

BÁO CÁO THỰC HÀNH LAB01

Họ tên: Nguyễn Mạnh Đức

MSSV: 20521196

Môn học: Thu thập và tiền xử lý dữ liệu







Lớp thực hành: DS103.M21.2

BÀI TẬP:

a) Code lại các ví dụ trong Phần 3.

```
rm(list = ls())
coronaData <- read.csv("data/covid_19_data.csv")
coronaData$ObservationDate <- as.Date(coronaData$ObservationDate, "%m/%d/%Y")
nrow(coronaData)
ncol(coronaData)
head(coronaData, 10)
names(coronaData)
countryCorona <- coronaData['Country.Region']
maxConfirmedCases <- max(coronaData['Confirmed'])
coronaChina <- coronaData[which(coronaData$Country.Region == 'Mainland China'),]
maxCountryConfirmedCorona <-
  coronaData[which(coronaData$Confirmed==maxConfirmedCases),]['Country.Region']
maxStateConfirmedCorona <-
  coronaData[which(coronaData$Confirmed==maxConfirmedCases),]['Province.State']
data_jan <- coronaData[which(coronaData$ObservationDate>=
  "2020-01-01" & coronaData$ObservationDate <= "2020-01-31"), ]
```

Sau khi chạy:

Data		
coronaChina	15758 obs. of 8 variables	
coronaData	306429 obs. of 8 variables	
countryCorona	306429 obs. of 1 variable	
data_jan	513 obs. of 8 variables	
maxCountryConfirmedCorona	1 obs. of 1 variable	
maxStateConfirmedCorona	1 obs. of 1 variable	
Values		
maxConfirmedCases	5863138	

```

> rm(list = ls())
> coronaData <- read.csv("data/covid_19_data.csv")
>
> coronaData$ObservationDate <- as.Date(coronaData$ObservationDate, "%m/%d/%Y")
>
> nrow(coronaData)
[1] 306429
>
> ncol(coronaData)
[1] 8
>
> head(coronaData, 10)
  SNo ObservationDate Province.State Country.Region Last.Update Confirmed Deaths Recovered
1    1      2020-01-22      Anhui Mainland China 1/22/2020 17:00         1         0         0
2    2      2020-01-22      Beijing Mainland China 1/22/2020 17:00        14         0         0
3    3      2020-01-22      Chongqing Mainland China 1/22/2020 17:00         6         0         0
4    4      2020-01-22      Fujian Mainland China 1/22/2020 17:00         1         0         0
5    5      2020-01-22      Gansu Mainland China 1/22/2020 17:00         0         0         0
6    6      2020-01-22      Guangdong Mainland China 1/22/2020 17:00        26         0         0
7    7      2020-01-22      Guangxi Mainland China 1/22/2020 17:00         2         0         0
8    8      2020-01-22      Guizhou Mainland China 1/22/2020 17:00         1         0         0
9    9      2020-01-22      Hainan Mainland China 1/22/2020 17:00         4         0         0
10  10      2020-01-22      Hebei Mainland China 1/22/2020 17:00         1         0         0

> names(coronaData)
[1] "SNo" "ObservationDate" "Province.State" "Country.Region" "Last.Update"
[6] "Confirmed" "Deaths" "Recovered"
>
> countryCorona <- coronaData['Country.Region']
>
> maxConfirmedCases <- max(coronaData['Confirmed'])
>
> coronaChina <- coronaData[which(coronaData$Country.Region == 'Mainland China'),]
>
> maxCountryConfirmedCorona <-
+   coronaData[which(coronaData$Confirmed==maxConfirmedCases),]['Country.Region']
>
> maxStateConfirmedCorona <-
+   coronaData[which(coronaData$Confirmed==maxConfirmedCases),]['Province.State']
>
> data_jan <- coronaData[which(coronaData$ObservationDate>=
+   "2020-01-01" & coronaData$ObservationDate <= "2020-01-31"), ]
>

```

b) Tìm dữ liệu về số ca lây nhiễm tại Vietnam (Country.Region == 'Vietnam') và lưu vào biến coronaVietnam.

```
coronaVietnam <- coronaData[which(coronaData$Country.Region == 'Vietnam'),]
```

