서울시 미세먼지량, 초미세먼지량에 따른 따릉이(자전거) 사용량 분석

AI로봇융합전공
20206617 박주희

프로젝트 주제 선정이유 및 목적

꾸준히 따릉이(자전거), 전동자전거, 전동 킥보드 등의 대중교통 수단이 학교나 거리에서 많이 보급되면서 도시 내에서의 이용량이 증가하는 것을 직접적으로 경험했다. 특히, 따릉이(자전거)를 자주 이용하고 있는데, 이에 대한 사용량이 궁금해지게 되었다. 또한, 미세먼지와 초미세먼지의 농도에 대한 관심이 높아져서 아침마다 미세먼지와 초미세먼지 수치를 주의 깊게 확인하고 있다. 이로 인해 미세먼지와 초미세먼지의 수준과 따릉이 사용량 간에 어떠한 상관관계가 있는지 파악해 보고 싶어서 이 주제를 프로젝트로 선정하게 되었다.

이 프로젝트를 통해 서울시 따릉이 사용량과 미세먼지, 초미세먼지 간의 상관관계를 확인하고자 한다. 결과적으로, 미세먼지와 초미세먼지 수준이 따릉이 사용량에 영향을 미치는지 여부를 알아보고, 미세먼지와 초미세먼지에 대한 인식과 함께 공용 자전거 이용의 중요성을 더욱 강조할 수 있을 것으로 기대한다.

1. 실제 데이터 파일 준비하기

```
getwd()
install.packages('dplyr')
library(dplyr)

#1번> csv파일 준비
#따릉이 사용량
bike <- read.csv("ddareung.csv",fileEncoding = "CP949",encoding="UTF-8",header=TRUE,na.strings = '.')
#서울시 미세먼지,초미세먼지량
microdust <- read.csv("대기오염.csv",fileEncoding = "CP949",encoding="UTF-8",header=FALSE,na.strings = '.')
```

데이터 정보

- 따릉이 대여량 데이터: 서울시 따릉이 2020년도 7~12월 자전거 대여량 정보
- 미세먼지, 초미세먼지 수준 데이터: 서울특별시기본통계에서 수집된 월별 평균 미세먼지, 초미세먼지 수준 정보

방법
- getwd()를 활용해 현재 작업 경로의 위치를 확인
- 'dplyr' 설치 후 library작성
- read.csv()함수 사용해 csv 파일을 각각 bike, microdust 변수에 저장

데이터 출처
- 서울시 따릉이 사용량 파일: https://www.kaggle.com/datasets/bong0506/seoul-public-share-bike-ddareung
- 서울시 미세먼지, 초미세먼지량 파일: 서울특별시,「서울특별시기본통계」, 2020.12, 2023.06.02, 대기오염

2. 데이터 정제 및 가공

- 서울시 따릉이 사용량 데이터

```r
#2,3> 데이터 정제 및 가공

#bike
bike<- subset(bike,select=-place_no) #place_no 제거
bike<- summarize(count = n(),group_by(bike,rental_month),rental_sum = sum(rentals))
count_target <- max(bike$count)  # count 열을 맞출 목표 값
bike$count <- ifelse(bike$count < count_target, count_target, bike$count)
# count가 목표 값보다 작으면 목표 값으로 변경
bike
```

```
> bike
# A tibble: 7 × 3
  rental_month count rental_sum
         <int> <int>      <int>
1       202007  2190    2582775
2       202008  2190    1734558
3       202009  2190    2811710
4       202010  2190    2891772
5       202011  2190    1990615
6       202012  2190    1196965
7       202101  2190     812249
>
```

방법

- Subset() 함수를 사용하여 place_no 제거
- Summarize()함수 내에서 count변수에 n()함수 사용해 월별 개수를 저장, Group_by()함수를 사용해 rental_month로 묶음, rental의 합을 저장할 'rental_sum' 변수를 생성함
- Count_target 변수에 가장 큰 count 값을 저장
- Count가 목표값보다 작으면 목표값으로 변경

=> 월별 count개수가 다 달랐기 때문에 이를 하나의 목표값으로 통일하여 정확한 평균을 구하기 위해 이 방법을 사용했다.

```r
# rentals의 평균 구하기
bike <- bike %>%mutate(rental_mean = rental_sum / count)  # rentals의 평균 계산
bike = bike %>% rename('month'= 'rental_month') #이름 재정의
head(bike,10)
```

방법

- mutate()함수를 사용해 rentals의 평균을 계산하는 'rental_mean' 변수를 추가
- bike의 'rental_month'를 'month'로 이름 재정의

=> 나중에 미세먼지량, 초미세먼지량을 저장한 데이터프레임과 month기준으로 합치기 위해 이름을 재정의했다.

```
> head(bike,10)
# A tibble: 7 × 4
   month count rental_sum rental_mean
   <int> <int>      <int>       <dbl>
1 202007  2190    2582775       1179.
2 202008  2190    1734558        792.
3 202009  2190    2811710       1284.
4 202010  2190    2891772       1320.
5 202011  2190    1990615        909.
6 202012  2190    1196965        547.
7 202101  2190     812249        371.
```

- 서울시 미세먼지량 & 초미세먼지량 데이터

```r
#microdust
microdust <- subset(microdust, select = -1) #기준1 삭제
microdust<-as_tibble(t(microdust)) #tibble형태로 변경
colnames(microdust) <- microdust[1,]
colnames(microdust)
```

방법
- subet() 함수를 사용해 '기준(1)' 열을 삭제
- microdust를 tibble형태로 변경
- microdust의 열 이름을 첫 번째 행의 값으로 설정

=> 열의 이름이 v1,v2 이렇게 저장되어 있기 때문에 첫번째 행을 이름으로 설정했다.

```r
microdust = microdust %>% rename('month'= '대기오염')
microdust = microdust[-1,]
microdust <- microdust[,c(1,6,7)]

microdust$month<- as.integer(microdust$month)
microdust$`미세먼지(㎍/㎥/년)`<- as.integer(microdust$`미세먼지(㎍/㎥/년)`) # 정수형으로 변환
microdust$`초미세먼지(㎍/㎥/년)`<- as.integer(microdust$`초미세먼지(㎍/㎥/년)`)
> microdust
# A tibble: 6 × 3
  month `미세먼지(㎍/㎥/년)` `초미세먼지(㎍/㎥/년)`
  <int>                <int>                 <int>
1 202007                  21                    13
2 202008                  23                    14
3 202009                  19                    11
4 202010                  33                    17
5 202011                  42                    24
6 202012                  42                    27
```

