BÀI TẬP THỰC HÀNH 4

Họ tên: Nguyễn Mạnh Đức

MSSV: 20521196

Lớp: DS103.M21

BÁO CÁO

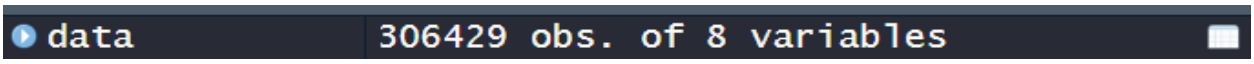
Bài 1:

a) Các bạn hiện thực lại các ví dụ ở trên.

CODE:

```
data <- read.csv("corona_virus/covid_19_data.csv")

data$ObservationDate <- as.Date(data$ObservationDate, tz = "UTC", "%d/%m/%Y")
# THAO TÁC SELECT TRÊN DỮ LIỆU

# 1. Lấy dữ liệu ở các thuộc tính bao gồm: ngày quan sát (ObservationDate), quốc gia
# nhiễm (Country.Region), số ca dương tính (Confirmed), số ca tử vong (Deaths).
data %>% select(ObservationDate, Country.Region, Confirmed, Deaths)
```

KẾT QUẢ




```
R 4.2.0 · D:/Môn Học/HK4-DS(3)-Thu thập và tiền xử lý dữ liệu/Lab04/BTTH4/
> data <- read.csv("corona_virus/covid_19_data.csv")
>
> data$ObservationDate <- as.Date(data$ObservationDate, tz = "UTC", "%d/%m/%Y")
> # THAO TÁC SELECT TRÊN DỮ LIỆU
>
> # 1. Lấy dữ liệu ở các thuộc tính bao gồm: ngày quan sát (ObservationDate), quốc gia
> # nhiễm (Country.Region), số ca dương tính (Confirmed), số ca tử vong (Deaths).
> data %>% select(ObservationDate, Country.Region, Confirmed, Deaths)
    ObservationDate Country.Region Confirmed Deaths
1              <NA> Mainland China         1      0
2              <NA> Mainland China        14      0
3              <NA> Mainland China         6      0
4              <NA> Mainland China         1      0
5              <NA> Mainland China         0      0
```

CODE

```
# 2. Lấy dữ liệu về số ca hồi phục ở từng quốc gia, thuộc tính hiển thị gồm: ngày quan
# sát (ObservationDate), quốc gia nhiễm (Country.Region) và số ca hồi phục
# (Recovered).
data %>% select(ObservationDate, Country.Region, Recovered)
```

KẾT QUẢ

```
> # 2. Lấy dữ liệu về số ca hồi phục ở từng quốc gia, thuộc tính hiển thị gồm: ngày quan
> # sát (ObservationDate), quốc gia nhiễm (Country.Region) và số ca hồi phục
> # (Recovered).
> data %>% select(ObservationDate, Country.Region, Recovered)
    ObservationDate Country.Region Recovered
1              <NA> Mainland China         0
2              <NA> Mainland China         0
3              <NA> Mainland China         0
4              <NA> Mainland China         0
5              <NA> Mainland China         0
6              <NA> Mainland China         0
7              <NA> Mainland China         0
```

CODE

```
# 3. Tính tổng số ca dương tính: sử dụng hàm sum(), thuộc tính: Confirmed
sum(data %>% select(Confirmed))
```

KẾT QUẢ

```
> # 3. Tính tổng số ca dương tính: sử dụng hàm sum(), thuộc tính: Confirmed
> sum(data %>% select(Confirmed))
[1] 26252051758
```

CODE

```
# 4. Tương tự câu 2, nhưng lấy ra 10 dòng đầu tiên:
head(data %>% select(ObservationDate, Country.Region, Confirmed, Deaths), 10)
```

KẾT QUẢ

```
> # 4. Tương tự câu 2, nhưng lấy ra 10 dòng đầu tiên:
> head(data %>% select(ObservationDate, Country.Region, Confirmed, Deaths), 10)
   ObservationDate Country.Region Confirmed Deaths
1             <NA> Mainland China         1      0
2             <NA> Mainland China        14      0
3             <NA> Mainland China         6      0
4             <NA> Mainland China         1      0
5             <NA> Mainland China         0      0
6             <NA> Mainland China        26      0
7             <NA> Mainland China         2      0
8             <NA> Mainland China         1      0
9             <NA> Mainland China         4      0
10            <NA> Mainland China         1      0
```

CODE

```
# THAO TÁC LỌC DỮ LIỆU VỚI FILTER:
# 1. Lấy dữ liệu về số ca nhiễm ở Trung Quốc (Mainland China):
data %>% filter(Country.Region == "Mainland China")
```

