in-class assignment2

Yingqi Huang

2023-10-08

1. Codebook Lookup

i. Indicators of education quality

This step is to set the R working dictionary the same as where the data is stored.

```
setwd("C:/Users/u3617194/Desktop/portfolio/_DataPublic_/vdem/Assignment2/1984_2022")
```

This step is to load the packages needed to tidy up the data later.

```
library(tidyverse)
```

This step is to read the file, and name is set as d.

```
d <- read_csv("vdem_1984_2022_external.csv")
names(d)
```

```
##   [1] "country_name"              "country_text_id"
##   [3] "country_id"                "year"
##   [5] "historical_date"           "project"
##   [7] "historical"                "histname"
##   [9] "codingstart"               "codingend"
##  [11] "codingstart_contemp"       "codingend_contemp"
##  [13] "codingstart_hist"          "codingend_hist"
##  [15] "gapstart1"                 "gapstart2"
##  [17] "gapstart3"                 "gapend1"
##  [19] "gapend2"                   "gapend3"
##  [21] "gap_index"                 "COWcode"
##  [23] "e_v2x_api_3C"              "e_v2x_api_4C"
##  [25] "e_v2x_api_5C"              "e_v2x_civlib_3C"
##  [27] "e_v2x_civlib_4C"           "e_v2x_civlib_5C"
##  [29] "e_v2x_clphy_3C"            "e_v2x_clphy_4C"
##  [31] "e_v2x_clphy_5C"            "e_v2x_clpol_3C"
##  [33] "e_v2x_clpol_4C"            "e_v2x_clpol_5C"
##  [35] "e_v2x_clpriv_3C"           "e_v2x_clpriv_4C"
##  [37] "e_v2x_clpriv_5C"           "e_v2x_corr_3C"
##  [39] "e_v2x_corr_4C"             "e_v2x_corr_5C"
##  [41] "e_v2x_cspart_3C"           "e_v2x_cspart_4C"
##  [43] "e_v2x_cspart_5C"           "e_v2x_delibdem_3C"
```

```
##  [45] "e_v2x_delibdem_4C"          "e_v2x_delibdem_5C"
##  [47] "e_v2x_EDcomp_thick_3C"      "e_v2x_EDcomp_thick_4C"
##  [49] "e_v2x_EDcomp_thick_5C"      "e_v2x_egal_3C"
##  [51] "e_v2x_egal_4C"              "e_v2x_egal_5C"
##  [53] "e_v2x_egaldem_3C"           "e_v2x_egaldem_4C"
##  [55] "e_v2x_egaldem_5C"           "e_v2x_elecoff_3C"
##  [57] "e_v2x_elecoff_4C"           "e_v2x_elecoff_5C"
##  [59] "e_v2x_execorr_3C"           "e_v2x_execorr_4C"
##  [61] "e_v2x_execorr_5C"           "e_v2x_feduni_3C"
##  [63] "e_v2x_feduni_4C"            "e_v2x_feduni_5C"
##  [65] "e_v2x_frassoc_thick_3C"     "e_v2x_frassoc_thick_4C"
##  [67] "e_v2x_frassoc_thick_5C"     "e_v2x_freexp_3C"
##  [69] "e_v2x_freexp_4C"            "e_v2x_freexp_5C"
##  [71] "e_v2x_freexp_altinf_3C"     "e_v2x_freexp_altinf_4C"
##  [73] "e_v2x_freexp_altinf_5C"     "e_v2x_gencl_3C"
##  [75] "e_v2x_gencl_4C"             "e_v2x_gencl_5C"
##  [77] "e_v2x_gencs_3C"             "e_v2x_gencs_4C"
##  [79] "e_v2x_gencs_5C"             "e_v2x_gender_3C"
##  [81] "e_v2x_gender_4C"            "e_v2x_gender_5C"
##  [83] "e_v2x_genpp_3C"             "e_v2x_genpp_4C"
##  [85] "e_v2x_genpp_5C"             "e_v2x_jucon_3C"
##  [87] "e_v2x_jucon_4C"             "e_v2x_jucon_5C"
##  [89] "e_v2x_libdem_3C"            "e_v2x_libdem_4C"
##  [91] "e_v2x_libdem_5C"            "e_v2x_liberal_3C"
##  [93] "e_v2x_liberal_4C"           "e_v2x_liberal_5C"
##  [95] "e_v2x_mpi_3C"               "e_v2x_mpi_4C"
##  [97] "e_v2x_mpi_5C"               "e_v2x_partip_3C"
##  [99] "e_v2x_partip_4C"            "e_v2x_partip_5C"
## [101] "e_v2x_partipdem_3C"         "e_v2x_partipdem_4C"
## [103] "e_v2x_partipdem_5C"         "e_v2x_polyarchy_3C"
## [105] "e_v2x_polyarchy_4C"         "e_v2x_polyarchy_5C"
## [107] "e_v2x_pubcorr_3C"           "e_v2x_pubcorr_4C"
## [109] "e_v2x_pubcorr_5C"           "e_v2x_suffr_3C"
## [111] "e_v2x_suffr_4C"             "e_v2x_suffr_5C"
## [113] "e_v2xcl_rol_3C"             "e_v2xcl_rol_4C"
## [115] "e_v2xcl_rol_5C"             "e_v2xcs_ccsi_3C"
## [117] "e_v2xcs_ccsi_4C"            "e_v2xcs_ccsi_5C"
## [119] "e_v2xdd_dd_3C"              "e_v2xdd_dd_4C"
## [121] "e_v2xdd_dd_5C"              "e_v2xdl_delib_3C"
## [123] "e_v2xdl_delib_4C"           "e_v2xdl_delib_5C"
## [125] "e_v2xeg_eqdr_3C"            "e_v2xeg_eqdr_4C"
## [127] "e_v2xeg_eqdr_5C"            "e_v2xeg_eqprotec_3C"
## [129] "e_v2xeg_eqprotec_4C"        "e_v2xeg_eqprotec_5C"
## [131] "e_v2xel_frefair_3C"         "e_v2xel_frefair_4C"
## [133] "e_v2xel_frefair_5C"         "e_v2xel_locelec_3C"
## [135] "e_v2xel_locelec_4C"         "e_v2xel_locelec_5C"
## [137] "e_v2xel_regelec_3C"         "e_v2xel_regelec_4C"
## [139] "e_v2xel_regelec_5C"         "e_v2xlg_legcon_3C"
## [141] "e_v2xlg_legcon_4C"          "e_v2xlg_legcon_5C"
## [143] "e_v2xme_altinf_3C"          "e_v2xme_altinf_4C"
## [145] "e_v2xme_altinf_5C"          "e_v2xps_party_3C"
## [147] "e_v2xps_party_4C"           "e_v2xps_party_5C"
## [149] "e_boix_regime"              "e_democracy_breakdowns"
## [151] "e_democracy_omitteddata"    "e_democracy_trans"
```

```
## [153] "e_fh_cl"                    "e_fh_pr"
## [155] "e_fh_rol"                   "e_fh_status"
## [157] "e_wbgi_cce"                 "e_wbgi_gee"
## [159] "e_wbgi_pve"                 "e_wbgi_rle"
## [161] "e_wbgi_rqe"                 "e_wbgi_vae"
## [163] "e_lexical_index"            "e_uds_median"
## [165] "e_uds_mean"                 "e_uds_pct025"
## [167] "e_uds_pct975"               "e_coups"
## [169] "e_legparty"                 "e_autoc"
## [171] "e_democ"                    "e_p_polity"
## [173] "e_polcomp"                  "e_polity2"
## [175] "e_bnr_dem"                  "e_chga_demo"
## [177] "e_ti_cpi"                   "e_vanhanen"
## [179] "e_peaveduc"                 "e_peedgini"
## [181] "e_area"                     "e_regiongeo"
## [183] "e_regionpol"                "e_regionpol_6C"
## [185] "e_cow_exports"              "e_cow_imports"
## [187] "e_gdp"                      "e_gdp_sd"
## [189] "e_gdppc"                    "e_gdppc_sd"
## [191] "e_miinflat"                 "e_pop"
## [193] "e_pop_sd"                   "e_total_fuel_income_pc"
## [195] "e_total_oil_income_pc"      "e_total_resources_income_pc"
## [197] "e_radio_n"                  "e_miferrat"
## [199] "e_mipopula"                 "e_miurbani"
## [201] "e_miurbpop"                 "e_pefeliex"
## [203] "e_peinfmor"                 "e_pelifeex"
## [205] "e_pematmor"                 "e_wb_pop"
## [207] "e_civil_war"                "e_miinteco"
## [209] "e_miinterc"                 "e_pt_coup"
## [211] "e_pt_coup_attempts"
```

According to the codebook of the V-dem background factors, there are two variables that can indicate the education quality, they are:

(1) Education 15+ (E) (e_peaveguc), and

(2) Eudcation inequality, Gini (E) (e_peedgini).