Sau khi chạy

SNo	ObservationDate	Province.State	Country.Region	Last.Update	Confirmed	Deaths	Recovered
82	2020-01-23		Vietnam	1/23/20 17:00	2	0	0
128	2020-01-24		Vietnam	1/24/20 17:00	2	0	0
171	2020-01-25		Vietnam	1/25/20 17:00	2	0	0
219	2020-01-26		Vietnam	1/26/20 16:00	2	0	0
268	2020-01-27		Vietnam	1/27/20 23:59	2	0	0
321	2020-01-28		Vietnam	1/28/20 23:00	2	0	0
375	2020-01-29		Vietnam	1/29/20 19:30	2	0	0
432	2020-01-30		Vietnam	1/30/20 16:00	2	0	0

c) In ra số ca lây nhiễm nhiều nhất tại Việt Nam (Sử dụng lệnh print() trong R)

```
print(max(coronaVietnam['Confirmed']))
```

Sau khi chạy:

```
> print(max(coronaVietnam['Confirmed']))  
[1] 6908
```

d) Tìm dữ liệu về số ca lây nhiễm tại Việt Nam trong tháng 02 năm 2021.

```
data_Vietnam_2 <- coronaVietnam[which(coronaVietnam$ObservationDate >= "2021-02-01"  
& coronaVietnam$ObservationDate < "2021-03-01"),]
```

Sau khi chạy:

```
> data_Vietnam_2 <- coronaVietnam[which(coronaVietnam$ObservationDate >= "2021-02-01"  
+ & coronaVietnam$ObservationDate < "2021-03-01"),]
```

	SNo	ObservationDate	Province.State	Country.Region	Last.Update	Confirmed	Deaths	Recovered
216327	216327	2021-02-01		Vietnam	2021-04-02 15:13:53	1850	35	1460
217092	217092	2021-02-02		Vietnam	2021-04-02 15:13:53	1882	35	1460
217857	217857	2021-02-03		Vietnam	2021-04-02 15:13:53	1948	35	1461
218622	218622	2021-02-04		Vietnam	2021-04-02 15:13:53	1957	35	1465
219387	219387	2021-02-05		Vietnam	2021-04-02 15:13:53	1976	35	1465
220152	220152	2021-02-06		Vietnam	2021-04-02 15:13:53	1985	35	1468
220917	220917	2021-02-07		Vietnam	2021-04-02 15:13:53	2001	35	1472
221682	221682	2021-02-08		Vietnam	2021-04-02 15:13:53	2050	35	1472
222447	222447	2021-02-09		Vietnam	2021-04-02 15:13:53	2064	35	1472
223212	223212	2021-02-10		Vietnam	2021-04-02 15:13:53	2091	35	1480
223977	223977	2021-02-11		Vietnam	2021-04-02 15:13:53	2140	35	1528
224742	224742	2021-02-12		Vietnam	2021-04-02 15:13:53	2142	35	1529
225507	225507	2021-02-13		Vietnam	2021-04-02 15:13:53	2195	35	1529
226272	226272	2021-02-14		Vietnam	2021-04-02 15:13:53	2228	35	1532

e) In ra số dữ liệu về ca lây nhiễm nhiều nhất trong tháng 01 và 02 tại Việt Nam (Lấy năm 2021).

```
# VietNam Thang 1
data_Vietnam_1 <- coronaVietnam[which(coronaVietnam$ObservationDate >= "2021-01-01"
& coronaVietnam$ObservationDate < "2021-02-01"),]

max_cofirmed_Vietnam_1 <- max(data_Vietnam_1['Confirmed'])

print(max_cofirmed_Vietnam_1)

# VietNam Thang 2
max_cofirmed_Vietnam_2 <- max(data_Vietnam_2['Confirmed'])

print(max_cofirmed_Vietnam_2)
```

