Data Analysis and Regression Model Building

Saranyan Vasudevan

Regression & Multivariate Analysis (PREDICT-410, Winter 2016)

Introduction

Our Objective

Predicting home prices has never been easy, it is a tradeoff among a set of housing parameters. Ames, Iowa dataset provides an exhaustive source of elements related to house buying/selling. The objective of our data analysis is to build a linear regression model to be able to predict the Sale Price of a housing property. For this, we are using a medium sized dataset comprising of 82 different housing variables. The data set is rich with facts and dimensions - and can be classified as Continuous, Ordinal, Nominal and discrete. The data dictionary is available for each of these variables (link provided in reference section), and it greatly helps with the data profiling. We will be SAS for this Exploratory Data Analysis exercise.

The objective of this report is to build regression models to predict Sale Price for Single family homes in Ames, Iowa – and assess the goodness of fit for those models. The assessment is based on analysis of plots generated by SAS procedures and metrics such as Coefficient of determination (adjusted R-squared), which indicates the closeness of data to our regression line and the F value of the model based on the predictors, and its corresponding p value. The 'p' value is a hypothetical test performed by SAS procedure, which we will using in this analysis. Additionally, we also assess the residual plots to uncover any important considerations that the model may pose.

For this analysis, we will build two simple regression models using the variables with best correlation coefficients as predictors, and then use these two predictors in a multiple regression model, followed by their reassessment after detecting and removing outliers. Our intent is to compare these models based on their goodness of fit. Finally, we re-do this exercise with a transformed response variable and re-assess the models with their counterpart.

Defining the Sample Population

Choosing candidate observations

As a part of defining our sample population, excluding observations that are not part of our intended analysis is an important step. Since, our broad objective is model for typical homes, we begin with identifying homes that are not typical, followed by excluding them from our working data set. From our preliminary analysis, we identify 4 types of exclusion criteria,

- Planned Unit Development Homes

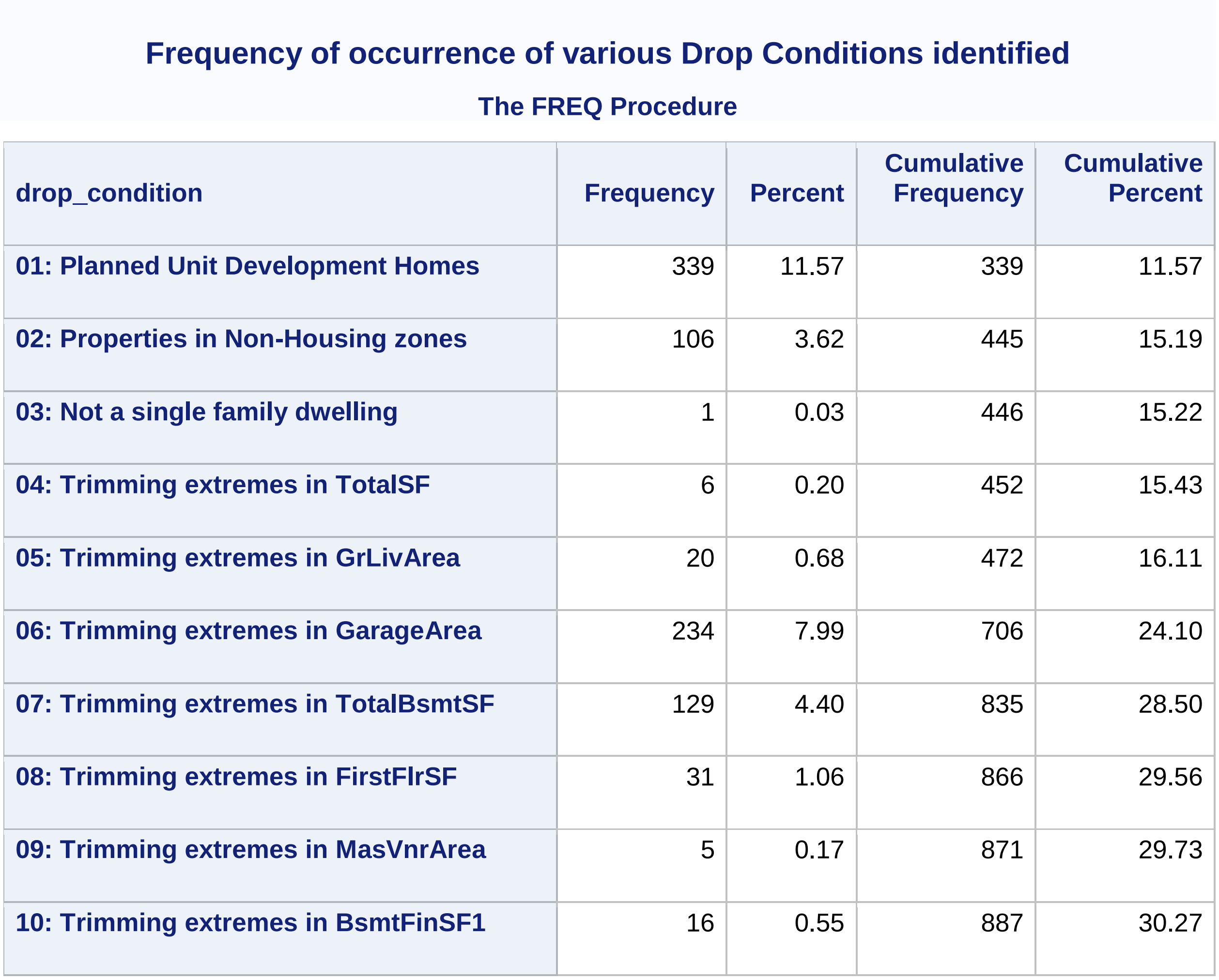
- Housing Properties in Non-Housing zones

- Not a single family dwelling based on Building type

- Extremes in various housing parameters (Given we are modeling for single family homes, having extremes might impact our model's prediction ability)

The observations not falling in any of these screening criteria will form part of our working dataset. Table A contains the frequency of occurrence based on the discussed criteria.

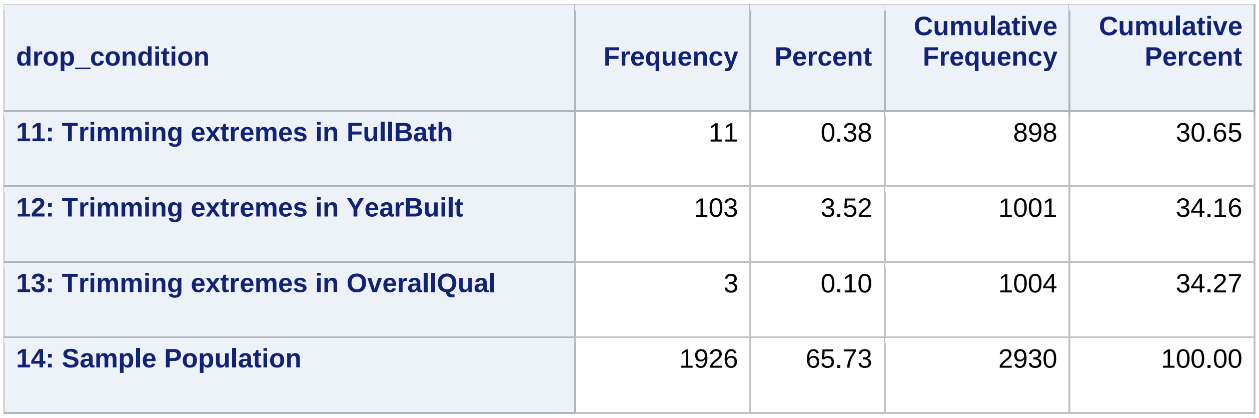
Table A



Frequency of occurrence of various Drop Conditions identified

The FREQ Procedure

drop_condition	Frequency	Percent	Cumulative Frequency	Cumulative Percent
01: Planned Unit Development Homes	339	11.57	339	11.57
02: Properties in Non-Housing zones	106	3.62	445	15.19
03: Not a single family dwelling	1	0.03	446	15.22
04: Trimming extremes in TotalSF	6	0.20	452	15.43
05: Trimming extremes in GrLivArea	20	0.68	472	16.11
06: Trimming extremes in GarageArea	234	7.99	706	24.10
07: Trimming extremes in TotalBsmtSF	129	4.40	835	28.50
08: Trimming extremes in FirstFlrSF	31	1.06	866	29.56
09: Trimming extremes in MasVnrArea	5	0.17	871	29.73
10: Trimming extremes in BsmtFinSF1	16	0.55	887	30.27



drop_condition	Frequency	Percent	Cumulative Frequency	Cumulative Percent
11: Trimming extremes in FullBath	11	0.38	898	30.65
12: Trimming extremes in YearBuilt	103	3.52	1001	34.16
13: Trimming extremes in OverallQual	3	0.10	1004	34.27
14: Sample Population	1926	65.73	2930	100.00

The resultant dataset contains 1926 observations, with computed variables for total square footage and log (SalePrice), log (TotalSF) and log (GrLivArea), which will be used later in our analysis.

Simple Linear Regression Model

We begin with running a PROC CORR procedure on the Ames dataset, and will work with the continuous variables that have the high correlation coefficients. Table B contains the output from the procedure.

Table B – PROC CORR Procedure Output

Variable	Pearson Correlation Coefficients	p Value	No. of Observations
OverallQual	0.81273	<.0001	1926
TotalSF	0.81997	<.0001	1926
GrLivArea	0.75518	<.0001	1926
YearBuilt	0.62565	<.0001	1926
GarageArea	0.62060	<.0001	1926
FullBath	0.60981	<.0001	1926
TotalBsmtSF	0.55569	<.0001	1926
FirstFlrSF	0.53767	<.0001	1926
MasVnrArea	0.50569	<.0001	1915
BsmtFinSF1	0.29474	<.0001	1926

Predictor: *TotalSF |* ***Response:*** *SalePrice*

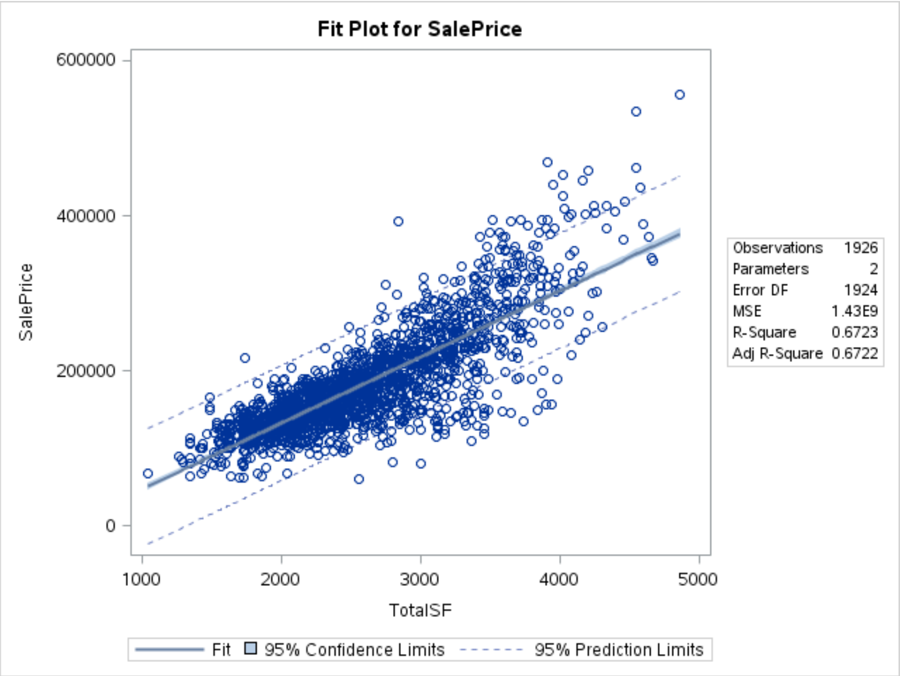
Our goal is to build a regression model to predict Selling price for typical homes based on TotalSF (Total Square Feet) from our dataset. We use PROC REG procedure in SAS to perform this regression. Below is the summarized SAS output from PROC REG procedure.

