Women in Parliament – Tidyverse Edition

Saghir Bashir

This version was compiled on April 7, 2021

We will use the World Bank's indicator data for "Women in Parliament" as a case study when working with the tidyverse suite of R packages. We will guide you through the geographical and time trends for the percentage of women in national parliaments. We will start by learning about and understanding the raw data, which we will then process ("wrangle") in preparation for some exploratory analysis.

Women in Parliament | World Bank Indicator | Tidyverse | dplyr | tidyr | ggplot2

1. Preface

We present a real-life case study for some of the `tidyverse`[1] package using the World Bank's "Women in Parliament" indicator data. To get the most out of this case-study guide, repeat the examples and do the exercises whilst reading it.

Guide materials. You can download materials for this guide from this link:

- https://ilustat.com/shared/WiP-tidyverse.zip

Unzip the file, which contains the data, this guide and an R script exercise file. We advise you to work with "`WiP-Exercise.R`" file to follow the examples and do the exercises. If you are using RStudio, you can double click on "`WiP-tv.Rproj`" to get started.

2. Objectives

Explore the geographical and time trends for the percentage of women[2] *in national parliaments.*

3. 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:[3]

- 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)".[4]

Source Data. The data originates from the "Inter-Parliamentary Union" (IPU)[5] 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.

4. About the data file

The data is stored in a file called:

- `API_SG.GEN.PARL.ZS_DS2_en_csv_v2_2163427.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.

Exercise. It is important to look at and understand the contents of the file before you start using it. Using a text editor or a spreadsheet software, open the `WB-WiP.csv` file (in the `data` directory). What do you observe in the contents of this file?

Content and Structure. The first four lines of WB-WiP.csv can be ignored, since they contain two lines of meta-information and two blank lines, as follows:

```
1  "Data Source","World Development Indicators",
2
3  "Last Updated Date","2021-03-19",
4
```

The fifth line contains the column (variable) names and the body of data starts in the sixth line. It is important to note that there was no collection of data for a majority of the years, which means that it is "missing".

5. Loading `tidyverse` packages

We will load the `tidyverse` packages we plan to use individually (messages have been suppressed).[6]

[1] For more information on the `tidyverse` see https://www.tidyverse.org/.

[2] The objective could be termed neutrally as "gender trends" but we will keep it per the World Bank data.

[3] The `wbstats` R package (https://cran.r-project.org/web/packages/wbstats/) gives access to a "tidier" version of the World Bank indicator data.

[4] https://datacatalog.worldbank.org/public-licenses#cc-by.

[5] Inter-Parliamentary Union: https://www.ipu.org/.

[6] We could have used `library(tidyverse)` but we prefer to load packages individually and only those that we will use.

```
library(here)
library(readr)
library(dplyr)
library(tidyr)
library(ggplot2)
library(gghighlight)
```

6. Importing the data

Based on our findings above, we can "skip" the first four lines and treat the fifth line as column (variable) names.

```
wip <- read_csv(here("data", "WB-WiP.csv"),
                skip = 4)
```

```
#  Warning: Missing column names filled in:
#  'X66' [66]
```

Messages. We have suppressed some of the messages but left the one about variable `X66` (we will come back to it below).

Exercise. Check what you have read by typing "`wip`" in the console window. What do you observe? Type "`class(wip)`" and "`glimpse(wip)`" to confirm that "`wip`" is of class "`tbl_df`".

"Fix" column names. Some of the column names contain spaces while others are numeric:

```
head(names(wip))
#  [1] "Country Name"   "Country Code"
#  [3] "Indicator Name" "Indicator Code"
#  [5] "1960"           "1961"
tail(names(wip))
#  [1] "2016" "2017" "2018" "2019" "2020" "X66"
```

By using the `make.names()` function we don't need to use back ticks (`` ` ``) around the column names (e.g. `` `col name` ``).

```
names(wip) <- make.names(names(wip))
head(names(wip))
#  [1] "Country.Name"   "Country.Code"
#  [3] "Indicator.Name" "Indicator.Code"
#  [5] "X1960"          "X1961"
tail(names(wip))
#  [1] "X2016" "X2017" "X2018" "X2019" "X2020"
#  [6] "X66"
```

7. Data Wrangling Aims

We can simplify the production of summaries and plots by restructuring the current `wip` dataset (which has 66 columns) to the following format:

```
Country       Year   pctWiP
Country AAA   1997     ##.#
Country AAA   1998     ##.#
Country AAA   1999     ##.#
...
```

pctWiP refers to the percentage of women in parliament.

