---  
title: "Assignment7\_Exercise12\_HousingData"  
author: "SumbarajuAditya"  
date: "1/29/2021"  
output: word\_document  
---

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:

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

library("readxl")  
housing\_data <- read\_excel("C:/BU/DSC520/assignment\_repo/dsc520/data/week-7-housing.xlsx")  
head(housing\_data)

## # A tibble: 6 x 24  
## `Sale Date` `Sale Price` sale\_reason sale\_instrument sale\_warning  
## <dttm> <dbl> <dbl> <dbl> <chr>   
## 1 2006-01-03 00:00:00 698000 1 3 <NA>   
## 2 2006-01-03 00:00:00 649990 1 3 <NA>   
## 3 2006-01-03 00:00:00 572500 1 3 <NA>   
## 4 2006-01-03 00:00:00 420000 1 3 <NA>   
## 5 2006-01-03 00:00:00 369900 1 3 15   
## 6 2006-01-03 00:00:00 184667 1 15 18 51   
## # ... with 19 more variables: sitetype <chr>, addr\_full <chr>, zip5 <dbl>,  
## # ctyname <chr>, postalctyn <chr>, lon <dbl>, lat <dbl>,  
## # building\_grade <dbl>, square\_feet\_total\_living <dbl>, bedrooms <dbl>,  
## # bath\_full\_count <dbl>, bath\_half\_count <dbl>, bath\_3qtr\_count <dbl>,  
## # year\_built <dbl>, year\_renovated <dbl>, current\_zoning <chr>,  
## # sq\_ft\_lot <dbl>, prop\_type <chr>, present\_use <dbl>

nrow(housing\_data)

## [1] 12865

str(housing\_data)

## 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\_data)

## 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

upd\_housing\_data <- housing\_data[(is.na(housing\_data$sale\_warning)) & (housing\_data$bedrooms != 0) & (housing\_data$bath\_full\_count !=23), ]  
upd\_housing\_data$`Sale Date` <- NULL  
upd\_housing\_data$sale\_warning <- NULL  
upd\_housing\_data$sitetype <- NULL  
upd\_housing\_data$addr\_full <- NULL  
upd\_housing\_data$ctyname <- NULL  
upd\_housing\_data$postalctyn <- NULL  
upd\_housing\_data$current\_zoning <- NULL  
upd\_housing\_data$prop\_type <- NULL  
upd\_housing\_data$present\_use <- NULL  
upd\_housing\_data$bath\_half\_count <- NULL  
upd\_housing\_data$bath\_3qtr\_count <- NULL  
head(upd\_housing\_data)

## # A tibble: 6 x 13  
## `Sale Price` sale\_reason sale\_instrument zip5 lon lat building\_grade  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 698000 1 3 98052 -122. 47.7 9  
## 2 649990 1 3 98052 -122. 47.7 9  
## 3 572500 1 3 98052 -122. 47.7 8  
## 4 420000 1 3 98052 -122. 47.6 8  
## 5 1050000 1 3 98053 -122. 47.7 10  
## 6 875000 1 3 98053 -122. 47.7 10  
## # ... with 6 more variables: square\_feet\_total\_living <dbl>, bedrooms <dbl>,  
## # bath\_full\_count <dbl>, year\_built <dbl>, year\_renovated <dbl>,  
## # sq\_ft\_lot <dbl>

nrow(upd\_housing\_data)

## [1] 10555

Thank you, Dr.Parajulee, for providing your inputs on removing the data points. The data frame housing\_data contains data points in the data set that can participate in Data Skewness, resulting in inaccurate linear models.

Following are the variables are “removed data points” from clean dataset week-7-housing.xlsx. Please find the reasoning as stated below

• Removed data points where bedrooms = 0; this will qualify as an outlier based on the myth - “a house cannot be complete without a bedroom.” Maybe this could be a bad record.

• Removed an outlier bath\_full\_count =23. Assuming this could be an error record.

• Removed data points where sale\_warning is not blank since sale warning could have impacted the sale price and 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, prop\_type.

• As per the understanding of the data points present\_use, bath\_half\_count,bath\_3qtr\_count.These are ok to be removed as we have a calculated field bath\_full\_count based on bath\_half\_count,bath\_3qtr\_count.

• Due to the lack of a data dictionary; It is challenging to determine other numeric variables’ impacts. Row Count after Anonymizing: The data set originally included 12865 rows, and after cleaning, the data set reduced to 10555 rows.

