Homework #01: Data Visualization

due January 27, 2021 11:59 PM

Dav King

1/20/22

Load Packages

```
library(tidyverse)
library(viridis)
library(ggridges)
library(scales)
```

Load Data

```
anes <- read_csv("anes2020_subset.csv")
```

Exercise 1

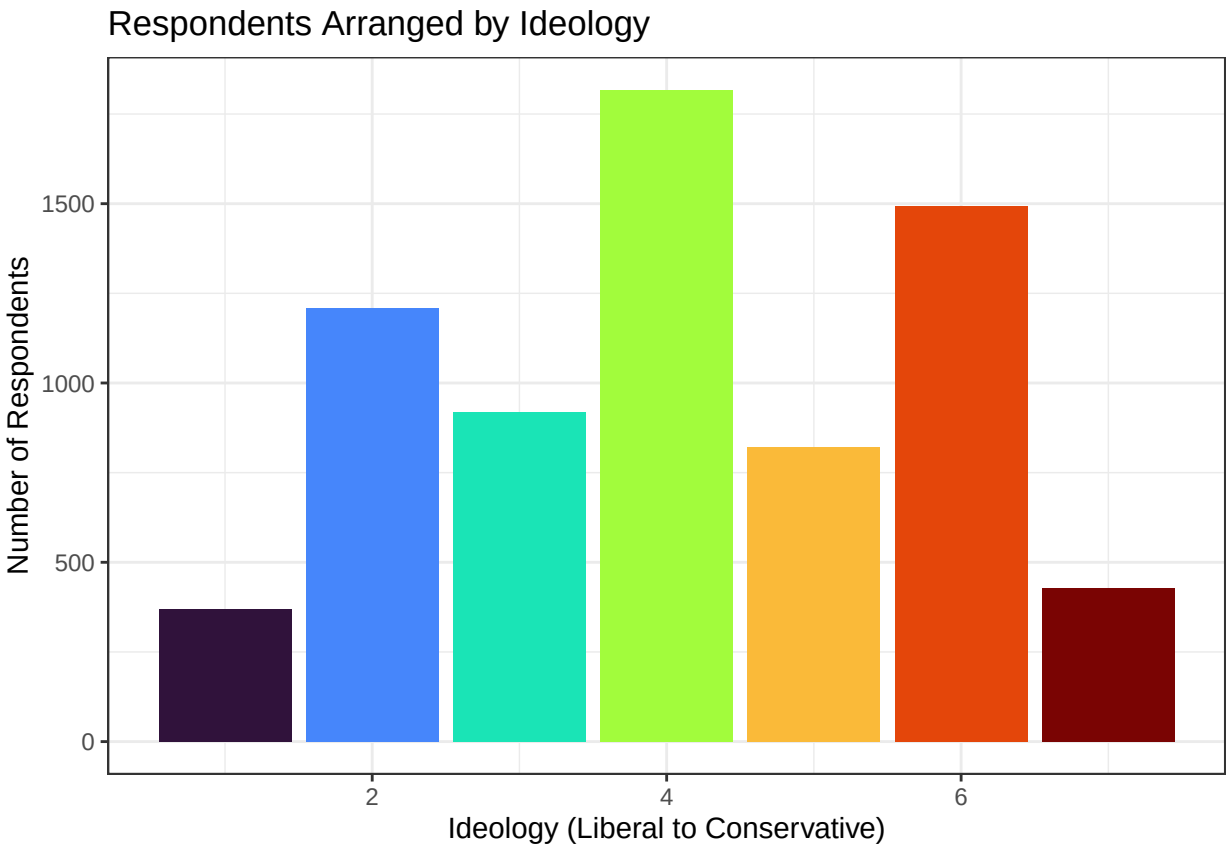
```
glimpse(anes)
```

```
## Rows: 8,280
## Columns: 6
## $ CASEID     <dbl> 200015, 200022, 200039, 200046, 200053, 200060, 200084, 200~
## $ hunt_fish  <dbl> 1, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0,~
## $ scientists <dbl> 100, 70, 100, 85, 60, 85, 85, NA, 60, 50, 85, 100, 70, 70, ~
## $ education  <dbl> 6, 3, 2, 4, 8, 3, 4, 2, 2, 4, 2, 2, 2, 7, 3, 3, 6, 6, 6, 2,~
## $ ideology   <dbl> 6, 4, 2, 3, 5, 4, 4, NA, NA, NA, NA, 4, 6, 3, 4, 6, 3, 3, 2~
## $ urbanrural <chr> "city", "suburb", "suburb", "small town", "city", "suburb",~
```

There are 8,280 rows and 6 columns in the `anes` dataset.

