SMI105, week 9 solutions

Week 9’s tasks were all designed to really test your ability to apply the more complex data tidying we did in the workshop. Some of these tasks were quite difficult, but if you tried them you will hopefully have gotten a good start on practicing the skills you’ll likely need to use to complete your final assessment for SMI105.

The first task was:

> Draw a graph of the relationship between the rule of law and government effectiveness, faceted by year. Make points more faint if the total number of sources across both measures is lower, and make points more opaque if the total number of sources across both measures is higher.

To complete this task, we needed to add data from the fourth and sixth sheet in our spreadsheet, as these related to Government Effectiveness and Rule of Law respectively. We needed two things from each of the datasets: the first is the estimates for both of the variables and the second is the number of sources that those estimates are based on (which was called NumSrc) in the dataset.

I’m going to reuse the code that we wrote several times to grab each of these variables. We could use the custom function, but I think using functions can be quite overwhelming for people! So let’s do this the slightly longwinded way:

```r
library(tidyverse)
library(readxl)


first_years <- seq(1996, 2002, 2)
second_years <- seq(2003, 2019, 1)
years <- c(first_years, second_years)

wb_goveff <- read_excel("wgidataset_2020.xlsx",
                       sheet = 4,
                       range = cell_rows(15:229),
                       na = "#N/A") %>%
              select(country = `Country/Territory`,
                     code = Code,
                     contains("Estimate")
                     ) %>%
              set_names(c("country", "code", years)) %>%
              pivot_longer(
                cols = c(-country, -code),
                names_to = "year",
                values_to = "gov_effectiveness"
              ) %>%
              mutate(
                year = as.numeric(year)
              )
```

```
wb_goveff_numsrc <- read_excel("wgidataset_2020.xlsx",
                       sheet = 4,
                       range = cell_rows(15:229),
                       na = "#N/A") %>%
              select(country = `Country/Territory`,
                     code = Code,
                     contains("NumSrc")
                     ) %>%
              set_names(c("country", "code", years)) %>%
              pivot_longer(
                cols = c(-country, -code),
                names_to = "year",
                values_to = "gov_effectiveness_nsrc"
              ) %>%
              mutate(
                year = as.numeric(year)
              )

wb_law <- read_excel("wgidataset_2020.xlsx",
                       sheet = 6,
                       range = cell_rows(15:229),
                       na = "#N/A") %>%
              select(country = `Country/Territory`,
                     code = Code,
                     contains("Estimate")
                     ) %>%
              set_names(c("country", "code", years)) %>%
              pivot_longer(
                cols = c(-country, -code),
                names_to = "year",
                values_to = "gov_law"
              ) %>%
              mutate(
                year = as.numeric(year)
              )

wb_law_numsrc <- read_excel("wgidataset_2020.xlsx",
                       sheet = 6,
                       range = cell_rows(15:229),
                       na = "#N/A") %>%
              select(country = `Country/Territory`,
                     code = Code,
                     contains("NumSrc")
                     ) %>%
              set_names(c("country", "code", years)) %>%
              pivot_longer(
                cols = c(-country, -code),
                names_to = "year",
                values_to = "gov_law_nsrc"
              ) %>%
              mutate(
                year = as.numeric(year)
```

```
               )
```

Okay, lots and lots of code but really all I was doing was copy and pasting and changing a few small things. So not really a big deal! The next stage would be to join the datasets together:

```
wb <- left_join(wb_goveff, wb_goveff_numsrc, by = c("country", "code", "year")) %
>%
  left_join(wb_law, by = c("country", "code", "year")) %>%
  left_join(wb_law_numsrc, by = c("country", "code", "year"))
```

Before we can create our plot, there was one more thing we need to do: we need a variable that can be used to measure the number of sources across both variables. At the moment we have the number of sources for each. There's a few different ways we might do this: we could take the average number of sources across both variables, or just the total, or something else (e.g. make it a percentage of the maxmum). The task itself asks for the total number of sources, so I'll just add them together.

I'm just going to create a new variable that's the average of the two columns:

```
wb <- wb %>%
  mutate(
    nsrc = gov_effectiveness_nsrc + gov_law_nsrc
  )
```

