

빅데이터 언어 프로젝트 서울시 대기환경 분석

씽크빅 조
오승준 이원빈 김준서 송민주

INDEX

- 001 프로젝트 소개
- 002 데이터 설명
- 003 데이터 분석
- 004 GUI 개발
- 005 결론

PART 1

프로젝트 소개



공기

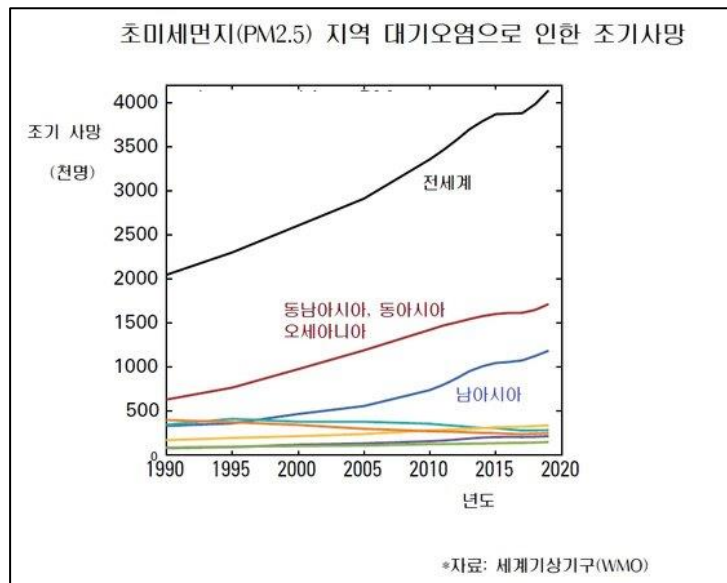
Air [空氣]

지구를 둘러싼 대기 하층을 구성하는 무색 투명한 기체로,
지구상 생물 존재에 꼭 필요한 역할을 한다.

WHO “대기오염으로 해마다 700만명 조기 사망”

f t v l s g

| ‘대기질 가이드라인’ 16년 만에 업데이트...“깨끗한 공기는 인간의 기본권”



“환경 오염으로 900만 명 조기 사망...미세먼지 영향 증가”

입력 2022.05.19 (08:17) | 수정 2022.05.23 (10:02)

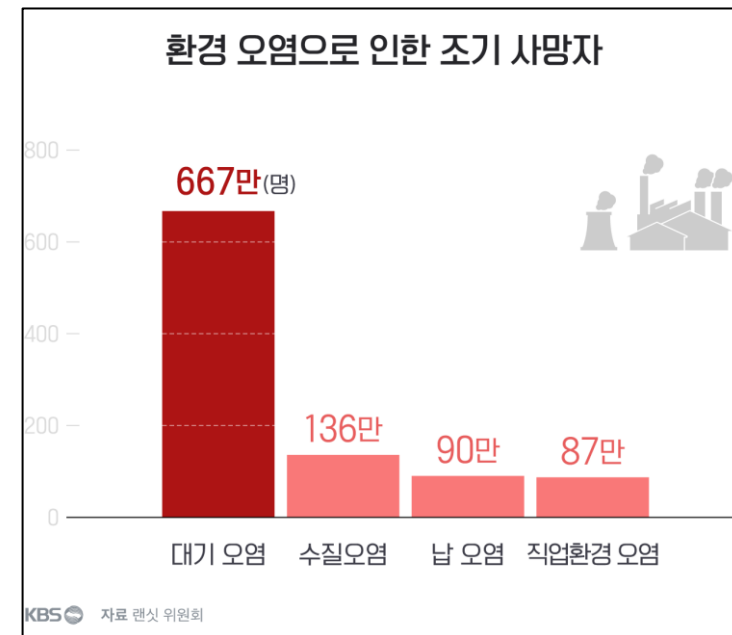
취재K

0 18

가



대기오염



주제

대기 오염이 많이 일어난 시기에는 시각 정보에 따른 특성이 존재할 것이다.

가설

대기 오염이 많이 된 시기에는 오염된 공기로 인해 호흡기 질환과 관련된 환자들이 증가할 것이다.

목표

시계열 데이터의 주기별 특성을 통해 오염된 공기에 대한 새로운 정책이나 서비스를 제공할 수 있지 않을까?

PART 2

데이터 설명



데이터셋

Home > 공공데이터 > 공공데이터

찾고 싶은 데이터를 입력해 주세요.



상세 검색

통합 검색

☐ 결과 내 재검색

공공데이터



환경

활용사례(갤러리) 등록

URL 복사

목록 이동

서울시 기간별 시간평균 대기환경 정보

대기 환경지수, 미세먼지, 오존, 이산화질소, 일산화탄소, 아황산가스 등의 기간별 시간평균 대기환경정보를 제공합니다.
※ Sheet 서비스는 최근 2달 이내의 데이터만 출력합니다.

자료 수집: 서울 열린 데이터 광장 [서울시 기간별 시간평균 대기환경 정보]



감염병 누리집
Infectious Disease Homepage

검색어를 입력하세요



법정감염병 | 전수감시감염병 | 표본감시감염병 | 지침 | 발간자료 | 자주묻는질문

전체
보기



급성호흡기감염증

표본감시감염병 통계정보를 안내 해드립니다.

인플루엔자

기생충감염증

수족구병

합병증동반수족구병

성매개감염병

해외유입기생충감염증

장관감염증

급성호흡기감염증

중증급성호흡기감염증

엔테로바이러스감염증

안과감염병

의료관련감염병

C형간염

기간

2022 년 01 주 ~ 2022 년 52 주

기간구분

☒ 주별 ☐ 연도별

감염병명

바이러스 전체 ☐ 사망

통계작성

자료 수집: 질병관리청 [급성호흡기감염증 통계자료]

Data Import

```
# [Dataset] : https://www.kaggle.com/datasets/bappekim/air-pollution-in-seoul
air_quality_dataset = pd.read_csv('{}seoul_air_1988_2021.csv'.format(file_path))
display(air_quality_dataset.head())
print(air_quality_dataset.info())
```

	dt	loc	lat	long	so2	no2	co	o3	pm10	pm2.5
0	1988010100	103	37.540037	127.002661	NaN	0.007	10.3	0.000	NaN	NaN
1	1988010100	105	37.593730	126.947561	0.340	0.055	12.6	0.043	NaN	NaN
2	1988010100	107	37.542043	127.047497	0.399	0.046	13.4	NaN	NaN	NaN
3	1988010100	108	37.547185	127.090304	0.261	0.034	5.4	0.000	NaN	NaN
4	1988010100	113	37.654140	127.026801	0.443	0.039	14.6	0.000	NaN	NaN

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5984782 entries, 0 to 5984781
Data columns (total 10 columns):
#   Column  Dtype
---  -
0    dt      int64
1    loc      int64
2    lat     float64
3    long     float64
4    so2      float64
5    no2      float64
6    co       float64
7    o3       float64
8    pm10     float64
9    pm2.5    float64
dtypes: float64(8), int64(2)
memory usage: 456.6 MB
None
```

[서울시 기간별 시간평균 대기환경 정보]

2015년 ~ 2021년 호흡기 환자 수 Data Import

```
# 2015년 ~ 2021년 호흡기 환자 수 데이터 import
# [DataSet] : https://www.kdca.go.kr/npt/biz/npp/iss/ariStatisticsMain.do
hospital_df = pd.read_csv('{}hospital.csv'.format(file_path))
st_row = hospital_df.columns # 1행의 column 지정되어있음
st = {'year':2015,'week':1,'germ1':1214,'germ2':161,'vir1':7,'vir2':70,'vi
col_list=['year','week','germ1','germ2','vir1','vir2','vir3','vir4','vir5'
hospital_df.columns=col_list # column 수정
hospital_df = hospital_df.append(st,ignore_index=True) # 1행 데이터 추가
hospital_df = hospital_df.drop(columns='vir10')
display(hospital_df.head())
print(hospital_df.info())