Key information retained. These three columns will contain the same information as the `wip` dataset but in a more usable format. We will also add a variable for the ratio of male to female MPs.

Superfluous columns. We will start by removing columns `X66`, `Indicator.Name` and `Indicator.Code`. There are years without any data but they will be removed automatically later (when restructuring from "wide" to "long" format).

Column `X66` is created automatically due to an extra comma at the end of the column names (fifth) line of `WB-WiP.csv`:

```
... ,"2017","2018","2019","2020",
```

Check. Before removing it check that all values are `NA`.

```
wip %>% pull(X66) %>% is.na(.) %>% all(.)
#  [1] TRUE
```

Column `Indicator.Name` has the unique value *"Proportion of seats held by women in national parliaments (%)"* and in `Indicator.Code` it is *"SG.GEN.PARL.ZS"*. As there is only one indicator in this dataset we will remove these two columns.

Exercise. Confirm that both `Indicator.Name` and `Indicator.Code` have the same values for all observations. Hint: Use either `count()` or `distinct()` functions.

Removing columns. The indicator and `X66` columns can be removed. We will also rename "`Country.Name`" as "`Country`" and "`Country.Code`" as "`Code`".

```
wip2 <- wip %>%
  select(-Indicator.Name, -Indicator.Code,
         -X66) %>%
  rename(Country=Country.Name, Code=Country.Code)
head(names(wip2))
#  [1] "Country" "Code"    "X1960"   "X1961"
#  [5] "X1962"   "X1963"
tail(names(wip2))
#  [1] "X2015" "X2016" "X2017" "X2018" "X2019"
#  [6] "X2020"
```

Reshape to long format. We want to transform the data so that for each country the year (column) data becomes a row. At the same time we will remove the missing data (with the `na.rm` option). We will also create a numeric `Year` variable and a `Ratio` of men to women in parliament.

```
WP <- wip2 %>%
  pivot_longer(starts_with("X"),
               names_to = "YearC",
               values_to = "pctWiP",
               values_drop_na = TRUE) %>%
  mutate(Year = parse_number(YearC),
         Ratio = (100-pctWiP)/pctWiP) %>%
  select(Country, Code, Year, pctWiP, Ratio) %>%
  arrange(Country, Year)
# Look at the contents of WP
glimpse(WP)
#  Rows: 5,391
#  Columns: 5
#  $ Country <chr> "Afghanistan", "Afghanistan~
#  $ Code    <chr> "AFG", "AFG", "AFG", "AFG",~
#  $ Year    <dbl> 2005, 2006, 2007, 2008, 200~
#  $ pctWiP  <dbl> 27.3, 27.3, 27.7, 27.7, 27.~
#  $ Ratio   <dbl> 2.66, 2.66, 2.61, 2.61, 2.6~
```

8. Questions

The objective is to look at the geographical and time trends in the data. We will answer the following questions.

- What are the time trends for Portugal?
- How does Portugal compare to other countries?
- Which countries have the highest percentage of women in parliament by year?
- How do continents compare?
- What are the global trends over time?

Exercise - Without Programming.

- Which country do you think has the highest percentage of women in parliament?
- In each continent (i.e. Africa, Americas, Asia, Europe and Oceania), which country has the highest percentage of women in parliament?
- What is the world percentage of women in parliament in 2020?

9. Exploratory Analysis

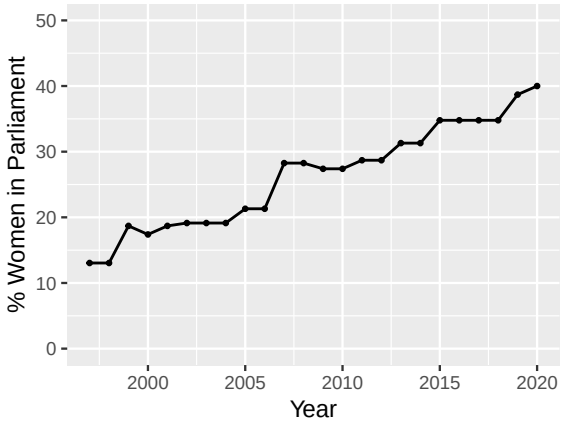
Select a country. This guide explores how Portugal performs over time and compared to other countries. Before continuing, select another country for yourself to repeat the examples and do the exercises.

