

Polishing Your Plots



```
scale_<MAPPING>_<KIND>()
```

MORE ABOUT
SCALE FUNCTIONS

**BUILD YOUR PLOTS
A PIECE AT A TIME**

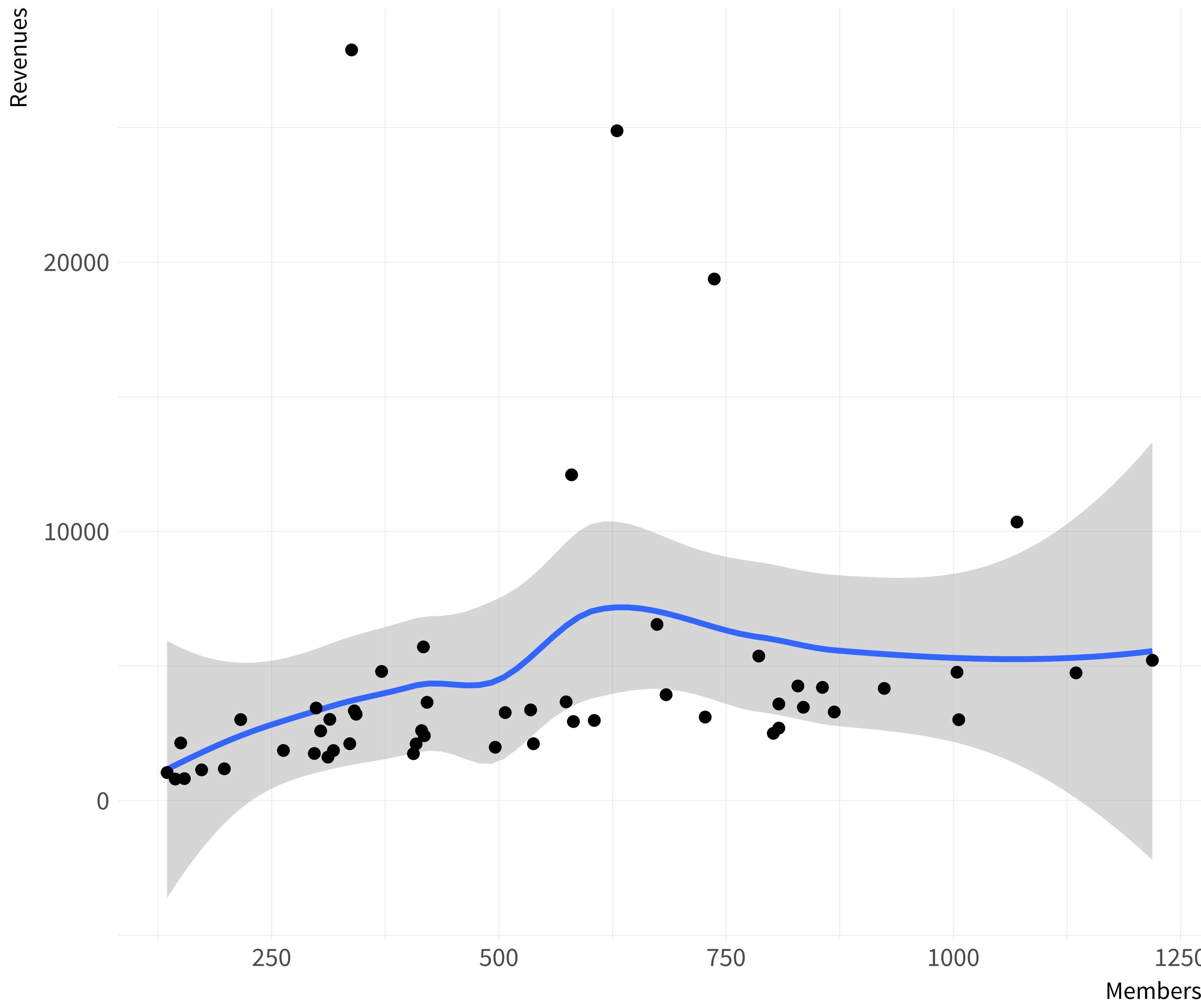
```
head(asasec)
```

```
##                                     Section      Sname Beginning Revenues Expenses Ending Journal Year Members
## 1   Aging and the Life Course (018)    Aging     12752    12104    12007  12849    No 2005    598
## 2   Alcohol, Drugs and Tobacco (030)  Alcohol/Drugs 11933     1144      400  12677    No 2005    301
## 3 Altruism and Social Solidarity (047) Altruism   1139     1862    1875   1126    No 2005      NA
## 4   Animals and Society (042)        Animals    473      820    1116    177    No 2005    209
## 5   Asia/Asian America (024)         Asia      9056    2116    1710  9462    No 2005    365
## 6   Body and Embodiment (048)        Body      3408    1618    1920  3106    No 2005      NA
```

```
dim(asasec)
```

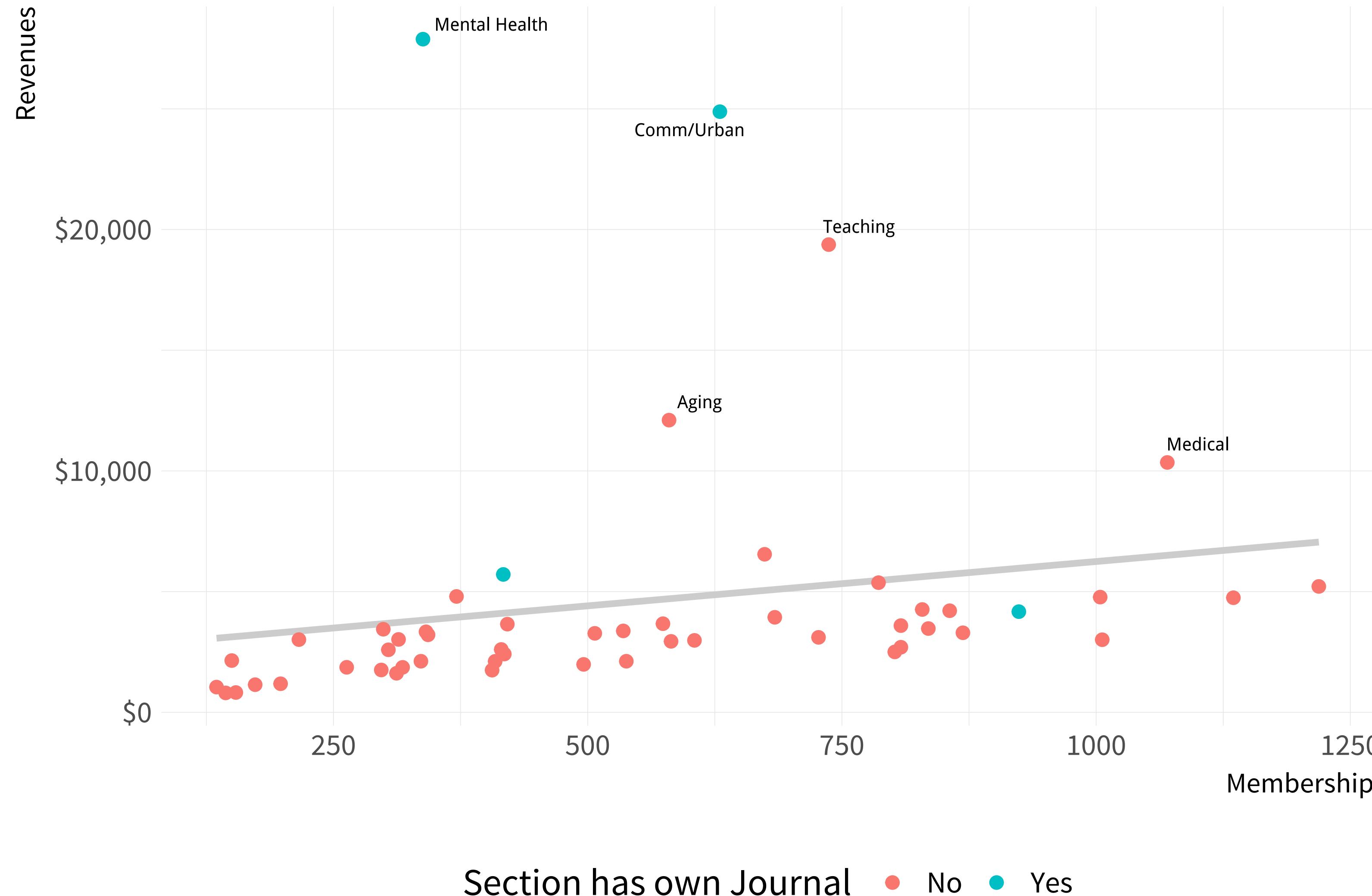
```
## [1] 572 9
```

```
p <- ggplot(subset(asasec, Year == 2014),  
            mapping = aes(x = Members, y = Revenues, label = Sname))  
p + geom_smooth() + geom_point()  
  
## `geom_smooth()` using method = 'loess'
```

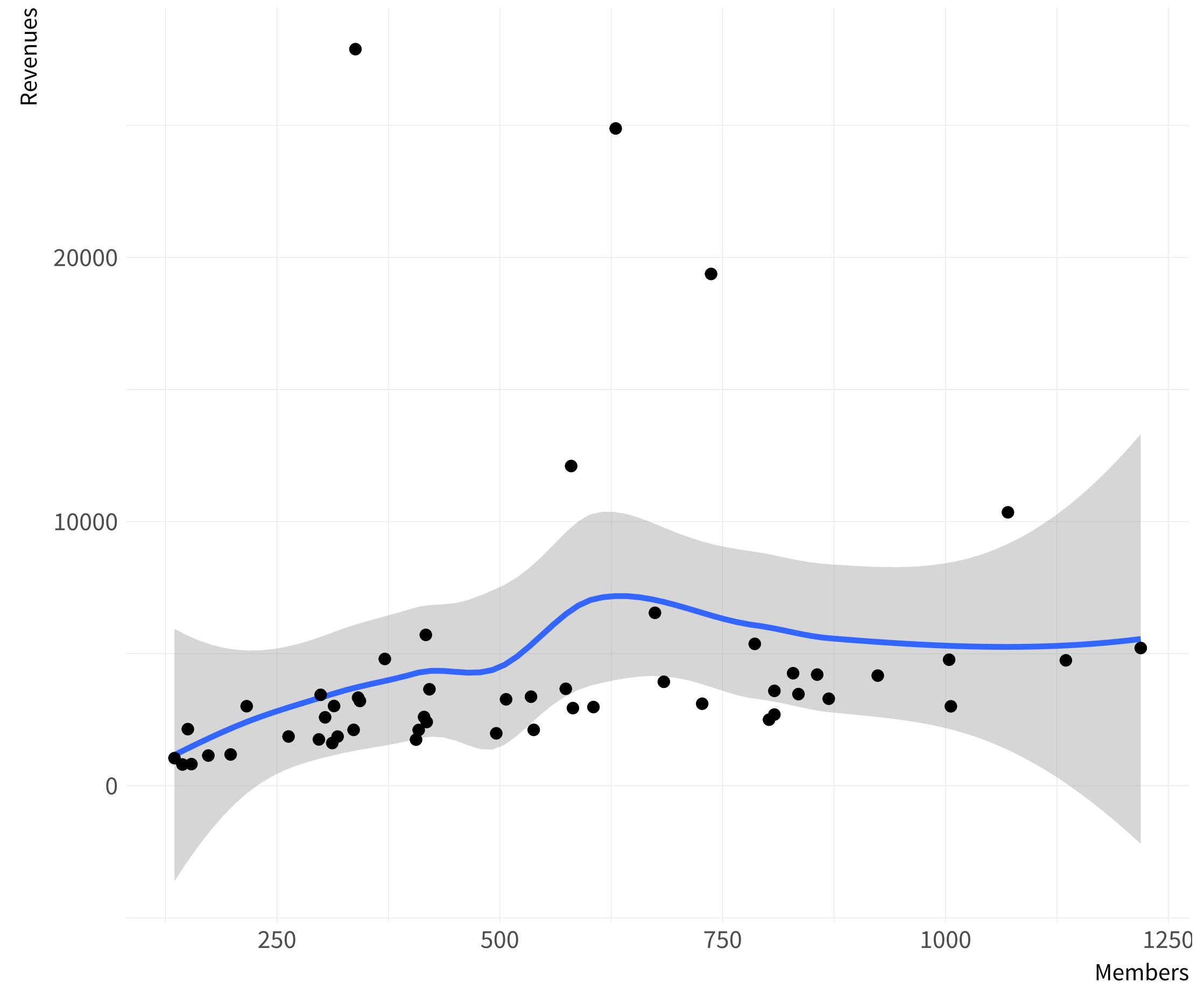


ASA Sections

2014 Calendar year.

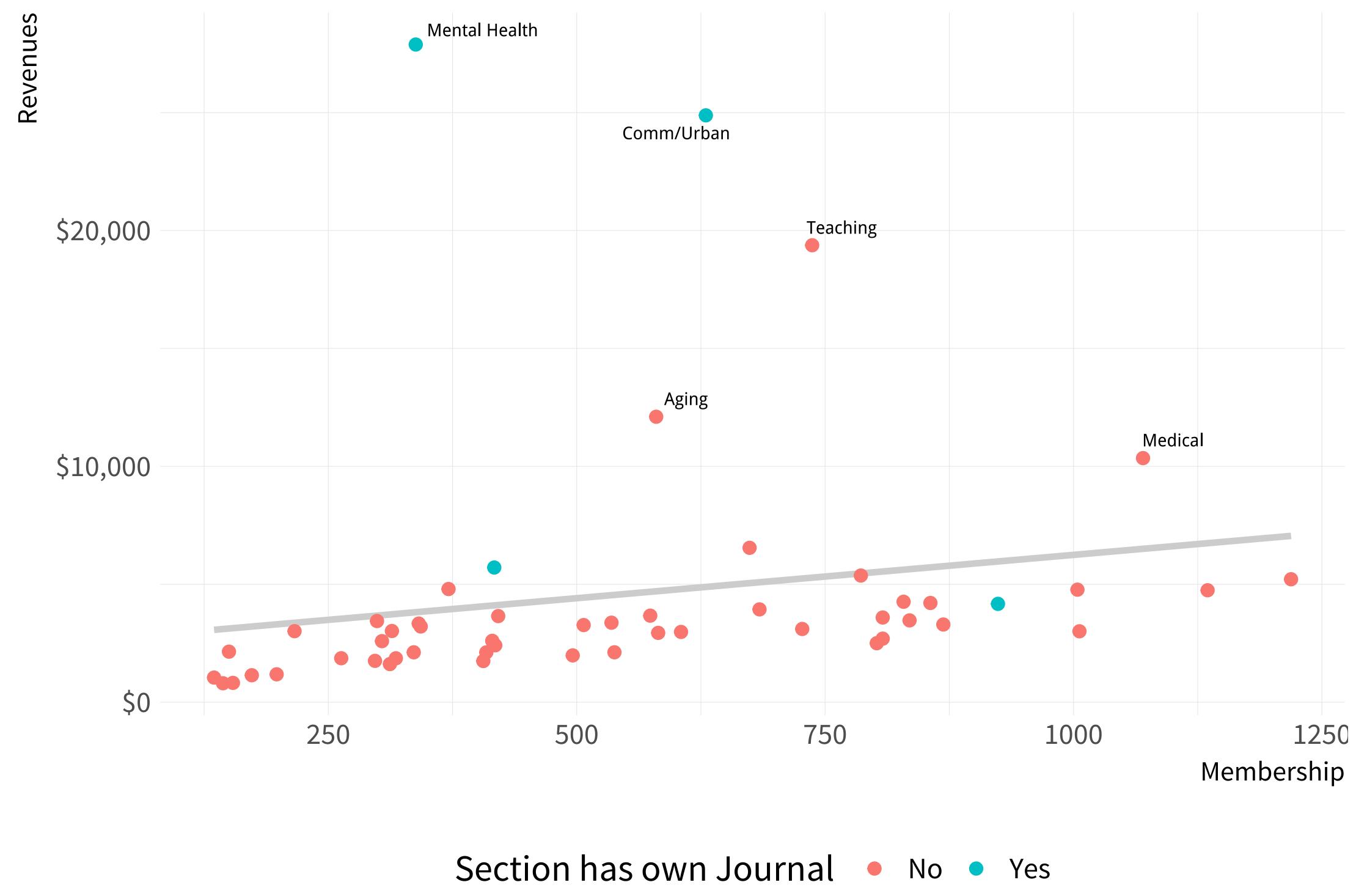


Source: ASA annual report.



ASA Sections

2014 Calendar year.



Source: ASA annual report.

```
p <- ggplot(subset(asasec, Year == 2014),  
            mapping = aes(x = Members, y = Revenues, label = Sname))  
  
p + geom_smooth(method = "lm", se = FALSE, color = "gray80") +  
  geom_point(mapping = aes(color = Journal)) +  
  geom_text_repel(data=subset(asasec,  
                           Year == 2014 & Revenues > 7000))
```

```
p0 <- ggplot(subset(asasec, Year == 2014),  
             mapping = aes(x = Members, y = Revenues, label = Sname))  
  
p1 <- p0 + geom_smooth(method = "lm", se = FALSE, color = "gray80") +  
      geom_point(mapping = aes(color = Journal))  
  
p2 <- p1 + geom_text_repel(data=subset(asasec,  
                                         Year == 2014 & Revenues > 7000),  
                           size = 2)
```

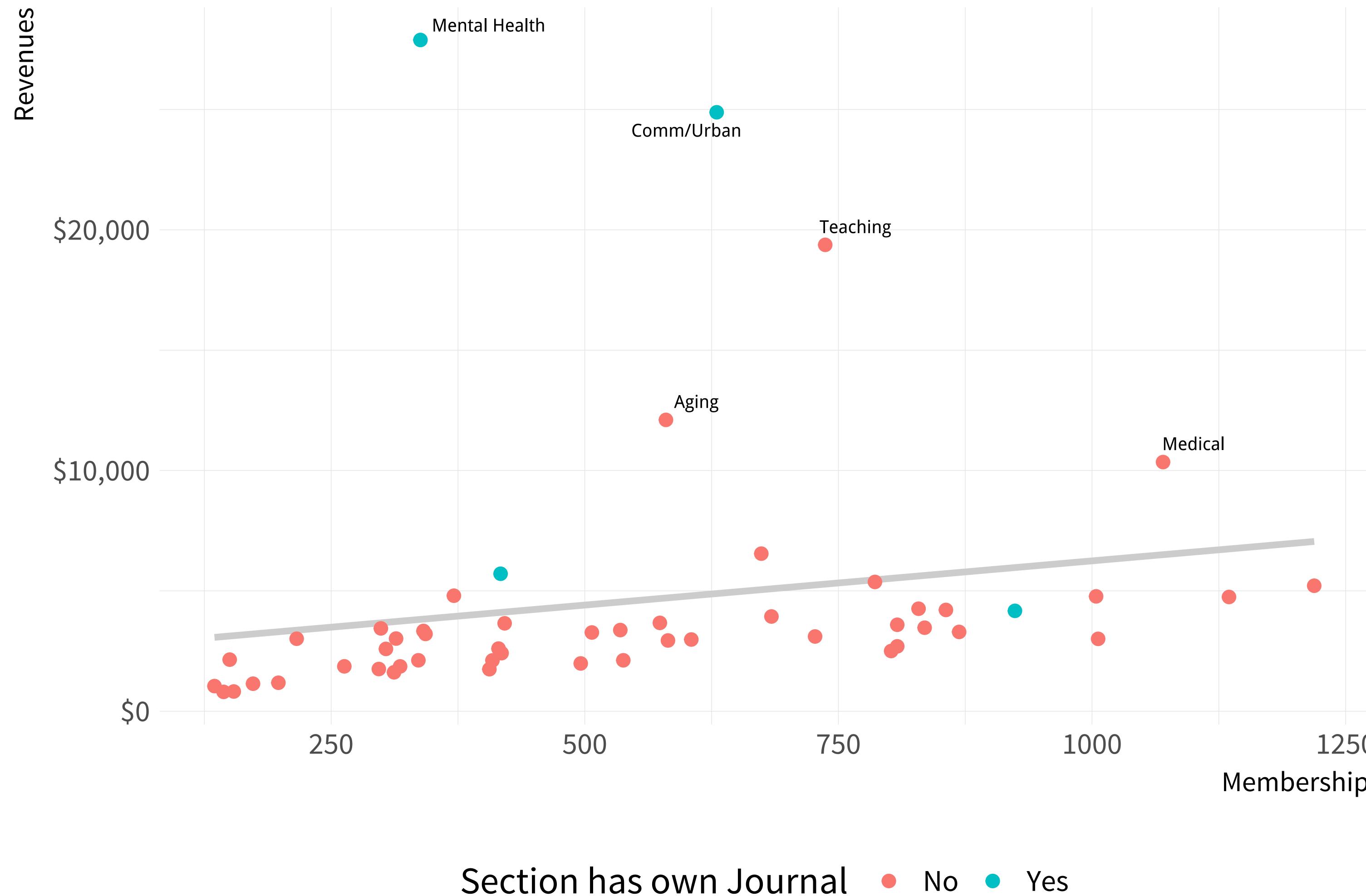
```
p3 <- p2 + labs(x="Membership",
                  y="Revenues",
                  color = "Section has own Journal",
                  title = "ASA Sections",
                  subtitle = "2014 Calendar year.",
                  caption = "Source: ASA annual report.")

p4 <- p3 + scale_y_continuous(labels = scales::dollar) +
      theme(legend.position = "bottom")

p4
```

ASA Sections

2014 Calendar year.



Source: ASA annual report.

**USE COLOR TO
YOUR ADVANTAGE**

color and fill SCALE FUNCTIONS

`scale_<MAPPING>_<KIND>(<ARGUMENTS>)`

```
p <- ggplot(data = organdata,  
             mapping = aes(x = roads,  
                            y = donors,  
                            color = world))
```

```
p + geom_point() + scale_color_hue(l = 40,  
                                         c = 40)
```

You can choose values
directly, but it's better to use
`RColorBrewer` or `colorspace`

`scale_<MAPPING>_<KIND>(<ARGUMENTS>)`

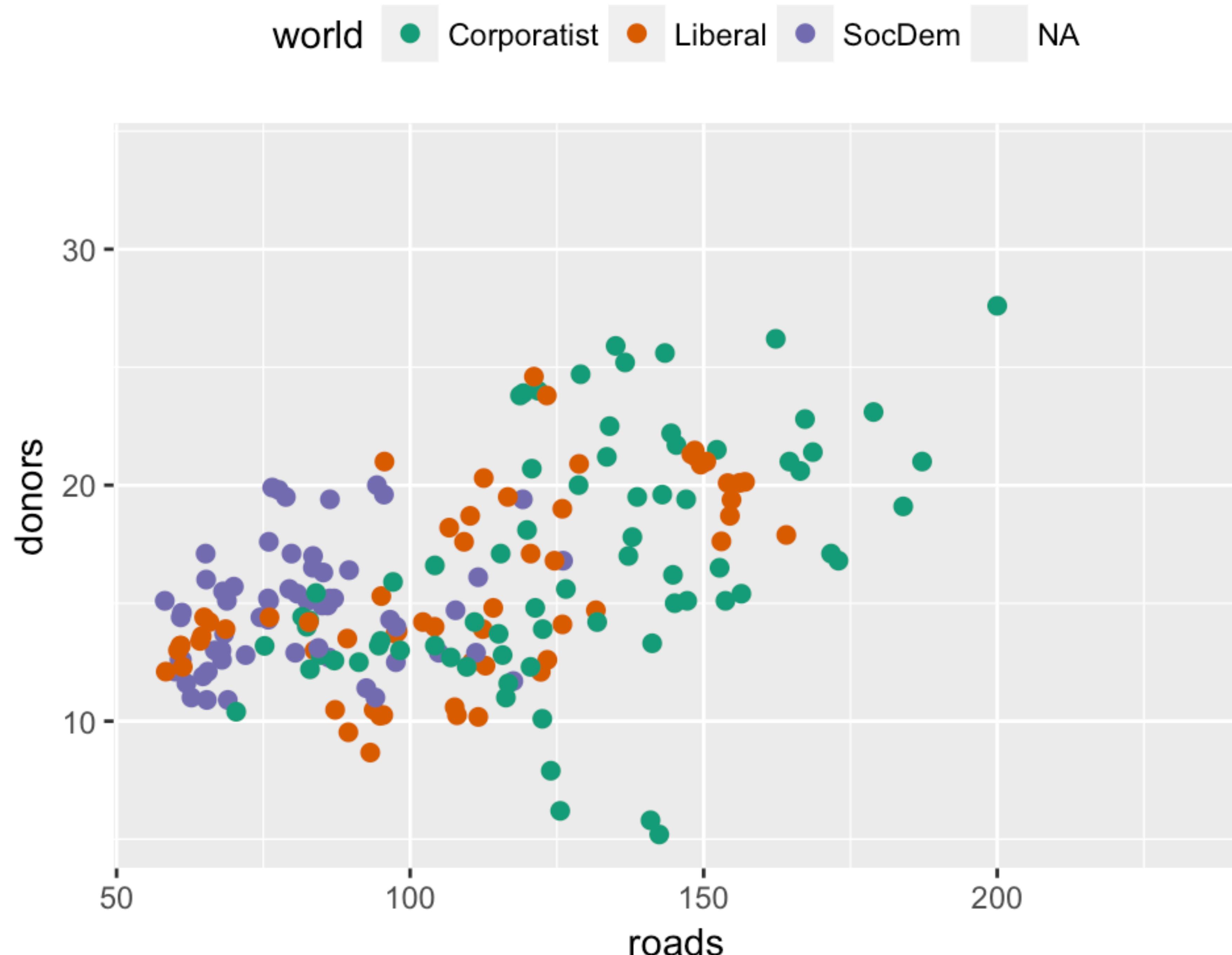
```
p + geom_point(size = 2) + scale_color_brewer(palette = "Set2") +  
  theme(legend.position = "top")
```

```
p + geom_point(size = 2) + scale_color_brewer(palette = "Pastel2") +  
  theme(legend.position = "top")
```

```
p + geom_point(size = 2) + scale_color_brewer(palette = "Dark2") +  
  theme(legend.position = "top")
```

```
p + geom_point(size = 2) + scale_color_brewer(palette = "Accent") +  
  theme(legend.position = "top")
```

Remember color
refers to the mapping



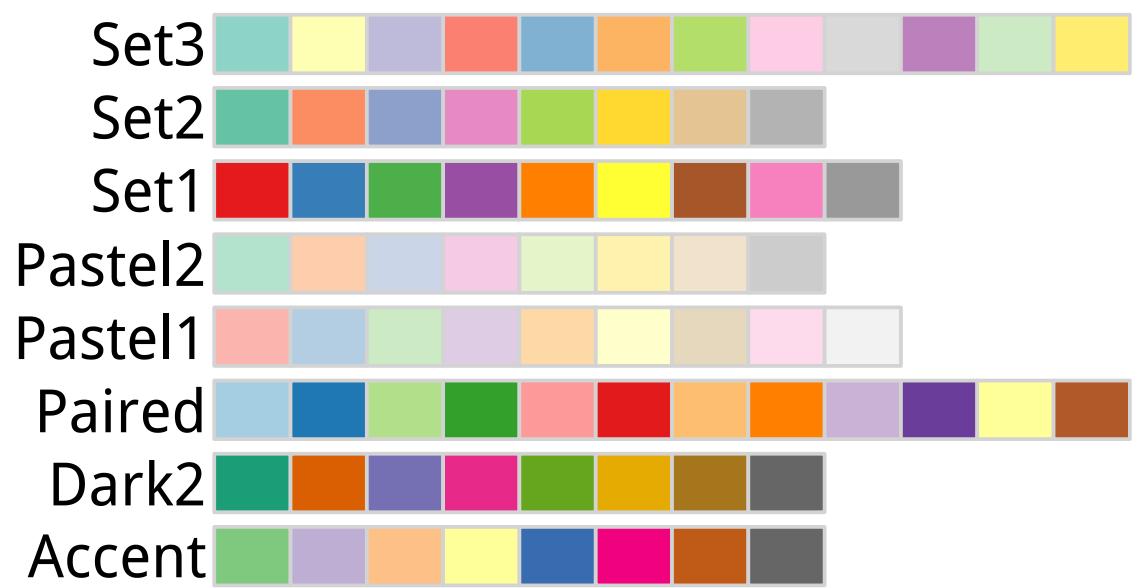
```
library(RColorBrewer)
```

```
display.brewer.all()
```

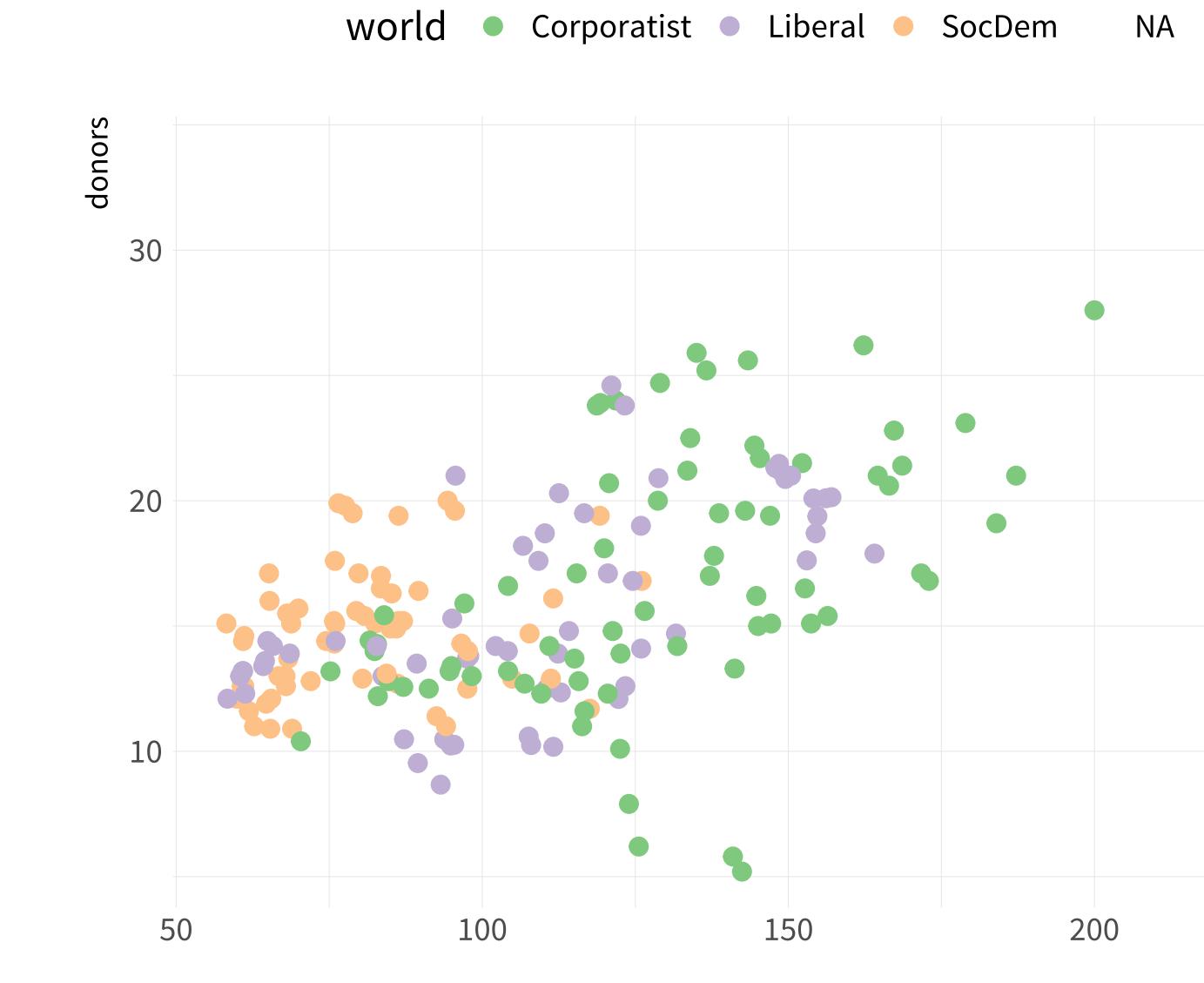
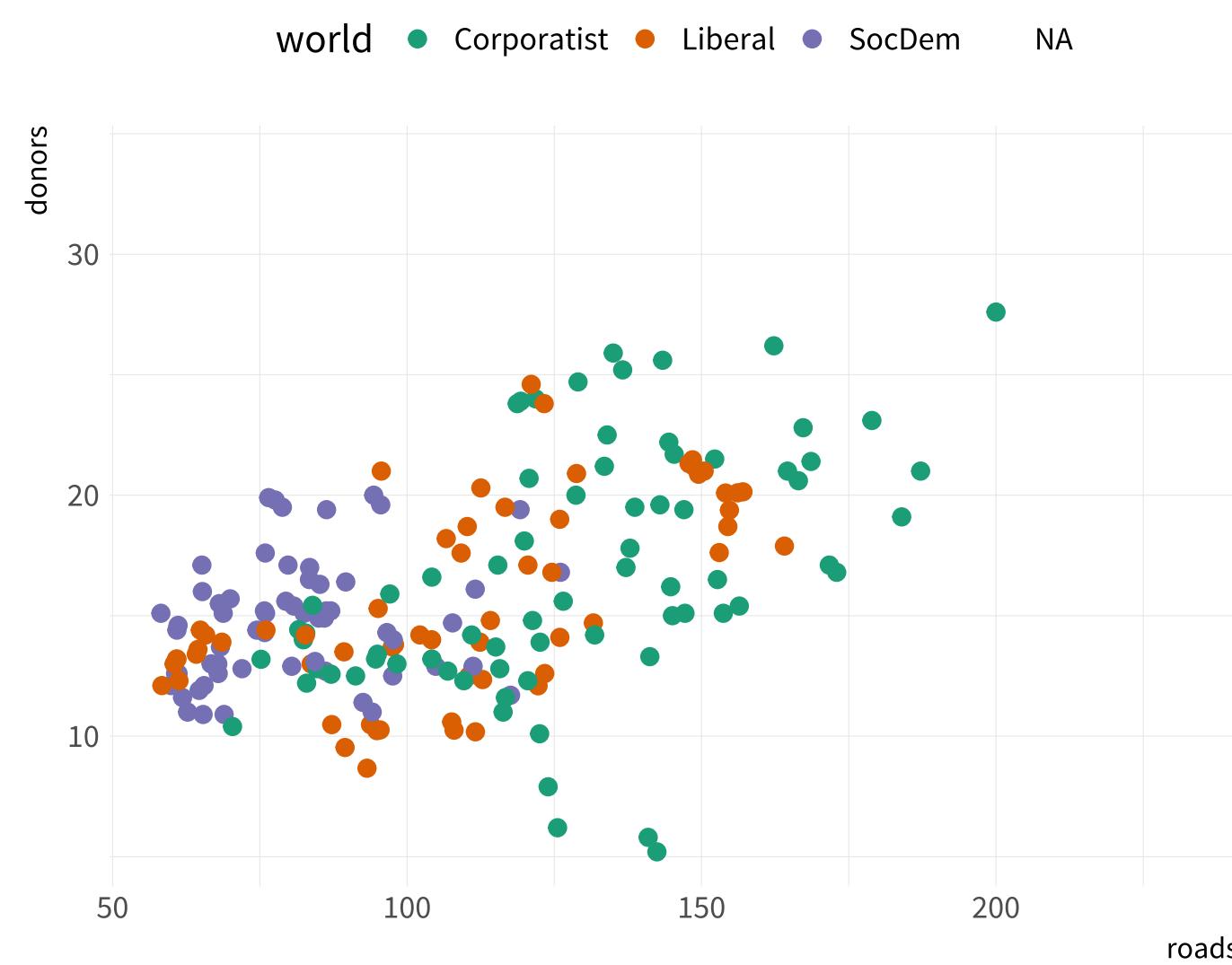
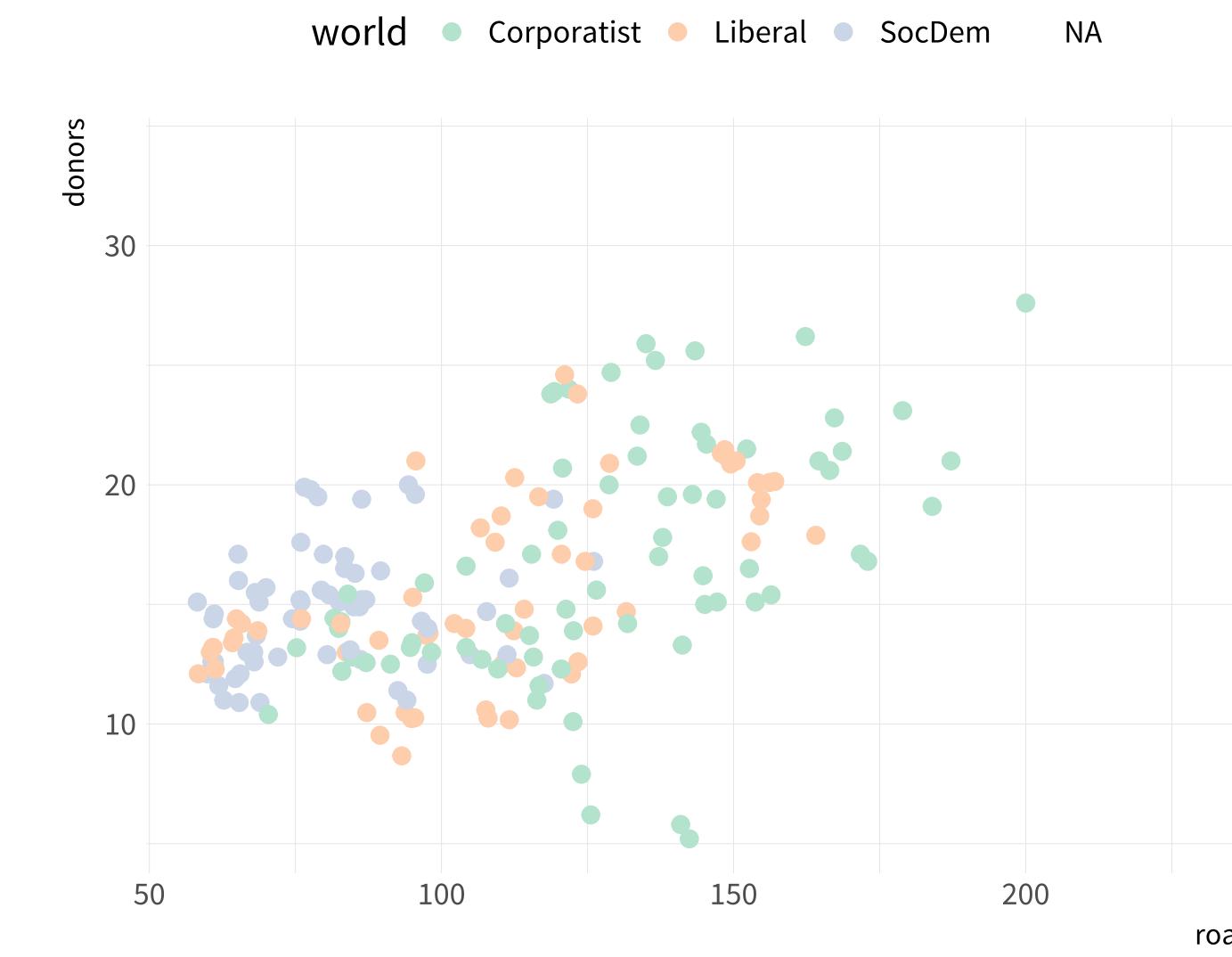
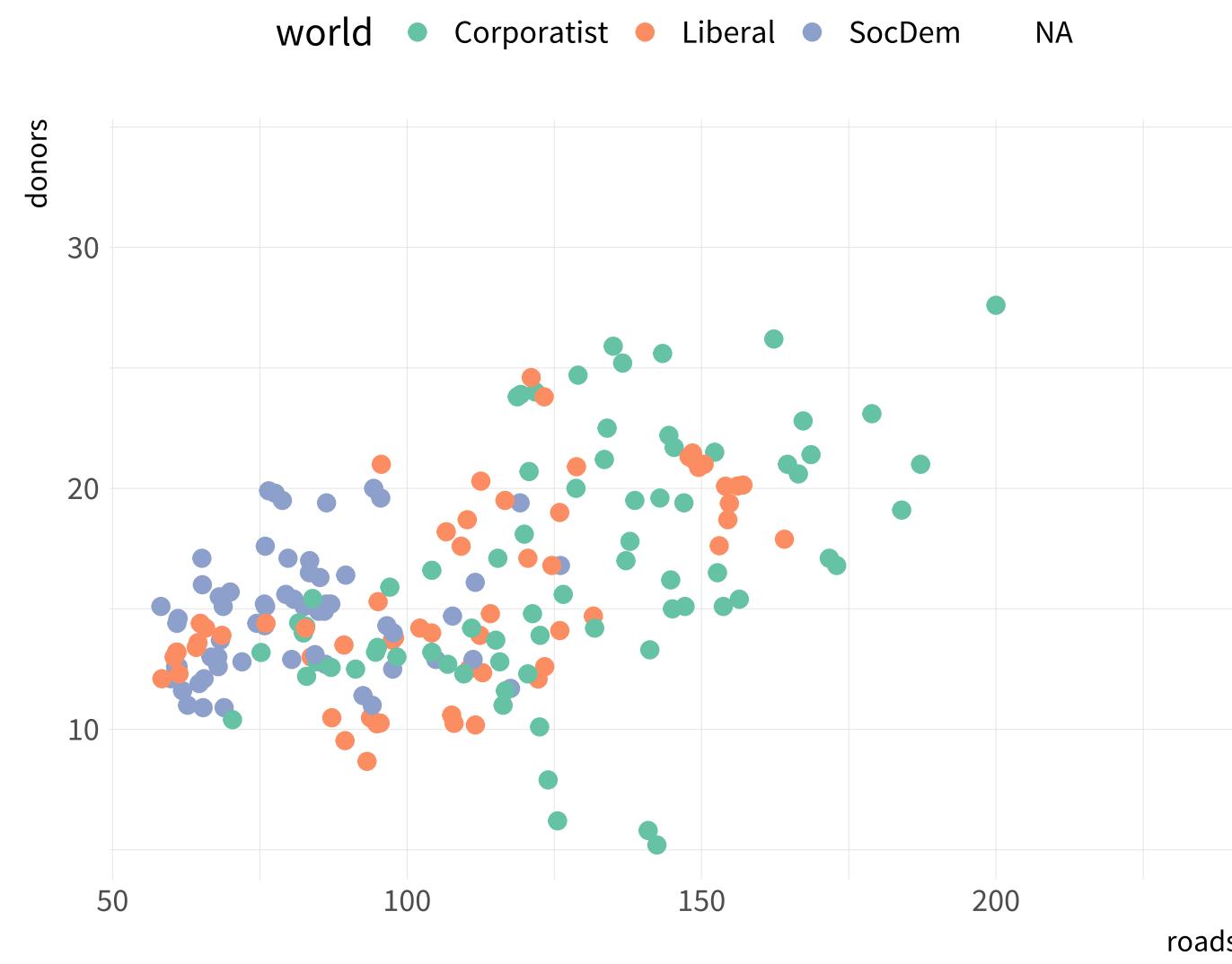
```
scale_color_brewer()
```

```
scale_fill_brewer()
```

```
scale_fill_manual(values = <NAME>)
```



```
p <- ggplot(data = organdata,  
             mapping = aes(x = roads, y = donors, color = world))  
  
p + geom_point(size = 2) + scale_color_brewer(palette = "Set2") +  
  theme(legend.position = "top")  
  
p + geom_point(size = 2) + scale_color_brewer(palette = "Pastel2") +  
  theme(legend.position = "top")  
  
p + geom_point(size = 2) + scale_color_brewer(palette = "Dark2") +  
  theme(legend.position = "top")  
  
p + geom_point(size = 2) + scale_color_brewer(palette = "Accent") +  
  theme(legend.position = "top")
```



```
library("colorspace")
hcl_palettes(plot = TRUE)
```

