Assignment #2
Prabhat Thakur

Introduction:

The goal of this assignment is to build simple and multiple regression models designed to predict house SalePrice values, evaluate goodness of fit for each model and compare them. We are also going to build simple and regression models for transformed response log(SalePrice) and compare results of these model. 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.

This report contains following sections:
(1) Defining Sample population
(2) Exploratory Data Analysis (EDA)
(2) Simple Linear Regression Models
(3) Multiple Linear Regression Models
(4) Regression models for the transformed response log(SalePrice)
(5) Conclusion

Data File: ames_housing_data.csv (Ames, Iowa housing data set posted in Canvas).

(1) Section 1: Define the Sample Population/EDA

The dataset contains information from the Ames Assessor's Office used in computing assessed values for individual residential properties sold in Ames, IA from 2006 to 2010. The data contains 2930 observations on various housing attributes and will be using full dataset to begin with. The dataset also includes the most recent SalePrice of housing property which will be the response variable for the all models we build here in.

<u>Sample Population</u>
Choosing right predictor variables which contribute significantly to the house price is very important for building efficient regression mode. We will begin with EDA on housing attributes to identify which attributes are best predictors of sales prices and also to exclude variables and observations from the population which doesn't confirm to majority of observations.

In assignment 1, I droped about 33% of the observations from population which is not a very good approach. This time I will keep the sample population at least 95% of the total observations. Instead of excluding some categories of housing attributes, I will focus on outliers which are way off from other observations.

Waterfall Drop Conditions:

1. SalePrice: There is a huge difference between 75 percentile and max house sale price. 75% of the house sale price are under $213500 and the max sale price is 755000. I have decided to exclude houses with sale price greater than $450000. This will drop 32 onservations.

2. GrLivArea: I also decided to exclude properties with GrLivArea (total square footage above ground) > 3500 sqft. This will drop 5 observations.

3. Zoning : The data contains few industrial and agricultural properties. I am excluding properties of Zoning types ‘C (all)’ , ‘I (all)’ and 'A (agr)'. This will drop 29 observations.

4. TotalBsmtSF : I have decided to excluded properties with TotalBsmtSF (total basement square footage) >3000 sqft. This will drop 3 observations.

Below is a summary of my waterfall process:

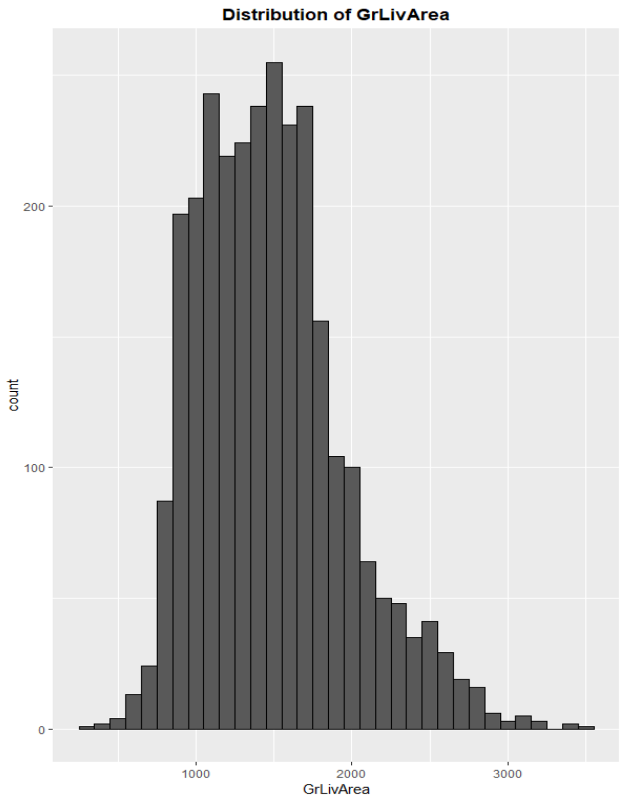
Waterfall of Drop Steps	Records dropped	Total Remaining
0: Original Data Set	0	2930
1: SalePrice > $450000	32	2898
2: GrLivArea > 3500	5	2893
3: Zoning != A or C or I	29	2864
4: TotalBsmtSF > 3000	3	2861

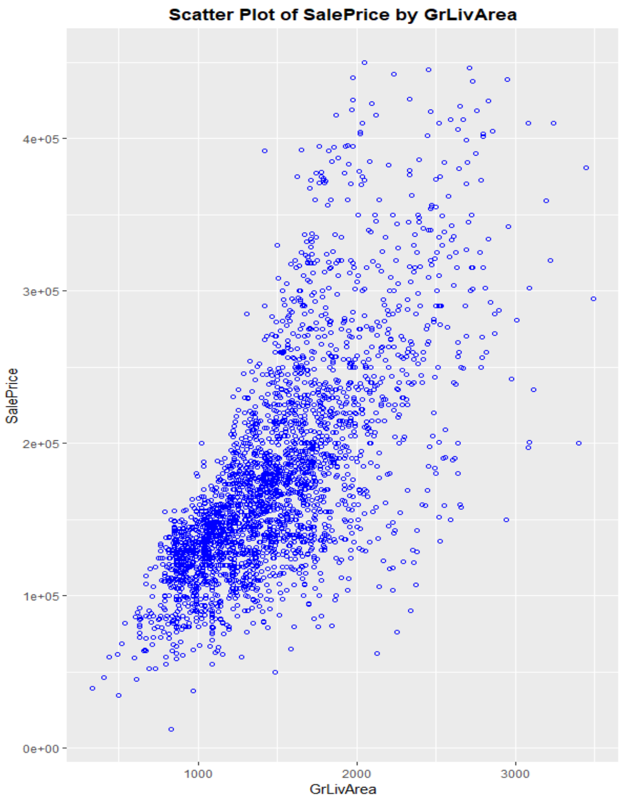
The remaining sample has 2861 total observations. The remaining sample dataset is approximately 97.6% of the original population. This sample will be used in the following sections for model building.

