Assignment #3

Scott M. Morgan

Introduction:

The purpose of this assignment is to continue building regression models to predict home sale price. To accomplish this objective, I use data from the Ames housing data set provided by the course instructor and SAS Studio to perform the analysis. Understanding the nuances of a data set and regression analysis is an integral part of the predictive modeling process. The importance of linear trends in regression analysis cannot be understated and in this assignment we first explore how variable transformation can assist in creating a more linear fit. Tangentially related is the appropriate handling of outliers which is becoming ever more important as the volume, variety and velocity of data grows. In the second portion of this assignment, we identify potential outliers and analyze what our models look like with and without them by "cleaning" the data. I expect that this exercise will demonstrate the usefulness of various transformation methods as well as introduce a logical approach for the handling of outliers. As practitioners of predictive analytics, we must keep at the forefront that there is seldom a singular metric that points to the "best" model, but rather a confluence of factors which includes statistical analysis and logical reasoning.

Results:

In the subsequent sections, we first use the functionality within SAS Studio to transform our variables and perform a brief exploratory data analysis (EDA) to build a rudimentary understanding of the modified predictors. Following this, we generate a series of regression models that both include and exclude the newly transformed variables. The second portion focuses on the existence potential outliers, rules for formal identification and their subsequent impact on a series of predictive models. We systematically summarize, discuss, compare and contrast our findings in each section. Please note all SAS code is included at the end of the assignment in the order in which it is discussed.

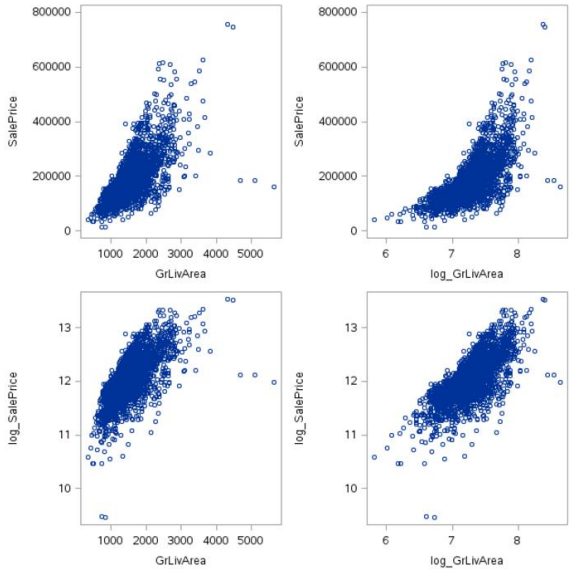
Exploratory Data Analysis (EDA). Before segueing into the initial model generation and comparison, we first need to transform our variables using SAS. The two new variables will be the logs of SalePrice and GrLivArea. For reference, SalePrice is a continuous variable which is the sales price in dollars of a home and GrLivArea is a continuous variable which is the above ground living area in square feet. Using PROC PRINT (obs=5) we generate Table 1 below which provides a snapshot of the newly created variables in the data set.

Table 1: Variables in Data Set

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

By plotting the predictors versus the responses, we notice that the transformed variables create a somewhat more linear pattern, especially in the case of log_SalePrice versus log_GrLivArea (Exhibit 1). We will keep this in mind as we progress through the analysis.

Exhibit 1: Normal and Transformed Variables



Model Transformations and Comparisons. In this section, we generate 4 simple linear regression models, report each model in equation form and interpret each coefficient. Then we construct a summative table for easy comparative analysis.

Model 1: GrLivArea to Predict SalePrice. Using the SAS simple linear regression procedure, we generate the following parameter estimates for Model 1 (Table 2):

Table 2: Parameter Estimates for GrLivArea to Predict SalePrice

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

This results in the following model in equation form:

SalePrice = 13290 + 111.694 x GrLivArea

Within the context of this model, if GrLivArea were equal to 0 then we expect SalePrice to be $13,290. There were no transformations in this model, therefore, for each increase of 1 square foot in GrLivArea we expect the SalePrice to increase by $111.69. This is in line with the expectation that more square footage should cost more money.

Model 2 (Exponential): Log of GrLivArea to Predict SalePrice. Using the SAS simple linear regression procedure, we generate the following parameter estimates for Model 2 (Table 3):

Table 3: Parameter Estimates for Log GrLivArea to Predict SalePrice

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

This results in the following model in equation form:

SalePrice = -1060765 + 171011 x log_GrLivArea

Within the context of this model, if log_GrLivArea were equal to 0 then we expect SalePrice to be -$1,060,765. Only the independent variable in this model is log transformed, therefore, if we increase log_GrLivArea by 1%, we expect SalePrice to increase by $1,710.11 (171,011/100 units of y). This is in line with the expectation that more square footage should cost more money, however, this is the first intercept we've encountered which is negative. This situation is common and generally a result of a strong positive relationship between the X and Y variables. We cannot reasonably expect a relationship identified over a limited range of data to be appropriate across all values, especially when one of the variables is transformed and the other is not.