Qualitative

Pastel 1	[Color Swatches]
Dark 2	[Color Swatches]
Dark 3	[Color Swatches]
Set 2	[Color Swatches]
Set 3	[Color Swatches]
Warm	[Color Swatches]
Cold	[Color Swatches]
Harmonic	[Color Swatches]
Dynamic	[Color Swatches]

Sequential (multi-hue)

Greens 3	[Color Swatches]
Oslo	[Color Swatches]
Purple-Blue	[Color Swatches]
Red-Purple	[Color Swatches]
Red-Blue	[Color Swatches]
Purple-Oran	[Color Swatches]
Purple-Yellow	[Color Swatches]
Blue-Yellow	[Color Swatches]
Green-Yellow	[Color Swatches]
Red-Yellow	[Color Swatches]
Heat	[Color Swatches]
Heat 2	[Color Swatches]
Terrain	[Color Swatches]
Terrain 2	[Color Swatches]
Viridis	[Color Swatches]
Plasma	[Color Swatches]
Inferno	[Color Swatches]
Reds 2	[Color Swatches]
Reds 3	[Color Swatches]
Greens 2	[Color Swatches]

Sequential (single-hue)

BluGrn	[Color Swatches]
Teal	[Color Swatches]
TealGrn	[Color Swatches]
Emrld	[Color Swatches]
BluYl	[Color Swatches]
ag_GrnYl	[Color Swatches]
Peach	[Color Swatches]
PinkYl	[Color Swatches]
Burg	[Color Swatches]
BurgYl	[Color Swatches]
RedOr	[Color Swatches]
OrYel	[Color Swatches]
Purp	[Color Swatches]
PurpOr	[Color Swatches]
Sunset	[Color Swatches]
Magenta	[Color Swatches]
SunsetDark	[Color Swatches]
ag_Sunset	[Color Swatches]
Dark Mint	[Color Swatches]
BrwnYl	[Color Swatches]
Mint	[Color Swatches]

Diverging

YlOrBr	[Color Swatches]
OrRd	[Color Swatches]
Oranges	[Color Swatches]
YlGn	[Color Swatches]
YlGnBu	[Color Swatches]
Reds	[Color Swatches]
RdPu	[Color Swatches]
PuRd	[Color Swatches]
Purples	[Color Swatches]
PuBuGn	[Color Swatches]
PuBu	[Color Swatches]
Greens	[Color Swatches]
BuGn	[Color Swatches]
GnBu	[Color Swatches]
BuPu	[Color Swatches]
Blues	[Color Swatches]
Lajolla	[Color Swatches]
Turku	[Color Swatches]

Blue-Red 2

[Color Swatches]

Blue-Red 3

[Color Swatches]

Red-Green

[Color Swatches]

Purple-Gree

[Color Swatches]

Purple-Brow

[Color Swatches]

Green-Brow

[Color Swatches]

Blue-Yellow 2

[Color Swatches]

Blue-Yellow 3

[Color Swatches]

Green-Oran

[Color Swatches]

Cyan-Mager

[Color Swatches]

Tropic

[Color Swatches]

Broc

[Color Swatches]

Cork

[Color Swatches]

Vik

[Color Swatches]

Berlin

[Color Swatches]

Lisbon

[Color Swatches]

Tofino

[Color Swatches]

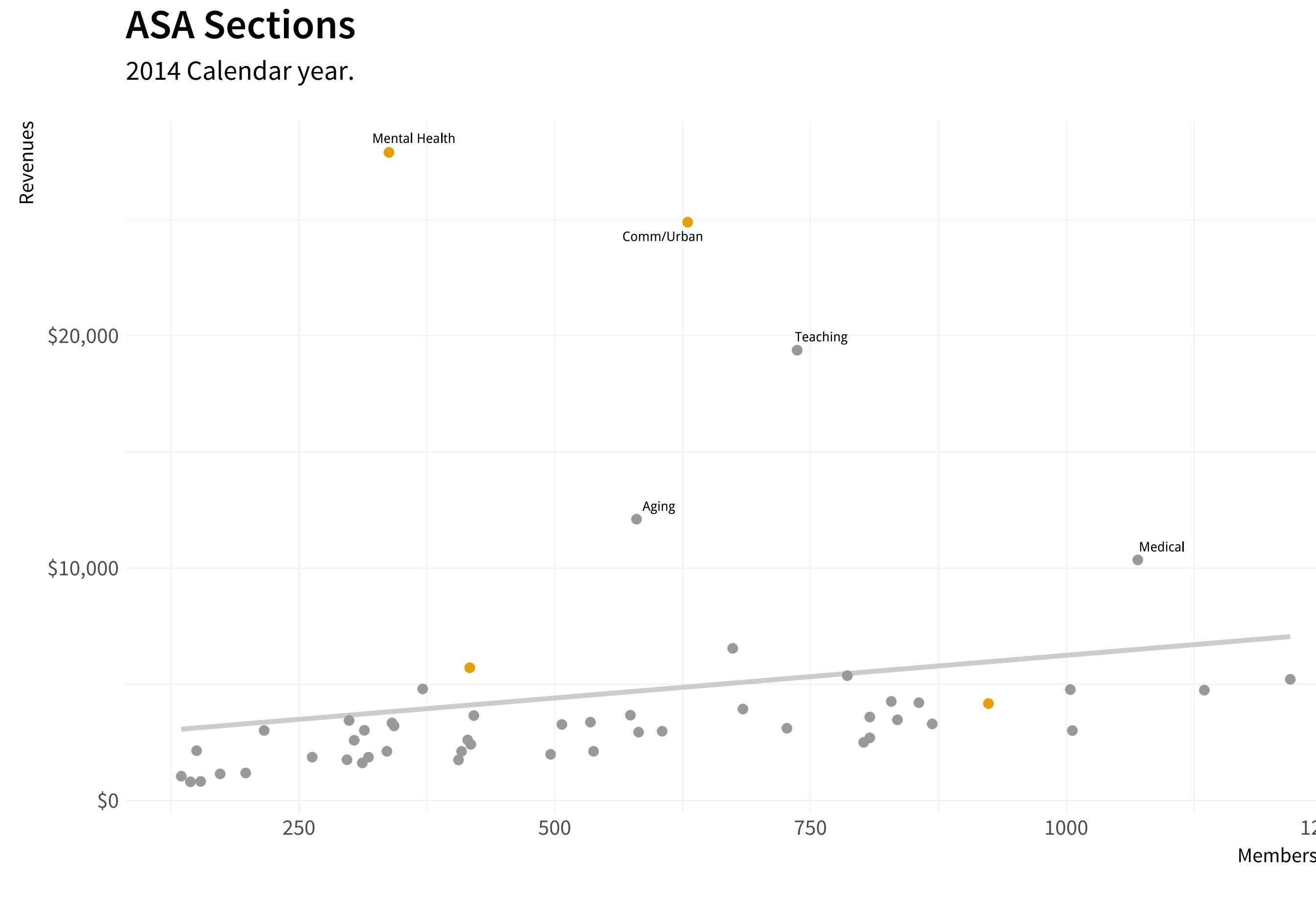
```
library("colorspace")
hcl_palettes(plot = TRUE)
```

scale_<mapping>_<kind>_<colorscale>()



```
cb_palette <- c("#999999", "#E69F00", "#56B4E9", "#009E73",
  "#F0E442", "#0072B2", "#D55E00", "#CC79A7")
```

```
p4 + scale_color_manual(values = cb_palette)
```



Source: ASA annual report.

```
Default <- brewer.pal(5, "Set2")

library(dichromat)

types <- c("deutan", "protan", "tritan")
names(types) <- c("Deuteronomia", "Protanopia", "Tritanopia")

color_table <- types %>%
  purrr::map(~ dichromat(Default, .x)) %>%
  as_tibble() %>%
  add_column(Default, .before = TRUE)

color_table

## # A tibble: 5 x 4
##   Default Deuteronomia Protanopia Tritanopia
##   <chr>    <chr>        <chr>        <chr>
## 1 #66C2A5  #AEAEA7      #BABAA5      #82BDBD
## 2 #FC8D62  #B6B661      #9E9E63      #F29494
## 3 #8DA0CB  #9C9CCB      #9E9ECB      #92ABAB
## 4 #E78AC3  #ACACC1      #9898C3      #DA9C9C
## 5 #A6D854  #CACAC5E     #D3D355      #B6C8C8
```

Default



color_comp(color_table)

Deuteranopia



Protanopia

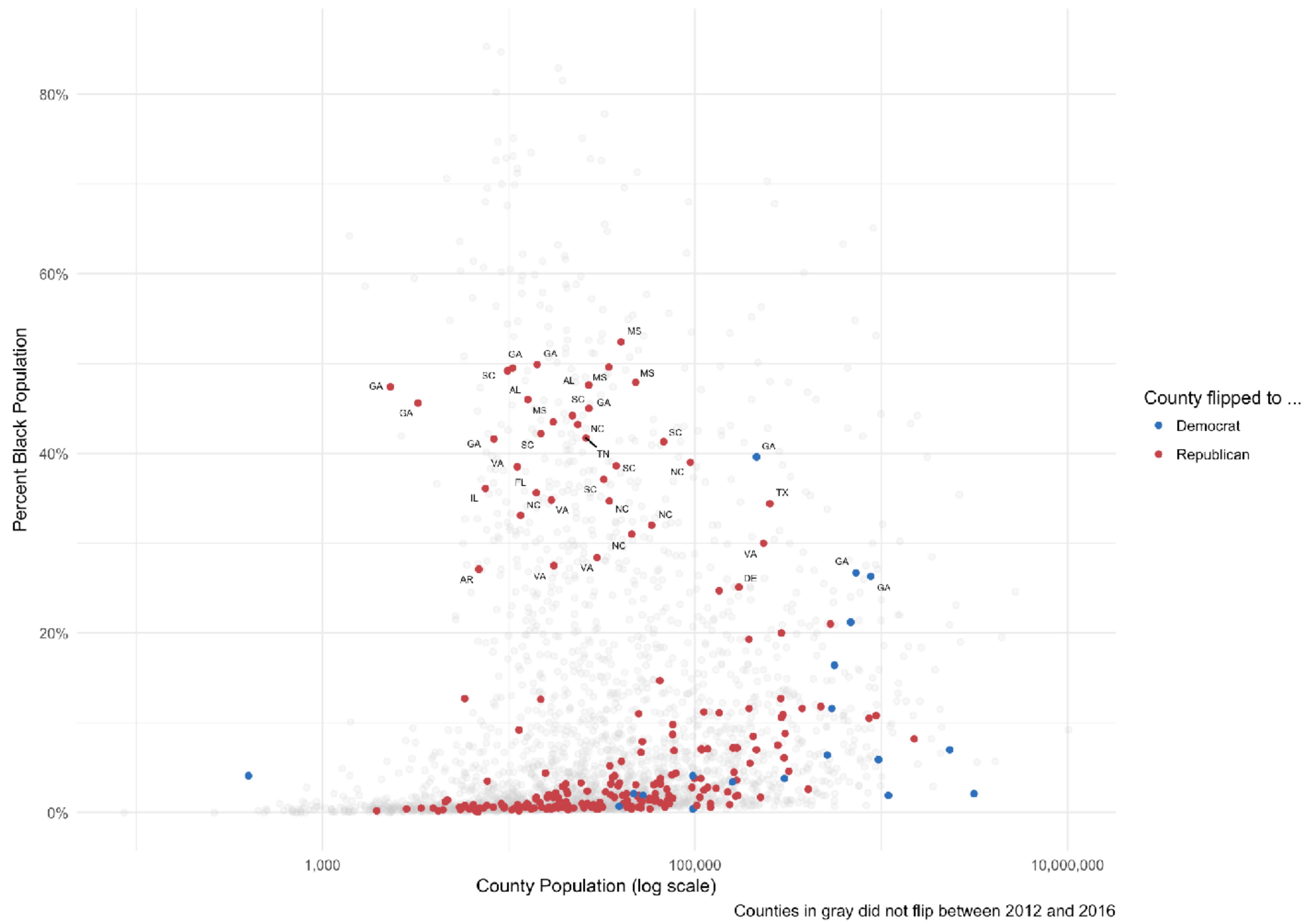


Tritanopia



**Layer Color and
Text Together**

Flipped counties, 2016

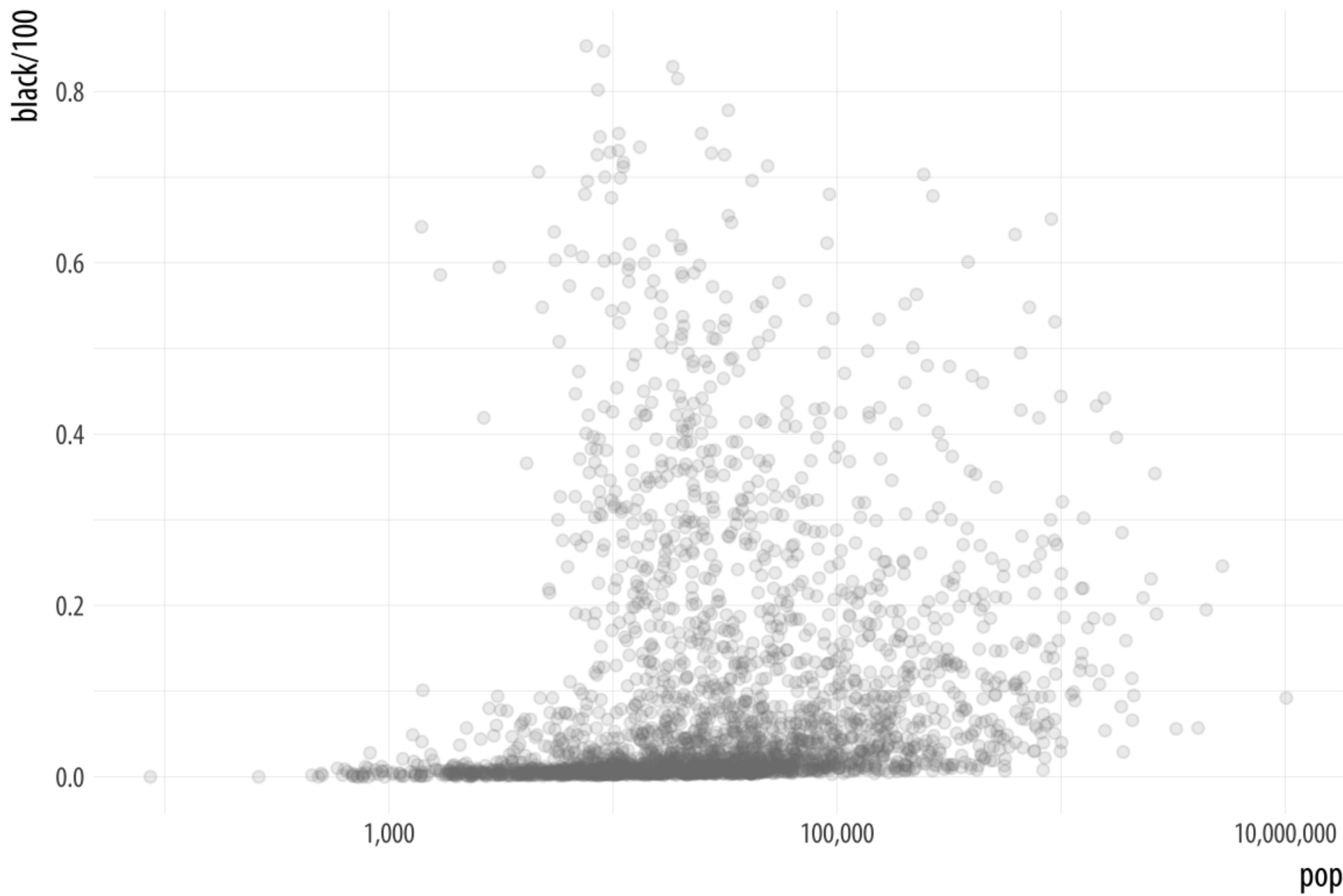


```
# Democrat Blue and Republican Red
party_colors <- c("#2E74C0", "#CB454A")

p0 <- ggplot(data = subset(county_data,
                           flipped == "No"),
              mapping = aes(x = pop,
                            y = black/100))

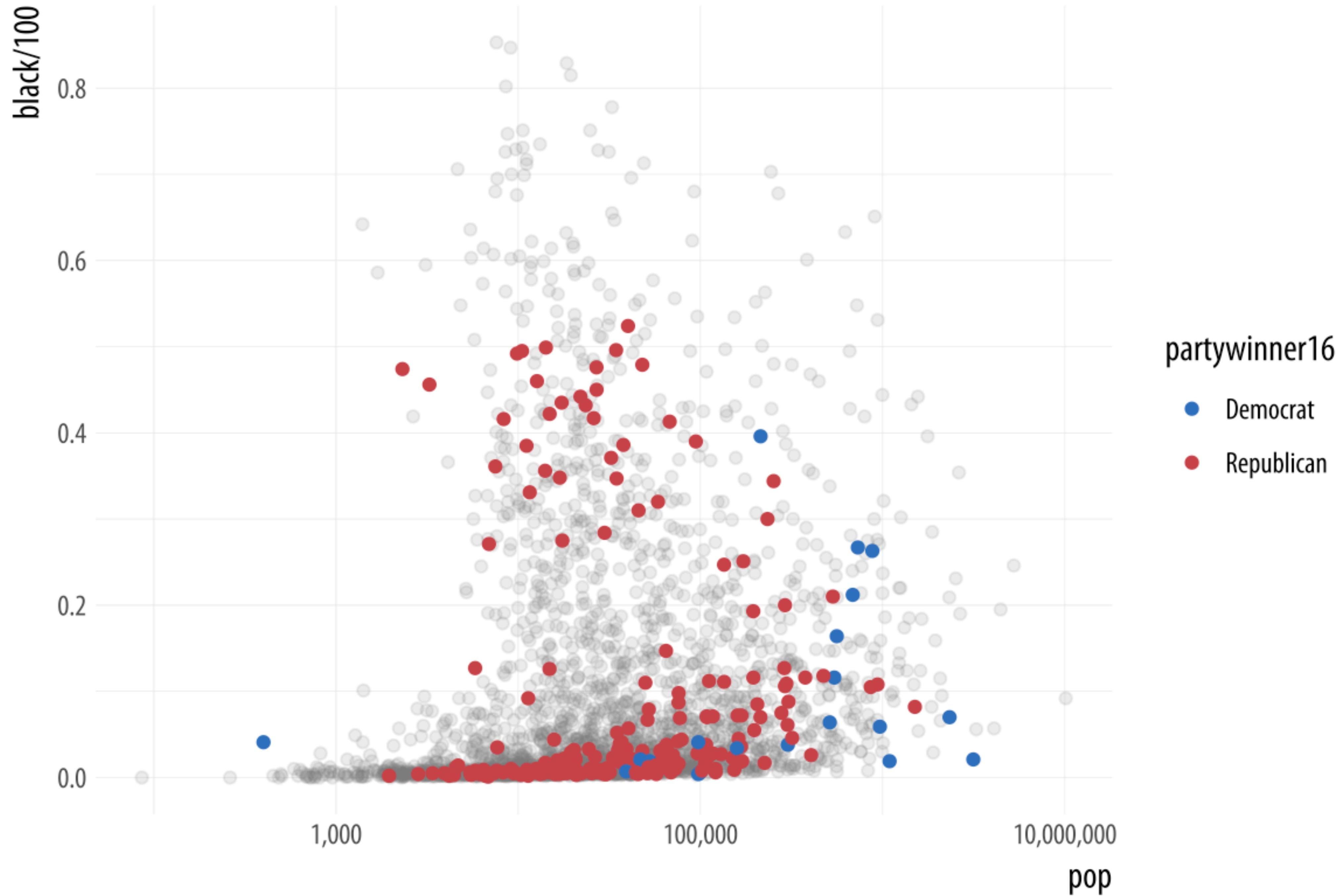
p1 <- p0 + geom_point(alpha = 0.15, color = "gray50") +
  scale_x_log10(labels=scales::comma)

p1
```



```
p2 <- p1 + geom_point(data = subset(county_data,
                                      flipped == "Yes"),
                        mapping = aes(x = pop, y = black/100,
                                      color = partywinner16)) +
scale_color_manual(values = party_colors)
```

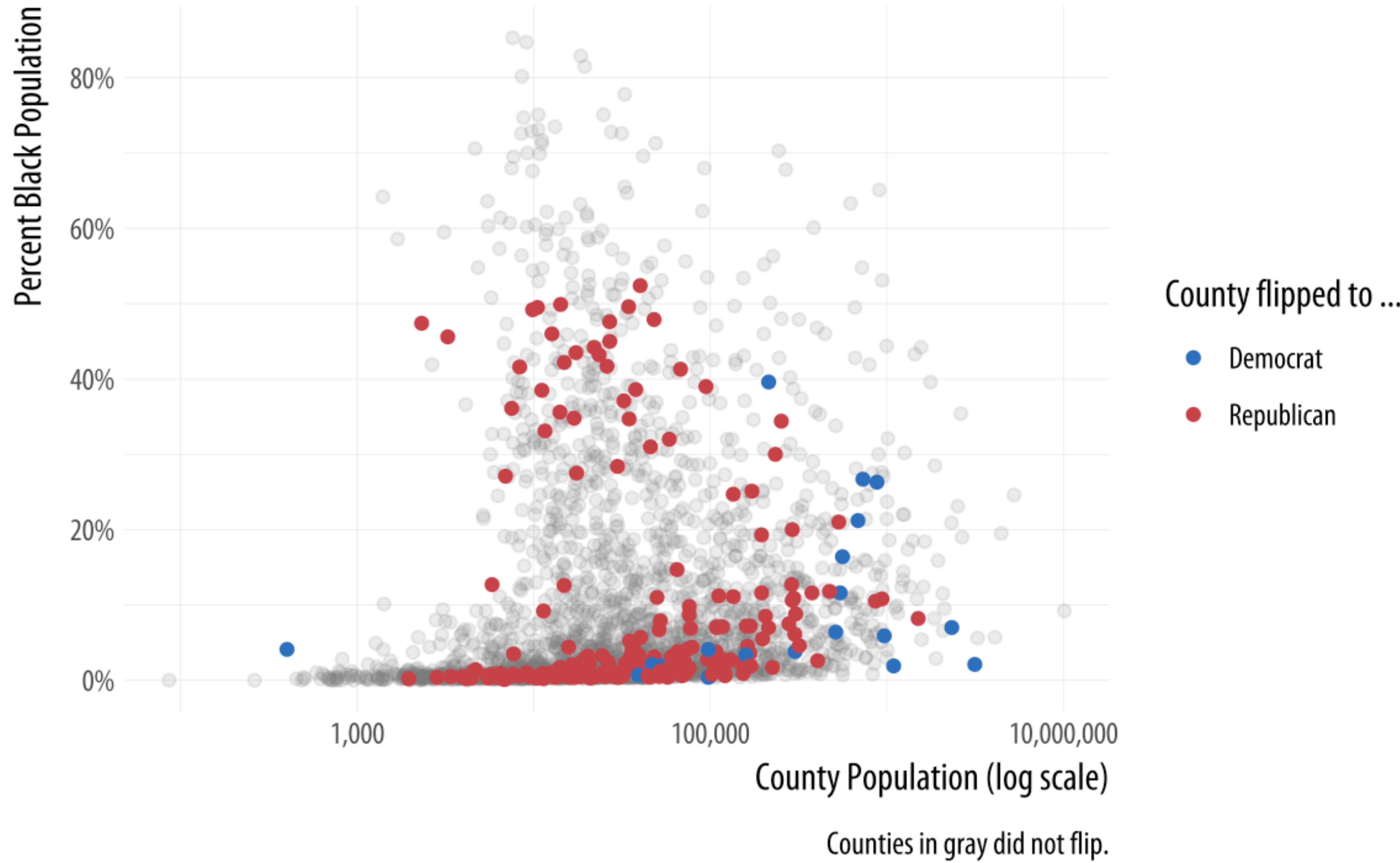
```
p2
```



```
p3 <- p2 + scale_y_continuous(labels=scales::percent) +  
  labs(color = "County flipped to ... ",  
        x = "County Population (log scale)",  
        y = "Percent Black Population",  
        title = "Flipped counties, 2016",  
        caption = "Counties in gray did not flip.")
```

p3

Flipped counties, 2016

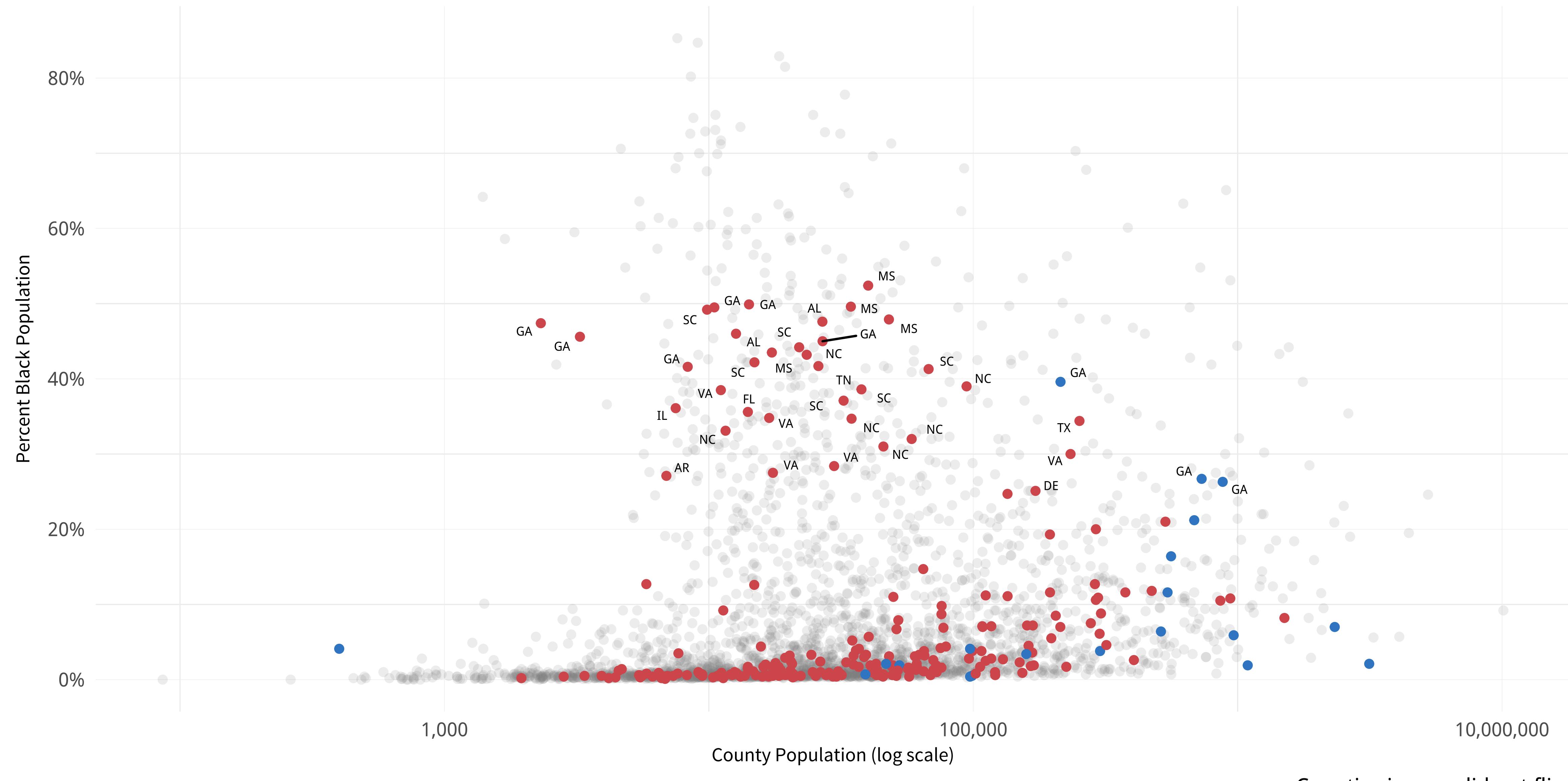


```
p4 <- p3 + geom_text_repel(data = subset(county_data,
                                         flipped == "Yes" &
                                         black > 25),
                           mapping = aes(x = pop,
                                         y = black/100,
                                         label = state), size = 2)
```

```
p4 + theme_minimal() +
  theme(legend.position="top")
```

Flipped counties, 2016

County flipped to ... ● Democrat ● Republican



THEMES

```
theme_set(theme_bw())
p4 + theme(legend.position="top")
```

```
theme_set(theme_dark())
p4 + theme(legend.position="top")
```

Built-in Themes can be
added per plot or globally set

```
install.packages("ggthemes")
```

```
library(ggthemes)
```

```
theme_set(theme_fivethirtyeight())
```

p4

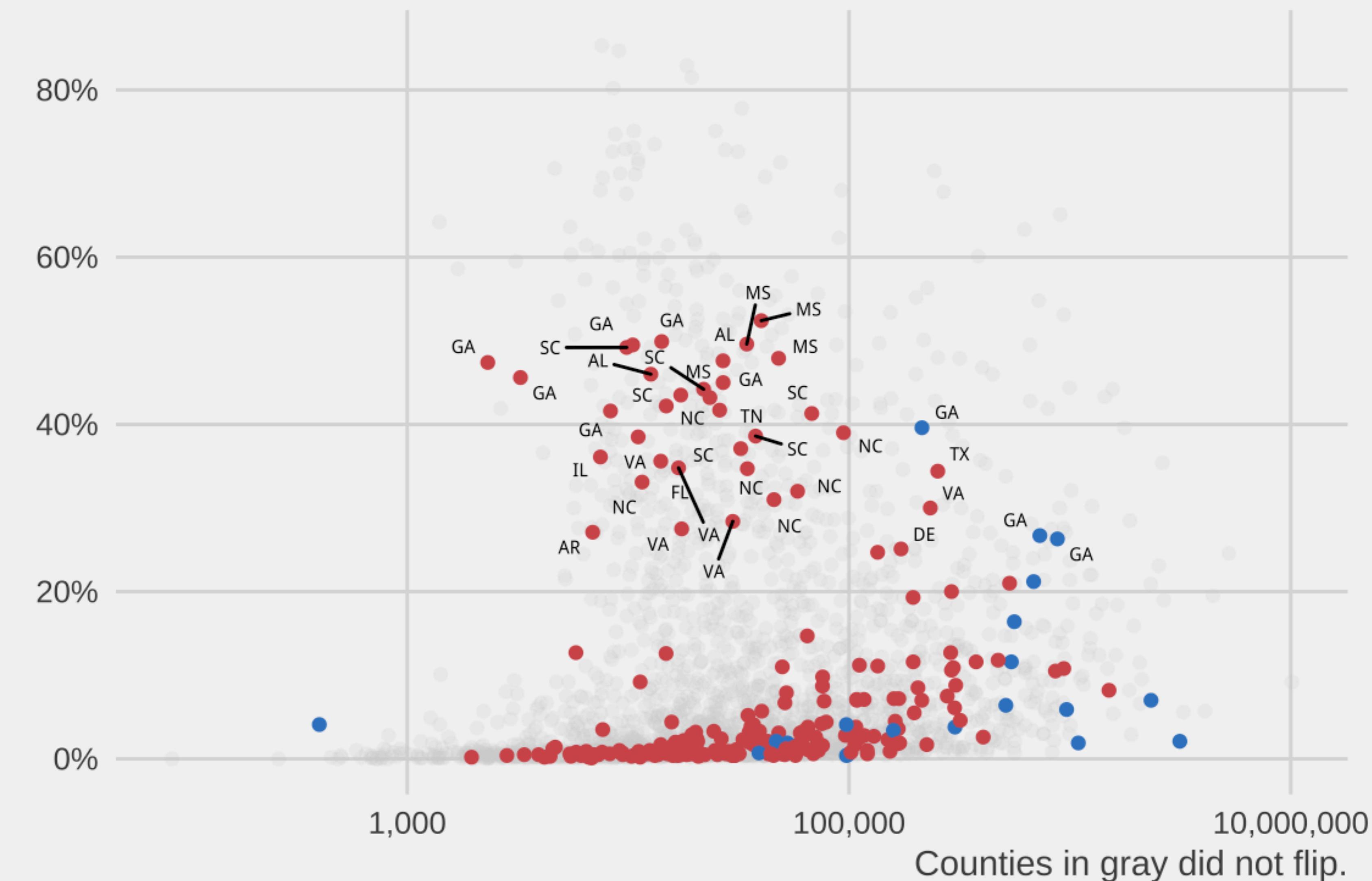
```
theme_set(theme_economist())
```

p4

The ggthemes library provides several additional alternatives

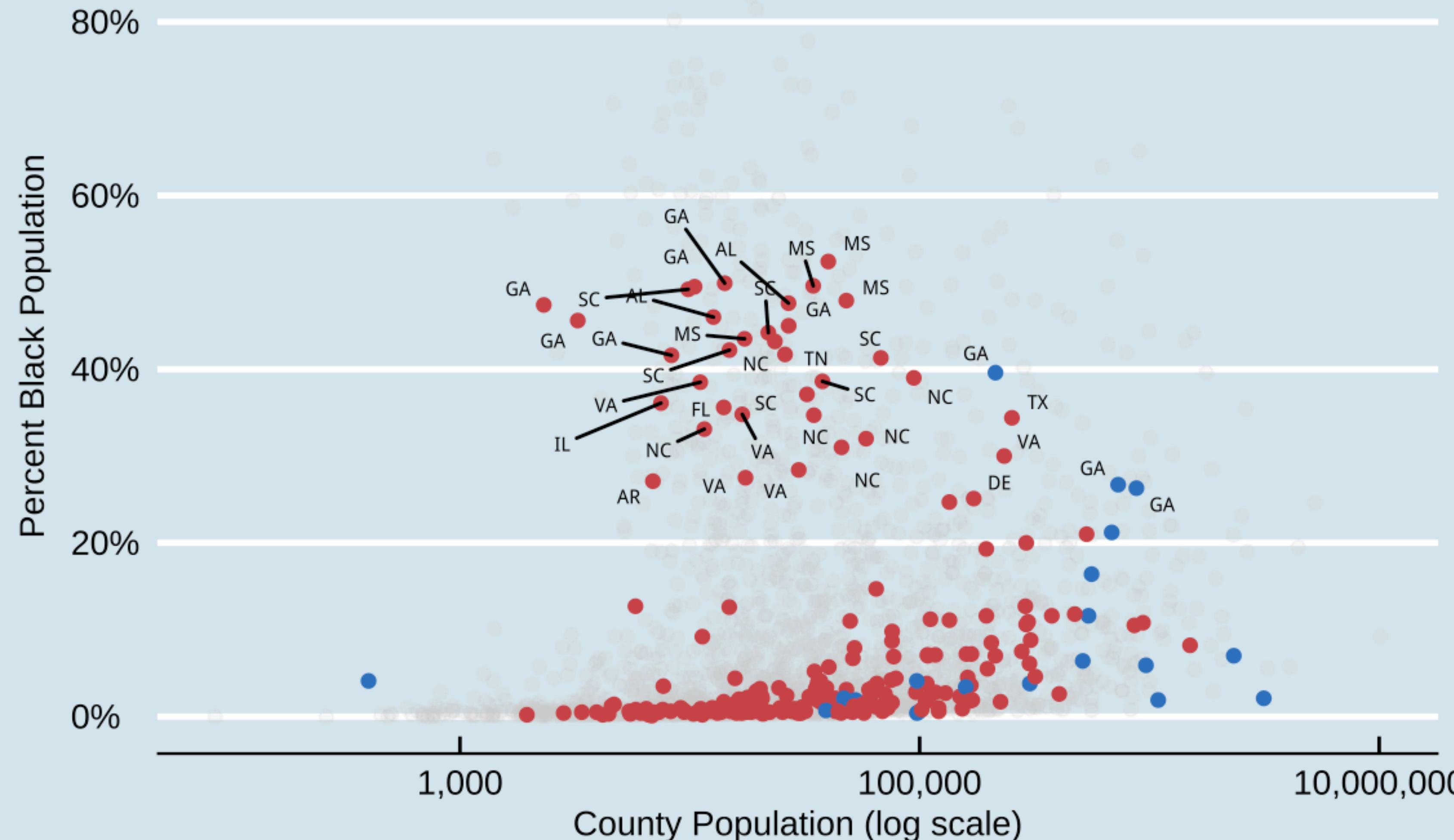
Flipped counties, 2016

County flipped to ... • Democrat • Republican



Flipped counties, 2016

County flipped to ... • Democrat • Republican



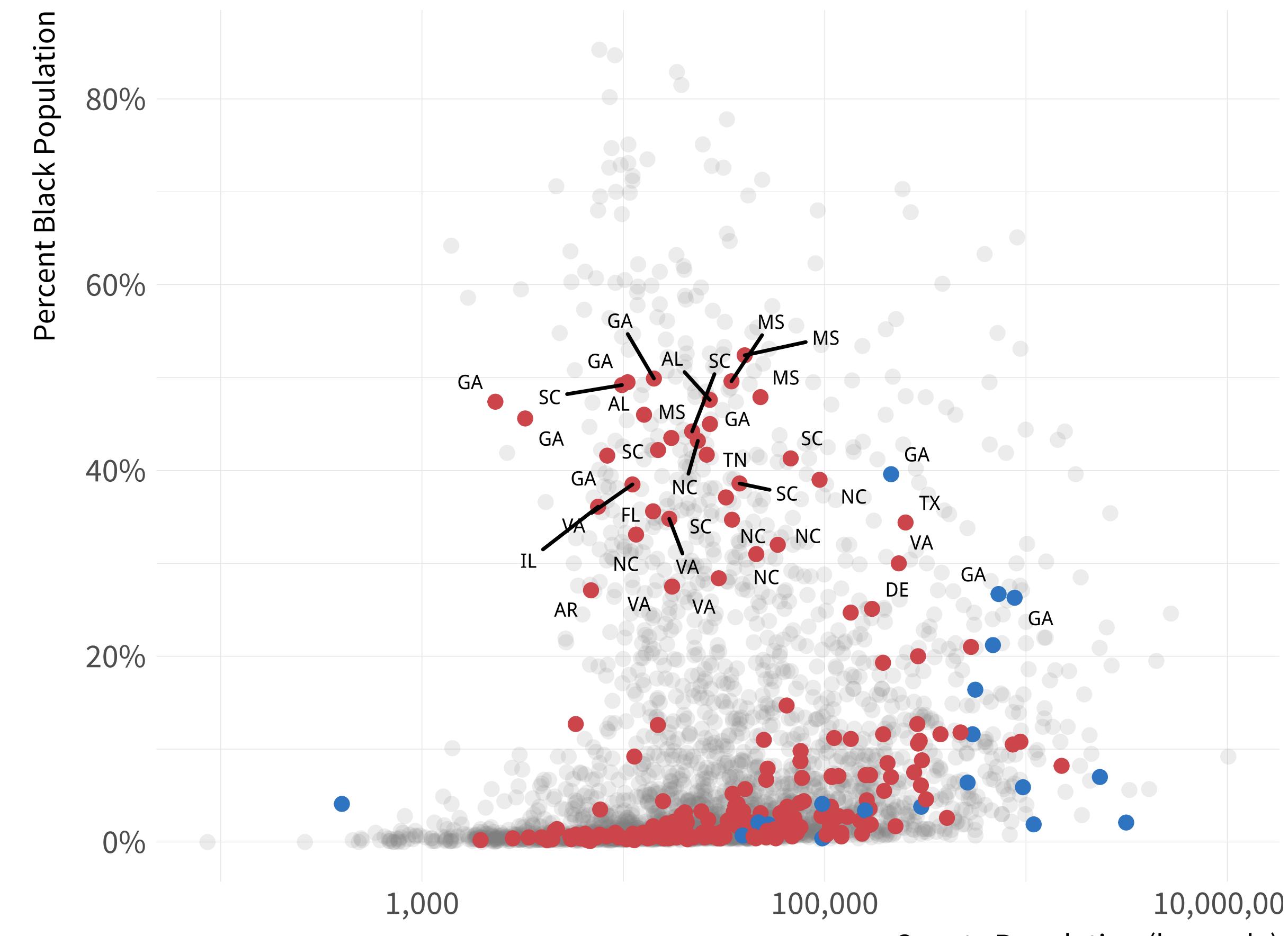
Counties in gray did not flip

```
p4 + theme(legend.position = "top")
```

```
p4 + theme(legend.position = "top",
plot.title = element_text(size=rel(2),
lineheight=.5,
family="Times",
face="bold.italic",
colour="orange"),
axis.text.x = element_text(size=rel(1.1),
family="Courier",
face="bold",
color="purple"))
```