KẾT QUẢ

```
> # 1. Lấy dữ liệu về số ca nhiễm ở Trung Quốc (Mainland China):
> data %>% filter(Country.Region == "Mainland China")
   SNo ObservationDate Province.State Country.Region     Last.Update Confirmed Deaths
1    1           <NA>          Anhui Mainland China 1/22/2020 17:00         1      0
2    2           <NA>        Beijing Mainland China 1/22/2020 17:00        14      0
3    3           <NA>      Chongqing Mainland China 1/22/2020 17:00         6      0
4    4           <NA>         Fujian Mainland China 1/22/2020 17:00         1      0
5    5           <NA>          Gansu Mainland China 1/22/2020 17:00         0      0
6    6           <NA>      Guangdong Mainland China 1/22/2020 17:00        26      0
7    7           <NA>        Guangxi Mainland China 1/22/2020 17:00         2      0
8    8           <NA>        Guizhou Mainland China 1/22/2020 17:00         1      0
9    9           <NA>         Hainan Mainland China 1/22/2020 17:00         4      0
10  10           <NA>          Hebei Mainland China 1/22/2020 17:00         1      0
```

CODE

```
# 2. Lấy dữ liệu về số ca nhiễm của Việt Nam trong tháng 03 và tháng 04.
data %>% filter(Country.Region == "Vietnam" &
                ObservationDate >= "2020-03-01" & ObservationDate <= "2020-04-30")
```

KẾT QUẢ

```
> # 2. Lấy dữ liệu về số ca nhiễm của Việt Nam trong tháng 03 và tháng 04.
> data %>% filter(Country.Region == "Vietnam" &
+                 ObservationDate >= "2020-03-01" & ObservationDate <= "2020-04-30")
     SNo ObservationDate Province.State Country.Region         Last.Update Confirmed
1    691      2020-03-02                      Vietnam 2020-02-03T21:43:02         8
2    761      2020-04-02                      Vietnam 2020-02-03T21:43:02         8
3   3223      2020-03-03                      Vietnam 2020-02-25T08:53:02        16
4   3381      2020-04-03                      Vietnam 2020-02-25T08:53:02        16
5  11653      2020-03-04                      Vietnam 2020-04-03 22:52:45       237
6  11973      2020-04-04                      Vietnam         4/4/20 9:38       240
7  21373      2020-03-05                      Vietnam 2020-05-04 02:32:28       271
8  21700      2020-04-05                      Vietnam 2020-05-05 02:32:34       271
9  34464      2020-03-06                      Vietnam 2021-04-02 15:13:53       328
```

CODE

```r
# THAO TÁC THỐNG KÊ DỮ LIỆU VỚI SUMMARIES
# 1. Thống kê dữ liệu về số ca dương tính của China: số ca dương tính trung bình, t
# phương sai và độ lệch chuẩn.
data %>% filter(Country.Region == "Mainland China") %>%
  summarise(
    Mean=mean(Confirmed, na.rm = TRUE),
    Median = median(Confirmed, na.rm = TRUE),
    Variance = var(Confirmed, na.rm = TRUE),
    SD = sd(Confirmed, na.rm = TRUE)
  )
```

KẾT QUẢ

```
> # THAO TÁC THỐNG KÊ DỮ LIỆU VỚI SUMMARIES
> # 1. Thống kê dữ liệu về số ca dương tính của China: số ca dương tính trung bình, trung
> # phương sai và độ lệch chuẩn.
> data %>% filter(Country.Region == "Mainland China") %>%
+   summarise(
+     Mean=mean(Confirmed, na.rm = TRUE),
+     Median = median(Confirmed, na.rm = TRUE),
+     Variance = var(Confirmed, na.rm = TRUE),
+     SD = sd(Confirmed, na.rm = TRUE)
+   )
      Mean Median  Variance       SD
1 2590.595    396 132298968 11502.13
```

CODE

```r
# THAO TÁC GOM NHÓM DỮ LIỆU VỚI GROUP BY

# 1. Hiển thị dữ liệu theo từng ngày quan sát (thuộc tính ObservationDate) của Việt Nam
# trong 2 tháng: tháng 3 và tháng 4 năm 2020.

data %>% filter(
  Country.Region == "Vietnam" &
    ObservationDate >= "2020-03-01" &
    ObservationDate <= "2020-04-30") %>%
  group_by(ObservationDate)
```

KẾT QUẢ

```
> # trong 2 tháng: tháng 3 và tháng 4 năm 2020.
>
> data %>% filter(
+   Country.Region == "Vietnam" &
+     ObservationDate >= "2020-03-01" &
+     ObservationDate <= "2020-04-30") %>%
+   group_by(ObservationDate)
# A tibble: 22 × 8
# Groups:   ObservationDate [22]
    SNo ObservationDate Province.State Country.Region Last.Update         Confirmed Deaths
  <int> <date>          <chr>          <chr>          <chr>                  <dbl>  <dbl>
1   691 2020-03-02      ""             Vietnam        2020-02-03T21:43:02        8      0
2   761 2020-04-02      ""             Vietnam        2020-02-03T21:43:02        8      0
3  3223 2020-03-03      ""             Vietnam        2020-02-25T08:53:02       16      0
4  3381 2020-04-03      ""             Vietnam        2020-02-25T08:53:02       16      0
```

