Data Science Project

Aicha Sidiya, Hanin Alzaher, Razan Almahdi

2023-06-01

```
#loading libraries
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.1     v purrr   1.0.1
## v tibble  3.2.1     v dplyr   1.1.0
## v tidyr   1.3.0     v stringr 1.5.0
## v readr   2.1.3     v forcats 1.0.0
```

```
## Warning: package ’tibble’ was built under R version 4.2.3
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(readr)
library(caret)
```

```
## Warning: package ’caret’ was built under R version 4.2.3
```

```
## Loading required package: lattice
##
## Attaching package: ’caret’
##
## The following object is masked from ’package:purrr’:
##
##     lift
```

```
library(RANN)
```

```
## Warning: package ’RANN’ was built under R version 4.2.3
```

```
library(skimr)
```

```
## Warning: package ’skimr’ was built under R version 4.2.3
```

```r
library(ggplot2)
library(stringr)
```

```r
#loading the data set
mortality_rate <- read.csv('data/Mortality rate, under-5 (per 1,000 live births).csv')
health_expenditure <- read.csv('data/Current health expenditure per capita (current US$).csv')
health_expenditure_per <- read.csv('data/Current health expenditure (% of GDP).csv')
education_expenditure <- read.csv('data/Current education expenditure, total (%).csv')
literacy_rate <- read.csv('data/literacy_rate.csv')
domestic_health_expenditure <- read.csv('data/Domestic private health expenditure (% of current health
economic_inequality <- read.csv('data/economic-inequality-gini-index.csv')
water_invest <- read.csv('data/Investment in water and sanitation (current US$).csv')
vacinnation <- read.csv('data/vaccination-coverage-by-income-in.csv')
water_productivity <- read.csv('data/Water productivity_per cubic meter of total freshwater withdrawal.
healthcare_access <- read.csv('data/healthcare-access-and-quality-index.csv')
```

```r
#selecting from year 2000 till 2020
mortality_rate <- select(mortality_rate, country, 'X2000':'X2020')
health_expenditure <- select(health_expenditure, country, 'X2000':'X2020')
health_expenditure_per<- select(health_expenditure_per, country, 'X2000':'X2020')
literacy_rate <- select(literacy_rate, country, 'X2000':'X2020')
education_expenditure <- select(education_expenditure, country, 'X2000':'X2020')
water_invest <- select(water_invest, country, 'X2000':'X2020')
water_productivity <- select(water_productivity, country, 'X2000':'X2020')
domestic_health_expenditure <- select(domestic_health_expenditure, country, 'X2000':'X2020')
economic_inequality <- filter(economic_inequality, year >= 2000)
vacinnation <- filter(vacinnation, year >= 2000)
healthcare_access <- filter(healthcare_access, year >= 2000)
```

```r
#renaming columns
mortality_rate_years <- select (mortality_rate, 'X2000':'X2020')
names(mortality_rate_years) <- str_sub(names(mortality_rate_years),2)
mortality_rate <- select(mortality_rate, country)
mortality_rate <- bind_cols(mortality_rate,mortality_rate_years)

health_expenditure_years <- select (health_expenditure, 'X2000':'X2020')
names(health_expenditure_years) <- str_sub(names(health_expenditure_years),2)
health_expenditure <- select(health_expenditure, country)
health_expenditure <- bind_cols(health_expenditure, health_expenditure_years)

health_expenditure_per_years <- select (health_expenditure_per, 'X2000':'X2020')
names(health_expenditure_per_years) <- str_sub(names(health_expenditure_per_years),2)
health_expenditure_per <- select(health_expenditure_per, country)
health_expenditure_per <- bind_cols(health_expenditure_per, health_expenditure_per_years)

education_expenditure_years <- select (education_expenditure, 'X2000':'X2020')
names(education_expenditure_years) <- str_sub(names(education_expenditure_years),2)
education_expenditure <- select(education_expenditure, country)
education_expenditure <- bind_cols(education_expenditure, education_expenditure_years)

domestic_health_expenditure_years <- select (domestic_health_expenditure, 'X2000':'X2020')
names(domestic_health_expenditure_years) <- str_sub(names(domestic_health_expenditure_years),2)
domestic_health_expenditure <- select(domestic_health_expenditure, country)
domestic_health_expenditure <- bind_cols(domestic_health_expenditure, domestic_health_expenditure_years)

literacy_rate_years <- select (literacy_rate, 'X2000':'X2020')
names(literacy_rate_years) <- str_sub(names(literacy_rate_years),2)
literacy_rate <- select(literacy_rate, country)
literacy_rate <- bind_cols(literacy_rate, literacy_rate_years)

water_invest_years <- select (water_invest, 'X2000':'X2020')
names(water_invest_years) <- str_sub(names(water_invest_years),2)
water_invest <- select(water_invest, country)
water_invest <- bind_cols(water_invest, water_invest_years)

water_productivity_years <- select (water_productivity, 'X2000':'X2020')
names(water_productivity_years) <- str_sub(names(water_productivity_years),2)
water_productivity <- select(water_productivity, country)
water_productivity <- bind_cols(water_productivity, water_productivity_years)
```

```
#pivoting tables
mortality_rate1 <- pivot_longer(mortality_rate, cols="2000":"2020",
                                 names_to = "year",