방법

- rename()함수 사용해 '대기오염'을 'month'로 이름 재정의

- 첫 번째 행을 제거하고, 그 후 첫 번째, 여섯 번째, 일곱 번째 열만 남긴 업데이트된 데이터프레임을 생성

- microdust의 'month', '미세먼지', '초미세먼지' 를 정수형으로 변경

```r
#데이터 합치기
bike_dust= left_join(bike,microdust,by="month")
bike_dust<- bike_dust %>% select(month, rental_mean, `미세먼지(㎍/㎥/년)`,`초미세먼지(㎍/㎥/년)` )
bike_dust
```

```
> bike_dust
# A tibble: 7 × 4
   month rental_mean `미세먼지(㎍/㎥/년)` `초미세먼지(㎍/㎥/년)`
   <int>       <dbl>                <int>                  <int>
1 202007       1179.                   21                     13
2 202008        792.                   23                     14
3 202009       1284.                   19                     11
4 202010       1320.                   33                     17
5 202011        909.                   42                     24
6 202012        547.                   42                     27
7 202101        371.                   NA                     NA
```

```r
#boxplot으로 이상값 확인
boxplot(bike_dust[,2])  #rental_mean
boxplot(bike_dust[,3:4]) #미세먼지량, 초미세먼지량
```

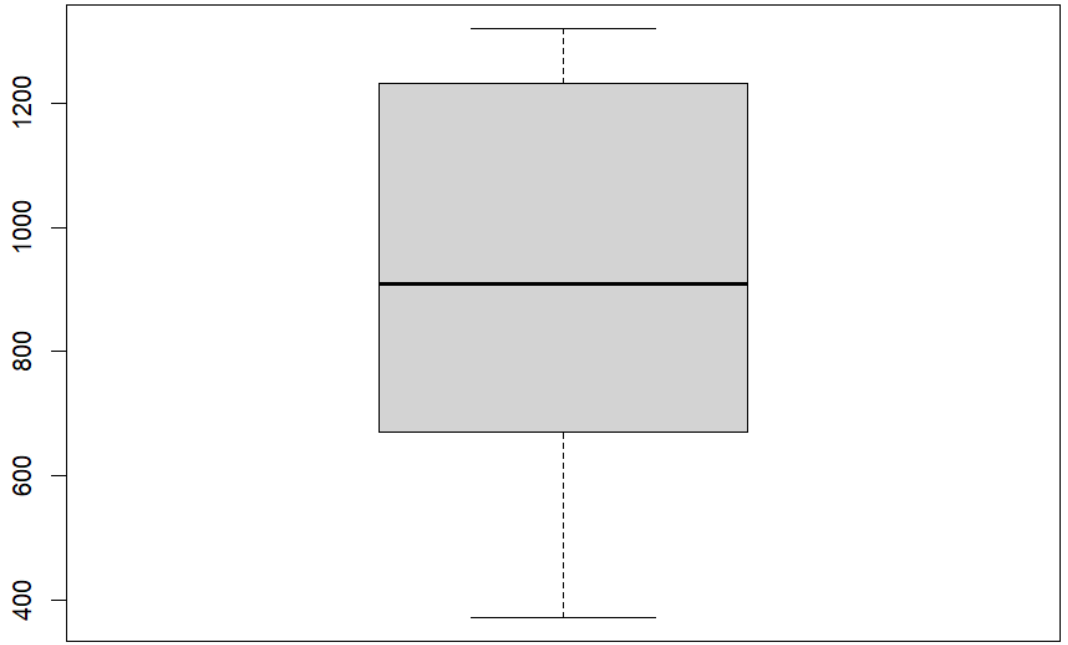
방법

- left_join()함수를 사용해 bike, microdust의 데이터 프레임을 'month' 기준으로 데이터 합쳐 bike_dust변수에 저장
- bike_dust에서 select()를 사용해 월, 미세먼지, 초미세먼지 열만 선택해 저장
- boxplot()을 사용해 bike_dust의 이상값을 확인

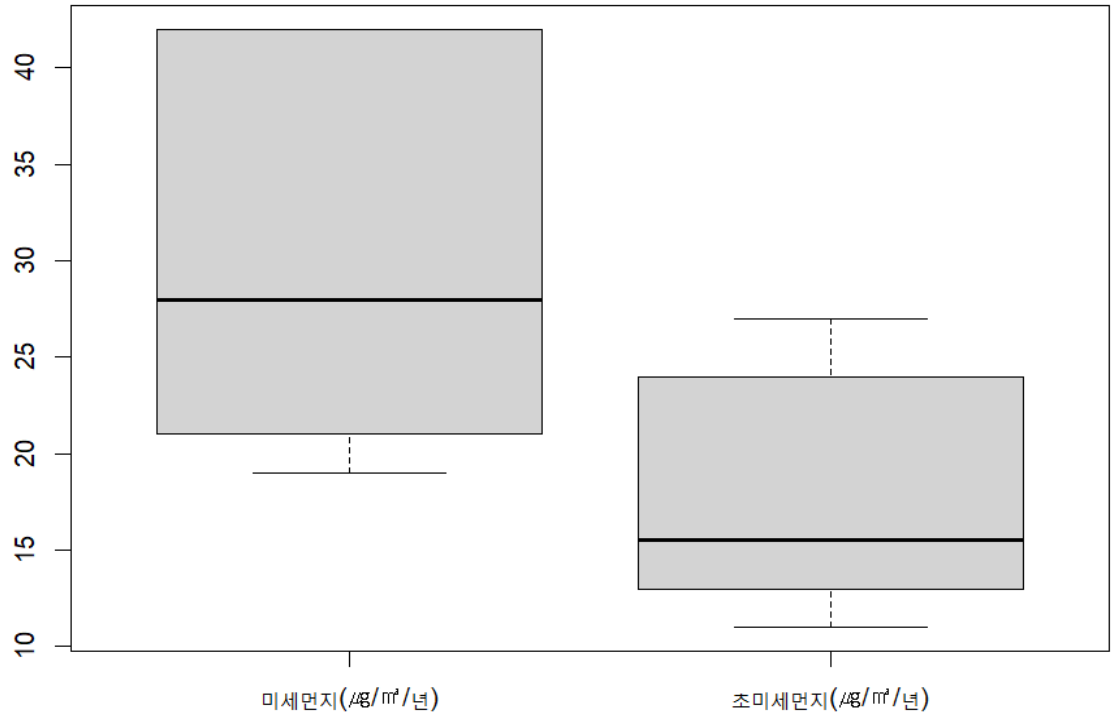
=> 하나의 데이터 프레임으로 만들어 데이터 분석을 실시하기 위해 데이터를 가공했다. 정상값과 이상값을 구분하기 위해 boxplot을 사용했다.

