The data from the 2018 wave of the [General Social Survey](http://gss.norc.org/) was released, leading to a flurry of graphs showing various trends. The GSS is one of the most important sources of information on various aspects of U.S. society. One of the best things about it is that the data is freely available for more than forty years worth of surveys. Here I’ll walk through my own quick look at the data, in order to show how R can tidily manage data from a complex survey. I decided to revisit a topic that came up a few years ago, in the *New York Times* and elsewhere, regarding the beliefs of young men about gender roles. The idea was that several surveys seemed to point to some increasing conservatism on this front. The GSS has a longstanding question named fefam:

It is much better for everyone involved if the man is the achiever outside the home and the woman takes care of the home and family.

Respondents may answer that they Strongly Agree, Agree, Disagree, or Strongly Disagree with the statement (as well as refusing to answer, or saying they don’t know).

We’ll grab the [GSS Cumulative Data File](http://gss.norc.org/get-the-data/stata) in Stata’s .dta format, and work from there in R, using the Tidyverse tools, Survey package, and srvyr package, which provides a set of wrappers to survey functions that allow them to be piped and worked with in a way familiar to tidyverse residents.

library(tidyverse)

library(forcats)

library(ggrepel)

library(haven)

library(survey)

library(srvyr)

library(tools)

We also load some libraries that aren’t strictly needed, but that will make our plots conform to the house style.

library(showtext)

showtext\_auto()

library(myriad)

import\_myriad\_semi()

This is a quick-and-dirty function we’ll use to clean some age group labels we’ll create in a minute.

convert\_agegrp <- function(x){

x <- gsub("\\(", "", x)

x <- gsub("\\[", "", x)

x <- gsub("\\]", "", x)

x <- gsub(",", "-", x)

x <- gsub("-89", "+", x)

regex <- "^(.\*$)"

x <- gsub(regex, "Age \\1", x)

x

}

my\_colors <- \_function (palette = "cb"){

cb.palette <- c("#000000", "#E69F00", "#56B4E9", "#009E73",

"#F0E442", "#0072B2", "#D55E00", "#CC79A7")

rcb.palette <- rev(cb.palette)

bly.palette <- c("#E69F00", "#0072B2", "#000000", "#56B4E9",

"#009E73", "#F0E442", "#D55E00", "#CC79A7")

if (palette == "cb")

return(cb.palette)

else if (palette == "rcb")

return(rcb.palette)

else if (palette == "bly")

return(bly.palette)

else stop("Choose cb, rcb, or bly only.")

}

The names of some of the weighting and stratifying variables.

cont\_vars <- c("year", "id", "ballot", "age")

cat\_vars <- c("race", "sex", "fefam")

wt\_vars <- c("vpsu",

"vstrat",

"oversamp",

"formwt", # weight to deal with experimental randomization

"wtssall", # weight variable

"sampcode", # sampling error code

"sample") # sampling frame and method

vars <- c(cont\_vars, cat\_vars, wt\_vars)

Load in the data to gss\_all and create a small subset of it, gss\_sm containing just the variables of interest.

gss\_all <- read\_stata("data/GSS7218\_R1.DTA")

gss\_sm <- gss\_all %>%

select(one\_of(vars))

Let’s take a look at it:

gss\_sm

## # A tibble: 64,814 x 14

## year id ballot age race sex fefam vpsu

##

## 1 1972 1 NA(i) [IAP] 23 1 [whi… 2 [fem… NA(i) [IAP] NA(i) [IAP]

## 2 1972 2 NA(i) [IAP] 70 1 [whi… 1 [mal… NA(i) [IAP] NA(i) [IAP]

## 3 1972 3 NA(i) [IAP] 48 1 [whi… 2 [fem… NA(i) [IAP] NA(i) [IAP]

## 4 1972 4 NA(i) [IAP] 27 1 [whi… 2 [fem… NA(i) [IAP] NA(i) [IAP]

## 5 1972 5 NA(i) [IAP] 61 1 [whi… 2 [fem… NA(i) [IAP] NA(i) [IAP]

## 6 1972 6 NA(i) [IAP] 26 1 [whi… 1 [mal… NA(i) [IAP] NA(i) [IAP]

## 7 1972 7 NA(i) [IAP] 28 1 [whi… 1 [mal… NA(i) [IAP] NA(i) [IAP]

## 8 1972 8 NA(i) [IAP] 27 1 [whi… 1 [mal… NA(i) [IAP] NA(i) [IAP]

## 9 1972 9 NA(i) [IAP] 21 2 [bla… 2 [fem… NA(i) [IAP] NA(i) [IAP]

## 10 1972 10 NA(i) [IAP] 30 2 [bla… 2 [fem… NA(i) [IAP] NA(i) [IAP]

## # … with 64,804 more rows, and 6 more variables: vstrat ,

## # oversamp , formwt , wtssall , sampcode ,

## # sample

The read\_stata() function has carried over the labeling information from Stata, which might be useful to us under other circumstances. Columns with, e.g., designations behave like regular (double-precision floating point numbers), but have the label information as metadata.

Now we clean up gss\_sm a bit, discarding some of the label and missing value information we don’t need. We also create some new variables: age quintiles, a variable flagging whether a respondent is 25 or younger, recoded fefam to binary “Agree” or “Disagree” (with non-responses dropped).

qrts <- quantile(as.numeric(gss\_sm$age), na.rm = TRUE)

quintiles <- quantile(as.numeric(gss\_sm$age), probs = seq(0, 1, 0.2), na.rm = TRUE)

gss\_sm <- gss\_sm %>%

modify\_at(vars(), zap\_missing) %>%

modify\_at(wt\_vars, as.numeric) %>%

modify\_at(cat\_vars, as\_factor) %>%

modify\_at(cat\_vars, fct\_relabel, toTitleCase) %>%

mutate(ageq = cut(x = age, breaks = unique(qrts), include.lowest=TRUE),

ageq = fct\_relabel(ageq, convert\_agegrp),

agequint = cut(x = age, breaks = unique(quintiles), include.lowest = TRUE),

agequint = fct\_relabel(agequint, convert\_agegrp),

year\_f = droplevels(factor(year)),

young = ifelse(age < 26, "Yes", "No"),

fefam = fct\_recode(fefam, NULL = "IAP", NULL = "DK", NULL = "NA"),

fefam\_d = fct\_recode(fefam,

Agree = "Strongly Agree",

Disagree = "Strongly Disagree"),

fefam\_n = car::recode(fefam\_d, "'Agree'=0; 'Disagree'=1;", as.factor=FALSE))

gss\_sm$compwt <- with(gss\_sm, oversamp \* formwt \* wtssall)

gss\_sm$samplerc <- with(gss\_sm, ifelse(sample %in% 3:4, 3,

ifelse(sample %in% 6:7, 6,

sample)))

Now we need to take this data and use the survey variables in it, so we can properly calculate population means and errors and so on. We use svyr’s wrappers to survey for this:

options(survey.lonely.psu = "adjust")

options(na.action="na.pass")

gss\_svy <- gss\_sm %>%

filter(year > 1974) %>%

drop\_na(fefam\_d) %>%

mutate(stratvar = interaction(year, vstrat)) %>%

as\_survey\_design(id = vpsu,

strata = stratvar,

weights = wtssall,

nest = TRUE)

Now gss\_svy is a survey object:

## Stratified 1 - level Cluster Sampling design (with replacement)

## With (3585) clusters.

## Called via srvyr

## Sampling variables:

## - ids: vpsu

## - strata: stratvar

## - weights: wtssall

## Data variables: year (dbl), id (dbl), ballot (dbl+lbl), age (dbl+lbl), race (fct), sex (fct), fefam

## (fct), vpsu (dbl), vstrat (dbl), oversamp (dbl), formwt (dbl), wtssall (dbl), sampcode (dbl),

## sample (dbl), ageq (fct), agequint (fct), year\_f (fct), young (chr), fefam\_d (fct), fefam\_n

## (dbl), compwt (dbl), samplerc (dbl), stratvar (fct)

We’re in a position to draw some pictures of fefam trends now.

