The Ames Housing dataset was compiled by Dean De Cock, and is available to download via Kaggle.com. While the entire training data set examines 1460 observations of 79 different variables of home ownership in Ames, Iowa, for example, square footage, lot size, number of bathrooms, number of bedrooms, etc, more information about all the variables can be found on the Kaggle website. For the first analysis we focused on what our client, Century 21, is interested in, which includes how the sales price of a home is related to the square footage of the living area of the house and if the SalesPrice (and its relationship to square footage) depends on which neighborhood the house is located in the three neighborhoods they sell in, which are the NAmes, Edwards, and BrkSide neighborhoods. For the second analysis, we will consider all neighborhoods and we conducted four separate types of regression stepwise, forward, and backward, and a custom model.

Analysis of Question 1

Restatement of Problem:

Century 21 Ames (a real estate company) in Ames, Iowa has commissioned us to analyze how the sale price of a house is related to the square footage of the living area of the house (GrLIvArea) based on its square footage of living area, and to see if the sales price (and relationship to square footage) depends on which neighborhood the house is located in.The company only sells houses in the NAmes, Edwards and BrkSide neighborhoods.In order to compete this analysis, we will restrict our model to only focus on these neighborhoods' variables.

Build and Fit the Model:

The first step was to examine a scatter plot of SalePrice vs GrLivArea by neighborhood [see Appendix, Figure 1.1].The results from this appear to demonstrate a positive linear relationship between the square footage living area and sale price. However, there are some clear outliers that we will need to review within the modeling stages.

1. **First tentative Model:**

Model 1: μ(SalePrice) = b0 + b1 (GrLIvArea)

The following observations were made after viewing Appendix Figures 1.1 and 1.3 to review the assumptions of regression.

- Linearity: There appears to be a linear trend. However, there appear to be some deviations at the higher end.
- Normality: Based on the histogram of residuals this appears relatively normal.
- Independence: Since we are looking at specific neighborhoods there could be a possible clustering effect, but we will assume independence, although not much is known about how these houses were selected.
- Constant Variance (Equal Spread): The QQ Plot appears mostly linear, while there is a significant amount of clustering within the residual plot, likely due to outliers. The confidence and prediction bands widening as GrLivArea increases, suggesting that the variance of the residuals may be increasing (heteroscedasticity).
- Leverage: The leverage plot identifies points that have more influence on the parameter estimates than is typical. Points with high leverage can have a large impact on the direction and slope of the regression line. It seems there are a few points with high leverage, but without numerical values, it's hard to quantify their exact influence.

We checked various model transformations such as log-linear and log-log, however these did not appear to improve the residual plots.

From plot 1.5 [Appendix], we can see four outliers with studentized residuals greater than 2.5 and one outlier with Cook's D greater than 5. The adjusted R-Square is 0.3406.We checked various model transformations such as log-linear and log-log, however these did not appear to improve the residual plots, therefore the outliers mentioned were removed in the following analysis.

2. **Second tentative Model: Re-ran the first model without the outliers.**

For the second model the following observations were made after viewing Appendix Figures 1.6 and 1.7 to check the assumptions of regressions below.

- Linearity: There appears to be a positive linear trend.
- Normality: The residuals largely follow the reference line, but there is some deviation at the ends. This could indicate that the residuals have heavier tails than a normal distribution.

- Independence: Since we are looking at specific neighborhoods there could be a possible clustering effect, but we will assume independence, although not much is known about how these houses were selected.
- Constant Variance (Equal Spread): The QQ Plot appears mostly linear. The plot provided does not show a clear pattern of increasing or decreasing variance, which is good. However, there seems to be a slight funnel shape, indicating potential heteroscedasticity.

From plot 1.8 [Appendix], we can see a straight line in the QQ plots and symmetric histogram that indicates the normal distribution. The adjusted R-Square is 0.449.

3. **Third tentative model including the Neighborhood variables with interactions:**

μ(SalePrice) = b0 + b1 (GrLIvArea) + b2 (GrLivArea*Neighborhood)

For the third model the following observations were made after viewing Appendix Figures Figures 2.2, 2.3, 2.4, and 2.5 to check the assumptions of regressions below.

- Linearity: The graphs appear to show a positive linear trend within each neighborhood, although the relationship may not be strong, especially for 'Edwards', where the data is more dispersed.
- Normality: The histogram of residuals appears relatively normal and improved with outliers removed and neighborhood interactions added.
- Independence: Same as above. The provided plots do not indicate a time component, so we would need additional information to assess this properly.
- Constant Variance (Equal Spread): The QQ Plot appears linear, there has been substantial improvement in the residual plot (more randomly distributed).
- Adjusted R-square = 0.5165.

This model appears to be the best fitting and it does not appear to need a transformation. Since we used interactions for each of the neighborhoods, a separate regression was written for each using the SA output in Appendix Figure 2.1.

- Regression model for NAmes neighborhood:
  - μ(SalePrice|NAmes) = 80325.71 +49.56*GrLivArea
- Regression model for BrkSide neighborhood:
  - μ(SalePrice|NAmes) =19971.51 +87.16*GrLivArea
- Regression model forEdwards neighborhood:
  - μ(SalePrice|NAmes = 37100.42+70.16*GrLivArea

Conclusion and Interpretation:

This model suggests that the linear regression is a good fit to the data set of the three neighborhoods, it's a good fit based on significant F-test= 81.76 and p-values is <.0001 with degree of freedom of (5, 373). The R-square= 0.5228, meaning that 52.28% of the variability of sale price can be explained by the living area square footage. It looks like neighborhood Edwards has the highest estimated mean of sale price followed by BrkSide and NAmes.

In the neighborhood for NAmes, every 100 sq. ft living area increase resulted in an estimated \$4,956 increase on sale price(see Appendix for valuation details), with 95% confidence interval from \$1,497 to \$8,404. In the neighborhood for BrkSide, every 100 sq. ft living area increase resulted in an estimated \$8,716 increase on sale price(see Appendix for valuation details), with 95% confidence interval from \$7,052 to \$10,340. In the neighborhood for Edwards, every 100 sq. ft living area increase resulted in an estimated \$7,016 increase on sale price, with 95% confidence interval from \$2,491 to \$10,334.

Since this was an observational study, we cannot make any causal inference. However there is a positive correlation between sale price, square footage and neighborhoods. There was no mention of random sampling so caution should be used in generalizing results.

Rshiny App:

R Shiny App :Scatterplot of price of the home v. square footage (GrLivArea)

Analysis of Question 2:

Restatement of Problem:

We have been commissioned to build the most predictive model for sales prices of homes in all of Ames, Iowa. This includes all neighborhoods. We will produce the following competing model: a simple linear regression model where we have the freedom to pick our explanatory variable, a multiple linear regression model (SalePrice~GrLivArea + FullBath) and at least one additional multiple linear regression model where we selected the explanatory variables. We will generate an adjusted R^2, CV Press, and Kaggle Score for each of these models and clearly describe which model we feel is best in terms of being able to predict future sale prices of homes in Ames, Iowa.

1. **Simple Linear Regression**

   Model Selection: Log(SalePrice) ~ Log(GrLivArea)

   Since we were constrained to a single explanatory variable for this model, we generated scatter plots for the variables that we believed would correlate with SalePrice. The scatter plots can be found in the appendix from figures 3.01 - 3.16. After generating said scatter plots, we noticed curvilinear association. We then proceeded to log-transform the most visually-promising variables and regenerated scatter plots against SalePrice as seen in figure 3.17. Because we still observed curvilinear association, we regenerated scatter plots based on the log-transformed variables but this time against the log-transformed dependent variable (SalePrice_log) as seen in figures 3.18-3.25. We proceeded to fit simple linear regression models of SalePrice_log against the following explanatory variables as seen in figures 3.26-3.37: GrLivArea_log, FirstFlrSf_log, TotalBsmtsf_log, and GarageArea_log. Out of the four, the best fitting model was

SalePrice_log~GrLivArea_log. However, as seen in figures 3.26-3.27, we can observe a couple of outliers that may be affecting the fit. As a result, we observed them and decided to remove them from the data set as no certain explanation was evident. After removing the outliers we generated the following simple linear regression model as seen in figures 3.38-3.4. We observe that GrLivArea is a statistically significant explanatory variable (p-value < 0.0001) (t-value: 41.46).

We are 95% confident that for each doubling of the GrLivArea the median sale price will increase between (1.83, 1.87). Our best estimate is an increase of 1.85 as seen in figure 3.41.

log(SalePrice) = 5.562069 + .889567 * log(GrLivArea)

Checking Assumptions: The following observations and assumptions were made as seen in figure 3.41.

- Linearity: The graphs appear to show a positive linear trend after log-transforming both SalePrice and GrLivArea
- Normality: The histogram of residuals appears relatively normal and improved with outliers removed.
- Independence: We will assume that the observations are independent as this does not seem to be compromised.
- Constant Variance (Equal Spread): The QQ Plot appears linear and there has been substantial improvement in the residual plot after the transformations
- Influential Point Analysis: Based on Cook's D all points seem to be under the .025 value meaning they have low influence on the regression model.

2. **Multiple Linear Regression**

Model Selection: Log(SalePrice) ~ Log(GrLivArea) + FullBath

For this analysis, we expand to a multiple linear regression model and fit SalePrice with respect to GrLiveArea + FullBath. As always, we first plot the data to look for a linear correlation, if any. The scatter plots with each log-transformations can be found in figures 3.42-3.47. After visually observing the correlation amongst the two variables, we fit the following model as seen in figures 3.50-3.51:

log(SalePrice) = 6.507627 + .728027 * log(GrLiveArea) + .144431 * FullBath

Although our statistics look favorable, there are a couple of outliers that we want to take care of before proceeding with the final model. Our final model gave us an Adjusted R-Squared of .5620 and a CV Press of 102.37840. As mentioned in our previous analysis we removed observation 1299.

Checking Assumptions: The following observations and assumptions were made as seen in figure 3.54.

- Linearity: The graphs appear to show a positive linear trend after log-transforming both SalePrice and GrLivArea and leaving FullBath in it's original scale
- Normality: The histogram of residuals appears relatively normal and improved with outlier ID 1299 removed.
- Independence: We will assume that the observations are independent as this does not seem to be compromised.
- Constant Variance (Equal Spread): The QQ Plot appears linear and there has been substantial improvement in the residual plot after the transformations although towards the bottom it may have a slight tail.
- Influential Point Analysis: Based on Cook's D all points seem to be under the .05 value meaning they have low influence on the regression model.

3. Custom Multiple Linear Regression Model

Model Selection: Stepwise - Log(SalePrice) ~ Log(OverallQual) + Log(GrLivArea) + Log(FirstFlrSf) + LotArea + FullBath

For this analysis, we expanded to a custom multiple linear regression model and fit Log(SalePrice) with respect Log(OverallQual) + Log(GrLivArea) + Log(FirstFlrSf) + LotArea + FullBath. Because we already had the scatter plots and previous domain knowledge based on our initial exploratory data analysis, we chose those variables. Out of the three models, this was the best fitting model with an adjusted r-squared of .7832 and a cv press of 51.67 as seen in figures 3.55 - 3.56.

Checking Assumptions: The following observations and assumptions were made as seen in figure 3.57.

- Linearity: The graphs appear to show a positive linear trend after log-transforming SalePrice OverallQual GrLivArea FirstFlrSf and LotArea and Fullbth in their original scale. Normality: The histogram of residuals appears relatively normal.
- Independence: We will assume that the observations are independent as this does not seem to be compromised.
- Constant Variance (Equal Spread): The QQ Plot appears linear and there has been substantial improvement in the residual plot after the transformations although towards the bottom it may have a slight tail.
- Influential Point Analysis: Based on Cook's D all points seem to be under the .3 value meaning they have low influence on the regression model.

4. Comparing Competing Models

Predictive Models	Adjusted R2	CV Press	Kaggle Score
Simple Linear Regression	.5431	103.34249	.28909
Multiple Linear Regression	.5620	102.37840	.2842
Custom MLR Model	.7825	51.67758	.1845

5. Conclusion:

Our preferred model was the custom multiple linear regression model utilizing a stepwise selection with a kaggle score of .1845, cv press of 51.67758, and an adjusted r-squared of .7825. Using our exploratory data analysis, domain knowledge, and stepwise selection, we found that the best fitting model was

Log(SalePrice) = 6.51035 + .84582 * Log(OverallQual) + .28415 * Log(GrLivArea) + .25908 * Log(FirstFlrSf) + .00000322 * LotArea + .05937 * FullBath.

All three models generated an adjusted r-square of over 50%, however, after using the stepwise selection, we were able to increase it to 78%. As a result, we feel this is the best fitting model based on our analysis.

Appendix:

SAS codes and Outputs for Analysis of Question 1

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

```
proc print data= test;
run;

proc print data= train;
run;
```

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

- Filter dataset with only 3 neighborhoods NAmes, BrkSide, and Edwards.

```
/*Question 1*/
/*Filter our dataset and Log Transform*/
data train2;
set train;
where Neighborhood contains "Edwards"
    or Neighborhood contains"NAmes"
    or Neighborhood contains "BrkSide";
run;

data train2;
set train2;
lPrice = log(SalePrice);
lLivArea = log(GrLivArea);
run;

proc print data= train2;
run;
```

Output (383 observations and 83 variables).

