GSS R: installation and first exploration

2024-09-05

ToC

```r
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`

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.

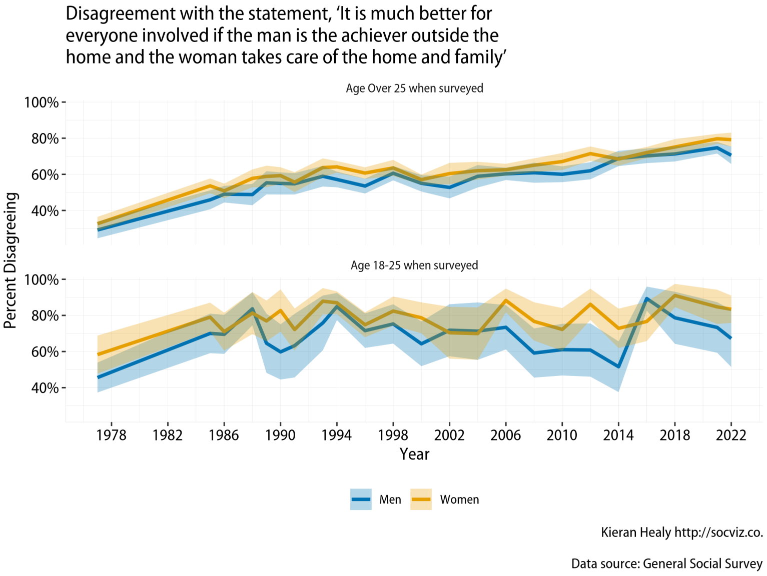
We work again with General Social Survey (GSS) data.We take advantage of R package gssr

```
if (!require(gssr)) {
  if (!require(remotes)
    install.packages("r
  }
  remotes::install_gith
}
```

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

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.



```r
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?

```r
dim(df_2018)
```

```
[1] 2348 1069
```

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.

```r
df_2018 |>
  dplyr::select(age, agekdbrn) |>
  skimr::skim() |>
  skimr::yank("numeric")
```

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
age	7	1.00	48.97	18.06	18	34	48	63	89	
agekdbrn	682	0.71	24.30	5.74	12	20	23	28	51	

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

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

How is sex encoded? Is it worth recoding it?

```
df_2018 |>
  mutate(sex=as_factor(sex)) |>
  skimr::skim(sex) |>
  skimr::yank("factor")
```

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
sex	0	1	FALSE	2	fem: 1296, mal: 1052, don: 0, iap: 0

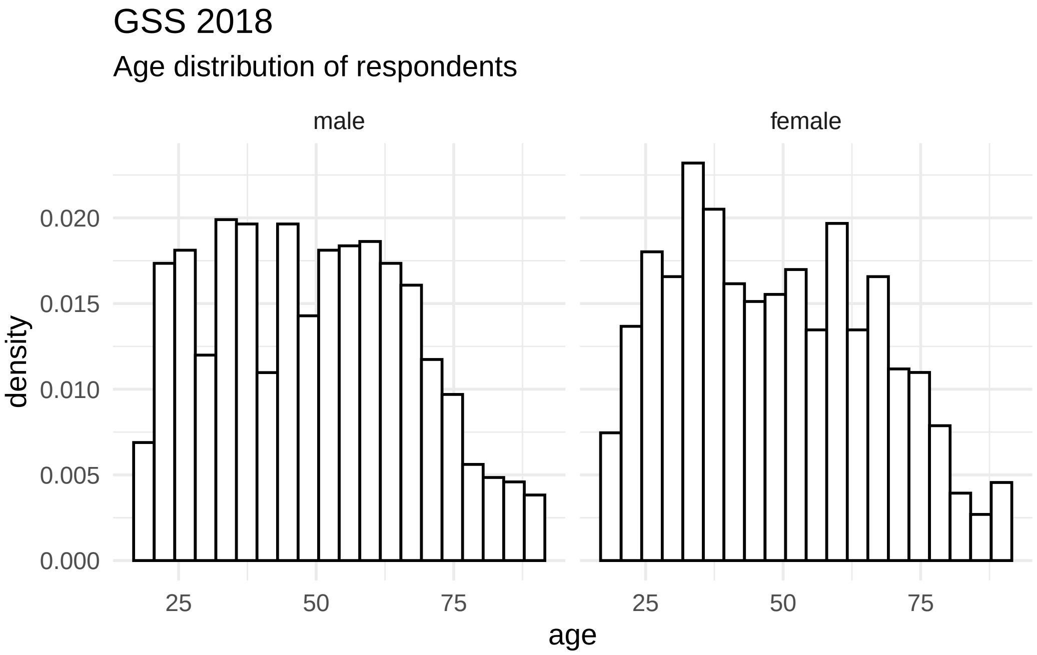
Histogram and density plots for age distribution/facet by sex

```
p_age <- df_2018 |>
  mutate(sex=as_factor(sex)) |>
  ggplot() +
  aes(x=age) +
  facet_wrap(~ sex, )

q_age <- df_2018 |>
  mutate(sex=as_factor(sex)) |>
  ggplot() +
  aes(x=age)
```

