[0110]_EDA방법론_9기_박서연

과제 1. 우리 팀 EDA 데이터에 EDA 방법론 세션 내용 적용

I. 모든 데이터에서 공통적으로 해보면 좋을 일들

1) (데이터 전처리 단계) 이상치 탐색

기상 데이터에서의 이상치
- 관측기기의 오작동으로 잘못 측정된 값
- '기상기후 품질검사 알고리즘' 기준을 넘는 값

이상치는 데이터의 전체적인 패턴에서 동떨어져 있는 관측값으로, 전체 데이터에 비정상적인 영향을 미칠 수 있으므로 파악이 필요하다. 그렇다면 이상치를 어떻게 찾아낼까?

(1) 시각화: 상자그림(Box Plot), 산점도(Scatter Plot)
-> -1.5*IQR(InterQuartile Range, 사분범위) ~ 1.5*IQR 범위를 벗어나는 값 탐색

(2) 평균을 기준으로 탐색: 평균으로부터 너무 멀리 떨어져 있는 값,
구체적으로는 상/하위 2.5%에 해당하는 값 탐색

(3) 특정 기준을 바탕으로 탐색: 기상 데이터의 경우 '기상기후 품질검사 알고리즘'을 기준으로 비정상적인 값을 확인할 수 있다.

기상기후 품질검사 알고리즘 기준

ASOS						AWS		
요소	물리한계검사		요소	물리한계검사		요소	물리한계검사	
	상한	하한		상한	하한		상한	하한
기온(℃)	60	-80	해면기압(hPa)	1080	500	기온(℃)	45	-35
최고기온(℃)	60	-80	최고해면기압(hPa)	1080	500	풍향(°)	360	0
최저기온(℃)	60	-80	최저해면기압(hPa)	1080	500	풍속(m/s)	75	0
이슬점온도(℃)	60	-80	습도(%)	100	0	강수유무	10	0
일강수량(㎜)	1000	0	최소습도(%)	100	0	일강수량(㎜)	1500	0
강수량(㎜)	300	0	지면온도(℃)	80	-80	습도(%)	100	0
풍향(deg)	360	0	초상온도(℃)	50	-50	기압(hPa)	1080	500
최대풍속풍향(deg)	360	0	5cm 지중온도(℃)	50	-50			
최대순간풍속풍향(deg)	100	0	10cm 지중온도(℃)	50	-50			
풍속(㎧)	100	0	20cm 지중온도(℃)	50	-50			
최대풍속(㎧)	100	0	30cm 지중온도(℃)	50	-50			
최대순간풍속(㎧)	100	0	0.5m 지중온도(℃)	50	-50			
3시간 신적설(㎝)	200	0	1.0m 지중온도(℃)	50	-50			
일 신적설(㎝)	200	0	1.5m 지중온도(℃)	50	-50			
일 적설(㎝)	200	0	3.0m 지중온도(℃)	50	-50			
일 최심적설(㎝)	2500	0	5.0m 지중온도(℃)	50	-50			
일 최심신적설(cm)	2500	0	가조시간(hr)	15	0			
16반위 풍향	16	0	합계 대형증발량(㎝)	15	0			
현지기압(hPa)	1080	500	합계 소형증발량(㎝)	15	0			
최고현지기압(hPa)	1080	500	전운량(1/10)	10	0			
최저현지기압(hPa)	1080	500	중하층운량(1/10)	10	0			

기상청

그림 1 기상기후 품질검사 알고리즘 기준

-> 관측기기가 측정 가능한 물리적인 한계를 벗어나는 값이라면 이상치로 판단한다. 기준 상한을 벗어나는 값이 있다면 상한값으로, 하한을 벗어나는 값이 있다면 하한값으로 대체한다.

ex) 기온은 최고 60도, 최저 -80도까지 측정할 수 있는데 데이터에 -90도라는 값이 찍혀 있다면, 해당 값은 하한을 벗어나기 때문에 물리한계검사 하한값인 -80도로 대체

위의 알고리즘 기준표에는 우리의 원래 데이터셋에 포함된 대기오염물질들(이산화질소, 오존, 아황산가스, 일산화탄소, 미세먼지, 초미세먼지)에 대한 기준은 포함되어 있지 않으나, 혹시라도 기타 기상 데이터와 연관 지어 분석할 것에 대비하여 함께 조사해 보았다.

