Reed Ballesteros
MSDS-410-DL, Summer 2022
Dr. Mickelson
7/17/2022

Modeling Assignment #4: Building Linear Regression Models – Diagnostics and Transformations

EDA

Similar to what was previously done in Modeling Assignment 3, we want to perform both data cleanup and waterfall dropdown.

Cleanup:

- Correct GarageCars with <NA> values to 0
- Correct MasVnrArea with <NA> values to 0
- Correct TotalBsmtSF with <NA> values to 0
- Correct TotRmsAbvGrd with <NA> values to 0
- Correct FullBath with <NA> values to 0

Waterfall dropdown:

- Narrow population to only single-family homes (BldgType = '1Fam')
- Remove GarageCars outlier with GarageCars = 5 (doesn't match with SalePrice)
- Remove LotArea outliers (3) with LotArea > 100000 sq ft (doesn't match with SalePrice)
- Remove TotRmsAbvGrd outlier with TotRmsAbvGrd = 15 sq ft (doesn't match with SalePrice)
- Remove FullBath outliers (7) with FullBath = 0 (extremely rare to find 0 bath in a single family home)

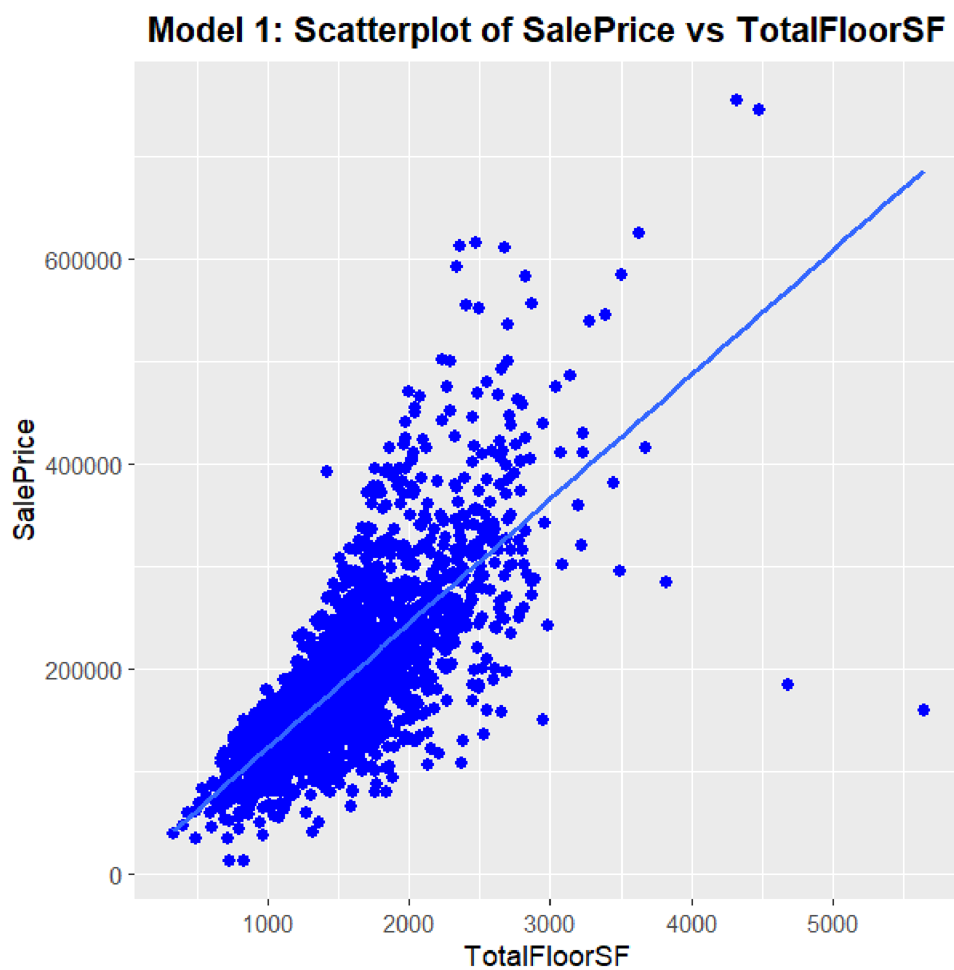
Assignment Tasks

1. Let Y = sale price be the dependent or response variable. Select what you consider to be "the best" continuous explanatory variable from the AMES data set to predict Y. Discuss what criteria you used to select this explanatory variable? Fit a simple linear regression model using your explanatory variable X to predict SALE PRICE(Y). Call this Model 1.

```
> model1

Call:
lm(formula = subdat$SalePrice ~ subdat$TotalFloorSF)

Coefficients:
        (Intercept)  subdat$TotalFloorSF
             1114.2                121.4
```

I selected TotalFloofSF as the 'best' continuous explanatory variable to use for the dependent variable SalePrice in the Ames dataset due their strong correlation between each other, as I will show in the following tasks below.

a. Make a scatterplot of Y and X, and overlay the regression line on the cloud of data.
The scatterplot of Y and X is the following:



b. Report the model in equation form and interpret each coefficient of the model in the context of this problem.
The simple linear regression equation for Model 1 is:

SalePrice = 1114.165 + 121.427*TotalFloorSF

The y-intercept is 1114.165. Model 1 predicts that an additional unit of TotalFloorSF will add $121.43 to the SalePrice of a single-family home.

c. Report and interpret the R-squared value in the context of this problem.

```
> summary(model1)

Call:
lm(formula = subdat$SalePrice ~ subdat$TotalFloorSF)

Residuals:
    Min      1Q  Median      3Q     Max 
-526208  -28227   -2013   21152  323488 

Coefficients:
                    Estimate Std. Error t value            Pr(>|t|)    
(Intercept)         1114.165   3404.253   0.327               0.743    
subdat$TotalFloorSF  121.427      2.132  56.959 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 53960 on 2412 degrees of freedom
Multiple R-squared:  0.5736,    Adjusted R-squared:  0.5734 
F-statistic:  3244 on 1 and 2412 DF,  p-value: < 0.00000000000000022
```

Given the R-based summary for Model 1, the R-squared value is 0.5736, in which the model indicates that TotalFloorSF in the model describes about 57.4% of the variability of SalePrice.

d. Report the coefficient and ANOVA Tables.

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

Response: subdat$SalePrice
                      Df        Sum Sq       Mean Sq F value                Pr(>F)    
subdat$TotalFloorSF    1 9447676578136 9447676578136  3244.3 < 0.00000000000000022 ***
Residuals           2412 7023898184788    2912063924                                  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The R-based ANOVA table is shown above, and the coefficient of the TotalFloorSF, as shown in the model equation and the summary table, is 121.427. The F-Value is a very high 3244.3 while the p-value is very close to 0.

e. Clearly specify the hypotheses associated with each coefficient of the model, as well as the hypothesis for the overall omnibus model. Conduct and interpret these hypothesis tests.
Hypothesis testing for Model 1 is defined as the following:

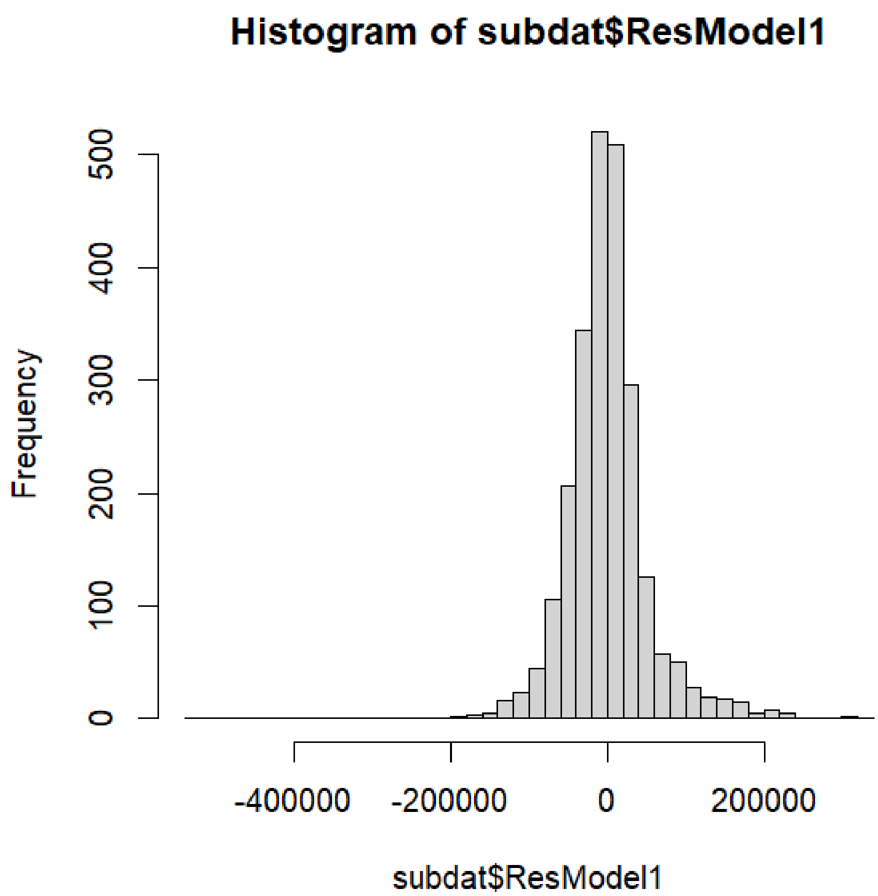
H_0: Beta1 = 0

H_A: Beta1 != 0

Given a 95% confidence interval and 2412 degrees of freedom, the critical T-Value is 1.96 (in R: `qt(0.05/2, 2412, lower.tail=FALSE)`). Based on the summary above, the T-Value for TotalFloorSF in Model 1 is 56.959, in which $T_{TotFloorSF} > T_{Critical}$. Also, The P-Value is very close to 0 and the F-Value is 3244.3, which higher than the calculated F-Critical value 0.0009822729 given the degrees of freedom of 1 and 2412 (in R: `qf(p=0.05/2, df1=1, df2=2412)`). Which these statistics so far, we can reject the null hypothesis H_0 and consider TotalFloorSF, and Model 1 with an overall F-statistic of 3244, as statistically significant.

