Draft Documents

Team 1 (Hedgehog, Callista, Imtiyaaz, Issac)

2022-10-30

2. Data Characteristic

2.1. Nature of Data

The data set is collection The World Bank Data, the variables of interest are extracted from the raw data files and combined into a single data frame for analysis. The final data set includes:

1. **country.code**: Country code
2. **country.name**: Country name
3. **year**: Year
4. **income**: Income class
    - Low income (L)
    - Lower middle income (LM)
    - Upper middle income (UM)
    - High income (H)
5. **reg**: Region
6. **pov**: Poverty headcount ratio
7. **mpi**: Multidimensional Poverty Index
8. **edu.total**: Total expenditure on education (% of GDP)
9. **edu.pri**: Total expenditure on primary education (% of total education expenditure)
10. **edu.sec**: Total expenditure on secondary education (% of total education expenditure)
11. **edu.ter**: Total expenditure on tertiary education (% of total education expenditure)
12. **hlth**: Total expenditure on health (% of GDP)
13. **mil**: Total expenditure on military (% of GDP)

14. **fdi**: Foreign Direct Investment
15. **lbr.part**: Labour force participation (% of population ages 15+)
16. **unemp**: Unemployment rate
17. **pop.gwth.total**: Total population growth rate
18. **pop.gwth.rural**: Total rural population growth rate
19. **pop.gwth.urban**: Total urban population growth rate
20. **gdp.dflt**: GDP deflator
21. **gdr.eql**: Gender equality rating
22. **gcf**: Gross Capital Formation
23. **trade**: Trade = import + export (% of GDP)

Data imports and combining:

```
# required library
library(knitr)
library(readr)
library(tidyr)
library(dplyr)
library(ggplot2)
```

```
# helper functions
importWDI <- function(filepath, value_name) {
  df <- read_csv(filepath, skip = 4)

  colnames(df) <- tolower(gsub(" ", ".", colnames(df)))

  df <- df %>%
    pivot_longer(5:ncol(.), names_to = "year", values_to = "value") %>%
    filter(!is.null(value) & !is.na(value)) %>%
    mutate(country.code = factor(country.code),
           country.name = factor(country.name),
           year = as.numeric(year)) %>%
    select(country.code, country.name, year, value)

  colnames(df)[4] <- value_name

  df
}

importRegionClass <- function(filepath) {
  df <- read_csv(filepath, skip = 4)

  colnames(df) <- tolower(gsub(" ", ".", colnames(df)))

  df %>% mutate(country.name = factor(country.name),
```

```
                region = factor(region)) %>%
    select(country.name, reg = region)
}

importIncomeClass <- function(filepath) {
  df <- read_csv(filepath, skip = 4)

  colnames(df) <- tolower(gsub(" ", ".", colnames(df)))

  df %>%
    pivot_longer(3:ncol(.), names_to = "year", values_to = "income") %>%
    filter(!is.null(income) & !is.na(income)) %>%
    mutate(country.code = factor(country.code),
           country.name = factor(country.name),
           year = as.numeric(year),
           income = factor(income)) %>%
    select(country.code, country.name, year, income)
}
```

```
# import and combine data
setwd("../data")

poverty.headcount <- importWDI("poverty.headcount.215dollar.csv", "pov")
mpi <- importWDI("mpi.csv", "mpi")
education.expenditure.total <- importWDI("total.education.expenditure.csv", "edu.total")
education.expenditure.primary <- importWDI("primary.education.expenditure.csv", "edu.pri")
education.expenditure.secondary <- importWDI("secondary.education.expenditure.csv", "edu.sec")
education.expenditure.tertiary <- importWDI("tertiary.education.expenditure.csv", "edu.ter")
health.expenditure <- importWDI("health.expenditure.csv", "hlth")
military.expenditure <- importWDI("military.expenditure.csv", "mil")
fdi <- importWDI("fdi.csv", "fdi")
labour.force.participation <- importWDI("labour.force.participation.csv", "lbr.part")
unemployment.rate <- importWDI("unemployment.csv", "unemp")
population.growth <- importWDI("population.growth.csv", "pop.gwth.total")
rural.population.growth <- importWDI("rural.population.growth.csv", "pop.gwth.rural")
urban.population.growth <- importWDI("urban.population.growth.csv", "pop.gwth.urban")
gdp.deflator <- importWDI("gdp.deflator.csv", "gdp.dflt")
gender.equality <- importWDI("gender.equality.csv", "gdr.eql")
gross.capital.formation <- importWDI("gross.capital.formation.csv", "gcf")
trade <- importWDI("trade.csv", "trade")
region.class <- importRegionClass("region.class.csv")
income.class <- importIncomeClass("income.class.csv")

setwd("../src")

countries <- income.class %>%
  full_join(region.class, by = "country.name") %>%
  right_join(poverty.headcount, by = c("country.name", "country.code", "year")) %>%
  full_join(mpi, by = c("country.name", "country.code", "year")) %>%
  left_join(education.expenditure.total, by = c("country.name", "country.code", "year")) %>%
  left_join(education.expenditure.primary, by = c("country.name", "country.code", "year")) %>%
  left_join(education.expenditure.secondary, by = c("country.name", "country.code", "year")) %>%
  left_join(education.expenditure.tertiary, by = c("country.name", "country.code", "year")) %>%
```

```
  left_join(health.expenditure, by = c("country.name", "country.code", "year")) %>%
  left_join(military.expenditure, by = c("country.name", "country.code", "year")) %>%
  left_join(fdi, by = c("country.name", "country.code", "year")) %>%
  left_join(labour.force.participation, by = c("country.name", "country.code", "year")) %>%
  left_join(unemployment.rate, by = c("country.name", "country.code", "year")) %>%
  left_join(population.growth, by = c("country.name", "country.code", "year")) %>%
  left_join(rural.population.growth, by = c("country.name", "country.code", "year")) %>%
  left_join(urban.population.growth, by = c("country.name", "country.code", "year")) %>%
  left_join(gdp.deflator, by = c("country.name", "country.code", "year")) %>%
  left_join(gender.equality, by = c("country.name", "country.code", "year")) %>%
  left_join(gross.capital.formation, by = c("country.name", "country.code", "year")) %>%
  left_join(trade, by = c("country.name", "country.code", "year"))
```

Data preview

```
head(countries)
```

```
## # A tibble: 6 x 23
##   count~1 count~2  year income reg     pov   mpi edu.t~3 edu.pri edu.sec edu.ter
##   <fct>   <fct>   <dbl> <fct>  <fct> <dbl> <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 ALB     Albania  1996 LM     Euro~   0.5    NA    3.08      NA      NA      NA
## 2 ALB     Albania  2002 LM     Euro~   1.1    NA    3.12      NA      NA      NA
## 3 ALB     Albania  2005 LM     Euro~   0.6    NA    3.28      NA      NA      NA
## 4 ALB     Albania  2008 LM     Euro~   0.2    NA   NA         NA      NA      NA
## 5 ALB     Albania  2012 UM     Euro~   0.6    NA    2.93      NA      NA      NA
## 6 ALB     Albania  2014 UM     Euro~   1      NA    3.05      NA      NA      NA
## # ... with 12 more variables: hlth <dbl>, mil <dbl>, fdi <dbl>, lbr.part <dbl>,
## #   unemp <dbl>, pop.gwth.total <dbl>, pop.gwth.rural <dbl>,
## #   pop.gwth.urban <dbl>, gdp.dflt <dbl>, gdr.eql <dbl>, gcf <dbl>,
## #   trade <dbl>, and abbreviated variable names 1: country.code,
## #   2: country.name, 3: edu.total
## # i Use `colnames()` to see all variable names
```

```
str(countries)
```

```
## tibble [2,390 x 23] (S3: tbl_df/tbl/data.frame)
##  $ country.code  : Factor w/ 272 levels "ABW","AFG","AGO",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ country.name  : Factor w/ 280 levels "Afghanistan",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ year          : num [1:2390] 1996 2002 2005 2008 2012 ...
##  $ income        : Factor w/ 4 levels "H","L","LM","UM": 3 3 3 3 4 4 4 4 4 4 ...
##  $ reg           : Factor w/ 7 levels "East Asia & Pacific",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ pov           : num [1:2390] 0.5 1.1 0.6 0.2 0.6 1 0.1 0.1 0.4 0 ...
##  $ mpi           : num [1:2390] NA NA NA NA NA NA NA NA 51.8 49 ...
##  $ edu.total     : num [1:2390] 3.08 3.12 3.28 NA 2.93 ...
##  $ edu.pri       : num [1:2390] NA NA NA NA NA ...
##  $ edu.sec       : num [1:2390] NA NA NA NA NA ...
##  $ edu.ter       : num [1:2390] NA NA NA NA NA ...
##  $ hlth          : num [1:2390] NA 6.91 6.34 5.14 5.06 ...
##  $ mil           : num [1:2390] 1.38 1.32 1.35 1.98 1.49 ...
##  $ fdi           : num [1:2390] 9.01e+07 1.35e+08 2.62e+08 1.25e+09 9.18e+08 ...
##  $ lbr.part      : num [1:2390] 38.8 59.6 34.5 53.2 57 ...
##  $ unemp         : num [1:2390] 12.3 15.8 14.1 13.1 13.4 ...
```

```
##  $ pop.gwth.total: num [1:2390] -0.622 -0.3 -0.512 -0.767 -0.165 ...
##  $ pop.gwth.rural: num [1:2390] -1.55 -2.17 -2.52 -2.92 -2.51 ...
##  $ pop.gwth.urban: num [1:2390] 0.812 2.181 1.826 1.435 1.848 ...
##  $ gdp.dflt      : num [1:2390] 38.17 3.65 3.31 4.12 1.04 ...
##  $ gdr.eql       : num [1:2390] NA NA 4 NA NA NA NA NA NA NA ...
##  $ gcf           : num [1:2390] 18.1 35.3 36.9 35.8 28.3 ...
##  $ trade         : num [1:2390] 44.9 68.5 70.9 77.5 76.5 ...
```

```
summary(countries)
```

```
##   country.code                     country.name       year       income
##  EAS    :  39   East Asia & Pacific      :  39   Min.   :1967   H   :600
##  LCN    :  39   Latin America & Caribbean:  39   1st Qu.:2000   L   :243
##  LMY    :  39   Low & middle income      :  39   Median :2008   LM  :477
##  UMC    :  39   Upper middle income      :  39   Mean   :2006   UM  :406
##  WLD    :  39   World                    :  39   3rd Qu.:2014   NA's:664
##  HIC    :  38   High income              :  38   Max.   :2021
##  (Other):2157   (Other)                  :2157
##                          reg           pov             mpi
##  Europe & Central Asia     :784   Min.   : 0.0   Min.   : 2.37
##  Latin America & Caribbean :402   1st Qu.: 0.5   1st Qu.:18.30
##  Sub-Saharan Africa        :189   Median : 3.6   Median :24.80
##  East Asia & Pacific       :161   Mean   :13.7   Mean   :27.06
##  Middle East & North Africa: 98   3rd Qu.:19.7   3rd Qu.:33.30
##  (Other)                   : 92   Max.   :91.5   Max.   :74.20
##  NA's                      :664   NA's   :68     NA's   :1935
##    edu.total         edu.pri           edu.sec          edu.ter
##  Min.   : 1.033   Min.   : 0.6578   Min.   : 2.724   Min.   : 0.00
##  1st Qu.: 3.475   1st Qu.:24.5817   1st Qu.:30.328   1st Qu.:16.91
##  Median : 4.330   Median :30.9588   Median :35.525   Median :20.34
##  Mean   : 4.451   Mean   :32.0324   Mean   :35.413   Mean   :20.80
##  3rd Qu.: 5.256   3rd Qu.:38.7398   3rd Qu.:40.904   3rd Qu.:24.06
##  Max.   :15.750   Max.   :70.0950   Max.   :71.587   Max.   :50.44
##  NA's   :808      NA's   :1509      NA's   :1512     NA's   :1334
##       hlth             mil              fdi                lbr.part
##  Min.   : 1.718   Min.   : 0.000   Min.   :-3.444e+11   Min.   :30.50
##  1st Qu.: 5.000   1st Qu.: 1.150   1st Qu.: 4.726e+08   1st Qu.:56.52
##  Median : 6.686   Median : 1.649   Median : 3.242e+09   Median :61.12
##  Mean   : 6.840   Mean   : 1.947   Mean   : 6.711e+10   Mean   :60.87
##  3rd Qu.: 8.502   3rd Qu.: 2.373   3rd Qu.: 2.204e+10   3rd Qu.:65.48
##  Max.   :17.733   Max.   :19.385   Max.   : 3.134e+12   Max.   :93.00
##  NA's   :681      NA's   :122      NA's   :28           NA's   :640
##      unemp        pop.gwth.total    pop.gwth.rural    pop.gwth.urban
##  Min.   : 0.250   Min.   :-3.6295   Min.   :-8.5607   Min.   :-4.078
##  1st Qu.: 4.567   1st Qu.: 0.4097   1st Qu.:-0.7193   1st Qu.: 0.694
##  Median : 6.795   Median : 1.1704   Median : 0.1027   Median : 1.834
##  Mean   : 7.981   Mean   : 1.1749   Mean   : 0.1450   Mean   : 1.905
##  3rd Qu.: 9.773   3rd Qu.: 1.9169   3rd Qu.: 1.1044   3rd Qu.: 2.979
##  Max.   :49.700   Max.   : 5.6145   Max.   : 4.5969   Max.   :13.805
##  NA's   :578      NA's   :11        NA's   :25        NA's   :33
##     gdp.dflt          gdr.eql           gcf             trade
##  Min.   : -26.300   Min.   :1.500   Min.   : 0.00   Min.   :  1.378
##  1st Qu.:   1.959   1st Qu.:3.000   1st Qu.:20.08   1st Qu.: 46.813
##  Median :   4.233   Median :3.464   Median :23.13   Median : 63.685
```

```
##  Mean   :  13.886   Mean   :3.481   Mean   :24.07   Mean   : 77.238
##  3rd Qu.:   8.488   3rd Qu.:4.000   3rd Qu.:26.94   3rd Qu.: 95.341
##  Max.   :3333.585   Max.   :5.000   Max.   :69.48   Max.   :380.104
##  NA's   :29         NA's   :1932    NA's   :83      NA's   :84
```

2.2. Missing values

As observed from the summary above, the data set contains a lot of missing values in some of the variables.

```
nCompleteObs <- sum(complete.cases(countries))
print(paste("No. of complete cases:", nCompleteObs))
```

```
## [1] "No. of complete cases: 3"
```

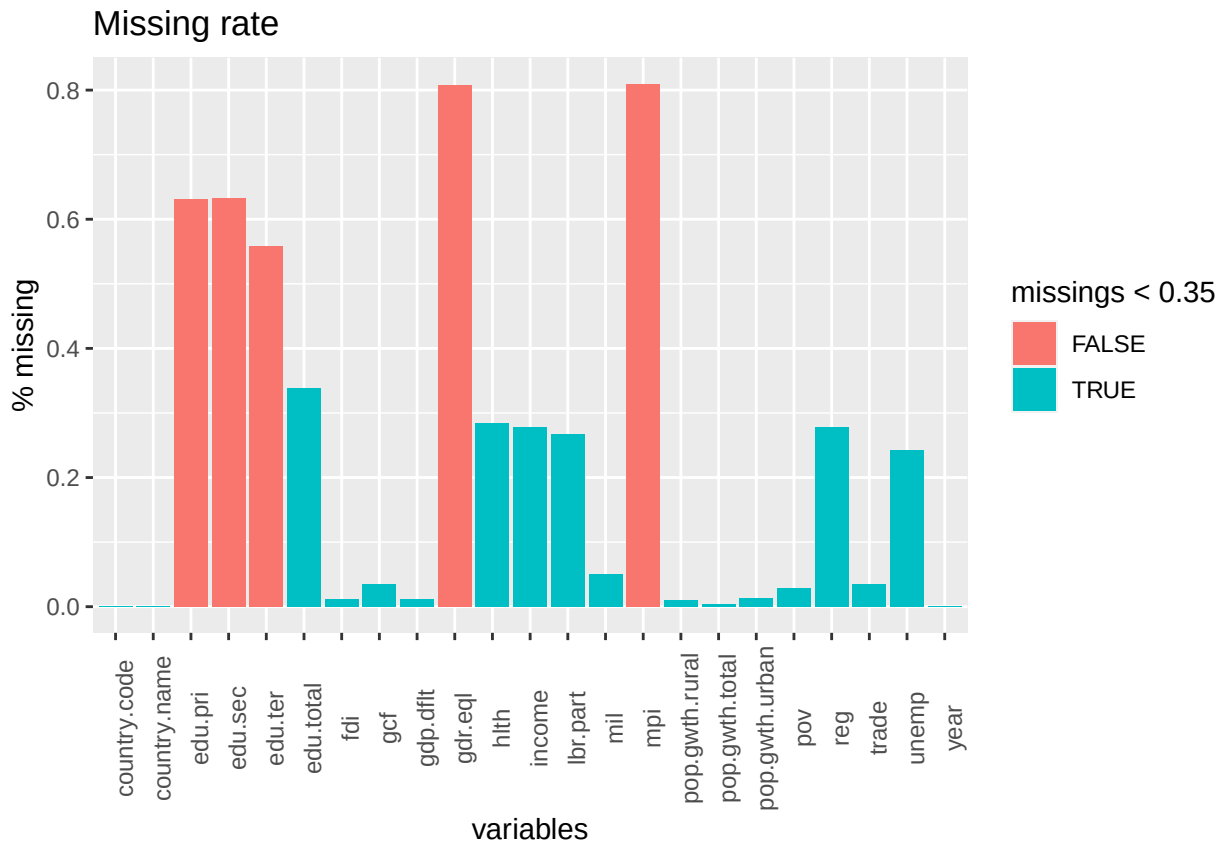
There are only 3 complete cases where all the variable is available. This is nowhere near acceptable to conduct any meaningful analysis. Therefore, we need to eliminate some variables for a more balance data set.

```
mean(is.na(countries))
```

```
## [1] 0.2317628
```