CODE

```
# THAO TÁC SẮP XẾP DỮ LIỆU VỚI ARRANGE
# 1. Hiển thị dữ liệu theo từng ngày quan sát (thuộc tính ObservationDate) của Việt
# Nam trong 2 tháng: tháng 3 và tháng 4 năm 2020. Sắp xếp theo số ca dương tính tăng
# dần.

data %>% filter(
  Country.Region == "Vietnam" &
    ObservationDate >= "2020-03-01" &
    ObservationDate <= "2020-04-30") %>%
  group_by(ObservationDate) %>%
  arrange(Confirmed)
```

KẾT QUẢ

```
+   group_by(ObservationDate) %>%
+   arrange(Confirmed)
# A tibble: 22 × 8
# Groups:   ObservationDate [22]
     SNo ObservationDate Province.State Country.Region Last.Update         Confirmed Deaths
   <int> <date>          <chr>          <chr>          <chr>                    <dbl>  <dbl>
 1   691 2020-03-02      ""             Vietnam        2020-02-03T21:43:02          8      0
 2   761 2020-04-02      ""             Vietnam        2020-02-03T21:43:02          8      0
 3  3223 2020-03-03      ""             Vietnam        2020-02-25T08:53:02         16      0
 4  3381 2020-04-03      ""             Vietnam        2020-02-25T08:53:02         16      0
 5 11653 2020-03-04      ""             Vietnam        2020-04-03 22:52:45        237      0
 6 11973 2020-04-04      ""             Vietnam        4/4/20 9:38                240      0
 7 21373 2020-03-05      ""             Vietnam        2020-05-04 02:32:28        271      0
 8 21700 2020-04-05      ""             Vietnam        2020-05-05 02:32:34        271      0
```

CODE

```
# 2. Tương tự như trên, nhưng sắp xếp số ca dương tính giảm dần. Để sắp xếp giảm dần
# một thuộc tính, ta dùng thao tác: desc(<thuộc_tính>) trong hàm arrange

data %>% filter(
  Country.Region == "Vietnam" &
    ObservationDate >= "2020-03-01" &
    ObservationDate <= "2020-04-30") %>%
  group_by(ObservationDate) %>%
  arrange(desc(Confirmed))
```

KẾT QUẢ

```
+   arrange(desc(Confirmed))
# A tibble: 22 × 8
# Groups:   ObservationDate [22]
     SNo ObservationDate Province.State Country.Region Last.Update         Confirmed Deaths
   <int> <date>          <chr>          <chr>          <chr>                    <dbl>  <dbl>
 1 170506 2020-03-12      ""             Vietnam        2021-04-02 15:13:53       1361     35
 2 171269 2020-04-12      ""             Vietnam        2021-04-02 15:13:53       1361     35
 3 148501 2020-04-11      ""             Vietnam        2021-04-02 15:13:53       1203     35
 4 147752 2020-03-11      ""             Vietnam        2021-04-02 15:13:53       1202     35
 5 124565 2020-03-10      ""             Vietnam        2021-04-02 15:13:53       1096     35
 6 125312 2020-04-10      ""             Vietnam        2021-04-02 15:13:53       1096     35
 7 102919 2020-04-09      ""             Vietnam        2021-04-02 15:13:53       1049     35
```

CODE

```r
# THÊM VÀO MỘT THUỘC TÍNH MỚI SỬ DỤNG MUTATE
# 1. Hiển thị dữ liệu theo từng ngày quan sát (thuộc tính ObservationDate) của Việt Nam
# trong 2 tháng: tháng 3 và tháng 4 năm 2020. Sắp xếp theo số ca dương tính giảm dần.
# Thêm thuộc tính Patients = số lượng ca dương tính (Confirmed) - số ca phục hồi
# (Recovered)
data %>% filter(
  Country.Region == "Vietnam" &
    ObservationDate >= "2020-03-01" &
    ObservationDate <= "2020-04-30") %>%
  group_by(ObservationDate) %>%
  arrange(desc(Confirmed)) %>%
  mutate(Patient = Confirmed - Recovered)
```

KẾT QUẢ

```
+   mutate(Patient = Confirmed - Recovered)
# A tibble: 22 x 9
# Groups:   ObservationDate [22]
     SNo ObservationDate Province.State Country.Region Last.Update         Confirmed Deaths
   <int> <date>          <chr>          <chr>          <chr>                   <dbl>  <dbl>
 1 170506 2020-03-12      ""             Vietnam        2021-04-02 15:13:53      1361     35
 2 171269 2020-04-12      ""             Vietnam        2021-04-02 15:13:53      1361     35
 3 148501 2020-04-11      ""             Vietnam        2021-04-02 15:13:53      1203     35
 4 147752 2020-03-11      ""             Vietnam        2021-04-02 15:13:53      1202     35
 5 124565 2020-03-10      ""             Vietnam        2021-04-02 15:13:53      1096     35
 6 125312 2020-04-10      ""             Vietnam        2021-04-02 15:13:53      1096     35
 7 102919 2020-04-09      ""             Vietnam        2021-04-02 15:13:53      1049     35
 8 102173 2020-03-09      ""             Vietnam        2021-04-02 15:13:53      1046     35
 9  79786 2020-04-08      ""             Vietnam        2021-04-02 15:13:53       672      8
```

