Exercise 12: Housing Data

Kolekar, Shilpa

October 18th, 2020

## 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:

library('readxl')

## Warning: package 'readxl' was built under R version 4.0.3

## Set the working directory to the root of your DSC 520 directory  
setwd("C:/Users/shilp/Documents/GitHub/dsc520")  
  
## Load the `data/week-7-housing.xlsx` to  
housing\_df <- read\_excel("data/week-7-housing.xlsx")

## a. Explain why you chose to remove data points from your ‘clean’ dataset.

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 NE" "3303 178TH AVE NE" ...  
## $ 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 ...

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

updated\_housing\_df <- housing\_df[(is.na(housing\_df$sale\_warning)),]  
updated\_housing\_df$`Sale Date` <- NULL  
updated\_housing\_df$sale\_warning <- NULL  
updated\_housing\_df$sitetype <- NULL  
updated\_housing\_df$addr\_full <- NULL  
updated\_housing\_df$ctyname <- NULL  
updated\_housing\_df$postalctyn <- NULL  
updated\_housing\_df$current\_zoning <- NULL  
updated\_housing\_df$prop\_type <- NULL  
summary(updated\_housing\_df)

## Sale Price sale\_reason sale\_instrument zip5   
## Min. : 2500 Min. : 0.000 Min. : 0.000 Min. :98052   
## 1st Qu.: 485075 1st Qu.: 1.000 1st Qu.: 3.000 1st Qu.:98052   
## Median : 605000 Median : 1.000 Median : 3.000 Median :98052   
## Mean : 645051 Mean : 1.107 Mean : 3.147 Mean :98053   
## 3rd Qu.: 749950 3rd Qu.: 1.000 3rd Qu.: 3.000 3rd Qu.:98053   
## Max. :4311000 Max. :18.000 Max. :26.000 Max. :98074   
## lon lat building\_grade square\_feet\_total\_living  
## Min. :-122.2 Min. :47.46 Min. : 2.000 Min. : 240   
## 1st Qu.:-122.1 1st Qu.:47.67 1st Qu.: 8.000 1st Qu.: 1870   
## Median :-122.1 Median :47.69 Median : 8.000 Median : 2450   
## Mean :-122.1 Mean :47.68 Mean : 8.273 Mean : 2545   
## 3rd Qu.:-122.0 3rd Qu.:47.71 3rd Qu.: 9.000 3rd Qu.: 3110   
## Max. :-121.9 Max. :47.73 Max. :13.000 Max. :13540   
## bedrooms bath\_full\_count bath\_half\_count bath\_3qtr\_count   
## Min. : 0.000 Min. : 0.000 Min. :0.0000 Min. :0.0000   
## 1st Qu.: 3.000 1st Qu.: 1.000 1st Qu.:0.0000 1st Qu.:0.0000   
## Median : 4.000 Median : 2.000 Median :1.0000 Median :0.0000   
## Mean : 3.482 Mean : 1.803 Mean :0.6175 Mean :0.5006   
## 3rd Qu.: 4.000 3rd Qu.: 2.000 3rd Qu.:1.0000 3rd Qu.:1.0000   
## Max. :11.000 Max. :23.000 Max. :6.0000 Max. :8.0000   
## year\_built year\_renovated sq\_ft\_lot present\_use   
## Min. :1900 Min. : 0.00 Min. : 785 Min. : 0.000   
## 1st Qu.:1980 1st Qu.: 0.00 1st Qu.: 5400 1st Qu.: 2.000   
## Median :1999 Median : 0.00 Median : 7850 Median : 2.000   
## Mean :1993 Mean : 21.93 Mean : 19921 Mean : 6.546   
## 3rd Qu.:2007 3rd Qu.: 0.00 3rd Qu.: 12037 3rd Qu.: 2.000   
## Max. :2016 Max. :2016.00 Max. :1631322 Max. :300.000

sale\_warning has total 12865 records and 10568 are blank records. Sale\_warning variable with values are removed. This variable can impact the sale price and if these values are not handled properly it can skew the data. I also removed all variables with non-numeric values.