Exercise 2

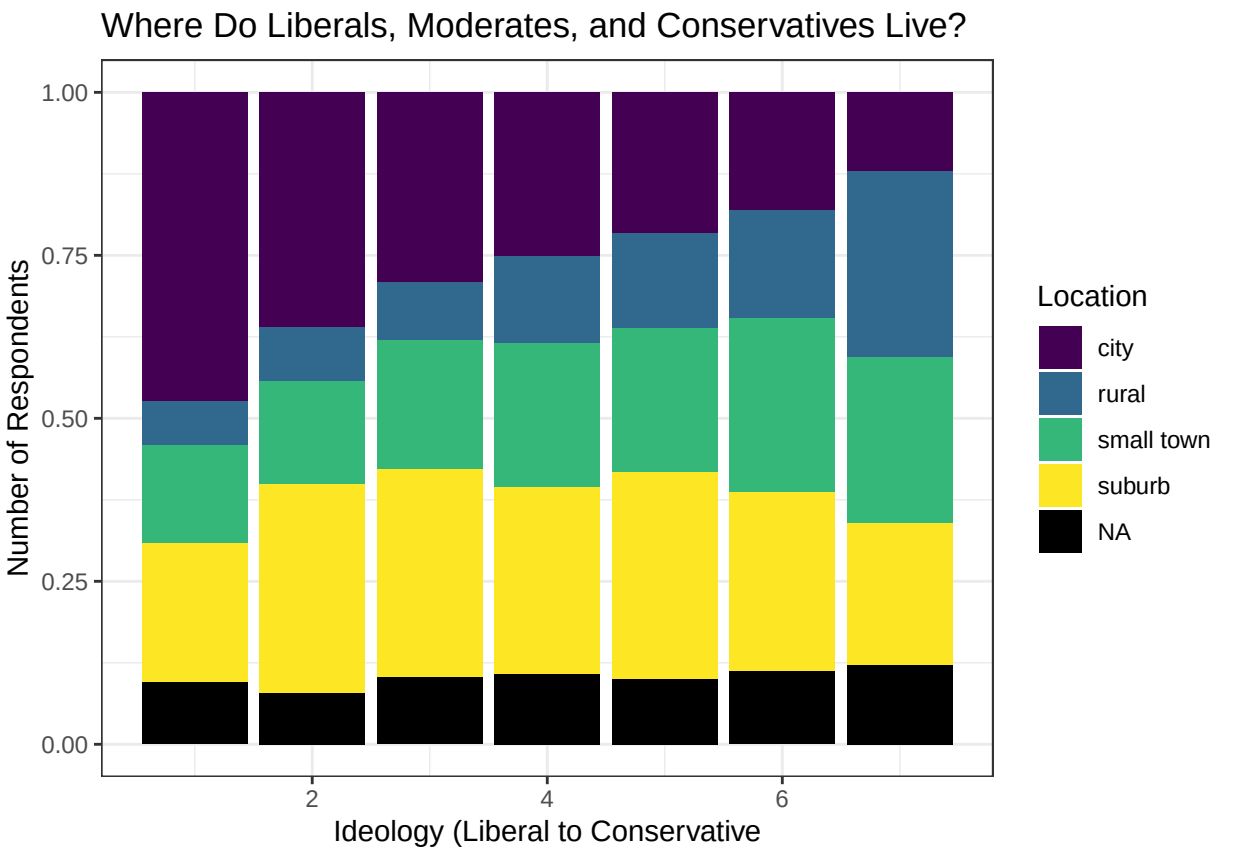
```
ggplot(anes, aes(x = ideology, fill = as.factor(ideology))) +
  geom_bar() +
  labs(title = "Respondents Arranged by Ideology",
       x = "Ideology (Liberal to Conservative)",
       y = "Number of Respondents") +
  theme_bw() +
  scale_fill_viridis(discrete = TRUE, option= "H") +
  guides(fill = "none")
```



The most common ideology is 4 (on a scale from 1-7), or the people who identify themselves as true moderates. In general, respondents tend to label themselves as more moderate than they do extreme (regardless of how true this is in reality).

