ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA KHOA HỌC & KỸ THUẬT MÁY TÍNH

CẤU TRÚC RỜI RẠC CHO KHMT (CO1007)

Thống kê khảo sát kết quả Covid-19 môn Cấu trúc rời rạc

GVHD: Huỳnh Tường Nguyên
Nguyễn Ngọc Lễ
SV thực hiện: Đỗ Văn Bâng – 2110813
Võ Văn Dũng – 2110102
Trần Hoàng Đại Sơn – 2110509
Nguyễn Tấn Dũng – 2110098
Hà Đình Nghĩa – 2114174
Trần Minh Khoa – 2110278

Tp. Hồ Chí Minh, Tháng 03/2022

Mục lục

1 Động cơ nghiên cứu

Bệnh Corona do virus gây ra còn gọi là COVID-19 đã tạo ra những tác động tiêu cực đến nền đời sống của cư dân trên thể giới. Các đợt bùng phát của COVID-19 hay những biến thể virus đã mang đến những thách thức chưa từng có và được dự báo sẽ có tác động đáng kể đến sự phát triển kinh tế. Nhiều thông tin, tin tức về tình hình dịch bệnh cũng như dữ liệu về COVID-19 được phổ biến rộng rải trong đời sống hay trên internet để giúp cho mọi người quan sát, phân tích, nghiên cứu đươc cập nhật hàng ngày.

Phân tích & thống kê dữ liệu về COVID-19 giúp cho ta thấy được số ca nhiễm bệnh, tử vong của một quốc gia, so sánh tình trạng của các quốc gia trong khu vực hay diễn biến dịch trên thế giới. Từ số liệu được báo cáo mơi chúng ta muốn biết các ca nhiễm bệnh có xu hướng tăng lên hay giảm xuống quy mô các đợt bùng phát ở mỗi quốc gia. Dữ liệu dùng cho bài tập lớn có tham khào từ nguồn có thể xử lý trước với một vài thống kê cơ bản trước khi nó được truyền đi để khai thác dữ liệu thông minh sâu hơn.

2 Mục tiêu

Trong bài tập lớn này, các sinh viên sẽ bắt đầu với các bài toán thống kê đơn giản từ những dữ liệu được cung cấp. Qua đó, các em sẽ tìm ra những con số thú vị, có ý nghĩa đối với các dữ liệu thực tế từ tình hình dịch corona. Những kết quả mà các em tìm ra sẽ là bước khởi đầu cho việc khai phá nguồn dữ liệu của hệ thống sau này, nhằm đạt tới mục tiêu nâng cao kỹ năng lập trình, kỹ năng giải quyết vấn đề cho người học, kỹ năng làm việc nhóm cũng như hướng tới mục tiêu cao hơn là đam mê trong làm việc, học tập và nghiên cứu.

3 Kiến thức và kết quả chuẩn bị

3.1 Thống kê và các đặc trưng của mẫu số liệu

3.1.1 Thống kê là gì

Thống kê là khoa học về các phương pháp thu thập, tổ chức, trình bày, phân tích số liệu. Thống kê giúp ta phân tích số liệu một cách khách quan và rút ra các tri thức, thông tin chứa đựng trong các số liệu đó. Trên cơ sở này, chúng ta mới có thể đưa ra những dự báo và quyết định đúng đắn. Thống kê cần thiết cho mọi lực lượng lao động, đặc biệt rất cần cho các ngành quản lí, hoạch định chính sách.

3.1.2 Các đặc trưng của mẫu số liệu

- **Mẫu** là một tập con hữu hạn các đơn vị điều tra.
- **Mẫu số liệu** là các giá trị của dấu hiệu thu được trên một mẫu.
- **Dữ liệu ngoại lệ (outliers)** là một điểm dữ liệu có sự khác biệt đáng kể so với các quan sát khác. Dữ liệu ngoại lệ có thể xuất hiện do sự thay đổi thang đo hoặc do lỗi từ dữ liệu thu thập (thông thường dữ liệu ngoại lệ dạng này sẽ bị loại khỏi tập dữ liệu). Một giá trị ngoại lệ có thể gây ra vấn đề nghiêm trọng trong quá trình phân tích dữ liệu.
- **Kích thước mẫu** N là số phần tử của một mẫu.
- **Tần số** n_i của một giá trị x_i là số lần xuất hiện của giá trị đó trong mẫu.
- **Mốt** của dấu hiệu là số liệu có tần số lớn nhất trong mẫu số liệu, kí hiệu M_o.
- **Tần suất** f_i của giá trị x_i là tỉ số giữa tần số n_i và kích thước mẫu N, thường được biểu diễn dưới dạng phần trăm.

$$f_i = \frac{n_i}{N} \tag{1}$$

- **Số trung bình** của một mẫu số liệu có vai trò đại diện cho các số liệu của mẫu. Giá trị số trung bình của một mẫu kích thước N gồm $m(m \leq N)$ phần tử $x_1, x_2, ..., x_n$ với tần số tương ứng là $n_1, n_2, ..., n_m$ được tính bởi công thức:

$$\overline{x} = \frac{1}{N} \sum_{i=1}^{m} n_i x_i \tag{2}$$

- **Số trung vị** M_e của một mẫu số liệu kích thước N được sắp xếp theo thứ tự không giảm là số liệu đứng ở vị trí thứ $\frac{N+1}{2}$ đối với trường hợp N là số lẻ. Nếu N là số chẵn thì số trung vị là trung bình cộng của hai số liệu ở vị trí thứ $\frac{N}{2}$ và $\frac{N}{2}+1$.

- **Bách phân vị (Percentile)** là đại lượng dùng để ước tính tỷ lệ dữ liệu trong một tập số liệu rơi vào vùng cao hơn hoặc thấp hơn so với một giá trị cho trước. Bách phân vị chia dữ liệu có thứ tự theo hàng trăm.Để xác định giá trị v_p của phân vị thứ p trong một tập dữ liệu, ta thực hiện các bước sau:

  1. Sắp xếp dữ liệu theo thứ tự từ bé đến lớn.
  2. Tính chỉ số i :
  $$i = \frac{p(n+1)}{100}$$
  Trong đó:
     - i là vị trí của giá trị dữ liệu tại phân vị p.
     - p là phân vị thứ p.
     - n là tổng số quan sát.
  3. Xác định vị trí v_p:
     - Nếu i là số nguyên thì phân vị thứ p là giá trị dữ liệu ở vị trí thứi trong tập dữ liệu.
     - Nếu i không phải là số nguyên thì làm tròn i lên và làm tròn i xuống số nguyên gần nhất, sau đó tính trung bình hai giá trị dữ liệu ở hai vị trí này trong tập dữ liệu.

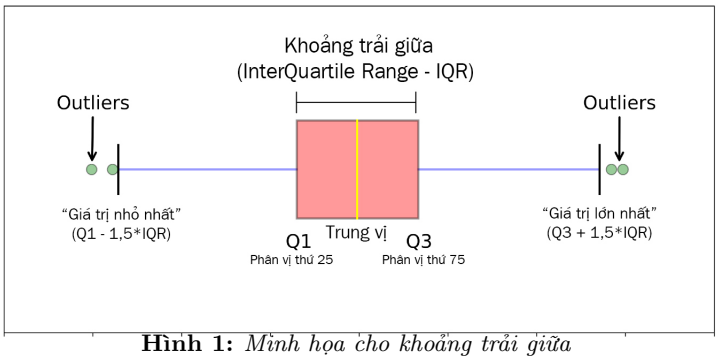
- **Tứ phân vị (Quartile)** là một trường hợp đặc biệt của bách phân vị. Tứ phân vị có 3 giá trị, đó là tứ phân vị thứ nhất, thứ nhì, và thứ ba. Ba giá trị này chia một tập hợp dữ liệu đã sắp xếp theo thứ tự thành 4 phần có số lượng quan sát đều nhau.

  - Tứ phân vị thứ nhất Q_1 bằng trung vị phần dưới, tương đương bách phân vị thứ 25.
  - Tứ phân vị thứ hai Q_2 chính bằng giá trị trung vị, tương đương bách phân vị thứ 50.
  - Tứ phân vị thứ ba Q_3 bằng giá trị trung vị phần trên, tương đương bách phân vị thứ 75.

- **Khoảng trải giữa (Interquartile Range-IQR)** hay còn gọi là khoảng tứ phân vị của tập dữ liệu. Khoảng trải giữa là một con số cho biết mức độ lan truyền của nửa giữa hoặc 50% phần giữa của tập dữ liệu. IQR thường được sử dụng thay cho khoảng biến thiên (Range) vì nó loại trừ hầu hết giá trị bất thường hay giá trị ngoại lệ (Outliers) của dữ liệu. Công thức tính IQR:

$$IQR = Q_3 - Q_1$$

  IQR có thể giúp xác định các giá trị ngoại lệ. Một giá trị bị nghi ngờ là một giá trị ngoại lệ nếu nó nhỏ hơn $1,5 * IQR$ dưới phần tư đầu tiên $(Q1 - 1,5 * IQR)$ hoặc lớn hơn $(1,5 * IQR)$ trên phần tư thứ ba $(Q3 + 1,5 * IQR)$ (Xem hình dưới). Các giá trị ngoại lệ luôn yêu cầu việc rà soát, kiểm tra lại dữ liệu. Những điểm dữ liệu đặc biệt này có thể do lỗi hoặc do sự bất thường trong dữ liệu nhưng cũng có thể là chìa khóa để hiểu dữ liệu.



Hình 1: *Minh họa cho khoảng trải giữa*

- **Phương sai (variance) và độ lệch chuẩn (standard deviation)** đo mức độ phân tán của các số liệu trong mẫu quanh số trung bình. Phương sai và độ lệch chuẩn càng lớn thì mức độ phân tán của số liệu càng lớn. Giả sử ta có mẫu số liệu kích thước N là $x_1, x_2, ..., x_n$. Phương sai của mẫu số liệu này, kí hiệu s^2 được tính bởi công thức:

$$s^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2 \tag{3}$$

  Căn bậc hai của phương sai được gọi là độ lệch chuẩn, kí hiệu s:

$$s = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2} \tag{4}$$

- **Độ méo lệch (skewness)** là một đại lượng đo lường mức độ mức độ bất đối xứng của phân phối xác suất của một biến ngẫu nhiên.

$$\gamma_1 = \frac{\sum_{i=1}^{N}(x_i - \overline{X})^3/N}{s^3} \tag{5}$$

  Trong đó:

  - γ_1 là độ méo lệch.
  - s là độ lệch chuẩn.
  - N là số phần tử của mẫu.
  - $\overline{X}$ là giá trị trung bình của mẫu.
  - x_i là giá trị thứ i của mẫu.

- **Độ nhọn (kurtosis)** là một đại lượng đo mức độ tập trung của phân phối xác suất của một biến ngẫu nhiên, cụ thể là mức độ tập trung của các quan sát quanh trung tâm của phân phối trong mối quan hệ với hai đuôi.

$$\gamma_2 = \frac{\sum_{i=1}^{N}(x_i - \overline{X})^4/N}{s^4} \tag{6}$$

  Trong đó:

  - γ_2 là độ méo lệch.
  - s là độ lệch chuẩn.
  - N là số phần tử của mẫu.
  - $\overline{X}$ là giá trị trung bình của mẫu.
  - x_i là giá trị thứ i của mẫu.

- **Hệ số tương quan Pearson (Pearson correlation coefficient)**: là số liệu thống kê kiểm tra đo lường mối quan hệ thống kê hoặc liên kết giữa các biến phụ thuộc với các biến liên tục. Hệ số tương quan sẽ trả lời cho các câu hỏi chẳng hạn như: Có mối quan hệ tương quan giữa nhiệt độ và doanh thu bán kem? Có mối quan hệ tương quan giữa sự hài lòng công việc, năng suất và thu nhập? Hay hai biến nào có mối liên hệ chặt chẽ nhất giữa tuổi, chiều cao, cân nặng, quy mô gia đình và thu nhập gia đình?
  Hệ số tương quan Pearson (r) có giá trị giao động trong khoảng liên tục từ -1 đến $+1$:
  - $r = 0$: Hai biến không có tương quan tuyến tính.
  - $r = 1$ hoặc $r = -1$: Hai biến có mối tương quan tuyến tính tuyệt đố.i
  - $r < 0$: Hệ số tương quan âm. Nghĩa là giá trị biến x tăng thì giá trị biến y giảm và ngược lại, giá trị biến y tăng thì giá trị biến x giảm.
  - $r > 0$: Hệ số tương quan dương. Nghĩa là giá trị biến x tăng thì giá trị biến y tăng và ngược lại, giá trị biến y tăng thì giá trị biến x cũng tăng.

3.2 Tìm hiểu ngôn ngữ lập trình R

3.2.1 Giới thiệu về R

R là một ngôn ngữ lập trình hàm cấp cao vừa là một môi trường dành cho tính toán thống kê. **R** hỗ trợ rất nhiều công cụ cho phân tích dữ liệu, khám phá tri thức và khai mỏ dữ liệu nhờ đó có thể phát triển nhanh các ứng dụng tính toán xác suất thống kê, phân tích dữ liệu.

Động lực ra đời của **R** là vào khoảng năm 1993, hai nhà thống kê học Ross Ihaka và Robert Gentleman ở University of Auckland (New Zealand) nhận thấy rằng các phần mềm thống kê thương mại sử dụng cho các tính toán thống kê vào thời điểm ấy còn đắt đỏ, không phù hợp và linh hoạt cho mục đích giảng dạy cũng như công việc.Sau đó hai ông đã quyết định lựa chọn ngôn ngữ S được phát triển bởi Bell Laboratories với nỗ lực viết một phần mềm thống kê mới. Cuối cùng Ross Ihaka và Robert Gentleman đã cho ra đời **R** và miễn phí cho tất cả người sử dụng vào năm 1995.

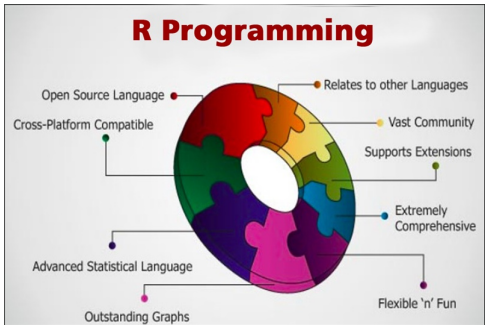
Hình 2: *Logo của **R***

3.2.2 Các điểm nổi bật của R

- **R** có tính mã nguồn mở (open - source), nên liên tục được nâng cấp, cập nhật bởi cộng đồng người sử dụng trên toàn thế giới.
- **R** tương thích đa nền tảng, do đó có thể chạy trên bất kỳ hệ điều hành nào
- **R** tương thích với nhiều ngôn ngữ lập trình khác như C, C++, và FORTRAN. Những ngôn ngữ lập trình khác như .NET, Java, Python cũng có thể thao tác trực tiếp trên đối tượng
- **R** chứa nhiều package cho phép tương tác với nhiều cơ sở dữ liệu. Một số trong đó là Roracle, RmySQL, ...
- **R** có một cộng đồng lớn và tích cực. Mọi người từ khắp nơi trên thế giới có thể giúp đỡ và hỗ trợ. Nhiều ý tưởng và công nghệ mới nhất xuất hiện trong cộng đồng R..

3.2.3 Các kiểu và cấu trúc dữ liệu cơ bản trong R

- **Character/String**: Là kiểu dữ liệu dùng để lưu kí tự.
- **Numeric**: Là kiểu dữ liệu dùng để lưu trữ các số thực trong hệ thập phân.
- **Integer**: Là kiểu dữ liệu dùng để lưu trữ số nguyên.
- **Complex**: Là kiểu dữ liệu dùng để lưu trữ số phức.

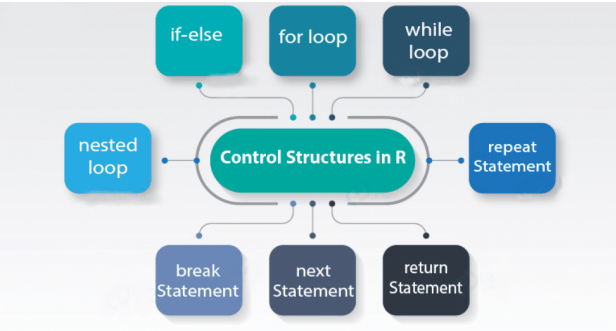


Hình 3: *Ưu điểm nổi bật của* ***R***

- **Logical**: Là kiểu dữ liệu dùng để lưu trữ giá trị kiểu Boolean (0 và 1).
- **Vector**: Là cấu trúc dữ liệu đơn giản nhất trong **R**. Một vecto là một tập hợp các phần tử có cùng kiểu dữ liệu. Một vectơ hỗ trợ các kiểu dữ liệu như logical, integer, double, character, complex, hoặc raw.
- **Matrix**: Là một tập hợp các vecto có cùng độ dài, hay còn gọi là mảng hai chiều hoặc ma trận.
- **List**: Là một tập hợp các phần tử có kiểu dữ liệu mà kiểu dữ liệu của chúng có thể khác nhau. Một phần tử của List có thể là 1 kiểu dữ liệu cơ bản, hoặc có thể là 1 vecto, 1 mattrix, 1 list,...
- **Data Frame**: Là một kiểu dữ liệu cao cấp hơn mattrix, được tích hợp từ nhiều cột hoặc nhiều hàng, tạo thành 1 bảng. Data Frames được tích hợp thêm rất nhiều hàm có ích cho việc thống kê, xử lí số liệu.
- **Array**: Là mảng nhiều chiều. Số chiều của array có thể lớn hơn 2. Người dùng có thể định nghĩa số chiều của Array.
- **Factor**: Là mảng 1 chiều giống vecto nhưng lưu trữ có phân loại. Ta có thể thống kê các giá trị và tần số của các phần tử trong factor.

3.2.4 Các cấu trúc điều khiển trong R

- **Cấu trúc tuần tự**: Đây là cấu trúc có cơ bản trong mọi ngôn ngữ. Trừ khi có rẽ nhánh hoặc vòng lặp, thứ tự thực hiện câu lệnh được là từ trên xuống dưới, một cách tuần tự.
- **Cấu trúc rẽ nhánh có điều kiện**: Tương tự như các ngôn ngữ khác, câu lệnh *if* được dùng để điều khiển việc rẽ nhánh có điều kiện. Các câu lệnh *if* có thể lồng nhau tạo ra *Nested if* và *if* còn có thể đi kèm với *else* hoặc *else if*, giúp cho việc điều khiển rẽ nhánh có điều kiện trở nên hiệu quả hơn. Ngoài ra, *R* còn có câu lệnh *switch* có chức năng tương tự như *switch case* trong ngôn ngữ *C*.
- **Cấu trúc điều khiển vòng lặp**: Tương tự như các các ngôn ngữ lập trình phổ biến, ngôn ngữ cũng có câu lệnh *for* và *while*. Các câu lệnh điều khiển việc kết thúc vòng lặp và rẽ nhánh không điều kiện trong *R* là *break* và *next* (giống continue của *C*). Ngoài ra *R* còn có *repeat* giúp tạo ra vòng lặp không cần điều kiện, chỉ kết thúc khi gặp câu lệnh *next*.



Hình 4: *Cấu trúc điều khiển của* ***R***

3.2.5 Function trong R

- Built-in funtion: Đây là những hàm có thể gọi trực tiếp trong chương trình bởi người dùng như: sq(), mean(), max(), min(),...
- User-defined function: Đây là những hàm do người dùng tự thiết kế.
  - **Cú pháp:**
    function_name = function(arguments) {
        statements
        ...
    }

3.2.6 Package trong R

- **R**-Package là tập hợp các function và dữ liệu mẫu được lưu trữ trong một thư mục gọi là "library" trong môi trường làm việc của **R**.
- Các package này thường được cài đặt tự động trong quá trình cài đặt **R**. Chúng ta cũng có thể cài đặt thêm một số package khác cho mục đích cụ thể.
- Các package của R được chứa trong các repository(kho chứa), trong đó repository phổ biến và đầy đủ nhất là **CRAN** (Comprehensive R Archive Network).
- Đưa package vào library: để sử dụng package nào đó thì chúng ta phải nó lên môi trường làm việc hiện tại của R bằng câu lệnh library("package Name", lib.loc = "path to library").

4 Mô tả dữ liệu

Dữ liệu gồm các thuộc tính chính **"iso_code, continent, location, date, new_cases, new_deaths"** được lưu tron file **csv**.

1. *iso_code*: Định danh đất nước
2. *continent* Tên châu lục
3. *location*: Tên quốc gia
4. *date*: Ngày quan sát với định dạng Month-Day-Year
5. *new_cases*: Số trường hợp COVID-19 mới được xác nhận
6. *new_deaths*: Số tử vong mới do COVID-19

5 Nhiệm vụ

Gọi MD là mã đề riêng cho mỗi nhóm (gồm 4 ký số) không trùng nhau, nhóm sinh viên sẽ thực hiện các yêu cầu dưới đây với các giá trị xác định như sau:

- Mỗi nhóm sẽ dùng R để thao tác trên số file dữ liệu khác nhau được chọn theo cột "STT" theo cách tính $kq = (kytu1 + kytu2 + kytu3 + kytu4)\%6$:
  - Nếu $kq = 0$ thì làm các stt là 1,2,3
  - Nếu $kq = 1$ thì làm các stt là 4,5,6
  - Nếu $kq = 2$ thì làm các stt là 7,8,9
  - Nếu $kq = 3$ thì làm các stt là 10,11,12
  - Nếu $kq = 4$ thì làm các stt là 13,14,15
  - Nếu $kq = 5$ thì làm các stt là 16,17,18

STT	đất nước	STT	đất nước
1	Kenya	10	Canada
2	Lesotho	11	Greenland
3	Morocco	12	United States
4	Indonesia	13	Australia
5	Japan	14	New Caledonia
6	Vietnam	15	New Zealand
7	Andorra	16	Brazil
8	Slovenia	17	Chile
9	United Kingdom	18	Venezuela

Hoàn thành các bài tập:

- Đối với các bài tập chung gồm các phần $\{i, ..., xiii\}$, tất cả các nhóm đều phải làm
- Mỗi nhóm sẽ thực hiện 4 câu riêng của mình trong phần ix bằng các lấy 4 ký số trong mã đề của mình là chỉ số câu hỏi tương ứng. Nếu ký số là 0 thì làm câu 10.

Do MD của nhóm là 6320, nên nhóm sẽ giải quyết vấn đề của:

- Ba quốc gia là Brazil, Chile và Venezuela.
- Đối với phần x nhóm sẽ giải quyết 4 câu là 2, 3, 6 và 10.

i) Nhóm câu hỏi liên quan đến tổng quát dữ liệu
Dùng tập dữ liệu để trả lời các câu hỏi và trình bày theo đinh dạng

1) Tập mẫu thu thập dữ liệu vào các năm nào

- Hiện thực trong R

```r
library("tidyverse")
Data <- read.csv("C:/Users/Asus/Documents/covidData.csv")
years = substring(Data$date,nchar(Data$date)-3,nchar(Data$date))
fac = factor(years)
print("Du lieu duoc thu thap vao nhung nam: ")
print(as.integer(levels(fac)))
```

- Kết quả

```
[1] "Du lieu duoc thu thap vao nhung nam: "
> print(as.integer(levels(fac)))
[1] 2020 2021 2022
```

2) Số lượng đất nước và định danh của mỗi đất nước (hiển thị 10 đất nước đầu tiên).

- Hiện thực trong R

```r
library("tidyverse")
library("gt")
data <- read.csv("owid-covid-data.csv",TRUE,",")
x <- distinct(data.frame((factor(data$iso_code))))
y <- distinct(data.frame(factor(data$location)))
g <- head(cbind(x,y), n = 10)
colnames(g) <- c("iso_code:", "Country")
g$"iso_code:" <- as.character(g$"iso_code:")
g$Country <- as.character(g$Country)
g[nrow(g)+1,] <- c("Count", nrow(x))
gt(g) %>%
cols_align(
  align = c("center"),
  columns = "iso_code:"
)
```

- Kết quả

iso_code:	Country
AFG	Afghanistan
OWID_AFR	Africa
ALB	Albania
DZA	Algeria
AND	Andorra
AGO	Angola
AIA	Anguilla
ATG	Antigua and Barbuda
ARG	Argentina
ARM	Armenia
Count	238

3) Số lượng châu lục trong tập mẫu

- Hiện thực trong R

```
library("tidyverse")
library("gt")
data <- read.csv("owid-covid-data.csv",TRUE,",")
x <- distinct(data.frame(factor(data$continent)))
x <- data.frame(x[-c(2),])
colnames(x) <- c("Continent:")
y <- data.frame(c("Chau A","Chau Au", "Chau Phi", "Chau Bac My", "Chau Nam My","Chau Dai Duong
    "))
colnames(y) <- c("So chau luc:")
gt(cbind(x,y)) %>%
  cols_align(
    align = c("center"),
    columns = "Continent:"
  )
```

- Kết quả

Continent:	So chau luc:
Asia	Chau A
Europe	Chau Au
Africa	Chau Phi
North America	Chau Bac My
South America	Chau Nam My
Oceania	Chau Dai Duong

4) Số lượng dữ liệu thu thập được trong từng từng châu lục và tổng số

- Hiện thực trong R

```
library("gt")
data <- read.csv("owid-covid-data.csv",TRUE,",")
w <- data.frame(table(data$continent))
colnames(w) <- c("Continent:", "Observations")
w <- w[-c(1),]
w$"Continent:" <- as.character(w$"Continent:")
w[nrow(w)+1,] <- c("Tong:", sum(w$Observations))
w$"Continent:" <- as.factor(w$"Continent:")
```

```
gt(w) %>%
  cols_align(
    align = c("center"),
    columns = "Continent:"
  )
```

- Kết quả

Continent:	Observations
Africa	38647
Asia	35528
Europe	36375
North America	24438
Oceania	8993
South America	9335
Tong:	153316

5) Số lượng dữ liệu thu thập được trong từng từng đất nước (hiển thị 10 dất nước cuối cùng) và tổng số

- Hiện thực trong R

```
library("gt")
data <- read.csv("owid-covid-data.csv",TRUE,",")
w <- data.frame(table(data$iso_code))
y = tail(w, n = 10)
colnames(y) <- c("iso_code", "Observations")
y$iso_code <- as.character(y$iso_code)
y[nrow(y)+1,] <- c("Tong:", sum(y$Observations))
y$iso_code <- as.factor(y$iso_code)
gt(y) %>%
  cols_align(
    align = c("center"),
    columns = "iso_code"
  )
```

- Kết quả

iso_code	Observations
VEN	708
VGB	694
VNM	759
VUT	467
WLF	489
WSM	459
YEM	681
ZAF	744
ZMB	704
ZWE	702
Tong:	6407

6) Cho biết các châu lục nào có lượng dữ liệu thu thập nhỏ nhất và giá trị nhó nhất đó?

- Hiện thực trong R

```
data <- read.csv("owid-covid-data.csv",TRUE,",")
w <- data.frame(table(data$continent))
colnames(w) <- c("Continent", "MinimumValue")
print(subset(w,w$MinimumValue == min(w$MinimumValue)))
min(w$Observations)
```

- Kết quả

```
  Continent MinimumValue
6   Oceania         8993
```

7) Cho biết các châu lục nào có lượng dữ liệu thu thập lớn nhất và giá trị lớn nhất đó?

- Hiện thực trong R

```
data <- read.csv("owid-covid-data.csv",TRUE,",")
w <- data.frame(table(data$continent))
colnames(w) <- c("Continent", "MaximumValue")
print(subset(w,w$MaximumValue == max(w$MaximumValue)))
```

- Kết quả

```
  Continent MaximumValue
2    Africa        38647
```

8) Cho biết các nước nào có lượng dữ liệu thu thập nhỏ nhất và giá trị nhó nhất đó?

- Hiện thực trong R

```
data <- read.csv("owid-covid-data.csv",TRUE,",")
w <- data.frame(table(data$location))
colnames(w) <- c("Country", "MinimumValue")
print(subset(w,w$MinimumValue == min(w$MinimumValue)))
```

- Kết quả

```
     Country MinimumValue
171 Pitcairn           85
```

9) Cho biết các nước nào có lượng dữ liệu thu thập lớn nhất và giá trị lớn nhất đó?

- Hiện thực trong R

```
data <- read.csv("owid-covid-data.csv",TRUE,",")
w <- data.frame(table(data$location))
colnames(w) <- c("Country", "MaximumValue")
print(subset(w,w$MaximumValue == max(w$MaximumValue)))
```

- Kết quả

```
      Country MaximumValue
9   Argentina          781
137    Mexico          781
```

10) Cho biết các date nào có lượng dữ liệu thu thập nhỏ nhất và giá trị nhó nhất đó?

- Hiện thực trong R

```
data <- read.csv("owid-covid-data.csv",TRUE,",")
w <- data.frame(table(data$date))
colnames(w) <- c("Date", "MinimumValue")
print(subset(w,w$MinimumValue == min(w$MinimumValue)))
```

- Kết quả

```
       Date MinimumValue
1  1/1/2020            2
```

```
34 1/2/2020            2
67 1/3/2020            2
```

11) Cho biết các date nào có lượng dữ liệu thu thập lớn nhất và giá trị lớn nhất đó?

- Hiện thực trong R

```r
data <- read.csv("owid-covid-data.csv",TRUE,",")
w <- data.frame(table(data$date))
colnames(w) <- c("Date", "MaximumValue")
print(subset(w,w$MaximumValue == max(w$MaximumValue)))
```

- Kết quả

```
          Date MaximumValue
689 8/22/2021           238
691 8/23/2021           238
693 8/24/2021           238
695 8/25/2021           238
697 8/26/2021           238
699 8/27/2021           238
701 8/28/2021           238
703 8/29/2021           238
```

12) Cho biết số lượng dữ liệu thu thập được theo date và châu lục.

- Hiện thực trong R

```r
library("tidyverse")
Data <- read.csv("C:/Users/Asus/Documents/covidData.csv")
continents <- Data$continent[Data$continent != ""]
continents <- c(levels(factor(continents)))
dates <- Data$date[Data$date != ""]
dates <- c(levels(factor(dates)))
size_c = NROW(continents)
size_d = NROW(dates)
size_res = size_c*size_d
res <- data.frame(matrix(NA, nrow = size_res, ncol = 3))
colnames(res) <- c("Continent", "Date", "Observations")
j = 1
k = 1
for (i in 1:size_res) {
  res$Continent[i] <- continents[j]
  res$Date[i] <- dates[k]
  k <- k + 1
  if (i == size_d) j <- j + 1
  if (k == size_d) k <- 1
}
for (i in 1:size_res) {
  temp <- subset(Data, Data$continent == res$Continent[i] & Data$date == res$Date[i])
  res$Observations[i] <- nrow(temp)
}
print(res)
```

- Kết quả sau khi thực hiện chương trình trên ("[...]" biểu thị còn nhiều giá trị do không đủ không gian nên nhóm không nhập vào):

```
    Continent       Date Observations
1      Africa   1/1/2020            0
2      Africa   1/1/2021           55
3      Africa   1/1/2022           55
4      Africa  1/10/2020            0
[...]
329    Africa  2/27/2021           55
330    Africa  2/28/2020            6
331    Africa  2/28/2021           55
```

```
332     Africa  2/29/2020            6
333     Africa   2/3/2020            0
 [ reached 'max' / getOption("max.print") -- omitted 4353 rows
```

13) Cho biết số lượng dữ liệu thu thập được là lớn nhất theo date và châu lục.

- Dựa vào kết quả câu 12, ta tái sử dụng code của câu 12 và dùng câu lệnh dưới đây để thực hiện yêu cầu đề bài:

```
result <- subset(res,res$Observations == max(res$Observations))
row.names(result) <- c(seq(1:NROW(result)))
print(result)
```

- Kết quả sau khi thực hiện như sau:

```
    Continent       Date Observations
1      Africa   1/1/2021           55
2      Africa   1/1/2022           55
3      Africa  1/10/2021           55
4      Africa  1/10/2022           55
[...]
329    Africa  4/13/2021           55
330    Africa  4/14/2021           55
331    Africa  4/15/2021           55
332    Africa  4/16/2021           55
333    Africa  4/17/2021           55
 [ reached 'max' / getOption("max.print") -- omitted 197 rows ]
```

14) Cho biết số lượng dữ liệu thu thập được là nhỏ nhất theo date và châu lục.

- Tương tự câu trên, ta giải quyết yêu cầu bài toán trên như sau:

```
result <- subset(res,res$Observations == min(res$Observations))
row.names(result) <- c(seq(1:NROW(result)))
print(result)
```

- Kết quả

```
   Continent      Date Observations
1     Africa  1/1/2020            0
2     Africa 1/10/2020            0
3     Africa 1/11/2020            0
4     Africa 1/12/2020            0
[...]
48      Asia  1/2/2020            0
49      Asia  1/3/2020            0
50      Asia  1/1/2020            0
51      Asia  1/2/2020            0
52      Asia  1/3/2020            0
53      Asia  1/1/2020            0
```

15) Với một date là k và châu lục t cho trước, hãy cho biết số lượng dữ liệu thu thập được.

- Tương tự ta sử dụng đoạn code ở câu 12, để hiện thực bài toán ta cần thêm các dòng lệnh sau:

```
k = readline("Nhap date ban muon tra cuu: ");
t = readline("Nhap continent ban muon tra cuu: ");
result <- res$Observations[res$Continent == t & res$Date == k]
print("Ket qua la: ")
print(result)
```

- Kết quả trả về màn hình với trường hợp ví dụ k = 11/1/2020, t = Africa:

```
> k = readline("Nhap date ban muon tra cuu: ");
Nhap date ban muon tra cuu: 11/1/2020
> t = readline("Nhap continent ban muon tra cuu: ");
Nhap continent ban muon tra cuu: Africa
> result <- res$Observations[res$Continent == t & res$Date == k]
> print("Ket qua la: ")
[1] "Ket qua la: "
> print(result)
[1] 55
```

16) Có đất nước nào mà số lượng dữ liệu thu thập được là bằng nhau không? Hãy cho biết các iso_code của đất nước đó.

- Hiện thực trong R

```r
library("tidyverse")
Data <- read.csv("C:/Users/Asus/Documents/covidData.csv")
countrys <- Data$location[is.na(Data$iso_code) == 0]
countrys <- c(levels(factor(countrys)))
num_of_data <- c()
for (i in 1:NROW(countrys)) {
  num_of_data[i] <- NROW(subset(Data,Data$location == countrys[i]))
}
isoCode <- c()
for (i in 1:NROW(countrys)) {
  isoCode[i] <- Data$iso_code[Data$location == countrys[i]][i]
}
res <- cbind(isoCode, countrys, num_of_data)
colnames(res) <- c("iso_code", "Country", "Num_of_data")
size_res <- nrow(res)
res <- data.frame(res)
res <- subset(res, is.na(res$iso_code) == 0)
row.names(res) <- c(seq(1:NROW(res)))
eq_data <- c(levels(factor(res$Num_of_data)))
for (i in 1:NROW(eq_data)) {
  temp <- c()
  k <- 1
  t <- FALSE
  for (j in 1:NROW(res)) {
    if (res$Num_of_data[j] == eq_data[i]) {
      temp[k] <- res$iso_code[j]
      k <- k + 1
    }
    if (k > 2) t <- TRUE
  }
  if (t == TRUE) {
    temp <- data.frame(temp)
    temp <- t(temp)
    colnames(temp) <- c(seq(1:NCOL(temp)))
    rownames(temp) <- c(paste(eq_data[i], ":"))
    print(temp)
  }
}
```

- Kết quả được thể hiện theo định dạng: với mỗi hàng (trừ hàng đánh thứ tự), con số đầu tiên biểu thị số lượng dữ liệu thu thập được ở các quốc gia, các chuỗi kí tự tiếp theo là các iso_code của các quốc gia có số lượng dữ liệu thu thập được bằng với con số đầu tiên mỗi hàng. Khi đó ta thấy có nhiều trường hợp số lượng dữ liệu thu thập được ở các quốc gia bằng nhau:

```
      1     2
686 : "SPM" "SSD"
      1     2
691 : "BDI" "SLE"
      1     2     3
694 : "AIA" "VGB" "TCA"
      1     2     3
```

```
697 : "GNB" "MLI" "KNA"
      1     2     3     4     5
700 : "DMA" "GRD" "MOZ" "SYR" "TLS"
      1     2
[...]
      1     2
781 : "ARG" "MEX"
```

17) Liệt kê iso_code, tên đất nước mà chiều dài iso_code lớn hơn 3.

- Hiện thực trong R

```r
library("tidyverse")
Data <- read.csv("C:/Users/Asus/Documents/covidData.csv")
iso_code <- Data$iso_code[is.na(Data$iso_code) == 0]
iso_code <- c(levels(factor(iso_code)))
countrys <- c()
for (i in 1:(NROW(iso_code))) {
  countrys[i] <- Data$location[Data$iso_code == iso_code[i]]
}
res <- cbind(iso_code,countrys)
res <- data.frame(res)
colnames(res) <- c("iso_code", "Country")
result <- subset(res, nchar(res$iso_code) > 3)
row.names(result) <- c(seq(1:NROW(result)))
print(result)
```

- Kết quả

```
   iso_code              Country
1  OWID_AFR               Africa
2  OWID_ASI                 Asia
3  OWID_CYN      Northern Cyprus
4  OWID_EUN       European Union
5  OWID_EUR               Europe
6  OWID_HIC          High income
7  OWID_INT        International
8  OWID_KOS               Kosovo
9  OWID_LIC           Low income
10 OWID_LMC Lower middle income
11 OWID_NAM        North America
12 OWID_OCE              Oceania
13 OWID_SAM        South America
14 OWID_UMC Upper middle income
15 OWID_WRL                World
```

ii) Nhóm câu hỏi liên quan đến mô tả thống kê cơ bản dữ liệu

Với mỗi quốc gia cần tính số liệu thống kê lần lượt cho nhiễm và tử vong do coronavirus được báo cáo mới:

- Chuẩn bị dữ liệu cho toàn bộ phần ii (phần code dùng chung cho cả phần ii)

```r
library(dataset)
library(tidyverse)
library(dplyr)
covid = read.csv("C:/Users/Admin/Downloads/covidData.csv")
covid = covid %>% filter(!continent =='') %>% mutate(new_cases = abs(new_cases), new_deaths= abs(
    new_deaths))
brazil = subset(covid, iso_code == "BRA")
chile = subset(covid, iso_code == "CHL")
venezuela = subset(covid, iso_code == "VEN")
brazil[is.na(brazil)]=0
chile[is.na(chile)]=0
venezuela[is.na(venezuela)]=0
```

```r
Countries = c("Brazil", "Chile", "Venezuela")
```

1) Tính giá trị nhỏ nhất, lớn nhất
   - Hướng giải quyết: Tìm giá trị nhỏ nhất/ lớn nhất bằng min/max.
   - Hiện thực trong R

```r
#for covid new_cases
infections = data.frame(Countries)
infections[["Min"]]<-NA
infections[1, "Min"] = min(brazil$new_cases)
infections[2,"Min"] = min(chile$new_cases)
infections[3, "Min"] = min(venezuela$new_cases)
infections[1, "Max"] = max(brazil$new_cases)
infections[2, "Max"] = max(chile$new_cases)
infections[3, "Max"] = max(venezuela$new_cases)


# for covid new_deaths
deaths = data.frame(Countries)
deaths[["Min"]]<-NA
deaths[1, "Min"] = min(brazil$ new_deaths)
deaths[2,"Min"] = min(chile$ new_deaths)
deaths[3, "Min"] = min(venezuela$ new_deaths)
deaths[1, "Max"] = max(brazil$ new_deaths)
deaths[2, "Max"] = max(chile$ new_deaths)
deaths[3, "Max"] = max(venezuela$ new_deaths)
```

   - Kết quả đối với các ca nhiễm

```
Countries Min    Max
   Brazil   0 287149
    Chile   0  41651
Venezuela   0   4418
```

   - Kết quả đối với các ca tử vong

```
Countries Min   Max
   Brazil   0  4148
    Chile   0  1057
Venezuela   0   35
```

2) Tính tứ phân vị thứ nhất(Q1), thứ hai(Q2), thứ ba(Q3)
   - Hướng giải quyết: Dùng quantile để tìm tứ phân vị.
   - Hiện thực trong R

```r
infections = data.frame(Countries)
infections[["Q1"]] =NA
infections[["Q2"]] =NA
infections[["Q3"]] =NA
for (i in c(2,3,4)) {
  infections[[i]][[1]] = quantile(brazil$new_cases)[[i]]
}
for (i in c(2,3,4)) {
  infections[[i]][[2]] = quantile(chile$new_cases)[[i]]
}
for (i in c(2,3,4)) {
  infections[[i]][[3]] = quantile(venezuela$new_cases)[[i]]
}

deaths = data.frame(Countries)
deaths[["Q1"]] =NA
deaths[["Q2"]] =NA
deaths[["Q3"]] =NA
for (i in c(2,3,4)) {
  deaths[[i]][[1]] = quantile(brazil$new_cases)[[i]]
}
```

```
for (i in c(2,3,4)) {
  deaths[[i]][[2]] = quantile(chile$new_cases)[[i]]
}
for (i in c(2,3,4)) {
  deaths[[i]][[3]] = quantile(venezuela$new_cases)[[i]]
}
```

- Kết quả đối với các ca nhiễm

```
  Countries       Q1      Q2       Q3
1    Brazil 13276.00 31149.0 53989.00
2     Chile  1179.75  1997.5  4315.25
3 Venezuela   245.00   590.0  1100.25
```

- Kết quả đối với các ca tử vong

```
  Countries       Q1      Q2       Q3
1    Brazil 13276.00 31149.0 53989.00
2     Chile  1179.75  1997.5  4315.25
3 Venezuela   245.00   590.0  1100.25
```

3) Tính giá trị trung bình (Avg)

- Hướng giải quyết: Dùng mean để tìm giá trị trung bình.
- Hiện thực trong R

```
infections = data.frame(Countries)
infections["Avg"] =NA
infections[1, "Avg"] = mean(brazil$ new_cases)
infections[2, "Avg"] = mean(chile$ new_cases)
infections[3, "Avg"] = mean(venezuela$ new_cases)

deaths = data.frame(Countries)
deaths["Avg"] =NA
deaths[1, "Avg"] = mean(brazil$ new_deaths)
deaths[2, "Avg"] = mean(chile$ new_deaths)
deaths[3, "Avg"] = mean(venezuela$ new_deaths)
```

- Kết quả đối với các ca nhiễm

```
  Countries        Avg
1    Brazil 38747.6979
2     Chile  3872.5907
3 Venezuela   721.1201
```

- Kết quả đối với các ca tử vong

```
  Countries        Avg
1    Brazil 888.544828
2     Chile  57.173077
3 Venezuela   7.929379
```

4) Tính giá trị độ lệch chuẩn (Std)

- Hướng giải quyết: Dùng sd để tìm giá trị độ lệch chuẩn.
- Hiện thực trong R

```
infections = data.frame(Countries)
infections["Std"] =NA
infections[[2]][[1]] = sd(brazil$new_cases)
infections[[2]][[2]] = sd(chile$new_cases)
infections[[2]][[3]] = sd(venezuela$new_cases)

deaths = data.frame(Countries)
deaths["Std"] =NA
deaths[[2]][[1]] = sd(brazil$new_deaths)
deaths[[2]][[2]] = sd(chile$new_deaths)
deaths[[2]][[3]] = sd(venezuela$new_deaths)
```

- Kết quả đối với các ca nhiễm

```
  Countries        Std
1    Brazil 37015.4040
2     Chile  6093.3562
3 Venezuela   631.3701
```

- Kết quả đối với các ca tử vong

```
  Countries        Std
1    Brazil 771.595312
2     Chile  67.164298
3 Venezuela   7.214482
```

5) Đếm xem có bao nhiêu outliers, một quan sát mà giá trị của nó nằm trong khoảng sau:
$IQR = Q3 - Q1$
$outliers < Q1 - 1.5 * IQR$ hoặc $outliers > Q3 + 1.5 * IQR$

- Hướng giải quyết: Dùng boxplot.stats(...) để tìm các outliers rồi dùng length(...) để đếm xem có bao nhiêu outliers.
- Hiện thực trong R

```r
infections = data.frame(Countries)
infections[["Outlier"]] =NA
infections[[2]][[1]] = length(boxplot.stats(brazil$new_cases)$out)
infections[[2]][[2]] = length(boxplot.stats(chile$new_cases)$out)
infections[[2]][[3]] = length(boxplot.stats(venezuela$new_cases)$out)

deaths = data.frame(Countries)
deaths[["Outlier"]] =NA
deaths[[2]][[1]] = length(boxplot.stats(brazil$new_deaths)$out)
deaths[[2]][[2]] = length(boxplot.stats(chile$new_deaths)$out)
deaths[[2]][[3]] = length(boxplot.stats(venezuela$new_deaths)$out)
```

- Kết quả đối với các ca nhiễm

```
  Countries Outlier
1    Brazil      26
2     Chile      38
3 Venezuela      15
```

- Kết quả đối với các ca tử vong

```
  Countries Outlier
1    Brazil      40
2     Chile      28
3 Venezuela      21
```

6) Lập bảng mô tả số liệu thống kê các đất nước cho từng thể loại:
infections[deaths]:

Countries	Min	Q1	Q2	Q3	Max	Avg	Std	Outlier
ctr_i	?	?	?	?	?	?	?	?

- Hướng giải quyết: Kết hợp thông tin từ các câu trên.
- Hiện thực trong R

```r
infections = data.frame(Countries)
infections[["Min"]]<-NA
infections[1, "Min"] = min(brazil$new_cases)
infections[2,"Min"] = min(chile$new_cases)
infections[3, "Min"] = min(venezuela$new_cases)
infections[["Q1"]] =NA
infections[["Q2"]] =NA
infections[["Q3"]] =NA
infections[["Max"]]=NA
infections[["Avg"]] =NA
infections[["Std"]] =NA
infections[["Outlier"]] =NA

infections[1, "Max"] = max(brazil$new_cases)
```

```
infections[2, "Max"] = max(chile$new_cases)
infections[3, "Max"] = max(venezuela$new_cases)
infections[1, "Avg"] = mean(brazil$new_cases)
infections[2, "Avg"] = mean(chile$new_cases)
infections[3, "Avg"] = mean(venezuela$new_cases)
for (i in c(3,4,5)) {
  infections[[i]][[1]] = quantile(brazil$new_cases)[[i-1]]
}
for (i in c(3,4,5)) {
  infections[[i]][[2]] = quantile(chile$new_cases)[[i-1]]
}
for (i in c(3,4,5)) {
  infections[[i]][[3]] = quantile(venezuela$new_cases)[[i-1]]
}
infections[[9]][[1]] = length(boxplot.stats(brazil$new_cases)$out)
infections[[8]][[1]] = sd(brazil$new_cases)


infections[[9]][[2]] = length(boxplot.stats(chile$new_cases)$out)
infections[[8]][[2]] = sd(chile$new_cases)


infections[[9]][[3]] = length(boxplot.stats(venezuela$new_cases)$out)
infections[[8]][[3]] = sd(venezuela$new_cases)
# for covid new_deaths
deaths = data.frame(Countries)
deaths[["Min"]]<-NA
deaths[1, "Min"] = min(brazil$ new_deaths)
deaths[2,"Min"] = min(chile$ new_deaths)
deaths[3, "Min"] = min(venezuela$ new_deaths)
deaths[["Q1"]] =NA
deaths[["Q2"]] =NA
deaths[["Q3"]] =NA
deaths[["Max"]]=NA
deaths[["Avg"]] =NA
deaths[["Std"]] =NA
deaths[["Outlier"]] =NA

deaths[1, "Max"] = max(brazil$ new_deaths)
deaths[2, "Max"] = max(chile$ new_deaths)
deaths[3, "Max"] = max(venezuela$ new_deaths)
deaths[1, "Avg"] = mean(brazil$ new_deaths)
deaths[2, "Avg"] = mean(chile$ new_deaths)
deaths[3, "Avg"] = mean(venezuela$ new_deaths)
for (i in c(3,4,5)) {
  deaths[[i]][[1]] = quantile(brazil$ new_deaths)[[i-1]]
}
for (i in c(3,4,5)) {
  deaths[[i]][[2]] = quantile(chile$ new_deaths)[[i-1]]
}
for (i in c(3,4,5)) {
  deaths[[i]][[3]] = quantile(venezuela$ new_deaths)[[i-1]]
}

deaths[[9]][[1]] = length(boxplot.stats(brazil$new_deaths)$out)
deaths[[8]][[1]] = sd(brazil$new_deaths)


deaths[[9]][[2]] = length(boxplot.stats(chile$new_deaths)$out)
deaths[[8]][[2]] = sd(chile$new_deaths)


deaths[[9]][[3]] = length(boxplot.stats(venezuela$new_deaths)$out)
deaths[[8]][[3]] = sd(venezuela$new_deaths)
```

- Thống kê cho các ca nhiễm

Countries	Min	Q1	Q2	Q3	Max	Avg	Std	Outlier
Brazil	0	13276.00	31149.0	53989.00	287149	38747.6979	37015.4040	26
Chile	0	1179.75	1997.5	4315.25	41651	3872.5907	6093.3562	38
Venezuela	0	245.00	590.0	1100.25	4418	721.1201	631.3701	15

- Thống kê cho các ca tử vong

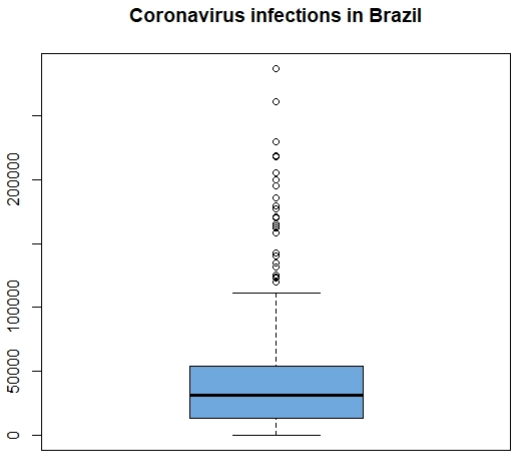
Countries	Min	Q1	Q2	Q3	Max	Avg	Std	Outlier
Brazil	0	317	721	1190	4148	888.544828	771.595312	40
Chile	0	17	40	82	1057	57.173077	67.164298	28
Venezuela	0	3	6	12	35	7.929379	7.214482	21

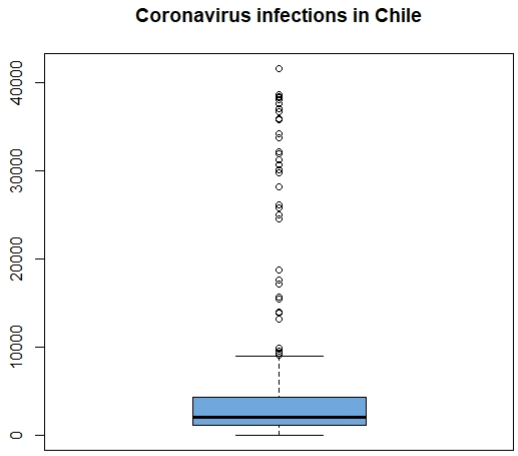
7) Vẽ biểu đồ boxplot hay còn được gọi là box-and-whisker cho nhiễm coronavirus

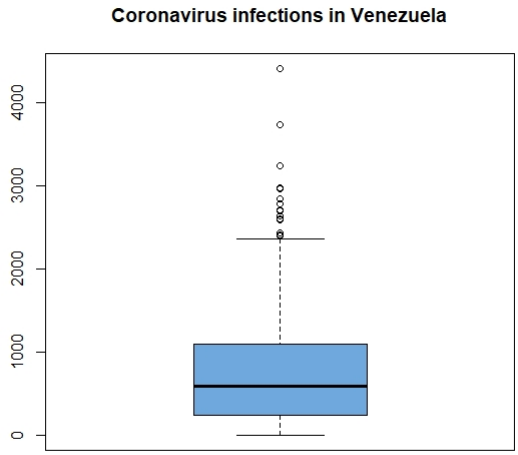
- Hiện thực trong R

```
boxplot(brazil$new_cases, main = "Coronavirus infections in Brazil", col ="#6fa8dc")
boxplot(chile$new_cases, main = "Coronavirus infections in Chile", col ="#6fa8dc")
boxplot(venezuela$new_cases, main = "Coronavirus infections in Venezuela", col ="#6fa8dc" )
```

- Biểu đồ boxplot cho các ca nhiễm:







iii) Nhóm câu hỏi liên quan đến dữ liệu thu thập

Với mỗi quốc gia cần tính số liệu thống kê lần lượt cho nhiễm và tử vong do coronavirus:

Chuẩn bị dữ liệu cho toàn bộ phần iii (phần code dùng chung cho cả phần iii)

```
library(tidyverse)
library(dplyr)
covid = read.csv("C:/Users/Admin/Downloads/covidData.csv")
covid = covid %>% filter(!continent =='') %>% mutate(new_cases = abs(new_cases), new_deaths= abs(
    new_deaths))
brazil = subset(covid, iso_code == "BRA")
chile = subset(covid, iso_code == "CHL")
venezuela = subset(covid, iso_code == "VEN")
brazil[is.na(brazil)]=0
chile[is.na(chile)]=0
venezuela[is.na(venezuela)]=0
Countries = c("Brazil", "Chile", "Venezuela")
```

1) Có bao nhiêu ngày có số lần dữ liệu không được báo cáo mới.
Note: Dữ liệu không được báo cáo mới là missing value (NA) hoặc số liệu bằng không
   - Hướng giải quyết: Đếm các hàng có dữ liệu không được báo cáo mới.
   - Hiện thực trong R

```
none_report = data.frame(Countries)
 none_report[["New cases"]] = NA
 none_report[["New deaths"]] = NA

 none_report[[2]][[1]]= nrow(brazil %>% filter(new_cases ==0))
 none_report[[2]][[2]]= nrow(chile %>% filter(new_cases ==0))
 none_report[[2]][[3]]= nrow(venezuela %>% filter(new_cases ==0))

 none_report[[3]][[1]]= nrow(brazil %>% filter(new_deaths ==0))
 none_report[[3]][[2]]= nrow(chile %>% filter(new_deaths ==0))
 none_report[[3]][[3]]= nrow(venezuela %>% filter(new_deaths ==0))
```

   - Kết quả

```
  Countries New cases New deaths
1    Brazil         9         22
2     Chile        14         33
3 Venezuela        63        129
```

2) Có bao nhiêu ngày có số lần thu thập dữ liệu là thấp nhất được báo cáo mới.
   - Hướng giải quyết: Đếm các hàng có dữ liệu bằng với giá trị nhỏ nhất của số ca nhiễm/ ca tử vong.
   - Hiện thực trong R

```
iii2 = data.frame(Countries)
brazil = brazil %>% filter(!new_cases ==0, !new_deaths==0)
chile = chile %>% filter(!new_cases ==0, !new_deaths==0)
venezuela = venezuela %>% filter(!new_cases ==0, !new_deaths==0)
iii2[["New cases"]]=NA
iii2[["New deaths"]]=NA
iii2[[2]][[1]]=nrow(brazil %>% filter( new_cases == min(brazil$new_cases)))
iii2[[2]][[2]]=nrow(chile %>% filter( new_cases == min(chile$new_cases)))
iii2[[2]][[3]]=nrow(venezuela %>% filter( new_cases == min(venezuela$new_cases)))
iii2[[3]][[1]]=nrow(brazil %>% filter( new_deaths == min(brazil$new_deaths)))
iii2[[3]][[2]]=nrow(chile %>% filter( new_deaths == min(chile$new_deaths)))
iii2[[3]][[3]]=nrow(venezuela %>% filter( new_deaths == min(venezuela$new_deaths)))
```

   - Kết quả

```
  Countries New cases New deaths
1    Brazil         1          1
2     Chile         1         11
3 Venezuela         3         14
```

3) Có bao nhiêu ngày có số lần thu thập dữ liệu là cao nhất được báo cáo mới
Tương tự câu 2

- Hiện thực trong R

```
iii3 = data.frame(Countries)
brazil = brazil %>% filter(!new_cases ==0, !new_deaths==0)
chile = chile %>% filter(!new_cases ==0, !new_deaths==0)
venezuela = venezuela %>% filter(!new_cases ==0, !new_deaths==0)
iii3[["New cases"]]=NA
iii3[["New deaths"]]=NA
iii3[[2]][[1]]=nrow(brazil %>% filter( new_cases == max(brazil$new_cases)))
iii3[[2]][[2]]=nrow(chile %>% filter( new_cases == max(chile$new_cases)))
iii3[[2]][[3]]=nrow(venezuela %>% filter( new_cases == max(venezuela$new_cases)))
iii3[[3]][[1]]=nrow(brazil %>% filter( new_deaths == max(brazil$new_deaths)))
iii3[[3]][[2]]=nrow(chile %>% filter( new_deaths == max(chile$new_deaths)))
iii3[[3]][[3]]=nrow(venezuela %>% filter( new_deaths == max(venezuela$new_deaths)))
```

- Kết quả

```
  Countries New cases New deaths
1    Brazil         1          1
2     Chile         1          1
3 Venezuela         1          3
```

4) Thể hiện bảng số liệu như sau:
Không được báo cáo mới:

Countries ctr_i	Infections value	Deaths value

Báo cáo mới:

Countries ctr_i	Infections value	Deaths value

- Hướng giải quyết: Đếm số ngày không có dữ liệu/ có dữ liệu về các ca nhiễm mới/ ca tử vong mới.
- Hiện thực trong R

```
unreport = data.frame(Countries)
brazil = subset(covid, iso_code == "BRA")
chile = subset(covid, iso_code == "CHL")
venezuela = subset(covid, iso_code == "VEN")
brazil[is.na(brazil)]=0
chile[is.na(chile)]=0
venezuela[is.na(venezuela)]=0
unreport[['Infections value']] = NA
unreport[['Deaths value']] = NA
unreport[[2]][[1]] = nrow(brazil %>% filter(new_cases ==0))
unreport[[2]][[2]] = nrow(chile %>% filter(new_cases ==0))
unreport[[2]][[3]] = nrow(venezuela %>% filter(new_cases ==0))
unreport[[3]][[1]] = nrow(brazil %>% filter(new_deaths ==0))
unreport[[3]][[2]] = nrow(chile %>% filter(new_deaths ==0))
unreport[[3]][[3]] = nrow(venezuela %>% filter(new_deaths ==0))

report = data.frame(Countries)
brazil = subset(covid, iso_code == "BRA")
chile = subset(covid, iso_code == "CHL")
venezuela = subset(covid, iso_code == "VEN")
brazil[is.na(brazil)]=0
chile[is.na(chile)]=0
venezuela[is.na(venezuela)]=0
report[['Infections value']] = NA
report[['Deaths value']] = NA
report[[2]][[1]] = nrow(brazil %>% filter(!new_cases ==0))
report[[2]][[2]] = nrow(chile %>% filter(!new_cases ==0))
report[[2]][[3]] = nrow(venezuela %>% filter(!new_cases ==0))
report[[3]][[1]] = nrow(brazil %>% filter(!new_deaths ==0))
report[[3]][[2]] = nrow(chile %>% filter(!new_deaths ==0))
report[[3]][[3]] = nrow(venezuela %>% filter(!new_deaths ==0))
```

- Không được báo cáo mới

Countries	Infections value	Deaths value
Brazil	9	22
Chile	14	33
Venezuela	63	129

- Báo cáo mới

Countries	Infections value	Deaths value
Brazil	716	703
Chile	714	695
Venezuela	645	579

5) Cho biết số ngày ngắn nhất liên tiếp mà không có dữ liệu được báo cáo

- Hướng giải quyết: Đánh dấu các dữ liệu khác không (tức là dữ liệu được báo cáo) bằng 1, giữ nguyên các dữ liệu bằng không, đếm số dòng có cùng giá trị (bằng 0 hoặc bằng 1) liên tiếp.
- Hiện thực trong R

```
#Nhung ngay co du lieu thi danh dau =1, khong co du lieu danh dau =0
bra = brazil
for (i in 1:nrow(bra)) {
  if(bra[[5]][[i]] != 0 ){
    bra[[5]][[i]] =1
  }
  if(bra[[6]][[i]] != 0 ){
    bra[[6]][[i]] =1
  }
}
chi = chile
for (i in 1:nrow(chi)) {
  if(chi[[5]][[i]] != 0 ){
    chi[[5]][[i]] =1
  }
  if(chi[[6]][[i]] != 0 ){
    chi[[6]][[i]] =1
  }
}
ven = venezuela
for (i in 1:nrow(ven)) {
  if(ven[[5]][[i]] != 0 ){
    ven[[5]][[i]] =1
  }
  if(ven[[6]][[i]] != 0 ){
    ven[[6]][[i]] =1
  }
}
iii5 = data.frame(Countries)
iii5[['New cases']]=NA
iii5[['New deaths']]=NA
iii5[[2]][[1]]= tapply(rle(bra$new_cases)$length, rle(bra$new_cases)$value, min)[[1]][[1]]
iii5[[2]][[2]]= tapply(rle(chi$new_cases)$length, rle(chi$new_cases)$value, min)[[1]][[1]]
iii5[[2]][[3]]= tapply(rle(ven$new_cases)$length, rle(ven$new_cases)$value, min)[[1]][[1]]
iii5[[3]][[1]]= tapply(rle(bra$new_deaths)$length, rle(bra$new_deaths)$value, min)[[1]][[1]]
iii5[[3]][[2]]= tapply(rle(chi$new_deaths)$length, rle(chi$new_deaths)$value, min)[[1]][[1]]
iii5[[3]][[3]]= tapply(rle(ven$new_deaths)$length, rle(ven$new_deaths)$value, min)[[1]][[1]]
```

- Kết quả: Số ngày ngắn nhất liên tiếp mà không có dữ liệu được báo cáo cho các ca nhiễm và các ca tử vong

```
  Countries New cases New deaths
1    Brazil         1          2
2     Chile         1          1
3 Venezuela         1          1
```

6) Cho biết số ngày dài nhất liên tiếp mà không có dữ liệu được báo cáo
Tương tự câu 5, thay đổi min thành max ở 6 dòng cuối

```
iii6 = data.frame(Countries)
iii6[['New cases']]=NA
```

```
iii6[['New deaths']]=NA
iii6[[2]][[1]]= tapply(rle(bra$new_cases)$length, rle(bra$new_cases)$value, max)[[1]][[1]]
iii6[[2]][[2]]= tapply(rle(chi$new_cases)$length, rle(chi$new_cases)$value, max)[[1]][[1]]
iii6[[2]][[3]]= tapply(rle(ven$new_cases)$length, rle(ven$new_cases)$value, max)[[1]][[1]]
iii6[[3]][[1]]= tapply(rle(bra$new_deaths)$length, rle(bra$new_deaths)$value, max)[[1]][[1]]
iii6[[3]][[2]]= tapply(rle(chi$new_deaths)$length, rle(chi$new_deaths)$value, max)[[1]][[1]]
iii6[[3]][[3]]= tapply(rle(ven$new_deaths)$length, rle(ven$new_deaths)$value, max)[[1]][[1]]
```

- Kết quả: Số ngày dài nhất liên tiếp mà không có dữ liệu được báo cáo cho các ca nhiễm và các ca tử vong

```
  Countries New cases New deaths
1    Brazil         3         20
2     Chile         6         28
3 Venezuela         1         24
```

7) Cho biết số ngày ngắn nhất liên tiếp mà không có người nhiễm bệnh mới
Tương tự hai câu trên

- Hiện thực trong R

```
iii7 = data.frame(Countries)
iii7[['New cases']]=NA
iii7[[2]][[1]]= tapply(rle(bra$new_cases)$length, rle(bra$new_cases)$value, min)[[1]][[1]]
iii7[[2]][[2]]= tapply(rle(chi$new_cases)$length, rle(chi$new_cases)$value, min)[[1]][[1]]
iii7[[2]][[3]]= tapply(rle(ven$new_cases)$length, rle(ven$new_cases)$value, min)[[1]][[1]]
```

- Kết quả: Số ngày ngắn nhất liên tiếp mà không có người nhiễm bệnh mới

```
  Countries New cases
1    Brazil         1
2     Chile         1
3 Venezuela         1
```

8) Cho biết số ngày dài nhất liên tiếp mà không có người nhiễm bệnh mới
Tương tự câu trên (thay min thành max)

- Hiện thực trong R

```
   iii8 = data.frame(Countries)
iii8[['New cases']]=NA
iii8[[2]][[1]]= tapply(rle(bra$new_cases)$length, rle(bra$new_cases)$value, max)[[1]][[1]]
iii8[[2]][[2]]= tapply(rle(chi$new_cases)$length, rle(chi$new_cases)$value, max)[[1]][[1]]
iii8[[2]][[3]]= tapply(rle(ven$new_cases)$length, rle(ven$new_cases)$value, max)[[1]][[1]]
```

- Kết quả: Số ngày dài nhất liên tiếp mà không có người nhiễm bệnh mới

```
  Countries New cases
1    Brazil         3
2     Chile         6
3 Venezuela         1
```

iv) Nhóm câu hỏi liên quan đến trực quan dữ liệu

1) Vẽ biểu đồ tần số tích lũy quốc gia cho các châu lục

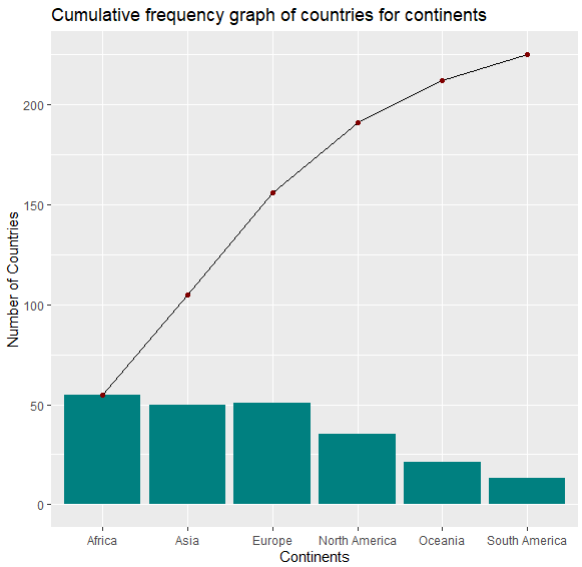
- Tần số tích lũy: Trong một tập dữ liệu, tần số tích lũy đối với một giá trị x là tổng số điểm mà nhỏ hơn hoặc bằng x.
- Biểu đồ tần số tích lũy: biểu thị những thông tin dạng tích lũy. Nó thể hiện số lượng hay tỉ lệ những quan sát nhỏ hơn hoặc bằng một giá trị cụ thể.
- Hiện thực trong R

```
continents = covid  %>% filter(!continent =='') %>% select(iso_code:location) %>% unique() %>%
     group_by(continent) %>% summarize(n = n())
ggplot(data = continents,aes(x = continent, y = n, group =1)) +
  xlab("Continents") +
  ylab("Number of Countries") +
```

```
  geom_line(aes(y = cumsum(n))) + geom_bar(stat="identity",fill = "#008080" ) +
  geom_point(aes(x = continent, y = cumsum(n)), color = "#800000") +
  labs(title = "Cumulative frequency graph of countries for continents" , substitute= "Tan so
    tich luy")

```

- Kết quả:

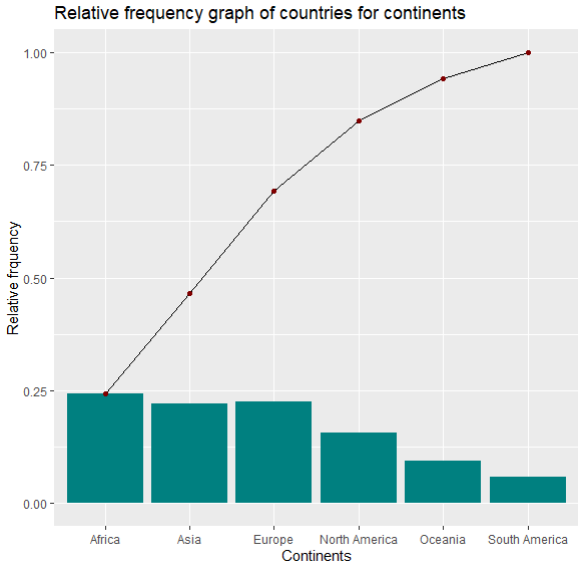


2) Vẽ biểu đồ tần số tương đối quốc gia cho các châu lục

- Biểu đồ tần số tương đối: Việc đếm tần số có thể được thể hiện dưới dạng số tuyệt đối hoặc tương đối (ví dụ phân số hoặc tỉ lệ phần trăm). Biểu đồ dưới đây thể hiện nội dung y nguyên như biểu đồ tần số tích lũy ở trên, chỉ khác một điều đơn vị tính của các cột là tỉ lệ phần trăm.
- Hiện thực trong R

```
continents =continents %>% mutate("m" = n/sum(n))
ggplot(data = continents, aes(x = continent, y = m, group =1)) +
  geom_line(aes(y = cumsum(m))) + geom_bar(stat="identity",fill = "#008080" )+
  xlab("Continents") +
  ylab("Relative frquency") +
  geom_point(aes(x = continent, y = cumsum(m)), color = "#800000") +
  labs(title = "Relative frequency graph of countries for continents" , substitute= "Tan so
    tuong doi" )
```

- Kết quả:



3) Vẽ biểu đồ thể hiện nhiễm bệnh đã báo cáo của các quốc gia trong 7 ngày cuối của năm cuối cùng

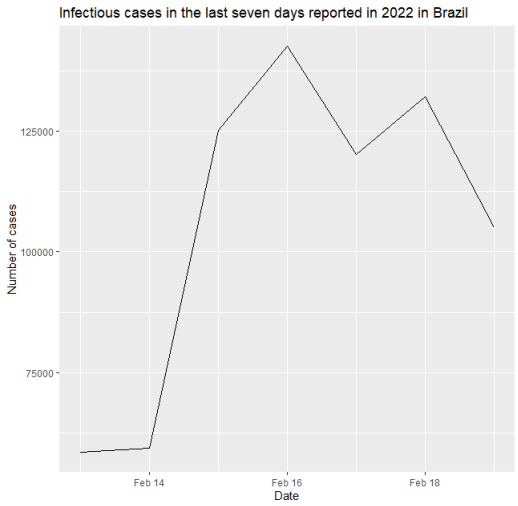
- Hướng giải quyết: Lọc dữ liệu của 7 ngày cuối được báo cáo của năm cuối. Vẽ biểu đồ dựa trên dữ liệu đó

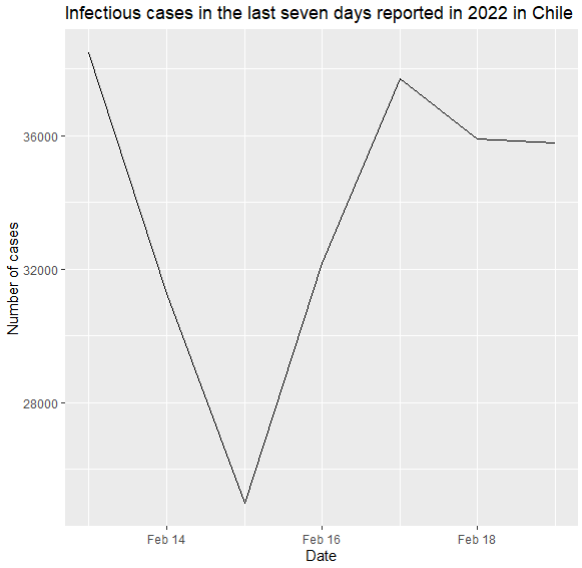
```
covid$date = as.Date(covid$date, format = "%m/%d/%Y")
bra = brazil[order(brazil$date, decreasing =  TRUE),]
bra = bra[1:7,] #select data from the last 7 days
ggplot(data= bra, aes(x= date, y =new_cases)) + geom_line()+ labs(title = "Infectious cases in
      the last seven days reported in 2022 in Brazil", x="Date", y="Number of cases")

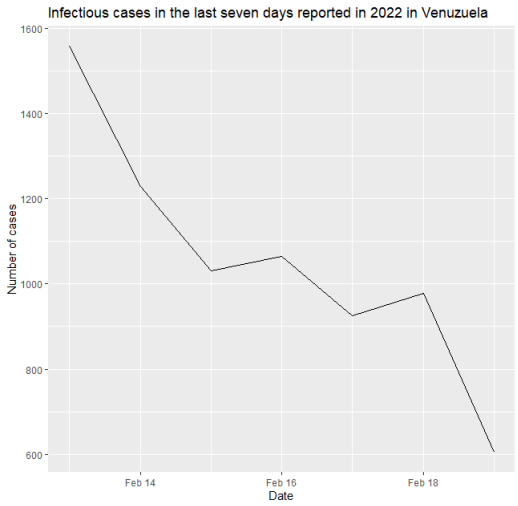
chi = chile[order(chile$date, decreasing =  TRUE),]
chi = chi[1:7,] #select the last 7 days
ggplot(data= chi, aes(x= date, y =new_cases)) + geom_line()+ labs(title = "Infectious cases in
      the last seven days reported in 2022 in Chile", x="Date", y="Number of cases")

ven = venezuela[order(venezuela$date, decreasing =  TRUE),]
ven = ven[1:7,] #select the last 7 days
ggplot(data= ven, aes(x= date, y =new_cases)) + geom_line()+ labs(title = "Infectious cases in
      the last seven days reported in 2022 in Venuzuela", x="Date", y="Number of cases")
```

- Kết quả:



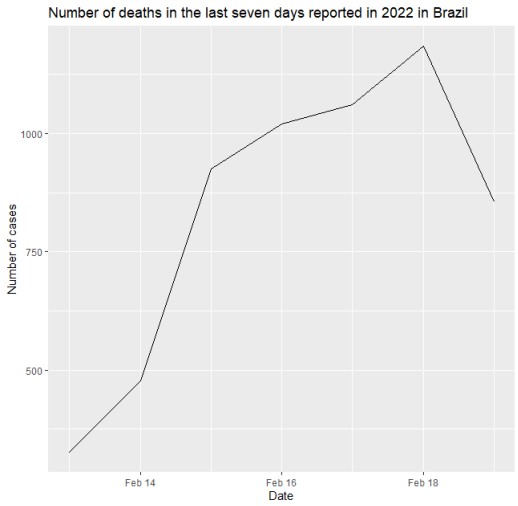


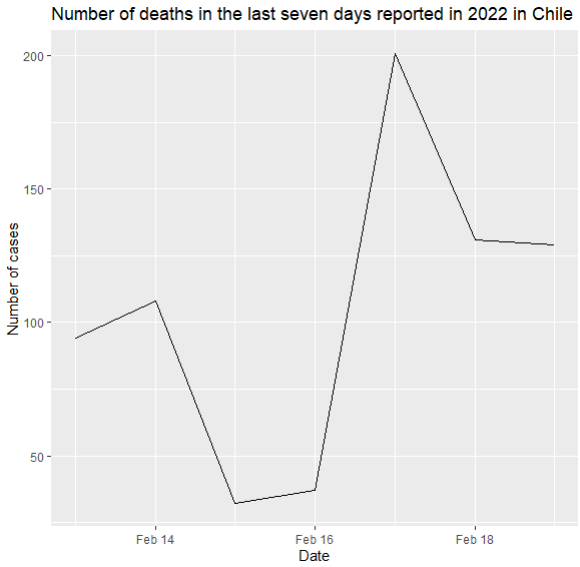


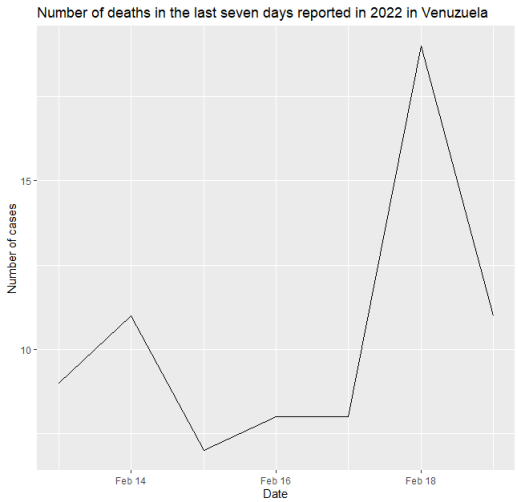
4) Vẽ biểu đồ thể hiện tử vong đã báo cáo của các quốc gia trong 7 ngày cuối của năm cuối cùng *Tương tự câu trên*Hiện thực trong R

- ```
  ggplot(data= bra, aes(x= date, y =new_deaths)) + geom_line()+ labs(title = "Number of deaths
      in the last seven days reported in 2022 in Brazil", x="Date", y="Number of cases")
  ggplot(data= chi, aes(x= date, y =new_deaths)) + geom_line()+ labs(title = "Number of deaths
      in the last seven days reported in 2022 in Chile", x="Date", y="Number of cases")
  ggplot(data= ven, aes(x= date, y =new_deaths)) + geom_line()+ labs(title = "Number of deaths
      in the last seven days reported in 2022 in Venuzuela", x="Date", y="Number of cases")
  ```

- Kết quả:
  ```







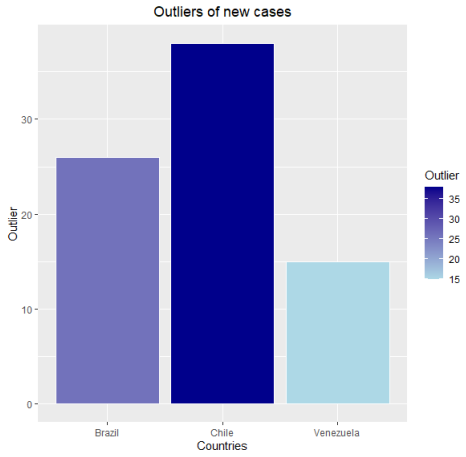
5) Vẽ biểu đồ phổ đất nước xuất hiện outliers cho nhiễm bệnh

- Dùng thông tin về Outliers ở phần ii) để vẽ biểu đồ
- Hiện thực trong R

```
ggplot(data = df_cases, aes(x= Countries, y = Outlier)) +
  geom_histogram(stat='identity',col="white", aes(fill = Outlier), binwidth = 7) +
  scale_fill_gradient("Outlier",low= "lightblue", high = "darkblue") +ggtitle("Outliers of
```

```
    new cases") +
  theme(plot.title = element_text(hjust = 0.5))
```

- Kết quả:

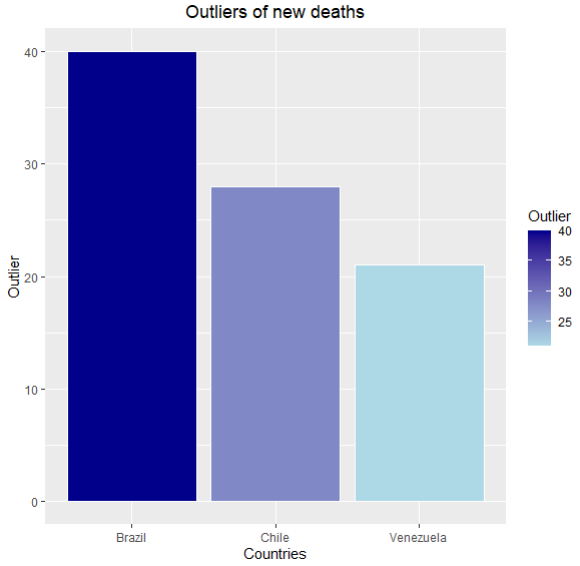


6) Vẽ biểu đồ phổ đất nước xuất hiện outliers cho tử vong
Tương tự câu trên

- Hiện thực trong R

```
ggplot(data = df_death, aes(x= Countries, y = Outlier)) +
  geom_histogram(stat='identity',col="white", aes(fill = Outlier), binwidth = 7) +
  scale_fill_gradient("Outlier",low= "lightblue", high = "darkblue") + ggtitle("Outliers of
   new deaths") +
  theme(plot.title = element_text(hjust = 0.5))
```

- Kết quả:



v) Nhóm câu hỏi liên quan đến trực quan dữ liệu theo thời gian là tháng
Trên từng năm hãy vẽ biểu đồ thể hiện trục Ox là thời gian, trục Oy là nhiễm bệnh/tử vong. Hãy dùng 4 ký số của mã đề để vẽ 4 tháng tương ứng theo ký số đó. Nếu ký số là 0 thì lấy tháng là 10.

- Chuẩn bị dữ liệu cho toàn bộ phần v (phần code dùng chung cho cả phần v)

```
library(tidyverse)
library(lubridate)
library(dplyr)
library(xlsx)
setwd("E:/BTL_CTRR")
```

```
data_covid = read.csv("owid-covid-data.csv", header = TRUE)
library("ggplot2")
data <- data_covid %>%
  separate(date, into = c("month", "day", "year"), sep = "/")
data = mutate(data, new_cases = abs(new_cases))
data = mutate(data, new_deaths = abs(new_deaths))
data <- subset(data,data$continent != "")
data[is.na(data)] <- 0
data <- transform(data,month=as.numeric(month))

year_2020_brazil <- data[data$location == "Brazil" & data$year=="2020",]
year_2020_chile <- data[data$location == "Chile" & data$year=="2020",]
year_2020_vene <- data[data$location == "Venezuela" & data$year=="2020",]

year_2021_brazil <- data[data$location == "Brazil"&data$year=="2021",]
year_2021_chile <- data[data$location == "Chile"&data$year=="2021",]
year_2021_vene <- data[data$location == "Venezuela"&data$year=="2021",]

year_2022_brazil <- data[data$location == "Brazil"&data$year=="2022",]
year_2022_chile <- data[data$location == "Chile"&data$year=="2022",]
year_2022_vene <- data[data$location == "Venezuela"&data$year=="2022",]

bra_2020 <- aggregate(cbind(new_cases,new_deaths) ~ month, data = year_2020_brazil, FUN = sum)
chi_2020 <- aggregate(cbind(new_cases,new_deaths) ~ month, data = year_2020_chile, FUN = sum)
ven_2020 <- aggregate(cbind(new_cases,new_deaths) ~ month, data = year_2020_vene, FUN = sum)

bra_2021 <- aggregate(cbind(new_cases,new_deaths) ~ month, data = year_2021_brazil, FUN = sum)
chi_2021 <- aggregate(cbind(new_cases,new_deaths) ~ month, data = year_2021_chile, FUN = sum)
ven_2021 <- aggregate(cbind(new_cases,new_deaths) ~ month, data = year_2021_vene, FUN = sum)

bra_2022 <- aggregate(cbind(new_cases,new_deaths) ~ month, data = year_2022_brazil, FUN = sum)
chi_2022 <- aggregate(cbind(new_cases,new_deaths) ~ month, data = year_2022_chile, FUN = sum)
ven_2022 <- aggregate(cbind(new_cases,new_deaths) ~ month, data = year_2022_vene, FUN = sum)

bra_2020_1<-subset(bra_2020,bra_2020$month==2 | bra_2020$month==10 |
                             bra_2020$month==3 | bra_2020$month==6)
bra_2021_1<-subset(bra_2021,bra_2021$month==2 | bra_2021$month==10 |
                    bra_2021$month==3 | bra_2021$month==6)
bra_2022_1<-subset(bra_2022,bra_2022$month==2 | bra_2022$month==10 |
                    bra_2022$month==3 | bra_2022$month==6)

chi_2020_1<-subset(chi_2020,bra_2020$month==2 | chi_2020$month==10 |
                    chi_2020$month==3 | chi_2020$month==6)
chi_2021_1<-subset(chi_2021,chi_2021$month==2 | chi_2021$month==10 |
                    chi_2021$month==3 | chi_2021$month==6)
chi_2022_1<-subset(chi_2022,chi_2022$month==2 | chi_2022$month==10 |
                    chi_2022$month==3 | chi_2022$month==6)

ven_2020_1<-subset(ven_2020,ven_2020$month==2 | ven_2020$month==10 |
                    ven_2020$month==3 | ven_2020$month==6)
ven_2021_1<-subset(ven_2021,ven_2021$month==2 | ven_2021$month==10 |
                    ven_2021$month==3 | ven_2021$month==6)
ven_2022_1<-subset(ven_2022,ven_2022$month==2 | ven_2022$month==10 |
                    ven_2022$month==3 | ven_2022$month==6)

bra_2020_2<-subset(bra_2020,bra_2020$month==11 | bra_2020$month==12)
bra_2021_2<-subset(bra_2021,bra_2021$month==11 | bra_2021$month==12)

chi_2020_2<-subset(chi_2020,chi_2020$month==11 | chi_2020$month==12)
chi_2021_2<-subset(chi_2021,chi_2021$month==11 | chi_2021$month==12)

ven_2020_2<-subset(ven_2020,ven_2020$month==11 | ven_2020$month==12)
ven_2021_2<-subset(ven_2021,ven_2021$month==11 | ven_2021$month==12)
new_row=c(2,0,0)
ven_2020_1 <- rbind(ven_2020_1,new_row)
temp=list(bra_2020_1,chi_2020_1,ven_2020_1)
data_2020=temp %>% reduce(full_join, by='month')
colnames(data_2020) <- c("month","brazil_cases","brazil_deaths","chile_cases","chile_deaths","
    vene_cases","vene_deaths")

temp=list(bra_2021_1,chi_2021_1,ven_2021_1)
```

```
data_2021=temp %>% reduce(full_join, by='month')
colnames(data_2021) <- c("month","brazil_cases","brazil_deaths","chile_cases","chile_deaths","
    vene_cases","vene_deaths")

temp=list(bra_2022_1,chi_2022_1,ven_2022_1)
data_2022=temp %>% reduce(full_join, by='month')
colnames(data_2022) <- c("month","brazil_cases","brazil_deaths","chile_cases","chile_deaths","
    vene_cases","vene_deaths")

temp=list(bra_2020_2,chi_2020_2,ven_2020_2)
data_2020_1=temp %>% reduce(full_join, by='month')
colnames(data_2020_1) <- c("month","brazil_cases","brazil_deaths","chile_cases","chile_deaths"
    ,"vene_cases","vene_deaths")

temp=list(bra_2021_2,chi_2021_2,ven_2021_2)
data_2021_1=temp %>% reduce(full_join, by='month')
colnames(data_2021_1) <- c("month","brazil_cases","brazil_deaths","chile_cases","chile_deaths"
    ,"vene_cases","vene_deaths")
```

1) Biểu đồ thu thập nhiễm bệnh cho từng tháng

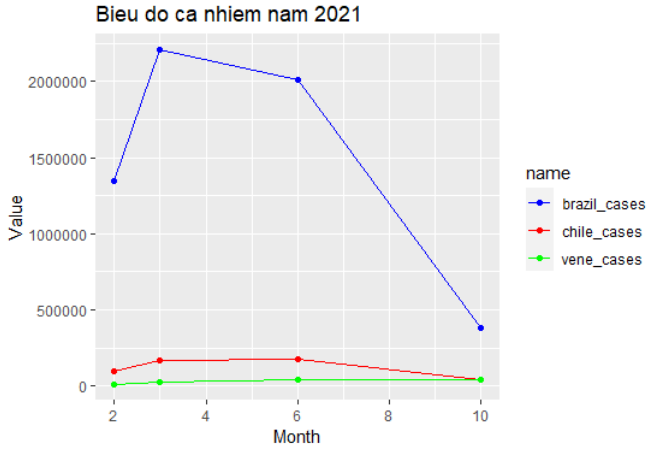
- Hiện thực trong R

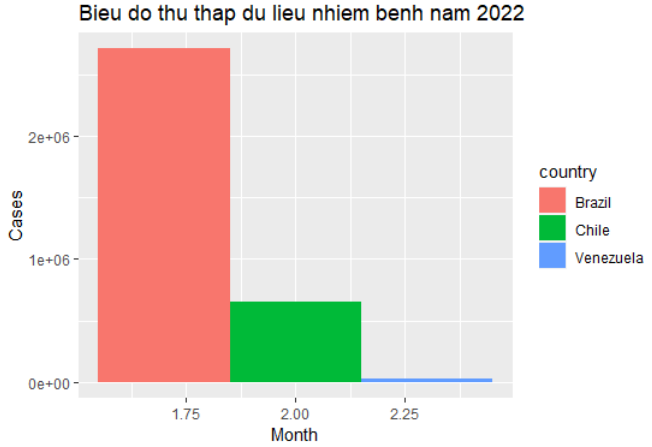
```
temp <- data_2020%>%
  tidyr::pivot_longer(cols = c(brazil_cases,chile_cases,vene_cases))
ggplot(data=temp,aes(x=month, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green"))+
  labs(title="Bieu do ca nhiem nam 2020",x="Month",y="Value")

temp <- data_2021%>%
  tidyr::pivot_longer(cols = c(brazil_cases,chile_cases,vene_cases))
ggplot(data=temp,aes(x=month, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green"))+
  labs(title="Bieu do ca nhiem nam 2021",x="Month",y="Value")

V_1_2022<- data.frame(month=c(2,2,2),
                      country=c('Brazil','Chile','Venezuela'),
                      cases=c(bra_2022_1$new_cases,chi_2022_1$new_cases,ven_2022_1$new_cases))
ggplot(V_1_2022, aes(fill=country, y=cases, x=month)) +
  geom_bar(position='dodge', stat='identity') +
  labs(x='Month', y='Cases', title='Bieu do thu thap du lieu tu vong nam 2022')
```

- Kết quả:





2) Biểu đồ thu thập tử vong cho từng tháng

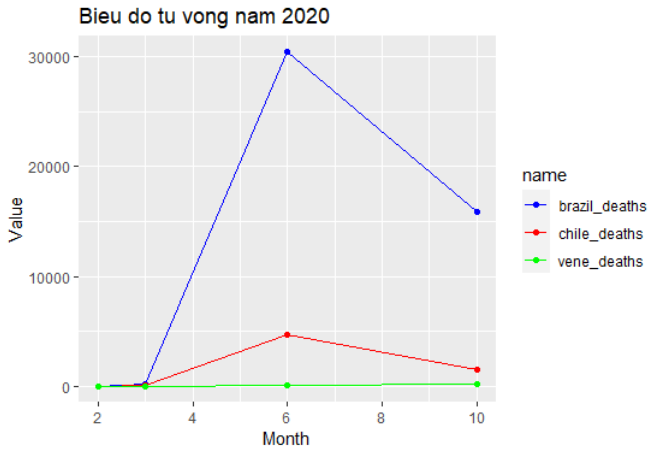
- Hiện thực trong R

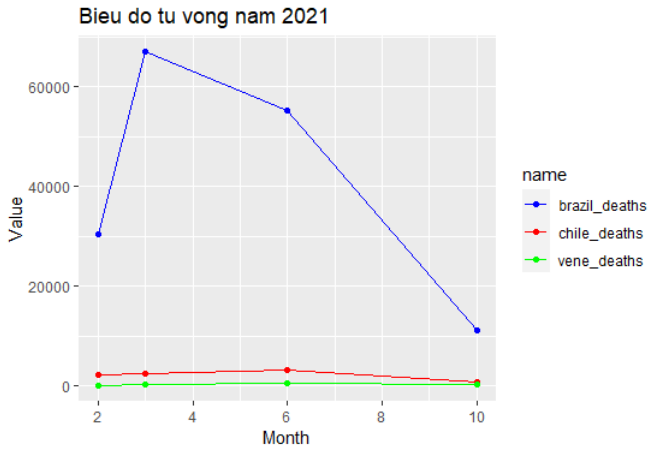
```
temp <- data_2020%>%
  tidyr::pivot_longer(cols = c(brazil_deaths,chile_deaths,vene_deaths))
ggplot(data=temp,aes(x=month, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green"))+
  labs(title="Bieu do tu vong nam 2020",x="Month",y="Value")

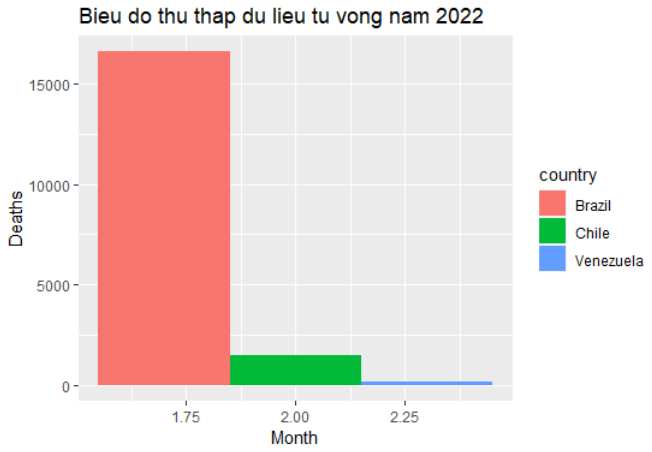
temp <- data_2021%>%
  tidyr::pivot_longer(cols = c(brazil_deaths,chile_deaths,vene_deaths))
ggplot(data=temp,aes(x=month, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green"))+
  labs(title="Bieu do tu vong nam 2021",x="Month",y="Value")


V_2_2022<-data.frame(month=c(2,2,2),
                     country=c('Brazil','Chile','Venezuela'),
                     deaths=c(bra_2022_1$new_deaths,chi_2022_1$new_deaths,ven_2022_1$
    new_deaths))
ggplot(V_2_2022, aes(fill=country, y=deaths, x=month)) +
  geom_bar(position='dodge', stat='identity') +
  labs(x='Month', y='Deaths', title='Bieu do thu thap du lieu tu vong nam 2022')
```

- Kết quả:







3) Biểu đồ thu thập gồm nhiễm bệnh và tử vong cho từng tháng

- Hiện thực trong R

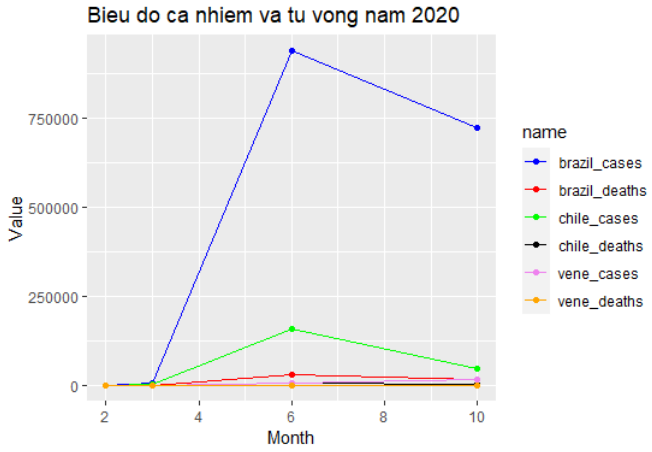
```
temp<- data_2020%>%
  tidyr::pivot_longer(cols = c(
     brazil_cases,brazil_deaths,chile_cases,chile_deaths,vene_cases,vene_deaths))
ggplot(data=temp,aes(x=month, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green","black","violet","orange"))+
  labs(title="Bieu do ca nhiem va tu vong nam 2020",x="Month",y="Value")

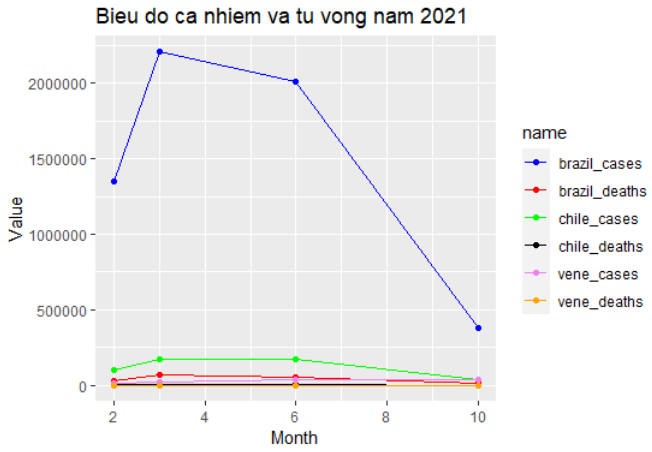
temp<- data_2021%>%
  tidyr::pivot_longer(cols = c(
     brazil_cases,brazil_deaths,chile_cases,chile_deaths,vene_cases,vene_deaths))
ggplot(data=temp,aes(x=month, y=value,color=name))+
```

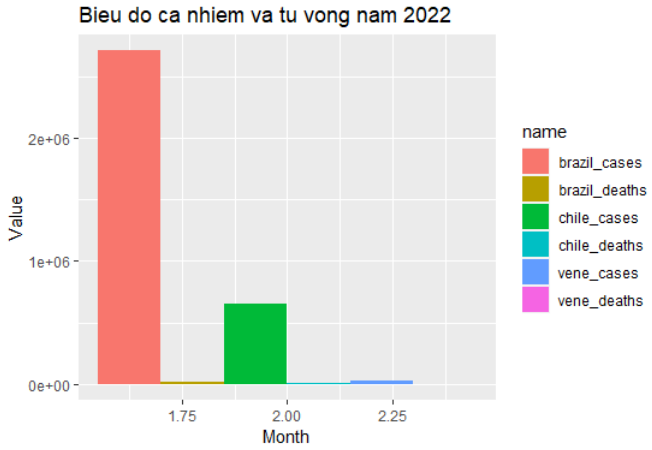
```
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green","black","violet","orange"))+
  labs(title="Bieu do ca nhiem va tu vong nam 2021",x="Month",y="Value")

temp<- data_2022%>%
  tidyr::pivot_longer(cols = c(
     brazil_cases,brazil_deaths,chile_cases,chile_deaths,vene_cases,vene_deaths))
ggplot(temp, aes(fill=name, y=value, x=month)) +
  geom_bar(position='dodge', stat='identity') +
  labs(x='Month', y='Value', title='Bieu do ca nhiem va tu vong nam 2022')
```

- Kết quả:







4) Biểu đồ thu thập nhiễm bệnh gồm 2 tháng cuối của năm

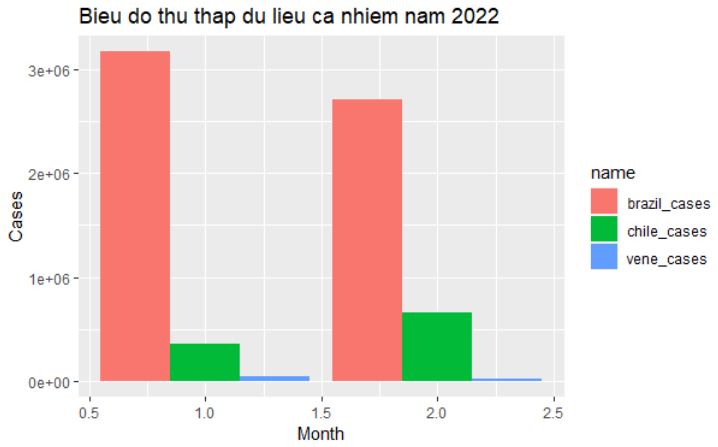
- Hiện thực trong R

```
temp<- data_2020_1%>%
  tidyr::pivot_longer(cols = c(brazil_cases,chile_cases,vene_cases))
ggplot(temp, aes(fill=name, y=value, x=month)) +
  geom_bar(position='dodge', stat='identity') +
  labs(x='Month', y='Cases', title='Bieu do thu thap du lieu ca nhiem nam 2020')

temp<- data_2021_1%>%
  tidyr::pivot_longer(cols = c(brazil_cases,chile_cases,vene_cases))
ggplot(temp, aes(fill=name, y=value, x=month)) +
  geom_bar(position='dodge', stat='identity') +
  labs(x='Month', y='Cases', title='Bieu do thu thap du lieu ca nhiem nam 2021')

temp<- data_2022_1%>%
  tidyr::pivot_longer(cols = c(brazil_cases,chile_cases,vene_cases))
ggplot(temp, aes(fill=name, y=value, x=month)) +
  geom_bar(position='dodge', stat='identity') +
  labs(x='Month', y='Cases', title='Bieu do thu thap du lieu ca nhiem nam 2022')
```

- Kết quả:



5) Biểu đồ thu thập tử vong gồm 2 tháng cuối của năm

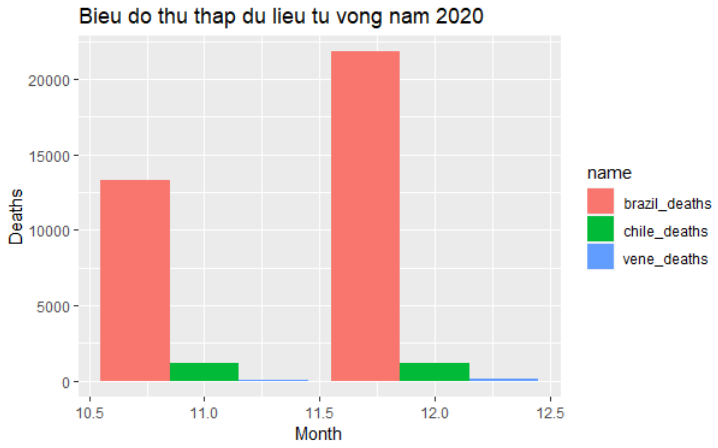
- Hiện thực trong R

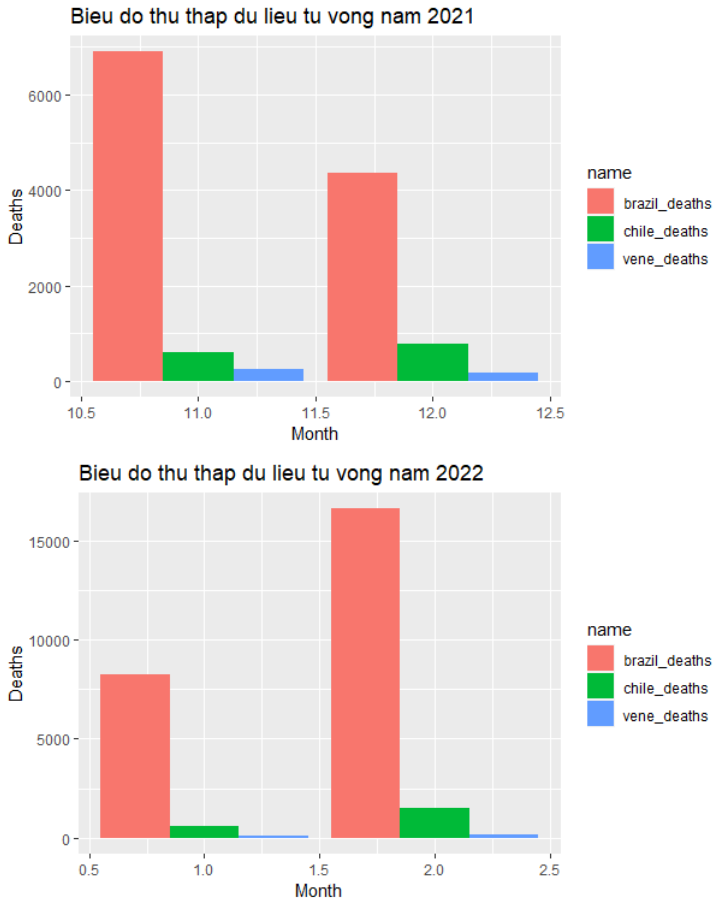
```
temp <- data_2020_1%>%
  tidyr::pivot_longer(cols = c(brazil_deaths,chile_deaths,vene_deaths))
ggplot(temp, aes(fill=name, y=value, x=month)) +
  geom_bar(position='dodge', stat='identity') +
  labs(x='Month', y='Deaths', title='Bieu do thu thap du lieu tu vong nam 2020')

temp <- data_2021_1%>%
  tidyr::pivot_longer(cols = c(brazil_deaths,chile_deaths,vene_deaths))
ggplot(temp, aes(fill=name, y=value, x=month)) +
  geom_bar(position='dodge', stat='identity') +
  labs(x='Month', y='Deaths', title='Bieu do thu thap du lieu tu vong nam 2021')

temp <- data_2022_1%>%
  tidyr::pivot_longer(cols = c(brazil_deaths,chile_deaths,vene_deaths))
ggplot(temp, aes(fill=name, y=value, x=month)) +
  geom_bar(position='dodge', stat='identity') +
  labs(x='Month', y='Deaths', title='Bieu do thu thap du lieu tu vong nam 2022')
```

- Kết quả:





6) Biểu đồ thu thập gồm nhiễm bệnh và tử vong gồm 2 tháng cuối của năm

- Hiện thực trong R

```
temp<- data_2020_1%>%
  tidyr::pivot_longer(cols = c(
     brazil_cases,brazil_deaths,chile_cases,chile_deaths,vene_cases,vene_deaths))
ggplot(temp, aes(fill=name, y=value, x=month)) +
  geom_bar(position='dodge', stat='identity') +
  labs(x='Month', y='Value', title='Bieu do ca nhiem va tu vong nam 2020')

temp<- data_2021_1%>%
  tidyr::pivot_longer(cols = c(
     brazil_cases,brazil_deaths,chile_cases,chile_deaths,vene_cases,vene_deaths))
ggplot(temp, aes(fill=name, y=value, x=month)) +
  geom_bar(position='dodge', stat='identity') +
  labs(x='Month', y='Value', title='Bieu do ca nhiem va tu vong nam 2021')

temp<- data_2022_1%>%
  tidyr::pivot_longer(cols = c(
     brazil_cases,brazil_deaths,chile_cases,chile_deaths,vene_cases,vene_deaths))
ggplot(temp, aes(fill=name, y=value, x=month)) +
  geom_bar(position='dodge', stat='identity') +
  labs(x='Month', y='Value', title='Bieu do ca nhiem va tu vong nam 2022')
```

- Kết quả:

7) Biểu đồ thu thập nhiễm bệnh tích lũy cho từng tháng

- Hiện thực trong R

```
temp=list(bra_2020,chi_2020,ven_2020)
sum_2020=temp %>% reduce(full_join, by='month')
colnames(sum_2020) <- c("month","brazil_cases","brazil_deaths","chile_cases","chile_deaths","
    vene_cases","vene_deaths")
sum_2020[is.na(sum_2020)] <- 0

temp=list(bra_2021,chi_2021,ven_2021)
sum_2021=temp %>% reduce(full_join, by='month')
colnames(sum_2021) <- c("month","brazil_cases","brazil_deaths","chile_cases","chile_deaths","
    vene_cases","vene_deaths")

temp=list(bra_2022,chi_2022,ven_2022)
sum_2022=temp %>% reduce(full_join, by='month')
```

```
colnames(sum_2022) <- c("month","brazil_cases","brazil_deaths","chile_cases","chile_deaths","
    vene_cases","vene_deaths")

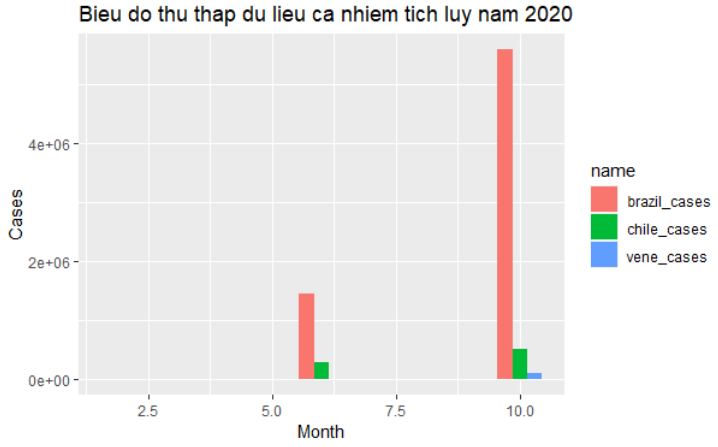
cumsum_2020=as.data.frame(lapply(sum_2020,cumsum))
cumsum_2020$month = c(2:12)
cumsum_2021=as.data.frame(lapply(sum_2021,cumsum))
cumsum_2021$month =c(1:12)
cumsum_2022=as.data.frame(lapply(sum_2022,cumsum))
cumsum_2022$month = c(1,2)

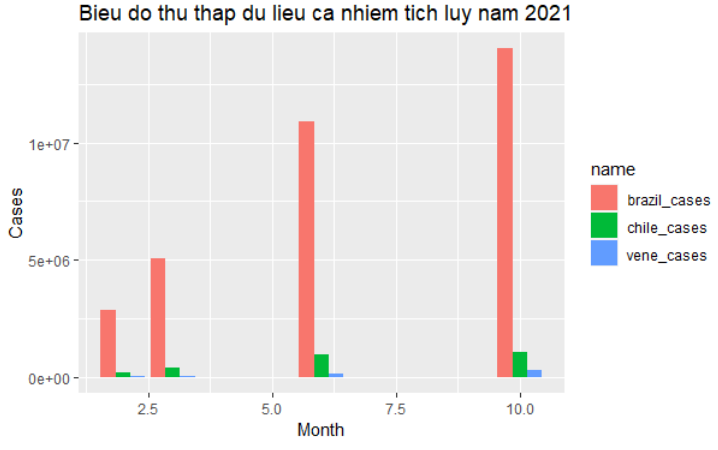
temp<- cumsum_2020%>%
  tidyr::pivot_longer(cols = c(brazil_cases,chile_cases,vene_cases))
temp <- subset(temp,temp$month==2 | temp$month==10 |
         temp$month==3 | temp$month==6)
ggplot(temp, aes(fill=name, y=value, x=month)) +
  geom_bar(position='dodge', stat='identity') +
  labs(x='Month', y='Cases', title='Bieu do thu thap du lieu ca nhiem tich luy nam 2020')

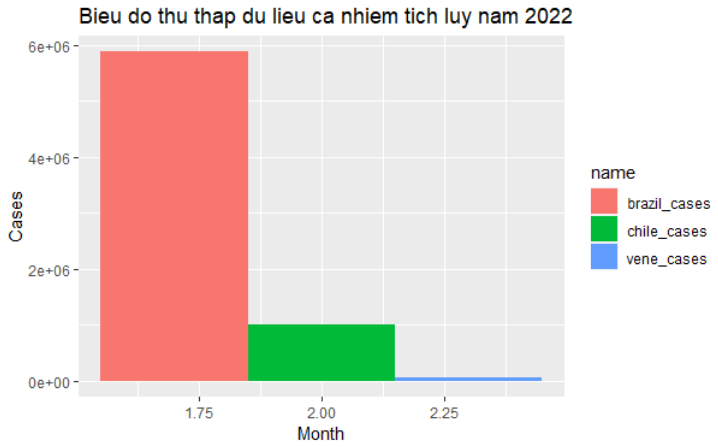
temp<- cumsum_2021%>%
  tidyr::pivot_longer(cols = c(brazil_cases,chile_cases,vene_cases))
temp <- subset(temp,temp$month==2 | temp$month==10 |
                 temp$month==3 | temp$month==6)
ggplot(temp, aes(fill=name, y=value, x=month)) +
  geom_bar(position='dodge', stat='identity') +
  labs(x='Month', y='Cases', title='Bieu do thu thap du lieu ca nhiem tich luy nam 2021')

temp<- cumsum_2022%>%
  tidyr::pivot_longer(cols = c(brazil_cases,chile_cases,vene_cases))
temp <- subset(temp,temp$month==2 | temp$month==10 |
                 temp$month==3 | temp$month==6)
ggplot(temp, aes(fill=name, y=value, x=month)) +
  geom_bar(position='dodge', stat='identity') +
  labs(x='Month', y='Cases', title='Bieu do thu thap du lieu ca nhiem tich luy nam 2022')
```

- Kết quả:





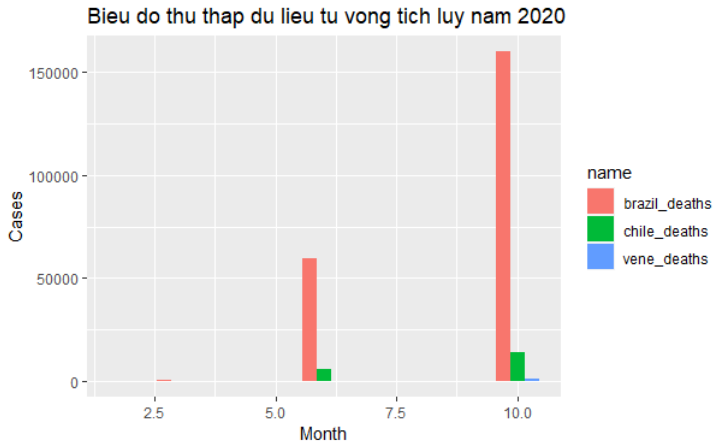


8) Biểu đồ thu thập tử vong tích lũy cho từng tháng

- Hiện thực trong R

```
temp<- cumsum_2020%>%
  tidyr::pivot_longer(cols = c(brazil_deaths,chile_deaths,vene_deaths))
temp <- subset(temp,temp$month==2 | temp$month==10 |
                 temp$month==3 | temp$month==6)
ggplot(temp, aes(fill=name, y=value, x=month)) +
  geom_bar(position='dodge', stat='identity') +
  labs(x='Month', y='Cases', title='Bieu do thu thap du lieu tu vong tich luy nam 2020')
temp<- cumsum_2021%>%
  tidyr::pivot_longer(cols = c(brazil_deaths,chile_deaths,vene_deaths))
temp <- subset(temp,temp$month==2 | temp$month==10 |
                 temp$month==3 | temp$month==6)
ggplot(temp, aes(fill=name, y=value, x=month)) +
  geom_bar(position='dodge', stat='identity') +
  labs(x='Month', y='Cases', title='Bieu do thu thap du lieu ca nhiem tich luy nam 2021')
temp<- cumsum_2022%>%
  tidyr::pivot_longer(cols = c(brazil_deaths,chile_deaths,vene_deaths))
temp <- subset(temp,temp$month==2 | temp$month==10 |
                 temp$month==3 | temp$month==6)
ggplot(temp, aes(fill=name, y=value, x=month)) +
  geom_bar(position='dodge', stat='identity') +
  labs(x='Month', y='Cases', title='Bieu do thu thap du lieu tu vong tich luy nam 2022')
```

- Kết quả:



vi) Nhóm câu hỏi liên quan đến trực quan dữ liệu theo trung bình 7 ngày gần nhất:

- Trên từng năm hãy vẽ biểu đồ thể hiện trục Ox là thời gian, trục Oy là nhiễm bệnh/tử vong. Hãy dùng 4 ký số của mã đề để vẽ 4 tháng tương ứng theo ký số đó. Nếu ký số là 0 thì lấy tháng là 10.

- Dùng trung bình của các ca nhiễm bệnh và tử vong được báo cáo trong 7 ngày gần nhất để loại trừ một số báo cáo không thường xuyên và đưa chúng ta đến gần hơn với con số hàng ngày.

- Chuẩn bị dữ liệu cho toàn bộ phần vi (phần code dùng chung cho cả phần vi)

