# REPORT

## 통계기반 데이터 분석

# 한국의 코로나 확진자 및 사망자 추이 파악 및 예측

## - CSSE 데이터셋을 바탕으로

2023.03.20

**B1 팀**

**김예지, 서영석, 이현빈, 전국림**

# <목차>

# 1. 서론

## 1) 데이터 분석 배경

지난 2020 년 부터 세계를 강타한 코로나 19 바이러스는 그 영향이 안정적으로 변화되고 있다. 이러한 양태는 한국에서도 비슷하게 진행되고 있다. 이 보고서에서는 데이터를 기반으로 한국의 일별 코로나 확진자 및 사망자 분석을 통해 한국의 코로나 19 확진자 및 사망자의 추이를 파악하고자 한다. 이를 통해 향후 한국의 코로나 19 확진자 및 사망자 발생 추이를 예측하여, 결과적으로 시기별 방역 대책 수립을 위한 데이터적 근거 기반을 형성하고자 한다.

## 2) 데이터 분석 설명

이 보고서에서는 R 을 활용하여 세계 코로나 19 확진자 및 사망자 데이터를 분석할 예정이다. 조사 기간을 '2020 년 3 월 ~ 2022 년 7 월'로 기간을 고정하여, 약 3 년 동안의 데이터만 활용한다. 데이터를 분석하는 과정에서 파악된 데이터 자체의 기재 오류는 '0' 으로 변환하여 분석했다. 이러한 제한을 통해 분석 결과에 대한 일관성과 정확성을 높였다. 여기서 활용한 데이터는 다양한 칼럼과 변수가 존재한다. 이해를 돕기 위해, 가장 활용도가 높은 칼럼과 변수명에 대해 우선적으로 선언하고자 한다.

| 칼럼 선언 | |
|---|---|
| **칼럼명** | **의미** |
| Deaths | 사망자 수 |
| Confirmed | 확인자 수 |
| dayDeaths | 일별 사망자 수 |
| dayConfirmed | 일별 확인자 수 |
| date | 데이터 해당 날짜 |

| 변수 선언 | |
|---|---|
| **변수명** | **의미** |
| korea20 | 2020 년 한국 코로나 환자 데이터 |
| korea21 | 2021 년 한국 코로나 환자 데이터 |
| korea22 | 2022 년 한국 코로나 환자 데이터 |
| koreas | 2020 년~2022 년 한국 코로나 환자 데이터 |
| subks | 일별 코로나 환자 확인을 위한 이전 일자 데이터 |
| onekoreas | 일별 코로나 환자 확인을 위한 이후 일자 데이터 |
| koreatotal / kor | 2020 년~2023 년 한국의 일별 코로나 환자 데이터 |

### 3) 데이터 정제

해당 데이터는 세계 200 여 국가을 표본으로 한다. 따라서 보고서의 목적인 '한국의 일별 코로나 확진자 및 사망자 추세'를 파악하기 위해서는 분석에 앞서 데이터 정제가 필요하다.

해당 부분에 대한 코드는 아래와 동일하며, 아래에서는 한국의 2020 년~2022 년 확진자 및 사망자에 대한 데이터셋의 정보를 'korea20','korea21','korea22'라는 이름의 변수에 각각 담았다.

```
#covid 라는 data.fram()형 변수 생성
covid20 <- data.frame()
covid21 <- data.frame()
covid22 <- data.frame()

#src_dir 에는 파일 위치를, src_file 에는 파일 리스트를 담음
src_dir <- c("C:/Rwork/project1/data")
src_file20 <- list.files(src_dir,pattern="*-2020.csv")
src_file21 <- list.files(src_dir,pattern="*-2021.csv")
src_file22 <- list.files(src_dir,pattern="*-2022.csv")

#for 문을 돌려서 모든 파일을 하나의 변수, covid 로
for (i in 1:length(src_file21)){
  if( i <=length(src_file20)){
    covid_temp20 <- read.csv(
      paste0(src_dir, "/", src_file20[i]),
      sep=",",
      header=T,
      stringsAsFactors=F)
    # 새로운 컬럼 date 생성
    covid_temp20$date <- src_file20[i]
    covid20 <-  bind_rows(covid20,covid_temp20)
  }

  if( i <=length(src_file21)){
    covid_temp21 <- read.csv(
      paste0(src_dir, "/", src_file21[i]),
      sep=",",
      header=T,
      stringsAsFactors=F)
    # 새로운 컬럼 date 생성
    covid_temp21$date <- src_file21[i]
    covid21 <-  rbind(covid21,covid_temp21)
  }

  if( i <=length(src_file22)){
    covid_temp22 <- read.csv(
      paste0(src_dir, "/", src_file22[i]),
      sep=",",
```

```
      header=T,
      stringsAsFactors=F)
    # 새로운 컬럼 date 생성
    covid_temp22$date <- src_file22[i]
    covid22 <-  rbind(covid22,covid_temp22)
  }
}
# 한국 추출
korea20 <- covid20 %>% subset(Country_Region == "Korea, South"|Country_Region == "Republic of Korea"|
                                Country.Region == "Korea, South"|Country.Region == "Republic of Korea"|Country_Region == "South Korea"|Country.Region == "South Korea")
korea21 <- covid21 %>% subset(Country_Region == "Korea, South"|Country_Region == "Republic of Korea"|Country_Region == "South Korea")
korea22 <- covid22 %>% subset(Country_Region == "Korea, South"|Country_Region == "Republic of Korea"|Country_Region == "South Korea")
```

이후 한국의 데이터는 ''koreas'에 담아 이후 활용하고 있다.

