A4

Kerui Du

08/10/2021

Q1

```
fun1 <- function(country) {
  country = tolower(gsub(' ','-',country))
  sprintf('%s-population.html',country)
}
Population=unlist(lapply(oceania, fun1));Population
```

```
##  [1] "australia-population.html"
##  [2] "papua-new-guinea-population.html"
##  [3] "new-zealand-population.html"
##  [4] "fiji-population.html"
##  [5] "solomon-islands-population.html"
##  [6] "vanuatu-population.html"
##  [7] "new-caledonia-population.html"
##  [8] "french-polynesia-population.html"
##  [9] "samoa-population.html"
## [10] "kiribati-population.html"
## [11] "tonga-population.html"
## [12] "marshall-islands-population.html"
## [13] "northern-mariana-islands-population.html"
## [14] "american-samoa-population.html"
## [15] "cook-islands-population.html"
## [16] "tuvalu-population.html"
## [17] "wallis-and-futuna-islands-population.html"
## [18] "nauru-population.html"
## [19] "niue-population.html"
## [20] "tokelau-population.html"
```

```
writeLines(Population,'Population.html')
writeLines(Population,'Population')
```

Q2

```
url = "https://www.worldometers.info/world-population/"
tableInfo <- function(input) {
  file = readLines(input,warn=F)
  pattern = grep('<table',file)

  # extract the indexes of interest(start) from each pattern
  tableStart_list = sapply(pattern, function(i) gregexpr('<table',file[i]))
  tableStart = unlist(tableStart_list) # convert list to vector

  # find length of each pattern(because some patterns have multiple indexes)
  lens=sapply(tableStart_list, length)
  lineNumber = rep(pattern,lens)  # extract correct number of pattern

  # extract the indexes of interest(end) for each pattern
  tableEnd = unlist(sapply(pattern, function(i) gregexpr('/table>',file[i])))
  cbind(lineNumber,tableStart,tableEnd) # bind them to a matrix
}
tableInfo(paste0(url,'new-zealand-population'))
```

```
##      lineNumber tableStart tableEnd
## [1,]        199       2177     2503
## [2,]        217       1642     6306
## [3,]        217       7248     9345
## [4,]        219       2438    14419
```

```
tableInfo(paste0(url,'cook-islands-population'))
```

```
##      lineNumber tableStart tableEnd
## [1,]        199       2148     2475
## [2,]        215       1356     4970
## [3,]        215       5873     7515
```

Q3

```
readCountryTable <- function(countryName,tableName) {
  web = sub('.html','',paste0(url,fun1(countryName))) # gain a website
  file = readLines(web,warn = F)
  line = tableInfo(web)[tableName]   # obtain index from previous function

  col_name = cols(file[line])        # get column names
  value = rows(file[line],tableName) # get value of each column by its 'tableName'

  # convert all character to numeric and transpose matrix to a dataframe
  df=as.data.frame(apply(value,2,as.numeric))
  colnames(df)=col_name
  df
}

# column function: extract all column names
cols <- function(line) {
  info = strsplit(line,'thead')[[1]][2]
```

```r
  sub_info = unlist(strsplit(info,'<th>')[[1]][-1])
  sub_info = gsub('</.+$','',sub_info)
  gsub(' <br> |<br> |<br>',' ',sub_info)
}

# value function: extract all values
rows <- function(line,tableName) {
  if (tableName==2) {
    info=strsplit(line,'<tr> <td>')[[1]][2:19]
    sub_info=strsplit(info,'</td>')
  }

  else {
    info=strsplit(line,'<tr> <td>')[[1]][20:26]
    sub_info=strsplit(info,'</td>')
  }

  m=do.call(rbind,sub_info)
  m=gsub(',| %|</.+$','',m[,-ncol(m)])
  m=gsub('| <.+>|<.+>','',m)
  ifelse(m=='N.A.', NA, m)
}

head(NZTable2 <- readCountryTable("French Polynesia", 2), 3)
```

```
##   Year Population Yearly % Change Yearly Change Migrants (net) Median Age
## 1 2020     280908            0.58          1621          -1000       33.6
## 2 2019     279287            0.58          1608          -1000       31.9
## 3 2018     277679            0.57          1577          -1000       31.9
##   Fertility Rate Density (P/Km²) Urban Pop % Urban Population
## 1           1.95              77        64.1           180188
## 2           2.02              76        63.9           178578
## 3           2.02              76        63.7           176757
##   Country’s Share of World Pop World Population French Polynesia Global Rank
## 1                            0       7794798739                          185
## 2                            0       7713468100                          185
## 3                            0       7631091040                          185
```

```r
head(CITable3 <- readCountryTable("Cook Islands", 3), 3)
```

```
##   Year Population Yearly % Change Yearly Change Density (P/Km²) Urban Pop %
## 1 2020      17564           -0.03            -4              73        75.3
## 2 2025      17544           -0.02            -4              73        77.4
## 3 2030      17524           -0.02            -4              73        79.3
##   Urban Population Country’s Share of World Pop World Population
## 1            13223                            0       7794798739
## 2            13571                            0       8184437460
## 3            13903                            0       8548487400
##   Cook Islands Global Rank
## 1                      223
## 2                      223
## 3                      223
```

