Capstone

Vinh

2025-06-13

knitr::opts\_chunk$set(warning = TRUE)

# **Phần 1. Gọi packages và đọc dữ liệu**

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.4.1

## Warning: package 'ggplot2' was built under R version 4.4.1

## Warning: package 'tibble' was built under R version 4.4.1

## Warning: package 'tidyr' was built under R version 4.4.1

## Warning: package 'readr' was built under R version 4.4.1

## Warning: package 'purrr' was built under R version 4.4.2

## Warning: package 'dplyr' was built under R version 4.4.1

## Warning: package 'stringr' was built under R version 4.4.1

## Warning: package 'forcats' was built under R version 4.4.1

## Warning: package 'lubridate' was built under R version 4.4.1

library(janitor)

## Warning: package 'janitor' was built under R version 4.4.2

library(gtsummary)

## Warning: package 'gtsummary' was built under R version 4.4.2

library(skimr)

## Warning: package 'skimr' was built under R version 4.4.1

library(scales)

## Warning: package 'scales' was built under R version 4.4.1

df <- read\_csv("E:/R-Together/day1/data/life-exp.csv")

# **Phần 2. Khám phá dữ liệu**

## 2.1. Xem và nhận xét dữ liệu

skimr::skim(df)

Data summary

|  |  |
| --- | --- |
| Name | df |
| Number of rows | 1649 |
| Number of columns | 22 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 2 |
| numeric | 20 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Country | 0 | 1 | 4 | 24 | 0 | 133 | 0 |
| Status | 0 | 1 | 9 | 10 | 0 | 2 | 0 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Year | 0 | 1 | 2007.84 | 4.09 | 2000.00 | 2005.00 | 2008.00 | 2011.00 | 2.015000e+03 | ▆▇▇▇▆ |
| Life expectancy | 0 | 1 | 69.30 | 8.80 | 44.00 | 64.40 | 71.70 | 75.00 | 8.900000e+01 | ▁▂▅▇▁ |
| Adult Mortality | 0 | 1 | 168.22 | 125.31 | 1.00 | 77.00 | 148.00 | 227.00 | 7.230000e+02 | ▇▆▂▁▁ |
| infant deaths | 0 | 1 | 32.55 | 120.85 | 0.00 | 1.00 | 3.00 | 22.00 | 1.600000e+03 | ▇▁▁▁▁ |
| Alcohol | 0 | 1 | 4.53 | 4.03 | 0.01 | 0.81 | 3.79 | 7.34 | 1.787000e+01 | ▇▅▂▂▁ |
| percentage expenditure | 0 | 1 | 698.97 | 1759.23 | 0.00 | 37.44 | 145.10 | 509.39 | 1.896135e+04 | ▇▁▁▁▁ |
| Hepatitis B | 0 | 1 | 79.22 | 25.60 | 2.00 | 74.00 | 89.00 | 96.00 | 9.900000e+01 | ▁▁▁▂▇ |
| Measles | 0 | 1 | 2224.49 | 10085.80 | 0.00 | 0.00 | 15.00 | 373.00 | 1.314410e+05 | ▇▁▁▁▁ |
| BMI | 0 | 1 | 38.13 | 19.75 | 2.00 | 19.50 | 43.70 | 55.80 | 7.710000e+01 | ▅▅▃▇▂ |
| under-five deaths | 0 | 1 | 44.22 | 162.90 | 0.00 | 1.00 | 4.00 | 29.00 | 2.100000e+03 | ▇▁▁▁▁ |
| Polio | 0 | 1 | 83.56 | 22.45 | 3.00 | 81.00 | 93.00 | 97.00 | 9.900000e+01 | ▁▁▁▂▇ |
| Total expenditure | 0 | 1 | 5.96 | 2.30 | 0.74 | 4.41 | 5.84 | 7.47 | 1.439000e+01 | ▂▇▆▂▁ |
| Diphtheria | 0 | 1 | 84.16 | 21.58 | 2.00 | 82.00 | 92.00 | 97.00 | 9.900000e+01 | ▁▁▁▂▇ |
| HIV/AIDS | 0 | 1 | 1.98 | 6.03 | 0.10 | 0.10 | 0.10 | 0.70 | 5.060000e+01 | ▇▁▁▁▁ |
| GDP | 0 | 1 | 5566.03 | 11475.90 | 1.68 | 462.15 | 1592.57 | 4718.51 | 1.191727e+05 | ▇▁▁▁▁ |
| Population | 0 | 1 | 14653625.89 | 70460393.40 | 34.00 | 191897.00 | 1419631.00 | 7658972.00 | 1.293859e+09 | ▇▁▁▁▁ |
| thinness 1-19 years | 0 | 1 | 4.85 | 4.60 | 0.10 | 1.60 | 3.00 | 7.10 | 2.720000e+01 | ▇▃▁▁▁ |
| thinness 5-9 years | 0 | 1 | 4.91 | 4.65 | 0.10 | 1.70 | 3.20 | 7.10 | 2.820000e+01 | ▇▃▁▁▁ |
| Income composition of resources | 0 | 1 | 0.63 | 0.18 | 0.00 | 0.51 | 0.67 | 0.75 | 9.400000e-01 | ▁▁▅▇▅ |
| Schooling | 0 | 1 | 12.12 | 2.80 | 4.20 | 10.30 | 12.30 | 14.00 | 2.070000e+01 | ▁▅▇▅▁ |