Exercise 3

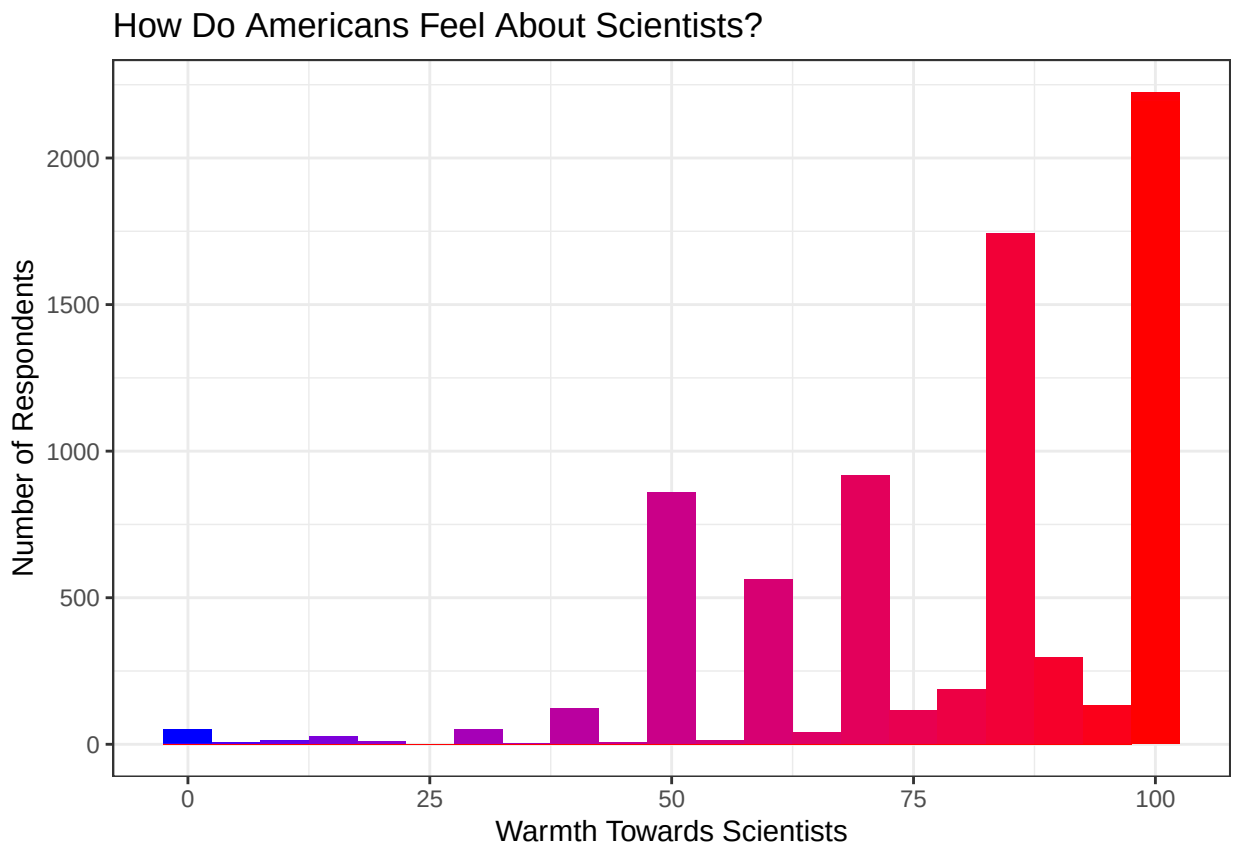
```
ggplot(anes, aes(x = ideology, fill = urbanrural)) +
  geom_bar(position = "fill") +
  labs(title = "Where Do Liberals, Moderates, and Conservatives Live?",
       x = "Ideology (Liberal to Conservative",
       y = "Number of Respondents",
       fill = "Location") +
  scale_fill_viridis(discrete = TRUE, option = "D", na.value = "Black") +
  theme_bw()
```



