Noé Flores
Data Analysis and Regression

Introduction:
Our target goal for this analysis is to build upon the initial exploratory analysis and regression models in an effort to identify a possible predictor variable in our dataset with a good fit for sale price. We previously incorporated Simple Linear Regression Models (SLR), and more complex Multiple Linear Regression Models (MLR) in our analysis. We move forward from this point and seek to incorporate different statistical methods to aid in our search for a model with a good fit and interpretability for our needs.

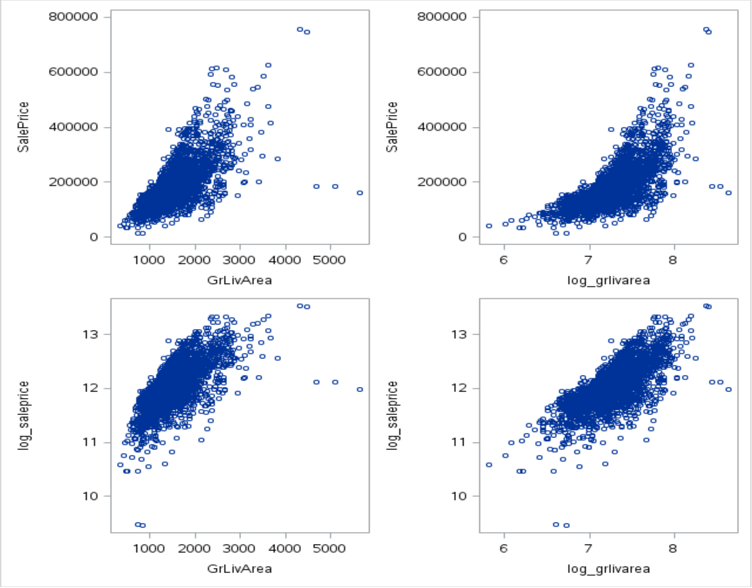
Results and Analysis:
We begin by transforming our dependent variable SalePrice, and our designated predictor variable GrLivAre by taking the log of each and adding the values to a new dataset. When running a regression analysis, taking the log of a variable can help reduce skewing. What we can immediately notice is the range of the log values is substantially tighter than the original values.

Figure 1: Log of SalePrice and GrLivArea.

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

Below is a plot comparing the new variables to their pre-transformed state.

Figure 2: SalePrice GrlivArea and Log of SalePrice and GrLivArea.



From the initial inspection of the four plot combinations, it would appear that the plot containing both transformed variables against one another portrays the most linearity.

We continue by fitting four new models, one for each pair-wise combination of SalePrice and GrLivArea. The equation for each of the models we are seeking to build is as follows:

Model (1) SalePrice = $\beta_0 + \beta_1$ (GrLivArea) + ε

Model (2) SalePrice = $\beta_0 + \beta_1$ (log_GrLivArea) + ε

Model (3) log_SalePrice = $\beta_0 + \beta_1$ (GrLivArea) + ε

Model (4) log_SalePrice = $\beta_0 + \beta_1$ (log_GrLivArea) + ε

The results of running the regression model on each combination produce the following parameter results:

Figure.3 Regression model #1 SalePrice with GrLivArea parameter estimates

Variable	Degrees of Freedom	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	13290	3269.70277	4.06	<.0001
GrLivArea	1	111.694	2.06607	54.06	<.0001

The complete fitted equation for model#1 is the following:

SalePrice = 13,290 + 111.694 * GrLivArea

The predicted value is SalePrice. Therefore the model coefficients of the equation indicate to us that if GrLivArea was 0, the final SalePrice of the house would be $13,290, and a one unit change in GrLivArea sqfootage would result in a SalePrice mean increase of $111.69. Our t-values are greater than zero, and p-values are below the benchmark .05.

Figure.4 Regression model #2 SalePrice with log_GrLivArea parameter estimates

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

The entire fitted equation for model#2 is the following:

SalePrice = -1,060,765 + 171,011 * log_GrLivArea

For this model since our independent variable is log-transformed, we can conclude that a one percent change in GrLivArea sq footage, would result in an average change in SalePrice of 1710.11% (171011/100). Our t-values, in this case, are no longer positive, for our intercept which is of concern, and our p-values are below the benchmark standard of .05.

Figure.5 Regression model #3log_ SalePrice with GrLivArea parameter estimates

Variable	Degrees of Freedom	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	11.17954	0.01694	660.12	<.0001
GrLivArea	1	0.00056107	0.0000107	52.43	<.0001

The complete fitted equation for model#3 is the following:

log_SalePrice = 11.178 + 0.0006 * GrLivArea

In this model, our dependent variable has been log transformed. We can therefore conclude that a one percent increase in GrLivArea SQ footage, would result in an average increase in SalePrice of 6% (.0006*100). Our t-values in this case are strong on the positive end and our p-values remain below .05.

Figure.6 Regression model #4 SalePrice with log_GrLivArea parameter estimates

Variable	Degrees of Freedom	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	5.43019	0.11644	46.63	<.0001
log_grlivarea	1	.90781	0.01602	56.66	<.0001

The complete fitted equation for model#3 is the following:

log_SalePrice = 5.43 + .908 * log_GrLivArea

Our final model has both the dependent and independent variable log transformed. We can interpret the results as a one percent increase in log_GrLivArea sqfootage, would result in an average change in the mean of SalePrice of 91%. Once again, our t-values, in this case, are strong and on the positive end and our p-values remain below the .05 threshold.