Model 3 (Logarithmic): GrLivArea to Predict Log of SalePrice. Using the SAS simple linear regression procedure, we generate the following parameter estimates for Model 3 (Table 4):

Table 4: Parameter Estimates for GrLivArea to Predict Log of SalePrice

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

This results in the following model in equation form:

log_SalePrice = 11.17954 + 0.00056107x GrLivArea

Within the context of this model, if GrLivArea were equal to 0 then we expect log_SalePrice to be 11.17954. Only the dependent variable in this model is log transformed, therefore, if we change GrLivArea by 1 square foot, we would expect log_SalePrice to change by 5.61% (100*0.00056107 percent). This is in line with the expectation that more square footage should cost more money, however, note how the scale is altered due to the log transformation of the dependent variable.

Model 4 (Goldilocks): Log of GrLivArea to Predict Log of SalePrice. Using the SAS simple linear regression procedure, we generate the following parameter estimates for Model 4 (Table 5):

Table 5: Parameter Estimates for Log of GrLivArea to Predict Log of SalePrice

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

This results in the following model in equation form:

log_SalePrice = 5.43019 + 0.90781 x log_GrLivArea

Within the context of this model, if log_GrLivArea were equal to 0 then we expect log_SalePrice to be 5.43019. Both independent and dependent variables are log transformed in this model, therefore, if we change log_GrLivArea by 1%, we expect log_SalePrice to change by 0.90781%. This is in line with expectations that more square footage costs more money and note again the scale is altered as a result of the log transformations.

Model Comparisons. Based on the diagnostics from the regression analysis and the automatically generated ODS output, Table 6 below provides a summative table to compare the fit of each of the four models. The metrics we use for comparing the 4 models are Adjusted R-squared, Predicted R-squared and the F-Value. While the Adjusted R-squared is commonly used to compare models with multiple variables, it is still applicable in simple linear regression. The Predicted R-squared indicates how well a regression model predicts responses for new observations. Lastly, the F-Value is used to compare the fits of the 4 different models.

Table 6: Summary Table of Models 1-4

Model	Equation Form	Adjusted R-Squared	Predicted R-Squared	F-Value
1	SalePrice = 13290 + 111.694 x GrLivArea	0.4994	0.4974	2922.59
2	SalePrice = -1060765 + 171011 x log_GrLivArea	0.4829	0.4820	2736.45
3	log_SalePrice = 11.17954 + 0.00056107 x GrLivArea	0.4840	0.4825	2748.89
4	log_SalePrice = 5.43019 + 0.90781 x log_GrLivArea	0.5228	0.5222	3209.97

If we rely purely on the statistical summary provided, we would conclude that Model 4 is the best overall fit. Model 4 accounts for the most variability in SalePrice according to the Adjusted R-squared. We could also expect this model to explain the highest amount of variability in predicting new observations. Lastly, Model 4 has the highest F-Value, indicating that the model is the best fit from the population it was sampled from.

Analyzing the graphical elements of the ODS output supports our earlier observation that the transformation of the independent and dependent variables created a superior linear pattern while also reducing the number of potential outliers from 3 to 2. With the introduction of log transformed variables, plots of the residuals versus the predicted values gradually becomes more randomized, albeit there is still pooling in the center of the plot. However, the obvious wedge shape from the non-transformed variables is eliminated. The Q-Q plots of the transformed variables suggest the residuals are normally distributed, though there remain 2 potential outliers that are causing negative skewness. This is confirmed when the histogram is taken into consideration, where Models 3 and 4 exhibit clear negative skewness. Interestingly, there appears to be more potentially influential cases in Model 4 according to Cook's D than the other models. All of the Fit-Mean Residual plots except for Model 2 (SalePrice and log_GrLivArea) have higher left panes, indicating that the predictor variables account for a meaningful portion of the variation. Finally, the range of values that fall within the 95% prediction limits in the Fit Plot graphs increase as the variables are transformed.

Several concerns arise when using these transformed variables. First, the log transformation appears to have increased the variability of the data. While the number of outliers seems to have decreased, Cook's D suggests more influential points. Second, the transformation exacerbated the negative skewness of the residuals in both Models 3 and 4. The interpretation of the log of SalePrice is different as well. With the price model, we interpret the X variables as the dollar impact on the overall dollar amount of the predicted sale price. In the case of a Y variable transformation, we interpret the coefficient's impact in terms of a percentage change of the predicted sale price. Lastly, its seems that in some cases comparing the variability of original data versus transformed variables is of little use as the scales are completely different.

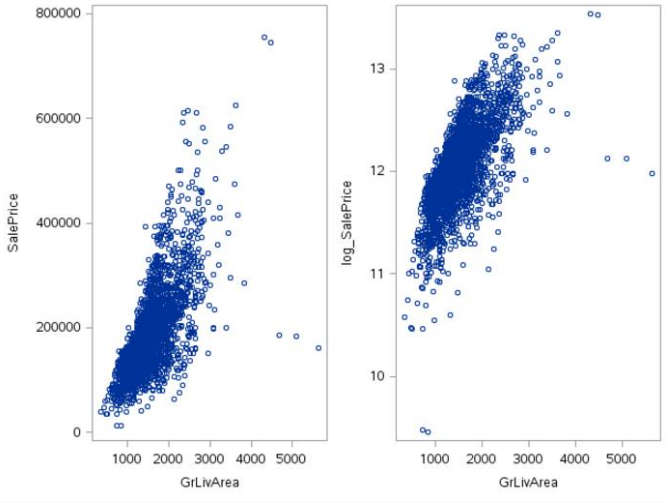
Correlations. To further examine the impact of the transformations, we correlate a sub-set of the continuous variables in the original data set with log_SalePrice. Table 7 below provides these correlations.

