미세먼지와 호흡기 질환 발생 간 영향분석

데이터베이스기반 빅데이터 분석가 양성과정

2019-11-26 ~ 2020-04-29 (800시간/100일)

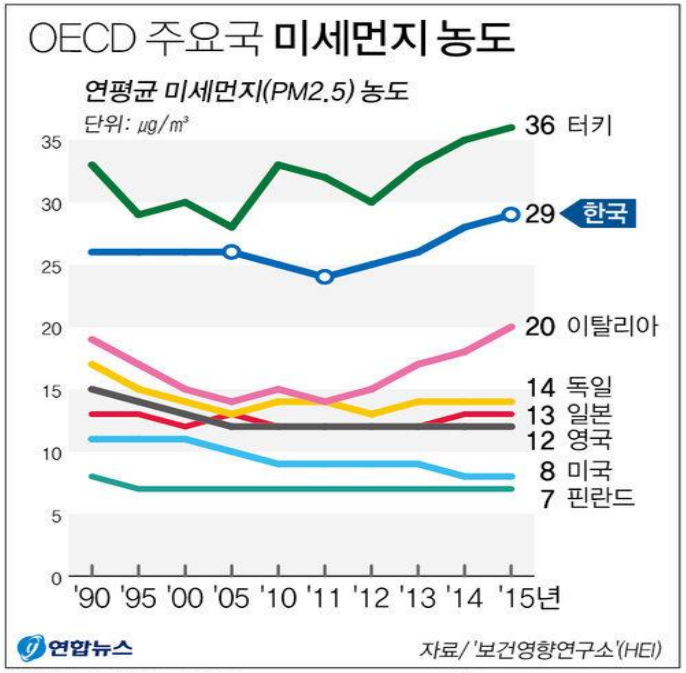
4조 – 강지* 백종* 신용* 이현* 황태*

목차
INDEX

주제 & 목적

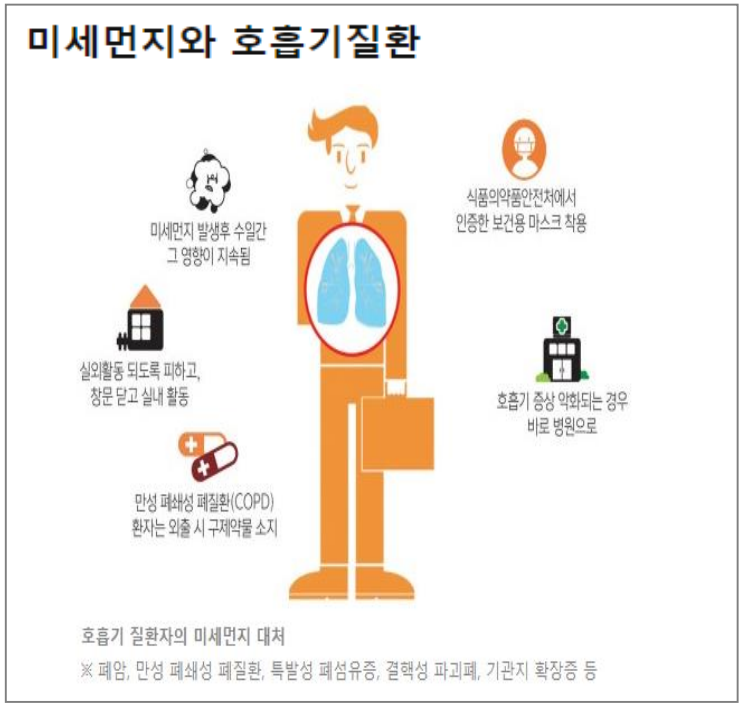
주제 및 목적

미세먼지 악화



박영석 기자 / 20170216

트위터 @yonhap_graphics, 페이스북 tuney.kr/LeYN1



2013년 이후 시작된 (초)미세먼지의 급증

호흡기 관련 질병 발생 간 연관성 논의 필요

주제 및 목적

국가정책사업으로 미세먼지 대두화

9대 국가전략 프로젝트 중 하나로 선정

2023년 까지 원인물질 배출량을 절반 수준으로 감축 및 미세먼지 해결을 위한 기술 개발을 목표하고 있음

두 요소(미세먼지, 질병)의 **시계열 데이터**를 통해 추이를 비교하여 **연관성 여부 확인**

정부의 향후 대책 및 예산 수립에 기여하고 궁극적으로 **국민 건강에 공헌**하기 위함.

주제 및 목적

호흡기 질환 발병률과 미세먼지 연관성

호흡기 원인 관련 뉴스 단어 빈도수

미세먼지 관련 뉴스 단어 빈도수

일정 계획 &
환경설정

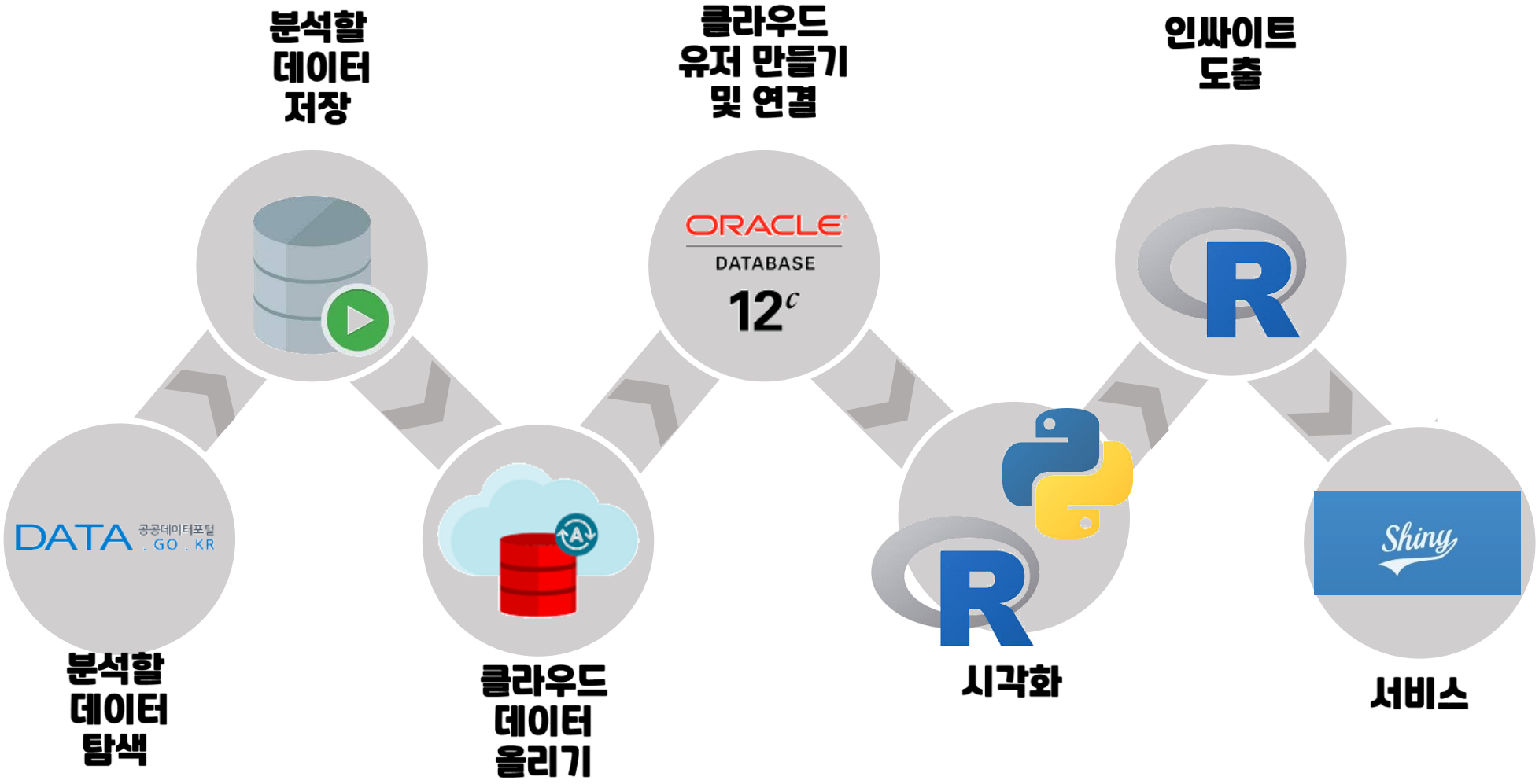
일정 계획 및 환경설정

일정 계획

활동 내용	1 주	2 주	3 주	4 주	5주	6 주	7주	8주
주제 선정/요구 사항 분석								
Dataset 수집 계획 수립								
분석환경 수립								
Data 전처리 & 시각화								
분석								
웹 서비스								
인사이트 도출								
보고서 작성								

일정 계획 및 환경설정

분석 환경설정



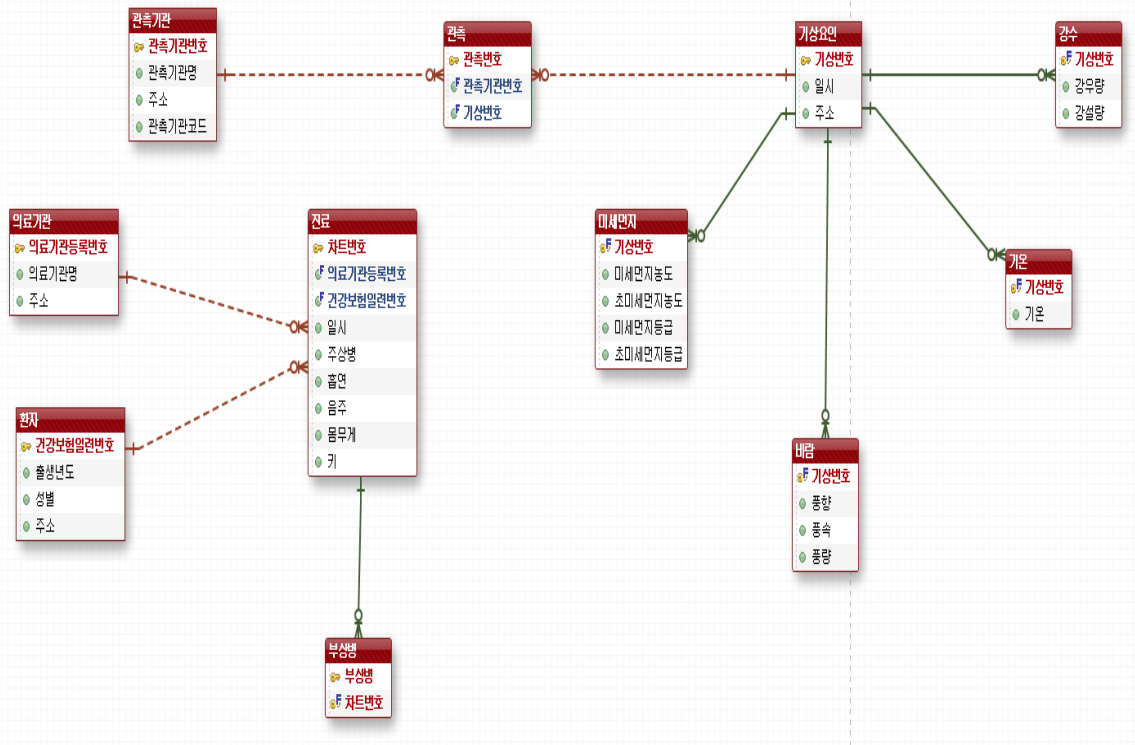
일정 계획 및 환경설정

분석 환경설정

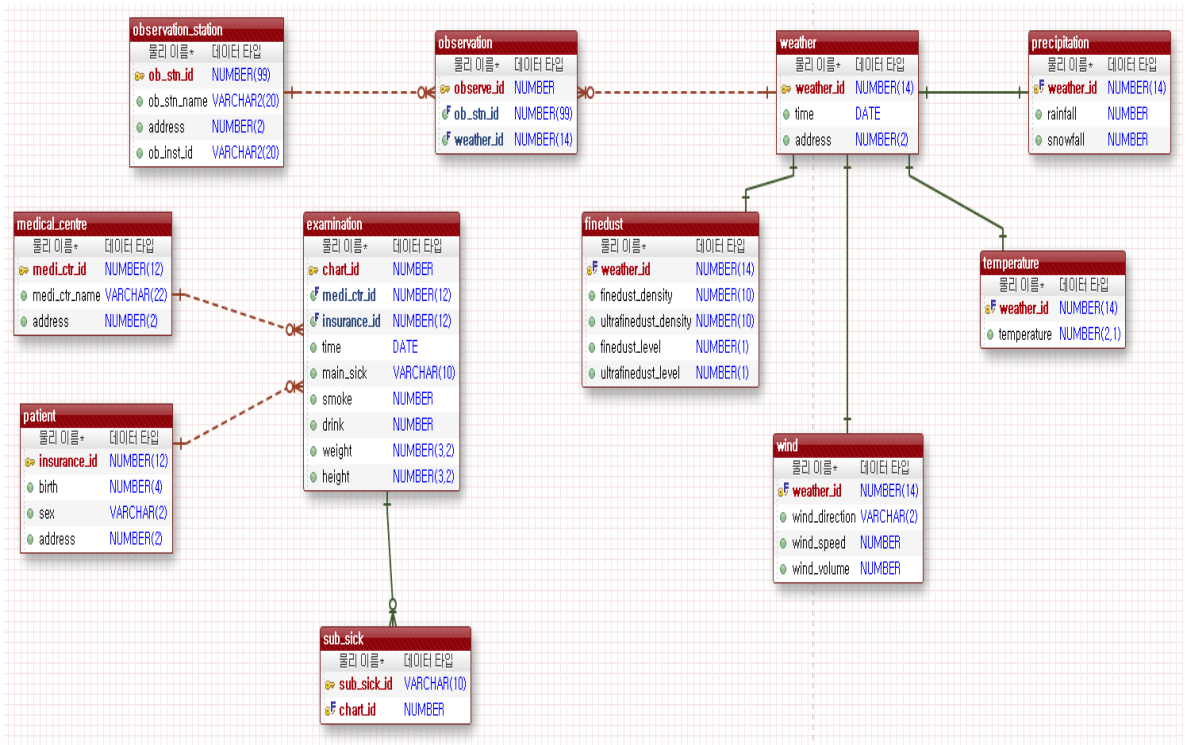
OS	DB	모듈	분석언어	형상관리	TOOL
					
					
					
