Exercise 12: Housing Data

Kolekar, Ajit

October 18th, 2020

# Assignment description

**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')  
housing\_df <- read\_excel("C:/Users/Shilp/Documents/GitHub/dsc520/data/week-7-housing.xlsx")

## Question 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

new\_housing\_df <- housing\_df[(is.na(housing\_df$sale\_warning)) & (housing\_df$bedrooms != 0), ]  
new\_housing\_df$`Sale Date` <- NULL  
new\_housing\_df$sale\_warning <- NULL  
new\_housing\_df$sitetype <- NULL  
new\_housing\_df$addr\_full <- NULL  
new\_housing\_df$ctyname <- NULL  
new\_housing\_df$postalctyn <- NULL  
new\_housing\_df$current\_zoning <- NULL  
new\_housing\_df$prop\_type <- NULL  
summary(new\_housing\_df)

## Sale Price sale\_reason sale\_instrument zip5   
## Min. : 2500 Min. : 0.000 Min. : 0.000 Min. :98052   
## 1st Qu.: 485000 1st Qu.: 1.000 1st Qu.: 3.000 1st Qu.:98052   
## Median : 605000 Median : 1.000 Median : 3.000 Median :98052   
## Mean : 644676 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. : 410   
## 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. : 1.000 Min. : 0.000 Min. :0.0000 Min. :0.000   
## 1st Qu.: 3.000 1st Qu.: 1.000 1st Qu.:0.0000 1st Qu.:0.000   
## Median : 4.000 Median : 2.000 Median :1.0000 Median :0.000   
## Mean : 3.486 Mean : 1.805 Mean :0.6181 Mean :0.501   
## 3rd Qu.: 4.000 3rd Qu.: 2.000 3rd Qu.:1.0000 3rd Qu.:1.000   
## Max. :11.000 Max. :23.000 Max. :6.0000 Max. :8.000   
## 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 : 7846 Median : 2.000   
## Mean :1993 Mean : 21.76 Mean : 19880 Mean : 6.551   
## 3rd Qu.:2007 3rd Qu.: 0.00 3rd Qu.: 12030 3rd Qu.: 2.000   
## Max. :2016 Max. :2016.00 Max. :1631322 Max. :300.000

There are certain data points in the data set that can skew the data and would make the linear models provide inaccurate outputs. Identifying these data points, such as outliers, is critical for successful data analysis.

Following are the variables I chose to clean/remove and the reasoning why.

* Removed data points where bedrooms = 0, since they appeared to be lands and not houses.
* Removed data points where sale\_warning is not blank, since sale warning could have impacted the sale price and thus could skew the data.
* Removed following variables defined as char so that correlation between variables can be determined: Sale Date, sale\_warning, sitetype, addr\_full, ctyname, postalctyn, current\_zoning, and prop\_type.
* It is difficult to determine the impacts of other numeric variables unless the meaning of each code is understood, which is not available.

The data set originally included 12865 rows and after cleaning the data set reduced to 10556 rows.

## Question 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.**

cor(new\_housing\_df)