Section 2: Exploratory Data Analysis for predictor variables

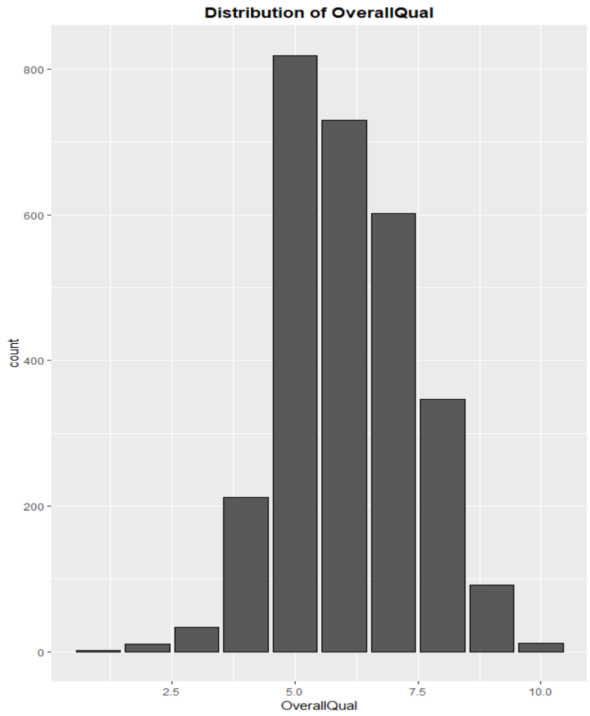
Based on Assignment 1 EDA, I ultimately chose to include the following two predictor variables for model building : GrLivArea, and OverallQual due to their strength of relationship to sales price according to the correlation coefficient.

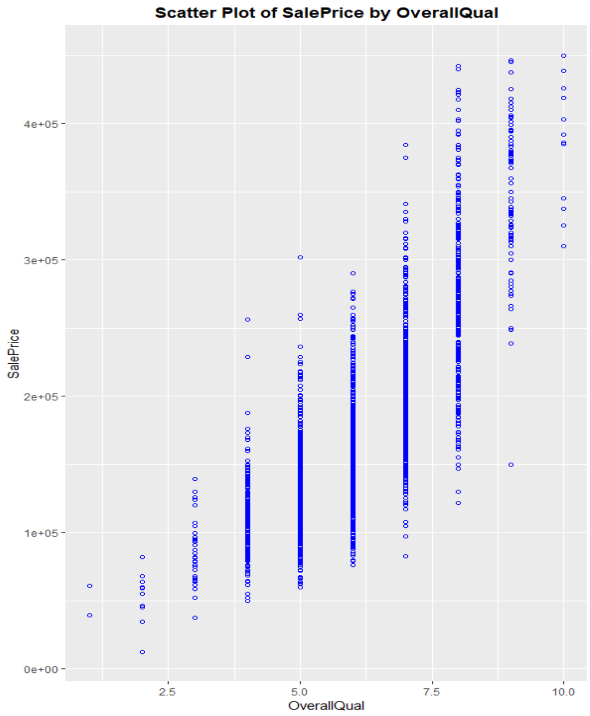
"SalePrice" vs " GrLivArea": We can see a strong positive relationship between Sale price and GrLivArea (Above grade (ground) living area square feet) variable.


Distribution of GrLivArea


Scatter Plot of SalePrice by GrLivArea

"SalePrice" vs " OverallQual ": Most of the houses are in good condition and we can see positive relationship between sales price and overall quality ratings.

