Exercise 12

Jyoti Sahu

01-29-2021

Housing Data

Problem Statement : Work individually on this assignment. You are encouraged to collaborate on ideas and strategies pertinent to this assignment. Data for this assignment is focused on real estate transactions recorded from 1964 to 2016 and can be found in Week 6 Housing.xlsx. Using your skills in statistical correlation, multiple regression and R programming, you are interested in the following variables: Sale Price and several other possible predictors.

Using your 'clean' data set from the previous week complete the following:

```
## Set the working directory to the root of your DSC 520 directory
setwd("/Users/sahujyot/Documents/DSC520")
## Load the `readxl` library
library(readxl)
## Load the `completed/Exercise 12/week-6-housing.xlsx` to
housing_df <- read_excel(path = 'Week 7/week-7-housing.xlsx' , skip = 0, sheet = 'Sheet2')
str(housing_df)
```

```
## tibble [12,865 x 24] (S3: tbl_df/tbl/data.frame)
##  $ Sale Date               : POSIXct[1:12865], format: "2006-01-03" "2006-01-03" ...
##  $ Sale Price              : num [1:12865] 698000 649990 572500 420000 369900 ...
##  $ sale_reason             : num [1:12865] 1 1 1 1 1 1 1 1 1 1 ...
##  $ sale_instrument         : num [1:12865] 3 3 3 3 3 15 3 3 3 3 ...
##  $ sale_warning            : chr [1:12865] NA NA NA NA ...
##  $ sitetype                : chr [1:12865] "R1" "R1" "R1" "R1" ...
##  $ addr_full               : chr [1:12865] "17021 NE 113TH CT" "11927 178TH PL NE" "13315 174TH AVE N
##  $ zip5                    : num [1:12865] 98052 98052 98052 98052 98052 ...
##  $ ctyname                 : chr [1:12865] "REDMOND" "REDMOND" NA "REDMOND" ...
##  $ postalctyn              : chr [1:12865] "REDMOND" "REDMOND" "REDMOND" "REDMOND" ...
##  $ lon                     : num [1:12865] -122 -122 -122 -122 -122 ...
##  $ lat                     : num [1:12865] 47.7 47.7 47.7 47.6 47.7 ...
##  $ building_grade          : num [1:12865] 9 9 8 8 7 7 10 10 9 8 ...
##  $ square_feet_total_living: num [1:12865] 2810 2880 2770 1620 1440 4160 3960 3720 4160 2760 ...
##  $ bedrooms                : num [1:12865] 4 4 4 3 3 4 5 4 4 4 ...
##  $ bath_full_count         : num [1:12865] 2 2 1 1 1 2 3 2 2 1 ...
##  $ bath_half_count         : num [1:12865] 1 0 1 0 0 1 0 1 1 0 ...
##  $ bath_3qtr_count         : num [1:12865] 0 1 1 1 1 1 1 0 1 1 ...
##  $ year_built              : num [1:12865] 2003 2006 1987 1968 1980 ...
##  $ year_renovated          : num [1:12865] 0 0 0 0 0 0 0 0 0 0 ...
##  $ current_zoning          : chr [1:12865] "R4" "R4" "R6" "R4" ...
##  $ sq_ft_lot               : num [1:12865] 6635 5570 8444 9600 7526 ...
```

```
##  $ prop_type              : chr [1:12865] "R" "R" "R" "R" ...
##  $ present_use            : num [1:12865] 2 2 2 2 2 2 2 2 2 2 ...
```

a. Explain why you chose to remove data points from your 'clean' dataset.

