Assignment #3

Prabhat Thakur

Introduction:

The goal of this assignment is to build simple and multiple regression models using continuous, categorical, and discrete predictor variables to predict house SalePrice, evaluate goodness of fit for each model, validate, and compare them. We will also build regression models for transformed response log(SalePrice) and compare its results with other models. These models are based on the Ames Housing Data and built based on the EDA analysis performed in Assignment 1 and 2.

I decided to use the same drop conditions as in Assignment #2 to create my sample data for this assignment. This will also help me to compare the models produced in this assignment with the fit obtained from the previous assignment.

Below is the summary of my waterfall drop conditions:

Waterfall Drop Condition 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 which is approximately 97.6% of the original population. This sample data set will be used in the following sections for model building.

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

(1) Section 1: Select a categorical variable and build a Simple liner regression model

For this task I have chosen **KitchenQual** categorical variable which has 5 levels. Below are the details about this variable.

KitchenQual (Ordinal): Kitchen quality

Kitchen Quality	Description	House Count	Mean Sale Price
Ex	Excellent	171	306968.4
Gd	Good	1154	209791.0
TA	Typical/Average	1470	140521.8
Fa	Fair	65	109690.8
Po	Poor	1	107500.0

From above table we can see that mean sale price is increasing as kitchen quality improves. There seems to be a positive correlation between kitchen quality and house SalePrice. One point to note is KitchenQual Po (Poor quality) only has single data point so it is difficult to make assumptions about additional houses with Poor category.

Now, let's build a simple linear regression model using KitchenQual as predictor variable and SalePrice as response variable. Below is the Analysis of Variance table and Regression statistics summary of our model.

```
Analysis of Variance Table

Response: SalePrice
               Df     Sum Sq    Mean Sq F value    Pr(>F)
KitchenQual     4 6.3833e+12 1.5958e+12  597.81 < 2.2e-16 ***
Residuals    2856 7.6239e+12 2.6694e+09
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(SLRresult)

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

Residuals:
    Min      1Q  Median      3Q     Max
-220968  -30622   -4022   25337  234478

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    306968       3951  77.693  < 2e-16 ***
KitchenQualFa -197278       7529 -26.204  < 2e-16 ***
KitchenQualGd  -97177       4234 -22.953  < 2e-16 ***
KitchenQualPo -199468      51818  -3.849 0.000121 ***
KitchenQualTA -166447       4174 -39.872  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 51670 on 2856 degrees of freedom
Multiple R-squared:  0.4557,    Adjusted R-squared:  0.455
F-statistic: 597.8 on 4 and 2856 DF,  p-value: < 2.2e-16
```

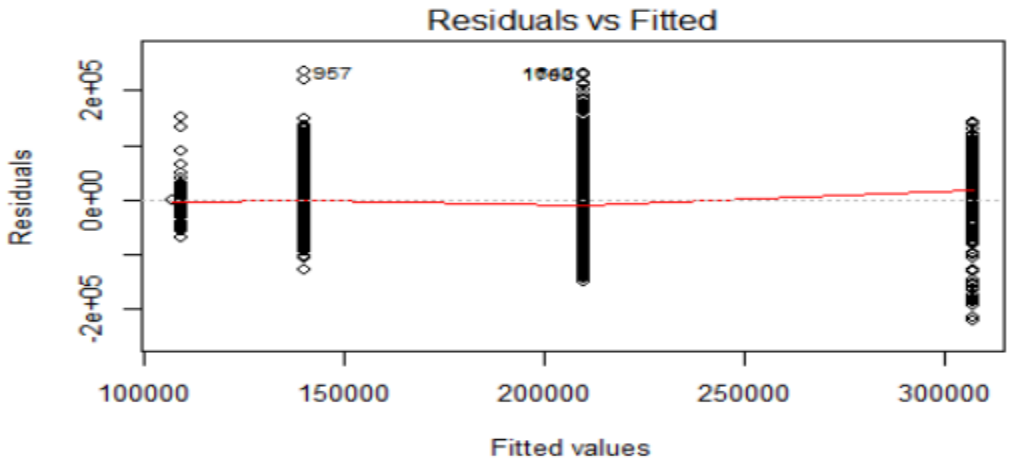
From analysis of variance table, we can see that KithcenQual is significant predictor in this model. From summary statistics, R-squared value is not very impressive though. From the fitted regression model coefficients we can interpret that, the basis group in this model is Kitchen quality EX and the estimated sale price of an excellent kitchen quality (EX) house is $306968 which is exactly same as mean sale price for excellent kitchen quality (EX) houses in the sample data.

Below is the estimated house sale price from each kitchen quality group from the fitted SLR model coefficients.

Kitchen Quality	Description	House Count	Mean Sale Price in $	Estimated SalePrice after fitting SLR model in $
Ex	Excellent	171	306968.4	306968
Gd	Good	1154	209791.0	306968-97177 = 209791
TA	Typical/Average	1470	140521.8	306968-166447 = 140521
Fa	Fair	65	109690.8	306968-197278 = 109690
Po	Poor	1	107500.0	306968-199468 = 107500

From above table we can confirm that predicted model go through the mean of Y (sale price) in each kitchen quality category. **One important conclusion we can make here is that, when a Simple Linear Regression model is fitted using categorical predictor variable, estimated response value for each category are same as mean value of response variable in each category in the sample.**

The mean of the residuals versus fitted values does appear to mostly go through the means of each category, there is slight deviation which seems to be acceptable.



Section 2: Create dummy variables for this categorical variable

For this task, I will use the same categorical variable as task 1 "KitchenQual" which has 5 levels. I have decided to keep KitchenQual EX as basic category.

Let's define our dummy (indicator) variable for KitchenQual categorical variable. Since there are 5 categories, we will need 4 dummy variable.

Kitchen Quality Value	KQ_GD	KQ_TA	KQ_FA	KQ_PO
Gd	1	0	0	0
TA	0	1	0	0
Fa	0	0	1	0
Po	0	0	0	1

When a house has Good Kitchen Quality, KQ_GD = 1; 0 otherwise
When a house has Typical Kitchen Quality, KQ_TA = 1; 0 otherwise
When a house has Fair Kitchen Quality, KQ_FA = 1; 0 otherwise
When a house has Poor Kitchen Quality, KQ_PO = 1; 0 otherwise
It is obvious that when all four dummy variable values are 0 then house has an excellent kitchen quality which is our basis category in the model.