Flipped counties, 2016

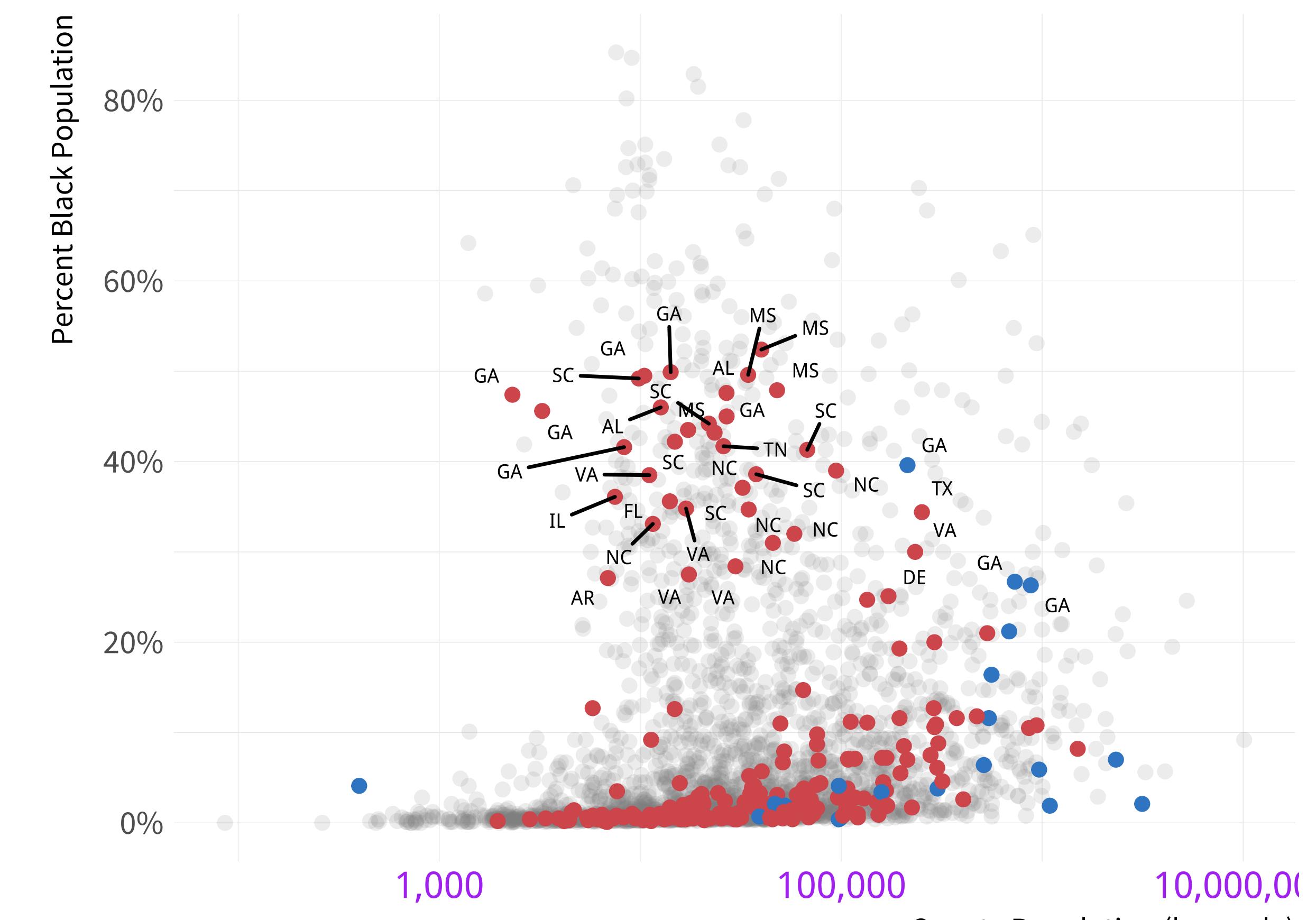
County flipped to ... • Democrat • Republican



Counties in gray did not flip.

Flipped counties, 2016

County flipped to ... • Democrat • Republican



Counties in gray did not flip.

```
install.packages("cowplot")
```

```
install.packages("hrbrthemes")
```

**Other theming to check out:
cowplot and hrbrthemes**

**Use Thematic Elements
in a Substantive Way**

```
p <- ggplot(subset(gss_lon, year %in% yrs), aes(x = age))

p1 <- p + geom_density(fill = "gray20", color = FALSE,
                       alpha = 0.9, mapping = aes(y = ..scaled..)) +
  geom_vline(data = subset(mean_age, year %in% yrs),
             aes(xintercept = xbar), color = "white", size = 0.5) +
  geom_text(data = subset(mean_age, year %in% yrs),
            aes(x = xbar, y = y, label = xbar), nudge_x = 6,
            color = "white", size = 4, hjust = 1) +
  geom_text(data = subset(yr_labs, year %in% yrs),
            aes(x = x, y = y, label = year)) +
  facet_grid(year ~ ., switch = "y")
```

```
yrs <- c(seq(1972, 1988, 4), 1993, seq(1996, 2016, 4))

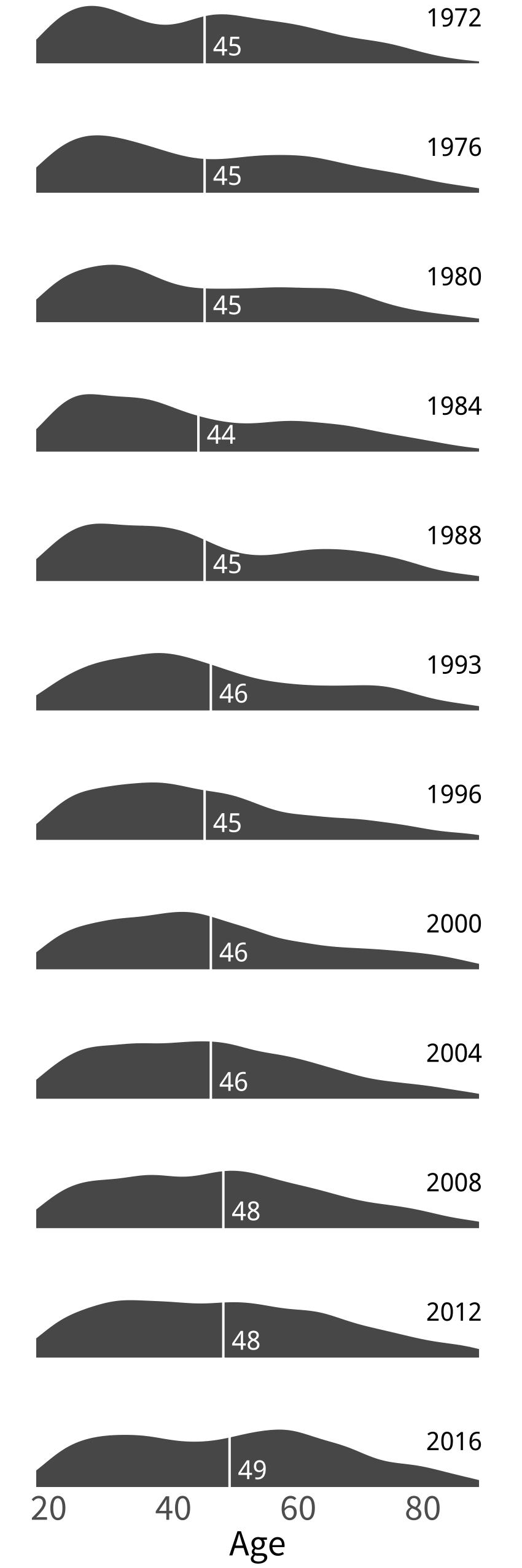
mean_age <- gss_lon %>%
  filter(age %in% NA && year %in% yrs) %>%
  group_by(year) %>%
  summarize(xbar = round(mean(age, na.rm = TRUE), 0))
mean_age$y <- 0.3

yr_labs <- data.frame(x = 85, y = 0.8,
                       year = yrs)
```

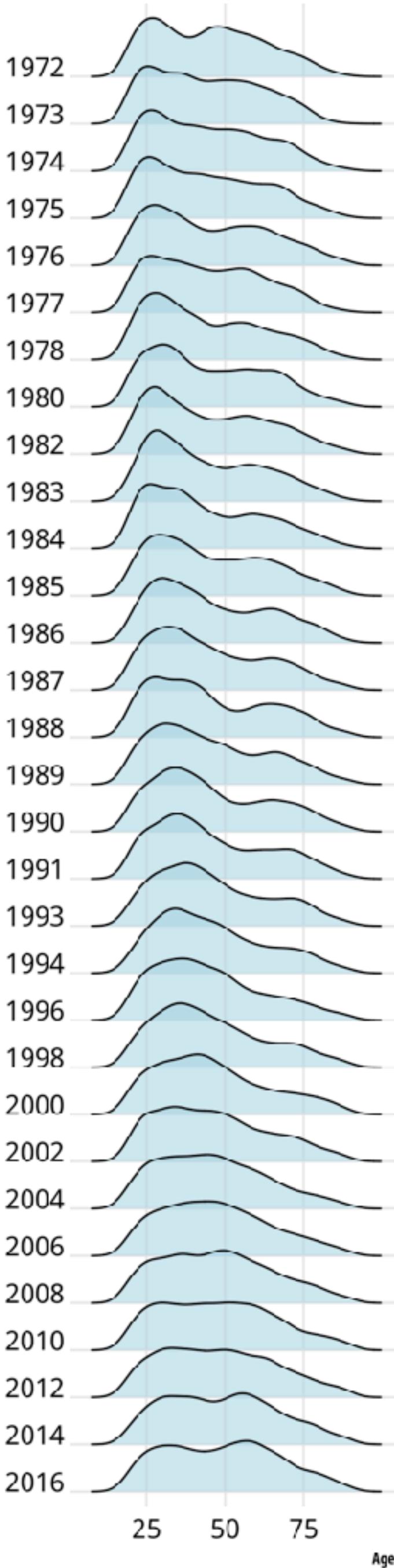
```
library(hrbrthemes)

p1 + theme_ipsum(base_size = 10, plot_title_size = 10,
                  strip_text_size = 32, panel_spacing = unit(0.1, "lines")) +
  theme(plot.title = element_text(size = 16),
        axis.text.x= element_text(size = 16),
        axis.title.x= element_text(size = 16, hjust = 0.5),
        axis.title.y=element_blank(),
        axis.text.y=element_blank(),
        axis.ticks.y = element_blank(),
        strip.background = element_blank(),
        strip.text.y = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank()) +
  labs(x = "Age",
       y = NULL,
       title = "Age Distribution of\nnGSS Respondents")
```

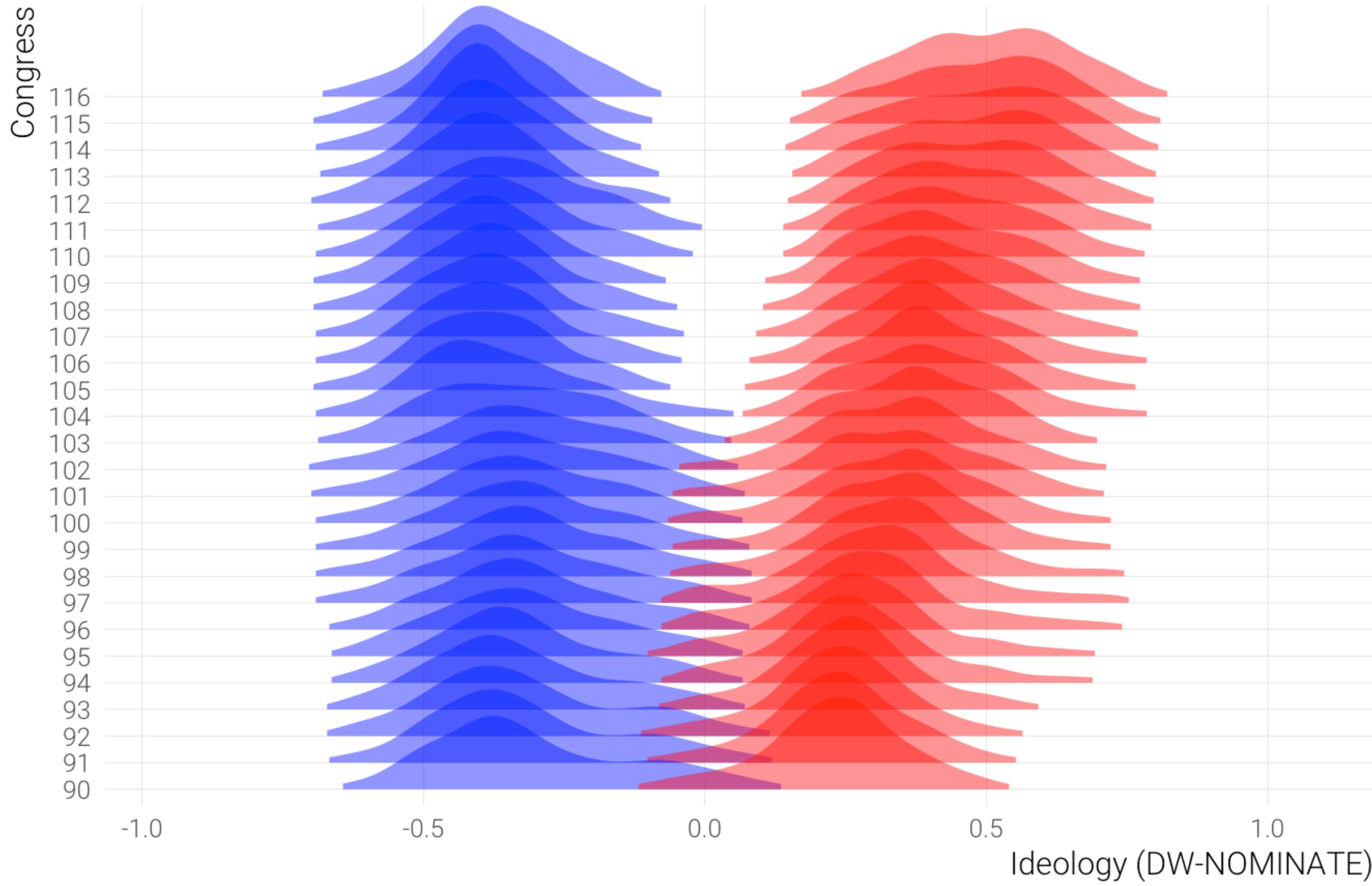
Age Distribution of GSS Respondents



Age Distribution of GSS Respondents

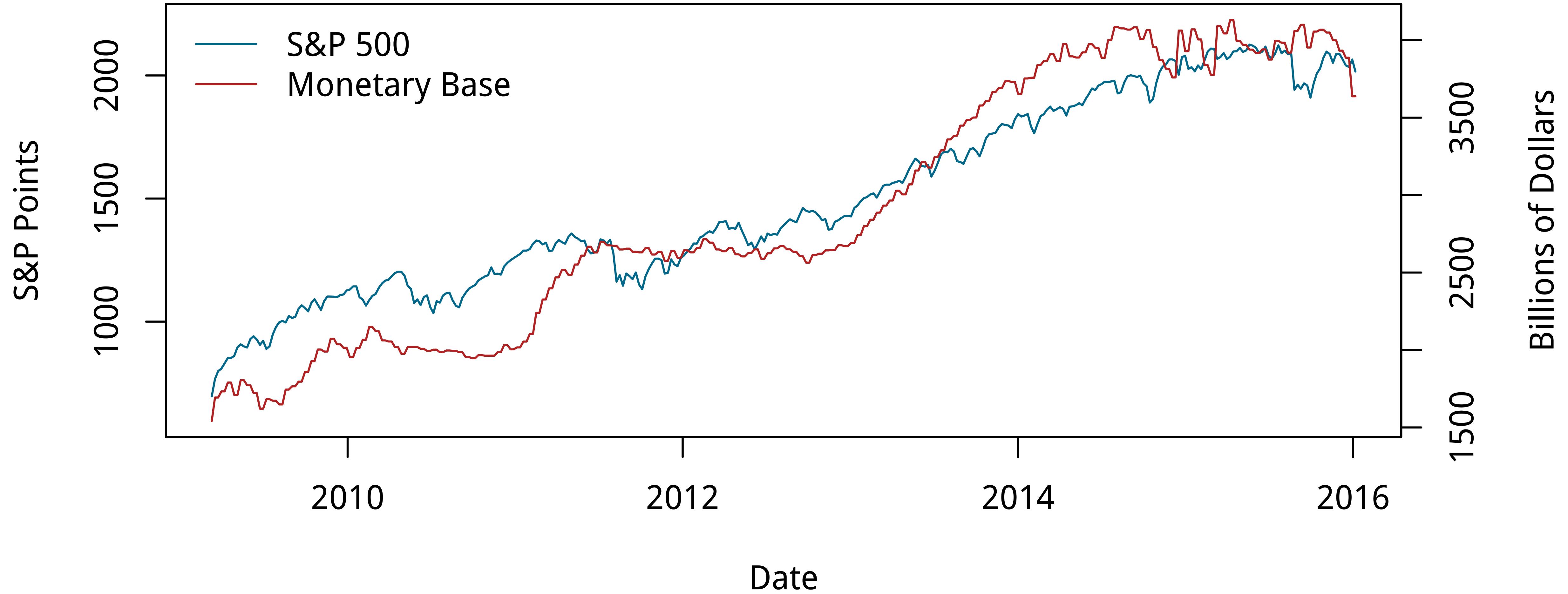


Polarization Between House Members, 1960-2018

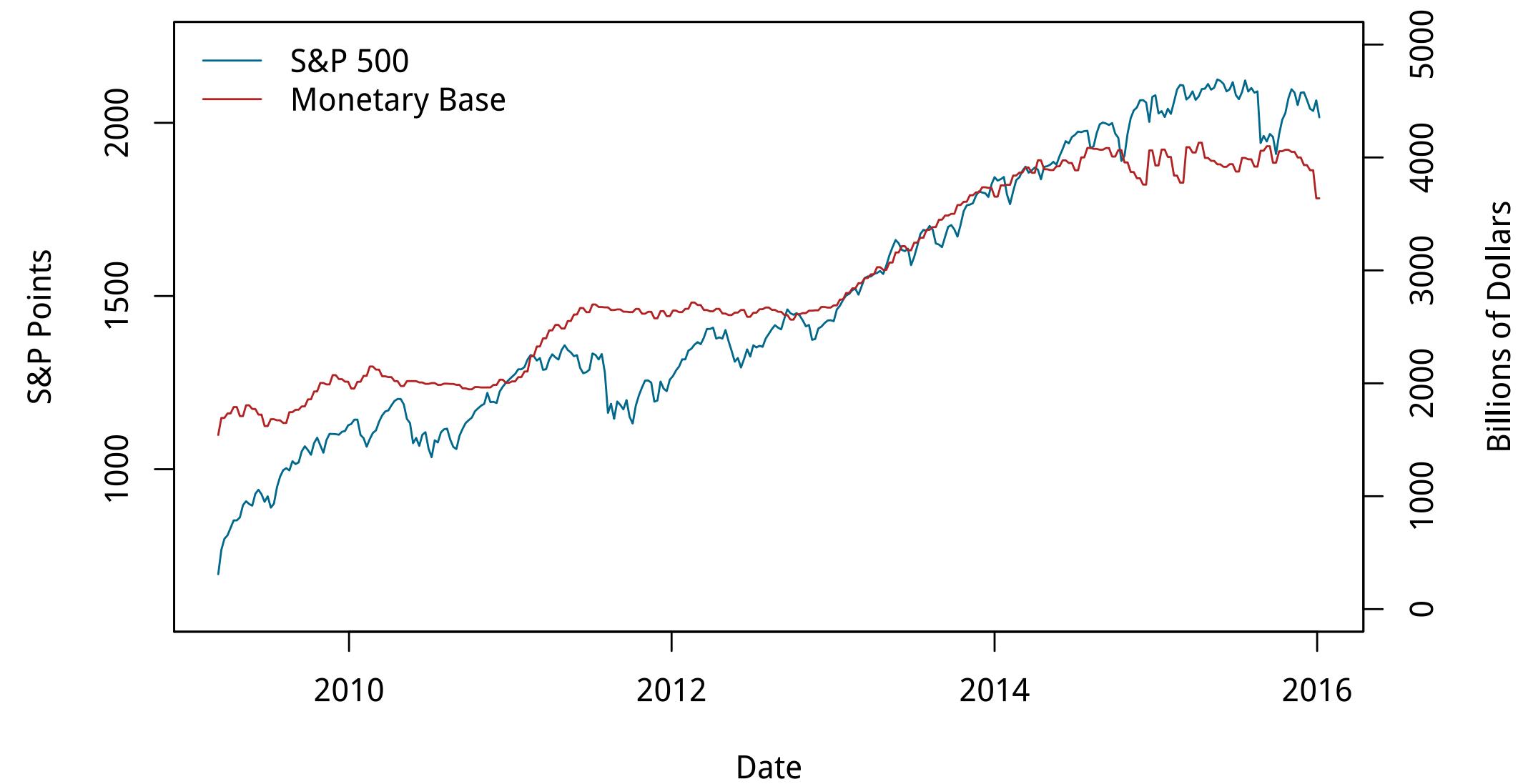


CASE STUDIES

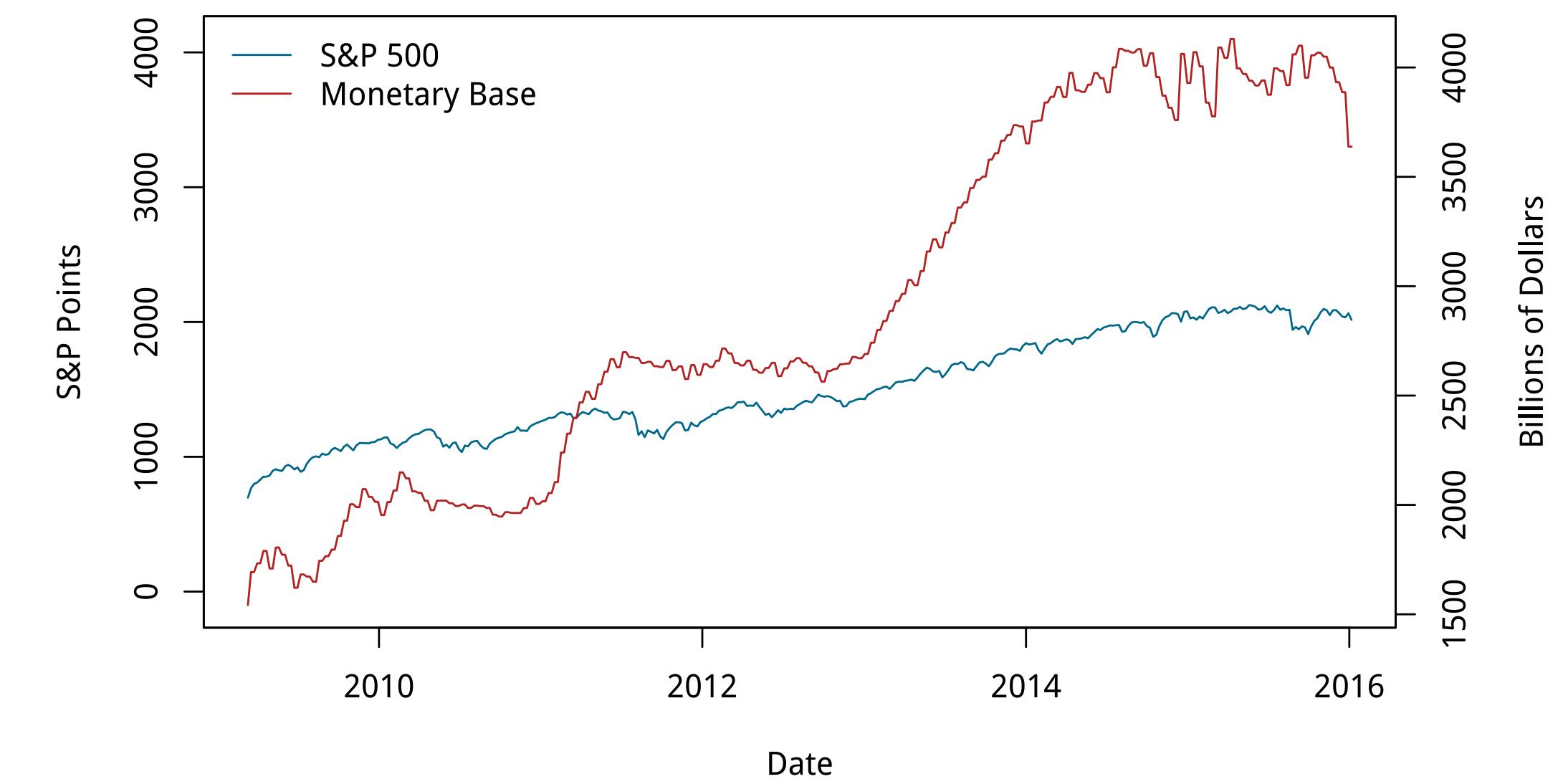
Two Y-Axes



Start y2 at Zero



Start y1 at Zero; Max both at Max y2



```
## Tidy the data with 'gather()'
fredts_m <- fredts %>% select(date, sp500_i, monbase_i) %>%
  gather(key = series, value = score, sp500_i:monbase_i)

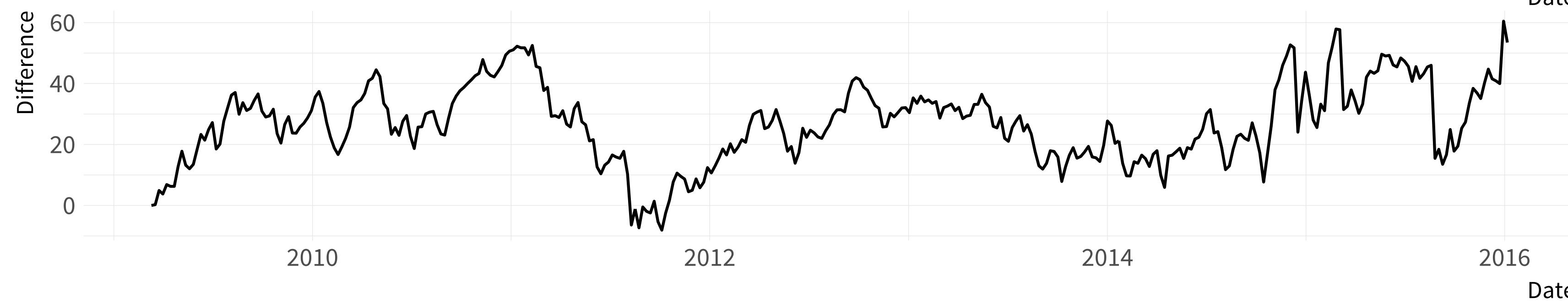
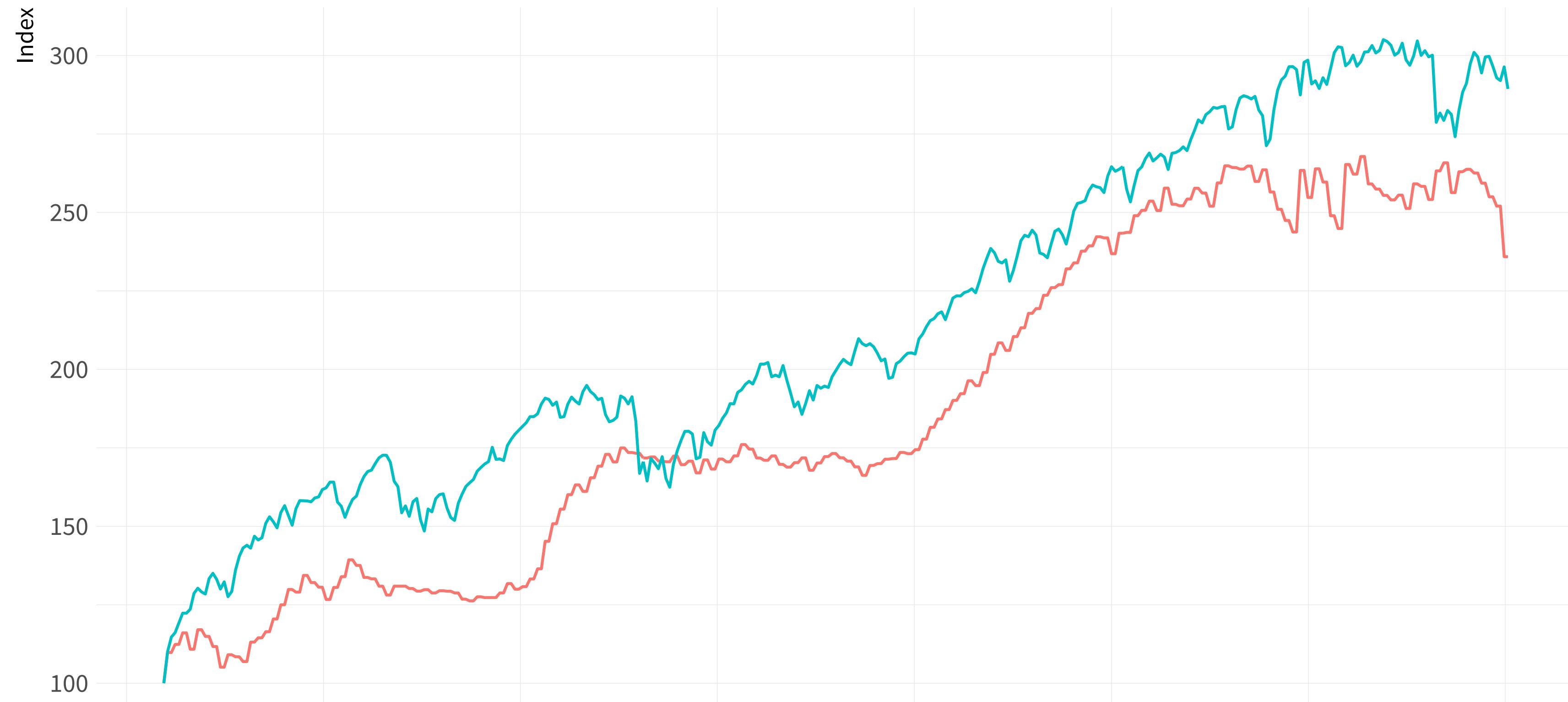
p <- ggplot(data = fredts_m,
             mapping = aes(x = date, y = score,
                             group = series,
                             color = series))
p1 <- p + geom_line() + theme(legend.position = "top") +
  labs(x = "Date",
        y = "Index",
        color = "Series")

p <- ggplot(data = fredts,
             mapping = aes(x = date, y = sp500_i - monbase_i))

p2 <- p + geom_line() +
  labs(x = "Date",
        y = "Difference")
```

```
library(gridExtra)
grid.arrange(p1, p2, nrow = 2, heights = c(0.75, 0.25))
```

Series — monbase_i — sp500_i



Multiple Plots

```
drat::addRepo("kjhealy")
install.packages("demog")
library(demog)
```

demog

Birth and Death rate data

```
... -->
> okboomer
# A tibble: 1,644 x 12
  year month n_days births total_pop births_pct births_pct_day date
  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <date>
1 1938     1     31 51820 41215000 0.00126      40.6 1938-01-01
2 1938     2     28 47421 41215000 0.00115      41.1 1938-02-01
3 1938     3     31 54887 41215000 0.00133      43.0 1938-03-01
4 1938     4     30 54623 41215000 0.00133      44.2 1938-04-01
5 1938     5     31 56853 41215000 0.00138      44.5 1938-05-01
6 1938     6     30 53145 41215000 0.00129      43.0 1938-06-01
7 1938     7     31 53214 41215000 0.00129      41.6 1938-07-01
8 1938     8     31 50444 41215000 0.00122      39.5 1938-08-01
9 1938     9     30 50545 41215000 0.00123      40.9 1938-09-01
10 1938    10     31 50079 41215000 0.00122     39.2 1938-10-01
# ... with 1,634 more rows
```

```
library(patchwork)
```

Patchwork is a ludicrously convenient package

```
title_txt <- "Average births per month per million people in the United States, 1933-73"  
start_date <- "1935-01-01"  
end_date <- "1973-12-01"  
end_year <- 1974  
by_unit <- "year"  
bar_width <- 0.5  
p_col <- "gray30"
```

```
break_vec <- seq(from=as.Date(start_date), to=as.Date(end_date), by = by_unit)  
break_vec <- break_vec[seq(1, length(break_vec), 5)]
```

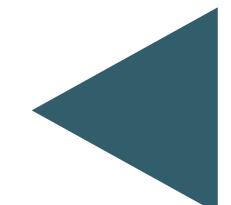
```
p <- ggplot(subset(okboomer, country == "United States" & year < end_year),  
            mapping = aes(x = date, y = births_pct_day))  
p1 <- p + geom_line(color = p_col) + ylab("Data") + xlab("") +  
      scale_x_date(breaks = break_vec) +  
      theme(axis.text.x = element_blank(),  
            axis.title.y = element_text(size = rel(0.8)),  
            plot.title = element_text(size = rel(1))) +  
      ggtitle(title_txt)
```

```
p <- ggplot(subset(okboomer, country == "United States" & year < end_year),  
            mapping = aes(x = date, y = trend))  
p2 <- p + geom_line(color = p_col) + ylab("Trend") + xlab("") +  
      scale_x_date(breaks = break_vec) +  
      theme(axis.text.x = element_blank(),  
            axis.title.y = element_text(size = rel(0.8)))
```

```
p <- ggplot(subset(okboomer, country == "United States" & year < end_year),  
            mapping = aes(x = date, y = seasonal))  
p3 <- p + geom_line(color = p_col) + ylab("Seasonal") + xlab("") +  
      scale_x_date(breaks = break_vec) +  
      theme(axis.text.x = element_blank(),  
            axis.title.y = element_text(size = rel(0.8)))
```

```
p <- ggplot(subset(okboomer, country == "United States" & year < end_year),  
            mapping = aes(x = date, ymax = remainder, ymin = 0))  
p4 <- p + geom_linerange(size = bar_width, color = p_col) +  
      scale_x_date(breaks = break_vec, date_labels = "%Y") +  
      ylab("Remainder") + xlab(by_unit) +  
      theme(axis.title.y = element_text(size = rel(0.8)))
```

