R Notebook

#QUESTION 3 The following code will download the data from the wiki page above and create a dataframe named top Cities  
  
#### Start Top Cities Code ####  
#library(reprex) -- I commented out the libraries I did not use  
library(tidyverse)

## ── Attaching packages ──────────────────────────────────────────────────────────────────────────────────────────────────────── tidyverse 1.3.0 ──

## ✓ ggplot2 3.3.2 ✓ purrr 0.3.4  
## ✓ tibble 3.0.3 ✓ dplyr 1.0.2  
## ✓ tidyr 1.1.1 ✓ stringr 1.4.0  
## ✓ readr 1.3.1 ✓ forcats 0.5.0

## ── Conflicts ─────────────────────────────────────────────────────────────────────────────────────────────────────────── tidyverse\_conflicts() ──  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(rvest)

## Loading required package: xml2

##   
## Attaching package: 'rvest'

## The following object is masked from 'package:purrr':  
##   
## pluck

## The following object is masked from 'package:readr':  
##   
## guess\_encoding

#library(readr)  
#library(dplyr)  
  
topCities <- read\_html("https://en.wikipedia.org/wiki/List\_of\_cities\_proper\_by\_population")  
topCities <- html\_table(  
html\_node(  
topCities, "table.sortable"  
), header=TRUE, trim=F, fill = TRUE  
)[-1,]  
names(topCities) <- c(  
"City", "Country", "Image", "Population", "City.Def", "City.Population", "City.Area", "Metro.Pop",  
"Metro.Area", "Urban.Pop", "Urban.Area"  
)  
  
topCities$Image <- NULL  
  
## Old String Substitute (gsub)  
topCities$Population <- as.numeric(gsub(",", "", topCities$Population))  
  
## tidyverse String Substitute (str\_replace\_all)  
topCities$City<-str\_replace\_all(topCities$City, "\n", "")  
topCities$Country<-str\_replace\_all(topCities$Country, "\n", "")  
topCities$City.Def<-str\_replace\_all(topCities$City.Def, "\n", "")  
topCities$City.Area<-str\_replace\_all(topCities$City.Area, "\n", "")  
topCities$Metro.Pop<-str\_replace\_all(topCities$Metro.Pop, "\n", "")  
topCities$Metro.Area<-str\_replace\_all(topCities$Metro.Area, "\n", "")  
topCities$Urban.Pop<-str\_replace\_all(topCities$Urban.Pop, "\n", "")  
topCities$Urban.Area<-str\_replace\_all(topCities$Urban.Area, "\n", "")  
#### End Top Cities Code ####

head(topCities)

## City Country Population City.Def City.Population  
## 2 Tokyo  Japan 37400068 Metropolis prefecture 13,515,271[14]  
## 3 Delhi  India 28514000 National capital territory 16,753,235[16]  
## 4 Shanghai  China 25582000 Municipality 24,183,000[18]  
## 5 São Paulo  Brazil 21650000 Municipality 12,252,023[19]  
## 6 Mexico City  Mexico 21581000 City-state 8,918,653[21]  
## 7 Cairo  Egypt 20076000 Urban governorate 9,500,000[23]  
## City.Area Metro.Pop Metro.Area Urban.Pop Urban.Area  
## 2 2,191[14] 37,274,000[15] 13,452[15] 37,977,000 8,230[e]  
## 3 1,484 29,000,000[17] 3,483[17] 29,617,000 2,232[f]  
## 4 6,341 N/A N/A 22,120,000 4,068[g]  
## 5 1,521 21,734,682[20] 7,947 22,046,000 3,116[h]  
## 6 1,485 20,892,724[22] 7,854 20,996,000 2,386  
## 7 3,085 N/A N/A 19,372,000 2,010