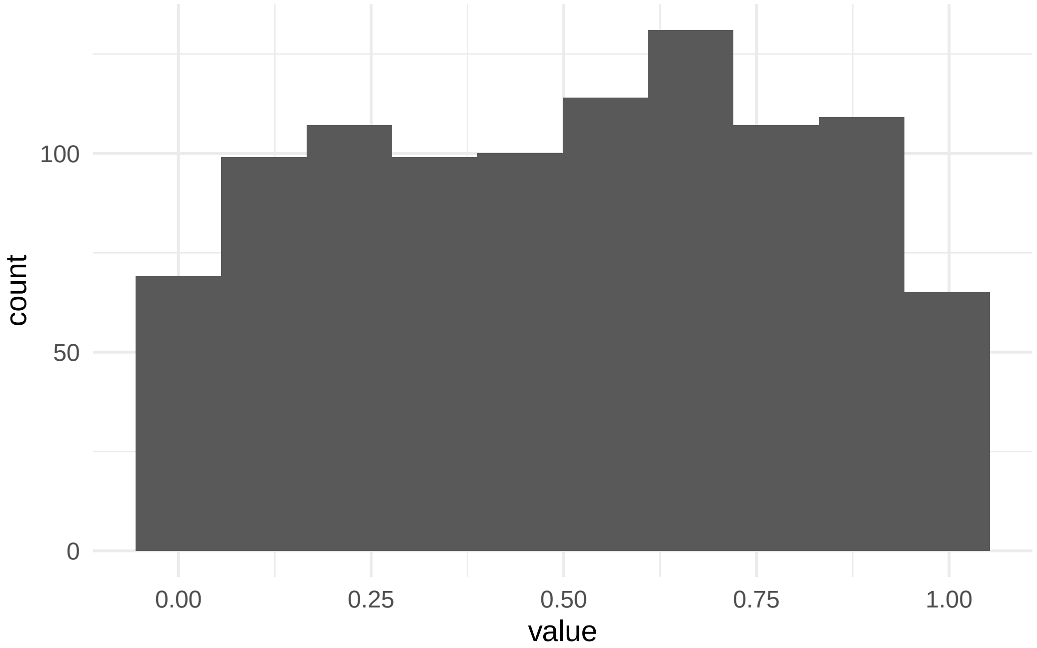
```
p_age +
  geom_histogram(aes(y=after_stat(density)),
                 fill="white",
                 color="black",
                 bins=20) +
  labs(
    title="GSS 2018",
    subtitle = "Age distribution of respondents"
  )
```

```
Warning: Removed 7 rows containing non-finite outside the scale range
(`stat_bin()`).
```



