Modeling: Building Linear Regression Models

Introduction:

Relevant housing data can support sale price prediction. For this modeling analysis, housing data was sourced from Ames, Iowa Assessor's Office. Our objective first is to identify sampling population that we will focus our modeling on. Then we will begin by fitting specific models and looking at diagnostic and model fit information.

Define the Sample Population:

Sample population has been defined as Single Family with residential zoning and the sale condition was normal. This is the same sample population previously defined in last project.

Missing values:

Replacing numeric variables where there is missing value with mean.

Replacing categorical variables where there is missing value with mode.

I do acknowledge this could potentially create bias in our sample. But for the purpose of modeling, we do not have any null values, and we have eliminated extreme outliers from our data set.

Drop functions waterfall table:

Waterfall drop funtions	**Records Dropped**	**Total Observation**
My data	0	2930
mydata, -Alley, -FireplaceQu, -PoolQC, -Fence, -MiscFeature	0	2930
BldgType == "1Fam"	505	2425
Zoning %in% c("RH", "RL", "RM", "FV")	26	2399
SaleCondition == "Normal"	411	1988
cleandata$GrLivArea<=4000	1	1987
cleandata$TotalFloorSF<=4000	0	1987
cleandata$SalePrice<=500000	7	1981

Adding these 5 additional variables to our data set

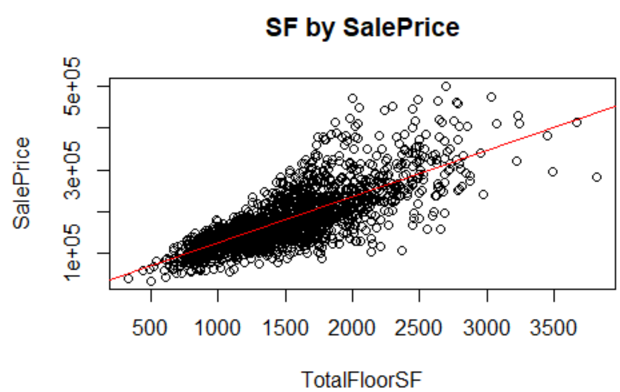
- mydata$TotalFloorSF <- mydata$FirstFlrSF + mydata$SecondFlrSF
- mydata$HouseAge <- mydata$YrSold - mydata$YearBuilt
- mydata$QualityIndex <- mydata$OverallQual * mydata$OverallCond
- mydata$price_sqft <- mydata$SalePrice/mydata$TotalFloorSF
- mydata$logSalePrice <- log(mydata$SalePrice)

PART A: Simple Linear Regression Models

(1) Let Y = sale price be the dependent or response variable. Select “the best” continuous explanatory variable from the AMES data set to predict Y.

rowname	SalePrice
OverallQual	0.801761801
TotalFloorSF	0.782481418
GrLivArea	0.77535809
GarageCars	0.660977076
TotalBsmtSF	0.650020052
GarageArea	0.640403126
FullBath	0.609247762
TotRmsAbvGrd	0.599288489
SF	0.558873904
MasVnrArea	0.554924441

a. Based on top 10 correlation table, TotalFloorSF is “the best” continuous explanatory.



b. Y=14505.205 + 110.249 β1

- This linear equation indicated, for every square footage increase, the total sales price will go up by $110.25, which is at $14.615.45.

c. Based on R-squared value of 0.6027, this means it can explain 60.27% of total variance of the model.

```
Call:
lm(formula = SalePrice ~ TotalFloorSF, data = cleandata)

Residuals:
    Min      1Q  Median      3Q     Max
-168516  -24527   -1331   19511  234996

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  14505.205   3143.648   4.614  4.2e-06 ***
TotalFloorSF   110.249      2.012  54.794  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 43150 on 1979 degrees of freedom
Multiple R-squared:  0.6027,    Adjusted R-squared:  0.6025
F-statistic:  3002 on 1 and 1979 DF,  p-value: < 2.2e-16
```

d. Report the coefficient and ANOVA Tables.

