Webscraping with Rvest

EOL

2022-11-07

Goal

I will use the rvest library to scrape data from wikipedia. More specifically, I will scrape a list of chief executive officers from this wikipedia page: https://en.wikipedia.org/wiki/List_of_chief_executive_officers (https://en.wikipedia.org/wiki/List_of_chief_executive_officers). On the wikipedia page, the list is described as the following: The following is a list of chief executive officers of notable companies. The list also includes lead executives with a position corresponding to chief executive officer (CEO), such as managing director (MD), and any concurrent positions held.

**My goal is to look at the overall gender distribution, i.e. the counts of males and females*. I will look at the gender distribution. However, the table doesn't contain gender labels so for that I'll use the R package called Gender: https://www.r-project.org/nosvn/pandoc/gender.html (https://www.r-project.org/nosvn/pandoc/gender.html)

Installing R packages

I'll install the following packages and load their libraries:

- `rvest` for scraping web data
- `tidyverse` , `stringr` , and `dplyr` all for data wrangling
- `tidyr` to create tidy data if needed
- `GenderInfer` to assign gender based on first name.

```
pacman::p_load(rvest,tidyverse,stringr,dplyr,tidyr,GenderInfer)
```

Scraping the data

```
url <- "https://en.wikipedia.org/wiki/List_of_chief_executive_officers"

# scraping
parsed_html <- read_html(url)
```

```
ceo_table <- parsed_html %>%
  html_elements("table") %>%
  html_table()

# retrieving the data
ceo_table <- ceo_table[[1]]

head(ceo_table)
```

```
## # A tibble: 6 × 6
##   Company             Executive             Title                 Since Notes Updated
##   <chr>               <chr>                 <chr>                 <chr> <chr> <chr>
## 1 Accenture           Julie Sweet           CEO[1]                2019  Succ… 2019-0…
## 2 Aditya Birla Group Kumar Mangalam Birla Chairman[2]            1995… Part… 2018-1…
## 3 Adobe Systems       Shantanu Narayen      Chairman, preside… 2007  Form… 2018-1…
## 4 Agenus              Garo H. Armen         Founder, chairman… 1994  Foun… 2018-1…
## 5 Airbus              Guillaume Faury       CEO[5]                2012  Succ… 2017-1…
## 6 Alibaba             Daniel Zhang          CEO[6]                2015  Prev… 2018-1…
```

Data cleaning

First, there are some name abbreviations that I would like to change.

```
ceo_df <- ceo_table

library(stringi)
ceo_withdots <- ceo_df[stri_detect_fixed(ceo_df$Executive, "."),]

length(ceo_withdots$Executive)
```

```
## [1] 29
```

```
ceo_df$Executive <- gsub('Garo H. Armen','Garo Armen',ceo_df$Executive)
ceo_df$Executive <- gsub('Joseph R. Swedish ','Joseph Swedish',ceo_df$Executive)
ceo_df$Executive <- gsub('Stephen A. Schwarzman ','Stephen Schwarzman',ceo_df$Executi
ve)
ceo_df$Executive <- gsub('Evan G. Greenberg','Evan GreenBerg',ceo_df$Executive)
ceo_df$Executive <- gsub('Brian L. Roberts','Brian Roberts',ceo_df$Executive)
ceo_df$Executive <- gsub('Roland Dickey Jr.','Roland Dickey',ceo_df$Executive)
ceo_df$Executive <- gsub('Edward D. Breen   ','Edward Breen',ceo_df$Executive)
ceo_df$Executive <- gsub('Lisa S. Jones','Lisa Jones',ceo_df$Executive)
ceo_df$Executive <- gsub('Frederick W. Smith','Frederick Smith',ceo_df$Executive)
ceo_df$Executive <- gsub('H. Lawrence Culp Jr.','Henry Lawrence Culp Jr',ceo_df$Execu
tive)
ceo_df$Executive <- gsub('Mary T. Barra','Mary Barra',ceo_df$Executive)
ceo_df$Executive <- gsub('David M. Solomon  ','David Solomon',ceo_df$Executive)
ceo_df$Executive <- gsub('John A. Kaneb ','John Kaneb',ceo_df$Executive)
ceo_df$Executive <- gsub('Richard B. Handler','Richard Handler',ceo_df$Executive)
ceo_df$Executive <- gsub('Andrew S. Rosen','Andrew Rosen',ceo_df$Executive)
ceo_df$Executive <- gsub('Charles G. Koch   ','Charles Koch',ceo_df$Executive)
ceo_df$Executive <- gsub('Steven A. Kandarian   ','Steven Kandarian',ceo_df$Executiv
e)
ceo_df$Executive <- gsub('Michael J. Saylor','Michael Saylor',ceo_df$Executive)
ceo_df$Executive <- gsub('James P. Gorman','James Gorman',ceo_df$Executive)
ceo_df$Executive <- gsub('David S. Taylor','David Taylor',ceo_df$Executive)
ceo_df$Executive <- gsub('David I. McKay','David McKay',ceo_df$Executive)
ceo_df$Executive <- gsub('Douglas L. Peterson','Douglas Peterson',ceo_df$Executive)
ceo_df$Executive <- gsub('Gary C. Kelly ','Gary Kelly',ceo_df$Executive)
ceo_df$Executive <- gsub('J. Clifford Hudson    ','Clifford Hudson',ceo_df$Executive)
ceo_df$Executive <- gsub('William H. Rogers Jr. ','William Rogers',ceo_df$Executive)
ceo_df$Executive <- gsub('Alan D. Schnitzer','Alan Schnitzer',ceo_df$Executive)
ceo_df$Executive <- gsub('Joseph C. Papa','Joseph Papa',ceo_df$Executive)
ceo_df$Executive <- gsub('Laura J. Alber    ','Laura Alber',ceo_df$Executive)
```

