Assignment #2
Andrew Knight

Introduction:

The sections that follow contain the problem description and exploratory data analysis for both simple and multiple regression models designed to predict SalePrice values for homes. These models are based on the Ames Housing Data and build on the analysis performed in Assignment 1. However, the same sample and EDA assumptions of Assignment 1 are not made here. Section 1 defines the new sample set with which the models will be built.

Data File: ames_housing_data.csv

Calculations File: Knight_Assignment2_code.R

Tasks

Section 1: Define the Sample Population & Perform EDA

We will start with the full housing dataset that contains a total 2930 samples each with 82 variables about houses, including most recent SalePrice. The variable SalePrice will be used as the response variable in the models that follow.

<u>Sample Population</u>

Based on the initial analysis from Assignment 1, I will restrict my model to properties of the same type, namely those identified as residential single-family homes. We will not be including agricultural, commercial or industrial properties.

The sample population obviously only includes homes in Ames, IA from the given dataset by the Assessor's Office. This data also only reports homes which had a reported sale in the years from 2006 to 2010 per the data document. The variable GrLivArea is being used as a continuous numeric variable to best represent total square footage above ground for each property.

<u>Drop Conditions</u>

The goal is to focus on residential, single-family properties so the two variables that will be used to qualify types are <u>Zoning</u> and <u>BldgType</u>. Zoning types 'C (all)' and 'I (all)' were dropped, a total of 27 records combined. Next BldgType records which were anything but '1Fam' were dropped, an additional total of 529 records. Also drop GrLivArea > 4500, & TotalBsmtSF outside of range 1 – 3000 square feet. Note the data contain basements with zero SF and those with very large values of SF (more than 3000) are being considered outliers here. The remaining sample has

2347 total observations. The remaining sample dataset is approximately 80% of the original population. This sample will be used in the following sections and is labeled sampledat.

Waterfall of Drop Steps	Records dropped	Total Remaining
1: Zoning != C or I	27	2903
2: BldgType == 1Fam	502	2401
3: GrLivArea > 4500	3	2398
4: TotalBsmtSF < 1 & TotalBsmtSF > 3000	51	**2347**

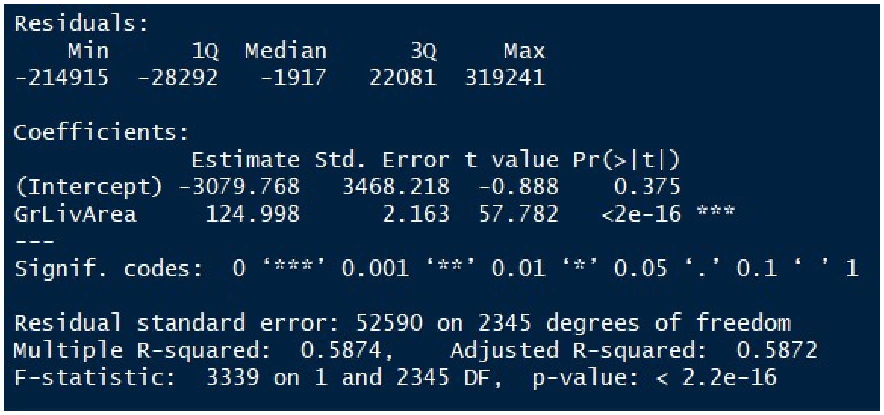
Section 2: Simple Linear Regression Models

The two continuous numeric variables I chose for this section are **GrLivArea** and **TotalBsmtSF**.

