마스크 없이 뛰고 싶었다…

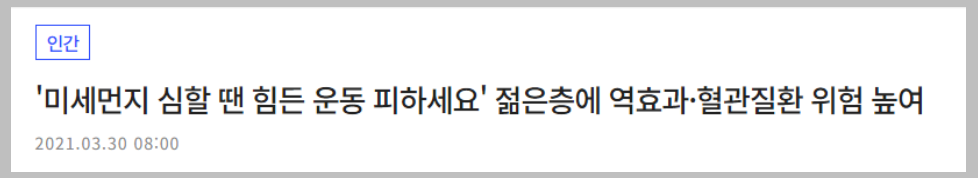
아시아 경제 2차 프로젝트

구성원: 조용준, 이석민

목차

주제 선정 배경 및 목적

배경



조승환, 동아사이언스

김희준 원장, 충청타임즈

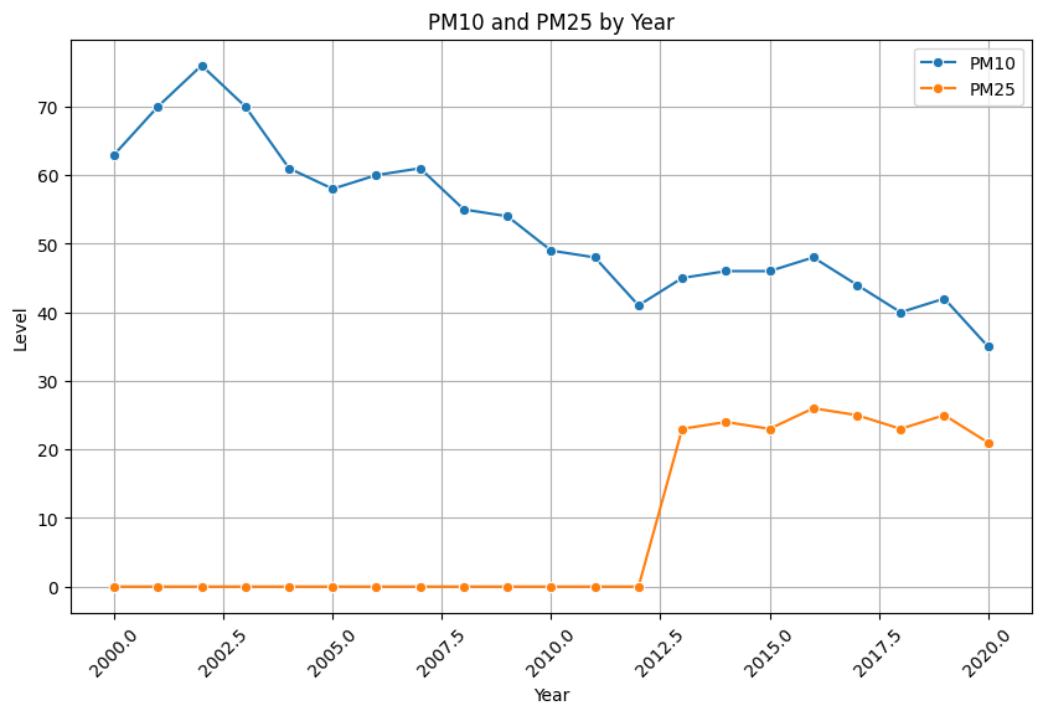
권서영, 아시아경제

목적

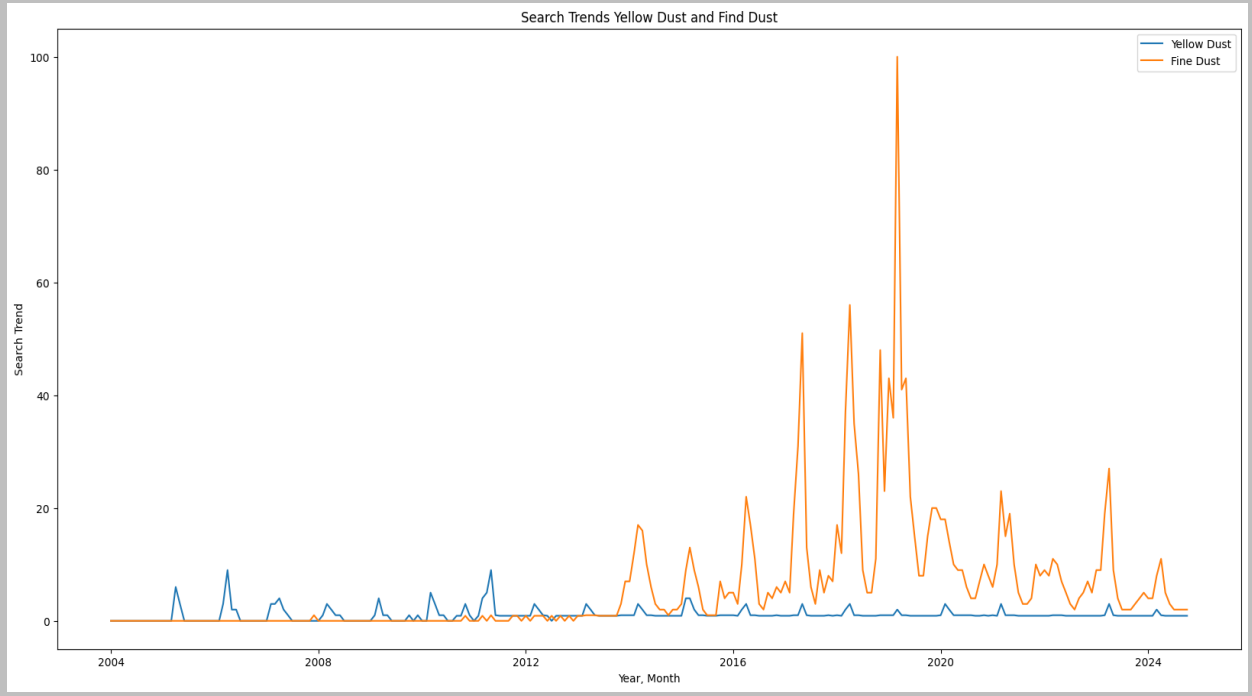
마스크 없이
안전하게 뛰고 싶다!!

주제 선정 배경 및 목적

미세먼지는 어떻게 흘러가고 있나?



연도별 (초)미세먼지 평균 농도



구글 트렌드 검색

과거에 비해 평균 농도는 **지속해서 감소**

관심도는 **과거에 비해 더욱 많아짐**

미세 먼지란…

케이웨더 제공, YTM사이언스

미세먼지란 대기에 있는 먼지 중
크기가 10마이크미터 이하인 먼지를 의미한다

보통 크기에 따라 미세먼지(10μm), 초미세먼지(2.5μm)로
구분된다

미세먼지는 크기가 너무 작아 인간의 기관지에서 걸러지지 못하고 폐포까지 침투하여 심혈관 질환과 호흡기 질환을
유발한다.

2022년 KBS 뉴스에 따르면 전세계 대기오염 사망자 수는
450만명에 이른다고 한다.

데이터 수집

에어코리아의 대기측정 자료와 기상청 종관기상관측 데이터를 확보해 사용

테이블 명	제공기관	사용기간	단위	변수명	변수 설명
최종확정 측정자료	한국환경공단 에어코리아	2016년 ~ 2020년	시간	지역	전국 시군구
