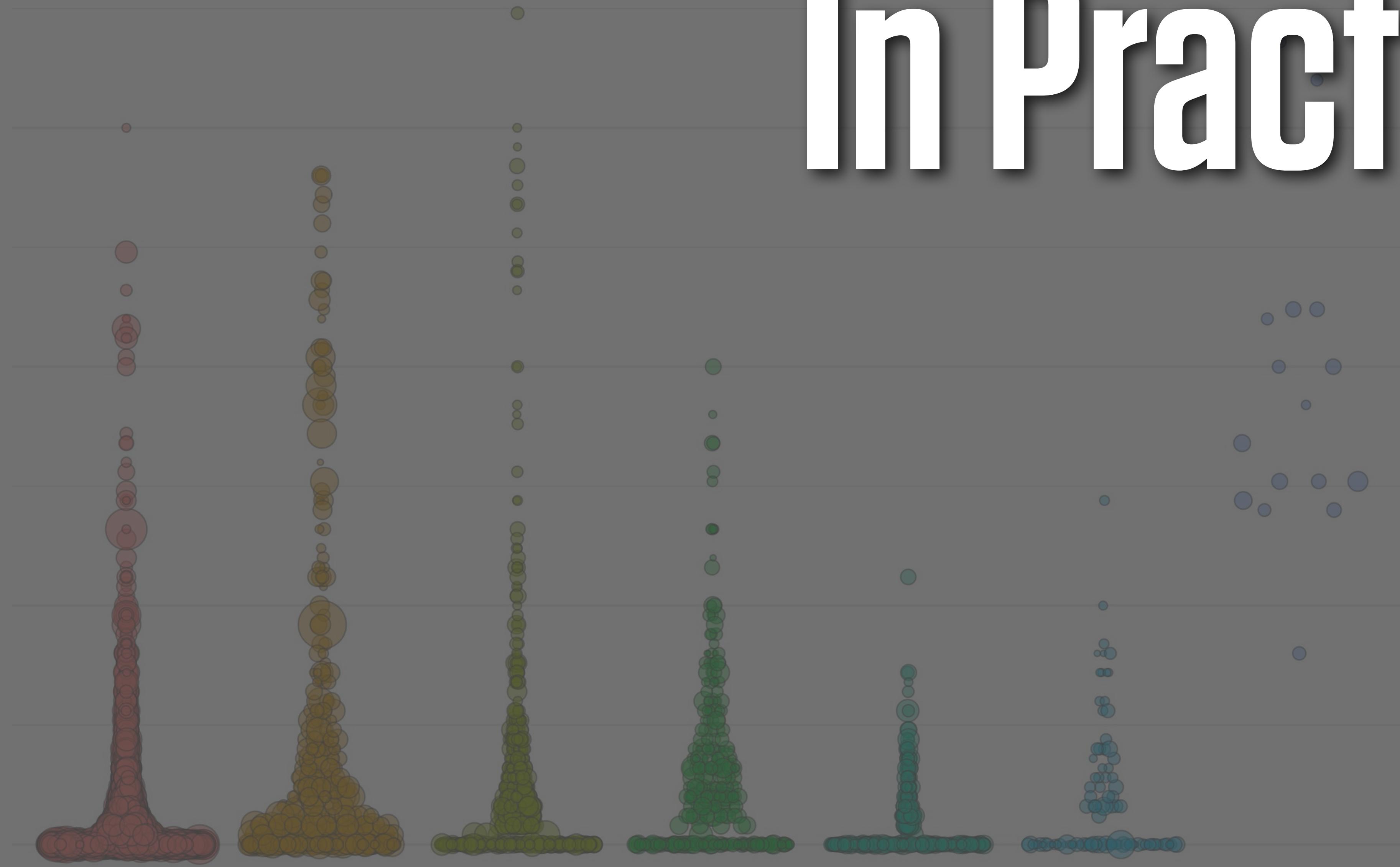


# Data Visualization



# In Practice



# Tidying Data

**Table A-1. Years of School Completed by People 25 Years and Over, by Age and Sex: Selected Years 1940 to 2016**

(Numbers in thousands. Noninstitutionalized population except where otherwise specified.)

Age, sex, and years	Total	Years of School Completed						
		Elementary		High school		College		Median
		0 to 4 years	5 to 8 years	1 to 3 years	4 years	1 to 3 years	4 years or more	

**25 YEARS AND OLDER**

**Male**

2016	103,372	1,183	3,513	7,144	30,780	26,468	34,283	(NA)
2015	101,887	1,243	3,669	7,278	30,997	25,778	32,923	(NA)
2014	100,592	1,184	3,761	7,403	30,718	25,430	32,095	(NA)
2013	99,305	1,127	3,836	7,314	30,014	25,283	31,731	(NA)
2012	98,119	1,237	3,879	7,388	30,216	24,632	30,766	(NA)
2011	97,220	1,234	3,883	7,443	30,370	24,319	29,971	(NA)
2010	96,325	1,279	3,931	7,705	30,682	23,570	29,158	(NA)
2009	95,518	1,372	4,027	7,754	30,025	23,634	28,706	(NA)
2008	94,470	1,310	4,136	7,853	29,491	23,247	28,433	(NA)
2007	93,421	1,458	4,249	8,294	29,604	22,219	27,596	(NA)
2006	92,233	1,472	4,395	7,940	29,380	22,136	26,910	(NA)
2005	90,899	1,505	4,402	7,787	29,151	21,794	26,259	(NA)



readxl part of the tidyverse

```
library(readxl)
```

edu

```
## # A tibble: 366 x 11
##   age   sex   year total elem4 elem8   hs3   hs4 coll3 coll4 median
##   <chr> <chr> <int> <int> <int> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 25-34 Male   2016 21845   116    468  1427  6386  6015  7432    NA
## 2 25-34 Male   2015 21427   166    488  1584  6198  5920  7071    NA
## 3 25-34 Male   2014 21217   151    512  1611  6323  5910  6710    NA
## 4 25-34 Male   2013 20816   161    582  1747  6058  5749  6519    NA
## 5 25-34 Male   2012 20464   161    579  1707  6127  5619  6270    NA
## 6 25-34 Male   2011 20985   190    657  1791  6444  5750  6151    NA
## 7 25-34 Male   2010 20689   186    641  1866  6458  5587  5951    NA
## 8 25-34 Male   2009 20440   184    695  1806  6495  5508  5752    NA
## 9 25-34 Male   2008 20210   172    714  1874  6356  5277  5816    NA
## 10 25-34 Male  2007 20024   246    757  1930  6361  5137  5593   NA
## # ... with 356 more rows
```

```
edu_t <- pivot_longer(data = edu,
                      cols = elem4:coll4,
                      names_to = "school",
                      values_to = "freq")
```

```
head(edu_t)  
tail(edu_t)
```

```
head(edu_t)
```

```
## # A tibble: 6 x 7  
##   age   sex   year total median school  freq  
##   <chr> <chr> <int> <dbl> <chr>    <dbl>  
## 1 25-34 Male   2016 21845     NA elem4     116  
## 2 25-34 Male   2015 21427     NA elem4     166  
## 3 25-34 Male   2014 21217     NA elem4     151  
## 4 25-34 Male   2013 20816     NA elem4     161  
## 5 25-34 Male   2012 20464     NA elem4     161  
## 6 25-34 Male   2011 20985     NA elem4     190
```

```
tail(edu_t)
```

```
## # A tibble: 6 x 7  
##   age   sex   year total median school  freq  
##   <chr> <chr> <int> <dbl> <chr>    <dbl>  
## 1 55> Female 1959 16263     8.30 coll4  
## 2 55> Female 1957 15581     8.20 coll4  
## 3 55> Female 1952 13662     7.90 coll4  
## 4 55> Female 1950 13150     8.40 coll4  
## 5 55> Female 1947 11810     7.60 coll4  
## 6 55> Female 1940 9777      8.30 coll4
```

```

gen_cats <- function(x, N = 1000) {
  sample(x, N, replace = TRUE)
}

set.seed(101)
N <- 1000
income <- rnorm(N, 100, 50)
vars <- list(stratum = c(1:8),
            sex = c("M", "F"),
            race = c("B", "W"),
            educ = c("HS", "BA"))

df <- as_tibble(map_dfc(vars, gen_cats))
df <- add_column(df, income)

```

	stratum	sex	race	educ	income
## 1	6	F	W	HS	83.7
## 2	5	F	W	BA	128.
## 3	3	F	B	HS	66.3
## 4	3	F	W	HS	111.
## 5	6	M	W	BA	116.
## 6	7	M	B	HS	159.
## 7	8	M	W	BA	131.
## 8	3	M	W	BA	94.4
## 9	7	F	B	HS	146.
## 10	2	F	W	BA	88.8
## # ... with 990 more rows					

```
df
## # A tibble: 1,000 x 5
##   stratum sex   race educ income
##   <int> <chr> <chr> <chr> <dbl>
## 1     6 F     W     HS      83.7
## 2     5 F     W     BA     128.
## 3     3 F     B     HS      66.3
## 4     3 F     W     HS      111.
## 5     6 M     W     BA     116.
## 6     7 M     B     HS     159.
## 7     8 M     W     BA     131.
## 8     3 M     W     BA     94.4
## 9     7 F     B     HS     146.
## 10    2 F     W     BA     88.8
## # ... with 990 more rows
```

# Lengthening Multiple Values

```
df %>%
  group_by(sex, race, stratum, educ) %>%
  summarize(mean_inc = mean(income),
            n = n()) %>%
  pivot_wider(names_from = educ,
              values_from = c(mean_inc, n))
```

```
## # A tibble: 32 x 7
## # Groups:   sex, race, stratum [32]
##   sex   race  stratum mean_inc_BA mean_inc_HS  n_BA  n_HS
##   <chr> <chr>    <int>      <dbl>       <dbl> <int> <int>
## 1 F     B          1        93.8       99.3    19     6
## 2 F     B          2        89.7       93.0    11    16
## 3 F     B          3       112.       95.0    13    16
## 4 F     B          4       108.       96.1    14    15
## 5 F     B          5       91.0       92.6    11    15
## 6 F     B          6       93.0      116.     15    15
## 7 F     B          7       102.      121.     13    13
## 8 F     B          8       105.      88.3     14     8
## 9 F     W          1       92.6      110.     19    13
## 10 F    W          2       98.5      101.     15    19
## # ... with 22 more rows
```

```
df %>%
  group_by(sex, race, stratum, educ) %>%
  summarize(mean_inc = mean(income),
            n = n())
```

```
# A tibble: 64 x 6
# Groups:   sex, race, stratum [32]
  sex   race  stratum educ mean_inc     n
  <chr> <chr>    <int> <chr>     <dbl> <int>
1 F     B          1 BA      93.8     19
2 F     B          1 HS      99.3      6
3 F     B          2 BA      89.7     11
4 F     B          2 HS      93.0     16
5 F     B          3 BA     112.      13
6 F     B          3 HS      95.0     16
7 F     B          4 BA     108.      14
8 F     B          4 HS      96.1     15
9 F     B          5 BA      91.0     11
10 F    B          5 HS      92.6     15
# ... with 54 more rows
```

# Date Formats

```
head(bad_date)
```

```
## # A tibble: 6 x 2
##   date      N
##   <chr>    <int>
## 1 9/1/11  44426
## 2 9/2/11  55112
## 3 9/3/11  19263
## 4 9/4/11  12330
## 5 9/5/11  8534
## 6 9/6/11  59490
```

```
head(bad_date)
```

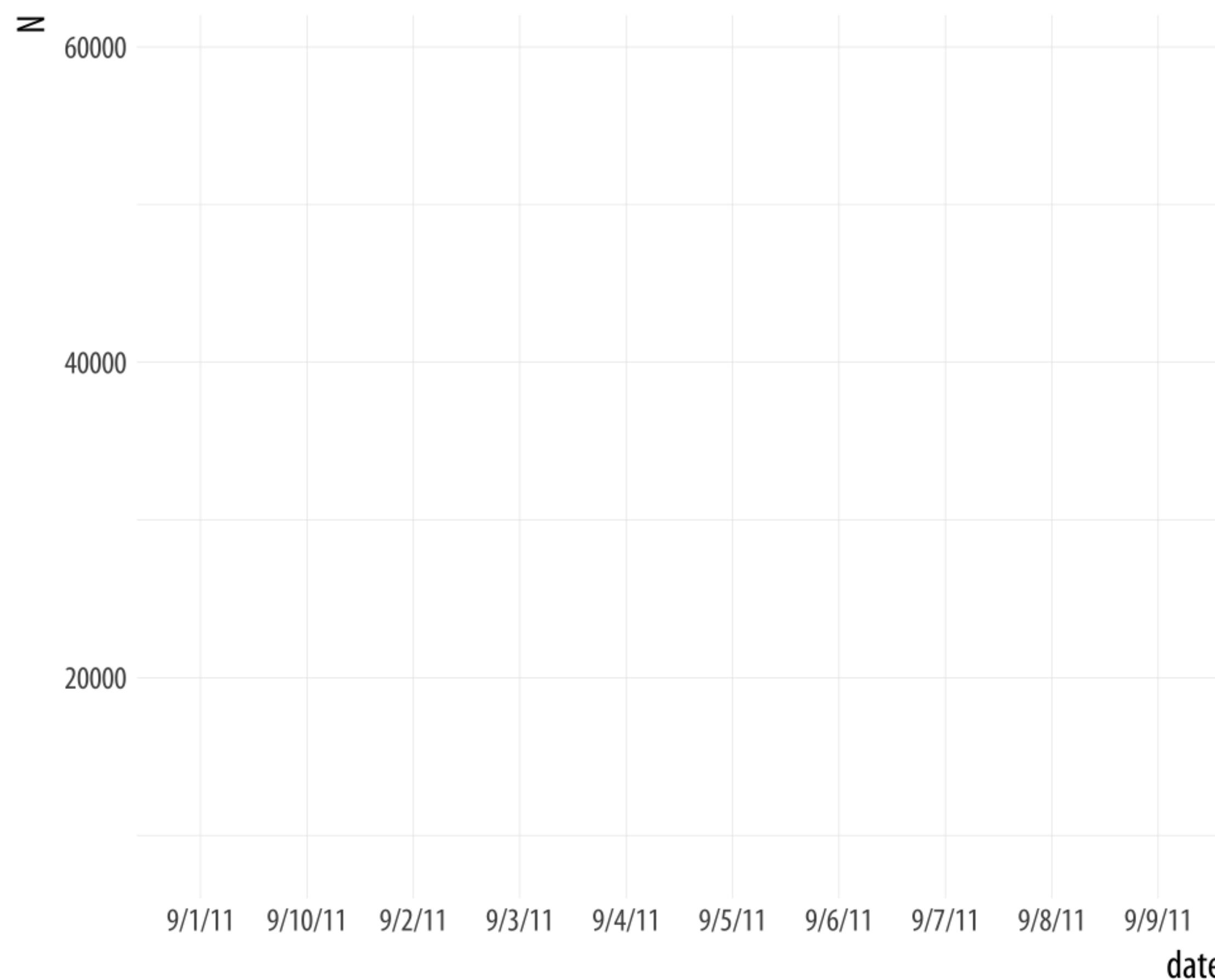
```
## # A tibble: 6 × 2
##   date      N
##   <chr>    <int>
## 1 9/1/11  44426
## 2 9/2/11  55112
## 3 9/3/11  19263
## 4 9/4/11  12330
## 5 9/5/11  8534
## 6 9/6/11  59490
```

```
p <- ggplot(data = bad_date, aes(x = date, y = N))
p + geom_line()
```

```
## geom_path: Each group consists of only one observation.
## Do you need to adjust the group aesthetic?
```

```
bad_date2 <- rbind(bad_date, bad_date)

p <- ggplot(data = bad_date2, aes(x = date, y = N))
p + geom_line()
```

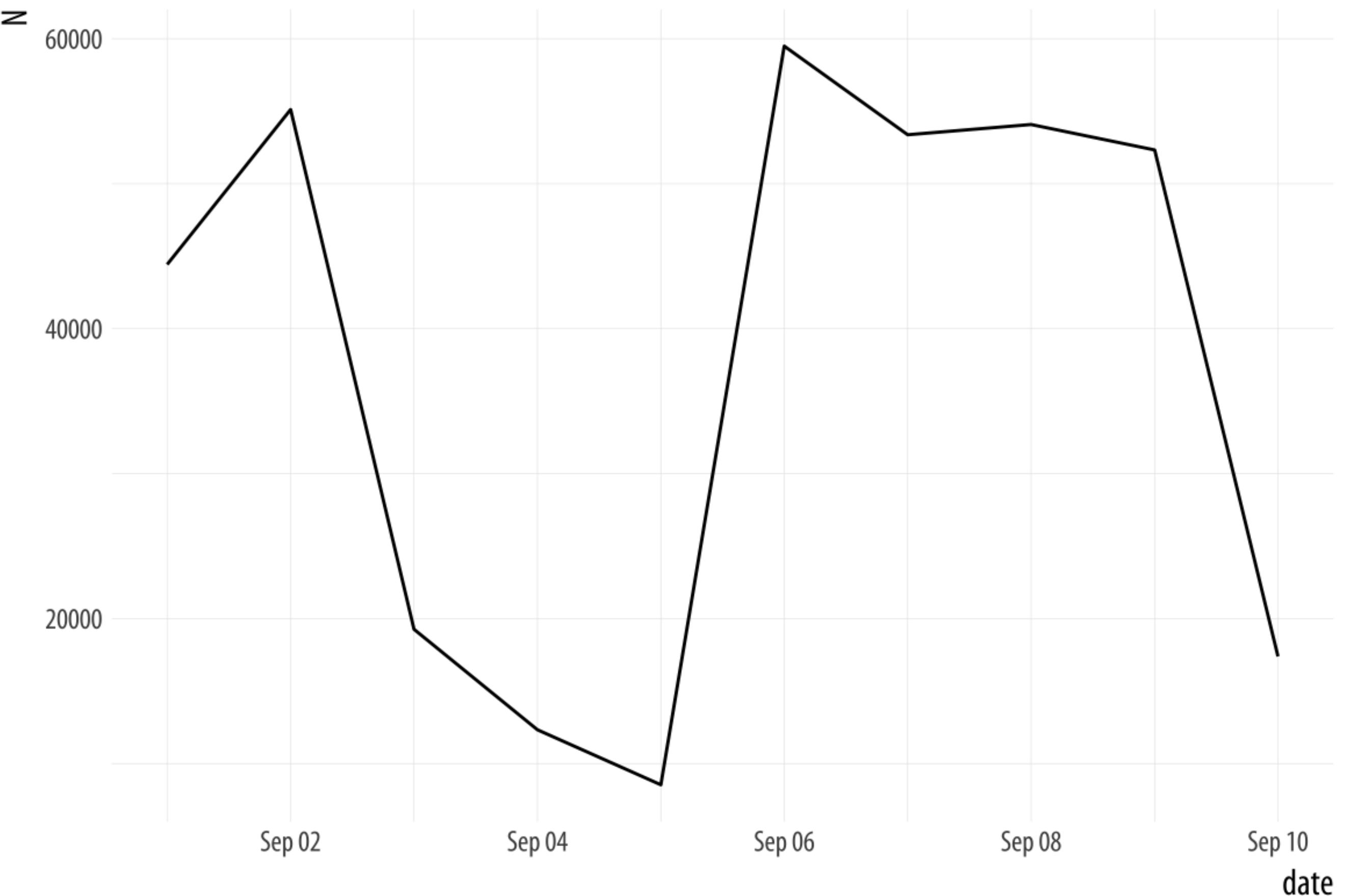


```
library(lubridate)

bad_date$date <- mdy(bad_date$date)
head(bad_date)

## # A tibble: 6 x 2
##   date      N
##   <date>  <int>
## 1 2011-09-01 44426
## 2 2011-09-02 55112
## 3 2011-09-03 19263
## 4 2011-09-04 12330
## 5 2011-09-05  8534
## 6 2011-09-06 59490

p <- ggplot(data = bad_date, aes(x = date, y = N))
p + geom_line()
```



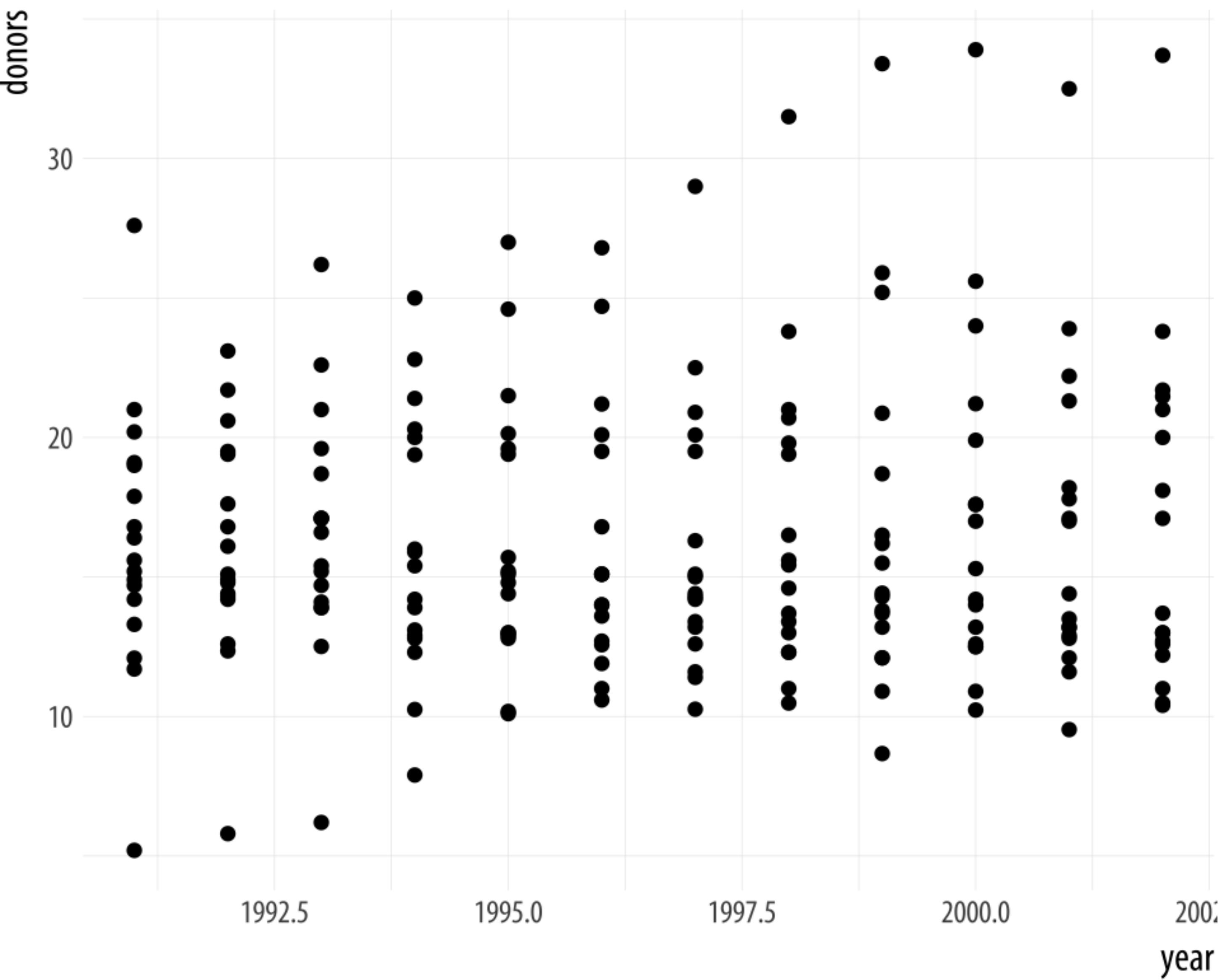
```

bad_year <- read_csv("data/organdonation")
bad_year %>% select(1:3) %>% sample_n(10)

## # A tibble: 10 x 3
##   country     year donors
##   <chr>       <int>  <dbl>
## 1 United States 1994  19.4
## 2 Australia    1999  8.67
## 3 Canada      2001  13.5
## 4 Australia    1994  10.2
## 5 Sweden       1993  15.2
## 6 Ireland      1992  19.5
## 7 Switzerland  1997  14.3
## 8 Ireland      2000  17.6
## 9 Switzerland  1998  15.4
## 10 Norway      NA    NA