The table below provides a numerical summary of each model, which allows us to compare the goodness-of-fit of the four models incorporated in the analysis.

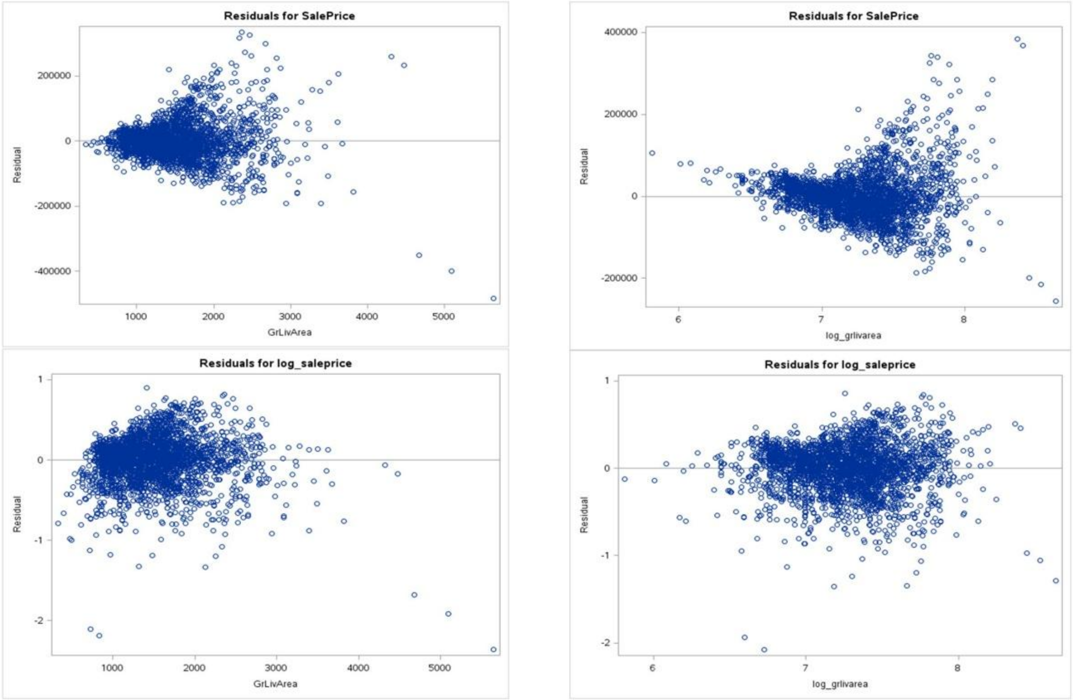
Figure.7 Regression models Summary table

Models	R-Square	Adj-R-Square	F-Statistic	Pr > F
(1) SalePrice = 13,290 + 111.694 * GrLivArea	0.4995	0.4994	2922.59	<.0001
(2) SalePrice = -1,060,765 + 171,011 * log_GrLivArea	0.4831	0.4829	2736.45	<.0001
(3) log_SalePrice = 11.178 + 0.0006 * GrLivArea	0.4842	0.484	2748.89	<.0001
(4) log_SalePrice = 5.43 + .908 * log_GrLivArea	0.523	0.5228	3209.97	<.0001

Based on the statistical output from our regression analysis, and solely considering the information in the table above, we could conclude that our fourth model had the best fit. This particular model, which incorporated the log-transformed variables for both SalePrice and GrLivArea, had the highest F-Statistic, which is used to decide whether the model as a whole has a statistically significant predictive capability. Generally speaking, we want F-Statistic to be high. The R-Square value was over 50%, which helps us identify the percentage of the dependent variable variation that is explained by the model. Our Adjusted R-Square was also over 50% and indicates how well our model fits, adjusting for the number of variables used.

Taking a look at the graphics of our models, it would appear that our initial assessment that the model with both log-transformed variables showed the most linearity. The histogram of the residuals associated with model shows a relatively normal distribution. The Q_Q-Plot of the log-transformed model shows some light tailing but is better than the other three models. What is worth noting is the presence of some outliers. This is of some concern, given that the model does appear to fit our needs, but the presence of the outliers could be having an effect on our model. In Figures.8 below you can see a breakdown of the residuals for each model. The log-transformed model once again appears to fit best, and we can also see some of the outliers previously mentioned.

Figure.8Regression models residual plots.



One of the most common reasons to use log-transformed variables is to convert skewed data to conform to normality. Therefore, use of transformation allows us to obtain residuals that are more normally distributed. Another reason we incorporate transformation of variables is to achieve linearity. There are however some concerns and considerations we should take into account when using log-transformed variables in the regression. One primary concern is that the log-transformed variables and subsequent analysis could hide limitations of the model. Perhaps forcefully making the variables fit. In our case, using a log-transformation variable for both our dependent and independent variables was beneficial. This log transformed model (4), provided more linearity than the original price model and improved our residuals. It did not, however, hide the fact that we have some outlying data points.

