# Task 4

Goal:

Build and test a regression model in R using average monthly incidents per capita as the dependent variable and our socioeconomic datapoints we previously joined in task 3 as our independent variables.

Data in:

Task 3 Dataset with aggregated city data and joined socioeconomic data.

Result:

Optimized regression model for analysis

Process:

**Model Building Steps:**

***#Installing and loading the packages***

*Packages <- c('tidyverse', 'mice', 'readxl', 'MASS','ggplot2', 'tidyr', 'lubridate','lattice', 'car')*

*lapply(Packages, library, character.only = TRUE)*

*install.packages("caret")*

*library(caret)*

*install.packages("olsrr")*

*library(olsrr)*

***#Reading the datafile***

*df <- read\_excel(file.choose())*

*head(df)*

*summary(df)*

***#Removing the missing values from the dataset***

*df2 <- df[complete.cases(df), ]*

***#Creating Training and Testing datasets for the model in 85:15 proportions of data respectively***

*sample <- sample.int(n = nrow(df2),*

*size = floor(.85\*nrow(df2)), replace = F)*

*train <- df2[sample, ]*

*test <- df2[-sample, ]*

*head(train)*

***#Model 1 including the variables from the dataset. Dependent Variable here is to predict Average incidents that happen per month per capita. Independent variables are socio economic variables.***

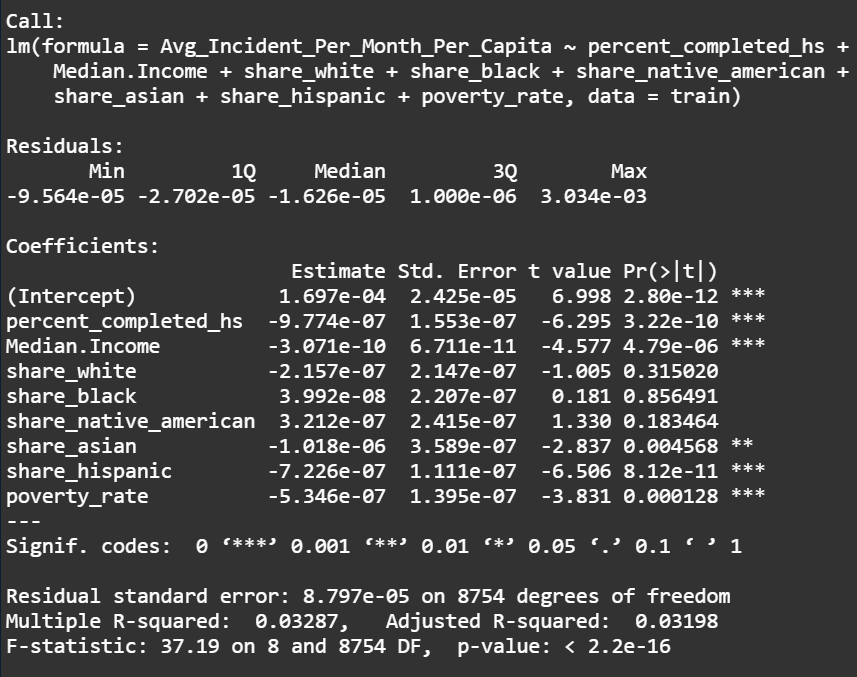
*reg1 <- lm(Avg\_Incident\_Per\_Month\_Per\_Capita ~ percent\_completed\_hs + Median.Income + share\_white + share\_black + share\_native\_american + share\_asian + share\_hispanic + poverty\_rate, train)*

*summary(reg1)*

*plot(reg1)*

*ols\_plot\_cooksd\_bar(reg1)*

***#Output***



*This Model had passed the F Test, but we can find that few variables are not jointly significant for the model “share\_white”, “share\_black”, “share\_native\_american”. The Adjusted R square value is 0.03198, which is incredibly low. So, for the next model we will run the model by removing the non-significant variables.*

***#Model 2***

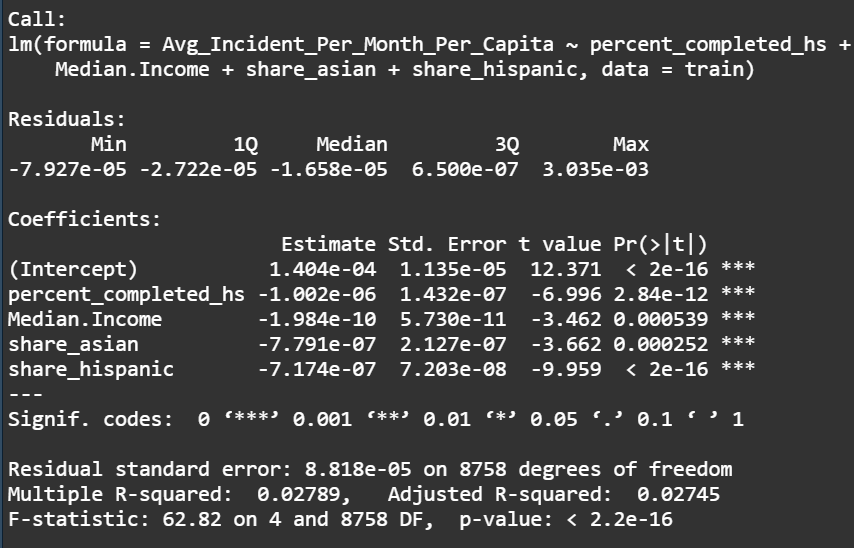
*reg2 <- lm(Avg\_Incident\_Per\_Month\_Per\_Capita ~ percent\_completed\_hs + Median.Income + share\_asian + share\_hispanic, train)*