Time trends for Portugal. First look at the raw data.

```
# Reset tibble print option to see more rows
options(tibble.print_max = 25)
WP %>% filter(Country=="Portugal")
#  # A tibble: 24 x 5
#     Country  Code   Year pctWiP Ratio
#     <chr>    <chr> <dbl>  <dbl> <dbl>
#   1 Portugal PRT    1997   13.0  6.67
#   2 Portugal PRT    1998   13.0  6.67
#   3 Portugal PRT    1999   18.7  4.35
#   4 Portugal PRT    2000   17.4  4.75
#   5 Portugal PRT    2001   18.7  4.35
#   6 Portugal PRT    2002   19.1  4.23
#   7 Portugal PRT    2003   19.1  4.23
#   8 Portugal PRT    2004   19.1  4.23
#   9 Portugal PRT    2005   21.3  3.69
#  10 Portugal PRT    2006   21.3  3.69
#  11 Portugal PRT    2007   28.3  2.54
#  12 Portugal PRT    2008   28.3  2.54
#  13 Portugal PRT    2009   27.4  2.65
#  14 Portugal PRT    2010   27.4  2.65
#  15 Portugal PRT    2011   28.7  2.48
#  16 Portugal PRT    2012   28.7  2.48
#  17 Portugal PRT    2013   31.3  2.19
#  18 Portugal PRT    2014   31.3  2.19
#  19 Portugal PRT    2015   34.8  1.87
#  20 Portugal PRT    2016   34.8  1.87
#  21 Portugal PRT    2017   34.8  1.87
#  22 Portugal PRT    2018   34.8  1.87
#  23 Portugal PRT    2019   38.7  1.58
#  24 Portugal PRT    2020   40    1.5
```

Visualisation. It is easier to find trends within a plot.

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

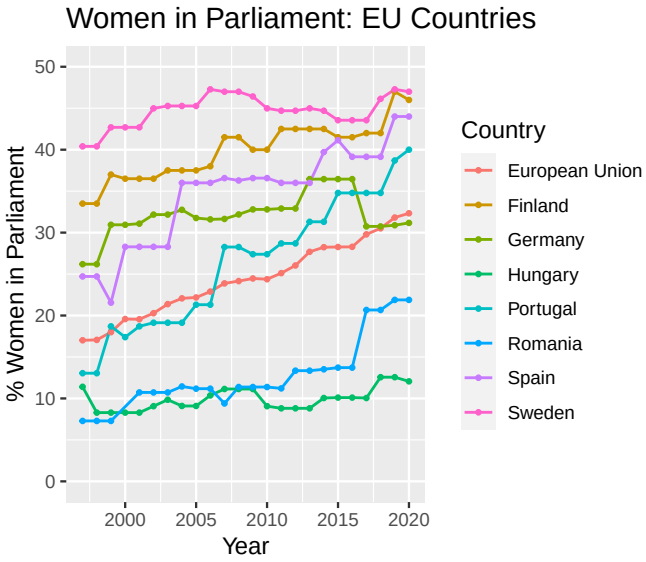


Interpretation. In 1997 Portugal had 13% women in parliament (i.e. 6.67 men for each woman), which increased to 40% (i.e. 1.5 men for each woman) in 2020. This still falls short of 50% (i.e. point of gender parity in parliament).

Exercise. For your chosen country look at the time trend data and the plot. What is your interpretation? How does it compare to Portugal?

Portugal versus European Union (EU) countries. We selected six EU countries (due to space limitations) for comparison. It would be better to compare all EU and/or all European countries.

```
WP %>%
  filter(Country %in% c("Portugal", "Sweden",
      "Spain", "Hungary", "Romania", "Finland",
      "Germany", "European Union")) %>%
  ggplot(aes(Year, pctWiP, colour=Country)) +
  geom_line() +
  geom_point() +
  scale_x_continuous(breaks=seq(1995, 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")
```



Interpretation. Since 2007 Portugal has had more women in parliament than the European Union average. The key point to note is that none of these countries reaches equality between males and females in parliament, although Sweden and Finland come closest.

A couple of points to note.