People who are more liberal tend to live in cities, while people who are more conservative tend to live in more rural places (each of those locations appears directly tied to ideology, with no exceptions, relative to other ideologies). However, people who are more moderate are relatively more prevalent in the suburbs. Though people in small towns are more or less evenly distributed ideologically, people who are more conservative tend to live in small towns more often than people who are very liberal. The percentage non-responses is almost entirely evenly distributed, though it might be a slight bit higher among conservatives than liberals.

Exercise 4

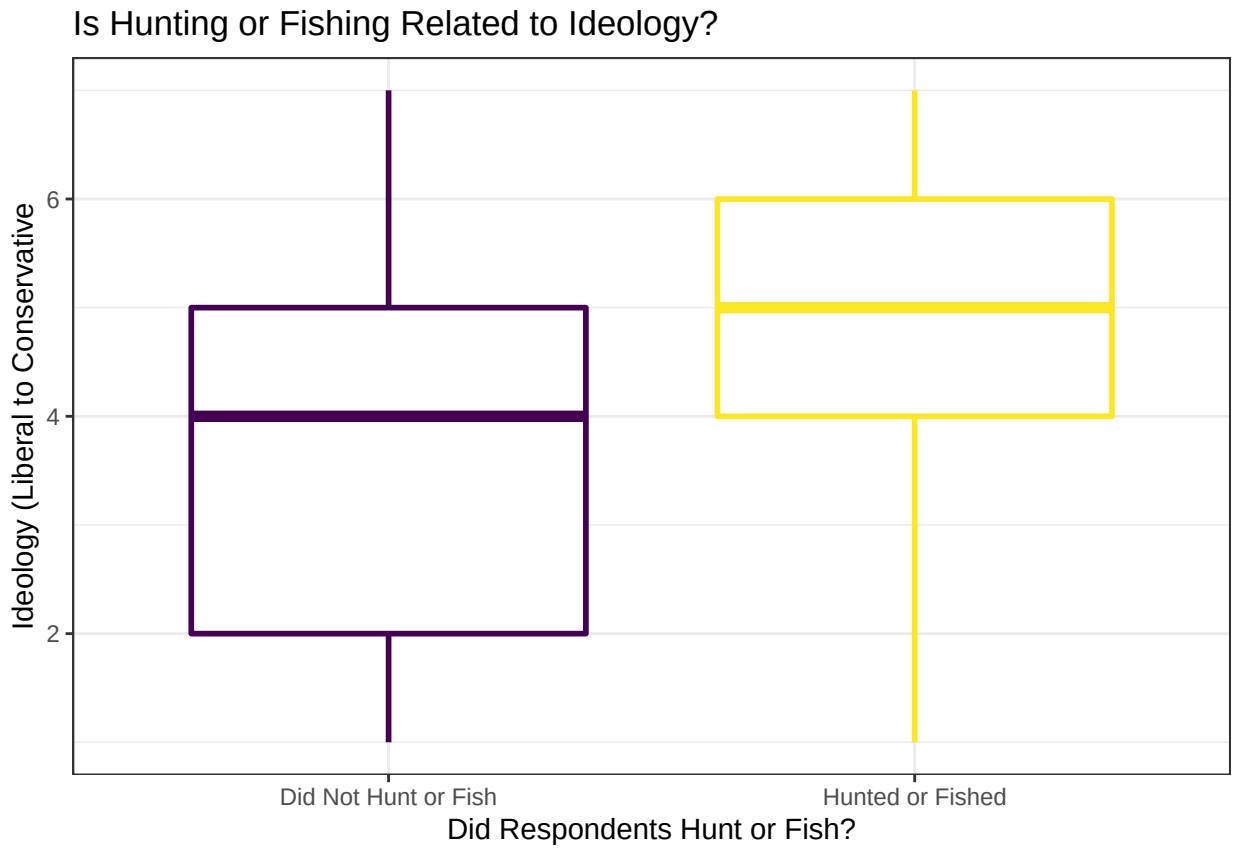
```
ggplot(anes, aes(x = scientists, fill = as.integer(scientists),
                 group = scientists)) +
  geom_histogram(binwidth = 5) +
  labs(title = "How Do Americans Feel About Scientists?",
       x = "Warmth Towards Scientists",
       y = "Number of Respondents") +
  theme_bw() +
  scale_fill_gradient(low = "Blue", high = "Red") +
  guides(fill = "none")
```