- boxplot()으로 정상값, 이상값 구분하기

평균 따릉이 대여량



미세먼지량, 초미세먼지량



```r
#summary() 통계량확인
summary(bike_dust)
bike_iqr <- IQR(bike_dust$rental_mean) #rental_mean의 IQR 구하기
bike_dust$rental_mean <- ifelse(bike_dust$rental_mean > summary(bike_dust$rental_mean)[5] + bike_dust_iqr*1.5, NA,
bike_dust$rental_mean)

dust1_iqr <- IQR(bike_dust$`미세먼지(㎍/㎥/년)`, na.rm = TRUE) #미세먼지의 IQR 구하기
bike_dust$`미세먼지(㎍/㎥/년)` <- ifelse(bike_dust$`미세먼지(㎍/㎥/년)` > summary(bike_dust$`미세먼지(㎍/㎥/년)`)[5] +
dust1_iqr*1.5, NA, bike_dust$`미세먼지(㎍/㎥/년)`)

dust2_iqr <- IQR(bike_dust$`초미세먼지(㎍/㎥/년)`, na.rm = TRUE) #초미세먼지의 IQR 구하기
bike_dust$`초미세먼지(㎍/㎥/년)` <- ifelse(bike_dust$`초미세먼지(㎍/㎥/년)` > summary(bike_dust$`초미세먼지(㎍/㎥/년)`)[5] +
dust2_iqr*1.5, NA, bike_dust$`초미세먼지(㎍/㎥/년)`)
```

방법

- summary() 함수를 이용해 통계적 요약 정보를 확인
- 평균 자전거 사용량, 미세먼지량, 초미세먼지량의 IQR을 구함
- (3사분면 + IQR * 1.5)보다 크면 NA(결측값)으로 변경하는 식 작성

=> 요약정보를 확인하고 최대값보다 큰 값을 결측값(NA)으로 변경하기 위해 IQR구하고 if 조건문 사용해 해당하는 데이터를 정제했다.

```
> #na 확인 및 처리
> table(is.na(bike_dust$rental_mean))#NA확인

FALSE
    7
> table(is.na(bike_dust$`미세먼지(㎍/㎥/년)`))#NA확인

FALSE  TRUE
    6     1
> table(is.na(bike_dust$`초미세먼지(㎍/㎥/년)`))#NA확인

FALSE  TRUE
    6     1
> bike_dust<-na.omit(bike_dust) #결측값제거
> bike_dust
# A tibble: 6 × 4
   month rental_mean `미세먼지(㎍/㎥/년)` `초미세먼지(㎍/㎥/년)`
   <int>       <dbl>                <int>                  <int>
1 202007       1179.                   21                     13
2 202008        792.                   23                     14
3 202009       1284.                   19                     11
4 202010       1320.                   33                     17
5 202011        909.                   42                     24
6 202012        547.                   42                     27
```

방법

- table()은 데이터의 빈도를 보여줌, is.na()는 NA의 개수를 알려줌 => NA가 몇 개 있는지 확인
- na.omit() 함수를 사용해 결측값을 제거

=> 2020년도 12월 이후의 미세먼지와 초미세먼지의 데이터는 NA로 작성되어 있었기 때문에 데이터를 제거해주었다.

3. 데이터 시각화

★ 목표: 2020년도 7~12월의 서울시 미세먼지량과 초미세먼지량에 따른 평균 따릉이 사용량을 분석

평균 따릉이 대여량과 미세먼지량, 초미세먼지량의 값의 크기 차이가 많이 나서 시각화 하는데 어려움이 있었다. 3가지 방법을 사용해 시각화를 해보았다.

① 따릉이 평균대여량과 미세먼지량 & 따릉이 평균대여량과 초미세먼지량의 막대, 선 그래프를 통한 시각화: 월별 따릉이 대여량은 선그래프로 나타내고 미세먼지와 초미세먼지는 막대그래프로 시각화하여 비교한다. 대여량에 따른 미세먼지,초미세먼지 수준의 변화를 확인할 수 있다.

② 따릉이 평균대여량, 미세먼지량, 초미세먼지량의 막대그래프 시각화: 월별 따릉이 대여량, 미세먼지량, 초미세먼지량의 관계를 막대그래프로 한번에 나타낸 시각화로 동시에 비교 할 수 있다.

