Assignment #3
Jeff Wuestling

Introduction:

The intent of this report is to summarize the second step of a data science project with a goal to provide estimates of home values for 'typical' homes in Ames, Iowa. This second step consisted of variable selection, fitting models, and model evaluation.

First, the two continuous variables with the most promise for predicting sale price were identified – living area SF and total basement SF. Those two variables were used to fit two simple regression models. The two variables were then combined to fit a multiple linear regression model. With a goal of improving the fit, outliers were identified and removed from the data set, the multiple linear regression model was fit to the new data set, and results were compared to the initial model. Last, the response variable, sale price, was log transformed and a comparison of a model with sale price and log transformed sale price was conducted. Model adequacy was evaluated in terms of goodness of fit (validation of the normality assumption and the homoscedasticity assumption) and performance was judged using the R^2 and Adjusted R^2 metrics.

A data set containing information on individual residential properties sold in Ames, Iowa between 2006 and 2010 was used (Ames dataset) and was provided by the Ames, Iowa Assessor's Office. This data is used for tax assessment purposes. The data set contains 2,930 observations with 82 variables. All coding was done using SAS (Appendix A).

Data:

For the purposes of this project, a "typical" home is defined as a single-family residence with a living area less than 4,000 square feet (SF). In order to isolate the sales of typical homes several drop conditions were applied to the Ames dataset. Table 1 lists those conditions along with a description of the logic used to implement it.

Drop Condition	Description
01: BldgType	BldgType is not "1Fam"
02: GTE 4000 SF	GrLivArea is greater than or equal to 4,000 SF
03: Functional	Functional is not 'Typ'
04: Sale Condition	SaleCondition is not 'Normal'
05: Zoning	Zoning is 'A', 'C','FV', or 'I'
06: Subclass	Subclass is '90' or '190'

Table 1

The total number of observations remaining after the drop conditions were applied to the data set was 1,793 or about 61% of the total observations (Figure 1). There were no observations dropped due to Drop Condition #06, however it was left in the waterfall for future reference. The final data set included one additional variable, the log transformed sale price.

drop_condition	Frequency	Percent	Cumulative Frequency	Cumulative Percent
01: Bldg Type	505	17.24	505	17.24
02: GTE 4000 SF	5	0.17	510	17.41
03: Functional	179	6.11	689	23.52
04: Sale Condition	394	13.45	1083	36.96
05: Zoning	54	1.84	1137	38.81
07: Sample Population	1793	61.19	2930	100.00

Figure 1

Variable Selection:

The variable selection process began by identifying all the continuous variables in addition to the response variable, sale price, in the Ames dataset (Table 2).

#	Variable	Description
1	BsmtFinSF1	(Continuous): Type 1 finished square feet
2	BsmtFinSF2	(Continuous): Type 2 finished square feet
3	BsmtUnfSF	(Continuous): Unfinished square feet of basement area
4	EnclosedPorch	(Continuous): Enclosed porch area in square feet
5	FirstFlrSF	(Continuous): First Floor square feet
6	GarageArea	(Continuous): Size of garage in square feet
7	GrLivArea	(Continuous): Above grade (ground) living area square feet
8	LotArea	(Continuous): Lot size in square feet
9	LotFrontage	(Continuous): Linear feet of street connected to property
10	LowQualFinSF	(Continuous): Low quality finished square feet (all floors)
11	MasVnrArea	(Continuous): Masonry veneer area in square feet
12	MiscVal	(Continuous): $Value of miscellaneous feature
13	OpenPorchSF	(Continuous): Open porch area in square feet
14	PoolArea	(Continuous): Pool area in square feet
15	ScreenPorch	(Continuous): Screen porch area in square feet
16	SecondFlrSF	(Continuous) : Second floor square feet
17	ThreeSsnPorch	(Continuous): Three season porch area in square feet
18	TotalBsmtSF	(Continuous): Total square feet of basement area
19	WoodDeckSF	(Continuous): Wood deck area in square feet

Table 2

The continuous variables seem to fall in to one of four general categories: value adjusted SF (e.g. added feature, low quality SF), home features (e.g. porch, garage, pool), lot size (e.g. lot area, lot frontage), and interior SF (e.g. first floor, second floor, basement).

To better understand the continuous variables as well as the relationship between each variable and the response variable, sale price, the continuous variables were examined by category. Simple statistics, the Pearson Correlation table, and a scatter plot of each

continuous variable along with the response variable, sale price, were generated and reviewed. The scatter plots included a LOESS line to aid in the visual examination of the plot.

The value adjusted SF variables were sporadic at best. As seen in Figure 2, there was not a high correlation between any these variables and sale price.

The CORR Procedure

4 Variables:	SalePrice LowQualFinSF MasVnrArea MiscVal

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
SalePrice	1793	180373	72432	323409048	35000	625000
LowQualFinSF	1793	3.47630	37.50129	6233	0	572.00000
MasVnrArea	1784	98.00056	172.22331	174833	0	1378
MiscVal	1793	53.86782	551.45733	96585	0	15500

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

	SalePrice	LowQualFinSF	MasVnrArea	MiscVal
SalePrice	1.00000 1793	-0.01921 0.4162 1793	0.56833 <.0001 1784	-0.01831 0.4383 1793
LowQualFinSF	-0.01921 0.4162 1793	1.00000 1793	-0.05291 0.0254 1784	-0.00265 0.9108 1793
MasVnrArea	0.56833 <.0001 1784	-0.05291 0.0254 1784	1.00000 1784	-0.01420 0.5490 1784
MiscVal	-0.01831 0.4383 1793	-0.00265 0.9108 1793	-0.01420 0.5490 1784	1.00000 1793

Figure 2

Many of the home features contained a lot of zeros, indicating the property did not have that specific feature, which was seen in many of the scatter plots. However, the Pearson Correlation table revealed that the garage area did have a modest correlation of .65 to sale price (Figure 3).

The CORR Procedure

8 Variables:	SalePrice	EnclosedPorch	GarageArea	OpenPorchSF	PoolArea	ScreenPorch	ThreeSsnPorch	WoodDeckSF

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
SalePrice	1793	180373	72432	323409048	35000	625000
EnclosedPorch	1793	24.35694	62.71041	43672	0	584.00000
GarageArea	1793	467.38483	198.67129	838021	0	1488
OpenPorchSF	1793	45.11991	61.80478	80900	0	502.00000
PoolArea	1793	1.17959	24.67723	2115	0	648.00000
ScreenPorch	1793	17.41383	59.76111	31223	0	576.00000
ThreeSsnPorch	1793	2.74958	26.91252	4930	0	508.00000
WoodDeckSF	1793	99.42052	132.69667	178261	0	1424