					

일정 계획 및 환경설정

DB구성



논리적 ERD

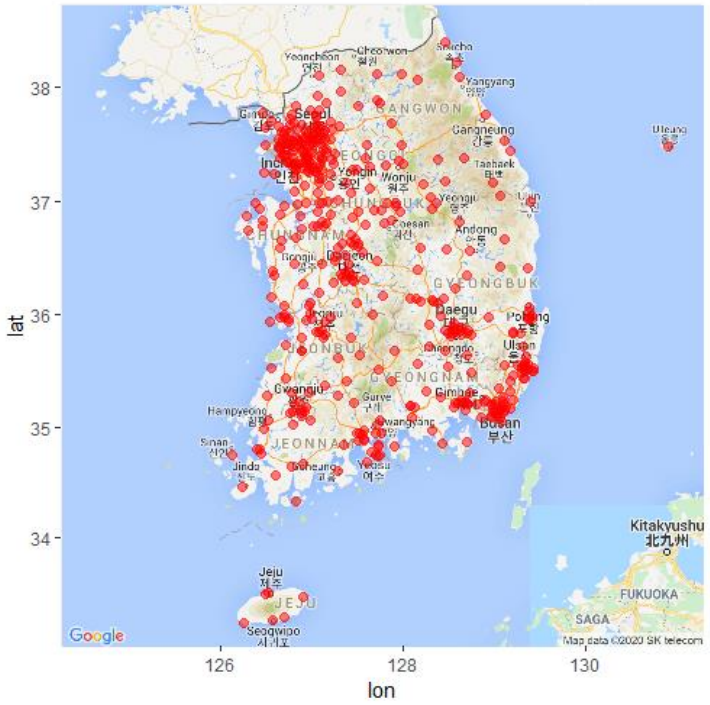


물리적 ERD

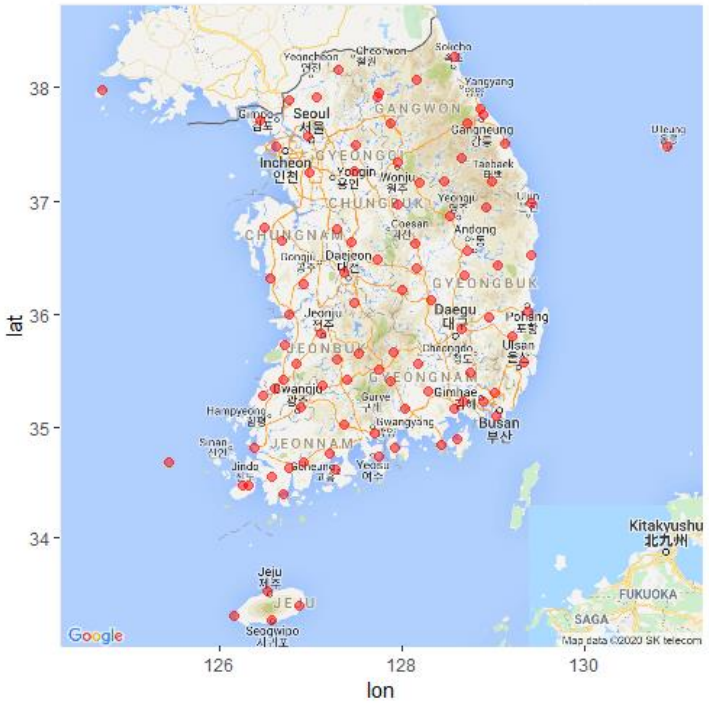
전처리

전처리

미세먼지 데이터(결측치 처리)



미세먼지 측정소 현황



기상 측정소 현황

전처리

미세먼지 데이터(결측치 처리)

- 지리적 자료 결측치 보완 기법 서치 -> 지역평균법 적용
- 전국 시 평균 면적(608.10 km2) -> 약 25 km를 영향반경으로 설정, 반경 내 측정된 데이터 사용

2.2. 공간 자료에 대한 주요 결측 보완 기법들

일반적인 결측치 예측 방법들과는 달리 공간 자료에 대한 결측치를 예측하기 위해서는 공간적 특성을 고려해야 한다. 이를 반영한 결측치 예측 기법들 중 본 연구에서 활용하는 방법들은 다음과 같다.

2.2.1. 지역평균법

지역평균법은 본 연구에서 적용한 방법은 아니다. 다만 공간 자료의 결측값을 예측하는 가장 기본적인 방법이며 뒤에서 소개하는 역거리가중치법의 특수한 사례이기에 여기서 소개한다. 지역평균법의 주요 특징은 다음과 같다.

- 결측점을 중심으로 영향반경을 정하고 반경 내의 모든 자료값들의 산술평균을 취하는 방법이다.

- 영향반경 이내에 속한 모든 점들에 대해 동일한 가중치를 부여한다. 그러기에 특이값 등의 문제가 발생할 가능성이 있다.

- 여러 공간 자료 결측치 예측 기법들 중 기본이 되는 방법으로, 방법의 우수성을 표현할 때 지역평균법 대비 어느 정도의 성능 향상이 있었는지로 표현하기도 한다.

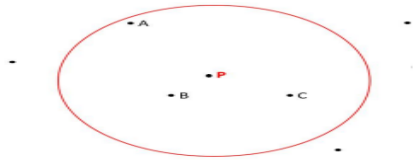


그림 2.2.1 지역평균법의 예

위의 그림 2.2.1에서 P는 결측값이다. P의 값을 계산하기 위해 적색의 영향반경 내에 있는 점 A, B, C를 고려하게 된다. 적색의 영향반경 밖에 있는 세 개의 점은 결측을 위해 사용되지 않는다.

지역 평균을 이용한 세종 결측치 처리

전처리

미세먼지 데이터(오염물질 기하평균 처리)

기하평균(geometric mean)

기하평균은 넓이, 부피, 비율 등 곱셈으로 이루어지는 값들의 평균을 구하는 데 주로 사용된다. 기하평균의 식은 다음과 같다.

$$\bar{x}_G = \sqrt{x_1 \cdot x_2 \cdot \cdots \cdot x_n}$$

기하평균 함수 적용

[36]:
```
def geo_mean(iterable):
    a = np.array(iterable)
    a = a[~np.isnan(a)]
    a = np.log(a)
    return np.exp(a.sum()/len(a))
```

[37]:
```
df = df.groupby(['시도','일시']).agg(geo_mean)
```

오염물질을 기하 평균으로 계산

전처리

의료 데이터(호흡기 질환 진료데이터 추출)

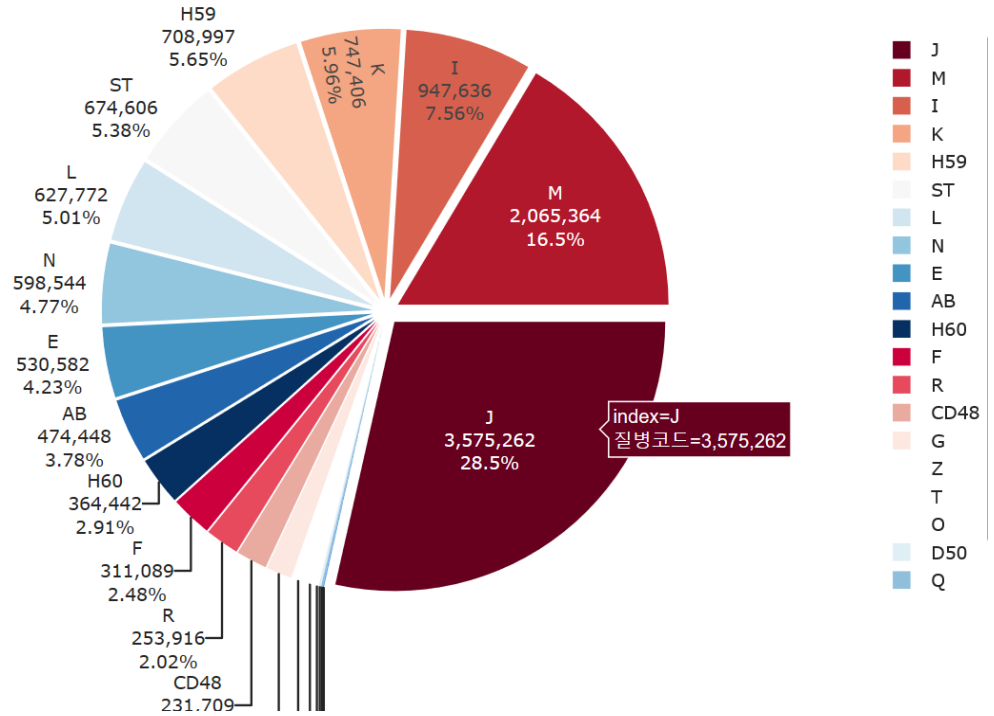
```
medi_2016 = pd.read_csv('../../lawdata/medical/NHIS_2016_Fixed.csv')
medi_2017 = pd.read_csv('../../lawdata/medical/NHIS_2017_Fixed.csv')
medi_2018 = pd.read_csv('../../lawdata/medical/NHIS_2018_Fixed.csv')

medi_2016 = medi_2016[medi_2016['질병코드']=='J']
medi_2017 = medi_2017[medi_2017['질병코드']=='J']
medi_2018 = medi_2018[medi_2018['질병코드']=='J']
```

	가입자일련번호	성별코드	연령대코드	시도코드	주상병코드	요양개시일자	시도	질병코드
0	23	1	9	45	J042	20160920	전북	J
1	35	1	13	45	J029	20161119	전북	J
2	117	1	11	45	J303	20161202	전북	J
3	117	1	11	45	J209	20161201	전북	J
4	143	2	14	45	J209	20160117	전북	J

```
# 2016~2018년도 호흡기질환 진료환자 데이터 저장

medi.to_csv('../../lawdata/medical/NHIS_J_Cases.csv', index = False)
```



호흡기 질환 관련 코드 추출

전처리

전체 데이터 셋(풍향)

	시도코드	일시	발생건수	시도	평균기온(°C)	최저기온(°C)	최고기온(°C)	평균 풍속(m/s)	평균 현지기압(hPa)	일 최심신적설(cm)	일강수량(mm)	강수 계속시간(hr)	최다풍향(16방위)
1	11	2016-01-01	217	서울	1.2	-3.3	4	1.6	1019.9	0	0	0	14
2	11	2016-01-02	2200	서울	5.7	1	9.5	2	1012	0	0	0	14
3	11	2016-01-03	267	서울	6.5	5.1	9.4	1.8	1008.9	0	0	0	14
4	11	2016-01-04	3244	서울	2	-2.5	5.3	3.1	1013.1	0	0	0	1
5	11	2016-01-05	2163	서울	-2.7	-4.8	1.5	2.3	1016.9	0	0	0	14
6	11	2016-01-06	2197	서울	-1.7	-4.9	1.7	1.8	1014.7	0	0	0	1
7	11	2016-01-07	2134	서울	-3.4	-5.9	1.4	2.5	1013.9	0	0	0	13
8	11	2016-01-08	2332	서울	-3.3	-6.9	1	2	1012.6	0	0	0	16
9	11	2016-01-09	1824	서울	-2.1	-6.2	2.4	2.1	1014.3	0	0	0	2
10	11	2016-01-10	207	서울	0.3	-2.7	3.8	2.6	1015.5	0	0	0	2
11	11	2016-01-11	3113	서울	-3.8	-6.5	0.9	2.8	1016.1	0	0	0	2
12	11	2016-01-12	2029	서울	-5.2	-9.1	0.7	2.5	1013.3	0	0	0	2
13	11	2016-01-13	1985	서울	-4.5	-8.2	0.1	2	1009.5	0.5	0.4	3.42	10

