Problem set 1

Siwei Dai

October 3, 2021

NOTE: Start with the file ***ps1_2021.Rmd*** *(available from the github repository at https://github.com/UChicago-pol-methods/IntroQSS-F21/tree/main/assignments). Save that file locally, open it with RStudio, and modify it to include your answers. To produce a pdf for submission, "knit" the file by clicking on the* ***Knit*** *button. Submit both the Rmd file and the knitted PDF via Canvas.*

Question 1: US presidential election results

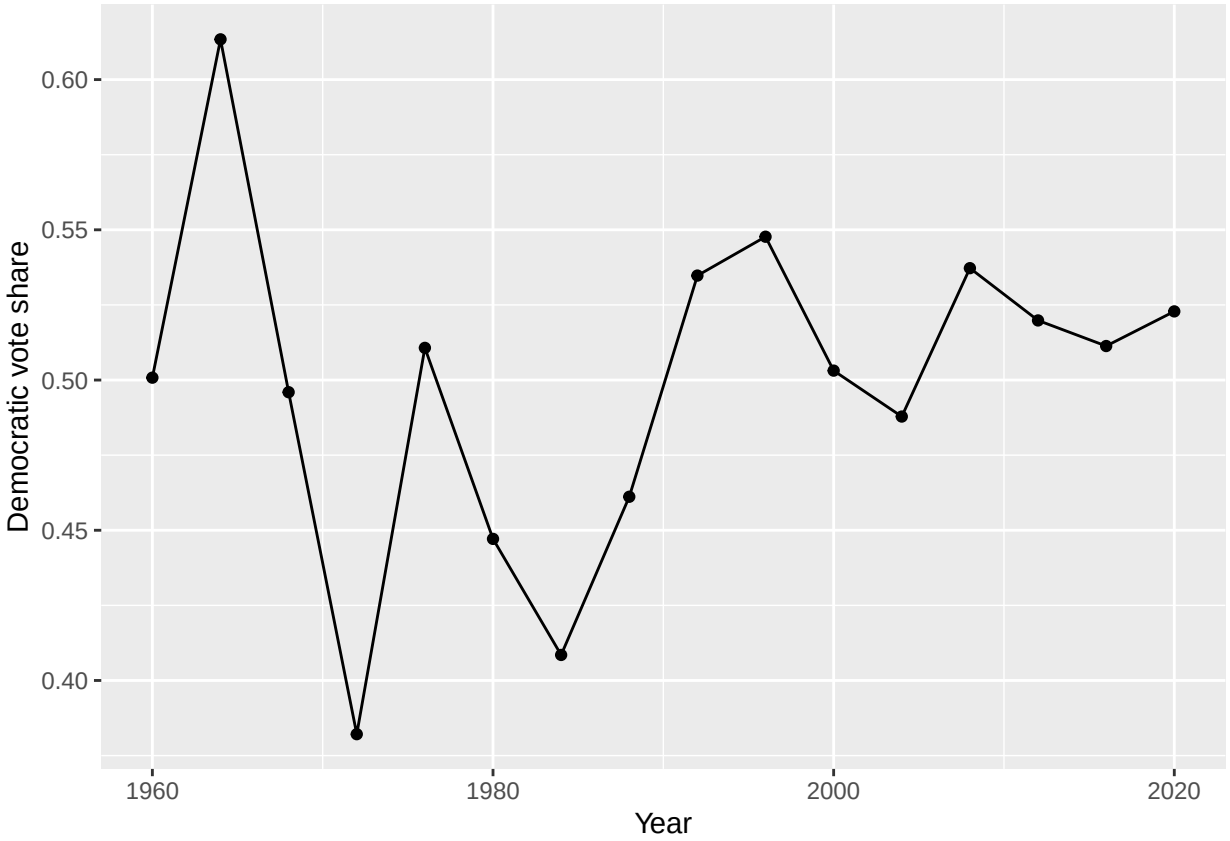
This week we will load data directly from the course github repository. If you have loaded the `tidyverse` library and you have an internet connection, the next code chunk will load the data. Don't worry if you don't understand this code yet.

```
data_path <- "https://raw.githubusercontent.com/UChicago-pol-methods/IntroQSS-F21/main/data/"
df <- read_csv(str_c(data_path, "yearly_county_pres_results_wide.csv"))
```

The object `df` is a dataset that includes `dem_vote_share` (the share of the two-party vote won by the Democratic candidate) and `dem_county_share` (the share of counties won by the Democratic candidate) in each U.S. presidential election since 1960.

