Assignment #3
Crystal M. Mosley

Introduction:

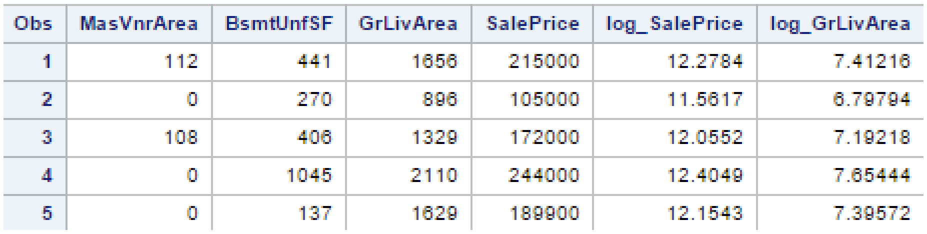
The purpose of this next task for the EDA of Ames, Iowa housing data set is to compare and contrast different data models with log transformations. First we'll conduct some transformations and only keep the variables that are of interest for this analysis. We'll then compare each model for SalePrice with the best predictor variable, GrLivArea. We'll also run correlation procedures for the log transformed SalePrice model.

Once we have completed the above steps, and outliers exist, we'll need to run a cleansing of the data set, and compare and contrast the differences between the non-manipulated data versus the manipulated data.

Results:

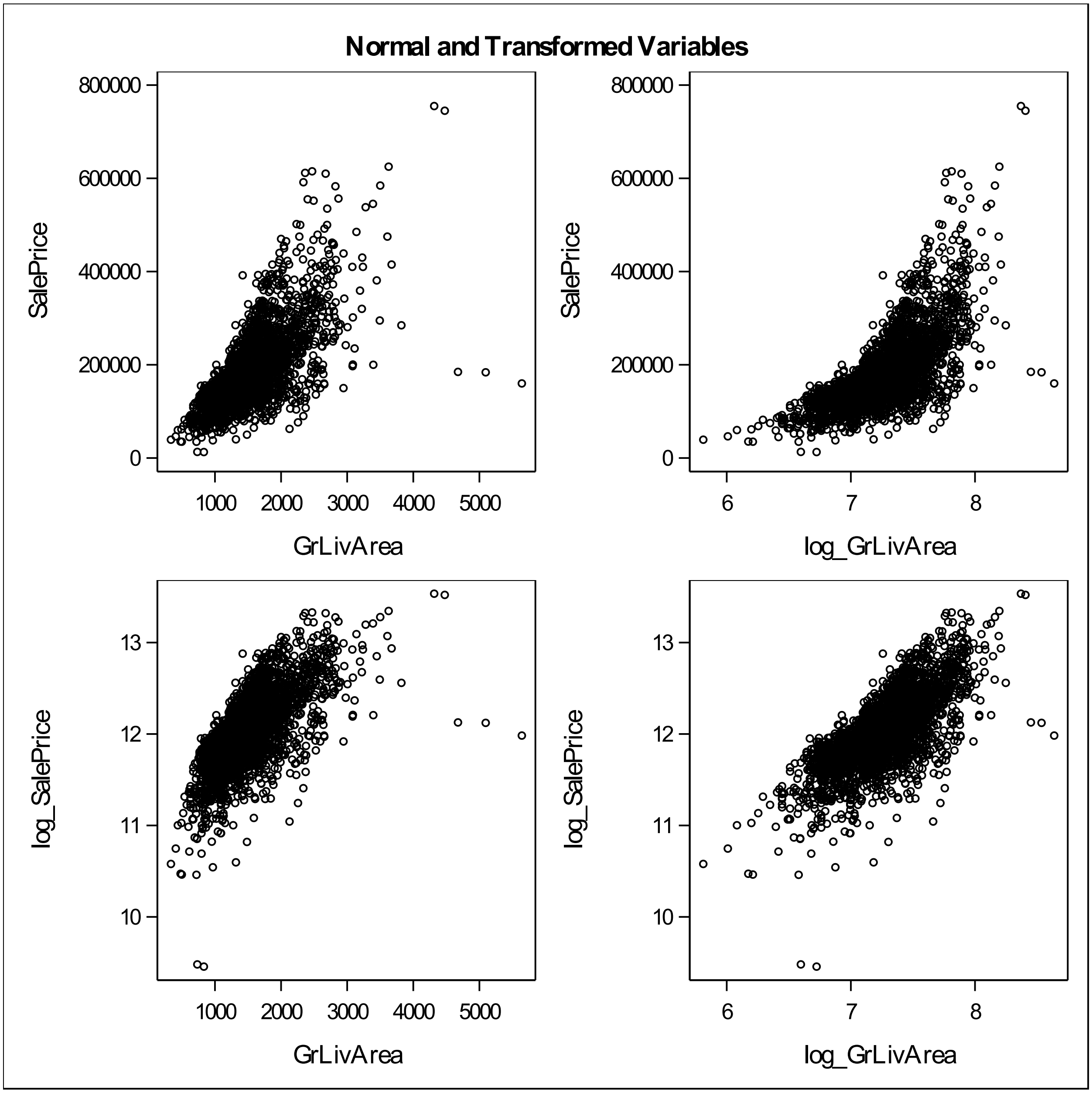
Transformations – Comparison of Y versus Log(Y)