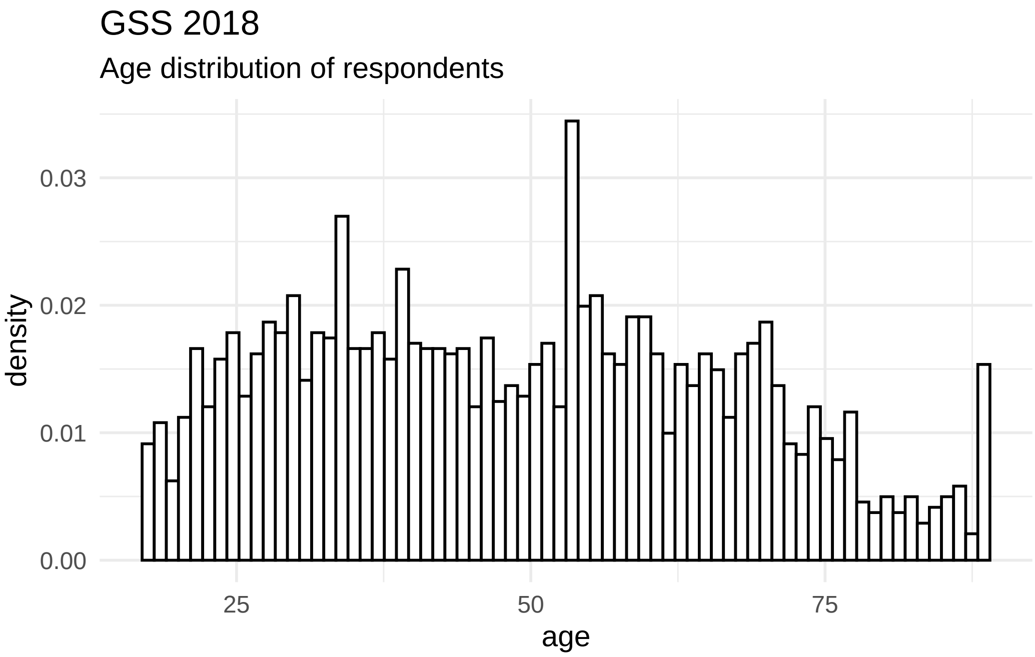
- Play with number of bins
- Spot the irregular behavior of the histograms
- Something special at the right edge of both histograms

```
enframe(x=runif(1000))  |>
  ggplot() +
  aes(x=value) +
  geom_histogram(bins=10)
```



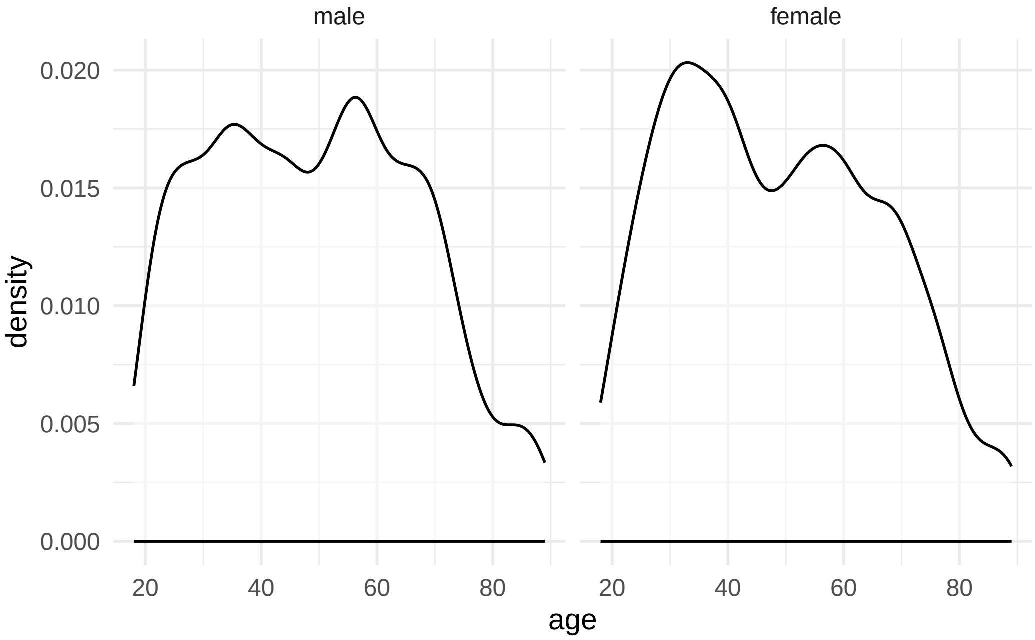
```
q_age +
  geom_histogram(aes(y=after_stat(density)),
                 fill="white",
                 color="black",
                 bins=70) +
  labs(
    title="GSS 2018",
    subtitle = "Age distribution of respondents"
  )
```

```
Warning: Removed 7 rows containing non-finite outside the scale range
(`stat_bin()`).
```



```
p_age +
  stat_density(aes(y=after_stat(density)),
               fill="white",
               alpha=.5,
               color="black",
               bw = "sj",
               adjust = 1
               )
```