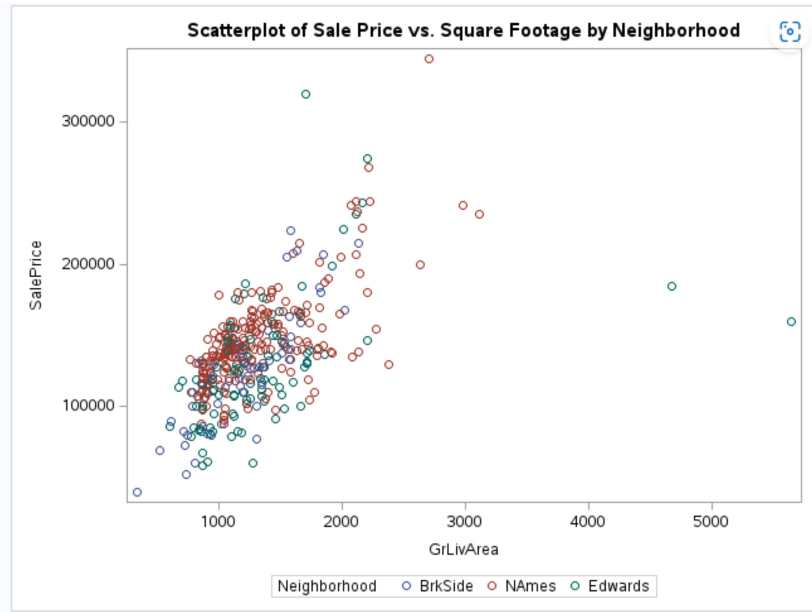
361	1385	50	RL	60	9060	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	Edwards	Norm	Norm	1Fam	1.5F
362	1390	50	RM	60	6000	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	BrkSide	Norm	Norm	1Fam	1.5F
363	1392	90	RL	65	8944	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	NAmes	Norm	Norm	Duplex	1Sto
364	1393	85	RL	68	7838	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	NAmes	Norm	Norm	1Fam	SFo
365	1398	70	RM	51	6120	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	BrkSide	Norm	Norm	1Fam	2Sto
366	1399	50	RL	60	7200	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	NAmes	Norm	Norm	1Fam	1.5F
367	1401	50	RM	50	6000	Pave	NA	Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Norm	Norm	1Fam	1.5F
368	1412	50	RL	80	9600	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	NAmes	Norm	Norm	1Fam	1.5F
369	1413	90	RL	60	7200	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	NAmes	Norm	Norm	Duplex	1Sto
370	1415	50	RL	64	13053	Pave	Pa	Reg	Bnk	AllPub	Inside	Gtl	BrkSide	Norm	Norm	1Fam	1.5F
371	1419	20	RL	71	9204	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	NAmes	Norm	Norm	1Fam	1Sto
372	1424	80	RL	NA	19690	Pave	NA	IR1	Lvl	AllPub	CulDSac	Gtl	Edwards	Norm	Norm	1Fam	SLv
373	1425	20	RL	NA	9503	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	NAmes	Norm	Norm	1Fam	1Sto
374	1426	20	RL	80	10721	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	NAmes	Norm	Norm	1Fam	1Sto
375	1428	50	RL	60	10930	Pave	Gr	Reg	Bnk	AllPub	Inside	Gtl	NAmes	Artery	Norm	1Fam	1.5F
376	1436	20	RL	80	8400	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	NAmes	Norm	Norm	1Fam	1Sto
377	1437	20	RL	60	9000	Pave	NA	Reg	Lvl	AllPub	FR2	Gtl	NAmes	Norm	Norm	1Fam	1Sto
378	1444	30	RL	NA	8854	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	BrkSide	Norm	Norm	1Fam	1.5U
379	1449	50	RL	70	11767	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	Edwards	Norm	Norm	1Fam	2Sto
380	1451	90	RL	60	9000	Pave	NA	Reg	Lvl	AllPub	FR2	Gtl	NAmes	Norm	Norm	Duplex	2Sto
381	1453	180	RM	35	3675	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	Edwards	Norm	Norm	TwnhsE	SLv
382	1459	20	RL	68	9717	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	NAmes	Norm	Norm	1Fam	1Sto
383	1460	20	RL	75	9937	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	Edwards	Norm	Norm	1Fam	1Sto

- Plot the data.

```
run;

/*Plot with Outliers*/
proc sgplot data=train2;
 scatter x=GrLivArea y=SalePrice / group=Neighborhood;
 title 'Scatterplot of Sale Price vs. Square Footage by Neighborhood';
run;
```

(Figure 1.1)



- Build first model:

```
/* Build Model 1 with outliers*/
proc reg data= train2;
model SalePrice = GrLIvArea / vif clb cli clm;
run;
```

(Table 1.2)

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

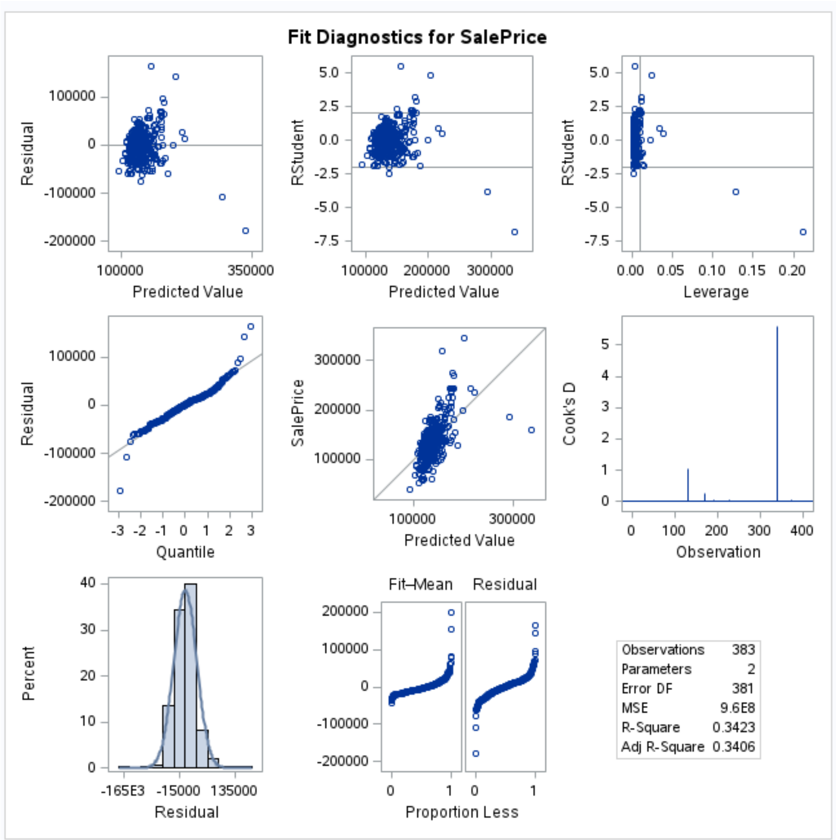
Number of Observations Read	383
Number of Observations Used	383

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			

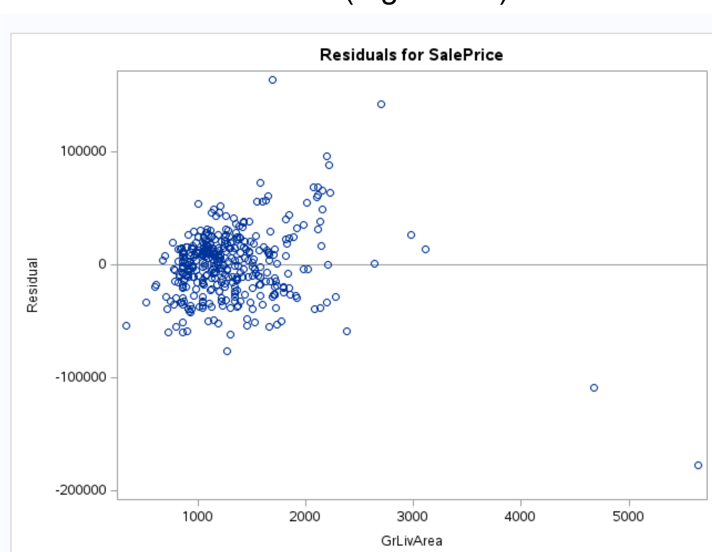
Root MSE	30985	R-Square	0.3423
Dependent Mean	138063	Adj R-Sq	0.3406
Coeff Var	22.44267		

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

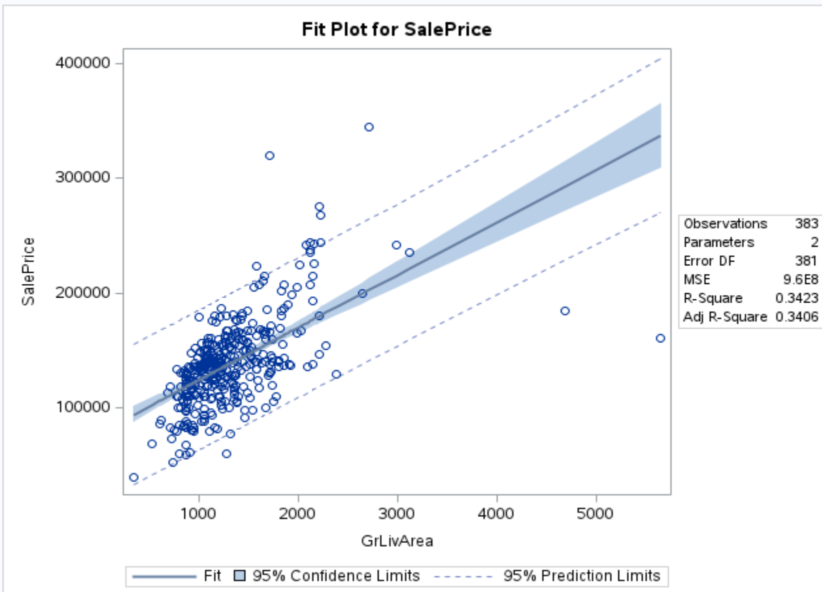
(Figure 1.3)



(Figure 1.4)



(Figure 1.5)



- Remove the 4 outliers.

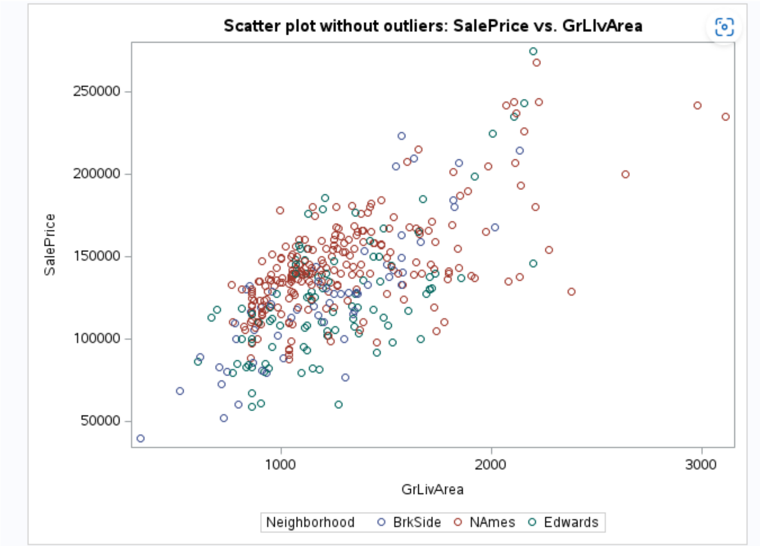
```
/* Remove Outliers */
data trainNoOutliers;
set train2;
where Id ~= 524 and Id ~= 643 and Id~= 725 and Id~= 1299 and Id~= 1299;
run;

proc print data= trainNoOutliers;
run;

/*Plot without Outliers*/
title 'Scatter plot without outlieers: SalePrice vs. GrLlvArea';
proc sgplot data=trainNoOutliers;
 scatter x=GrLivArea y=SalePrice / group=Neighborhood;
run;
```

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

(Figure 1.6)



- Build the second model without outliers.

```
/* Run Model Without Outliers */
Proc reg data= trainNoOutliers;
model SalePrice = GrLivArea/ vif clb cli clm;
run;
```

(Figure 1.7)

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Number of Observations Read	379
Number of Observations Used	379

Analysis of Variance

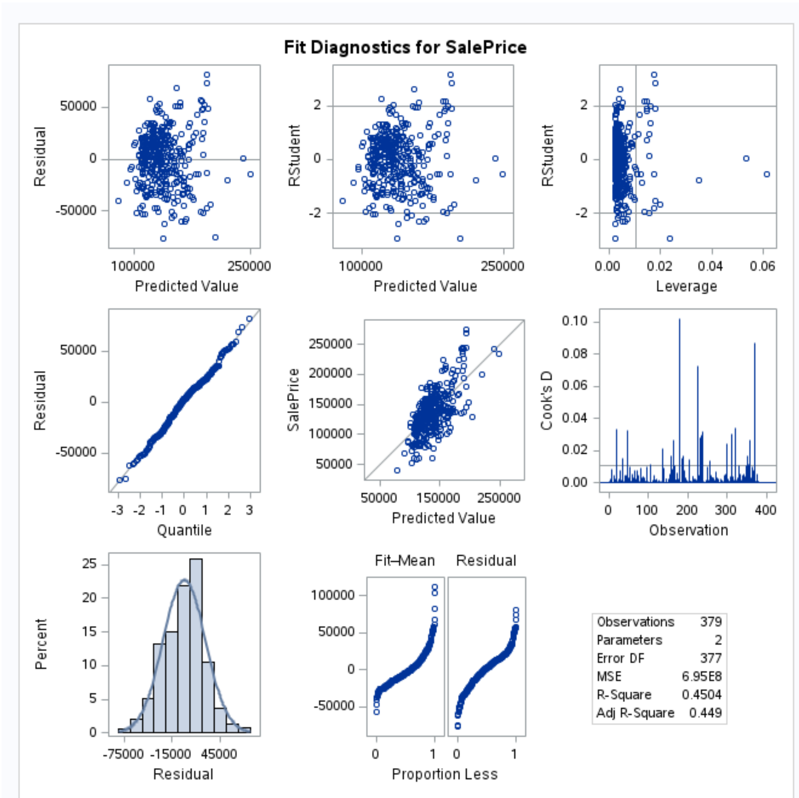
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2.148668E11	2.148668E11	309.00	<.0001
Error	377	2.621477E11	695351977		
Corrected Total	378	4.770145E11			

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

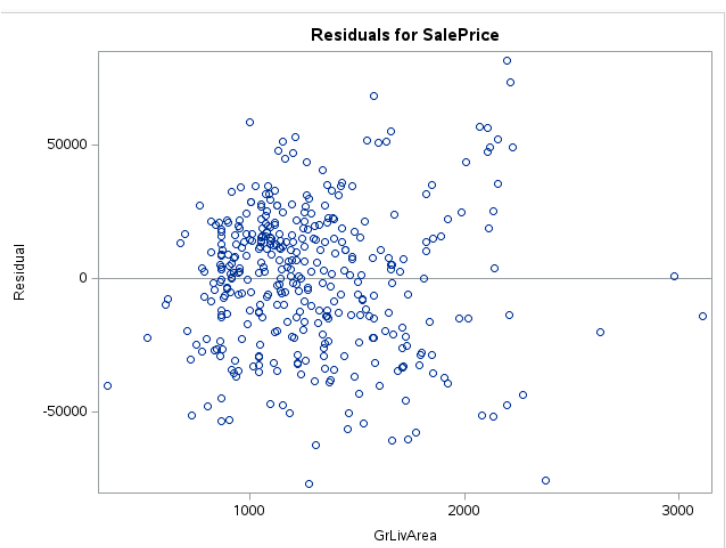
Parameter Estimates

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

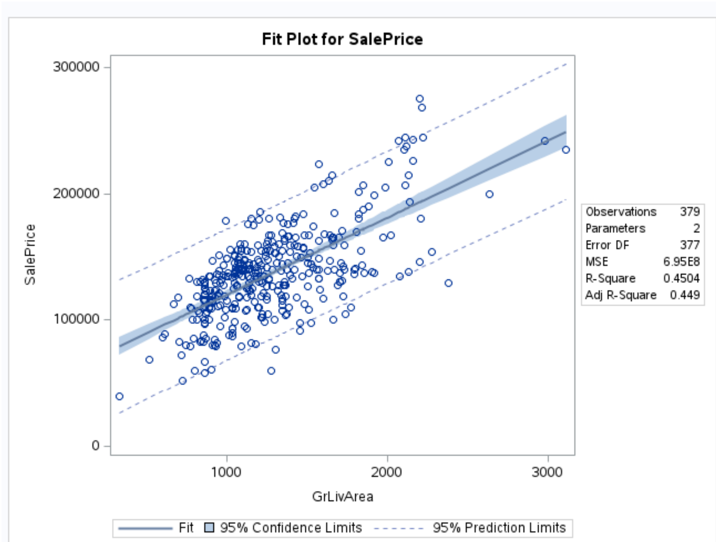
(Figure 1.8)



