STAT 231: Problem Set 8B

Sean Wei

due by 5 PM on Friday, November 6

This homework assignment is designed to help you further ingest, practice, and expand upon the material covered in class over the past week(s). You are encouraged to work with other students, but all code and text must be written by you, and you must indicate below who you discussed the assignment with (if anyone).

Steps to proceed:

1. In RStudio, go to File > Open Project, navigate to the folder with the course-content repo, select the course-content project (course-content.Rproj), and click "Open"
2. Pull the course-content repo (e.g. using the blue-ish down arrow in the Git tab in upper right window)
3. Copy ps8B.Rmd from the course repo to your repo (see page 6 of the GitHub Classroom Guide for Stat231 if needed)
4. Close the course-content repo project in RStudio
5. Open YOUR repo project in RStudio
6. In the ps8B.Rmd file in YOUR repo, replace "YOUR NAME HERE" with your name
7. Add in your responses, committing and pushing to YOUR repo in appropriate places along the way
8. Run "Knit PDF"
9. Upload the pdf to Gradescope. Don't forget to select which of your pages are associated with each problem. *You will not get credit for work on unassigned pages (e.g., if you only selected the first page but your solution spans two pages, you would lose points for any part on the second page that the grader can't see).*

If you discussed this assignment with any of your peers, please list who here:

ANSWER:

1. Mapping spatial data

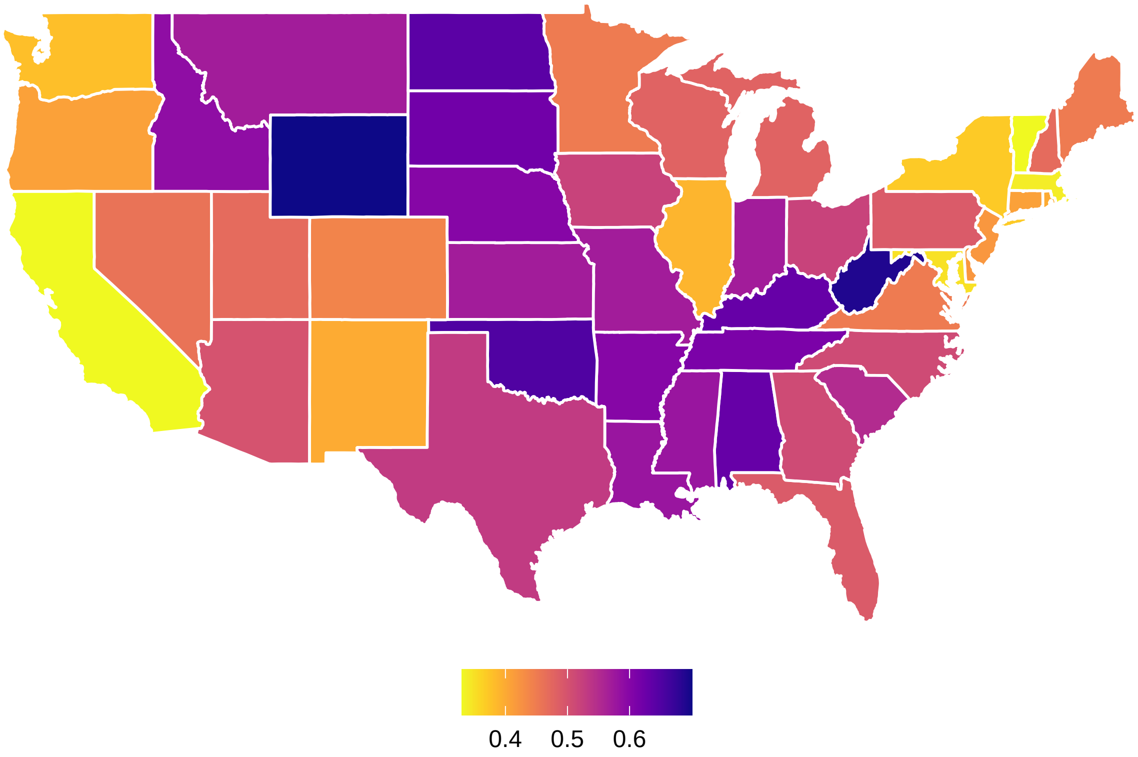
Reproduce the map you created for Lab08-spatial (and finish it if you didn't in class). In 2-4 sentences, interpret the visualization. What stands out as the central message?