CODE

```r
# LÀM SẠCH DỮ LIỆU VỚI TIDYR
data_mt <- read.csv("mtcars/mtcars.csv")

# THAOTÁC GOM DỮ LIỆU VỚI GATHER
gathered <- data_mt %>% gather(attribute, value, -model)

spread <- gathered %>% spread(attribute, value)

set.seed(1)
date <- as.Date('2016-01-01') + 0:14
hour <- sample(1:24, 15)
min <- sample(1:60, 15)
second <- sample(1:60, 15)
event <- sample(letters, 15)
data_mt <- data_mt.frame(date, hour, min, second, event)
```

KẾT QUẢ

```
# THAOTÁC GOM DỮ LIỆU VỚI GATHER
gathered <- data_mt %>% gather(attribute, value, -model)

spread <- gathered %>% spread(attribute, value)

set.seed(1)
date <- as.Date('2016-01-01') + 0:14
hour <- sample(1:24, 15)
min <- sample(1:60, 15)
second <- sample(1:60, 15)
event <- sample(letters, 15)
data_mt <- data_mt.frame(date, hour, min, second, event)
```

CODE

```
fullTime <- data_mt %>%
  unite(datehour, date, hour, sep = ' ') %>%
  unite(datetime, datehour, min, second, sep = ':')

fullTime %>%
  separate(datetime, c('date', 'time'), sep = ' ') %>%
  separate(time, c('hour', 'min', 'second'), sep = ':')
```

CODE

```
print(data[1,])
leap_year(data$ObservationDate[1])

print(data[1,])
year(data$ObservationDate[1])

print(data[1,])
month(data$ObservationDate[1])

print(data[1,])
day(data$ObservationDate[1])
```

KẾT QUẢ

```
> print(data[1,])
  SNo ObservationDate Province.State Country.Region     Last.Update Confirmed Deaths
1   1            <NA>          Anhui Mainland China 1/22/2020 17:00         1      0
  Recovered
1         0
> leap_year(data$ObservationDate[1])
[1] NA
> print(data[1,])
  SNo ObservationDate Province.State Country.Region     Last.Update Confirmed Deaths
1   1            <NA>          Anhui Mainland China 1/22/2020 17:00         1      0
  Recovered
1         0
> leap_year(data$ObservationDate[1])
[1] NA
```

```
> print(data[1,])
  SNo ObservationDate Province.State Country.Region     Last.Update Confirmed Deaths
1   1            <NA>          Anhui Mainland China 1/22/2020 17:00         1      0
  Recovered
1         0
> year(data$ObservationDate[1])
[1] NA
```

```
> print(data[1,])
  SNo ObservationDate Province.State Country.Region     Last.Update Confirmed Deaths
1   1            <NA>          Anhui Mainland China 1/22/2020 17:00         1      0
  Recovered
1         0
> month(data$ObservationDate[1])
[1] NA
```

```
> print(data[1,])
  SNo ObservationDate Province.State Country.Region     Last.Update Confirmed Deaths
1   1            <NA>          Anhui Mainland China 1/22/2020 17:00         1      0
  Recovered
1         0
> day(data$ObservationDate[1])
[1] NA
```

CODE

```
data %>% filter(Country.Region == "Vietnam" & month(ObservationDate) %in% c(1,3))

data %>% filter(Country.Region == "Vietnam" & wday(ObservationDate, label=TRUE) == "Wed")
```

KẾT QUẢ

```
> data %>% filter(Country.Region == "Vietnam" & month(ObservationDate) %in% c(1,3))
      SNo ObservationDate Province.State Country.Region         Last.Update Confirmed
1     554      2020-01-02                       Vietnam       2/1/2020 7:38         6
2     691      2020-03-02                       Vietnam 2020-02-03T21:43:02         8
3    2948      2020-01-03                       Vietnam 2020-02-25T08:53:02        16
4    3223      2020-03-03                       Vietnam 2020-02-25T08:53:02        16
5   11014      2020-01-04                       Vietnam 2020-04-01 22:04:58       218
6   11653      2020-03-04                       Vietnam 2020-04-03 22:52:45       237
7   20719      2020-01-05                       Vietnam 2020-05-02 02:32:27       270
8   21373      2020-03-05                       Vietnam 2020-05-04 02:32:28       271
9   33129      2020-01-06                       Vietnam 2021-04-02 15:13:53       328
10  34464      2020-03-06                       Vietnam 2021-04-02 15:13:53       328
11  54625      2020-01-07                       Vietnam 2021-04-02 15:13:53       355
```

