1. Khai báo thư viện
    - Sử dụng hàm p_load của thư viện Pacman (đã được tích hợp vào R ở các phiên bản trên 4.1.3) để cài các thư viện chưa có sẵn trong máy vào thêm các thư viện vào chương trình
    - Các thư viện sử dụng:
        - rio: Nhập xuất file
        - here: Tạo đường dẫn
        - janitor: làm sạch dữ liệu
        - lubridate: làm việc với ngày tháng
        - tidyverse: Xử lý và trực quan hóa dữ liệu
        - zoo: Cung cấp thêm một vài hàm xử lý thời gian

```
pacman::p_load(
  rio,          # importing data
  here,         # relative file pathways
  janitor,      # data cleaning and tables
  lubridate,    # working with dates
  tidyverse,    # data management and visualization
  zoo,          # additional date/time functions
)
```

2. Nội dung code
    a. Nhập file và thiết lập một số giá trị ban đầu
        - Sử dụng hàm import kết hợp here để nhập file "owid-covid-data.csv"
        - Sử dụng thư viện lubridate định dạng lại cột “date” về kiểu dữ liệu date thích hợp
        - Tạo vector “name.country” lưu tên ba nước của nhóm
        - Tạo 3 dataframe “covid.data.first”, “covid.data.second”, “covid.data.third” lần lượt lưu giữ liệu của 3 nước nhóm được giao

```
#nhap file du lieu va lam sach so qua
########
covid.data <- import(here("R", "owid-covid-data.csv"))
covid.data <- covid.data %>%
  mutate(date = lubridate::mdy(date))
name.country <- c("Canada", "Greenland", "United States")
########
#tao bo du lieu cua ba nuoc
####################################
covid.data.first <- covid.data %>%
  filter(location == name.country[1])
covid.data.second <- covid.data %>%
  filter(location == name.country[2])
covid.data.third <- covid.data %>%
  filter(location == name.country[3])
```

    b. Xử lí missing value của ba nước
        - Dùng vòng lặp for lồng nhau để xử lí dữ liệu:
            - 7 hàng đầu tiên:
                - Nếu hàng đầu tiên là dữ liệu trống, gán giá trị không cho hàng
                - Nếu hàng thứ 2 là dữ liệu trống, gán giá trị của hàng thứ nhất cho hàng

        - Các hàng sau nếu là dữ liệu trống thì lần lượt lấy trung bình của các ngày trước đó
    - Các hàng còn lại: Nếu có giá trị trống thì gán giá trị trung bình 7 ngày gần nhất
    - Lần lượt xử lí cho dữ liệu của ba nước

c. Câu 1: Biểu đồ thể hiện thu thập dữ liệu nhiễm bệnh cho từng tháng
   - Sử dụng các hàm của thư viện tidyverse và lubridate để tạo ra dataframe "covid.data.first.ncases.1.CND" chứa dữ liệu của các tháng cần biểu diễn về thời gian, số ca nhiễm của Canada đồng thời tạo thêm 2 cột chứa dữ liệu tháng và năm để thuận tiện chia nhóm khi vẽ biểu đổ

```
covid.data.first.ncases.1.CND<-covid.data.first %>%
    select(-c(iso_code, continent, location, new_deaths)) %>%#xoa cac cot khong can thiet
    filter(month(date)==2|month(date)==3|month(date)==4|month(date)==6) %>%#giu lai cac thang can bieu dien
    mutate(month = month(date), year = year(date))#tao cot month va year de lat phan nhom
```

   - Vẽ biểu đồ "cau1.ncases.CND" trực quan cho dataframe vừa tạo
     - Sử dụng hàm "ggplot" của ggplot2 để vẽ
     - Trục x là thời gian, trục y là số ca nhiễm mới
     - Dùng biểu đồ cột với thiết lập width=1 để dễ quan sát tính liên tục hơn
     - Sử dụng theme bw
     - Chia nhóm theo 2 cột tháng và năm tạo nhiều biểu đồ và sắp xếp theo quy luật, cho thang đo cột y tự do giữa các biểu đồ để dễ quan sát hơn
     - Lưu ý: một vài biểu đồ không có giá trị được biểu diễn vì giá trị quá bé với các biểu đồ còn lại hoặc không có dữ liệu tháng đó ở trong dataframe

```
cau1.ncases.CND<-ggplot(covid.data.first.ncases.1.CND, aes(x=date, y=new_cases)) +
    geom_col(width = 1) +
  labs(
    x="Time",
    y="Cases",
    title = "Number of cases for each month in Canada"
  ) +
  theme_bw() +
  scale_x_date(
    # labels should show month then date
    date_labels = "%b"
  ) +
  facet_grid(year ~ month, scales = "free_y")#phan nhom bang thang va nam
```