(Figure 1.9)



(Figure 2.0)



- Build the third model without outliers.

```
/*Develop a third model without outliers*/
proc glm data= trainNoOutlier plots = all;
class neighborhood (REF = "BrkSide");
model SalePrice = GrLIvArea|Neighborhood / solution clparm cli;
run;
```

(Figure 2.1)

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			

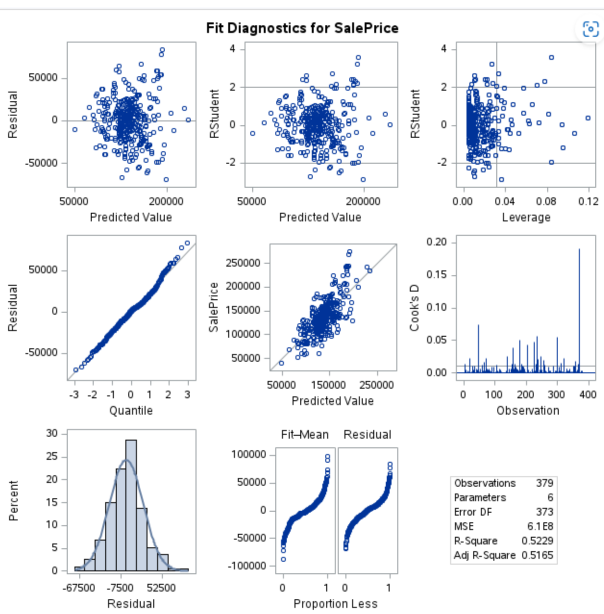
R-Square	Coeff Var	Root MSE	SalePrice Mean
0.522897	18.04909	24701.16	136855.4

Source	DF	Type I SS	Mean Square	F Value	Pr > F
GrLivArea	1	214866815955	214866815955	352.16	<.0001
Neighborhood	2	23080953286	11540476643	18.91	<.0001
GrLivArea*Neighborho	2	11481870833	5740935416.3	9.41	0.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
GrLivArea	1	187984048744	187984048744	308.10	<.0001
Neighborhood	2	20322662167	10161331083	16.65	<.0001
GrLivArea*Neighborho	2	11481870833	5740935416.3	9.41	0.0001

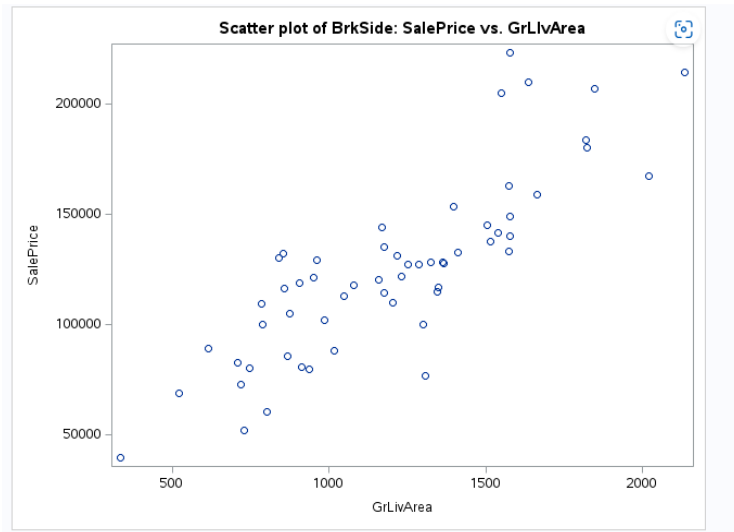
Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	19971.51379	B	10685.20222	1.87	0.0624	-1039.27266	40982.30025
GrLivArea	87.16253	B	8.46257	10.30	<.0001	70.52222	103.80285
Neighborhood Edwards	17128.90777	B	14154.89029	1.21	0.2270	-10704.48003	44962.29557
Neighborhood NAmes	60354.19850	B	12060.03479	5.00	<.0001	36640.01788	84068.37913
Neighborhood BrkSide	0.00000	B	.	.	.	.	.
GrLivArea*Neighborho Edwards	-17.00416	B	11.05135	-1.54	0.1247	-38.73493	4.72660
GrLivArea*Neighborho NAmes	-37.60128	B	9.40218	-4.00	<.0001	-56.08921	-19.11336
GrLivArea*Neighborho BrkSide	0.00000	B	.	.	.	.	.

(Figure 2.2)



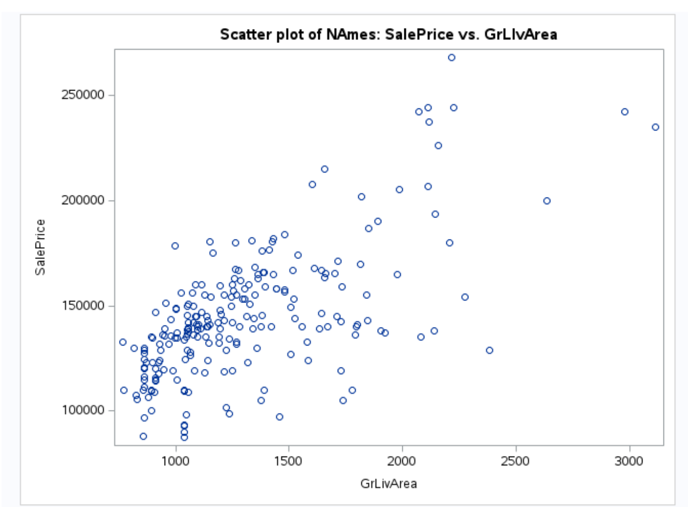
```
/* Model 3 */
/*Plot third model without Outliers*/
title 'Scatter plot of BrkSide: SalePrice vs. GrLlvArea';
proc sgplot data=trainNoOutliers;
where neighborhood= 'BrkSide';
 scatter x=GrLivArea y=SalePrice;
run;
```

(Figure 2.3)



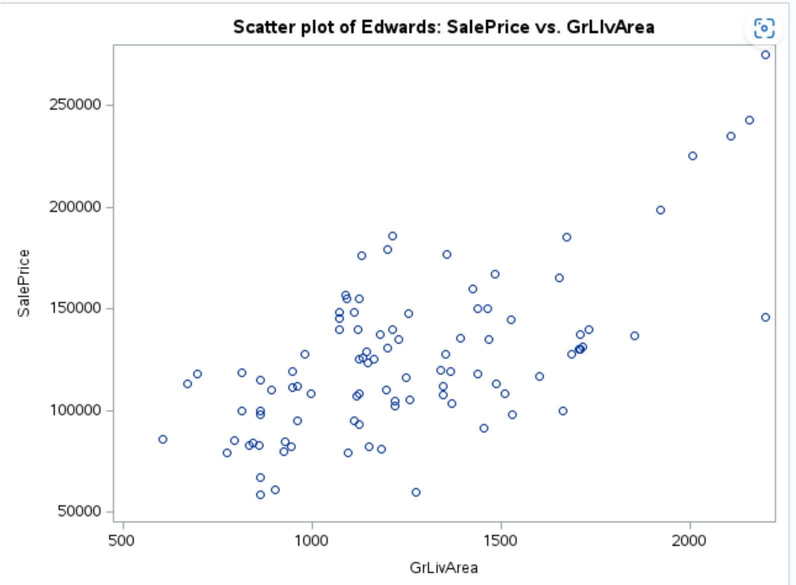

```
/*Plot third model without Outliers*/
title 'Scatter plot of NAmes: SalePrice vs. GrLlvArea';
proc sgplot data=trainNoOutliers;
where neighborhood= 'NAmes';
 scatter x=GrLivArea y=SalePrice;
run;
```

(Figure 2.4)



```
/*Plot third model without Outliers*/
title 'Scatter plot of Edwards: SalePrice vs. GrLlvArea';
proc sgplot data=trainNoOutliers;
where neighborhood= 'Edwards';
 scatter x=GrLivArea y=SalePrice;
run;
```

(Figure 2.5)



(Data 2.6)

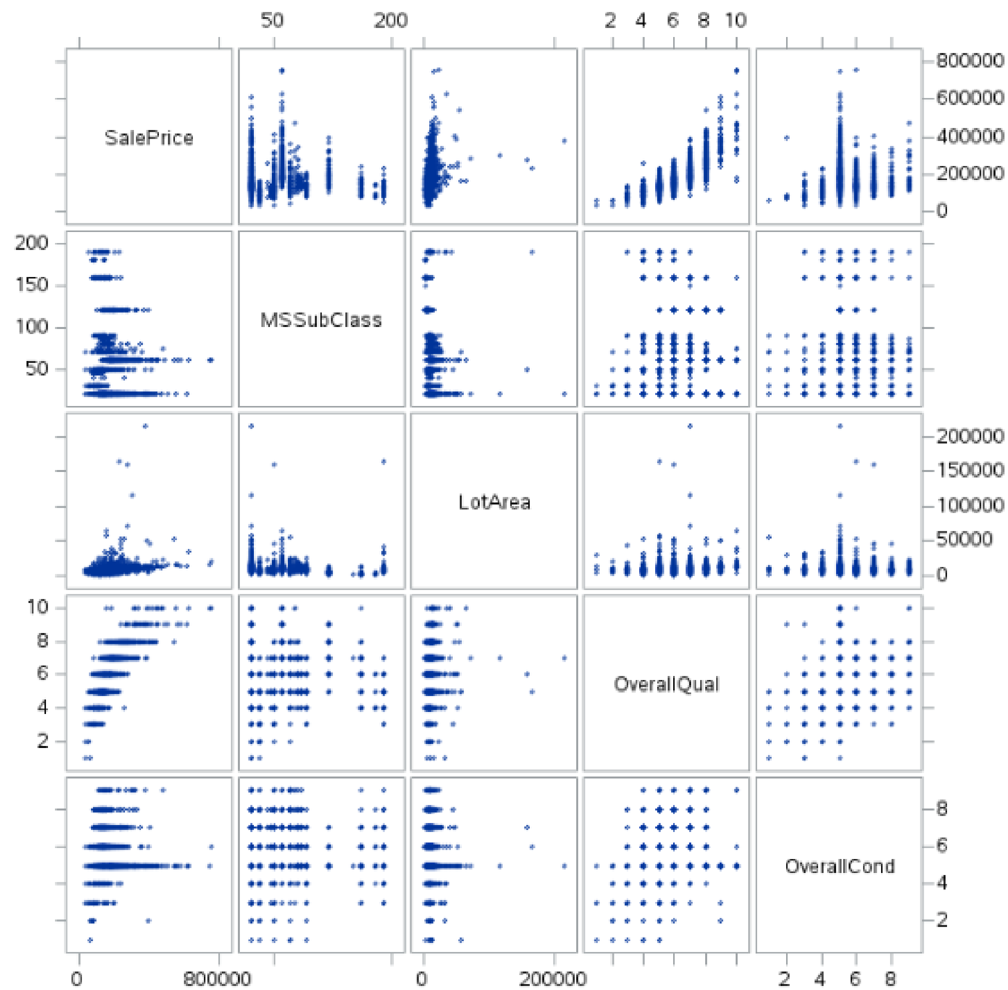
- Regression model for NAmes neighborhood:
    - μ(SalePrice|NAmes) = 80325.71 +49.56*GrLivArea
    - = 49.56*(100)=$4,956
- Regression model for BrkSide neighborhood:
    - μ(SalePrice|NAmes) =19971.51 +87.16*GrLivArea
    - =87.16*100= $8,716
- Regression model forEdwards neighborhood:
    - μ(SalePrice|NAmes = 37100.42+70.16*GrLivArea
    - =70.16*100=$7,016
- `> qt(.975, 373)`
- `[1] 1.966344`

25138.77 +/- 1.966 * 14113.06= 27726.28

$$\text{Coefficient} \pm (t_{\alpha/2} \times \text{SE})$$

SAS Code and Analysis for Question 2

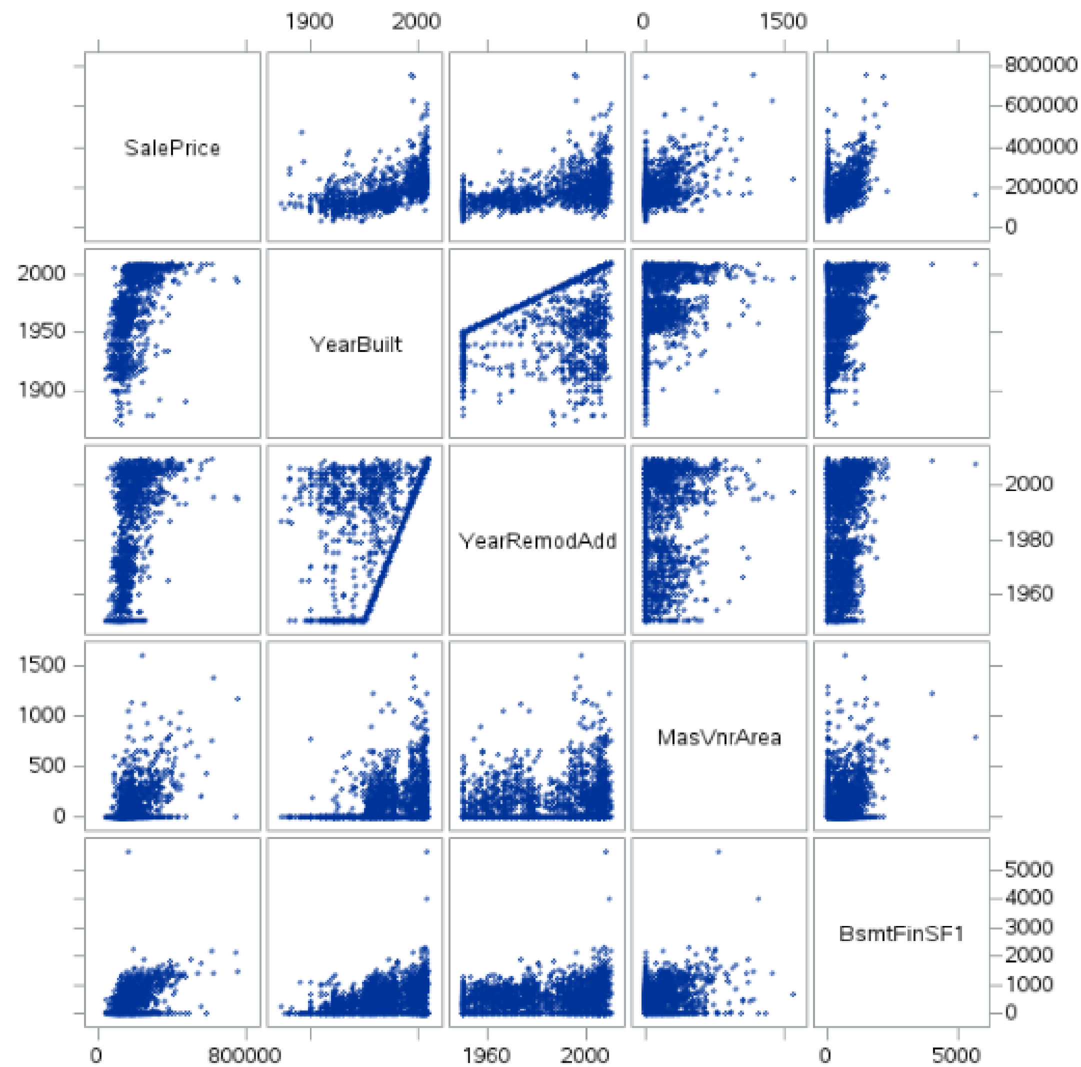
1. Simple Linear Regression
    a. Exploratory Data Analysis: building scatter plots to identify correlations
       Figure 3.01
        i.
```
/* Scatter Plot Matrix */
proc sgscatter data=train2;
    matrix SalePrice MSSubClass LotArea OverallQual OverallCond;
run;
```
       Figure 3.02



        ii.

