410-57, Assignment #3
Anamitra Bhattacharyya

INTRODUCTION

This assignment specifically deals with building linear regression models (LRMs), applying and comparing transformations on variables, to improve fit, in addition to identifying and removing select outliers. The overall goal of the assignment and project as a whole is to identify suitable predictors of house sale price in Ames, Iowa. The variables are derived from an observational data set from the Ames Assessor's Office used in obtaining values for individual residential properties sold in Ames, Iowa between 2006 and 2010. The transformed LRMs and multiple regression models (MRMs) are characterized separately and compared, as well as evaluated for goodness-of-fi. Furthermore, identification and removal of outliers in the models of house sales price is performed and evaluated using analysis of variance (ANOVA) metrics, such as the F-statistic and p-value, to determine whether they can be further improved.

RESULTS

Part A: Transformations – Comparison of Y versus Log(Y)

1. Transformations of X and Y

For the purposes of this section, various transformations of SalePrice (*e.g.* log and square root) and GrLivArea (*e.g.* Log) were performed. The transformations were appended to the Ames housing data set as shown in the table below.

Obs	BldgType	OverallQual	BsmtFinSF1	CentralAir	GrLivArea	FullBath	BedroomAbvGr	KitchenQual	GarageCars	WoodDeckSF	SalePrice	TotalFlrSF	houseage	LogSalePrice	SqrtSalePrice	LogGrLivArea
1	1Fam	6	639	Y	1656	1	3	TA	2	210	215000	1656	50	12.2784	463.681	7.41216
2	1Fam	5	468	Y	896	1	2	TA	1	140	105000	896	49	11.5617	324.037	6.79794
3	1Fam	6	923	Y	1329	1	3	Gd	1	393	172000	1329	52	12.0552	414.729	7.19218
4	1Fam	7	1065	Y	2110	2	3	Ex	2	0	244000	2110	42	12.4049	493.964	7.65444
5	1Fam	5	791	Y	1629	2	3	TA	2	212	189900	1629	13	12.1543	435.775	7.39572
6	1Fam	6	602	Y	1604	2	3	Gd	2	360	195500	1604	12	12.1833	442.154	7.38026
7	TwnhsE	8	616	Y	1338	2	2	Gd	2	0	213500	1338	9	12.2714	462.061	7.19893
8	TwnhsE	8	263	Y	1280	2	2	Gd	2	0	191500	1280	18	12.1626	437.607	7.15462
9	TwnhsE	8	1180	Y	1616	2	2	Gd	2	237	236500	1616	15	12.3737	486.313	7.38771
10	1Fam	7	0	Y	1804	2	3	Gd	2	140	189000	1804	11	12.1495	434.741	7.49776

Conclusion: Transformations of SalePrice and GrLivArea were created and will be evaluated and used in the following sections of this assignment.

2. Model fitting transformations

Four models were fitted using SalePrice and GrLivArea and the various transformations on predictor and response variables. These models comprised:

a) GrLivArea (X) *versus* SalePrice (untransformed, starting point)
b) GrLivArea (X) *versus* LogSalePrice
c) LogGrLivArea (X) *versus* SalePrice
d) LogGrLivArea (X) *versus* LogSalePrice

Each of these models were systematically analyzed for goodness-of-fit (GOF) and adequacy (see data below).

(a) GrLivArea (X) versus SalePrice (Y)

For the purposes of this section, Y= SalePrice X=GrLivArea (R^2=0.4995). The R-squared for the fitted LRM was approximately 50% (see table below), which is good and suggests that 50% of the variance in SalePrice is explained by GrLivArea. The equation of the fitted, single LRM is described below. In the format y=b0 + b1x1 + e, where x=GrLivArea, y=SalePrice and e = error term.

SalePrice = 13290 + 111GrLivArea, that is,

y = 13290 + 111x

The table of parameter estimates shown below allows one to derive the formula of the linear regression model for SalePrice and GrLivArea (noted above). From the formula slope for the fitted model (above), we interpret this as: for every 1 sq. foot increase in a property GrLivArea there is an increase in SalePrice of $111.

SLR - SalePrice vs GrLivArea

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

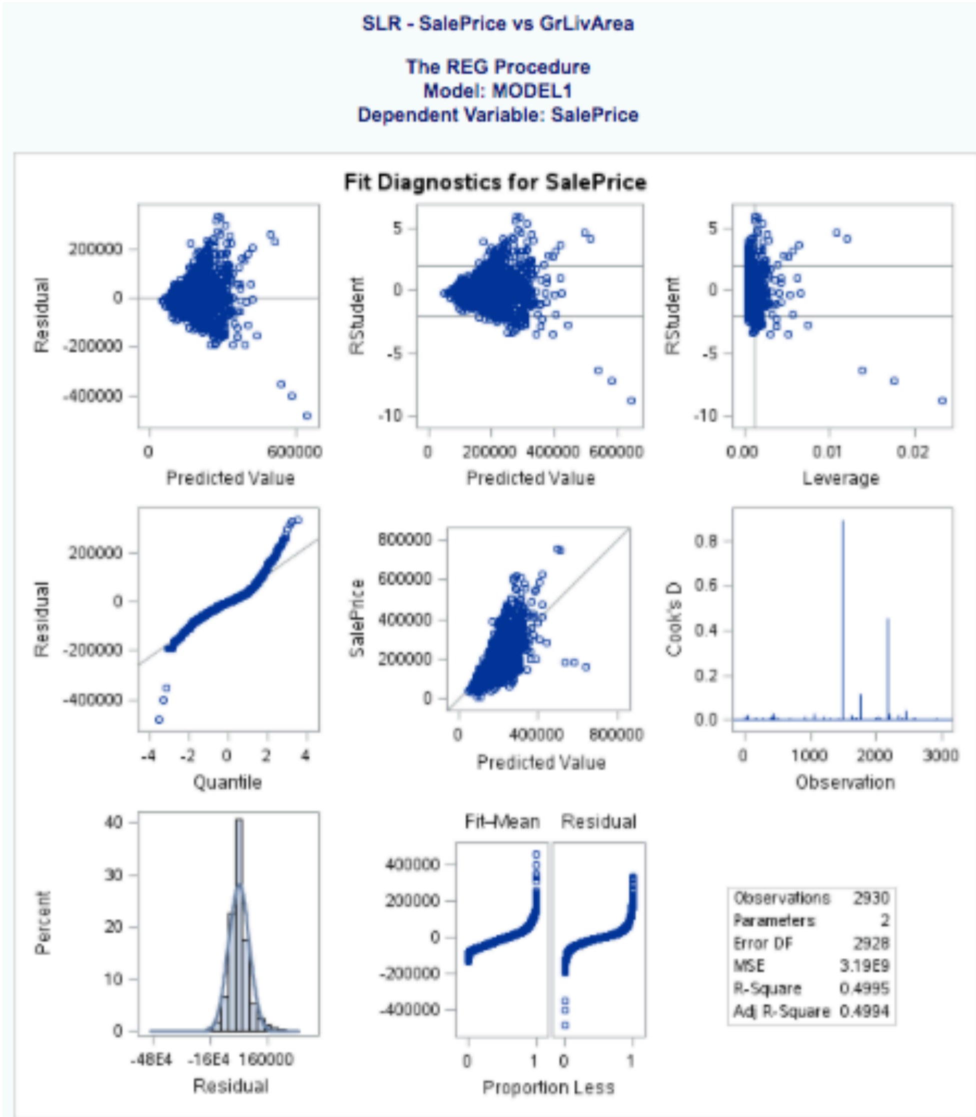
Number of Observations Read	2930
Number of Observations Used	2930

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|	Variance Inflation
Intercept	1	13290	3269.70277	4.06	<.0001	0
GrLivArea	1	111.69400	2.06607	54.06	<.0001	1.00000

From the analysis of variance (ANOVA) table, to assess the quality of the fitted model, the F-statistic is significantly high and the p-value is low, indicating to reject the null hypothesis. This indicates that there is a correlation between SalePrice and GrLivArea. The R-squared value is approximately 0.5, indicating that approximately 50% of the variance in SalePrice is explained by GrLivArea. Reviewing the diagnostics for goodness-of-fit (see graphic below), the quantile-quantile plot of residuals (QQplot; central left plot), shows deviation away from the ideal line, that is predicted to be obtained if it were a normal distribution. Instead there is a deviation of points, especially at the upper end away from normality.



Looking at the residuals (top left panel above) rather than having a random distribution points around the zero level, there is a funnel-shape distribution of residuals. The QQ plot shows (center left panel above) indicating significant deviation away from the normality distribution line, especially at the higher quartile range. The SalePrice against predicted value plot (middle center above) shows a line through the data points, which does not align with the best-fit 45^{o} line.

(b) GrLivArea (X) versus LogSalePrice (Y)

For the purposes of this section, Y= LogSalePrice X=GrLivArea. The R-squared (R^2=0.4842) for the fitted LRM was approximately 48% (see table below), which has decreased from the untransformed data see (a) above (approx. 0.5). The equation of the fitted, single LRM is described below. In the format y=b0 + b1x1 + e, where x=GrLivArea, y=LogSalePrice and e = error term.

LogSalePrice = 11 + 0.00056107GrLivArea, that is,

y = 11 + 0.00056107x

The table of parameter estimates shown below allows one to derive the formula of the linear regression model for LogSalePrice and GrLivArea (noted above). From the formula slope for the fitted model (above), we interpret this as: for every 1 sq. foot increase in a property GrLivArea there is an increase in LogSalePrice of $0.00056107.