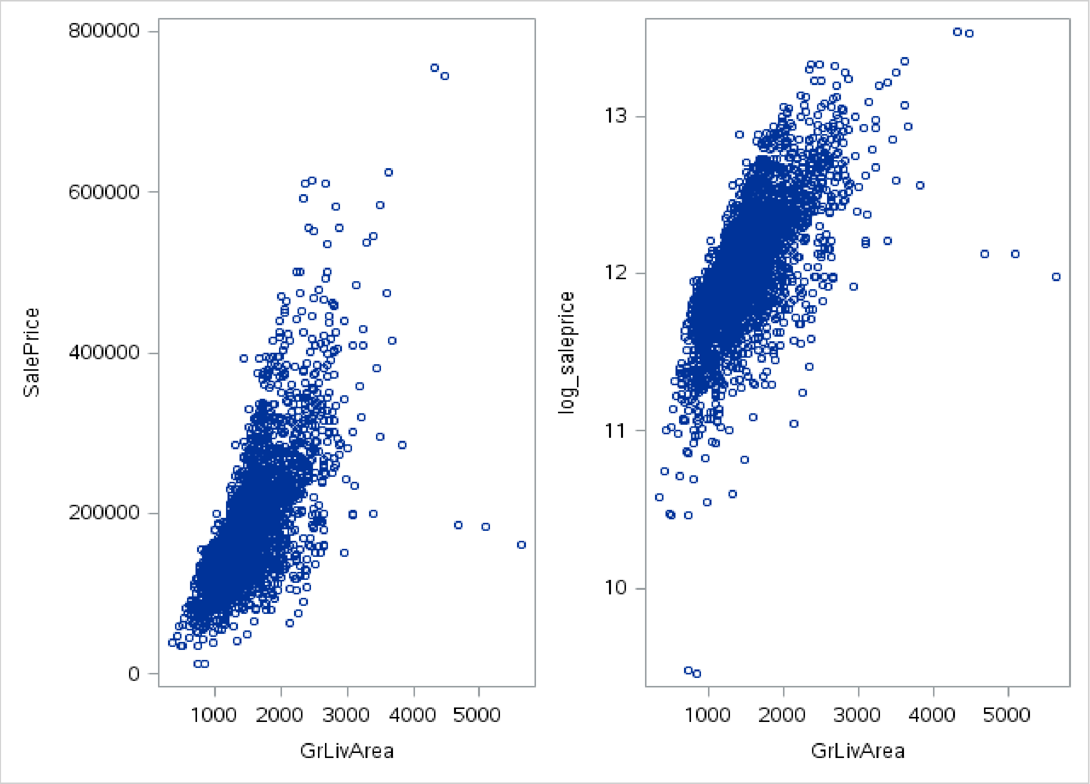
We will now observe the relationship of our log transformed SalePrice variable, and how it correlates to the other continuous variables in our dataset. The correlation results are below.

Figure.8Regression models residual plots

Variable	GrLivArea	GarageArea	TotalBsmtSF	FirstFlrSF	MasVnrArea
Correlation to Sales Price	0.696	0.651	0.625	0.602	0.450
P- Value	<.0001	<.0001	<.0001	<.0001	<.0001
Observations	2930	2929	2929	2930	2907

Similar to our original finding, GrLivArea continues to be the most correlated variables in the dataset, with respect to the log-transformed variable SalePrice. We will plot a side by side comparison of GrLivArea with respect to SalePrice and the log of SalePrice.

Figure 9.Plot of GrLivArea vs SalePrice and Log_SalePrice.



Since we have a log-transformed variable, we see that there is a difference in the scale of the two plots. In both plots, we see the signs of linearity. In both graphs, we also see the presence of outliers. Transforming the dependent variable to log form did not fix the issue. We still have some noticeable outlying data, and that should be investigated. It highlights our concern that using a transformation for the purposes of making our data fit or look better is likely unjustified. There is also the added issue of the results being harder to interpret by individuals who aren't trained in statistical analysis.

Transforming the dependent or response variable is typically used when we want to achieve greater normality and reduce variance. Specifically, transforming the dependent variable is useful for attaining greater linearity and producing a "straighter" line. This is on display in Figure.9, where see the straightening of the curvature present in the original non-transformed plot.

Building upon our previous observations, one option we have is to take the square root of one of our variables. Taking the square of a variable is another option to normalize data. We will use the square root of our dependent variable SalePrice, and run the regression model against GrLivArea.

The equation for the new model we are seeking to build is as follows:

$$\sqrt{\boldsymbol{SalePrice}} = \boldsymbol{\beta_0} {}_{+}\boldsymbol{\beta_1}\ \textbf{(log\_GrLivArea)} + \boldsymbol{\varepsilon}$$

The results of running the regression model on each combination produce the following parameter results:

Figure.10 Regression model #1 SalePrice with GrLivArea parameter estimates

Variable	Degrees of Freedom	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 complete fitted equation for model#1 is the following:

$$\sqrt{\boldsymbol{SalePrice}} = \mathbf{232.93 + 0.122} * \boldsymbol{GrLivArea}$$

The predicted value is now $\sqrt{SalePrice}$, therefore the model coefficients of the equation indicate to us that if GrLivArea was 0, the final SalePrice of the house would be equal to the $\sqrt{\$232.93}$, and a one unit change in GrLivArea SQ footage would result in a SalePrice mean increase of $\$\surd\$0.12$. Our t-values are greater than zero, and p-values are below the benchmark .05.

The table below provides a numerical summary of our new model, along with the original model containing non-transformed variables for SalePrice and GrLivArea, and the model we previously found to have the best fit, which includes transformed variables of both SalePrice and GrLivArea.