       Figure 3.03
        iii.
```
proc sgscatter data=train2;
    matrix SalePrice YearBuilt YearRemodAdd MasVnrArea BsmtFinSF1;
run;
```

Figure 3.04



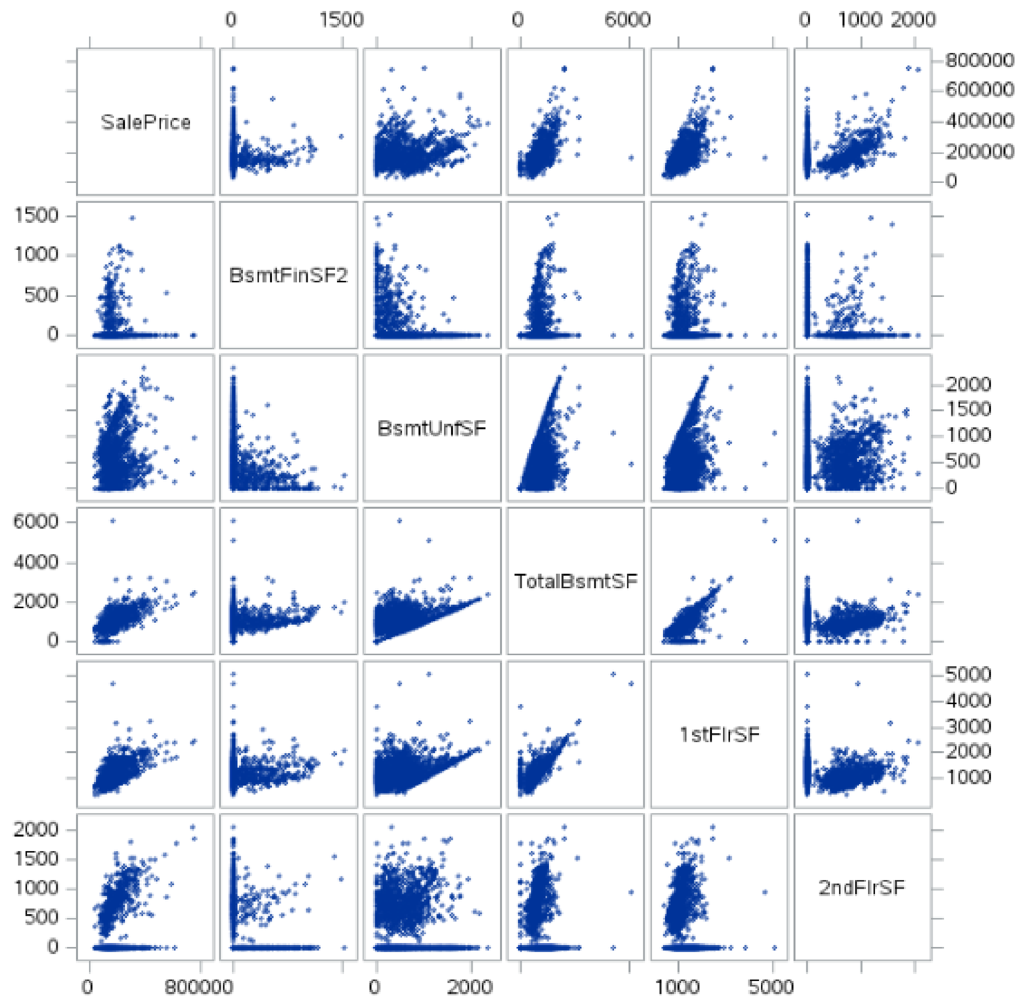
iv.

Figure 3.05

```
proc sgscatter data=train2;
    matrix SalePrice BsmtFinSF2 BsmtUnfSf TotalBsmtSF '1stFlrSF'n '2ndFlrSF'n;
run;
```

v.

Figure 3.06



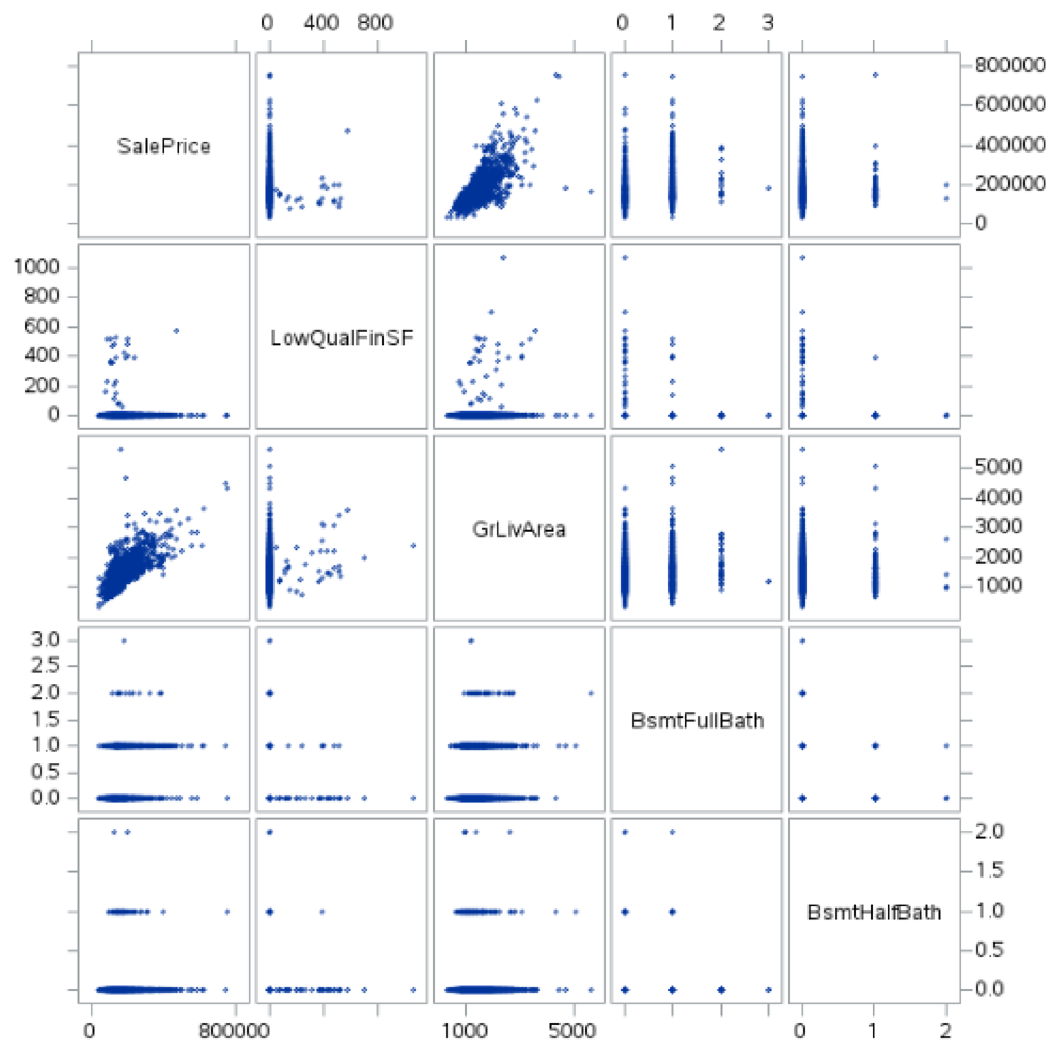
vi.

Figure 3.07

vii.

```
proc sgscatter data=train2;
    matrix SalePrice LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath ;
run;
```

Figure 3.08



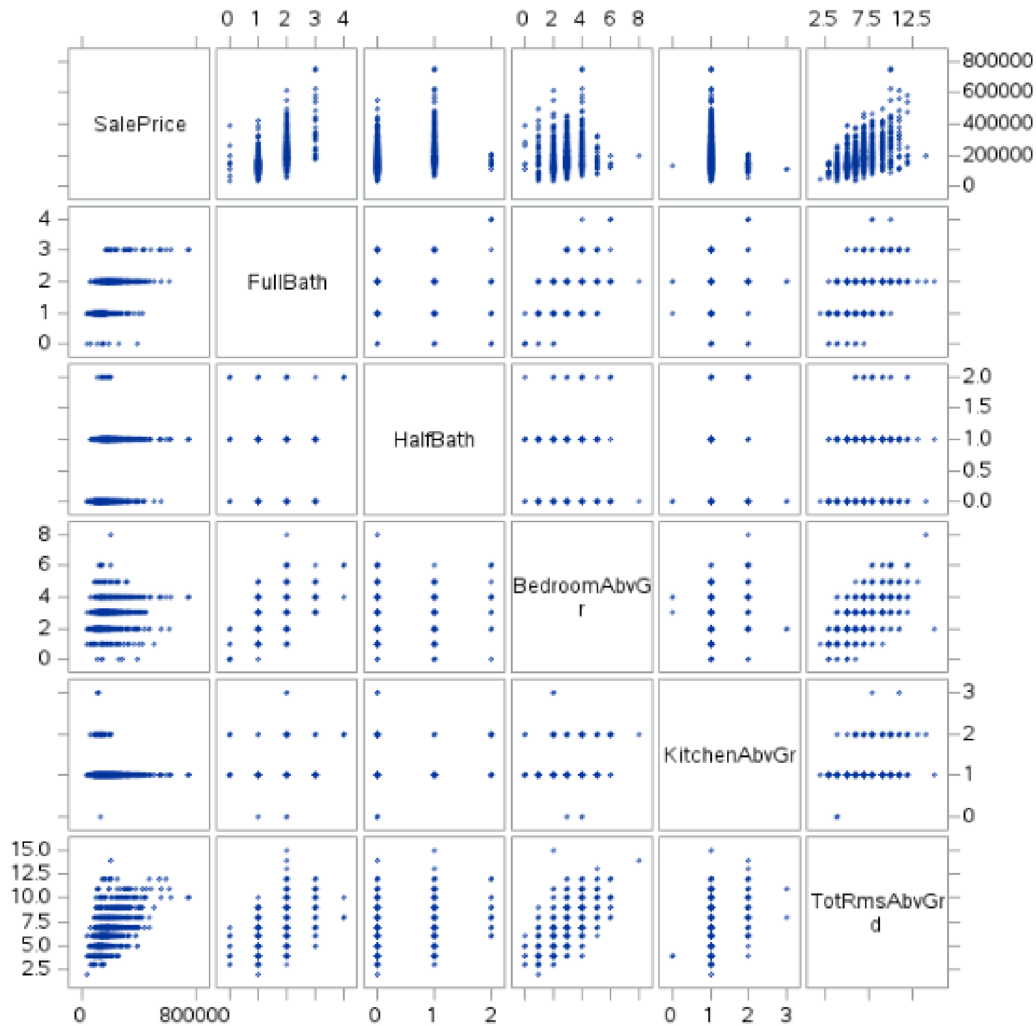
viii.

Figure 3.09

```
proc sgscatter data=train2;
    matrix SalePrice FullBath HalfBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd;
run;
```

ix.

Figure 3.1



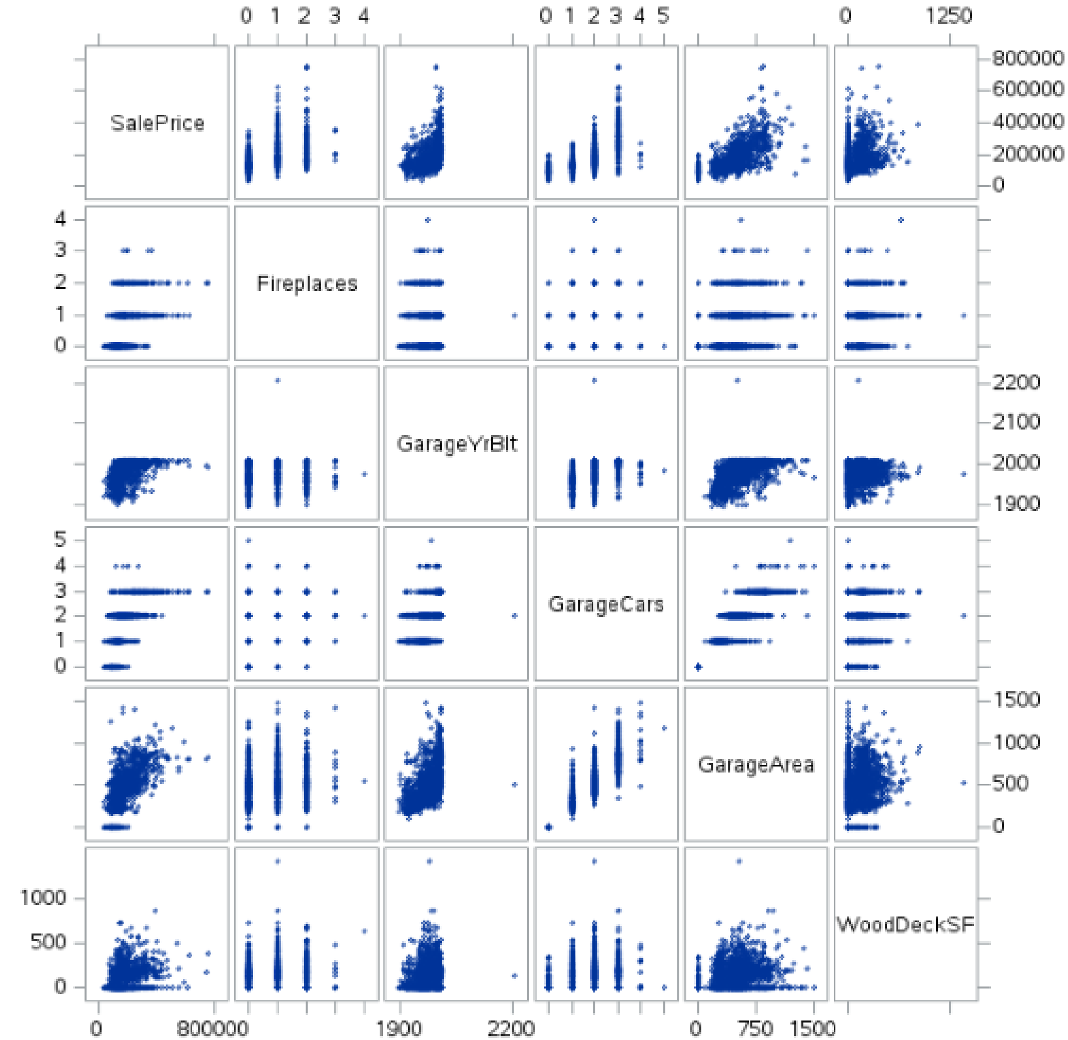
x.