Sau khi chạy:

```
> data_Vietnam_1 <- coronaVietnam[which(coronaVietnam$ObservationDate >= "2021-01-01"
+ & coronaVietnam$ObservationDate < "2021-02-01"),]
>
> max_cofirmed_Vietnam_1 <- max(data_Vietnam_1['Confirmed'])
>
> print(max_cofirmed_Vietnam_1)
[1] 1817
>
> # VietNam Thang 2
> max_cofirmed_Vietnam_2 <- max(data_Vietnam_2['Confirmed'])
>
> print(max_cofirmed_Vietnam_2)
[1] 2448
>
```

f) Thực hiện tương tự câu e) cho Indonesia và Philipine.

```
# Indonesia Thang 1
data_Indonesia <- coronaData[which(coronaData$Country.Region == 'Indonesia'),]

data_Indonesia_1 <- data_Indonesia[which(data_Indonesia$ObservationDate >= '2021-01-01'
& data_Indonesia$ObservationDate < '2021-02-01' ),]

max_cofirmed_Indonesia_1 <- max(data_Indonesia_1['Confirmed'])

print(max_cofirmed_Indonesia_1)

# Indonesia Thang 2
data_Indonesia_2 <- data_Indonesia[which(data_Indonesia$ObservationDate >= '2021-02-01'
& data_Indonesia$ObservationDate < '2021-03-01' ),]

max_cofirmed_Indonesia_2 <- max(data_Indonesia_2['Confirmed'])

print(max_cofirmed_Indonesia_2)
```

```

#Philippines Thang 1
data_Philippines <- coronaData[which(coronaData$Country.Region == 'Philippines'),]

data_Philippines_1 <- data_Philippines[which(data_Philippines$ObservationDate >= '2021-01-01'
& data_Philippines$ObservationDate < '2021-02-01' ),]

max_cofirmed_Philippines_1 <- max(data_Philippines_1['Confirmed'])

print(max_cofirmed_Philippines_1)

#Philippines Thang 2

data_Philippines_2 <- data_Philippines[which(data_Philippines$ObservationDate >= '2021-02-01'
& data_Philippines$ObservationDate < '2021-03-01' ),]

max_cofirmed_Philippines_2 <- max(data_Philippines_2['Confirmed'])

print(max_cofirmed_Philippines_2)

```

Sau khi chạy:

```

> # Indonesia Thang 1
> data_Indonesia <- coronaData[which(coronaData$Country.Region == 'Indonesia'),]
>
> data_Indonesia_1 <- data_Indonesia[which(data_Indonesia$ObservationDate >= '2021-01-01'
+ & data_Indonesia$ObservationDate < '2021-02-01' ),]
>
> max_cofirmed_Indonesia_1 <- max(data_Indonesia_1['Confirmed'])
>
> print(max_cofirmed_Indonesia_1)
[1] 1078314
>
> # Indonesia Thang 2
>
> data_Indonesia_2 <- data_Indonesia[which(data_Indonesia$ObservationDate >= '2021-02-01'
+ & data_Indonesia$ObservationDate < '2021-03-01' ),]
>
> max_cofirmed_Indonesia_2 <- max(data_Indonesia_2['Confirmed'])
>
> print(max_cofirmed_Indonesia_2)
[1] 1334634

```

```

> #Philippines Thang 1
> data_Philippines <- coronaData[which(coronaData$Country.Region == 'Philippines'),]
>
> data_Philippines_1 <- data_Philippines[which(data_Philippines$ObservationDate >= '2021-01-01'
+ & data_Philippines$ObservationDate < '2021-02-01' ),]
>
> max_cofirmed_Philippines_1 <- max(data_Philippines_1['Confirmed'])
>
> print(max_cofirmed_Philippines_1)
[1] 525618
>
> #Philippines Thang 2
>
> data_Philippines_2 <- data_Philippines[which(data_Philippines$ObservationDate >= '2021-02-01'
+ & data_Philippines$ObservationDate < '2021-03-01' ),]
>
> max_cofirmed_Philippines_2 <- max(data_Philippines_2['Confirmed'])
>
> print(max_cofirmed_Philippines_2)
[1] 576352