The following step is to select the country and the corresponding indicators of education quality.

```
d_education <- d |>
  select(country_name, year, e_peaveduc, e_peedgini) |> distinct()
```

The following step is to rename the variables to make them informative.

```
edu_quality <- d_education |>
  rename("Country" = "country_name", "Year" = "year", "Education_level" = "e_peaveduc", "Education_ineq

edu_quality
```

```
## # A tibble: 6,789 x 4
##    Country  Year Education_level Education_inequality
```

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

ii. Data's coverage

The data is available in a total of 181 countries listed below.

```
edu_quality |> select(Country) |> distinct()
```

```
## # A tibble: 181 x 1
##    Country
##    <chr>
##  1 Mexico
##  2 Suriname
##  3 Sweden
##  4 Switzerland
##  5 Ghana
##  6 South Africa
##  7 Japan
##  8 Burma/Myanmar
##  9 Russia
## 10 Albania
## # i 171 more rows
```

In the current data set, we have data covering from 1984 to 2013.

```
edu_quality |> select(Year) |> distinct()
```

```
## # A tibble: 39 x 1
##     Year
##    <dbl>
##  1  1984
##  2  1985
##  3  1986
##  4  1987
##  5  1988
##  6  1989
##  7  1990
##  8  1991
##  9  1992
## 10  1993
## # i 29 more rows
```

Besides, we have some additional information about the dates that the variables are coded, for instance, we know the coding start for the countries coded by Contemporary V-Dem as followed:

```
d|> select(country_name,codingstart_contemp) |> distinct()
```

```
## # A tibble: 181 x 2
##    country_name  codingstart_contemp
##    <chr>                       <dbl>
##  1 Mexico                       1900
##  2 Suriname                     1900
##  3 Sweden                       1900
##  4 Switzerland                  1900
##  5 Ghana                        1902
##  6 South Africa                 1900
##  7 Japan                        1900
##  8 Burma/Myanmar                1900
##  9 Russia                       1900
## 10 Albania                      1912
## # i 171 more rows
```

iii. The sources of data

According to the codebook, most indicators of education quality question have at least one sources. For instance, the indicator "Education 15+" is cited from Clio Infra:

The website of Clio Infra

And the sources of the "Educational inequality" data are multiple, examples include:

Clio Infra (The website of Clio Infra), United States Census Bureau (2021) (Link of website), and academic paper by Földvári and van Leeuwen (2011)(link as followed)

- Földvári, & van Leeuwen, B. (2011). Should less inequality in education lead to a more equal income distribution? *Education Economics, 19*(5), 537–554. https://doi.org/10.1080/09645292.2010.488472

2. Subset by columns

i. Country-year identifiers

The following step is to create a new data set containing only country-year identifiers and indicators of education quality. It is already created above, the "edu_quality" data set.

```
print(edu_quality)
```

```
## # A tibble: 6,789 x 4
##    Country  Year Education_level Education_inequality
##    <chr>   <dbl>           <dbl>                <dbl>
##  1 Mexico   1984            6.08                 32.7
##  2 Mexico   1985            6.22                 32.4
##  3 Mexico   1986            6.36                 31.9
##  4 Mexico   1987            6.5                  31.4
```

```
##  5 Mexico   1988           6.64                 31.1
##  6 Mexico   1989           6.78                 30.1
##  7 Mexico   1990           6.92                 30.0
##  8 Mexico   1991           7.03                 29.7
##  9 Mexico   1992           7.14                 29.5
## 10 Mexico   1993           7.25                 29.3
## # i 6,779 more rows
```

ii. Rename the variables.

