1. Sample population.

The Ames data sets contain all sorts of buildings but in this study I am only interested to include residential facilities. Classification variables for those are RH, RL, RP, RM.

```
MS Zoning (Nominal): Identifies the general zoning classification of the sale.

       A       Agriculture
       C       Commercial
       FV      Floating Village Residential
       I       Industrial
       RH      Residential High Density
       RL      Residential Low Density
       RP      Residential Low Density Park
       RM      Residential Medium Density
```

Another criteria used in selection of sample population was building type. Only one family houses had been chosen.

```
Bldg Type (Nominal): Type of dwelling

       1Fam    Single-family Detached
       2FmCon  Two-family Conversion; originally built as one-family dwelling
       Duplx   Duplex
       TwnhsE  Townhouse End Unit
       TwnhsI  Townhouse Inside Unit
```

Buildings with bathroom will be excluded from the sample population. The result will be fairly homogenous sample population that I will use to explain sale price of one family, residential house in Ames.

```
data drop;
        set ames;
        format drop_condition $40.;
        if      (BldgType ne '1Fam')                        then drop_condition='01: Not a 
Single Family';
        else if (Zoning not in ('RH','RL','RM','FV'))      then drop_condition='02: Non-
Residential Zoning';
        else if (FullBath < 1)                              then drop_condition='04: No Bath';
    else drop_condition='99: Sample Population';
```

2. Simple linear regression model

Since simple linear regression require continuous variables I investigate which could two of them could be most relevant for the studie.

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations	
	SalePrice
TotalBsmtSF	0.63228 <.0001 2929
GrLivArea	0.70678 <.0001 2930

TotalBsmtSF and GrLivArea have both low p-value and strong pearson correlation coefficient.

```
proc corr data=mydata.ames_housing_data  nosimple;
   var saleprice;
   with LotFrontage LotArea MasVnrArea BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF
   FirstFlrSF SecondFlrSF LowQualFinSF GrLivArea GarageArea WoodDeckSF OpenPorchSF
   EnclosedPorch ThreeSsnPorch ScreenPorch PoolArea MiscVal
```

I will now compare following two models.
Model1 - SalePrice = β_0 + β_1GrLivArea + ε
Model2 – SalePrice = β_0 + β_1TotalBsmtSF + ε

```
proc reg data=ames_samp;
  model SalePrice = GrLivArea;
  model SalePrice = TotalBsmtSF;

run; quit;
```

This yields following results.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	7948.02012	3480.42036	2.28	0.0225
GrLivArea	1	116.88332	2.16146	54.08	<.0001

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	57750	3460.83152	16.69	<.0001
TotalBsmtSF	1	120.21509	3.00583	39.99	<.0001

Simple comparison of the models.

Model	Adj-Rsq	F Value
SalePrice=7948+116.9*GrLivArea	0.55	2924.22
SalePrice=57750+120.2*TotalBsmtSF	0.4006	1599.52

If we solely look at these two variables then GrLivArea would be the best to explain variability in SalePrices. We can also look at the F Value to see that the model best fit the population from which the data were sampled.

3. Multiple regression model

This time I will use both variables GrLivArea and TotalBsmtSF in a new model to see if multiple regression model will better explain variation in sale price. The result yields

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-26269	3277.36826	-8.02	<.0001	0
GrLivArea	1	90.05960	2.12033	42.47	<.0001	1.26729
TotalBsmtSF	1	70.37780	2.55490	27.55	<.0001	1.26729

Model3 – SalePrice = 90.1*GrLivArea + 70.4* TotalBsmtSF - 26269

I have also added the Variance Inflation Factor (VIF), a multicollinearity diagnostic, to the model output. VIF values greater than 3 indicate multicollinearity. We can conclude that there is no multicollinearity in this case.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1.066281E13	5.331403E12	2304.90	<.0001
Error	2390	5.528247E12	2313074195		
Corrected Total	2392	1.619105E13			

From the Anova table above we can see that the the F value is statistically significant at the p < 0.01 level, indicating that model has predictive power in explaining the variability in SalePrice.