# 전체 환자 수에 대한 파생변수 sum 생성
hospital_df['germ_sum'] = hospital_df['germ1'] + hospital_df['germ2']
hospital_df['vir_sum'] = hospital_df['vir1'] + hospital_df['vir2'] + hospi
hospital_df['all_sum'] = hospital_df['vir_sum'] + hospital_df['germ_sum']
```

	year	week	germ1	germ2	vir1	vir2	vir3	vir4	vir5	vir6	vir7	vir8	vir9
0	2015	2	1114	149	13	90	13	8	519	136	14	70	102
1	2015	3	1139	141	7	70	9	19	387	130	17	53	306
2	2015	4	1025	128	15	78	11	14	281	164	43	43	248
3	2015	5	1002	145	10	56	9	19	299	170	19	43	232
4	2015	6	1052	161	10	60	17	18	213	189	22	37	325

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 365 entries, 0 to 364
Data columns (total 13 columns):
#   Column  Non-Null Count  Dtype
---  -
0    year    365 non-null    int64
1    week    365 non-null    int64
2    germ1    365 non-null    int64
3    germ2    365 non-null    int64
4    vir1     365 non-null    int64
5    vir2     365 non-null    int64
6    vir3     365 non-null    int64
7    vir4     365 non-null    int64
8    vir5     365 non-null    int64
9    vir6     365 non-null    int64
10   vir7     365 non-null    int64
11   vir8     365 non-null    int64
12   vir9     365 non-null    int64
dtypes: int64(13)
memory usage: 37.2 KB
```

[급성호흡기감염증 통계자료]



아황산가스

불쾌하고 자극적인 냄새가 나는 무색의 불연성 기체
발전소, 난방장치 등에서 발생



오존

무색, 무미, 또는 해초냄새의 산화력이 강한 기체
자동차, 도로포장 등에서 발생




일산화탄소

무색, 무취의 맹독성 기체
주방, 담배연기, 산불 등에서 발생



이산화질소


적갈색의 자극성 냄새가 있는 유독한 기체
고온 연소공정과 화학물질 제조공정 등에서 발생

A diagram for PM-10. It features a large blue circle at the top containing the text "PM-10". Below the circle is a smaller blue rounded rectangle containing the text "미세먼지". At the bottom is a larger light blue rounded rectangle containing two lines of text: "직경이 10μm이하인 먼지" and "자동차, 난방 및 에너지 사용 등으로 발생".

PM-10

미세먼지

직경이 10 μ m이하인 먼지
자동차, 난방 및 에너지 사용 등으로 발생

A diagram for PM-2.5. It features a large blue circle at the top containing the text "PM-2.5". Below the circle is a smaller blue rounded rectangle containing the text "초미세먼지". At the bottom is a larger light blue rounded rectangle containing two lines of text: "직경이 2.5μm이하인 먼지" and "자동차, 난방 및 에너지 사용 등으로 발생".

PM-2.5

초미세먼지

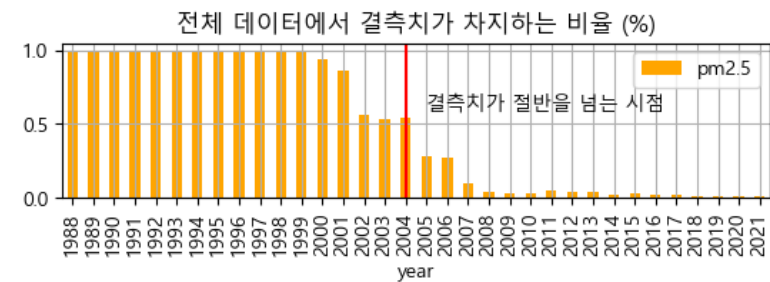
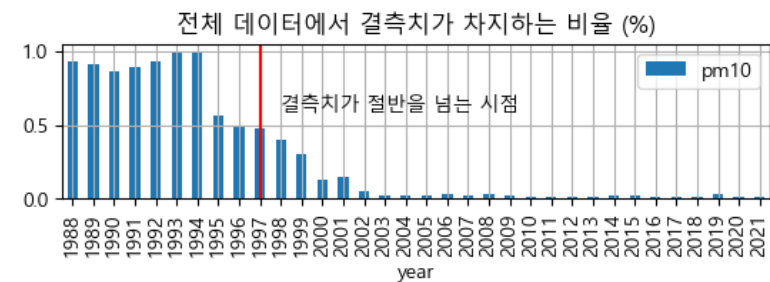
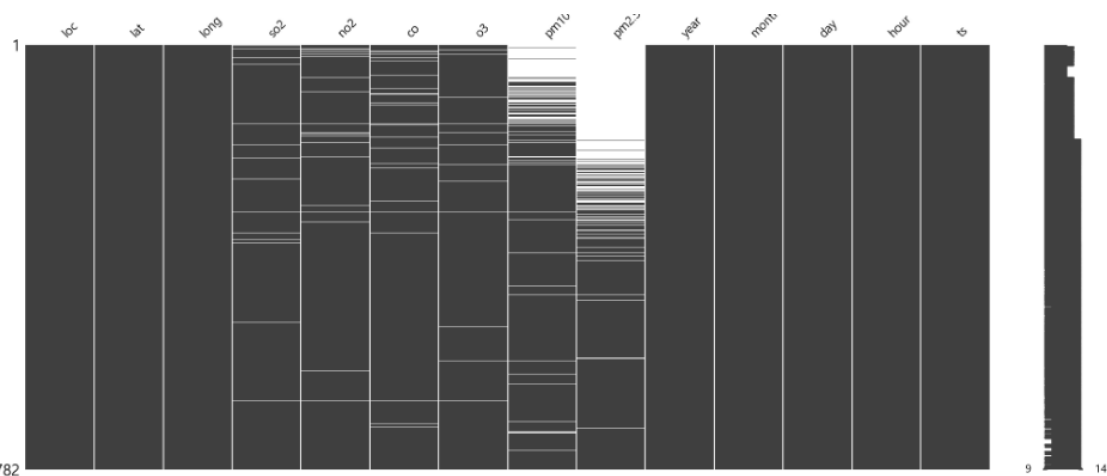
직경이 2.5 μ m이하인 먼지
자동차, 난방 및 에너지 사용 등으로 발생

PART 3

데이터 분석

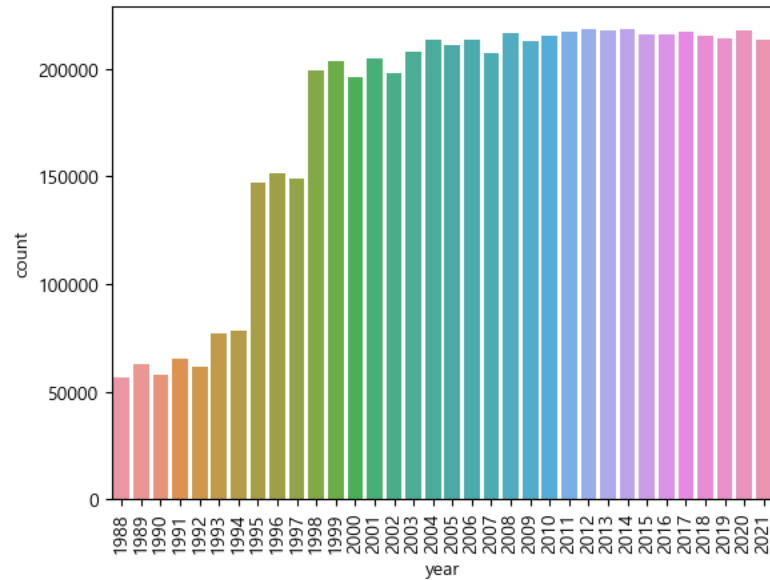


결측치 확인

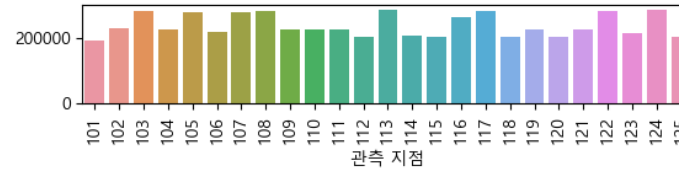


원인 파악

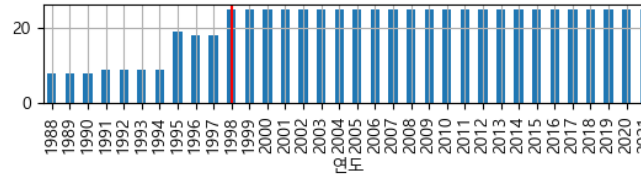
연도에 따른 전체 데이터 수 (개)