out\_ff <- gss\_svy %>%

group\_by(year, sex, young, fefam\_d) %>%

summarize(prop = survey\_mean(na.rm = TRUE, vartype = "ci"))

out\_ff

## # A tibble: 168 x 7

## year sex young fefam\_d prop prop\_low prop\_upp

##

## 1 1977 Male No Agree 0.726 0.685 0.766

## 2 1977 Male No Disagree 0.274 0.234 0.315

## 3 1977 Male Yes Agree 0.551 0.469 0.633

## 4 1977 Male Yes Disagree 0.449 0.367 0.531

## 5 1977 Female No Agree 0.674 0.639 0.709

## 6 1977 Female No Disagree 0.326 0.291 0.361

## 7 1977 Female Yes Agree 0.415 0.316 0.514

## 8 1977 Female Yes Disagree 0.585 0.486 0.684

## 9 1985 Male No Agree 0.542 0.496 0.587

## 10 1985 Male No Disagree 0.458 0.413 0.504

## # … with 158 more rows

facet\_names <- c("No" = "Age Over 25 when surveyed", "Yes" = "Age 18-25 when surveyed")

p <- ggplot(subset(out\_ff, fefam\_d == "Disagree"),

aes(x = year, y = prop,

ymin = prop\_low, ymax = prop\_upp,

color = sex, group = sex, fill = sex)) +

geom\_line(size = 1.2) +

geom\_ribbon(alpha = 0.3, color = NA) +

scale\_x\_continuous(breaks = seq(1978, 2018, 4)) +

scale\_y\_continuous(labels = scales::percent\_format(accuracy = 1)) +

scale\_color\_manual(values = my\_colors("bly")[2:1],

labels = c("Men", "Women"),

guide = guide\_legend(title=NULL)) +

scale\_fill\_manual(values = my\_colors("bly")[2:1],

labels = c("Men", "Women"),

guide = guide\_legend(title=NULL)) +

facet\_wrap(~ young, labeller = as\_labeller(facet\_names),

ncol = 1) +

coord\_cartesian(xlim = c(1977, 2017)) +

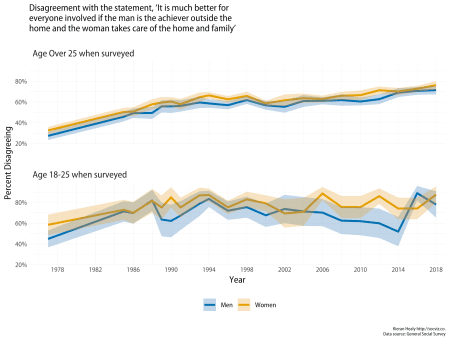
labs(x = "Year",

y = "Percent Disagreeing",

subtitle = "Disagreement with the statement, ‘It is much better for\neveryone involved if the man is the achiever outside the\nhome and the woman takes care of the home and family’) +

theme(legend.position = "bottom")

ggsave("figures/fefam\_svy.png", p, width = 8, height = 6, dpi = 300)



Let’s take a closer look at the age breakdown.

out\_ff\_agequint <- gss\_svy %>%

group\_by(year, agequint, fefam\_d) %>%

summarize(prop = survey\_mean(na.rm = TRUE, vartype = "se")) %>%

mutate(end\_label = if\_else(year == max(year),

socviz::prefix\_strip(as.character(agequint), "Age "), NA\_character\_),

start\_label = if\_else(year == min(year),

socviz::prefix\_strip(as.character(agequint), "Age "), NA\_character\_))

## Warning: Factor `agequint` contains implicit NA, consider using

## `forcats::fct\_explicit\_na`

p <- ggplot(subset(out\_ff\_agequint, fefam\_d == "Disagree"),

aes(x = year, y = prop, ymin = prop - prop\_se, ymax = prop + prop\_se,

color = agequint, group = agequint, fill = agequint)) +

geom\_line(size = 1.2) +

geom\_ribbon(alpha = 0.3, color = NA) +

scale\_x\_continuous(breaks = seq(1978, 2018, 4)) +

scale\_y\_continuous(labels = scales::percent\_format(accuracy = 1)) +

scale\_fill\_manual(values = man\_cols) +

scale\_color\_manual(values = man\_cols) +

guides(fill = FALSE, color = FALSE) +

annotate("text", x = 1974.5, y = 0.58, label = "Age at time\nof survey", size = 3, hjust = 0, fontface = "bold", lineheight = 0.9) +

