Political Analysis of Social Media Data

Visualization

Instructor: Gregory Eady
Office: 18.2.10
Office hours: Fridays 13-15

Why look at data?

Scatterplot

Mean and correlation are identical, but:

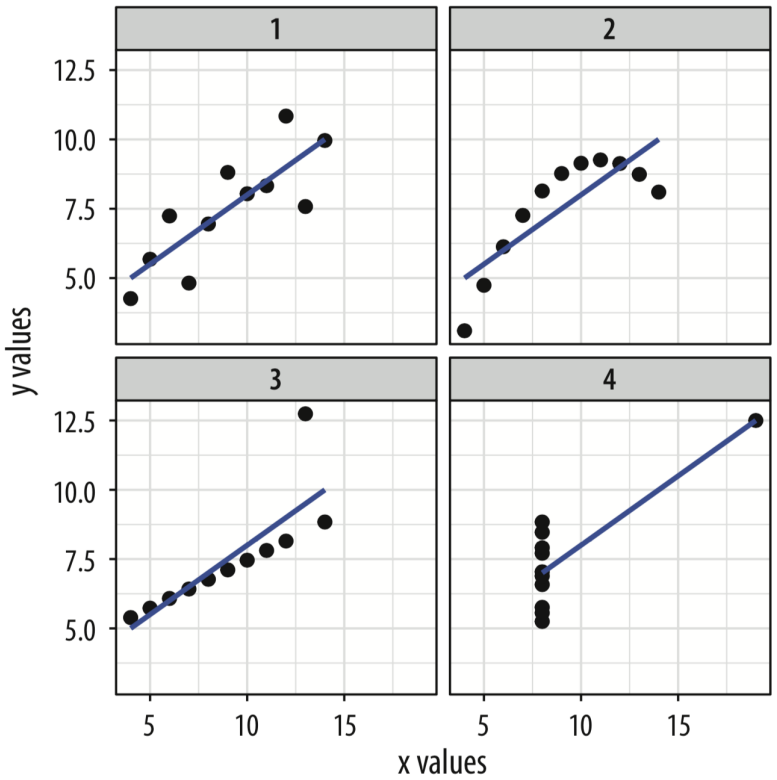
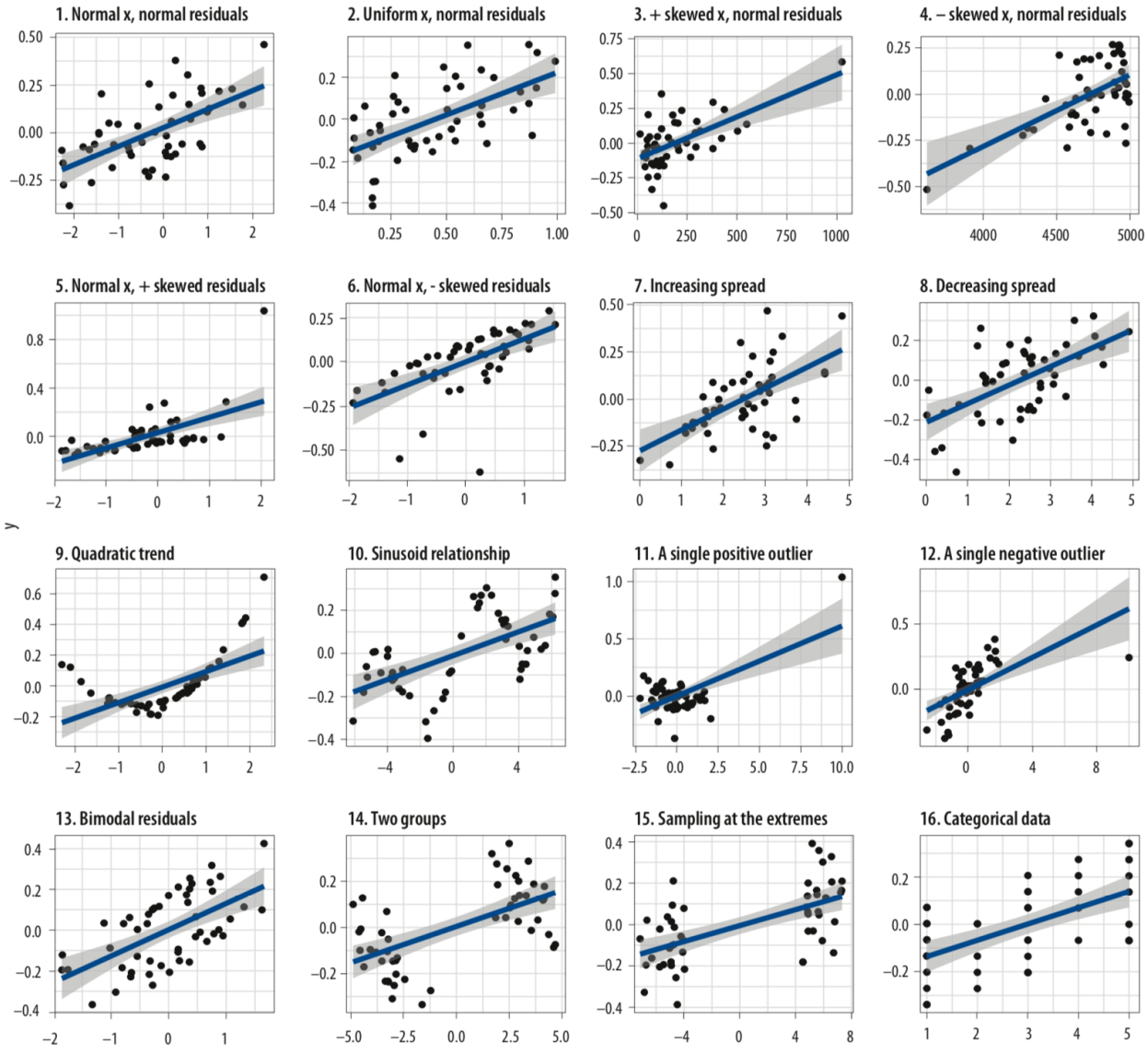
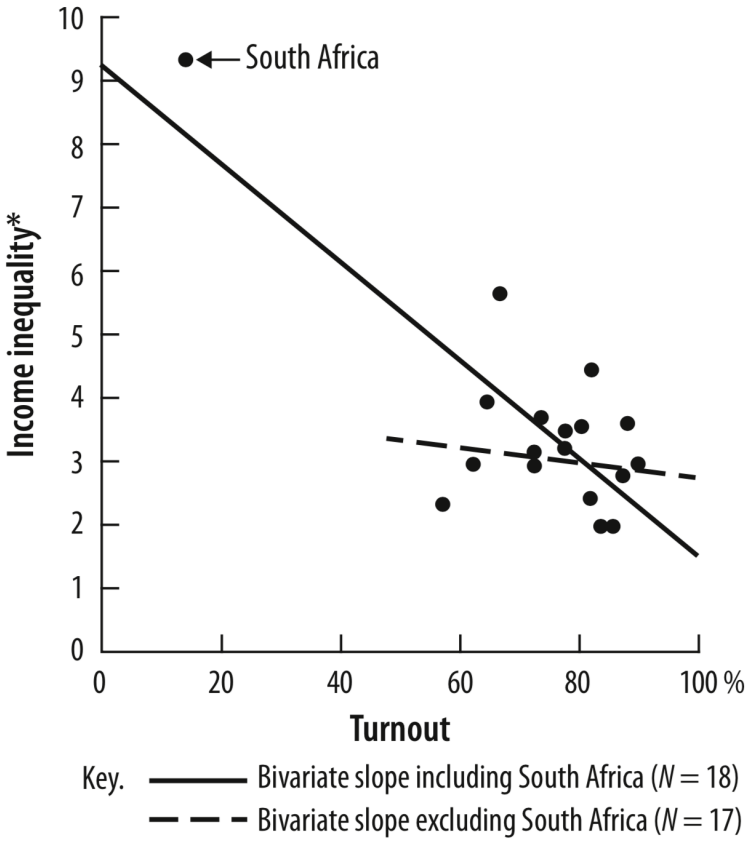


Figure 1.1: Plots of Anscombe's quartet.


1. Normal x, normal residuals
2. Uniform x, normal residuals
3. + skewed x, normal residuals
4. – skewed x, normal residuals
5. Normal x, + skewed residuals
6. Normal x, - skewed residuals
7. Increasing spread
8. Decreasing spread
9. Quadratic trend
10. Sinusoid relationship
11. A single positive outlier
12. A single negative outlier
13. Bimodal residuals
14. Two groups
15. Sampling at the extremes
16. Categorical data
y

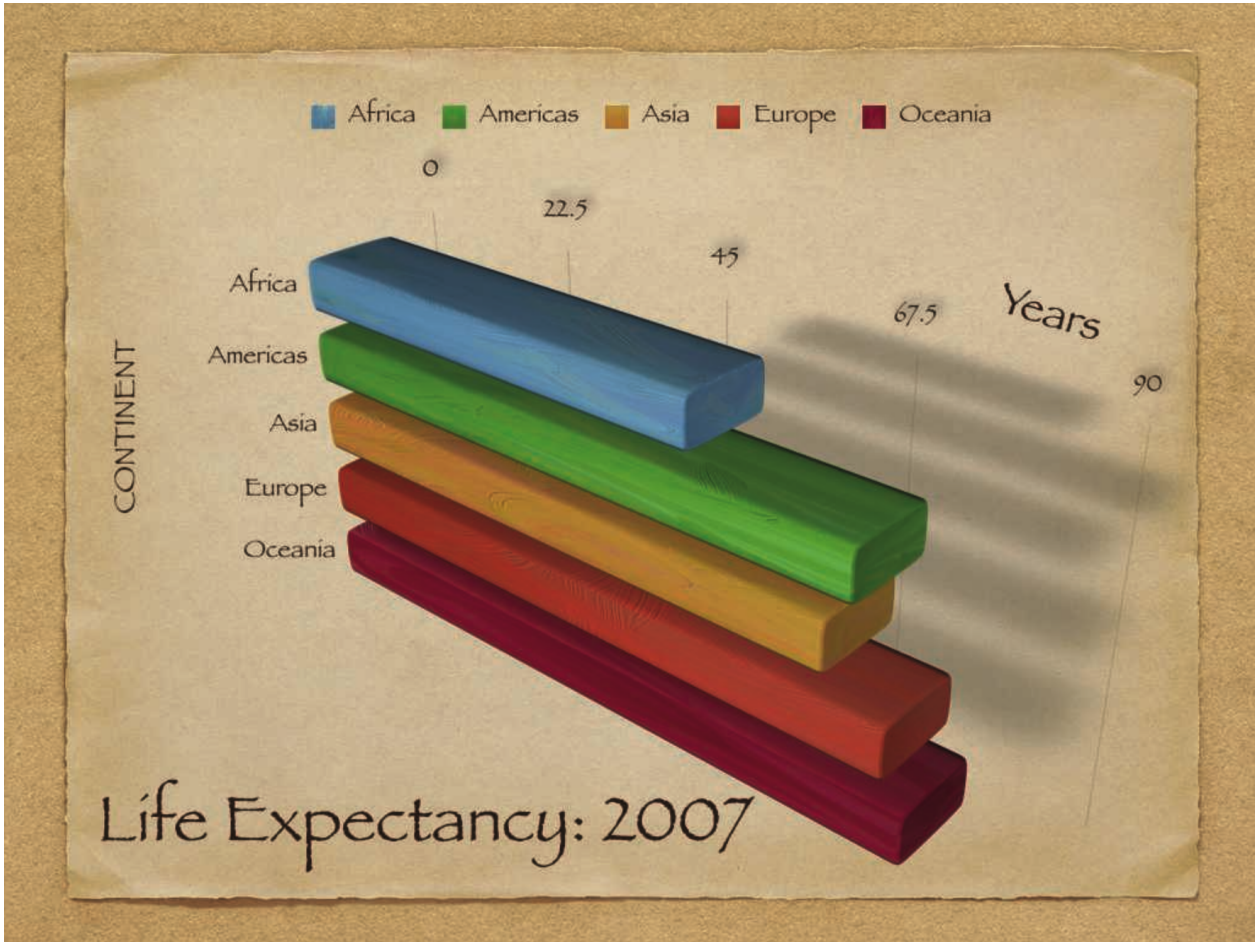
A real case


South Africa
Income inequality*
10
9
8
7
6
5
4
3
2
1
0
0
20
40
60
80
100 %
Turnout
Key.
Bivariate slope including South Africa (N = 18)
Bivariate slope excluding South Africa (N = 17)

What makes a "bad" visualization?

1. Aesthetic concerns
2. Substantive concerns

Bad aesthetics



Bad aesthetics

- Bars are hard to read and compare
- Duplicates labels
- Pointless use of 3D effects
- Useless drop shadows

Tufte & the conventional wisdom

- Maximize the "data-to-ink" ratio
- i.e. Simplify as much as possible

Unfortunately (?), infographic-style visualizations have some benefits

- Are easier to recall (even if they are harder to interpret)
- Are more memorable

Example

Figure 1.6: "Monstrous Costs" by Nigel Holmes (1982). Also a classic of its kind.

Furthermore, simplicity can go too far

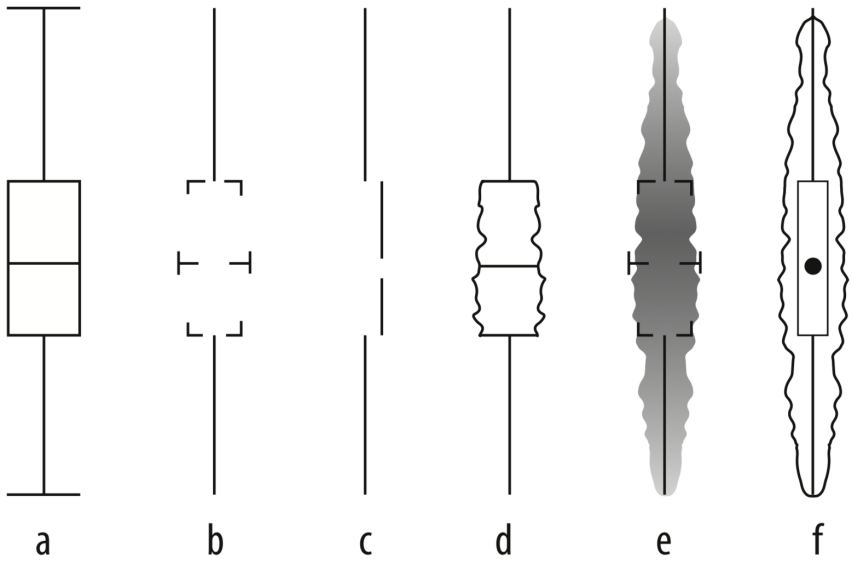
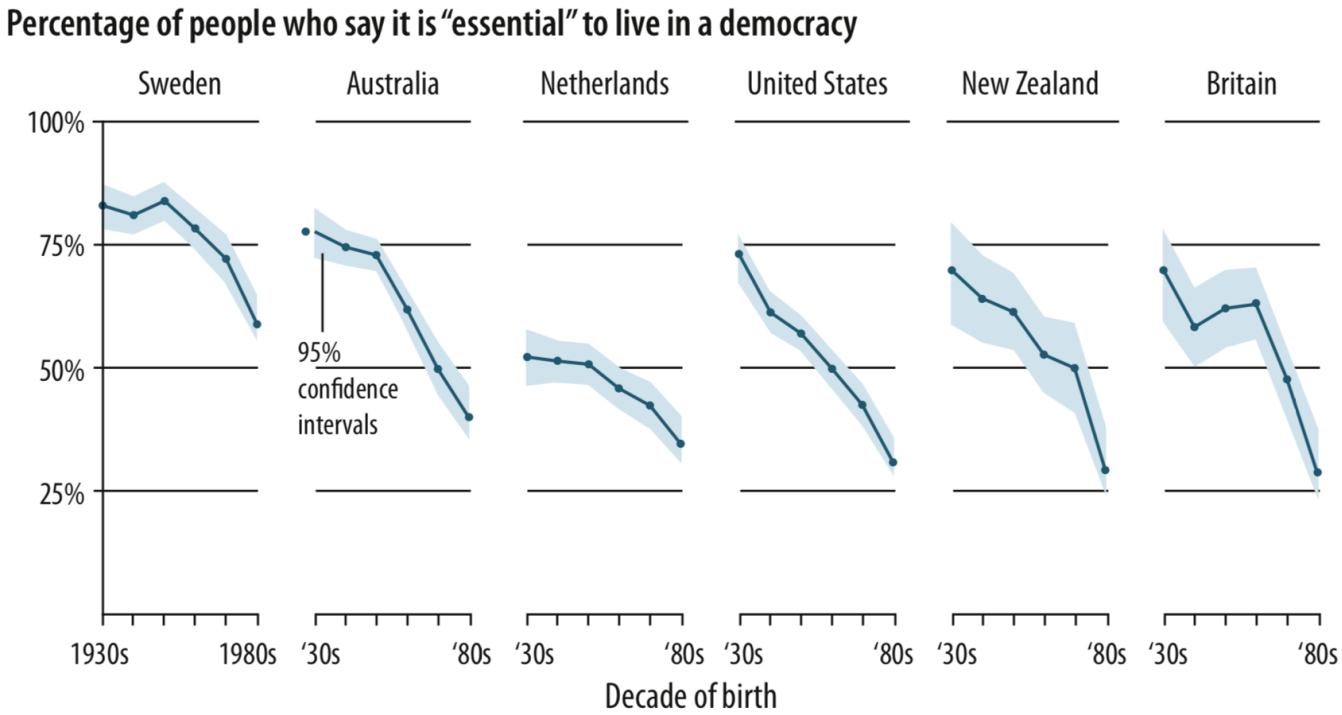


Figure 1.7: Six kinds of summary boxplots. Type (c) is from Tufte.