## Sale Price sale\_reason sale\_instrument  
## Sale Price 1.000000000 -0.032409183 -0.0667470552  
## sale\_reason -0.032409183 1.000000000 0.2191629030  
## sale\_instrument -0.066747055 0.219162903 1.0000000000  
## zip5 0.072836524 -0.007206362 -0.0074858546  
## lon 0.042020337 -0.003037295 -0.0018861286  
## lat -0.005363584 0.018576458 -0.0368878795  
## building\_grade 0.653412765 -0.029167368 -0.0109631257  
## square\_feet\_total\_living 0.710440521 -0.026844450 0.0177542270  
## bedrooms 0.339592447 -0.031438476 0.0125537836  
## bath\_full\_count 0.388416107 -0.027205404 0.0090721407  
## bath\_half\_count 0.226803882 -0.002869406 -0.0071825924  
## bath\_3qtr\_count 0.098611305 -0.005973520 0.0103592113  
## year\_built 0.261871790 -0.030665451 -0.0288476357  
## year\_renovated 0.054437260 0.003748229 0.0187682236  
## sq\_ft\_lot 0.240103556 -0.001874820 0.1113761922  
## present\_use -0.008306268 -0.007952966 -0.0002830689  
## zip5 lon lat building\_grade  
## Sale Price 0.0728365236 0.042020337 -0.005363584 0.65341277  
## sale\_reason -0.0072063616 -0.003037295 0.018576458 -0.02916737  
## sale\_instrument -0.0074858546 -0.001886129 -0.036887879 -0.01096313  
## zip5 1.0000000000 0.366332808 -0.122872048 0.09040821  
## lon 0.3663328075 1.000000000 -0.021633426 0.05014112  
## lat -0.1228720479 -0.021633426 1.000000000 0.03421125  
## building\_grade 0.0904082125 0.050141120 0.034211251 1.00000000  
## square\_feet\_total\_living 0.0875187633 0.081895066 -0.037475291 0.73361961  
## bedrooms -0.0425745763 -0.242605438 -0.168157236 0.33178478  
## bath\_full\_count 0.0989004376 0.139628246 0.086316826 0.44784733  
## bath\_half\_count 0.0140743467 0.014059005 -0.036617004 0.25688523  
## bath\_3qtr\_count -0.0708125768 -0.124841456 -0.094538140 0.04224956  
## year\_built 0.1065940889 0.348337251 0.404542514 0.36608949  
## year\_renovated -0.0002392475 -0.019364510 -0.092333798 -0.01865125  
## sq\_ft\_lot 0.0915354156 0.222523679 -0.164201442 0.18081649  
## present\_use 0.0344885804 0.134640501 0.029897345 0.02559069  
## square\_feet\_total\_living bedrooms bath\_full\_count  
## Sale Price 0.71044052 0.339592447 0.388416107  
## sale\_reason -0.02684445 -0.031438476 -0.027205404  
## sale\_instrument 0.01775423 0.012553784 0.009072141  
## zip5 0.08751876 -0.042574576 0.098900438  
## lon 0.08189507 -0.242605438 0.139628246  
## lat -0.03747529 -0.168157236 0.086316826  
## building\_grade 0.73361961 0.331784781 0.447847330  
## square\_feet\_total\_living 1.00000000 0.586633933 0.512735342  
## bedrooms 0.58663393 1.000000000 0.272715456  
## bath\_full\_count 0.51273534 0.272715456 1.000000000  
## bath\_half\_count 0.28587810 0.126560722 0.195465514  
## bath\_3qtr\_count 0.17685036 0.241459428 -0.399012171  
## year\_built 0.33066314 0.004035635 0.459345634  
## year\_renovated 0.03634355 0.019257201 0.023665812  
## sq\_ft\_lot 0.25726410 0.067745272 0.062338636  
## present\_use 0.01639448 -0.045759657 0.033114962  
## bath\_half\_count bath\_3qtr\_count year\_built  
## Sale Price 0.226803882 0.098611305 0.261871790  
## sale\_reason -0.002869406 -0.005973520 -0.030665451  
## sale\_instrument -0.007182592 0.010359211 -0.028847636  
## zip5 0.014074347 -0.070812577 0.106594089  
## lon 0.014059005 -0.124841456 0.348337251  
## lat -0.036617004 -0.094538140 0.404542514  
## building\_grade 0.256885228 0.042249560 0.366089490  
## square\_feet\_total\_living 0.285878102 0.176850363 0.330663144  
## bedrooms 0.126560722 0.241459428 0.004035635  
## bath\_full\_count 0.195465514 -0.399012171 0.459345634  
## bath\_half\_count 1.000000000 -0.358909270 0.181697960  
## bath\_3qtr\_count -0.358909270 1.000000000 -0.155258064  
## year\_built 0.181697960 -0.155258064 1.000000000  
## year\_renovated -0.026629288 0.020249732 -0.214223291  
## sq\_ft\_lot 0.049525393 0.050210958 -0.089460628  
## present\_use 0.007911082 -0.008755589 0.129974958  
## year\_renovated sq\_ft\_lot present\_use  
## Sale Price 0.0544372599 0.24010356 -0.0083062676  
## sale\_reason 0.0037482293 -0.00187482 -0.0079529663  
## sale\_instrument 0.0187682236 0.11137619 -0.0002830689  
## zip5 -0.0002392475 0.09153542 0.0344885804  
## lon -0.0193645101 0.22252368 0.1346405011  
## lat -0.0923337978 -0.16420144 0.0298973450  
## building\_grade -0.0186512507 0.18081649 0.0255906854  
## square\_feet\_total\_living 0.0363435500 0.25726410 0.0163944809  
## bedrooms 0.0192572005 0.06774527 -0.0457596566  
## bath\_full\_count 0.0236658116 0.06233864 0.0331149620  
## bath\_half\_count -0.0266292876 0.04952539 0.0079110820  
## bath\_3qtr\_count 0.0202497322 0.05021096 -0.0087555886  
## year\_built -0.2142232912 -0.08946063 0.1299749575  
## year\_renovated 1.0000000000 0.04916700 -0.0159403201  
## sq\_ft\_lot 0.0491670004 1.00000000 0.0552090693  
## present\_use -0.0159403201 0.05520907 1.0000000000