```
library(tidyverse)
library(datasets)
library(dplyr)
library("ggplot2")
library(lubridate)
setwd("E:/BTL_CTRR")
covid = read.csv("owid-covid-data.csv", header = TRUE)
covid$date = as.Date(covid$date, format = "%m/%d/%Y")
covid = covid %>% filter(!continent ==’’) %>% mutate(new_cases = abs(new_cases), new_deaths= abs(
    new_deaths))
covid[is.na(covid)]=0
brazil = subset(covid, iso_code == "BRA")
chile = subset(covid, iso_code == "CHL")
venezuela = subset(covid, iso_code == "VEN")

brazil =  brazil %>% mutate(day =as.numeric(format(as.Date(brazil$date),"%d"))) %>%
  mutate(month = as.numeric(format(as.Date(brazil$date),"%m"))) %>%
  mutate(year = as.numeric(format(as.Date(brazil$date),"%Y")))
brazil[["Newcases"]]=NA
brazil[["Newdeaths"]]=NA
for(i in 1:nrow(brazil)) {

  if(i <=7) {
    brazil$Newcases[[i]] = mean(brazil[[5]][1:i])
    brazil$newdeaths[[i]] = mean(brazil[[6]][1:i])
```

```
  }
  else{
    brazil$Newcases[[i]] = sum( brazil[[5]][(i-6):i]  )/7
    brazil$Newdeaths[[i]] = sum( brazil[[6]][(i-6):i]  )/7
  }
}
brazil = brazil %>% filter(!Newcases==0) %>% filter(!Newdeaths==0)

chile =  chile %>% mutate(day =as.numeric(format(as.Date(chile$date),"%d"))) %>%
  mutate(month = as.numeric(format(as.Date(chile$date),"%m"))) %>%
  mutate(year = as.numeric(format(as.Date(chile$date),"%Y")))
# thay du lieu new cases, deaths bang gia tri trung binh 7 ngay gan nhat
chile[["Newcases"]]=NA
chile[["Newdeaths"]]=NA
for(i in 1:nrow(chile)) {

  if(i <=7) {
    chile$Newcases[[i]] = mean(chile[[5]][1:i])
    chile$newdeaths[[i]] = mean(chile[[6]][1:i])
  }
  else{
    chile$Newcases[[i]] = sum( chile[[5]][(i-6):i]  )/7
    chile$Newdeaths[[i]] = sum( chile[[6]][(i-6):i]  )/7
  }
}
chile = chile %>% filter(!Newcases==0) %>% filter(!Newdeaths==0)

#Dung cot Newcases thay cho new_cases v Newdeaths thay cho new_deaths

venezuela =  venezuela %>% mutate(day =as.numeric(format(as.Date(venezuela$date),"%d"))) %>%
  mutate(month = as.numeric(format(as.Date(venezuela$date),"%m"))) %>%
  mutate(year = as.numeric(format(as.Date(venezuela$date),"%Y")))
# thay du lieu new cases, deaths bang gia tri trung binh 7 ngay gan nhat
venezuela[["Newcases"]]=NA
venezuela[["Newdeaths"]]=NA
for(i in 1:nrow(venezuela)) {

  if(i <=7) {
    venezuela$Newcases[[i]] = mean(venezuela[[5]][1:i])
    venezuela$newdeaths[[i]] = mean(venezuela[[6]][1:i])
  }
  else{
    venezuela$Newcases[[i]] = sum( venezuela[[5]][(i-6):i]  )/7
    venezuela$Newdeaths[[i]] = sum( venezuela[[6]][(i-6):i]  )/7
  }
}
venezuela = venezuela %>% filter(!Newcases==0) %>% filter(!Newdeaths==0)

#Dung cot Newcases thay cho new_cases v Newdeaths thay cho new_deaths

brazil_2020 <- subset(brazil,brazil$year == "2020")
brazil_2021 <- subset(brazil,brazil$year == "2021")
brazil_2022 <- subset(brazil,brazil$year == "2022")

chile_2020 <- subset(chile,chile$year == "2020")
chile_2021 <- subset(chile,chile$year == "2021")
chile_2022 <- subset(chile,chile$year == "2022")
vene_2020 <- subset(venezuela,venezuela$year == "2020")
vene_2021 <- subset(venezuela,venezuela$year == "2021")
vene_2022 <- subset(venezuela,venezuela$year == "2022")

temp=list(brazil_2020,chile_2020,vene_2020)
data_2020=temp %>% reduce(full_join, by='date')
data_2020 = rename(data_2020,brazil_cases=Newcases.x,brazil_deaths=Newdeaths.x,
                   chile_cases=Newcases.y,chile_deaths=Newdeaths.y,
                   venezuela_cases=Newcases,venezuela_deaths=Newdeaths)

temp=list(brazil_2021,chile_2021,vene_2021)
data_2021=temp %>% reduce(full_join, by='date')
data_2021 = rename(data_2021,brazil_cases=Newcases.x,brazil_deaths=Newdeaths.x,
                   chile_cases=Newcases.y,chile_deaths=Newdeaths.y,
```

```
                    venezuela_cases=Newcases,venezuela_deaths=Newdeaths)

temp=list(brazil_2022,chile_2022,vene_2022)
data_2022=temp %>% reduce(full_join, by='date')
data_2022 = rename(data_2022,brazil_cases=Newcases.x,brazil_deaths=Newdeaths.x,
                    chile_cases=Newcases.y,chile_deaths=Newdeaths.y,
                    venezuela_cases=Newcases,venezuela_deaths=Newdeaths)
```

1) Biểu đồ thu thập nhiễm bệnh cho từng tháng

- Hiện thực trong R

```
data_2020_3 = subset(data_2020, as.numeric(format(as.Date(data_2020$date),"%m")) == "3")
temp <- data_2020_3%>%
  tidyr::pivot_longer(cols = c(brazil_cases,chile_cases,venezuela_cases))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green"))+
  labs(title="Bieu do ca nhiem thang 3 nam 2020",x="Date of month",y="Value")

data_2020_6 = subset(data_2020, as.numeric(format(as.Date(data_2020$date),"%m")) == "6")
temp <- data_2020_6%>%
  tidyr::pivot_longer(cols = c(brazil_cases,chile_cases,venezuela_cases))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green"))+
  labs(title="Bieu do ca nhiem thang 6 nam 2020",x="Date of month",y="Value")

data_2020_10 = subset(data_2020, as.numeric(format(as.Date(data_2020$date),"%m")) == "10")
temp <- data_2020_10%>%
  tidyr::pivot_longer(cols = c(brazil_cases,chile_cases,venezuela_cases))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green"))+
  labs(title="Bieu do ca nhiem thang 10 nam 2020",x="Date of month",y="Value")

data_2021_2 = subset(data_2021, as.numeric(format(as.Date(data_2021$date),"%m")) == "2")
temp <- data_2021_2%>%
  tidyr::pivot_longer(cols = c(brazil_cases,chile_cases,venezuela_cases))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green"))+
  labs(title="Bieu do ca nhiem thang 2 nam 2021",x="Date of month",y="Value")

data_2021_3 = subset(data_2021, as.numeric(format(as.Date(data_2021$date),"%m")) == "3")
temp <- data_2021_3%>%
  tidyr::pivot_longer(cols = c(brazil_cases,chile_cases,venezuela_cases))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green"))+
  labs(title="Bieu do ca nhiem thang 3 nam 2021",x="Date of month",y="Value")

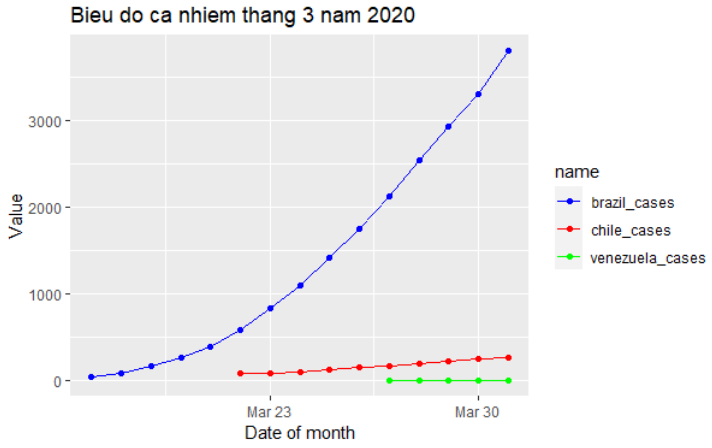
data_2021_6 = subset(data_2021, as.numeric(format(as.Date(data_2021$date),"%m")) == "6")
temp <- data_2021_6%>%
  tidyr::pivot_longer(cols = c(brazil_cases,chile_cases,venezuela_cases))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green"))+
  labs(title="Bieu do ca nhiem thang 6 nam 2021",x="Date of month",y="Value")

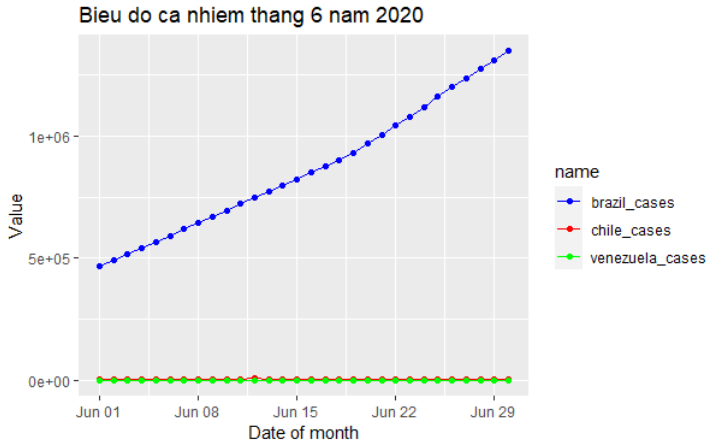
data_2021_10 = subset(data_2021, as.numeric(format(as.Date(data_2021$date),"%m")) == "10")
temp <- data_2021_10%>%
  tidyr::pivot_longer(cols = c(brazil_cases,chile_cases,venezuela_cases))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
```

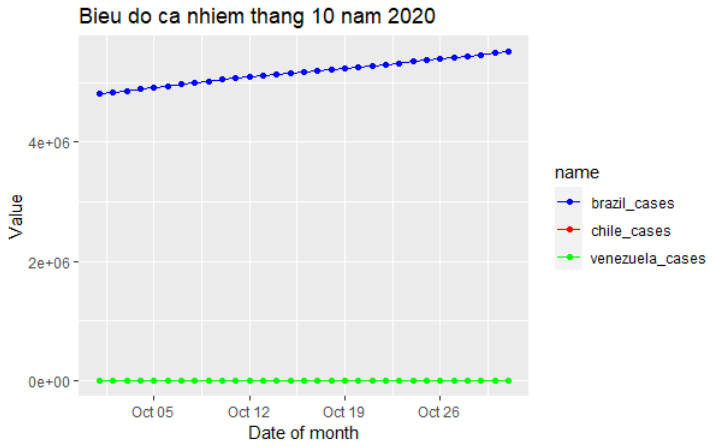
```
  geom_point()+
  scale_color_manual(values =c("blue","red","green"))+
  labs(title="Bieu do ca nhiem thang 10 nam 2021",x="Date of month",y="Value")

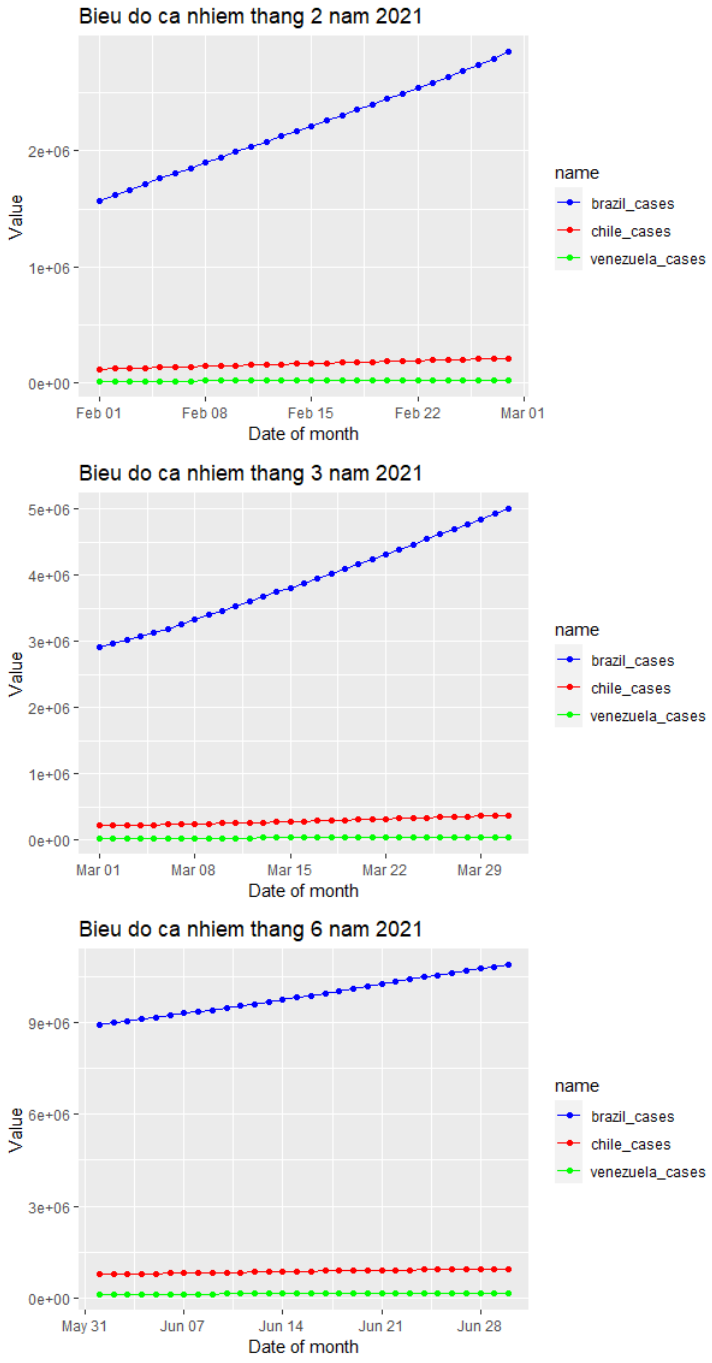
data_2022_2 = subset(data_2022, as.numeric(format(as.Date(data_2022$date),"%m")) == "2")
temp <- data_2022_2%>%
  tidyr::pivot_longer(cols = c(brazil_cases,chile_cases,venezuela_cases))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green"))+
  labs(title="Bieu do ca nhiem thang 2 nam 2022",x="Date of month",y="Value")
```

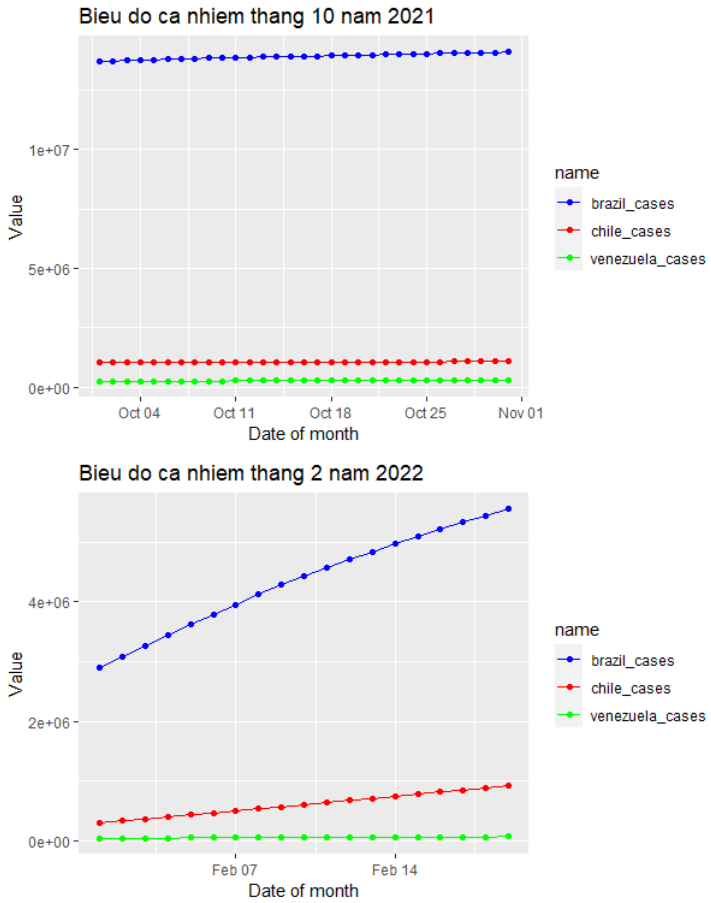
- Kết quả:








Bieu do ca nhiem thang 2 nam 2021
Value
2e+06
1e+06
0e+00
Feb 01
Feb 08
Feb 15
Feb 22
Mar 01
Date of month
name
brazil_cases
chile_cases
venezuela_cases
Bieu do ca nhiem thang 3 nam 2021
Value
5e+06
4e+06
3e+06
2e+06
1e+06
0e+00
Mar 01
Mar 08
Mar 15
Mar 22
Mar 29
Date of month
name
brazil_cases
chile_cases
venezuela_cases
Bieu do ca nhiem thang 6 nam 2021
Value
9e+06
6e+06
3e+06
0e+00
May 31
Jun 07
Jun 14
Jun 21
Jun 28
Date of month
name
brazil_cases
chile_cases
venezuela_cases



2) Biểu đồ thu thập tử vong cho từng tháng

- Hiện thực trong R

```
#Vi_2
temp <- data_2020_3%>%
  tidyr::pivot_longer(cols = c(brazil_deaths,chile_deaths,venezuela_deaths))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green"))+
  labs(title="Bieu do tu vong thang 3 nam 2020",x="Date of month",y="Value")

temp <- data_2020_6%>%
  tidyr::pivot_longer(cols = c(brazil_deaths,chile_deaths,venezuela_deaths))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green"))+
  labs(title="Bieu do tu vong thang 6 nam 2020",x="Date of month",y="Value")

temp <- data_2020_10%>%
  tidyr::pivot_longer(cols = c(brazil_deaths,chile_deaths,venezuela_deaths))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green"))+
  labs(title="Bieu do tu vong thang 10 nam 2020",x="Date of month",y="Value")

temp <- data_2021_2%>%
  tidyr::pivot_longer(cols = c(brazil_deaths,chile_deaths,venezuela_deaths))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green"))+
```

```
  labs(title="Bieu do tu vong thang 2 nam 2021",x="Date of month",y="Value")

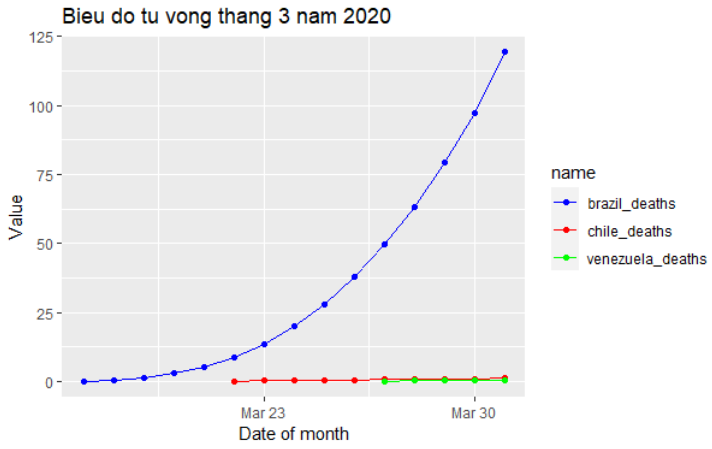
temp <- data_2021_3%>%
  tidyr::pivot_longer(cols = c(brazil_deaths,chile_deaths,venezuela_deaths))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green"))+
  labs(title="Bieu do tu vong thang 3 nam 2021",x="Date of month",y="Value")

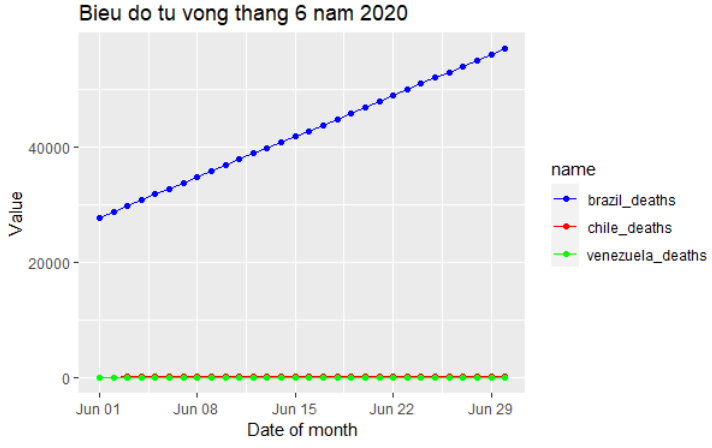
temp <- data_2021_6%>%
tidyr::pivot_longer(cols = c(brazil_deaths,chile_deaths,venezuela_deaths))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green"))+
  labs(title="Bieu do tu vong thang 6 nam 2021",x="Date of month",y="Value")

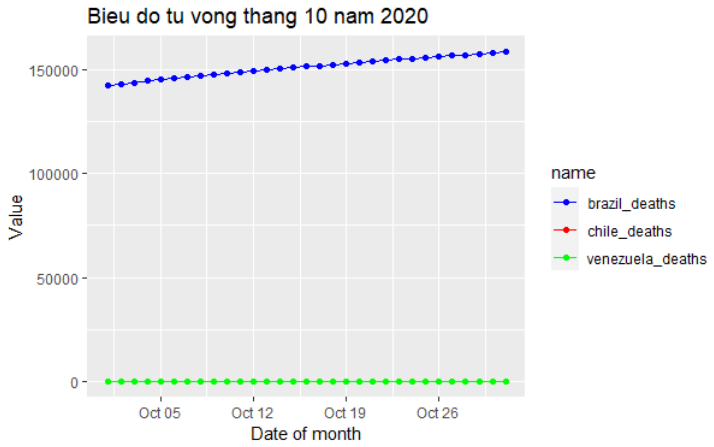
temp <- data_2021_10%>%
  tidyr::pivot_longer(cols = c(brazil_deaths,chile_deaths,venezuela_deaths))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green"))+
  labs(title="Bieu do tu vong thang 10 nam 2021",x="Date of month",y="Value")

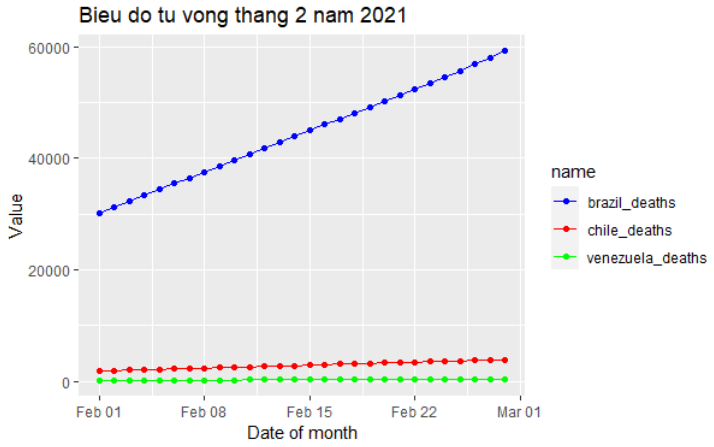
temp <- data_2022_2%>%
  tidyr::pivot_longer(cols = c(brazil_deaths,chile_deaths,venezuela_deaths))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green"))+
  labs(title="Bieu do tu vong thang 2 nam 2022",x="Date of month",y="Value")
```

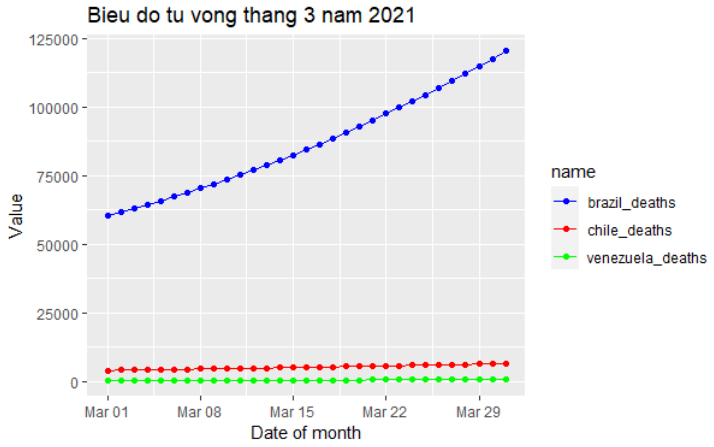
- Kết quả:

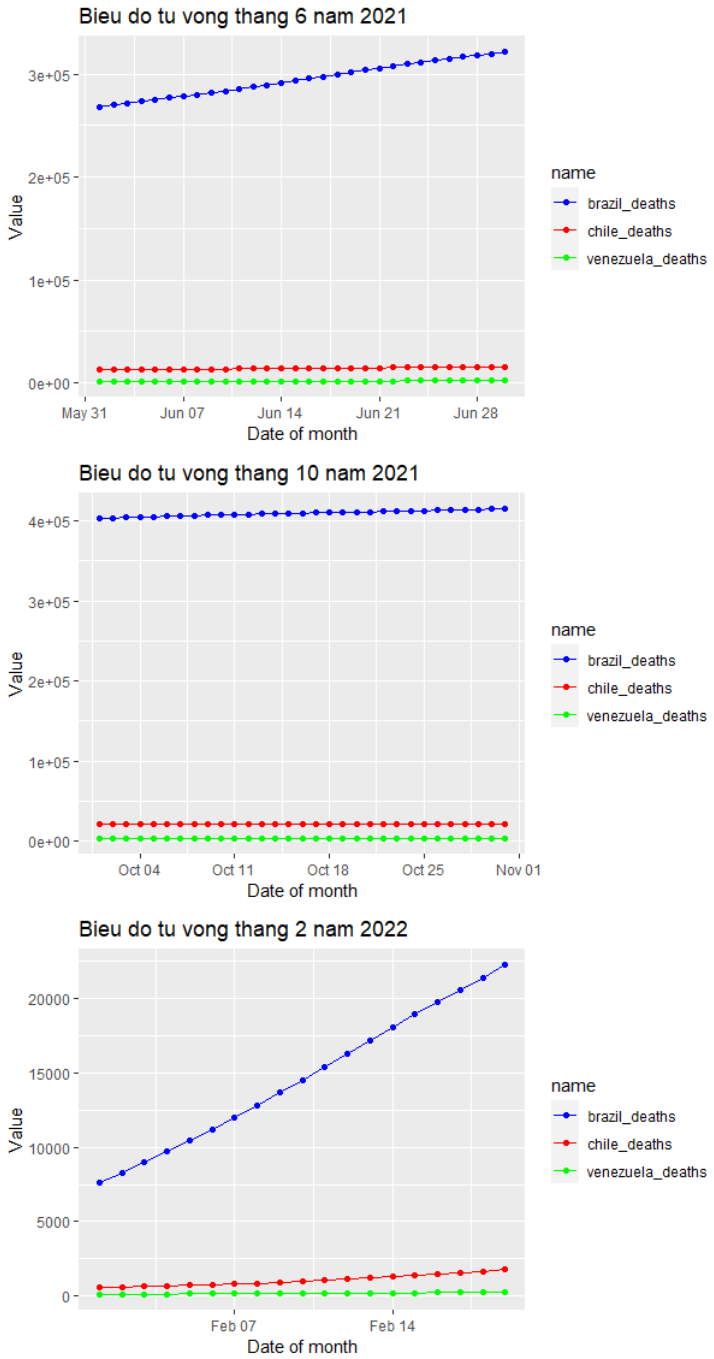





Bieu do tu vong thang 10 nam 2020
150000
100000
50000
0
Value
Oct 05
Oct 12
Oct 19
Oct 26
Date of month
name
brazil_deaths
chile_deaths
venezuela_deaths


Bieu do tu vong thang 2 nam 2021
60000
40000
20000
0
Value
Feb 01
Feb 08
Feb 15
Feb 22
Mar 01
Date of month
name
brazil_deaths
chile_deaths
venezuela_deaths


Bieu do tu vong thang 3 nam 2021
125000
100000
75000
50000
25000
0
Value
Mar 01
Mar 08
Mar 15
Mar 22
Mar 29
Date of month
name
brazil_deaths
chile_deaths
venezuela_deaths



3) Biểu đồ thu thập gồm nhiễm bệnh và tử vong cho từng tháng

- Hiện thực trong R

```
temp <- data_2020_3%>%
  tidyr::pivot_longer(cols = c(
     brazil_cases,brazil_deaths,chile_cases,chile_deaths,venezuela_cases,venezuela_deaths))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green","brown","violet","pink"))+
  labs(title="Bieu do ca nhiem va tu vong thang 3 nam 2020",x="Date of month",y="Value")

temp <- data_2020_6%>%
  tidyr::pivot_longer(cols = c(
     brazil_cases,brazil_deaths,chile_cases,chile_deaths,venezuela_cases,venezuela_deaths))
ggplot(data=temp,aes(x=date, y=value,color=name))+
```

```
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green","brown","violet","pink"))+
  labs(title="Bieu do ca nhiem va tu vong thang 6 nam 2020",x="Date of month",y="Value")

temp <- data_2020_10%>%
  tidyr::pivot_longer(cols = c(
     brazil_cases,brazil_deaths,chile_cases,chile_deaths,venezuela_cases,venezuela_deaths))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green","brown","violet","pink"))+
  labs(title="Bieu do ca nhiem va tu vong thang 10 nam 2020",x="Date of month",y="Value")

temp <- data_2021_2%>%
  tidyr::pivot_longer(cols = c(
     brazil_cases,brazil_deaths,chile_cases,chile_deaths,venezuela_cases,venezuela_deaths))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green","brown","violet","pink"))+
  labs(title="Bieu do ca nhiem va tu vong thang 2 nam 2021",x="Date of month",y="Value")

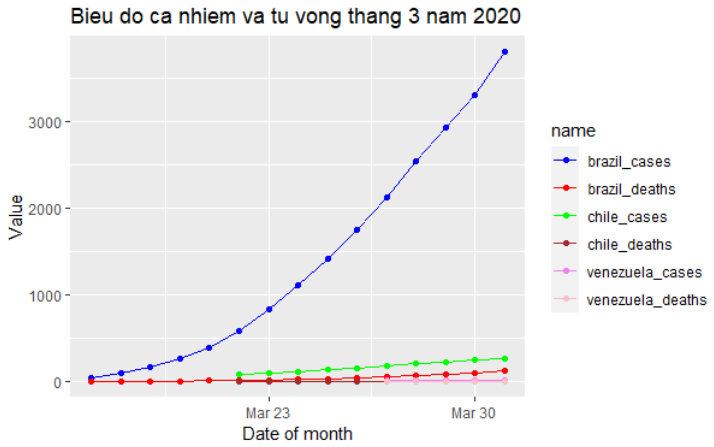
temp <- data_2021_3%>%
  tidyr::pivot_longer(cols = c(
     brazil_cases,brazil_deaths,chile_cases,chile_deaths,venezuela_cases,venezuela_deaths))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green","brown","violet","pink"))+
  labs(title="Bieu do ca nhiem va tu vong thang 3 nam 2021",x="Date of month",y="Value")

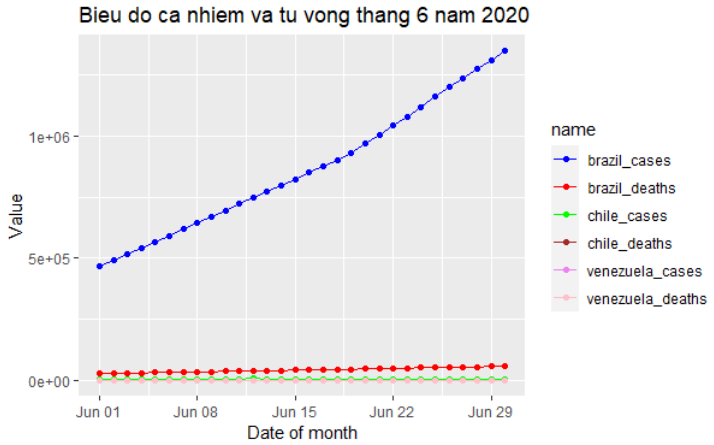
temp <- data_2021_6%>%
tidyr::pivot_longer(cols = c(
     brazil_cases,brazil_deaths,chile_cases,chile_deaths,venezuela_cases,venezuela_deaths))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green","brown","violet","pink"))+
  labs(title="Bieu do ca nhiem va tu vong thang 6 nam 2021",x="Date of month",y="Value")

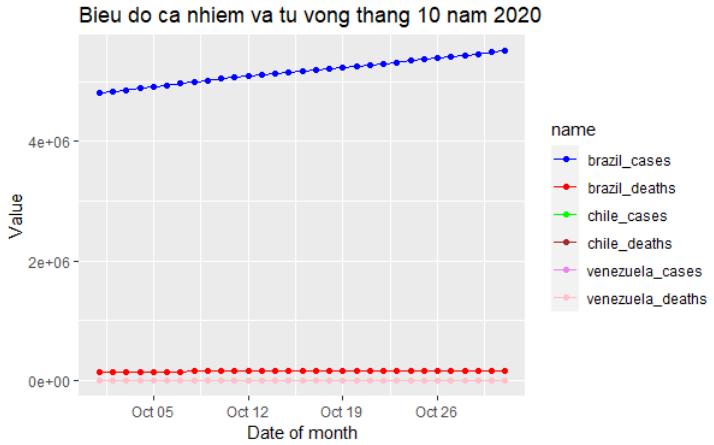
temp <- data_2021_10%>%
  tidyr::pivot_longer(cols = c(
     brazil_cases,brazil_deaths,chile_cases,chile_deaths,venezuela_cases,venezuela_deaths))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green","brown","violet","pink"))+
  labs(title="Bieu do ca nhiem va tu vong thang 10 nam 2021",x="Date of month",y="Value")

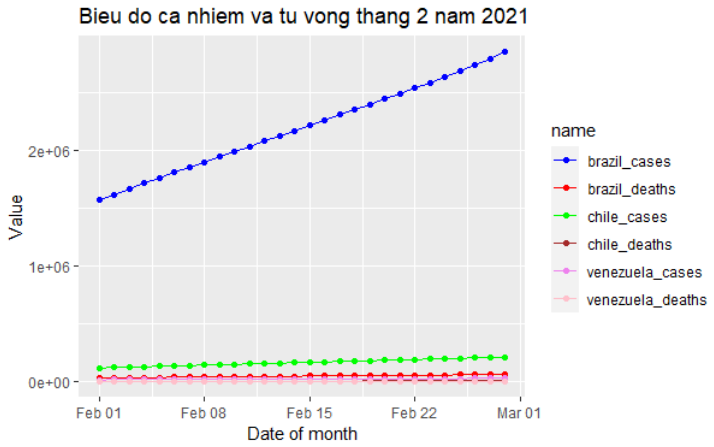
temp <- data_2022_2%>%
  tidyr::pivot_longer(cols = c(
     brazil_cases,brazil_deaths,chile_cases,chile_deaths,venezuela_cases,venezuela_deaths))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green","brown","violet","pink"))+
  labs(title="Bieu do ca nhiem va tu vong thang 2 nam 2022",x="Date of month",y="Value")
```

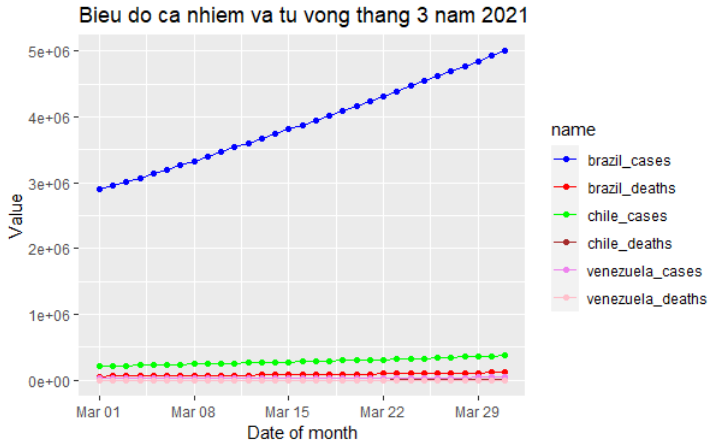
- Kết quả:


Bieu do ca nhiem va tu vong thang 3 nam 2020
Value
3000
2000
1000
0
Mar 23
Mar 30
Date of month
name
brazil_cases
brazil_deaths
chile_cases
chile_deaths
venezuela_cases
venezuela_deaths

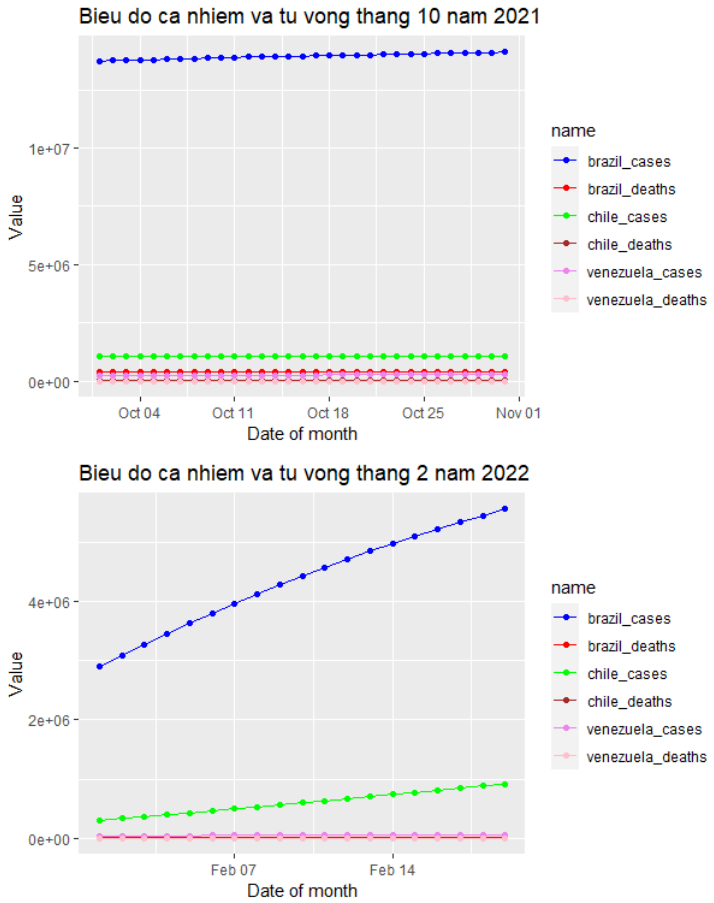

Bieu do ca nhiem va tu vong thang 6 nam 2020
Value
1e+06
5e+05
0e+00
Jun 01
Jun 08
Jun 15
Jun 22
Jun 29
Date of month
name
brazil_cases
brazil_deaths
chile_cases
chile_deaths
venezuela_cases
venezuela_deaths


Bieu do ca nhiem va tu vong thang 10 nam 2020
Value
4e+06
2e+06
0e+00
Oct 05
Oct 12
Oct 19
Oct 26
Date of month
name
brazil_cases
brazil_deaths
chile_cases
chile_deaths
venezuela_cases
venezuela_deaths


Bieu do ca nhiem va tu vong thang 2 nam 2021
Value
2e+06
1e+06
0e+00
Feb 01
Feb 08
Feb 15
Feb 22
Mar 01
Date of month
name
brazil_cases
brazil_deaths
chile_cases
chile_deaths
venezuela_cases
venezuela_deaths


Bieu do ca nhiem va tu vong thang 3 nam 2021
Value
5e+06
4e+06
3e+06
2e+06
1e+06
0e+00
Mar 01
Mar 08
Mar 15
Mar 22
Mar 29
Date of month
name
brazil_cases
brazil_deaths
chile_cases
chile_deaths
venezuela_cases
venezuela_deaths


Bieu do ca nhiem va tu vong thang 6 nam 2021
Value
9e+06
6e+06
3e+06
0e+00
May 31
Jun 07
Jun 14
Jun 21
Jun 28
Date of month
name
brazil_cases
brazil_deaths
chile_cases
chile_deaths
venezuela_cases
venezuela_deaths



4) Biểu đồ thu thập nhiễm bệnh gồm 2 tháng cuối của năm

- Hiện thực trong R

```
data_2020_11 = subset(data_2020, as.numeric(format(as.Date(data_2020$date),"%m")) == "11")
data_2020_12 = subset(data_2020, as.numeric(format(as.Date(data_2020$date),"%m")) == "12")

data_2021_11 = subset(data_2021, as.numeric(format(as.Date(data_2021$date),"%m")) == "11")
data_2021_12 = subset(data_2021, as.numeric(format(as.Date(data_2021$date),"%m")) == "12")

data_2022_1 = subset(data_2022, as.numeric(format(as.Date(data_2022$date),"%m")) == "1")
data_2022_2 = subset(data_2022, as.numeric(format(as.Date(data_2022$date),"%m")) == "2")

temp <- data_2020_11%>%
  tidyr::pivot_longer(cols = c(brazil_cases,chile_cases,venezuela_cases))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green"))+
  labs(title="Bieu do ca nhiem thang 11 nam 2020",x="Date of month",y="Value")

temp <- data_2020_12%>%
  tidyr::pivot_longer(cols = c(brazil_cases,chile_cases,venezuela_cases))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green"))+
  labs(title="Bieu do ca nhiem thang 12 nam 2020",x="Date of month",y="Value")

temp <- data_2021_11%>%
  tidyr::pivot_longer(cols = c(brazil_cases,chile_cases,venezuela_cases))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green"))+
```

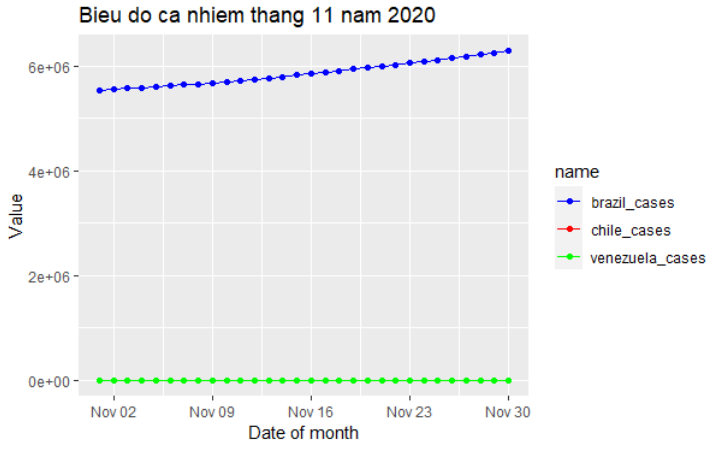
```
  labs(title="Bieu do ca nhiem thang 11 nam 2021",x="Date of month",y="Value")

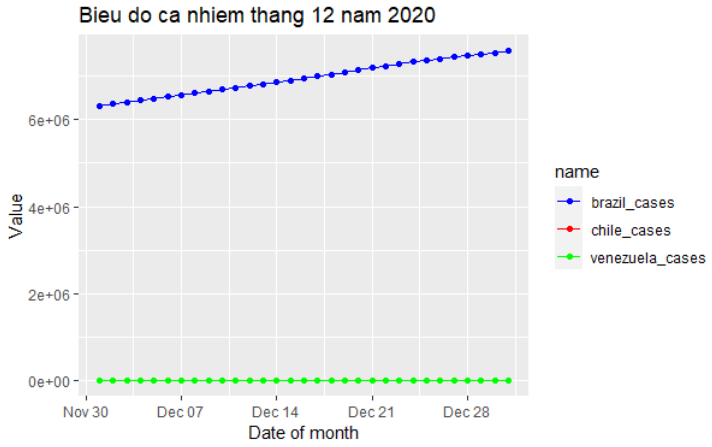
temp <- data_2021_12%>%
  tidyr::pivot_longer(cols = c(brazil_cases,chile_cases,venezuela_cases))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green"))+
  labs(title="Bieu do ca nhiem thang 12 nam 2021",x="Date of month",y="Value")

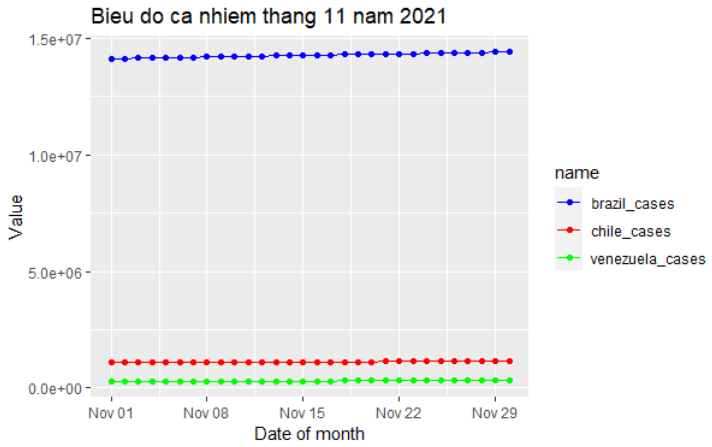
temp <- data_2022_1%>%
tidyr::pivot_longer(cols = c(brazil_cases,chile_cases,venezuela_cases))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green"))+
  labs(title="Bieu do ca nhiem thang 1 nam 2022",x="Date of month",y="Value")

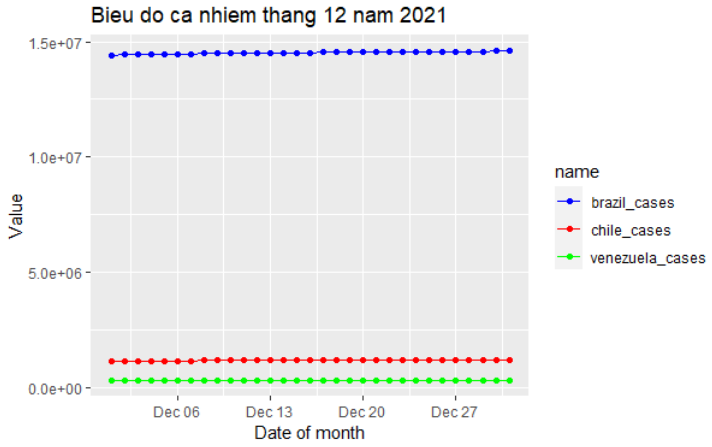
temp <- data_2022_2%>%
  tidyr::pivot_longer(cols = c(brazil_cases,chile_cases,venezuela_cases))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green"))+
  labs(title="Bieu do ca nhiem thang 2 nam 2022",x="Date of month",y="Value")
```

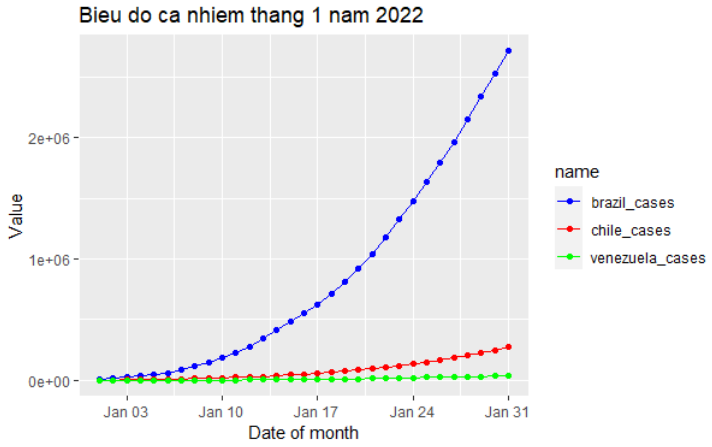
- Kết quả:

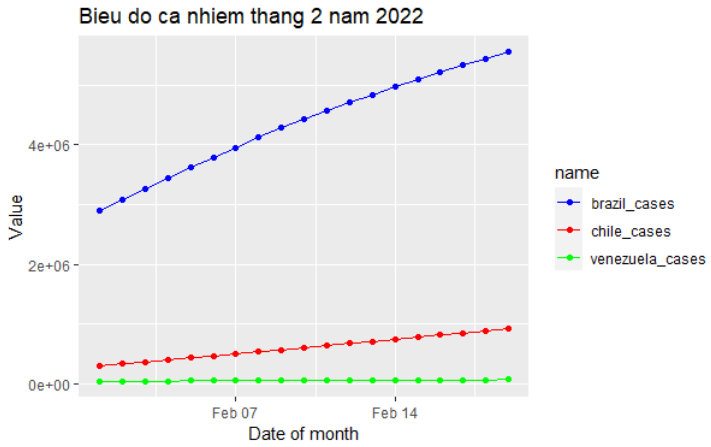





Bieu do ca nhiem thang 11 nam 2021
1.5e+07
1.0e+07
5.0e+06
0.0e+00
Value
Nov 01
Nov 08
Nov 15
Nov 22
Nov 29
Date of month
name
brazil_cases
chile_cases
venezuela_cases


Bieu do ca nhiem thang 12 nam 2021
1.5e+07
1.0e+07
5.0e+06
0.0e+00
Value
Dec 06
Dec 13
Dec 20
Dec 27
Date of month
name
brazil_cases
chile_cases
venezuela_cases


Bieu do ca nhiem thang 1 nam 2022
2e+06
1e+06
0e+00
Value
Jan 03
Jan 10
Jan 17
Jan 24
Jan 31
Date of month
name
brazil_cases
chile_cases
venezuela_cases



5) Biểu đồ thu thập tử vong gồm 2 tháng cuối của năm

- Hiện thực trong R

```
temp <- data_2020_11%>%
  tidyr::pivot_longer(cols = c(brazil_deaths,chile_deaths,venezuela_deaths))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green"))+
  labs(title="Bieu do tu vong thang 11 nam 2020",x="Date of month",y="Value")

temp <- data_2020_12%>%
  tidyr::pivot_longer(cols = c(brazil_deaths,chile_deaths,venezuela_deaths))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green"))+
  labs(title="Bieu do tu vong thang 12 nam 2020",x="Date of month",y="Value")

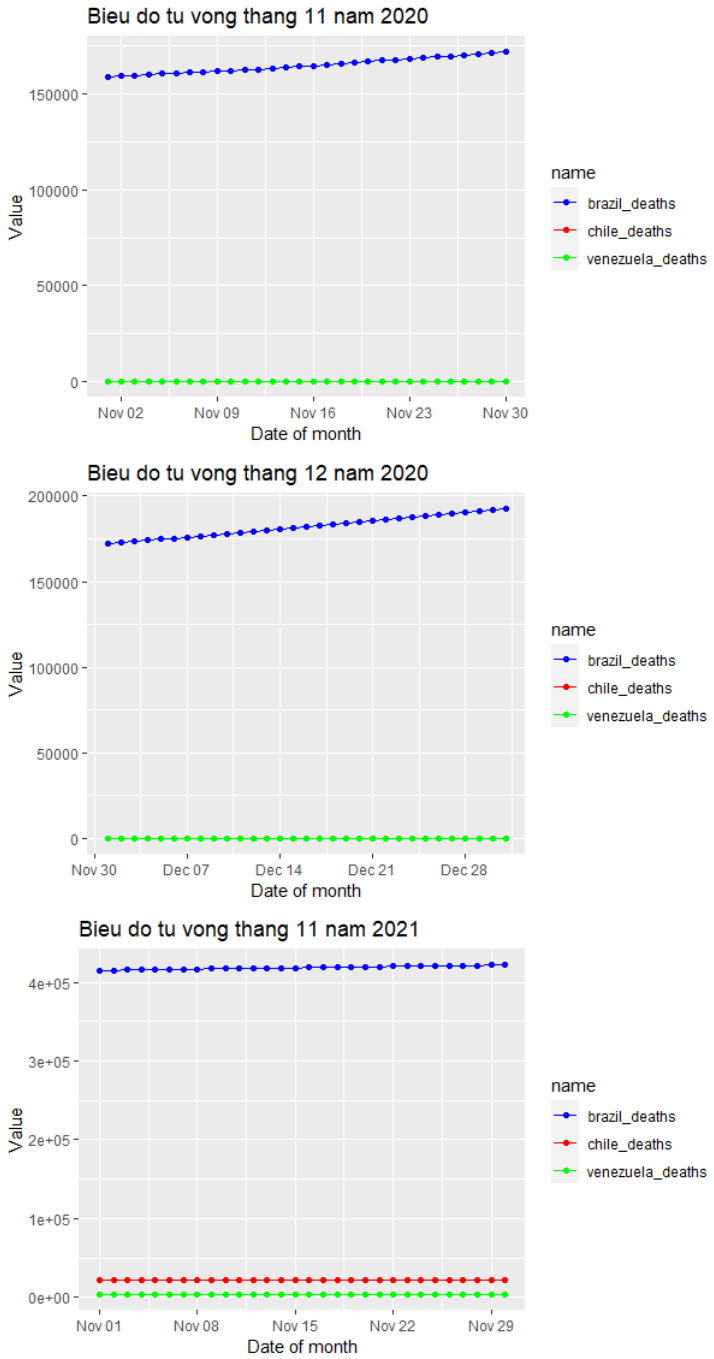
temp <- data_2021_11%>%
  tidyr::pivot_longer(cols = c(brazil_deaths,chile_deaths,venezuela_deaths))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green"))+
  labs(title="Bieu do tu vong thang 11 nam 2021",x="Date of month",y="Value")

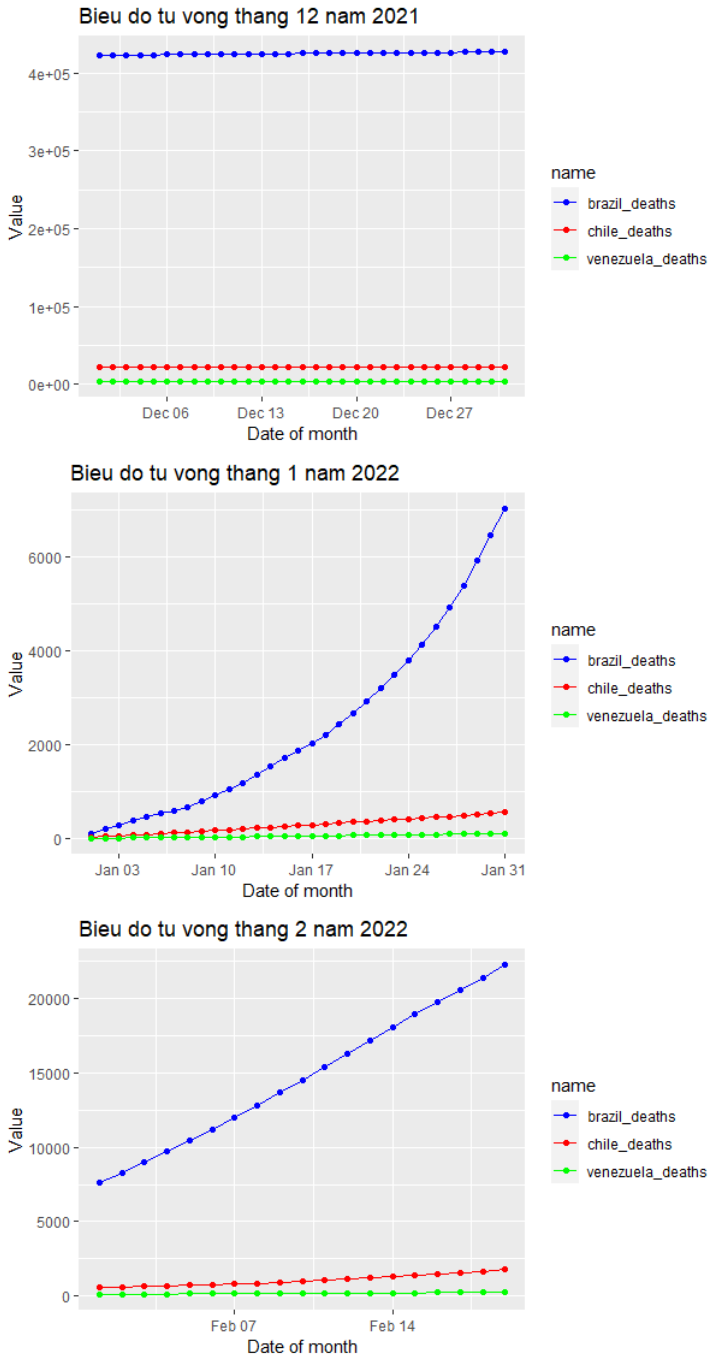
temp <- data_2021_12%>%
  tidyr::pivot_longer(cols = c(brazil_deaths,chile_deaths,venezuela_deaths))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green"))+
  labs(title="Bieu do tu vong thang 12 nam 2021",x="Date of month",y="Value")

temp <- data_2022_1%>%
  tidyr::pivot_longer(cols = c(brazil_deaths,chile_deaths,venezuela_deaths))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green"))+
  labs(title="Bieu do tu vong thang 1 nam 2022",x="Date of month",y="Value")

temp <- data_2022_2%>%
  tidyr::pivot_longer(cols = c(brazil_deaths,chile_deaths,venezuela_deaths))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green"))+
  labs(title="Bieu do tu vong thang 2 nam 2022",x="Date of month",y="Value")
```

- Kết quả:


Bieu do tu vong thang 11 nam 2020
150000
100000
50000
0
Value
Nov 02
Nov 09
Nov 16
Nov 23
Nov 30
Date of month
name
brazil_deaths
chile_deaths
venezuela_deaths
Bieu do tu vong thang 12 nam 2020
200000
150000
100000
50000
0
Value
Nov 30
Dec 07
Dec 14
Dec 21
Dec 28
Date of month
name
brazil_deaths
chile_deaths
venezuela_deaths
Bieu do tu vong thang 11 nam 2021
4e+05
3e+05
2e+05
1e+05
0e+00
Value
Nov 01
Nov 08
Nov 15
Nov 22
Nov 29
Date of month
name
brazil_deaths
chile_deaths
venezuela_deaths



6) Biểu đồ thu thập gồm nhiễm bệnh và tử vong gồm 2 tháng cuối của năm

- Hiện thực trong R

```
temp <- data_2020_11%>%
  tidyr::pivot_longer(cols = c(
     brazil_cases,brazil_deaths,chile_cases,chile_deaths,venezuela_cases,venezuela_deaths))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green","brown","violet","pink"))+
  labs(title="Bieu do ca nhiem va tu vong thang 11 nam 2020",x="Date of month",y="Value")

temp <- data_2020_12%>%
tidyr::pivot_longer(cols = c(
     brazil_cases,brazil_deaths,chile_cases,chile_deaths,venezuela_cases,venezuela_deaths))
ggplot(data=temp,aes(x=date, y=value,color=name))+
```

```
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green","brown","violet","pink"))+
  labs(title="Bieu do ca nhiem va tu vong thang 12 nam 2020",x="Date of month",y="Value")

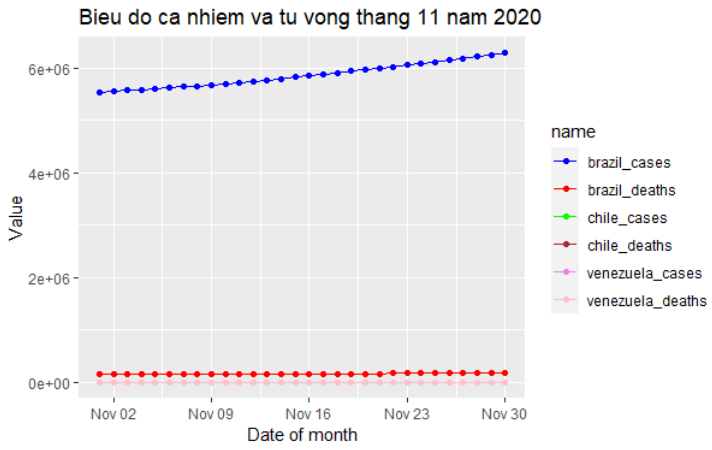
temp <- data_2021_11%>%
  tidyr::pivot_longer(cols = c(
    brazil_cases,brazil_deaths,chile_cases,chile_deaths,venezuela_cases,venezuela_deaths))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green","brown","violet","pink"))+
  labs(title="Bieu do ca nhiem va tu vong thang 11 nam 2021",x="Date of month",y="Value")

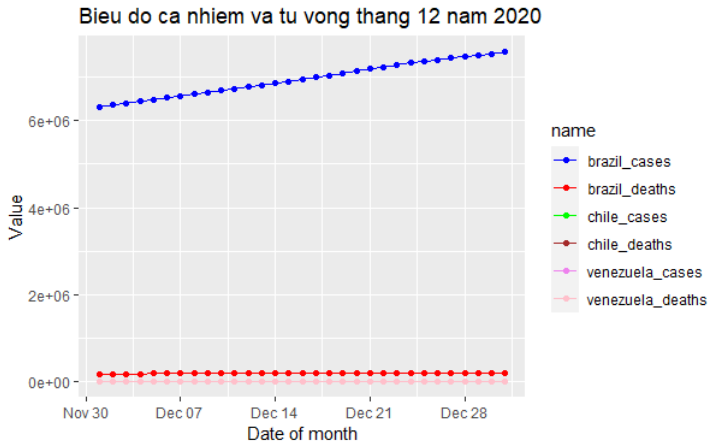
temp <- data_2021_12%>%
  tidyr::pivot_longer(cols = c(
    brazil_cases,brazil_deaths,chile_cases,chile_deaths,venezuela_cases,venezuela_deaths))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green","brown","violet","pink"))+
  labs(title="Bieu do ca nhiem va tu vong thang 12 nam 2021",x="Date of month",y="Value")

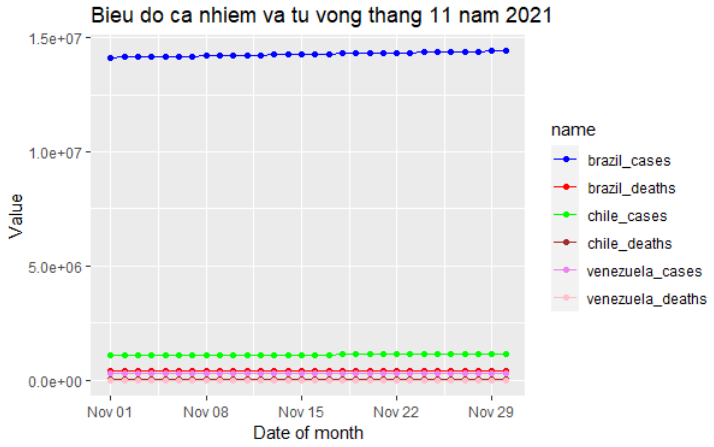
temp <- data_2022_1%>%
  tidyr::pivot_longer(cols = c(
    brazil_cases,brazil_deaths,chile_cases,chile_deaths,venezuela_cases,venezuela_deaths))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green","brown","violet","pink"))+
  labs(title="Bieu do ca nhiem va tu vong thang 1 nam 2022",x="Date of month",y="Value")

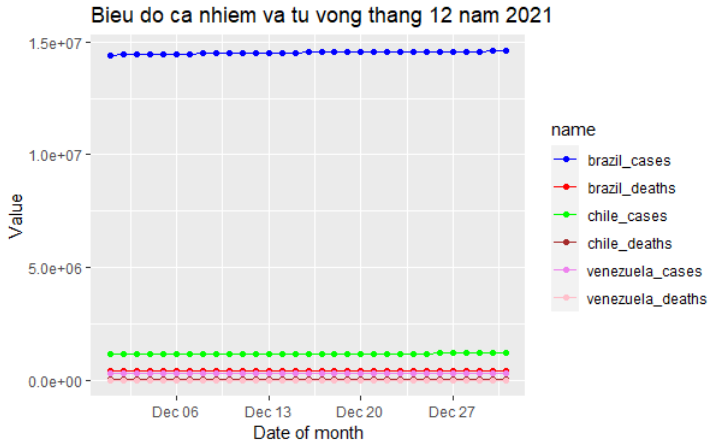
temp <- data_2022_2%>%
  tidyr::pivot_longer(cols = c(
    brazil_cases,brazil_deaths,chile_cases,chile_deaths,venezuela_cases,venezuela_deaths))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green","brown","violet","pink"))+
  labs(title="Bieu do ca nhiem va tu vong thang 2 nam 2022",x="Date of month",y="Value")
```

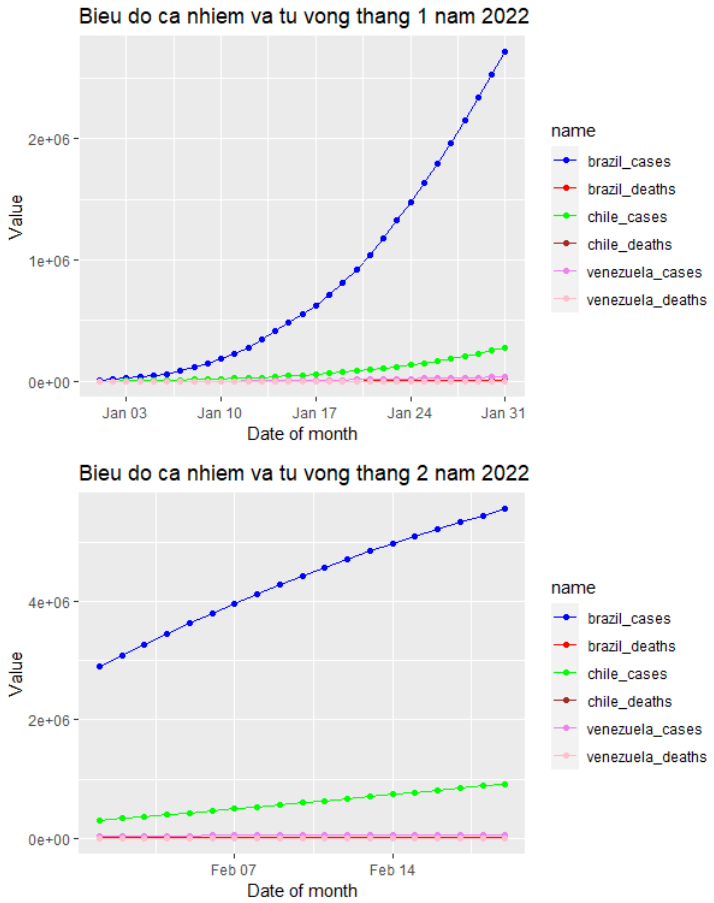
- Kết quả:




Bieu do ca nhiem va tu vong thang 12 nam 2020
Value
6e+06
4e+06
2e+06
0e+00
Nov 30
Dec 07
Dec 14
Dec 21
Dec 28
Date of month
name
brazil_cases
brazil_deaths
chile_cases
chile_deaths
venezuela_cases
venezuela_deaths


Bieu do ca nhiem va tu vong thang 11 nam 2021
Value
1.5e+07
1.0e+07
5.0e+06
0.0e+00
Nov 01
Nov 08
Nov 15
Nov 22
Nov 29
Date of month
name
brazil_cases
brazil_deaths
chile_cases
chile_deaths
venezuela_cases
venezuela_deaths


Bieu do ca nhiem va tu vong thang 12 nam 2021
Value
1.5e+07
1.0e+07
5.0e+06
0.0e+00
Dec 06
Dec 13
Dec 20
Dec 27
Date of month
name
brazil_cases
brazil_deaths
chile_cases
chile_deaths
venezuela_cases
venezuela_deaths



7) Biểu đồ thu thập nhiễm bệnh tích lũy cho từng tháng

- Hiện thực trong R

```
data_2020[is.na(data_2020)] =0
data_2020$brazil_cases = cumsum(data_2020$brazil_cases)
data_2020$brazil_deaths = cumsum(data_2020$brazil_deaths)
data_2021$brazil_cases = cumsum(data_2021$brazil_cases)
data_2021$brazil_deaths = cumsum(data_2021$brazil_deaths)
data_2022$brazil_cases = cumsum(data_2022$brazil_cases)
data_2022$brazil_deaths = cumsum(data_2022$brazil_deaths)

data_2020$chile_cases = cumsum(data_2020$chile_cases)
data_2020$chile_deaths = cumsum(data_2020$chile_deaths)
data_2021$chile_cases = cumsum(data_2021$chile_cases)
data_2021$chile_deaths = cumsum(data_2021$chile_deaths)
data_2022$chile_cases = cumsum(data_2022$chile_cases)
data_2022$chile_deaths = cumsum(data_2022$chile_deaths)

data_2020$venezuela_cases = cumsum(data_2020$venezuela_cases)
data_2020$venezuela_deaths = cumsum(data_2020$venezuela_deaths)
data_2021$venezuela_cases = cumsum(data_2021$venezuela_cases)
data_2021$venezuela_deaths = cumsum(data_2021$venezuela_deaths)
data_2022$venezuela_cases = cumsum(data_2022$venezuela_cases)
data_2022$venezuela_deaths = cumsum(data_2022$venezuela_deaths)

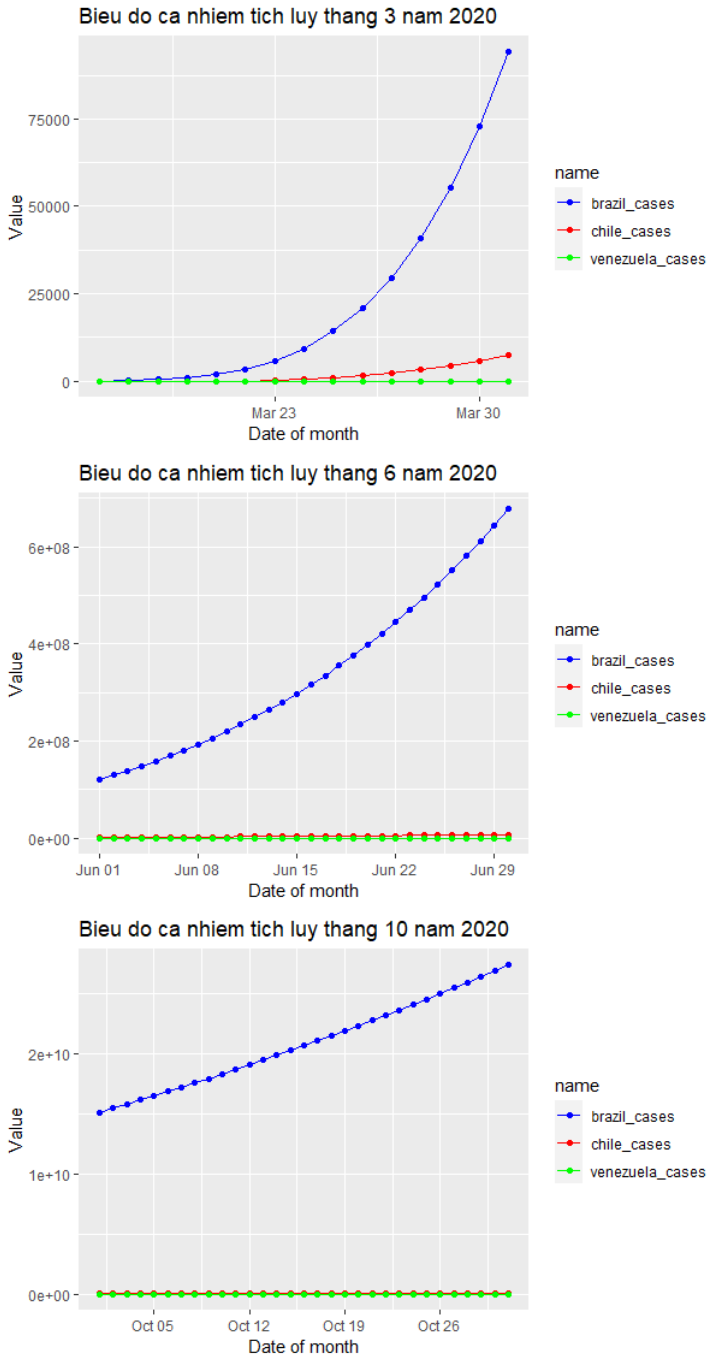
cumsum_2020_3 = subset(data_2020, as.numeric(format(as.Date(data_2020$date),"%m")) == "3")
cumsum_2020_6 = subset(data_2020, as.numeric(format(as.Date(data_2020$date),"%m")) == "6")
cumsum_2020_10 = subset(data_2020, as.numeric(format(as.Date(data_2020$date),"%m")) == "10")

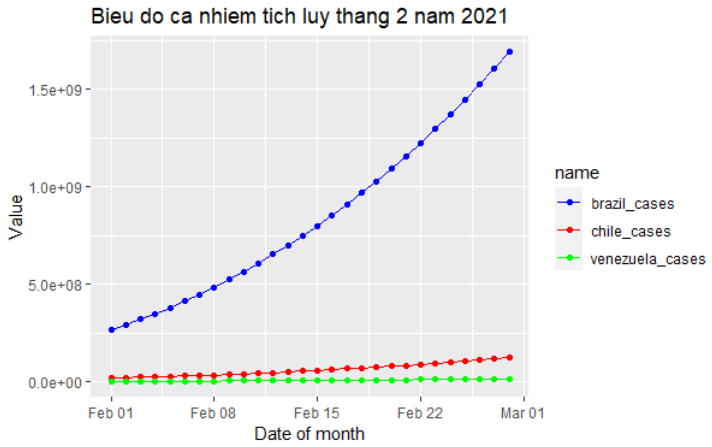
cumsum_2021_2 = subset(data_2021, as.numeric(format(as.Date(data_2021$date),"%m")) == "2")
cumsum_2021_3 = subset(data_2021, as.numeric(format(as.Date(data_2021$date),"%m")) == "3")
cumsum_2021_6 = subset(data_2021, as.numeric(format(as.Date(data_2021$date),"%m")) == "6")
cumsum_2021_10 = subset(data_2021, as.numeric(format(as.Date(data_2021$date),"%m")) == "10")

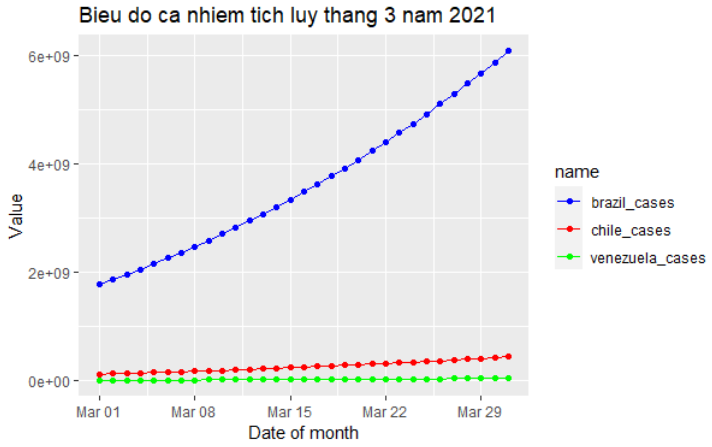
```

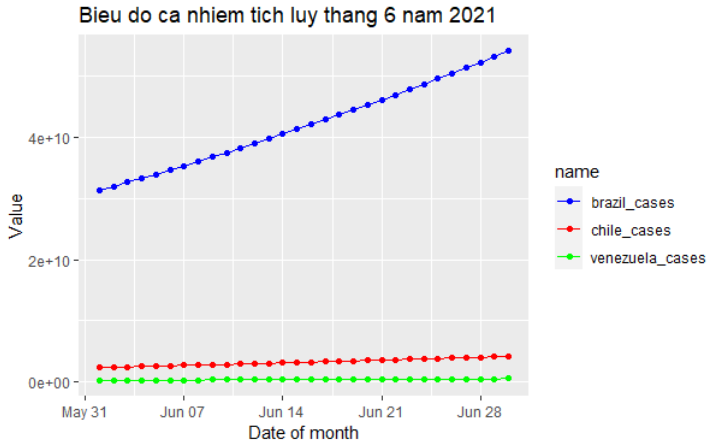
```r
cumsum_2022_2 = subset(data_2022, as.numeric(format(as.Date(data_2022$date),"%m")) == "2")

temp <- cumsum_2020_3%>%
  tidyr::pivot_longer(cols = c(brazil_cases,chile_cases,venezuela_cases))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green"))+
  labs(title="Bieu do ca nhiem tich luy thang 3 nam 2020",x="Date of month",y="Value")

temp <- cumsum_2020_6%>%
  tidyr::pivot_longer(cols = c(brazil_cases,chile_cases,venezuela_cases))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green"))+
  labs(title="Bieu do ca nhiem tich luy thang 6 nam 2020",x="Date of month",y="Value")

temp <- cumsum_2020_10%>%
  tidyr::pivot_longer(cols = c(brazil_cases,chile_cases,venezuela_cases))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green"))+
  labs(title="Bieu do ca nhiem tich luy thang 10 nam 2020",x="Date of month",y="Value")

temp <- cumsum_2021_2%>%
  tidyr::pivot_longer(cols = c(brazil_cases,chile_cases,venezuela_cases))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green"))+
  labs(title="Bieu do ca nhiem tich luy thang 2 nam 2021",x="Date of month",y="Value")

temp <- cumsum_2021_3%>%
  tidyr::pivot_longer(cols = c(brazil_cases,chile_cases,venezuela_cases))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green"))+
  labs(title="Bieu do ca nhiem tich luy thang 3 nam 2021",x="Date of month",y="Value")

temp <- cumsum_2021_6%>%
  tidyr::pivot_longer(cols = c(brazil_cases,chile_cases,venezuela_cases))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green"))+
  labs(title="Bieu do ca nhiem tich luy thang 6 nam 2021",x="Date of month",y="Value")

temp <- cumsum_2021_10%>%
  tidyr::pivot_longer(cols = c(brazil_cases,chile_cases,venezuela_cases))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green"))+
  labs(title="Bieu do ca nhiem tich luy thang 10 nam 2021",x="Date of month",y="Value")

temp <- cumsum_2022_2%>%
  tidyr::pivot_longer(cols = c(brazil_cases,chile_cases,venezuela_cases))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green"))+
  labs(title="Bieu do ca nhiem tich luy thang 2 nam 2022",x="Date of month",y="Value")
```

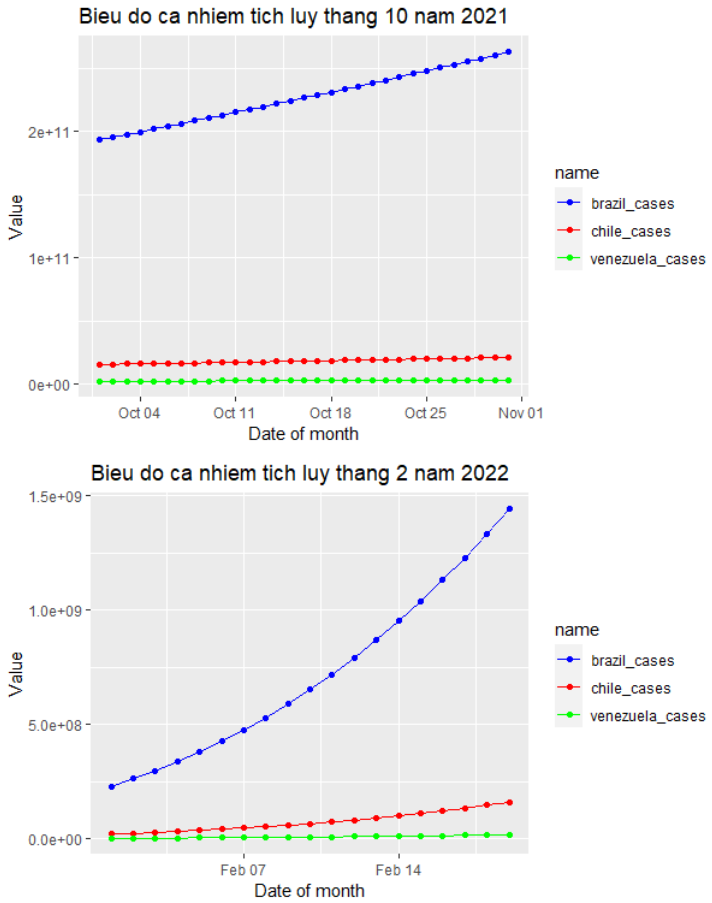
- Kết quả:


Bieu do ca nhiem tich luy thang 3 nam 2020
75000
50000
25000
0
Value
Mar 23
Mar 30
Date of month
name
brazil_cases
chile_cases
venezuela_cases
Bieu do ca nhiem tich luy thang 6 nam 2020
6e+08
4e+08
2e+08
0e+00
Value
Jun 01
Jun 08
Jun 15
Jun 22
Jun 29
Date of month
name
brazil_cases
chile_cases
venezuela_cases
Bieu do ca nhiem tich luy thang 10 nam 2020
2e+10
1e+10
0e+00
Value
Oct 05
Oct 12
Oct 19
Oct 26
Date of month
name
brazil_cases
chile_cases
venezuela_cases


Bieu do ca nhiem tich luy thang 2 nam 2021
1.5e+09
1.0e+09
5.0e+08
0.0e+00
Value
Feb 01
Feb 08
Feb 15
Feb 22
Mar 01
Date of month
name
brazil_cases
chile_cases
venezuela_cases


Bieu do ca nhiem tich luy thang 3 nam 2021
6e+09
4e+09
2e+09
0e+00
Value
Mar 01
Mar 08
Mar 15
Mar 22
Mar 29
Date of month
name
brazil_cases
chile_cases
venezuela_cases


Bieu do ca nhiem tich luy thang 6 nam 2021
4e+10
2e+10
0e+00
Value
May 31
Jun 07
Jun 14
Jun 21
Jun 28
Date of month
name
brazil_cases
chile_cases
venezuela_cases



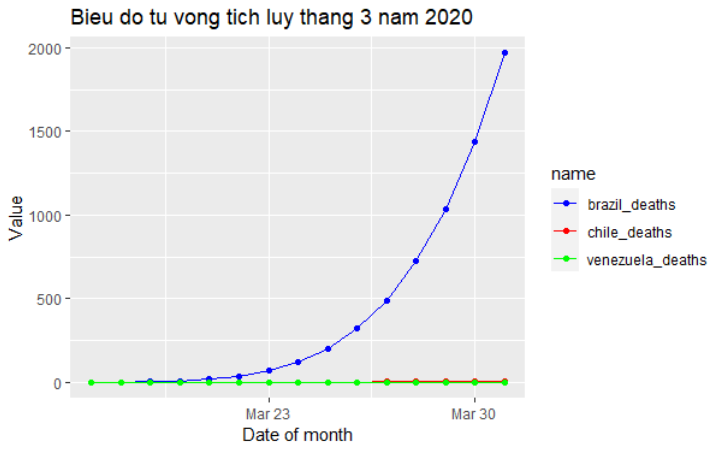
8) Biểu đồ thu thập tử vong tích lũy cho từng tháng

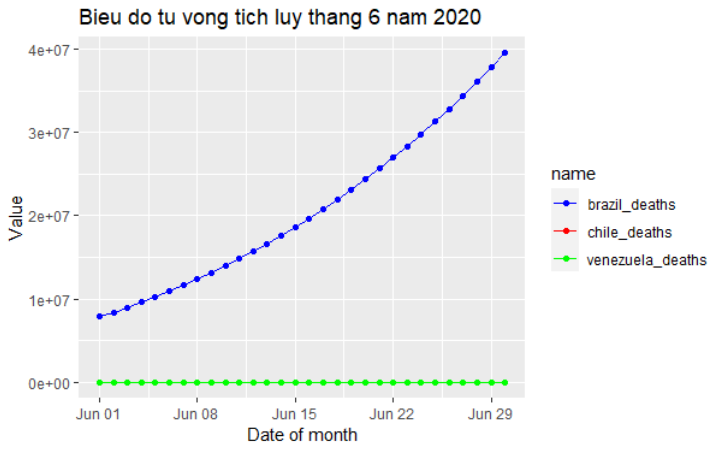
- Hiện thực trong R

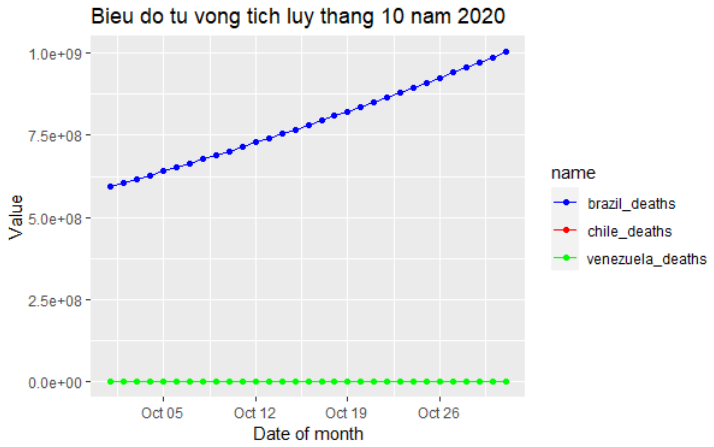
```
temp <- cumsum_2020_3%>%
  tidyr::pivot_longer(cols = c(brazil_deaths,chile_deaths,venezuela_deaths))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green"))+
  labs(title="Bieu do tu vong tich luy thang 3 nam 2020",x="Date of month",y="Value")

temp <- cumsum_2020_6%>%
  tidyr::pivot_longer(cols = c(brazil_deaths,chile_deaths,venezuela_deaths))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green"))+
  labs(title="Bieu do tu vong tich luy thang 6 nam 2020",x="Date of month",y="Value")

temp <- cumsum_2020_10%>%
  tidyr::pivot_longer(cols = c(brazil_deaths,chile_deaths,venezuela_deaths))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green"))+
  labs(title="Bieu do tu vong tich luy thang 10 nam 2020",x="Date of month",y="Value")

temp <- cumsum_2021_2%>%
  tidyr::pivot_longer(cols = c(brazil_deaths,chile_deaths,venezuela_deaths))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green"))+
  labs(title="Bieu do tu vong tich luy thang 2 nam 2021",x="Date of month",y="Value")
```

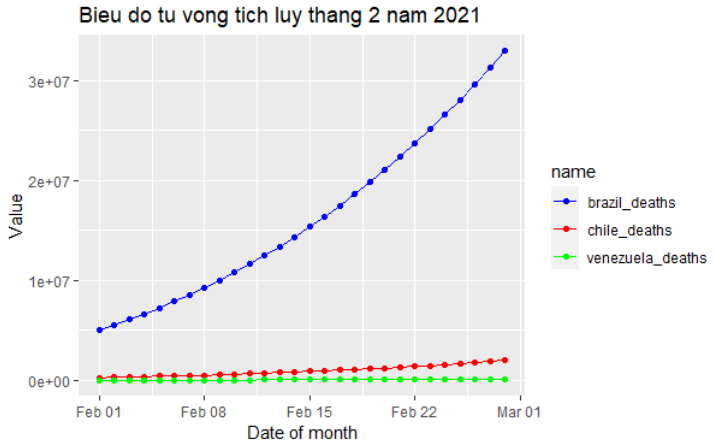
```

temp <- cumsum_2021_3%>%
  tidyr::pivot_longer(cols = c(brazil_deaths,chile_deaths,venezuela_deaths))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green"))+
  labs(title="Bieu do tu vong tich luy thang 3 nam 2021",x="Date of month",y="Value")

temp <- cumsum_2021_6%>%
  tidyr::pivot_longer(cols = c(brazil_deaths,chile_deaths,venezuela_deaths))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green"))+
  labs(title="Bieu do tu vong tich luy thang 6 nam 2021",x="Date of month",y="Value")

temp <- cumsum_2021_10%>%
  tidyr::pivot_longer(cols = c(brazil_deaths,chile_deaths,venezuela_deaths))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green"))+
  labs(title="Bieu do tu vong tich luy thang 10 nam 2021",x="Date of month",y="Value")

temp <- cumsum_2022_2%>%
  tidyr::pivot_longer(cols = c(brazil_deaths,chile_deaths,venezuela_deaths))
ggplot(data=temp,aes(x=date, y=value,color=name))+
  geom_line()+
  geom_point()+
  scale_color_manual(values =c("blue","red","green"))+
  labs(title="Bieu do tu vong tich luy thang 2 nam 2022",x="Date of month",y="Value")
```

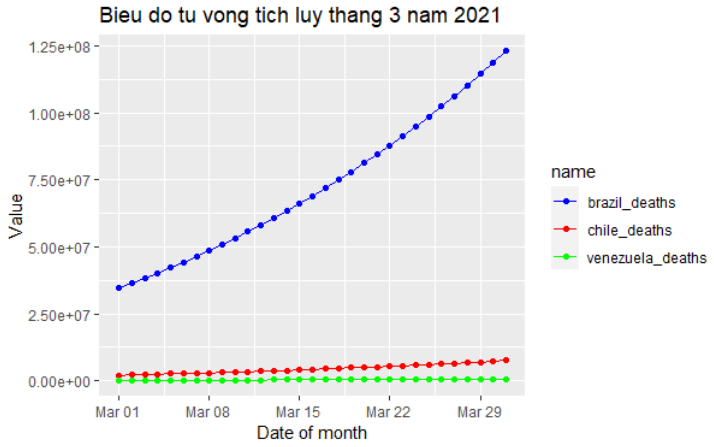
- Kết quả:

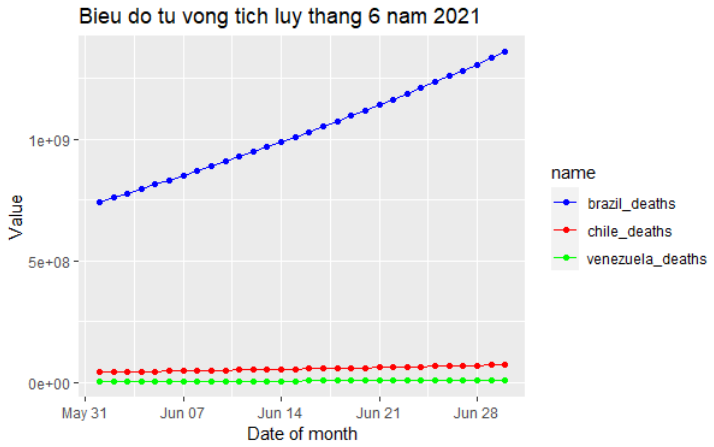


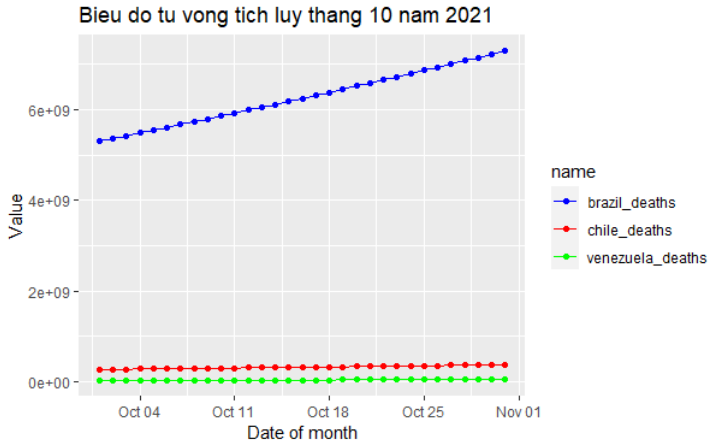


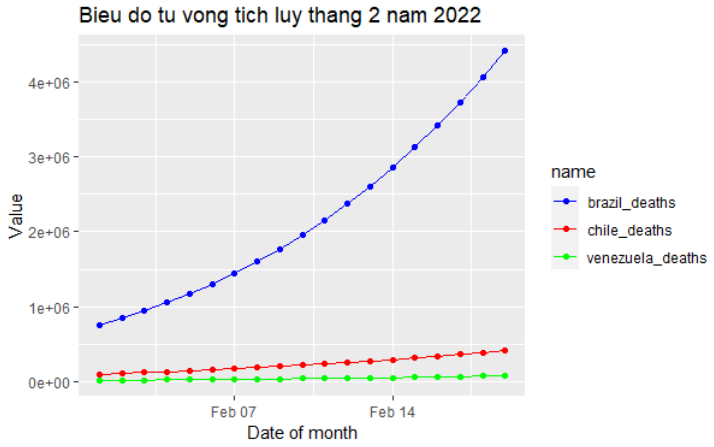

Bieu do tu vong tich luy thang 10 nam 2020
1.0e+09
7.5e+08
5.0e+08
2.5e+08
0.0e+00
Value
Oct 05
Oct 12
Oct 19
Oct 26
Date of month
name
brazil_deaths
chile_deaths
venezuela_deaths


Bieu do tu vong tich luy thang 2 nam 2021
3e+07
2e+07
1e+07
0e+00
Value
Feb 01
Feb 08
Feb 15
Feb 22
Mar 01
Date of month
name
brazil_deaths
chile_deaths
venezuela_deaths


Bieu do tu vong tich luy thang 3 nam 2021
1.25e+08
1.00e+08
7.50e+07
5.00e+07
2.50e+07
0.00e+00
Value
Mar 01
Mar 08
Mar 15
Mar 22
Mar 29
Date of month
name
brazil_deaths
chile_deaths
venezuela_deaths







vii) Nhóm câu hỏi liên quan đến tất cả quốc gia theo thời gian là tháng

- Trên từng năm hãy vẽ biểu đồ thể hiện trục Ox là thời gian, trục Oy là nhiễm bệnh/tử vong. Hãy dùng 4 ký số của mã đề để vẽ 4 tháng tương ứng theo ký số đó. Nếu ký số là 0 thì lấy tháng là 10. Số đường thể hiện trên đồ thị là số quốc gia.

  - Chuẩn bị dữ liệu cho toàn bộ phần vii (phần code dùng chung cho cả phần vii)

```
library(tidyverse)
library(lubridate)
library(dplyr)
library(xlsx)
setwd("E:/BTL_CTRR")
data_covid = read.csv("owid-covid-data.csv", header = TRUE)
library("ggplot2")
data_covid = mutate(data_covid, new_cases = abs(new_cases))
data_covid = mutate(data_covid, new_deaths = abs(new_deaths))
```

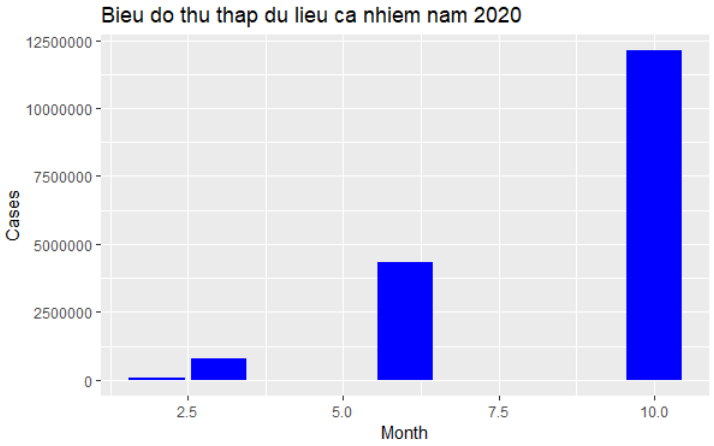
```
data_covid <- subset(data_covid,data_covid$continent != "")
data_covid[is.na(data_covid)] <- 0
data <- data_covid %>%
  separate(date, into = c("month", "day", "year"), sep = "/")
data <- transform(data,month=as.numeric(month))

year_2020 <- data[data$year=="2020",]
year_2021 <- data[data$year=="2021",]
year_2022 <- data[data$year=="2022",]

year_2020 <- aggregate(cbind(new_cases,new_deaths) ~ month, data = year_2020, FUN = sum)
year_2021 <- aggregate(cbind(new_cases,new_deaths) ~ month, data = year_2021, FUN = sum)
year_2022 <- aggregate(cbind(new_cases,new_deaths) ~ month, data = year_2022, FUN = sum)

data_2020<-subset(year_2020,year_2020$month==2 | year_2020$month==10 |
                    year_2020$month==3 | year_2020$month==6)
data_2021<-subset(year_2021,year_2021$month==2 | year_2021$month==10 |
                    year_2021$month==3 | year_2021$month==6)
data_2022<-subset(year_2022,year_2022$month==2 | year_2022$month==10 |
                    year_2022$month==3 | year_2022$month==6)

data_2020_1<-subset(year_2020,year_2020$month==11 | year_2020$month==12)
data_2021_1<-subset(year_2021,year_2021$month==11 | year_2021$month==12)
data_2022_1<-subset(year_2022,year_2022$month==2 | year_2022$month==1)
```

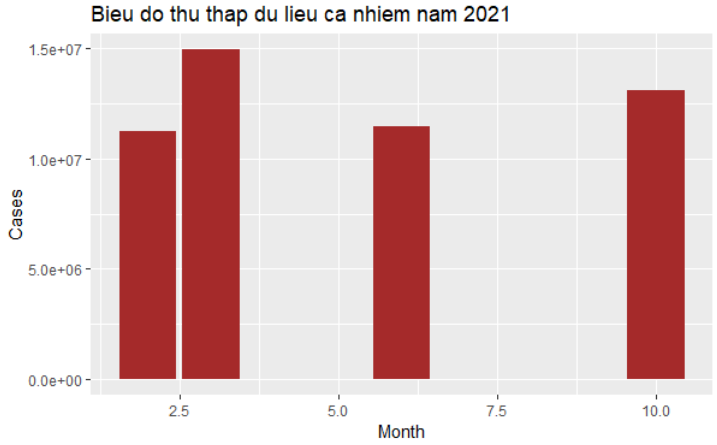
1) Biểu đồ thể hiện thu thập dữ liệu nhiễm bệnh theo thời gian là tháng của tất cả quốc gia

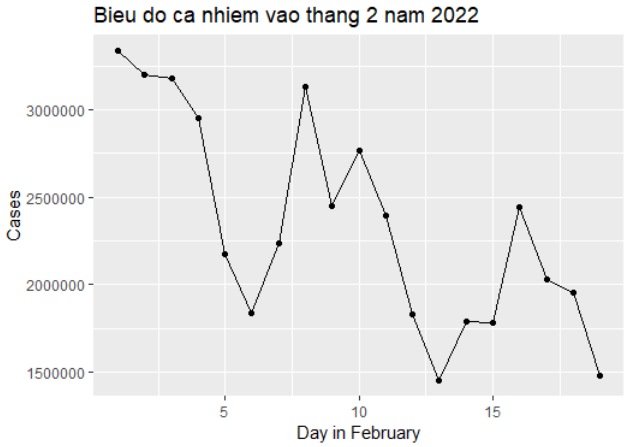
- Hiện thực trong R

```
ggplot(data_2020, aes(y=new_cases, x=month)) +
  geom_bar(position='dodge', stat='identity', fill ="blue") +
  labs(x='Month', y='Cases', title='Bieu do thu thap du lieu ca nhiem nam 2020')

ggplot(data_2021, aes(y=new_cases, x=month)) +
  geom_bar(position='dodge', stat='identity', fill ="brown") +
  labs(x='Month', y='Cases', title='Bieu do thu thap du lieu ca nhiem nam 2021')

temp = data[data$year=="2022",]
temp = temp[temp$month==2,]
data_2020_new<- aggregate(cbind(new_cases,new_deaths) ~ day, data = temp, FUN = sum)
data_2020_new <- arrange(data_2020_new,day=as.numeric(day))
ggplot(data=data_2020_new,aes(x=as.numeric(day), y=new_cases))+
  geom_line()+
  geom_point()+
  labs(title="Bieu do ca nhiem vao thang 2 nam 2022",x="Day in February",y="Cases")
```

- Kết quả:





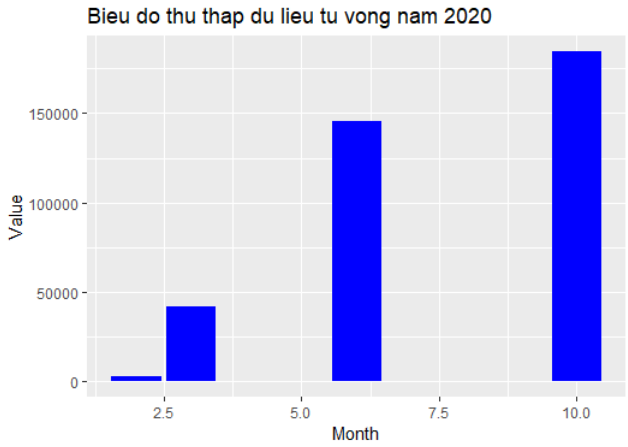


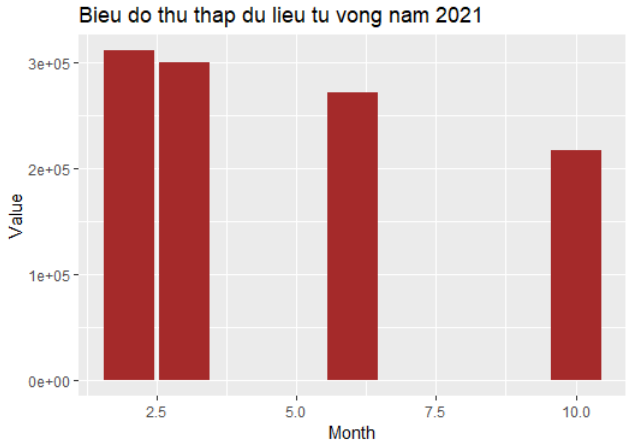
2) Biểu đồ thể hiện thu thập dữ liệu tử vong theo thời gian là tháng của tất cả quốc gia

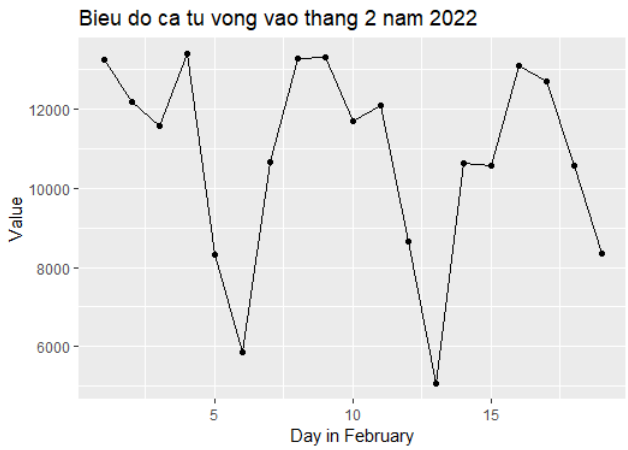
- Hiện thực trong R

```
ggplot(data_2020, aes(y=new_deaths, x=month)) +
  geom_bar(position='dodge', stat='identity', fill ="blue") +
  labs(x='Month', y='Value', title='Bieu do thu thap du lieu tu vong nam 2020')

ggplot(data_2021, aes(y=new_deaths, x=month)) +
  geom_bar(position='dodge', stat='identity', fill ="brown") +
  labs(x='Month', y='Value', title='Bieu do thu thap du lieu tu vong nam 2021')

ggplot(data=data_2020_new,aes(x=as.numeric(day), y=new_deaths))+
  geom_line()+
  geom_point()+
  labs(title="Bieu do ca tu vong vao thang 2 nam 2022",x="Day in February",y="Value")
```

- Kết quả:





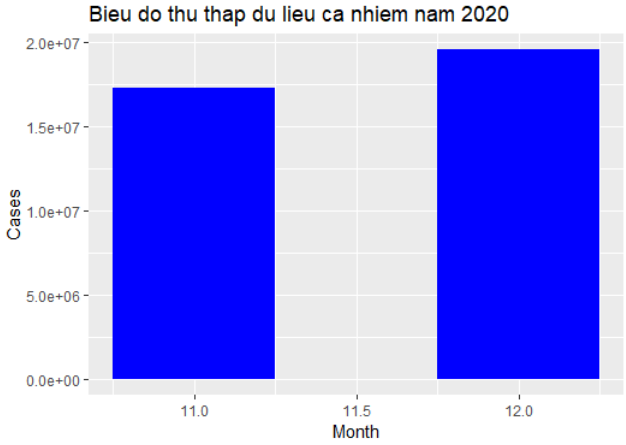


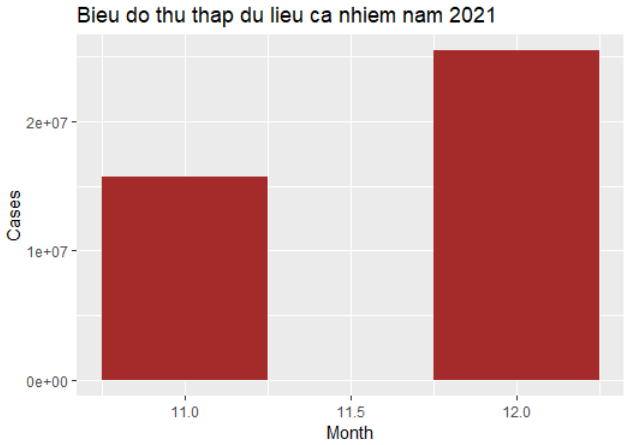
3) Biểu đồ thể hiện thu thập dữ liệu nhiễm bệnh theo thời gian là 2 tháng cuối của năm của tất cả quốc gia

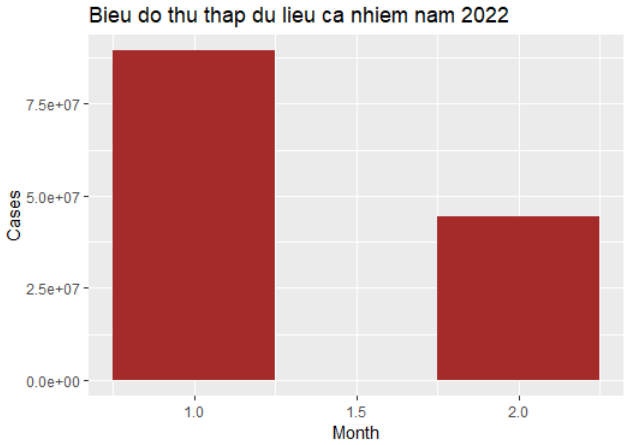
- Hiện thực trong R

```
ggplot(data_2020_1, aes(y=new_cases, x=month,width = 0.5)) +
  geom_bar(position='dodge', stat='identity', fill ="blue") +
  labs(x='Month', y='Cases', title='Bieu do thu thap du lieu ca nhiem nam 2020')

ggplot(data_2021_1, aes(y=new_cases, x=month,width = 0.5)) +
  geom_bar(position='dodge', stat='identity', fill ="brown") +
  labs(x='Month', y='Cases', title='Bieu do thu thap du lieu ca nhiem nam 2021')

ggplot(data_2022_1, aes(y=new_cases, x=month,width = 0.5)) +
  geom_bar(position='dodge', stat='identity', fill ="brown") +
  labs(x='Month', y='Cases', title='Bieu do thu thap du lieu ca nhiem nam 2022')
```

- Kết quả:





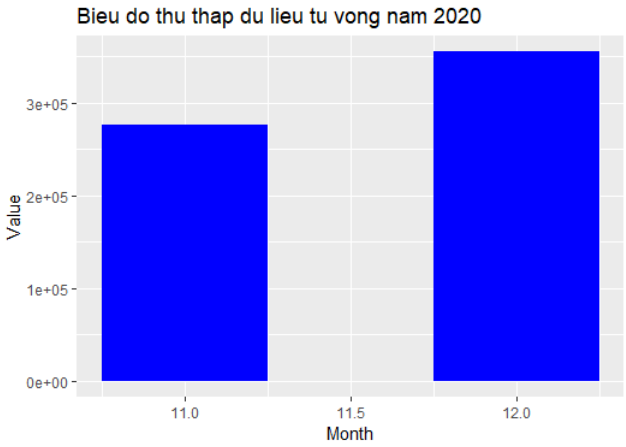


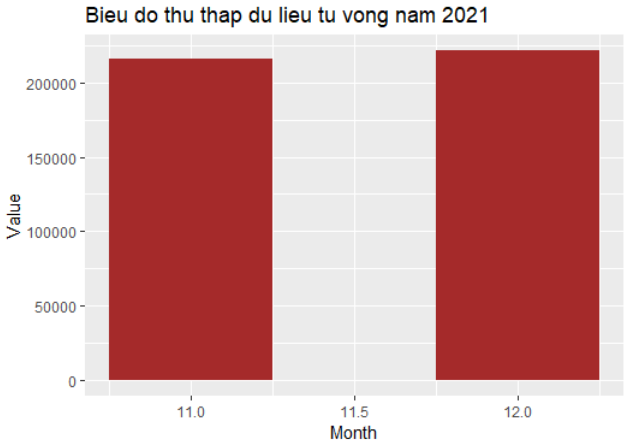
4) Biểu đồ thể hiện thu thập dữ liệu tử vong theo thời gian là 2 tháng cuối của năm của tất cả quốc gia

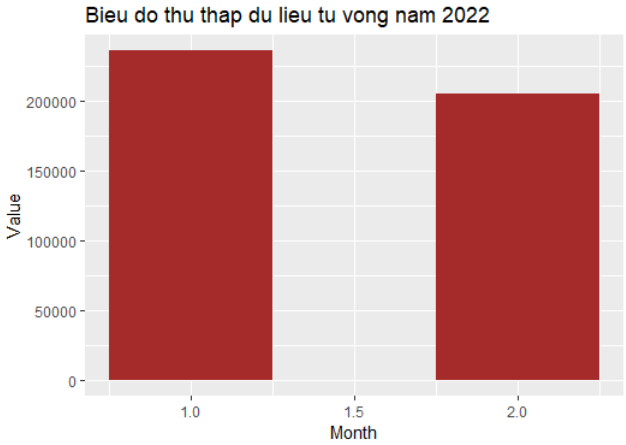
- Hiện thực trong R

```
ggplot(data_2020_1, aes(y=new_deaths, x=month, width = 0.5)) +
  geom_bar(position='dodge', stat='identity', fill ="blue") +
  labs(x='Month', y='Value', title='Bieu do thu thap du lieu tu vong nam 2020')

ggplot(data_2021_1, aes(y=new_deaths, x=month,width = 0.5)) +
  geom_bar(position='dodge', stat='identity', fill ="brown") +
  labs(x='Month', y='Value', title='Bieu do thu thap du lieu tu vong nam 2021')

ggplot(data_2022_1, aes(y=new_deaths, x=month,width = 0.5)) +
  geom_bar(position='dodge', stat='identity', fill ="brown") +
  labs(x='Month', y='Value', title='Bieu do thu thap du lieu tu vong nam 2022')
```

- Kết quả:







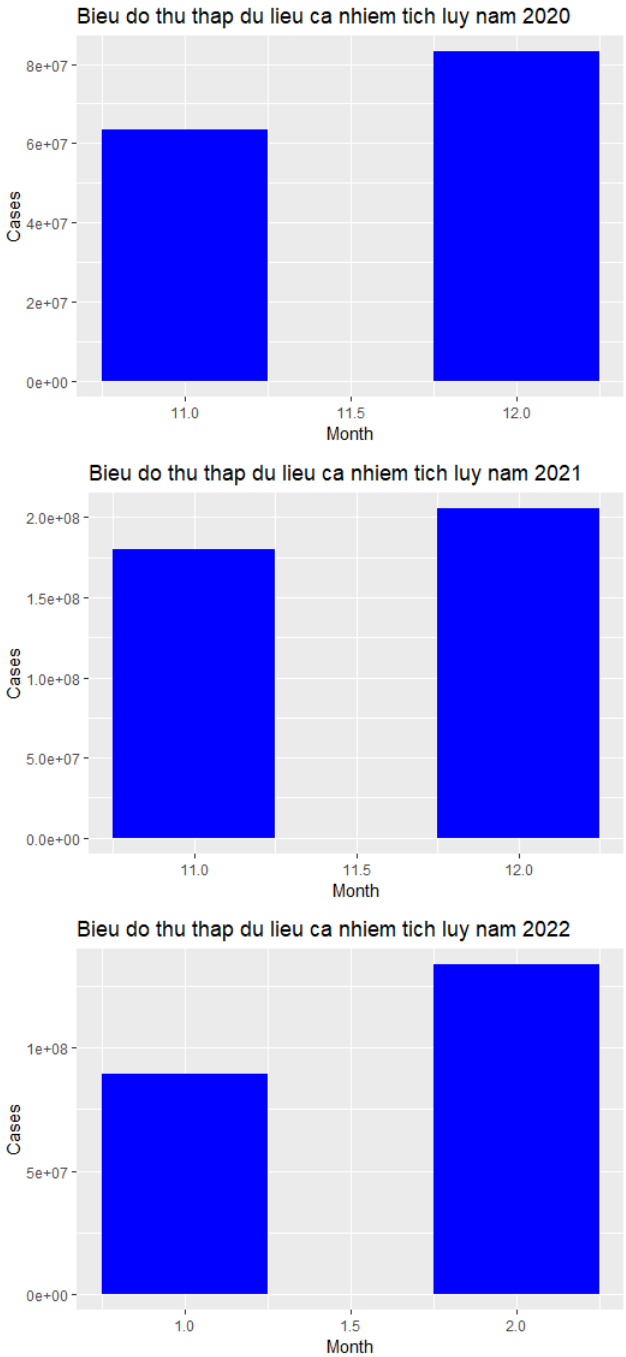
5) Biểu đồ thể hiện thu thập dữ liệu nhiễm bệnh tương đối tích lũy theo thời gian là 2 tháng cuối của năm của tất cả quốc gia

- Hiện thực trong R

```
cumsum_2020=as.data.frame(lapply(year_2020,cumsum))
cumsum_2020$month = c(1:12)
cumsum_2021=as.data.frame(lapply(year_2021,cumsum))
cumsum_2021$month = c(1:12)
cumsum_2022=as.data.frame(lapply(year_2022,cumsum))
cumsum_2022$month = c(1,2)

cumsum_2020_1<-subset(cumsum_2020,cumsum_2020$month==11 | cumsum_2020$month==12)
cumsum_2021_1<-subset(cumsum_2021,cumsum_2021$month==11 | cumsum_2021$month==12)
cumsum_2022_1<-subset(cumsum_2022,cumsum_2022$month==2 | cumsum_2022$month==1)

```