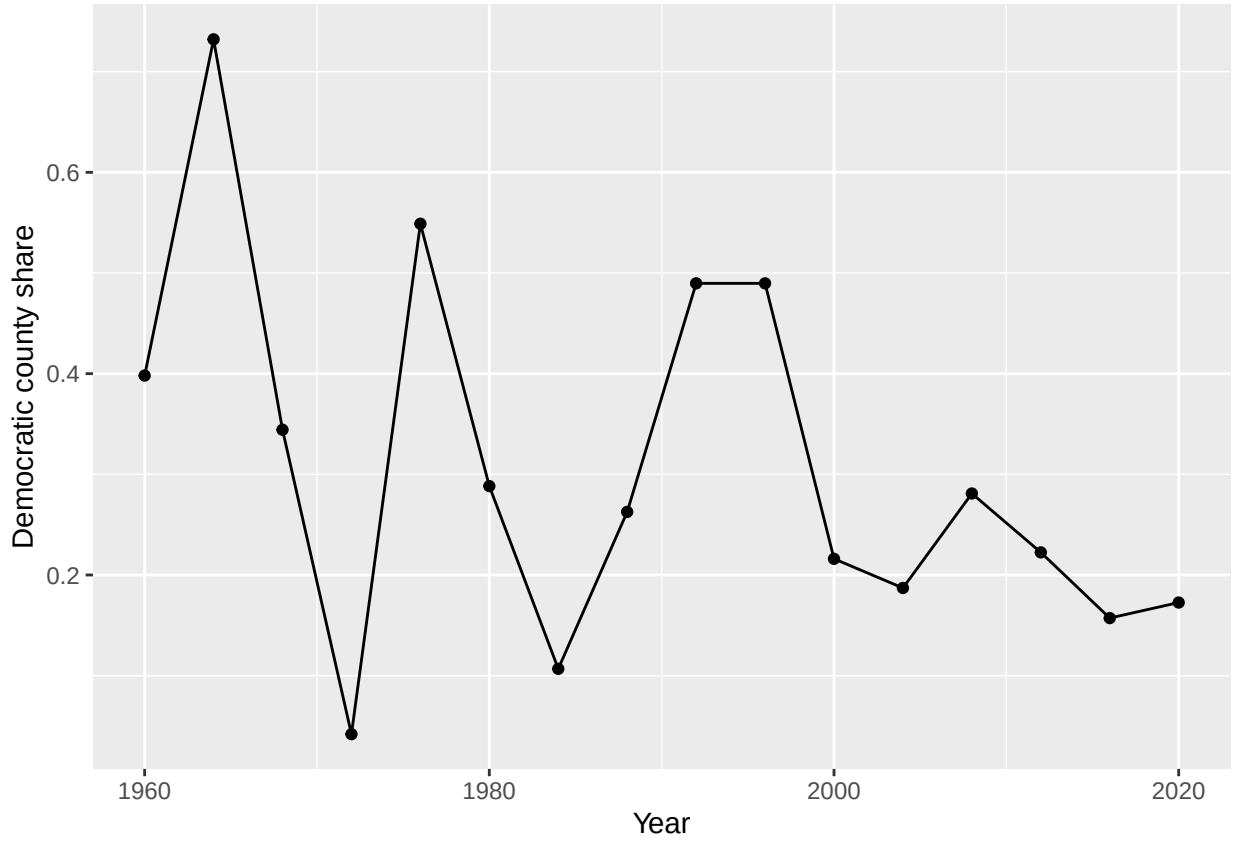
1a) Using this data, make a plot showing the Democratic vote share (vertical axis) in each year (horizontal axis). Draw a point for each year and connect them with a line.

```
plot1a <- ggplot(data = df,
                 mapping = aes(x = year, y = dem_vote_share)) +
          geom_point() +
          geom_line() +
          labs (x = 'Year', y = 'Democratic vote share')
plot1a
```