"Germany". In October 1997, the process of "German reunification" lead to the creation of Germany, which united the former "German Democratic Republic" (East Germany) and the "Federal Republic of Germany" (West Germany). Therefore, since reunification, the data is presented for the reunified "Germany" only. Careful thought should be given to handling, analysing and interpreting any pre-reunification data (if available).

"European Union". The "European Union" has changed over time (unlike the "continent of Europe"). It started in the 1950s as a block of six European countries (known as the "European Community") and has expanded over the years to 28 countries (with the United Kingdom having departed now). This raises the question of how the European Union average is calculated. For a given year, is it calculated based on the actual member states in that year or on all of the current member states?

Exercises. Compare the country of your choice to four or five other countries by plotting a line graph similar to the one above.

Countries with the highest percentage of women in parliament. A quick answer can be obtained by looking at the highest percentages.

```
WP %>%
  arrange(-pctWiP) %>%
  head(10)
#  # A tibble: 10 x 5
#     Country Code   Year pctWiP Ratio
#     <chr>   <chr> <dbl>  <dbl> <dbl>
#   1 Rwanda  RWA    2013   63.8 0.569
#   2 Rwanda  RWA    2014   63.8 0.569
#   3 Rwanda  RWA    2015   63.8 0.569
#   4 Rwanda  RWA    2016   63.8 0.569
#   5 Rwanda  RWA    2017   61.2 0.633
#   6 Rwanda  RWA    2018   61.2 0.633
#   7 Rwanda  RWA    2019   61.2 0.633
#   8 Rwanda  RWA    2020   61.2 0.633
#   9 Rwanda  RWA    2008   56.2 0.778
#  10 Rwanda  RWA    2009   56.2 0.778
```

Data speaks. Are you surprised? Data can be very enlightening.

Highest percentage by year. Which countries have the highest percentage of women in parliament by year?

```
WP %>%
  group_by(Year) %>%
  arrange(Year, -pctWiP) %>%
  filter(row_number()==1)
#  # A tibble: 24 x 5
#  # Groups:   Year [24]
#     Country Code   Year pctWiP Ratio
#     <chr>   <chr> <dbl>  <dbl> <dbl>
#   1 Sweden  SWE    1997   40.4 1.48
#   2 Sweden  SWE    1998   40.4 1.48
#   3 Sweden  SWE    1999   42.7 1.34
#   4 Sweden  SWE    2000   42.7 1.34
#   5 Sweden  SWE    2001   42.7 1.34
#   6 Sweden  SWE    2002   45.0 1.22
#   7 Rwanda  RWA    2003   48.8 1.05
#   8 Rwanda  RWA    2004   48.8 1.05
#   9 Rwanda  RWA    2005   48.8 1.05
#  10 Rwanda  RWA    2006   48.8 1.05
#  11 Rwanda  RWA    2007   48.8 1.05
#  12 Rwanda  RWA    2008   56.2 0.778
#  13 Rwanda  RWA    2009   56.2 0.778
#  14 Rwanda  RWA    2010   56.2 0.778
#  15 Rwanda  RWA    2011   56.2 0.778
#  16 Rwanda  RWA    2012   56.2 0.778
#  17 Rwanda  RWA    2013   63.8 0.569
#  18 Rwanda  RWA    2014   63.8 0.569
#  19 Rwanda  RWA    2015   63.8 0.569
#  20 Rwanda  RWA    2016   63.8 0.569
#  21 Rwanda  RWA    2017   61.2 0.633
#  22 Rwanda  RWA    2018   61.2 0.633
#  23 Rwanda  RWA    2019   61.2 0.633
#  24 Rwanda  RWA    2020   61.2 0.633
```

Merging continent. The variable `Country` in the `WP` dataset is a mix of countries and regions (e.g. "European Union", "South Asia" and "World"). To present the highest percentages grouped by continent we need to add it. Luckily, given the large number of R packages available, we can merge the "`continent`" from the "`codelist`" dataset in the "`countrycode`" package.

```
# Ensure that 'countrycode' package is installed.
# install.packages("countrycode")
library(countrycode)
cl <- codelist %>%
  select(continent, wb) %>%
  rename(Code = wb, Continent = continent)
cWP <- WP %>%
  left_join(cl, by = "Code")
```

