# week 8

2022-05-16

library (readxl)

```
setwd("/Users/sarah/documents/hello_world")
housing <- read_xlsx("/Users/sarah/documents/hello_world/data/week-7-housing.xlsx")
```

#Explain any transformations or modifications you made to the dataset #remove human behavior variables such as sale reason, instrument, address, etc.
options(scipen = 999)
Price.out1 <- which(housing‘$SalePrice$‘`Sale Price`, col = “springgreen”)$out)sqft.out1 < -which(housing$sq_ft_lot
%in% boxplot(housing$sq_ft_lot, col = "skyblue"$)out)
length(square_feet_total_living1)

```
#remove outliers
outliers1 <- c(Price.out1,sqft.out1)
Sale.Price1 <- c(housing$‘Sale Price‘[-outliers1])
sq_ft_lot1 <- housing$sq_ft_lot[-outliers1]
hist(housing$‘Sale Price‘, col = "green")
hist(housing$sq_ft_lot, col = "blue")
```

#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.

model1 <- lm(housing‘$SalePrice$‘ $housing$sq_ft_lot)
temp1 <- c(housing$square_feet_total_living, housing$outliers1)
square_feet_total_living1 <- c(temp1[-outliers1])
temp2 <- c(housing$bedrooms, outliers1)bedrooms1 < -c(temp2[-outliers1])temp3 < -c(housing$year_renovated[outliers1])
year_renovated1 <- c(temp3[-outliers1])
length(year_renovated1)
model2 <- lm(Sale.Price1~square_feet_total_living1)

Besides the normal variables, I created several variables such as saleprice1, sq_ft_lot1, square_feet_total_living1 and bedrooms1 after removing the outliers. The outliers are the extreme value ones. I want to compare how the outliers would affect the results.

```
#Execute a summary() function on two variables defined in the previous step to compare the model results
summary (model1)
```

lm(formula = housing‘$SalePrice$‘ $housing$sq_ft_lot)

Residuals: Min 1Q Median 3Q Max -2016064 -194842 -63293 91565 3735109

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 641821.40609 3799.91526 168.90 <0.0000000000000002 ***housing$sq_ft_lot 0.85099 0.06217 13.69 <0.0000000000000002*** — 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: < 0.00000000000000022

```
summary (model2)
lm(formula = Sale.Price1 ~ square_feet_total_living1)

Residuals:
    Min      1Q  Median      3Q     Max 
-843931  -77311   -5525   70849  623565 

Coefficients:
                            Estimate Std. Error t value            Pr(>|t|)    
(Intercept)               182564.404   4139.361    44.1 <0.0000000000000002 ***
square_feet_total_living1    170.459      1.668   102.2 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 130500 on 10122 degrees of freedom
Multiple R-squared:  0.5079,    Adjusted R-squared:  0.5078 
F-statistic: 1.045e+04 on 1 and 10122 DF,  p-value: < 0.00000000000000022
```

#Calculate the confidence intervals for the parameters in your model and explain what the results indicate.
model3 <- lm(Sale.Price1~scale(square_feet_total_living1)+scale(bedrooms1)+scale(year_renovated1))
summary(model3)
#These values indicate that an increase of one standard deviation of a predictor variable keeping all the other predictor variables constant, results in an expected change of the respective regression coefficient in Sale.price1.

```
#Perform casewise diagnostics to identify outliers and/or influential cases, storing each function’s ou
outliers1 <- c(Price.out1,sqft.out1)
Sale.Price1 <- c(Sale.Price[-outliers1])
sq_ft_lot1 <- sq_ft_lot[-outliers1]
hist(Sale.Price1, col = "green")
hist(sq_ft_lot1, col = "blue")
#In the second model, we use square_feet_total_living, bedrooms, and year_renovated as our additional p
```

#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(model1)