최대 빈도와 풍속을 이용한 최다풍향 평균

전처리

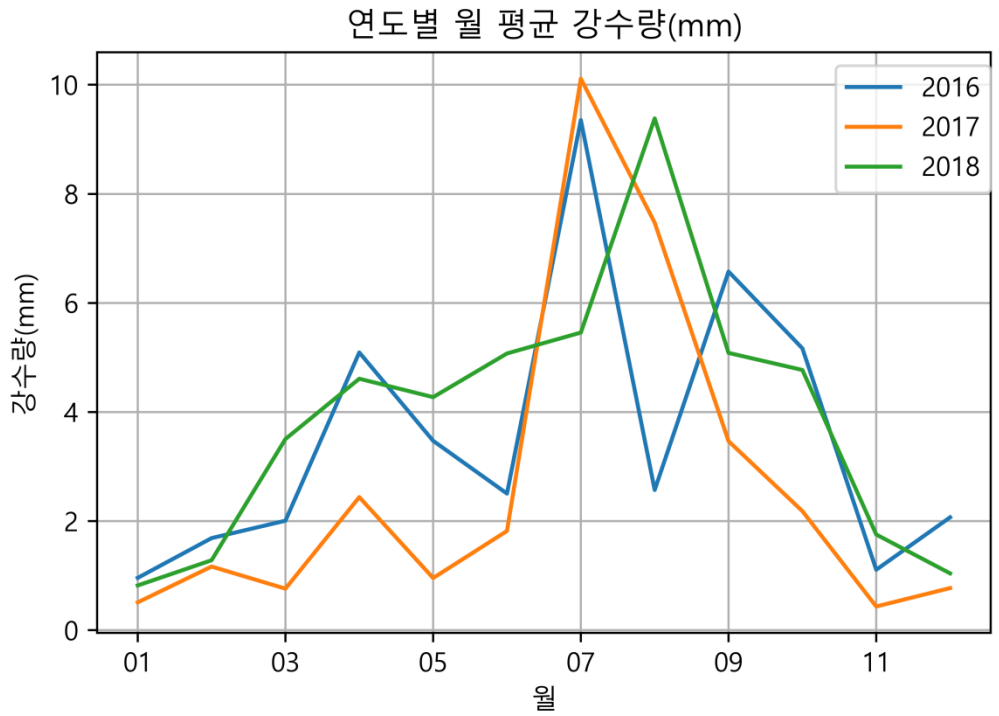
전체 데이터 셋(발병률)

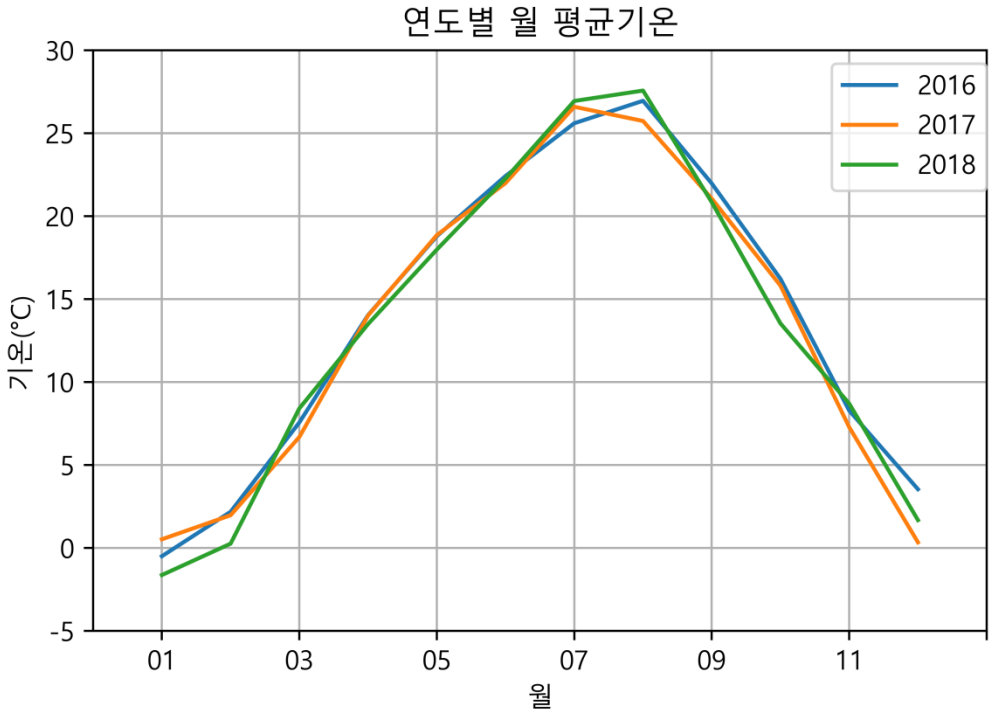
	시도코드	일시	발생건수	시도	년도	인구수	발병률
1	11	2016-01-01	217	서울	2016	9805506	0.002213
2	11	2016-01-02	2200	서울	2016	9805506	0.022436
3	11	2016-01-03	267	서울	2016	9805506	0.002723
4	11	2016-01-04	3244	서울	2016	9805506	0.033083
5	11	2016-01-05	2163	서울	2016	9805506	0.022059
6	11	2016-01-06	2197	서울	2016	9805506	0.022406
7	11	2016-01-07	2134	서울	2016	9805506	0.021763
8	11	2016-01-08	2332	서울	2016	9805506	0.023783
9	11	2016-01-09	1824	서울	2016	9805506	0.018602
10	11	2016-01-10	207	서울	2016	9805506	0.002111
11	11	2016-01-11	3113	서울	2016	9805506	0.031747
12	11	2016-01-12	2029	서울	2016	9805506	0.020692
13	11	2016-01-13	1985	서울	2016	9805506	0.020244
14	11	2016-01-14	2056	서울	2016	9805506	0.020968
15	11	2016-01-15	2364	서울	2016	9805506	0.024109
16	11	2016-01-16	1847	서울	2016	9805506	0.018836