Table C – PROC REG Procedure Output

Metric	Value
Adjusted R-squared	0.6722
F-Value	3948.09
‘p’ Value	< 0.0001
Model	SalePrice = -37881 + (85.33337 * TotalSF) + e

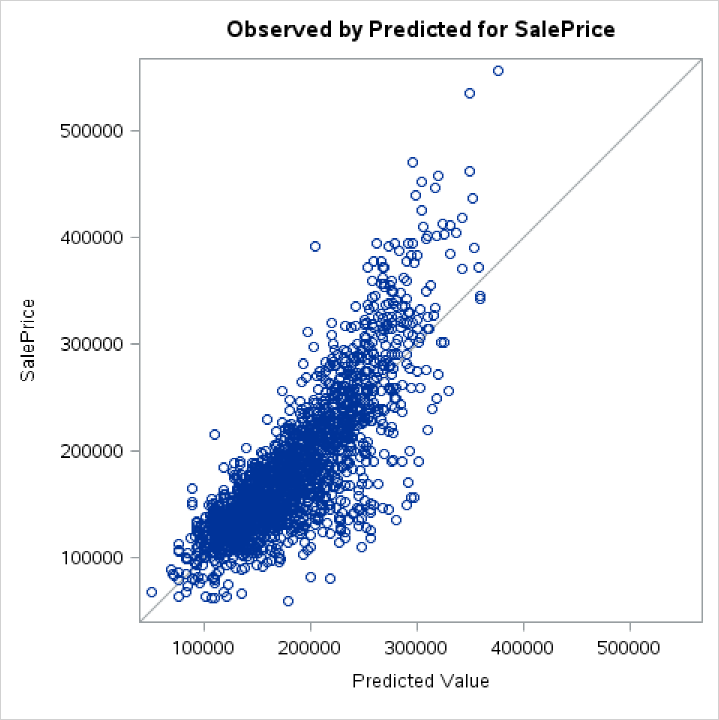
This model has an Adjusted R-squared of 0.6722 and has a negative intercept, which technically means the SalePrice will take a negative value of Total Square feet is set to zero, however this value does not fall in our range, and is only an adjustment to our prediction based on Total Square feet. Since we are trying to determine the variance that TotalSF produces on SalePrice, the sign of intercept can be disregarded.

Figure 1 – SalePrice vs. TotalSF



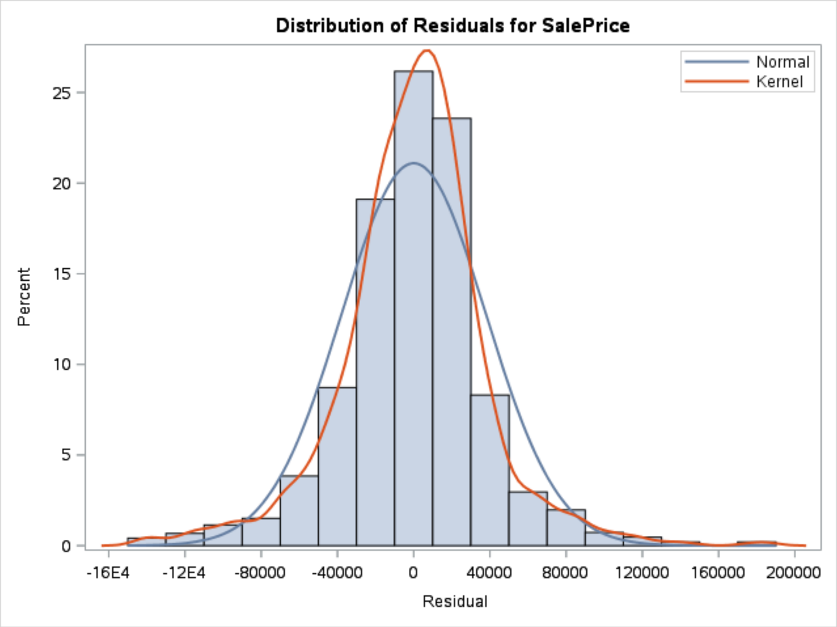
The plot above indicates a positive relationship between Total Square Feet and Sale Price. The sale price increases by an average of 85.33 for each unit increase in Total Square Feet. This model brings up some concerns about outliers in the data, which we will deal with in the upcoming sections. The predicted vs Actuals plot (Figure 2) indicates a drop in prediction accuracy as the Total Square foot increases, this could be due to presence of influential observations at the right end of the distribution.

Figure 2 – Predicted vs Observed (SalePrice)



The Residual histogram (Figure 3) takes the shape of a bell curve, which is in line with our linearity assumption. But we also observe a slightly extended right tail.

Figure 3 – Residual Distribution



The QQ plot (Figure 4) indicates a departure from normality at the ends, with a longer right tail i.e. presence of extreme data points at both the ends. They can be better identified by 'UNPACK'ing the Cook's D plot, which points to the influential observations in the dataset.

Figure 4 – QQ Plot

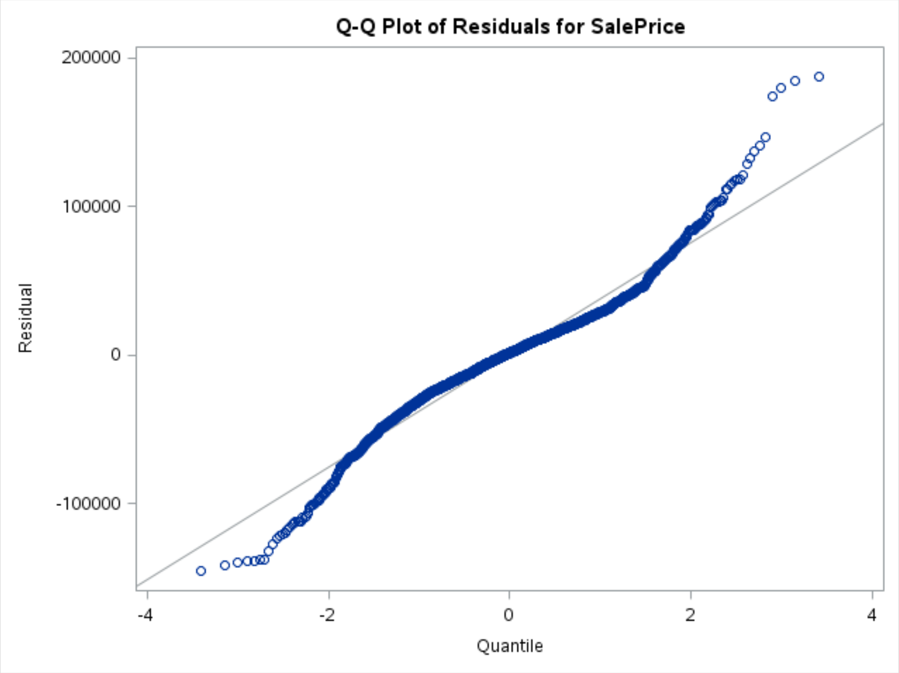
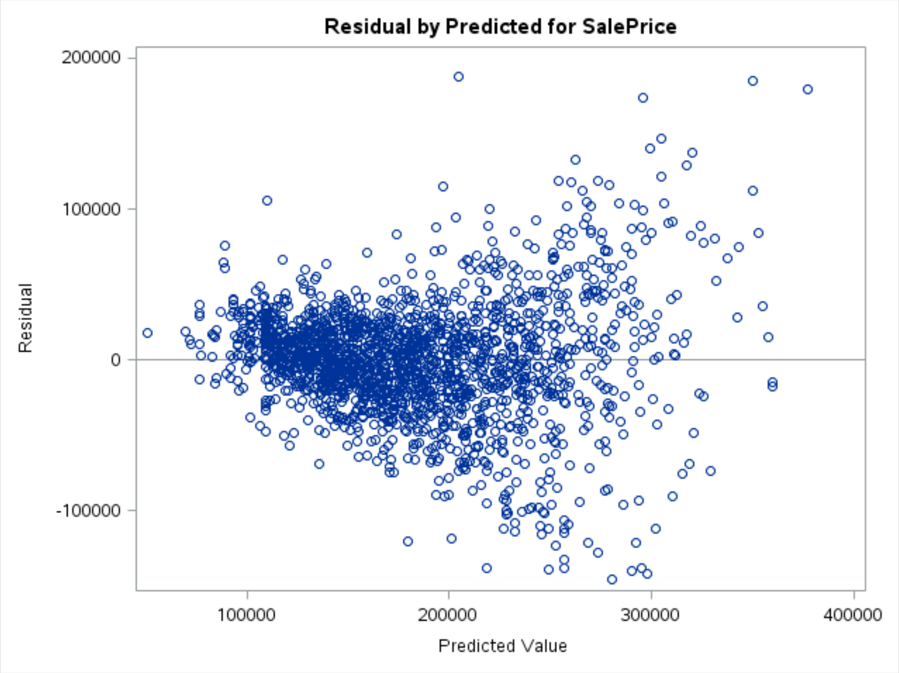


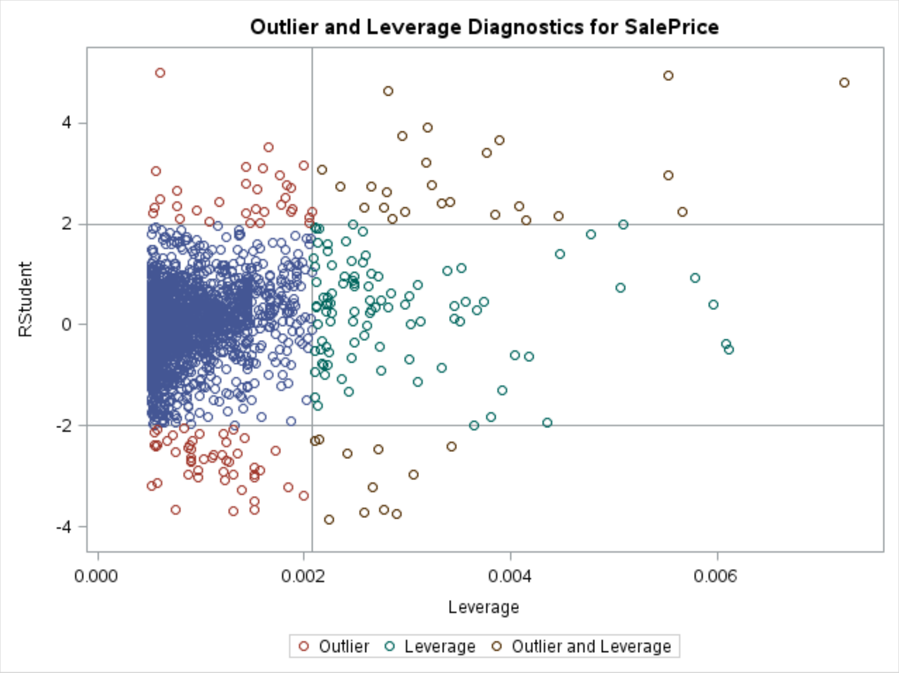
Figure 5



The Residual by Predicted Values plot in Figure 5, in this model is clustered towards the middle line, but indicates a slightly expanding scatter towards the right, pointing to slight heteroscedasticity. Adding another predictor variable or transforming the response variable can change this behavior.