## 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.

saleprice\_simple\_lm <- lm(updated\_housing\_df$`Sale Price` ~ updated\_housing\_df$sq\_ft\_lot, data = updated\_housing\_df)  
cor(updated\_housing\_df)

## Sale Price sale\_reason sale\_instrument  
## Sale Price 1.000000000 -0.032484121 -0.066754031  
## sale\_reason -0.032484121 1.000000000 0.219170195  
## sale\_instrument -0.066754031 0.219170195 1.000000000  
## zip5 0.072773893 -0.007221279 -0.007500265  
## lon 0.042330262 -0.003100288 -0.001947785  
## lat -0.006475514 0.018608422 -0.036816103  
## building\_grade 0.648578055 -0.028983798 -0.010848380  
## square\_feet\_total\_living 0.707278014 -0.026791780 0.017724281  
## bedrooms 0.329989828 -0.030730733 0.012844784  
## bath\_full\_count 0.382787449 -0.026845156 0.009283723  
## bath\_half\_count 0.224632618 -0.002758342 -0.007071653  
## bath\_3qtr\_count 0.097513036 -0.005903850 0.010421069  
## year\_built 0.259561628 -0.030594032 -0.028779557  
## year\_renovated 0.057477945 0.003659769 0.018618139  
## sq\_ft\_lot 0.240818204 -0.001948701 0.111168649  
## present\_use -0.008513231 -0.007936924 -0.000267675  
## zip5 lon lat building\_grade  
## Sale Price 0.0727738929 0.042330262 -0.006475514 0.64857806  
## sale\_reason -0.0072212795 -0.003100288 0.018608422 -0.02898380  
## sale\_instrument -0.0075002650 -0.001947785 -0.036816103 -0.01084838  
## zip5 1.0000000000 0.366373965 -0.122924853 0.09003670  
## lon 0.3663739654 1.000000000 -0.022525285 0.04890966  
## lat -0.1229248529 -0.022525285 1.000000000 0.03511233  
## building\_grade 0.0900367046 0.048909660 0.035112335 1.00000000  
## square\_feet\_total\_living 0.0873212955 0.081108514 -0.036480952 0.73421133  
## bedrooms -0.0428492907 -0.243051692 -0.164502449 0.33090837  
## bath\_full\_count 0.0981772679 0.137495966 0.086821837 0.44702861  
## bath\_half\_count 0.0139185642 0.013491147 -0.036318807 0.25706207  
## bath\_3qtr\_count -0.0708690768 -0.124931372 -0.094327829 0.04213351  
## year\_built 0.1062900121 0.347082919 0.404863606 0.36719622  
## year\_renovated -0.0005502049 -0.019477828 -0.093690857 -0.02124717  
## sq\_ft\_lot 0.0916982417 0.222732461 -0.164298940 0.17856326  
## present\_use 0.0344609753 0.134444427 0.029950288 0.02562454  
## square\_feet\_total\_living bedrooms bath\_full\_count  
## Sale Price 0.70727801 0.329989828 0.382787449  
## sale\_reason -0.02679178 -0.030730733 -0.026845156  
## sale\_instrument 0.01772428 0.012844784 0.009283723  
## zip5 0.08732130 -0.042849291 0.098177268  
## lon 0.08110851 -0.243051692 0.137495966  
## lat -0.03648095 -0.164502449 0.086821837  
## building\_grade 0.73421133 0.330908367 0.447028607  
## square\_feet\_total\_living 1.00000000 0.580316824 0.509873836  
## bedrooms 0.58031682 1.000000000 0.280253412  
## bath\_full\_count 0.50987384 0.280253412 1.000000000  
## bath\_half\_count 0.28533595 0.130034380 0.197936875  
## bath\_3qtr\_count 0.17614365 0.242104801 -0.395605314  
## year\_built 0.33191015 0.005438607 0.457758585  
## year\_renovated 0.03395984 0.015847783 0.021204393  
## sq\_ft\_lot 0.25554914 0.063734370 0.060729984  
## present\_use 0.01636895 -0.044648587 0.033406417  
## bath\_half\_count bath\_3qtr\_count year\_built  
## Sale Price 0.224632618 0.097513036 0.259561628  
## sale\_reason -0.002758342 -0.005903850 -0.030594032  
## sale\_instrument -0.007071653 0.010421069 -0.028779557  
## zip5 0.013918564 -0.070869077 0.106290012  
## lon 0.013491147 -0.124931372 0.347082919  
## lat -0.036318807 -0.094327829 0.404863606  
## building\_grade 0.257062074 0.042133508 0.367196219  
## square\_feet\_total\_living 0.285335946 0.176143648 0.331910151  
## bedrooms 0.130034380 0.242104801 0.005438607  
## bath\_full\_count 0.197936875 -0.395605314 0.457758585  
## bath\_half\_count 1.000000000 -0.357769366 0.181815485  
## bath\_3qtr\_count -0.357769366 1.000000000 -0.155081892  
## year\_built 0.181815485 -0.155081892 1.000000000  
## year\_renovated -0.027459627 0.019532116 -0.214618861  
## sq\_ft\_lot 0.048702612 0.049523348 -0.090627512  
## present\_use 0.008084105 -0.008641109 0.129866805  
## year\_renovated sq\_ft\_lot present\_use  
## Sale Price 0.0574779452 0.240818204 -0.008513231  
## sale\_reason 0.0036597689 -0.001948701 -0.007936924  
## sale\_instrument 0.0186181388 0.111168649 -0.000267675  
## zip5 -0.0005502049 0.091698242 0.034460975  
## lon -0.0194778280 0.222732461 0.134444427  
## lat -0.0936908572 -0.164298940 0.029950288  
## building\_grade -0.0212471750 0.178563260 0.025624544  
## square\_feet\_total\_living 0.0339598434 0.255549144 0.016368948  
## bedrooms 0.0158477832 0.063734370 -0.044648587  
## bath\_full\_count 0.0212043929 0.060729984 0.033406417  
## bath\_half\_count -0.0274596271 0.048702612 0.008084105  
## bath\_3qtr\_count 0.0195321156 0.049523348 -0.008641109  
## year\_built -0.2146188606 -0.090627512 0.129866805  
## year\_renovated 1.0000000000 0.048806331 -0.015991754  
## sq\_ft\_lot 0.0488063310 1.000000000 0.055016811  
## present\_use -0.0159917537 0.055016811 1.000000000