SLR - LogSalePrice vs GrLivArea

The REG Procedure
Model: MODEL1
Dependent Variable: LogSalePrice

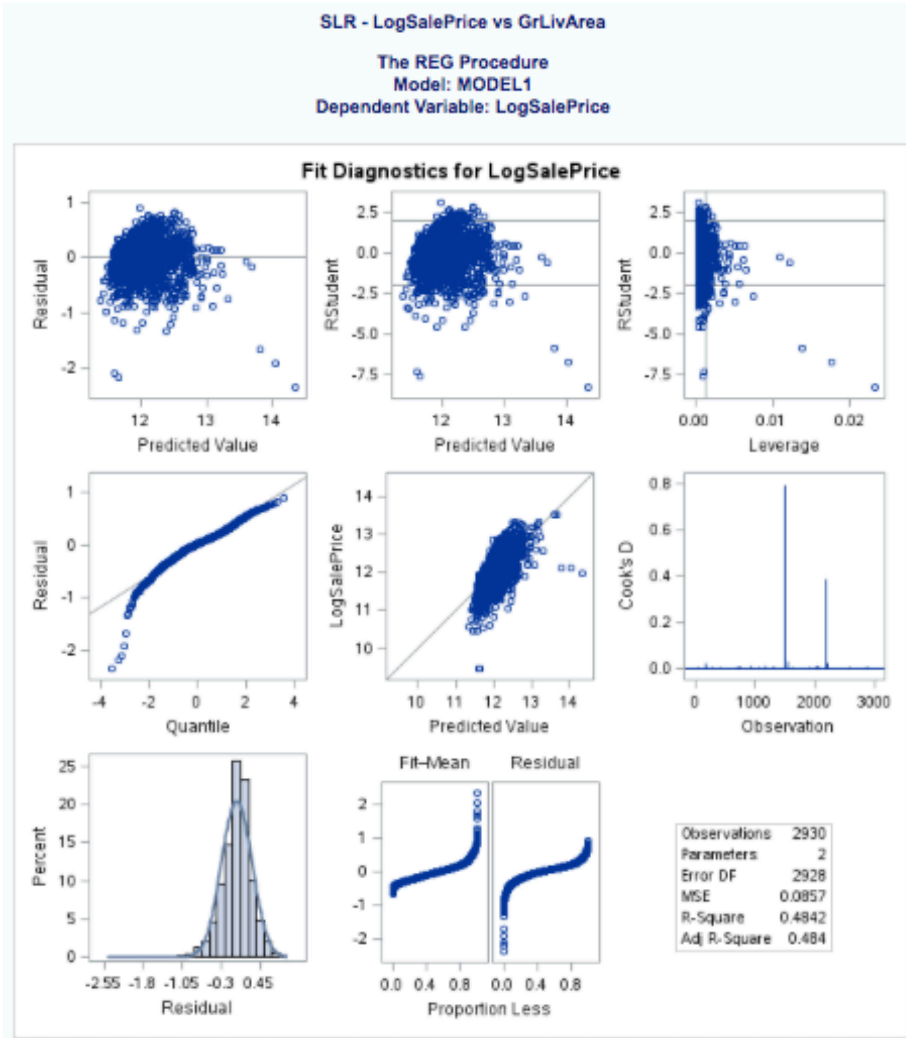
Number of Observations Read	2930
Number of Observations Used	2930

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		

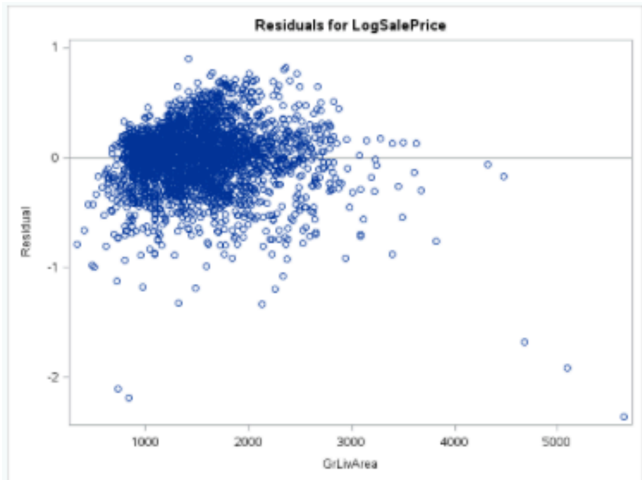
Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	11.17954	0.01694	660.12	<.0001	0
GrLivArea	1	0.00056107	0.00001070	52.43	<.0001	1.00000

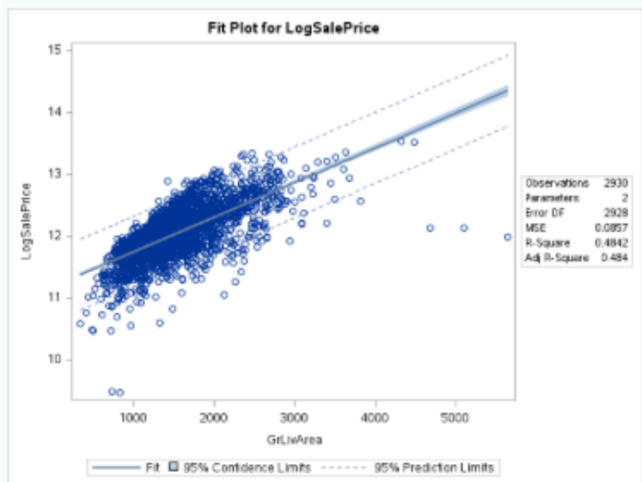
From the analysis of variance (ANOVA) table, to assess the quality of the fitted model, the F-statistic is significantly high and the p-value is low, indicating to reject the null hypothesis. This indicates that there is a correlation between LogSalePrice and GrLivArea. The R-squared value is approximately 0.48, indicating that approximately 48% of the variance in LogSalePrice is explained by GrLivArea. Reviewing the diagnostics for goodness-of-fit (see graphic below), the quantile-quantile plot of residuals (QQplot; central left plot), has improved at the upper end of the distribution but still shows deviation away from the ideal line at the lower range, that is predicted to be obtained if it were a normal distribution.



Looking at the residuals (top left panel above) the funnel-shaped distribution has diminished so there is a more random distribution of points around the zero level. The QQ plot shows (center left panel above) indicating some deviation away from the normality distribution line, especially at the lower quartile range. The LogSalePrice against predicted value plot (middle center above) shows a line through the data points, which has improved its alignment with the best-fit 45^{o} line, but still there is room for improvement.

Inspection of the scatter-plot of LogSalePrice *versus* GrLivArea, there is a greater proportion of data points within the prediction range (see below), though there continues to numerous points outside that range as well.





***(c)* LogGrLivArea (X) *versus* SalePrice**

For the purposes of this section, Y=SalePrice X=LogGrLivArea. The R-squared (R^2=0.4831) for the fitted LRM was approximately 48% (see table below), which has decreased from the untransformed data see (a) above (approx. 0.5) and is the same as the R^2value for (b) LogSalePrice vs GrLivArea. The equation of the fitted, single LRM is described below. In the format y=b0 + b1x1 + e, where x=LogGrLivArea, y=SalePrice and e = error term.

SalePrice = -1060765 + 171011LogGrLivArea, that is,

y = -1060765 + 171011x

The table of parameter estimates shown below allows one to derive the formula of the linear regression model for SalePrice and LogGrLivArea (noted above). From the formula slope for the fitted model (above), we interpret this as: for every 1 unit increase in LogGrLivArea there is an increase in SalePrice of $171,011.

From the analysis of variance (ANOVA) table, to assess the quality of the fitted model, the F-statistic is significantly high and the p-value is low, indicating to reject the null hypothesis. This indicates that there is a correlation between SalePrice and LogGrLivArea. The R-squared value is approximately 0.48, indicating that approximately 48% of the variance in LogSalePrice is explained by LogGrLivArea. Reviewing the diagnostics for goodness-of-fit (see graphic below),

SLR - SalePrice vs LogGrLivArea

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			

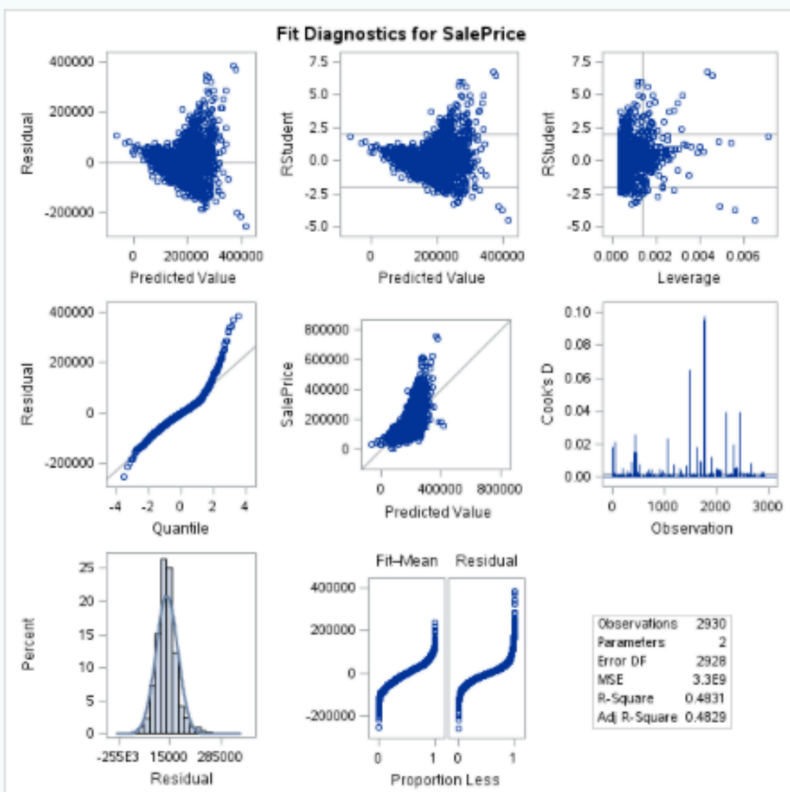
Root MSE	57445	R-Square	0.4831
Dependent Mean	180796	Adj R-Sq	0.4829
Coeff Var	31.77358		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-1060765	23758	-44.65	<.0001	0
LogGrLivArea	1	171011	3269.11261	52.31	<.0001	1.00000

SLR - SalePrice vs LogGrLivArea

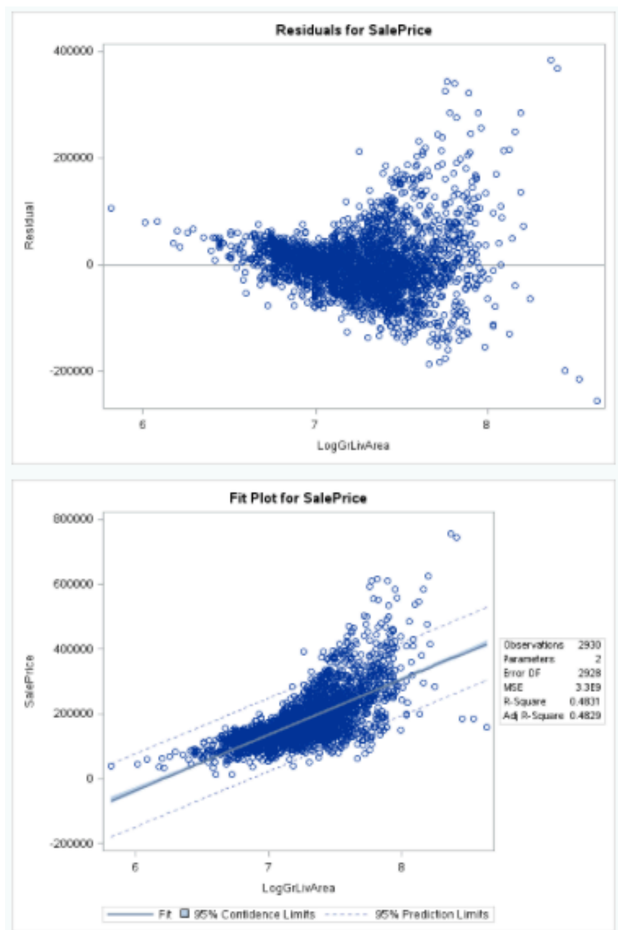
The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice



the quantile-quantile plot of residuals (QQplot; central left plot), deviates away from the normal distribution ideal line in the upper range.

