

Tên: Ngô Đức Vũ
MSSV: 20520950
Môn: Thu thập và tiền xử lý dữ liệu

BÁO CÁO

THỰC HÀNH LAB1: LÀM QUEN VỚI RSTUDIO

BÀI THỰC HÀNH 1:

- a) Code lại các ví dụ trong Phần 3.
- b) Tìm dữ liệu về số ca lây nhiễm tại Vietnam (`Country.Region == 'Vietnam'`) và lưu vào biến `coronaVietnam`.
- c) In ra số ca lây nhiễm nhiều nhất tại Việt Nam (Sử dụng lệnh `print()` trong R)
- d) Tìm dữ liệu về số ca lây nhiễm tại Việt Nam trong tháng 02.
- e) In ra số dữ liệu về ca lây nhiễm nhiều nhất trong tháng 01 và 02 tại Việt Nam (Lấy năm 2021).
- f) Thực hiện tương tự câu e) cho Indonesia và Philipine.
- 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()`.
- h) Đếm số lượng ca ghi nhận theo từng tỉnh của Trung Quốc trong tháng 02/2021. Gợi ý: Dùng hàm `table()`.
- 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()`.
- k) * Có 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.
- l) * 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

```
#Đọc file và gán vào biến coronaData
coronaData <- read.csv("dataset/covid_19_data.csv")
```

```
coronaData      306429 obs. of 8 variables
```

```
#Hiển thị số dòng, cột
```

```
ncol(coronaData)
```

```
nrow(coronaData)
```

```
> ncol(coronaData)
```

```
[1] 8
```

```
> nrow(coronaData)
```

```
[1] 306429
```

```
>
```

```
#Hiển thị 10 dòng dữ liệu đầu tiên
```

```
head(coronaData,10)
```

	SNO	ObservationDate	Province.State	Country.Region	Last.Update	Confirmed	Deaths	Recovered
1	1	01/22/2020	Anhui	Mainland China	1/22/2020 17:00	1	0	0
2	2	01/22/2020	Beijing	Mainland China	1/22/2020 17:00	14	0	0
3	3	01/22/2020	Chongqing	Mainland China	1/22/2020 17:00	6	0	0
4	4	01/22/2020	Fujian	Mainland China	1/22/2020 17:00	1	0	0
5	5	01/22/2020	Gansu	Mainland China	1/22/2020 17:00	0	0	0
6	6	01/22/2020	Guangdong	Mainland China	1/22/2020 17:00	26	0	0
7	7	01/22/2020	Guangxi	Mainland China	1/22/2020 17:00	2	0	0
8	8	01/22/2020	Guizhou	Mainland China	1/22/2020 17:00	1	0	0
9	9	01/22/2020	Hainan	Mainland China	1/22/2020 17:00	4	0	0
10	10	01/22/2020	Hebei	Mainland China	1/22/2020 17:00	1	0	0

```
#In ra các tên cột
```

```
names(coronaData)
```

```
> #In ra tên các cột
```

```
> names(coronaData)
```

```
[1] "SNO" "observationDate" "Province.State" "Country.Region" "Last.Update"
```

```
[6] "Confirmed" "Deaths" "Recovered"
```

```
#Tạo biến các quốc gia có dịch corona
```

```
countryCorona <- unique(coronaData["Country.Region"])
```

```
coronaData      306429 obs. of 8 variables
```

```
countryCorona    229 obs. of 1 variable
```

```
coronaData      306429 obs. of 8 variables
```

```
countryCorona    306429 obs. of 1 variable
```

```
#Liệt kê số ca nhiễm đã được xác nhận
```

```
maxConfirmedCases <- max(coronaData["Confirmed"])
```

```
values
```

```
maxConfirmedCa... 5863138
```

#Liệt kê các dữ liệu về covid-19 tại quốc gia Trung Quốc đại lục (Mainland China) và lưu vào biến coronaChina.

```
coronaChina <- coronaData[which(coronaData$Country.Region=="Mainland China"),]
```

