HW#4_2229027_이지민

JiminLee_2229027

2023-10-10

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.
3.2 ──
## ✓ ggplot2 3.3.6     ✓ purrr   1.0.2
## ✓ tibble  3.2.1     ✓ dplyr   1.1.3
## ✓ tidyr   1.2.1     ✓ stringr 1.4.0
## ✓ readr   2.1.2     ✓ forcats 0.5.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflict
s() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

- USA-covid19.csv 자료를 이용하여 각 문항별로 R code 와 수행결과, 그리고 결과 해석을 모두 제출하시오.

1. USA-covid19.csv 자료를 읽어 USAcovid 라는 이름으로 저장하시오.

```
USAcovid = read.csv("/Users/jimin/Desktop/데스크탑 - 이지민의 MacBook Air/ㅈ ㅣ
ㅁ ㅣㄴ/ewha/2023-2/BD/HW/USA-covid19.csv")
head(USAcovid, digits=5)
```

```
##   iso_code     continent      location       date total_cases new_cases
## 1      USA North America United States 2020-01-03          NA         0
## 2      USA North America United States 2020-01-04          NA         0
## 3      USA North America United States 2020-01-05          NA         0
## 4      USA North America United States 2020-01-06          NA         0
## 5      USA North America United States 2020-01-07          NA         0
## 6      USA North America United States 2020-01-08          NA         0
##   total_deaths new_deaths total_vaccinations new_vaccinations population
## 1           NA          0                 NA               NA  338289856
## 2           NA          0                 NA               NA  338289856
## 3           NA          0                 NA               NA  338289856
## 4           NA          0                 NA               NA  338289856
## 5           NA          0                 NA               NA  338289856
## 6           NA          0                 NA               NA  338289856
```

2. 자료의 개수는?

```
nrow(USAcovid)
```

```
## [1] 1350
```

3. 변수에는 어떤 것들이 있으며 변수의 type 은 무엇인가?

```
str(USAcovid)
```

```
## 'data.frame':    1350 obs. of  11 variables:
##  $ iso_code          : chr  "USA" "USA" "USA" "USA" ...
##  $ continent         : chr  "North America" "North America" "North America
" "North America" ...
##  $ location          : chr  "United States" "United States" "United States
" "United States" ...
##  $ date              : chr  "2020-01-03" "2020-01-04" "2020-01-05" "2020-0
1-06" ...
##  $ total_cases       : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_cases         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ total_deaths      : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_deaths        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ total_vaccinations: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ new_vaccinations  : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ population        : int  338289856 338289856 338289856 338289856 338289
856 338289856 338289856 338289856 338289856 338289856 ...
```

iso_code, continent, location, date : 범주형 자료 total_cases, new_cases, total_deaths, new_deaths, total_vaccinations, new_vaccinations, population : 연속형 변수

```
USAcovid$date = as.Date(USAcovid$date) # str type date 를 date 타입으로 변환한
다.
head(USAcovid, 2)
```

```
##   iso_code     continent      location       date total_cases new_cases
## 1      USA North America United States 2020-01-03          NA         0
## 2      USA North America United States 2020-01-04          NA         0
##   total_deaths new_deaths total_vaccinations new_vaccinations population
## 1           NA          0                 NA               NA  338289856
## 2           NA          0                 NA               NA  338289856
```

```
summary(USAcovid)
```

```
##    iso_code          continent           location              date
##
##  Length:1350        Length:1350        Length:1350        Min.   :2020-01-
03
##  Class :character   Class :character   Class :character   1st Qu.:2020-12-
05
##  Mode  :character   Mode  :character   Mode  :character   Median :2021-11-
07
##                                                           Mean   :2021-11-
07
##                                                           3rd Qu.:2022-10-
10
```