Looking at the residuals (top left panel above and top panel below) there is an unusual U-shaped distribution unlike the more random distribution we saw in (b) above with LogSalePrice *versus* GrLivArea. This suggests a non-random distribution of points around the zero level. The SalePrice against predicted value plot (middle center above) shows a line through the data points, which does not align with the best-fit 45^{o} line.

Inspection of the scatter-plot of SalePrice *versus* LogGrLivArea, there is a funnel shape distribution observed and a greater proportion of data points outside the prediction range (see below), compared to the same plot in (b) above, especially in the upper range of LogGrLivArea.



(d) LogGrLivArea (X) versus LogSalePrice

For the purposes of this section, Y=LogSalePrice X=LogGrLivArea. The R-squared (R^2=0.5230) for the fitted LRM was approximately 52% (see table below), which has increased from the untransformed data see (a) above (approx. 0.5). The equation of the fitted, single LRM is

described below. In the format y=b0 + b1x1 + e, where x=LogGrLivArea, y=LogSalePrice and e = error term.
LogSalePrice = 5 + 0.9LogGrLivArea, that is,
y = 5 + 0.9x

The table of parameter estimates shown below allows one to derive the formula of the linear regression model for SalePrice and LogGrLivArea (noted above). From the formula slope for the fitted model (above), we interpret this as: for every 1 unit increase in LogGrLivArea there is an increase in LogSalePrice of $0.9.

SLR - LogSalePrice vs LogGrLivArea

The REG Procedure
Model: MODEL1
Dependent Variable: LogSalePrice

Number of Observations Read	2930
Number of Observations Used	2930

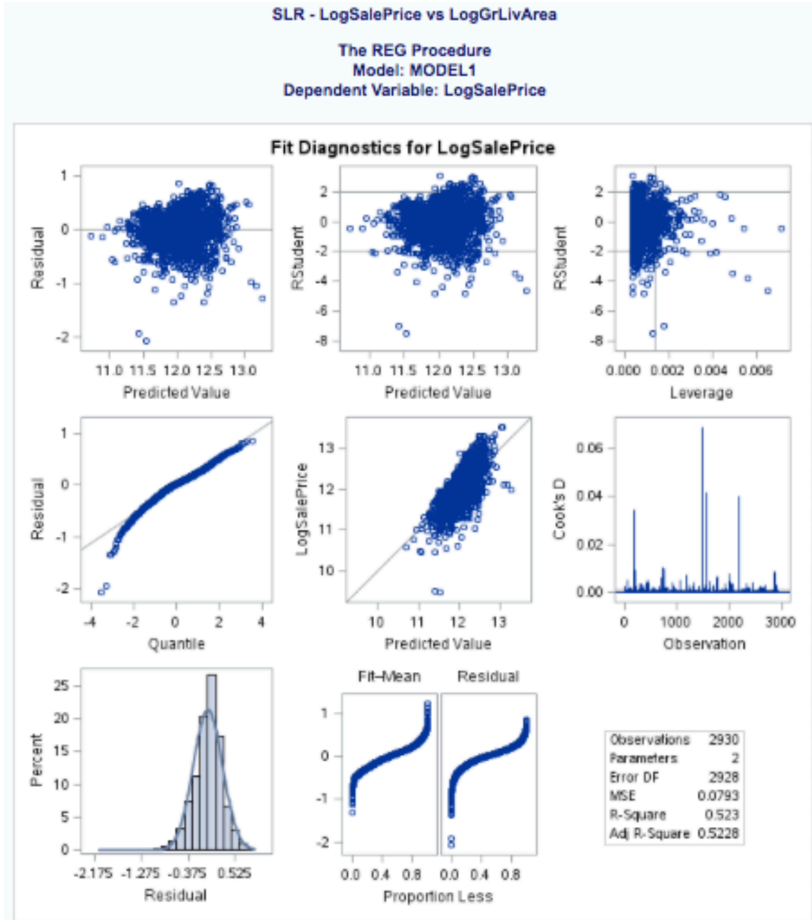
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	254.46967	254.46967	3209.97	<.0001
Error	2928	232.11659	0.07927		
Corrected Total	2929	486.58626			

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|	Variance Inflation
Intercept	1	5.43019	0.11644	46.63	<.0001	0
LogGrLivArea	1	0.90781	0.01602	56.66	<.0001	1.00000

From the analysis of variance (ANOVA) table, to assess the quality of the fitted model, the F-statistic is significantly high and the p-value is low, indicating to reject the null hypothesis. This indicates that there is a correlation between LogSalePrice and LogGrLivArea. The R-squared value is approximately 0.52, indicating that approximately 52% of the variance in LogSalePrice is explained by LogGrLivArea. Reviewing the diagnostics for goodness-of-fit (see graphic below), the quantile-quantile plot of residuals (QQplot; central left plot), there is a good fit but there continues to be deviation away from the normal distribution ideal line in the lower range.

Looking at the residuals (top left panel below) there is no unusual shaped distribution unlike in (a) and (c) above, indicating a more random distribution like we saw in (b) above with LogSalePrice *versus* GrLivArea. This suggests a non-random distribution of residuals around the zero level. The SalePrice against predicted value plot (middle center above) shows a line through the data points, which is much more in alignment with the best-fit 45^{o} line as in (b), LogSalePrice versus GrLivArea.

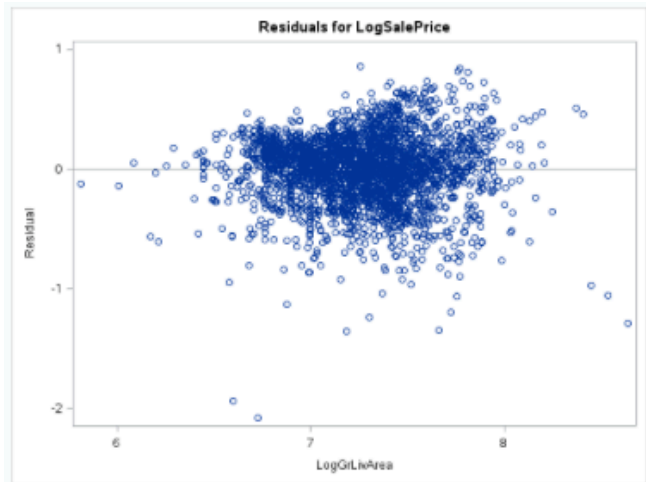


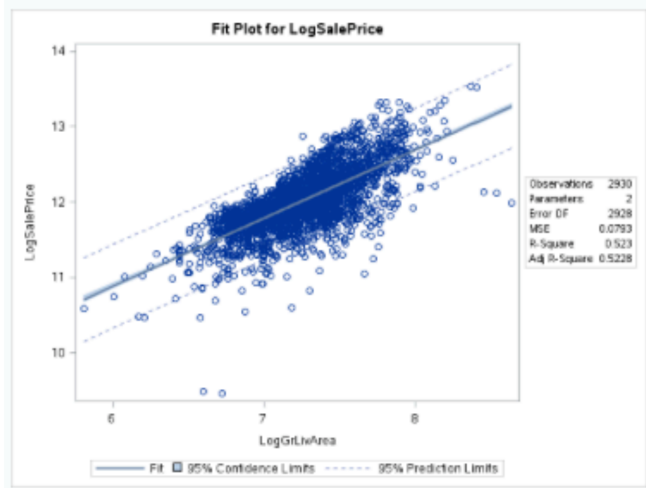
Looking at the residuals in the plot below versus LogGrLivArea shows a relatively random distribution of points either side of the zero point. Inspection of the scatter-plot of LogSalePrice *versus* LogGrLivArea, there is a good distribution of data points within the prediction range (see below), better than (a) and (c), although it is slightly worse than (b) which appeared to have a greater portion of observations with the prediction range compared to (d).

Summary

A summary table of the four models created is shown below for comparison.

Model	**R^2 coefficient metric**
a) GrLivArea (X) *versus* SalePrice	0.4995
b) GrLivArea (X) *versus* LogSalePrice	0.4842
c) LogGrLivArea (X) *versus* SalePrice	0.4831
d) LogGrLivArea (X) *versus* LogSalePrice	0.5230





Conclusion: Model (d) based on the R-squared metrics scores highest (0.52) of the four models, and is the best fitting model. Based on the scatter plot of residuals, models (b) and (d) are best, though the only reservation with (d) is that there appears to be more observations outside the prediction range in (d) compared to (b), where the observations are more tightly clustered at the lower range. One of the issues with using a transformed variable(s) is they have been transformed and are thus more difficult to interpret back to reality, they are an abstraction from reality. So for instance, while the original untransformed model in (a) is unaltered it is simple to interpret the effect of a change in the predictor variable on the real-world response variable - sale price. In contrast, when using the Log of the SalePrice one always has to remember to reverse translate the log value (log Y) to Y to derive a meaningful result about SalePrice.