Pearson Correlation Coefficients, N = 1793
Prob > |r| under H0: Rho=0

	SalePrice	EnclosedPorch	GarageArea	OpenPorchSF	PoolArea	ScreenPorch	ThreeSsnPorch	WoodDeckSF
SalePrice	1.00000	-0.12953 <.0001	0.65108 <.0001	0.35389 <.0001	0.03885 0.1001	0.08920 0.0002	0.01278 0.5886	0.35082 <.0001
EnclosedPorch	-0.12953 <.0001	1.00000	-0.09698 <.0001	-0.08769 0.0002	-0.00468 0.8430	-0.09173 0.0001	-0.03802 0.1076	-0.12541 <.0001
GarageArea	0.65108 <.0001	-0.09698 <.0001	1.00000	0.24199 <.0001	0.01084 0.6466	0.03950 0.0945	0.00815 0.7303	0.23415 <.0001
OpenPorchSF	0.35389 <.0001	-0.08769 0.0002	0.24199 <.0001	1.00000	-0.00488 0.8363	0.02243 0.3426	-0.01492 0.5278	0.05005 0.0341
PoolArea	0.03885 0.1001	-0.00468 0.8430	0.01084 0.6466	-0.00488 0.8363	1.00000	0.08424 0.0004	-0.00489 0.8362	0.02553 0.2800
ScreenPorch	0.08920 0.0002	-0.09173 0.0001	0.03950 0.0945	0.02243 0.3426	0.08424 0.0004	1.00000	-0.02979 0.2074	-0.08381 0.0004
ThreeSsnPorch	0.01278 0.5886	-0.03802 0.1076	0.00815 0.7303	-0.01492 0.5278	-0.00489 0.8362	-0.02979 0.2074	1.00000	-0.02357 0.3185
WoodDeckSF	0.35082 <.0001	-0.12541 <.0001	0.23415 <.0001	0.05005 0.0341	0.02553 0.2800	-0.08381 0.0004	-0.02357 0.3185	1.00000

Figure 3

In Figure 4, a positive linear relationship is seen in the scatter plot of garage area against sale price, but the variance increases substantially as the garage size approaches 750 SF.

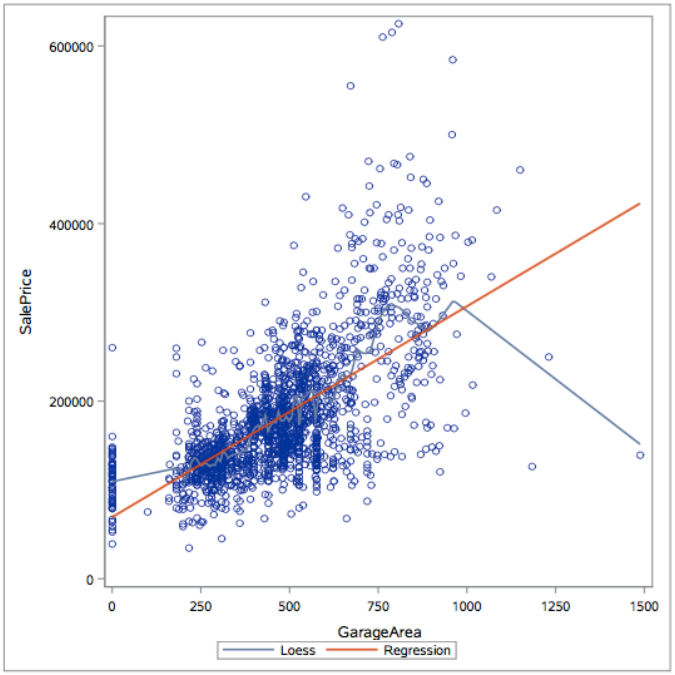

Figure 4

The results of the analysis of the lot related variables were mixed. Although there appears to be a positive linear relationship between both lot area (Figure 5) and lot frontage (Figure 6) to sale price, neither variable showed a high correlation to sale price (Figure 7).

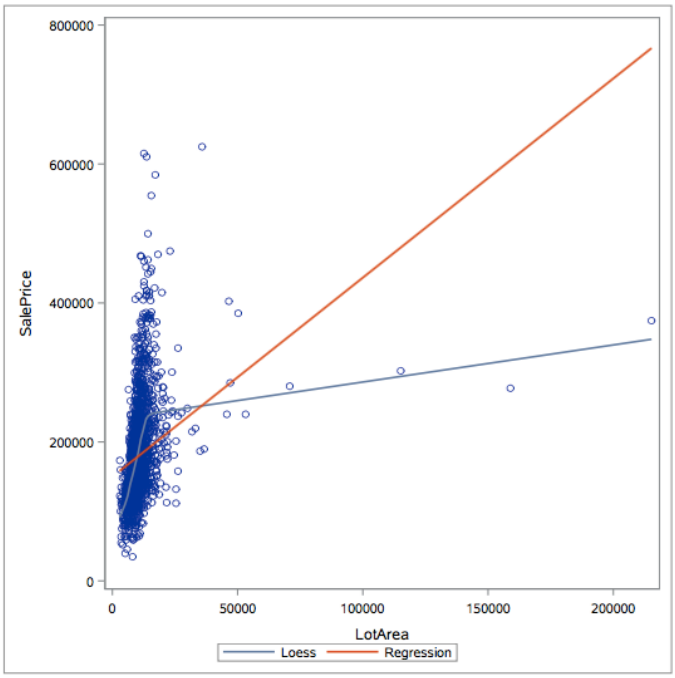

Figure 5

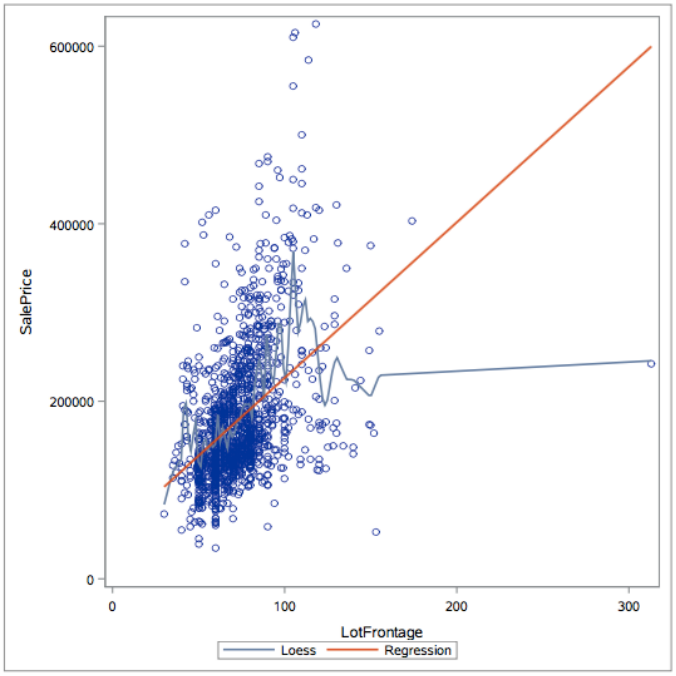

Figure 6

The CORR Procedure

3 Variables:	SalePrice	LotArea	LotFrontage

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
SalePrice	1793	180373	72432	323409048	35000	625000
LotArea	1793	10678	7793	19146216	2887	215245
LotFrontage	1443	72.64588	19.51652	104828	30.00000	313.00000

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

	SalePrice	LotArea	LotFrontage
SalePrice	1.00000 1793	0.30840 <.0001 1793	0.45414 <.0001 1443
LotArea	0.30840 <.0001 1793	1.00000 1793	0.33229 <.0001 1443
LotFrontage	0.45414 <.0001 1443	0.33229 <.0001 1443	1.00000 1443

Figure 7

The last category contained the two variables that looked the most promising (living area SF and the total basement SF). As illustrated in Figure 8, there is a strong positive linear relationship between sale price and living area SF. However, it was noted that the plot begins to funnel out when the SF of the home nears 2,000 SF.