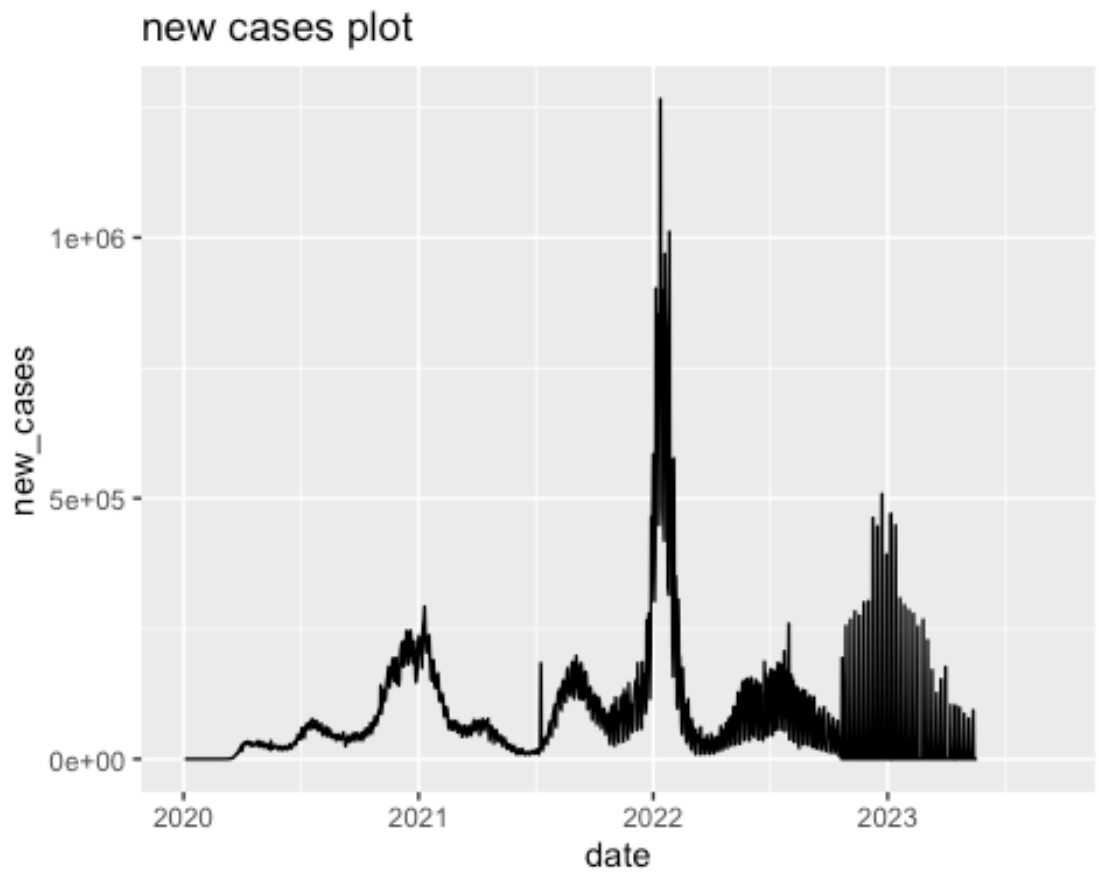
```
##                                                                 Max.   :2023-09-
13
##
```

```
##   total_cases          new_cases         total_deaths       new_deaths
##  Min.   :        1   Min.   :      0   Min.   :      1   Min.   :   0.0
##  1st Qu.: 17047944   1st Qu.:  14298   1st Qu.: 407379   1st Qu.: 121.0
##  Median : 46809945   Median :  47558   Median : 784507   Median : 661.0
##  Mean   : 54359311   Mean   :  83822   Mean   : 713128   Mean   : 916.1
##  3rd Qu.: 95602703   3rd Qu.: 111388   3rd Qu.:1056838   3rd Qu.:1368.0
##  Max.   :103436829   Max.   :1265520   Max.   :1127152   Max.   :5061.0
##  NA's   :17          NA's   :116       NA's   :57        NA's   :117
##  total_vaccinations  new_vaccinations    population
##  Min.   :    45620   Min.   :   2556   Min.   :338289856
##  1st Qu.:349277488   1st Qu.: 198058   1st Qu.:338289856
##  Median :560841574   Median : 475619   Median :338289856
##  Mean   :471846075   Mean   : 771589   Mean   :338289856
##  3rd Qu.:627458475   3rd Qu.: 985889   3rd Qu.:338289856
##  Max.   :676728782   Max.   :4581777   Max.   :338289856
##  NA's   :472         NA's   :473
```

4. new_cases 의 추이를 알아보려고 한다. 이에 알맞는 그림을 그리고 해석하시오.

```
ggplot(USAcovid, aes(x=date, y=new_cases)) +
  geom_line() +
  ggtitle("new cases plot")
```

```
## Warning: Removed 116 row(s) containing missing values (geom_path).
```

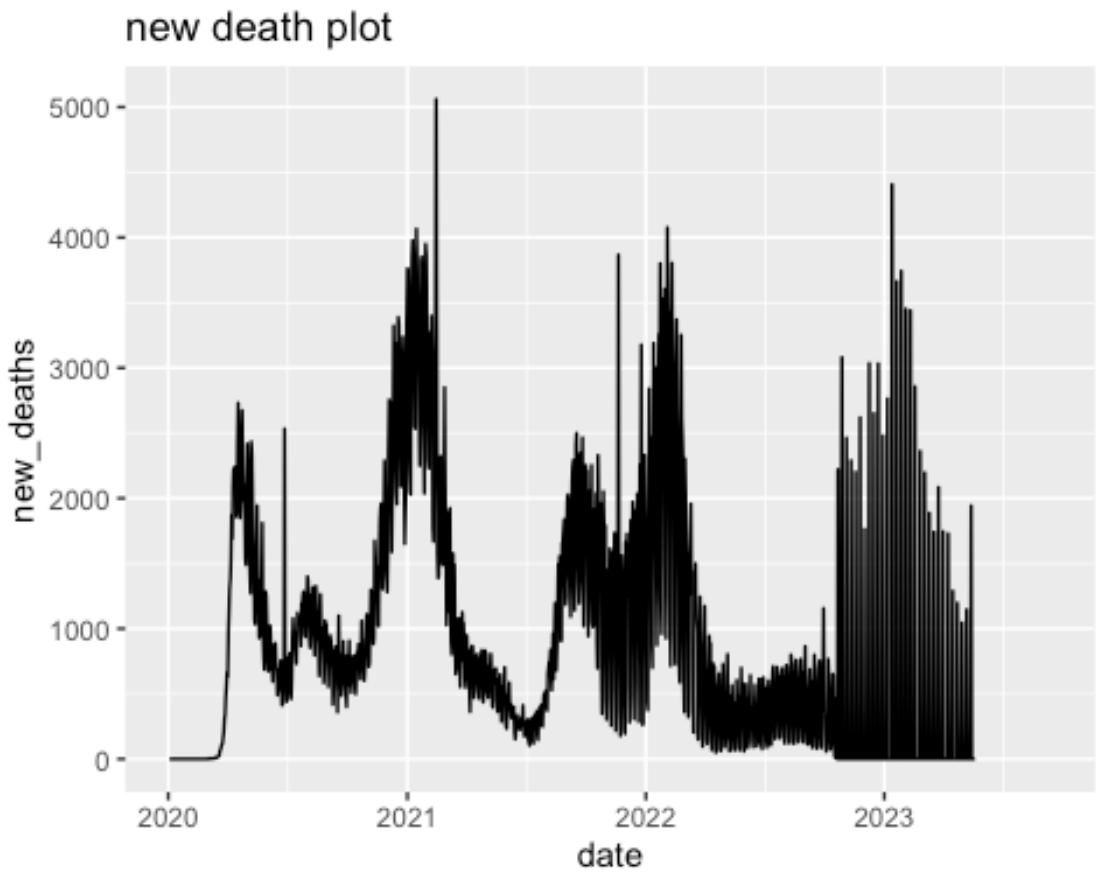


신규 확진자는 20 년도 말, 21 년도 초에 증가하는 경향을 보였으며, 2022 년 초에 폭증했다. 이후 22 년도 초만큼의 증가폭은 존재하지 않았지만, 22 년도 중순, 23 년도 초에 상승하는 경향을 보였다.

5. new_death 의 추이를 알아보려고 한다. 이에 알맞는 그림을 그리고 해석하시오.

```
ggplot(USAcovid, aes(x=date, y=new_deaths)) +
  geom_line() +
  ggtitle("new death plot")
```

```
## Warning: Removed 116 row(s) containing missing values (geom_path).
```

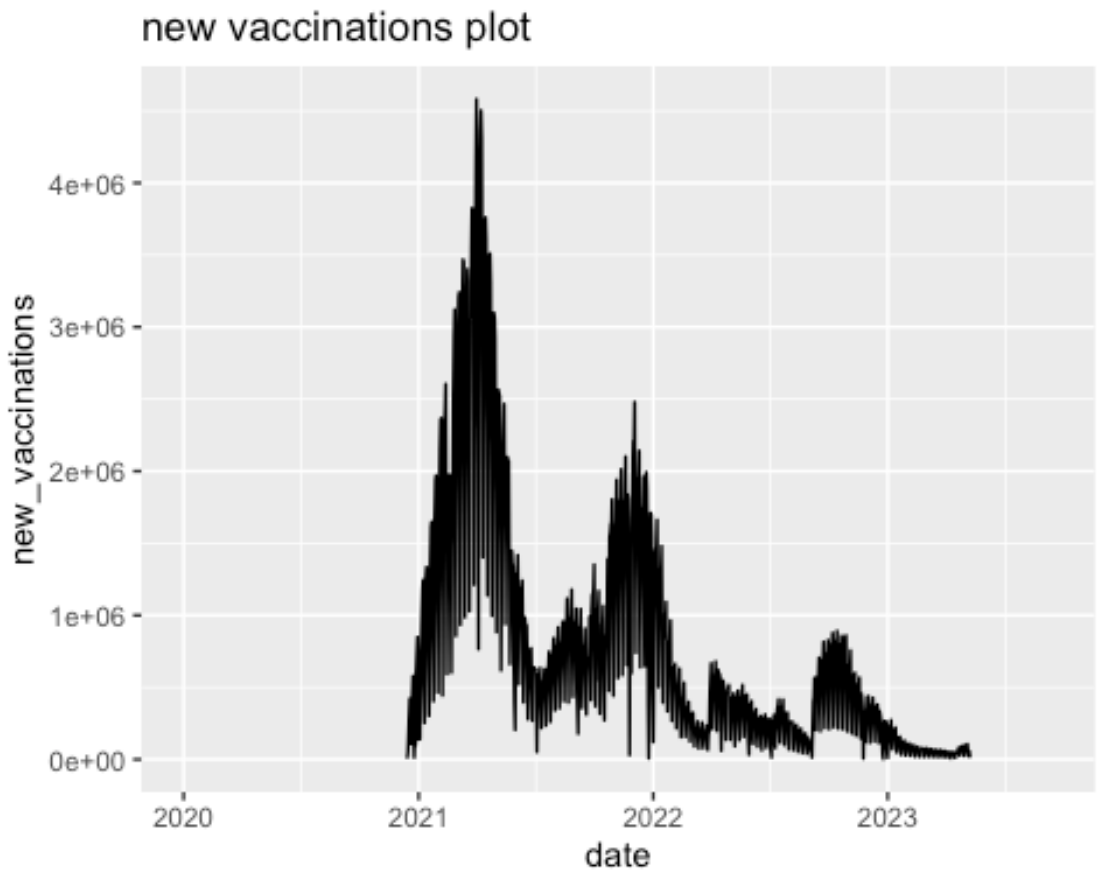


신규 사망자 수는 20 년 중순, 21 년 초, 22 년 초, 23 년 초에 증가하는 모습을 보였다.

6. new_vaccinations 의 추이를 알아보려고 한다. 이에 알맞은 그림을 그리고 해석 하시오.

```
ggplot(USAcovid, aes(x=date, y=new_vaccinations)) +
  geom_line() +
  ggtitle("new vaccinations plot")
```

```
## Warning: Removed 473 row(s) containing missing values (geom_path).
```



21 년도 상반기에 가장 많은 백신 접종이 있었고, 21 말에 두 번째로 높은 백신 접종이 있었다. 이후 22 년도 하반기에도 적지만 평소보단 많은 백신 접종이 있었다. 전체적으로 정리해보면, 20 년도 후반기 전에는 백신 접종이 없었고, 21 년도 상반기에 가장 많은 접종이 있고나선 점차 감소하는 추세를 보였다.

7. new_cases, new_death, new_vaccinations 의 추이를 함께 살펴보려고 한다. 이에 알맞은 그림을 그리고 해석하시오.

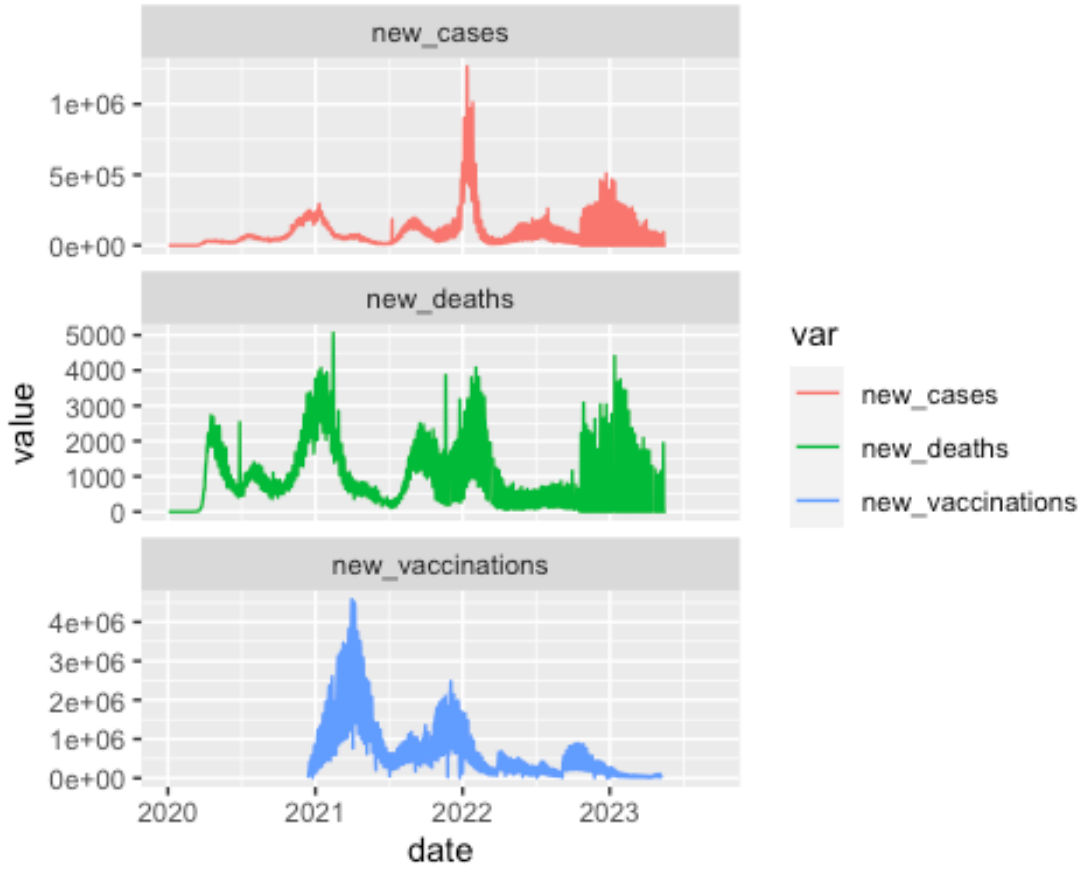
```
data.for.plot = data.frame(date = USAcovid$date,
                           var = c(rep("new_cases", length(USAcovid$new_case
s)),
                                   rep("new_deaths", length(USAcovid$new_deat
hs)),
                                   rep("new_vaccinations", length(USAcovid$ne
w_vaccinations))),
                           value = c(USAcovid$new_cases, USAcovid$new_deaths,
USAcovid$new_vaccinations))

head(data.for.plot)
```

```
##         date       var value
## 1 2020-01-03 new_cases     0
## 2 2020-01-04 new_cases     0
## 3 2020-01-05 new_cases     0
## 4 2020-01-06 new_cases     0
## 5 2020-01-07 new_cases     0
## 6 2020-01-08 new_cases     0
```

```
ggplot(data.for.plot, aes(date, value, color = var)) +
  geom_line() +
  facet_wrap(~var, ncol=1, scale="free_y")
```

```
## Warning: Removed 705 row(s) containing missing values (geom_path).
```



신규 백신 접종자는 21 년도 상반기에 제일 많았고, 신규 확진자는 22 년 초에 급증했다. 사망자 수는 확진자가 늘어나는 시기인 20 말-21 초, 22 초, 23 초에 증가하는 양상을 보인다. 눈에 띄는 점은 22 년도의 확진자 수가 다른 시기보다 압도적으로 많음에도 불구하고, 확진자 수가 증가하는 다른 시기와 사망자 수가 크게 다르지 않다는 점이다. 8. total_cases, total_deaths, total_vaccinations 의 관계를 살펴보기 위한 알맞은 그 림을 그리고 해석하시오.

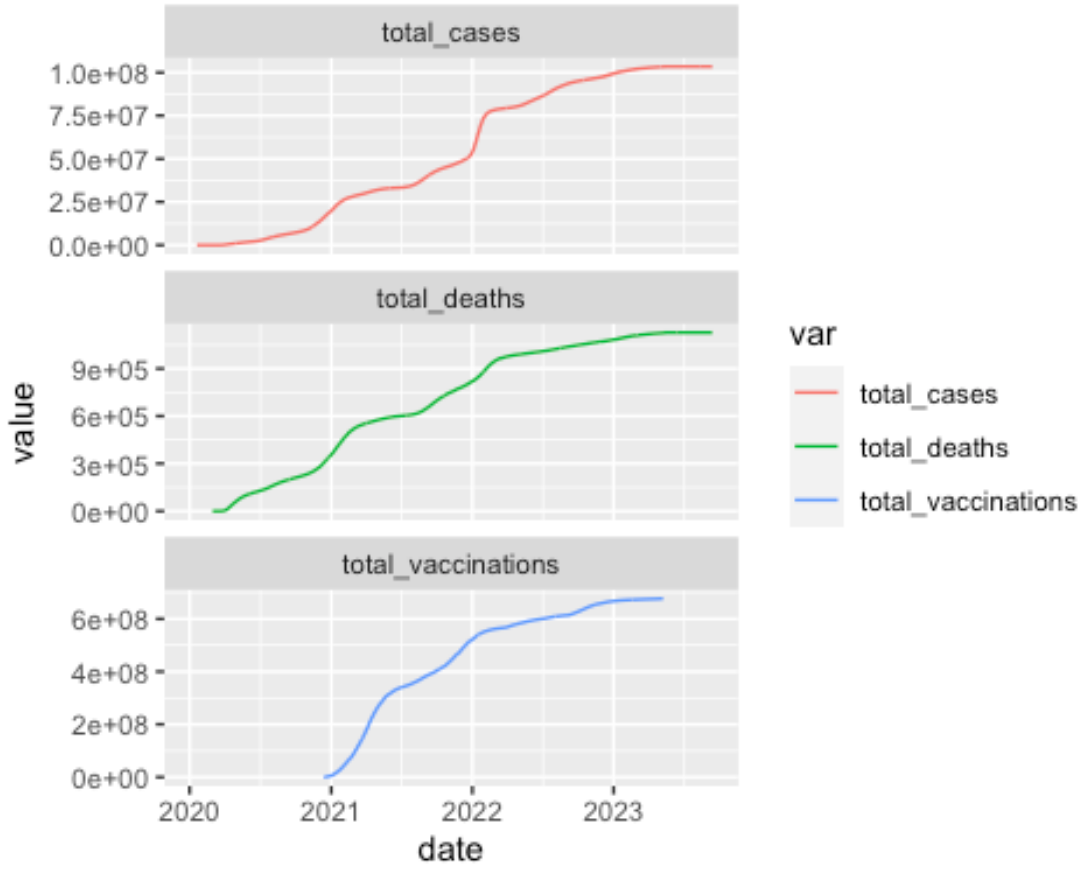
```r
data.for.plot.total = data.frame(date = USAcovid$date,
                                 var = c(rep("total_cases", length(USAcovid$total_cases)),
                                         rep("total_deaths", length(USAcovid$total_deaths)),
                                         rep("total_vaccinations", length(USAcovid$total_vaccinations))),
                                 value = c(USAcovid$total_cases, USAcovid$total_deaths, USAcovid$total_vaccinations))

head(data.for.plot.total)
```

```
##         date         var value
## 1 2020-01-03 total_cases    NA
## 2 2020-01-04 total_cases    NA
## 3 2020-01-05 total_cases    NA
## 4 2020-01-06 total_cases    NA
## 5 2020-01-07 total_cases    NA
## 6 2020-01-08 total_cases    NA
```

```r
ggplot(data.for.plot.total, aes(date, value, color = var)) +
  geom_line() +
  facet_wrap(~var, ncol=1, scale="free_y")
```

```
## Warning: Removed 546 row(s) containing missing values (geom_path).
```



누적

그래프이기 때문에 기울기를 기준으로 양상에 대해 설명한다. 먼저 전체 케이스에서 가장 급격하게 기울기가 변화하는 구간은 22 년 초다. 이를 바탕으로 22 년 초에 엄청난 수의 확진자가 생겼음을 알 수 있다. 전체 사망자 수는 21 상반기, 22 년 초에 기울기가 급격하게 변했다. 이를 통해 두 시점에서 많은 사망자가 생겼음을 알 수 있다. 전체 백신 접종의 경우, 20 년도 후부터 접종을 시작해 21 상반기에 집중적으로 접종이 이뤄졌음을 알 수 있고, 22 년 초와 하반기에도 평소보다 많은 접종이 이뤄졌음을 알 수 있다.