EDA

Derek Situ

2024-10-10

```
pacman::p_load(tidyverse, here, janitor, fixest, modelsummary)

# Load data
data <- read_csv(here("data", "processed", "data.csv"))
```

```
## Rows: 27689 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (3): institution, state, level
## dbl (18): institution_id, ban_state, ban_start, years_after_ban, year, white...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# Load top institutions data
top_institutions <- read_csv(here("data", "raw", "top_institutions.csv"))
```

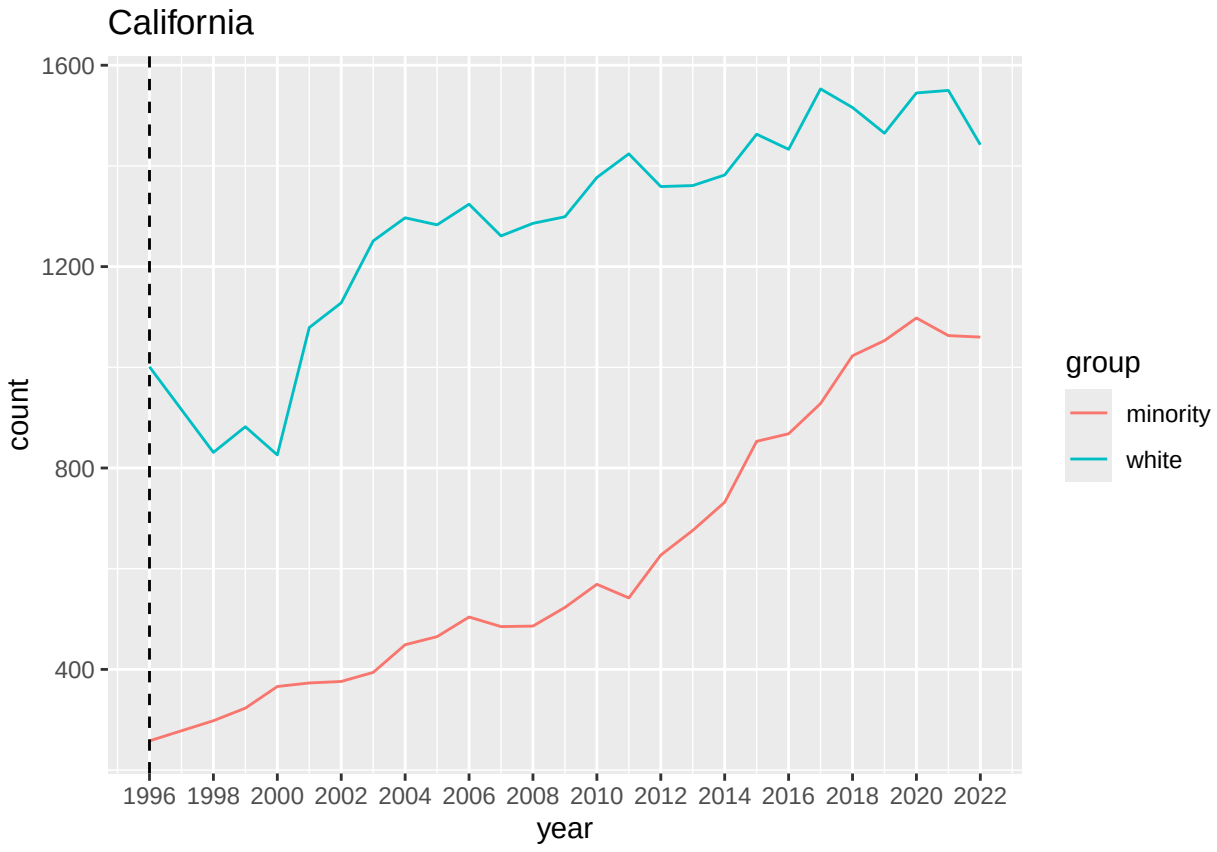
```
## Rows: 75 Columns: 1
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): institution
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

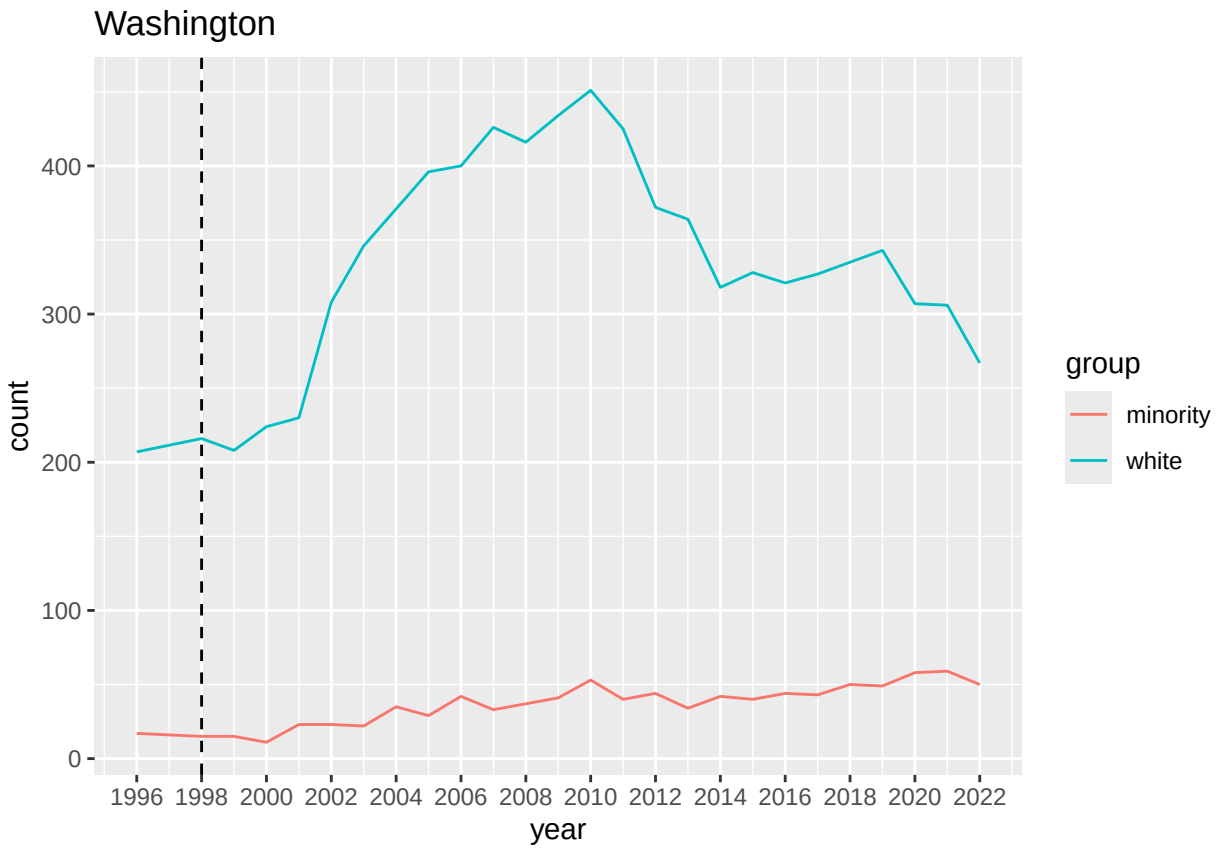
```
# Load bans data
bans <- read_csv(here("data", "raw", "bans.csv"))
```

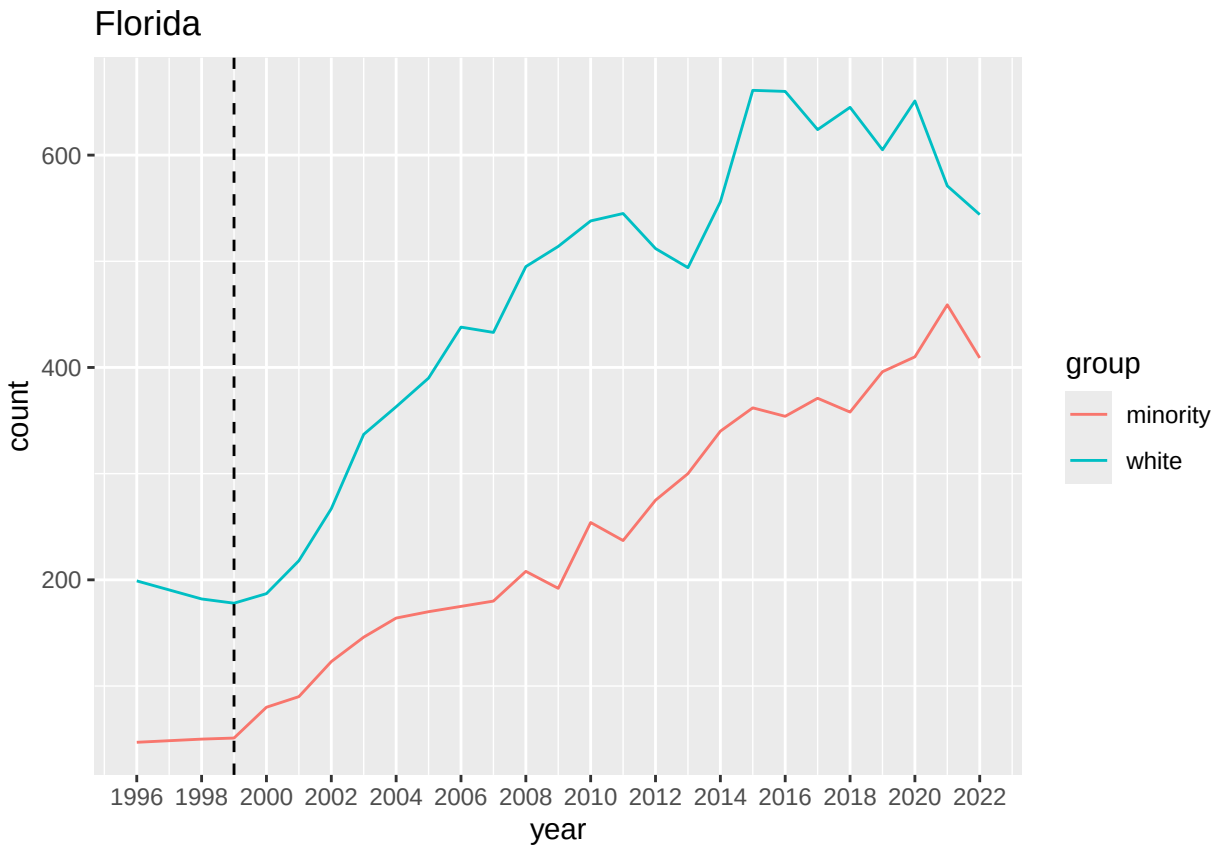
```
## Rows: 9 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): state
## dbl (1): ban_start
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