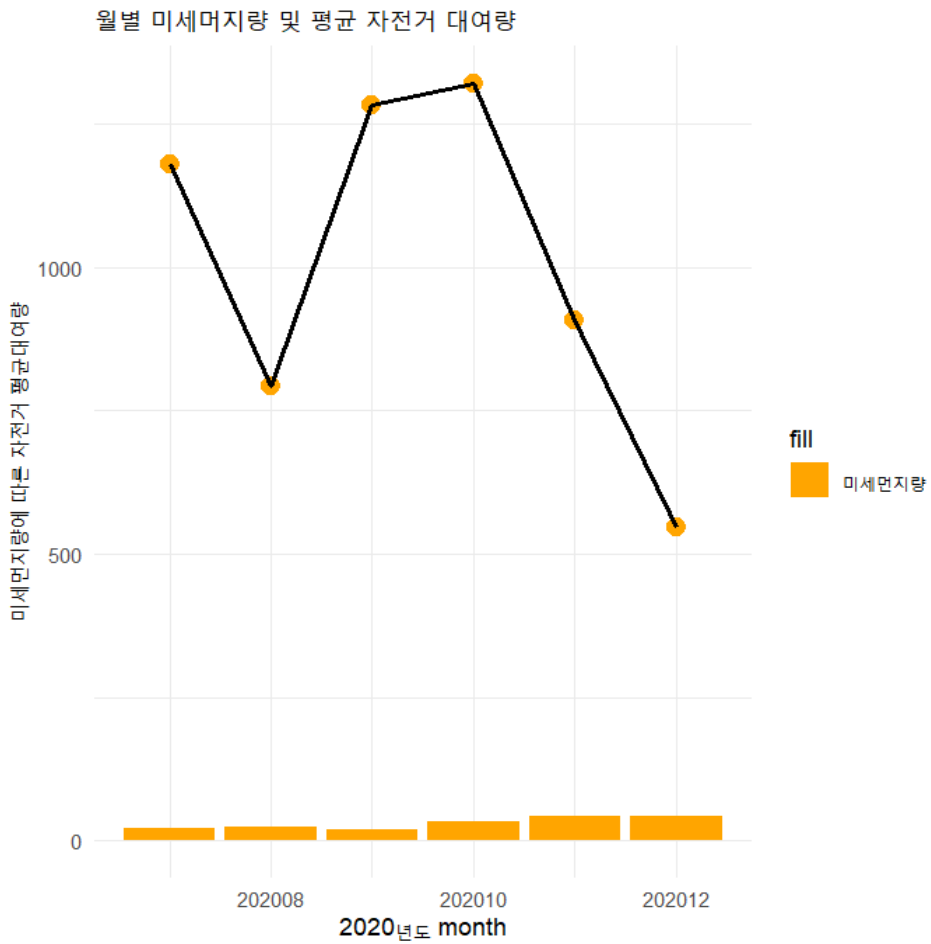
③ 월별 따릉이 평균 대여량 & 월별 미세먼지량 & 월별 초미세먼지량의 선그래프를 통한 시각화: 평균 대여량, 미세먼지량, 초미세먼지량을 각각 따로 시각화를 실시해 값의 크기 차이가 나지 않지만 하나씩 따로 비교해야한다.

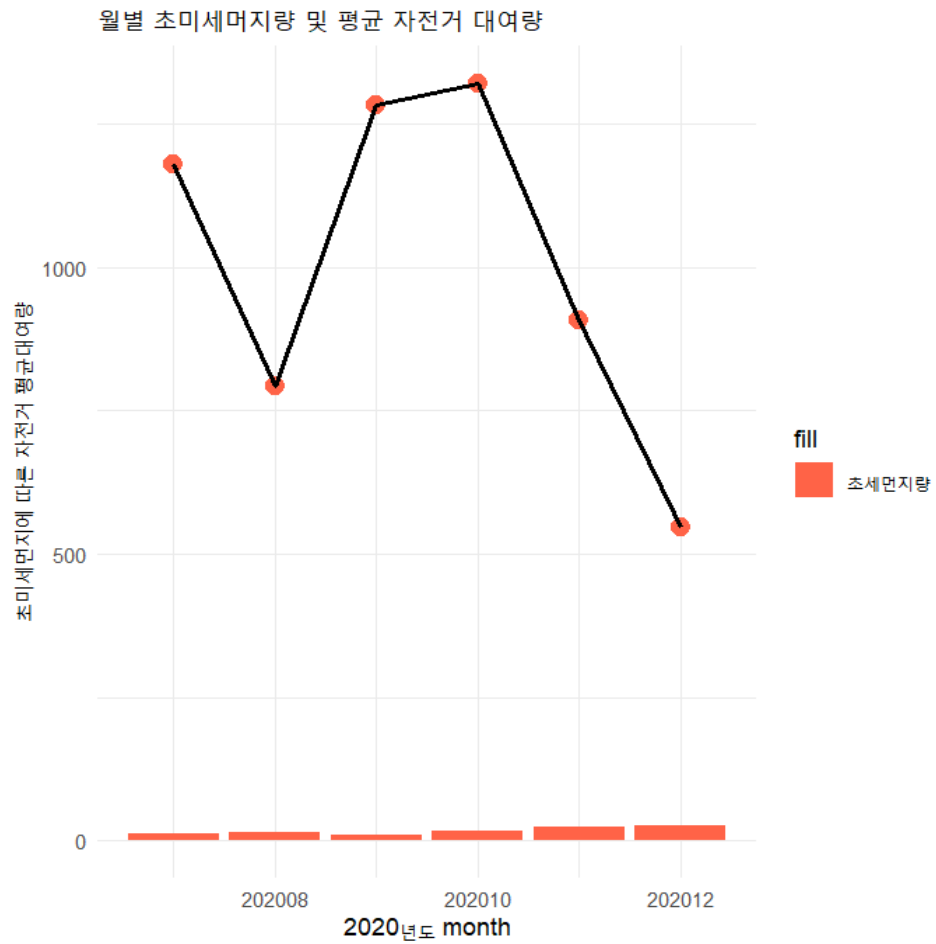
① 따릉이 평균대여량과 미세먼지량 & 따릉이 평균대여량과 초미세먼지량의 막대, 선 그래프

```r
#4번> 데이터시각화하기
install.packages("ggplot2")
library(ggplot2)

dev.off() #plots 팔레트 리셋해주기기

# bar, line그래프
p1<-bike_dust %>% ggplot(aes(month,rental_mean)) +
  geom_col(aes(y = `미세먼지(㎍/㎥/년)`, fill = "미세먼지량"))+  #막대그래프요약표
  geom_point(color = "orange", size = 3.5)+geom_line(size = 1.0)+
  labs(x = "2020년도 month", y = "미세먼지량에 따른 자전거 평균대여량") +
  ggtitle("월별 미세머지량 및 평균 자전거 대여량") +
  scale_fill_manual(values = c("미세먼지량" = "orange"))+theme_minimal()

p2 <-bike_dust %>% ggplot(aes(month,rental_mean)) +
  geom_col(aes(y = `초미세먼지(㎍/㎥/년)`, fill = "초세먼지량"))+
  geom_point(color = "tomato1", size = 3.5)+geom_line(size = 1.0)+
  labs(x = "2020년도 month", y = "초미세먼지에 따른 자전거 평균대여량")+
  ggtitle("월별 초미세머지량 및 평균 자전거 대여량") +
  scale_fill_manual(values = c("초세먼지량" = "tomato1"))+theme_minimal()

p1+p2
```