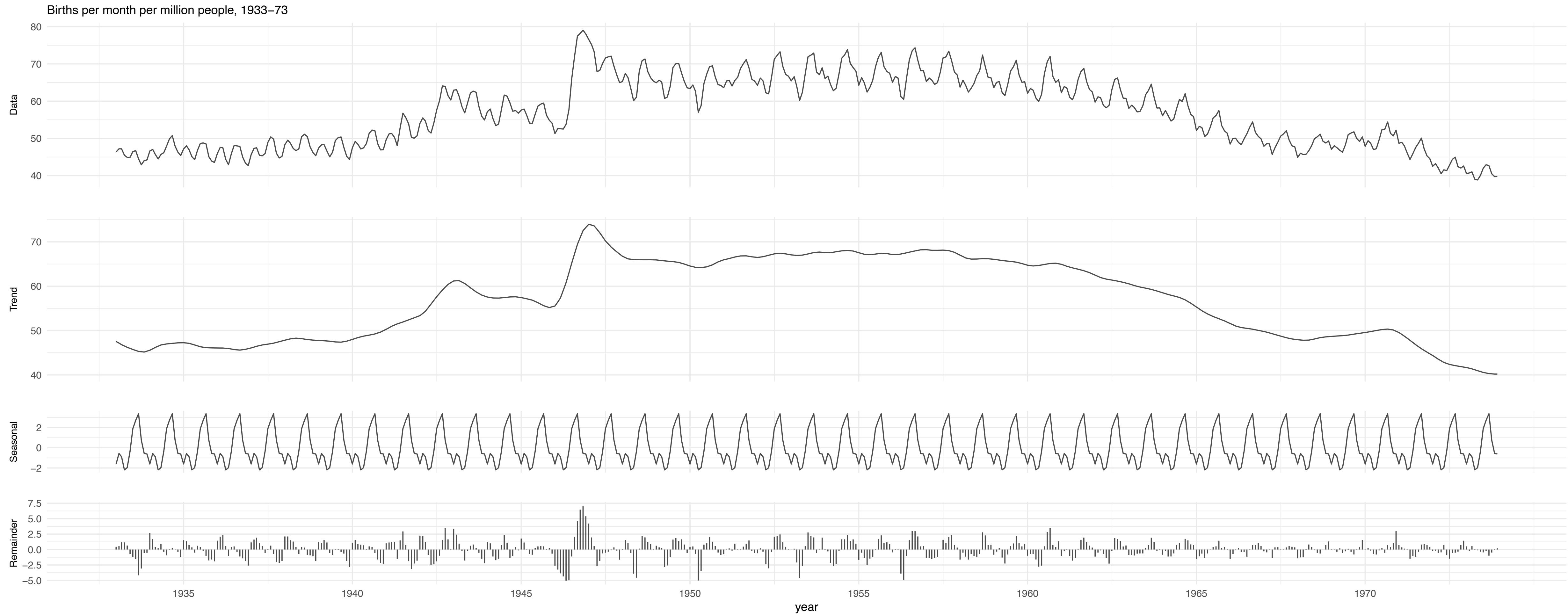
```
patch_plot <- (p1 / p2 / p3 / p4)
```



Vertically stack the plots! Automatically aligned!

```
patch_plot + plot_layout(heights = c(2, 2, 0.75, 1))
```

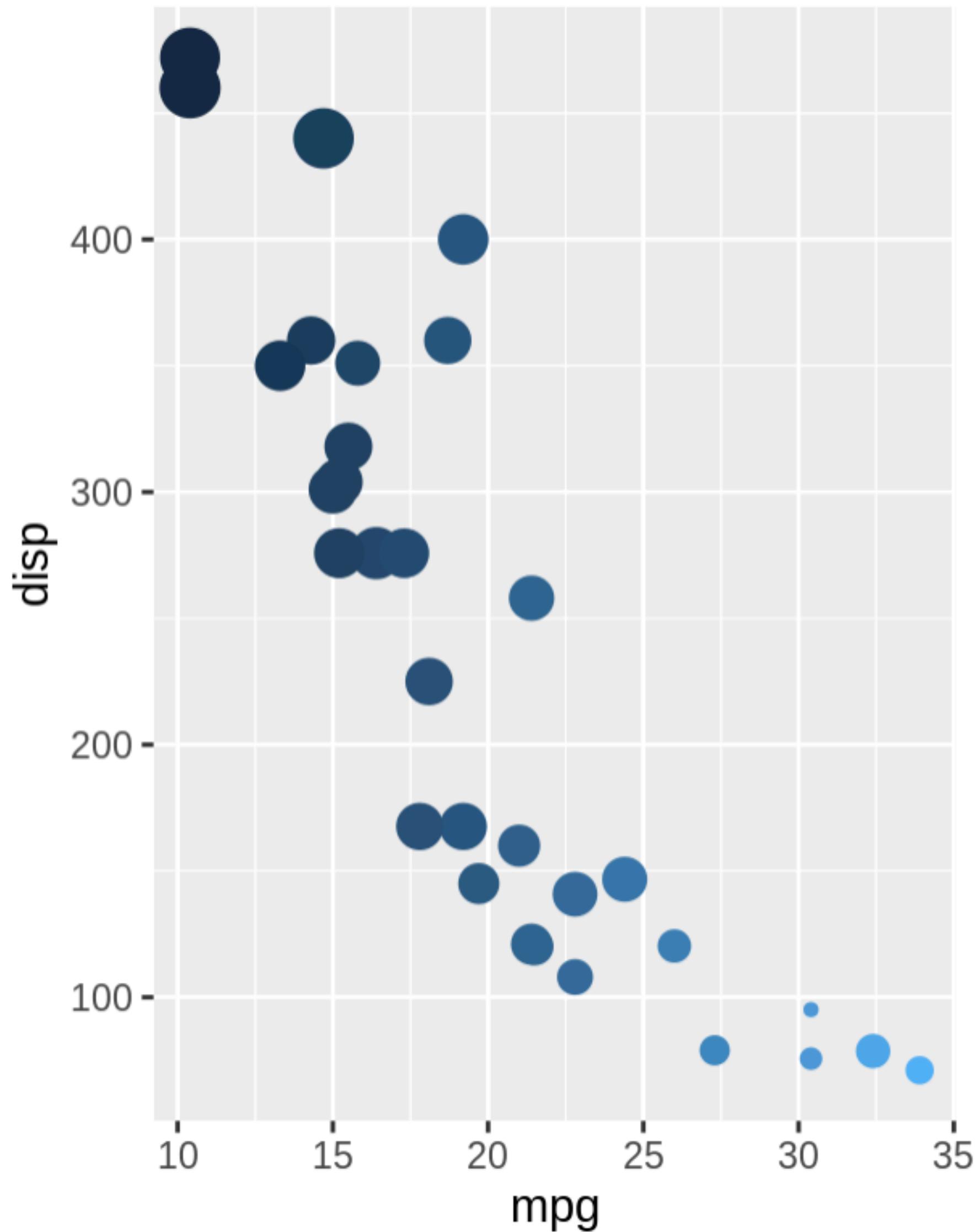
◀ Specify relative heights



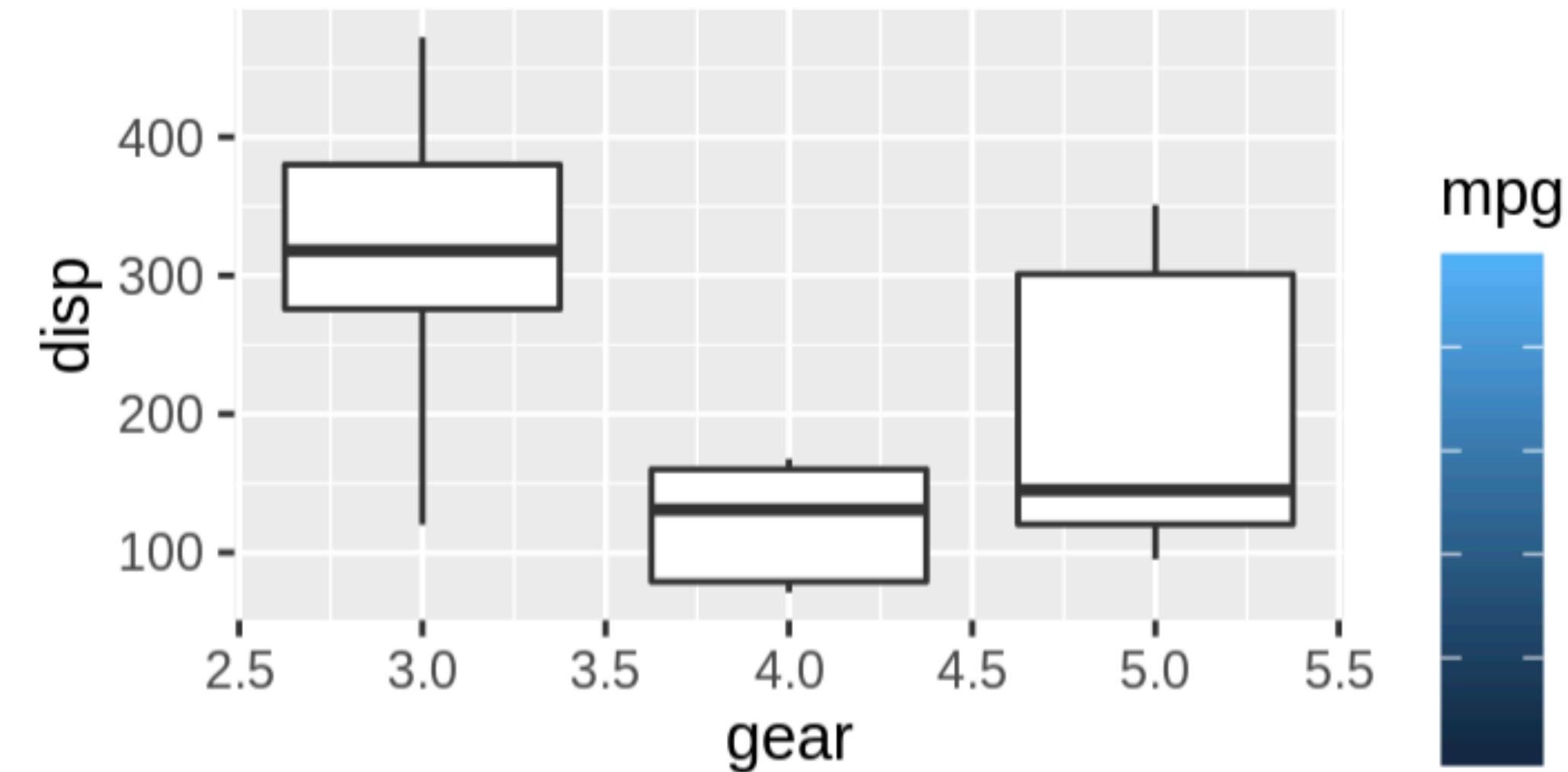
```
(p1a | (p2 / p3)) + plot_layout(guides = 'collect')
```

I told you it was ludicrously useful

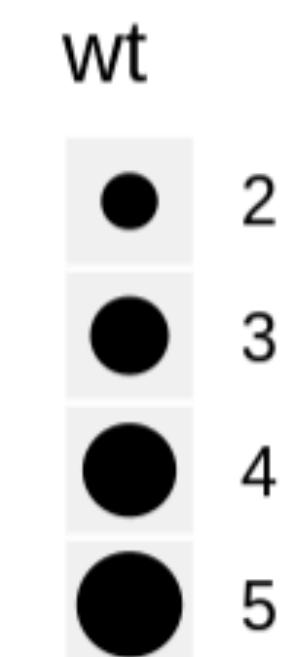
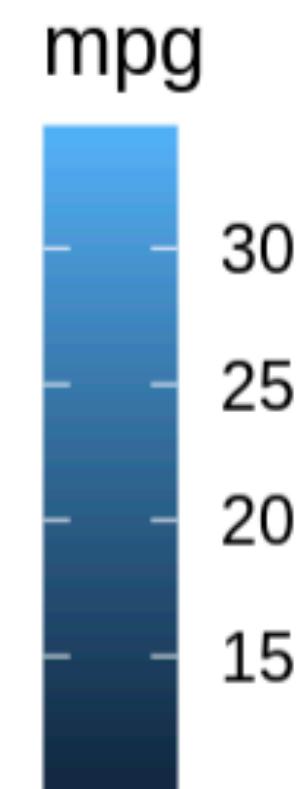
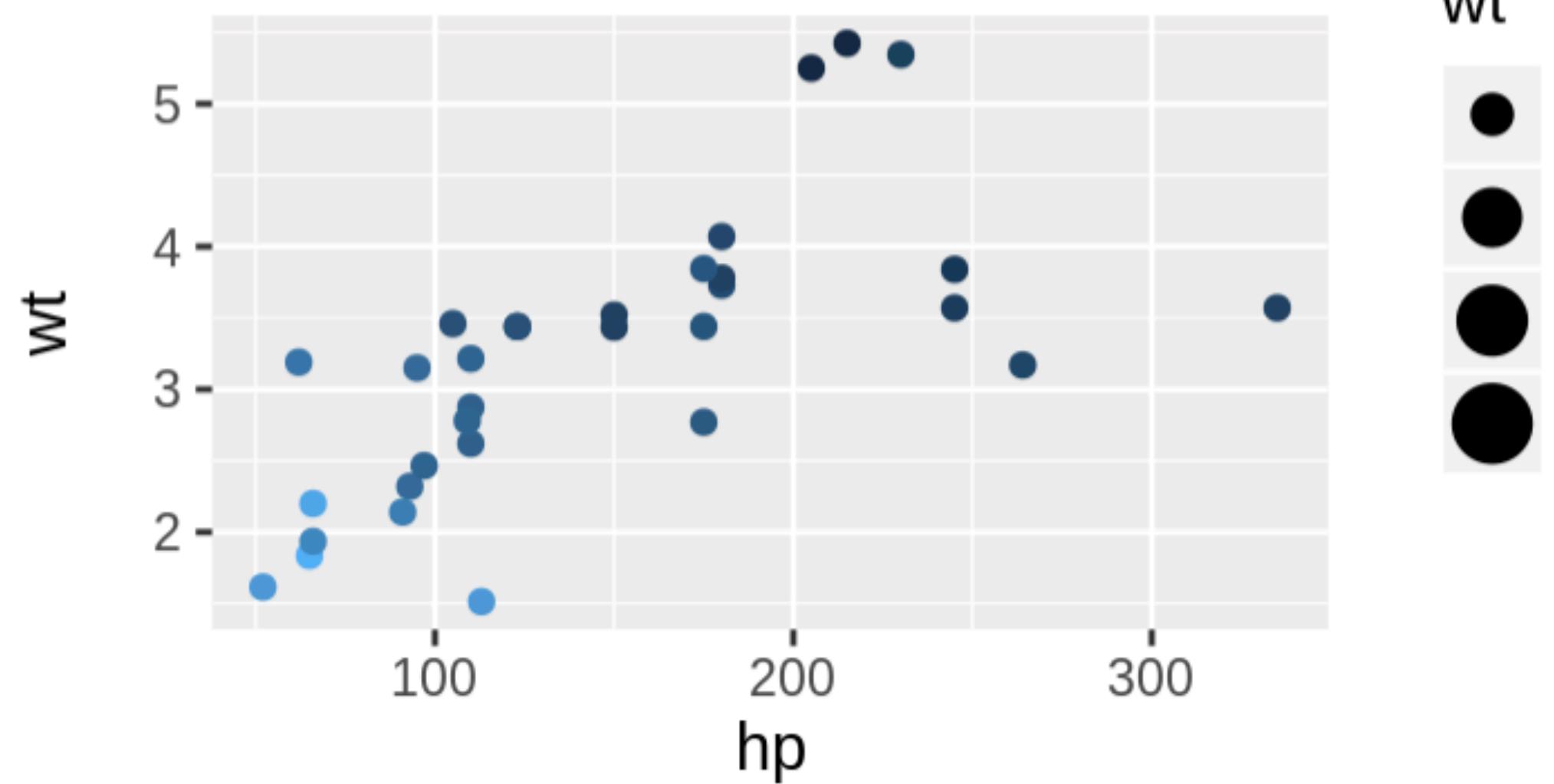
Plot 1a



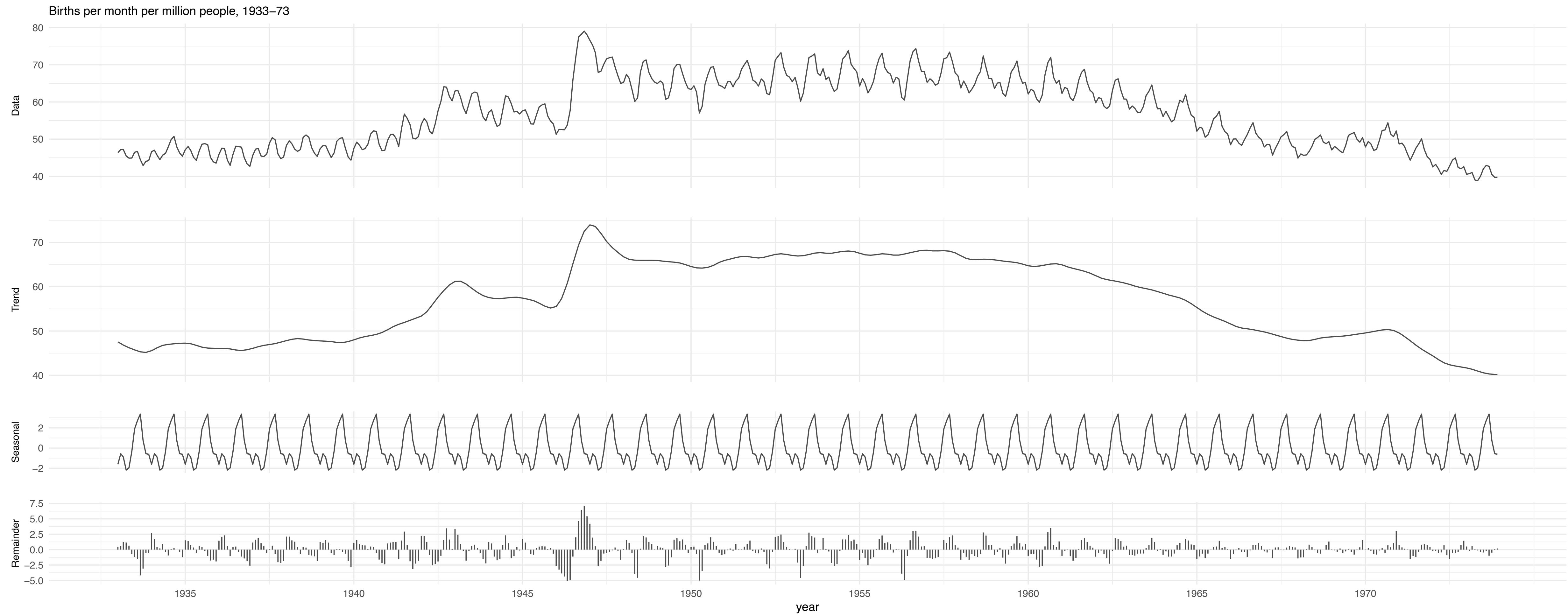
Plot 2



Plot 3



```
patch_plot + plot_layout(heights = c(2, 2, 0.75, 1))
```



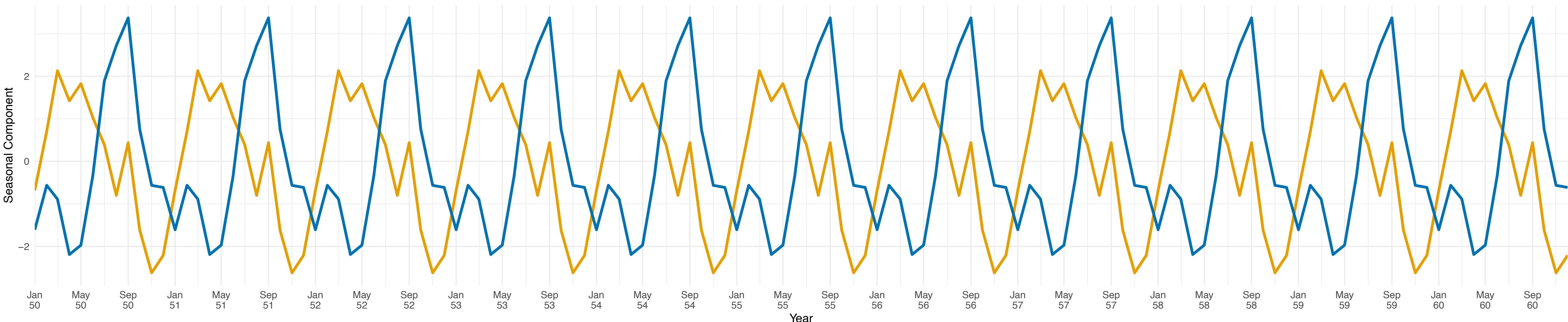
```
(p_out <- okboomer %>%
  filter(date > lubridate::ymd("1949-12-01"),
        date < lubridate::ymd("1961-01-01")) %>%
  ggplot(mapping = aes(x = date, y = seasonal, color = country)) +
  geom_line(size = 1.2) +
  scale_x_date(date_breaks = "4 months",
               date_labels = "%b\n%y", expand = expand_scale()) +
  scale_color_manual(values = my_colors("bly")) +
  theme(legend.position = "top") +
  labs(x = "Year",
       y = "Seasonal Component",
       color = "Country",
       title = "Comparative Seasonality of Births in the US and England/Wales, 1950-1960")
)
```

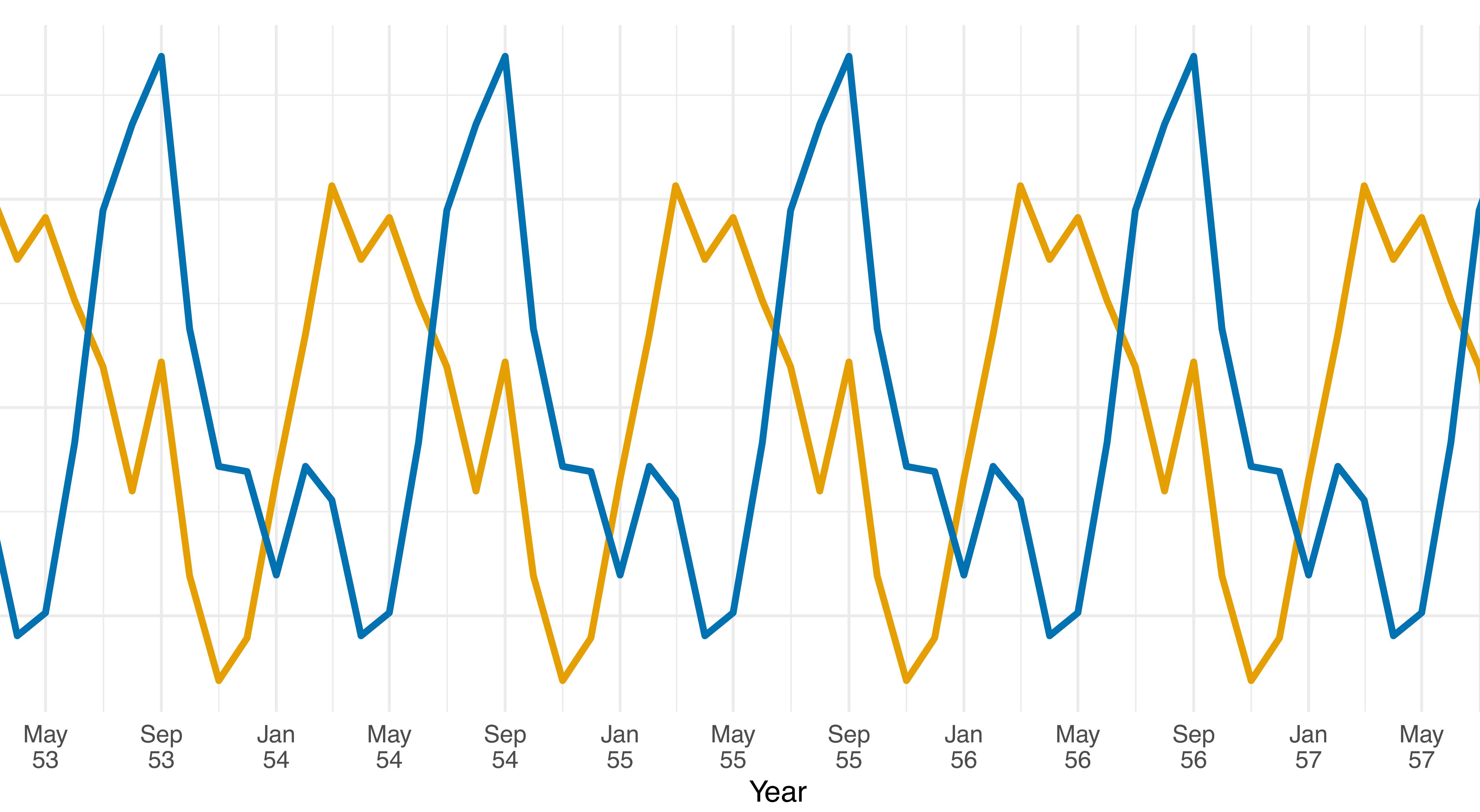
Date coercion with lubridate

Breaks and labels

Comparative Seasonality of Births in the US and England/Wales, 1950–1960

Country — England and Wales — United States

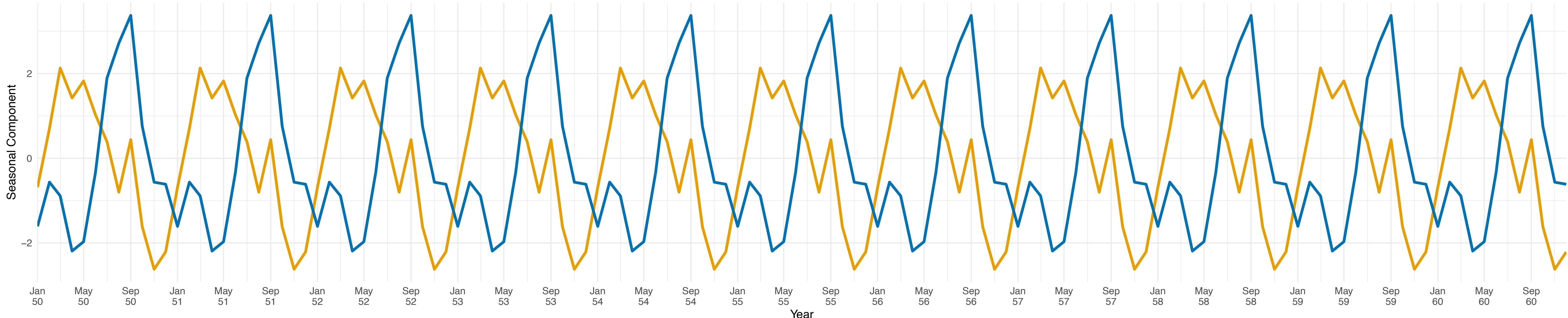




```
(p_out <- okboomer %>%
  filter(date > lubridate::ymd("1949-12-01"),
         date < lubridate::ymd("1961-01-01")) %>%
  ggplot(mapping = aes(x = date, y = seasonal, color = country)) +
  geom_line(size = 1.2) +
  scale_x_date(date_breaks = "4 months",
               date_labels = "%b\n%y", expand = expand_scale()) +
  scale_color_manual(values = my_colors("bly")) +
  theme(legend.position = "top") +
  labs(x = "Year",
       y = "Seasonal Component",
       color = "Country",
       title = "Comparative Seasonality of Births in the US and England/Wales, 1950-1960")
)
```

Comparative Seasonality of Births in the US and England/Wales, 1950–1960

Country — England and Wales — United States



%a Abbreviated weekday name in the current locale on this platform.

%A Full weekday name in the current locale.

%b Abbreviated month name in the current locale on this platform. **%B** Full month name in the current locale.

%c Date and time. Locale-specific on output, "%a %b %e %H:%M:%S %Y" on input.

%C Century (00-99): the integer part of the year divided by 100.

%d Day of the month as decimal number (01-31).

%D Date format such as "%m/%d/%y": the C99 standard says it should be that exact format (but not all OSes comply).

%e Day of the month as decimal number (1-31), with a leading space for a single-digit number.

%F Equivalent to "%Y-%m-%d" (the ISO 8601 date format).

%g The last two digits of the week-based year (see %V). (Accepted but ignored on input.)

%G The week-based year (see %V) as a decimal number. (Accepted but ignored on input.)

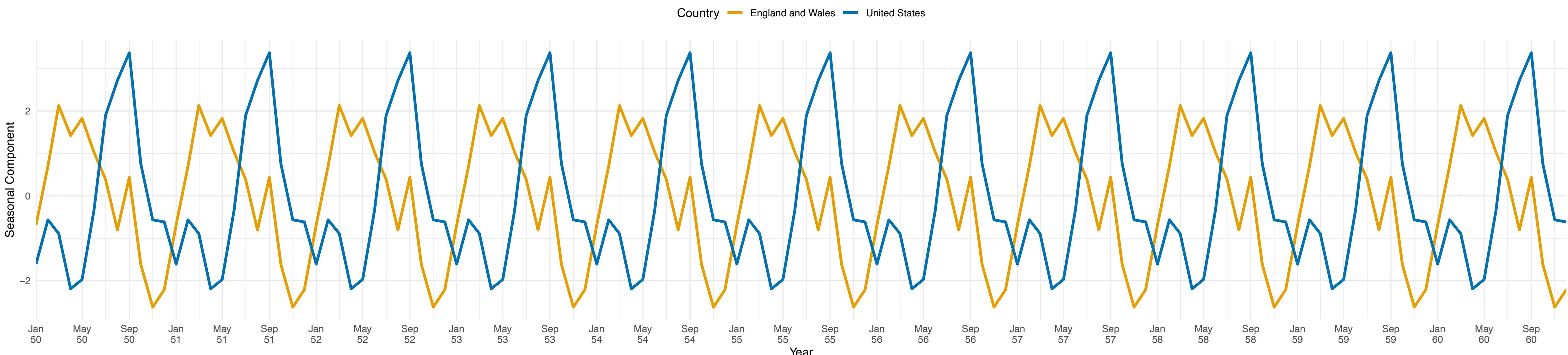
%h Equivalent to %b.

%H Hours as decimal number (00-23).

%I Hours as decimal number (01-12).

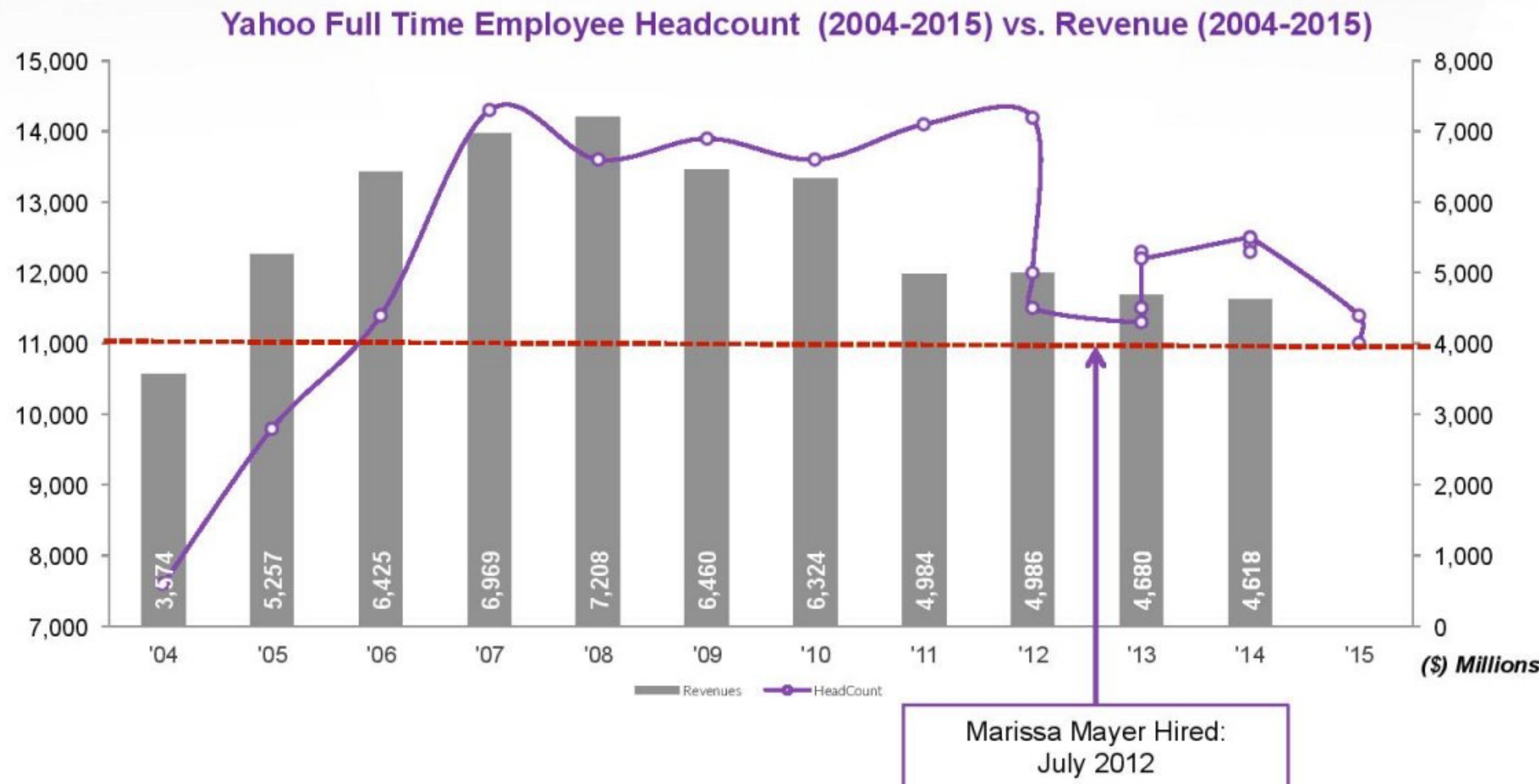
%j Day of year as decimal number (001-366).

Comparative Seasonality of Births in the US and England/Wales, 1950–1960



Redrawing a Bad Slide

Yahoo's Headcount Still Excessively High Given Revenues:



Source: Company Filings (10K), Analyst calls

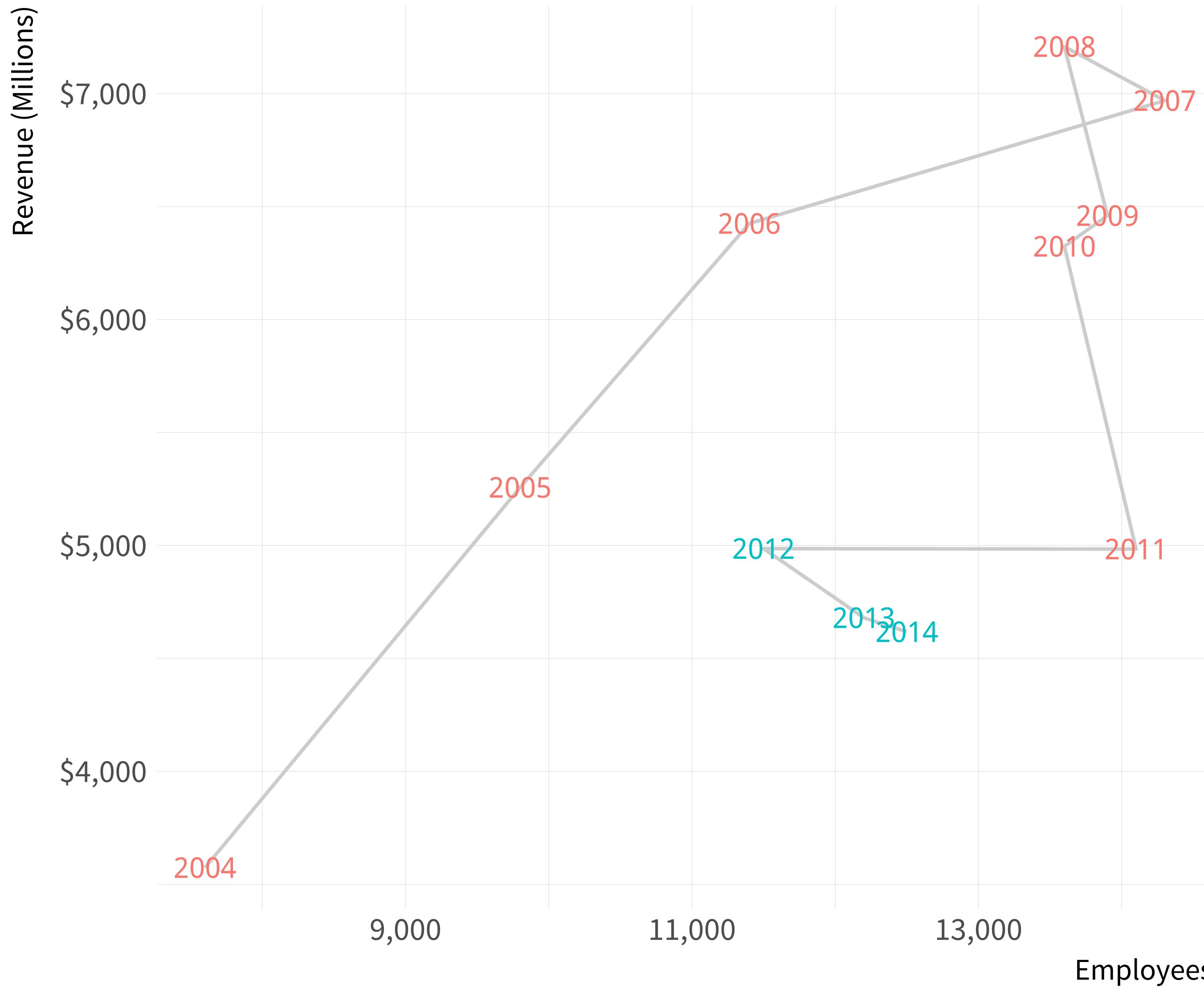
head(yahoo)

	##	Year	Revenue	Employees	Mayer
	## 1	2004	3574	7600	No
	## 2	2005	5257	9800	No
	## 3	2006	6425	11400	No
	## 4	2007	6969	14300	No
	## 5	2008	7208	13600	No
	## 6	2009	6460	13900	No

```
p <- ggplot(data = yahoo,
             mapping = aes(x = Employees, y = Revenue))

p + geom_path(color = "gray80") +
  geom_text(aes(color = Mayer, label = Year),
            size = 3, fontface = "bold") +
  theme(legend.position = "bottom") +
  labs(color = "Mayer is CEO",
       x = "Employees", y = "Revenue (Millions)",
       title = "Yahoo Employees vs Revenues, 2004-2014") +
  scale_y_continuous(labels = scales::dollar) +
  scale_x_continuous(labels = scales::comma)
```