```

**p** <- **ggplot**(data = bad\_year,  
 mapping = **aes**(x = year,  
 y = donors))  
**p** + **geom\_point**()



```
bad_year <- bad_year %>%
  mutate(year = int_to_year(year))
bad_year %>% select(1:3)
```

```
# A tibble: 238 x 3
  country    year     donors
  <chr>      <date>   <dbl>
1 Australia  NA        NA
2 Australia  1991-06-15 12.1
3 Australia  1992-06-15 12.4
4 Australia  1993-06-15 12.5
5 Australia  1994-06-15 10.2
6 Australia  1995-06-15 10.2
7 Australia  1996-06-15 10.6
8 Australia  1997-06-15 10.3
9 Australia  1998-06-15 10.5
10 Australia 1999-06-15  8.67
# ... with 228 more rows
```

# Zero Counts in dplyr

**[https://github.com/kjhealy/fc\\_sample](https://github.com/kjhealy/fc_sample)**

```
library(tidyverse)
```

```
## Hex colors for sex
sex_colors <- c("#E69F00", "#993300")
```

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

```
## Group labels
mf_labs <- tibble(M = "Men", F = "Women")
```

```
theme_set(theme_minimal())
```

```
## Character vectors only, by default
df <- read_csv("data/fc_sample.csv")
```

```
df
```

```
#> #> df
#> # A tibble: 280 x 4
#>   pid start_year party      sex
#>   <int> <date>    <chr>    <chr>
#> 1 3160 2013-01-03 Republican M
#> 2 3161 2013-01-03 Democrat   F
#> 3 3162 2013-01-03 Democrat   M
#> 4 3163 2013-01-03 Republican M
#> 5 3164 2013-01-03 Democrat   M
#> 6 3165 2013-01-03 Republican M
#> 7 3166 2013-01-03 Republican M
#> 8 3167 2013-01-03 Democrat   F
#> 9 3168 2013-01-03 Republican M
#> 10 3169 2013-01-03 Democrat   M
#> # ... with 270 more rows
```

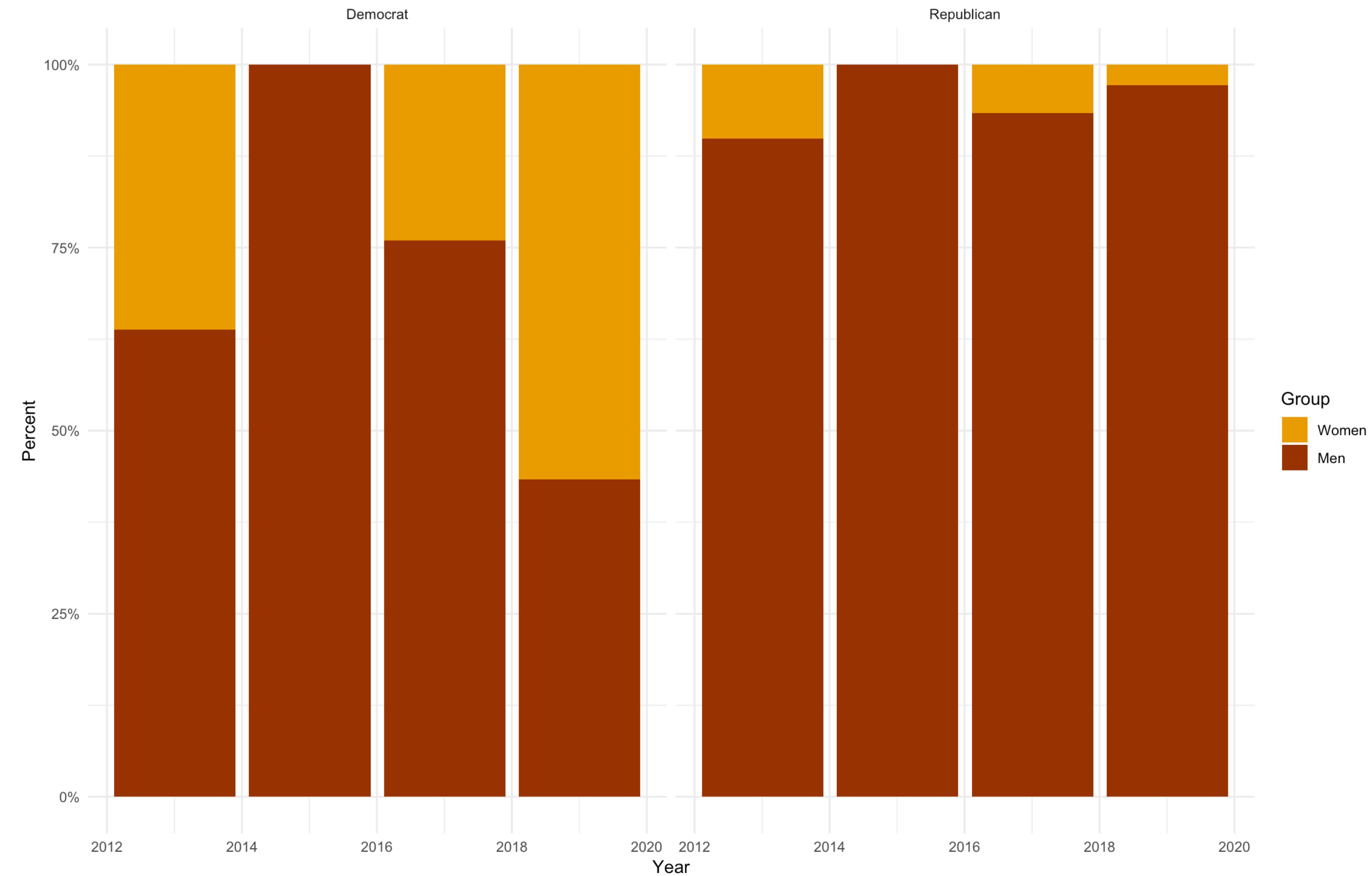
```
df %>%  
  group_by(start_year, party, sex) %>%  
  summarize(N = n()) %>%  
  mutate(freq = N / sum(N))
```

```
#> # A tibble: 14 x 5  
#> # Groups:   start_year, party [8]  
#>   start_year party      sex     N   freq  
#>   <date>     <chr>    <chr> <int> <dbl>  
#> 1 2013-01-03 Democrat   F     21 0.362  
#> 2 2013-01-03 Democrat   M     37 0.638  
#> 3 2013-01-03 Republican F      8 0.101  
#> 4 2013-01-03 Republican M     71 0.899  
#> 5 2015-01-03 Democrat   M      1 1  
#> 6 2015-01-03 Republican M      5 1  
#> 7 2017-01-03 Democrat   F      6 0.24  
#> 8 2017-01-03 Democrat   M     19 0.76  
#> 9 2017-01-03 Republican F      2 0.0667  
#> 10 2017-01-03 Republican M     28 0.933  
#> 11 2019-01-03 Democrat   F     33 0.647  
#> 12 2019-01-03 Democrat   M     18 0.353  
#> 13 2019-01-03 Republican F      1 0.0323  
#> 14 2019-01-03 Republican M     30 0.968
```

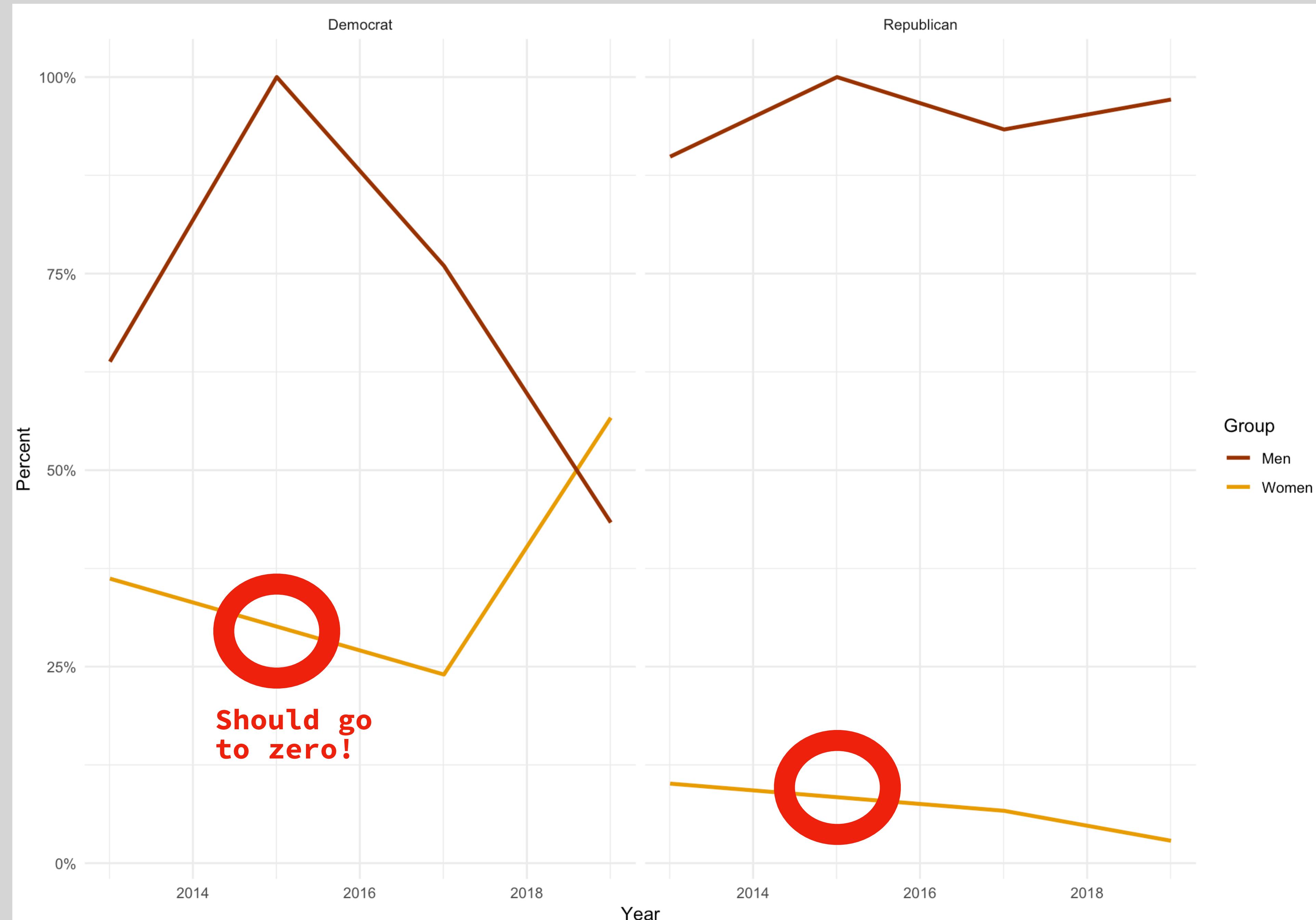
# Not in the table

```
#>      start_year party       sex     N freq
#> 5' 2015-01-03 Democrat   F      0 0
#> 6' 2015-01-03 Republican F      0 0
```

```
df %>%
  group_by(start_year, party, sex) %>%
  summarize(N = n()) %>%
  mutate(freq = N / sum(N)) %>%
  ggplot(aes(x = start_year,
              y = freq,
              fill = sex)) +
  geom_col() +
  scale_y_continuous(labels = scales::percent) +
  scale_fill_manual(values = sex_colors,
                    labels = c("Women", "Men")) +
  labs(x = "Year", y = "Percent", fill = "Group") +
  facet_wrap(~ party)
```



```
df %>%
  group_by(start_year, party, sex) %>%
  summarize(N = n()) %>%
  mutate(freq = N / sum(N)) %>%
  ggplot(aes(x = start_year,
              y = freq,
              color = sex)) +
  geom_line(size = 1.1) +
  scale_y_continuous(labels = scales::percent) +
  scale_color_manual(values = sex_colors,
                     labels = c("Women", "Men")) +
  guides(color = guide_legend(reverse = TRUE)) +
  labs(x = "Year", y = "Percent", color = "Group") +
  facet_wrap(~ party)
```



```
df_f <- df %>% modify_if(is.character, as.factor)

df_f %>%
  group_by(start_year, party, sex, .drop = FALSE) %>%
  tally()

#> # A tibble: 16 x 4
#> # Groups:   start_year, party [8]
#>   start_year party       sex     n
#>   <date>     <fct>     <fct> <int>
#> 1 2013-01-03 Democrat   F      21
#> 2 2013-01-03 Democrat   M      37
#> 3 2013-01-03 Republican F      8
#> 4 2013-01-03 Republican M     71
#> 5 2015-01-03 Democrat   F      0
#> 6 2015-01-03 Democrat   M      1
#> 7 2015-01-03 Republican F      0
#> 8 2015-01-03 Republican M      5
```

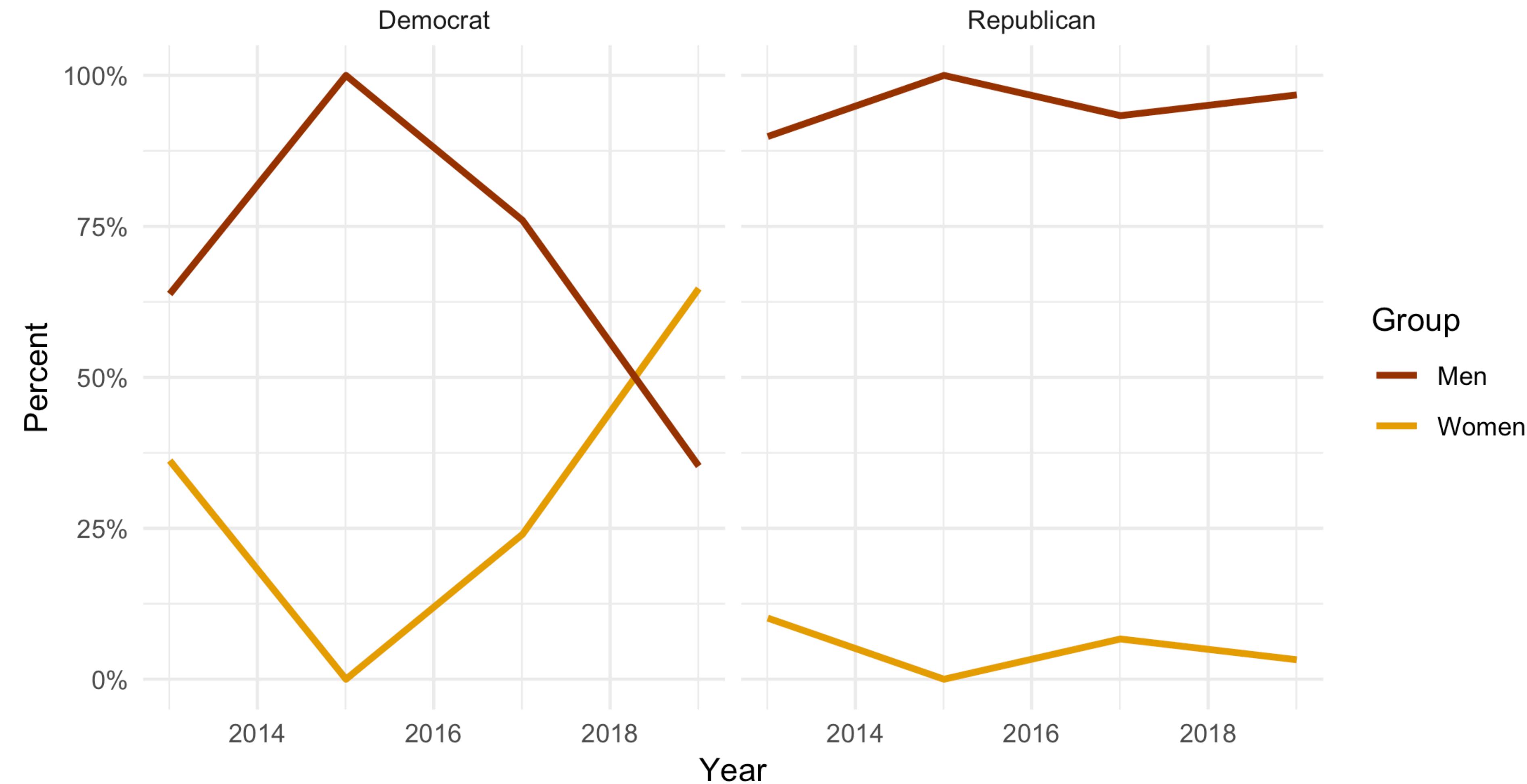
Option 1: Convert to Factor + .drop = FALSE

```
df %>%
  group_by(start_year, party, sex) %>%
  summarize(N = n()) %>%
  mutate(freq = N / sum(N)) %>%
  ungroup() %>%
  complete(start_year, party, sex,
    fill = list(N = 0, freq = 0))

#> # A tibble: 16 x 5
#>   start_year party       sex     N   freq
#>   <date>     <chr>     <chr> <dbl> <dbl>
#> 1 2013-01-03 Democrat   F     21  0.362
#> 2 2013-01-03 Democrat   M     37  0.638
#> 3 2013-01-03 Republican F      8  0.101
#> 4 2013-01-03 Republican M     71  0.899
#> 5 2015-01-03 Democrat   F      0  0
#> 6 2015-01-03 Democrat   M      1  1
#> 7 2015-01-03 Republican F      0  0
#> 8 2015-01-03 Republican M      5  1
```

## Option 2: `ungroup()` & `complete()`

```
df_f %>%
  group_by(start_year, party, sex) %>%
  summarize(N = n()) %>%
  mutate(freq = N / sum(N)) %>%
  ggplot(aes(x = start_year,
              y = freq,
              color = sex)) +
  geom_line(size = 1.1) +
  scale_y_continuous(labels = scales::percent) +
  scale_color_manual(values = sex_colors,
                     labels = c("Women", "Men")) +
  guides(color = guide_legend(reverse = TRUE)) +
  labs(x = "Year", y = "Percent", color = "Group") +
  facet_wrap(~ party)
```



# Visualizing Missing Data

```
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(naniar)  
library(visdat)
```

```
vis_dat(congress)
```

naniar 0.4.2.9000

Getting Started

Gallery

Articles ▾

Reference

News

# naniar

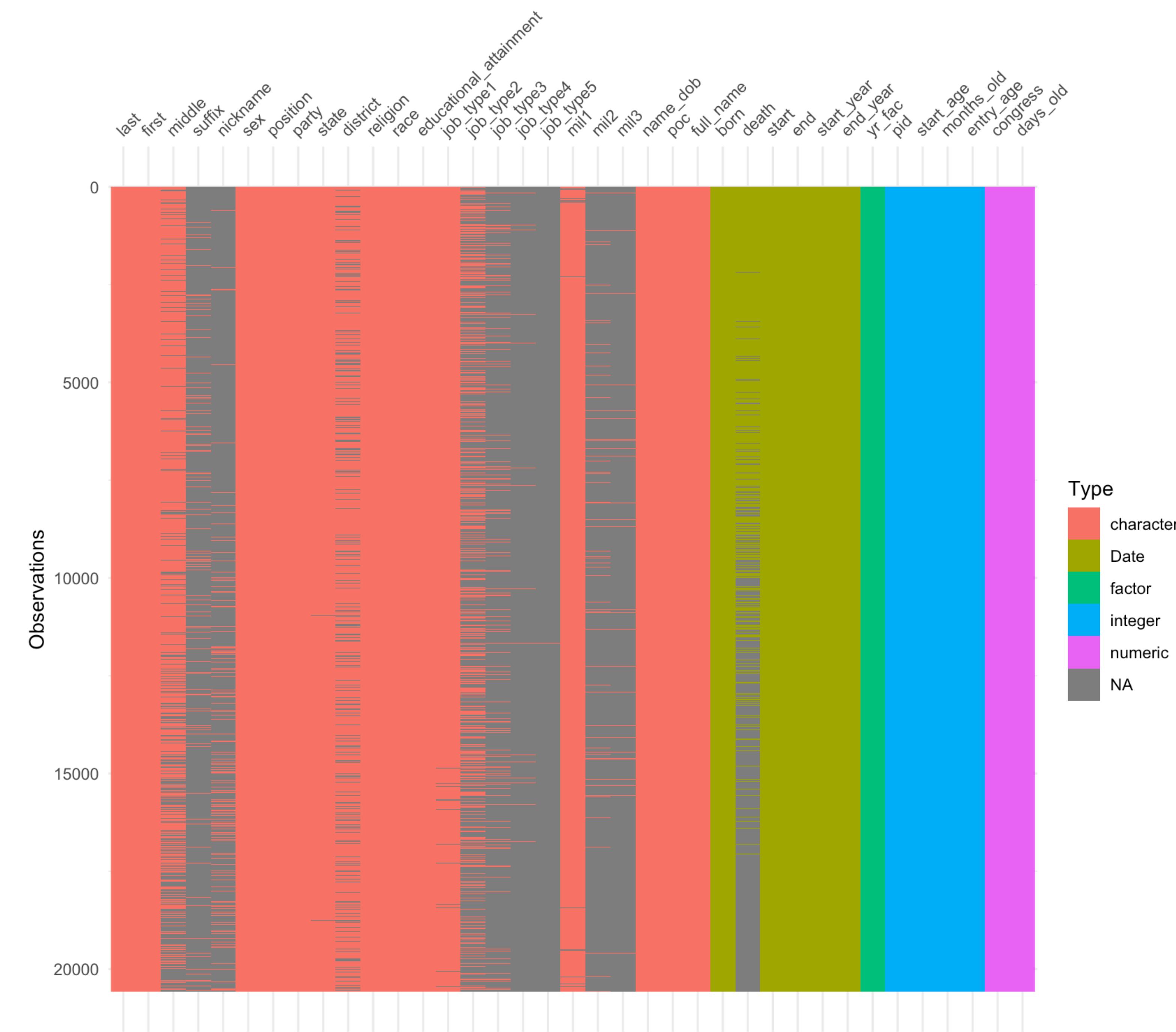
`naniar` provides principled, tidy ways to summarise, visualise, and manipulate missing data with minimal deviations from the workflows in `ggplot2` and tidy data. It does this by providing:

- Shadow matrices, a tidy data structure for missing data:
  - `bind_shadow()` and `nabular()`
- Shorthand summaries for missing data:
  - `n_miss()` and `n_complete()`
  - `pct_miss()` and `pct_complete()`
- Numerical summaries of missing data in variables and cases:
  - `miss_var_summary()` and `miss_var_table()`
  - `miss_case_summary()`, `miss_case_table()`
- Visualisation for missing data:
  - `geom_miss_point()`
  - `gg_miss_var()`
  - `gg_miss_case()`
  - `gg_miss_fct()`



For more details on the workflow and theory underpinning naniar, read the vignette [Getting started with naniar](#).