Variables within this data set of interest:



Obs	MasVnrArea	BsmtUnfSF	GrLivArea	SalePrice	log_SalePrice	log_GrLivArea
1	112	441	1656	215000	12.2784	7.41216
2	0	270	896	105000	11.5617	6.79794
3	108	406	1329	172000	12.0552	7.19218
4	0	1045	2110	244000	12.4049	7.65444
5	0	137	1629	189900	12.1543	7.39572

Next, we'll review the transformed variables in relation to SalePrice.



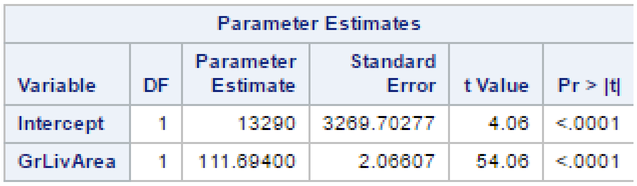
All combinations are showing a positive linearity for the most part.

Model Comparisons

Next steps are to compare each of the four models:

$$SalePrice = \beta_0 + \beta_1 GrLivArea + \varepsilon$$
$$SalePrice = \beta_0 + \beta_1 log\_GrLivArea + \varepsilon$$
$$Log\_SalePrice = \beta_0 + \beta_1 GrLivArea + \varepsilon$$
$$Log\_SalePrice = \beta_0 + \beta_1 log\_GrLivArea + \varepsilon$$

Model Equation: SalePrice = 13290 + 111.694 x GrLivArea

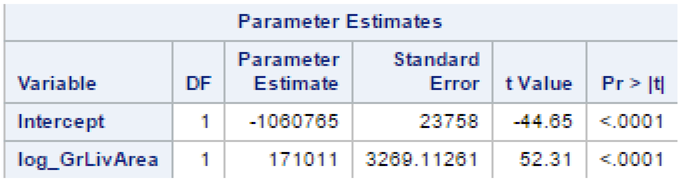


Parameter Estimates

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

The independent variable is not transformed in this model. A 1% increase in the independent variable would result in an average change in the mean of the *dependent* variable by 111.694.

Model Equation: SalePrice = 171011 x log_GrLivArea - 1060765

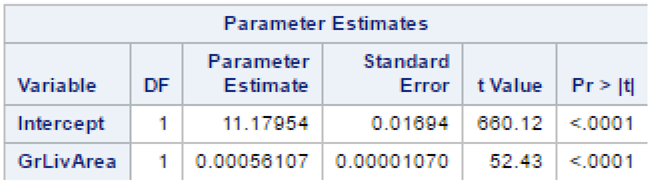


Parameter Estimates

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

The independent variable is log-transformed in this model. A 1% increase in the independent variable would result in an average change in the mean of the dependent variable by 1710.11%. [Note: 171011/100]

Model Equation: log_SalePrice = 11.17954 + 0.00056107 x GrLivArea

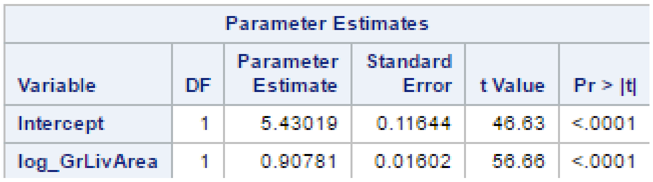


Parameter Estimates

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

The dependent variable is log-transformed in this model. A 1% increase in the independent variable would result in an average change in the mean of the dependent variable by 0.056107%. [Note: 0.00056107x100]

Model Equation: log_SalePrice = 5.43019 + 0.90781 x log_GrLivArea



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

Both the dependent and independent variables are log-transformed in this model. A 1% increase in the independent variable would result in an average change in the mean of the dependent variable by 0.90781%.

