relational data_D

Team D

2022-03-11

```
library(tidyverse)
library(ggplot2)
library(readxl)
library(countrycode)
library(gapminder)
```

```
gdp <- read_excel("API_NY.GDP.PCAP.KD_DS2_en_excel_v2_3731742.xls",
  skip = 3,
  sheet = "Data"
)
le <- read_excel("API_SP.DYN.LE00.IN_DS2_en_excel_v2_3731513.xls",
  skip = 3,
  sheet = "Data"
)
pop <- read_excel("API_SP.POP.TOTL_DS2_en_excel_v2_3759026.xls",
  skip = 3,
  sheet = "Data"
)
```

(1) Import the World Bank data for GDP per capita, life expectancy and population.

```
all(gdp$"Country Code" == le$"Country Code")
```

(2) Is the column with three-letter country codes (second column from the left) the same in all three spreadsheets?

```
## [1] TRUE
```

```
all(pop$"Country Code" == le$"Country Code")
```

```
## [1] TRUE
```

```
all(gdp$"Country Code" == pop$"Country Code")
```

```
## [1] TRUE
```

They are all the same.

```
gdp_n <- gdp |>
  pivot_longer(
    c("1960":"2020"),
    names_to = "year",
```

```r
    values_to = "gdp"
  ) |>
  select("Country Name", "Country Code", "year", "gdp")

le_n <- le |>
  pivot_longer(
    c("1960":"2020"),
    names_to = "year",
    values_to = "le"
  ) |>
  select("Country Name", "Country Code", "year", "le")

pop_n <- pop |>
  pivot_longer(
    c("1960":"2020"),
    names_to = "year",
    values_to = "pop"
  ) |>
  select("Country Name", "Country Code", "year", "pop")


wb <- left_join(gdp_n, le_n) |>
  left_join(pop_n)

glimpse(wb)
```

(3) Merge three spreadsheets into a single tibble `wb`

```
## Rows: 16,226
## Columns: 6
## $ `Country Name` <chr> "Aruba", "Aruba", "Aruba", "Aruba", "Aruba", "Aruba", "~
## $ `Country Code` <chr> "ABW", "ABW", "ABW", "ABW", "ABW", "ABW", "ABW", "ABW",~
## $ year           <chr> "1960", "1961", "1962", "1963", "1964", "1965", "1966",~
## $ gdp            <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ le             <dbl> 65.662, 66.074, 66.444, 66.787, 67.113, 67.435, 67.762,~
## $ pop            <dbl> 54208, 55434, 56234, 56699, 57029, 57357, 57702, 58044,~
```

```r
wb |>
  anti_join(codelist, by = c("Country Code" = "iso3c")) |>
  select("Country Code", "Country Name") |>
  distinct()
```

(4) Perform an anti-join to find out which three-letter country codes in the World Bank spreadsheets do not have a matching code in `codelist`. What are the corresponding 'country names'?

```
## # A tibble: 51 x 2
##    `Country Code` `Country Name`
##    <chr>          <chr>
##  1 AFE            Africa Eastern and Southern
##  2 AFW            Africa Western and Central
##  3 ARB            Arab World
##  4 CEB            Central Europe and the Baltics
##  5 CHI            Channel Islands
```

```
##  6 CSS            Caribbean small states
##  7 EAP            East Asia & Pacific (excluding high income)
##  8 EAR            Early-demographic dividend
##  9 EAS            East Asia & Pacific
## 10 ECA            Europe & Central Asia (excluding high income)
## # ... with 41 more rows
```

Do the results make sense? Yes, the results make sense because the country names showed are not actually names of countries. They are names of groups of countries and these names correspond to many ways of grouping a set of countries like geographical location (i.e. North America, Asia) and income (i.e. High Income, Low Income). Since these are not country names, it makes sense that they do not have a matching code in `codelist`.

```
wb <- wb |>
  semi_join(codelist, by = c("Country Code" = "iso3c"))
```

(5) Use a `dplyr` 'join' function to remove all rows from `wb` that do not match any country code in `codelist`.

```
missing_values <- wb |>
  filter(is.na(gdp) | is.na(le) | is.na(pop)) |>
  group_by(year) |>
  count()

missing_values
```