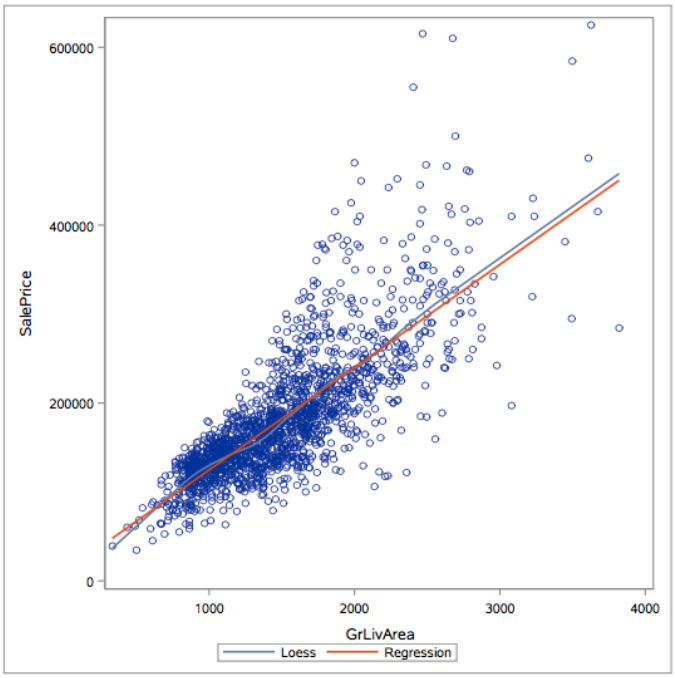

Figure 8

There is also a positive linear relationship between sale price and total basement SF (Figure 9).

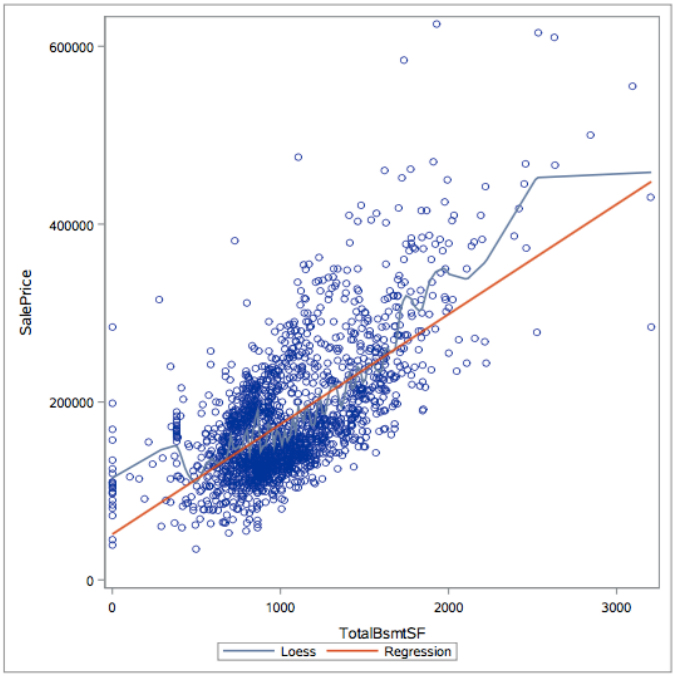

Figure 9

At .788, the living area SF had the highest correlation to sale price of any of the continuous variables (Figure 10).

The CORR Procedure

8 Variables:	SalePrice BsmtFinSF1 BsmtFinSF2 BsmtUnfSF FirstFlrSF GrLivArea SecondFlrSF TotalBsmtSF

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
SalePrice	1793	180373	72432	323409048	35000	625000
BsmtFinSF1	1793	449.84941	420.42789	806580	0	2288
BsmtFinSF2	1793	55.68600	178.13413	99845	0	1526
BsmtUnfSF	1793	538.06414	397.47849	964749	0	2336
FirstFlrSF	1793	1133	358.26446	2031768	334.00000	3820
GrLivArea	1793	1482	494.35559	2656768	334.00000	3820
SecondFlrSF	1793	345.10151	430.33536	618767	0	1836
TotalBsmtSF	1793	1044	387.46567	1871174	0	3206

Pearson Correlation Coefficients, N = 1793 Prob > |r| under H0: Rho=0								
	SalePrice	BsmtFinSF1	BsmtFinSF2	BsmtUnfSF	FirstFlrSF	GrLivArea	SecondFlrSF	TotalBsmtSF
SalePrice	1.00000	0.45296 <.0001	0.02313 0.3277	0.15537 <.0001	0.68074 <.0001	0.78792 <.0001	0.34008 <.0001	0.66152 <.0001
BsmtFinSF1	0.45296 <.0001	1.00000	-0.06938 0.0033	-0.50469 <.0001	0.46129 <.0001	0.17037 <.0001	-0.18320 <.0001	0.53544 <.0001
BsmtFinSF2	0.02313 0.3277	-0.06938 0.0033	1.00000	-0.26325 <.0001	0.09420 <.0001	-0.02886 0.2219	-0.11101 <.0001	0.11440 <.0001
BsmtUnfSF	0.15537 <.0001	-0.50469 <.0001	-0.26325 <.0001	1.00000	0.26634 <.0001	0.22128 <.0001	0.03062 0.1951	0.35718 <.0001
FirstFlrSF	0.68074 <.0001	0.46129 <.0001	0.09420 <.0001	0.26634 <.0001	1.00000	0.52362 <.0001	-0.22975 <.0001	0.81706 <.0001
GrLivArea	0.78792 <.0001	0.17037 <.0001	-0.02886 0.2219	0.22128 <.0001	0.52362 <.0001	1.00000	0.70557 <.0001	0.39859 <.0001
SecondFlrSF	0.34008 <.0001	-0.18320 <.0001	-0.11101 <.0001	0.03062 0.1951	-0.22975 <.0001	0.70557 <.0001	1.00000	-0.21841 <.0001
TotalBsmtSF	0.66152 <.0001	0.53544 <.0001	0.11440 <.0001	0.35718 <.0001	0.81706 <.0001	0.39859 <.0001	-0.21841 <.0001	1.00000

Figure 10

Linear Regression Models:

Simple Linear Regression – Model # 1

The first simple linear regression model was fitted with living area SF. The least squares fit is

$$y = 9313.49 + 115.44x$$

Visual inspection of the QQ plot (Figure 11) revealed a mostly normal distribution of the residuals. However, the tails are both "thick" which indicates possible outliers or influential observations.

