



BÁO CÁO BÀI TẬP LỚN

# Thống kê khảo sát kết quả Covid-19

Môn: Cấu trúc rời rạc cho KHMT (CO1007)

*Thành phố Hồ Chí Minh, ngày 24 tháng 04 năm 2022*

Huynh Tuong Nguyen, Nguyen Ngoc Le  
Faculty of Computer Science and Engineering  
University of Technology - VNUHCM  
[htnguyen@hcmut.edu.vn](mailto:htnguyen@hcmut.edu.vn)



Động cơ nghiên cứu

Mục tiêu

Kiến thức chuẩn bị

Nhiệm vụ

Kết luận

## Nhóm SV thực hiện

Nguyễn Thái Tân – 2112256  
Lê Nguyên Chương – 2112945  
Trương Hoàng Nhật – 2114303  
Nguyễn Ngọc Khánh My – 2114094  
Trần Minh Thuận – 2114939  
Nguyễn Danh Thành – 2114782



Động cơ nghiên cứu

Mục tiêu

Kiến thức chuẩn bị

Nhiệm vụ

Kết luận

① Động cơ nghiên cứu

② Mục tiêu

③ Kiến thức chuẩn bị

④ Nhiệm vụ

⑤ Kết luận

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 thế 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 rãi trong đời sống hay trên internet để giúp cho mọi người quan sát, phân tích, nghiên cứu được cập nhật hàng ngày.

Phân tích và 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 mới 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 khảo 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.



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.





## Độ lệch chuẩn

là một đại lượng thống kê mô tả dùng để đo mức độ phân tán của một tập dữ liệu đã được lập thành bảng tần số. Được tính bằng cách lấy căn bậc hai của phương sai.

$$s = \sqrt{\frac{\sum_{i=0}^n (X_i - \bar{X})^2}{n - 1}}$$

- $s$ : là độ lệch chuẩn
- $\bar{X}$ : là giá trị trung bình của mẫu
- $X_i$ : là thành phần thứ  $i$  của mẫu
- $n$ : là số thành phần của mẫu

**Hiện thực trên R:** `sd(na.omit(x))`

- $x$ : một vector số hoặc một đối tượng R nhưng không phải factor được ép kiểu thành số bởi `as.double`.
- `na.omit()`: loại bỏ các giá trị NA.



## Tứ phân vị

là đại lượng mô tả sự phân bố và sự phân tán của tập dữ liệu.

Trong đó:

- Giá trị tứ phân vị thứ nhất Q1 bằng trung vị phần dưới
- Giá trị tứ phân vị thứ hai Q2 chính bằng giá trị trung vị
- Giá trị tứ phân vị thứ ba Q3 bằng trung vị phần trên

**Hiện thực trên R:** `quantile(na.omit(x))`

Trong đó:

- $x$ : một vector đầu vào.
- `na.omit()`: loại bỏ các giá trị NA.



## Tương quan

Hệ số tương quan là chỉ số thống kê đo lường mức độ mạnh yếu của mối quan hệ giữa hai biến số. Trong đó, hệ số tương quan có giá trị từ -1.0 đến 1.0.

- Hệ số tương quan có giá trị âm cho thấy hai biến có mối quan hệ nghịch biến hoặc tương quan âm (nghịch biến tuyệt đối khi giá trị bằng -1)
- Hệ số tương quan có giá trị dương cho thấy mối quan hệ đồng biến hoặc tương quan dương (đồng biến tuyệt đối khi giá trị bằng 1)
- Tương quan bằng 0 cho hai biến độc lập với nhau.





## Tương quan

Có nhiều loại hệ số tương quan, nhưng trong bài tập lớn này, ta xét đến loại tương quan Pearson.

$$\rho_{xy} = \frac{\text{cov}(x, y)}{s_x s_y}$$

Trong đó:

- $\rho_{xy}$ : hệ số tương quan Pearson
- $\text{cov}(x, y)$ : hiệp phương sai của biến  $x$  và  $y$
- $s_x, s_y$ : độ lệch chuẩn của  $x, y$

**Hiện thực trên R:**  $\text{cor}(x, y)$

## Các bước ban đầu

### MADE

Nhóm có  $MD$  là 4315 nên  $kq = (4 + 3 + 1 + 5) \% 6 = 1$ . Vậy nhóm sẽ xử lý 3 quốc gia *Indonesia, Japan, Vietnam*.

### Các packages

Sau khi đọc sơ lược qua các nhiệm vụ, nhóm nhận thấy rằng để giải quyết các nhiệm vụ một cách thuận tiện hơn, thì chúng ta nên thêm các packages như sau vào R.

- `library(readr)`: cung cấp nhiều hàm để đọc dữ liệu từ các tập tin csv.
- `library(stringr)`: cung cấp các hàm trong việc xử lý text.
- `library(ggplot2)`: một package rất mạnh trong việc vẽ biểu đồ, bản đồ với nhiều tùy biến.
- `library(lubridate)`: hỗ trợ thao tác với dữ liệu thời gian (ngày, tháng, năm, giờ,...)
- `library(here)`: hỗ trợ tìm working directory cho R



# Các bước ban đầu



## Các packages

- `library(scales)`: cung cấp các thao tác tùy chỉnh đồ thị
- `library(dplyr)`: cung cấp khả năng thao tác với dữ liệu một cách dễ dàng hơn, với các tính năng và hàm bổ sung.
- `library(zoo)`: hỗ trợ tạo định dạng time-series trên R.

## Đọc file

Nhóm sẽ đọc file dữ liệu vào một biến đặt tên là `dataFile`. Ta cũng chuyển các dữ liệu âm thành dương ở bước đầu tiên này.

```
dataFile <- read_csv("owid-covid-data.csv", show_col_
  types = FALSE)
```

```
dataFile$new_cases <- abs(dataFile$new_cases)
dataFile$new_deaths <- abs(dataFile$new_deaths)
```

## Nhiệm vụ i: Nhóm câu hỏi liên quan đến tổng quát dữ liệu

- ❶ Tập mẫu thu thập dữ liệu vào các năm nào  
Từ cột *date* ở bảng dữ liệu, chúng ta lấy dữ liệu năm làm gốc phân loại.

```
i1<-function()  
{  
  time <- dataFile %>% select(date)  
  mdy <- strptime(time$date,format="%m/%d/%Y")  
  year <- format(mdy,"%Y")  
  cat("Tap mau du lieu thu duoc vao cac nam: ",  
      unique(year))  
}  
i1()
```

Kết quả

```
> i1()  
Tap mau du lieu thu duoc vao cac nam: 2020 2021  
2022
```



## Nhiệm vụ i: Nhóm câu hỏi liên quan đến tổng quát dữ liệu

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

Thông qua *dataframe*, chúng ta đếm số lượng đất nước qua *isocode* và liệt kê ra 10 nước đầu tiên.

```
i2<-function()  
{  
  isoCode<-dataFile$iso_code  
  cnames<-dataFile$location  
  conn<-dataFile$continent  
  i2.1<-data.frame(isoCode, cnames, conn,  
    stringsAsFactors=FALSE)  
  i2.2<-subset(i2.1, i2.1$conn!="")  
  a<-unique(i2.2)  
  index<-dim(a)[1]  
  data1<-a[1:10,c(1,2)]  
  colnames(data1)<-c("iso_code:", "Country")  
  rownames(data1)<-c("1", "2", "3", "4", "5", "6", "7", "  
    8", "9", "10")  
  prmatrix(data1, left = TRUE, quote = FALSE)  
  cat("Count: ", index)  
}
```

i2()



## Nhiệm vụ i: Nhóm câu hỏi liên quan đến tổng quát dữ liệu

### Kết quả

```
iso_code: Country
1 AFG Afghanistan
2 ALB Albania
3 DZA Algeria
4 AND Andorra
5 AGO Angola
6 AIA Anguilla
7 ATG Antigua and Barbuda
8 ARG Argentina
9 ARM Armenia
10 ABW Aruba
Count: 225
```

**Hình:** Số lượng đất nước và định danh của mỗi đất nước



## Nhiệm vụ i: Nhóm câu hỏi liên quan đến tổng quát dữ liệu

### ③ Số lượng châu lục trong tập mẫu

Chúng ta lấy dữ liệu từ cột *continent* từ tệp gốc rồi loại bỏ những dữ liệu trống, đồng thời sắp xếp, đếm, phiên dịch cùng liệt kê dữ liệu trong đó.

```
i3<-function()  
{  
  Con <- dataFile %>% select(continent)  
  Con<-unique(Con)  
  Con <- Con[Con != ""]  
  Con<-sort(Con, decreasing = FALSE)  
  Trans<-c("Chau Phi", "Chau A", "Chau Au", "Nam  
    My", "Chau Dai Duong", "Bac My")  
  m<-data.frame(unlist(Con),unlist(Trans),  
    stringsAsFactors = FALSE)  
  colnames(m)<-c("Continent:", "6")  
  rownames(m)<-c("1", "2", "3", "4", "5", "6")  
  prmatrix(m, left = TRUE, quote = FALSE)  
}  
i3()
```



## Nhiệm vụ i: Nhóm câu hỏi liên quan đến tổng quát dữ liệu

### Kết quả

Continent:	6
1 Africa	Châu Phi
2 Asia	Chau A
3 Europe	Chau Au
4 North America	Nam My
5 Oceania	Chau Dai Duong
6 South America	Bac My

**Hình:** *Số lượng châu lục trong tập mẫu*





## Nhiệm vụ i: Nhóm câu hỏi liên quan đến tổng quát dữ liệu

- ④ Số lượng dữ liệu thể hiện thu thập dữ liệu được trong từng từng châu lục và tổng số

Chúng ta nhóm dữ liệu theo *continent*, sau đó đếm số lượng dòng có trong *continent*, số dòng bằng số dữ liệu thu thập được. Từ đó ta tổng hợp thành một dataframe mới.

```
i4<-function()  
{  
  formattedData <- dataFile %>% filter(nchar(as.  
    character(continent))>0)  
  table <- formattedData %>% group_by(continent)  
    %>% summarise(observation = length(continent  
    ))  
  ti4<-sum(table$observation)  
  a<-c("Tong:",ti4)  
  table<-rbind(table,a)  
  rownames(table)<-c("1","2","3","4","5","6","7")  
  prmatrix(table, left = TRUE, quote = FALSE)  
  return(table)  
}  
table<-i4()
```



## Nhiệm vụ i: Nhóm câu hỏi liên quan đến tổng quát dữ liệu

### Kết quả

	continent	observation
1	Africa	38647
2	Asia	35528
3	Europe	36375
4	North America	24438
5	Oceania	8993
6	South America	9335
7	Tong:	153316

**Hình:** *Số lượng dữ liệu thu thập được trong từng từng châu lục và tổng số*



## Nhiệm vụ i: Nhóm câu hỏi liên quan đến tổng quát dữ liệu

- ⑤ Số lượng dữ liệu thể hiện thu thập dữ liệu được trong từng từng đất nước (hiển thị 10 đất nước cuối cùng) và tổng số Tương tự như câu 4 nhưng chúng ta lấy dữ liệu theo *location* và xuất ra 10 đất nước (theo isocode) cuối cùng. Kết quả

	iso_code	observation
1	VEN	708
2	VGB	694
3	VNM	759
4	VUT	467
5	WLF	489
6	WSM	459
7	YEM	681
8	ZAF	744
9	ZMB	704
10	ZWE	702
11	Tong:	153316

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



## Nhiệm vụ i: Nhóm câu hỏi liên quan đến tổng quát dữ liệu

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

Từ bảng dữ liệu ở câu 4, ta sử dụng hàm *min()* để tìm giá trị nhỏ nhất giữa các châu lục

```
i6<-function(table)
{
  mini6<-min(as.numeric(table$observation))
  tmini6 <- table %>% filter(observation == min(as
    .numeric(observation)))
  cat("Chau luc co luong thu thap du lieu nho nhat
    la",tmini6$continent,"va gia tri nho nhat
    do la",tmini6$observation)
}
i6(table)
```

Kết quả

```
> i6()
Chau luc co luong thu thap du lieu nho nhat la
  Oceania va gia tri nho nhat do la 8993
```



## Nhiệm vụ i: Nhóm câu hỏi liên quan đến tổng quát dữ liệu

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

Từ bảng dữ liệu ở câu 4, ta sử dụng hàm `max()` để tìm giá trị lớn nhất giữa các châu lục

```
i7<-function(table)
{
  cuttable<-table[1:6,c(1,2)]
  maxi7<-max(as.numeric(cuttable$observation))
  tmaxi7 <- cuttable %>% filter(observation == max
    (as.numeric(observation)))
  cat("Chau luc co luong thu thap du lieu lon nhat
    la",tmaxi7$continent,"va gia tri lon nhat
    do la",tmaxi7$observation)
}
i7(table)
```

Kết quả

```
> i7()
Chau luc co luong thu thap du lieu lon nhat la
  Africa va gia tri lon nhat do la 38647
```



## Nhiệm vụ i: Nhóm câu hỏi liên quan đến tổng quát dữ liệu

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

Từ bảng dữ liệu ở câu 5, ta tìm giá trị thu thập dữ liệu nhỏ nhất, sau đó tìm *isocode* tương ứng, đồng thời ta tìm được tên quốc gia tương ứng với giá trị nhỏ nhất đó.

```
i8<-function(i5data)
{
  minData <- min(as.numeric(i5data$observation))
  i8result <- i5data %>% filter (as.numeric(
    observation)==minData)
  colnames(i8result)<-c("iso_code","Country Name",
    "Min observation")
  i8result<-i8result[c(2,3)]
  prmatrix(i8result, left = TRUE, quote = FALSE)
}
i8(i5data)
```

Kết quả

```
> i8(i5data)
      Country Name Min observation
[1,] Pitcairn      85
```



## Nhiệm vụ i: Nhóm câu hỏi liên quan đến tổng quát dữ liệu

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

Tương tự câu 8 nhưng chúng ta tìm giá trị lớn nhất.

Kết quả

```
> i9(i5data)
      Country Name Max observation
[1,] Argentina      781
[2,] Mexico         781
```



## Nhiệm vụ i: Nhóm câu hỏi liên quan đến tổng quát dữ liệu

- 10 Cho biết các date nào có lượng dữ liệu thu thập nhỏ nhất và giá trị nhỏ nhất đó?  
Đầu tiên chúng ta sẽ tạo một bảng số liệu thống kê dữ liệu theo ngày, sau đó dựa vào bảng số liệu trên để tìm ra ngày có lượng dữ liệu thu thập nhỏ nhất.

```
dte <- table(dataFile$date)
dte <- as.data.frame(dte)
dte_min <- dte$Var1[dte$Freq==min(dte$Freq)]
cat("Cac ngay co luong du lieu thu thap nho nhat
    la: ", levels(droplevels(dte_min)), "\n")
cat("Luong du lieu thu thap nho nhat la: ", min(
    dte$Freq), "\n")
```

Kết quả

```
> i10 ()
Cac ngay co luong du lieu thu thap nho nhat la: 1
/23/2020 2/18/2020
Luong du lieu thu thap nho nhat la: 10
```





## Nhiệm vụ i: Nhóm câu hỏi liên quan đến tổng quát dữ liệu

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

Dựa vào bảng số liệu ở câu 10 để tìm ra ngày có lượng dữ liệu thu thập lớn nhất.

```
dte_max <- dte$Var1[dte$Freq==max(dte$Freq)]
cat("Cac ngay co luong du lieu thu thap lon nhat
    la: ", levels(droplevels(dte_max)), "\n")
cat("Luong du lieu thu thap lon nhat la: ", max(
    dte$Freq), "\n")
```

Kết quả

```
> i11 ()
Cac ngay co luong du lieu thu thap lon nhat la: 1
/31/2022
Luong du lieu thu thap lon nhat la: 196
```



## Nhiệm vụ i: Nhóm câu hỏi liên quan đến tổng quát dữ liệu

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

Ta sẽ tổng hợp và đếm số lượng dữ liệu dựa theo ngày và châu lục

```
cont_dte <- table(dataFile$continent, dataFile$
  date)
cont_dte <- as.data.frame(cont_dte,
  stringsAsFactors = FALSE)
cont_dte$continent <- cont_dte$Var1
cont_dte$date <- cont_dte$Var2
cont_dte$Number_of_Data <- cont_dte$Freq
cont_dte$Var1 <- NULL
cont_dte$Var2 <- NULL
cont_dte$Freq <- NULL
cont_dte = subset(cont_dte, cont_dte$continent !=
  " ")
View(cont_dte)
```



## Nhiệm vụ i: Nhóm câu hỏi liên quan đến tổng quát dữ liệu

### Kết quả

continent	date	Number_of_Data
Africa	1/1/2021	39
Asia	1/1/2021	42
Europe	1/1/2021	41
North America	1/1/2021	16
Oceania	1/1/2021	1
South America	1/1/2021	12
Africa	1/1/2022	33
Asia	1/1/2022	42
Europe	1/1/2022	38
North America	1/1/2022	20

**Hình:** Số lượng dữ liệu thu thập được theo ngày và châu lục



## Nhiệm vụ i: Nhóm câu hỏi liên quan đến tổng quát dữ liệu

Thông kê khảo sát kết quả Covid-19

Huynh Tuong Nguyen,  
Nguyen Ngoc Le



Động cơ nghiên cứu

Mục tiêu

Kiến thức chuẩn bị

Nhiệm vụ

Kết luận

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

Dựa vào bảng đã tổng hợp từ câu 12, ta có thể tìm được số lượng dữ liệu lớn nhất theo date và châu lục.

```
cat("So luong du lieu thu thap lon nhat theo  
date va chau luc la: ", max(cont_dte$  
Number_of_Data), "\n")
```

Kết quả

```
> i13 ()  
So luong du lieu thu thap lon nhat theo date va  
chau luc la: 48
```

## Nhiệm vụ i: Nhóm câu hỏi liên quan đến tổng quát dữ liệu

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

Tương tự câu 13, dựa vào bảng số liệu ở câu 12, ta có thể tìm được số lượng dữ liệu nhỏ nhất theo date và châu lục.

```
cat("So luong du lieu thu thap nho nhat theo  
date va chau luc la: ", min(cont_dte$  
Number_of_Data), "\n")
```

Kết quả

```
> i14 ()  
So luong du lieu thu thap nho nhat theo date va  
chau luc la: 0
```



## Nhiệm vụ i: Nhóm câu hỏi liên quan đến tổng quát dữ liệu

- 15 Với một date là k và châu lục t cho trước, hãy cho biết số lượng dữ liệu thể hiện thu thập dữ liệu được. Dựa vào bảng dữ liệu ở câu 12 ta có thể tìm được số lượng dữ liệu thu thập được trong ngày k ở một châu lục t (VD: Asia ngày 1/1/2021)

```
k = readline()
1/1/2021
t = readline()
Asia
val <- cont_dte$Number_of_Data[(cont_dte$date==k)
& (cont_dte$continent==t)]
cat("So luong du lieu thu thap duoc trong ngay",k,
    "o chau luc",t, "la: ", val, "\n")
```

Kết quả

```
> i15 ()
So luong du lieu thu thap duoc trong ngay 1/1/2021
o chau luc Asia la: 42
```



## Nhiệm vụ i: Nhóm câu hỏi liên quan đến tổng quát dữ liệu

- 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. Trước tiên từ dữ liệu ban đầu ta sẽ lọc ra những nước có cùng iso\_code và tìm được số lượng dữ liệu thu thập được của mỗi iso\_code, sau đó ta sẽ sắp xếp chúng theo thứ tự tăng dần về số lượng dữ liệu và lọc ra những iso\_code có lượng dữ liệu bằng nhau

```
loc <- table(dataFile$iso_code)
loc <- as.data.frame(loc, stringsAsFactors = FALSE)
country1 <- loc[order(loc$Freq),]
country <- subset(country1, duplicated(Freq) |
  duplicated(Freq, fromLast = TRUE))
colnames(country) <- c("iso_code", "Num_of_Data")
print(country, row.names = FALSE)
```



## Nhiệm vụ i: Nhóm câu hỏi liên quan đến tổng quát dữ liệu

### Kết quả

iso_code	Num_of_Data
TON	12
WSM	12
MSR	49
TZA	49
WLF	49
CMR	194
GRL	194
LBR	317
MUS	317
BDI	397
ERI	397
GHA	449
TCD	449
GIN	515
MCO	515
ISL	516
LIE	516
LUX	550
MNG	550

**Hình:** *iso\_code của các đất nước có lượng dữ liệu thu thập được bằng nhau*





## Nhiệm vụ i: Nhóm câu hỏi liên quan đến tổng quát dữ liệu

- 17 Liệt kê iso\_code, tên đất nước mà chiều dài iso\_code lớn hơn 3.