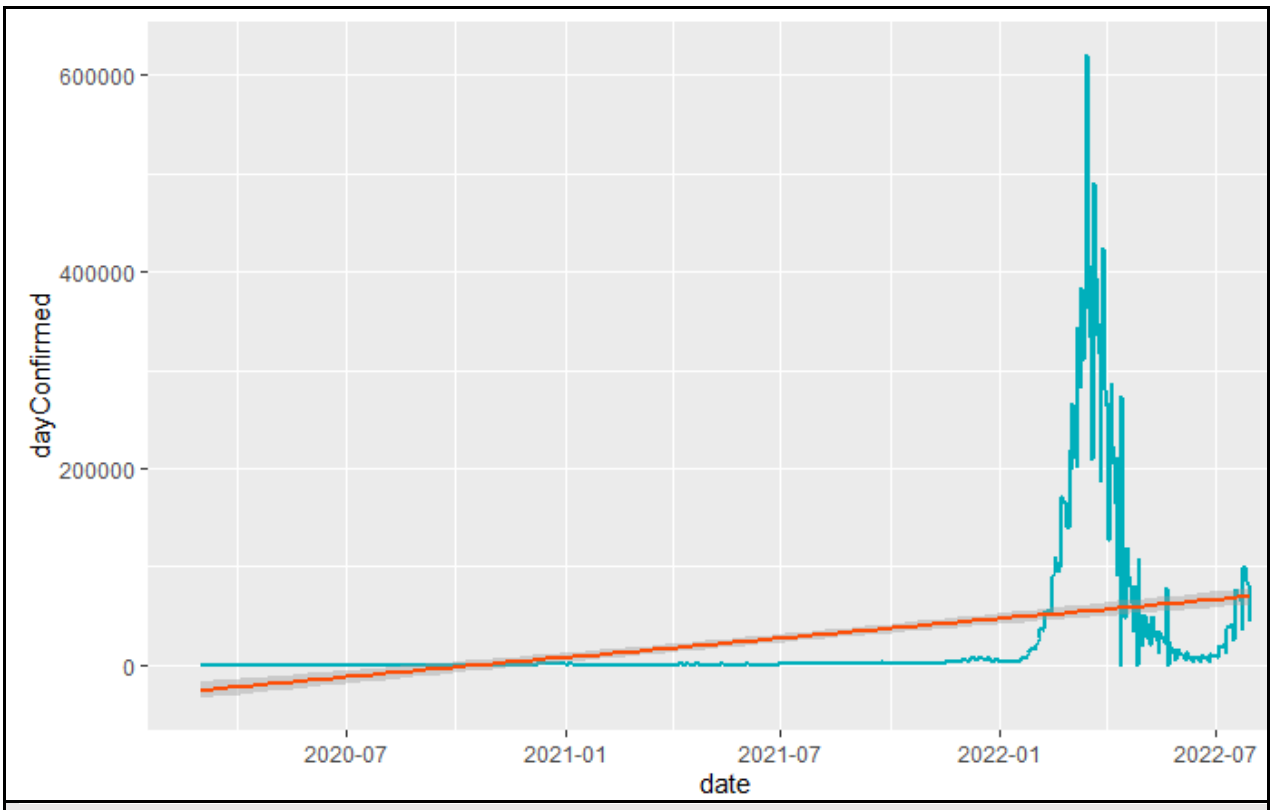
## 2. 본론

### 1) 한국의 코로나 환자 추세 분석

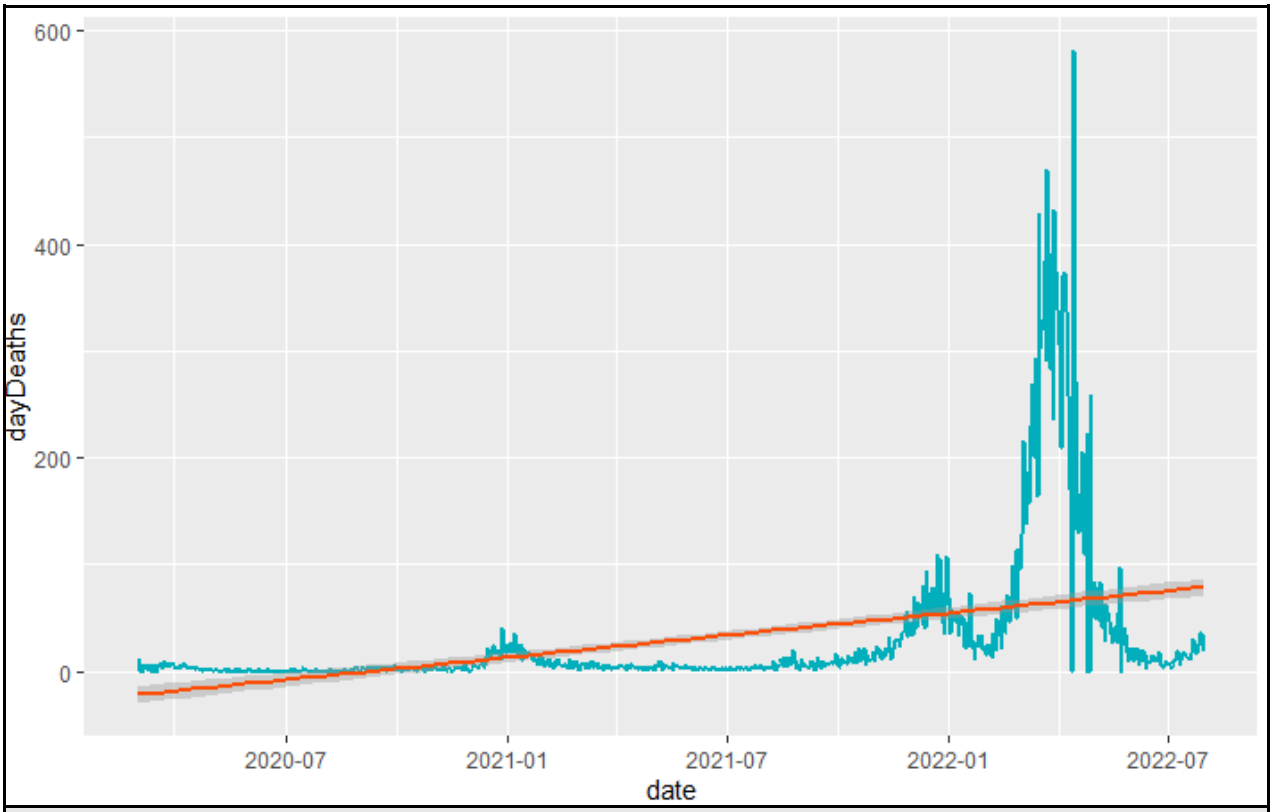
한국의 조사 기간 동안 코로나 환자 추세에 대하여 확인하고자 한다. 따라서 추세선 및 시계열 그래프를 통해 코로나 확진자 및 사망자에 대해서 파악했다. 여기서 x 축은 시간(일자), y 축은 각각 일별 확진자, 일별 사망자 수를 의미한다. 이를 통해 일자에 따른 일별 확진자 및 사망자 수에 대한 추세를 확인할 수 있다. 코드는 아래와 동일하다.

```
# 시계열 그래프 그리기
dayConfGrape <- ggplot(data = kor, aes(x = date, y = dayConfirmed))
                                    +geom_line(color = "#00AFBB", size = 1)
dayDGrape <- ggplot(data = kor, aes(x = date, y = dayDeaths))
                                     +geom_line(color = "#00AFBB", size = 1)

##선형 추세선 추가
dayConfChu <- dayConfGrape+ stat_smooth(color = "#FC4E07", method = "lm")
dayConfChu
dayDChu <- dayDGrape+ stat_smooth(color = "#FC4E07", method = "lm")
dayDChu
```

위의 코드에 대한 결과 값은 아래 사진과 동일하게 출력된다.




**일별 코로나 확진자 추세**




**일별 코로나 사망자 추세**

여기서 붉은 색으로 표시된 그래프는 추세선을 의미한다. 또한 파란색의 그래프는 정확한 수치를 선으로 이은 그래프이다. 이를 통해 일별 코로나 확진자 및 사망자는 2020 년 3 월부터 지속적으로 증가 추세에 있음을 알 수 있다.

이 때, 2022 년 2 월~4 월이라는 기간 동안 코로나 확진자 및 사망자는 급증한 것을 확인할 수 있다. 이는 당시 한국을 비롯한 세계에 오미크론 변이 바이러스가 확산되고 있었던 결과로 볼 수 있다.