Table 7: Correlations with log_SalePrice

Variable	Correlation
GrLivArea	0.696
GarageArea	0.651
TotalBsmtSF	0.625
FirstFlrSF	0.603
MasVnrArea	0.449
BsmtFinSF1	0.411
LotFrontage	0.350
LotArea	0.255
BsmtUnfSF	0.194
PoolArea	0.054

Of the 4 highly correlated variables, we select GrLivArea as it has the strongest correlation with log_SalePrice. Exhibit 2 below provides scatterplots of GrLivArea with SalePrice and log_SalePrice.

Exhibit 2: GrLivArea vs. SalePrice and log_SalePrice



From these visuals, the most obvious observation is the scaling difference in the y axis as a result of the log transformation. While the graph of log_SalePrice exhibits a more positive linear relationship between the two variables, there are still numerous influential outliers. We should conclude that log transformation alone might not generate the ideal model.

Model 5: Alternative Transformations. In this section, we will fit one more regression model using an alternative transformation of the response variable. According to Montegomery, Peck and Vinning (2012), transformations are employed to stabilize variances. Log transformations of dependent variables are effective when the measurements cannot be negative, the model exhibits unequal variances and a curved (i.e. non-linear) relationship exists between the X and Y variables. In the case of

our original equation (Model 1), we can categorically state there is a linear relationship between SalePrice and GrLivArea.

There are many mathematical functions to transform the response variable Y or a predictor variable X. From our earlier regression models, we know that log transformations of left skewed data can increase skewness. As the residuals for Model 1 exhibit moderate negative skewness, we apply the square transformation on the dependent variable to reduce negative skewness. The independent variable, GrLivArea, will not be transformed in this model. Using the SAS simple linear regression procedure, we generate the following parameter estimates for Model 5 (Table 8):

Table 8: Parameter Estimates for GrLivArea to Predict Square of SalePrice

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-4.13149E10	1858691833	-22.23	<.0001
GrLivArea	1	53598922	1174478	45.64	<.0001

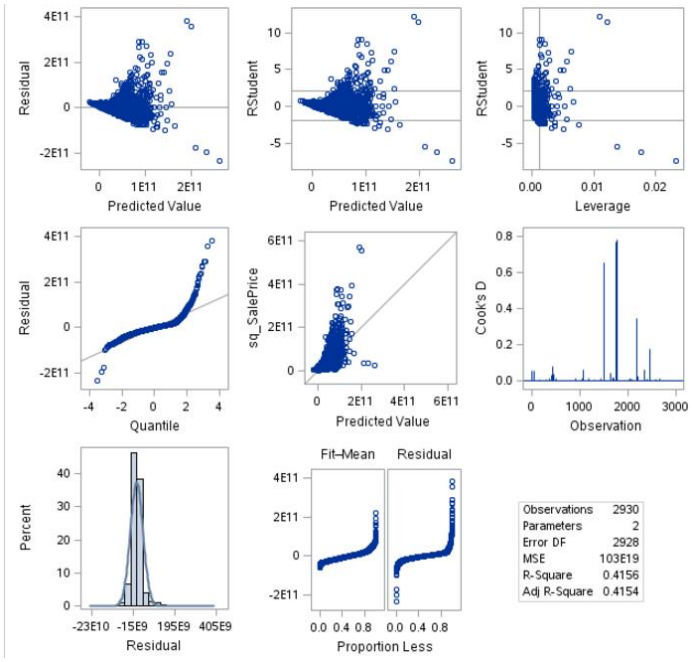
This results in the following model in equation form:

sq_SalePrice = -4.13149E10 + 53598922 x GrLivArea

Within the context of this model, if GrLivArea were equal to 0 then we expect sq_SalePrice to be -4.13149E10. Only the dependent variable in this model is square transformed, therefore, if we change GrLivArea by 1 square foot, we would expect sq_SalePrice to be positively impacted by a large factor. This is in line with the expectation that more square footage should cost more money, however, the squared transformation makes the coefficient difficult to interpret.

Studying the automatically generated ODS output (Exhibit 3) from SAS we assess the goodness-of-fit. The scatterplot for the residual values appears to be concentrated and wedge-shaped; not entirely random. The Q-Q plot of the residuals suggests a non-normal distribution. The presence of multiple possible outliers is also apparent. The histogram of the residuals suggests a somewhat normal distribution with a slight positive skewness. Cook's D denotes the presence of multiple possible outliers and numerous smaller influential points. The left pane of the Fit-Mean Residual plot is much lower than the right, indicating that there is a large amount of variation not explained by the model. From these observations, we can conclude that this model is not homoscedastic.

Exhibit 3: Fit Diagnostics for sq_SalePrice Using GrLivArea



Model Comparison: Alternative Transformations. Based on the output from the regression analysis and the automatically generated ODS output, Table 9 below provides a summative table to compare the fit of each of the five models. We compare the models based on the same criteria used earlier.

