A3

Kerui Du

04/10/2021

```
url = 'https://www.stat.auckland.ac.nz/~yongwang/stats380/population-by-country.html'
readin = readLines(url)
col_names = strsplit(readin[11],'thead')[[1]][2] # all column names with uninterested parts
col_names_mod = strsplit(col_names,'</th> <th>')[[1]][-1] # split by common pattern

cols <- function(i) { # write function to remove each pattern
  if (i==1)
    col_names_mod[i]
  else if (i!=1&i!=11)
    sub('<br>','',col_names_mod[i])
  else if (i==11)
    sub('<.+$','',sub('<br>','',col_names_mod[i]))
}
col_name = sapply(1:11, cols) # all column names

interest = grep('/world-population/',readin) # return 11
sep_part = strsplit(readin[interest],'/world-population/')[[1]][-1:-2] # common pattern
info = sub('^.+population/\">','',sep_part) # all information we want
matrix = matrix(0,nr=235,nc=11) # create a matrix to help us build a data frame easily

country <- function(infos) { # write a function to extract all country names
  country = sub('<.+$', '', infos) # extract country
  mod = grep('&amp;',country) # some countries need to be modified
  country[mod] <- sub('&amp;','&',country[mod]) # replace '&amp;' by '&'
  country[grep(';',country)] <- paste('Country', grep(';',country)) # modify some special letters
  country # our final result
}

info1 = sub('^.+;\">','',info)
info1 = strsplit(info1,'<td>') # separate rest allow us extract columns simply

for (i in 1:nrow(matrix)){
  matrix[i,1] = country(info)[i] # all Country with modification embed to 1st column
  for (j in 2:ncol(matrix)){
    index = j-1
    col = info1[[i]][index]
    if (index %in% c(2, 9, 10)){
      infos = sub(' %<.+$', '', col) # extract numbers with '%' and some may contain NA
      infos = sub('<.+$', '', infos) # get rid of unwanted
      infos = ifelse(infos=='N.A.',NA,infos) # get rid of unwanted
      matrix[i,j] = infos
    }
```

```
    else if (index %in% c(1, 3, 5, 6)){
      infos = sub('<.+$', '', col)  # get rid of unwanted
      infos = gsub(',','',infos)    # extract numbers with ',' but no NA
      infos = ifelse(infos==" ",NA,infos) # get rid of unwanted
      matrix[i,j] = infos
    }
    else if (index %in% c(4, 7, 8)){
      infos = sub('<.+$', '', col)  # get rid of unwanted
      infos = gsub(',','',infos)    # extract numbers with ',' and it has NA
      infos = ifelse(infos=='N.A.',NA,infos) # get rid of unwanted
      matrix[i,j] = infos
    }
  }
}

population <- as.data.frame(apply(matrix[,2:11], 2, as.numeric)) # convert all characters to numeric
population = cbind(matrix[,1],population) # add the country column
colnames(population) <- col_name # final data frame

dim(population)
```

```
## [1] 235  11
```

```
class(population)
```

```
## [1] "data.frame"
```

```
head(population)
```

```
##   Country (or dependency) Population (2020) Yearly Change Net Change
## 1                   China        1439323776          0.39    5540090
## 2                   India        1380004385          0.99   13586631
## 3           United States         331002651          0.59    1937734
## 4               Indonesia         273523615          1.07    2898047
## 5                Pakistan         220892340          2.00    4327022
## 6                  Brazil         212559417          0.72    1509890
##   Density (P/Km²) Land Area (Km²) Migrants (net) Fert. Rate Med. Age
## 1             153         9388211        -348399        1.7       38
## 2             464         2973190        -532687        2.2       28
## 3              36         9147420         954806        1.8       38
## 4             151         1811570         -98955        2.3       30
## 5             287          770880        -233379        3.6       23
## 6              25         8358140          21200        1.7       33
##   Urban Pop % World Share
## 1          61       18.47
## 2          35       17.70
## 3          83        4.25
## 4          56        3.51
## 5          35        2.83
## 6          88        2.73
```

```
tail(population)
```

```
##     Country (or dependency) Population (2020) Yearly Change Net Change
## 230 Saint Pierre & Miquelon              5794         -0.48        -28
## 231              Montserrat              4992          0.06          3
## 232        Falkland Islands              3480          3.05        103
## 233                    Niue              1626          0.68         11
## 234                 Tokelau              1357          1.27         17
## 235               Holy See               801          0.25          2
##     Density (P/Km²) Land Area (Km²) Migrants (net) Fert. Rate Med. Age
## 230              25             230             NA         NA       NA
## 231              50             100             NA         NA       NA
## 232               0           12170             NA         NA       NA
## 233               6             260             NA         NA       NA
## 234             136              10             NA         NA       NA
## 235            2003               0             NA         NA       NA
##     Urban Pop % World Share
## 230         100           0
## 231          10           0
## 232          66           0
## 233          46           0
## 234           0           0
## 235          NA           0
```

```
sum(population[[2]])
```

```
## [1] 7795232630
```