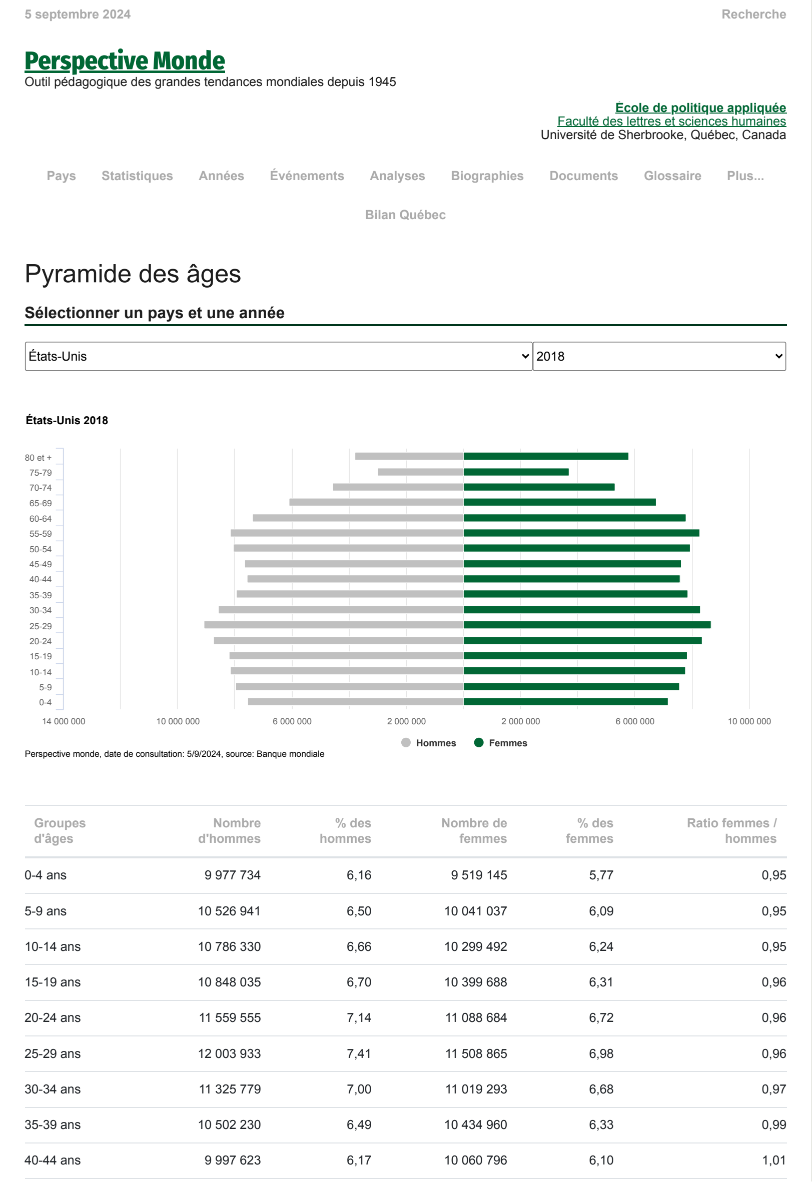
```
Warning: Removed 7 rows containing non-finite outside the scale range
(`stat_density()`).
```



- Play with arguments `bw` and `adjust` of `stat_density`
- Same comments

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

```
knitr::include_url("https://perspective.usherbrooke.ca/bilan/servlet/BMPagePyramide/USA/20
```



5 septembre 2024 Recherche

Perspective Monde

Outil pédagogique des grandes tendances mondiales depuis 1945

Ecole de politique appliquée
Faculté des lettres et sciences humaines
Université de Sherbrooke, Québec, Canada

