

2024 기계학습 Team Project

노원구 대기오염 예측

TEAM 굿머닝

Hello / Bonjour / Ohayo / Aloha / Hola
김서경 / 맹도현 / 문혜진 / 백지우 / 최수현

(click) 굿머닝의 NOTION



01

주제 소개

연구 배경 및 목적을 소개합니다

02

데이터 설명

사용한 데이터에 대해
설명합니다.

03

데이터 전처리

연구 진행에 앞서 진행한
데이터 전처리를 소개합니다.

04

알고리즘1_회귀

Multiple Linear Regression,
RandomForest, Lidge,
Lasso, CatBoost

05

알고리즘2_분류

RandomForest, AdaBoost,
Gradient Boosting,
XGBoost, LightGBM,
CatBoost

06

평가 및 해석

알고리즘을 종합적으로 평가하
고 해석합니다.

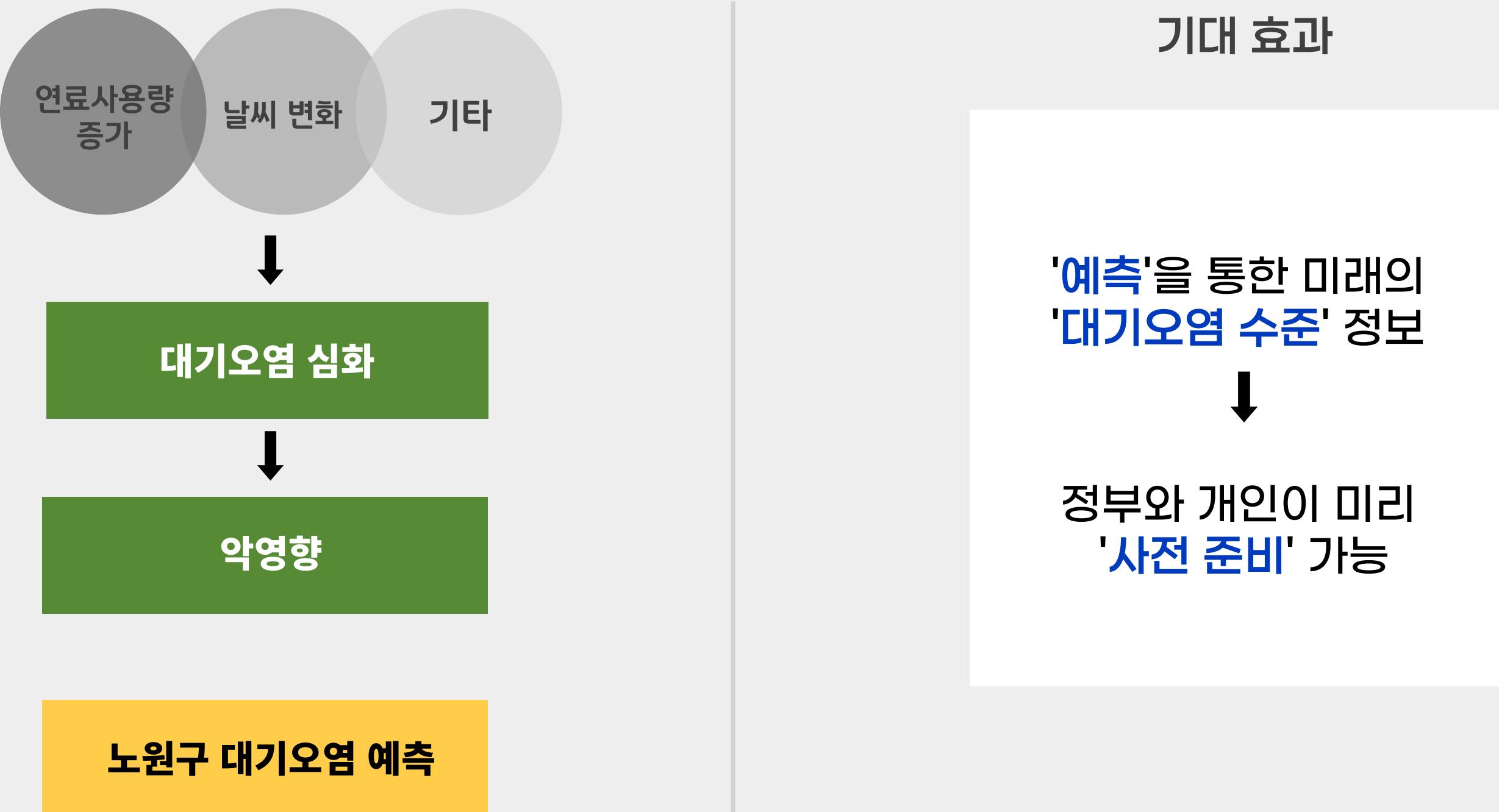
07

결론

이번 프로젝트를 마무리하며...

01 주제 소개

주제 선정 배경





① 대기오염 데이터

에어코리아 '한국 대기 오염 데이터'
(국립환경과학원의 최종확정자료)

SO2

CO

O3

NO2

PM10

PM2.5

② 날씨 데이터

기상청 자료개방포털 '날씨데이터'
(총관기상관측(ASOS))

③ 인구 특성 데이터

주민등록 인구통계 행정안전부
'인구 특성 데이터'

④ 녹지 데이터

서울 열린 데이터 광장 사이트
'녹지현황 통계'

⑤ 자동차 데이터

국토교통 통계누리 사이트
'자동차등록현황보고 데이터'

기온(°C), 강수량(mm), 풍속(m/s), 풍향(16방위), 풍향습도(%), 증기압(hPa), 이슬점온도(°C),
현지기압(hPa), 해면기압(hPa), 일조(hr), 일사(MJ/m²), 전운량(10분위), 중하층운량(10분위), 시정(10m),
지면온도(°C)5cm, 지중온도(°C)10cm, 지중온도(°C)20cm, 지중온도(°C)30cm 지중온도(°C)

인구 수

시설녹지(개소), 일반녹지(개소), 분리대(개소), 수벽(개소), 수림대(개소), 하천변조경(개소), 간이휴게소(개소),
지하철환기구주변(개소), 건물주변(개소), 아파트 및 학교(개소), 침수공간조성(개소), 기타(개소), 총합계(면적),
시설녹지(면적), 일반녹지(면적), 분리대(면적), 수벽(면적), 수림대(면적), 하천변조경(면적), 간이휴게소(면적),
지하철환기구주변(면적), 건물주변(면적), 아파트 및 학교(면적), 침수공간조성(면적), 기타(면적)

휘발유경유, LPG전기, CNG하이브리드(휘발유+전기), 하이브리드(경유+전기), 하이브리드(LPG+전기),
하이브리드(CNG+전기), 수소기타연료

03 데이터 전처리

① 대기오염 데이터

Target 변수

SO2	CO	O3
NO2	PM10	PM2.5

$$I_p = \frac{I_{HI} - I_{LO}}{BP_{HI} - BP_{LO}} \times (C_p - BP_{LO}) + I_{LO}$$

[별첨0] 참고

파생 변수 생성

오염도

수치형
(6가지 요인을 반영한
통합 오염정도)

CAI

연속형 오염지수

파생 변수 생성

오염상태

범주형
(좋음 / 보통 /
나쁨 / 매우나쁨)