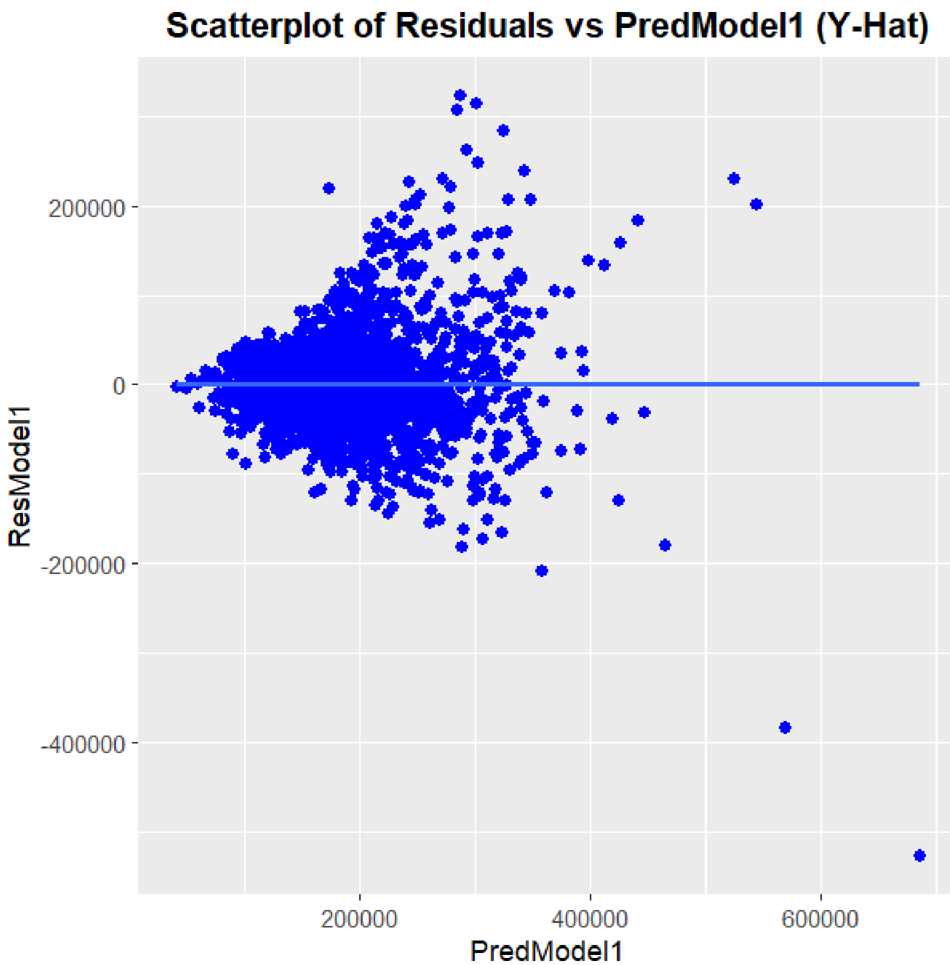
f. The validity of the hypothesis tests are dependent on the underlying assumptions of Independence, Normality, and Homoscedasticity being well met. Check on these underlying assumptions by plotting:

• Histogram of the standardized residuals



The histogram above shows that the residuals for Model 1 generally look normally distributed.

• **Scatterplot of standardized residuals (Y) by predicted values (Y_hat)**



As we can see in the scatterplot of residuals vs. predicted values in Model 1, the residuals 'fan-out' as the predicted price increases, which indicates heteroscedasticity, which means that the variance is not constant throughout the predicted values.

Discuss any deviations from normality or patterns in the residuals that indicate heteroscedasticity.

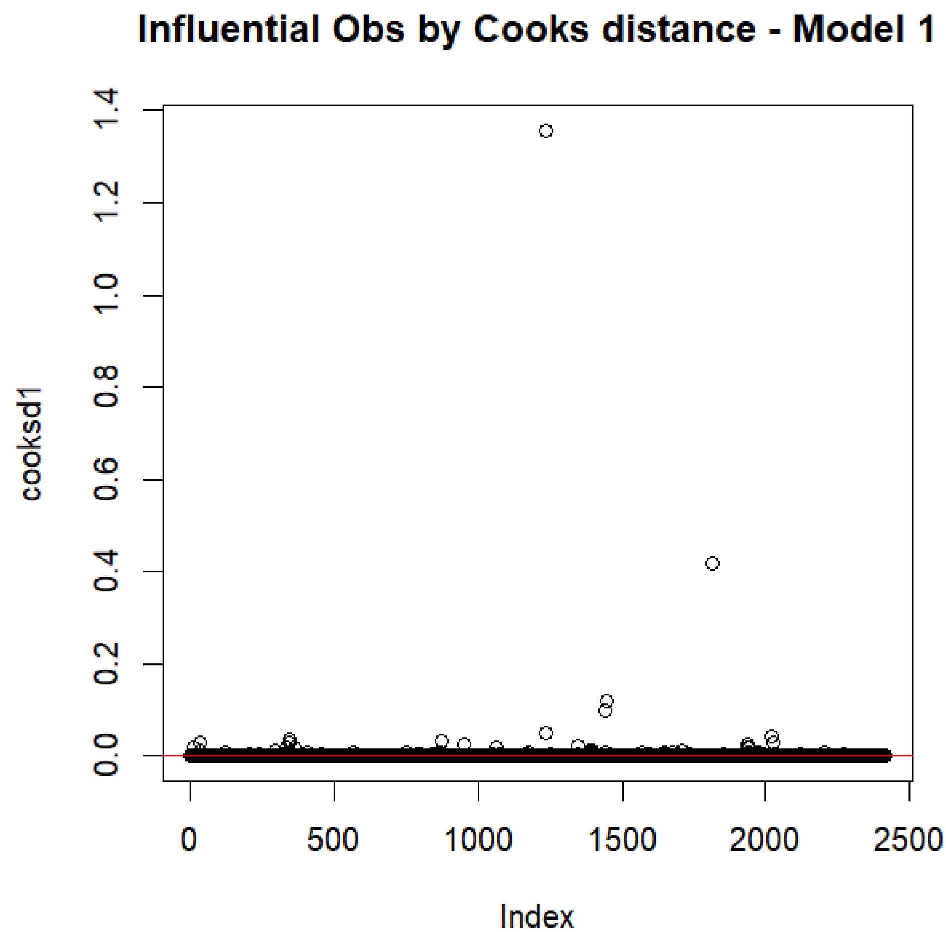
Evaluating the global validation of linear model assumptions (GVLMA) gives us the following:

```
> gvmodel1 <- gvlma(model1)
> summary(gvmodel1)

Call:
lm(formula = subdat$SalePrice ~ subdat$TotalFloorSF)

Residuals:
    Min      1Q  Median      3Q     Max 
-526208  -28227   -2013   21152  323488 

Coefficients:
                    Estimate Std. Error t value            Pr(>|t|)    
(Intercept)         1114.165   3404.253   0.327               0.743    
subdat$TotalFloorSF  121.427      2.132  56.959 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 53960 on 2412 degrees of freedom
Multiple R-squared:  0.5736,	Adjusted R-squared:  0.5734 
F-statistic:  3244 on 1 and 2412 DF,  p-value: < 0.00000000000000022


ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05 

Call:
 gvlma(x = model1) 

                     Value    p-value                   Decision
Global Stat        7468.60 0.00000000 Assumptions NOT satisfied!
Skewness             88.03 0.00000000 Assumptions NOT satisfied!
Kurtosis           7351.85 0.00000000 Assumptions NOT satisfied!
Link Function        15.82 0.00006956 Assumptions NOT satisfied!
Heteroscedasticity   12.90 0.00032850 Assumptions NOT satisfied!
```

The high evidence of heteroscedasticity in both the scatterplot and the GVLMA summary above indicate that Model 1 violates the assumptions of linear regression modeling, such that modeling errors should be generally constant, in which the scatterplot shows that is it not so.

g. Check on leverage, influence, and outliers. These points can be identified by several statistics such as DFFITS, Cook’s Distance, Leverage, and Influence. Discuss any issues or concerns. Describe what course of action should be taken.

```
> cooksd1 <- cooks.distance(model1)
> plot(cooksd1,main="Influential Obs by Cooks distance - Model 1")  # plot cook's distance
> abline(h = 4/nrow(subdat), col="red")  # add cutoff line
```



Using Cook's distance, we find there are 143 data observations that can be a potential influence on the dataset:

```
> influential1 <- as.numeric(names(cooksd1)[(cooksd1 > (4/nrow(subdat)))])
> influential1
  [1]   12   14   25   27   33   34   93  122  131  169  203  237  261  278  293  294  303  318
 [19]  331  332  333  337  339  340  341  342  343  344  346  357  358  361  362  388  397  406
 [37]  414  457  465  569  580  662  750  792  801  833  842  858  859  860  861  863  864  865
 [55]  867  868  871  872  875  951  959 1049 1065 1070 1074 1078 1167 1173 1174 1237 1238 1254
 [73] 1295 1310 1343 1345 1346 1347 1348 1379 1385 1386 1388 1390 1391 1394 1395 1396 1401 1440
 [91] 1441 1443 1447 1460 1510 1568 1597 1607 1644 1646 1649 1675 1684 1691 1708 1737 1739 1815
[109] 1840 1853 1894 1897 1935 1936 1938 1940 1941 1942 1943 1945 1954 1970 1973 1975 1986 1988
[127] 1989 1990 1991 2018 2019 2023 2029 2034 2079 2110 2118 2196 2202 2220 2273 2402 2403
> length(influential1)
[1] 143
```

Removing these 143 data points can reduce the dataset to 2271 observations, but doing so can potentially have an impact on improving heteroscedasticity.

2. For Task 2, you will fit a multiple regression model that uses 2 continuous explanatory (X) variables to predict Sale Price (Y). Call this Model 2. The explanatory variables for Model 2

should be the explanatory variable you had in Model 1, plus the OVERALL QUALITY variable. To report the results for Model 2, you are to:

a. Report the prediction equation and interpret each coefficient of the model in the context of this problem. Is there something different about the coefficient interpretations here relative to the simple linear regression model in Task 1?

```
> model2

Call:
lm(formula = subdat$SalePrice ~ subdat$TotalFloorSF + subdat$OverallQual)

Coefficients:
        (Intercept)  subdat$TotalFloorSF   subdat$OverallQual
         -109684.51                63.46             32566.61
```

The multiple linear regression equation for Model 2 is:

SalePrice = -109684.61 + 63.46*TotalFloorSF + 32566.61*OverallQual

The y-intercept is -109684.51. Model 2 predicts that an additional unit of TotalFloorSF will add $63.46 to the SalePrice of a single-family home, and a single rating increase of OverallQual can add $32566.61 to the SalePrice.

In this scenario, the y-intercept in Model 2 is a very large negative number, TotalFloorSF's slope is just over half its value compared to Model 1, and OverallQual has a very large slope that can greatly affect the prediction of SalePrice, given that its range is only in a small range of small numbers between 1 and 10.

b. Report and interpret the R-squared value in the context of this problem. Calculate and report the difference in R-squared between Model 2 and Model 1. Interpret this difference.

```
> summary(model2)

Call:
lm(formula = subdat$SalePrice ~ subdat$TotalFloorSF + subdat$OverallQual)

Residuals:
    Min      1Q  Median      3Q     Max 
-413997  -23327   -1416   19408  278234 

Coefficients:
                      Estimate  Std. Error t value            Pr(>|t|)    
(Intercept)        -109684.506    3706.197  -29.59 <0.0000000000000002 ***
subdat$TotalFloorSF      63.455       2.134   29.73 <0.0000000000000002 ***
subdat$OverallQual    32566.606     778.613   41.83 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 41090 on 2411 degrees of freedom
Multiple R-squared:  0.7529,    Adjusted R-squared:  0.7527 
F-statistic:  3673 on 2 and 2411 DF,  p-value: < 0.00000000000000022
```

Given the R-based summary for Model 2, the R-squared value is 0.7529, in which the model indicates that TotalFloorSF along with OverallQual in the model describes about 75.3% of the variability of SalePrice.