                                 values_to = "mortality_rate")
health_expenditure1 <- pivot_longer(health_expenditure, cols="2000":"2020",
                                 names_to = "year",
                                 values_to = "health_expenditure")
health_expenditure_per1 <- pivot_longer(health_expenditure_per, cols="2000":"2020",
                                 names_to = "year",
                                 values_to = "health_expenditure_per")
education_expenditure1 <- pivot_longer(education_expenditure, cols="2000":"2020",
                                 names_to = "year",
                                 values_to = "education_expenditure")
domestic_health_expenditure1 <- pivot_longer(domestic_health_expenditure, cols="2000":"2020",
                                 names_to = "year",
                                 values_to = "domestic_health_expenditure")
literacy_rate1 <- pivot_longer(literacy_rate, cols="2000":"2020",
                                 names_to = "year",
                                 values_to = "literacy_rate")
water_invest1 <- pivot_longer(water_invest, cols="2000":"2020",
                                 names_to = "year",
                                 values_to = "water_invest")
water_productivity1 <- pivot_longer(water_productivity, cols="2000":"2020",
                                 names_to = "year",
                                 values_to = "water_productivity")
```

```
#merging data
merge_data <- merge(mortality_rate1, health_expenditure1, by = c("country", "year"), all = TRUE)
merge_data <- merge(merge_data, health_expenditure_per1, by = c("country", "year"), all = TRUE)
merge_data <- merge(merge_data, education_expenditure1, by = c("country", "year"), all = TRUE)
merge_data <- merge(merge_data, domestic_health_expenditure1, by = c("country", "year"), all = TRUE)
merge_data <- merge(merge_data, literacy_rate1, by = c("country", "year"), all = TRUE)
merge_data <- merge(merge_data, water_invest1, by = c("country", "year"), all = TRUE)
merge_data <- merge(merge_data, water_productivity1, by = c("country", "year"), all = TRUE)
merge_data <- merge(merge_data, vacinnation, by = c("country", "year"), all = TRUE)
```

```
skimmed <- skim_to_wide(merge_data)
```

```
## Warning: ’skim_to_wide’ is deprecated.
## Use ’skim()’ instead.
## See help("Deprecated")
```

```
all_countries <- c("Afghanistan", "Albania", "Algeria", "Andorra", "Angola", "Antigua and Barbuda",
    "Argentina", "Armenia", "Australia", "Austria", "Azerbaijan", "Bahamas", "Bahrain",
    "Bangladesh", "Barbados", "Belarus", "Belgium", "Belize", "Benin", "Bhutan",
    "Bolivia", "Bosnia and Herzegovina", "Botswana", "Brazil", "Brunei", "Bulgaria",
    "Burkina Faso", "Burundi", "Cabo Verde", "Cambodia", "Cameroon", "Canada",
    "Central African Republic", "Chad", "Chile", "China", "Colombia", "Comoros",
    "Congo", "Costa Rica", "Croatia", "Cuba", "Cyprus", "Czech Republic", "Denmark",
    "Djibouti", "Dominica", "Dominican Republic", "East Timor", "Ecuador", "Egypt",
    "El Salvador", "Equatorial Guinea", "Eritrea", "Estonia", "Eswatini", "Ethiopia",
    "Fiji", "Finland", "France", "Gabon", "Gambia", "Georgia", "Germany", "Ghana",
    "Greece", "Grenada", "Guatemala", "Guinea", "Guinea-Bissau", "Guyana", "Haiti",
    "Honduras", "Hungary", "Iceland", "India", "Indonesia", "Iran", "Iraq", "Ireland",
    "Israel", "Italy", "Jamaica", "Japan", "Jordan", "Kazakhstan", "Kenya", "Kiribati",
    "Korea, North", "Korea, South", "Kosovo", "Kuwait", "Kyrgyzstan", "Laos", "Latvia",
    "Lebanon", "Lesotho", "Liberia", "Libya", "Liechtenstein", "Lithuania", "Luxembourg",
    "Madagascar", "Malawi", "Malaysia", "Maldives", "Mali", "Malta", "Marshall Islands",
    "Mauritania", "Mauritius", "Mexico", "Micronesia", "Moldova", "Monaco", "Mongolia",
    "Montenegro", "Morocco", "Mozambique", "Myanmar", "Namibia", "Nauru", "Nepal",
    "Netherlands", "New Zealand", "Nicaragua", "Niger", "Nigeria", "North Macedonia",
    "Norway", "Oman", "Pakistan", "Palau", "Panama", "Papua New Guinea", "Paraguay",
    "Peru", "Philippines", "Poland", "Portugal", "Qatar", "Romania", "Russia", "Rwanda",
    "Saint Kitts and Nevis", "Saint Lucia", "Saint Vincent and the Grenadines", "Samoa",
    "San Marino", "Sao Tome and Principe", "Saudi Arabia", "Senegal", "Serbia", "Seychelles",
    "Sierra Leone", "Singapore", "Slovakia", "Slovenia", "Solomon Islands", "Somalia",
    "South Africa", "South Sudan", "Spain", "Sri Lanka", "Sudan", "Suriname", "Sweden",
    "Switzerland", "Syria", "Taiwan", "Tajikistan", "Tanzania", "Thailand", "Togo",
    "Tonga", "Trinidad and Tobago", "Tunisia", "Turkey", "Turkmenistan", "Tuvalu",
    "Uganda", "Ukraine", "United Arab Emirates", "United Kingdom", "United States",
    "Uruguay", "Uzbekistan", "Vanuatu", "Vatican City", "Venezuela", "Vietnam",
    "Yemen", "Zambia", "Zimbabwe")

merge_data <- subset(merge_data, country %in% all_countries)
```

```
## Rows: 7403 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): country
## dbl (11): year, mortality_rate, health_expenditure, health_expenditure_per, ...
##
## 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.
```