PoliDATA

Yifei Zhu

2023-09-21

Poli3148 class assignment 2

```
library(tidyverse)
```

```
## -- 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
```

```
##data import
setwd('D:/OneDrive - The University of Hong Kong - Connect/dataprojects/POLIDATA/Poliassignment')
vdem <- read_csv("D:/OneDrive - The University of Hong Kong - Connect/dataprojects/POLIDATA/Poliassignme
```

```
## Rows: 6789 Columns: 3196
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr    (18): country_name, country_text_id, histname, v2lpname, v2slpname, v...
## dbl  (3177): country_id, year, project, historical, codingstart, codingend, ...
## 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.
```

Codebook lookup

1. quality of education variable: v2peedueq
2. Data coverage: 1900-2022
3. Sources: World bank indicators: https://databank.worldbank.org/

Subset by columns

```r
  #select variables
vdem_edu <- vdem |>
  select(country_name, country_id,year, v2peedueq)
```

```r
  #rename
vdem_edu <- vdem_edu |> rename(edu_quality=v2peedueq)
```

Subset by rows

```r
  #5 countries_years have the highest edu
vdem_edu |>
  group_by(country_id) |>
  slice_max(order_by = edu_quality,n=5)
```

```
## # A tibble: 2,536 x 4
## # Groups:   country_id [181]
##    country_name country_id  year edu_quality
##    <chr>             <dbl> <dbl>       <dbl>
##  1 Mexico                3  2018      -0.686
##  2 Mexico                3  2019      -0.686
##  3 Mexico                3  2020      -0.686
##  4 Mexico                3  2017      -0.753
##  5 Mexico                3  2021      -0.815
##  6 Mexico                3  2022      -0.815
##  7 Suriname              4  2010       0.79
##  8 Suriname              4  2011       0.79
##  9 Suriname              4  2012       0.79
## 10 Suriname              4  2013       0.782
## # i 2,526 more rows
```

```r
  #5 countries_years have the lowest edu
vdem_edu |>
  group_by(country_id) |>
  slice_min(order_by = edu_quality,n=5)
```

```
## # A tibble: 2,431 x 4
## # Groups:   country_id [181]
##    country_name country_id  year edu_quality
##    <chr>             <dbl> <dbl>       <dbl>
##  1 Mexico                3  1984       -1.09
##  2 Mexico                3  1985       -1.09
##  3 Mexico                3  1986       -1.09
##  4 Mexico                3  1987       -1.09
##  5 Mexico                3  1988       -1.09
##  6 Mexico                3  1989       -1.09
##  7 Mexico                3  1990       -1.09
##  8 Mexico                3  1991       -1.09
##  9 Mexico                3  1992       -1.09
## 10 Mexico                3  1993       -1.09
## # i 2,421 more rows
```

Summarize the data

```
  #data availibility
vdem_edu |>
  mutate(edu_missing=as.numeric(is.na(edu_quality))) |>
  group_by(country_id) |>
  summarize(edu_missing)
```

```
## Warning: Returning more (or less) than 1 row per ‘summarise()‘ group was deprecated in
## dplyr 1.1.0.
## i Please use ‘reframe()‘ instead.
## i When switching from ‘summarise()‘ to ‘reframe()‘, remember that ‘reframe()‘
##   always returns an ungrouped data frame and adjust accordingly.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

```
## ‘summarise()‘ has grouped output by ’country_id’. You can override using the
## ‘.groups‘ argument.
```

```
## # A tibble: 6,789 x 2
## # Groups:   country_id [181]
##    country_id edu_missing
##         <dbl>       <dbl>
##  1          3           0
##  2          3           0
##  3          3           0
##  4          3           0
##  5          3           0
##  6          3           0
##  7          3           0
##  8          3           0
##  9          3           0
## 10          3           0
## # i 6,779 more rows
```

```
  #Country level indicaters
vdem_edu |>
  group_by(country_id) |>
  mutate(avgedu=mean(edu_quality),na.rm=TRUE)
```