방법

1. bike_dust을 사용하여 ggplot 객체 p1,p2을 생성. x축으로 month 열을, y축으로 rental_mean 열을 사용하도록 설정
2. geom_col 함수를 사용하여 막대 그래프를 생성. y는 '미세먼지' or '초미세먼지' 열을 사용하고, fill은 '미세먼지량' or '초미세먼지'을 색상으로 채워 표현
3. Geom_point() 함수를 사용하여 'orange' / 'tomato1'색상으로 크기가 3.5인 점을 추가
4. geom_line() 함수를 사용하여 선을 추가
5. x축 레이블을 '2020년도 month'로, y축 레이블을 '미세먼지량' or '초미세먼지량'에 따른 자전거 평균대여량'으로 설정
6. 그래프의 제목을 '월별 미세머지량 / 초미세먼지 및 평균 자전거 대여량 ' 으로 설정
7. '미세먼지량 / 초미세먼지'에 대한 색상을 설정
8. 그래프의 테마를 ' minimal ' 로 설정





⇒ 미세먼지와 초미세먼지가 가장 낮은 수준일 때(2020.09) 따릉이 평균 대여량이 증가한 것을 볼 수 있다.

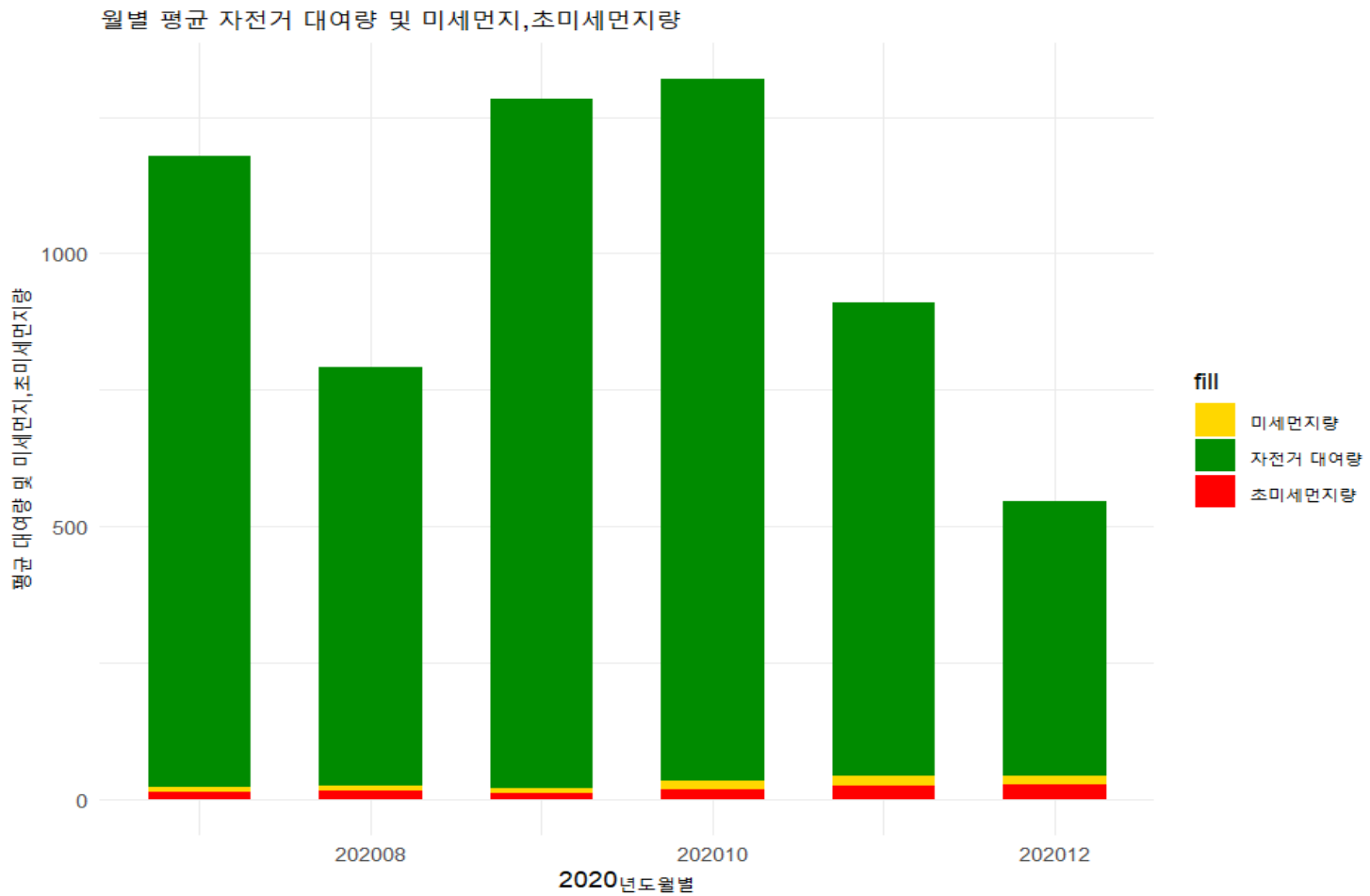
⇒ 미세먼지와 초미세먼지가 증가하는 시기인 10월부터 12월까지 대여량이 점점 줄어드는 것을 볼 수 있다

② 따릉이 평균대여량, 미세먼지량, 초미세먼지량의 막대그래프 시각화

```r
#자전거 사용량, 미세먼지, 초미세먼지 막대그래프시각화
ggplot(bike_dust, aes(x = month)) +
  geom_col(aes(y = rental_mean, fill = "자전거 대여량"), width = 0.6) +
  geom_col(aes(y = `미세먼지(㎍/㎥/년)`, fill = "미세먼지량"), width = 0.6) +
  geom_col(aes(y = `초미세먼지(㎍/㎥/년)`, fill = "초미세먼지량"), width = 0.6) +
  labs(x = "2020년도월별", y = "평균 대여량 및 미세먼지,초미세먼지량") +
  scale_fill_manual(values = c("자전거 대여량" = "green4", "미세먼지량" = "gold",
"초미세먼지량" = "red")) +
  ggtitle("월별 평균 자전거 대여량 및 미세먼지,초미세먼지량") +
  theme_minimal()
```