Comparison of all models via Adj R-Square and F values:

Model	Adj R-Square	F Value
SalePrice = 13290 + 111.694 x GrLivArea	0.4994	2922.59
SalePrice = 171011 x log_GrLivArea - 1060765	0.4829	2736.45
log_SalePrice = 11.17954 + 0.00056107 x GrLivArea	0.4840	2748.89
log_SalePrice = 5.43019 + 0.90781 x log_GrLivArea	0.5228	3209.97

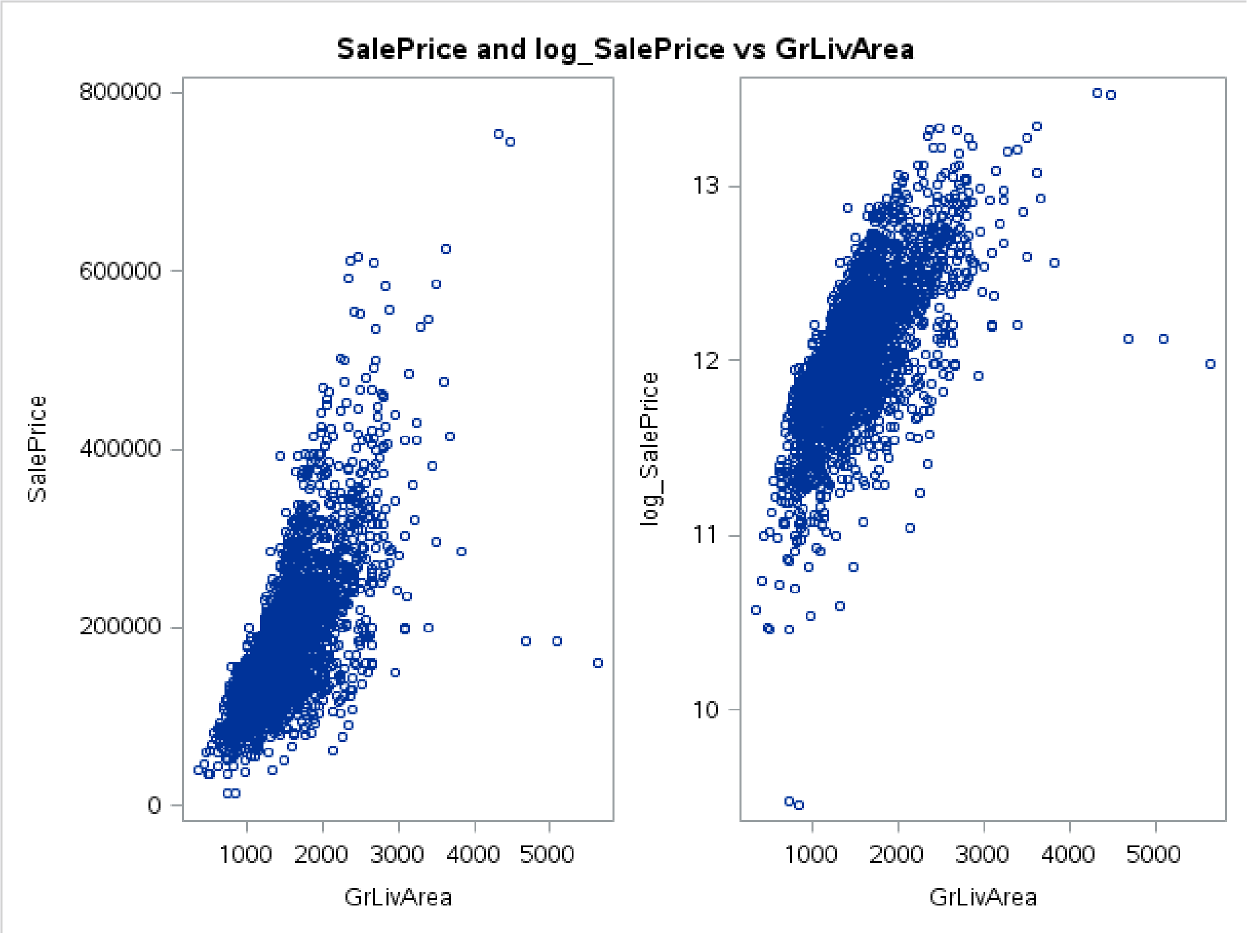
In reviewing the above criteria, the model that was 'best', based on explanation of variability in SalePrice would be the model that uses a log-transform of both SalePrice and GrLivArea. The F Value for the independent and dependent variable from this log-transformed model also shows that the population from which the data were sampled is the best. Each model graphically shows that the transformation of both dependent and independent variables looked to be much more linear than the alternatives. The log-transformed dependent and independent variable model looks to be lightly tailed, but appears to be more normal than the other models. Outliers *do* exist with this model, so those should be noted for concern and possible investigation. This model may have powerful explanatory power, however its current fit certainly incorporates some outliers from the data set.