I have removed/not used below data points in my analysis. Sale Date- It does not affect housing price so there wont be any pattern. sale_reason-I do not think this will play a major role on house pricing. sale_instrument- do not think this will play a major role on house pricing. sale_warning- do not think this will play a major role on house pricing. sitetype- I do not think this will play a major role on house pricing.It is always great to know the area and criminal activities in deciding a house. AAlso this field is not useful in my current data set addr_full-I do not think this will play a major role on house pricing.It is always great to know the area and criminal activities in deciding a house. Also this field is not useful in my current data set. zip5-I do not think this will play a major role on house pricing.It is always great to know the area and criminal activities in deciding a house. Also this field is not useful in my current data set. ctyname-I do not think this will play a major role on house pricing.It is always great to know the area and criminal activities in deciding a house. Also this field is not useful in my current data set. postalctyn-I do not think this will play a major role on house pricing.It is always great to know the area and criminal activities in deciding a house. Also this field is not useful in my current data set. lon-I do not think this will play a major role on house pricing.It is always great to know the area and criminal activities in deciding a house. Also this field is not useful in my current data set. lat-I do not think this will play a major role on house pricing.It is always great to know the area and criminal activities in deciding a house. Also this field is not useful in my current data set. building_grade- It has very minimal effect on housing price current_zoning- Current does not matter here because it is all R (Residence) prop_type-This field does not matter here because it is all R (Residence) present_use- It does not matter.

```
summary(housing_df)
```

```
##    Sale Date                     Sale Price       sale_reason
##  Min.   :2006-01-03 00:00:00   Min.   :    698   Min.   : 0.00
##  1st Qu.:2008-07-07 00:00:00   1st Qu.: 460000   1st Qu.: 1.00
##  Median :2011-11-17 00:00:00   Median : 593000   Median : 1.00
##  Mean   :2011-07-28 15:07:32   Mean   : 660738   Mean   : 1.55
##  3rd Qu.:2014-06-05 00:00:00   3rd Qu.: 750000   3rd Qu.: 1.00
##  Max.   :2016-12-16 00:00:00   Max.   :4400000   Max.   :19.00
##  sale_instrument  sale_warning         sitetype          addr_full
##  Min.   : 0.000   Length:12865       Length:12865       Length:12865
##  1st Qu.: 3.000   Class :character   Class :character   Class :character
##  Median : 3.000   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 3.678
##  3rd Qu.: 3.000
##  Max.   :27.000
##       zip5         ctyname           postalctyn             lon
##  Min.   :98052   Length:12865       Length:12865       Min.   :-122.2
##  1st Qu.:98052   Class :character   Class :character   1st Qu.:-122.1
##  Median :98052   Mode  :character   Mode  :character   Median :-122.1
##  Mean   :98053                                         Mean   :-122.1
##  3rd Qu.:98053                                         3rd Qu.:-122.0
##  Max.   :98074                                         Max.   :-121.9
##       lat        building_grade  square_feet_total_living    bedrooms
##  Min.   :47.46   Min.   : 2.00   Min.   :  240            Min.   : 0.000
##  1st Qu.:47.67   1st Qu.: 8.00   1st Qu.: 1820            1st Qu.: 3.000
##  Median :47.69   Median : 8.00   Median : 2420            Median : 4.000
##  Mean   :47.68   Mean   : 8.24   Mean   : 2540            Mean   : 3.479
```

```
##  3rd Qu.:47.70   3rd Qu.: 9.00   3rd Qu.: 3110            3rd Qu.: 4.000
##  Max.   :47.73   Max.   :13.00   Max.   :13540            Max.   :11.000
##  bath_full_count  bath_half_count  bath_3qtr_count   year_built
##  Min.   : 0.000   Min.   :0.0000   Min.   :0.000   Min.   :1900
##  1st Qu.: 1.000   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:1979
##  Median : 2.000   Median :1.0000   Median :0.000   Median :1998
##  Mean   : 1.798   Mean   :0.6134   Mean   :0.494   Mean   :1993
##  3rd Qu.: 2.000   3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:2007
##  Max.   :23.000   Max.   :8.0000   Max.   :8.000   Max.   :2016
##  year_renovated    current_zoning       sq_ft_lot        prop_type
##  Min.   :   0.00   Length:12865       Min.   :    785   Length:12865
##  1st Qu.:   0.00   Class :character   1st Qu.:   5355   Class :character
##  Median :   0.00   Mode  :character   Median :   7965   Mode  :character
##  Mean   :  26.24                      Mean   :  22229
##  3rd Qu.:   0.00                      3rd Qu.:  12632
##  Max.   :2016.00                      Max.   :1631322
##   present_use
##  Min.   :  0.000
##  1st Qu.:  2.000
##  Median :  2.000
##  Mean   :  6.598
##  3rd Qu.:  2.000
##  Max.   :300.000
```

```
cleaned_housing_df <- subset(housing_df,select=c("Sale Price","square_feet_total_living","bedrooms","ba
```

```
summary(cleaned_housing_df)
```

```
##    Sale Price      square_feet_total_living    bedrooms      bath_full_count
##  Min.   :    698   Min.   :  240            Min.   : 0.000   Min.   : 0.000
##  1st Qu.: 460000   1st Qu.: 1820            1st Qu.: 3.000   1st Qu.: 1.000
##  Median : 593000   Median : 2420            Median : 4.000   Median : 2.000
##  Mean   : 660738   Mean   : 2540            Mean   : 3.479   Mean   : 1.798
##  3rd Qu.: 750000   3rd Qu.: 3110            3rd Qu.: 4.000   3rd Qu.: 2.000
##  Max.   :4400000   Max.   :13540            Max.   :11.000   Max.   :23.000
##  bath_half_count  bath_3qtr_count   year_built   year_renovated
##  Min.   :0.0000   Min.   :0.000   Min.   :1900   Min.   :   0.00
##  1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:1979   1st Qu.:   0.00
##  Median :1.0000   Median :0.000   Median :1998   Median :   0.00
##  Mean   :0.6134   Mean   :0.494   Mean   :1993   Mean   :  26.24
##  3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:2007   3rd Qu.:   0.00
##  Max.   :8.0000   Max.   :8.000   Max.   :2016   Max.   :2016.00
##    sq_ft_lot
##  Min.   :    785
##  1st Qu.:   5355
##  Median :   7965
##  Mean   :  22229
##  3rd Qu.:  12632
##  Max.   :1631322
```