Trước tiên chúng ta sẽ tổng hợp danh sách các iso\_code sau đó lọc ra những nước có chiều dài iso\_code lớn hơn 3.

```
i_c <- table(dataFile$iso_code, dataFile$location)
i_c <- as.data.frame(i_c)
i_l <- subset(i_c, i_c$Freq!=0)
i_l <- as.data.frame(i_l, stringsAsFactors = FALSE)
)
i <- subset(i_l, str_length(i_l$Var1)>3)
i <- as.data.frame(i)
i$Freq <- NULL
colnames(i) <- c("iso_code", "location")
cat("iso_code của những dat nuoc co do dai iso_
    code lon hon 3: \n")
print(i, row.names = FALSE)
```



## Nhiệm vụ i: Nhóm câu hỏi liên quan đến tổng quát dữ liệu

### Kết quả

iso_code	location
OWID_AFR	Africa
OWID_ASI	Asia
OWID_EUR	Europe
OWID_EUN	European Union
OWID_HIC	High income
OWID_INT	International
OWID_KOS	Kosovo
OWID_LIC	Low income
OWID_LMC	Lower middle income
OWID_NAM	North America
OWID_OCE	Oceania
OWID_SAM	South America
OWID_UMC	Upper middle income
OWID_WRL	world

**Hình:** Những nước có độ dài iso\_code lớn hơn 3



## Nhiệm vụ ii: Nhóm câu hỏi liên quan đến mô tả thống kê cơ bản dữ liệu

**Xử lý chung:** Trước hết, ta cần trích lọc dữ liệu cần thiết của 3 quốc gia cần xử lý (Indonesia, Japan, Vietnam), để thuận tiện hơn trong khi thực hiện chương trình. Các câu hỏi là tương đương nhau cho mỗi quốc gia, do đó, chúng ta có thể dùng vòng *for* để xử lý.

*Lưu ý: Ở phần trình bày câu ii này, ta quy ước số thứ tự 1, 2, 3 lần lượt tương ứng với Indonesia, Japan, Vietnam.*

```
indoFile <- subset(dataFile, dataFile$location == "
  Indonesia")
japanFile <- subset(dataFile, dataFile$location == "
  Japan")
vietnamFile <- subset(dataFile, dataFile$location == "
  Vietnam")

ii_File <- list(indoFile, japanFile, vietnamFile)
ii_string <- cbind("Indonesia", "Japan", "Vietnam")
```



## Nhiệm vụ ii: Nhóm câu hỏi liên quan đến mô tả thống kê cơ bản dữ liệu

### ❶ Tính giá trị nhỏ nhất, lớn nhất

Chúng ta chỉ cần dùng hàm *min()*, *max()* đơn giản, với lưu ý là phải bỏ qua các giá trị NA.

```
cases_min <- vector(length = 3)

for (i in 1:3) {
  cases_min[i] = min(na.omit(data.frame(ii_File[i])$new
    _cases))
  cat(ii_string[i], "min new cases=", cases_min[i], "\n
    ")
}
```

### Kết quả

```
Indonesia min new cases: 0
Indonesia max new cases: 64718
Indonesia min new deaths: 0
Indonesia max new deaths: 2069

Japan min new cases: 0
Japan max new cases: 104345
Japan min new deaths: 0
Japan max new deaths: 271

vietnam min new cases: 0
vietnam max new cases: 54830
vietnam min new deaths: 0
vietnam max new deaths: 804
```

**Hình:** Giá trị lớn nhất, nhỏ nhất



## Nhiệm vụ ii: Nhóm câu hỏi liên quan đến mô tả thống kê cơ bản dữ liệu

② Tính tứ phân vị thứ nhất(Q1), thứ hai(Q2), thứ ba(Q3)  
Để tính tứ phân vị, chúng ta sử dụng hàm *quantile()*.

```
cases_Q1 <- vector(length = 3)

for (i in 1:3) {
  cases_Q1[i] = unname(quantile(na.omit(data.frame(ii_
    File[i])$new_cases))[2])
  cat(ii_string[i], "Q1 new cases =", cases_Q1[i], "\n
    ")
}
```

Kết quả

```
Indonesia Q1 new cases: 766
Indonesia Q3 new cases: 6816.5
Indonesia Q1 new deaths: 33
Indonesia Q3 new deaths: 187

Japan Q1 new cases: 225
Japan Q3 new cases: 3342.5
Japan Q1 new deaths: 4
Japan Q3 new deaths: 46

Vietnam Q1 new cases: 1
Vietnam Q3 new cases: 4758
vietnam Q1 new deaths: 0
vietnam Q3 new deaths: 113
```

**Hình:** Giá trị tứ phân vị



## Nhiệm vụ ii: Nhóm câu hỏi liên quan đến mô tả thống kê cơ bản dữ liệu

### ③ Tính giá trị trung bình (Avg)

Chúng ta sử dụng hàm *mean()* và bỏ qua các giá trị NA để tính giá trị trung bình.

```
cases_avg <- vector(length = 3)

for (i in 1:3) {
  cases_avg[i] = mean(na.omit(data.frame(ii_File[i])$
    new_cases))
  cat(ii_string[i], "average new cases =", cases_avg[i],
    ], "\n")
}
```

Kết quả

```
Indonesia average new cases: 7078.772
Indonesia average new deaths: 205.6287

Japan average new cases: 5822.466
Japan average new deaths: 29.38347

vietnam average new cases: 3610.399
vietnam average new deaths: 69.28822
```

**Hình:** Giá trị trung bình



## Nhiệm vụ ii: Nhóm câu hỏi liên quan đến mô tả thống kê cơ bản dữ liệu

### ④ Tính giá trị độ lệch chuẩn (Std)

Chúng ta sử dụng hàm `sd()` và bỏ qua các giá trị NA để tính độ lệch chuẩn.

```
cases_std <- vector(length = 3)

for (i in 1:3) {
  cases_std[i] = sd(na.omit(data.frame(ii_File[i])$new_
    _cases))
  cat(ii_string[i], "standard deviation new cases =",
    cases_std[i], "\n")
}
```

Kết quả

```
Indonesia standard deviation new cases: 10904.26
Indonesia standard deviation new deaths: 348.4646

Japan standard deviation new cases: 16231.87
Japan standard deviation new deaths: 36.63266

vietnam standard deviation new cases: 6917.646
vietnam standard deviation new deaths: 116.4545
```

**Hình:** Giá trị độ lệch chuẩn



## Nhiệm vụ ii: Nhóm câu hỏi liên quan đến mô tả thống kê cơ bản dữ liệu

- ⑤ Đếm xem có bao nhiêu outliers, một quan sát mà giá trị của nó nằm trong khoảng sau:

$$IQR = Q3 - Q1$$

$$outliers < Q1 - 1.5 * IQR \text{ hoặc } outliers > Q3 + 1.5 * IQR$$

Với giá trị Q1, Q3 đã tính ở câu trên, ta dễ dàng tính được giá trị IQR. Sau đó kết hợp `subset()` để trích xuất dữ liệu thỏa mãn outlier và `nrow()` để xác định số hàng trong `subset` vừa thực hiện.

```
cases_outlier <- vector(length = 3)

for (i in 1:3) {
  cases_IQR = cases_Q3[i] - cases_Q1[1]
  cases_outlier[i] = nrow(subset(data.frame(ii_File[i
    ]),
    new_cases < cases_Q1[i] - 1.5*cases_IQR |
    new_cases > cases_Q3[i] + 1.5*cases_IQR))

  cat(ii_string[i], "outliers new cases =", cases_
    outlier[i], "\n")
}
```





## Nhiệm vụ ii: Nhóm câu hỏi liên quan đến mô tả thống kê cơ bản dữ liệu

### Kết quả

```
Indonesia outliers new cases = 80  
Indonesia outliers new deaths = 74
```

```
Japan outliers new cases = 93  
Japan outliers new deaths = 93
```

```
Vietnam outliers new cases = 115  
Vietnam outliers new deaths = 63
```

**Hình:** *Số lượng outlier*



## Nhiệm vụ ii: Nhóm câu hỏi liên quan đến mô tả thống kê cơ bản dữ liệu

- ⑥ Lập bảng mô tả số liệu thống kê cho từng đất nước thuộc về nhóm:

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

Chúng ta lập bảng bằng cách sử dụng *cbind()* và *rbind()* để kết hợp các dữ liệu lại với nhau.

```
cases_table <- vector()
for (i in 1:3) {
  cases_table = rbind(cases_table, cbind("Countries" =
    ii_string[i],
    "Min"=cases_min[i], "Q1"=cases_Q1[i], "Q2"=cases_Q2[
    i],
    "Q3"=cases_Q3[i], "Max"=cases_max[i], "Avg"=cases_
    avg[i],
    "Std"=cases_std[i], "Outlier"=cases_outlier[i]))
}
```



## Nhiệm vụ ii: Nhóm câu hỏi liên quan đến mô tả thống kê cơ bản dữ liệu

### Kết quả

Countries <chr>	Min <chr>	Q1 <chr>	Q2 <chr>	Q3 <chr>	Max <chr>	Avg <chr>	Std <chr>	Outlier <chr>
Indonesia	0	766	3874	6816.5	64718	7078.7719054242	10904.2606094201	80
Japan	0	225	1032	3342.5	104345	5822.46640316206	16231.8661554278	93
Vietnam	0	1	10	4758	54830	3610.39920948617	6917.64550707318	115

Countries <chr>	Min <chr>	Q1 <chr>	Q2 <chr>	Q3 <chr>	Max <chr>	Avg <chr>	Std <chr>	Outlier <chr>
Indonesia	0	33	100	187	2069	205.628691983122	348.46457167239	74
Japan	0	4	14	46	271	29.3834688346883	36.6326603229071	93
Vietnam	0	0	0	113	804	69.2882249560633	116.454478889211	63

**Hình:** Bảng số liệu new cases (phía trên) và new deaths (phía dưới)



## Nhiệm vụ ii: Nhóm câu hỏi liên quan đến mô tả thống kê cơ bản dữ liệu

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

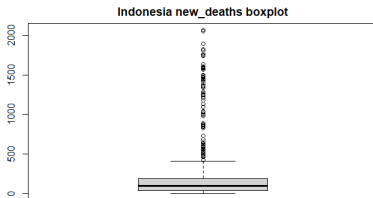
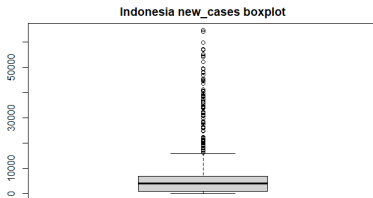
Rất rõ ràng, chúng ta sử dụng hàm *boxplot()* để vẽ biểu đồ boxplot cho dữ liệu.

```
for (i in 1:3) {  
  boxplot(data.frame(ii_File[i])$new_cases, main=paste  
    (ii_string[i], "new_cases boxplot"))  
  boxplot(data.frame(ii_File[i])$new_deaths, main=  
    paste(ii_string[i], "new_deaths boxplot"))  
}
```



## Nhiệm vụ ii: Nhóm câu hỏi liên quan đến mô tả thống kê cơ bản dữ liệu

### Kết quả



## Nhiệm vụ ii: Nhóm câu hỏi liên quan đến mô tả thống kê cơ bản dữ liệu

Thống kê khảo sát kết quả Covid-19

Huỳnh Tường Nguyễn,  
Nguyễn Ngọc Lê



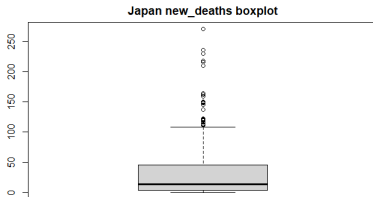
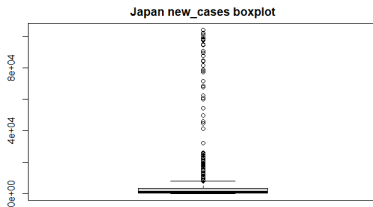
Động cơ nghiên cứu

Mục tiêu

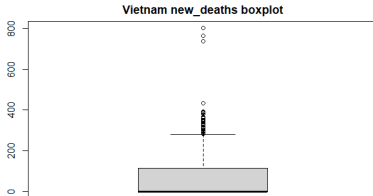
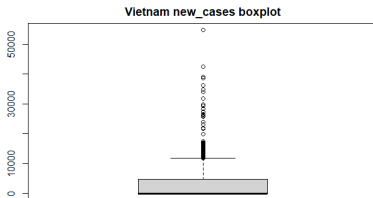
Kiến thức chuẩn bị

Nhiệm vụ

Kết luận



## Nhiệm vụ ii: Nhóm câu hỏi liên quan đến mô tả thống kê cơ bản dữ liệu



**Hình:** Biểu đồ boxplot của new cases và new deaths



## Nhiệm vụ iii: Nhóm câu hỏi liên quan đến dữ liệu thể hiện thu thập dữ liệu

**Xử lý chung:** Đầu tiên chúng ta nhập dữ liệu vào bảng, sửa những giá trị âm lại thành giá trị dương và sau đó lọc tiếp dữ liệu của từng nước cần được xử lý ra bảng.

Vì câu hỏi tính số liệu thống kê lần lượt cho nhiễm và tử vong như nhau (trừ câu 7 và 8) nên báo cáo chỉ giới thiệu về cách xử lý đối với lượt nhiễm; làm tương tự đối với lượt tử vong.

```
dataFile <- read_csv("owid-covid-data.csv", show_col_
  types = FALSE)
```

```
dataFile$new_cases <- abs(dataFile$new_cases)
dataFile$new_deaths <- abs(dataFile$new_deaths)
```

```
dataFile_ISO <- subset(dataFile, iso_code==country_
  code)
```





## Nhiệm vụ iii: Nhóm câu hỏi liên quan đến dữ liệu thể hiện thu thập dữ liệu

❶ Có bao nhiêu ngày có số lần dữ liệu không được báo cáo mới.

Từ bảng dữ liệu cho từng nước, ta lọc ra các ngày có dữ liệu được báo cáo hợp lệ (khác 0 và khác NA), sau đó loại những ngày hợp lệ ra khỏi bảng dữ liệu chung của nước đó, ta được bảng những ngày không được báo cáo mới.

```
dataFile_cases <- subset(dataFile_ISO, dataFile_ISO$
  new_cases>0)
invalid_cases <- subset(dataFile_ISO, !(new_cases %in%
  dataFile_cases$new_cases) | is.na(new_cases),
  select = c(location, new_cases, new_deaths))
cat("1. Số ngày dữ liệu không được báo cáo mới:", nrow
  (invalid_cases), "ngày \n")
```



## Nhiệm vụ iii: Nhóm câu hỏi liên quan đến dữ liệu thể hiện thu thập dữ liệu

- ② Có bao nhiêu ngày có số ca nhiễm/ tử vong là thấp nhất được báo cáo mới.

Từ bảng dữ liệu được báo cáo hợp lệ, ta tìm số ca nhiễm trong ngày thấp nhất rồi thống kê xem có bao nhiêu ngày có số ca nhiễm bằng với số ca vừa tìm được.

```
cases_min <- min(dataFile_cases$new_cases)
cases_min_Freq <- table(dataFile_ISO$new_cases==cases_min)
cat("2. Số ngày có số ca nhiễm thấp nhất:", unname(
  cases_min_Freq["TRUE"]), "\n")
```



## Nhiệm vụ iii: Nhóm câu hỏi liên quan đến dữ liệu thể hiện thu thập dữ liệu

- ③ Có bao nhiêu ngày có số ca nhiễm/ tử vong là cao nhất được báo cáo mới

Từ bảng dữ liệu được báo cáo hợp lệ, ta tìm số ca nhiễm trong ngày cao nhất rồi thống kê xem có bao nhiêu ngày có số ca nhiễm bằng với số ca vừa tìm được.

```
cases_max <- max(dataFile_cases$new_cases)
cases_max_Freq <- table(dataFile_ISO$new_cases==cases_max)
cat("3. Số ngày có số ca nhiễm cao nhất:", unname(
  cases_max_Freq["TRUE"]), "\n")
```



## Nhiệm vụ iii: Nhóm câu hỏi liên quan đến dữ liệu thể hiện thu thập dữ liệu

- ④ Thể hiện bảng số liệu về dữ liệu được báo cáo mới và không được báo cáo mới.

Không được báo cáo mới: Xuất ra bảng những ngày không có báo cáo mới (ở câu 1)

```
colnames(invalid_cases) <- c("Countries", "Infections",  
                             , "Deaths")  
cat("4. \nKhông được báo cáo mới: \n")  
print(invalid_cases)
```

Báo cáo mới: Từ bảng dữ liệu chung của từng nước, kết hợp với số ca thấp nhất/cao nhất tìm được ở câu 2) và 3), chúng ta tạo bảng gồm những ngày có số ca nhiễm thấp nhất và cao nhất, sau đó xuất bảng ra màn hình.

```
min_max_cases <- subset(dataFile_ISO, new_cases==cases  
                        _min | new_cases==cases_max, select = c(location,  
                        new_cases, new_deaths))  
colnames(min_max_cases) <- c("Countries", "Infections",  
                             , "Deaths")  
cat("\nBáo cáo mới: \n")  
print(min_max_cases)
```



### Nhiệm vụ iii: Nhóm câu hỏi liên quan đến dữ liệu thể hiện thu thập dữ liệu

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

Hướng giải quyết: Tìm chuỗi ngày ngắn nhất mà có số ca nhiễm bằng NA

Trong R: Tạo một hàm mới *condition\_NA* xác định xem ngày hôm đó số ca nhiễm có phải NA hay không (hàm trả về *TRUE* hoặc *FALSE*), sau đó áp dụng hàm lên toàn bộ cột *new\_cases* của bảng dữ liệu của từng nước. Ta được một chuỗi kí tự bao gồm *TRUE* và *FALSE* nối tiếp với nhau, tương ứng với kết quả trả về của hàm. Sau đó dùng hàm *rle* để thống kê số lần xuất hiện liên tiếp của từng kết quả (*TRUE* hoặc *FALSE*). Cuối cùng ta tìm giá trị nhỏ nhất của số lần xuất hiện *TRUE*.



## Nhiệm vụ iii: Nhóm câu hỏi liên quan đến dữ liệu thể hiện thu thập dữ liệu

```
condition_NA <- function(x) is.na(x)
dataFile_cases_NA <- rle(condition_NA(dataFile_ISO$new
  _cases))
cases_NA_minFreq <- min(dataFile_cases_NA$lengths[
  dataFile_cases_NA$values == TRUE], na.rm = TRUE)
cases_NA_minFreq[!is.finite(cases_NA_minFreq)] <- 0
cat("5. Số ngày gần nhất liên tiếp không có dữ liệu
    được báo cáo:", cases_NA_minFreq, "\n")
```

Thông kê khảo sát kết quả Covid-19

Huỳnh Tường Nguyên,  
Nguyễn Ngọc Lê



Động cơ nghiên cứu

Mục tiêu

Kiến thức chuẩn bị

Nhiệm vụ

Kết luận

## Nhiệm vụ iii: Nhóm câu hỏi liên quan đến dữ liệu thể hiện thu thập dữ liệu

- ⑥ Cho biết số ngày dài nhất liên tiếp mà không có dữ liệu được báo cáo

Tương tự câu 5), nhưng chúng ta tìm giá trị lớn nhất của số lần xuất hiện *TRUE*

```
condition_NA <- function(x) is.na(x)
dataFile_cases_NA <- rle(condition_NA(dataFile_ISO$new
  _cases))
cases_NA_maxFreq <- max(dataFile_cases_NA$lengths[
  dataFile_cases_NA$values == TRUE], na.rm = TRUE)
cases_NA_maxFreq[!is.finite(cases_NA_minFreq)] <- 0
cat("6. Số ngày dài nhất liên tiếp không có dữ liệu
    được báo cáo:", cases_NA_maxFreq, "\n")
```



## Nhiệm vụ iii: Nhóm câu hỏi liên quan đến dữ liệu thể hiện thu thập dữ liệu

- ⑦ Cho biết số ngày ngắn nhất liên tiếp mà không có người nhiễm bệnh mới

Hướng giải quyết: Tìm chuỗi ngày ngắn nhất mà có số ca nhiễm bằng 0

Trong R: Tạo một hàm mới *condition\_no\_new\_cases* xác định xem ngày hôm đó số ca nhiễm có bằng 0 hay không (hàm trả về *TRUE* hoặc *FALSE*). Sau đó thực hiện tương tự như câu 5)

```
condition_no_new_cases <- function(x) x==0
dataFile_cases_zero <- rle(condition_no_new_cases(
  dataFile_IS0$new_cases))
cases_zero_minFreq <- min(dataFile_cases_zero$lengths[
  dataFile_cases_zero$values == TRUE], na.rm = TRUE)
cases_zero_minFreq[!is.finite(cases_zero_minFreq)] <-
0
cat("7. Số ngày ngắn nhất liên tiếp không có người
    nhiễm bệnh mới:", cases_zero_minFreq, "\n")
```





## Nhiệm vụ iii: Nhóm câu hỏi liên quan đến dữ liệu thể hiện thu thập dữ liệu

- ⑧ Cho biết số ngày dài nhất liên tiếp mà không có người nhiễm bệnh mới

Tương tự câu 7), nhưng chúng ta tìm giá trị lớn nhất của số lần xuất hiện *TRUE*

```
condition_no_new_cases <- function(x) x==0
dataFile_cases_zero <- rle(condition_no_new_cases(
  dataFile_IS0$new_cases))
cases_zero_maxFreq <- max(dataFile_cases_zero$lengths[
  dataFile_cases_zero$values == TRUE], na.rm = TRUE)
cases_zero_maxFreq[!is.finite(cases_zero_maxFreq)] <-
0
cat("8. Số ngày dài nhất liên tiếp không có người
nhiễm bệnh mới:", cases_zero_maxFreq, "\n")
```



# Nhiệm vụ iii: Nhóm câu hỏi liên quan đến dữ liệu thể hiện thu thập dữ liệu

## Kết quả

### Indonesia

---Ca nhiễm---

1. Số ngày đủ liệu không được báo cáo mỗi: 8 ngày
2. Số ngày có số ca nhiễm thấp nhất: 3
3. Số ngày có số ca nhiễm cao nhất: 1
- 4.