saleprice\_multi\_lm <- lm(updated\_housing\_df$`Sale Price` ~ updated\_housing\_df$sq\_ft\_lot + updated\_housing\_df$building\_grade + updated\_housing\_df$square\_feet\_total\_living + updated\_housing\_df$bedrooms + updated\_housing\_df$bath\_full\_count + updated\_housing\_df$bath\_half\_count + updated\_housing\_df$year\_built, data = updated\_housing\_df)

After calculating the correlation between sale price and other variables, there are few variables that have strong correlation than other variables. I picked these variables as predictors: sq\_ft\_lot, building\_grade, square\_feet\_total\_living, bedrooms, bath\_full\_count, bath\_half\_count, and year\_built co create multiple regression model.

## 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\_simple\_lm)

##   
## Call:  
## lm(formula = updated\_housing\_df$`Sale Price` ~ updated\_housing\_df$sq\_ft\_lot,   
## data = updated\_housing\_df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2615922 -151493 -35572 106230 3293158   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.205e+05 2.598e+03 238.9 <2e-16 \*\*\*  
## updated\_housing\_df$sq\_ft\_lot 1.232e+00 4.830e-02 25.5 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 248100 on 10566 degrees of freedom  
## Multiple R-squared: 0.05799, Adjusted R-squared: 0.0579   
## F-statistic: 650.5 on 1 and 10566 DF, p-value: < 2.2e-16

summary(saleprice\_multi\_lm)

