ANA 515 Assignment 2, Loading, Saving, and Describing Data

Mel Sparks

4/10/2022

## Greenhouse Gas Emissions Dataset

This dataset contains information regarding how many tons of greenhouse gas emissions are produced annually in the world and by country from 1990 to 2018. I’m hoping to ascertain whether greenhouse gas production has increased or decreased worldwide, as well as if any top producing countries have made progress towards decreasing their annual output. The dataset is saved in a csv, thusly the delimiter in this case is the comma. Additionally, we can open csv files in excel, which I frequently do before reading them into R so I already know approximately what they should look like.

ghgemissions <- read.csv("C:/Users/melin/OneDrive/Desktop/Data Analytics M.S/ANA515/Week 4/ALL GHG\_historical\_emissions.csv")  
#I used the read.csv function because I knew the file I was using was a csv; this function should be a base R function.

#I want to drop columns I don't need; all of these columns have the same entry for all observations  
ghgdrop <- subset(ghgemissions, select = -c(Data.source, Sector,Gas,Unit))  
#Now I'll limit my data to the top producing countries, any who produced less than 1000 tons of ghg emissions in 2018  
ghgtopprod <- filter(ghgdrop, X2018>1000)  
#Having the World total is useful, but it will skew my summary statistics at the end of this, so I'll put it in it's own dataset  
ghgworld <- filter(ghgtopprod, X2018>20000)  
ghgtopprod <- filter(ghgtopprod, X2018<20000)  
#Apparently X1990 was accepted as a character class, I need to change that to integer  
ghgtopprod$X1990 <- as.integer(ghgtopprod$X1990)

## This dataframe has 8 rows and 30 columns. The names of the columns and a brief description of each are in the table below:

#this makes a new data.frame called text\_tbl with two columns, Names and Description  
text\_tbl <- data.frame(  
Names = c("Country", "X2018", "X2017", "...", "X1990"),  
Description = c("Specific country or world overview", "All emissions in 2018 measured in tons", "All emissions in 2017 measured in tons", "...", "All emissions in 1990 measured in tons")  
)  
text\_tbl #prints the table

## Names Description  
## 1 Country Specific country or world overview  
## 2 X2018 All emissions in 2018 measured in tons  
## 3 X2017 All emissions in 2017 measured in tons  
## 4 ... ...  
## 5 X1990 All emissions in 1990 measured in tons

#For the three columns, I will select the most recent year, the oldest year, and a year approximately midspaced between the two  
datapick3 <- select(ghgtopprod, X2018, X2008, X1990)  
#And now I'll produce a table with the summary statistics  
SummaryTable<-summary(datapick3)  
SummaryTable

## X2018 X2008 X1990   
## Min. : 1155 Min. :1088 Min. :1009   
## 1st Qu.: 1633 1st Qu.:1579 1st Qu.:1220   
## Median : 2663 Median :2174 Median :2258   
## Mean : 3806 Mean :3352 Mean :2575   
## 3rd Qu.: 3959 3rd Qu.:4436 3rd Qu.:3234   
## Max. :11706 Max. :8456 Max. :5543

#Note: I accidentally picked a dataset with no missing values, so I was not able to provide a summary of that info