Không được báo cáo mỗi:

# A tibble: 8 x 3

	Countries	Infections	Deaths
	<chr>	<dbl>	<dbl>
1	Indonesia	0	NA
2	Indonesia	0	NA
3	Indonesia	0	NA
4	Indonesia	0	NA
5	Indonesia	0	0
6	Indonesia	0	0
7	Indonesia	0	0
8	Indonesia	NA	158

Báo cáo mỗi:

# A tibble: 4 x 3

	Countries	Infections	Deaths
	<chr>	<dbl>	<dbl>
1	Indonesia	2	NA
2	Indonesia	2	NA
3	Indonesia	2	NA
4	Indonesia	64718	167

5. Số ngày ngắn nhất liên tiếp không có đủ liệu được báo cáo: 1
6. Số ngày dài nhất liên tiếp không có đủ liệu được báo cáo: 1
7. Số ngày ngắn nhất liên tiếp không có người nhiễm bệnh mỗi: 1
8. Số ngày dài nhất liên tiếp không có người nhiễm bệnh mỗi: 3

**Hình:** Kết quả đối với ca nhiễm



## Nhiệm vụ iii: Nhóm câu hỏi liên quan đến dữ liệu thể hiện thu thập dữ liệu

```
---Ca tu vong---
1. So ngay du lieu khong duoc bao cao moi: 15 ngay
2. So ngay co so ca tu vong thap nhat: 5
3. So ngay co so ca tu vong cao nhat: 1
4.
Khong duoc bao cao moi:# A tibble: 15 x 3
  Countries Infections Deaths
  <chr>      <dbl>    <dbl>
1 Indonesia      2      NA
2 Indonesia      0      NA
3 Indonesia      0      NA
4 Indonesia      0      NA
5 Indonesia      2      NA
6 Indonesia      0      NA
7 Indonesia      2      NA
8 Indonesia     13      NA
9 Indonesia      8      NA
10 Indonesia     0      0
11 Indonesia    21      0
12 Indonesia    17      0
13 Indonesia    38      0
14 Indonesia     0      0
15 Indonesia     0      0
```

**Hình:** Kết quả đối với ca tử vong



## Nhiệm vụ iii: Nhóm câu hỏi liên quan đến dữ liệu thể hiện thu thập dữ liệu

Bao cao moi:

```
# A tibble: 6 x 3
```

	Countries	Infections	Deaths
	<chr>	<dbl>	<dbl>
1	Indonesia	7	1
2	Indonesia	27	1
3	Indonesia	65	1
4	Indonesia	45203	2069
5	Indonesia	264	1
6	Indonesia	174	1

5. So ngay ngan nhat lien tiep khong co du lieu duoc bao cao: 9

6. So ngay dai nhat lien tiep khong co du lieu duoc bao cao: 9

**Hình:** *Kết quả đối với ca tử vong*

Thông kê khảo sát kết quả Covid-19

Huynh Tuong Nguyen,  
Nguyen Ngoc Le



Động cơ nghiên cứu

Mục tiêu

Kiến thức chuẩn bị

Nhiệm vụ

Kết luận

## Nhiệm vụ iv: Nhóm câu hỏi liên quan đến trực quan dữ liệu

Thông kê khảo sát kết quả Covid-19

Huynh Tuong Nguyen,  
Nguyen Ngoc Le



Động cơ nghiên cứu

Mục tiêu

Kiến thức chuẩn bị

Nhiệm vụ

Kết luận

### Xử lý chung cho câu 1 và 2

Chúng ta tính tổng số quốc gia dựa trên châu lục, tính tỉ lệ số đất nước từng châu lục so với số đất nước toàn thế giới rồi đưa chúng vào bảng.

```
Countries <- dataFile %>% select(location)
Con <- dataFile %>% select(continent)
temp<- cbind(Countries,Con)
temp <- distinct(temp)
Countries <- count(temp, 'continent')
probability <- prop.table(Countries[,2])
cumulative <- cumsum(Countries[,2])
Countries <- cbind(Countries, probability, cumulative)
```

## Nhiệm vụ iv: Nhóm câu hỏi liên quan đến trực quan dữ liệu

Thông kê khảo sát kết quả Covid-19

Huỳnh Tường Nguyễn,  
Nguyễn Ngọc Lê



Động cơ nghiên cứu

Mục tiêu

Kiến thức chuẩn bị

Nhiệm vụ

Kết luận

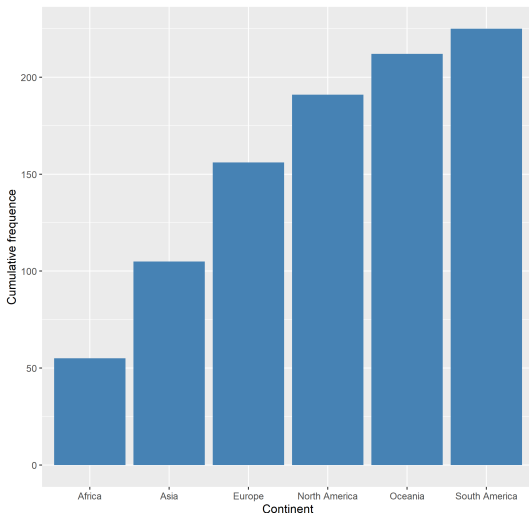
### ① Vẽ biểu đồ tần số tích lũy quốc gia cho các châu lục

```
graph1 <- ggplot(data = Countries, aes(x=continent
, y=cumulative)) +
geom_bar(stat = "identity", position = "dodge",
fill = "steelblue") +
labs(title = "", x="Continent", y="Cumulative
frequency")
```

```
graph1
ggsave("iv.1) Cumulative frequency.png", plot =
graph1)
```

# Nhiệm vụ iv: Nhóm câu hỏi liên quan đến trực quan dữ liệu

## Kết quả



**Hình:** Biểu đồ tần số tích lũy quốc gia cho các châu lục



## Nhiệm vụ iv: Nhóm câu hỏi liên quan đến trực quan dữ liệu

Thống kê khảo sát kết quả Covid-19

Huynh Tuong Nguyen,  
Nguyen Ngoc Le



Động cơ nghiên cứu

Mục tiêu

Kiến thức chuẩn bị

Nhiệm vụ

Kết luận

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

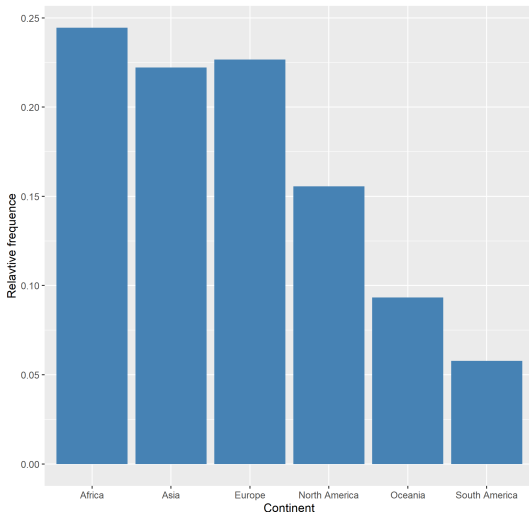
```
graph2 <- ggplot(data = Countries, aes(x=continent  
, y=probability)) +  
  geom_bar(stat = "identity", position = "dodge",  
    fill = "steelblue") +  
  labs(title = "", x="Continent", y="Relative  
    frequency")
```

```
graph2  
ggsave("iv.2) Relative frequency.png", plot =  
  graph2)
```



# Nhiệm vụ iv: Nhóm câu hỏi liên quan đến trực quan dữ liệu

## Kết quả



**Hình:** Biểu đồ tần số tương đối quốc gia cho các châu lục

Thống kê khảo sát kết quả Covid-19

Huỳnh Tường Nguyễn,  
Nguyễn Ngọc Lê



Động cơ nghiên cứu

Mục tiêu

Kiến thức chuẩn bị

**Nhiệm vụ**

Kết luận

## Nhiệm vụ iv: Nhóm câu hỏi liên quan đến trực quan dữ liệu

### Xử lý chung cho câu 3 và 4

Chúng ta lấy dữ liệu *newcases* và số *newdeaths* theo quốc gia, cùng với *date*, rồi tách bộ phận dữ liệu gồm 7 ngày cuối cùng của năm cuối cùng của từng quốc gia.

```
dataFile$new_cases <- abs(dataFile$new_cases)
dataFile$new_deaths <- abs(dataFile$new_deaths)
InJaVi <- dataFile %>% filter(location == "Indonesia"
  | location == "Japan" | location == "Vietnam")
InJaVi <- InJaVi %>% select(location | date | new_
  cases | new_deaths)

tmp <- InJaVi
formattedDate <- as.Date(tmp$date, format = "%m/%d/%Y")
tmp[, "date"] <- formattedDate

thelastday <- max(formattedDate)
thelastsevendays <- tmp %>% group_by(location) %>%
  filter(date > thelastday - 7)
```



## Nhiệm vụ iv: Nhóm câu hỏi liên quan đến trực quan dữ liệu

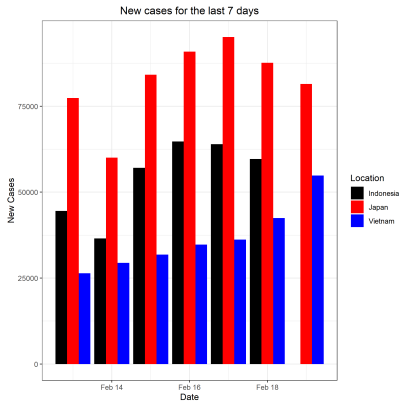
- ③ Vẽ biểu đồ thể hiện nhiễm bệnh đã báo cáo của các quốc gia trong 7 ngày cuối của năm cuối cùng

```
graph3 <- ggplot(data = thelastsevendays, aes(x =  
  date, y = new_cases, fill = factor(location)))  
  +  
  theme_bw() +  
  geom_bar(stat = "identity", position = "dodge")  
  +  
  labs(title = "New cases for the last 7 days", x =  
    "Date", y = "New Cases") +  
  theme(plot.title = element_text(hjust = 0.5)) +  
    scale_fill_manual("Location", values = c("  
      Indonesia" = "black", "Japan" = "red", "  
      Vietnam" = "blue"))  
  
graph3  
ggsave("iv.3) New cases for the last 7 days.png",  
  plot = graph3)
```



## Nhiệm vụ iv: Nhóm câu hỏi liên quan đến trực quan dữ liệu

### Kết quả



**Hình:** Biểu đồ thể hiện nhiễm bệnh đã báo cáo của các quốc gia trong 7 ngày cuối của năm cuối cùng

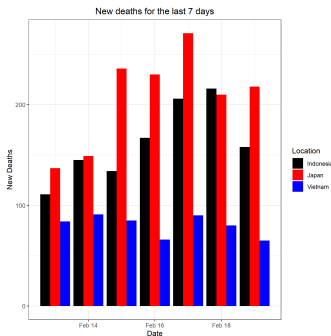


## Nhiệm vụ iv: Nhóm câu hỏi liên quan đến trực quan dữ liệu

- ④ Vẽ biểu đồ thể hiện tử vong đã báo cáo của các quốc gia trong 7 ngày cuối của năm cuối cùng

Hoàn toàn tương tự câu 3, chỉ đổi *new\_cases* thành *new\_deaths*.

Kết quả



**Hình:** Biểu đồ thể hiện tử vong đã báo cáo của các quốc gia trong 7 ngày cuối của năm cuối cùng



## Nhiệm vụ iv: Nhóm câu hỏi liên quan đến trực quan dữ liệu

Thông kê khảo sát kết quả Covid-19

Huynh Tuong Nguyen,  
Nguyen Ngoc Le



Động cơ nghiên cứu

Mục tiêu

Kiến thức chuẩn bị

Nhiệm vụ

Kết luận

### Xử lý chung cho câu 5 và 6

Chúng ta lấy dữ liệu từ câu 5 phần *ii* để làm cơ sở xử lý yêu cầu bài toán.

```
datafromii.5 <- cbind(c("Indonesia", "Japan", "Vietnam")  
  ,as.data.frame(cases_outlier),as.data.frame(deaths_outlier))  
colnames(datafromii.5) <- c("Country", "casesOutliers",  
  "deathsOutliers")
```

## Nhiệm vụ iv: Nhóm câu hỏi liên quan đến trực quan dữ liệu

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

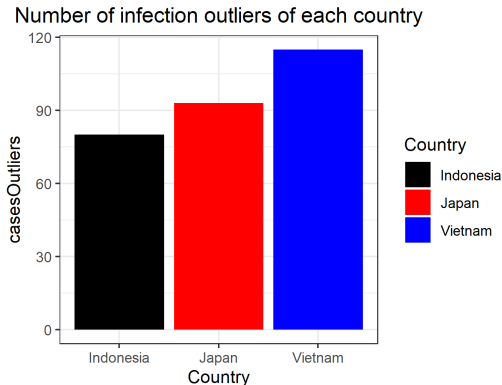
```
graph5 <- ggplot(data = datafromii.5, aes(x=
  Country, y=casesOutliers, fill = factor(
    Country))) +
  geom_bar(stat="identity") +
  theme_bw() +
  labs(title = "Number of infection outliers of
    each country") +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_fill_manual("Country", values = c("
    Indonesia" = "black", "Japan" = "red", "
    Vietnam" = "blue"))

graph5
ggsave("iv.5) caseOutPlot.png", plot = graph5)
```



## Nhiệm vụ iv: Nhóm câu hỏi liên quan đến trực quan dữ liệu

### Kết quả



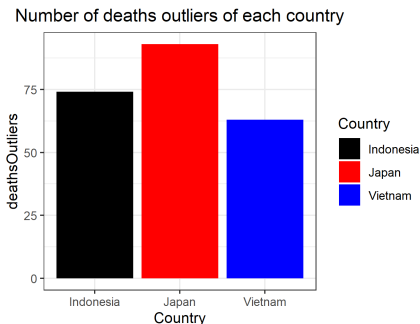
**Hình:** Biểu đồ phổ đất nước xuất hiện outliers cho nhiễm bệnh





## Nhiệm vụ iv: Nhóm câu hỏi liên quan đến trực quan dữ liệu

⑥ Vẽ biểu đồ phổ đất nước xuất hiện outliers cho tử vong  
Hoàn toàn tương tự như câu 5, chỉ thay đổi *new\_cases* thành *new\_deaths*  
Kết quả



**Hình:** Biểu đồ phổ đất nước xuất hiện outliers cho tử vong



## Nhiệm vụ v: Nhóm câu hỏi liên quan đến trực quan dữ liệu theo thời gian là tháng

**Xử lý chung:** Đầu tiên chúng ta nhập dữ liệu vào bảng, sửa những giá trị âm lại thành giá trị dương, đồng thời định dạng ngày tháng lại để dễ dàng xử lý và sau đó lọc tiếp dữ liệu của từng nước cần được xử lý ra bảng.

```
dataFile <- read_csv("owid-covid-data.csv", show_col_
  types = FALSE)

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

dataFile$date <- as.Date(dataFile$date, format="%m/%d/
  %Y")

dataFile_ISO <- subset(dataFile, iso_code==country_
  code)
```



## Nhiệm vụ v: Nhóm câu hỏi liên quan đến trực quan dữ liệu theo thời gian là tháng

Vì các câu hỏi như nhau đối với từng năm cần xử lý dữ liệu nên báo cáo chỉ giới thiệu qua cách xử lý đối với năm đầu tiên.

❶ Biểu đồ thể hiện thu thập dữ liệu nhiễm bệnh cho từng tháng. Đầu tiên ta tách dữ liệu về các ca nhiễm bệnh của tất cả các tháng cần được xử lý ra bảng riêng, sau đó thêm một cột *month* ứng với tên tháng cho từng ngày được báo cáo. Cuối cùng ta dùng *ggplot* để vẽ đồ thị và *ggsave* để lưu về máy.

```
dataFile_cases <- subset(dataFile_ISO, new_cases>0 &
  ((format(date, "%m")=="01") | (format(
    date, "%m")=="03") |
  (format(date, "%m")=="04") | (format(
    date, "%m")=="05"))) &
  format(date, "%Y")=="2020",
  select = c(date, new_cases))
dataFile_cases <- dataFile_cases %>% mutate(month = as
  .numeric(format(dataFile_cases$date, "%m"))) %>%
  mutate(month = month.name[month])
dataFile_cases$month = factor(dataFile_cases$month,
  levels = c("January", "March", "April", "May"))
```



## Nhiệm vụ v: Nhóm câu hỏi liên quan đến trực quan dữ liệu theo thời gian là tháng

```
dataFile_cases_plot <- ggplot(data = dataFile_cases,
  mapping = aes(x = date, y = new_cases, label = new_
    _cases)) + geom_line() + geom_point() +
  facet_grid(~ dataFile_cases$month, scales = "free_
    x", drop = FALSE) +
  labs(x = "", y = "Number of new cases", title = "
    New COVID-19 cases in Indonesia", subtitle = "
    Number of newly reported COVID-19 cases by
    date in January, March, April and May of 2020"
  ) +
  theme_bw() + theme(text = element_text(size = 14))
  +
  theme(plot.title = element_text(face = "bold")) +
  theme(plot.subtitle = element_text(face = "italic"
  )) +
  theme(axis.text.x = element_text(angle = 0, size =
    9)) +
  theme(plot.margin = margin(1,1.2,0.5,1, "cm")) +
  theme(panel.spacing.x = unit(4, "mm")) +
  scale_y_continuous(labels = label_number())
ggsave(dataFile_cases_plot, filename = paste(country_
  code, "2020_new_cases.pdf", sep = "_"), width = 12,
  height = 6)
```



# Nhiệm vụ v: Nhóm câu hỏi liên quan đến trực quan dữ liệu theo thời gian là tháng

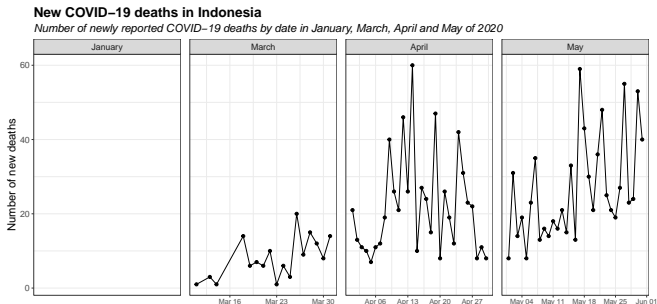
## Kết quả



**Hình:** Biểu đồ thể hiện thu thập dữ liệu nhiễm bệnh cho từng tháng

## Nhiệm vụ v: Nhóm câu hỏi liên quan đến trực quan dữ liệu theo thời gian là tháng

② Biểu đồ thể hiện thu thập dữ liệu tử vong cho từng tháng  
Tương tự câu 1), ta tách dữ liệu về các ca tử vong cho tất cả các tháng cần được xử lí.  
Kết quả



**Hình:** Biểu đồ thể hiện thu thập dữ liệu tử vong cho từng tháng

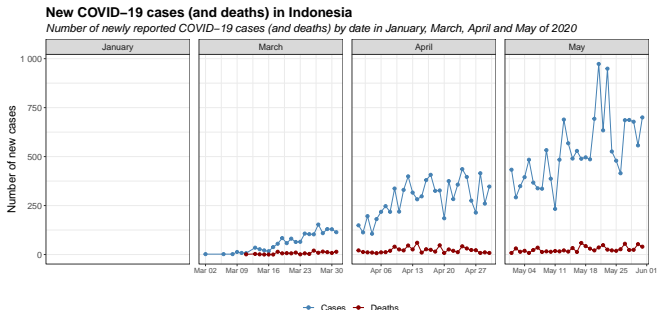


## Nhiệm vụ v: Nhóm câu hỏi liên quan đến trực quan dữ liệu theo thời gian là tháng

- ③ Biểu đồ thể hiện thu thập dữ liệu gồm nhiễm bệnh và tử vong cho từng tháng

Tương tự câu 2), ta tách dữ liệu về các ca nhiễm bệnh và tử vong cho tất cả các tháng cần được xử lí.

Kết quả



**Hình:** Biểu đồ thể hiện thu thập dữ liệu gồm nhiễm bệnh và tử vong cho từng tháng

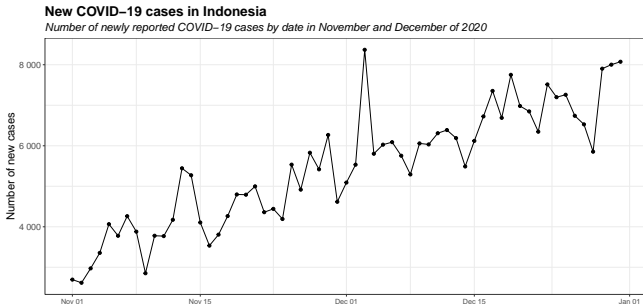


## Nhiệm vụ v: Nhóm câu hỏi liên quan đến trực quan dữ liệu theo thời gian là tháng

- ④ Biểu đồ thể hiện thu thập dữ liệu nhiễm bệnh gồm 2 tháng cuối của năm

Tương tự câu 1), ta tách dữ liệu về các ca nhiễm bệnh của 2 tháng cuối của năm..