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

After calculating the correlation between sale price and other variables, I noticed that several variables had significant relationship with sale price. I used all these variables as predictors. Their correlation coefficient is more than 0.20 in either positive or negative direction. The predictors I chose are sq\_ft\_lot, building\_grade, square\_feet\_total\_living, bedrooms, bath\_full\_count, bath\_half\_count, and year\_built.

## Question 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(housing\_simple\_lm)

##   
## Call:  
## lm(formula = new\_housing\_df$`Sale Price` ~ new\_housing\_df$sq\_ft\_lot,   
## data = new\_housing\_df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2608418 -151495 -35459 106127 3293524   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.203e+05 2.596e+03 238.94 <2e-16 \*\*\*  
## new\_housing\_df$sq\_ft\_lot 1.227e+00 4.830e-02 25.41 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 247800 on 10554 degrees of freedom  
## Multiple R-squared: 0.05765, Adjusted R-squared: 0.05756   
## F-statistic: 645.7 on 1 and 10554 DF, p-value: < 2.2e-16

summary(housing\_multi\_lm)

##   
## Call:  
## lm(formula = new\_housing\_df$`Sale Price` ~ new\_housing\_df$sq\_ft\_lot +   
## new\_housing\_df$building\_grade + new\_housing\_df$square\_feet\_total\_living +   
## new\_housing\_df$bedrooms + new\_housing\_df$bath\_full\_count +   
## new\_housing\_df$bath\_half\_count + new\_housing\_df$year\_built,   
## data = new\_housing\_df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2215012 -79656 -13269 61726 3682162   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 3.456e+05 2.402e+05 1.439 0.1503  
## new\_housing\_df$sq\_ft\_lot 2.816e-01 3.568e-02 7.893 3.25e-15  
## new\_housing\_df$building\_grade 6.676e+04 2.427e+03 27.503 < 2e-16  
## new\_housing\_df$square\_feet\_total\_living 1.461e+02 3.388e+00 43.134 < 2e-16  
## new\_housing\_df$bedrooms -2.182e+04 2.529e+03 -8.630 < 2e-16  
## new\_housing\_df$bath\_full\_count 6.081e+03 3.283e+03 1.852 0.0640  
## new\_housing\_df$bath\_half\_count 5.404e+03 3.408e+03 1.585 0.1129  
## new\_housing\_df$year\_built -2.855e+02 1.219e+02 -2.342 0.0192  
##   
## (Intercept)   
## new\_housing\_df$sq\_ft\_lot \*\*\*  
## new\_housing\_df$building\_grade \*\*\*  
## new\_housing\_df$square\_feet\_total\_living \*\*\*  
## new\_housing\_df$bedrooms \*\*\*  
## new\_housing\_df$bath\_full\_count .   
## new\_housing\_df$bath\_half\_count   
## new\_housing\_df$year\_built \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 171300 on 10548 degrees of freedom  
## Multiple R-squared: 0.5498, Adjusted R-squared: 0.5495   
## F-statistic: 1840 on 7 and 10548 DF, p-value: < 2.2e-16