Air_Quality_Status

좋음, 보통,
나쁨, 매우나쁨

Label
인코딩

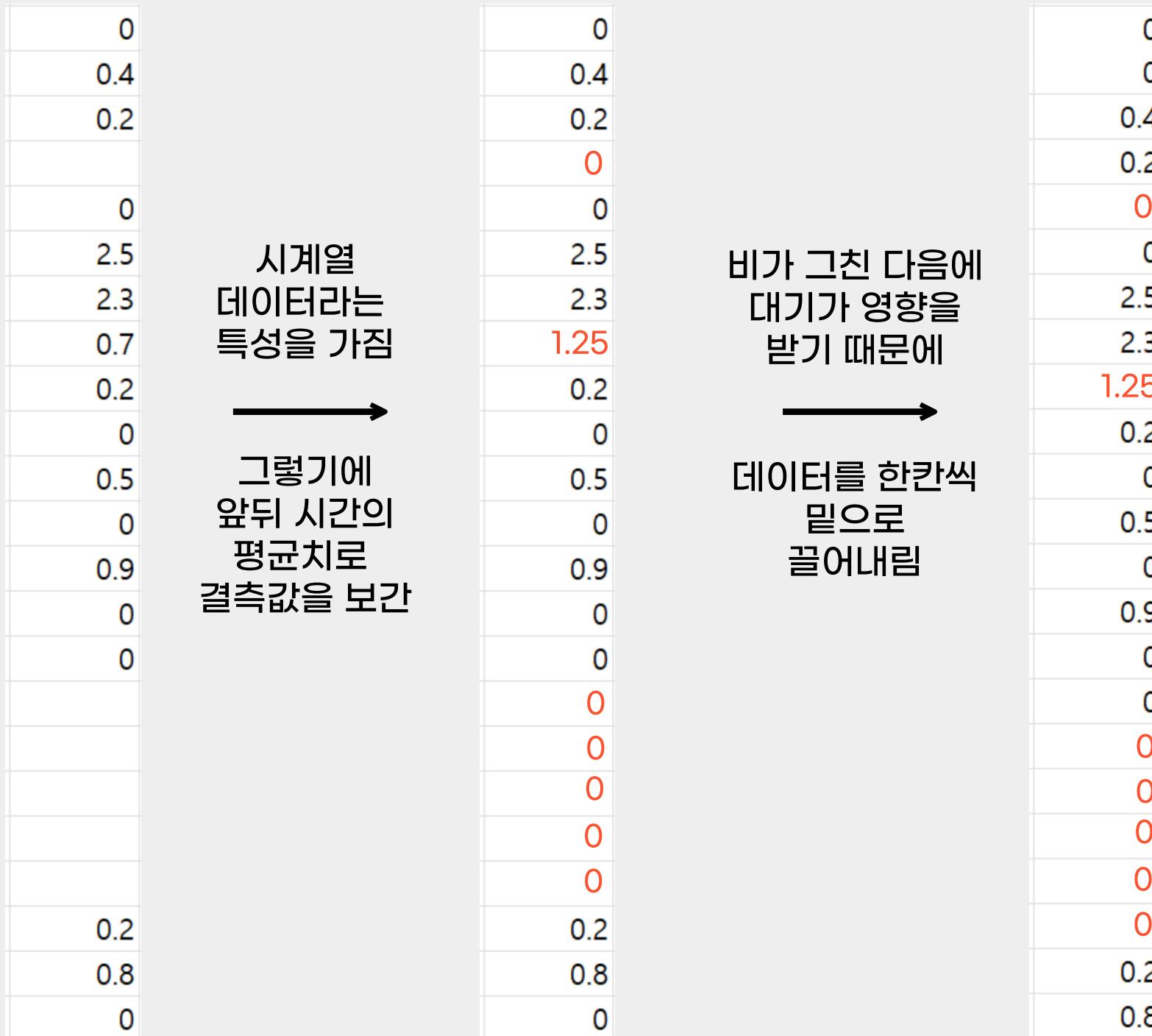
Air_Quality_Status_Label

{ 좋음:0, 보통:1, 나쁨:2, 매우나쁨:3 }

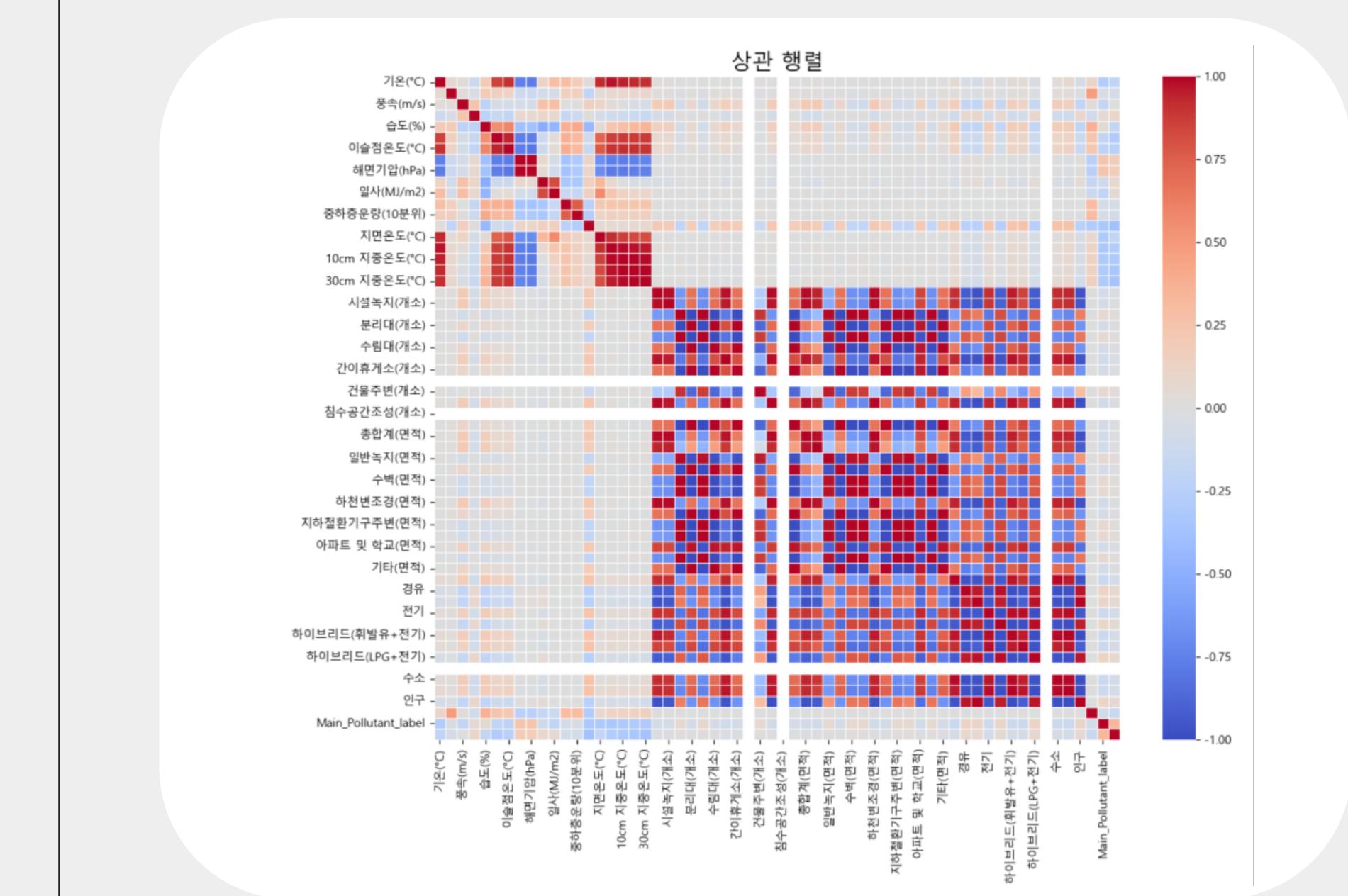
	좋음	보통	나쁨	매우나쁨
I _{LO}	0	51	101	251
I _{HI}	50	100	250	500

[별첨0] 참고

강수량 결측값 처리 및 시계열 조정

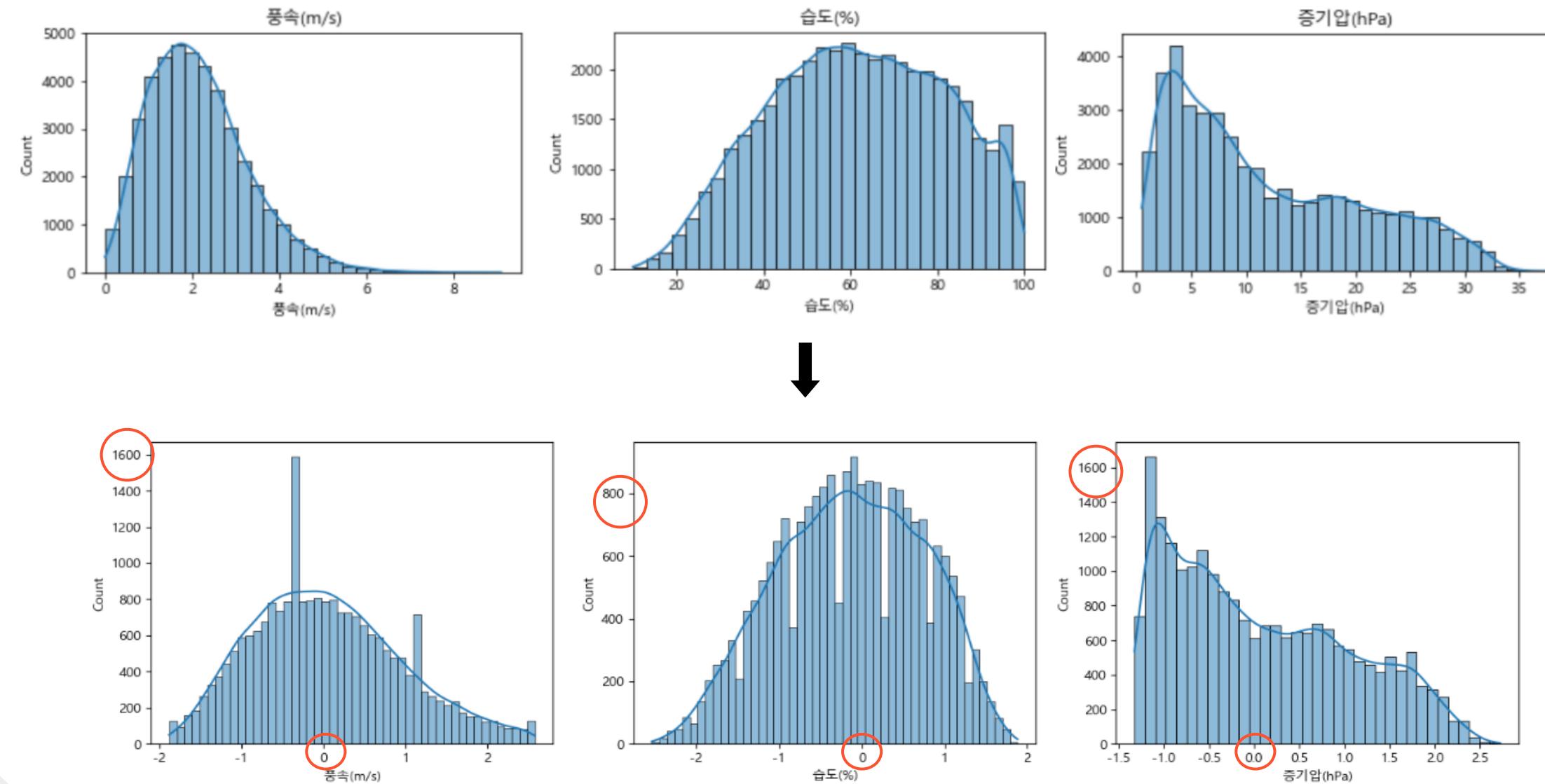


EDA를 통해 데이터 분포 확인



전처리가 필요한 feature를 찾기 위해
Correlation [별첨1], Histogram [별첨2],
Boxplot [별첨3]을 통한 데이터 EDA 진행

앞의 EDA를 통해 (연속형 변수+이상치의 위험이 있음)의 column들에 대해 standardScaler() 진행 별첨[4]



