DS Homework #3: 538 Assignment

Adam Haertter & Brennan Mulligan

538

For those of you who don't know, 538 is a website that focuses on opinion poll analysis, politics, economics, and sports blogging. It was created and is currently run by Nate Silver, who has become famous for his writing and work on the site. He has a famous book called "The Signal and the Noise", among other works.

`fivethirtyeight` package

A nice feature of many articles articles on the site is that the data they use is freely available for people to use and reproduce their analysis. Some professors and students at Smith College have compiled this data into an R package called `fivethirtyeight`. Even nicer, they link to the article page as well!

The Assignment

For this assignment, you are going to work in groups of two or three to recreate a production-level plot an article of your choice!

- Assignment Structure:
  - Choose one or two other classmates to work with
  - Browse the offerings from the `fivethirtyeight` package. You can run `data(package = "fivethirtyeight")` to get all of the datasets.
  - Choose an article that has a reasonable plot that seems possible to replicate (or at least get close!)
  - Contact me as a team and let me know which data you plan to use. I will vet if it's at the appropriate level (too easy or too hard).
  - Make the plot and comment on how you plot is similar / how it is different.
  - Due Date: March 18th, 2022
- Grading:

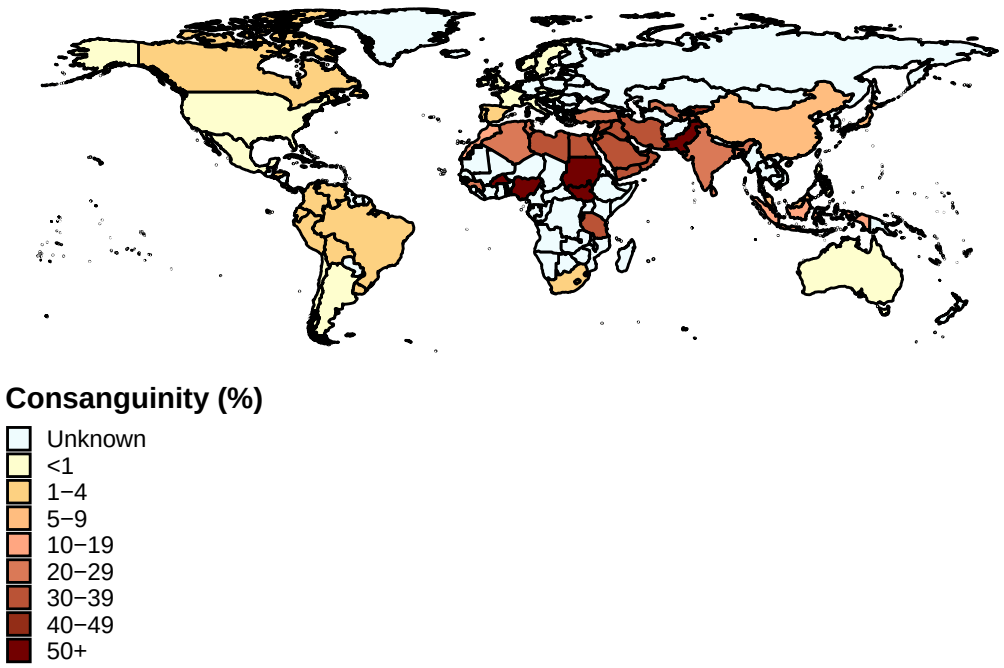
- A: Almost complete recreation of the plot! Only very minor differences
- B: Missing some of the finer details but has a strong resemblance to the original plot
- C: Recreates the main substance of the plot but is missing all finer details
- D and below: Fails to recreate any plot

Code

```
#install.packages("tidyverse")
#install.packages("fivethirtyeight")
library("tidyverse")
library("fivethirtyeight")
library("dplyr")
library("magrittr")
require("maps")
require("viridis")
```

```
# code goes here!
# data(package = "fivethirtyeight")
cousin <- fivethirtyeight::cousin_marriage
world_map <- map_data("world")

# Data Manipulation
world_map %<>% filter(region != "Antarctica")

cousin %<>% mutate(region = country) %>% .[, 2:3]
cousin[64, "region"] <- "UK"
cousin[67, "region"] <- "Netherlands"
cousin[68, "region"] <- "USA"

cousin[nrow(cousin) + 1,] <- list(cousin$percent[cousin$region == "Sudan"], "South Sudan")

make_pct <- function(x) {
  if(is.na(x))
      return("Unknown")
  if(x < 1)
      return("<1")
  else if(x >= 1 && x < 5)
      return("1-4")
  else if(x >= 5 && x < 10)
```