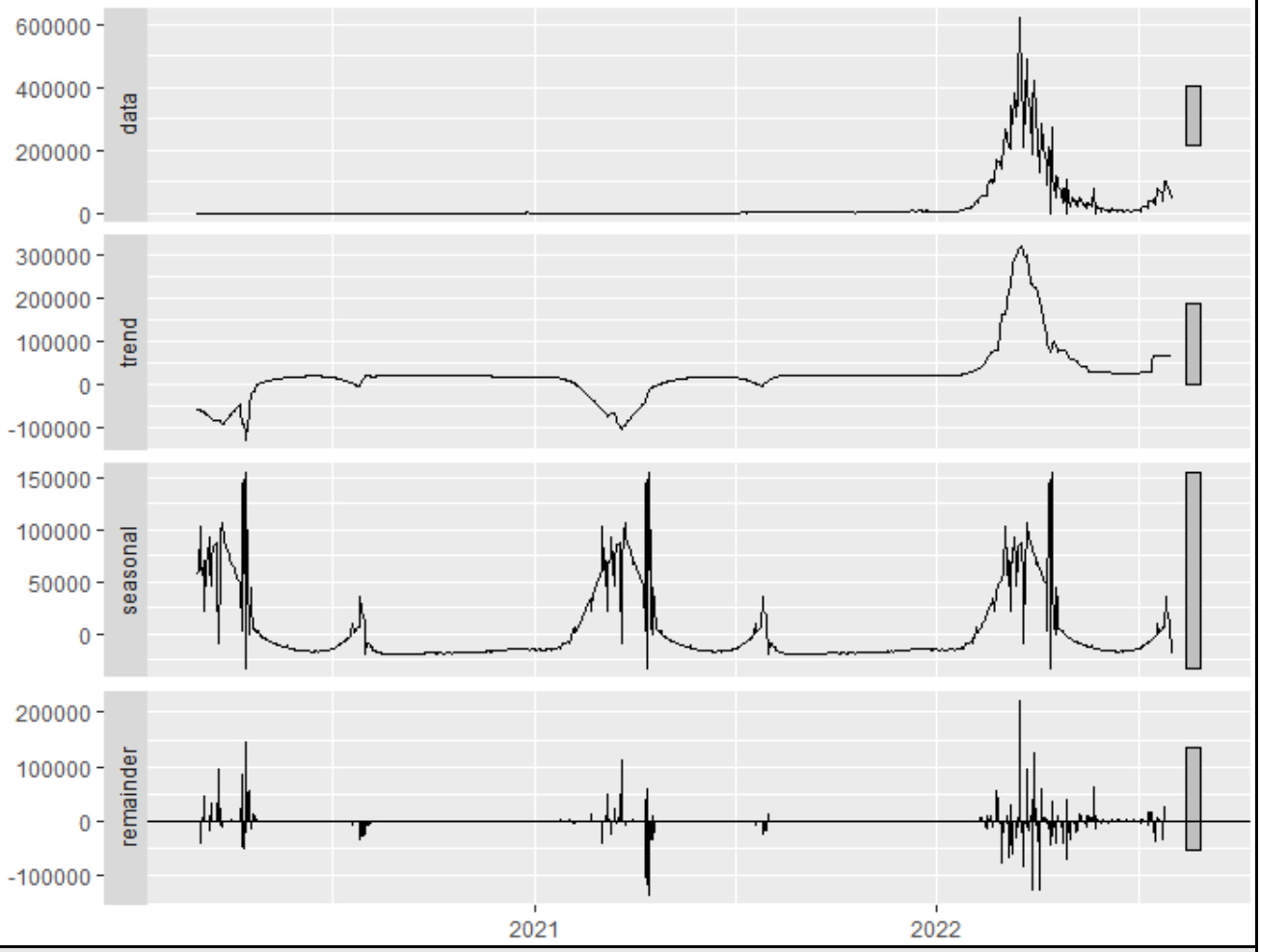
### 2) 변동 요인에 따른 추세 분석 및 예측

이어서 기본, 계절변동, 추세변동, 잔차(오차)에 따른 특징을 시각화하고자 한다. 이러한 4 가지 변동요인 분해를 통해, 변동 요소들을 찾아내고 시계열자료를 그 요소들의 결합으로 표현한 후 장래시점에 대해 예측한다. 계절 변동은 1 년 단위의 반복되어지는 특징을 가지며, 추세변동은 증가를, 잔차는 관측치와 예측치 사이의 오차를 의미한다.

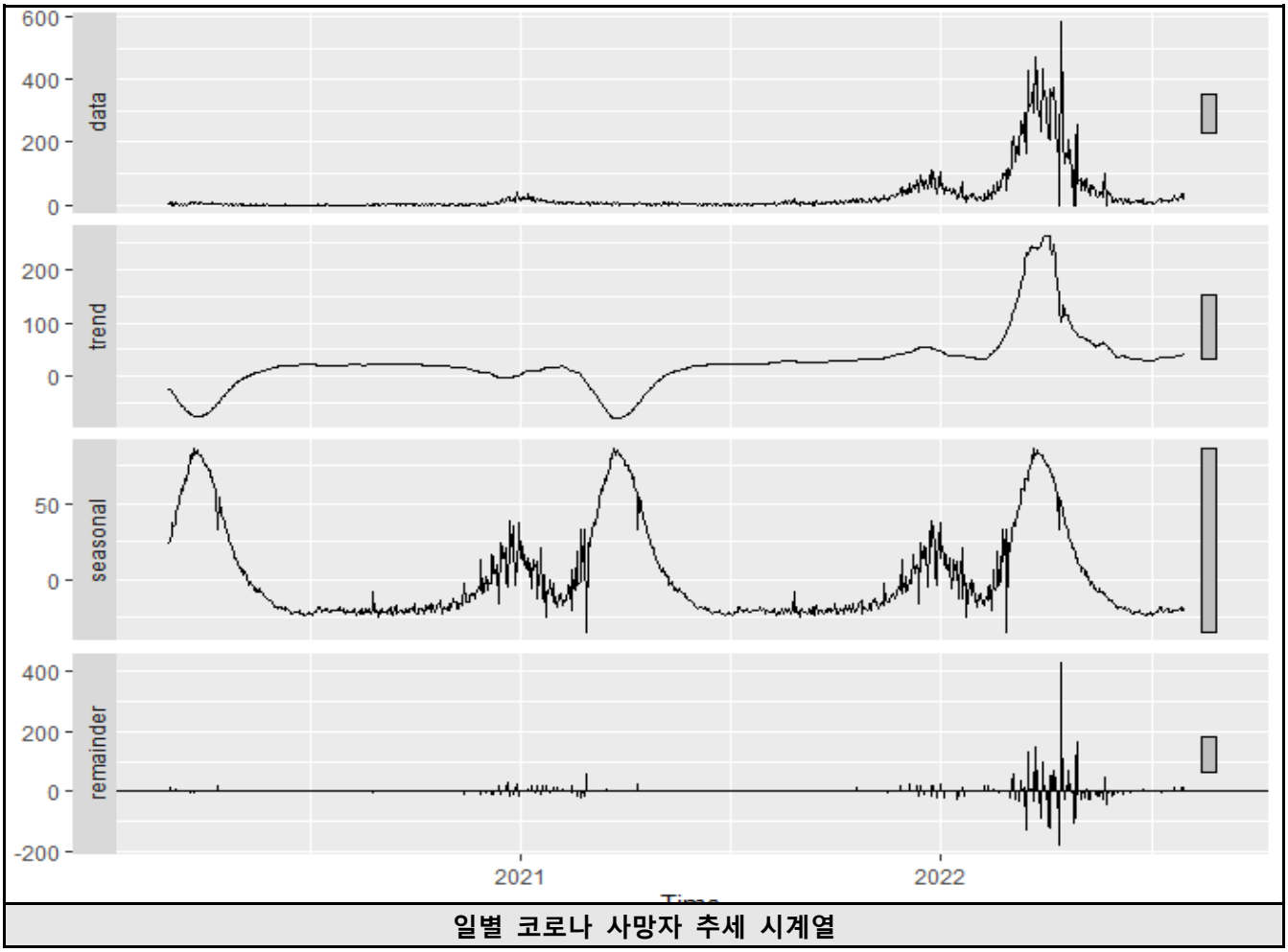
서론에서 밝힌 바와 같이, 여기서 분석에 사용되는 데이터 자료는 2020 년 3 월 1 일부터 시작된다. 또한 3 년 동안의 데이터 셋이므로 반복주기(N)을 365 로 설정하였다. 이를 바탕으로 한 코드는 아래와 동일하다.

```
library(lubridate)
library(seasonal)
library(zoo)
zooCon <- kor$dayConfirmed
zooDea <- kor$dayDeaths

# time series 변환
tsCon <- as.ts(zooCon) %>% ts(start = decimal_date(as.Date("2020-03-01")), frequency = 365)
tsDea <- as.ts(zooDea) %>% ts(start = decimal_date(as.Date("2020-03-01")), frequency = 365)

#  4 가지 변동요인 분해
tsCon %>% stl(t.window = 13, s.window = "periodic", robust=T) %>% autoplot()
tsDea %>% stl(t.window = 13, s.window = "periodic", robust=T) %>% autoplot()
```

tsCon 은 일별 확진자 수를 의미하며, tsDea 는 일별 사망자 수를 의미한다. 또한 위의 코드에 대한 결과 값은 아래 사진과 동일하게 출력된다.