   - Tương tự với các nước còn lại

d. Câu 2: Biểu đồ thể hiện thu thập dữ liệu tử vong cho từng tháng
   - Tương tự câu 1, thay thế cột số ca nhiễm bằng số ca tử vong

e. Câu 4+5: Biểu đồ thể hiện thu thập dữ liệu nhiễm bệnh gồm 2 tháng cuối của năm/ Biểu đồ thể hiện thu thập dữ liệu tử vong gồm 2 tháng cuối của năm
   - Tương tự câu 1 và 2, ở bước tạo dataframe ban đầu thay đổi các giá trị tháng cần chọn

f. Câu 7: Biểu đồ thể hiện thu thập dữ liệu nhiễm bệnh tích lũy cho từng tháng
   - Thiết lập các giá trị ban đầu:
     - "months.present" chứa giá trị các tháng cần biểu diễn
     - "years.present" chứa giá trị các năm cần biểu diễn

- “covid.data.first.ncases.7.CND” với giá trị rỗng để chứa các xử lí ở bước sau

```
months.present <-c(2,3,4,6)
years.present <-c(2020,2021,2022)
covid.data.first.ncases.7.CND <- data.frame(NULL)
```

- Tạo dataframe “covid.data.first.ncases.7.CND” để vẽ biểu đồ
  - Sử dụng vòng lặp for lồng nhau với giá trị chạy trong 2 vector vừa tạo ở trên để lần lượt chạy qua các tháng và các năm cần xử lý
  - Ở mỗi tháng, sử dụng thư viện tidyverse để lọc giữ liệu của tháng đó đồng thời kết hợp hàm cumsum của base R để thay thế giá trị của cột “new_cases” chứa số ca nhiễm thành số ca nhiễm tích lũy. Thêm lần lượt các dữ liệu vừa xử lý vào “dataframe covid.data.first.ncases.7.CND” rỗng tạo ở đầu

```
#tong hop tong so ca nhiem theo tung thang va bo vao dataframe vua tao
for(i in years.present){
  for(j in months.present){
    temp <- covid.data.first.ncases.1.CND %>%
      filter(month == j&year==i) %>%
      mutate(new_cases = cumsum(new_cases))
    covid.data.first.ncases.7.CND <- bind_rows(covid.data.first.ncases.7.CND, temp)
  }
}
```

- Vẽ biểu đồ: Tương tự như các biểu đồ đã vẽ trước đó
- Làm lần lượt cho các nước còn lại

g. Câu 8: Biểu đồ thể hiện thu thập dữ liệu tử vong tích lũy cho từng tháng
   - Tương tự câu 7, thay thế số ca nhiễm bằng số ca tử vong

h. Câu 3: Biểu đồ thể hiện thu thập dữ liệu gồm nhiễm bệnh và tử vong cho từng tháng
   - Tạo dữ liệu để vẽ biểu đồ
     - Tạo dataframe “covid.data.first.ncases.3.CND” từ dataframe “covid.data.first.ncases.1.CND” đã được tạo với cột mới “Type” chứa chuỗi “new cases” đồng thời đổi tên cột “new_cases” thành number
     - Tương tự tạo dataframe “covid.data.first.ndeathss.3.CND” từ dataframe “covid.data.first.ndeaths.1.CND”
     - Tạo dataframe “covid.data.first.3” tạo từ 2 bộ dữ liệu trên ghép với nhau qua hàm “bind_rows”
   - Vẽ biểu đồ:
     - Tương tự các bài trước nhưng thêm tham số fill = Type và position =”dodge” để vẽ biểu đồ cột kề nhau của ca nhiễm và tử vong

```
cau3.CND<-ggplot(covid.data.first.3, aes(x=date, y=number, fill = Type)) +
  geom_col(
    width = 1,
    position = "dodge"
    ) +
  labs(
    x="Time",
    y="Number",
    title = "Number of new cases and new deaths for each month in Canada"
  ) +
  theme_bw() +
  scale_x_date(
    # labels should show month then date
    date_labels = "%b"
  ) +
  facet_grid(year ~ month, scales = "free_y")#phan nhom bang thang va nam
```

- Lần lượt vẽ tiếp tương tự cho các nước còn lại

i. Câu 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, tạo các dataframe để vẽ từ các dataframe đã tạo ở câu 4

j. Xuất file

- Sử dụng hàm ggsave của thư viện ggplot2, lần lượt xuất kết quả của các câu ở định dạng pdf với tên và thông số như dưới

```
#cau1
##
ggsave( filename = "task6_subtask1_Cananda.pdf",
        plot = cau1.ncases.CND,
        scale = 1,
        units = c("in"),
        dpi = 300)
```