Assignment #3
Nitin Gaonkar

Introduction:

The purpose of this assignment is to build regression models for the home sale price, in this assignment I will be transforming the variables and build regression models and then compare all the models to see which is the best fit, I will be using the ODS SAS outputs to compare the models, along with that I will be working on finding the potential outliers in the observations and and work on few cases where we may to delete or retain the outliers.

Results:

1. Transforming the variables:

Here we have transformed the variable saleprice (y) and the best continuous predictor(x) Grlivarea, below table gives us the sample five observations of the actual variable and the transformed variable.

Proc print has been used to print the observation.

Obs	GrLivArea	SalePrice	log_saleprice	log_grlivarea
1	1656	215000	12.2784	7.41216
2	896	105000	11.5617	6.79794
3	1329	172000	12.0552	7.19218
4	2110	244000	12.4049	7.65444
5	1629	189900	12.1543	7.39572

2.

a. Model saleprice vs grlivarea

Model salesprice vs. Grlivarea

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Number of Observations Read	2930
Number of Observations Used	2930

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	9.33763E12	9.33763E12	2922.59	<.0001
Error	2928	9.354907E12	3194981962		
Corrected Total	2929	1.869254E13			

Root MSE	56524	R-Square	0.4995
Dependent Mean	180796	Adj R-Sq	0.4994
Coeff Var	31.26405		

Parameter Estimates

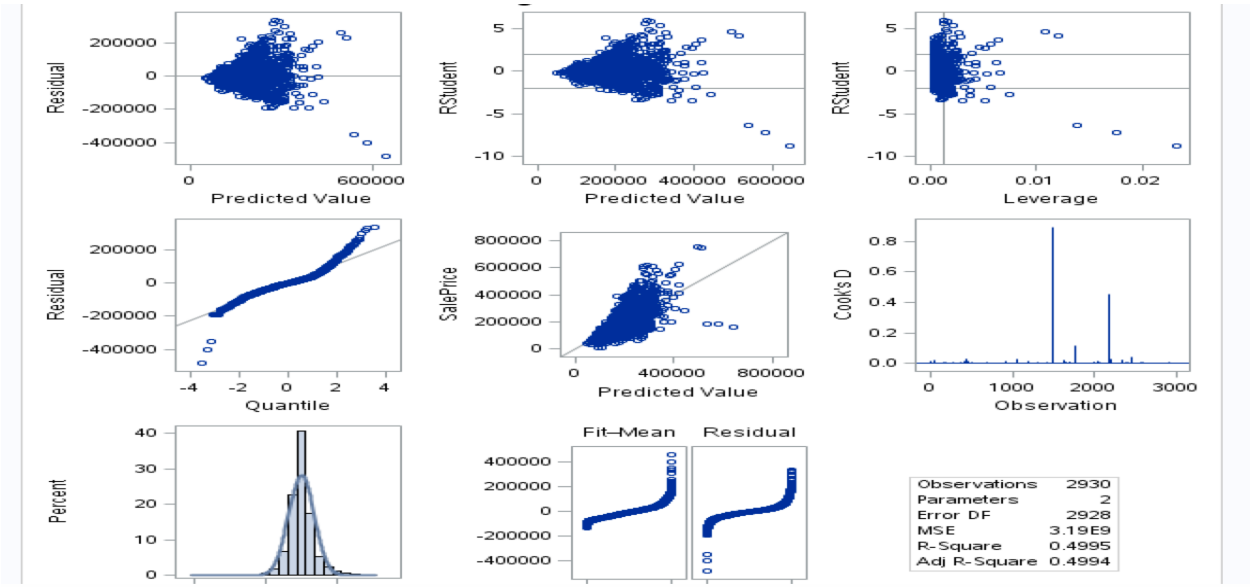
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	13290	3269.70277	4.06	<.0001
GrLivArea	1	111.69400	2.06607	54.06	<.0001

Model in equation form:

saleprice= 13290 + 111.694 * GrLivArea

Here for each increase of 1 sq foot in living area, the sale price will increase by $111.694

ODS SAS outputs for the mode saleprice vs GrlivArea:



b. Model saleprice vs log(grlivarea)

Model saleprice vs. Log_Grlivarea

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Number of Observations Read	2930
Number of Observations Used	2930

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	9.030218E12	9.030218E12	2736.45	<.0001
Error	2928	9.662319E12	3299972433		
Corrected Total	2929	1.869254E13			

Root MSE	57445	R-Square	0.4831
Dependent Mean	180796	Adj R-Sq	0.4829
Coeff Var	31.77358		

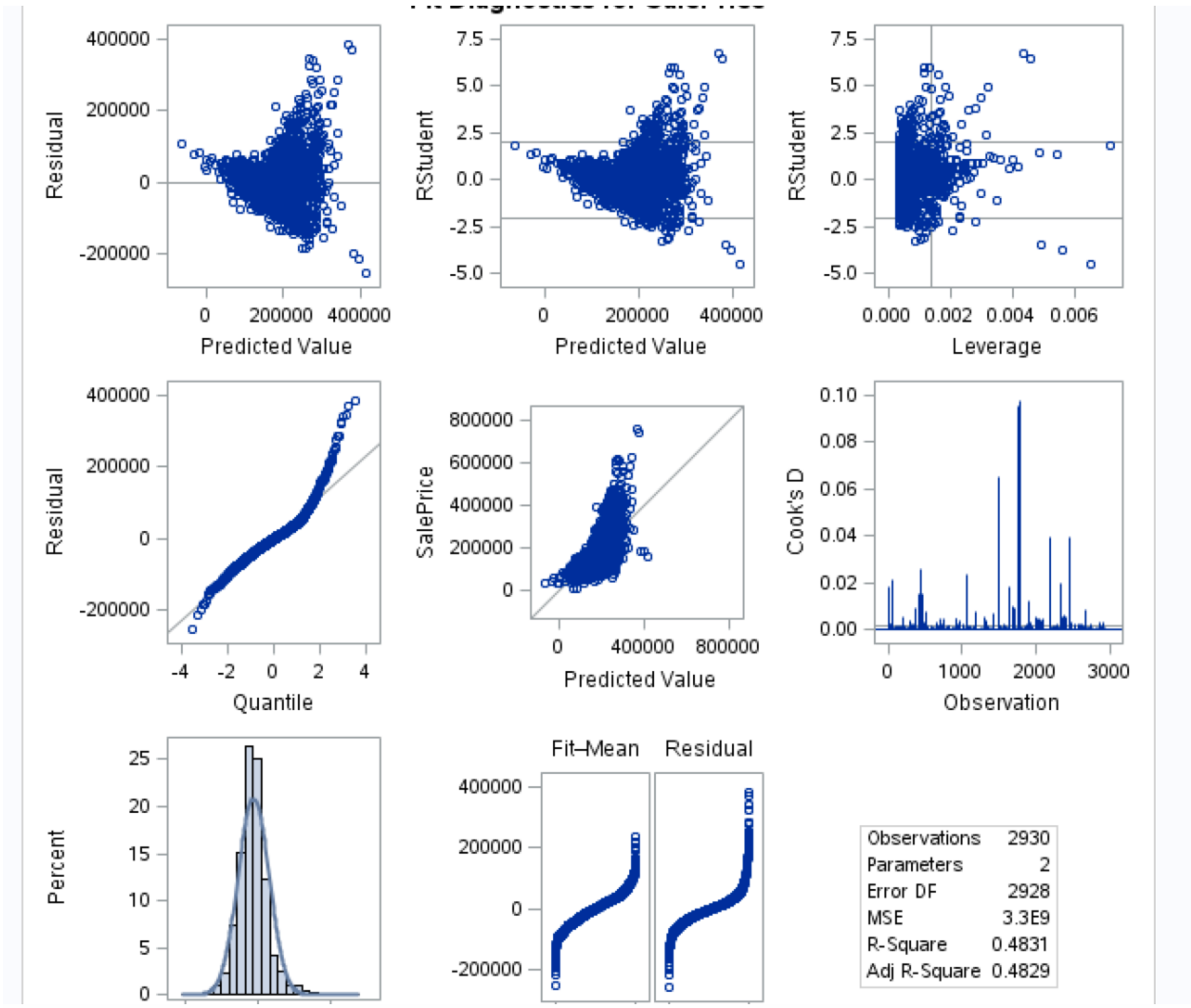
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-1060765	23758	-44.65	<.0001
log_grlivarea	1	171011	3269.11261	52.31	<.0001

Model in equation form:

saleprice= -1060765 + 171011 * log(GrLivArea)

Here for each unit increase in log(grlivarea) will increase sale price by 171011 $


Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
SalePrice
Predicted Value
Cook's D
Observation
Percent
Fit–Mean
Residual
Observations 2930
Parameters 2
Error DF 2928
MSE 3.3E9
R-Square 0.4831
Adj R-Square 0.4829

c. Model log(saleprice) vs grlivarea

Model log_saleprice vs. grlivearea

The REG Procedure
Model: MODEL1
Dependent Variable: log_saleprice

Number of Observations Read	2930
Number of Observations Used	2930

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	235.61694	235.61694	2748.89	<.0001
Error	2928	250.96931	0.08571		
Corrected Total	2929	486.58626			

Root MSE	0.29277	R-Square	0.4842
Dependent Mean	12.02097	Adj R-Sq	0.4840
Coeff Var	2.43548		