While most of the observations fall have conformed leverage, the RStudent leverage plot in Figure 6 below, indicates a handful of observations that have a relatively lower/higher Square footage compared to neighborhood data points, they are either influential or considered outliers.

Figure 6



Predictor: *GrLivArea |* ***Response:*** *SalePrice*

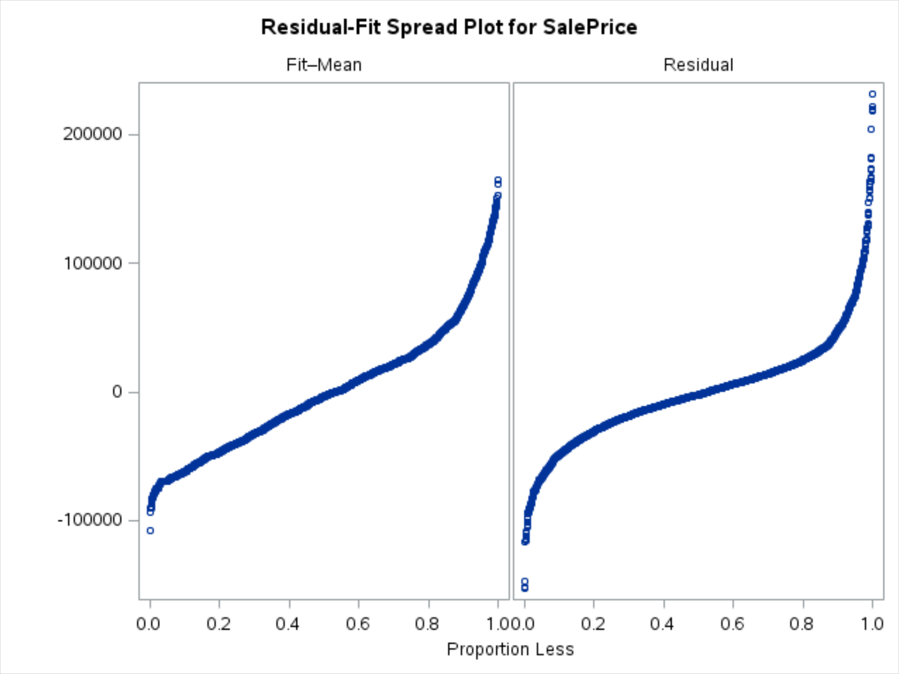
Now, we use GrLivArea, which has the next higher correlation coefficient, to predict SalePrice. Below is the summarized SAS output from PROC REG procedure.

Table D: PROC REG Output

Metric	**Value**
Adjusted R-squared	0.5701
F-Value	2553.48
‘p’ Value	< 0.0001
Model	SalePrice = 17206 + (110.58 * GrLivArea) + e

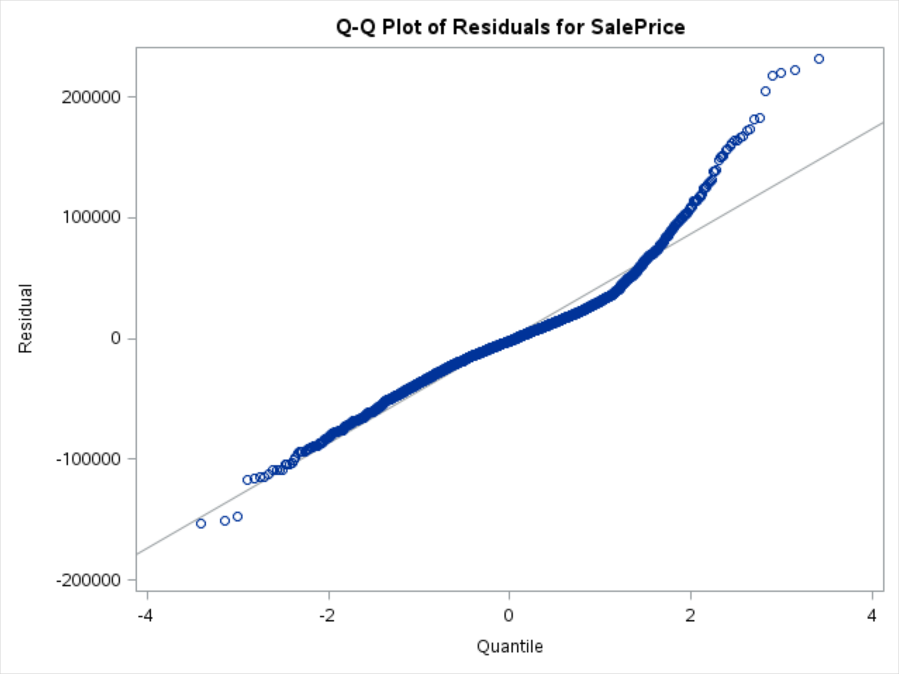
The residual plot indicates a wide spread distribution, and this model does not explain the variation in the response variable well and attributes to a good percentage of variation in residuals (as indicated in Figure 7).

Figure 7 – R-F plot (GrLivArea Vs SalePrice)



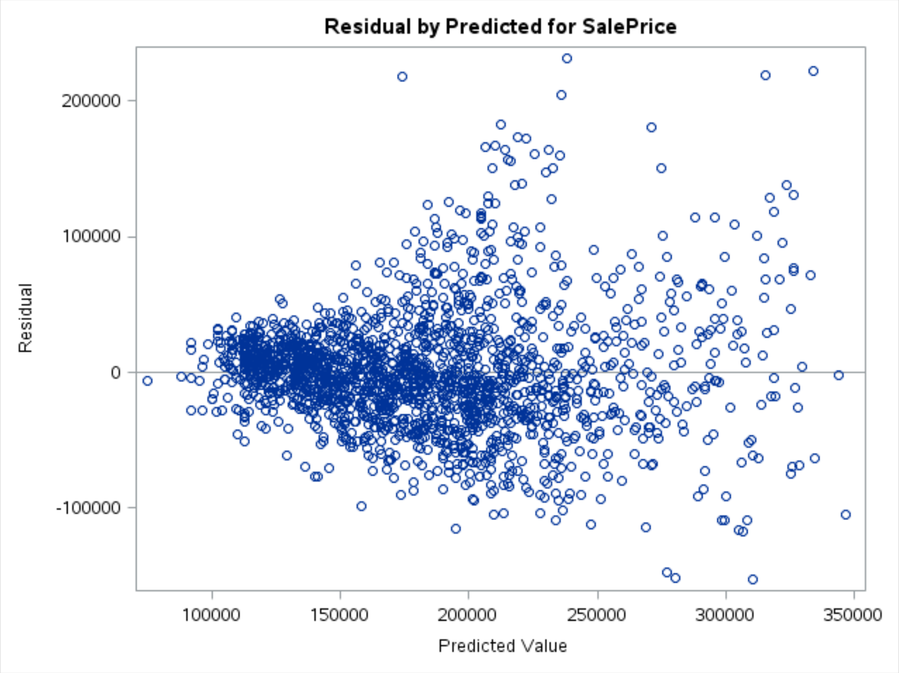
A prominent departure from normality is observed in right tail of distribution.

Figure 8 – QQ plot (GrLivArea Vs SalePrice)



This behavior of QQ plot (Figure 8) could be attributed to the presence of outliers which is in agreement with RF plot. The residual has a skewed distribution, and departs from normality, with a long right tail, and violates the assumption of homoscedasticity – this is further confirmed by the presence of outliers in residual vs Predicted (Figure 9), leverage and Cook’s D plots.

Figure 9 (GrLivArea Vs SalePrice)



Multiple Linear Regression Model

Predictor: *TotalSF, GrLivArea* | ***Response:*** *SalePrice*

We combine the Simple Linear Models 1 & 2 to make it a Multiple Regression model with 2 continuous predictor variables (TotalSF & GrLivArea), and 1 response variable (SalePrice).

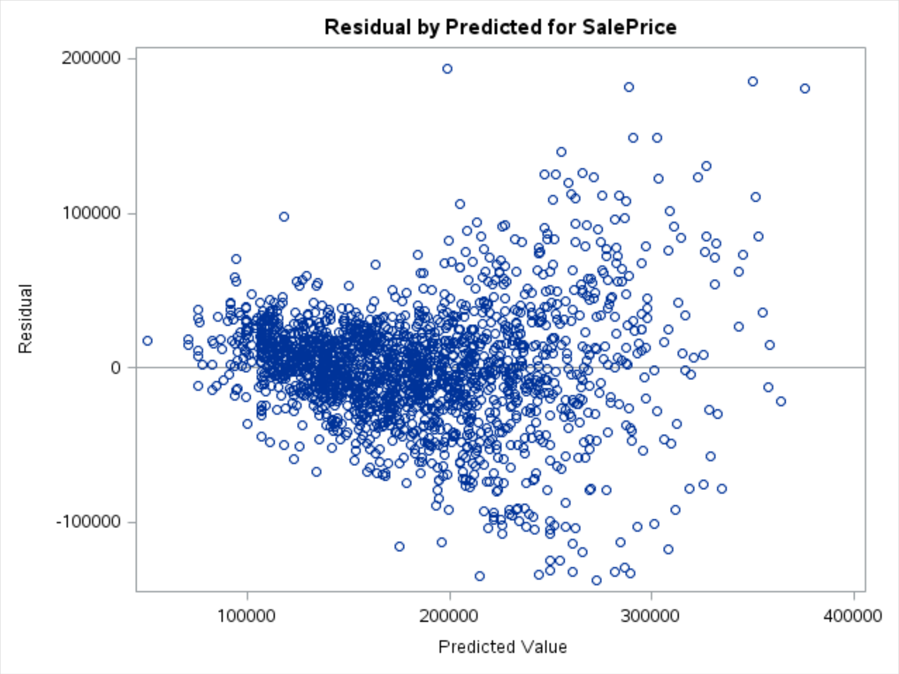
Table E: PROC REG Output

Metric	Value
Adjusted R-squared	0.6779
F-Value	2026.80
‘p’ Value	< 0.0001
Model	SalePrice = -35240 + (70.83 * TotalSF) + (23.29 * GrLivArea)+ e

The summarized output from PROC REG in above table, indicates an Adjusted R-Squared of 0.6779, i.e. much of the variation is explained by this model based on this Adjusted R-Square metric – which is also greater than R-square values in Simple regression models we saw earlier. The p value for the F statistic is less than the significance level. Given this is a multiple regression model, this indicator tells us that our model is statistically significant. However, we will revalidate the same by looking at the plots generated.