3. Best Predictor with Transformations of SalePrice

The continuous variables used in this part of the assignment were WoodDeckSF, GrLivArea, BsmtFinSF1 and OverallQual. These variables were use to develop correlations with transformations of the response variable – SalePrice, specifically: SalePrice, LogSalePrice and SqrtSalePrice. The results of this analysis are shown below.

Correlation of Various Continuous Variables against Transformations of SalePrice

The CORR Procedure

4 With Variables:	WoodDeckSF GrLivArea BsmtFinSF1 OverallQual
3 Variables:	SalePrice LogSalePrice SqrtSalePrice

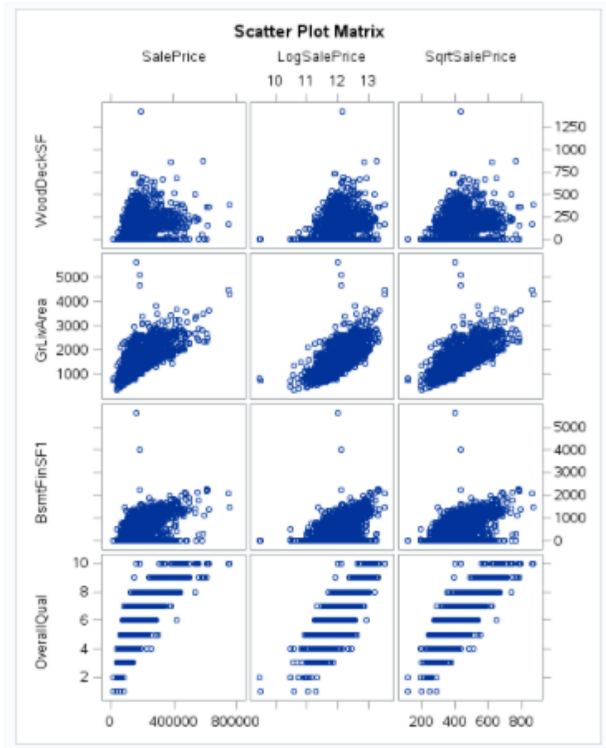
Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
WoodDeckSF	2930	93.75188	126.36156	274693	0	1424
GrLivArea	2930	1500	505.50889	4394093	334.00000	5642
BsmtFinSF1	2929	442.62957	455.59084	1296462	0	5644
OverallQual	2930	6.09488	1.41103	17858	1.00000	10.00000
SalePrice	2930	180796	79887	529732456	12789	755000
LogSalePrice	2930	12.02097	0.40759	35221	9.45634	13.53447
SqrtSalePrice	2930	416.26208	86.74391	1219648	113.08846	868.90736

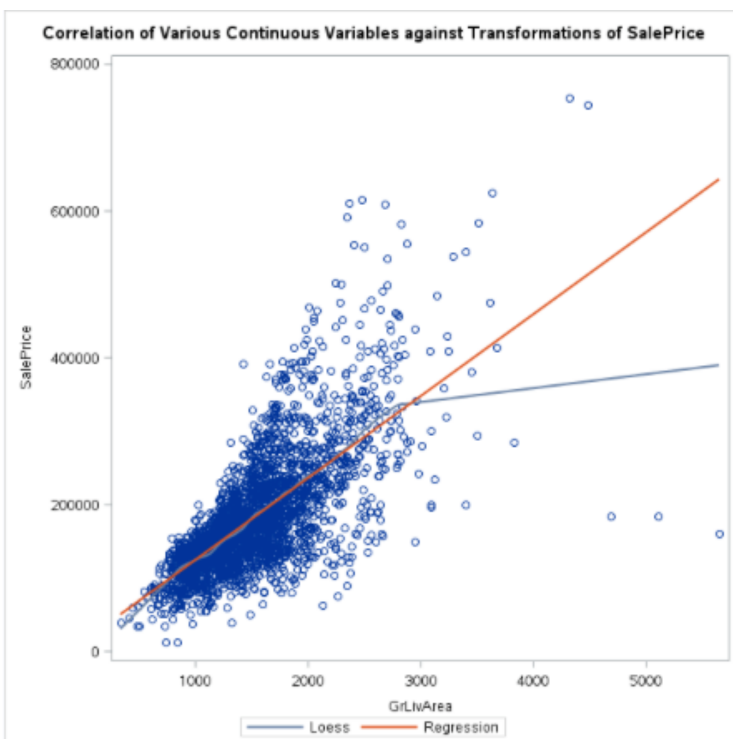
Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	SalePrice	LogSalePrice	SqrtSalePrice
WoodDeckSF	0.32714 <.0001 2930	0.33332 <.0001 2930	0.33566 <.0001 2930
GrLivArea	0.70678 <.0001 2930	0.69586 <.0001 2930	0.71240 <.0001 2930
BsmtFinSF1	0.43291 <.0001 2929	0.41080 <.0001 2929	0.42719 <.0001 2929
OverallQual	0.79926 <.0001 2930	0.82564 <.0001 2930	0.82460 <.0001 2930

For each of the predictor variables (rows) in the table above, the R-squared value stays approximately the same irrespective of the SalePrice or LogSalePrice (transformed). The variables GrLivArea and OverallQual score most highly in terms of R-squared. For GrLivArea the R-squared using LogSalePrice goes down fractionally from 0.70678 (untransformed) to 0.69586 (LogSalePrice). For OverallQual the R-squared value increases with the log transformation 0.79926 (untransformed) to 0.82564. A scatter plot matrix for all these variables is shown below; note that the GrLivArea scatter shows a strong correlation and the scatter is less pronounced compared to WoodDeckSF and BsmtFinSF1.

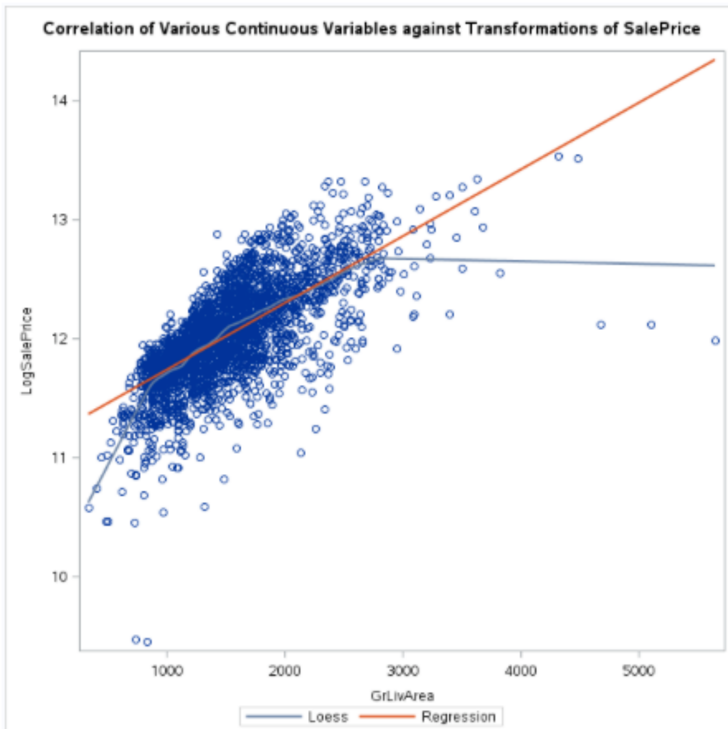


A scatterplot matrix using one of the highly scoring continuous variables (*e.g.* GrLivArea) and transformations of SalePrice (*e.g.* SalesPrice, LogSalePrice, SqrtSalePrice) were employed for this analysis. The results are shown in the graphics below.



For the untransformed plot above there is a good correlation of variables until approximately x=2,600 for GrLivArea, with the fitted line approximately passing through the origin. The linear model (brown line) and LOESS (blue line) are collinear below x=2,600. Above x=2,600 there are outliers that act to distort the best fit line as shown by the LOESS line (blue) diverging at an acute angle.

For the correlation plot with LogSalePrice *versus* GrLivArea there is a similar effect, except that the y-intercept is pushed up further and the point of divergence of the linear fitted model and LOESS is at approximately x=2,400. The observations are also more tightly clustered with the linear fitted model (brown line). Also it should be noted that the LOESS line diverges not only at the higher ranges of x but also at the lower end.



Conclusion: The GrLivArea is an example of a continuous variable that correlates strongly with both SalePrice and LogSalePrice in terms of R-squared specifically 0.71 and 0.69, respectively. The plot of the LogSalePrice *versus* GrLivArea produces a much more clustered set of observations, though above x=2,400 there is a divergence in the LOESS line due to the presence of potential outlier observations at the higher ranges of GrLivArea, that can affect the trajectory of the fitted model.

4. Other transformations

From the analysis performed in part 3 above, the log transformation improves response variable (*e.g.* SalePrice) when there is a lot of variability or scatter, and acts to tighten the spread of the observations. Also from the analysis performed, we only know that the fit is improved in a simple linear regression model (LRM), we have not evaluated it in a multiple regression model (MRM) yet.

Another transformation was chosen to transform SalePrice specifically, square root of SalePrice (SqrtSalePrice). The correlation with GrLivArea and SalePrice, LogSalePrice and SqrtSalePrice is shown and compared below. The R-squared value improves slightly for SqrtSalePrice (0.71240) over LogSalePrice (0.69586) and SalePrice (0.70678). The scatter plots of the observations of GrLivArea *versus* the untransformed and transformed SalePrice are similar to each other.