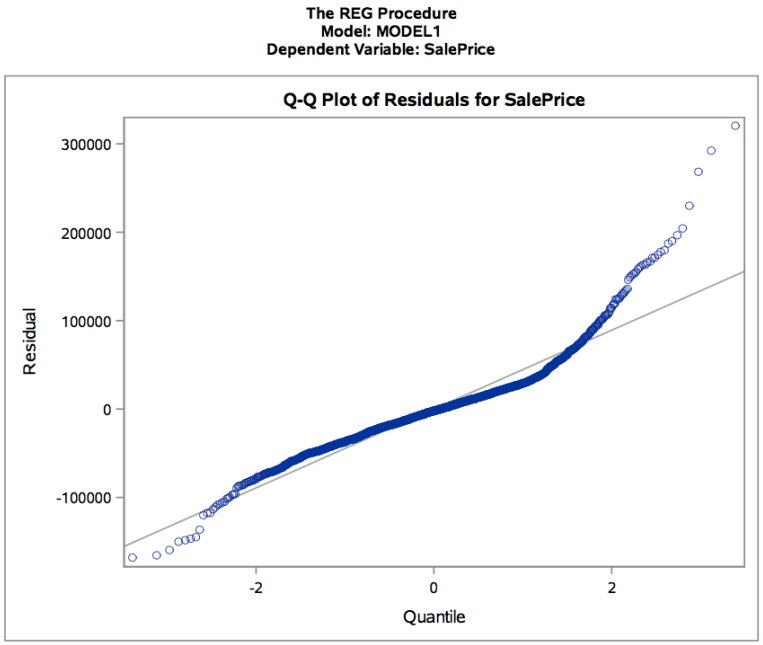

Figure 11

No pattern was seen in the scatter plot of the residuals against living area SF, but it was noted that there is a substantial increase in the variance of the residuals beginning around 2,000 SF (Figure 12).

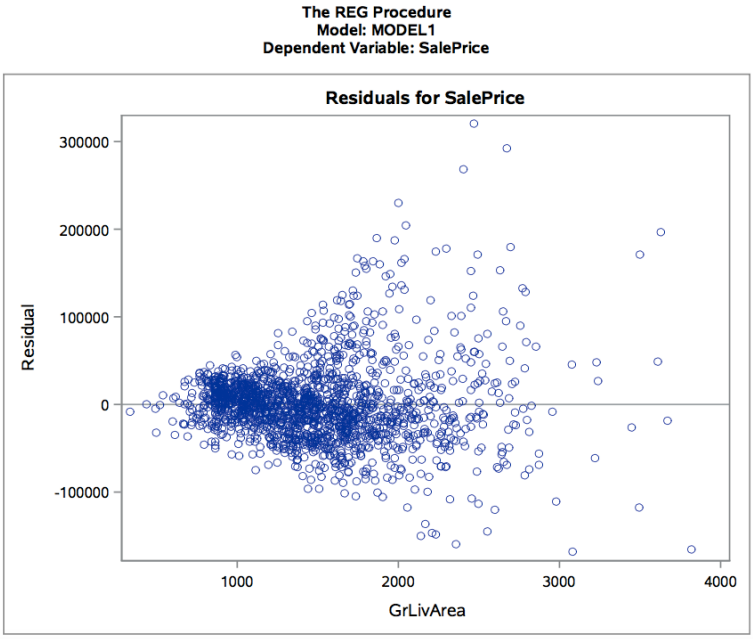

Figure 12

The SAS output from this model is shown in Figure 13. The Overall F-test, found in the Analysis of Variance (ANOVA) table, indicates that this is a significant model. The null hypothesis that the predictor variable's coefficient is equal to zero can be rejected with the F Statistic of 2932.30 and a p-value of less than .0001. The t-test for the predictor variable found in the Parameter Estimates table had the same conclusion. The R^2 and Adjusted R^2 were extremely close and indicate that 62% of the change in y can be predicted by the change in x.

GrLivArea

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Number of Observations Read	1793
Number of Observations Used	1793

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	5.836674E12	5.836674E12	2932.30	<.0001
Error	1791	3.564944E12	1990476573		
Corrected Total	1792	9.401617E12			

Root MSE	44615	R-Square	0.6208
Dependent Mean	180373	Adj R-Sq	0.6206
Coeff Var	24.73470		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	9313.48782	3330.03508	2.80	0.0052
GrLivArea	1	115.44477	2.13192	54.15	<.0001

Figure 13

Simple Linear Regression – Model # 2

The second simple linear regression model was fitted with total basement SF. The least squares fit is

$$\mathbf{y = 51{,}318 + 123.67x}$$

As seen in Figure 14, the QQ plot of the residuals shows a significant departure from a normal distribution. This finding poses a risk that the assumption of normality cannot be validated. The statistical inference must be used with caution.

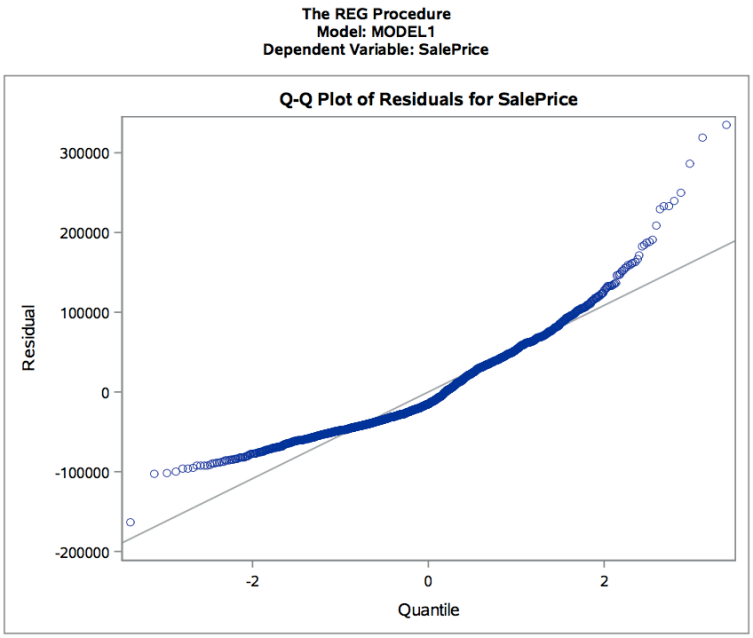

Figure 14

The scatter plot of the residuals against the total basement SF did not show any patterns but the plot does highlight several outliers (Figure 15).