OBS: `G. V. Prasad is a male` although I couldn’t find what first name G. stands for

I now want to split intermixed names into first, middle, and last names. This step is necessary because I’ll be using the `GenderInfer` library to infer the gender of a CEO based on her/his first name.

```
library(stringr)
ceo_df$firstname <- stringr::str_extract(ceo_df$Executive, '\\w*')
ceo_df$lastname <- str_extract(ceo_df$Executive, "\\w+$")
```

```
head(ceo_df)
```

```
## # A tibble: 6 × 8
##   Company             Executive      Title Since Notes Updated firstname lastname
##   <chr>               <chr>          <chr> <chr> <chr> <chr>   <chr>     <chr>
## 1 Accenture           Julie Sweet    CEO[… 2019  Succ… 2019-0… Julie     Sweet
## 2 Aditya Birla Group Kumar Mangala… Chai… 1995… Part… 2018-1… Kumar     Birla
## 3 Adobe Systems       Shantanu Nara… Chai… 2007  Form… 2018-1… Shantanu  Narayen
## 4 Agenus              Garo Armen     Foun… 1994  Foun… 2018-1… Garo      Armen
## 5 Airbus              Guillaume Fau… CEO[… 2012  Succ… 2017-1… Guillaume Faury
## 6 Alibaba             Daniel Zhang   CEO[… 2015  Prev… 2018-1… Daniel    Zhang
```

Using GenderInfer

About GenderInfer: *GenderInfer (Giordano et al. (2021) is a package developed to investigate gender differences within a data set. This package is based on the work of Dr. A. Day et al. Chem. Sci., 2020,11, 2277-2301. This has been developed for analysing differences in publishing authorship by gender. This package could also be useful for other analyses where there might be differences between male and female percentages from a specified baseline. The gender is assigned based on the first name, using the following data set as a corpus: https://github.com/OpenGenderTracking/globalnamedata (https://github.com/OpenGenderTracking/globalnamedata)* (Giordano et al. (2021))

```
# Assigning Gender
ceo_df <- assign_gender(ceo_df,"firstname")
head(ceo_df)
```

```
##              Company            Executive                          Title
## 1          Accenture          Julie Sweet                         CEO[1]
## 2 Aditya Birla Group Kumar Mangalam Birla                    Chairman[2]
## 3      Adobe Systems     Shantanu Narayen Chairman, president and CEO[3]
## 4             Agenus           Garo Armen      Founder, chairman, CEO[4]
## 5             Airbus      Guillaume Faury                         CEO[5]
## 6            Alibaba         Daniel Zhang                         CEO[6]
##     Since                                            Notes    Updated firstname
## 1    2019                  Succeeded Pierre Nanterme, died 2019-01-31     Julie
## 2 1995[2] Part of the Birla family business house in India 2018-10-01     Kumar
## 3    2007                              Formerly with Apple 2018-10-01  Shantanu
## 4    1994   Founder of the Children of Armenia Fund (COAF) 2018-10-01      Garo
## 5    2012                          Succeeded Louis Gallois 2017-11-14 Guillaume
## 6    2015                           Previously with Taobao 2018-10-01    Daniel
##   lastname gender
## 1    Sweet      U
## 2    Birla      U
## 3  Narayen      U
## 4    Armen      U
## 5    Faury      U
## 6    Zhang      U
```

```
ceo_df %>% count(gender)
```

```
##   gender   n
## 1      U 176
```