당해 년도의 시도의 인구수를 이용한 발병률 계산

시각화

시각화
기상데이터

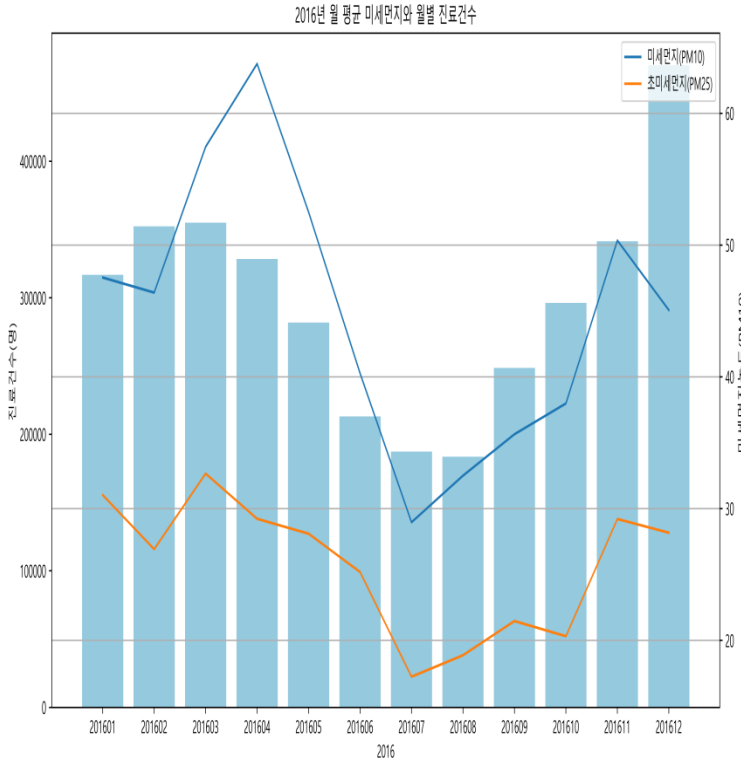


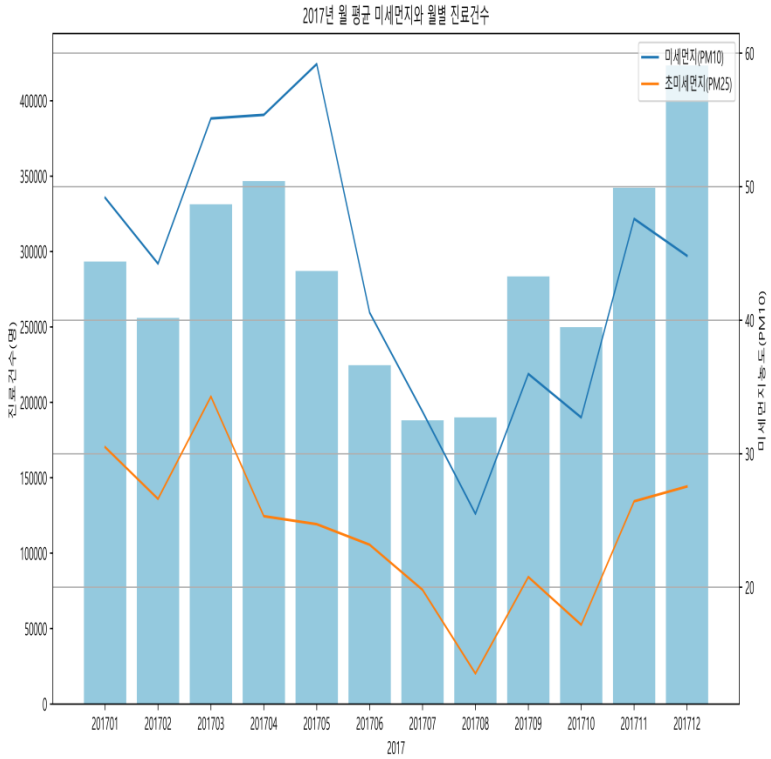


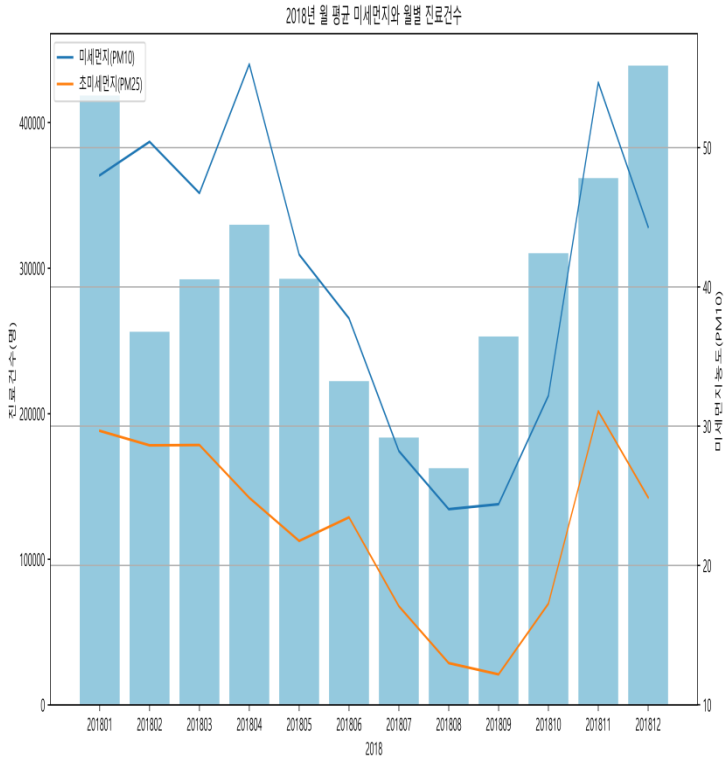
연도별 기상요인의 변화

시각화

미세먼지 & 진료데이터



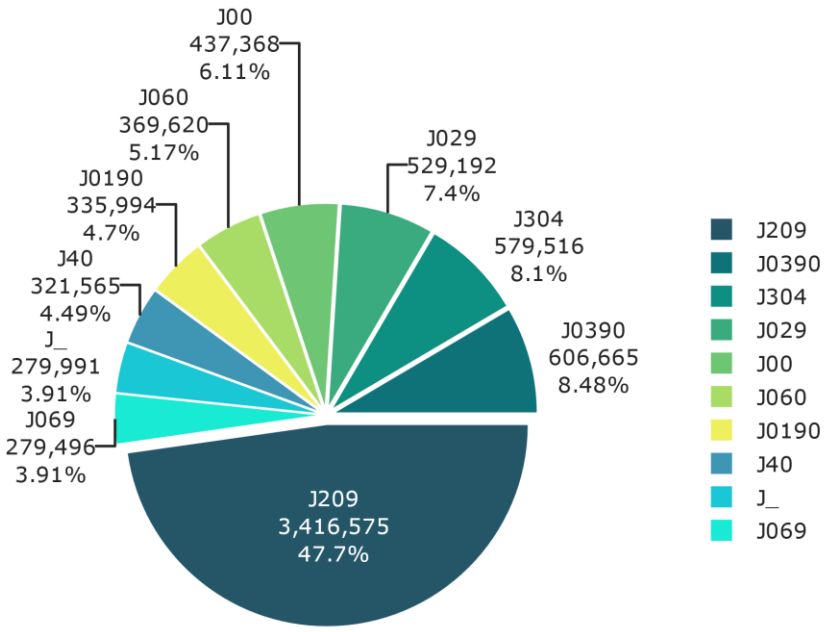




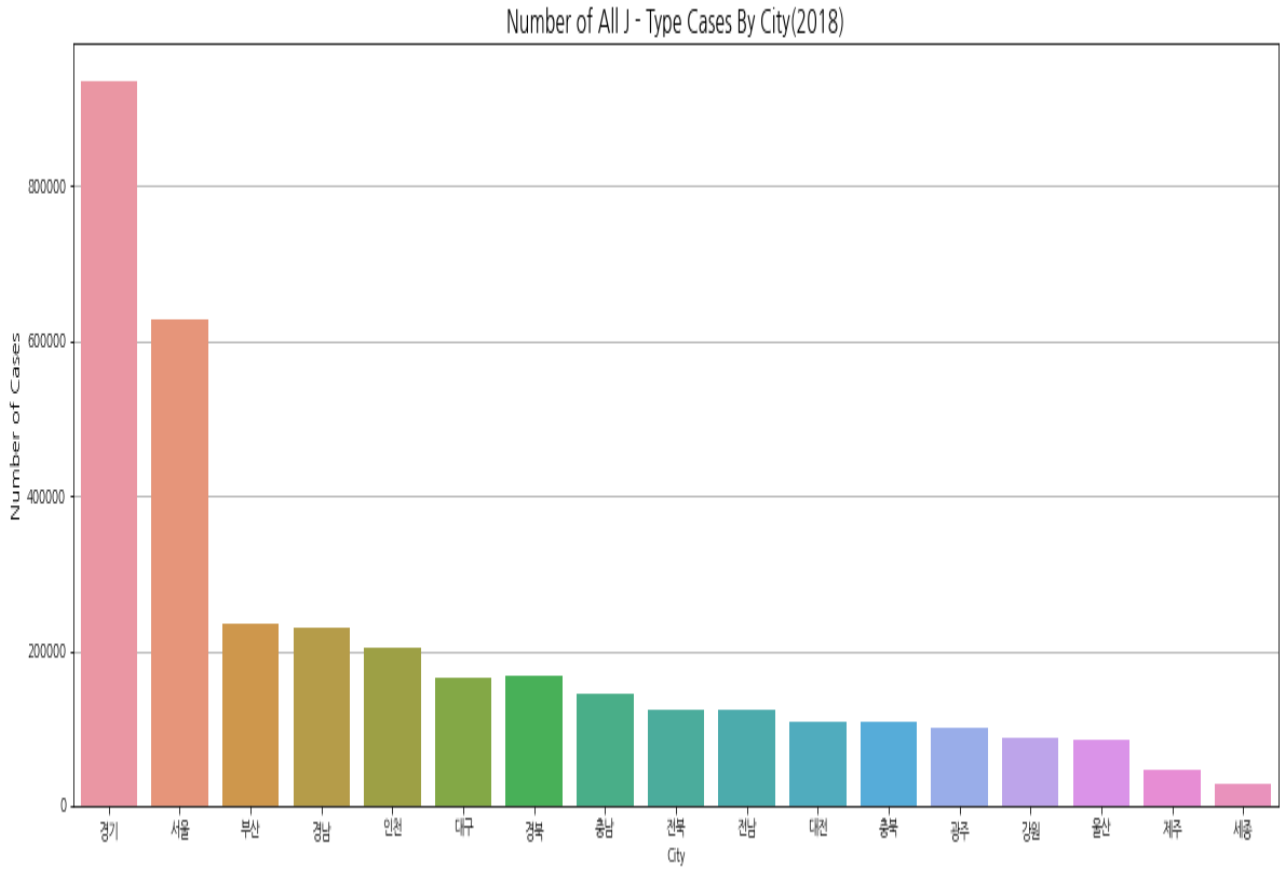
연도별 진료건수와 미세먼지간의 관계

시각화

의료데이터



호흡기 질환 질병코드의 비율

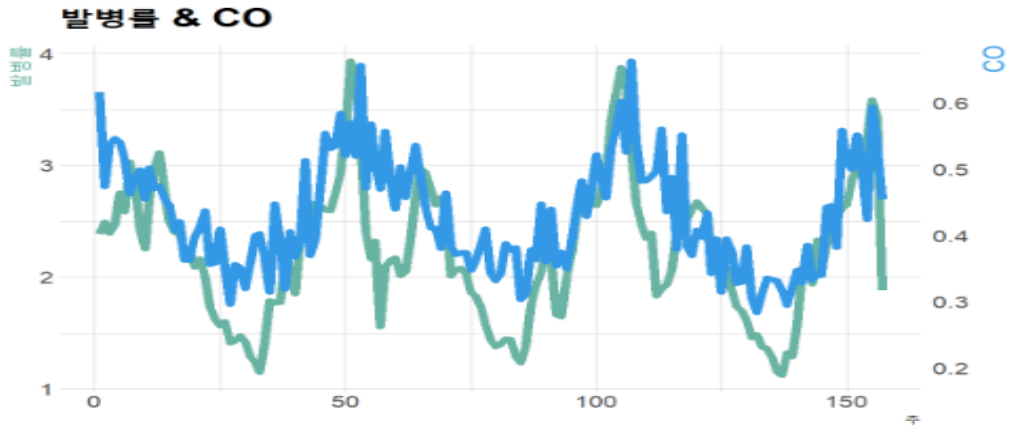


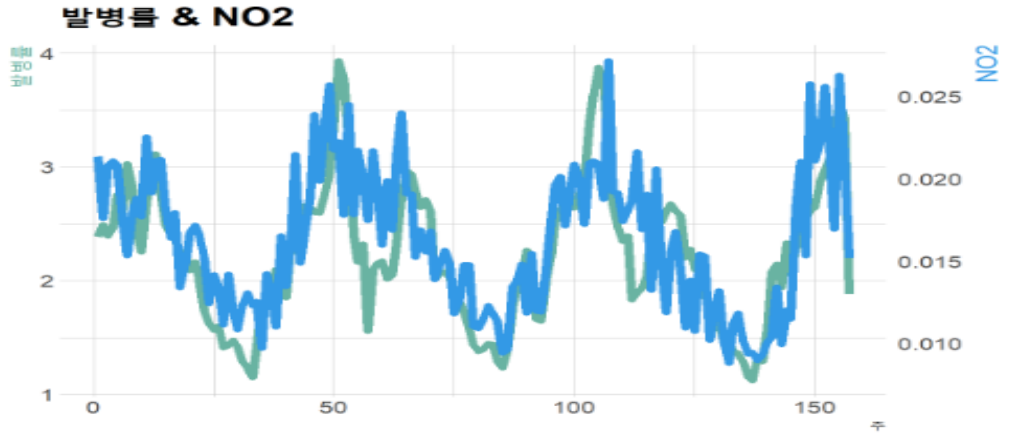
시도별 호흡기질환 발생건수

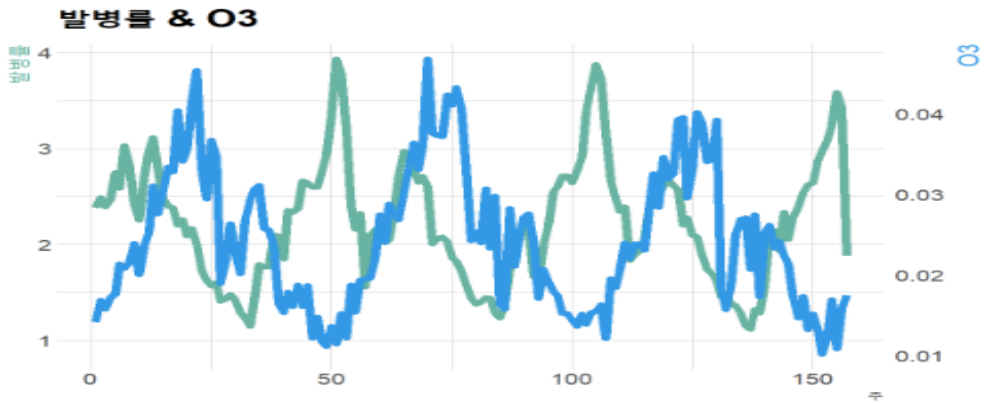
분석

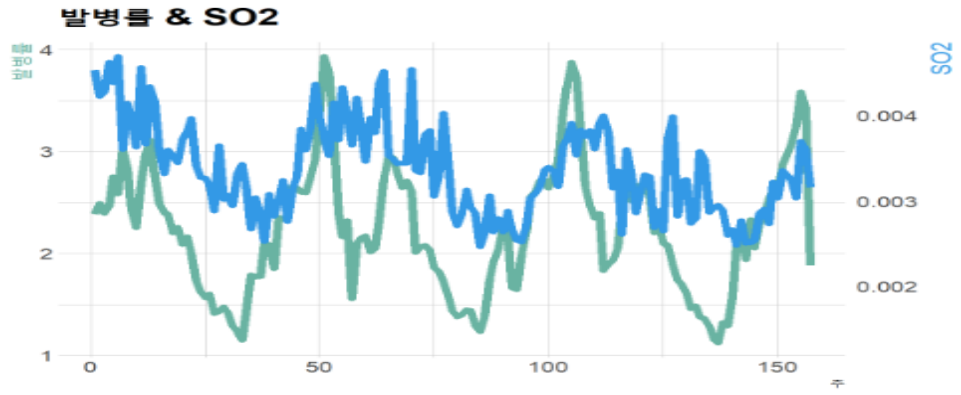
분석

요인 분석(시계열 분석)





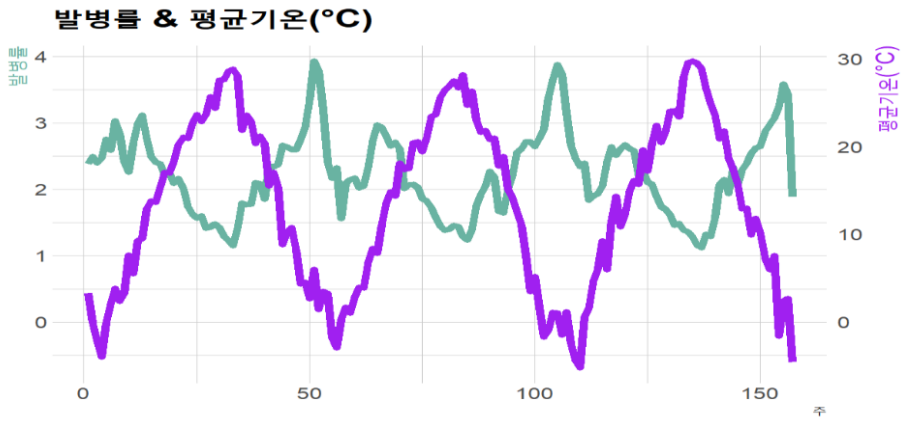


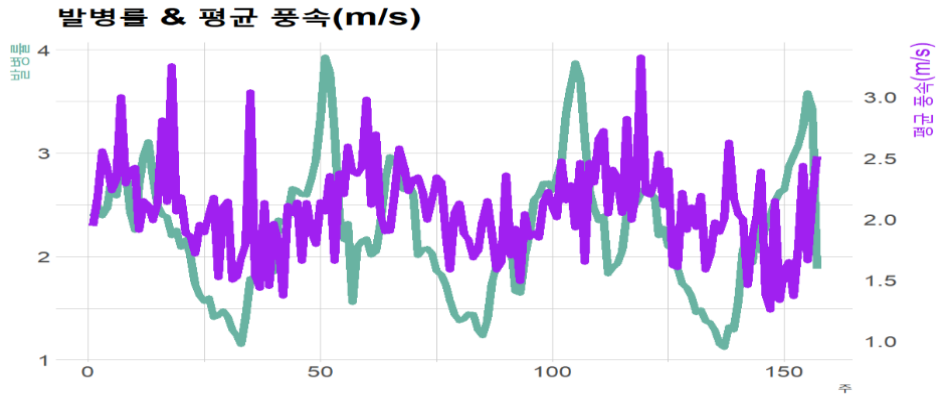


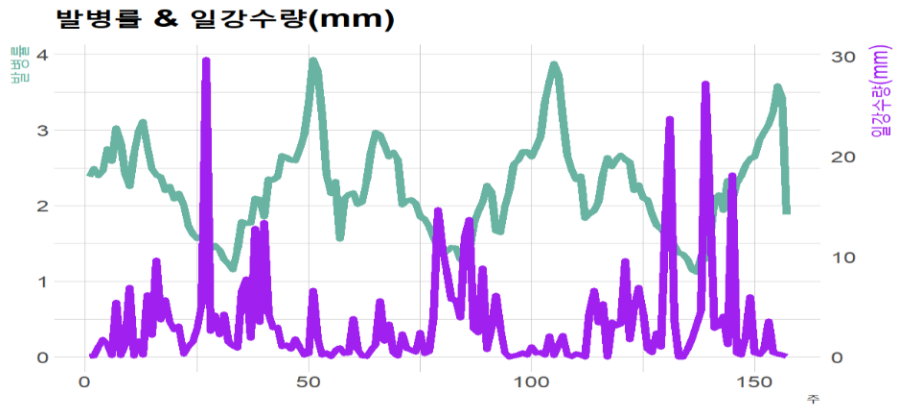
오염물질중 발병률과 O3는 지연효과를 더했을때 거의 비슷한 추이를 보인다.
O3를 제외한 나머지 오염물질은 비슷한 패턴을 보인다.