2) (데이터 전처리 단계) 결측치 탐색

기상 데이터에서의 결측치
- 미관측 되었거나 비어있는 값
- '결측'은 보통 정보가 존재하지 않으나, 가끔 정보를 추론할 수 있는 경우가 있음.
아래 표에서 Non Random에 해당하는 결측치의 경우 관측 signal에서 trend를 고려하는 등 얻어낼 수 있는 정보를 활용해 결측치 표현 가능

	결측치 종류
Random	Missing completely at random(MCAR)
	Missing at random(MAR)
Non Random	Non missing at random(NMAR)

표 1 결측치 종류

결측치가 포함된 모형은 편향적인 모델 구축에 영향을 주며, 잘못된 결론에 도달할 가능성이 존재하므로 처리해주는 과정이 필요하다. 그렇다면 결측치를 어떻게 처리할까?

결측치 처리 방법[1)]

결측치 비율	처리 방법
10% 미만	삭제 또는 대치
10~20%	Hot dect[2)], regression, model based imputation
20% 이상	Regression, model based imputation
50% 이상	결측율이 50% 이상인 자료라면 사용하지 않는다

표 2 결측치 비율에 따른 처리 방법

특히 Hot dect의 경우 기상 데이터, 특히 시간 평균과 같이 작은 단위의 데이터일 때 사용하기 좋은 결측치 처리 방법일 것 같다. 기상 데이터는 매년 매월 매일 매시 마다 꾸준히 측

1) Hari etal.(2006, pp.49-73)에 제시된 가이드라인 기반
2) Hot dect: 매년 조사해오던 자료에 대해서 올해 값이 결측이라면 작년 자료를 이용해서 채운다.

정되어 데이터가 쌓여 오고 있기 때문에 어느 일자의 데이터가 없더라도 바로 작년 데이터를 구해서 대체하기가 용이하리라 생각한다.

결측치 대체 방법
- 결측치가 포함된 열의 평균 또는 중앙값 활용
- Conditional imputation(조건부 대체): 다른 여러 조건들이 비슷한 값들을 추출해서 그 값들의 평균으로 대체
- Multiple imputation(가중 대체)

기상청의 관측 자료는 실시간으로 품질 관리가 이루어지므로 기기 오작동 및 점검 외에는 결측치가 발생할 이유가 없으나 관측 주기 혹은 기상 상황에 의해 값이 없는 경우 결측치를 제거하고 분석하는 것이 좋다.

3) (데이터 전처리 단계) 데이터 shape, type 파악

우리가 현재 다루고 있는 파일은 '.csv' 확장자의 비교적 간단한 형태이지만, 기상 분야 데이터 중에는 '.nc'라는 확장자의, netCDF라는 이름의 특이한 형태를 가진 것도 있다. 이러한 데이터는 보통 time(시간), lev(고도), lat(위도), lon(경도)로 이루어진 4차원의 복잡한 형태를 가지고 있기 때문에, 본격적으로 데이터를 다루기 전 shape를 파악해보는 것이 좋을 듯하다.

데이터의 타입 또한 미리 확인해보는 것이 좋다. 우리의 데이터셋(그중에서도 서울시 월별 평균 대기오염도 정보)을 가지고 나름의 EDA를 해보다가 그 필요성을 절감하게 되는 일이 발생했는데, 바로 해당 데이터셋에 포함된 컬럼 중 '측정월'의 데이터 타입과 관련해서였다.

데이터의 전체적인 추세를 파악하고자 시계열 그래프로 시각화를 하려고 했었는데, 측정월 컬럼을 그대로 x축으로 설정하자 202301을 2023년 1월의 날짜가 아닌 20만 2천 3백 1로 인식해서 축 정보가 엉망이 되는 일이 발생했다. 데이터 타입을 미리 파악했더라면 이러한 시행착오 없이 바로 datetime 모듈을 사용해서 해당 데이터 타입을 날짜 형식으로 바꿔주어야 한다는 것을 알 수 있었을 것이다.

4) (데이터 시각화 단계) 데이터 시각화

기상 데이터는 기본적으로 시계열 데이터의 속성을 띠고 있으므로, 가로축을 시간, 세로축을 우리가 보고자 하는 정보(우리가 사용하는 데이터셋의 경우 대기오염물질의 농도)로 하여 시계열 그래프를 그리면 추세를 파악하기에 용이할 것이다.

