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)

housing_data = read_excel("C:/Users/brean/OneDrive/Desktop/NucampFolder/projects/dsc520-1/data/week-7-h
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)
setwd("C:/Users/brean/OneDrive/Desktop/NucampFolder/projects/dsc520-1")
housing_data = read_excel("data/week-7-housing.xlsx")
print(housing_data)
```

```
## # A tibble: 12,865 x 24
##    `Sale Date`         Sale Pric~1 sale_~2 sale_~3 sale_~4 sitet~5 addr_~6  zip5
##    <dttm>                    <dbl>   <dbl>   <dbl> <chr>   <chr>   <chr>   <dbl>
##  1 2006-01-03 00:00:00      698000       1       3 <NA>    R1      17021 ~ 98052
##  2 2006-01-03 00:00:00      649990       1       3 <NA>    R1      11927 ~ 98052
##  3 2006-01-03 00:00:00      572500       1       3 <NA>    R1      13315 ~ 98052
##  4 2006-01-03 00:00:00      420000       1       3 <NA>    R1      3303 1~ 98052
##  5 2006-01-03 00:00:00      369900       1       3 15      R1      16126 ~ 98052
##  6 2006-01-03 00:00:00      184667       1      15 18 51   R1      8101 2~ 98053
##  7 2006-01-04 00:00:00     1050000       1       3 <NA>    R1      21634 ~ 98053
##  8 2006-01-04 00:00:00      875000       1       3 <NA>    R1      21404 ~ 98053
##  9 2006-01-04 00:00:00      660000       1       3 <NA>    R1      7525 2~ 98053
## 10 2006-01-04 00:00:00      650000       1       3 <NA>    R1      17703 ~ 98052
## # ... with 12,855 more rows, 16 more variables: 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>, and abbreviated variable names
## #   1: ‘Sale Price‘, 2: sale_reason, 3: sale_instrument, 4: sale_warning, ...
```

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?

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?

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

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. Perform casewise diagnostics to identify outliers and/or influential cases, storing each function's output in a dataframe assigned to a unique variable name. Calculate the standardized residuals using the appropriate command, specifying those that are +-2, storing the results of large residuals in a variable you create. Use the appropriate function to show the sum of large residuals. Which specific variables have large residuals (only cases that evaluate as TRUE)? 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. Perform the necessary calculations to assess the assumption of no multicollinearity and state if the condition is met or not. 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. Overall, is this regression model unbiased? If an unbiased regression model, what does this tell us about the sample vs. the entire population model?