**Pooja Deshpande**

# Modeling Assignment 2
# MSDS – 410 Data Modeling for Supervised Learning, Summer 2020
# Northwestern University

**Contents:**

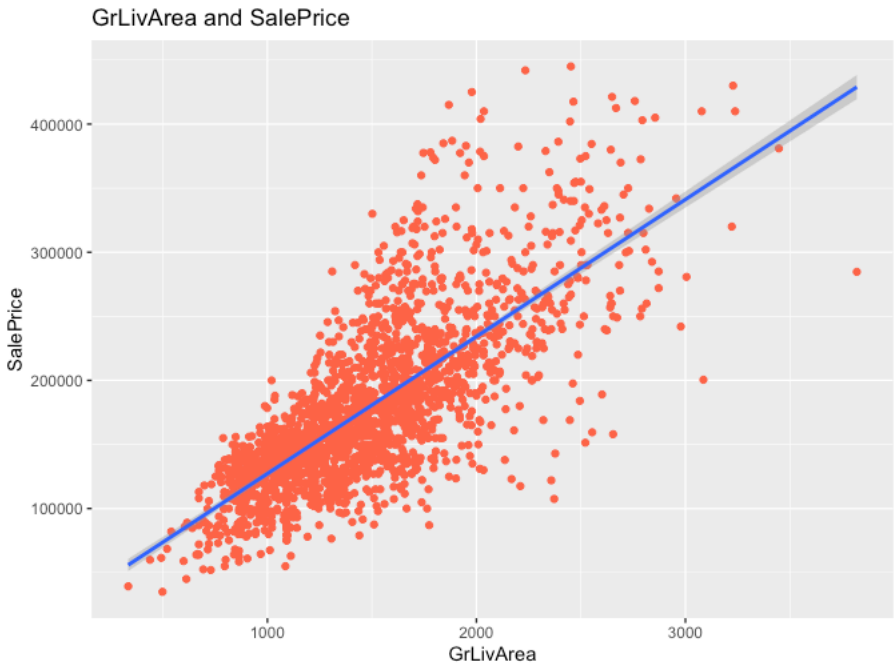
## Q.1

Response variable – SalePrice

Explanatory variable – GrLivArea

This variable was chosen based on the intuition that the above grade living area of a residence would highly determine its price level.

### a. Scatterplot of SalePrice and GrLivArea:




As it can be seen here, the response variable is positively correlated with the explanatory variable. The variance increases with an increase in the explanatory variable.

### b. Model1 and the Regression coefficient:

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

Residuals:
    Min      1Q  Median      3Q     Max
-166495  -25335   -1997   20005  194922

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept) 20243.503   3051.146   6.635      0.0000000000408 ***
GrLivArea     106.978      1.989  53.777 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 42480 on 2190 degrees of freedom
Multiple R-squared:  0.5691,    Adjusted R-squared:  0.5689
F-statistic:  2892 on 1 and 2190 DF,  p-value: < 0.00000000000000022
```

**Model equation:**

SalePrice = 20243.503 + 106.978*GrLivArea

**Coefficient interpretation:**

With every unit increase the above grade living area, the sale price of a residence in Ames increases by 106.978.

**c. R-squared and its interpretation:**

R-square = 0.5691

Interpretation – About 57% of the variance in Sale Price is explained by the above grade living area.

**d. ANOVA table results:**

```
> anova(model1)
Analysis of Variance Table

Response: SalePrice
            Df        Sum Sq       Mean Sq F value                Pr(>F)
GrLivArea    1 5218999900157 5218999900157    2892 < 0.00000000000000022 ***
Residuals 2190 3952124782842    1804623189
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

**Hypothesis tests for the coefficients:**

H0: $\beta 1 = 0$

Ha: $\beta 1 \; != 0$

p-value of the t-test < 0.0001,

Null hypothesis can be rejected as there is evidence that the relationship between the explanatory and response variable is significant.

**Omnibus F-test:**

Ho: Reduced model is adequate

Ha: Full model is adequate

RM: Ho: $Y = \beta 0$

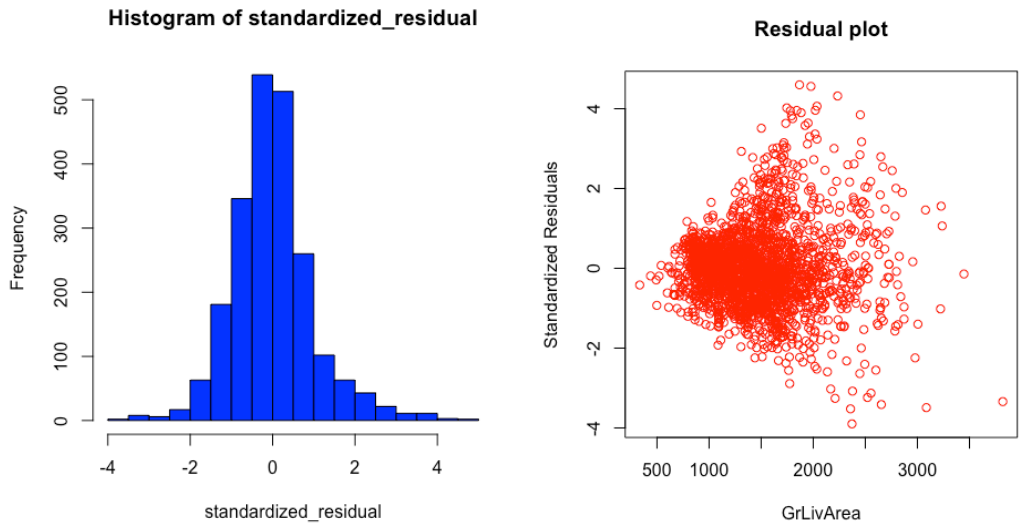
FM: Ha: $Y = \beta 0 + \beta 1 X1$

p-value < 0.0001.

Null hypothesis can be rejected which means we have scientific evidence that the predictor variable has a significant explanatory power. The full model is adequate.