```

g) In ra dữ liệu về ca tử vong của Trung Quốc trong khoảng thời gian từ 01/02/2021 cho đến 15/02/2021. In ra màn hình sử dụng lệnh print().

```
data_MainlandChina <- coronaData[which(coronaData$Country.Region == 'Mainland China'),]  
data_MainlandChina_1.2_15.2 <- data_MainlandChina[which(data_MainlandChina$ObservationDate >= '2021-02-01'  
                                                         & data_MainlandChina$ObservationDate <= '2021-02-15'),]  
print(data_MainlandChina_1.2_15.2['Deaths'])
```

Sau khi chạy:

```
> data_MainlandChina <- coronaData[which(coronaData$Country.Region == 'Mainland China'),]  
>  
> data_MainlandChina_1.2_15.2 <- data_MainlandChina[which(data_MainlandChina$ObservationDate >= '2021-02-01'  
+                                                         & data_MainlandChina$ObservationDate <= '2021-02-15'),]  
>  
> print(data_MainlandChina_1.2_15.2['Deaths'])  
Deaths  
216354      6  
216388      9  
216438      6  
216480      1  
216485      2  
216503      8  
216504      2  
216507      2  
216510      6  
216516      7  
216517     13  
216518     22  
216527    4512  
216529      4  
216537      1  
216551      0  
216552      1  
216553      3  
216601      2  
216685      0  
216750      0  
216796      3  
216797      7
```

h) Đếm số lượng ca ghi nhận theo từng tỉnh của Trung Quốc trong tháng 02/2021. Gợi ý: Dùng hàm table().

```
data_MainlandChina_2 <- data_MainlandChina[which(data_MainlandChina$ObservationDate >= '2021-02-01'  
                                                  & data_MainlandChina$ObservationDate < '2021-03-01'),]  
count_data_province_MainlandChina_2 <- table(data_MainlandChina_2$Province.State)  
list_province_MainlandChina_2 <- unique(data_MainlandChina_2$Province.State)  
for (i in 1:length(list_province_MainlandChina_2)){  
  cur_data <- data_MainlandChina_2[which(data_MainlandChina_2$Province.State == list_province_MainlandChina_2[i]),]  
  min_date <- min(cur_data$ObservationDate)  
  max_date <- max(cur_data$ObservationDate)  
  case <- cur_data[which(cur_data$ObservationDate == max_date),]$Confirmed  
  - cur_data[which(cur_data$ObservationDate == min_date),]$Confirmed  
  print(paste(list_province_MainlandChina_2[i], ": ", case))  
}
```

Sau khi chạy:

```

> data_MainlandChina_2 <- data_MainlandChina[which(data_MainlandChina$ObservationDate >= '2021-02-01'
+ & data_MainlandChina$ObservationDate < '2021-03-01'),]
>
> count_data_province_MainlandChina_2 <- table(data_MainlandChina_2$Province.State)
> list_province_MainlandChina_2 <- unique(data_MainlandChina_2$Province.State)
> for (i in 1:length(list_province_MainlandChina_2)){
+   cur_data <- data_MainlandChina_2[which(data_MainlandChina_2$Province.State == list_province_MainlandChina_2[i]),]
+   min_date <- min(cur_data$ObservationDate)
+   max_date <- max(cur_data$ObservationDate)
+   case <- cur_data[which(cur_data$ObservationDate == max_date),]$Confirmed
+     - cur_data[which(cur_data$ObservationDate == min_date),]$Confirmed
+   print(paste(list_province_MainlandChina_2[i], ": ", case))
+ }
[1] "Anhui : 994"
[1] "Beijing : 1049"
[1] "Chongqing : 591"
[1] "Fujian : 551"
[1] "Gansu : 187"
[1] "Guangdong : 2212"
[1] "Guangxi : 267"
[1] "Guizhou : 147"
[1] "Hainan : 171"
[1] "Hebei : 1317"
[1] "Heilongjiang : 1610"
[1] "Henan : 1304"
[1] "Hubei : 68151"
[1] "Hunan : 1036"

```

- i) Tìm dữ liệu ca tử vong của Trung Quốc trong khoảng thời gian từ 01/02/2021 cho đến 15/02/2021. In ra màn hình sử dụng lệnh print().