**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.**

options(scipen = 999)  
cor(upd\_housing\_data$'Sale Price', upd\_housing\_data$square\_feet\_total\_living)^2 \* 100

## [1] 50.57093

cor.test(upd\_housing\_data$'Sale Price', upd\_housing\_data$square\_feet\_total\_living)

##   
## Pearson's product-moment correlation  
##   
## data: upd\_housing\_data$"Sale Price" and upd\_housing\_data$square\_feet\_total\_living  
## t = 103.91, df = 10553, p-value < 0.00000000000000022  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.7015727 0.7204362  
## sample estimates:  
## cor   
## 0.7111324

cor(upd\_housing\_data$'Sale Price', upd\_housing\_data$building\_grade)^2 \* 100

## [1] 42.7787

cor.test(upd\_housing\_data$'Sale Price', upd\_housing\_data$building\_grade)

##   
## Pearson's product-moment correlation  
##   
## data: upd\_housing\_data$"Sale Price" and upd\_housing\_data$building\_grade  
## t = 88.822, df = 10553, p-value < 0.00000000000000022  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.6429997 0.6648363  
## sample estimates:  
## cor   
## 0.6540542

cor(upd\_housing\_data$'Sale Price', upd\_housing\_data$year\_built)^2 \* 100

## [1] 6.870237

cor.test(upd\_housing\_data$'Sale Price', upd\_housing\_data$year\_built)

##   
## Pearson's product-moment correlation  
##   
## data: upd\_housing\_data$"Sale Price" and upd\_housing\_data$year\_built  
## t = 27.902, df = 10553, p-value < 0.00000000000000022  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.2442550 0.2797901  
## sample estimates:  
## cor   
## 0.2621114

cor(upd\_housing\_data$'Sale Price', upd\_housing\_data$bedrooms)^2 \* 100

## [1] 11.54069

cor.test(upd\_housing\_data$'Sale Price', upd\_housing\_data$bedrooms)

##   
## Pearson's product-moment correlation  
##   
## data: upd\_housing\_data$"Sale Price" and upd\_housing\_data$bedrooms  
## t = 37.105, df = 10553, p-value < 0.00000000000000022  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.3227297 0.3564833  
## sample estimates:  
## cor   
## 0.3397159

**Answer:**  
Looking at the correlation we can see all the variables shown above are positively correlated. However, building grade and square feet of living room share 50% and 42% variation and sample estimates share 71% and 65% of correlaton in determining sales price. Hence building grade and square feet of living are chosen as predictors for the model over year\_built and bedrooms.

sales\_price\_with\_sq\_ft\_lot <- lm(upd\_housing\_data$'Sale Price' ~ upd\_housing\_data$sq\_ft\_lot, data = upd\_housing\_data)  
sales\_price\_with\_others <- lm(upd\_housing\_data$'Sale Price' ~ upd\_housing\_data$sq\_ft\_lot + upd\_housing\_data$square\_feet\_total\_living + upd\_housing\_data$building\_grade, data = upd\_housing\_data)

**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(sales\_price\_with\_sq\_ft\_lot)

##   
## Call:  
## lm(formula = upd\_housing\_data$"Sale Price" ~ upd\_housing\_data$sq\_ft\_lot,   
## data = upd\_housing\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2610431 -151426 -35485 106113 3293469   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 620296.9135 2595.6838 238.97 <0.0000000000000002  
## upd\_housing\_data$sq\_ft\_lot 1.2285 0.0483 25.44 <0.0000000000000002  
##   
## (Intercept) \*\*\*  
## upd\_housing\_data$sq\_ft\_lot \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 247800 on 10553 degrees of freedom  
## Multiple R-squared: 0.05777, Adjusted R-squared: 0.05768   
## F-statistic: 647 on 1 and 10553 DF, p-value: < 0.00000000000000022

summary(sales\_price\_with\_others)