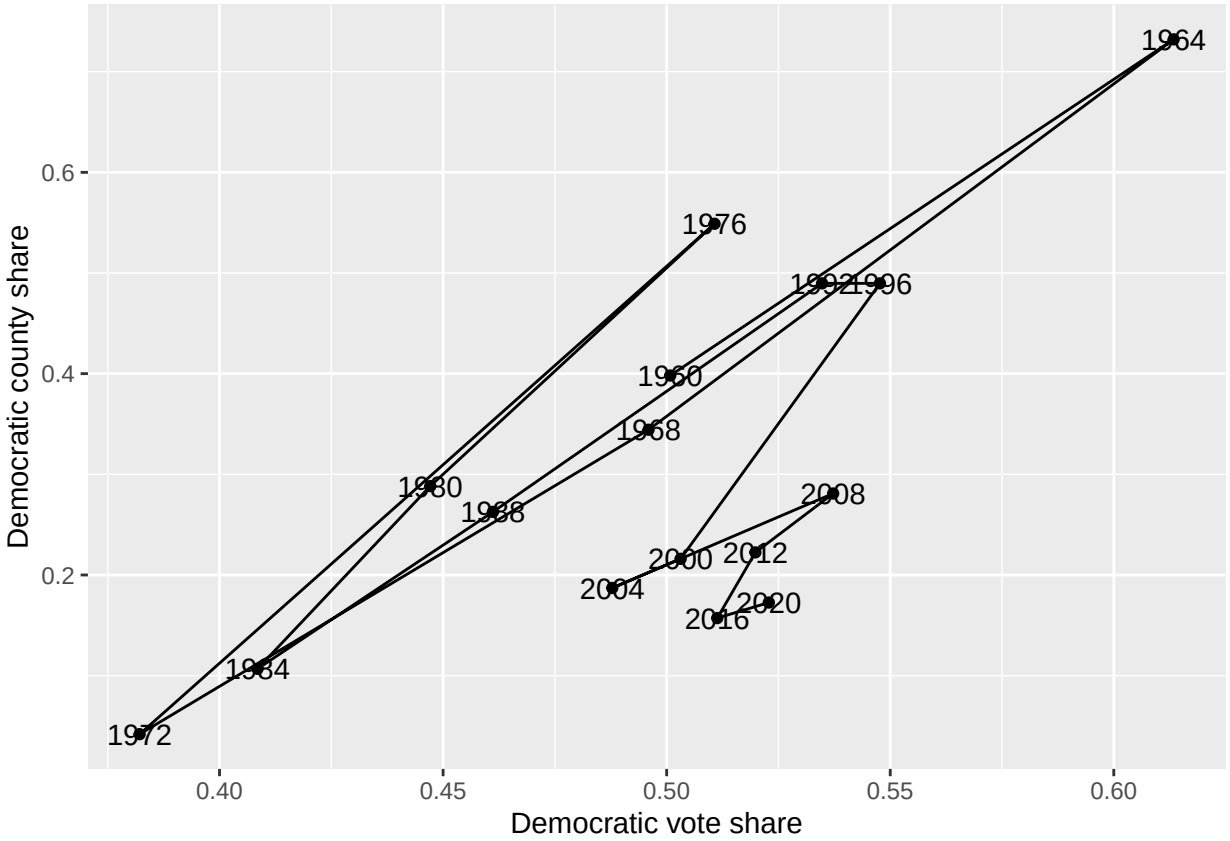
1b) Make a plot showing the Democratic county share (vertical axis) in each year (horizontal axis).

```
plot1b <- ggplot(data = df,
                 mapping = aes(x = year, y = dem_county_share)) +
          geom_point() +
          geom_line() +
          labs (x = 'Year', y = 'Democratic county share')
plot1b
```



1c) Now make a plot showing Democratic county share (vertical axis) and Democratic vote share (horizontal axis), again connecting the points with a line. (Hint: use `geom_path()`.) Label each point with the corresponding year. (Hint: use `geom_text()`.)

```
## replace this with your 1c plot code
plot1c <- ggplot(data = df,
                 mapping = aes(x = dem_vote_share, y = dem_county_share)) +
          geom_point() +
          geom_path() +
          labs (x = 'Democratic vote share', y = 'Democratic county share') +
          geom_text(aes(label = year))
plot1c
```