For a short primer on the data visualisation available in naniar, read the vignette [Gallery of Missing Data Visualisations](#).

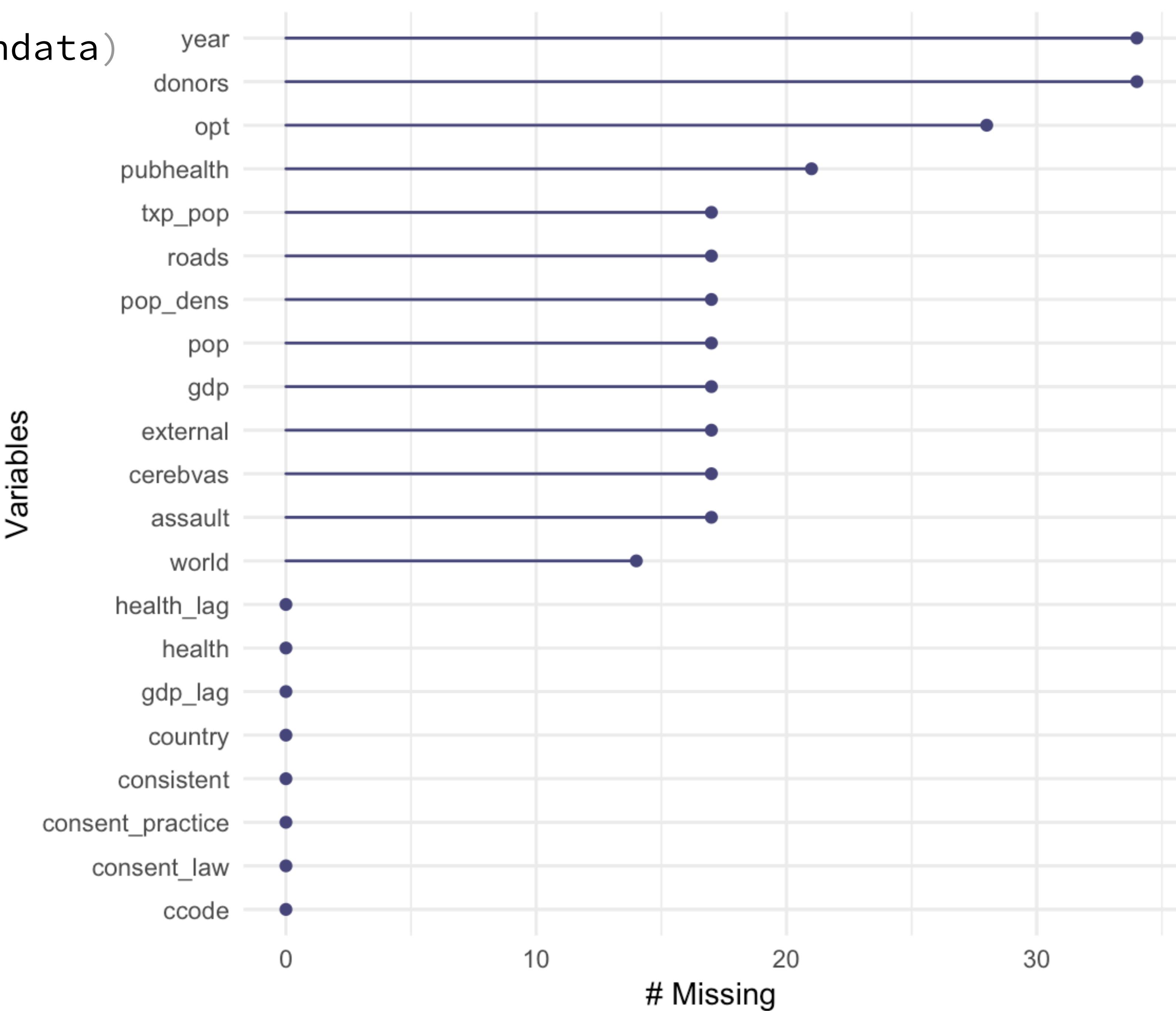


```
library(socviz)
```

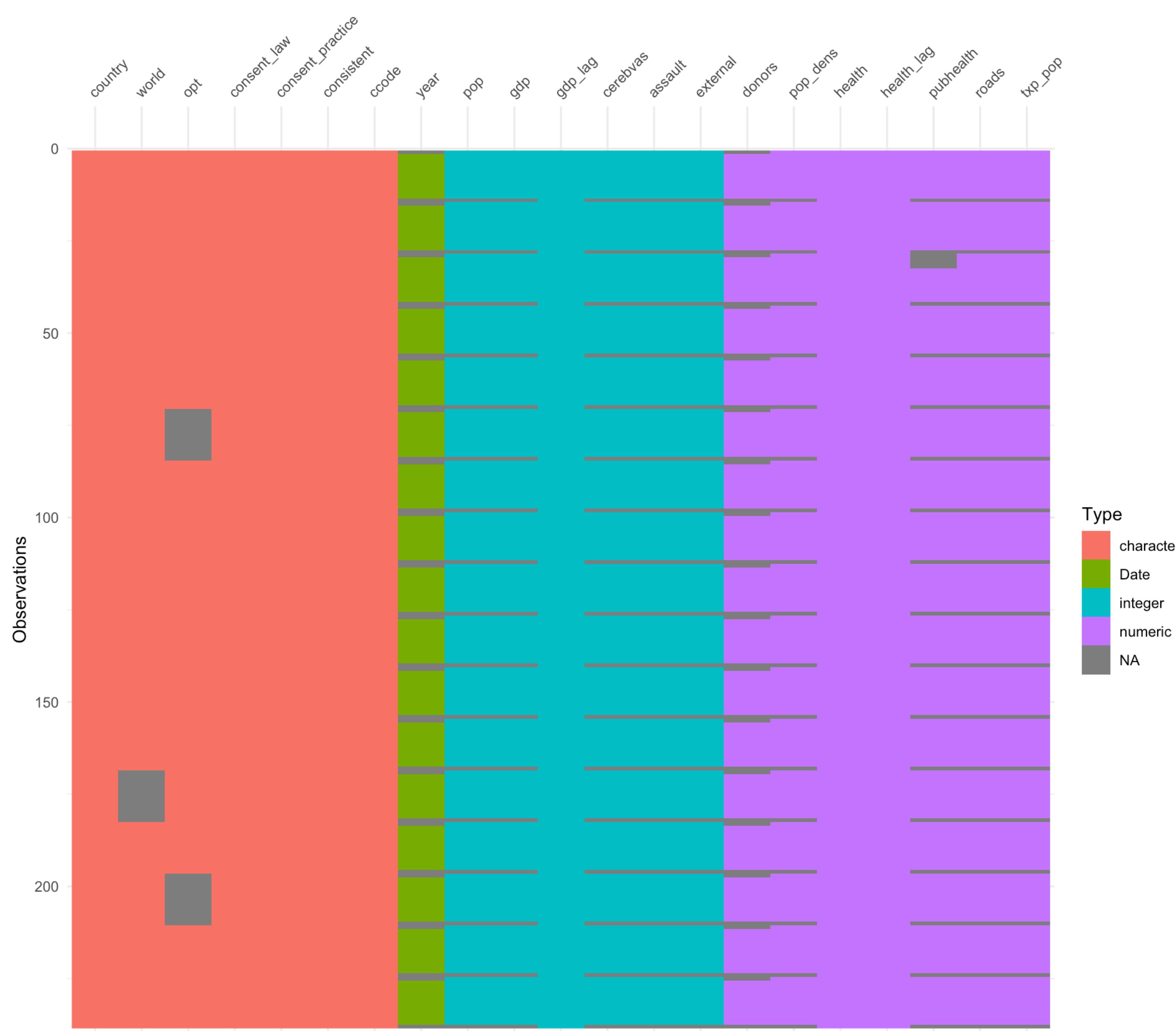
# organdata

```
# A tibble: 238 x 21
  country     year   donors    pop  pop_dens     gdp  gdp_lag health
  <chr>     <date>   <dbl>  <int>     <dbl>  <int>     <int>     <dbl>
1 Austra... NA        NA    17065      0.220  16774    16591    1300
2 Austra... 1991-01-01 12.1    17284      0.223  17171    16774    1379
3 Austra... 1992-01-01 12.4    17495      0.226  17914    17171    1455
4 Austra... 1993-01-01 12.5    17667      0.228  18883    17914    1540
5 Austra... 1994-01-01 10.2    17855      0.231  19849    18883    1626
6 Austra... 1995-01-01 10.2    18072      0.233  21079    19849    1737
7 Austra... 1996-01-01 10.6    18311      0.237  21923    21079    1846
8 Austra... 1997-01-01 10.3    18518      0.239  22961    21923    1948
9 Austra... 1998-01-01 10.5    18711      0.242  24148    22961    2077
10 Austra... 1999-01-01 8.67   18926      0.244  25445    24148    2231
# ... with 228 more rows, and 13 more variables: health_lag <dbl>,
# pubhealth <dbl>, roads <dbl>, cerebvas <int>, assault <int>,
# external <int>, txp_pop <dbl>, world <chr>, opt <chr>,
# consent_law <chr>, consent_practice <chr>, consistent <chr>,
# ccode <chr>
```

gg\_miss\_var(organdata)



# vis\_dat(organdata)



```
miss_var_summary(organdata)
```

A tibble: 21 x 3

	variable	n_miss	pct_miss
	<chr>	<int>	<dbl>
1	year	34	14.3
2	donors	34	14.3
3	opt	28	11.8
4	pubhealth	21	8.82
5	pop	17	7.14
6	pop_dens	17	7.14
7	gdp	17	7.14
8	roads	17	7.14
9	cerebvas	17	7.14
10	assault	17	7.14
# ... with 11 more rows			

```
miss_case_summary(organdata)
```

A tibble: 238 x 3

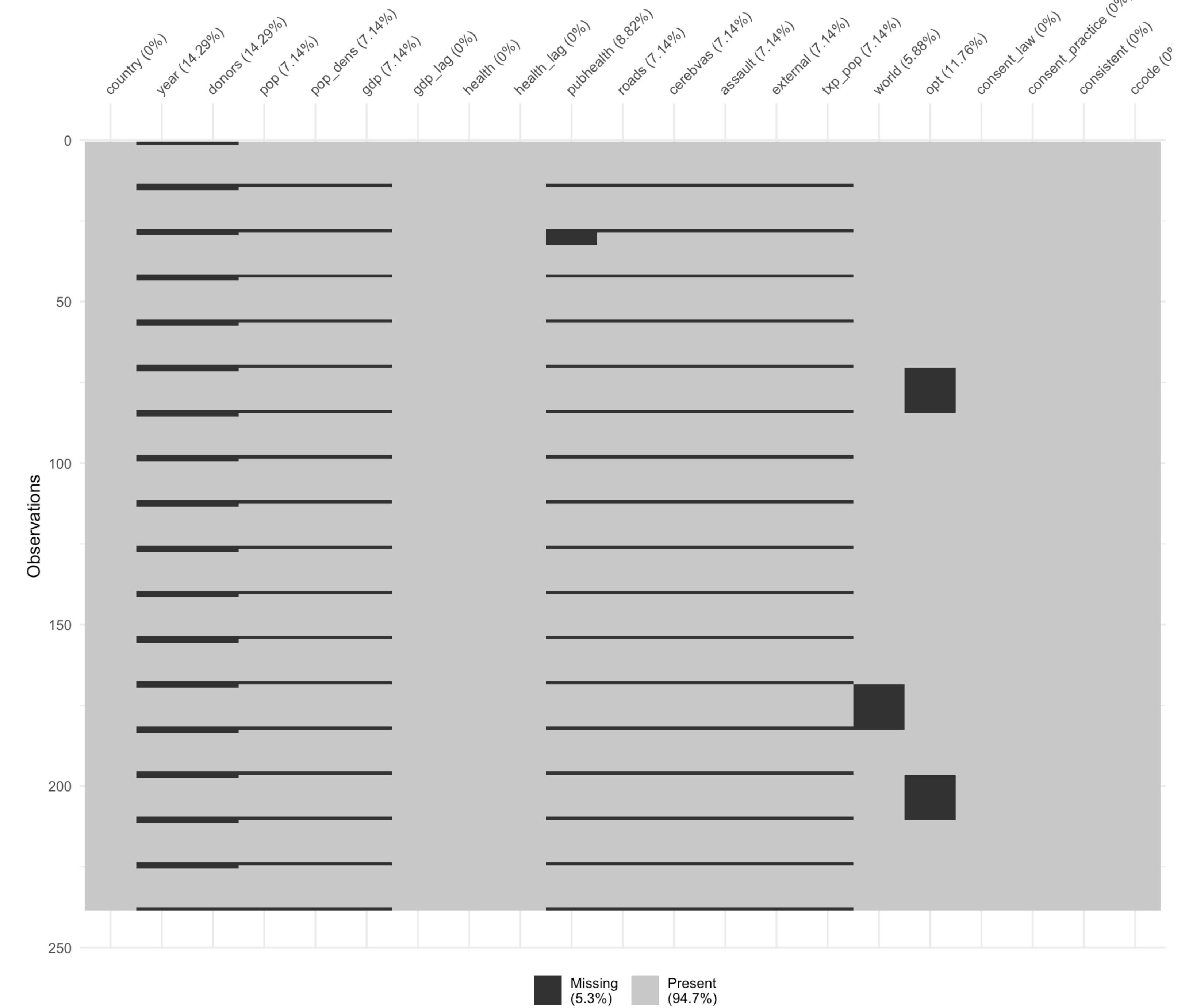
	case	n_miss	pct_miss
	<int>	<int>	<dbl>
1	84	12	57.1
2	182	12	57.1
3	210	12	57.1
4	14	11	52.4
5	28	11	52.4
6	42	11	52.4
7	56	11	52.4
8	70	11	52.4
9	98	11	52.4
10	112	11	52.4
# ... with 228 more rows			

```
organdata %>%
  select(consent_law, year, pubhealth, roads) %>%
  group_by(consent_law) %>%
  miss_var_summary()
```

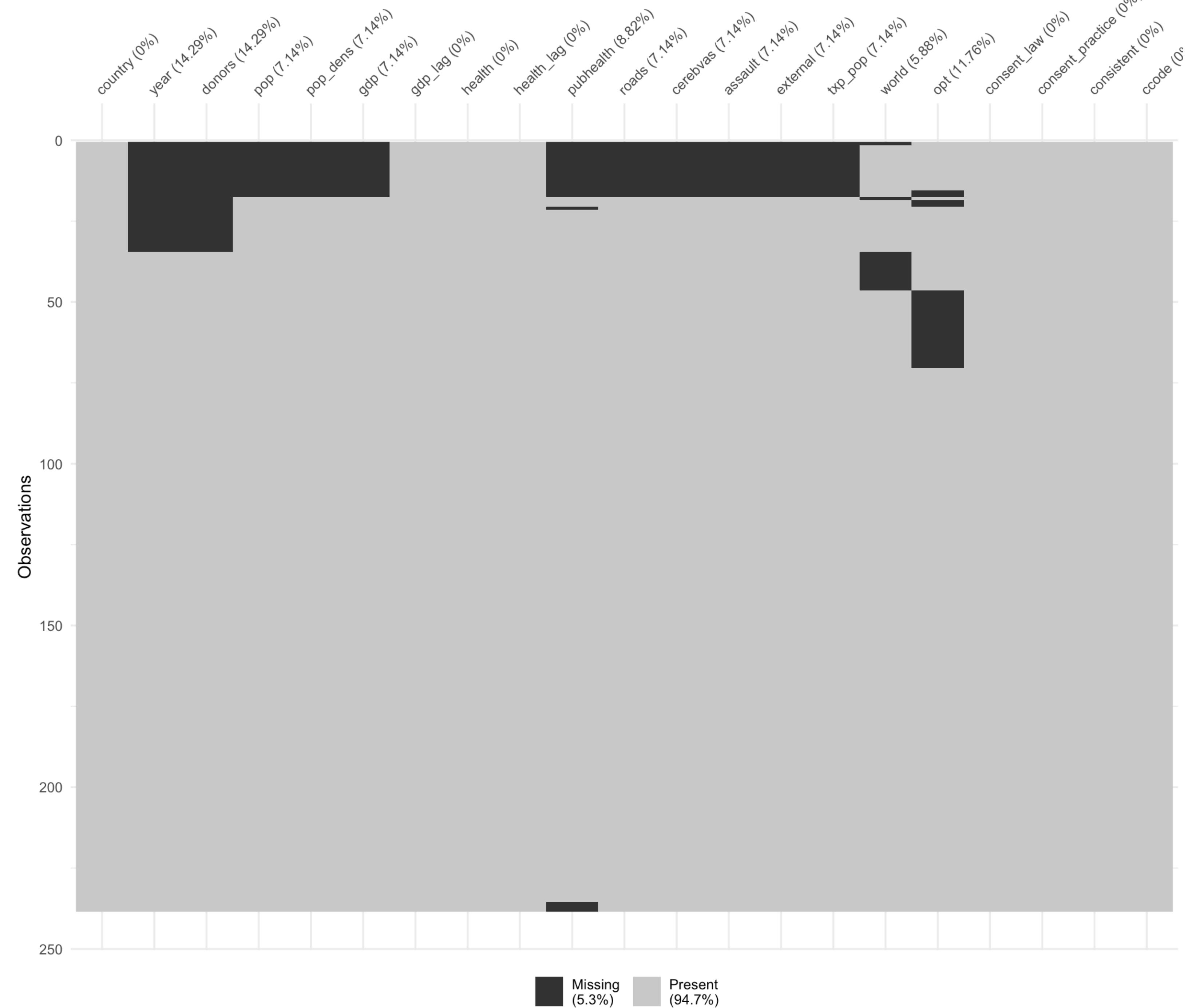
A tibble: 6 x 4

	consent_law	variable	n_miss	pct_miss
	<chr>	<chr>	<int>	<dbl>
1	Informed	year	16	14.3
2	Informed	pubhealth	8	7.14
3	Informed	roads	8	7.14
4	Presumed	year	18	14.3
5	Presumed	pubhealth	13	10.3
6	Presumed	roads	9	7.14