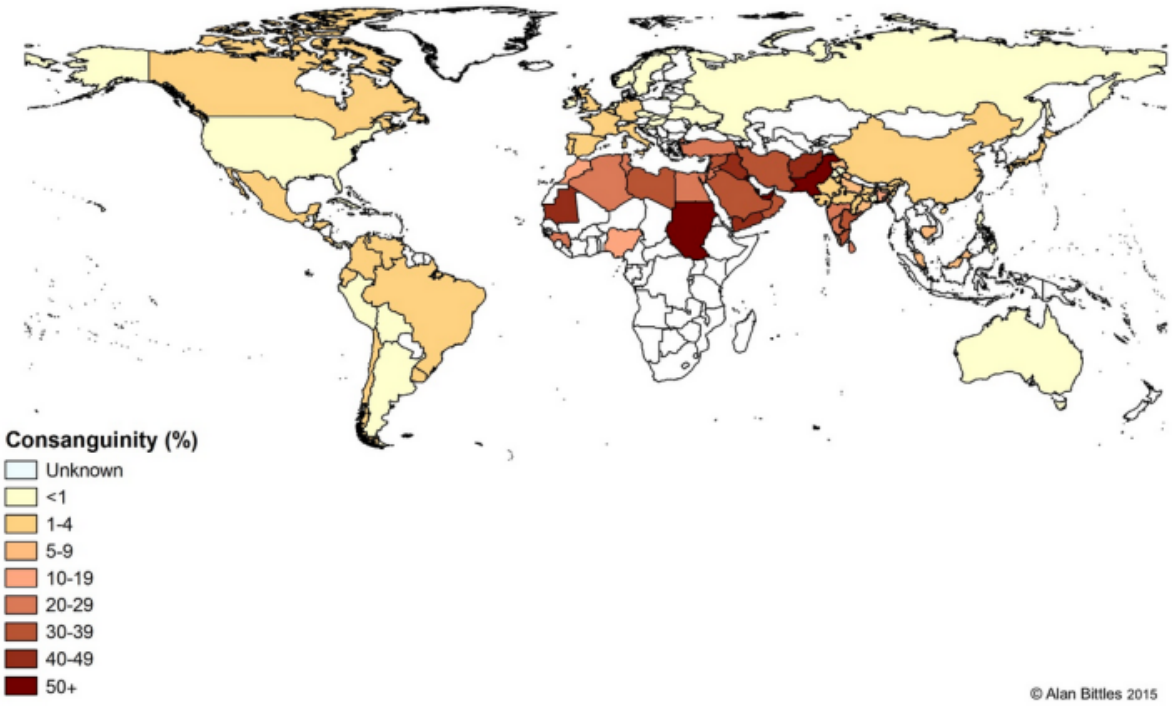
```
      return("5-9")
  else if(x >= 10 && x < 20)
      return("10-19")
  else if(x >= 20 && x < 30)
      return("20-29")
  else if(x >= 30 && x < 40)
      return("30-39")
  else if(x >= 40 && x < 50)
      return("40-49")
  else if(x >= 50)
      return("50+")
  else
      return("Unknown")
}

cousin$percent_set <- sapply(cousin$percent, make_pct)

world_map %<>% left_join(cousin, by = "region")
world_map["percent_set"][is.na(world_map["percent_set"])] <- "Unknown"

cat_order <- c("Unknown", "<1", "1-4", "5-9", "10-19", "20-29", "30-39", "40-49", "50+")
world_map$percent_set <- factor(world_map$percent_set, levels = cat_order)
world_map <- world_map[order(world_map$percent_set), ]

#world_map %>% group_by(region) %>% summarize(mean = mean(long)) # Use to see countries

# Graphing
theme_set(theme_void())

colorscale = scale_fill_manual(values=c( "#EFFCFE", "#FEFECE", "#FDD181",
    "#FFBD7F", "#FFA681", "#DA7957", "#BA5336", "#932D17", "#720000"))

ggplot(world_map, aes(x = long, y = lat, group = group)) +
  geom_polygon(aes(fill=percent_set), colour = "black") + colorscale + theme(legend.positi
```



Comment: Our graph is fairly close to the original. We had to learn a lot about formatting the legend to make it like the original and we also combined Sudan and South Sudan to be more accurate. We discovered that many countries that had data in the original graphic did not have data in the dataset causing some differences between the two. Also, the original plot has India split up into multiple subregions which is odd considering no other country is split the same way. The data for the subregions was also not provided in the set so it was impossible to replicate.

Despite the lack of data in some areas, we can see a trend form where some regions have a higher rate of consanguinity than others. For example, it is more common in areas such as North Africa and the Middle East in comparison to places such as the West Hemisphere. It would be interesting to see how this data has changed over time as the article claims that it was gathered in 2001. That means that this is over two decades old and has possibly seen changes since then. Our hypothesis is that this trend has gone down overall.

We recreated the original plot to the best of our ability. However, the fundamental differences between the data in the dataset and the data used to create the original graph did prevent us from making it wholly accurate (Ex. Russia, Mauritania missing from the data from 538, data from South Africa included in the dataset but not the graph) . This is a little disappointing, but we thought it was best not to make up data or alter the original data in large ways to match these changes. We are still very proud of the plot we created. The original plot can be seen below for reference:


Consanguinity (%)
Unknown
<1
1-4
5-9
10-19
20-29
30-39
40-49
50+
© Alan Bittles 2015