* Dữ liệu có bao nhiêu dòng và bao nhiêu cột? Number of rows 1649  
  Number of columns 22
* Dữ liệu gồm những loại biến nào? Column type frequency:  
  character 2  
  numeric 20
* Dữ liệu có cột nào có missing không? Nếu có thì gồm những cột nào? Không có missing  
  ## 2.2. Đổi tên các biến
* Tên các cột có đúng quy tắc chưa? Nếu chưa thì đổi tên cột theo đúng quy tắc. Quy tắc Tên ngắn Không có khoảng trắng (thay thế bằng dấu gạch dưới\_) Không có ký tự lạ (&, #, <, >, …) Thống nhất cách định danh

names(df)

## [1] "Country" "Year"   
## [3] "Status" "Life expectancy"   
## [5] "Adult Mortality" "infant deaths"   
## [7] "Alcohol" "percentage expenditure"   
## [9] "Hepatitis B" "Measles"   
## [11] "BMI" "under-five deaths"   
## [13] "Polio" "Total expenditure"   
## [15] "Diphtheria" "HIV/AIDS"   
## [17] "GDP" "Population"   
## [19] "thinness 1-19 years" "thinness 5-9 years"   
## [21] "Income composition of resources" "Schooling"

df <- clean\_names(df)  
names(df)

## [1] "country" "year"   
## [3] "status" "life\_expectancy"   
## [5] "adult\_mortality" "infant\_deaths"   
## [7] "alcohol" "percentage\_expenditure"   
## [9] "hepatitis\_b" "measles"   
## [11] "bmi" "under\_five\_deaths"   
## [13] "polio" "total\_expenditure"   
## [15] "diphtheria" "hiv\_aids"   
## [17] "gdp" "population"   
## [19] "thinness\_1\_19\_years" "thinness\_5\_9\_years"   
## [21] "income\_composition\_of\_resources" "schooling"

## 2.3. Xoay trục dữ liệu

df <- df %>%  
 pivot\_longer(  
 cols = starts\_with("thinness"),  
 names\_to = "age\_group",  
 values\_to = "thinness"  
 )

# Phần 3. Xử lý và phân tích dữ liệu

## 3.1. Sử dụng df truy xuất ra những cột Measles, Polio, Diphtheria và gán thành df\_diseases. Sau đó in kết quả những dòng đầu tiên

df\_diseases <- df %>% select(measles, polio, diphtheria)  
head(df\_diseases)

## # A tibble: 6 × 3  
## measles polio diphtheria  
## <dbl> <dbl> <dbl>  
## 1 1154 6 65  
## 2 1154 6 65  
## 3 492 58 62  
## 4 492 58 62  
## 5 430 62 64  
## 6 430 62 64

## 3.2. Sử dụng df truy xuất hàng 2, cột 5.

a <- df[2, 5]  
a

## # A tibble: 1 × 1  
## adult\_mortality  
## <dbl>  
## 1 263

## 3.3. Sử dụng df truy nhất hàng 2,5 và cột Measles, Polio.

b <- df[c(2, 5), c("measles", "polio")]  
b

## # A tibble: 2 × 2  
## measles polio  
## <dbl> <dbl>  
## 1 1154 6  
## 2 430 62

## 3.4. Thay thế những giá trị missing của cột Life Expectancy bằng giá trị trung bình của những giá trị

table(is.na(df$life\_expectancy))