##   
## Call:  
## lm(formula = upd\_housing\_data$"Sale Price" ~ upd\_housing\_data$sq\_ft\_lot +   
## upd\_housing\_data$square\_feet\_total\_living + upd\_housing\_data$building\_grade,   
## data = upd\_housing\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2241756 -80357 -14345 60390 3668452   
##   
## Coefficients:  
## Estimate Std. Error t value  
## (Intercept) -280798.18842 15366.80930 -18.273  
## upd\_housing\_data$sq\_ft\_lot 0.32763 0.03464 9.458  
## upd\_housing\_data$square\_feet\_total\_living 132.26579 2.68207 49.315  
## upd\_housing\_data$building\_grade 70396.39532 2356.46724 29.874  
## Pr(>|t|)   
## (Intercept) <0.0000000000000002 \*\*\*  
## upd\_housing\_data$sq\_ft\_lot <0.0000000000000002 \*\*\*  
## upd\_housing\_data$square\_feet\_total\_living <0.0000000000000002 \*\*\*  
## upd\_housing\_data$building\_grade <0.0000000000000002 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 171700 on 10551 degrees of freedom  
## Multiple R-squared: 0.5475, Adjusted R-squared: 0.5474   
## F-statistic: 4256 on 3 and 10551 DF, p-value: < 0.00000000000000022

**Answer:**

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 0.05777. This indicates that the sq\_ft\_lot accounted for only 5.77% of the variation in sale price. The value of adjusted R2 is 0.5768 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 0.5475. 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 0.5474 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 54.75% 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")

## Loading required package: boot

## Loading required package: MASS

##   
## Attaching package: 'QuantPsyc'

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

lm.beta(sales\_price\_with\_others)

## upd\_housing\_data$sq\_ft\_lot   
## 0.06409636   
## upd\_housing\_data$square\_feet\_total\_living   
## 0.48356015   
## upd\_housing\_data$building\_grade   
## 0.28781857

**Answer:**

The standardized beta estimates tell us the number of standard deviations by which the outcome will change as a result of one standard deviation change in the predictor.

In this case, 1 standard deviation of change in Sq\_ft\_lot causes sales price to change by 0.064 standard deviation. 1 standard deviation change in square\_feet\_total\_living causes sales price to change by 0.483 standard deviation and 1 standard deviation change in building\_grade can cause 0.287 standard deviation change in sale price.

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

confint(sales\_price\_with\_others, level = 0.95)

## 2.5 % 97.5 %  
## (Intercept) -310920.0366424 -250676.3401935  
## upd\_housing\_data$sq\_ft\_lot 0.2597312 0.3955315  
## upd\_housing\_data$square\_feet\_total\_living 127.0084194 137.5231654  
## upd\_housing\_data$building\_grade 65777.2745241 75015.5161177

**Answer:**

The confidence interval shows that there is a positive relationship between all predictors and outcomes. Also, the 95% confidence interval range is not very big, 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 terrible model. But in our sample, predictors are having a positive relationship between all predictors and their outcome.

**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(sales\_price\_with\_sq\_ft\_lot, sales\_price\_with\_others)

## Analysis of Variance Table  
##   
## Model 1: upd\_housing\_data$"Sale Price" ~ upd\_housing\_data$sq\_ft\_lot  
## Model 2: upd\_housing\_data$"Sale Price" ~ upd\_housing\_data$sq\_ft\_lot +   
## upd\_housing\_data$square\_feet\_total\_living + upd\_housing\_data$building\_grade  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 10553 647851958564378   
## 2 10551 311105109684125 2 336746848880254 5710.3 < 0.00000000000000022 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**Answer:** The value in the column Pr(>F) is 0.00000000000000022 is an insignificant value. It refers to the p-value for F statistics. From the analysis of variance table We can say that Model 2 -sales\_price\_with\_others significantly improved the fit of the model to the data compared to Model 1- sales\_price\_with\_sq\_ft\_lot.

**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  
upd\_housing\_data$residuals <- resid(sales\_price\_with\_others)  
upd\_housing\_data$standardized\_residuals<- rstandard(sales\_price\_with\_others)  
upd\_housing\_data$studentized\_residuals<-rstudent(sales\_price\_with\_others)  
# Influential cases  
upd\_housing\_data$cooks\_distance<-cooks.distance(sales\_price\_with\_others)  
upd\_housing\_data$dfbeta<-dfbeta(sales\_price\_with\_others)  
upd\_housing\_data$dffit<-dffits(sales\_price\_with\_others)  
upd\_housing\_data$leverage<-hatvalues(sales\_price\_with\_others)  
upd\_housing\_data$covariance\_ratios<-covratio(sales\_price\_with\_others)  
  
summary(upd\_housing\_data)

## 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 : 644712 Mean : 1.107 Mean : 3.148 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 year\_built year\_renovated   
## Min. : 1.000 Min. :0.000 Min. :1900 Min. : 0.00   
## 1st Qu.: 3.000 1st Qu.:1.000 1st Qu.:1980 1st Qu.: 0.00   
## Median : 4.000 Median :2.000 Median :1999 Median : 0.00   
## Mean : 3.486 Mean :1.803 Mean :1993 Mean : 21.77   
## 3rd Qu.: 4.000 3rd Qu.:2.000 3rd Qu.:2007 3rd Qu.: 0.00   
## Max. :11.000 Max. :6.000 Max. :2016 Max. :2016.00   
## sq\_ft\_lot residuals standardized\_residuals  
## Min. : 785 Min. :-2241756 Min. :-13.755762   
## 1st Qu.: 5400 1st Qu.: -80357 1st Qu.: -0.468009   
## Median : 7846 Median : -14345 Median : -0.083546   
## Mean : 19873 Mean : 0 Mean : -0.000017   
## 3rd Qu.: 12028 3rd Qu.: 60390 3rd Qu.: 0.351726   
## Max. :1631322 Max. : 3668452 Max. : 21.442244   
## studentized\_residuals cooks\_distance   
## Min. :-13.880136 Min. :0.000000   
## 1st Qu.: -0.467992 1st Qu.:0.000002   
## Median : -0.083542 Median :0.000010   
## Mean : 0.000123 Mean :0.001052   
## 3rd Qu.: 0.351711 3rd Qu.:0.000033   
## Max. : 21.924233 Max. :5.213614   
## dfbeta.(Intercept) dfbeta.upd\_housing\_data$sq\_ft\_lot dfbeta.upd\_housing\_data$square\_feet\_total\_living dfbeta.upd\_housing\_data$building\_grade  
## Min. :-10156.973 Min. :-0.15445116 Min. :-2.2812274 Min. :-959.3112   
## 1st Qu.: -25.476 1st Qu.:-0.00002346 1st Qu.:-0.0051910 1st Qu.: -5.5802   
## Median : 1.988 Median : 0.00000051 Median :-0.0000664 Median : -0.2533   
## Mean : -0.055 Mean :-0.00000083 Mean :-0.0000174 Mean : 0.0139   
## 3rd Qu.: 35.165 3rd Qu.: 0.00002849 3rd Qu.: 0.0055220 3rd Qu.: 3.9154   
## Max. : 4644.950 Max. : 0.06371872 Max. : 2.5839038 Max. :1441.8984   
## dffit leverage covariance\_ratios  
## Min. :-4.607958 Min. :0.0001015 Min. :0.8432   
## 1st Qu.:-0.006964 1st Qu.:0.0001555 1st Qu.:1.0004   
## Median :-0.001247 Median :0.0002371 Median :1.0005   
## Mean : 0.000327 Mean :0.0003790 Mean :1.0004   
## 3rd Qu.: 0.005196 3rd Qu.:0.0003491 3rd Qu.:1.0006   
## Max. : 1.881932 Max. :0.0992713 Max. :1.0472

**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.**

upd\_housing\_data$large\_residual <- upd\_housing\_data$standardized\_residuals > 2 | upd\_housing\_data$standardized\_residuals < -2

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

sum(upd\_housing\_data$large\_residual)

## [1] 350

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

upd\_housing\_data[upd\_housing\_data$large\_residual, c('Sale Price', 'sq\_ft\_lot', 'square\_feet\_total\_living', 'building\_grade', "standardized\_residuals")]

## # A tibble: 350 x 5  
## `Sale Price` sq\_ft\_lot square\_feet\_total\_l~ building\_grade standardized\_resi~  
## <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 165000 278891 1850 9 -3.06  
## 2 265000 112650 4920 10 -4.93  
## 3 1392000 17291 3740 9 3.14  
## 4 1053649 8517 2680 9 2.00  
## 5 1080135 7694 2700 9 2.14  
## 6 732500 10200 5710 9 -2.21  
## 7 1390000 225640 3280 10 2.68  
## 8 370000 11780 4000 9 -3.00  
## 9 390000 63162 5800 11 -5.19  
## 10 1588359 8752 3360 9 4.59  
## # ... with 340 more rows

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

upd\_housing\_data[upd\_housing\_data$large\_residual, c("cooks\_distance", "leverage", "covariance\_ratios","standardized\_residuals",'Sale Price', 'sq\_ft\_lot', 'square\_feet\_total\_living', 'building\_grade')]