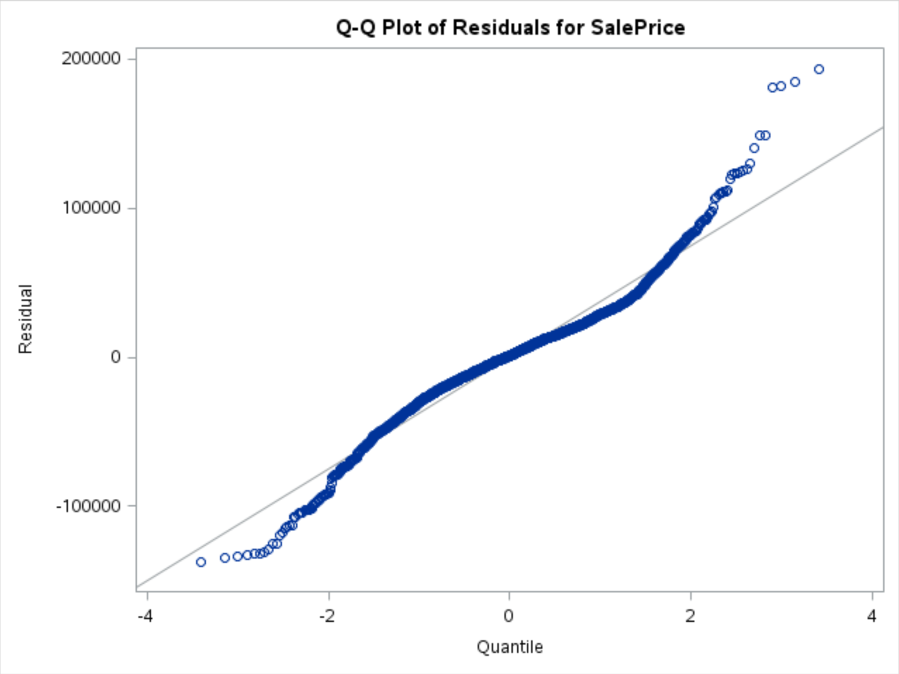
The adjusted R-squared is slightly better compared to our 1st simple regression model with TotalSF as predictor, but all three models have a skewed residual distribution. The residual distribution in this case is well spread, but is skewed and departs from normality, and is slightly heteroskedastic or approximately homoscedastic. The departure from normality is well pronounced at both the ends, with long tails.

Figure 10 (TotalSF, GrLivArea Vs SalePrice)



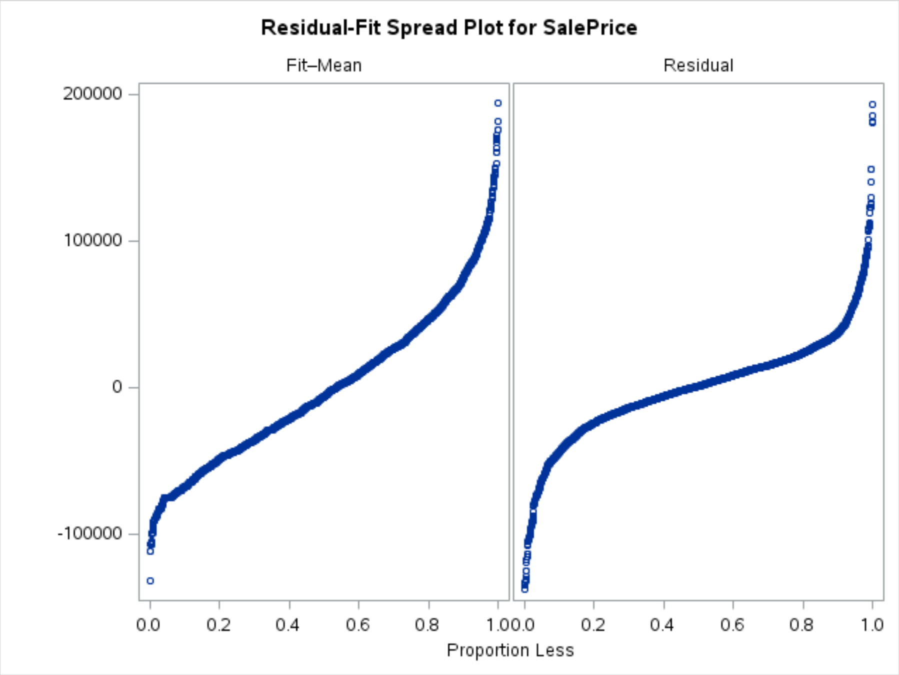
The impact of Outliers and Influential points are evident in the lower and upper tails of the data distribution. The Predicted to Actual values plot indicates a tighter relationship, when compared to our simple regression models.

Figure 11 – QQ Plot (TotalSF, GrLivArea Vs SalePrice)



Comparing the RF spread plots, it seems that a good percentage of variation is not explained by the model and it significantly accounts for the residual variation.

Figure 12 (TotalSF, GrLivArea Vs SalePrice)

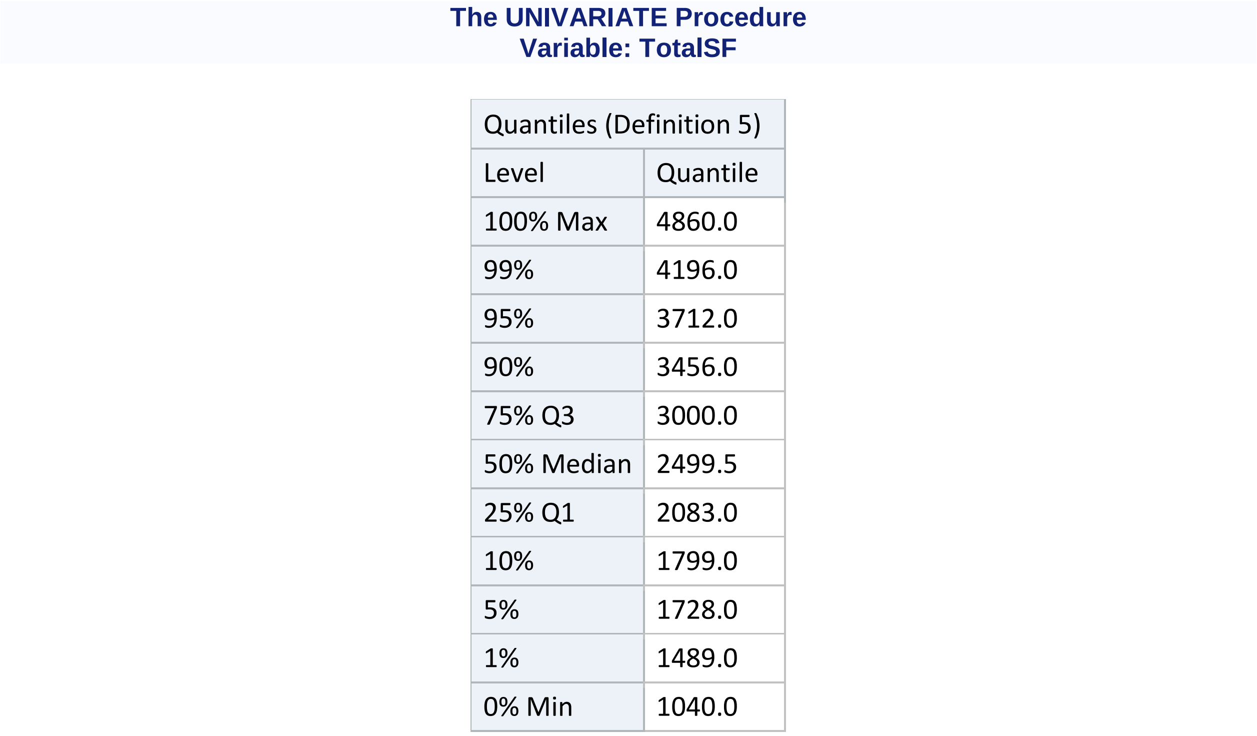


However, this multiple regression model violates the linearity assumptions and does not significantly better either of the simple models we discussed – and there is room for improvement in this model. There is a possibility of addition of another predictor variable to this model. But first, we would like to see how it performs when a transformation is applied in following sections.

Outlier Identification

The models produced in the earlier steps have outliers, which greatly influence the regression line and associated metrics. Outliers are extreme values, which typically lie on the tails of a data distribution, and have the potential to impact the overall regression exercise. Identification of outliers in our case is to look for observations that do not fit our model. Since linear regression assumes normality, we can assume that values lying outside of 99% from either side of the mean as outliers (Empirical Rule). The PROC UNIVARIATE procedure in SAS can summarize the data distribution in the dataset, based on quantiles for the predictor variables in our multiple regression model. Below is the PROC UNIVARIATE output for TotalSF variable:

Table F – SAS Univariate Procedure Output (TotalSF)



The UNIVARIATE Procedure
Variable: TotalSF

Quantiles (Definition 5)	
Level	Quantile
100% Max	4860.0
99%	4196.0
95%	3712.0
90%	3456.0
75% Q3	3000.0
50% Median	2499.5
25% Q1	2083.0
10%	1799.0
5%	1728.0
1%	1489.0
0% Min	1040.0

Figure 13 (SAS Univariate Procedure Output - TotalSF)