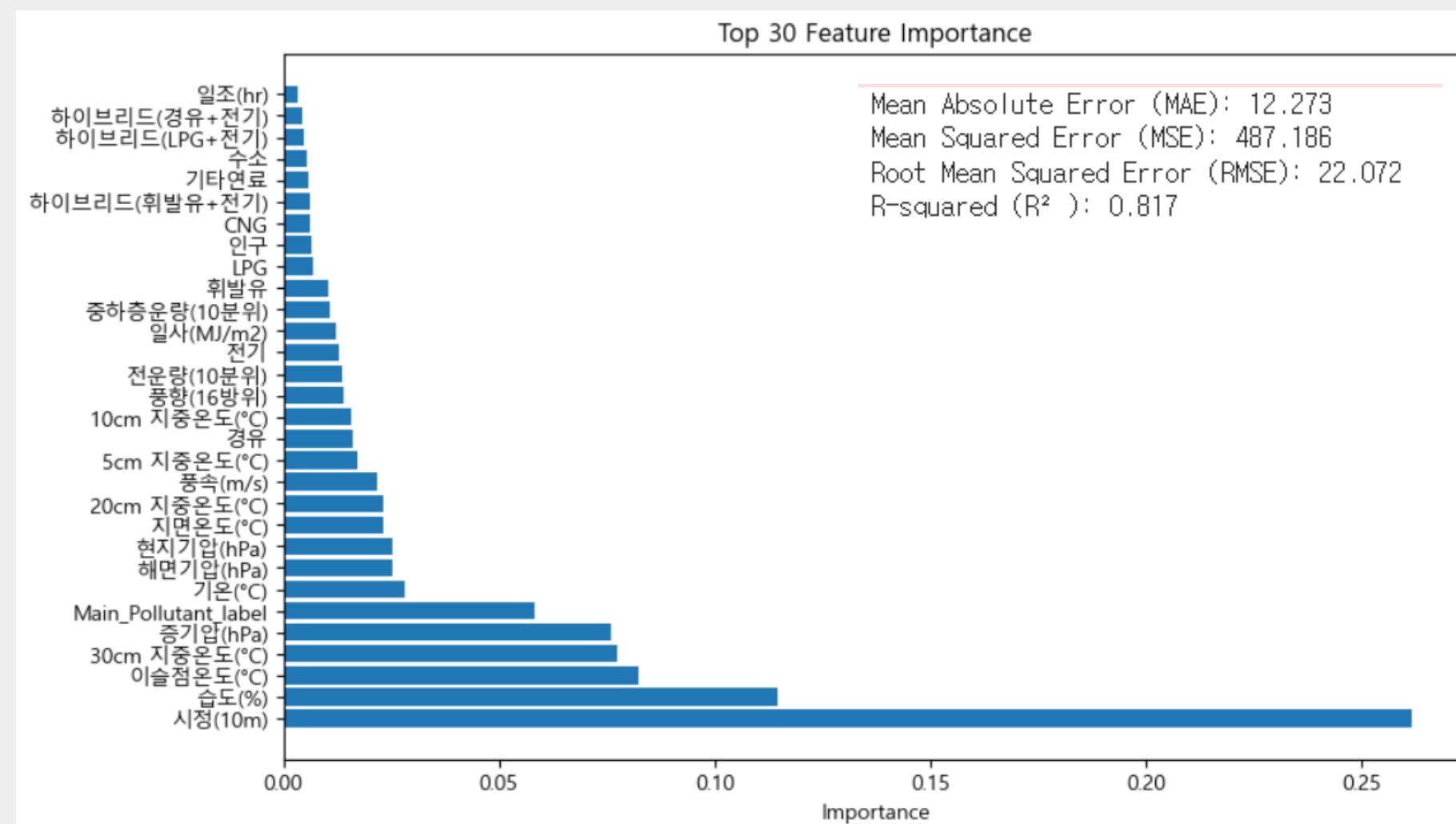
다중선행회귀 1차 모델링

Mean Absolute Error (MAE): 24.419
Mean Squared Error (MSE): 1412.235
Root Mean Squared Error (RMSE): 37.580
R-squared (R^2): 0.469

R-squared값이 낮고 예측이 틀릴때
MSE와 MAE의 차이가 극명해지기 때문에
해당 모델은 오차가 존재한다고 할 수 있다.



RandomForest 주요 변수 추출하기



다중선형회귀 재구축

Mean Absolute Error (MAE): 24.419
Mean Squared Error (MSE): 1412.235
Root Mean Squared Error (RMSE): 37.580
R-squared (R^2): 0.469

(주요 변수 추출 이전)



Mean Absolute Error (MAE): 20.458
Mean Squared Error (MSE): 746.460
Root Mean Squared Error (RMSE): 27.321
R-squared (R^2): 0.719

(주요 변수 추출 이후)

[MAE (Mean Absolute Error)]

예측값과 실제 값의 절대적인 차이를 평균화
24.219 -> 20.458로 감소. 예측 성능 향상

[MSE (Mean Squared Error)]

MSE는 오차의 제곱을 평균화
1412.235 -> 746.460으로 감소. 예측 정확도가 개선

[RMSE (Root Mean Squared Error)]

MSE의 제곱근으로, 실제 값과 예측 값의 평균적인 차이
37.580 -> 27.321로 감소. 예측 정확도가 향상

[R^2 (R-squared)]

모델이 설명하는 데이터의 분산 비율, 0에서 1 사이의 값
0.469 -> 0.719로 증가. 데이터의 변동성에 대한 **설명력 향상**

[결론]

주요 변수를 선택하여 모델을 재구성하니 예측 성능이 개선되었음을 확인
선택한 변수가 모델의 예측 정확도를 높이는데 기여했다고 추론 가능

확인 결과 주요 변수간 **다중공선성** 발견되어 하이퍼파라미터로 규제 강도 조절하는 릿지, 라쏘 회귀 시행

VIF 값이 10 이상인 피처와 이하인 피처 구분

Features with VIF > 10:

['하천변조경(면적)', '아파트 및 학교(면적)', '수림대(면적)', '수벽(면적)', '침수공간조성(면적)', '일반녹지(면적)', '간이휴게소(면적)', '분리대(면적)', '아파트 및 학교(개소)', '종합계(개소)', '일반녹지(개소)', '하천변조경(개소)', '건물주변(개소)', '시설녹지(개소)', '간이휴게소(개소)', '수벽(개소)', '5cm 지중온도(° C)', '30cm 지중온도(° C)', '하이브리드(경유+전기)', 'LPG', '경유', '휘발유', '전기']

Features with VIF <= 10:

['지하철환기구주변(개소)', '침수공간조성(개소)', '해면기압(hPa)', 'Air_Quality_Status_Label', '이슬점온도(° C)', '지면온도(° C)', 'Main_Pollutant_Label']



GridSearchCV를 활용해 cross validation 값을 기반으로 하이퍼파라미터 α (알파) 최적값으로 시행

BUT

Lasso Regression:	Ridge Regression:
MAE: 19.864	MAE: 20.454
MSE: 783.924	MSE: 746.299
RMSE: 27.999	RMSE: 27.318
R-squared: 0.705	R-squared: 0.719

다중공선성 문제를 고려하여 릿지(Ridge)와 라쏘(Lasso) 모델을 평가한 결과
기존의 다중선형회귀 결과와 비슷하거나 더 나빠졌다



여전한 다중공선성

아! 데이터 해석에 범주형 변수 처리와 성능 최적화에 강점을 가진 그래디언트 부스팅모델 **CatBoost**를 적용해보자

여전한 다중공선성

성능 최적화에 강점을 가진 그래디언트 부스팅모델 **CatBoost**를 적용해보자



Lasso Regression:	Ridge Regression:
MAE: 19.864	MAE: 20.454
MSE: 783.924	MSE: 746.299
RMSE: 27.999	RMSE: 27.318
R-squared: 0.705	R-squared: 0.719

VS

CatBoost Regression Evaluation:
MAE: 9.928
MSE: 215.304
RMSE: 14.673
R-squared: 0.919

기존 selected feature로 진행한 다중선행회귀, 라쏘, 릿지 회귀들보다
눈에 띠는 **성능 향상**

[CAI를 예측할 때, 어떤 모델을 선택해야 할까?]

다중공선성 문제가 변수의 비선행 관계를 효과적으로 다룰 수 있는 Tree 기반의 양상을 모델인 CatBoost을 사용하자.

RandomForest

Accuracy: 0.999				
Confusion Matrix:				
<code>[[2584 0 0 0]</code>				
<code>[0 8202 0 0]</code>				
<code>[0 2 1698 3]</code>				
<code>[0 0 13 373]]</code>				
Classification Report:				
	precision	recall	f1-score	support
0- 좋음	1.00	1.00	1.00	2584
1- 보통	1.00	1.00	1.00	8202
2- 나쁨	0.99	1.00	0.99	1703
3- 매우나쁨	0.99	0.97	0.98	386
	accuracy		1.00	12875
	macro avg	1.00	0.99	12875
	weighted avg	1.00	1.00	12875

Feature importance in descending order:		
	Feature	Importance
59	CAI	0.605173
13	시정(10m)	0.041434
18	30cm 지중온도(° C)	0.027465
60	Main_Pollutant_Label	0.023355
4	습도(%)	0.022561
..
26	간이휴게소(개소)	0.000104
31	기타(개소)	0.000090
53	하이브리드(CNG+전기)	0.000000
27	지하철환기구주변(개소)	0.000000
30	침수공간조성(개소)	0.000000
[61 rows x 2 columns]		

Precision (정밀도)

feature 중요도 순위

Positive로 예측한 값 중 실제로 Positive인 비율

Recall (재현율)

실제 Positive 값 중에서 모델이 Positive로 예측한 비율

F1 Score

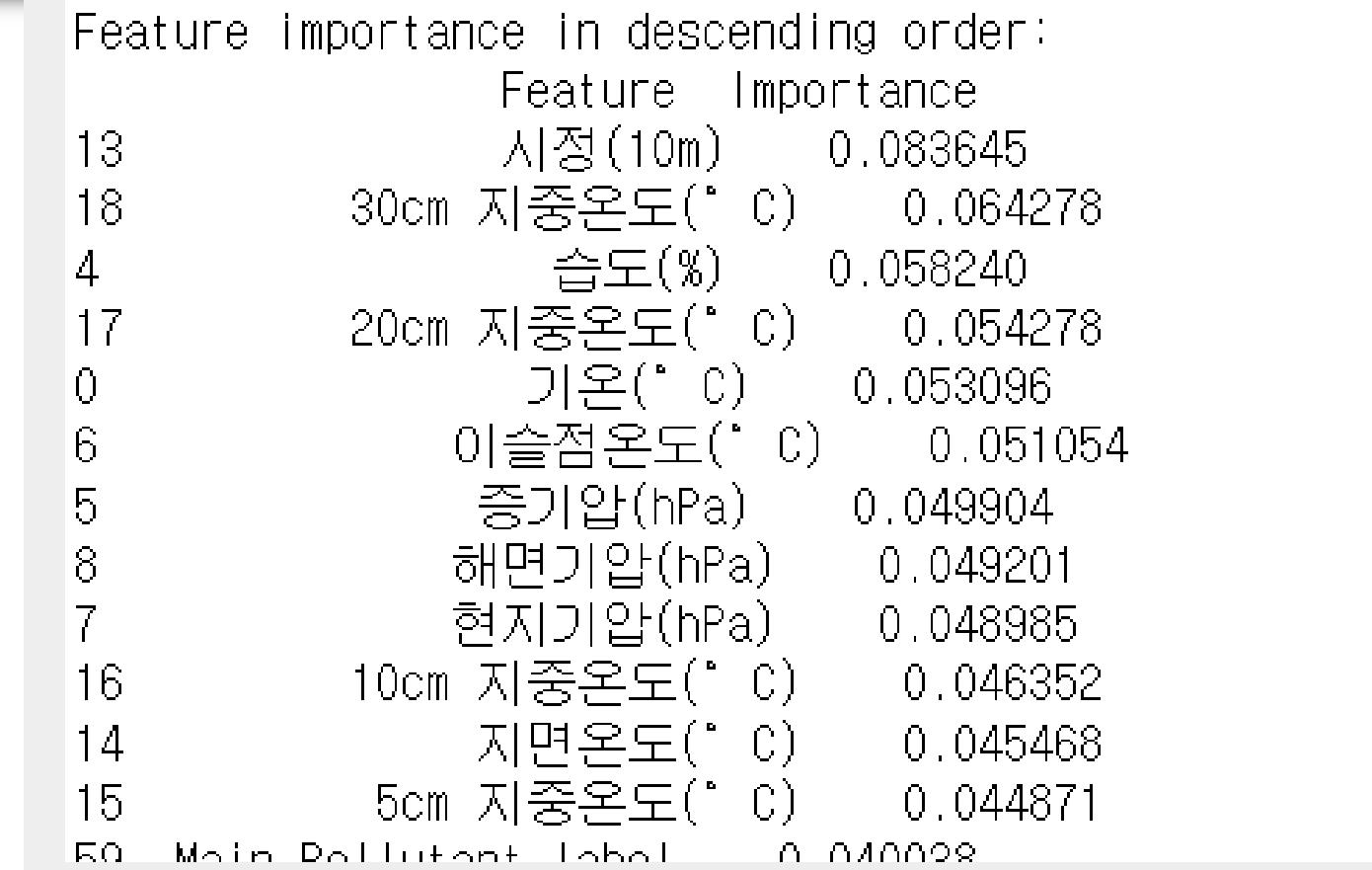
정밀도와 재현율의 조화평균, 모델의 전반적인 성능을 나타냄

→ 정확도 1.00 -> CAI가 종속변수를 만들때 사용한 컬럼이라 과적합 발생했다고 예상
 CAI 컬럼 독립변수에서 제외 후 다시 RandomForest 시행