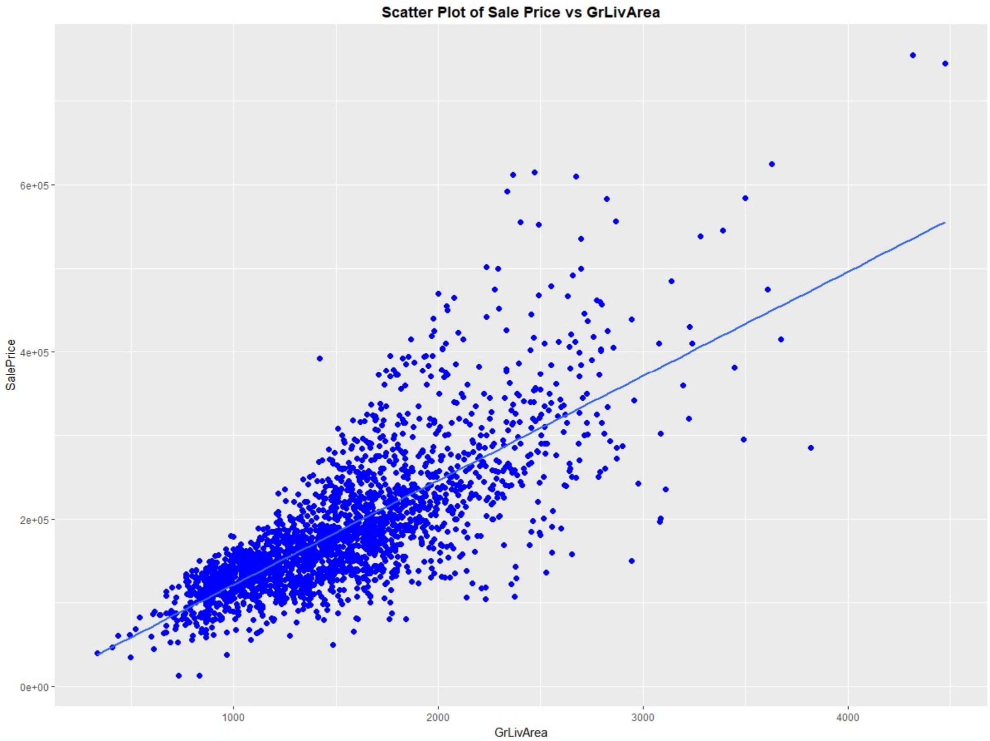
The Residuals & Scatter Plots for each:

GrLivArea

Linear Model Summary for SalePrice ~ GrLivArea:



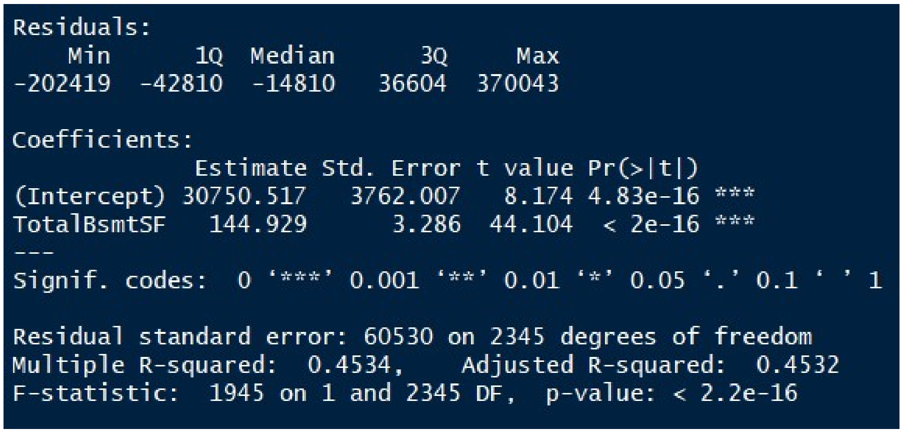
```
Residuals:
    Min      1Q  Median      3Q     Max 
-214915  -28292   -1917   22081  319241 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -3079.768   3468.218  -0.888    0.375    
GrLivArea     124.998      2.163  57.782   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 52590 on 2345 degrees of freedom
Multiple R-squared:  0.5874,	Adjusted R-squared:  0.5872 
F-statistic:  3339 on 1 and 2345 DF,  p-value: < 2.2e-16
```



