HW-webscraping

Lin Pin Tzu (Ruby)

2022-07-08

1) 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."https://www.worldometers.info/world-population/population-by-country/"

This website link contains both countries and their dependencies. The figures are based on the most recent United Nations Population Division estimates. For current estimates (live population clock), historical statistics, and forecast figures, we can click on the name of the nation or dependent. The world population of each country in 2022, and data has not only each country's population but also lots of information about each country, for example, net change, density, and land area. Which can make people know about and let us analyze and learn more about the information we want.

2) 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.

```
worldpop <- read_html(
  "https://www.worldometers.info/world-population/population-by-country/")
datatables <- worldpop%>%
  html_table(., fill = T)
datatables[[1]] -> world2022
row.names(world2022) <- NULL
names(world2022)[3] <- "Population"
names(world2022)[2] <- "Country"
world2022$Population <-parse_number(world2022$Population)
tibble(world2022)
```

```
## # A tibble: 235 x 12
##      `#` Country        Population `Yearly Change` `Net Change` `Density (P/Km~`
##    <int> <chr>               <dbl> <chr>           <chr>        <chr>
##  1     1 Honduras          9904607 1.63 %          158,490      89
##  2     2 United Arab E~    9890402 1.23 %          119,873      118
##  3     3 Djibouti           988000 1.48 %          14,440       43
##  4     4 Saint Barthel~       9877 0.30 %          30           470
##  5     5 Seychelles          98347 0.62 %          608          214
##  6     6 Antigua and B~      97929 0.84 %          811          223
##  7     7 Vietnam          97338579 0.91 %          876,473      314
##  8     8 Hungary           9660351 -0.25 %         -24,328      107
##  9     9 Tajikistan        9537645 2.32 %          216,627      68
```

```
## 10    10 Belarus           9449323 -0.03 %         -3,088       47
## # ... with 225 more rows, and 6 more variables: `Land Area (Km²)` <chr>,
## #   `Migrants (net)` <chr>, `Fert. Rate` <chr>, `Med. Age` <chr>,
## #   `Urban Pop %` <chr>, `World Share` <chr>
```

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

```
world2022 %>%
    select(Country,Population)%>%
   filter(Country =="China"|Country =="United States"|Country =="Russia"
          |Country =="Vietnam"|Country =="South Africa") %>%
  arrange(desc(Population))
```

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