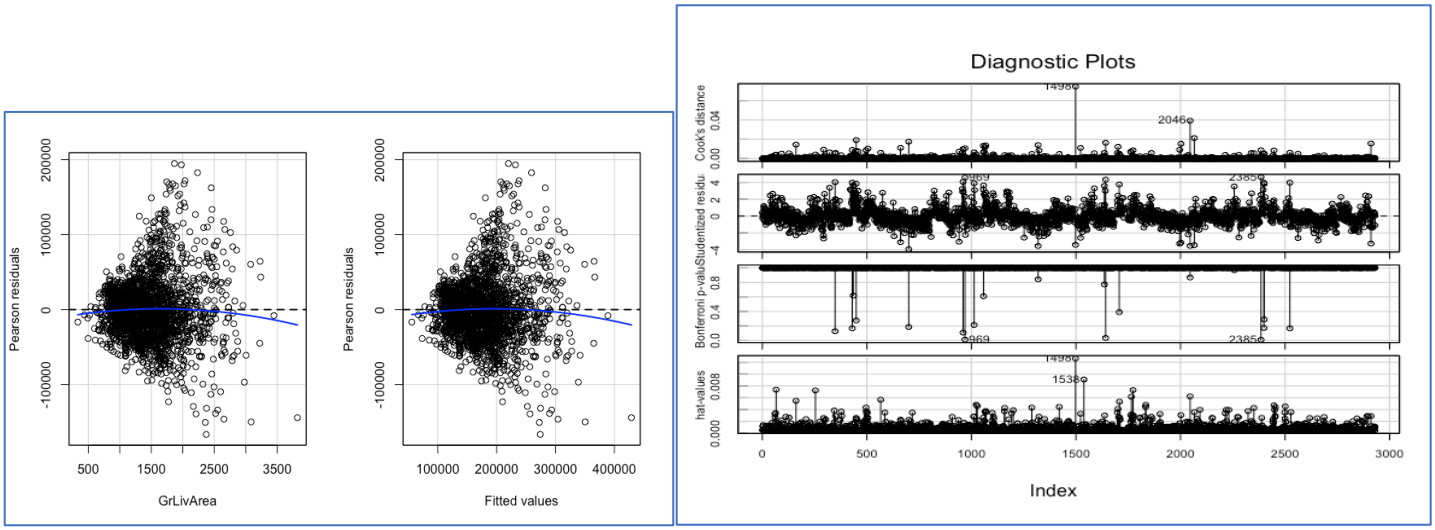
**e. Residual analysis:**

```
#Calculating fiited, residuals and standardized residuals
fitted <- 20243 + 106*final_df$GrLivArea
residual <- final_df$SalePrice - fitted
mean_r <- mean(residual)
std_r <- sd(residual)
standardized_residual <- (residual - mean_r)/std_r

#Plotting standardized residuals
hist(standardized_residual, col = 'blue')
plot(final_df$GrLivArea,standardized_residual,col = 'red',main = 'Residual plot',
     xlab = 'GrLivArea',ylab = 'Standardized Residuals')
```




- The histogram of the standardized residuals looks approximately normal.
- The plot of the predictor vs the residuals suggests a pattern. The plot resembles a funnel which violates the homoscedasticity assumption.

**Residual and Index plots for Influential points:**



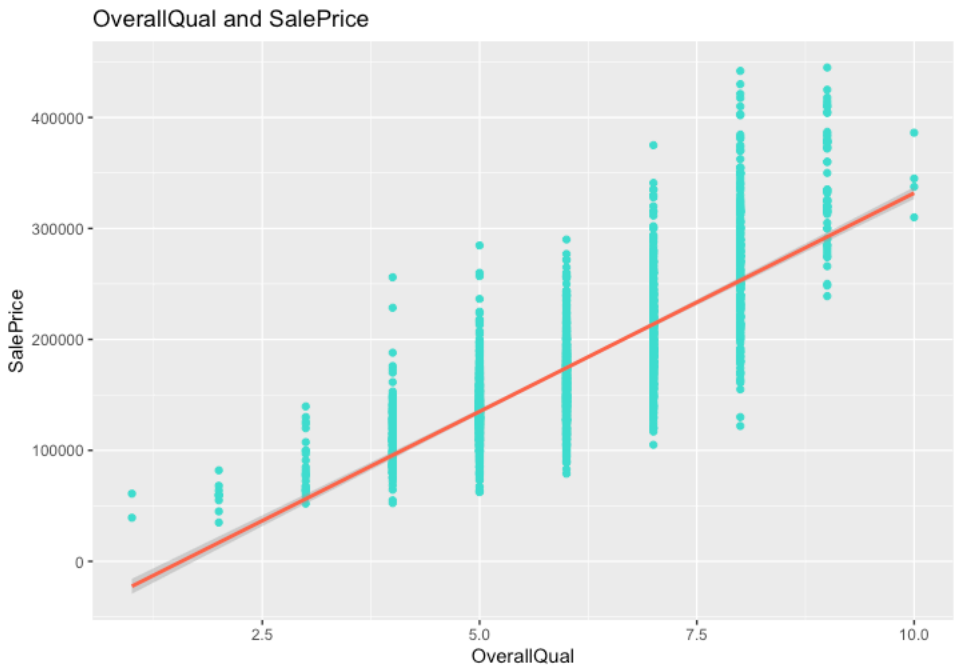

- Based on the index plots of Cook's distance and hat-values, there seem to be outliers in the response and outliers in the predictor which could be categorized as Influential Points.

## Q.2

Response variable – SalePrice

Explanatory variable – OverallQual

### a. Scatterplot of SalePrice and OverallQual (discrete variable):




As it can be seen here, there is definitely a pattern between OverallQual and SalePrice. As the overall quality index of a home increases, its price range increases too.

### b. Model2 and the Regression coefficient:

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

Residuals:
    Min      1Q  Median      3Q     Max
-131101  -25006   -2506   20280  188828

Coefficients:
             Estimate Std. Error t value            Pr(>|t|)
(Intercept) -61819.5     4056.8  -15.24 <0.0000000000000002 ***
OverallQual  39365.1      654.2   60.18 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 39730 on 2190 degrees of freedom
Multiple R-squared:  0.6231,    Adjusted R-squared:  0.623
F-statistic:  3621 on 1 and 2190 DF,  p-value: < 0.00000000000000022
```

**Model equation:**

SalePrice = -61819.5 + 39365.1*OverallQual

**Coefficient interpretation:**

With every unit increase in the overall quality, the sale price of a residence in Ames increases by 39365.1. Since OverallQual is a discrete variable and the distribution of SalePrice and OverallQual suggests that a 1 unit increase in OverallQual doesn't actually cause a 39365.1 increase in the sale price. It probably applies to only those values which are near the median within each quality index group.

**c. R-squared and its interpretation:**

R-square = 0.6231

Interpretation – About 62% of the variance in Sale Price is explained by the overall quality index

**d. ANOVA table results:**

```
Analysis of Variance Table

Response: SalePrice
              Df        Sum Sq       Mean Sq F value                Pr(>F)
OverallQual    1 5714855806893 5714855806893  3621.1 < 0.00000000000000022 ***
Residuals   2190 3456268876106    1578204966
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

**Hypothesis tests for the coefficients:**

H0: $\beta 1 = 0$

Ha: $\beta 1 \neq 0$

p-value of the t-test < 0.0001,

Null hypothesis can be rejected as there is evidence that the relationship between the explanatory and response variable is significant.