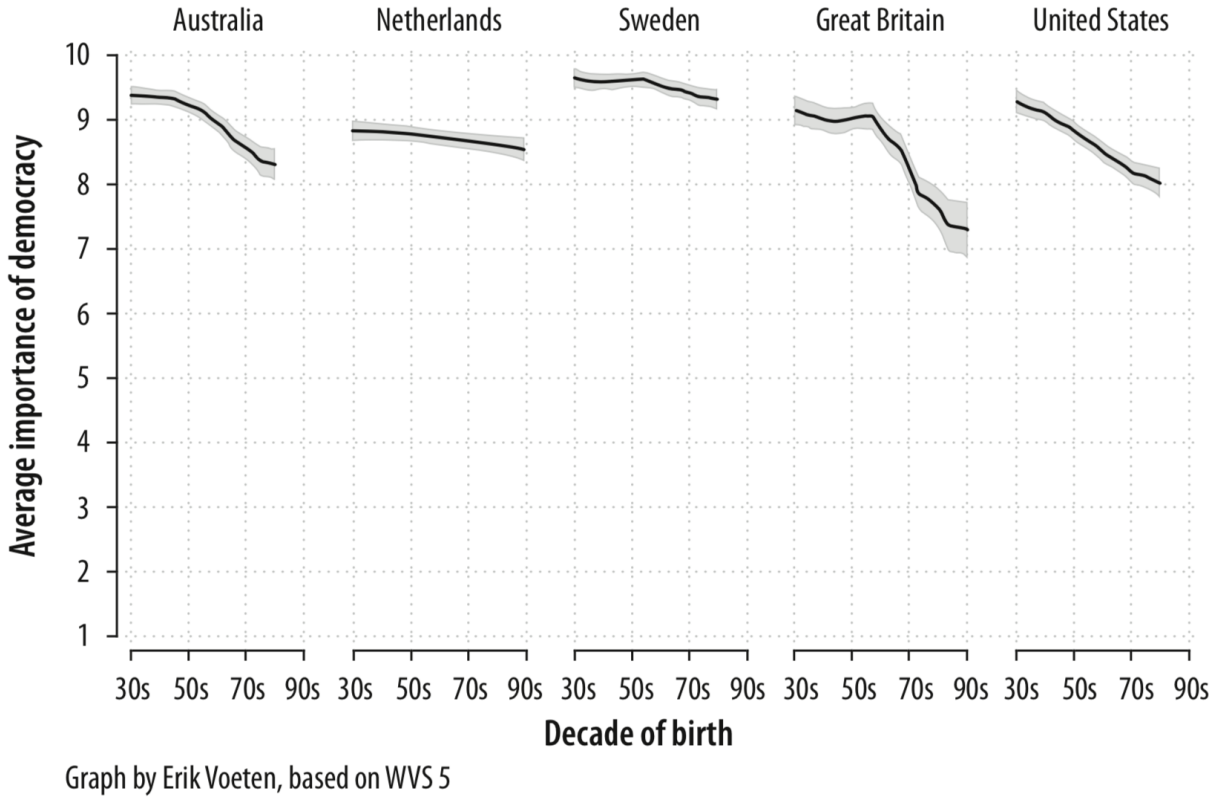
Tufte's own preference (boxplot c) is the least understood

Bad data



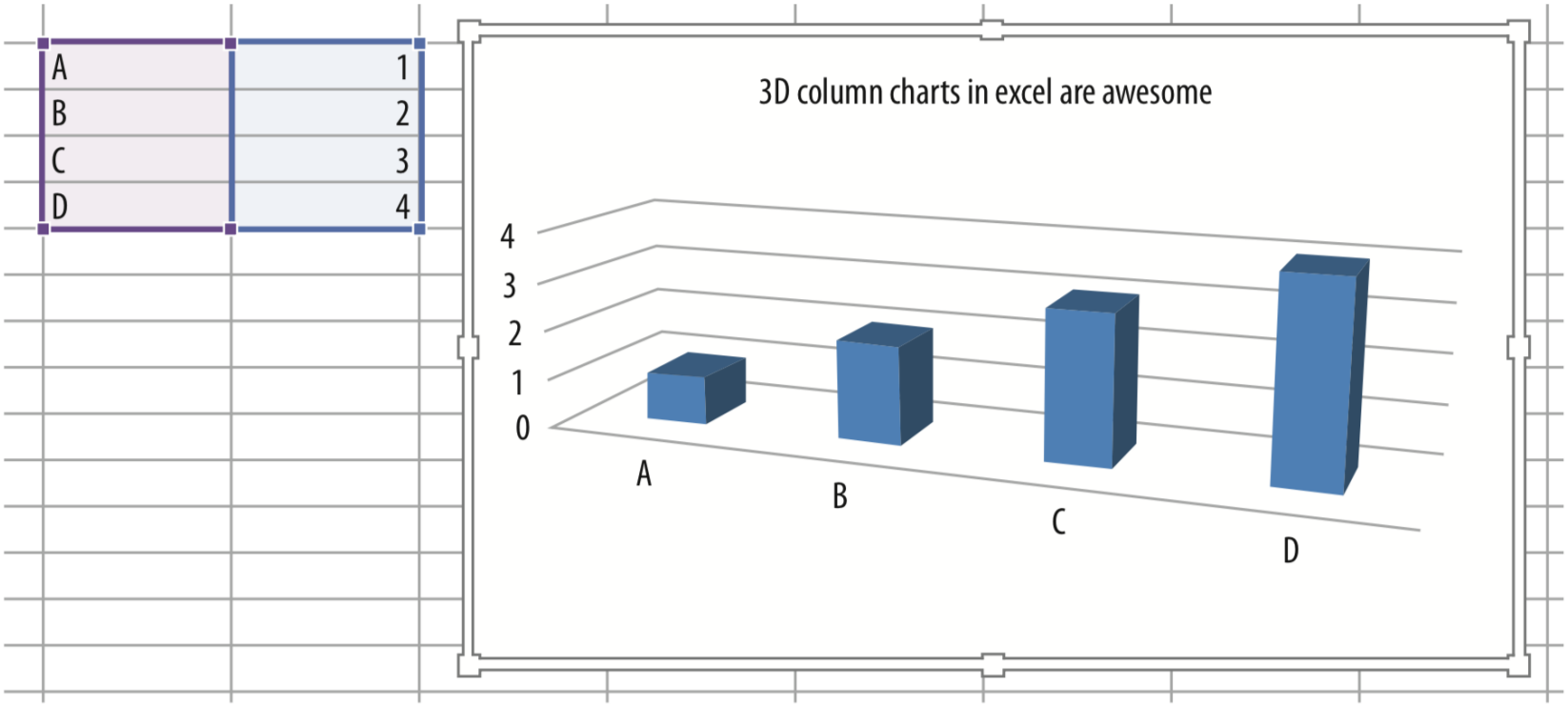
The y-axis suggests percentage of people agreeing...

But on original scale...



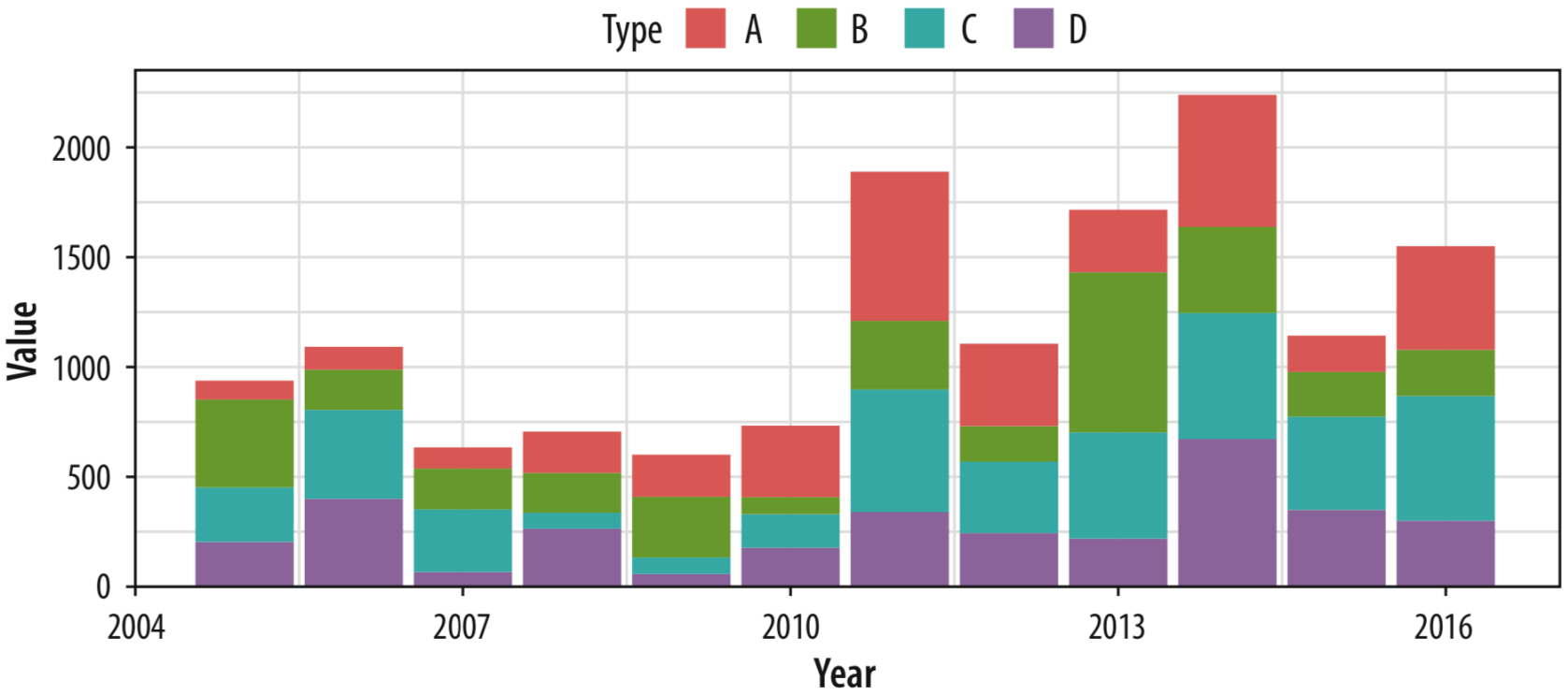
Problem is the data themselves

Bad perceptions



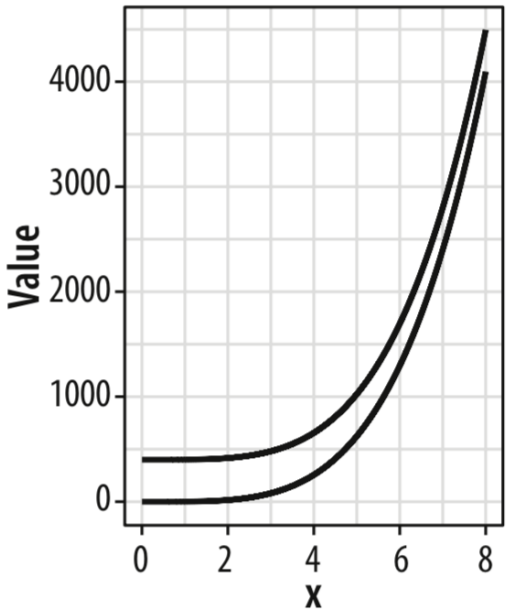
Values even appear “too low’ in the graph