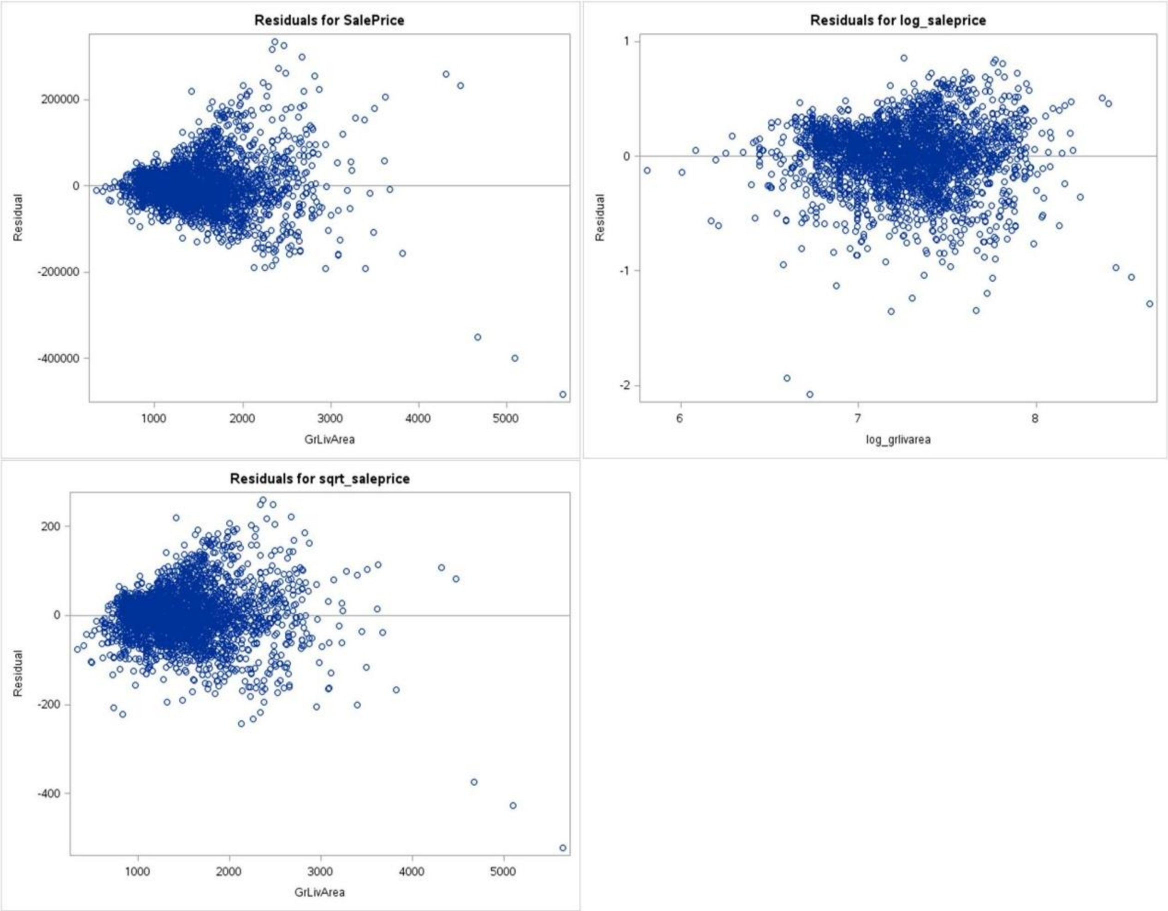
Figure.11 Regression models Summary table2

Models	R-Square	Adj-R-Square	F-Statistic	Pr > F
(1) SalePrice = 13,290 + 111.694 * GrLivArea	0.4995	0.4994	2922.59	<.0001
(3) log_SalePrice = 5.43 + .908 * log_GrLivArea	0.523	0.5228	3209.97	<.0001
(3) √SalePrice = 232.93 + 0.122 * GrLivArea	0.508	0.507	3017.28	<.0001

Considering the information in the table above, we could conclude that our model which incorporated the log-transformed variables for both SalePrice and GrLivArea continues to perform better than the rest. The new model using the Square Root transformed dependent variable performed well, with higher R-Square, Adjusted-R square, and F-Statistic than our original non-transformed model, but falls short of the log-transformed model in our opinion.

Comparing the graphics of all three models leads us to maintain our position on which model worked best. The new model has graphics that show a histogram with a normal distribution, and the Q_Q-Plot still has some light tailing. We still see outliers present in our plots. The new transformation clearly wouldn't take those away, but more importantly, the new transformation did not do much to improve the results either. Figures.12 has a breakdown of the residuals for each model we just reviewed. The log-transformed model once again appears to fit best, and we see those ever prevalent outliers we mentioned.

Figure.12 Plot of SalePrice,Sqrt_SalePrice vs GrLivArea and log_SalePrice vs GrLivArea



A recurrent theme in every model we ran so far in this analysis is the prevalence of the outliers in our graphics outputs. Outliers present in our data present us with a difficult set of questions, which is how we should handle and treat their presence? Are the outliers legitimate records, and therefore should be kept, or are they indeed anomalies that should be kept out. To begin, we will have to perform some additional exploratory analysis on our SalePrice variable, and attempt to address the outliers in our data.

We provided below a table of the Quantiles for SalePrice figures in our Ames Housing data set. The information in the Quantiles table below paints a picture of the range of prices for our SalePrice. The Quantiles give us a description of where our SalePrice data points are located. It also provides a visual of the range.

Figure.13 Quantiles Summary table

Quantiles	
Level	**Quantile**
100% Max	$ 755,000
99%	$ 457,347
95%	$ 335,000
90%	$ 281,357
75% Q3	$ 213,500
50% Median	$ 160,000
25% Q1	$ 129,500
10%	$ 105,250
5%	$ 87,500
1%	$ 61,500
0% Min	$ 12,789

The next table present expands on the idea of the range for our data and shows the top 5 and bottom 5 values within the SalePrice variable.