**Omnibus F-test:**

Ho: Reduced model is adequate

Ha: Full model is adequate

RM: Ho: $Y = \beta 0$

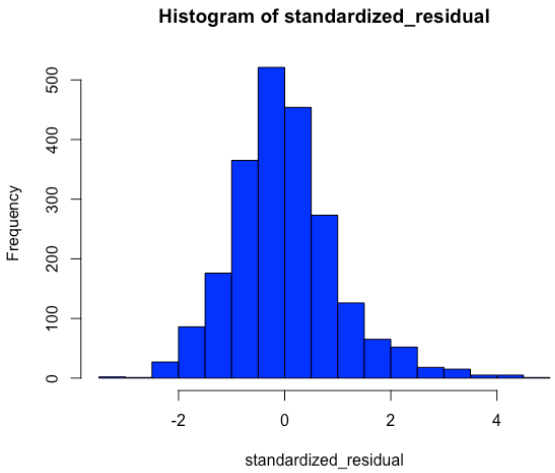
FM: Ha: $Y = \beta 0 + \beta 1 X1$

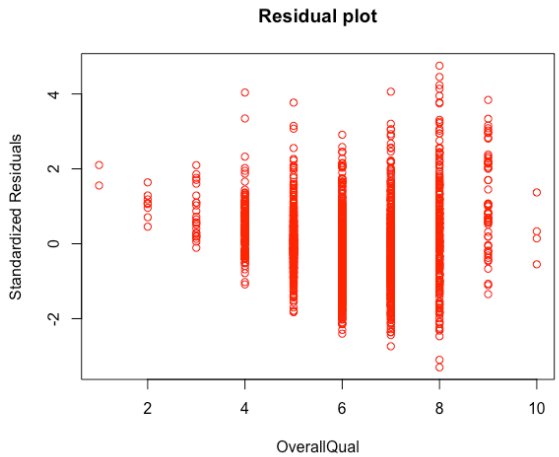
p-value < 0.0001.

Null hypothesis can be rejected which means we have scientific evidence that the predictor variable has a significant explanatory power. The full model is adequate.

**e. Residual analysis:**

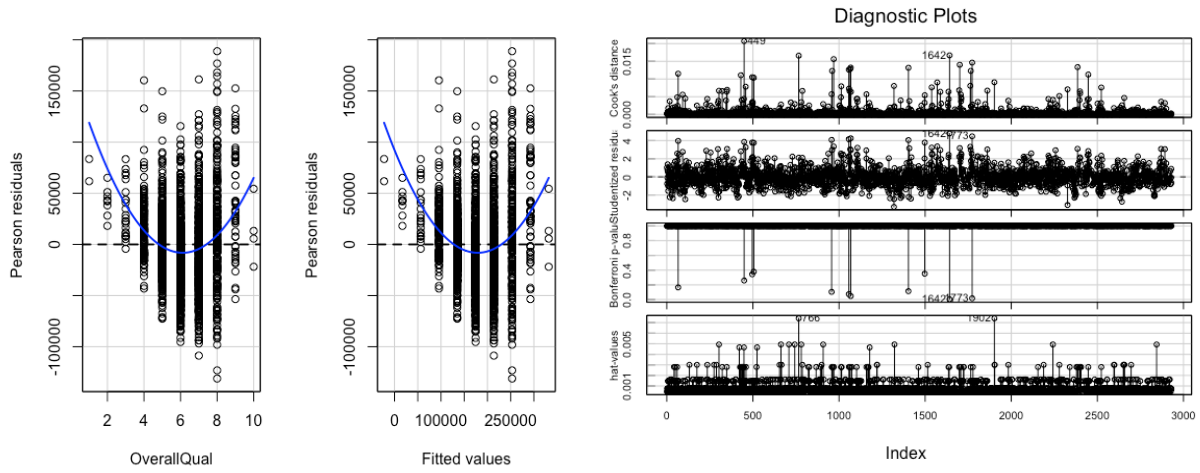
```
#Calculating fiited, residuals and standardized residuals
fitted <- 39365.1*final_df$OverallQual - 61819.5
residual <- final_df$SalePrice - fitted
mean_r <- mean(residual)
std_r <- sd(residual)
standardized_residual <- (residual - mean_r)/std_r
```



- The histogram of the standardized residuals looks more or less normal.
- The plot of the predictor vs the residuals suggests a pattern. The plot resembles a funnel which violates the homoscedasticity assumption.

**Residual and Index plots for Influential points:**




- From the Pearson residual plots, it can be observed that there is curved pattern which provides evidence for the presence of heteroscedasticity.
- Based on the index plots of Cook's distance and hat-values, there seem to be outliers in the response and outliers in the predictor which could be categorized as Influential Points.

**Q.3**

Both Model1 and Model2 are better than the baseline model where Y = y_bar. However, Model1 seems to have a better fit because there is a clear indication of a positive correlation between

GrLivArea and the SalePrice. The coefficient in the model applies to all the values of X. Although there is evidence of heteroscedasticity, the lines in the residual plots are not very curved which indicates that the errors in variance can be corrected with the help of transformations. There seem to be fewer influential points in Model1.

## Q.4

Response variable – SalePrice

Explanatory variables – GrLivArea, OverallQual

### a. Model3 and the Regression coefficient:

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

Residuals:
    Min      1Q  Median      3Q     Max
-163972  -19884     113   17686  163939

Coefficients:
             Estimate Std. Error t value            Pr(>|t|)
(Intercept) -78854.577   3245.715  -24.30 <0.0000000000000002 ***
GrLivArea       64.372      1.781   36.14 <0.0000000000000002 ***
OverallQual  26629.524    626.364   42.51 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 31450 on 2189 degrees of freedom
Multiple R-squared:  0.764,     Adjusted R-squared:  0.7637
F-statistic:  3543 on 2 and 2189 DF,  p-value: < 0.00000000000000022
```

**Model equation:**

SalePrice = -78854.577 + 26629.524*OverallQual + 64.374*GrLivArea

**Coefficient interpretation:**

With every unit increase in the overall quality, the sale price of a residence in Ames increases by 26629.524 when adjusted for other predictors. With every unit increase in the above grade living

area, the sale price of a residence in Ames increases by 64.374 when adjusted for other predictors.

**b. R-squared and its interpretation:**

R-square = 0.764

Interpretation – About 76% of the variance in Sale Price is explained by the overall quality index and above grade living area. It may be better to see the adjusted R-square since it penalizes the model every time a new predictor gets added to it.

Adjusted R-square = 0.7637, there isn't much of a difference between the R-square and adjusted R-square.

Difference between R-square of Model1 and Model 3 = 0.764 – 0.5691 = 0.1949.

About 19.4% of the variance is explained by the added predictor.

Adding an extra predictor has increased R-square since it accounts for some part of the variance in the response variable. It has a significant explanatory power.

**c. ANOVA table results:**

```
Analysis of Variance Table

Response: SalePrice
              Df        Sum Sq       Mean Sq F value                Pr(>F)
GrLivArea      1 5218999900157 5218999900157  5277.6 < 0.00000000000000022 ***
OverallQual    1 1787418247138 1787418247138  1807.5 < 0.00000000000000022 ***
Residuals   2189 2164706535704     988902026
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

**Hypothesis tests for the coefficients:**

H0: $\beta 1 = 0$

Ha: $\beta 1 \neq 0$

p-value of the t-test < 0.0001,

Null hypothesis can be rejected as there is evidence that the relationship between the explanatory (GrLivArea) and response variable is significant.