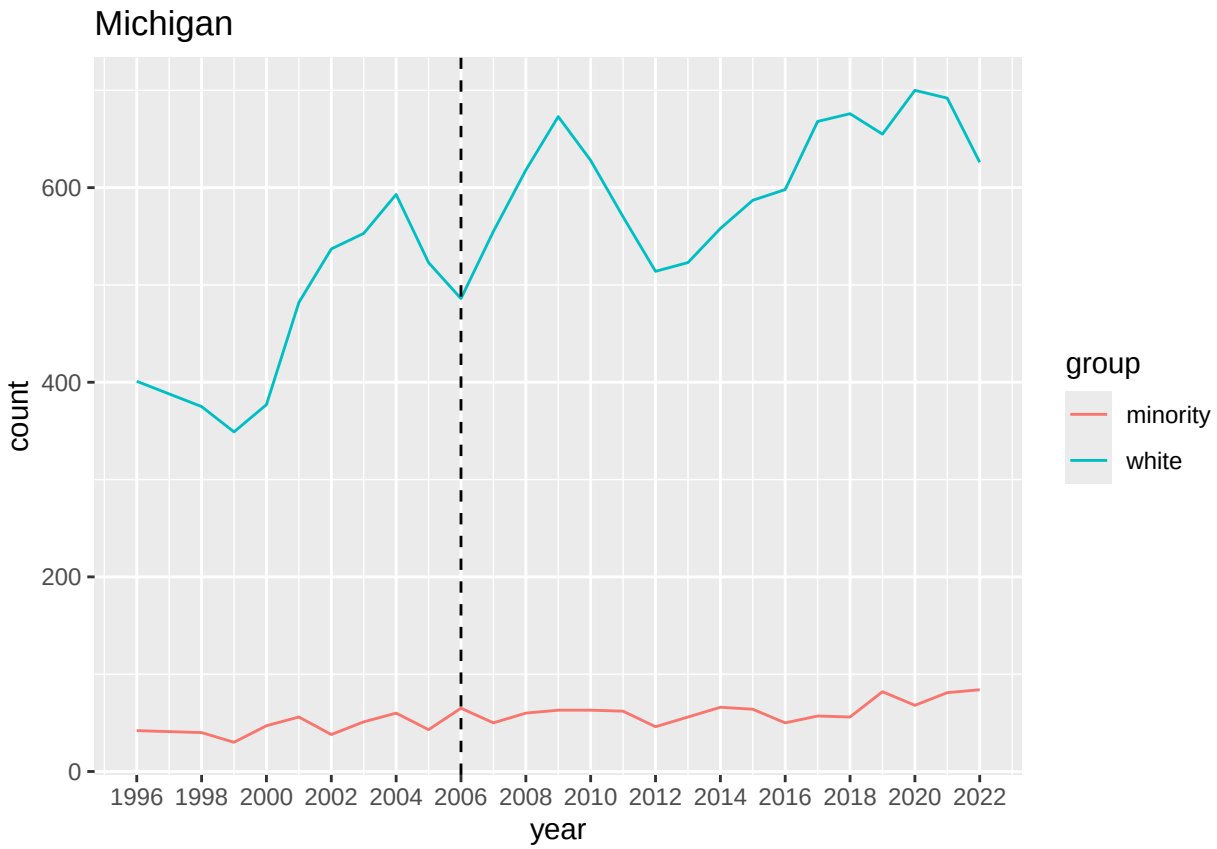
```
# Plot ----------------------------------------------------------------------
sums_by_state <- data %>%
  #filter(top_institution == 1) %>%
```