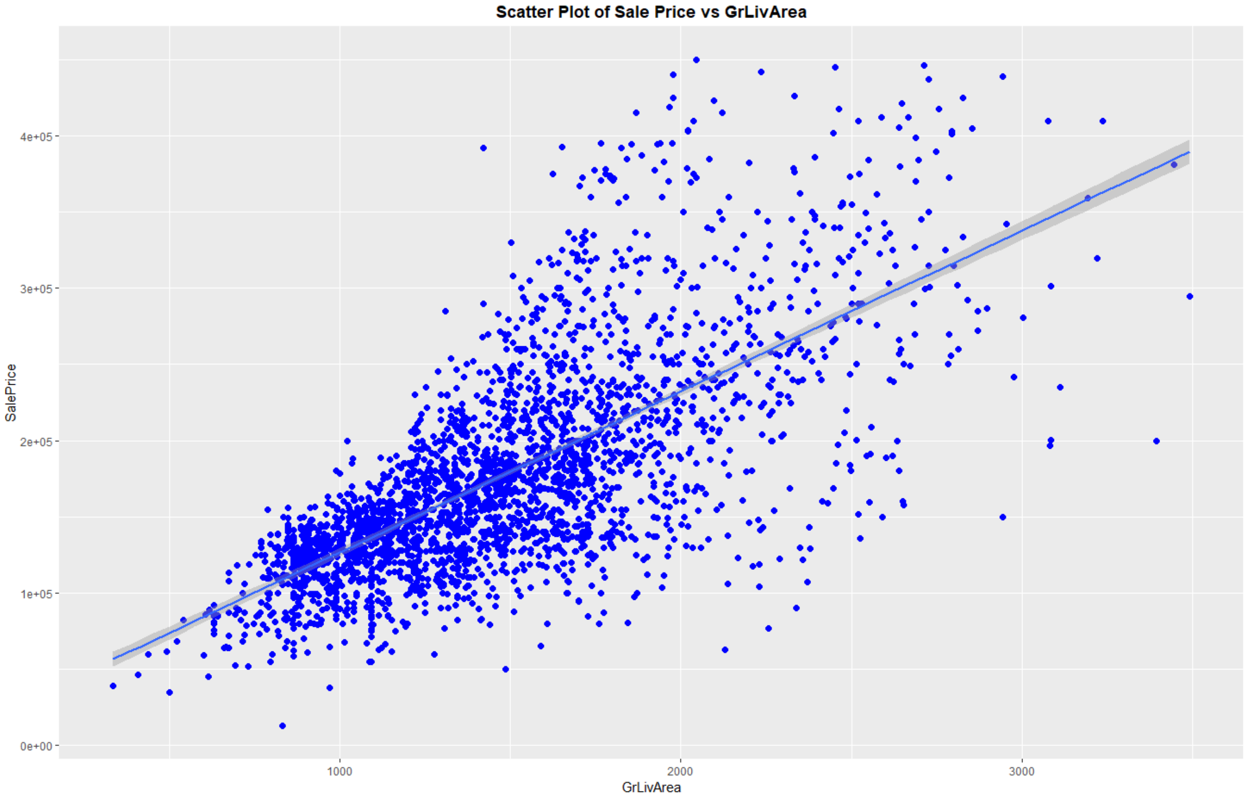

Distribution of OverallQual


Scatter Plot of SalePrice by OverallQual

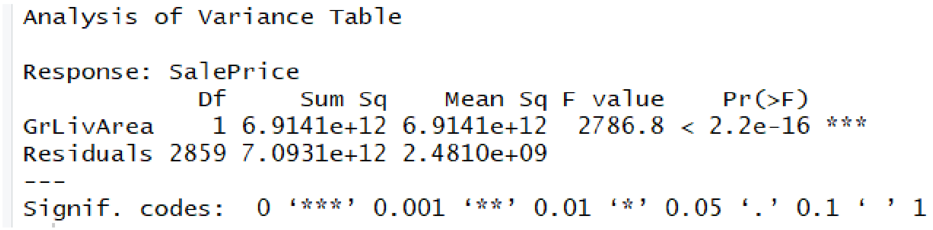
Section 3: Simple Linear Regression Models

In this section we will build two simple linear regression models using the predictor variables selected in above section 2.

- Section 3.1: Model #1 (**GrLivArea**)
  Below is the fitted model between Sale Price and GrLivArea.