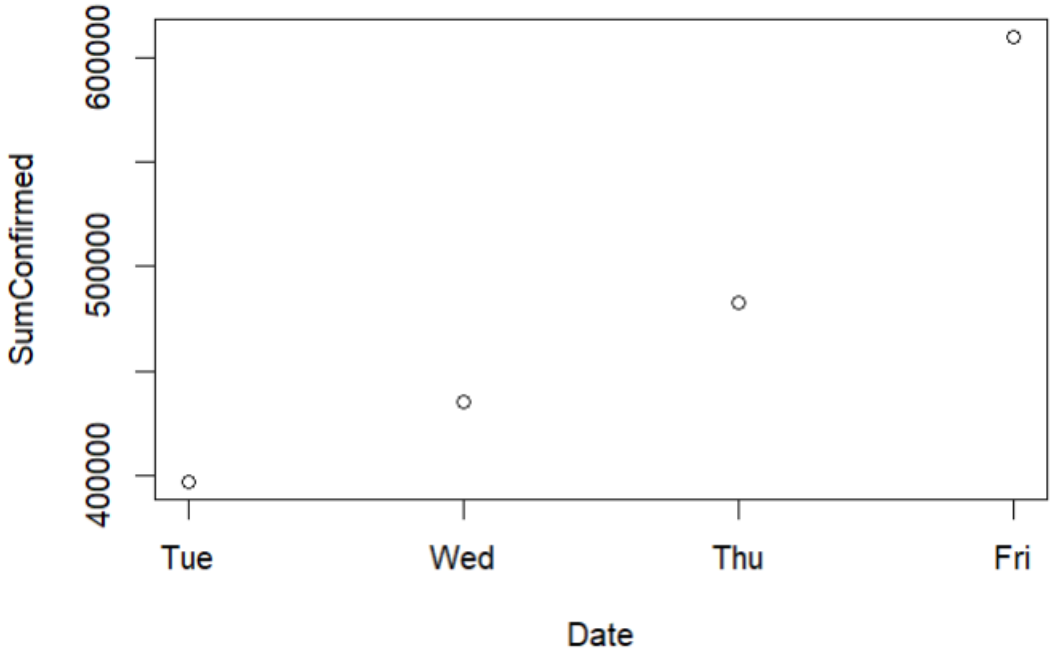
```
> data %>% filter(Country.Region == "Vietnam" & wday(ObservationDate, label=TRUE) == "Wed")
      SNo ObservationDate Province.State Country.Region         Last.Update Confirmed
1    1127      2020-09-02                       Vietnam 2020-02-08T07:23:04        13
2    1350      2020-12-02                       Vietnam 2020-02-11T16:43:06        15
3    3732      2020-06-03                       Vietnam 2020-02-25T08:53:02        16
4   11653      2020-03-04                       Vietnam 2020-04-03 22:52:45       237
5   14240      2020-11-04                       Vietnam 2020-04-11 22:52:46       258
6   21046      2020-02-05                       Vietnam 2020-05-03 02:32:28       270
7   23008      2020-08-05                       Vietnam 2021-04-02 15:13:53       288
8   35799      2020-05-06                       Vietnam 2021-04-02 15:13:53       328
9   61222      2020-10-07                       Vietnam 2021-04-02 15:13:53       370
10  77545      2020-01-08                       Vietnam 2021-04-02 15:13:53       590
11  79786      2020-04-08                       Vietnam 2021-04-02 15:13:53       672
```

b) Tìm dữ liệu về số ca nhiễm của Nhật Bản từ ngày 02/3/2021 đến ngày 15/03/2021. Vẽ biểu số ca nhiễm theo từng ngày. (sử dụng hàm plot)

CODE

```
# b) Tìm dữ liệu về số ca nhiễm của Nhật Bản từ ngày 02/3/2021 đến ngày 15/03/2021
# Vẽ biểu số ca nhiễm theo từng ngày. (sử dụng hàm plot)
data_NB_days <- data %>% filter(
  Country.Region == "Japan" &
    ObservationDate >= "2021-03-02" &
    ObservationDate <= "2021-03-15") %>%
  group_by(ObservationDate) %>% summarise(SumConfirmed = sum(Confirmed))

plot(data_NB_days$ObservationDate,data_NB_days$SumConfirmed,
      xlab = "Date", ylab = "SumConfirmed")
```