Data

coronaChina 15758 obs. of 8 variables

	SNo	ObservationDate	Province.State	Country.Region	Last.Update	Confirmed	Deaths	Recovered
1	1	01/22/2020	Anhui	Mainland China	1/22/2020 17:00	1	0	0
2	2	01/22/2020	Beijing	Mainland China	1/22/2020 17:00	14	0	0
3	3	01/22/2020	Chongqing	Mainland China	1/22/2020 17:00	6	0	0
4	4	01/22/2020	Fujian	Mainland China	1/22/2020 17:00	1	0	0
5	5	01/22/2020	Gansu	Mainland China	1/22/2020 17:00	0	0	0
6	6	01/22/2020	Guangdong	Mainland China	1/22/2020 17:00	26	0	0
7	7	01/22/2020	Guangxi	Mainland China	1/22/2020 17:00	2	0	0
8	8	01/22/2020	Guizhou	Mainland China	1/22/2020 17:00	1	0	0
9	9	01/22/2020	Hainan	Mainland China	1/22/2020 17:00	4	0	0
10	10	01/22/2020	Hebei	Mainland China	1/22/2020 17:00	1	0	0
11	11	01/22/2020	Heilongjiang	Mainland China	1/22/2020 17:00	0	0	0
12	12	01/22/2020	Henan	Mainland China	1/22/2020 17:00	5	0	0
14	14	01/22/2020	Hubei	Mainland China	1/22/2020 17:00	444	17	28
15	15	01/22/2020	Hunan	Mainland China	1/22/2020 17:00	4	0	0
16	16	01/22/2020	Inner Mongolia	Mainland China	1/22/2020 17:00	0	0	0
17	17	01/22/2020	Jiangsu	Mainland China	1/22/2020 17:00	1	0	0
18	18	01/22/2020	Jiangxi	Mainland China	1/22/2020 17:00	2	0	0

Showing 1 to 17 of 15,758 entries, 8 total columns

#Quốc gia có số ca nhiễm nhiều nhất

```
maxCountryConfirmedCorona <- coronaData[which(coronaData$Confirmed==maxConfirmedCases),][['Country.Region']]
```

code.R* x maxCountryConfirmedCorona x coron

Filter

Country.Region

298069 France

#Tìm tỉnh có số ca nhiễm nhiều nhất

```
maxStateConfirmedCorona <- coronaData[which(coronaData$Confirmed==maxConfirmedCases),][['Province.State']]
```

Kết quả: NULL vì (Province.State) của nước Pháp được để trống

code.R* x maxStateConfirmedCorona x

Filter

Province.State

298069

#Định dạng lại dữ liệu ngày tháng

```
coronaData$ObservationDate <- as.Date(coronaData$ObservationDate, "%m/%d/%Y")
```

	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
11	11	2020-01-22	Heilongjiang	Mainland China	1/22/2020 17:00	0	0	0
12	12	2020-01-22	Henan	Mainland China	1/22/2020 17:00	5	0	0
13	13	2020-01-22	Hong Kong	Hong Kong	1/22/2020 17:00	0	0	0
14	14	2020-01-22	Hubei	Mainland China	1/22/2020 17:00	444	17	28
15	15	2020-01-22	Hunan	Mainland China	1/22/2020 17:00	4	0	0
16	16	2020-01-22	Inner Mongolia	Mainland China	1/22/2020 17:00	0	0	0

Lấy dữ liệu trong tháng 1/2020: bắt đầu từ 01/01/2020 đến 31/01/2020.

```
data_jan <- coronaData[which(coronaData$ObservationDate>= "2020-01-01" & coronaData$ObservationDate <= "2020-01-31"), ]
```