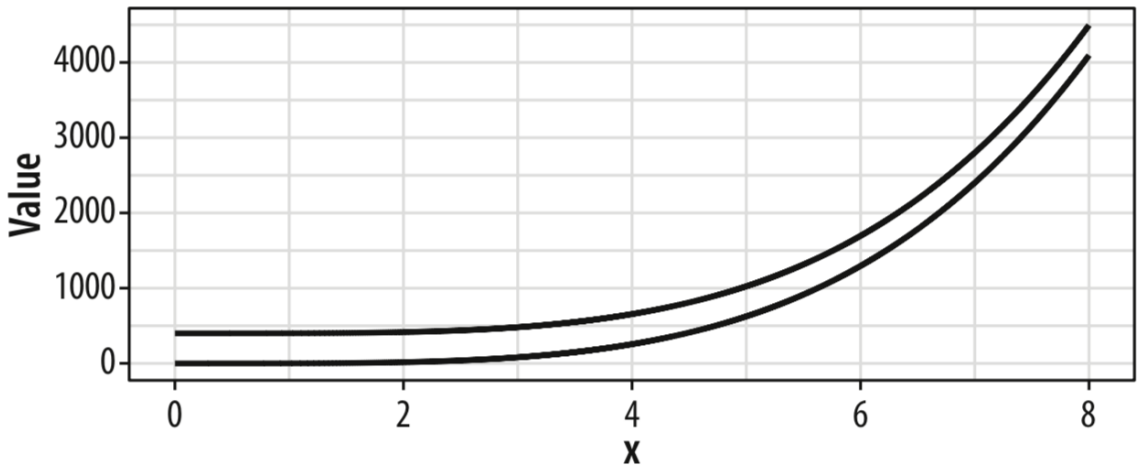
Bad perceptions



Flat, but still difficult to interpret

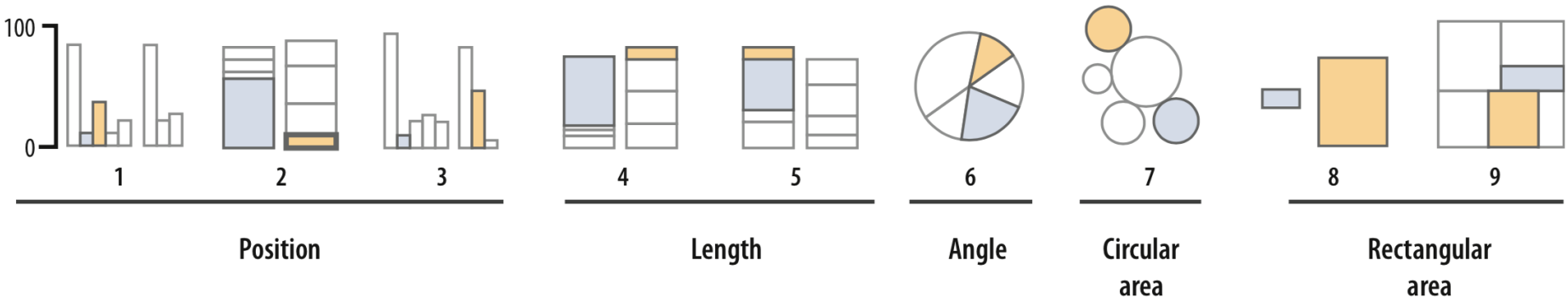
Bad perceptions



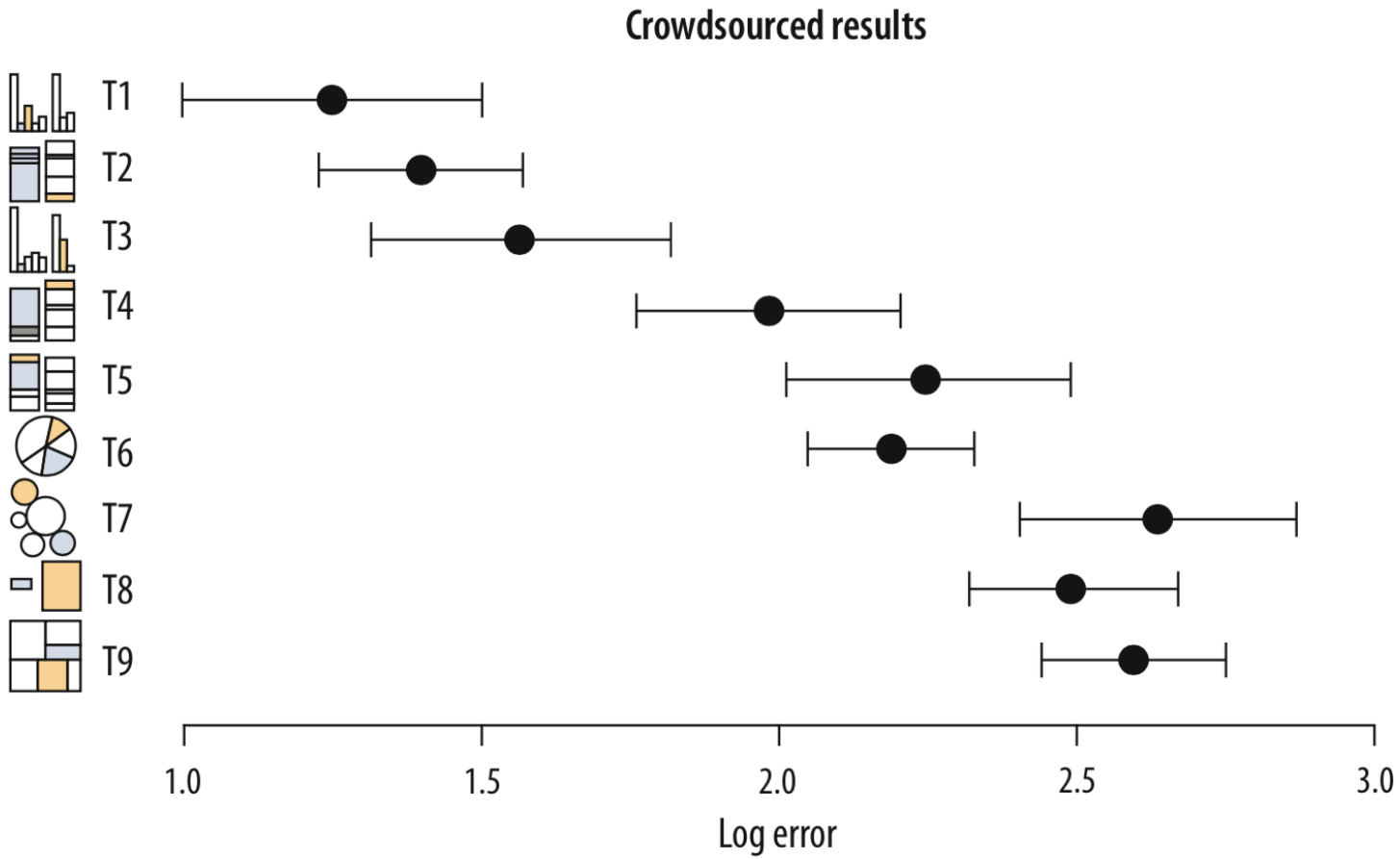


Same data, different aspect ratio

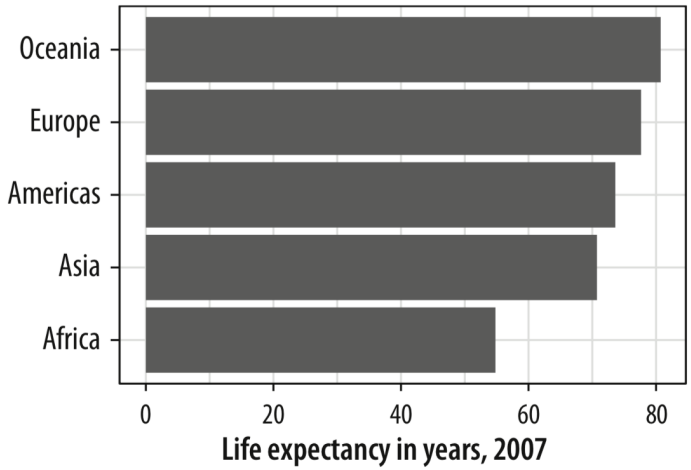
What graphs are most interpretable?

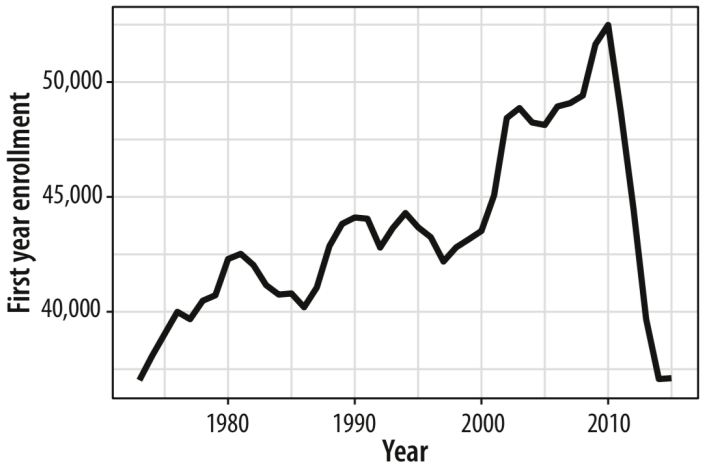


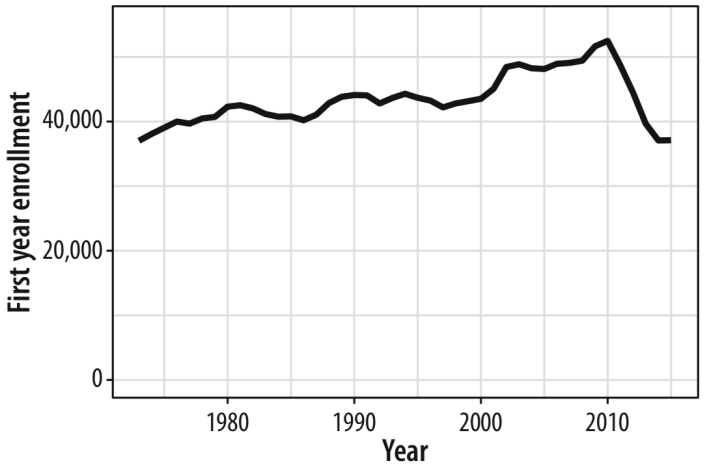
What graphs are most interpretable?



Honesty and good judgment

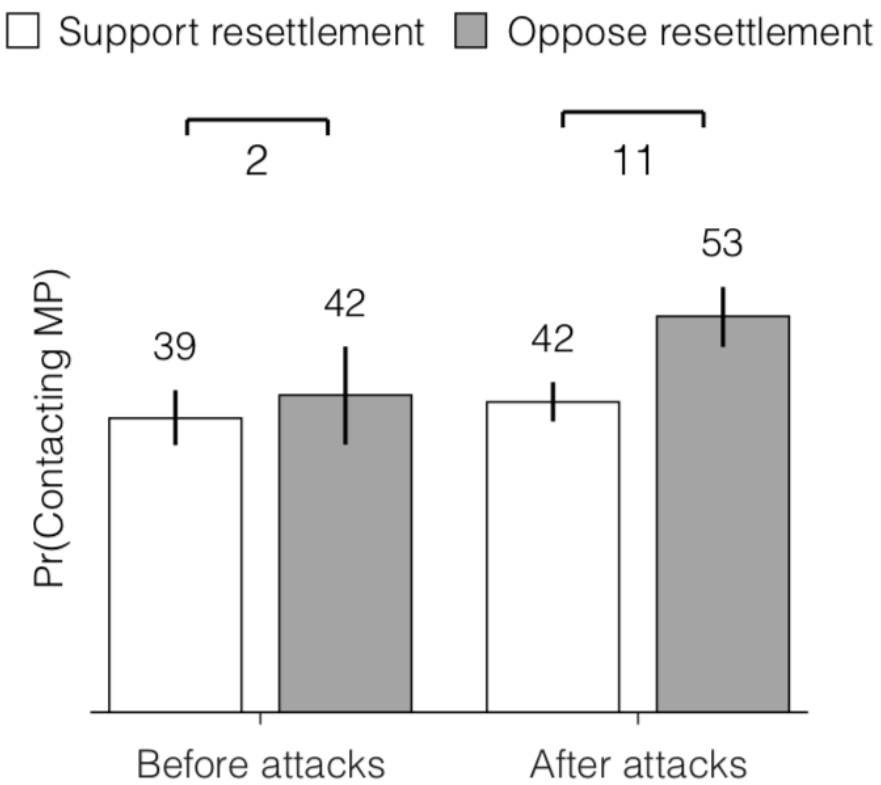




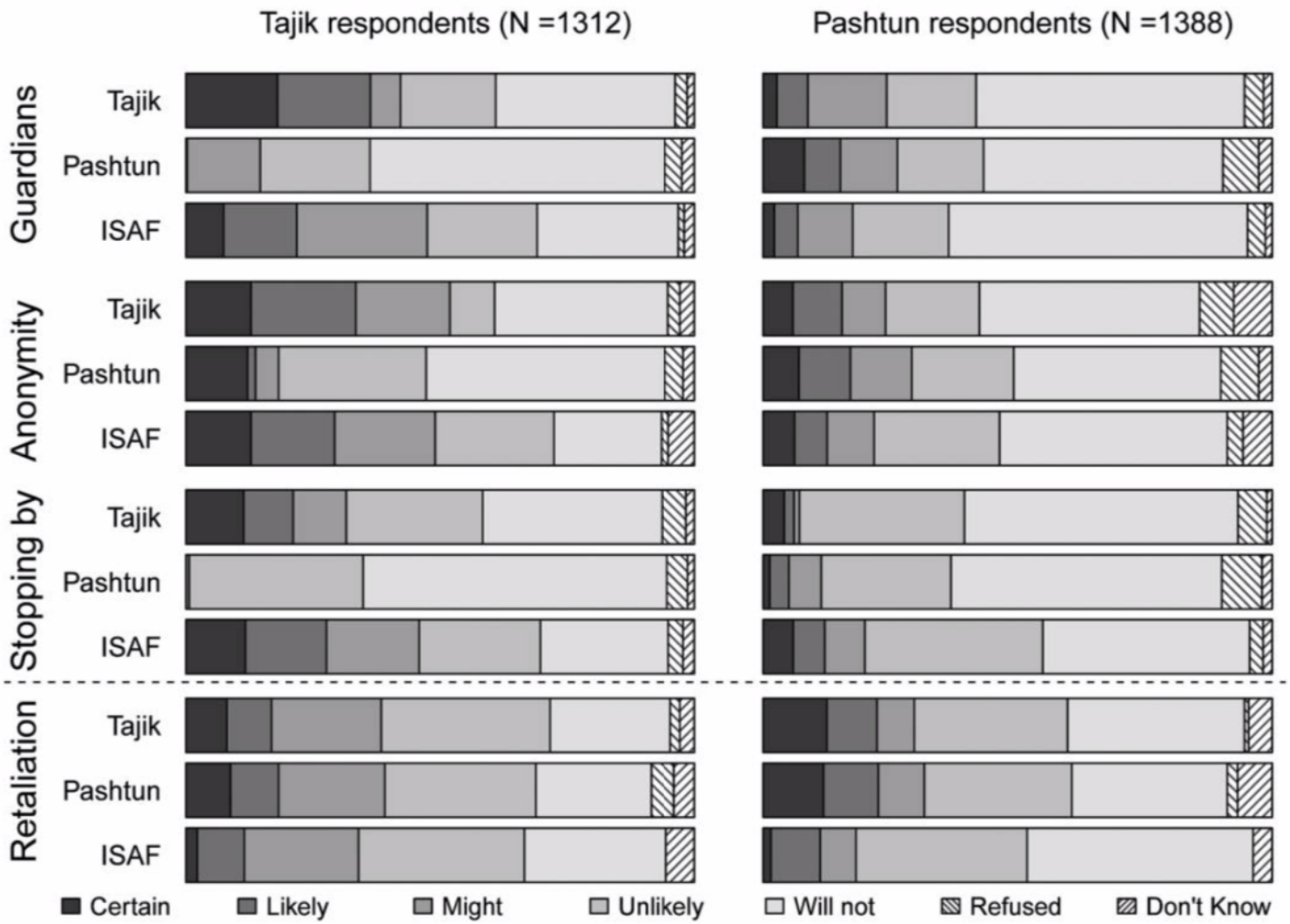


ggplot2

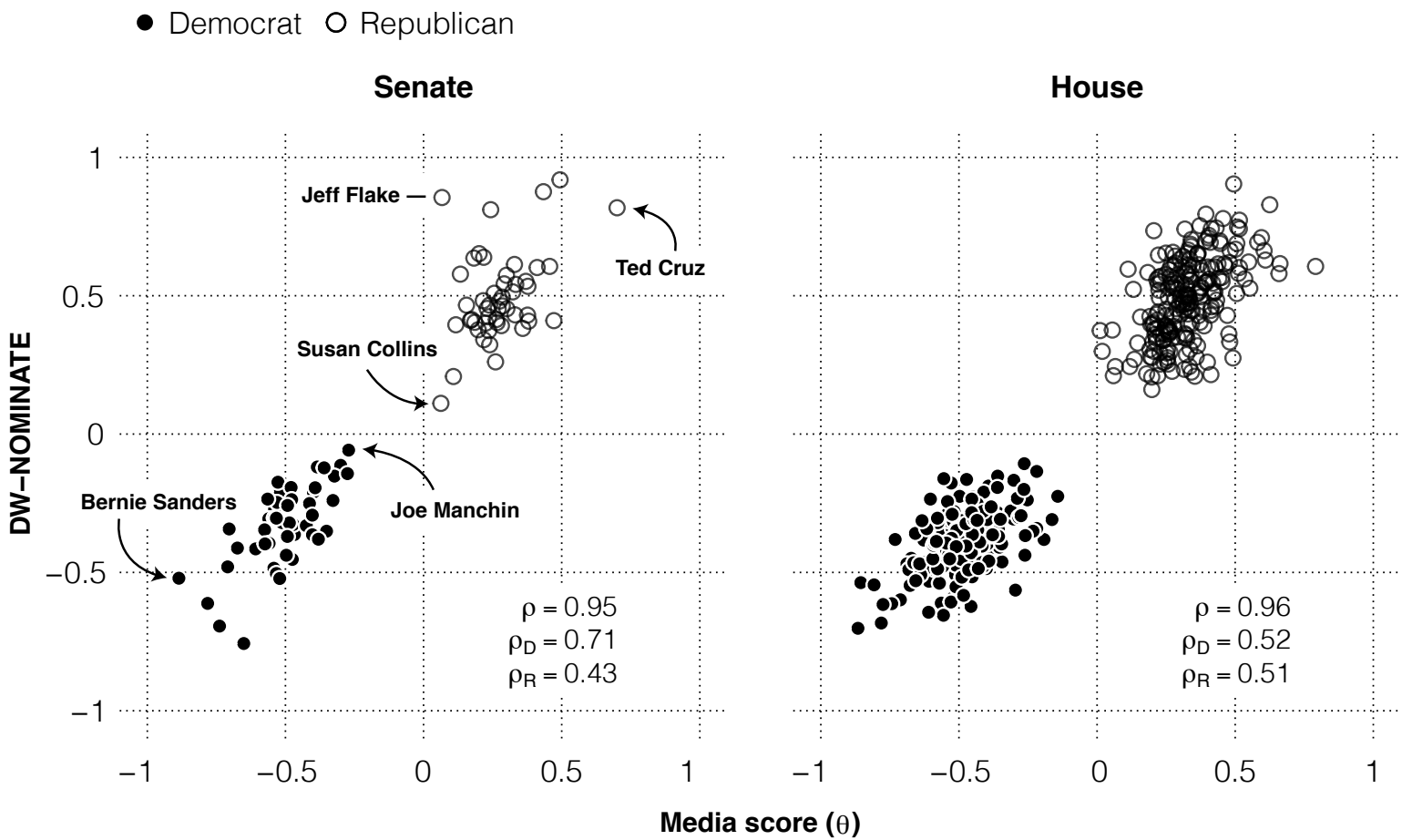
Bar graph: `geom_bar()`



Bar graph (stacked): `geom_bar()`

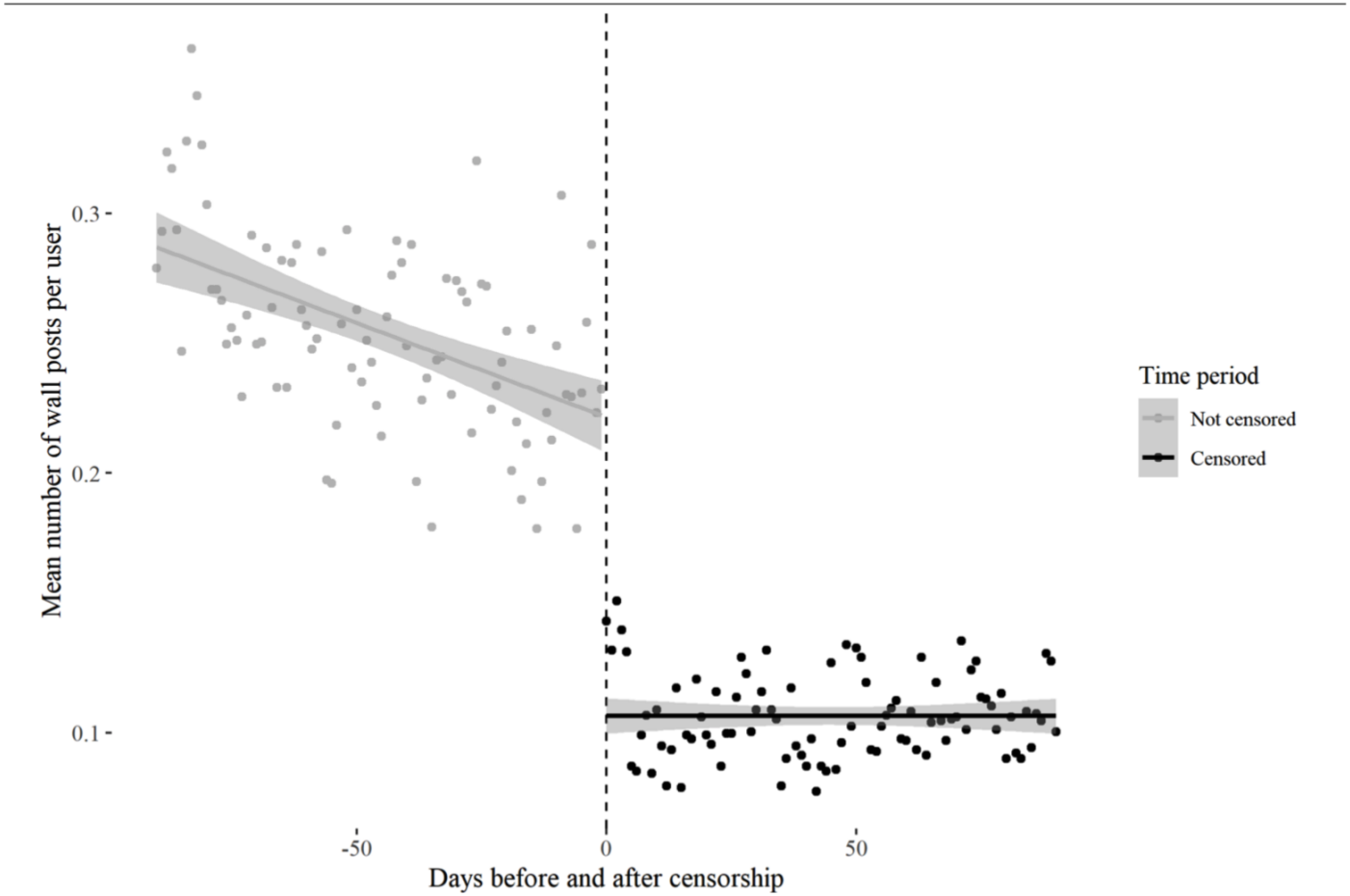


Scatterplot: `geom_point()`

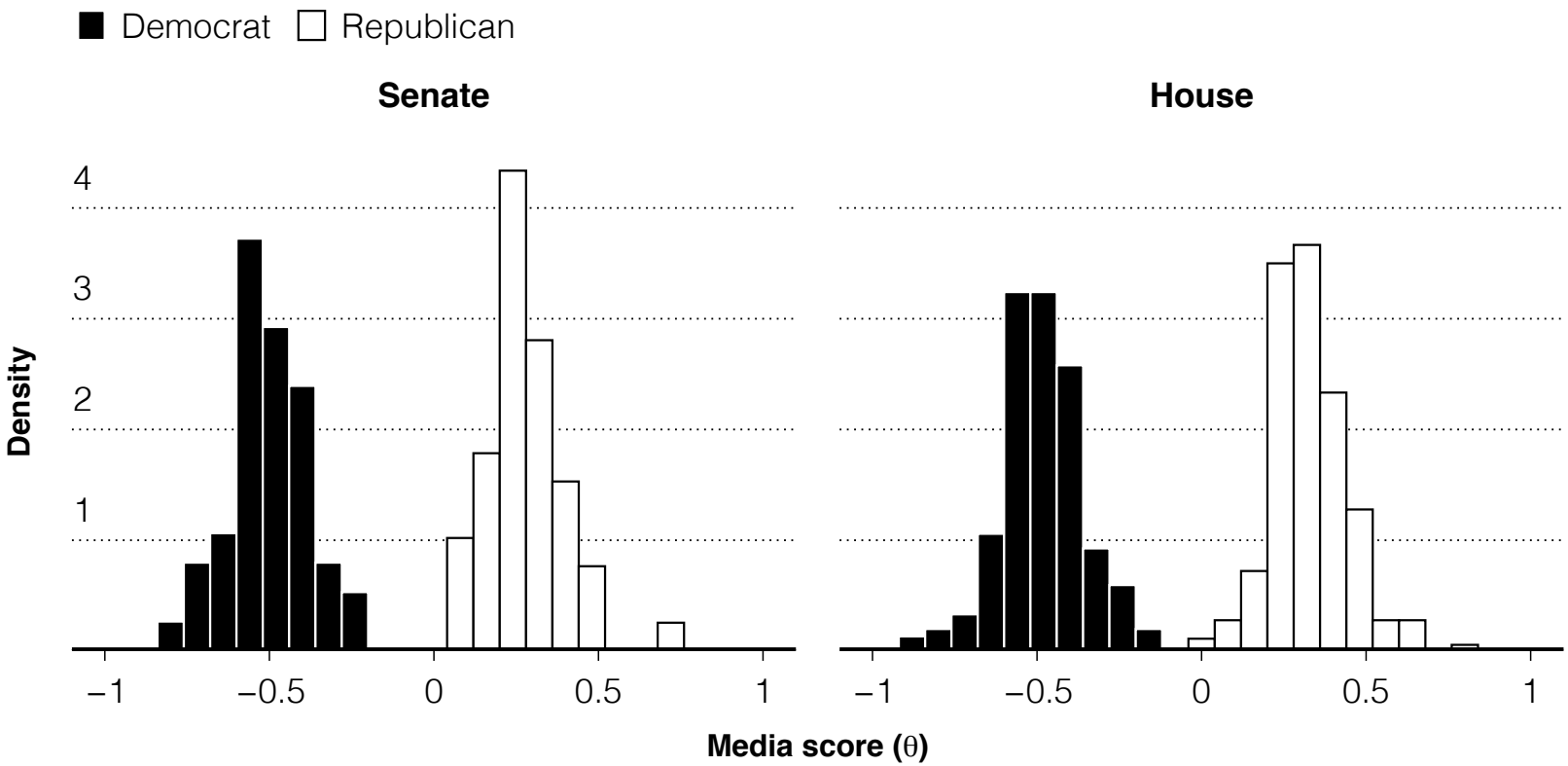


Scatterplot with regression line:
`geom_point() + geom_smooth()`

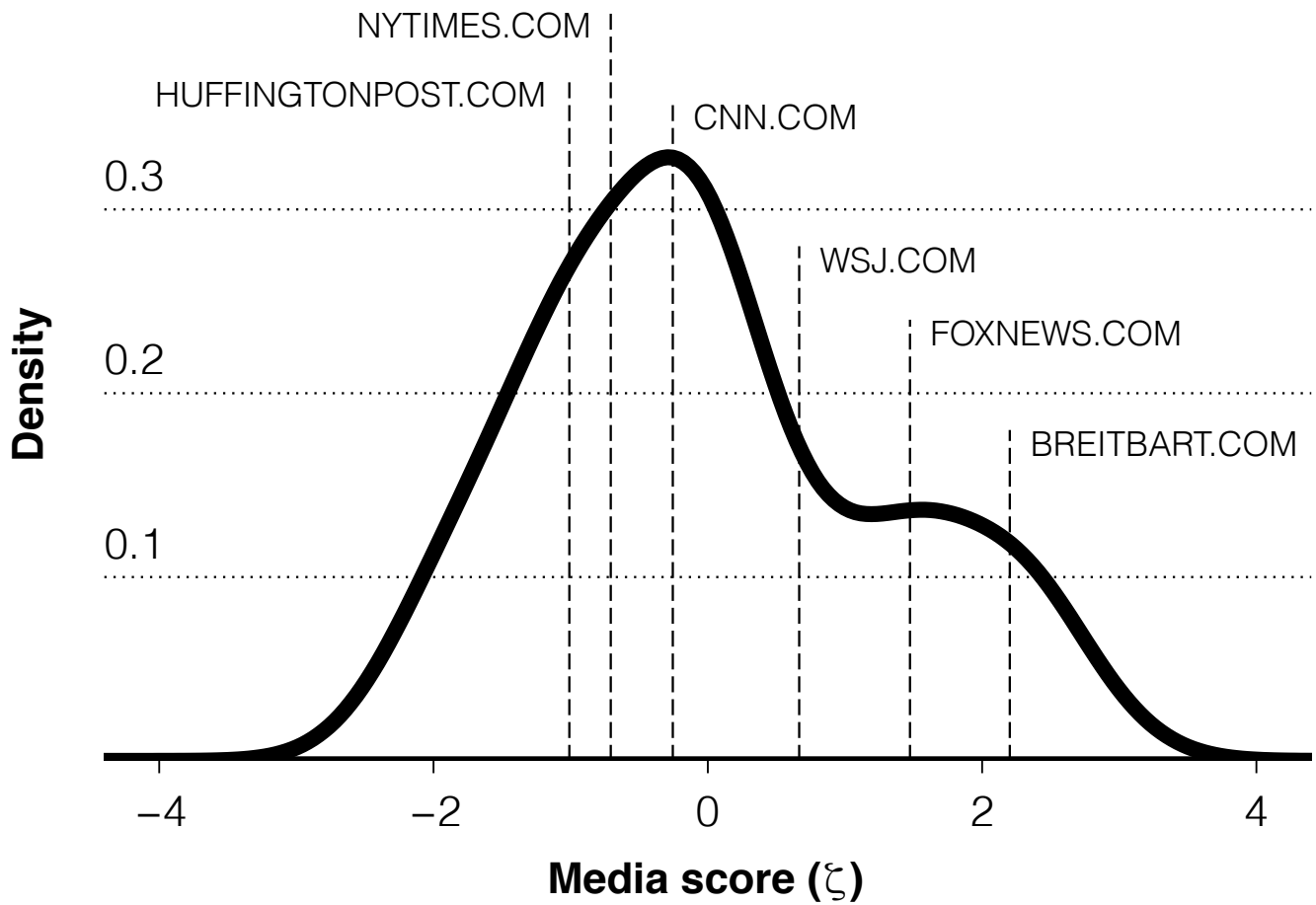
FIGURE 5. Regression discontinuity in posting activity 90 days before and after the ban (95% confidence interval)



Histogram: `geom_histogram()`

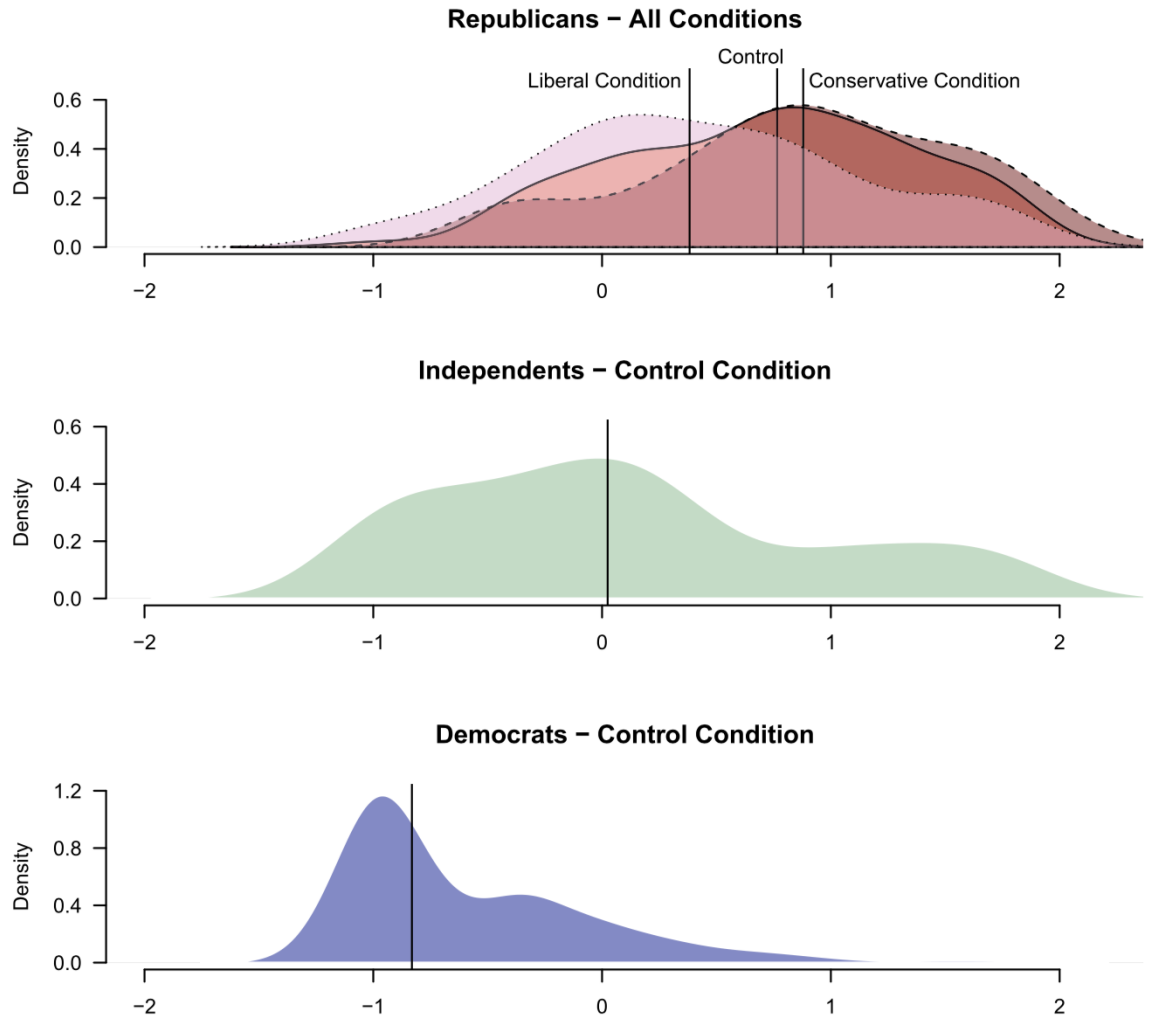


Density plot: `geom_density()`

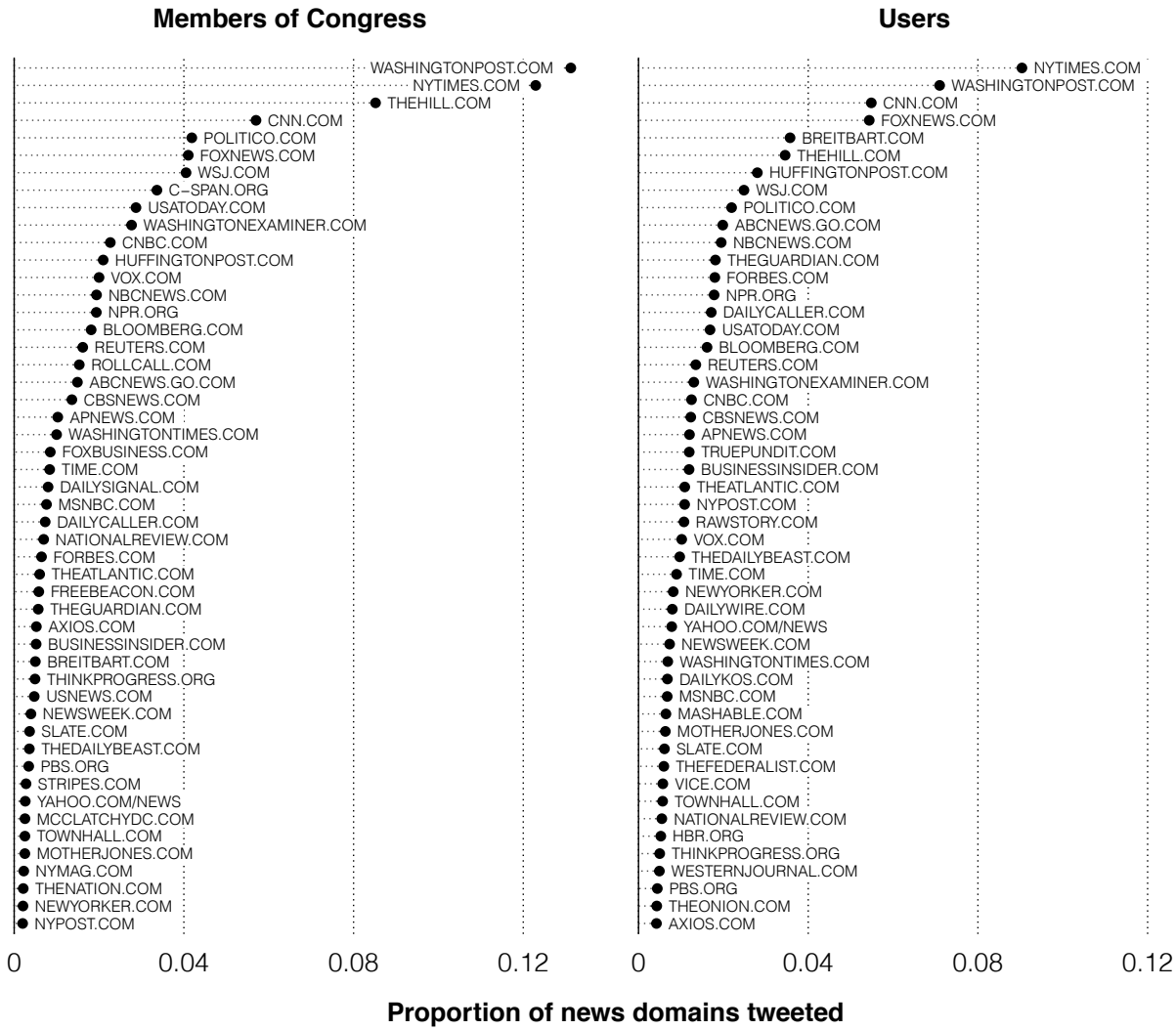


Density plot: `geom_density()`

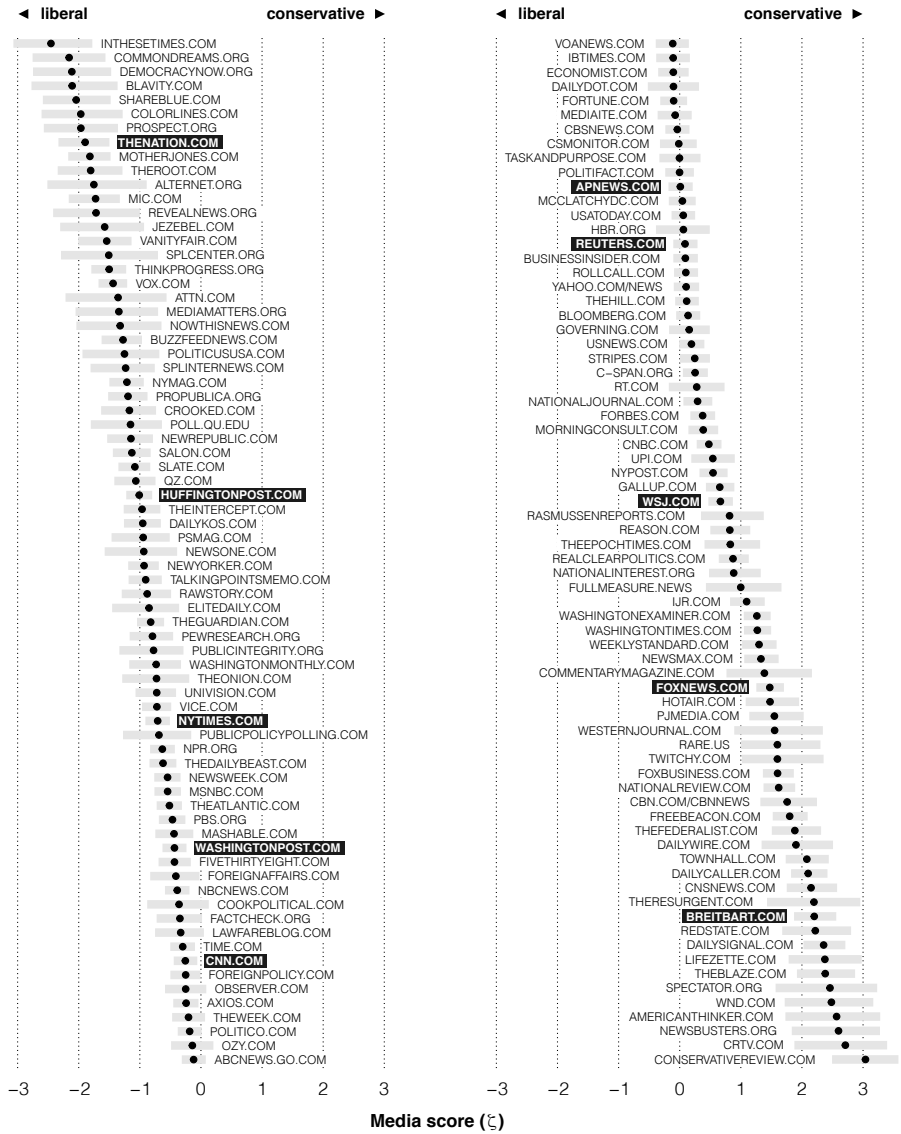
FIGURE 7. Ideological Distribution by Condition



Dot plot: `geom_point()`

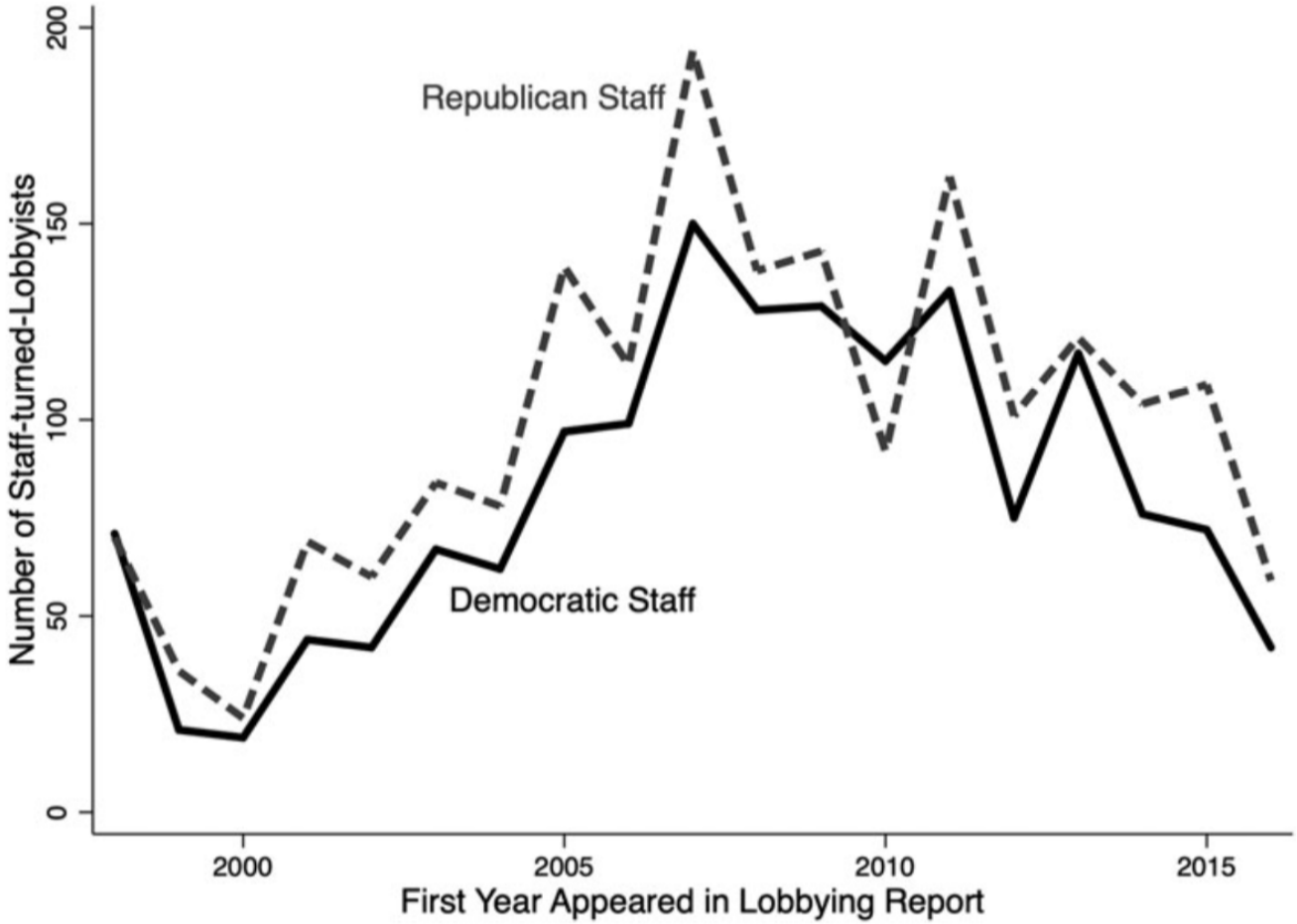


Dot plot (for coefficients): `geom_point()`

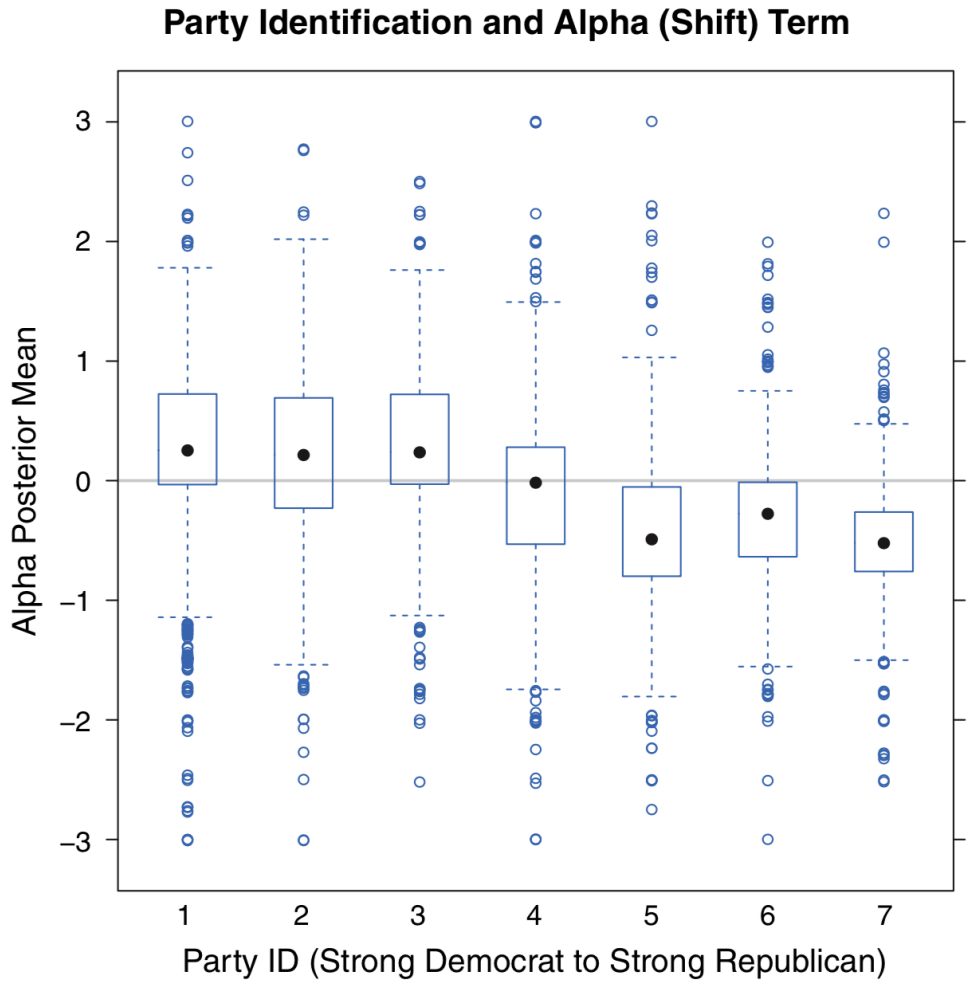


Line plot: `geom_line()`

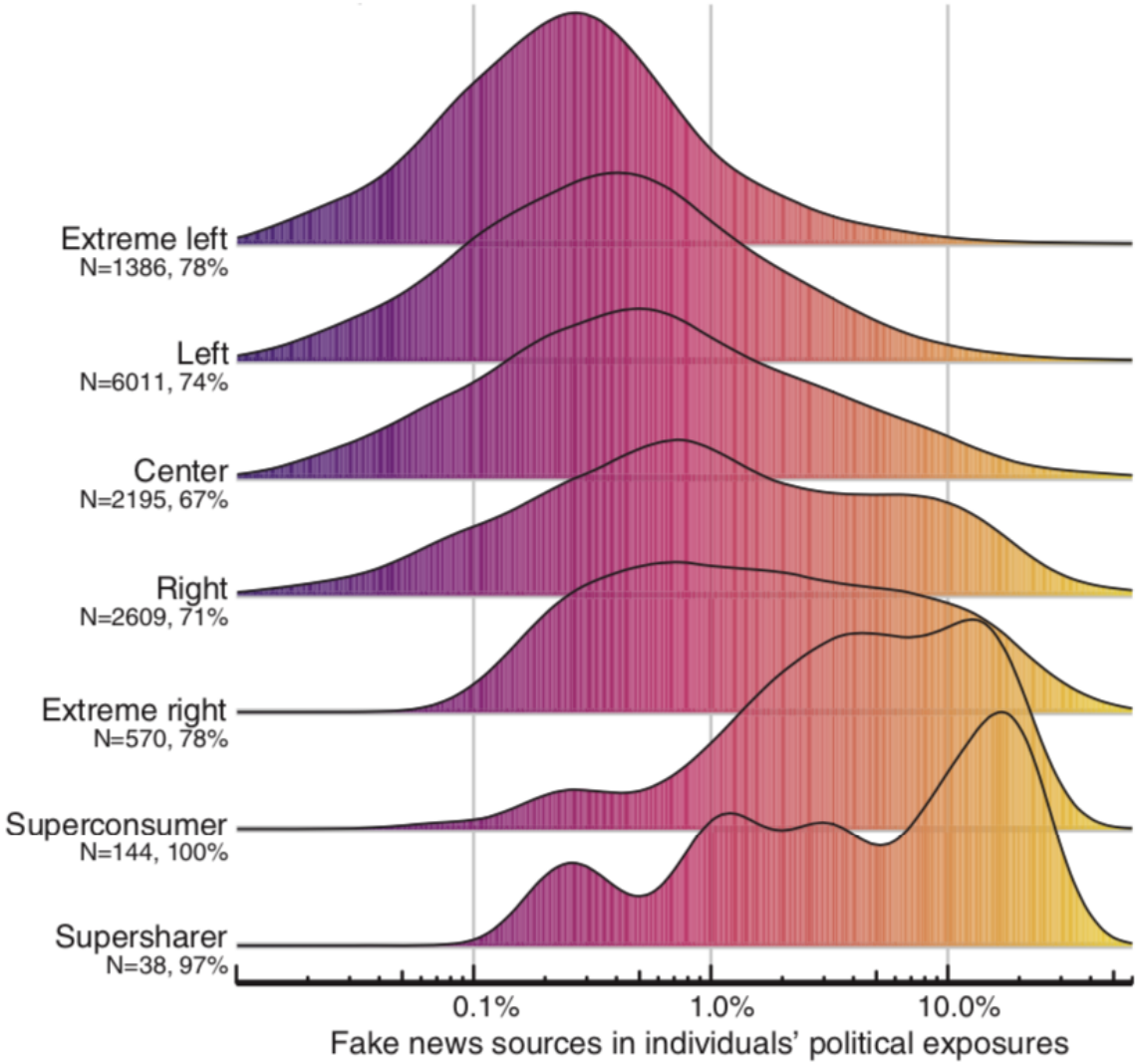
FIGURE 1. Number of Congressional Staffers-Turned-Lobbyists, 1998–2016



