GSS R: installation and first exploration

2024-09-02

ToC


```
if (!require(gssr)) {
  if (!require(remotes)){
    install.packages("remotes")
  }
  remotes::install_github("kjhealy/gssr")
}
```

- **L3 MIASHS**
- **Université Paris Cité**
- Année 2024-2025
- Course Homepage

- Moodle

> **Objectives**

Install and use package `gssr`

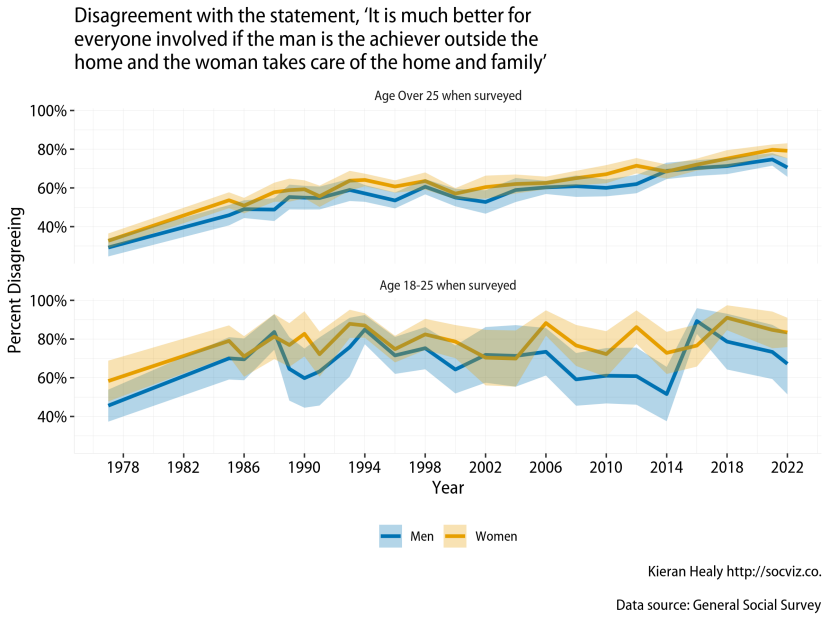
```
PhantomJS not found. You can install it with webshot::install_phantomj
```

gssr

The General Social Survey Cumulative Data (1972-2022, release 2a) and Panel Data files packaged for easy use in R. The companion package to gssr (https://github.com/kjhealy/gssr) is gssrdoc (https://kjhealy.github.io/gssrdoc), which integrates the GSS codebook into R's help system. I recommend you install both packages.

We work again with General Social Survey (GSS) data.We take advantage of `R` package `gssr`

```
if (!require(gssr)) {
  if (!require(remotes)){
    install.packages("rem
  }
  remotes::install_github
}
```



Get data for year 2018

The GSS is carried out every two years. It offers both *cross-sectional* data and *panel* data.

Package `gssr` offers a simple way to retrieve yearly data.

```
df_2018 <- gssr::gss_get_yr(2018)
```

```
Fetching: https://gss.norc.org/documents/stata/2018_stata.zip
```

Inspect the data

- How many observations?
- How many variables?
- Are the data tidy/messy?

Numerical summaries for `age` and `agekdbrn`

The `2018` data provide (among too many other things) columns named `age` abd `agekdbrn`. Get numerical summaries about these two columns.