	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
11	11	2020-01-22	Heilongjiang	Mainland China	1/22/2020 17:00	0	0	0
12	12	2020-01-22	Henan	Mainland China	1/22/2020 17:00	5	0	0
13	13	2020-01-22	Hong Kong	Hong Kong	1/22/2020 17:00	0	0	0
14	14	2020-01-22	Hubei	Mainland China	1/22/2020 17:00	444	17	28
15	15	2020-01-22	Hunan	Mainland China	1/22/2020 17:00	4	0	0
16	16	2020-01-22	Inner Mongolia	Mainland China	1/22/2020 17:00	0	0	0
17	17	2020-01-22	Jiangsu	Mainland China	1/22/2020 17:00	1	0	0

Showing 1 to 17 of 513 entries, 8 total columns

#4.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"),]
```

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
496	2020-01-31		Vietnam	1/31/2020 23:59	2	0	0
554	2020-02-01		Vietnam	2/1/2020 7:38	6	0	1
623	2020-02-02		Vietnam	2020-02-01T07:38:12	6	0	1
691	2020-02-03		Vietnam	2020-02-03T21:43:02	8	0	1
761	2020-02-04		Vietnam	2020-02-03T21:43:02	8	0	1
833	2020-02-05		Vietnam	2020-02-03T21:43:02	8	0	1
906	2020-02-06		Vietnam	2020-02-06T01:13:08	10	0	1
980	2020-02-07		Vietnam	2020-02-06T01:13:08	10	0	1

#4.c, In ra số ca lây nhiễm nhiều nhất tại Việt Nam (Sử dụng lệnh print())

```
maxConfirmedVietnam<-max(coronaVietnam$Confirmed)
```

```
print(maxConfirmedVietnam)
```

```
R 4.1.2 · V/Ngôn ngữ R/20520950_Lab1/
> View(coronaVietnam)
> maxConfirmedVietnam<-max(coronaVietnam$Confirmed)
> maxConfirmedVietnam<-max(coronaVietnam$Confirmed)
> print(maxConfirmedVietnam)
[1] 6908
> |
```

#4.d, Tìm dữ liệu về số ca lây nhiễm tại Việt Nam trong tháng 02

```
Vietnam_feb<-coronaVietnam[which(coronaVietnam$ObservationDate>= "2020-02-01" & coronaVietnam$ObservationDate<= "2020-02-29"),]
```

SNo	ObservationDate	Province.State	Country.Region	Last.Update	Confirmed	Deaths	Recovered
1578	2020-02-15		Vietnam	2020-02-13T06:33:02	16	0	7
1655	2020-02-16		Vietnam	2020-02-13T06:33:02	16	0	7
1732	2020-02-17		Vietnam	2020-02-13T06:33:02	16	0	7
1809	2020-02-18		Vietnam	2020-02-13T06:33:02	16	0	7
1886	2020-02-19		Vietnam	2020-02-13T06:33:02	16	0	7
1964	2020-02-20		Vietnam	2020-02-13T06:33:02	16	0	7
2044	2020-02-21		Vietnam	2020-02-21T06:03:04	16	0	14
2130	2020-02-22		Vietnam	2020-02-21T06:03:04	16	0	14
2216	2020-02-23		Vietnam	2020-02-21T06:03:04	16	0	14
2304	2020-02-24		Vietnam	2020-02-21T06:03:04	16	0	14
2397	2020-02-25		Vietnam	2020-02-25T08:53:02	16	0	16
2495	2020-02-26		Vietnam	2020-02-25T08:53:02	16	0	16
2598	2020-02-27		Vietnam	2020-02-25T08:53:02	16	0	16
2708	2020-02-28		Vietnam	2020-02-25T08:53:02	16	0	16
2825	2020-02-29		Vietnam	2020-02-25T08:53:02	16	0	16

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

```
Vietnam_jan_feb<-coronaVietnam[which(coronaVietnam$ObservationDate>= "2021-01-01" &  
coronaVietnam$ObservationDate <= "2021-02-28"),]
```