Boxplot: `geom_boxplot()`

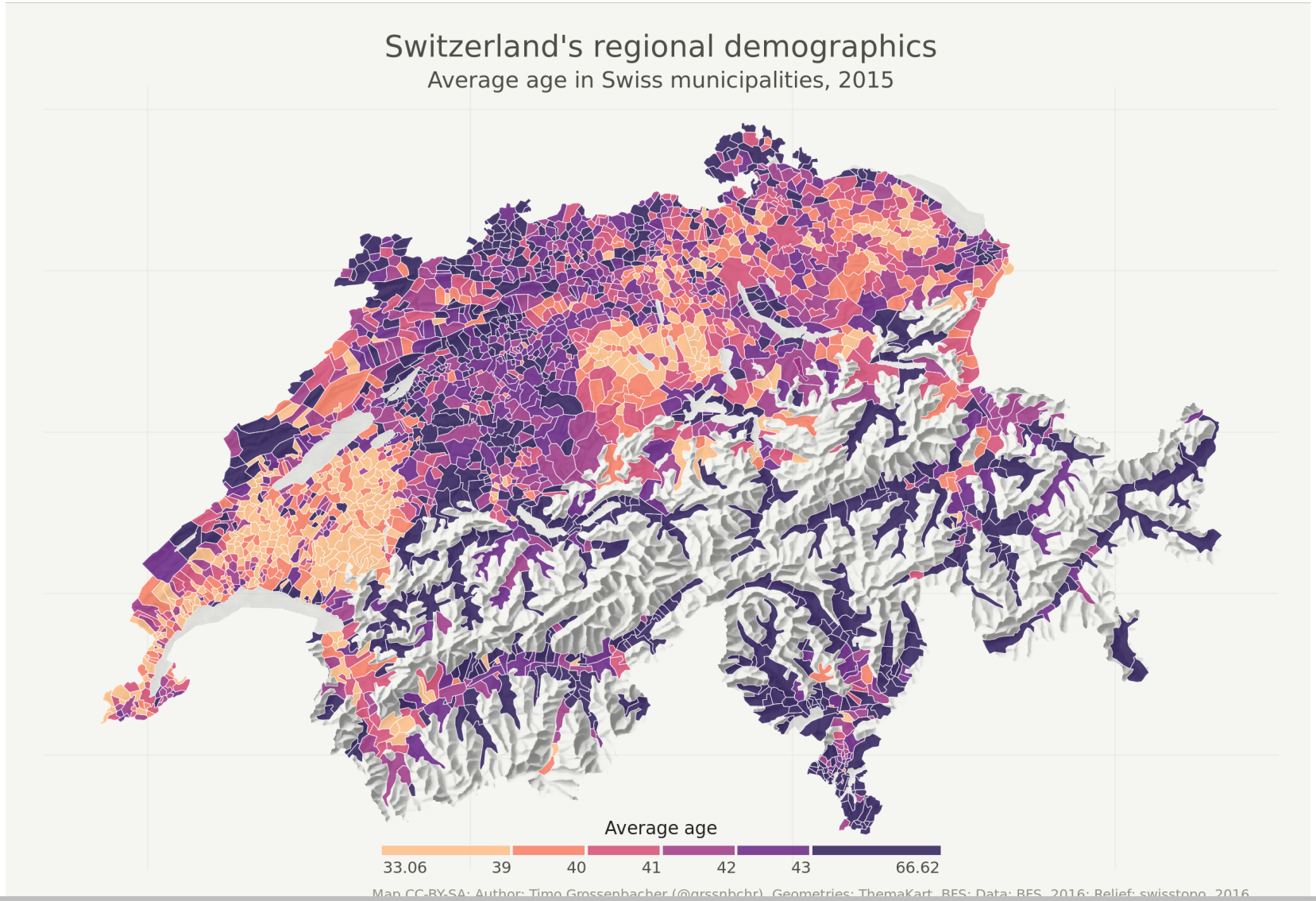


Ridge plot: `geom_density_ridges()`



Maps: `geom_polygon()`

Maps (choropleth): `geom_polygon()`



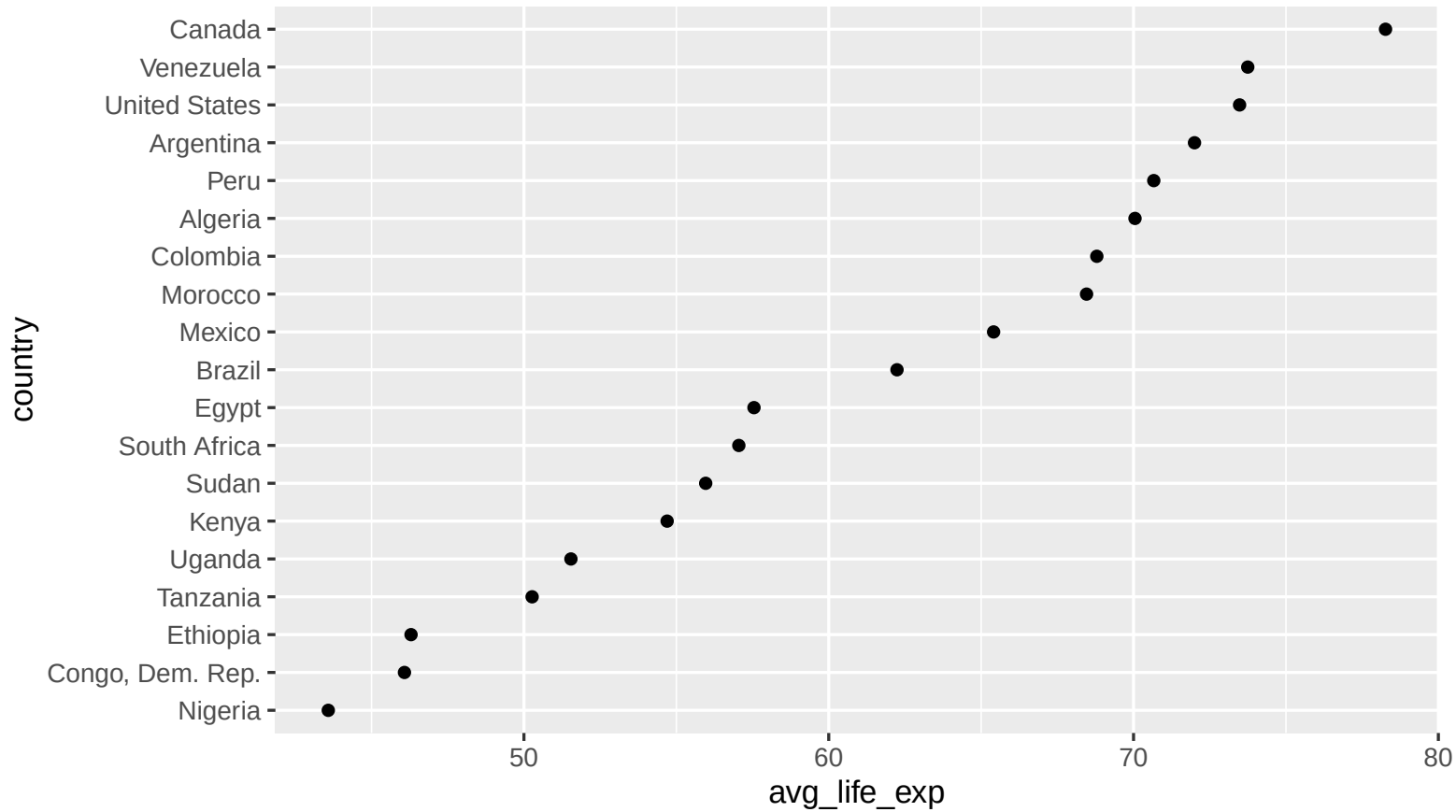
ggplot2 in practice

```
library(tidyverse)

# Set working directory
setwd("~/Downloads/")

# Load data
GM <- read_csv("GM.csv")

# Create data.frame for graph
# (1) Filter for only observations in Africa and the Americas with a
# population greater than 25 million people
# (2) Group by country
# (3) and take the average life expectancy across all years (for each country)
#     and keep the continent variable from each observation
# (4) Order the resulting country variable by its average life expectancy
# (5) Order the resulting continent variable manually, first yb "Africa" and
#     then by countries in the "Americas"
G1 <- GM %>%
      filter(continent %in% c("Africa", "Americas") & pop > 25000000) %>%
      group_by(country) %>%
      summarize(continent = unique(continent),
                avg_life_exp = mean(lifeExp)) %>%
      mutate(country = fct_reorder(country, avg_life_exp),
             continent = fct_relevel(continent, c("Africa", "Americas")))
```

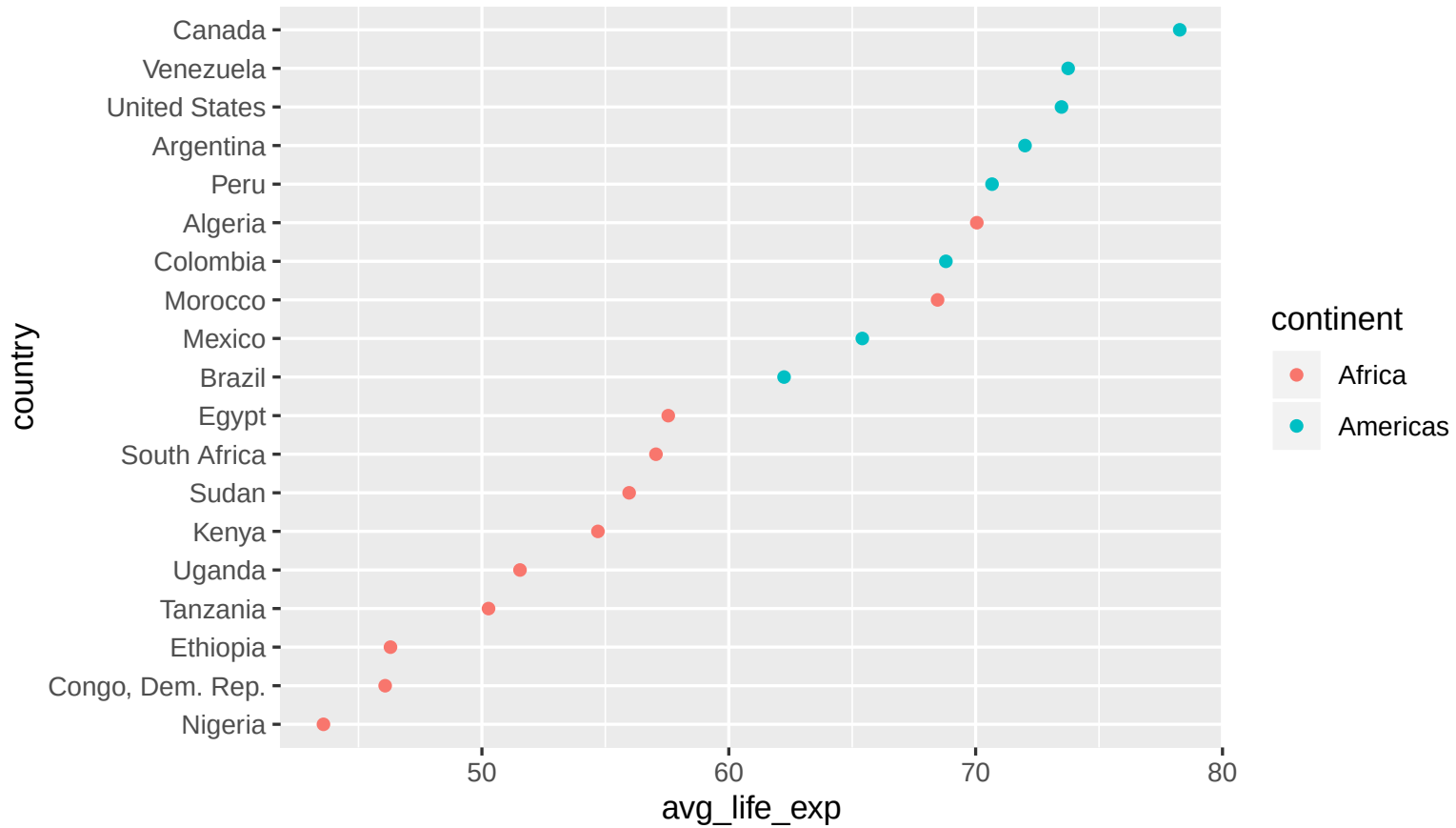
ggplot2 in practice

```
> G1
# A tibble: 19 x 3
   country           continent avg_life_exp
   <fct>             <fct>             <dbl>
 1 Algeria           Africa             70.0
 2 Argentina         Americas           72
 3 Brazil            Americas           62.2
 4 Canada            Americas           78.3
 5 Colombia          Americas           68.8
 6 Congo, Dem. Rep.  Africa             46.1
 7 Egypt             Africa             57.5
 8 Ethiopia          Africa             46.3
 9 Kenya             Africa             54.7
10 Mexico            Americas           65.4
11 Morocco           Africa             68.5
12 Nigeria           Africa             43.6
13 Peru              Americas           70.7
14 South Africa      Africa             57.0
15 Sudan             Africa             56.0
16 Tanzania          Africa             50.3
17 Uganda            Africa             51.5
18 United States     Americas           73.5
19 Venezuela         Americas           73.7
```

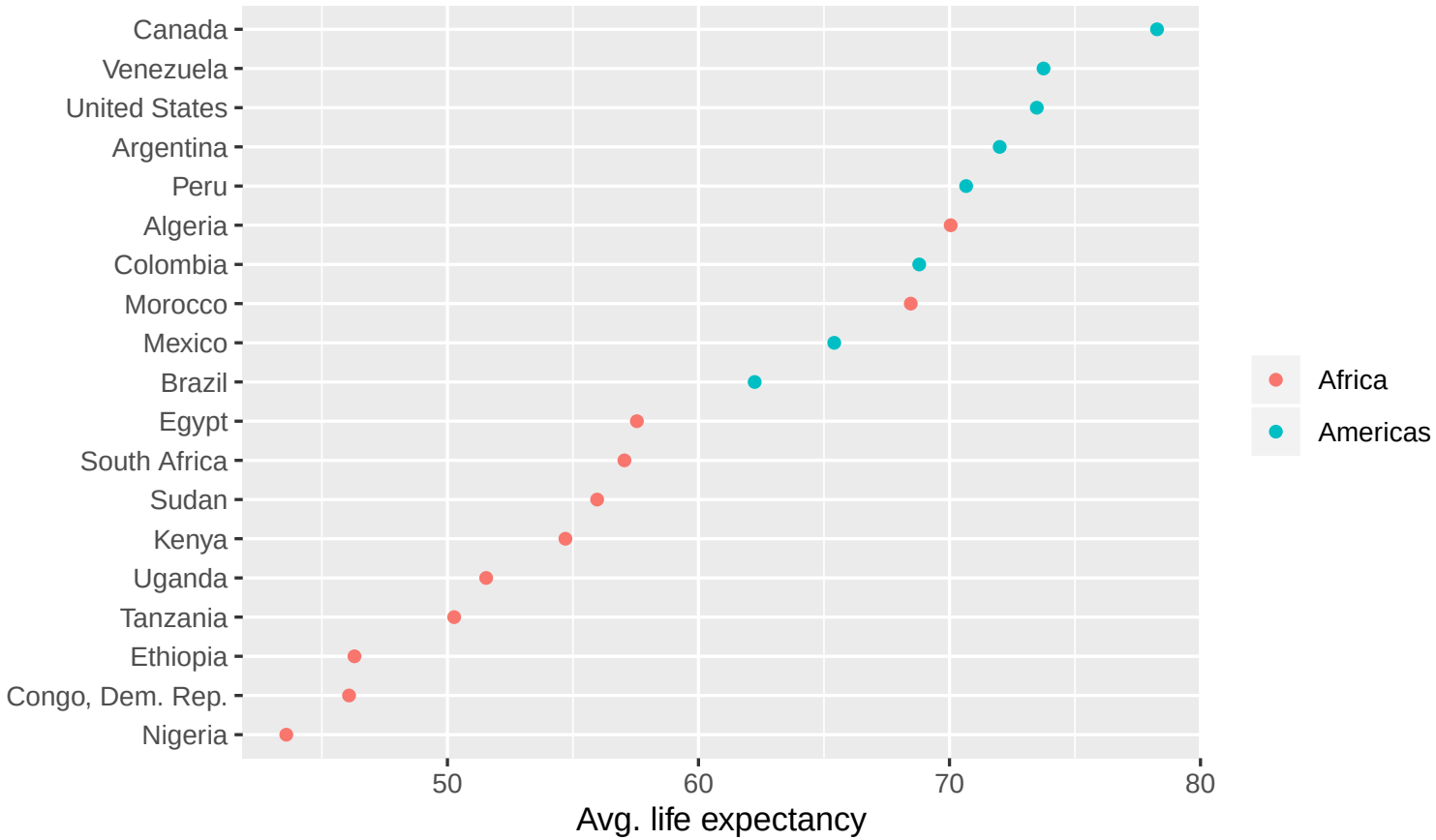
```
ggplot(G1, aes(x = avg_life_exp, y = country)) +
  geom_point()
```



