Women in Parliament - Tidy Data

Carolina Silva

Fri 06 Dec 2019 (11:24:13)

Contents

Objectives **1**

Understanding the Data **1**

The World Bank Data . . . 1

Source Data . . . 2

Data limitations . . . 2

Data definitions & assumptions . . . 2

"Women" . . . 2

"Country (Region)" . . . 2

Pro tip . . . 2

About the data file **2**

Pro tip . . . 3

Importing the data **3**

Objectives

Explore the geographical and time trends for the percentage of women in national parliaments.

Understanding the Data

The World Bank Data

The raw data for *"Proportion of seats held by women in national parliaments"* includes the percentage of women in parliament (*"single or lower parliamentary chambers only"*) by country (region) and year. It can be downloaded from:

- https://data.worldbank.org/indicator/SG.GEN.PARL.ZS

As part of its "open data" mission the World Bank offers *"free and open access to global development data"* kindly licensed under the "Creative Commons Attribution 4.0 (CC-BY 4.0)".

Source Data

The data originates from the "Inter-Parliamentary Union" (IPU)[^ipuwebsite] which provides an "Archive of statistical data on the percentage of women in national parliaments" going back to 1997 on a monthly basis:

- http://archive.ipu.org/wmn-e/classif-arc.htm

The World Bank data is for "single or lower parliamentary chambers only", while the IPU also presents data for "Upper Houses or Senates". Moreover, the IPU provides the actual numbers used to calculate the percentages (which the World Bank does not).

Data limitations

Take caution when interpreting these data, as parliamentary systems vary from country to country, and in some cases over time. Some of the issues to consider include:

- Who has, and who does not have, the right to become a Member of Parliament (MP)?
- How does someone become an MP? Through democratic elections? How is "democratic election" defined?
- What is the real power of MPs and their parliament? Can MPs make a difference?

Data definitions & assumptions

"Women"

The definition for "women" is not given, so we will assume that it refers to a binary classification for gender (sex).

"Country (Region)"

The definition of countries and regions can change over time. (e.g. formation of new countries after conflicts, new member states joining a pre-existing collective). How are these changes reflected in the data? How do they affect the interpretation?

Pro tip

Understand the limitations of your data before anybody else points them out to you.

About the data file

The data is stored in a file called: `API_SG.GEN.PARL.ZS_DS2_en_csv_v2_511371.csv`

To simplify things we have copied it to `WB-WiP.csv` (which also allows us to maintain the original file in case something goes wrong).

Pro tip

Always keep a backup copy of the data. Alternatively, set the data file(s) to "read-only" to protect it from being overwritten or modified.