About 23% of the data set is missing.

```
missings <- colMeans(is.na(countries))
ggplot(mapping = aes(x = names(missings), y = missings, fill = missings < 0.35)) +
  geom_bar(stat = "identity") +
  ggtitle("Missing rate") +
  xlab("variables") +
  ylab("% missing") +
  theme(axis.text.x = element_text(size=9, angle=90))
```



There are 5 variables with missing rate >35%.

```
missings[missings > 0.35]
```

```
##       mpi   edu.pri   edu.sec   edu.ter   gdr.eql
## 0.8096234 0.6313808 0.6326360 0.5581590 0.8083682
```

These can be very useful and relevant information (Akbar et al. 2019). However, we would like to exclude these variables from some first analyses to make use of the richer set of data. We can conduct a separate analysis with these variable to gain more insight.

```
# variables with high missing rate
hMiss <- names(missings[missings > 0.35])
countries1 <- countries %>% select(!hMiss)
```

```
## Note: Using an external vector in selections is ambiguous.
## i Use ‘all_of(hMiss)‘ instead of ‘hMiss‘ to silence this message.
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This message is displayed once per session.
```

```
str(countries1)
```

```
## tibble [2,390 x 18] (S3: tbl_df/tbl/data.frame)
##  $ country.code  : Factor w/ 272 levels "ABW","AFG","AGO",..: 4 4 4 4 4 4 4 4 4 4 ...
```

```
##  $ country.name  : Factor w/ 280 levels "Afghanistan",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ year          : num [1:2390] 1996 2002 2005 2008 2012 ...
##  $ income        : Factor w/ 4 levels "H","L","LM","UM": 3 3 3 3 4 4 4 4 4 4 ...
##  $ reg           : Factor w/ 7 levels "East Asia & Pacific",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ pov           : num [1:2390] 0.5 1.1 0.6 0.2 0.6 1 0.1 0.1 0.4 0 ...
##  $ edu.total     : num [1:2390] 3.08 3.12 3.28 NA 2.93 ...
##  $ hlth          : num [1:2390] NA 6.91 6.34 5.14 5.06 ...
##  $ mil           : num [1:2390] 1.38 1.32 1.35 1.98 1.49 ...
##  $ fdi           : num [1:2390] 9.01e+07 1.35e+08 2.62e+08 1.25e+09 9.18e+08 ...
##  $ lbr.part      : num [1:2390] 38.8 59.6 34.5 53.2 57 ...
##  $ unemp         : num [1:2390] 12.3 15.8 14.1 13.1 13.4 ...
##  $ pop.gwth.total: num [1:2390] -0.622 -0.3 -0.512 -0.767 -0.165 ...
##  $ pop.gwth.rural: num [1:2390] -1.55 -2.17 -2.52 -2.92 -2.51 ...
##  $ pop.gwth.urban: num [1:2390] 0.812 2.181 1.826 1.435 1.848 ...
##  $ gdp.dflt      : num [1:2390] 38.17 3.65 3.31 4.12 1.04 ...
##  $ gcf           : num [1:2390] 18.1 35.3 36.9 35.8 28.3 ...
##  $ trade         : num [1:2390] 44.9 68.5 70.9 77.5 76.5 ...
```

Re-evaluate the `countries1` set.

```
mean(is.na(countries1))
```

```
## [1] 0.1050209
```

```
sum(complete.cases(countries1))
```

```
## [1] 916
```

```
mean(complete.cases(countries1))
```

```
## [1] 0.3832636
```

On average, each column has 10% missing rate, that results in 916 complete data point (i.e. 38%).
This can be a sufficient number for the analysis. However, the high missing rate (62%) might introduce some bias to the data set ()

Akbar, Muhammad, Mukaram Khan, Haidar Farooqe, and Kaleemullah. 2019. "Public Spending, Education and Poverty: A Cross Country Analysis" 4 (April): 12–20.