Figure 3.11

```
proc sgscatter data=train2;
    matrix SalePrice Fireplaces GarageYrBlt GarageCars GarageArea WoodDeckSF;
run;
```

xi.

Figure 3.12



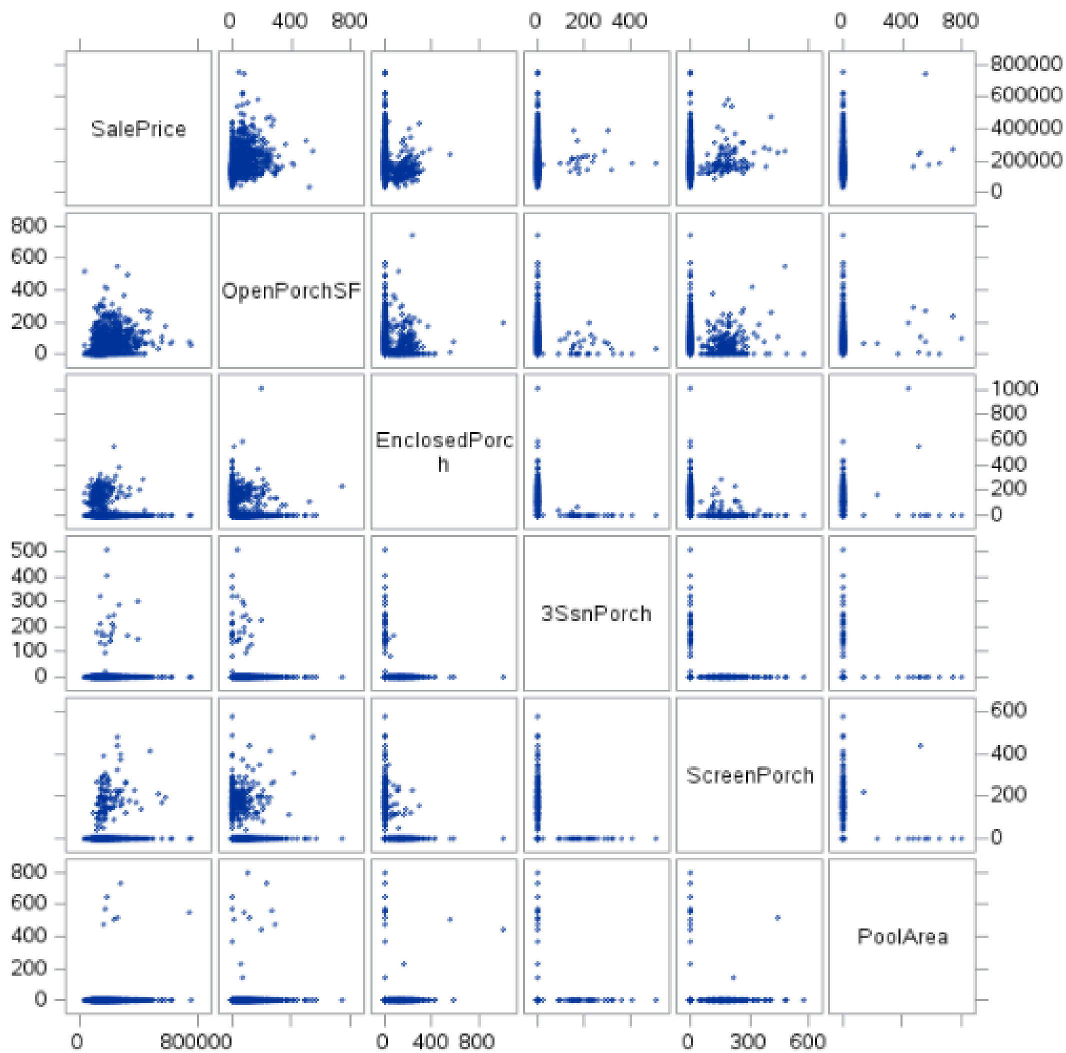
xii.

Figure 3.13

```
proc sgscatter data=train2;
    matrix SalePrice OpenPorchSF EnclosedPorch '3SsnPorch'n ScreenPorch PoolArea;
run;
```

xiii.

Figure 3.14



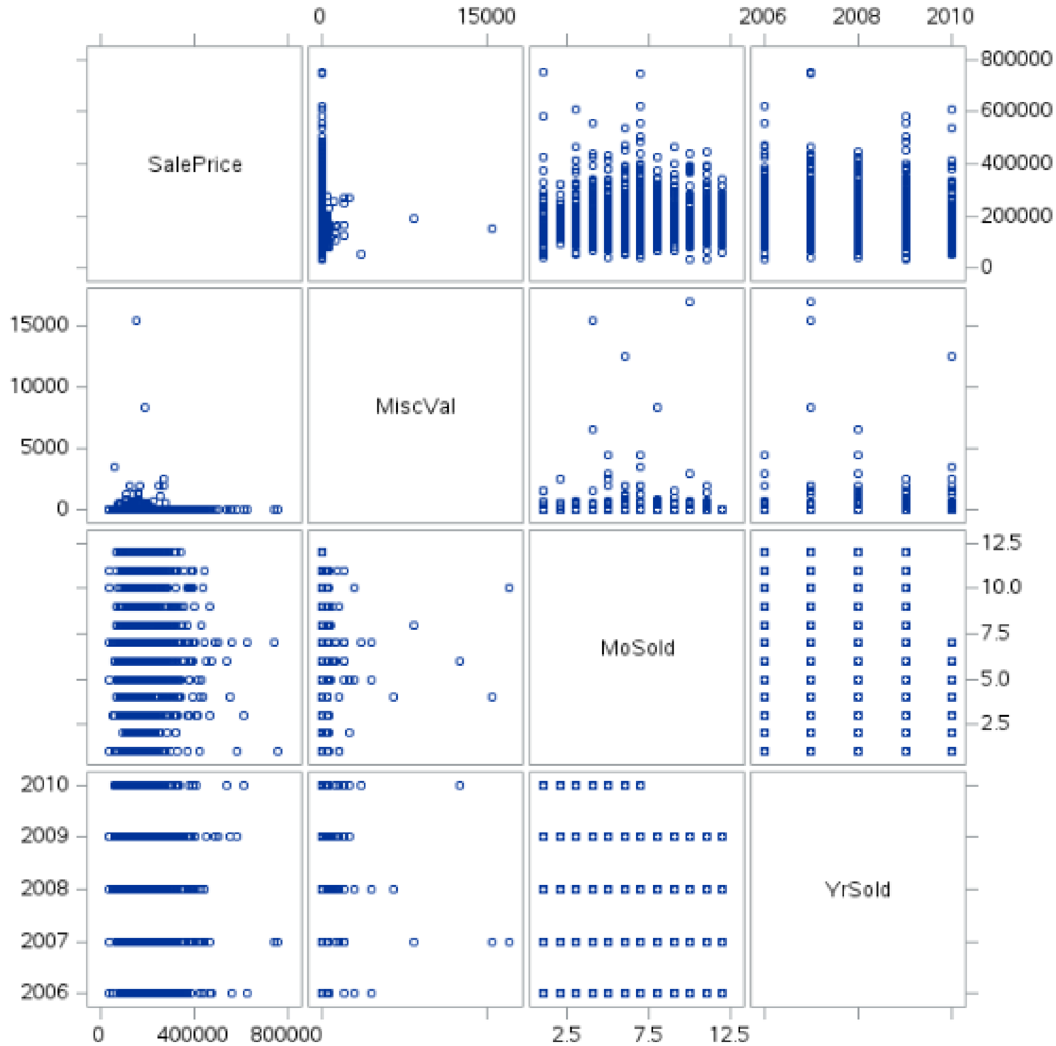
xiv.

Figure 3.15

```
proc sgscatter data=train2;
    matrix SalePrice MiscVal MoSold YrSold;
run;
```

xv.

Figure 3.16



xvi.

b. Log-transforming explanatory variables that show potential correlation

Figure 3.17

```
/* Logging the most promising explanatory variables */
data train2;
set train2;
OverallQual_log = log(OverallQual);
OverallCond_log = log(OverallCond);
TotalBsmtSF_log = log(TotalBsmtSF);
FirstFlrSf_log = log('1stFlrSf'n);
SecondFlrSF_log = log('2ndFlrSf'n);
GrLivArea_log = log(GrLivArea);
FullBath_log = log(FullBath);
TotRmsAbvGrd_log = log(TotRmsAbvGrd);
GarageArea_log = log(GarageArea);
SalePrice_log = log(SalePrice);
run;
```

i.

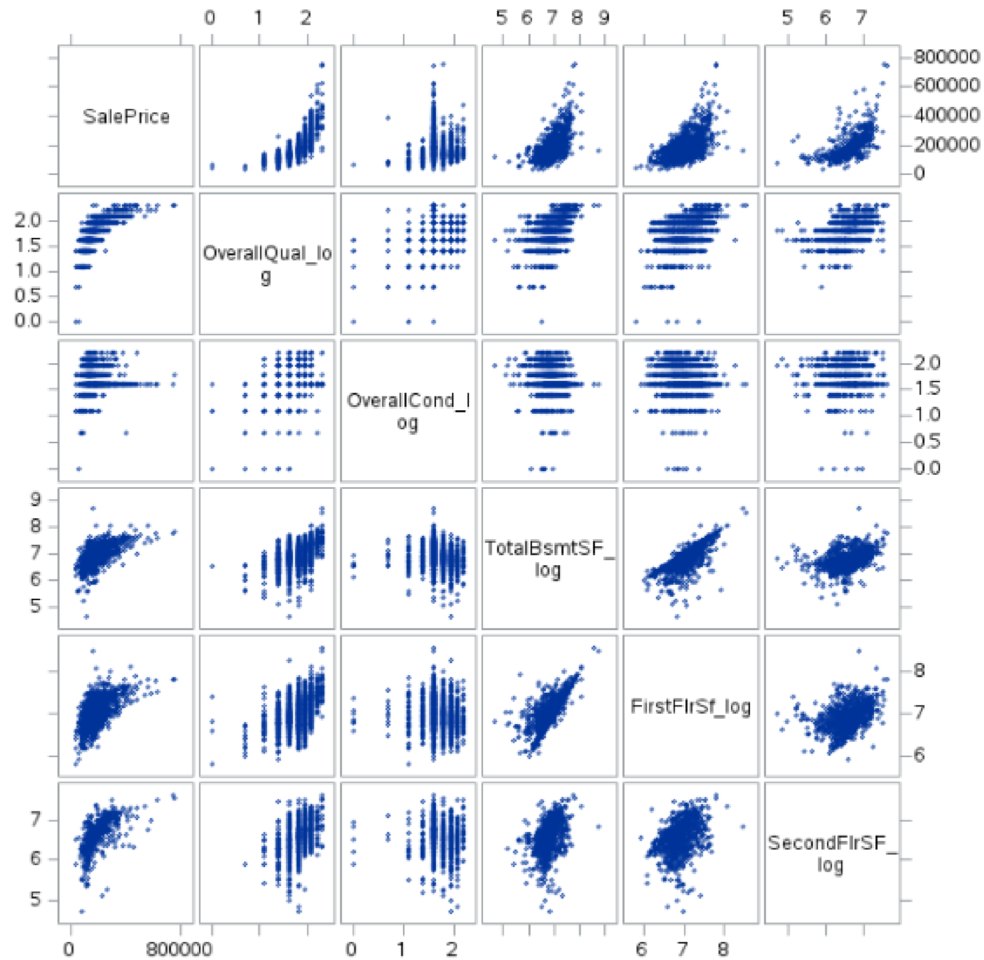
c. Regenerating scatter plots of log-transformations

Figure 3.18

i.

```
proc sgscatter data=train2;
    matrix SalePrice OverallQual_log OverallCond_log TotalBsmtSF_log FirstFlrSf_log SecondFlrSF_log;
run;
```

Figure 3.19



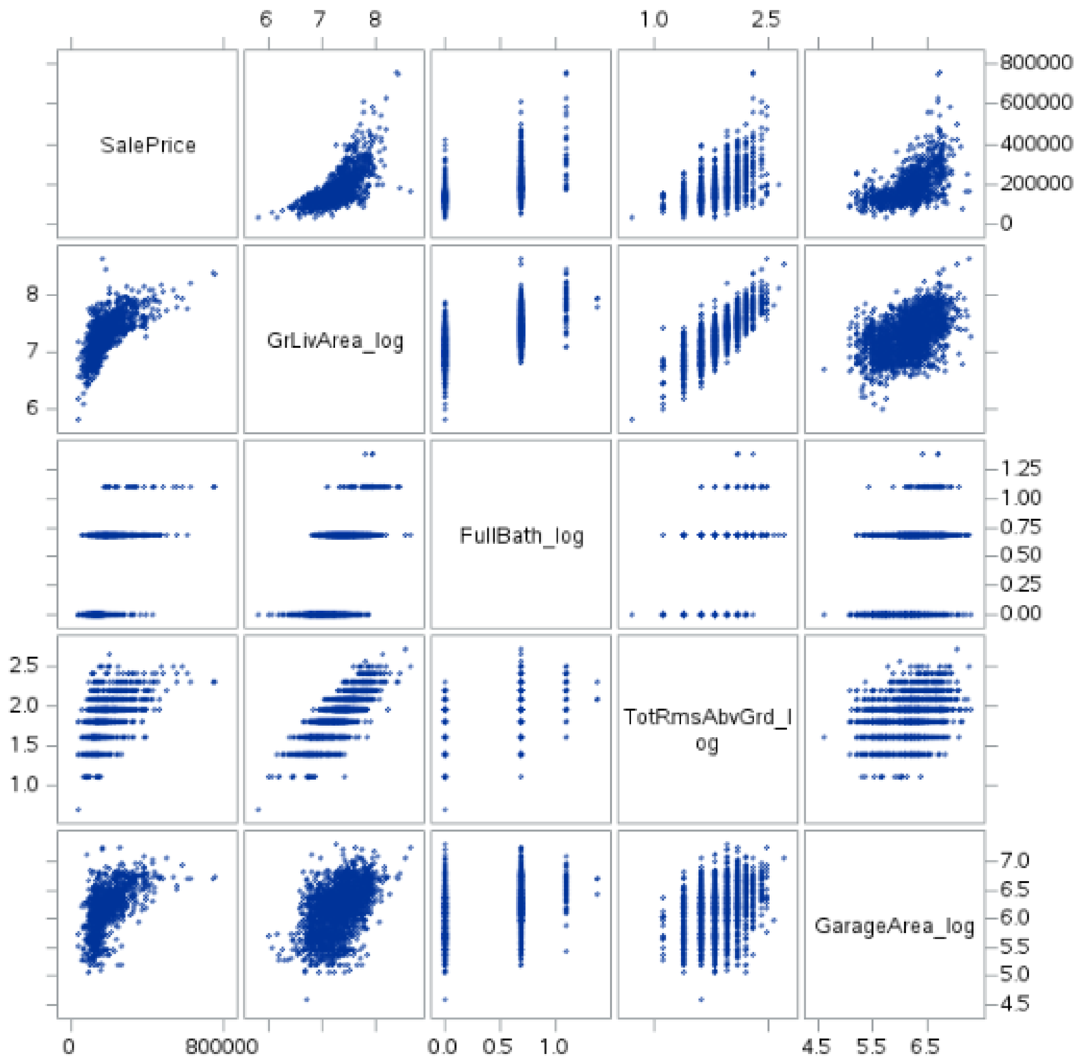
ii.