Pyramide des âges

Sélectionner un pays et une année

États-Unis | 2018

États-Unis 2018

Perspective monde, date de consultation: 5/9/2024, source: Banque mondiale

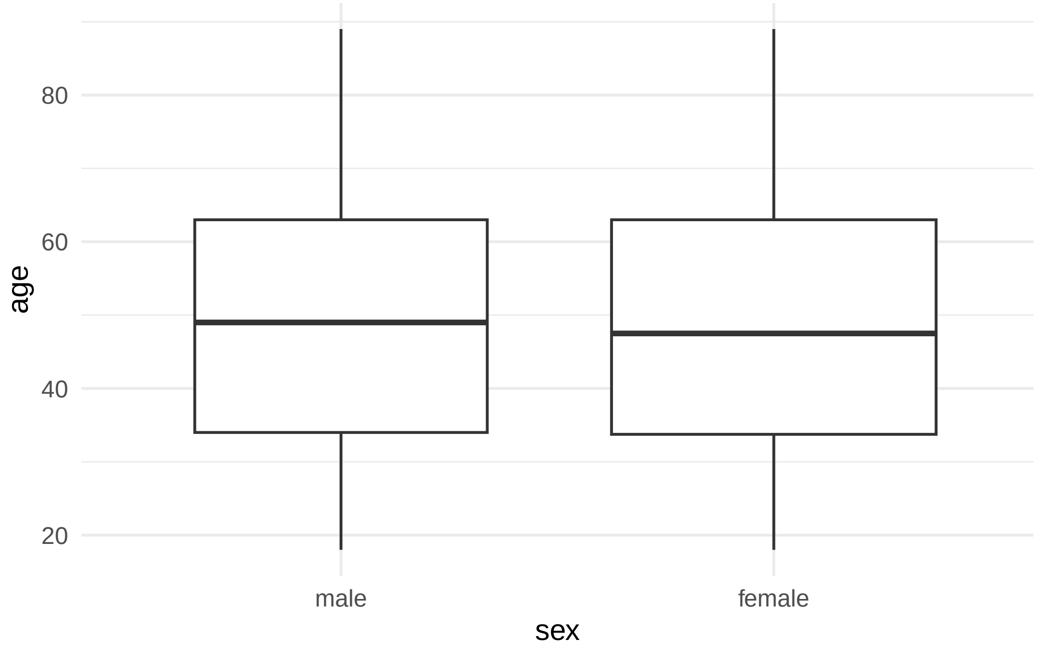
Groupes d'âges	Nombre d'hommes	% des hommes	Nombre de femmes	% des femmes	Ratio femmes / hommes
0-4 ans	9 977 734	6,16	9 519 145	5,77	0,95
5-9 ans	10 526 941	6,50	10 041 037	6,09	0,95
10-14 ans	10 786 330	6,66	10 299 492	6,24	0,95
15-19 ans	10 848 035	6,70	10 399 688	6,31	0,96
20-24 ans	11 559 555	7,14	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

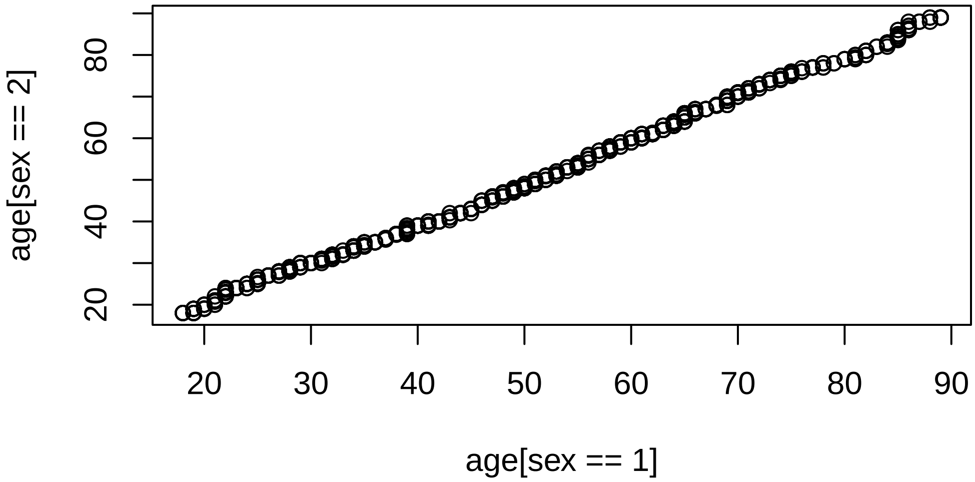
```
df_2018 |>
  mutate(sex=as_factor(sex)) |>
  ggplot() +
  aes(y=age, x=sex) +
  geom_boxplot(varwidth = T) +
  xlab("sex")
```

```
Warning: Removed 7 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```



QQplot comparing sample male and female age distributions