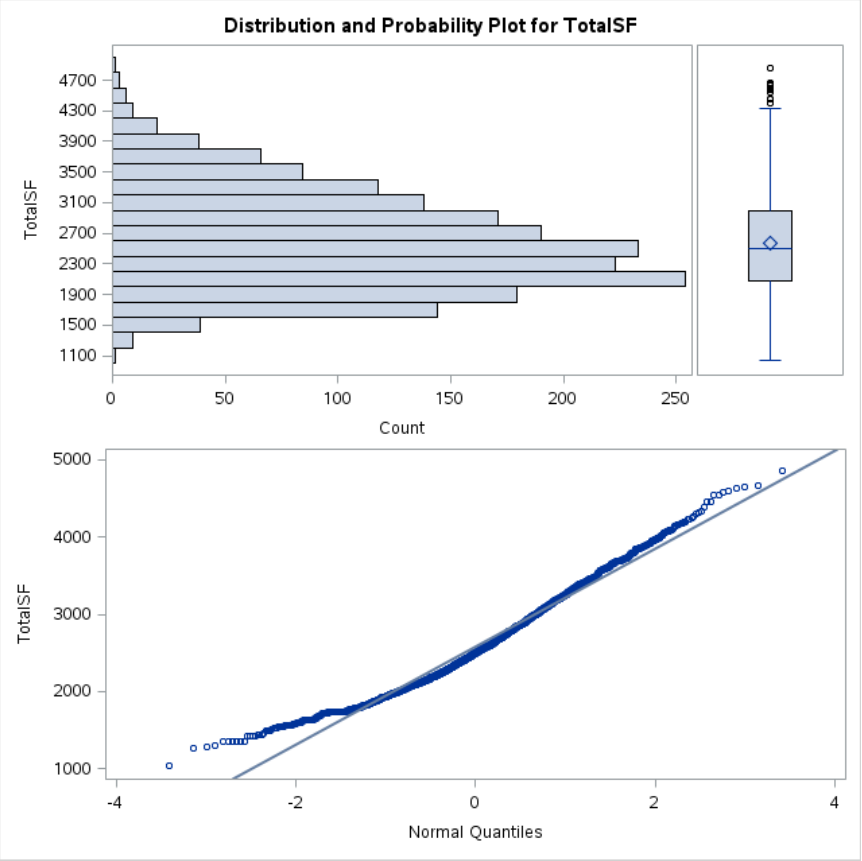
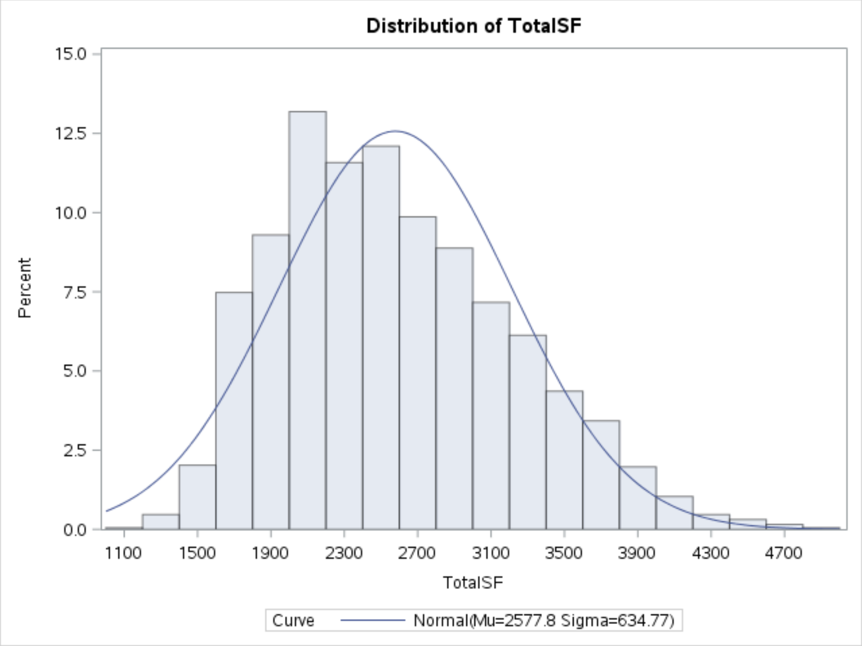


Figure 14 (SAS Univariate Procedure Output – TotalSF)



Based on this output, we can trim the 1% from either side of the mean, i.e. TotalSF >= 4196 or TotalSF <= 1489. Likewise, we run the procedure for the other regressor variable, i.e. GrLivArea

Table G – SAS Univariate Procedure Output (GrLivArea)

The UNIVARIATE Procedure
Variable: GrLivArea

Quantiles (Definition 5)	
Level	Quantile
100% Max	2978.0
99%	2728.0
95%	2376.0
90%	2093.0
75% Q3	1740.0
50% Median	1458.5
25% Q1	1130.0
10%	935.0
5%	864.0
1%	768.0
0% Min	520.0

Figure 15 (SAS Univariate Procedure Output – GrLivArea)

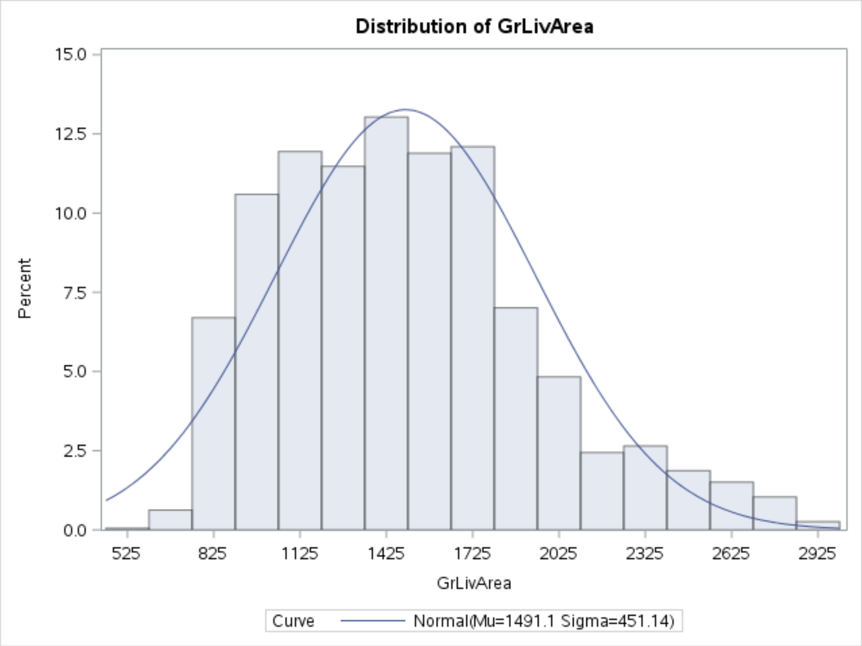
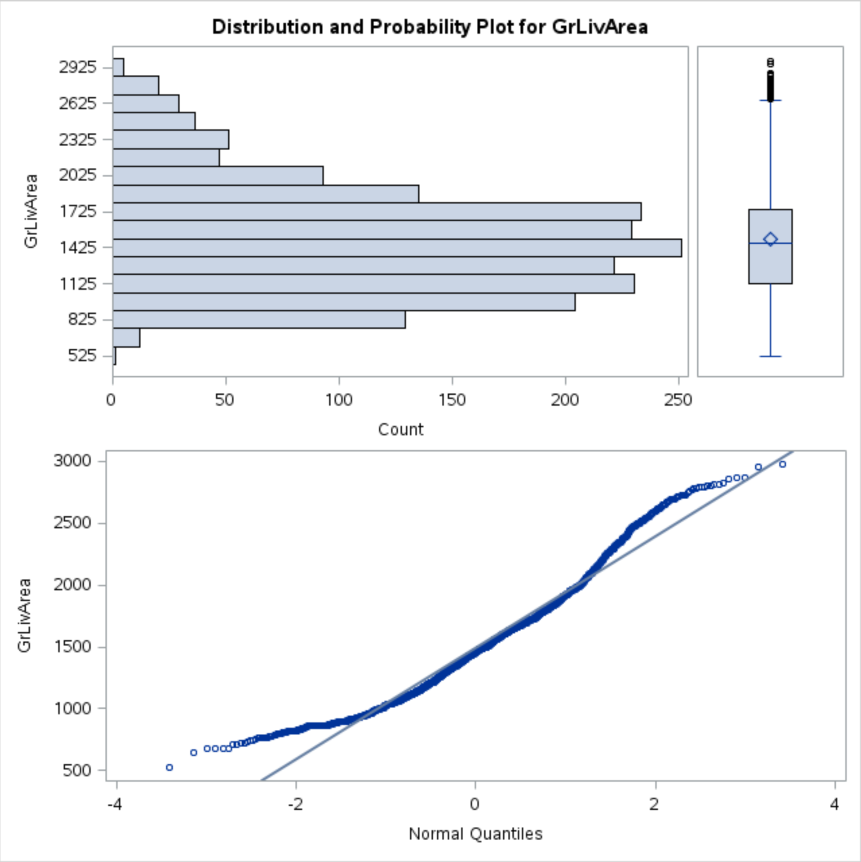


Figure 16 (SAS Univariate Procedure Output – GrLivArea)



These results can be complimented visually, using Cook's D or RStudent plots from the output of PROC REG in SAS, however for the purposes of accuracy, we will export the regression results to another SAS output and then use numbers instead of visuals for outlier detection. The PROC FREQ output in table H gives the frequency of occurrence of RStudent based outlier detection.

Table H – Distribution of Outliers by RStudent

The FREQ Procedure

Rstudent_Range	Frequency
00: more than 3	15
01: 2 to 3	44
02: 1 to 2	133
03: 0 to 1	804
04: -1 to 0	693
05: -2 to -1	176
06: -3 to -2	45

Rstudent_Range	Frequency
08: Less than -3	16

Below table contains the frequency of occurrence of Cook's D based outlier detection.

Table I – Distribution of Outliers by Cooks D

The FREQ Procedure

CookD_Range	Frequency
01: CookD Potential outlier	156
02: Not an outlier	1770

Now that we have identified outliers using various methods, we combine all these conditions in a waterfall of outlier definitions. In total, 189 observations are considered outliers and below is their frequency distribution (Table J).

Table J – Distribution of Outliers – All Categories

The FREQ Procedure

Type_of_Outlier	Frequency	Percent	Cumulative Frequency	Cumulative Percent
01: CookD Potential outlier	156	8.10	156	8.10
03: TotalSF outlier	25	1.30	181	9.40
04: GrLivArea outlier	8	0.42	189	9.81
05: Not an outlier	1737	90.19	1926	100.00

Refitting Multiple Regression Model without Outliers

Predictor: *TotalSF, GrLivArea |* ***Response:*** *SalePrice*

To see the impact of these outliers in our previously constructed multiple regression model, we exclude the outliers identified in the above step, and rerun the PROC REG procedure in SAS on the TotalSF and GrLivArea.

Below is the summarized SAS output. The adjusted R-squared value has improved, also the F-value is highly significant. With the 'p' value, we can reject the null hypotheses – provided all the goodness of fit tests are checked out.

Table K: PROC REG Output (TotalSF, GrLivArea Vs SalePrice) – Without outliers

Metric	Value
Adjusted R-squared	0.7375 (Up from 0.6779)
F-Value	2439.17
‘p’ Value	< 0.0001
Model	SalePrice = - 24761+ (65.40 * TotalSF) + (24.72 * GrLivArea)+ e

The Mean squared error value also reduced in the model without outliers. In the upcoming section, we will compare the multiple regression models (with and without outliers), to determine if the fit has improved after the trimming the outlier records.

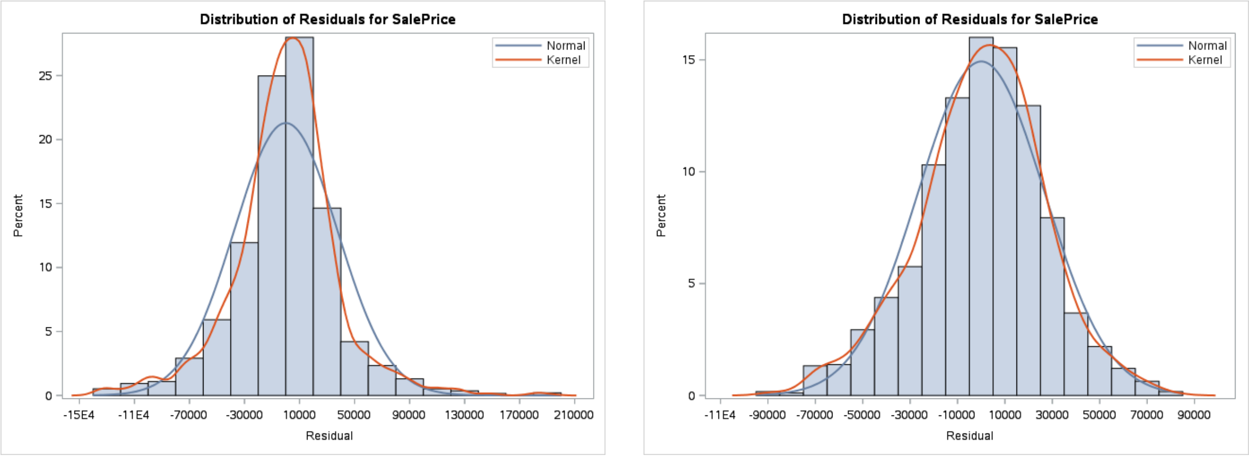
Model Comparison

Below are the models that we are trying to compare in this section.