**일별 코로나 확진자 추세 시계열**




**일별 코로나 사망자 추세 시계열**

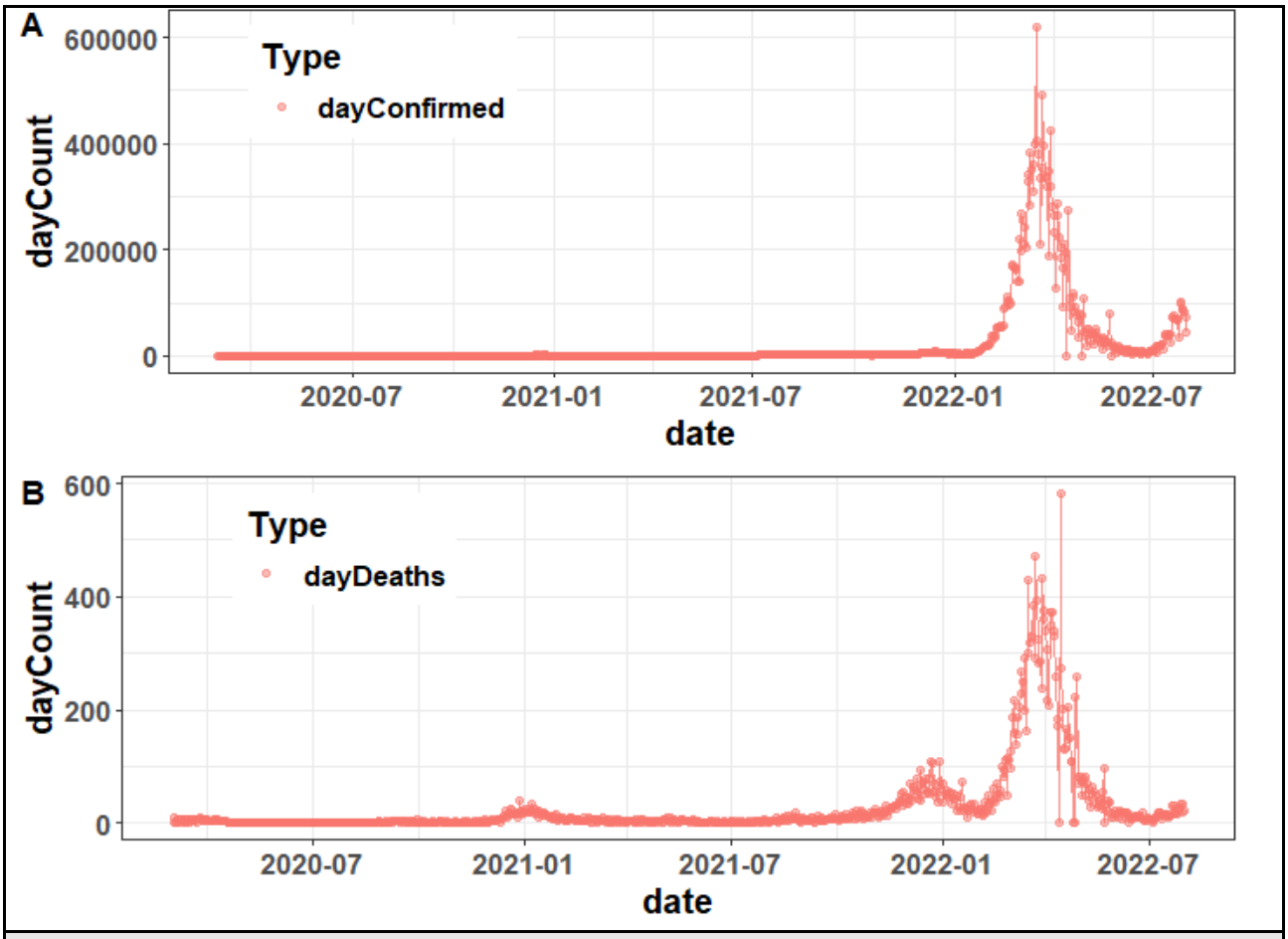
이러한 결과로 보아, 코로나 확진자는 계절적 추세를 지니고 있는 것으로 확인할 수 있다. 즉 겨울이 여름보다 높은 사망자 및 확진자가 나타나고 있다. 이를 위해서는 월별 데이터로 파악하는 것이 보다 용이하다. 이는 일자별 시각화 이후 구현했다.

한국의 일별 사망자와 확진자의 추세에 대한 보다 명확한 이해를 위해, 분기별 일자를 포함하여 시각화하였다. 해당 코드는 아래와 동일하다. 여기서 A 는 일별 확진자 수, B 는 일별 사망자 수에 대한 그래프를 의미한다.

```
library(reshape)
library(cowplot)
A = kor   %>%
  select(-dayDeaths) %>%
  melt(id.vars = c("date")) %>%
  ggplot() +
  geom_point(aes(x = date, y = value, col = variable),
             alpha =  0.5) +
  geom_line(aes(x = date, y = value, col = variable, group = variable),
```

```
                alpha = 0.8) +
  xlab("date") + ylab("dayCount") +
  labs(col = "Type") +
  theme_bw() +
  theme(text = element_text(size = 15, face = "bold"),
        legend.position = c(0.2,0.8))

B = kor   %>%
  select(-dayConfirmed) %>%
  melt(id.vars = c("date")) %>%
  ggplot() +
  geom_point(aes(x = date, y = value, col = variable),
             alpha =  0.5) +
  geom_line(aes(x = date, y = value, col = variable, group = variable),
            alpha = 0.8) +
  xlab("date") + ylab("dayCount") +
  labs(col = "Type") +
  theme_bw() +
  theme(text = element_text(size = 15, face = "bold"),
        legend.position = c(0.2,0.8))

cowplot::plot_grid(A,B,ncol = 1,labels = c("A","B"))
```

코드에 대한 결과값은 아래 사진과 동일하다.