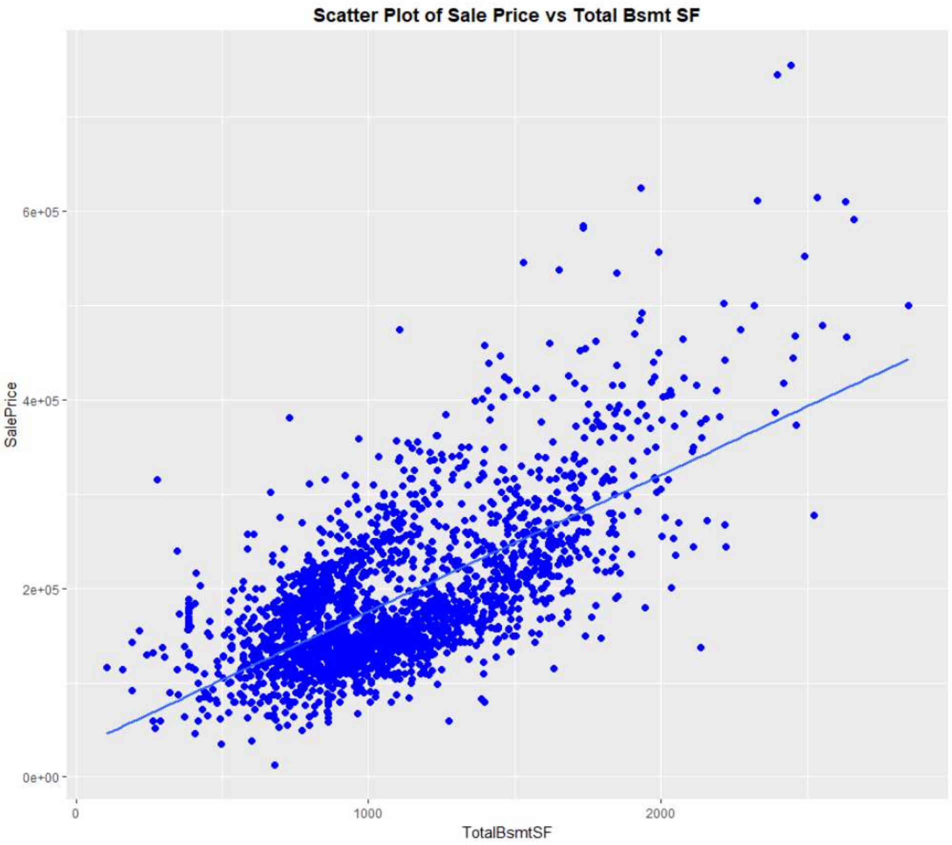
This information shows that for each square foot increment in GrLivArea, the SalePrice goes up by $125 and the p-value verifies statistical significance. The residual standard error is about $52,600 for the single variable model based on GrLivArea and the R squared value is about 0.587 meaning that almost 59% of the variation in the SalePrice is attributed to the GrLivArea.

TotalBsmtSF

Linear Model SUmmary for SalePrice ~ TotalBsmtSF:



```
Residuals:
    Min      1Q  Median      3Q     Max
-202419  -42810  -14810   36604  370043

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 30750.517   3762.007   8.174 4.83e-16 ***
TotalBsmtSF   144.929      3.286  44.104  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 60530 on 2345 degrees of freedom
Multiple R-squared:  0.4534,    Adjusted R-squared:  0.4532
F-statistic:  1945 on 1 and 2345 DF,  p-value: < 2.2e-16
```