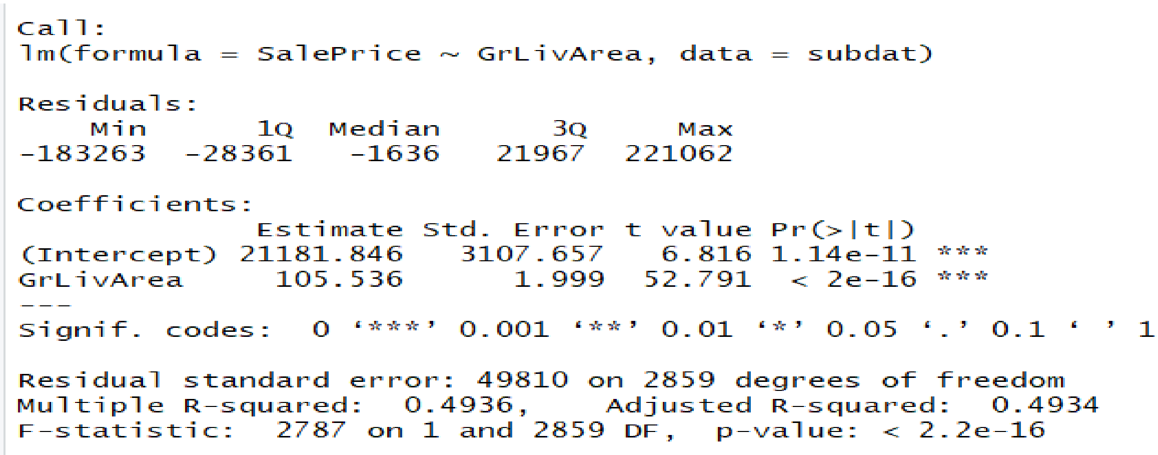


From analysis of Variance table we can see that this model is significant. Which means GrLivArea is indeed a good predictor of Sale Price.



```
Analysis of Variance Table

Response: SalePrice
            Df     Sum Sq    Mean Sq F value    Pr(>F)    
GrLivArea    1 6.9141e+12 6.9141e+12  2786.8 < 2.2e-16 ***
Residuals 2859 7.0931e+12 2.4810e+09                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Let's look at the Linear Model Summary for SalePrice ~ GrLivArea:



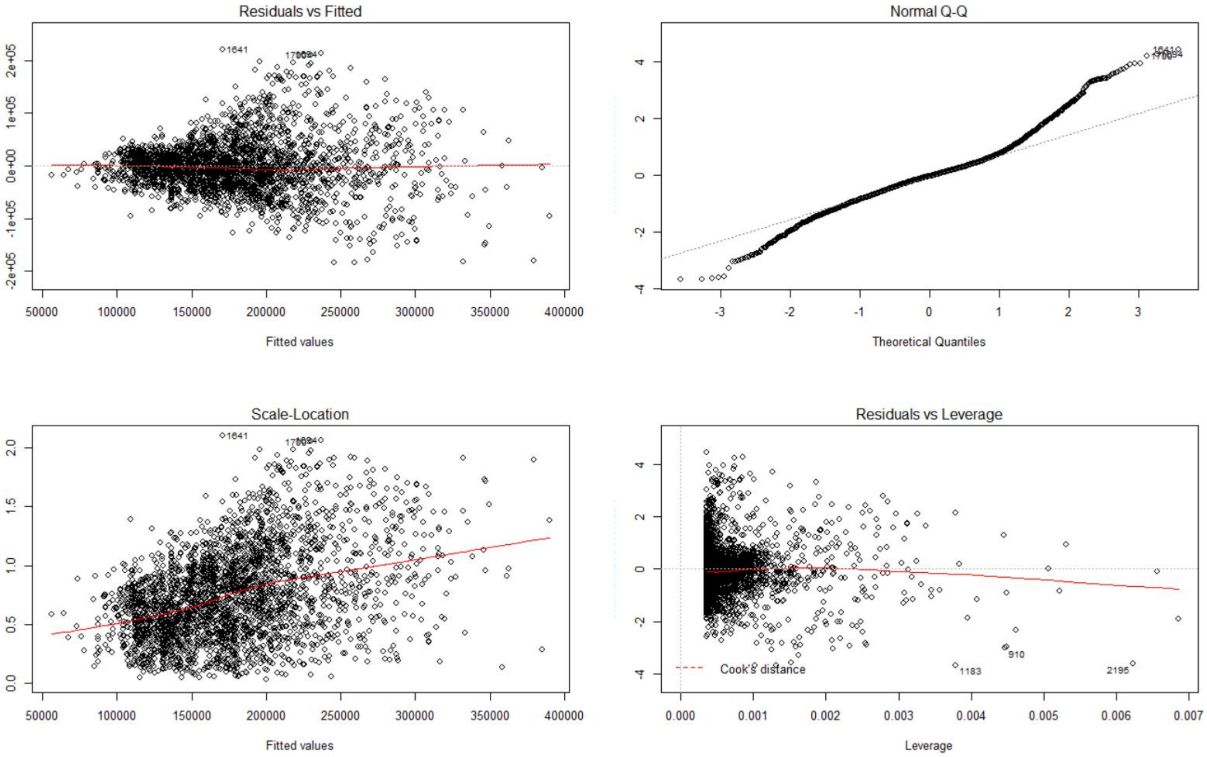
```
Call:
lm(formula = SalePrice ~ GrLivArea, data = subdat)

