Solution to exercise at bottom of `web_scraping_and_string_cleaning.R`

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```r
library(rvest)
```

```
##
## Attaching package: 'rvest'
##
## The following object is masked from 'package:readr':
##
##     guess_encoding
```

```r
wiki_country <- 'https://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_area'

country <- read_html(wiki_country)
tbl_list <- country |>
  html_elements("table") |>
  html_table()
tbl <- tbl_list[[2]]
tbl2 <- tbl |> select(-1,-7)

clean1string <- function(x){
  x <- str_extract(x, "\\(.*\\)") # Find pattern "(ANYTHING)"
  x <- str_remove_all(x, "\\(|\\,|\\)") # Remove anything matching "(", "," or ")"
  return(as.numeric(x))
}

tbl3 <- tbl2 |> mutate(
  total = clean1string(`Totalin km2 (mi2)`),
  land = clean1string(`Landin km2 (mi2)`),
  water = clean1string(`Waterin km2 (mi2)`))

tbl3 |> select(-(2:4))
```

```
## # A tibble: 264 x 5
##    `Country / dependency` `%water`     total      land      water
##    <chr>                     <dbl>     <dbl>    <dbl>     <dbl>
##  1 Earth                      70.8 196940000 57506000 139434000
##  2 Russia                      4.2   6601667  6323142    278530
##  3 Antarctica                  0     5480000  5480000        NA
##  4 Canada                      8.9   3855100  3511021    344080
##  5 China                       2.8   3705410  3600950    104460
##  6 United States               4     3677647  3531904    145724
##  7 Brazil                      0.6   3285862  3266583     21372
##  8 Australia                   0.8   2988900  2966200     22750
##  9 India                       9.6   1269219  1147960    121260
## 10 Argentina                   1.6   1073500  1056640     16880
## # i 254 more rows
```