Thanks to `gssr`, you can get meta-information about the columns

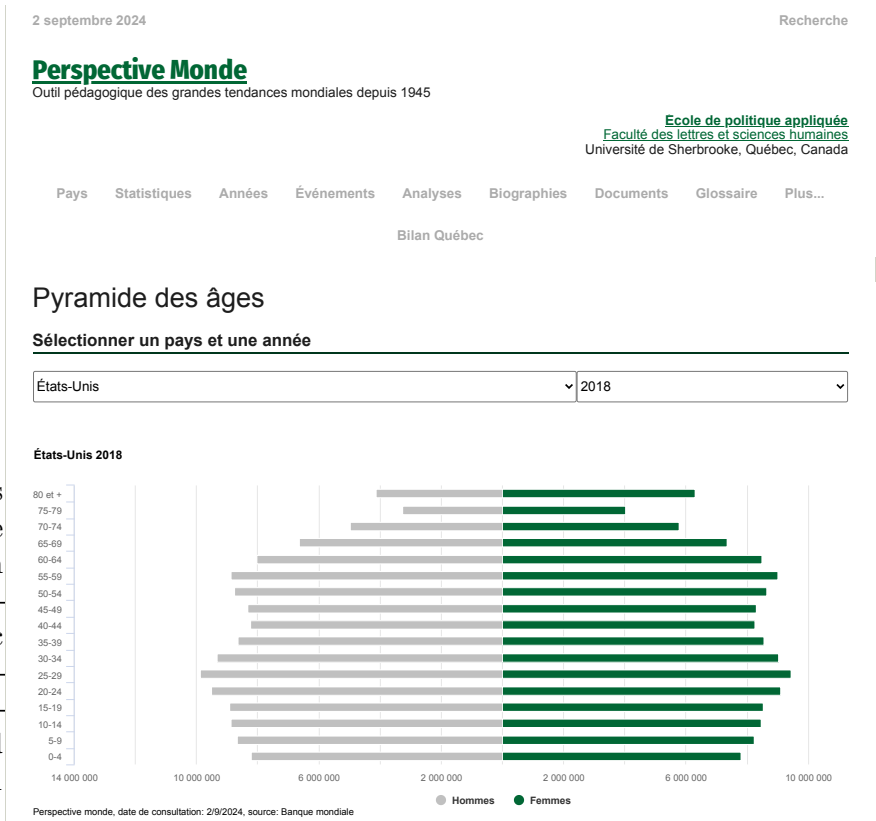
```
?aged
?agekdbrn
?sex
```

How is sex encoded? Is it worth recoding it?

Histogram and density plots for age distribution/facet by sex

Compare *sample* age distribution with *population* age distribution

```
knitr::include_url("https://perspective.usherbrooke.ca/bilan/servlet/BMPagePyramide/USA/2018/?
```



Groupes d'âges	Nombre d'hommes	% des hommes	Nombre de femmes	% des femmes	Ratio femmes / hommes
0-4 ans	9 977 734	6,16	9 519 145	5,77	0,95
5-9 ans	10 526 941	6,50	10 041 037	6,09	0,95
10-14 ans	10 786 330	6,66	10 299 492	6,24	0,95
15-19 ans	10 848 035	6,70	10 399 688	6,31	0,96
20-24 ans	11 559 555	7,14	11 088 684	6,72	0,96
25-29 ans	12 003 933	7,41	11 508 865	6,98	0,96
30-34 ans	11 325 779	7,00	11 019 293	6,68	0,97
35-39 ans	10 502 230	6,49	10 434 960	6,33	0,99
40-44 ans	9 997 623	6,17	10 060 796	6,10	1,01

Sherbrooke University offers visual information about the age structure of population of a wide range of countries.Following demographic usage, the age structure is presented through an age pyramid.Note that an age pyramid is a special kind of histogram

Parallel boxplots of `age` with respect to `sex`

QQplot comparing sample male and female age distributions

Make your own qqplot

Scatterplot for `age` and `agekdbrn`, facet by `sex` ‘

Working with `gss_sub`

```
data("gss_sub")

gss_sub |>
  glimpse()
```

```
Rows: 72,390
Columns: 20
$ year     <dbl+lbl> 1972, 1972, 1972, 1972, 1972, 1972, 1972, 1972, 1972, 197~
$ id       <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18~
$ ballot   <dbl+lbl> NA(i), NA(i), NA(i), NA(i), NA(i), NA(i), NA(i), NA(i), N~
$ age      <dbl+lbl> 23, 70, 48, 27, 61, 26, 28, 27, 21, 30, 30, 56, 54, 49, 4~
$ race     <dbl+lbl> 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 1, 1, 1, ~
$ sex      <dbl+lbl> 2, 1, 2, 2, 2, 1, 1, 1, 2, 2, 2, 1, 1, 2, 1, 1, 1, 2, 2, ~
$ degree   <dbl+lbl> 3, 0, 1, 3, 1, 1, 1, 3, 1, 1, 1, 0, 0, 0, 0, 1, 1, 0, 3, ~
$ padeg    <dbl+lbl>     0,     0,     0,     3,     0,     3,     3,     3,  ~
$ madeg    <dbl+lbl> NA(i),     0,     0,     1,     0,     4,     1,     1,  ~
$ relig    <dbl+lbl> 3, 2, 1, 5, 1, 1, 2, 3, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
$ polviews <dbl+lbl> NA(i), NA(i), NA(i), NA(i), NA(i), NA(i), NA(i), NA(i), N~
$ fefam    <dbl+lbl> NA(i), NA(i), NA(i), NA(i), NA(i), NA(i), NA(i), NA(i), N~
$ vpsu     <dbl+lbl> NA(i), NA(i), NA(i), NA(i), NA(i), NA(i), NA(i), NA(i), N~
$ vstrat   <dbl+lbl> NA(i), NA(i), NA(i), NA(i), NA(i), NA(i), NA(i), NA(i), N~
$ oversamp <dbl+lbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
$ formwt   <dbl+lbl> NA(y), NA(y), NA(y), NA(y), NA(y), NA(y), NA(y), NA(y), N~
$ wtssall  <dbl+lbl> 0.4446, 0.8893, 0.8893, 0.8893, 0.8893, 0.4446, 0.4446, 0~
$ wtssps   <dbl+lbl> 0.6631963, 0.9173700, 0.8974125, 1.0663408, 0.9443237, 0.~
$ sampcode <dbl+lbl> NA(i), NA(i), NA(i), NA(i), NA(i), NA(i), NA(i), NA(i), N~
$ sample   <dbl+lbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
```

Education through generations

What kind of information do we get through variables `degree` and `padeg`?

```
?degree
?padeg
```

Compute contingency table for `degree` and `padeg`

Visualize contingency table for `degree` and `padeg`

Rearrange the levels of `degree` and `padeg`