```
ggplot(cumsum_2020_1, aes(y=new_cases, x=month, width = 0.5)) +
  geom_bar(position='dodge', stat='identity', fill = "blue") +
  labs(x='Month', y='Cases', title='Bieu do thu thap du lieu ca nhiem tich luy nam 2020')

ggplot(cumsum_2021_1, aes(y=new_cases, x=month, width = 0.5)) +
  geom_bar(position='dodge', stat='identity', fill = "blue") +
  labs(x='Month', y='Cases', title='Bieu do thu thap du lieu ca nhiem tich luy nam 2021')

ggplot(cumsum_2022_1, aes(y=new_cases, x=month, width = 0.5)) +
  geom_bar(position='dodge', stat='identity', fill = "blue") +
  labs(x='Month', y='Cases', title='Bieu do thu thap du lieu ca nhiem tich luy nam 2022')
```

- Kết quả:

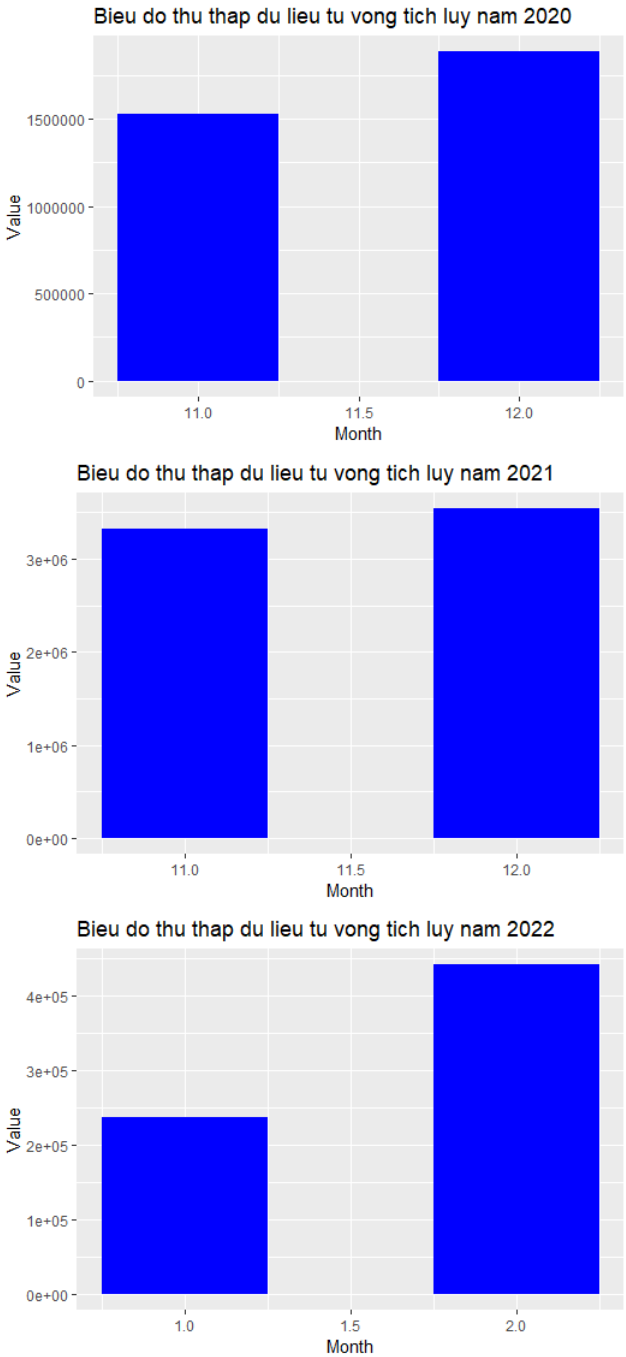


6) Biểu đồ thu thập tử vong tương đối tích lũy theo thời gian là 2 tháng cuối của tất cả quốc gia

- Hiện thực trong R

```
ggplot(cumsum_2020_1, aes(y=new_deaths, x=month, width = 0.5)) +
  geom_bar(position='dodge', stat='identity', fill = "blue") +
  labs(x='Month', y='Value', title='Bieu do thu thap du lieu tu vong tich luy nam 2020')

ggplot(cumsum_2021_1, aes(y=new_deaths, x=month, width = 0.5)) +
  geom_bar(position='dodge', stat='identity', fill = "blue") +
  labs(x='Month', y='Value', title='Bieu do thu thap du lieu tu vong tich luy nam 2021')

ggplot(cumsum_2022_1, aes(y=new_deaths, x=month, width = 0.5)) +
  geom_bar(position='dodge', stat='identity', fill = "blue") +
  labs(x='Month', y='Value', title='Bieu do thu thap du lieu tu vong tich luy nam 2022')
```

- Kết quả:



viii) Nhóm câu hỏi liên quan đến tất cả quốc gia theo trung bình 7 ngày gần nhất
Trên từng năm hãy vẽ biểu đồ thể hiện trục Ox là thời gian, trục Oy là nhiễm bệnh/tử vong. Hãy

dùng 4 ký số của mã đề để vẽ 4 tháng tương ứng theo ký số đó. Nếu ký số là 0 thì lấy tháng là 10.

- Chuẩn bị dữ liệu cho toàn bộ phần viii (phần code dùng chung cho cả phần viii)

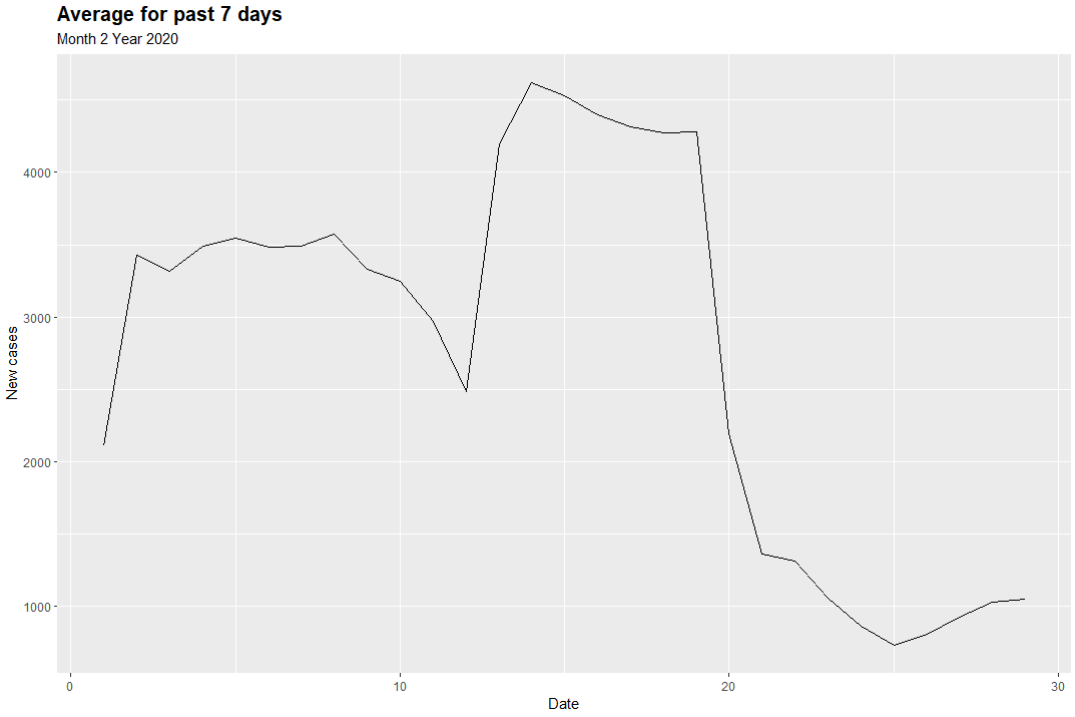
```
library(ggplot2)
library(readr)
library(runner)
library(dplyr)

covid <- read_csv("owid-covid-data.csv")
covid$new_cases <- abs(covid$new_cases)
covid$new_deaths <- abs(covid$new_deaths)
covid$date <- as.Date(covid$date, "%m/%d/%Y")
covid$month <- as.numeric(format(covid$date,'%m'))
covid$year <- as.numeric(format(covid$date,'%Y'))
covid <- subset(covid, !is.na(continent))
covid$new_cases[is.na(covid$new_cases)] <- 0
covid$new_deaths[is.na(covid$new_deaths)] <- 0
```

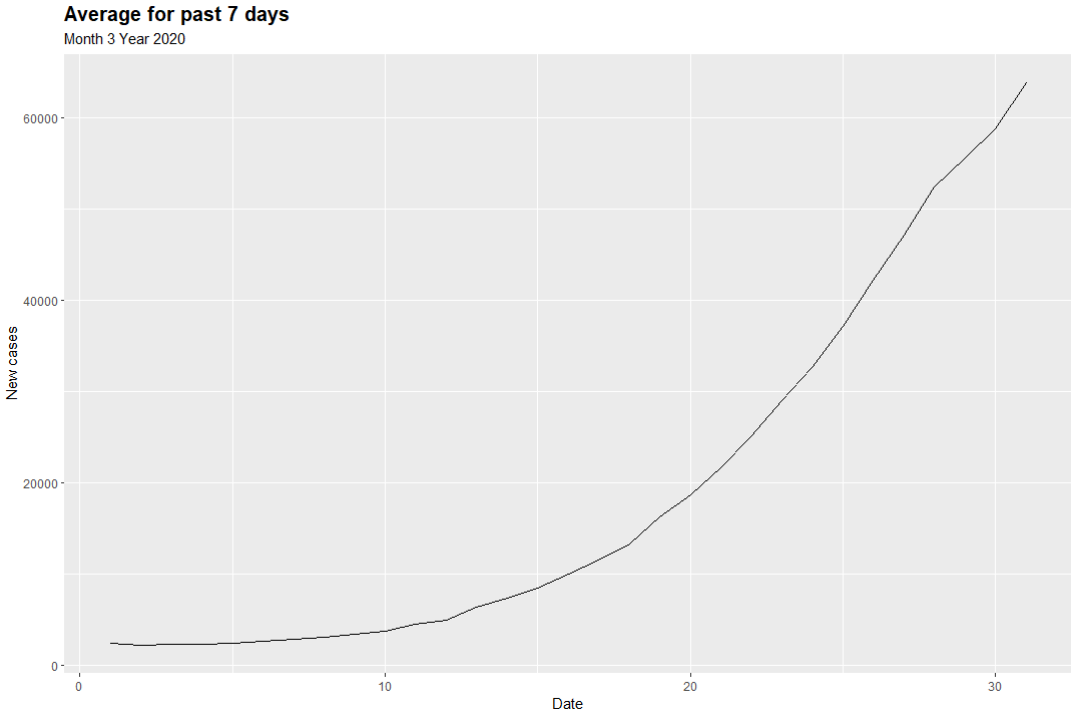
1) Biểu đồ thể hiện thu thập dữ liệu nhiễm bệnh theo thời gian là tháng của tất cả quốc gia theo trung bình 7 ngày gần nhất

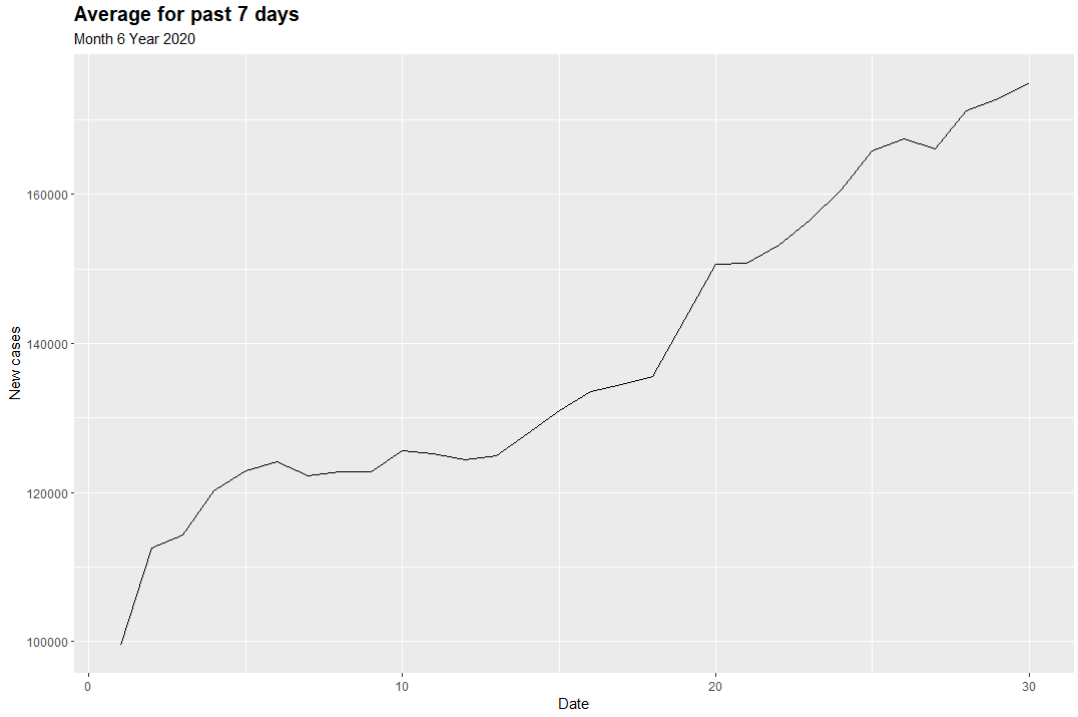
- Hiện thực trong R

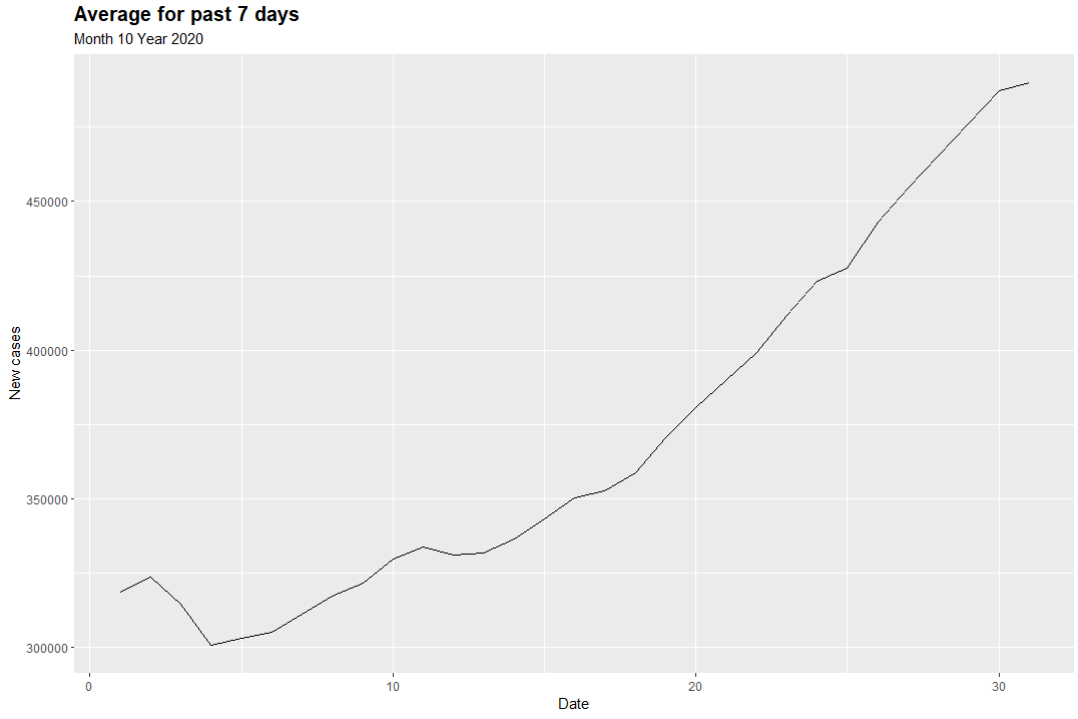
```
figure <- function(m,y,name)
{
  task <- subset(covid, month == m & year == y)
  task <- aggregate(task$new_cases, list(task$date), FUN=sum)
  colnames(task) <- c("date","new_cases")
  task$rec <- 1
  task$rec <- sum_run(x = task$rec,k = 7,i = as.Date(task$date, format = "%m/%d/%Y"))
  task$sum <- sum_run(x = task$new_cases,k = 7,i = as.Date(task$date, format = "%m/%d/%Y"))
  task$sum <- task$sum/task$rec
  task$day <- as.numeric(format(task$date,'%d'))

  graph <- ggplot(task, aes(x=day, y=sum)) + geom_line() + theme(plot.title = element_text(
     size = 15, face = "bold")) + labs(title = "Average for past 7 days", subtitle=name, x= "
     Date", y= "New cases")
  graph
}

figure(2,2020,"Month 2 Year 2020")
figure(3,2020,"Month 3 Year 2020")
figure(6,2020,"Month 6 Year 2020")
figure(10,2020,"Month 10 Year 2020")
figure(2,2021,"Month 2 Year 2021")
figure(3,2021,"Month 3 Year 2021")
figure(6,2021,"Month 6 Year 2021")
figure(10,2021,"Month 10 Year 2021")
figure(2,2022,"Month 2 Year 2021")
```

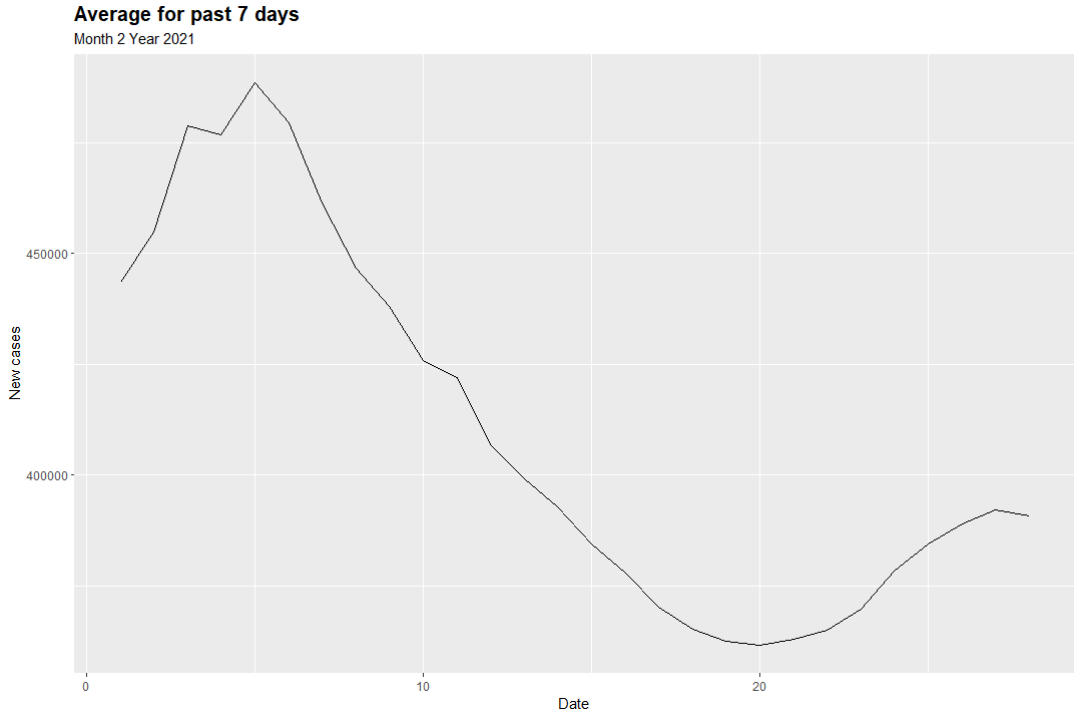
- Kết quả:

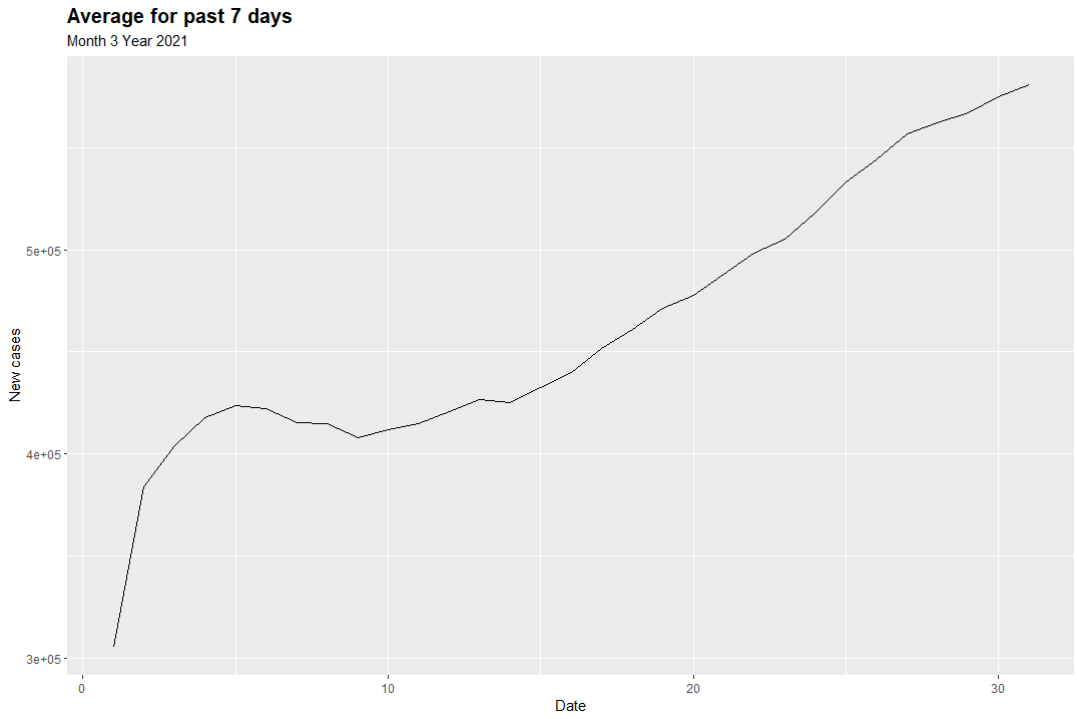

Average for past 7 days
Month 2 Year 2020
New cases
4000
3000
2000
1000
0
10
20
30
Date

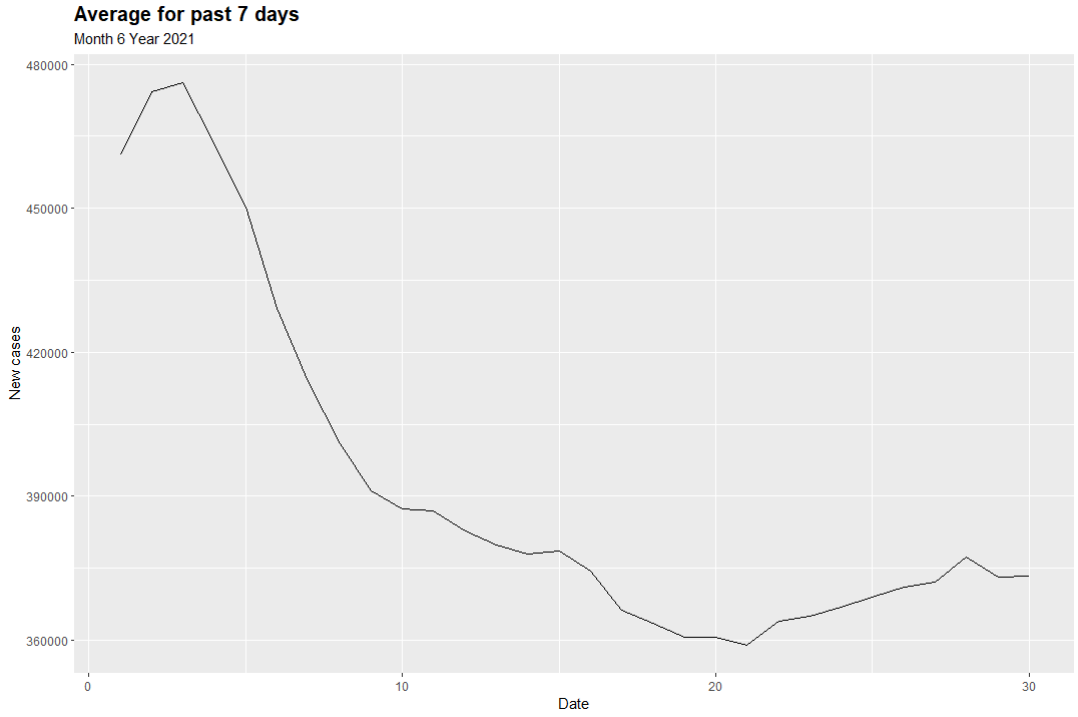

Average for past 7 days
Month 3 Year 2020
New cases
60000
40000
20000
0
0
10
20
30
Date

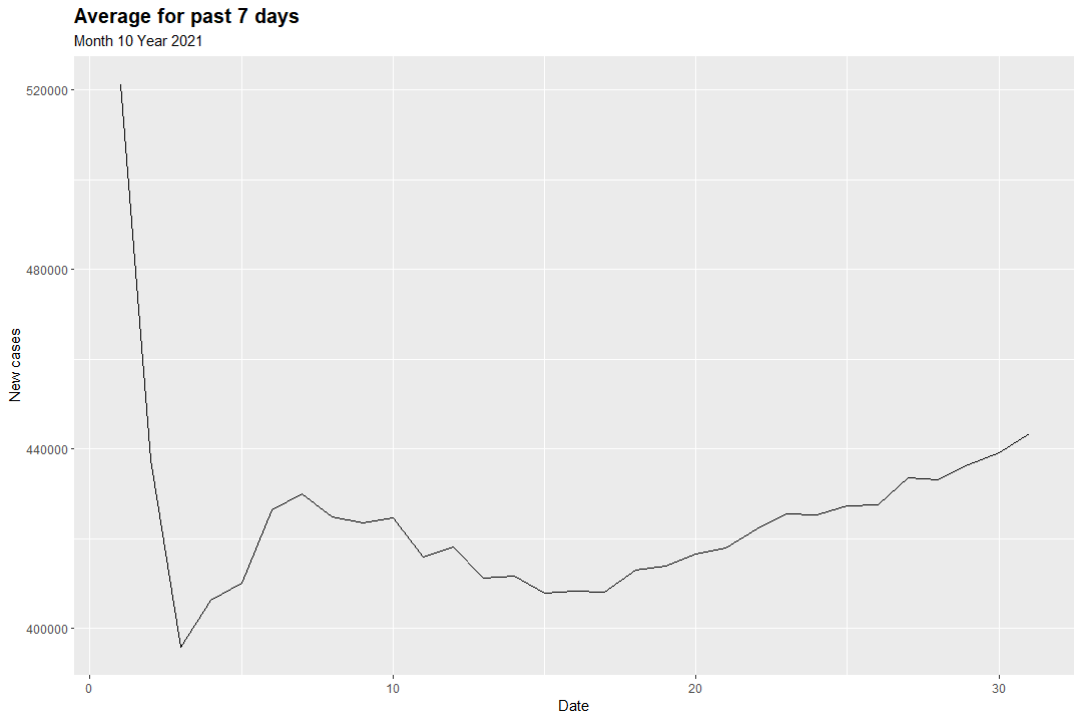

Average for past 7 days
Month 6 Year 2020
New cases
160000
140000
120000
100000
0
10
20
30
Date

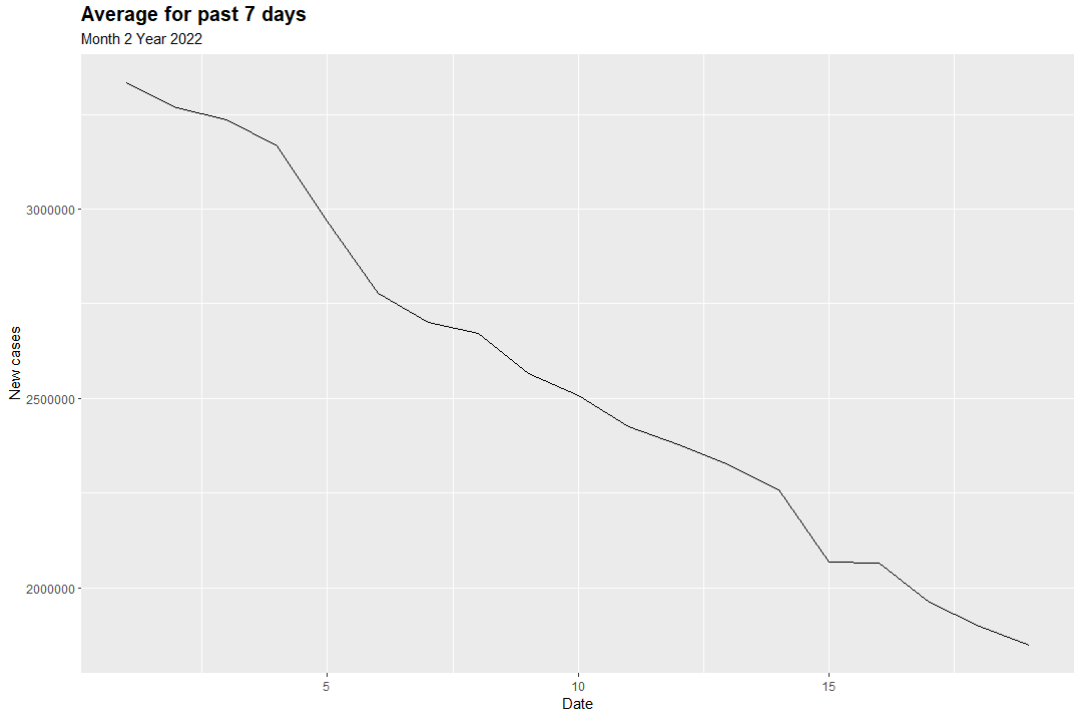

Average for past 7 days
Month 10 Year 2020
New cases
450000
400000
350000
300000
0
10
20
30
Date


Average for past 7 days
Month 2 Year 2021
New cases
450000
400000
0
10
20
Date


Average for past 7 days
Month 3 Year 2021
New cases
5e+05
4e+05
3e+05
0
10
20
30
Date


Average for past 7 days
Month 6 Year 2021
480000
450000
420000
390000
360000
New cases
0
10
20
30
Date


Average for past 7 days
Month 10 Year 2021
520000
480000
440000
400000
New cases
0
10
20
30
Date

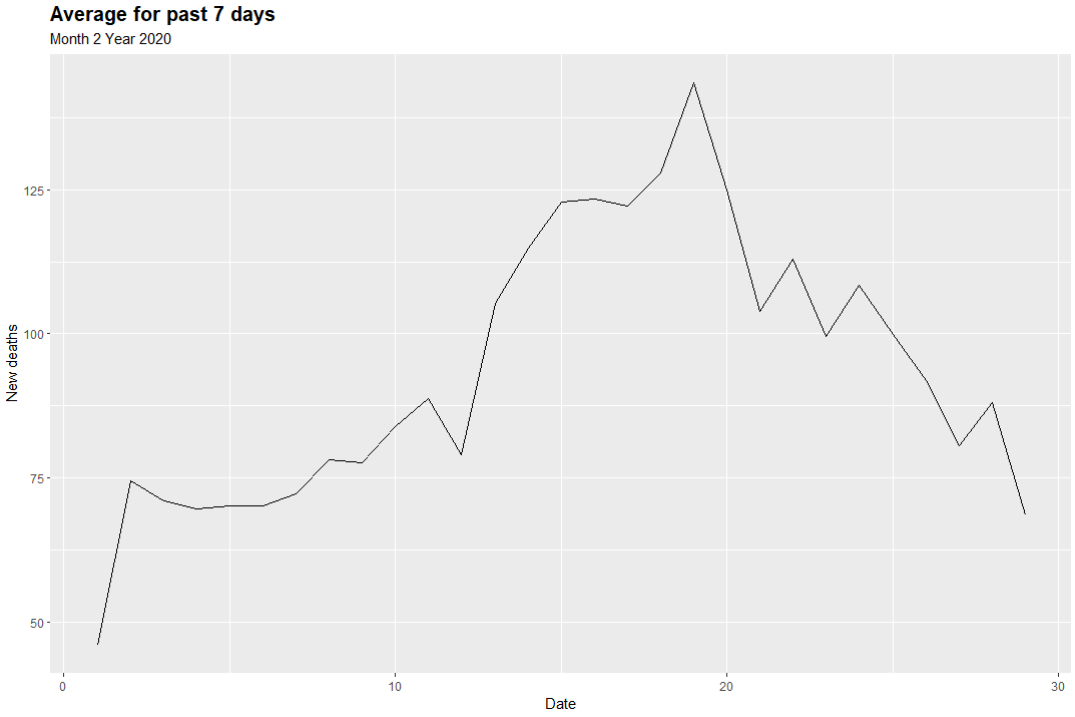


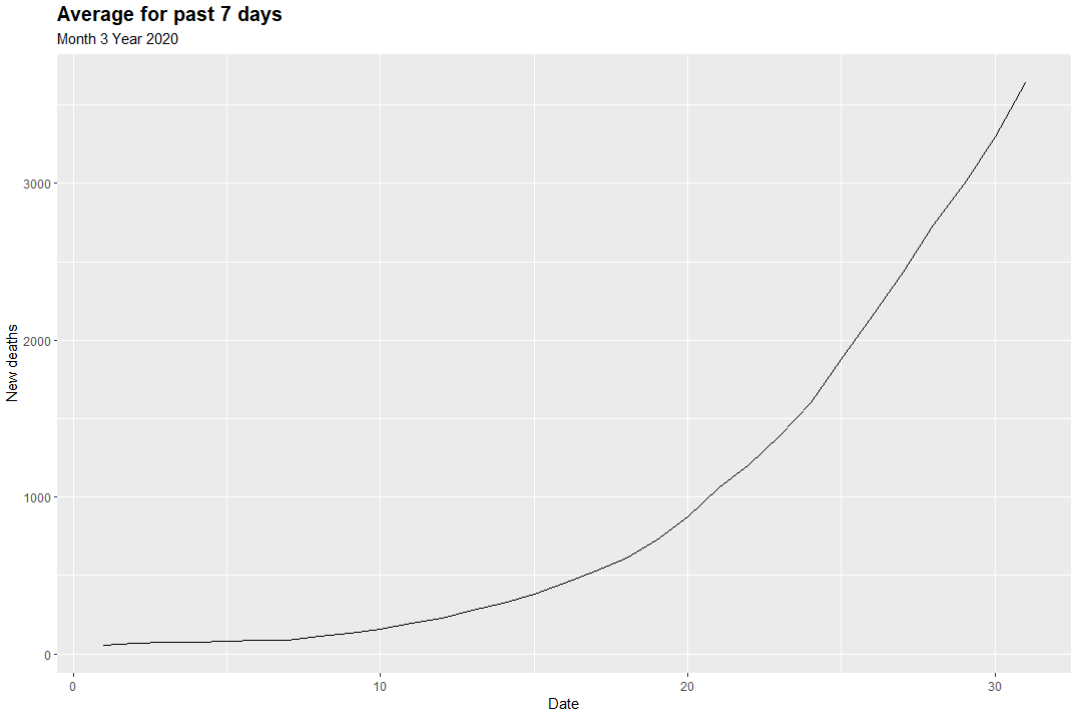
2) Biểu đồ thể hiện thu thập dữ liệu tử vong theo thời gian là tháng của tất cả quốc gia theo trung bình 7 ngày gần nhất

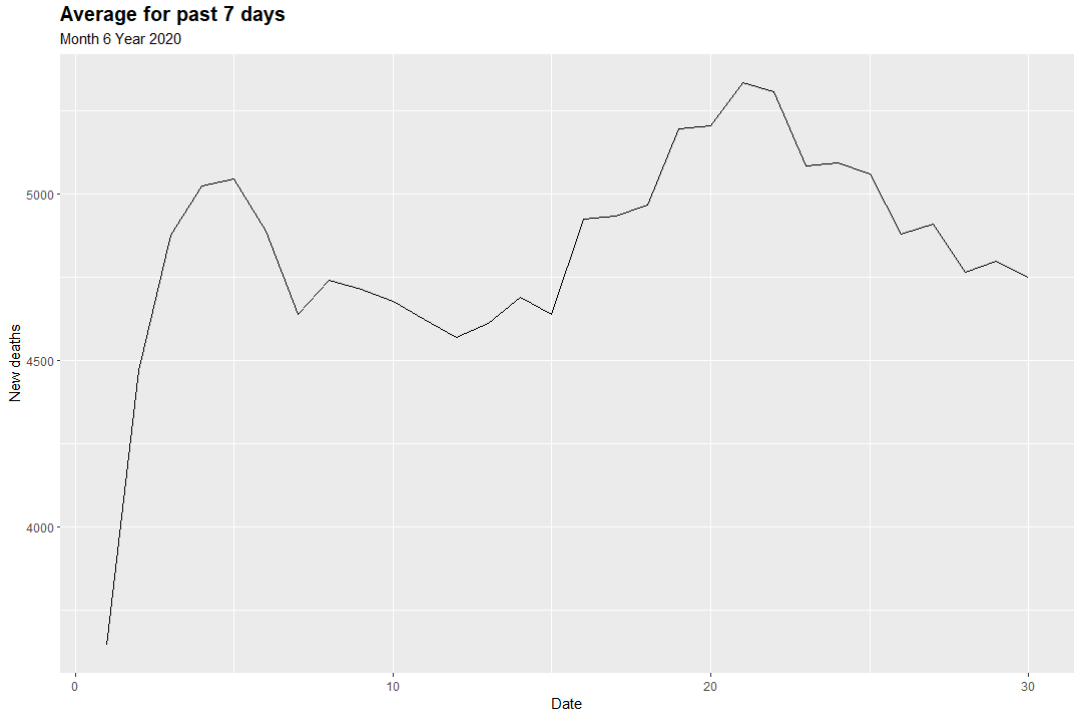
- Hiện thực trong R

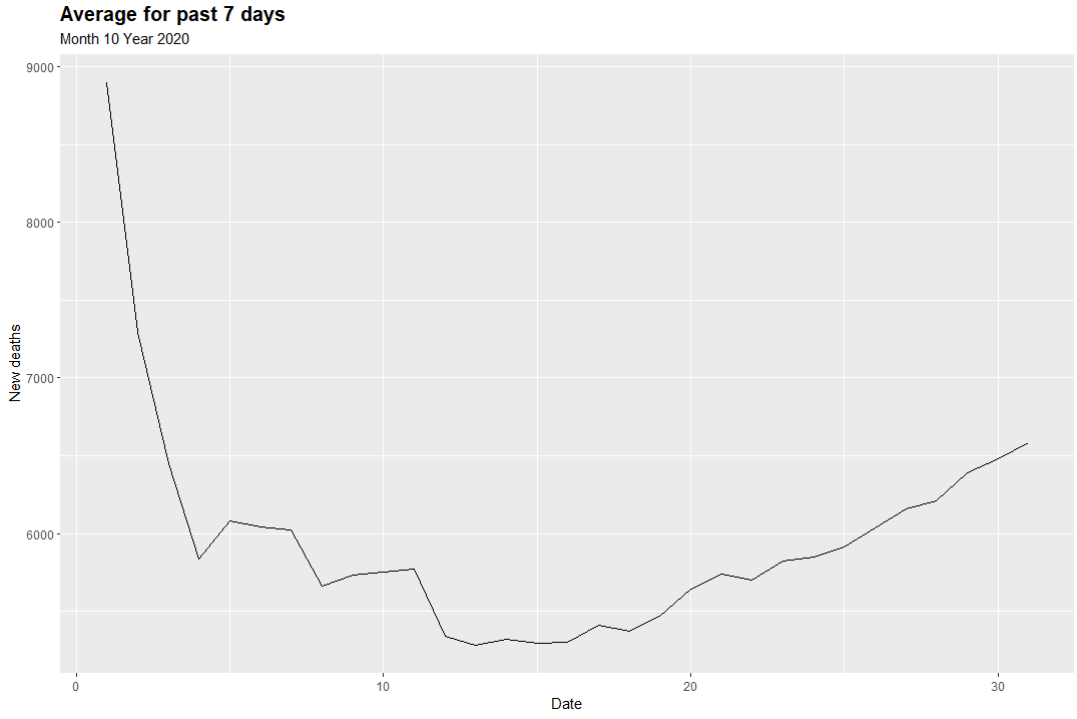
```
figure <- function(m,y,name)
{
  task <- subset(covid, month == m & year == y)
  task <- aggregate(task$new_deaths, list(task$date), FUN=sum)
  colnames(task) <- c("date","new_deaths")
  task$rec <- 1
  task$rec <- sum_run(x = task$rec,k = 7,i = as.Date(task$date, format = "%m/%d/%Y"))
  task$sum <- sum_run(x = task$new_deaths,k = 7,i = as.Date(task$date, format = "%m/%d/%Y"))
  task$sum <- task$sum/task$rec
  task$day <- as.numeric(format(task$date,'%d'))

  graph <- ggplot(task, aes(x=day, y=sum)) + geom_line() + theme(plot.title = element_text(
     size = 15, face = "bold")) + labs(title="Average for past 7 days", subtitle=name, x= "Date
     ", y= "New deaths")
  graph
}

figure(2,2020,"Month 2 Year 2020")
figure(3,2020,"Month 3 Year 2020")
figure(6,2020,"Month 6 Year 2020")
figure(10,2020,"Month 10 Year 2020")
figure(2,2021,"Month 2 Year 2021")
figure(3,2021,"Month 3 Year 2021")
figure(6,2021,"Month 6 Year 2021")
figure(10,2021,"Month 10 Year 2021")
figure(2,2022,"Month 2 Year 2021")
```

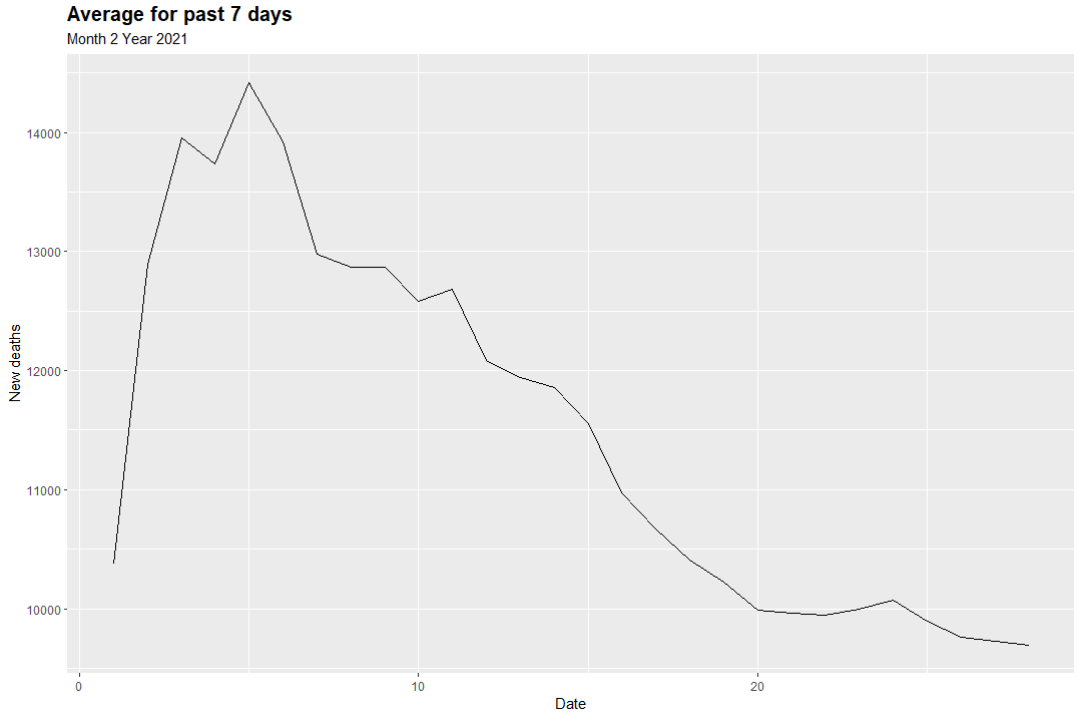
- Kết quả:

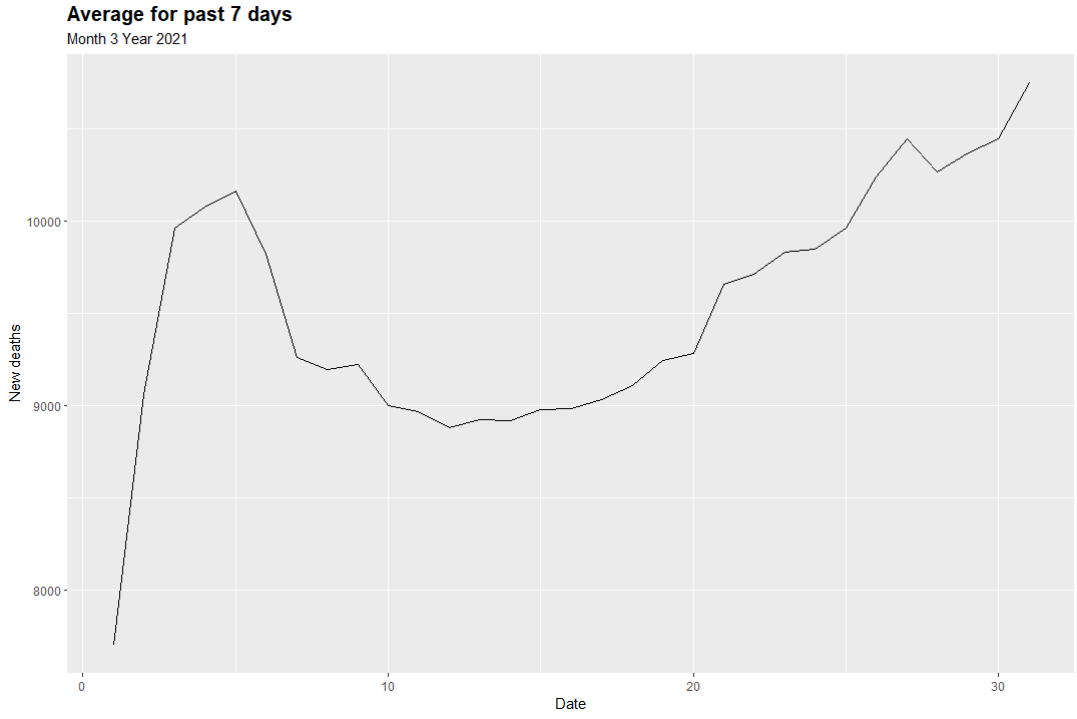

Average for past 7 days
Month 2 Year 2020
New deaths
125
100
75
50
0
10
20
30
Date

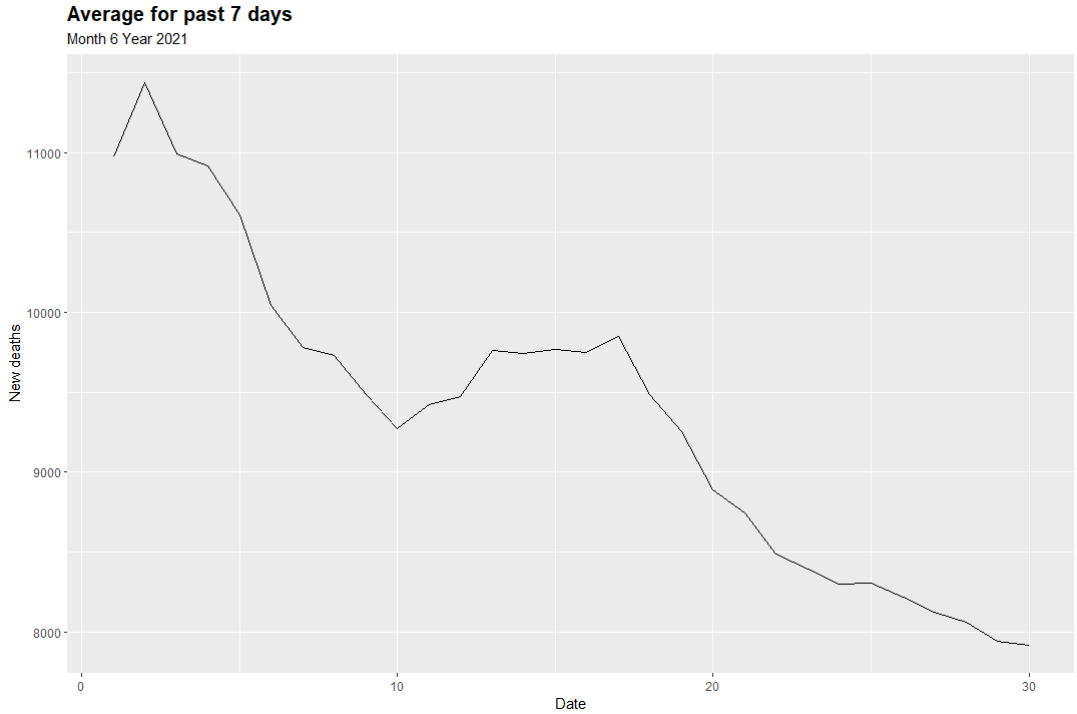

Average for past 7 days
Month 3 Year 2020
New deaths
3000
2000
1000
0
0
10
20
30
Date

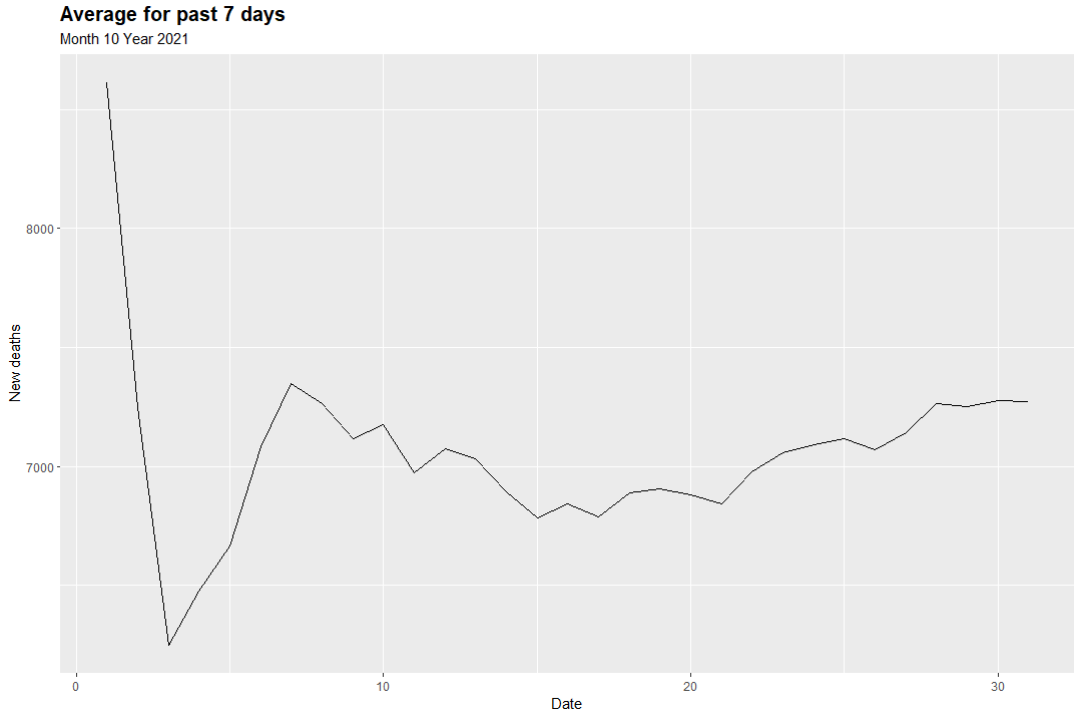

Average for past 7 days
Month 6 Year 2020
New deaths
5000
4500
4000
0
10
20
30
Date


Average for past 7 days
Month 10 Year 2020
New deaths
9000
8000
7000
6000
0
10
20
30
Date


Average for past 7 days
Month 2 Year 2021
New deaths
14000
13000
12000
11000
10000
0
10
20
Date


Average for past 7 days
Month 3 Year 2021
New deaths
10000
9000
8000
0
10
20
30
Date





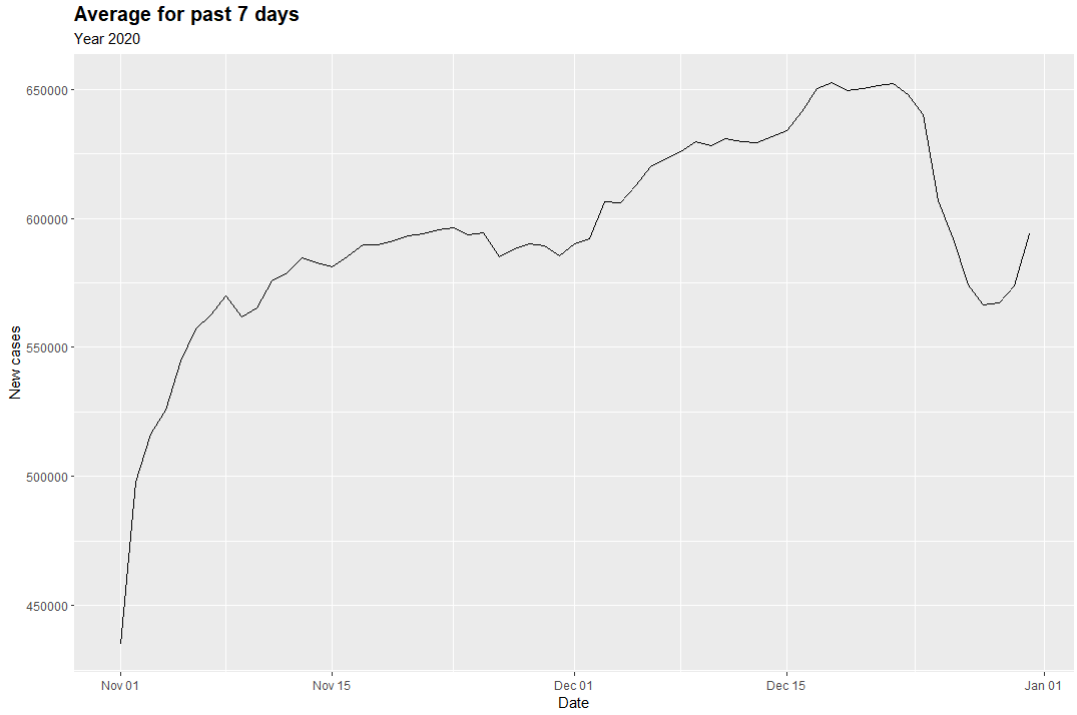
3) Biểu đồ thể hiện thu thập dữ liệu nhiễm bệnh theo thời gian là 2 tháng của năm của tất cả quốc gia theo trung bình 7 ngày gần nhất

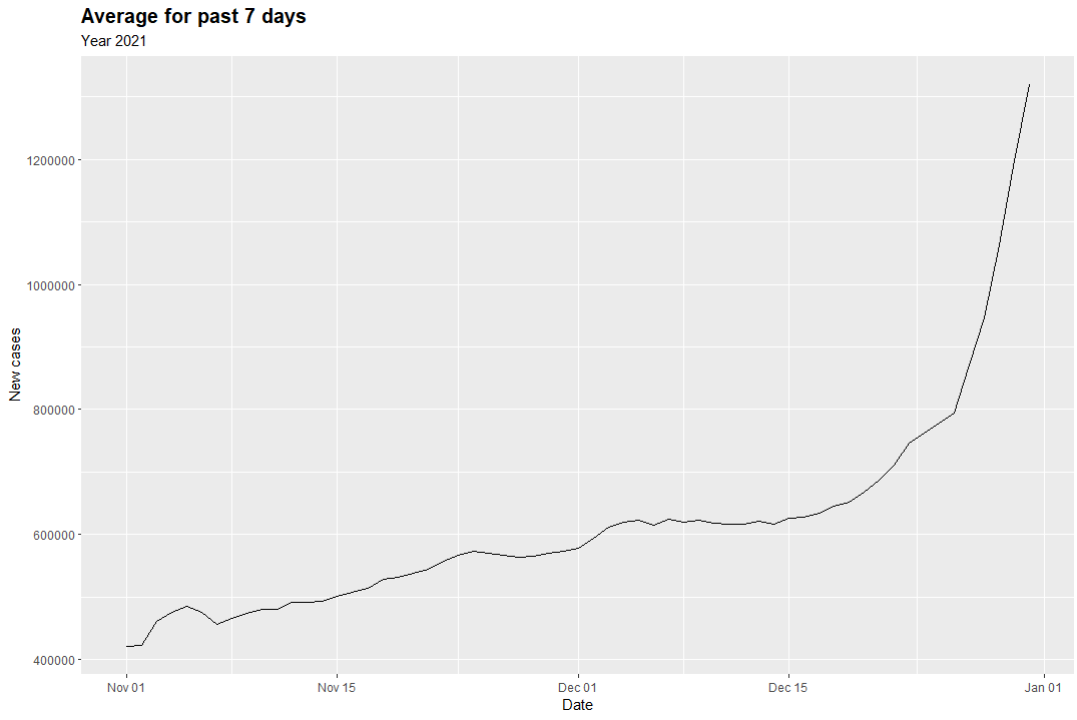
- Hiện thực trong R

```
figure <- function(y,name)
{
  task <- subset(covid, (month == 11|month == 12) & year == y)
  task <- aggregate(task$new_cases, list(task$date), FUN=sum)
  colnames(task) <- c("date","new_cases")

  task$rec <- 1
  task$rec <- sum_run(x = task$rec,k = 7,i = as.Date(task$date, format = "%m/%d/%Y"))
  task$sum <- sum_run(x = task$new_cases,k = 7,i = as.Date(task$date, format = "%m/%d/%Y"))
```