##   
## Call:  
## lm(formula = updated\_housing\_df$`Sale Price` ~ updated\_housing\_df$sq\_ft\_lot +   
## updated\_housing\_df$building\_grade + updated\_housing\_df$square\_feet\_total\_living +   
## updated\_housing\_df$bedrooms + updated\_housing\_df$bath\_full\_count +   
## updated\_housing\_df$bath\_half\_count + updated\_housing\_df$year\_built,   
## data = updated\_housing\_df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2231685 -79883 -13674 61708 3676548   
##   
## Coefficients:  
## Estimate Std. Error t value  
## (Intercept) 4.300e+05 2.416e+05 1.780  
## updated\_housing\_df$sq\_ft\_lot 2.918e-01 3.588e-02 8.133  
## updated\_housing\_df$building\_grade 6.520e+04 2.436e+03 26.762  
## updated\_housing\_df$square\_feet\_total\_living 1.476e+02 3.375e+00 43.739  
## updated\_housing\_df$bedrooms -2.371e+04 2.506e+03 -9.461  
## updated\_housing\_df$bath\_full\_count 5.889e+03 3.288e+03 1.791  
## updated\_housing\_df$bath\_half\_count 5.228e+03 3.428e+03 1.525  
## updated\_housing\_df$year\_built -3.196e+02 1.226e+02 -2.606  
## Pr(>|t|)   
## (Intercept) 0.07516 .   
## updated\_housing\_df$sq\_ft\_lot 4.65e-16 \*\*\*  
## updated\_housing\_df$building\_grade < 2e-16 \*\*\*  
## updated\_housing\_df$square\_feet\_total\_living < 2e-16 \*\*\*  
## updated\_housing\_df$bedrooms < 2e-16 \*\*\*  
## updated\_housing\_df$bath\_full\_count 0.07331 .   
## updated\_housing\_df$bath\_half\_count 0.12729   
## updated\_housing\_df$year\_built 0.00917 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 172500 on 10560 degrees of freedom  
## Multiple R-squared: 0.5447, Adjusted R-squared: 0.5444   
## F-statistic: 1805 on 7 and 10560 DF, p-value: < 2.2e-16

For the simple regression model, the value of R2 is 0.05799. This indicates that the sq\_ft\_lot accounted for only 5.80% of the variation in sale price. The value of adjusted R2 is 0.0579 which is very close to R2 value, and that indicates that the sample is a good representation of population.

For the multiple regression model, the value of R2 is 0.5447. This indicates that the the model with multiple predictors accounted for 54.47% of the variation in sale price. The value of adjusted R2 is 0.5444 which is very close to R2 value, and that indicates that the sample is a good representation of population.

The prediction percentage went up from 5.80% to 54.47% which indicates that the sale price can be better predicted with the multiple predictors than only with sq\_ft\_lot variable.

## 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')

## Warning: package 'QuantPsyc' was built under R version 4.0.3

## Loading required package: boot

## Loading required package: MASS

##   
## Attaching package: 'QuantPsyc'

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

lm.beta(saleprice\_multi\_lm)

## updated\_housing\_df$sq\_ft\_lot   
## 0.05704603   
## updated\_housing\_df$building\_grade   
## 0.26710418   
## updated\_housing\_df$square\_feet\_total\_living   
## 0.53975845   
## updated\_housing\_df$bedrooms   
## -0.08064082   
## updated\_housing\_df$bath\_full\_count   
## 0.01470705   
## updated\_housing\_df$bath\_half\_count   
## 0.01051910   
## updated\_housing\_df$year\_built   
## -0.02070563

The beta value tells us the number of standard deviations by which the outcome will change as a result of one standard deviation of change in the predictor. Based on the standardized beta values for predictors, it looks like building\_grade and square\_feet\_total\_living are the only important predictors since they have comparable degree of importance in the model. Other predictors (sq\_ft\_lot, bedrooms, bath\_full\_count, bath\_half\_count, and year\_built) do not have comparable degree of importance.

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

confint(saleprice\_multi\_lm)

## 2.5 % 97.5 %  
## (Intercept) -4.362094e+04 903589.851212  
## updated\_housing\_df$sq\_ft\_lot 2.214616e-01 0.362111  
## updated\_housing\_df$building\_grade 6.042284e+04 69973.798616  
## updated\_housing\_df$square\_feet\_total\_living 1.409873e+02 154.217073  
## updated\_housing\_df$bedrooms -2.862264e+04 -18798.069869  
## updated\_housing\_df$bath\_full\_count -5.560603e+02 12334.690905  
## updated\_housing\_df$bath\_half\_count -1.492102e+03 11948.909001  
## updated\_housing\_df$year\_built -5.599320e+02 -79.225681

Small confidence interval indicates that the value of beta in the sample is close to the true value of the beta in the population. The positive or negative sign indicates the direction of the relationship between the predictor and the outcome. If the confidence interval crosses zero, then that is a sign of a very bad model.

sq\_ft\_lot, building\_grade, square\_feet\_total\_living, bedrooms, and year\_built have the confidence interval on one side of zero, which is good. sq\_ft\_lot and square\_feet\_total\_living have tight gap, so their estimates seem to be more likely true representatives of population. building\_grade, bedrooms, and year\_built are less representative of the population.

bath\_full\_count and bath\_half\_count are bad predictors in the model as the confidence interval for them crosses zero.