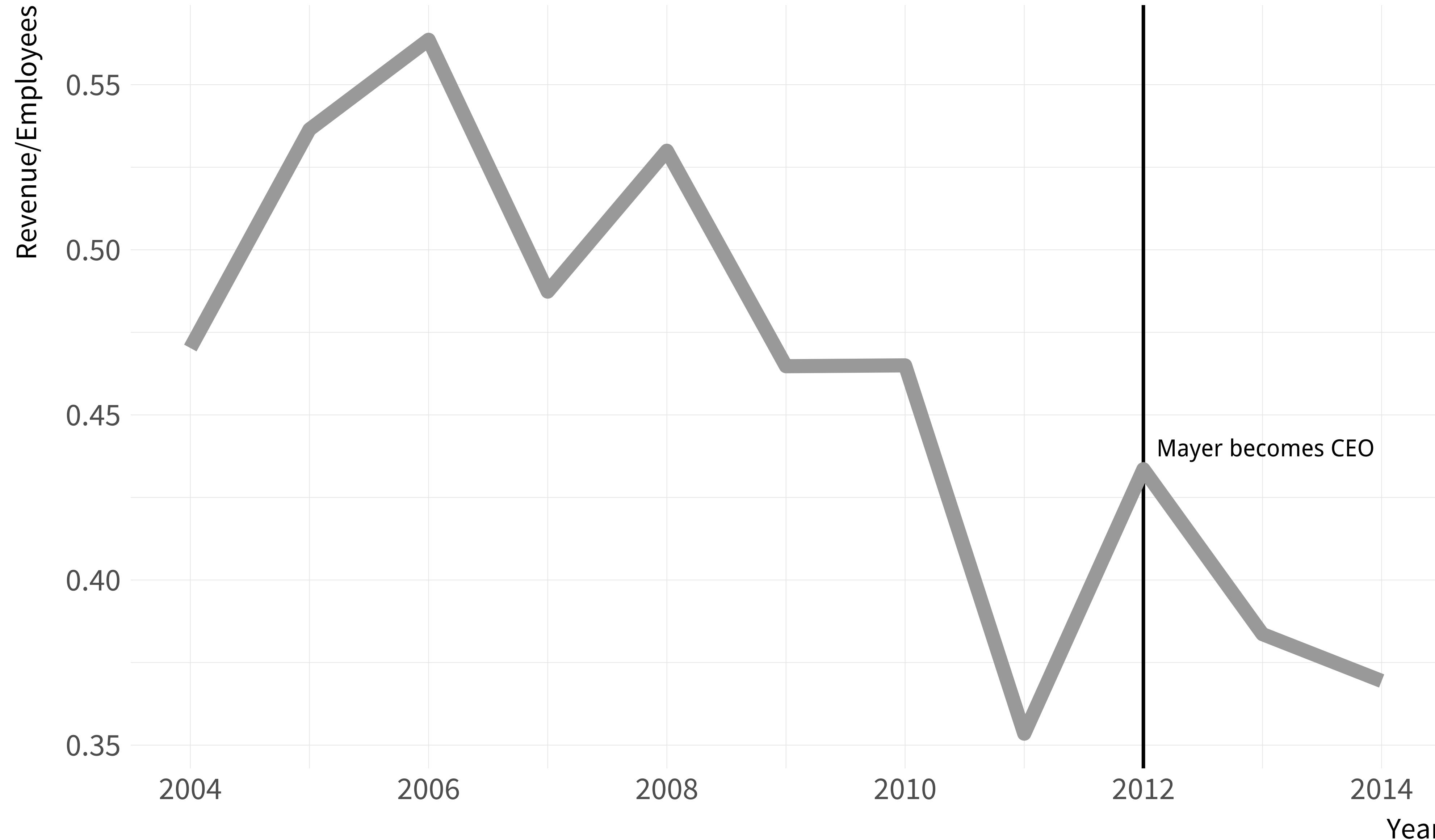
Yahoo Employees vs Revenues, 2004-2014



Mayer is CEO a No a Yes

```
p <- ggplot(data = yahoo,  
             mapping = aes(x = Year, y = Revenue/Employees))  
  
p + geom_vline(xintercept = 2012) +  
  geom_line(color = "gray60", size = 2) +  
  annotate("text", x = 2013, y = 0.44,  
           label = " Mayer becomes CEO", size = 2.5) +  
  labs(x = "Year\n",  
        y = "Revenue/Employees",  
        title = "Yahoo Revenue to Employee Ratio, 2004-2014")
```

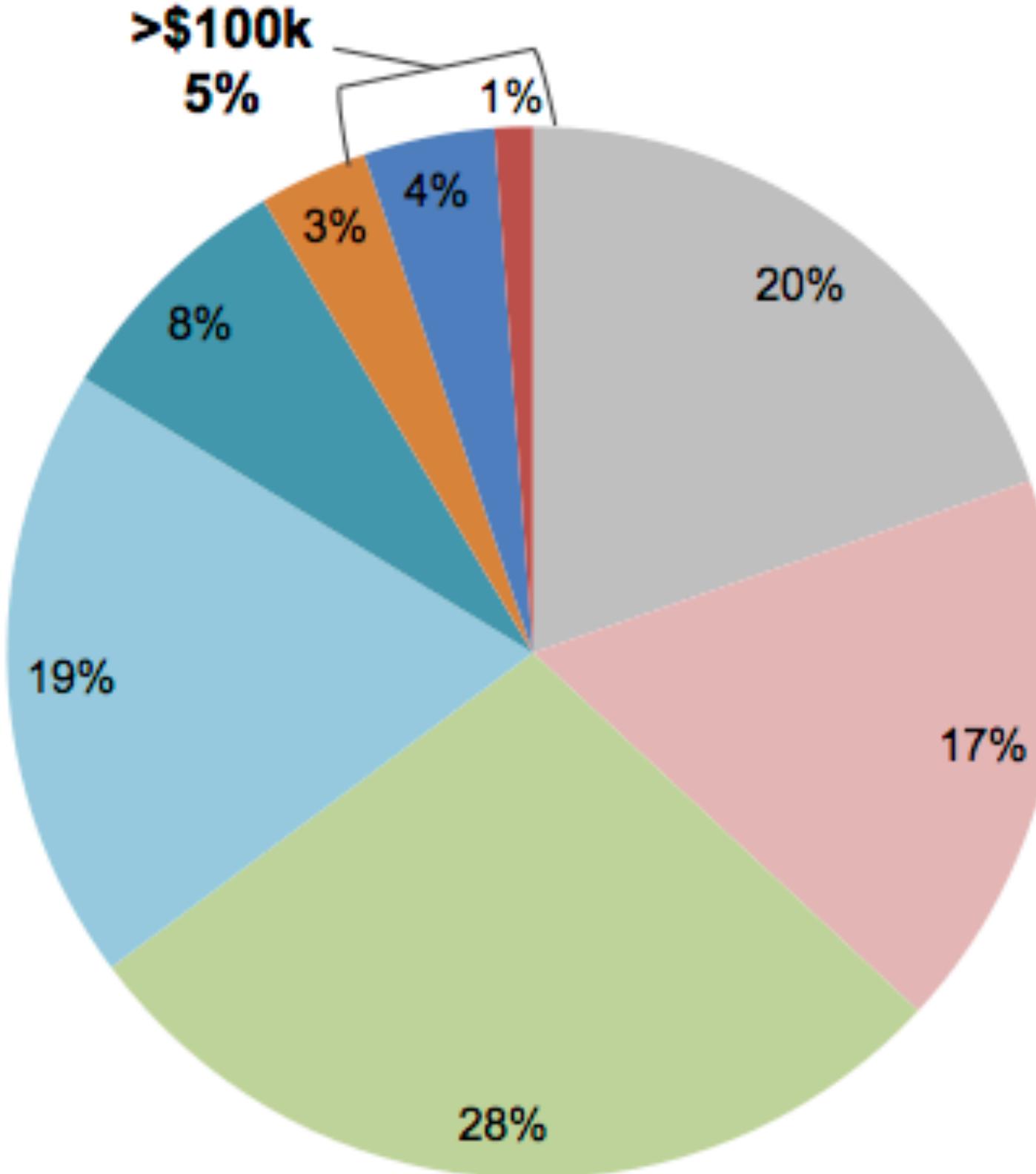
Yahoo Revenue to Employee Ratio, 2004-2014



Say No to Pie

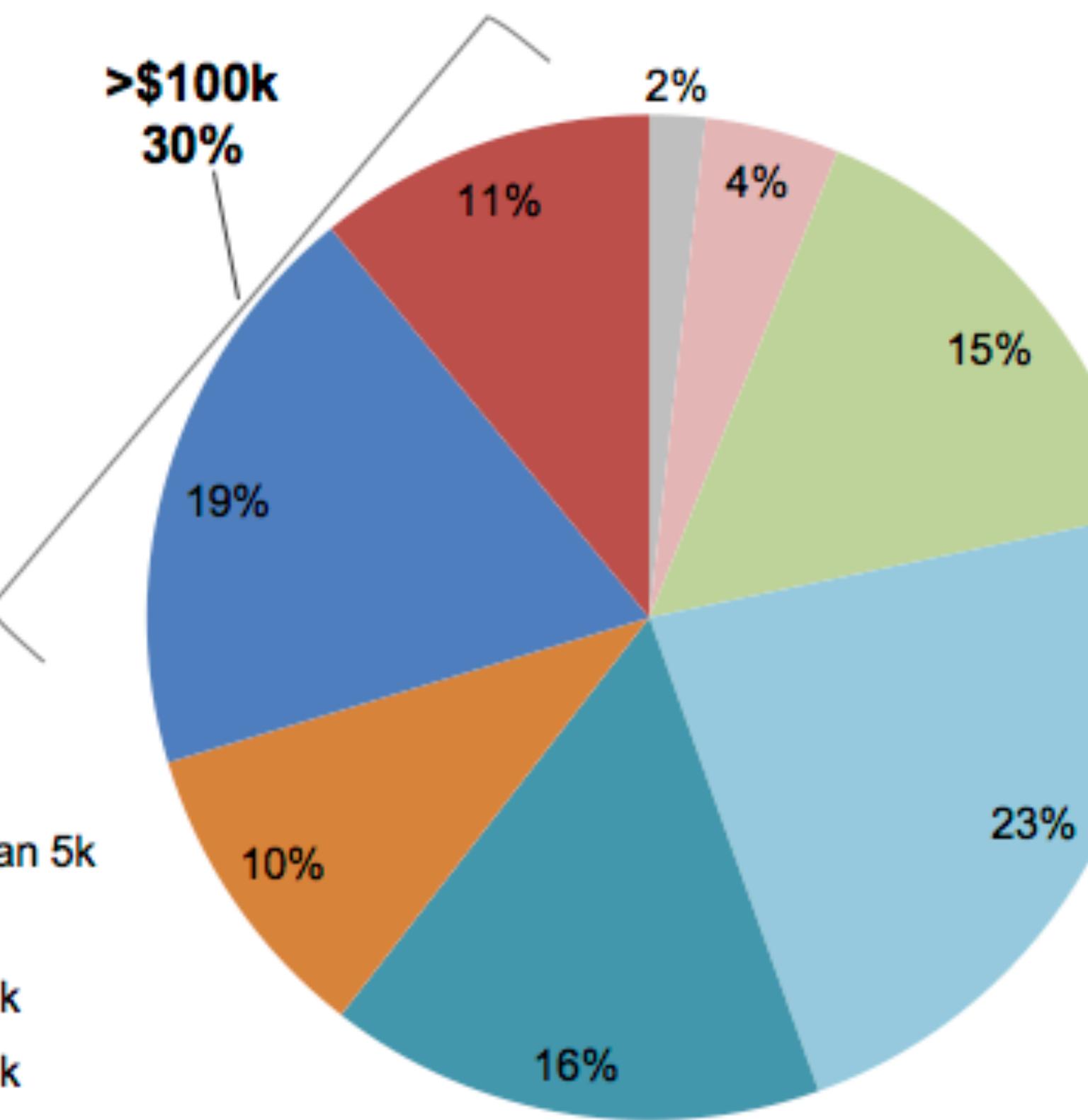
Borrower Distribution by Outstanding Balance

out of 44 million borrowers in 2016



Debt Distribution by Outstanding Balance

out of \$1.3 trillion in 2016



- less than 5k
- 5k-10k
- 10k-25k
- 25k-50k
- 50k-75k
- 75k-100k
- 100k-200k
- >100k
- 200k+

```
head(studebt)
```

```
## # A tibble: 6 x 4
##       Debt      type   pct Debtrc
##       <ord>     <fctr> <int>  <ord>
## 1 Under $5 Borrowers     20 Under $5
## 2 $5-$10 Borrowers      17 $5-$10
## 3 $10-$25 Borrowers     28 $10-$25
## 4 $25-$50 Borrowers     19 $25-$50
## 5 $50-$75 Borrowers      8 $50-$75
## 6 $75-$100 Borrowers      3 $75-$100
```

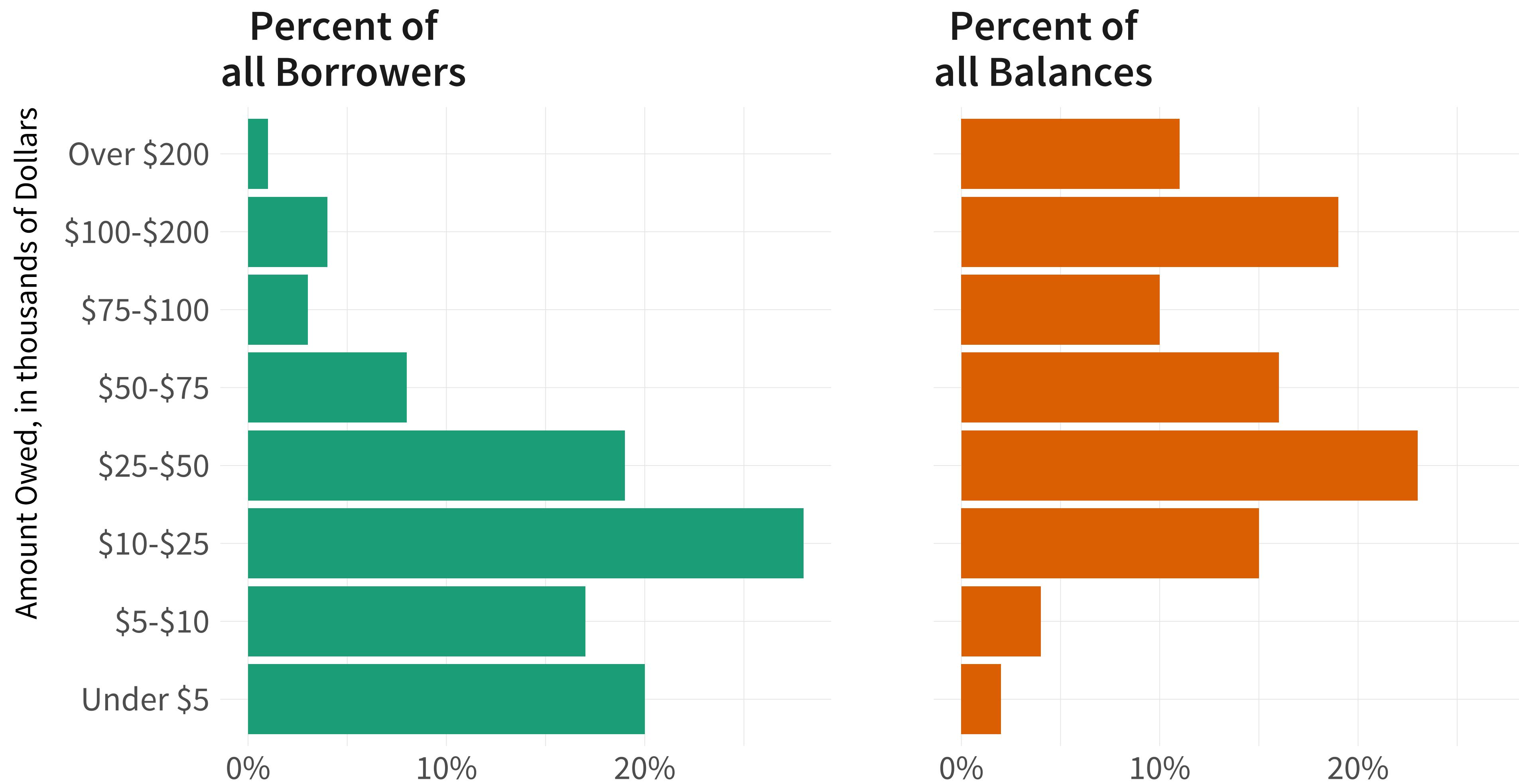
```
p_xlab <- "Amount Owed, in thousands of Dollars"
p_title <- "Outstanding Student Loans"
p_subtitle <- "44 million borrowers owe a total of $1.3 trillion"
p_caption <- "Source: FRB NY"

f_labs <- c(Borrowers = "Percent of\nall Borrowers",
            Balances = "Percent of\nall Balances")

p <- ggplot(data = studebt,
             mapping = aes(x = Debt, y = pct/100, fill = type))
p + geom_bar(stat = "identity") +
  scale_fill_brewer(type = "qual", palette = "Dark2") +
  scale_y_continuous(labels = scales::percent) +
  guides(fill = FALSE) +
  theme(strip.text.x = element_text(face = "bold")) +
  labs(y = NULL, x = p_xlab,
       caption = p_caption,
       title = p_title,
       subtitle = p_subtitle) +
  facet_grid(~ type, labeller = as_labeller(f_labs)) +
  coord_flip()
```

Outstanding Student Loans

44 million borrowers owe a total of \$1.3 trillion



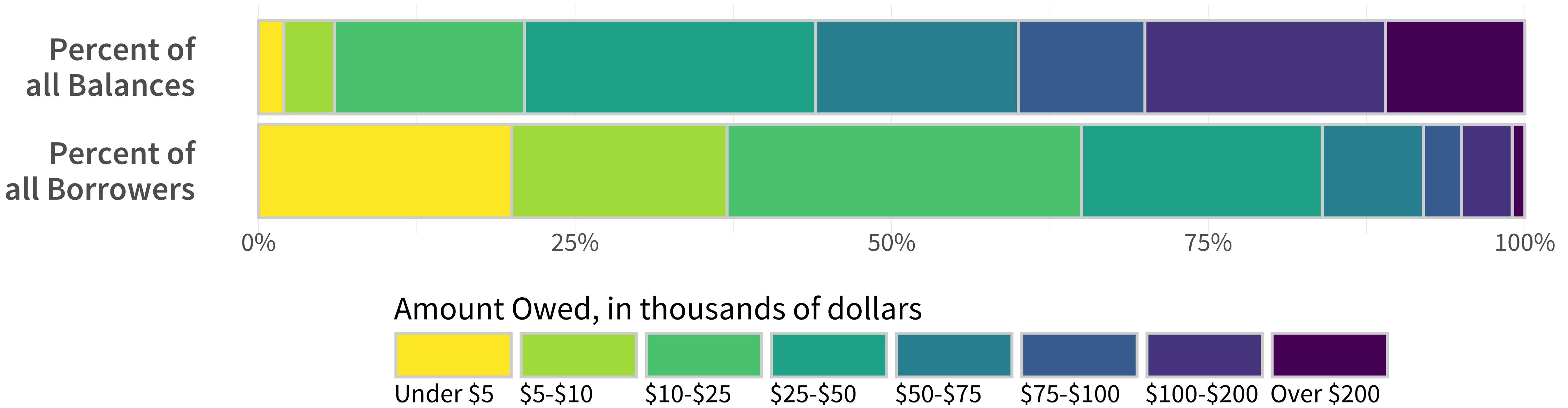
Source: FRB NY

```
library(viridis)

p <- ggplot(studebt, aes(y = pct/100, x = type, fill = Debtrc))
p + geom_bar(stat = "identity", color = "gray80") +
  scale_x_discrete(labels = as_labeller(f_labs)) +
  scale_y_continuous(labels = scales::percent) +
  scale_fill_viridis(discrete = TRUE) +
  guides(fill = guide_legend(reverse = TRUE,
                             title.position = "top",
                             label.position = "bottom",
                             keywidth = 3,
                             nrow = 1)) +
  labs(x = NULL, y = NULL,
       fill = "Amount Owed, in thousands of dollars",
       caption = p_caption,
       title = p_title,
       subtitle = p_subtitle) +
  theme(legend.position = "bottom",
        axis.text.y = element_text(face = "bold", hjust = 1, size = 12),
        axis.ticks.length = unit(0, "cm"),
        panel.grid.major.y = element_blank()) +
  coord_flip()
```

Outstanding Student Loans

44 million borrowers owe a total of \$1.3 trillion



Source: FRB NY

```
p <- ggplot(studebt, aes(y = pct/100, x = type, fill = Debtrc))  
p + geom_bar(stat = "identity", color = "gray80") +  
  scale_x_discrete(labels = as_labeller(f_labs)) +  
  scale_y_continuous(labels = scales::percent) +  
  scale_fill_viridis_d() +  
  guides(fill = guide_legend(reverse = TRUE,  
                             title.position = "top",  
                             label.position = "bottom",  
                             keywidth = 3,  
                             nrow = 1)) +  
  labs(x = NULL, y = NULL,  
        fill = "Amount Owed, in thousands of dollars",  
        caption = p_caption,  
        title = p_title,  
        subtitle = p_subtitle) +  
  theme(legend.position = "bottom",  
        axis.text.y = element_text(face = "bold", hjust = 1, size = 12),  
        axis.ticks.length = unit(0, "cm"),  
        panel.grid.major.y = element_blank()) +  
  coord_flip()
```

Three GSS Plots

gssr

The General Social Survey, 1972-2018

```
install.packages("drat")
drat::addRepo("kjhealy")
install.packages("gssr")
library(gssr)
data(gss_all)
```

```
> gss_all
# A tibble: 64,814 x 6,108
  year    id wrkstat hrs1 hrs2 evwork occ prestige wrkslf wrkgovt commute industry occ80 prestg80 indus80 indus07 occonet found      occ10
  <dbl> <dbl> <dbl+lbl> <dbl> <dbl> <dbl+lb> <dbl> <dbl+lb> <dbl> <dbl+lb> <dbl> <dbl+lb> <dbl> <dbl+lb> <dbl> <dbl+lb> <dbl> <dbl+lb>
1 1972     1 1 [WOR... NA NA NA NA 205   50 2 [SOM... NA NA 609  NA NA NA NA NA 1 [Fou... 520 [Who...
2 1972     2 5 [RET... NA NA 1 [YES] 441   45 2 [SOM... NA NA 338  NA NA NA NA NA 1 [Fou... 7700 [Fir...
3 1972     3 2 [WOR... NA NA NA NA 270   44 2 [SOM... NA NA 718  NA NA NA NA NA 0 [Not... 4920 [Rea...
4 1972     4 1 [WOR... NA NA NA NA 1    57 2 [SOM... NA NA 319  NA NA NA NA NA 1 [Fou... 800 [Acc...
5 1972     5 7 [KEE... NA NA 1 [YES] 385   40 2 [SOM... NA NA 448  NA NA NA NA NA 1 [Fou... 5020 [Tel...
6 1972     6 1 [WOR... NA NA NA NA 281   49 2 [SOM... NA NA 209  NA NA NA NA NA 1 [Fou... 4850 [Sal...
7 1972     7 1 [WOR... NA NA NA NA 522   41 2 [SOM... NA NA 69   NA NA NA NA NA 0 [Not... 6440 [Pip...
8 1972     8 1 [WOR... NA NA NA NA 314   36 2 [SOM... NA NA 587  NA NA NA NA NA 1 [Fou... 5350 [Ord...
9 1972     9 2 [WOR... NA NA NA NA 912   26 2 [SOM... NA NA 669  NA NA NA NA NA 0 [Not... 4020 [Coo...
10 1972    10 1 [WOR... NA NA NA NA 984   18 2 [SOM... NA NA 769  NA NA NA NA NA 1 [Fou... 4230 [Mai...
# ... with 64,804 more rows, and 6,089 more variables: occindv <dbl+lbl>, occstatus <dbl+lbl>, occtag <dbl+lbl>, prestg10 <dbl+lbl>,
# prestg105plus <dbl+lbl>, indus10 <dbl+lbl>, indstatus <dbl+lbl>, indtag <dbl+lbl>, marital <dbl+lbl>, martype <dbl+lbl>, agewed <dbl+lbl>,
# divorce <dbl+lbl>, widowed <dbl+lbl>, spwrksta <dbl+lbl>, sphrs1 <dbl+lbl>, sphrs2 <dbl+lbl>, spevwork <dbl+lbl>, cowrksta <dbl+lbl>,
# cowrkslf <dbl+lbl>, coevwork <dbl+lbl>, cohrls1 <dbl+lbl>, cohrls2 <dbl+lbl>, spocc <dbl+lbl>, sppres <dbl+lbl>, spwrkslf <dbl+lbl>, spind <dbl+lbl>,
# spocc80 <dbl+lbl>, sppres80 <dbl+lbl>, spind80 <dbl+lbl>, spocc10 <dbl+lbl>, spoccindv <dbl+lbl>, spoccstatus <dbl+lbl>, spocctag <dbl+lbl>,
```

```
library(ggrepel)  
library(forcats)  
library(haven)
```

◀ For help cleaning some Stata elements

```
library(survey)
```

◀ Thomas Lumley's survey library, and Grad Freedman Ellis's compatibility layer

```
options(survey.lonely.psu = "adjust")  
options(na.action="na.pass")
```

◀ Set some survey options relevant to the GSS

```
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")  
  error <- "Choose cb, rcb, or bly only."  
  
  case_when(  
    palette == "cb" ~ cb_palette,  
    palette == "rcb" ~ rcb_palette,  
    palette == "bly" ~ bly_palette  
  )  
}
```

◀ A very rough
custom colors
function

```
convert_agegrp <- function(x){  
  x <- stringr::str_replace_all(x, "\\\", "")  
  x <- stringr::str_replace_all(x, "\\[", "")  
  x <- stringr::str_replace_all(x, "\\]", "")  
  x <- stringr::str_replace_all(x, ",\"", "-")  
  x <- stringr::str_replace_all(x, "-89", "+")  
  regex <- "^\.*$"  
  x <- stringr::str_replace_all(x, regex, "Age \\\1")  
  x  
}
```

◀ A rough label-cleaning function

```
cont_vars <- c("year", "id", "ballot", "age")  
  
cat_vars <- c("race", "sex", "fefam")
```

◀ The variables that we want from
the big GSS file. Categorical,
Continuous, and Weights

```
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)
```

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

Calculate the breaks for Age quintiles

```
gss_svy <- gss_all %>%  
  select(vars) %>%  
  filter(year > 1974) %>%  
  modify_at(vars(), zap_missing) %>%  
  modify_at(wt_vars, as.numeric) %>%  
  modify_at(cat_vars, as_factor) %>%  
  mutate(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 = fct_relabel(fefam, snakecase::to_title_case),  
        fefam_d = fct_recode(fefam,  
                           Agree = "Strongly Agree",  
                           Disagree = "Strongly Disagree"))
```

Select the variables we want,
and do some recoding

gss_svy

```
## # A tibble: 60,213 x 18
##       year     id ballot   age race   sex   fefam   vpsu vstrat oversamp formwt
##       <dbl> <dbl> <dbl+> <dbl> <fct> <fct> <fct> <dbl> <dbl> <dbl> <dbl>
## 1 1975     1      NA    38  WHITE MALE <NA>     1    7001     1     1
## 2 1975     2      NA    20  WHITE FEMA... <NA>     1    7001     1     1
## 3 1975     3      NA    61  WHITE FEMA... <NA>     1    7001     1     1
## 4 1975     4      NA    19  WHITE MALE <NA>     1    7001     1     1
## 5 1975     5      NA    28  WHITE MALE <NA>     1    7001     1     1
## 6 1975     6      NA    28  WHITE FEMA... <NA>     1    7002     1     1
## 7 1975     7      NA    35  WHITE FEMA... <NA>     1    7002     1     1
## 8 1975     8      NA    64  WHITE FEMA... <NA>     1    7002     1     1
## 9 1975     9      NA    53  WHITE MALE <NA>     1    7002     1     1
## 10 1975    10      NA    34  WHITE FEMA... <NA>     1    7002     1     1
## # ... with 60,203 more rows, and 7 more variables: wtssall <dbl>, sampcode <dbl>,
## #       sample <dbl>, agequint <fct>, year_f <fct>, young <chr>, fefam_d <fct>
```

```
gss_svy <- gss_svy %>%
  mutate(compwt = oversamp * formwt * wtssall,
        samplerc = ifelse(sample %in% 3:4, 3,
                           ifelse(sample %in% 6:7, 6,
                                 sample)))
```

```
gss_svy <- gss_svy %>%
  drop_na(fefam_d) %>%
  mutate(stratvar = interaction(year, vstrat)) %>%
  as_survey_design(id = vpsu,
                    strata = stratvar,
                    weights = wtssall,
                    nest = TRUE)
```

gss_svy

```
## 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), agequint (fct),
##   year_f (fct), young (chr), fefam_d (fct), compwt (dbl), samplerc (dbl),
##   stratvar (fct)
```



More GSS-specific weight recoding



Now make the tibble a survey object

```
out_ff <- gss_svy %>%  
  group_by(year, sex, young, fefam_d) %>%  
  summarize(prop = survey_mean(na.rm = TRUE,  
                                vartype = "ci")) %>%  
  drop_na()
```

out_ff

```
## # A tibble: 168 x 7  
##   year   sex   young fefam_d   prop prop_low prop_upp  
##   <dbl> <fct> <chr> <fct>     <dbl>     <dbl>     <dbl>  
## 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
```

Calculate survey-weighted means and confidence intervals for subgroups

Hurray! A tidy table of results

Survey-weighted trends, by sex and young/old groups

```
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) +
  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\\nthe\\nhome and the woman takes care of the home and family''",
       caption = "Kieran Healy http://socviz.co.\nData source: General Social Survey") +
  theme(legend.position = "bottom")
```

Nicer facet labels

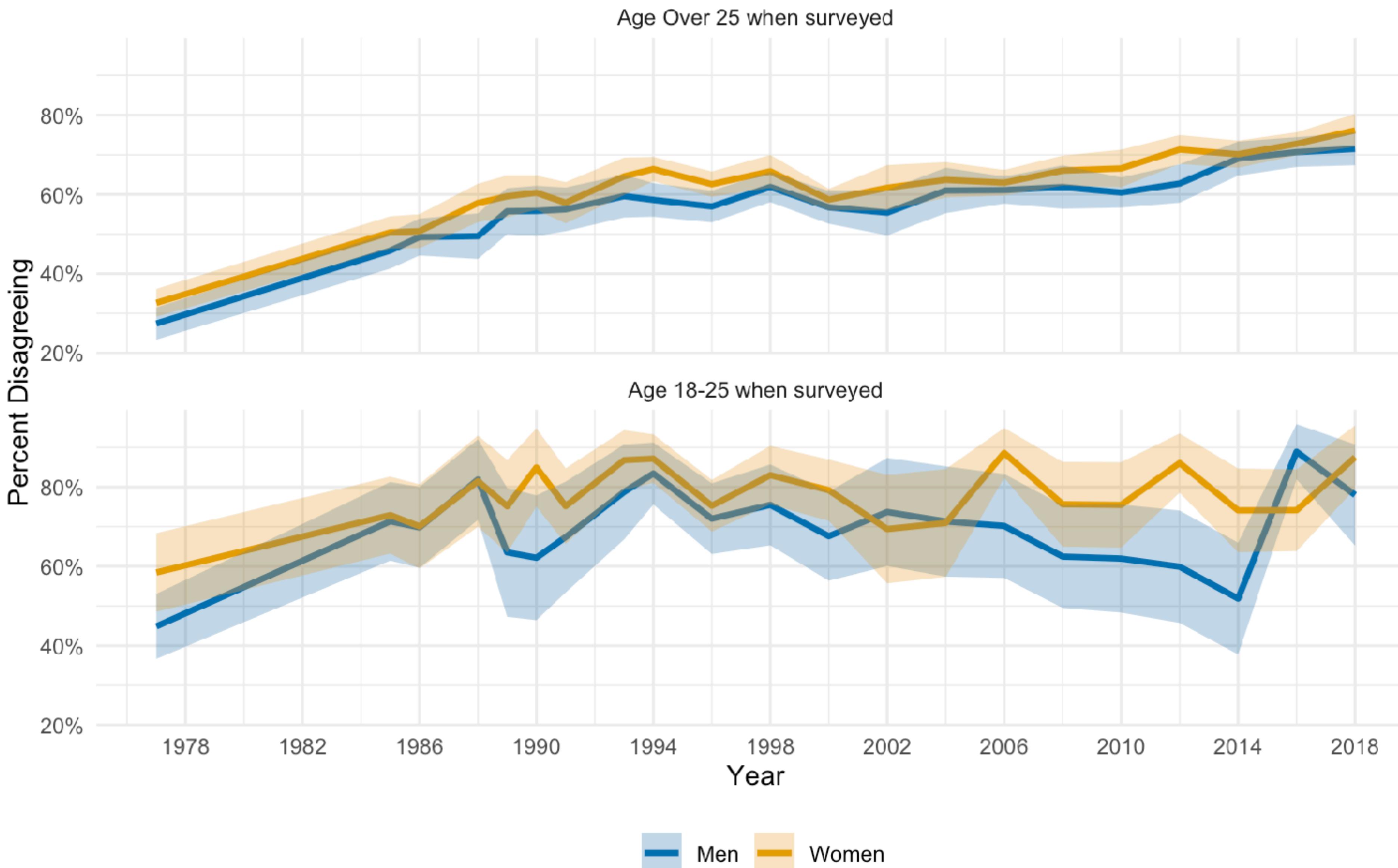
Subset to “Disagree”, then map both color and fill

Manually set x-axis breaks

Manually set color and fill, remember you need to do both, or you'll get two guides

Use our nicer facet labels

Disagreement with the statement, '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'



```

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_))

```

```

> out_ff_agequint
# A tibble: 247 x 7
  year agequint fefam_d    prop  prop_se end_label start_label
  <dbl> <fct>     <fct>    <dbl>   <dbl> <chr>      <chr>
1 1977 Age 18-29 Agree     0.477  0.0244 NA          18-29
2 1977 Age 18-29 Disagree  0.523  0.0244 NA          18-29
3 1977 Age 29-38 Agree     0.527  0.0365 NA          29-38
4 1977 Age 29-38 Disagree  0.473  0.0365 NA          29-38
5 1977 Age 38-49 Agree     0.667  0.0293 NA          38-49
6 1977 Age 38-49 Disagree  0.333  0.0293 NA          38-49
7 1977 Age 49-63 Agree     0.812  0.0228 NA          49-63
8 1977 Age 49-63 Disagree  0.188  0.0228 NA          49-63
9 1977 Age 63+   Agree     0.885  0.0237 NA          63+
10 1977 Age 63+   Disagree  0.115  0.0237 NA          63+
# ... with 237 more rows

```

See how we calculate
start and end labels here

Survey-Weighted trend by Age Quintiles

```
man_cols <- RColorBrewer::brewer.pal(9, "Blues")
man_cols <- man_cols[4:9]

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'",
       caption = "Kieran Healy http://socviz.co.\nData source: General Social Survey. Shaded ranges are population-adjusted standard errors for each age group.") +
  theme(legend.position = "right")
```

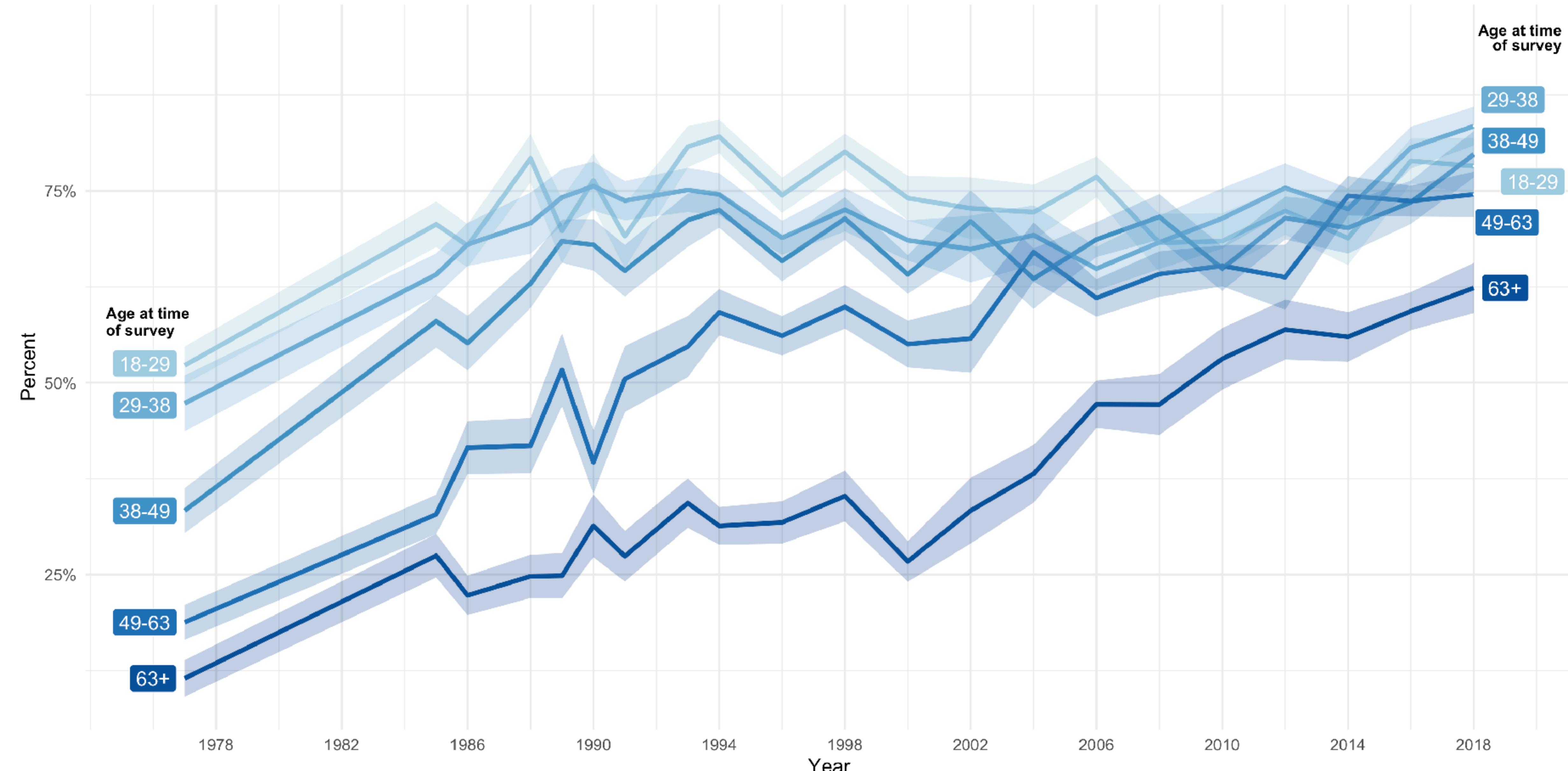
◀ **annotate() for a helpful note**

◀ **geom_label_repel() for the labels**

◀ **coord_cartesian() to make space**

Changing Attitudes to Gender Roles, by Age Quintiles

Percent Disagreeing with the statement, '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'



Quasi-Cohort Comparison

```
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, ~ stringr::str_replace_all(.x, "\\.", " in ")),
        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_))
```

◀ **case_when()** is very handy

Comparing Approximately the Same Cohort in 1977 and 2018,

Comparing the Old in 1977 vs the Young in 2018"),

Quasi-Cohort Comparison

Here we've done a lot of the plot setup inside the tibble
In the actual plot we'll use compare as a facet, and this
will make end_label work as expected.