Correlations for log_SalePrice

After running the corr procedure on the continuous variables, the following variables correlate with log_SalePrice: GrLivArea, GarageArea, TotalBsmtSF, FirstFlrSF, and MasVnrArea

Variable	Correlation Coefficient	Prob > |r| under H_0: ρ=0	Num. Observations
GrLivArea	0.69586	<.0001	2930
GarageArea	0.65113	<.0001	2929
TotalBsmtSF	0.62510	<.0001	2929
FirstFlrSF	0.60263	<.0001	2930
MasVnrArea	0.44861	<.0001	2907

Next, examine how the GrLivArea variable looks with the non-transformed and log-transformed SalePrice variable:



Immediately there's a noticeable difference in the scaling after the transformation. The same amount of outliers is still visible; however, the transformed graph makes these outliers seem less. Making bad data appear good, or doing transformations specifically for visualization purposes would not be reason enough to accept the transform.

Transformation of variable on SalePrice, utilizing square root

Since the relationship between SalePrice and GrLivArea are linear, we'll create a transformation on the dependent variable:

$$\sqrt{SalePrice} = \beta_0 + \beta_1 GrLivArea + \varepsilon$$

We'll also compare the below models:

$$\sqrt{SalePrice} = \beta_0 + \beta_1 GrLivArea + \varepsilon$$

$$SalePrice = \beta_0 + \beta_1 GrLivArea + \varepsilon$$

Lastly, we'll use the best performing model from the previous steps:

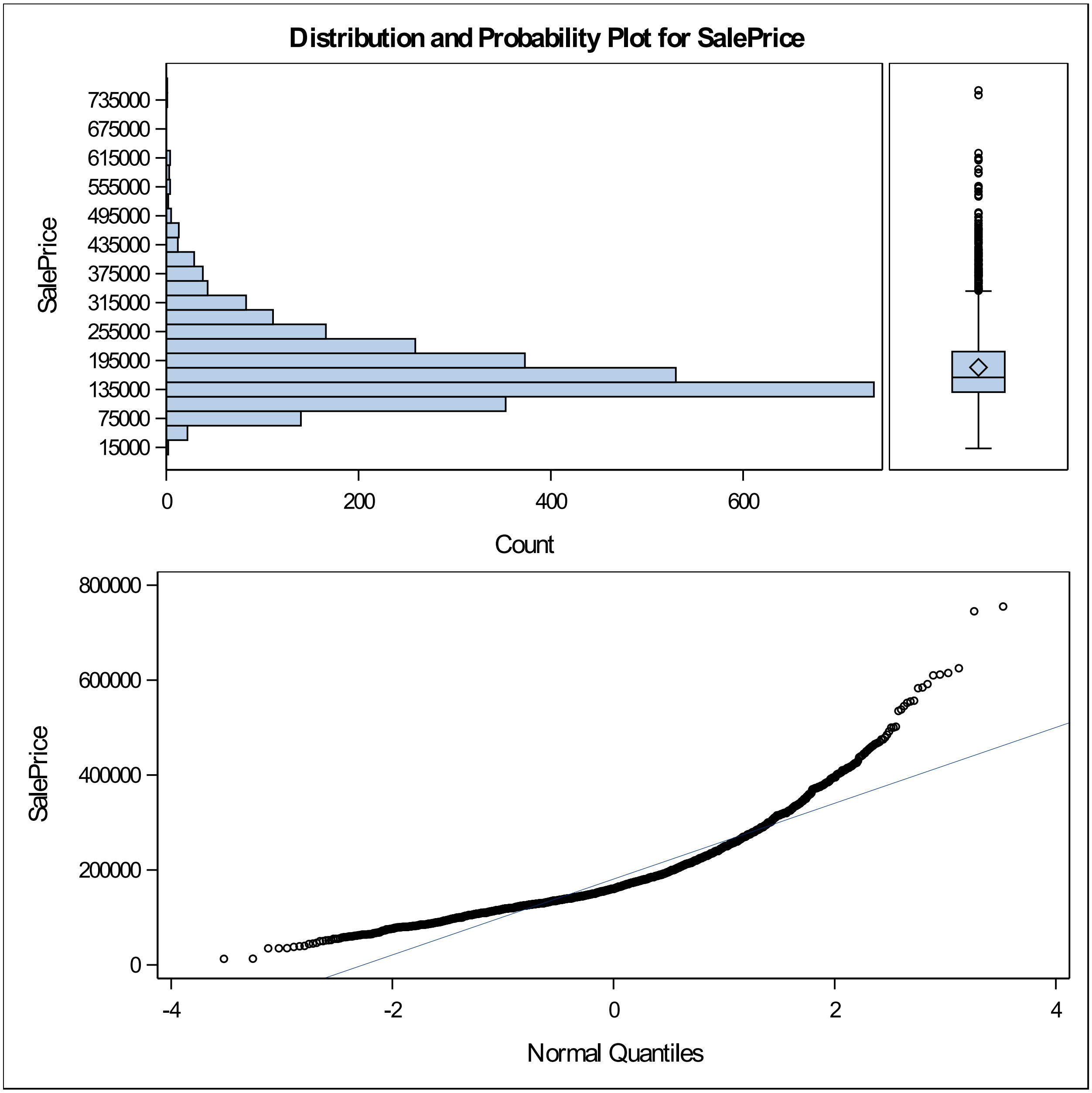
$$log\_SalePrice = 5.43019 + 0.90781 \times log\_GrLivArea$$

Model	Adj R-Square	F Value
SalePrice = 13290 + 111.694 x GrLivArea	0.4994	2922.59
√SalePrice = 232.93209 + 0.12225 x GrLivArea	0.5073	3017.28
log_SalePrice = 5.43019 + 0.90781 x log_GrLivArea	0.5228	3209.97

The chosen transform of the dependent variable shows to be a good choice due to approximate symmetry and homoscedasticity of the residuals. This transformation performed better than the non-re-expressed model; however, it performs slightly worse than the log transformed model from above.

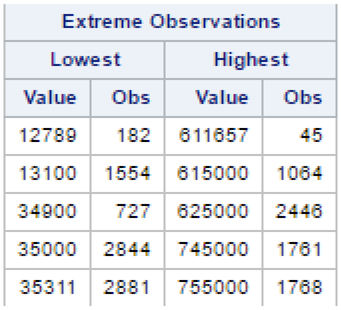
Outliers – Identify observations that are potential outliers for the SalePrice variable

Utilize univariate procedure to identify outliers and extreme observations —



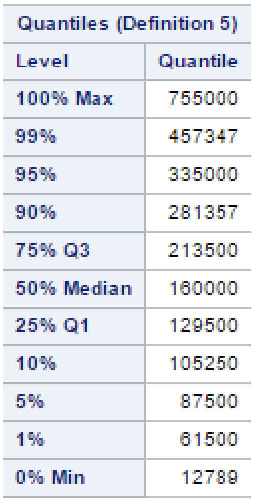
A single outlier is visible at the top of the histogram. Looking at the normal probability plot for SalePrice, the line is curved which suggests a skewed distribution. Few outliers also exist in this depiction of the data.