Giống câu g)

- k) *Cố nhận xét gì về số ca nhiễm mới tại Việt Nam giữa tháng 05/2020 và tháng 05/2021. Vẽ biểu đồ đường thể hiện số ca nhiễm mới trong 2 tháng trên. Gợi ý: Dùng hàm plot() trong R.

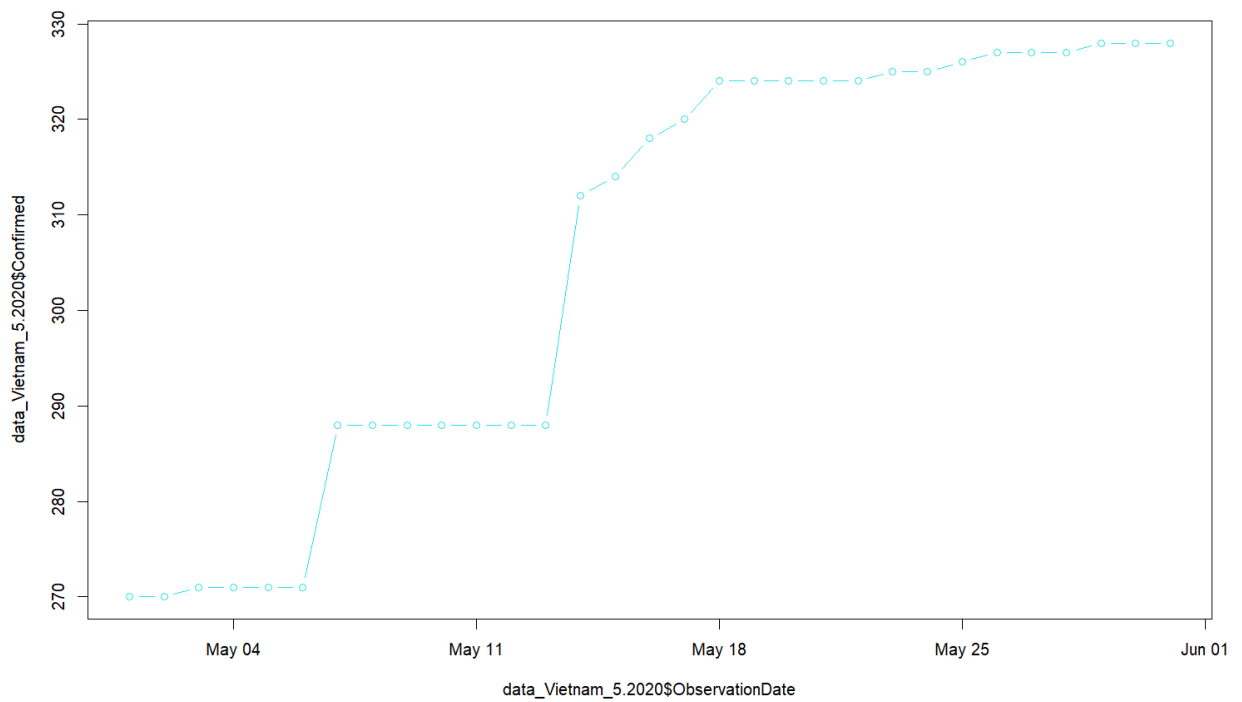
Tháng 5/2020:

```

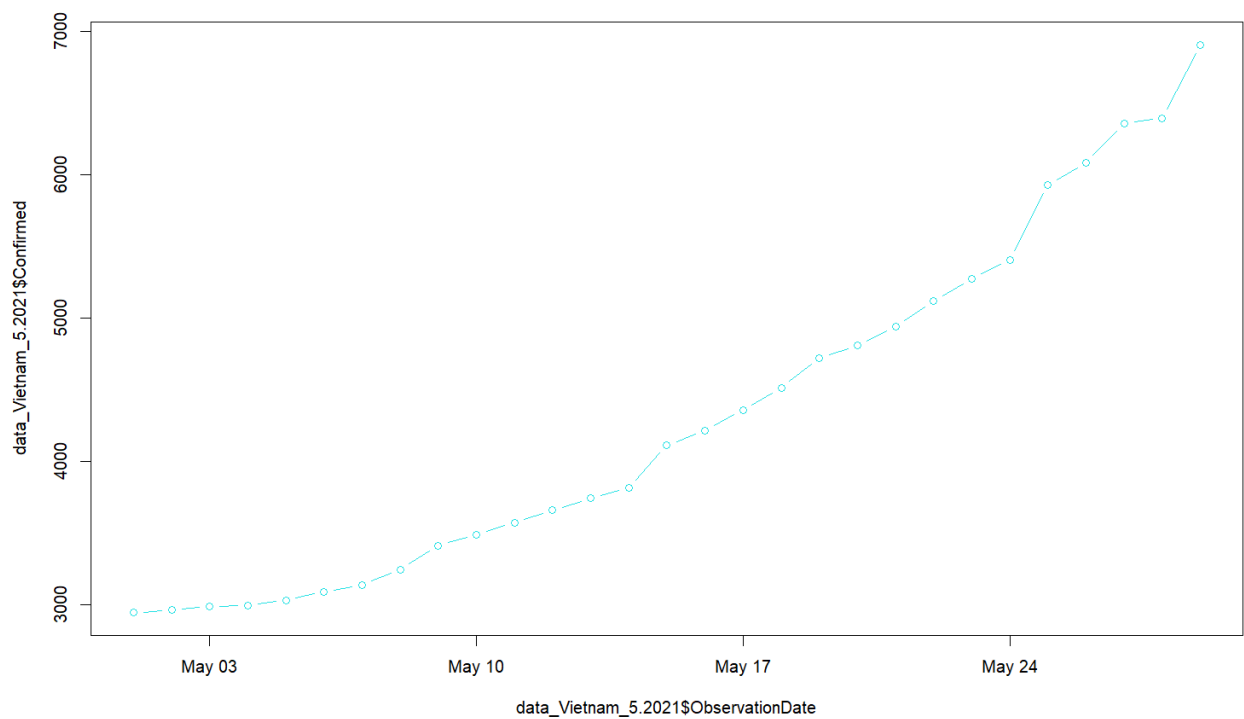
data_Vietnam_5.2020 <- coronaVietnam[which(coronaVietnam$ObservationDate >= '2020-05-01'
& coronaVietnam$ObservationDate < '2020-06-01'),]

plot(data_Vietnam_5.2020$ObservationDate, data_Vietnam_5.2020$Confirmed, type = "b" , col = 5)