b. Create two variables; one that will contain the variables Sale Price and Square Foot of Lot (same variables used from previous assignment on simple regression) and one that will contain Sale Price and several additional predictors of your choice. Explain the basis for your additional predictor selections.

```
# This is Simple Linear Regression Model
saleprice_slm <- lm(cleaned_housing_df$`Sale Price` ~ cleaned_housing_df$sq_ft_lot, cleaned_housing_df)
summary(saleprice_slm )
```

```
##
## Call:
## lm(formula = cleaned_housing_df$`Sale Price` ~ cleaned_housing_df$sq_ft_lot,
##     data = cleaned_housing_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2016064  -194842   -63293    91565  3735109
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  6.418e+05  3.800e+03  168.90   <2e-16 ***
## cleaned_housing_df$sq_ft_lot 8.510e-01  6.217e-02   13.69   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 401500 on 12863 degrees of freedom
## Multiple R-squared:  0.01435,    Adjusted R-squared:  0.01428
## F-statistic: 187.3 on 1 and 12863 DF,  p-value: < 2.2e-16
```

```
print("Calculating coefficient of other Predictos")
```

```
## [1] "Calculating coefficient of other Predictos"
```

```
print("Correlation of Sale Price and square_feet_total_living ")
```

```
## [1] "Correlation of Sale Price and square_feet_total_living "
```

```
cor(cleaned_housing_df$`Sale Price`,cleaned_housing_df$square_feet_total_living)
```

```
## [1] 0.4545876
```

```
print("Correlation of Sale Price and bedrooms")
```

```
## [1] "Correlation of Sale Price and bedrooms"
```

```
cor(cleaned_housing_df$`Sale Price`,cleaned_housing_df$bedrooms)
```

```
## [1] 0.2254675
```

```
print("Correlation of Sale Price and bath_full_count")
```

```
## [1] "Correlation of Sale Price and bath_full_count"
```

```
cor(cleaned_housing_df$`Sale Price`,cleaned_housing_df$bath_full_count)
```

```
## [1] 0.284849
```

```
print("Correlation of Sale Price and bath_half_count")
```

```
## [1] "Correlation of Sale Price and bath_half_count"
```

```
cor(cleaned_housing_df$`Sale Price`,cleaned_housing_df$bath_half_count)
```

```
## [1] 0.1658284
```

```
print("Correlation of Sale Price and bath_3qtr_count")
```

```
## [1] "Correlation of Sale Price and bath_3qtr_count"
```

```
cor(cleaned_housing_df$`Sale Price`,cleaned_housing_df$bath_3qtr_count)
```

```
## [1] 0.03574175
```

```
print("Correlation of Sale Price and year_built")
```

```
## [1] "Correlation of Sale Price and year_built"
```

```
cor(cleaned_housing_df$`Sale Price`,cleaned_housing_df$year_built)
```

```
## [1] 0.2426713
```

```
print("Correlation of Sale Price and year_renovated")
```

```
## [1] "Correlation of Sale Price and year_renovated"
```

```
cor(cleaned_housing_df$`Sale Price`,cleaned_housing_df$year_renovated)
```

```
## [1] 0.03286429
```

Based on the Correlation between Sales price and other variables, I am picking the fields with correlation over 0.2 because of strong relationship and feeding them into the model.

```
# This is Multiple Linear Regression Model
saleprice_mlm <- lm(cleaned_housing_df$`Sale Price` ~ cleaned_housing_df$square_feet_total_living + clea
```

c. Execute a summary() function on two variables defined in the previous step to compare the model results. What are the R2 and Adjusted R2 statistics? Explain what these results tell you about the overall model. Did the inclusion of the additional predictors help explain any large variations found in Sale Price?

```
summary(saleprice_slm)
```

```
##
## Call:
## lm(formula = cleaned_housing_df$`Sale Price` ~ cleaned_housing_df$sq_ft_lot,
##     data = cleaned_housing_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2016064  -194842   -63293    91565  3735109
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  6.418e+05  3.800e+03  168.90   <2e-16 ***
## cleaned_housing_df$sq_ft_lot 8.510e-01  6.217e-02   13.69   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 401500 on 12863 degrees of freedom
## Multiple R-squared:  0.01435,    Adjusted R-squared:  0.01428
## F-statistic: 187.3 on 1 and 12863 DF,  p-value: < 2.2e-16
```

```
summary(saleprice_mlm)
```

```
##
## Call:
## lm(formula = cleaned_housing_df$`Sale Price` ~ cleaned_housing_df$square_feet_total_living +
##     cleaned_housing_df$bedrooms + cleaned_housing_df$bath_full_count +
##     cleaned_housing_df$year_built)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1719151  -120511   -42398    45744  3904824
##
## Coefficients:
##                                           Estimate Std. Error t value
## (Intercept)                             -4.430e+06  4.195e+05 -10.559
## cleaned_housing_df$square_feet_total_living  1.744e+02  4.423e+00  39.424
## cleaned_housing_df$bedrooms             -1.375e+04  4.517e+03  -3.045
## cleaned_housing_df$bath_full_count       1.730e+04  6.095e+03   2.838
## cleaned_housing_df$year_built            2.340e+03  2.117e+02  11.053
##                                           Pr(>|t|)
```

```
## (Intercept)                                  < 2e-16 ***
## cleaned_housing_df$square_feet_total_living  < 2e-16 ***
## cleaned_housing_df$bedrooms                  0.00234 **
## cleaned_housing_df$bath_full_count           0.00454 **
## cleaned_housing_df$year_built                < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 357300 on 12860 degrees of freedom
## Multiple R-squared:  0.2194, Adjusted R-squared:  0.2192
## F-statistic: 903.7 on 4 and 12860 DF,  p-value: < 2.2e-16
```

The R2 of model tells us the prediction % of the model. Higher the R2 value, means better the Correlation coefficient, which is square root of R2.So based on the values from two models, the first model which has value of 0.01435 , which means square foot of the lot only contributes 1.4% to the sales price.However in the other model, other predictors together the R2 value is 0.2194 contribute approx 21% towards the sale price. However I noticed bedrooms is negatively co-related.