Kết quả



**Hình:** Biểu đồ thể hiện thu thập dữ liệu nhiễm bệnh gồm 2 tháng cuối của năm



## Nhiệm vụ v: Nhóm câu hỏi liên quan đến trực quan dữ liệu theo thời gian là tháng

⑤ Biểu đồ thể hiện thu thập dữ liệu tử vong gồm 2 tháng cuối của năm

Tương tự câu 2), ta tách dữ liệu về các ca tử vong của 2 tháng cuối của năm.

Kết quả



**Hình:** Biểu đồ thể hiện thu thập dữ liệu tử vong gồm 2 tháng cuối của năm

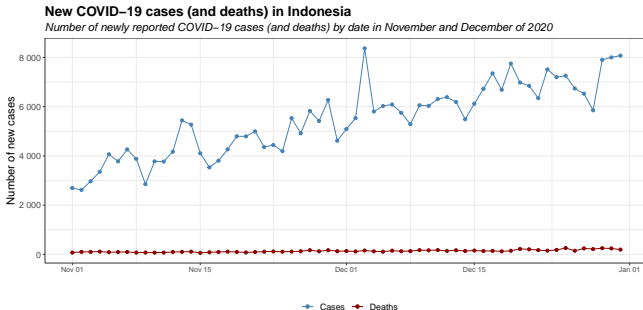


## Nhiệm vụ v: Nhóm câu hỏi liên quan đến trực quan dữ liệu theo thời gian là tháng

- ⑥ Biểu đồ thể hiện thu thập dữ liệu gồm nhiễm bệnh và tử vong gồm 2 tháng cuối của năm

Tương tự câu 3), ta tách dữ liệu về các ca nhiễm bệnh và tử vong cho 2 tháng cuối của năm.

Kết quả



**Hình:** Biểu đồ thể hiện thu thập dữ liệu gồm nhiễm bệnh và tử vong gồm 2 tháng cuối của năm



## Nhiệm vụ v: Nhóm câu hỏi liên quan đến trực quan dữ liệu theo thời gian là tháng

- ⑦ Biểu đồ thể hiện thu thập dữ liệu nhiễm bệnh tích lũy cho từng tháng

Tương tự câu 1), sau đó ta tiến hành cộng tích lũy số ca nhiễm với nhau bằng *cumsum*

```
dataFile_cases$new_cases <- cumsum(dataFile_cases$new_cases)
```

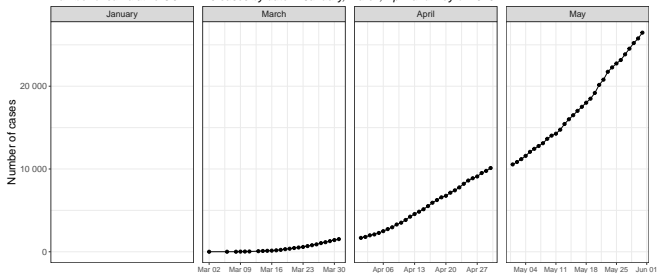


# Nhiệm vụ v: Nhóm câu hỏi liên quan đến trực quan dữ liệu theo thời gian là tháng

## Kết quả

### COVID-19 cases in Indonesia

Number of cumulative COVID-19 cases by date in January, March, April and May of 2020



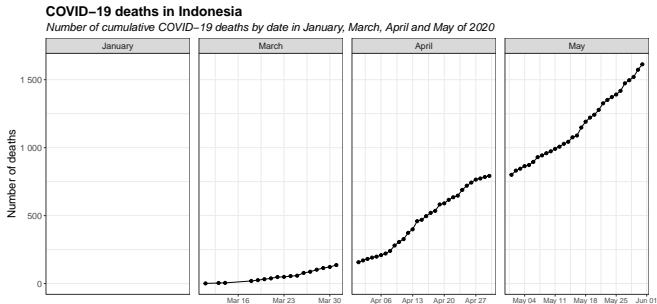
**Hình:** Biểu đồ thể hiện thu thập dữ liệu nhiễm bệnh tích lũy cho từng tháng

## Nhiệm vụ v: Nhóm câu hỏi liên quan đến trực quan dữ liệu theo thời gian là tháng

⑧ Biểu đồ thể hiện thu thập dữ liệu tử vong tích lũy cho từng tháng

Tương tự câu trên.

Kết quả



**Hình:** Biểu đồ thể hiện thu thập dữ liệu tử vong tích lũy cho từng tháng



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

- Với mỗi quốc gia mà thuộc về nhóm, 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.
- Xử lý chung:** Dùng trung bình của các ca nhiễm bệnh và tử vong được báo cáo trong 7 ngày gần nhất. Ta dùng vòng lặp *for*.

```
avg_nc_Jp_1_2020 <- c()  
avg_nc_Jp_1_2020[1] <- Japan_nc_1_2020$new_cases[1]  
avg_nc_Jp_1_2020[2] <- (Japan_nc_1_2020$new_cases[1] +  
  Japan_nc_1_2020$new_cases[2])/2  
avg_nc_Jp_1_2020[3] <- (Japan_nc_1_2020$new_cases[1] +  
  Japan_nc_1_2020$new_cases[2] + Japan_nc_1_2020$  
  new_cases[3])/3
```



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

```
avg_nc_Jp_1_2020[4] <- (Japan_nc_1_2020$new_cases [1] +  
  Japan_nc_1_2020$new_cases [2] + Japan_nc_1_2020$  
  new_cases [3] + Japan_nc_1_2020$new_cases [4])/4  
avg_nc_Jp_1_2020[5] <- (Japan_nc_1_2020$new_cases [1] +  
  Japan_nc_1_2020$new_cases [2] + Japan_nc_1_2020$  
  new_cases [3] + Japan_nc_1_2020$new_cases [4] +  
  Japan_nc_1_2020$new_cases [5])/5  
avg_nc_Jp_1_2020[6] <- (Japan_nc_1_2020$new_cases [1] +  
  Japan_nc_1_2020$new_cases [2] + Japan_nc_1_2020$  
  new_cases [3] + Japan_nc_1_2020$new_cases [4] +  
  Japan_nc_1_2020$new_cases [5] + Japan_nc_1_2020$new  
  _cases [6])/6  
for(i in 7:length(Japan_nc_1_2020$new_cases))  
{  
  avg_nc_Jp_1_2020[i]=  
  (Japan_nc_1_2020$new_cases [i] +  
  Japan_nc_1_2020$new_cases [i-1] +  
  Japan_nc_1_2020$new_cases [i-2] +  
  Japan_nc_1_2020$new_cases [i-3] +  
  Japan_nc_1_2020$new_cases [i-4] +  
  Japan_nc_1_2020$new_cases [i-5] +  
  Japan_nc_1_2020$new_cases [i-6])/7  
}
```



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

Để tính số lượng tích lũy, ta cũng dùng vòng lặp for:

```
acml_nc_Jp_1_2020 <- c()
acml_nc_Jp_1_2020[1] <- avg_nc_Jp_1_2020[1]
for(i in 2:length(avg_nc_Jp_1_2020))
{
    acml_nc_Jp_1_2020[i] <- avg_nc_Jp_1_2020[i] +
        acml_nc_Jp_1_2020[i-1]
}
```

Kết hợp 2 biến trung bình và tích lũy trên vào bảng trên:

```
Japan_nc_1_2020 <- data.frame(Japan_nc_1_2020, avg_nc_
    Jp_1_2020, acml_nc_Jp_1_2020)
```

Thực hiện tương tự các bước trên đối với những tháng khác và những quốc gia còn lại, cũng thực hiện tương tự khi lọc số liệu theo *new\_deaths*.

Từ các bảng số liệu đã lập, ta đã có đầy đủ dữ kiện để vẽ biểu đồ. Khi vẽ biểu đồ, với các bảng số liệu rộng, ta bỏ qua không xét.





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

### ❶ Biểu đồ thu thập nhiễm bệnh cho từng tháng

Với trục Ox là trục thời gian, trục Oy là nhiễm bệnh, ta vẽ được biểu đồ đường, mỗi đường đại diện cho số ca nhiễm bệnh của 1 nước.

Với mỗi *geom\_line* là một đường biểu thị cho số liệu *new\_cases* của 1 bảng số liệu không rỗng

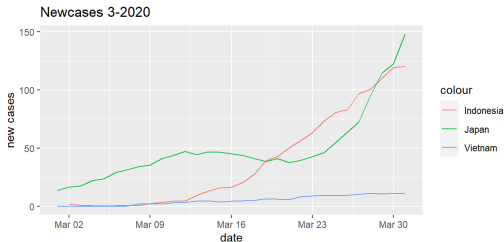
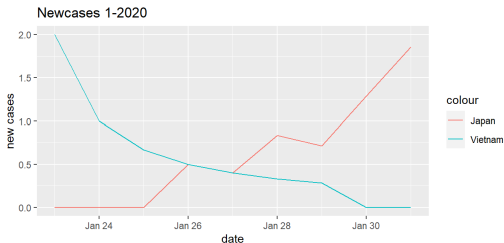
Ví dụ: Số ca nhiễm của tháng 1/2020

```
Newcases_1_2020 <- ggplot() +  
  geom_line(data=Japan_nc_1_2020, aes(x=datetime  
    , y=avg_nc_Jp_1_2020, color = 'Japan')) +  
  geom_line(data=Vietnam_nc_1_2020, aes(x=  
    datetime, y=avg_nc_Vn_1_2020, color = '  
    Vietnam')) +  
  labs(title = "Newcases 1-2020", x = "date", y  
    = "new cases")
```

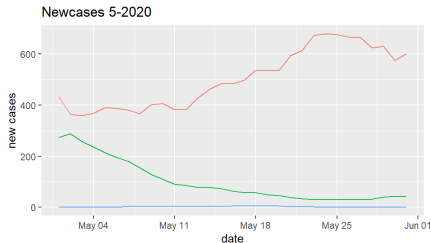
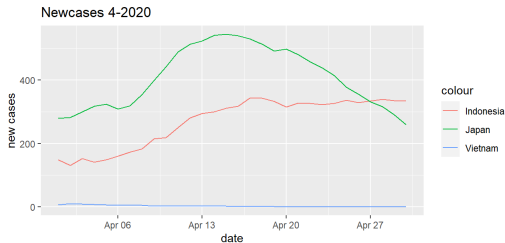


# Nhiệm vụ vi: Nhóm câu hỏi liên quan đến trực quan dữ liệu theo trung bình 7 ngày gần nhất

## Kết quả



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



Thông kê khảo sát kết quả Covid-19

Huỳnh Tường Nguyễn,  
Nguyễn Ngọc Lê



Động cơ nghiên cứu

Mục tiêu

Kiến thức chuẩn bị

Nhiệm vụ

Kết luận

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

Thông kê khảo sát kết quả Covid-19

Huỳnh Tường Nguyễn,  
Nguyễn Ngọc Lê



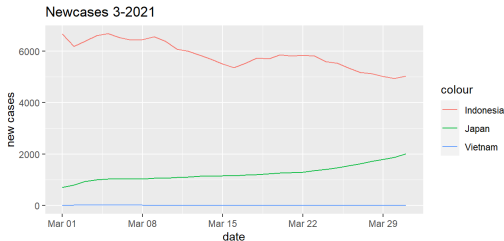
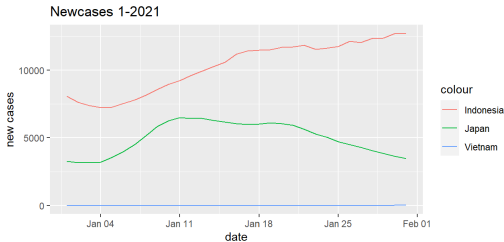
Động cơ nghiên cứu

Mục tiêu

Kiến thức chuẩn bị

Nhiệm vụ

Kết luận



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

Thống kê khảo sát kết quả Covid-19

Huỳnh Tường Nguyên,  
Nguyễn Ngọc Lê



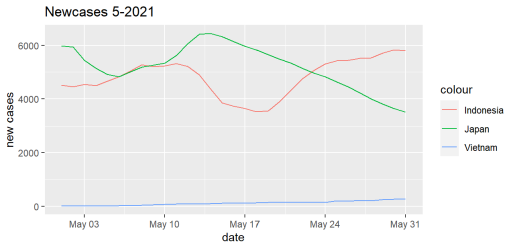
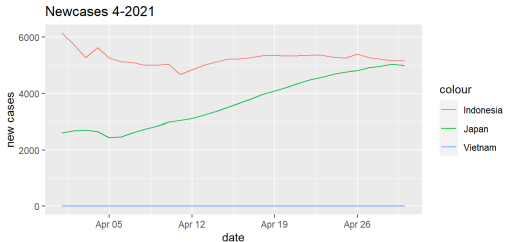
Động cơ nghiên cứu

Mục tiêu

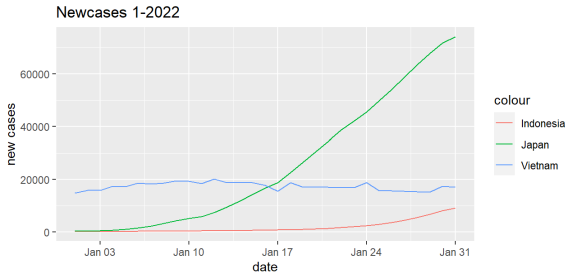
Kiến thức chuẩn bị

Nhiệm vụ

Kết luận



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



**Hình:** Biểu đồ thu thập nhiễm bệnh theo từng tháng

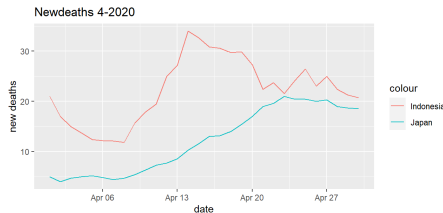
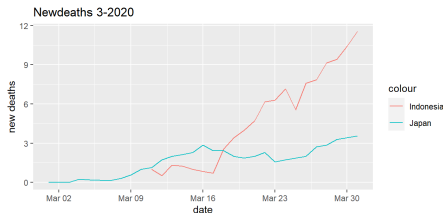


## Nhiệm vụ vi: Nhóm câu hỏi liên quan đến trực quan dữ liệu theo trung bình 7 ngày gần nhất

### ② Biểu đồ thu thập tử vong cho từng tháng

Hoàn toàn tương tự như câu 1, thay dữ liệu *new\_cases* thành *new\_deaths*.

Kết quả



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

Thông kê khảo sát kết quả Covid-19

Huynh Tuong Nguyen,  
Nguyen Ngoc Le



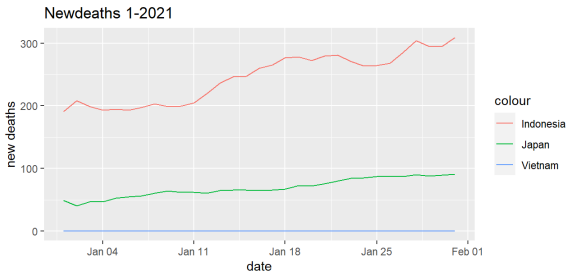
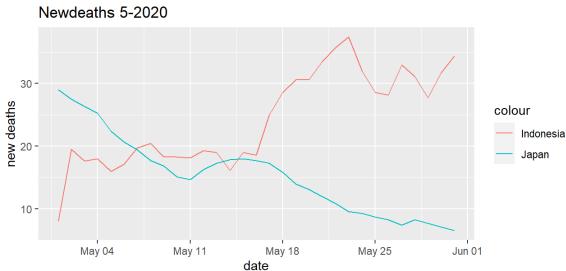
Động cơ nghiên cứu

Mục tiêu

Kiến thức chuẩn bị

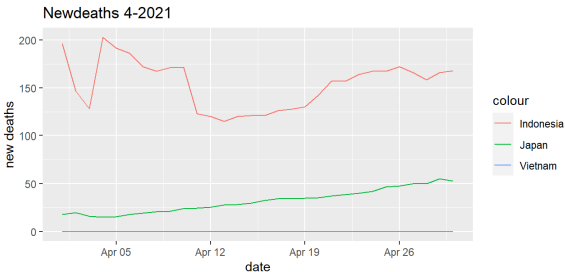
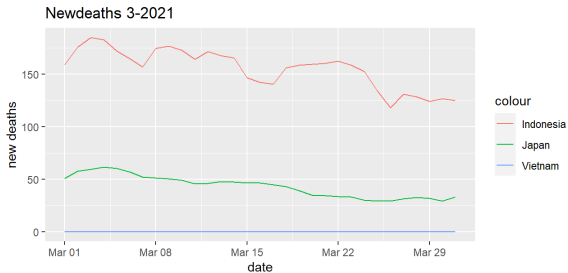
Nhiệm vụ

Kết luận





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



Thông kê khảo sát kết quả Covid-19

Huỳnh Tường Nguyễn,  
Nguyễn Ngọc Lê



Động cơ nghiên cứu

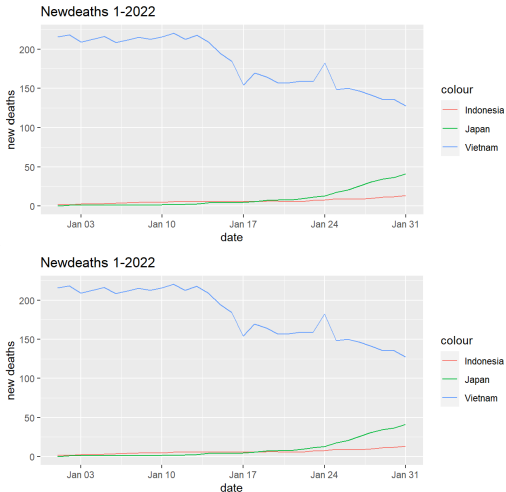
Mục tiêu

Kiến thức chuẩn bị

Nhiệm vụ

Kết luận

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



**Hình:** Biểu đồ thu thập tử vong theo từng tháng



Động cơ nghiên cứu

Mục tiêu

Kiến thức chuẩn bị

Nhiệm vụ

Kết luận

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

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

Ở câu này, ta sẽ kết hợp biểu diễn số ca nhiễm và số ca tử vong trong cùng một biểu đồ bằng cách thêm các *geom\_line* của *new\_cases* và *new\_deaths* (của các bảng dữ liệu khác rỗng) vào cùng một biểu đồ.

Ví dụ: Đối với tháng 1/2020, các quốc gia đều không có ghi nhận ca tử vong nào, nên biểu đồ cần vẽ chính là biểu đồ thu thập ca nhiễm.