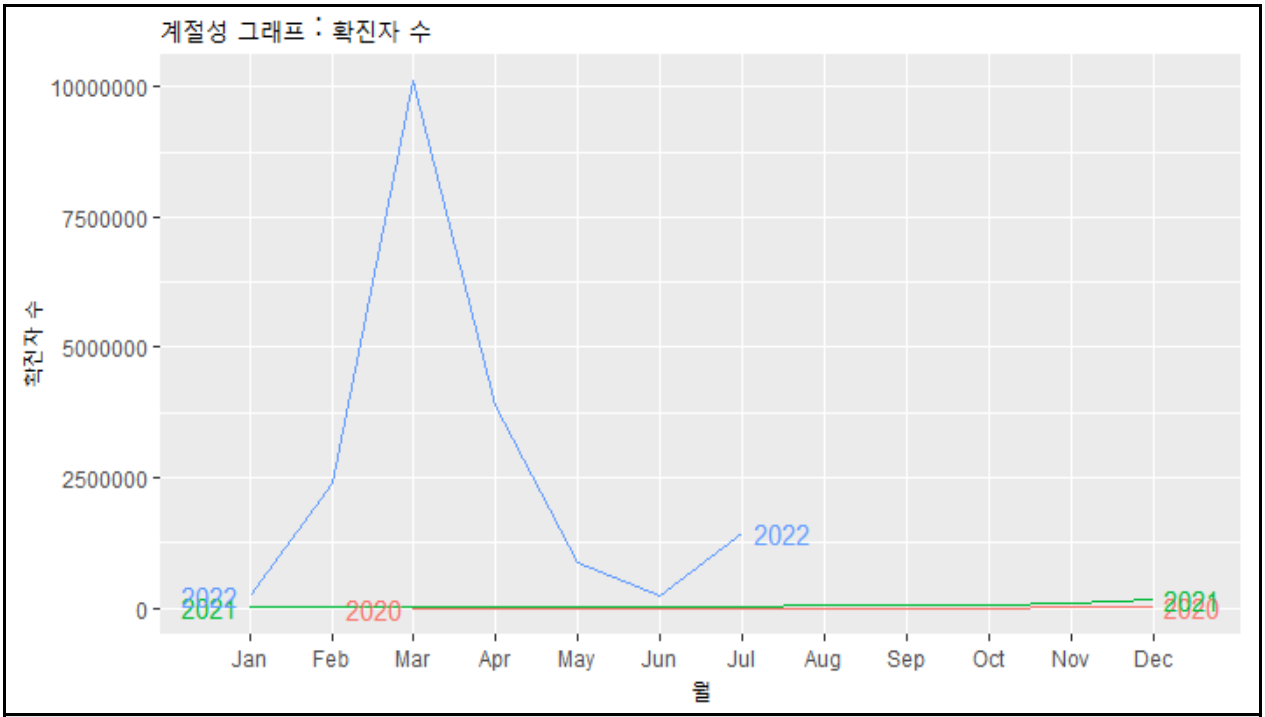

**일별 확진자 및 사망자 그래프**

앞서 언급한, 점차 증가하는 계절 요인에 대해서 보다 자세히 살펴보도록 한다. 이를 위해 먼저 일별로 나누어져 있는 데이터를 월별 기준으로 통합하였다. 이후 월별 확진자 및 사망자 수에 대한 그래프를 구현했다. 이를 위한 코드는 아래와 동일하다.

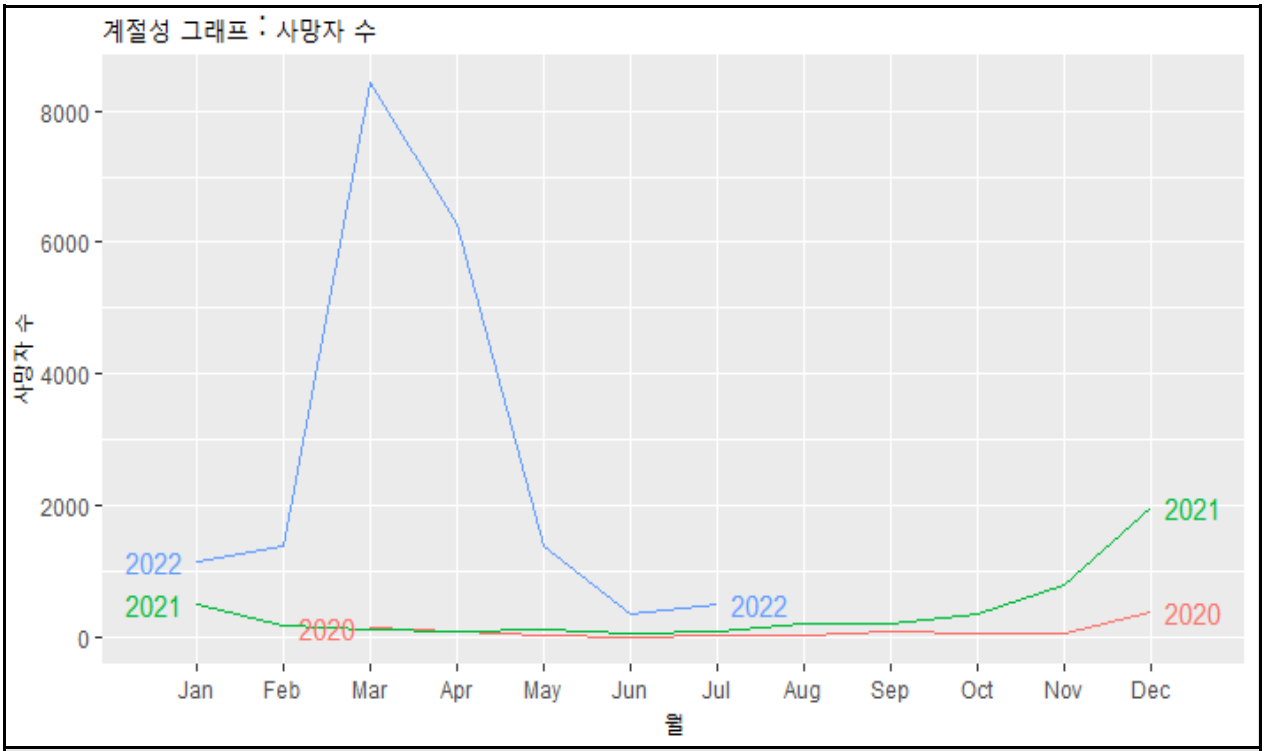
```
#월 단위로 데이터 묶기
kor$month <- factor(format(kor$date, "%Y-%m"))
#월별로 사망자 확진자 합침
kormonth <- aggregate(kor[,c('dayDeaths','dayConfirmed')],
          by=list(kor$month),FUN = sum)
names(kormonth) <- c("month","Deaths","Confirmed")
# 그래프 구현
zooCon1 <- kormonth$Confirmed
zooDea1 <- kormonth$Deaths
kormonth

tsCon1 <- as.ts(zooCon1) %>% ts(start = c(2020,3), frequency = 12)
tsDea1 <- as.ts(zooDea1) %>% ts(start = c(2020,3), frequency = 12)
```

```
tsCon1 %>% stl(t.window = 13, s.window = "periodic", robust=T) %>% autoplot()
tsDea1 %>% stl(t.window = 13, s.window = "periodic", robust=T) %>% autoplot()

# install.packages("fpp2")
library(fpp2)
plot(tsCon1)
ggseasonplot(tsCon1, year.labels = TRUE, year.labels.left = TRUE) + ylab("확진자 수") + xlab("월") + ggtitle("계절성 그래프 : 확진자 수")
ggseasonplot(tsDea1, year.labels = TRUE, year.labels.left = TRUE) + ylab("사망자 수") + xlab("월") + ggtitle("계절성 그래프 : 사망자 수")
```

또한, 이에 대한 결과는 아래 사진과 동일하다.