*summary(reg2)*

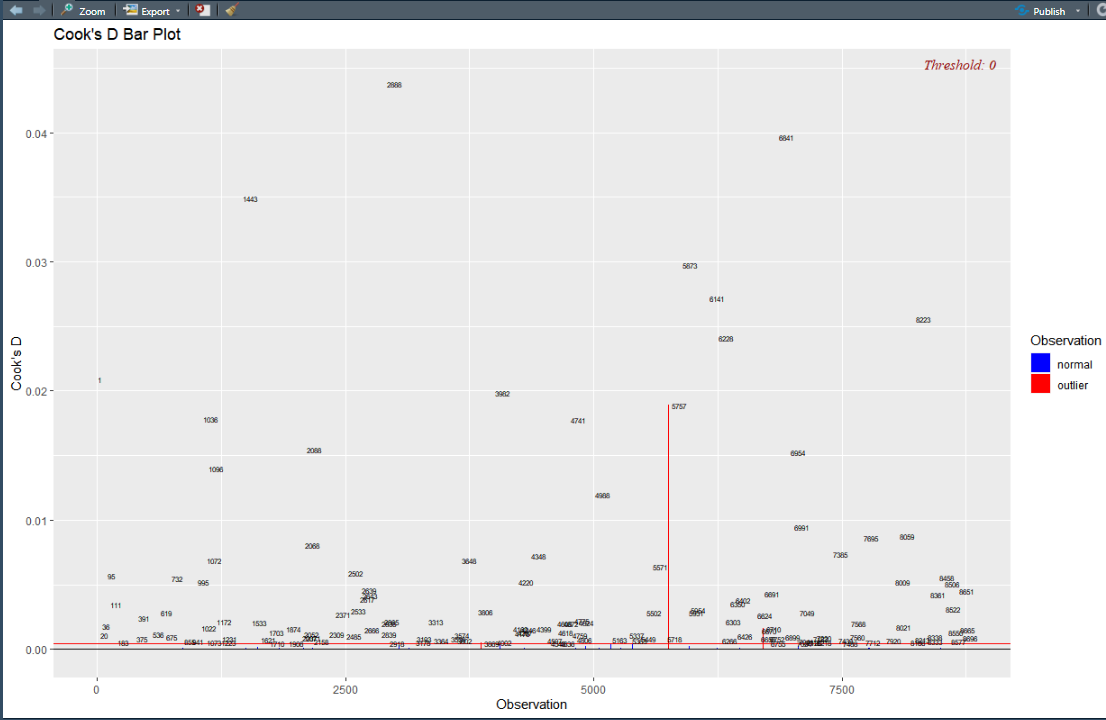
*par(mfrow=c(2,2)) #0.02745*

*plot(reg2) #769 478 4294 6469*

*ols\_plot\_cooksd\_bar(reg2)*



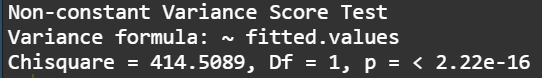
*After running the model, we see that the variables are jointly significant for the model, however the R square for the real dataset is extremely low which is a problem. Considering this into the account, we have checked for the* ***outliers*** *in the model, and it shows that there are too many outliers in the dataset for each column.*



*The above figure is the cooks plot for the regression 2, we find many outliers in the model, however if we delete the outliers, we see some significant increase in the t test of non-significant variables and R square.*

***#Testing heteroskedasticity for the model reg2***

*ncvTest(reg2)*



*We see that there is heteroskedasticity in the model.*

Observations:

Our initial model had an incredibly low adjusted R-squared value of 0.01, and when we filtered out lower population cities, we saw incremental improvement in our adjusted R-squared as we filtered cities out with less than 50, 100, 500, 1000, 5000, and 10000 total population. We used the Excel geography data type to retrieve population statistics for all our cities, and some of the values returned were incorrect. (e.g., Mandan, North Dakota returned a population of 50 via Excel but a quick google search revealed a population over 22,000) This massive and unpredictable inaccuracy strongly affected our dependent variable value and rendered our R-squared nearly unusable. This issue seemed to happen more often with smaller cities with populations under 1000, which in turn led us to filtering out low population cities and observing R-squared improvements up to nearly 0.4 from a measly 0.01.