c. Report the coefficient and ANOVA Tables.

```
> anova(model2)
Analysis of Variance Table

Response: subdat$SalePrice
                      Df        Sum Sq       Mean Sq F value                Pr(>F)
subdat$TotalFloorSF    1 9447676578136 9447676578136  5596.1 < 0.0000000000000022 ***
subdat$OverallQual     1 2953515962617 2953515962617  1749.4 < 0.0000000000000022 ***
Residuals           2411 4070382222171    1688254758
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The R-based ANOVA table is shown above, and the coefficient of the TotalFloorSF, as shown in the model equation and the summary table, is 63.455, with an F-Value of 5596.1 and a P-Value very close to 0. The coefficient of OverallQual is 32566.606 with an F-Value of 1749.4 and a P-Value very close to 0 as well.

d. Specify the hypotheses associated with each coefficient of the model and the hypothesis for the overall omnibus model. Conduct and interpret these hypothesis tests.

Hypothesis testing for Model 2 is defined as the following:

TotalFloorSF:

H_0: Beta1 = 0

H_A: Beta1 != 0

OverallQual:

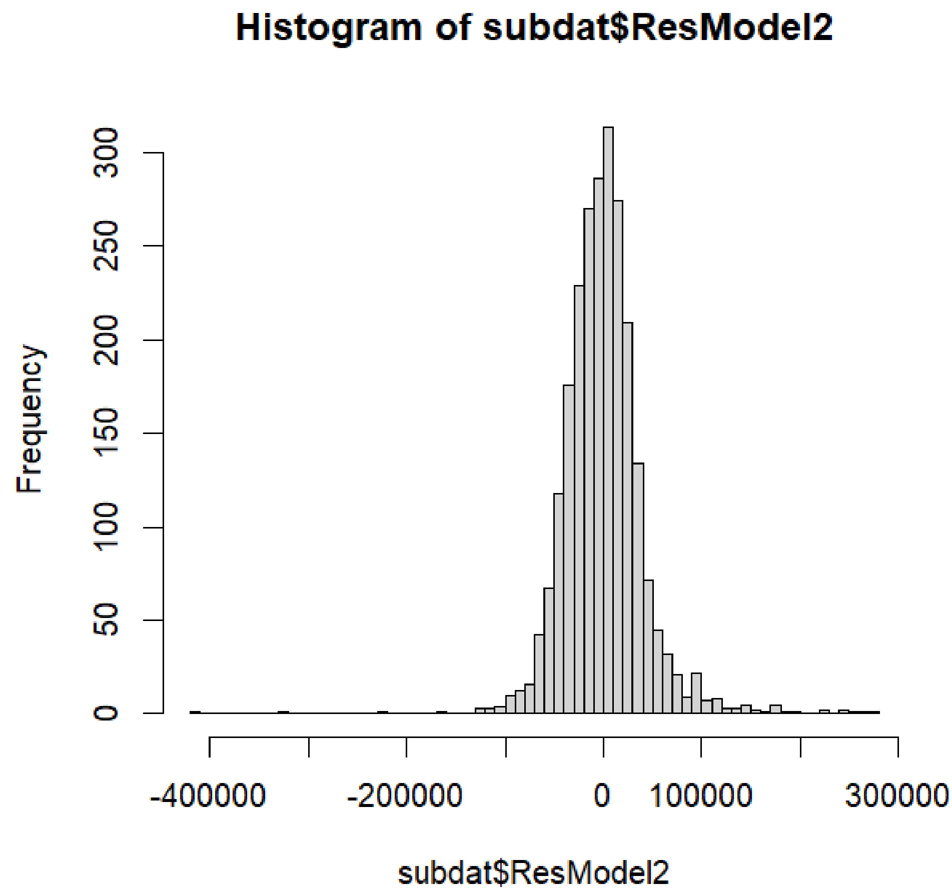
H_0: Beta2 = 0

H_A: Beta2 != 0

Given a 95% confidence interval and 2411 degrees of freedom, the critical T-Value is 1.96 (in R: `qt(0.05/2, 2411, lower.tail=FALSE)`). Based on the summary above, the T-Value for TotalFloorSF in Model 1 is 29.73, in which $T_{TotFloorSF} > T_{Critical}$. Also, The P-Value is very close to 0 and the F-Value is 5596.1, which is higher than the calculated F-Critical value 0.02531807 given the degrees of freedom of 2 and 2411 (in R: `qf(p=0.05/2, df1=2, df2=2411)`). The T-Value for OverallQual in Model 2 is 41.83, in which $T_{OverallQual} > T_{Critical}$. Also, The P-Value is very close to 0 and the F-Value is 1749.4, which is higher than the calculated F-Critical value 0.02531807. Which these statistics so far, we can reject both null hypotheses H_0 and consider TotalFloorSF, OverallQual, and Model 2 with an overall F-Statistic of 3673, as statistically significant.

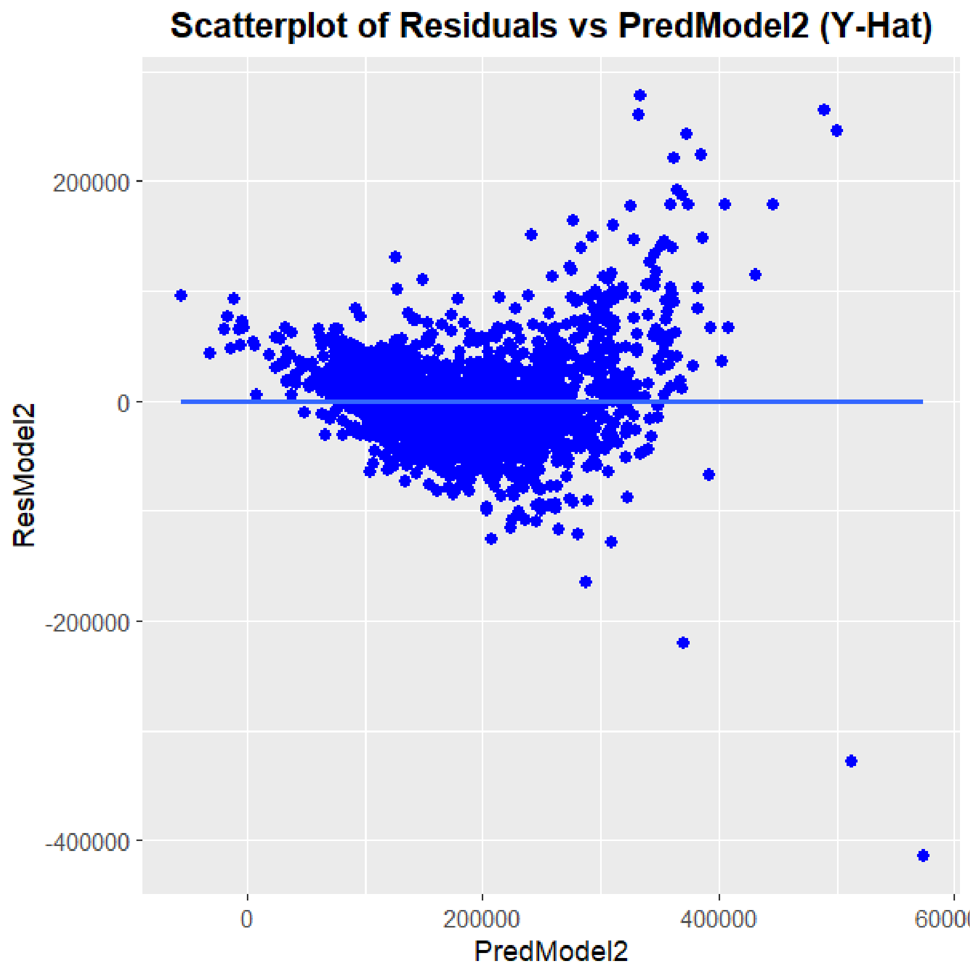
e. The validity of the hypothesis tests are dependent on the underlying assumptions of Independence, Normality, and Homoscedasticity being well met. Check on these underlying assumptions by plotting:

- **Histogram of the standardized residuals**



We see a generally normal distribution of residuals in Model 2.

• **Scatterplot of standardized residuals (Y) by predicted values (Y_hat)**



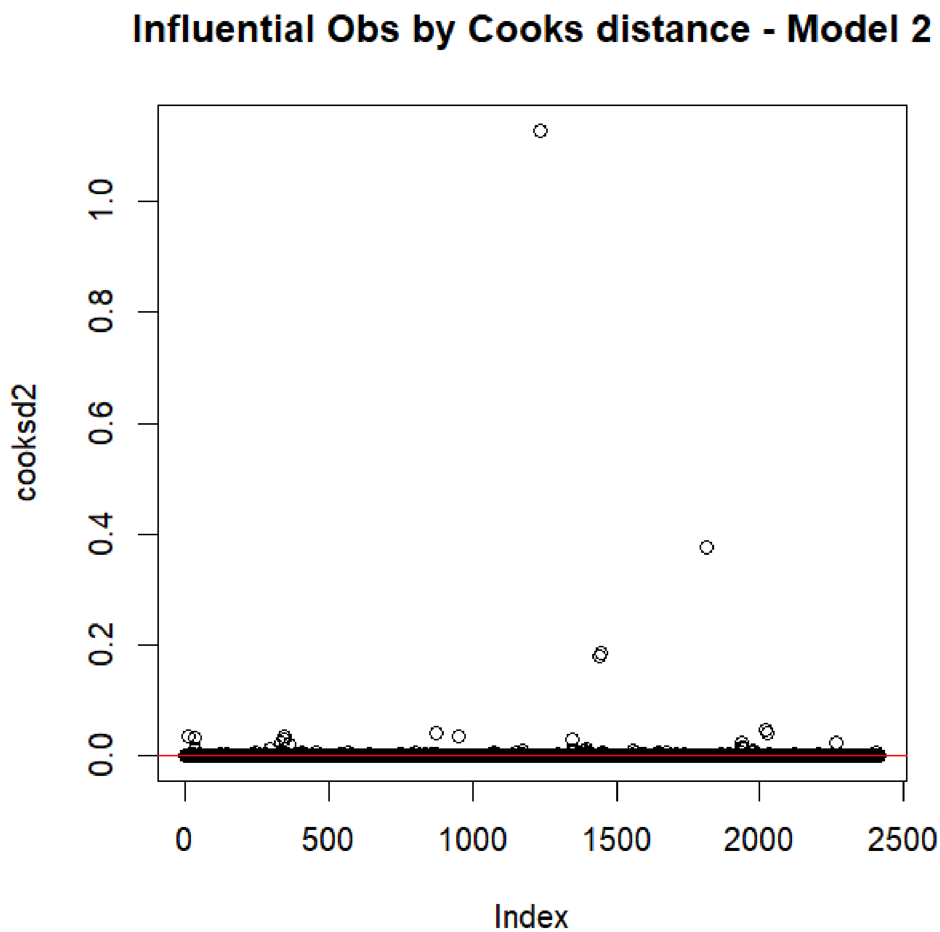
As we can see in the scatterplot of residuals vs. predicted values in Model 2, the residuals still 'fan-out' as the predicted price increases, which indicates heteroscedasticity, which means that the variance is not constant throughout the predicted values. The residuals show a more pronounced upward curve as well.

Discuss any deviations from normality or patterns in the residuals that indicate heteroscedasticity.