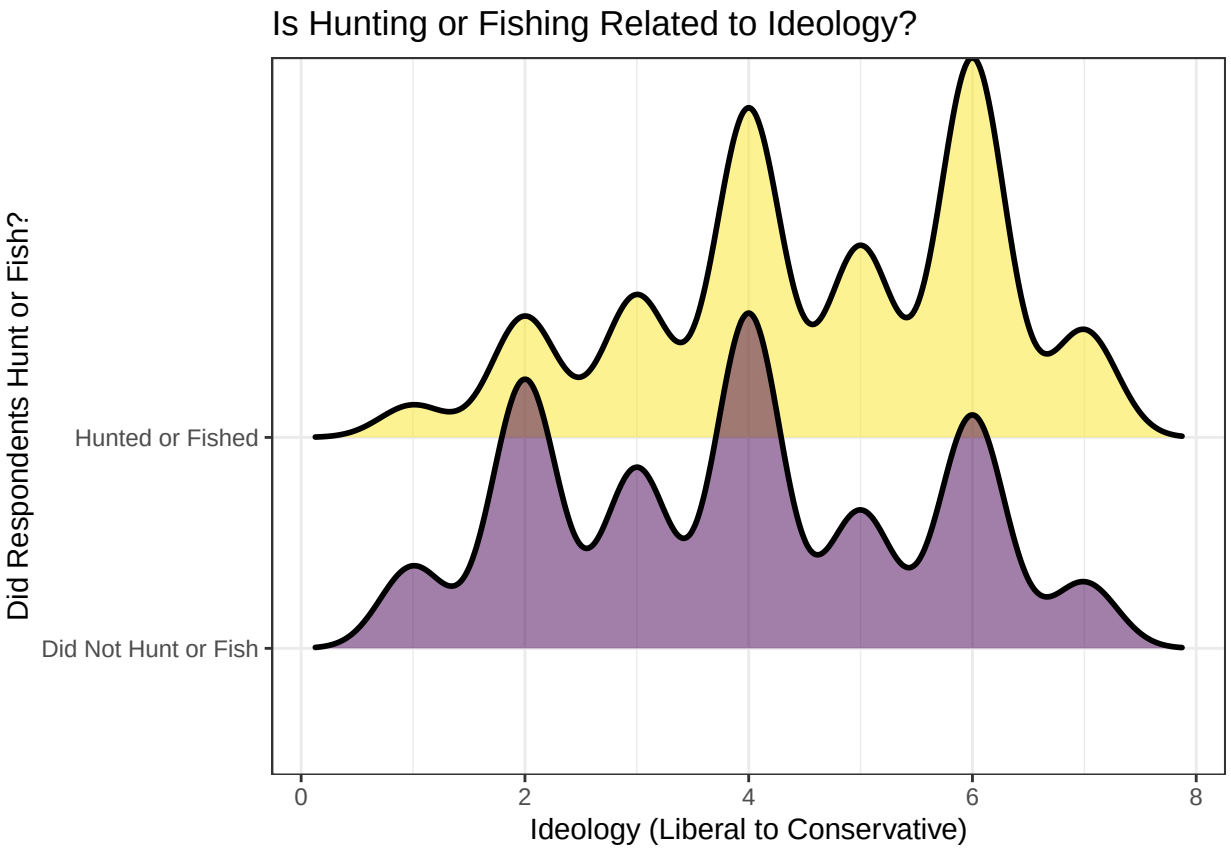
In general, people report feeling very warmly towards scientists - the data are heavily skewed left. While unimodal, the peak of the data occurs at 100 - the maximum possible score. Additionally, responses seem to be clustered at a few specific points (almost regardless of bin width). Though this is not in the question itself, we would be ill-advised to draw any conclusions from data this heavily skewed - it suggests that perhaps there was a flaw in the methodology from which this data set was created.