Top 5 highest and lowest observations:



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

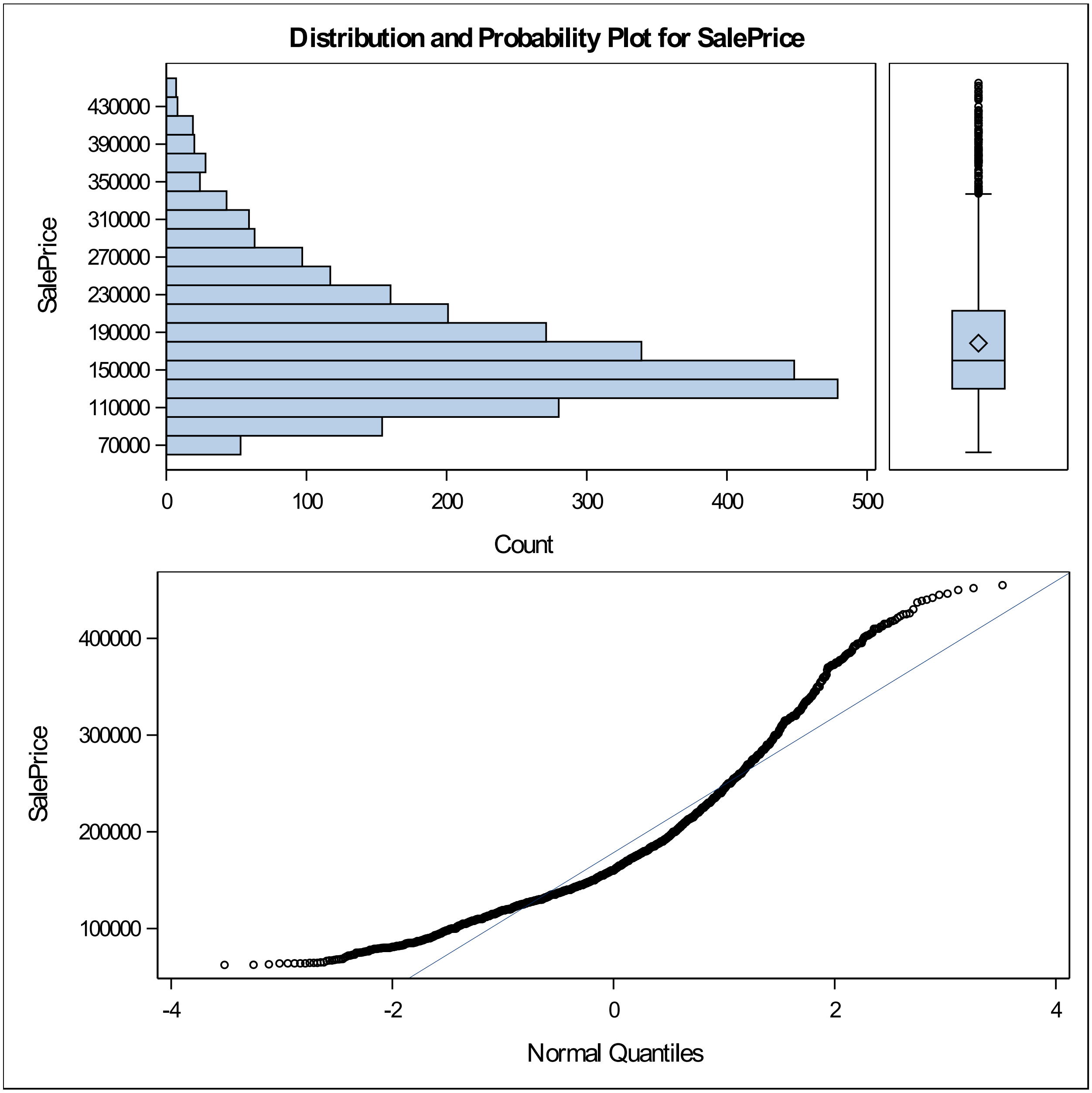
Quantiles:



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

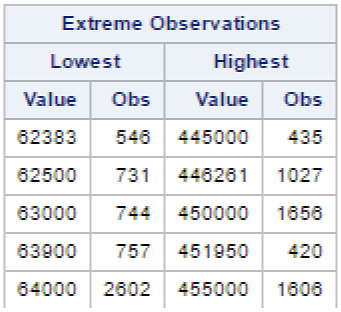
After reviewing the lowest and highest observations and quantiles, we see that there was possibility of at least one substantive outlier. On the top end, there were two observations made a few measurements apart that have close to the same value—both are also outliers.