- Hypothesis testing for β1

H0 : β1 = 0

Ha : β1 ≠ 0

- T-statistics for β1 = 110.249/2.012=54.794, and with P value is <2e-16. We reject the null hypothesis. This explanatory variable has statistical significance to the prediction of our target response variable SalePrice(Y).

```
Analysis of Variance Table

Response: SalePrice
               Df     Sum Sq    Mean Sq F value    Pr(>F)
TotalFloorSF    1 5.5903e+12 5.5903e+12  3002.4 < 2.2e-16 ***
Residuals    1979 3.6847e+12 1.8619e+09
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

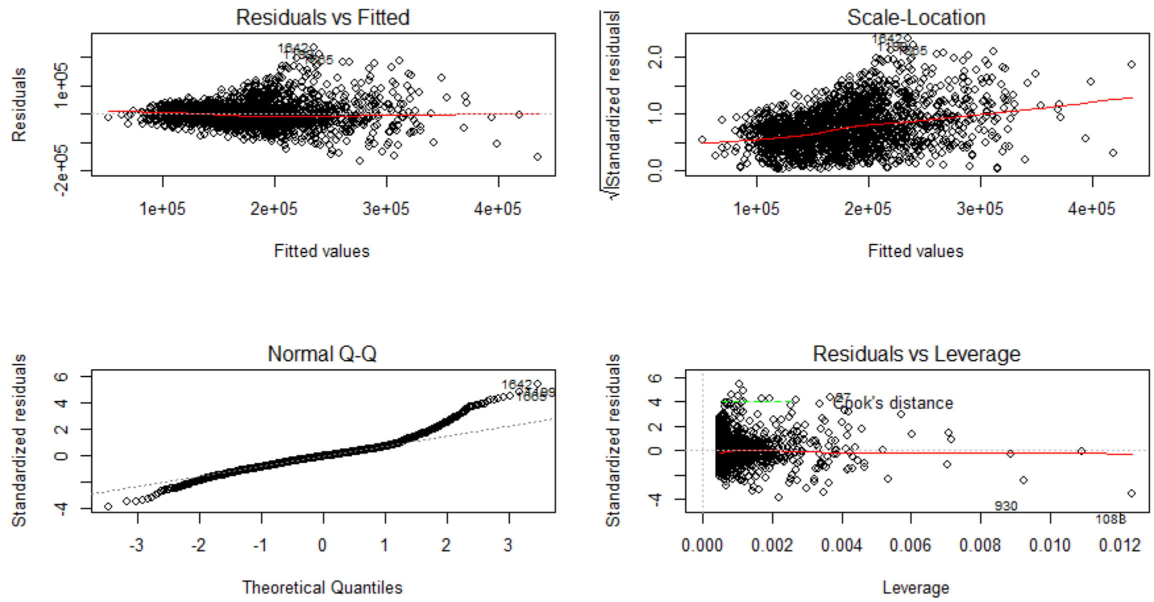
- Omnibus model

H0 : β1 = 0

Ha : βj ≠ 0, for at least one j, j = 1

F0=3002.4 with p value <2.2e-16. We reject the null hypothesis.

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



- This scatterplot for Residuals vs Fitted indicates a distinct parabola shape, so we can say the model does not meet the homoscedasticity assumption since the residuals are not equally spread around the y = 0 line. linear model assumption is there since the redline through our scatterplot is fairly straight.

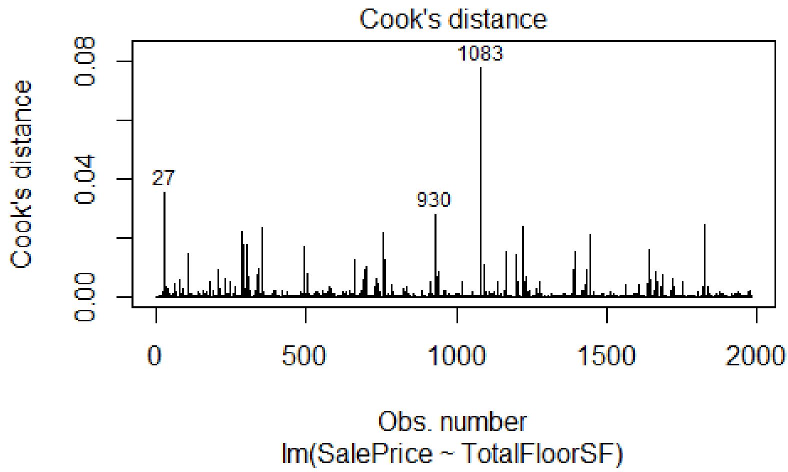
- Our QQ line is measuring the normality. We observed three outliers (1642, 1199, and 1665). Majority is aligned with 45-degree line, but we do have 2 tails deviating away from the normality.

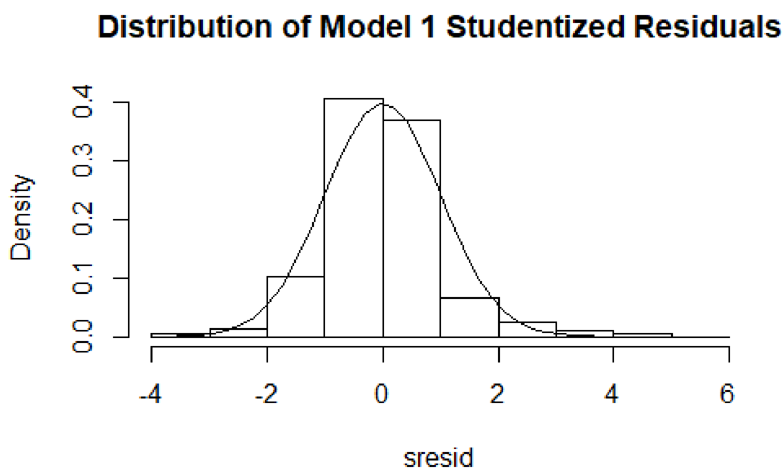
- Scale location graph is supporting the evaluation of homoscedasticity. We see the red line is sloping slightly up and the data points are not randomly spread out. So, this model violated the assumption.

- Cook's distance line is slightly crossed by three influential points (27, 930,1083) presenting the issue with bias in the model.

outlierTest(model1)			
	rstudent	unadjusted p-value	Bonferroni p
1642	5.488936	4.5639e-08	0.00009041
1199	4.896069	1.0570e-06	0.00209390
1665	4.532754	6.1700e-06	0.01222300
702	4.483155	7.7759e-06	0.01540400
27	4.391204	1.1868e-05	0.02351000
293	4.294375	1.8367e-05	0.03638400

- Looking further at the outlier test, we see there are 6 extreme observations. 1642 is the most extreme among the 6.

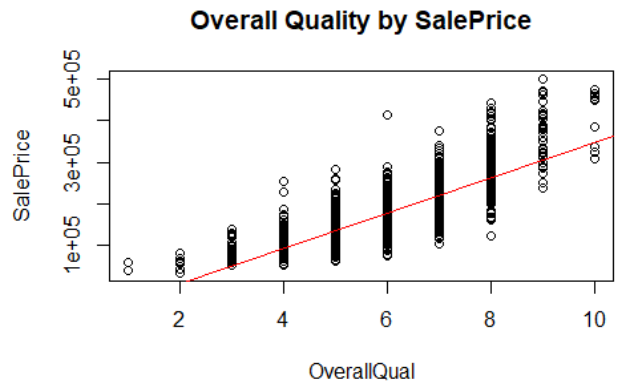




Histogram of studentized residuals suggest that it's normally distributed.

2) Let Y = sale price be the dependent or response variable. Use the OVERALL QUALITY variable as the explanatory variable (X) to predict Y. Fit a simple linear regression model using X to predict Y. Call this Model 2.

- Make a scatterplot of Y and X, and overlay the regression line on the cloud of data



b. Y= -75196.1 + 42256.4 β1

- The incremental per OverallQual, which is ordinal variable, obviously not the same as continuous variable of TotalFloorSF. Per 1 OverallQual based on this equation will add $42,256.40 to the Sale Price. Intercept at -$75,196.10, that would mean -$32,939.70. This certainly does not make sense. Since per quality score can have various price point, this variable can create challenge as the sole predictor for the linear model.

c. Based on R-squared value of 0.6521, this means variable represents overall quality of the house can explain 65.21% of total variation of the model.

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

Residuals:
    Min      1Q  Median      3Q     Max
-140855  -25599   -3586   19401  236658

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -75196.1     4261.3  -17.65   <2e-16 ***
OverallQual  42256.4      693.9   60.90   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 40380 on 1979 degrees of freedom
Multiple R-squared:  0.6521,	Adjusted R-squared:  0.6519
F-statistic:  3709 on 1 and 1979 DF,  p-value: < 2.2e-16
```

d. Hypothesis testing for β1

H0 : β1 = 0

Ha : β1 ≠ 0

- T-statistics for β1 = 42256.4/693.9=60.90, and with P value is <2e-16. We reject the null hypothesis. This explanatory "OverallQual" variable has statistical significance to the prediction of our target response variable SalePrice(Y).

Analysis of Variance Table					
Response: SalePrice					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
OverallQual	1	6.0478e+12	6.0478e+12	3708.7	< 2.2e-16 ***
Residuals	1979	3.2272e+12	1.6307e+09		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

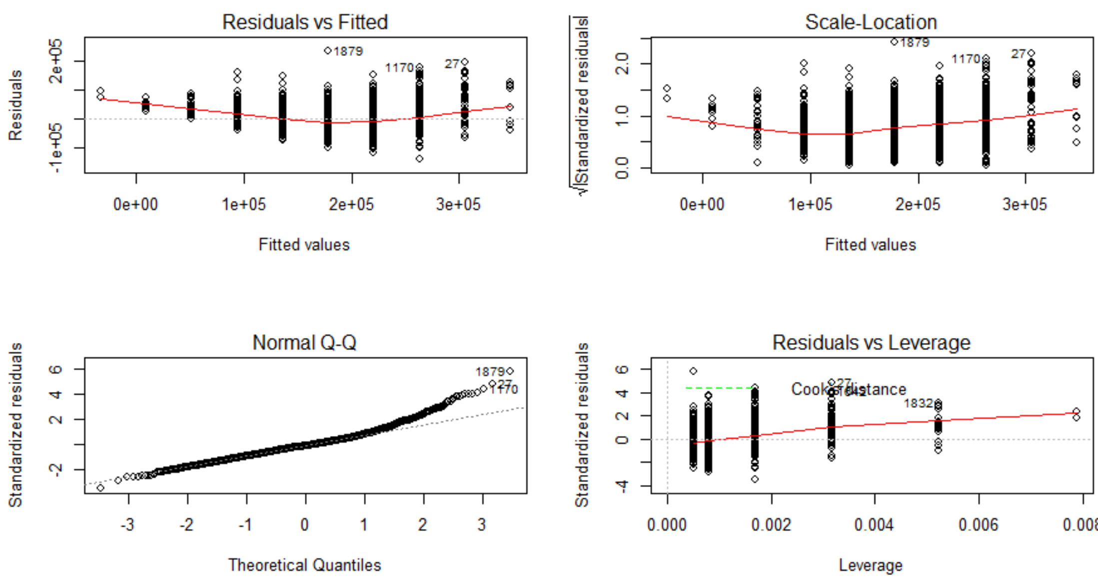
- Omnibus testing

H0 : β1 = 0

Ha : βj ≠ 0, for at least one j, j = 1

F0=3708.7 with P value is <2e-16. We reject the null hypothesis.

e. Check on the underlying assumptions.