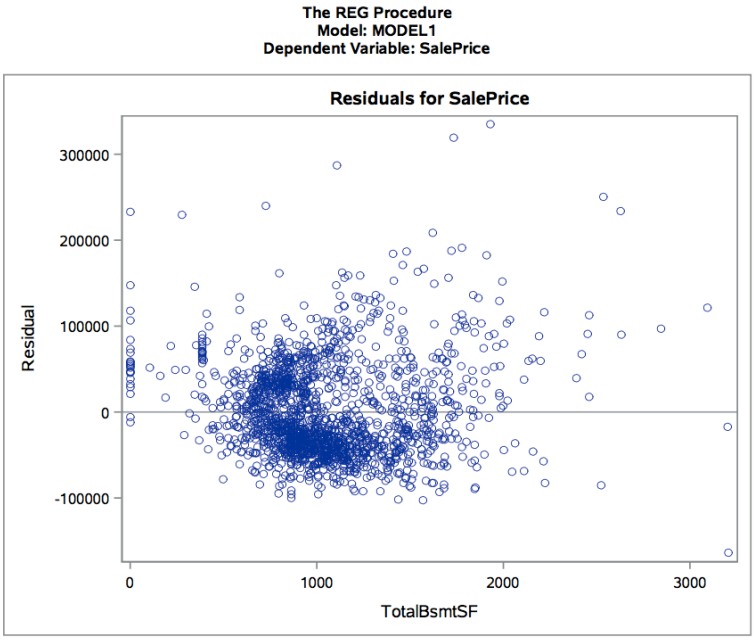

Figure 15

The SAS output of this model is shown in Figure 16. The regression effect was examined with caution. The null hypothesis that the predictor variable's coefficient is equal to zero can be rejected with the F Statistic of 1393.61 and a p-value of less than .0001. The t-test for the predictor variable found in the Parameter Estimates table had the same conclusion.

The R^2 and Adjusted R^2 were extremely close and indicate that close to 44% of the change in y can be predicted by the change in x. This is almost 20% lower than the first model.

TotalBsmtSF

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Number of Observations Read	1793
Number of Observations Used	1793

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	4.114221E12	4.114221E12	1393.61	<.0001
Error	1791	5.287396E12	2952203420		
Corrected Total	1792	9.401617E12			

Root MSE	54334	R-Square	0.4376
Dependent Mean	180373	Adj R-Sq	0.4373
Coeff Var	30.12321		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	51318	3687.50092	13.92	<.0001
TotalBsmtSF	1	123.66351	3.31261	37.33	<.0001

Figure 16

Multiple Linear Regression Model – Model # 3

The third model fitted both the living area SF and total basement SF. The least squares fit is

$$y = -35{,}529 + 91.32x1 + 77.22x2$$

Similar to model #1, the QQ plot for this model (Figure 17) shows “thick” tails. This suggests there are outliers in the data set that should be investigated.

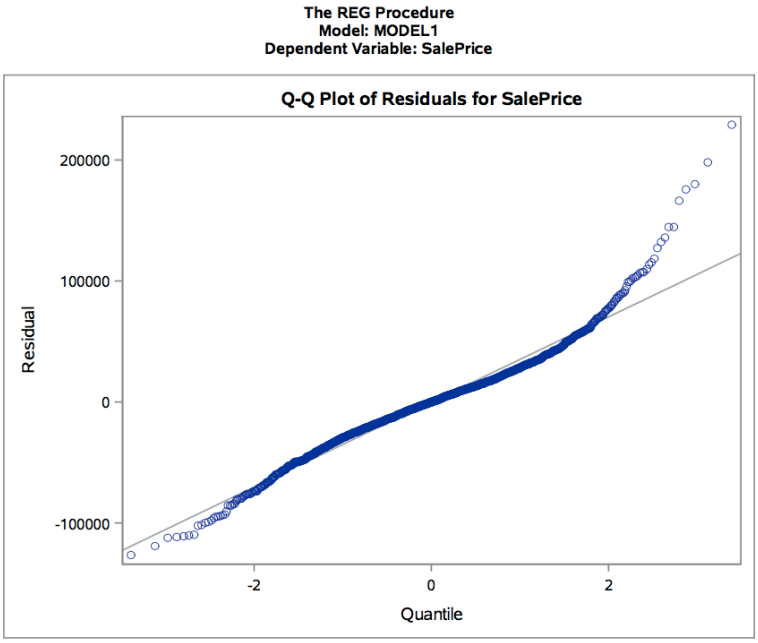

Figure 17

A moderate funnel was detected in the scatter plot of the residuals against the total living area SF (Figure 18).

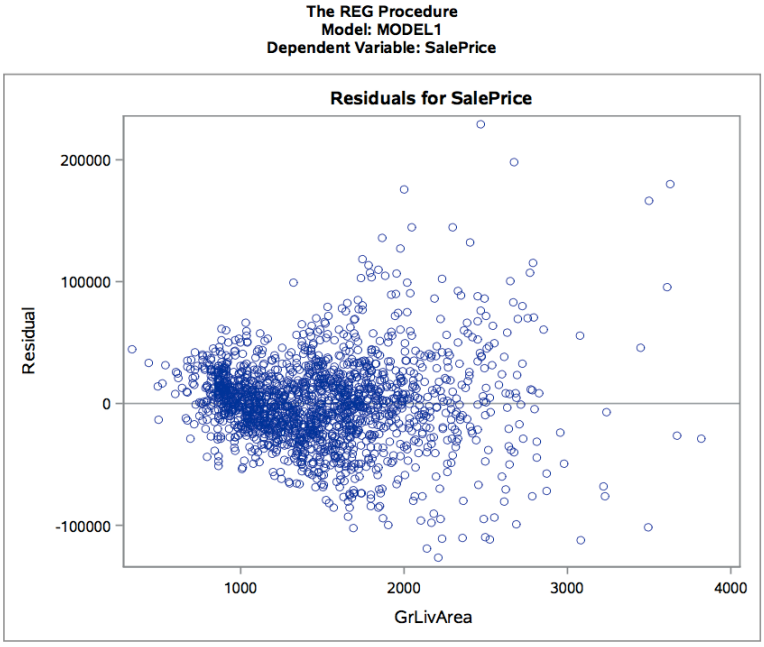

Figure 18

The scatter plot of the residuals against the total basement SF had a slight funnel effect and also highlighted several outliers (Figure 19).

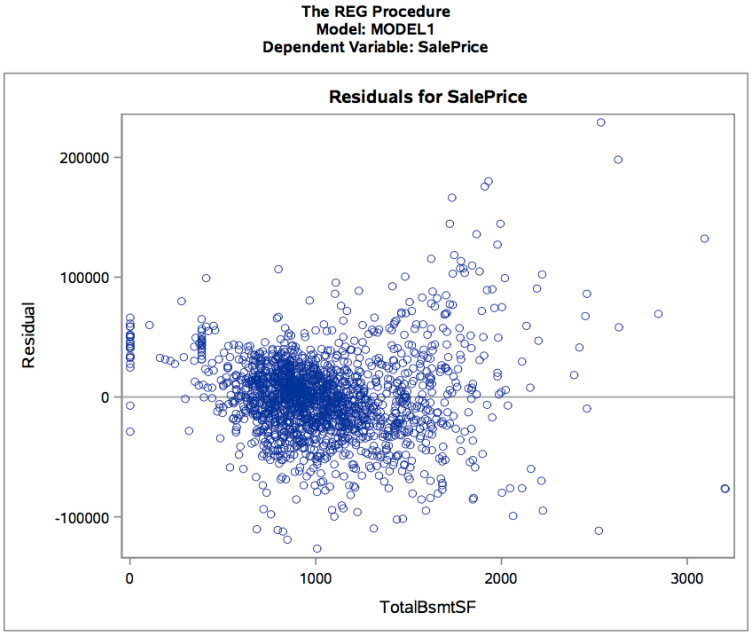

Figure 19

The SAS output of this model is shown in Figure 20. The null hypothesis that at least one of the predictor variables' coefficients is equal to zero can be rejected with the F Statistic of 2902.97 and a p-value of less than .0001. The t-test for both predictor variables found in the Parameter Estimates indicated that both contribute to the model. In each case, the null hypothesis that the coefficient is equal to zero can be rejected. The R^2 and Adjusted R^2 were extremely close and indicate that about 76.4% of the change in y can be predicted by the change in x. In this situation, adding another predictor variable greatly improved the performance of the model.