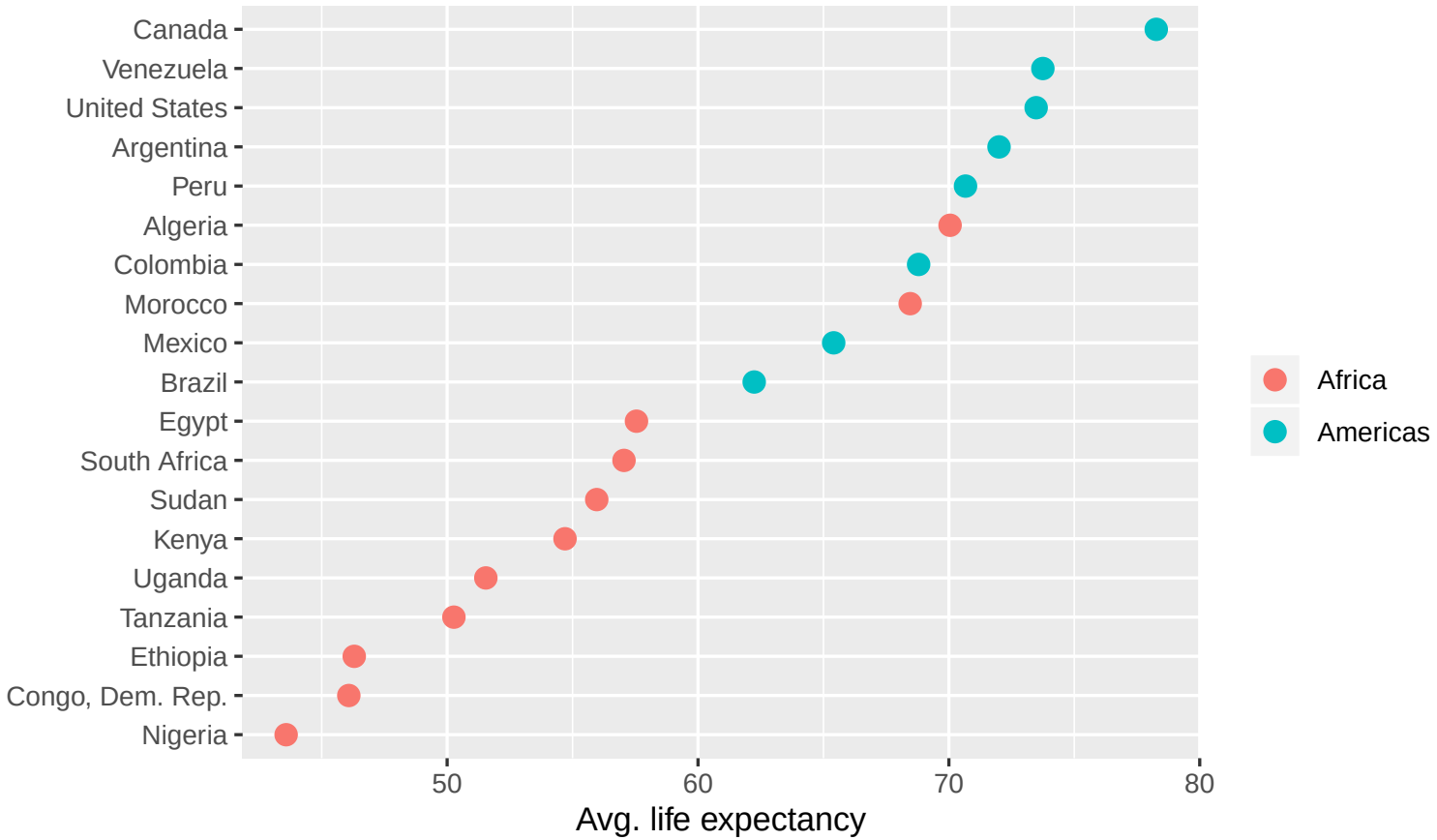
```
ggplot(G1, aes(x = avg_life_exp, y = country, color = continent)) +
  geom_point()
```



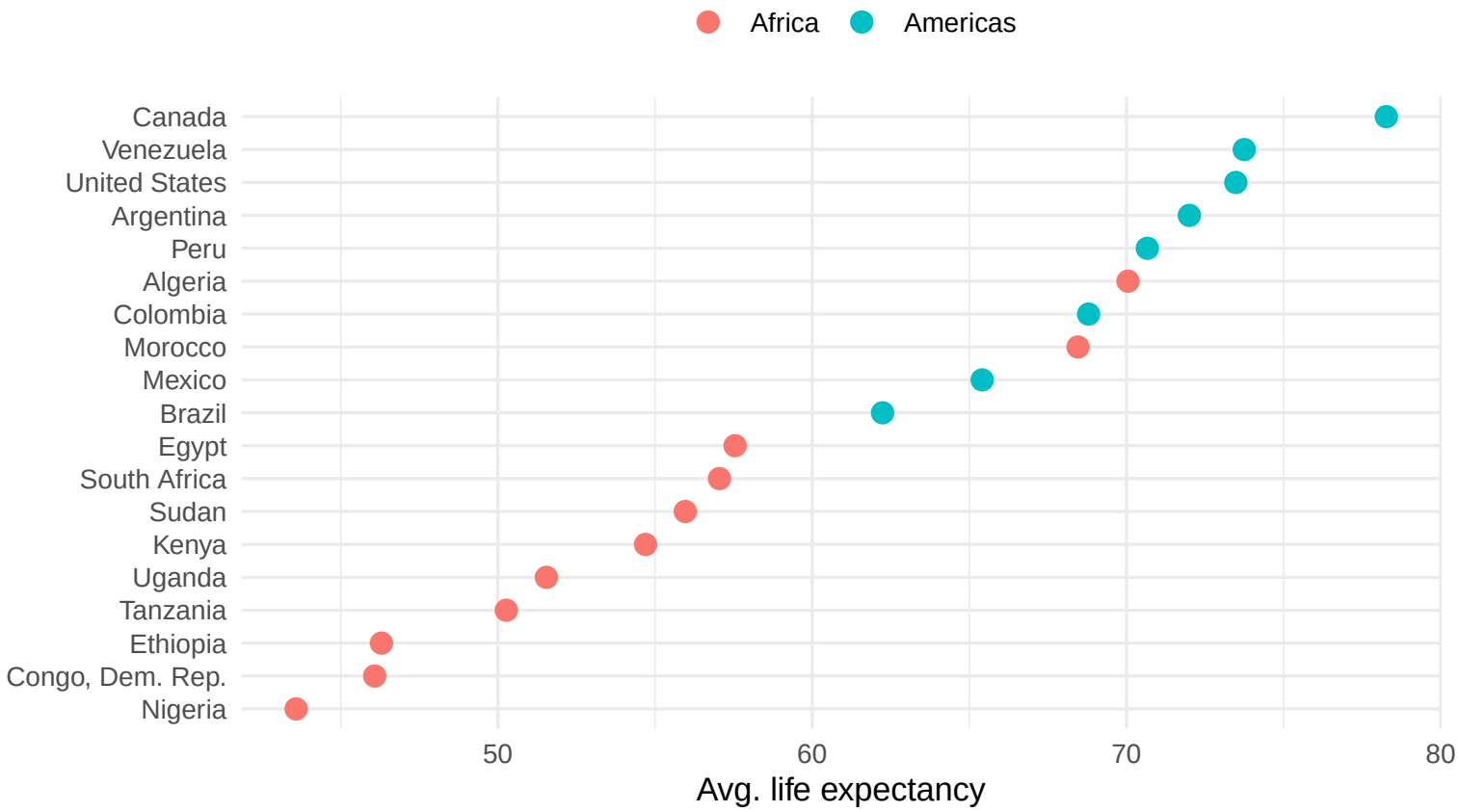
```
ggplot(G1, aes(x = avg_life_exp, y = country, color = continent)) +
  geom_point() +
  labs(x = "Avg. life expectancy", y = "", color = "")
```



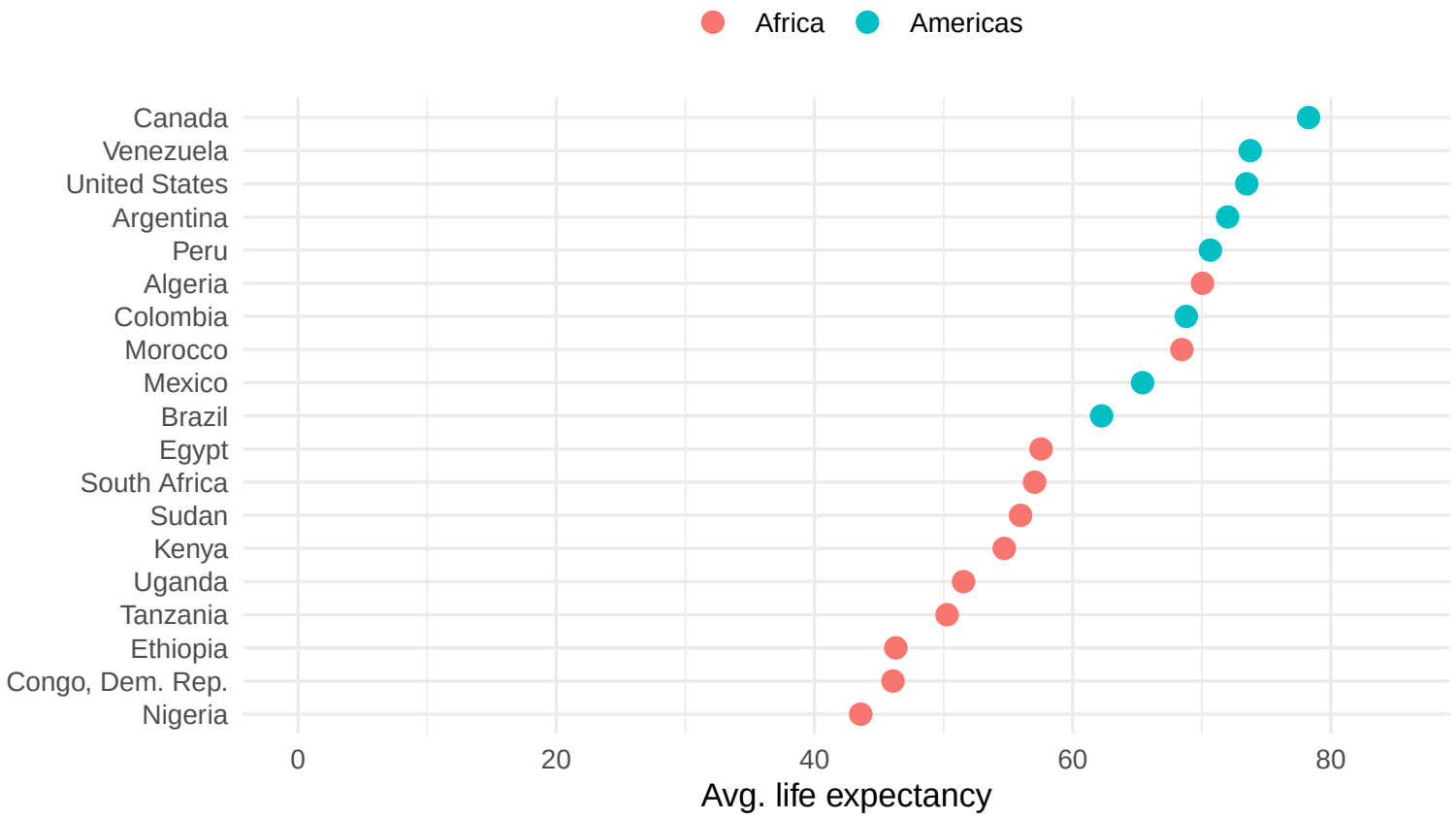
```
ggplot(G1, aes(x = avg_life_exp, y = country, color = continent)) +
  geom_point(size = 3) +
  labs(x = "Avg. life expectancy", y = "", color = "")
```



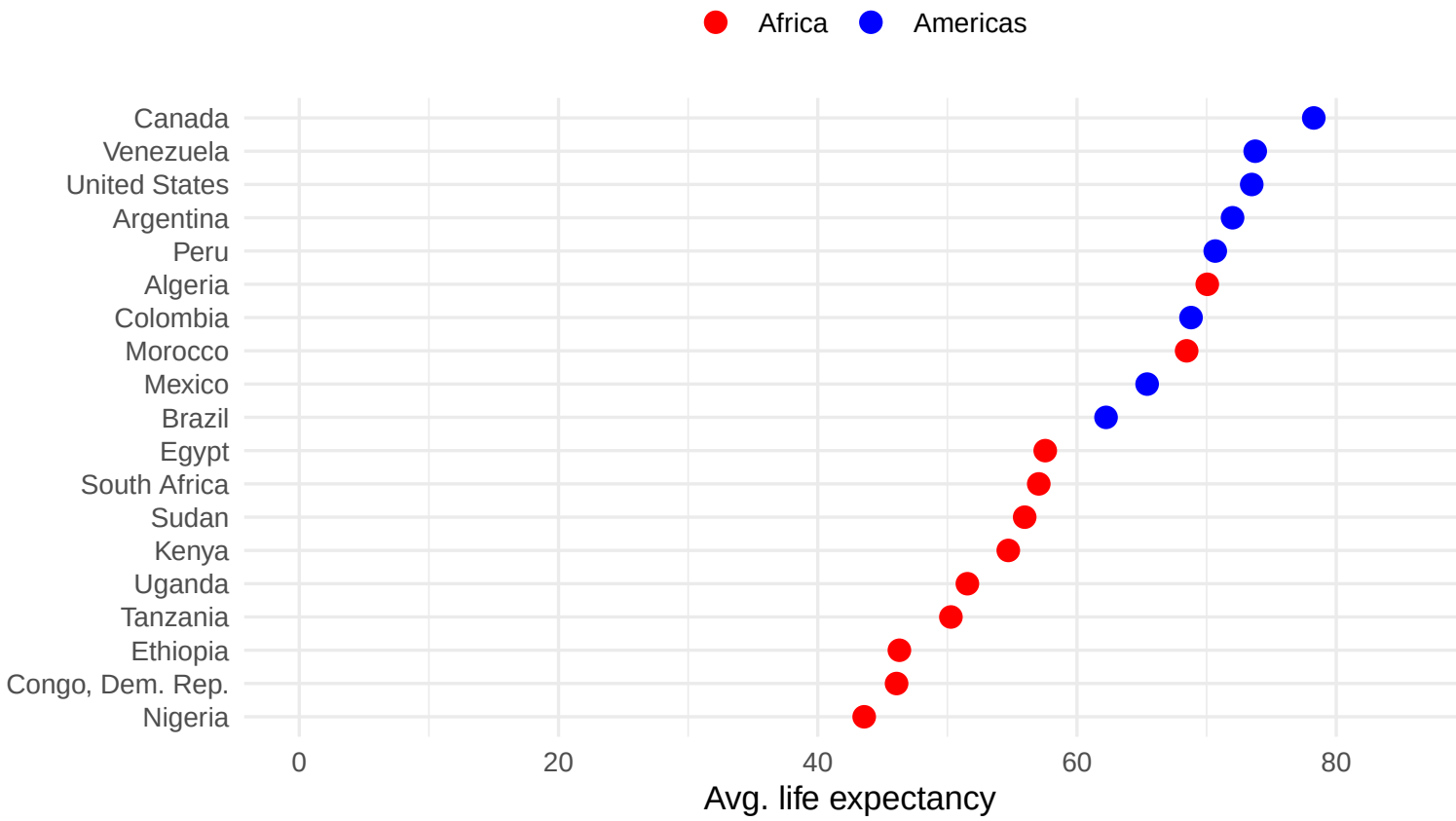
```
ggplot(G1, aes(x = avg_life_exp, y = country, color = continent)) +
  geom_point(size = 3) +
  labs(x = "Avg. life expectancy", y = "", color = "") +
  theme_minimal() +
  theme(legend.position = "top")
```