- We observed three outliers (1879, 27, and 1170). Majority is aligned with 45-degree line, but we do have a tail curving upwards away from the normality.
- Linearity is violated since the line in residuals versus fitted is slightly curving.
- QQline suggests normality except three outliers, which is extreme validated by Cook's distance test which can impact our model.

	rstudent	unadjusted p-value	Bonferroni p
1879	5.912032	3.9705e-09	7.8655e-06
27	4.861338	1.2579e-06	2.4918e-03
1170	4.459373	8.6809e-06	1.7197e-02

3) Of the above 2 models, which one fits better? On what criteria are you assessing the model fit?

Model	Y ~ X	R-Squared	Adj. R-Sqaured	RSE	Delta R-Squared
Model 1	SalePrice ~ TotalFloorSF	0.6027	0.6025	43150	0%
Model 2	SalePrice ~ OverallQual	0.6521	0.6519	40380	5%

- Based on the 2 models, using R-Squared value as well as RSE, they suggest model 2 is better, and improved by about 5% while RSE decrease by 2770.

PART B: Multiple Linear Regression Models

4) Fit a multiple regression model that uses 2 continuous explanatory (X) variables to predict Sale Price (Y). These two explanatory(X) variables should be: the explanatory variables from Model 1 and Model 2 above. Call this Model 3. You should:

a. Y= -81875.149 + 63.34 β1 + 27680.949 β2

- Per incremental of total floor sf can add $63.34 to the sale price. Per overall qual it can add $27,680.949 to the total sale price.
- Both variables' coefficient values have decreased comparing to the simple linear equation.

	Simple Linear	Multiple linear	Delta
TotalFloorSF	110.25	63.34	(46.91)
OverallQual	42,256.40	27,680.95	(69,937.35)

```
Call:
lm(formula = SalePrice ~ TotalFloorSF + OverallQual, data = cleandata)

Residuals:
    Min      1Q  Median      3Q     Max
-153943  -20243     190   17380  176067

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -81875.149   3445.735  -23.76   <2e-16 ***
TotalFloorSF     63.340      1.946   32.55   <2e-16 ***
OverallQual   27680.949    717.085   38.60   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 32600 on 1978 degrees of freedom
Multiple R-squared:  0.7734,    Adjusted R-squared:  0.7732
F-statistic:  3376 on 2 and 1978 DF,  p-value: < 2.2e-16
```

b. R-Squared at 0.7734 with 2 variables in model 3 suggests it's better combined in explaining 77.34% of the variation of the total model.

- Model 3 improved the power in explaining the variance of the model by 12%. OverallQual is statistically significant for the model of Sale Price.

Model	Y ~ X	R-Squared	Adj. R-Sqaured	RSE	Delta R-Squared
Model 1	SalePrice ~ TotalFloorSF	0.6027	0.6025	43150	0%
Modle 3	SalePrice ~ TotalFloorSF + OverallQual	0.7734	0.7732	32600	12%

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

Response: SalePrice
              Df     Sum Sq    Mean Sq F value    Pr(>F)
TotalFloorSF    1 5.5903e+12 5.5903e+12  5261.6 < 2.2e-16 ***
OverallQual     1 1.5832e+12 1.5832e+12  1490.1 < 2.2e-16 ***
Residuals    1978 2.1015e+12 1.0625e+09
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

c. Hypothesis testing for β1, β2

H0 : β1 = 0
Ha : β1 ≠ 0

- T-statistics for β1 = 32.55 , and with P value is <2.2e-16. We reject the null hypothesis.

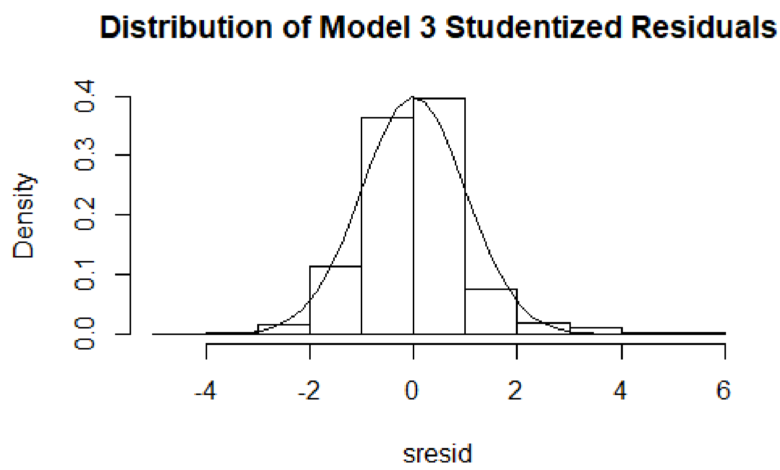
H0 : β2 = 0
Ha : β2 ≠ 0

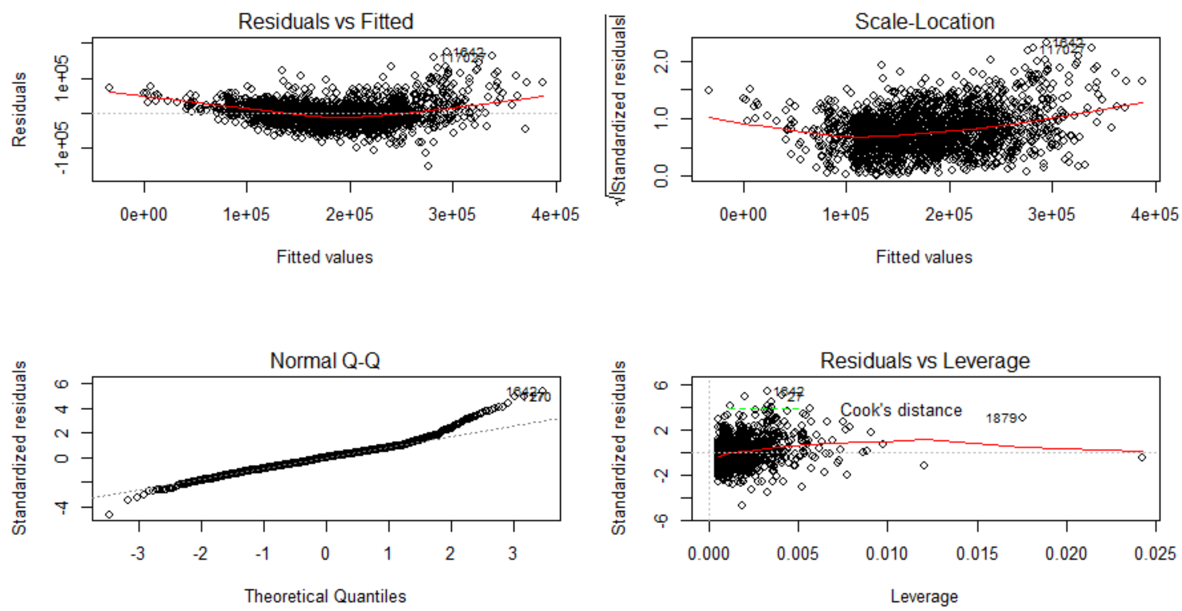
- T-statistics for β2 = 38.60 , and with P value is <2.2e-16. We reject the null hypothesis.

- Omnibus Model

  H0 : β1 = β2 = 0
  Ha : βj ≠ 0, for at least one j, j = 1, 2
- F0=3376 with p value<2.2e-16 we reject hypothesis, at least one beta does not equal to 0.

d. Underlying assumption





- Model is normally distributed based on studentized residual histogram
- Linearity is violated given it's concaved in residuals vs Fitted graph. But residuals are little more distributed than model 1 except some extreme observations.
- QQ line indicate normal distribution except extreme observations
- We can identify 27, 1642 crossed Cook's distance. Outlier test listed 5 below as extreme.

	rstudent	unadjusted p-value	Bonferroni p
1642	5.449558	5.6807e-08	0.00011253
27	5.010013	5.9267e-07	0.00117410
1170	4.969305	7.2978e-07	0.00144570
938	-4.752918	2.1496e-06	0.00425840
352	4.416297	1.0583e-05	0.02096500

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

Model	Y ~ X	R-Squared	Adj. R-Sqaured	RSE	Delta R-Squared
Model 1	SalePrice ~ TotalFloorSF	0.6027	0.6025	43150	0%
Model 2	SalePrice ~ OverallQual	0.6521	0.6519	40380	5%
Modle 3	SalePrice ~ TotalFloorSF + OverallQual	0.7734	0.7732	32600	12%

- Based on R-Squared and RSE values, both variables indicated significance in explaining the variance of the model at 77.34%, increased by 12 % comparing to Model2, 17% comparing to Model1. Both variables should be retained.