```
> gvmodel2 <- gvlma(model2)
> summary(gvmodel2)

Call:
lm(formula = subdat$SalePrice ~ subdat$TotalFloorSF + subdat$OverallQual)

Residuals:
    Min      1Q  Median      3Q     Max
-413997  -23327   -1416   19408  278234

Coefficients:
                      Estimate  Std. Error t value           Pr(>|t|)
(Intercept)        -109684.506    3706.197  -29.59 <0.0000000000000002 ***
subdat$TotalFloorSF     63.455       2.134   29.73 <0.0000000000000002 ***
subdat$OverallQual   32566.606     778.613   41.83 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 41090 on 2411 degrees of freedom
Multiple R-squared:  0.7529,    Adjusted R-squared:  0.7527
F-statistic:  3673 on 2 and 2411 DF,  p-value: < 0.00000000000000022


ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = model2)

                       Value  p-value                   Decision
Global Stat        13002.475 0.000000 Assumptions NOT satisfied!
Skewness              76.445 0.000000 Assumptions NOT satisfied!
Kurtosis           12648.914 0.000000 Assumptions NOT satisfied!
Link Function        270.013 0.000000 Assumptions NOT satisfied!
Heteroscedasticity     7.102 0.007698 Assumptions NOT satisfied!
```

The high evidence of heteroscedasticity in both the scatterplot and the GVLMA summary above indicate that Model 2 still violates the assumptions of linear regression modeling, such that modeling errors should be generally constant, in which the scatterplot shows that is it not so. While the addition of OverallQual in Model 2 lowers the GVLMA heteroscedasticity score, it's still not enough.

f. Check on leverage, influence and outliers, and discuss any issues or concerns.



Using Cook's distance, we find there are 144 data observations that can be a potential influence on the dataset:

```
> influential2 <- as.numeric(names(cooksd2)[(cooksd2 > (4/nrow(subdat)))])
> influential2
  [1]   12   13   14   27   33   34   48  122  131  145  237  244  261  278  293  294  299  331
 [19]  332  333  339  340  341  342  343  344  346  358  361  362  388  397  406  414  457  530
 [37]  538  569  574  580  584  607  637  643  670  748  750  780  792  801  833  858  859  861
 [55]  863  864  865  867  871  872  875  899  951 1061 1065 1070 1074 1076 1078 1112 1153 1158
 [73] 1173 1174 1236 1238 1283 1294 1310 1343 1345 1346 1347 1348 1379 1385 1386 1388 1390 1391
 [91] 1394 1395 1396 1401 1440 1441 1443 1447 1452 1460 1462 1556 1568 1597 1632 1644 1646 1649
[109] 1675 1708 1736 1737 1815 1842 1865 1894 1935 1936 1938 1940 1945 1970 1973 1975 1988 1989
[127] 1990 1991 2018 2023 2028 2029 2118 2140 2181 2192 2196 2202 2263 2355 2379 2402 2403 2404
> length(influential2)
[1] 144
```

Removing these 144 data points can reduce the dataset to 2270 observations, but like Model 1, doing so can potentially have an impact on improving heteroscedasticity.

g. Based on the information, should you want to retain both variables as predictor variables of Y? Discuss why or why not.

While the T-Test, F-Test, and P-Values show that TotalFloorSF and OverallQual are two explanatory variables that are statistically significant, the Model 1's and Model 2's heteroscedasticity show a different story in that their respective residuals are not constant. Because of this added insight of heteroscedasticity, we should consider other continuous variables in the Ames dataset instead, or see if adding another variable can improve things (which we will see in section 3).

3. Select any other continuous explanatory variable you wish. Fit a multiple regression model that uses 3 continuous explanatory (X) variables to predict Sale Price (Y). These three variables should be the explanatory variables from Model 2 plus your choice of an additional explanatory variable. Call this Model 3. To report the results for Model 3, you are to:

a. Report Model 3 in equation form and interpret each coefficient of the model in the context of this problem. Is there something different about the coefficient interpretations here to Models 1 and 2?

```
> model3

Call:
lm(formula = subdat$SalePrice ~ subdat$TotalFloorSF + subdat$OverallQual +
    subdat$LotArea)

Coefficients:
        (Intercept)  subdat$TotalFloorSF   subdat$OverallQual       subdat$LotArea
        -119812.876                56.156            33131.628                1.652
```

The multiple linear regression equation for Model 3 is:

SalePrice = -119812.876 + 56.156*TotalFloorSF + 33131.628*OverallQual + 1.652*LogArea

The y-intercept is -119812.876. Model 3 predicts that an additional unit of TotalFloorSF will add $ 56.16 to the SalePrice of a single-family home, and a single rating increase of OverallQual can add $ 33131.63, and a, extra square foot in LotArea can add $1.66.

Compared to Model 2, Model 3' y-intercept is a larger negative value while TotalFloorSF's impact is slightly lessened and OverallQual's impact is slightly increased. LogArea, though, doesn't seem to have a much of an impact compared to the other explanatory variables is a very large negative number, TotalFloorSF's slope is just over half its value compared to Model 1, and OverallQual has a very large slope that can greatly affect the prediction of SalePrice, given that its range is only in a small range of small numbers between 1 and 10.

b. Report and interpret R-squared value in the context of this problem. Calculate the difference in R-squared between Model 3 and Model 2. How would you interpret this difference? Does your variable of choice help to improve the model's explanatory ability?

```
> summary(model3)

Call:
lm(formula = subdat$SalePrice ~ subdat$TotalFloorSF + subdat$OverallQual +
    subdat$LotArea)

Residuals:
    Min      1Q  Median      3Q     Max
-473892  -22163   -1335   18827  279188

Coefficients:
                      Estimate  Std. Error t value           Pr(>|t|)
(Intercept)        -119812.876    3799.963 -31.530 <0.0000000000000002 ***
subdat$TotalFloorSF     56.156       2.238  25.087 <0.0000000000000002 ***
subdat$OverallQual   33131.628     767.461  43.170 <0.0000000000000002 ***
subdat$LotArea           1.652       0.177   9.334 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 40370 on 2410 degrees of freedom
Multiple R-squared:  0.7615,    Adjusted R-squared:  0.7612
F-statistic:  2565 on 3 and 2410 DF,  p-value: < 0.00000000000000022
```

Given the R-based summary for Model 2, the R-squared value is 0.7615, in which the model indicates that the combination of TotalFloorSF, OverallQual, and LotARea in the model describes about 76.2% of the variability of SalePrice. Compared to Model 2, the increase in R-squared is marginal compared to the increase of R-squared between Model 1 and Model 2.

c. Report the coefficient and ANOVA Tables for Model 3.

```
> anova(model3)
Analysis of Variance Table

Response: subdat$SalePrice
                      Df        Sum Sq       Mean Sq  F value                Pr(>F)
subdat$TotalFloorSF    1 9447676578136 9447676578136 5796.006 < 0.00000000000000022 ***
subdat$OverallQual     1 2953515962617 2953515962617 1811.937 < 0.00000000000000022 ***
subdat$LotArea         1  142004338269  142004338269   87.118 < 0.00000000000000022 ***
Residuals           2410 3928377883902    1630032317
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The R-based ANOVA table is shown above, and the coefficient of the TotalFloorSF, as shown in the model equation and the summary table, is 56.156, with an F-Value of 5796.006 and a P-Value very close to 0. The coefficient of OverallQual is 33131.628 with an F-Value of 1811.937 and a P-Value very close to 0. The coefficient of LotArea is 1.652 with an F-Value of 87.118 and a P-Value very close to 0.

d. Specify the hypotheses associated with each coefficient of the model and the hypothesis for the omnibus model. Conduct and interpret these hypothesis tests.

Hypothesis testing for Model 3 is defined as the following:

TotalFloorSF:

H_0: Beta1 = 0

H_A: Beta1 != 0

OverallQual:

H_0: Beta2 = 0

H_A: Beta2 != 0

LotArea:

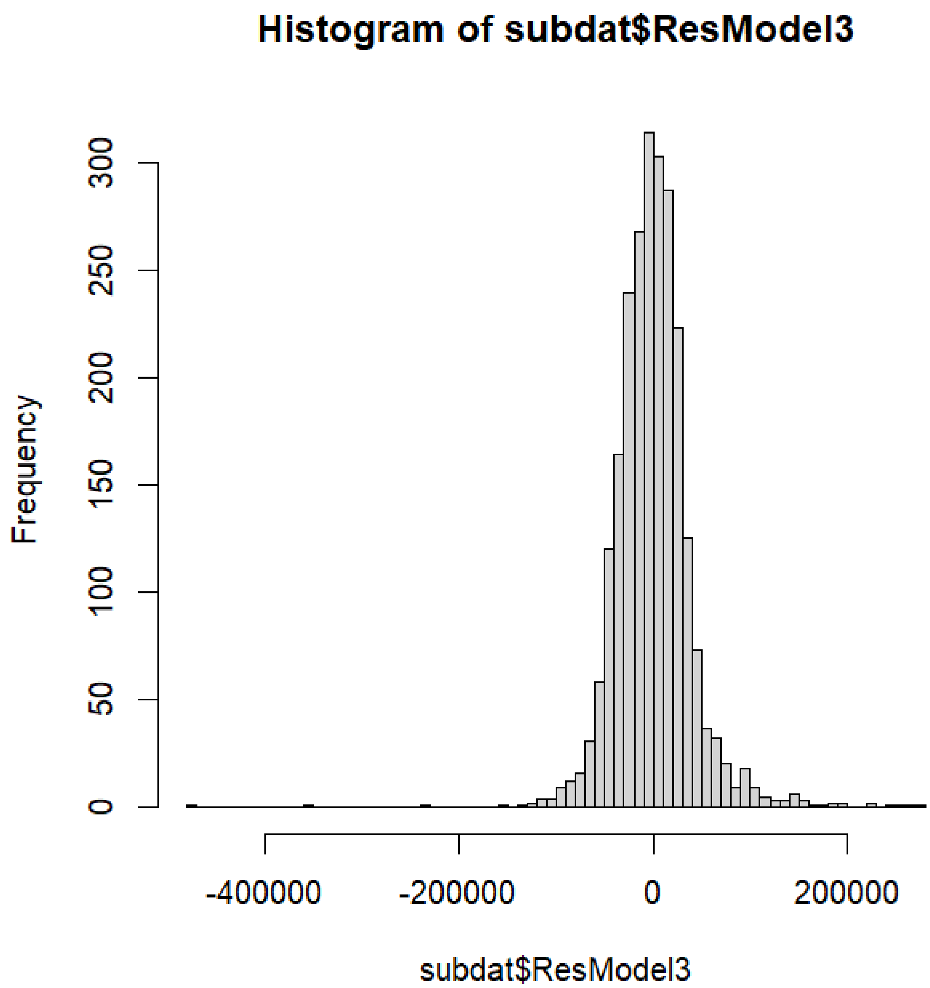
H_0: Beta3 = 0

H_A: Beta3 != 0

Given a 95% confidence interval and 2410 degrees of freedom, the critical T-Value is 1.96 (in R: `qt(0.05/2, 2410, lower.tail=FALSE)`). Based on the summary above, the T-Value for TotalFloorSF in Model 1 is 25.087, in which $T_{TotFloorSF} > T_{Critical}$. Also, The P-Value is very close to 0 and the F-Value is 5796.006, which is higher than the calculated F-Critical value 0.07192006 given the degrees of freedom of 3 and 2410 (in R: `qf(p=0.05/2, df1=3, df2=2410)`). The T-Value for OverallQual in Model 3 is 43.170, in which $T_{OverallQual} > T_{Critical}$. Also, The P-Value is very close to 0 and the F-Value is 1749.4, which is higher than the calculated $F_{Critical}$. The T-Value for LotArea is 9.334, where $T_{LotArea} > T_{Critical}$. Also, The P-Value is very close to 0 and the F-Value is 87.118 in which $F_{LotArea} > F_{Critical}$.