Correlation of GrLivArea against SqrtSalePrice

The CORR Procedure

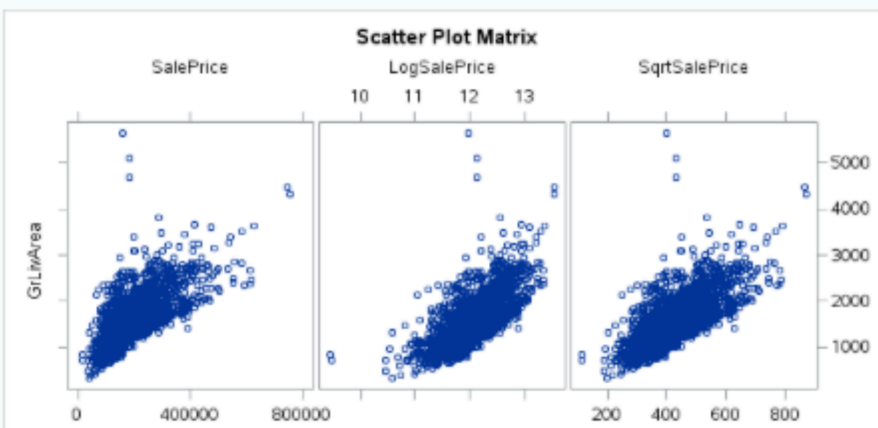
1 With Variables:	GrLivArea
3 Variables:	SalePrice LogSalePrice SqrtSalePrice

Simple Statistics

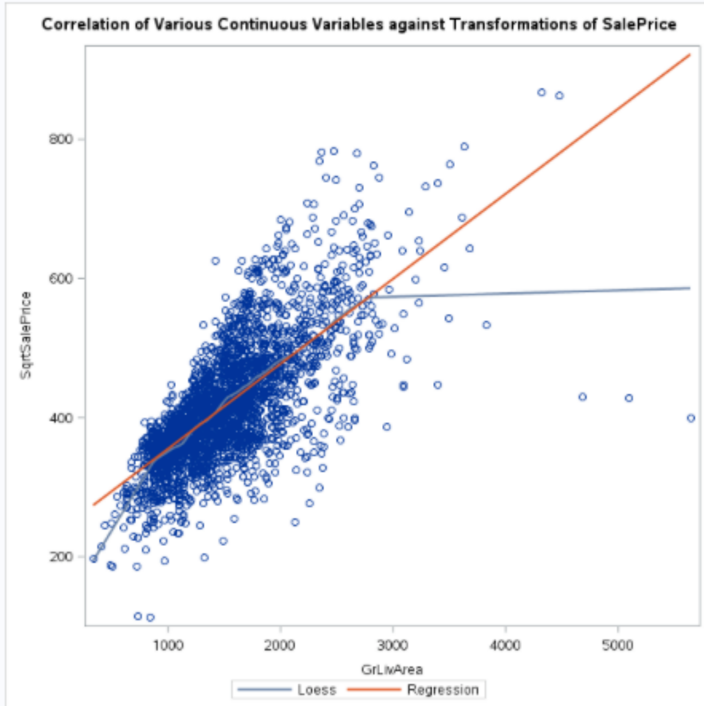
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
GrLivArea	2930	1500	505.50889	4394093	334.00000	5642
SalePrice	2930	180796	79887	529732456	12789	755000
LogSalePrice	2930	12.02097	0.40759	35221	9.45634	13.53447
SqrtSalePrice	2930	416.26208	86.74391	1219648	113.08846	868.90736

Pearson Correlation Coefficients, N = 2930
Prob > |r| under H0: Rho=0

	SalePrice	LogSalePrice	SqrtSalePrice
GrLivArea	0.70678 <.0001	0.69586 <.0001	0.71240 <.0001



The scatter plot of SqrtSalePrice against GrLivArea has a strong correlation (R^2=0.71240), see plot below. The scatter below illustrates that the LRM fitted line (brown) and the LOESS (blue) lines diverge at a value of GrLivArea of approximately 2,600 sq. ft. due to outliers.



The equation of the fitted, single LRM is described below. In the format y=b0 + b1x1 + e, where x=GrLivArea, y=SqrtSalePrice and e = error term.
SqrtSalePrice = 232 + 0.1GrLivArea, that is,
Sqrt(y) = 232 + 0.1x

SLR - Sqrtsaleprice versus GrLivArea

The REG Procedure
Model: MODEL1
Dependent Variable: SqrtSalePrice

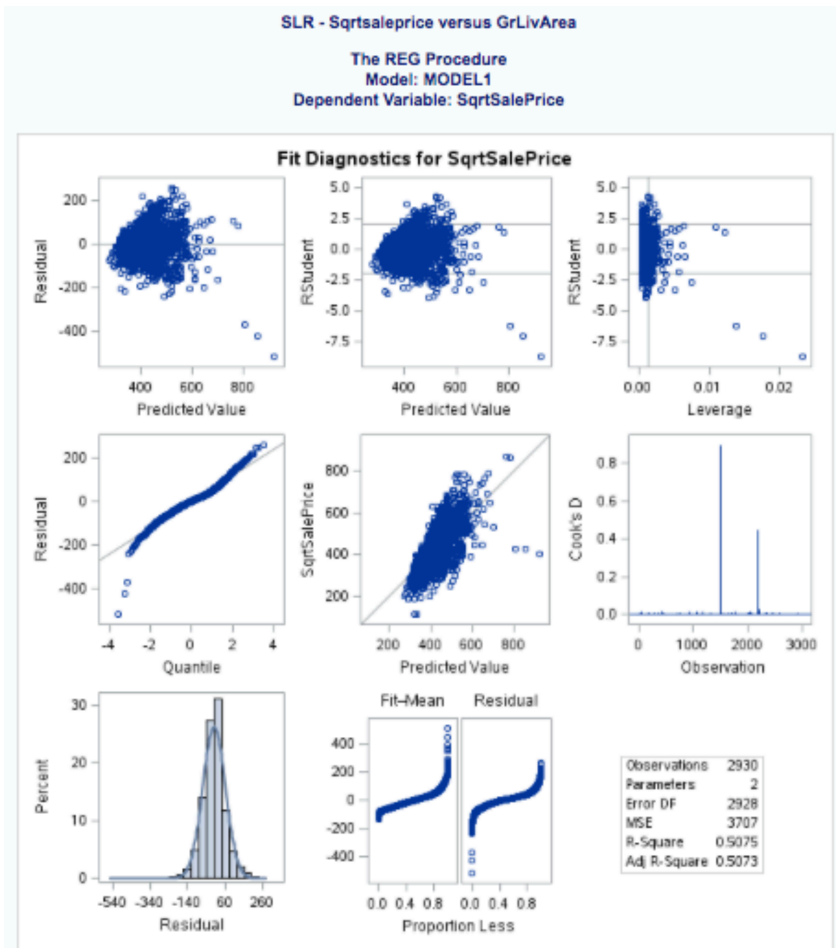
Number of Observations Read	2930
Number of Observations Used	2930

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|	Variance Inflation
Intercept	1	232.93209	3.52198	66.14	<.0001	0
GrLivArea	1	0.12225	0.00223	54.93	<.0001	1.00000

The table of parameter estimates shown above allows one to derive the formula of the linear regression model for SqrtSalePrice and GrLivArea (noted above). From the formula slope for the fitted model (above), we interpret this as: for every 1 unit increase in GrLivArea there is an increase in SqrtSalePrice of $0.1. From the analysis of variance (ANOVA) table, to assess the quality of the fitted model, the F-statistic is significantly high and the p-value is low, indicating to reject the null hypothesis. This indicates that there is a correlation between SqrtSalePrice and GrLivArea. The R-squared value is approximately 0.5075, indicating that approximately 51% of the variance in SqrtSalePrice is explained by GrLivArea. Reviewing the diagnostics for goodness-of-fit (see graphic below), the quantile-quantile plot of residuals (QQplot; central left plot), there is a good fit but there continues to be deviation away from the normal distribution ideal line in the upper and especially lower ranges.



Looking at the residuals (top left panel above) there is a little funnel-shaped distribution indicating less than random distribution The SqrtSalePrice against predicted value plot (middle center above) shows a line through the data points, which is not alignment with the best-fit 45^{o} line.

Another regression model was fitted using SqrtSalePrice and SqrtGrLivArea. The equation of the fitted, single LRM is described below. In the format y=b0 + b1x1 + e, where x=SqrtGrLivArea, y=SqrtSalePrice and e = error term.
SqrtSalePrice = 33 + 10SqrtGrLivArea, that is,

Sqrt(y) = 232 + 0.1Sqrt(x)

The table of parameter estimates shown above allows one to derive the formula of the linear regression model for SqrtSalePrice and SqrtGrLivArea (noted above). From the formula slope for the fitted model (above), we interpret this as: for every 1 unit increase in SqrtGrLivArea there is an increase in SqrtSalePrice of $10. From the analysis of variance (ANOVA) table, to assess the quality of the fitted model, the F-statistic is significantly high and the p-value is low, indicating to reject the null hypothesis. This indicates that there is a correlation between SqrtSalePrice and SqrtGrLivArea. The R-squared value is 0.5223, indicating that approximately 52% of the variance in SqrtSalePrice is explained by SqrtGrLivArea.

SLR - Sqrtsaleprice versus SqrtGrLivArea

The REG Procedure
Model: MODEL1
Dependent Variable: SqrtSalePrice

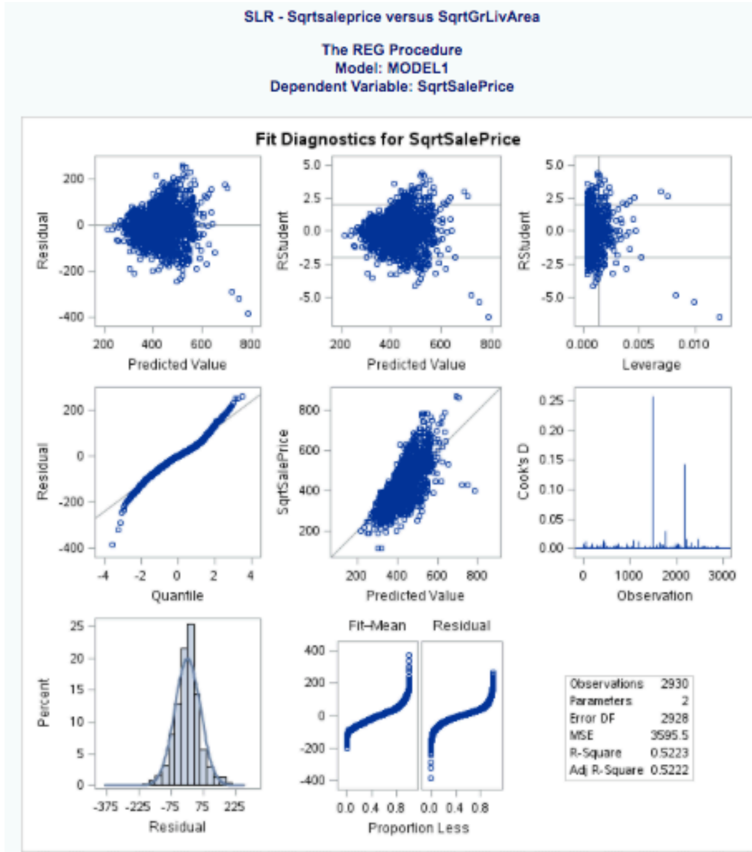
Number of Observations Read	2930
Number of Observations Used	2930

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	11511534	11511534	3201.61	<.0001
Error	2928	10527745	3595.54129		
Corrected Total	2929	22039279			

Root MSE	59.96283	R-Square	0.5223
Dependent Mean	416.26208	Adj R-Sq	0.5222
Coeff Var	14.40507		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	33.80285	6.84946	4.94	<.0001	0
SqrtGrLivArea	1	10.00783	0.17687	56.58	<.0001	1.00000

Reviewing the diagnostics for goodness-of-fit (see graphic below), the QQplot of residuals (central left plot), there is a reasonable fit but there continues to be deviation away from the normal distribution ideal line in the upper and lower ranges. Looking at the residuals (top left panel below) there is a little funnel-shaped distribution indicating less than random distribution The SqrtSalePrice against predicted value plot (middle center above) shows a line through the data points, which is not aligned with the best-fit 45^{o} line through the origin.