R2 value describes whether the regression model is successful in predicting the outcome. Adjusted R2 is used to compare with R2 to determine whether the sample was a good representation of the population. If the difference between R2 and adjusted R2 values is small, then that would indicate that the sample is a good representation of population.

For the simple regression model, the value of R2 is 0.05765. This indicates that the sq\_ft\_lot accounted for only 5.77% of the variation in sale price. The value of adjusted R2 is 0.5756 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 0.5498. This indicates that the the model with multiple predictors accounted for 54.98% of the variation in sale price. The value of adjusted R2 is 0.5495 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 5.77% to 54.98% which indicates that the sale price can be better predicted with the multiple predictors than only with sq\_ft\_lot variable,

## Question 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(housing\_multi\_lm)

## new\_housing\_df$sq\_ft\_lot new\_housing\_df$building\_grade   
## 0.05509897 0.27301464   
## new\_housing\_df$square\_feet\_total\_living new\_housing\_df$bedrooms   
## 0.53443500 -0.07367565   
## new\_housing\_df$bath\_full\_count new\_housing\_df$bath\_half\_count   
## 0.01515463 0.01088315   
## new\_housing\_df$year\_built   
## -0.01850610

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\_grader 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.

## Question e

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

confint(housing\_multi\_lm)

## 2.5 % 97.5 %  
## (Intercept) -1.252666e+05 8.164022e+05  
## new\_housing\_df$sq\_ft\_lot 2.116977e-01 3.515892e-01  
## new\_housing\_df$building\_grade 6.200267e+04 7.151891e+04  
## new\_housing\_df$square\_feet\_total\_living 1.395044e+02 1.527875e+02  
## new\_housing\_df$bedrooms -2.677880e+04 -1.686583e+04  
## new\_housing\_df$bath\_full\_count -3.542317e+02 1.251592e+04  
## new\_housing\_df$bath\_half\_count -1.277279e+03 1.208439e+04  
## new\_housing\_df$year\_built -5.243798e+02 -4.653322e+01

A good model would have small confidence interval which 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 since the confidence interval for them crosses zero.

## Question 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(housing\_simple\_lm, housing\_multi\_lm)

## Analysis of Variance Table  
##   
## Model 1: new\_housing\_df$`Sale Price` ~ new\_housing\_df$sq\_ft\_lot  
## Model 2: new\_housing\_df$`Sale Price` ~ new\_housing\_df$sq\_ft\_lot + new\_housing\_df$building\_grade +   
## new\_housing\_df$square\_feet\_total\_living + new\_housing\_df$bedrooms +   
## new\_housing\_df$bath\_full\_count + new\_housing\_df$bath\_half\_count +   
## new\_housing\_df$year\_built  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 10554 6.4806e+14   
## 2 10548 3.0961e+14 6 3.3845e+14 1921.8 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

F(6, 10548) = 1921.8 with p < 0.001. This indicates that the multiple regression model significantly improved the fit of the model.

## Question 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  
new\_housing\_df$residuals <- resid(housing\_multi\_lm)  
new\_housing\_df$standardized.residuals <- rstandard(housing\_multi\_lm)  
new\_housing\_df$studentized.residuals <- rstudent(housing\_multi\_lm)  
  
# Influential cases  
new\_housing\_df$cooks.distance <- cooks.distance(housing\_multi\_lm)  
new\_housing\_df$dfbeta <- dfbeta(housing\_multi\_lm)  
new\_housing\_df$dffit <- dffits(housing\_multi\_lm)  
new\_housing\_df$leverage <- hatvalues(housing\_multi\_lm)  
new\_housing\_df$covariance.ratios <- covratio(housing\_multi\_lm)  
  
summary(new\_housing\_df)

## Sale Price sale\_reason sale\_instrument zip5   
## Min. : 2500 Min. : 0.000 Min. : 0.000 Min. :98052   
## 1st Qu.: 485000 1st Qu.: 1.000 1st Qu.: 3.000 1st Qu.:98052   
## Median : 605000 Median : 1.000 Median : 3.000 Median :98052   
## Mean : 644676 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. : 410   
## 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. : 1.000 Min. : 0.000 Min. :0.0000 Min. :0.000   
## 1st Qu.: 3.000 1st Qu.: 1.000 1st Qu.:0.0000 1st Qu.:0.000   
## Median : 4.000 Median : 2.000 Median :1.0000 Median :0.000   
## Mean : 3.486 Mean : 1.805 Mean :0.6181 Mean :0.501   
## 3rd Qu.: 4.000 3rd Qu.: 2.000 3rd Qu.:1.0000 3rd Qu.:1.000   
## Max. :11.000 Max. :23.000 Max. :6.0000 Max. :8.000   
## 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 : 7846 Median : 2.000   
## Mean :1993 Mean : 21.76 Mean : 19880 Mean : 6.551   
## 3rd Qu.:2007 3rd Qu.: 0.00 3rd Qu.: 12030 3rd Qu.: 2.000   
## Max. :2016 Max. :2016.00 Max. :1631322 Max. :300.000   
## residuals standardized.residuals studentized.residuals  
## Min. :-2215012 Min. :-13.643701 Min. :-13.765058   
## 1st Qu.: -79656 1st Qu.: -0.465148 1st Qu.: -0.465130   
## Median : -13269 Median : -0.077476 Median : -0.077472   
## Mean : 0 Mean : -0.000059 Mean : 0.000075   
## 3rd Qu.: 61726 3rd Qu.: 0.360384 3rd Qu.: 0.360369   
## Max. : 3682162 Max. : 21.574159 Max. : 22.065462   
## cooks.distance   
## Min. :0.0000000   
## 1st Qu.:0.0000025   
## Median :0.0000119   
## Mean :0.0007224   
## 3rd Qu.:0.0000392   
## Max. :2.6451351   
## dfbeta.(Intercept) dfbeta.new\_housing\_df$sq\_ft\_lot dfbeta.new\_housing\_df$building\_grade dfbeta.new\_housing\_df$square\_feet\_total\_living dfbeta.new\_housing\_df$bedrooms dfbeta.new\_housing\_df$bath\_full\_count dfbeta.new\_housing\_df$bath\_half\_count dfbeta.new\_housing\_df$year\_built  
## Min. :-218458.93 Min. :-0.15845345 Min. :-918.8991 Min. :-3.480046 Min. :-993.9973 Min. :-9252.178 Min. :-747.2516 Min. :-86.49017   
## 1st Qu.: -641.98 1st Qu.:-0.00002130 1st Qu.: -6.2727 1st Qu.:-0.006269 1st Qu.: -4.8709 1st Qu.: -6.201 1st Qu.: -11.0819 1st Qu.: -0.25762   
## Median : -24.52 Median : 0.00000092 Median : -0.3309 Median :-0.000021 Median : 0.1270 Median : 0.054 Median : -0.3669 Median : 0.01348   
## Mean : -2.08 Mean :-0.00000086 Mean : 0.0223 Mean :-0.000001 Mean : 0.0089 Mean : -0.115 Mean : -0.0024 Mean : 0.00105   
## 3rd Qu.: 502.78 3rd Qu.: 0.00003120 3rd Qu.: 3.8840 3rd Qu.: 0.006519 3rd Qu.: 6.4546 3rd Qu.: 6.261 3rd Qu.: 11.2366 3rd Qu.: 0.33260   
## Max. : 168890.22 Max. : 0.06890463 Max. :1415.6825 Max. : 3.397754 Max. :1261.6813 Max. : 1600.828 Max. : 958.6665 Max. :113.36708   
## dffit leverage covariance.ratios  
## Min. :-4.641034 Min. :0.0001866 Min. :0.7028   
## 1st Qu.:-0.010615 1st Qu.:0.0004308 1st Qu.:1.0009   
## Median :-0.001784 Median :0.0005602 Median :1.0011   
## Mean : 0.000301 Mean :0.0007579 Mean :1.0008   
## 3rd Qu.: 0.008428 3rd Qu.:0.0007368 3rd Qu.:1.0013   
## Max. : 1.962093 Max. :0.1512764 Max. :1.1396