We can now compare all three models to see if which performs the best. We can see from the p value that we have statistically significant predictor variables. We can also see that the multiple linear regression is able to explain almost 66% of variation in sale prices with adj-rsq value 0.6583. In this case using adj-rsq value is safer since it only increases if adding a variable makes improvement in explaining the predicted value whereas r-square will always increase when adding new variable.

Model	R-square	Adj-Rsq	F Value	Pr > F
SalePrice=7948+116.9*GrLivArea	0.5502	0.55	2924.22	<.0001
SalePrice=57750+120.2*TotalBsmtSF	0.4008	0.4006	1599.52	<.0001
SalePrice = 90.1*GrLivArea + 70.4* TotalBsmtSF - 26269	0.6586	0.6583	2304.9	<.0001

4. Outliers

By using proc univariate I continue analysis of outliers. We could earlier observe some potential outlier values by studying output from QQ plot and Cook's D. The output from proc univariate yields.

Quantiles (Definition 5)	
Level	**Quantile**
100% Max	755000
99%	470000
95%	345474
90%	290000
75% Q3	220000
50% Median	165325
25% Q1	131500
10%	110000
5%	94000
1%	64000
0% Min	12789

There is quite big jump between 1 and 0, and 99 and 100 percentile. I will exclude those value from the data set.

```
data outliers;

  set ames_samp;

  keep SalePrice price_outlier GrLivArea TotalBsmtSF;

  if SalePrice <= 64000 then price_outlier = 1;

    else if SalePrice > 64000 & SalePrice < 470000 then price_outlier = 2;
    else if SalePrice >= 470000 then price_outlier = 3;

    run; quit;
```

I will then define new data set base on above criterias.

```
data pruned;

  set outliers;

  if price_outlier = 1 then delete;

    else if price_outlier = 3 then delete;
```

5. Refit multiple regression model without outliers

Root MSE	44416	**R-Square**	0.6244
Dependent Mean	183545	**Adj R-Sq**	0.6241
Coeff Var	24.19914		

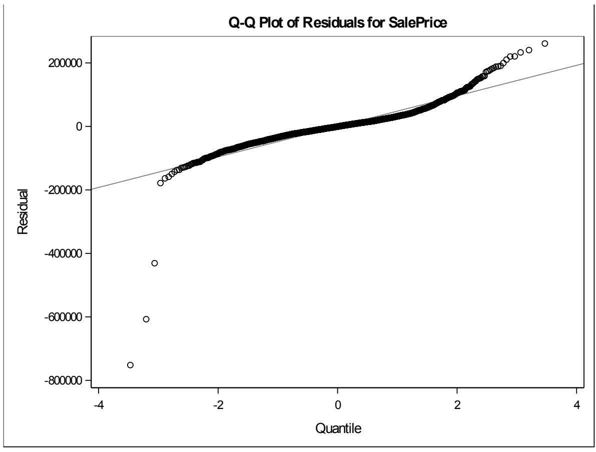
Comparing the models so far shows that model without outliers performs slightly worse than the model that contains outliers. We cannot do direct comparison of the Adj R-sqr and F-value between the two data sets of models, as these calculations are within the context of the model fit to the data set.

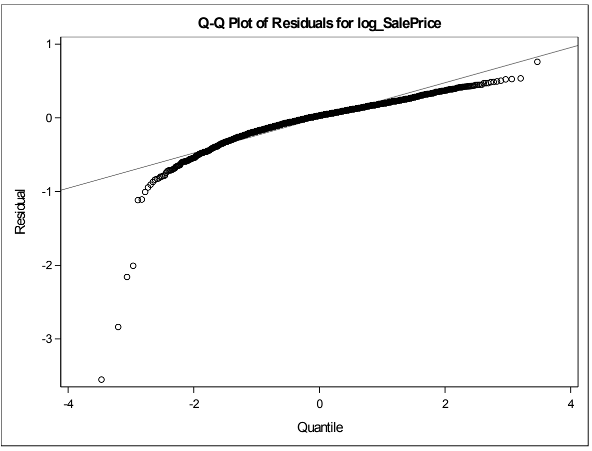
If the decision to exclude the outliers was correct and based on the domain knowledge then the model was adequate to the phenomena we wanted to measure. Data set with outliers could then measure entirely different phenomena.