For some reason, the above chunk needs to be run twice to work?

```
# Assigning Gender
ceo_df <- assign_gender(ceo_df,"firstname")
head(ceo_df)
```

```
##                   Company        Executive                       Title Since
## 1    Fidelity Investments  Abigail Johnson Chairman, president and CEO  2014
## 2                  Toyota      Akio Toyoda President and director[137]  2009
## 3 The Travelers Companies   Alan Schnitzer       Chairman and CEO[136]  2015
## 4                  Qantas       Alan Joyce             CEO and MD[103]  2008
## 5                  Pfizer    Albert Bourla        Chairman and CEO[99]  2019
## 6                     BHP Andrew Mackenzie                     CEO[22]  2013
##                                                                      Notes
## 1               Granddaughter of the firm's founder, Edward C. Johnson II
## 2                            Son of Shoichiro Toyoda, the former chairman
## 3 Previously over the firm's Business and International Insurance segment
## 4                           Formerly with Aer Lingus and Ansett Australia
## 5                              Succeeded Jeff Kindler and Henry McKinnell
## 6                                    Previously with BP and the Rio Tinto
##      Updated firstname  lastname gender
## 1 2017-11-14   Abigail   Johnson      F
## 2 2017-11-11      Akio    Toyoda      M
## 3 2017-11-11      Alan Schnitzer      M
## 4 2017-11-12      Alan     Joyce      M
## 5               Albert    Bourla      M
## 6 2017-11-15    Andrew Mackenzie      M
```

```
ceo_df %>% count(gender)
```

```
##   gender   n
## 1      F  20
## 2      M 140
## 3      U  16
```

```
#which(ceo_df$gender == "U")
ceo_unknowngender<- ceo_df[ceo_df$gender=="U",]
ceo_unknowngender[,c("firstname","lastname","gender")]
```

```
##     firstname    lastname gender
## 16      Börje      Ekholm      U
## 27          C Vijayakumar      U
## 51     Dikesh    Malhotra      U
## 64          G      Prasad      U
## 69   Gunupati       Reddy      U
## 77          J      Hudson      U
## 107        Li   Dongsheng      U
## 125        Oh        Kwon      U
## 129       Pat   Gelsinger      U
## 131     Pekka    Lundmark      U
## 133     Phiwa    Nkambule      U
## 151     Safra        Catz      U
## 163    Sundar      Pichai      U
## 168   Tidjane       Thiam      U
## 172     Toxey        Haas      U
## 173    Vasant  Narasimhan      U
```

Changing gender of `G. V. Prasad` to male although I couldn’t find what first name G. stands for

```
ceo_unknowngender$gender <- ifelse(ceo_unknowngender$firstname=="G","M",
                                   ifelse(ceo_unknowngender$firstname =="Börje","M",
                                          ifelse(ceo_unknowngender$firstname =="C",
"M",
                                                 ifelse(ceo_unknowngender$firstname==
"Dikesh","M",
                                                        ifelse(ceo_unknowngender$firs
tname =="Gunupati","M",
                                                               ifelse(ceo_unknowngend
er$firstname =="J","F",
                                                                      ifelse(ceo_unkn
owngender$firstname =="Li","M",
                                                                             ifelse(c
eo_unknowngender$firstname =="Oh","M",
                                                                                    i
felse(ceo_unknowngender$firstname =="Pekka","M",

ifelse(ceo_unknowngender$firstname =="Phiwa","M",

ifelse(ceo_unknowngender$firstname =="Safra" ,"F",

ifelse(ceo_unknowngender$firstname =="Sundar","M",

ifelse(ceo_unknowngender$firstname =="Tidjane","M",

ifelse(ceo_unknowngender$firstname =="Toxey","M",

ifelse(ceo_unknowngender$firstname=="Pat","M",

ifelse(ceo_unknowngender$firstname =="Vasant","M","U")))))))))))))))
```

