Part B – Explanation of Importing and Cleansing the Data

I started the process by setting my working directory to the same file path as where I placed the dataset using *"setwd("C:/Users/tedda/Desktop/C997")".* Next, I imported the dataset using readxl::read_excel into the object raw_data.

Using dplyr::select and dplyr::filter, I selected the specific columns I needed initially which were NAME and the Year columns (2010:2019) using *"data <- raw_data %>% select(NAME, "2010":"2019")".* After, I filtered to only rows that contained Pennsylvania data using *"data <- filter(data, NAME == "Pennsylvania")".* Seeing as I only needed the data from the years, I removed the NAME column containing "Pennsylvania" by using *"data2 <- data %>% select("2010":"2019")".*

The data is still in a horizontal format, but also not useable with lm(). I created a new dataframe from data2 by using data.frame and transposing the data to show vertically using t(). The code was "*df2<-data.frame(x=2010:2019, y=t(data2))".* The rownames were still years so I reset the rownames using *"rownames(df2) <- NULL*".

```
# Set the working directory to where you have the excel file located
setwd("C:/Users/tedda/Desktop/C997")

# Import the raw dataset using readxl::read_excel
raw_data <- read_excel("nst-est2019-alldata.xlsx")

#Cleanse the data for Pennsylvania 2010 - 2019 using dplyr::select & dplyr::filter
data <- raw_data %>% select(NAME, "2010":"2019")
data <- filter(data, NAME == "Pennsylvania")

#Remove the NAME column using dplyr::select
data2 <- data %>% select("2010":"2019")

#Create a dataframe from data2 using data.frame
#t(data2) is used to transpose the data from horizontal to vertical
#2010:2019 creates a column of Years
df2<-data.frame(x=2010:2019, y=t(data2))
rownames(df2) <- NULL #This resets the rownames
```