II. 데이터셋별로 해보면 좋을 일들

1) 서울시 연도별 평균 대기오염도 정보

데이터셋 설명: 1987(다만 지역별로 측정 시작 연도가 다름)년부터 2023년까지 서울시 각 측정소별 대기오염도를 '연평균' 내어 보여주고 있다.

연도별 평균 데이터의 특징
- 과거부터 현재까지의 전체적인 추세를 파악하기에 좋다
- 좀 더 구체적으로는, 데이터의 전반적인 추세를 사회적인 현상(각 대기오염물질에 대한 정부 규제 발생, 중국의 올림픽 개최 및 코로나로 인한 공장 가동 중지 및 그로 인한 한반도 미세먼지, 초미세먼지 유입 감소)과 결부시켜 분석하기에 좋다.

```
### 이산화질소 농도 내림차순 정렬

ATM_yearly_df.sort_values(by='이산화질소농도(ppm)', ascending=False, inplace=False).head()
```
[22]

	측정년도	측정소명	이산화질소농도(ppm)	오존농도(ppm)	일산화탄소농도(ppm)	아황산가스(ppm)	미세먼지(㎍/㎥)	초미세먼지(㎍/㎥)
1174	1992	천호대로	0.105	0.014	1.4	0.019	NaN	NaN
1088	1996	신촌로	0.087	0.009	2.2	0.017	NaN	NaN
1096	1996	청계천로	0.084	0.002	2.2	0.018	NaN	NaN
1029	1998	신촌로	0.081	0.008	1.7	0.009	NaN	NaN
1068	1997	청계천로	0.080	NaN	1.7	NaN	NaN	NaN

```
### 오존 농도 내림차순 정렬

ATM_yearly_df.sort_values(by='오존농도(ppm)', ascending=False, inplace=False).head()
```
[23]

	측정년도	측정소명	이산화질소농도(ppm)	오존농도(ppm)	일산화탄소농도(ppm)	아황산가스(ppm)	미세먼지(㎍/㎥)	초미세먼지(㎍/㎥)
1041	1998	홍릉로	0.063	0.088	1.5	0.006	NaN	NaN
108	2021	관악산	0.009	0.047	0.3	0.003	34.0	16.0
58	2022	관악산	0.009	0.045	0.4	0.003	29.0	16.0
462	2013	관악산	0.018	0.040	NaN	0.006	34.0	23.0
137	2021	자연사박물관	0.020	0.037	0.6	0.004	36.0	19.0

위와 같은 방식으로 sort_values를 이용해서 이산화질소, 오존, 일산화탄소, 아황산가스, 미세먼지, 초미세먼지 각각에 대한 내림차순 정렬을 해보았고, 이를 통해 특정 대기오염물질이 언제, 어느 지역의 측정소에서 최고치를 나타냈는지, 그리고 높은 순위를 기록한 행들 사이에 공통점이 있는지 등을 보려고 시도했다.

2) 서울시 월별 평균 대기오염도 정보

데이터셋 설명: 1987(다만 지역별로 측정 시작 연도가 다름)년부터 2023년까지 서울시 각 측정소별 대기오염도를 '월평균' 내어 보여주고 있다.

월별 평균 데이터의 특징
- 연도별 데이터와 달리 월별 데이터는 1년 중 월별로 어떠한 변화가 나타나는지 볼 수 있다.
- 특히 우리의 데이터셋의 경우 대기오염물질의 '계절성'을 파악해볼 수 있다.

```
[63]
### 이산화질소 농도 내림차순 정렬

seodaemun_monthly_22.sort_values(by='이산화질소농도(ppm)', ascending=False, inplace=False).head()

# 이산화질소 농도는 주로 겨울에 높은 것으로 보인다
```

	측정월	측정소명	이산화질소농도(ppm)	오존농도(ppm)	일산화탄소농도(ppm)	아황산가스(ppm)	미세먼지(㎍/㎥)	초미세먼지(㎍/㎥)
622	202201	서대문구	0.031	0.016	0.8	0.005	42.0	28.0
72	202212	서대문구	0.026	0.016	0.7	0.004	36.0	17.0
572	202202	서대문구	0.025	0.026	0.7	0.004	39.0	26.0
122	202211	서대문구	0.023	0.020	0.8	0.004	37.0	21.0
522	202203	서대문구	0.022	0.032	0.7	0.003	42.0	21.0

```
[64]
### 오존 농도 내림차순 정렬

seodaemun_monthly_22.sort_values(by='오존농도(ppm)', ascending=False, inplace=False).head()

# 오존 농도는 주로 봄, 여름에 높은 것으로 보인다
```