```



Tháng 5/2021




```
data_Vietnam_5.2021 <- coronaVietnam[which(coronaVietnam$ObservationDate >= '2021-05-01'
& coronaVietnam$ObservationDate < '2021-06-01'),]

plot(data_Vietnam_5.2021$ObservationDate, data_Vietnam_5.2021$Confirmed, type = "b" , col = 5)
```

Nhận xét: Biểu đồ ở tháng 5/2021 tăng dần đều hơn so với biểu đồ ở tháng 5/2020.

Ở Biểu đồ tháng 5/2020 chủ yếu đi ngang + có các đường chéo lên rõ rệt → biểu thị việc có ngăn chặn dịch covid nhưng tới 1 giai đoạn thì không hiệu quả.

Ở Biểu đồ tháng 5/2021 thì tăng dần khá đều → biểu thị việc lây nhiễm covid

1) * Vẽ biểu đồ về số ca lây nhiễm nhiều nhất của 3 quốc gia: Vietnam, Indonesia và Philippine trong tháng 01 và tháng 02 năm 2021.

```
Ten <- c("VN_max_1","VN_max_2", "ID_max_1", "ID_max_2", "PH_max_1", "PH_max_2")

SoLuong <- c(max_cofirmed_Vietnam_1,max_cofirmed_Vietnam_2,max_cofirmed_Indonesia_1,
max_cofirmed_Indonesia_2,max_cofirmed_Philippines_1, max_cofirmed_Philippines_2)

barplot(SoLuong, names.arg = Ten)
```

Sau khi chạy:

