Khoa Khoa học
và Kỹ thuật Thông tin

BÁO CÁO BÀI TẬP THỰC HÀNH 4:
THAO TÁC VÀ LÀM SẠCH DỮ LIỆU ĐÃ THU THẬP

Môn học: Thu thập và tiền xử lý dữ liệu - DS103.N21

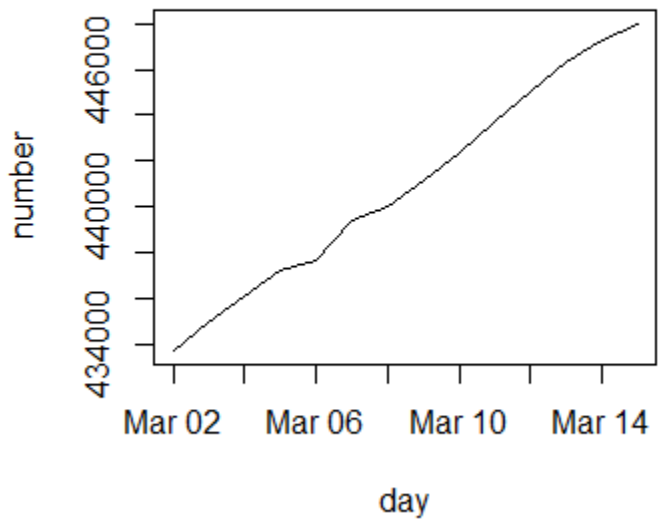
Tên: Nguyễn Thị Huyền Trang

MSSV: 21520488

Bài 1:

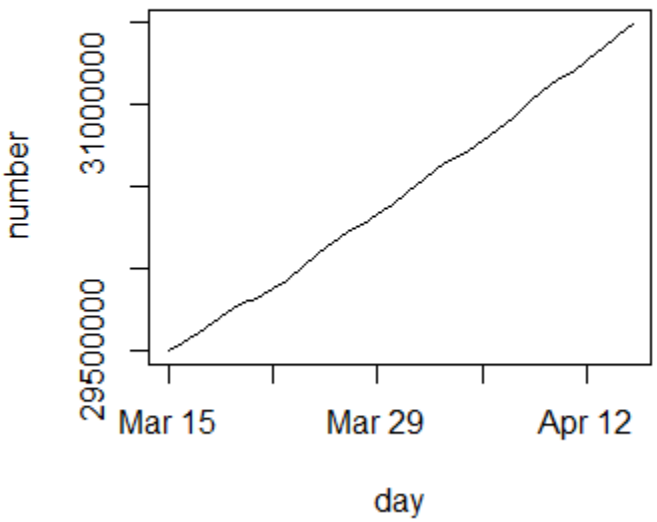
b. Tìm dữ liệu về số ca nhiễm của Nhật Bản từ ngày 02/3/2021 đến ngày 15/03/2021. Vẽ biểu số ca nhiễm theo từng ngày. (sử dụng hàm plot)

```
jp<-data %>% filter(
  Country.Region == "Japan" &
    ObservationDate >= "2021-03-02" &
    ObservationDate <= "2021-03-15") %>%
  group_by(ObservationDate)
temp <-data.frame( day=unique(jp$ObservationDate),  num=c(0))
for (i in 1:length(temp$day)){
          for (j in 1:length(jp$ObservationDate)){
                    if(temp$day[i]==jp$ObservationDate[j]){
                              temp$num[i]=temp$num[i]+jp$Confirmed[j]
                    }
           }
}
plot(x=temp$day,y=temp$num,type="l",xlab="day",ylab="number")
```

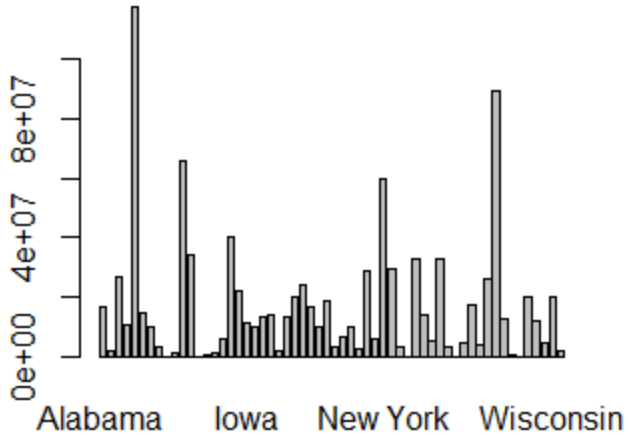


c. Tìm dữ liệu về số ca nhiễm của Hoa Kỳ từ ngày 15/03/2021 đến ngày 15/04/2021. Vẽ biểu đồ số ca nhiễm theo từng ngày (biểu đồ đường - hàm plot) và vẽ biểu đồ đếm số ca nhiễm được ghi nhận theo từng bang (biểu đồ cột - hàm barplot).

```
usa<-data %>% filter(
  Country.Region == "US" &
    ObservationDate >= "2021-03-15" &
    ObservationDate <= "2021-04-15") %>%
  group_by(ObservationDate)
temp <-data.frame( day=unique(usa$ObservationDate),  num=c(0))
for (i in 1:length(temp$day)){
            for (j in 1:length(usa$ObservationDate)){
                        if(temp$day[i]==usa$ObservationDate[j]){
                                    temp$num[i]=temp$num[i]+usa$Confirmed[j]
                        }
             }
}
plot(x=temp$day,y=temp$num,type="l",xlab="day",ylab="number")
```



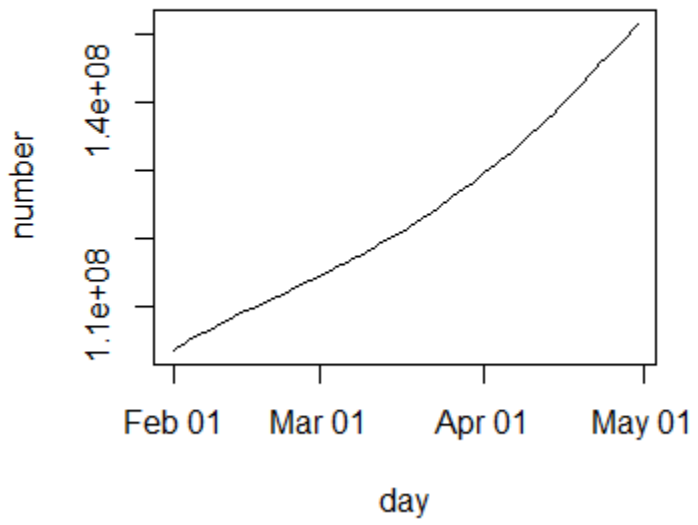
```
temp <-data.frame(  Province.State=unique(usa$Province.State), num=c(0) )
for (i in 1:length(temp$Province.State)){
        for (j in 1:length(usa$Province.State)){
                if(temp$Province.State[i]==usa$Province.State[j]){
                        temp$num[i]=temp$num[i]+usa$Confirmed[j]
                 }
        }
}
barplot(temp$num,names.arg=temp$Province.State)
```



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).

```
word<-data %>% filter(
                ObservationDate >= "2021-02-01" &
                ObservationDate <= "2021-04-30") %>%
        group_by(ObservationDate)
temp <-data.frame( day=unique(word$ObservationDate),num=c(0))
```

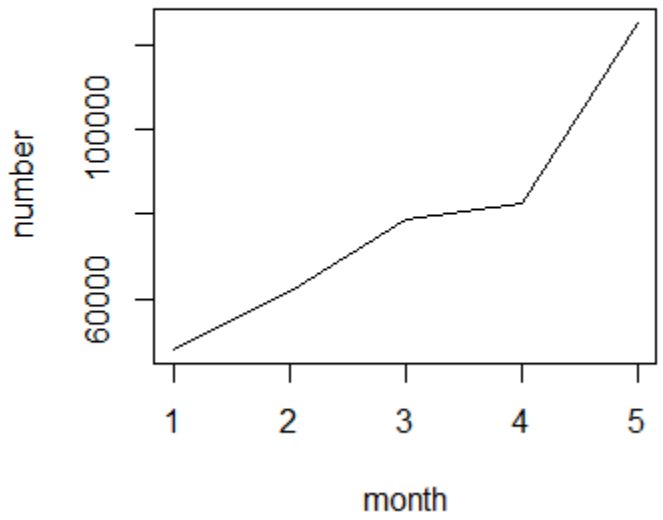
```
for (i in 1:length(temp$day)){
        for (j in 1:length(word$ObservationDate)){
                if(temp$day[i]==word$ObservationDate[j]){
                        temp$num[i]=temp$num[i]+word$Confirmed[j]
                }
        }
}
plot(x=temp$day,y=temp$num,type='l',xlab="day",ylab="number")
```



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).

```
vn<-data %>% filter(
        Country.Region == "Vietnam"
        & year(ObservationDate)==2021
 )
temp <-data.frame( month=unique(month(vn$ObservationDate)), num=c(0))
for (i in 1:length(temp$month)){
        for (j in 1:length(vn$ObservationDate)){
```

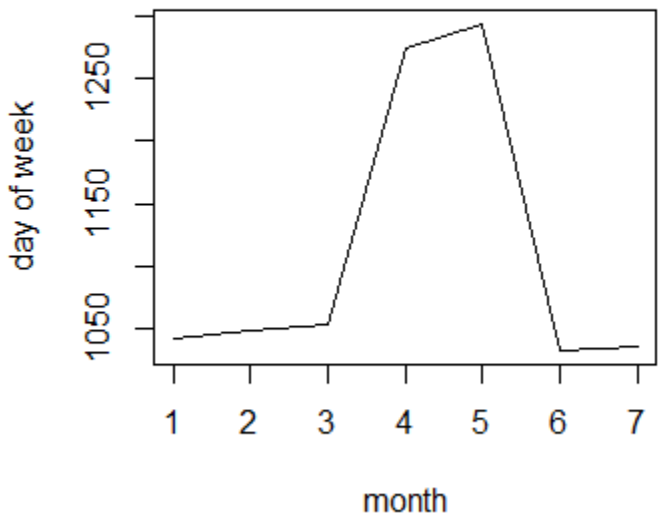
```
                if(temp$month[i]==month(vn$ObservationDate[j])){
                        temp$num[i]=temp$num[i]+vn$Confirmed[j]
                }
        }
}
plot(x=temp$month,y=temp$num,type='l',xlab="month",ylab="number")
```



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

```
vn_4<-data %>% filter(
        Country.Region == "Vietnam"
        & year(ObservationDate)==2020
        & month(ObservationDate)==4
)
temp <-data.frame( day=unique(wday(vn$ObservationDate,label=FALSE)), num=c(0))
temp<-temp[order(temp$day),]
```

```
for (i in 1:length(temp$day)){
        for (j in 1:length(vn_4$ObservationDate)){
                if(temp$day[i]==wday(vn_4$ObservationDate[j],label=FALSE)){
                 temp$num[i]=temp$num[i]+vn_4$Confirmed[j]
                }
         }
}
plot(x=temp$day,y=temp$num,type='l',xlab="month",ylab="day of week")
```



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)

```
bai_5_2021<-data %>% filter(
        Country.Region == "Vietnam"
        & ObservationDate>="2021-01-01"
        & ObservationDate<="2021-03-31"
)
quantile(bai_5_2021$Confirmed)
```

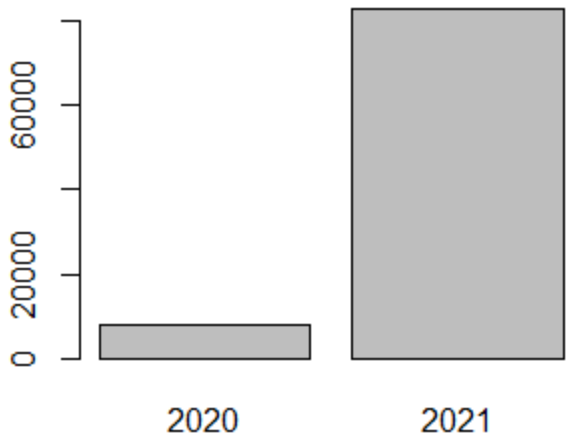
```
> quantile(bai_5_2021$Confirmed)
     0%     25%     50%     75%    100%
1474.0 1548.0 2248.5 2525.5 2603.0
```

```
bai_5_2020<-data %>% filter(
        Country.Region == "Vietnam"
        & ObservationDate>="2020-01-01"
        & ObservationDate<="2020-03-31"
)
quantile(bai_5_2020$Confirmed)
```

```
> quantile(bai_5_2020$Confirmed)
  0%  25%  50%  75% 100%
   2   13   16   53  212
```

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.

```
bai_6<-data %>% filter(
        Country.Region == "Vietnam"
        & (year(ObservationDate)==2020 || year(ObservationDate)==2021)
        & month(ObservationDate)==4
)
temp <-data.frame( year=unique(year(bai_6$ObservationDate)), num=c(0))
for (i in 1:length(temp$year)){
        for (j in 1:length(cau_6$ObservationDate)){
                if(temp$year[i]==year(cau_6$ObservationDate[j])){
                        temp$num[i]=temp$num[i]+cau_6$Confirmed[j]
                }
        }
}
barplot(temp$num,names.arg=temp$year)
```



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.

```
bai_7_2021<-data %>% filter(
        Country.Region == "Vietnam"
        & ObservationDate>="2021-01-01"
        & ObservationDate<="2021-03-31"
)
bai_7_2020<-data %>% filter(
        Country.Region == "Vietnam"
        & ObservationDate>="2020-01-01"
        & ObservationDate<="2020-03-31"
)
temp_2020 <- bai_7_2020["Confirmed"]
temp_2021 <- bai_7_2021["Confirmed"]

jpeg(filename="boxplot.jpg")
```

```
op <- par(mfrow=c(1,2))
boxplot(temp_2020,ylab="Confirmed",xlab="2020")
boxplot(temp_2021,ylab="Confirmed",xlab="2021")
par(op)
dev.off()
```