Table 9: Summary Table of Models 1-5

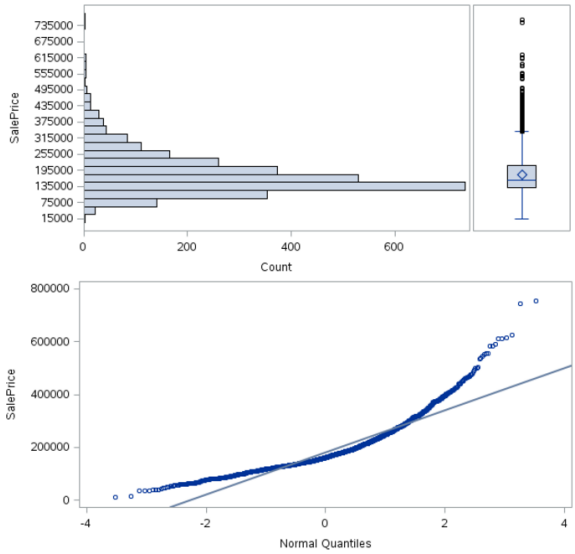
Model	Equation Form	Adjusted R-Squared	Predicted R-Squared	F-Value
1	SalePrice = 13290 + 111.694 x GrLivArea	0.4994	0.4974	2922.59
2	SalePrice = -1060765 + 171011 x log_GrLivArea	0.4829	0.4820	2736.45
3	log_SalePrice = 11.17954 + 0.00056107 x GrLivArea	0.4840	0.4825	2748.89
4	log_SalePrice = 5.43019 + 0.90781 x log_GrLivArea	0.5228	0.5222	3209.97
5	sq_SalePrice = -4.13149E10 + 53598922 x GrLivArea	0.4154	0.4120	2082.68

From a statistical summary standpoint, Model 5 is an inferior solution when compared to Model 1 through 4. As we also discovered, Model 5's problems are multi-faceted from a goodness of fit standpoint, demonstrating that variable transformation and model building, in general, is a very iterative process. The best model within the context of this exercise thus far remains Model 4 where both X and Y variables are log transformed; also known as the Goldilocks model.

Outlier Identification and Removal. In the second portion of this assignment, we shift our focus to identifying and working with outliers using SAS. To accomplish this, we start by identifying observations that are potential outliers for the SalePrice variable, Y. Using the SAS univariate procedure, we can begin

identifying outliers as well as extreme observations. Exhibit 4 below provides a histogram, boxplot and QQ plot for SalePrice.

Exhibit 4: Distribution and Probability Plot for SalePrice



The histogram in Exhibit 4 suggests the distribution of SalePrice is positively skewed (right) and peaked. This corroborated with the boxplot. Lastly, the Q-Q plot shows there is a systematic pattern of progressive departure from normality as both ends curve up, signifying positive skewness. Investigating this further using the Moments table from the ODS output, a skewness of 1.74 and kurtosis of 5.19 confirms departures from normality. The distribution is positively skewed with a leptokurtic shape (i.e. peaked). Each graphic also indicates the existence of possible outliers, especially at large SalePrice values; contributing to the skewness. This supports our earlier findings using the Cook's D and fit plots.

Using PROC SQL, we take the analysis a step further and inspect the individual records of concern. The smallest and largest 5 records are presented in Table 10 below.

Table 10: Bottom/Top 5 SalePrice Values

Observation	SalePrice
182	12789
1554	13100
727	34900
2844	35000
2881	35311
45	611657
1064	615000
2446	625000
1761	745000
1768	755000

The ODS output also provides a quantile breakdown. These figures are provided in Table 11 below.

Table 11: Quantiles

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

The definition of an outlier is relatively vague and can be subjective especially when relying solely on visual cues, so to formally identify potential extreme points in our sample we need to establish a criteria. For purposes of this study, we will use the interquartile rule for outliers where an outlier is defined as any point that is over 1.5x IQRs (interquartile range) below the first quartile or above the third quartile. Alternatively, we could use the 2 standard deviation rule, which was a common approach before advances in modern computing technology. This rule states that outliers are defined as any point outside of +/- 2 standard deviations from the mean. It also assumes a normal distribution. As we have the luxury of powerful analytical software and what we've established is a non-normal distribution, we elect to use the interquartile rule. This is written in equation form as follows:

High: Q3 + 1.5 x IQR

Low: Q1 - 1.5 x IQR

Fortunately, the interquartile range, $84,000, is provided in the ODS output. Per the equations above, the low limit is $3,500 and the high limit is $339,500. Even before performing further calculations, we can already deduce given Tables 10 and 11 that we have several large values that qualify as outliers while no values violate the lower limit. Next, we will bucket the data into three separate tranches using SAS to create an if-then-else ladder to assign each observation to a category. The SAS coding syntax for our indicators are 1 (low outlier), 2 (not outlier) and 3 (high outlier). Table 12 shows the frequency of the outliers; as suspected, there are no values that are below the lower threshold while 4.68% of the sample (137 observations) are above the upper threshold. It important to note that if this were for a professional activity, we would be conducting similar analysis on all variables in the data set. However, this is outside of the scope of this project.