5) Select any other continuous 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 your variable of choice plus the explanatory variables from Model 3. Call this Model 4. You should:

rowname	**SalePrice**
OverallQual	0.801761801
TotalFloorSF	0.782481418
GrLivArea	0.77535809
GarageCars	0.660977076
TotalBsmtSF	0.650020052
GarageArea	0.640403126

- Adding GrLivArea, TotalBsmtSF, and GarageArea continuous explanatory into Model 4, which is nesting Model 3.

```
lm(formula = SalePrice ~ TotalFloorSF + OverallQual + GrLivArea +
    TotalBsmtSF + GarageArea, data = cleandata)

Residuals:
    Min      1Q  Median      3Q     Max
-105786  -16207    -926   14482  156887

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -83815.656   2885.917 -29.043  < 2e-16 ***
TotalFloorSF     60.008     12.625   4.753 2.15e-06 ***
OverallQual   18856.037    668.062  28.225  < 2e-16 ***
GrLivArea        -5.663     12.431  -0.456    0.649
TotalBsmtSF      41.200      1.856  22.199  < 2e-16 ***
GarageArea       55.181      3.828  14.417  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 27100 on 1975 degrees of freedom
Multiple R-squared:  0.8436,	Adjusted R-squared:  0.8432
F-statistic:  2131 on 5 and 1975 DF,  p-value: < 2.2e-16
```

a. Y= -83815.656+ 60.008 β1 + 18856.037 β2 - 5.663 β3 + 41.20 β4+ 55.181 β5

- Per incremental of total floor sf adds $60.008 to the sale price.
- Per overall qual adds $18,856.037 to the total sale price.
- Per GrLivArea it adds $-5.663 to the total sale price.
- Per incremental total basement sf adds $41.2 to the total sale price.
- Per GrLivArea it adds $-55.181 to the total sale price.

b. β1 = TotalFloorSF, β2 = OverallQual, β3 = GrLivArea, β4 = TotalBsmtSF, β5 = GarageArea,

Hypothesis testing for β1

H0 : β1 = 0

Ha : β1 ≠ 0

- T-statistics for β1 = 60.008/12.625=4.753, and with P value is 2.15e-06. We reject the null hypothesis. TotalFloorSF should be retained for explaining the model.

Hypothesis testing for β2

H0 : β2 = 0

Ha : β2 ≠ 0

- T-statistics for β2 = 18856.037/668.062=28.225, and with P value is 2e-16. We reject the null hypothesis. OverallQual should be retained for explaining the model.

Hypothesis testing for β3

H0 : β3 = 0

Ha : β3 ≠ 0

- T-statistics for β3 = -5.663/3.828=14.417, and with P value is at 0.649. We fail to reject the null hypothesis. GrLivArea should not be included for modeling target response of SalePrice Y.

Hypothesis testing for β4

H0 : β4 = 0

Ha : β4 ≠ 0

- T-statistics for β4 = 41.2/1.856=22.199, and with P value is 2e-16. We reject the null hypothesis. TotalBsmtSF should be retained for explaining the model.

Hypothesis testing for β5

H0 : β5 = 0

Ha : β5 ≠ 0

- T-statistics for β5 = 55.181/1.856=22.199, and with P value is 2e-16. We reject the null hypothesis. GarageArea should be retained for explaining the model.

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

Response: SalePrice
               Df     Sum Sq    Mean Sq   F value Pr(>F)
TotalFloorSF    1 5.5903e+12 5.5903e+12 7613.0717 <2e-16 ***
OverallQual     1 1.5832e+12 1.5832e+12 2156.0589 <2e-16 ***
GrLivArea       1 2.3675e+08 2.3675e+08    0.3224 0.5702
TotalBsmtSF     1 4.9845e+11 4.9845e+11  678.8164 <2e-16 ***
GarageArea      1 1.5262e+11 1.5262e+11  207.8416 <2e-16 ***
Residuals    1975 1.4502e+12 7.3430e+08
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

c. Omnibus Model:

H0 : β1 = β2 = β3 = β4 = β5 = 0
Ha : βj ≠ 0, for at least one j, j = 1, 2, 3, 4, 5

SSR = 7.82E+12

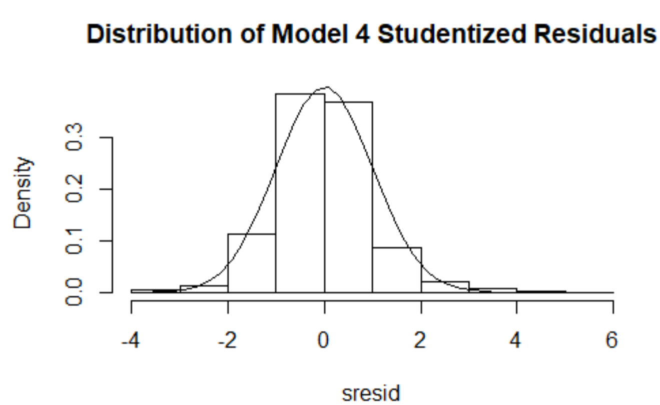
SSE = 1.45E+12

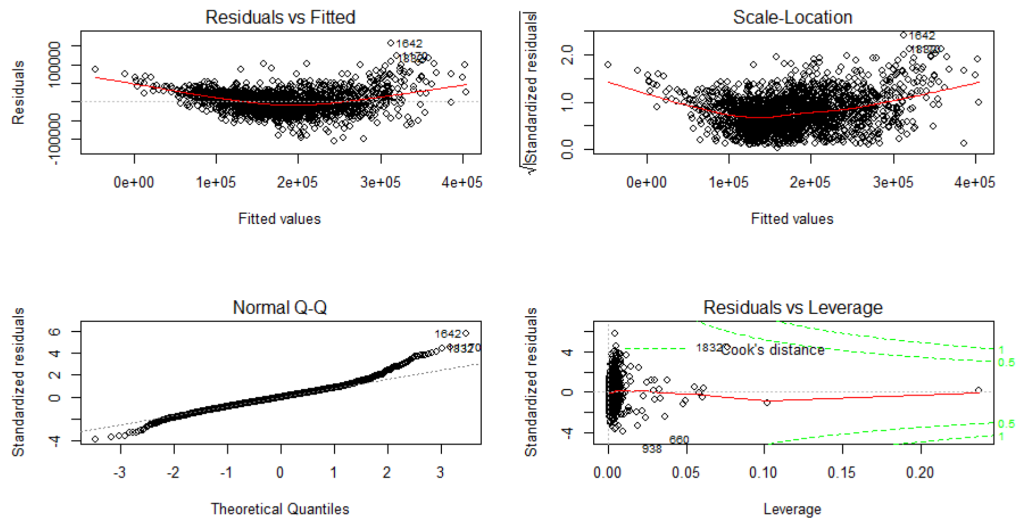
F0 = [SSR/p] / [SSE/(n − p − 1)] = MSR MSE ∼ Fp,n−p−1

= (7.82E+12/5)/(1.45E+12/(1975))=2131

- We reject the null hypothesis conclude that at least one of β1 or β2 or β3 or β4 or β5 is not equal to 0, with p-value 2.2e-16 These variables can explain 84.36 % of the model.