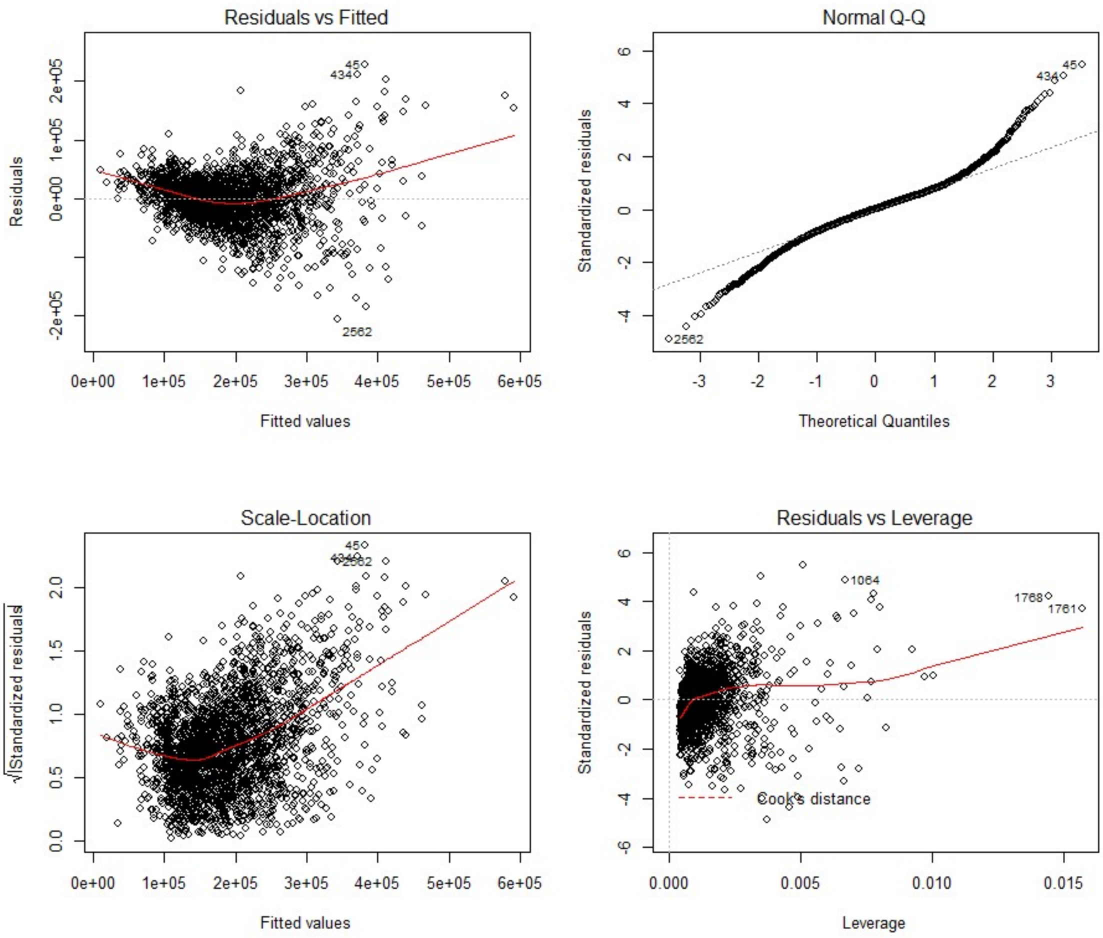


This information shows that for each square foot increment in TotalBsmtSF, the SalePrice goes up by $145 and the p-value verifies statistical significance. The residual standard error is about $60,500 for the single variable model based on TotalBsmtSF and the R squared value is about 0.453 meaning that 45% of the variation in the SalePrice is attributed to the TotalBsmtSF.

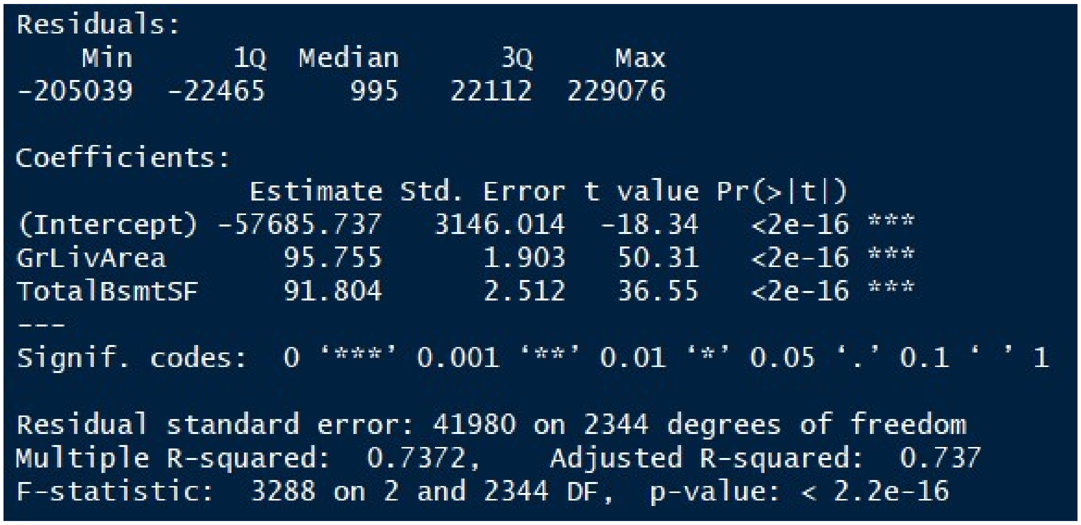
Section 3: Multiple Linear Regression Models

The multiple linear regression model results for the two chosen variables GrLivArea and TotalBsmtSF is given below. This was run on the same sampledat data as the single linear regression tests above.