This step is also done in the first question, using the function "*rename()*". Refer to Question 1:i

```
print(edu_quality)
```

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

3. Subset by rows

i. 5 countries-years with highest education level

The following step is to list five countries-years observations with highest educational level.

Given that all top 5 observations with highest educational level is the UK and the level remains the same, therefore, there are 13 observations listed in the output.

```
edu_quality |> slice_max(order_by = Education_level, n = 5)
```

```
## # A tibble: 13 x 4
##    Country         Year Education_level Education_inequality
##    <chr>          <dbl>           <dbl>                <dbl>
##  1 United Kingdom  2010            13.3                 6.07
##  2 United Kingdom  2011            13.3                NA
##  3 United Kingdom  2012            13.3                NA
##  4 United Kingdom  2013            13.3                NA
##  5 United Kingdom  2014            13.3                NA
##  6 United Kingdom  2015            13.3                NA
##  7 United Kingdom  2016            13.3                NA
##  8 United Kingdom  2017            13.3                NA
```

```
##  9 United Kingdom  2018            13.3                NA
## 10 United Kingdom  2019            13.3                NA
## 11 United Kingdom  2020            13.3                NA
## 12 United Kingdom  2021            13.3                NA
## 13 United Kingdom  2022            13.3                NA
```

ii. 5 countries-years with most severe educational inequality

```
edu_quality |> slice_min(order_by = Education_inequality, n = 5)
```

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

Results show that the top 5 countries-years observations are Barbados in 2008, 2003, 2007 and Austria in 2007 and 2008.

4. Summary the data

i. Data availability check

(1) The following task is to check the countries with missing values

First create a new column that indicates whether the value is missing for Educational level and Educational inequality.

```
edu_quality |>
  mutate(level_missing = as.numeric(is.na(Education_level)), .afrer = Education_level, inequality_missi
  group_by(Country) |>
  summarise(N_level_missing = sum(level_missing),
            N_inequality_missing = sum(inequality_missing))
```

```
## # A tibble: 181 x 3
##    Country     N_level_missing N_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
```

As shown in the output above, all countries being surveyed contain at least 7 years of which the data of the educational inequality is missing, while Albania, Bahrain, Bhutan, Bosnia and Herzegovina, Burma/Myanmar, Cape Verde, Comoros, Croatia, Congo, Dijbouti, Equatorial Guinea, Eritrea, Ethiopia, German Democratic Republic, Guinea-Bissau, Hong Kong, Indonesia, Ivory Coast, Kosovo, Kuwait, Luxembourg, Maldives, Malta, Mongolia, Montenegro, North Macedonia, Oman, Palestine, Papua Ne Guinea, Qatar, Republic of the Congo, Sao Tome and Principle, Serbia, Slovakia, Slovenia, Solomon Islands, Somaliland, South Sudan, South Yemen, Sudan, Suriname, Taiwan, Timor-Leste, Turkmenistan, United Arab Emirates, USA, Vanuatu, Vietnam, Yemen and Zanzibar have data regarding the educational level not available for at least 7 years.

(2) The following task is to check which years are the indicators available

```
edu_quality |>
  mutate(level_missing = as.numeric(is.na(Education_level)), .afrer = Education_level, inequality_missi
  group_by(Year) |>
  summarise(N_level_missing = sum(level_missing),
            N_inequality_missing = sum(inequality_missing))
```

```
## # A tibble: 39 x 3
##     Year N_level_missing N_inequality_missing
##    <dbl>           <dbl>                <dbl>
##  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
```

As shown in the output above, for the all data set, there are indicators not available very year across the surveyed years.

To get a idea of the country-year relationship of the data availability, the following task is performed.

```
edu_quality |>
  mutate(level_missing = as.numeric(is.na(Education_level)), .afrer = Education_level, inequality_missi
  group_by(Country, Year)|>
  summarise(N_level_missing = sum(level_missing),
            N_inequality_missing = sum(inequality_missing))
```

```
## # A tibble: 6,789 x 4
## # Groups:   Country [181]
##    Country      Year N_level_missing N_inequality_missing
##    <chr>       <dbl>           <dbl>                <dbl>
##  1 Afghanistan  1984               0                    0
##  2 Afghanistan  1985               0                    0
##  3 Afghanistan  1986               0                    0
```

```
##  4 Afghanistan  1987              0                   0
##  5 Afghanistan  1988              0                   0
##  6 Afghanistan  1989              0                   0
##  7 Afghanistan  1990              0                   0
##  8 Afghanistan  1991              0                   0
##  9 Afghanistan  1992              0                   0
## 10 Afghanistan  1993              0                   0
## # i 6,779 more rows
```