GrLivArea - TotalBsmtSF

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Number of Observations Read	1793
Number of Observations Used	1793

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	7.186103E12	3.593052E12	2902.97	<.0001
Error	1790	2.215514E12	1237717184		
Corrected Total	1792	9.401617E12			

Root MSE	35181	R-Square	0.7643
Dependent Mean	180373	Adj R-Sq	0.7641
Coeff Var	19.50468		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-35529	2956.31269	-12.02	<.0001
GrLivArea	1	91.31967	1.83304	49.82	<.0001
TotalBsmtSF	1	77.22242	2.33872	33.02	<.0001

Figure 20

Outlier Identification

During the goodness-of-fit analysis several outliers were observed (Figures 12, 15, 18 and 19). With the goal of improving how well the multiple linear regression model fit the data, an attempt was made to identify those outliers so that they could be excluded from the data set. Since the model contained both the living area SF and total basement SF, this was the main area of focus. The garage area was also included because it was found to be a relevant variable during the variable selection process.

Four outlier definitions were used to identify and remove these influential observations. In total, 471 observations were removed, leaving 1,322 or about 74% of the original observations. Figure 21 contains these details.

outlier_def	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Def 1 - Greater than 2,000 SF	248	13.83	248	13.83
Def 2 - Less than 900 SF	175	9.76	423	23.59
Def 3 - No basement	13	0.73	436	24.32
Def 4 - No Garage	35	1.95	471	26.27
Not an outlier	1322	73.73	1793	100.00

Figure 21

Multiple Linear Regression Model – Outliers Removed

The third model was re-fitted with both the living area SF and total basement SF using the new data set with the outliers removed. The least squares fit is

$$y = -34{,}136 + 92.9x1 + 71.7x2$$

The distribution of the residuals from this model more closely resembles a normal distribution (Figure 22) but there is still a "thick" tail at the high end. This suggests some outliers remain in the data set and should be investigated further.

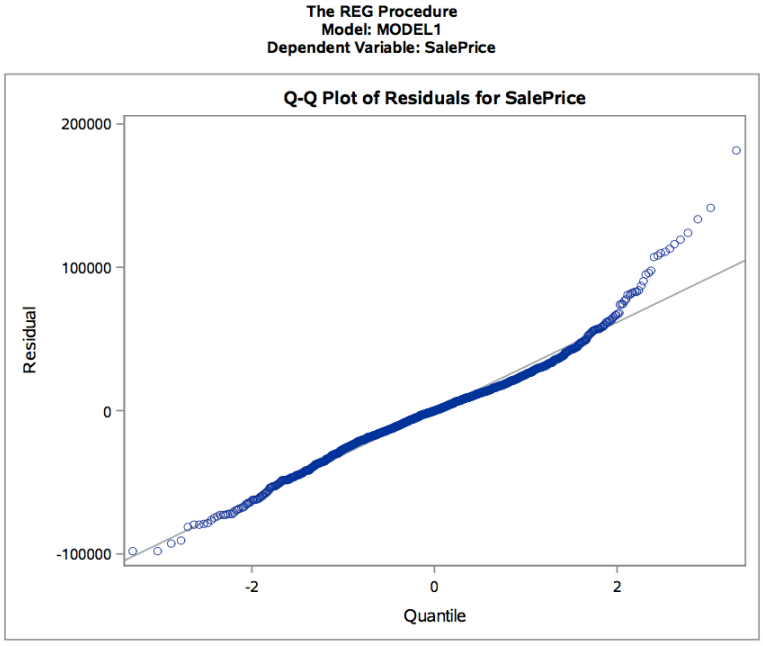

Figure 22

The scatter plot of sale price against living area SF (Figure 23) reveals a much tighter funnel than the previous model.

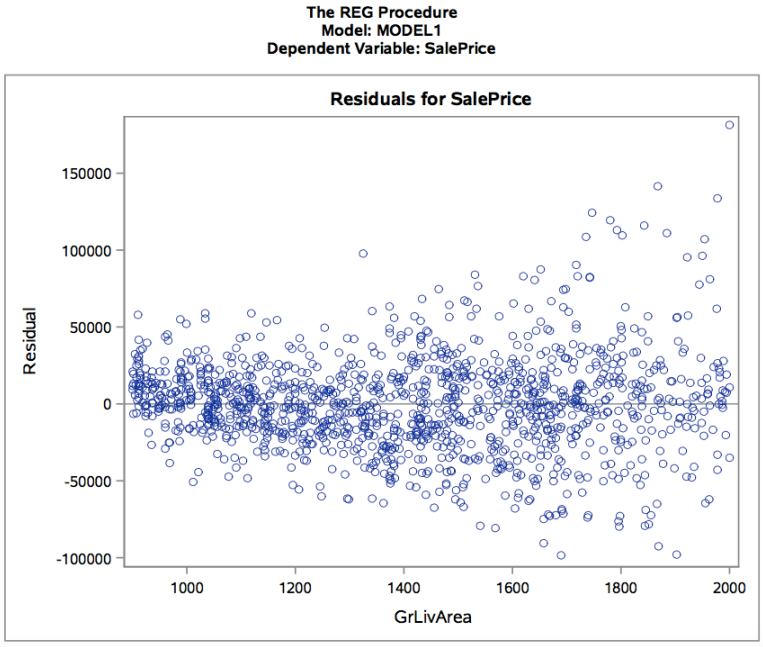

Figure 23

As Figure 24 illustrates, most of the variance in total basement SF is seen just below 2,000 SF.