Figure 3.2

iii.

```
proc sgscatter data=train2;
    matrix SalePrice GrLivArea_log FullBath_log TotRmsAbvGrd_log GarageArea_log;
run;
```

Figure 3.21



iv.

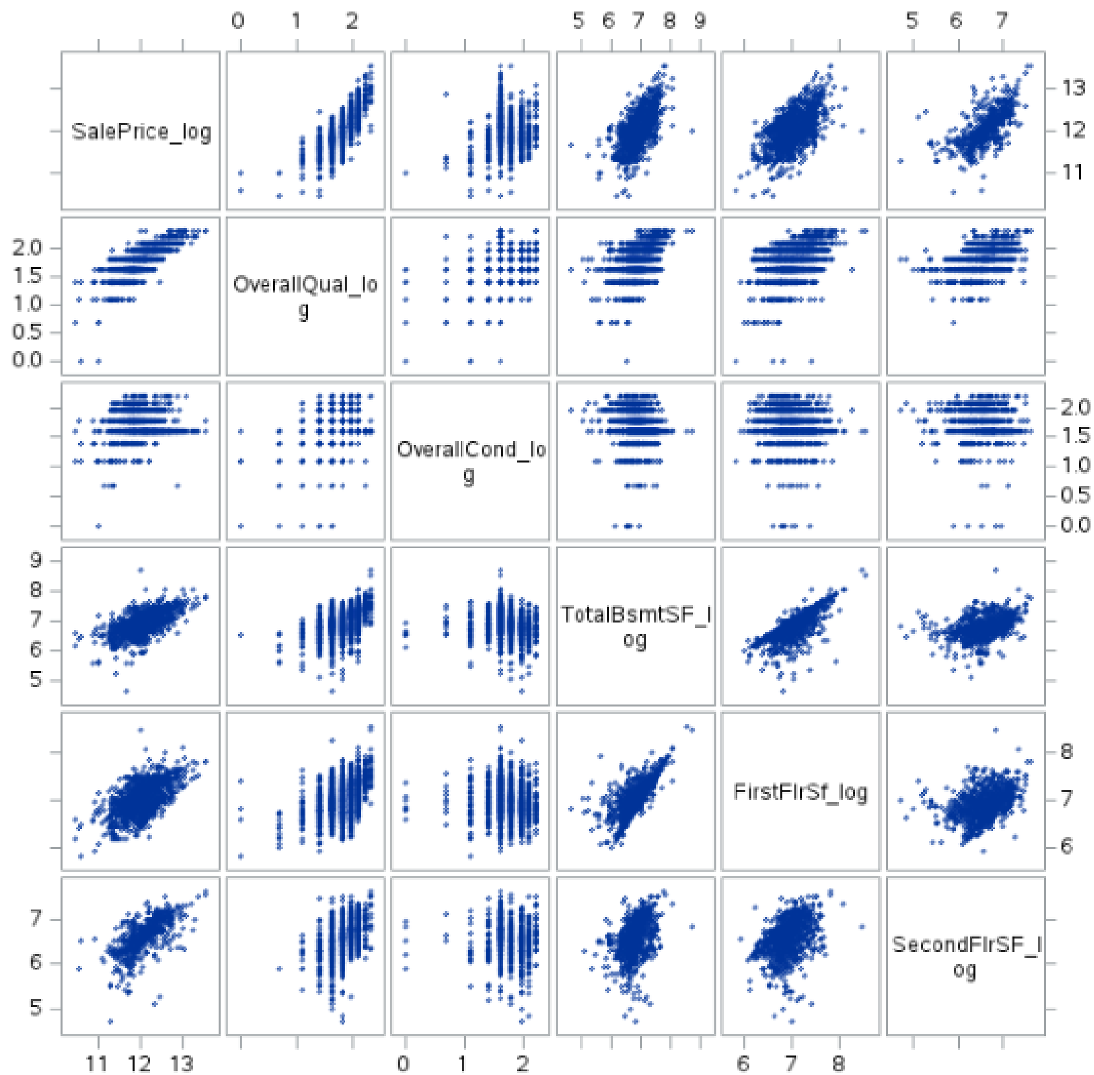
d. Regenerating scatter plots of log transformations with log-transformed SalePrice because the previous plots still looked curvilinear

Figure 3.22

i.

```
proc sgscatter data=train2;
    matrix SalePrice_log OverallQual_log OverallCond_log TotalBsmtSF_log FirstFlrSf_log SecondFlrSF_log;
run;
```

Figure 3.23



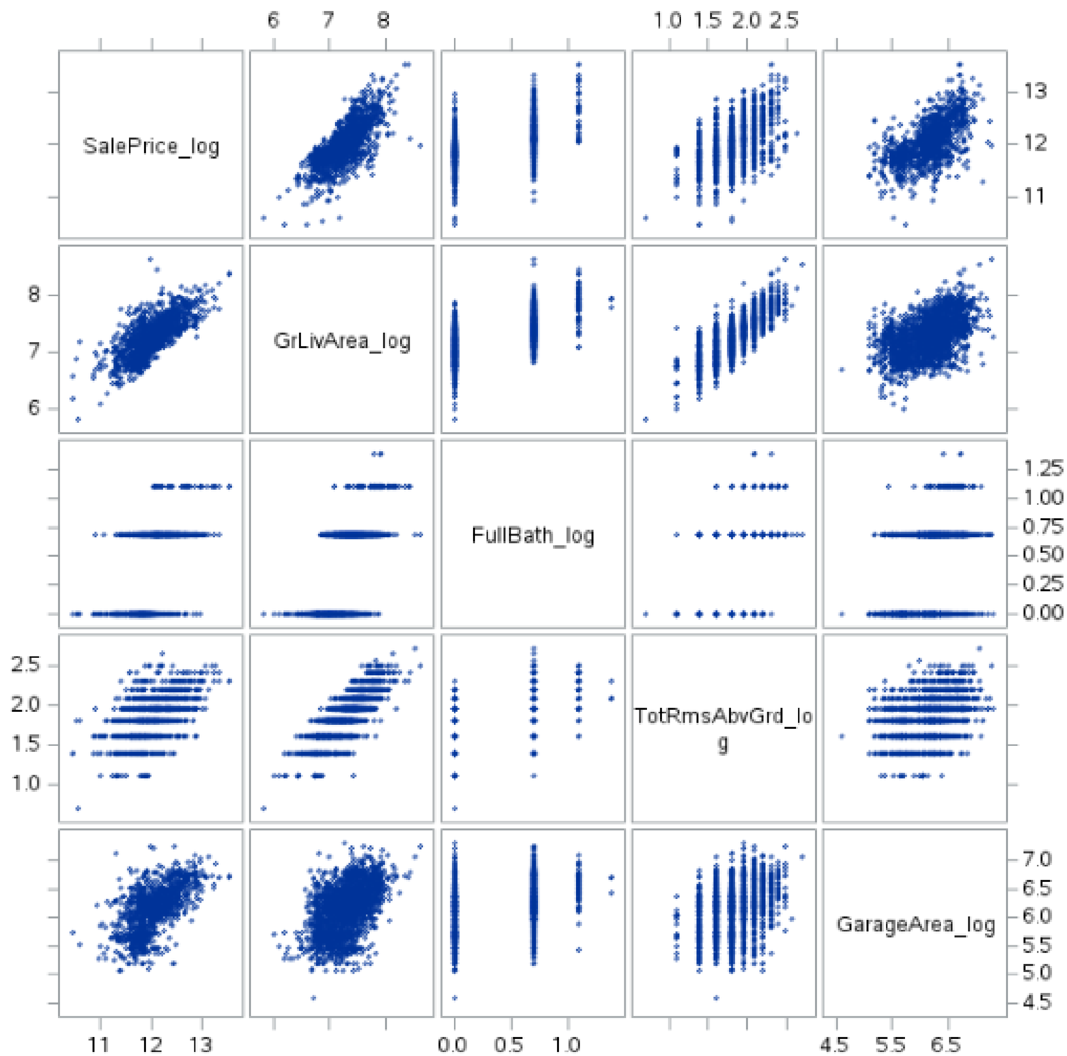
ii.

Figure 3.24

```
proc sgscatter data=train2;
    matrix SalePrice_log GrLivArea_log FullBath_log TotRmsAbvGrd_log GarageArea_log;
run;
```

iii.

Figure 3.25



iv.

e. Building Simple Linear Regression Models of the top explanatory variables

Figure 3.26

i.

```
proc glm data = train2 plots = all;
    model SalePrice_log = GrLivArea_Log / cli solution;
run;
```

Figure 3.27

ii.

The GLM Procedure

Dependent Variable: SalePrice_log

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	124.1461074	124.1461074	1665.88	<.0001
Error	1458	108.6545516	0.0745230		
Corrected Total	1459	232.8006590			

R-Square	Coeff Var	Root MSE	SalePrice_log Mean
0.533272	2.270358	0.272989	12.02405

Source	DF	Type I SS	Mean Square	F Value	Pr > F
GrLivArea_log	1	124.1461074	124.1461074	1665.88	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
GrLivArea_log	1	124.1461074	124.1461074	1665.88	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	5.668124686	0.15588850	36.36	<.0001
GrLivArea_log	0.874535433	0.02142674	40.82	<.0001

Figure 3.28

iii.

Sum of Residuals	-0.0000000
Sum of Squared Residuals	108.6545516
Sum of Squared Residuals - Error SS	0.0000000
PRESS Statistic	109.0202511
First Order Autocorrelation	-0.0107726
Durbin-Watson D	2.0215006

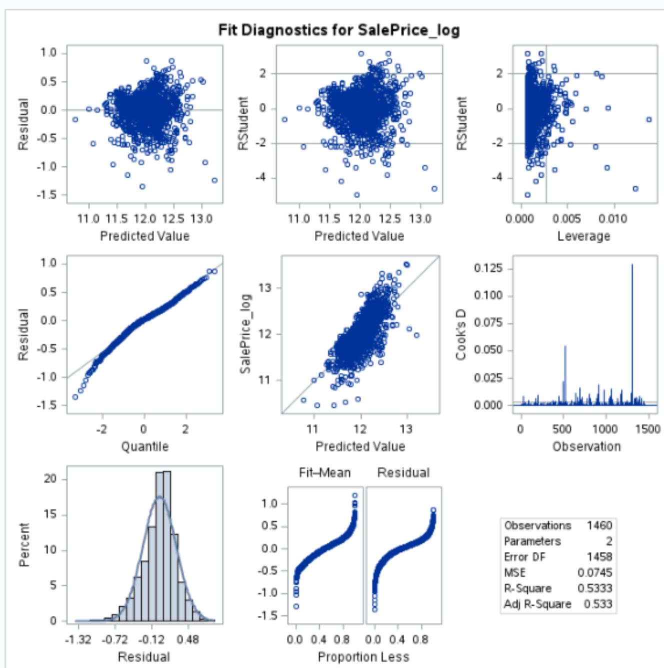


Figure 3.29

iv.

```
proc glm data = train2 plots = all;
    model SalePrice_log = FirstFlrSF_log / cli solution;
run;
```

v. Figure 3.3

The GLM Procedure

Dependent Variable: SalePrice_log

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	86.3262158	86.3262158	859.29	<.0001
Error	1458	146.4744432	0.1004626		
Corrected Total	1459	232.8006590			

R-Square	Coeff Var	Root MSE	SalePrice_log Mean
0.370816	2.636036	0.316958	12.02405

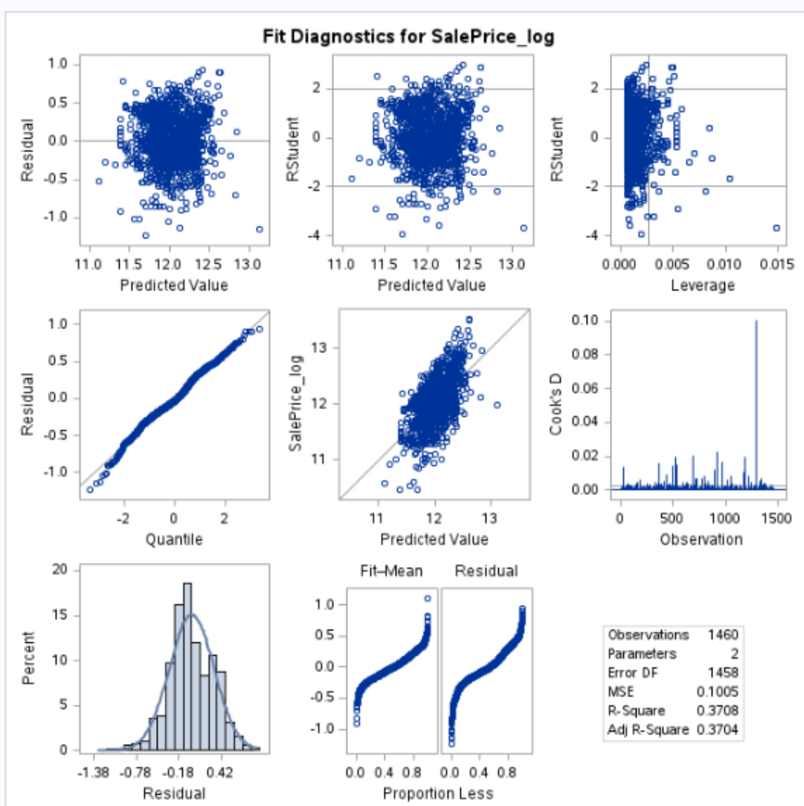
Source	DF	Type I SS	Mean Square	F Value	Pr > F
FirstFlrSf_log	1	86.32621580	86.32621580	859.29	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
FirstFlrSf_log	1	86.32621580	86.32621580	859.29	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	6.659312840	0.18319980	36.35	<.0001
FirstFlrSf_log	0.765570739	0.02611657	29.31	<.0001

vi.