Ví dụ: Đối với tháng 3/2020, ta có:



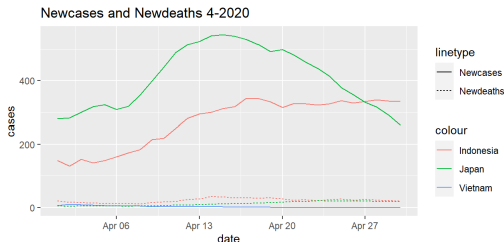
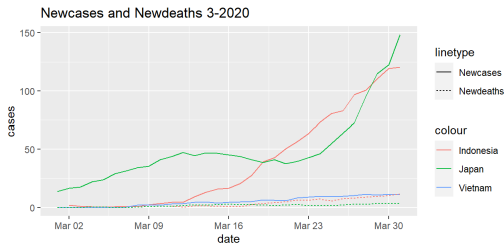
## Nhiệm vụ vi: Nhóm câu hỏi liên quan đến trực quan dữ liệu theo trung bình 7 ngày gần nhất

```
New_3_2020 <- ggplot() +  
  geom_line(data=Indonesia_nc_3_2020, aes(x=  
    datetime, y=avg_nc_Indo_3_2020, linetype =  
    'Newcases', color='Indonesia')) +  
  geom_line(data=Japan_nc_3_2020, aes(x=datetime  
    , y=avg_nc_Jp_3_2020, linetype = 'Newcases'  
    , color='Japan')) +  
  geom_line(data=Vietnam_nc_3_2020, aes(x=  
    datetime, y=avg_nc_Vn_3_2020, linetype = '  
    Newcases', color='Vietnam')) +  
  geom_line(data=Indonesia_nd_3_2020, aes(x=  
    datetime, y=avg_nd_Indo_3_2020, linetype =  
    'Newdeaths', color='Indonesia')) +  
  geom_line(data=Japan_nd_3_2020, aes(x=datetime  
    , y=avg_nd_Jp_3_2020, linetype = '  
    Newdeaths', color='Japan')) +  
  labs(title = "Newcases and Newdeaths 3-2020",  
    x = "date", y = "cases")
```



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

## Kết quả



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

Thông kê khảo sát kết quả Covid-19

Huỳnh Tường Nguyên,  
Nguyễn Ngọc Lê



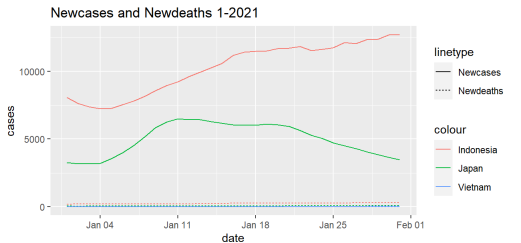
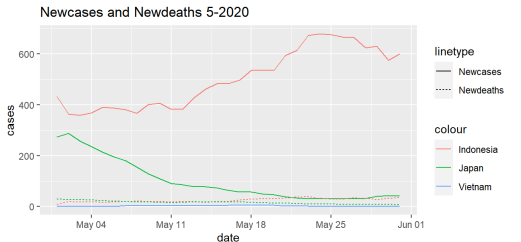
Động cơ nghiên cứu

Mục tiêu

Kiến thức chuẩn bị

Nhiệm vụ

Kết luận



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

Thông kê khảo sát kết quả Covid-19

Huỳnh Tường Nguyên,  
Nguyễn Ngọc Lê



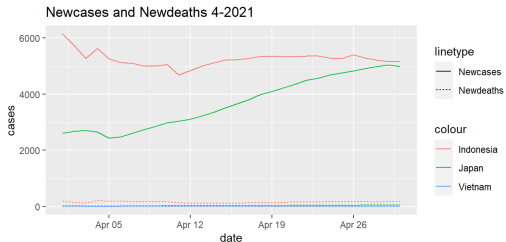
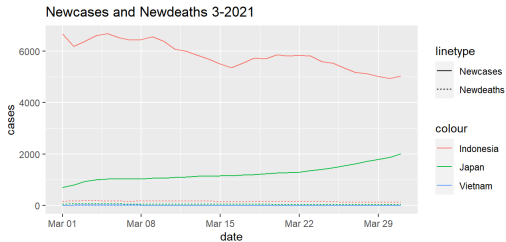
Động cơ nghiên cứu

Mục tiêu

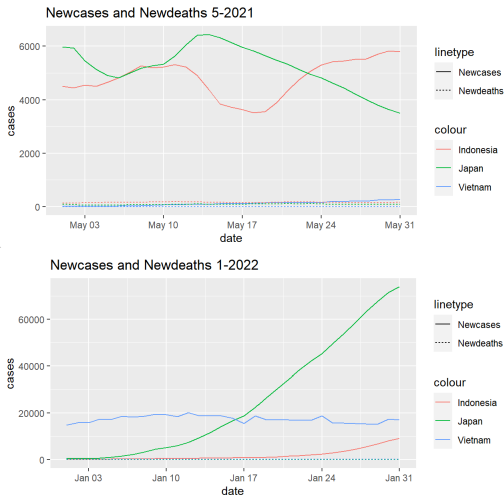
Kiến thức chuẩn bị

Nhiệm vụ

Kết luận



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



**Hình:** Biểu đồ thu thập nhiễm bệnh và tử vong cho từng tháng





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

④ Biểu đồ thu thập nhiễm bệnh gồm 2 tháng cuối của năm. Đối với 2 tháng cuối năm, đầu tiên ta cũng lọc dữ liệu như các tháng khác đã làm.

Ví dụ: Đối với 2 tháng cuối năm 2020 của Indonesia  
Ta có:

```
Indonesia_nc_11_12_2020 <- na.omit(Indonesia_nc[
  Indonesia_nc$datetime >= "2020-11-01" & Indonesia_
  nc$datetime <= "2020-12-31",])
avg_nc_Indo_11_12_2020 <- c()
avg_nc_Indo_11_12_2020[1] <- Indonesia_nc_11_12_2020$
  new_cases[1]
avg_nc_Indo_11_12_2020[2] <- (Indonesia_nc_11_12_2020$
  new_cases[1] + Indonesia_nc_11_12_2020$new_cases
  [2])/2
avg_nc_Indo_11_12_2020[3] <- (Indonesia_nc_11_12_2020$
  new_cases[1] + Indonesia_nc_11_12_2020$new_cases
  [2] + Indonesia_nc_11_12_2020$new_cases[3])/3
```



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

```
avg_nc_Indo_11_12_2020[4] <- (Indonesia_nc_11_12_2020$  
  new_cases[1] + Indonesia_nc_11_12_2020$new_cases  
  [2] + Indonesia_nc_11_12_2020$new_cases[3] +  
  Indonesia_nc_11_12_2020$new_cases[4])/4  
avg_nc_Indo_11_12_2020[5] <- (Indonesia_nc_11_12_2020$  
  new_cases[1] + Indonesia_nc_11_12_2020$new_cases  
  [2] + Indonesia_nc_11_12_2020$new_cases[3] +  
  Indonesia_nc_11_12_2020$new_cases[4] + Indonesia_  
  nc_11_12_2020$new_cases[5])/5  
avg_nc_Indo_11_12_2020[6] <- (Indonesia_nc_11_12_2020$  
  new_cases[1] + Indonesia_nc_11_12_2020$new_cases  
  [2] + Indonesia_nc_11_12_2020$new_cases[3] +  
  Indonesia_nc_11_12_2020$new_cases[4] + Indonesia_  
  nc_11_12_2020$new_cases[5] + Indonesia_nc_11_12_  
  2020$new_cases[6])/6
```



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

```
for(i in 7:length(Indonesia_nc_11_12_2020$new_cases))  
{  
  avg_nc_Indo_11_12_2020[i]=(Indonesia_nc_11_12_2020  
    $new_cases[i] + Indonesia_nc_11_12_2020$new_  
    cases[i-1] + Indonesia_nc_11_12_2020$new_cases  
    [i-2] + Indonesia_nc_11_12_2020$new_cases[i-3]  
    + Indonesia_nc_11_12_2020$new_cases[i-4] +  
    Indonesia_nc_11_12_2020$new_cases[i-5] +  
    Indonesia_nc_11_12_2020$new_cases[i-6])/7  
}  
acml_nc_Indo_11_12_2020 <- c()  
acml_nc_Indo_11_12_2020[1] <- avg_nc_Indo_11_12_  
  2020[1]  
for(i in 2:length(avg_nc_Indo_11_12_2020))  
{  
  acml_nc_Indo_11_12_2020[i] <- avg_nc_Indo_11_  
    12_2020[i] + acml_nc_Indo_11_12_2020[i-1]  
}  
Indonesia_nc_11_12_2020 <- data.frame(Indonesia_nc_11_  
  12_2020, avg_nc_Indo_11_12_2020, acml_nc_Indo_11_  
  12_2020)
```



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

Thực hiện tương tự cho Japan và Vietnam ta cũng được 2 bảng dữ liệu nữa.

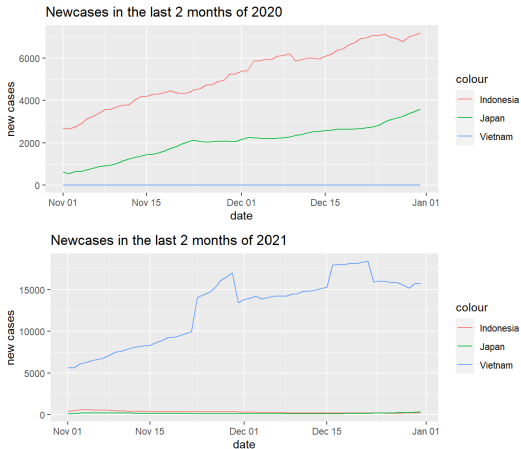
Sau đó ta tiến hành vẽ biểu đồ dựa trên các bảng dữ liệu vừa tìm được

```
Newcases_11_12_2020 <- ggplot() +  
  geom_line(data=Indonesia_nc_11_12_2020, aes(x=  
    datetime, y=avg_nc_Indo_11_12_2020, color  
    = 'Indonesia')) +  
  geom_line(data=Japan_nc_11_12_2020, aes(x=  
    datetime, y=avg_nc_Jp_11_12_2020, color =  
    'Japan')) +  
  geom_line(data=Vietnam_nc_11_12_2020, aes(x=  
    datetime, y=avg_nc_Vn_11_12_2020, color =  
    'Vietnam')) +  
  labs(title = "Newcases in the last 2 months of  
    2020", x = "date", y = "new cases")
```



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

Kết quả



**Hình:** Biểu đồ thu thập nhiễm bệnh cho 2 tháng cuối năm



Động cơ nghiên cứu

Mục tiêu

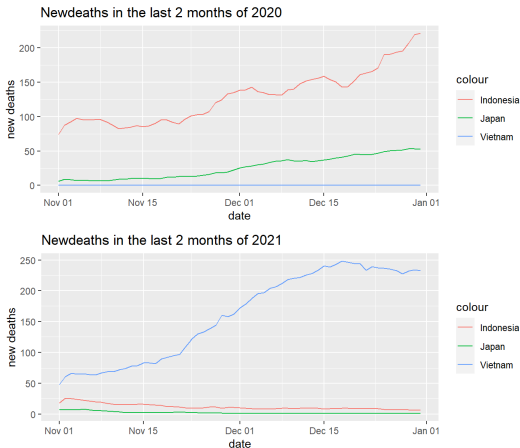
Kiến thức chuẩn bị

Nhiệm vụ

Kết luận

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

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



**Hình:** Biểu đồ thu thập tử vong cho 2 tháng cuối năm



Động cơ nghiên cứu

Mục tiêu

Kiến thức chuẩn bị

Nhiệm vụ

Kết luận

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

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

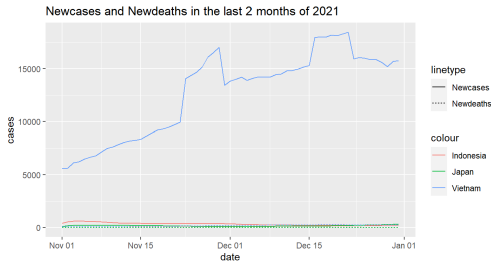
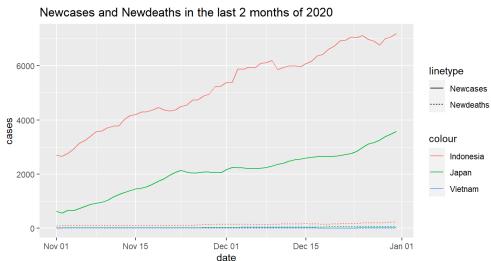
Kết hợp các đường biểu diễn ca nhiễm và các đường biểu diễn tử vong trong cùng một biểu đồ như sau:

```
New_11_12_2020 <- ggplot() +  
  geom_line(data=Indonesia_nc_11_12_2020, aes(x=  
    datetime, y=avg_nc_Indo_11_12_2020,  
    linetype = 'Newcases', color='Indonesia'))  
  +  
  geom_line(data=Japan_nc_11_12_2020, aes(x=  
    datetime, y=avg_nc_Jp_11_12_2020, linetype  
      = 'Newcases', color='Japan')) +  
  geom_line(data=Vietnam_nc_11_12_2020, aes(x=  
    datetime, y=avg_nc_Vn_11_12_2020, linetype  
      = 'Newcases', color='Vietnam')) +  
  ..... +  
  labs(title = "Newcases and Newdeaths in the  
    last 2 months of 2020", x = "date", y = "  
    cases")
```



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

## Kết quả





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

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

Với đề yêu cầu là biểu đồ thu thập tích lũy, chỉ khác một tí là ta sẽ vẽ dựa trên biến tích lũy đã tạo thay vì các biến giá trị trung bình như các câu trên.

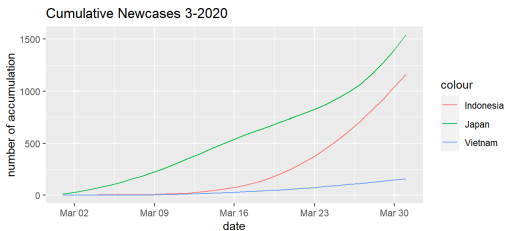
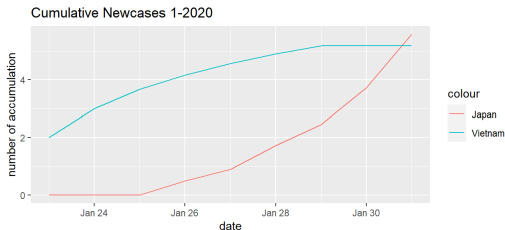
Ví dụ với tháng 1/2020:

```
Acml_Newcases_1_2020 <- ggplot() +  
  geom_line(data=Japan_nc_1_2020, aes(x=datetime  
    , y=acml_nc_Jp_1_2020, color = 'Japan')) +  
  geom_line(data=Vietnam_nc_1_2020, aes(x=  
    datetime, y=acml_nc_Vn_1_2020, color = '  
    Vietnam')) +  
  labs(title = "Cumulative Newcases 1-2020", x =  
    "date", y = "number of accumulation")
```

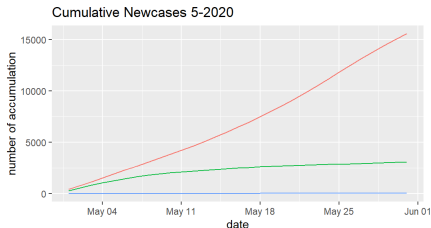
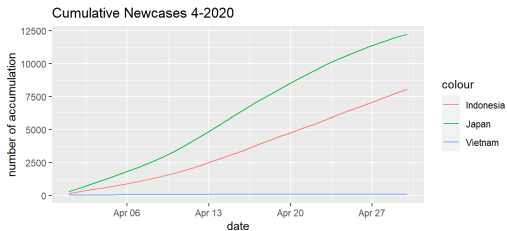


# Nhiệm vụ vi: Nhóm câu hỏi liên quan đến trực quan dữ liệu theo trung bình 7 ngày gần nhất

## Kết quả



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



Động cơ nghiên cứu

Mục tiêu

Kiến thức chuẩn bị

Nhiệm vụ

Kết luận

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

Thông kê khảo sát kết quả Covid-19

Huỳnh Tường Nguyên,  
Nguyễn Ngọc Lê



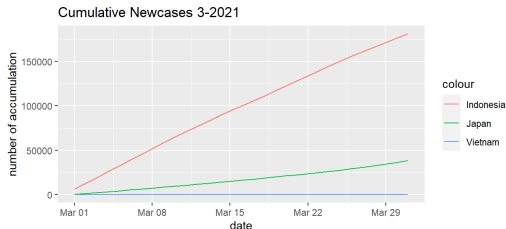
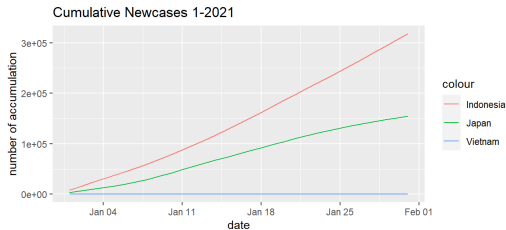
Động cơ nghiên cứu

Mục tiêu

Kiến thức chuẩn bị

Nhiệm vụ

Kết luận



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

Thông kê khảo sát kết quả Covid-19

Huỳnh Tường Nguyễn,  
Nguyễn Ngọc Lê



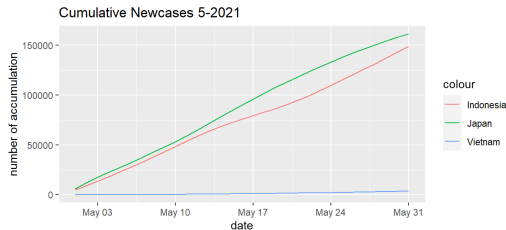
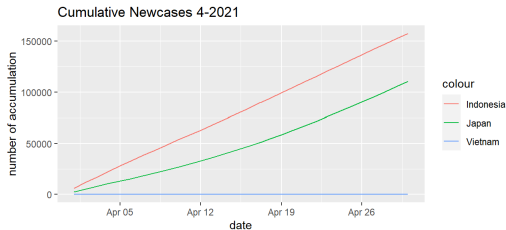
Động cơ nghiên cứu

Mục tiêu

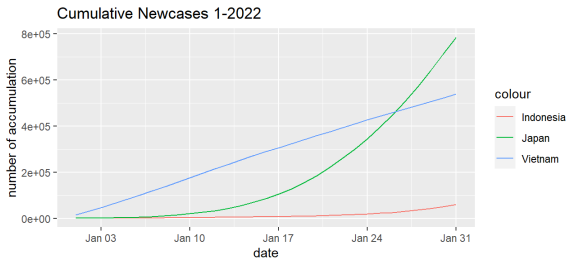
Kiến thức chuẩn bị

Nhiệm vụ

Kết luận



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



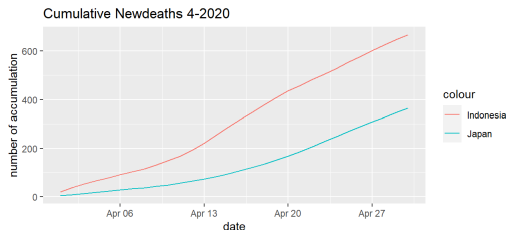
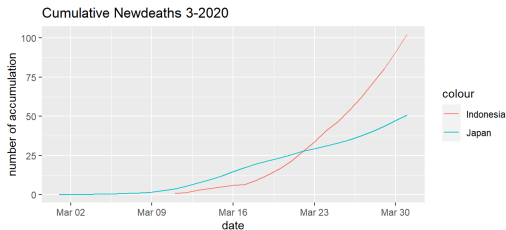
**Hình:** Biểu đồ thu thập nhiễm bệnh tích lũy cho từng tháng



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

⑧ Biểu đồ thu thập tử vong tích lũy cho từng tháng  
Thực hiện hoàn toàn tương tự như câu 7, thay dữ liệu *new\_cases* thành *new\_deaths*.

Kết quả



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

Thông kê khảo sát kết quả Covid-19

Huỳnh Tường Nguyễn,  
Nguyễn Ngọc Lê



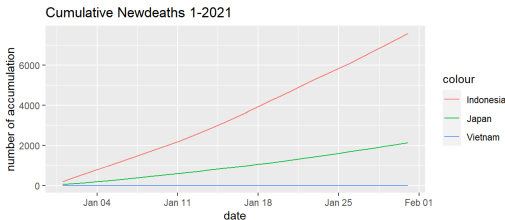
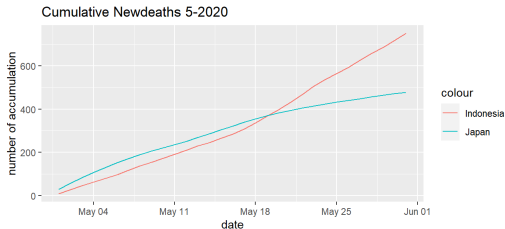
Động cơ nghiên cứu

Mục tiêu

Kiến thức chuẩn bị

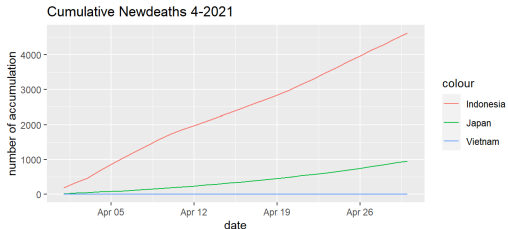
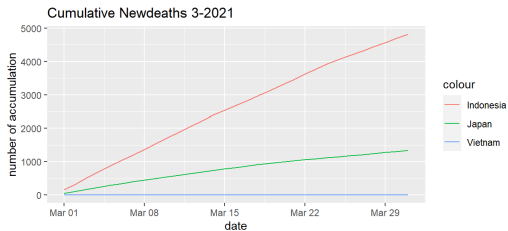
Nhiệm vụ

Kết luận





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



Thông kê khảo sát kết quả Covid-19

Huỳnh Tường Nguyễn,  
Nguyễn Ngọc Lê



Động cơ nghiên cứu

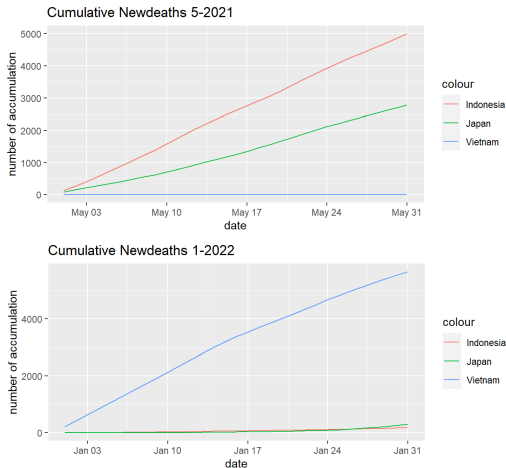
Mục tiêu

Kiến thức chuẩn bị

Nhiệm vụ

Kết luận

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



**Hình:** Biểu đồ thu thập tử vong tích lũy cho từng tháng



## Nhiệm vụ 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. Đây là nhóm câu hỏi liên quan đến tháng, nên bước đầu tiên ta đưa format chuẩn về ngày tháng năm để tiện xử lý.

```
dataFile$date <- strptime(dataFile$date, format="%m/%d  
/%Y")
```



## Nhiệm vụ vii: Nhóm câu hỏi liên quan đến tất cả quốc gia theo thời gian là tháng

- 1 Vẽ biểu đồ thể hiện thu thập dữ liệu nhiễm bệnh theo thời gian là tháng của tất cả quốc gia.

