POLI3148 Lecture 4 in class exercise 2

Lau Wai Fung Raymond

2023-10-05

```
library(tidyverse)
```

Load the `tidyverse` package

```
## -- Attaching core tidyverse packages ----------------------- tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
d <- read_csv("_DataPublic_/vdem/1984_2022/vdem_1984_2022_external.csv")
```

Import and Export V-dem data

```
## Rows: 6789 Columns: 211
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr    (3): country_name, country_text_id, histname
## dbl  (207): country_id, year, project, historical, codingstart, codingend, c...
## date   (1): historical_date
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

1. Codebook look up

i. What indicators regarding the quality of education are available in the V-Dem datasets?

In the V-Dem dataset, there is two indicators that relates to the quality of education, which is Educational inequality (e_peedgini), and Education 15+ (e_peaveduc).

ii. What are the data's coverage (i.e., for which countries and years do we have data?)

```
# Find out which country and year are available on the dataset
d |> select(country_name, country_id, year) |> distinct()
```

```
## # A tibble: 6,789 x 3
##    country_name country_id  year
##    <chr>             <dbl> <dbl>
##  1 Mexico                3  1984
##  2 Mexico                3  1985
##  3 Mexico                3  1986
##  4 Mexico                3  1987
##  5 Mexico                3  1988
##  6 Mexico                3  1989
##  7 Mexico                3  1990
##  8 Mexico                3  1991
##  9 Mexico                3  1992
## 10 Mexico                3  1993
## # i 6,779 more rows
```

Based on the codebook, the data coverage (year) for Education 15+ indicator is available from 1820 to 2022, while the data coverage (year) for Educational inequality indicator is available form 1850 to 2010.

iii. What are their sources? Provide the link to least 1 source.

The V-Dem data is download from V-Dem Dataset website.

```
# Load the `tidyverse` package before importing the dataset
library(tidyverse)
```

```
# Import the V-Dem dataset that used for this in-class exercise 2
d <- read_csv("_DataPublic_/vdem/1984_2022/vdem_1984_2022_external.csv")
```

```
## Rows: 6789 Columns: 211
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr    (3): country_name, country_text_id, histname
## dbl  (207): country_id, year, project, historical, codingstart, codingend, c...
## date   (1): historical_date
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

2. Subset by columns

i. Create a dataset containing only the country-year identifiers and indicators of education quality.

```
# create a new dataset that contains country-year identifiers, and education quality indicators

d_edu <- d |>
  select(country_name, country_id, year, e_peaveduc, e_peedgini)

d_edu
```

```
## # A tibble: 6,789 x 5
##    country_name country_id  year e_peaveduc e_peedgini
##    <chr>             <dbl> <dbl>      <dbl>      <dbl>
##  1 Mexico                3  1984       6.08       32.7
##  2 Mexico                3  1985       6.22       32.4
##  3 Mexico                3  1986       6.36       31.9
##  4 Mexico                3  1987       6.5        31.4
##  5 Mexico                3  1988       6.64       31.1
##  6 Mexico                3  1989       6.78       30.1
##  7 Mexico                3  1990       6.92       30.0
##  8 Mexico                3  1991       7.03       29.7
##  9 Mexico                3  1992       7.14       29.5
## 10 Mexico                3  1993       7.25       29.3
## # i 6,779 more rows
```

ii. **Rename the columns of education quality to make them informative.**

```
# Rename the columns of educational quality indicators
d_edu <- d_edu |>
  rename("avg_year_education" = "e_peaveduc", "education_inequality" ="e_peedgini",
         "Country" = "country_name", "ID" = "country_id",
         "Year" = "year")

d_edu
```

```
## # A tibble: 6,789 x 5
##    Country    ID  Year avg_year_education education_inequality
##    <chr>   <dbl> <dbl>              <dbl>                <dbl>
##  1 Mexico      3  1984               6.08                 32.7
##  2 Mexico      3  1985               6.22                 32.4
##  3 Mexico      3  1986               6.36                 31.9
##  4 Mexico      3  1987               6.5                  31.4
##  5 Mexico      3  1988               6.64                 31.1
##  6 Mexico      3  1989               6.78                 30.1
##  7 Mexico      3  1990               6.92                 30.0
##  8 Mexico      3  1991               7.03                 29.7
##  9 Mexico      3  1992               7.14                 29.5
## 10 Mexico      3  1993               7.25                 29.3
## # i 6,779 more rows
```

3. Subset by rows

i. List 5 countries-years that have the highest education level among its population.

```
# List five countries-years with highest education level among the population
d_edu |> slice_max(order_by = avg_year_education, n = 5)
```

```
## # A tibble: 13 x 5
##    Country          ID  Year avg_year_education education_inequality
##    <chr>         <dbl> <dbl>              <dbl>                <dbl>
##  1 United Kingdom  101  2010               13.3                 6.07
##  2 United Kingdom  101  2011               13.3                 NA
```

3

```
##  3 United Kingdom   101  2012                  13.3                NA
##  4 United Kingdom   101  2013                  13.3                NA
##  5 United Kingdom   101  2014                  13.3                NA
##  6 United Kingdom   101  2015                  13.3                NA
##  7 United Kingdom   101  2016                  13.3                NA
##  8 United Kingdom   101  2017                  13.3                NA
##  9 United Kingdom   101  2018                  13.3                NA
## 10 United Kingdom   101  2019                  13.3                NA
## 11 United Kingdom   101  2020                  13.3                NA
## 12 United Kingdom   101  2021                  13.3                NA
## 13 United Kingdom   101  2022                  13.3                NA
```

ii. List 5 countries-years that suffer from the most severe inequality in education.

```
# Get 5 country-years that have the most severe educational inequality
d_edu |> slice_min(order_by = education_inequality, n = 5)
```

```
## # A tibble: 5 x 5
##    Country      ID  Year avg_year_education education_inequality
##    <chr>     <dbl> <dbl>              <dbl>                <dbl>
## 1 Barbados    147  2008               9.57                 3.77
## 2 Barbados    147  2003               9.32                 3.80
## 3 Barbados    147  2007               9.52                 4.01
## 4 Austria     144  2007              11.4                  4.03
## 5 Austria     144  2008              11.4                  4.04
```