		2016년 ~ 2020년	시간	망	측정장소특징
		2016년 ~ 2020년	시간	측정소코드	측정소구분코드
		2016년 ~ 2020년	시간	측정소명	측정소 이름
		2016년 ~ 2020년	시간	측정일시	시간별 시계열
		2016년 ~ 2020년	시간	SO2	아황산 가스, 통합대기환경지수변수
		2016년 ~ 2020년	시간	CO	일산화탄소, 통합대기환경지수변수
		2016년 ~ 2020년	시간	O3	오존, 통합대기환경지수변수
		2016년 ~ 2020년	시간	NO2	이산화질소, 통합대기환경지수변수
		2016년 ~ 2020년	시간	PM10	미세먼지, 통합대기환경지수변수
		2016년 ~ 2020년	시간	PM25	초미세먼지, 통합대기환경지수변수
		2016년 ~ 2020년	시간	주소	측정소 위치 주소
종관기상관측	기상청	2016년 ~ 2020년	시간	일시	시간별 시계열
		2016년 ~ 2020년	시간	기온	시간별 기온
		2016년 ~ 2020년	시간	강수량(mm)	시간별 강수량
		2016년 ~ 2020년	시간	풍향(16방위)	시간별 풍향, 0 ~ 360 사이의 숫자값
비고	2016 ~ 2019년 자료는 학습용으로 사용 예정 / 2020년 자료는 검증용으로 사용 예정				

데이터 전처리 (1)

- 에어코리아 미세먼지 데이터

1. 데이터 수 축소

기존	개선
전국 도,시,군,구,동,읍,면,리	서울시 한강 주변 11개구로 축소
전국 모든 지역 데이터 연간 약 280만행	구별로 구분 및 연간 약 8.7천개행