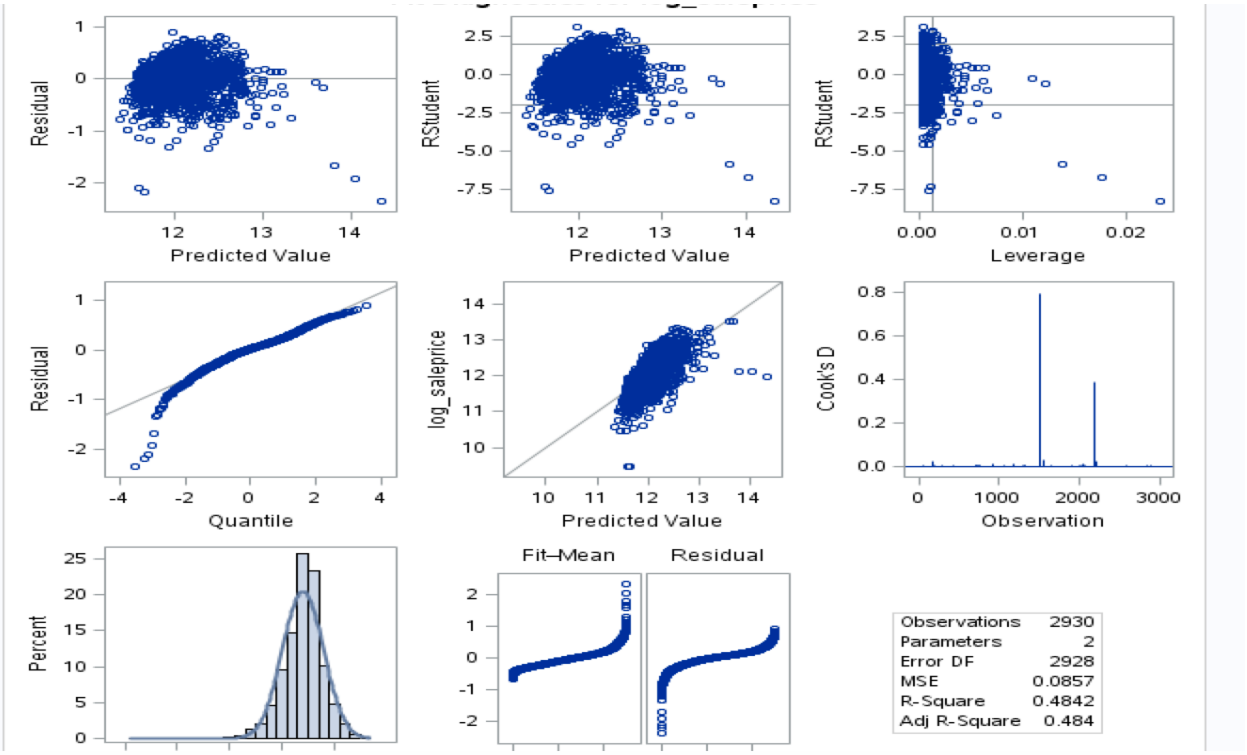
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	11.17954	0.01694	660.12	<.0001
GrLivArea	1	0.00056107	0.00001070	52.43	<.0001

Model in equation form:

Log(saleprice)= 11.17954 + 0.00056107 * GrLivArea

Here for each increase of 1 sq foot in living area, the sale price will increase by log (saleprice) by 0.000567 units.


Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
log_saleprice
Predicted Value
Cook's D
Observation
Percent
Fit–Mean
Residual
Observations 2930
Parameters 2
Error DF 2928
MSE 0.0857
R-Square 0.4842
Adj R-Square 0.484

D. Model log(saleprice) vs log(grlivarea)

Model log_saleprice vs. log_grlivarea

The REG Procedure
Model: MODEL1
Dependent Variable: log_saleprice

Number of Observations Read	2930
Number of Observations Used	2930

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	254.46967	254.46967	3209.97	<.0001
Error	2928	232.11659	0.07927		
Corrected Total	2929	486.58626			

Root MSE	0.28156	R-Square	0.5230
Dependent Mean	12.02097	Adj R-Sq	0.5228
Coeff Var	2.34222		

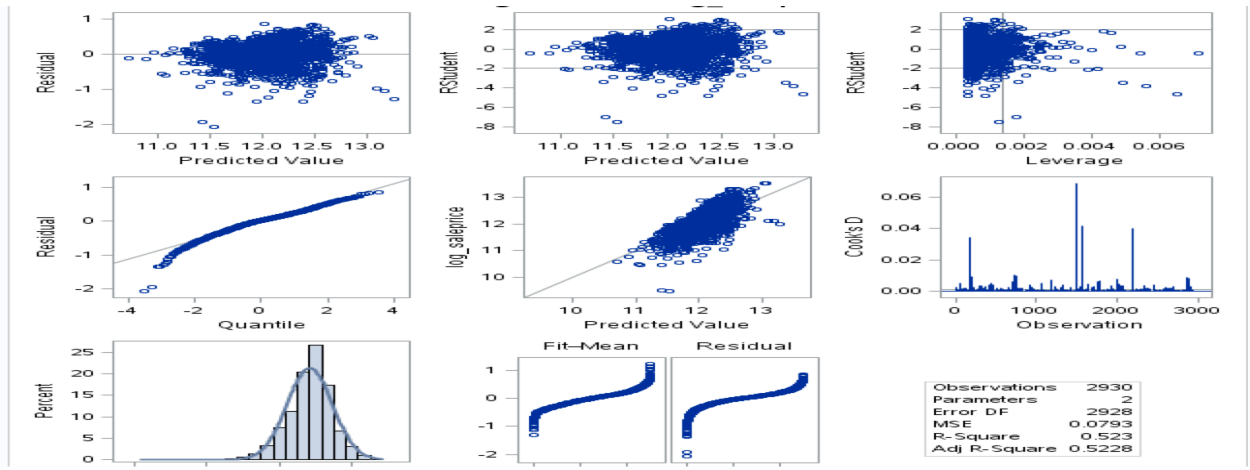
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	5.43019	0.11644	46.63	<.0001
log_grlivarea	1	0.90781	0.01602	56.66	<.0001

Model in equation form:

Log(saleprice)= 5.43019 + 0.90781 * log(GrLivArea)

Here for each unit increase of log(grlivarea) will increase log saleprice by 0.90781 units.



	r-squared value	fit-mean/residual	cook's D	qq plot	scatter plot	Histogram
Model1 (saleprice vs grlivarea)	0.4995	fit mean is greater than the residual	three dominant outliers	points are way off the line	few of the points are away from the line	except few a the points ar covered by tl SAS curve
Model 2 (saleprice vs log(grlivarea))	0.4831	residual is greated than the fit mean	lot of outliers	the points are way off the line	Points are not distributed in a straight line	except few a the points ar covered by tl SAS curve
Model 3(log(saleprice) vs grlivarea)	0.4842	fit mean is greater than the residual	couple of outliers	most of the points are on the line	most of the points are on the line except few of the outliers.	most of the points are covered by tl SAS curve
Model 4 log(saleprice) vs log(grlivarea)	0.523	fit mean is greater than the residual	lot of outliers	all the points are on the line	graphs looks good and all the points are distributed in a straight line	All the points are covered l the SAS curv

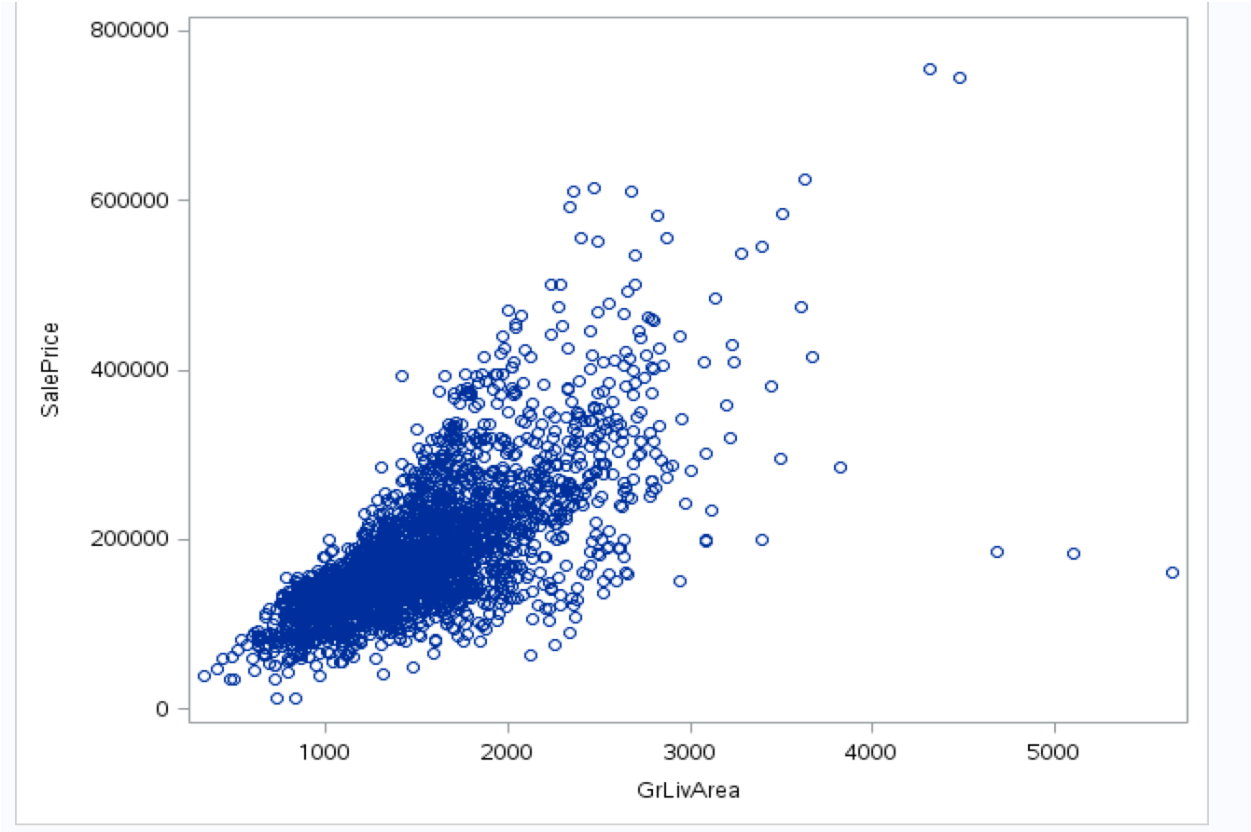
By looking at the metric I think model 4 is the best fit, the only concern about this fit is the cook's D plot, I could see a lot of outliers in this graph. The natural way to interpret the log model is exponentiation of regression coefficients, exp(β), since exponentiation is the inverse of logarithm function.