```
max_vietnam_confirmed_jan_feb<-max(Vietnam_jan_feb$Confirmed)
```

```
print(max_vietnam_confirmed_jan_feb)
```

```
> Vietnam_jan_feb<-coronaVietnam[which(coronaVietnam$ObservationDate>= "2021-01-01" & coronaVietnam$Observa  
tionDate <= "2021-02-28"),]  
> max_vietnam_confirmed_jan_feb<-max(Vietnam_jan_feb$Confirmed)  
> print(max_vietnam_confirmed_jan_feb)  
[1] 2448  
>
```

#4.f, Đối với Philippines

```
coronaPhilippines<-coronaData[which(coronaData$Country.Region=="Philippines"),]
```

```
Philippines_jan_feb<-coronaPhilippines[which(coronaPhilippines$ObservationDate>= "2021-01-01" &  
coronaPhilippines$ObservationDate <= "2021-02-28"),]
```

```
max_philippines_confirmed_jan_feb<-max(Philippines_jan_feb$Confirmed)
```

```
print(max_philippines_confirmed_jan_feb)
```

```
> Philippines_jan_feb<-coronaPhilippines[which(coronaPhilippines$ObservationDate>= "2021-01-01" & coronaPhi  
lippines$ObservationDate <= "2021-02-28"),]  
> max_philippines_confirmed_jan_feb<-max(Philippines_jan_feb$Confirmed)  
> print(max_philippines_confirmed_jan_feb)  
[1] 576352  
>
```

#4.f, Đối với Indonesia

```
coronaIndonesia<-coronaData[which(coronaData$Country.Region=="Indonesia"),]
```

```
Indonesia_jan_feb<-coronaIndonesia[which(coronaIndonesia$ObservationDate>= "2021-01-01" &  
coronaIndonesia$ObservationDate <= "2021-02-28"),]
```

```
max_indonesia_confirmed_jan_feb<-max(Indonesia_jan_feb$Confirmed)
```

```
print(max_indonesia_confirmed_jan_feb)
```

```
> Indonesia_jan_feb<-coronaIndonesia[which(coronaIndonesia$ObservationDate>= "2021-01-01" & coronaIndonesi  
a$ObservationDate <= "2021-02-28"),]  
> max_indonesia_confirmed_jan_feb<-max(Indonesia_jan_feb$Confirmed)  
> print(max_indonesia_confirmed_jan_feb)  
[1] 1334634  
>
```

#4.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()

```
coronaChina$ObservationDate<-as.Date(coronaChina$ObservationDate, "%m/%d/%Y")
```

```
coronaChinaDeaths<-coronaChina[which(coronaChina$ObservationDate>= "2021-02-01" & coronaChina$ObservationDate  
<= "2021-02-15"),][["Deaths"]
```

```
print(coronaChinaDeaths)
```

```
R 4.1.2 - V:\Ngôn ngữ R\20520950_Lab1/ <-  
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  
216798      7  
216799      0  
216803      3  
216835      3  
216836      0  
216864      0  
216908      3  
216916      2  
216922      1  
217119      6  
217153      9  
217203      6  
217245      1  
217250      2
```

#4.h, Đếm số lượng ca ghi nhận theo từng tỉnh của Trung Quốc trong tháng 02/2021.

```
coronaChina_feb<-coronaChina[which(coronaChina$ObservationDate>= "2021-02-01" & coronaChina$ObservationDate <=
"2021-02-15"),]
```

```
count=table(coronaChina_feb$Province.State)
```

```
list_province<-unique(coronaChina_feb$Province.State)
```

```
for(i in 1:length(list_province))
```

```
{
```

```
  data_new <- coronaChina_feb[which(coronaChina_feb$Province.State==list_province[i]),]
```

```
  min_date <- min(data_new$ObservationDate)
```

```
  max_date <- max(data_new$ObservationDate)
```

```
  case<-data_new[which(data_new$ObservationDate ==max_date),]$Confirmed -
```