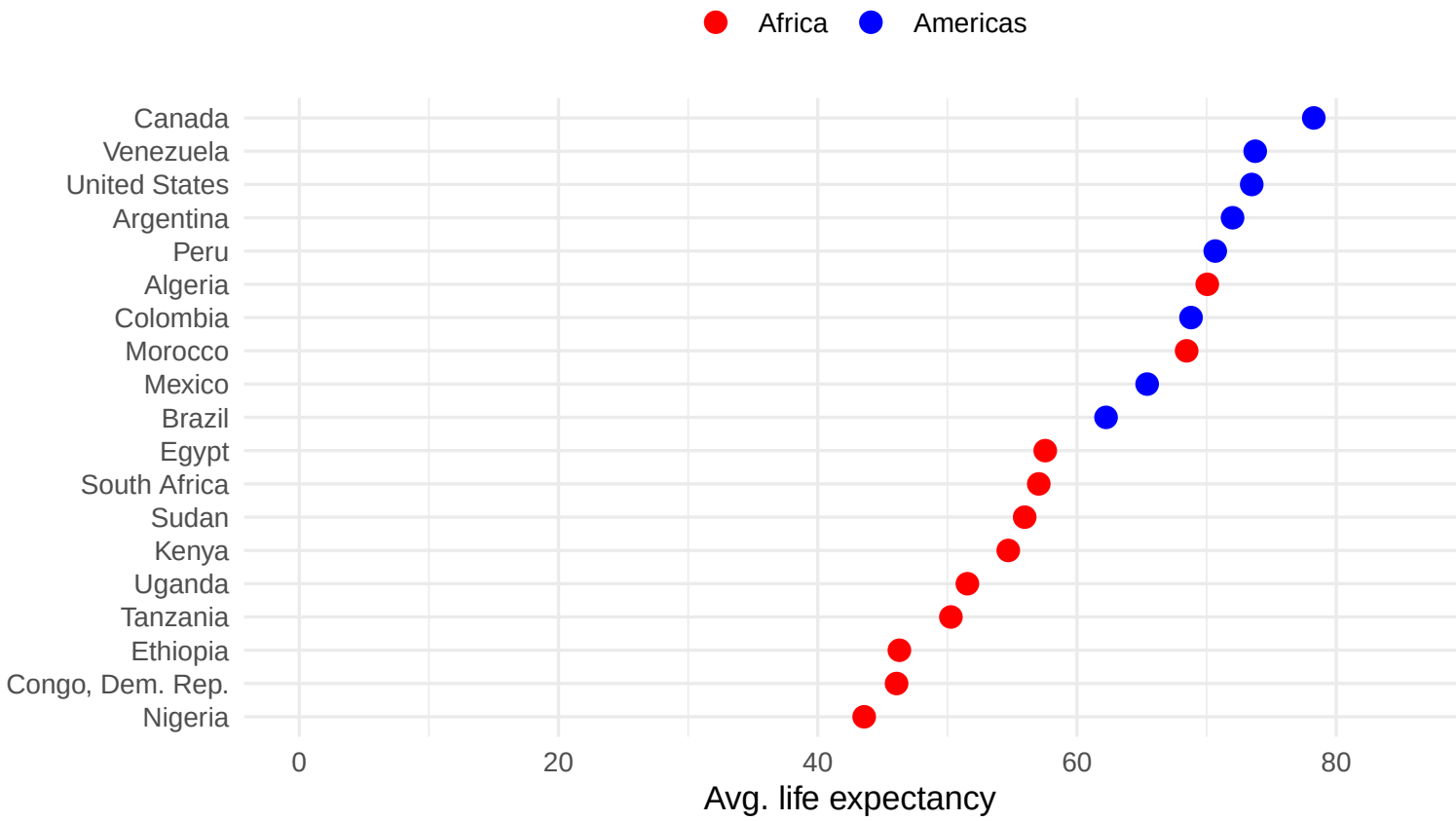
```
ggplot(G1, aes(x = avg_life_exp, y = country, color = continent)) +
  geom_point(size = 3) +
  coord_cartesian(xlim = c(0, 85)) +
  labs(x = "Avg. life expectancy", y = "", color = "") +
  theme_minimal() +
  theme(legend.position = "top")
```



```
ggplot(G1, aes(x = avg_life_exp, y = country, color = continent)) +
  geom_point(size = 3) +
  coord_cartesian(xlim = c(0, 85)) +
  labs(x = "Avg. life expectancy", y = "", color = "") +
  scale_color_manual(values = c("Africa" = "red", "Americas" = "blue")) +
  theme_minimal() +
  theme(legend.position = "top")
```



```
pdf("/Path/On/My/Computer/My_Graph.pdf, 7, 4) # Or png("...", 700, 400)
ggplot(G1, aes(x = avg_life_exp, y = country, color = continent)) +
  geom_point(size = 3) +
  coord_cartesian(xlim = c(0, 85)) +
  labs(x = "Avg. life expectancy", y = "", color = "") +
  scale_color_manual(values = c("Africa" = "red", "Americas" = "blue")) +
  theme_minimal() +
  theme(legend.position = "top")
dev.off()
```

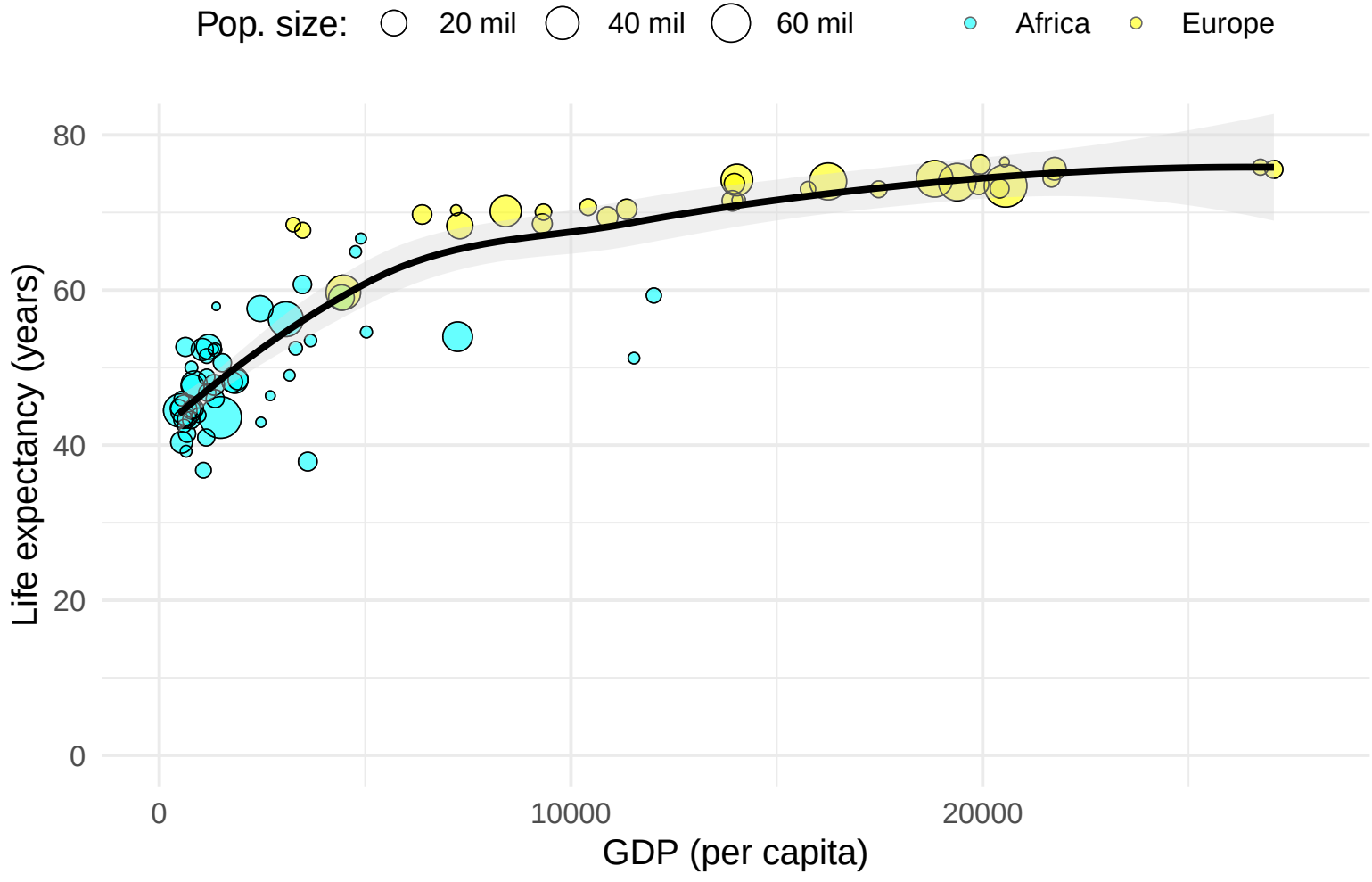


Exercises

Exercise 1 solution

```
# Generate data for Exercise 1
G2 <- GM %>%
      filter(continent %in% c("Europe", "Africa")) %>%
      group_by(country) %>%
      summarize(continent = unique(continent),
                avg_life_exp = mean(lifeExp),
                population = mean(pop),
                gdp = mean(gdpPercap)) %>%
      arrange(avg_life_exp)

# Highest and lowest life expectancy?
head(G2) # To see the lower life expectancy
tail(G2) # To see the lower life expectancy

# Graph for Exercise 1
pdf("Exercise_1_Graph.pdf", 6, 4)
ggplot(G2, aes(x = gdp, y = avg_life_exp,
               size = population, fill = continent)) +
  geom_point(shape = 21, alpha = 0.6, stroke = 0.3) +
  stat_smooth(color = "black", fill = "grey85", size = 1) +
  coord_cartesian(xlim = c(0, 28000), ylim = c(0, 80)) +
  labs(x = "GDP (per capita)", y = "Life expectancy (years)",
       size = "Pop. size:", fill = "") +
  scale_fill_manual(values = c("Africa" = "cyan", "Europe" = "yellow")) +
  scale_size(breaks = c(20000000, 40000000, 60000000),
             labels = c("20 mil", "40 mil", "60 mil")) +
  theme_minimal() +
  theme(legend.position = "top")
dev.off()
```

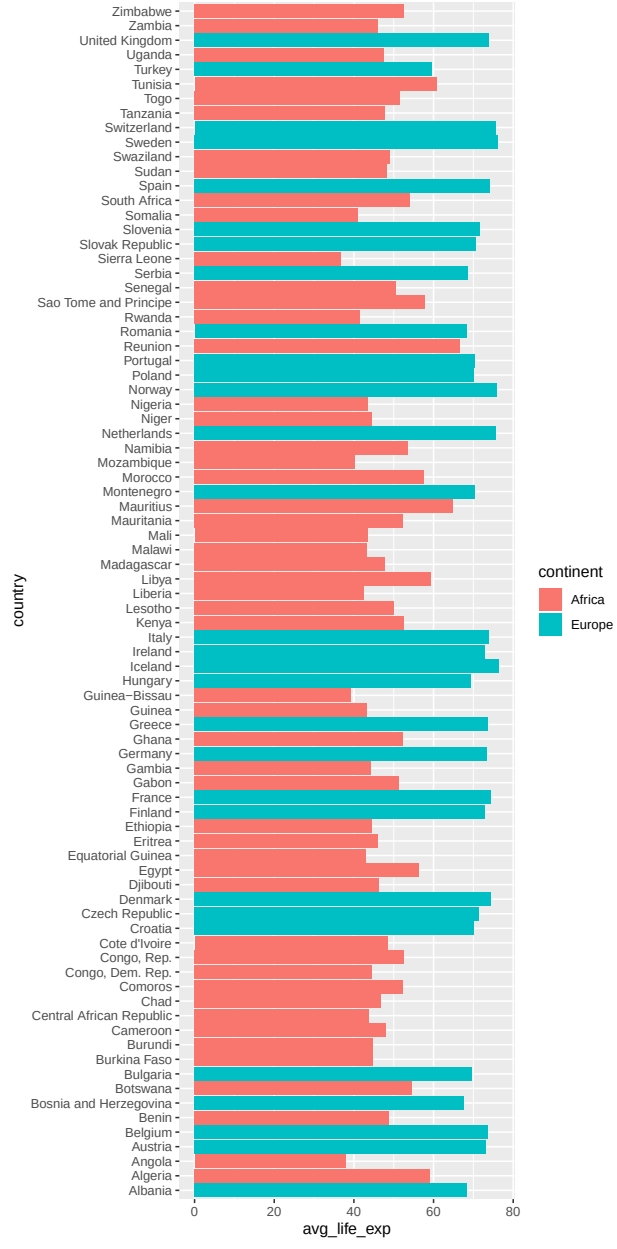
Exercise 1 graph



Exercise 2 solution

```
# Graph for Exercise 2
ggplot(G1, aes(x = country, y = avg_life_exp, fill = continent)) +
  geom_col() +
  coord_flip()
```

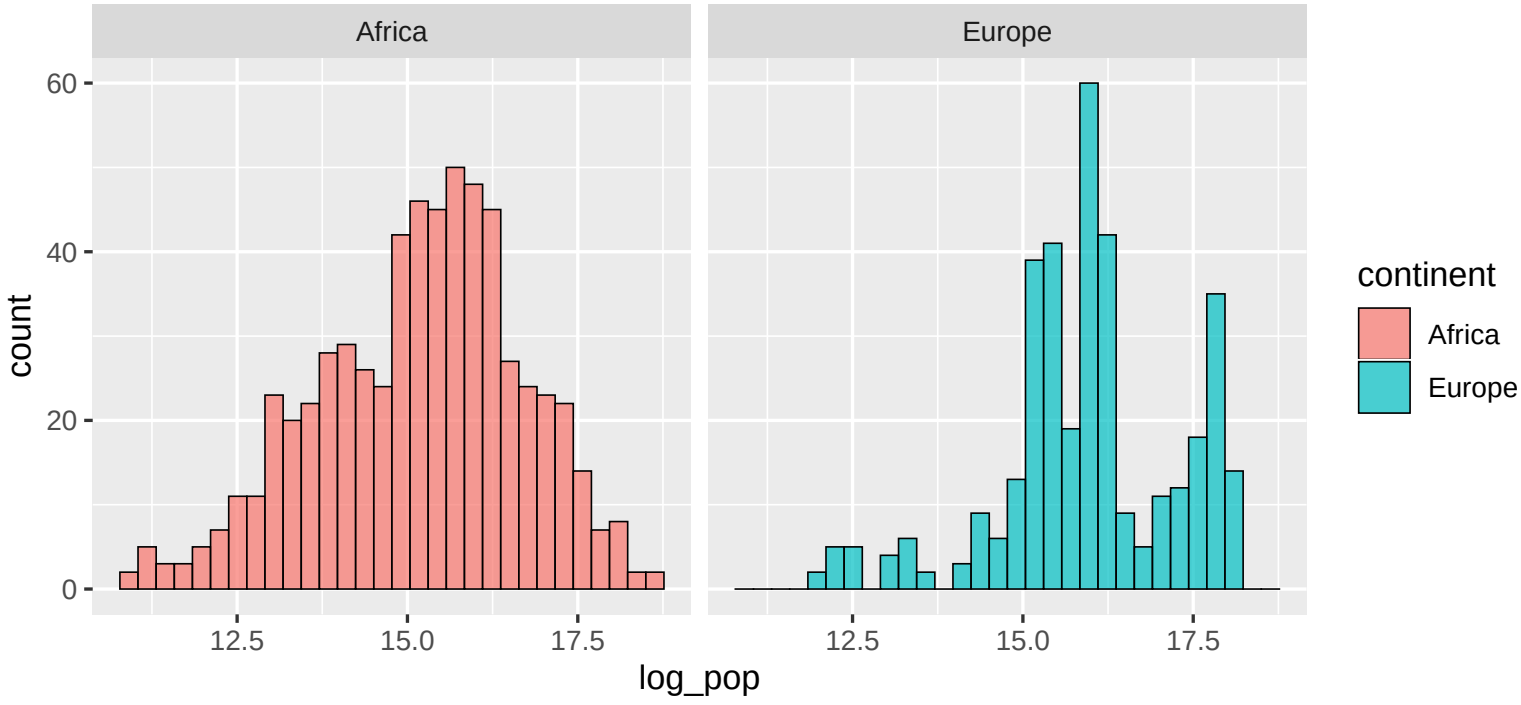
Exercise 2 graph



Exercise 3 solution

```
# Generate data for Exercise 3
G3 <- GM %>%
      filter(continent %in% c("Europe", "Africa")) %>%
      mutate(log_pop = log(pop))

# Graph for Exercise 3
ggplot(G3, aes(x = log_pop, fill = continent)) +
  geom_histogram(alpha = 0.7, color = "black", size = 0.25) +
  facet_wrap(~ continent)
```

Exercise 3 graph



Exercise 4 solution

```
# Generate data for Exercise 4
G4 <- GM %>%
      filter(country %in% c("Denmark", "Canada", "United States"))

# Graph for Exercise 4
ggplot(G4, aes(x = year, y = gdpPercap, color = country, shape = country)) +
  geom_line() +
  geom_point()
```

Exercise 4 graph