H0: $\beta 2 = 0$

Ha: $\beta 2 \; != 0$

p-value of the t-test < 0.0001,

Null hypothesis can be rejected as there is evidence that the relationship between the explanatory (Overall Qual) and response variable is significant.

**Omnibus F-test:**

Ho: Reduced model is adequate

Ha: Full model is adequate

RM: Ho: $Y = \beta 0$ or $\beta 1 = \beta 2 = 0$

FM: Ha: $Y = \beta 0 + \beta 1X1 + \beta 2X2$ or At least one coefficient is not equal to zero

p-value < 0.0001.

Null hypothesis can be rejected which means we have scientific evidence that at least one predictor variable has a significant explanatory power. The full model is adequate.

**ANOVA results:**

Anova function performs sequential tests where it compares models.

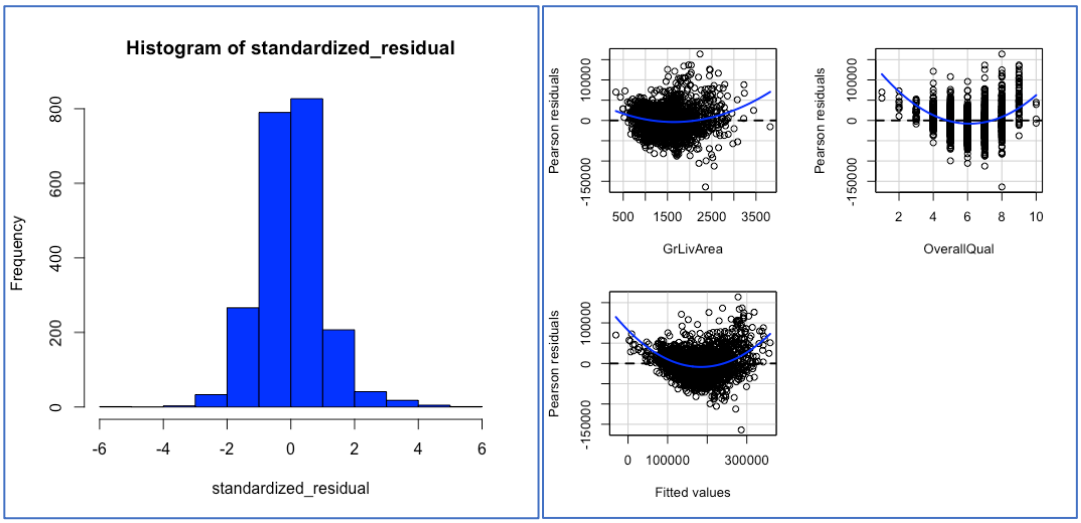
Here it compares,

SalePrice ~ 1 versus SalePrice ~ GrLivArea

SalePrice ~ GrLivArea versus SalePrice ~ GrLivArea + OverallQual.

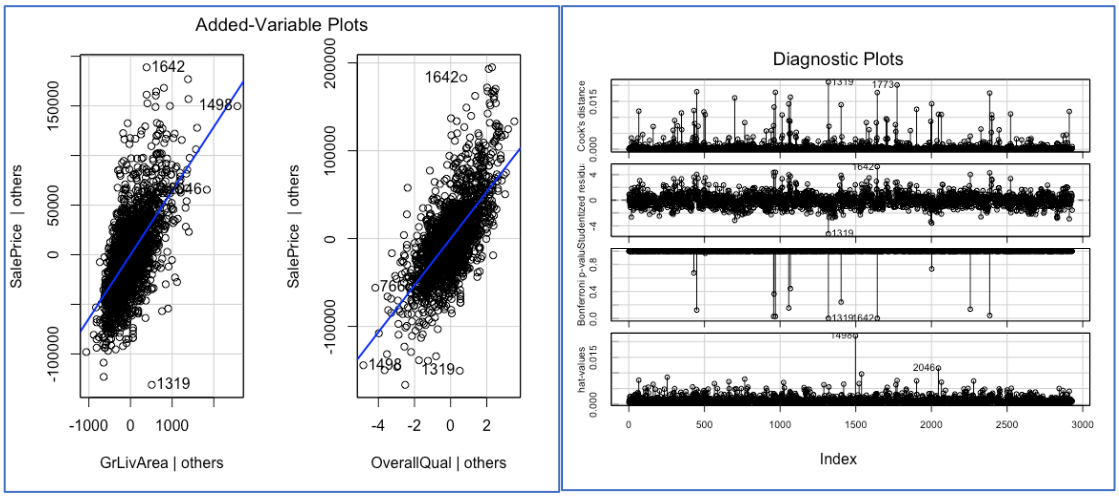
Since the p-value for the second test is low, we can conclude that Model3 is significant.

**d. Residual analysis:**




- The histogram of the standardized residuals follows a normal distribution.
- The plots of the predictors vs the residuals suggest a curved pattern. This indicates violation of the homoscedasticity assumption.

**Added Variable plots and Index plots for Influential points:**



- The added variable plots measure the residuals of the response variable and the residuals of each predictor. A strong positive linear relationship indicates that the predictor is significant.
- Based on the index plots of Cook's distance, there seem to be a few Influential Points.

**e. Retention of both predictors or not:**

Both the variables should be retained since the adjusted r-square value has increased from models 1 and 2. The omnibus F-test and the anova test results also help in concluding that both the variables are significant explanators of the response variable, SalePrice. Inclusion of both the variables doesn't cause a large change in their coefficients or the t-test statistics. However, there is some indication of heteroscedasticity and influential points.

**Q.5**

Response variable – SalePrice

Explanatory variables – GrLivArea, OverallQual, LotArea

**a. Model4 and the Regression coefficients:**

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

Residuals:
    Min      1Q  Median      3Q     Max
-154009  -18832       5   16378  162211

Coefficients:
              Estimate   Std. Error t value            Pr(>|t|)
(Intercept) -86556.14097   3153.10064  -27.45 <0.0000000000000002 ***
GrLivArea       58.15104      1.75995   33.04 <0.0000000000000002 ***
OverallQual  27369.28181    601.69551   45.49 <0.0000000000000002 ***
LotArea          1.23635      0.08698   14.21 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 30100 on 2188 degrees of freedom
Multiple R-squared:  0.7839,    Adjusted R-squared:  0.7836
F-statistic:  2646 on 3 and 2188 DF,  p-value: < 0.00000000000000022
```

**Model equation:**

SalePrice = -86556.14097 + 27369.2818*OverallQual + 58.151*GrLivArea + 1.23*LotArea

**Coefficient interpretation:**

With every unit increase in the overall quality, the sale price of a residence in Ames increases by 27369.2818 when adjusted for other predictors. With every unit increase in the above grade living area, the sale price of a residence in Ames increases by 58.151 when adjusted for other predictors. With every unit increase in the lot area, the sale price of a residence in Ames increases by 1.23 when adjusted for other predictors. There is a larger change in the coefficient of GrLivArea when compared to the simple linear regression model.

**b. R-squared and its interpretation:**

R-square = 0.7834

Interpretation – About 78% of the variance in Sale Price is explained by the overall quality index, above grade living area and lot area. It may be better to see the adjusted R-square since it penalizes the model every time a new predictor gets added to it.

Adjusted R-square = 0.7836, there isn't much of a difference between the R-square and adjusted R-square.

Difference between R-square of Model4 and Model 3 = 0.7834 - 0.764 = 0.0194.

Only 1% of variance is explained by 'LotArea'.

Adding 'LotArea' hasn't increased the R-square by a lot. However, the adjusted R-Square has increased which means that LotArea adds to the predictability of SalePrice more than expected by chance.

**c. ANOVA table results:**