3 Correlation:

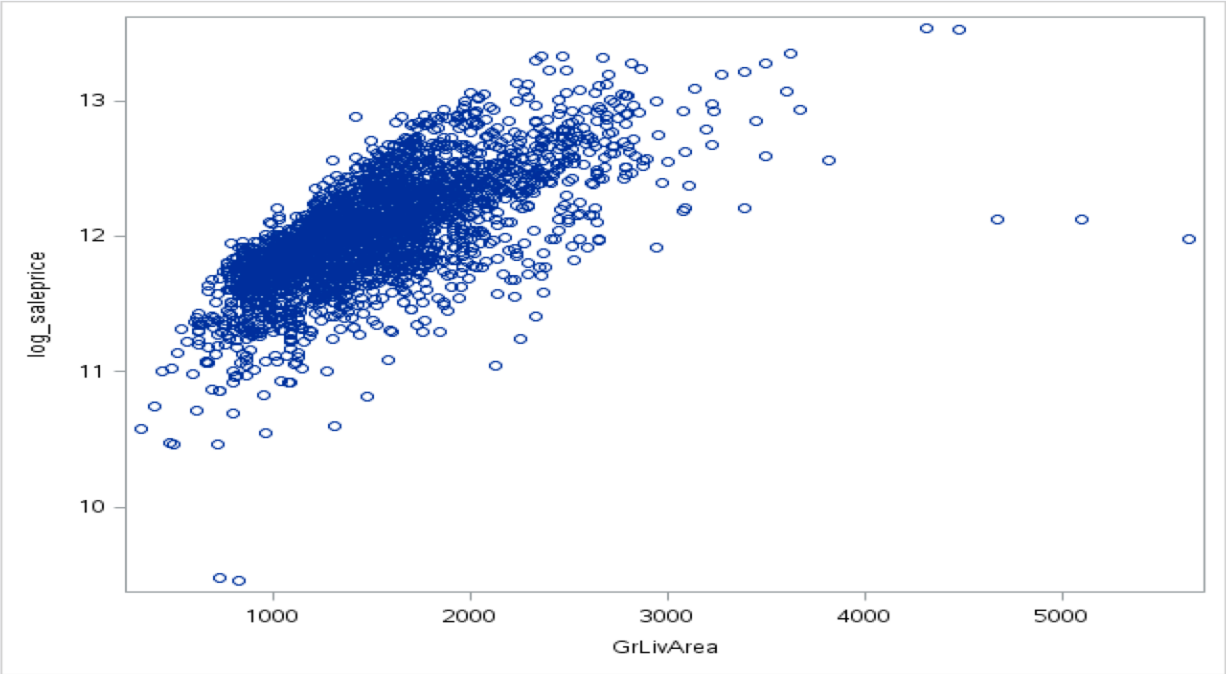
	log_saleprice	SalePrice	OverallQual	log_grlivarea	GrLivArea	GarageCars	GarageArea	TotalBsmtSF	YearBuilt	FirstFlrSF	YearRemodel	GarageYrBlt	FullBath
log_saleprice	1.00000	0.94630	0.82564	0.72317	0.69586	0.67532	0.65113	0.62510	0.61548	0.60263	0.58615	0.58050	0.57733
		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
	2930	2930	2930	2930	2930	2929	2929	2929	2930	2930	2930	2771	2930

By using the proc corr, we can see that the overallQual has the highest correlation coefficient, but since its not a continuous variable I am choosing the next one that is Grlivarea as X.

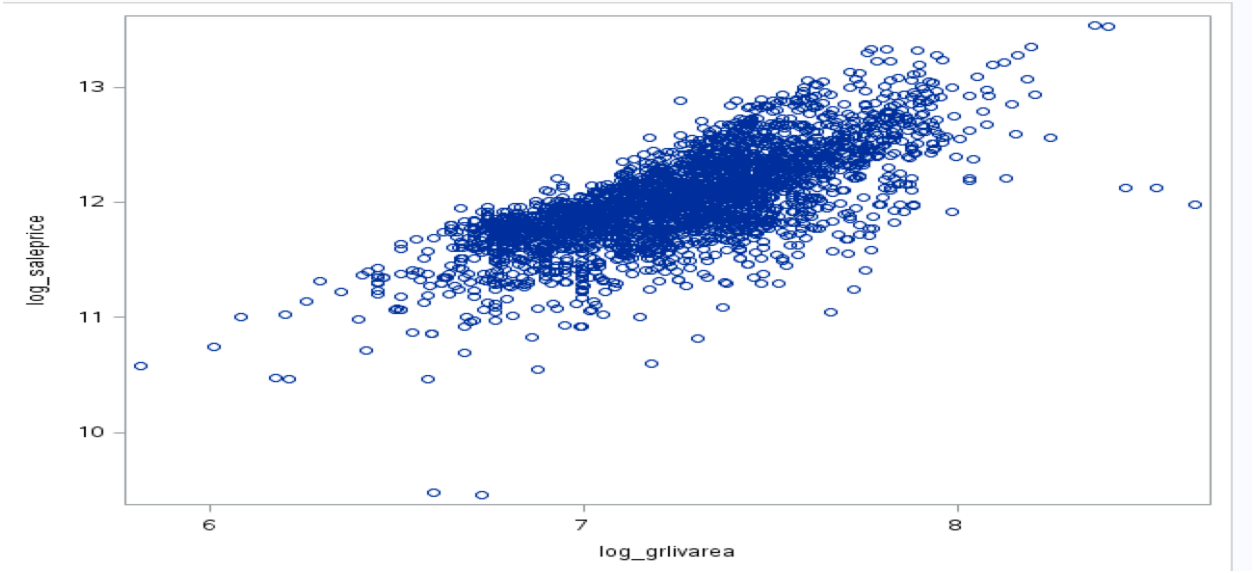
Scatter plot of saleprice and grlivarea



Scatter plot of log_saleprice and grlivarea



4. **We can see from the scatter plot that log transforming both sales price and X variable grlivarea we get the best fit, also by looking at the R value and evaluating the ODS SAS output when we transform both the variables we get the best fit.**



Model log_saleprice vs. log_grlivarea

The REG Procedure
Model: MODEL1
Dependent Variable: log_saleprice

Number of Observations Read	2930
Number of Observations Used	2930

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	254.46967	254.46967	3209.97	<.0001
Error	2928	232.11659	0.07927		
Corrected Total	2929	486.58626			

Root MSE	0.28156	R-Square	0.5230
Dependent Mean	12.02097	Adj R-Sq	0.5228
Coeff Var	2.34222		

Parameter Estimates

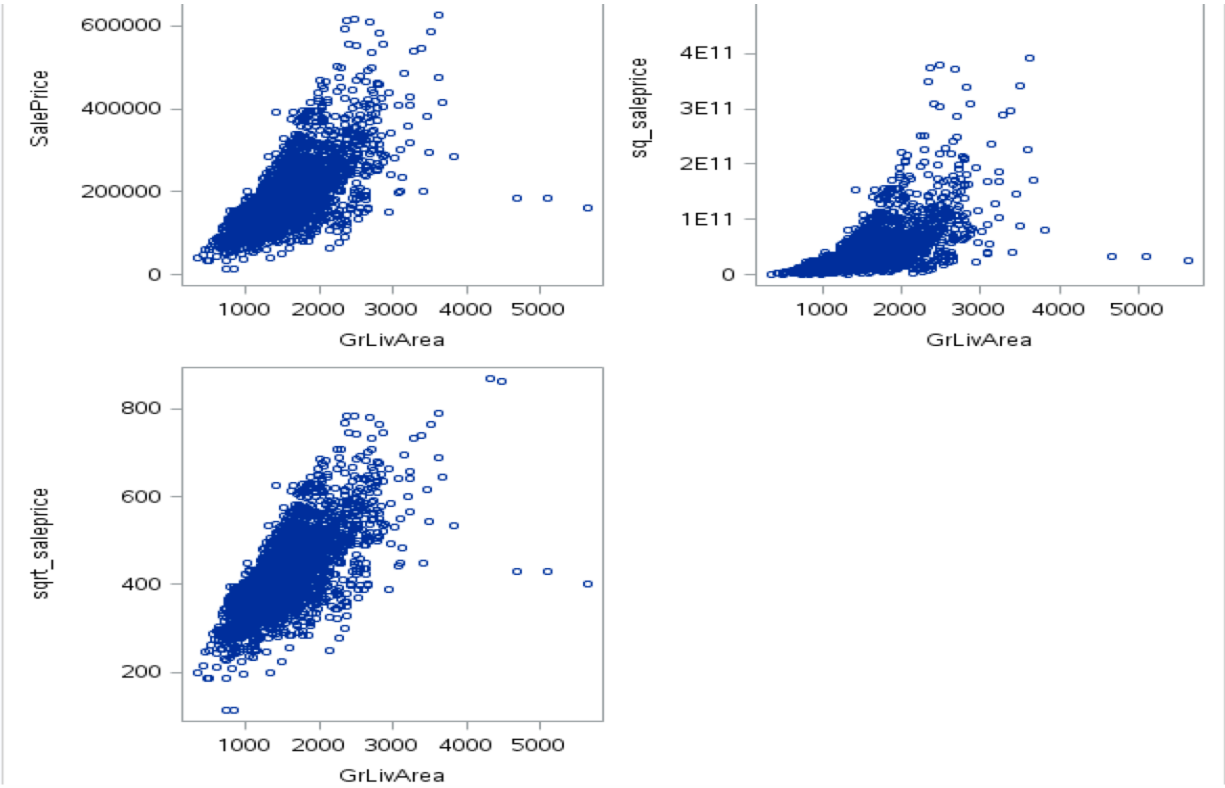
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	5.43019	0.11644	46.63	<.0001
log_grlivarea	1	0.90781	0.01602	56.66	<.0001

now I have transformed the y variables by using square and square root and store in the variables sq_saleprice and sqrt_saleprice.

Sample data after the transformation:

Obs	GrLivArea	SalePrice	log_saleprice	log_grlivarea	sq_saleprice	sqrt_saleprice
1	1656	215000	12.2784	7.41216	46225000000	463.681
2	896	105000	11.5617	6.79794	11025000000	324.037
3	1329	172000	12.0552	7.19218	29584000000	414.729
4	2110	244000	12.4049	7.65444	59536000000	493.964
5	1629	189900	12.1543	7.39572	36062010000	435.775

Below are the scatter plots for transformed saleprice and grlivarea:



above scatter plots are saleprice vs grlivarea, sq_saleprice vs grlivarea and sqrt_saleprice vs grlivarea,

by looking at those scatter plots we can see that the sqrt_saleprice vs grlivarea looks a good fit, let us run the regression for these two variables and investigate further to see how this model fits.