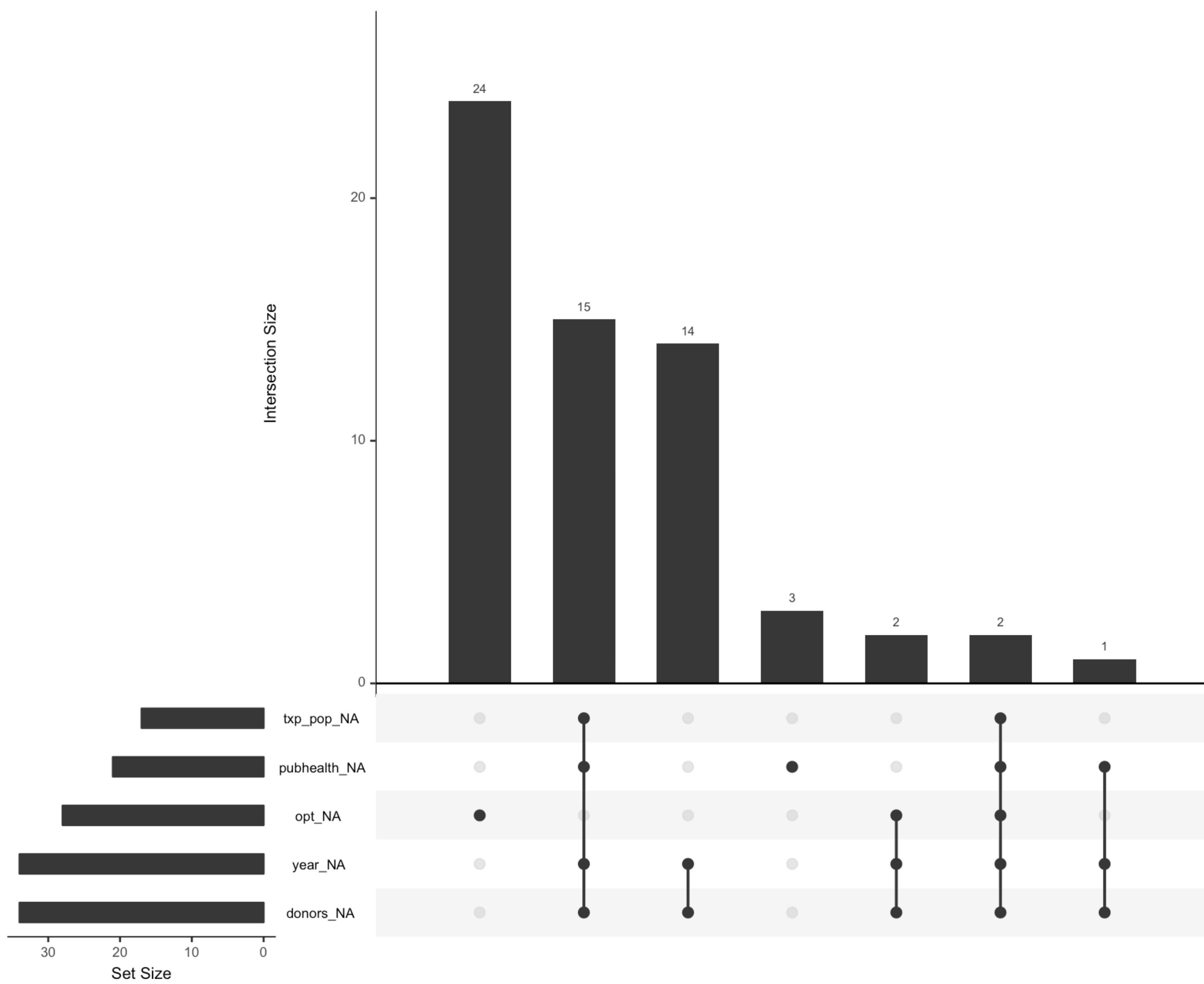
## vis\_miss(organdata)



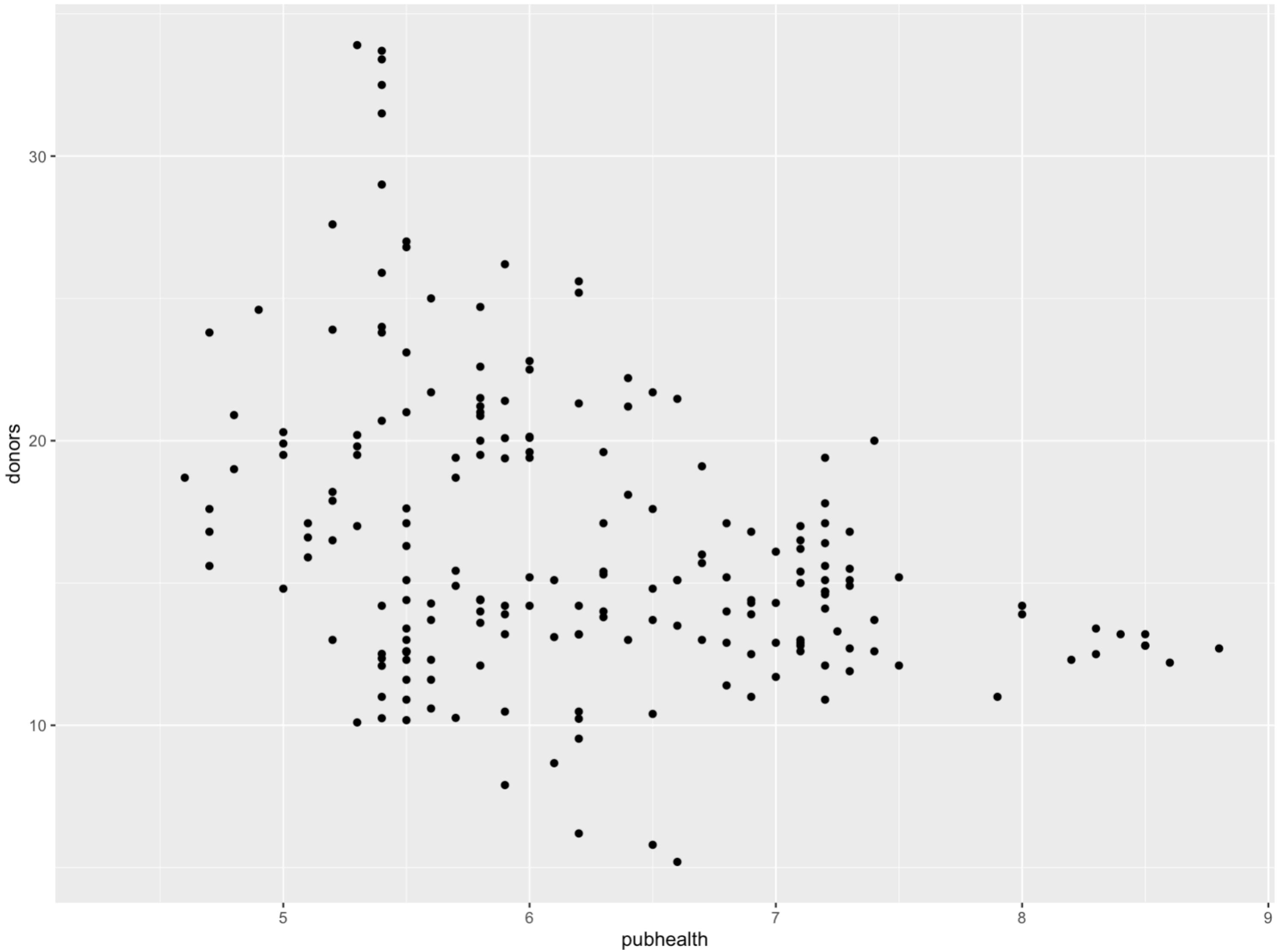
vis\_miss(organdata,  
cluster = TRUE)



gg\_miss\_upset(organdata)



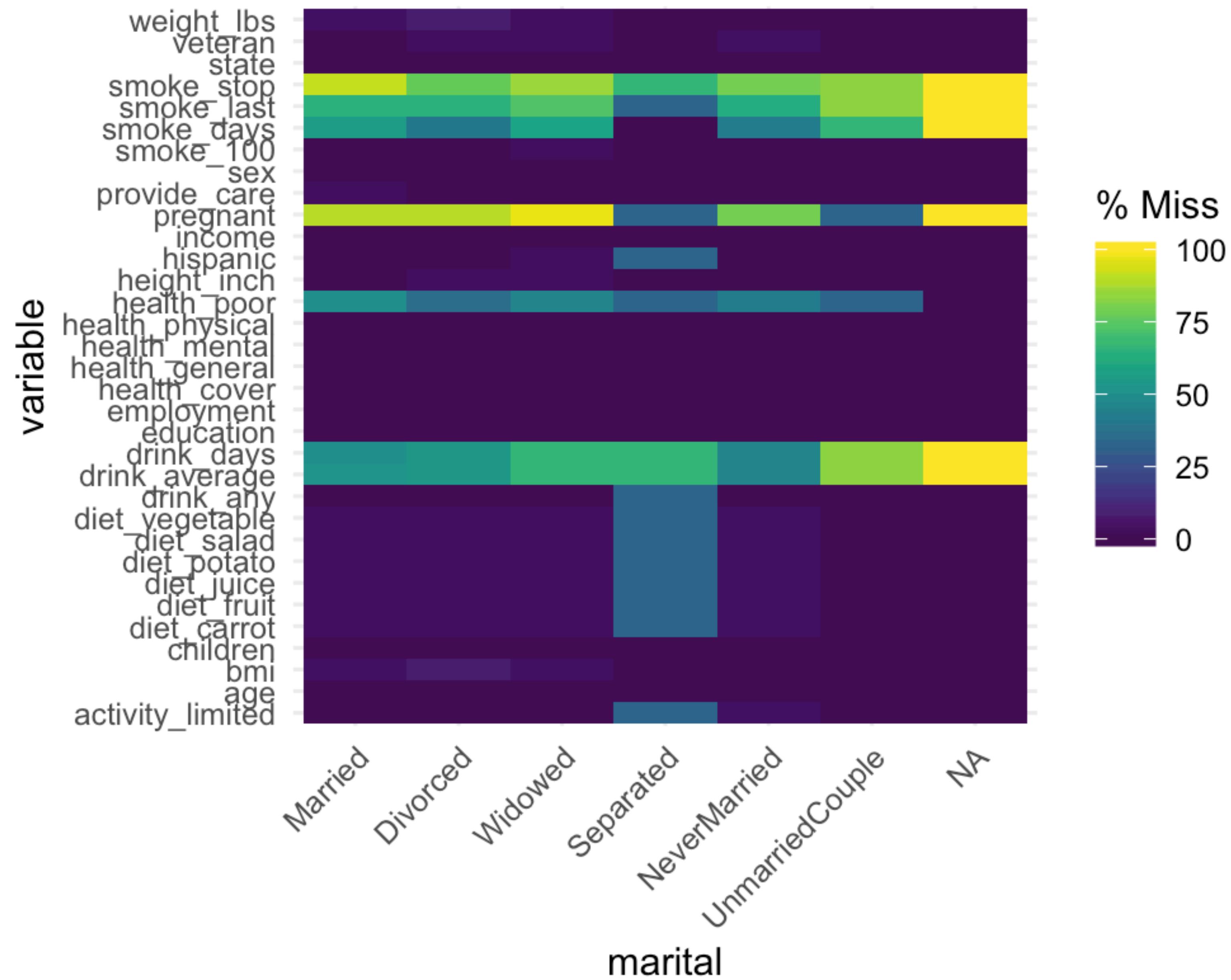
```
ggplot(data = organdata, mapping = aes(x = pubhealth, y = donors)) +  
  geom_point()  
  
## Warning message:  
## Removed 37 rows containing missing values (geom_point).
```



```
ggplot(data = organdata, mapping = aes(x = pubhealth, y = donors)) +  
  geom_miss_point()
```



```
gg_miss_fct(x = riskfactors, fct = marital)
```



# Visualizing Network Data

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

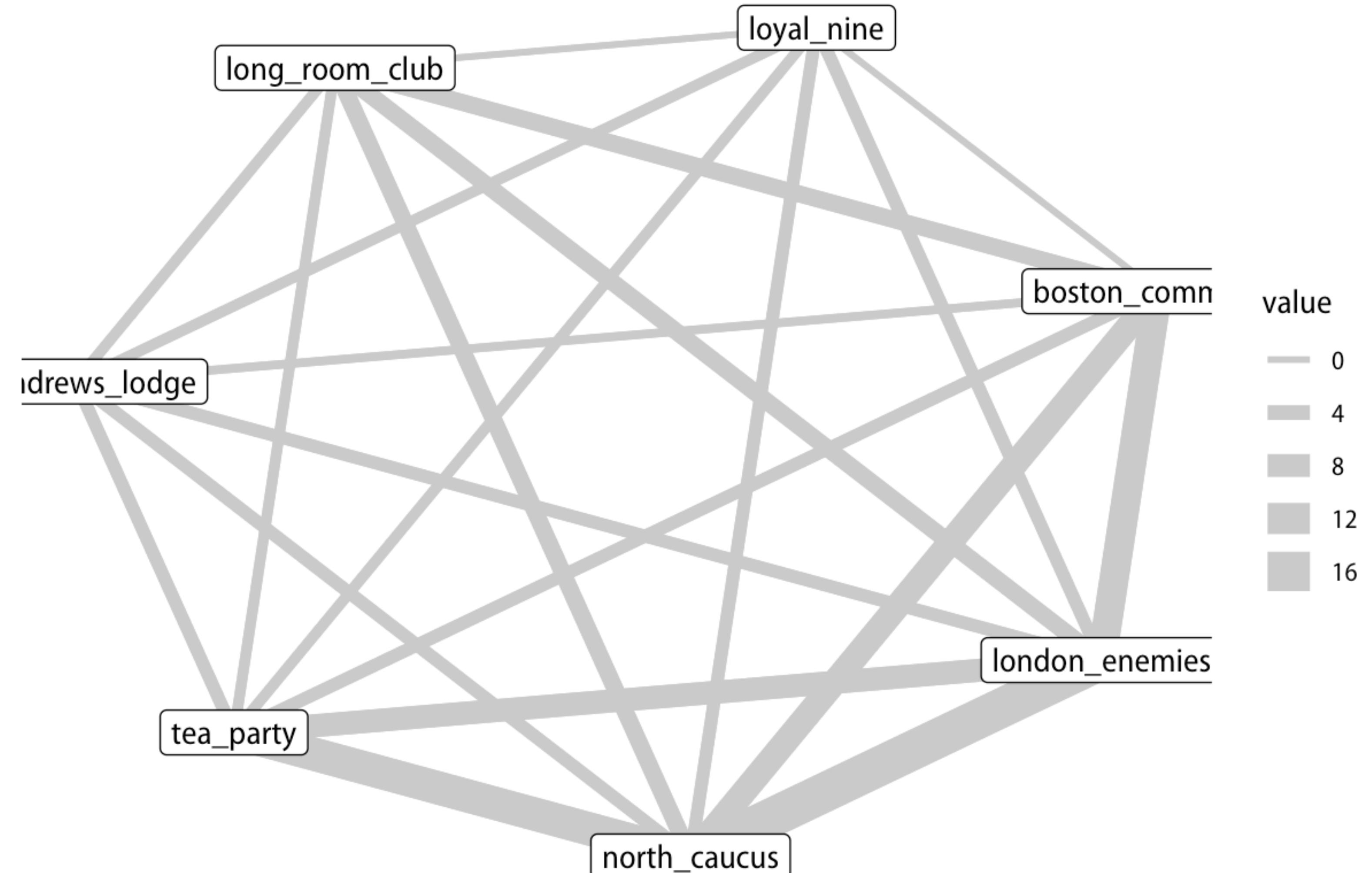
```
library(here)
library(tidyverse)
library(ggraph)
library(tidygraph)
library(graphlayouts)
library(janitor)
library(kjhnet)
```

Paul Revere

```
revere_groups
```

```
## # A tbl_graph: 7 nodes and 49 edges
## #
## # An undirected multigraph with 1 component
## #
## # Node Data: 7 x 2 (active)
##   id name
##   <int> <chr>
## 1 1 st_andrews_lodge
## 2 2 loyal_nine
## 3 3 north_caucus
## 4 4 long_room_club
## 5 5 tea_party
## 6 6 boston_committee
## # with 1 more row
## #
## # Edge Data: 49 x 3
##   from to value
##   <int> <int> <dbl>
## 1 1     1     53
## 2 1     2     2
## 3 1     3     3
## # with 46 more rows
```

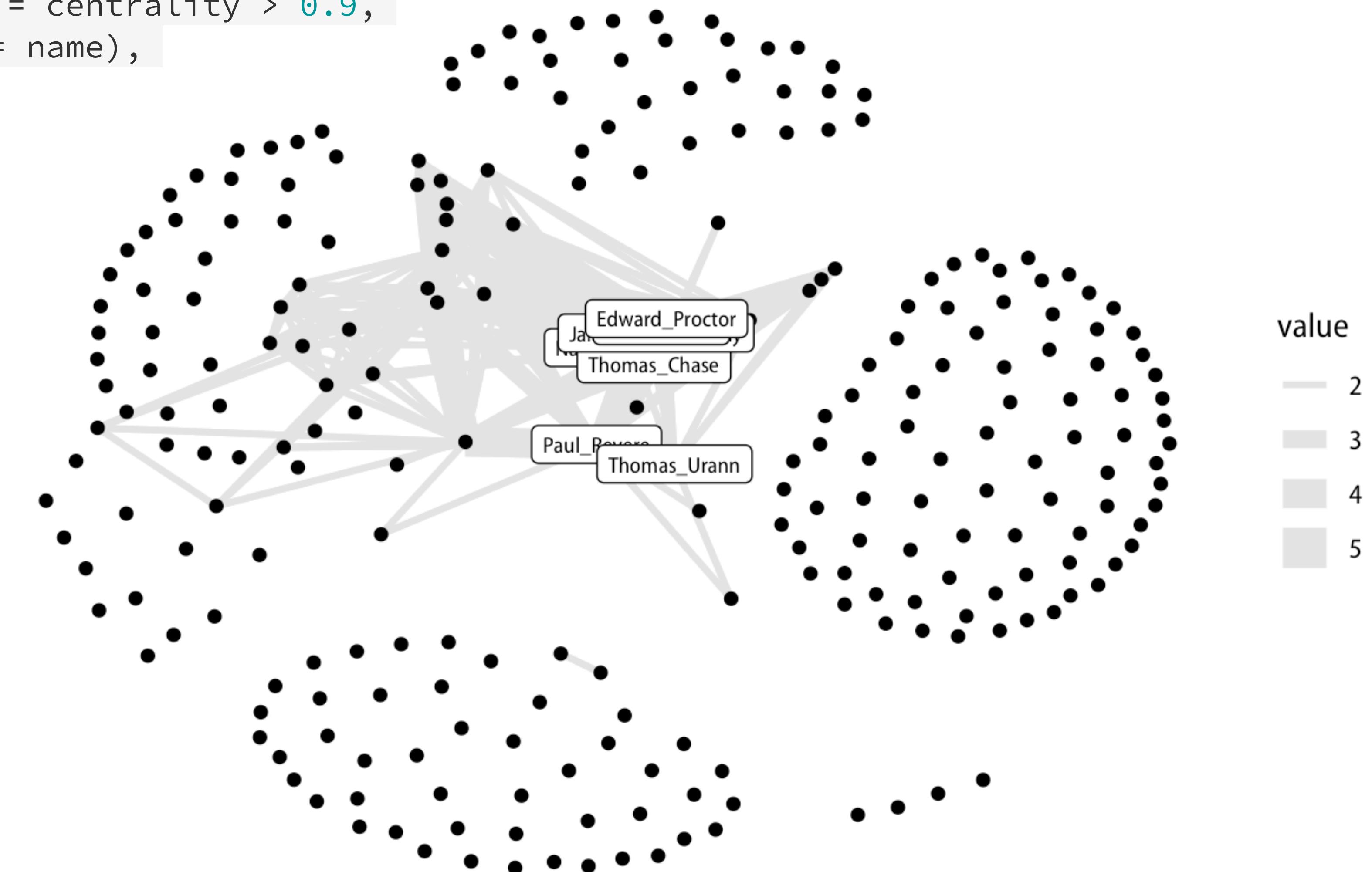
```
revere_groups %>%
  ggraph(layout = "kk") +
  geom_edge_link(aes(width = value), color = "gray80") +
  geom_node_label(aes(label = name))
```



```
revere_persons
```

```
## # A tbl_graph: 254 nodes and 19666 edges
## #
## # An undirected multigraph with 1 component
## #
## # Node Data: 254 x 2 (active)
##     id name
##     <int> <chr>
## 1   1 John_Adams
## 2   2 Samuel_Adams
## 3   3 Dr_Allen
## 4   4 Nathaniel_Appleton
## 5   5 Gilbert_Ash
## 6   6 Benjamin_Austin
## # â€¦ with 248 more rows
## #
## # Edge Data: 19,666 x 3
##     from    to value
##     <int> <int> <dbl>
## 1   1     1     2
## 2   1     2     2
## 3   1     3     1
## # # â€¦ with 1.966e+04 more rows
```

```
revere_persons %>%
  mutate(centrality = centrality_eigen()) %>%
  ggraph(layout = "stress") +
  geom_edge_link0(mapping = aes(filter = value > 1,
                                 width = value),
                  color = "gray90") +
  geom_node_point() +
  geom_node_label(mapping = aes(filter = centrality > 0.9,
                                label = name),
                  size = rel(2.5))
```



# Killings in The Iliad

Hector waiting, watching  
the greathearted Patroclus trying to stagger free,  
seeing him wounded there with the sharp bronze  
came rushing into him right across the lines  
and rammed his spearshaft home,  
stabbing deep in the bowels, and the brazen point  
went jutting straight out through Patroclus' back

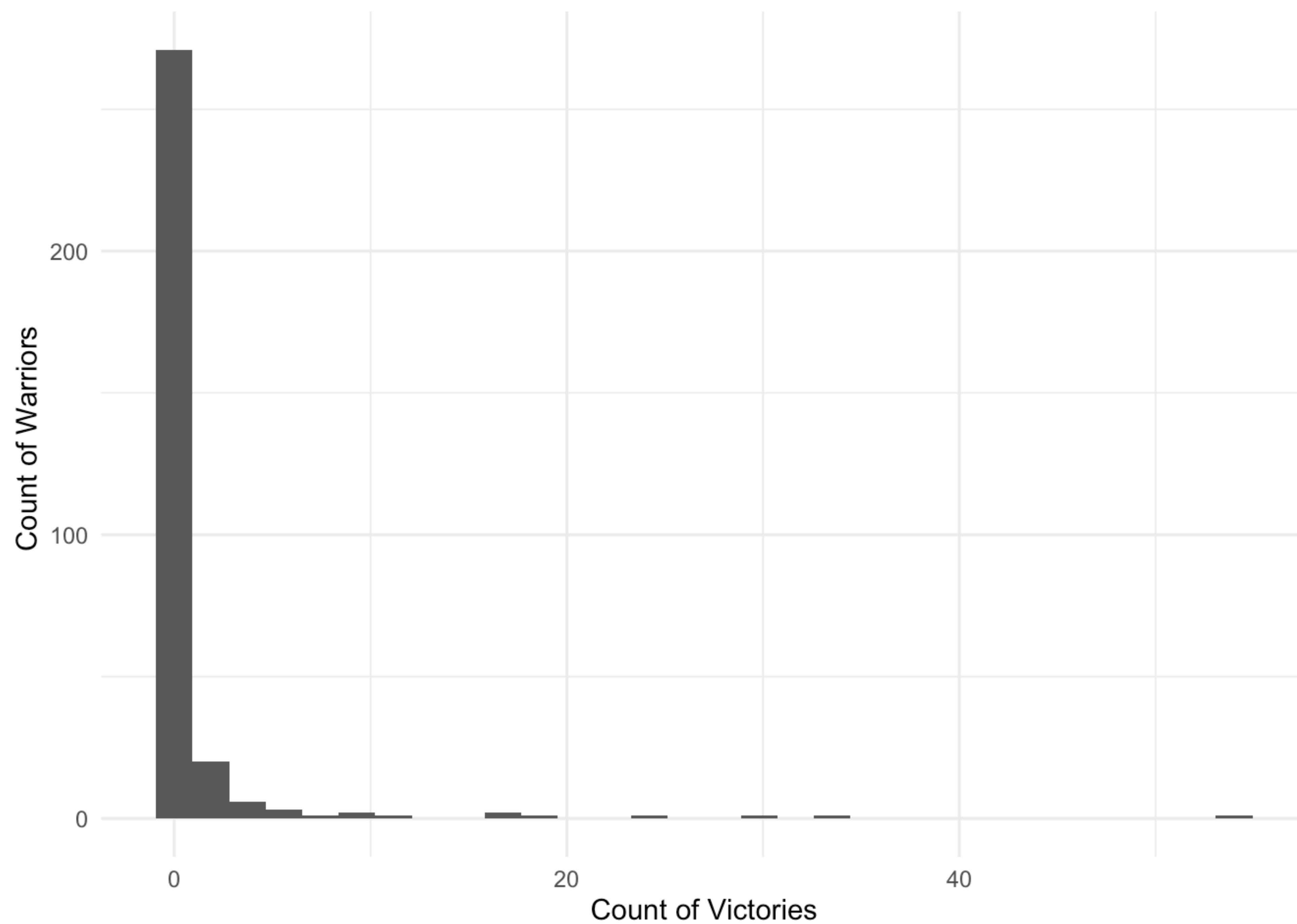
Hector! Now is your time to glory to the skies

```
il_tidy
## # A tbl_graph: 311 nodes and 292 edges
## #
## # A directed multigraph with 23 components
## #
## # Node Data: 311 x 3 (active)
##   id name          affil
##   <int> <chr>        <chr>
## 1 1 Antilochus      A
## 2 2 Agenor          T
## 3 3 Telamonian Ajax A
## 4 4 Antiphus        T
## 5 5 Odysseus        A
## 6 6 Peirous          T
## # â€¦ with 305 more rows
## #
## # Edge Data: 292 x 3
##   from    to act
##   <int> <int> <chr>
## 1 1      41 kills
## 2 2      42 kills
## 3 3      43 kills
## # â€¦ with 289 more rows
```