	측정월	측정소명	이산화질소농도(ppm)	오존농도(ppm)	일산화탄소농도(ppm)	아황산가스(ppm)	미세먼지(㎍/㎥)	초미세먼지(㎍/㎥)
422	202205	서대문구	0.013	0.052	0.4	0.004	33.0	15.0
472	202204	서대문구	0.017	0.047	0.5	0.003	44.0	18.0
322	202207	서대문구	0.011	0.037	0.4	0.002	22.0	12.0
372	202206	서대문구	0.010	0.035	0.4	0.003	20.0	10.0
522	202203	서대문구	0.022	0.032	0.7	0.003	42.0	21.0

이번에도 아까와 같은 방식으로 sort_values를 이용해서 이산화질소, 오존, 일산화탄소, 아황산가스, 미세먼지, 초미세먼지 각각에 대한 내림차순 정렬을 해보았고, 이를 통해 특정 대기오염물질이 언제, 어느 지역의 측정소에서 최고치를 나타냈는지를 확인해보았다. 이번에는 특히 계절성 파악을 위해 '몇월'에 주로 최고치를 달성했는지를 위주로 살펴보았다.

또한 x축을 시간으로 두어 시계열 그래프를 그려서 전 측정 범위 동안의 전체적인 추세 및 2020~2022 최근 3년 동안의 월별 추세를 파악해보려고 했다.