RandomForest

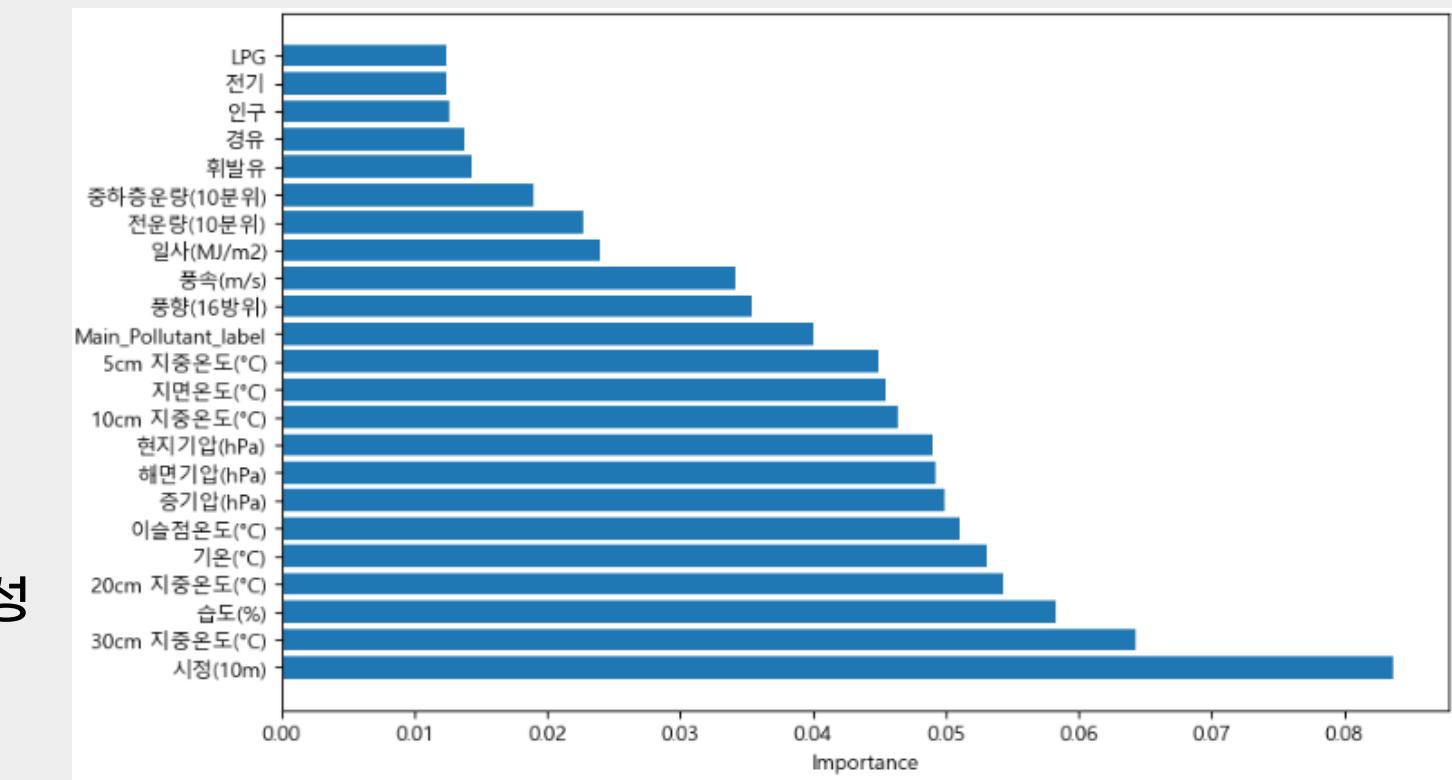
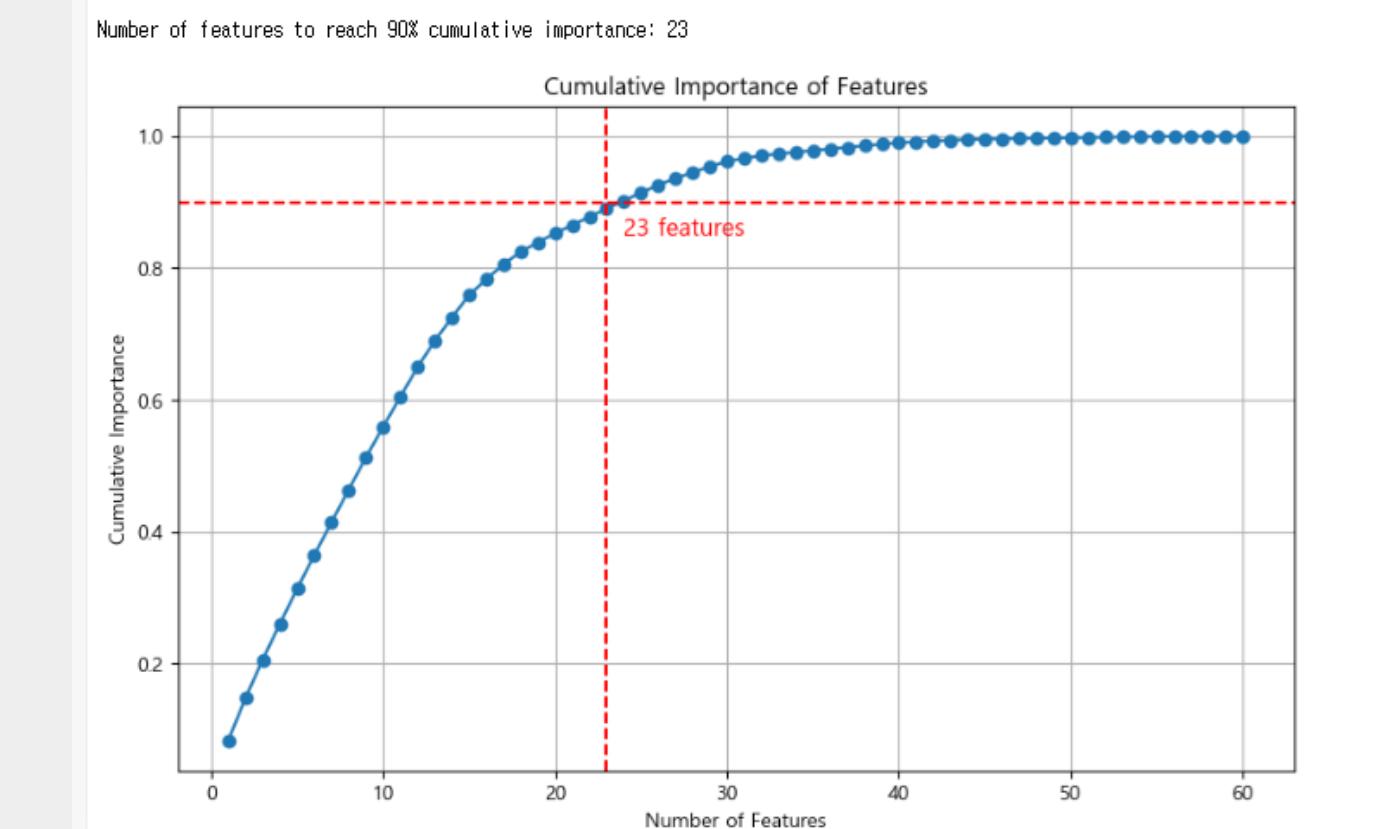
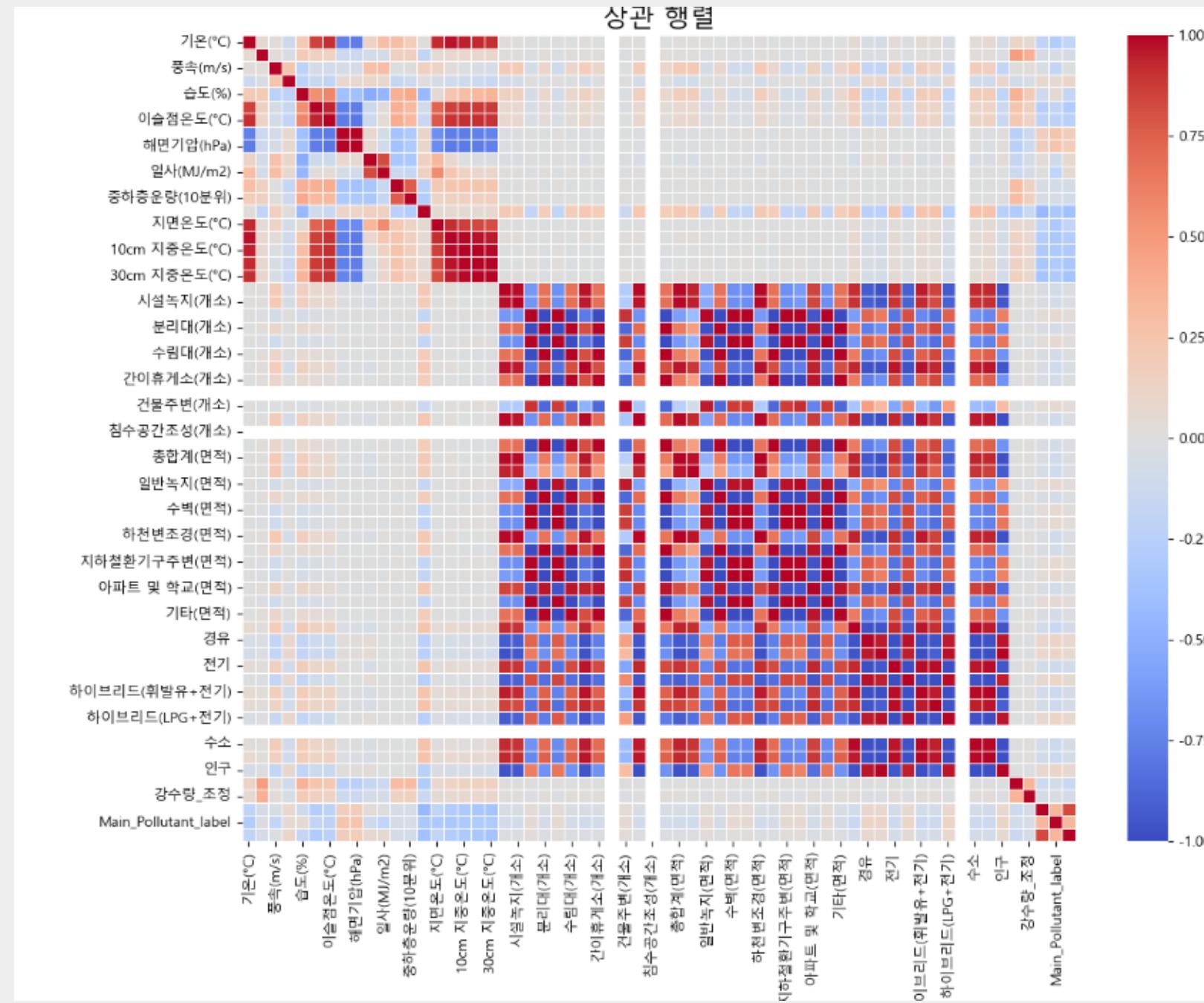
혼돈행렬
0- 좋음
1- 보통
2- 나쁨
3- 매우나쁨

```
Accuracy: 0.869
Confusion Matrix:
[[1847  737   0   0]
 [ 250 7838 111   3]
 [  0  491 1192  20]
 [  0   18   60 308]]
Classification Report:
          precision    recall   f1-score   support
          0         0.88     0.71      0.79     2584
          1         0.86     0.96      0.91     8202
          2         0.87     0.70      0.78     1703
          3         0.93     0.80      0.86     386
          accuracy           0.87     12875
          macro avg       0.89     0.79      0.83     12875
          weighted avg    0.87     0.87      0.86     12875
```