```
  data_new[which(data_new$ObservationDate ==min_date),]$Confirmed
```

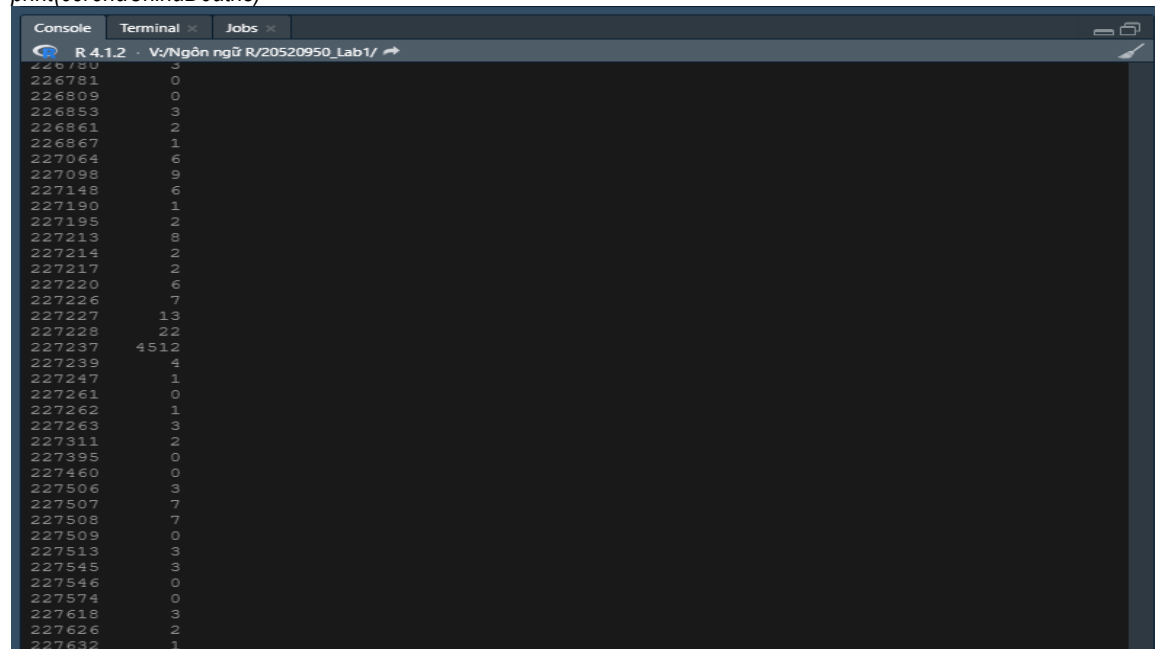
```
  print(paste(list_province[i], " :",case))
```