Highest percentages by year and continent. Which countries have the highest percentages in 1997 and 2020?

```
cWP %>%
  filter(Year %in% c(1997, 2020) &
           !is.na(Continent)) %>%
  group_by(Continent, Year) %>%
  arrange(Continent, Year, -pctWiP) %>%
  filter(row_number()==1) %>%
  select(Continent, Country, Year, pctWiP, Ratio)
#  # A tibble: 10 x 5
#  # Groups:   Continent, Year [10]
#     Continent Country        Year pctWiP Ratio
#     <chr>     <chr>         <dbl>  <dbl> <dbl>
#   1 Africa    Seychelles     1997   27.3 2.67
#   2 Africa    Rwanda         2020   61.2 0.633
#   3 Americas  Argentina      1997   27.6 2.62
#   4 Americas  Cuba           2020   53.2 0.879
#   5 Asia      Vietnam        1997   26.2 2.81
#   6 Asia      United Arab ~  2020   50   1
#   7 Europe    Sweden         1997   40.4 1.48
#   8 Europe    Sweden         2020   47.0 1.13
#   9 Oceania   New Zealand    1997   29.2 2.43
#  10 Oceania   New Zealand    2020   40.8 1.45
```

Decline in percentage. Which countries have had a decline in percentage since their first measurement (not always 1997)?

```
dWP <- cWP %>%
  group_by(Country) %>%
  arrange(Country, Year) %>%
  filter(row_number()==1 | row_number()==n()) %>%
  mutate(pctDiff = pctWiP -
           lag(pctWiP, order_by=Country)) %>%
  filter(pctDiff<0  & !is.na(Continent)) %>%
```

```
  arrange(pctDiff)
dWP %>% select(Country, pctDiff)
#  # A tibble: 11 x 2
#  # Groups:   Country [11]
#     Country                    pctDiff
#     <chr>                        <dbl>
#   1 Seychelles                 -6.06
#   2 Korea, Dem. Peoples Rep. -2.47
#   3 Bahamas, The               -2.18
#   4 Tuvalu                     -2.08
#   5 Papua New Guinea           -1.83
#   6 Maldives                   -1.65
#   7 Haiti                      -1.07
#   8 Congo, Rep.                -0.742
#   9 Yemen, Rep.                -0.332
#  10 Afghanistan                -0.293
#  11 Oman                       -0.0841
```

Ranked status. Another way to look at the data is to look at the ranking of countries, which could be done at a global level or by continent. Nonetheless, the results should be interpreted with caution and an understanding of the actual percentages. For example, if most countries were around the 50% mark, rankings could be misleading and subject to random fluctuations.

Global ranks by year. We will rank the countries by year based on the percentage of women in parliaments. The countries with the highest percentage will be ranked first and the lowest last. A total for the number of countries with data is included as it varies by year.

```
cWPrankG <- cWP %>%
  filter(!is.na(Continent)) %>%
  group_by(Year) %>%
  mutate(RankG = rank(-pctWiP),
         TotalG = n())
```

Global ranking – Portugal.

```
cWPrankG %>%
  filter(Country=="Portugal") %>%
  select(Country, Year, pctWiP, Ratio, RankG,
         TotalG) %>%
  arrange(Year)
#  # A tibble: 24 x 6
#  # Groups:   Year [24]
#     Country   Year pctWiP Ratio RankG TotalG
#     <chr>    <dbl>  <dbl> <dbl> <dbl>  <int>
#   1 Portugal  1997   13.0  6.67  41.5    160
#   2 Portugal  1998   13.0  6.67  49.5    162
#   3 Portugal  1999   18.7  4.35  24      153
#   4 Portugal  2000   17.4  4.75  34      157
#   5 Portugal  2001   18.7  4.35  33      167
#   6 Portugal  2002   19.1  4.23  42      161
#   7 Portugal  2003   19.1  4.23  46      174
#   8 Portugal  2004   19.1  4.23  54      181
#   9 Portugal  2005   21.3  3.69  46      185
#  10 Portugal  2006   21.3  3.69  50      189
#  11 Portugal  2007   28.3  2.54  28      188
#  12 Portugal  2008   28.3  2.54  28      187
#  13 Portugal  2009   27.4  2.65  33      187
#  14 Portugal  2010   27.4  2.65  34      187
#  15 Portugal  2011   28.7  2.48  31      188
#  16 Portugal  2012   28.7  2.48  35      189
#  17 Portugal  2013   31.3  2.19  36      184
#  18 Portugal  2014   31.3  2.19  36      187
#  19 Portugal  2015   34.8  1.87  29      189
#  20 Portugal  2016   34.8  1.87  27      192
#  21 Portugal  2017   34.8  1.87  28      190
#  22 Portugal  2018   34.8  1.87  29      191
#  23 Portugal  2019   38.7  1.58  28      192
#  24 Portugal  2020   40    1.5   22.5    190
```

Interpretation. Portugal has generally been ranked in the first quartile (25%) of countries in the world, with the fluctuations of its ranking most likely due to random variation.