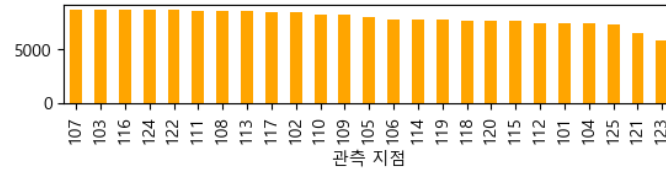
측정 장소에 따른 전체 데이터 수 (개)



연도별 관측 지점 수 (개)



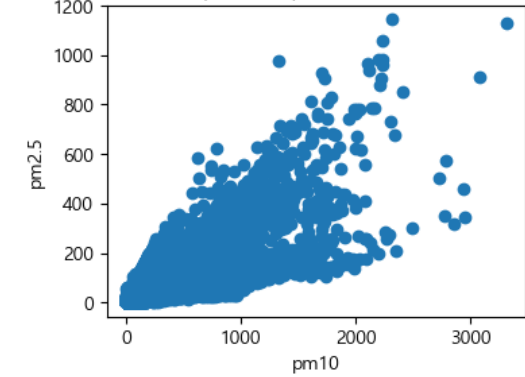
1998년의 관측 지점에 따른 데이터 수 (개)



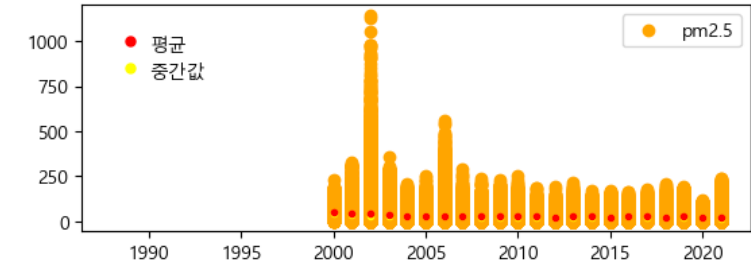
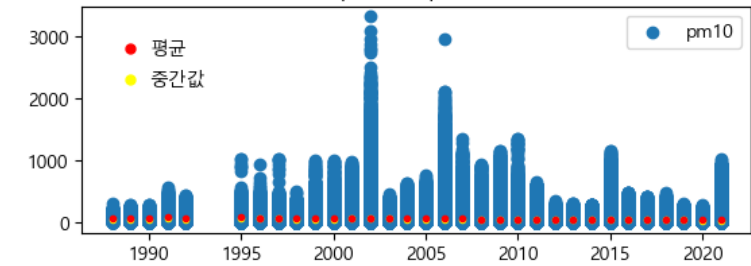
2021년의 관측 지점에 따른 데이터 수 (개)



pm10과 pm2.5의 관계



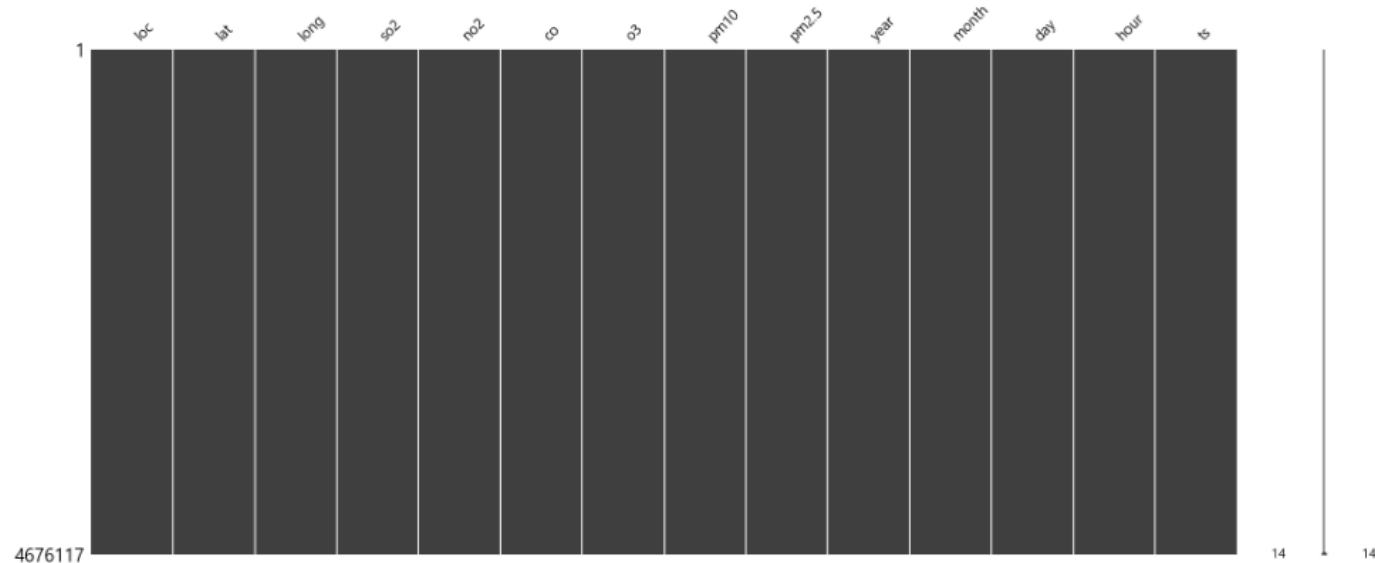
연도에 따른 pm10과 pm2.5의 농도 분포



전처리

so2 column별 평균값에 대한 최대 표준편차는 0.0010449034545253067입니다.
 no2 column별 평균값에 대한 최대 표준편차는 0.005828205250020761입니다.
 co column별 평균값에 대한 최대 표준편차는 0.13197734685423929입니다.
 o3 column별 평균값에 대한 최대 표준편차는 0.00849448443797602입니다.
 pm10 column별 평균값에 대한 최대 표준편차는 12.543551841219942입니다.
 pm2.5 column별 평균값에 대한 최대 표준편차는 6.808615519356849입니다.

<AxesSubplot:>



-> column별 결측치를 처리하기 위해 column별 가장 민감한 값을 가지는 범주형 column을 범주별 평균값에 대한 표준편차를 비교하는 방식으로 확인한다. 범주별 평균값간의 표준편차가 적은 데이터는 해당 column의 평균값으로, 표준편차가 어느정도 나타나는 데이터는 범주별 평균값으로 결측치를 대체한다.