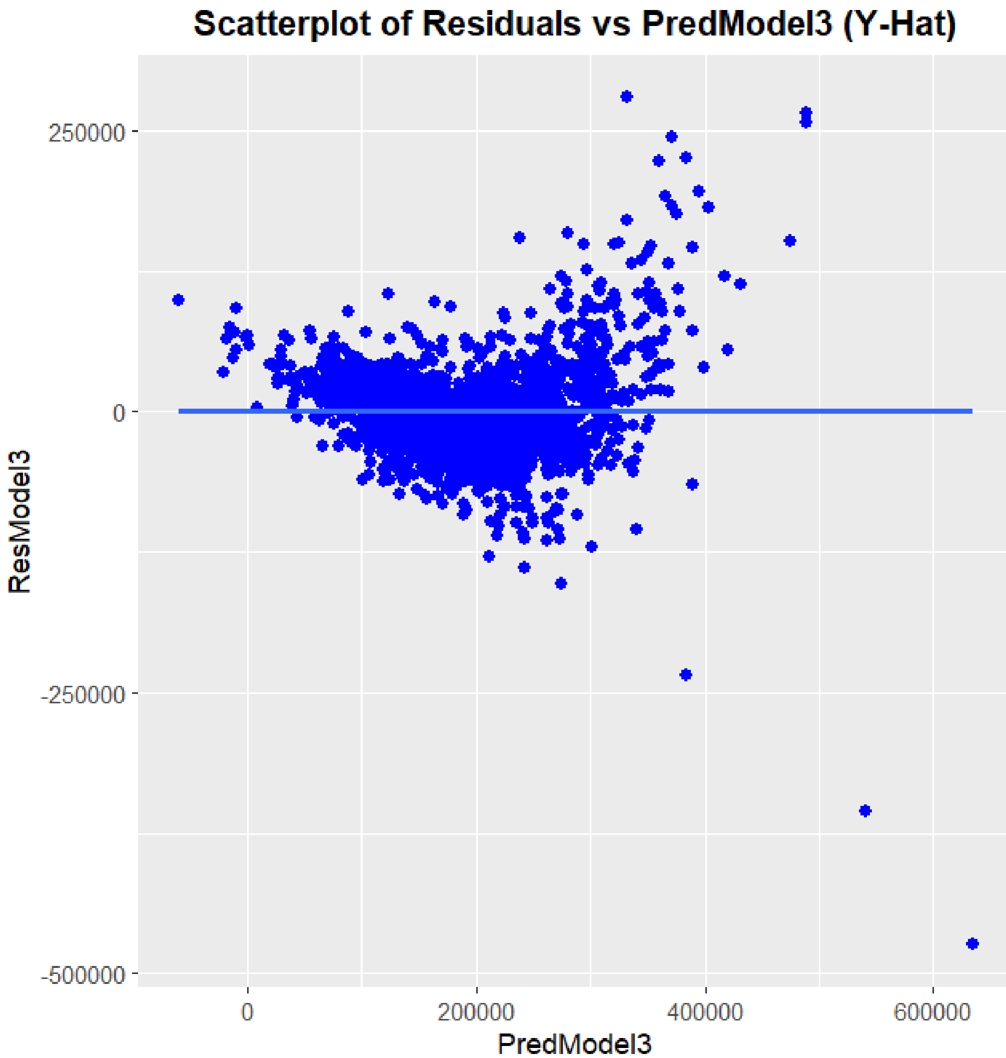
Which these statistics so far, we can reject the null hypotheses H_0 for each variableand consider TotalFloorSF, OverallQual, LotARea, and Model 3 with an overall F-Statistic of 2565, as statistically significant.

e. Check on the underlying assumptions. Discuss any deviations from normality or patterns in the residuals that indicate heteroscedasticity.



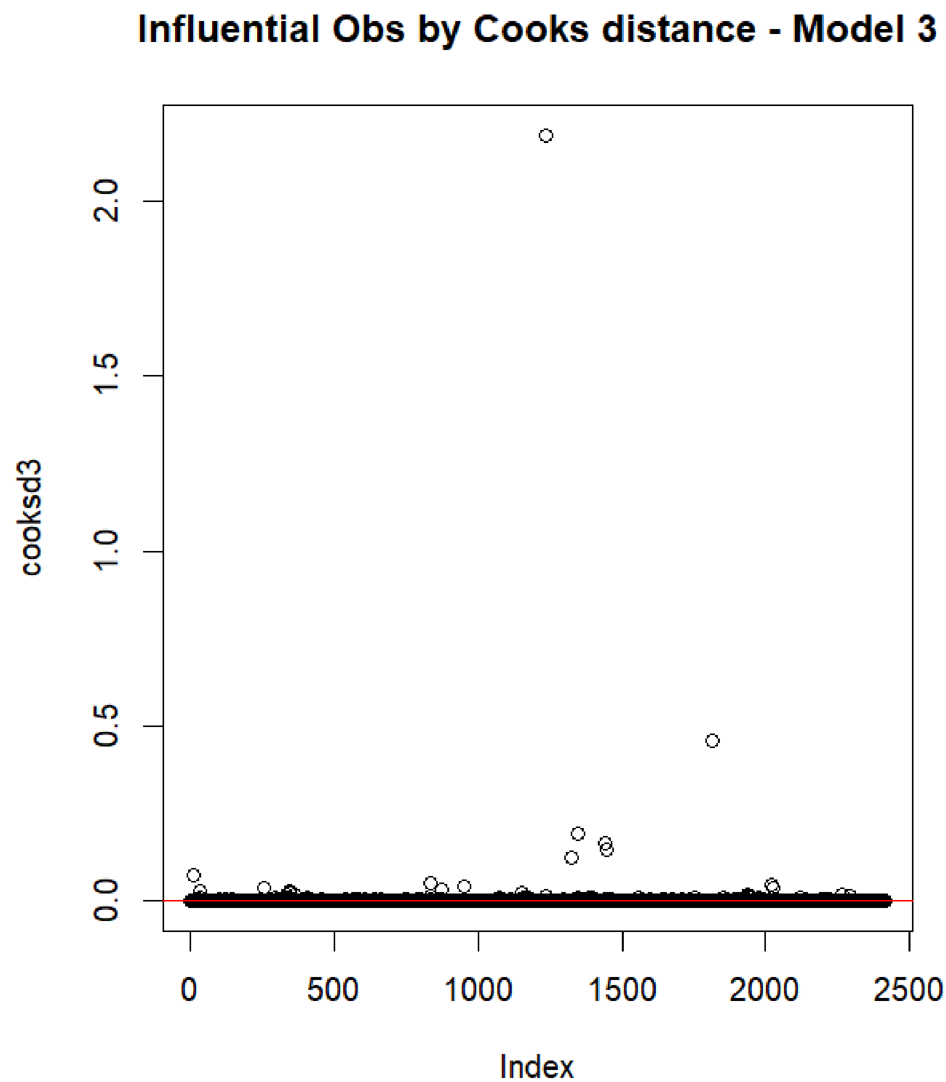
We see a generally normal distribution of residuals in Model 3.

• **Scatterplot of standardized residuals (Y) by predicted values (Y_hat)**



As we can see in the scatterplot of residuals vs. predicted values in Model 3, the residuals still ‘fan-out’ as the predicted price increases, which indicates heteroscedasticity, which means that the variance is not constant throughout the predicted values.

f. Check on leverage, influence and outliers, and discuss any issues or concerns.



Using Cook’s distance, we find there are 145 data observations that can be a potential influence on the dataset:

```
> influential3 <- as.numeric(names(cooksd3)[(cooksd3 > (4/nrow(subdat)))])
> influential3
  [1]   12   13   14   27   33   34   48  101  120  122  145  191  237  244  254  278  293  299
 [19]  331  332  333  339  340  341  342  343  344  346  358  361  362  388  397  406  414  457
 [37]  530  538  569  574  580  584  607  616  637  662  748  750  792  801  803  833  834  835
 [55]  858  859  861  863  864  865  867  871  875  899  951 1061 1065 1074 1076 1078 1101 1112
 [73] 1153 1157 1158 1173 1174 1236 1237 1238 1294 1323 1343 1345 1346 1347 1348 1385 1386 1388
 [91] 1390 1391 1394 1395 1396 1401 1440 1441 1443 1447 1452 1460 1462 1556 1568 1597 1632 1644
[109] 1646 1649 1675 1708 1709 1736 1737 1755 1815 1842 1853 1865 1894 1897 1906 1916 1935 1936
[127] 1938 1940 1945 1973 1975 1988 1989 1990 1991 2018 2019 2023 2029 2118 2136 2181 2192 2196
[145] 2202 2220 2263 2291 2379 2404
> length(influential3)
[1] 150
```

Removing these 150 data points, like with the previous models, can potentially have an impact on improving heteroscedasticity.

g. Based on this information, should you want to retain all three variables as predictor variables of Y? Discuss why or why not.

Evaluating the GVLMA of Model 3:

```
> summary(gvmodel3)

Call:
lm(formula = subdat$SalePrice ~ subdat$TotalFloorSF + subdat$OverallQual +
    subdat$LotArea)

Residuals:
    Min      1Q  Median      3Q     Max
-473892  -22163   -1335   18827  279188

Coefficients:
                      Estimate  Std. Error t value           Pr(>|t|)
(Intercept)        -119812.876    3799.963 -31.530 <0.0000000000000002 ***
subdat$TotalFloorSF     56.156       2.238  25.087 <0.0000000000000002 ***
subdat$OverallQual   33131.628     767.461  43.170 <0.0000000000000002 ***
subdat$LotArea           1.652       0.177   9.334 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 40370 on 2410 degrees of freedom
Multiple R-squared:  0.7615,    Adjusted R-squared:  0.7612
F-statistic:  2565 on 3 and 2410 DF,  p-value: < 0.00000000000000022


ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = model3)

                      Value  p-value                   Decision
Global Stat       24548.0676 0.000000 Assumptions NOT satisfied!
Skewness              0.4913 0.483338    Assumptions acceptable.
Kurtosis          24356.4215 0.000000 Assumptions NOT satisfied!
Link Function       183.9270 0.000000 Assumptions NOT satisfied!
Heteroscedasticity    7.2277 0.007179 Assumptions NOT satisfied!
```

The addition of LotArea to the regression model doesn’t really improve the heteroscedasticity, which makes us question the validity of using TotalFloorSF, OverallQual, and LotArea together in a linear regression model.

4. Refit Model 3 using the Natural Log of SALEPRICE as the response variable. Call this Model 4. This is LOG base e, or LN() on your calculator. You’ll have to find the appropriate function

using R. Perform an analysis of goodness-of-fit to compare the Natural Log of SALEPRICE model, Model 4, to the original Model 3. Does the transformed model fit better? Provide evidence in your discussion. Discuss if the improvement of model fit justifies the use of the transformed response variable, Log(SALEPRRICE).

```
> summary(model4)

Call:
lm(formula = subdat$LogSalePrice ~ subdat$TotalFloorSF + subdat$OverallQual +
    subdat$LotArea)

Residuals:
     Min       1Q   Median       3Q      Max
-2.21758 -0.08933  0.01907  0.11520  0.63841

Coefficients:
                      Estimate    Std. Error t value           Pr(>|t|)
(Intercept)         10.4763381614  0.0180943643 578.983 <0.0000000000000002 ***
subdat$TotalFloorSF  0.0002509380  0.0000106586  23.543 <0.0000000000000002 ***
subdat$OverallQual   0.1804970765  0.0036544341  49.391 <0.0000000000000002 ***
subdat$LotArea       0.0000078797  0.0000008429   9.348 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1922 on 2410 degrees of freedom
Multiple R-squared:  0.7842,    Adjusted R-squared:  0.7839
F-statistic:  2919 on 3 and 2410 DF,  p-value: < 0.00000000000000022
```

The multiple linear regression equation for Model 4 is:

LogSalePrice = 10.4763 + 0.00025*TotalFloorSF + 0.1805*OverallQual + 0.000008*LogArea

The y-intercept is 10.4763. Model 4 predicts that an additional square foor of TotalFloorSF will add 0.00025 to the LogSalePrice, a single rating increase of OverallQual can add 0.1805 to the LogSalePrice, and an additional square foot of LotArea can increase LogSalePrice by 0.000008.