Exercise 5

```
anes %>%
  drop_na(hunt_fish) %>%
  mutate(hunted_fished = ifelse(hunt_fish == 0,
                                "Did Not Hunt or Fish", "Hunted or Fished")) %>%
  ggplot(., aes(x = hunted_fished, y = ideology, color = hunted_fished)) +
  geom_boxplot(lwd = 1) +
  labs(title = "Is Hunting or Fishing Related to Ideology?",
       x = "Did Respondents Hunt or Fish?",
       y = "Ideology (Liberal to Conservative") +
  theme_bw() +
  scale_color_viridis(discrete = TRUE, option = "D")+
  guides(color = "none")
```




```
anes %>%
  drop_na(hunt_fish) %>%
  mutate(hunted_fished = ifelse(hunt_fish == 0,
                                "Did Not Hunt or Fish", "Hunted or Fished")) %>%
  ggplot(., aes(x = ideology, y = hunted_fished, fill = hunted_fished)) +
  geom_density_ridges(alpha = 0.5, color = "black", lwd = 1) +
  labs(title = "Is Hunting or Fishing Related to Ideology?",
       x = "Ideology (Liberal to Conservative)",
       y = "Did Respondents Hunt or Fish?") +
  theme_bw() +
  scale_fill_viridis(discrete = TRUE, option = "D") +
  guides(fill = "none")
```



In the graph of the boxplots, it is easy to see the median ideology of each group, which is 4 or moderate for `Did Not Hunt or Fish` and 5 or slightly conservative for `Hunted or Fished`. It also shows the rest of the summary statistics, which suggest a wide distribution in ideology of people who did not hunt or fish but a much narrower distribution concentrated on the conservative side of the spectrum. However, there are no outliers, perhaps to be expected with such a narrow range of possible values.

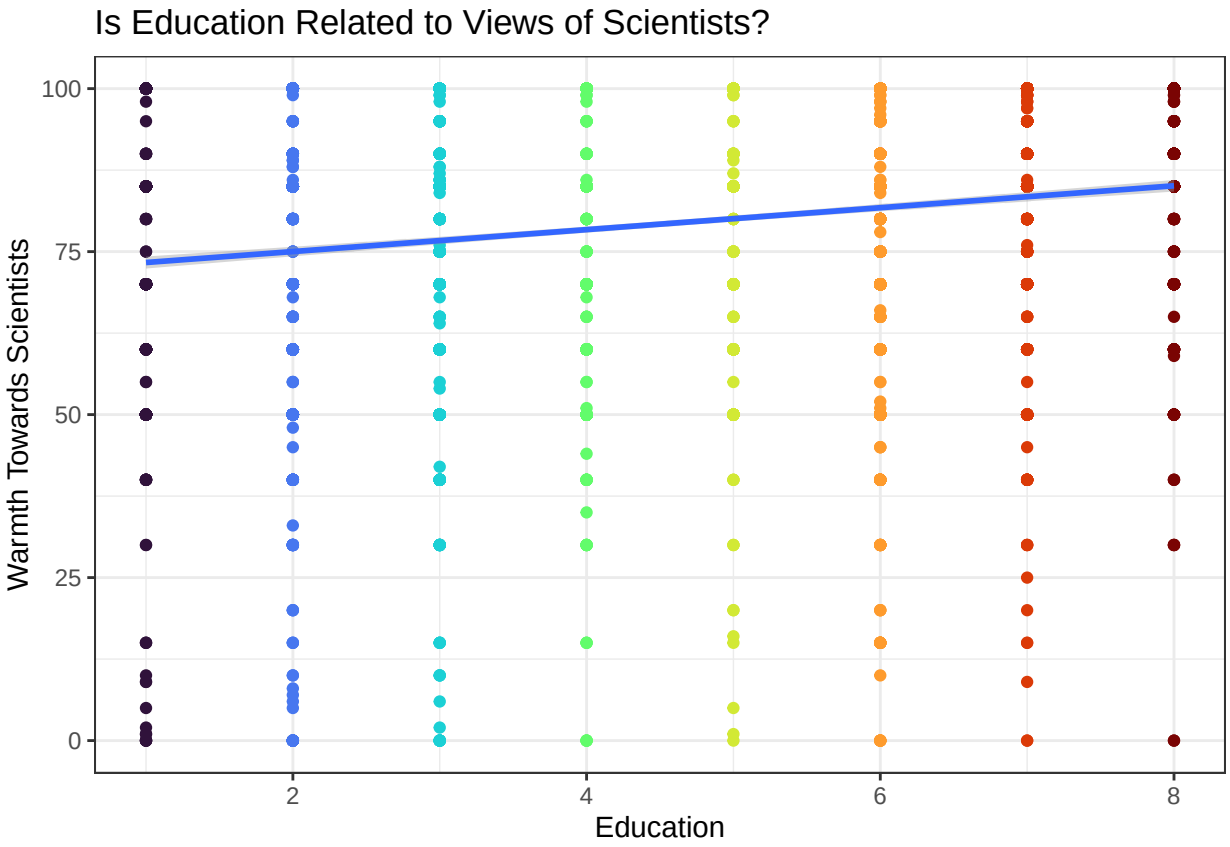
In the graph of the ridgelines, we do not get to see summary statistics in the same way. However, we do get to see individual data points. This shows the higher concentration of people who hunted or fished on the Conservative side of the spectrum, and reveals that the median value at 5 is actually a number chosen by relatively few respondents within that category. It also shows the more normal distribution of people who did not hunt or fish, who had surprisingly low rates of `3` and `5`. Unfortunately, this also shows some of the limitation of ridgeline plots - those drapes are entirely meaningless (in fact, they provide confusing data points that do not actually exist) and smoothing the curve to remove them would also eliminate any of the benefits from the ridgeline plot.