Exercise. For your chosen country, interpret its ranking over the years. How does it compare to Portugal?

Continent ranks by year. We will rank the countries by year within a continent based on the percentage of women in parliaments. The countries with the highest percentage will be ranked first and the lowest last. A total for the number of countries with data, within each continent, is included as it varies by year.

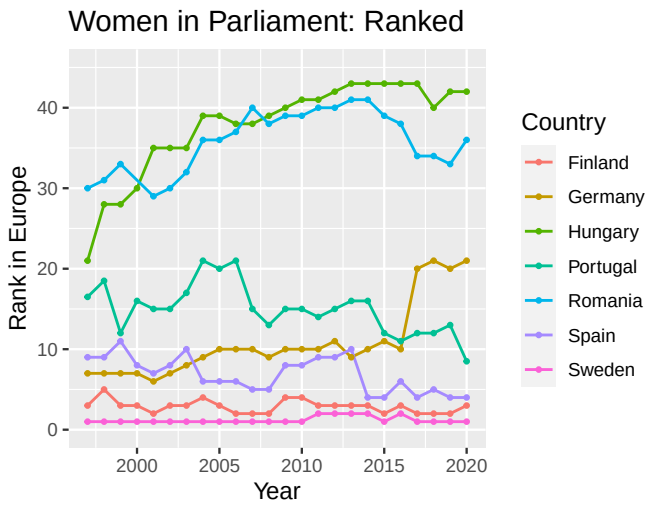
```
cWPx <- cWPrankG %>%
  filter(!is.na(Continent)) %>%
  group_by(Continent, Year) %>%
  mutate(RankC = rank(-pctWiP),
         TotalC = n())
```

Portugal's ranking in Europe.

```
cWPx %>%
  ungroup() %>%
  filter(Country=="Portugal") %>%
  select(Country, Year, pctWiP, Ratio, RankC,
         TotalC) %>%
  arrange(Year)
#  # A tibble: 24 x 6
#     Country   Year pctWiP Ratio RankC TotalC
#     <chr>    <dbl>  <dbl> <dbl> <dbl>  <int>
#   1 Portugal  1997   13.0  6.67  16.5     38
#   2 Portugal  1998   13.0  6.67  18.5     36
#   3 Portugal  1999   18.7  4.35  12       37
#   4 Portugal  2000   17.4  4.75  16       37
#   5 Portugal  2001   18.7  4.35  15       40
#   6 Portugal  2002   19.1  4.23  15       38
#   7 Portugal  2003   19.1  4.23  17       40
#   8 Portugal  2004   19.1  4.23  21       41
#   9 Portugal  2005   21.3  3.69  20       41
#  10 Portugal  2006   21.3  3.69  21       43
#  11 Portugal  2007   28.3  2.54  15       43
#  12 Portugal  2008   28.3  2.54  13       43
#  13 Portugal  2009   27.4  2.65  15       43
#  14 Portugal  2010   27.4  2.65  15       43
#  15 Portugal  2011   28.7  2.48  14       43
#  16 Portugal  2012   28.7  2.48  15       43
#  17 Portugal  2013   31.3  2.19  16       43
#  18 Portugal  2014   31.3  2.19  16       43
#  19 Portugal  2015   34.8  1.87  12       43
#  20 Portugal  2016   34.8  1.87  11       43
#  21 Portugal  2017   34.8  1.87  12       43
#  22 Portugal  2018   34.8  1.87  12       43
#  23 Portugal  2019   38.7  1.58  13       43
#  24 Portugal  2020   40    1.5    8.5     43
```