## # A tibble: 350 x 8  
## cooks\_distance leverage covariance\_rati~ standardized\_re~ `Sale Price`  
## <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 0.00802 0.00342 1.00 -3.06 165000  
## 2 0.00525 0.000864 0.992 -4.93 265000  
## 3 0.000683 0.000277 0.997 3.14 1392000  
## 4 0.000175 0.000175 0.999 2.00 1053649  
## 5 0.000198 0.000173 0.999 2.14 1080135  
## 6 0.00240 0.00196 1.00 -2.21 732500  
## 7 0.00352 0.00196 1.00 2.68 1390000  
## 8 0.000901 0.000399 0.997 -3.00 370000  
## 9 0.00843 0.00125 0.991 -5.19 390000  
## 10 0.000990 0.000188 0.993 4.59 1588359  
## # ... with 340 more rows, and 3 more variables: sq\_ft\_lot <dbl>,  
## # square\_feet\_total\_living <dbl>, building\_grade <dbl>

#Check if any problematic cases exist, with cooks.distance greater than 1  
cook\_dist <- upd\_housing\_data$cooks\_distance > 1.0000000000  
sum(cook\_dist)

## [1] 1

**Answer:**

Out of 350 large residuals only 1 value has cook distance greater than 1. Hence only 1 influential observation in the model. problematic record: As per my analysis, the sale price is tagged as $14,000 and i am assuming that based on the other factors such as 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 (-13.755762). This indicates that the pricing was is too low for the listed property. (Ref. - Discovering Statistics Using R (Field, Miles, and Field 2012, 424) )

k <- 3  
sample\_size <- 10555  
leverage <- k/sample\_size  
leverage

## [1] 0.0002842255

leverage\_data <- upd\_housing\_data$cooks\_distance > 0.0006 | upd\_housing\_data$cooks\_distance < 0.0009  
sum(leverage\_data)

## [1] 10555

Leverage we have is 0.0002 which means we have to look for values between 0.0006 (twice of leverage) and 0.0009 (thrice of leverage). We can see that all the observations are within the required range and hence there is no problem with the observation used in the model. (Ref. - Discovering Statistics Using R (Field, Miles, and Field 2012, 424) )

sample\_size <- 10555  
covr\_min <- 1 + 12/sample\_size  
covr\_min

## [1] 1.001137

covr\_max <- 1 - 12/sample\_size  
covr\_max

## [1] 0.9988631

covr\_data <- upd\_housing\_data$cooks\_distance > covr\_min | upd\_housing\_data$cooks\_distance < covr\_max  
sum(covr\_data)

## [1] 10555

**Answer:**

From the covariance data above we can see that all the observation in the dataset lies between cvr\_min, cvr\_mx and hence there is no outlier as per the covarince ratio calculation. (Ref. - Discovering Statistics Using R (Field, Miles, and Field 2012, 425) )

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

#Need to import car library as Durbin Watson Test was not working.  
library("car")

## Loading required package: carData

##   
## Attaching package: 'car'

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

dwt(sales\_price\_with\_others)

## lag Autocorrelation D-W Statistic p-value  
## 1 0.265446 1.46908 0  
## Alternative hypothesis: rho != 0

**Answer:**

The Durbin-Watson statistic should be between 0 and 4 and should be closer to 2: Below are the indicators - 2 is no autocorrelation. - ‘0 to <2’ is positive autocorrelation (common in time series data). - ‘>2 to 4’ is negative autocorrelation (less common in time series data).

Based on the Durbin watson test we can see the DWT-statistic value for the model is 1.46. Which means there is positive autocorrelation and the condition is met. (Ref. - Discovering Statistics Using R (Field, Miles, and Field 2012, 426) )

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

vif(sales\_price\_with\_others)

## upd\_housing\_data$sq\_ft\_lot   
## 1.070889   
## upd\_housing\_data$square\_feet\_total\_living   
## 2.242080   
## upd\_housing\_data$building\_grade   
## 2.164527

mean(vif(sales\_price\_with\_others))