**2020 년 ~ 2022 년 계절성 그래프 : 확진자 수에 따라**

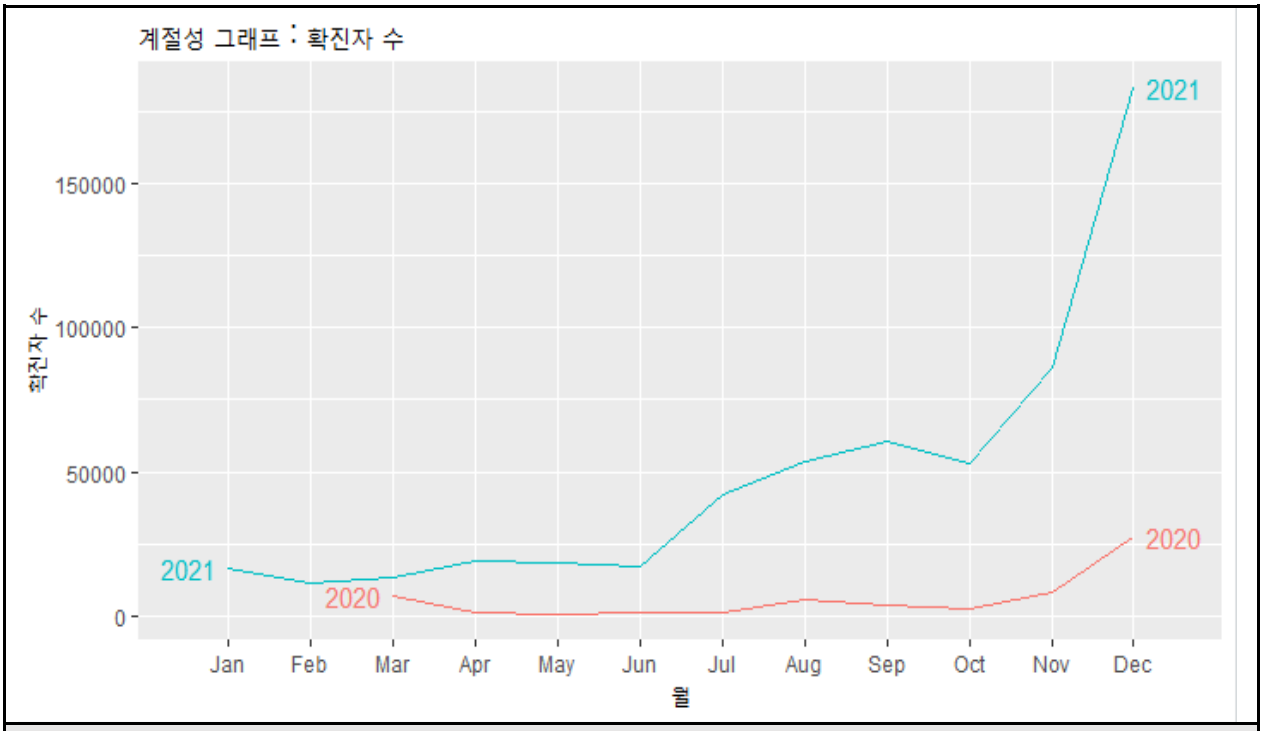



**2020 년 ~ 2022 년 계절성 그래프 : 사망자 수에 따라**

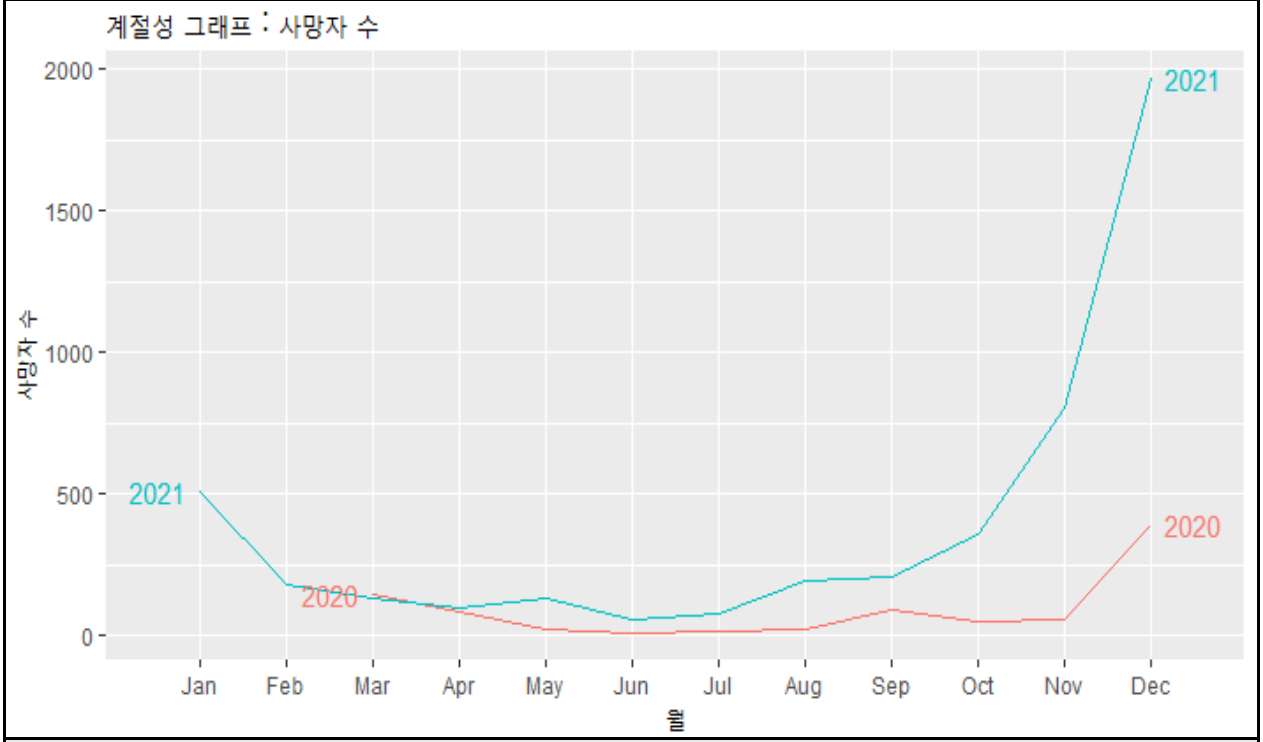
여기서 2022 년의 데이터 수치가 커, 정확한 값을 확인하기 어렵다. 따라서 아래 코드를 추가하여 2020 년과 2021 년을 기준으로 명확한 이해를 돕고자 한다.

```
#2020 년, 2021 년 그래프 시각화
zooCon2 <- kormonth1$Confirmed
zooDea2 <- kormonth1$Deaths
kormonth
tsCon2 <- as.ts(zooCon2) %>% ts(start = c(2020,3), frequency = 12)
tsDea2 <- as.ts(zooDea2) %>% ts(start = c(2020,3), frequency = 12)

# install.packages("fpp2")
library(fpp2)
plot(tsCon1)
ggseasonplot(tsCon2, year.labels = TRUE, year.labels.left = TRUE) + ylab("확진자 수") +
xlab("월") + ggtitle("계절성 그래프 : 확진자 수")
ggseasonplot(tsDea2, year.labels = TRUE, year.labels.left = TRUE) + ylab("사망자 수") +
xlab("월") + ggtitle("계절성 그래프 : 사망자 수")
```

해당 코드의 결과는 아래 사진과 일치한다.




**2020 년 ~ 2021 년 계절성 그래프 : 확진자 수에 따라**




**2020 년 ~ 2021 년 계절성 그래프 : 사망자 수에 따라**

위의 결과를 보면, 2020 년과 2021 년 모두 10 월부터 2 월에 해당하는 겨울에 확진자 및 사망자 수가 증가한다. 즉 한국의 코로나 19 환자는 겨울이라는 계절적 요인의 영향을 받고 있다고 할 수 있다.