2. 결측치 제거

```
while unpre_Gu['미세먼지(PM10)'].isna().sum() and
      unpre_Gu['초미세먼지(PM25)'].isna().sum() != 0:
        unpre_Gu[['미세먼지(PM10)', '초미세먼지(PM25)']] =
        unpre_Gu[['미세먼지(PM10)', '초미세먼지(PM25)']].fillna(unpre_Gu[['미세먼지(PM10)', '초미세먼지(PM25)']].rolling(window = 3, min_periods=1).mean())
    if unpre_Gu['미세먼지(PM10)'].isna().sum() and
       unpre_Gu['초미세먼지(PM25)'].isna().sum() != 0:
        unpre_Gu[['미세먼지(PM10)', '초미세먼지(PM25)']] =
        unpre_Gu[['미세먼지(PM10)', '초미세먼지(PM25)']].fillna(0)
```

PM10 : 약 3만개의 결측치 / PM25 : 약 2.8만개의 결측치
-> 지역별 이전 3시간 평균치로 대체

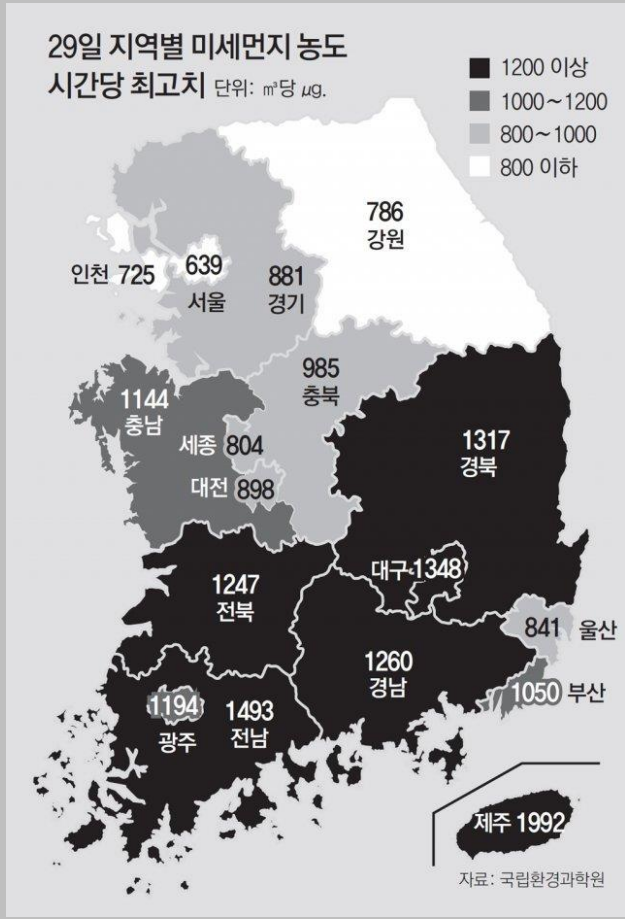
이유
1. 미세먼지는 앞 전 시간들의 수준을 유지 할 것이다.

2. 데이터는 시간별 데이터인데 하루 평균이나
한달 평균을 넣게 된다면 데이터 편차가 심할 것이다.

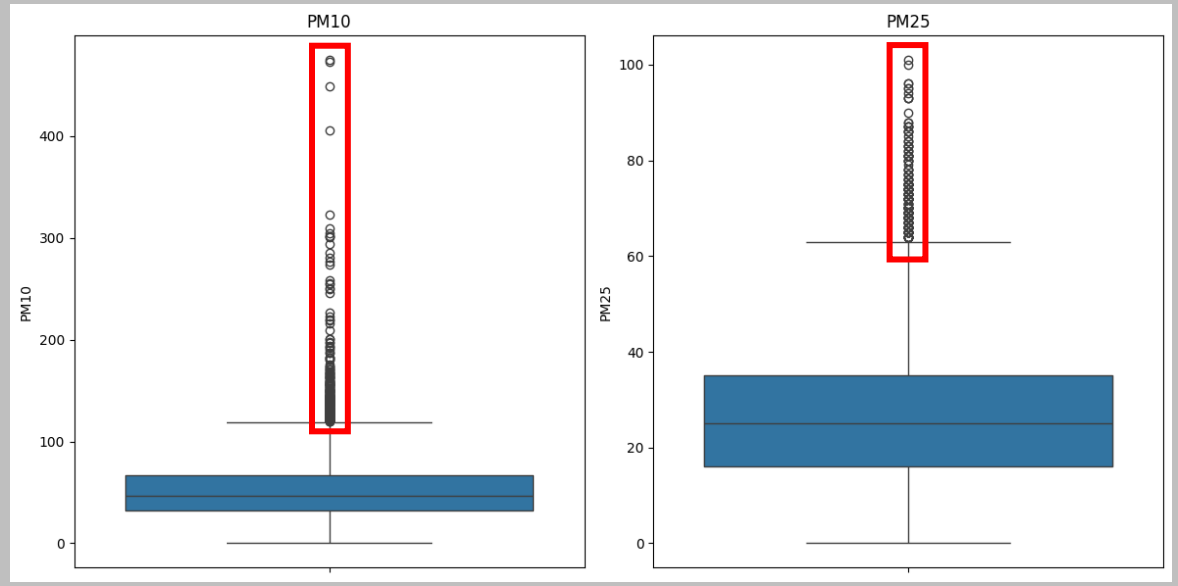
데이터 전처리 (1)