*SalePrice = -35240 + (70.83 * TotalSF) + (23.29 * GrLivArea) (With outliers)*
*SalePrice = -24761+ (65.40 * TotalSF) + (24.72 * GrLivArea) (Without outliers)*

We saw the adjusted R squared value improved and Mean Squared Value dropped, after removing the outliers, so we will compare the plots to confirm the improvement. The plots to the left are from the datasets that contain outliers.

Figure 17 – Residual Histogram Comparison (Multiple Regression Models)

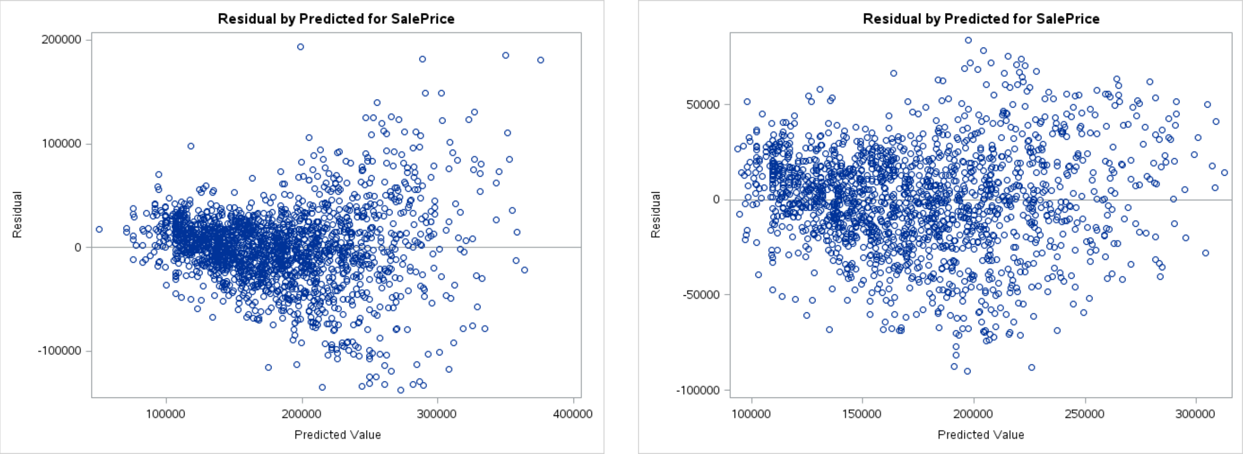

With outliers

without outliers

The plot to the right, has a better normal distribution compared to the one on left. Also, the skew and long right tail is not present in the one without outliers.

Figure 18 – Residual vs Predicted plots Comparison (Multiple Regression Models)

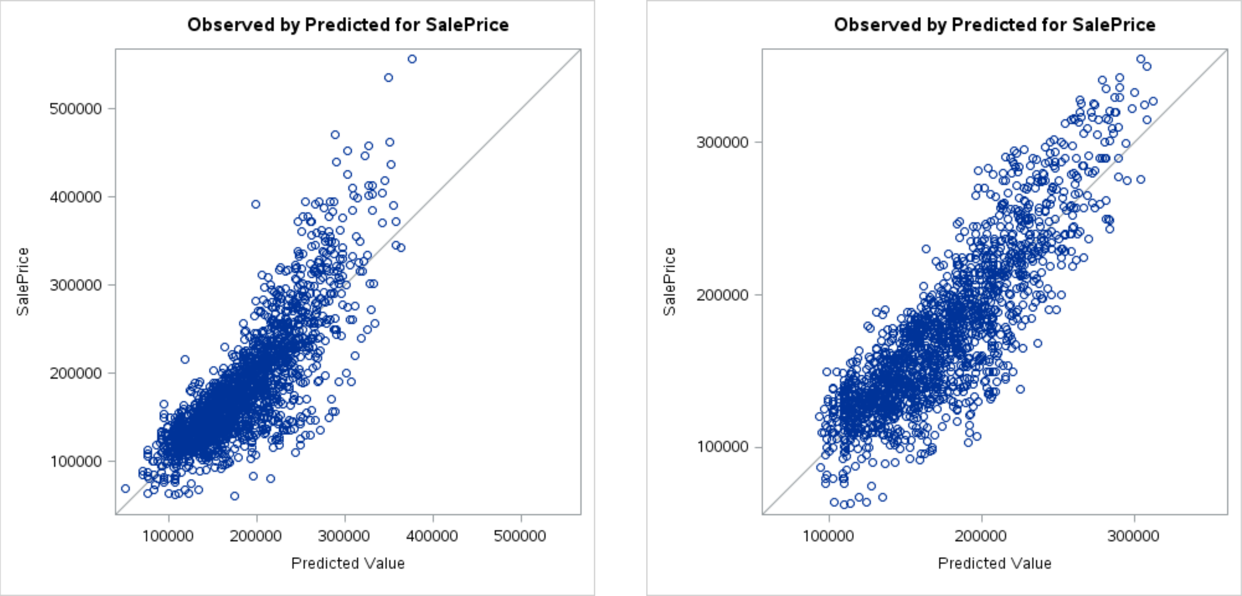
With outliers without outliers



The error variances in the model without outliers appears to be constant, i.e. is relatively homoscedastic compared to the one on its left. The below observed vs predicted plot comparison (Figure 19) indicates a better positive linear relationship, when the outliers are removed.

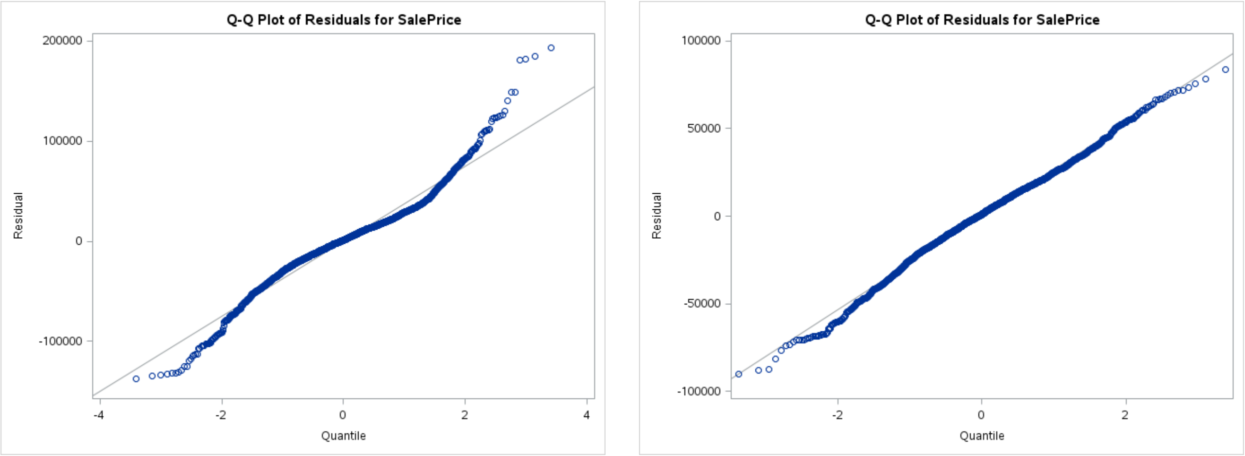
Figure 19 - Observed vs Predicted plot comparison (Multiple Regression Models)

With outliers without outliers



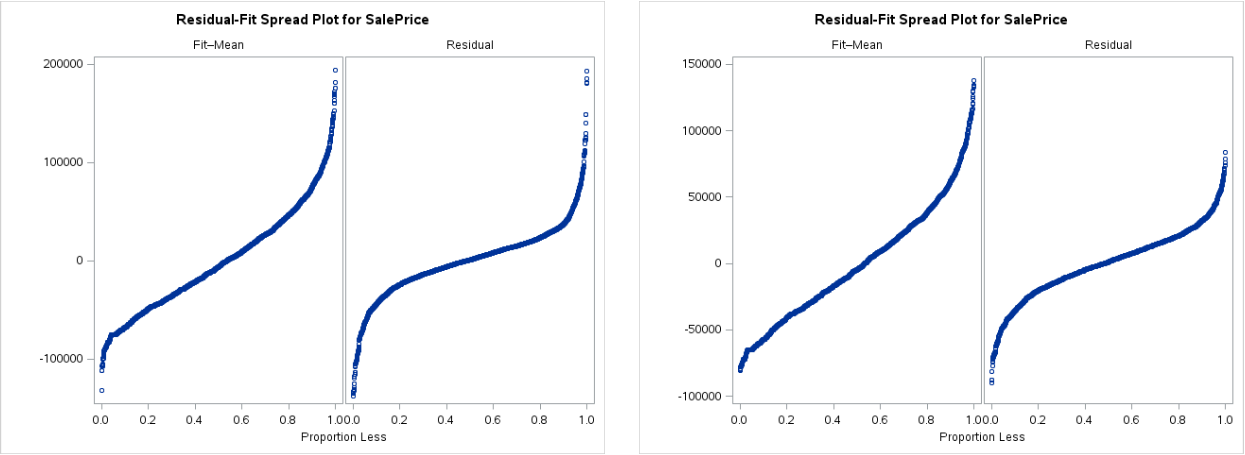
The QQ plot comparison below, shows a great reduction in departure from normality, and the residuals sticking close to the regression line, this indicates that the residuals are nearly normal in shape. The improvement is also noticed in R-F plot.

Figure 20 – QQ Plot comparison (Multiple Regression Models)



With outliers without outliers

Figure 21 – R-F Plot Comparison (Multiple Regression Models)



With outliers without outliers

With these goodness of fit observations and adjusted r-squared metric, we can say that the model without outliers is a better fit than the one with outliers.

Multiple Regression model (with log transformation)

***Predictor**: TotalSF, GrLivArea| **Response:** log (SalePrice)*

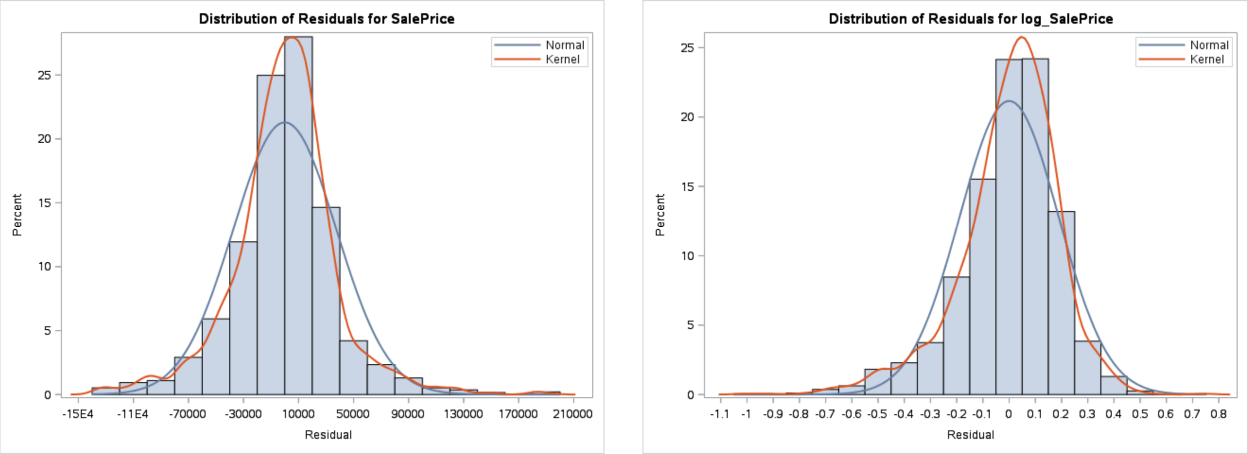
We now regress the combination of 2 continuous variables used in our previous regression model, with the logarithmic value of Sale Price. With the logarithmic transformation, we hope to see high variation in the dependent variable for a unit increase in one of the independent variables, while the other independent variable is constant. Below is the summarized SAS output from PROC REG procedure.

