In-class ex

Nicole Wong 3035785697

Last update: 9 Oct 2023

1. Codebook Lookup

1. **What indicators regarding the quality of education are available in the V-Dem datasets?**
   - *Education 15+ (E) (e_peaveduc)*
     - *The Average years of education in the total population aged 15 years and older.*
   - *Educational inequality, Gini (E) (e_peedgini)*
     - *Gini coefficient of educational inequality estimated from average education data*
2. **What are the data's coverage (i.e., for which countries and years do we have data?)**
   - For *Education 15+ (E) (e_peaveduc)*
     - Years: 1820-2022
   - For *Educational inequality, Gini (E) (e_peedgini)*
     - Years: 1850-2010
3. What are their sources? Provide the link to least 1 source.
   - Sources : *Clio Infra (clio-infra.eu), drawing on Mitchell (1998a, 1998b, 1998c), United States Census Bureau (2021), UNESCO, Földvári and van Leeuwen (2010a), Leeuwen, van Leeuwen- Li, Földvári (2011), Leeuwen, van Leeuwen-Li, Földvári (2012a), Leeuwen, van Leeuwen-Li, Földvári (2012b), Didenko, Foldvari, van Leeuwen (2012).*
   - Link: https://clio-infra.eu/Indicators/AverageYearsofEducation.html

Data Pre-processing

```
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 1.9.2     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
```

```
library(dplyr)
d <- read_csv("_DataPublic_/vdem/1984_2022/vdem_1984_2022_external.csv", show_col_types = FALSE)

#names(d)
```

2. Subset by Columns

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

```
#create new dataset
d_edu <- d |>
  select(country_name, country_id, year, e_peaveduc, e_peedgini)
```

2. Rename the columns of education quality to make them informative.

```
#rename columns
d_edu <- d_edu |>
  rename("Average_Education" = "e_peaveduc", "Gini" = "e_peedgini", "Country" = "country_name", "ID" =
```

3. Subset by Rows

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

```
#List top 5 Country-Years (Avg. edu)
d_edu |>
  select(Country, Year, Average_Education) |>
  arrange(-Average_Education) |>
  slice_head(n = 5)
```

```
## # A tibble: 5 x 3
##   Country         Year Average_Education
##   <chr>          <dbl>             <dbl>
## 1 United Kingdom  2010              13.3
## 2 United Kingdom  2011              13.3
## 3 United Kingdom  2012              13.3
## 4 United Kingdom  2013              13.3
## 5 United Kingdom  2014              13.3
```

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

```
#List bottom 5 country-years (Gini)
d_edu |>
  select(Country, Year, Gini) |>
  arrange(-Gini) |>
  slice_head(n = 5)
```

```
## # A tibble: 5 x 3
##   Country       Year  Gini
##   <chr>        <dbl> <dbl>
```

```
## 1 Burkina Faso  1984  97.0
## 2 Burkina Faso  1985  96.9
## 3 Burkina Faso  1986  96.7
## 4 Burkina Faso  1987  96.4
## 5 Burkina Faso  1988  96.1
```

4. Summarize the Data

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

```
# Data availability/ integrity check

#Reference list: If indicators are available, mark as true
d_edu_filtered <- d_edu |>
  group_by(Country, Year) |>
  mutate (Average_Edu_available = !any(is.na(Average_Education)), Average_Gini_available = !any(is.
```

```
#missing by country
d_edu|>
  group_by(Country) |>
  #create a column to indicate missing Average_Education
  mutate (Average_Education_missing = is.na(Average_Education), Gini_missing = is.na(Gini)) |>
  summarize(n_Average_Education_missing = sum(Average_Education_missing), n_Gini_missing = sum(Gini
```

```
## # A tibble: 181 x 3
##    Country     n_Average_Education_missing n_Gini_missing
##    <chr>                             <int>          <int>
##  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
```

```
#missing by year
d_edu|>
  group_by(Year) |>
mutate (Average_Education_missing = is.na(Average_Education), Gini_missing = is.na(Gini)) |>
  summarize(n_Average_Education_missing = sum(Average_Education_missing), n_Gini_missing = sum(Gini
```

```
## # A tibble: 39 x 3
##     Year n_Average_Education_missing n_Gini_missing
##    <dbl>                       <int>          <int>
##  1  1984                          40             42
##  2  1985                          40             42
##  3  1986                          40             42
```

```
##  4  1987                          40              42
##  5  1988                          40              42
##  6  1989                          41              43
##  7  1990                          42              44
##  8  1991                          43              45
##  9  1992                          44              46
## 10  1993                          45              47
## # i 29 more rows
```