- 에어코리아 미세먼지 데이터

3. 이상치



명민준, 동아일보



2016년 (초)미세먼지 이상치

최대치를 벗어난 이상치의 수가 많았다.
대부분의 이상치들은 봄과 겨울에 집중되어 있다

데이터 전처리 (2)

- 기상청 종관기상관측

1. 강수량 결측치 제거

기존 강수량 컬럼의 결측치를 0으로
강수 0을 0.01로 대체

2. 풍향 컬럼 세분화 및 범주화

풍향을 세분화 하여 상관관계 분석 시도 및 가설 검증
(미세먼지가 많은 봄과 가을 겨울의 서풍과의 연관성)

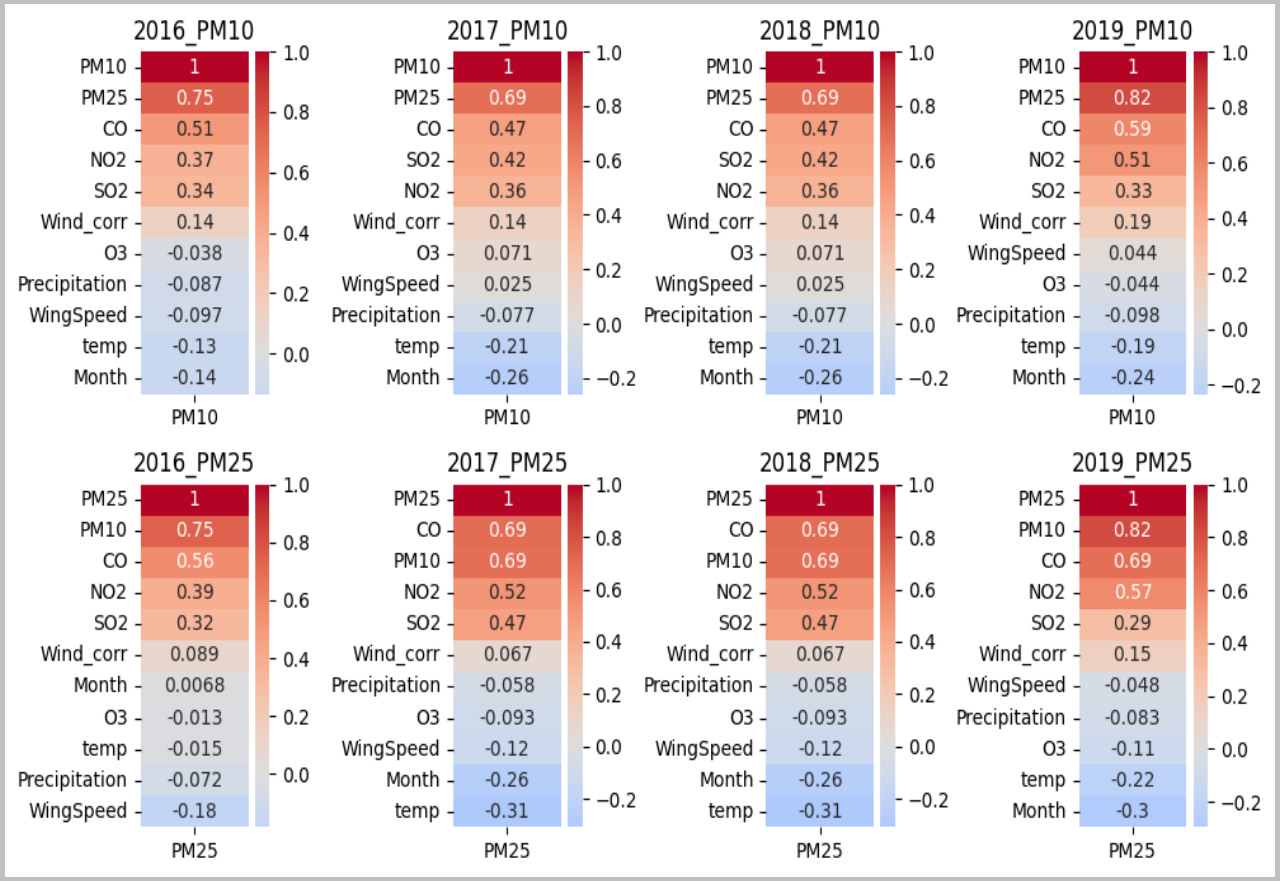
A	B	C	D	E	F
지점	지점명	일시	기온(°C)	강수량(mm)	풍향(16방위)
108	서울	2016-01-01 0:00	-1.9		0
108	서울	2016-01-01 1:00	-2.1		90
108	서울	2016-01-01 2:00	-2.2		0
108	서울	2016-01-01 3:00	-2.5		90
108	서울	2016-01-01 4:00	-2.9		70
108	서울	2016-01-01 5:00	-3.2		90
108	서울	2016-01-01 6:00	-3.1		90
108	서울	2016-01-01 7:00	-2.6		90
108	서울	2016-01-01 8:00	-2.4		90
108	서울	2016-01-01 9:00	-2		90
108	서울	2016-01-01 10:00	-0.6		90