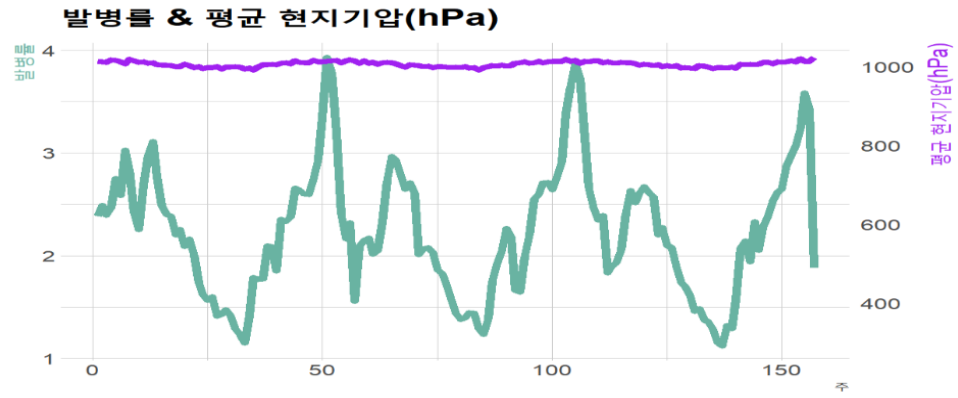
분석

요인 분석(시계열 분석)





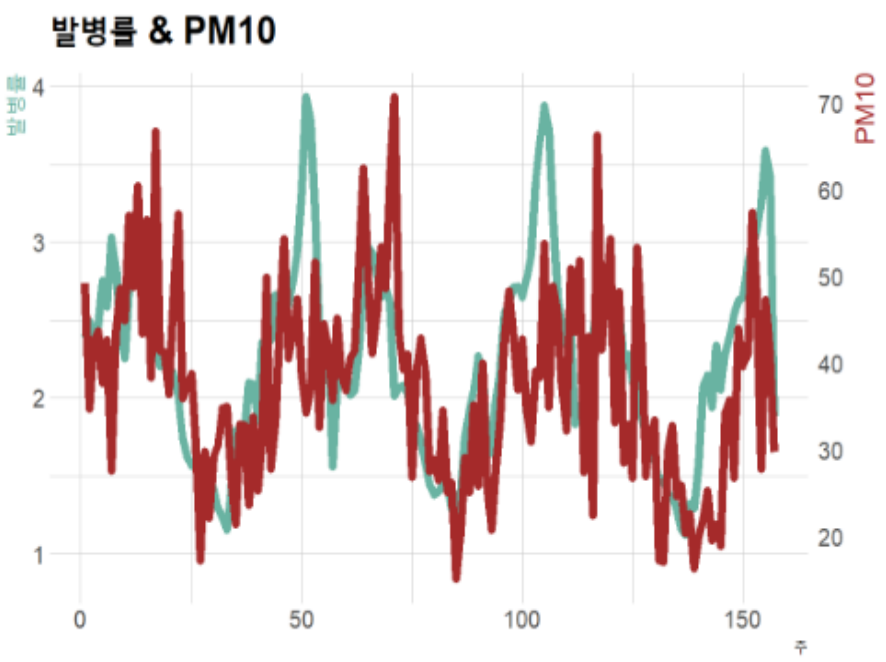


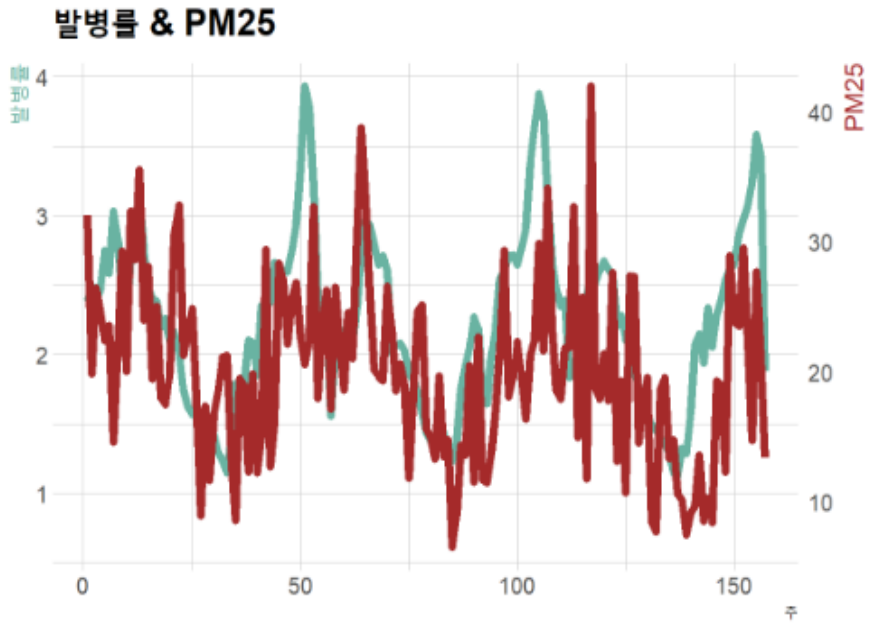


기상요인중 발병률과 가장 비슷한 패턴을 보이는 것은 없는 것으로 보인다

분석

요인 분석(시계열 분석)

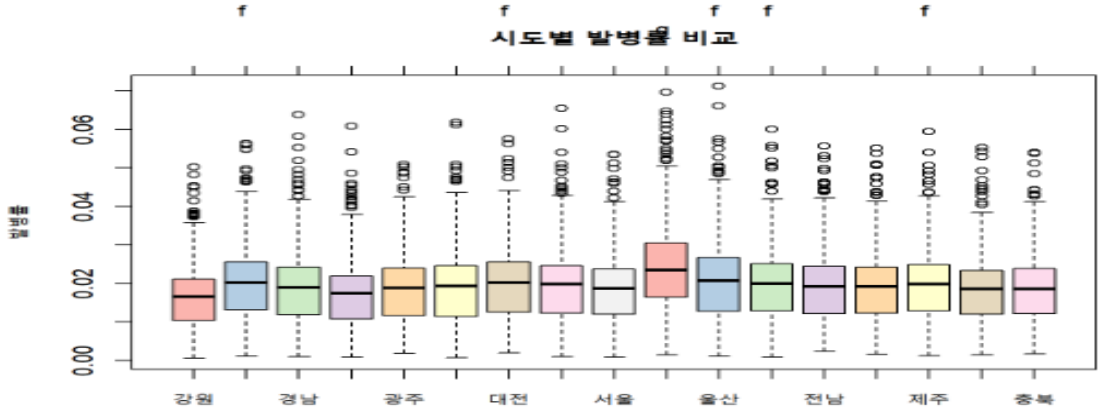


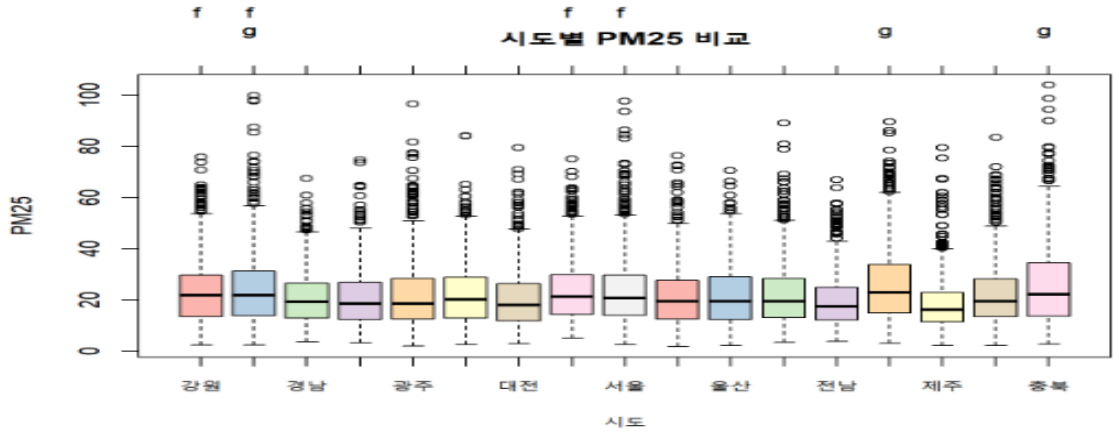


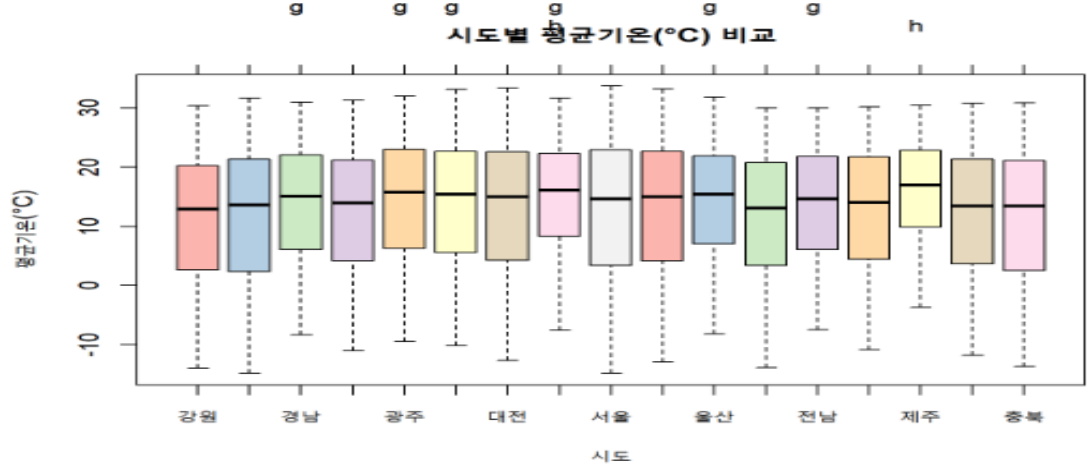
미세먼지와 발병률간에는 거의 비슷한 패턴은 아니지만 두번의 봉오리가 보이는 패턴이 보인다
PM10과 PM25와는 거의 비슷한 패턴을 보인다.

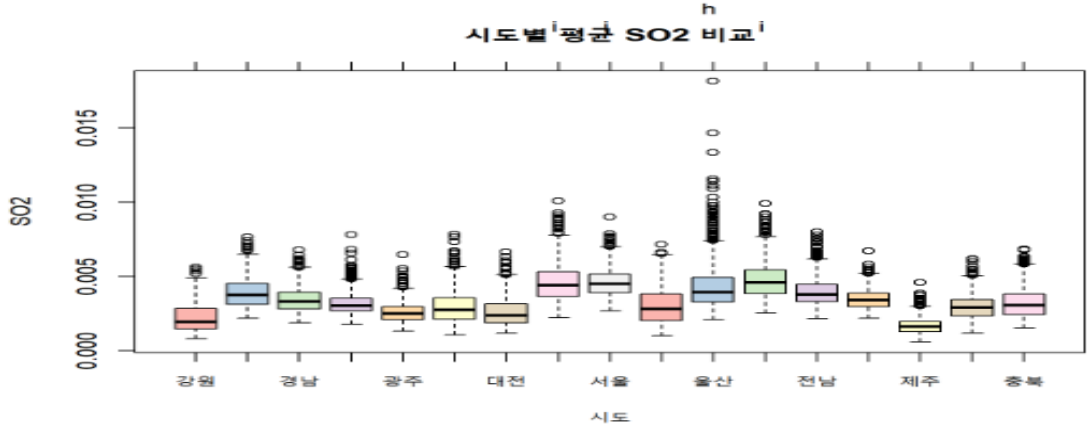
분석

요인 분석(분산분석)