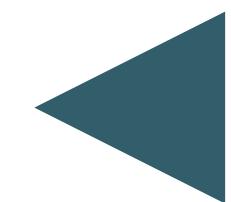
```
> cohort_comp
# A tibble: 16 x 6
  cohort      fefam      prop  prop_se compare                               end_label
  <ord>      <fct>     <dbl>   <dbl> <chr>
  1 Age 18-29 in 1977 Strongly Agree 0.101    0.0171 Comparing Approximately the Same Cohort in 1977 and 2018 NA
  2 Age 18-29 in 1977 Agree        0.377    0.0225 Comparing Approximately the Same Cohort in 1977 and 2018 NA
  3 Age 18-29 in 1977 Disagree     0.435    0.0268 Comparing Approximately the Same Cohort in 1977 and 2018 NA
  4 Age 18-29 in 1977 Strongly Disagree 0.0879  0.0178 Comparing Approximately the Same Cohort in 1977 and 2018 18-29 in 1977
  5 Age 63+ in 1977 Strongly Agree  0.372    0.0372 Comparing the Old in 1977 vs the Young in 2018 NA
  6 Age 63+ in 1977 Agree        0.513    0.0448 Comparing the Old in 1977 vs the Young in 2018 NA
  7 Age 63+ in 1977 Disagree      0.108    0.0237 Comparing the Old in 1977 vs the Young in 2018 NA
  8 Age 63+ in 1977 Strongly Disagree 0.00703 0.00524 Comparing the Old in 1977 vs the Young in 2018 63+ in 1977
  9 Age 18-29 in 2018 Strongly Agree 0.0504  0.0167 Comparing the Old in 1977 vs the Young in 2018 NA
 10 Age 18-29 in 2018 Agree       0.167    0.0314 Comparing the Old in 1977 vs the Young in 2018 NA
 11 Age 18-29 in 2018 Disagree     0.472    0.0382 Comparing the Old in 1977 vs the Young in 2018 NA
 12 Age 18-29 in 2018 Strongly Disagree 0.310    0.0355 Comparing the Old in 1977 vs the Young in 2018 18-29 in 2018
 13 Age 63+ in 2018 Strongly Agree 0.0947  0.0166 Comparing Approximately the Same Cohort in 1977 and 2018 NA
 14 Age 63+ in 2018 Agree        0.282    0.0296 Comparing Approximately the Same Cohort in 1977 and 2018 NA
 15 Age 63+ in 2018 Disagree      0.413    0.0272 Comparing Approximately the Same Cohort in 1977 and 2018 NA
 16 Age 63+ in 2018 Strongly Disagree 0.210    0.0257 Comparing Approximately the Same Cohort in 1977 and 2018 63+ in 2018
```

Quasi-Cohort Comparison

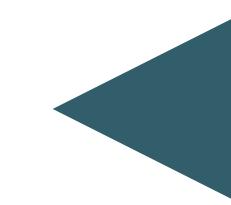
```
p <- ggplot(cohort_comp,
             aes(x = fefam, y = prop, group = cohort,
                  color = cohort, fill = cohort,
                  ymin = prop - prop_se,
                  ymax = prop + prop_se)) +
  geom_point(size = 3) +
  geom_line(size = 1.2) +
  geom_ribbon(alpha = 0.2, color = 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 = 0.7) +
  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\nnachiever outside the home and the woman takes care of the home and family'",
       caption = "Kieran Healy http://socviz.co.\nData source: General Social Survey. Shaded ranges are population-adjusted standard errors for each age group.")
```



Turn off the ribbon border



Manual color and fill palette again

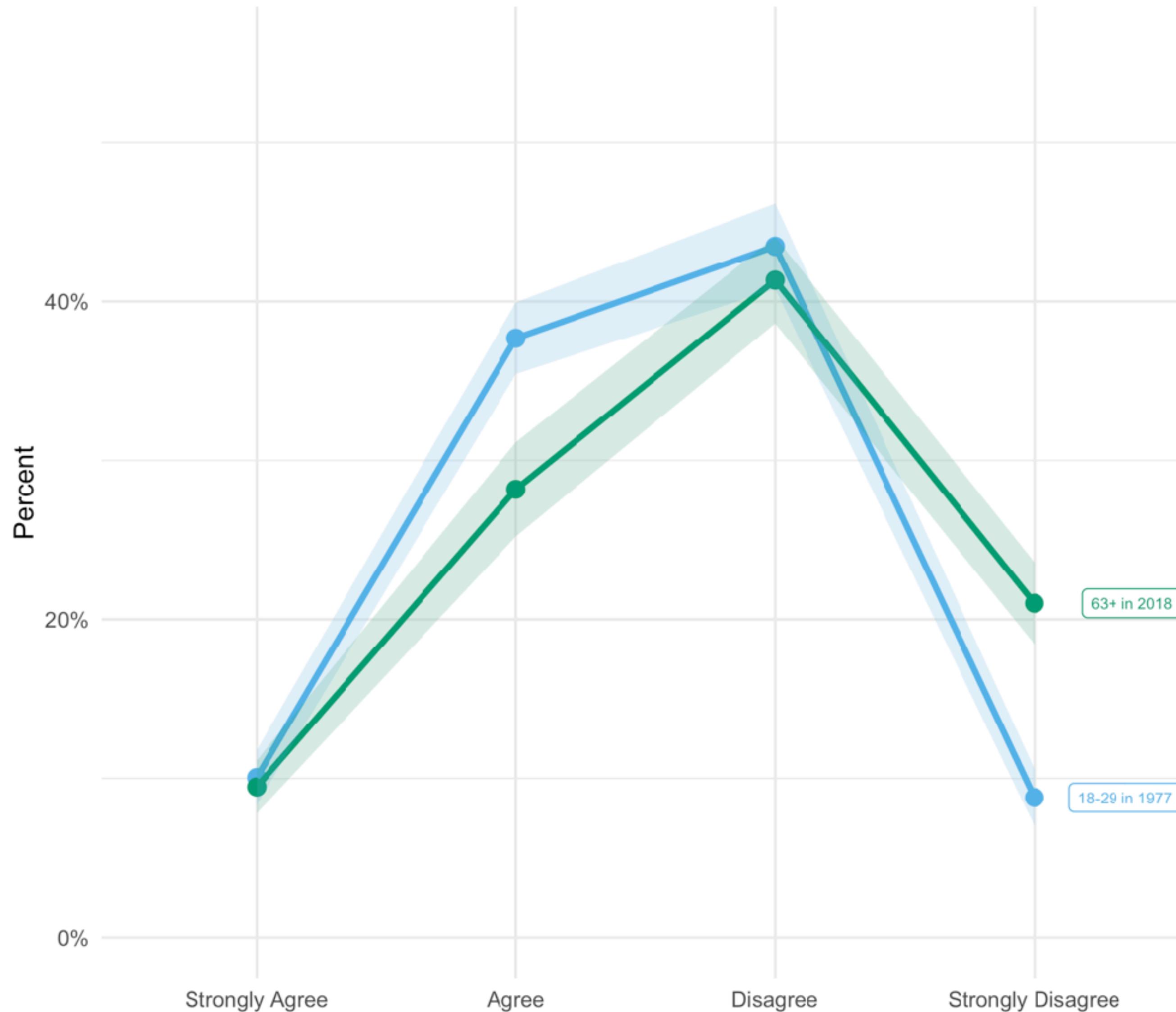


Tweak the end-labels

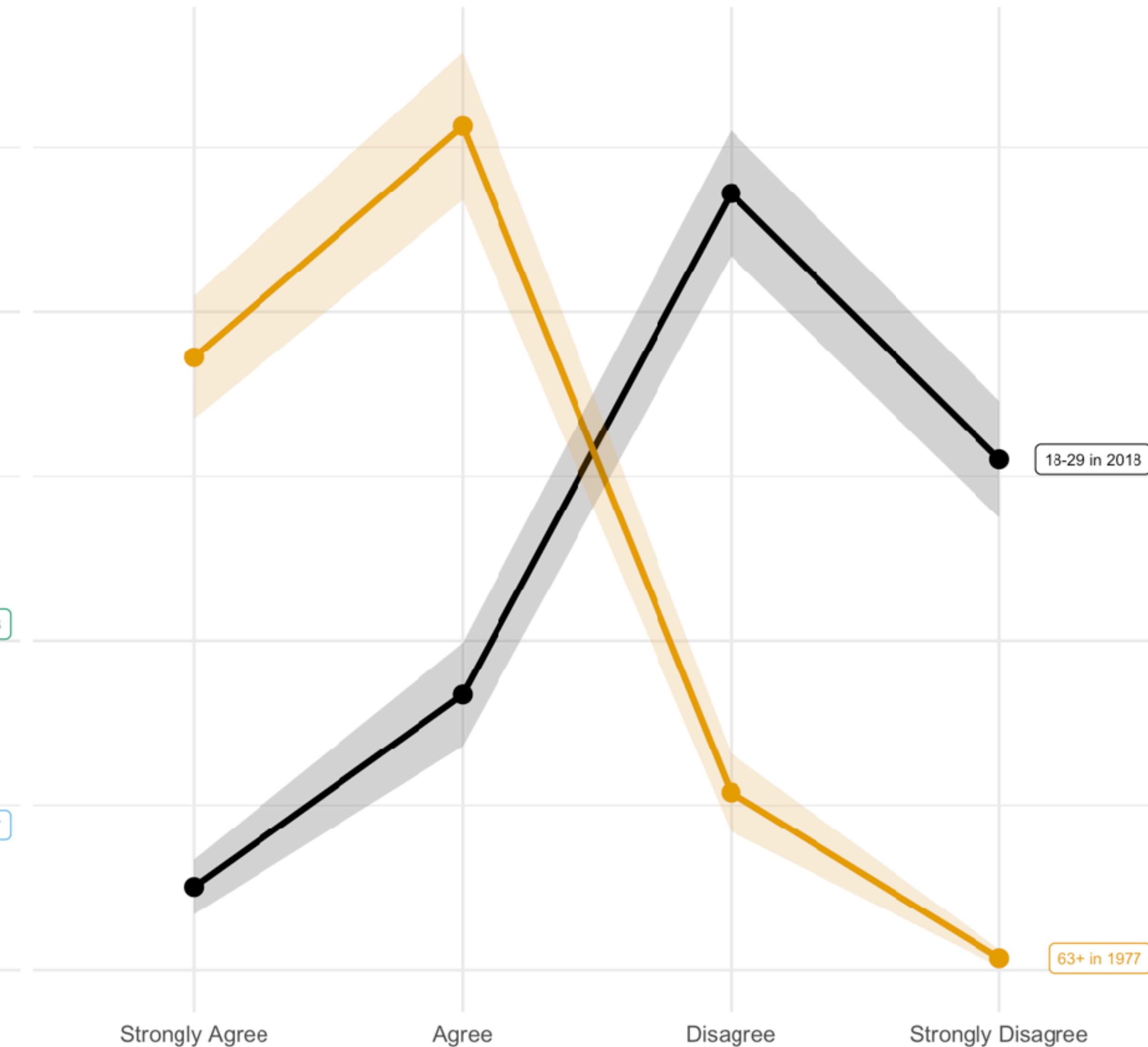
Generational Replacement, or, People Don't Change Much, They Just Get Old

Responses to the statement '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'

Comparing Approximately the Same Cohort in 1977 and 2018



Comparing the Old in 1977 vs the Young in 2018



But I want a Pony

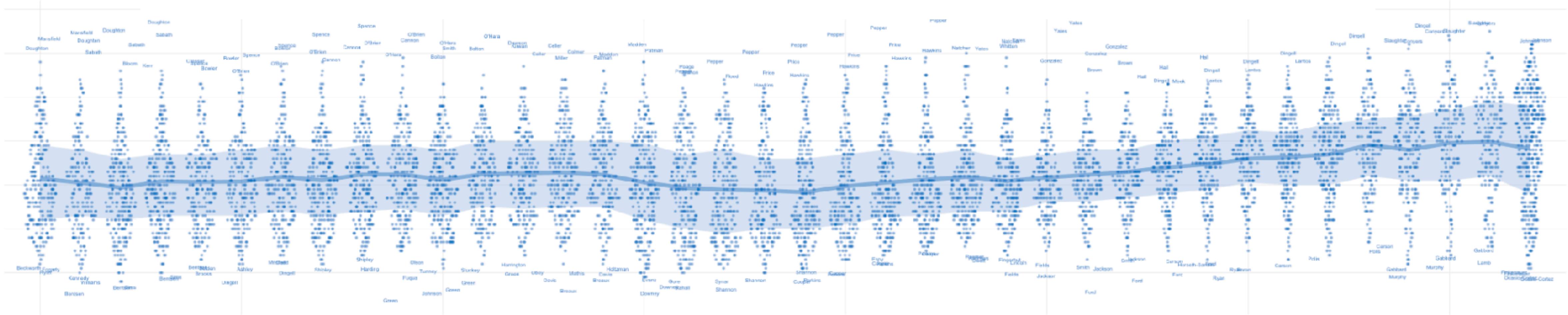
Age Distribution of Congressional Representatives, 1945-2019

Trend line is mean age; bands are 25th and 75th percentiles of the range.

Youngest and oldest percentiles are named instead of being shown by points.

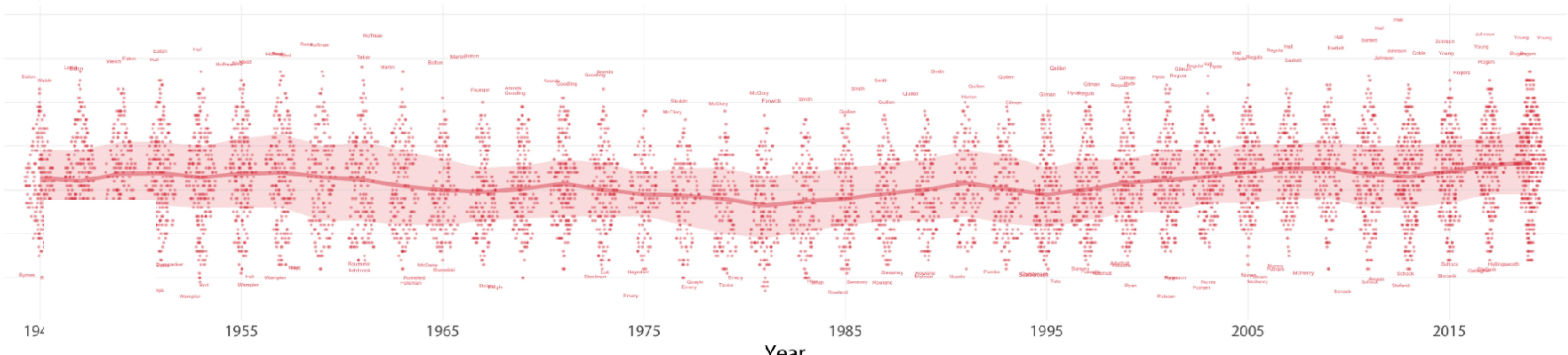
Show Ponies

Democrats



Age

Republicans



194

1955

1965

1975

1985

1995

2005

2015

Year

MORTALITY IN FRANCE 1816 – 2016

100 | Males

75

50

25

0

100 | Females

Age

75

50

25

0

1820

1845

1870

1895

1920

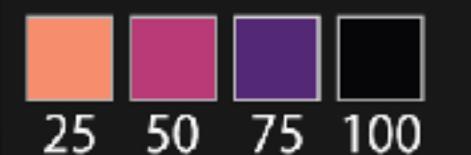
1945

1970

1995

2015

Death Rate Percentile



Mortality rates are calculated for each age in each year and binned by percentile. The darker the color at any particular point, the more people of that age die in that year. The lighter the color, the more people of that age survive in that year.

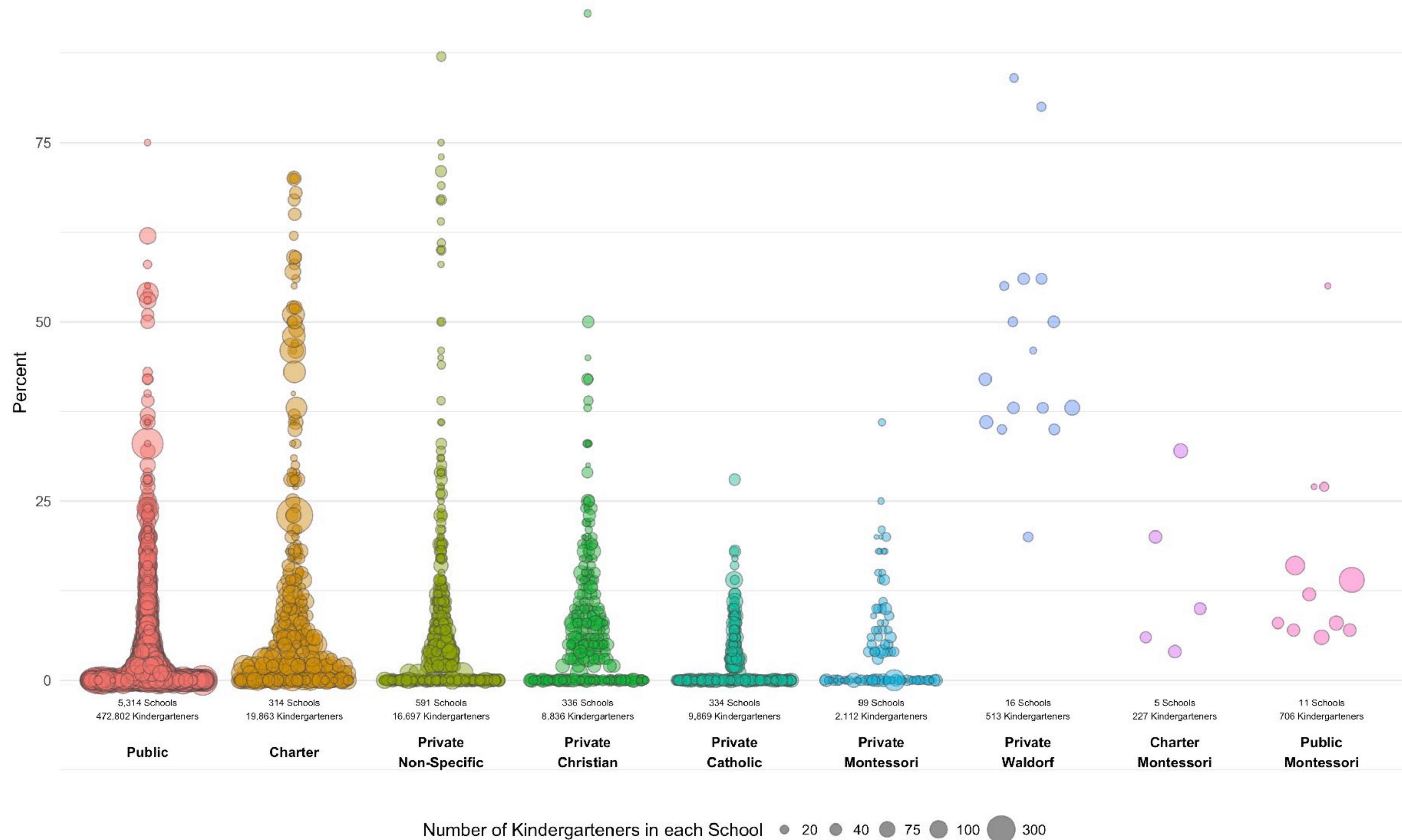
Historical trends are visible, such as the rapid decrease in infant mortality rates after World War II, as well as increased life expectancy overall. Specific events show up as vertical streaks in the graph. The death toll due to wars is evident for Males. Pandemics are also visible, most notably the 1918 Influenza pandemic and the death toll due to Smallpox outbreaks after the Franco-Prussian war of 1870.

Diagonal streaks in the data are visible in some parts of the graph. These are artifacts due to the estimation of the mortality rate in some years. Single-year-of-age figures prior to 1900 are calculated from five-year age groups, as no single-year data is available in the original mortality tabulations from which the rates were derived.

Kieran Healy / [socrata.co](#)
Data Visualization: A Practical Introduction
is published by Princeton University Press

Vaccination Exemption Rates in California Kindergartens

Percent of Kindergarteners with a Personal Belief Exemption, by Type and Size of School.



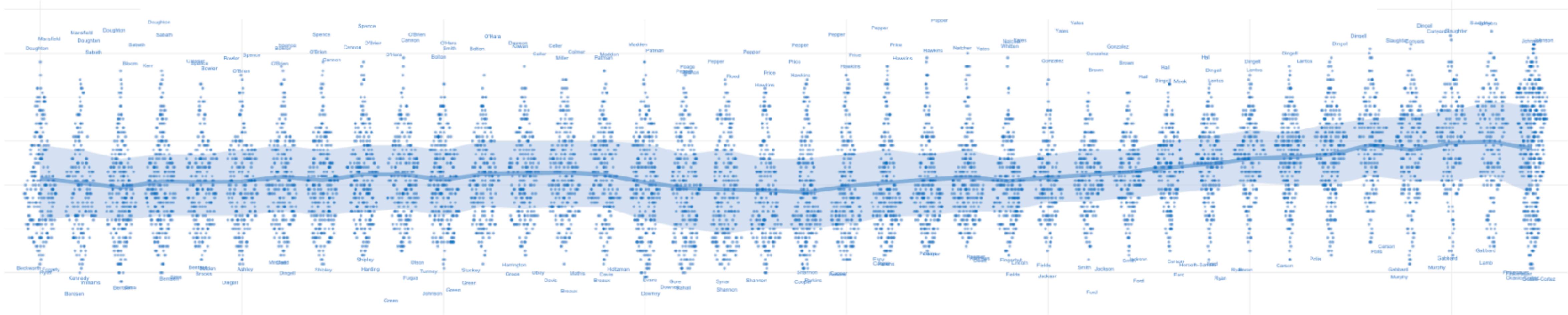
Age Distribution of Congressional Representatives, 1945-2019

Trend line is mean age; bands are 25th and 75th percentiles of the range.

Youngest and oldest percentiles are named instead of being shown by points.

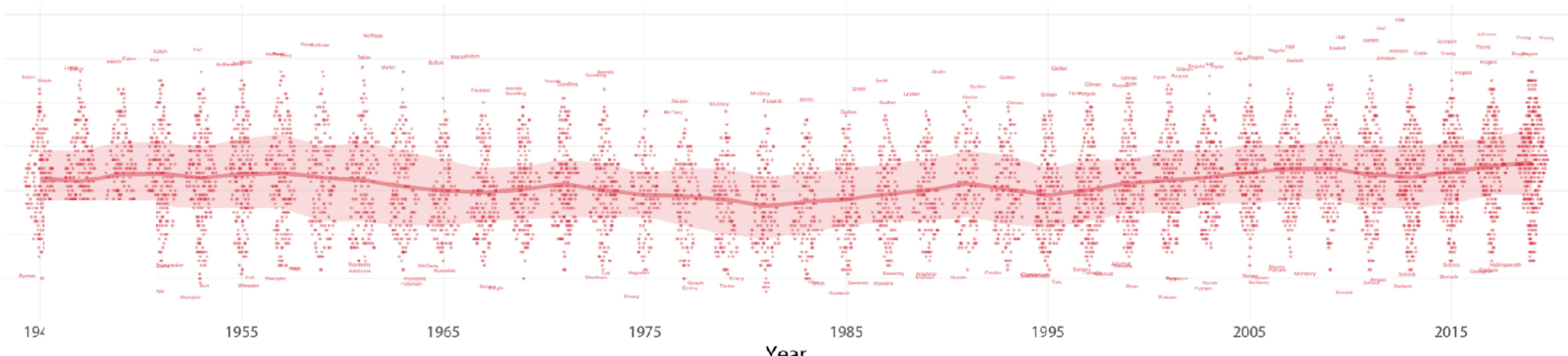
Show Ponies

Democrats



Age

Republicans



194

1955

1965

1975

1985

1995

2005

2015

Year

```
install.packages("drat")
drat::addRepo("kjhealy")
install.packages("congress")
library(congress)
```

congress

Representatives and Senators since 1945

```
> congress
# A tibble: 21,009 x 38
   congress last first middle suffix nickname born      death     sex position party state district start      end religion race
   <dbl> <chr> <chr> <chr> <chr> <chr> <date>    <date>    <chr> <chr> <chr> <chr> <chr> <date>    <chr> <chr> <chr>
1       79 Aber... Thom... Gerst... NA     NA     1903-05-16 1953-01-23 M     U.S. Re... Demo... MS     4     1945-01-03 01/0... Methodi... White
2       79 Adams Sher... NA     NA     1899-01-08 1986-10-27 M     U.S. Re... Repu... NH     2     1945-01-03 01/0... Not spe... White
3       79 Aiken Geor... David NA     NA     1892-08-20 1984-11-19 M     U.S. Se... Repu... VT     NA    1945-01-03 01/0... Protest... White
4       79 Allen Asa  Leona... NA     NA     1891-01-05 1969-01-05 M     U.S. Re... Demo... LA     8     1945-01-03 01/0... Not spe... White
5       79 Allen Leo  Elwood NA     NA     1898-10-05 1973-01-19 M     U.S. Re... Repu... IL    13    1945-01-03 01/0... Presbyt... White
6       79 Almo... J. Linds... Jr. NA     NA     1898-06-15 1986-04-14 M     U.S. Re... Demo... VA     6     1946-02-04 04/1... Lutheran White
7       79 Ande... Herm... Carl  NA     NA     1897-01-27 1978-07-26 M     U.S. Re... Repu... MN     7     1945-01-03 01/0... Lutheran White
8       79 Ande... Clin... Presba NA     NA     1895-10-23 1975-11-11 M     U.S. Re... Demo... NM    AL    1941-01-03 06/3... Presbyt... White
9       79 Ande... John  Zuing... NA     NA     1904-03-22 1981-02-09 M     U.S. Re... Repu... CA     8     1945-01-03 01/0... Not spe... White
10      79 Andr... Augu... Herman NA     NA     1890-10-11 1958-01-14 M     U.S. Re... Repu... MN     1     1945-01-03 01/1... Not spe... White
# ... with 20,999 more rows, and 21 more variables: educational_attainment <chr>, job_type_1 <chr>, job_type_2 <chr>, job_type_3 <chr>,
#   job_type_4 <chr>, job_type_5 <chr>, mil_1 <chr>, mil_2 <chr>, mil_3 <chr>, start_year <date>, end_year <date>, name_dob <chr>,
#   pid <dbl>, start_age <int>, poc <chr>, days_old <dbl>, months_old <int>, full_name <chr>, end_career <date>, entry_age <int>,
#   yr_fac <fct>
> |
```

```

library(lubridate)
library(tidyverse)
library(ggbeeswarm)
library(ggrepel)
library(congress)

oldest_group_by_year <- congress %>%
  filter(party %in% c("Democrat", "Republican"),
         position == "U.S. Representative") %>%
  group_by(congress, party) %>%
  filter(start_age > quantile(start_age, 0.99, na.rm = TRUE))

```

```

> oldest_group_by_year
# A tibble: 181 x 32
# Groups:   congress, party [75]
  congress    pid last first middle suffix born      death     sex position party state district start      end
  <dbl> <dbl> <chr> <chr> <chr> <chr> <date> <date> <chr> <chr> <chr> <chr> <chr> <date> <chr>
1      79    136 Doug... Robe... Lee    NA 1863-11-07 1954-10-01 M U.S. Re... Demo... NC     9 1945-01-03 01/0...
2      79    146 Eaton Char... Aubrey NA 1868-03-29 1953-01-23 M U.S. Re... Repu... NJ     5 1945-01-03 01/0...
3      79    326 Mans... Jose... J.    NA 1861-02-09 1947-07-12 M U.S. Re... Demo... TX     9 1917-04-02 07/1...
4      79    529 Welch Rich... Joseph NA 1869-02-13 1949-09-10 M U.S. Re... Repu... CA     5 1926-08-31 09/1...
5      80    136 Doug... Robe... Lee    NA 1863-11-07 1954-10-01 M U.S. Re... Demo... NC     9 1945-01-03 01/0...
6      80    146 Eaton Char... Aubrey NA 1868-03-29 1953-01-23 M U.S. Re... Repu... NJ     5 1945-01-03 01/0...
7      80    624 Lewis Will... NA      NA 1868-09-22 1959-08-08 M U.S. Re... Repu... KY     9 1948-05-03 01/0...
8      80    326 Mans... Jose... J.    NA 1861-02-09 1947-07-12 M U.S. Re... Demo... TX     9 1917-04-02 07/1...
9      80    449 Saba... Adol... Joach... NA 1866-04-04 1952-11-06 M U.S. Re... Demo... IL     5 1945-01-03 01/0...
10     81     43 Bloom Sol    NA      NA 1870-03-09 1949-03-07 M U.S. Re... Demo... NY    20 1923-01-30 03/0...
# ... with 171 more rows, and 17 more variables: religion <chr>, race <chr>, education <chr>, occ1 <chr>, occ2 <chr>,
#   occ3 <chr>, mil1 <chr>, mil2 <chr>, mil3 <chr>, start_year <date>, end_year <date>, start_age <int>, poc <chr>,
#   days_old <dbl>, entry_age <int>, end_career <date>, yr_fac <fct>

```

```

youngest_group_by_year <- congress %>%
  filter(party %in% c("Democrat", "Republican"),
         position == "U.S. Representative") %>%
  group_by(congress, party) %>%
  filter(start_age < quantile(start_age, 0.01, na.rm = TRUE))

```

```

> youngest_group_by_year
# A tibble: 164 x 32
# Groups:   congress, party [71]
  congress pid last first middle suffix born       death      sex position party state district start      end
  <dbl> <dbl> <chr> <chr> <chr> <chr> <date>     <date>    <chr> <chr> <chr> <chr> <chr> <chr> <date>     <chr>
 1      79  33 Beck.. Lind.. Gary  NA  1913-06-30 1984-03-09 M  U.S. Re.. Demo.. TX    3  1945-01-03 01/0...
 2      79  37 Benn.. Mari.. Tinsl.. NA  1914-06-06 2000-09-06 M  U.S. Re.. Repu.. MO    6  1945-01-03 01/0...
 3      79  75 Byrn.. John Willi.. NA  1913-06-12 1985-01-12 M  U.S. Re.. Repu.. WI    8  1945-01-03 01/0...
 4      79  167 Foga.. John Edward NA  1913-03-23 1967-01-10 M  U.S. Re.. Demo.. RI    2  1945-01-03 01/1...
 5      79  448 Ryter John Franc.. NA  1914-02-04 1978-02-05 M  U.S. Re.. Demo.. CT    AL  1945-01-03 01/0...
 6      80  563 Bent.. Lloyd Milla.. Jr. 1921-02-11 2006-05-23 M  U.S. Re.. Demo.. TX   15  1948-12-04 01/0...
 7      80  621 Kenn.. John Fitzg.. NA  1917-05-29 1963-11-22 M  U.S. Re.. Demo.. MA   11  1947-01-03 01/0...
 8      80  656 Nodar Robe.. Joseph Jr. 1916-03-23 1974-09-11 M  U.S. Re.. Repu.. NY    6  1947-01-03 01/0...
 9      80  664 Pott.. Char.. Edward NA  1916-10-30 1979-11-23 M  U.S. Re.. Repu.. MI   11  1947-08-26 11/0...
10      80  677 Sarb.. Geor.. Willi.. Jr. 1919-09-30 1973-03-04 M  U.S. Re.. Repu.. PA    5  1947-01-03 01/0...
# ... with 154 more rows, and 17 more variables: religion <chr>, race <chr>, education <chr>, occ1 <chr>, occ2 <chr>,
#   occ3 <chr>, mil1 <chr>, mil2 <chr>, mil3 <chr>, start_year <date>, end_year <date>, start_age <int>, poc <chr>,
#   days_old <dbl>, entry_age <int>, end_career <date>, yr_fac <fct>

```

```
mean_age_swarm <- congress %>%
  filter(position == "U.S. Representative",
         party %in% c("Democrat", "Republican")) %>%
  group_by(congress, party) %>%
  summarize(year = first(start_year),
            mean_age = mean(start_age, na.rm = TRUE),
            lo = quantile(start_age, 0.25, na.rm = TRUE),
            hi = quantile(start_age, 0.75, na.rm = TRUE)) %>%
  mutate(yr_fac = as.factor(year(year)))
```

```
> mean_age_swarm
```

A tibble: 76 x 7

Groups: congress [38]

	congress	party	year	mean_age	lo	hi	yr_fac
	<dbl>	<chr>	<date>	<dbl>	<dbl>	<dbl>	<chr>
1	79	Democrat	1945-01-03	51.5	42	59	1945
2	79	Republican	1945-01-03	52.8	46	59	1945
3	80	Democrat	1947-01-03	50.5	43	58	1947
4	80	Republican	1947-01-03	52.0	45	59	1947
5	81	Democrat	1949-01-03	49.4	42	56	1949
6	81	Republican	1949-01-03	53.7	47	61	1949
7	82	Democrat	1951-01-03	50.8	43	57	1951
8	82	Republican	1951-01-03	53.8	46.5	61	1951
9	83	Democrat	1953-01-03	50.7	43	57	1953
10	83	Republican	1953-01-03	52.9	46	60	1953

... with 66 more rows

##	[1]	136	146	326	529	136	146	624	326	449	43	136	146	449	529	136
##	[16]	146	260	288	449	920	117	260	477	920	249	979	378	423	477	920
##	[31]	249	979	378	423	477	80	249	378	423	477	80	249	378	477	492
##	[46]	80	329	378	788	44	711	329	788	471	44	711	788	1013	15	88
##	[61]	124	1252	294	15	88	106	1252	354	15	1252	320	389	320	389	322
##	[76]	1349	392	402	1372	166	1349	392	1776	1336	1349	392	406	1336	392	406
##	[91]	1837	1336	392	406	1362	1837	1336	392	406	1362	1837	1336	392	1362	1837
##	[106]	1337	978	1362	829	1690	978	1362	535	829	1690	1251	1362	829	1313	1690
##	[121]	1251	1796	1728	829	1313	1690	1251	2057	1796	1728	1030	1690	2057	1796	2485
##	[136]	1728	1030	2057	2057	1796	2068	1728	1030	2057	1796	2068	1728	2400	1030	2057
##	[151]	2068	1728	2400	1030	2057	2400	1030	2057	2366	2203	1405	1030	2057	2366	2290
##	[166]	1405	2366	2090	2290	1748	1405	2366	2090	2290	1748	2459	2090	1748	1748	2090
##	[181]	2459	33	37	75	167	448	563	621	656	664	677	700	707	563	664
##	[196]	808	707	563	852	865	808	563	922	957	994	1007	1016	1030	957	1007
##	[211]	1082	1030	957	1107	1112	1007	1217	1233	1256	1291	1217	1325	1328	202	1358
##	[226]	1368	202	1431	1350	1368	1472	202	1530	1539	1540	202	1577	1590	1616	1625
##	[241]	1641	1616	1625	1699	1712	1736	1768	1772	1774	1784	1768	1772	1876	1908	1909
##	[256]	1772	2003	2013	2014	2003	2129	2063	2168	2003	2099	2129	2168	2226	2232	2129
##	[271]	2251	2168	2226	2232	2129	2318	2168	2226	2342	2373	2382	2386	2433	2435	2472
##	[286]	2493	2535	2433	2566	2589	2595	2607	2630	2631	2566	2677	2682	2607	2630	2631
##	[301]	2566	2721	2677	2745	2755	2630	2781	2721	2630	2829	2842	2781	2849	2856	2881
##	[316]	2848	2781	2848	2809	2881	2848	2781	2826	2844	2854	2809	2826	2844	2818	2861

```
party_names <- c('Democrat' = "Democrats",
                 'Republican' = "Republicans")
```

```
party_names
```

```
##      Democrat    Republican
##  "Democrats" "Republicans"
```

```
# Hex color codes for Dem Blue and Rep Red
party_colors <- c("#2E74C0", "#CB454A")
```

```
p <- ggplot(data = subset(congress,
                           party %in% c("Democrat", "Republican") &
                               position == "U.S. Representative" &
                               pid %nin% exclude_pid),
             mapping = aes(x = yr_fac,
                           y = start_age,
                           color = party,
                           label = last))
```

```
p_out <- p + geom_quasirandom(size = 0.1, alpha = 0.4,  
                               method = "pseudorandom", dodge.width = 1) +  
  geom_line(data = mean_age_swarm,  
            mapping = aes(x = yr_fac, y = mean_age,  
                           color = party, group = party),  
            inherit.aes = FALSE, size = 1, alpha = 0.5) +  
  geom_ribbon(data = mean_age_swarm,  
              mapping = aes(x = yr_fac,  
                             ymin = lo,  
                             ymax = hi,  
                             color = NULL,  
                             fill = party,  
                             group = party),  
              inherit.aes = FALSE, alpha = 0.2) +  
  geom_text(data = oldest_group_by_year,  
            size = 0.9, alpha = 1,  
            position = position_jitter(width = 0.4, height = 0.4)) +  
  geom_text(data = youngest_group_by_year,  
            size = 0.9, alpha = 1,  
            position = position_jitter(width = 0.4, height = 0.4)) +  
  scale_x_discrete(breaks = levels(congress$yr_fac)[c(T, rep(F, 4))]) +  
  scale_color_manual(values = party_colors) +  
  scale_fill_manual(values = party_colors) +  
  guides(color = FALSE, fill = FALSE) +  
  labs(x = "Year", y = "Age",  
       title = "Age Distribution of Congressional Representatives, 1945-2019",  
       subtitle = "Trend line is mean age; bands are 25th and 75th percentiles  
                  of the range.\n\nYoungest and oldest percentiles are named  
                  instead of being shown by points.") +  
  facet_wrap(~ party, ncol = 1,  
            labeller = as_labeller(party_names)) +  
  theme(plot.subtitle = element_text(size = 10))
```

◀ The yearly point swarms

◀ The mean age line
◀ The hi-lo age ribbon

◀ The oldsters

◀ The youngsters

Slightly hacky
scale adjustment

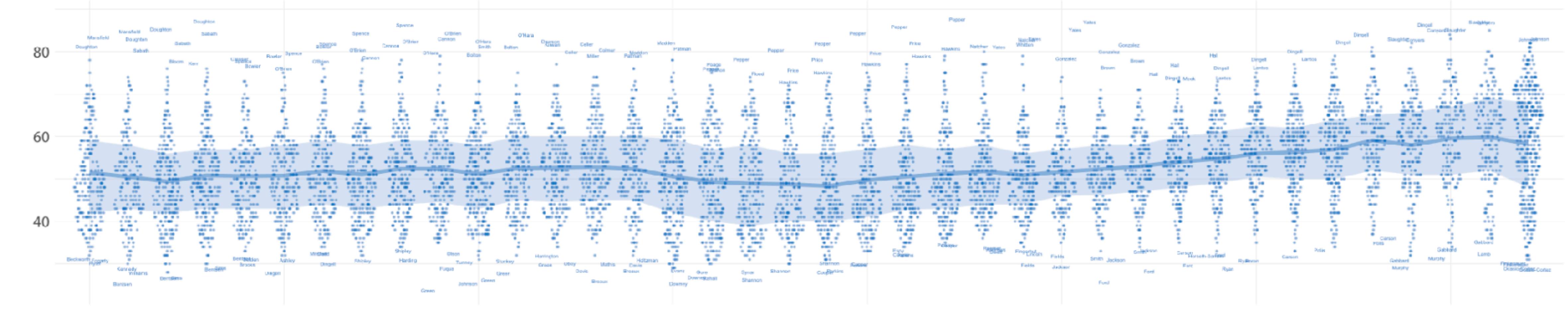
◀ Facet and label

Age Distribution of Congressional Representatives, 1945-2019

Trend line is mean age; bands are 25th and 75th percentiles of the range.

Youngest and oldest percentiles are named instead of being shown by points.

Democrats



Age

Republicans



Year

MORTALITY IN FRANCE 1816 – 2016

100 | Males

75

50

25

0

100 | Females

75

50

25

0

1820

1845

1870

1895

1920

1945

1970

1995

2015

Made with R and ggplot2. Source data: mortality.org

Death Rate Percentile



Mortality rates are calculated for each age in each year and binned by percentile. The darker the color at any particular point, the more people of that age die in that year. The lighter the color, the more people of that age survive in that year.

Historical trends are visible, such as the rapid decrease in infant mortality rates after World War II, as well as increased life expectancy overall. Specific events show up as vertical streaks in the graph. The death toll due to wars is evident for Males. Pandemics are also visible, most notably the 1918 Influenza pandemic and the death toll due to Smallpox outbreaks after the Franco-Prussian war of 1870.

Diagonal streaks in the data are visible in some parts of the graph. These are artifacts due to the estimation of the mortality rate in some years. Single-year-of-age figures prior to 1900 are calculated from five-year age groups, as no single-year data is available in the original mortality tabulations from which the rates were derived.