## 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\_simple\_lm, saleprice\_multi\_lm)

## Analysis of Variance Table  
##   
## Model 1: updated\_housing\_df$`Sale Price` ~ updated\_housing\_df$sq\_ft\_lot  
## Model 2: updated\_housing\_df$`Sale Price` ~ updated\_housing\_df$sq\_ft\_lot +   
## updated\_housing\_df$building\_grade + updated\_housing\_df$square\_feet\_total\_living +   
## updated\_housing\_df$bedrooms + updated\_housing\_df$bath\_full\_count +   
## updated\_housing\_df$bath\_half\_count + updated\_housing\_df$year\_built  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 10566 6.5023e+14   
## 2 10560 3.1425e+14 6 3.3598e+14 1881.7 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

F(6, 10560) = 1881.7 with p < 0.001. This indicates that the multiple regression model significantly improved the fit of the 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  
updated\_housing\_df$residuals <- resid(saleprice\_multi\_lm)  
updated\_housing\_df$standardized.residuals <- rstandard(saleprice\_multi\_lm)  
updated\_housing\_df$studentized.residuals <- rstudent(saleprice\_multi\_lm)  
  
# Influential cases  
updated\_housing\_df$cooks.distance <- cooks.distance(saleprice\_multi\_lm)  
updated\_housing\_df$dfbeta <- dfbeta(saleprice\_multi\_lm)  
updated\_housing\_df$dffit <- dffits(saleprice\_multi\_lm)  
updated\_housing\_df$leverage <- hatvalues(saleprice\_multi\_lm)  
updated\_housing\_df$covariance.ratios <- covratio(saleprice\_multi\_lm)  
  
summary(updated\_housing\_df)