```r
with(df_2018,
     qqplot(x=age[sex==1], y=age[sex==2]))
```



Make your own qqplot

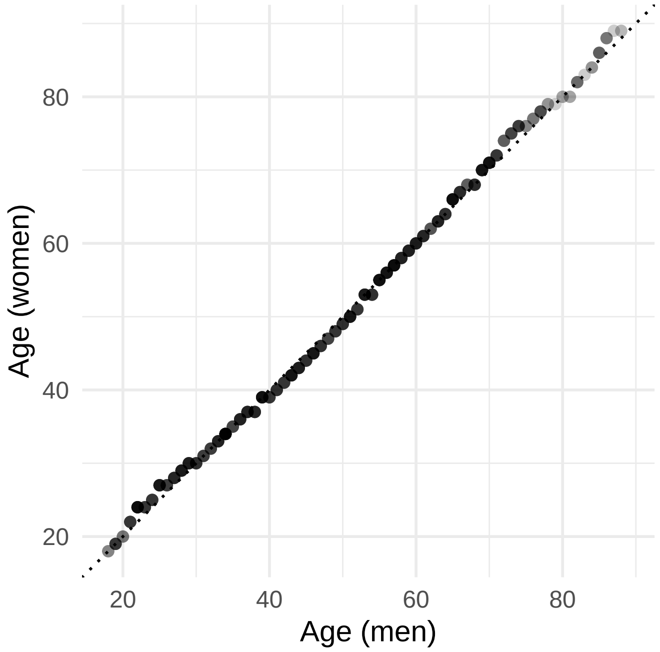
```r
cdf_age_2018_1 <- ecdf(df_2018$age[df_2018$sex==1])

tb <- df_2018 |>
  dplyr::filter(sex==2) |>
  dplyr::select(age) |>
  mutate(Fn=rank(age, ties.method = "max")/n()) |>
  distinct() |>
  arrange(age)

eqf_age_2018_2 <- with(tb,
     stepfun(x=Fn, y=c(age, max(age)), right = T, f = 1))
```

```
filter(df_2018, sex==1) |>
  ggplot() +
  aes(x=age, y=eqf_age_2018_2(cdf_age_2018_1(age))) +
  geom_point(alpha=.1, fill="white") +
  geom_abline(intercept = 0, slope=1, linetype="dotted") +
  coord_fixed() +
  xlab("Age (men)") +
  ylab("Age (women)")
```

```
Warning: Removed 15 rows containing missing values or values outside the scale range
(`geom_point()`).
```



```
# data(gss_all)
```

```
data(gss_dict)
```

```
gss_dict |>
  filter(variable=="age")
```

```
# A tibble: 1 x 13
    pos variable label      missing var_doc_label value_labels var_text years
  <int> <chr>    <chr>        <int> <chr>         <chr>        <chr>    <list>
1    90 age      age of re~     769 age of respo~ [89] 89 or ~ 13. Res~ <tibble>
# i 5 more variables: var_yrtab <list>, var_ballots <list>, col_type <chr>,
#   var_type <chr>, var_na_codes <chr>
```