## Question 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.**

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

## Question i

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

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

## [1] 356

## Question j

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

new\_housing\_df[new\_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: 356 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 10 3280 3 2  
## 7 370000 9 4000 4 3  
## 8 390000 11 5800 5 4  
## 9 1588359 9 3360 2 2  
## 10 1450000 8 3480 3 2  
## # ... with 346 more rows, and 4 more variables: bath\_half\_count <dbl>,  
## # year\_built <dbl>, sq\_ft\_lot <dbl>, standardized.residuals <dbl>

## Question k

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

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

## # A tibble: 356 x 3  
## cooks.distance leverage covariance.ratios  
## <dbl> <dbl> <dbl>  
## 1 0.00455 0.00419 0.998  
## 2 0.00551 0.00171 0.983  
## 3 0.00128 0.00110 0.995  
## 4 0.000242 0.000432 0.998  
## 5 0.00220 0.00310 1.00   
## 6 0.00202 0.00225 0.998  
## 7 0.000709 0.000601 0.994  
## 8 0.00609 0.00175 0.982  
## 9 0.00237 0.000995 0.987  
## 10 0.00227 0.00119 0.990  
## # ... with 346 more rows

There is 1 problematic record out of 356 records since cooks distance is greater than 1 for the problematic record and is less than 1 for the remaining records. When I looked at the data for that record, 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 (-13.643701).

## Question 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(housing\_multi\_lm)

## lag Autocorrelation D-W Statistic p-value  
## 1 0.2639746 1.472023 0  
## Alternative hypothesis: rho != 0

The assumption of independence can be tested using Durbin-Watson Test. The Durbin-Watson statistic should be between 1 and 3 and should be closer to 2. In this case, it is 1.472023 which means that the assumption of independence has met. The p-value is 0 which is a good for the model since it is less than 0.05.

## Question 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(housing\_multi\_lm)

## new\_housing\_df$sq\_ft\_lot new\_housing\_df$building\_grade   
## 1.141755 2.308675   
## new\_housing\_df$square\_feet\_total\_living new\_housing\_df$bedrooms   
## 3.596795 1.707484   
## new\_housing\_df$bath\_full\_count new\_housing\_df$bath\_half\_count   
## 1.568313 1.104013   
## new\_housing\_df$year\_built   
## 1.462950

print("1/vif")

## [1] "1/vif"

1/vif(housing\_multi\_lm)

## new\_housing\_df$sq\_ft\_lot new\_housing\_df$building\_grade   
## 0.8758450 0.4331489   
## new\_housing\_df$square\_feet\_total\_living new\_housing\_df$bedrooms   
## 0.2780253 0.5856572   
## new\_housing\_df$bath\_full\_count new\_housing\_df$bath\_half\_count   
## 0.6376276 0.9057862   
## new\_housing\_df$year\_built   
## 0.6835504

print("mean")

## [1] "mean"

mean(vif(housing\_multi\_lm))