Q4

```
# get a list that contains second dataframe of 20 countries from previous function
df = lapply(oceania, readCountryTable, 2)

# sublist(1-3) the list from last step
mod_df = lapply(1:length(oceania), function(i) cbind(Country=oceania[i],df[[i]][1:3]))

# convert list to dataframe named f_df(for next question)
f_df = do.call(rbind,mod_df)
head(f_df);tail(f_df)
```

```
##     Country Year Population Yearly % Change
## 1 Australia 2020   25499884            1.18
## 2 Australia 2019   25203198            1.23
## 3 Australia 2018   24898152            1.28
## 4 Australia 2017   24584620            1.33
## 5 Australia 2016   24262712            1.38
## 6 Australia 2015   23932502            1.56

##     Country Year Population Yearly % Change
## 355 Tokelau 1980       1553           -0.24
## 356 Tokelau 1975       1572           -0.61
## 357 Tokelau 1970       1621           -3.35
## 358 Tokelau 1965       1922            0.52
## 359 Tokelau 1960       1873            3.11
## 360 Tokelau 1955       1607            0.52
```

Q5

```
# find the each 10-year
sub_df = f_df[f_df$Year %% 10 == 0,]

# use tapply to get a contingency table
final_df=(with(sub_df,tapply(`Yearly % Change`, list(Country,as.factor(Year)), mean)))

# convert to dataframe
final_df = as.data.frame(final_df)

# modification
final_df=cbind(Country=rownames(final_df),final_df)
rownames(final_df)=1:nrow(final_df);final_df
```

```
##                      Country 1960  1970  1980  1990  2000  2010  2020
## 1             American Samoa 0.37  2.94  1.67  3.68  1.69 -1.20 -0.22
## 2                  Australia 2.25  2.49  1.16  1.60  1.09  1.89  1.18
## 3               Cook Islands 2.18  2.43 -2.85  0.66 -1.25 -0.73 -0.03
## 4                       Fiji 3.26  2.33  1.96  0.47  0.90  0.91  0.73
## 5           French Polynesia 2.55  3.41  3.16  2.46  1.99  0.59  0.58
## 6                   Kiribati 2.41  1.96  1.48  2.50  1.66  2.20  1.57
## 7           Marshall Islands 1.08  3.37  3.64  4.28  0.12  0.40  0.60
## 8                      Nauru 2.76  2.27  1.53  2.26 -0.55  0.31  0.84
## 9              New Caledonia 2.87  2.92  2.19  1.78  2.37  1.41  0.97
## 10               New Zealand 2.12  1.42  0.41  0.78  0.98  1.11  0.82
## 11                      Niue 0.56  0.14 -3.06 -3.05 -2.59 -0.78  0.09
## 12  Northern Mariana Islands 4.56  3.35  2.19  9.76  4.61 -0.93  0.63
## 13          Papua New Guinea 1.58  2.29  2.61  2.49  2.43  2.39  1.95
## 14                     Samoa 2.90  2.41  0.54  0.35  0.51  0.68  0.67
## 15           Solomon Islands 2.93  3.21  3.58  2.87  2.81  2.35  2.55
## 16                   Tokelau 3.11 -3.35 -0.24 -1.21  0.48 -1.15  1.62
## 17                     Tonga 2.17  2.56  1.03  0.25  0.41  0.60  1.15
## 18                    Tuvalu 0.98  1.26  4.73  1.63  0.21  1.04  1.22
## 19                   Vanuatu 3.01  2.83  2.97  2.43  1.92  2.45  2.42
## 20 Wallis and Futuna Islands 1.24  0.29  3.80  0.26  0.76 -3.21 -1.73
```