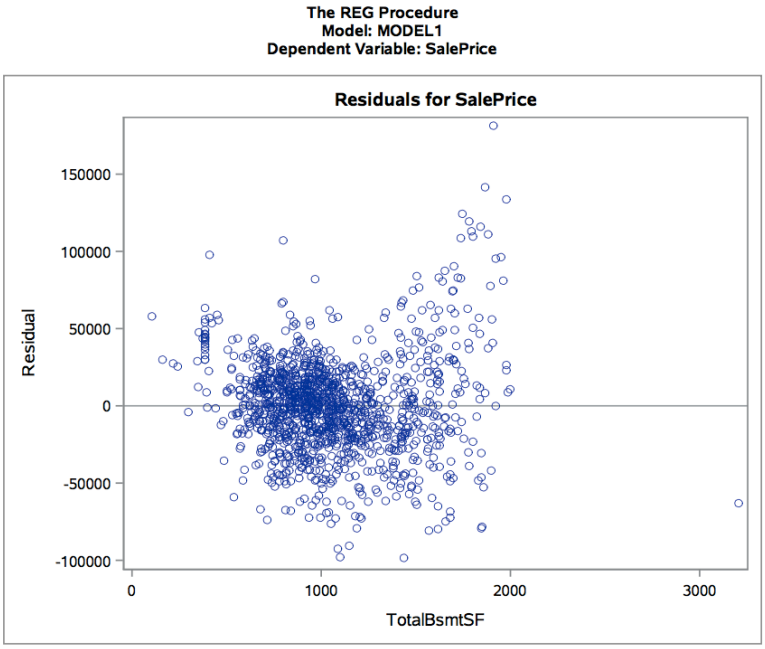

Figure 24

The SAS output of this model is shown in Figure 25. The null hypothesis that at least one of the predictor variables' coefficients is equal to zero can be rejected with the F Statistic of 1167.44 and a p-value of less than .0001. The t-test for both predictor variables found in the Parameter Estimates indicated that both contribute to the model. The R^2 and

Adjusted R^2 were extremely close and indicate that about 64% of the change in y can be predicted by the change in x.

GrLivArea - TotalBsmtSF

The REG Procedure
Model: MODEL1
Dependent Variable: SalePrice

Number of Observations Read	1322
Number of Observations Used	1322

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2.217951E12	1.108976E12	1167.44	<.0001
Error	1319	1.252941E12	949917293		
Corrected Total	1321	3.470892E12			

Root MSE	30821	R-Square	0.6390
Dependent Mean	172253	Adj R-Sq	0.6385
Coeff Var	17.89268		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-34136	4455.19523	-7.66	<.0001
GrLivArea	1	92.89757	2.94033	31.59	<.0001
TotalBsmtSF	1	71.69778	2.56359	27.97	<.0001

Figure 25

In terms of R^2 and Adjusted R^2, the model fit with the data set that had the outliers removed did not perform as well as the model fit with the data set that contained the outliers. The model with the outliers present had an Adjusted R^2 of .7641 compared to the Adjusted R^2 of .6385 for the model with the outliers removed.

Metric	Model with Outliers Present	Model with Outliers Removed
R^2	.7643	.6369
Adjusted R^2	.7641	.6385

Table 3

Based on the results of this exercise it would be prudent to take another look at the potential outliers and to revise the outlier definitions.

Model Comparison of Y versus log(Y)

The final task of this second step was to explore the use of a transform response variable sale price. The multiple linear regression model was re-fit using the log transform sale price (log(saleprice)).

A side by side comparison of the QQ plots (Figure 26) reveals a closer to normal distribution of the residuals from the log(saleprice).

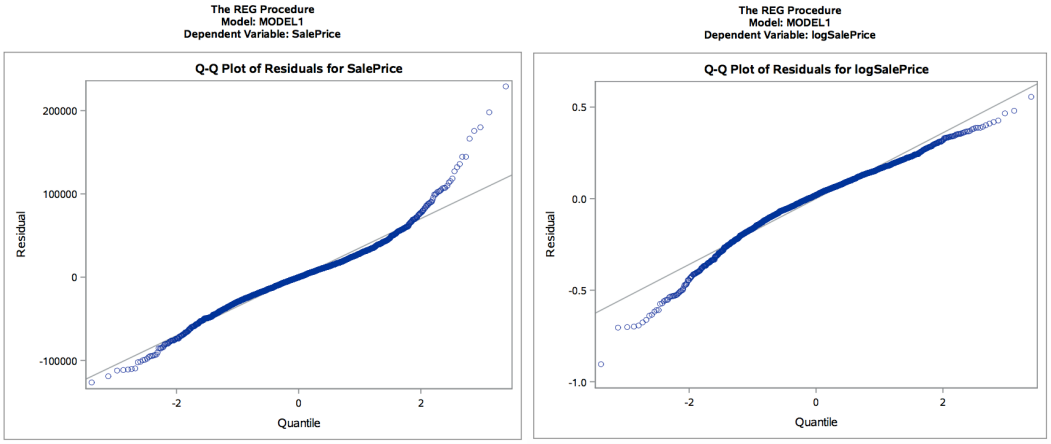

Figure 26

The funnel effect seen in the scatter plot of the residuals against living area SF was not present in the log(saleprice) scatter plot (Figure 27).

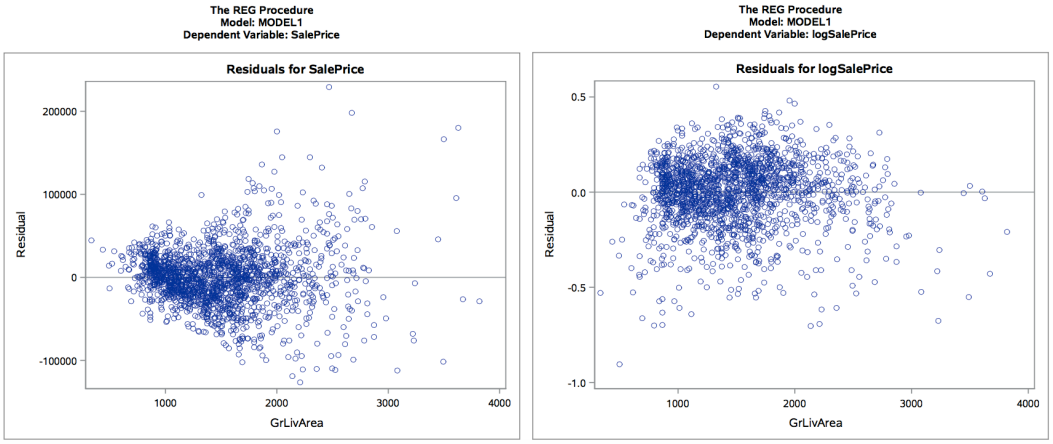

Figure 27

The comparison of the scatter plots of the residuals against the total basement SF had a similar outcome to that of the comparison between the plots with the living area SF. The plot with the log(saleprice) did not show a funnel effect (Figure 28).