ii. Two types of country-level indicators of education quality

a. Average level of education quality from 1984 to 2022

The average level of education quality can be explained by the average level of education and the average level of inequality in education. While the two indicators cannot be simply merged, they are shown seperately below:

```
edu_quality |>
 filter(Year >= 1984 & Year <= 2022)|>
 group_by(Country) |>
 arrange(Year) |>
 summarise(average_level = mean(Education_level, na.rm = TRUE),
 average_inequality = mean(Education_inequality, na.rm = TRUE))
```

```
## # A tibble: 181 x 3
##    Country     average_level average_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
```

b. Change of education quality from 1984 to2022

(1) Change between the first year and the most recent year being surveyed Data regarding the education quality only available from 1984 to 2013, therefore, the following output show the ratio between the education quality in 2013 and the education quality in 1984 to compare the values of that from the most recent year and the earliest year.

```
edu_quality |>
  filter(Year >= 1984 & Year <= 2022) |>
  group_by(Country) |>
  arrange(Year) |>
  summarise(Edu_level_compare = (last(Education_level, na_rm = TRUE) - first(Education_level, na_rm = TR
            Edu_inequality_compare = (last(Education_inequality, na_rm = TRUE) - first(Education_inequal
```

```
  ungroup() |>
  arrange(Country)
```

```
## # A tibble: 181 x 3
##    Country     Edu_level_compare Edu_inequality_compare
##    <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 171 more rows
```

(2) Year-on-year changes of the education equality The following output shows the changes of education quality year by year in each countries.

```
edu_quality |>
group_by(Country) |>
arrange(Year) |>
mutate(Edulevel_yoy_change = Education_level - lag(Education_level, n = 1), Eduinequality_yoy_change = I
ungroup() |>
arrange(Country, Year)
```

```
## # A tibble: 6,789 x 6
##    Country      Year Education_level Education_inequality Edulevel_yoy_change
##    <chr>       <dbl>           <dbl>                <dbl>                <dbl>
##  1 Afghanistan  1984            1.30                 85.4               NA
##  2 Afghanistan  1985            1.35                 84.8                0.0510
##  3 Afghanistan  1986            1.40                 84.8                0.0510
##  4 Afghanistan  1987            1.45                 84.6                0.0510
##  5 Afghanistan  1988            1.50                 84.5                0.0510
##  6 Afghanistan  1989            1.55                 84.1                0.0510
##  7 Afghanistan  1990            1.60                 83.8                0.0510
##  8 Afghanistan  1991            1.69                 82.8                0.091
##  9 Afghanistan  1992            1.78                 81.9                0.0900
## 10 Afghanistan  1993            1.88                 81.0                0.091
## # i 6,779 more rows
## # i 1 more variable: Eduinequality_yoy_change <dbl>
```

iii. Which countries perform the best and the worst

Provided with the data that compare the latest education quality and the earliest education quality, we can see that Nepal improve their national education level for around 3 (exactly 2.78) times during the past four decades, decreasing the educational inequality for about 0.45 compared with the very first data. A backward in national education level is witnessed in Tajikistan for about 0.03 compared with the earliest data.

According toThe World Bank (2023), Nepal can make rapid improvements in its educational quality partly because its School Sector Development Program. However, although progress have been made, Nepal still facing challenges such as the inconsistent education quality and the cultural and household factors that prevent children from school. The relatively worse education conditions may also explain the high improvement rate.

Regarding Tajikistan, a report by The World Bank (2015) pointed out the current increasingly barriers for people there to get into higher educations, including the high personal diffferences that stop them from school, the incresingly high drop-out rate, and the severe educational situation for women. The report revealed that only 13 percent of general secondary students are from the bottom quintile of consumption, and 1 of 3 women stooping their studies before finishing the secondary education. Therefore, with increasing inequality of the decesive factors for education including gender and economic backgrounds, it is plausible that the level of national education is experiencing a backward.