**So, Who's The  
Most Important  
Character in  
The Iliad?**

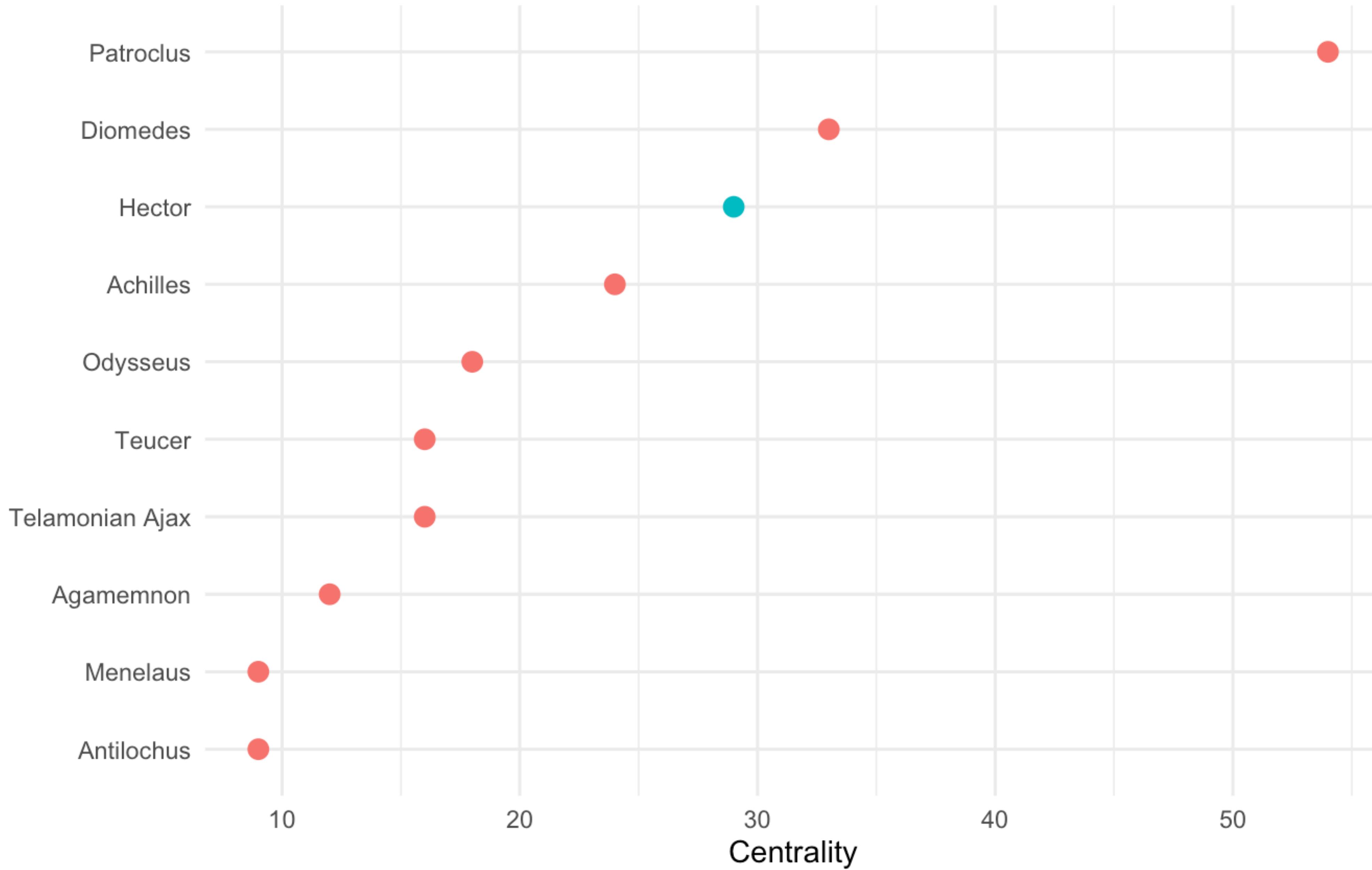
```
il_tidy %>%
  activate(nodes) %>%
  mutate(centrality = centrality_degree()) %>%
  ggplot(mapping = aes(x = centrality)) +
  geom_histogram() +
  labs(x = "Count of Victories", y = "Count of Warriors ") +
  theme_minimal() +
  theme(legend.position = "top")
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

Notice which class of people  
is missing from the data



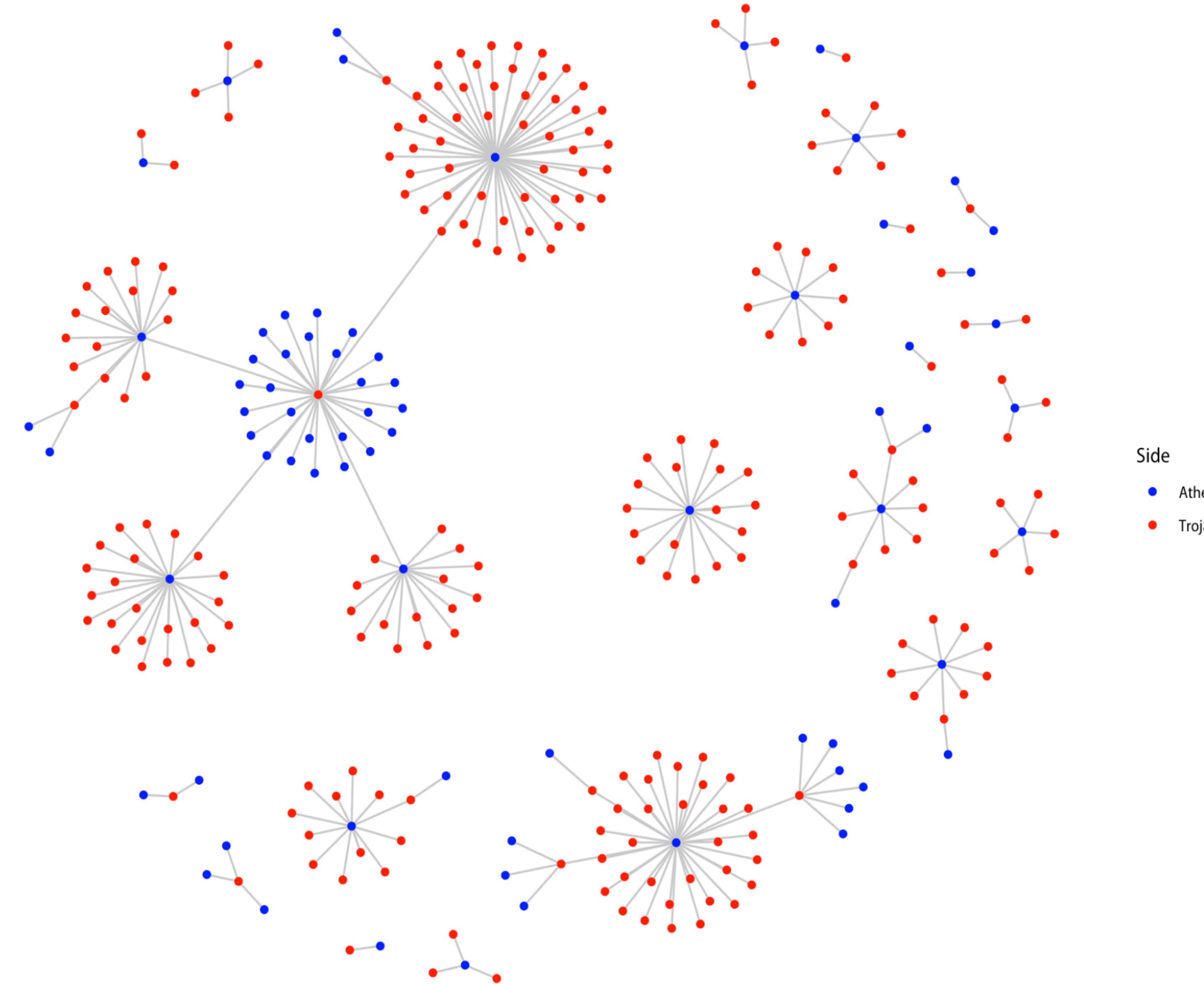
```
il_tidy %>%
  activate(nodes) %>%
  mutate(centrality = centrality_degree()) %>%
  arrange(desc(centrality)) %>%
  top_n(10, wt = centrality) %>%
  ggplot(mapping = aes(x = centrality,
                        y = reorder(name, centrality),
                        color = affil)) +
  geom_point(size = 3) +
  labs(x = "Centrality", y = NULL, color = "Side") +
  theme_minimal() +
  theme(legend.position = "top")
```

Side A T



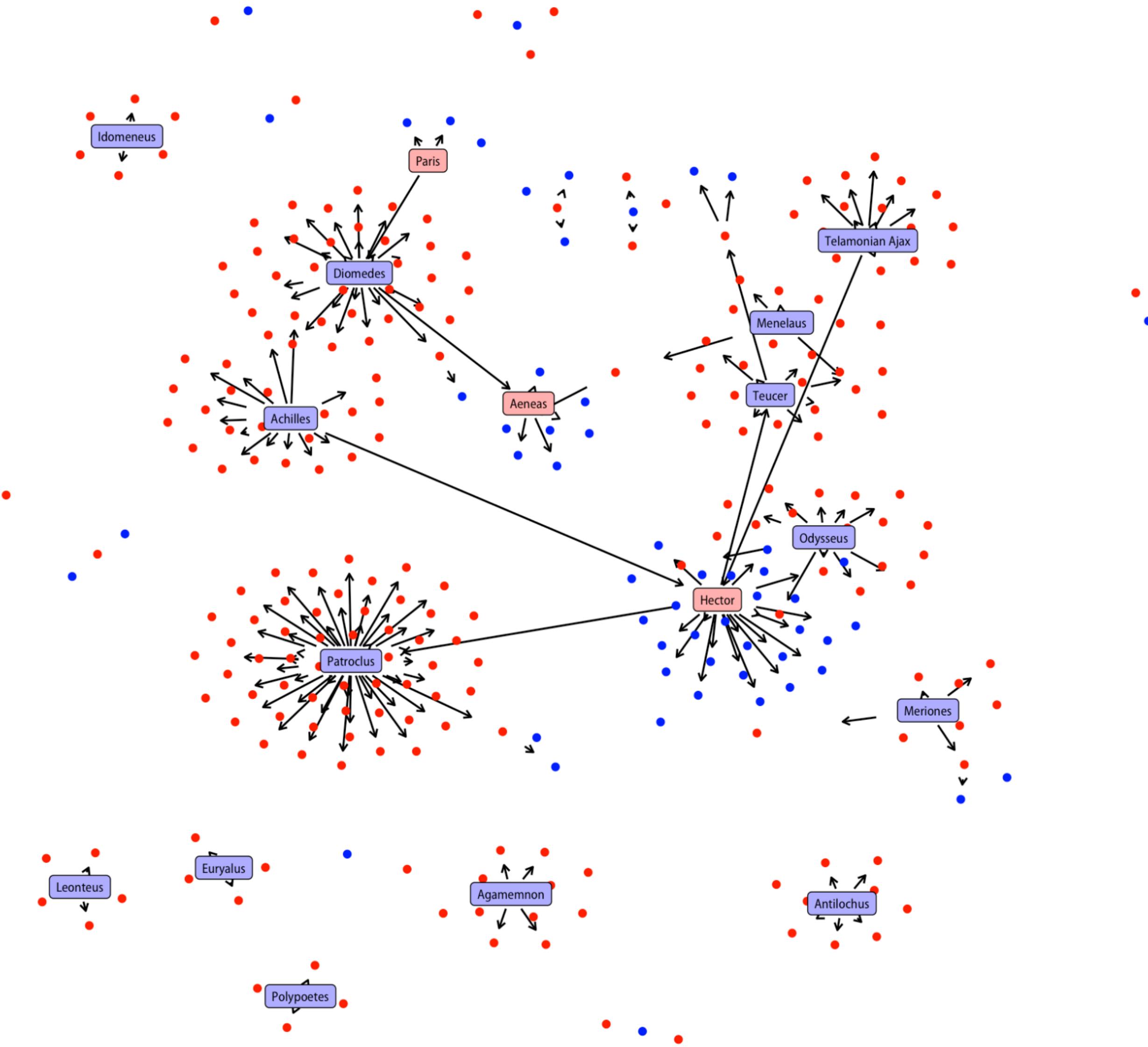
```
il_tidy %>%
  activate(nodes) %>%
  ggraph(layout = "fr") +
  geom_edge_link(color = "gray80") +
  geom_node_point(aes(color = affil)) +
  scale_color_manual(values = c("blue", "red"),
                     labels = c("Athenian", "Trojan")) +
  guides(color = guide_legend(title = "Side", )) +
  labs(title = "Killings in The Iliad") +
  theme(plot.title = element_text(size = rel(3)))
```

# Killings in The Iliad

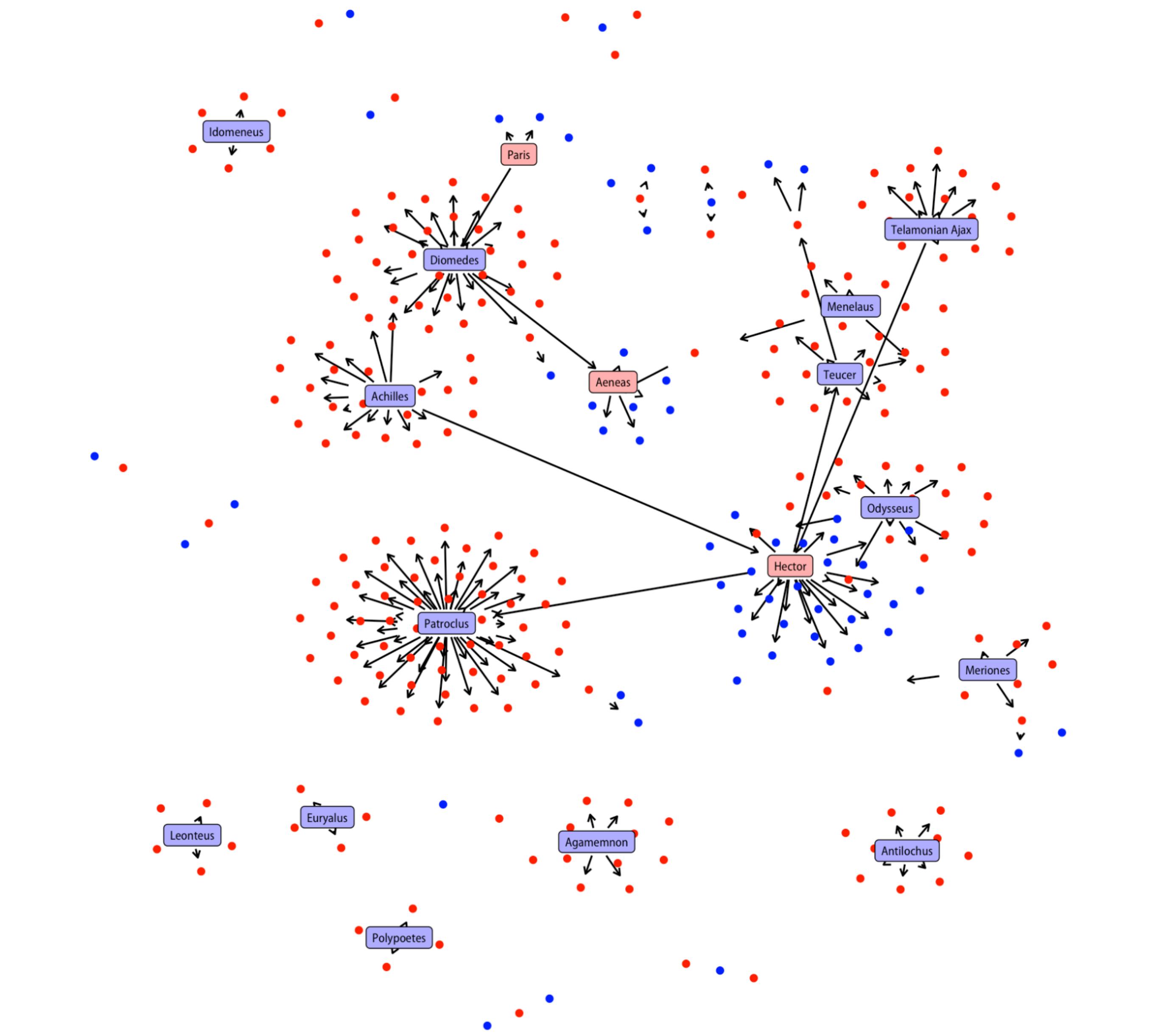


```
label_colors <- c(prismatic::clr_lighten("blue", 0.7),  
                  prismatic::clr_lighten("red", 0.7))  
il_tidy %>%  
  activate(edges) %>%  
  filter(act == "kills") %>%  
  activate(nodes) %>%  
  mutate(centrality = centrality_degree(mode = "out")) %>%  
  ggraph(layout = 'graphopt') +  
  geom_edge_link(aes(start_cap = label_rect(node1.name),  
                    end_cap = label_rect(node2.name)),  
                 arrow = arrow(length = unit(1.5, 'mm'))) +  
  geom_node_point(aes(color = affil)) +  
  scale_color_manual(values = c("blue", "red"),  
                      labels = c("Athenian", "Trojan")) +  
  guides(color = FALSE, fill = FALSE) +  
  geom_node_label(aes(filter = centrality > 3,  
                     label = name, fill = affil),  
                 size = rel(2.5)) +  
  scale_fill_manual(values = label_colors) +  
  labs(title = "Killings in The Iliad") +  
  theme(plot.title = element_text(size = rel(3)))
```

# Killings in The Iliad

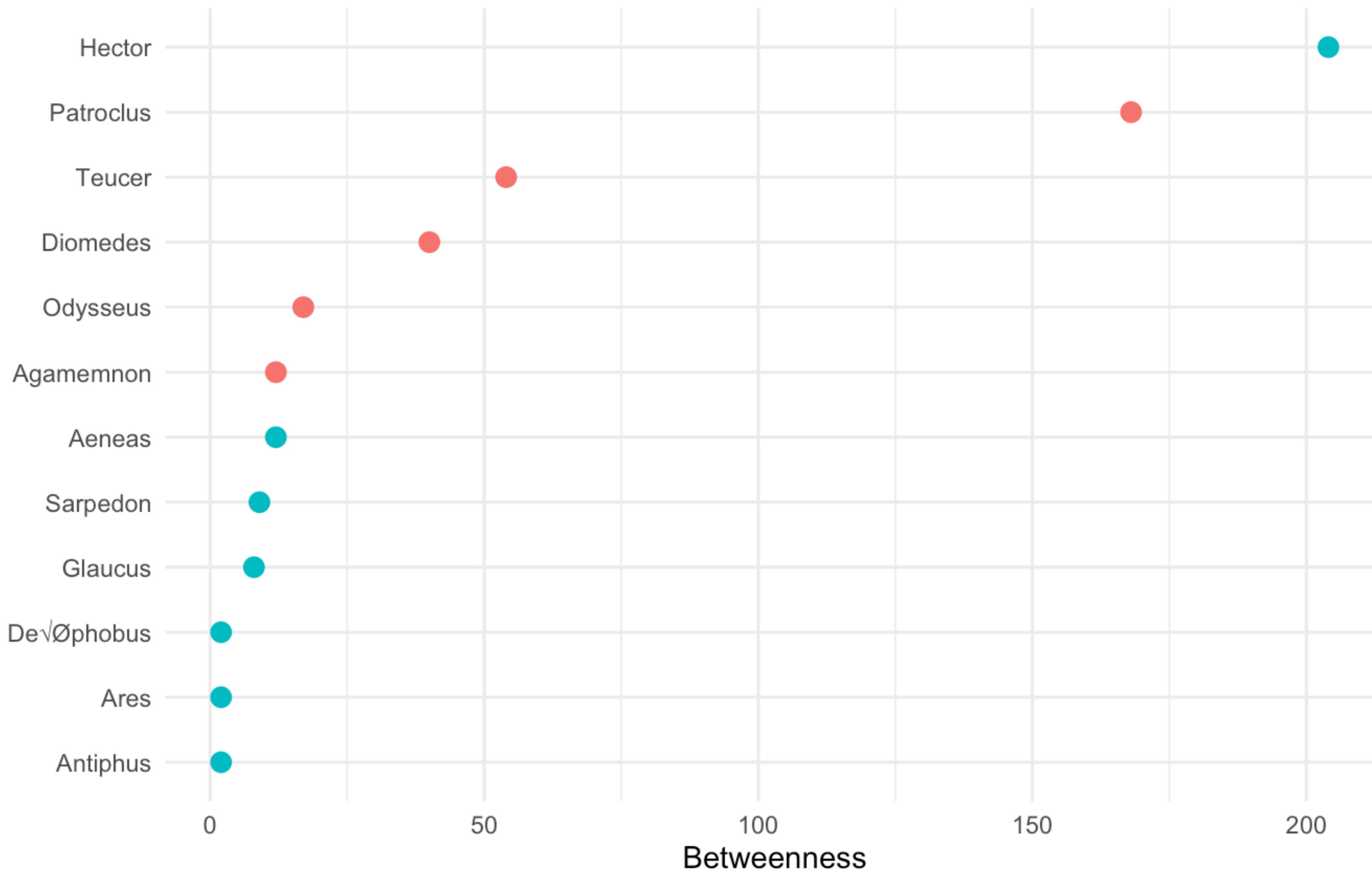


# Killings in The Iliad