## Sale Price sale\_reason sale\_instrument zip5   
## Min. : 2500 Min. : 0.000 Min. : 0.000 Min. :98052   
## 1st Qu.: 485075 1st Qu.: 1.000 1st Qu.: 3.000 1st Qu.:98052   
## Median : 605000 Median : 1.000 Median : 3.000 Median :98052   
## Mean : 645051 Mean : 1.107 Mean : 3.147 Mean :98053   
## 3rd Qu.: 749950 3rd Qu.: 1.000 3rd Qu.: 3.000 3rd Qu.:98053   
## Max. :4311000 Max. :18.000 Max. :26.000 Max. :98074   
## lon lat building\_grade square\_feet\_total\_living  
## Min. :-122.2 Min. :47.46 Min. : 2.000 Min. : 240   
## 1st Qu.:-122.1 1st Qu.:47.67 1st Qu.: 8.000 1st Qu.: 1870   
## Median :-122.1 Median :47.69 Median : 8.000 Median : 2450   
## Mean :-122.1 Mean :47.68 Mean : 8.273 Mean : 2545   
## 3rd Qu.:-122.0 3rd Qu.:47.71 3rd Qu.: 9.000 3rd Qu.: 3110   
## Max. :-121.9 Max. :47.73 Max. :13.000 Max. :13540   
## bedrooms bath\_full\_count bath\_half\_count bath\_3qtr\_count   
## Min. : 0.000 Min. : 0.000 Min. :0.0000 Min. :0.0000   
## 1st Qu.: 3.000 1st Qu.: 1.000 1st Qu.:0.0000 1st Qu.:0.0000   
## Median : 4.000 Median : 2.000 Median :1.0000 Median :0.0000   
## Mean : 3.482 Mean : 1.803 Mean :0.6175 Mean :0.5006   
## 3rd Qu.: 4.000 3rd Qu.: 2.000 3rd Qu.:1.0000 3rd Qu.:1.0000   
## Max. :11.000 Max. :23.000 Max. :6.0000 Max. :8.0000   
## year\_built year\_renovated sq\_ft\_lot present\_use   
## Min. :1900 Min. : 0.00 Min. : 785 Min. : 0.000   
## 1st Qu.:1980 1st Qu.: 0.00 1st Qu.: 5400 1st Qu.: 2.000   
## Median :1999 Median : 0.00 Median : 7850 Median : 2.000   
## Mean :1993 Mean : 21.93 Mean : 19921 Mean : 6.546   
## 3rd Qu.:2007 3rd Qu.: 0.00 3rd Qu.: 12037 3rd Qu.: 2.000   
## Max. :2016 Max. :2016.00 Max. :1631322 Max. :300.000   
## residuals standardized.residuals studentized.residuals  
## Min. :-2231685 Min. :-13.650696 Min. :-13.772102   
## 1st Qu.: -79883 1st Qu.: -0.463190 1st Qu.: -0.463172   
## Median : -13674 Median : -0.079285 Median : -0.079281   
## Mean : 0 Mean : -0.000057 Mean : 0.000077   
## 3rd Qu.: 61708 3rd Qu.: 0.357792 3rd Qu.: 0.357777   
## Max. : 3676548 Max. : 21.393726 Max. : 21.871950   
## cooks.distance   
## Min. :0.0000000   
## 1st Qu.:0.0000025   
## Median :0.0000116   
## Mean :0.0007152   
## 3rd Qu.:0.0000385   
## Max. :2.6413062   
## dfbeta.(Intercept) dfbeta.updated\_housing\_df$sq\_ft\_lot dfbeta.updated\_housing\_df$building\_grade dfbeta.updated\_housing\_df$square\_feet\_total\_living dfbeta.updated\_housing\_df$bedrooms dfbeta.updated\_housing\_df$bath\_full\_count dfbeta.updated\_housing\_df$bath\_half\_count dfbeta.updated\_housing\_df$year\_built  
## Min. :-216952.03 Min. :-0.15927429 Min. :-911.8551 Min. :-3.424907 Min. :-944.5483 Min. :-9098.876 Min. :-739.1081 Min. :-87.62030   
## 1st Qu.: -647.02 1st Qu.:-0.00002126 1st Qu.: -6.1500 1st Qu.:-0.006177 1st Qu.: -4.7024 1st Qu.: -6.171 1st Qu.: -11.0292 1st Qu.: -0.25369   
## Median : -26.76 Median : 0.00000091 Median : -0.3114 Median :-0.000019 Median : 0.1886 Median : 0.058 Median : -0.3151 Median : 0.01425   
## Mean : -1.98 Mean :-0.00000087 Mean : 0.0219 Mean :-0.000001 Mean : 0.0092 Mean : -0.112 Mean : -0.0020 Mean : 0.00100   
## 3rd Qu.: 498.84 3rd Qu.: 0.00003107 3rd Qu.: 3.9711 3rd Qu.: 0.006337 3rd Qu.: 6.3648 3rd Qu.: 6.233 3rd Qu.: 11.2365 3rd Qu.: 0.33446   
## Max. : 171340.71 Max. : 0.06799227 Max. :1413.7942 Max. : 3.344290 Max. :1212.4387 Max. : 1570.894 Max. : 961.5785 Max. :112.37185   
## dffit leverage covariance.ratios  
## Min. :-4.637669 Min. :0.0001863 Min. :0.7074   
## 1st Qu.:-0.010578 1st Qu.:0.0004285 1st Qu.:1.0009   
## Median :-0.001792 Median :0.0005571 Median :1.0011   
## Mean : 0.000336 Mean :0.0007570 Mean :1.0008   
## 3rd Qu.: 0.008260 3rd Qu.:0.0007305 3rd Qu.:1.0013   
## Max. : 1.925179 Max. :0.1498742 Max. :1.1388

## 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.

updated\_housing\_df$large.residual <- updated\_housing\_df$standardized.residuals > 2 | updated\_housing\_df$standardized.residuals < -2

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

sum(updated\_housing\_df$large.residual)

## [1] 357

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

updated\_housing\_df[updated\_housing\_df$large.residual, c("Sale Price", "building\_grade", "square\_feet\_total\_living", "bedrooms", "bath\_full\_count", "bath\_half\_count", "year\_built", "sq\_ft\_lot", "standardized.residuals")]