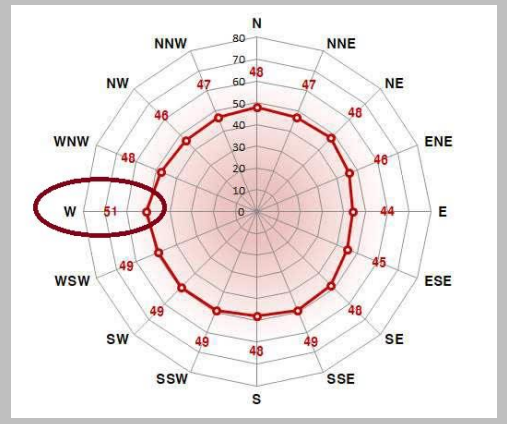
→

A	B	C	D	E	F	G	H	I
	지점	지점명	일시	기온(°C)	강수량(mm)	풍향(16방위)	풍향범주_corr	풍향범주
0	108	서울	2019-01-01 0:00	-5.5	0	290	8	북서
1	108	서울	2019-01-01 1:00	-5.9	0	270	7	서
2	108	서울	2019-01-01 2:00	-6.5	0	290	8	북서
3	108	서울	2019-01-01 3:00	-6.9	0	270	7	서
4	108	서울	2019-01-01 4:00	-7.2	0	270	7	서
5	108	서울	2019-01-01 5:00	-7.6	0	270	7	서
6	108	서울	2019-01-01 6:00	-7.9	0	290	8	북서
7	108	서울	2019-01-01 7:00	-7.7	0	320	8	북서
8	108	서울	2019-01-01 8:00	-7.7	0	360	1	북
9	108	서울	2019-01-01 9:00	-7	0	290	8	북서
10	108	서울	2019-01-01 10:00	-4.9	0	290	8	북서
11	108	서울	2019-01-01 11:00	-3.7	0	290	8	북서
12	108	서울	2019-01-01 12:00	-2.8	0	290	8	북서

시각화

미세먼지 / 초미세먼지는 누구와 가장 큰 상관성이 있을까?