→ 1.0 → 0.87 과적합 해결

feature 중요도 순위



누적 중요도를 최소 90% 만족하는 특성 개수 찾기 => **23개**

23개 + 종속변수('Air_Quality_Status_label')로 새 데이터 프레임 생성

AdaBoost

	precision	recall	f1-score	support
좋음	0.49	0.53	0.51	2584
보통	0.74	0.76	0.75	8202
나쁨	0.48	0.37	0.42	1703
매우나쁨	0.56	0.46	0.50	386
accuracy			0.65	12875
macro avg	0.57	0.53	0.55	12875
weighted avg	0.65	0.65	0.65	12875

Training time: 4.77 seconds
Prediction time: 0.27 seconds

단순 로직 boosting인 AdaBoost 시행

Gradient Boosting

	precision	recall	f1-score	support
좋음	0.78	0.51	0.61	2584
보통	0.77	0.93	0.84	8202
나쁨	0.67	0.42	0.52	1703
매우나쁨	0.79	0.54	0.64	386
accuracy			0.76	12875
macro avg	0.75	0.60	0.65	12875
weighted avg	0.76	0.76	0.75	12875

Training time: 48.74 seconds
Prediction time: 0.08 seconds

오분류 가중치에 가중을 두는 Ada와 다르게 직전 단계의 오차를 학습해 확률값을 추정하는 Gradient Boosting 수행

→ Gradient descent 를 통해 잔차가 감소되어 정확도 향상

XGBoost

	precision	recall	f1-score	support
좋음	0.78	0.57	0.66	2584
보통	0.79	0.92	0.85	8202
나쁨	0.71	0.49	0.58	1703
매우나쁨	0.80	0.63	0.71	386
accuracy			0.78	12875
macro avg	0.77	0.65	0.70	12875
weighted avg	0.78	0.78	0.77	12875

Training time: 0.59 seconds
Prediction time: 0.02 seconds

정형화된 데이터 분류에 성능 좋은 XGBoost 수행

Accuracy : 0.76 -> 0.78

LightGBM

	precision	recall	f1-score	support
좋음	0.81	0.65	0.72	2584
보통	0.83	0.92	0.87	8202
나쁨	0.77	0.58	0.66	1703
매우나쁨	0.86	0.74	0.79	386
accuracy			0.82	12875
macro avg			0.72	12875
weighted avg	0.82	0.82	0.81	12875

Training time: 0.92 seconds
Prediction time: 0.05 seconds

Accuracy : 0.78 -> 0.82

학습 속도 개선을 위해 LightGBM 사용

CatBoost

	precision	recall	f1-score	support
좋음	0.76	0.63	0.69	1706
보통	0.82	0.91	0.86	5493
나쁨	0.73	0.55	0.63	1142
매우나쁨	0.76	0.65	0.70	243
accuracy			0.80	8584
macro avg	0.77	0.68	0.72	8584
weighted avg	0.79	0.80	0.79	8584

Training time: 1.59 seconds
Prediction time: 0.10 seconds

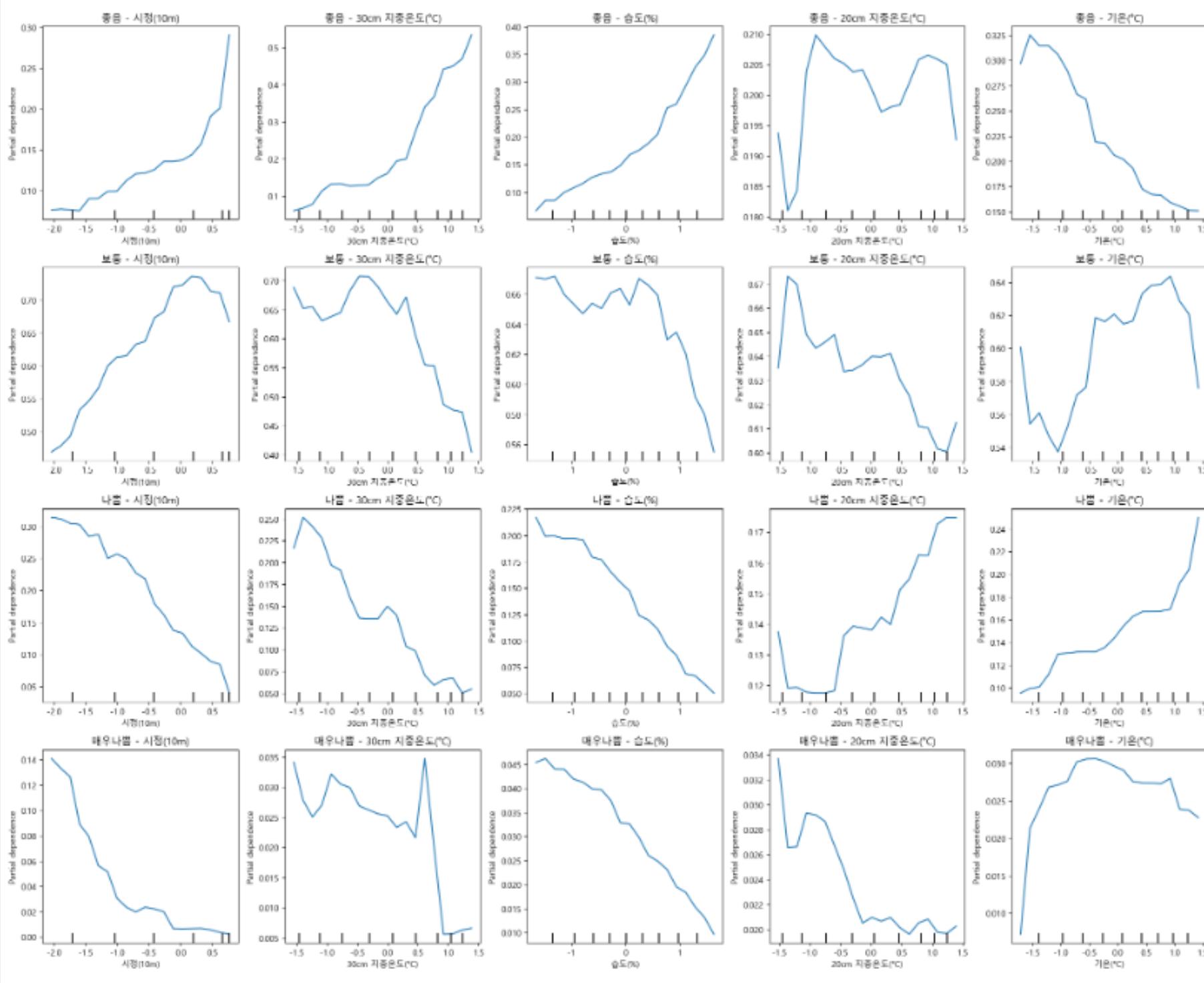
Gradient Boosting의 과적합을 해결하고 학습 속도가 빠른 범주형 변수 자체를 처리하는 CatBoost를 사용해 보자

Accuracy : 0.76 -> 0.80 (Gradient와 비교)

→ 모델 발전 순서대로 높은 accuacy 확인 가능하다!