Figure.14 Extreme Observations in SalePrice

Extreme Observations			
Highest		**Lowest**	
Value	**Obs**	**Value**	**Obs**
$ 611,657	45	$ 12,789	182
$ 615,000	1064	$ 13,100	1554
$ 625,000	2446	$ 34,900	727
$ 745,000	1761	$ 35,000	2844
$ 755,000	1768	$ 35,311	2881

Taking a quick look at the extreme values, it doesn't appear that the data points are completely implausible, meaning that in theory, the data points fit more or less within the range. However, if we take a more critical look, we could probably justify saying that there are 2 data points on each end of the spectrum that could be influencing our analysis, and don't quite fit in. These values should more likely than not be considered outliers in our dataset.

A conventional method used to deal with outliers in data is referred to as the Interquartile Range Rule (IQR). This rule basically takes the Interquartile range of the data, multiplies that figure by 1.5 times, and add the result to the third quartile, and subtracts it from the first quartile. The table below provides the Interquartile range value and the results of applying the Interquartile rule for outliers.

Figure.15 Interquartile Rangeand Results

Interquartile Range Calculation	Interquartile Range Results
IQR= Q3 - Q1	Top = \$213,500 + (\$84,500 * 1.5) **= \$339,500**
IQR = \$213,500 - \$129,500 = **\$84,500**	Bottom = \$129,500 - (\$84,500* 1.5) **= \$3,500**

Given the results of the IQR results, and looking at our data range, and the extreme values information, I don't believe the IQR rule is a suitable solution for our dataset. It would omit what appears to be a sizeable amount of data.

This brings us to another possible solution for our outliers, which is the 2 standard deviation rule. This rule advocate's taking the standard deviation of our variable, multiplying it by 2, and adding and subtracting that total from the mean value. The breakdown of those results is below.

Figure.16 2STD Range and Results

2 Standard deviation Calculation	Interquartile Range Results
2 * STD	Top = \$159.774 + \$180,796 **= \$340,570**
2STD = \$79,887 * 2 = **\$159,774**	Bottom = -\$180,796 - \$159,774 **= \$21,022**

This method appears to work well with the lowest extremes of our data, but once again it looks as though it would omit a significant portion of the high-end data points. Still, considering the fact that the IQR rule appears to continue to include the lowest extreme of our data, we will move forward with the 2 Standard deviation rule for outlier omission. This will ensure removal of the smallest and most significant data points from SalePrice we previously observed.

Figure.17 Outlier Definitions

Outlier Definition
if saleprice <= **\$21,022** then price_outlier = 1;
else if saleprice > **\$21,022** & saleprice < **\$340,570** then price_outlier = 2;
else if saleprice >= **\$340,570** then price_outlier = 3;

These rules we set forward may not be perfect, but it will suit our needs of removing the extreme values on the top and bottom range of the SalePrice data, and still maintain a great amount of information contained in SalePrice. After running the definitions through our SalePrice data, we can summarize precisely how many points lie within each newly defined bucket.

Figure.18 Outlier Definitions

Price_Outlier	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	2	0.07%	30	0.07%
2	2794	95%	2794	95%
3	134	4.57%	134	4.57%

The information in the table above gives us some valuable figures and some confidence in the outlier definition we chose. If we omit Price_Outlier data points 1 and 3 we will still have 2,794 data points to run in our analysis, a substantial portion. We can now move forward and remove price_outliers 1 and 3 and create a cleaned SalePrice dataset without outliers. We will use this new version of SalePrice data to re-run a few models and observe their fit.

Our first model incorporating the new trimmed and cleaned dataset will fit a simple linear regression model using GrLivArea as our predictor variable. Since this model is integrating the new SalePrice data, and to avoid confusion, we will mark the change with a superscript to SalePrice, ($\boldsymbol{Saleprice^1}$).

$$\mathbf{Saleprice^1} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}\ \mathbf{(GrLivArea)} + \boldsymbol{\varepsilon}$$

The results of running the model produce the following parameter results:

Figure 19: Parameter estimates of Regression model $Saleprice^1$with GrLIvArea.

Variable	Degrees of Freedom	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	47,397	2806.90605	16.89	<.0001
GrLivArea	1	83.77971	1.83758	45.59	<.0001

The complete fitted equation of the model is the following:

$$\mathbf{Saleprice^1} = \mathbf{47,397} + \mathbf{83.78(GrLivArea)}$$

The desired predicted value is $Saleprice^1$. The model coefficients in the equation indicate to us that if GrLivArea was 0, the final $Saleprice^1$ of the house would be $47,397, and a one unit change in GrLivArea square footage would result in a $Saleprice^1$ mean increase of $83.78. Our t-values are greater than zero, and p-values are significantly below the benchmark .05 level

Our second model will incorporate a multiple linear regression model using MasVnrArea and GrLivArea as our predictor variables for $Saleprice^1$.

$$\mathbf{Saleprice^1} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}\ \mathbf{(GrLivArea)} + \boldsymbol{\beta_2}\ \mathbf{(MasVnrArea)} + \boldsymbol{\varepsilon}$$

The results of running the model produce the following parameter results:

Figure 20: Parameter estimates of MLR Regression model $Saleprice^1$with GrLIvArea + MasVnrArea

Variable	Degrees of Freedom	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	51,625	2761.19424	18.7	<.0001
GrLivArea	1	76.73699	1.88495	40.71	<.0001
MasVnrArea	1	66.53444	5.57388	11.94	<.0001

The complete fitted equation of the model is the following:

$$\mathbf{Saleprice^1 = 51{,}625 + 76.74 * GrLivArea + 66.53 * MasVnrArea}$$

The model coefficients in the equation indicate to us that if GrLivArea was 0, and MasVnrArea was 0, the final Saleprice^1 of the house would be $51,625. If GrLiveArea is fixed, then for each increase of 1 SQ foot in MasVnrArea, the sale price will increase by $66.53. If MasVnrArea is fixed, then for each increase of 1 SQ foot in GrLivArea, the sale price will increase by $76.47. Our t-values are greater than zero and our p-values are below .05. The coefficients have more impact relative to the final Saleprice^1 , meaning that in the strange case where our predictor are could be 0, our final Saleprice^1 is significantly depleted.