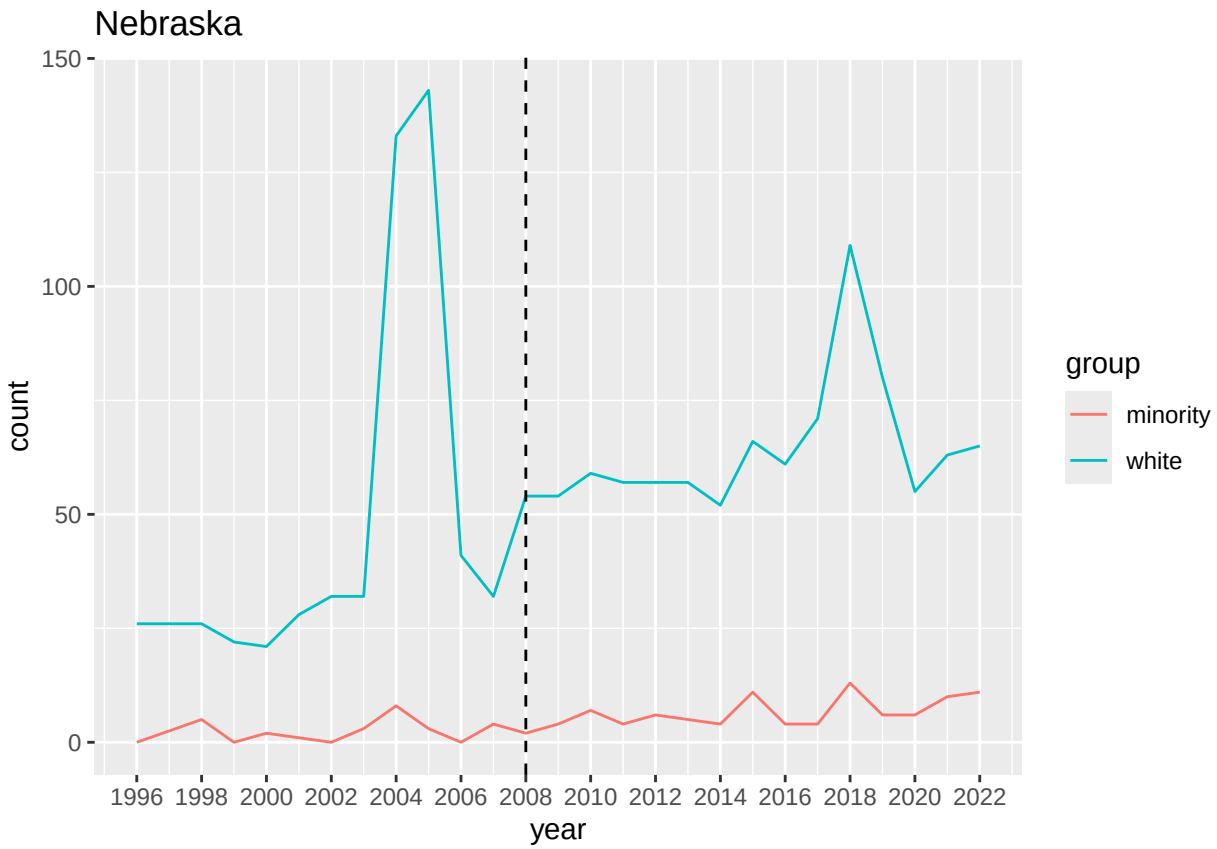
```r
  group_by(state, ban_start, year, level) %>%
  summarize_at(vars(white_men:white), sum, na.rm = TRUE) %>%
  pivot_longer(cols = white_men:white,
               names_to = "group", values_to = "count")

# Plots for minority/white
for (stat in unique(bans$state)) {
  sums_by_state_plot <- sums_by_state %>%
    filter(level == "Bachelor's degree",
           state == stat,
           group %in% c("white", "minority"))

  print(ggplot(sums_by_state_plot,
               aes(x = year,
                   y = count,
                   colour = group)) +
          geom_line() +
          # Add a vertical line at the year of the ban
          geom_vline(aes(xintercept = ban_start), linetype = "dashed") +
          scale_x_continuous(breaks = seq(1996, 2022, 2)) +
          labs(title = stat))
}
```

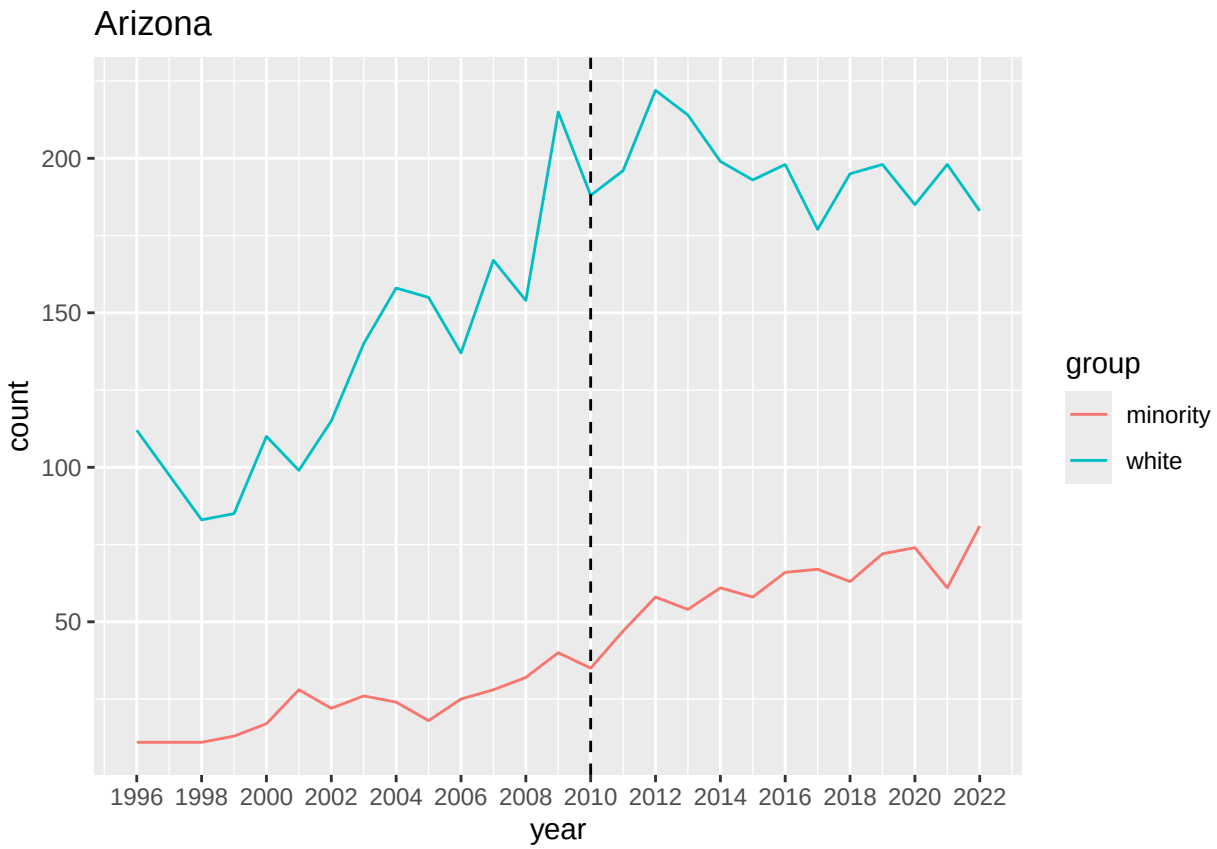


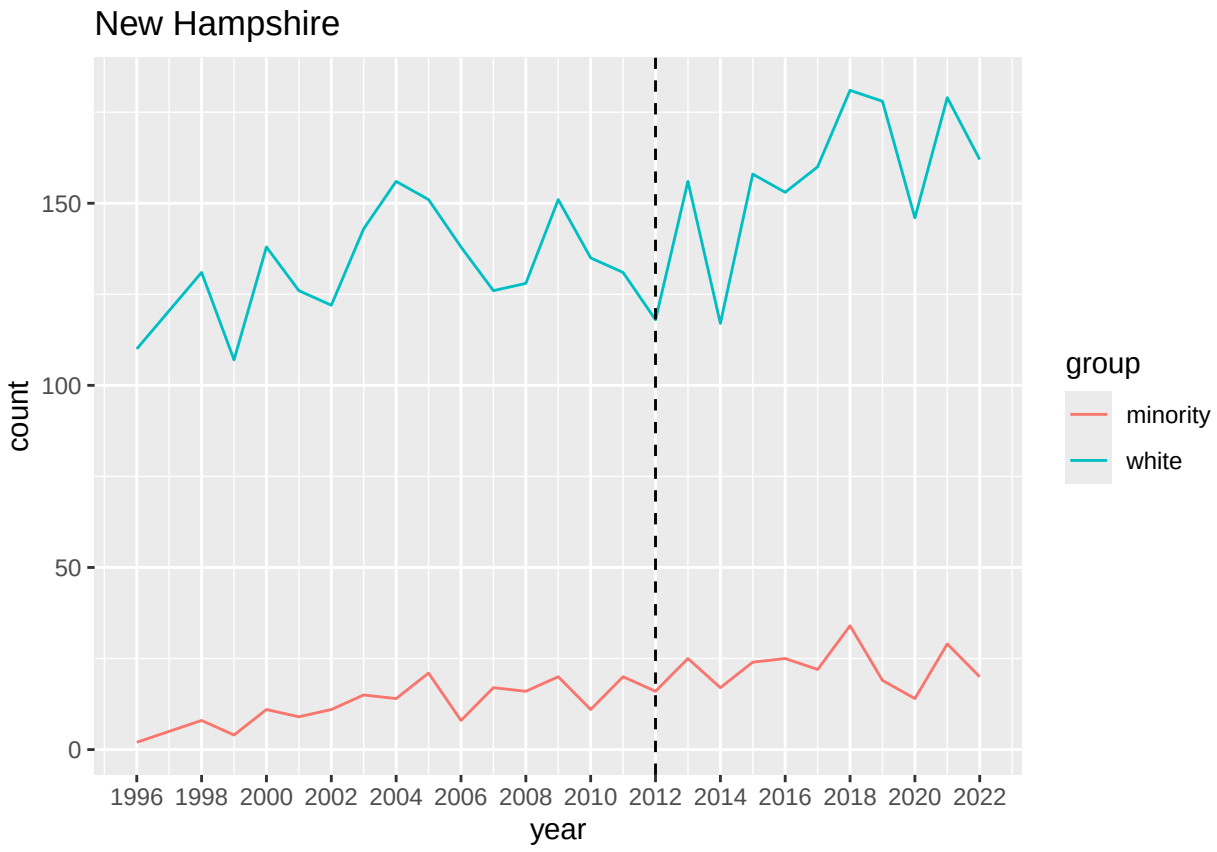

Washington
count
400
300
200
100
0
1996
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
2018
2020
2022
year
group
minority
white