방법

1. `bike_dust` 데이터프레임을 사용하여 `ggplot` 객체를 생성. x축으로 'month' 열을 사용하도록 설정
2. `geom_col` 함수를 사용하여 막대 그래프를 생성하고 `y`는 'rental_mean' , '자전거 대여량'으로 색상을 채움. 막대의 너비를 0.6으로 설정
3. `geom_col` 함수를 사용하여 막대 그래프를 생성하고 `y`는 ' 미세먼지(㎍/㎥/년)', ' 미세먼지량 ' 으로 색상을 채움. 막대의 너비를 0.6으로 설정
4. `geom_col` 함수를 사용하여 막대 그래프를 생성하고 `y`는 ' 초미세먼지(㎍/㎥/년)', ' 초미세먼지량 ' 으로 색상을 채움.. 막대의 너비를 0.6으로 설정
5. x축 레이블을 '2020년도월별'로, y축 레이블을 '평균 대여량 및 미세먼지,초미세먼지량'으로 설정
6. 그래프의 제목을 '월별 평균 자전거 대여량 및 미세먼지,초미세먼지량'으로 설정
7. 그래프의 테마를 'minimal'로 설정



⇒ 미세먼지와 초미세먼지량이 가장 높을 때 자전거 대여량이 줄어들었고, 가장 낮을 때 자전거 대여량이 늘어났다.

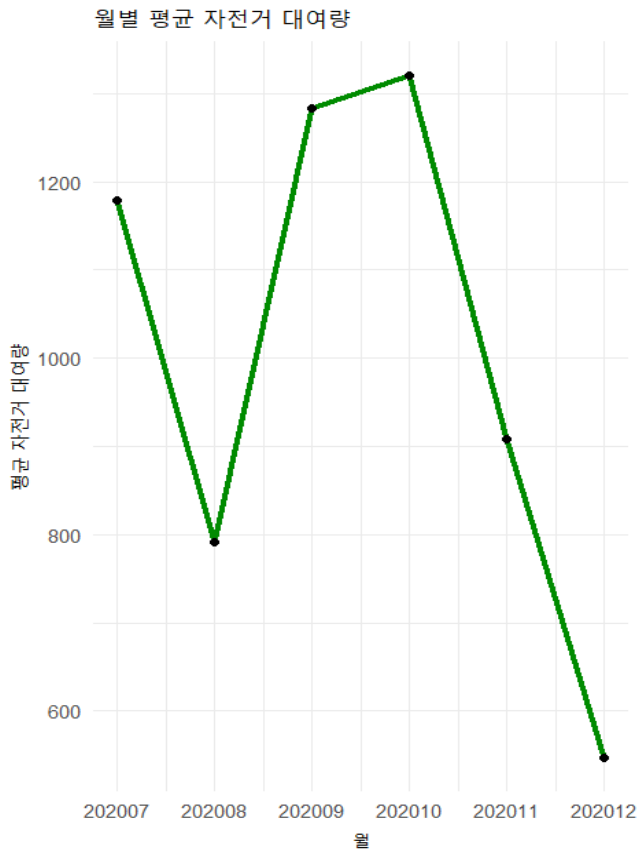
⇒ 자전거 대여량과 미세먼지량, 초미세먼지량의 값의 크기 차이가 많이 나기 때문에 시각화한 결과를 해석하는데 어려움이 있지만 한번에 값들을 비교할 수 있었다.

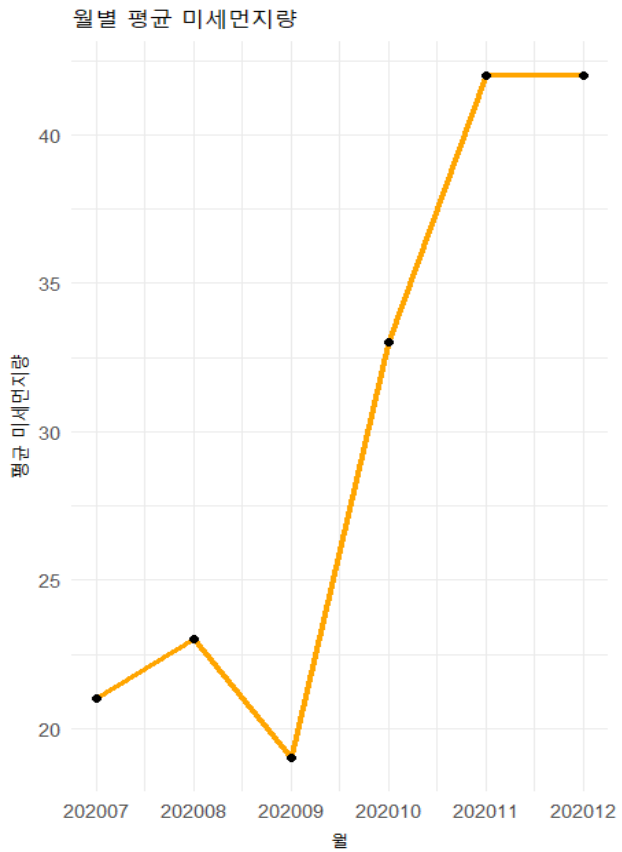
③ 월별 따릉이 평균 대여량 & 월별 미세먼지량 & 월별 초미세먼지량 line그래프 시각화