KẾT QUẢ

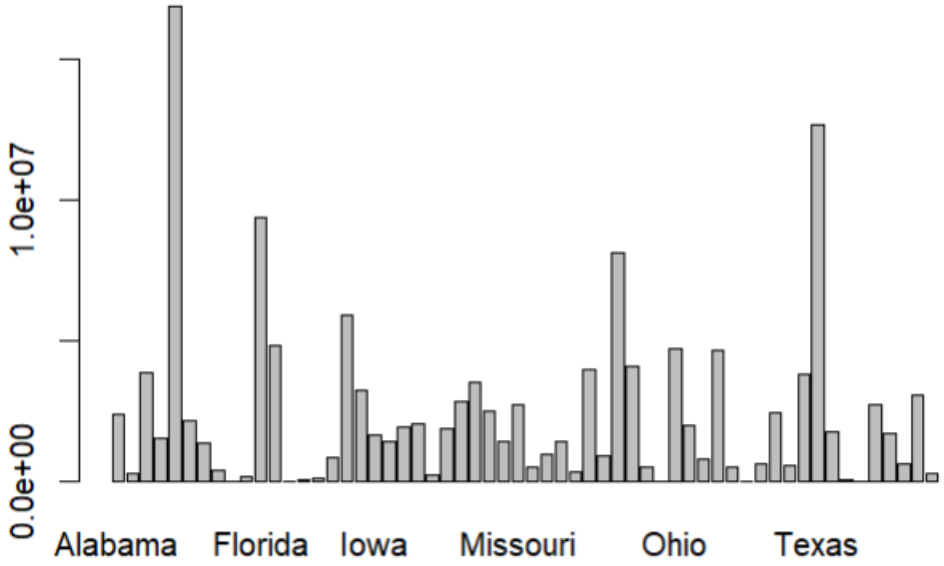


c) Tìm dữ liệu về số ca nhiễm của Hoa Kỳ từ ngày 15/03/2021 đến ngày 15/04/2021. Vẽ biểu đồ số ca nhiễm theo từng ngày (biểu đồ đường - hàm plot) và vẽ biểu đồ đếm số ca nhiễm được ghi nhận theo từng bang (biểu đồ cột - hàm barplot).

CODE

```
# c) Tìm dữ liệu về số ca nhiễm của Hoa Kỳ từ ngày 15/03/2021 đến ngày 15/04/2021.
# Vẽ biểu đồ số ca nhiễm theo từng ngày (biểu đồ đường - hàm plot) và vẽ biểu đồ đếm
# số ca nhiễm được ghi nhận theo từng bang (biểu đồ cột - hàm barplot).

data_HK_days <-data %>% filter(
  Country.Region == "US" &
    ObservationDate >= "2021-03-15" &
    ObservationDate <= "2021-04-15") %>%
  group_by(ObservationDate) %>% summarise(SumConfirmed = sum(Confirmed))

plot(data_HK_days$ObservationDate,data_HK_days$SumConfirmed,type = "l")

data_HK_Bang <-data %>% filter(
  Country.Region == "US" &
    ObservationDate >= "2021-03-15" &
    ObservationDate <= "2021-04-15") %>%
  group_by(Province.State) %>% summarise(SumConfirmed = sum(Confirmed))

barplot(data_HK_Bang$SumConfirmed, names.arg = data_HK_Bang$Province.State)
```

KẾT QUẢ

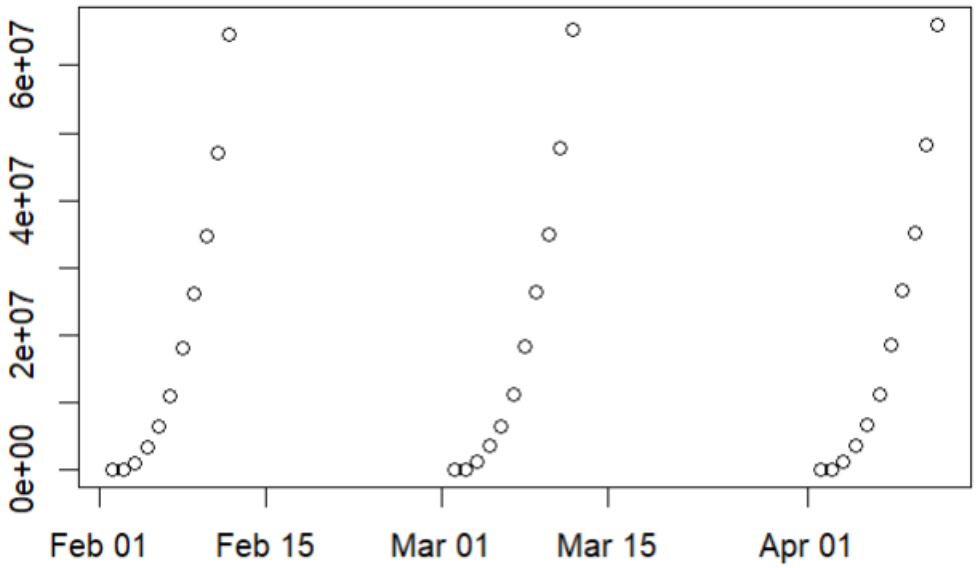


Bài 2: Thống kê số ca nhiễm mới của thế giới theo từng ngày, từ tháng 02 đến tháng 04 của năm 2020. Vẽ biểu đồ số ca nhiễm theo từng ngày (biểu đồ đường).