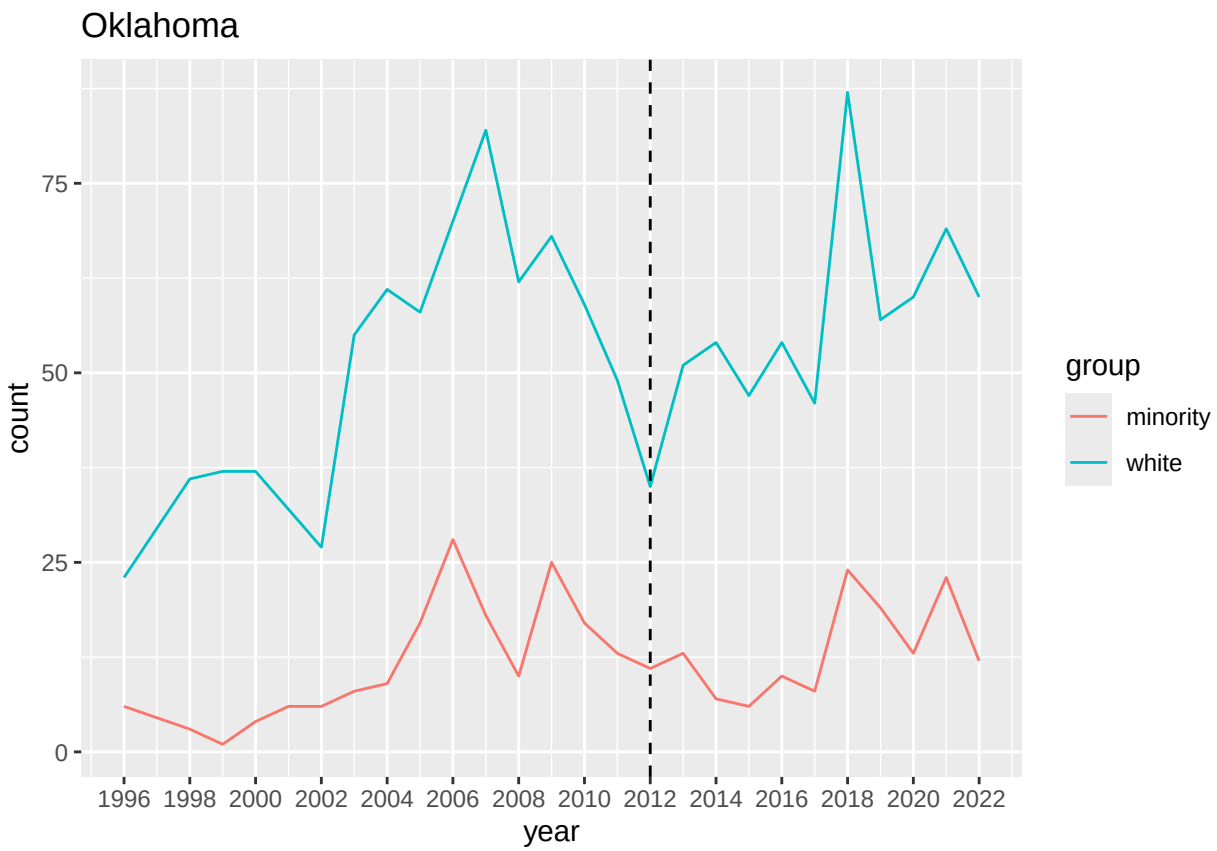

Florida
600
400
200
count
1996
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
2018
2020
2022
year
group
minority
white

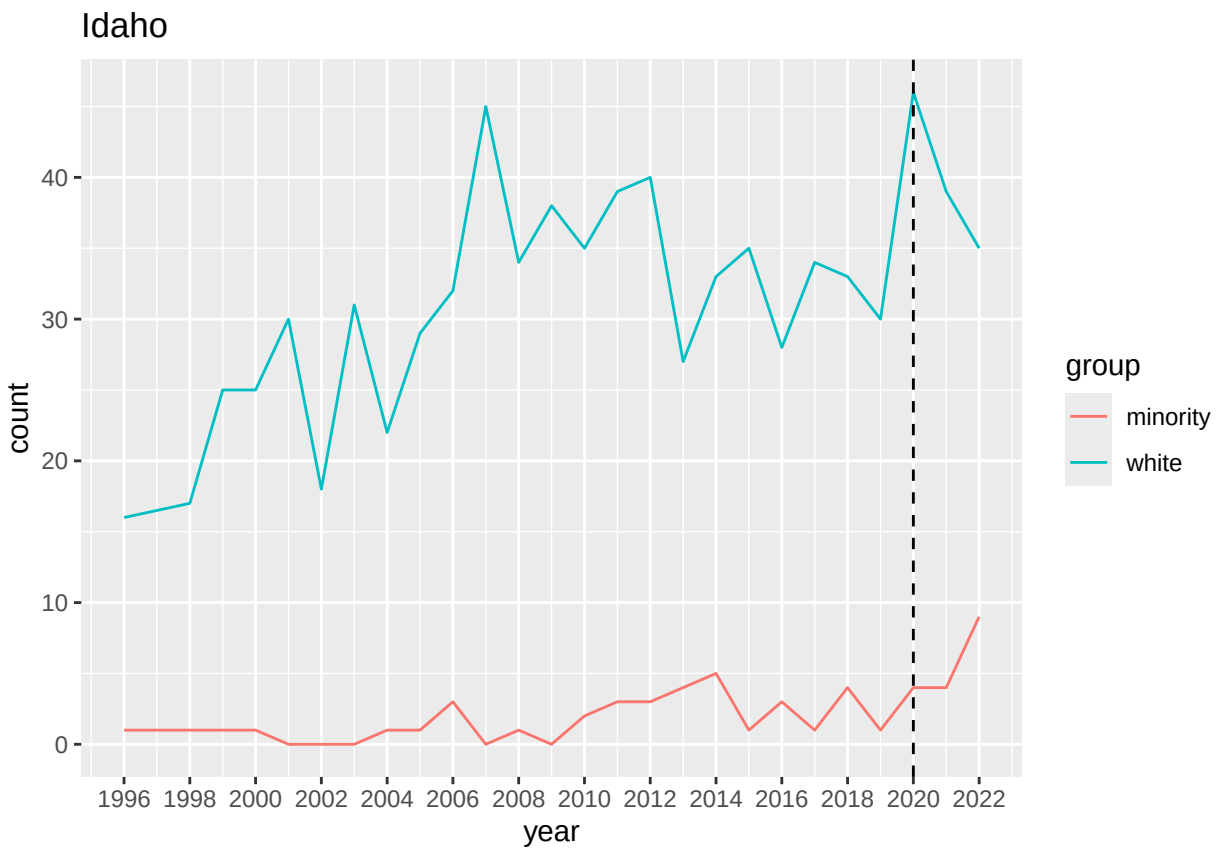

Michigan
count
600
400
200
0
1996 1998 2000 2002 2004 2006 2008 2010 2012 2014 2016 2018 2020 2022
year
group
minority
white