PDP

Top 5 features: ['시정(10m)', '30cm 지중온도(°C)', '습도(%)', '20cm 지중온도(°C)', '기온(°C)']

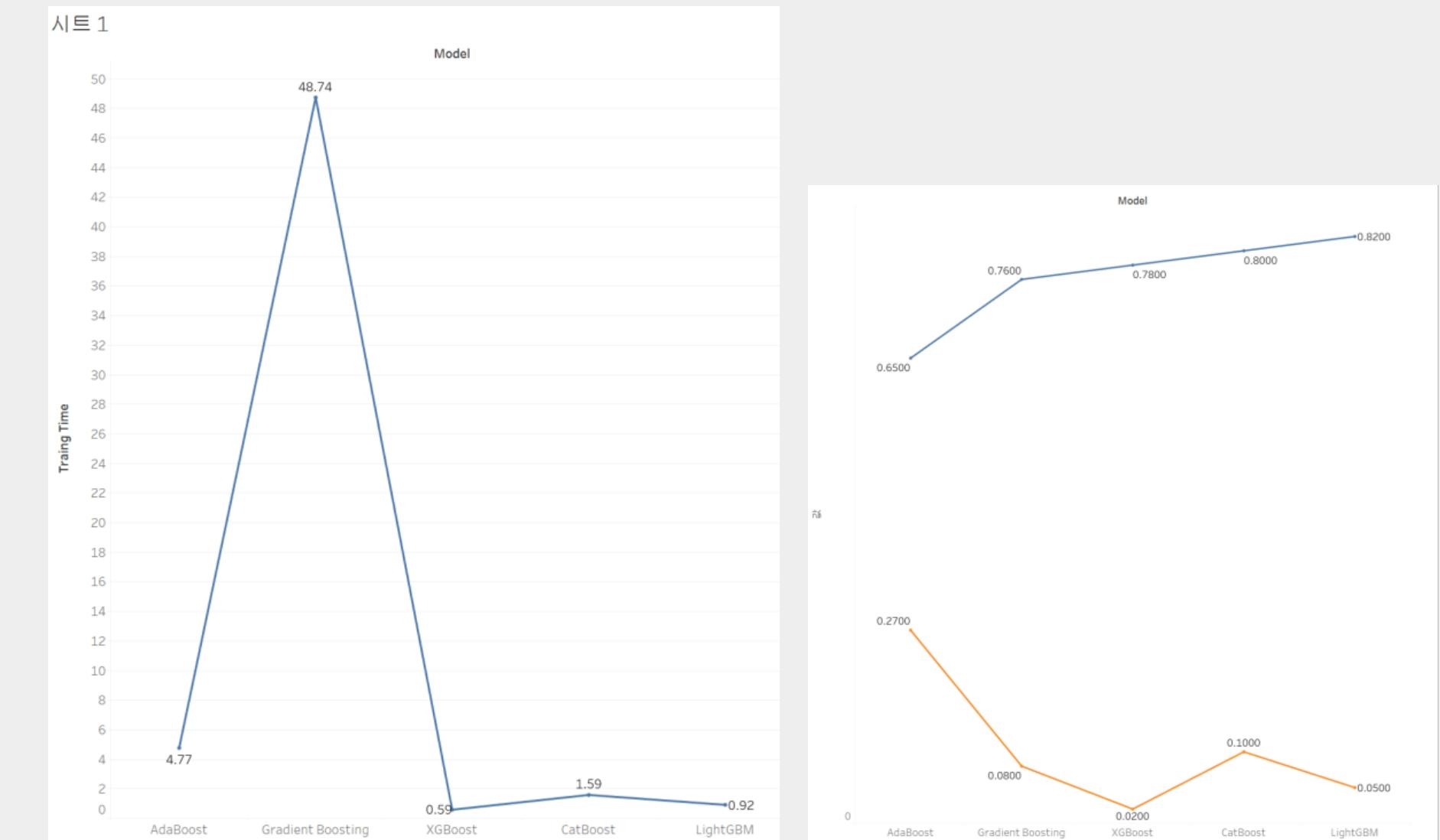


주요 변수 TOP5에 대해서 PDP를 진행

회귀 및 분류 모델 성능 비교

	R-squared
1차 다중선형회귀	0.469
2차 다중선형회귀	0.719
릿지 / 라쏘	0.719 / 0.790
CatBoost	0.919

CatBoost 모델이 가장 데이터 분포를 잘 설명한다.



Accuracy, Training time, Prediction time을 비교해보자

→ 모델 발전 순서대로 높은 accuracy 확인 가능하다!

우리의 최적 모델은

대기오염 지수 예측시



대기오염 상태 예측시



--> 회귀에서는 CatBoost가, 분류에서는 LightGBM이
가장 좋은 성능을 보였다

감사합니다

* 뒤 페이지부터는 별첨이 이어집니다.

[별첨]

개요

통합대기환경지수(CAI, Comprehensive air-quality index)는 대기오염도 측정치를 국민이 쉽게 알 수 있도록 하고 대기오염으로부터 피해를 예방하기 위한 행동지침을 국민에게 제시하기 위하여 대기오염도에 따른 인체 영향 및 체감오염도를 고려하여 개발된 대기오염도 표현방식

지수 산출방법

- 6개 대기오염물질별로 통합대기환경지수 점수를 산정하며 가장 높은 점수를 통합 지수값으로 사용
- 산출된 각각의 오염물질별 지수점수가 '나쁨' 이상의 등급이 2개 물질 이상일 경우 통합지수값에 가산점을 부여
 - 1개일 경우 : 점수가 가장 높은 지수점수를 통합지수로 사용
 - 2개일 경우 : 가장 높은 점수가 나온 오염물질을 영향 오염물질로 표시하고 그 오염물질의 점수에 50점을 가산
 - 3개 이상일 경우 : 가장 높은 점수가 나온 오염물질을 영향 오염물질로 표시하고 그 오염물질의 점수에 75점 가산
- 통합대기환경지수는 0에서 500까지의 지수를 4단계로 나누어 점수가 커질수록 대기상태가 좋지 않음을 나타냄

$$I_p = \frac{I_{HI} - I_{LO}}{BP_{HI} - BP_{LO}} \times (C_p - BP_{LO}) + I_{LO}$$

- I_p = 대상 오염물질의 대기지수 점수
- C_p = 대상 오염물질의 대기중 농도
- BP_{HI} = 대상 오염물질의 오염도 해당 구간에 대한 최고 오염도
- BP_{LO} = 대상 오염물질의 오염도 해당 구간에 대한 최저 오염도
- I_{HI} = BP_{HI} 에 해당하는 지수값(구간 최고 지수값)
- I_{LO} = BP_{LO} 에 해당하는 지수값(구간 최저 지수값)

지수산출에 필요한 변수

지수구분	좋음		보통		나쁨		매우나쁨	
	I_{LO}	0	51	101	251	500	101	251
점수구분 값	I_{HI}	50	100	250	500	101	251	500
	BP_{LO}	BP_{HI}	BP_{LO}	BP_{HI}	BP_{LO}	BP_{HI}	BP_{LO}	BP_{HI}
농도구분								
아황산가스(ppm)	1hr	0	0.02	0.021	0.05	0.051	0.15	0.151
일산화탄소(ppm)	1hr	0	2	2.01	9	9.01	15	15.01
오존(ppm)	1hr	0	0.03	0.031	0.09	0.091	0.15	0.151
이산화질소(ppm)	1hr	0	0.03	0.031	0.06	0.061	0.2	0.201
미세먼지 PM-10($\mu\text{g}/\text{m}^3$)	24hr ^{주1)}	0	30	31	80	81	150	151
초미세먼지 PM-2.5($\mu\text{g}/\text{m}^3$)	24hr ^{주2)}	0	15	16	35	36	75	76

* 측정된 농도값(C_p)이 정의된 농도값(BP_{HI})을 초과하는 경우에 BP_{HI} 값은 매우나쁨의 BP_{HI} 값으로 갈음한다.

- 주1) 미세먼지 PM-10 24hr은 미세먼지 PM-10 24시간 예측 이동평균임.
- 주2) 초미세먼지 PM-2.5 24hr은 초미세먼지 PM-2.5 24시간 예측 이동평균임.

별첨 [1]

오염도(CAI)와 데이터 전체 column의 상관계수

CAI	1.000000	증기압(hPa)	0.264610	CNG	0.064217	
PM2.5_CAI	0.908313	이슬점온도(° C)	0.213210	강수량_조정	0.060469	
PM2.5_24hr	0.897827	기온(° C)	0.200890	강수량(mm)	0.057634	
PM10_CAI	0.871342	지면온도(° C)	0.195664	중하층운량(10분위)	0.055222	
PM10_24hr	0.852922	해면기압(hPa)	0.171280	전운량(10분위)	0.052547	
PM2.5	0.839653	현지기압(hPa)	0.167534	시설녹지(면적)	0.045910	간이휴게소(개소)
PM10	0.779737	강수량_범주	0.114472	하천변조경(개소)	0.044937	수림대(개소)
CO_CAI	0.557584	습도(%)	0.107352	하천변조경(면적)	0.044345	수벽(면적)
CO	0.557562	경유	0.088359	총합계(면적)	0.043221	기타(면적)
N02	0.435167	기타연료	0.086301	아파트 및 학교(면적)	0.043018	수벽(개소)
N02_CAI	0.430735	인구	0.083382	시설녹지(개소)	0.042962	간이휴게소(면적)
시정(10m)	0.421378	LPG	0.082821	일반녹지(면적)	0.040848	기타(개소)
S02_CAI	0.365672	휘발유	0.082716	건물주변(면적)	0.040158	분리대(개소)
S02	0.365672	풍향(16방위)	0.081065	건물주변(개소)	0.039947	지하철환기구주변(면적)
30cm 지중온도(° C)	0.348648	전기	0.080887	일반녹지(개소)	0.039939	풍속(m/s)
20cm 지중온도(° C)	0.327699	하이브리드(LPG+전기)	0.080513	아파트 및 학교(개소)	0.039481	총합계(개소)
10cm 지중온도(° C)	0.306258	하이브리드(휘발유+전기)	0.079940	수림대(면적)	0.039277	일조(hr)
5cm 지중온도(° C)	0.280427	수소	0.074020	침수공간조성(면적)	0.039277	03_CAI
증기압(hPa)	0.264610	하이브리드(경유+전기)	0.073025	분리대(면적)	0.039277	일사(MJ/m2)
						03