Table L: PROC REG Output (TotalSF, GrLivArea vs. log (SalePrice))

Metric	Value
Adjusted R-squared	0.6799 (up from 0.6779)
F-Value	2045.54
‘p’ Value	< 0.0001
Model	log_SalePrice = 10.96190 + (0.00033874* TotalSF) + (0.00014714 * GrLivArea) + e

The first glance indicates a left skewed and double peaked distribution of residuals with transformation applied, whereas the original multiple regression model was right skewed. It is unclear at this point, if this effect could be because of the different distributions of the independent variables.

Figure 22: Comparison Residual Histogram Plots: SalePrice vs log_SalePrice



The QQ plot (Figure 23) indicates a slightly better percentage of data points falling on the regression line. However, departures from normality is seen on both the ends (although, not as aggressive as the original model), so the linearity assumption is not met, even in this case. However, because a logarithmic transformation was applied, the error variances tend to be constant, checking out the homoscedasticity assumption in Figure 24. The R-F plot (Figure 25) for the model with transformed response variable, doesn’t seem to explain the variation in predictor variables like the one on the left does.

Figure 23: Comparison QQ Plots: SalePrice vs log_SalePrice

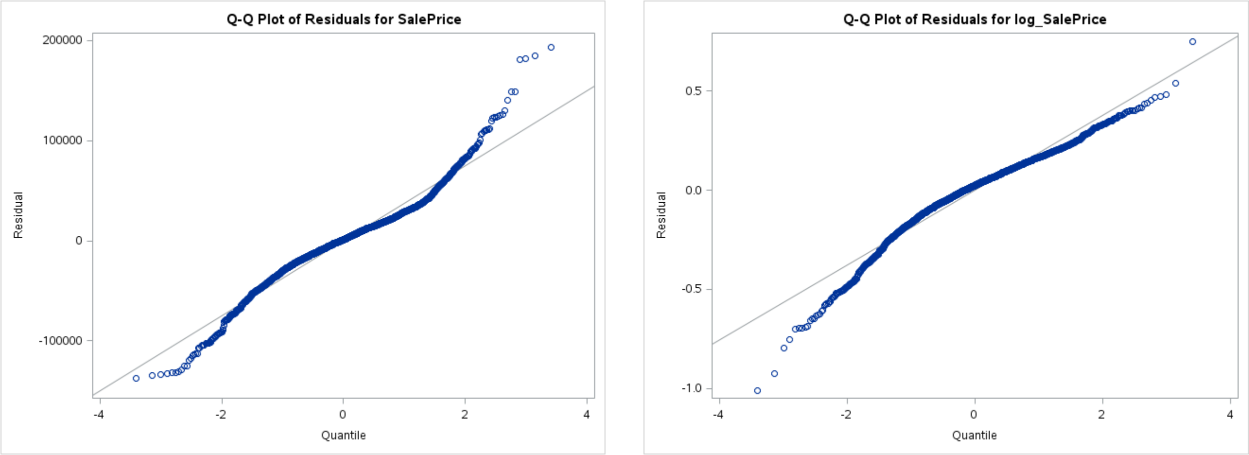


Figure 24: Comparison Residual vs Predicted Plots: SalePrice vs log_SalePrice

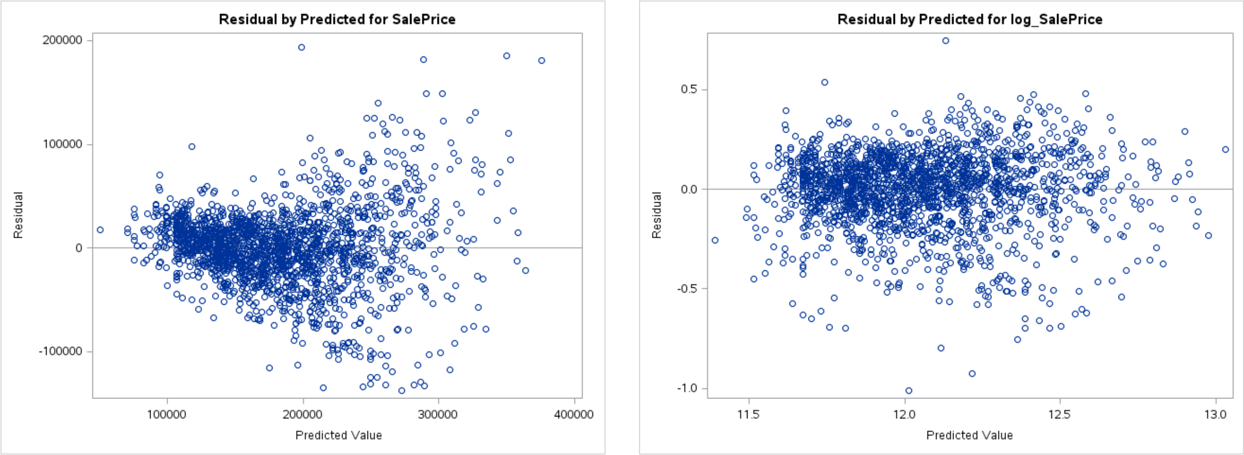
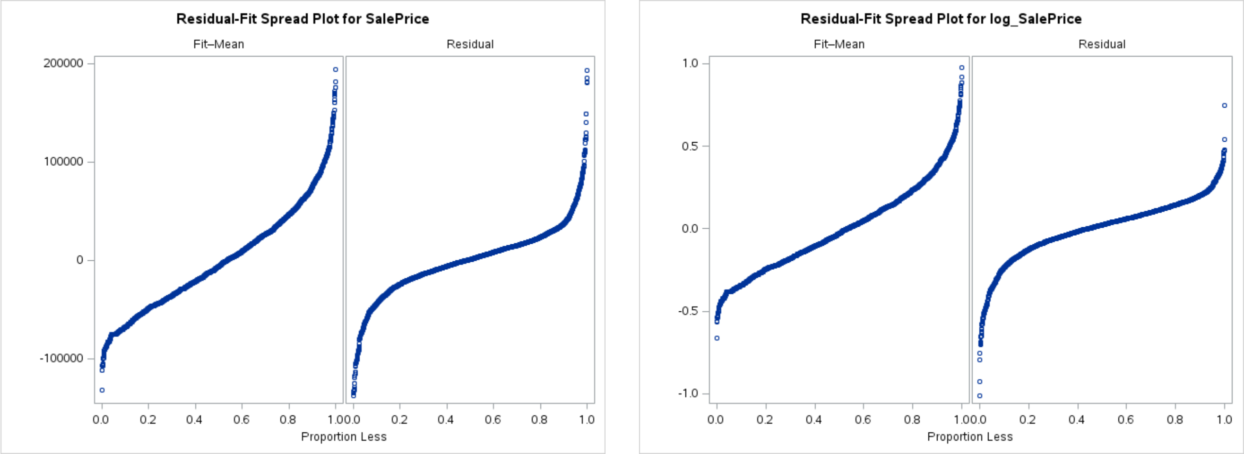


Figure 25: Comparison R-F Plots: SalePrice vs log_SalePrice



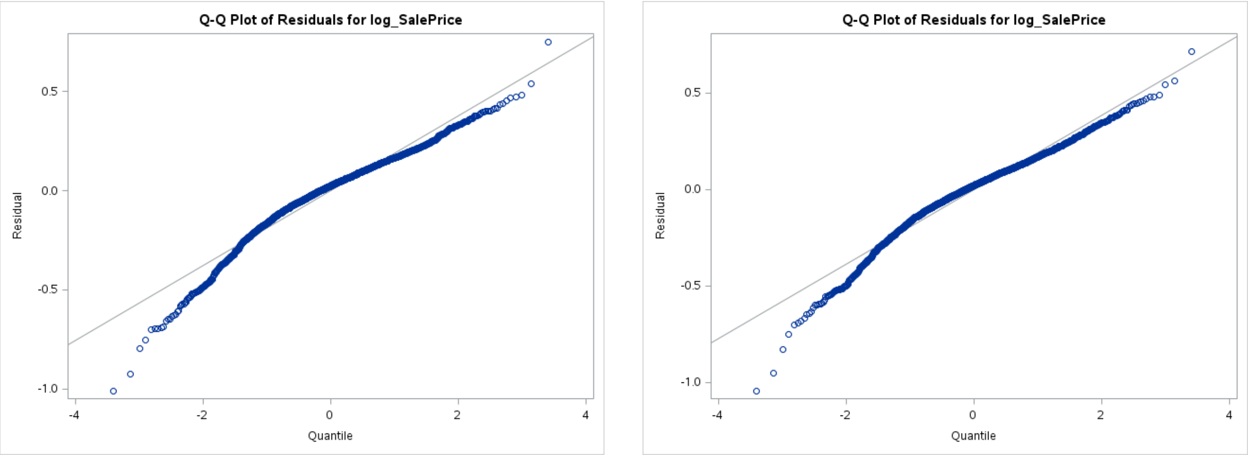
Conclusion:

The marginal increase in R-Squared value is not very well complimented in the residual goodness of fit validations, and the model with log transformation applied to the response variable does not seem to be significantly better than its counterpart. Transformations to the response variable can be applied to linearize the relationship with the predictor variable(s) or when the residual spread is skewed or to retain the influential observations in the dataset, while suppressing their effect. But, it didn’t help much in this specific case, which could be attributed to the presence of outliers in the dataset. Since, we choose to not remove the outliers for this test, we can try to transform the predictor variable and see how the model performs. To start with, we can consider applying logarithmic transformation to the predictor variable(s).

***Predictor**: log (TotalSF), log (GrLivArea)| **Response:** log (SalePrice)*

There is a marginal drop observed in the R-Squared value after applying logarithmic transformation to the predictor variables, also, the basic assumption of normality still remains violated, as the departure from normality is observed. The heavy tail still exists. Below QQ Plots compare the model with transformed response variable (left) and model with transformed response and predictor variables (right).

Figure 26: Comparison QQ Plots:
Transformed Response Variable vs. Transformed Response & Predictor Variables



So, the application of transformation to predictor variables did not help much to fit the model or to mask the effect of outliers.

Conclusion

We started our analysis with 2930 observations, but reduced to 1926 for building most of our models, we further trimmed outliers and worked with nearly 60% of the original count. This brings up a very important point that there can’t be a model that caters to every candidate observation in the dataset, or in other words, modeling for regular observations and modeling for extremes are very different, and the models cannot be reused in most cases.