Next, we'll use a histogram, scatterplots and sort procedure on SalePrice to determine which records should be of concern. The 99% and 1% quantile values will be used for the high and low sale prices to isolate the records that are potential outliers. Once we have this data set, we will then clean all of the outliers within the data.



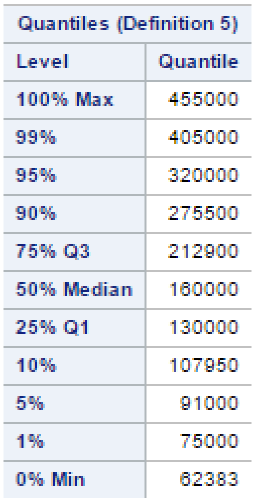
The outliers on both side of this tailed data set have been cleaned. The histogram shows a shift, and the line is still curved, indicating that we have data that is not normal in distribution.

Top 5 highest and lowest observations:



Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
62383	546	445000	435
62500	731	446261	1027
63000	744	450000	1656
63900	757	451950	420
64000	2602	455000	1606

Quantiles:



Quantiles (Definition 5)	
Level	Quantile
100% Max	455000
99%	405000
95%	320000
90%	275500
75% Q3	212900
50% Median	160000
25% Q1	130000
10%	107950
5%	91000
1%	75000
0% Min	62383

The outlier removal procedure executed correctly; however, from comparing the tables and looking at the histograms, we should continue with this data set so that we can examine how the performance of linear regression is affected.

Comparing models using the modified data set

Models to compare:

$$\text{SalePrice} = \beta_0 + \beta_1\text{GrLivArea} + \varepsilon$$
$$\text{SalePrice} = \beta_0 + \beta_1\text{GrLivArea} + \beta_2\text{MasVnrArea} + \varepsilon$$
$$\text{SalePrice} = \beta_0 + \beta_1\text{GrLivArea} + \beta_2\text{MasVnrArea} + \beta_3\text{BsmtUnfSF} + \varepsilon$$

SalePrice = non-manipulated data set
SalePrice$_0$ *= manipulated data set*

Non-manipulated results:

$$\text{SalePrice} = 13316 + 111.50225 \times \text{GrLivArea}$$
$$\text{SalePrice} = 26596 + 94.57781 \times \text{GrLivArea} + 118.54037 \times \text{MasVnrArea}$$
$$\text{SalePrice} = 25788 + 93.91175 \times \text{GrLivArea} + 118.57390 \times \text{MasVnrArea} + 3.23138 \times \text{BsmtUnfSF}$$

Manipulated results:

$$\text{SalePrice}_0 = 31604 + 98.20384 \times \text{GrLivArea}$$
$$\text{SalePrice}_0 = 40449 + 85.89009 \times \text{GrLivArea} + 997.44693 \times \text{MasVnrArea}$$
$$\text{SalePrice}_0 = 39321 + 84.92176 \times \text{GrLivArea} + 97.45046 \times \text{MasVnrArea} + 4.61086 \times \text{BsmtUnfSF}$$

Model	Adj R-Square	F Value
$\text{SalePrice} = 13316 + 111.50225 \times \text{GrLivArea}$	0.5002	2908.60
$\text{SalePrice} = 26596 + 94.57781 \times \text{GrLivArea} + 118.54037 \times \text{MasVnrArea}$	0.5594	1844.94
$\text{SalePrice} = 25788 + 93.91175 \times \text{GrLivArea} + 118.57390 \times \text{MasVnrArea} + 3.23138 \times \text{BsmtUnfSF}$	0.5595	1231.02
$\text{SalePrice}_0 = 31604 + 98.20384 \times \text{GrLivArea}$	0.4577	2403.02
$\text{SalePrice}_0 = 40449 + 85.89009 \times \text{GrLivArea} + 997.44693 \times \text{MasVnrArea}$	0.5063	1460.07
$\text{SalePrice}_0 = 39321 + 84.92176 \times \text{GrLivArea} + 97.45046 \times \text{MasVnrArea} + 4.61086 \times \text{BsmtUnfSF}$	0.5069	976.10

After reviewing both results, the performance of the models were degraded by the outlier removal. By having two different data sets and comparing regression performance, the analyst can look at the respective model performance accurately. Observing degradation in model performance tells us that our initial models were fitting well due to the influenced by the outliers.