데이터 일부만 사용

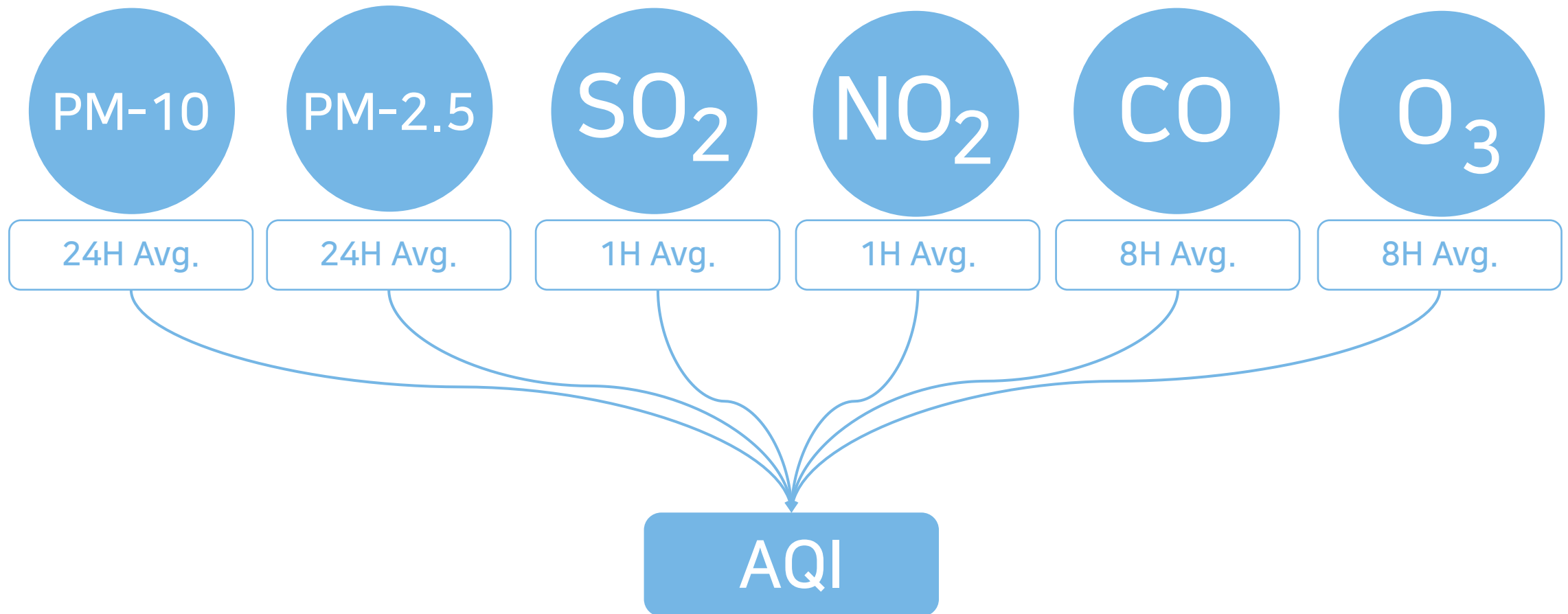
2001~2021 데이터만 사용

결측치 대체

범주형 Column별 평균값에 대한 표준편차 비교 후
해당 범주에 대한 평균값으로 대체

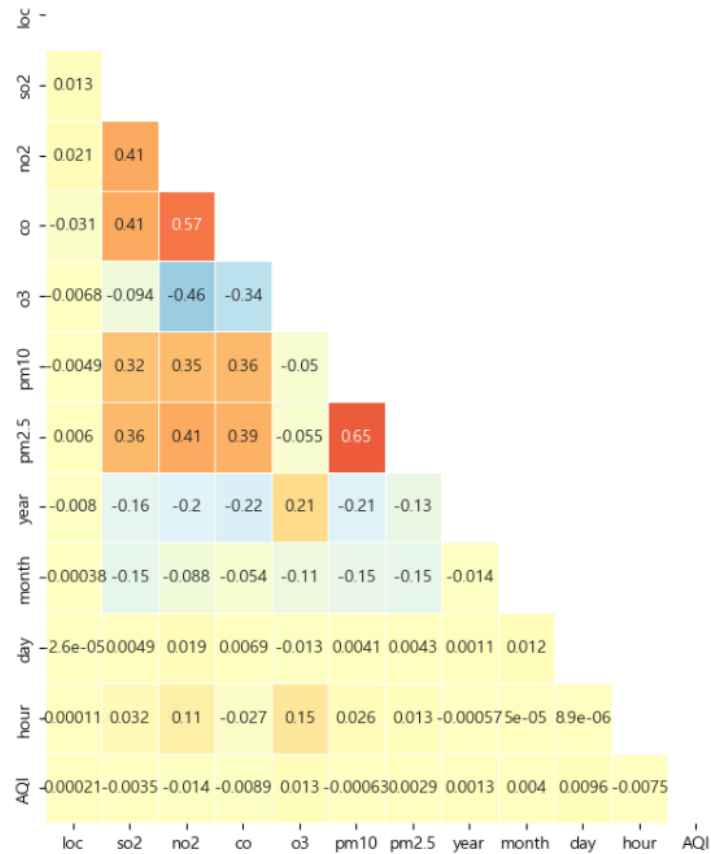
AQI Air Quality Index

대기환경지수