Table 12: Frequency of SalePrice Outliers

price_outlier	Frequency	Percent	Cumulative Frequency	Cumulative Percent
2	2793	95.32	2793	95.32
3	137	4.68	2930	100

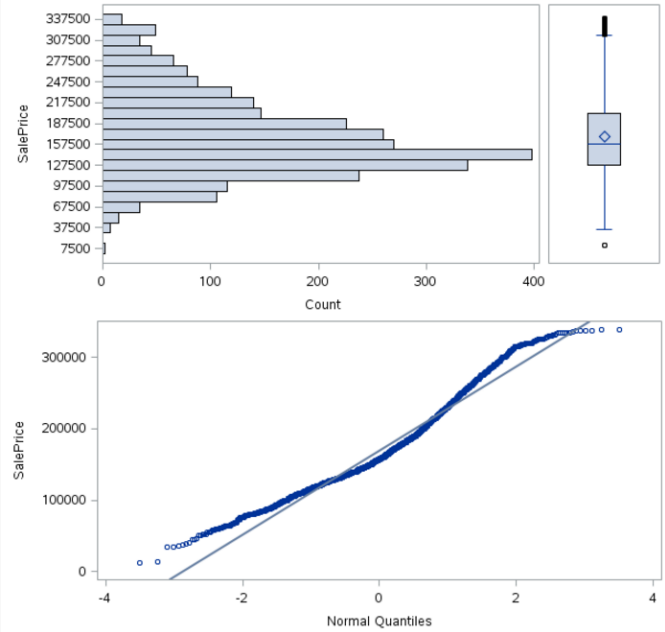
Suggesting the removal of roughly 5% of the sample necessitates us exploring more about this potential impact. Table 13 provides the mean, median, standard deviation, kurtosis, skewness, range, minimum and maximum values by outlier indicator. We take each one of these into consideration as our ultimate the goal is to improve the regression model with variables that are more normally distributed and have a linear relationship to the predictors. Bifurcating the sample by price_outlier grants us insight into what the distribution would look like with the removal of the identified outliers. Positive skewness is evident in the higher outlier portion of the sample. The standard deviation is also notably higher in the outliers, as is the range and kurtosis across a much smaller number of data points. The portion of the sample deemed not outliers appears to be largely normally distributed. Given these observations, we posit that the regression model will be improved by the removal of the 137 outliers.

Table 13: Summary Table for price_outliers

	price_outlier =2 (Not Outlier)	price_outlier =3 (High Outlier)
Mean	169,115.50	418,926.01
Median	157,500.00	394,617.00
Std Dev	58,989.05	77,999.88
Kurtosis	0.09	4.23
Skewness	0.67	1.90
Range	326,142.00	415,250.00
Minimum	12,789.00	339,750.00
Maximum	338,931.00	755,000.00

We create a new data set by removing the outliers previously identified. Exhibit 5 below provides a histogram, boxplot and Q-Q plot for our new SalePrice sample. The removal of the outliers appears to have generally improved normality, however, the Q-Q plot suggests a light tailed distribution. There also seems to be a handful of large values but no evidence of outliers or other extreme points.

Exhibit 5: Distribution and Probability Plot for SalePrice (No Outliers)



Model Comparison on Modified Data Set. In this final section, we use the newly cleaned data set to refit a select number regression models. The models we will be refitting come from the previous assignment. The regression results of the refitted models will be compared to the original output. The nomenclature $SalePrice_{New}$ is used to represent the new data set. The original models from the previous assignment are designated Models 6, 7 and 8 while their refitted forms are Models 9, 10 and 11, respectively. No variables in the following models have been transformed. Overall, we expect a better fit given the removal of the outliers. The original models from the previous assignment are:

Model 6: $SalePrice = \beta_0 + \beta_1 GrLivArea$

Model 7: $SalePrice = \beta_0 + \beta_1 MasVnrArea + \beta_2 GrLivArea$

Model 8: $SalePrice = \beta_0 + \beta_1 MasVnrArea + \beta_2 GrLivArea + \beta_3 BsmtFinSF2$

Model 9: Simple Linear Regression. Using the SAS simple linear regression procedure, we generate the following parameter estimates (Table 14):

Table 14: Parameter Estimates for GrLivArea to Predict $SalePrice_{New}$

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	47500	2805.88096	16.93	<.0001
GrLivArea	1	83.58790	1.83872	45.46	<.0001

This results in the following model in equation form:

$SalePrice_{New}$ = 47500 + 83.58790 x GrLivArea

Within the context of this model, if GrLivArea were equal to 0 then we expect the SalePrice to be $47,500. Additionally, for every increase of 1 square foot in GrLivArea we expect the SalePrice to increase by $83.59. This is in line with the expectation that more square footage should cost more money.

Model 10: Multiple Linear Regression (2 Independent Variables). Using the SAS simple linear regression procedure, we generate the following parameter estimates (Table 15):

Table 15: Parameter Estimates for MasVnrArea and GrLivArea to Predict $SalePrice_{New}$

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	51693	2760.03246	18.73	<.0001
MasVnrArea	1	66.54310	5.58584	11.91	<.0001
GrLivArea	1	76.57315	1.88567	40.61	<.0001

This results in the following model in equation form:

$SalePrice_{New}$ = 51693+ 66.54310 x MasVnrArea + 76.57315x GrLivArea

Within the context of this model, if MasVnrArea and GrLivArea were equal to 0 then we expect the SalePrice to be $51,693. If GrLivArea is fixed, then for each increase of 1 square foot in MasVnrArea, $SalePrice_{New}$ will increase by $66.54. Additionally, if MasVnrArea is fixed, then for each increase of 1 square foot in GrLivArea, $SalePrice_{New}$ will increase by $76.57. Both coefficients are positively associated with the response which intuitively makes sense as more square footage should cost more money.