Model sqrt saleprice vs. grlivarea;run;OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;ODS HTML CLOSE;&GRAPHTERM; ;*

The REG Procedure
Model: MODEL1
Dependent Variable: sqrt_saleprice

Number of Observations Read	2930
Number of Observations Used	2930

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	11185123	11185123	3017.28	<.0001
Error	2928	10854156	3707.02050		
Corrected Total	2929	22039279			

Root MSE	60.88531	R-Square	0.5075
Dependent Mean	416.26208	Adj R-Sq	0.5073
Coeff Var	14.62668		

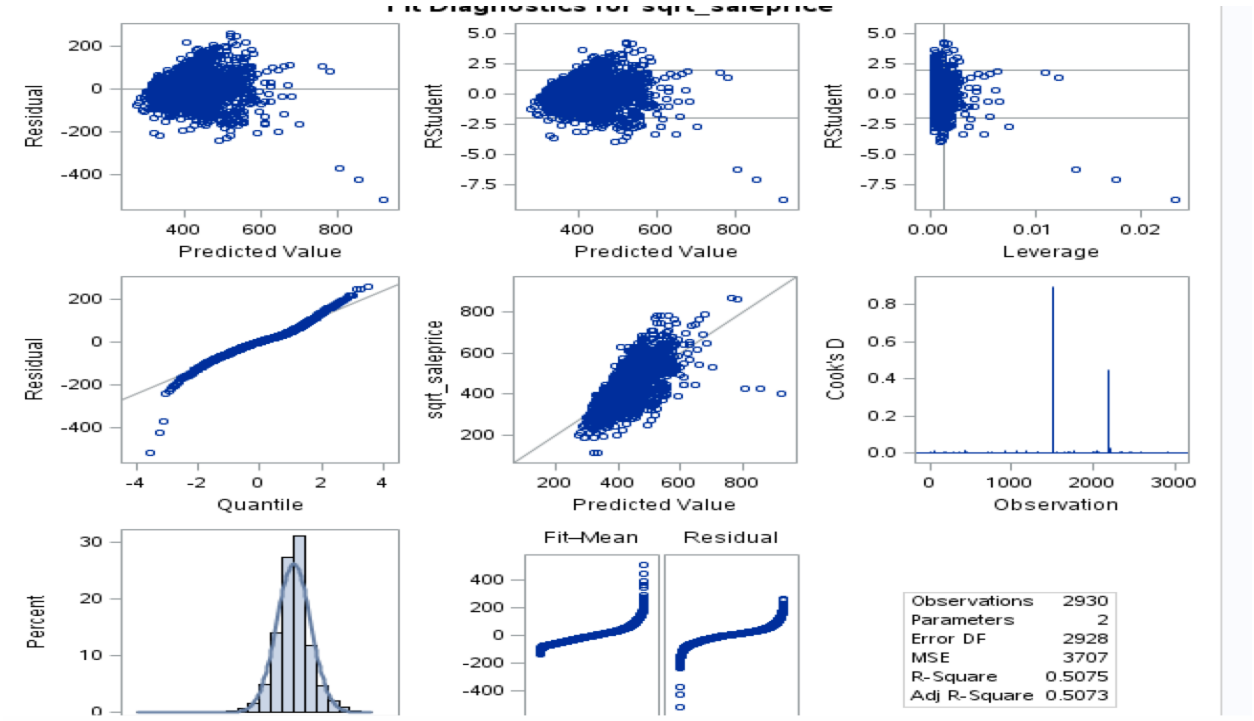
Parameter Estimates

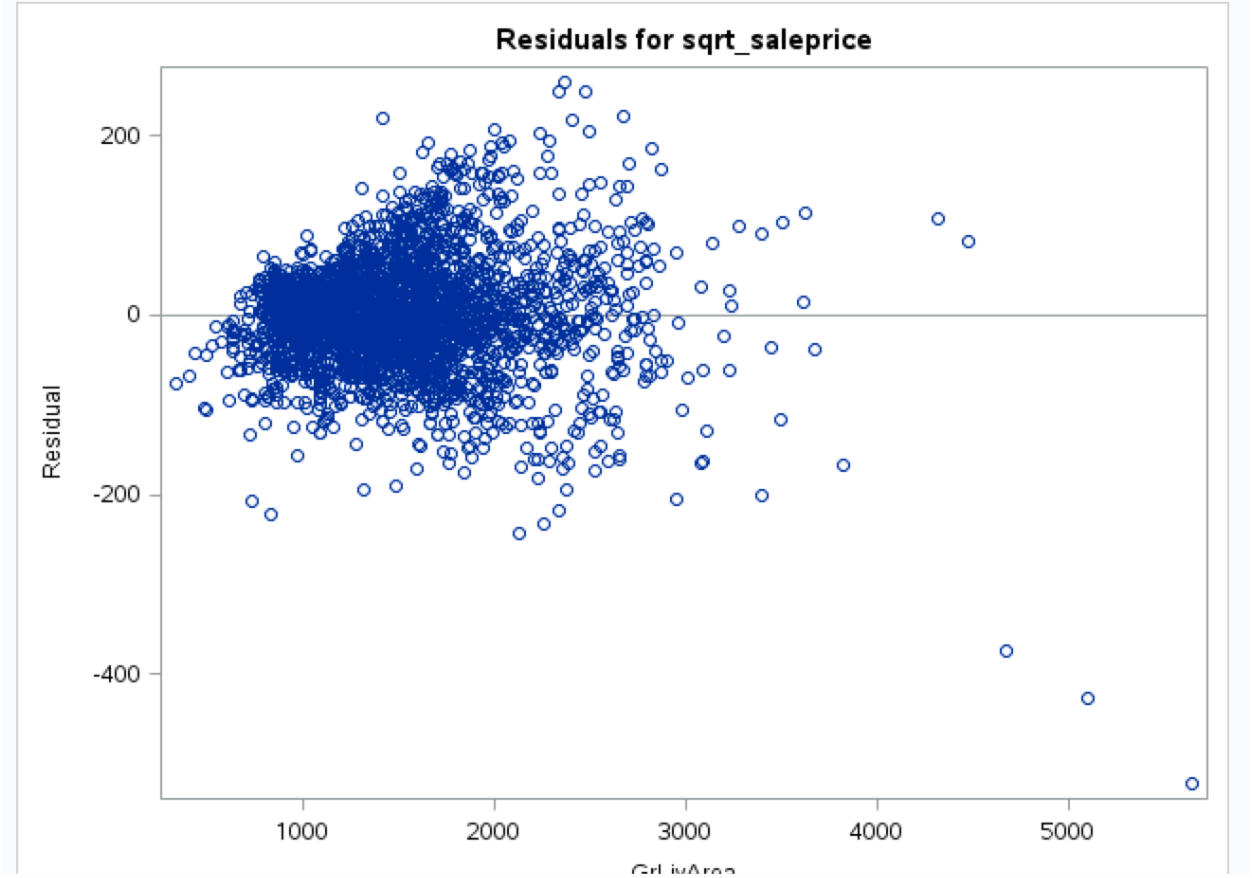
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	232.93209	3.52198	66.14	<.0001
GrLivArea	1	0.12225	0.00223	54.93	<.0001

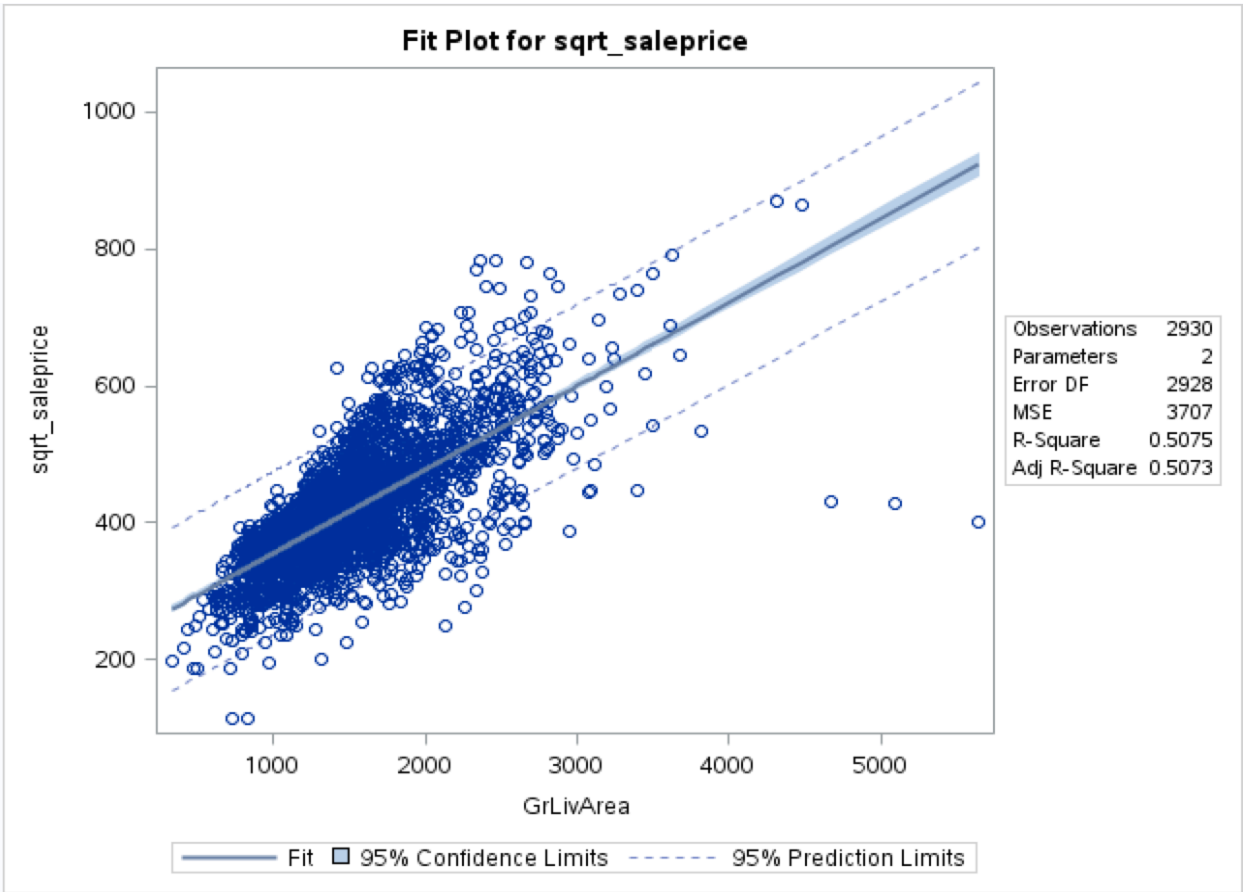
The equation for the model is:

Saleprice= 232.93209 + 0.1225 Grlivarea

Here for each increase of 1 sq foot in living area, the sale price will increase by $0.907


Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
sqrt_saleprice
Predicted Value
Cook's D
Observation
Percent
Fit–Mean
Residual
Observations 2930
Parameters 2
Error DF 2928
MSE 3707
R-Square 0.5075
Adj R-Square 0.5073


Residuals for sqrt_saleprice
Residual
GrLivArea


Fit Plot for sqrt_saleprice
sqrt_saleprice
1000
800
600
400
200
1000
2000
3000
4000
5000
GrLivArea
Observations 2930
Parameters 2
Error DF 2928
MSE 3707
R-Square 0.5075
Adj R-Square 0.5073
Fit
95% Confidence Limits
95% Prediction Limits

5. **Identifying outliers:**

The UNIVARIATE Procedure
Variable: SalePrice

Moments			
N	2930	Sum Weights	2930
Mean	180796.06	Sum Observations	529732456
Std Deviation	79886.6924	Variance	6381883616
Skewness	1.74350008	Kurtosis	5.11889995
Uncorrected SS	1.14466E14	Corrected SS	1.86925E13
Coeff Variation	44.1860803	Std Error Mean	1475.8446

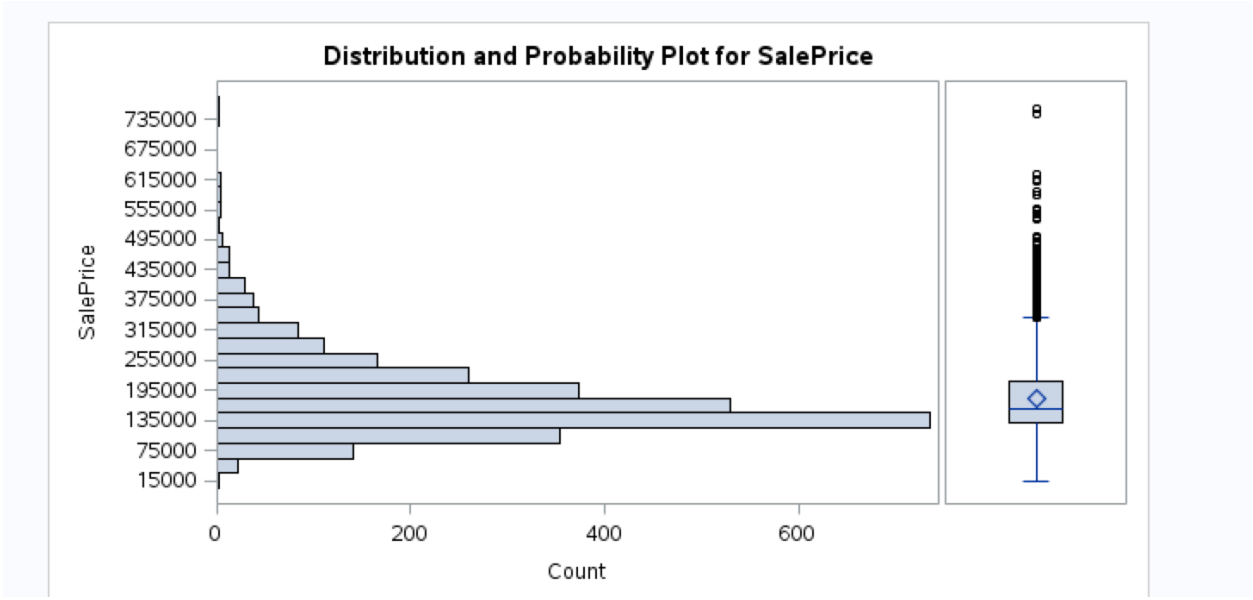
Basic Statistical Measures			
Location		Variability	
Mean	180796.1	Std Deviation	79887
Median	160000.0	Variance	6381883616
Mode	135000.0	Range	742211
		Interquartile Range	84000

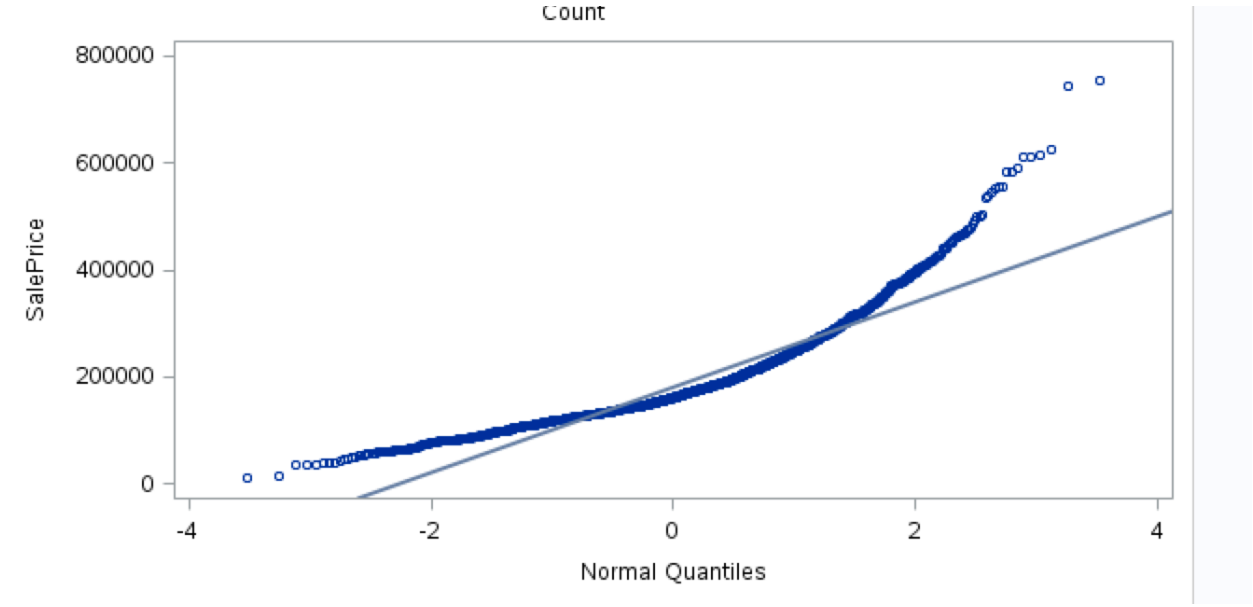
Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	122.5035	Pr > |t|	<.0001
Sign	M	1465	Pr >= |M|	<.0001
Signed Rank	S	2146958	Pr >= |S|	<.0001

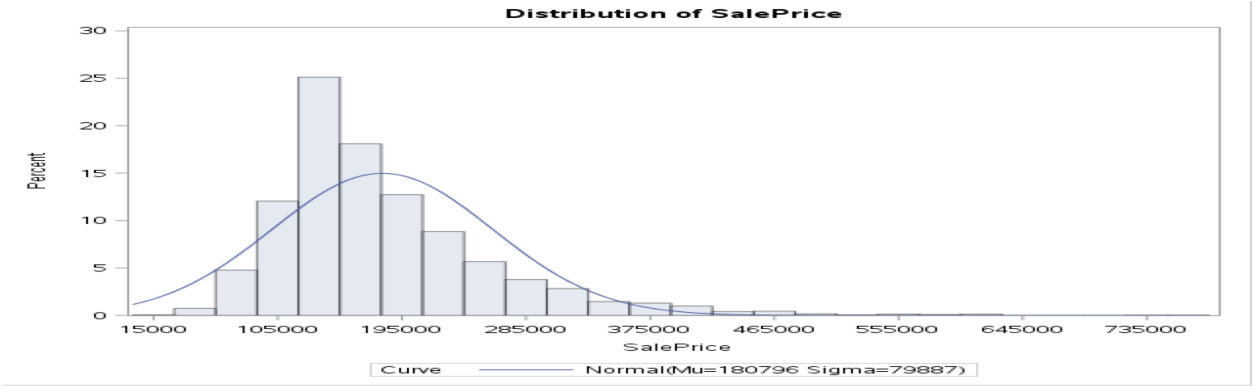
Tests for Normality				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.123422	Pr > D	<0.0100
Cramer-von Mises	W-Sq	15.38003	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	87.36744	Pr > A-Sq	<0.0050

Quantiles (Definition 5)	
Level	Quantile
100% Max	755000
99%	457347
95%	335000
90%	281357
75% Q3	213500
50% Median	160000
25% Q1	129500
10%	105250
5%	87500
1%	61500
0% Min	12789

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
12789	182	611657	45
13100	1554	615000	1064
34900	727	625000	2446
35000	2844	745000	1761
35311	2881	755000	1768


Distribution and Probability Plot for SalePrice
SalePrice
735000
675000
615000
555000
495000
435000
375000
315000
255000
195000
135000
75000
15000
0
200
400
600
Count


Count
SalePrice
800000
600000
400000
200000
0
-4
-2
0
2
4
Normal Quantiles


Distribution of SalePrice
Percent
30
25
20
15
10
5
0
15000
105000
195000
285000
375000
465000
555000
645000
735000
SalePrice
Curve
Normal(Mu=180796 Sigma=79887)

By observing the above graphs and tables we can see that the Max and min and median quantiles, we also have the table where the mean, median and other basic statistical measure have been show.