NOTE: you do NOT need to say what colors are representing what feature (e.g, NOT: "In this map, I've colored the countries by GDP, with green representing low values and red representing high values") – this is obvious to the viewer, assuming there's an appropriate legend and title. Rather, what *information* do you extract from the visualization? (e.g., "From the choropleth below, we can see that the percent change in GDP per capita between 1957-2007 varies greatly across countries in Central America. In particular, Panama and Costa Rica stand out as having GDPs per capita that increased by over 200% across those 50 years. In contrast, Nicaragua's GDP per capita decreased by a small percentage during that same time span.")

> ANSWER: The cloropleth below conveys the proportion of voters who voted for Trump during the 2016 election. We can see that for the most part, the visualization mirrors red and blue states - for example, California, New York, and Massachussets (traditional blue states) had very few people voting for Trump overall, while Wyoming, West Virginia, and Oklahoma (traditional red states) had a very large number of people voting for Trump. From this visualization, we can also see that Trump barely won a lot of the swing states - specifically Wisconsin, Michigan, and Pennsylvania, which resulted in his election as president.

```r
library(fivethirtyeight)
data(state)
hate_crimes <- fivethirtyeight::hate_crimes
usa_states <- map_data(map = "state", region = ".")

state_info <- data.frame(state_full = tolower(state.name),
                         state = state.abb,
                         region = state.region)

hate_crimes <- hate_crimes %>%
  filter(state_abbrev != "DC") %>%
  select(-c(state)) %>%
  rename(state = state_abbrev)

map <- hate_crimes %>%
  left_join(state_info, by = "state") %>%
  right_join(usa_states, by = c("state_full" = "region"))

ggplot(map, aes(x = long, y = lat, group = group, fill = share_vote_trump)) +
  geom_polygon(color = "white") +
  theme_void() +
  coord_fixed(ratio = 1.3) +
  labs(fill = "", title = "Proportion of Trump Votes in U.S in 2016") +
  theme(legend.position="bottom") +
  scale_fill_viridis(option = "plasma", direction = -1)
```

Proportion of Trump Votes in U.S in 2016



2. Mapping spatial data at a different level

Create a map at the world, country, or county level based on the choices provided in lab08-spatial, that is at a DIFFERENT level than the map you created for the lab (and included above). For instance, if you created a map of US counties for the lab, then choose a country or world map to create here.