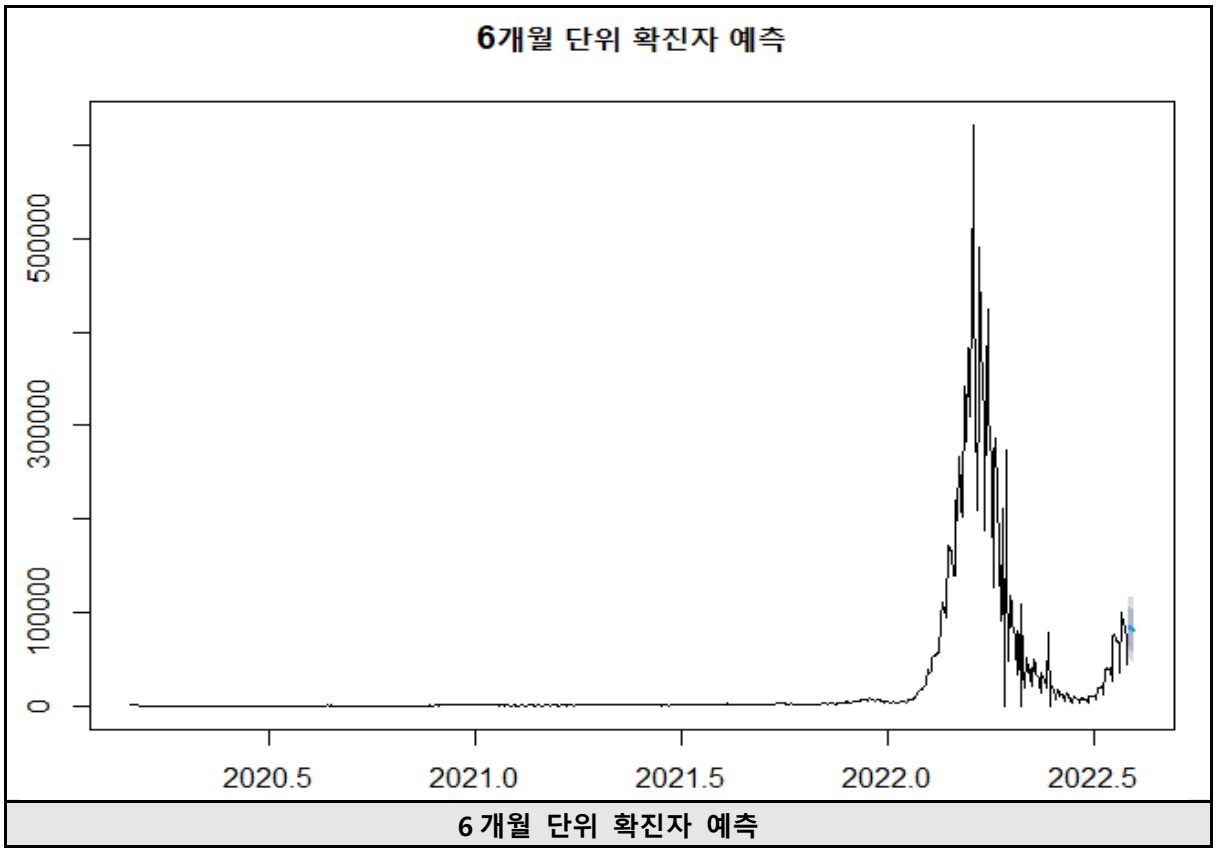
## 3. 분석 결과 및 결론

### 1) 데이터 분석 결과 및 결론 도출

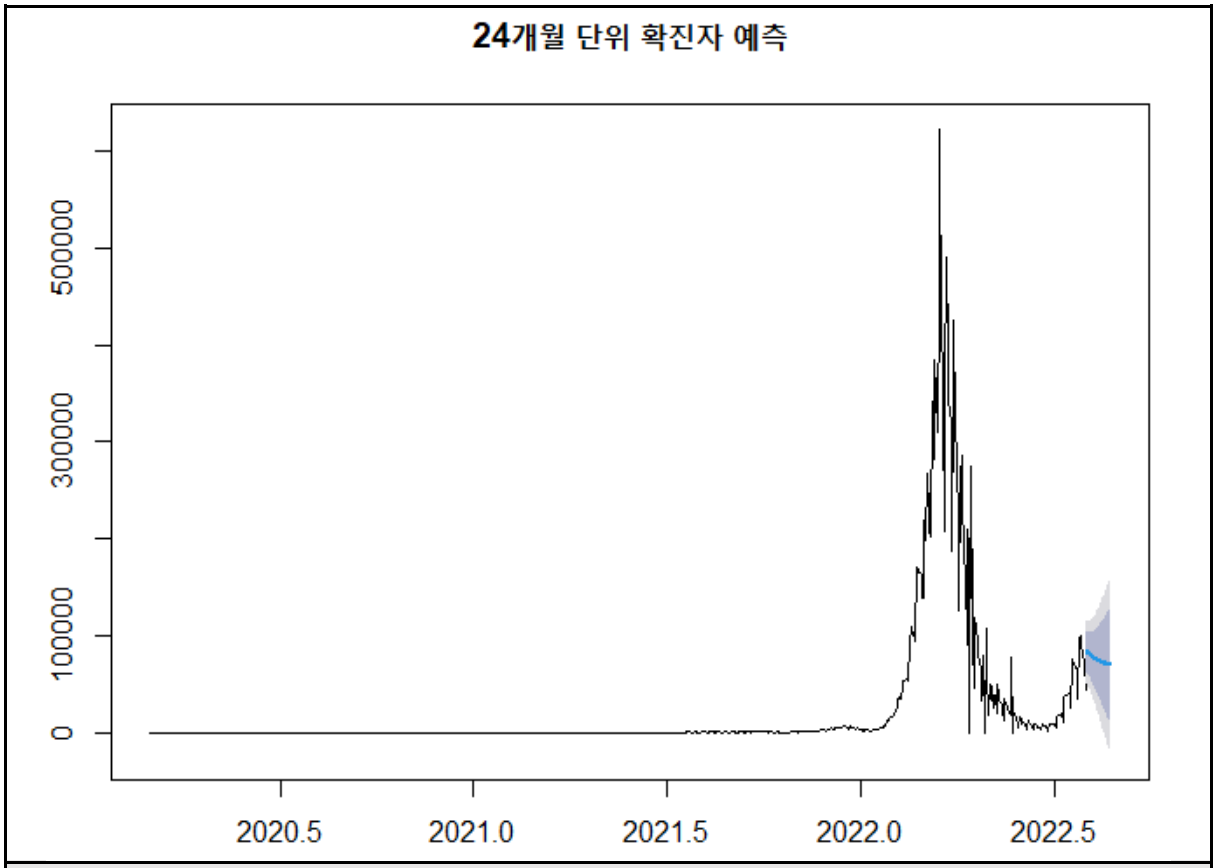
한국의 코로나 19 환자에 대한 추세는 점진적인 증가로 볼 수 있다. 특히, 계절적 요인을 받으며 겨울이 여름보다 확진자 및 사망자가 더욱 발생하는 것을 알 수 있다. 이를 통해 앞으로도 지속적으로 한국의 코로나 19 확진자 및 사망자는 계절적 요인을 받으며 증가할 것이라고 볼 수 있다. 이에 보다 객관적인 지표를 근거로 활용하고자 한다. 따라서 코드를 통해 향후 6 개월과 24 개월(2 년) 후의 확진자 및 사망자에 대한 예측을 실시했다. 코드는 아래와 동일하다.

```
fore <- forecast(tsCon, h = 24) #24 개월(2 년) 예측
plot(fore, main="24 개월 단위 확진자 예측")
fore2 <- forecast(tsCon, h = 6) #6 개월 예측
plot(fore2, main="6 개월 단위 확진자 예측")
fore3 <- forecast(tsDea, h = 24) #24 개월(2 년) 예측
plot(fore3, main="24 개월 단위 사망자 예측")
fore4 <- forecast(tsDea, h = 6) #6 개월 예측
plot(fore4, main="6 개월 단위 사망자 예측")
```