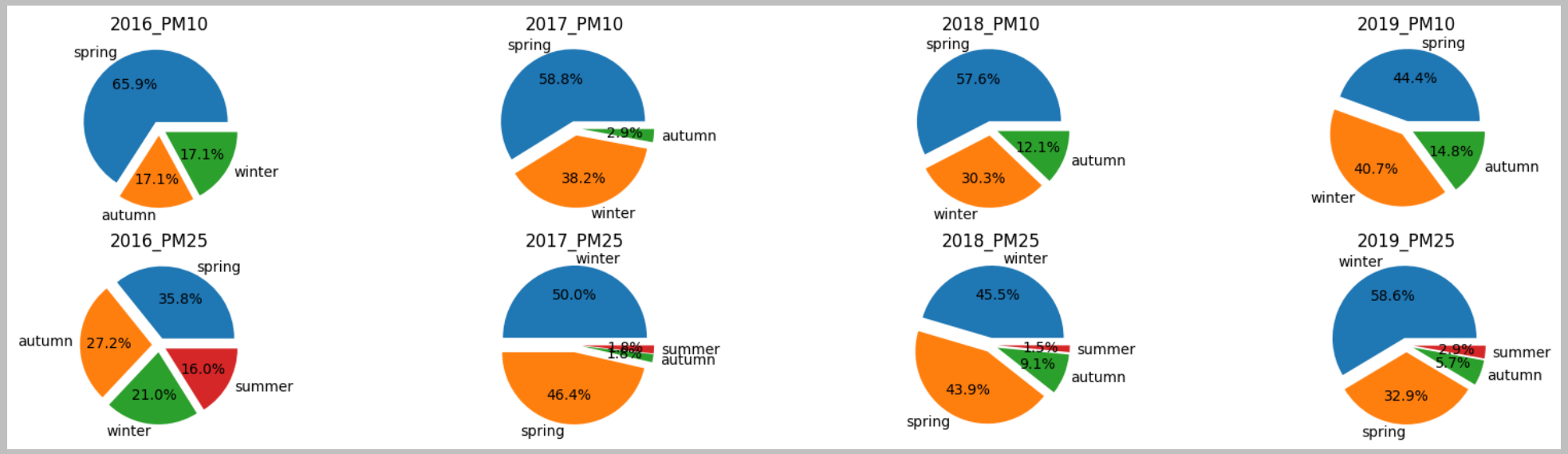




- 이산화화황(SO2), 이산화질소(NO2), 일산화탄소(CO) 변수가 초미세먼지, 미세먼지와의 상호 영향력이 높았다.
- ***의외로, 풍향과 미세먼지의 연관성은 높지 않았다***
- '2016년 서울 대기질 평가 보고서'에 따르면, 16방위 가운데 서풍(W)이 불때, 평균 미세먼지 농도가 51㎍/㎥로 가장 높았다.
- 서풍이 불면 동풍이 불 때 보다 평균 16%정도 농도가 높았다.
- *전처리 과정에서 기존 8방위를 16방위로 너무 세분화 했기 때문에 영향이 적은게 아닐까 추정된다.*

시각화

미세먼지 / 초미세먼지는 어떤 계절에 가장 심했을까?



연도별 (초)미세먼지 나쁨 이상인 날의 계절 비율

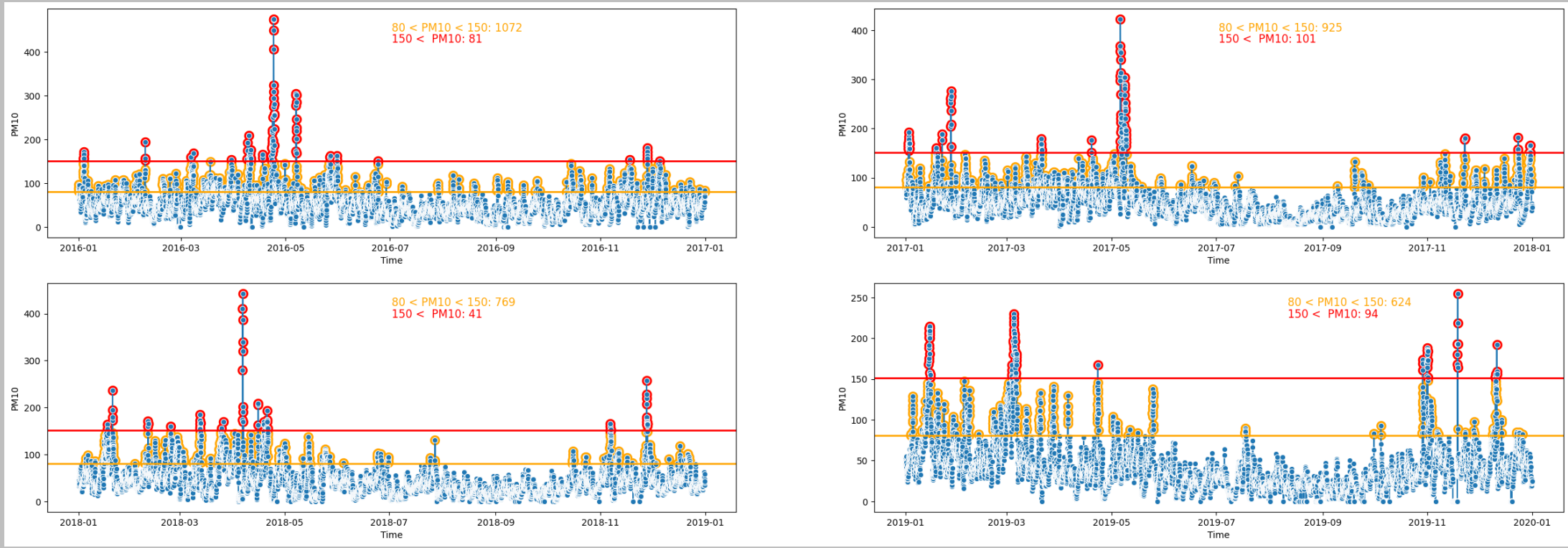
미세먼지의 경우 대부분의 연도에서 **봄과 겨울에 나쁨 / 매우 나쁜 날이 많았다.**

초미세먼지 또한 **미세먼지와 비슷한 분포**를 보였다.