Our final model with the clean data is a multiple linear regression model using MasVnrArea, GrLivArea, and BsmtUnfSf as our predictor variables for Saleprice^1.

$$\textbf{Saleprice}^{\mathbf{1}} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}\ \textbf{(MasVnrArea)} + \boldsymbol{\beta_2}\ (\textbf{GrLivArea}) + \boldsymbol{\beta_3}\ (\textbf{BsmtUnfSF}) + \boldsymbol{\varepsilon}$$

The results of running the model produce the following parameter results:

Figure 21: Parameter estimates of MLR model #.

Variable	Degrees of Freedom	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	50,269	2798.91568	17.96	<.0001
GrLivArea	1	75.43146	1.92948	39.09	<.0001
MasVnrArea	1	66.51072	5.56596	11.95	<.0001
BsmtUnfSF	1	5.94211	1.97907	3	0.0027

The complete fitted equation of the model is the following:

Saleprice^1 = 50,269 + 75.43 * GrLivArea + 66.51 * MasVnrArea+ 5.94 * BsmtUnSF

If GrLivArea was 0, MasVnrArea was 0, and BsmtUnSF was 0, the final Saleprice^1 of the house would be $50,269. If GrLiveArea and MasVnrArea are fixed, then an increase of 1 SQ foot in BsmtUnfSF will increase Saleprice^1 by $5.94. If MasVnrArea and BsmtUnSF are fixed, then for each increase of 1 sqfoot in GrLivArea, the Saleprice^{1}will increase by $75.43. If GrLiveArea and BsmtUnfSF are fixed, then an increase of 1 SQ foot in MasVnrArea will increase Saleprice^{1}by $66.51. Our t-values are greater than zero for and the p-values remain below the benchmark of .05.

In order to assess the effect of removing data points, we observed and considered to be outliers; we will compare the results of the three models we ran against our new Saleprice^1 variable vs the models using the original SalePrice data.

Figure 22: Summative Table of Models.

Models With trimmed Outliers of SalePrice	R-Square	F-Statistic	Pr > F
Saleprice^1 = 47.397 + 83.78(GrLivArea)	0.4268	2078.67	<.0001
Saleprice^1 = 51,625 + 76.74 * GrLivArea + 66.53 * MasVnrArea	0.4558	1160.23	<.0001
Saleprice^1 = 50,269 + 75.43 * GrLivArea + 66.51 * MasVnrArea+ 5.94 * BsmtUnSF	0.4573	777.57	<.0001
Models Without trimmed Outliers for SalePrice			
SalePrice = 13,290 + 111.694 * GrLivArea	0.4995	2922.59	<.0001
SalePrice = 26,547 + 94.60 * GrLivArea + 118.55 * MasVnrArea	0.5599	1846.98	<.0001
SalePrice = 25,788 + 93.91 * GrLivArea + 118.57 * MasVnrArea+ 3.23 * BsmtUnSF	0.56	1231.02	<.0001

Summative Analysis for models:

Considering the information in the table above, it does not appear like the new models using the manipulated SalePrice data (Saleprice^{1}) have improved the general fit of our models. The numerical assessments for goodness-of-fit, R-Square, and F-Statistic are both worse off in the newer models. The difference isn't very substantial, but the obvious lack of improvement is visible.

Comparing the graphics of the new versus old models, we don't see any significant differences there either. The distributions look relatively normal in the histogram, and the residual data seems fairly uniform and in line with desired results of a regression analysis. What is interesting is that our Q_Q plots in the models with manipulated Saleprice^{1} data appear to have more normal results, meaning the plots are lying flatter at the upper range of the plot, but there is still significant tailing and outliers present.

Figure.23 Q_Q Plots for Manipulated Models in Respective order:

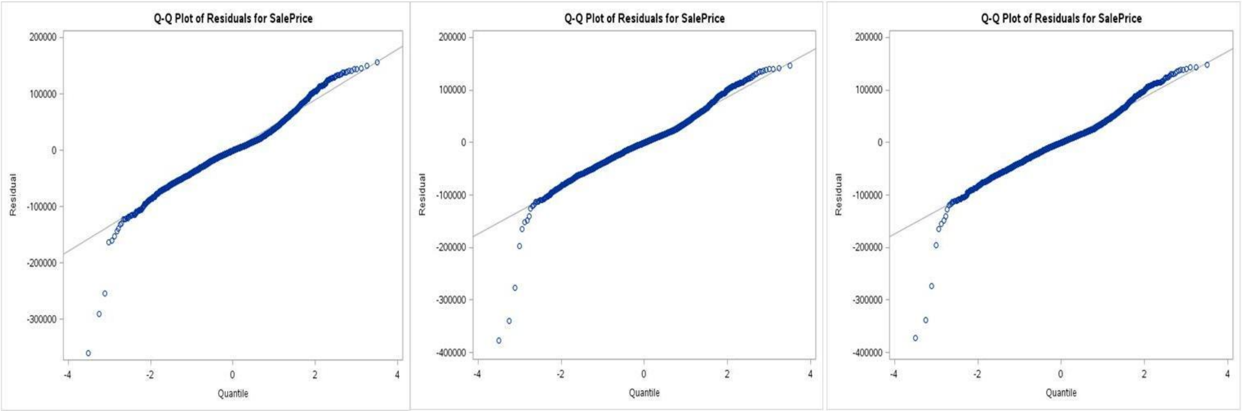
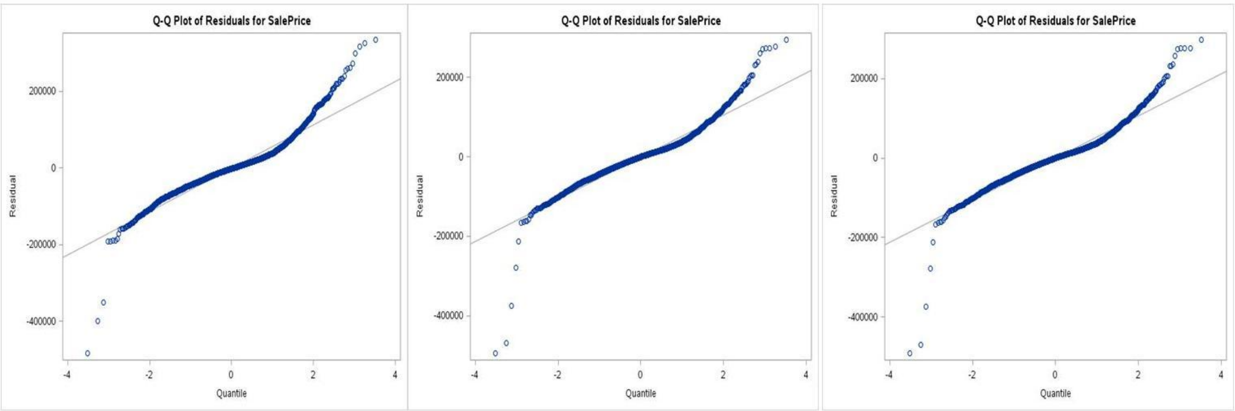


Figure.23 Q_Q Plots for Non-Manipulated Models in Respective order



The general overall assessment given all the data and models we have executed so far leads me to believe that our original models were achieving a sufficiently acceptable goodness-of-fit, and there wasn't much need for manipulating or removing outliers.

Conclusion:

After going through the exercise of analyzing various models throughout this investigation, it has become evident that transformation and outlier deletion are statistical techniques that have a tremendous amount of impact on our models. Transformation is useful in cases where we want to fit our regression models to reduce skewness, conform to a more normal distribution, and improve linearity. There are many obvious reasons for removing outliers from data to achieve similar results as well, with the most obvious being erroneous or misplaced values within the data. In our analysis, we found that transforming both dependent and independent variables into Log form helped us produce the model with the best goodness-of-fit. By comparison, removing the outliers in this analysis did not improve our models, and potentially omitted data points that should be included in the overall assessment.

This brings us to the issues of subjectivity and individual interpretation that comes up in the removal of outliers and transformation. In each situation, performing an initial exploratory analysis and running models without prior manipulation provide insight as to the possible need for incorporating one of these techniques into the models. It is certain to impact the goodness-of-fit when transformation and manipulation is involved, and its best to have a non-manipulated, non-transformed baseline for assessment.

The next step for modeling this particular data set would be to expand model validation. The validation process of these initial models could benefit from Cross Validation, which would allow us to assess how well the model will perform in the future, based on the true values in the dataset.