```
> anova(model4)
Analysis of Variance Table

Response: SalePrice
              Df        Sum Sq       Mean Sq F value                Pr(>F)
GrLivArea      1 5218999900157 5218999900157 5762.30 < 0.00000000000000022 ***
OverallQual    1 1787418247138 1787418247138 1973.49 < 0.00000000000000022 ***
LotArea        1  183001304005  183001304005  202.05 < 0.00000000000000022 ***
Residuals   2188 1981705231699     905715371
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

**Hypothesis tests for the coefficients:**

H0: $\beta 1 = 0$

Ha: $\beta 1 \neq 0$

p-value of the t-test < 0.0001,

Null hypothesis can be rejected as there is evidence that the relationship between the explanatory (GrLivArea) and response variable is significant.

H0: $\beta 2 = 0$

Ha: $\beta 2 \neq 0$

p-value of the t-test < 0.0001,

Null hypothesis can be rejected as there is evidence that the relationship between the explanatory (Overall Qual) and response variable is significant.

H0: $\beta 3 = 0$

Ha: $\beta 3$ != 0

p-value of the t-test < 0.0001,

Null hypothesis can be rejected as there is evidence that the relationship between the explanatory (LotArea) and response variable is significant.

**Omnibus F-test:**

Ho: Reduced model is adequate

Ha: Full model is adequate

RM: Ho: Y = $\beta 0$ or $\beta 1 = \beta 2 = \beta 3 = 0$

FM: Ha: Y = $\beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X3$ or At least one coefficient is not equal to zero

p-value < 0.0001.

Null hypothesis can be rejected which means we have scientific evidence that at least one predictor variable has a significant explanatory power. The full model is adequate.

**ANOVA results:**

Anova function performs sequential tests where it compares models.

Here it compares,

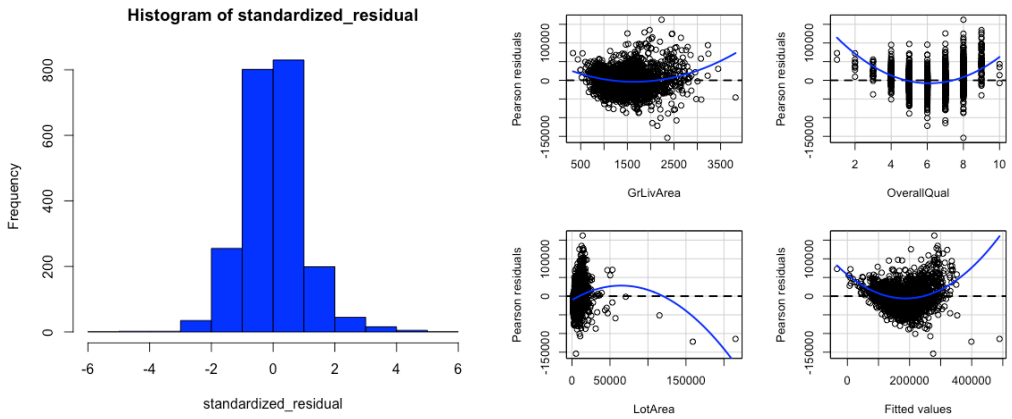
SalePrice ~ 1 versus SalePrice ~ GrLivArea

SalePrice ~ GrLivArea versus SalePrice ~ GrLivArea + OverallQual.

SalePrice ~ GrLivArea + OverallQual versus SalePrice ~ GrLivArea + OverallQual + LotArea

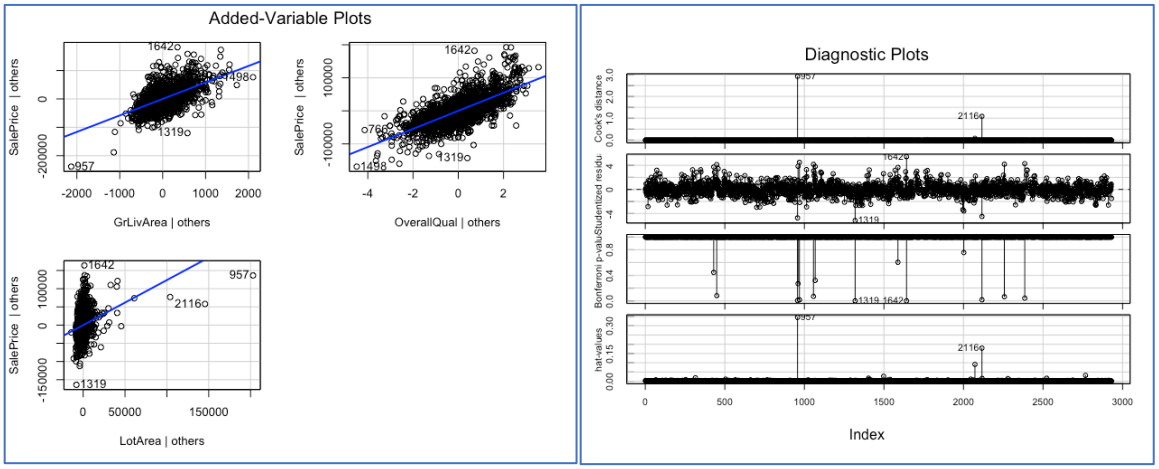
Since the p-value for the third test is low, we can conclude that Model4 is significant.

**d. Residual analysis:**




- The histogram of the standardized residuals follows a normal distribution.
- The plots of the predictors vs the residuals suggest a curved pattern especially the LotArea. This indicates violation of the homoscedasticity assumption.

## Added Variable plots and Index plots for Influential points:




- The added variable plots measure the residuals of the response variable and the residuals of each predictor. A strong positive linear relationship indicates that the predictor is significant. As we see here in the plots, LotArea residuals does not have a linear relationship with SalePrice residuals.
- Based on the index plots of Cook's distance, there seem to be **TWO** Influential Points.

**e. Retention of both predictors or not:**

Although LotArea has a significant t-test statistic, it may be safe to remove it since the R-square of the model doesn't increase as much. The residual plots indicate extreme deviation from homoscedasticity. The added variable plot doesn't indicate a positive linear relationship between SalePrice and LotArea. In addition, there might be a collinearity problem between GrLivArea and LotArea which violates another assumption.