Nebraska
150
100
50
0
count
1996
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
2018
2020
2022
year
group
minority
white

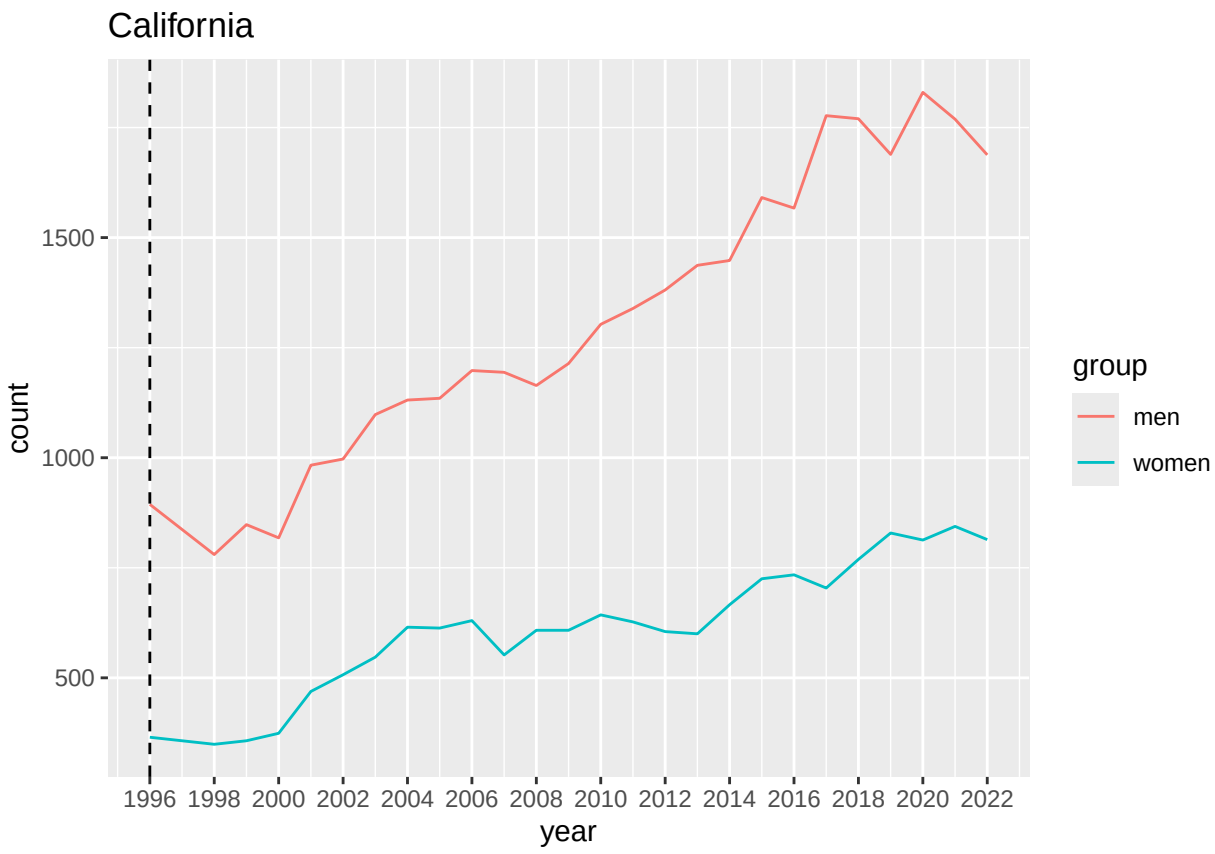

Arizona
count
200
150
100
50
1996
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
2018
2020
2022
year
group
minority
white

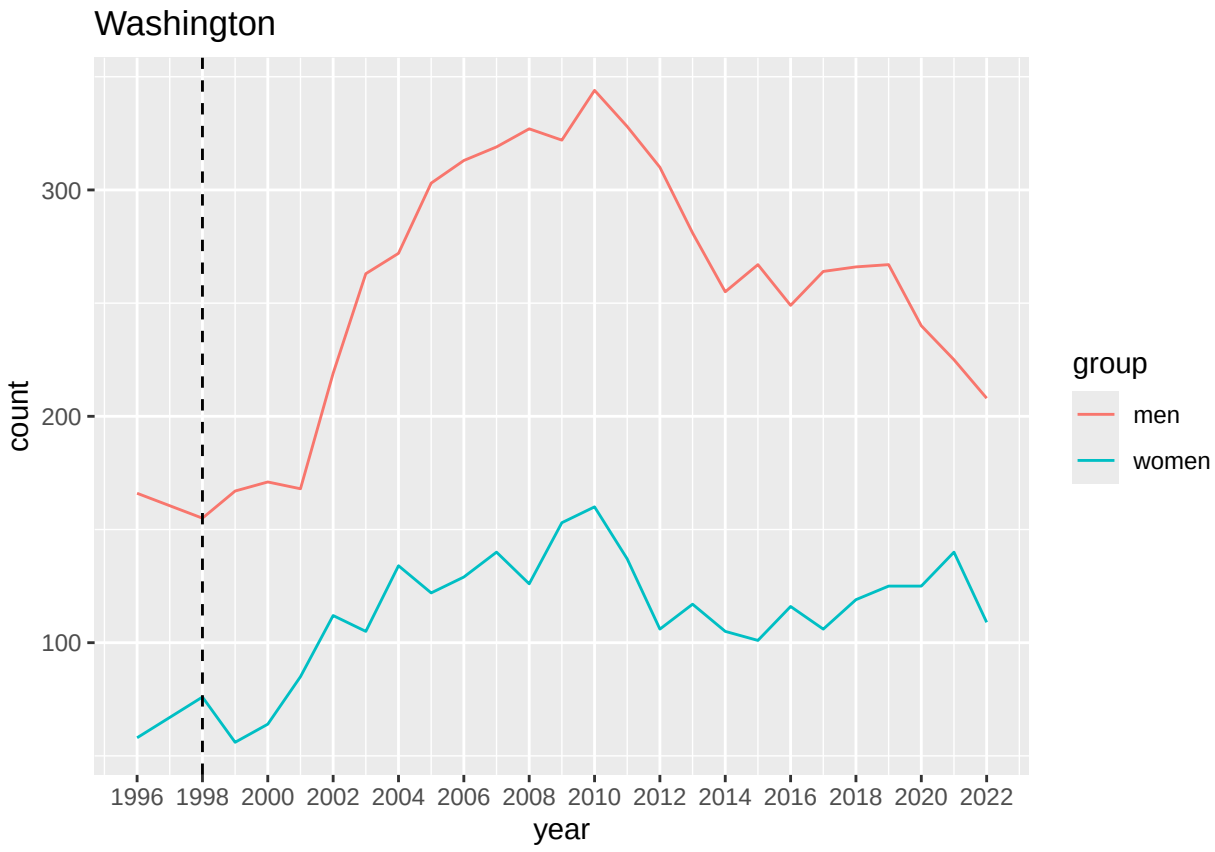

New Hampshire
count
150
100
50
0
1996 1998 2000 2002 2004 2006 2008 2010 2012 2014 2016 2018 2020 2022
year
group
minority
white