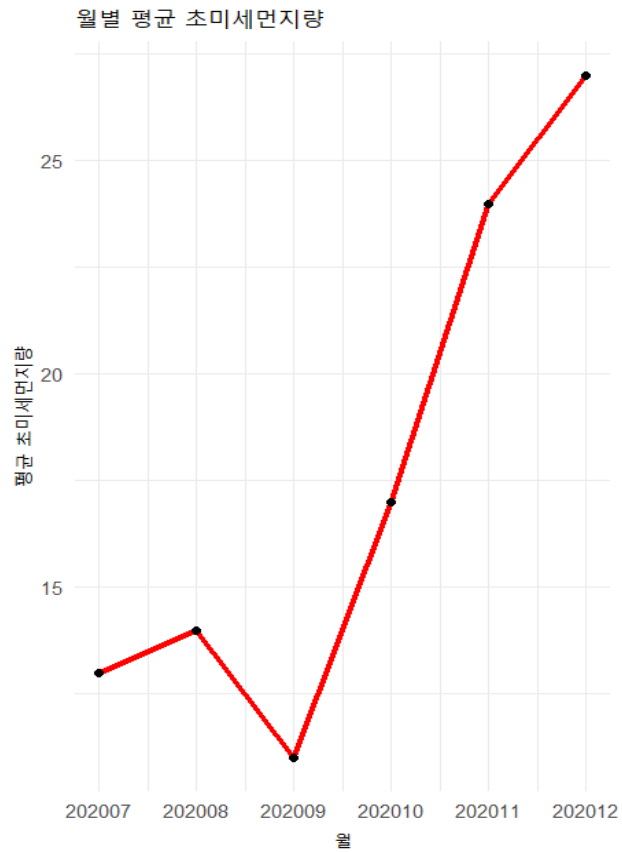
```r
# 월별 평균 자전거 대여량 시각화
plot1 <- ggplot(bike_dust, aes(x = month, y = rental_mean)) +
  geom_line(color = "green4", size = 1.2) +
  labs(x = "월", y = "평균 자전거 대여량") +
  ggtitle("월별 평균 자전거 대여량") + geom_point()+
  theme_minimal()

# 월별 미세먼지량 시각화
plot2 <- ggplot(bike_dust, aes(x = month, y = `미세먼지(㎍/㎥/년)`)) +
  geom_line(color = "orange", size = 1.2) +
  labs(x = "월", y = "평균 미세먼지량") +
  ggtitle("월별 평균 미세먼지량") + geom_point()+
  theme_minimal()

# 월별 초미세먼지량 시각화
plot3 <- ggplot(bike_dust, aes(x = month, y = `초미세먼지(㎍/㎥/년)`)) +
  geom_line(color = "red", size = 1.2) +
  labs(x = "월", y = "평균 초미세먼지량") +
  ggtitle("월별 평균 초미세먼지량") + geom_point()+
  theme_minimal()

plot1 + plot2 + plot3
```

방법

1. bike_dust 데이터프레임을 사용하여 ggplot 객체 plot1, plot2, plot3을 생성
2. x축으로 ' month ' 열을, y축으로 rental_mean, '미세먼지', '초미세먼지 ' 열을 사용하도록 설정
3. geom_line() 함수를 사용하여 선을 추가하고 선의 색상과 크기는 1.2로 설정
4. x축 레이블을 "월"로, y축 레이블을 "평균 자전거 대여량, 평균 미세먼지량, 평균 초미세먼지량"으로 설정
5. 그래프의 제목을 "월별 평균 자전거 대여량, 미세먼지량, 초미세먼지량"으로 설정
6. geom_point 함수를 사용하여 점을 추가
7. 그래프의 테마를 ' minimal ' 로 설정
8. plot1, plot2, plot3을 한번에 나타냄







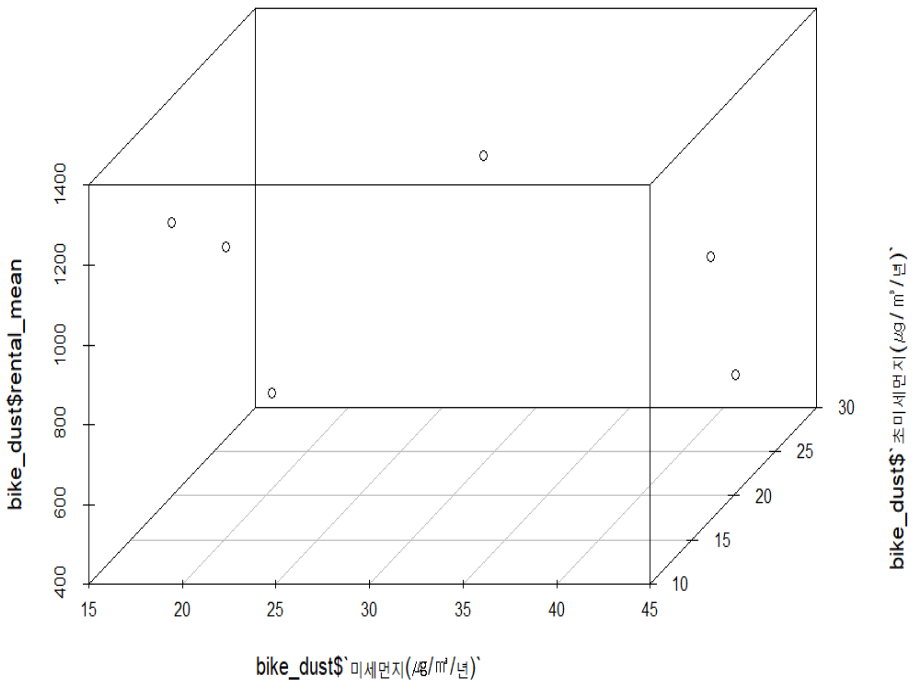
⇒ 월별로 미세먼지와 초미세먼지는 유사한 패턴을 보이고 있는 것을 확인할 수 있다. 그러나 미세먼지는 11월과 12월에 유사한 수준을 유지하는 반면, 초미세먼지는 증가하는 추세를 보인다.

⇒ 자전거 대여량은 8월부터 증가하다가 10월부터 감소하고 있는 것을 확인할 수 있다.

⇒ 각각 그래프를 통해 값의 변화와 증감 추세를 잘 파악할 수 있었으나 서로 비교하기에는 어려움이 있었다.

+3차원으로 시각화하기

```
#3차원 시각화
install.packages("scatterplot3d")
library(scatterplot3d)

sp3=scatterplot3d(bike_dust$`미세먼지(㎍/㎥/년)`, bike_dust$`초미세먼지(㎍/㎥/년)`
,bike_dust$rental_mean, angle=50)
```



방법

- scatterplot3d 패키지를 사용하여 3차원 산점도를 생성
- scatterplot3d(): 3차원 산점도를 생성하는 함수
- bike_dust$rental_mean: z축에 해당하는 따릉이 대여량 데이터 (반응변수)
- bike_dust$미세먼지(㎍/㎥/년): x축에 해당하는 미세먼지 데이터
- bike_dust$초미세먼지(㎍/㎥/년): y축에 해당하는 초미세먼지 데이터
- angle = 50: 3차원 그래프의 각도를 설정하는 매개변수로 50도로 설정

⇒ rental_mean을 반응변수로 결정했다.

⇒ 따릉이 대여량, 미세먼지량, 초미세먼지량을 산점도로 확인할 수 있다.

4. 가설 검정 단계 & 예측

가설검정1

- 귀무가설 H0: 미세먼지량이 높을 수록 월별 평균 따릉이사용량이 높다.
- 대립가설 H1: 미세먼지량이 높을 수록 월별 평균 따릉이사용량이 낮다.