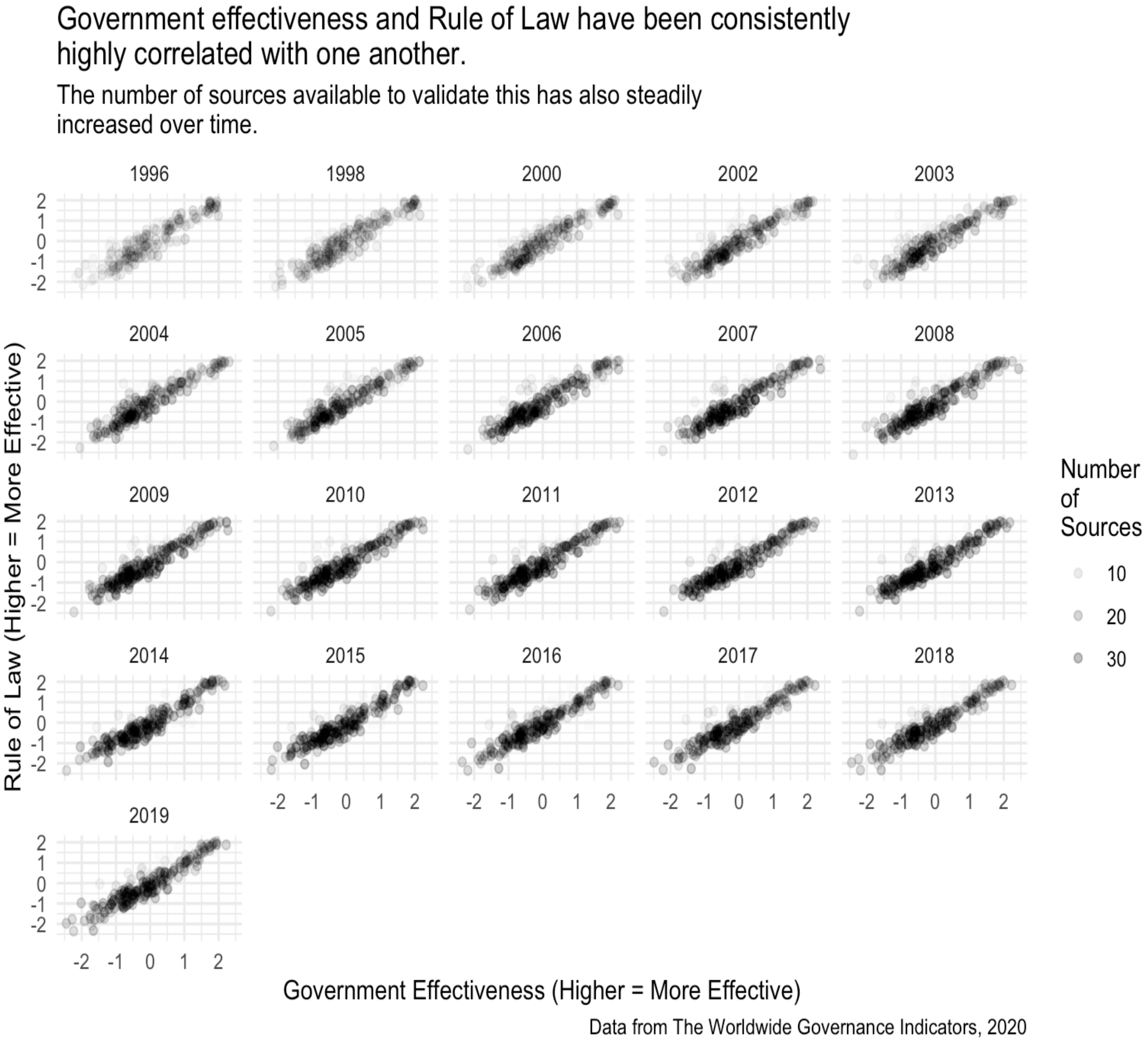
Okay, have we checked off everything we need?

- relationship between the **rule of law** — check
- and **government effectiveness** — check
- faceted by **year** — check
- Make points more faint if the **total number of sources** across both measures is lower — check

We're ready to make our plot.

Both the rule of law variable and the government effectiveness variable are continuous measures, so a scatterplot would be appropriate (week 4).

```
wb %>%
  ggplot() +
  aes(x = gov_effectiveness, y = gov_law, alpha = nsrc) +
  geom_point() +
  facet_wrap(~year) +
  scale_alpha_continuous(range = c(0.01, 0.3)) +
  theme_minimal() +
  xlab("Government Effectiveness (Higher = More Effective)") +
  ylab("Rule of Law (Higher = More Effective)") +
  labs(alpha = "Number\nof\nSources",
       caption = "Data from The Worldwide Governance Indicators, 2020") +
  ggtitle("Government effectiveness and Rule of Law have been consistently\nhighly
correlated with one another.",
          subtitle = "The number of sources available to validate this has also st
eadily\nincreased over time.")
```



Government effectiveness and Rule of Law have been consistently highly correlated with one another.

The number of sources available to validate this has also steadily increased over time.

Data from The Worldwide Governance Indicators, 2020

Okay, anything new here? Well, one thing that's new is I've limited my range for the transparency between 0.01 and 0.3, meaning even the points with the most sources will still be quite transparent. I did this because I noticed there was quite a lot of "overplotting", that is, points that all end up plotted on top of one another and can hide the pattern. By limiting the highest transparency to 0.3, it helps me to see the pattern a little better.

If I were interpreting this plot I might say:

"Government effectiveness and the rule of law have been very highly correlated with one another in every year that the data have been published, suggesting that countries with more effective governments tend to also have more effective legal systems. The number of sources that are able to inform the measurement of both of these concepts has increased substantially over the last three decades. There is some suggestion that the majority of countries are clustered around average to low levels of both government effectiveness and rule of law, but there has also been a consistent apparent cluster of countries with high government effectiveness and high rule of law, and another cluster with very high government effectiveness and very high rule of law. There were some years where it appears the rule of law and government effectiveness were slightly less strongly related to one another, namely 2003, 2004, 2007, 2008, and 2009."

The second task for the week was to try finding a dataset that you would like to make a graph from and seeing how easy or difficult it would be to tidy it in a way that means we could make a graph. I've incorporated feedback from that below.

Feedback

- The big stumbling block that people keep coming across is not being able to read their data into R because R cannot find the file. This is why setting up R project files is so important. Remember that you should set up an R project directory (folder), **move the data you download into the R project folder**, and then try reading it. Remember each time you close and reopen R, to either do so from the "blue box" of the project you want to work on, or by choosing your project from the menu in the top right.
- There was quite an interesting thing that happened where someone discovered that `readxl` will read the hidden sheets in an excel file. So while it looked like sheet 3 has the data that you might need, you end up getting something that looks totally different! One way around this is to right click on the sheet tabs in Excel and then click "Unhide", then make everything visible. That way you can see which sheet the one you want *really* is. An alternative is to use the sheet's name, in quotation marks (e.g. "Sheet1") instead of using the number that it is in your read_xlsx() function.
- Remember to double-check the question: some people just made all of their points transparent rather than making the transparency based on the number of sources.