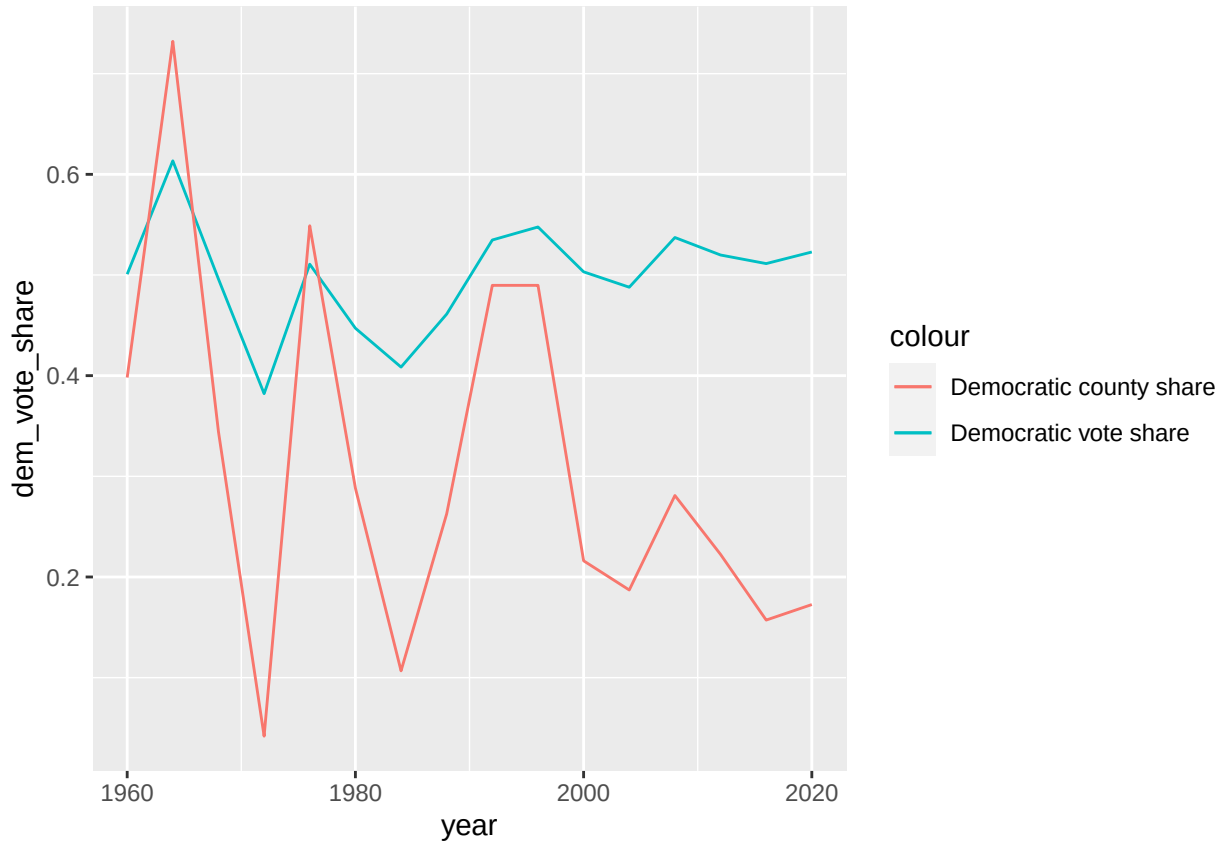


Now load a different dataset, which is the same data organized differently:

```
df2 <- read_csv(str_c(data_path, "yearly_county_pres_results_long.csv"))
```

1d) Make a plot showing both the Democratic vote share and Democratic county share (vertical axis) in each year (horizontal axis), with a different color for each series. Your figure should include a legend.

```
plot1d <- ggplot (data = df, mapping = aes(x = year)) +
          geom_line(mapping = aes(y = dem_vote_share, color = 'Democratic vote share')) +
          geom_line(mapping = aes(y = dem_county_share, color = 'Democratic county share'))
plot1d
```



1e) Which of these two plots do you prefer, and why?
I prefer the 1d plot as it shows the variance in which democratic county vote share and vote share changes over time.