The Adjusted R2 gives an idea how well our model generalizes, and ideally we expect a similar value or close to R2.In both of our model the the difference is minimal which is a good sign For both of our models R2 and Adjusted R2 is very similar which indicates that cross-validity of the model is good.

d. Considering the parameters of the multiple regression model you have created. What are the standardized betas for each parameter and what do the values indicate?

```
library(QuantPsyc)
```

```
## Loading required package: boot
```

```
## Loading required package: MASS
```

```
##
## Attaching package: 'QuantPsyc'
```

```
## The following object is masked from 'package:base':
##
##     norm
```

```
lm.beta(saleprice_mlm)
```

```
## cleaned_housing_df$square_feet_total_living
##                                   0.42677620
##                 cleaned_housing_df$bedrooms
##                                  -0.02979645
##          cleaned_housing_df$bath_full_count
##                                   0.02783809
##               cleaned_housing_df$year_built
##                                   0.09966661
```

In general, it tells that if the specific attribute changes by one standard deviation, then the sales price(or outcome variable) increase by the Standardized Beta times(the value it displays) the standard deviation. If Beta is negative, it means decreases by same factor of Standard Deviation.

e. Calculate the confidence intervals for the parameters in your model and explain what the results indicate.

```
confint(saleprice_mlm)
```

```
##                                                   2.5 %       97.5 %
## (Intercept)                                -5252187.0182 -3607553.8714
## cleaned_housing_df$square_feet_total_living      165.6868      183.0244
## cleaned_housing_df$bedrooms                   -22607.0742    -4898.3341
## cleaned_housing_df$bath_full_count              5351.3294    29243.8018
## cleaned_housing_df$year_built                   1925.4259     2755.5146
```

From the confidence interval values here we can say that square_feet_total_living,bedrooms,bath_full_count and bath_half_count are on the same side of Zero. So these are fine.

The gap between square_feet_total_living is tight, so seems its estimates using this are more likely representing the true population. However the bedrooms, bath_full_count ,bedrooms are less representatives.

f. Assess the improvement of the new model compared to your original model (simple regression model) by testing whether this change is significant by performing an analysis of variance.

```
anova(saleprice_slm, saleprice_mlm)
```

```
## Analysis of Variance Table
##
## Model 1: cleaned_housing_df$`Sale Price` ~ cleaned_housing_df$sq_ft_lot
## Model 2: cleaned_housing_df$`Sale Price` ~ cleaned_housing_df$square_feet_total_living +
##     cleaned_housing_df$bedrooms + cleaned_housing_df$bath_full_count +
##     cleaned_housing_df$year_built
##   Res.Df        RSS Df  Sum of Sq    F    Pr(>F)
## 1  12863 2.0734e+15
## 2  12860 1.6420e+15  3 4.3134e+14 1126 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F(3,12860) = 1126 for p<0.001 So the Fit of the model has significantly improved from the original model.

g. Perform casewise diagnostics to identify outliers and/or influential cases, storing each function's output in a dataframe assigned to a unique variable name.

```
# Outliers
cleaned_housing_df$residuals <- resid(saleprice_mlm)
cleaned_housing_df$standardized.residuals <- rstandard(saleprice_mlm)
cleaned_housing_df$rstudent <- rstudent(saleprice_mlm)
# Influential Cases
```

```
cleaned_housing_df$cooks.distance <- cooks.distance(saleprice_mlm)
cleaned_housing_df$dfbeta <- dfbeta(saleprice_mlm)
cleaned_housing_df$dffits <- dffits(saleprice_mlm)
cleaned_housing_df$leverage <- hatvalues(saleprice_mlm)
cleaned_housing_df$covariance.ratios <- covratio(saleprice_mlm)
```

h. Calculate the standardized residuals using the appropriate command, specifying those that are +-2, storing the results of large residuals in a variable you create.

```
cleaned_housing_df$large.residuals<-cleaned_housing_df$standardized.residuals > 2 | cleaned_housing_df$s
```

i. Use the appropriate function to show the sum of large residuals.

```
sum(cleaned_housing_df$large.residuals)
```

```
## [1] 329
```

j. Which specific variables have large residuals (only cases that evaluate as TRUE)?

```
cleaned_housing_df[cleaned_housing_df$large.residuals, c("Sale Price", "square_feet_total_living", "bedi
```

```
## # A tibble: 329 x 6
##    `Sale Price` square_feet_tot~ bedrooms bath_full_count bath_half_count
##           <dbl>            <dbl>    <dbl>           <dbl>           <dbl>
##  1       184667             4160        4               2               1
##  2       265000             4920        4               4               1
##  3      1390000              660        0               1               0
##  4       390000             5800        5               4               1
##  5      1588359             3360        2               2               1
##  6      1450000              900        2               1               0
##  7       163000             4710        4               2               1
##  8       270000             5060        4              23               1
##  9       200000             6880        5               1               1
## 10       300000             4490        4               2               1
## # ... with 319 more rows, and 1 more variable: year_built <dbl>
```

k. Investigate further by calculating the leverage, cooks distance, and covariance rations. Comment on all cases that are problematics.

```
cleaned_housing_df[cleaned_housing_df$large.residuals, c("cooks.distance", "leverage", "covariance.ratio
```

```
## # A tibble: 329 x 3
##    cooks.distance leverage covariance.ratios
##             <dbl>    <dbl>             <dbl>
##  1       0.000328 0.000341             0.999
##  2       0.00143  0.00119              0.999
##  3       0.00359  0.00185              0.998
##  4       0.00162  0.00130              0.999
##  5       0.000570 0.000677             0.999
##  6       0.00471  0.00194              0.998
##  7       0.000836 0.000628             0.998
##  8       0.376    0.120                1.13
##  9       0.00700  0.00300              0.999
## 10       0.000407 0.000485             0.999
## # ... with 319 more rows
```

l. Perform the necessary calculations to assess the assumption of independence and state if the condition is met or not.