Với câu hỏi này, ta sử dụng hàm `sum()` với điều kiện để tính theo yêu cầu.

```
data_newcases <- c(rep(0,times=12))
data_newcases[1] <- sum(dataFile[which(dataFile$new_
cases>0 & dataFile$iso_code!="OWID_WRL" & format
(dataFile$date,"%Y")=="2020" & format(dataFile$
date,"%m")=="01"),5])
```

Những tháng sau hiện thực code hoàn toàn tương tự như trên.



## Nhiệm vụ vii: Nhóm câu hỏi liên quan đến tất cả quốc gia theo thời gian là tháng

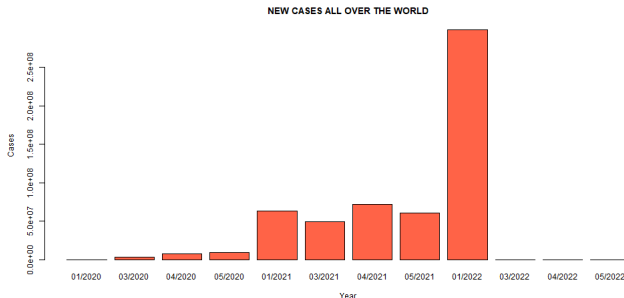
Khi đã tổng hợp dữ liệu, ta vẽ biểu đồ cột với hàm *barplot()* và xuất hình ảnh bằng hàm *png()*, cuối cùng kết thúc bằng hàm *dev.off()* để đóng file png.

```
png(file = "newcase.png",width=1000)
barplot(data_newcases ,
        main="NEW CASES ALL OVER THE WORLD",
        beside=TRUE,
        col="tomato",
        names.arg=c("01/2020","03/2020","04/2020","
                    05/2020","01/2021","03/2021","04/2021",
                    "05/2021","01/2022","03/2022","04/2022",
                    "05/2022"),
        ylab="Cases",
        xlab="Year",
    )
dev.off()
```



## Nhiệm vụ vii: Nhóm câu hỏi liên quan đến tất cả quốc gia theo thời gian là tháng

### Kết quả



**Hình:** *Biểu đồ nhiễm bệnh theo từng tháng của tất cả quốc gia*

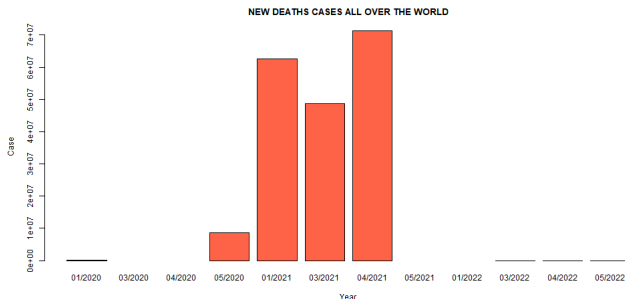


## Nhiệm vụ vii: Nhóm câu hỏi liên quan đến tất cả quốc gia theo thời gian là tháng

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

Tương tự với ý 1, ta chỉ cần thay vì thu thập dữ liệu *new\_cases*, ta sẽ lấy dữ liệu là *new\_deaths*.

Kết quả



**Hình:** Biểu đồ tử vong theo từng tháng của tất cả quốc gia



## Nhiệm vụ vii: Nhóm câu hỏi liên quan đến tất cả quốc gia theo thời gian là tháng

- 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.

Với câu này, trong hàm *sum()*, ta lấy điều kiện là tháng 11 và 12 của từng năm. Sau đó lưu vào một ma trận để vẽ biểu đồ.

```
data_newcases_2months_2020 <- sum(dataFile[which(
  dataFile$new_cases>0 &
  dataFile$iso_code!="OWID_WRL"&
  format(dataFile$date,"%Y")== "2020"&
  (format(dataFile$date,"%m")== "12" |
  format(dataFile$date,"%m")== "11")),5])
```

Những tháng sau hiện thực code hoàn toàn tương tự như trên.





## Nhiệm vụ vii: Nhóm câu hỏi liên quan đến tất cả quốc gia theo thời gian là tháng

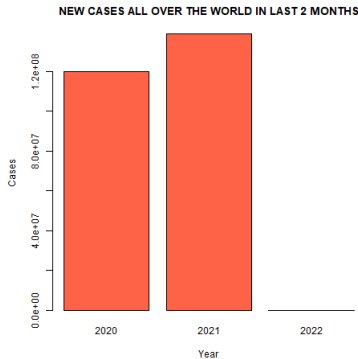
Ta tiếp tục vẽ biểu đồ bằng hàm barplot và xuất ra file png.

```
png(file="vii3.png")
barplot(data_newcases_2months,
        main="NEW CASES ALL OVER THE WORLD IN LAST 2
              MONTHS",
        col="tomato",
        ylab="Cases",
        xlab="Year",
        names.arg=c("2020", "2021", "2022"),
        )
dev.off()
```



## Nhiệm vụ vii: Nhóm câu hỏi liên quan đến tất cả quốc gia theo thời gian là tháng

Kết quả:



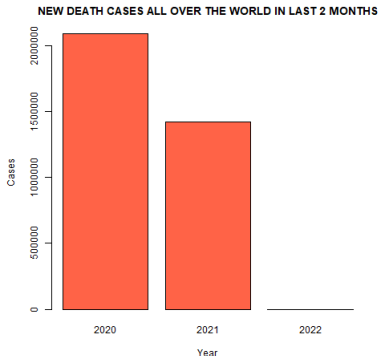
**Hình:** Biểu đồ nhiễm bệnh 2 tháng cuối của mỗi năm.



## Nhiệm vụ vii: Nhóm câu hỏi liên quan đến tất cả quốc gia theo thời gian là tháng

- 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.

Tương tự với ý 3, ta thay dữ liệu *new\_cases* thành *new\_deaths*.  
Kết quả



**Hình:** Biểu đồ tử vong 2 tháng cuối của mỗi năm.



## Nhiệm vụ vii: Nhóm câu hỏi liên quan đến tất cả quốc gia theo thời gian là tháng

- ⑤ Biểu đồ thể hiện thu thập dữ liệu nhiễm bệnh tương đối tích lũy 2 tháng cuối của năm của tất cả quốc gia.

Với bài toán tương đối tích lũy, ta sẽ tính dữ liệu cộng dồn.

```
data_newcases_20 <- c(0,0)
data_newcases_20[1] <- sum(dataFile[which(dataFile$
  new_cases>0 &
  dataFile$iso_code!="OWID_WRL" &
  format(dataFile$date,"%Y")== "2020" &
  format(dataFile$date,"%m")== "11"),5])
```

Những tháng sau hiện thực code hoàn toàn tương tự như trên.



## Nhiệm vụ vii: Nhóm câu hỏi liên quan đến tất cả quốc gia theo thời gian là tháng

Vì đây là nhiều vector nên ta tạo một dataframe lưu dữ liệu để vẽ biểu đồ

```
data_newcase_multi <- data.frame(data_newcases_20, data_newcases_21, data_newcases_22)
```

Cuối cùng là dùng hàm barplot để vẽ biểu đồ.

```
png(file="vii5.png")
barplot(as.matrix(data_newcase_multi),
        main="DATA NEW CASE MULTI",
        ylab="Cases", xlab="Year", beside=TRUE,
        col=c("tomato", "steelblue2"), legend.text=c(
            "November", "December"),
        args.legend=list(x="topright"), names.arg=c(
            "2020", "2021", "2022"),)
dev.off()
```



## Nhiệm vụ vii: Nhóm câu hỏi liên quan đến tất cả quốc gia theo thời gian là tháng

### Kết quả



**Hình:** Biểu đồ thể hiện nhiễm bệnh tương đối tích lũy 2 tháng cuối năm tất cả quốc gia.

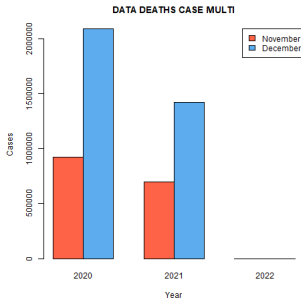


## Nhiệm vụ vii: Nhóm câu hỏi liên quan đến tất cả quốc gia theo thời gian là tháng

⑥ Biểu đồ thể hiện thu thập dữ liệu tử vong tương đối tích lũy theo thời gian là 2 tháng cuối của năm của tất cả quốc gia

Tương tự với ý 5, ta thay dữ liệu từ *new\_cases* thành *new\_deaths*.

Kết quả



**Hình:** Biểu đồ tử vong tương đối tích lũy 2 tháng cuối năm tất cả quốc gia.



## Nhiệm vụ viii: Nhóm câu hỏi liên quan đến tất cả quốc gia theo trung bình 7 ngày gần nhất

Đầu tiên ta sẽ trích xuất một file data để xử lý.

```
data_viii <- function(year) {  
  subset(dataFile, year(dataFile$date) == year &  
    (month(dataFile$date) == 1 | month(dataFile$date)  
      == 3 |  
    month(dataFile$date) == 4 | month(dataFile$date)  
      == 5 |  
    month(dataFile$date) == 11 | month(dataFile$date)  
      == 12))  
}  
  
data_viii_2020 <- data_viii(2020)
```





## Nhiệm vụ viii: Nhóm câu hỏi liên quan đến tất cả quốc gia theo trung bình 7 ngày gần nhất

- 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

Ta bắt đầu bằng việc tính tổng các ca nhiễm của tất cả quốc gia theo đơn vị ngày. Việc này được hỗ trợ bằng hàm *aggregate()*, lưu ý bỏ qua các giá trị trống NA.

```
sum_cases <- function(data) {  
  aggregate(x = data$new_cases, by = list(data$date),  
    FUN = sum, na.rm = TRUE)  
}
```

```
sum_cases_2020 <- sum_cases(data_viii_2020)  
names(sum_cases_2020)[1] = 'Date'
```

Ở trên ta đặt một cột có tên là *Date* để dễ dàng xử lý hơn.



## Nhiệm vụ viii: Nhóm câu hỏi liên quan đến tất cả quốc gia theo trung bình 7 ngày gần nhất

Để tính giá trị trung bình theo 7 ngày gần nhất, chúng ta có thể dùng hàm *rollapply()* có trong thư viện *zoo*.

```
avg_7d <- function(data){  
  data %>% group_by(format.Date(Date, "%Y/%m")) %>%  
    mutate(avg_7 = rollapply(x, width=7,  
      FUN=function(x) mean(na.omit(x)),  
      fill=NA, by=1, partial=TRUE, align="right"))  
}  
  
sum_cases_2020 <- avg_7d(sum_cases_2020)
```



## Nhiệm vụ viiii: Nhóm câu hỏi liên quan đến tất cả quốc gia theo trung bình 7 ngày gần nhất

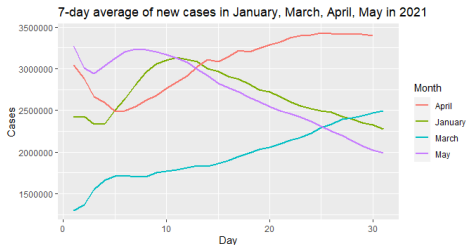
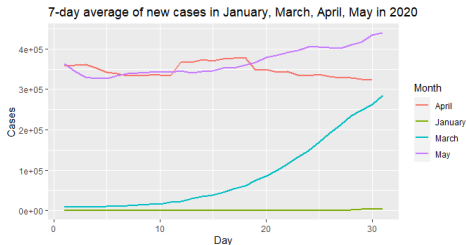
Cuối cùng ta cần vẽ đồ thị, dùng hàm *ggplot()*.

```
p <- function(data.fr, mth, str){  
  geom_line(data = subset(data.fr, month(data.fr$Date)  
    == mth),  
    mapping = aes(x=day(Date), y=avg_7, color=str),  
    size = 1)  
}  
  
p_cases_2020 <- ggplot() +  
  p(sum_cases_2020, 1, 'January') +  
  p(sum_cases_2020, 3, 'March') +  
  p(sum_cases_2020, 4, 'April') +  
  p(sum_cases_2020, 5, 'May') +  
  labs(title="7-day average of new cases in January,  
    March, April, May in 2020", x = "Day", y = "  
    Cases") +  
  scale_color_discrete(name="Month")
```



# Nhiệm vụ viii: Nhóm câu hỏi liên quan đến tất cả quốc gia theo trung bình 7 ngày gần nhất

## Kết quả



Động cơ nghiên cứu

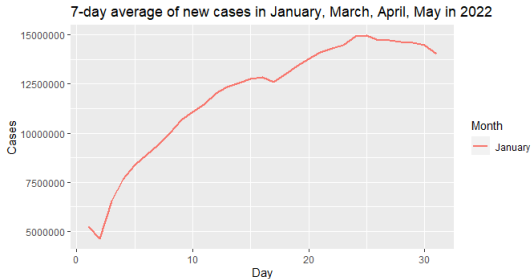
Mục tiêu

Kiến thức chuẩn bị

Nhiệm vụ

Kết luận

## Nhiệm vụ viii: Nhóm câu hỏi liên quan đến tất cả quốc gia theo trung bình 7 ngày gần nhất



**Hình:** Biểu đồ nhiễm bệnh theo trung bình 7 ngày gần nhất

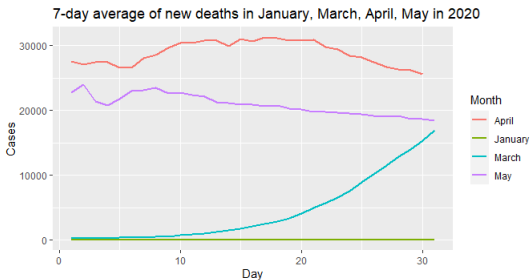


## Nhiệm vụ viii: Nhóm câu hỏi liên quan đến tất cả quốc gia theo trung bình 7 ngày gần nhất

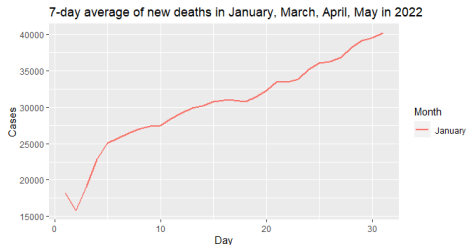
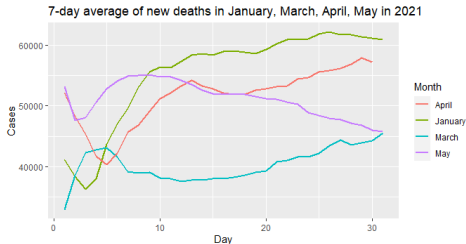
- ② Biểu đồ thể hiện thu thập dữ liệu tử vong theo thời gian là tháng của tất cả quốc gia theo trung bình 7 ngày gần nhất

Câu hỏi này cũng hoàn toàn tương tự như câu 1, chỉ đổi *new\_cases* thành *new\_deaths*.

Kết quả



## Nhiệm vụ viii: Nhóm câu hỏi liên quan đến tất cả quốc gia theo trung bình 7 ngày gần nhất



**Hình:** Biểu đồ tử vong theo trung bình 7 ngày gần nhất



## Nhiệm vụ viii: Nhóm câu hỏi liên quan đến tất cả quốc gia theo trung bình 7 ngày gần nhất

- 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 theo trung bình 7 ngày gần nhất

Câu hỏi này cũng tương tự như câu 1, chỉ đổi dữ liệu về tháng.

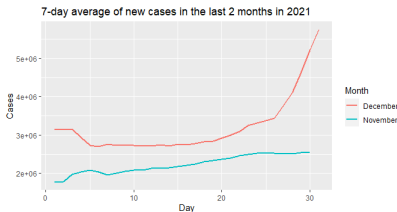
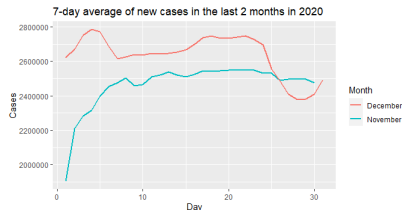
```
p_2last_cases_2020 <- ggplot() +  
  p(sum_cases_2020, 11, 'November') +  
  p(sum_cases_2020, 12, 'December') +  
  labs(title="7-day average of new cases in the last  
    2 months in 2020", x = "Day", y = "Cases") +  
  scale_color_discrete(name="Month")
```





# Nhiệm vụ viii: Nhóm câu hỏi liên quan đến tất cả quốc gia theo trung bình 7 ngày gần nhất

Kết quả



**Hình:** Biểu đồ nhiễm bệnh theo trung bình 7 ngày gần nhất trong 2 tháng cuối năm

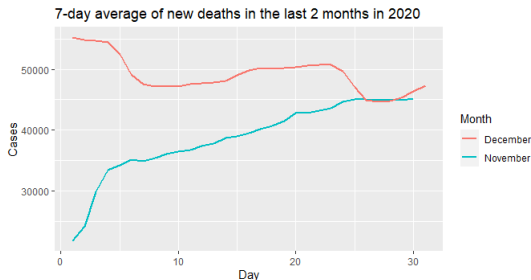


## Nhiệm vụ viii: Nhóm câu hỏi liên quan đến tất cả quốc gia theo trung bình 7 ngày gần nhất

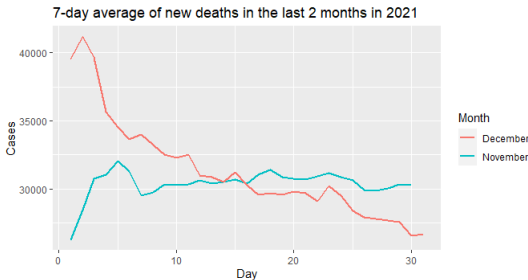
- ④ Biểu đồ thể hiện thu thập dữ liệu tử vong theo thời gian là 2 tháng cuối của năm của tất cả quốc gia theo trung bình 7 ngày gần nhất

Câu hỏi này tương tự như câu 3 ở trên, đổi *new\_cases* thành *new\_deaths*.

Kết quả



## Nhiệm vụ viii: Nhóm câu hỏi liên quan đến tất cả quốc gia theo trung bình 7 ngày gần nhất



**Hình:** Biểu đồ tử vong theo trung bình 7 ngày gần nhất trong 2 tháng cuối năm



## Nhiệm vụ viii: Nhóm câu hỏi liên quan đến tất cả quốc gia theo trung bình 7 ngày gần nhất

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

Với dữ liệu trung bình 7 ngày đã có sẵn ở trên, ta chỉ cần tính thêm dữ liệu tích lũy, việc này được thực hiện bằng hàm *cumsum()*.

```
sum_cases_2020 <- sum_cases_2020 %>%  
  mutate(cummulative = cumsum(avg_7))
```



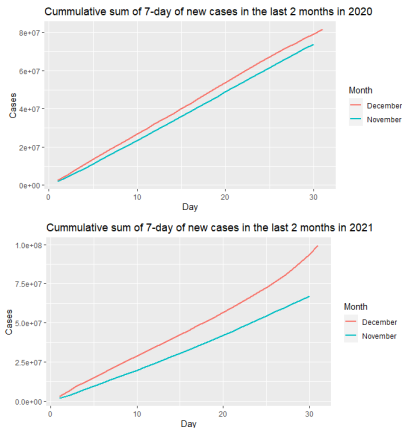
## Nhiệm vụ viiii: Nhóm câu hỏi liên quan đến tất cả quốc gia theo trung bình 7 ngày gần nhất

```
p_cum <- function(data.fr, mth, str){  
  geom_line(data = subset(data.fr, month(data.fr$Date)  
    == mth), mapping = aes(x=day(Date), y=  
    cumulative, color = str), size = 1)  
}  
  
p_cum_cases_2020 <- ggplot() +  
  p_cum(sum_cases_2020, 11, 'November') +  
  p_cum(sum_cases_2020, 12, 'December') +  
  labs(title="Cumulative sum of 7-day of new cases  
    in the last 2 months in 2020", x = "Day", y =  
    "Cases") +  
  scale_color_discrete(name="Month")
```



# Nhiệm vụ viii: Nhóm câu hỏi liên quan đến tất cả quốc gia theo trung bình 7 ngày gần nhất

Kết quả



**Hình:** Biểu đồ nhiễm bệnh tích lũy theo trung bình 7 ngày gần nhất trong 2 tháng cuối năm

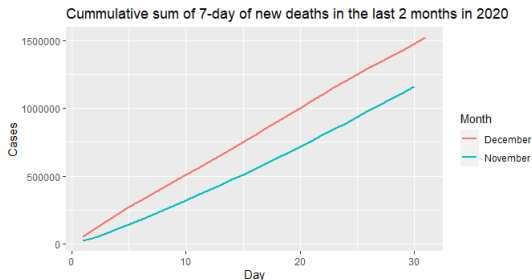


## Nhiệm vụ viii: Nhóm câu hỏi liên quan đến tất cả quốc gia theo trung bình 7 ngày gần nhất

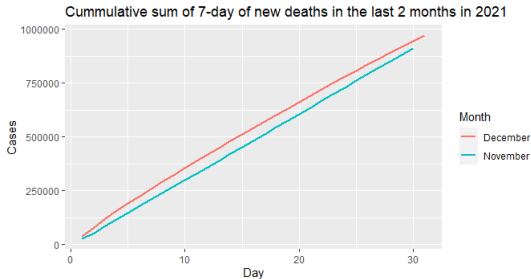
- ⑥ Biểu đồ thể hiện thu thập dữ liệu tử vong tích lũy theo thời gian là 2 tháng cuối của năm của tất cả quốc gia theo trung bình 7 ngày gần nhất

Câu hỏi này tương tự như câu 5 phía trên, đổi *new\_cases* thành *new\_deaths*.