Figure 3.31



vii.

Figure 3.32

```
proc glm data = train2 plots = all;
    model SalePrice_log = TotalBsmtSF_log / cli solution;
run;
```

viii.

Figure 3.33

The GLM Procedure

Dependent Variable: SalePrice_log

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	80.6779079	80.6779079	819.69	<.0001
Error	1421	139.8619519	0.0984250		
Corrected Total	1422	220.5398598			

R-Square	Coeff Var	Root MSE	SalePrice_log Mean
0.365820	2.606381	0.313728	12.03691

Source	DF	Type I SS	Mean Square	F Value	Pr > F
TotalBsmtSF_log	1	80.67790790	80.67790790	819.69	<.0001

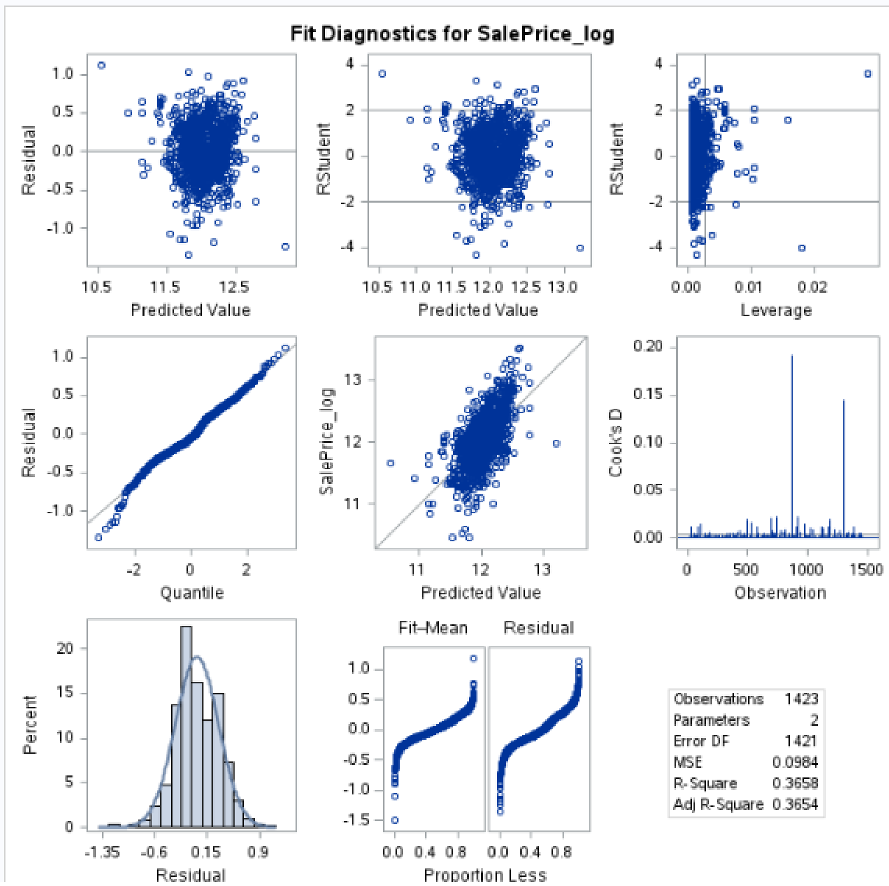
Source	DF	Type III SS	Mean Square	F Value	Pr > F
TotalBsmtSF_log	1	80.67790790	80.67790790	819.69	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	7.471996401	0.15966052	46.80	<.0001
TotalBsmtSF_log	0.659189633	0.02302427	28.63	<.0001

ix.

Figure 3.34

Sum of Residuals	0.0000000
Sum of Squared Residuals	139.8619519
Sum of Squared Residuals - Error SS	0.0000000
PRESS Statistic	140.4412453
First Order Autocorrelation	0.0167172
Durbin-Watson D	1.9652744



x.

Figure 3.35

```
proc glm data = train2 plots = all;
    model SalePrice_log = GarageArea_log / cli solution;
run;
```

xi.

Figure 3.36

The GLM Procedure

Dependent Variable: SalePrice_log

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	74.2247912	74.2247912	818.11	<.0001
Error	1377	124.9306596	0.0907267		
Corrected Total	1378	199.1554508			

R-Square	Coeff Var	Root MSE	SalePrice_log Mean
0.372698	2.498556	0.301209	12.05531

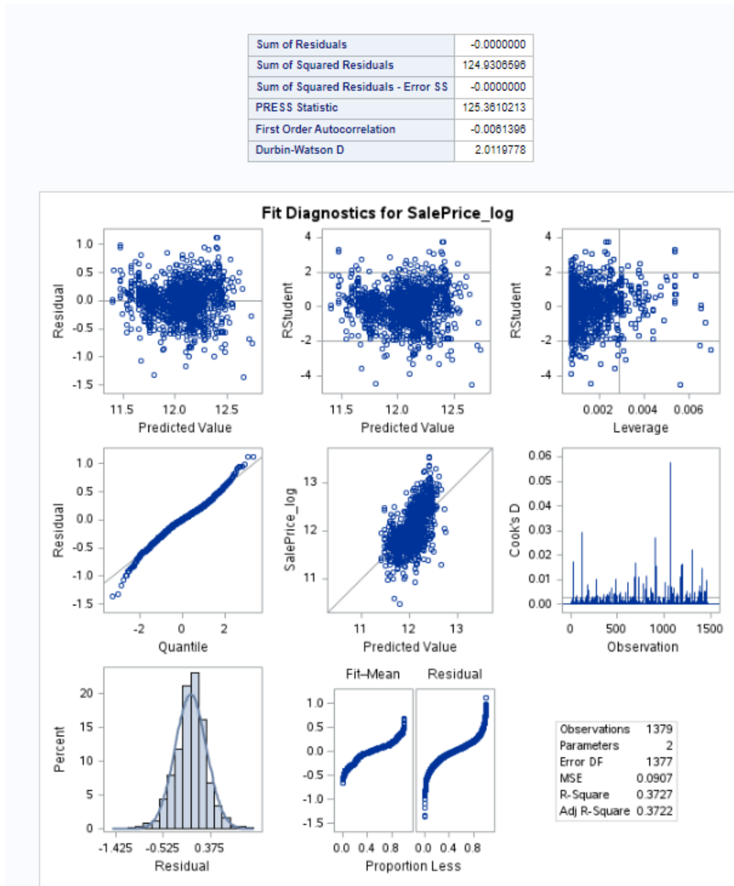
Source	DF	Type I SS	Mean Square	F Value	Pr > F
GarageArea_log	1	74.22479120	74.22479120	818.11	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
GarageArea_log	1	74.22479120	74.22479120	818.11	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	8.283775330	0.13210866	62.70	<.0001
GarageArea_log	0.613555305	0.02145096	28.60	<.0001

xii.

Figure 3.37



Sum of Residuals	-0.0000000
Sum of Squared Residuals	124.9306596
Sum of Squared Residuals - Error SS	-0.0000000
PRESS Statistic	125.3610213
First Order Autocorrelation	-0.0061396
Durbin-Watson D	2.0119778

xiii.

f. Removing Outliers from Selected Model : SalePrice_log~GrLivArea_log

Figure 3.38

i.

```
data train2Q1NoOutliers;
    set train2;
    where ID ~= 1299 and ID ~= 524 and ID ~= 31 and ID ~= 643
        and ID ~= 725 and ID ~= 913 and ID ~= 495 and ID ~= 1095
        and ID~= 494 and ID ~= 911 and ID ~= 1039 and ID ~= 798
        and ID ~= 536 and ID ~= 534 ;
run;

proc glm data = train2Q1NoOutliers plots = all;
    model SalePrice_log = GrLivArea_Log / cli solution;
run;
```

g. Regenerating model after removing outliers

Figure 3.39

i.

```
proc glmselect data=train2Q1NoOutliers;
  model SalePrice_log = GrLivArea_Log /  selection=Stepwise(stop=CV) cvmethod = random(5) stats = adjrsq;
run;
```

Figure 3.4

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 1).

Effects:	Intercept GrLivArea_log

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	1	122.92276	122.92276	1718.88
Error	1444	103.26533	0.07151	
Corrected Total	1445	226.18809		

Root MSE	0.26742
Dependent Mean	12.02715
R-Square	0.5435
Adj R-Sq	0.5431
AIC	-2384.36229
AICC	-2384.34565
SBC	-3801.80918
CV PRESS	103.41770

ii.

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	5.562069	0.156096	35.63
GrLivArea_log	1	0.889567	0.021456	41.46

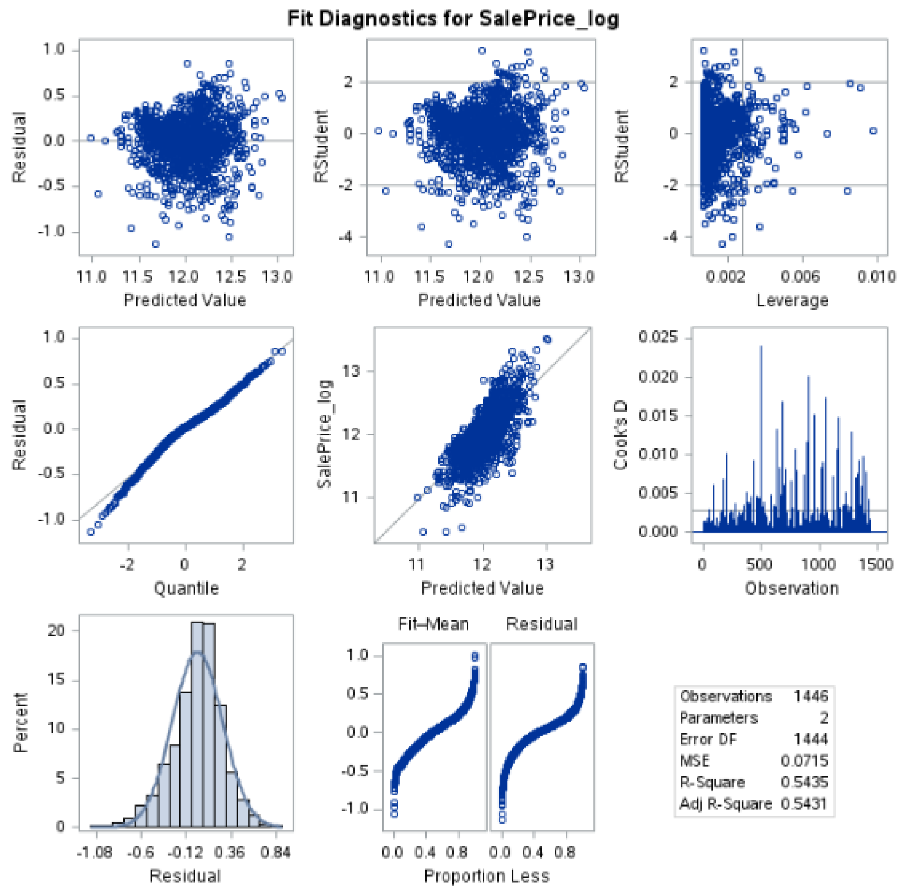
2^**.889567 +-** 0.02145638

1.85262 +- (.02145638)

iii. (1.83, 1.87)

h. Observing Assumptions

Figure 3.41



i.

2. Multiple Linear Regression: SalePrice~GrLiveArea + FullBath

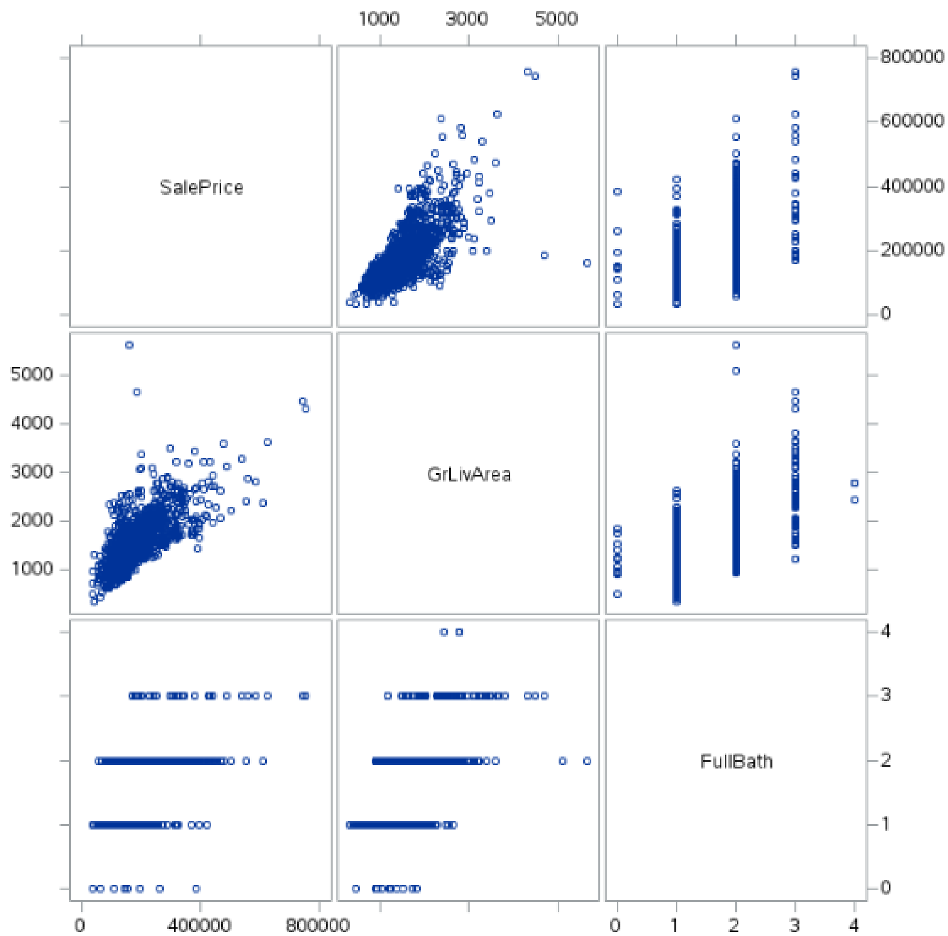
a. Exploratory Data Analysis: Visual Scatter Plot

Figure 3.42

```
/* Scatter Plot */
proc sgscatter data=train2;
    matrix SalePrice GrLivArea FullBath;
run;
```

i.