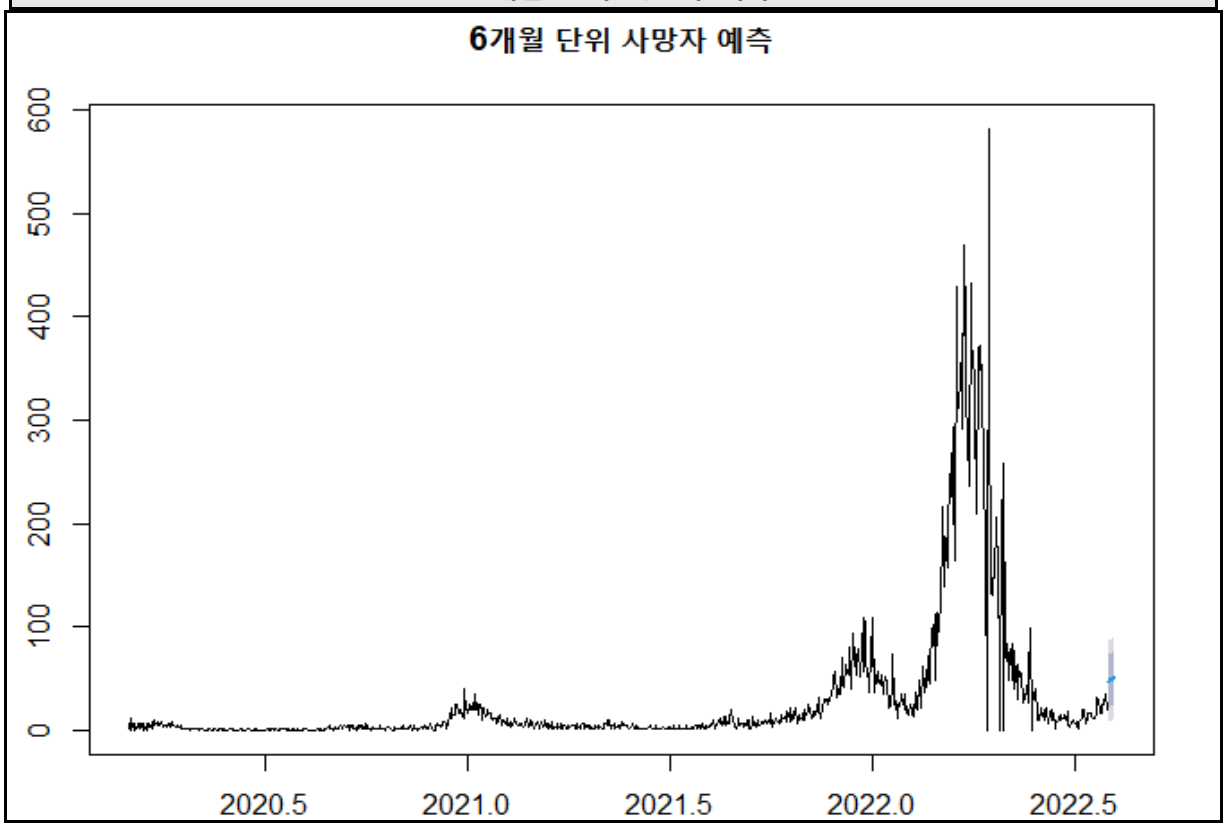
이를 통한 최종 결과는 아래 사진과 동일하다.



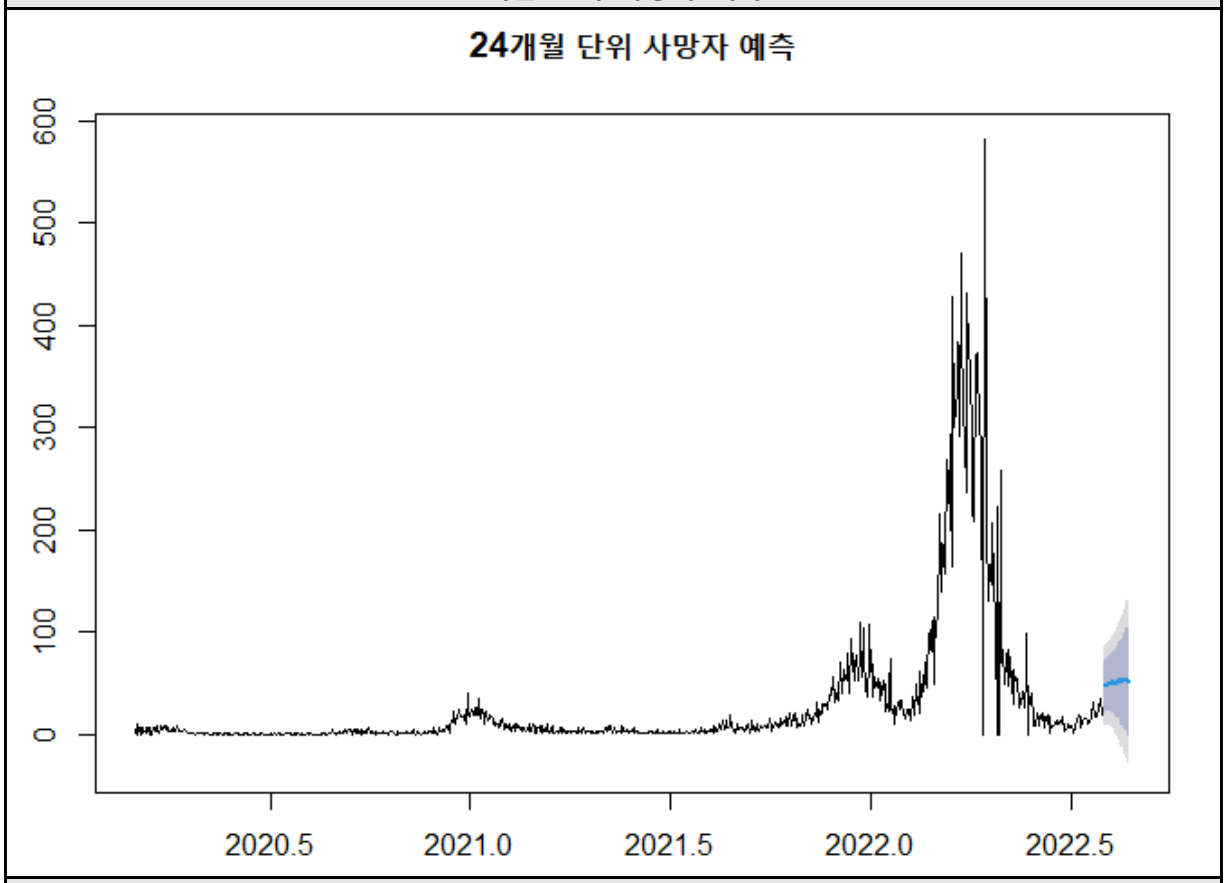

6 개월 단위 확진자 예측




24 개월 단위 확진자 예측



**6 개월 단위 사망자 예측**




**24 개월 단위 사망자 예측**

이러한 그래프를 통해 알 수 있듯, 향후 6 개월과 24 개월 동안의 한국의 코로나 19 확진자 및 사망자의 수는 계절적 요인 등의 영향을 받아 점진적으로 증가할 것으로 예측된다.

## 참고 자료


1) **데이터셋: Johns' Hopkis 대학 내 The Center For Systems Science and Engineering(CSSE) :**

https://github.com/CSSEGISandData/COVID-19/tree/770dafdb73e9dc31140db77b13b1b92cfd8241f9/csse_covid_19_data/csse_covid_19_daily_reports