d. underlying assumption



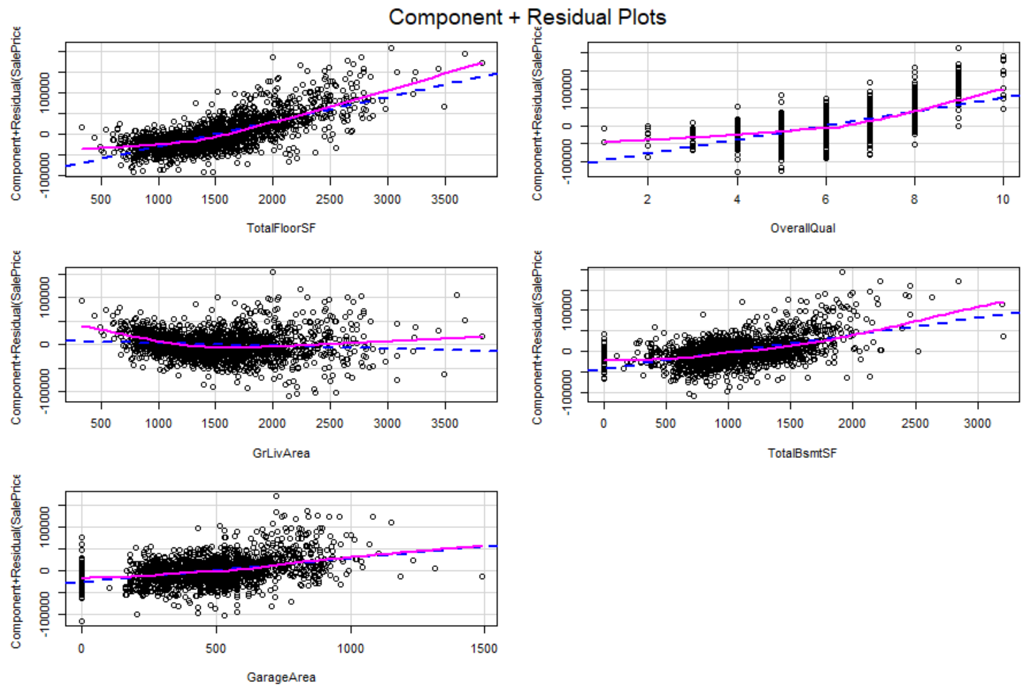


	rstudent	unadjusted p-value	Bonferroni p
1642	5.850294	5.7307e-09	1.1352e-05
1170	4.539246	5.9857e-06	1.1858e-02
1832	4.497761	7.2663e-06	1.4395e-02

- Model is normally distributed based on studentized residual histogram
- Linearity is violated given it's concaved in residuals vs Fitted graph. But residuals are little more distributed than model 1 except some extreme observations.
- Homoscedasticity is violated sine the scale location graph shows the concave curve of the line indicating the data points is not equally or randomly distributed.
- QQ line indicate distribution is slightly abnormal due to extreme observations
- Cook's distance suggests 3 extreme observations crossing the line.

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

Model	Y ~ X	R-Squared	Adj. R-Sqaured	RSE	Delta R-Squared
Model 1	SalePrice ~ TotalFloorSF	0.6027	0.6025	43150	0%
Model 2	SalePrice ~ OverallQual	0.6521	0.6519	40380	5%
Modle 3	SalePrice ~ TotalFloorSF + OverallQual	0.7734	0.7732	32600	12%
Model 4	SalePrice ~ TotalFloorSF + OverallQual + GrLivArea + TotalBsmtSF + GarageArea	0.8436	0.8432	27100	7%



- Based on R-Squared, we see 7% improvement. However, based on individual hypothesis testing "GrLivArea" is not statistically significant to the model, so even though we have an over 84.36% R-squared value explaining the total model, we would recommend excluding β3 from the model.
- CR plot further indicate "GrLivArea" does not show a strong linearity as other variables.

PART C: Multiple Linear Regression Models on Transformed Response Variable

6) Refit Model 1, Model 3 and Model 4 using the Natural Log of SALEPRICE as the response variable. 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 models to the original models. Which transformed model fits the best? Do the transformed models fit better than the original models? You do not need to report all of the output like was done in Parts A and B. Rather, you should construct a table to summarize your findings so that the comparisons can be made easily. What is the best way or statistic to use, to make comparisons between models? You may need more than one table to do this adequately, if you have more than 1 criteria.

Model 1 REFIT:

Model	Y ~ X	R-Squared	Adj. R-Sqaured	RSE
Model 1	SalePrice ~ TotalFloorSF	0.6027	0.6025	43150
Model 1b	logSalePrice ~ TotalFloorSF	0.6066	0.6064	0.2277

Model 3 REFIT

Model	Y ~ X	R-Squared	Adj. R-Sqaured	RSE
Modle 3	SalePrice ~ TotalFloorSF + OverallQual	0.7734	0.7732	32600
Modle 3b	logSalePrice ~ TotalFloorSF + OverallQual	0.7907	0.7905	0.1661

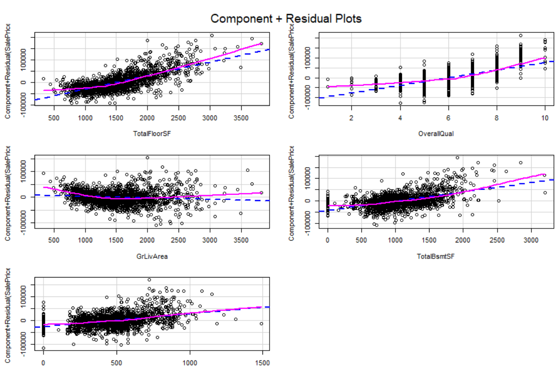
Model 4 REFIT:

Model	Y ~ X	R-Squared	Adj. R-Sqaured	RSE
Model 4	SalePrice ~ TotalFloorSF + OverallQual + GrLivArea + TotalBsmtSF + GarageArea	0.8436	0.8432	27100
Model 4b	logSalePrice ~ TotalFloorSF + OverallQual + GrLivArea + TotalBsmtSF + GarageArea	0.8524	0.852	0.1396

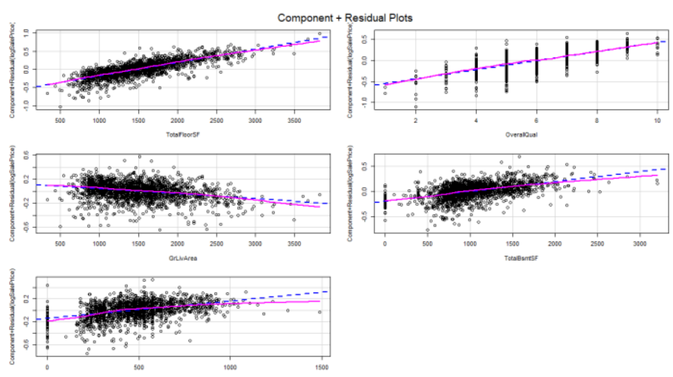
How is the interpretation of the LN(SalePrice) models different from the SalePrice models? Discuss if the improvement of model fit justifies the use of the Log(SALEPRRICE) response variable, relative to interpretation and explanation to a non-technical audience, like your manager or other executives.

- RSE is no longer a valid indicator for performance so we can simply rely on R-Squared value. Improvement across all 3 models during the refit, and we should use Log(SalePrice) as the response variable instead. Log function transforming skewed variable, such as our dataset has some extreme outliers, and normalize them. Side by side before and after chart using model 4 which has 5 variables, the linearity line is smoothed out more.

BEFORE using log



AFTER using log

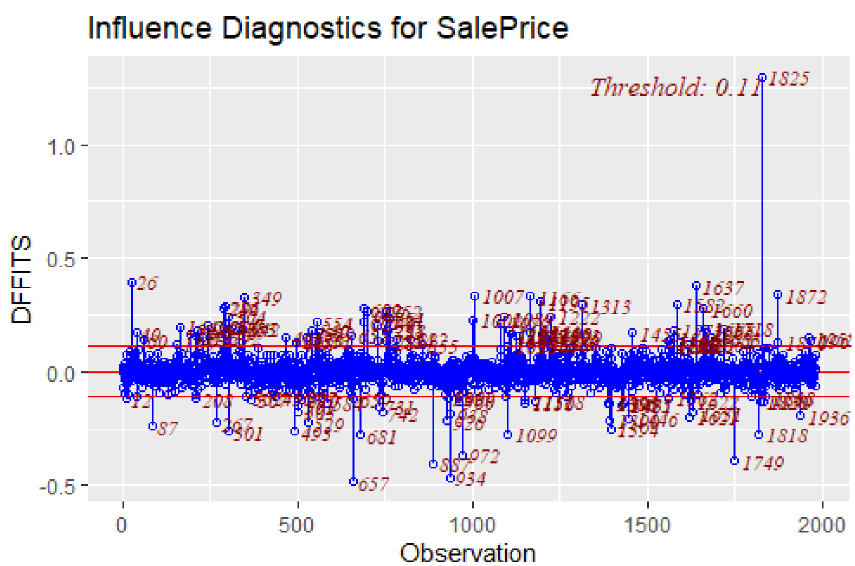


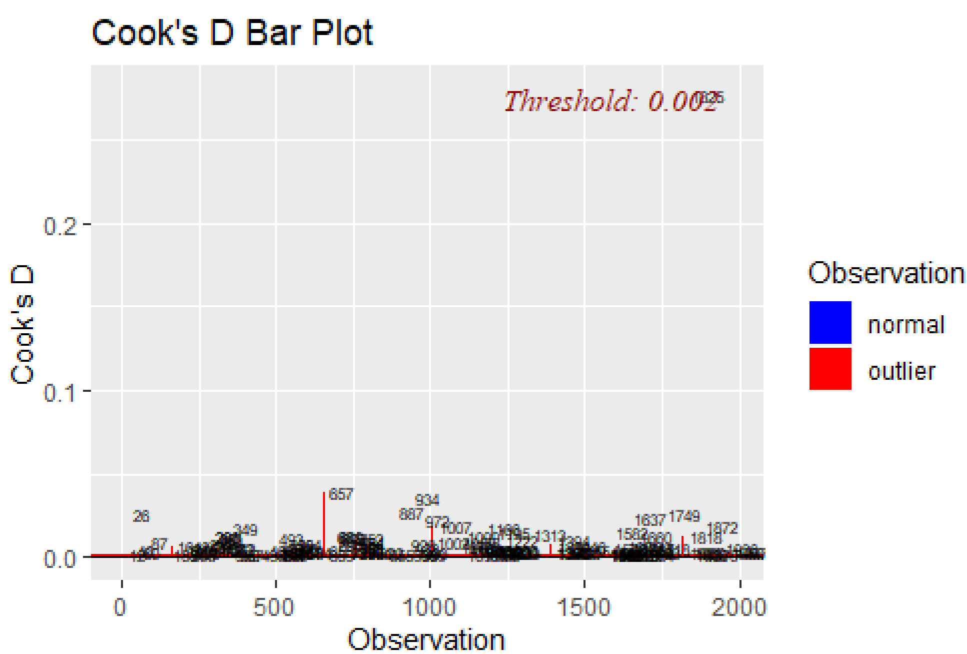
PART D: Multiple Linear Regression and Influential Points

7) Use Model 4 for this part. Even after you have cleaned your data, still you may have unusually large residuals, which you can see from the residual plots. These are called ‘influential’ points. Sometimes, we find that a small subset of ‘influential’ points exerts a disproportionate influence on the model coefficients. These points can be identified by several statistics such as DFFITS, Cook’s Distance, Leverage, and Influence. Fit Model 4 using a regression function from one of the comprehensive regression packages (like lessR).

Obtain output data with these statistics (DFFITS, etc.) for individual records so that you can identify the influential points. Use the threshold value given in the text book (Like that on Page 112 of Chatterjee and Hadi). 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? The other side of the coin is that if you remove data points due to them being "influential" and not looking like you might want them to look, some would argue that such an action is the modeler biasing the data. 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.