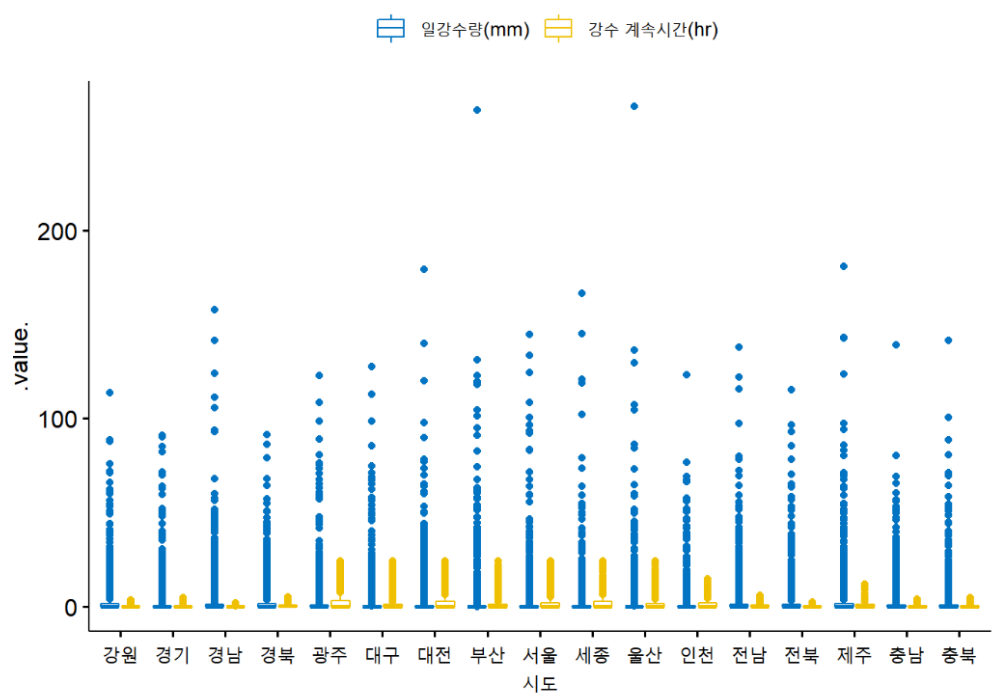


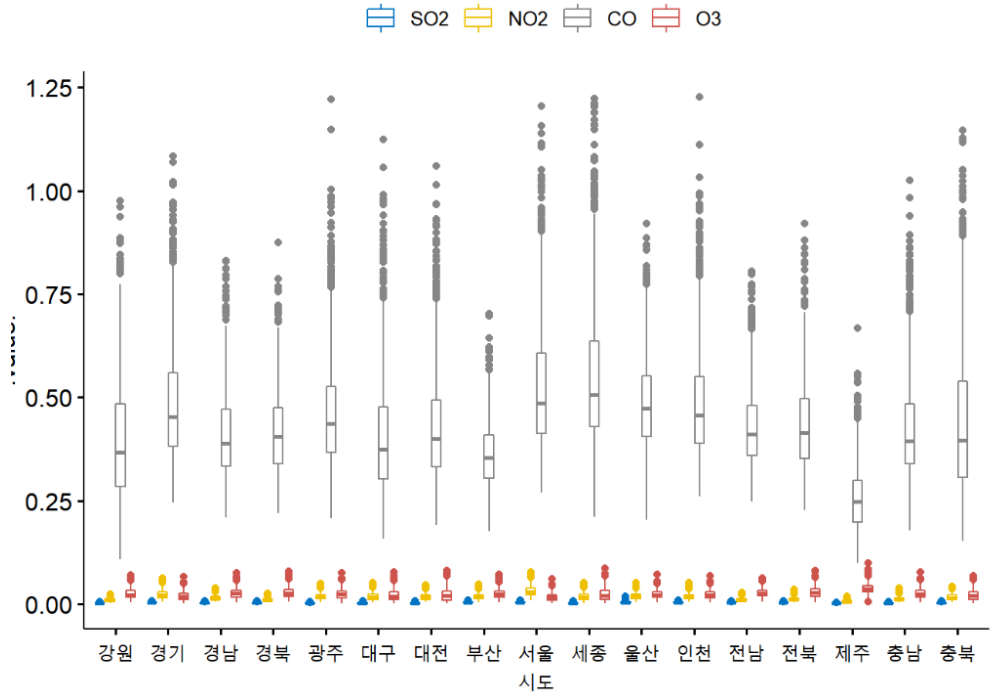


시도별 요인들의 차이가 들어난다. 근처에 위치한 지역끼리 비슷한 값을 가지고 있다.

분석

요인 분석(분산분석)

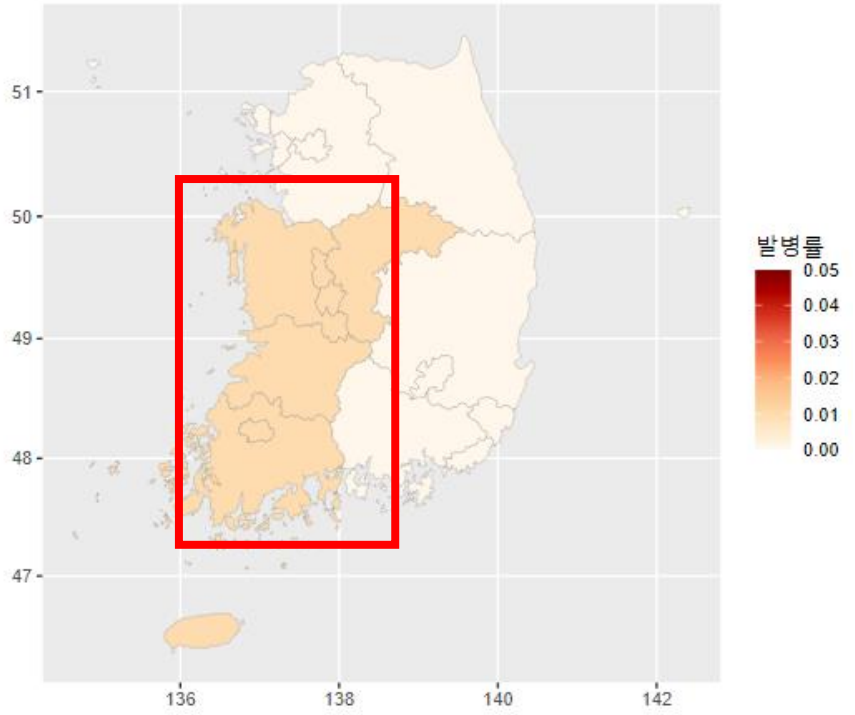


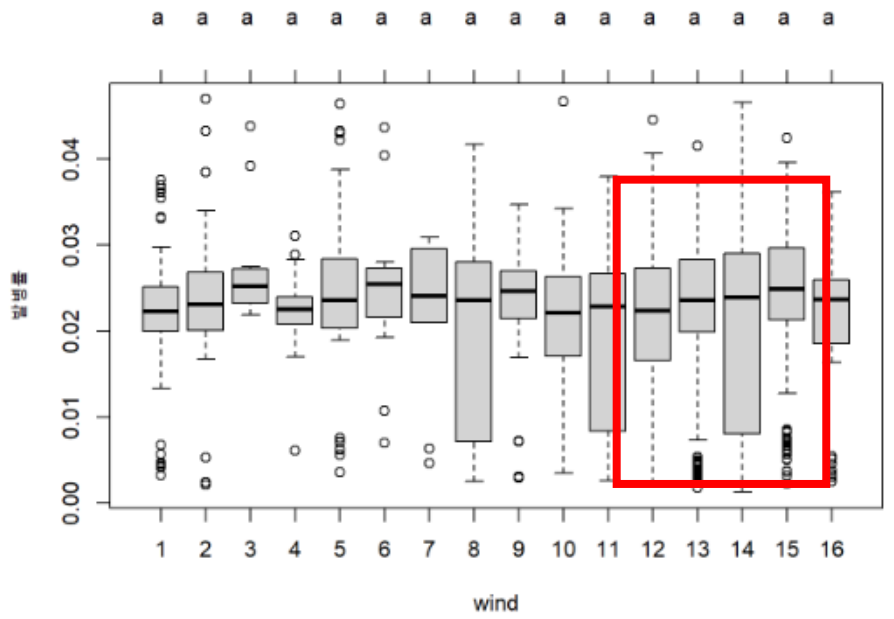


일강수량과 강수 계속시간이 같은 지역에서 높은 값을 보인다.
오염물질 간에도 같은 지역에서 높은 값을 보인다.

분석

요인분석(풍향별 미세먼지 농도 분석)



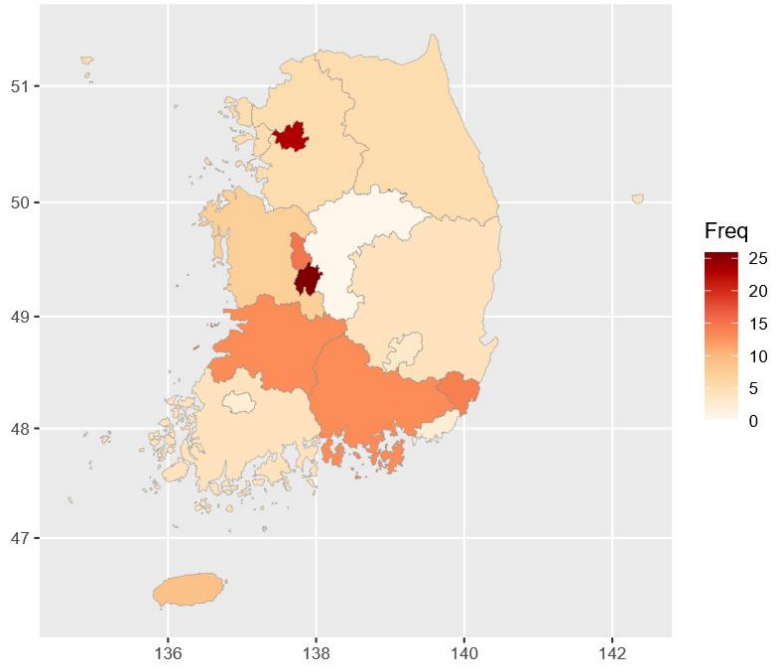


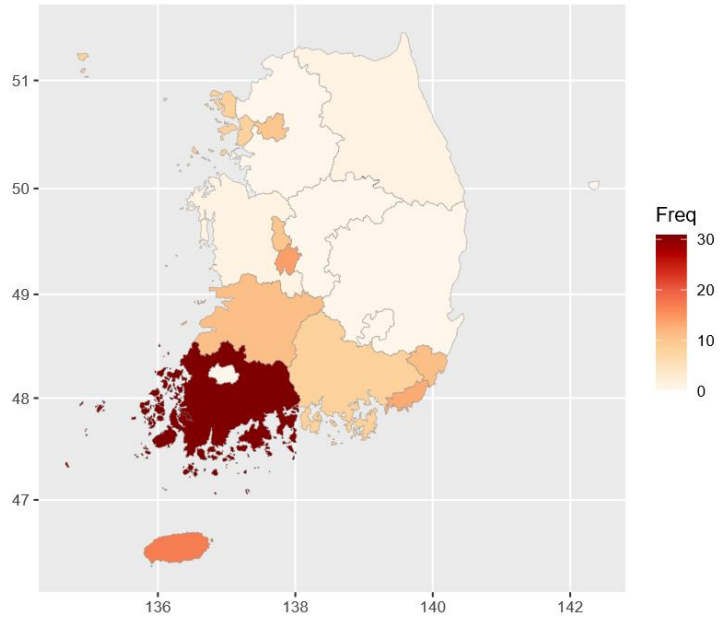
2016년도 1분기 지역별,풍향별 발병률 차이

중국과 맞닿는 북서풍 방향에서 발병률 평균이 높은 것으로 보인다

분석

요인분석(풍향별 미세먼지 농도 분석)