CODE

```
# BÀI 2. Thống kê số ca nhiễm mới của thế giới theo từng ngày, từ tháng 02 đến tháng
# 04 của năm 2020. Vẽ biểu đồ số ca nhiễm theo từng ngày (biểu đồ đường)

data_World_days <-data %>% filter(
    ObservationDate >= "2020-02-1" &
    ObservationDate < "2020-05-1") %>%
  group_by(ObservationDate) %>% summarise(SumConfirmed = sum(Confirmed))

plot(data_World_days$ObservationDate, data_World_days$SumConfirmed)
```

KẾT QUẢ

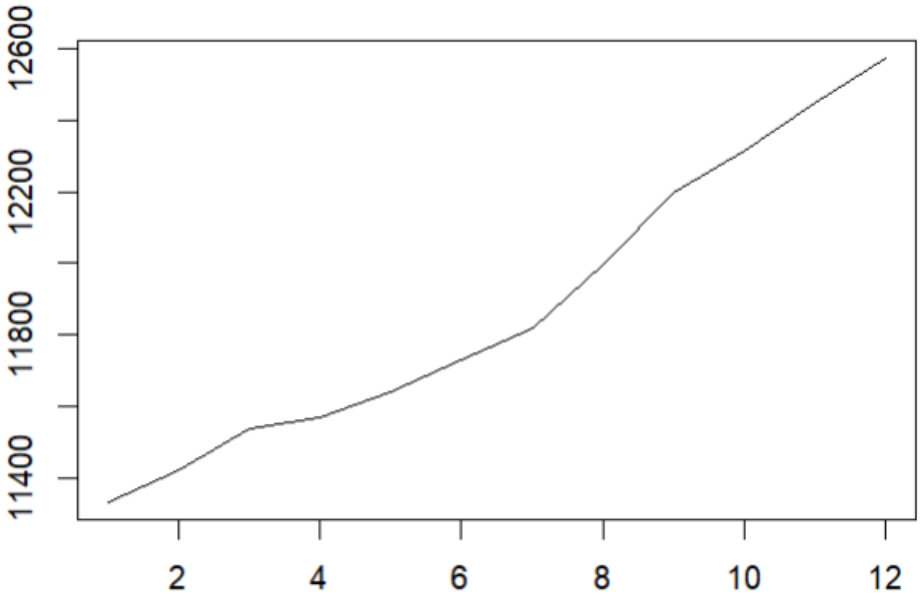


Bài 3: Thống kê số ca nhiễm mới của Việt Nam theo từng tháng trong năm 2021. Vẽ biểu đồ số ca nhiễm theo từng tháng (biểu đồ đường).Gợi ý: Tách thuộc tính ObservationDate ra thành 3 phần: ngày / tháng / năm. Sau đó gom nhóm (group by) theo từng tháng và tính tổng số ca nhiễm mới.

CODE

```r
# Bài 3: Thống kê số ca nhiễm mới của Việt Nam theo từng tháng trong năm 2021. Vẽ
# biểu đồ số ca nhiễm theo từng tháng (biểu đồ đường).

data_VN_months <-data %>% filter(
  Country.Region == "Vietnam" &
  ObservationDate >= "2021-01-1" &
    ObservationDate < "2022-01-1")
data_VN_months$month <- months(data_VN_months$ObservationDate)


data_VN_months <-data_VN_months %>%  group_by(month) %>% summarise(SumConfirmed = sum(Confirmed)) %>%
  arrange(SumConfirmed)

plot(data_VN_months$SumConfirmed, type = "l")
```

KẾT QUẢ



Bài 4: Thống kê số ca nhiễm mới của Việt nam theo từng thứ trong tháng 04 năm 2020. Gợi ý: sử dụng group by theo hàm wday (hàm wday thuộc về thư viện lubridate).

CODE

```
# Bài 4. Bài 4: Thống kê số ca nhiễm mới của Việt nam theo từng thứ trong tháng 04 năm 2020

data$Thu <- wday(data$ObservationDate)

data_VN_Thu <-data %>% filter(
  Country.Region == "Vietnam" &
    ObservationDate >= "2021-04-1" &
    ObservationDate < "2021-05-1") %>%
  group_by(Thu) %>% summarise(SumConfirmed = sum(Confirmed)) %>%
  arrange(Thu)
```

KẾT QUẢ

20521196_Nguyễn Mạnh Đức_BT4.R* | data_VN_Thu

Filter

	Thu	SumConfirmed
1	1	2631
2	2	2995
3	5	1497
4	6	1957
5	7	2488

Bài 5: Hãy thống kê số ca nhiễm mới tại Việt Nam trong khoảng tháng 01 - 03/2020 và tháng 01 - 03/2021, sử dụng tứ phân vị (dùng hàm quantile)

CODE

```
# Bài 5: Hãy thống kê số ca nhiễm mới tại Việt Nam trong khoảng tháng 01 - 03/2020
# và tháng 01 - 03/2021, sử dụng tứ phân vị (dùng hàm quantile)

data_VN_1_3_2020_2021 <-data %>% filter(
  Country.Region == "Vietnam" & ((
    ObservationDate >= "2020-01-1" &
    ObservationDate < "2020-04-1")
    |
    ObservationDate >= "2021-01-1" &
    ObservationDate < "2021-04-1"))
```