가설검정2

- 귀무가설 H0:초미세먼지량이 높을 수록 월별 평균 따릉이사용량이 높다.
- 대립가설 H1:초초미세먼지량이 높을 수록 월별 평균 따릉이사용량이 낮다.

```
> #다중회귀모델생성
> bike_bust_model = lm(rental_mean~`미세먼지(㎍/㎥/년)`+`초미세먼지(㎍/㎥/년)`,data = bike_dust)
> #가설 검정
> summary(bike_bust_model)

Call:
lm(formula = rental_mean ~ `미세먼지(㎍/㎥/년)` + `초미세먼지(㎍/㎥/년)`,
    data = bike_dust)

Residuals:
      1       2       3       4       5       6
 127.11 -254.93   86.59   20.31  -18.00   38.92

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               1457.07     234.41   6.216  0.00839 **
`미세먼지(㎍/㎥/년)`        67.25      30.64   2.195  0.11577
`초미세먼지(㎍/㎥/년)`    -139.77      49.82  -2.806  0.06754
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 174.1 on 3 degrees of freedom
Multiple R-squared:  0.8076,    Adjusted R-squared:  0.6793
F-statistic:  6.296 on 2 and 3 DF,  p-value: 0.08439
```

분석

lm()함수를 이용해 최적의 모델을 찾는다.

- 반응변수: rental_mean
- 설명변수: ‘미세먼지량’ + ‘초미세먼지량’

summary()함수를 사용해 모델의 통계적 요약 정보를 확인한다.

- 미세먼지 p-값: 0.11577
- 초미세먼지 p-값: 0.06754
- R-squared 결정계수: 0.6793 => 0.6793은 1에 가깝기 때문에 설명력이 높다고 해석된다.
- F-statistic: 6.296

```
> #6번> 모델링, 예측 판별
> p_value1 <- summary(bike_bust_model)$coefficients[2, "Pr(>|t|)"]
> p_value1
[1] 0.1157651
> p_value2 <- summary(bike_bust_model)$coefficients[3, "Pr(>|t|)"]
> p_value2
[1] 0.06753973
> alpha <- 0.05  # 유의 수준 설정
```

분석

Coefficients 속성을 통해 회귀 계수에 대한 정보를 가져온다.

- P_value1변수: 3번째 행의 p-값을 선택한다. (미세먼지)
- P_value2변수: 2번째 행의 p-값을 선택한다. (초미세먼지)
- 유의 수준 'alpha'를 0.05로 설정한다.

⇒제 1종 오류를 범할 확률의 허용한계를 5%라 설정하고 분석을 진행한다.

⇒대립가설을 기각시킬 확률(p_value)을 미세먼지와 초미세먼지 각각 구한다.

```
> if (p_value1 < alpha) {
+   cat("귀무 가설(H0)을 기각합니다. 즉, 미세먼지량이 높을수록 월별 평균 따릉이 사용량이 낮을 가능성이 있습니다.")
+ } else {
+   cat("귀무 가설(H0)을 채택합니다. 즉, 미세먼지량이 높을수록 월별 평균 따릉이 사용량이 높을 가능성이 있습니다.")
+ }
귀무 가설(H0)을 채택합니다. 즉, 미세먼지량이 높을수록 월별 평균 따릉이 사용량이 높을 가능성이 있습니다.
```

```
> if (p_value2 < alpha) {
+   cat("귀무 가설(H0)을 기각합니다. 즉, 초미세먼지량이 높을수록 월별 평균 따릉이 사용량이 낮을 가능성이 있습니다.")
+ } else {
+   cat("귀무 가설(H0)을 채택합니다. 즉, 초미세먼지량이 높을수록 월별 평균 따릉이 사용량이 높을 가능성이 있습니다.")
+ }
귀무 가설(H0)을 채택합니다. 즉, 초미세먼지량이 높을수록 월별 평균 따릉이 사용량이 높을 가능성이 있습니다.
```

분석

미세먼지, 초미세먼지의 p-값과 유의수준(0.05)을 비교하여 귀무가설을 기각하거나 채택하는 if 조건문을 작성했다.

- 미세먼지량이 높아질수록 따릉이 대여량이 낮아진다는 대립가설은 기각되고 **귀무가설이 채택**된다.

- 초미세먼지량이 높아질수록 따릉이 대여량이 낮아진다는 대립가설은 기각되고 **귀무가설이 채택**된다.

- 두 가설 모두 유의 수준 0.05보다 크므로 통계적으로 유의한 차이가 없다고 판정된다.

5. 결과해석 & 느낀점

미세먼지량과 초미세먼지량이 평균 따릉이 사용량과 통계적으로 유의미한 관련성이 나타나지 않은 것으로 분석되었다. 데이터 분석 결과는 미세먼지량과 초미세먼지량이 평균 따릉이 사용량에 직접적인 영향을 미치지 않는 것을 의미한다. 이 결과를 통해 미세먼지, 초미세먼지가 따릉이 사용량에 주요한 영향을 미치기 보다는 다른 요인들이 더 큰 영향을 미치고 있다고 볼 수 있다. 예를 들면, 날씨(온도, 강수량), 교통 체증상황 등이 따릉이 사용량에 영향을 미칠 것으로 생각된다. 따라서, 미세먼지량과 초미세먼지량이 따릉이 사용량을 직접적으로 예측하는데 제한적이라는 결론을 낼 수 있다.

5. 결과해석 & 느낀점

월별로 미세먼지와 초미세먼지 수준이 좋은 달과 좋지 않을 달이 존재했을 가능성을 간과했던 것 같다. 또한, 따릉이는 외부에서 이용되기 때문에 온도와 강수량과 같은 기상 조건에 크게 영향을 많이 받을 것으로 예상했다. 그래서 일상생활에서 항상 신경 쓰는 미세먼지와 초미세먼지 수준을 활용해 따릉이 대여량과의 상관관계를 확인하고자 했다. 하지만 예상했던 것과 다른 결과가 나와서 아쉬움을 느꼈다. 현재의 분석으로 마무리하는 것이 아니라 더 많은 데이터를 수집하고 추가적인 분석을 시도하여 다시 도전하고 싶은 욕구가 생겼다. 이 프로젝트를 통해 데이터 분석의 한계와 가능성을 경험하며 더욱 성장할 수 있을 것이라고 생각한다.