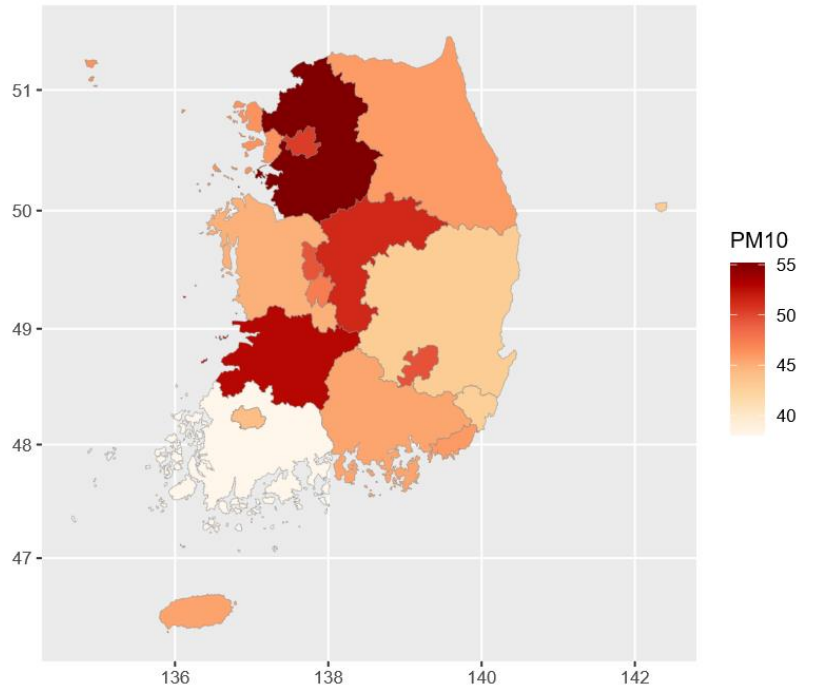


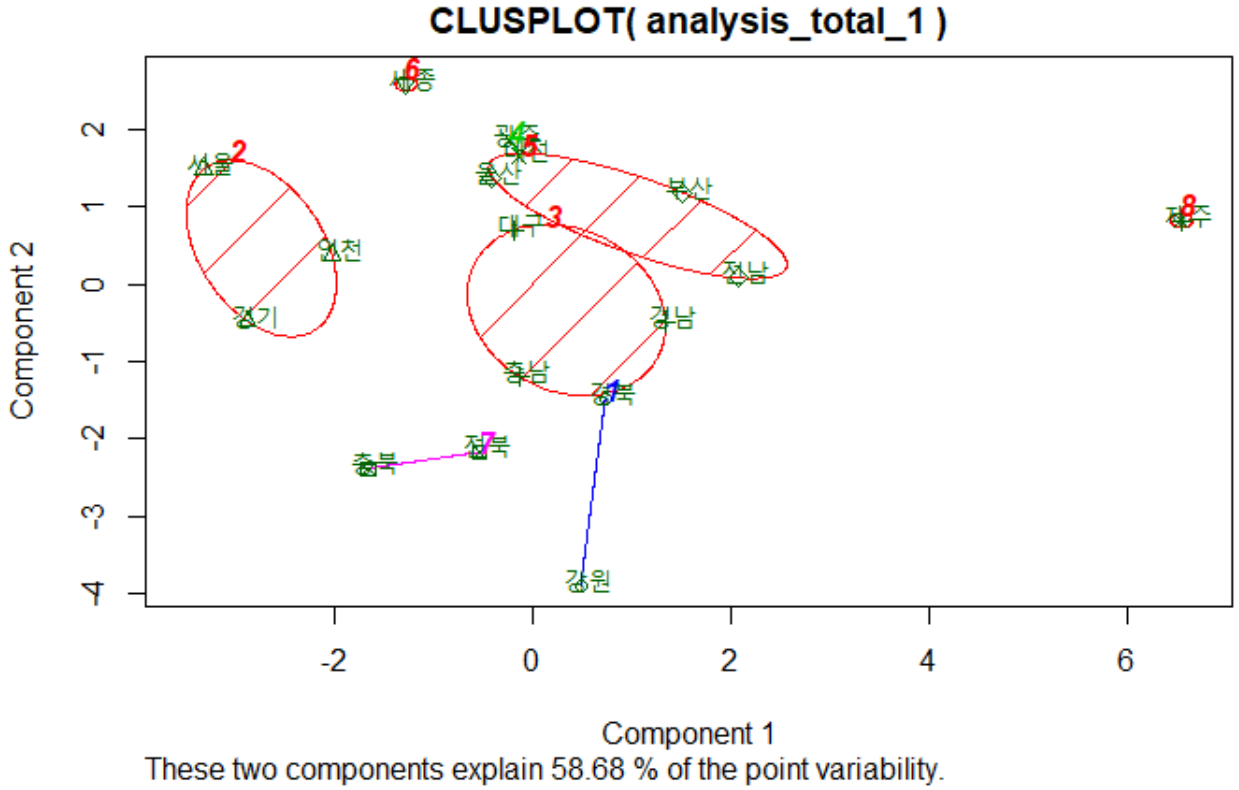


중국풍 바람 방향 일때 세종의 풍향의 빈도수가 높은 것으로 보인다
중국풍 바람과 발병률간의 관계를 추론할 수 있다

분석

요인분석(군집 분석)

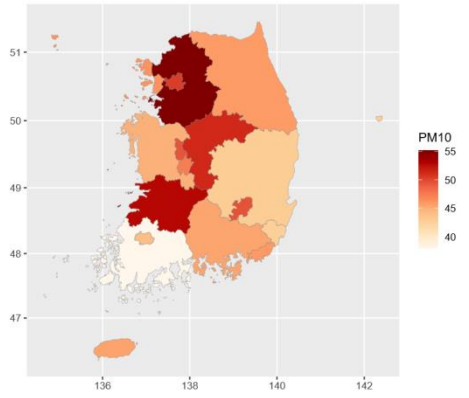
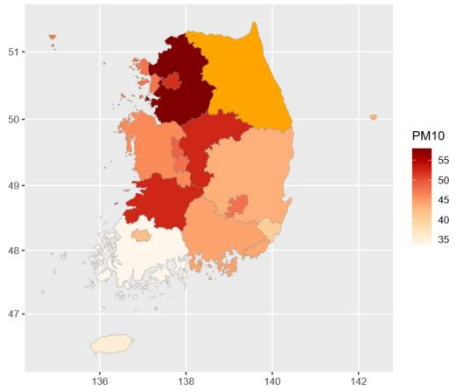
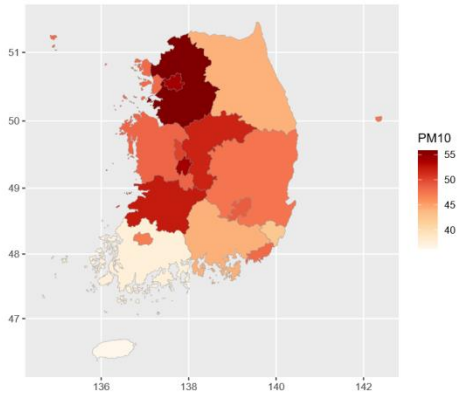
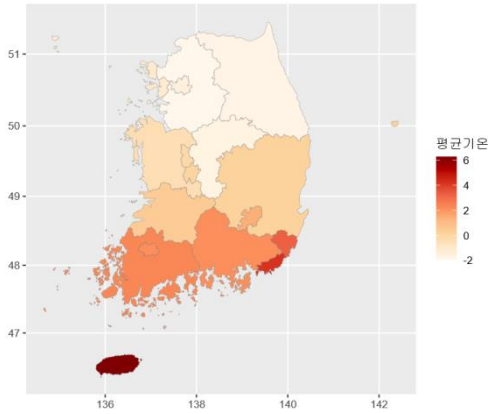
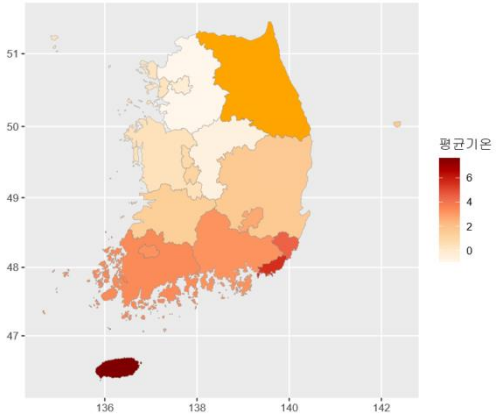
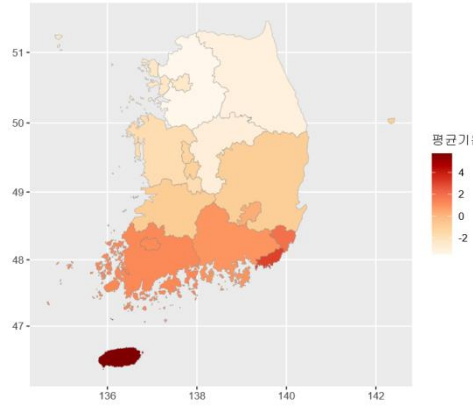




발병률이 특히 높은 세종 및 여러 지역별 특징을 비교

분석

요인분석(군집 분석)

	2016	2017	2018
PM10			
평균기온			

연도별로는 군집별 같은 특징을 보이고 분기별로는 차이를 보인다

분석

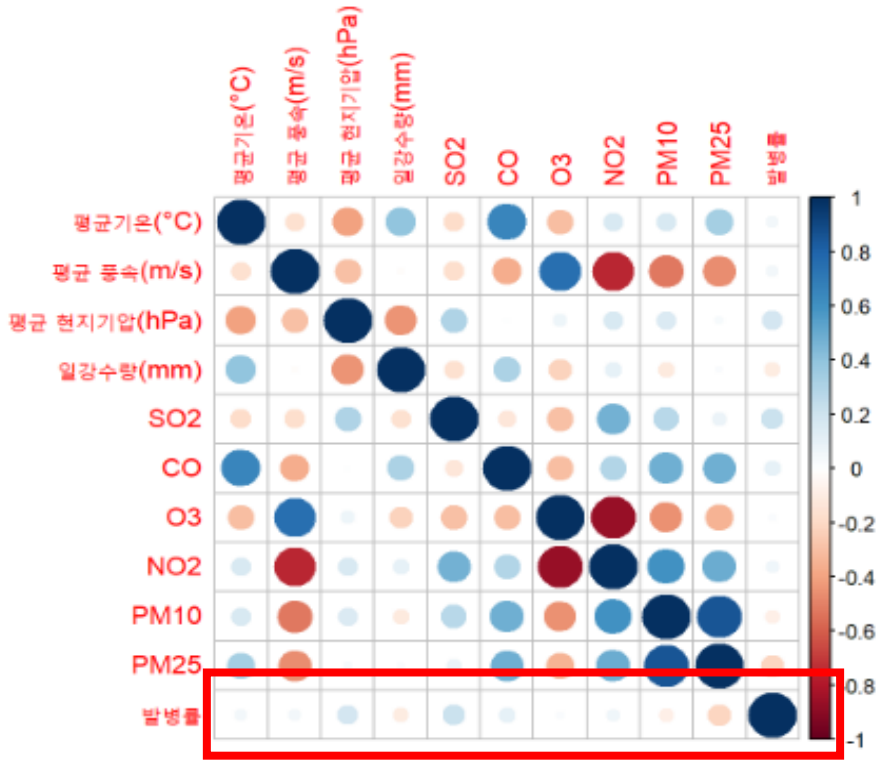
요인분석(군집 분석)

통합														
군집	발병률	발생건수	PM10	PM25	SO2	CO	O3	NO2	평균기온	평균풍속	평균현지기압	일 최심신적설(봄)	일강수량	강수 계속시간
1	2	2	3	4	3	4	3	2	3	2	2	4		
2	3	5	5	5	5	4	2	4	2	2	5		2	2
3	2	2	4	4	2	2		2	3	2	4	1	2	
4	3	2	2	2	1	4	3	2		2	4	5	2	
5	2	2		3	2	4	2	2	1		4	1	3	
6	5	2	2	4	2	5	2	2	3	1	3	5	2	3
7	2	3	4	4	2	3		2	2	2	3	1	3	2
8	4	2	2	2	1	1	4	1	5	5	5	1	2	2

군집별 특성을 찾아보고 발병률이 높은 6번 군집 세종과 발병률이 낮은 타군집간의 차이를 비교.
세종은 타군집에 비해 PM25,CO,일최심신 적설이 높고 평균 풍속이 낮은 것으로 나타났다.
이결과가 회귀분석에서도 나타나는지 확인 해보았다.

분석

군집 회귀분석



2016년 1분기 세종의 상관분석표

연도별 분기별 군집데이터 추출

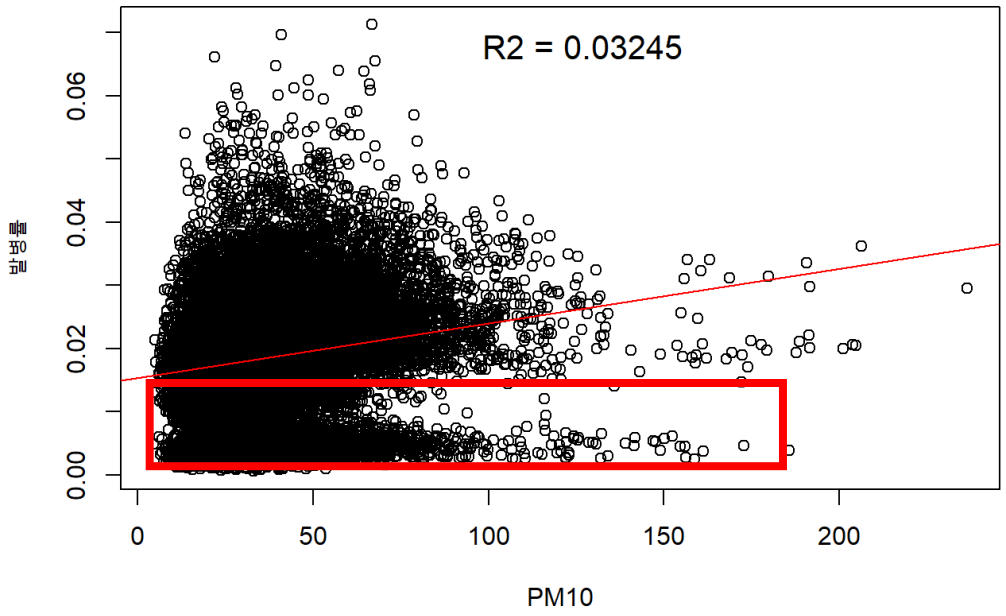
요인의 상관관계 분석

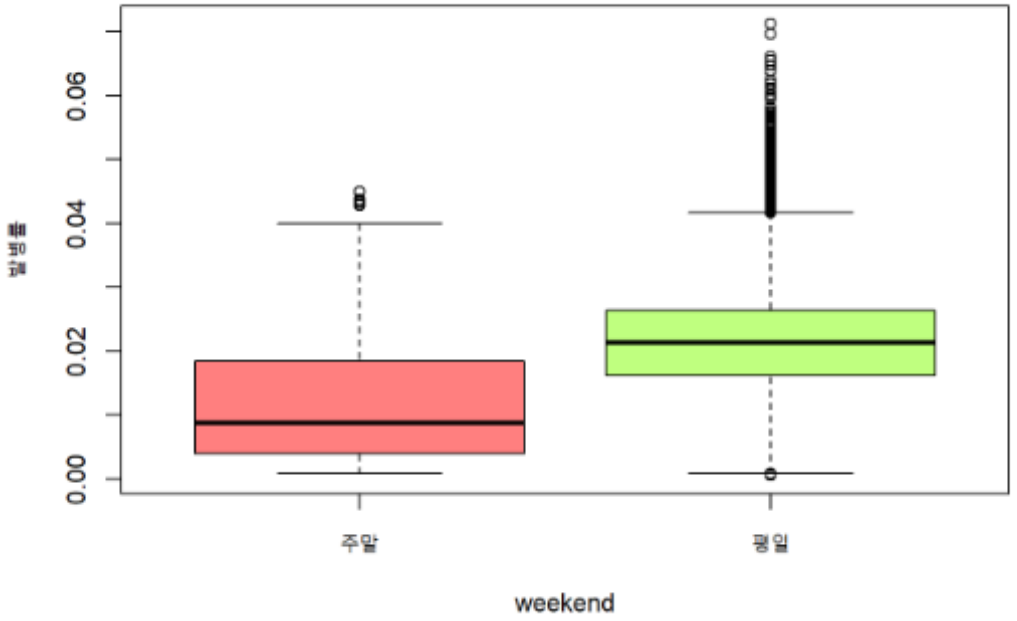
StepAIC를 이용한 회귀분석 실행

세종 및 여러 군집의 발병률과 다른 요인간의 관계가 특별하게 드러나지 않았다

분석

시계열 회귀분석(일별->주별)

