Housing Data 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 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. If you worked with the Housing dataset in previous week – you are in luck, you likely have already found any issues in the dataset and made the necessary transformations. If not, you will want to take some time looking at the data with all your new skills and identifying if you have any clean up that needs to happen. Complete the following: Explain any transformations or modifications you made to the dataset

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(readxl)  
# tinytex::install\_tinytex() to fix the error with pdf knit  
  
housing\_data = read\_excel("C:/Users/brean/OneDrive/Desktop/NucampFolder/projects/dsc520-1/data/week-7-housing.xlsx")  
colnames(housing\_data)[2] <- "sale\_price"  
colnames(housing\_data)[14] <- "square\_feet"  
colnames(housing\_data)[1] <- "sale\_date"  
#I changed the column names because it made it difficult to grab that data.

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.

library(ggplot2)  
library(readxl)  
library(readr)  
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✔ tibble 3.1.8 ✔ stringr 1.5.0  
## ✔ tidyr 1.2.1 ✔ forcats 0.5.2  
## ✔ purrr 0.3.5   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(knitr)  
library(car)

## Loading required package: carData  
##   
## Attaching package: 'car'  
##   
## The following object is masked from 'package:purrr':  
##   
## some  
##   
## The following object is masked from 'package:dplyr':  
##   
## recode

sale\_sq <- lm(formula = sale\_price ~ sq\_ft\_lot, data=housing\_data)  
sale\_bed <- lm(formula = sale\_price ~ bedrooms + square\_feet + year\_built, data=housing\_data)  
#I used befrooms, square feet, and year\_built because these could all affect sale price.

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(sale\_sq)

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

# r2 is 0.01435 and adjusted r2 is 0.01428.   
  
summary(sale\_bed)

##   
## Call:  
## lm(formula = sale\_price ~ bedrooms + square\_feet + year\_built,   
## data = housing\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1743544 -120674 -42905 44943 3905060   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -4.858e+06 3.916e+05 -12.41 < 2e-16 \*\*\*  
## bedrooms -1.255e+04 4.499e+03 -2.79 0.00528 \*\*   
## square\_feet 1.784e+02 4.183e+00 42.66 < 2e-16 \*\*\*  
## year\_built 2.564e+03 1.967e+02 13.04 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 357400 on 12861 degrees of freedom  
## Multiple R-squared: 0.2189, Adjusted R-squared: 0.2187   
## F-statistic: 1202 on 3 and 12861 DF, p-value: < 2.2e-16

# r2 is 0.2189 and adjusted r2 is 0.2187  
  
#the low r2 of sale price and square feet of lot means that they aren't super connected. But if you look at the sale price based on the square foot, number of bedrooms, and year built, the r2 is still low but it means that they are connected but not enough for precise predictions. The sale price is based on more factors than what was used.

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(lm.beta)  
  
#anova(sale\_sq, sale\_bed)  
lm.beta(sale\_sq)

##   
## Call:  
## lm(formula = sale\_price ~ sq\_ft\_lot, data = housing\_data)  
##   
## Standardized Coefficients::  
## (Intercept) sq\_ft\_lot   
## NA 0.1198122

lm.beta(sale\_bed)

##   
## Call:  
## lm(formula = sale\_price ~ bedrooms + square\_feet + year\_built,   
## data = housing\_data)  
##   
## Standardized Coefficients::  
## (Intercept) bedrooms square\_feet year\_built   
## NA -0.02719051 0.43677937 0.10916845

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

library(MASS, pos = 18)

##   
## Attaching package: 'MASS'

## The following object is masked \_by\_ 'package:dplyr':  
##   
## select

with(housing\_data, (t.test(sale\_price, sq\_ft\_lot, alternative = "two.sided", conf.level = .95)))

##   
## Welch Two Sample t-test  
##   
## data: sale\_price and sq\_ft\_lot  
## t = 177.35, df = 13374, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 631451.9 645566.4  
## sample estimates:  
## mean of x mean of y   
## 660737.75 22228.57

#the mean of x falls within the confidence interval so it is the true mean. The 95% confidence interval means that 95% of the data contains the true mean.

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.

library(QuantPsyc)

## Loading required package: boot

##   
## Attaching package: 'boot'

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

##   
## Attaching package: 'QuantPsyc'

## The following object is masked from 'package:lm.beta':  
##   
## lm.beta

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

#compareCoefs(sale\_sq, sale\_bed)  
anova(sale\_sq, sale\_bed)

## Analysis of Variance Table  
##   
## Model 1: sale\_price ~ sq\_ft\_lot  
## Model 2: sale\_price ~ bedrooms + square\_feet + year\_built  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 12863 2.0734e+15   
## 2 12861 1.6431e+15 2 4.3031e+14 1684.1 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

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

all\_variables <-   
 lm(sale\_price ~ square\_feet + bath\_3qtr\_count + bath\_full\_count + bath\_half\_count + bedrooms + building\_grade + lat + lon + present\_use + sale\_instrument + sale\_reason + sq\_ft\_lot + year\_built + year\_renovated + sale\_reason,  
 data=housing\_data)  
summary(all\_variables)