```
  task$sum <- task$sum/task$rec

  graph <- ggplot(task, aes(x=date, y=sum)) + geom_line() + theme(plot.title = element_text(
     size = 15, face = "bold")) + labs(title="Average for past 7 days", subtitle=name, x= "Date
     ", y= "New cases")
  graph
}

figure(2020,"Year 2020")
figure(2021,"Year 2021")
```

- Kết quả:



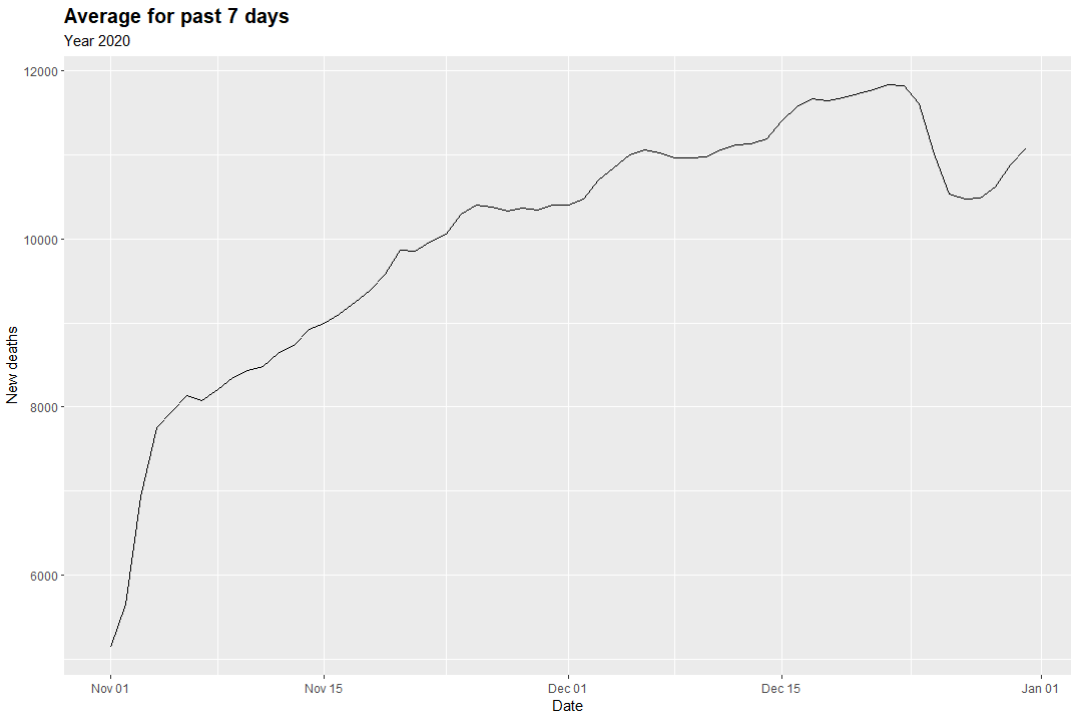


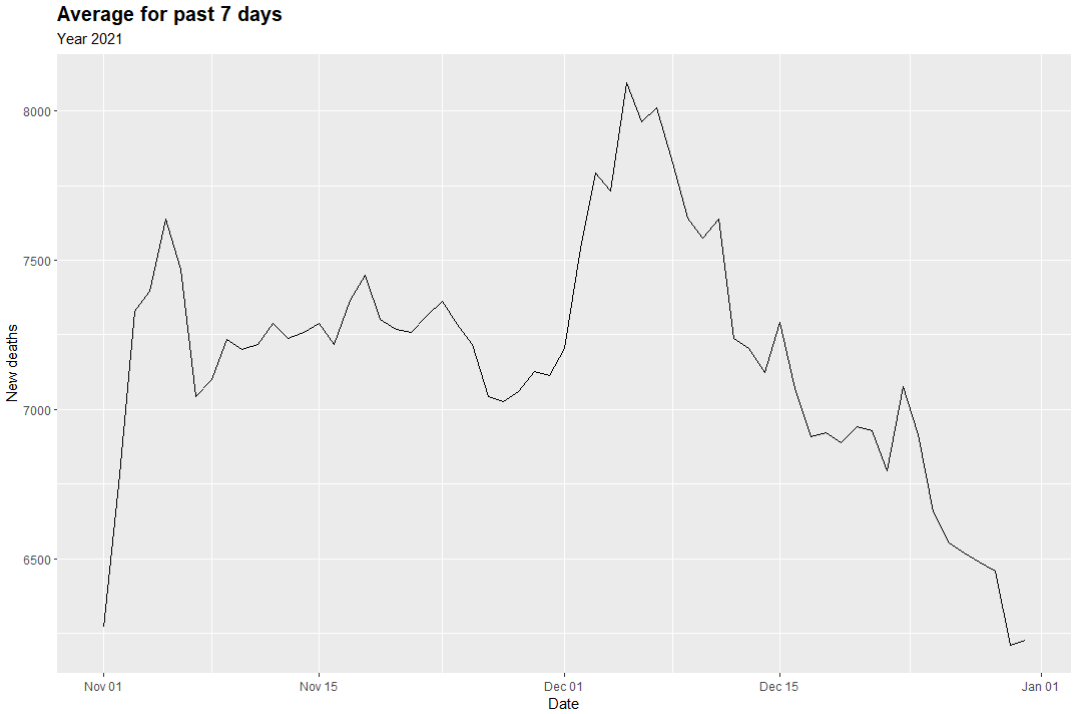
4) Biểu đồ thể hiện thu thập dữ liệu tử vong theo thời gian là 2 tháng của năm của tất cả quốc gia theo trung bình 7 ngày gần nhất

- Hiện thực trong R

```
figure <- function(y,name)
{
  task <- subset(covid, (month == 11|month == 12) & year == y)
  task <- aggregate(task$new_deaths, list(task$date), FUN=sum)
  colnames(task) <- c("date","new_deaths")
  task$rec <- 1
  task$rec <- sum_run(x = task$rec,k = 7,i = as.Date(task$date, format = "%m/%d/%Y"))
  task$sum <- sum_run(x = task$new_deaths,k = 7,i = as.Date(task$date, format = "%m/%d/%Y"))
  task$sum <- task$sum/task$rec

  graph <- ggplot(task, aes(x=date, y=sum)) + geom_line() + theme(plot.title = element_text(
     size = 15, face = "bold")) + labs(title="Average for past 7 days", subtitle=name, x= "Date
     ", y= "New deaths")
  graph
}

figure(2020,"Year 2020")
figure(2021,"Year 2021")
```

- Kết quả:



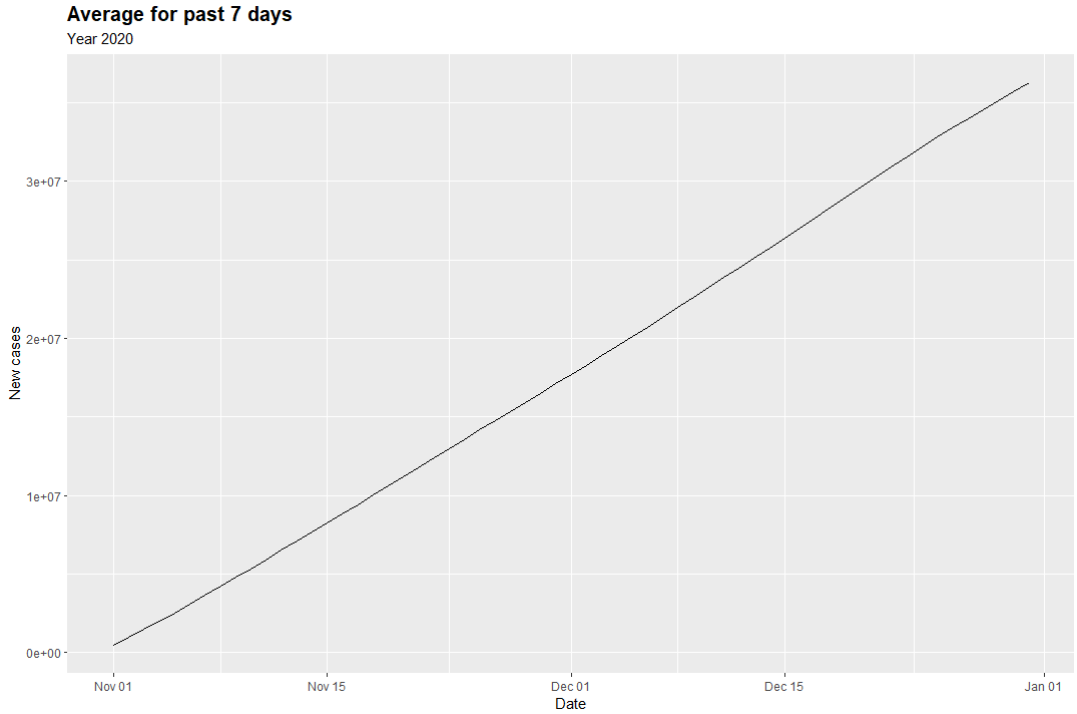


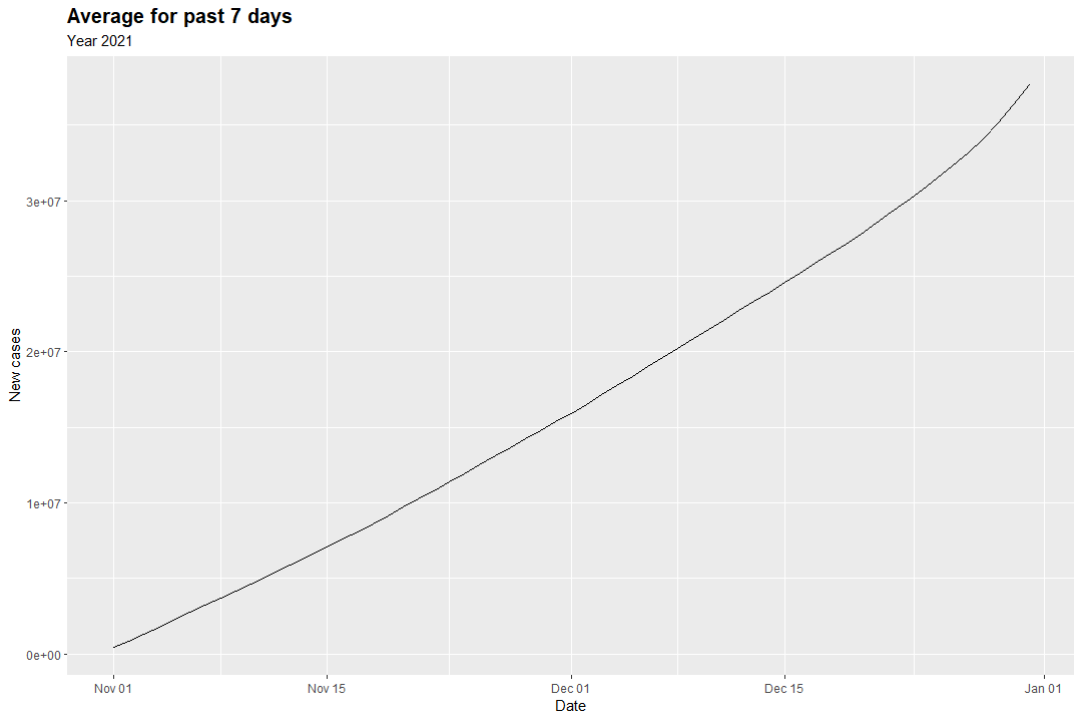
5) Biểu đồ thể hiện thu thập dữ liệu nhiễm bệnh tích lũy theo thời gian là 2 tháng của năm của tất cả quốc giai theo trung bình 7 ngày gần nhất

- Hiện thực trong R

```
figure <- function(y,name)
{
  task <- subset(covid, (month == 11|month == 12) & year == y)
  task <- aggregate(task$new_cases, list(task$date), FUN=sum)
  colnames(task) <- c("date","new_cases")

  task$rec <- 1
  task$rec <- sum_run(x = task$rec,k = 7,i = as.Date(task$date, format = "%m/%d/%Y"))
  task$sum <- sum_run(x = task$new_cases,k = 7,i = as.Date(task$date, format = "%m/%d/%Y"))
  task$ave <- task$sum/task$rec

  task$sum <- cumsum(task$ave)

  graph <- ggplot(task, aes(x=date, y=sum)) + geom_line() + theme(plot.title = element_text(
     size = 15, face = "bold")) + labs(title="Average for past 7 days", subtitle=name, x= "Date
     ", y= "New cases")
  graph
}

figure(2020,"Year 2020")
figure(2021,"Year 2021")
```

- Kết quả:





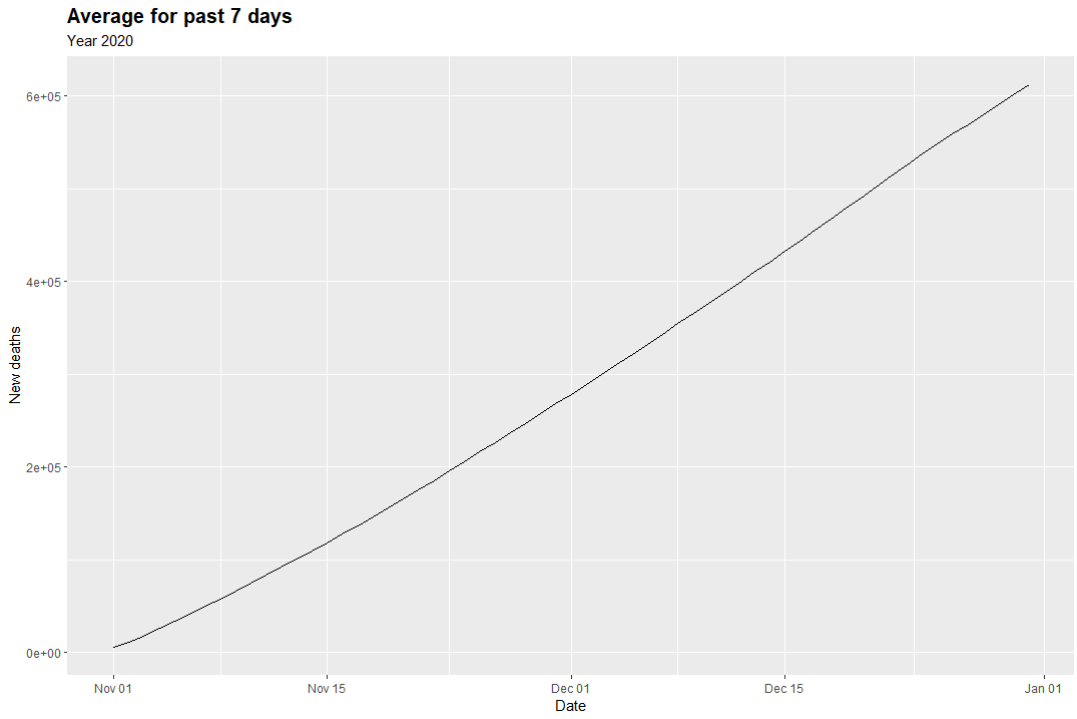
6) Biểu đồ thể hiện thu thập dữ liệu tử vong tích lũy theo thời gian là 2 tháng của năm của tất cả quốc gia theo trung bình 7 ngày gần nhấtt

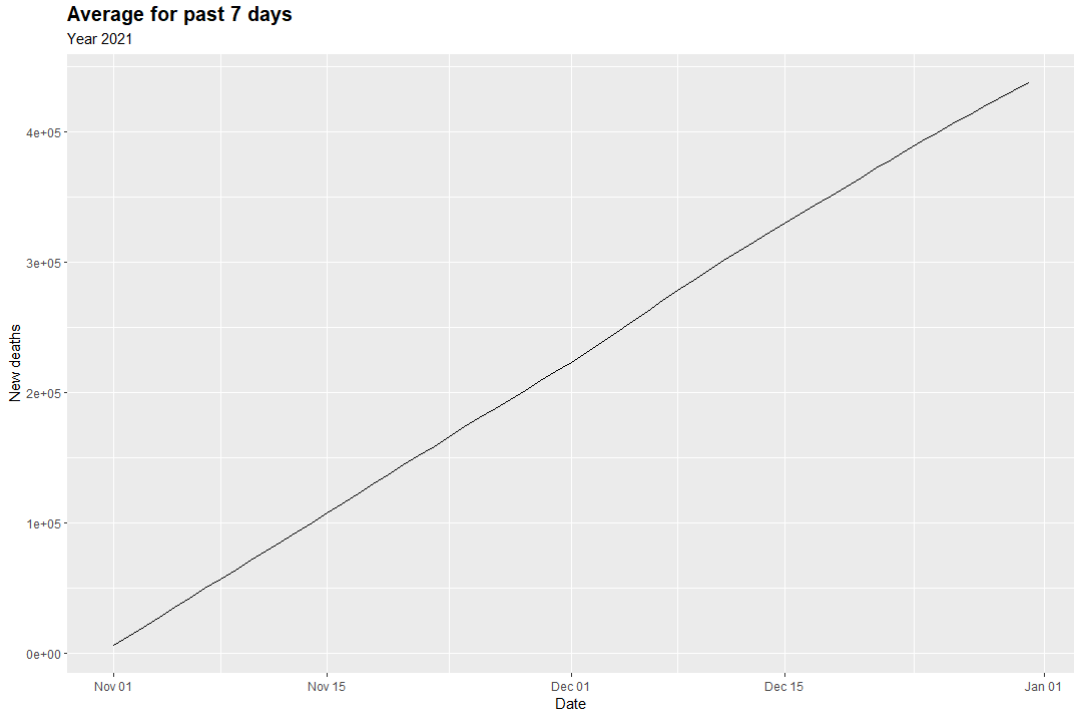
- Hiện thực trong R

```
figure <- function(y,name)
{
  task <- subset(covid, (month == 11|month == 12) & year == y)
  task <- aggregate(task$new_deaths, list(task$date), FUN=sum)
  colnames(task) <- c("date","new_deaths")

  task$rec <- 1
  task$rec <- sum_run(x = task$rec,k = 7,i = as.Date(task$date, format = "%m/%d/%Y"))
  task$sum <- sum_run(x = task$new_deaths,k = 7,i = as.Date(task$date, format = "%m/%d/%Y"))
```

```
  task$ave <- task$sum/task$rec

  task$sum <- cumsum(task$ave)

  graph <- ggplot(task, aes(x=date, y=sum)) + geom_line() + theme(plot.title = element_text(
     size = 15, face = "bold")) + labels(title="Average for past 7 days", subtitle=name, x= "Date
     ", y= "New deaths")
  graph
}

figure(2020,"Year 2020")
figure(2021,"Year 2021")
```

- Kết quả:





ix) Nhóm câu hỏi liên quan đến sự tương quan giữa nhiễm bệnh và tử vong

- Chuẩn bị dữ liệu cho toàn bộ phần ix (phần code dùng chung cho cả phần ix)

```
library(ggplot2)
library(readr)
library(runner)
library(dplyr)
library(zoo)

covid <- read_csv("C:/Users/Asus/Documents/covidData.csv")
covid$new_cases <- abs(covid$new_cases)
covid$new_deaths <- abs(covid$new_deaths)
covid$date <- as.Date(covid$date, "%m/%d/%Y")
covid$month <- as.numeric(format(covid$date,'%m'))
covid$year <- as.numeric(format(covid$date,'%Y'))
covid <- subset(covid, !is.na(continent))
covid$new_cases[is.na(covid$new_cases)] <- 0
covid$new_deaths[is.na(covid$new_deaths)] <- 0
year <- unique(format(covid$date, format= "%Y"))
lcountry <- c("Brazil","Chile","Venezuela")
lmonth <- c("02","03","06","10")
```

1) Vẽ biểu đồ thể hiện phần trăm giữa nhiễm bệnh tích lũy trên tổng nhiễm bệnh và phần trăm tử vong tích lũy trên tổng số tử vong cho từng quốc gia theo thời gian. Vẽ 2 đường trên cùng biểu đồ

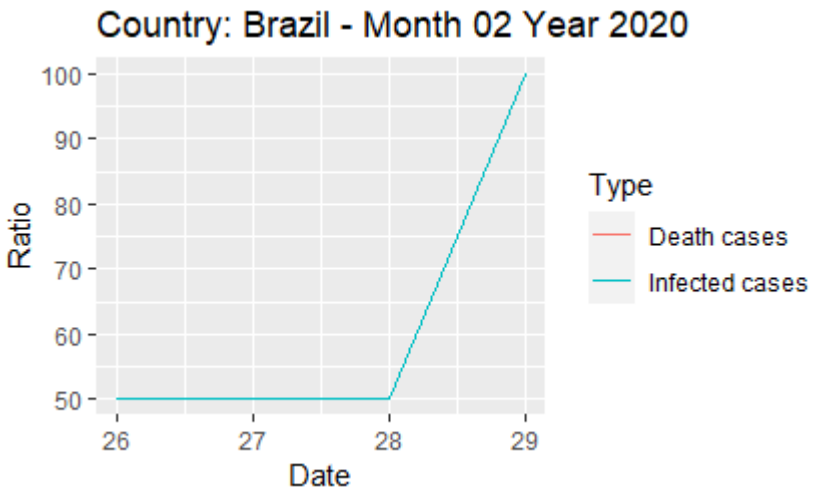
   Trên từng quốc gia hãy vẽ biểu đồ thể hiện trục Ox là nhiễm bệnh, trục Oy là tử vong. Hãy lấy 4 tháng theo 4 ký số mã đề thể hiện. Nếu ký số là 0 thì lấy tháng là 10.

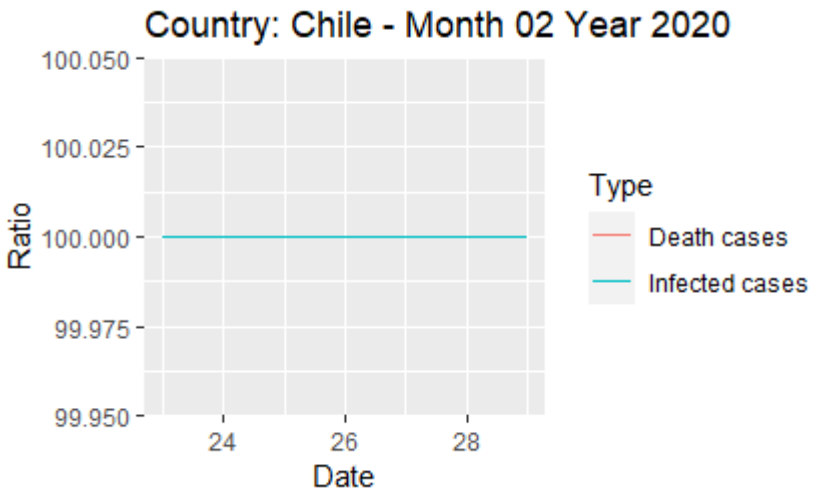
   - Hiện thực trong R:

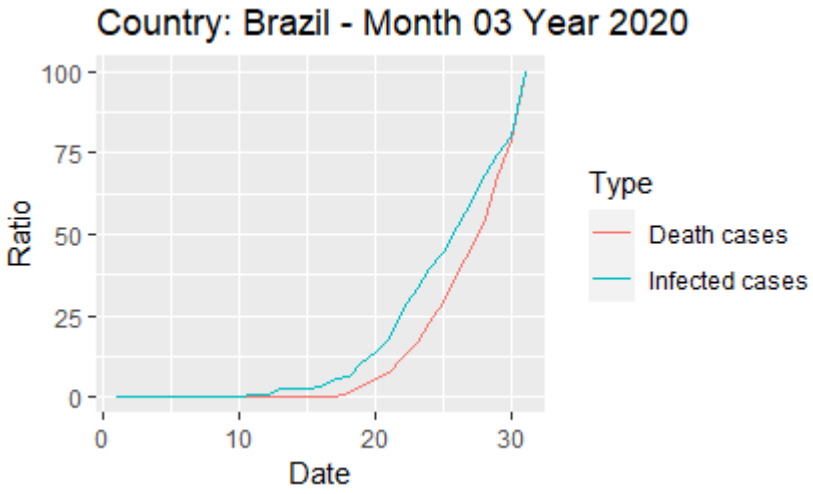
```
for (i in year){
  for (k in lmonth){
    for (c in lcountry){
      l1 <- NULL
      l2 <- NULL
      tab <- subset(covid, format(covid$date, format= "%Y") == i & format(covid$date, format= "%m") == k & covid$location == c)
      if (nrow(tab) != 0){
        l1 <- tab
```

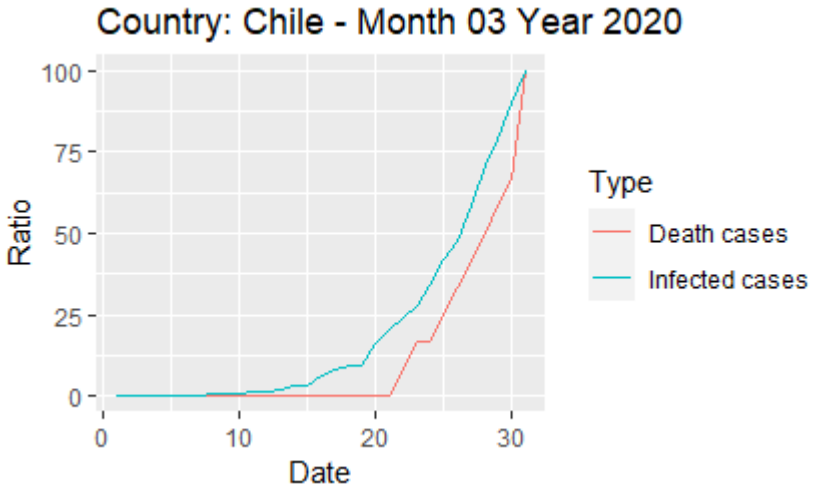
```
        l1$day <- format(l1$date,"%d")
        l1$date <- "Infected cases"
        l1$new <- l1$new_cases

        l2 <- tab
        l2$day <- format(l2$date,"%d")
        l2$date <- "Death cases"

        l2$new <- l2$new_deaths
        keeps <- c("day","date","new")
        l1 <- subset(l1, select = keeps)
        l2 <- subset(l2, select = keeps)
        df <- rbind(l1,l2)
        print(ggplot(df,aes(x=as.integer(day), y=new, fill = date)) +
                geom_point(aes(col=date, size = new)) +
                geom_smooth(formula = y ~ x, method = "lm") +
                labs(title = paste("Country:",c,"- Month",k,"Year",i), x ="Date", y = "Number
    of cases",fill="Type",size = "Number of cases) +
                guides(col = FALSE))
        cat(paste("Country:",c,"- Month",k,"Year",i))
        cat("\n")
        cat("Correlation coefficients of infected cases: ")
        cat(cor(as.integer(df$day[df$date=="Infected cases"]),df$new[df$date=="Infected cases"
    ]))
        cat("\n")
        cat("Correlation coefficients of death cases: ")
        cat(cor(as.integer(df$day[df$date=="Death cases"]),df$new[df$date=="Death cases"]))
        cat("\n")
      }
    }
  }
}
```

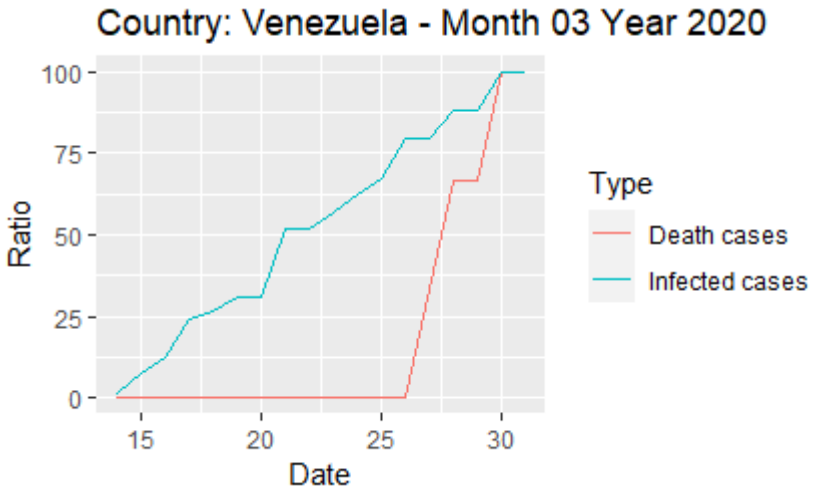
- Kết quả

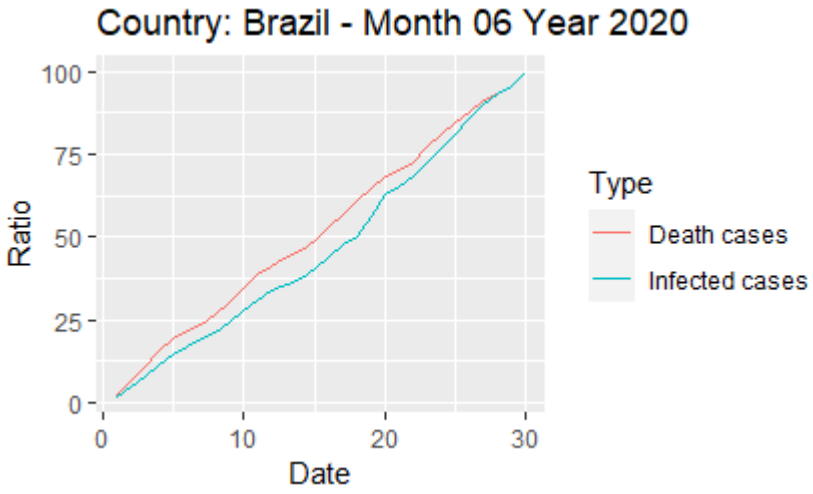


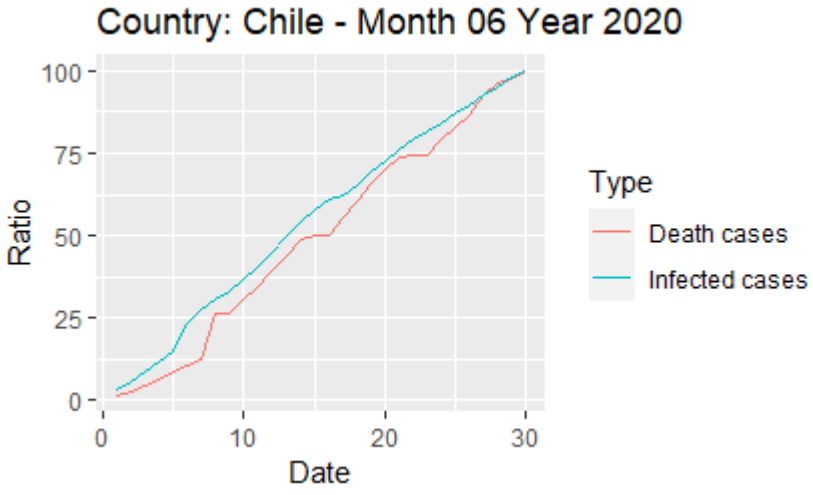

Country: Chile - Month 02 Year 2020
100.050
100.025
100.000
99.975
99.950
Ratio
24
26
28
Date
Type
Death cases
Infected cases

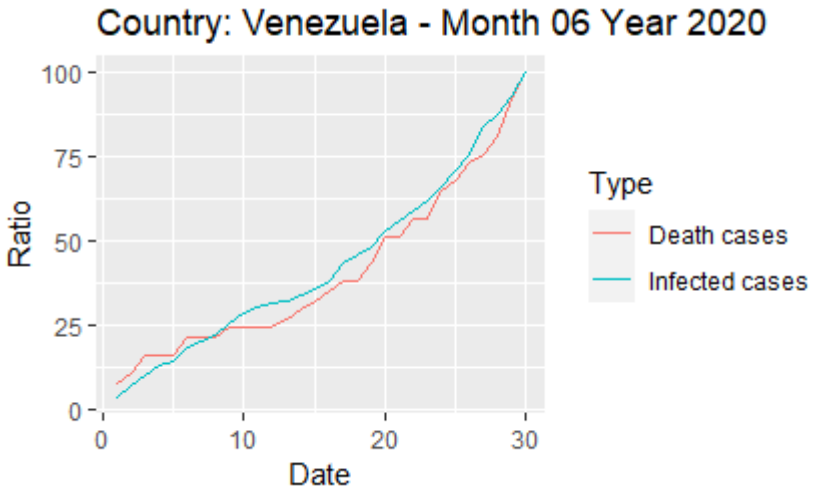

Country: Brazil - Month 03 Year 2020
100
75
50
25
0
Ratio
0
10
20
30
Date
Type
Death cases
Infected cases

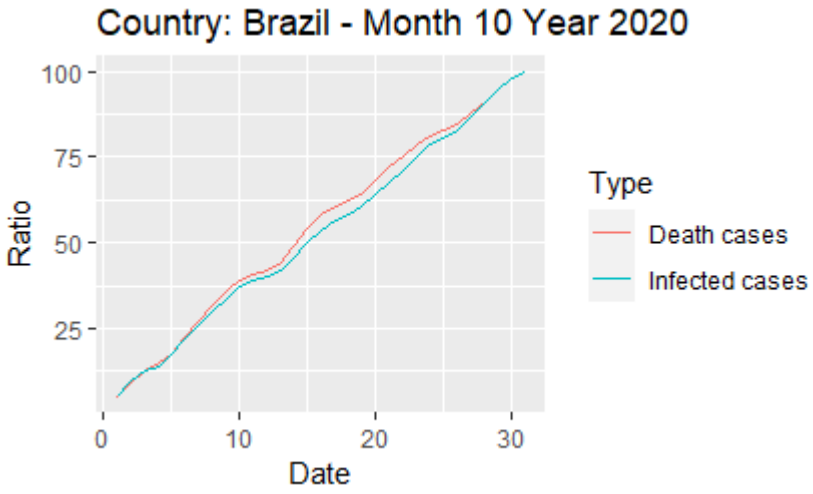

Country: Chile - Month 03 Year 2020
100
75
50
25
0
Ratio
0
10
20
30
Date
Type
Death cases
Infected cases

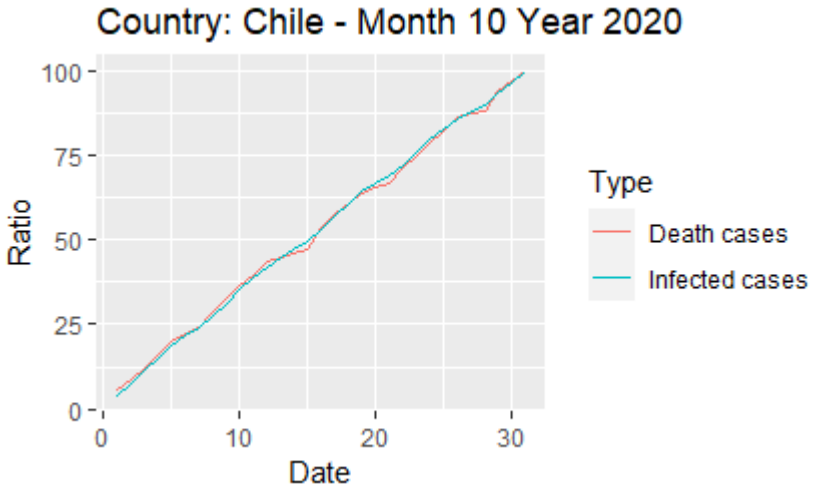

Country: Venezuela - Month 03 Year 2020
Ratio
Date
Type
Death cases
Infected cases

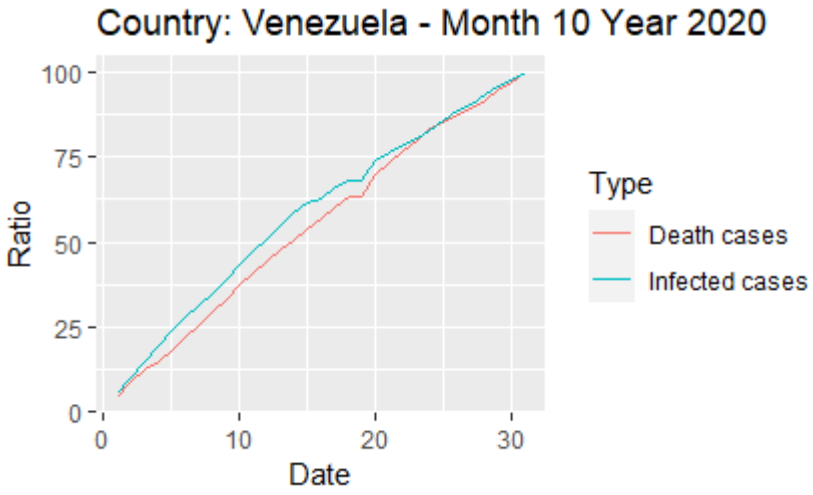

Country: Brazil - Month 06 Year 2020
Ratio
Date
Type
Death cases
Infected cases

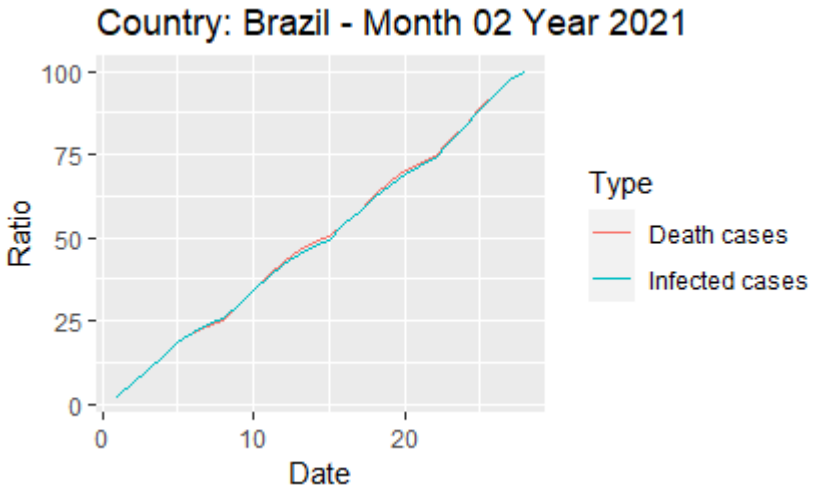

Country: Chile - Month 06 Year 2020
Ratio
Date
Type
Death cases
Infected cases

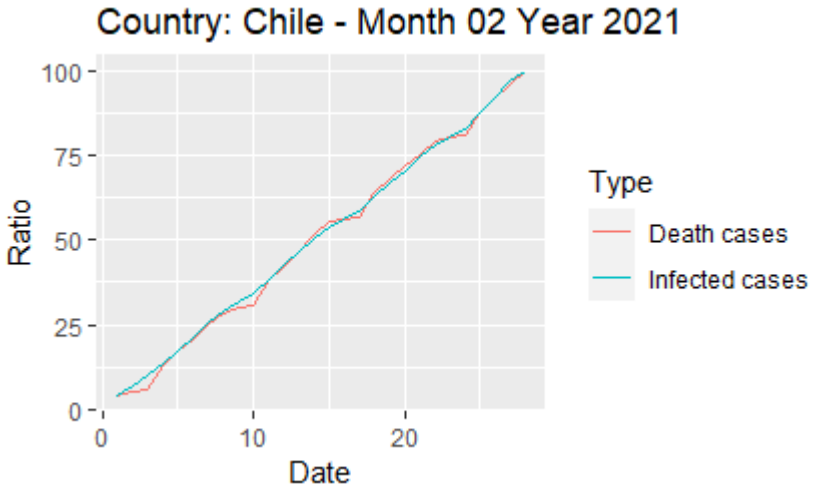

Country: Venezuela - Month 06 Year 2020
Ratio
Date
Type
Death cases
Infected cases

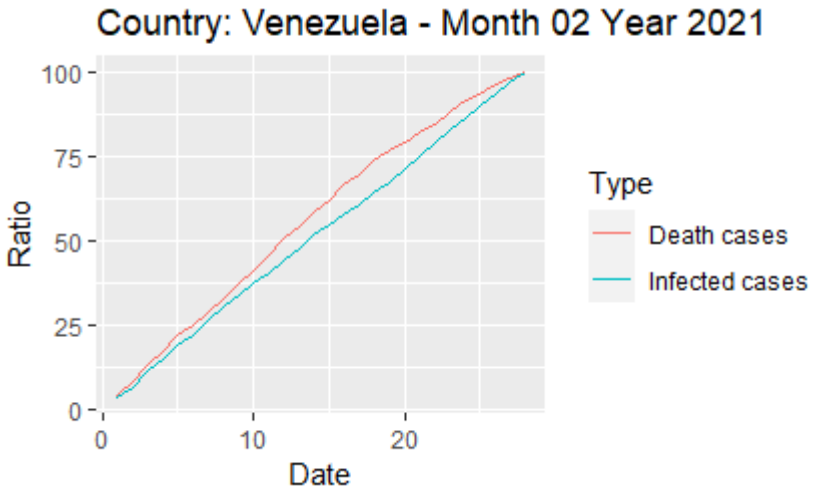

Country: Brazil - Month 10 Year 2020
Ratio
Date
Type
Death cases
Infected cases

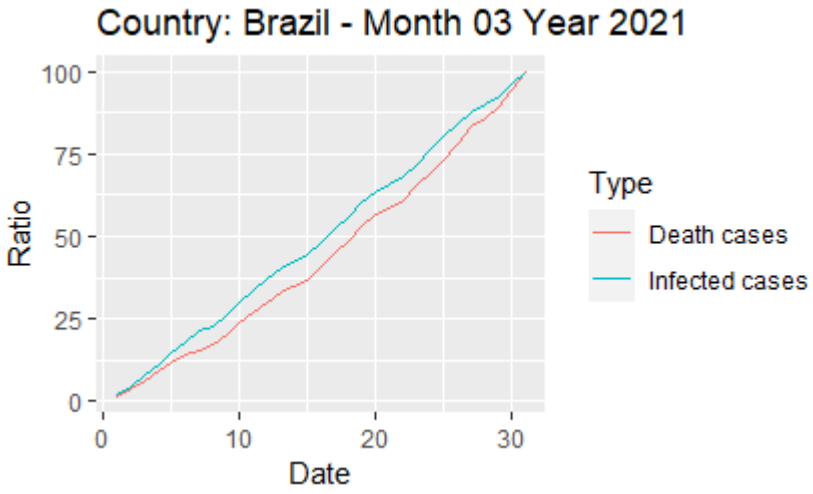

Country: Chile - Month 10 Year 2020
Ratio
Date
Type
Death cases
Infected cases

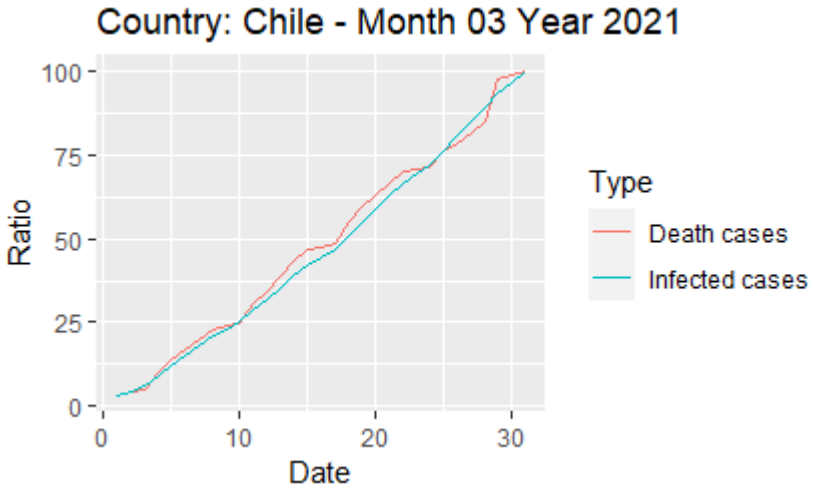

Country: Venezuela - Month 10 Year 2020
Ratio
Date
Type
Death cases
Infected cases

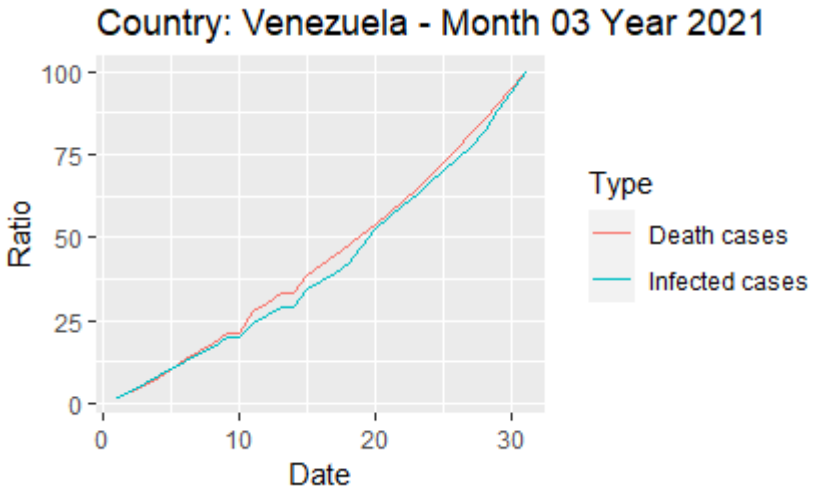

Country: Brazil - Month 02 Year 2021
Ratio
Date
Type
Death cases
Infected cases

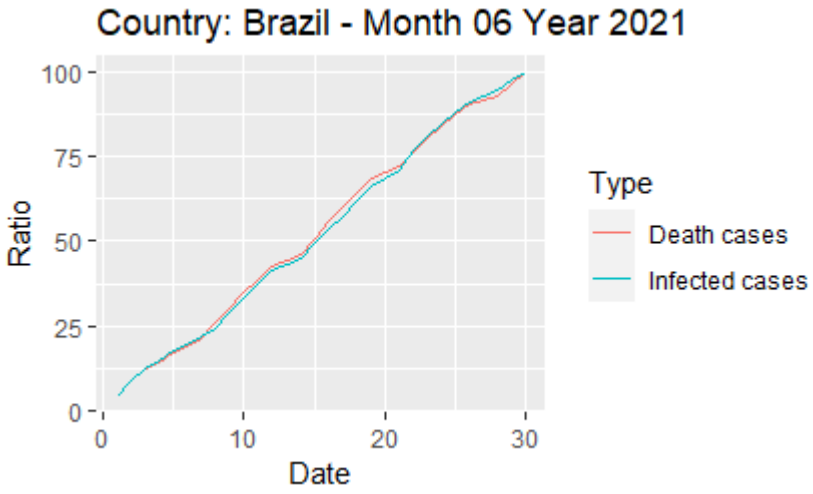

Country: Chile - Month 02 Year 2021
Ratio
Date
Type
Death cases
Infected cases

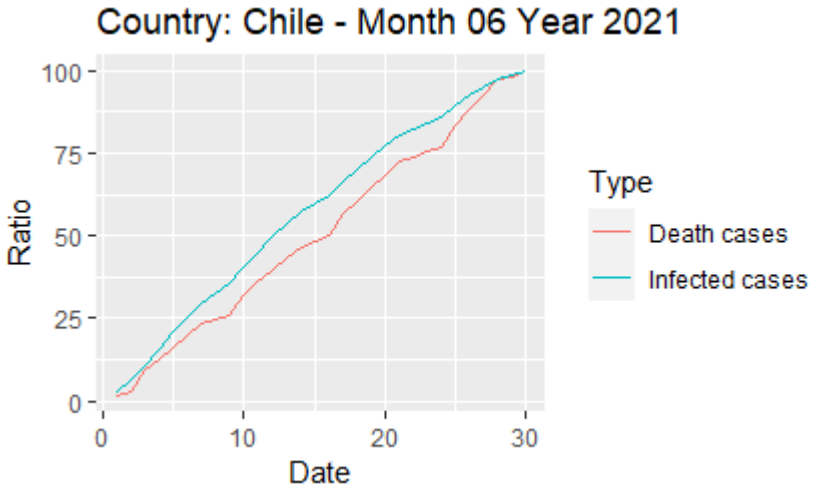

Country: Venezuela - Month 02 Year 2021
Ratio
Date
Type
Death cases
Infected cases

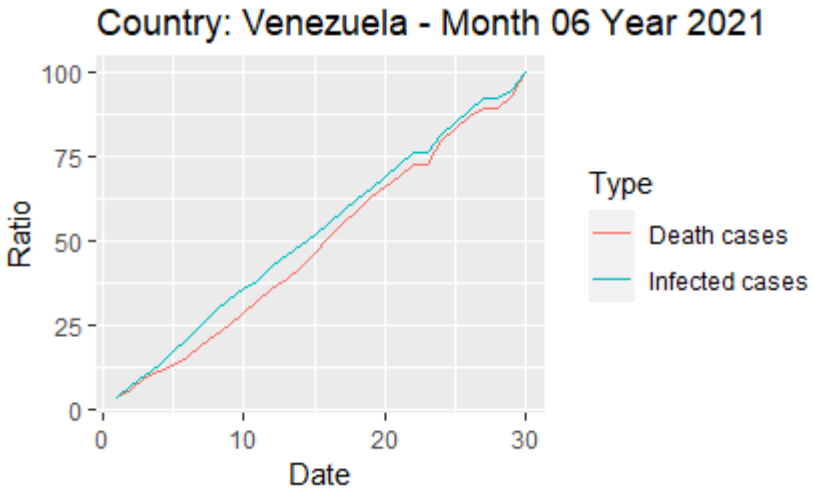

Country: Brazil - Month 03 Year 2021
Ratio
Date
Type
Death cases
Infected cases

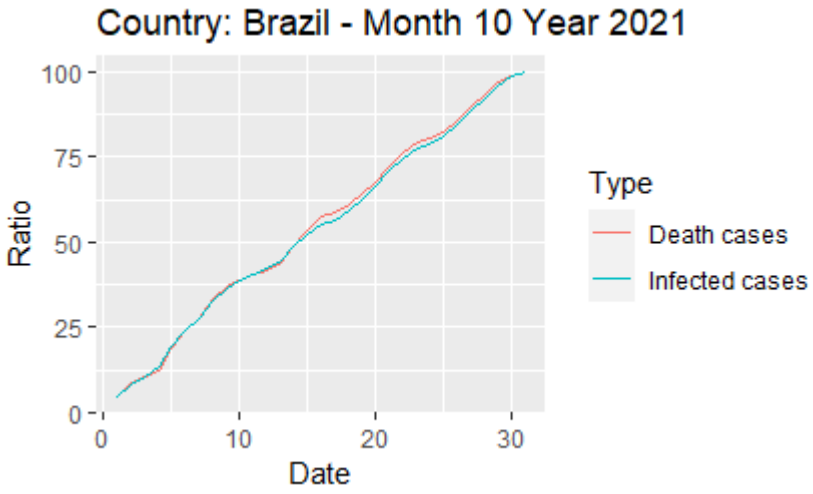

Country: Chile - Month 03 Year 2021
Ratio
Date
Type
Death cases
Infected cases

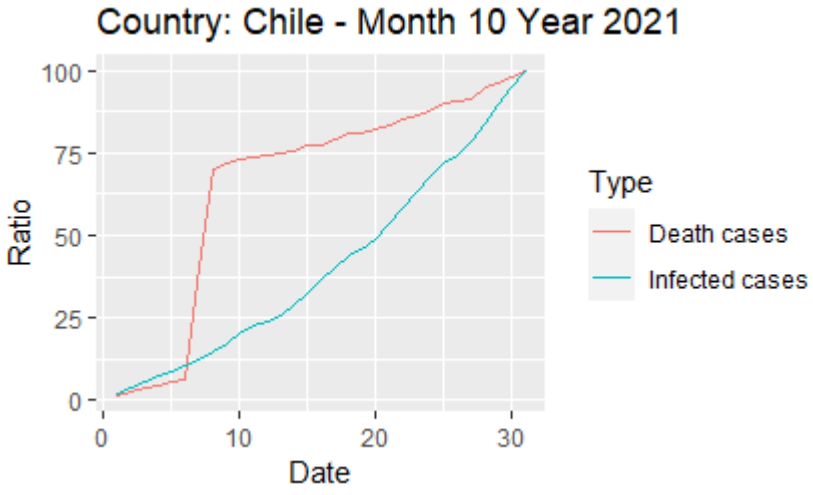

Country: Venezuela - Month 03 Year 2021
Ratio
Date
Type
Death cases
Infected cases

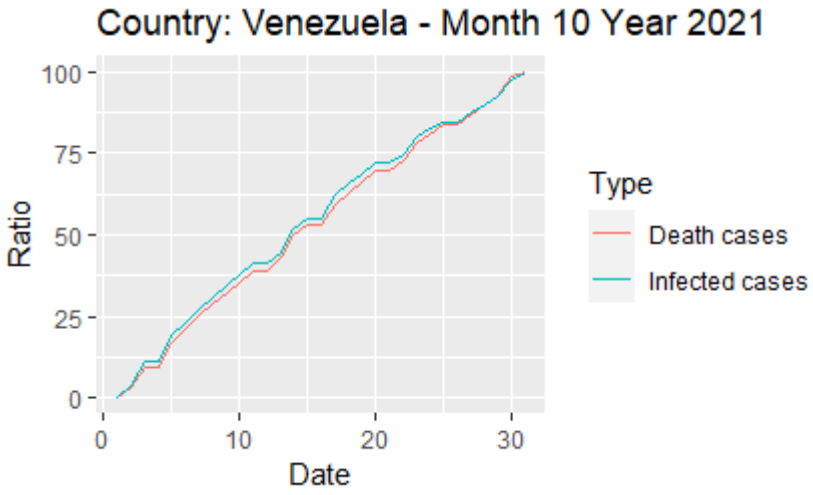

Country: Brazil - Month 06 Year 2021
Ratio
Date
Type
Death cases
Infected cases

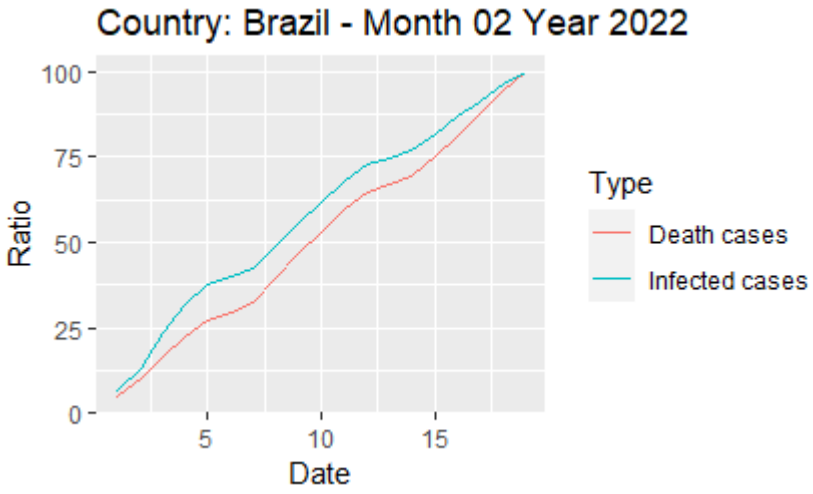

Country: Chile - Month 06 Year 2021
Ratio
Date
Type
Death cases
Infected cases

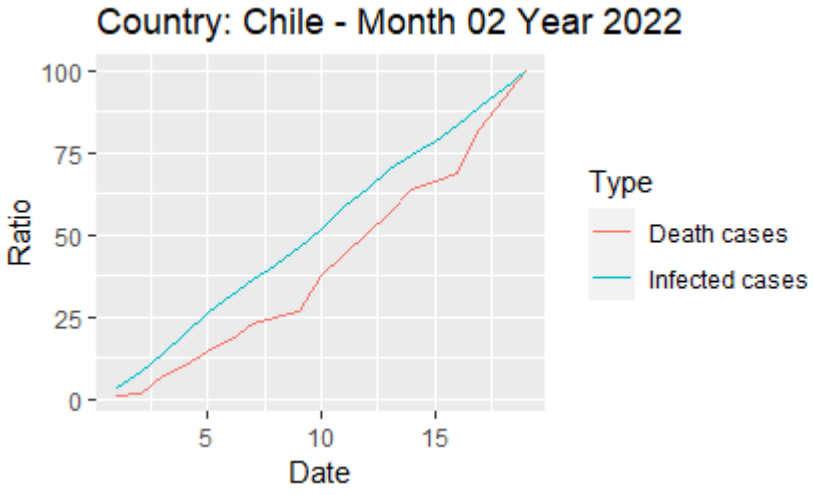

Country: Venezuela - Month 06 Year 2021
Ratio
Date
Type
Death cases
Infected cases

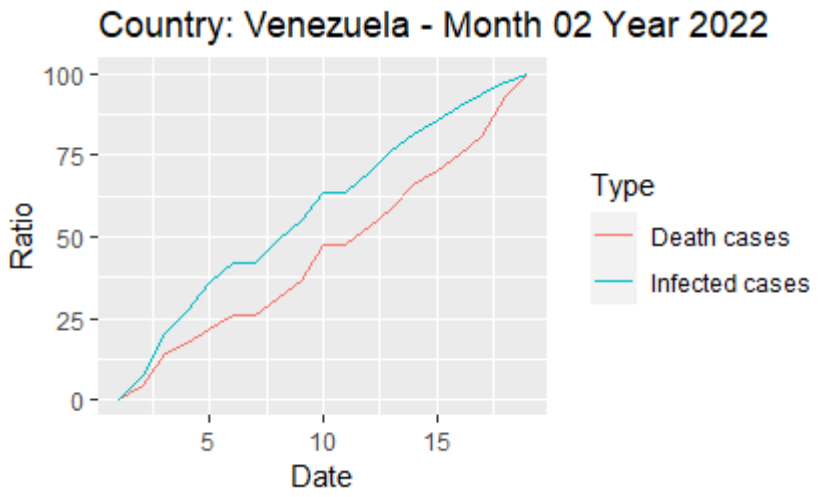

Country: Brazil - Month 10 Year 2021
Ratio
Date
Type
Death cases
Infected cases


Country: Chile - Month 10 Year 2021
Ratio
Date
Type
Death cases
Infected cases


Country: Venezuela - Month 10 Year 2021
Ratio
Date
Type
Death cases
Infected cases


Country: Brazil - Month 02 Year 2022
Ratio
Date
Type
Death cases
Infected cases


Country: Chile - Month 02 Year 2022
Ratio
Date
Type
Death cases
Infected cases

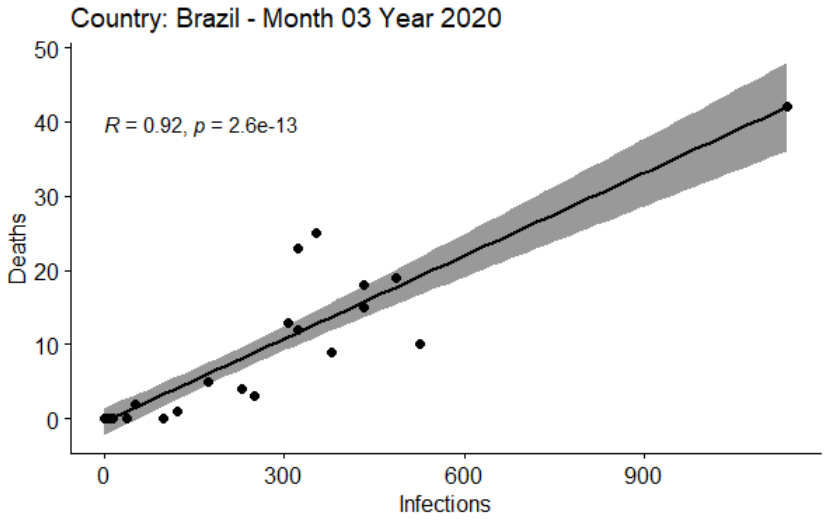


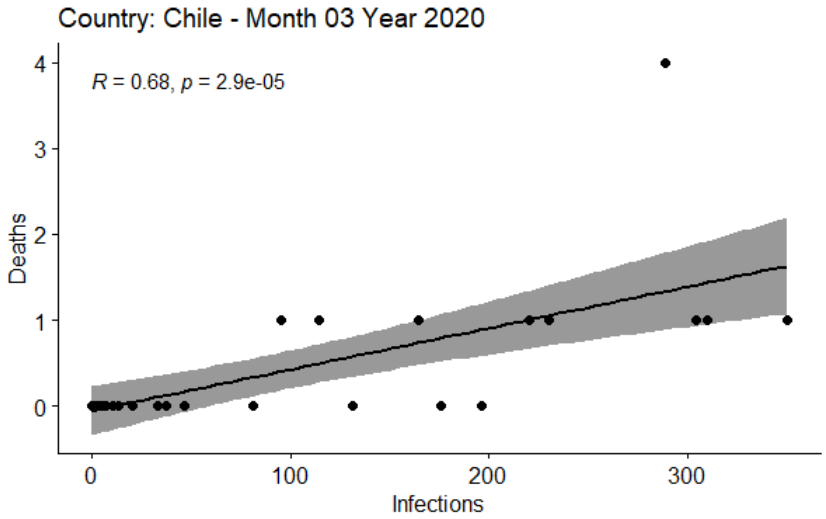
2) Xét tương quan trong mỗi tháng,

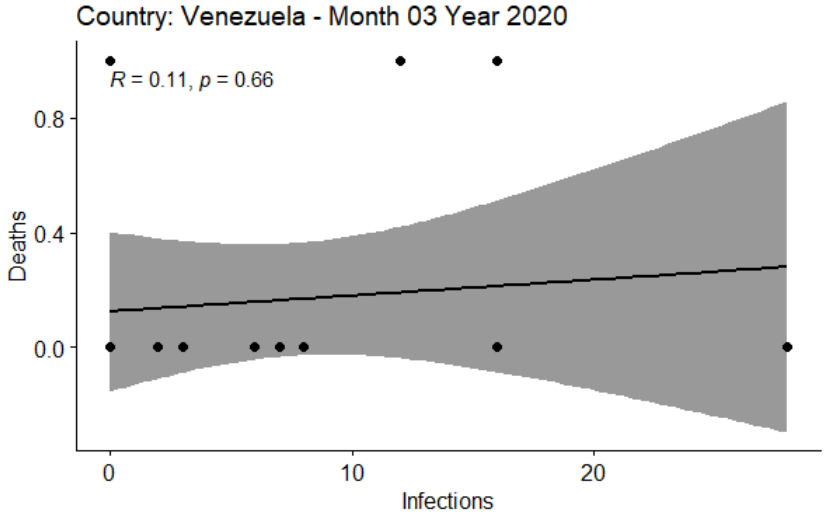
- Hiện thực trong R:

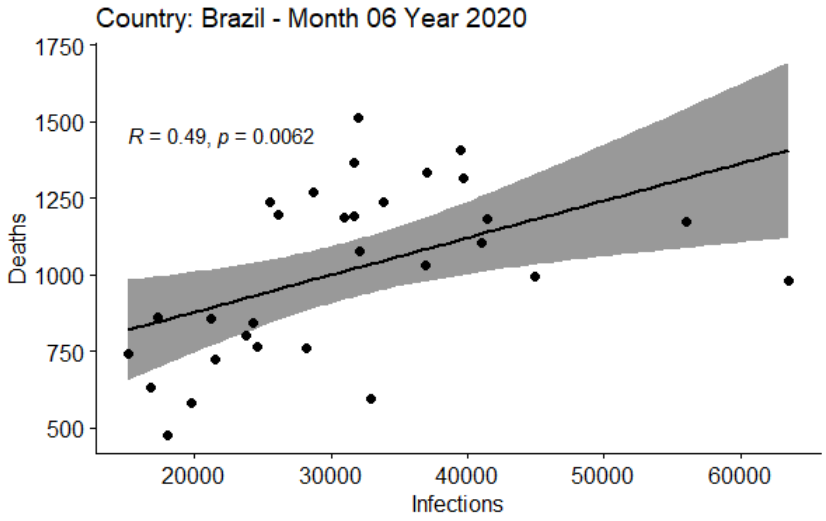
```
library(tidyverse)
library(dplyr)
library(datasets)
library("ggplot2")