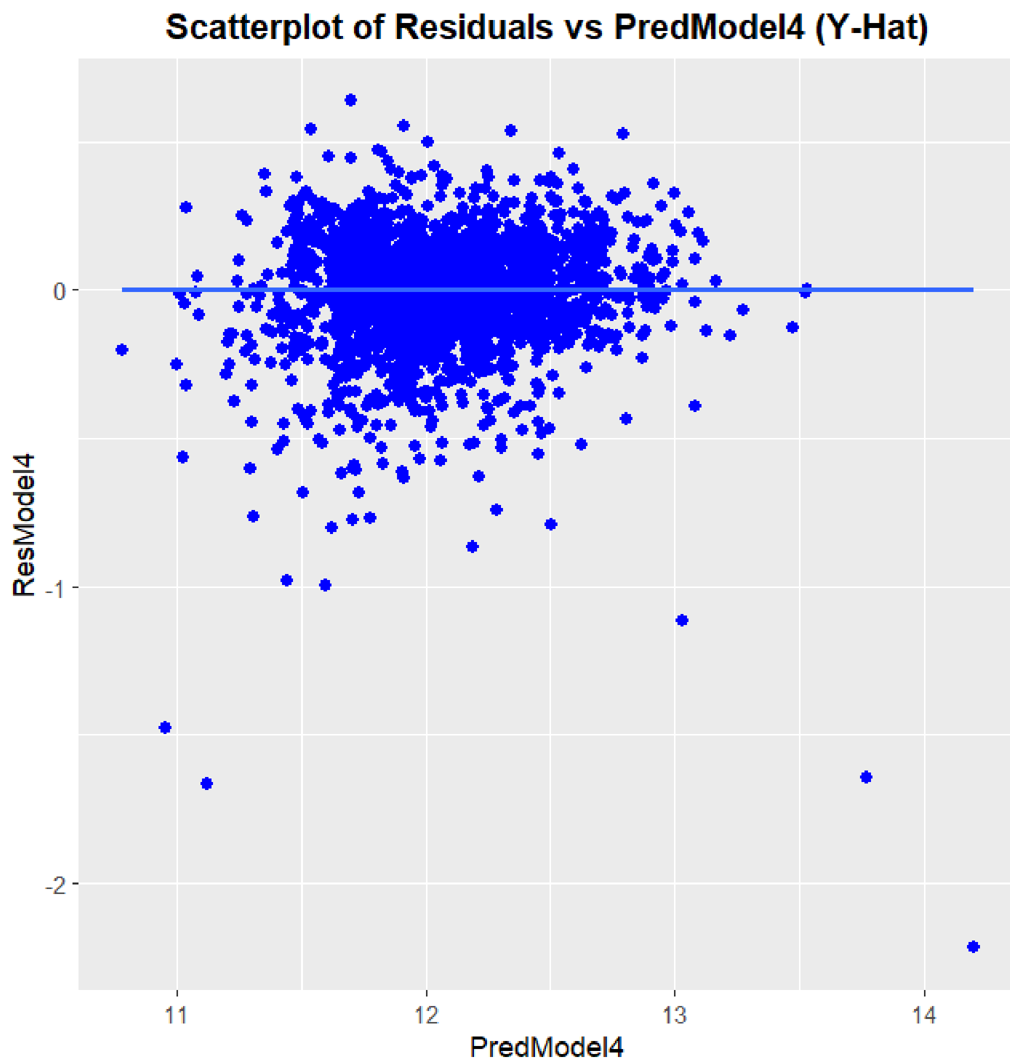
Model 4’s coefficient is different from the other models since we’re dealing with the log transformation of SalePrice.

Along with the ANOVA table:

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

Response: subdat$LogSalePrice
                      Df  Sum Sq Mean Sq  F value                Pr(>F)
subdat$TotalFloorSF    1 232.411 232.411 6288.303 < 0.00000000000000022 ***
subdat$OverallQual     1  88.038  88.038 2382.023 < 0.00000000000000022 ***
subdat$LotArea         1   3.230   3.230   87.384 < 0.00000000000000022 ***
Residuals           2410  89.072   0.037
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We can infer that the respective T-Values and F-Values of each of the exploratory variables are larger than the calculated T-Critical value of 1.96 and F-Critical value of 0.0719, such that they can be considered statistically significant.

The heteroscedasticity graphically shows a different story:

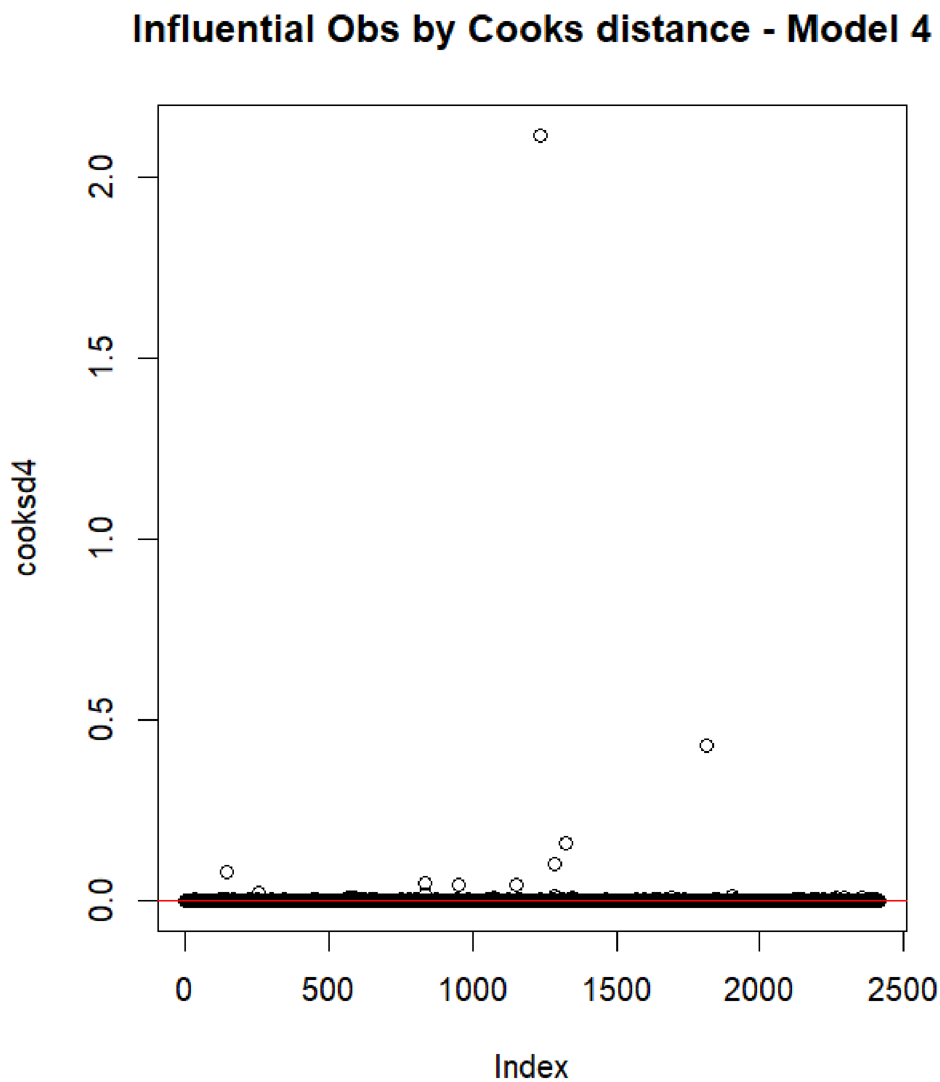


While the residuals in Model 4 do not ‘fan-out’ compared to the other models, it doesn’t necessarily show general constant variance. The GVLMA assessment shows that its calculated heteroscedasticity is still high:

```
> gvmodel4 <- gvlma(model4)
> summary(gvmodel4)

Call:
lm(formula = subdat$LogSalePrice ~ subdat$TotalFloorSF + subdat$OverallQual +
    subdat$LotArea)

Residuals:
     Min       1Q   Median       3Q      Max
-2.21758 -0.08933  0.01907  0.11520  0.63841

Coefficients:
                       Estimate    Std. Error t value             Pr(>|t|)
(Intercept)         10.4763381614  0.0180943643 578.983 <0.0000000000000002 ***
subdat$TotalFloorSF  0.0002509380  0.0000106586  23.543 <0.0000000000000002 ***
subdat$OverallQual   0.1804970765  0.0036544341  49.391 <0.0000000000000002 ***
subdat$LotArea       0.0000078797  0.0000008429   9.348 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1922 on 2410 degrees of freedom
Multiple R-squared:  0.7842,	Adjusted R-squared:  0.7839
F-statistic:  2919 on 3 and 2410 DF,  p-value: < 0.00000000000000022


ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = model4)

                      Value   p-value                   Decision
Global Stat        23633.45 0.0000000 Assumptions NOT satisfied!
Skewness            1713.10 0.0000000 Assumptions NOT satisfied!
Kurtosis           21829.02 0.0000000 Assumptions NOT satisfied!
Link Function         80.05 0.0000000 Assumptions NOT satisfied!
Heteroscedasticity    11.28 0.0007835 Assumptions NOT satisfied!
```

Due to the heteroscedasticity evaluation, we cannot validate Model 4 as a linear regression model despite the log transformation of SalePrice.