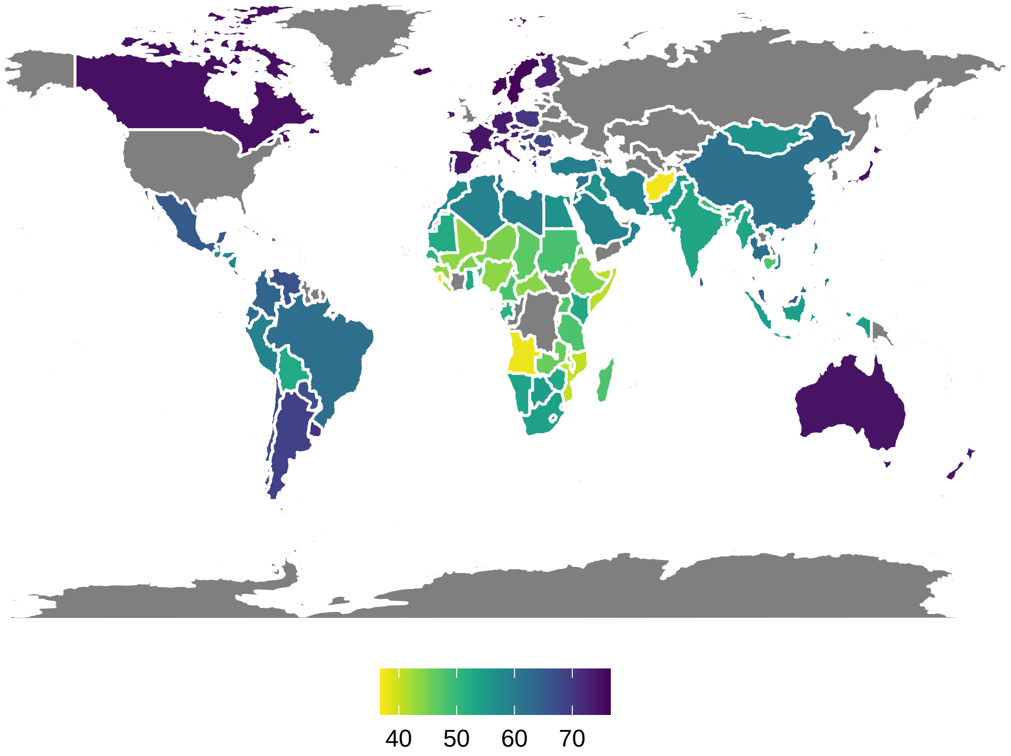
Note: While I recommend using one of the datasets provided in the lab so you don't spend a lot of time searching for data, you are not strictly required to use one of those datasets. You could, for instance, create a static map that might be relevant to your project (so long as it's at a different level than your map above).

Describe one challenge you encountered (if any) while creating this map.

> ANSWER: I created a map at the world level that looks at average world life expectancy from the years 1952-2007. As you can see below, it is clear that Canada, Australia, and Europe had the highest average life expectancies during this time period, while countries in South America and Africa had lower average life expectancies. I did not encounter any challenges while creating this map.

```r
data(gapminder)
avg_life_exp <- gapminder %>%
  group_by(country) %>%
  summarise(avgLifeExp = mean(lifeExp))

world_map <- map_data(map = "world", region = ".")

map2 <- avg_life_exp %>%
  right_join(world_map, by = c("country" = "region"))

ggplot(map2, aes(x = long, y = lat, group = group, fill = avgLifeExp)) +
  geom_polygon(color = "white") +
  theme_void() +
  coord_fixed(ratio = 1.3) +
  labs(fill = "",
       title = "Average World Life Expectancy from 1952-2007",
       caption = "Countries in gray do not have data") +
  theme(legend.position="bottom") +
  scale_fill_viridis(option = "viridis", direction = -1)
```

Average World Life Expectancy from 1952−2007



Countries in gray do not have data

3. Ethics follow-up

(a) Thinking about the discussion you had with the first group you were with during class last Thursday (focused on either "Predicting Policing & Recidivism" or "Predicting Financial Risk"), did your perspective on, or understanding of, any of the questions shift? If so, please describe. If not, was there anything you found surprising in the resources or your first group discussion?

ANSWER: No, our group on "Predicting Policing & Recidivism" agreed on the existence of algorithmic bias. I personally found it surprising in Buolamwini's facial recognition example that the algorithm didn't consider black indviduals as faces, especially when the example is so straightforward and simple (expanding the training set to include more diverse individuals). We ultimately came to the conclusion that there could potentially be an underrepresentation of black people in the field, and if training sets are based on people you work with, the algorithm will mirror these social inequalities.

(b) Thinking about the discussion you had with the second group you were with during class last Thursday (focused on considering the use of algorithms in the college admissions processs), did your perspective on, or understanding of, the use of algorithms in these contexts shift? If not, was there anything you found surprising in the resources or your second group discussion?

ANSWER: In our second discussion, I eventually came to the conclusion that although algorithms can make the admissions process quicker, I would prefer a human to make the decision in my college admission process. The main point that brought me to this conclusion was the idea of how jobs and schools are looking to hire/admit people who have more personality than perfect grades. We discussed how we don't think that algorithms are advanced enough yet to determine personality from a candidate's previous work experiences, life experiences, or even college essay.