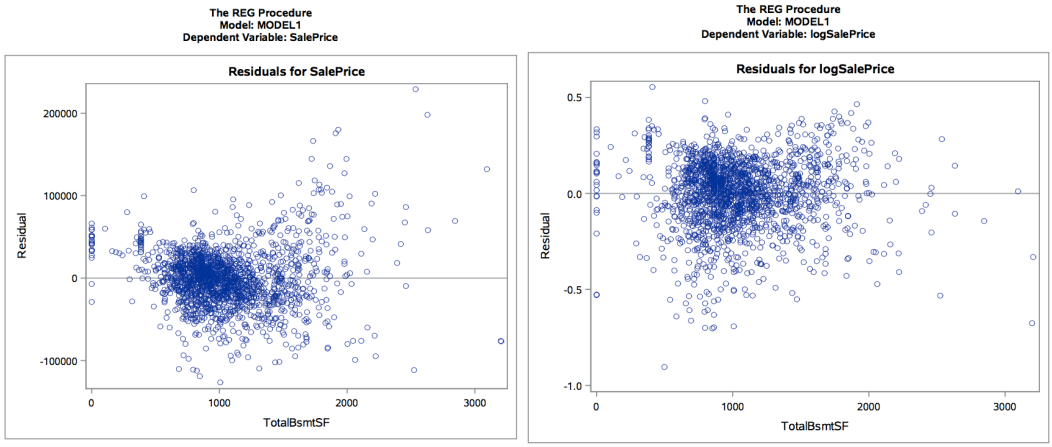

Figure 28

The SAS output of this model is shown in Figure 29. The null hypothesis that at least one of the predictor variables' coefficients is equal to zero can be rejected with the F Statistic of 2816.94 and a p-value of less than .0001. The t-test for both predictor variables found in the Parameter Estimates indicated that both contribute to the model. The R^2 and Adjusted R^2 were extremely close and indicate that just about 76% of the change in y can be predicted by the change in x.

logSalePrice = GrLivArea - TotalBsmtSF

The REG Procedure
Model: MODEL1
Dependent Variable: logSalePrice

Number of Observations Read	1793
Number of Observations Used	1793

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	182.16490	91.08245	2816.94	<.0001
Error	1790	57.87749	0.03233		
Corrected Total	1792	240.04239			

Root MSE	0.17982	R-Square	0.7589
Dependent Mean	12.03357	Adj R-Sq	0.7586
Coeff Var	1.49429		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	10.94884	0.01511	724.60	<.0001
GrLivArea	1	0.00047638	0.00000937	50.85	<.0001
TotalBsmtSF	1	0.00036303	0.00001195	30.37	<.0001

Figure 29

Although the model with log(saleprice) seems to fit the data better than the raw sale price, the model was slightly less accurate in terms of R^2 and Adjusted R^2. A side-by-side comparison on those metrics is shown in Table 4.

Metric	saleprice	log(saleprice)
R^2	.7643	.7589
Adjusted R^2	.7641	.7586

Table 4

Conclusion:

The results from this phase of the project were very insightful. The two predictor variables selected for modeling were the living area SF and total basement SF. The simple linear regression models, one with each predictor variable, were significant but did not prove to be very predictive. The multiple linear regression model with both predictor variables was significant and had an R^2 of almost 76%. It was clearly the best

performing model. Further exploration could be done using a predictor variable with the home's total SF (living area SF + total basement SF) and additional predictor variables. The model with the log transform sale price fit better than the model with the raw sale price, but was slightly less accurate. It is important to remember that the transformed sale price must be converted back before it can be used in the appropriate context. The model comparison also revealed that the data set might contain sales for larger homes that may not be representative of a 'typical' home. Further investigation should be done in this area.

Appendix A: SAS Code

```
/* waterfall */
data temp;
		set mydata.ames_housing_data;
		format drop_condition $30.;
		if (BldgType ne '1Fam') then drop_condition='01: Bldg Type';
		else if (GrLivArea >= 4000) then drop_condition='02: GTE 4000 SF';
		else if (Functional ne 'Typ') then drop_condition='03: Functional';
		else if (SaleCondition ne 'Normal') then drop_condition='04: Sale Condition';
		else if (Zoning = 'A' or Zoning = 'C' or Zoning = 'FV' or Zoning = 'I') then drop_condition='05: Zoning';
		else if (Subclass = '90' or Subclass = '190') then drop_condition='06: Subclass';
		else drop_condition='07: Sample Population';
run;

proc freq data=temp;
tables drop_condition;
run; quit;

/* final data set */
data temp2;
set temp (where=(drop_condition='07: Sample Population'));
logSalePrice=log(SalePrice);
run;

/* value adjusted SF */
proc corr data = temp2;
var SalePrice LowQualFinSF MasVnrArea MiscVal;
run;

/* home features */
proc corr data = temp2;
var SalePrice EnclosedPorch GarageArea GrLivArea OpenPorchSF PoolArea ScreenPorch ThreeSsnPorch WoodDeckSF;
run;

/* interior SF */
proc corr data = temp2;
var SalePrice BsmtFinSF1 BsmtFinSF2 BsmtUnfSF FirstFlrSF SecondFlrSF TotalBsmtSF;
run;

/* lot size */
proc corr data = temp2;
var SalePrice LotArea LotFrontage;
run;

/* scatter plot */
ods graphics on;
proc sgscatter data=temp2;
compare x=WoodDeckSF y=SalePrice / loess reg;
run; quit;

/* models */
ods graphics on;
proc reg data=temp2 plots(unpack)=(diagnostics fitplot residuals);
model SalePrice = GrLivArea;
title 'GrLivArea';
run; quit;
ods graphics off;

ods graphics on;
proc reg data=temp2 plots(unpack)=(diagnostics fitplot residuals);
model SalePrice = TotalBsmtSF;
title 'TotalBsmtSF';
run; quit;
ods graphics off;

ods graphics on;
proc reg data=temp2 plots(unpack)=(diagnostics fitplot residuals);
```

```
model SalePrice = GrLivArea TotalBsmtSF;
title 'GrLivArea - TotalBsmtSF';
run; quit;
ods graphics off;

/* models with log transformed SalePrice */
ods graphics on;
proc reg data=temp2 plots(unpack)=(diagnostics fitplot residuals);
model logSalePrice = GrLivArea;
title 'logSalePrice = GrLivArea';
run; quit;
ods graphics off;

ods graphics on;
proc reg data=temp2 plots(unpack)=(diagnostics fitplot residuals);
model logSalePrice = TotalBsmtSF;
title 'logSalePrice = TotalBsmtSF';
run; quit;
ods graphics off;

ods graphics on;
proc reg data=temp2 plots(unpack)=(diagnostics fitplot residuals);
model logSalePrice = GrLivArea TotalBsmtSF;
title 'logSalePrice = GrLivArea - TotalBsmtSF';
run; quit;
ods graphics off;

/* outliers */
data temp3;
        set temp2;
        format outlier_def $30.;
        if (GrLivArea > 2000) then do;
                outlier_def='Def 1 - Greater than 2,000 SF';
                outlier_code = 1;
        end;
        else if (GrLivArea < 900) then do;
                outlier_def='Def 2 - Less than 900 SF';
                outlier_code = 2;
        end;
        else if (TotalBsmtSF = 0) then do;
                outlier_def='Def 3 - No basement';
                outlier_code = 3;
        end;
        else if (GarageArea = 0) then do;
                outlier_def='Def 4 - No Garage';
                outlier_code = 4;
        end;
        else do;
                outlier_def='Not an outlier';
                outlier_code = 0;
        end;
run;

proc freq data=temp3;
tables outlier_def;
run; quit;
```