Kết quả



## Nhiệm vụ viii: Nhóm câu hỏi liên quan đến tất cả quốc gia theo trung bình 7 ngày gần nhất



**Hình:** Biểu đồ tử vong tích lũy theo trung bình 7 ngày gần nhất trong 2 tháng cuối năm





## Nhiệm vụ ix: Nhóm câu hỏi liên quan đến sự tương quan giữa nhiễm bệnh và tử vong

- 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

Để thực hiện được yêu cầu trên ta cần phải tổng hợp số ca nhiễm và tử vong của 3 nước Việt Nam, Indonesia, Nhật bản từ dữ liệu ban đầu, lưu ý bỏ các giá trị trống NA.

```
sum_cases_VN <- sum(dataFile[which(dataFile$iso_code ==  
  "VNM" & dataFile$new_cases!=""),5])  
sum_deaths_VN <- sum(dataFile[which(dataFile$iso_code  
  == "VNM" & dataFile$new_deaths!=""),6])  
sum_cases_IDN <- sum(dataFile[which(dataFile$iso_code  
  == "IDN" & dataFile$new_cases!=""),5])  
sum_deaths_IDN <- sum(dataFile[which(dataFile$iso_code  
  == "IDN" & dataFile$new_deaths!=""),6])  
sum_cases_JPN <- sum(dataFile[which(dataFile$iso_code  
  == "JPN" & dataFile$new_cases!=""),5])  
sum_deaths_JPN <- sum(dataFile[which(dataFile$iso_code  
  == "JPN" & dataFile$new_deaths!=""),6])
```



## Nhiệm vụ ix: Nhóm câu hỏi liên quan đến sự tương quan giữa nhiễm bệnh và tử vong

Tiếp theo ta sẽ tính số ca nhiễm và tử vong theo tỉ lệ phần trăm dựa trên tổng số ca nhiễm và tử vong, sau đó dùng hàm *cumsum()* để tính phần trăm tích lũy.

```
VN <- subset(dataFile, dataFile$iso_code=="VNM")
VN <- as.data.frame(VN, stringsAsFactors = FALSE)
VN <- VN[order(as.Date(VN$date, format="%m/%d/%Y")),]
VN$iso_code <- NULL
VN$continent <- NULL
VN$location <- NULL
VN[is.na(VN)] <- 0
VN$new_cases <- VN$new_cases*100/sum_cases_VN
VN$new_cases <- cumsum(VN$new_cases)
VN$new_deaths <- VN$new_deaths*100/sum_deaths_VN
VN$new_deaths <- cumsum(VN$new_deaths)
```



## Nhiệm vụ ix: Nhóm câu hỏi liên quan đến sự tương quan giữa nhiễm bệnh và tử vong

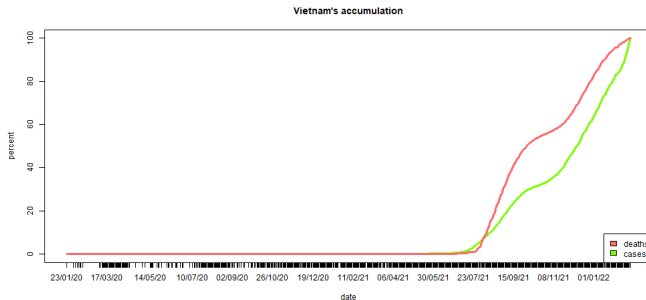
Cuối cùng ta sẽ vẽ đồ thị bằng hàm *plot()*.

```
VN_filler <- c("indianred1", "lawngreen")
VN_label <- c("deaths", "cases")
VN$date <- as.Date(VN$date, "%m/%d/%Y")
plot(VN$new_cases ~ VN$date, xlab = "date", ylab = "
    percent", type = "l", lwd = 3, xaxt = "n", col = "
    lawngreen", main = "Vietnam's accumulation")
lines(VN$new_deaths ~ VN$date, type = "l", lwd = 3,
    xaxt = "n", col = "indianred1")
axis(1, VN$date, format(VN$date, "%d/%m/%y"))
legend("bottomright", VN_label, fill = VN_filler)
```



# Nhiệm vụ ix: Nhóm câu hỏi liên quan đến sự tương quan giữa nhiễm bệnh và tử vong

## Kết quả



## Nhiệm vụ ix: Nhóm câu hỏi liên quan đến sự tương quan giữa nhiễm bệnh và tử vong

Thông kê khảo sát kết quả Covid-19

Huỳnh Tường Nguyễn,  
Nguyễn Ngọc Lê



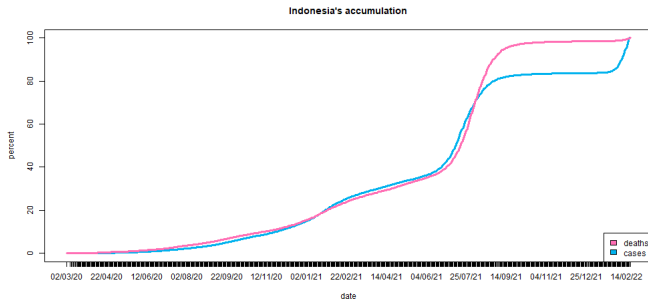
Động cơ nghiên cứu

Mục tiêu

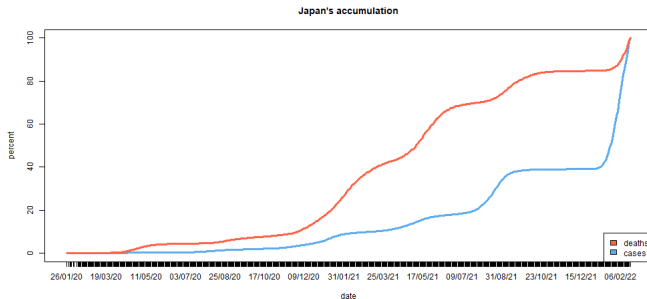
Kiến thức chuẩn bị

Nhiệm vụ

Kết luận



## Nhiệm vụ ix: Nhóm câu hỏi liên quan đến sự tương quan giữa nhiễm bệnh và tử vong



**Hình:** Biểu đồ tích lũy ca nhiễm và tử vong theo thời gian



## Nhiệm vụ ix: Nhóm câu hỏi liên quan đến sự tương quan giữa nhiễm bệnh và tử vong

Câu 2, 3: Trên từng quốc gia riêng của nhóm 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.

### ② Xét tương quan trong mỗi tháng.

Trước tiên ta sẽ lập 3 bảng số liệu của 3 nước về số ca nhiễm và tử vong:

```
VN_2 <- subset(dataFile, dataFile$iso_code=="VNM")
VN_2 <- as.data.frame(VN_2, stringsAsFactors = FALSE)
VN_2 <- VN_2[order(as.Date(VN_2$date, format="%m/%d/%Y"),),]
VN_2$date <- as.Date(VN_2$date, "%m/%d/%Y")
VN_2$iso_code <- NULL
VN_2$continent <- NULL
VN_2$location <- NULL
VN_2[is.na(VN_2)] <- 0
```



## Nhiệm vụ ix: Nhóm câu hỏi liên quan đến sự tương quan giữa nhiễm bệnh và tử vong

Sau đó từ bảng số liệu trên chúng ta sẽ lọc ra những tháng cần vẽ đồ thị:

```
VN_01_2020 <- subset(VN_2, format(date, "%m-%Y")==  
  "01-2020")  
IDN_01_2020 <- subset(IDN_2, format(date, "%m-%Y")==  
  "01-2020")  
JPN_01_2020 <- subset(JPN_2, format(date, "%m-%Y")==  
  "01-2020")
```

Tiếp theo là vẽ đồ thị cho từng tháng:

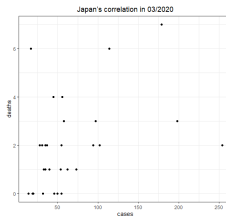
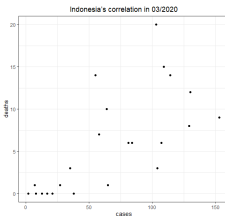
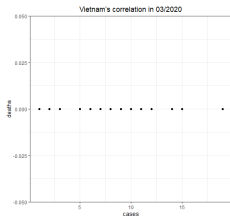
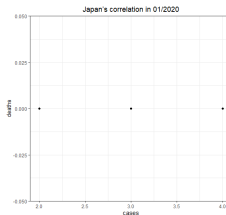
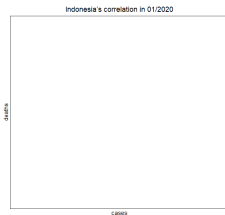
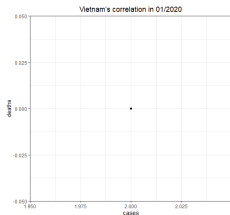
```
r_VN_01_2020 <- cor(VN_01_2020$new_cases, VN_01_2020$  
  new_deaths)  
plot1 <- ggplot(VN_01_2020, aes(x=new_cases, y=new_  
  deaths)) +  
  geom_point() +  
  theme_bw() +  
  xlab("cases") +  
  ylab("deaths") +  
  ggtitle(paste0("Vietnam's correlation in 01/2020")  
    ) +  
  theme(plot.title = element_text(hjust = 0.5))  
plot1
```





# Nhiệm vụ ix: Nhóm câu hỏi liên quan đến sự tương quan giữa nhiễm bệnh và tử vong

## Kết quả



Động cơ nghiên cứu

Mục tiêu

Kiến thức chuẩn bị

Nhiệm vụ

Kết luận

# Nhiệm vụ ix: Nhóm câu hỏi liên quan đến sự tương quan giữa nhiễm bệnh và tử vong

Thống kê khảo sát kết quả Covid-19

Huỳnh Tường Nguyễn,  
Nguyễn Ngọc Lê



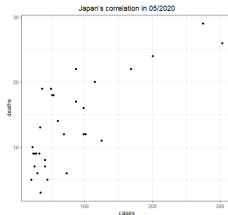
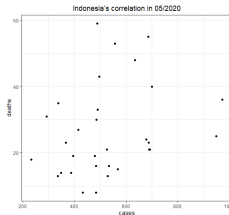
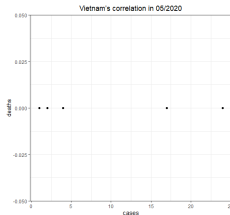
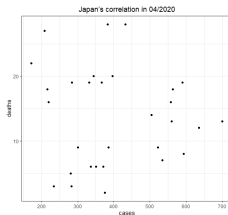
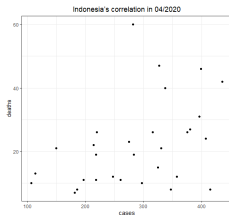
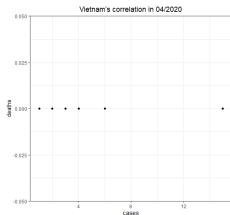
Động cơ nghiên cứu

Mục tiêu

Kiến thức chuẩn bị

Nhiệm vụ

Kết luận



# Nhiệm vụ ix: Nhóm câu hỏi liên quan đến sự tương quan giữa nhiễm bệnh và tử vong

Thống kê khảo sát kết quả Covid-19

Huỳnh Tường Nguyễn,  
Nguyễn Ngọc Lê



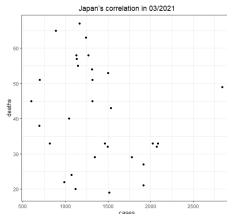
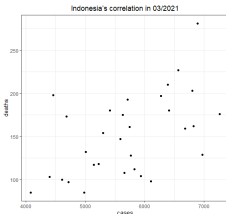
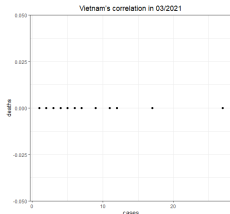
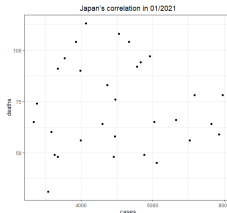
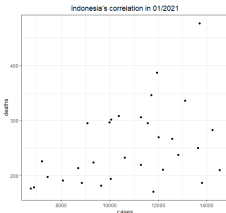
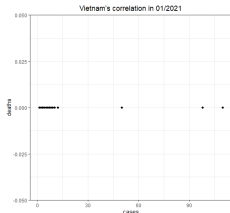
Động cơ nghiên cứu

Mục tiêu

Kiến thức chuẩn bị

Nhiệm vụ

Kết luận



# Nhiệm vụ ix: Nhóm câu hỏi liên quan đến sự tương quan giữa nhiễm bệnh và tử vong

Thống kê khảo sát kết quả Covid-19

Huynh Tuong Nguyen,  
Nguyen Ngoc Le



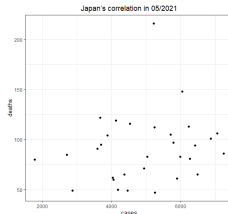
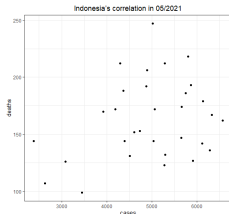
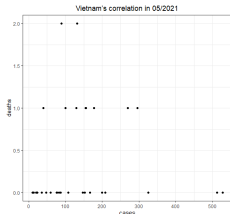
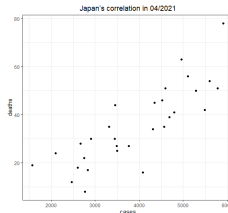
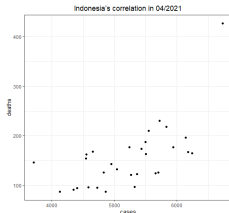
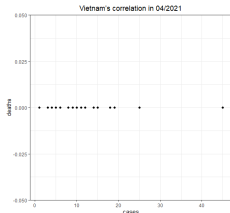
Động cơ nghiên cứu

Mục tiêu

Kiến thức chuẩn bị

Nhiệm vụ

Kết luận



## Nhiệm vụ ix: Nhóm câu hỏi liên quan đến sự tương quan giữa nhiễm bệnh và tử vong

Thông kê khảo sát kết quả Covid-19

Huynh Tuong Nguyen,  
Nguyen Ngoc Le



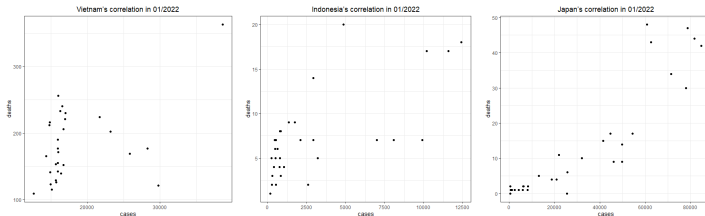
Động cơ nghiên cứu

Mục tiêu

Kiến thức chuẩn bị

Nhiệm vụ

Kết luận



**Hình:** Biểu đồ thể hiện tương quan mỗi tháng của từng quốc gia

## Nhiệm vụ ix: Nhóm câu hỏi liên quan đến sự tương quan giữa nhiễm bệnh và tử vong

Cuối cùng là xét hệ số tương quan của mỗi tháng:

```
cat("He so tuong quan cua VN thang 01/2020: ", r_VN_01_2020, "\n")
cat("He so tuong quan cua IDN thang 01/2020: ", r_IDN_01_2020, "\n")
cat("He so tuong quan cua JPN thang 01/2020: ", r_JPN_01_2020, "\n")
```

### Kết quả

```
He so tuong quan cua VN thang 01/2020: NA
He so tuong quan cua IDN thang 01/2020: NA
He so tuong quan cua JPN thang 01/2020: NA
He so tuong quan cua VN thang 03/2020: NA
He so tuong quan cua IDN thang 03/2020: 0.7270849
He so tuong quan cua JPN thang 03/2020: 0.3999013
He so tuong quan cua VN thang 04/2020: NA
He so tuong quan cua IDN thang 04/2020: 0.416476
He so tuong quan cua JPN thang 04/2020: -0.06970152
He so tuong quan cua VN thang 05/2020: NA
He so tuong quan cua IDN thang 05/2020: 0.2708646
He so tuong quan cua JPN thang 05/2020: 0.7601386
```



## Nhiệm vụ ix: Nhóm câu hỏi liên quan đến sự tương quan giữa nhiễm bệnh và tử vong

He so tuong quan cua VN thang 01/2021:	NA
He so tuong quan cua IDN thang 01/2021:	0.419901
He so tuong quan cua JPN thang 01/2021:	-0.008507115
He so tuong quan cua VN thang 03/2021:	NA
He so tuong quan cua IDN thang 03/2021:	0.5284843
He so tuong quan cua JPN thang 03/2021:	-0.2413034
He so tuong quan cua VN thang 04/2021:	NA
He so tuong quan cua IDN thang 04/2021:	0.6400672
He so tuong quan cua JPN thang 04/2021:	0.825705
He so tuong quan cua VN thang 05/2021:	0.009510053
He so tuong quan cua IDN thang 05/2021:	0.2896207
He so tuong quan cua JPN thang 05/2021:	0.1818131
He so tuong quan cua VN thang 01/2022:	0.4631564
He so tuong quan cua IDN thang 01/2022:	0.688997
He so tuong quan cua JPN thang 01/2022:	0.9020827

**Chú ý:** Những hệ số NA là do trong tháng không có ca tử vong nào



## Nhiệm vụ ix: Nhóm câu hỏi liên quan đến sự tương quan giữa nhiễm bệnh và tử vong

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

Tương tự như câu 2, nhưng ở đây chúng ta sẽ tạo một hàm để tìm giá trị trung bình 7 ngày gần nhất:

```
avrg_7_days <- function(data1, data2){  
  for (i in 1:length(data1)) {  
    count = 1  
    for (j in (i-1):(i-6)) {  
      if (j>0) {  
        count <- count+1  
        data2[i] <- data2[i] + data1[j]  
      }  
    }  
    data2[i] <- data2[i]/count  
  }  
  return(data2)  
}
```





## Nhiệm vụ ix: Nhóm câu hỏi liên quan đến sự tương quan giữa nhiễm bệnh và tử vong

Sử dụng hàm đã tạo ở trên để lập bảng dữ liệu, các bước tiếp theo làm như câu 2.

```
VN_01_2020_avrg <- VN_01_2020
VN_01_2020_avrg <- as.data.frame(VN_01_2020_avrg,
  stringsAsFactors = FALSE)
VN_01_2020_avrg$new_cases <- avrg_7_days(VN_01_2020$
  new_cases, VN_01_2020_avrg$new_cases)
VN_01_2020_avrg$new_deaths <- avrg_7_days(VN_01_2020$
  new_deaths, VN_01_2020_avrg$new_deaths)
```

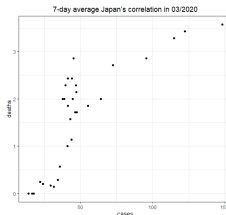
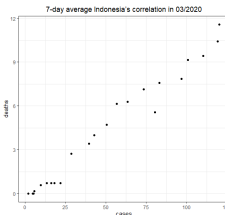
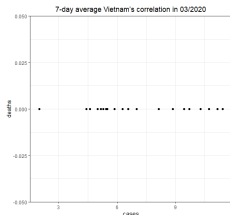
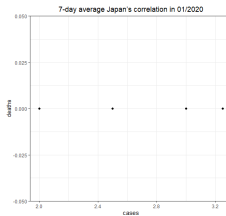
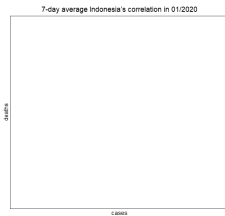
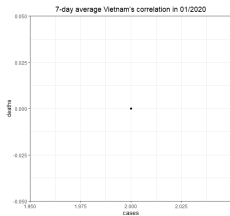
Vẽ đồ thị:

```
r_VN_01_2020_avrg <- cor(VN_01_2020_avrg$new_cases, VN
  _01_2020_avrg$new_deaths)
plot1 <- ggplot(VN_01_2020_avrg, aes(x=new_cases, y=
  new_deaths)) +
  geom_point() +
  theme_bw() +
  xlab("cases") +
  ylab("deaths") +
  ggtitle(paste0("7-day average Vietnam's
    correlation in 01/2020")) +
  theme(plot.title = element_text(hjust = 0.5))
plot1
```



# Nhiệm vụ ix: Nhóm câu hỏi liên quan đến sự tương quan giữa nhiễm bệnh và tử vong

## Kết quả



# Nhiệm vụ ix: Nhóm câu hỏi liên quan đến sự tương quan giữa nhiễm bệnh và tử vong

Thông kê khảo sát kết quả Covid-19

Huynh Tuong Nguyen,  
Nguyen Ngoc Le



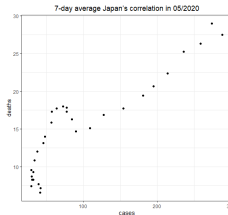
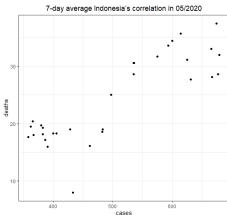
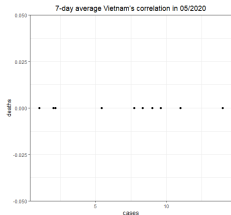
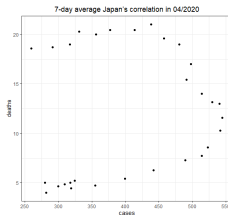
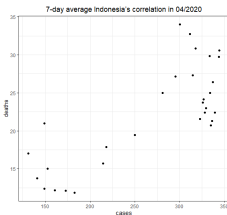
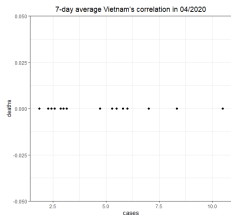
Động cơ nghiên cứu