Importing the data

Based on our findings above, we can "skip" the first four lines and treat the fifth line as column (variable) names. Also note the use of the `check.names` argument to ensure that the column names are compliant in R.

```
library(data.table)
library(here)
wip <- fread(here("data", "WB-WiP.csv"),
             skip = 4, header = TRUE,
             check.names = TRUE)
```

```
wip[, .N, by=.(V65)] #verifies that all values are NA
```

```
##    V65   N
## 1:  NA 264
```

```
wip[, c("Indicator.Name", "Indicator.Code",
        "V65"):=NULL]

setnames(wip, c("Country.Name", "Country.Code"),
              c("Country","Code"))

head(names(wip))
```

```
## [1] "Country" "Code"    "X1960"   "X1961"   "X1962"   "X1963"
```

```
tail(names(wip))
```

```
## [1] "X2014" "X2015" "X2016" "X2017" "X2018" "X2019"
```

```
WP <- melt(wip,
           id.vars = c("Country", "Code"),
           measure = patterns("^X"),
           variable.name = "YearC",
           value.name = c("pctWiP"),
           na.rm = TRUE)
```

```
## Warning in melt.data.table(wip, id.vars = c("Country", "Code"), measure =
## patterns("^X"), : 'measure.vars' [X1960, X1961, X1962, X1963, ...] are not all
## of the same type. By order of hierarchy, the molten data value column will be of
## type 'double'. All measure variables not of type 'double' will be coerced too.
## Check DETAILS in ?melt.data.table for more on coercion.
```

```
WP
```

```
##                    Country Code YearC    pctWiP
##    1:           Afghanistan  AFG X1990  3.700000
##    2:                Angola  AGO X1990 14.500000
##    3:               Albania  ALB X1990 28.800000
##    4:            Arab World  ARB X1990  3.891439
##    5: United Arab Emirates  ARE X1990  0.000000
##   ---
## 5105:                 Samoa  WSM X2018 10.000000
## 5106:           Yemen, Rep.  YEM X2018  0.000000
## 5107:          South Africa  ZAF X2018 42.300000
## 5108:                Zambia  ZMB X2018 18.000000
## 5109:              Zimbabwe  ZWE X2018 31.500000
```

```
WP[, ':='(Year=as.numeric(gsub("[^[:digit:].]",
                               "", YearC)),
          Ratio = (100-pctWiP)/pctWiP)][
            , YearC:=NULL]
setcolorder(WP, c("Country","Code","Year",
                  "pctWiP","Ratio"))

WP
```

```
##                    Country Code Year    pctWiP     Ratio
##    1:           Afghanistan  AFG 1990  3.700000 26.027027
##    2:                Angola  AGO 1990 14.500000  5.896552
##    3:               Albania  ALB 1990 28.800000  2.472222
##    4:            Arab World  ARB 1990  3.891439 24.697433
##    5: United Arab Emirates  ARE 1990  0.000000       Inf
##   ---
## 5105:                 Samoa  WSM 2018 10.000000  9.000000
## 5106:           Yemen, Rep.  YEM 2018  0.000000       Inf
## 5107:          South Africa  ZAF 2018 42.300000  1.364066
## 5108:                Zambia  ZMB 2018 18.000000  4.555556
## 5109:              Zimbabwe  ZWE 2018 31.500000  2.174603
```