annotate("text", x = 2020.8, y = 0.95, label = "Age at time\nof survey", size = 3, hjust = 1, fontface = "bold", lineheight = 0.8) +

geom\_label\_repel(aes(label = end\_label), color = "white", nudge\_x = 1) +

geom\_label\_repel(aes(label = start\_label), color = "white", nudge\_x = -1) +

coord\_cartesian(xlim = c(1976, 2019)) +

labs(x = "Year",

y = "Percent",

title = "Changing Attitudes to Gender Roles, by Age Quintiles",

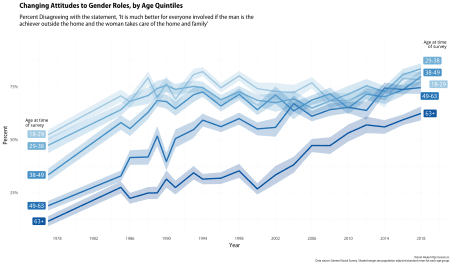
subtitle = "Percent Disagreeing with the statement, ‘It is much better for everyone involved if the man is the\nachiever outside the home and the woman takes care of the home and family’) +

theme(legend.position = "right")

ggsave("figures/fefam\_age\_quin\_svy.png", p, height = 7, width = 12, dpi = 300)

## Warning: Removed 100 rows containing missing values (geom\_label\_repel).

## Warning: Removed 100 rows containing missing values (geom\_label\_repel).



Finally, we can make a plot to get a sense of generational replacement and cohort effects. We’ll make two panels. First, a comparison of more or less the same cohort (though not of course the same individuals): these are people who answered the fefam question in 1977 when they were aged 18-25 and those who answered in 2018 and were aged 63 or older. We’ll also look at two very different cohorts: people who were over 63 in 1977, and people who were aged 18-25 in 2018.

cohort\_comp <- gss\_svy %>%

filter(year %in% c(1977, 2018) &

agequint %in% c("Age 18-29", "Age 63+")) %>%

mutate(cohort = interaction(agequint, year),

cohort = droplevels(cohort)) %>%

group\_by(cohort, fefam) %>%

summarize(prop = survey\_mean(na.rm = TRUE, vartype = "se")) %>%

mutate(cohort = fct\_relabel(cohort, ~ gsub("\\.", " in ", .x)),

cohort = factor(cohort,

levels = c("Age 18-29 in 2018", "Age 63+ in 1977",

"Age 18-29 in 1977", "Age 63+ in 2018"),

ordered = TRUE),

compare = case\_when(cohort %in% c("Age 18-29 in 1977",

"Age 63+ in 2018") ~ "Comparing Approximately the Same Cohort in 1977 and 2018",

cohort %in% c("Age 18-29 in 2018",

"Age 63+ in 1977") ~ "Comparing the Old in 1977 vs the Young in 2018"),

end\_label = if\_else(fefam == "Strongly Disagree",

socviz::prefix\_strip(as.character(cohort), "Age "), NA\_character\_))

p <- ggplot(cohort\_comp,

aes(x = fefam, y = prop, group = cohort,

color = cohort, fill = cohort, ymin = prop - prop\_se, ymax = prop + prop\_se)) +

geom\_point(size = 3) + geom\_line(size = 1.2) +

geom\_ribbon(alpha = 0.2, color = NA) +

scale\_color\_manual(values = my\_colors()) +

scale\_fill\_manual(values = my\_colors()) +

guides(fill = FALSE, color = FALSE) +

scale\_y\_continuous(labels = scales::percent\_format(accuracy = 1)) +

geom\_label\_repel(aes(label = end\_label), fill = "white",

size = 2.2, segment.colour = NA, nudge\_x = 0.6) +

facet\_wrap(~ compare) +

labs(y = "Percent", x = NULL,

title = "Generational Replacement, or, People Don't Change Much, They Just Get Old",

subtitle = "Responses to the statement ‘It is much better for everyone involved if the man is the\nachiever outside the home and the woman takes care of the home and family’")

ggsave("figures/fefam\_age\_quin\_svy\_synth.png", p, height = 7, width = 13, dpi = 300)

## Warning: Removed 12 rows containing missing values (geom\_label\_repel).