KẾT QUẢ

20521196_Nguyễn Mạnh Đức_BT4.R* | data_VN_1_3_2020_2021 | data_VN_Thu

Filter

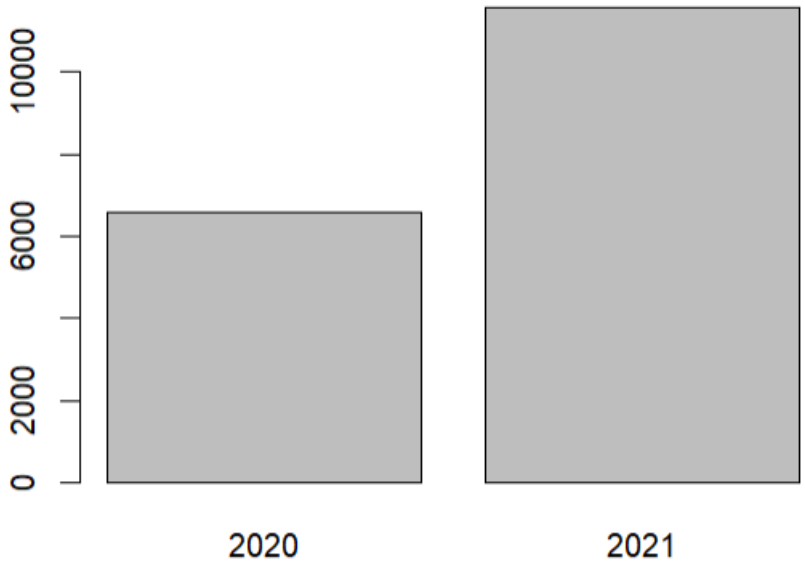
	SNo	ObservationDate	Province.State	Country.Region	Last.Update	Confirmed	Deaths	Recovered	Thu
1	554	2020-01-02		Vietnam	2/1/2020 7:38	6	0	1	5
2	623	2020-02-02		Vietnam	2020-02-01T07:38:12	6	0	1	1
3	691	2020-03-02		Vietnam	2020-02-03T21:43:02	8	0	1	2
4	2948	2020-01-03		Vietnam	2020-02-25T08:53:02	16	0	16	6
5	3078	2020-02-03		Vietnam	2020-02-25T08:53:02	16	0	16	2
6	3223	2020-03-03		Vietnam	2020-02-25T08:53:02	16	0	16	3
7	11014	2020-01-04		Vietnam	2020-04-01 22:04:58	218	0	63	7
8	11333	2020-02-04		Vietnam	4/2/20 8:53	233	0	75	3
9	11653	2020-03-04		Vietnam	2020-04-03 22:52:45	237	0	85	4
10	20719	2020-01-05		Vietnam	2020-05-02 02:32:27	270	0	219	1
11	21046	2020-02-05		Vietnam	2020-05-03 02:32:28	270	0	219	4
12	21373	2020-03-05		Vietnam	2020-05-04 02:32:28	271	0	219	5
13	33129	2020-01-06		Vietnam	2021-04-02 15:13:53	328	0	293	2

Bài 6: *Vẽ biểu đồ so sánh tổng số ca nhiễm mới của Việt nam giữa tháng 04 năm 2019, tháng 04 năm 2020 và tháng 04 năm 2021.

CODE

```r
# Bài 6: *Vẽ biểu đồ so sánh tổng số ca nhiễm mới của Việt nam giữa tháng 04 năm
# 2019, tháng 04 năm 2020 và tháng 04 năm 2021.

data_VN_4_2019_2020_2021 <-data %>% filter(
  Country.Region == "Vietnam" & ((
      ObservationDate >= "2019-04-1" &
      ObservationDate < "2019-05-1")
    |
      (ObservationDate >= "2020-04-1" &
      ObservationDate < "2020-05-1")
    |
      (ObservationDate >= "2021-04-1" &
      ObservationDate < "2021-05-1")))

data_VN_4_2019_2020_2021$Year <- year(data_VN_4_2019_2020_2021$ObservationDate)
data_VN_4_2019_2020_2021 <- data_VN_4_2019_2020_2021 %>% group_by(Year) %>%
                          summarise(SumConfirmed = sum(Confirmed))

barplot(data_VN_4_2019_2020_2021$SumConfirmed, names.arg = data_VN_4_2019_2020_2021$Year)
```

KẾT QUẢ



Bài 7: *Vẽ biểu đồ boxplot, so sánh số ca nhiễm tại Việt Nam trong khoảng tháng 01 - 03/2020 và tháng 01 - 03/2021.

CODE

```r
# Bài 7: *Vẽ biểu đồ boxplot, so sánh số ca nhiễm tại Việt Nam trong khoảng tháng 01
# - 03/2020 và tháng 01 - 03/2021.

data_VN_1_3_2020_2021$Year <- year(data_VN_1_3_2020_2021$ObservationDate)

boxplot(data_VN_1_3_2020_2021$Confirmed ~ data_VN_1_3_2020_2021$Year,
        xlab = "Year" , ylab = "Confirmed")
```

KẾT QUẢ