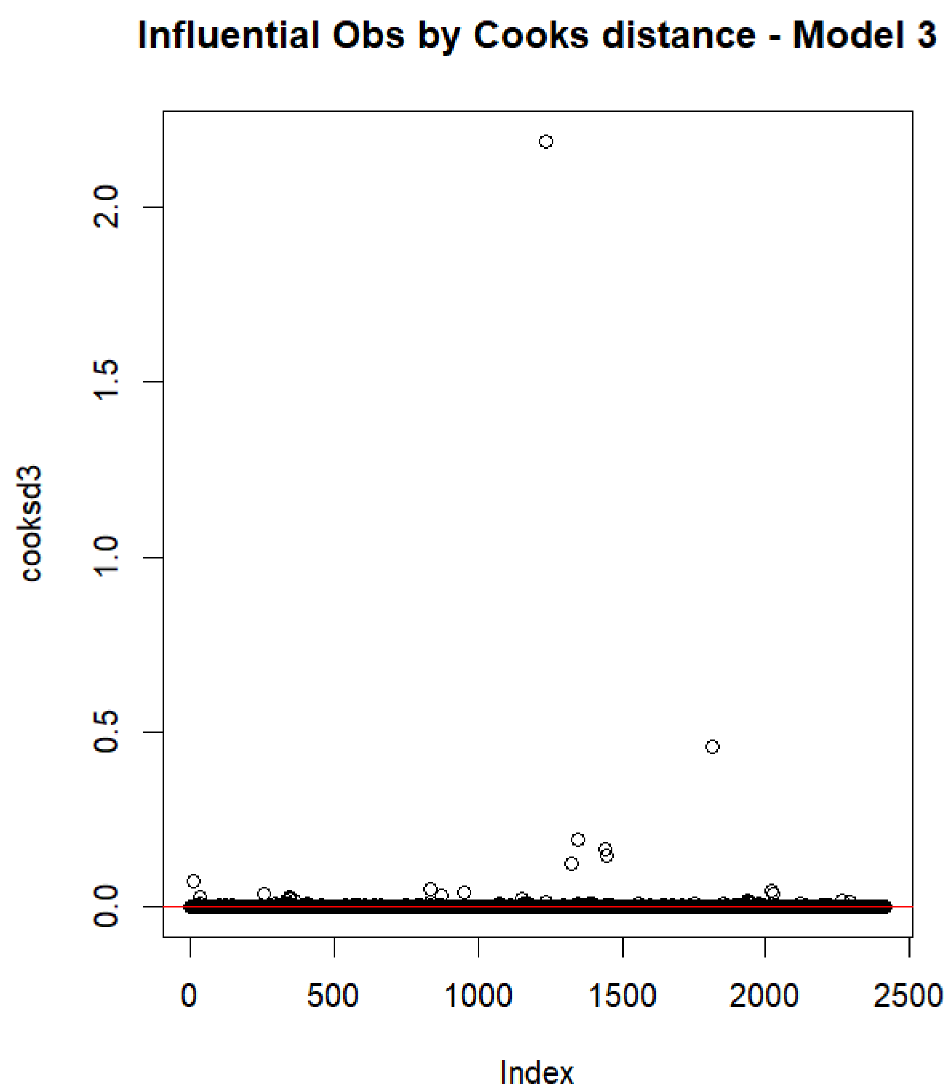
In terms of influence, calculating Cook's distance on Model 4 gives us 123 potential data points we can remove to potentially improve heteroscedasticity:




```
> influential4 <- as.numeric(names(cooksd4)[(cooksd4 > (4/nrow(subdat)))])
> influential4
  [1]   13   15   33   62  101  120  131  143  145  169  191  224  233  244  248  249  254  293
 [19]  299  342  343  443  451  457  565  569  574  575  580  584  585  591  592  593  616  624
 [37]  643  647  662  750  780  803  834  835  871  947  951  985 1045 1050 1055 1061 1064 1065
 [55] 1074 1075 1076 1077 1078 1084 1103 1112 1122 1153 1157 1158 1164 1236 1237 1238 1259 1272
 [73] 1276 1283 1285 1294 1312 1323 1345 1346 1394 1462 1568 1597 1601 1632 1644 1649 1675 1689
 [91] 1708 1709 1716 1736 1737 1815 1842 1851 1853 1863 1906 1916 1938 1945 2029 2118 2136 2147
[109] 2172 2181 2192 2194 2196 2197 2222 2225 2260 2263 2291 2355 2379 2388 2405
> length(influential4)
[1] 123
```

5. For either Model 3 or Model 4, your choice, identify the influential, high leverage, or outlier data points. Remove these data points from the dataset, then refit the model after removing the influential points. How many influential points did you find & remove? When you refitted the model, did the model improve? Comment on whether or not you find the improvement of model fit justifies the potential for the modeler biasing the result by removing potentially legitimate data points.

We will attempt to improve Model 3 by removing the potential 150 influencers identified in section 3f:



```
> influential3 <- as.numeric(names(cooksd3)[(cooksd3 > (4/nrow(subdat)))])
> influential3
  [1]   12   13   14   27   33   34   48  101  120  122  145  191  237  244  254  278  293  299
 [19]  331  332  333  339  340  341  342  343  344  346  358  361  362  388  397  406  414  457
 [37]  530  538  569  574  580  584  607  616  637  662  748  750  792  801  803  833  834  835
 [55]  858  859  861  863  864  865  867  871  875  899  951 1061 1065 1074 1076 1078 1101 1112
 [73] 1153 1157 1158 1173 1174 1236 1237 1238 1294 1323 1343 1345 1346 1347 1348 1385 1386 1388
 [91] 1390 1391 1394 1395 1396 1401 1440 1441 1443 1447 1452 1460 1462 1556 1568 1597 1632 1644
[109] 1646 1649 1675 1708 1709 1736 1737 1755 1815 1842 1853 1865 1894 1897 1906 1916 1935 1936
[127] 1938 1940 1945 1973 1975 1988 1989 1990 1991 2018 2019 2023 2029 2118 2136 2181 2192 2196
[145] 2202 2220 2263 2291 2379 2404
> length(influential3)
[1] 150
```

With the influencers above removed, we create the following linear regression Model 5:

```
subdat_cleanup <- subdat
subdat_cleanup <- subdat[-influential3, ]
```

```
> model5 <- lm(subdat_cleanup$SalePrice ~
+                subdat_cleanup$TotalFloorSF
+              + subdat_cleanup$OverallQual
+              + subdat_cleanup$LotArea)
> model5

Call:
lm(formula = subdat_cleanup$SalePrice ~ subdat_cleanup$TotalFloorSF +
    subdat_cleanup$OverallQual + subdat_cleanup$LotArea)

Coefficients:
                (Intercept)  subdat_cleanup$TotalFloorSF   subdat_cleanup$OverallQual
                -100974.473                       55.233                    28128.316
     subdat_cleanup$LotArea
                      2.626
```

SalePrice = -100974.473+ 55.233*TotalFloorSF + 28128.316*OverallQual + 2.626*LotArea

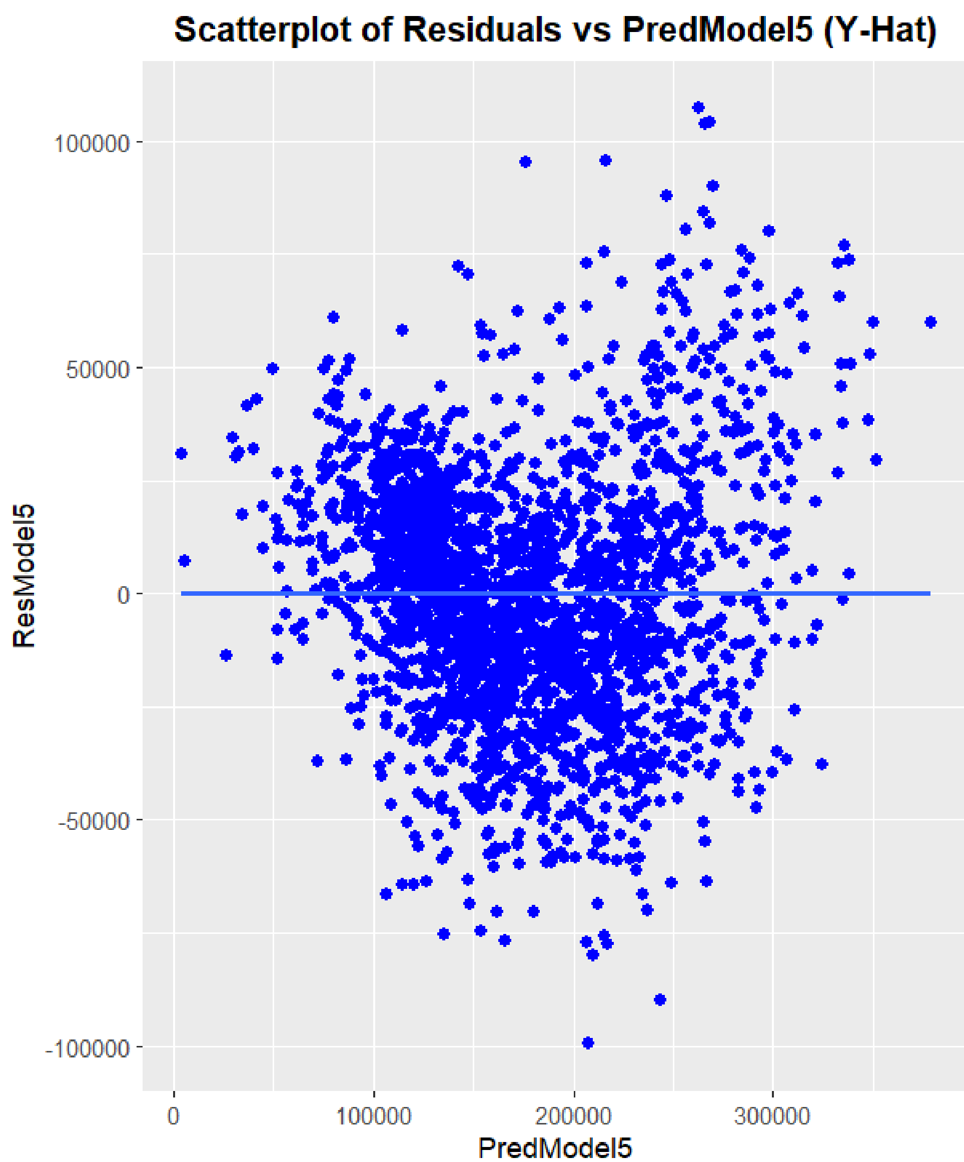
Compared to Model 3:

```
> model3

Call:
lm(formula = subdat$SalePrice ~ subdat$TotalFloorSF + subdat$OverallQual +
    subdat$LotArea)

Coefficients:
        (Intercept)  subdat$TotalFloorSF   subdat$OverallQual       subdat$LotArea
        -119812.876               56.156            33131.628                1.652
```

The y-intercept and the coefficients have increased after the removal of the influence points.

```
> anova(model5)
Analysis of Variance Table

Response: subdat_cleanup$SalePrice
                              Df        Sum Sq       Mean Sq F value                Pr(>F)
subdat_cleanup$TotalFloorSF    1 5922028294340 5922028294340 8045.97 < 0.00000000000000022 ***
subdat_cleanup$OverallQual     1 1744346066541 1744346066541 2369.96 < 0.00000000000000022 ***
subdat_cleanup$LotArea         1  176765060196  176765060196  240.16 < 0.00000000000000022 ***
Residuals                   2260 1663413761926     736023788
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

High F-Values for each variable along with P-Values near zero in the ANOVA table above gives us reason to reject their respective null hypotheses where each Beta = 0 and accept their respective alternative hypotheses where Beta != 0 such that each variable is statistically significant.



We also notice that the range of residuals in the scatterplot above have also become narrower compared to Model 3, showing more of a constant variance, and visually showing better signs of homoscedasticity.

```
> gvmodel5 <- gvlma(model5)
> summary(gvmodel5)

Call:
lm(formula = subdat_cleanup$SalePrice ~ subdat_cleanup$TotalFloorSF +
    subdat_cleanup$OverallQual + subdat_cleanup$LotArea)

Residuals:
   Min     1Q Median     3Q    Max
-99551 -17387    365  17108 107410

Coefficients:
                              Estimate   Std. Error t value           Pr(>|t|)
(Intercept)               -100974.4731    2955.3246  -34.17 <0.0000000000000002 ***
subdat_cleanup$TotalFloorSF     55.2329      1.6889   32.70 <0.0000000000000002 ***
subdat_cleanup$OverallQual   28128.3161    574.4066   48.97 <0.0000000000000002 ***
subdat_cleanup$LotArea           2.6261      0.1695   15.50 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27130 on 2260 degrees of freedom
Multiple R-squared:  0.825,     Adjusted R-squared:  0.8248
F-statistic:  3552 on 3 and 2260 DF,  p-value: < 0.00000000000000022


ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = model5)

                     Value     p-value                   Decision
Global Stat        311.023 0.000000000 Assumptions NOT satisfied!
Skewness             8.384 0.003786009 Assumptions NOT satisfied!
Kurtosis            22.522 0.000002077 Assumptions NOT satisfied!
Link Function      278.357 0.000000000 Assumptions NOT satisfied!
Heteroscedasticity   1.760 0.184612195    Assumptions acceptable.
```

Calculating heteroscedasticity via GVLMA gives a value of 1.76, which is an acceptable value near 1, and a large improvement over Model 3.

After the removal of influencers in Model 3, Model 5 seems to be a much more valid linear regression model in terms of much-improved heteroscedasticity/homoscedasticity. The summary also shows R-Squared value has also improved by over 6% compared to Model 3 in explaining the variance to SalePrice as well. The GVLMA also suggests that we should also find ways to improve other linear regression model properties such as Skewness and Kurtosis.

6. So far, we have fit a few models to predict SALEPRICE(Y). But there are many other continuous variables in the data set, with many different possible combinations of variables that could be used in a regression model. You could use theory, or your background knowledge, to select variables for inclusion in a multiple regression model. Many modelers do this. It gives a nice place to start the search process. On the technical side, in this assignment, we know about correlation between variables and have been looking at change

in R-squared when a new variable has been added to an existing model to isolate the explanatory contribution of that new variable. We have also been looking at hypothesis tests on the individual coefficients.

Use the concept of Change in R-squared, plus anything else you wish, to put together a reasonable approach to find a good, comprehensive multiple regression model to predict SALEPRICE(Y). Any of the continuous variables can be considered fair game as explanatory variables. This can feel like an overwhelming task. You don't need to go overboard, or kill yourself, in doing this. We will learn about automated approaches to do this shortly. But, for now, I'd like you to think about how you would do this by hand.

Use your approach to identify a good multiple regression model to predict SALEPRICE(Y) from the set of continuous explanatory variables available to you in the AMES dataset.

For this task you need to:

a. Explain your approach.

We created Model 6 using continuous explanatory variables from the dataset that covered general square footage areas in a single-family home and are highly correlated to SalePrice, such as TotalFloorSF, TotalBsmtSF, MasVnrArea, and GarageArea, as well as OverallQual which was used in Models 2, 3, 4, and 5.

b. Report the model you determined and interpret the coefficients.