```
il_tidy %>%
  activate(nodes) %>%
  mutate(centrality = centrality_degree(),
        betweenness = centrality_betweenness()) %>%
  arrange(desc(betweenness)) %>%
  top_n(10, wt = betweenness) %>%
  ggplot(mapping = aes(x = betweenness,
                        y = reorder(name,
                                    betweenness), color = affil)) +
  geom_point(size = 3) +
  labs(x = "Betweenness", y = NULL, color = "Side") +
  theme_minimal() +
  theme(legend.position = "top")
```

Side A T



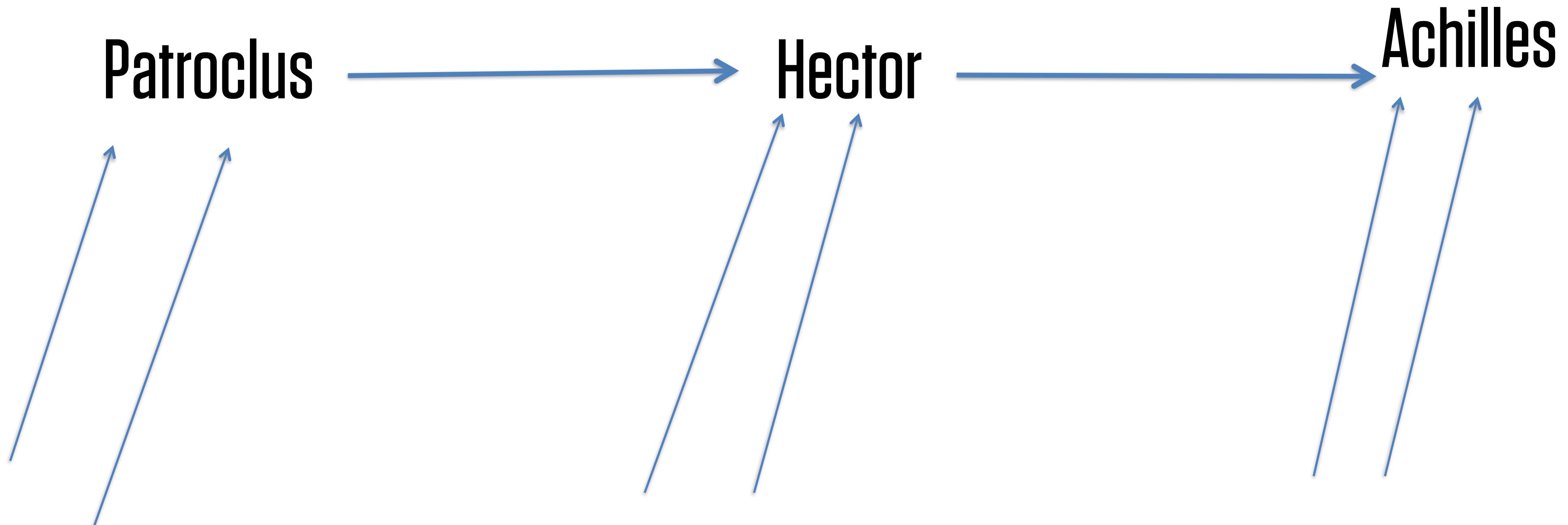


Rage, sing goddess of the rage of **Achilles** son of Peleus

**Patroclus**

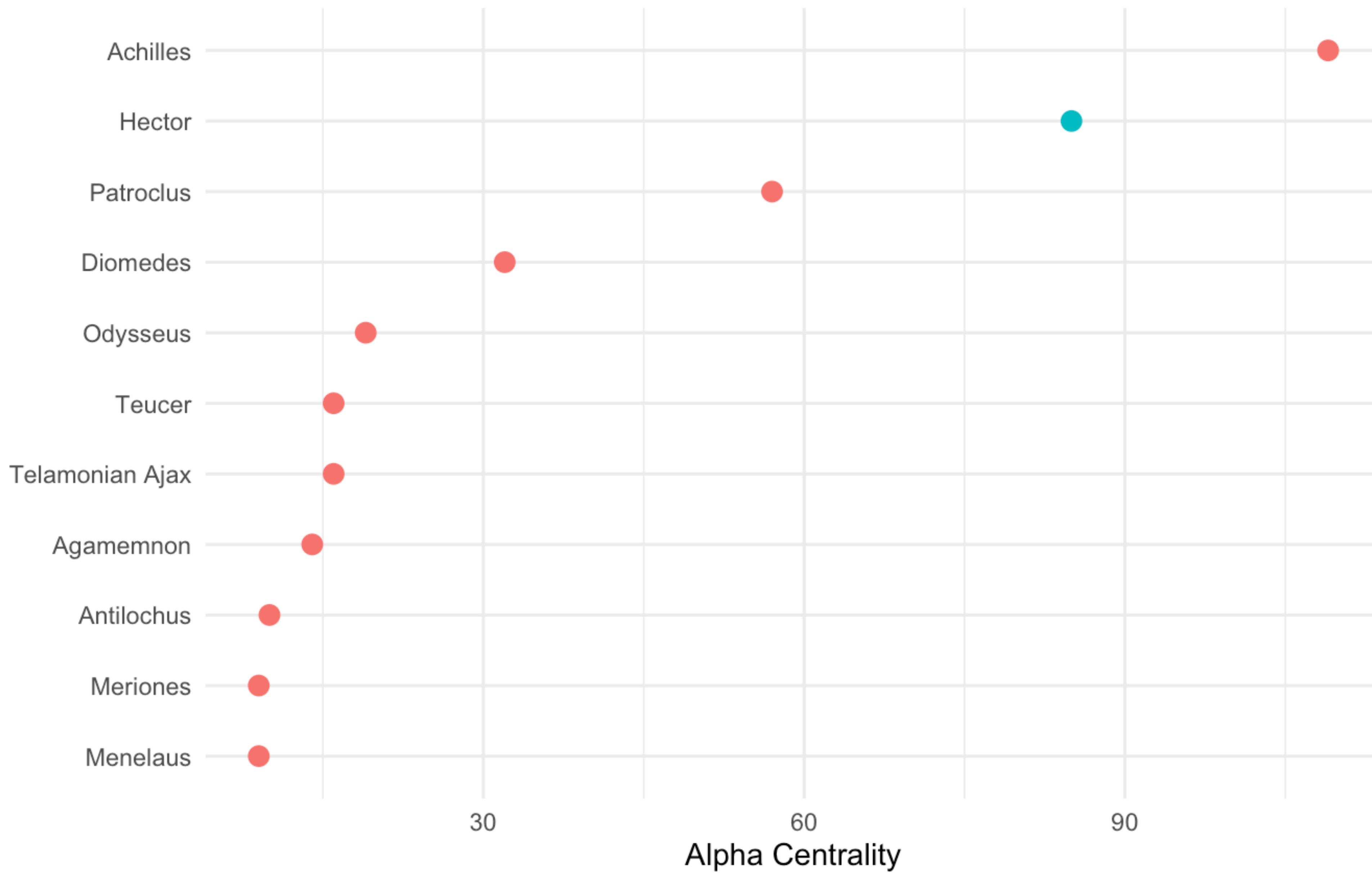
**Hector**

**Achilles**



```
il_tidy %>%
  activate(edges) %>%
  filter(act == "kills") %>%
  reroute(from = to, to = from) %>%
  activate(nodes) %>%
  mutate(alpha = centrality_alpha()) %>%
  arrange(desc(alpha)) %>%
  top_n(10, wt = alpha) %>%
  ggplot(mapping = aes(x = alpha,
                        y = reorder(name, alpha),
                        color = affil)) +
  geom_point(size = 3) +
  labs(x = "Alpha Centrality", y = NULL, color = "Side") +
  theme_minimal() +
  theme(legend.position = "top")
```

Side A T



# Animation

```
library(babynames)
library(gganimate)

> babynames
# A tibble: 1,924,665 x 5
  year sex   name         n     prop
  <dbl> <chr> <chr>     <int>   <dbl>
1 1880 F    Mary      7065 0.0724
2 1880 F    Anna     2604 0.0267
3 1880 F    Emma     2003 0.0205
4 1880 F    Elizabeth 1939 0.0199
5 1880 F    Minnie   1746 0.0179
6 1880 F    Margaret 1578 0.0162
7 1880 F    Ida      1472 0.0151
8 1880 F    Alice    1414 0.0145
9 1880 F    Bertha   1320 0.0135
10 1880 F   Sarah    1288 0.0132
# ... with 1,924,655 more rows
```

```
## Create the plot object
p <- babynames %>%
  filter(sex == "M") %>%
  mutate(endletter = stringr::str_sub(name, -1)) %>%
  group_by(year, endletter) %>%
  summarize(letter_count = n()) %>%
  mutate(letter_prop = letter_count / sum(letter_count),
         rank = min_rank(-letter_prop) * 1) %>%
  ungroup() %>%
  ggplot(aes(x = factor(endletter, levels = letters, ordered = TRUE),
             y = letter_prop,
             group = endletter,
             fill = factor(endletter),
             color = factor(endletter))) +
  geom_col(alpha = 0.8) +
  scale_y_continuous(labels = scales::percent_format(accuracy = 1)) +
  guides(color = FALSE, fill = FALSE) +
  labs(title = "Distribution of Last Letters of U.S. Girls' Names over Time",
       subtitle = '{closest_state}',
       x = "", y = "Names ending in letter",
       caption = "Data: US Social Security Administration. @kjhealy / socviz.co") +
  theme(plot.title = element_text(size = rel(2)),
        plot.subtitle = element_text(size = rel(3)),
        plot.caption = element_text(size = rel(2)),
        axis.text.x = element_text(face = "bold", size = rel(3)),
        axis.text.y = element_text(size = rel(3)),
        axis.title.y = element_text(size = rel(2))) +
  transition_states(year, transition_length = 4, state_length = 1) +
  ease_aes('cubic-in-out')
```

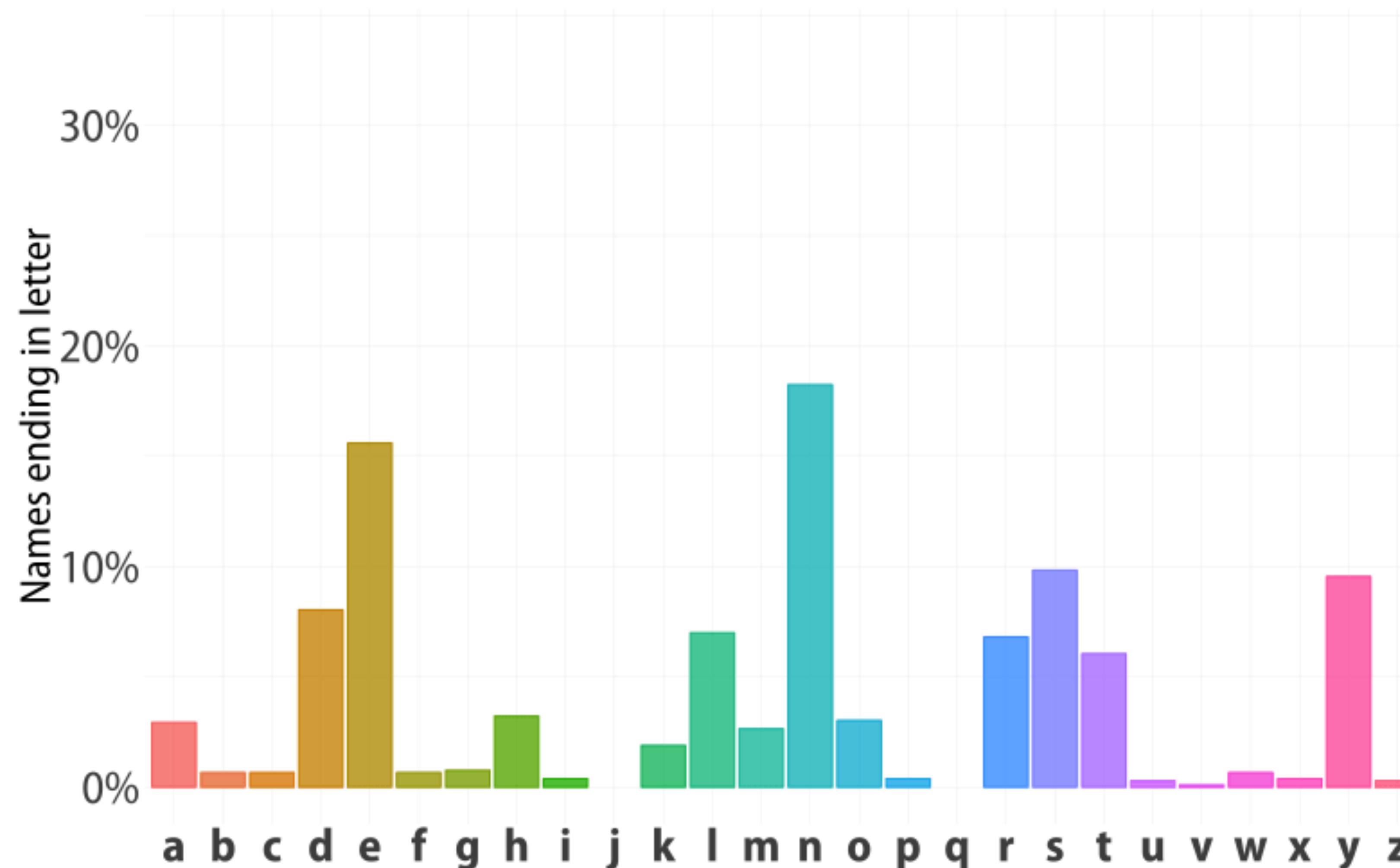
```
# A tibble: 3,424 x 5
  year endletter letter_count letter_prop rank
  <dbl> <chr>          <int>        <dbl> <dbl>
1 1880 a              31       0.0293     11
2 1880 b              7        0.00662    15
3 1880 c              7        0.00662    15
4 1880 d             85       0.0803     5
5 1880 e             165      0.156      2
6 1880 f              7        0.00662    15
7 1880 g              8        0.00756    14
8 1880 h             34       0.0321     9
9 1880 i              4        0.00378   19
10 1880 k             20       0.0189    13
# ... with 3,414 more rows
```

```
## Create the plot object
p <- babynames %>%
  filter(sex == "M") %>%
  mutate(endletter = stringr::str_sub(name, -1)) %>%
  group_by(year, endletter) %>%
  summarize(letter_count = n()) %>%
  mutate(letter_prop = letter_count / sum(letter_count),
         rank = min_rank(-letter_prop) * 1) %>%
  ungroup() %>%
  ggplot(aes(x = factor(endletter, levels = letters, ordered = TRUE),
             y = letter_prop,
             group = endletter,
             fill = factor(endletter),
             color = factor(endletter))) +
  geom_col(alpha = 0.8) +
  scale_y_continuous(labels = scales::percent_format(accuracy = 1)) +
  guides(color = FALSE, fill = FALSE) +
  labs(title = "Distribution of Last Letters of U.S. Girls' Names over Time",
       subtitle = '{closest_state}',
       x = "", y = "Names ending in letter",
       caption = "Data: US Social Security Administration. @kjhealy / socviz.co") +
  theme(plot.title = element_text(size = rel(2)),
        plot.subtitle = element_text(size = rel(3)),
        plot.caption = element_text(size = rel(2)),
        axis.text.x = element_text(face = "bold", size = rel(3)),
        axis.text.y = element_text(size = rel(3)),
        axis.title.y = element_text(size = rel(2))) +
  transition_states(year, transition_length = 4, state_length = 1) +
  ease_aes('cubic-in-out')
```

```
animate(p, fps = 25, duration = 20, width = 800, height = 600,  
       renderer = gifski_renderer("figures/name_endings_boys.gif"))
```

## Distribution of Last Letters of U.S. Boys' Names over Time

1880



Data: US Social Security Administration. @kjhealy / socviz.co

# Tiles and Labels

## CLASS I.

Institutions whose graduates would ordinarily be able to take the master's degree at any of the large graduate schools in one year after receiving the bachelor's degree, without necessarily doing more than the amount of work regularly prescribed for such higher degree.

```
> data_allu
# A tibble: 206 x 13
  Rank School Babcock PubPriv Tuition Enrollment Acceptance Retention Graduation Type Dummy sname
  <dbl> <fct>   <chr>    <chr>    <dbl>      <dbl>      <dbl>      <dbl>      <dbl> <chr> <dbl> <chr>
1    152 Adelp... Class 2 Private 30800       7859       66.5        81        66 Univ... 1 Adel...
2     75 Ameri... Not Ra... Private 40649      12904       44.2        90        77 Univ... 1 Amer...
3    181 Andre... Not Ra... Private 25470       3551       37.5        79        59 Univ... 1 Andr...
4    142 Arizo... Not Ra... Public 10002       73378       87.9        82        57 Univ... 1 Ariz...
5     91 Aubur... Not Ra... Public  9852       25134       77.2        88        68 Univ... 1 Aubu...
6    173 Azusa... Not Ra... Private 32256      10184       52.3        85        63 Univ... 1 Azus...
7    181 Ball ... Not Ra... Public  9250       21053       61.2        79        57 Univ... 1 Ball...
8     75 Baylo... Class 2 Private 35972      15364       60.7        85        75 Univ... 1 Bayl...
9     97 Bingh... Not Ra... Public  8144       15308       42.9        91        79 Univ... 1 Bing...
10   177 Biola... Not Ra... Private 32142       6302       74.7        85        65 Univ... 1 Biola
# ... with 196 more rows, and 1 more variable: usnwr_grp <fct>
```

## CLASS IV.

Institutions whose bachelor's degree would be approximately two years short of equivalency with the standard bachelor's degree of a standard college as described above. It should be said in connection with this class that the information upon which to base judgment of individual institutions is less sufficient and satisfactory, and in larger proportion drawn from catalogues, than is the case for the other classes, since a relatively smaller proportion of the graduates of institutions in this class appears in the registration in graduate and professional schools. Presumably a much larger number of institutions will

```
> data_allu
# A tibble: 206 x 13
  Rank School Babcock PubPriv Tuition Enrollment Acceptance Retention Graduation Type Dummy sname
  <dbl> <fct>   <chr>    <chr>     <dbl>      <dbl>      <dbl>      <dbl>      <dbl> <chr>    <dbl> <chr>
1 152 Adelp... Class 2 Private 30800       7859       66.5        81        66 Univ... 1 Adel...
2 75 Ameri... Not Ra... Private 40649       12904       44.2        90        77 Univ... 1 Amer...
3 181 Andre... Not Ra... Private 25470       3551        37.5        79        59 Univ... 1 Andr...
4 142 Arizo... Not Ra... Public 10002       73378       87.9        82        57 Univ... 1 Ariz...
5 91 Aubur... Not Ra... Public  9852        25134       77.2        88        68 Univ... 1 Aubu...
6 173 Azusa... Not Ra... Private 32256       10184       52.3        85        63 Univ... 1 Azus...
7 181 Ball ... Not Ra... Public  9250        21053       61.2        79        57 Univ... 1 Ball...
8 75 Baylo... Class 2 Private 35972       15364       60.7        85        75 Univ... 1 Bayl...
9 97 Bingh... Not Ra... Public  8144        15308       42.9        91        79 Univ... 1 Bing...
10 177 Biola... Not Ra... Private 32142       6302        74.7        85        65 Univ... 1 Biola
# ... with 196 more rows, and 1 more variable: usnwr_grp <fct>
```