4. Summarise the data

i. Check data availability: For which countries and years are the indicators of education quality available

```
# Data integrity check
# first step: check whether the educational quality are available in different countries and year
d_edu |>
  mutate(avg_year_education_missing = as.numeric(is.na(avg_year_education)),
         .after = avg_year_education) |>
  mutate(education_inequality_missing = as.numeric(is.na(education_inequality)),
         .after = education_inequality) |>
  arrange(Country)
```

```
## # A tibble: 6,789 x 7
##    Country       ID  Year avg_year_education avg_year_education_missing
##    <chr>      <dbl> <dbl>              <dbl>                      <dbl>
##  1 Afghanistan   36  1984               1.30                          0
##  2 Afghanistan   36  1985               1.35                          0
##  3 Afghanistan   36  1986               1.40                          0
##  4 Afghanistan   36  1987               1.45                          0
##  5 Afghanistan   36  1988               1.50                          0
##  6 Afghanistan   36  1989               1.55                          0
##  7 Afghanistan   36  1990               1.60                          0
##  8 Afghanistan   36  1991               1.69                          0
##  9 Afghanistan   36  1992               1.78                          0
```

```
## 10 Afghanistan    36  1993                    1.88                               0
## # i 6,779 more rows
## # i 2 more variables: education_inequality <dbl>,
## #   education_inequality_missing <dbl>
```

```r
# second step
# For each country, calculate the missing data in education quality
d_edu |>
    mutate(avg_year_education_missing = as.numeric(is.na(avg_year_education)),
           .after = avg_year_education) |>
    mutate(education_inequality_missing = as.numeric(is.na(education_inequality)),
           .after = education_inequality) |>
    group_by(Country) |>
    summarise(N_avg_years_education_missing = sum(avg_year_education_missing),
              N_education_inequality_missing = sum(education_inequality_missing))
```

```
## # A tibble: 181 x 3
##    Country     N_avg_years_education_missing N_education_inequality_missing
##    <chr>                           <dbl>                         <dbl>
##  1 Afghanistan                         0                            12
##  2 Albania                            39                            39
##  3 Algeria                             0                            12
##  4 Angola                              0                            12
##  5 Argentina                           0                            12
##  6 Armenia                             0                            12
##  7 Australia                           0                            12
##  8 Austria                             0                            12
##  9 Azerbaijan                          0                            12
## 10 Bahrain                            39                            39
## # i 171 more rows
```

ii. **Create two types of country-level indicators of education quality**

```r
# Average level of education quality from 1984 to 2022
d_edu |>
  group_by(Country) |>
  summarise(average_education_years = mean(avg_year_education, na.rm = TRUE),
            average_education_inequality = mean(education_inequality, na.rm = TRUE))
```

```
## # A tibble: 181 x 3
##    Country     average_education_years average_education_inequality
##    <chr>                         <dbl>                        <dbl>
##  1 Afghanistan                    2.80                         77.8
##  2 Albania                         NaN                          NaN
##  3 Algeria                        6.31                         45.8
##  4 Angola                         2.46                         53.9
##  5 Argentina                      8.37                         16.6
##  6 Armenia                       10.7                          16.5
##  7 Australia                     12.9                           9.60
##  8 Austria                       11.2                           6.35
##  9 Azerbaijan                    10.7                          14.5
## 10 Bahrain                         NaN                          NaN
## # i 171 more rows
```

```
# Question: change of education quality from 1984 to 2022
# Due to data availability issue, comparing 2010 with 1984
d_edu |>
  filter(Year >= 1984 & Year <= 2010) |>
  group_by(Country) |>
  arrange(Year) |>
  summarise(educationyears_2010_1984 = (last(avg_year_education) - first(avg_year_education))/
              first(avg_year_education),
            education_inequality_2010_1984 = (last(education_inequality) - first(education_inequality))
            / first(education_inequality)) |>
  ungroup() |>
  arrange(Country)
```

```
## # A tibble: 180 x 3
##    Country     educationyears_2010_1984 education_inequality_2010_1984
##    <chr>                          <dbl>                          <dbl>
##  1 Afghanistan                   1.94                          -0.246
##  2 Albania                       NA                            NA
##  3 Algeria                       0.847                         -0.335
##  4 Angola                        1.22                          -0.440
##  5 Argentina                     0.138                         -0.185
##  6 Armenia                       0.0321                        -0.154
##  7 Australia                     0.0716                        -0.551
##  8 Austria                       0.112                         -0.575
##  9 Azerbaijan                    0.0239                        -0.132
## 10 Bahrain                       NA                            NA
## # i 170 more rows
```

iii. Examine the data and briefly discuss: Which countries perform the best and the worst in terms of education quality in the past four decades?

Comparing the country's performance in education, Germany has been the country that performs the best in the average years of education in the past decades, citizens older than 15 had received 12.939 years of education on average. On the other hand, Burkina Faso represents a country with the lowest average years of education. On average, less than a year of education for citizens over 15 years old (value = 0.982).

Likewise, Burkina Faso is the country that has suffered the most severe educational inequality on average, compared to other countries. The Gini coefficient of unequal education in Burkina Faso is 91.26. In contrast, Austria is the country that had the lowest average educational inequality, whereas their Gini coefficient only consists of 6.35.