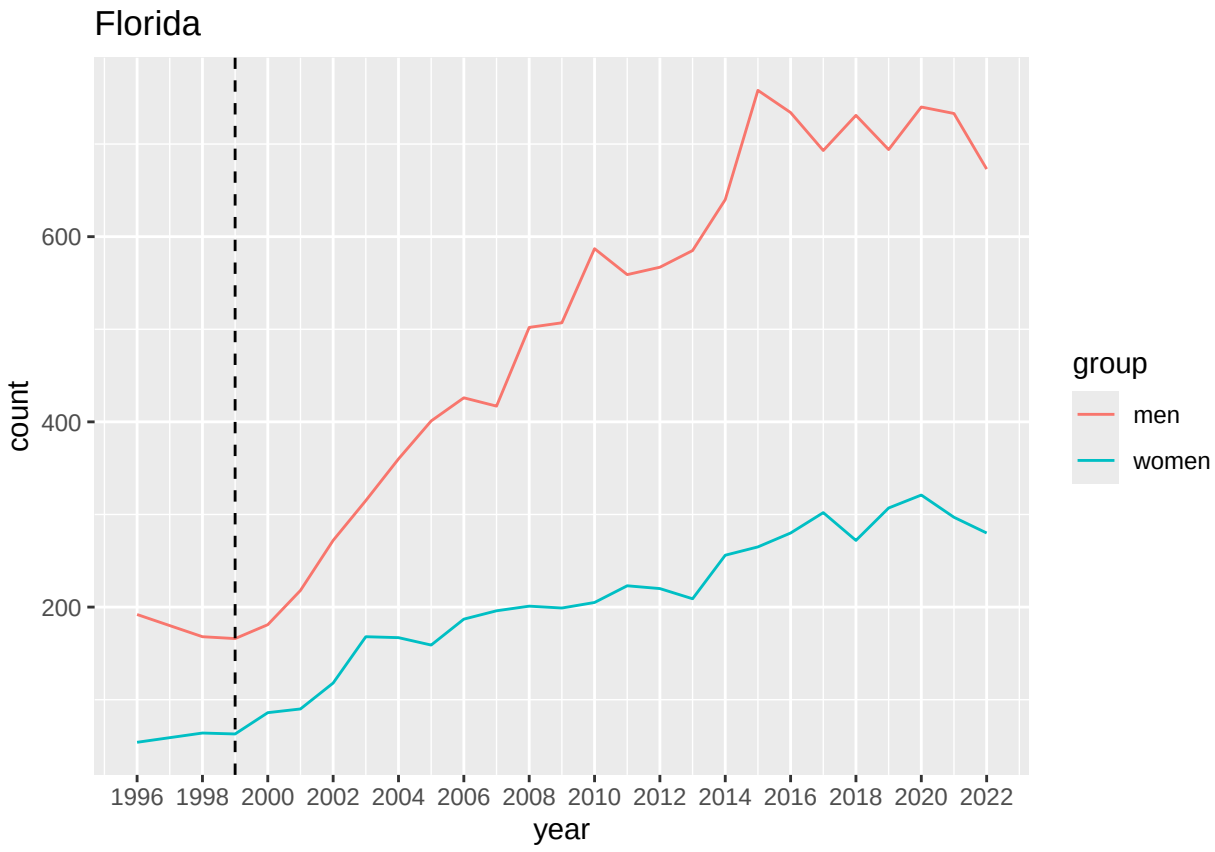

Oklahoma
count
75
50
25
0
1996
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
2018
2020
2022
year
group
minority
white

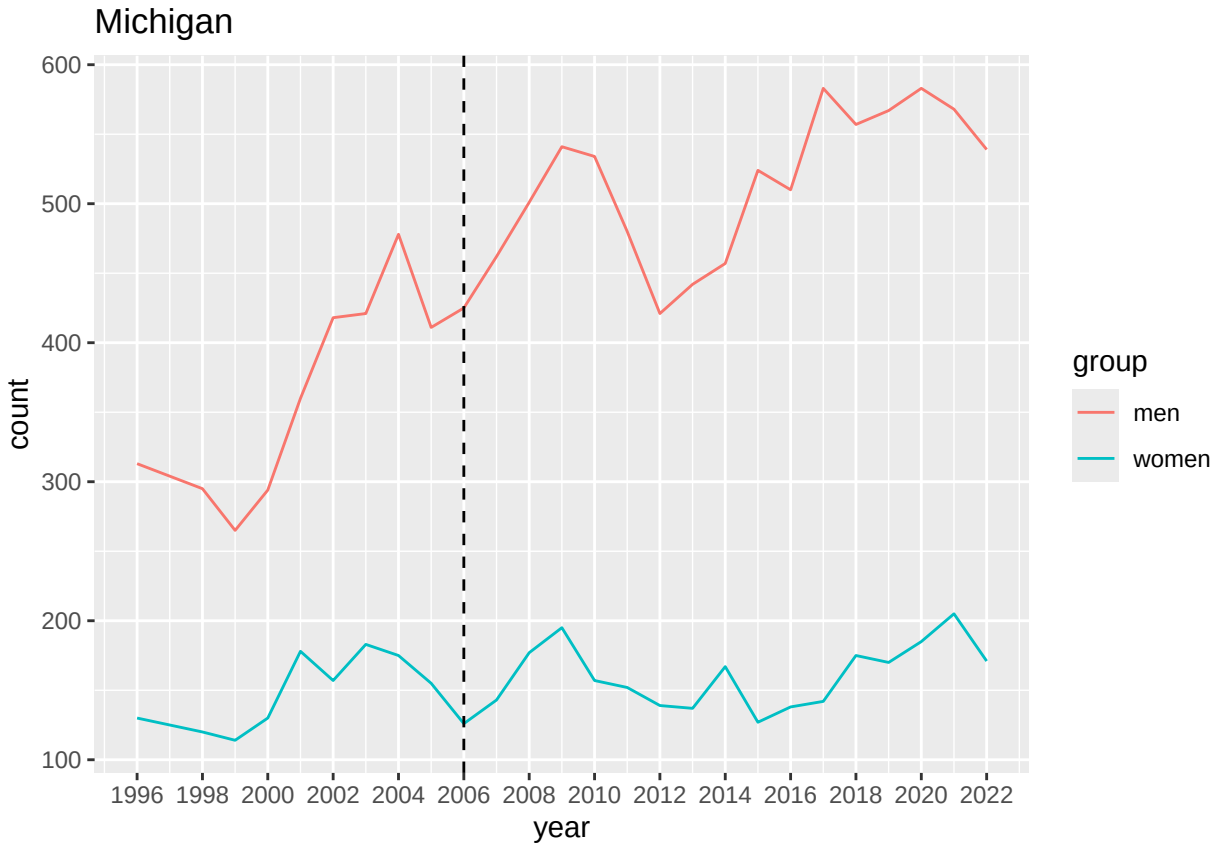


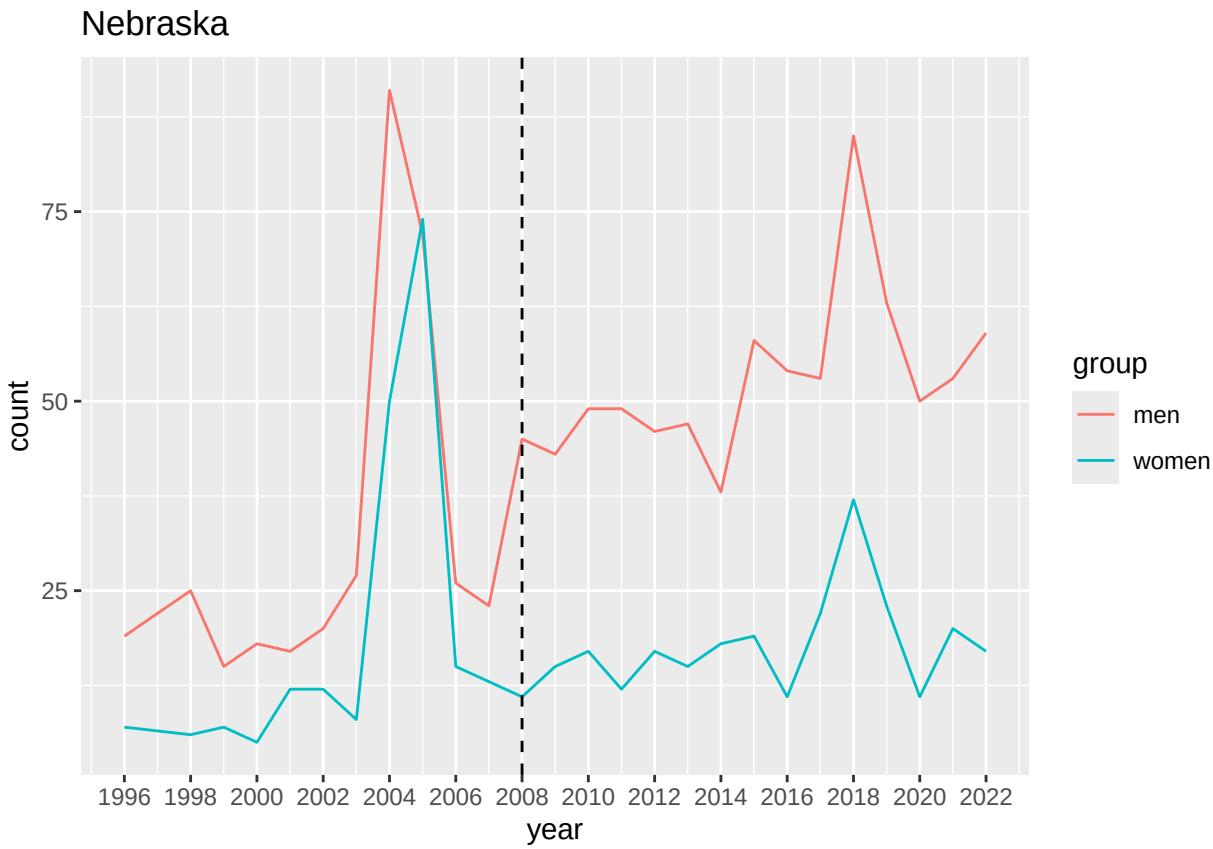

```
# Plots for men/women
for (stat in unique(bans$state)) {
  sums_by_state_plot <- sums_by_state %>%
    filter(level == "Bachelor's degree",
           state == stat,
           group %in% c("men", "women"))

  print(ggplot(sums_by_state_plot,
               aes(x = year,
                   y = count,
                   colour = group)) +
          geom_line() +
          # Add a vertical line at the year of the ban
          geom_vline(aes(xintercept = ban_start), linetype = "dashed") +
          scale_x_continuous(breaks = seq(1996, 2022, 2)) +
          labs(title = stat))
}
```