The distribution and probability plot for sale price gives us the distribution of the sale prices and the box plot shows us the outliers in the data.

The standard deviation is 79887 so I have considered the values outside the 2 standard deviations from mean as outliers.

So by using the if conditions below criteria has been set for the outliers:

if saleprice <= 21022 then price_outlier=1;

else if saleprice >= 21022 & saleprice < 340570 then price_outlier=2;

else if saleprice >= 340570 then price_outlier=3;

Obs	GrLivArea	SalePrice	price_outlier
1	1656	215000	2
2	896	105000	2
3	1329	172000	2
4	2110	244000	2
5	1629	189900	2
6	1604	195500	2
7	1338	213500	2
8	1280	191500	2
9	1616	236500	2
10	1804	189000	2
11	1655	175900	2
12	1187	185000	2
13	1465	180400	2
14	1341	171500	2
15	1502	212000	2
16	3279	538000	3
17	1752	164000	2
18	1856	394432	3
19	864	141000	2
20	2073	210000	2

above table gives the group to which the outlier belongs, the groups have been set by the if conditions in the SAS code.

Below frequency tables gives the information about the frequency of if the data points in these groups. As we can see from the table that the 95% of the data falls under group **two** which is within 2 standard deviations from the mean.

The FREQ Procedure

price_outlier	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	2	0.07	2	0.07
2	2794	95.36	2796	95.43
3	134	4.57	2930	100.00

Below tables gives us the information about mean, standard deviation and min and max values of all the three groups.

The MEANS Procedure

price_outlier=1

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
2	12944.50	219.9102089	12789.00	13100.00

price_outlier=2

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
2794	169410.68	59095.23	34900.00	340000.00

price_outlier=3

Analysis Variable : SalePrice				
N	Mean	Std Dev	Minimum	Maximum
134	420694.88	77956.75	341000.00	755000.00

6. **After identifying the outliers, I have removed the outliers, group 1 and group 3 data have been removed which contained the outliers**

The FREQ Procedure

price_outlier	Frequency	Percent	Cumulative Frequency	Cumulative Percent
2	2794	100.00	2794	100.00

a. **Below reg model is with the data without outlier's response variable saleprice with the predictor variable grlivarea.**

The equation of model is :

Saleprice=47397 + 83.77 * grlivarea

For every 1 sq foot increase in grlivarea the sale price increases by $83.77

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Number of Observations Read	2794
Number of Observations Used	2794

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	4.162679E12	4.162679E12	2078.67	<.0001
Error	2792	5.591165E12	2002566160		
Corrected Total	2793	9.753843E12			

Root MSE	44750	R-Square	0.4268
Dependent Mean	169411	Adj R-Sq	0.4266
Coeff Var	26.41512		

Parameter Estimates

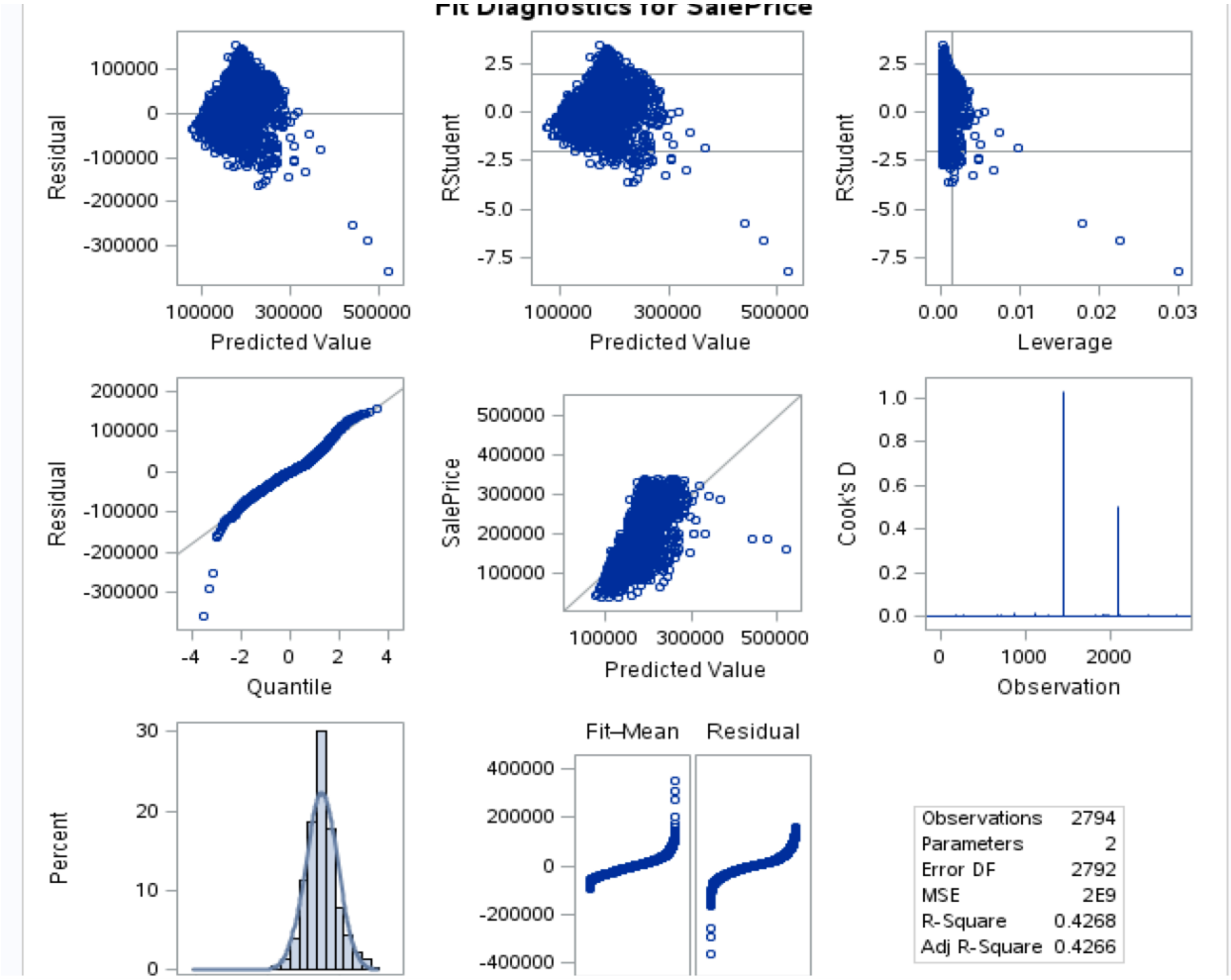
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	47397	2806.90605	16.89	<.0001
GrLivArea	1	83.77971	1.83758	45.59	<.0001

By looking at the below ODS SAS outputs :

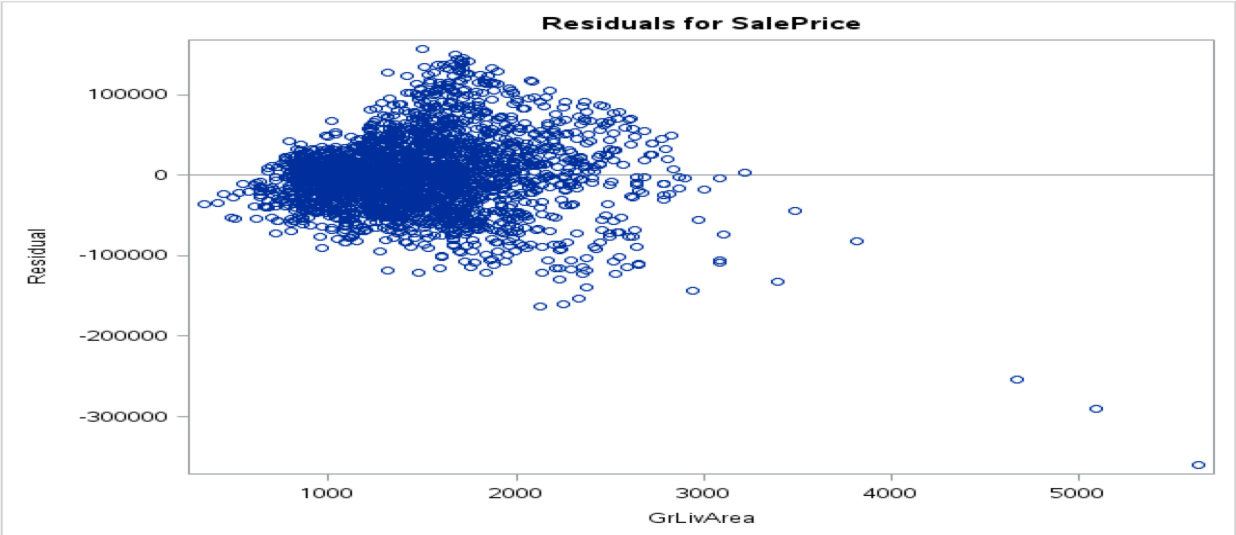
By observing the below graphs, we can see that in the residual graph data points are scattered and we can see few outliers.
In the Cook's D we can see couple of outliers around observation 2000