## # A tibble: 357 x 9  
## `Sale Price` building\_grade square\_feet\_tot~ bedrooms bath\_full\_count  
## <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 165000 9 1850 3 2  
## 2 265000 10 4920 4 4  
## 3 1392000 9 3740 4 3  
## 4 1080135 9 2700 3 2  
## 5 732500 9 5710 5 3  
## 6 1390000 6 660 0 1  
## 7 1390000 10 3280 3 2  
## 8 370000 9 4000 4 3  
## 9 390000 11 5800 5 4  
## 10 1588359 9 3360 2 2  
## # ... with 347 more rows, and 4 more variables: bath\_half\_count <dbl>,  
## # year\_built <dbl>, sq\_ft\_lot <dbl>, standardized.residuals <dbl>

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

updated\_housing\_df[updated\_housing\_df$large.residual, c("cooks.distance", "leverage", "covariance.ratios")]

## # A tibble: 357 x 3  
## cooks.distance leverage covariance.ratios  
## <dbl> <dbl> <dbl>  
## 1 0.00450 0.00418 0.998  
## 2 0.00543 0.00171 0.983  
## 3 0.00126 0.00109 0.995  
## 4 0.000237 0.000428 0.998  
## 5 0.00218 0.00309 1.00   
## 6 0.0178 0.00400 0.978  
## 7 0.00198 0.00225 0.998  
## 8 0.000697 0.000601 0.994  
## 9 0.00596 0.00175 0.982  
## 10 0.00225 0.000966 0.988  
## # ... with 347 more rows

There is 1 problematic record out of 357 records. cooks distance for that record is greater than 1. The data shows that the sale price was only $14,000 but other factors indicated that the price is too low. The square\_feet\_total\_living is 8750, there are 5 bedrooms, 2 full bathrooms, 2 half bathrooms, and the sq\_ft\_lot is 1631322. The standardized residual is too high (-13.650696).

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

library('car')

## Warning: package 'car' was built under R version 4.0.3

## Loading required package: carData

## Warning: package 'carData' was built under R version 4.0.3

##   
## Attaching package: 'car'

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

durbinWatsonTest(saleprice\_multi\_lm)

## lag Autocorrelation D-W Statistic p-value  
## 1 0.2670775 1.465819 0  
## Alternative hypothesis: rho != 0

The assumption of independence is tested using Durbin-Watson Test. The D-W statistic should be between 1 and 3 and should be closer to 2. In this case, it is 1.465819 and the p-value is 0 which is less than 0.05. This means that the model meets assumption of independence.

## m. Perform the necessary calculations to assess the assumption of no multicollinearity and state if the condition is met or not.

print("VIF")

## [1] "VIF"

vif(saleprice\_multi\_lm)

## updated\_housing\_df$sq\_ft\_lot   
## 1.141163   
## updated\_housing\_df$building\_grade   
## 2.310647   
## updated\_housing\_df$square\_feet\_total\_living   
## 3.532411   
## updated\_housing\_df$bedrooms   
## 1.685048   
## updated\_housing\_df$bath\_full\_count   
## 1.563972   
## updated\_housing\_df$bath\_half\_count   
## 1.103657   
## updated\_housing\_df$year\_built   
## 1.463967

print("Tolerance")

## [1] "Tolerance"

1/vif(saleprice\_multi\_lm)

## updated\_housing\_df$sq\_ft\_lot   
## 0.8762987   
## updated\_housing\_df$building\_grade   
## 0.4327791   
## updated\_housing\_df$square\_feet\_total\_living   
## 0.2830927   
## updated\_housing\_df$bedrooms   
## 0.5934551   
## updated\_housing\_df$bath\_full\_count   
## 0.6393977   
## updated\_housing\_df$bath\_half\_count   
## 0.9060789   
## updated\_housing\_df$year\_built   
## 0.6830753

print("Mean")

## [1] "Mean"

mean(vif(saleprice\_multi\_lm))