**** EX. 추가 데이터와 엮어서 분석 1**

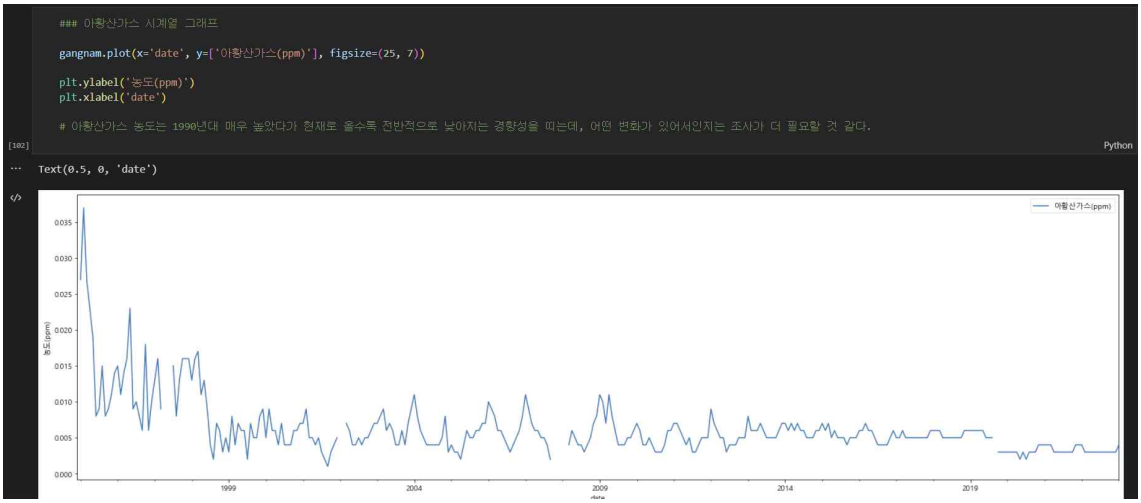

그림 6 아황산가스 시계열 그래프

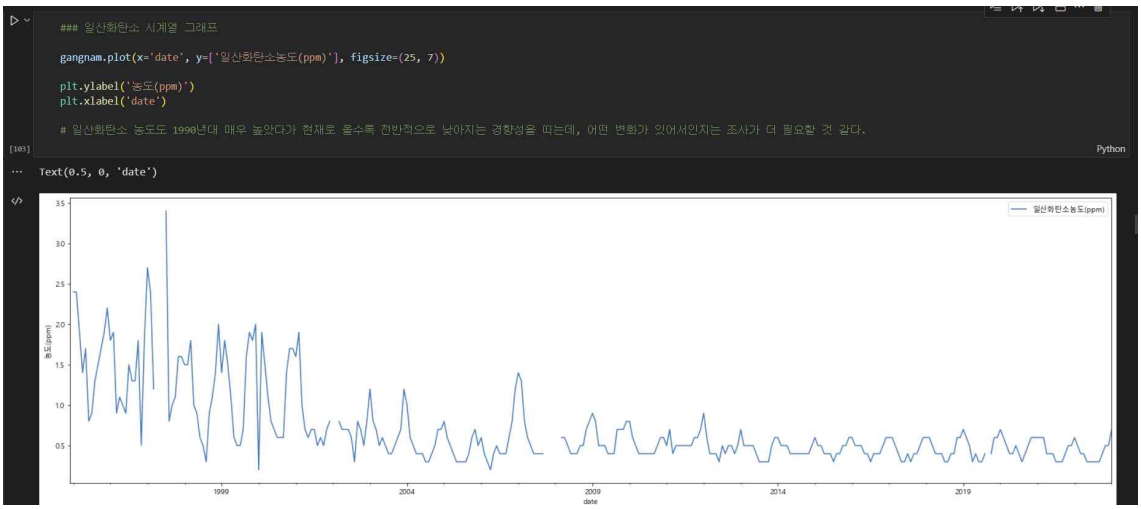

그림 7 일산화탄소 시계열 그래프

아황산가스, 일산화탄소 시계열 그래프를 그려본 결과 1990년대에 매우 높다가 현재로 올수록 전반적으로 급격하게 낮아지는 경향성을 띠었다. 이렇게 급격한 변화를 발생시킬 수 있는 유력한 요인은 '정부 규제'라고 생각하였고, 이에 대해서 찾아본 결과 실제로 아래 표와 같이 1990년대~2000년대 초반에 해당 대기오염물질들에 대한 대기환경기준이 크게 강화된 것을 확인할 수 있었다.[3)]

---

3) 대기환경기준의 체계변경 및 강화내역 표 출처:
대기환경연보(2020) 2021 : Annual Report of Air Quality in Korea, 2020 / 환경부 ; 대기미래전략과 ; 국립환경과학원 ; 기후대기연구부 ; 대기환경연구과

환경부 디지털도서관(https://library.me.go.kr/#/)에 위의 대기환경연보 외에도 환경통계연감, 환경백서, 대기환경월보 등 우리가 분석하고자 하는 데이터셋의 대기오염물질들을 포함한 여러 환경 요인들에 대한 분석을 담은 자료들이 많이 있으니 확인해보면 좋을 것 같다.

우리의 데이터셋에는 80년대 데이터가 많이 포함되어 있지 않아 직접 확인하기는 어려우나, 실제로 『2020 서울 대기질 평가 보고서[4]』에 따르면 아황산가스와 일산화탄소의 경우 에너지 전환정책 등의 영향으로 80년대 말 이후 농도가 빠르게 감소하는 추세였다. 1990년대 대기환경기준 강화와 더불어 큰 폭으로 줄었으며, 2000년대 이후에도 점차 감소하여 일산화탄소는 0.5ppm 정도, 아황산가스의 경우 0.005ppm 정도의 연평균 농도를 유지하고 있다.

<표 1-1> 대기환경기준의 체계변경 및 강화내역

항 목	'78	'83	'91	'93주3	'01	'07	'12	'18
아황산가스 (ppm)	0.05/년 0.15/일	0.05/년 0.15/일	0.05/년 0.15/일	0.03/년 0.14/일 0.25/시간	0.02/년 0.05/일 0.15/시간	0.02/년 0.05/일 0.15/시간	0.02/년 0.05/일 0.15/시간	0.02/년 0.05/일 0.15/시간
일산화탄소 (ppm)	-	8/월 20/8시간	8/월 20/8시간	9/8시간 25/시간	9/8시간 25/시간	9/8시간 25/시간	9/8시간 25/시간	9/8시간 25/시간
이산화질소 (ppm)	-	0.05/년 0.15/시간	0.05/년 0.15/일	0.05/년 0.08/일 0.15/시간	0.05/년 0.08/일 0.15/시간	0.03/년 0.06/일 0.1/시간	0.03/년 0.06/일 0.1/시간	0.03/년 0.06/일 0.1/시간