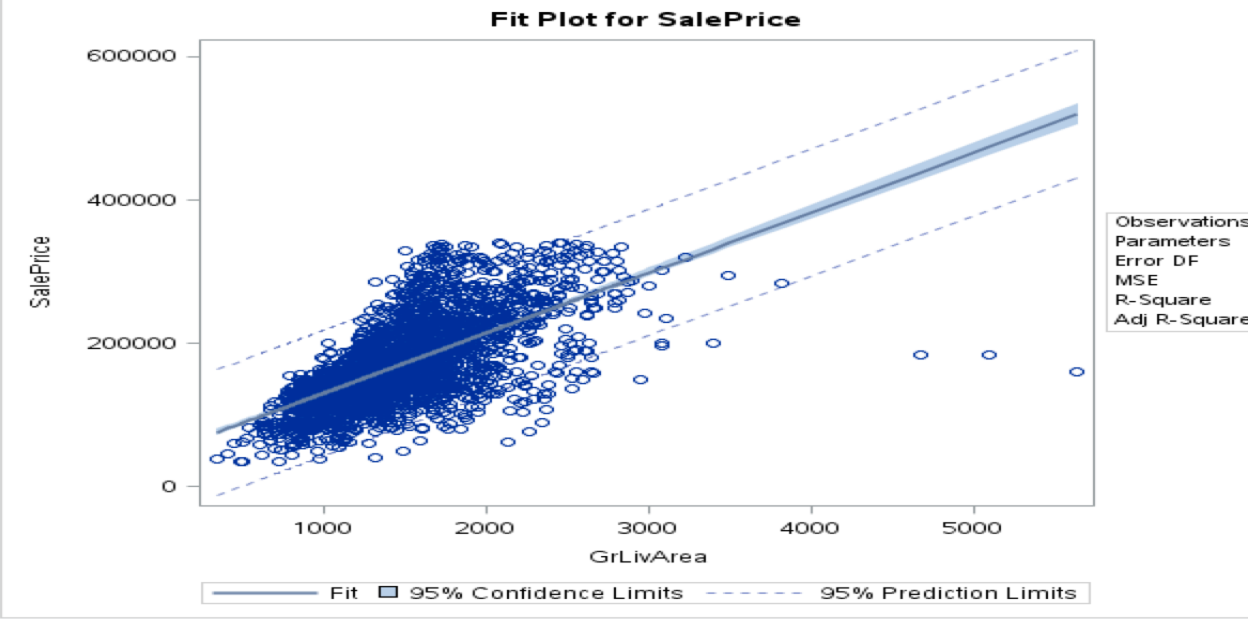
Also in fit mean and residual plot we can see the fit mean is higher than the residual which is a good sign

And also the qq plots looks pretty ok and also in the histogram the points are around the SAS curve.



The scatter plot looks ok, few of the points are way off and in the fir plot most of the data points fall within the prediction limits





b. Multiple regression model with the new data without outliers.

Model equation is:

Saleprice= 51625 + 66.5344 MasVnrArea + 76.73 GrLivArea

If GrLiveArea is fixed, then for each increase of 1 sq foot in masonary veneer area, the sale price will increase by $66.5344

If MasVnrArea is fixed, then for each increase of 1 sq foot in living area, the sale price will increase by $76.73

Forward Selection: Step 2

Variable MasVnrArea Entered: R-Square = 0.4558 and C(p) = 3.0000

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	4.402497E12	2.201248E12	1160.23	<.0001
Error	2770	5.255402E12	1897257069		
Corrected Total	2772	9.657899E12			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	51625	2761.19424	6.63215E11	349.57	<.0001
MasVnrArea	66.53444	5.57388	2.703365E11	142.49	<.0001
GrLivArea	76.73699	1.88495	3.144377E12	1657.33	<.0001

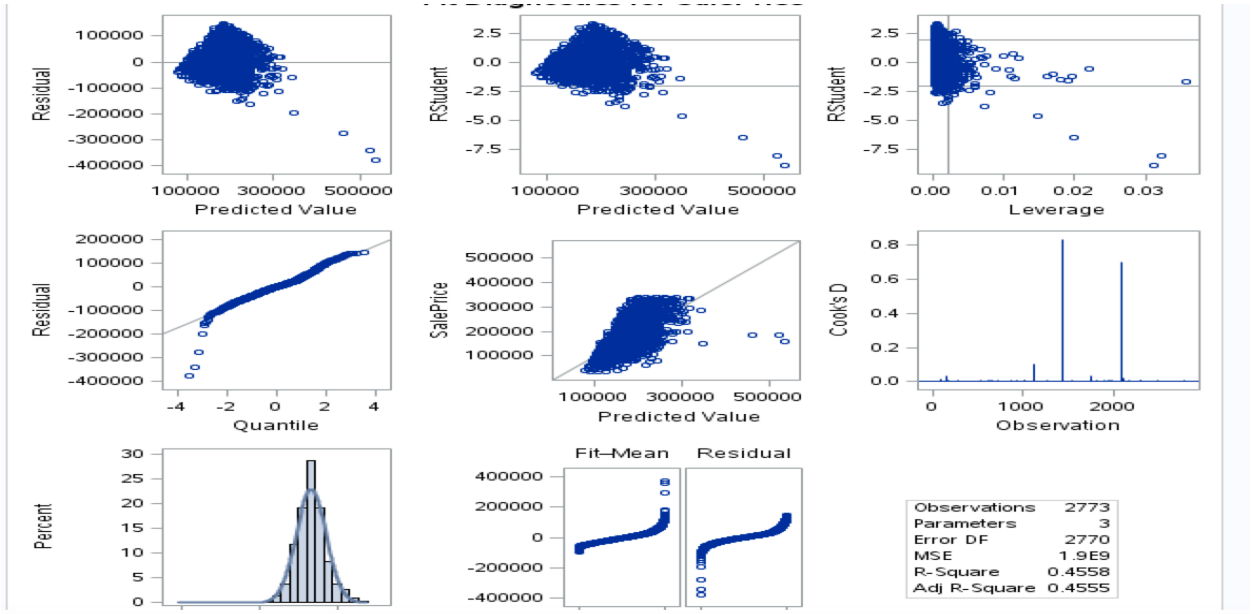
Bounds on condition number: 1.1052, 4.4207

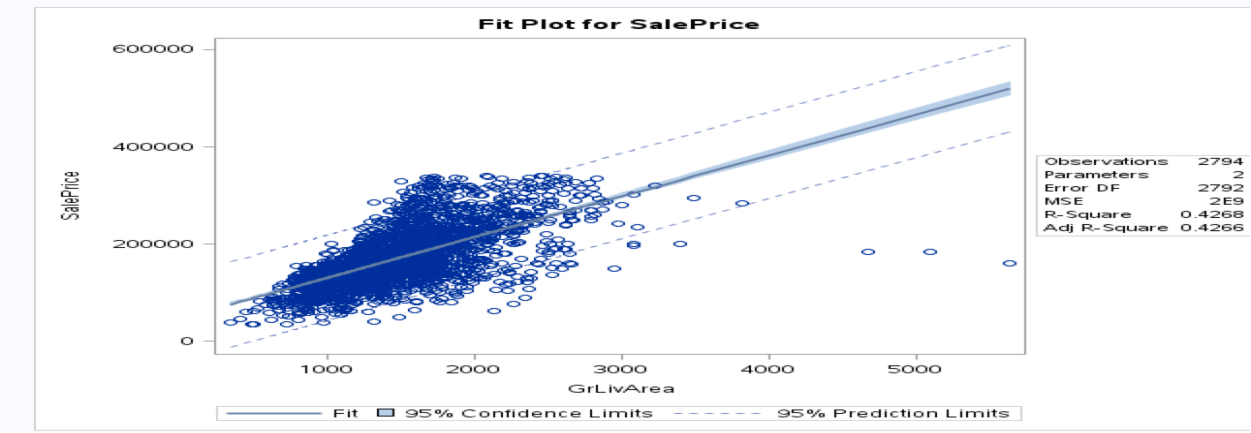
All variables have been entered into the model.

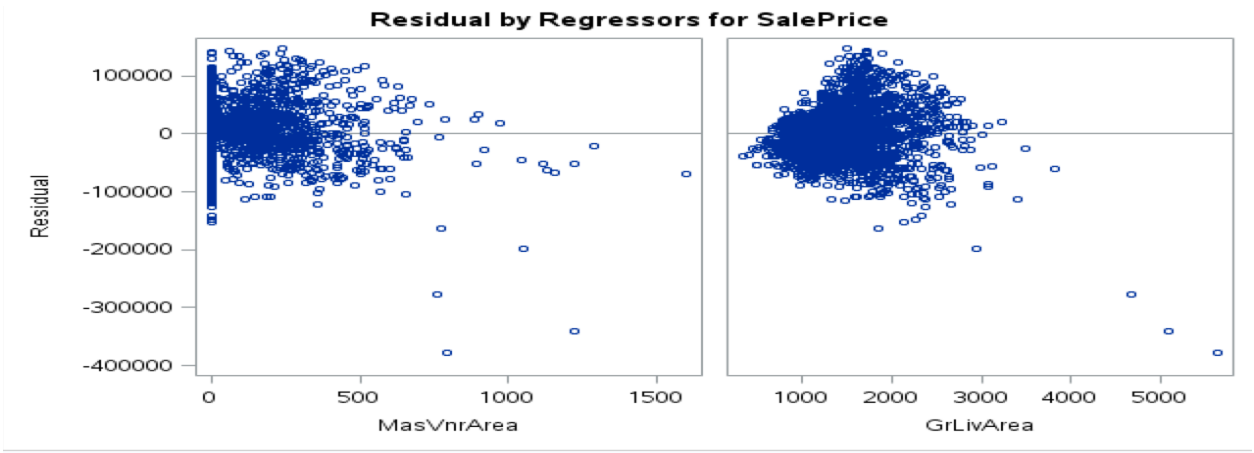
Summary of Forward Selection							
Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	GrLivArea	1	0.4279	0.4279	143.488	2072.16	<.0001
2	MasVnrArea	2	0.0280	0.4558	3.0000	142.49	<.0001

QQ plots looks pretty ok, most of the points are on the line and we could see few point away at the start rest of the points are pretty much on the line, the histogram looks ok as all the points are close to SAS curve, we could see couple of outliers in the cook'd D graph around observations 2000 and 1000

Also in the fit plot most of the points are well within the predictions limits. For the scatter plot apart from few outliers rest of the graph looks ok.







c. I added the variable BsmtFinSF2 into my model since this had the least correlation with the Y. After running the model, I did not see much changes in the model after adding the new variable, the r square value did not change much and also the residual plots and the other plots did not change much.