Model 11: Multiple Linear Regression (3 Independent Variables). Using the SAS simple linear regression procedure, we generate the following parameter estimates (Table 16):

Table 16: Parameter Estimates for MasVnrArea, GrLivArea and BsmtFinSF2 to Predict $SalePrice_{New}$

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	51150	2776.22907	18.42	<.0001
MasVnrArea	1	66.54487	5.58288	11.92	<.0001
GrLivArea	1	76.61361	1.88545	40.63	<.0001
BsmtFinSF2	1	9.91782	4.91478	2.02	0.0437

This results in the following model in equation form:

$SalePrice_{New}$ = 51150+ 66.54487 x MasVnrArea + 76.61361 x GrLivArea + 9.91782 x BsmtFinSF2

Within the context of this model, if MasVnrArea, GrLivArea and BsmtFinSF2 were equal to 0 then we expect the SalePrice to be $51,150. If GrLivArea and BsmtFinSF2 are fixed, then for each increase of 1 square foot in MasVnrArea, $SalePrice_{New}$ will increase by $66.54. Additionally, if MasVnrArea and BsmtFinSF2 are fixed, then for each increase of 1 square foot in GrLivArea, $SalePrice_{New}$ will increase by $76.61. Lastly, if MasVnrArea and GrLivArea are fixed, then for each increase of 1 square foot in BsmtFinSF2, $SalePrice_{New}$ will increase by $9.92. All coefficients are positively associated with the response which intuitively makes sense as more square footage should cost more money. There are likely few circumstances where GrLivArea would equal 0 but it's possible for MasVnrArea and BsmtFinSF2 to equal 0 if the home doesn't have a masonry veneer or a basement.

Model Comparison: Refitted Models. Based on the output from the regression analysis and the automatically generated ODS output, Table 17 below provides a summative table to compare the fit of the original and refitted models. The models are arranged in order by original and refitted for ease of assessment. As a point of clarification Model 1 and 6 in this assignment are identical. We compare the models based on the same criteria used earlier.

Table 17: Summary Table of Original and Refitted Models

Model	Equation Form	Adjusted R-Squared	Predicted R-Squared	F-Value
6	SalePrice = 13290+ 111.69400 x GrLivArea	0.4994	0.4974	2922.59
9	$SalePrice_{New}$ = 47500 + 83.58790 x GrLivArea	0.4252	0.4231	2066.60
7	SalePrice = 26547+ 118.54695 x MasVnrArea + 94.60302 x GrLivArea	0.5596	0.5567	1846.98
10	$SalePrice_{New}$= 51693+ 66.54310 x MasVnrArea + 76.57315 x GrLivArea	0.4541	0.4509	1153.35
8	SalePrice = 25997+ 118.65010 x MasVnrArea + 94.62084 x GrLivArea + 10.46253 x BsmtFinSF2	0.5597	0.5567	1232.01
11	$SalePrice_{New}$ = 51150+ 66.54487 x MasVnrArea + 76.61361 x GrLivArea + 9.91782 x BsmtFinSF2	0.4544	0.4511	769.97

Based on the specified criteria, all of the models (9, 10 and 11) using the modified data set are inferior to their original forms (6, 7 and 8). The Adjusted and Predicted R-squared values were less negatively impacted in the simple linear regression models (6 and 9) while a ~10% decrease in these metrics is observed between Models 7 and 10 and Models 8 and 11. The F-Values were also lower in every refitted model. Overall, removal of the outliers did not improve fit. This is contrary to expectations.

Conclusions:

Building and validating regression models is a cornerstone of predictive analytics. For this exercise, the Ames, Iowa data set was used to construct simple and multiple regression models. As part of the construction and validation process, we analyzed the impacts of variable transformations and removal of outliers on these models. We primarily used the Adjusted R-squared, Predicted R-squared, and F-Value statistics to compare our models but recognize that these are not the sole determinants in the selection process. Transformations and outlier deletion are intended to improve the robustness of models and can have powerful effects on fit. While the effects are sometimes not in the desired direction, as this week's exercise demonstrated, these analytical activities can be beneficial if they improve the linear relationship between two or more variables.

Assignments #1 and #2 were valuable introductions to data manipulation using SAS, interpreting statistical outputs to identify potential predictor variables and initial model building. This week's assignment expounded upon this by introducing variable transformations and outlier removal, two key tools in predictive modeling. While these practices can seem more of an art than a science at times, especially in the case of transformations, it's important to understand what techniques (squared roots, logs, etc.) are applicable to the data and when they should be implemented. Modifying data just because you can does not make for effective predictions; every action should be supported by logic. This

exercise also reinforced the importance of data visualization in validation activities, especially when attempting to transform X and Y variables. As this exercise focused on continuous variables, the next steps in the model building process would be to investigate categorical variables followed by estimation of the model, validation, production of a final result, communication with business users, implementation and performance monitoring.