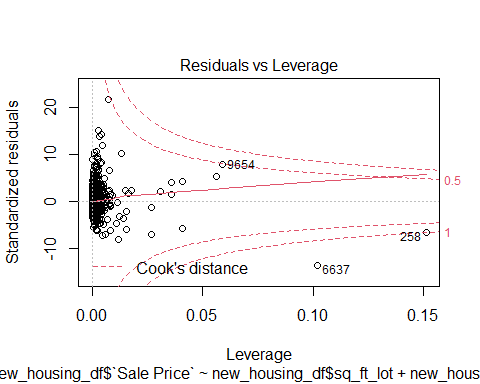
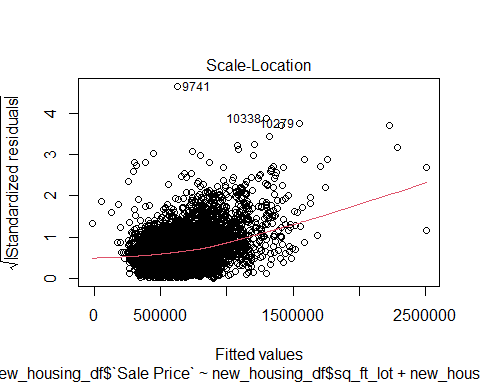
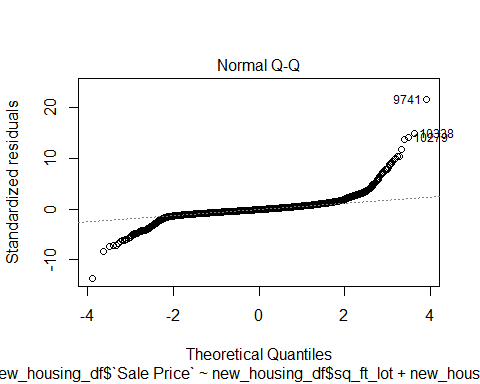
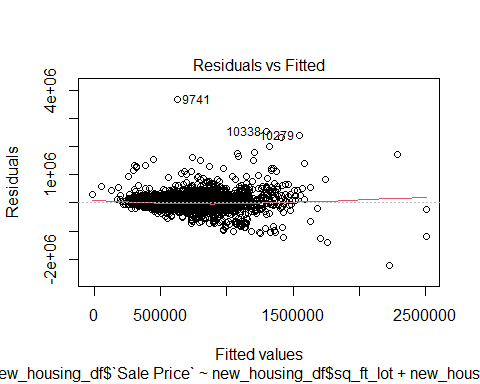
## [1] 1.841426

If the largest VIF is greater than 10, then there is a cause for concern. In this case, the largest VIF 3.59, so there are no concerns. If the tolerance (1/vif) is below 0.2, then its a potential problem. In this case, the smallest tolerance value is 0.27, so there are no concerns. If the average (mean) is substantially greater than 1, then the regression may be biased. In this case, it is not too far from 1, so there are no concerns.

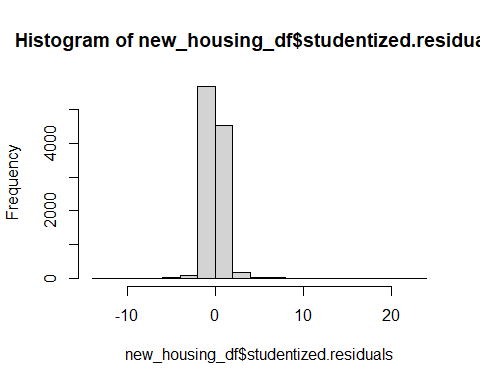
## Question n

**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(housing\_multi\_lm)



hist(new\_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 roughly normal or skewed a little to right.

## Question 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?**

While working on question k, I found a problematic record that skewed the data. The record had cooks distance greater than 1. When I looked at the data for that record, 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 (-13.643701).

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.

Based on these observations, I feel that the regression model is biased.

The extreme values (outliers) need to be removed from the data set. It would also be beneficial to have the meaning of the codes of each variable available so that analysis can be done on them. The model should be re-created after the problematic record and outliers removed from the data set.