Conclusion: The fitted models of SqrtSalePrice *versus* GrLivArea (R^2=0.5075) and SqrtSalePrice *versus* SqrtGrLivArea (R^2=0.5223) are comparable, though the latter is a slightly better correlation. Both the latter models have slight funnel-shaped distribution of residuals, which is absent in the model for LogSalePrice *versus* LogGrLivArea (R^2=0.5228), suggesting the log-log model is perhaps a better model from several perspectives.

Part B: Outliers

5. Identifying outliers

I used a the 2 standard deviation rule to filter out outliers, which is that 95% of the data under a normal distribution curve resides under (- 2σ from the mean)) from the mean. The mean for the SalePrice is $180,000. Since sigma is ~$80,000, then 2σ value is $160,000. The lower range would be properties less than $20,000 (-2σ from the mean) while upper range would be above $340,000 (+2σ from the mean). The counts of outlier by this definition is noted below:

Criteria	Range	Count
SalePrice < $20,000	- 2σ	2 (removed)
SalePrice > $340,000	+ 2σ	134 (removed)
SalePrice within range		2794

These criteria for outliers were also corroborated by the EDA that was performed in Assignment 1 earlier.

Conclusion: Using the 2σ rule, it is possible to filter out outliers from the lower and upper SalePrice ranges, such that we are left with 2794 properties for consideration in further analysis. This results in 136 observations being removed for further analysis.

6. Removing outliers

Using the cleaned data to re-fit models 2, 5 and 6 (from assignment 2), again OverallQual is the best fit as shown below. Also the table confirms that number of observations in SalePrice has been reduced to 2794 from 2930.

(a) Model #2 (from assignment 2)

Model 2 - SLR with best R-Squared

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice
R-Square Selection Method

Number of Observations Read	2794
Number of Observations Used	2793
Number of Observations with Missing Values	1

Number in Model	R-Square	AIC	BIC	MSE	Parameter Estimates Intercept	TotalFlrSF	GrLivArea	houseage	OverallQual	FullBath	WoodDeckSF	BsmtFinSF1
1	0.6175	58687.8081	58687.8159	1334398957	-43696	.	.	.	35691	.	.	.

Model 2 - SLR with best R-Squared

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Number of Observations Read	2794
Number of Observations Used	2793
Number of Observations with Missing Values	1

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	6.01201E12	6.01201E12	4505.41	<.0001
Error	2791	3.724307E12	1334398957		
Corrected Total	2792	9.736317E12			

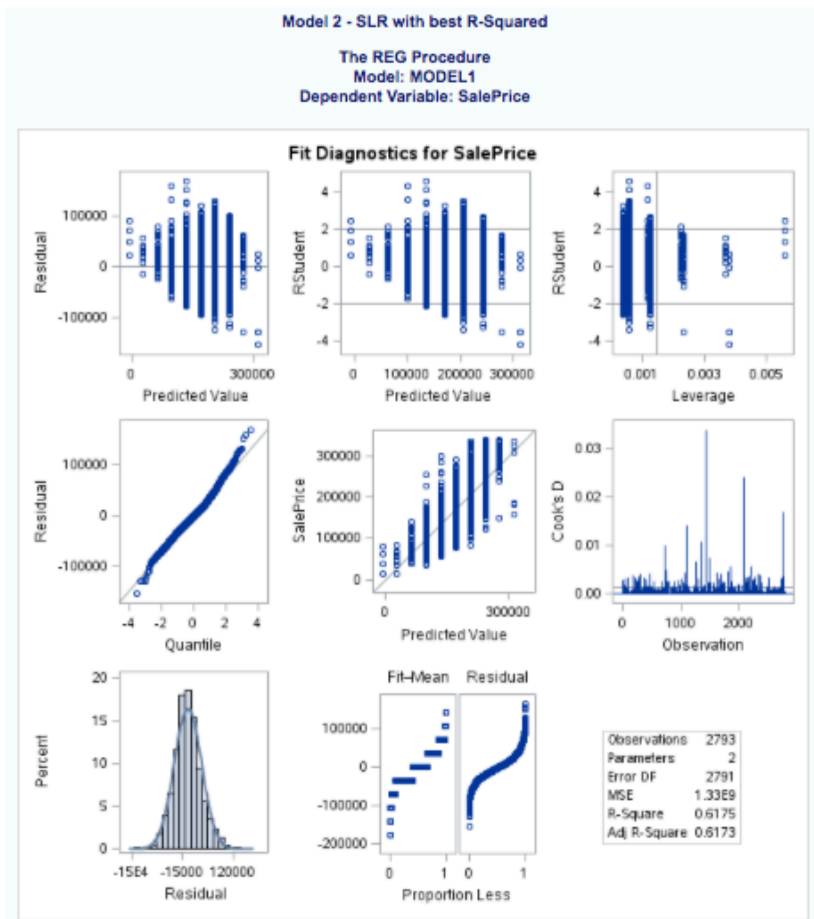
Root MSE	36529	R-Square	0.6175
Dependent Mean	169209	Adj R-Sq	0.6173
Coeff Var	21.58836		

Parameter Estimates

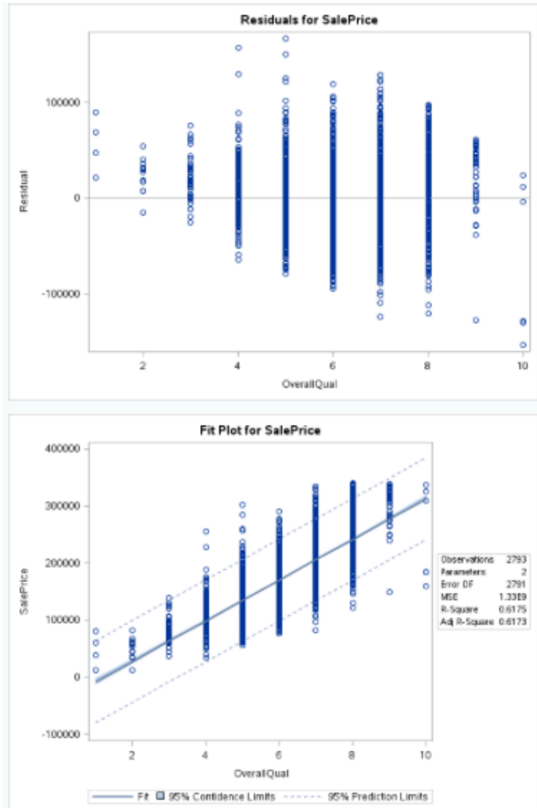
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-43696	3246.33484	-13.46	<.0001	0
OverallQual	1	35691	531.72712	67.12	<.0001	1.00000

From the parameter estimates table above the equation for the cleaned fitted model is:
SalePrice(Y)= -43696 + 35691OverallQual(X)
y = -43696 + 35691x

So for every unit increase in OverallQual, the SalePrice increases by $35,691. From the analysis of variance (ANOVA) table, to assess the quality of the fitted model, the F-statistic is significantly high and the p-value is low, indicating to reject the null hypothesis. This indicates that there is a correlation between SalePrice and OverallQual with the cleaned data. The R-squared value is 0.6175, indicating that approximately 61% of the variance in SalePrice is explained by OverallQual. Reviewing the diagnostics for goodness-of-fit (see graphic below), the QQplot of residuals (central left plot), there is a reasonable fit but there continues to be deviation away from the normal distribution ideal line in the upper and lower ranges but less so than previously. Looking at the residuals (top left panel below) there is a little funnel-shaped distribution indicating less than random distribution The cleaned SalePrice against predicted value plot (middle center above) shows a line through the data points, which is quite well aligned with the best-fit 45^{o} line through the origin.



The fit plot below (lower panel) still shows that there continues to be observations outside the prediction limits. Interestingly, there continue to be outliers in the middle ranges especially.



(b) Model #5 (from assignment 2)

Model 5 - MLR with FullBath and OverallQual

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	2	6.407911E12	3.203956E12	2680.09	<.0001
Error	2791	3.336541E12	1195464248		
Corrected Total	2793	9.744452E12			

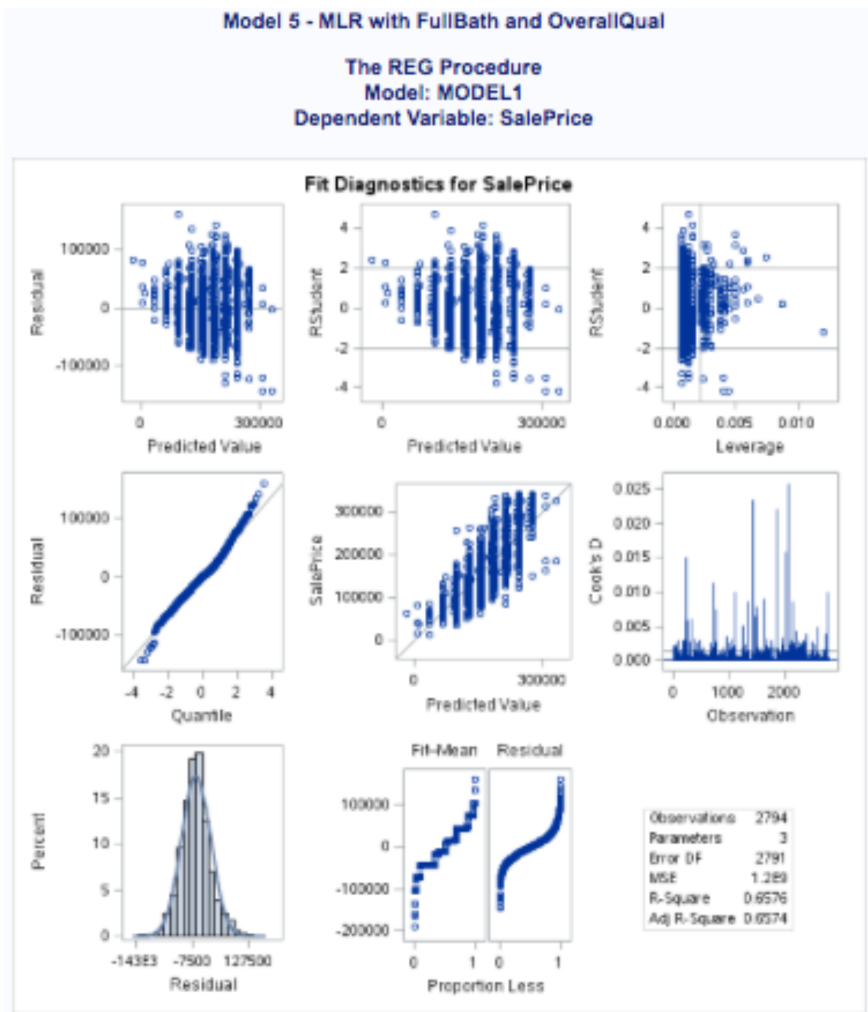
Root MSE	34575	R-Square	0.6576
Dependent Mean	169177	Adj R-Sq	0.6574
Coeff Var	20.43751		

Parameter Estimates

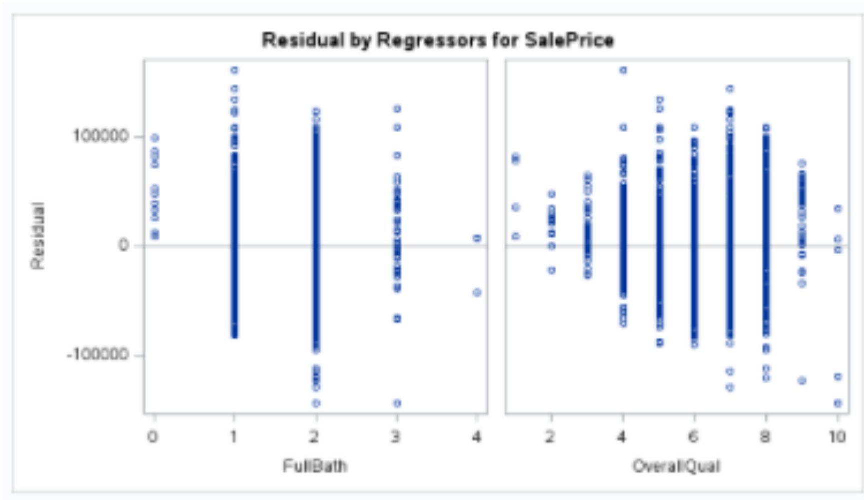
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-51223	3098.97319	-16.53	<.0001	0
FullBath	1	25075	1391.57442	18.02	<.0001	1.33167
OverallQual	1	30478	580.54324	52.50	<.0001	1.33167

From the parameter estimates table above the equation for the cleaned fitted model is:
SalePrice(Y)= -51223 + 25075FullBath(x1) + 30478OverallQual(x2)
y = -51223 + 25075x1 + 30478x2

So for every unit increase in FullBath and OverallQual, the SalePrice increases by $25,075 and $30,478, respectively. From the analysis of variance (ANOVA) table above, to assess the quality of the fitted model, the F-statistic is significantly high and the p-value is low, indicating to reject the null hypothesis. This indicates that there is a correlation between SalePrice and OverallQual with the cleaned data. The R-squared value is 0.6576, FullBath and OverallQual explain approximately 65% of the variance in SalePrice. Reviewing the diagnostics for goodness-of-fit (see graphic below), the QQplot of residuals (central left plot), there is a reasonable fit but there continues to be deviation away from the normal distribution ideal line in the upper and lower ranges. Looking at the residuals (top left panel below) there is a little funnel-shaped distribution indicating less than random distribution The cleaned SalePrice against predicted value plot (middle center above) shows a line through the data points, which is quite well aligned with the best-fit 45^{o} line through the origin, though there are still outliers present.



The fit plot below (lower panel) for residuals still shows a funnel-shape for the OverallQual suggesting some non-randomness, and non-random distribution of residuals at the upper and lower ranges for FullBath.



(c) Model #6 (from assignment 2)

Model 6 - MLR with FullBath, OverallQual and BedroomAbvGr

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Number of Observations Read	2794
Number of Observations Used	2794

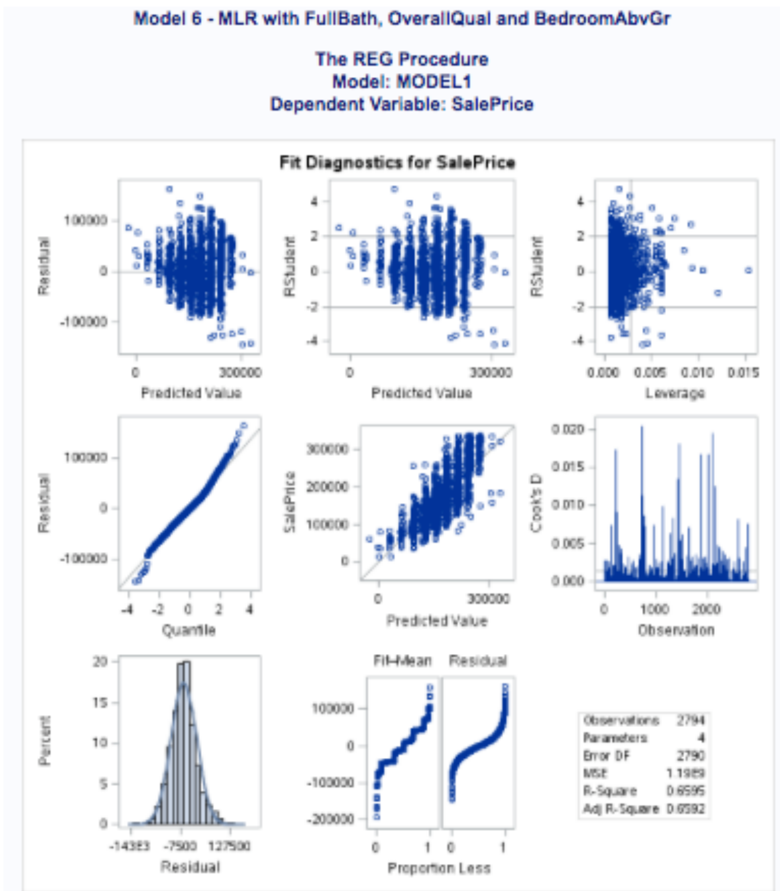
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	6.426822E12	2.142274E12	1801.57	<.0001
Error	2790	3.317631E12	1189114882		
Corrected Total	2793	9.744452E12			

Root MSE	34484	R-Square	0.6595
Dependent Mean	169177	Adj R-Sq	0.6592
Coeff Var	20.38317		

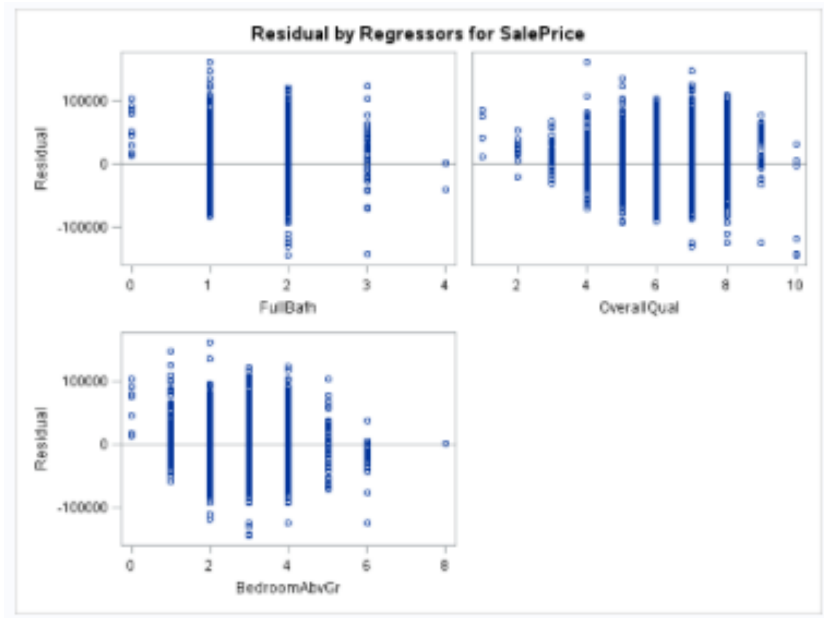
Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-59598	3736.75399	-15.95	<.0001	0
FullBath	1	22884	1492.68697	15.33	<.0001	1.54040
OverallQual	1	30811	584.98757	52.67	<.0001	1.35935
BedroomAbvGr	1	3430.79908	860.31836	3.99	<.0001	1.16054

From the parameter estimates table above the equation for the cleaned fitted model is:
SalePrice(Y)= -59598 + 22884FullBath(x1) + 30811OverallQual(x2) + 3430BedroonAbvGr(x3)
y = -59598 + 22884x1)+ 30811x2 + 3430x3

So for every unit increase in FullBath, OverallQual and BedroomAbvGr the SalePrice increases by \$22,884, \$30,811 and \$3,430, respectively. From the analysis of variance (ANOVA) table above, to assess the quality of the fitted model, the F-statistic is significantly high and the p-value is low, indicating to reject the null hypothesis. This indicates that there is a correlation between SalePrice and OverallQual with the cleaned data. The R-squared value is 0.6595, FullBath, OverallQual and BedroomAbvGr explain approximately 65% of the variance in SalePrice. Reviewing the diagnostics for goodness-of-fit (see graphic below), the QQplot of residuals (central left plot), there is a reasonable fit but there continues to be deviation away from the normal distribution ideal line in the upper and lower ranges. Looking at the residuals (top left panel below) there is a little funnel-shaped distribution indicating less than random distribution. The cleaned SalePrice against predicted value plot (middle center above) shows a line through the data points, which is quite well aligned with the best-fit 45° line through the origin, though there are still outliers present especially at the upper range.