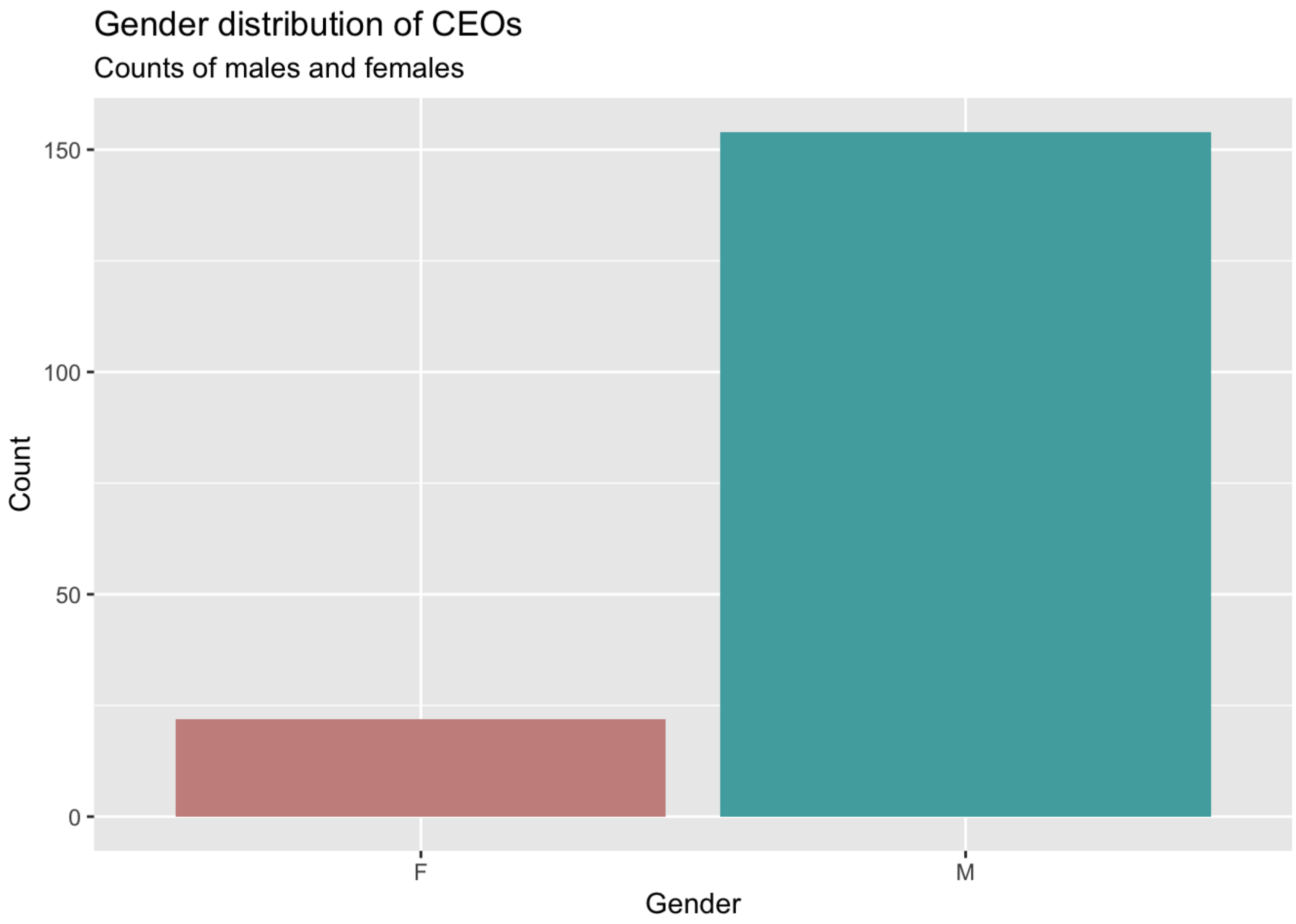
Using `match` in `Executive` column to select the elements of `gender`.

```
ceo_df$gender[match(ceo_unknowngender$Executive,ceo_df$Executive)] <- ceo_unknowngend
er$gender
```

Data visualisation

Count the variable gender

```
ggplot(ceo_df, aes(x=as.factor(gender), fill=as.factor(gender) )) +
  geom_bar( ) +
  scale_fill_hue(c=40) +
  theme(legend.position="none")+
  ggtitle("Count of males and females")+
  labs(title = "Gender distribution of CEOs",subtitle = "Counts of males and females"
,
  x = "Gender",
  y = "Count")
```



```
ceo_df %>% count(gender)
```

```
##   gender   n
## 1      F  22
## 2      M 154
```

```
ceo_df %>% count(gender) %>%
  mutate(percent=n/sum(n)) %>%
  select(-n) %>%
  spread(gender,percent)
```

```
##       F     M
## 1 0.125 0.875
```

As seen in the blot above and the summary, there are 154 men on the list and 22 women on the list, corresponding to 87.5% of the CEO's on the list being males. Women are severely underrepresented by making up only 12.5% on the list, reflecting an unequal gender distribution at top positions of well knowned US companies.

References

Giordano et al. (2021). gender: Predict Gender from Names Using Historical Data. https://github.com/ropensci/gender (https://github.com/ropensci/gender)