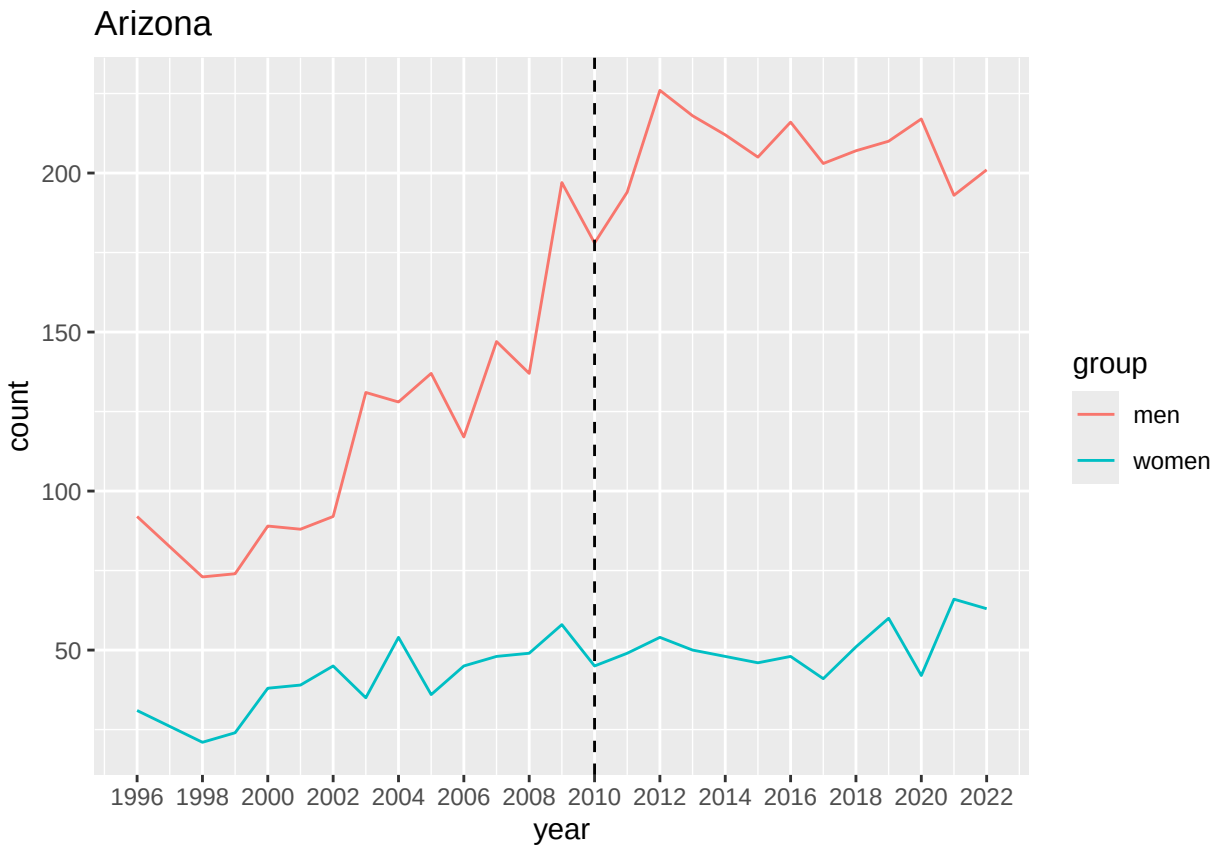

California
count
1500
1000
500
1996 1998 2000 2002 2004 2006 2008 2010 2012 2014 2016 2018 2020 2022
year
group
men
women

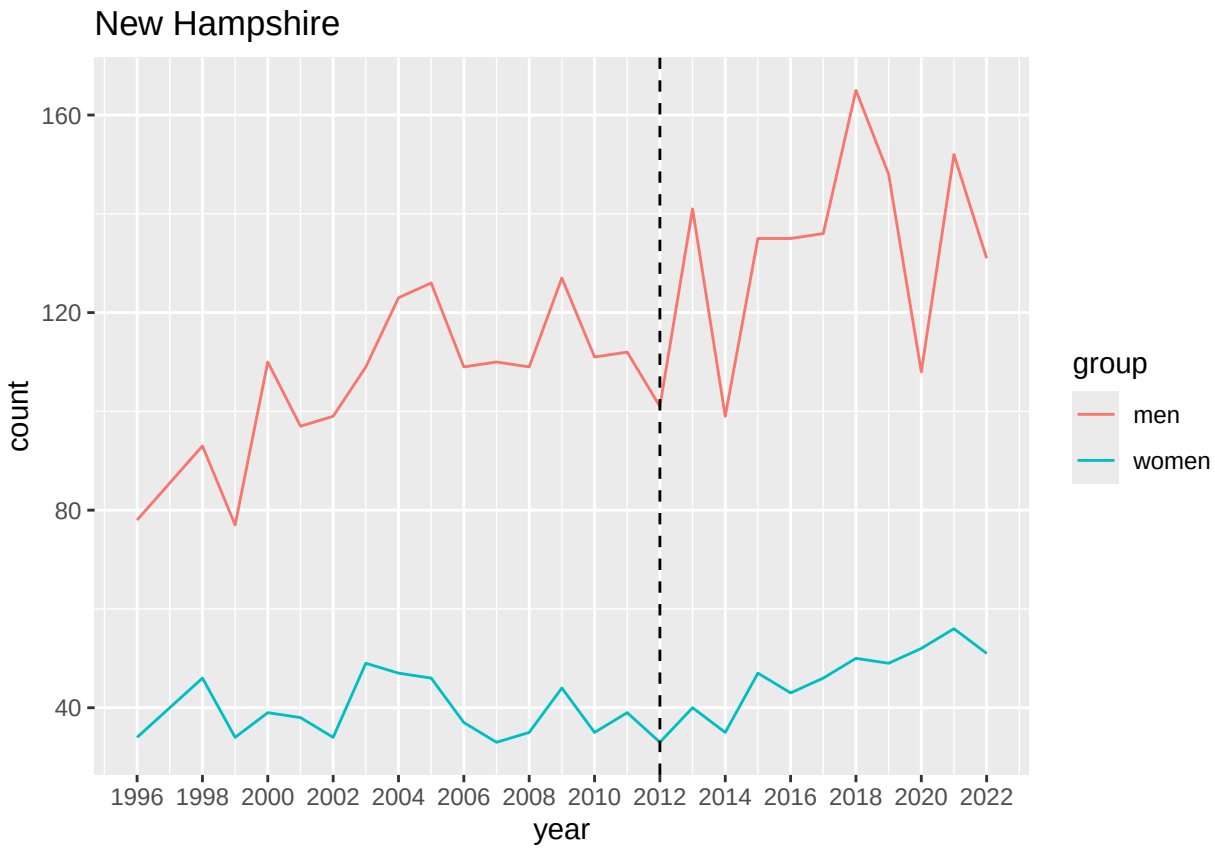

Washington
count
300
200
100
1996
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
2018
2020
2022
year
group
men
women