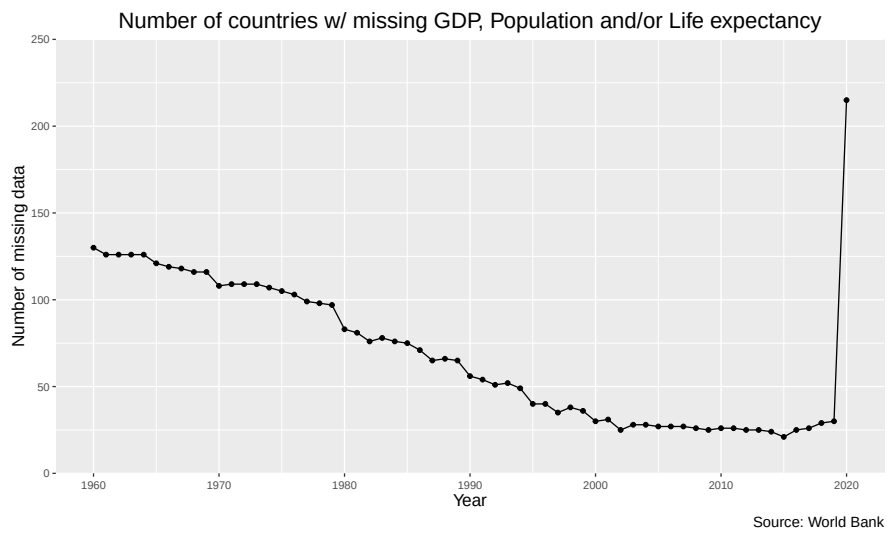
(6) Summarise the number of countries per year that cannot be plotted on the basis of the World Bank data.

```
## # A tibble: 61 x 2
## # Groups:   year [61]
##    year      n
##    <chr> <int>
##  1 1960    130
##  2 1961    126
##  3 1962    126
##  4 1963    126
##  5 1964    126
##  6 1965    121
##  7 1966    119
##  8 1967    118
##  9 1968    116
## 10 1969    116
## # ... with 51 more rows
```

```
m_pl <- missing_values |>
  ggplot(aes(as.integer(year), n)) +
  geom_point() +
  geom_line() +
  theme(axis.text.x = element_text(vjust = 0.5)) +
  labs(
    title = "Number of countries w/ missing GDP, Population and/or Life expectancy",
```

```
    x = "Year",
    y = "Number of missing data",
    caption = "Source: World Bank"
  ) +
  scale_x_continuous(
    breaks = seq(1960, 2020, 10),
    limits = c(1960, 2020),
  ) +
  scale_y_continuous(
    limits = c(0, 250),
    expand = expansion(0)
  ) +
  theme(
    plot.title = element_text(hjust = 0.5, size = 18),
    axis.title.x = element_text(size = 14),
    axis.title.y = element_text(size = 14),
    plot.caption = element_text(size = 12)
  )
m_pl
```

(7) Plot the number of missing countries per year.



Comment on the result. From the graph above, the general trend observed is that the number of countries with missing data decreased over time, with exception of Year 2020. This trend is indicative of the increased ability and capacity of countries to collect census data as they become more developed. Additionally, we observe that there seems to be a periodic cycle between Year 1977 and Year 2003, where the number of missing values decrease before slightly increasing, followed by a decrease again. This might be due to some indicators being derived from sporadic surveys and are only available every few years.[1]

```
gap <-
  gapminder_unfiltered |>
  filter(country != "Netherlands Antilles") |>
  mutate(country_code = countrycode(country, "country.name", "iso3c"))
```

[1] https://datahelpdesk.worldbank.org/knowledgebase/articles/191133-why-are-some-data-not-available. Accessed 16th March 2022.

```r
#Countries without a country code
sum(is.na(gap$country_code))
```

(8) Are there countries in `gap` without a country code? Are there countries that share the same country code?

```
## [1] 0
```

```r
# Obtaining distinct combinations of country name and code
# Then grouping rows by country code and filter out country codes with more
# than one corresponding country name to identify countries with the same code
gap |>
  distinct(country, country_code) |>
  group_by(country_code) |>
  filter(n() > 1) |>
  nrow()
```

```
## [1] 0
```

There is no country in `gap` without a country code. There are no countries in `gap` that share the same country code.

```r
# countries in gap but not in wb
anti_join(gap, wb, by = c("country" = "Country Name")) |>
  distinct(country)
```