Variable BsmtFinSF2 Entered: R-Square = 0.4563 and C(p) = 4.0000

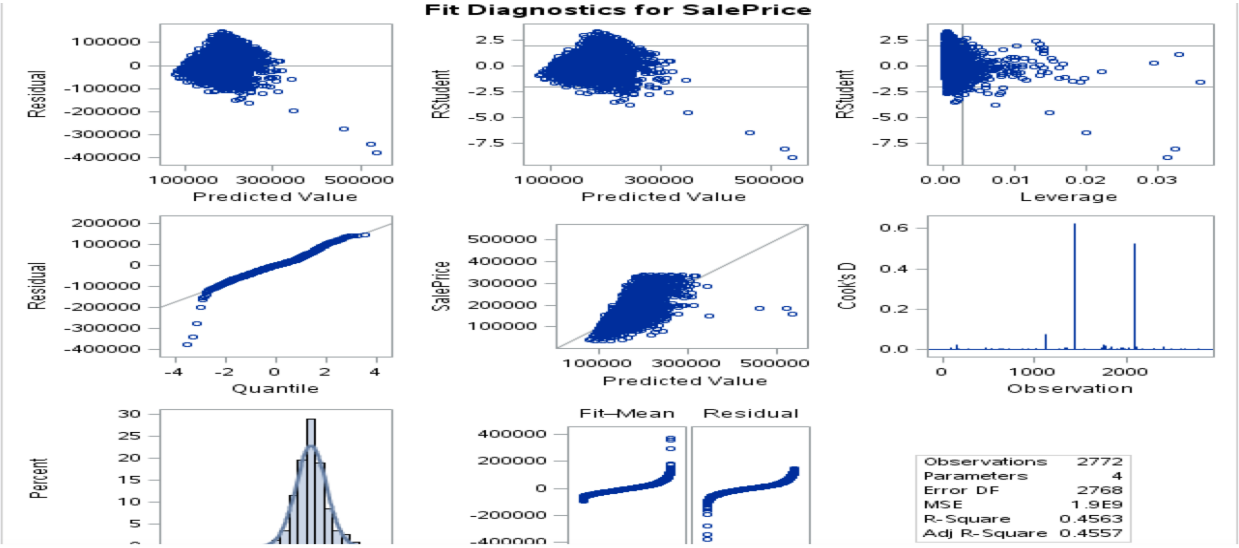
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	4.403361E12	1.467787E12	774.40	<.0001
Error	2768	5.246426E12	1895385252		
Corrected Total	2771	9.649787E12			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	51097	2777.76300	6.413438E11	338.37	<.0001
MasVnrArea	66.53789	5.57117	2.703599E11	142.64	<.0001
GrLivArea	76.77792	1.88485	3.144959E12	1659.27	<.0001
BsmtFinSF2	9.62602	4.91785	7261729789	3.83	0.0504

Bounds on condition number: 1.1056, 9.6335

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	GrLivArea	1	0.4276	0.4276	146.376	2068.99	<.0001
2	MasVnrArea	2	0.0280	0.4556	5.8313	142.40	<.0001
3	BsmtFinSF2	3	0.0008	0.4563	4.0000	3.83	0.0504


Fit Diagnostics for SalePrice
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
SalePrice
Predicted Value
Cook's D
Observation
Percent
Fit–Mean
Residual
Observations 2772
Parameters 4
Error DF 2768
MSE 1.9E9
R-Square 0.4563
Adj R-Square 0.4557

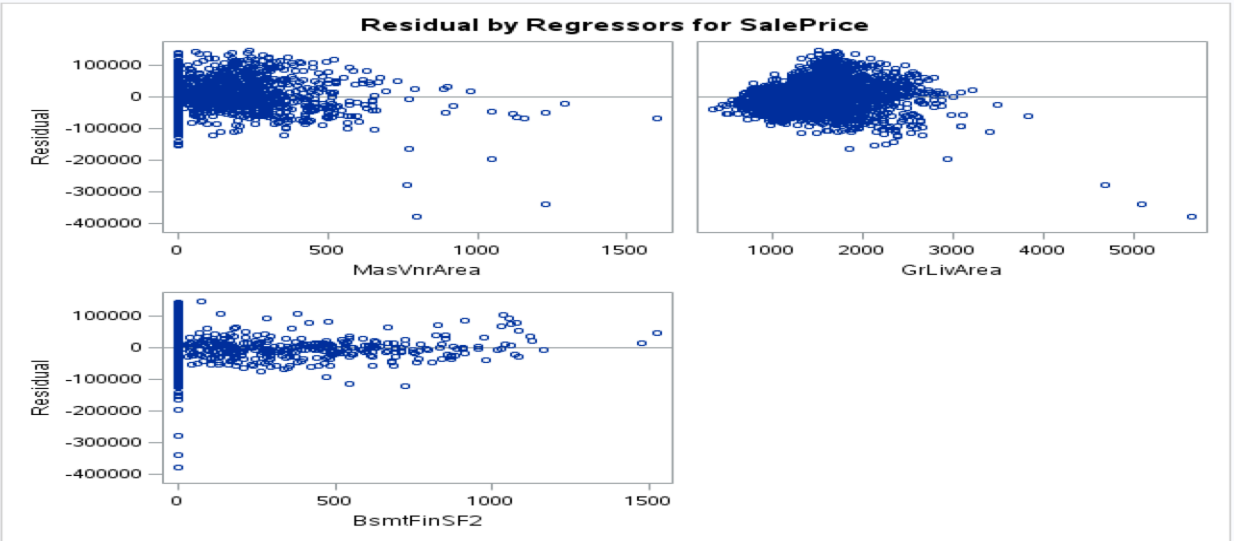

Residual by Regressors for SalePrice
Residual
MasVnrArea
GrLivArea
Residual
BsmtFinSF2

Table comparing the metrics from assignment# and assignment 3

	r-square value	fit-mean/residual	cook's D	qq plot	scatter plot	Histogram
Model1 -assignment #2(saleprice vs grlivarea)	0.5006	fit mean is greater than the residual	three dominant outliers	points are way off the line	few of the points are away from the line	except few all the points are covered by the SAS curve
Model1 -assignment #3(saleprice vs grlivarea)	0.4268	residual is greaterthan the fit mean	two dominant outliers	qq plot looks good all the points are on the line	Points are distributed in a straight line	most of the points are covered by the SAS curve
Model-2 assignment#2 saleprice vs masaarea,grlivarea)	0.559	fit mean is greater than the residual	lot of outliers	most of the points are on the line, bit off end the end	most of the points are on the line except few of the outliers.	most of the points are covered by the SAS curve
Model-2 assignment#3saleprice vs masaarea,grlivarea)	0.4558	fit mean is greater than the residual	two dominant outliers	all the points are on the line	graphs looks good and all the points are distributed in a straight line	All the points are covered by the SAS curve
Model-3 assignment#2	0.5602	fit mean is greater than the residual	lot of outliers	most of the points are on the line, little off at the end	most of the points are on the line except few of the outliers.	most of the points are covered by the SAS curve
Model3-assignment#3	0.4563	fit mean is greater than the residual	two dominant outliers	all the points are on the line, except few at the start	graphs looks good and all the points are distributed in a straight line	All the points are covered by the SAS curve

Conclusion:

Overall in this assignment I have tried to fit simple and multi regression by transforming the variables and in the second part I worked on identifying/deleting the outliers and fitting a model with the new data. it is often useful to do a transformation such as log-transformation for the dependent variable to achieve better normal distribution conformation. Often it is also useful to inspect beta's from the regression to better assess the effect size/real relevance of the results. More commonly, the outlier affects both results and assumptions. it is not legitimate to simply drop the outlier. We may run the analysis both with and without it, but you should state in at least a footnote the dropping of any such data points and how the results changed, the next steps would be validating the models and produce the final model results.

Code:

Paste your code in at the end.

```
libname mydata "/scs/wtm926/" access=readonly;
proc datasets library=mydata;
run;
 quit;

data my_assign;
set mydata.ames_housing_data;

data my_assign1;
set my_assign;
log_saleprice=log(saleprice);
log_grlivarea= log(grlivarea);
keep grlivarea log_grlivarea saleprice log_saleprice;

proc print data=my_assign1 (obs=5);
run;

libname mydata "/scs/wtm926/" access=readonly;
proc datasets library=mydata;
run;
 quit;
```

```
data my_assign;
set mydata.ames_housing_data;

data my_assign1;
set my_assign;
log_saleprice=log(saleprice);
log_grlivarea= log(grlivarea);
keep grlivarea log_grlivarea saleprice log_saleprice;

proc print data=my_assign1 (obs=5);
run;

proc reg data=my_assign1;
 model saleprice=log_grlivarea;
 title 'Model saleprice vs. Log_Grlivarea';
 run;

proc reg data=my_assign1;
 model log_saleprice=grlivearea;
 title 'Model log_saleprice vs. grlivearea';
 run;
```

b. Correlation:

```
data combined;
set my_assign;
```

```
set my_assign1;
run;
proc corr data=combined rank;
var _numeric_ ;
with log_saleprice ;
run;



proc sgscatter data=my_assign1;
plot log_saleprice=log_grlivarea;
run;



data  my_assign2;
  set my_assign1;
    sq_saleprice = saleprice * saleprice;
    sqrt_saleprice = sqrt(saleprice);
run;
proc print data=my_assign2 (obs=5);
run;

proc sgscatter data=my_assign2;
plot (saleprice sq_saleprice sqrt_saleprice) * grlivarea;
run;
```

```
proc reg data=my_assign2;
model sqrt_saleprice = grlivarea;
title 'Model sqrt saleprice vs. grlivarea;
run;
proc univariate NORMAL PLOT DATA=my_assign;
var saleprice;
HISTOGRAM saleprice/NORMAL (color=RED W=5);
run;
data part5;
set my_assign1;
keep saleprice grlivarea price_outlier;
if saleprice = . then delete;

if saleprice <= 21022 then price_outlier=1;
else if saleprice >= 21022 & saleprice < 340570 then price_outlier=2;
else if saleprice >= 340570 then price_outlier=3;

proc print data=part5(obs=20);
run;

proc freq data=part5;
tables price_outlier;
run;
data part6;
set part5;
```

```
if price_outlier =1 then delete;
else if price_outlier =3 then delete;
run;

proc freq data=part6;
tables price_outlier;
run;

proc sort data=part6;
by price_outlier;
run;

ods graphics on;
proc reg data=part6;
model saleprice=grlivarea;
run;
ods graphics on;
proc reg data=part6;
model saleprice = MasVnrArea grlivarea /
selection=forward;

ods graphics on;
proc reg data=part6;
model saleprice = MasVnrArea grlivarea BsmtFinSF2  /
selection=forward;
```