Adjusted R-squared metric is a summary statistic, and is not a good indicator when used standalone in the identification of goodness of fit for models. It can explain the variability in the response variables using the predictor variables, but has to be complimented with residual analysis, the F-test and Mean Squared Error value to make an accurate judgement. The metric can be misleading with multiple predictor variables.

Transformation of the response variable resulted in the model to align with homoscedasticity, however the influential and outlier observations impacted the model more than they did to the original model. The application of logarithmic transformation did not fully take care of the skewed residual distribution. Overall, the models with log transformation applied to response variables does not seem to be a better fit, when compared to their counterparts. Transformation to predictor variables did not produce better results either.

By far, the multiple regression model without outliers performed much better than the rest of the models discussed here, which underlines an important step on outlier detection and removal. Normality, homoscedasticity and linearity were observed, and makes us believe that this model could be a good fit for our prediction needs. However, it cannot be used for modeling outliers.

One thing to remember in our analysis, is that we chose the variable that reported a top correlation figure. This does not indicate that one of our models will be the best candidate to predict housing prices in Ames, Iowa.

Appendix:

```
title 'Predict 410 Winter 2016 Sec 55 Assignment 3';
libname mydata '/scs/crb519/PREDICT_410/SAS_Data/' access=readonly;

data ames_stg0;
set mydata.ames_housing_data;
run;

* Create additional variables for computation and to qualify sample population;
data ames_stg1;
set ames_stg0;
format drop_condition $50.;
TotalSF = TotalBsmtSF + FirstFlrSF + SecondFlrSF;
if subclass >= 120 and subclass <= 180 then drop_condition='01: Planned Unit Development Homes';
else if Zoning ='A' or Zoning ='C' or Zoning ='FV' or Zoning ='I' then drop_condition='02: Properties in Non-Housing zones';
else if BldgType = '2FmCon' or BldgType = 'Duplx' or BldgType = 'TwnhsE' or BldgType = 'TwnhsI' then drop_condition='03: Not a single family dwelling';
else if TotalSF > 6000 then drop_condition='04: Trimming extremes in TotalSF';
else if GrLivArea < 400 OR GrLivArea > 3000 then drop_condition='05: Trimming extremes in GrLivArea';
else if GarageArea < 220 OR GarageArea > 1000 then drop_condition='06: Trimming extremes in GarageArea';
else if TotalBsmtSF < 400 OR TotalBsmtSF > 2100 then drop_condition='07: Trimming extremes in TotalBsmtSF';
else if FirstFlrSF < 400 OR FirstFlrSF > 2000 then drop_condition='08: Trimming extremes in FirstFlrSF';
else if MasVnrArea > 1000 then drop_condition='09: Trimming extremes in MasVnrArea';
else if BsmtFinSF1 > 1500 then drop_condition='10: Trimming extremes in BsmtFinSF1';
else if FullBath < 1 OR FullBath > 3 then drop_condition='11: Trimming extremes in FullBath';
else if YearBuilt < 1920 then drop_condition='12: Trimming extremes in YearBuilt';
else if OverallQual < 3 then drop_condition='13: Trimming extremes in OverallQual';
else drop_condition='14: Sample Population';
run;

* Keep only qualifed sample population in our work dataset;
data ames_stg2;
set ames_stg1;
where drop_condition='14: Sample Population';
run;

* Initial Data Exploratory Analysis;
data ames_stg3;
set ames_stg2 (keep=OverallQual TotalSF GrLivArea GarageArea TotalBsmtSF FirstFlrSF YearBuilt MasVnrArea FullBath BsmtFinSF1 SalePrice);
run;quit;
```

```
* Create log(SalePrice),log(TotalSF), log(GrLivArea) to be used later in the analysis;
data ames_stg4;
set ames_stg3;
log_SalePrice= log(SalePrice);
log_TotalSF= log(TotalSF);
log_GrLivArea= log(GrLivArea);
run;


proc corr data=ames_stg3 nosimple rank;
var BsmtFinSF1 FirstFlrSF FullBath GarageArea GrLivArea MasVnrArea OverallQual TotalBsmtSF TotalSF
YearBuilt;
with SalePrice;

proc reg data=ames_stg3 plots=(RESIDUALBYPREDICTED RESIDUALS(UNPACK));
model SalePrice=BsmtFinSF1 FirstFlrSF GarageArea GrLivArea MasVnrArea TotalBsmtSF TotalSF
/ selection=rsquare ADJRSQ;
run;


/* Simple linear regression with TotalSF and original response variable */
proc reg data=ames_stg4 PLOTS=(DIAGNOSTICS(UNPACK) RESIDUALBYPREDICTED QQPLOT
OBSERVEDBYPREDICTED);
model SalePrice = TotalSF;
output out=temp_TotalSF
    p=PREDICT_SalePricehat
    r=RESID_SalePriceresid
    RSTUDENT=RSTUDENT_SalePrice
          COOKD=COOKD_SalePrice;
run;


/* Simple linear regression with GrLivArea and original response variable */
proc reg data=ames_stg4 PLOTS=(DIAGNOSTICS(UNPACK) RESIDUALBYPREDICTED QQPLOT
OBSERVEDBYPREDICTED);
model SalePrice = GrLivArea;
output out=temp_GrLivArea
    p=PREDICT_SalePricehat
    r=RESID_SalePriceresid
    RSTUDENT=RSTUDENT_SalePrice
          COOKD=COOKD_SalePrice;
run;


/* Multiple linear regression with 2 Predictors */
proc reg data=ames_stg4 PLOTS=(DIAGNOSTICS(UNPACK) RESIDUALBYPREDICTED QQPLOT
OBSERVEDBYPREDICTED);
```

```
model SalePrice = TotalSF GrLivArea;
output out=temp_multi1
    p=PREDICT_SalePricehat
    r=RESID_SalePriceresid
    RSTUDENT=RSTUDENT_SalePrice
           COOKD=COOKD_SalePrice;
run;


proc univariate normal plot data=ames_stg4 ;
var TotalSF;
histogram TotalSF / normal;
run;quit;


proc univariate normal plot data=ames_stg4 ;
var GrLivArea;
histogram GrLivArea/ normal;
run;quit;

proc format;
value rstudent_sfmt
          3 - high  = '00: more than 3'
          2 -< 3    = '01: 2 to 3'
          1 -< 2    = '02: 1 to 2'
          0 -< 1    = '03: 0 to 1'
          -1 -< 0   = '04: -1 to 0'
          -2 -< -1  = '05: -2 to -1'
          -3 -< -2  = '06: -3 to -2'
          low -< -3 = '08: Less than -3'
          ;
run;


data temp_multi2;
set temp_multi1;
format Rstudent_Range $50.;
format CookD_Range $50.;
          Rstudent_Range = put(RSTUDENT_SalePrice,rstudent_sfmt.);
          if COOKD_SalePrice > 4/(1926-(3+1)) then CookD_Range='01: CookD Potential outlier';
          else CookD_Range='02: Not an outlier';
run;


title 'Distribution of Outliers by RStudent';
proc freq data=temp_multi2;
tables Rstudent_Range / nocol nocum nopercent norow;
```

```
run; quit;


title 'Distribution of Outliers by Cooks D';
proc freq data=temp_multi2;
tables CookD_Range / nocol nocum nopercent norow;
run; quit;


title 'Distribution of Outliers by Cooks D and RStudent';
proc freq data=temp_multi2;
tables Rstudent_Range * CookD_Range / nocol nocum nopercent norow;
run; quit;


data ames_stg5;
set temp_multi2;
format Type_of_Outlier $50.;
        if CookD_Range='01: CookD Potential outlier' then Type_of_Outlier='01: CookD Potential outlier';
        else if Rstudent_Range='00: more than 3' or Rstudent_Range='08: Less than -3' then
Type_of_Outlier='02: RStudent outlier';
        else if TotalSF >= 4196 or TotalSF <= 1489 then Type_of_Outlier='03: TotalSF outlier';
        else if GrLivArea >= 2728 or GrLivArea <= 768 then Type_of_Outlier='04: GrLivArea outlier';
        else Type_of_Outlier='05: Not an outlier';
run;


title 'Distribution of Outliers';
proc freq data=ames_stg5;
tables Type_of_Outlier;
run; quit;

data ames_stg6;
set ames_stg5;
where Type_of_Outlier='05: Not an outlier';
run;

/* Multiple linear regression with 2 Predictors, Without outliers */
proc reg data=ames_stg6 PLOTS=(DIAGNOSTICS(UNPACK) RESIDUALBYPREDICTED QQPLOT
OBSERVEDBYPREDICTED);
model SalePrice = TotalSF GrLivArea;
output out=temp_multi1
    p=PREDICT_SalePricehat
    r=RESID_SalePriceresid
    RSTUDENT=RSTUDENT_SalePrice
          COOKD=COOKD_SalePrice;
run;
```

```
/* Multiple linear regression with 2 Predictors and transformed response variable */
proc reg data=ames_stg4 PLOTS=(DIAGNOSTICS(UNPACK) RESIDUALBYPREDICTED QQPLOT
OBSERVEDBYPREDICTED);
model log_SalePrice = TotalSF GrLivArea;
output out=temp_multi2
     p=PREDICT_SalePricehat
     r=RESID_SalePriceresid
     RSTUDENT=RSTUDENT_SalePrice
              COOKD=COOKD_SalePrice;
run;


/* Multiple linear regression with 2 transformed Predictors and transformed response variable */
proc reg data=ames_stg4 PLOTS=(DIAGNOSTICS(UNPACK) RESIDUALBYPREDICTED QQPLOT
OBSERVEDBYPREDICTED);
model log_SalePrice = log_TotalSF log_GrLivArea;
output out=temp_multi3
     p=PREDICT_SalePricehat
     r=RESID_SalePriceresid
     RSTUDENT=RSTUDENT_SalePrice
              COOKD=COOKD_SalePrice;
run;


/* Multiple linear regression with 1 transformed Predictor and transformed response variable */
proc reg data=ames_stg4 PLOTS=(DIAGNOSTICS(UNPACK) RESIDUALBYPREDICTED QQPLOT
OBSERVEDBYPREDICTED);
model log_SalePrice = TotalSF log_GrLivArea;
output out=temp_multi3
     p=PREDICT_SalePricehat
     r=RESID_SalePriceresid
     RSTUDENT=RSTUDENT_SalePrice
              COOKD=COOKD_SalePrice;
run;

proc reg data=ames_stg4 PLOTS=(DIAGNOSTICS(UNPACK) RESIDUALBYPREDICTED QQPLOT
OBSERVEDBYPREDICTED);
model log_SalePrice = log_TotalSF GrLivArea;
output out=temp_multi3
     p=PREDICT_SalePricehat
     r=RESID_SalePriceresid
     RSTUDENT=RSTUDENT_SalePrice
              COOKD=COOKD_SalePrice;
run;
```

References:

Montgomery, Douglas C., Elizabeth A. Peck, and G. Geoffrey. Vining. Introduction to Linear Regression Analysis. Hoboken, NJ: Wiley, Fifth Edition. Chapters 1, 2 & 3

Ken Black, Business Statistics for Contemporary Decision Making, Wiley, Eigth Edition. Unit IV

http://www.amstat.org/publications/jse/v19n3/decock/DataDocumentation.txt