2. Create two types of country-level indicators of education quality

- Average level of education quality from 1984 to 2022

```
d_edu_country1 <- d_edu |>
  group_by(Country) |>
  summarise(Mean_Average_Education = mean(Average_Education, na.rm = TRUE), Mean_Gini = mean(Gini, na.r

d_edu_country1
```

```
## # A tibble: 181 x 3
##    Country     Mean_Average_Education Mean_Gini
##    <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
```

- Change of education quality from 1984 to 2022

```
d_edu_country2 <- d_edu |>
  filter(Year >= 1984 & Year <= 2010) |>
  group_by(Country) |>
  arrange(Year) |>
  summarize(Average_Education_Growth = (last(Average_Education) - first(Average_Education))/last(Averag
  ungroup() |>
  arrange(Country)

d_edu_country2
```

```
## # A tibble: 180 x 3
##    Country     Average_Education_Growth Gini_Growth
##    <chr>                          <dbl>       <dbl>
##  1 Afghanistan                    0.660      -0.326
##  2 Albania                       NA          NA
##  3 Algeria                        0.459      -0.503
```

```
##  4 Angola                         0.550       -0.785
##  5 Argentina                      0.121       -0.227
##  6 Armenia                        0.0311      -0.182
##  7 Australia                      0.0668      -1.23
##  8 Austria                        0.101       -1.35
##  9 Azerbaijan                     0.0233      -0.152
## 10 Bahrain                       NA           NA
## # i 170 more rows
```

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

Best and Worst Ranking

First, we will rank all of the countries in terms of the above 4 indicators to get the top 10 and bottom 10 countries in terms of each category.

```
merged_table <- merge(d_edu_country1, d_edu_country2, by = "Country")
```

```
ranked_data <- merged_table |>
  #mean average
  arrange(-Mean_Average_Education)|>
  mutate(Ranking_Mean_Ave = row_number()) |>
  #mean geni
  arrange(Mean_Gini) |>
  mutate(Ranking_Mean_Gini = row_number()) |>
  #growth average
  arrange(-Average_Education_Growth) |>
  mutate(Ranking_Ave_Growth = row_number()) |>
  #Gini growth
  arrange(Gini_Growth) |>
  mutate(Ranking_Gini_Growth = row_number())
```

Then we can try to create an overall ranking.

```
overall_ranking <- ranked_data |>
  mutate(Total_Score = (Ranking_Mean_Ave + Ranking_Mean_Gini + Ranking_Ave_Growth + Ranking_Gini_Growth
  arrange(Total_Score) |>
  mutate(Overall_Rank = row_number()) |>
  select(Country, Overall_Rank)
```

Hence, the best performing countries are:

```
slice_head(overall_ranking, n = 10)
```

```
##           Country Overall_Rank
## 1        Botswana            1
## 2  United Kingdom            2
## 3         Austria            3
## 4       Australia            4
## 5          Canada            5
```

```
## 6         Hungary            6
## 7          Norway            7
## 8        Barbados            8
## 9         Denmark            9
## 10        Iceland           10
```

Whereas, the worst performing countries are:

```
slice_tail(overall_ranking, n = 10)
```

```
##                     Country Overall_Rank
## 1                  Suriname          171
## 2                    Taiwan          172
## 3               Timor-Leste          173
## 4              Turkmenistan          174
## 5      United Arab Emirates          175
## 6  United States of America          176
## 7                   Vanuatu          177
## 8                   Vietnam          178
## 9                     Yemen          179
## 10                 Zanzibar          180
```

Additional Notes

Something interesting to note is that there is a **strong inverse correlation between the average education level and the Gini coefficient** (-0.889). In other words, the higher the average education level, the lower the education inequality. Similarly, there is a strong correlation between the average education level and the growth in average education level. While we cannot define the causation, it is noted that countries with higher education levels also experience higher growth in education levels.

On the other hand, there is a weak inverse correlation between the average education level growth and growth (change) in Gini coefficient (-0.195). In addition, there is no correlation between the Gini coefficient and growth in Gini coefficient (0.032). It can be concluded that the improvement in inequality is not related to the inequality of education. Hence, the level inequality does not impact the rate of improvement.

Correlation analysis (for reference)

```
#correlation between average education and gini coefficient
cor(d_edu_country1$Mean_Average_Education, d_edu_country1$Mean_Gini, use = "complete.obs")
```

```
## [1] -0.888669
```

```
cor(d_edu_country2$Average_Education_Growth, d_edu_country2$Gini_Growth, use = "complete.obs")
```

```
## [1] -0.1954076
```

```
#correlation between average education and and average educaiton growth
cor(merged_table$Mean_Average_Education, merged_table$Average_Education_Growth, use = "complete.obs")
```

```
## [1] -0.8212475
```

```r
#correlation between gini and and gini growth
cor(merged_table$Mean_Gini, merged_table$Gini_Growth, use = "complete.obs")
```

```
## [1] 0.03280402
```