covid = read.csv("owid-covid-data.csv", header = TRUE)

covid$new_cases <- abs(covid$new_cases)
covid$new_deaths <- abs(covid$new_deaths)

covid$date <- as.Date(covid$date, "%m/%d/%Y")
covid$month <- as.numeric(format(covid$date,'%m'))
covid$year <- as.numeric(format(covid$date,'%Y'))

covid <- subset(covid, !is.na(continent))
covid$new_cases[is.na(covid$new_cases)] <- 0
covid$new_deaths[is.na(covid$new_deaths)] <- 0
year <- unique(format(covid$date, format= "%Y"))
lcountry <- c("Brazil","Chile","Venezuela")
lmonth <- c("02","03","06","10")



for (i in year){
  for (k in lmonth){
    for (c in lcountry){
      temp = subset(covid, format(covid$date, format= "%Y") == i & format(covid$date, format= "%m") == k & covid$location == c)
      print(ggscatter(temp, x = "new_cases", y = "new_deaths",
                      add = "reg.line", conf.int = TRUE,
                      cor.coef = TRUE, cor.method = "pearson",
                      xlab = "Infections", ylab = "Deaths") + labs(title = paste("Country:",c, "- Month",k,"Year",i)) )
    }
  }
}
```

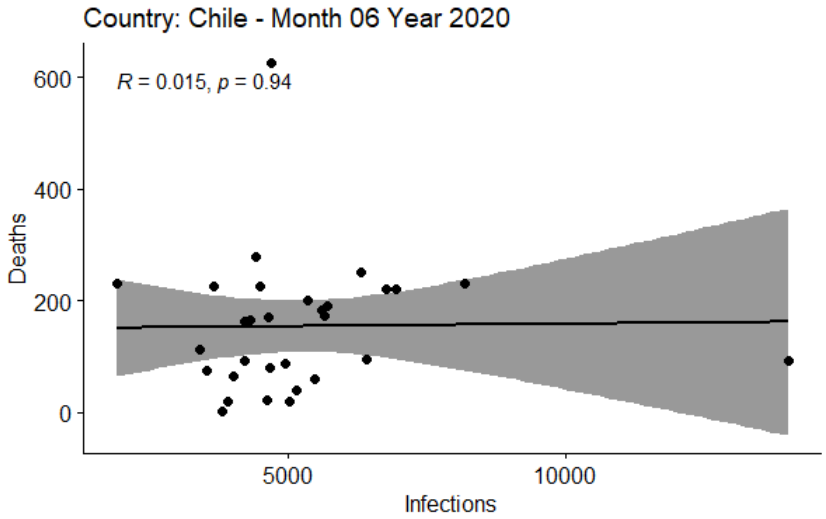
- Kết quả

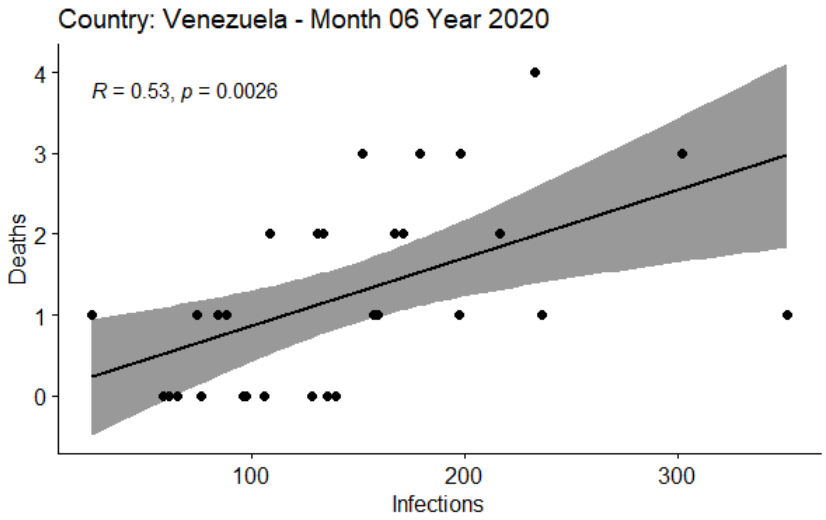

Country: Brazil - Month 03 Year 2020
R = 0.92, p = 2.6e-13
Deaths
Infections
0
10
20
30
40
50
0
300
600
900

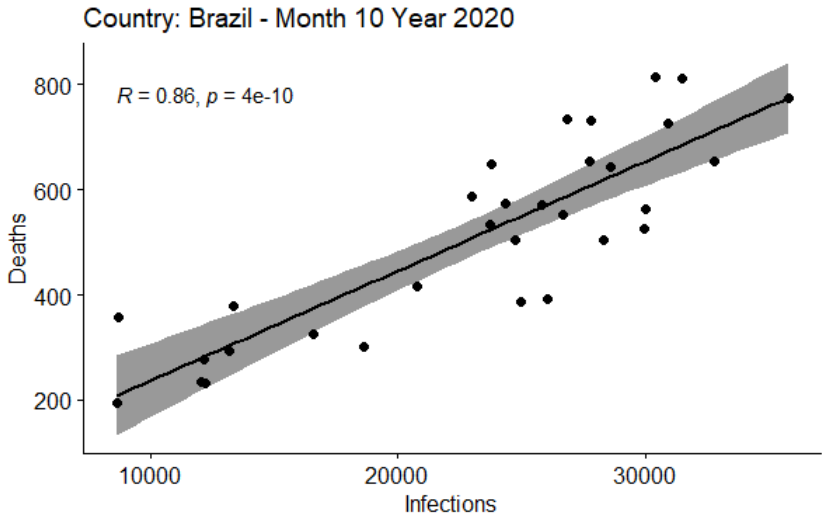

Country: Chile - Month 03 Year 2020
R = 0.68, p = 2.9e-05
Deaths
Infections
0
1
2
3
4
0
100
200
300

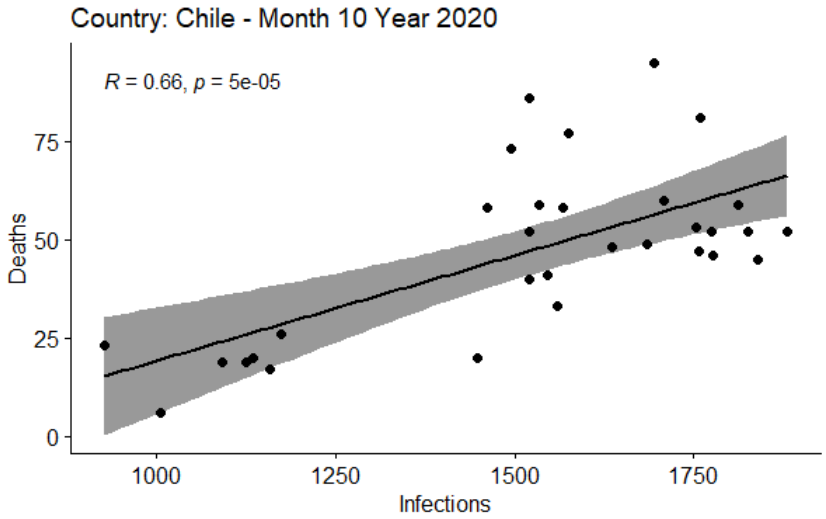

Country: Venezuela - Month 03 Year 2020
R = 0.11, p = 0.66
Deaths
Infections
0.0
0.4
0.8
0
10
20

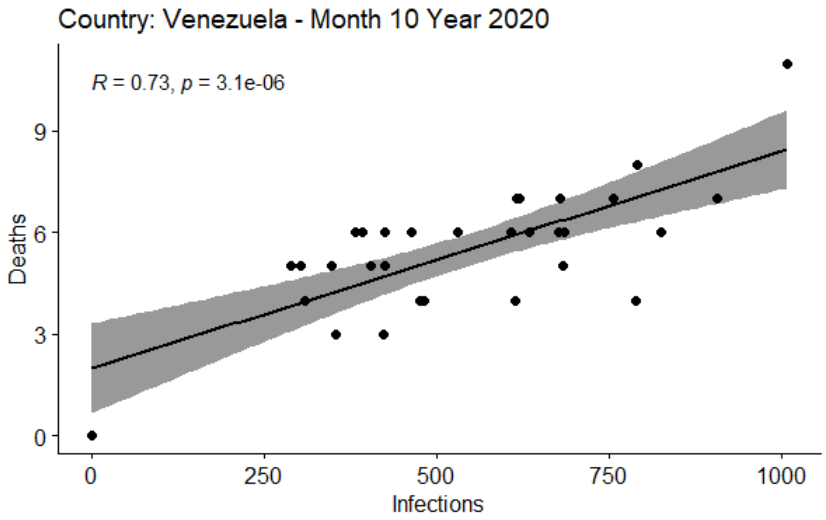

Country: Brazil - Month 06 Year 2020
R = 0.49, p = 0.0062
Deaths
Infections
1750
1500
1250
1000
750
500
20000
30000
40000
50000
60000

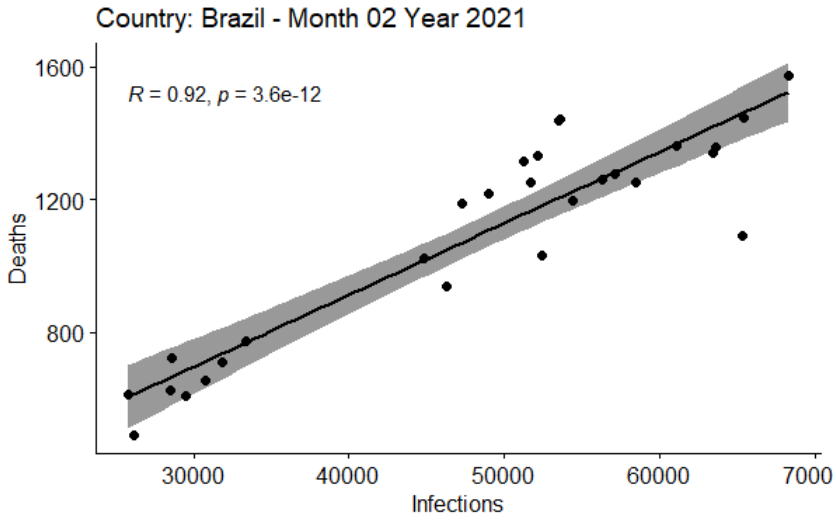

Country: Chile - Month 06 Year 2020
R = 0.015, p = 0.94
Deaths
Infections
600
400
200
0
5000
10000

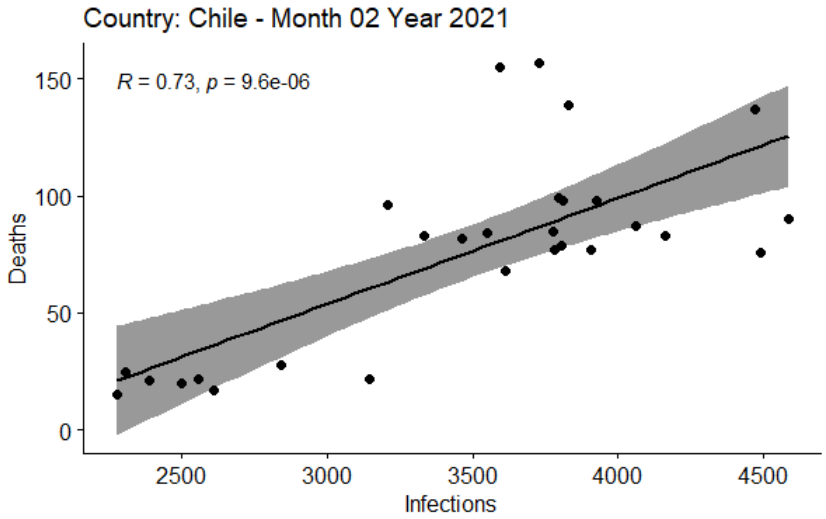

Country: Venezuela - Month 06 Year 2020
R = 0.53, p = 0.0026
Deaths
Infections
4
3
2
1
0
100
200
300

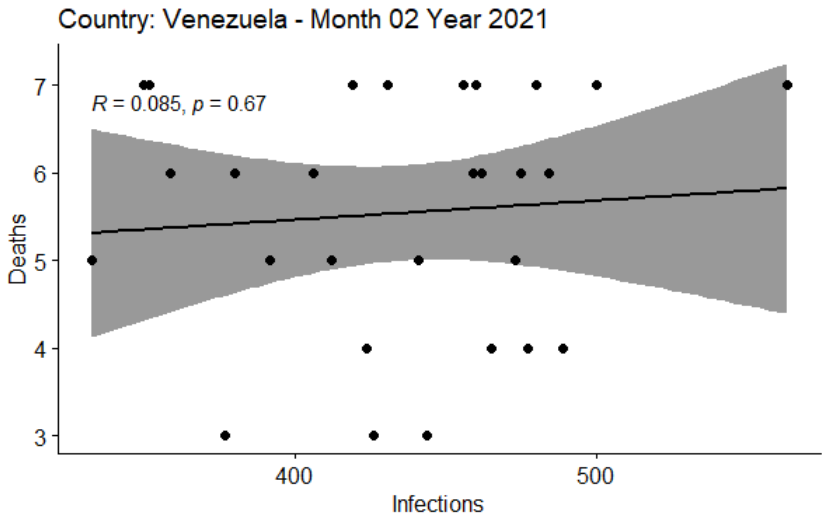

Country: Brazil - Month 10 Year 2020
R = 0.86, p = 4e-10
Deaths
800
600
400
200
10000
20000
30000
Infections

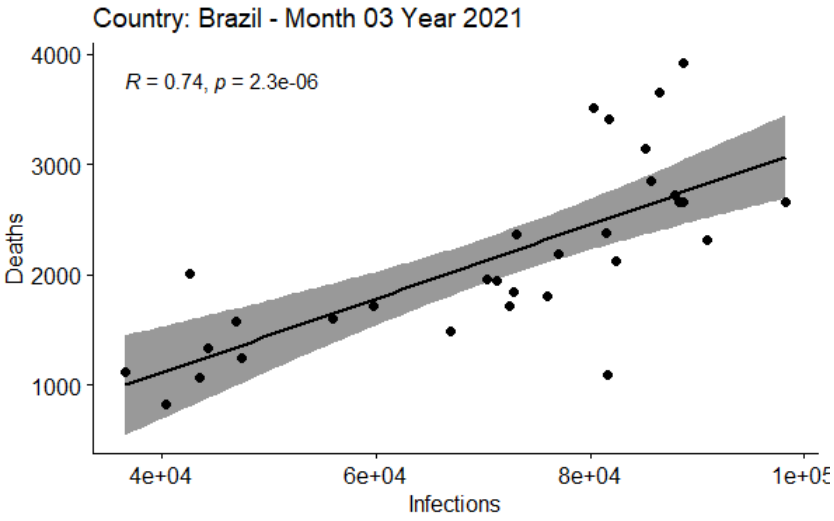

Country: Chile - Month 10 Year 2020
R = 0.66, p = 5e-05
Deaths
75
50
25
0
1000
1250
1500
1750
Infections

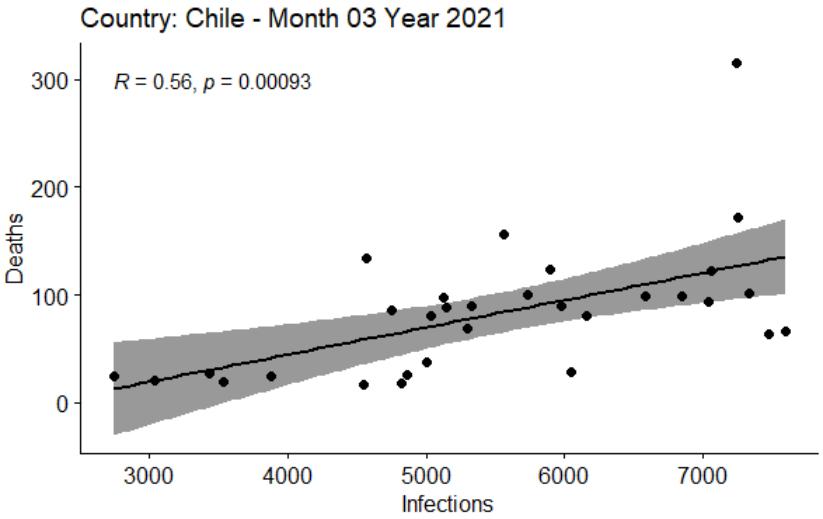

Country: Venezuela - Month 10 Year 2020
R = 0.73, p = 3.1e-06
Deaths
9
6
3
0
0
250
500
750
1000
Infections

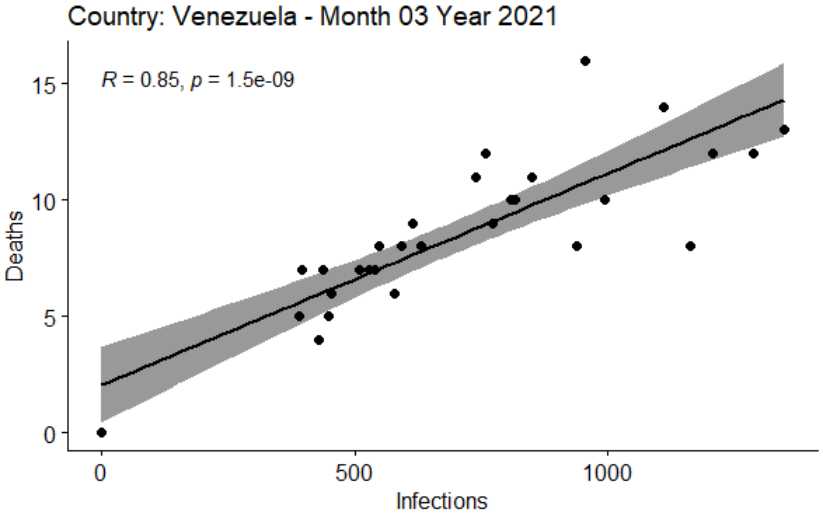

Country: Brazil - Month 02 Year 2021
R = 0.92, p = 3.6e-12
Deaths
1600
1200
800
30000
40000
50000
60000
7000
Infections

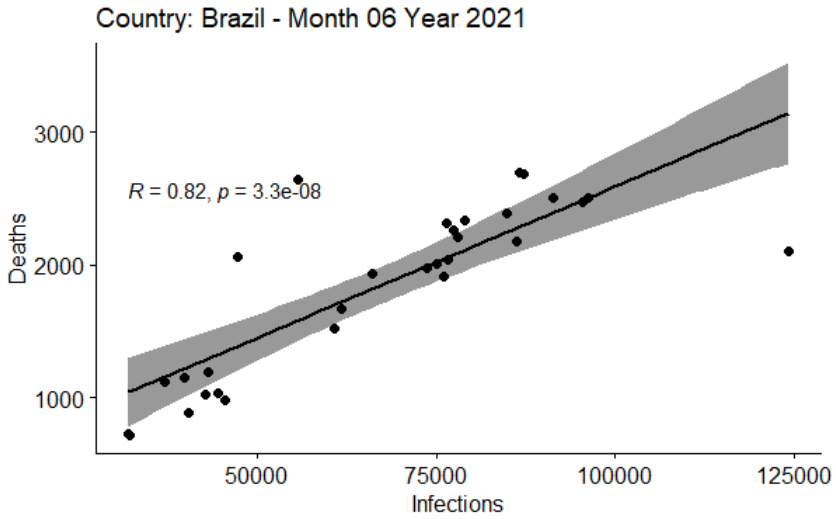

Country: Chile - Month 02 Year 2021
R = 0.73, p = 9.6e-06
Deaths
150
100
50
0
2500
3000
3500
4000
4500
Infections

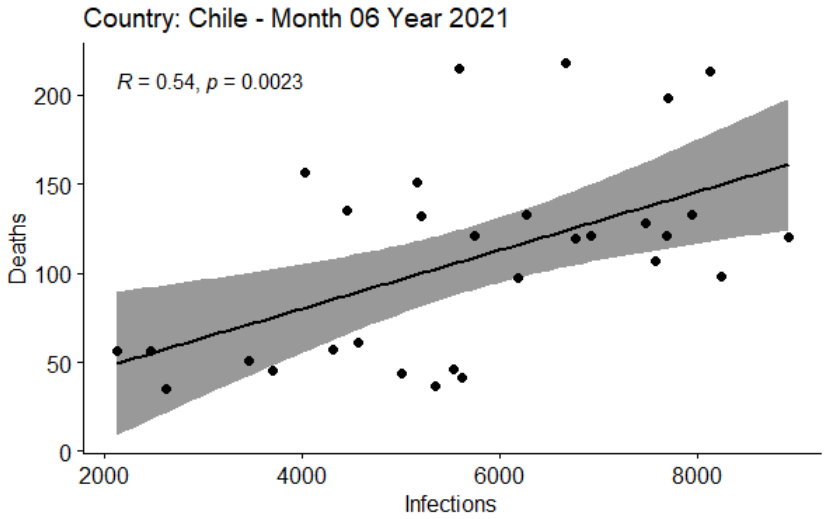

Country: Venezuela - Month 02 Year 2021
R = 0.085, p = 0.67
Deaths
7
6
5
4
3
400
500
Infections

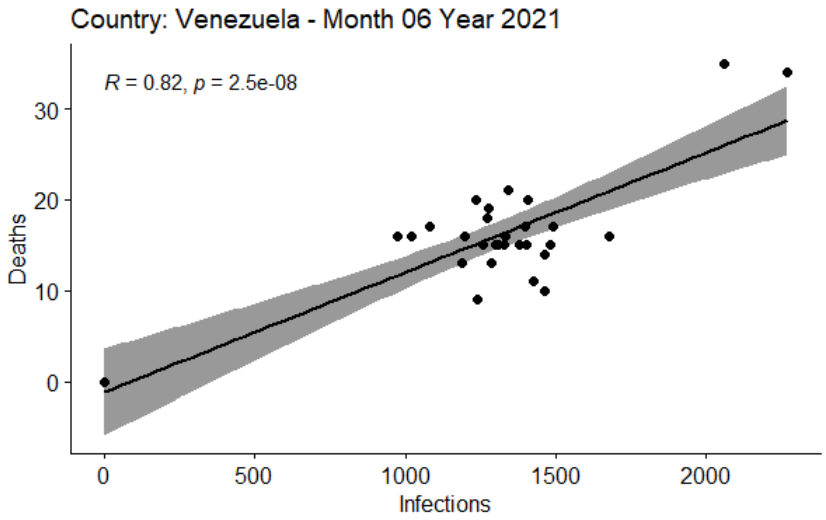

Country: Brazil - Month 03 Year 2021
R = 0.74, p = 2.3e-06
Deaths
Infections
4000
3000
2000
1000
4e+04
6e+04
8e+04
1e+05

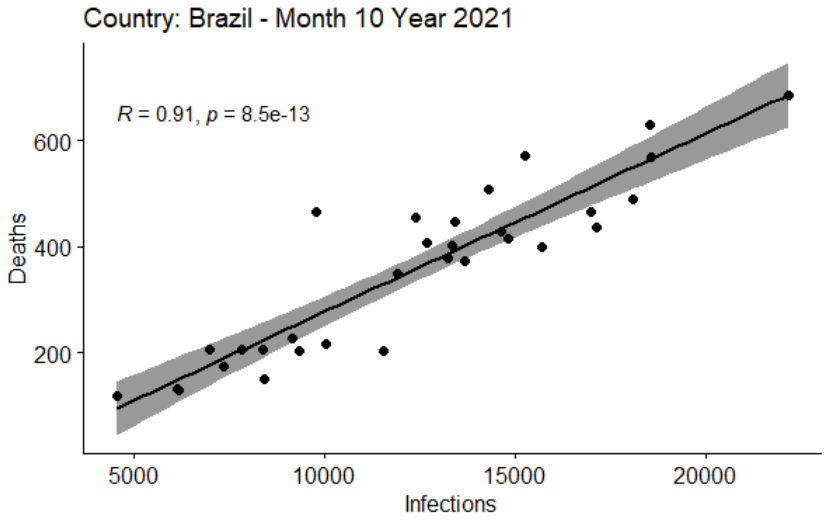

Country: Chile - Month 03 Year 2021
R = 0.56, p = 0.00093
Deaths
Infections
300
200
100
0
3000
4000
5000
6000
7000

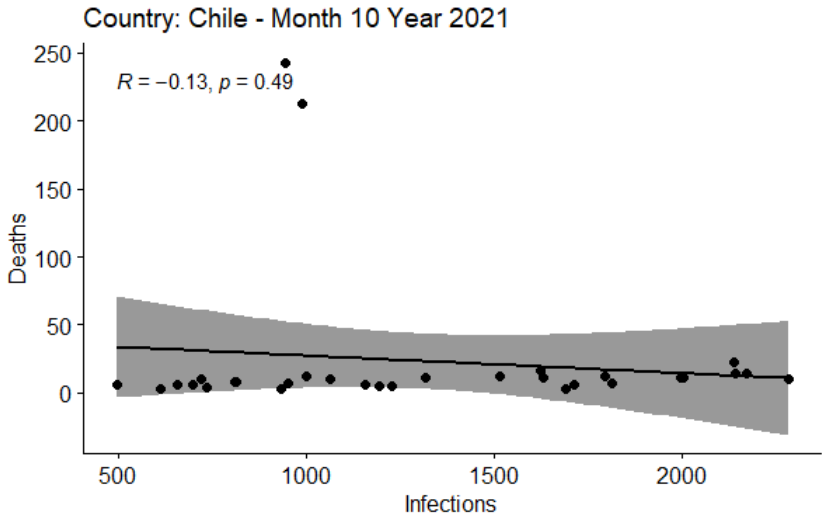

Country: Venezuela - Month 03 Year 2021
R = 0.85, p = 1.5e-09
Deaths
Infections
15
10
5
0
0
500
1000

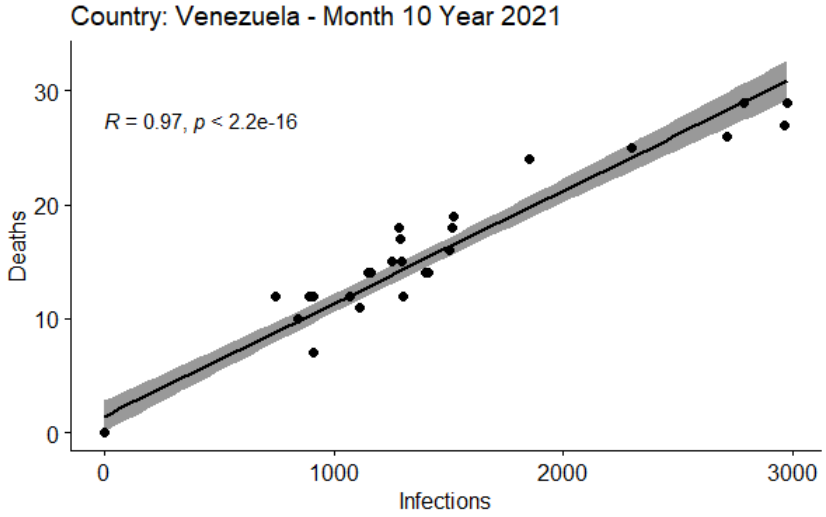

Country: Brazil - Month 06 Year 2021
R = 0.82, p = 3.3e-08
Deaths
Infections
3000
2000
1000
50000
75000
100000
125000

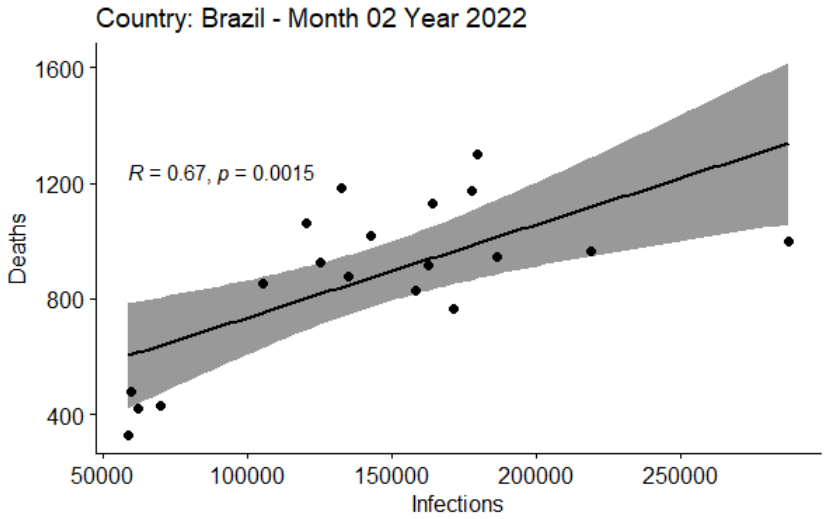

Country: Chile - Month 06 Year 2021
R = 0.54, p = 0.0023
Deaths
Infections
200
150
100
50
0
2000
4000
6000
8000

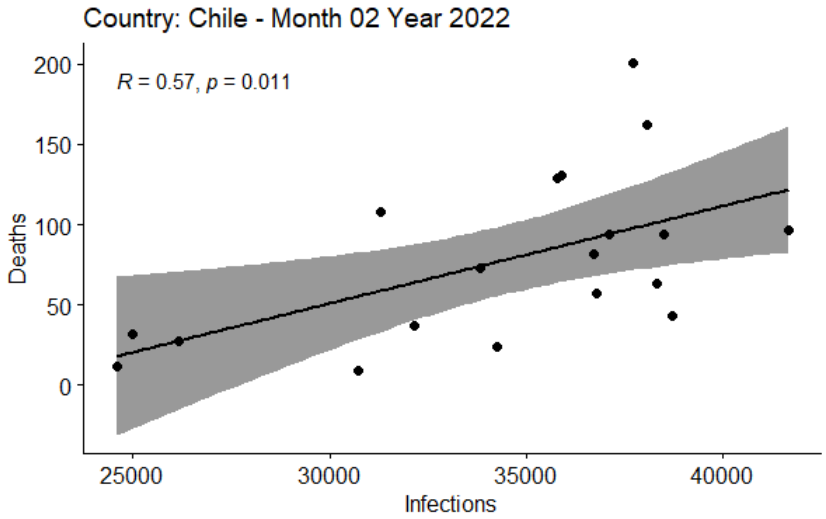

Country: Venezuela - Month 06 Year 2021
R = 0.82, p = 2.5e-08
Deaths
Infections
30
20
10
0
0
500
1000
1500
2000

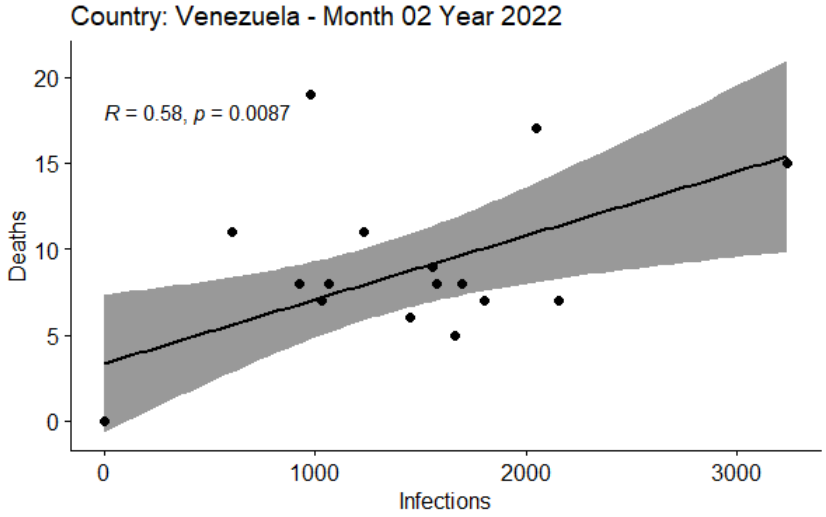

Country: Brazil - Month 10 Year 2021
R = 0.91, p = 8.5e-13
Deaths
600
400
200
5000
10000
15000
20000
Infections


Country: Chile - Month 10 Year 2021
R = −0.13, p = 0.49
Deaths
250
200
150
100
50
0
500
1000
1500
2000
Infections


Country: Venezuela - Month 10 Year 2021
R = 0.97, p < 2.2e-16
Deaths
30
20
10
0
0
1000
2000
3000
Infections







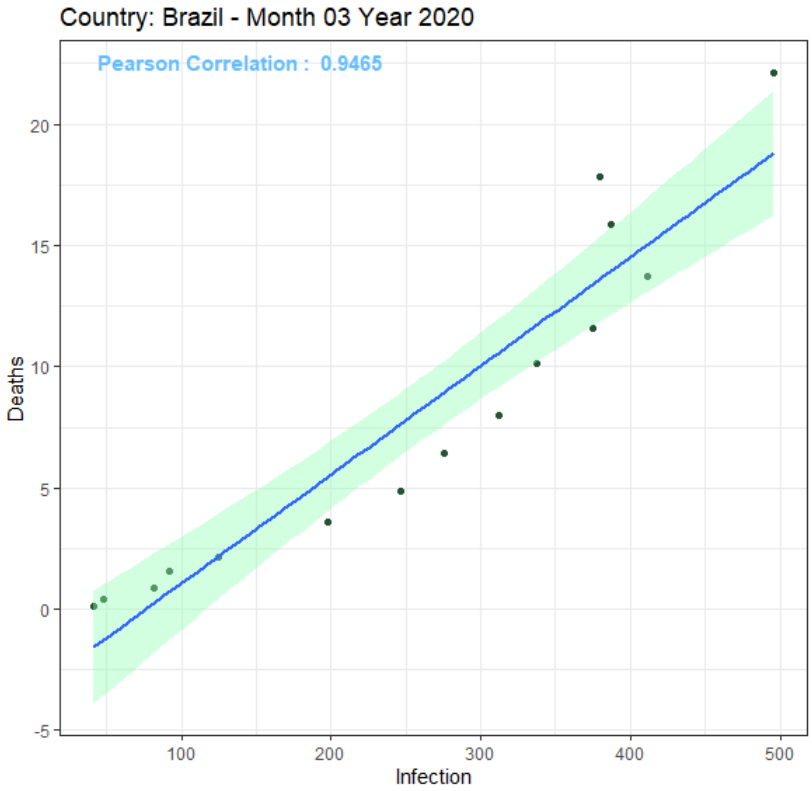
3) Xét tương quan trong mỗi tháng theo trung bình 7 ngày gần nhất

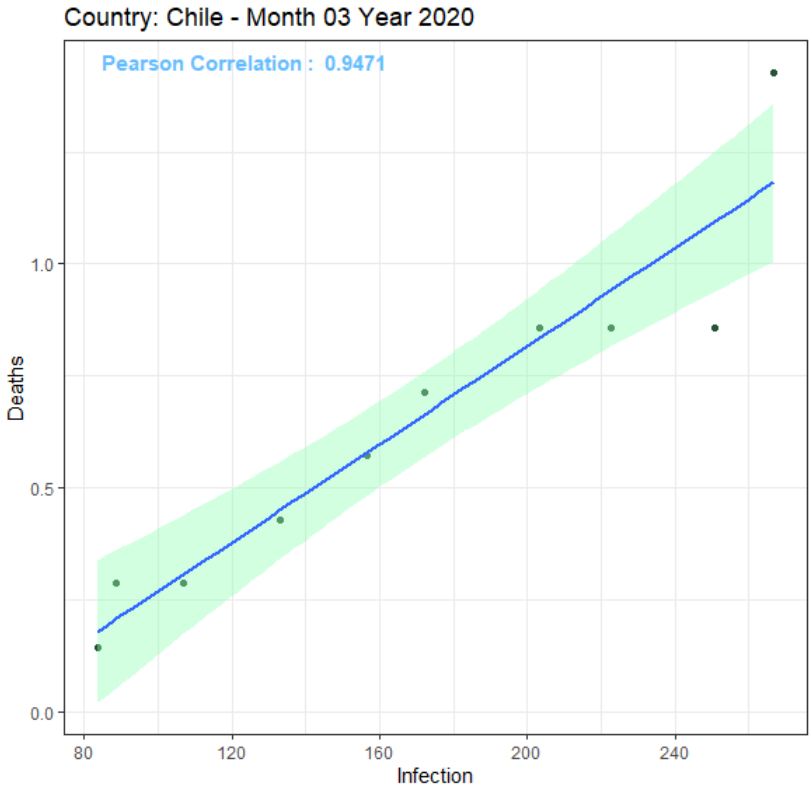
- Hiện thực trong R:

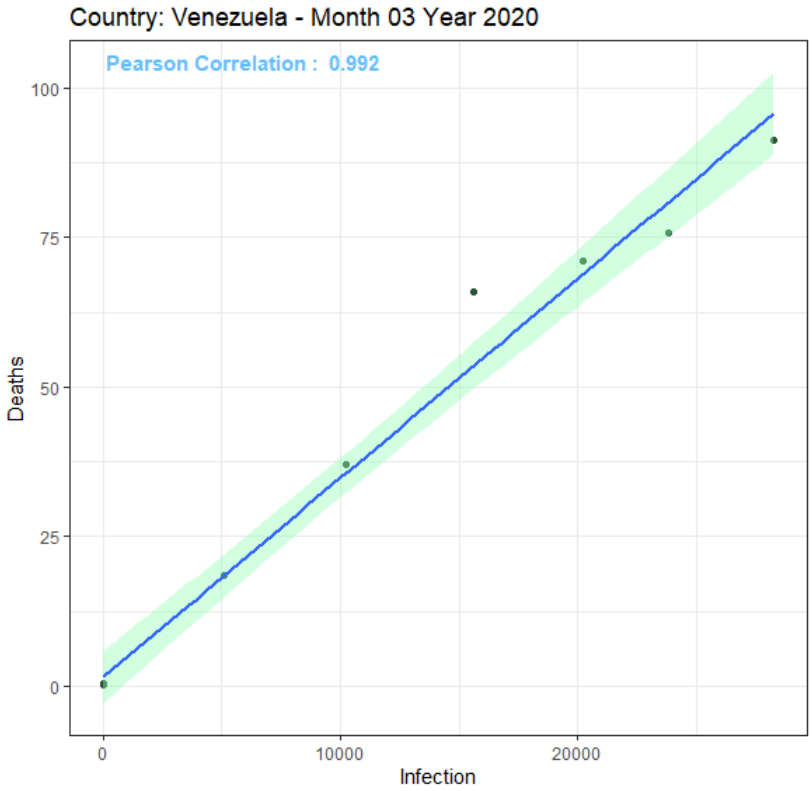
```
library(tidyverse)
library(dplyr)
library(datasets)
library("ggplot2")
```

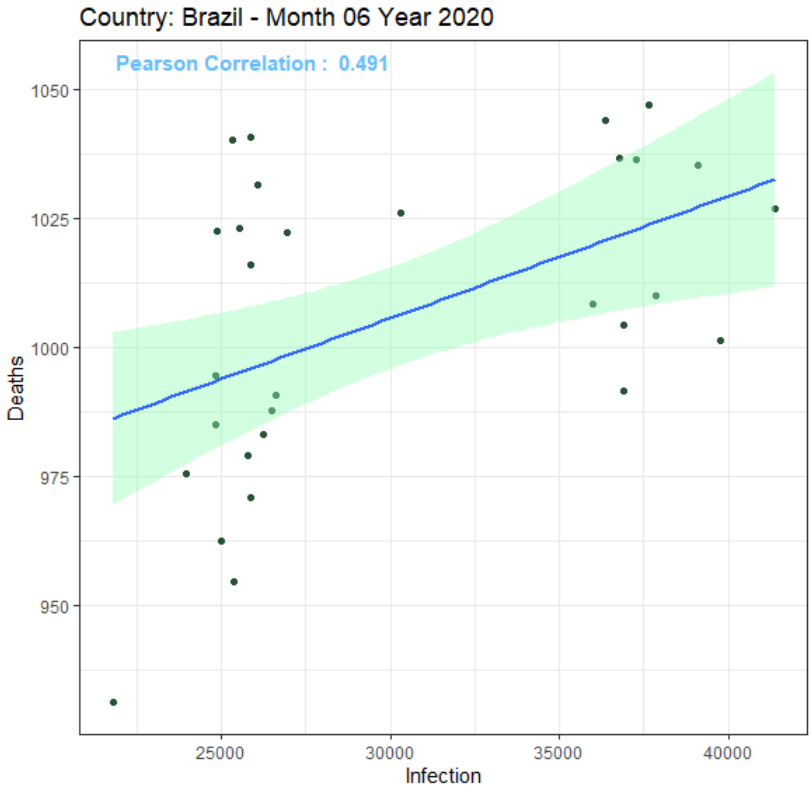
```
library(grid)
library(gridExtra)
covid = read.csv("C:/Users/Admin/Downloads/covidData.csv")

covid$new_cases <- abs(covid$new_cases)
covid$new_deaths <- abs(covid$new_deaths)

covid$date <- as.Date(covid$date, "%m/%d/%Y")
covid$month <- as.numeric(format(covid$date,'%m'))
covid$year <- as.numeric(format(covid$date,'%Y'))

covid <- subset(covid, !is.na(continent))
covid$new_cases[is.na(covid$new_cases)] <- 0
covid$new_deaths[is.na(covid$new_deaths)] <- 0
year <- unique(format(covid$date, format= "%Y"))
lcountry <- c("Brazil","Chile","Venezuela")
lmonth <- c("02","03","06","10")

covid[["Infection"]] = NA
covid[["Deaths"]] = NA
covid = covid %>% filter(iso_code== "BRA" | iso_code =="CHL"| iso_code =="VEN")
for(j in 1:nrow(covid)) {
  if(j <=7) {
    covid$Infection[[j]] = mean(covid$new_cases[1:j])
    covid$Deaths[[j]] = mean(covid$new_deaths[1:j])
  }
  else{
    covid$Infection[[j]] = sum( covid$new_cases[(j-6):j]  )/7
    covid$Deaths[[j]] = sum( covid$new_deaths[(j-6):j]  )/7
  }
}

for (i in year){
  for (k in lmonth){
    if(i==2022 & k >2) break
    for (c in lcountry){
      #temp = covid %>% subset( location == c) %>% filter()

      temp = subset(covid, format(covid$date, format= "%Y") == i & format(covid$date, format=
    "%m") == k & covid$location == c)
      temp = temp %>% filter(!Infection==0 | !Deaths ==0)

     temp = temp %>% filter(!Infection ==0) %>% filter(!Deaths ==0)
      #remove Outliers