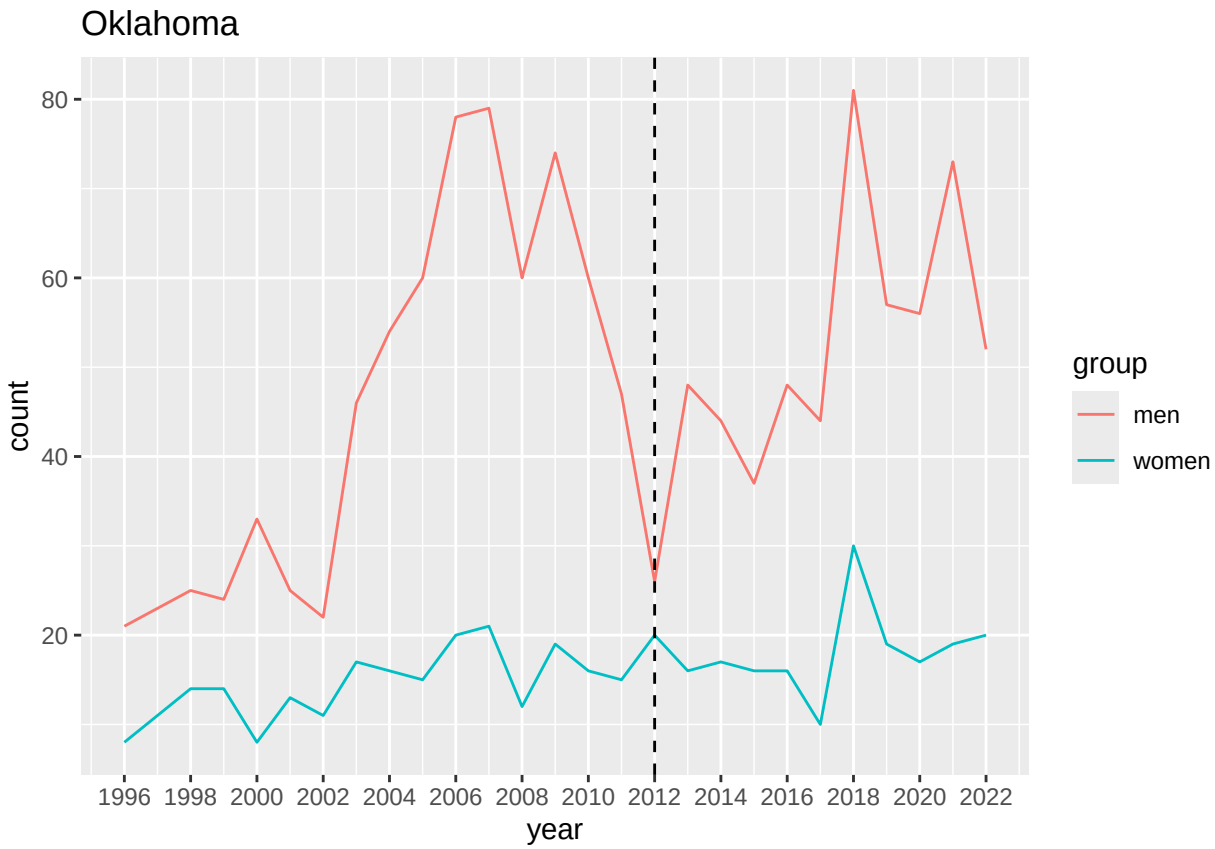

Florida
count
600
400
200
1996
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
2018
2020
2022
year
group
men
women

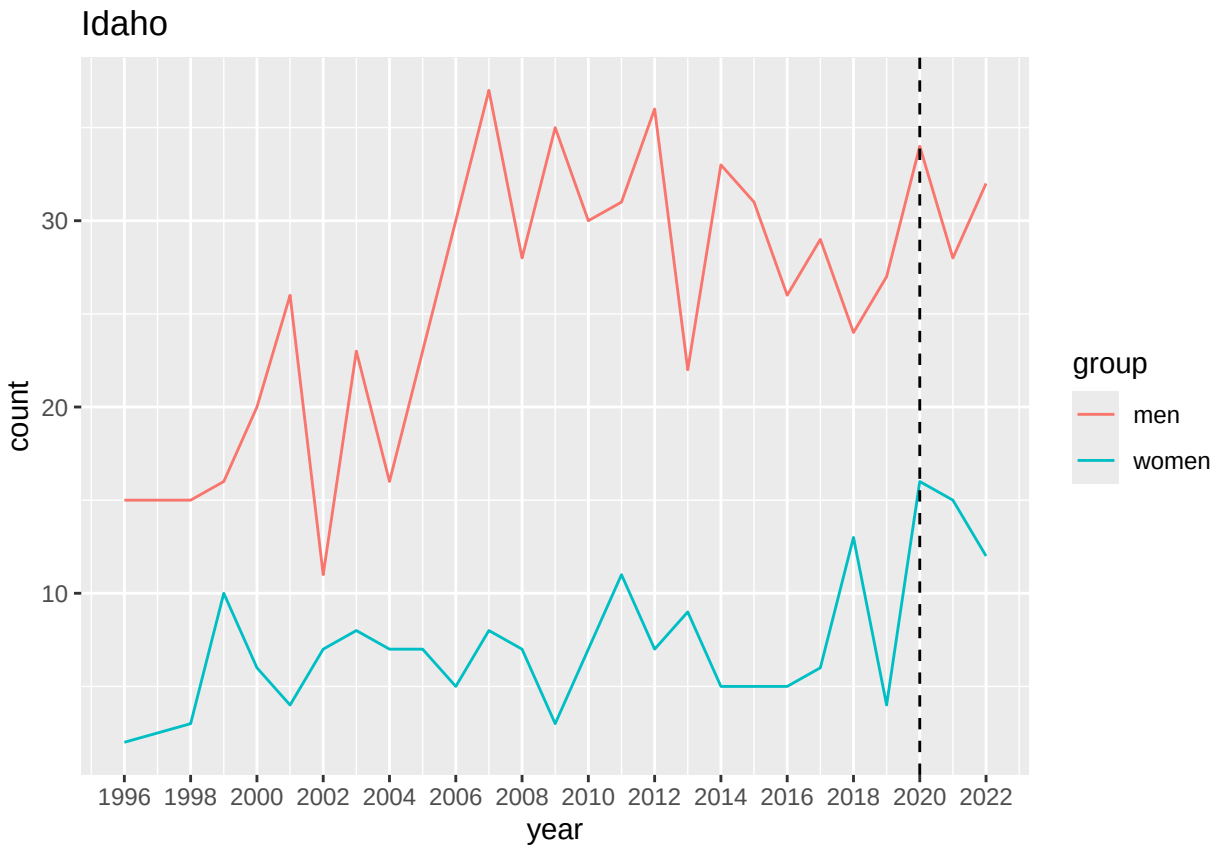

Michigan
600
500
400
300
200
100
count
1996
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
2018
2020
2022
year
group
men
women


Nebraska
count
75
50
25
1996
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
2018
2020
2022
year
group
men
women


Arizona
count
200
150
100
50
1996
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
2018
2020
2022
year
group
men
women


New Hampshire
160
120
80
40
count
1996
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
2018
2020
2022
year
group
men
women


Oklahoma
count
80
60
40
20
1996
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
2018
2020
2022
year
group
men
women


Idaho
count
30
20
10
1996
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
2018
2020
2022
year
group
men
women