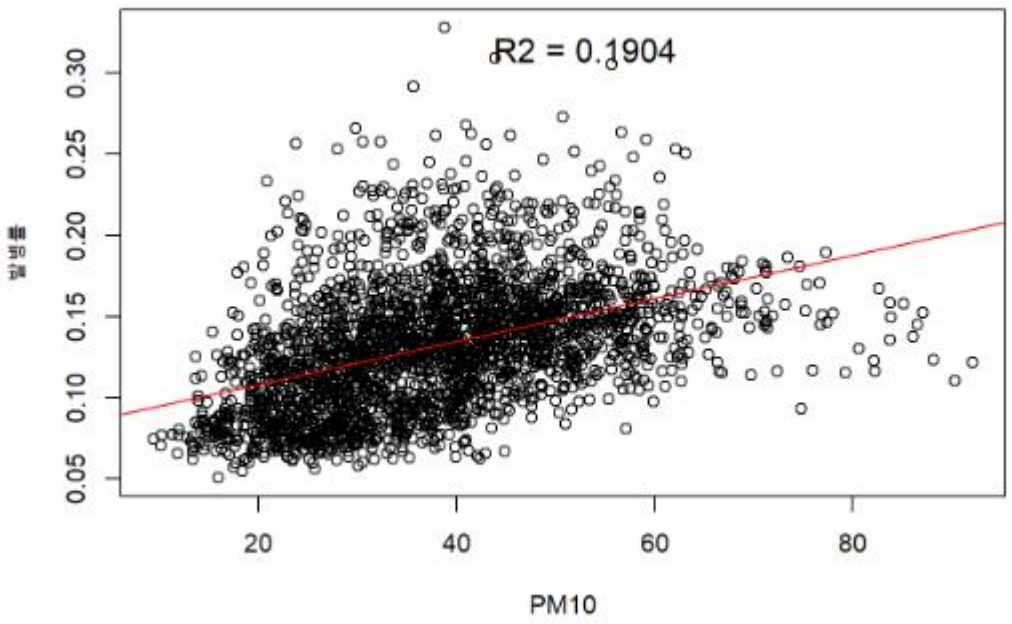


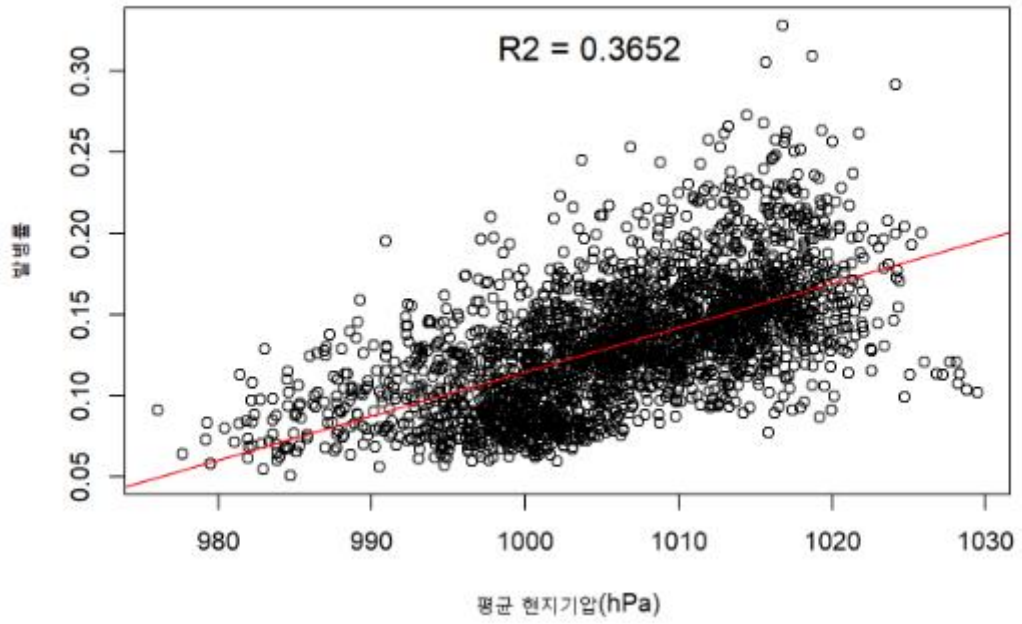


발병률이 0값에 가까운 값이 존재하는 것을 확인
평일과 주말의 발병률 차이를 확인

분석

시계열 회귀분석(일별->주별)

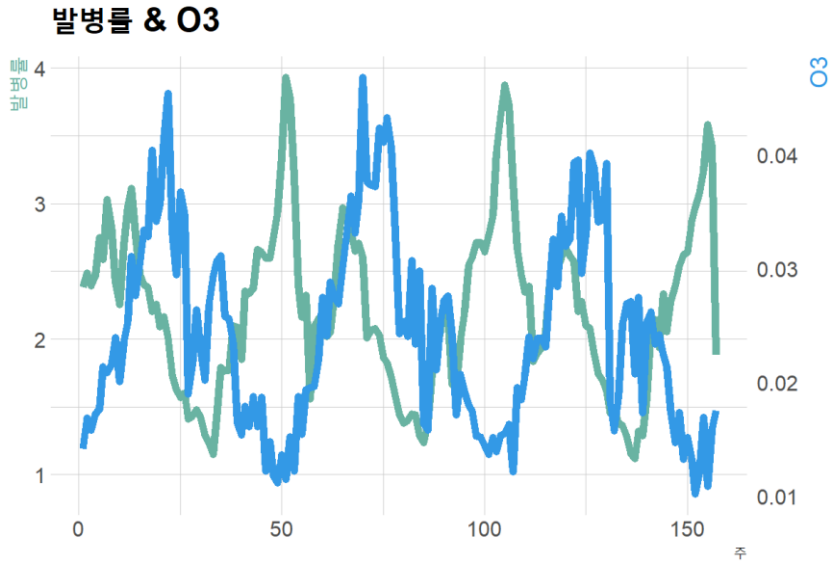


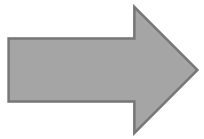


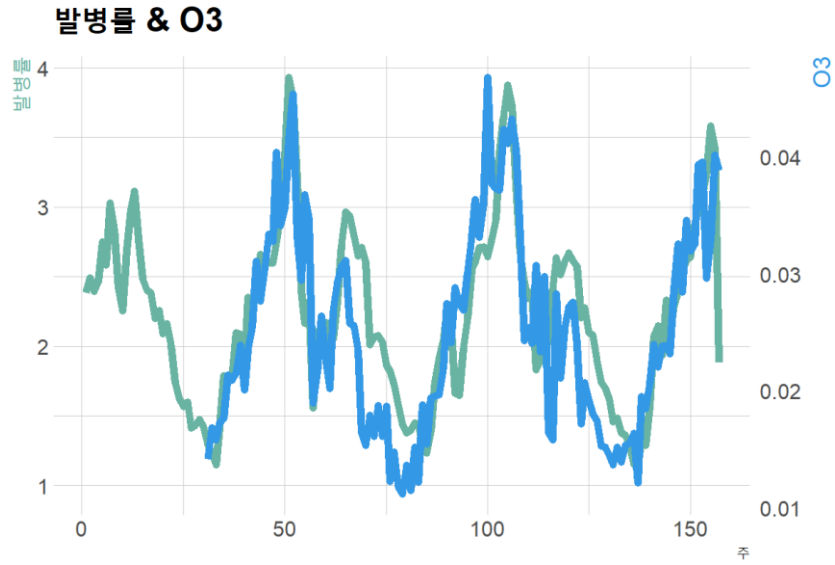
주별로 평균을 내렸을 때 발병률의 값이 개선된 것을 확인

분석

시계열 회귀분석(지연효과)







O3등 여러 요인들이 지연효과를 더 하였을때 발병률과 패턴이 비슷해지는 경우 발생

분석

시계열 회귀분석(지연효과)

```
ts_O3_original <- ts(lag(analysis_total_week$O3,0), start = c(2016, 1), freq = 52)#omnibus o
ts_O3 <- ts(lag(analysis_total_week$O3,30), start = c(2016, 1), freq = 52)#omnibus o
ts_NO2 <- ts(lag(analysis_total_week$NO2,0), start = c(2016, 1), freq = 52)#omnibus x
ts_CO <- ts(lag(analysis_total_week$CO,0), start = c(2016, 1), freq = 52)#omnibus o
ts_SO2 <- ts(lag(analysis_total_week$SO2,0), start = c(2016, 1), freq = 52)#omnibus x
ts_PM10 <- ts(lag(analysis_total_week$PM10,0), start = c(2016, 1), freq = 52)#omnibus x
ts_PM25 <- ts(lag(analysis_total_week$PM25,0), start = c(2016, 1), freq = 52)#omnibus x
ts_rain <- ts(lag(analysis_total_week$`일강수량(mm)`,25), start = c(2016, 1), freq = 52)#omnibus o
ts_temperature <- ts(lag(analysis_total_week$`평균기온(°C)`,0), start = c(2016, 1), freq = 52)#omnibus x
ts_air <- ts(lag(analysis_total_week$`평균 현지기압(hPa)`,1), start = c(2016, 1), freq = 52)#omnibus o
ts_wind <- ts(lag(analysis_total_week$`평균 풍속(m/s)`,36), start = c(2016, 1), freq = 52)#omnibus o
```

각 요인별 지연효과 적용

분석

시계열 회귀분석(최종결과)

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -20.017457   7.974557  -2.510  0.01340 *
## ts_O3        39.478557   6.708323   5.885 3.69e-08 ***
## ts_CO         0.681800   1.042304   0.654  0.51428
## ts_SO2      -64.997544 106.793460  -0.609  0.54392
## ts_PM10       0.018649   0.006345   2.939  0.00395 **
## ts_PM25      -0.013119   0.012804  -1.025  0.30762
## ts_air        0.020626   0.008135   2.535  0.01252 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3482 on 120 degrees of freedom
##   (30 observations deleted due to missingness)
## Multiple R-squared:  0.7225, Adjusted R-squared:  0.7086
## F-statistic: 52.07 on 6 and 120 DF,  p-value: < 2.2e-16
```

```
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = fit)
##
##                       Value p-value                Decision
## Global Stat        0.928224  0.9205 Assumptions acceptable.
## Skewness           0.009454  0.9225 Assumptions acceptable.
## Kurtosis           0.798750  0.3715 Assumptions acceptable.
## Link Function      0.102876  0.7484 Assumptions acceptable.
## Heteroscedasticity 0.017144  0.8958 Assumptions acceptable.
```

VIF

```
##    ts_O3    ts_CO   ts_SO2  ts_PM10  ts_PM25   ts_air
## 3.554605 8.167667 2.972158 5.240235 7.953823 3.003587
```

최적의 회귀식과 테스트 결과

결론

결론

분석 결론

미세먼지는 중국 방향에서 오는 바람과 연관이 있을 것으로 보인다.

미세먼지 오염물질 간에는 상관관계가 있다. 특히 NO2와 O3간의 연관작용이 존재한다.

지연 효과에 의한 기상요인과 발병률 간의 상관관계가 보인다.

호흡기 질환의 발병률은 O3 , PM10은 고도로 유의미하다.

호흡기 질환의 발병률은 평균 현지기압과 유의미한 관계이다.
(추가적인 연구가 필요할 것으로 보인다.)

결론

향후 발전 방향

1. 주간 시도별 회귀분석 진행

2. 웹을 통한 분석결과 제공

3.코로나와 같은 현재 문제와 관련 분석

결론

팀원 역할

팀원 이름	역 할
강지혜	미세먼지 데이터 전처리 & 시각화, Raw data munging, 분산 분석
백종호	진료내역 part, Raw data 추출, 회귀 분석
신용준	기상 part, 군집 분석, PPT 자료 작성
황태성	기상 part, Raw data munging, 시계열 분석
이현빈	진료내역 데이터 전처리 & 시각화, 상관 분석, 웹 서비스

Q & A

감사합니다