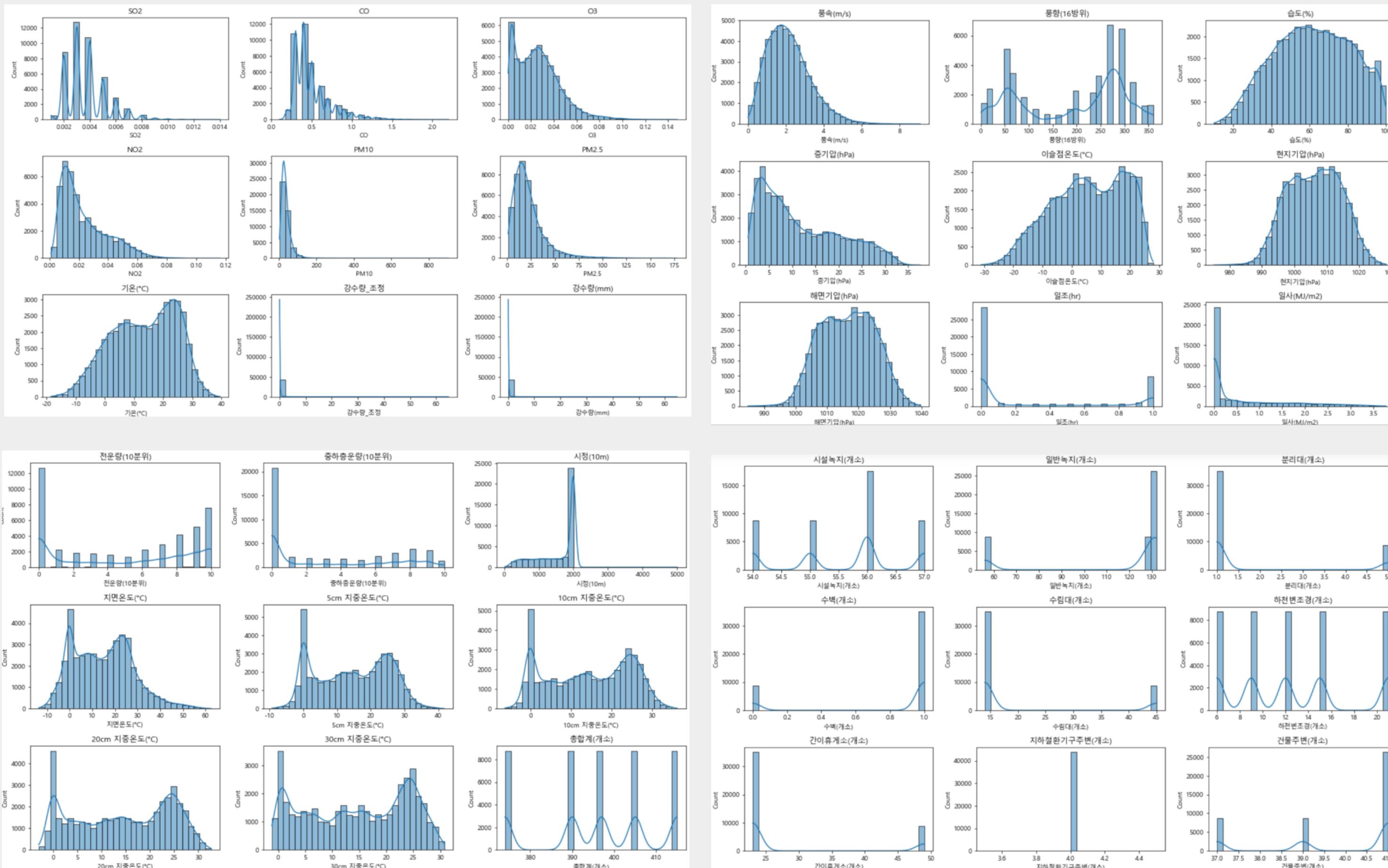
correlation 상대적 높음

correlation 상대적 낮음

내림차순

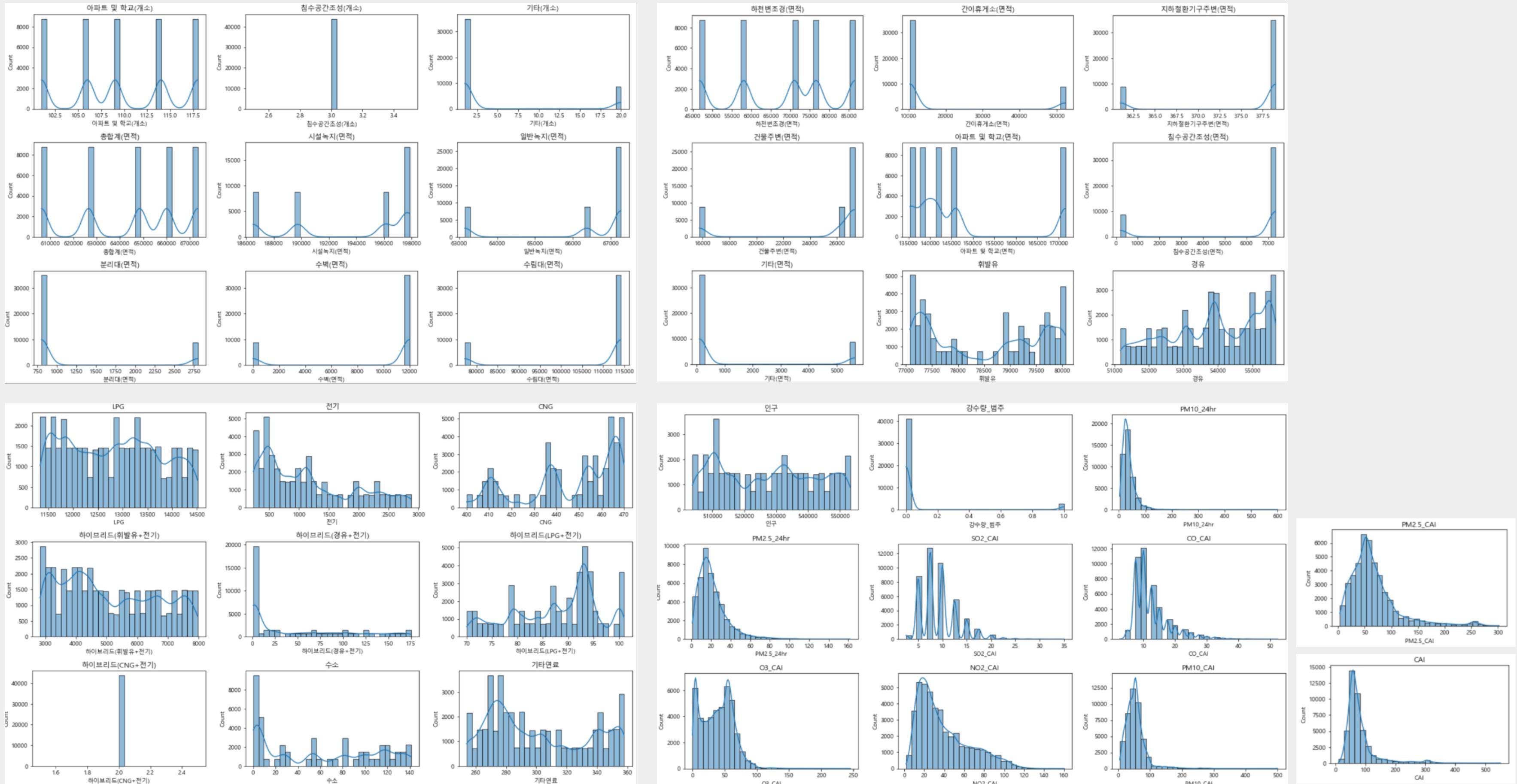
별첨 [2]

데이터 전체 column의 분포 Histogram



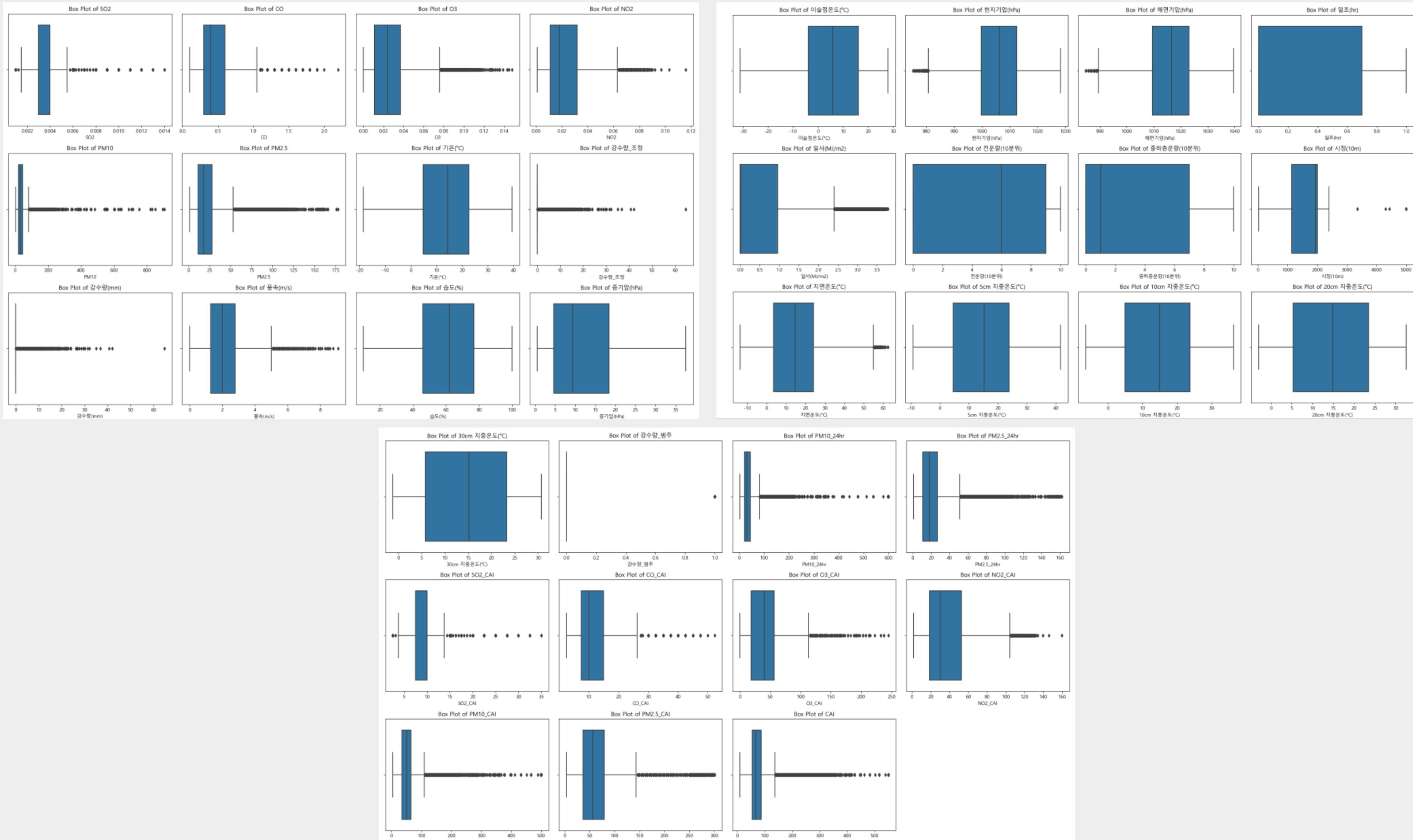
별첨 [2]