The fit plot below (lower panel) for residuals shows an elliptical shape for the OverallQual suggesting some non-randomness, and non-random distribution of residuals at the upper and lower ranges for FullBath and similarly for BedroonAbvGr.



Summary
Using the cleaned data for SalePrice that has been refit to the models in assignment 2 a summative table of the R-squared coefficients is shown below for comparison.

Model	**R^2 coefficient**	
	Assignment 2	**Assignment 3**
(2) LRM Best variable (GrLivArea)	0.6386	0.6175
(5) MLR with FullBath and OverallQual	0.6614	0.6576
(6) MLR with FullBath OverallQual BedroomAbvGr	0.6629	0.6595

Conclusion: While the R-squared coefficients for the models using the cleaned data are comparable (though fractionally less than from assignment 2, there are improvement in the distribution of the QQ plots using the cleaned versus not cleaned data. It is likely that additional cleaning of the SalePrice needs to be performed, but this is also likely for the predictor variables as well.

7. Model-based outliers (Influential outliers)
After removing 'outliers' for SalePrice, there are still unusually large residuals observed in the residual plots earlier. This is in part since we only removed some outliers in SalePrice but we did not remove any with the other predictor variables. These are due to the 'influential' points, which exert a disproportionate affect on the model coefficients. These points are identified by several statistics such as DFFITS. The DFFITS statistic was used to identify and remove these influential points and re-fit the model. The results from before and after refitting the model are shown below.

A total of 147 observations were removed after cleaning further with DFFITS.

Before refitting (2794 observations):

Three variable Regression Model with log transformation

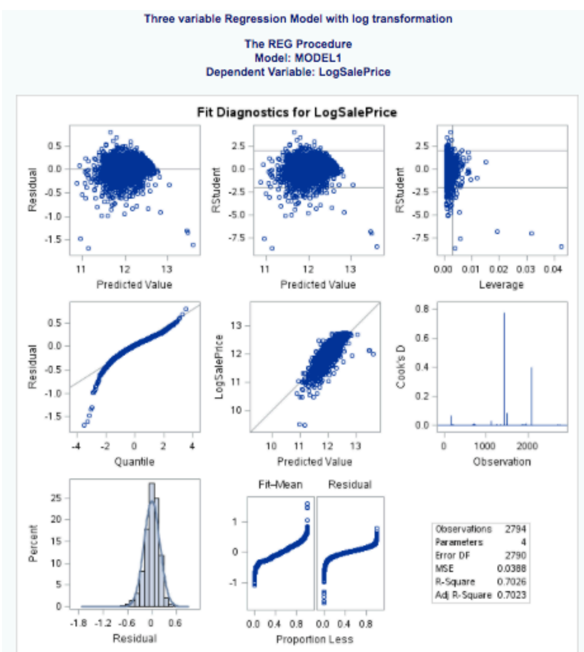
The REG Procedure
Model: MODEL1
Dependent Variable: LogSalePrice

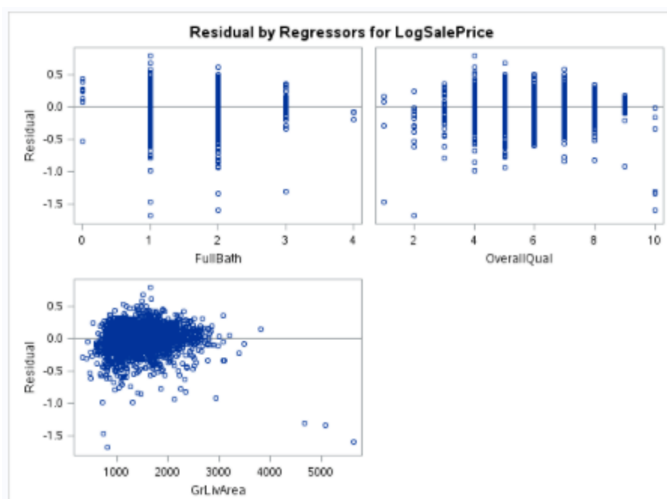
Number of Observations Read	2794
Number of Observations Used	2794

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	256.08257	85.36086	2197.24	<.0001
Error	2790	108.38892	0.03885		
Corrected Total	2793	364.47149			

Root MSE	0.19710	R-Square	0.7026
Dependent Mean	11.97661	Adj R-Sq	0.7023
Coeff Var	1.64572		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	10.56716	0.01791	590.13	<.0001	0
FullBath	1	0.06018	0.00904	6.65	<.0001	1.73082
OverallQual	1	0.16768	0.00346	48.53	<.0001	1.45179
GrLivArea	1	0.00021758	0.00001068	20.37	<.0001	1.73783





After refitting (2647 observations):

Three variable Regression Model with log transformation-Outliers Removed

The REG Procedure
Model: MODEL1
Dependent Variable: LogSalePrice

Number of Observations Read	2647
Number of Observations Used	2647

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	225.07280	75.02427	3203.45	<.0001
Error	2643	61.89853	0.02342		
Corrected Total	2646	286.97134			

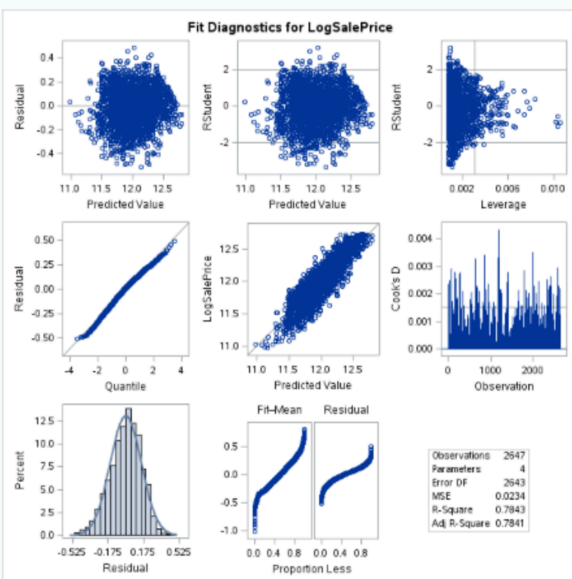
Root MSE	0.15304	R-Square	0.7843
Dependent Mean	11.99478	Adj R-Sq	0.7841
Coeff Var	1.27585		

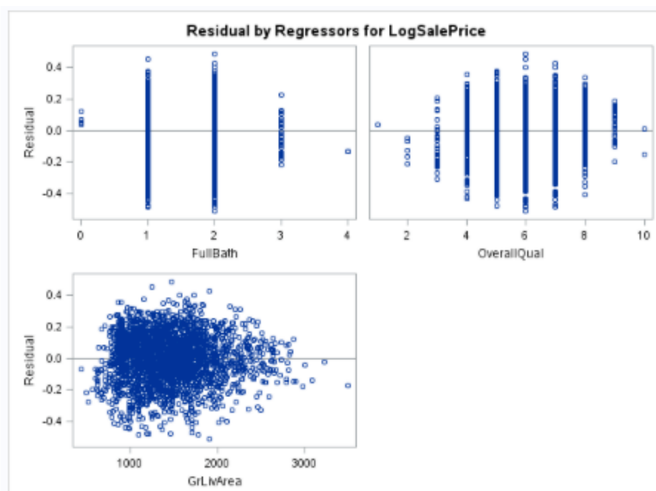
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	10.58913	0.01491	710.37	<.0001	0
FullBath	1	0.05447	0.00765	7.12	<.0001	1.86390
OverallQual	1	0.15725	0.00291	54.10	<.0001	1.49829
GrLivArea	1	0.00026124	0.00000932	28.02	<.0001	1.81235

Three variable Regression Model with log transformation-Outliers Removed

The REG Procedure
Model: MODEL1
Dependent Variable: LogSalePrice





Conclusion: The before and after removal of the influential points using DFFITS the R-squared value for the multiple linear regression model (MRM) were 0.7026 and 0.7843, respectively. A total of 147 observations were removed during this process. Though only 147 data points were deleted there was a substantial improvement in the fitted model afterwards suggesting that some of these 'model-based' outliers had a disproportionate effect on the overall fit of the model. Furthermore comparing before and after models, the residuals improved so that the plot of residuals versus predicted value after cleaning showed no funnel-shaped distribution, indicating a random disposition of residuals. Furthermore, the QQplot after cleaning showed a much tighter fit to the ideal normality line than before the cleaning. A similar effect was observed in the plot of LogSalePrice *versus* Predicted value (center middle panel above), which displayed a line through the data points, which is very well aligned with the best-fit 45^{o} line through the origin. We conclude that the removal of influential points did improve the fit of the model based on the R-squared comparison and the disposition of the residuals.

CONCLUSIONS

In what ways do variable transformation and outlier deletion impact the modeling process and the results? Are these analytical activities a benefit or do they create additional difficulties?

The use of variable transformation and outlier detection and removal, separately and in combination, helped to drastically improve the fitted models. This helped not only in tightening the observations such that there was less scatter and variability but also the distribution of residuals was improved and become more random (ideal assumption), so that no unusual shapes (*e.g.* funnel-shape) observed in residual plots. While the fits to the linear models and their goodness-of-fit improved, producing a 'better' model, there were observations removed from the analysis. From the observations removed from the SalePrice, most were deleted from the upper price range. While absent it is unclear how useful these observations are, though that depends on the nature of the question we are asking. Specifically, if we wanted to know what makes these expensive houses more valuable than most residences in Ames, then we are no longer in a position to address that.

What do you consider to be next steps in the modeling process?

Since we have not yet been able to incorporate categorical variables into any fitted MRM model, this remains another step. In addition further iterations of model adequacy and validation may be required in order to identify and remove additional outliers and influential points in the predictor variable data set, in addition to SalePrice. Once we are completed with these steps and we are satisfied with the model, we can set about using the model – the original purpose.