```
## # A tibble: 6,789 x 6
## # Groups:   country_id [181]
##    country_name country_id  year edu_quality avgedu na.rm
##    <chr>             <dbl> <dbl>       <dbl>  <dbl> <lgl>
##  1 Mexico                3  1984       -1.09  -1.01 TRUE
##  2 Mexico                3  1985       -1.09  -1.01 TRUE
##  3 Mexico                3  1986       -1.09  -1.01 TRUE
##  4 Mexico                3  1987       -1.09  -1.01 TRUE
##  5 Mexico                3  1988       -1.09  -1.01 TRUE
##  6 Mexico                3  1989       -1.09  -1.01 TRUE
##  7 Mexico                3  1990       -1.09  -1.01 TRUE
##  8 Mexico                3  1991       -1.09  -1.01 TRUE
##  9 Mexico                3  1992       -1.09  -1.01 TRUE
## 10 Mexico                3  1993       -1.09  -1.01 TRUE
## # i 6,779 more rows
```

```
  ## Change of education quality
vdem_edu<-
  vdem_edu |>
  group_by(country_id) |>
  arrange(year) |>
  mutate(edu_year_change = edu_quality-first(edu_quality), edu_total_change =last(edu_quality)-first(edu
  ungroup() |>
  arrange(country_id, year)
```

Which countries perform the best and the worst in terms of education quality in the past four decades?

It is hard to decide which country did the best and the worst in the past four decades. Below are two aspects regarding education quality.

United Arab Emirates improved most in education quality, their education quality increased from -0.583 in 1984 to 2.281 in 2022. In contrast, Tajikistan decreased from 2.464 in 1984 to -0.526 in 2022. Norway scores the highest in 2022 with 3.475, and Yemen scores the lowest with -2.736.

```
## which country improved most in 4 decades
vdem_edu |>
  slice_max(order_by = edu_total_change,n=1)
```

```
## # A tibble: 39 x 6
##    country_name    country_id  year edu_quality edu_year_change edu_total_change
##    <chr>                <dbl> <dbl>       <dbl>           <dbl>            <dbl>
##  1 United Arab Em~        207  1984      -0.583           0                 2.88
##  2 United Arab Em~        207  1985      -0.583           0                 2.88
##  3 United Arab Em~        207  1986      -0.583           0                 2.88
##  4 United Arab Em~        207  1987      -0.583           0                 2.88
##  5 United Arab Em~        207  1988      -0.583           0                 2.88
##  6 United Arab Em~        207  1989      -0.583           0                 2.88
##  7 United Arab Em~        207  1990      -0.583           0                 2.88
##  8 United Arab Em~        207  1991      -0.583           0                 2.88
##  9 United Arab Em~        207  1992      -0.583           0                 2.88
## 10 United Arab Em~        207  1993       0.074           0.657             2.88
## # i 29 more rows
```

```
## which country least in 4 decades
vdem_edu |>
  slice_min(order_by = edu_total_change,n=1)
```

```
## # A tibble: 33 x 6
##    country_name country_id  year edu_quality edu_year_change edu_total_change
##    <chr>             <dbl> <dbl>       <dbl>           <dbl>            <dbl>
##  1 Tajikistan          133  1990       2.46             0               -2.99
##  2 Tajikistan          133  1991       2.46             0               -2.99
##  3 Tajikistan          133  1992       1.21            -1.26            -2.99
##  4 Tajikistan          133  1993       0.116           -2.35            -2.99
##  5 Tajikistan          133  1994       0.116           -2.35            -2.99
##  6 Tajikistan          133  1995       0.116           -2.35            -2.99
##  7 Tajikistan          133  1996       0.116           -2.35            -2.99
##  8 Tajikistan          133  1997       0.116           -2.35            -2.99
##  9 Tajikistan          133  1998       0.116           -2.35            -2.99
## 10 Tajikistan          133  1999       0.365           -2.10            -2.99
## # i 23 more rows
```

```
## which country has the best edu in 2022
vdem_edu |>
  group_by(country_id) |>
  slice_max(order_by = year,n=1) |>
  ungroup() |>
  slice_max(order_by = edu_quality, n=1)
```

```
## # A tibble: 1 x 6
##   country_name country_id  year edu_quality edu_year_change edu_total_change
##   <chr>             <dbl> <dbl>       <dbl>           <dbl>            <dbl>
## 1 Norway              186  2022        3.48               0                0
```

```
## which country has the worst edu in 2022
vdem_edu |>
  group_by(country_id) |>
  slice_max(order_by = year,n=1) |>
  ungroup() |>
  slice_min(order_by = edu_quality, n=1)
```

```
## # A tibble: 1 x 6
##   country_name country_id  year edu_quality edu_year_change edu_total_change
##   <chr>             <dbl> <dbl>       <dbl>           <dbl>            <dbl>
## 1 Yemen                14  2022       -2.74          -0.997           -0.997
```