Kieran Healy / socviz.co
Data Visualization: A Practical Introduction
is published by Princeton University Press

```
drat::addRepo("kjhealy")
install.packages("demog")
library(demog)
```

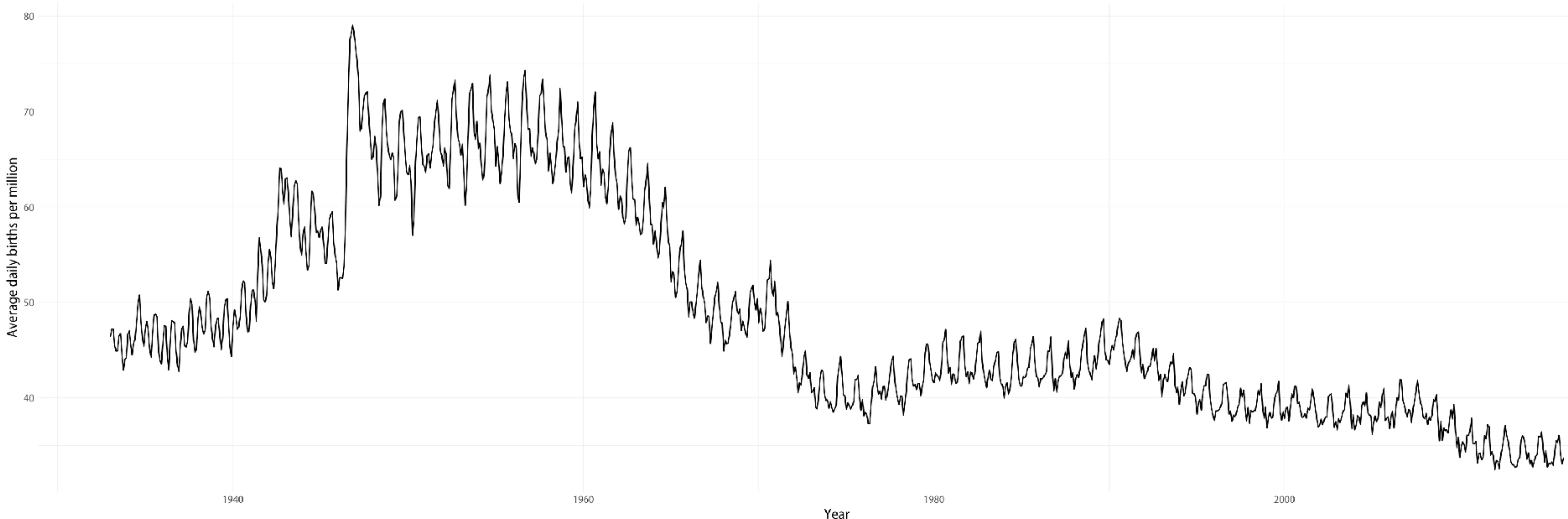
demog

Birth and Death rate data

```
... -->
> okboomer
# A tibble: 1,644 x 12
  year month n_days births total_pop births_pct births_pct_day date
  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <date>
1 1938     1     31 51820 41215000 0.00126      40.6 1938-01-01
2 1938     2     28 47421 41215000 0.00115      41.1 1938-02-01
3 1938     3     31 54887 41215000 0.00133      43.0 1938-03-01
4 1938     4     30 54623 41215000 0.00133      44.2 1938-04-01
5 1938     5     31 56853 41215000 0.00138      44.5 1938-05-01
6 1938     6     30 53145 41215000 0.00129      43.0 1938-06-01
7 1938     7     31 53214 41215000 0.00129      41.6 1938-07-01
8 1938     8     31 50444 41215000 0.00122      39.5 1938-08-01
9 1938     9     30 50545 41215000 0.00123      40.9 1938-09-01
10 1938    10     31 50079 41215000 0.00122     39.2 1938-10-01
# ... with 1,634 more rows
```

```
okboomer %>%
```

```
  filter(country == "United States") %>%
  ggplot(aes(x = date, y = births_pct_day)) +
  geom_line(size = 0.5) +
  labs(x = "Year",
       y = "Average daily births per million")
```



```

okboomer <- okboomer %>%
  mutate(year_fct = factor(year,
                           levels = unique(year),
                           ordered = TRUE),
        month_fct = factor(month,
                           levels = rev(c(1:12)),
                           labels = rev(c("Jan", "Feb", "Mar", "Apr",
                                         "May", "Jun", "Jul", "Aug",
                                         "Sep", "Oct", "Nov", "Dec")),
                           ordered = TRUE)) %>%
  select(year, month,
         year_fct, month_fct, everything())

```

```

> okboomer
# A tibble: 1,644 x 14
   year month year_fct month_fct n_days births total_pop births_pct births_pct_day date
   <dbl> <dbl> <ord>    <ord>    <dbl> <dbl>    <dbl>    <dbl> <dbl> <date>
 1 1938     1 1938      Jan       Jan     31 51820 41215000 0.00126  40.6 1938-01-01
 2 1938     2 1938      Feb       Feb     28 47421 41215000 0.00115  41.1 1938-02-01
 3 1938     3 1938      Mar       Mar     31 54887 41215000 0.00133  43.0 1938-03-01
 4 1938     4 1938      Apr       Apr     30 54623 41215000 0.00133  44.2 1938-04-01
 5 1938     5 1938      May       May     31 56853 41215000 0.00138  44.5 1938-05-01
 6 1938     6 1938      Jun       Jun     30 53145 41215000 0.00129  43.0 1938-06-01
 7 1938     7 1938      Jul       Jul     31 53214 41215000 0.00129  41.6 1938-07-01
 8 1938     8 1938      Aug       Aug     31 50444 41215000 0.00122  39.5 1938-08-01
 9 1938     9 1938      Sep       Sep     30 50545 41215000 0.00123  40.9 1938-09-01
10 1938    10 1938     Oct       Oct     31 50079 41215000 0.00122  39.2 1938-10-01
# ... with 1,634 more rows
> |

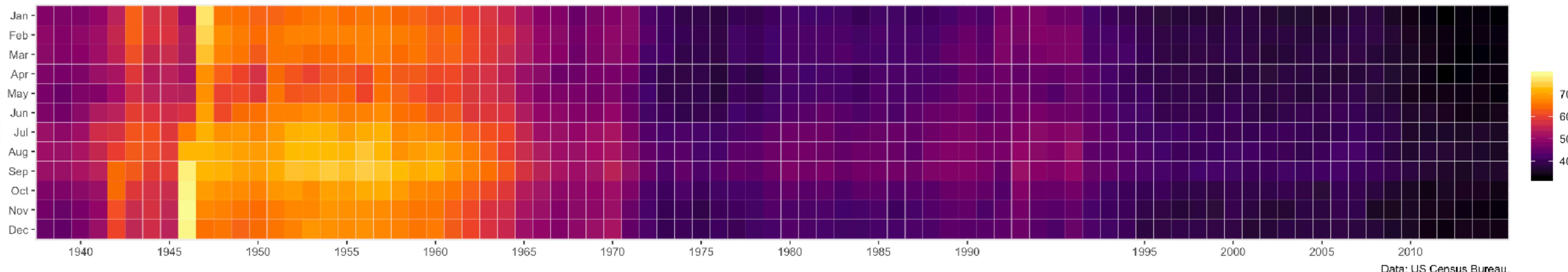
```

```
okboomer %>%
```

```
  filter(country == "United States") %>%
  ggplot(aes(x = year_fct, y = month_fct)) +
  geom_tile(mapping = aes(fill = births_pct_day),
            color = "white") +
  scale_x_discrete(breaks = seq(1940, 2010, 5)) +
  scale_y_discrete(position = "top") +
  scale_fill_viridis_c(option = "B") +
  labs(x = NULL, y = NULL, fill = NULL, title = "Monthly Birth Rates",
       subtitle = "Average births per million people per day.",
       caption = "Data: US Census Bureau.")
```

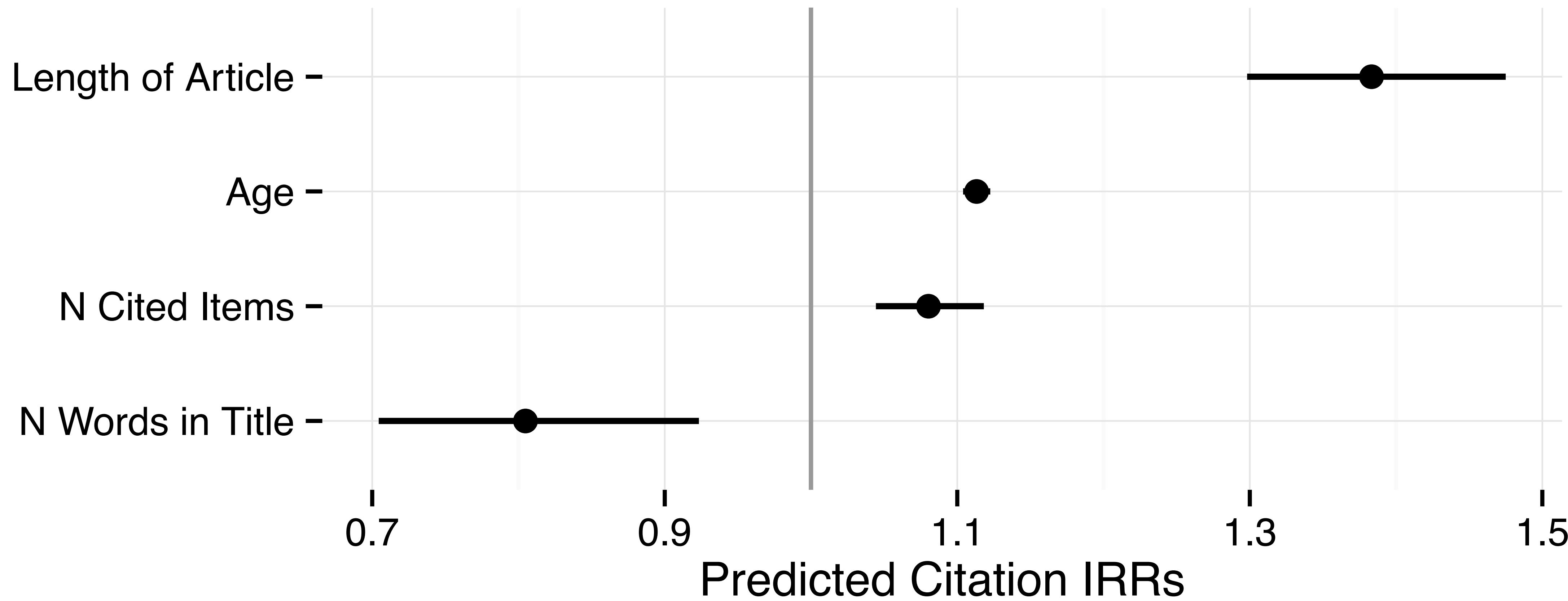
Monthly Birth Rates

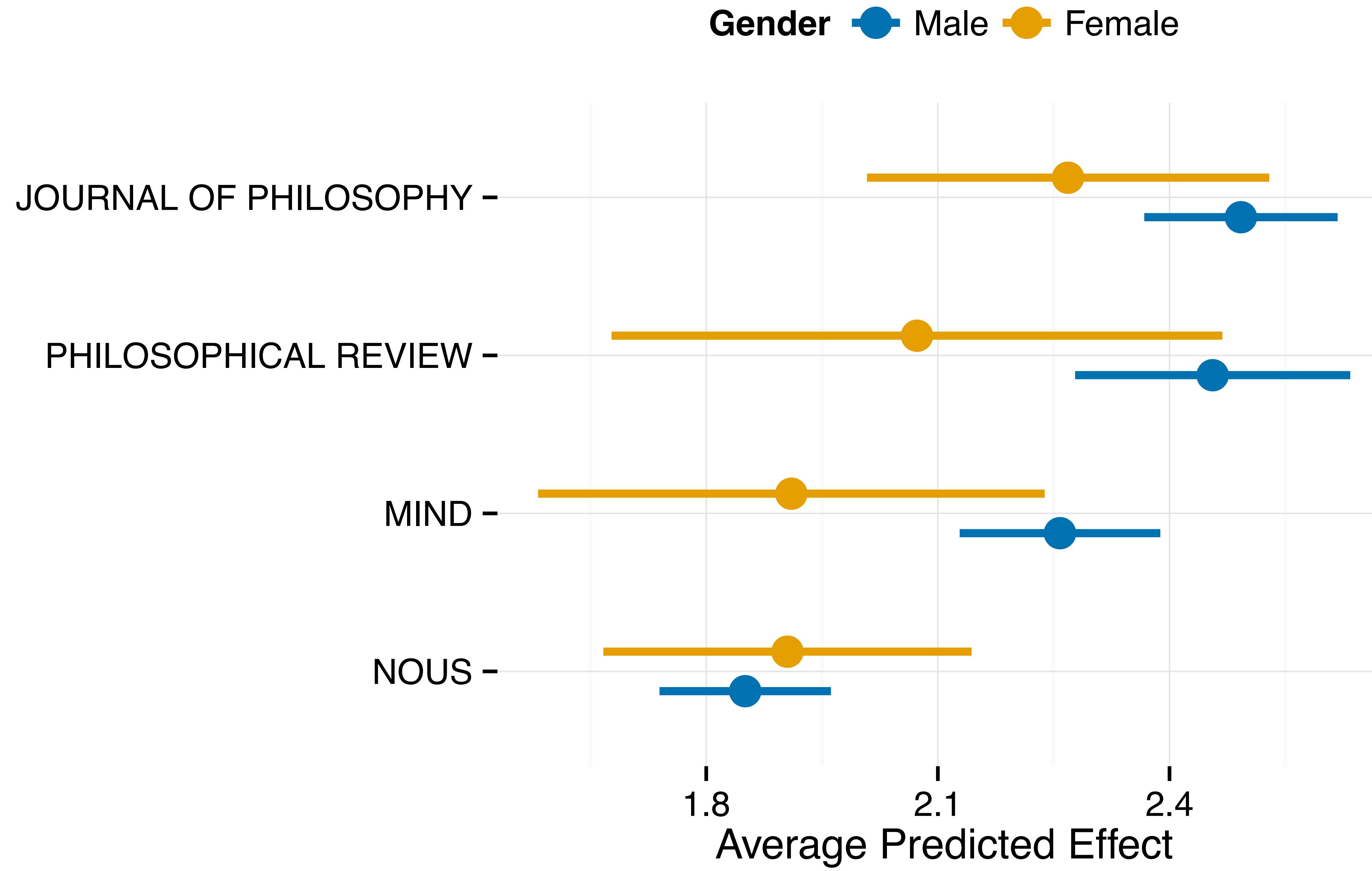
Average births per million people per day.



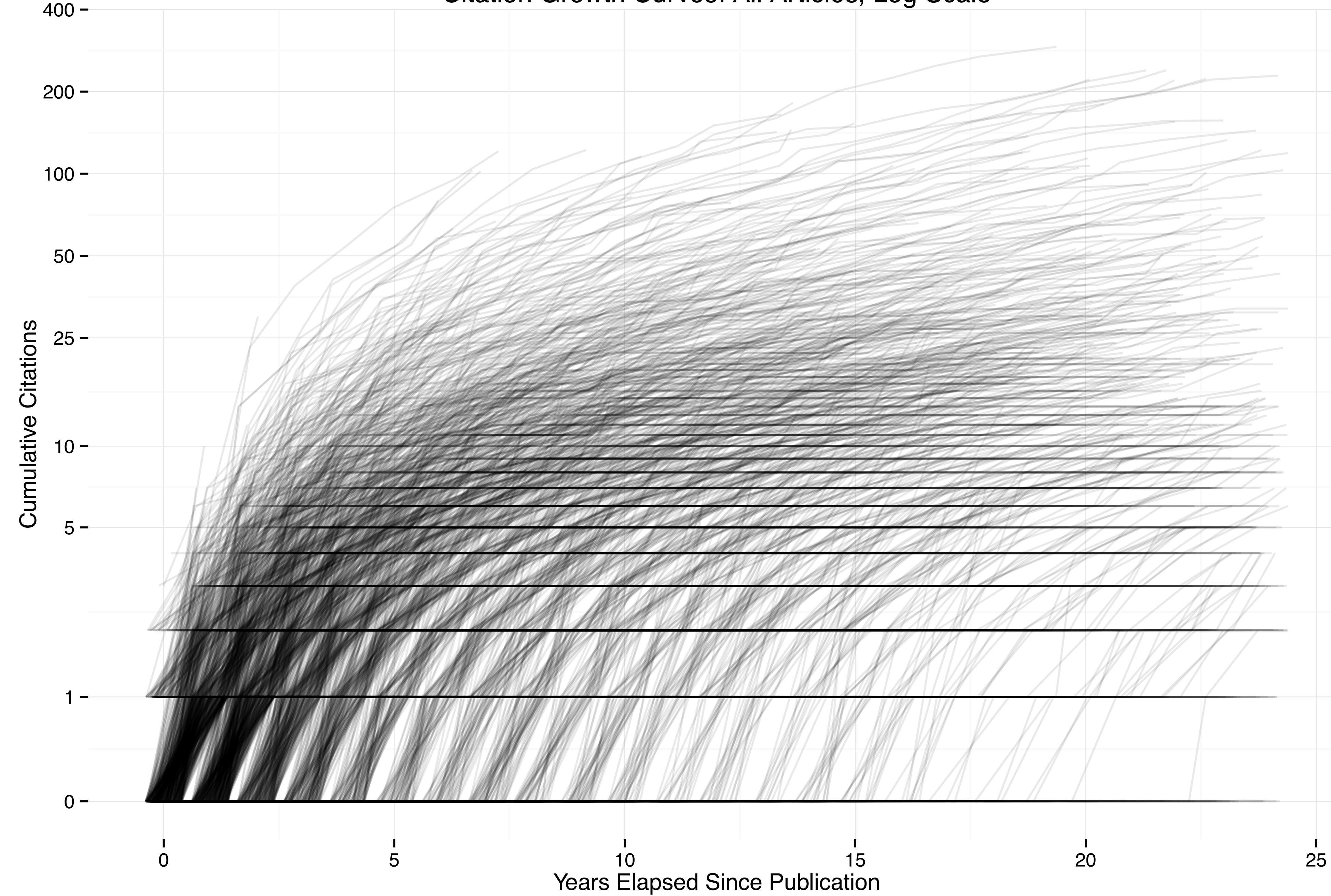
PRESENTING RESULTS

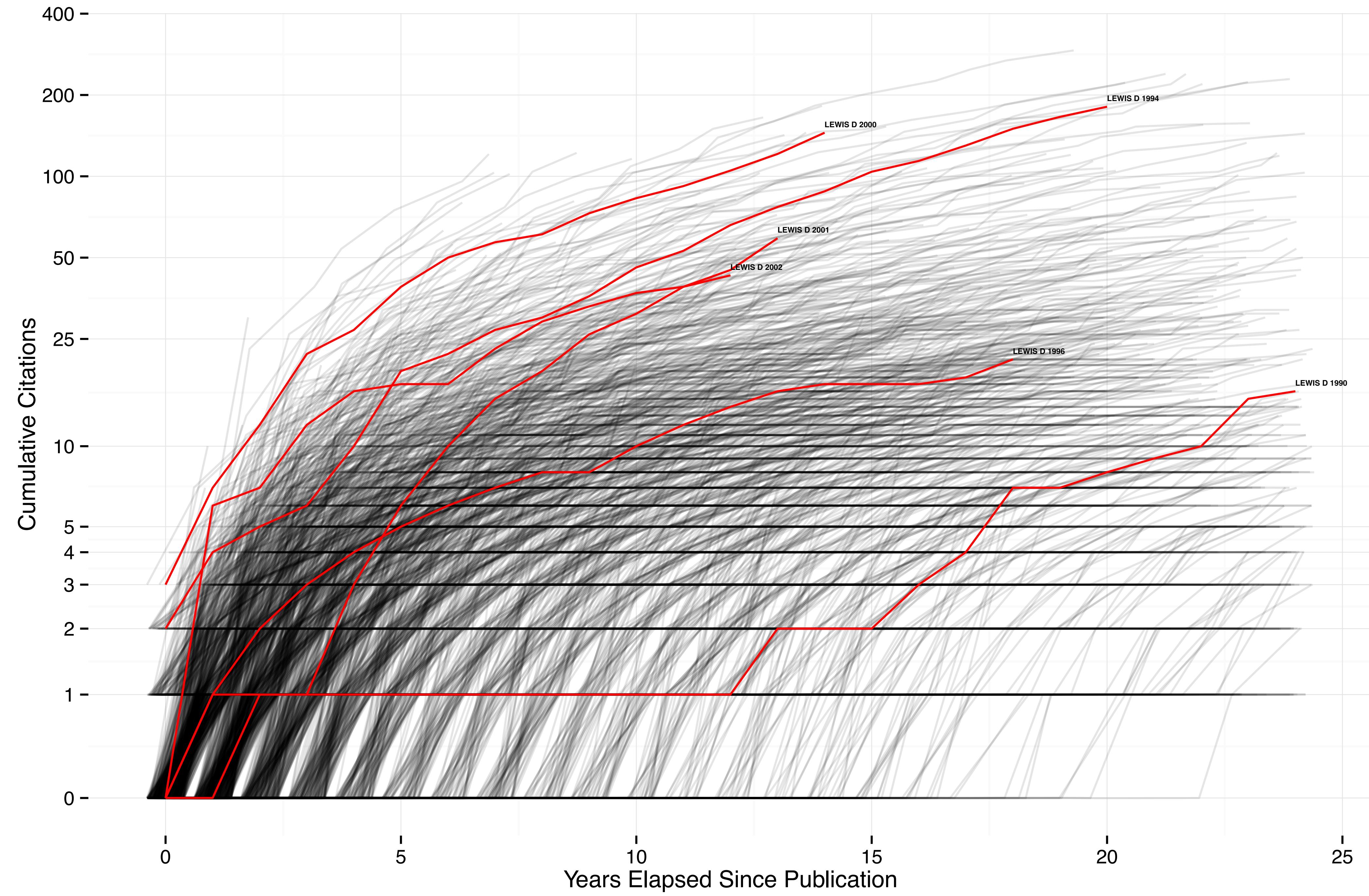
Leverage ggplot's
grammar, and its
layered approach

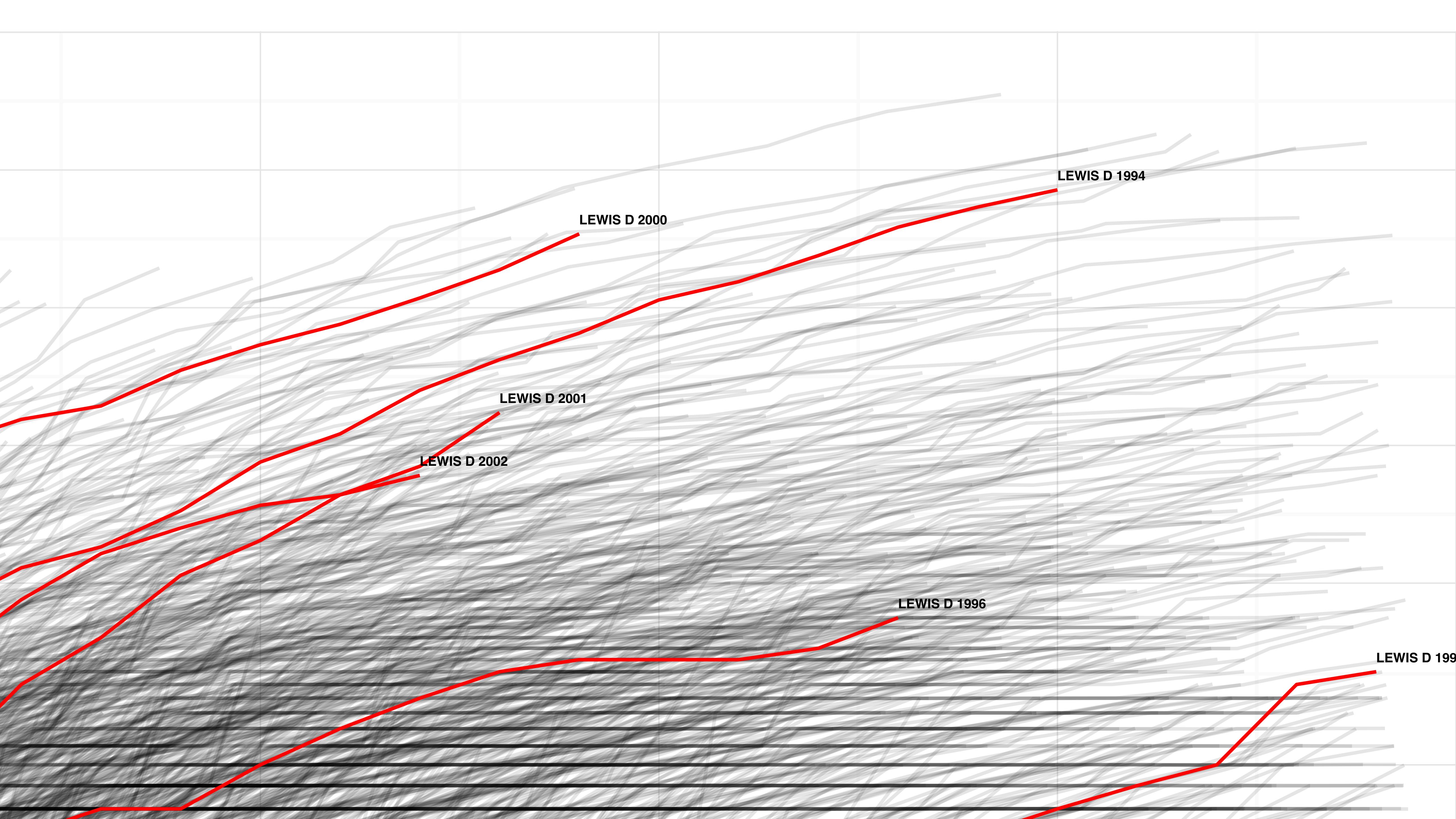


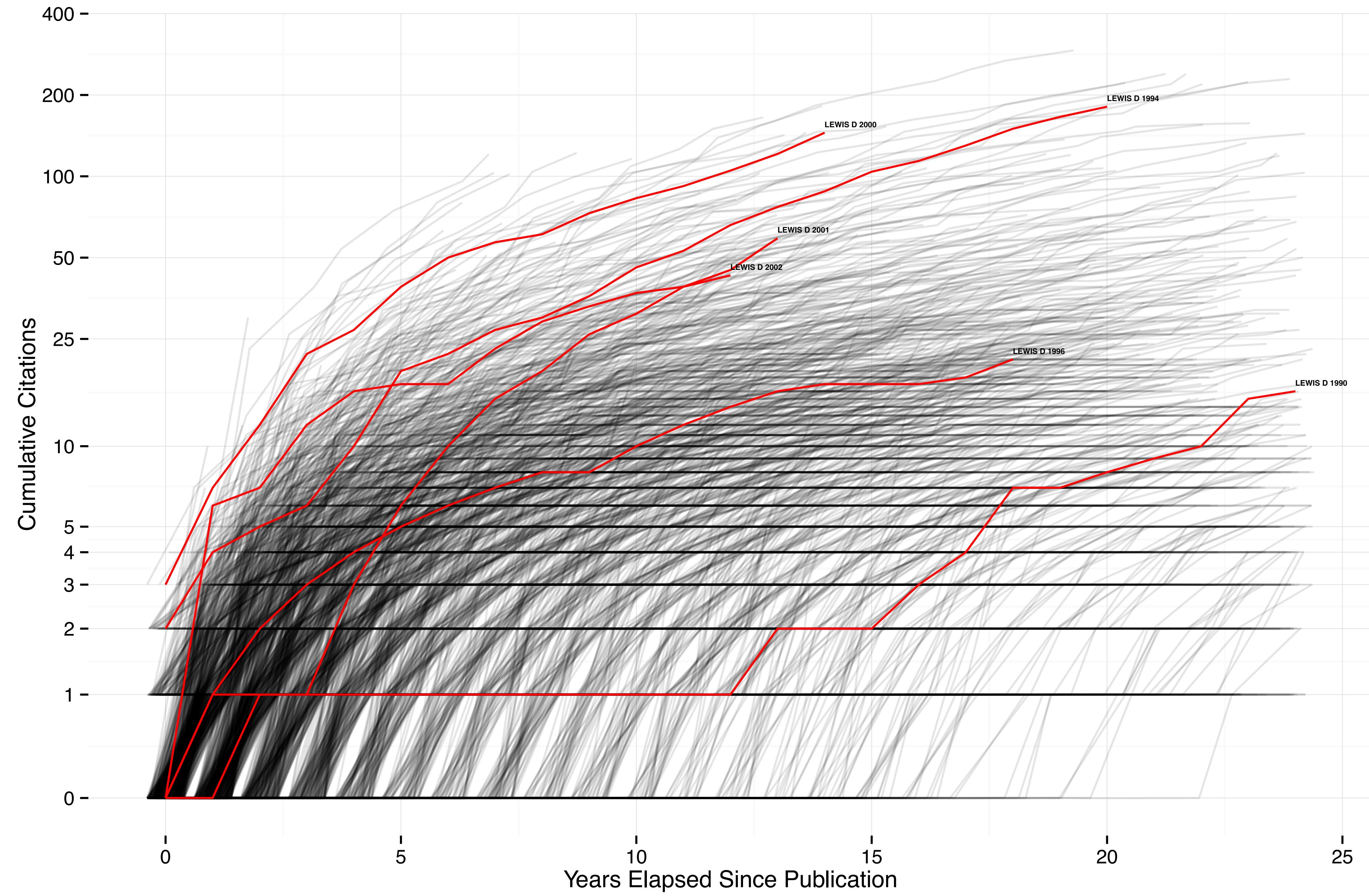


Citation Growth Curves: All Articles, Log Scale





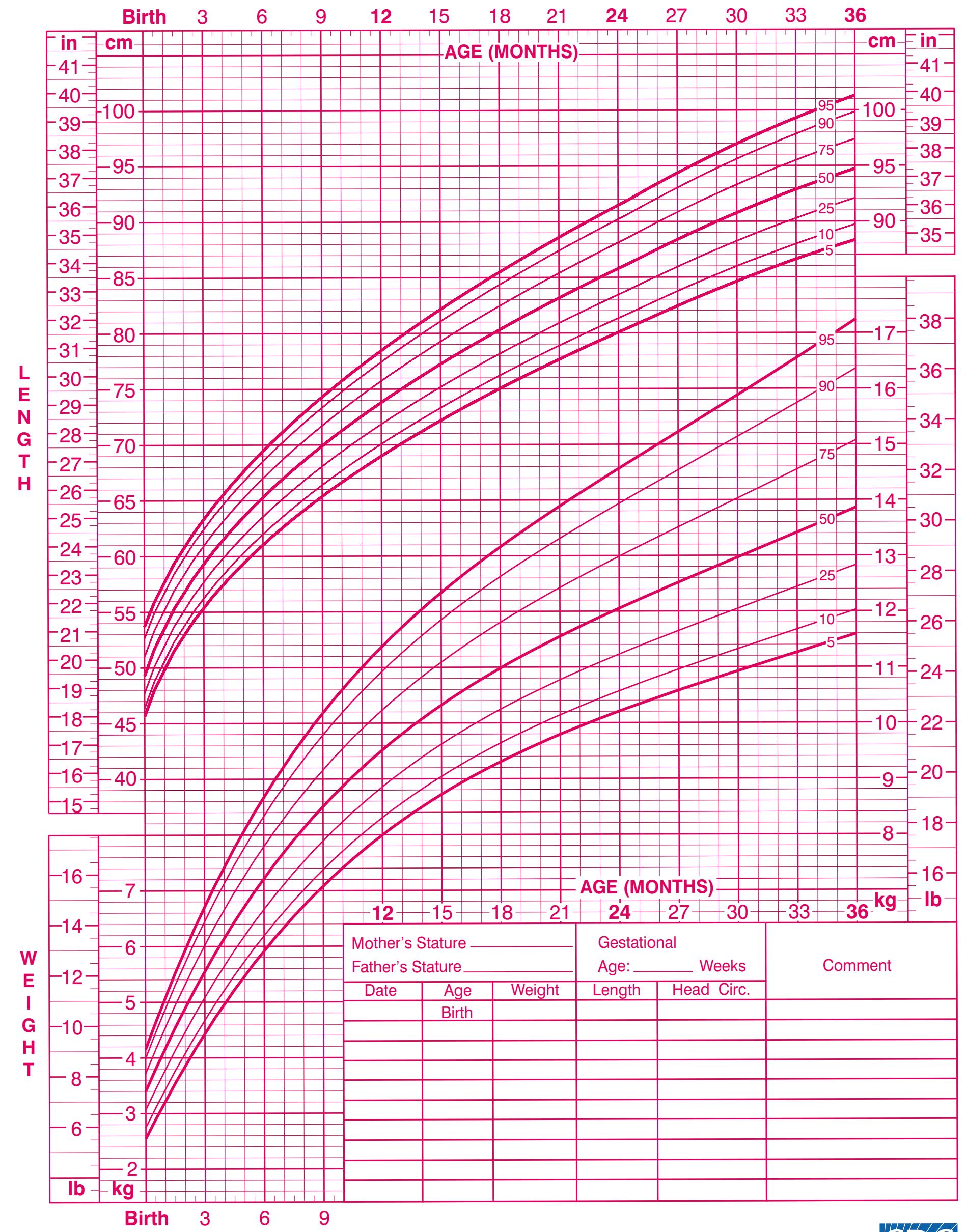




Birth to 36 months: Girls

Length-for-age and Weight-for-age percentiles

NAME _____ RECORD # _____



Published May 30, 2000 (modified 4/20/01).

SOURCE: Developed by the National Center for Health Statistics in collaboration with the National Center for Chronic Disease Prevention and Health Promotion (2000).
<http://www.cdc.gov/growthcharts>

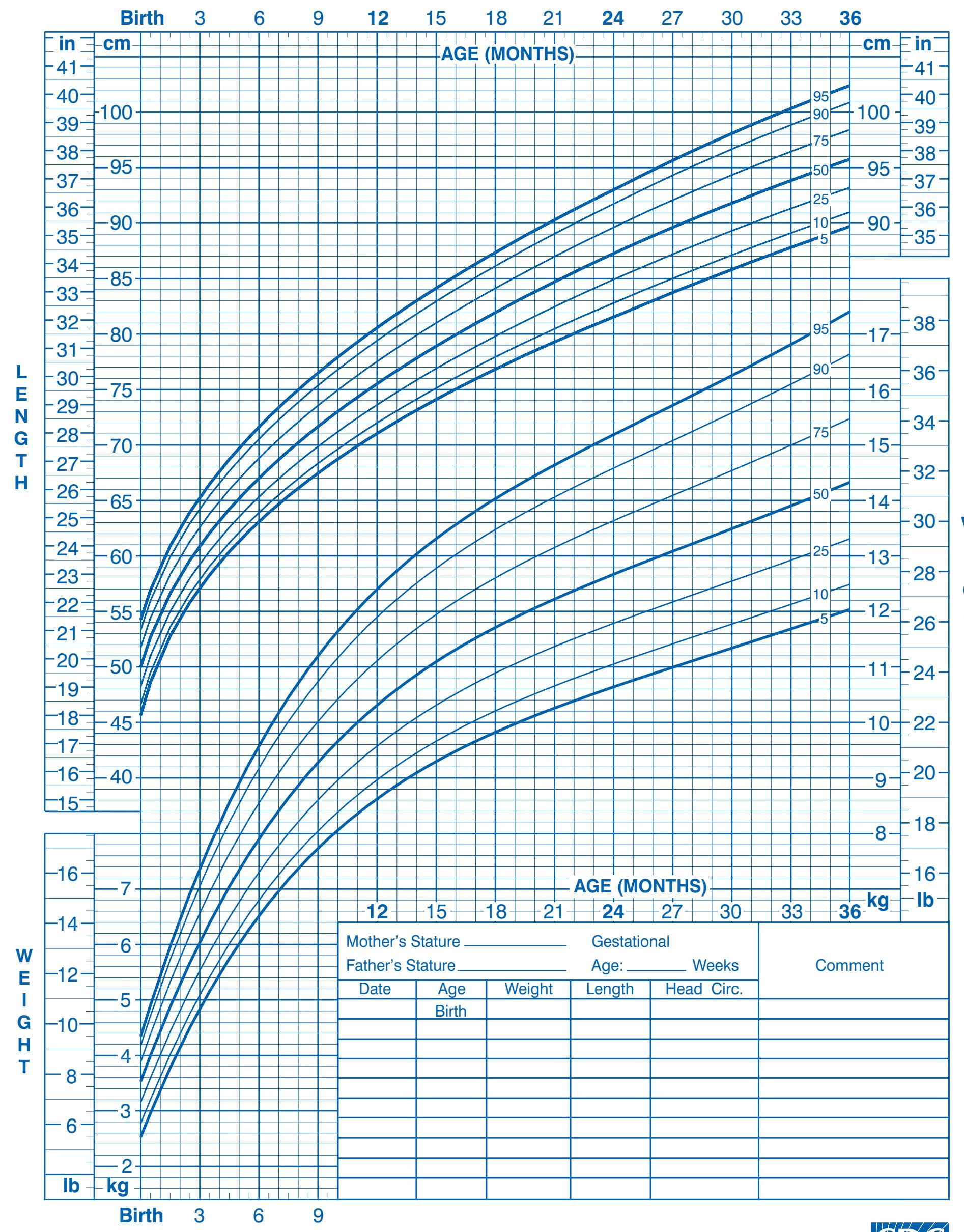


SAFER • HEALTHIER • PEOPLE™

Birth to 36 months: Boys

Length-for-age and Weight-for-age percentiles

NAME _____ RECORD # _____



Published May 30, 2000 (modified 4/20/01).