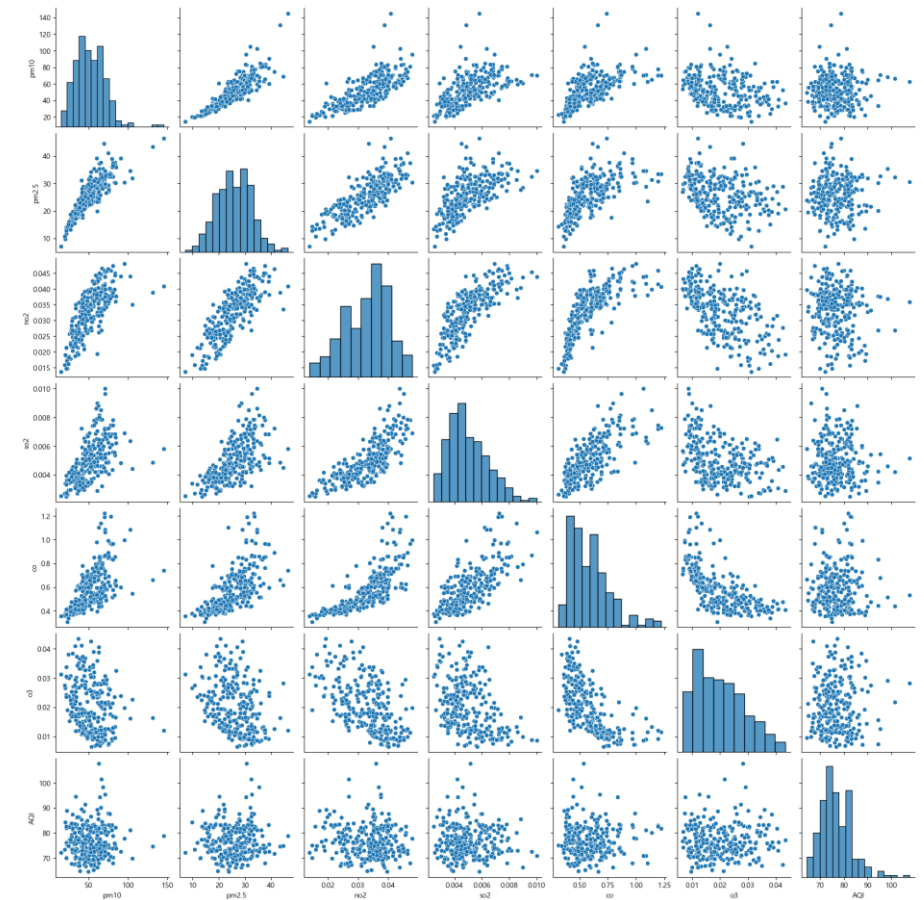
HeatMap, Scatter

< 2000년 이후 데이터의 상관관계에 따른 HeatMap >



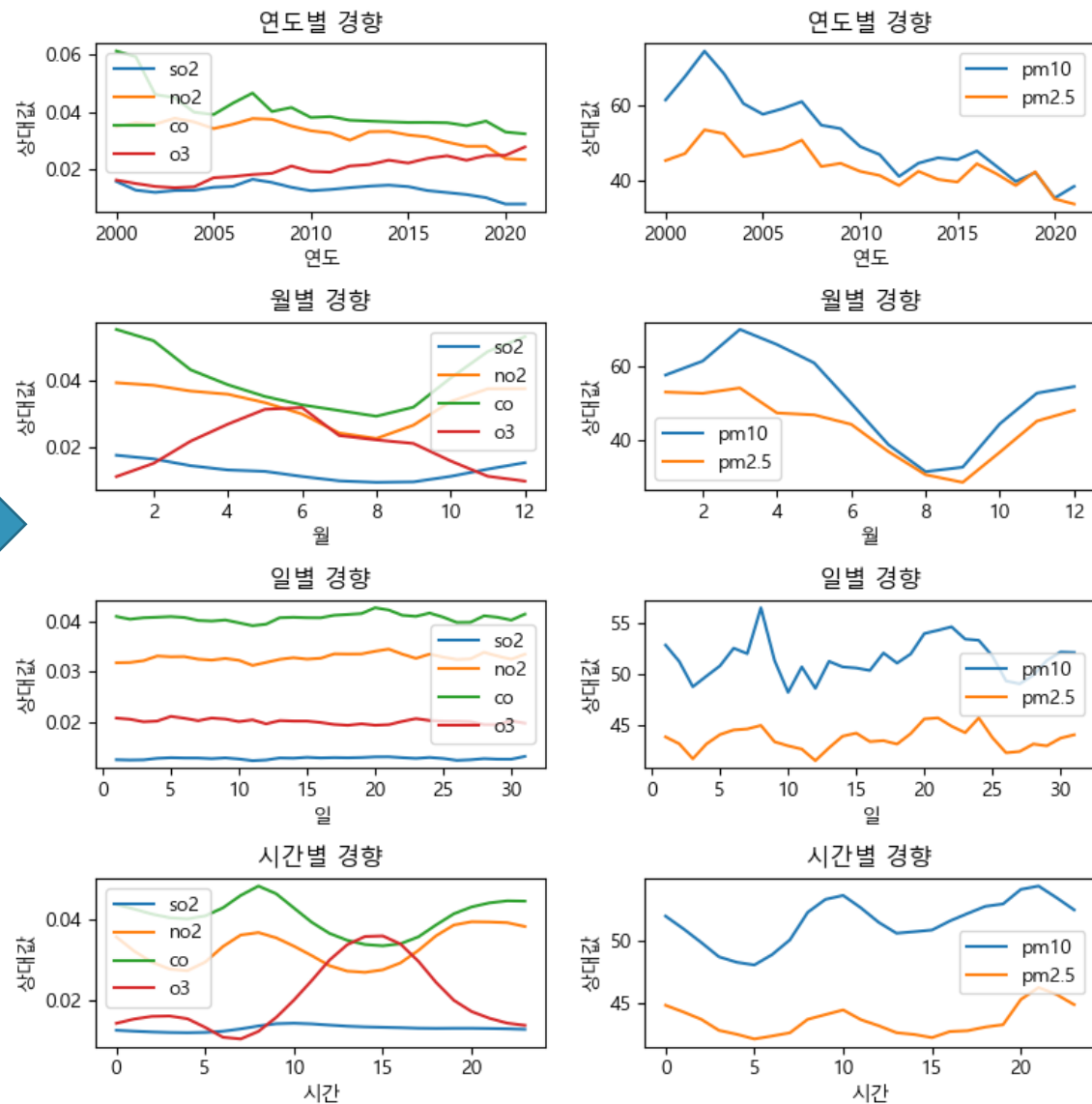
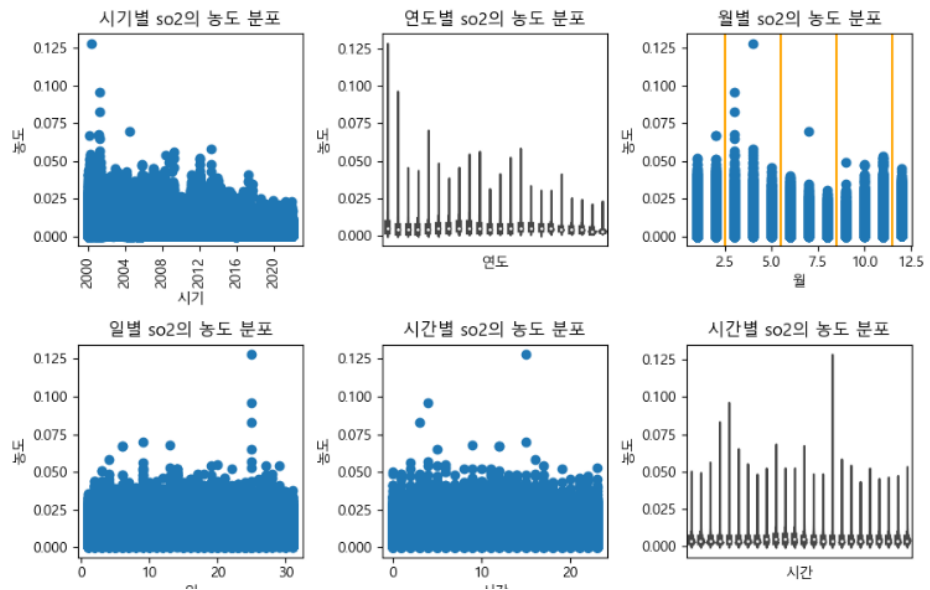
오염물질 간 상관관계 1st