```
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:boot':
##
##     logit
```

```
durbinWatsonTest(saleprice_mlm)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.7210338     0.5579232       0
##  Alternative hypothesis: rho != 0
```

As per the Durbin Watson Test, if the values is in between 1-3, the model is considered good. Closer the value to 2, better the model. This model is bad

m. Perform the necessary calculations to assess the assumption of no multi-collinearity and state if the condition is met or not.

```
vif(saleprice_mlm)
```

```
## cleaned_housing_df$square_feet_total_living
##                                    1.930570
##                 cleaned_housing_df$bedrooms
##                                    1.577994
```

```
##          cleaned_housing_df$bath_full_count
##                                    1.584923
##               cleaned_housing_df$year_built
##                                    1.339428
```

```
print("Tolerance = 1/VIF")
```

```
## [1] "Tolerance = 1/VIF"
```

```
1/vif(saleprice_mlm)
```

```
## cleaned_housing_df$square_feet_total_living
##                                   0.5179818
##                 cleaned_housing_df$bedrooms
##                                   0.6337161
##          cleaned_housing_df$bath_full_count
##                                   0.6309454
##               cleaned_housing_df$year_built
##                                   0.7465875
```

```
print("Mean VIF")
```

```
## [1] "Mean VIF"
```

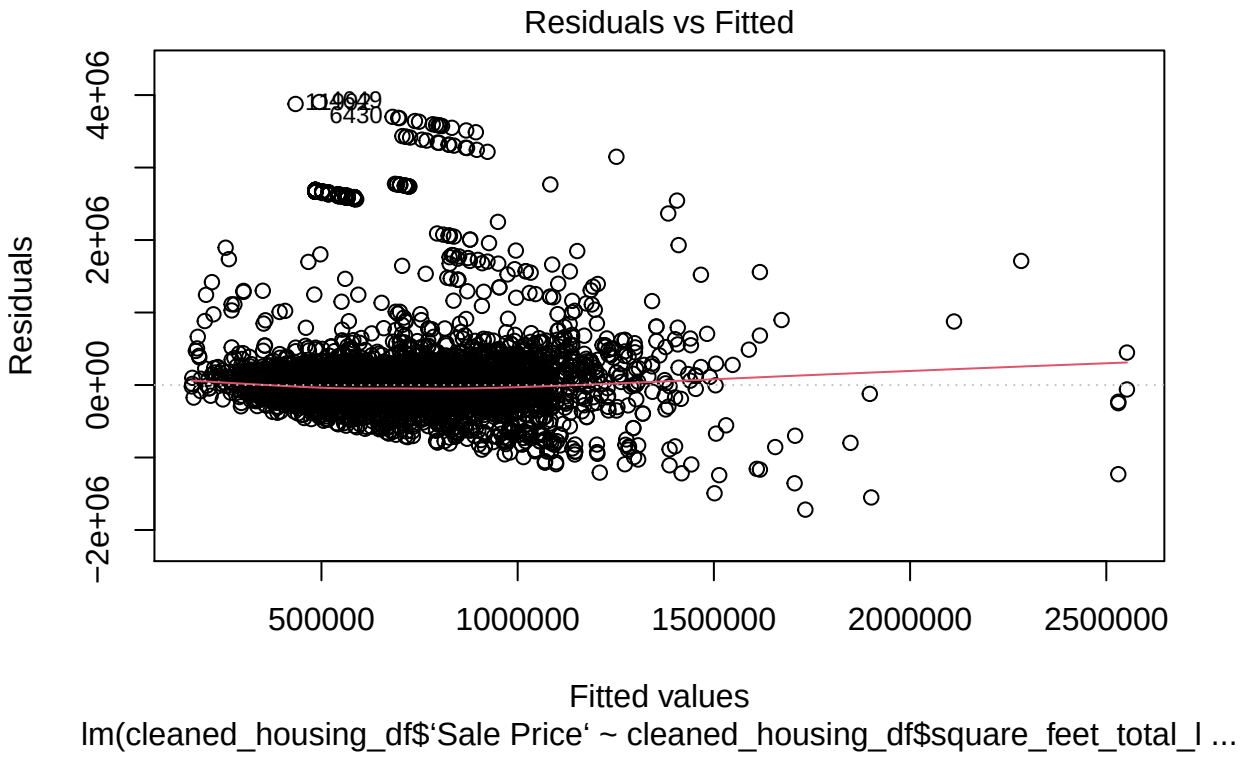
