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
```

ANS:
They are all the same.

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

(3) Merge three spreadsheets

```r
wb |>
  anti_join(codelist, by = c("Country Code" = "iso3c")) |>
  select("Country Name") |>
  unique() |>
  head(n = 10)
```

(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’? Do the results make sense?

```
## # A tibble: 10 x 1
##    ‘Country Name‘
##    <chr>
##  1 Africa Eastern and Southern
##  2 Africa Western and Central
##  3 Arab World
##  4 Central Europe and the Baltics
##  5 Channel Islands
##  6 Caribbean small states
##  7 East Asia & Pacific (excluding high income)
##  8 Early-demographic dividend
##  9 East Asia & Pacific
## 10 Europe & Central Asia (excluding high income)
```

ANS:
The names are not country names but are names of regions, of which has a number of countries. The result makes sense as these regions' codes cannot be found in the list of country code.

```r
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.

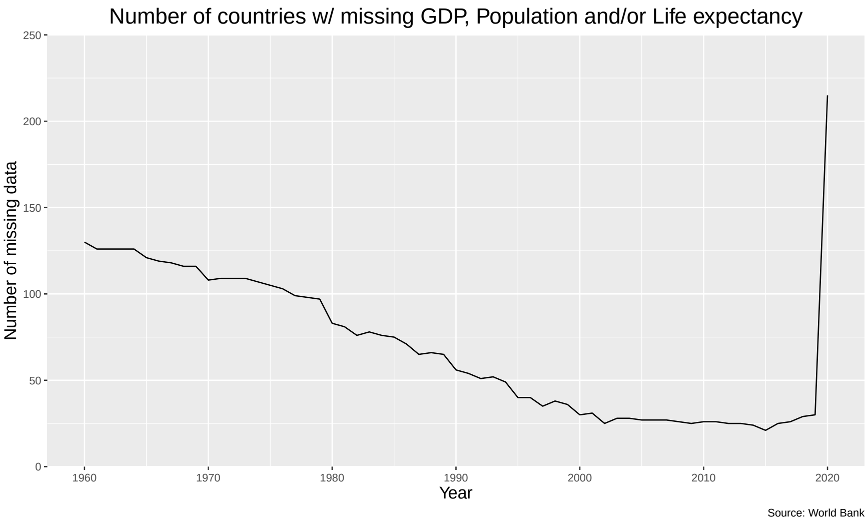
```r
missing_values <- wb |>
  mutate(na = ifelse(is.na(gdp) | is.na(le) | is.na(pop), 1, 0)) |>
  group_by(year) |>
  summarise(na_countries = sum(na))

head(missing_values)
```

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

```
## # A tibble: 6 x 2
##   year  na_countries
##   <chr>        <dbl>
## 1 1960           130
## 2 1961           126
## 3 1962           126
## 4 1963           126
## 5 1964           126
## 6 1965           121
```

```r
m_pl <- missing_values |>
  ggplot(aes(as.integer(year), na_countries)) +
  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)
  )
m_pl
```

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



ANS:
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. Addiionally, throughout the graph, 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"))
sum(is.na(gap$country_code))
```

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

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

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

```
# Arranging the data according to country and country_code,
# filtering all repeated rows such that no two rows will be the same
compare <- gap |>
  select("country", "country_code") |>
  distinct()

# Find the number of countries with the same code.
dim(compare[duplicated(compare$country_code), ])[1]
```

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

ANS:
There is no country in `gap` without a country code. There is no country in `gap` that shares the same country code.

```


# 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
```

```
# 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(c(1, 2, 6, 7)) |>
  drop_na()

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

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

```
wb_gap <- wb_gap |>
  mutate(per_chge = (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.

```
head(wb_gap[order(wb_gap$per_chge, decreasing = TRUE), ], 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 6
##   `Country Name` `Country Code`   gdp continent gdpPercap per_chge
##   <chr>          <chr>          <dbl> <fct>         <dbl>    <dbl>
## 1 Chad           TCD             656. Africa        1704.     160.
## 2 Ukraine        UKR            2528. FSU           6549.     159.
## 3 Bhutan         BTN            1840. Asia          4745.     158.
## 4 Afghanistan    AFG             393. Asia           975.     148.
## 5 Timor-Leste    TLS             933. Asia          2286.     145.
```

```
head(wb_gap[order(wb_gap$per_chge), ], n = 5)
```

```
## # A tibble: 5 x 6
##   `Country Name` `Country Code`    gdp continent gdpPercap per_chge
##   <chr>          <chr>           <dbl> <fct>         <dbl>    <dbl>
## 1 Zimbabwe       ZWE             1042. Africa         470.    -54.9
## 2 Switzerland    CHE            81805. Europe       37506.    -54.2
## 3 Maldives       MDV             8535. Asia          5167.    -39.5
## 4 Norway         NOR            75624. Europe       49357.    -34.7
## 5 Denmark        DNK            53936. Europe       35278.    -34.6
```

ANS:
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.