오염물질과 시간 사이의 상관관계 2nd

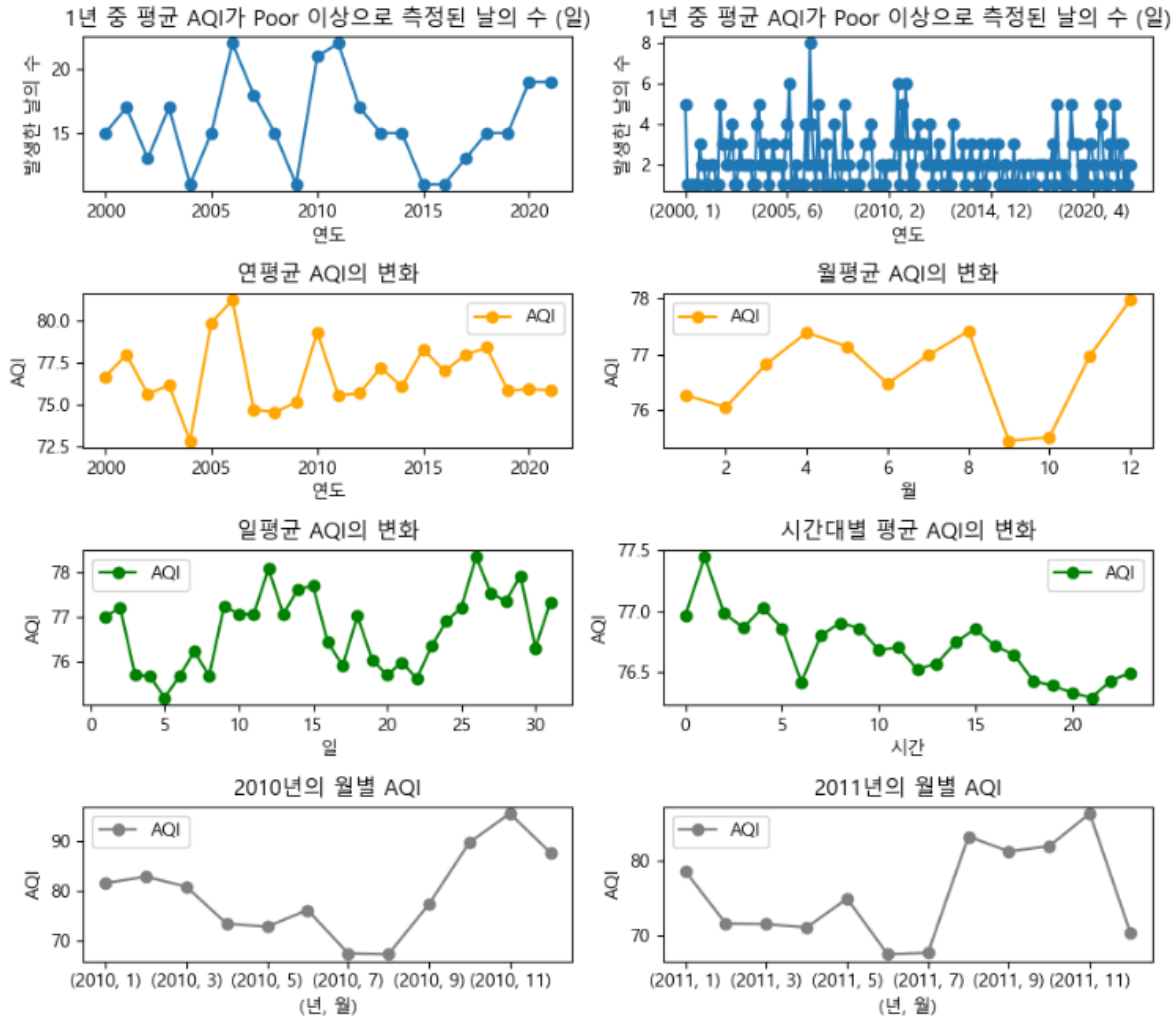


시계열 패턴 및 경향 분석

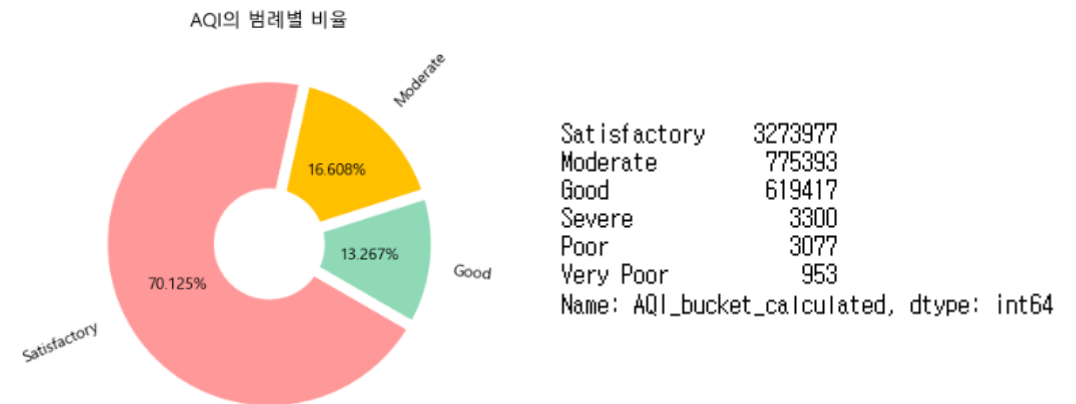
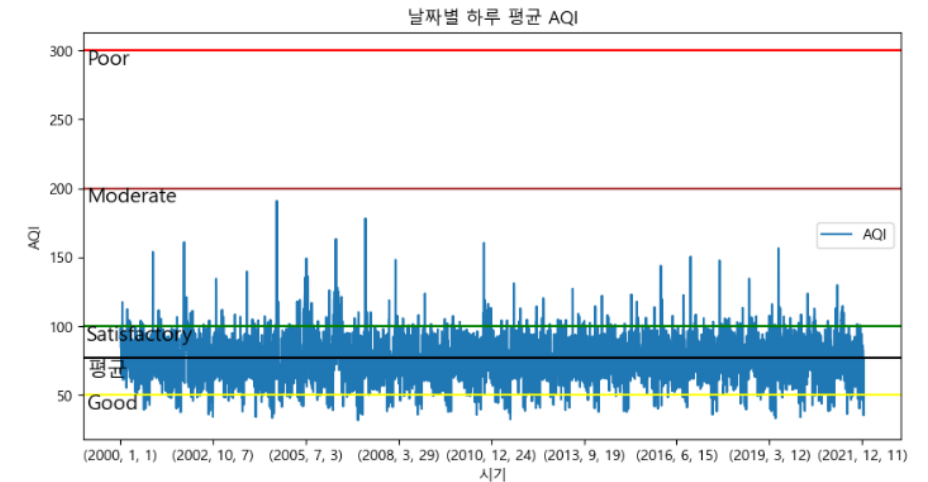
모든 오염물질들에 대해 분석



