ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

THU THẬP VÀ TIỀN XỬ LÝ DỮ LIỆU
BÁO CÁO LAB1

Họ và tên : Lưu Quang Tiến Hoàng
MSSV : 20521342
Lớp : DS103.M21

Dữ liệu mẫu:

Bộ dữ liệu: **Novel Corona Virus 2019**

Link tải: https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset

Bước 1: Khởi tạo Project trong R Studio và đặt tên là *Lab1_2.*
Bước 2: Tạo file chứa source code và đặt tên là *btapso2.R.*
Bước 3: Tạo thư mục *dataset/* chứa trong thư mục project.
Bước 4: Giải nén bộ dữ liệu Novel Corona Virus 2019 và copy file covid_19_data.csv vào thư mục dataset/ đã tạo ở Bước 3.

Xóa bỏ các đối tượng có trong bộ nhớ

```
rm(list=ls())
```

Đọc dữ liệu và gán vào biến coronaData.

```
crData <- read.csv("dataset/covid_19_data.csv")
```

Định dạng ngày tháng năm.

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

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

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

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

```
maxConfirmedvietnam <- max(coronaVietnam['Confirmed'])

print(maxConfirmedvietnam)
```

Tìm dữ liệu về số ca lây nhiễm tại Việt Nam trong tháng 02.

```
data_vn <- crData[which(crData$ObservationDate>="2021-01-01"&
crData$ObservationDate<="2021-02-28"& crData$Country.Region=='Vietnam'), ]
```

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

```
max_vn <- max(data_feb_vn['Confirmed'])

print(max_vn)
```

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

```
data_id <- crData[which(crData$ObservationDate>="2021-01-01"& crData$ObservationDate<="2021-02-28"& crData$Country.Region=='India'), ]

max_id <- max(data_jan_id['Confirmed'])

print(max_id)
```

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

```
data_sg <- crData[which(crData$ObservationDate>="2021-01-01"& crData$ObservationDate<="2021-02-28"& crData$Country.Region=='Singapore'), ]

max_sg <- max(data_sg['Confirmed'])

print(max_sg)
```

In ra dữ liệu về ca tử vong của Trung Quốc trong khoảng thời gian từ 01/02/2021

```
 tq <- crData[which(crData$Country.Region=='Mainland China'),]

deaths_tq <- crData[which(crData$ObservationDate>="2021-02-01"& crData$ObservationDate<="2021-02-15"& crData$Country.Region=='Mainland China'), ]

print(deaths_tq)
```

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.

data_may20_vn <- crData[which(crData$ObservationDate>="2020-05-01"& crData$ObservationDate<="2020-05-31"& crData$Country.Region=='Vietnam'),]

data_may21_vn <- crData[which(crData$ObservationDate>="2021-05-01"& crData$ObservationDate<="2021-05-31"& crData$Country.Region=='Vietnam'),]

plot(data_may20_vn,data_may21_vn)