##   
## FALSE   
## 3298

df <- df %>%  
 mutate(life\_expectancy = ifelse(  
 is.na(life\_expectancy),  
 mean(life\_expectancy, na.rm = TRUE),  
 life\_expectancy  
 ))

## 3.5. Tạo 1 data frame chứa 2 đối tượng top\_5 và bot\_5 và gán thành df\_top\_bot

# 5 quốc gia có tuổi thọ trung bình cao nhất năm 2015  
top\_5 <- df %>%  
 filter(year == 2015) %>%  
 distinct(country, .keep\_all = TRUE) %>% # Loại trùng lặp quốc gia  
 slice\_max(order\_by = life\_expectancy, n = 5, with\_ties = T)  
  
# 5 quốc gia có tuổi thọ trung bình thấp nhất năm 2015  
bot\_5 <- df %>%  
 filter(year == 2015) %>%  
 distinct(country, .keep\_all = TRUE) %>%  
 slice\_min(order\_by = life\_expectancy, n = 5, with\_ties = FALSE)  
  
 # Tạo 1 data frame chứa 2 đối tượng top\_5 và bot\_5 và gán thành df\_top\_bot  
df\_top\_bot <- bind\_rows(top\_5, bot\_5)  
head(df\_top\_bot)

## # A tibble: 4 × 22  
## country year status life\_expectancy adult\_mortality infant\_deaths alcohol  
## <chr> <dbl> <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 Albania 2015 Devel… 77.8 74 0 4.6   
## 2 Afghanistan 2015 Devel… 65 263 62 0.01  
## 3 Afghanistan 2015 Devel… 65 263 62 0.01  
## 4 Albania 2015 Devel… 77.8 74 0 4.6   
## # ℹ 15 more variables: percentage\_expenditure <dbl>, hepatitis\_b <dbl>,  
## # measles <dbl>, bmi <dbl>, under\_five\_deaths <dbl>, polio <dbl>,  
## # total\_expenditure <dbl>, diphtheria <dbl>, hiv\_aids <dbl>, gdp <dbl>,  
## # population <dbl>, income\_composition\_of\_resources <dbl>, schooling <dbl>,  
## # age\_group <chr>, thinness <dbl>

## 3.6. Tạo biến mới bmi\_lv là biến thứ tự gồm các nhóm: dưới 18.5, từ 18.5 đến 23, trên 23

df <- df %>% mutate(bmi= case\_when(  
 bmi < 18.5 ~ "Dưới 18.5",  
 bmi >= 18.5 & bmi <= 23 ~ "Từ 18.5 đến 23",  
 bmi > 23 ~ "Trên 23",  
 TRUE ~ NA\_character\_) # Nếu bị thiếu dữ liệu  
 )

## 3.7. Sử dụng df nhóm theo country để tính trung bình GDP và gán thành df\_gdp\_by\_country

df\_gdp\_by\_country <- df %>%  
 group\_by(country) %>%  
 summarise(mean\_gdp = mean(gdp, na.rm = TRUE))

# **Phần 4: Tạo bảng thống kê năm 2015**

Nhập lại dữ liệu

## Tạo bảng kết quả

tbl <- df %>%  
 select(status, life\_expectancy, adult\_mortality, gdp, schooling, bmi) %>%  
 tbl\_summary(  
 by = status,  
 label = list(  
 life\_expectancy ~ "\*\*Tuổi thọ\*\*",  
 adult\_mortality ~ "\*\*Tử vong người lớn\*\*",  
 gdp ~ "\*\*GDP\*\*",  
 schooling ~ "\*\*Đi học\*\*",  
 bmi ~ "\*\*BMI\*\*"  
 ),  
 missing = "no",  
 statistic = c(  
 all\_continuous() ~ "{mean} [{min}-{max}]",  
 bmi ~ "{median} ({p25}-{p75})"  
 )) %>%   
 add\_overall() %>%  
 add\_p(pvalue\_fun = ~style\_pvalue(.x, digits = 3)) %>%  
 bold\_labels() %>%  
 modify\_caption("\*\*Bảng mô tả các đặc trưng theo Status\*\*")  
# In bảng kết quả  
tbl