```
mean(vif(saleprice_mlm))
```

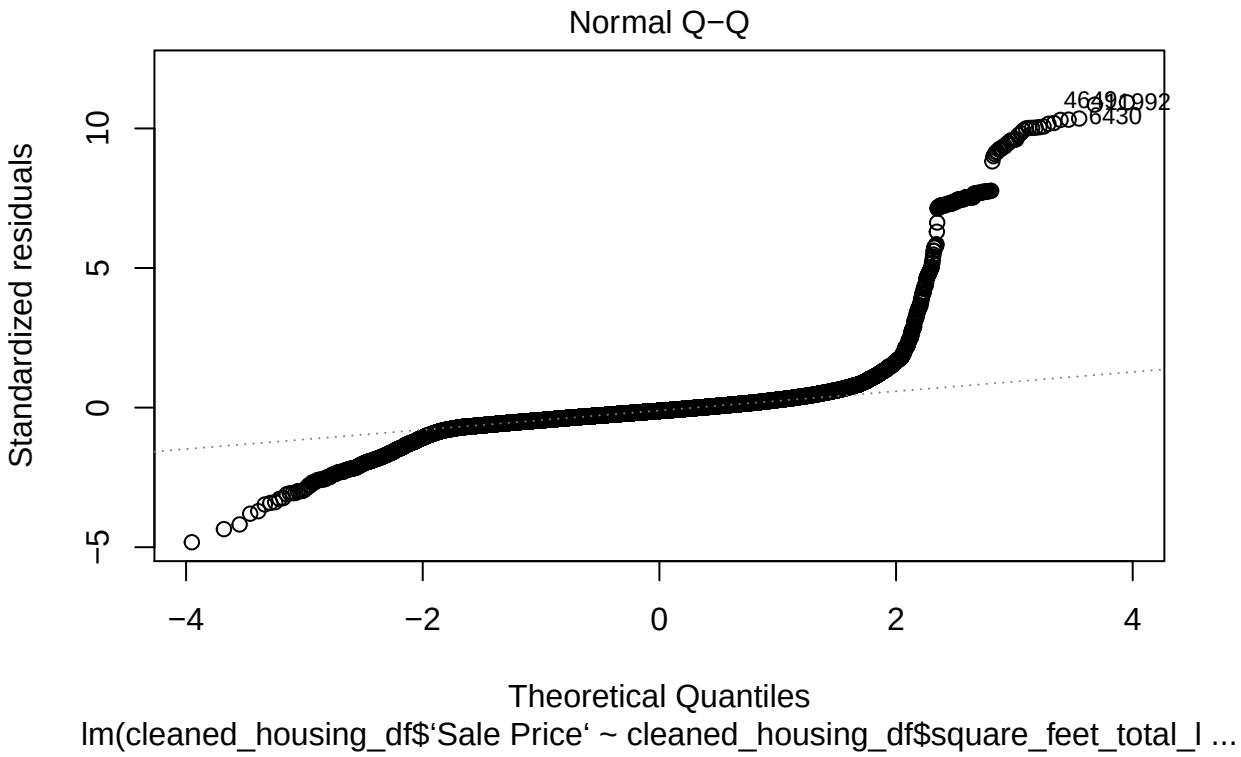
```
## [1] 1.608229
```

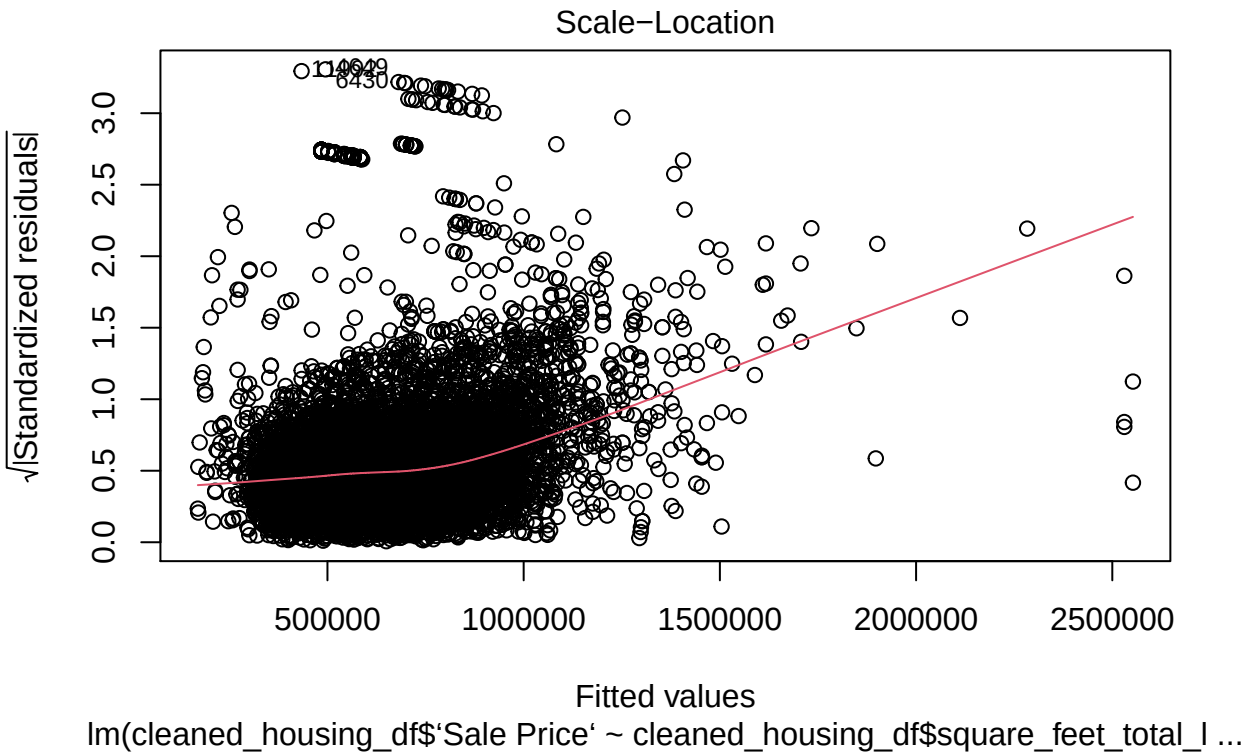
Is Largest VIF > 10 ? NO - So no cause for concern Avg VIF is 1.60, which is not substantially greater than 1. (Substantially more is considered more than 2.5, as from https://statisticalhorizons.com/multicollinearity) All Tolerance are above 0.2, meaning it should be fine.(Less than 0.2 is potential problem, less than 0.1 is significant problem. Its same as VIF >10, as tolerance $= 1/\text{VIF}$)

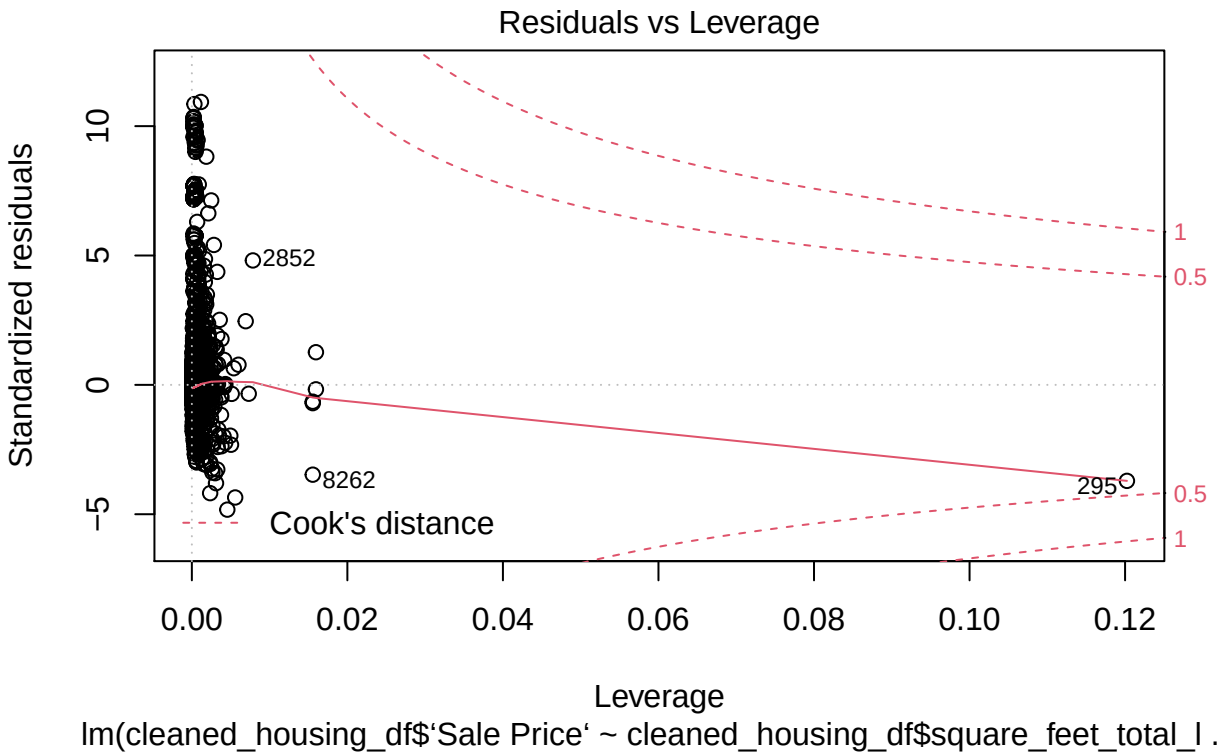
n. Visually check the assumptions related to the residuals using the plot() and hist() functions. Summarize what each graph is informing you of and if any anomalies are present.