Conclusions:

Transformation and outlier deletion are two influential techniques that can be used to influence the model fit, and potentially model strength. There are characteristics of data that give indication that this technique may be required; however, the application of the technique is highly procedural and requires interpretation. What can appear as outliers may be observations driven by a separate process. When examining the Goodness-of-Fit criteria, there are indicators that mean we should employ either of these techniques. There are some cases where it is absolutely necessary to use these techniques, as in scenarios where you can confirm that during collection, you've received observations that truly are not representative of the occurrences you're modeling, or if the variable transformation is required within the professional domain.

The next steps in the modeling process for this data set would be to see if there is other data that can be used to begin an attempt at model validation. It may be valuable at this time to present the initial model findings to the stakeholders. In this case, our initial assessment is that a categorical variable performs the best for explaining variability of Sale Price. If the stakeholder(s) don't want us to use a categorical variable, or would rather we use a multi-variable approach to include both continuous and categorical variables, it would be best to know that before proceeding to validation.

Code:

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

proc datasets library=mydata;
run;

data temp1;
set mydata.ames_housing_data;
log_SalePrice = log(SalePrice);
log_GrLivArea = log(GrLivArea);
sqrt_SalePrice = sqrt(SalePrice);
keep SalePrice log_salePrice sqrt_SalePrice GrLivArea log_GrLivArea MasVnrArea BsmtUnfSF BsmtFinSF1
FirstFlrSF TotalBsmtSF GarageArea;

ods graphics on;

proc print data=temp1 (obs=5);
run;

proc sgscatter data=temp1;
   title 'Normal and Transformed Variables';
   plot (SalePrice log_SalePrice) * (GrLivArea log_GrLivArea);
run;

proc reg data=temp1;
   model SalePrice = GrLivArea;
   model SalePrice = log_GrLivArea;
   model log_SalePrice = GrLivArea;
   model log_SalePrice = log_GrLivArea;
run;

proc corr data=temp1 nosimple rank;
        var GrLivArea MasVnrArea BsmtUnfSF BsmtFinSF1 FirstFlrSF TotalBsmtSF GarageArea;
        with log_saleprice;
run;

proc sgscatter data=temp1;
   title 'SalePrice and log_SalePrice vs GrLivArea';
   plot (SalePrice log_SalePrice) * GrLivArea;
run;

proc reg data=temp1;
   model SalePrice = GrLivArea;
run;
```

```
proc reg data=temp1;
   model sqrt_SalePrice = GrLivArea;
run;

proc univariate normal plot data=temp1;
        var SalePrice;
        histogram SalePrice / normal;
run;

data outliers;
        set temp1;
        keep SalePrice price_outlier GrLivArea MasVnrArea BsmtUnfSF;
        if SalePrice <= 61500  then price_outlier = 1;
        else if SalePrice > 61500  & SalePrice < 457347  then price_outlier = 2;
        else if SalePrice >= 457347  then price_outlier = 3;

proc sort data=outliers;
        by price_outlier;

proc means data=outliers;
        by price_outlier;
        var SalePrice;

data cleaned;
        set outliers;
         if price_outlier = 1 then delete;
        else if price_outlier = 3 then delete;

proc univariate normal plot data=cleaned;
        var SalePrice;
        histogram SalePrice / normal;

proc reg data=temp1;
        model SalePrice = GrLivArea;
        model SalePrice = GrLivArea MasVnrArea;
        model SalePrice = GrLivArea MasVnrArea BsmtUnfSF;

proc reg data=cleaned;
        model SalePrice = GrLivArea;
        model SalePrice = GrLivArea MasVnrArea;
        model SalePrice = GrLivArea MasVnrArea BsmtUnfSF;
run;
```