시계열 패턴 및 경향 분석

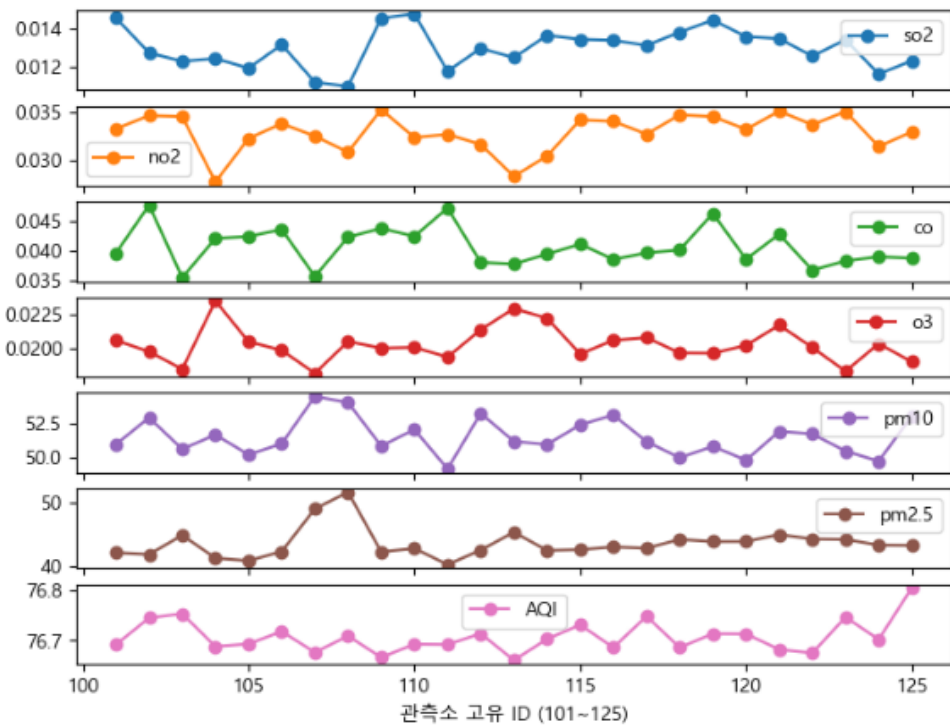


서울시의 공기는 대부분
Satisfactory

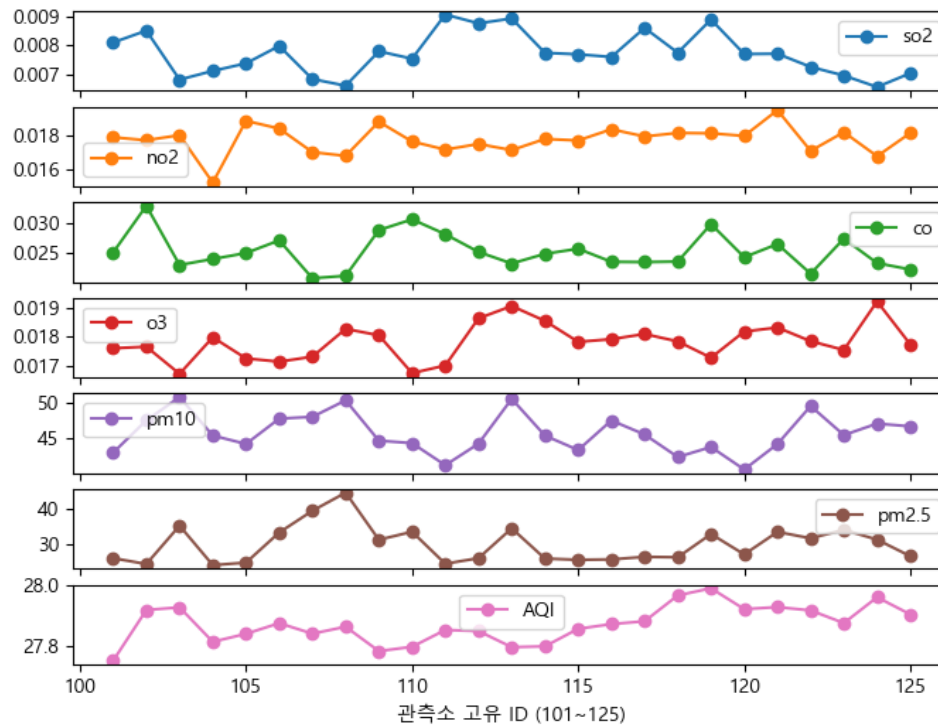


시계열 패턴 및 경향 분석

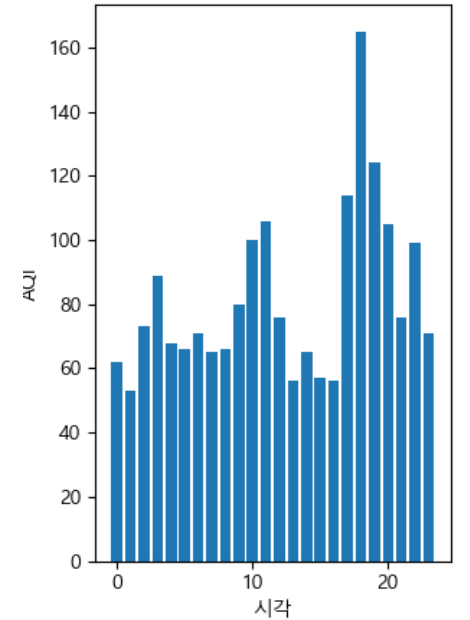
관측소별 관측값의 평균



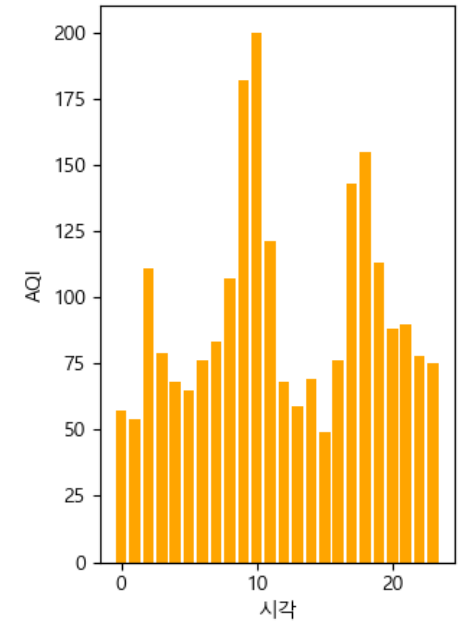
관측소별 관측값의 편차



2016년 4월 22일의 101번 관측소



2016년 4월 22일의 111번 관측소



PART 4

GUI 개발



GUI 개발에 사용된 라이브러리



유저와의 상호작용을 위한
Tkinter

Folium

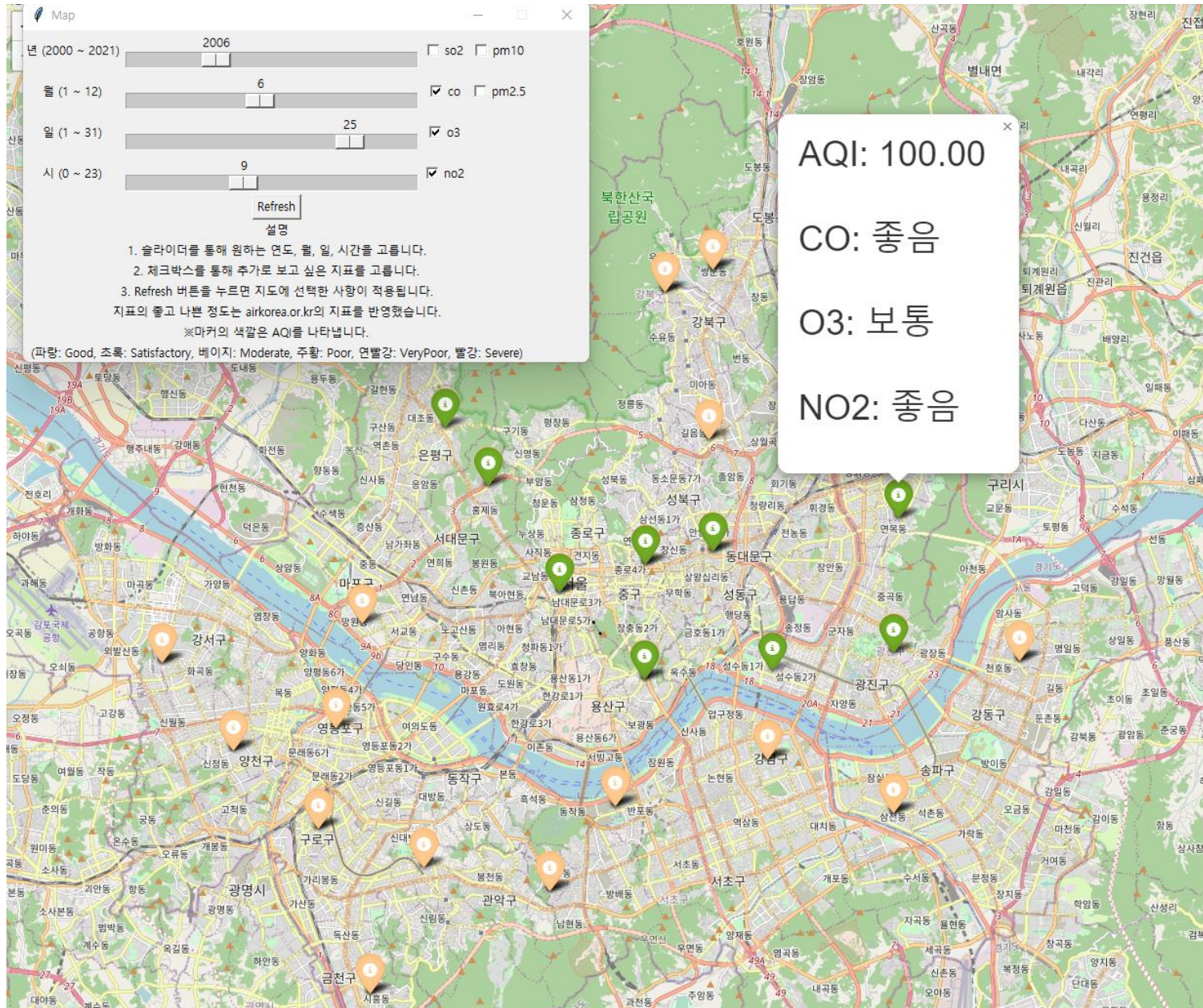


Map과 Marker을 구현해주는
Folium



Folium과 Tkinter를 연동하기 위한
Selenium

004 GUI 개발



Tkinter

- ▶ 슬라이더를 통해 유저에게 [년, 월, 일, 시간]을 입력 받는다.
- ▶ 체크박스를 통해 추가로 보여줄 데이터를 입력 받는다.

Folium

- ▶ 지도와 마커를 구현해서 HTML파일로 저장해준다.
- ▶ AQI 수치에 따라 마커의 색깔을 바꾸어 시각효과를 준다.
- ▶ 마커 클릭 시 Popup을 통해 추가적인 정보를 준다.