```
plot(saleprice_mlm)
```


Residuals vs Fitted
Residuals
Fitted values
lm(cleaned_housing_df$‘Sale Price‘ ~ cleaned_housing_df$square_feet_total_l ...
4e+06
2e+06
0e+00
−2e+06
500000
1000000
1500000
2000000
2500000
6430


Normal Q−Q
Standardized residuals
10
5
0
−5
−4
−2
0
2
4
Theoretical Quantiles
lm(cleaned_housing_df$‘Sale Price‘ ~ cleaned_housing_df$square_feet_total_l ...

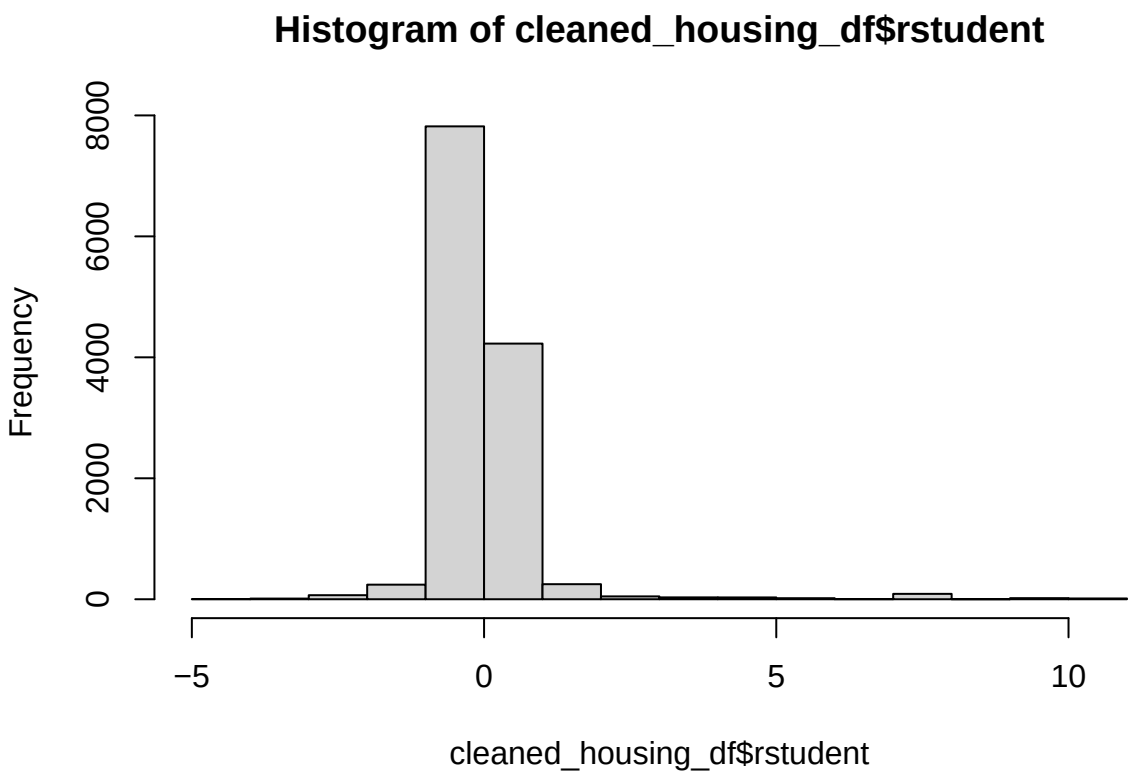

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
500000
1000000
1500000
2000000
2500000
Fitted values
lm(cleaned_housing_df$‘Sale Price‘ ~ cleaned_housing_df$square_feet_total_l ...



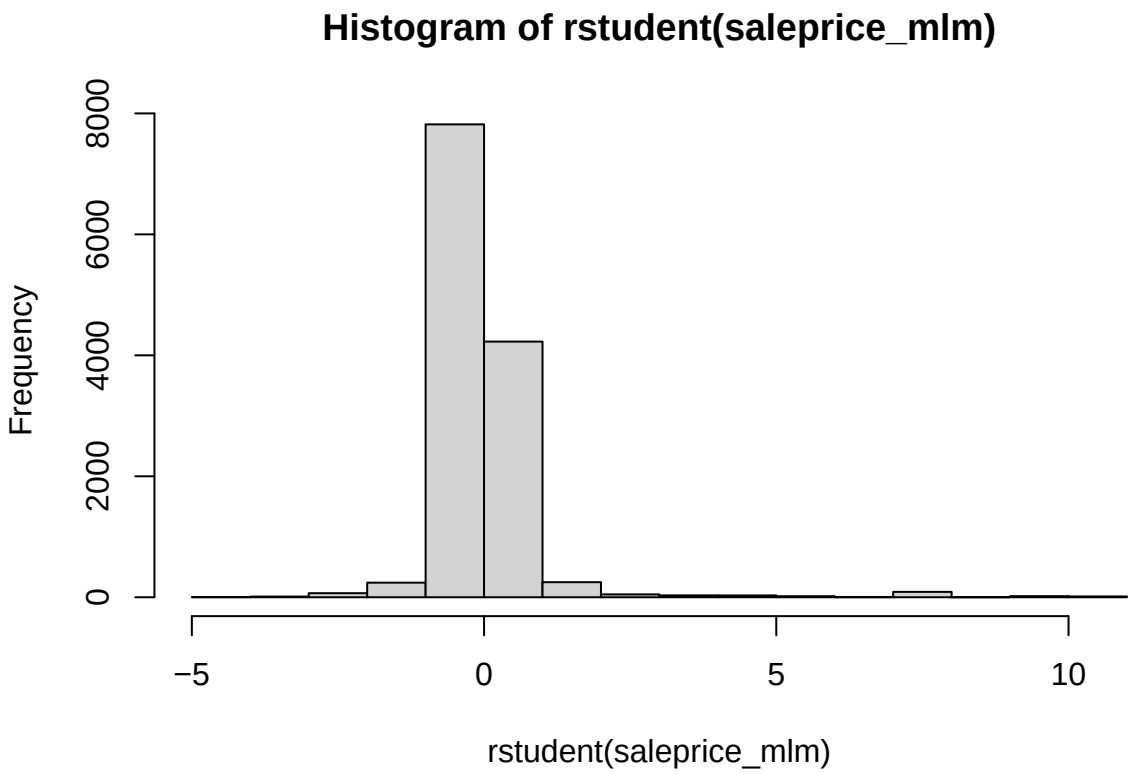
The Residuals Vs Fitted Graph shows random dots evenly dispersed around 0. Though not fully dispersed but evenly dispersed. It does not funnel out, so there is no heteroscedasticity.The data points also do not form a curve, so should be linear.

With the QQ plot we see that the plot curves of at extremes, so it means has more extreme values than would be expected if they truly came from a Normal distribution.

```
hist(cleaned_housing_df$rstudent)
```




```
hist(rstudent(saleprice_mlm))
```



Looks like a Bell slight skewed towards right. It could be assumed as normal.

o. Overall, is this regression model unbiased? If an unbiased regression model, what does this tell us about the sample vs. the entire population model?

We can say that we have bias present in this model due to following reason 1. The QQ plot that the plot curves away in opposite directions when approaching extreme values. This means there are outliers present at extremes. This tells that the model could be biased. 2. As we saw with year_built and bathrooms attribute the confint() output shows to affect the model in a bad way.

If the model is unbiased, it means that it holds true for both sample as well as it could be used confidently over the entire population.

To make this model better 1. We should try to clean the outliers based on the analysis so far. 2. We should also try to re-look at the parameters being used in the model. The one's which have bad effect on the model, should be removed. Additional parameters should also be added, if needed to improve the model. 3. We should try testing if these predictors are overfitting.