      grob1 = grobTree(textGrob(paste("Pearson Correlation : ", round(cor(temp$Infection, temp
    $Deaths ), 4) ), x = 0.05, y = 0.97, hjust = 0, gp = gpar(col = "#63BDFF", fontsize = 11,
    fontface = "bold")))
      print(      ggplot(temp, aes(x =Infection, y=Deaths) ) +  geom_point(color = "#275235")
    + geom_smooth(formula = y ~ x, method = "lm", fill = "#91FFB4")+annotation_custom(grob1)+
    theme_bw() +labs(title = paste("Country:",c,"- Month",k,"Year",i) )
)
      cat(paste("Country:",c,"- Month",k,"Year",i))
    }
  }
}
```

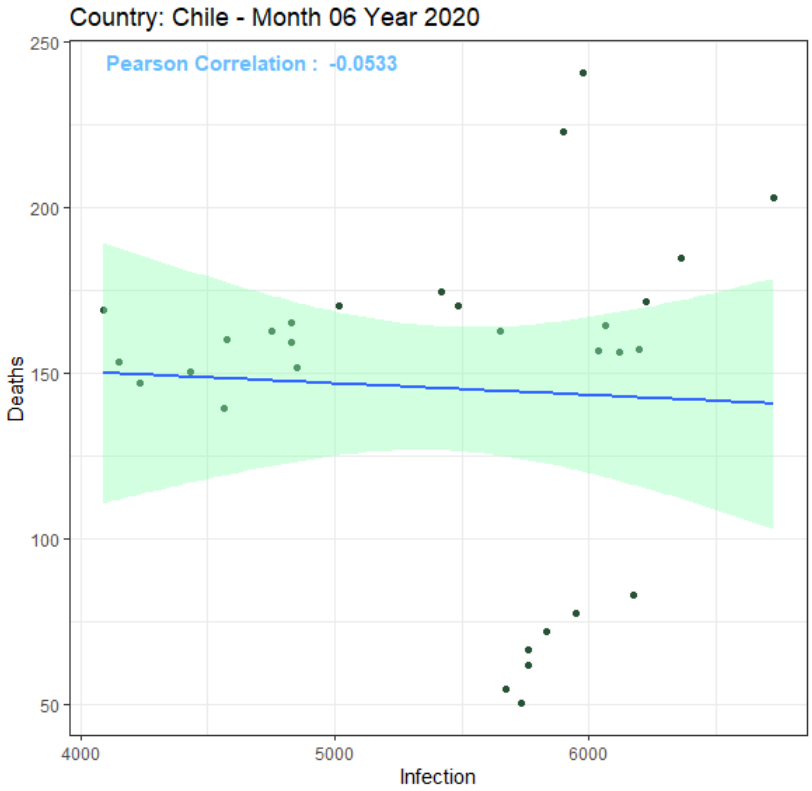
- Kết quả

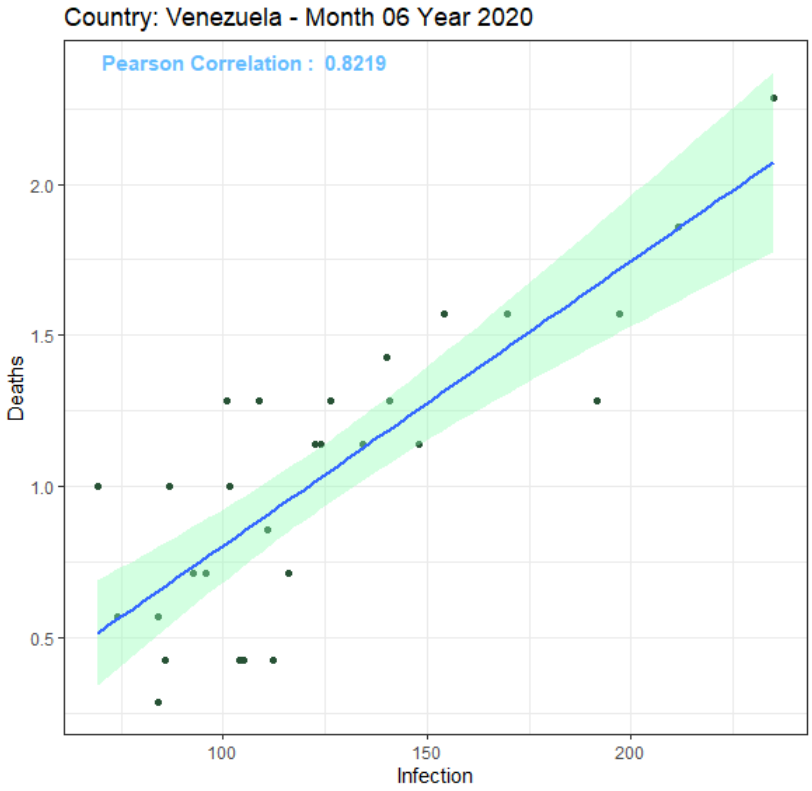

Country: Brazil - Month 03 Year 2020
Pearson Correlation : 0.9465
Deaths
Infection
20
15
10
5
0
-5
100
200
300
400
500

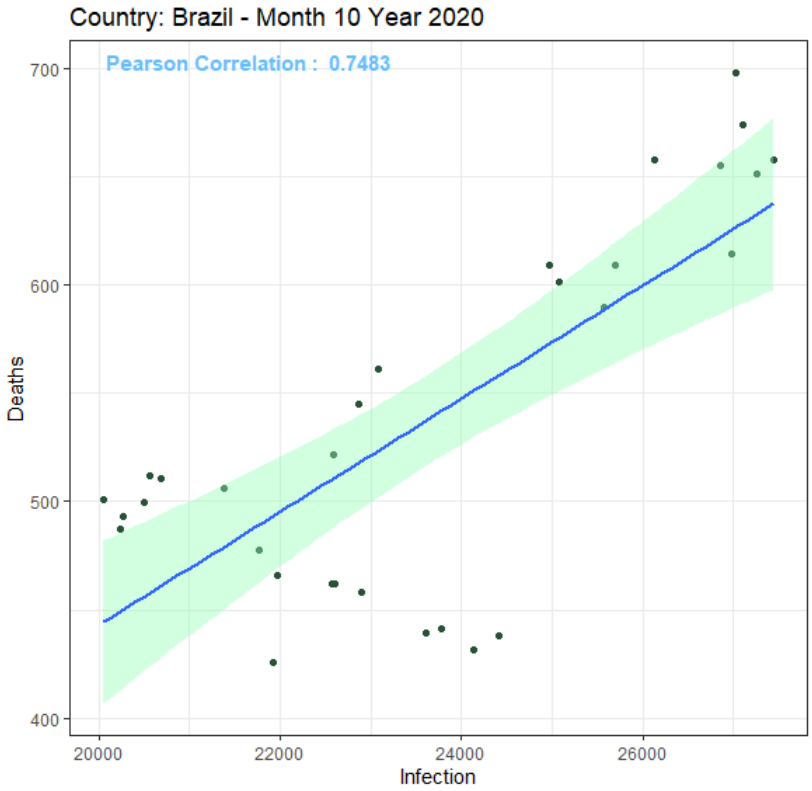

Country: Chile - Month 03 Year 2020
Pearson Correlation : 0.9471
Deaths
Infection
1.0
0.5
0.0
80
120
160
200
240

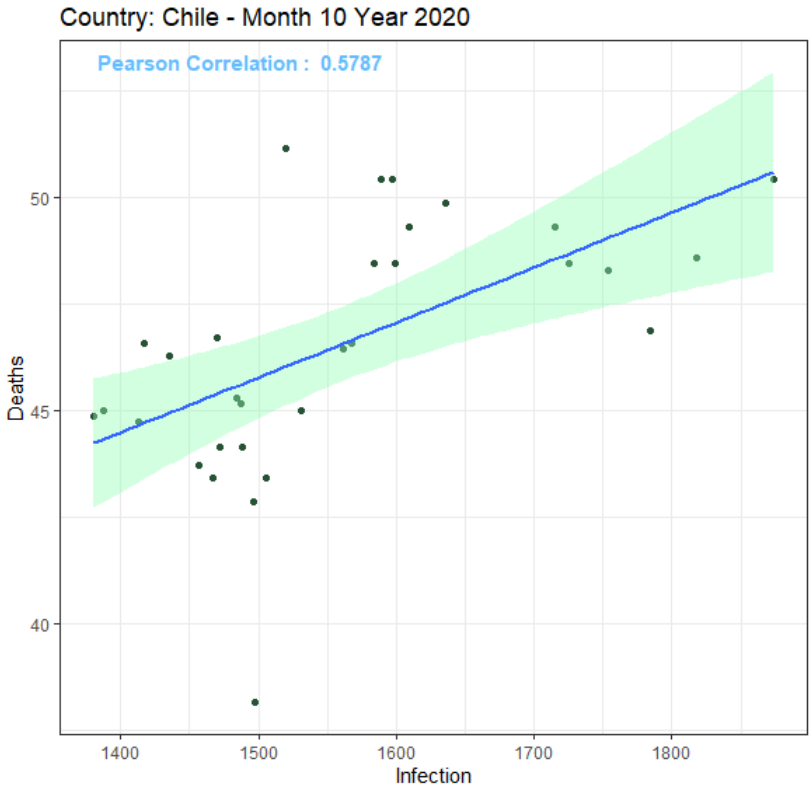

Country: Venezuela - Month 03 Year 2020
Pearson Correlation : 0.992
Deaths
Infection
0
25
50
75
100
0
10000
20000

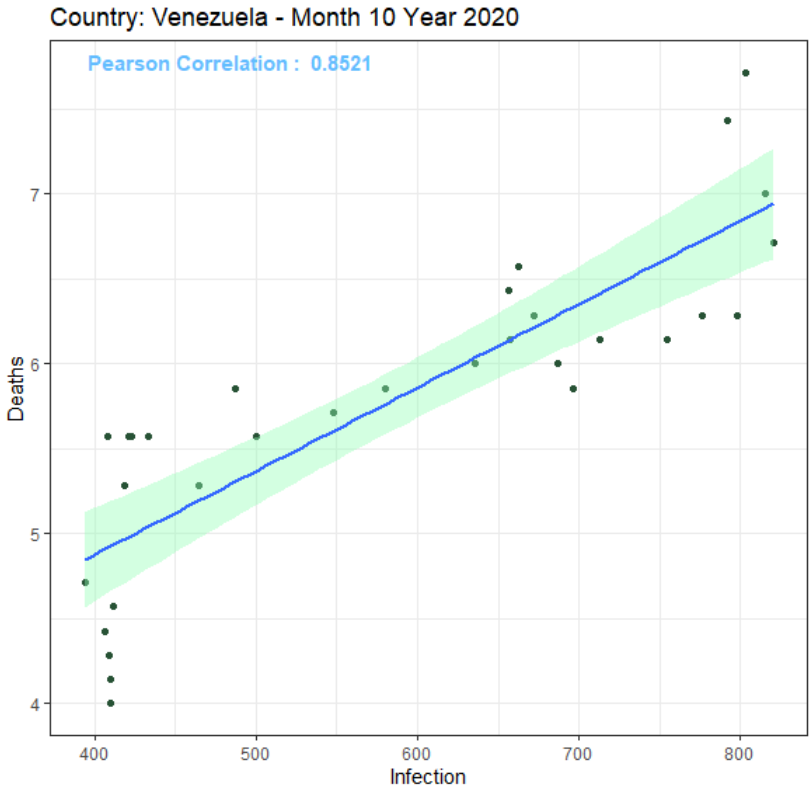

Country: Brazil - Month 06 Year 2020
Pearson Correlation : 0.491
Deaths
Infection
950
975
1000
1025
1050
25000
30000
35000
40000

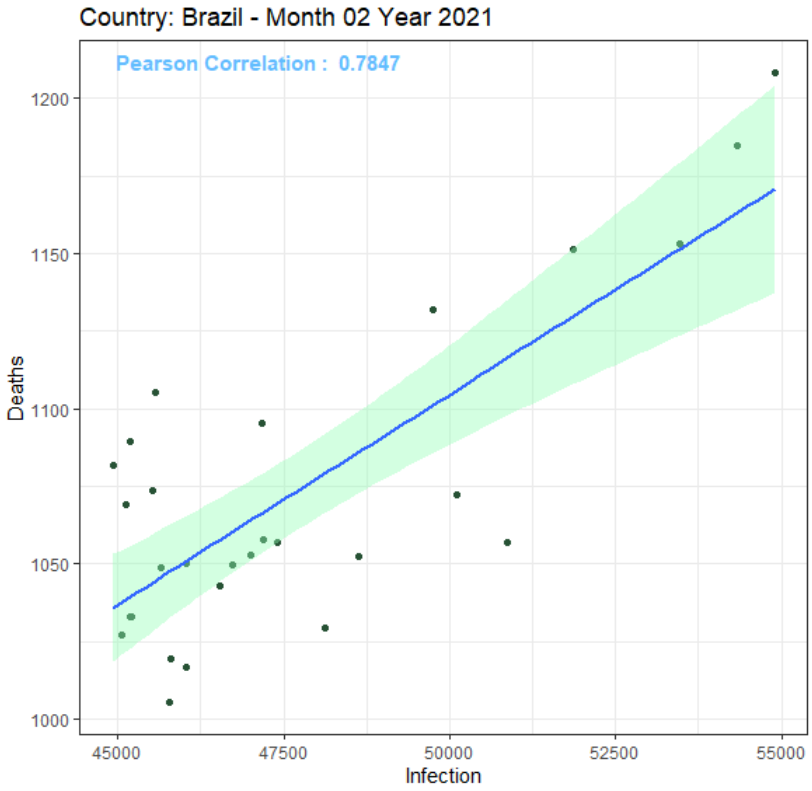

Country: Chile - Month 06 Year 2020
Pearson Correlation : -0.0533
Deaths
Infection
250
200
150
100
50
4000
5000
6000

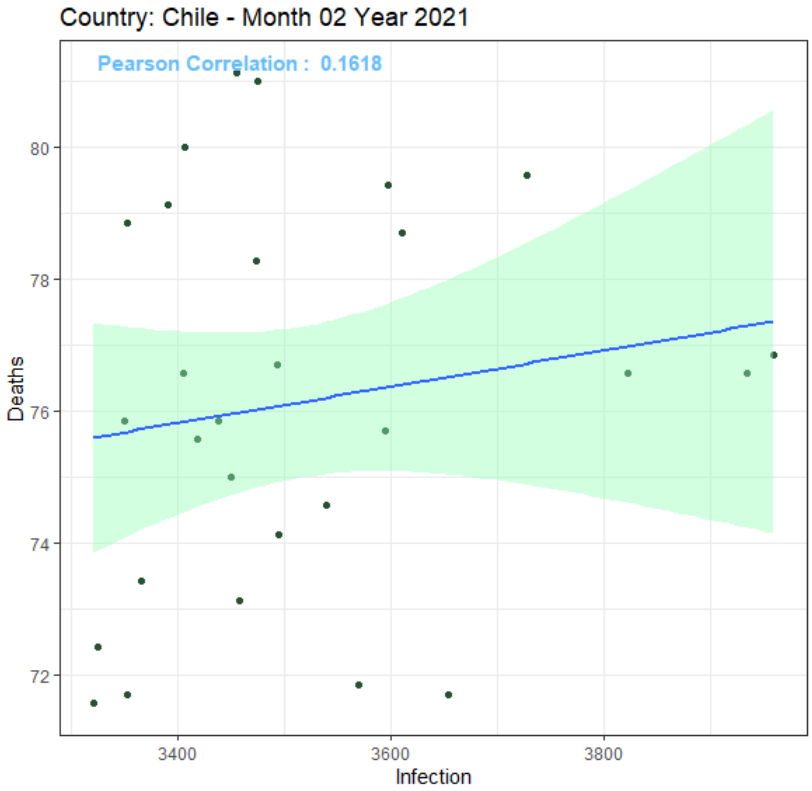

Country: Venezuela - Month 06 Year 2020
Pearson Correlation : 0.8219
Deaths
Infection
2.0
1.5
1.0
0.5
100
150
200

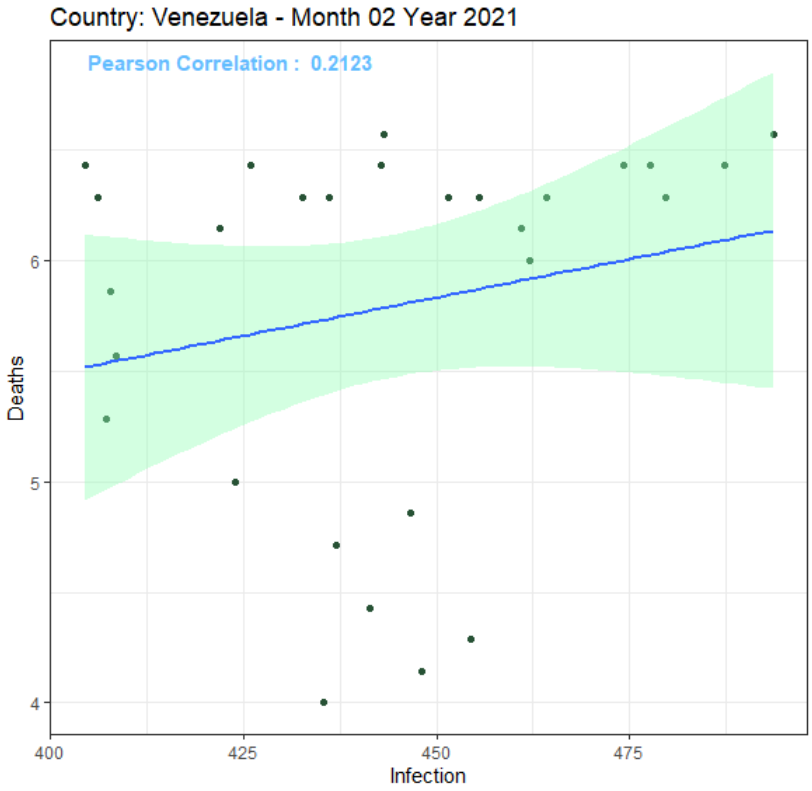

Country: Brazil - Month 10 Year 2020
Pearson Correlation : 0.7483
Deaths
Infection
700
600
500
400
20000
22000
24000
26000

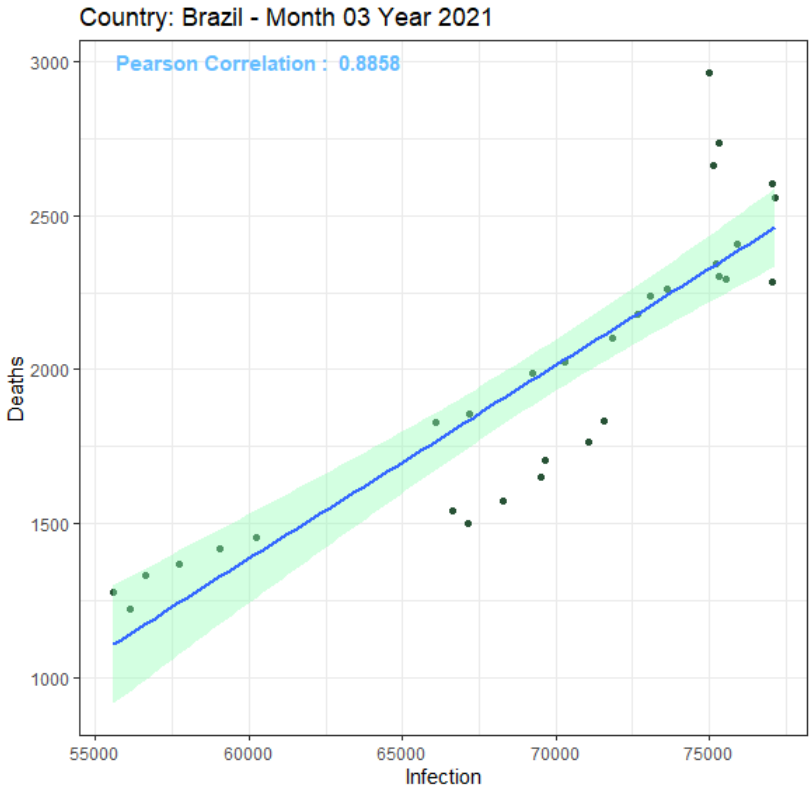

Country: Chile - Month 10 Year 2020
Pearson Correlation : 0.5787
Deaths
Infection
50
45
40
1400
1500
1600
1700
1800

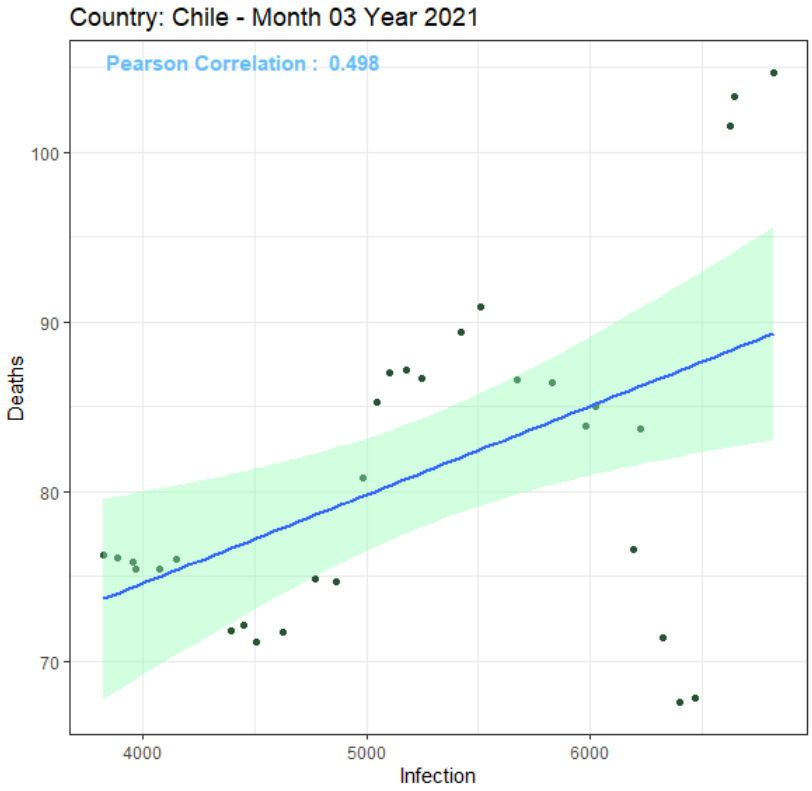

Country: Venezuela - Month 10 Year 2020
Pearson Correlation : 0.8521
Deaths
Infection
4
5
6
7
400
500
600
700
800

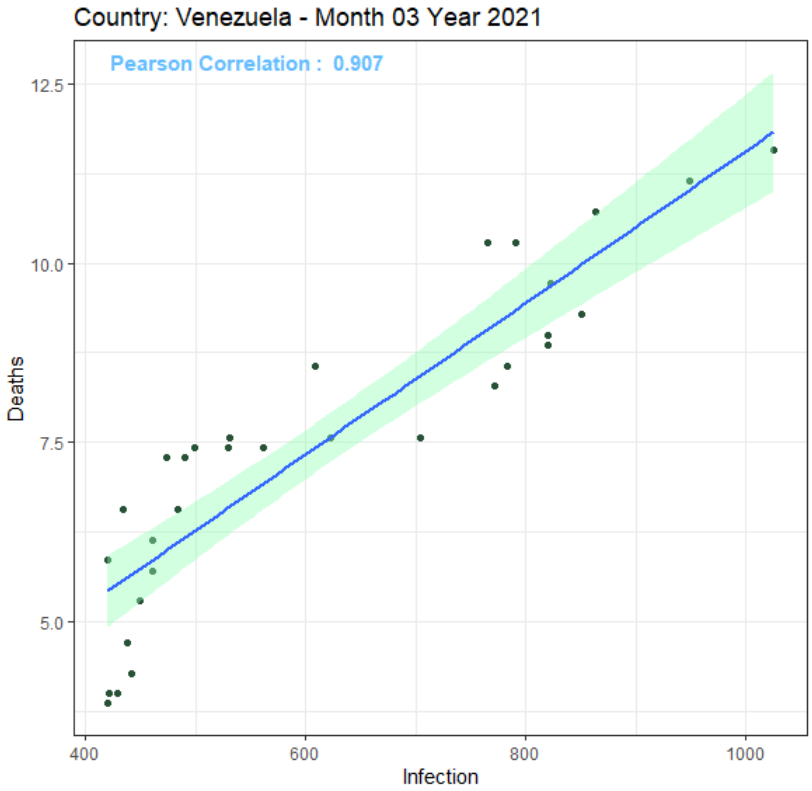

Country: Brazil - Month 02 Year 2021
Pearson Correlation : 0.7847
Deaths
Infection
1000
1050
1100
1150
1200
45000
47500
50000
52500
55000

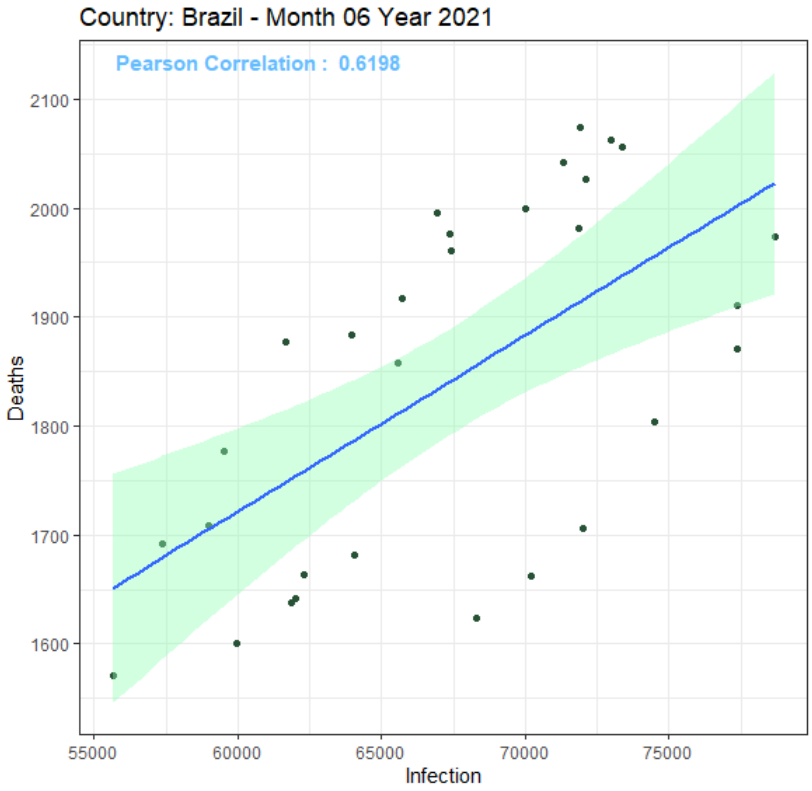

Country: Chile - Month 02 Year 2021
Pearson Correlation : 0.1618
Deaths
80
78
76
74
72
3400
3600
3800
Infection

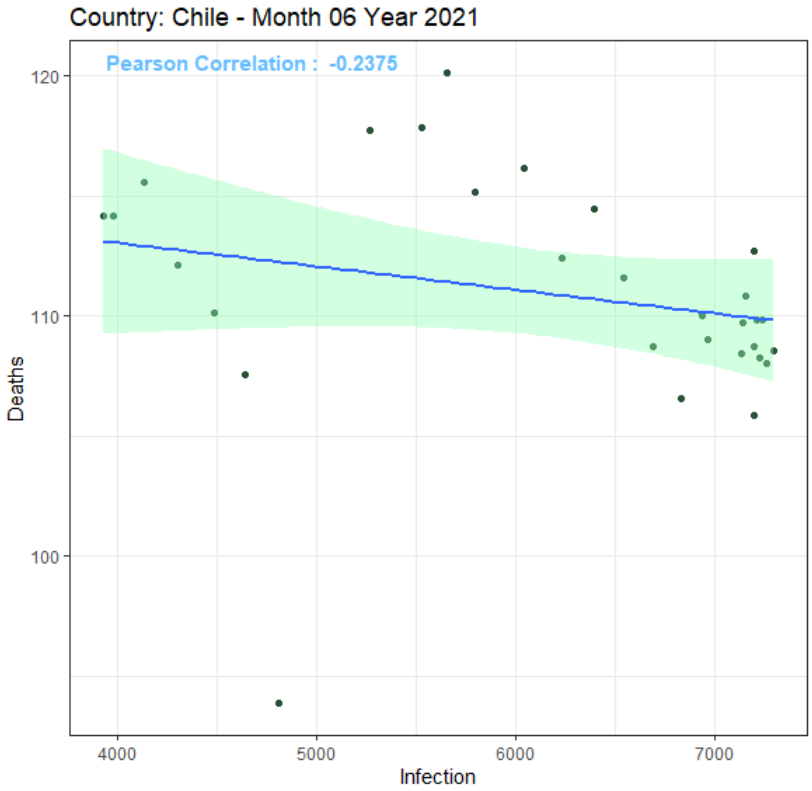

Country: Venezuela - Month 02 Year 2021
Pearson Correlation : 0.2123
Deaths
6
5
4
400
425
450
475
Infection

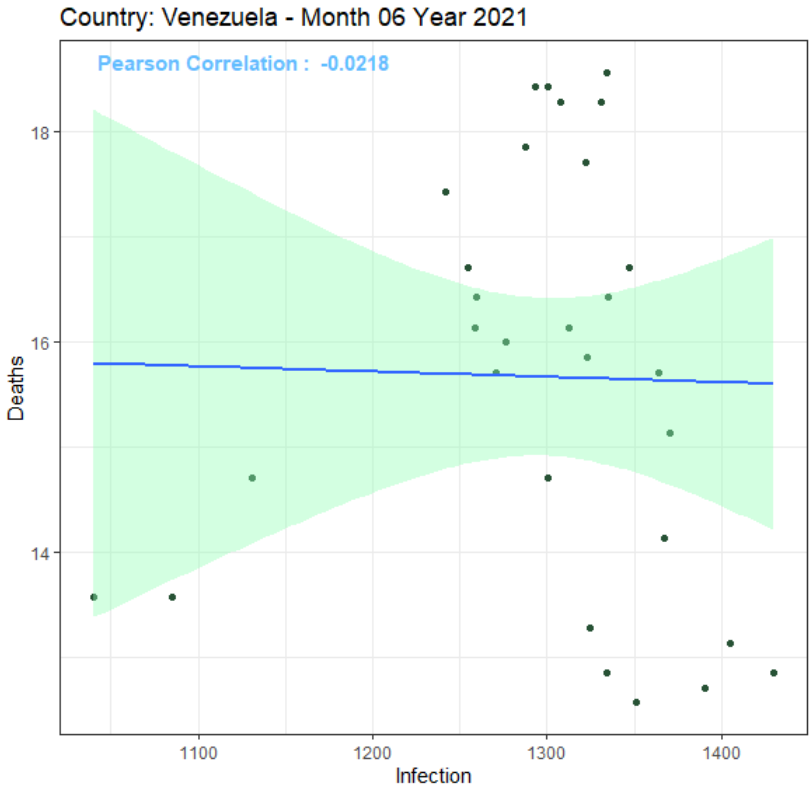

Country: Brazil - Month 03 Year 2021
Pearson Correlation : 0.8858
Deaths
Infection
3000
2500
2000
1500
1000
55000
60000
65000
70000
75000

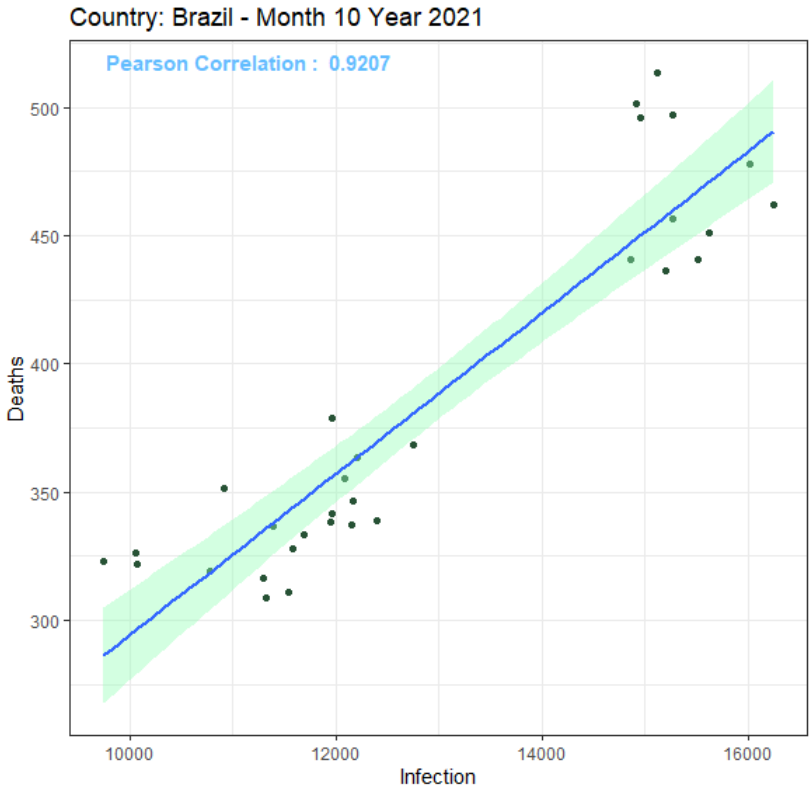

Country: Chile - Month 03 Year 2021
Pearson Correlation : 0.498
Deaths
Infection
100
90
80
70
4000
5000
6000

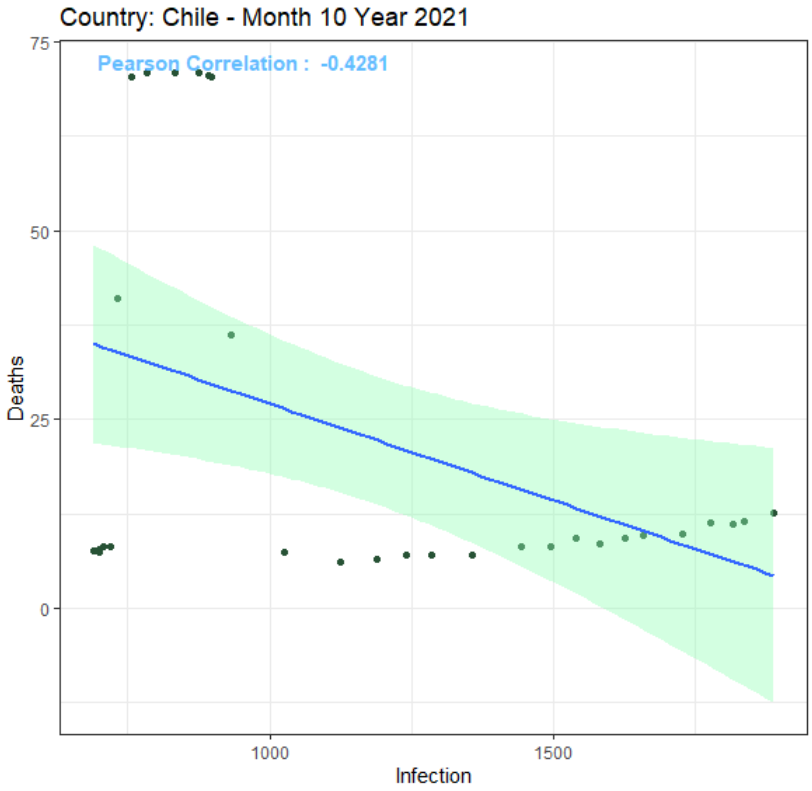

Country: Venezuela - Month 03 Year 2021
Pearson Correlation : 0.907
Deaths
Infection
12.5
10.0
7.5
5.0
400
600
800
1000

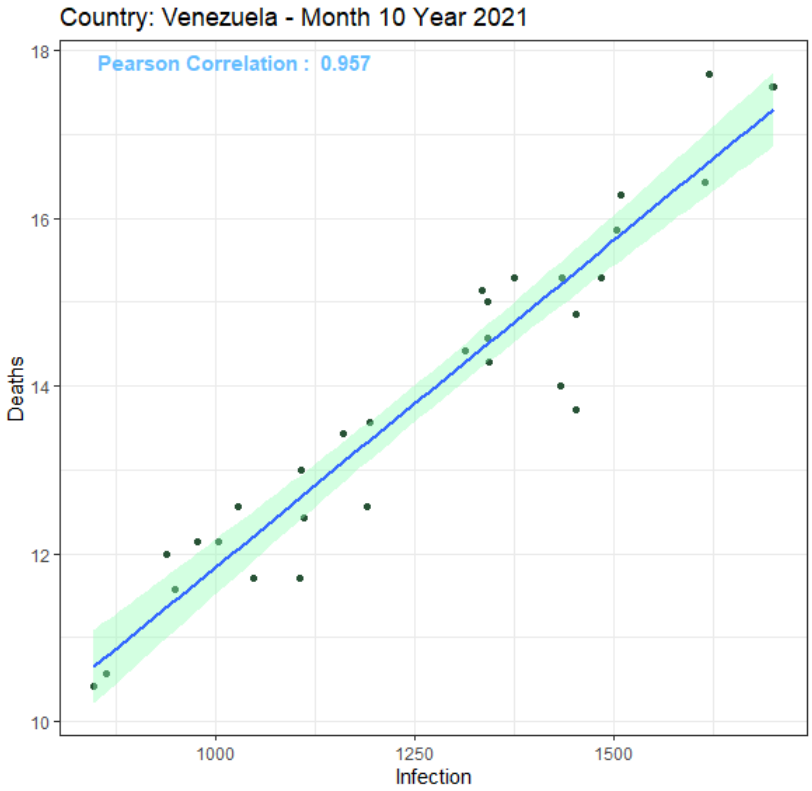

Country: Brazil - Month 06 Year 2021
Pearson Correlation : 0.6198
Deaths
Infection
2100
2000
1900
1800
1700
1600
55000
60000
65000
70000
75000

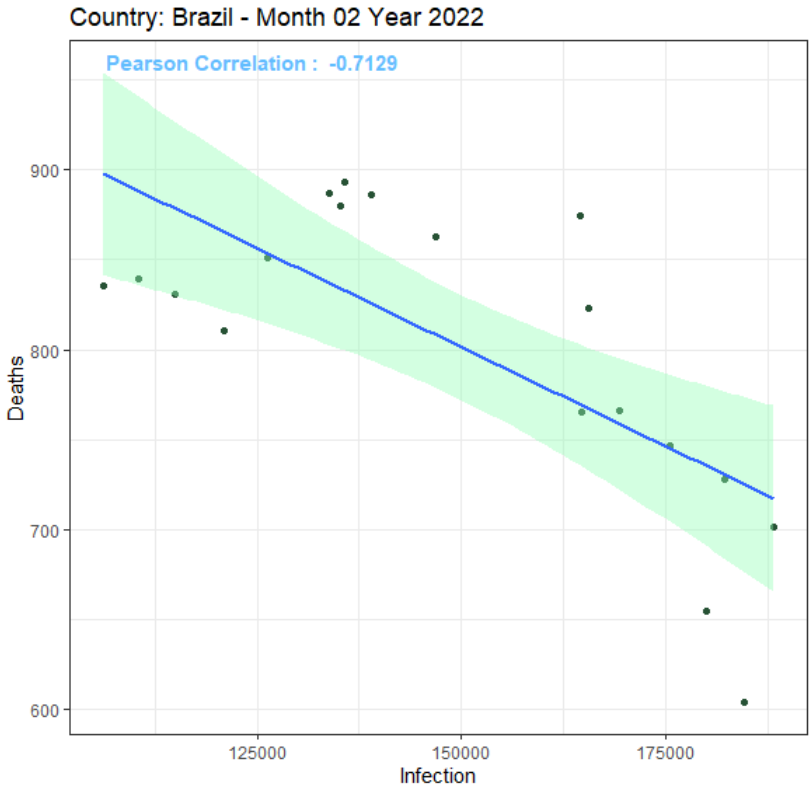

Country: Chile - Month 06 Year 2021
Pearson Correlation : -0.2375
120
110
100
Deaths
4000
5000
6000
7000
Infection

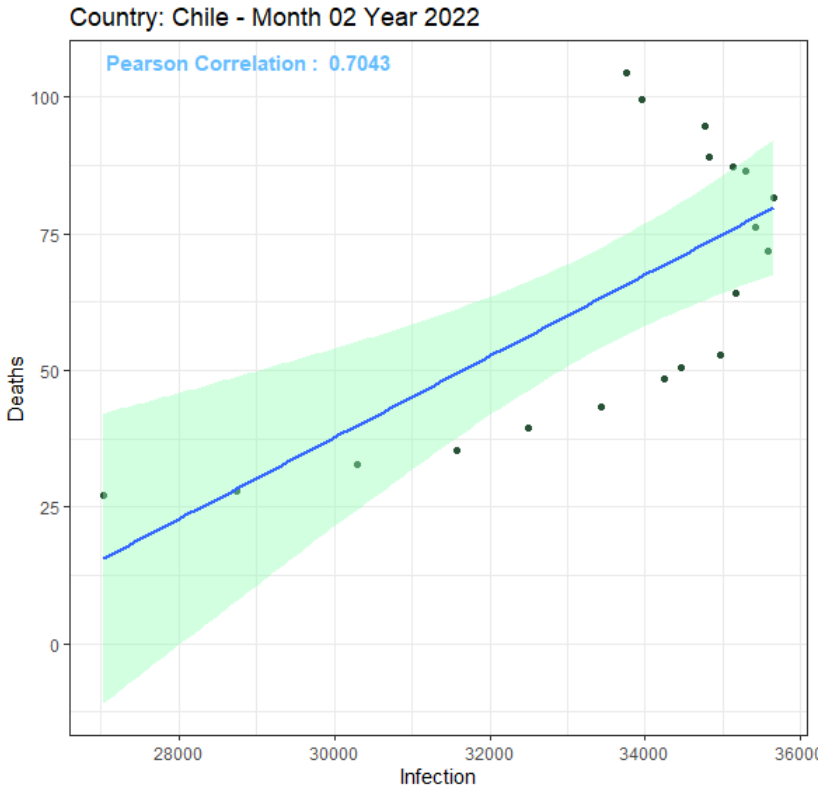

Country: Venezuela - Month 06 Year 2021
Pearson Correlation : -0.0218
18
16
14
Deaths
1100
1200
1300
1400
Infection

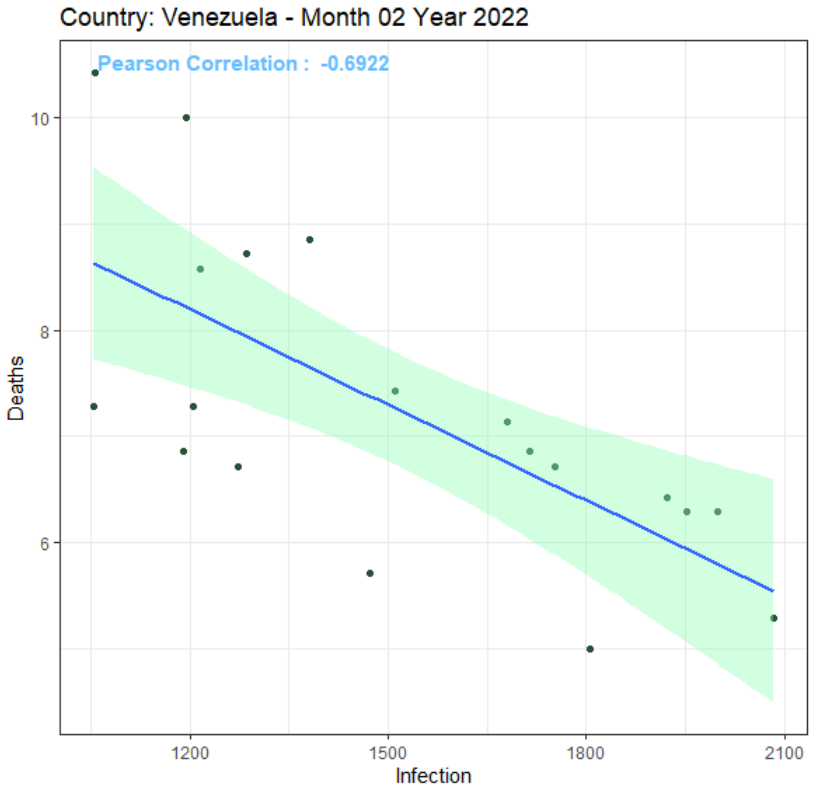

Country: Brazil - Month 10 Year 2021
Pearson Correlation : 0.9207
Deaths
Infection
500
450
400
350
300
10000
12000
14000
16000


Country: Chile - Month 10 Year 2021
Pearson Correlation : -0.4281
Deaths
Infection
75
50
25
0
1000
1500


Country: Venezuela - Month 10 Year 2021
Pearson Correlation : 0.957
Deaths
Infection
18
16
14
12
10
1000
1250
1500

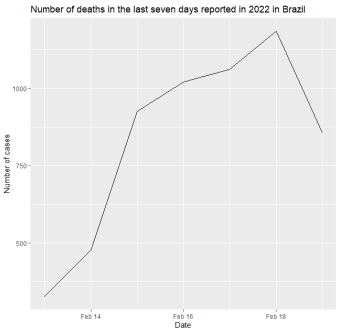

Country: Brazil - Month 02 Year 2022
Pearson Correlation : -0.7129
Deaths
Infection
900
800
700
600
125000
150000
175000

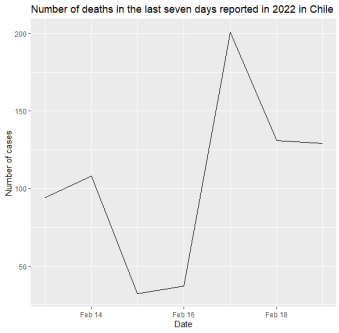


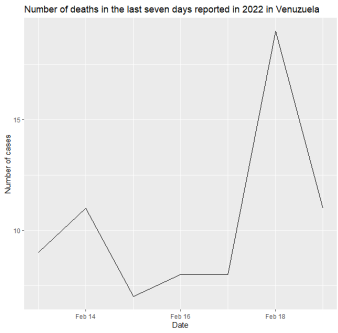


x) Nhóm câu hỏi riêng

2) So sánh tình trạng tử vong của các quốc gia trong 7 ngày cuối của năm cuối cùng
Sử dụng biểu đồ ở câu iv phần 4







- So sánh tình trạng tử vong:
  Nhìn chung, tình trạng tử vong của cả ba nước đều đã đỉnh rồi bắt đầu giảm mạnh vào những ngày cuối. Trong đó, Brazil là nước có số ca tử vong lớn nhất, và rất lớn so với 2 nước còn lại, ngược lại số ca tử vong ở Venezuela trong 7 ngày cuối được cung cấp số liệu của năm 2022 là thấp nhất.

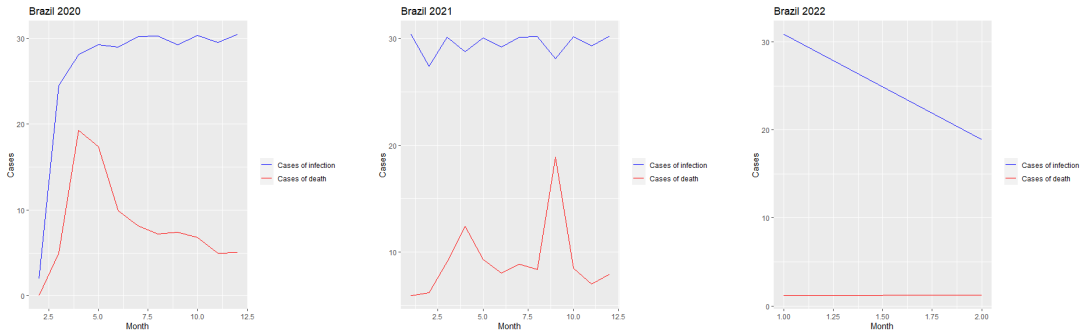
  Ở Brazil, vào 7 ngày cuối được thu thập dữ liệu, số ca tử vong có xung hướng tăng mạnh đến khi đạt đỉnh với số ca khoảng 1180 vào ngày 18/2 thì giảm mạnh vào ngày cuối -19/2. Biểu đồ củaChile và Venezuela chia sẻ xu thế giống nhau khi mà số ca tử vong dao động mạnh trước khi tăng nhanh đến điểm trong nửa sau của thời gian thực hiện khảo sát bằng biểu đồ. Sau khi chạm đỉnh với khoảng 200 ca và 19 ca ở lần lượt Chile và Venezuela, thì số ca tử vong bắt đầu giảm.

3) Cho biết các khoảng thời gian nào mà tỉ lệ tử vong tích lũy giảm mạnh nhưng tỉ lệ nhiễm bệnh tích lũy tăng mạnh hoặc ngược lại cho các quốc gia.

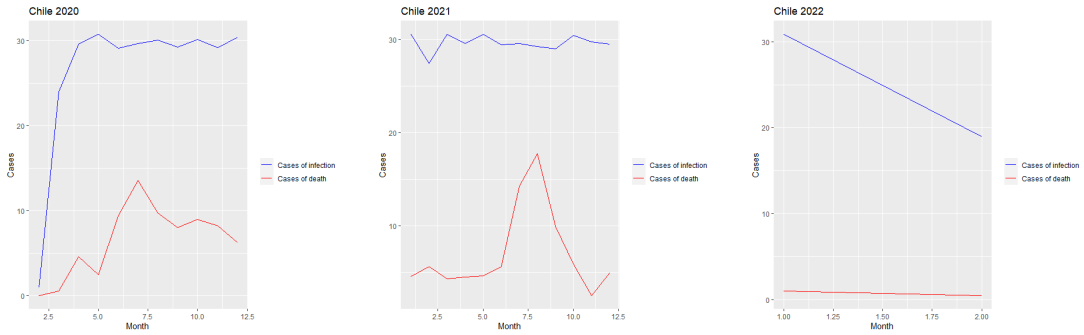
- Đối với Brazil

```
library(tidyverse)
library(datasets)
library(dplyr)
library("ggplot2")
library(lubridate)
covid = read.csv("C:/Users/Admin/Downloads/covidData.csv")
covid$date = as.Date(covid$date, format = "%m/%d/%Y")
covid = covid %>% filter(!continent =='') %>% mutate(new_cases = abs(new_cases), new_deaths=
    abs(new_deaths))
covid[is.na(covid)]=0
brazil = subset(covid, iso_code == "BRA")
chile = subset(covid, iso_code == "CHL")
venezuela = subset(covid, iso_code == "VEN")


#Doi voi brazil
brazil = brazil %>% mutate( totalcases = new_cases + new_deaths) %>% filter(!totalcases == 0)
brazil = brazil %>% mutate(new_cases = new_cases/totalcases , new_deaths = new_deaths/
    totalcases)

brazil = brazil %>% mutate(month = as.numeric(format(as.Date(brazil$date),"%m")), year =
    as.numeric(format(as.Date(brazil$date),"%Y")))
bra =  aggregate(cbind(brazil$new_cases, brazil$new_deaths), by = list(brazil$month, brazil$
    year), sum)
bra$V2 = bra$V2 * 10
ggplot(data=bra%>%filter(Group.2 == 2020), aes(x= Group.1))  + geom_line(aes(y= V1,color= "
    Cases of infection")) +
  geom_line(aes(y= V2,color = "Cases of death")) + xlab("Month") + ylab("Cases") + ggtitle("
    Brazil 2020") +
  scale_colour_manual("",
                      breaks = c("Cases of infection", "Cases of death"),
                      values = c("blue", "red"))

ggplot(data=bra%>%filter(Group.2 == 2021), aes(x= Group.1))  + geom_line(aes(y= V1,color= "
    Cases of infection")) +
  geom_line(aes(y= V2,color = "Cases of death")) + xlab("Month") + ylab("Cases") + ggtitle("
    Brazil 2021") +
  scale_colour_manual("",
```

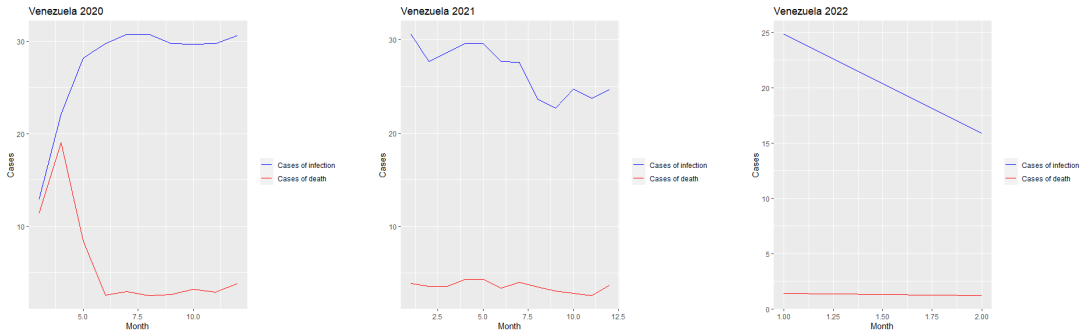
```
                      breaks = c("Cases of infection", "Cases of death"),
                      values = c("blue", "red"))
ggplot(data=bra%>%filter(Group.2 == 2022), aes(x= Group.1))  + geom_line(aes(y= V1,color= "
    Cases of infection")) +
  geom_line(aes(y= V2,color = "Cases of death")) + xlab("Month") + ylab("Cases") + ggtitle("
    Brazil 2022") +
  scale_colour_manual("",
                      breaks = c("Cases of infection", "Cases of death"),
                      values = c("red", "green"))

```



- Nhận xét từ đồ thị của Brazil:
  Khoảng thời gian nào mà tỉ lệ tử vong tích lũy giảm mạnh nhưng tỉ lệ nhiễm bệnh tích lũy tăng mạnh ở Brazil: Vào khoảng tháng 4 năm 2020 đến hết năm 2020 và khoảng tháng 9 đến tháng 10 năm 2021.
  Khoảng thời gian nào mà tỉ lệ nhiễm bệnh tích lũy giảm mạnh nhưng tỉ lệ tử vong tích lũy tăng mạnh ở Brazil: Vào khoảng tháng 3 đến tháng 4 và khoảng tháng 8 đến tháng 9 năm 2021.
  Thực hiện đoạn code tương tự với Chile và Venezuela
- Đối với Chile



- Nhận xét từ đồ thị của Chile:
  Khoảng thời gian nào mà tỉ lệ tử vong tích lũy giảm mạnh nhưng tỉ lệ nhiễm bệnh tích lũy tăng mạnh ở Chile: tháng 4 đến tháng 5 và tháng 7 đến tháng 9 năm 2020. Ngoài ra tỉ lệ tử vong tích lũy giảm mạnh nhất vào tháng 8 đến tháng 11 năm 2021.
  Khoảng thời gian nào mà tỉ lệ nhiễm bệnh tích lũy giảm mạnh nhưng tỉ lệ tử vong tích lũy tăng mạnh ở Chile: tháng 5 đến tháng 6 năm 2020 và tháng 1 đến tháng 2 năm 2021.
- Đối với Venezuela



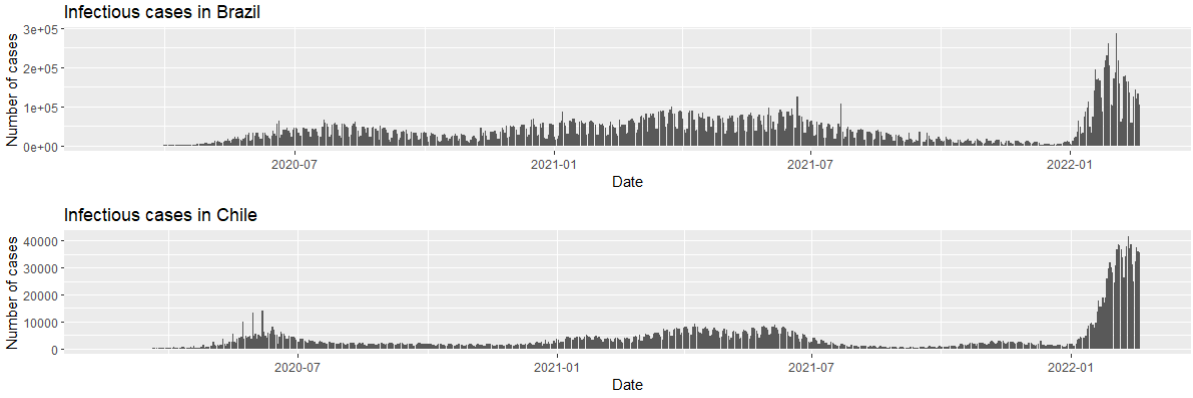
- Nhận xét từ đồ thị của Venezuela:
  Khoảng thời gian nào mà tỉ lệ tử vong tích lũy giảm mạnh nhưng tỉ lệ nhiễm bệnh tích lũy tăng mạnh ở Venezuela: tháng 4 đến tháng 6 năm 2020.
  Khoảng thời gian nào mà tỉ lệ nhiễm bệnh tích lũy giảm mạnh nhưng tỉ lệ tử vong tích lũy tăng mạnh ở Venezuela: tỉ lệ nhiễm bệnh tích lũy có xu hướng giảm kể tử lần tăng đạt đỉnh vào giữa tháng 7 năm 2020 trong khi tỉ lệ tử vong tích lũy lại dao động không đáng kể t

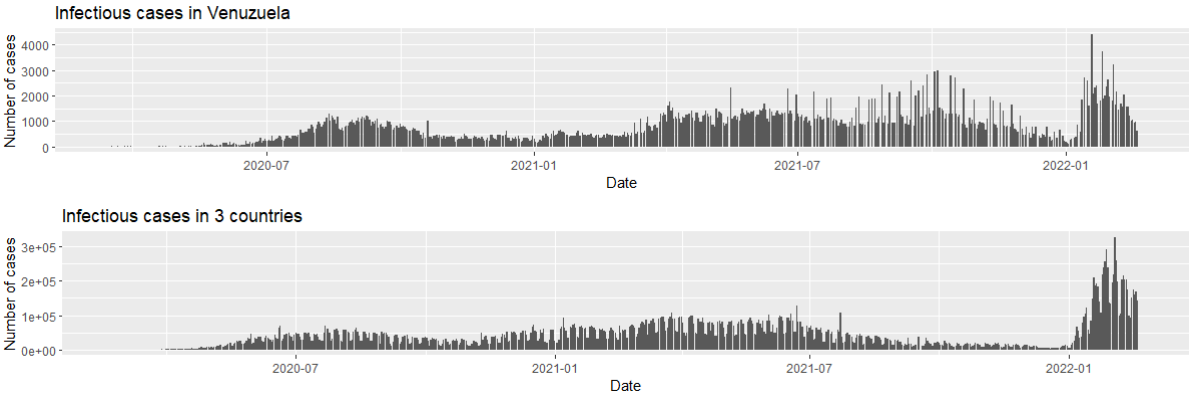
6) Khoảng thời bùng phát nhiễm bệnh lớn nhất giữa các quốc gia có chồng lên nhau không, Cho biết khoảng thời gian giao nhau đó?

- Hiện thực trong R:

```r
brazil = subset(covid, iso_code == "BRA")
bra = brazil[order(brazil$date),]
ggplot(data = bra, aes(x = date, y = new_cases)) + geom_col() + labs(title = "Infectious cases
      in Brazil", x = "Date", y = "Number of cases")

chile = subset(covid, iso_code == "CHL")
chi = chile[order(chile$date),]
ggplot(data = chi, aes(x = date, y = new_cases)) + geom_col() + labs(title = "Infectious cases
      in Chile", x = "Date", y = "Number of cases")

venezuela = subset(covid, iso_code == "VEN")
ven = venezuela[order(venezuela$date),]
ggplot(data= ven, aes(x= date, y = new_cases)) + geom_col() + labs(title = "Infectious cases
     in Venuzuela", x = "Date", y = "Number of cases")

total <- rbind(brazil, venezuela, chile)
tot = total[order(total$date),]
ggplot(data= tot, aes(x= date, y = new_cases)) + geom_col() + labs(title = "Infectious cases
     in 3 countries", x = "Date", y = "Number of cases")
}
```

- Kết quả





- Nhận xét: Dựa vào các biểu đồ, khoảng thời gian bùng phát nhiễm bệnh lớn nhất ở cả 3 quốc gia Brazil, Chile, Venezuela có sự giao nhau vào khoảng tháng 2 năm 2022.

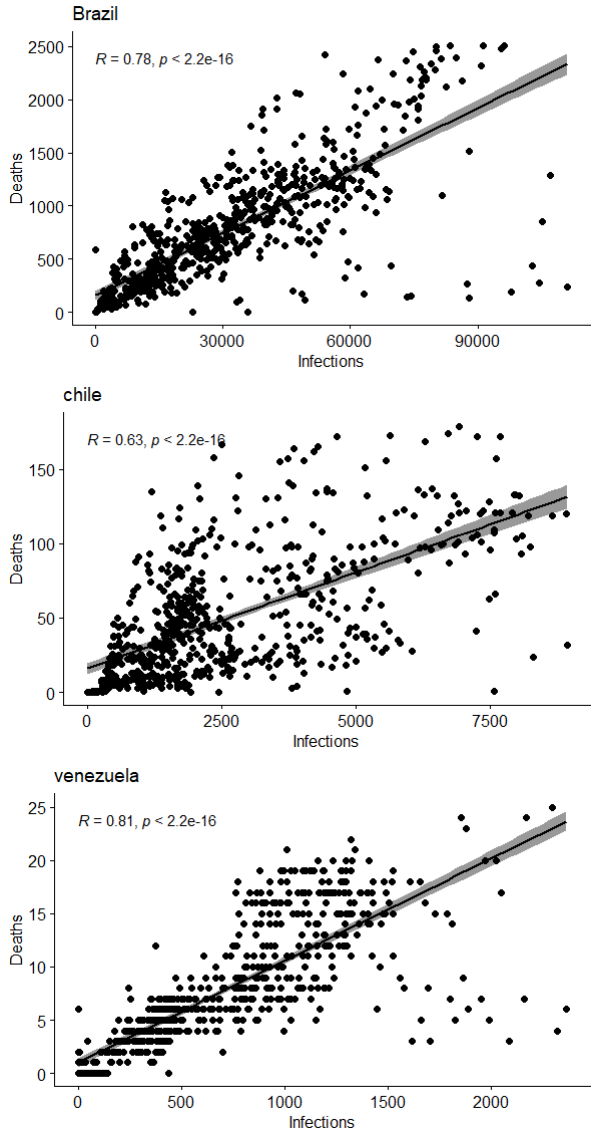
10) Hãy mô tả mối quan hệ tuyến tính giữa nhiễm bệnh và tử vong bằng cách đo độ kết hợp của mối quan hệ dùng correlation r (correlation coefficient) và hướng kết hợp.

- Hiện thực trong R:

```
library(tidyverse)
library(dplyr)
library(datasets)
library("ggplot2")
library("ggpubr")
setwd("E:/BTL_CTRR")
covid = read.csv("owid-covid-data.csv", header = TRUE)
covid = covid %>% filter(!continent =='') %>% mutate(new_cases = abs(new_cases), new_deaths=
    abs(new_deaths))
brazil = subset(covid, iso_code == "BRA")
chile = subset(covid, iso_code == "CHL")
venezuela = subset(covid, iso_code == "VEN")
brazil[is.na(brazil)]=0
chile[is.na(chile)]=0
venezuela[is.na(venezuela)]=0

Countries = c("Brazil", "Chile", "Venezuela")

#remove outliers from new cases col
q1 = quantile(brazil$new_cases, 0.25)
q3 = quantile(brazil$new_cases, 0.75)
iqr =IQR(brazil$new_cases)
brazil = brazil %>% subset(brazil$new_cases > (q1 - 1.5*iqr) & brazil$new_cases< (q3 + 1.5*iqr
    ) )

q1 = quantile(brazil$new_deaths, 0.25)
q3 = quantile(brazil$new_deaths, 0.75)
iqr =IQR(brazil$new_deaths)
brazil = brazil %>% subset(brazil$new_deaths > (q1 - 1.5*iqr) & brazil$new_deaths< (q3 + 1.5*
    iqr) )


ggscatter(brazil, x = "new_cases", y = "new_deaths",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "Infections", ylab = "Deaths", title = "brazil")

q1 = quantile(chile$new_cases, 0.25)
q3 = quantile(chile$new_cases, 0.75)
iqr =IQR(chile$new_cases)
chile = chile %>% subset(chile$new_cases > (q1 - 1.5*iqr) & chile$new_cases< (q3 + 1.5*iqr) )

q1 = quantile(chile$new_deaths, 0.25)
q3 = quantile(chile$new_deaths, 0.75)
iqr =IQR(chile$new_deaths)
chile = chile %>% subset(chile$new_deaths > (q1 - 1.5*iqr) & chile$new_deaths< (q3 + 1.5*iqr)
    )

```

```
ggscatter(chile, x = "new_cases", y = "new_deaths",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "Infections", ylab = "Deaths", title = "chile")

#remove outliers from new cases col
q1 = quantile(venezuela$new_cases, 0.25)
q3 = quantile(venezuela$new_cases, 0.75)
iqr =IQR(venezuela$new_cases)
venezuela = venezuela %>% subset(venezuela$new_cases > (q1 - 1.5*iqr) & venezuela$new_cases< (
    q3 + 1.5*iqr) )

q1 = quantile(venezuela$new_deaths, 0.25)
q3 = quantile(venezuela$new_deaths, 0.75)
iqr =IQR(venezuela$new_deaths)
venezuela = venezuela %>% subset(venezuela$new_deaths > (q1 - 1.5*iqr) & venezuela$new_deaths<
     (q3 + 1.5*iqr) )

ggscatter(venezuela, x = "new_cases", y = "new_deaths",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "Infections", ylab = "Deaths", title = "venezuela")
}
```

- Kết quả



- Nhận xét:
  - Đối với cả 3 biểu đồ tương ứng với 3 quốc gia Brazil, Chile, Venezuela, hệ số tương quan Pearson(R) đều dương (>0), nghĩa là số ca nhiễm(infections) và số ca tử vong(deaths) ở 3 quốc gia có mối tương quan tuyến tính với nhau. Nói cách khác, số ca nhiễm(infections) tăng thì số ca tử vong(deaths) cũng tăng.
  - Ta thấy, hệ số tương quan Pearson(R) ứng với Venezuela là 0.81 lớn nhất trong 3 nước nên mối tương quan tuyến tính giữa số ca nhiễm và số ca tử vong được thu thập tại Venezuela là chặt chẽ nhất.
  - Ngược lại, hệ số tương quan Pearson(R) ứng với Chile là 0.63 bé nhất nên mối tương quan tuyến tính giữa số ca nhiễm và số ca tử vong được thu thập tại đây ít chặt chẽ hơn 2 nước còn lại.

Tài liệu

[Dal] Dalgaard, P. *Introductory Statistics with R.* Springer 2008.

[K-Z] Kenett, R. S. and Zacks, S. *Modern Industrial Statistics: with applications in R, MINITAB and JMP,* 2nd ed., John Wiley and Sons, 2014.

[Ker] Kerns, G. J. *Introduction to Probability and Statistics Using R,* 2nd ed., CRC 2015.

[4] CSETI. *Thống kê mô tả trong nghiên cứu - Các đại lượng về độ phân tán.* <http://thongke.cesti.gov.vn/dich-vu-thong-ke/tai-lieu-phan-tich-thong-ke/845-thong-ke-mo-ta-trong-nghien-cuu-dai-luong-do-phan-tan>.

[5] Sololearn. *R course,* <https://www.sololearn.com/learning/1147>.

[6] Đoàn Quỳnh, Nguyễn Huy Đoan, Nguyễn Xuân Liêm, Đặng Hùng Thắng, Trần Văn Vuông. *Đại Số 10 nâng cao,* tái bản lần thứ 14. NXB Giáo dục Việt Nam 2020.

[7] Luanvan2S. *Lý thuyết về hệ số tương quan Pearson - Phân tích tương quan trong SPSS,* <https://luanvan2s.com/he-so-tuong-quan-pearson-trong-spss-bid61.html>.