(9) Compare data between `gap` and `wb`.

```
## # A tibble: 18 x 1
##    country
##    <fct>
##  1 Bahamas
##  2 Brunei
##  3 Cape Verde
##  4 Egypt
##  5 French Guiana
##  6 Gambia
##  7 Guadeloupe
##  8 Hong Kong, China
##  9 Iran
## 10 Korea, Dem. Rep.
## 11 Macao, China
## 12 Martinique
## 13 Reunion
## 14 Russia
## 15 Swaziland
## 16 Syria
## 17 Taiwan
## 18 Venezuela
```

```r
# countries in wb but not in gap
anti_join(wb, gap, by = c("Country Name" = "country")) |>
  distinct(`Country Name`)
```

```
## # A tibble: 47 x 1
##    `Country Name`
```

```
##    <chr>
##  1 Andorra
##  2 American Samoa
##  3 Antigua and Barbuda
##  4 Bahamas, The
##  5 Bermuda
##  6 Brunei Darussalam
##  7 Cabo Verde
##  8 Curacao
##  9 Cayman Islands
## 10 Dominica
## # ... with 37 more rows
```

```
wb_2007 <- wb |>
  filter(year == "2007") |>
  select("Country Name", "Country Code", "gdp") |>
  drop_na()

gap_2007 <- gap |>
  filter(year == 2007) |>
  select(country, gdpPercap, country_code) |>
  drop_na()

wb_gap <- inner_join(
  wb_2007,
  gap_2007,
  by = c("Country Name" = "country", "Country Code" = "country_code")
  )

wb_gap
```

(10) Compare GDP data in `wb` and `gap` for the year 2007. Merge the information from `wb` and `gap` into a tibble `wb_gap` such that only those countries are included that appear in both tibbles.

```
## # A tibble: 163 x 4
##    `Country Name`       `Country Code`    gdp gdpPercap
##    <chr>                <chr>           <dbl>     <dbl>
##  1 Aruba                ABW            30161.    27231.
##  2 Afghanistan          AFG              393.      975.
##  3 Angola               AGO             3807.     4797.
##  4 Albania              ALB             3045.     5937.
##  5 United Arab Emirates ARE            45389.    36954.
##  6 Argentina            ARG            12919.    12779.
##  7 Armenia              ARM             3093.     4943.
##  8 Australia            AUS            52539.    34435.
##  9 Austria              AUT            43920.    36126.
## 10 Azerbaijan           AZE             4327.     7709.
## # ... with 153 more rows
```

```
wb_gap <- wb_gap |>
  mutate(perc_change = (gdpPercap - gdp) / gdp * 100)
```

(11) Append a column to `wb_gap` that shows the percentage difference of Gapminder's GDP estimate compared to the World Bank estimate.

```
slice_max(wb_gap, perc_change, n = 5)
```

(12) For which five countries is the percentage difference largest? For which five countries is it smallest (i.e. most strongly negative).

```
## # A tibble: 5 x 5
##   `Country Name` `Country Code`   gdp gdpPercap perc_change
##   <chr>          <chr>          <dbl>     <dbl>       <dbl>
## 1 Chad           TCD             656.     1704.        160.
## 2 Ukraine        UKR            2528.     6549.        159.
## 3 Bhutan         BTN            1840.     4745.        158.
## 4 Afghanistan    AFG             393.      975.        148.
## 5 Timor-Leste    TLS             933.     2286.        145.
```

```
slice_min(wb_gap, perc_change, n = 5)
```

```
## # A tibble: 5 x 5
##   `Country Name` `Country Code`    gdp gdpPercap perc_change
##   <chr>          <chr>           <dbl>     <dbl>       <dbl>
## 1 Zimbabwe       ZWE             1042.      470.       -54.9
## 2 Switzerland    CHE            81805.    37506.       -54.2
## 3 Maldives       MDV             8535.     5167.       -39.5
## 4 Norway         NOR            75624.    49357.       -34.7
## 5 Denmark        DNK            53936.    35278.       -34.6
```

The five countries with the greatest percentage difference (in descending order) are: Chad, Ukraine, Bhutan, Afghanistan and Timor-Leste.
The five countries with the most strongly negative percentage difference (in increasing order) are: Zimbabwe, Switzerland, Maldives, Norway and Denmark.