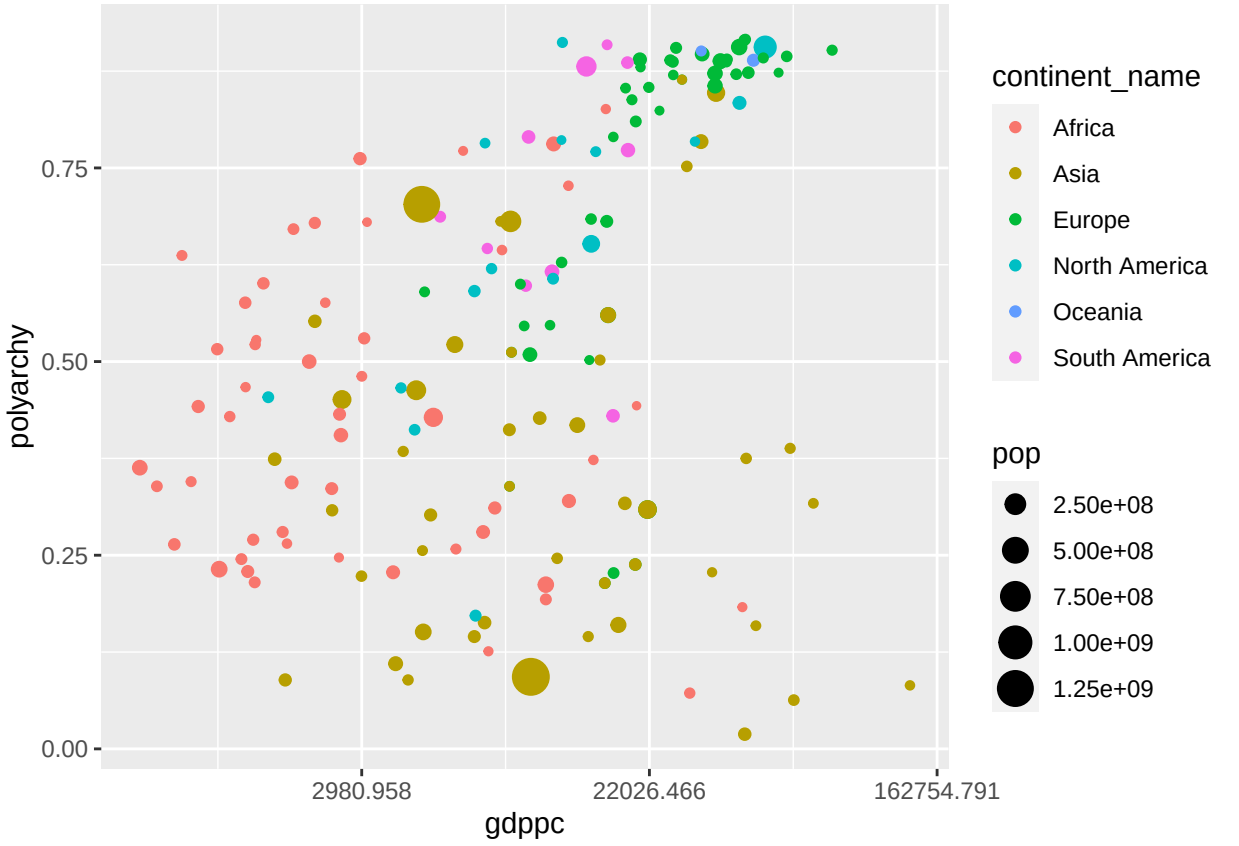
Question 2: democracy and GDP

The code chunk below loads an extract from the V-Dem dataset (https://www.v-dem.net/). Variables include `country_name` and `continent_name` (self-explanatory), `polyarchy` (V-Dem's measure of democracy), `pop` (World Bank measure of population), and `gdppc` (GDP per capita), all from 2010. The full dataset (available in the `vdemdata` `R` package) contains many more variables and years.

```
vd <- read_csv(str_c(data_path, "vdem_2010_extract.csv"))
```

2a) Make a scatterplot of the V-Dem polyarchy score (vertical axis) against GDP per capita (horizontal axis). Make the color of the dots reflect the continent, and the size reflect the population. Show the horizontal axis on the log scale.