Residuals:
    Min      1Q  Median      3Q     Max 
-183263  -28361   -1636   21967  221062 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 21181.846   3107.657   6.816 1.14e-11 ***
GrLivArea     105.536      1.999  52.791  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 49810 on 2859 degrees of freedom
Multiple R-squared:  0.4936,	Adjusted R-squared:  0.4934 
F-statistic:  2787 on 1 and 2859 DF,  p-value: < 2.2e-16
```

Simple Linear Regression equation for this model would be:

Sale Price = 21181.8 + 105.5 * GrLiveArea

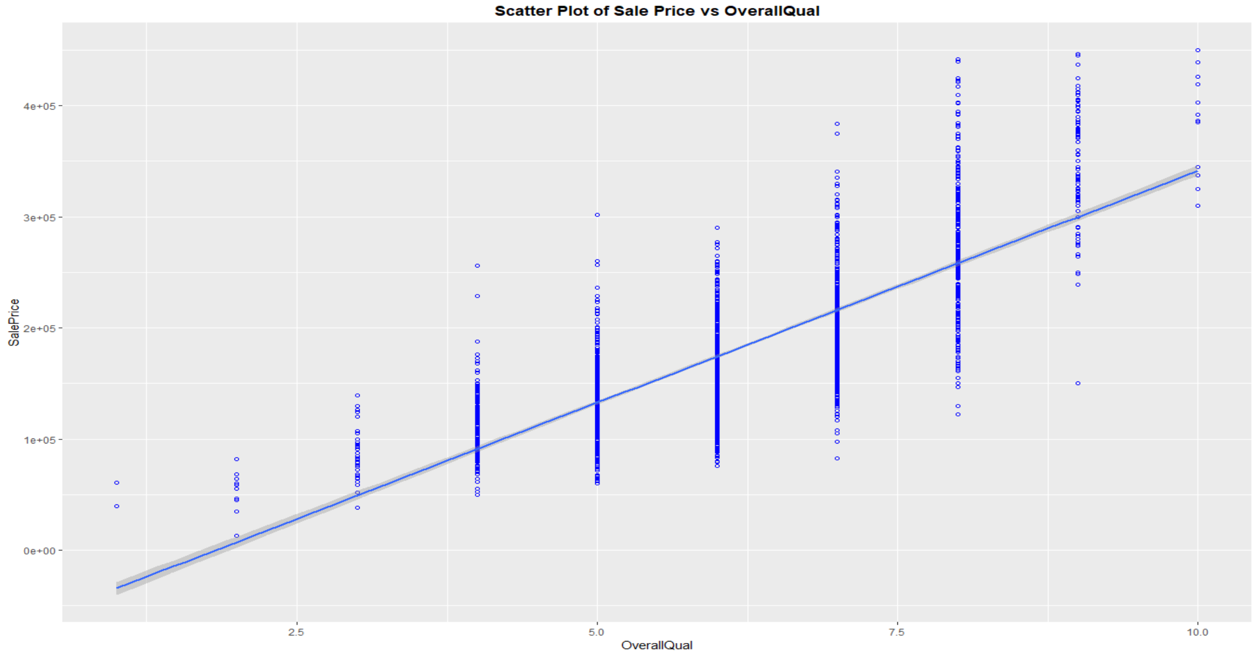
This information shows that for each square foot increment in GrLivArea, the SalePrice goes up by $105.5 and the p-value verifies models statistical significance. The residual standard error is about $49810 for the single variable model based on GrLivArea and the R squared value is about 0.49 meaning that almost 49% of the variation in the Sale Price is explained by this model.

Below plots show results related to the assumptions made for the model.

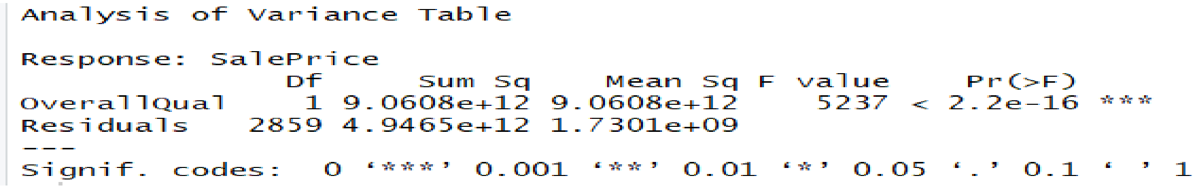


Due to small R- squared value and high residual standard error, we can expect large predictive error.

- Section 3.2: Model #2 (**OverallQual**)
  Below is the fitted model between Sale Price and OverallQual.

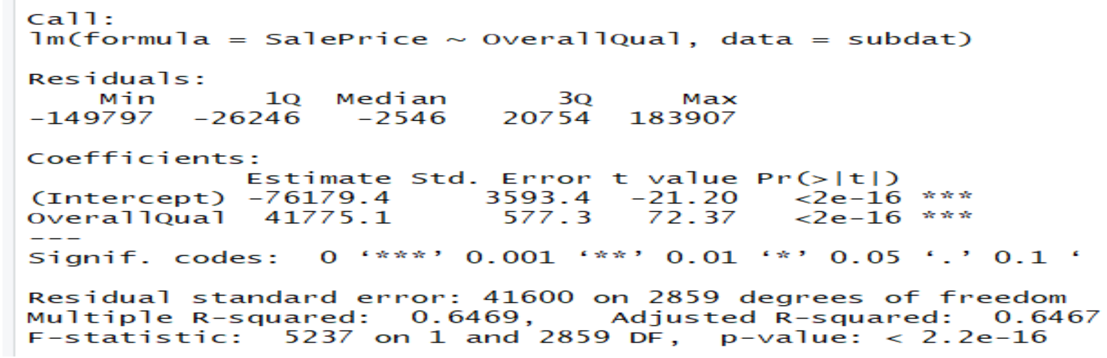


From analysis of Variance table we can see that this model is significant. Which means GrLivArea is indeed a good predictor of Sale Price.



```
Analysis of Variance Table

Response: SalePrice
              Df     Sum Sq    Mean Sq F value    Pr(>F)    
OverallQual    1 9.0608e+12 9.0608e+12    5237 < 2.2e-16 ***
Residuals   2859 4.9465e+12 1.7301e+09                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Now let's look at the Linear Model Summary for SalePrice ~ OVerallQual:



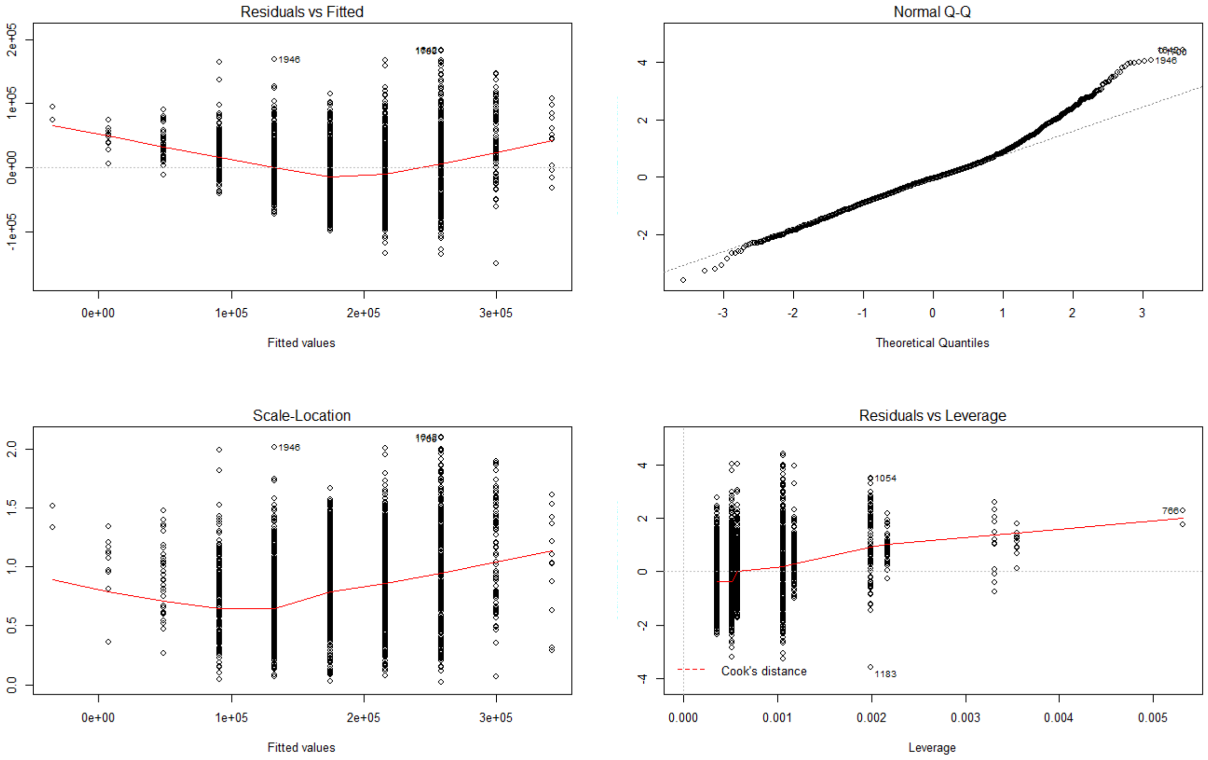
```
Call:
lm(formula = SalePrice ~ OverallQual, data = subdat)

Residuals:
    Min      1Q  Median      3Q     Max 
-149797  -26246   -2546   20754  183907 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -76179.4     3593.4  -21.20   <2e-16 ***
OverallQual  41775.1      577.3   72.37   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 41600 on 2859 degrees of freedom
Multiple R-squared:  0.6469,	Adjusted R-squared:  0.6467 
F-statistic:  5237 on 1 and 2859 DF,  p-value: < 2.2e-16
```

Simple Linear Regression equation for this model would be:

Sale Price = -76179.4 + 41775.1 * OverallQual

This information shows that for each increment in Overall Quality rating , the SalePrice goes up by $41775.1 and the p-value verifies models statistical significance. The residual standard error is about $41600 for the single variable model based on OverallQual and the R squared value is about 0.65 meaning that almost 65% of the variation in the Sale Price is explained by this model.

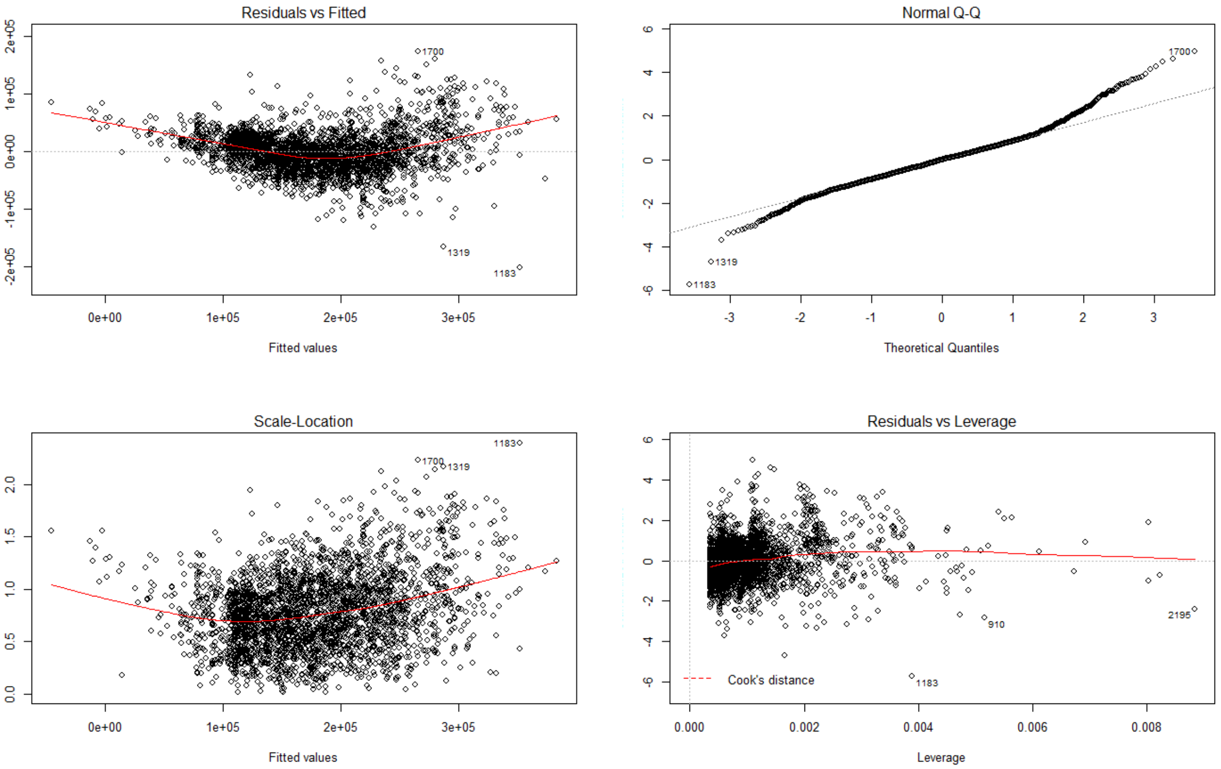
Below plots show results related to the assumptions made for the model.