**Q.6**

| Model Name | Log transformation | Adjusted R-square | T-test p-values | F-test p-value | Residuals normal? | Residuals vs predictors (Homoscedasticity check) | Influential Observations |
|---|---|---|---|---|---|---|---|
| Model 1 | No | 0.5689 | < 0.001 (all) | < 0.001 | Approximately | Slightly violated | A few |
| Model 1 | Yes | 0.5622 | < 0.001 (all) | < 0.001 | Approximately | Curved pattern in variance | A few |

| Model Name | Log transformation | Adjusted R-square | T-test p-values | F-test p-value | Residuals normal? | Residuals vs predictors (Homoscedasticity check) | Influential Observations |
|---|---|---|---|---|---|---|---|
| Model 3 | No | 0.7637 | < 0.001 (all) | < 0.001 | Yes | Curved pattern in variance | Many |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Model 4 | Yes | 0.7766 | < 0.001 (all) | < 0.001 | Yes | Slightly curved only in GrLivArea | A few |

| **Model Name** | **Log transformation** | **Adjusted R-square** | **T-test p-values** | **F-test p-value** | **Residuals normal?** | **Residuals vs predictors (Homoscedasticity check)** | **Influential Observations** |
|---|---|---|---|---|---|---|---|
| Model 4 | No | 0.7836 | < 0.001 (all) | < 0.001 | Yes | Curved pattern in variance | Very few |
| Model 4 | Yes | 0.7943 | < 0.001 (all) | < 0.001 | Yes | Curved in all except OverallQual | Only two |

Models 3 and 4 after the log transformation have a higher adjusted r-square values. However, Model 3 with log transformation considerably reduces the heteroscedasticity. 'LotArea' doesn't add much to the predictability of the response variable. In order to have reliable results, all the assumptions need to be satisfied. Hence, Model 3 with the log transformation could be considered as the best fitting model.

## Q.7

The interpretation of models with transformed variables is different from those without any transformed variables. For example: Interpretation of model 3 with log transformed response variable.

log(SalePrice) = 10.605 + 0.0003*GrLivArea + 0.1522*OverallQual

For one unit change in OverallQual, there is (e^(0.1522) -1 )* 100 % increase in SalePrice i.e;

For one-unit change in OverallQual there is 16.44% increase in SalePrice when adjusted for other predictors. Although, transformed models are harder to interpret, they provide rather reliable results since transformations help in satisfying the linearity, normality and homoscedasticity assumptions in linear models.

## Q.8

### Results from Model4 built using lessR

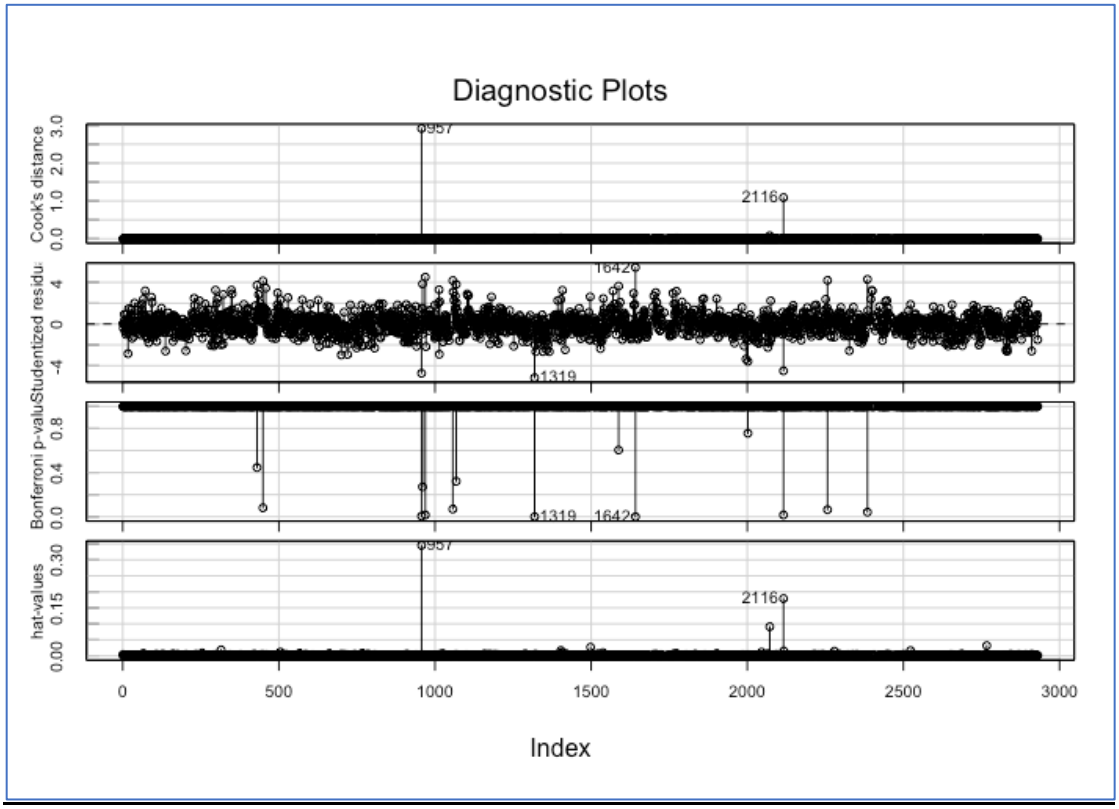
```
  BACKGROUND

Data Frame:  final_df

Response Variable: SalePrice
Predictor Variable 1: GrLivArea
Predictor Variable 2: OverallQual
Predictor Variable 3: LotArea

Number of cases (rows) of data:  2192
Number of cases retained for analysis:  2192


  BASIC ANALYSIS

Estimated Model

             Estimate    Std Err  t-value  p-value    Lower 95%    Upper 95%
(Intercept) -86556.141   3153.101  -27.451    0.000   -92739.525   -80372.757
  GrLivArea     58.151      1.760   33.041    0.000       54.700       61.602
OverallQual  27369.282    601.696   45.487    0.000    26189.328    28549.236
    LotArea      1.236      0.087   14.214    0.000        1.066        1.407
```

```
Model Fit

Standard deviation of residuals:  30095.105 for 2188 degrees of freedom

R-squared:  0.784    Adjusted R-squared:  0.784    PRESS R-squared:  0.780

Null hypothesis that all population slope coefficients are 0:
  F-statistic: 2645.945     df: 3 and 2188     p-value:  0.000


Analysis of Variance

              df             Sum Sq            Mean Sq   F-value   p-value
  GrLivArea    1  5218999900156.658  5218999900156.658  5762.296     0.000
OverallQual    1  1787418247137.720  1787418247137.720  1973.488     0.000
    LotArea    1   183001304005.000   183001304005.000   202.052     0.000

Model          3  7189419451299.379  2396473150433.126  2645.945     0.000
Residuals   2188  1981705231699.497      905715370.978
SalePrice   2191  9171124682998.875     4185816833.865
```