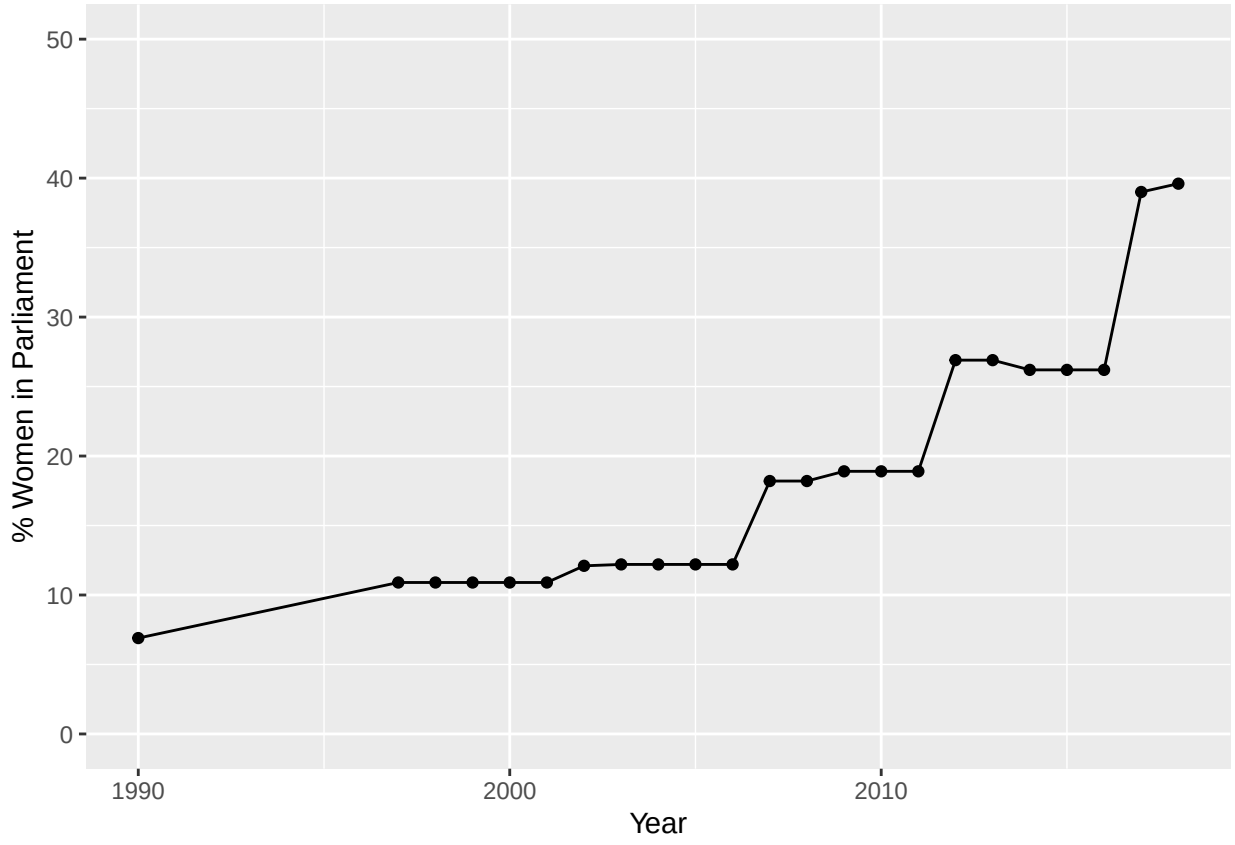
```
WP[Country %in% "France"]
```

```
##     Country Code Year pctWiP     Ratio
##  1:  France  FRA 1990    6.9 13.492754
##  2:  France  FRA 1997   10.9  8.174312
##  3:  France  FRA 1998   10.9  8.174312
##  4:  France  FRA 1999   10.9  8.174312
##  5:  France  FRA 2000   10.9  8.174312
##  6:  France  FRA 2001   10.9  8.174312
##  7:  France  FRA 2002   12.1  7.264463
##  8:  France  FRA 2003   12.2  7.196721
##  9:  France  FRA 2004   12.2  7.196721
## 10:  France  FRA 2005   12.2  7.196721
## 11:  France  FRA 2006   12.2  7.196721
## 12:  France  FRA 2007   18.2  4.494505
```

```
## 13:  France  FRA 2008   18.2  4.494505
## 14:  France  FRA 2009   18.9  4.291005
## 15:  France  FRA 2010   18.9  4.291005
## 16:  France  FRA 2011   18.9  4.291005
## 17:  France  FRA 2012   26.9  2.717472
## 18:  France  FRA 2013   26.9  2.717472
## 19:  France  FRA 2014   26.2  2.816794
## 20:  France  FRA 2015   26.2  2.816794
## 21:  France  FRA 2016   26.2  2.816794
## 22:  France  FRA 2017   39.0  1.564103
## 23:  France  FRA 2018   39.6  1.525253
##     Country Code Year pctWiP     Ratio
```

```
library(ggplot2)
library(magrittr)

WP[Country %in% "France"] %>%
ggplot(aes(Year, pctWiP)) +
  geom_line() + geom_point() +
  scale_y_continuous(limits = c(0, 50)) +
  ylab("% Women in Parliament")
```




```
WP[Country %in% c("France", "Portugal", "United Kingdom","Norway","Denmark","Poland")] %>%

  ggplot(aes(Year, pctWiP, colour=Country)) +
```

```
  geom_line() +
  geom_point() +
  scale_x_continuous(breaks = seq(1990, 2020, 5)) +
  scale_y_continuous(limits = c(0,50),
                     breaks=seq(0,50,by=10)) +
  ggtitle("Women in Parliament: EU Countries") +
  ylab("% Women in Parliament")
```