Appendix: SAS code

```
/*Read file into SAS and identify it as mydata*/
libname mydata '/scs/wtm926/' access = readonly;

Data temp1;
   set mydata.ames_housing_data;

proc sgscatter data=temp1;
    plot saleprice*grlivarea;
run;

/****(1)Use continuous predictor from assignment 2 and create two new variables***/
data temp2;
   set temp1;
   log_saleprice = log(saleprice);
   log_grlivarea = log(grlivarea);
   keep log_grlivarea saleprice log_saleprice Lotfrontage LotArea MasVnrArea BsmtFinSF1
BsmtFinSF2 BsmtUnfSF TotalBsmtSF FirstFlrSF SecondFlrSF
    LowQualFinSF GrLivArea GarageArea WoodDeckSF OpenPorchSF EnclosedPorch
ThreeSsnPorch ScreenPorch
    PoolArea MiscVal;
/***(1.b)Print out new data***/
proc print data=temp2 (obs=5);
run;

proc sgscatter data=temp2;
  plot(saleprice log_saleprice) *( grlivarea log_grlivarea);
run;

/***2. Fit four models for Saleprice to Predictor variable grlivarea***/
/*** Model 1 Saleprice vs GrlivArea***/
proc reg data=temp2 plots(unpack)=(diagnostics fitplot residualplot);
    model SalePrice = GrLivArea;
    title 'Model Saleprice vs. Grlivarea';
run;
/****model 2 Saleprice vs Log_Grlivarea***/
proc reg data=temp2 plots(unpack)=(diagnostics fitplot residualplot);
    model SalePrice = log_GrLivArea;
    title 'Model Saleprice vs. log_Grlivarea';
run;
/***Model 3 Log_saleprice vs Grlivarea***/
proc reg data=temp2 plots(unpack)=(diagnostics fitplot residualplot);
    model log_SalePrice = GrLivArea;
    title 'Model Log_Saleprice vs. Grlivarea';
run;
/***Model 4 Log_Saleprice vs Log_Grlivarea***/
```

```
proc reg data=temp2 plots(unpack)=(diagnostics fitplot residualplot);
   model log_SalePrice = log_GrLivArea;
   title 'Model Log_Saleprice vs. Log_Grlivarea';
run;

/***3. Correlate continuous variables with Log_Saleprice***/
proc corr data=temp2 nosimple rank;
   var log_SalePrice;
   with Lotfrontage LotArea MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF
    FirstFlrSF SecondFlrSF LowQualFinSF GrLivArea GarageArea WoodDeckSF OpenPorchSF
    EnclosedPorch ThreeSsnPorch ScreenPorch PoolArea MiscVal;
run;

proc sgscatter data=temp2;
   plot (saleprice log_SalePrice) * grlivarea;
run;


/***4. Create another transformation variable on Y(Saleprice)***/
data temp3;
   set temp2;
   sq_saleprice = saleprice*saleprice;
   sqrt_saleprice = sqrt(saleprice);
   keep sq_saleprice sqrt_saleprice log_grlivarea saleprice log_saleprice MasVnrArea BsmtUnfSF
GrLivArea;
run;

proc print data=temp3 (obs=5);
run;

/***Check plot of new variables with SalePrice to asses fit***/
proc sgscatter data=temp3;
   plot (saleprice sq_saleprice sqrt_saleprice) * grlivarea;
run;

proc reg data=temp3;
   model sqrt_saleprice = grlivarea;
   title 'Model Sqrt_Saleprice vs. Grlivarea';
run;

proc sgscatter data=temp3;
   plot (sqrt_saleprice saleprice log_SalePrice) * grlivarea;
run;


/****Part 5 identify outliers***/
proc univariate normal plot data=temp1;
```

```
   var SalePrice;
   histogram SalePrice / normal (color=red w=5);

/***Double Check mean of SalePrice***/
proc means data=temp1;
   var saleprice;
run;

/***Double Check high and Low values of SalePrice***/
proc sort data=temp1;
   by descending saleprice;
proc print data=temp1 (obs=10);
   run;

proc sort data=temp1;
   by saleprice;
proc print data=temp1 (obs=10);
   run;

/***outlier classification***/
data part5;
   set temp1;
   keep sq_saleprice sqrt_saleprice log_grlivarea saleprice log_saleprice MasVnrArea
   BsmtUnfSF 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;
run;

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

proc freq data=part5;
  tables price_outlier;
run;

proc sort data=part5;
   by price_outlier;
run;

proc means data=part5;
   by price_outlier;
   var saleprice;
run;
```

```
/***Part 6. Create a dataset without outliers by including a delete statement***/

data part6;
   set part5;
   if price_outlier = 1 then delete;
   if price_outlier = 3 then delete;
run;
proc univariate normal plot data=part6;
   var SalePrice;
   histogram SalePrice / normal (color=red w=5);

proc freq data=part6;
   tables price_outlier;
run;

proc sort data=part6;
   by price_outlier;
run;
proc means data=part6;
   by price_outlier;
   var saleprice;
run;
/***Model SalePrice vs GrLivArea***/
proc reg data=part6 plots(unpack)=(diagnostics fitplot residualplot);
   model saleprice = GrLivArea;
run;
/***Model SalePrice vs GrLivArea and MasVnr Area***/
proc reg data=part6 plots(unpack)=(diagnostics fitplot residualplot);
  model SalePrice = GrLivArea MasVnrArea;
run;
/***Model SalePrice vs GrLivArea + MasVnr Area + BsmtUnfSF***/
proc reg data=part6 plots(unpack)=(diagnostics fitplot residualplot);
  model SalePrice = GrLivArea MasVnrArea BsmtUnfSf;
run;
/***Old Models***/
proc reg data=temp1 plots(unpack)=(diagnostics fitplot residualplot);
   model saleprice = GrLivArea;
run;
/***Model SalePrice vs GrLivArea and MasVnr Area***/
proc reg data=temp1 plots(unpack)=(diagnostics fitplot residualplot);
  model SalePrice = GrLivArea MasVnrArea;
run;

/***Model SalePrice vs GrLivArea + MasVnr Area + BsmtUnfSF***/
proc reg data=temp1 plots(unpack)=(diagnostics fitplot residualplot);
  model SalePrice = GrLivArea MasVnrArea BsmtUnfSf;
run;
```

References

(1) Black, K. (2008). *Business statistics: For contemporary decision making*. Hoboken, NJ: Wiley.

(2) Montgomery, D. C., Peck, E. A., Vinning, G. G., (2012). *Introduction to Linear Regression Analysis* Hoboken, NJ: Wiley.

(3) Wilcox, R. R. (2009). *Basic statistics: Understanding conventional methods and modern insights*.Oxford: Oxford University Press.

(4) Cody, R. (2011). *SAS: Statistics by Example*. Carey, NC: SAS Institute Inc.

(5) Cooks distance. https://en.wikipedia.org/wiki/Cook's_distance (accessed January 13, 2017)

(6) Data Transformations. https://onlinecourses.science.psu.edu/stat501/node/320(accessed January 19, 2017)

(7) IQR Rule for Outliers. http://www.unc.edu/~rls/s151-09/class4.pdf (accessed January 20, 2017)