```
}
```

```
[1] "Anhui      : 0"
[1] "Beijing    : 7"
[1] "Chongqing  : 0"
[1] "Fujian     : 7"
[1] "Gansu      : 0"
[1] "Guangdong  : 44"
[1] "Guangxi    : 0"
[1] "Guizhou    : 0"
[1] "Hainan     : 0"
[1] "Hebei      : 4"
[1] "Heilongjiang : 18"
[1] "Henan      : 1"
[1] "Hubei      : 0"
[1] "Hunan      : 1"
[1] "Inner Mongolia : 1"
[1] "Jiangsu    : 6"
[1] "Jiangxi    : 0"
[1] "Jilin      : 18"
[1] "Liaoning   : 2"
[1] "Ningxia    : 0"
[1] "Qinghai    : 0"
[1] "Shaanxi    : 7"
[1] "Shandong   : 1"
[1] "Shanghai   : 60"
[1] "Shanxi     : 0"
[1] "Sichuan    : 8"
[1] "Tianjin    : 5"
[1] "Tibet      : 0"
[1] "Unknown    : 0"
[1] "Xinjiang   : 0"
[1] "Yunnan     : 0"
[1] "Zhejiang   : 4"
```

#4.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().

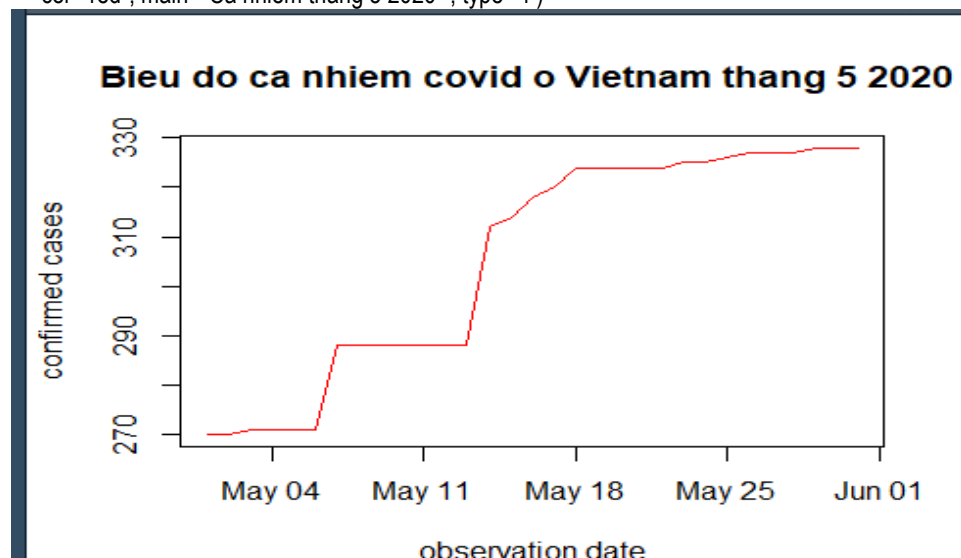
```
print(coronaChinaDeaths)
```



```
R 4.1.2 - V:/Ngôn ngữ R/20520950_Lab1/
226780      3
226781      0
226809      0
226853      3
226861      2
226867      1
227064      6
227098      9
227148      6
227190      1
227195      2
227213      8
227214      2
227217      2
227220      6
227226      7
227227     13
227228     22
227237    4512
227239      4
227247      1
227261      0
227262      1
227263      3
227311     21
227395      0
227460      0
227506      3
227507      7
227508      7
227509      0
227513      3
227545      3
227546      0
227574      0
227618      3
227626      2
227632      1
```

#4.k, Có nhận xét gì về số ca nhiễm mới tại Việt Nam giữa tháng 05/2020 và tháng 05/2021.

```
coronaVietnam20_may<-coronaVietnam[which(coronaVietnam$ObservationDate>= "2020-05-01" &
coronaVietnam$ObservationDate <= "2020-05-31"),]
coronaVietnam21_may<-coronaVietnam[which(coronaVietnam$ObservationDate>= "2021-05-01" &
coronaVietnam$ObservationDate <= "2021-05-31"),]
plot(x=coronaVietnam20_may$ObservationDate, xlab="observation date",
y=coronaVietnam20_may$Confirmed, ylab="confirmed cases",
col="red", main="Ca nhiễm tháng 5 2020 ", type="l")
plot(x=coronaVietnam21_may$ObservationDate, xlab="observation date",
y=coronaVietnam21_may$Confirmed, ylab="confirmed cases",
col="red", main="Ca nhiễm tháng 5 2021 ", type="l")
```



Nhận xét: Số lượng ca xác nhận dương tính với covid trong tháng 5/2020 thấp hơn so nhiều số với tháng 5/2021

Tại biểu đồ của tháng 5/2020, số ca nhiễm mới tăng chậm, biên độ nhỏ khi trong tháng 5 tăng khoảng 60 ca, so với tổng số ca đầu tháng 5 là 270 ca

Tại biểu đồ của tháng 5/2021, số ca nhiễm mới tăng mạnh, biên độ lớn khi trong tháng 5 đã tăng khoảng 4000 ca, so với tổng số ca đầu tháng là 3000 ca, lượng ca nhiễm mới tháng 5 cao hơn gấp đôi so với lúc đầu tháng 5.