Now let's fit a multiple regression model using above dummy coded variables. Below are the Regression summary statistics and Analysis of variance (ANOVA) table.

```
Call:
lm(formula = SalePrice ~ KQ_GD + KQ_TA + KQ_FA + KQ_PO, data = subdat)

Residuals:
    Min      1Q  Median      3Q     Max
-220968  -30622   -4022   25337  234478

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   306968       3951  77.693  < 2e-16 ***
KQ_GD         -97177       4234 -22.953  < 2e-16 ***
KQ_TA        -166447       4174 -39.872  < 2e-16 ***
KQ_FA        -197278       7529 -26.204  < 2e-16 ***
KQ_PO        -199468      51818  -3.849 0.000121 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 51670 on 2856 degrees of freedom
Multiple R-squared:  0.4557,    Adjusted R-squared:  0.455
F-statistic: 597.8 on 4 and 2856 DF,  p-value: < 2.2e-16

> anova(MLRKQ_result)
Analysis of Variance Table

Response: SalePrice
            Df     Sum Sq    Mean Sq F value    Pr(>F)
KQ_GD        1 1.9921e+12 1.9921e+12 746.246 < 2.2e-16 ***
KQ_TA        1 2.5373e+12 2.5373e+12 950.509 < 2.2e-16 ***
KQ_FA        1 1.8144e+12 1.8144e+12 679.685 < 2.2e-16 ***
KQ_PO        1 3.9556e+10 3.9556e+10  14.818  0.000121 ***
Residuals 2856 7.6239e+12 2.6694e+09
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We can see that model coefficients, R-square and other values are exactly same as Simple linear regression model from task 1. Also we can see that basis category and other 4 categories are significant. It seems we may not have to create indicator variable explicitly to build the model but we can do that if we want to control how different categories influence the model output.

From the fitted regression model coefficients we can interpret that, the basis group in this model is Kitchen quality EX and the estimated sale price of an excellent kitchen quality (EX) house is $306968 which is exactly same as mean sale price for excellent kitchen quality (EX) houses in the sample data. Since all coefficients are same as task 1 Simple linear regression model and we already concluded that the predicted response values go through the mean of Y (sale price) in each kitchen quality category, we can conclude that this holds true for task 2 MLR model as well.

Section 3: Report on the hypothesis tests for each of the betas

Following linear equation could be stated for the model in task 1 and 2.
Y = β0 + β1 * KQ_GD + β2 * KQ_TA + β3 * KQ_FA + β4 * KQ_PO

Using the model coefficients, the response estimate can be calculated using coefficient estimates as below:
Yhat = 306968 - 97177 * KQ_GD - 166447 * KQ_TA - 197278 * KQ_FA - 199468 * KQ_PO

Below is the results summary for quick reference:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    306968       3951  77.693  < 2e-16 ***
KQ_GD          -97177       4234 -22.953  < 2e-16 ***
KQ_TA         -166447       4174 -39.872  < 2e-16 ***
KQ_FA         -197278       7529 -26.204  < 2e-16 ***
KQ_PO         -199468      51818  -3.849 0.000121 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 51670 on 2856 degrees of freedom
Multiple R-squared:  0.4557,    Adjusted R-squared:  0.455
F-statistic: 597.8 on 4 and 2856 DF,  p-value: < 2.2e-16
```

Hypotheses Test1:
Null hypothesis H0: βj = 0 against an alternate hypothesis H1: βj != 0 where j=1,2..5.

From the above regression summary statistics, we can see that all five regression coefficients are Significant which is supported by significant t-test values. We can safely reject the null hypothesis and conclude that the all 5 coefficients are significant.

Hypotheses Test2:
Null hypothesis H0: β1 = β2 = β3 = β4 = β5 against an alternate hypothesis H1: β1 != β2 != β3 != β4 != β5

From the ANOVA test and regression summary statistics we can see that the F-statistics is significant which means that model is significant and fits the population. Also the β coefficients are different which means predicted sale price for each category is not same as mean sale price of entire population. We can safely reject the null hypothesis to conclude that all coefficients are significantly different from each other.

Section 4: Multiple Linear Regression Model:

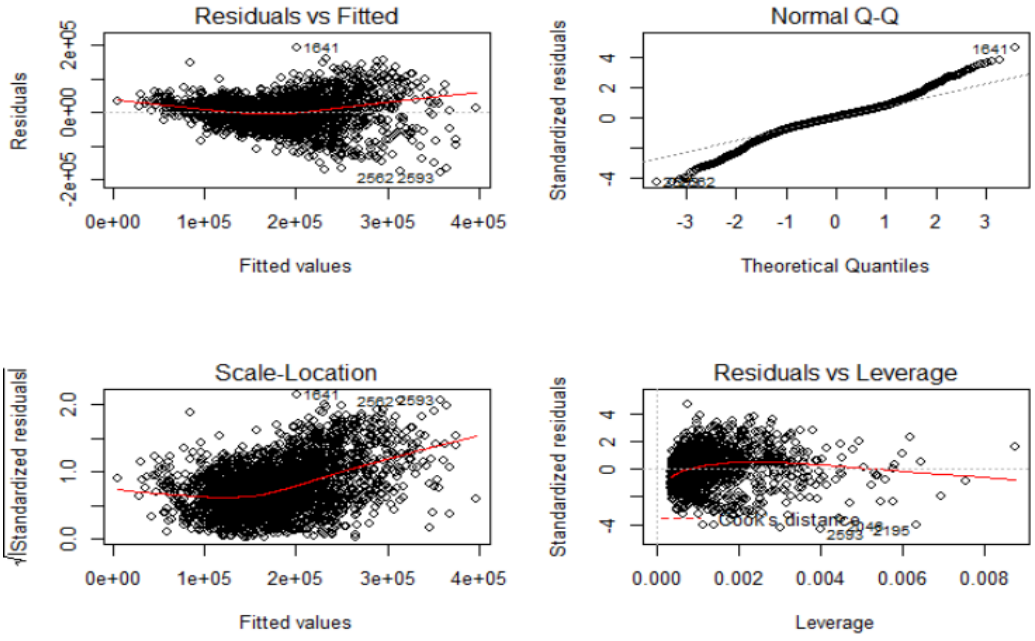
For this task, I have chosen following two continuous variables GrLivArea and TotalBsmtSF. In previous assignments I have found these variables to be positively correlated with house sale price hence using them to build the multiple linear regression model. Same dataset has been used as in above steps.

Model1: SalePrice ~ GrLivArea + TotalBsmtSF

The output of the MLR Model 1 is as below:

```
> MLRresult = lm(SalePrice ~ GrLivArea+TotalBsmtSF, data=subdat)
> anova(MLRresult)
Analysis of Variance Table

Response: SalePrice
              Df     Sum Sq     Mean Sq F value    Pr(>F)
GrLivArea      1 6.9141e+12 6.9141e+12  4066.2 < 2.2e-16 ***
TotalBsmtSF    1 2.2334e+12 2.2334e+12  1313.5 < 2.2e-16 ***
Residuals   2858 4.8597e+12 1.7004e+09
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(MLRresult)

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

Residuals:
    Min      1Q  Median      3Q     Max
-177289  -20659     833   21622  191054

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -21628.627   2830.960   -7.64 2.94e-14 ***
GrLivArea       81.853      1.779   46.00  < 2e-16 ***
TotalBsmtSF     75.001      2.069   36.24  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 41240 on 2858 degrees of freedom
Multiple R-squared:  0.6531,    Adjusted R-squared:  0.6528
F-statistic:  2690 on 2 and 2858 DF,  p-value: < 2.2e-16
```

Diagnostic plots for Model 1 using residuals.



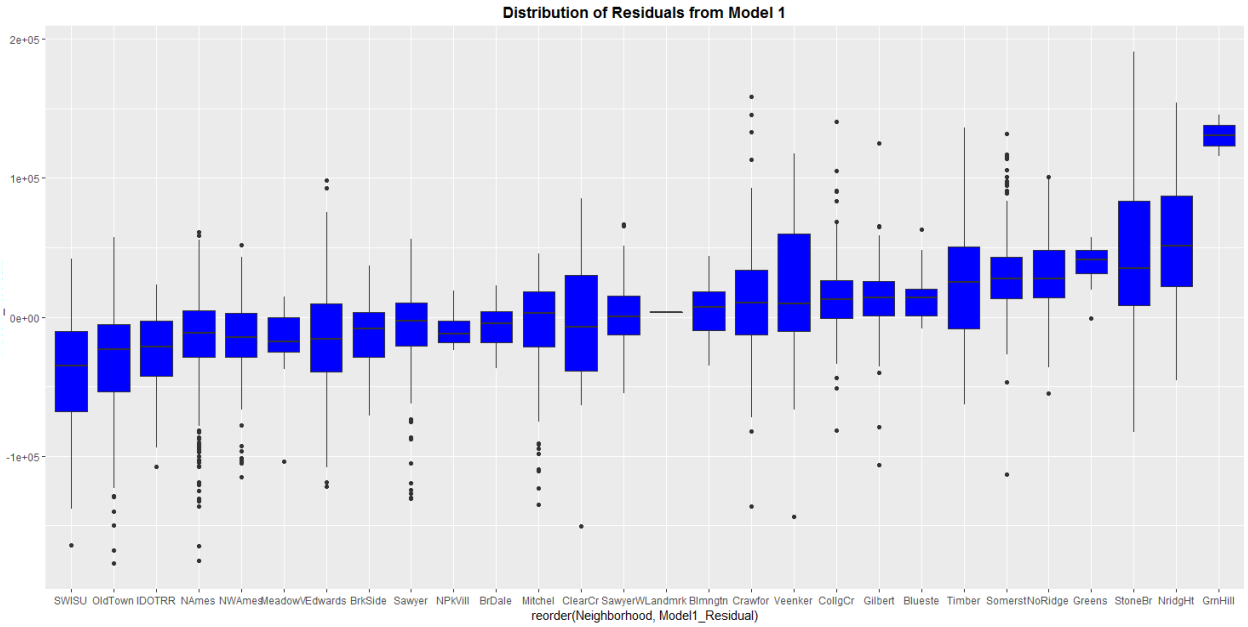
In residuals vs Fitted values scatter plot, under the standard assumptions, the standardized residuals are uncorrelated with the fitted values therefore we expect a random pattern but it seems like a funnel shape distribution. Also the red line seems to be quadratic which suggest that our model violates the linearity and normality assumptions for residuals.
From the QQ plot as well, we can see that residuals are deviating from the mean line which again shows that our model is violating the normality assumption. Points in bottom left corner suggest that there are some outliers as well.
From Scale-Location plot as well can see that the red line going upwards suggesting constant residual variance assumption violated.
Residual vs Leverage plot gives more insight into outliers and leverage data points in the sample dataset. Data points beyond the accepted limit of cooks distance can be treated as outliers. From the plot we can see few outliers in the dataset.

In summary, the model doesn't meet the standard assumptions completely. More/different predictors and transformations of variables should be tried to improve the models goodness of fit.

Comparing with simple linear regression in task 1, this model performs little better. While adding additional predictor variables does not always mean a better fitting model, it can have a positive effect on the fit. 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.

Section 5: Neighborhood Accuracy:

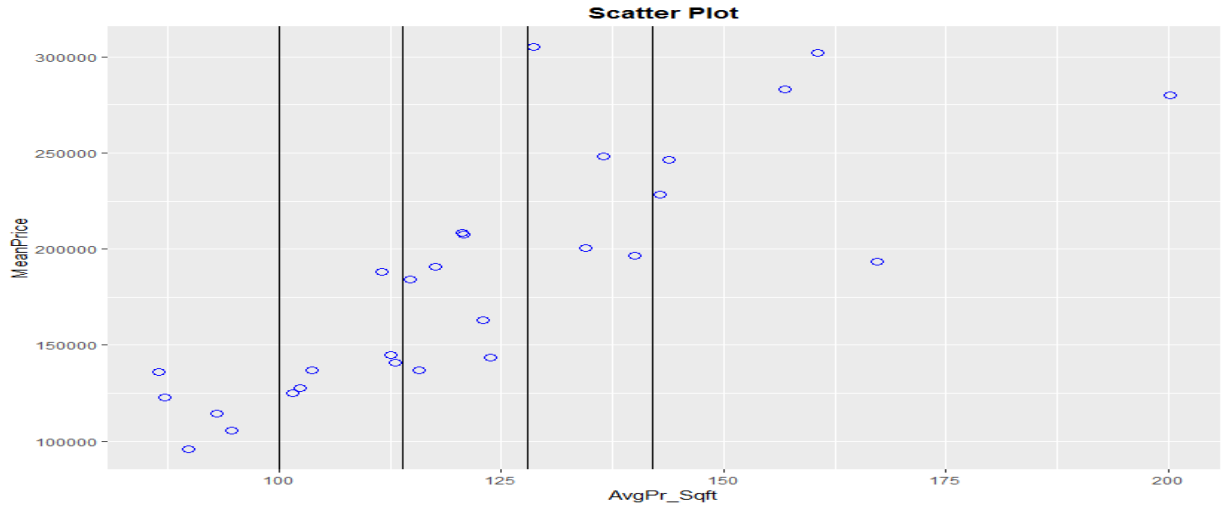
Below is the boxplot of neighborhoods by model 1 residuals shorted by mean neighborhoods residual:



From the plot we can see that neighborhoods has relation to the SalePrice. It shows that Sawyer West neighborhoods is best fitted by this model. This model does over predicted and under predicted few neighborhoods. The boxplots indicate quite a bit of movement in terms of average SalePrice by neighborhood which means that this categorical variable has the potential to be a good predictor of price.

Since the number of houses in each neighborhoods may not be same as there are many neighborhood with similar predictions, we need to group them into reasonable indicator predictors. One approach is to plot Mean Sale Price vs Average SalePrice per sqft for these neighborhood and group from there.

Below is the plot for the same. We can see a positive correlation between these two quantities. We can see that in general as average price per sqft increases neighborhood mean price also increases.



Based on above plot I've grouped neighborhoods into 5 buckets Ngrp1, Ngrp2, Ngrp3, Ngrpb4,and Ngrp5. Average Price per sqft range for each neighborhood group is given in dollars below.
Ngrp1 = $0-100, Ngrp2 = $111 – 114, Ngrp3 = $115 – 128, Ngrp3 = $129 – 142, and Ngrp5 > $143

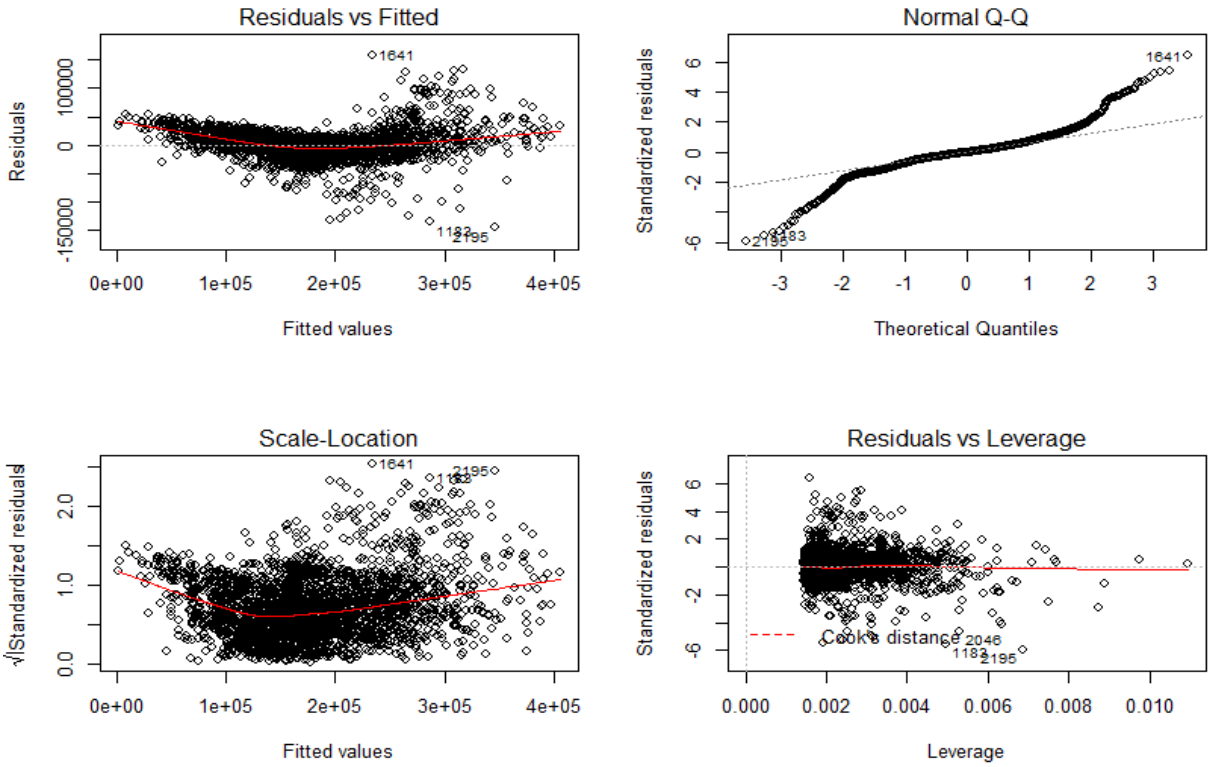
Let's define our indicator variable for neighborhoods groups. Since there are 5 groups, we will need 4 indicator variable.

Neighborhoods groups	Ngrp2	Ngrp3	Ngrp4	Ngrp5
$111 – 114	1	0	0	0
$115 – 128	0	1	0	0
$129 – 142	0	0	1	0
> $143	0	0	0	1

Base category is neighborhoods group from $0-100 average Price per sqft range. Let's add these indicator variables in task 4 multiple regression model. We will call it Model2.
The output of the MLR Model 2 is as below:

```
Call:
lm(formula = SalePrice ~ GrLivArea + TotalBsmtSF + Ngrp2 + Ngrp3 +
    Ngrp4 + Ngrp5, data = subdat)

Residuals:
     Min       1Q   Median       3Q      Max
-145445  -10138    -338   10594  157344

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -64636.188   1884.613  -34.30   <2e-16 ***
GrLivArea      112.173      1.147   97.76   <2e-16 ***
TotalBsmtSF     20.315      1.455   13.96   <2e-16 ***
Ngrp2        37862.863   1459.244   25.95   <2e-16 ***
Ngrp3        56678.993   1447.134   39.17   <2e-16 ***
Ngrp4        71982.179   1568.193   45.90   <2e-16 ***
Ngrp5       111291.566   1563.798   71.17   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24540 on 2854 degrees of freedom
Multiple R-squared:  0.8773,    Adjusted R-squared:  0.877
F-statistic:  3400 on 6 and 2854 DF,  p-value: < 2.2e-16


Analysis of Variance Table

Response: SalePrice
             Df     Sum Sq     Mean Sq    F value Pr(>F)
GrLivArea     1 6.9141e+12 6.9141e+12 11479.7558 <2e-16 ***
TotalBsmtSF   1 2.2334e+12 2.2334e+12  3708.2036 <2e-16 ***
Ngrp2         1 4.3446e+10 4.3446e+10    72.1344 <2e-16 ***
Ngrp3         1 6.5554e+08 6.5554e+08     1.0884 0.2969
Ngrp4         1 4.6213e+10 4.6213e+10    76.7296 <2e-16 ***
Ngrp5         1 3.0505e+12 3.0505e+12  5064.8121 <2e-16 ***
Residuals  2854 1.7189e+12 6.0229e+08
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Diagnostic plots for Model 2:



Looking at Regression statistics summary we can see that R-squared value has increased substantially compared to Model 1 and now it is 0.877, also MAE value for model 1 was $29787 which has reduced to $16163 in Model2, an improvement of $13623. This means that Model 2 prediction accuracy is much better then model 1.

However when we see the residual value for both models, they are pretty much same and also the diagnostic plots for model 2 appears to be same as model 1. That means Model 2 also violates the standard linearity and normality assumptions.

In summary, the model2 has better prediction accuracy but doesn't meet the standard linearity and normality assumptions completely. More/different predictors and transformations of variables should be tried to improve the models goodness of fit.

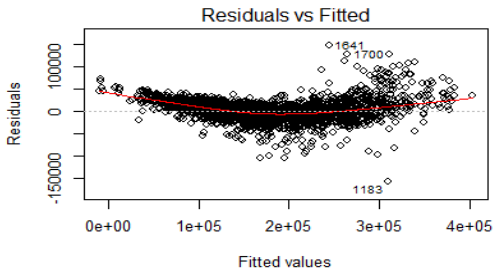
Section 6: Model Comparison of Y versus log(Y):

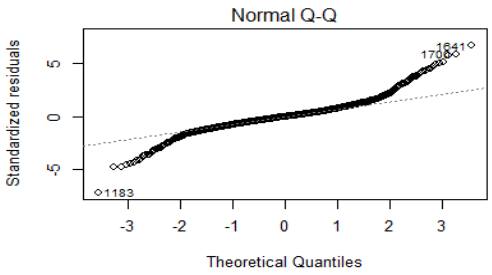
In this section we will fit 2 more models using the same set of predictor variables, but the response variables will be SalePrice and log(SalePrice). I will keep the same predictor variables used in Model 2 and add two more continuous predictor variables and one discrete variable.

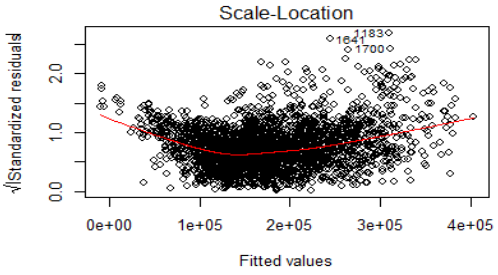
With this my model will contain 4 continuous predictor variables, 1 categorical variable and 1 discrete variable.

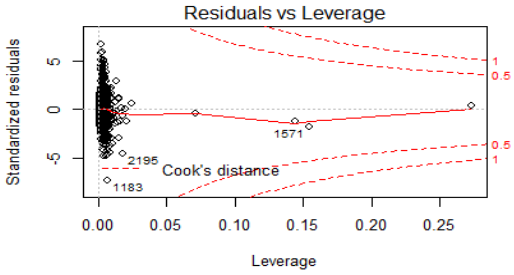
Model3 : SalePrice ~ GrLivArea+TotalBsmtSF + LotArea+ GarageArea+ OverallQual + Ngrp2 + Ngrp3 + Ngrp4 + Ngrp5

```
Call:
lm(formula = SalePrice ~ GrLivArea + TotalBsmtSF + LotArea +
    GarageArea + OverallQual + Ngrp2 + Ngrp3 + Ngrp4 + Ngrp5,
    data = subdat)

Residuals:
    Min      1Q  Median      3Q     Max
-159290  -11400    -471    9884  147607

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.138e+04  2.057e+03 -44.431  < 2e-16 ***
GrLivArea    8.707e+01  1.433e+00  60.744  < 2e-16 ***
TotalBsmtSF  1.513e+01  1.333e+00  11.347  < 2e-16 ***
LotArea      3.441e-01  5.764e-02   5.970 2.67e-09 ***
GarageArea   2.056e+01  2.594e+00   7.924 3.26e-15 ***
OverallQual  1.145e+04  4.891e+02  23.420  < 2e-16 ***
Ngrp2        2.867e+04  1.362e+03  21.043  < 2e-16 ***
Ngrp3        4.303e+04  1.408e+03  30.558  < 2e-16 ***
Ngrp4        5.376e+04  1.582e+03  33.977  < 2e-16 ***
Ngrp5        8.469e+04  1.751e+03  48.356  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22100 on 2850 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.9006,    Adjusted R-squared:  0.9003
F-statistic:  2868 on 9 and 2850 DF,  p-value: < 2.2e-16
```





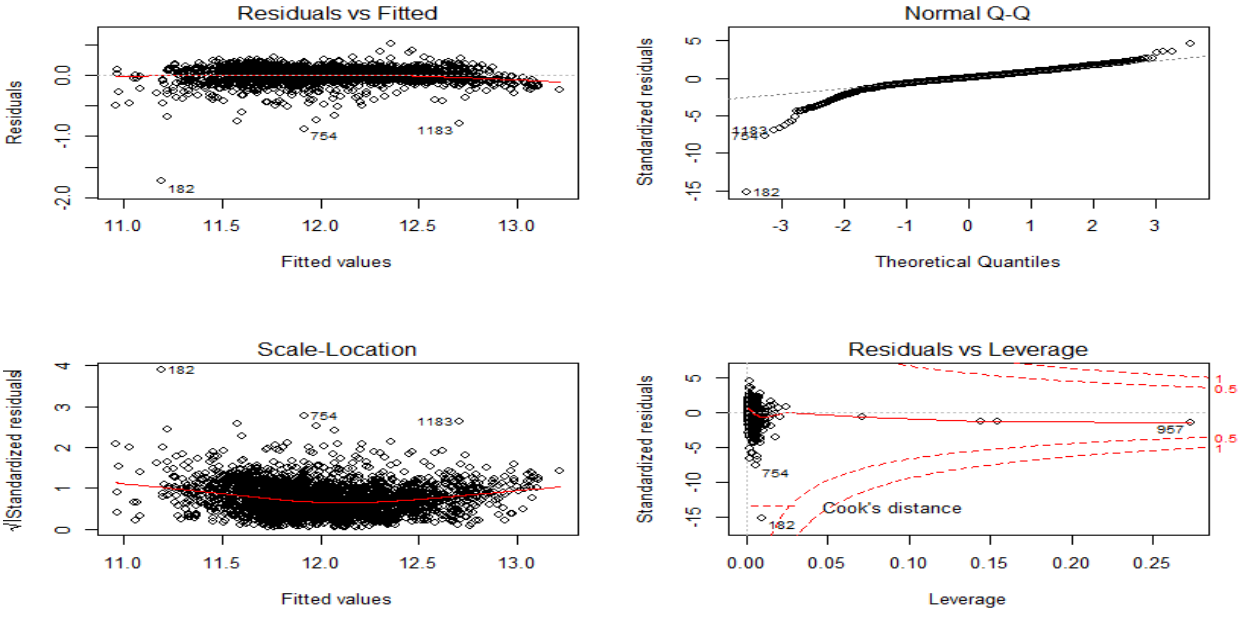




Interpretation of Model 3: When all the predictor variables are unchanged, for neighborhoods group of $0-100 average Price per sqft range predicted sale price increase by approximately $11450 for each increment in Overall Quality rating.

Model4: log(SalePrice) ~ GrLivArea+TotalBsmtSF + LotArea+ GarageArea+ OverallQual + Ngrp2 + Ngrp3 + Ngrp4 + Ngrp5

```
Call:
lm(formula = logSalePrice ~ GrLivArea + TotalBsmtSF + LotArea +
    GarageArea + OverallQual + Ngrp2 + Ngrp3 + Ngrp4 + Ngrp5,
    data = subdat)

Residuals:
      Min        1Q    Median        3Q       Max
-1.73591 -0.05018  0.00609  0.06426  0.51900

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.051e+01  1.069e-02 983.103  < 2e-16 ***
GrLivArea   4.615e-04  7.453e-06  61.926  < 2e-16 ***
TotalBsmtSF 7.551e-05  6.932e-06  10.892  < 2e-16 ***
LotArea     1.723e-06  2.997e-07   5.749 9.93e-09 ***
GarageArea  1.154e-04  1.349e-05   8.553  < 2e-16 ***
OverallQual 6.855e-02  2.543e-03  26.957  < 2e-16 ***
Ngrp2       2.087e-01  7.084e-03  29.461  < 2e-16 ***
Ngrp3       2.862e-01  7.322e-03  39.085  < 2e-16 ***
Ngrp4       3.376e-01  8.228e-03  41.033  < 2e-16 ***
Ngrp5       4.569e-01  9.106e-03  50.173  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1149 on 2850 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.9085,    Adjusted R-squared:  0.9082
F-statistic:  3144 on 9 and 2850 DF,  p-value: < 2.2e-16
```



Interpretation of Model 4: When all the predictor variables are unchanged, for neighborhoods group of $0-100 average Price per sqft range predicted log sale price increase by approximately 0.06855 for each

increment in Overall Quality rating. To get the normal sale price, exponent of log sale price should be taken.

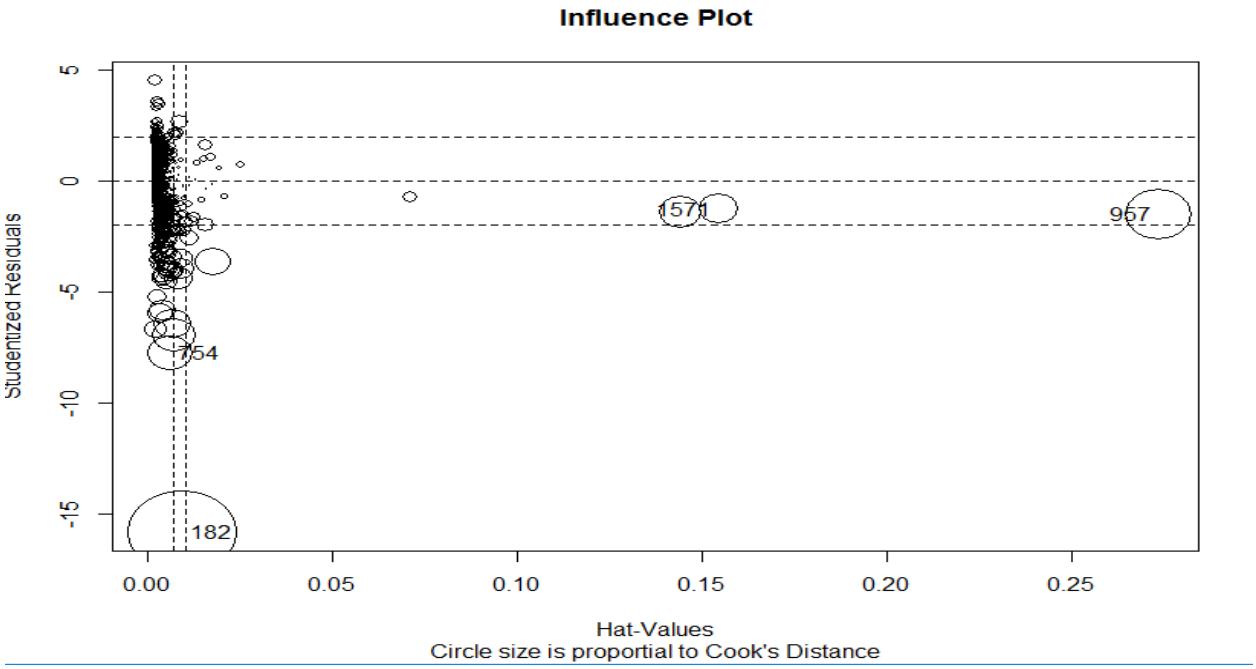
Model3 has better prediction accuracy then model 2 but Model4 has even better prediction accuracy compared to model 3. When validating goodness of fit using diagnostic plots for Model 3 and Model 4, model 4 residuals versus fitted plot of the logSalePrice shows a more linear fit then the non-transformed model 3. The Scale-Location plot for model 4 shows similar corrective behavior over model 3. I noticed that the Q-Q Plot shows improvement on positive tail in model 4 but does not appear to have much closer of a normal standardized residual approximation on the lower negative tail between model 3 and model 4. The Residuals Leverage plot interestingly shows the main cluster of residual points as expected to the left, but it also shows a small but mostly separate cluster between the .02 and the .025 leverage distance. These will likely need further investigation in the next section regarding influential points.

From this experiment, we can conclude that transformation of the response to log(SalePrice) did improve the model fit. In general we can use transformation of the response variable if model residual doesn't meet linearity and normality assumptions also known as heteroscedasticity.

Different transformation methods can be used based on variable type (values) and I don't think we can just use any transformation method on our data, it should meet the need of the model. The logarithmic transformation is a strong transformation with a major effect on distribution shape and it suits best for our need to reduce residual variation.

Section 7: Model based Outliers aka influential points:

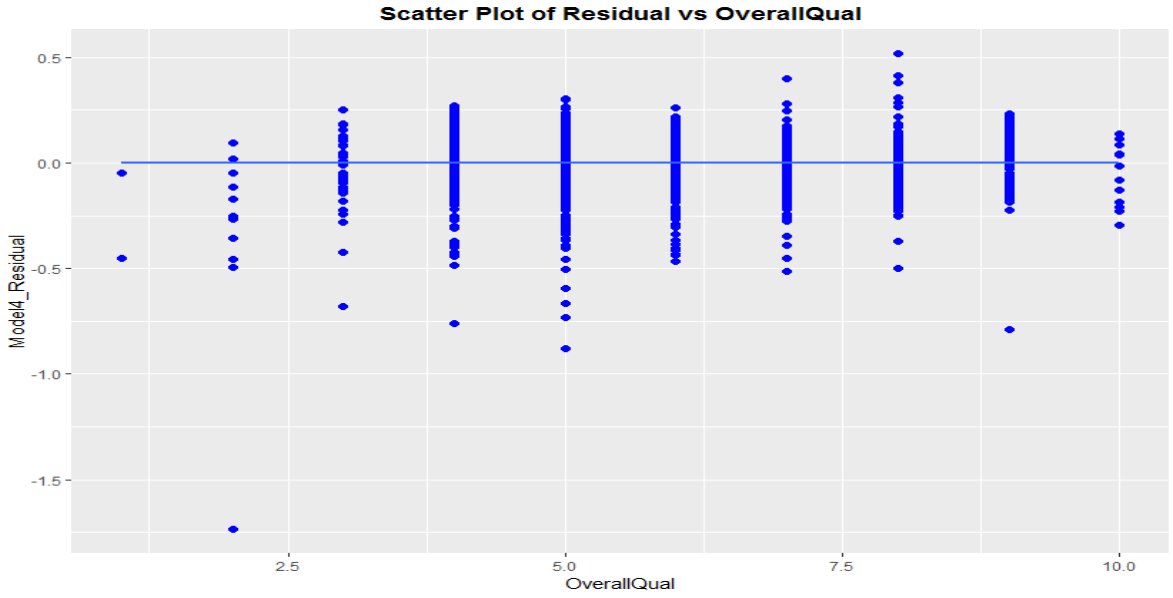
The initial Influence Plot from Model 4 is shown below:



The Influence plot reveals residuals which have high leverage and/or high influence. When checking the results from my Model 4 above, I see few residuals that sticks out with much higher than normal DFITS value indicating high leverage. For example observation 182, dffits value is -1.547. This value from observation 182 can be seen here.

	Model4_Residual	dffitslog
35	-0.24324465	-0.1223017
66	-0.17412953	-0.1475995
83	-0.33127366	-0.1554571
84	-0.25378970	-0.1602166
126	-0.45771767	-0.3330623
170	-0.43333421	-0.2295132
182	-1.73591302	-1.5470241
183	-0.26119953	-0.1285625
187	-0.73127691	-0.5247763
190	-0.27139434	-0.1784960
207	-0.38610462	-0.2133802
278	-0.36969210	-0.1679269
288	-0.24344120	-0.1196507
295	0.25690933	0.1196699
315	0.12729301	0.1472533
373	-0.51177468	-0.3159626
380	-0.16961808	-0.1470544
427	-0.18447407	-0.1804189
448	-0.21094325	-0.1367120
562	-0.33485007	-0.1701496
578	-0.21855845	-0.1463042

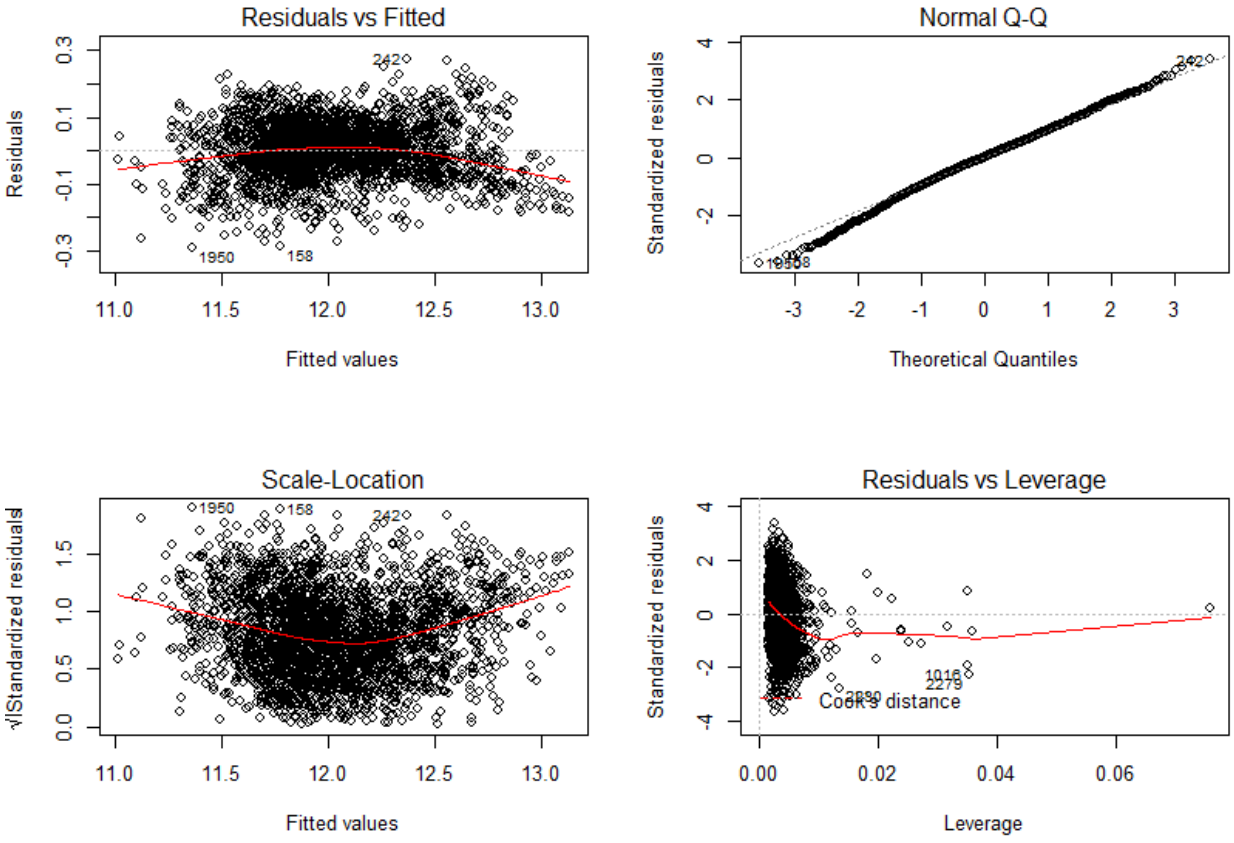
The below scatter plot between residual vs overall quality is also showing point way below or above the mean line which is shifting the mean residual for each quality ratings above or below the mean line.



The cutoff DFITS value for this data set is |0.1185| and there are about 133 observations which are larger than this and which can be usually classified as influential points. **I will go ahead and remove those 133 rows from my sample dataset and refit the model 4.**

The output of the MLR Model 4 after removing the influential points is as below:

```
Call:
lm(formula = logSalePrice ~ GrLivArea + TotalBsmtSF + LotArea +
    GarageArea + OverallQual + Ngrp2 + Ngrp3 + Ngrp4 + Ngrp5,
    data = subdatinf)

Residuals:
      Min        1Q    Median        3Q       Max
-0.292175 -0.049126  0.001565  0.051214  0.272572

Coefficients:
              Estimate Std. Error  t value Pr(>|t|)
(Intercept) 1.056e+01  8.115e-03 1300.678  < 2e-16 ***
GrLivArea   4.927e-04  5.849e-06   84.235  < 2e-16 ***
TotalBsmtSF 6.992e-05  5.121e-06   13.651  < 2e-16 ***
LotArea     2.745e-06  3.682e-07    7.455  1.2e-13 ***
GarageArea  9.660e-05  1.016e-05    9.512  < 2e-16 ***
OverallQual 5.944e-02  1.933e-03   30.758  < 2e-16 ***
Ngrp2       1.798e-01  5.133e-03   35.032  < 2e-16 ***
Ngrp3       2.629e-01  5.320e-03   49.420  < 2e-16 ***
Ngrp4       3.176e-01  5.981e-03   53.097  < 2e-16 ***
Ngrp5       4.421e-01  6.729e-03   65.699  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08051 on 2717 degrees of freedom
Multiple R-squared:  0.9489,    Adjusted R-squared:  0.9487
F-statistic:  5606 on 9 and 2717 DF,  p-value: < 2.2e-16
```



The model has improved in all fronts. The R-Squared value is now .95 and diagnostics plots are showing much better results. Residual vs Fitted and Scale-Location plots has mostly random distribution. In QQ Plot, most of the points are on mean line.

Conclusion & Reflections:
In summary our model using the log-transformed SalePrice as the dependent variable with a total of six predictors seems to give the best fit. Removing influential points improves the model furthers but we know that it is not advisable to automatically drop observations without a careful review. The 133 records that were removed should receive adequate checking. Which properties types are they, what were the sale conditions, where are they located? It was nice to see the model improve as we took steps to increase the number of predictors, test transformation, and remove specific observations as it helps reinforce the concepts that each step provides.

There are couple of things we can do as next step in modeling process.

1. We added 6 predictors in our model but we did not perform multicollinearity test.
2. We randomly selected predictors for model building. Of course we used the findings from EDA analysis but more analysis should be performed to select the best predictors.