| **Characteristic** | **Overall** N = 1,649*1* | **Developed** N = 242*1* | **Developing** N = 1,407*1* | **p-value***2* |
| --- | --- | --- | --- | --- |
| **\*\*Tuổi thọ\*\*** | 69 [44-89] | 79 [70-89] | 68 [44-89] | <0.001 |
| **\*\*Tử vong người lớn\*\*** | 168 [1-723] | 84 [1-229] | 183 [1-723] | <0.001 |
| **\*\*GDP\*\*** | 5,566 [2-119,173] | 18,977 [12-119,173] | 3,259 [2-52,497] | <0.001 |
| **\*\*Đi học\*\*** | 12.12 [4.20-20.70] | 15.57 [11.50-20.70] | 11.53 [4.20-17.30] | <0.001 |
| **\*\*BMI\*\*** | 44 (20-56) | 58 (55-61) | 36 (19-53) | <0.001 |
| *1*Mean [Min-Max]; Median (Q1-Q3) | | | | |
| *2*Wilcoxon rank sum test | | | | |

# Phần 5. Trực quan hoá dữ liệu

## 5.1. Plot 1

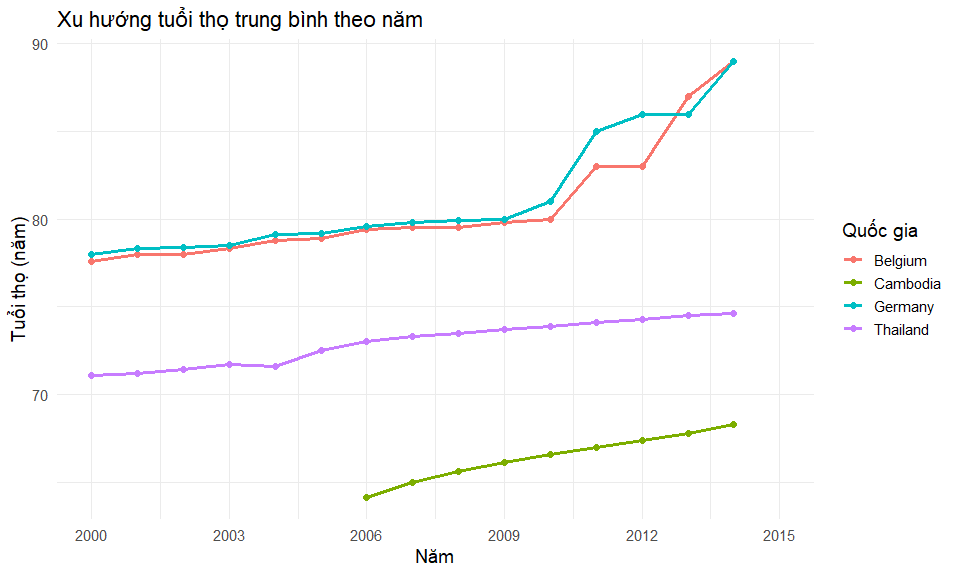
* Chuẩn bị data

plot5 <- df %>%  
 select(country, life\_expectancy, year) %>%  
 filter(country %in% c("Belgium","Germany", "Thailand", "Cambodia")) %>%  
 group\_by(country, year)

* Vẽ biểu đồ

ggplot(plot5, aes(x = year, y = life\_expectancy, color = country)) +  
 geom\_line(size = 1.2) +  
 geom\_point(size = 2) +  
 scale\_x\_continuous(  
 limits = c(2000, 2015), # Giới hạn trục x từ 2000 đến 2015  
 breaks = seq(2000, 2015, by = 3) # Nhãn năm cách nhau mỗi 3 năm  
 ) +  
 labs(  
 title = "Xu hướng tuổi thọ trung bình theo năm",  
 x = "Năm",  
 y = "Tuổi thọ (năm)",  
 color = "Quốc gia"  
 ) +  
 theme\_minimal(base\_size = 14)

## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.  
## ℹ Please use `linewidth` instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.