```
p <- ggplot(mapping = data_allu, aes(x = Dummy, y = reorder(sname, -Rank),  
fill = Babcock,  
label = sname))  
  
p + geom_tile() +  
facet_wrap( ~ usnwr_grp, nrow = 1, scales = "free_y") +  
geom_label(fill = "#FFFFFF", alpha = 0.9, size = rel(1.8)) +  
scale_fill_viridis_d(option = "D", direction = -1) +  
guides(fill = guide_legend(title="Babcock Class in 1911",  
title.position = "top")) +  
labs(x = NULL, y = NULL,  
title = "The Persistence of the Old Regime",  
subtitle = "1911 Babcock Classification and 2014 US News Rankings",  
caption = "Kieran Healy. http://kieranhealy.org") +  
theme(strip.text.x = element_text(size = rel(0.8), face = "bold"),  
axis.ticks=element_blank(),  
axis.text.x = element_blank(),  
axis.text.y = element_blank(),  
legend.title = element_text(size = rel(0.9)),  
panel.grid.major.x = element_blank(),  
panel.grid.minor.x = element_blank(),  
legend.position = "top",  
legend.justification = "left")
```

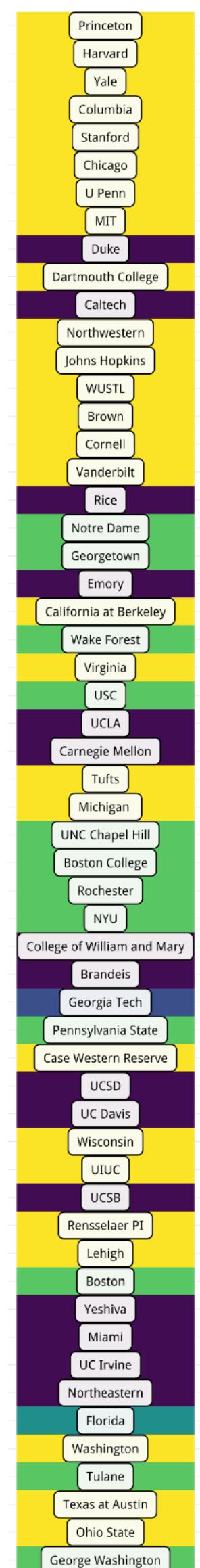
# The Persistence of the Old Regime

1911 Babcock Classification and 2014 US News Rankings

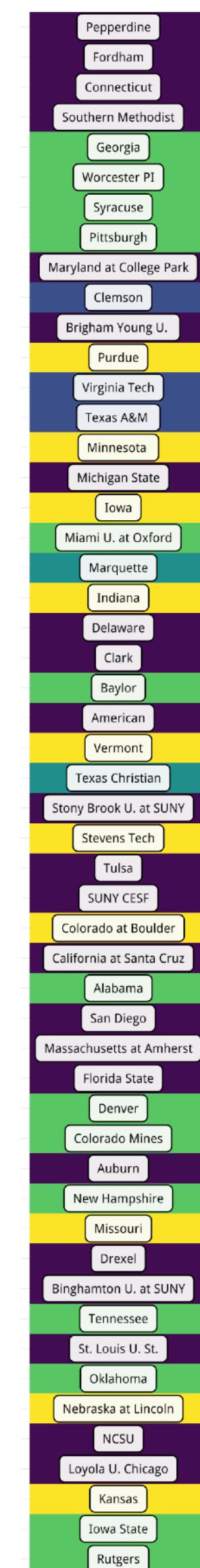
Babcock Class in 1911

Class 1   Class 2   Class 3   Class 4   Not Rated/Not Yet Founded

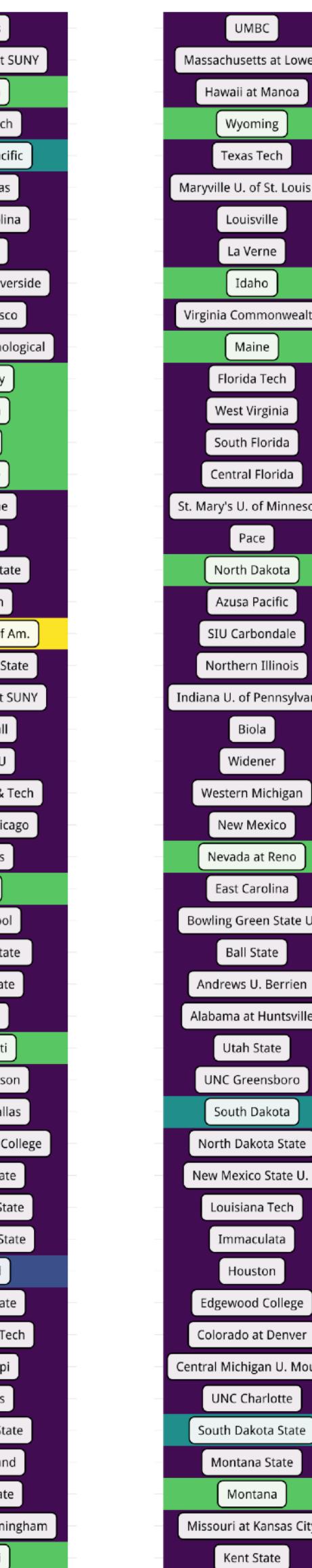
USNWR 1-52



USNWR 53-101



USNWR 101-152



USNWR 153-200



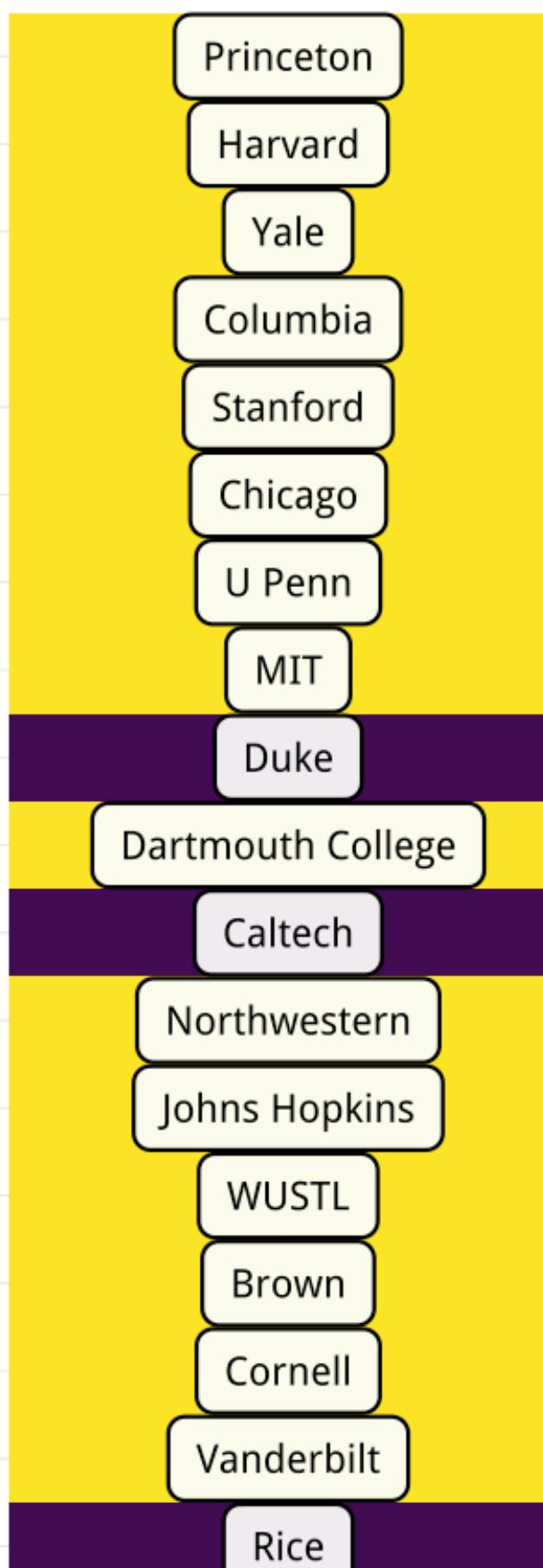
# The Persistence of the Old Regime

1911 Babcock Classification and 2014 US News Rankings

Babcock Class in 1911



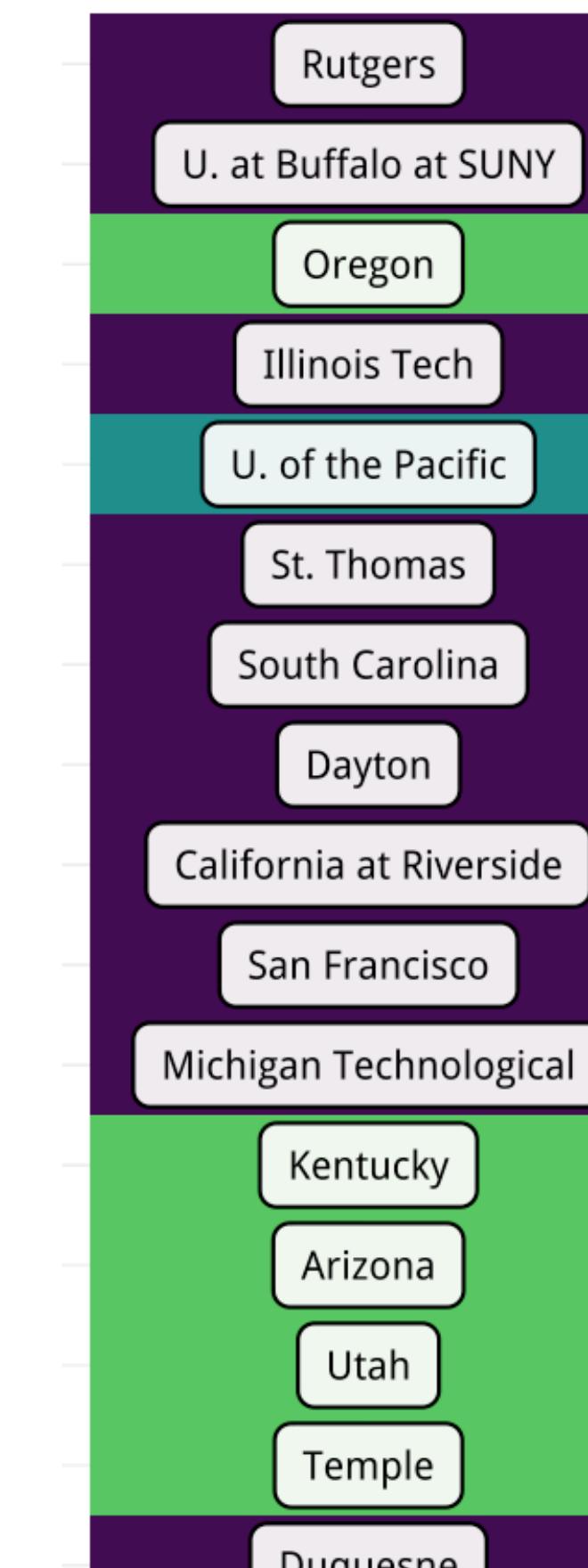
## USNWR 1-52



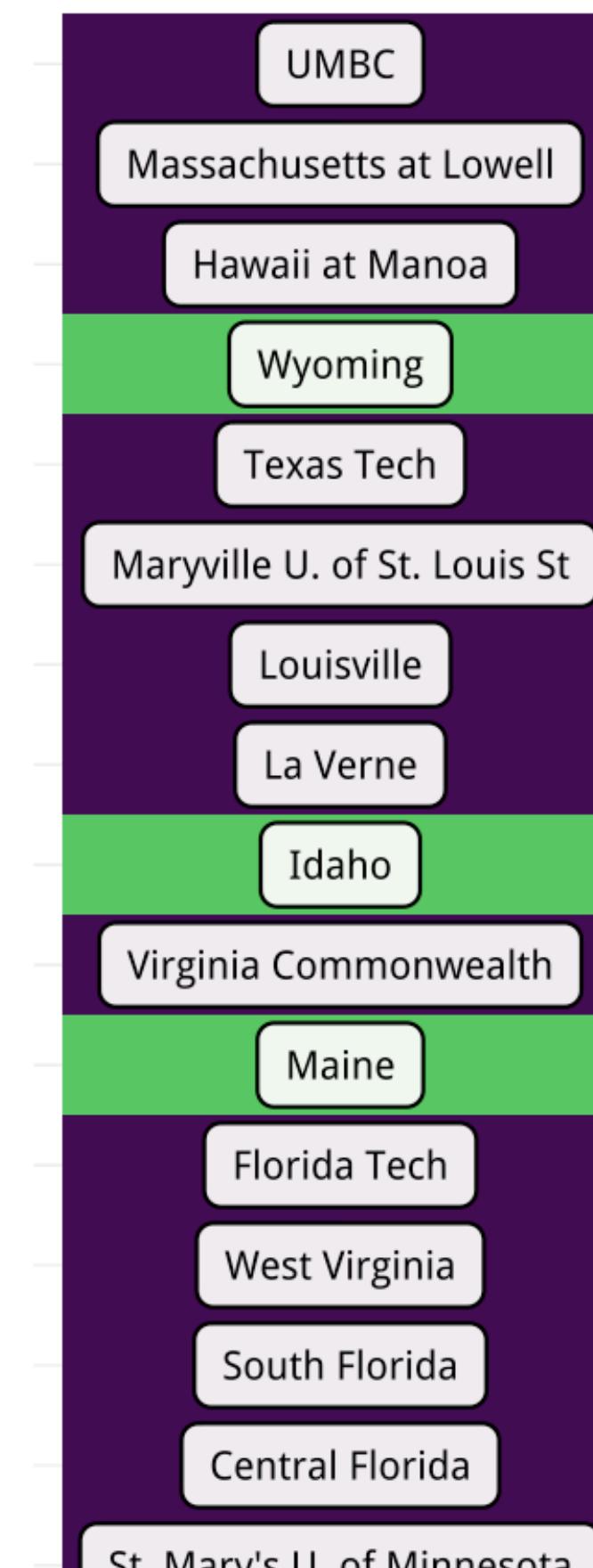
## USNWR 53-101



## USNWR 101-152



## USNWR 153-200



```
add_xy(x = 1, y = 7)
```

```
## [1] 8
```

```
add_xy <- function(x, y) {  
  x + y  
}
```

```
add_xy(x = 5, y = 2)
```

```
## [1] 7
```

```
plot_section <- function(section="Culture", x = "Year",
                        y = "Members", data = asasec,
                        smooth=FALSE){

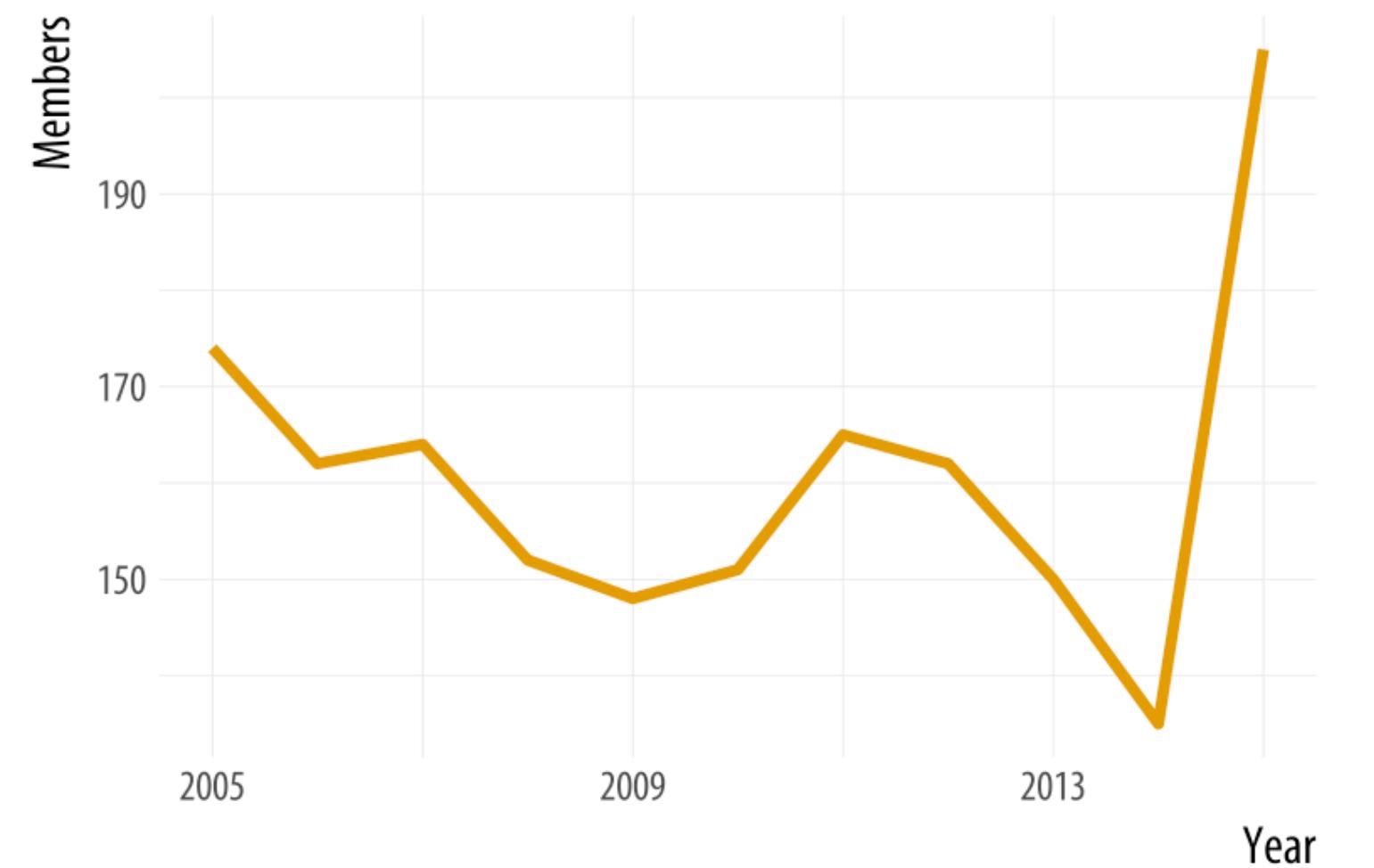
  require(ggplot2)
  require(splines)
  # Note use of aes_string() rather than aes()
  p <- ggplot(subset(data, Sname==section),
               mapping = aes_string(x=x, y=y))

  if(smooth == TRUE) {
    p0 <- p + geom_smooth(color = "#999999",
                            size = 1.2, method = "lm",
                            formula = y ~ ns(x, 3)) +
    scale_x_continuous(breaks = c(seq(2005, 2015, 4))) +
    labs(title = section)
  } else {
    p0 <- p + geom_line(color= "#E69F00", size=1.2) +
    scale_x_continuous(breaks = c(seq(2005, 2015, 4))) +
    labs(title = section)
  }