다만 초미세먼지의 경우 여름에도 어느정도 관측이 되었다.

시각화

미세먼지 / 초미세먼지는 어떤 계절에 가장 심했을까?

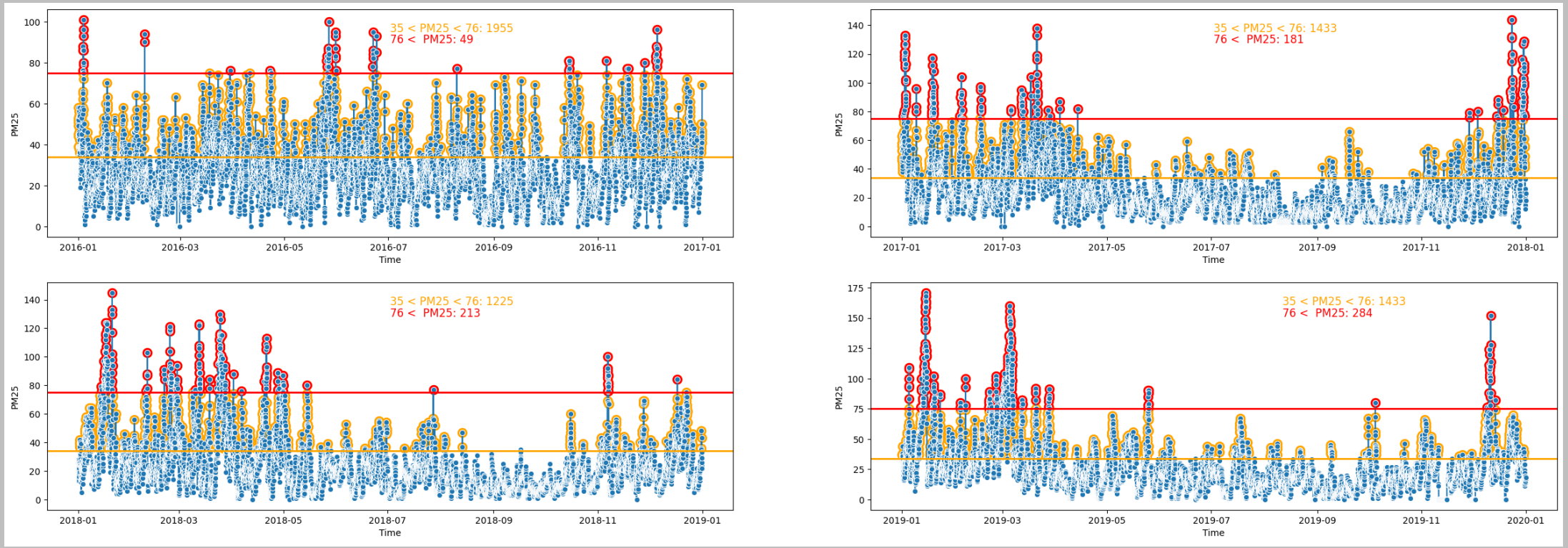


시간별 미세먼지가 나쁨 / 매우나쁨 분포

미세먼지는 봄과 겨울에 집중되었지만…

시각화

미세먼지 / 초미세먼지는 어떤 계절에 가장 심했을까?