Reference(s):

Montgomery, D.C., Peck, E.A., and Vining, G.G. (2012). Introduction to Linear Regression Analysis. (5th Edition). New York, NY: Wiley

Code:

```
libname mydata "/scs/wtm926/"   access=readonly;


proc datasets library=mydata; run;


Data temp1;
        set mydata.ames_housing_data;


/* PART A - Transformations - Comparisons of Y versus Log(Y) */;


/**Question 1**/;


Data temp2;
        set temp1;
        log_SalePrice = log(SalePrice);
        log_GrLivArea = log(GrLivArea);
        keep GrLivArea log_GrLivArea SalePrice log_SalePrice MasVnrArea BsmtFinSF2;


Proc Print data=temp2 (obs=5);
run; /***Table 1***/;


Proc sgscatter data=temp2;
        plot (SalePrice log_SalePrice) * (GrLivArea log_GrLivArea);
run; /***Exhibit 1***/;
```

```
/**Question 2**/;



proc reg data=temp2;
                    model SalePrice = GrLivArea ;
                    model SalePrice = log_GrLivArea;
                    model log_SalePrice = GrLivArea ;
                    model log_SalePrice = log_GrLivArea;
run; /***Tables 2,3,4,5***/;


/*** Table 6 Data***/;


/*** Model 1: Predicted R-Squared***/;
proc reg data=temp2 outest=outest plots=none;
model SalePrice = GrLivArea / rsquare press sse adjrsq;
run;
quit;


data outestPlus;
set outest;
_PRSQ_ = 1 - _PRESS_ * (1 - _RSQ_)/_SSE_;
label _PRSQ_ = "Predicted r-squared";
run;
```

```
proc print data=outestPlus label;
var _RSQ_ _ADJRSQ_ _PRSQ_;
run;

/*** Model 2: Predicted R-Squared***/;

proc reg data=temp2 outest=outest plots=none;
model SalePrice = log_GrLivArea / rsquare press sse adjrsq;
run;
quit;

data outestPlus;
set outest;
_PRSQ_ = 1 - _PRESS_ * (1 - _RSQ_)/_SSE_;
label _PRSQ_ = "Predicted r-squared";
run;

proc print data=outestPlus label;
var _RSQ_ _ADJRSQ_ _PRSQ_;
run;

/*** Model 3: Predicted R-Squared***/;

proc reg data=temp2 outest=outest plots=none;
model log_SalePrice = GrLivArea / rsquare press sse adjrsq;
```

```
run;

quit;


data outestPlus;

set outest;

_PRSQ_ = 1 - _PRESS_ * (1 - _RSQ_)/_SSE_;

label _PRSQ_ = "Predicted r-squared";

run;


proc print data=outestPlus label;

var _RSQ_ _ADJRSQ_ _PRSQ_;

run;


/*** Model 4: Predicted R-Squared***/;


proc reg data=temp2 outest=outest plots=none;

model log_SalePrice = log_GrLivArea / rsquare press sse adjrsq;

run;

quit;


data outestPlus;

set outest;

_PRSQ_ = 1 - _PRESS_ * (1 - _RSQ_)/_SSE_;

label _PRSQ_ = "Predicted r-squared";

run;
```

```
proc print data=outestPlus label;

var _RSQ_ _ADJRSQ_ _PRSQ_;

run;



/**Question 3**/;



data temp3;

        set mydata.ames_housing_data;

        log_SalePrice = log(SalePrice);

        log_GrLivArea = log(GrLivArea);

        keep PoolArea LotFrontage LotArea MasVnrArea BsmtUnfSF BsmtFinSF1 FirstFlrSF TotalBsmtSF
GarageArea GrLivArea log_GrLivArea SalePrice log_SalePrice;

proc print data=temp3 (obs=5);


/*** Table 7 Data***/;


proc corr data=temp3 nosimple rank;

        var log_SalePrice;

        with PoolArea LotFrontage LotArea GrLivArea MasVnrArea BsmtUnfSF BsmtFinSF1 FirstFlrSF
TotalBsmtSF GarageArea;

run;


/*** Exhibit 2 Data***/;
```

```
Proc sgscatter data=temp3;
	plot (SalePrice log_SalePrice) * (GrLivArea);
run; /***Exhibit 2***/;



/**Question 4**/;



data temp4;
	set mydata.ames_housing_data;
	sqrt_SalePrice = sqrt(saleprice);
	sq_SalePrice = saleprice * saleprice;
	log_GrLivArea = log(GrLivArea);
run;

proc print data=temp4 (obs=5);

/***Model 5***/;

proc reg data=temp4;
		model sq_SalePrice = GrLivArea;
run; /***Table 8 and Exhibit 3***/;

/***Model 5 Predicted R Squared***/;
```

```
/***Table 9***/;


proc reg data=temp4 outest=outest plots=none;
model sq_SalePrice = GrLivArea/ rsquare press sse adjrsq;
run;
quit;


data outestPlus;
set outest;
_PRSQ_ = 1 - _PRESS_ * (1 - _RSQ_)/_SSE_;
label _PRSQ_ = "Predicted r-squared";
run;


proc print data=outestPlus label;
var _RSQ_ _ADJRSQ_ _PRSQ_;
run;


proc sgscatter data=temp4;
        plot (saleprice sq_SalePrice) * grlivarea;
run;



/* PART B - Outliers */;
```

```
/**Question 5**/;

/***Exhibit 4 and Table 11***/;

proc univariate normal plot data=temp1;
        var saleprice;
run;

/***Table 10***/;

/***Smallest 5 SalePrice***/;

proc sql number outobs=5;
create table jj as
select sid,saleprice from temp1 group by saleprice order by saleprice asc;
quit;

/***Largest 5 SalePrice***/;

proc sql number outobs=5;
create table kk as
select sid,saleprice from temp1 group by saleprice order by saleprice descending;
quit;
```

```
/***Interquartile Ranges***/;

Data Part5;
set temp1;
keep saleprice grlivarea price_outlier MasVnrArea BsmtFinSF2;
if saleprice = . then delete;

if saleprice <= 3500 then price_outlier = 1;
else if saleprice > 3500 & saleprice < 339500 then price_outlier = 2;
else if saleprice >= 339500 then price_outlier = 3;

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

/***Table 12***/;

proc freq data=part5;
                tables price_outlier;
        run;

proc sort data=part5;
                by price_outlier;
        run;

/****Confirm Ranges worked****/;
```