그림 8 대기환경기준의 체계변경 및 강화내역

< 월별 평균 대기오염도 정보를 통해 파악한 대기오염물질의 계절성 >

- 이산화질소: 주로 겨울에 높은 것으로 보인다
- 오존: 주로 봄, 여름에 높은 것으로 보인다
- 일산화탄소: 주로 겨울에 높은 것으로 보인다
- 아황산가스: 주로 겨울에 높은 것으로 보인다
- 미세먼지: 주로 봄에 높은 것으로 보인다
- 초미세먼지: 주로 겨울, 봄에 높은 것으로 보인다

3) 서울시 시간 평균 대기오염도 정보

데이터셋 설명: 2023년 1월 1일부터 1월 8일까지 서울시 각 측정소별 대기오염도를 1시간 단위로 '시간 평균' 내어 보여주고 있다.

시간 평균 데이터의 특징
- 하루 중 시간별로 어떠한 변화가 나타나는지 볼 수 있다.
- 특히 우리의 데이터셋의 경우 대기오염물질의 '일변화'를 파악해볼 수 있다.

4) 출처: https://news.seoul.go.kr/env/files/2022/02/620cb1eadc4908.19590134.pdf

```
[77]
### 이산화질소 농도 내림차순 정렬

seodaemun_hourly_22.sort_values(by='이산화질소농도(ppm)', ascending=False, inplace=False).head()

# 이산화질소 농도는 주로 밤, 새벽 시간대에 높은 것으로 보인다
```

	측정일시	측정소명	이산화질소농도(ppm)	오존농도(ppm)	일산화탄소농도(ppm)	아황산가스(ppm)	미세먼지(㎍/㎥)	초미세먼지(㎍/㎥)
588	202301072100	서대문구	0.049	0.005	1.8	0.006	120.0	100.0
1029	202301070300	서대문구	0.047	0.004	1.3	0.004	89.0	84.0
601	202301072000	서대문구	0.045	0.010	1.7	0.006	122.0	92.0
561	202301072200	서대문구	0.043	0.011	1.5	0.006	108.0	87.0
1068	202301070200	서대문구	0.042	0.008	1.2	0.004	81.0	77.0

```
[78]
### 오존 농도 내림차순 정렬

seodaemun_hourly_22.sort_values(by='오존농도(ppm)', ascending=False, inplace=False).head()

# 오존 농도는 주로 낮, 오후 시간대에 높은 것으로 보인다
```

	측정일시	측정소명	이산화질소농도(ppm)	오존농도(ppm)	일산화탄소농도(ppm)	아황산가스(ppm)	미세먼지(㎍/㎥)	초미세먼지(㎍/㎥)
743	202301071500	서대문구	0.018	0.037	1.1	0.005	153.0	104.0
704	202301071600	서대문구	0.020	0.036	1.2	0.005	137.0	104.0
676	202301071700	서대문구	0.023	0.034	1.2	0.005	127.0	104.0
755	202301071400	서대문구	0.018	0.032	1.1	0.005	158.0	104.0
653	202301071800	서대문구	0.029	0.030	1.2	0.004	128.0	98.0

이번에도 아까와 같은 방식으로 sort_values를 이용해서 이산화질소, 오존, 일산화탄소, 아황산가스, 미세먼지, 초미세먼지 각각에 대한 내림차순 정렬을 해보았고, 이를 통해 특정 대기오염물질이 어느 시간대에서 최고치를 나타냈는지를 확인해보았다. 이번에는 특히 일변화 파악을 위해 '몇 시'에 주로 최고치를 달성했는지를 위주로 살펴보았다.

< 시간 평균 대기오염도 정보를 통해 파악한 대기오염물질의 일변화 경향성 >

- 이산화질소: 주로 밤, 새벽 시간대에 높은 것으로 보인다
- 오존: 주로 낮, 오후 시간대에 높은 것으로 보인다
- 일산화탄소: 주로 밤 시간대에 높은 것으로 보인다
- 아황산가스: 주로 밤 시간대에 높은 것으로 보인다
- 미세먼지: 주로 낮 시간대에 높은 것으로 보인다
- 초미세먼지: 주로 아침 및 낮 시간대에 높은 것으로 보인다