##   
## Call:  
## lm(formula = sale\_price ~ square\_feet + bath\_3qtr\_count + bath\_full\_count +   
## bath\_half\_count + bedrooms + building\_grade + lat + lon +   
## present\_use + sale\_instrument + sale\_reason + sq\_ft\_lot +   
## year\_built + year\_renovated + sale\_reason, data = housing\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2258363 -120578 -44201 41877 3690935   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3.996e+07 9.718e+06 -4.112 3.94e-05 \*\*\*  
## square\_feet 1.479e+02 6.530e+00 22.650 < 2e-16 \*\*\*  
## bath\_3qtr\_count -1.654e+04 6.938e+03 -2.384 0.0171 \*   
## bath\_full\_count -1.171e+03 7.598e+03 -0.154 0.8775   
## bath\_half\_count -2.421e+03 7.155e+03 -0.338 0.7351   
## bedrooms -1.038e+04 4.909e+03 -2.114 0.0345 \*   
## building\_grade 2.808e+04 4.488e+03 6.257 4.06e-10 \*\*\*  
## lat -5.680e+04 1.388e+05 -0.409 0.6823   
## lon -2.997e+05 7.229e+04 -4.146 3.41e-05 \*\*\*  
## present\_use -7.525e+02 1.049e+02 -7.176 7.57e-13 \*\*\*  
## sale\_instrument 1.040e+02 1.038e+03 0.100 0.9202   
## sale\_reason -1.161e+04 1.281e+03 -9.064 < 2e-16 \*\*\*  
## sq\_ft\_lot 3.921e-01 6.121e-02 6.406 1.55e-10 \*\*\*  
## year\_built 3.110e+03 2.677e+02 11.617 < 2e-16 \*\*\*  
## year\_renovated 8.070e+01 1.433e+01 5.632 1.82e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 353700 on 12850 degrees of freedom  
## Multiple R-squared: 0.2359, Adjusted R-squared: 0.235   
## F-statistic: 283.3 on 14 and 12850 DF, p-value: < 2.2e-16

outlierTest(all\_variables)

## rstudent unadjusted p-value Bonferroni p  
## 11992 10.50876 1.0009e-25 1.2877e-21  
## 6430 10.44967 1.8597e-25 2.3925e-21  
## 6438 10.41514 2.6666e-25 3.4306e-21  
## 6437 10.40765 2.8831e-25 3.7091e-21  
## 6431 10.30227 8.5933e-25 1.1055e-20  
## 6436 10.27046 1.1923e-24 1.5339e-20  
## 6441 10.25983 1.3299e-24 1.7110e-20  
## 6432 10.22902 1.8242e-24 2.3468e-20  
## 6442 10.19140 2.6799e-24 3.4477e-20  
## 6433 10.16186 3.6211e-24 4.6586e-20

housing\_data\_updated <- housing\_data[-c(11992,6430,6438,6437,6431,6436,6441,6432,6442,6433),  
 ]

str((housing\_data\_updated))

## tibble [12,855 × 24] (S3: tbl\_df/tbl/data.frame)  
## $ sale\_date : POSIXct[1:12855], format: "2006-01-03" "2006-01-03" ...  
## $ sale\_price : num [1:12855] 698000 649990 572500 420000 369900 ...  
## $ sale\_reason : num [1:12855] 1 1 1 1 1 1 1 1 1 1 ...  
## $ sale\_instrument: num [1:12855] 3 3 3 3 3 15 3 3 3 3 ...  
## $ sale\_warning : chr [1:12855] NA NA NA NA ...  
## $ sitetype : chr [1:12855] "R1" "R1" "R1" "R1" ...  
## $ addr\_full : chr [1:12855] "17021 NE 113TH CT" "11927 178TH PL NE" "13315 174TH AVE NE" "3303 178TH AVE NE" ...  
## $ zip5 : num [1:12855] 98052 98052 98052 98052 98052 ...  
## $ ctyname : chr [1:12855] "REDMOND" "REDMOND" NA "REDMOND" ...  
## $ postalctyn : chr [1:12855] "REDMOND" "REDMOND" "REDMOND" "REDMOND" ...  
## $ lon : num [1:12855] -122 -122 -122 -122 -122 ...  
## $ lat : num [1:12855] 47.7 47.7 47.7 47.6 47.7 ...  
## $ building\_grade : num [1:12855] 9 9 8 8 7 7 10 10 9 8 ...  
## $ square\_feet : num [1:12855] 2810 2880 2770 1620 1440 4160 3960 3720 4160 2760 ...  
## $ bedrooms : num [1:12855] 4 4 4 3 3 4 5 4 4 4 ...  
## $ bath\_full\_count: num [1:12855] 2 2 1 1 1 2 3 2 2 1 ...  
## $ bath\_half\_count: num [1:12855] 1 0 1 0 0 1 0 1 1 0 ...  
## $ bath\_3qtr\_count: num [1:12855] 0 1 1 1 1 1 1 0 1 1 ...  
## $ year\_built : num [1:12855] 2003 2006 1987 1968 1980 ...  
## $ year\_renovated : num [1:12855] 0 0 0 0 0 0 0 0 0 0 ...  
## $ current\_zoning : chr [1:12855] "R4" "R4" "R6" "R4" ...  
## $ sq\_ft\_lot : num [1:12855] 6635 5570 8444 9600 7526 ...  
## $ prop\_type : chr [1:12855] "R" "R" "R" "R" ...  
## $ present\_use : num [1:12855] 2 2 2 2 2 2 2 2 2 2 ...