Figure 3.43



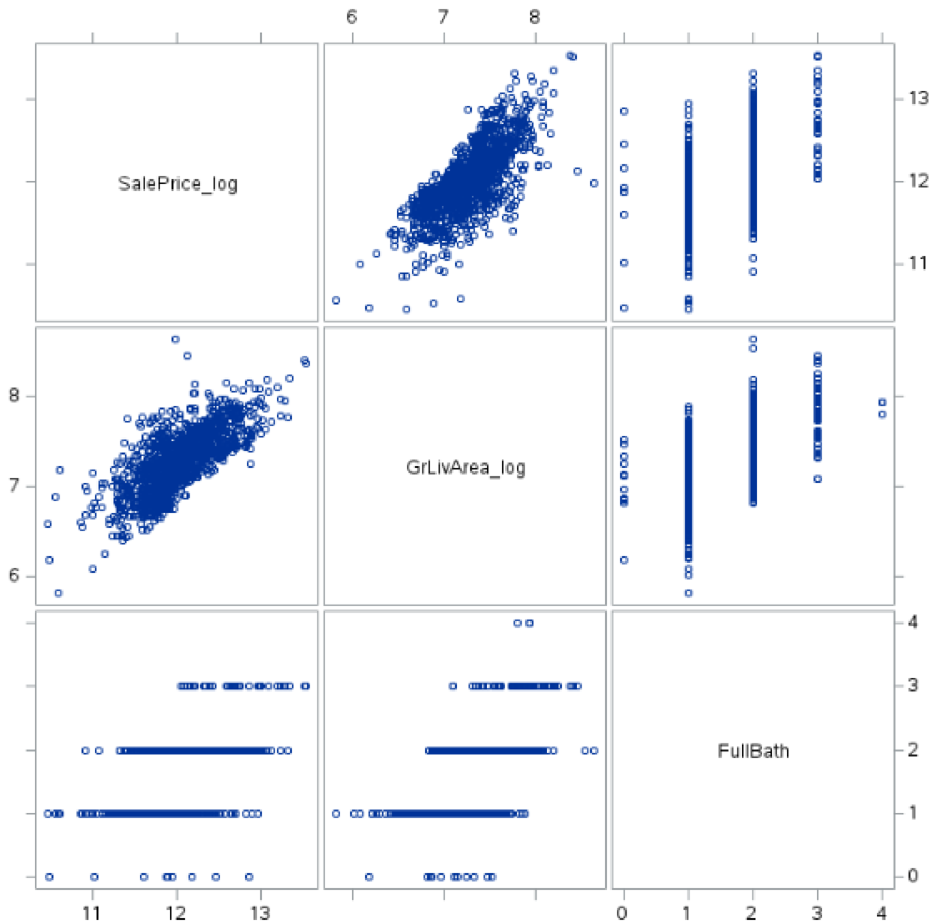
ii.

Figure 3.44

```
proc sgscatter data=train2;
    matrix SalePrice_log GrLivArea_log FullBath;
run;
```

iii.

Figure 3.45



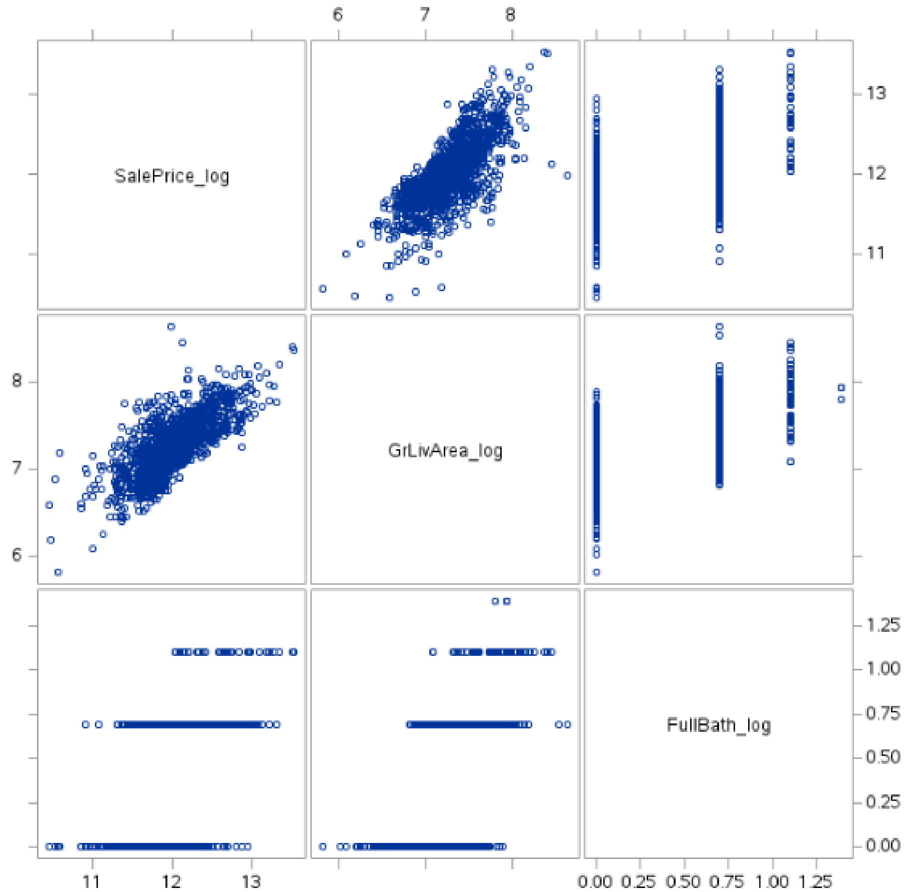
iv.

Figure 3.46

```
proc sgscatter data=train2;
    matrix SalePrice_log GrLivArea_log FullBath_log;
run;
```

v.

Figure 3.47



vi.

b. Fitting the model: log(SalePrice)~log(GrLivArea)+log(FullBath)

Figure 3.48

```
proc glm data = train2 plots = all;
    model SalePrice_log = GrLivArea_log FullBath_log / cli solution;
run;
```

i.

Figure 3.49

The GLM Procedure

Dependent Variable: SalePrice_log

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	127.4498766	63.7249383	915.61	<.0001
Error	1448	100.7782745	0.0695983		
Corrected Total	1450	228.2281511			

R-Square	Coeff Var	Root MSE	SalePrice_log Mean
0.558432	2.193818	0.263815	12.02537

Source	DF	Type I SS	Mean Square	F Value	Pr > F
GrLivArea_log	1	121.0198192	121.0198192	1738.83	<.0001
FullBath_log	1	6.4300575	6.4300575	92.39	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
GrLivArea_log	1	44.44628993	44.44628993	638.61	<.0001
FullBath_log	1	6.43005745	6.43005745	92.39	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	6.874573219	0.19373814	35.48	<.0001
GrLivArea_log	0.695320952	0.02751482	25.27	<.0001
FullBath_log	0.245702336	0.02556237	9.61	<.0001

ii.

c. Revised model: log(SalePrice)~log(GrLivArea)+ FullBath

i. Figure 3.50

```
/* Runnning model without outlier */
proc glm data = train2Q2NoOutliers plots = all;
    model SalePrice_log = GrLivArea_log FullBath / cli solution;
run;
```

ii. Figure 3.51

```
proc glmselect data=train2Q2NoOutliers;
  model SalePrice_log = GrLivArea_log FullBath /  selection=Stepwise(stop=CV) cvmethod = random(5) stats = adjrsq;
run;
```

iii.

Figure 3.52

The GLMSELECT Procedure

Data Set	WORK.TRAIN2Q2NOOUTLIERS
Dependent Variable	SalePrice_log
Selection Method	Stepwise
Select Criterion	SBC
Stop Criterion	Cross Validation
Cross Validation Method	Random
Cross Validation Fold	5
Effect Hierarchy Enforced	None
Random Number Seed	600749668

Number of Observations Read	2918
Number of Observations Used	1459

Dimensions	
Number of Effects	3
Number of Parameters	3

The GLMSELECT Procedure

Stepwise Selection Summary						
Step	Effect Entered	Effect Removed	Number Effects In	Adjusted R-Square	SBC	CV PRESS
0	Intercept		1	0.0000	-2670.4628	232.9921
1	GrLivArea_log		2	0.5396	-3795.9722	107.1999
2	FullBath		3	0.5620*	-3862.3565*	102.1023*
* Optimal Value of Criterion						

Selection stopped because all effects are in the final model.

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 2).

Effects:	Intercept GrLivArea_log FullBath

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	2	130.97320	65.48660	936.39
Error	1456	101.82577	0.06994	
Corrected Total	1458	232.79897		

iv.

Figure 3.53

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 2).

Effects:	Intercept GrLivArea_log FullBath

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	2	130.97320	65.48660	936.39
Error	1456	101.82577	0.06994	
Corrected Total	1458	232.79897		

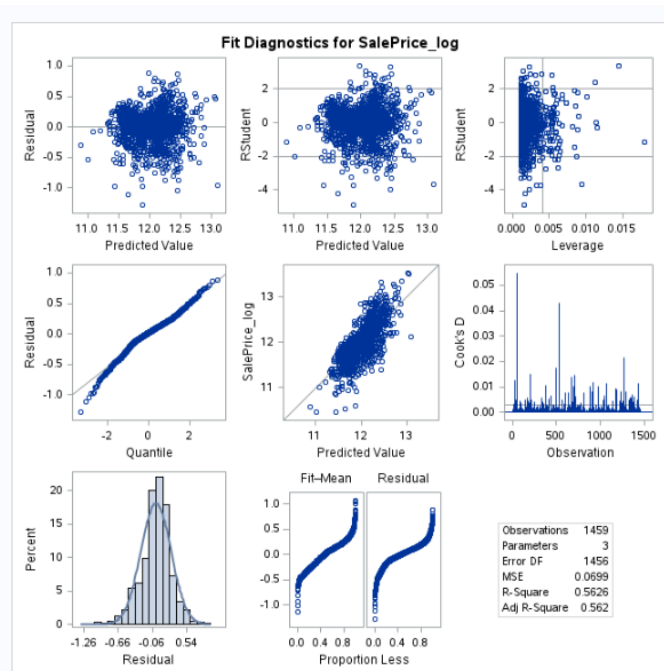
Root MSE	0.26445
Dependent Mean	12.02408
R-Square	0.5626
Adj R-Sq	0.5620
AIC	-2417.21304
AICC	-2417.18553
SBC	-3862.35652
CV PRESS	102.10225

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	6.507627	0.184889	35.20
GrLivArea_log	1	0.728027	0.027626	26.35
FullBath	1	0.144431	0.016633	8.68

v.

Figure 3.54



vi.

3. Custom Multiple Linear Regression Mode:Log(SalePrice) ~ Log(OverallQual) + Log(GrLivArea) + Log(FirstFlrSf) + LotArea + FullBath

Figure 3.55

a.

```
/* Question 3 */
proc glmselect data=train2;
  model SalePrice_log = OverallQual_log GrLivArea_log FirstFlrSf_log LotArea FullBath
        / selection=Stepwise(stop=CV)
          cvmethod=random(5)
          stats=adjrsq;
run;
```

Figure 3.56

b.

The GLMSELECT Procedure

Data Set	WORK.TRAIN2Q2NOOUTLIERS
Dependent Variable	SalePrice_log
Selection Method	Stepwise
Select Criterion	SBC
Stop Criterion	Cross Validation
Cross Validation Method	Random
Cross Validation Fold	5
Effect Hierarchy Enforced	None
Random Number Seed	609508673

Number of Observations Read	2918
Number of Observations Used	1459

Dimensions	
Number of Effects	6
Number of Parameters	6

The GLMSELECT Procedure

Stepwise Selection Summary

Step	Effect Entered	Effect Removed	Number Effects In	Adjusted R-Square	SBC	CV PRESS
0	Intercept		1	0.0000	-2670.4628	233.2471
1	OverallQual_log		2	0.6325	-4124.5695	85.9367
2	GrLivArea_log		3	0.7362	-4602.2770	61.9195
3	FirstFlrSf_log		4	0.7726	-4812.7693	53.3818
4	LotArea		5	0.7797	-4852.3264	52.2520
5	FullBath		6	0.7825*	-4864.7378*	51.6776*

* Optimal Value of Criterion

Selection stopped because all effects are in the final model.

The GLMSELECT Procedure
Selected Model

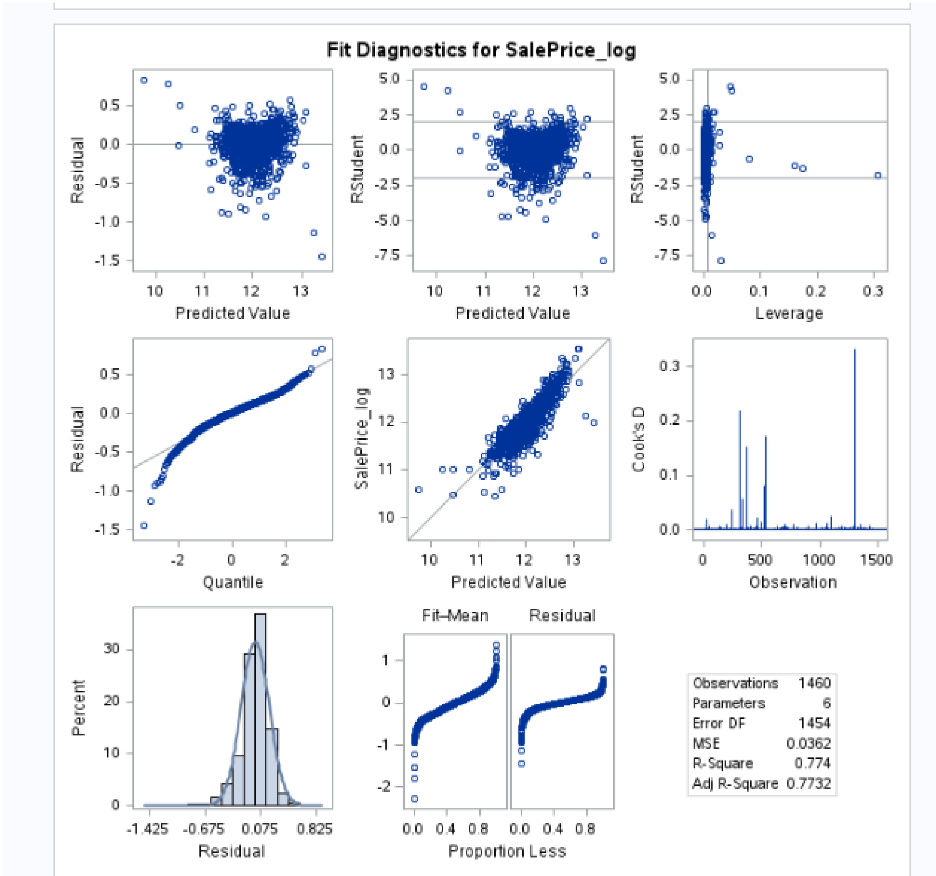
The selected model is the model at the last step (Step 5).

Effects: Intercept OverallQual_log GrLivArea_log FirstFlrSf_log LotArea FullBath

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	5	182.33550	36.46710	1050.00

Figure 3.57

c. 

GitHub Link: [stedua22/MSDS-6371-Stats-Kaggle-Project (github.com)]()