Data 413: Homework (Web Scraping)

Frankie Tyndall

2022-07-08

## Run the following link that will direct you to a table on the internet. Using 4 or 5 sentences, discuss the content of the table.

The table above has a sample size of 235 countries and dependencies with 12 variables that looks at specific aspects of each country or dependency. The variables from range from population which looks at the number of people that reside in each country or dependency to the change is population and the land area of each. The density, amount of migrants, and fertility rate of each of the 235 countries/dependencies can also be found in the table. The fertility rate is average amount of children born (alive) to women of a country over their lifetime. The last three variables of the data set is the average age in each country/dependency, as well as, the urban population percentage, and world share.

## Using the R coding structure that was illustrated in class, transfer the table observed on the internet into R Studio.The table will be imported in tibble form.

x <- read\_html("https://www.worldometers.info/world-population/population-by-country/")  
world\_pop <- x %>%  
 html\_table(., fill = T)  
  
world\_pop[[1]] -> world\_population   
  
world\_population

## # A tibble: 235 × 12  
## `#` `Country (or dependency)` `Population (2…` `Yearly Change` `Net Change`  
## <int> <chr> <chr> <chr> <chr>   
## 1 1 Honduras 9,904,607 1.63 % 158,490   
## 2 2 United Arab Emirates 9,890,402 1.23 % 119,873   
## 3 3 Djibouti 988,000 1.48 % 14,440   
## 4 4 Saint Barthelemy 9,877 0.30 % 30   
## 5 5 Seychelles 98,347 0.62 % 608   
## 6 6 Antigua and Barbuda 97,929 0.84 % 811   
## 7 7 Vietnam 97,338,579 0.91 % 876,473   
## 8 8 Hungary 9,660,351 -0.25 % -24,328   
## 9 9 Tajikistan 9,537,645 2.32 % 216,627   
## 10 10 Belarus 9,449,323 -0.03 % -3,088   
## # … with 225 more rows, and 7 more variables: `Density (P/Km²)` <chr>,  
## # `Land Area (Km²)` <chr>, `Migrants (net)` <chr>, `Fert. Rate` <chr>,  
## # `Med. Age` <chr>, `Urban Pop %` <chr>, `World Share` <chr>

## Now use R coding and dplyr functions to modify your table, until you get the exact final representation shown below

country\_namechange <- world\_population  
names(country\_namechange)[2]<-paste("Country")  
names(country\_namechange)[3]<-paste("Population")  
  
select\_country\_pop <- country\_namechange %>%   
 select(c(Country, Population))  
  
select\_country\_pop$Population <-  
 as.numeric(gsub(",","",select\_country\_pop$Population))  
  
select\_country\_pop

## # A tibble: 235 × 2  
## Country Population  
## <chr> <dbl>  
## 1 Honduras 9904607  
## 2 United Arab Emirates 9890402  
## 3 Djibouti 988000  
## 4 Saint Barthelemy 9877  
## 5 Seychelles 98347  
## 6 Antigua and Barbuda 97929  
## 7 Vietnam 97338579  
## 8 Hungary 9660351  
## 9 Tajikistan 9537645  
## 10 Belarus 9449323  
## # … with 225 more rows

country\_population <- select\_country\_pop %>%   
 filter(Country == "China"|Country == "United States"|Country == "Russia"  
 |Country == "Vietnam"|Country == "South Africa") %>%  
 group\_by(Country) %>%  
 arrange(desc(Population))  
  
country\_population

## # A tibble: 5 × 2  
## # Groups: Country [5]  
## Country Population  
## <chr> <dbl>  
## 1 China 1439323776  
## 2 United States 331002651  
## 3 Russia 145934462  
## 4 Vietnam 97338579  
## 5 South Africa 59308690