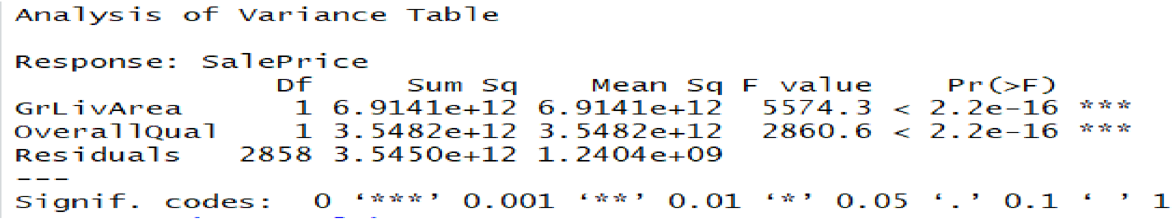
Section 4: Multiple Linear Regression Model – Model #3

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

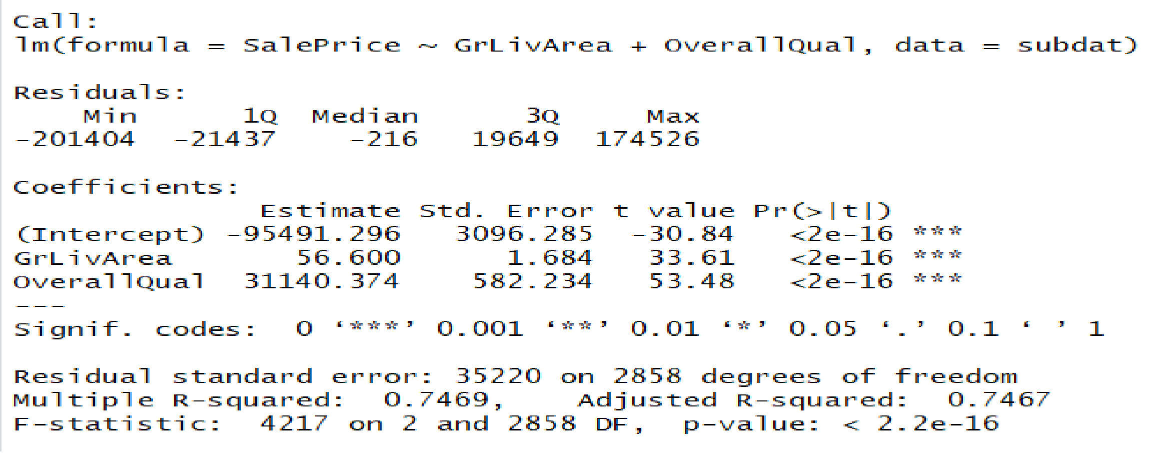
MLR Output for SalePrice ~ GrLivArea + OverallQual:



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


```
Analysis of Variance Table

Response: SalePrice
              Df     Sum Sq    Mean Sq F value    Pr(>F)    
GrLivArea      1 6.9141e+12 6.9141e+12  5574.3 < 2.2e-16 ***
OverallQual    1 3.5482e+12 3.5482e+12  2860.6 < 2.2e-16 ***
Residuals   2858 3.5450e+12 1.2404e+09                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Below is the summary statistics for this Multiple Linear Regression mdel.


```
Call:
lm(formula = SalePrice ~ GrLivArea + OverallQual, data = subdat)