DFFITS threshold: 2*sqrt((p+1)/(n-p-1)) = 2*sqrt((5+1)/(1981-5-1)) =0.11

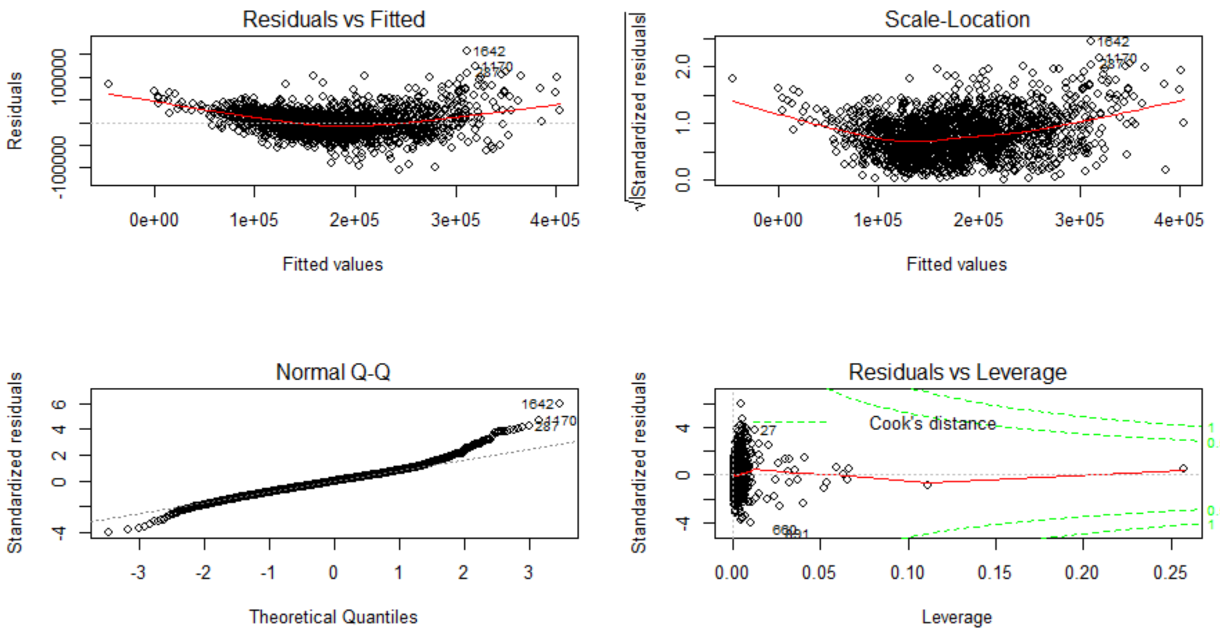


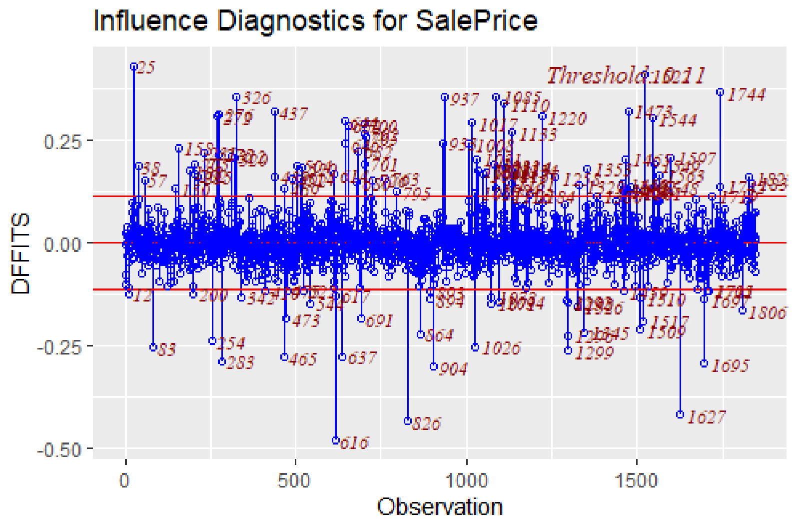


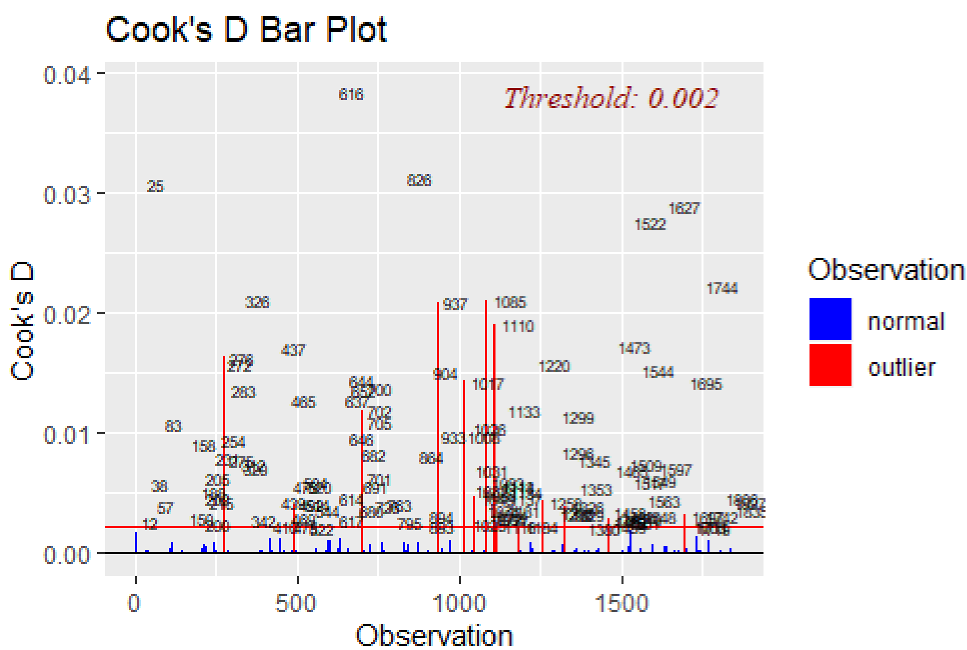
Using Model 4 for this evaluation running cook's distance and DFFITS. We see 1 influential point: 1825 while cook's SD identified 1832 as the influential point.

Model	Y ~ X	R-Squared	Adj. R-Sqaured	RSE
Model 4	SalePrice ~ TotalFloorSF + OverallQual + GrLivArea + TotalBsmtSF + GarageArea	0.8436	0.8432	27100
Model 4b	logSalePrice ~ TotalFloorSF + OverallQual + GrLivArea + TotalBsmtSF + GarageArea	0.8524	0.852	0.1396
Model 4c	SalePrice ~ TotalFloorSF + OverallQual + GrLivArea + TotalBsmtSF + GarageArea	0.8448	0.8444	26450

- Model 4c indicating the removal of one influential point 1825, which decreased our sample size by 133 records. (1981 to 1848 observations) This is 6.7% reduction of the data. Our rerun of the model suggests a 0.12% improvement than original Model 4 but decline of 0.76% in explaining the variance of the model comparing to using Log SalePrice as the target variable. Reviewing the underlying assumption post removal of the influential points did not yield assumption Homoscedasticity or Linearity.







- Post Cooks distance graph by removal of 1832 influential point, we see more arise as the influential point now, such as 660, and 891. While looking at the new DFFITS graph, more popping up.
- Based on these observation and comparison, I do not feel the removal of the influential point justifies the minimal improvement in our model. And we did not gain any more accuracy than prior to the cleanup.

PART E: Beginning to Think About a Final Model

8) Use Model 4 to start with for this part.
   a. Given the use of logSalePrice as the predictor, we were able to achieve better R-Squared value, I am using logSalePrice as my target variable for this analysis. Below is our performance benchmark

Model	Y ~ X	R-Squared	Adj. R-Sqaured	RSE
Model 4b	logSalePrice ~ TotalFloorSF + OverallQual + GrLivArea + TotalBsmtSF + GarageArea	0.8524	0.852	0.1396

- Since GrLivArea was not significant, we are removing it from this model and selecting additional 5. I will call this Model 5. Additional models and justification listed below.

	Variable	SalePrice	Reason
5	LotArea	0.299702741	**Appealing for expansion or build out**
9	YearRemodel	0.495729963	**Disclose any recent updates to the interior/exterior**
14	TotalBsmtSF	0.650020052	**High Correlation**
21	FullBath	0.609247762	**High Correlation**
26	Fireplaces	0.480994782	**Personal preference. Adds character and can be energy efficient**
28	GarageCars	0.660977076	**High Correlation**

- Now let β1 = TotalFloorSF, β2 = OverallQual, β3 = TotalBsmtSF, β4 = GarageArea, B5 = FullBath, β6= LotArea, B7 = Fireplaces, β8 = GarageCars, β9=YearRemodel

```
Call:
lm(formula = logSalePrice ~ TotalFloorSF + OverallQual + TotalBsmtSF +
    GarageArea + FullBath + LotArea + Fireplaces + GarageCars +
    YearRemodel, data = cleandata)

Residuals:
     Min       1Q   Median       3Q      Max
-0.70329 -0.06389  0.00902  0.08148  0.40594

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.174e+00  3.183e-01  16.253  < 2e-16 ***
TotalFloorSF 2.298e-04  9.223e-06  24.916  < 2e-16 ***
OverallQual  8.531e-02  3.373e-03  25.292  < 2e-16 ***
TotalBsmtSF  1.813e-04  8.676e-06  20.896  < 2e-16 ***
GarageArea   1.788e-04  2.998e-05   5.966 2.88e-09 ***
FullBath     9.372e-03  7.625e-03   1.229 0.219198
LotArea      3.259e-06  3.845e-07   8.475  < 2e-16 ***
Fireplaces   4.893e-02  5.127e-03   9.542  < 2e-16 ***
GarageCars   3.348e-02  8.780e-03   3.813 0.000141 ***
YearRemodel  2.818e-03  1.645e-04  17.128  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1246 on 1971 degrees of freedom
Multiple R-squared:  0.8826,    Adjusted R-squared:  0.8821
F-statistic:  1647 on 9 and 1971 DF,  p-value: < 2.2e-16
```

b. Report the model you determined and interpret the coefficients

- Y=5.174e+2.298e-04 β1+8.531e-02 β2+1.813e-04 β3+1.788e-04 β4+9.372e-03 β5+ 3.259e-06β6+4.893e-02 β7+3.348e-02 β8+2.818e-03 β9
- Per incremental of total floor sf adds 2.298e-04 to the log sale price.
- Per overall qual adds 8.531e-02 to the log sale price.
- Per incremental total basement sf adds 1.813e-02 to the log sale price.
- Per incremental of garage area adds 1.788e-04 to the log sale price.
- Per incremental full bathroom adds 9.3721e-03 to the log sale price.
- Per incremental lot area adds 3.2591e-06 to the log sale price.
- Per incremental of Fireplace adds 4.893e-02 to the log sale price.
- Per incremental garage car adds 3.348e-02 to the log sale price.
- Per incremental lot area adds 2.8182-03 to the log sale price.

Analysis of Variance Table

Response: logSalePrice

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
TotalFloorSF	1	158.134	158.134	10185.188	< 2.2e-16	***
OverallQual	1	47.994	47.994	3091.241	< 2.2e-16	***
TotalBsmtSF	1	11.463	11.463	738.309	< 2.2e-16	***
GarageArea	1	4.595	4.595	295.955	< 2.2e-16	***
FullBath	1	0.483	0.483	31.082	2.815e-08	***
LotArea	1	1.577	1.577	101.590	< 2.2e-16	***
Fireplaces	1	0.945	0.945	60.892	9.704e-15	***
GarageCars	1	0.352	0.352	22.692	2.041e-06	***
YearRemodel	1	4.555	4.555	293.354	< 2.2e-16	***
Residuals	1971	30.602	0.016			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

c. Report the coefficient and ANOVA tables.