MLR Output for SalePrice ~ GrLivArea + TotalBsmtSF:



The ANOVA test p-values show that both variables are highly significant and the model summary output is given here.



```
Residuals:
    Min      1Q  Median      3Q     Max 
-205039  -22465     995   22112  229076 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -57685.737   3146.014  -18.34   <2e-16 ***
GrLivArea       95.755      1.903   50.31   <2e-16 ***
TotalBsmtSF     91.804      2.512   36.55   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 41980 on 2344 degrees of freedom
Multiple R-squared:  0.7372,    Adjusted R-squared:  0.737 
F-statistic:  3288 on 2 and 2344 DF,  p-value: < 2.2e-16
```

Both the GrLiveArea and the TotalBsmtSF contributed to the SalePrice with about a $96 and $92 increase for each square foot added respectively. The Residual Standard Error gives almost $42,000 and the R squared shows an improved fit using two variables of almost 74%.

Section 4: Regression Models for the Transformed Response Using Log Transformation

Next, we will perform a Log transformation of SalePrice and run Simple Linear Regression and Multiple Linear Regression Models again.

SLR Comparison:

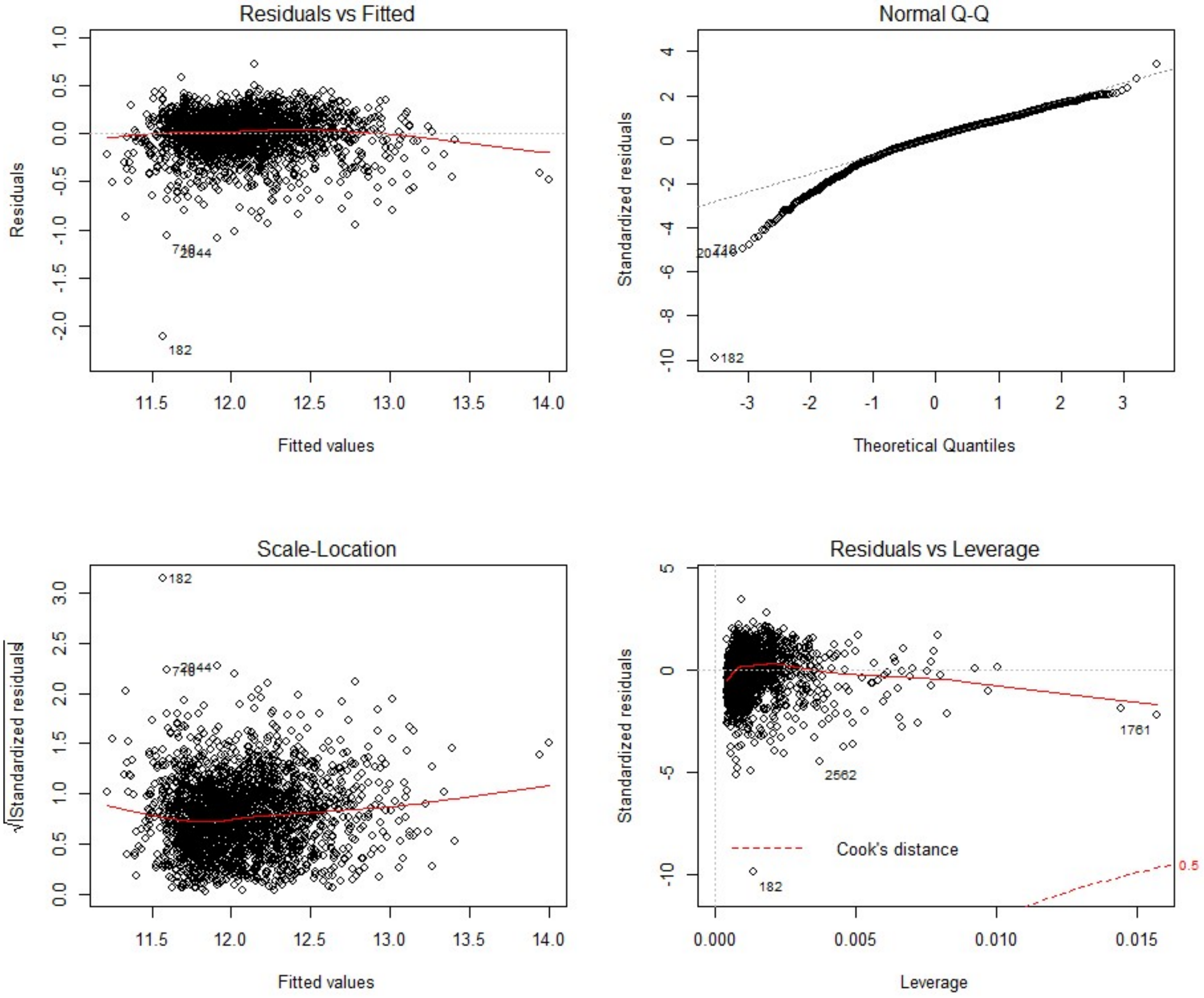
	GrLivArea	**TotalBsmtSF**	
SalePrice	R^2: 0.5874 R^2a: 0.5872 RSE: 52590	R^2: 0.4534 R^2a: 0.4532 RSE: 60530	
Log(SalePrice)	R^2: 0.5836 R^2a: 0.5834 RSE: 0.256	R^2: 0.4172 R^2a: 0.4169 RSE: 0.3029	