Residuals:
    Min      1Q  Median      3Q     Max 
-201404  -21437    -216   19649  174526 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -95491.296   3096.285  -30.84   <2e-16 ***
GrLivArea       56.600      1.684   33.61   <2e-16 ***
OverallQual  31140.374    582.234   53.48   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 35220 on 2858 degrees of freedom
Multiple R-squared:  0.7469,	Adjusted R-squared:  0.7467 
F-statistic:  4217 on 2 and 2858 DF,  p-value: < 2.2e-16
```

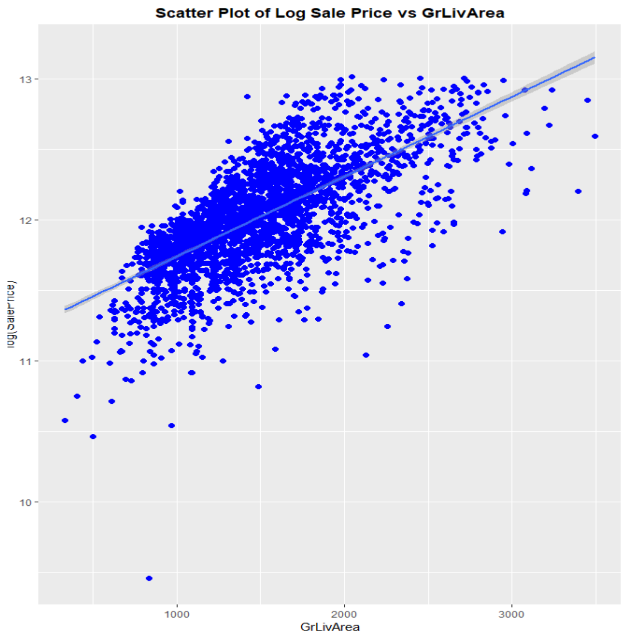
Both the GrLiveArea and the OverallQual contributed to the SalePrice with about a $56 increase for each square foot added and about $31140 for each increment in Overall Quality rating. The Residual Standard Error gives almost $35,220 and the R squared shows an improved fit using two variables of almost 74.7%.

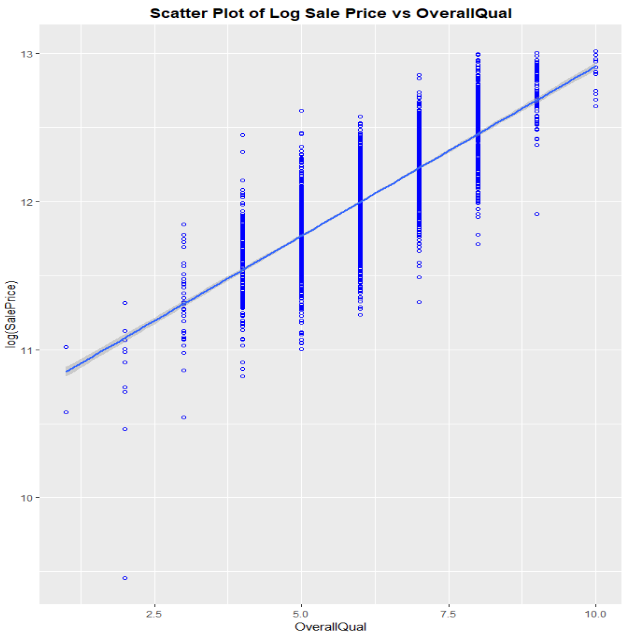
This multiple regression model fits better than the two simple regression model. Adding more predictor variables to a model increases the predictive power of models but there is always a risk of over fitting. The model should not be too simple to not incorporate important predictors and should not be too complex to over fit it. Standard residual error and R-square statistics are good values to compare the models and clarity these statistics are better than earlier simple regression models.

Model1 (SLR) GrLivArea	Midel2 (SLR) OverallQual	Model3 (MLR)
R2: 0.4936 R2a: 0.4934 RSE: 49810	R2: 0.6469 R2a: 0.6467 RSE: 41600	R2: 0.7469 R2a: 0.7467 RSE: 35220

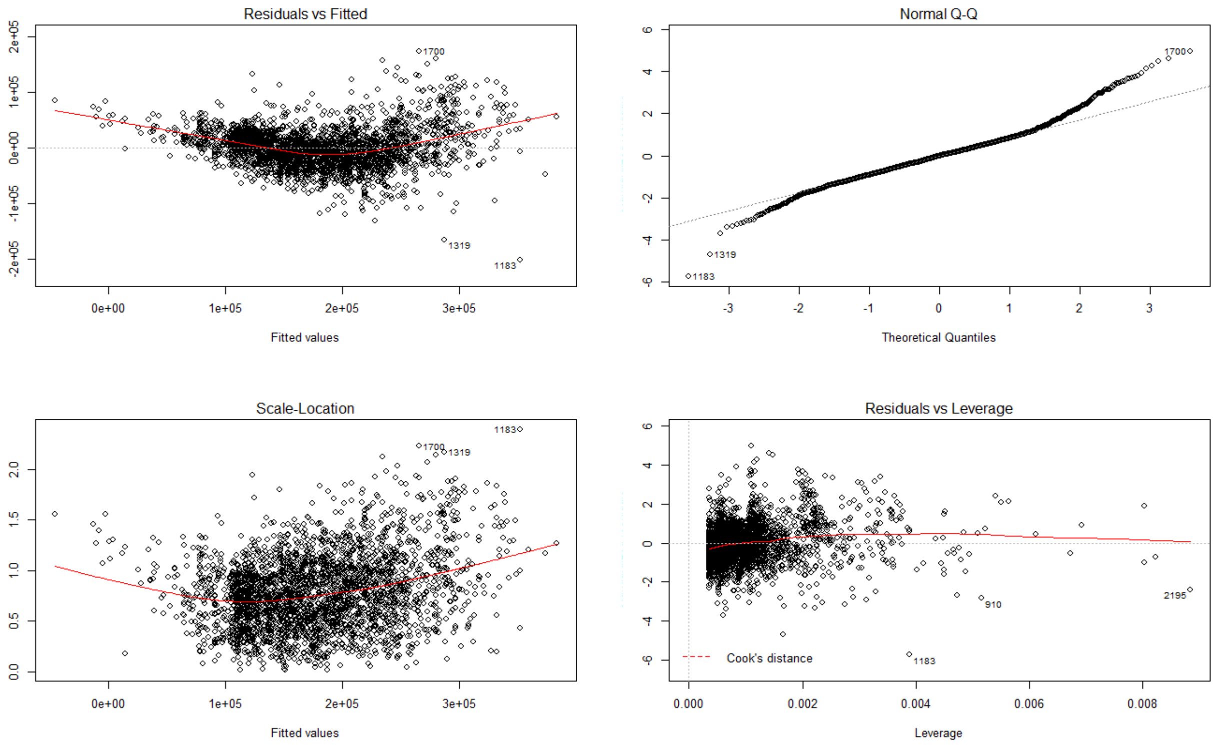
Section 5: Regression models for the transformed response log(SalePrice)

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


Scatter Plot of Log Sale Price vs GrLivArea


Scatter Plot of Log Sale Price vs OverallQual

MLR Output for log(SalePrice) ~ GrLivArea + OverallQual:



Model Comparison:

Response Variable	Model1 (SLR) GrLivArea	Midel2 (SLR) OverallQual	Model3 (MLR)
SalePrice	R2: 0.4936 R2a: 0.4934 RSE: 49810	R2: 0.6469 R2a: 0.6467 RSE: 41600	R2: 0.7469 R2a: 0.7467 RSE: 35220
Log(SalePrice)	R2: 0.4856 R2a: 0.4854 RSE: 0.2721	R2: 0.6638 R2a: 0.6637 RSE: 0.2199	R2: 0.7555 R2a: 0.7553 RSE: 0.1876

It turns out that for the Simple Models using only one predictor variable, the log transformation underperforms the non-transformed model. The R2 goodness-of-fit values are lower for the log of SalePrice. On the other hand when comparing MLR models, 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 a few 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 6: 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 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 OverallQual, OverallQual is a discrete variable with 10 levels which concentrates the fitted values around those 10 points. 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.