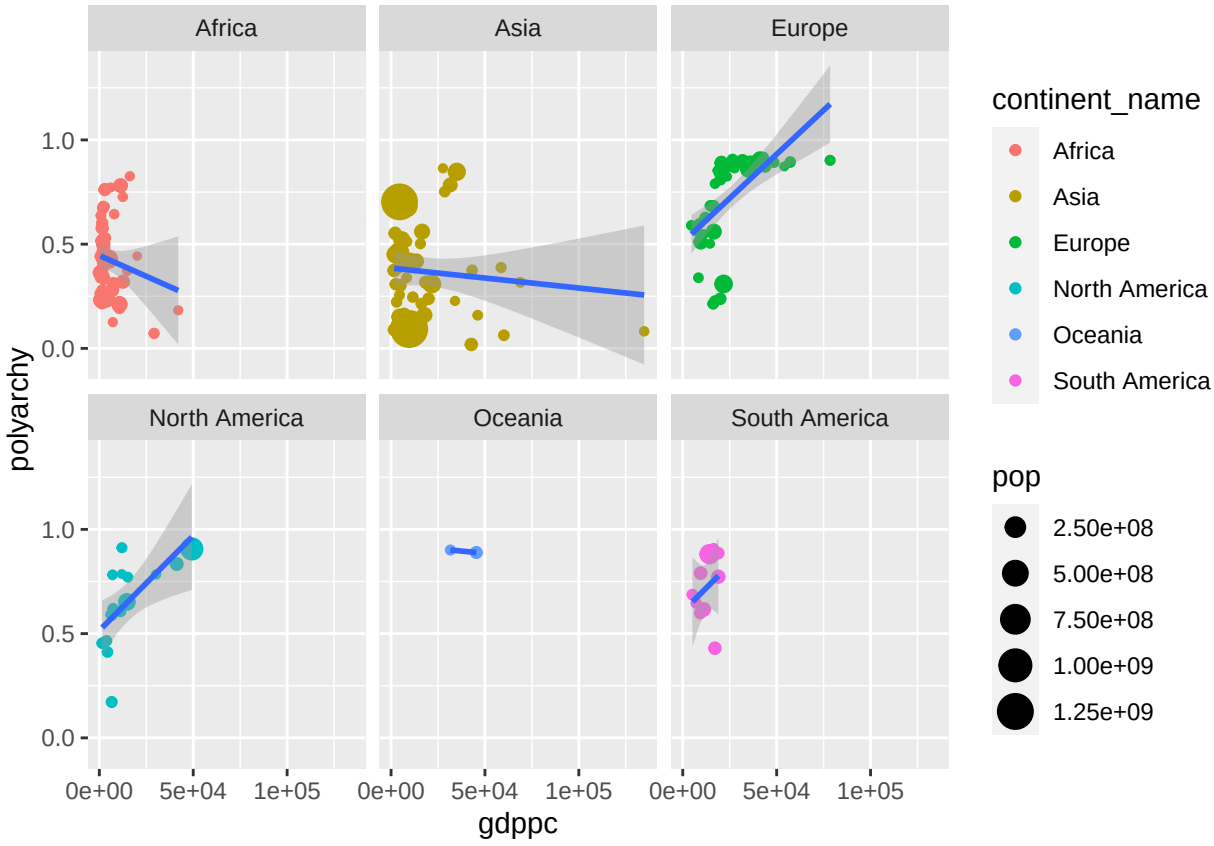
```
plot2a <- ggplot(data = vd, mapping = aes(x = gdppc, y = polyarchy)) +
        geom_point(mapping = aes(color = continent_name, size = pop)) +
        scale_x_continuous(trans = 'log')
plot2a
```



2b) Now make the same figure faceted by continent. Add a linear regression line (use `geom_smooth(method = lm)`). How does the relationship between GDP per capita and democracy differ across continents?

```
plot2b <- ggplot(data = vd,
                 mapping = aes(x = gdppc, y = polyarchy)) +
        geom_point(mapping = aes(size = pop, color = continent_name)) +
        geom_smooth(method = 'lm') +
        facet_wrap(facets = vars(continent_name))
plot2b
```

```
## `geom_smooth()` using formula 'y ~ x'
```



In Asia and in Africa, democracy and GDP per capita have a negative correlation whereas in Europe, North America and South America, they have a positive correlation. The correlation is hard to identify in Oceania.

Question 3: independent project brainstorming (not graded)

For your independent project you will find and analyze a dataset using the tools we learn in this course. At this stage we want you to think about possible datasets. Identify three datasets you might like to work with, and assess how practical and appropriate the dataset might be for our class. It should have many observations (but not e.g. billions, so that you can't easily work with it) and it should have more than one variable/attribute for each observation.

Dataset 1: Chinese Political Elite Database
This dataset needs some recoding but could be used in this course.

Dataset 2: V-Dem

Dataset 3: Global Transitional Justice Dataset (1946 - 2016)
Or the Dataset on Authoritarian Elites to be published Dec. 2021.