The boxplots give the summary statistics, which is helpful because they provide the clearest summary for the data's quartiles and thus make the two groups immediately comparable. However, what they do not give that the ridgeline plot does is individual data points, which remind us that the median can settle on a value that is not particularly common.

Exercise 6

```
ggplot(anes, aes(x = education, y = scientists)) +
  geom_point(aes(color = as.factor(education))) +
  scale_color_viridis(discrete = TRUE, option = "H") +
```

```
  guides(color = "none") +
  labs(title = "Is Education Related to Views of Scientists?",
       x = "Education", y = "Warmth Towards Scientists") +
  theme_bw() +
  geom_smooth(method = lm)
```

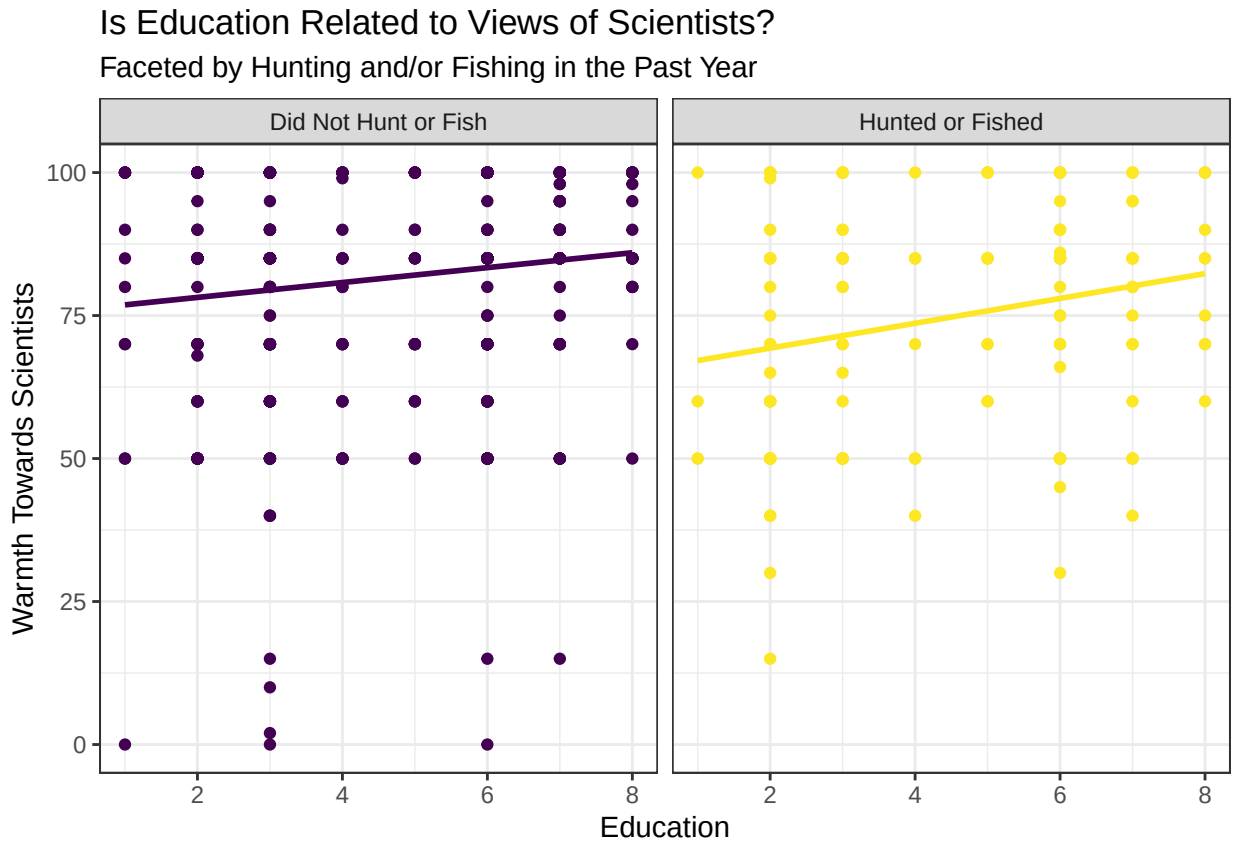


This is a very poor way to visualize data and isn't helpful in the slightest. When making scatterplots, we do so with the intention of seeing whether an increase in one variable is related to an increase (or decrease) in another. However, while this does the same thing, `Education` is not a continuous variable in this dataset. Instead, it is an ordinal categorical variable, which therefore presents in column-like structures on a scatterplot and is almost unreadable as a result. It also requires more specific labels across the x-axis, as the discrete points of the variable are functionally meaningless. Additionally, since the gaps in years of education or any other continuous variable are not consistent from one discrete marker to another in this dataset, there is no interpretable structure in the data and the regression line is misleading at best. Scatterplots with discrete data such as these are not useful.

Exercise 7

```
set.seed(18)
anes2 <- anes %>%
  sample_frac(.10)
facet.labs <- c("Hunted or Fished",
                "Did Not Hunt or Fish",
```

```
                "Unknown")
names(facet.labs) <- c("1", "0", NA)
ggplot(na.omit(anes2), aes(x = education, y = scientists,
                           color = as.factor(hunt_fish))) +
  geom_point() +
  scale_color_viridis(discrete = TRUE, option = "D") +
  guides(color = "none") +
  facet_grid(cols = vars(hunt_fish),
             labeller = labeller(hunt_fish = facet.labs)) +
  labs(title = "Is Education Related to Views of Scientists?",
       subtitle = "Faceted by Hunting and/or Fishing in the Past Year",
       x = "Education", y = "Warmth Towards Scientists") +
  theme_bw()+
  geom_smooth(method = lm, se = FALSE)
```



In these graphs, almost the same trend as before comes into play regardless of facet. Though using `Education`, a discrete variable, makes for very poor scatterplots, once again a slight linear correlation appears to exist between education and views of scientists. This is seen in both groups - those who did and did not hunt or fish in the past year - and while the correlation is the slightest bit stronger (and begins and ends at slightly higher intercepts) for the group that did not hunt or fish, this difference would not be significant if we ran statistical tests. Both groups start out at a high level of warmth towards scientists, and this increases slightly with education in the sample.