  print(p0)
}
```

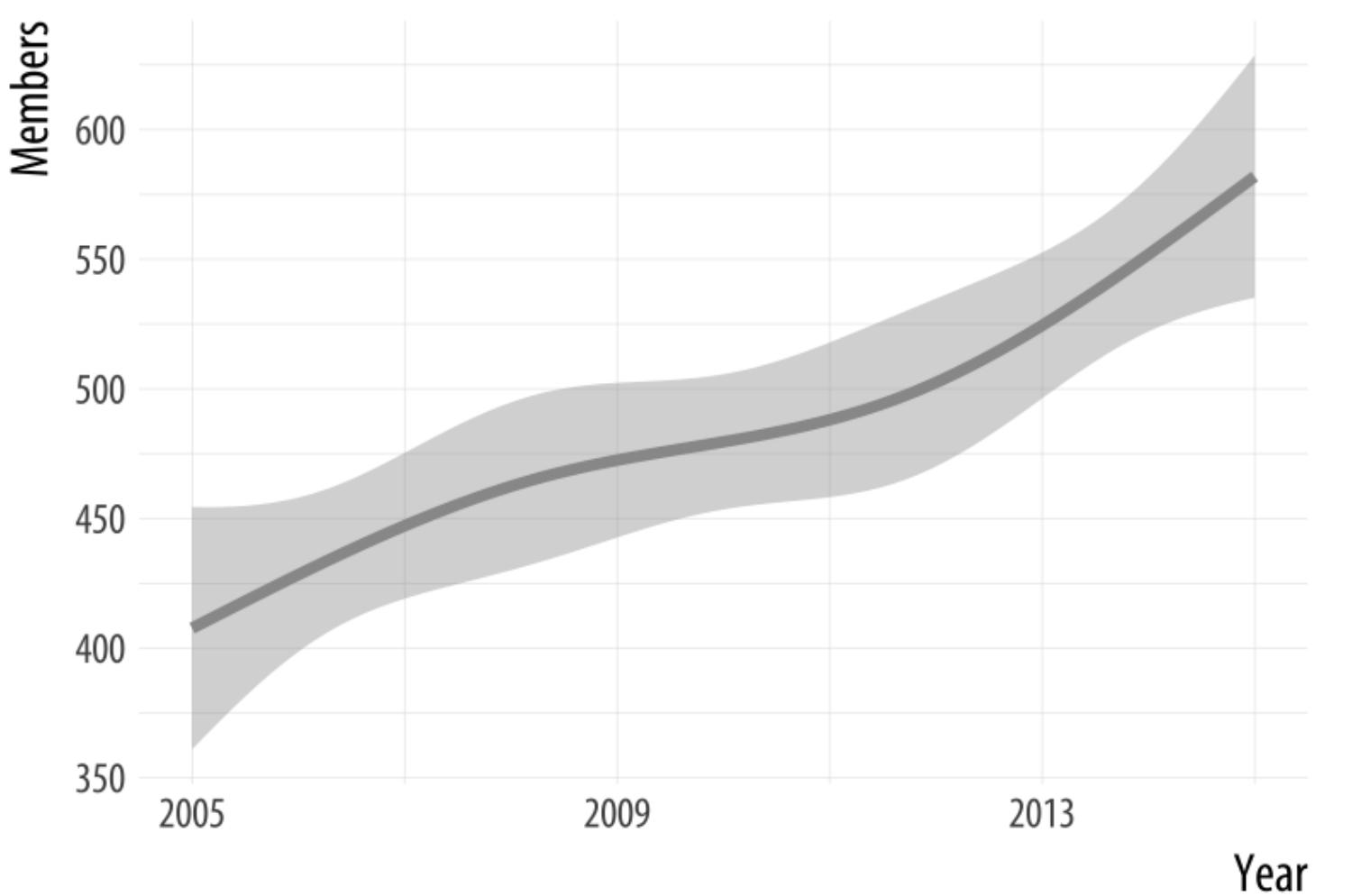
```
plot_section("Rationality")
```

### Rationality



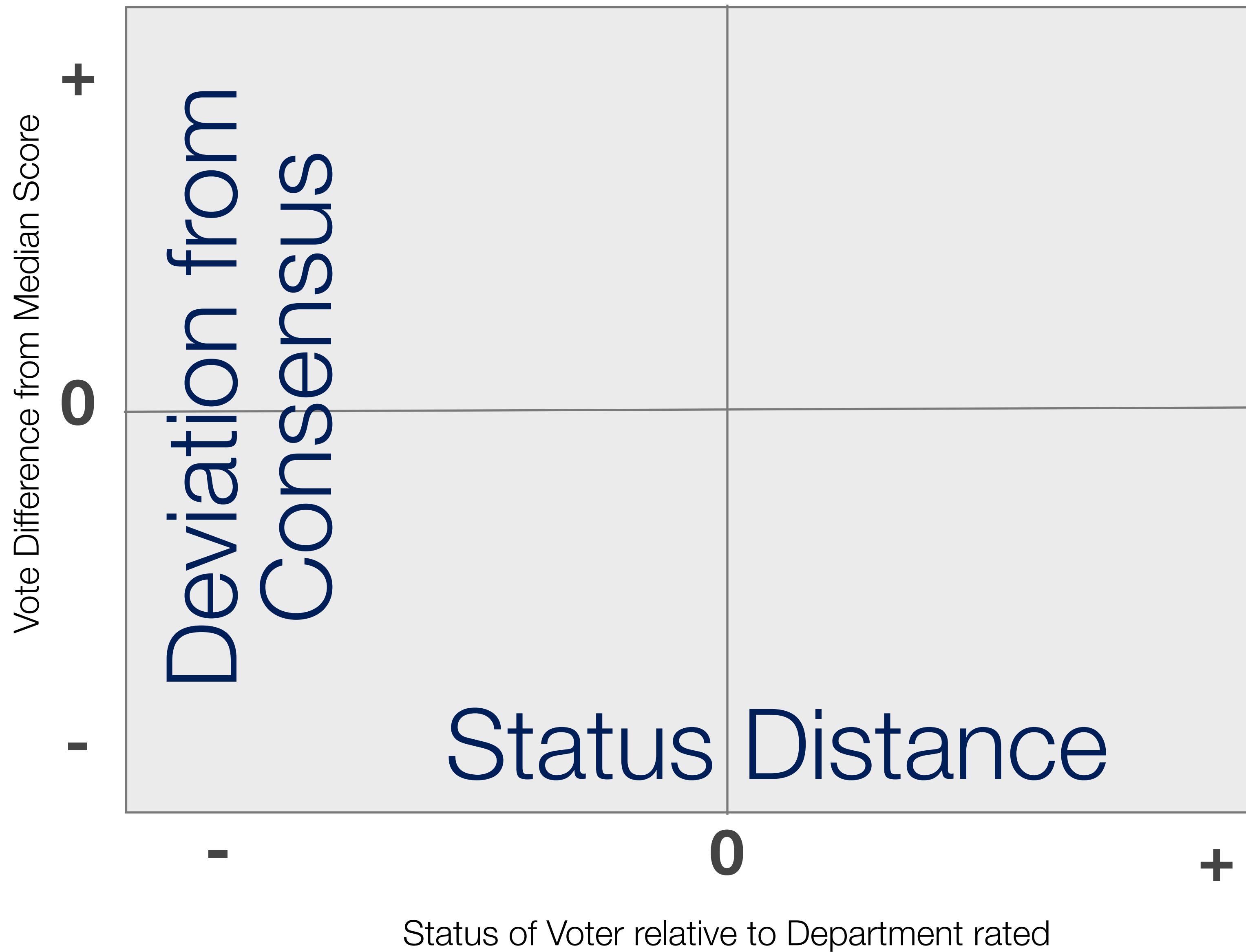
```
plot_section("Sexualities", smooth = TRUE)
```

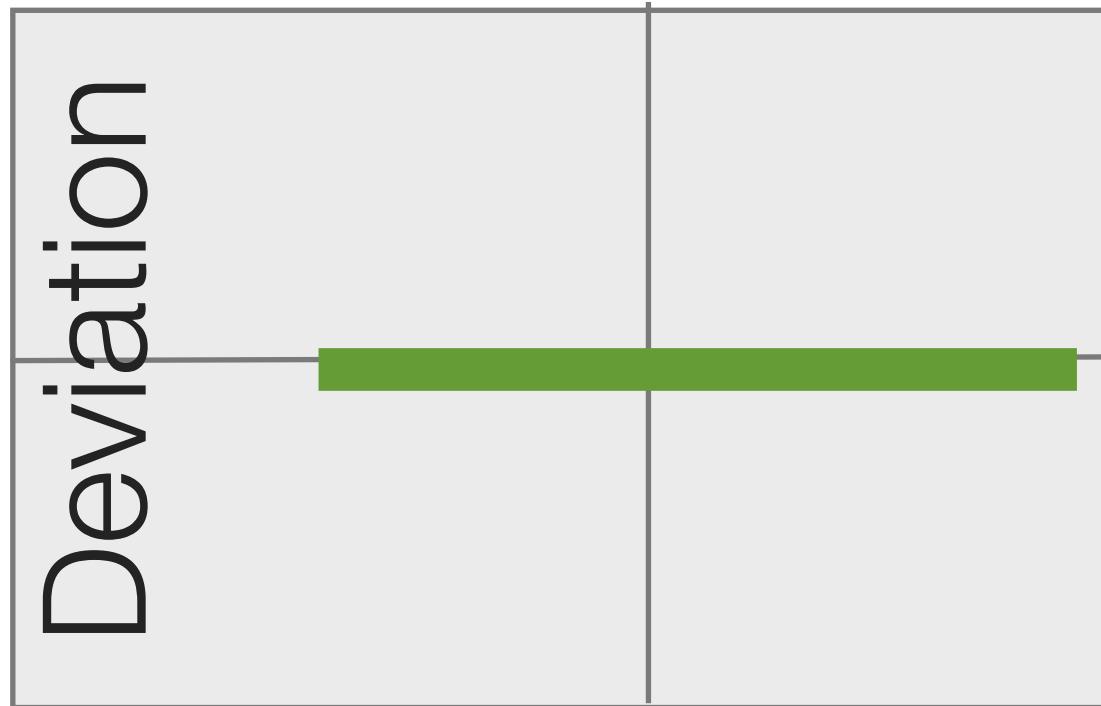
### Sexualities



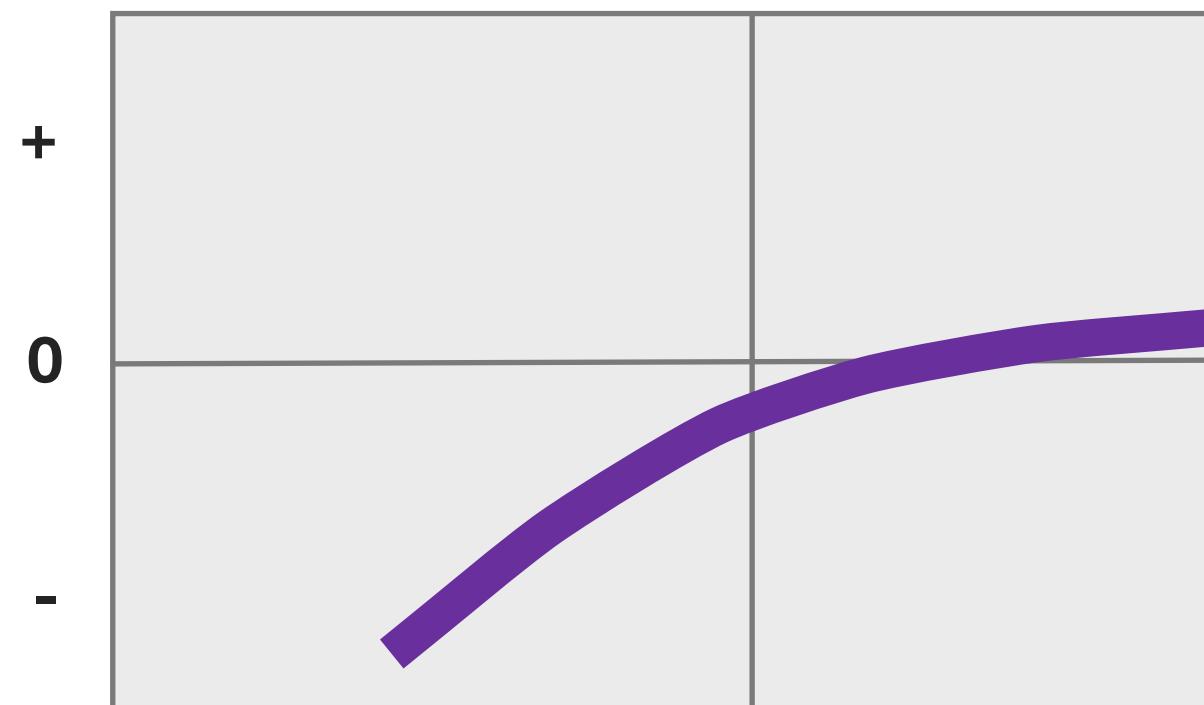
**Layer  
Highlight  
Repeat**

# Build from Ideas to Data

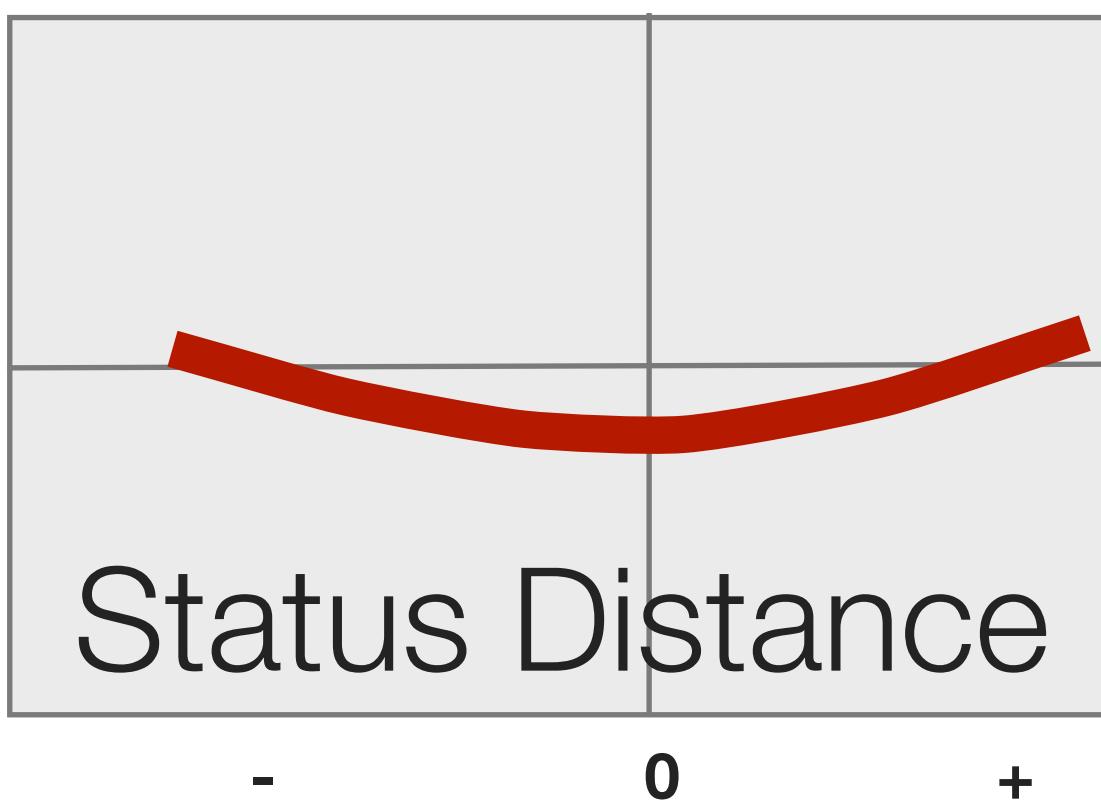




## 1. Pure Objectivity

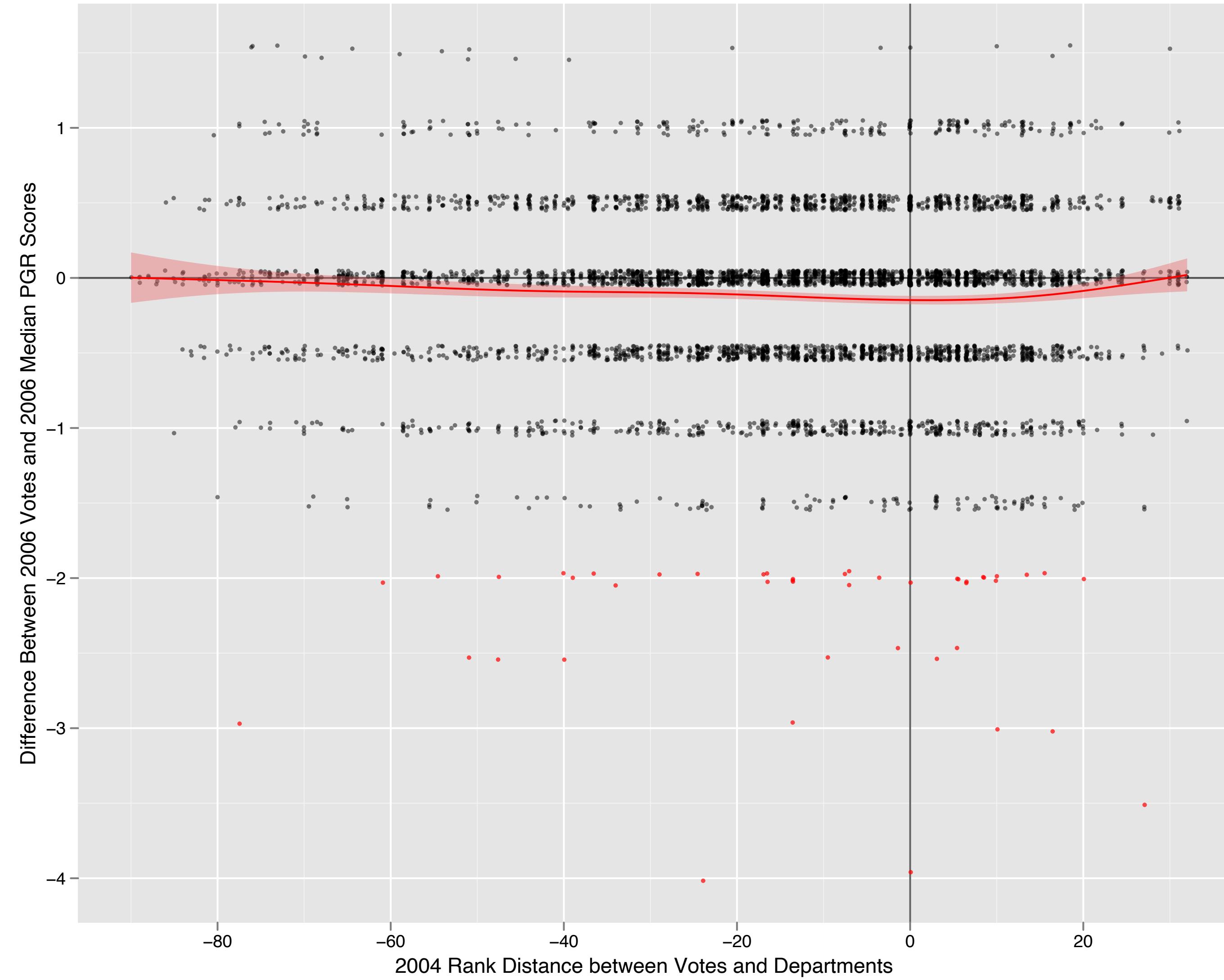


## 2. Distant Envy



## 3. Local Competition

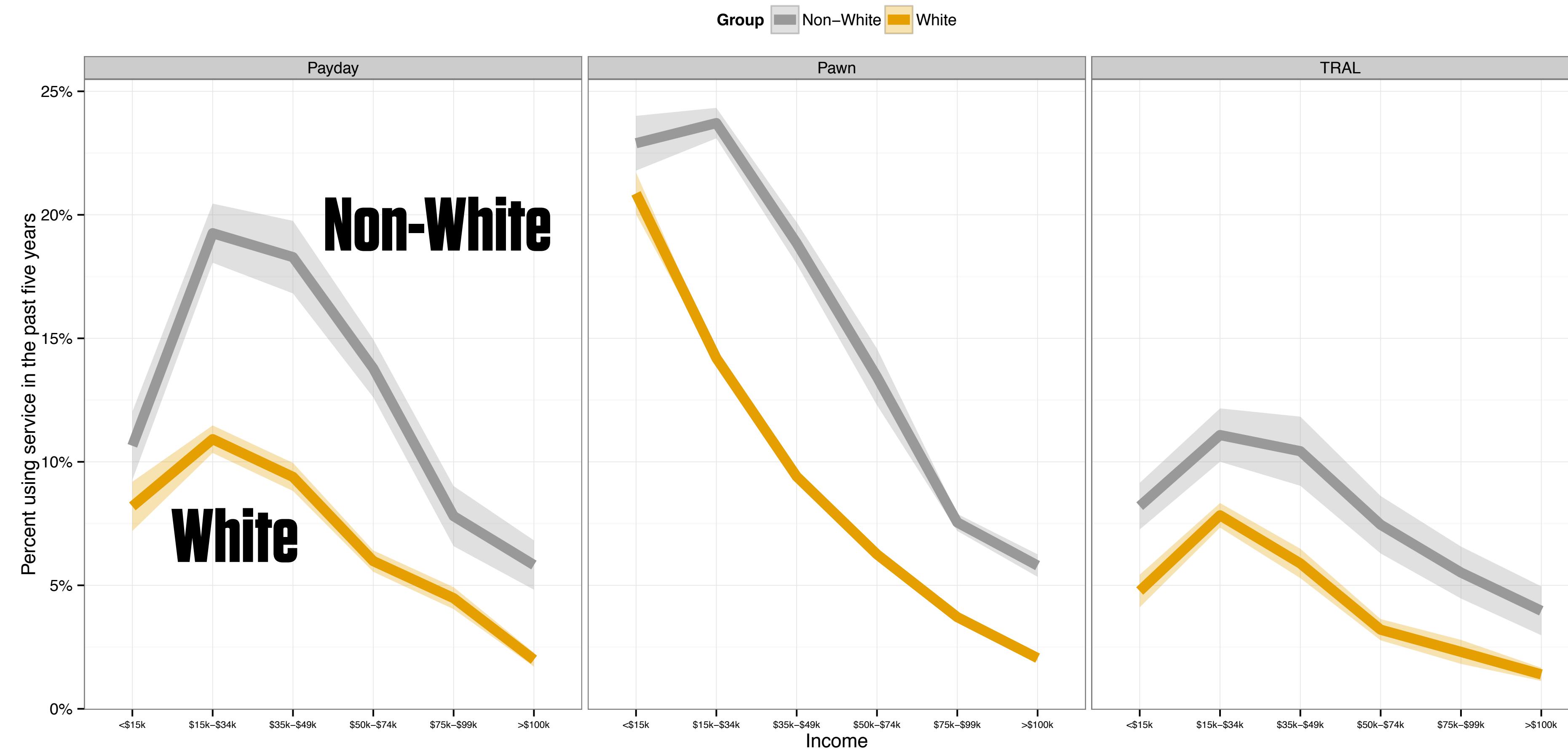
# Deviation



# Status Distance

Leverage ggplot's  
grammar, and its  
layered approach

# Categorical Gaps: Alternative Financial Services

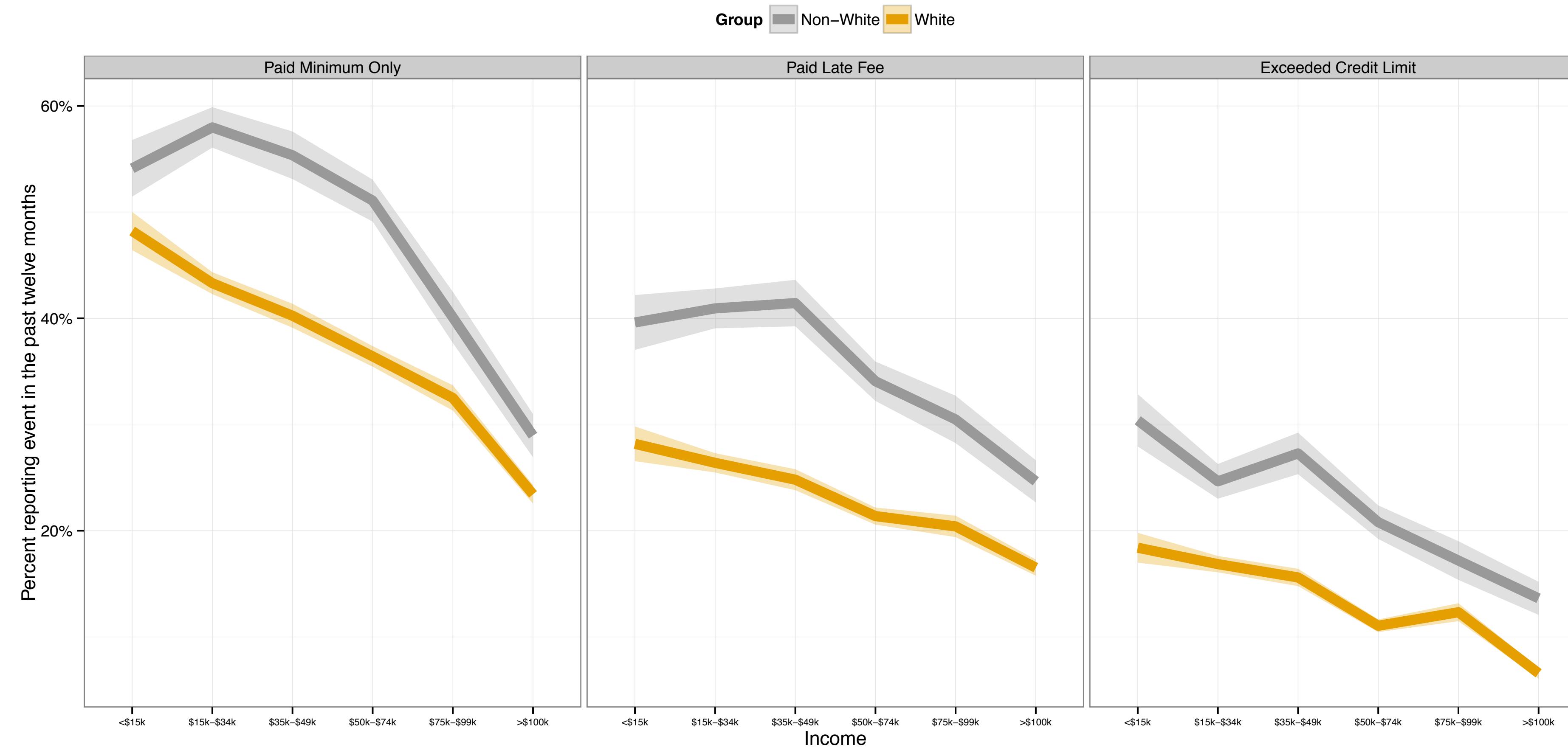


## Payday Loan

## Pawn Shop

## TRAL

# Categorical Gaps: Adverse Credit Events



Minimum Only

Late Fee

Over Limit