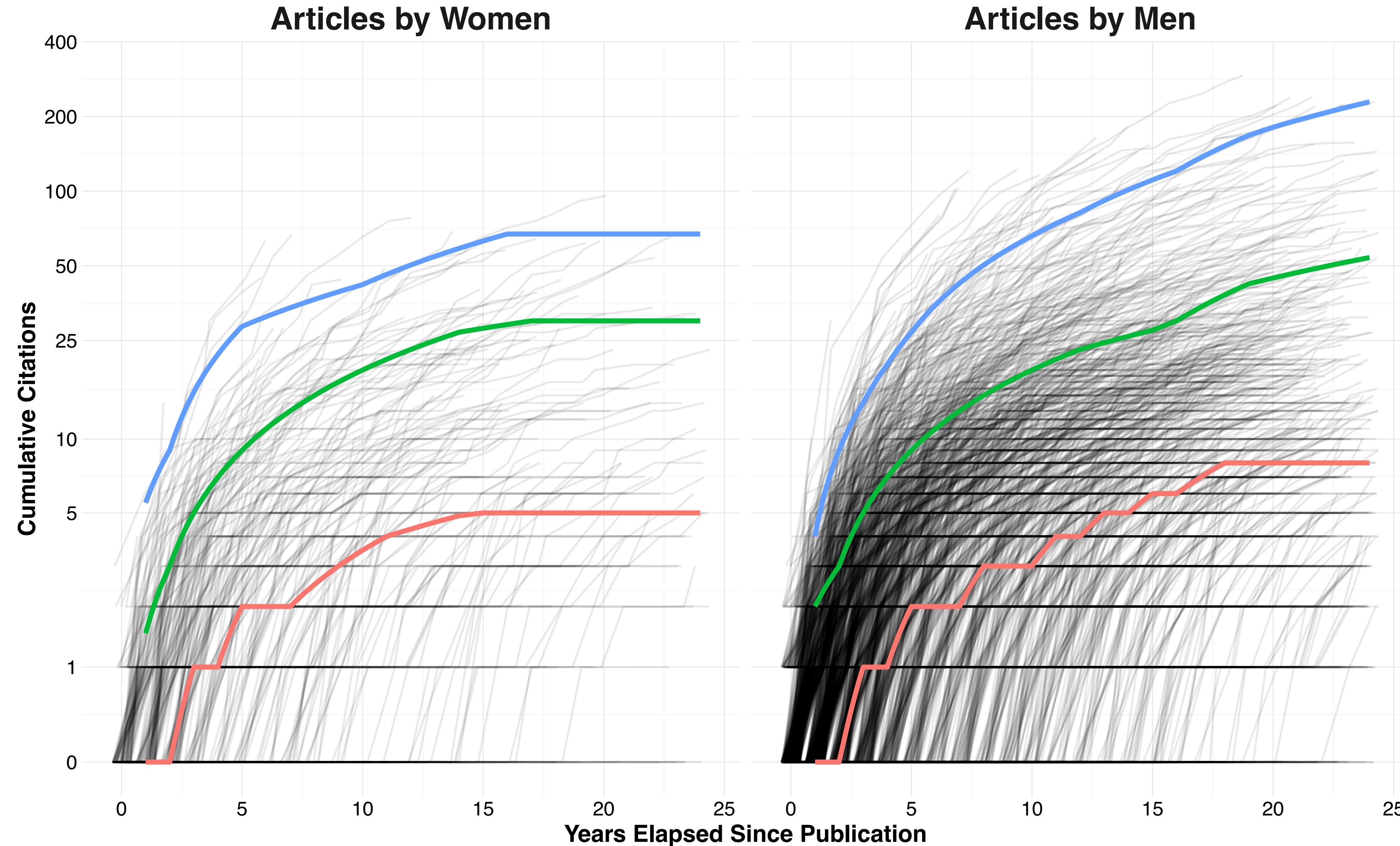
SOURCE: Developed by the National Center for Health Statistics in collaboration with the National Center for Chronic Disease Prevention and Health Promotion (2000).
<http://www.cdc.gov/growthcharts>



SAFER • HEALTHIER • PEOPLE™

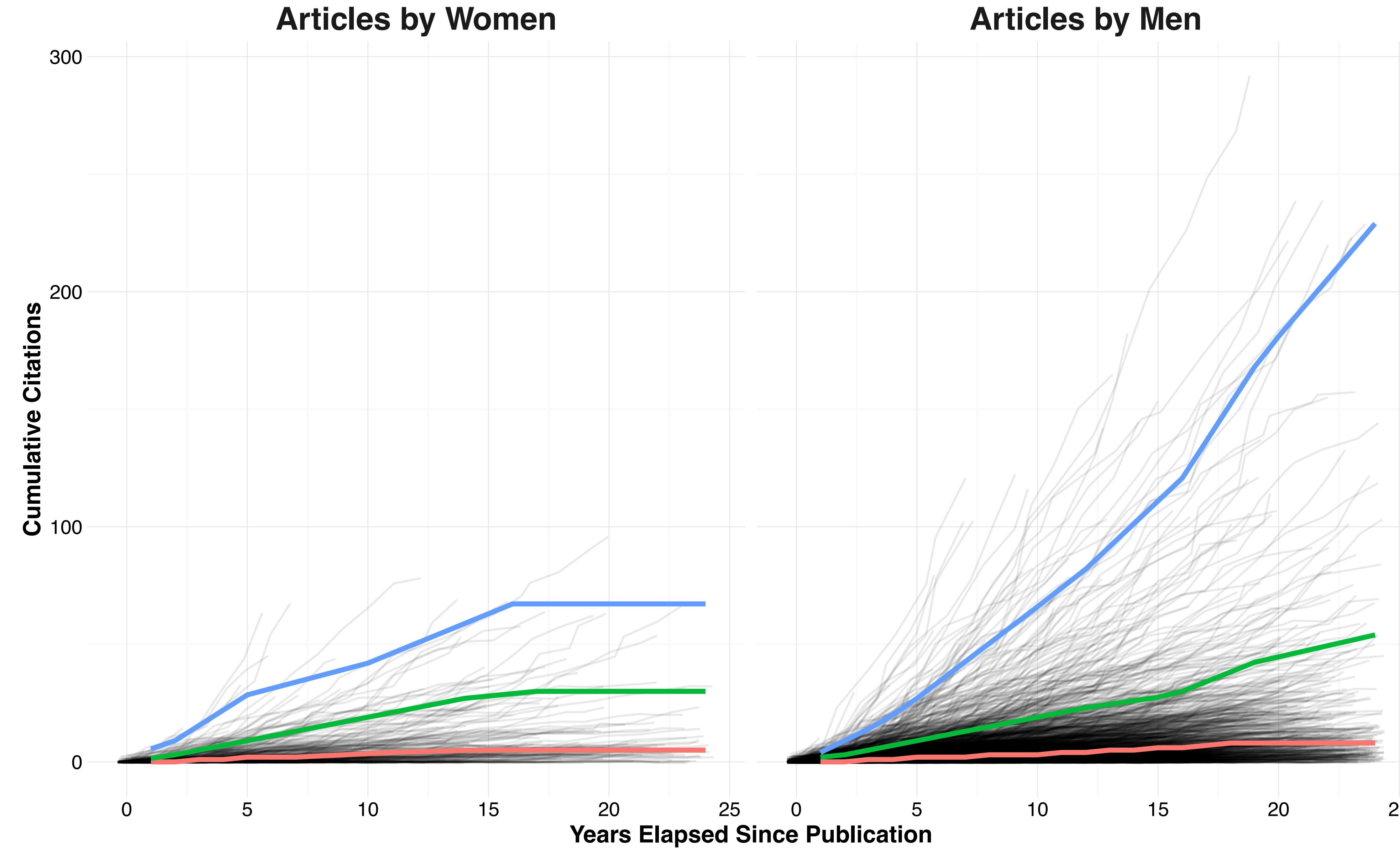
Citation Growth Curves for all Articles

Percentile — at the Median — at the 90th Percentile — in the top One Percent



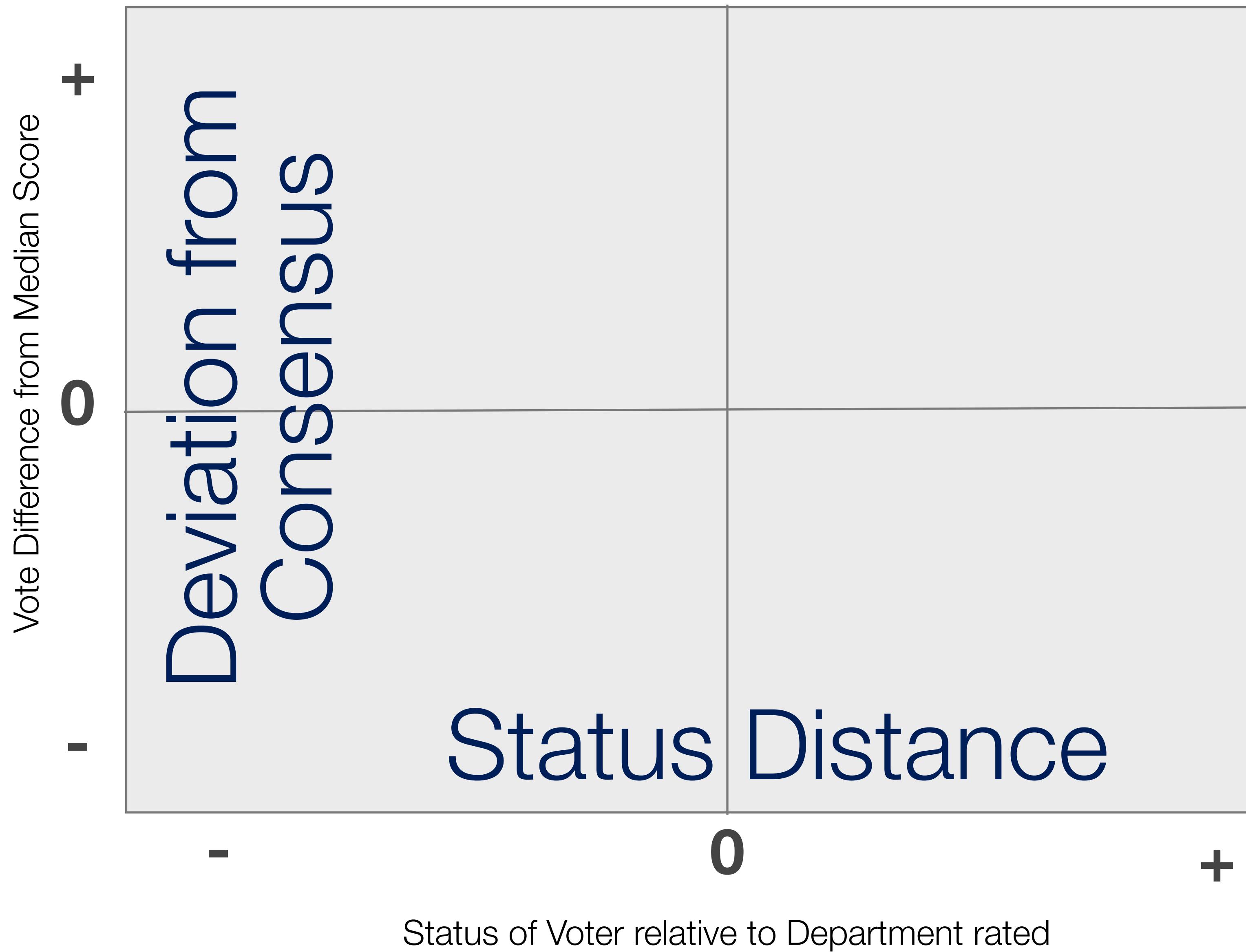
Citation Growth Curves for all Articles

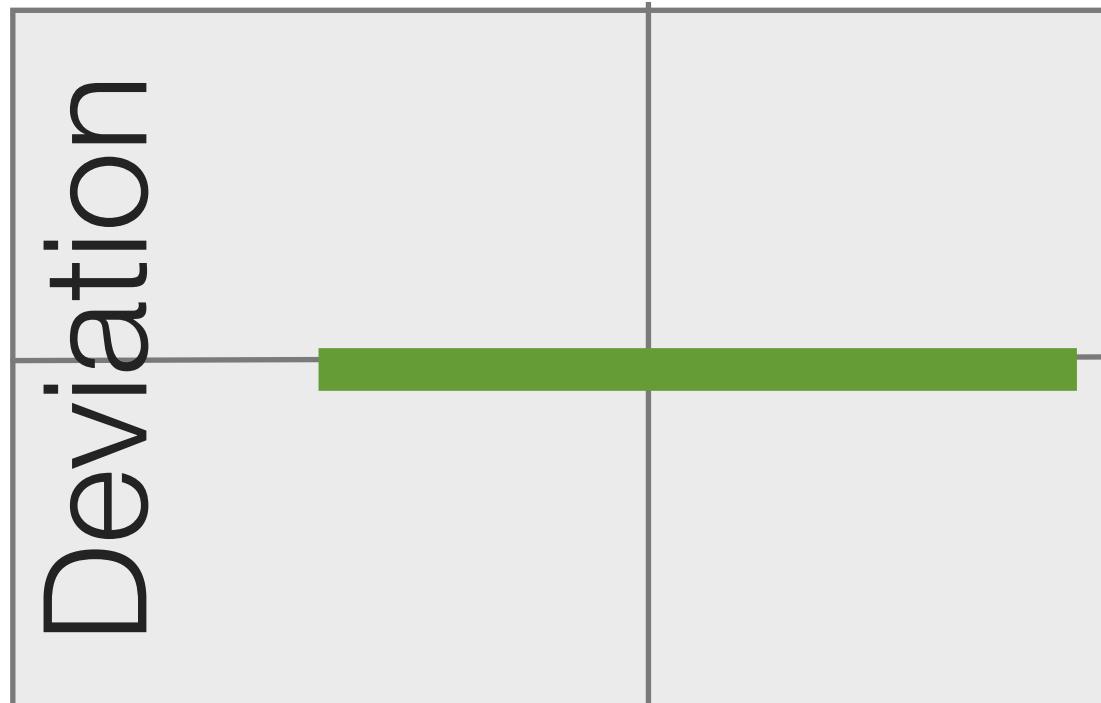
Percentile — at the Median — at the 90th Percentile — in the top One Percent



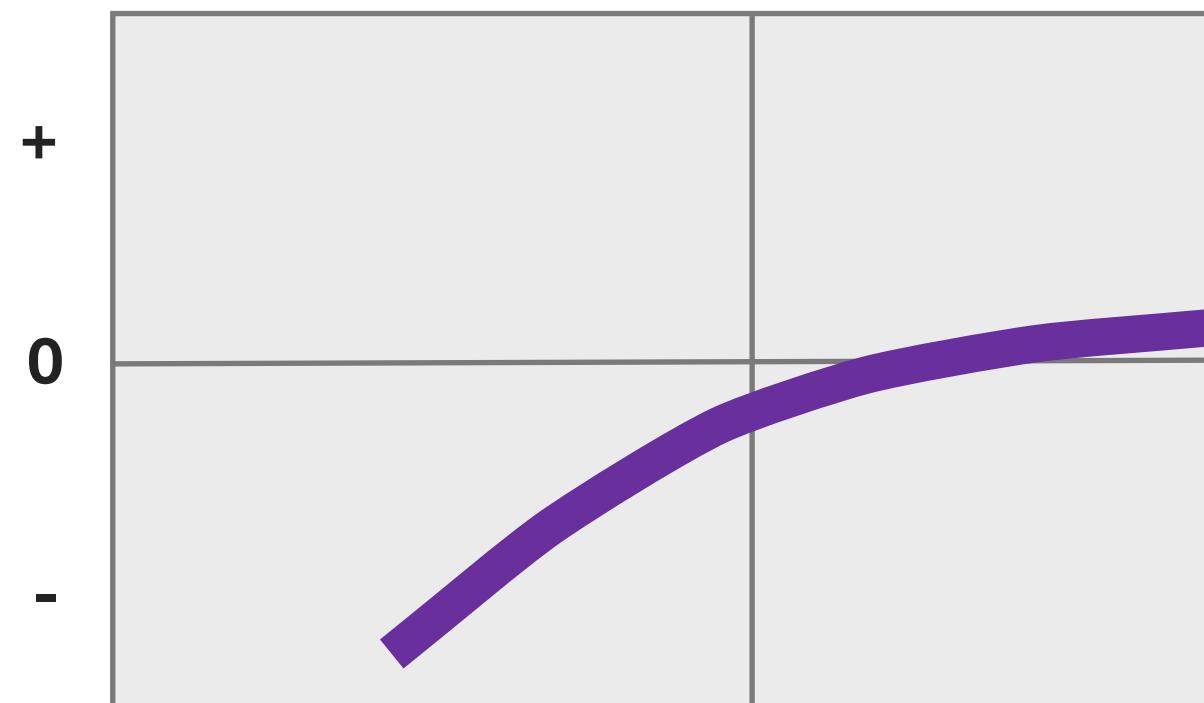
Layer
Highlight
Repeat

Build from Ideas to Data

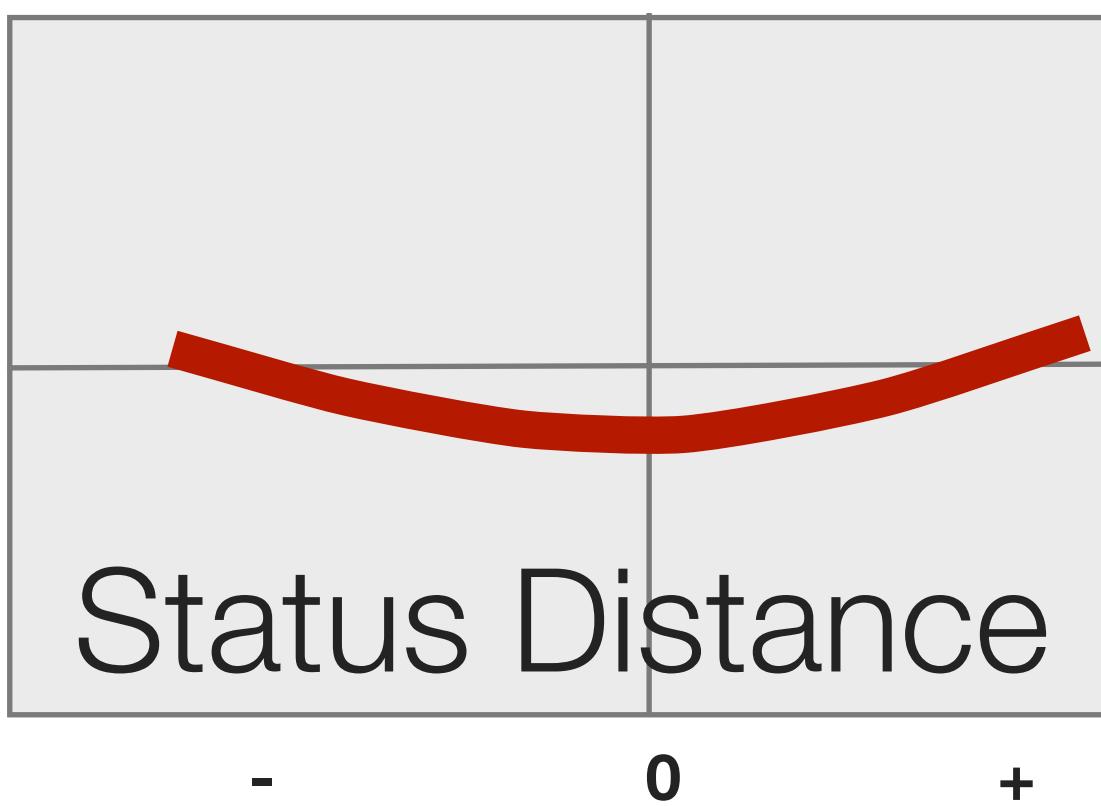




1. Pure Objectivity



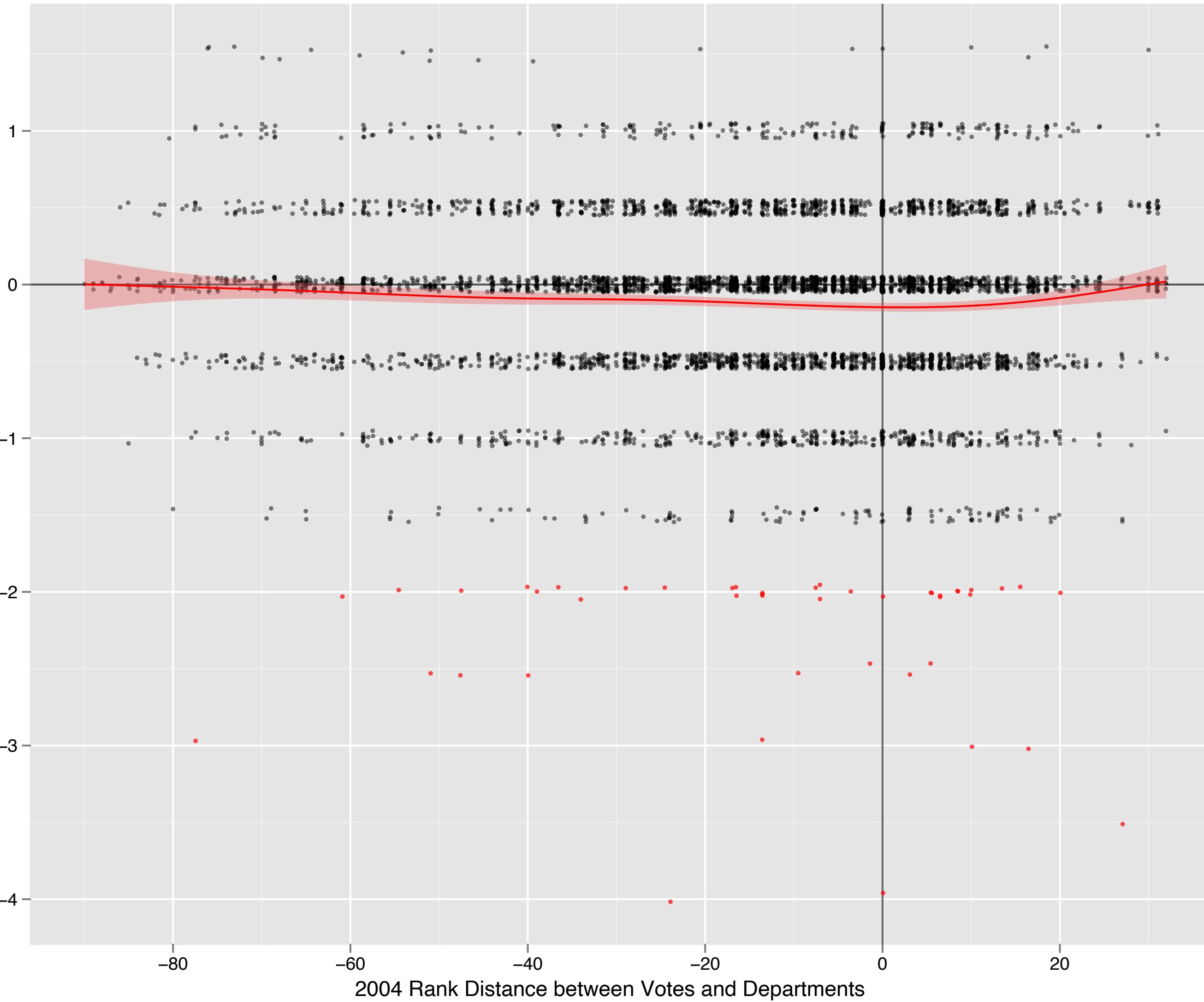
2. Distant Envy



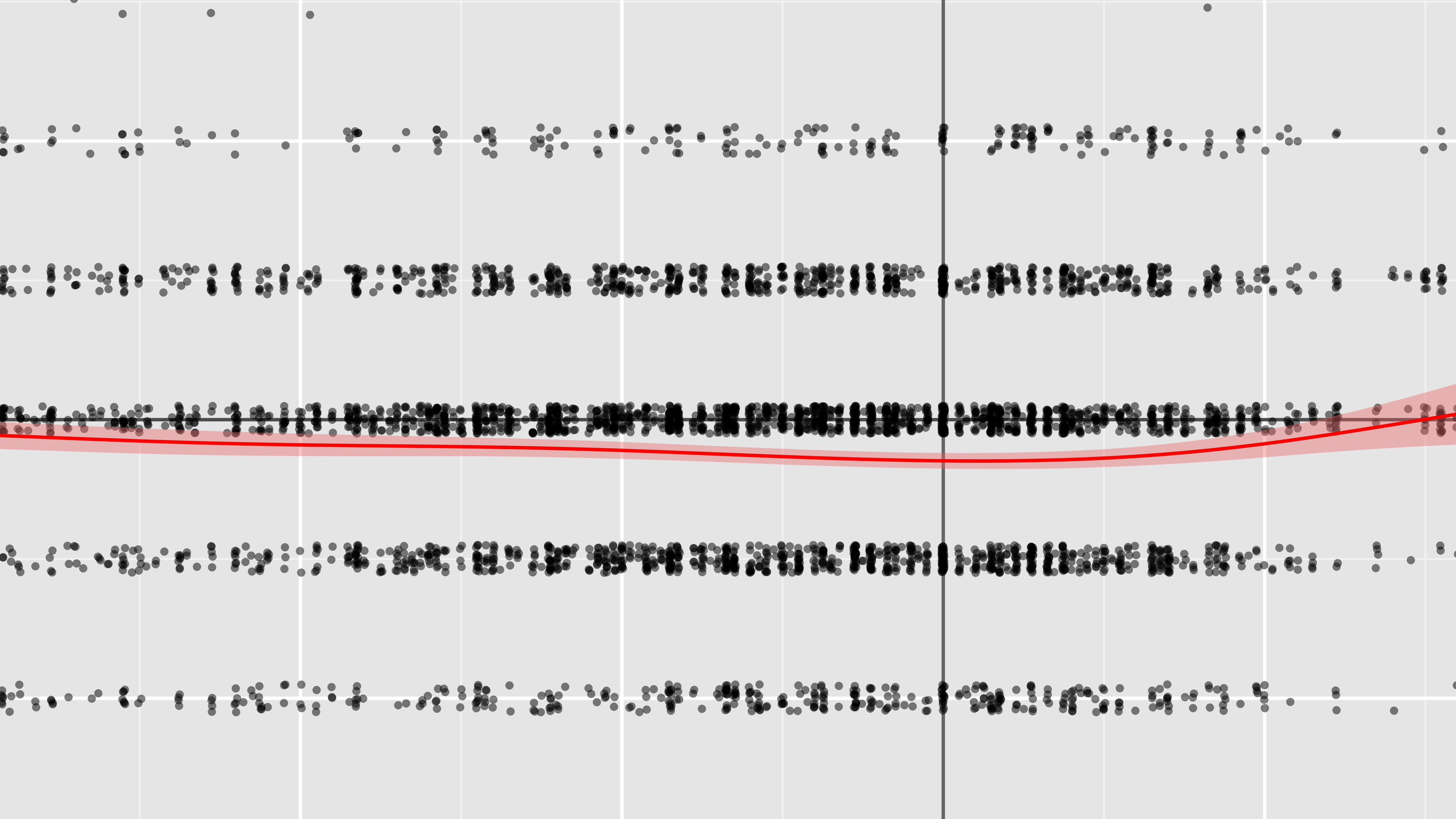
3. Local Competition

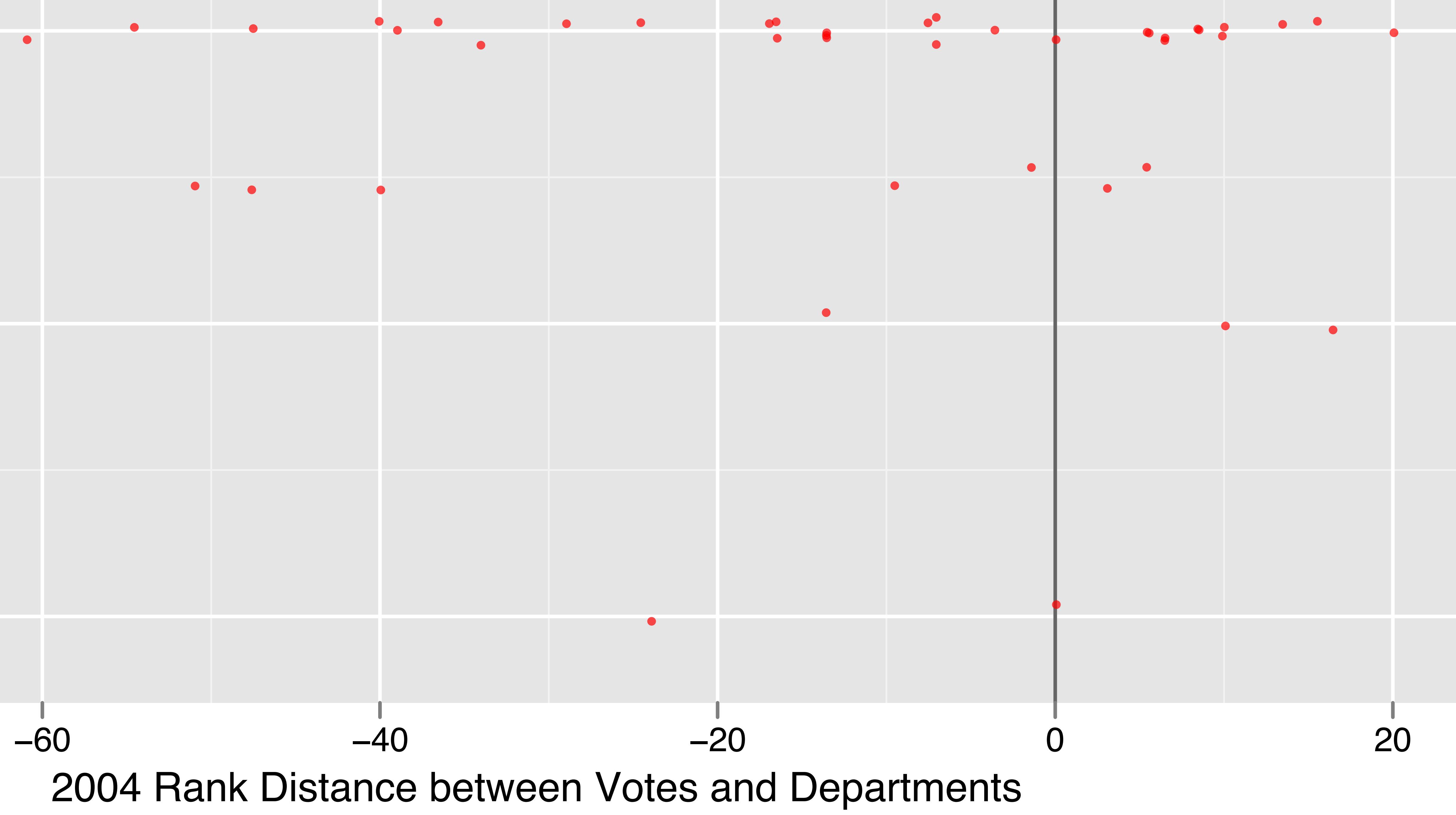
Deviation

Difference Between 2006 Votes and 2006 Median PGR Scores

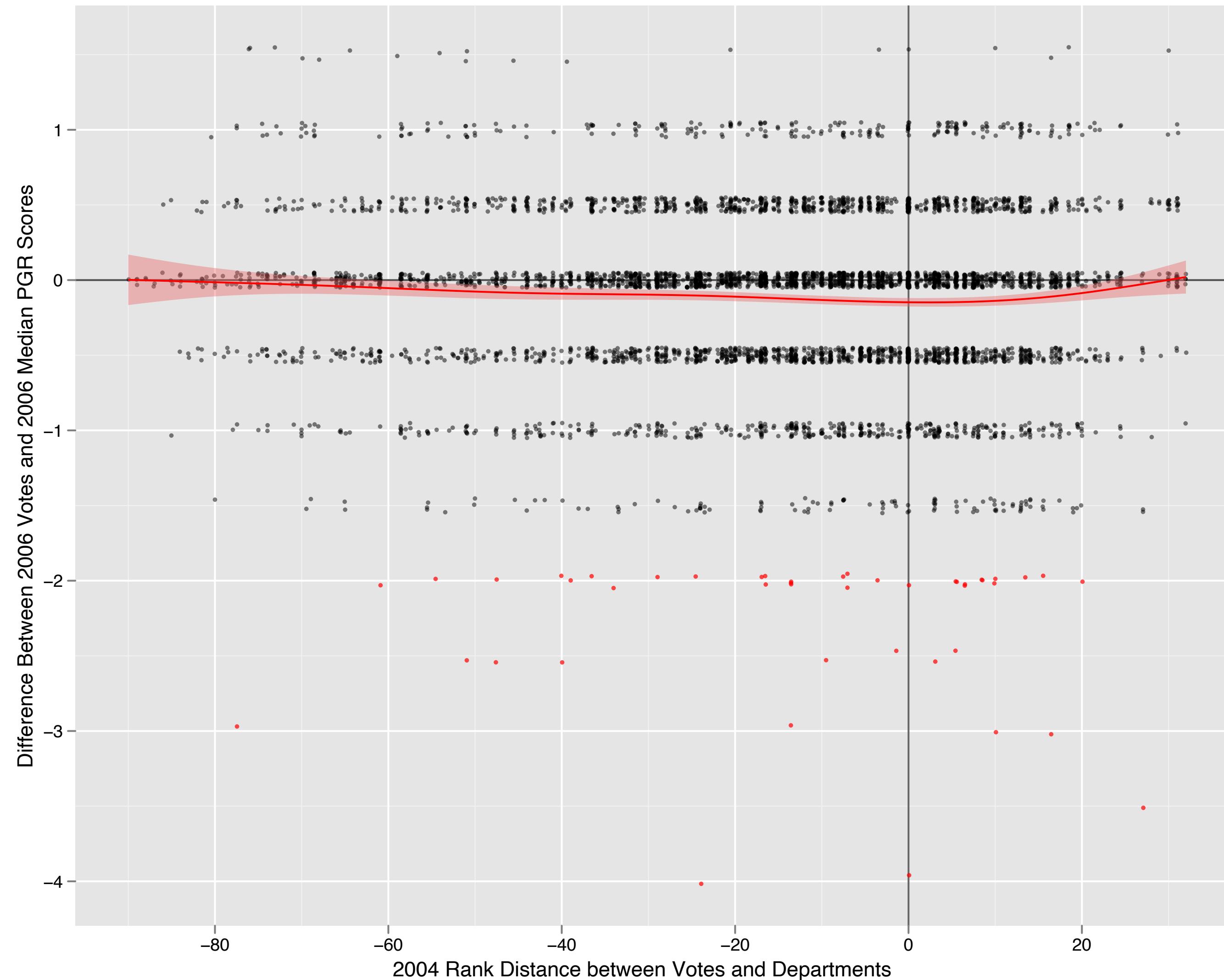


Status Distance



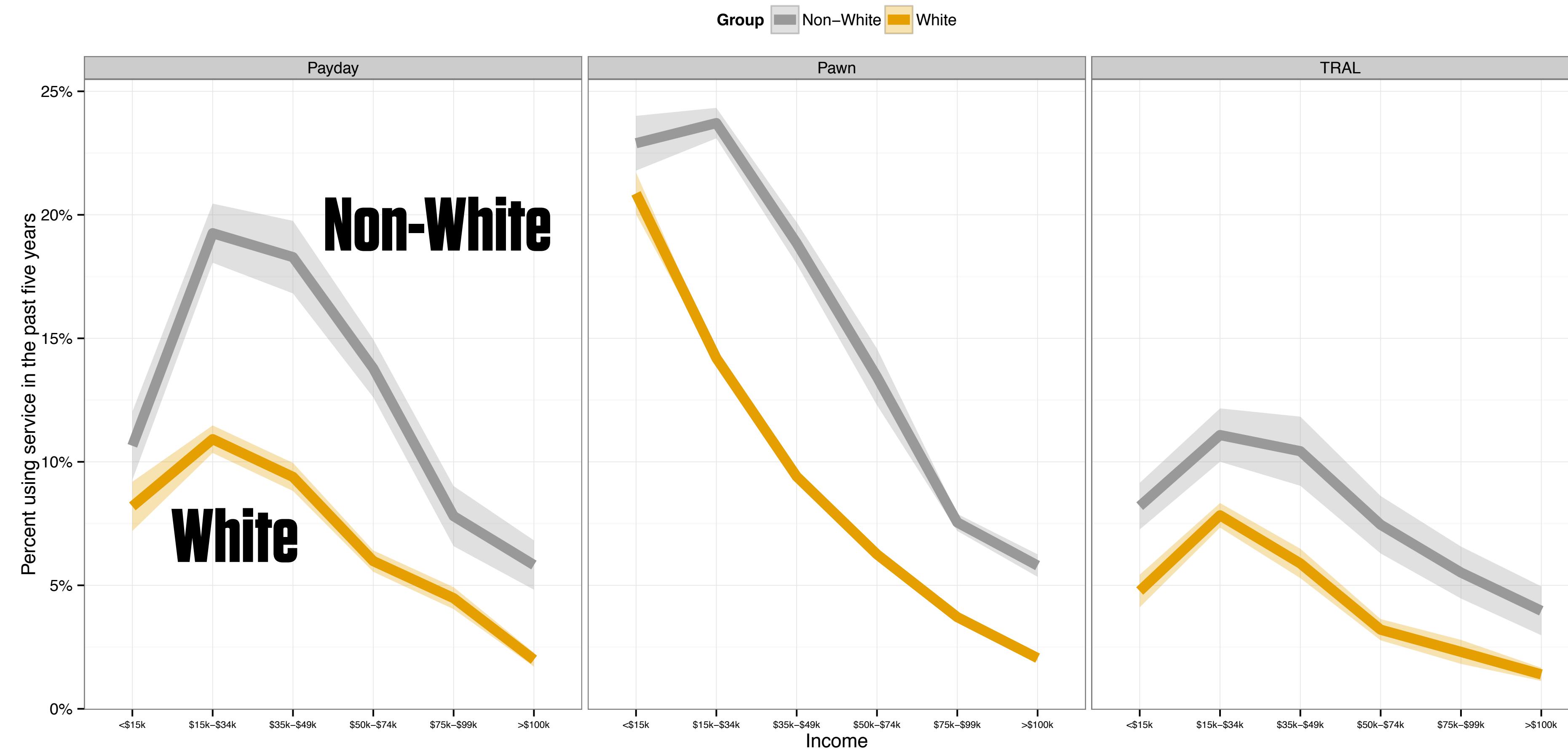


Deviation



Status Distance

Categorical Gaps: Alternative Financial Services

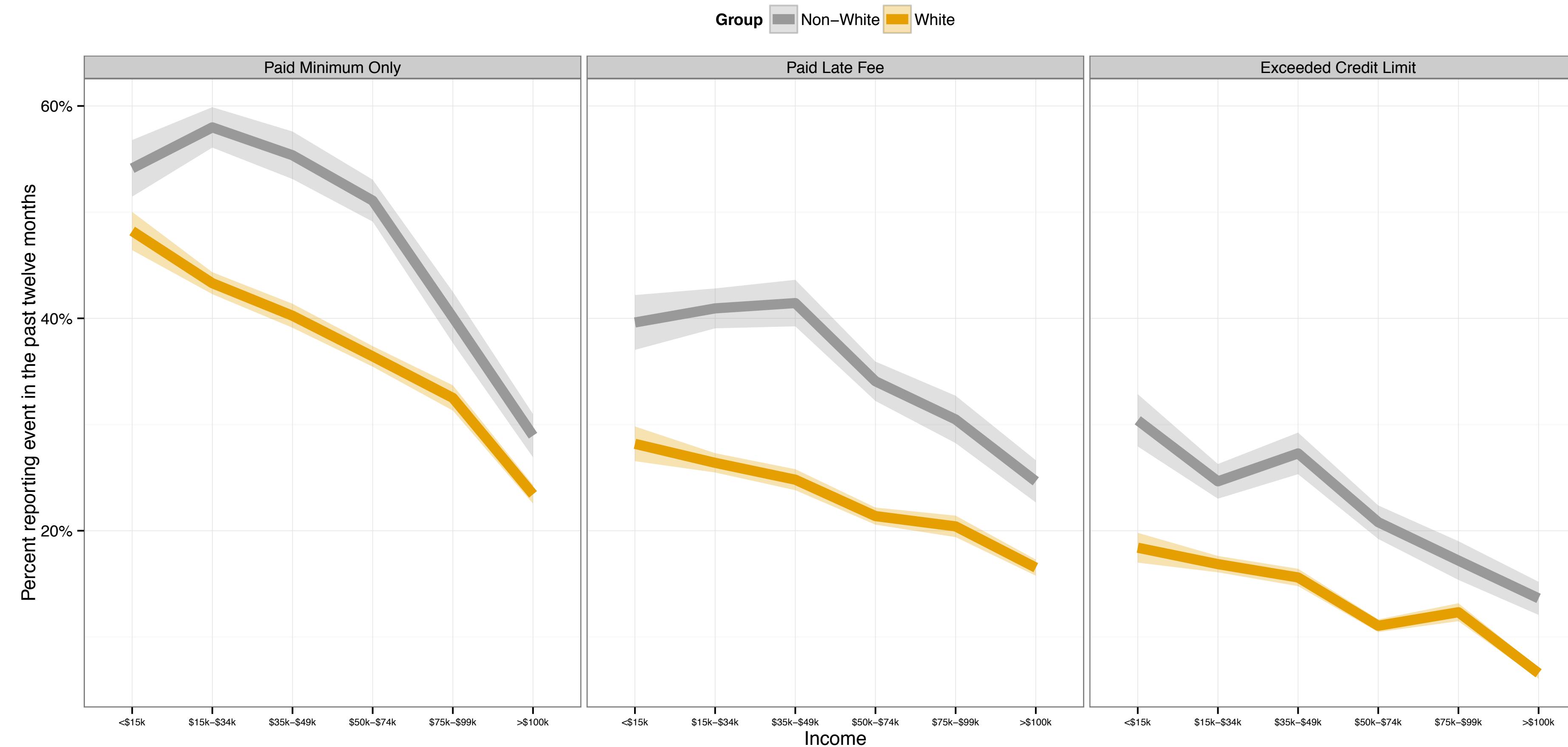


Payday Loan

Pawn Shop

TRAL

Categorical Gaps: Adverse Credit Events



Minimum Only

Late Fee

Over Limit