## Collinearity

| | Tolerance | VIF |
|---|---|---|
| GrLivArea | 0.641 | 1.560 |
| OverallQual | 0.678 | 1.474 |
| LotArea | 0.934 | 1.071 |

```
  RESIDUALS AND INFLUENCE

Data, Fitted, Residual, Studentized Residual, Dffits, Cook's Distance
   [sorted by Cook's Distance]
   [res_rows = 20, out of 2192 rows of data, or do res_rows="all"]
-------------------------------------------------------------------------------------------
      GrLivArea OverallQual    LotArea  SalePrice     fitted       resid rstdnt dffits cooks
  957      2036           7     215245     375000 489542.807 -114542.807 -4.726 -3.432 2.916
 2116      2144           6     159000     277000 398915.255 -121915.255 -4.490 -2.097 1.090
 2072      1824           7     115149     302000 353460.959  -51460.959 -1.794 -0.568 0.081
 1403      1663           4      53227     256000 185433.446   70566.554  2.367  0.310 0.024
 1773      3228           8      12692     430000 335801.453   94198.547  3.149  0.279 0.019
 1319      2358           8       5250     122000 276009.118 -154009.118 -5.155 -0.276 0.019
  315      1687           5      57200     160000 219110.377  -59110.377 -1.984 -0.273 0.019
 1498      3820           5      47007     284700 330544.422  -45844.422 -1.545 -0.257 0.017
  505      2448           8      46589     402000 332352.242   69647.758  2.330  0.253 0.016
 2523      1842           9      50271     385000 329034.237   55965.763  1.876  0.242 0.015
 1642      2234           8      14082     441929 279717.844  162211.156  5.430  0.241 0.014
  449      2452           9      15274     445000 321237.784  123762.216  4.134  0.240 0.014
  969      1978           9      12633     425000 290408.986  134591.014  4.498  0.240 0.014
 2385      1868           9      14780     415000 286666.817  128333.183  4.288  0.238 0.014
 1068      2649           8      16900     421250 307334.565  113915.435  3.804  0.228 0.013
  700      2372           5       6600     107500 196384.462  -88884.462 -2.967 -0.218 0.012
 1057      2464           8      13693     417500 292611.643  124888.357  4.171  0.217 0.012
 1902       334           1       5000      39300 -33582.654   72882.654  2.434  0.212 0.011
 2002      2554           7      10800     159500 266899.191 -107399.191 -3.585 -0.211 0.011
   66      3238           8      14720     410000 338890.284   71109.716  2.375  0.210 0.011
```




Based on the above plot and the cook's distance values in the previous table, there are two influential points (observation 957 and observation 2116). They have cook's distance values of 2.91 and 1.09. These could be considered as influential since they are greater than 1 (cut-off) and also lie away from the rest of the cook's distance values.

A new model was constructed after removing the entries with SID = 957 and SID = 2116. An increase in the adjusted R-square value was observed (0.788). After examining the diagnostic plots, another observation (SID = 2072) was removed. This increased the adjusted R-square value to 0.791.

```
  BASIC ANALYSIS

Estimated Model

              Estimate    Std Err  t-value  p-value    Lower 95%    Upper 95%
(Intercept) -93462.920   3181.085  -29.381    0.000   -99701.188   -87224.652
  GrLivArea     54.178      1.779   30.449    0.000       50.689       57.667
OverallQual  27872.332    592.973   47.004    0.000    26709.482    29035.182
    LotArea      2.241      0.137   16.332    0.000        1.972        2.510


Model Fit

Standard deviation of residuals:  29521.437 for 2185 degrees of freedom

R-squared:  0.791    Adjusted R-squared:  0.791    PRESS R-squared:  0.790

Null hypothesis that all population slope coefficients are 0:
  F-statistic: 2754.540     df: 3 and 2185     p-value:  0.000
```

```
  RESIDUALS AND INFLUENCE

Data, Fitted, Residual, Studentized Residual, Dffits, Cook's Distance
   [sorted by Cook's Distance]
   [res_rows = 20, out of 2189 rows of data, or do res_rows="all"]
-----------------------------------------------------------------------------------------
       GrLivArea OverallQual   LotArea  SalePrice     fitted       resid rstdnt dffits cooks
   315      1687           5     57200     160000 265470.501 -105470.501 -3.670 -0.813 0.164
  2767      1533           7     70761     280000 343259.104  -63259.104 -2.238 -0.667 0.111
  1498      3820           5     47007     284700 358192.056  -73492.056 -2.540 -0.497 0.062
  2117      1953           6     53107     240000 298582.655  -58582.655 -2.025 -0.402 0.040
  2279      2034           3     43500     130000 197826.805  -67826.805 -2.331 -0.383 0.037
  1014      1474           6     31220     115000 223587.133 -108587.133 -3.708 -0.379 0.036
  1407      2358           6     45600     240000 303703.175  -63703.175 -2.187 -0.351 0.031
  1319      2358           8      5250     122000 269031.908 -147031.908 -5.017 -0.297 0.022
  1773      3228           8     12692     430000 332842.846   97157.154  3.312  0.296 0.022
  1642      2234           8     14082     441929 282104.490  159824.510  5.455  0.247 0.015
  2385      1868           9     14780     415000 291711.700  123288.300  4.199  0.245 0.015
   969      1978           9     12633     425000 292860.315  132139.685  4.502  0.244 0.015
   449      2452           9     15274     445000 324458.676  120541.324  4.105  0.243 0.015
   807      2016           5     26400     131000 214278.731  -83278.731 -2.835 -0.232 0.013
  2687      2486           6     33120     220000 282672.899  -62672.899 -2.137 -0.231 0.013
  1068      2649           8     16900     421250 310902.962  110347.038  3.756  0.230 0.013
  1902       334           1      5000      39300 -36291.138   75591.138  2.574  0.225 0.013
  1057      2464           8     13693     417500 293693.791  123806.209  4.216  0.220 0.012
   700      2372           5      6600     107500 189198.475  -81698.475 -2.780 -0.217 0.012
    66      3238           8     14720     410000 337928.952   72071.048  2.454  0.217 0.012
```

Three influential points were removed (Sid = 957, SID = 2116, SID = 2072). None of the cook's distance values are greater than one.

Since we are trying to model the average price of a typical family home in Ames, it is justifiable to remove certain outliers/ influential points which greatly affect the regression line.