Plot of Portugal's ranking in Europe. Below we reproduce the percentage plot to show how Portugal ranks in relation to six other European countries. Note that the highest percentage is ranked first and the lowest last.

```
cWPx %>%
  filter(Country %in% c("Portugal", "Sweden",
    "Spain", "Hungary", "Romania", "Finland",
    "Germany")) %>%
  ggplot(aes(Year, RankC, colour=Country)) +
  geom_line() +
  geom_point() +
  scale_x_continuous(breaks=seq(1995, 2020, 5)) +
  scale_y_continuous(limits=c(0, 45),
                     breaks=seq(0, 45, by=10)) +
  ggtitle("Women in Parliament: Ranked") +
  ylab("Rank in Europe")
```



Interpretation. A total of 38 European countries had data in 1997, 38 in 1997 and 43 in 2020. Within Europe, Portugal was typically ranked in the second quartile (25-50%) with the fluctuations of its ranking most likely due to random variation.

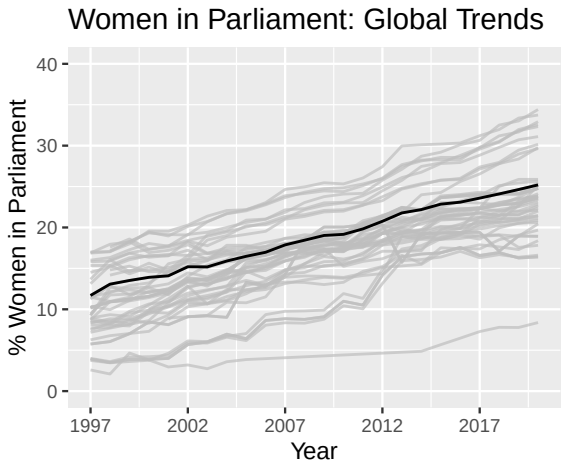
Exercise. How does your chosen country rank within its continent?

Highest rank by year and continent. Which countries have the highest rank in 1997 and 2020? The answer will coincide with the highest percentages (see above).

```
cWPx %>%
  filter(Year %in% c(1997, 2020) & RankC==1) %>%
  arrange(Continent, Year) %>%
  select(Continent, Year, Country, pctWiP, Ratio)
#  # A tibble: 10 x 5
#  # Groups:   Continent, Year [10]
#     Continent  Year Country       pctWiP Ratio
#     <chr>     <dbl> <chr>          <dbl> <dbl>
#   1 Africa     1997 Seychelles      27.3 2.67
#   2 Africa     2020 Rwanda          61.2 0.633
#   3 Americas   1997 Argentina       27.6 2.62
#   4 Americas   2020 Cuba            53.2 0.879
#   5 Asia       1997 Vietnam         26.2 2.81
#   6 Asia       2020 United Arab ~   50   1
#   7 Europe     1997 Sweden          40.4 1.48
#   8 Europe     2020 Sweden          47.0 1.13
#   9 Oceania    1997 New Zealand     29.2 2.43
#  10 Oceania    2020 New Zealand     40.8 1.45
```

Overall picture. What are the trends globally? There are various regions defined in the World Bank data. We can plot them and highlight the world "average".

```
cWP %>%
  filter(is.na(Continent)) %>%
  ggplot(aes(Year, pctWiP, group=Country)) +
  geom_line() +
  gghighlight(Country=="World",
              use_direct_label = FALSE,
              use_group_by = FALSE) +
  scale_x_continuous(breaks=seq(1997, 2020, 5)) +
  scale_y_continuous(limits=c(0, 40),
                     breaks=seq(0, 40, by=10)) +
  ggtitle("Women in Parliament: Global Trends") +
  ylab("% Women in Parliament")
```



Interpretation. The grey lines show that regardless of how we define region the general trends are upwards. The "World" percentage (black line) increased between 1997 and 2020. In 2020, women in parliament represented 24% (i.e. a ratio of 3.17 men to each woman), which is still less than half the level before gender parity can be claimed.

10. Conclusion

This guide presented an analysis of the percentage of women in parliament as a real-life case study for some of the tidyverse package. Although the format limited what could be presented, we can conclude that the percentage of women in parliament is increasing but that gender parity in parliaments is still far-off.

There is a lot more that can be said and discussed about the limitations, interpretation and potential impact of this data which the World Bank has nicely summarised.[7] You are strongly encouraged to read their discussion for a more complete understanding.

[7] https://databank.worldbank.org/data/reports.aspx?source=2&type=metadata&series=SG.GEN.PARL.ZS.