```
> model6

Call:
lm(formula = subdat$SalePrice ~ subdat$TotalFloorSF + subdat$TotalBsmtSF +
    subdat$MasVnrArea + subdat$GrLivArea + subdat$GarageArea +
    subdat$LotArea + subdat$OverallQual)

Coefficients:
        (Intercept)  subdat$TotalFloorSF   subdat$TotalBsmtSF    subdat$MasVnrArea     subdat$GrLivArea
        -98139.9580              52.6888              32.0954              44.2532              -5.2526
  subdat$GarageArea       subdat$LotArea   subdat$OverallQual
            52.1882               0.6547           23079.3591
```

We created Model 6 with the following:

SalePrice = -98139.9580 + 52.6888*TotalFloorSF + 32.0954*TotalBsmtSF + 44.2532 *MasVnrArea + 52.1882*GarageArea + 23079.3591* OverallQual

An extra square foot to the total flooring would add $56.69 to the SalePrice, and extra square foot from the total basement would add $32.10, an extra square foot of masonry veneer would add $44.25, and extra square foot of garage space would add $52.19, and just one higher rating in overall quality would add $23079.36.

c. Report the coefficient and ANOVA tables and goodness of fit.

```
> anova(model6)
Analysis of Variance Table

Response: subdat$SalePrice
                    Df        Sum Sq       Mean Sq F value                Pr(>F)
subdat$TotalFloorSF    1 9447676578136 9447676578136 7214.03 < 0.00000000000000022 ***
subdat$TotalBsmtSF     1 1932918488448 1932918488448 1475.93 < 0.00000000000000022 ***
subdat$MasVnrArea      1  306471053246  306471053246  234.01 < 0.00000000000000022 ***
subdat$GarageArea      1  564581922998  564581922998  431.10 < 0.00000000000000022 ***
subdat$OverallQual     1 1066349612149 1066349612149  814.24 < 0.00000000000000022 ***
Residuals           2408 3153577107947    1309625045
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Each explanatory value in Model 6 has overwhelmingly high values greater than the $F_{Critical}$ value of 0.1662 with 5 and 2408 degrees of freedom and a 95% confidence interval, and P-Values very close to 0.

```
> summary(model6)

Call:
lm(formula = subdat$SalePrice ~ subdat$TotalFloorSF + subdat$TotalBsmtSF +
    subdat$MasVnrArea + subdat$GarageArea + subdat$OverallQual)

Residuals:
    Min      1Q  Median      3Q     Max
-570718  -18330    -854   16087  266160

Coefficients:
                     Estimate Std. Error t value            Pr(>|t|)
(Intercept)        -93896.436   3521.955 -26.660 <0.0000000000000002 ***
subdat$TotalFloorSF     49.824      1.968  25.318 <0.0000000000000002 ***
subdat$TotalBsmtSF      33.684      2.218  15.184 <0.0000000000000002 ***
subdat$MasVnrArea       44.812      4.960   9.034 <0.0000000000000002 ***
subdat$GarageArea       54.565      4.523  12.064 <0.0000000000000002 ***
subdat$OverallQual   22465.740    787.308  28.535 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 36190 on 2408 degrees of freedom
Multiple R-squared:  0.8085,    Adjusted R-squared:  0.8081
F-statistic:  2034 on 5 and 2408 DF,  p-value: < 0.00000000000000022
```

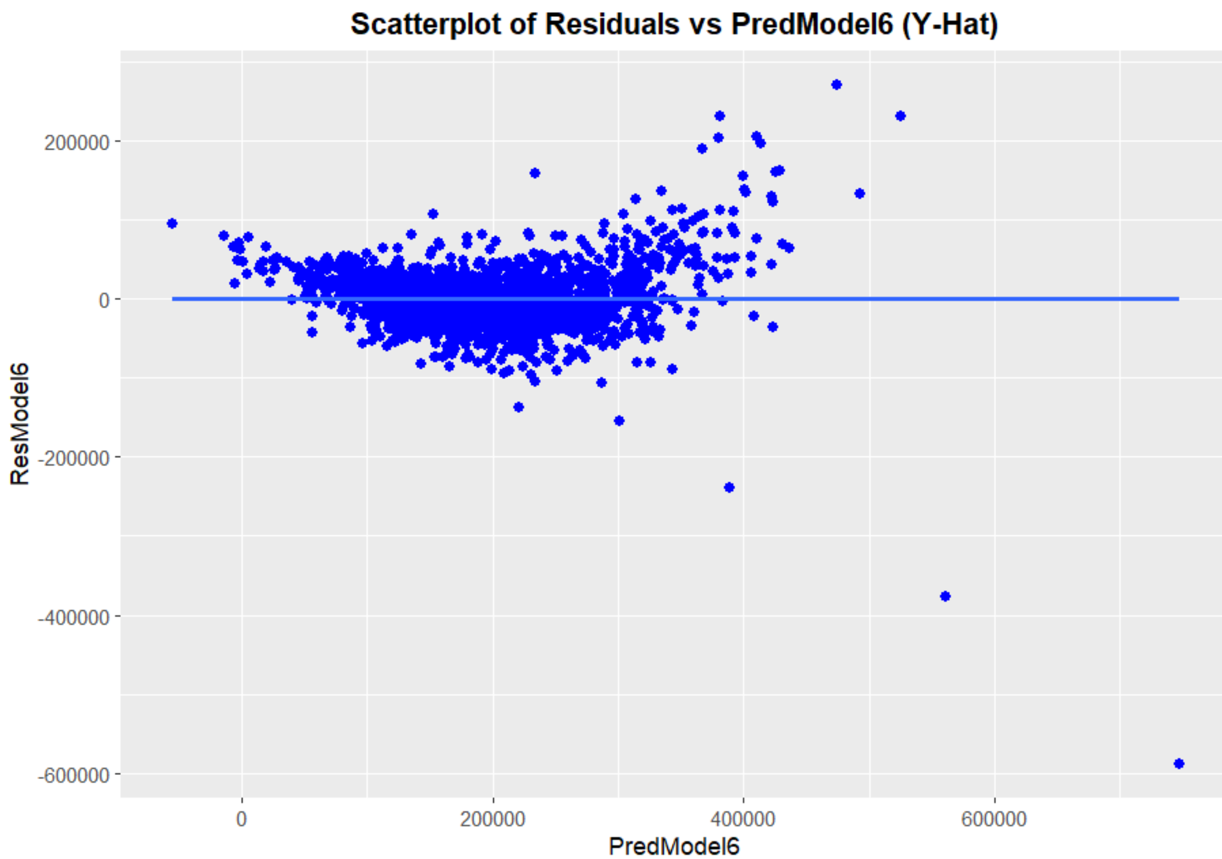
Each explanatory value in Model 6 has have high T-values that are greater than the $T_{Critical}$ value of 1.96 with 2408 degrees of freedom and a 95% confidence interval.

With the information above and given the null hypothesis H0 for each coefficient to their respective explanatory variable where each beta = 0, we can reject each of these null hypotheses and accept their alternative hypothesis where each beta !=0 and are statistically significant.

As shown in the summary above, the five explanatory variables in Model 6 have an R-squared value of 0.8095, in which the model can explain 80.1% of the variability to SalePrice.

d. Check on underlying model assumptions.

The scatterplot of residuals to the predicted values of Model 6 are the following:



Compared to the other models, with the exception of a few outliers, most of the dataset has more of a constant difference in residuals, which indicate more homoscedasticity than heteroscedasticity.

GLVMA also confirms greater homoscedasticity with a calculated value near 1 (0.8513), which is much better than Models 1 to 4.

```
> gvmodel6

Call:
lm(formula = subdat$SalePrice ~ subdat$TotalFloorSF + subdat$TotalBsmtSF +
    subdat$MasVnrArea + subdat$GarageArea + subdat$OverallQual)

Coefficients:
        (Intercept)  subdat$TotalFloorSF   subdat$TotalBsmtSF    subdat$MasVnrArea
          -93896.44                49.82                33.68                44.81
  subdat$GarageArea   subdat$OverallQual
              54.56             22465.74


ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = model6)

                         Value             p-value                   Decision
Global Stat        125154.8530 0.000000000000000000 Assumptions NOT satisfied!
Skewness              544.5605 0.000000000000000000 Assumptions NOT satisfied!
Kurtosis           124546.6395 0.000000000000000000 Assumptions NOT satisfied!
Link Function          62.8017 0.000000000000002331 Assumptions NOT satisfied!
Heteroscedasticity      0.8513 0.356184334292671267    Assumptions acceptable.
```

With Model 6 exhibiting better heteroscedasticity/homoscedasticity like Model 5, we can also validate it as a linear regression model, but GVLMA indicates that we should also consider other ways to improve the model in terms of other linear regression model attributes such as Skewness and Kurtosis.

7. Please write a conclusion / reflection section that, at minimum, addresses the questions:

• In what ways do variable transformation and outlier deletion impact the modeling process and the results?

• Are these analytical activities a benefit or do they create additional difficulties?

• Can you trust statistical hypothesis test results in regression?

• What do you consider to be next steps in the modeling process?

In regard to the Ames dataset and an initial waterfall dropdown limiting our study to primarily single-family homes, we found the deletion of influencers from the dataset did have an impact in improving the model's heteroscedasticity/homoscedasticity for linear model regression. We also find that taking the route of adding more highly correlated continuous variables to SalePrice greatly improved the model's homoscedasticity/homoscedasticity as well. The extra effort taken to determine a model's heteroscedasticity/homoscedasticity helps in producing a better, more valid linear regression model overall. From this exercise we find that hypothesis tests alone do not fully cover all the required assumptions when it comes to creating linear regression models. Going forward, we should integrate more discrete, ordinal, and nominal variables from the dataset into the model via proper transformations, normalizations, and scaling. A model cannot be only dependent on continuous variables and integrating interpretations of other non-continuous variables available in the dataset could potentially make it more robust.