## [1] 1.828695

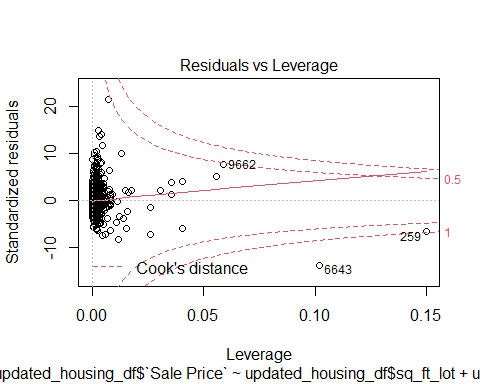
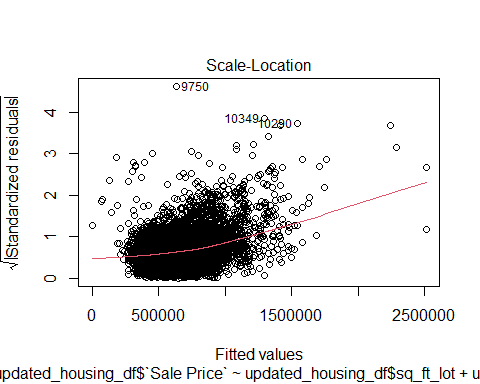
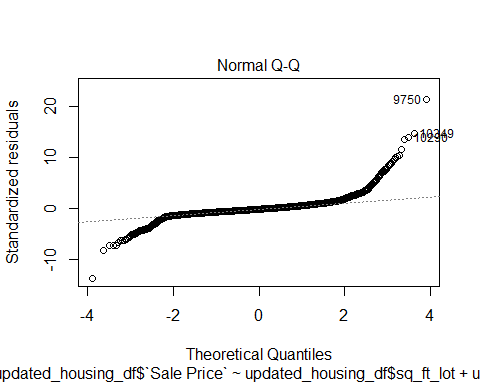
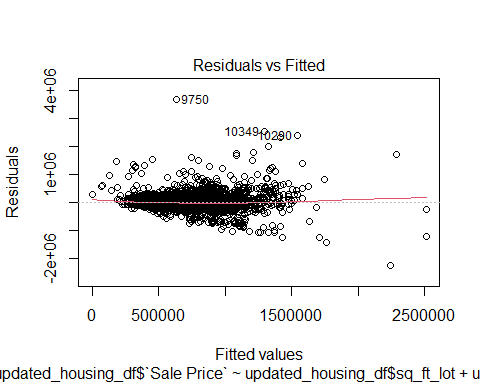
If the largest VIF is greater than 10, then there is a cause for concern. In this case, the largest VIF 3.532411, so there are no concerns.

If the tolerance is below 0.2, then its a potential problem. In this case, the smallest tolerance value is 0.2830927, so there are no concerns.

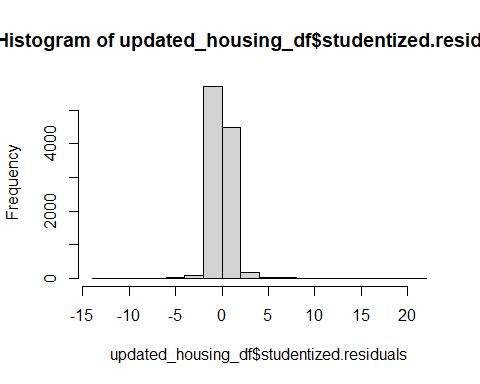
If the mean is substantially greater than 1, then the regression may be biased. In this case, it is not too far from 1, so there may not be any concerns.

## 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\_multi\_lm)



hist(updated\_housing\_df$studentized.residuals)



The Residuals vs Fitted graph above looks like a random array of dots evenly dispersed around zero. It does not funnel out, so there is no heteroscedasticity in the data. There is no curve in the graph, so it is not violating any assumptions of linearity.

The Q-Q plot should show deviations from normality. In the plot above, it deviates from both the ends of the line, which indicates deviation of normality at the extreme values.

The histogram indicates that the distribution is skewed to right to some degree.

## 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?

The problematic record I found had cooks distance greater than 1. The sale price was only $14,000 but other factors indicated that the price is too low. The square\_feet\_total\_living is 8750, there are 5 bedrooms, 2 full bathrooms, 2 half bathrooms, and the sq\_ft\_lot is 1631322. The standardized residual is too high.

The Q-Q plot also showed significant curves at the ends which could indicate that there are extreme values in the data set that make the model deviate from normality.

The mean VIF is greater that 1, which could indicate that the model is biased.

Based on these things, I conclude that the regression model is biased.

The model needs to be recreated after removing the outliers and problematic records.