## Q.9

## Approach to build a multiple regression model:

- The following predictors were considered while building successive multiple regression models to finally choose the best one:
- MasVnrArea, TotalBsmtSF, FirstFlrSF, SecondFlrSF, GarageArea, WoodDeckSF, OpenPorchSF, PoolArea, age.
- Model 4 is used as a starting point.
- One new predictor is added to a model in each step.
- It is retained only if the adjusted R-squared increases by a minimum of 0.01, AND the anova test is significant AND there isn't much difference in the coefficients/t-tests of the other existing predictors.
- After building 9 successive models, the final predictors were chosen to be : GrLivArea + OverallQual + LotArea + MasVnrArea + TotalBsmtSF + GarageArea + age.

```
Call:
lm(formula = SalePrice ~ GrLivArea + OverallQual + LotArea +
    MasVnrArea + TotalBsmtSF + GarageArea + age, data = final_df)

Residuals:
   Min     1Q Median     3Q    Max
-97730 -15185  -1228  12888 115629

Coefficients:
               Estimate   Std. Error t value             Pr(>|t|)
(Intercept) -46826.90974   3683.96305 -12.711 < 0.0000000000000002
GrLivArea       52.19044      1.49357  34.943 < 0.0000000000000002
OverallQual  15662.05878    609.79894  25.684 < 0.0000000000000002
LotArea          0.87995      0.07184  12.248 < 0.0000000000000002
MasVnrArea      21.30208      3.47013   6.139       0.000000000986
TotalBsmtSF     34.23968      1.65271  20.717 < 0.0000000000000002
GarageArea      37.96414      3.52540  10.769 < 0.0000000000000002
age           -310.46367     23.56303 -13.176 < 0.0000000000000002

Residual standard error: 24300 on 2173 degrees of freedom
  (11 observations deleted due to missingness)
Multiple R-squared:  0.8597,    Adjusted R-squared:  0.8593
F-statistic:  1902 on 7 and 2173 DF,  p-value: < 0.00000000000000022
```

**Coefficient Interpretation:**

With every unit increase in the overall quality, the sale price of a residence in Ames increases by 15662.05 when adjusted for other predictors. With every unit increase in the above grade living area, the sale price of a residence in Ames increases by 52.191 when adjusted for other predictors. With every unit increase in the lot area, the sale price of a residence in Ames increases by 0.87 when adjusted for other predictors. With every unit increase in the mas veneer area, the sale price of a residence in Ames increases by 21.3 when adjusted for other predictors. With every unit increase in the total basement square feet, the sale price of a residence in Ames increases by 34.23 when adjusted for other predictors. With every unit increase in the garage area, the sale price of a residence in Ames increases by 37.96 when adjusted for other predictors. With every unit increase in the age, the sale price of a residence in Ames decreases by 310.46 when adjusted for other predictors.

**ANOVA table:**

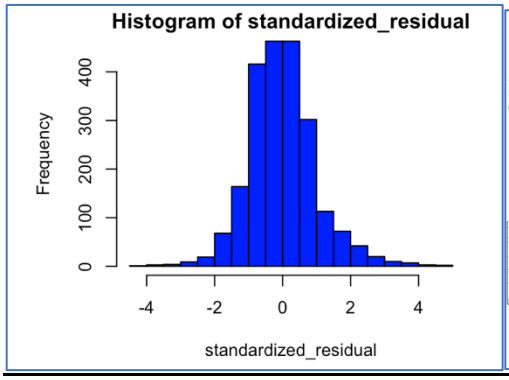
```
> anova(model13)
Analysis of Variance Table

Response: SalePrice
              Df        Sum Sq       Mean Sq F value                Pr(>F)
GrLivArea      1 5203746612103 5203746612103 8810.74 < 0.00000000000000022
OverallQual    1 1784341219557 1784341219557 3021.16 < 0.00000000000000022
LotArea        1  182893955577  182893955577  309.67 < 0.00000000000000022
MasVnrArea     1   90889463708   90889463708  153.89 < 0.00000000000000022
TotalBsmtSF    1  383477950536  383477950536  649.29 < 0.00000000000000022
GarageArea     1  117364089242  117364089242  198.72 < 0.00000000000000022
age            1  102532897410  102532897410  173.60 < 0.00000000000000022
Residuals   2173 1283404471200     590614115
```

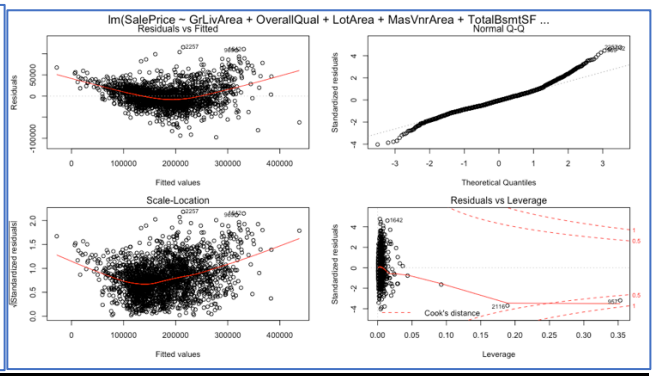
**R-square and adjusted r-square:**

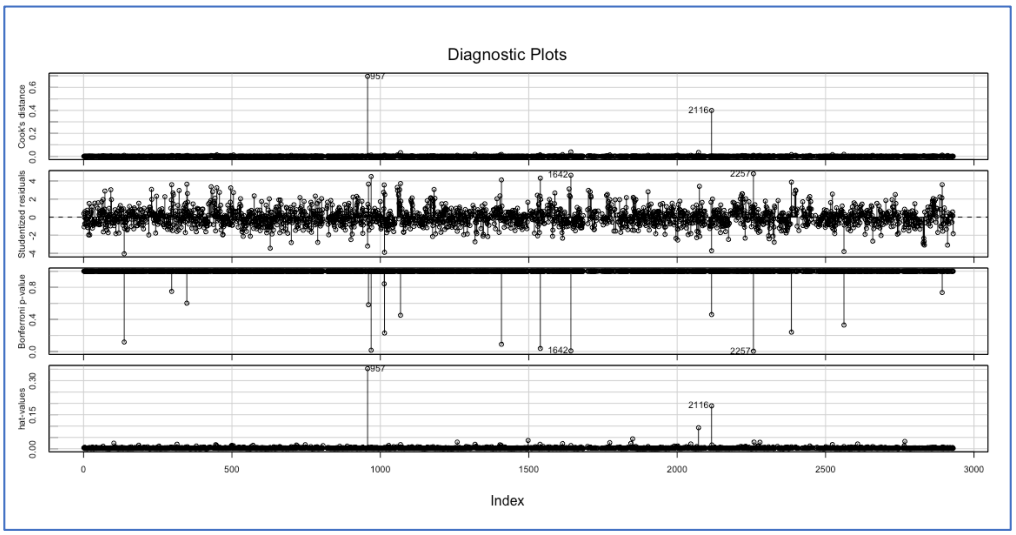
**R-square- 85.97%** of the variance in sale price is explained by these predictors.

Adjusted r-square – 85.93%

## Residual and Influence plots:




- Residuals follow a normal distribution. Evidence of heteroscedasticity in a few predictors.

## CONCLUSION:

In order to have reliable results in linear models, all the assumptions need to be satisfied. The violation of some assumptions may not affect the model as much however, the violation of assumptions like heteroscedasticity can adversely affect the goodness of a model. Hence, transformations can be applied to either the response variable or the predictors or both. It may

not be necessary to apply three different transformations for three assumption violations. Usually, one transformation helps with linearity, normality and homoscedasticity. Even if the interpretation of these models might be challenging, it is necessary to perform this. The next steps in the modeling process would be testing the model on a different dataset, evaluating it using a metric like RMSE and then deploying it (if it's going to be used for prediction purposes).