```
# gss_which_years(gss_all, c("age", "agekdbrn"))
```

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, ~
```

```r
gss_sub |>
  head()
```

```
# A tibble: 6 x 20
  year          id ballot      age    race    sex     degree  padeg   madeg
  <dbl+lbl> <dbl> <dbl+lbl>   <dbl+> <dbl+l> <dbl+l> <dbl+l> <dbl+l> <dbl+lbl>
1 1972          1 NA(i) [iap] 23     1 [whi~ 2 [fem~ 3 [bac~ 0 [les~ NA(i) [iap]
2 1972          2 NA(i) [iap] 70     1 [whi~ 1 [mal~ 0 [les~ 0 [les~     0 [les~
3 1972          3 NA(i) [iap] 48     1 [whi~ 2 [fem~ 1 [hig~ 0 [les~     0 [les~
4 1972          4 NA(i) [iap] 27     1 [whi~ 2 [fem~ 3 [bac~ 3 [bac~     1 [hig~
5 1972          5 NA(i) [iap] 61     1 [whi~ 2 [fem~ 1 [hig~ 0 [les~     0 [les~
6 1972          6 NA(i) [iap] 26     1 [whi~ 1 [mal~ 1 [hig~ 3 [bac~     4 [gra~
# i 11 more variables: relig <dbl+lbl>, polviews <dbl+lbl>, fefam <dbl+lbl>,
#   vpsu <dbl+lbl>, vstrat <dbl+lbl>, oversamp <dbl+lbl>, formwt <dbl+lbl>,
#   wtssall <dbl+lbl>, wtssps <dbl+lbl>, sampcode <dbl+lbl>, sample <dbl+lbl>
```

```r
gss_sub |>
  dplyr::select(-id, -year) |>
  summarise(across(everything(), n_distinct)) |>
  pivot_longer(cols = everything(), names_to="name_col", values_to = "n_distct") |>
  arrange(n_distct) |>
  filter(n_distct < 15) |>
  left_join(gss_dict, by=c("name_col"="variable"))
```

```
# A tibble: 11 x 14
   name_col n_distct   pos label     missing var_doc_label value_labels var_text
   <chr>       <int> <int> <chr>       <int> <chr>         <chr>        <chr>
 1 sex             3   125 responde~     112 respondents ~ [1] male; [~ 23. Cod~
 2 race            4   126 race of ~     107 race of resp~ [1] white; ~ 24. Wha~
 3 ballot          5  6072 ballot u~   21875 ballot used ~ [1] ballot ~ 1659. B~
 4 fefam           5   784 better f~   37259 better for m~ [1] strongl~ 252. No~
 5 oversamp        5  6078 weights ~       0 weights for ~ [1] not 198~ None
 6 degree          6    98 r's high~     196 r's highest ~ [0] less th~ 19. If ~
 7 padeg           6    99 father's~   17881 father's hig~ [0] less th~ 20. If ~
 8 madeg           6   100 mothers ~    8971 mothers high~ [0] less th~ 21. If ~
 9 polviews        8   227 think of~    9672 think of sel~ [1] extreme~ 67a. We~
10 sample         11  6077 sampling~    4032 sampling fra~ [1] 1960 sa~ 1664. T~
11 relig          14   336 r's reli~     437 r's religiou~ [1] protest~ 104. Wh~
# i 6 more variables: years <list>, var_yrtab <list>, var_ballots <list>,
#   col_type <chr>, var_type <chr>, var_na_codes <chr>
```

Education through generations

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

```
?degree
?padeg
```

Compute contingency table for `degree` and `padeg`

```
tab_degree_padeg <- gss_sub |>
  dplyr::select(degree, padeg) |>
  mutate(across(everything(), as_factor)) |>
  table()
```

```
tab_degree_padeg |>
  chisq.test()
```

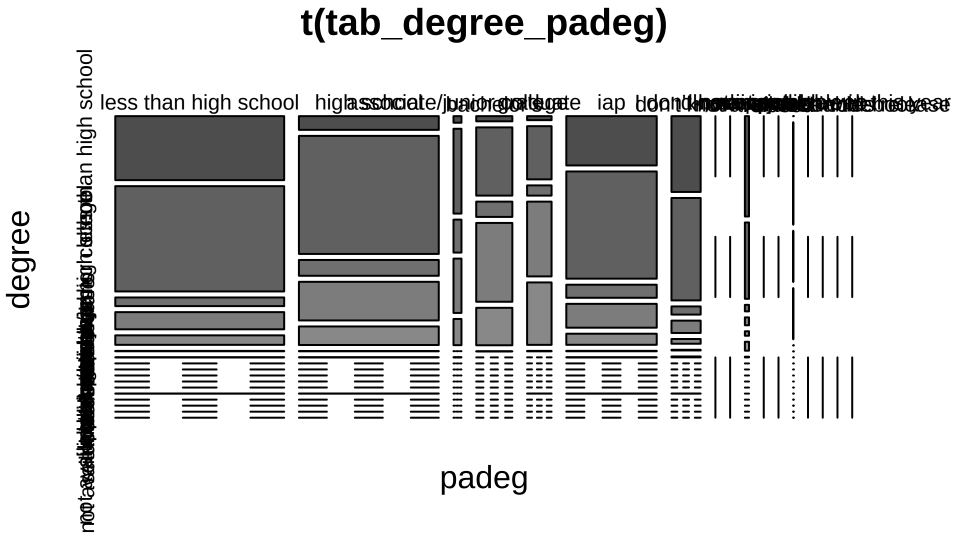
```
Warning in chisq.test(tab_degree_padeg): Chi-squared approximation may be
incorrect

    Pearson's Chi-squared test

data:  tab_degree_padeg
X-squared = NaN, df = 256, p-value = NA
```

Visualize contingency table for `degree` and `padeg`

```
tab_degree_padeg |>
  t() |>
  mosaicplot(color = T)
```



Rearrange the levels of `degree` and `padeg`