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

/****Table 13****/;

proc summary print mean median std kurtosis skewness range min max;
	by price_outlier;
	var SalePrice;
run;

/****Removal of outliers****/;
data part6;
	set part5;
	if price_outlier = 1 then delete;
	if price_outlier = 3 then delete;
	keep SalePrice price_outlier GrLivArea MasVnrArea BsmtFinSF2;
run;

/***Exhibit 5***/;

proc univariate normal plot data=part6;
	var SalePrice;
```

```
        histogram saleprice / normal;



/**Question 6**/;


/***For Table 17***/;


/***Model 9 Refitted SLR***/;
proc reg data=part6;
                model SalePrice = GrLivArea;
run; /****Table 14 ****/;



/***Model 10 Refitted MLR 2 variables***/;
proc reg data=part6;
        model saleprice = MasVnrArea GrLivArea;
run; /****Table 15 ****/;


/***Model 11 Refitted MLR 3 variables***/;
proc reg data=part6;
        model saleprice = MasVnrArea GrLivArea BsmtFinSF2 ;
run; /***Table 16 ***/;



/**** Model 6: Diagnostics****/;
```

```
proc reg data=temp2 outest=outest plots=none;
model SalePrice = GrLivArea / rsquare press sse adjrsq;
run;
quit;

data outestPlus;
set outest;
_PRSQ_ = 1 - _PRESS_ * (1 - _RSQ_)/_SSE_;
label _PRSQ_ = "Predicted r-squared";
run;

proc print data=outestPlus label;
var _RSQ_ _ADJRSQ_ _PRSQ_;
run;



**** Model 9: Diagnostics****/;
proc reg data=part6 outest=outest plots=none;
model SalePrice = GrLivArea / rsquare press sse adjrsq;
run;
quit;

data outestPlus;
set outest;
_PRSQ_ = 1 - _PRESS_ * (1 - _RSQ_)/_SSE_;
```

```
label _PRSQ_ = "Predicted r-squared";
run;

proc print data=outestPlus label;
var _RSQ_ _ADJRSQ_ _PRSQ_;
run;



/**** Model 7: Diagnostics****/;
proc reg data=temp2 outest=outest plots=none;
model SalePrice = GrLivArea MasVnrArea / rsquare press sse adjrsq;
run;
quit;

data outestPlus;
set outest;
_PRSQ_ = 1 - _PRESS_ * (1 - _RSQ_)/_SSE_;
label _PRSQ_ = "Predicted r-squared";
run;

proc print data=outestPlus label;
var _RSQ_ _ADJRSQ_ _PRSQ_;
run;
```

```
/**** Model 10: Diagnostics****/;
proc reg data=part6 outest=outest plots=none;
model SalePrice = GrLivArea MasVnrArea  / rsquare press sse adjrsq;
run;
quit;

data outestPlus;
set outest;
_PRSQ_ = 1 - _PRESS_ * (1 - _RSQ_)/_SSE_;
label _PRSQ_ = "Predicted r-squared";
run;

proc print data=outestPlus label;
var _RSQ_ _ADJRSQ_ _PRSQ_;
run;



/**** Model 8: Diagnostics****/;
proc reg data=temp2 outest=outest plots=none;
model SalePrice = GrLivArea MasVnrArea BsmtFinSF2 / rsquare press sse adjrsq;
run;
quit;

data outestPlus;
set outest;
```

```
_PRSQ_ = 1 - _PRESS_ * (1 - _RSQ_)/_SSE_;
label _PRSQ_ = "Predicted r-squared";
run;

proc print data=outestPlus label;
var _RSQ_ _ADJRSQ_ _PRSQ_;
run;



/**** Model 11: Diagnostics****/;
proc reg data=part6 outest=outest plots=none;
model SalePrice = GrLivArea MasVnrArea BsmtFinSF2 / rsquare press sse adjrsq;
run;
quit;

data outestPlus;
set outest;
_PRSQ_ = 1 - _PRESS_ * (1 - _RSQ_)/_SSE_;
label _PRSQ_ = "Predicted r-squared";
run;

proc print data=outestPlus label;
var _RSQ_ _ADJRSQ_ _PRSQ_;
run;
```