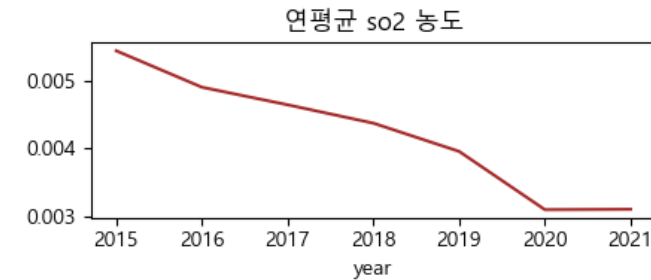
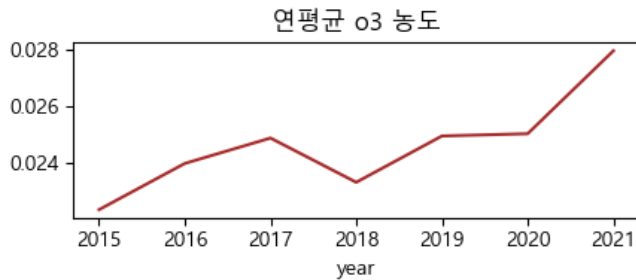
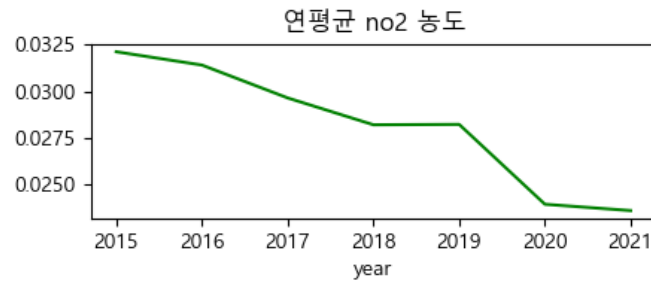
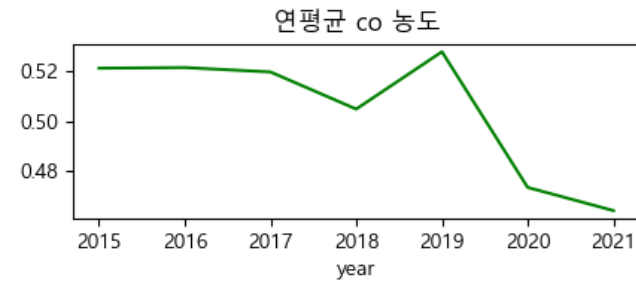
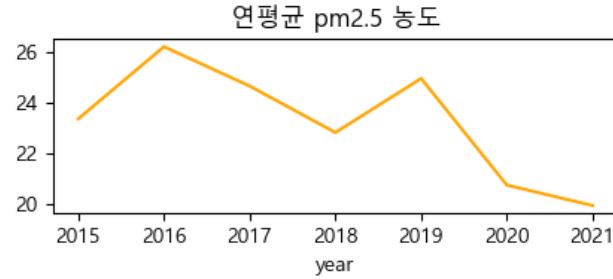
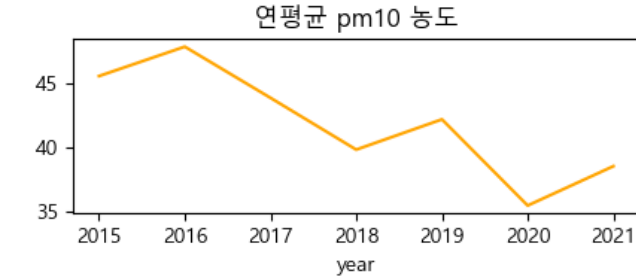
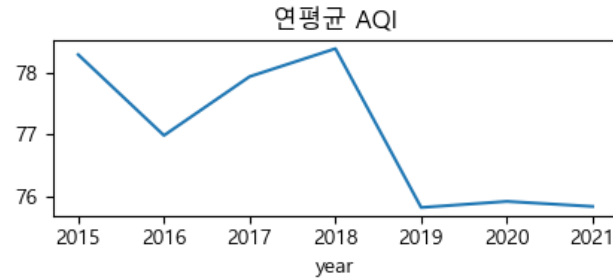
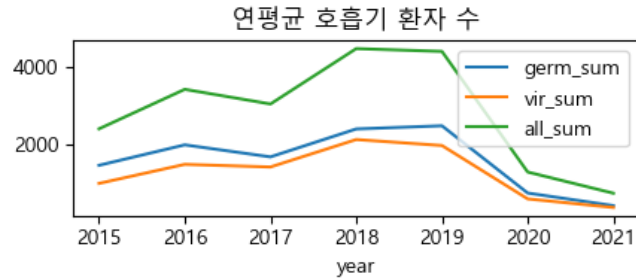
Selenium

- ▶ 저장된 HTML파일을 Chrome 으로 열고 새로고침 해준다.

PART 5

결론



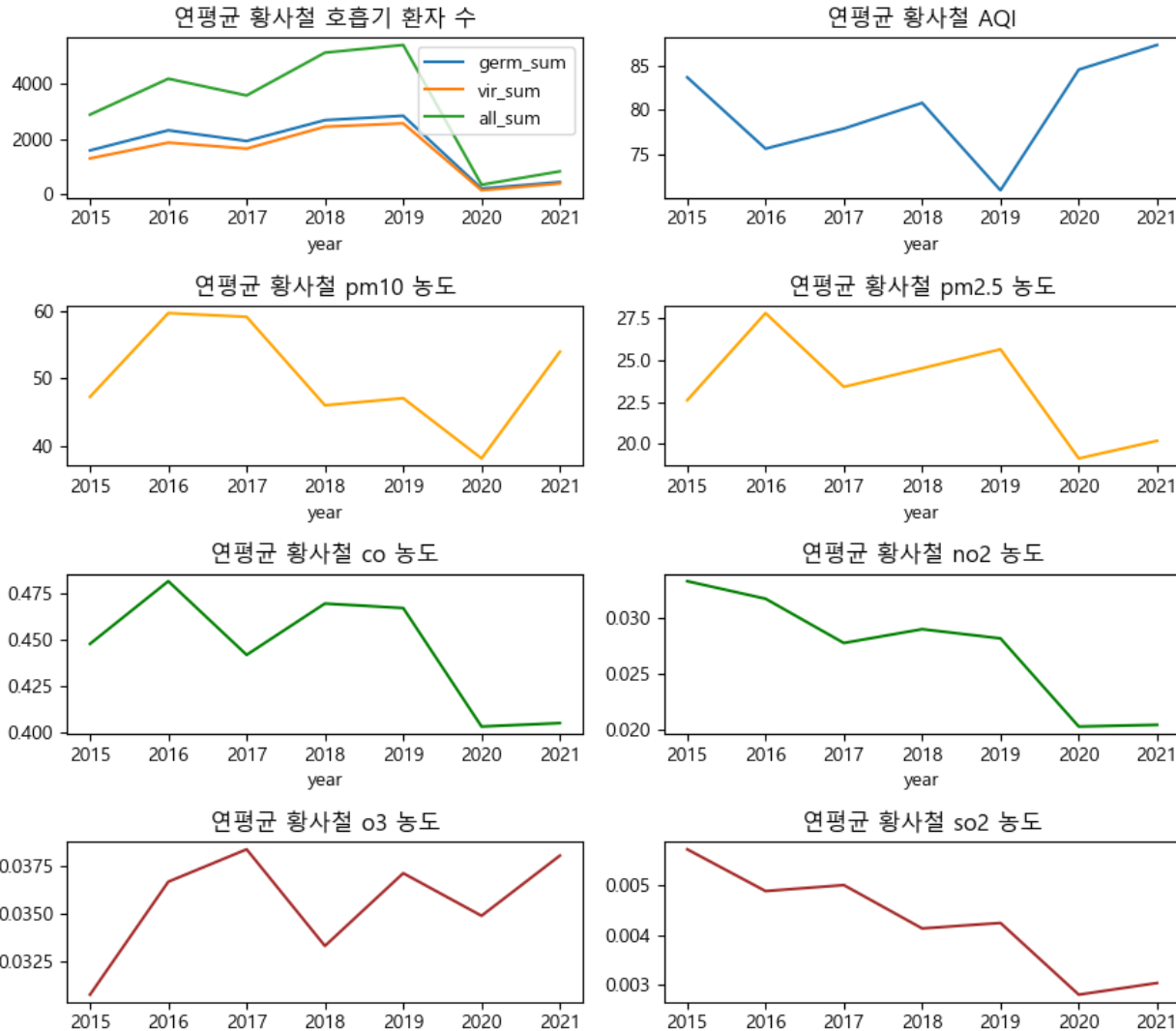


가설 검증

인과성 x

▶ 두 그래프 간에 유사성이
충분하지 않다.

005 결론



따라서

▶ 안좋은 공기는 시민들의
호흡기에 안좋은 영향을 미친다.

봄철 [15주~22주]에 한정하여 비교 시

연평균 pm10, pm2.5,
co 농도의 증감 방향



연평균 호흡기
환자 수의 증감 방향

가설

대기 오염이 많이 된 시기의 사람들은 오염된 공기로 인해 호흡기가 약해져, 호흡기 질환과 관련된 환자들이 증가할 것이며, 환경 요인만큼 시각 정보에 따른 특성이 존재할 것이다.



결론

높은 대기 중 오염물질 농도는 호흡기를 약하게 만들어 호흡기 환자의 수를 늘릴 것이라는 가설은 실제와 일치하며, 각 오염물질들은 시간 속성에 대해 일정한 패턴이나 경향성을 나타낸다.

END

Thank you