all\_variables\_2 <-   
 lm(sale\_price ~ square\_feet + bath\_3qtr\_count + bath\_full\_count + bath\_half\_count + bedrooms + building\_grade + lat + lon + present\_use + sale\_instrument + sale\_reason + sq\_ft\_lot + year\_built + year\_renovated + sale\_reason,  
 data=housing\_data\_updated)

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

housing\_data\_updated$standardized.residuals <- rstandard(all\_variables\_2)  
housing\_data\_updated$studentized.residuals <- rstudent(all\_variables\_2)  
housing\_data\_updated$cooks.distance <- cooks.distance(all\_variables\_2)  
housing\_data\_updated$dfbeta <- dfbeta(all\_variables\_2)  
housing\_data\_updated$leverage <- hatvalues(all\_variables\_2)  
housing\_data\_updated$covariance.ratios <- covratio(all\_variables\_2)

Use the appropriate function to show the sum of large residuals. Which specific variables have large residuals (only cases that evaluate as TRUE)?

housing\_data\_updated$large.residual <- housing\_data\_updated$standardized.residuals > 2 | housing\_data\_updated$studentized.residuals < -2  
  
sum(housing\_data\_updated$large.residual)

## [1] 327

Investigate further by calculating the leverage, cooks distance, and covariance rations. Comment on all cases that are problematics. Perform the necessary calculations to assess the assumption of independence and state if the condition is met or not.

housing\_data\_updated[housing\_data\_updated$large.residual , c("leverage" , "cooks.distance","covariance.ratios") ]

## # A tibble: 327 × 3  
## leverage cooks.distance covariance.ratios  
## <dbl> <dbl> <dbl>  
## 1 0.00145 0.000633 0.995  
## 2 0.0110 0.00297 1.01   
## 3 0.00311 0.00201 0.993  
## 4 0.00174 0.000819 0.995  
## 5 0.000850 0.000272 0.996  
## 6 0.00115 0.000321 0.997  
## 7 0.00241 0.00228 0.987  
## 8 0.00170 0.000723 0.995  
## 9 0.183 0.147 1.21   
## 10 0.00536 0.00471 0.991  
## # … with 317 more rows

dwt(all\_variables\_2)

## lag Autocorrelation D-W Statistic p-value  
## 1 0.7203714 0.5592475 0  
## Alternative hypothesis: rho != 0

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

vif(all\_variables\_2)

## square\_feet bath\_3qtr\_count bath\_full\_count bath\_half\_count bedrooms   
## 4.297176 2.092272 2.514983 1.457975 1.902750   
## building\_grade lat lon present\_use sale\_instrument   
## 2.474186 1.314517 1.465197 1.032218 1.221431   
## sale\_reason sq\_ft\_lot year\_built year\_renovated   
## 1.212256 1.249764 2.186311 1.092614

mean(vif(all\_variables\_2))

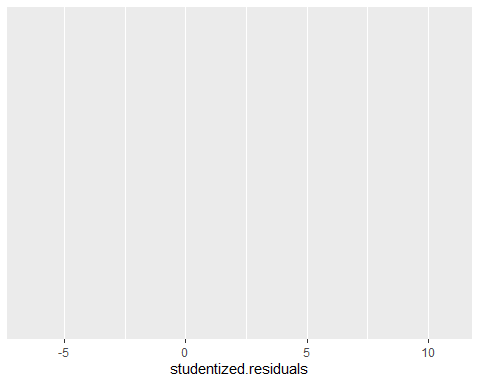
## [1] 1.822404

1/vif(all\_variables\_2)

## square\_feet bath\_3qtr\_count bath\_full\_count bath\_half\_count bedrooms   
## 0.2327110 0.4779493 0.3976170 0.6858827 0.5255551   
## building\_grade lat lon present\_use sale\_instrument   
## 0.4041734 0.7607360 0.6825021 0.9687876 0.8187120   
## sale\_reason sq\_ft\_lot year\_built year\_renovated   
## 0.8249084 0.8001511 0.4573914 0.9152360

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.

library(ggplot2)  
all\_variables\_2\_fitted <- all\_variables\_2$fitted.values  
  
histogram <- ggplot(housing\_data\_updated , aes(studentized.residuals))  
histogram



#with(housing\_data\_updated, discretePlot(bedrooms, by = prop\_type, scale = "frequency"))  
  
#with(housing\_data\_updated, Hist(standardized.residuals, scale="frequency", breaks="Sturges", col="red", xlab="Studentized Residuals"))  
  
  
#could not get plots to work.

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