It turns out that for the Simple Models using only one predictor variable, the log transformation underperforms the non-transformed model. The R^2 goodness-of-fit values are lower for the log of SalePrice.

Now run Multiple Linear Regression model and compare to non-transformed results to compare results between SalePrice and log(SalePrice). The R-squared and the adjusted R-squared values are given below.

MLR Comparison:

	GrLivArea + TotalBsmtDF	**Reg Coefficients**
SalePrice	R^2: 0.7371 R^2a: 0.7368 RSE: 39,930	GrLivArea: 67.72 TotalBsmtSF: 32.62
Log(SalePrice)	R^2: 0.7117 R^2a: 0.7115 RSE: 0.2131	GrLivArea: 0.00047 TotalBsmtSF: 0.00041



The R-squared values show that the log-transformed model does not fit the data as well as the non-transformed SalePrice and our linear models have not improved as a result. There are afew things we note about both MLR Outputs.

In the residuals plot we should not see any kind of pattern that would indicate non-linearity. In the case of the log(SalePrice) output it is certainly less pronounced than in the non-transformed SalePrice output, which shows a slight parabolic shape. The Q-Q plot for the SalePrice MLR is reasonable but the log(SalePrice) shows quite a bit of deviation on the lower 2nd and 3rd quantiles.

In the Scale-Location plot of the SalePrice MLR output we can see that the residuals are not spread equally between the two predictors whereas in the log(SalePrice) MLR output they do appear evenly spread.

Section 5: Summary & Conclusions

In summary, more work needs to be done in identifying the right model. While the focus has been on numeric metrics about the size of the house for the predictor, we may need to go back and assess the data. Some things that should be re-considered:

- Should other property types or sale types be further restricted?
- Should additional outliers be removed? For example, should we be considering basements with zero SF in the TotalBsmtSF analysis? Should the high-end outliers for GrLivArea be reduced further?
- Should we be considering more predictor variables or choosing different variables?
- Can we perform transformations on the variable data prior to running the model so that assumptions about normality, skewness, etc are more precise?

One reason that our linear model may not be performing as expected is that in using GrLivArea and TotalBsmtSF, we are at risk of violating the 'No Mulicollinearity' assumption. It could be that these two variables are too-highly correlated. Another reason our model is not performing as well is because the sample data do not give a close enough approximation to normality. Other assumptions about the linear variables were not satisfied as well. We actually saw this in some of the initial EDA analysis in Assignment 1. Going back and assessing the three points above should help improve this as well. Using **gvlma** package I get an output asserting that the assumptions were not met. Clearly the model can still be improved.