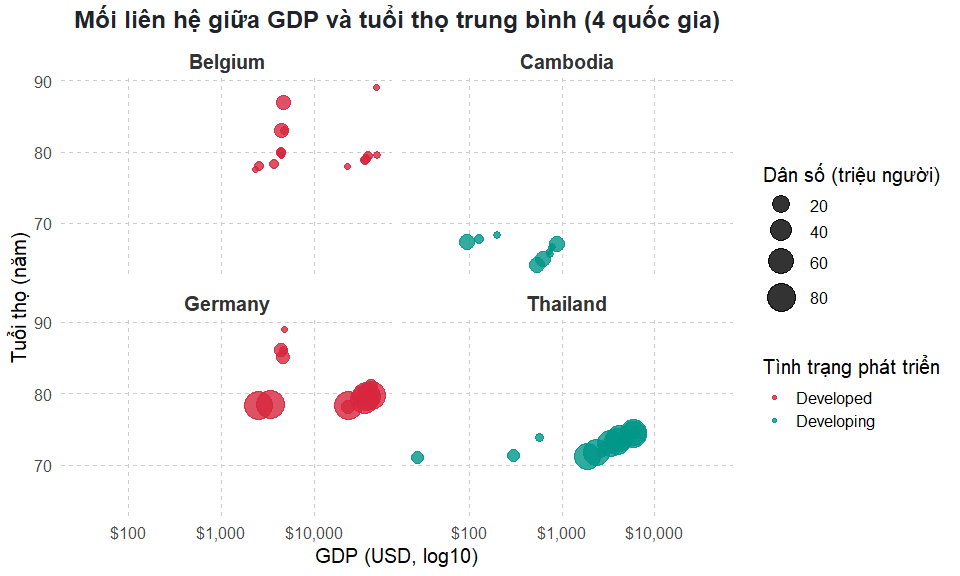
## 5.2. Plot 2

* Chuẩn bị data

df\_sub <- df %>%  
 filter(country %in% c("Belgium","Germany", "Thailand", "Cambodia")) %>%  
 filter(gdp > 0)

* Vẽ biểu đồ

ggplot(df\_sub, aes(  
 x = gdp,   
 y = life\_expectancy,   
 color = status,   
 size = population/1e6  
)) +  
 geom\_point(alpha = 0.8) +  
 scale\_x\_log10(labels = dollar\_format()) +  
 scale\_color\_manual(  
 name = "Tình trạng phát triển",  
 values = c("Developed" = "#d7263d", "Developing" = "#009688")  
 ) +  
 scale\_size\_continuous(  
 name = "Dân số (triệu người)",  
 range = c(2, 10)  
 ) +  
 facet\_wrap(~ country, ncol = 2, strip.position = "top") +  
 labs(  
 title = "Mối liên hệ giữa GDP và tuổi thọ trung bình (4 quốc gia)",  
 x = "GDP (USD, log10)",  
 y = "Tuổi thọ (năm)"  
 ) +  
 theme\_minimal(base\_size = 15) +  
 theme(  
 strip.text = element\_text(face = "bold", size = 15, color = "#333"),  
 plot.title = element\_text(face = "bold", size = 18, hjust = 0.5, color = "#20232a"),  
 legend.position = "right",  
 legend.background = element\_rect(fill = "white", color = NA),  
 panel.grid.major = element\_line(linetype = "dashed", color = "grey80"),  
 panel.grid.minor = element\_blank()  
 )



# Phần 6. Function

plot\_country <- function(df, country\_name, indicator) {  
 # Kiểm tra country\_name có trong df không  
 if (!(country\_name %in% unique(df$country))) {  
 stop("Tên quốc gia không tồn tại trong dữ liệu!")  
 }  
 # Kiểm tra indicator có trong df không  
 if (!(indicator %in% names(df))) {  
 stop("Chỉ số không tồn tại trong dữ liệu!")  
 }  
   
 df\_sub <- df %>%  
 filter(country == country\_name) %>%  
 filter(gdp > 0)  
   
 p <- ggplot(df\_sub, aes(  
 x = gdp,   
 y = .data[[indicator]], # Truy xuất biến động  
 color = status,   
 size = population / 1e6  
 )) +  
 geom\_point(alpha = 0.8) +  
 scale\_x\_log10(labels = dollar\_format()) +  
 scale\_color\_manual(  
 name = "Tình trạng phát triển",  
 values = c("Developed" = "#d7263d", "Developing" = "#009688")  
 ) +  
 scale\_size\_continuous(  
 name = "Dân số (triệu người)",  
 range = c(2, 10)  
 ) +  
 labs(  
 title = paste("Mối liên hệ giữa GDP và", indicator, "của", country\_name),  
 x = "GDP (USD, log10)",  
 y = indicator  
 ) +  
 theme\_minimal(base\_size = 15) +  
 theme(  
 plot.title = element\_text(face = "bold", size = 18, hjust = 0.5, color = "#20232a"),  
 legend.position = "right",  
 legend.background = element\_rect(fill = "white", color = NA),  
 panel.grid.major = element\_line(linetype = "dashed", color = "grey80"),  
 panel.grid.minor = element\_blank()  
 )  
   
 print(p)  
}

Ví dụ

plot\_country(df, "Australia", "schooling")