Model	R-square	Adj-Rsq	F Value	Pr > F
SalePrice=7948+116.9*GrLivArea	0.5502	0.55	2924.22	<.0001
SalePrice=57750+120.2*TotalBsmtSF	0.4008	0.4006	1599.52	<.0001
SalePrice = 90.1*GrLivArea + 70.4* TotalBsmtSF - 26269	0.6586	0.6583	2304.9	<.0001
(Without outliers)SalePrice = 90.1*GrLivArea + 70.4* TotalBsmtSF - 26269	0.6244	0.6241	1946.82	<.0001

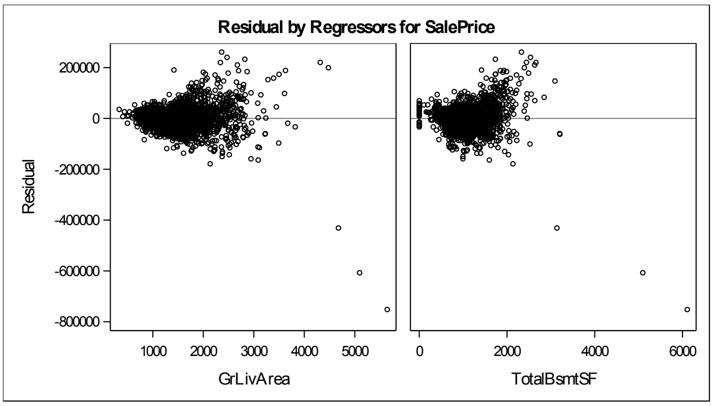
6. Model comparison of Y versus log(Y)

I will now compare two models with SalePrice and log(SalePrice) as predictor variables. First I will compare the distribution of residuals. that seems normally distributed. The log function has straightened out the line on the right side, but since it returns NaN on negative numbers the extreme values on the left are left untouched. Maybe sqrt is more adequate in this case.

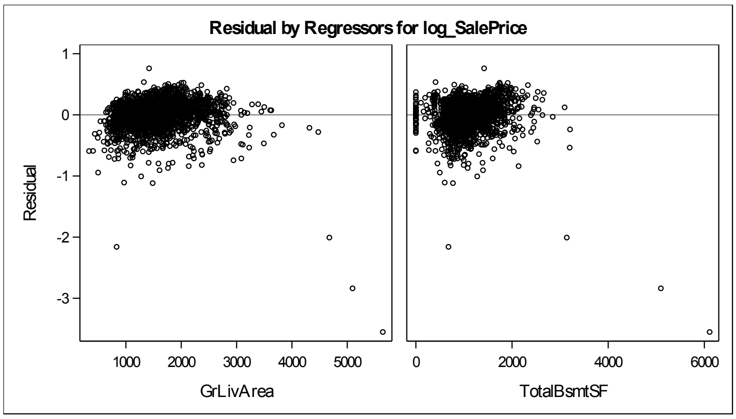




Next I will examine homoscedasticity. Residuals should be randomly distributed without any particular pattern. It seems that in both cases residuals are randomly distributed. So there is linear relation between predictors and predicted variable. If that was not the case we could try to transform the predictor that does not satisfy that condition.



.



Finally the R-squared values. It looks that model performs slightly worse then the model without transformations. Maybe one could try different transformations but results are not discouraging. The model explains almost 65% of variation in the sale price.

Model	R-square	Adj-Rsq	F Value	Pr > F
SalePrice=7948+116.9*GrLivArea	0.5502	0.55	2924.22	<.0001
SalePrice=57750+120.2*TotalBsmtSF	0.4008	0.4006	1599.52	<.0001
SalePrice = 90.1*GrLivArea + 70.4* TotalBsmtSF - 26269	0.6586	0.6583	2304.9	<.0001
(Without outliers)SalePrice = 90.1*GrLivArea + 70.4* TotalBsmtSF - 26269	0.6244	0.6241	1946.82	<.0001
Log(SalePrice)=11.02+0.00044659*GrLivArea +0.00032608*TotalBsmtSF	0.6477	0.6474	2197.41	<.0001