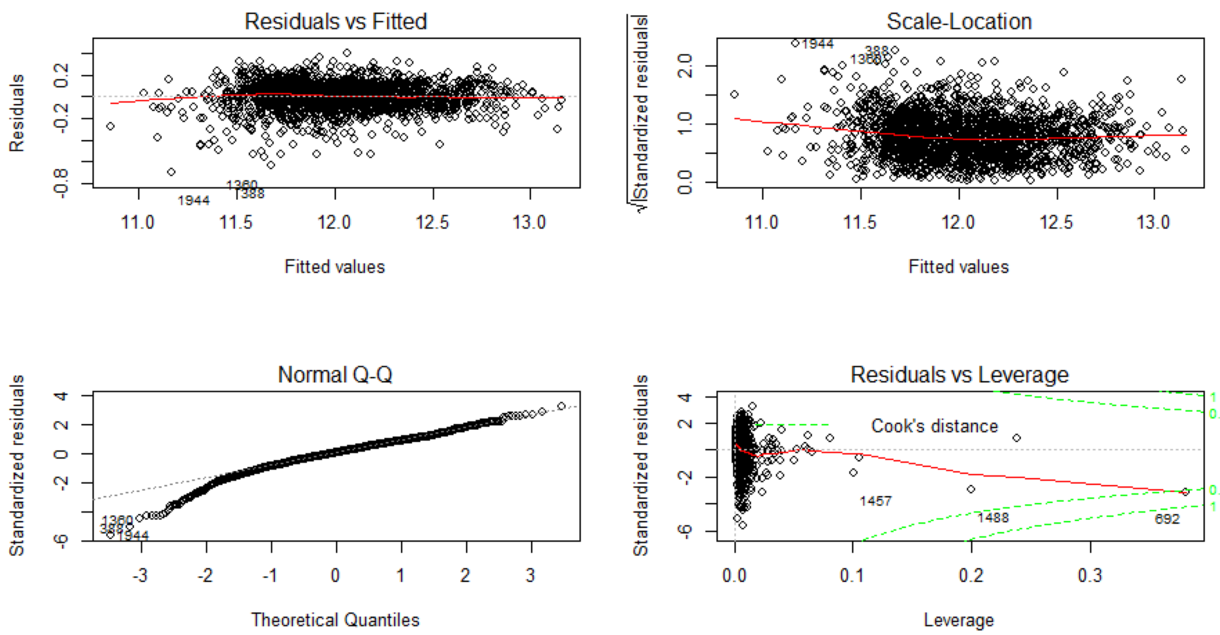
- Based on coefficients. β5 =FullBath, we fail to reject bull hypothesis. Rest of the variables show statistically significance. Combined can explain 88.26% of the total variance of the model.
- Omnibus test with F statistic at 1482 while P-value at <2.2e_16 indicates we reject null hypothesis.

d. Report goodness of fit

Goodness of fit using R-Squared, Adj R-Squared and RSE shows the 10 variables is improving our model by **3%,** and decrease RSE by **0.015.**

Model	Y ~ X	R-Squared	Adj. R-Sqaured	RSE
Model 4b	logSalePrice ~ TotalFloorSF + OverallQual + GrLivArea + TotalBsmtSF + GarageArea	0.8524	0.852	0.1396
Model 5	logSalePrice ~ TotalFloorSF+OverallQual+GrLivArea+TotalBsmtSF +GarageArea+FullBath+LotArea+Fireplaces+GarageCars+YearRemodel	0.8826	0.8821	0.1246

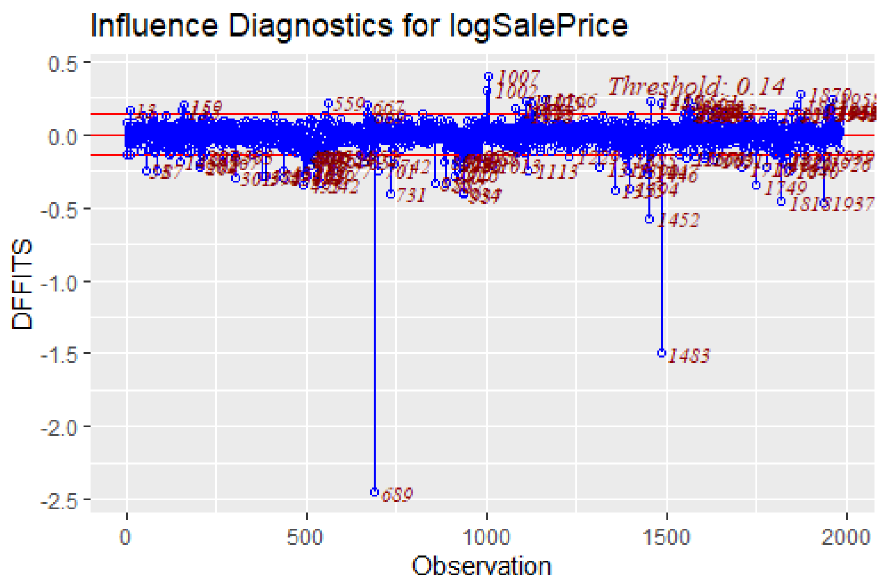
e. Check on underlying model assumptions.



- Residual vs fitted show that we have a good linearity, while some outliers seems consistent in our other plots for 1944, 1360, and 388. Linearity assumption is met.
- Homoscedasticity is violated since the line is curving down
- QQ plot shows extreme observations but otherwise aligned with 45-degree line, and shows the best fit comparing to previous models.

- Cook's distance shows influential points at 1457, 1488, 692

	rstudent	unadjusted p-value	Bonferroni p
1944	-5.707974	1.3169e-08	2.6087e-05
388	-5.138316	3.0461e-07	6.0342e-04
1360	-4.564425	5.3177e-06	1.0534e-02
1825	-4.317959	1.6530e-05	3.2746e-02
734	-4.314855	1.6762e-05	3.3205e-02
913	-4.297590	1.8108e-05	3.5872e-02
545	-4.292590	1.8517e-05	3.6682e-02



CONCLUSION / REFLECTION

- 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?

We evaluated many possible models for the Ames Housing data in this modeling assignment. And started from simple linear to multiple linear and variable transformation. The focus has been understating the underlying consumption by evaluating the residual distribution, linearity and homoscedasticity of the model through various approach while evaluating the model results by removing influential points.

- During our variable transformation, we saw a consistent improvement over all three models, but we need to be mindful on how to explain the results to non-technical managers.
- During the outlier identification, and removal process, we noticed decline in R-Squared value yet did not improve our underlying assumptions. Since the removal of influential points will

eliminate 6.7% of the total sample observation, I decided to not move forward with that approach. What I learned was that we need to balance various parameters and make the best decision based on the data we have on hand.

- Additional analysis was beneficial to keep us in mind that there are multiple evaluation needs to be conducted so we can understand bias and normality of our model instead of only chasing after R-Squared value.
- Hypothesis testing is a quick way to evaluate our selected variables especially when progressed to the final model. It's far more reliable than just the correlation chart. Some variables I selected had low correlation but prove significant in the model. I eliminated variable from previous testing and able to see how significant each variable is when explaining the total model. But we should understand other factors such as linearity, outliers, residual distribution, etc.

Next step

- Further evaluating the outliers, what are those records, should they be transformed?
- I initially dropped some variables had high % of missing values, perhaps we can apply different rules to impute and include them into the model
- Review categorical variables transform them into dummy variables?
- Learning a more automated fashion to select variable rather than doing it by hand. This will be useful when dealing with dataset that we can lean on based on intuition or expertise.