## [1] 1.825832

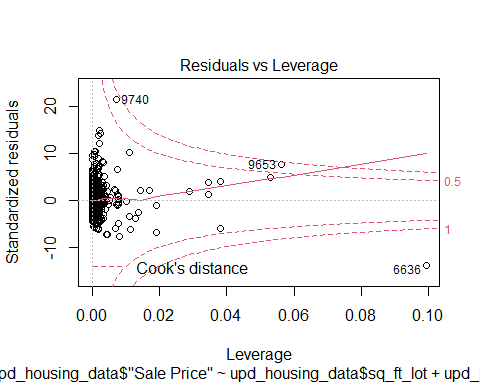
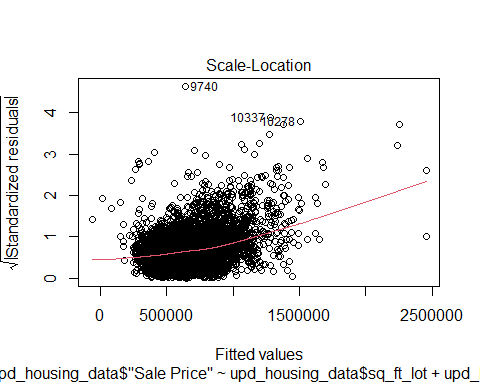
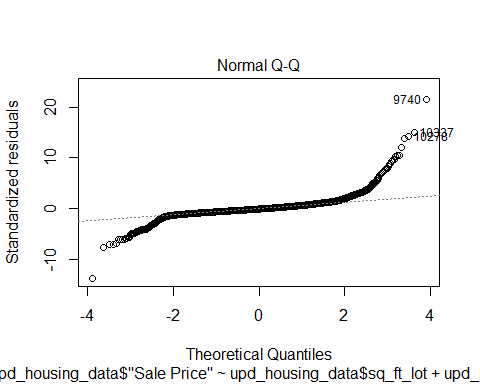
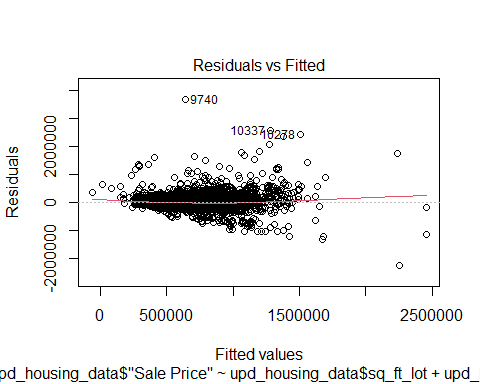
#Tolerence Statistics  
1/vif(sales\_price\_with\_others)

## upd\_housing\_data$sq\_ft\_lot   
## 0.9338039   
## upd\_housing\_data$square\_feet\_total\_living   
## 0.4460145   
## upd\_housing\_data$building\_grade   
## 0.4619948

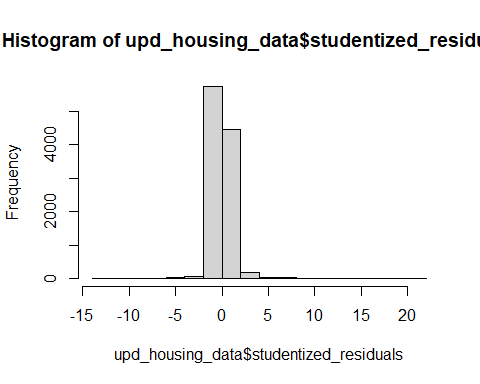
**Answer:** VIF values are all well below 10 and the tolerance statistics all well above 0.2. Also, the average VIF is very close to 1. Based on these measures we can safely conclude that there is no multicollinearity within our data.(Ref. - Discovering Statistics Using R (Field, Miles, and Field 2012, 428) )

**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(sales\_price\_with\_others)



hist(upd\_housing\_data$studentized\_residuals)



Residuals vs. fitted plots depict a random pattern, which means the assumptions of randomness are met. Primarily 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 shows less normality, but this can happen due to a smaller sample than the population. The Q-Q plot shows less normality but this can happen due to smaller sample than the population. The histogram indicates that the distribution is roughly normal or skewed a little to right; tail is almost on 10.

(Ref. - Discovering Statistics Using R (Field, Miles, and Field 2012, 428) )

**0. 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 Q-Q plot also showed significant curves at the ends, indicating that there are extreme values in the data set that make the model deviate from normality. Based on the above histogram, which looks like a normal distribution, we can say that the model is both accurate and generalizable to the population. But we need to keep an eye on problematic record observed in assignment question -K; It would be beneficial that if the model is re-created after the problematic record and outliers removed from the data set

References:

Field, A., J. Miles, and Z. Field. 2012. *Discovering Statistics Using R*. SAGE Publications. <https://books.google.com/books?id=wd2K2zC3swIC>.