데이터 전체 column의 분포 Histogram



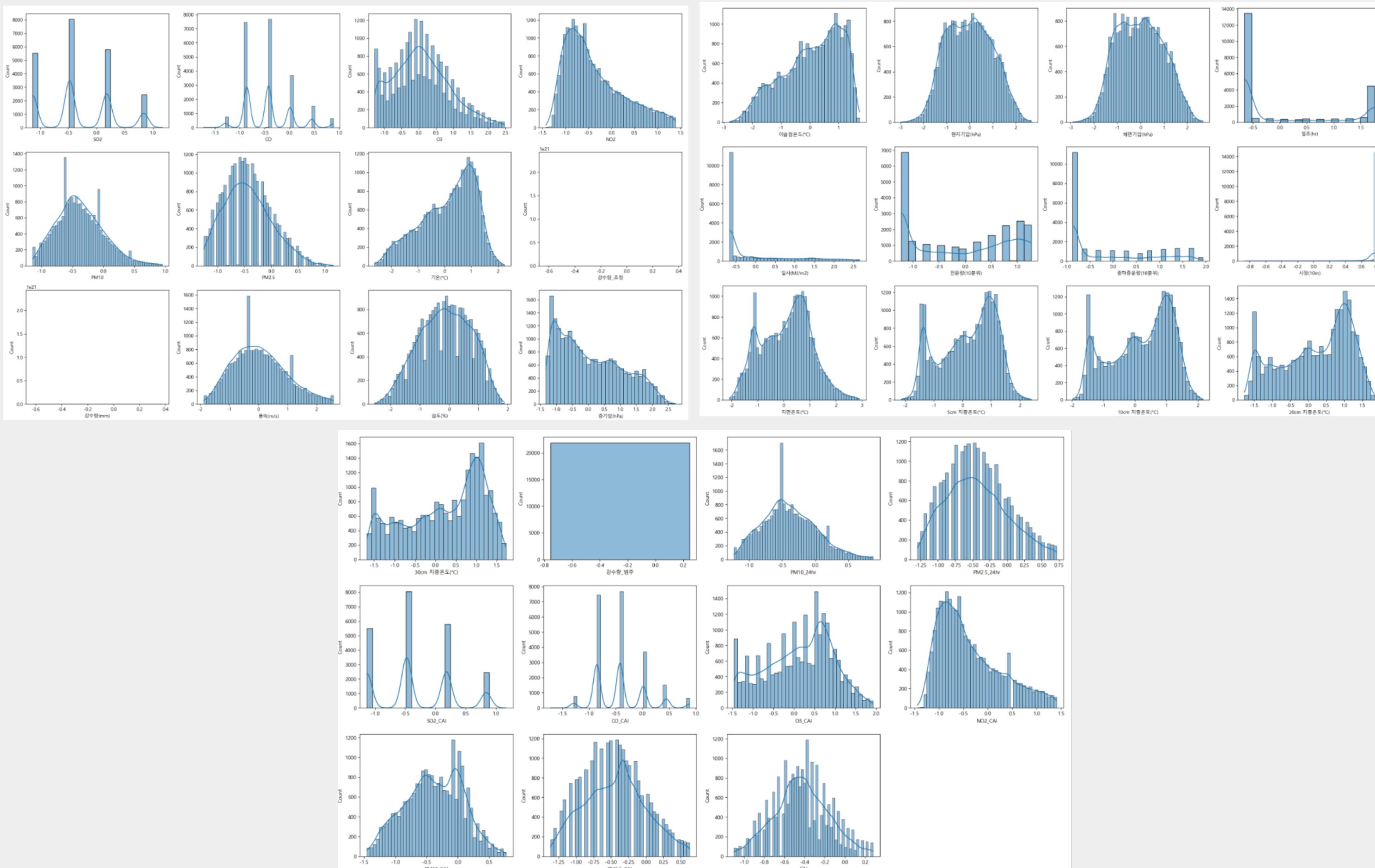
별첨 [3]

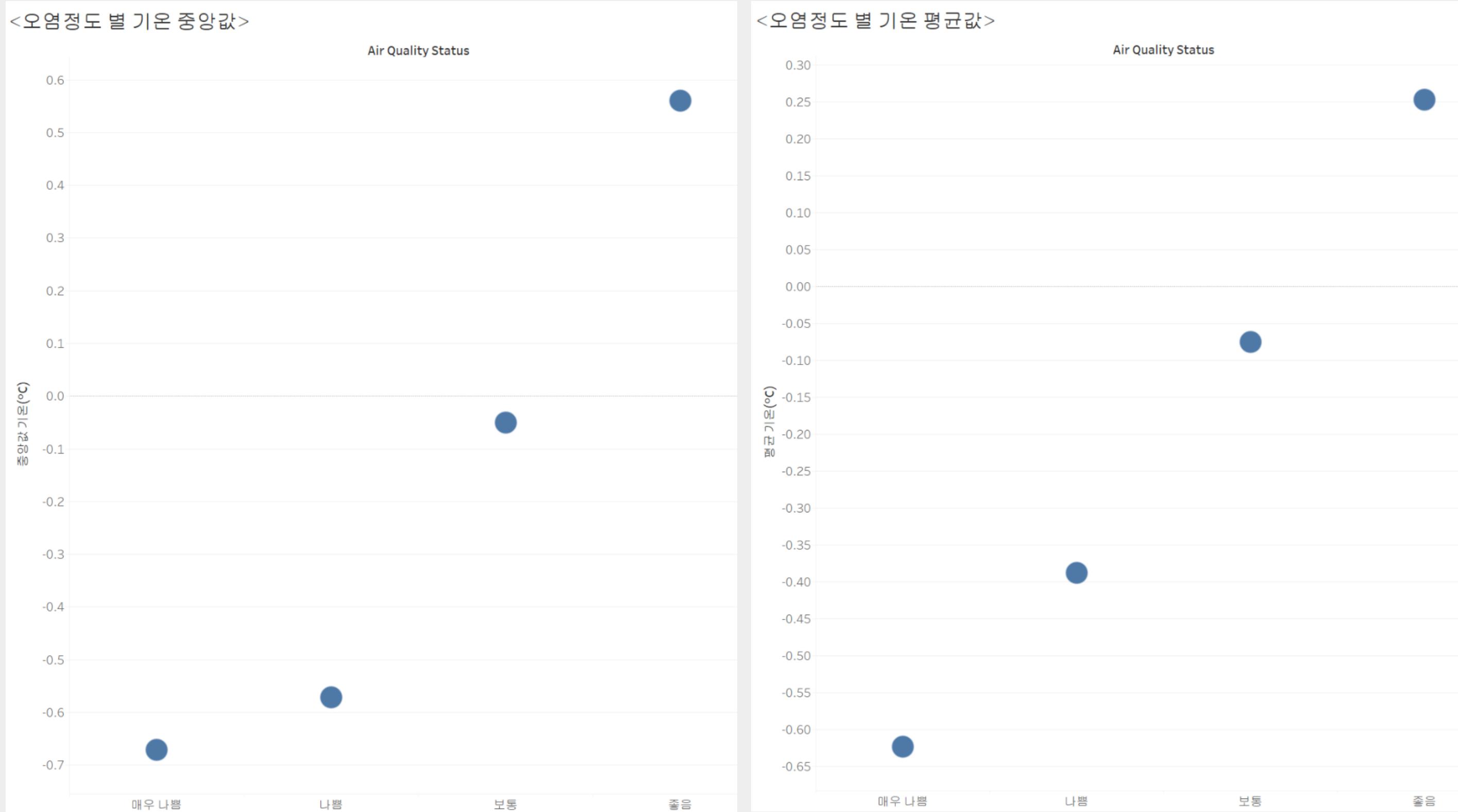
데이터 전체 column의 Boxplot



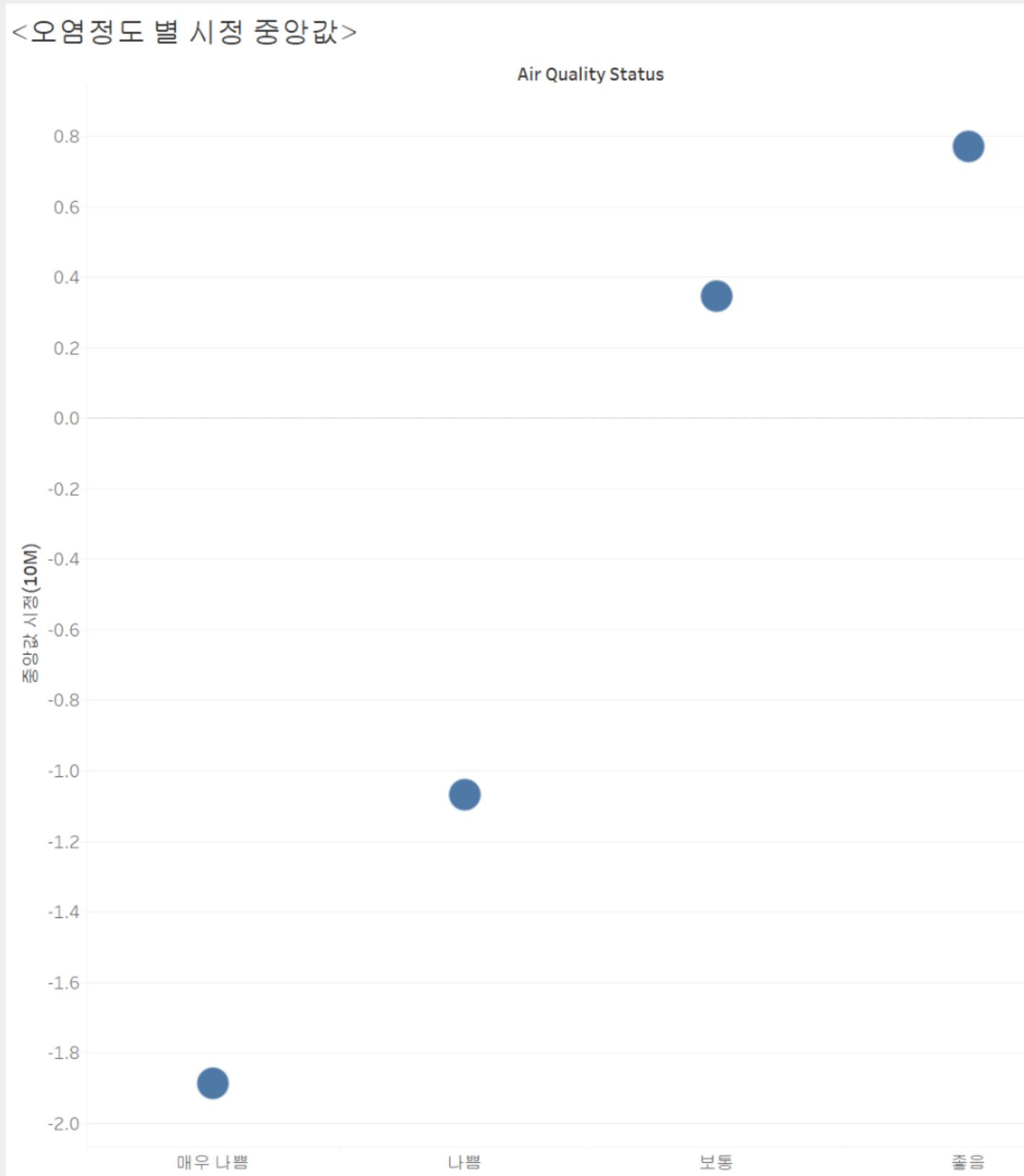
별첨 [4]

표준화 된 columns 시각화





기온이 상승할 수록 대기의 오염 수준은 좋아짐



대기상태가 좋을수록 시정값이 상승