Mục tiêu

Kiến thức chuẩn bị

Nhiệm vụ

Kết luận



# Nhiệm vụ ix: Nhóm câu hỏi liên quan đến sự tương quan giữa nhiễm bệnh và tử vong

Thống kê khảo sát kết quả Covid-19

Huynh Tuong Nguyen,  
Nguyen Ngoc Le



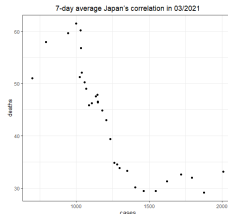
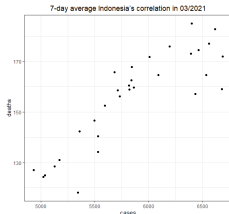
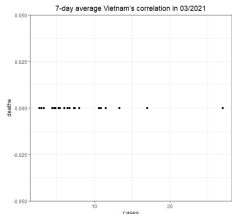
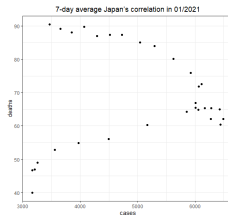
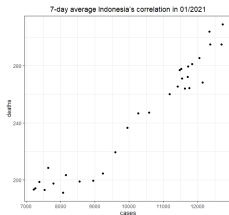
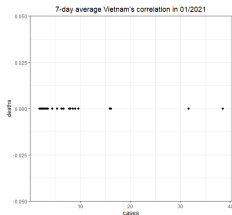
Động cơ nghiên cứu

Mục tiêu

Kiến thức chuẩn bị

Nhiệm vụ

Kết luận



# Nhiệm vụ ix: Nhóm câu hỏi liên quan đến sự tương quan giữa nhiễm bệnh và tử vong

Thống kê khảo sát kết quả Covid-19

Huỳnh Tường Nguyễn,  
Nguyễn Ngọc Lê



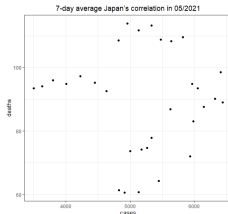
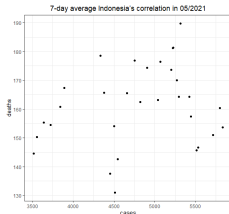
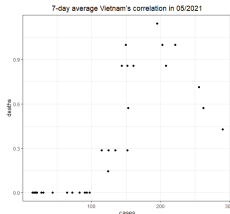
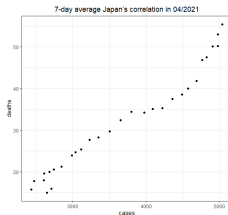
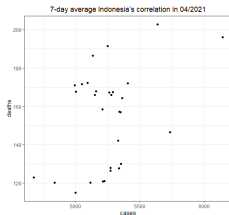
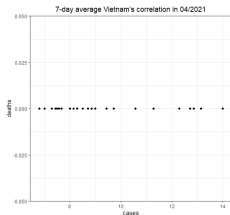
Động cơ nghiên cứu

Mục tiêu

Kiến thức chuẩn bị

Nhiệm vụ

Kết luận



## Nhiệm vụ ix: Nhóm câu hỏi liên quan đến sự tương quan giữa nhiễm bệnh và tử vong

Thông kê khảo sát kết quả Covid-19

Huỳnh Tường Nguyễn,  
Nguyễn Ngọc Lê



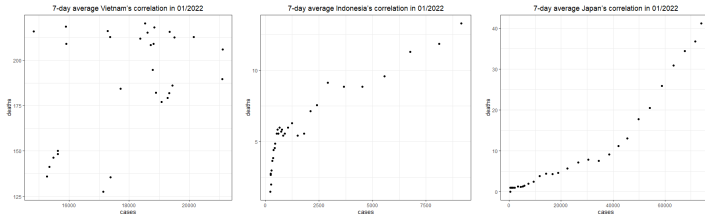
Động cơ nghiên cứu

Mục tiêu

Kiến thức chuẩn bị

Nhiệm vụ

Kết luận



**Hình:** Biểu đồ thể hiện tương quan mỗi tháng của từng quốc gia theo trung bình 7 ngày gần nhất

## Nhiệm vụ ix: Nhóm câu hỏi liên quan đến sự tương quan giữa nhiễm bệnh và tử vong

Tìm hệ số tương quan:

```
cat("He so tuong quan cua VN thang 01/2020 theo trung  
  binh 7 ngay gan nhat: ", r_VN_01_2020_avrg, "\n")  
cat("He so tuong quan cua IDN thang 01/2020 theo trung  
  binh 7 ngay gan nhat: ", r_IDN_01_2020_avrg, "\n")  
cat("He so tuong quan cua JPN thang 01/2020 theo trung  
  binh 7 ngay gan nhat: ", r_JPN_01_2020_avrg, "\n")
```

Kết quả

```
He so tuong quan cua VN thang 01/2020 theo trung binh  
 7 ngay gan nhat:  NA  
He so tuong quan cua IDN thang 01/2020 theo trung binh  
 7 ngay gan nhat:  NA  
He so tuong quan cua JPN thang 01/2020 theo trung binh  
 7 ngay gan nhat:  NA  
He so tuong quan cua VN thang 03/2020 theo trung binh  
 7 ngay gan nhat:  NA  
He so tuong quan cua IDN thang 03/2020 theo trung binh  
 7 ngay gan nhat:  0.9879779  
He so tuong quan cua JPN thang 03/2020 theo trung binh  
 7 ngay gan nhat:  0.7937885
```



## Nhiệm vụ ix: Nhóm câu hỏi liên quan đến sự tương quan giữa nhiễm bệnh và tử vong

He so tuong quan cua VN thang 04/2020 theo trung binh  
7 ngay gan nhat: NA

He so tuong quan cua IDN thang 04/2020 theo trung binh  
7 ngay gan nhat: 0.7776329

He so tuong quan cua JPN thang 04/2020 theo trung binh  
7 ngay gan nhat: 0.1518605

He so tuong quan cua VN thang 05/2020 theo trung binh  
7 ngay gan nhat: NA

He so tuong quan cua IDN thang 05/2020 theo trung binh  
7 ngay gan nhat: 0.8449888

He so tuong quan cua JPN thang 05/2020 theo trung binh  
7 ngay gan nhat: 0.9083547

He so tuong quan cua VN thang 01/2021 theo trung binh  
7 ngay gan nhat: NA

He so tuong quan cua IDN thang 01/2021 theo trung binh  
7 ngay gan nhat: 0.9683285

He so tuong quan cua JPN thang 01/2021 theo trung binh  
7 ngay gan nhat: 0.09324637

He so tuong quan cua VN thang 03/2021 theo trung binh  
7 ngay gan nhat: NA

He so tuong quan cua IDN thang 03/2021 theo trung binh  
7 ngay gan nhat: 0.8624254

He so tuong quan cua JPN thang 03/2021 theo trung binh  
7 ngay gan nhat: -0.823936





## Nhiệm vụ ix: Nhóm câu hỏi liên quan đến sự tương quan giữa nhiễm bệnh và tử vong

He so tuong quan cua VN thang 04/2021 theo trung binh  
7 ngay gan nhat: NA

He so tuong quan cua IDN thang 04/2021 theo trung binh  
7 ngay gan nhat: 0.4009489

He so tuong quan cua JPN thang 04/2021 theo trung binh  
7 ngay gan nhat: 0.980602

He so tuong quan cua VN thang 05/2021 theo trung binh  
7 ngay gan nhat: 0.764611

He so tuong quan cua IDN thang 05/2021 theo trung binh  
7 ngay gan nhat: 0.2233713

He so tuong quan cua JPN thang 05/2021 theo trung binh  
7 ngay gan nhat: -0.03916217

He so tuong quan cua VN thang 01/2022 theo trung binh  
7 ngay gan nhat: 0.4266099

He so tuong quan cua IDN thang 01/2022 theo trung binh  
7 ngay gan nhat: 0.9194707

He so tuong quan cua JPN thang 01/2022 theo trung binh  
7 ngay gan nhat: 0.9505387

**Chú ý:** Những hệ số NA là do trong tháng không có ca tử vong nào



## Nhiệm vụ x: Nhóm câu hỏi riêng

- ① So sánh tình trạng nhiễm bệnh của các quốc gia trong 7 ngày cuối của năm cuối cùng

Dễ dàng nhận thấy 7 ngày cần khảo sát là từ ngày 13/2/2022 cho đến ngày 19/2/2022.

Ta lọc dữ liệu từ các ngày đó như sau:

```
Indo_nc_last_7d <- na.omit(Indonesia_nc[Indonesia_nc$  
  datetime >= "2022-02-13" & Indonesia_nc$datetime  
  <= "2022-02-19",])
```

Các quốc gia khác tương tự.

Để trực quan, ta vẽ biểu đồ:

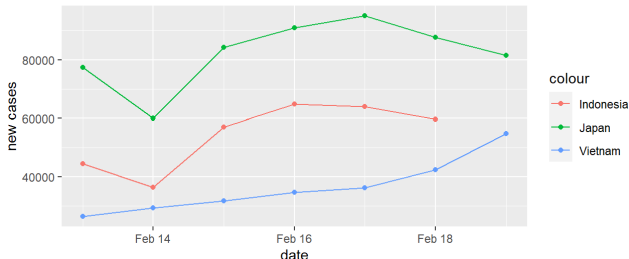
```
Newcases_last_7d <- ggplot() +  
  geom_line(data=Indo_nc_last_7d, aes(x=datetime, y=new_  
    cases, color = 'Indonesia')) + geom_point(data=  
    Indo_nc_last_7d, aes(x=datetime, y=new_cases,  
    color = 'Indonesia')) +  
  labs(title = "Newcases in the last 7 days", x = "date"  
    , y = "new cases")
```



## Nhiệm vụ x: Nhóm câu hỏi riêng



Newcases in the last 7 days



### Nhận xét

- Trong 3 nước thì nước có số ca mắc cao nhất từng ngày là Nhật Bản và thấp nhất là Việt Nam.
- Indonesia và Nhật Bản có cả sự tăng và giảm số ca nhiễm bệnh qua từng ngày.
- Việt Nam có số ca nhiễm bệnh tăng qua từng ngày.
- Số ca nhiễm bệnh đạt cực tiểu vào ngày 14/2/2022 đối với Indonesia và Nhật Bản, đạt bé nhất vào ngày 13/2/2022 đối với Việt Nam.

## Nhiệm vụ x: Nhóm câu hỏi riêng



### Nhận xét

- Số ca nhiễm bệnh của Indonesia đạt cực đại vào ngày 16/2/2022, của Nhật Bản đạt cực đại vào ngày 17/2/2022 và của Việt Nam đạt lớn nhất vào ngày 19/2/2022.
- Số ca nhiễm tăng mạnh nhất từ ngày 14/2/2022 sang ngày 15/2/2022 đối với Indonesia và Nhật Bản, từ ngày 18/2/2022 sang ngày 19/2/2022 đối với Việt Nam
- Nhìn chung, số ca nhiễm của Nhật Bản mỗi ngày gần gấp đôi của Indonesia và gấp 3 Việt Nam.
- Tại ngày cuối cùng được ghi nhận (19/2/2022), số ca nhiễm của Việt Nam đang có xu hướng tăng, số ca nhiễm của Nhật Bản đang có xu hướng giảm.

## Nhiệm vụ x: Nhóm câu hỏi riêng

- 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.

Về sự biến thiên của các đường trong đồ thị tích lũy:

- Khi tỉ lệ tích lũy giảm thì tại vị trí đó, đường tích lũy sẽ cong vồng lên phía trên (bề lõm hướng xuống). Giảm mạnh thì độ cong càng lớn.
- Ngược lại khi tỉ lệ tích lũy tăng thì tại vị trí đó đường tích lũy sẽ cong vồng xuống phía dưới (bề lõm hướng lên). Tăng mạnh thì độ cong càng lớn.

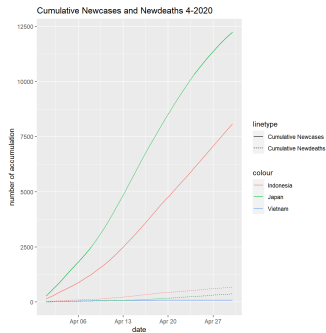
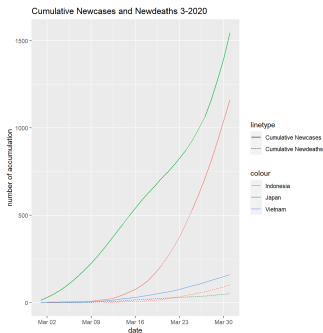
Dựa vào đó, ta vẽ các đường tích lũy của ca nhiễm và tử vong trong cùng một biểu đồ để dễ quan sát và so sánh.

```
Acml_3_2020 <- ggplot() +  
  geom_line(data=Indonesia_nc_3_2020, aes(x=datetime,  
    y=acml_nc_Indo_3_2020, color = 'Indonesia',  
    linetype = 'Cumulative Newcases')) +  
  labs(title = "Cumulative Newcases and Newdeaths  
    3-2020", x = "date", y = "number of accumulation  
    ")
```



# Nhiệm vụ x: Nhóm câu hỏi riêng

## Kết quả



# Nhiệm vụ x: Nhóm câu hỏi riêng

Thông kê khảo sát kết quả Covid-19

Huỳnh Tường Nguyễn,  
Nguyễn Ngọc Lê



Động cơ nghiên cứu

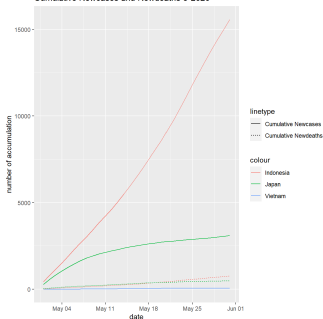
Mục tiêu

Kiến thức chuẩn bị

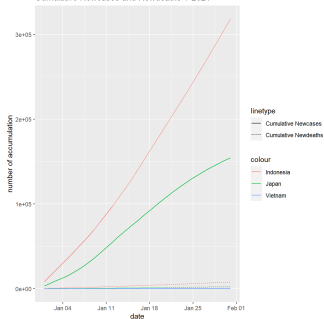
Nhiệm vụ

Kết luận

Cumulative Newcases and Newdeaths 5-2020



Cumulative Newcases and Newdeaths 1-2021



# Nhiệm vụ x: Nhóm câu hỏi riêng

Thống kê khảo sát kết quả Covid-19

Huỳnh Tường Nguyễn,  
Nguyễn Ngọc Lê



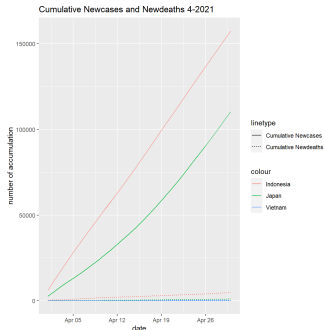
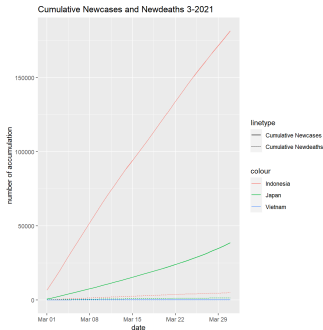
Động cơ nghiên cứu

Mục tiêu

Kiến thức chuẩn bị

Nhiệm vụ

Kết luận





# Nhiệm vụ x: Nhóm câu hỏi riêng

Thông kê khảo sát kết quả Covid-19

Huynh Tuong Nguyen,  
Nguyen Ngoc Le



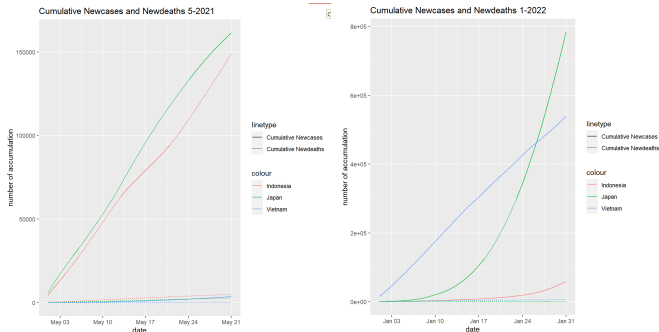
Động cơ nghiên cứu

Mục tiêu

Kiến thức chuẩn bị

Nhiệm vụ

Kết luận



**Hình:** Biểu đồ tử vong tích lũy và nhiễm bệnh tích lũy

## Nhiệm vụ x: Nhóm câu hỏi riêng

Từ các biểu đồ trên, ta rút ra được các khoảng thời gian thỏa mãn điều kiện đề bài như sau:

- Indonesia: Từ 14/3/2020 đến 17/3/2020; Từ 17/5/2021 đến 31/5/2021
- Japan: Từ 13/4/2020 đến 27/4/2020; Từ 18/1/2021 đến 31/1/2021; Tháng 3/2021
- Vietnam: Không xác định được



## Nhiệm vụ x: Nhóm câu hỏi riêng

- ④ Với  $k$  là mốc bùng phát dịch, hãy xác định  $k$  và cho biết các khoảng thời gian bùng phát

Ta chọn mốc  $k=10000$  là mốc bùng dịch. Tại các thời điểm có ca mới nhiễm bệnh lớn hơn mốc  $k$  thì đây sẽ là mốc bùng phát. Đầu tiên ta tính tổng các ca nhiễm theo ngày bằng hàm *aggregate()*, sau đó xử lý dữ liệu bằng dòng *while* đơn giản.



## Nhiệm vụ x: Nhóm câu hỏi riêng

```
data_x_4 <- aggregate(x = dataFile$new_cases, by =  
  list(dataFile$date), FUN = sum, na.rm = TRUE)  
data_x_4 <- data_x_4[order(as.Date(data_x_4$Group.1,  
  format="%m/%d/%Y")),]  
  
i=1  
k=10000  
while (i < nrow(data_x)) {  
  if (i == 1) cat ('Ngày bat dau    ', 'Ngày ket thuc',  
    "\n")  
  temp = c()  
  if (data_x_4[i,2] >= k) {  
    temp = cbind(temp, toString(data_x_4[i,1]))  
    while (data_x_4[i,2] >= k) {  
      i = i + 1  
      if (i > nrow(data_x_4)) break  
    }  
    temp = cbind(temp, "    ", toString(data_x_4[i  
      -1,1]))  
    cat(temp, "\n")  
  }  
  else i = i + 1  
}
```



## Nhiệm vụ x: Nhóm câu hỏi riêng

Kết quả:

Ngày bắt đầu	Ngày kết thúc
1/28/2020	1/28/2020
2/2/2020	2/10/2020
2/13/2020	2/14/2020
3/1/2020	3/1/2020
3/3/2020	2/19/2022

**Nhận xét:** với cách chọn mốc bùng dịch này, ta có thể thấy rằng dịch đã bùng lên ở một khoảng rất lâu, từ đầu năm 2020 đến đầu năm 2022.



## Nhiệm vụ x: Nhóm câu hỏi riêng

- ⑤ Với  $k$  là mốc bùng tử vong, hãy xác định  $k$  và cho biết các khoảng thời gian bùng phát

Tương tự với ý 4, ta chọn  $k=10000$  là mốc bùng phát và thay dữ liệu *new\_cases* thành *new\_deaths*.

Kết quả:

Ngày bắt đầu	Ngày kết thúc
3/24/2020	5/24/2020
5/26/2020	2/19/2022


**Nhận xét:** với cách chọn mốc bùng tử vong này, cũng giống như bùng dịch, ta có thể thấy rằng số ca tử vong đã bùng lên ở một khoảng rất lâu, từ giữa năm 2020 đến đầu năm 2022.






- Nhóm cơ bản thực hiện được các thao tác với ngôn ngữ R để thực hành phân tích và thống kê dữ liệu một cách khoa học, chính xác hơn.
- Thông qua các câu hỏi được đưa ra, ta phần nào có được một cái nhìn tổng quan về ý nghĩa của thống kê trong thực tiễn và ứng dụng ngôn ngữ R trong thống kê (Ở trong bài tập lớn này chính là chủ đề dịch bệnh, một đề tài nhức nhối thời gian rất lâu vừa qua).

 Dalgaard, P. *Introductory Statistics with R*. Springer 2008.

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

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

 <https://vietnambiz.vn/he-so-tuong-quan-correlation-coefficient-la-gi-ung-dung-cua-he-htm>

 <https://rpubs.com/>

 <https://cran.r-project.org/index.html>







[Động cơ nghiên cứu](#)

[Mục tiêu](#)

[Kiến thức chuẩn bị](#)

[Nhiệm vụ](#)

[Kết luận](#)

Cảm ơn mọi người đã lắng nghe!