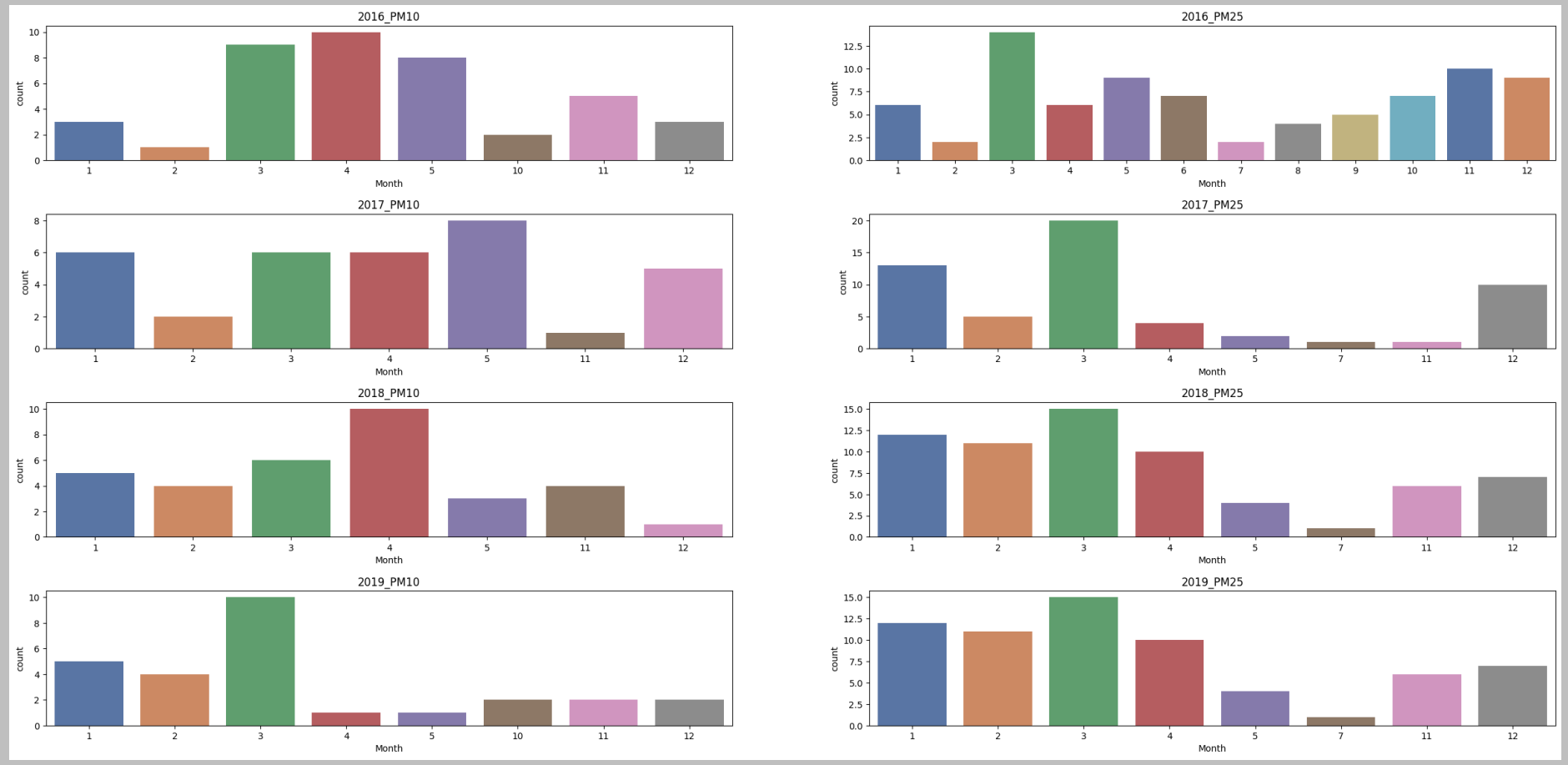


시간별 미세먼지가 나쁨 / 매우나쁨 분포

초미세먼지 역시 봄과 겨울에 가장 심하지만 미세먼지에 비해 여름과 가을에도 고르게 분포되어있다.

시각화

미세먼지 / 초미세먼지는 어떤 달에 가장 심했을까?



일 평균 미세먼지 초미세먼지가 나쁨 / 매우나쁨 수

미세먼지의 경우 3월 ~ 5월, 즉 **봄철에** 가장 많은 수를 기록했다.
초미세먼지의 경우 여름에도 있는 경우가 있었지만 **봄과 겨울철에 더욱 더 높은 수치를 보였다.**

모델링(ARIMA)

ARIMA모델

시계열 데이터의 예측과 분석을 위해 사용하는 통계모델

자기회귀(AR) / 미분(I) / 이동평균(MA)의 각 요소 **(p, d, q)를 조합**하여 구성

자기회귀(AR) : **현재시점**의 데이터는 **과거 시점**의 데이터의 **영향**을 받는다.
p : 과거 몇 개의 시점을 고려할지 결정하는 파라미터 / p = 1 이면 직전 시점 까지만 반영

미분(I) : 시계열 데이터가 **비정상성**을 보일 때 (추세가 있거나 변화량이 일정X) **차분** 적용
차분 : **데이터의 변화**를 구하는 것 / 1차 차분은 현재에서 한 시점 전 데이터를 뺀 값
d : 차분을 수행한 횟수 / d = 1 이면 1차 차분, d = 2 이면 2차 차분

이동평균(MA) : **과거 예측 값**과 **실제 값**의 **차이**를 (잔차)를 반영하여 현재 값을 보정
q : 과거 몇 개의 오차를 고려할지 결정하는 파라미터 / q = 1 이면 직전 시점의 오차 까지만 반영

ARIMA 적용 순서

1. 정상성 확인 : 비정상적이면 차분을 통해 정상성 확보
2. 모델 파라미터 선택(p, d, q) : 자기 회귀 / 차분 / 이동평균에 대한 적절한 파라미터 선택(ACF / PACF / AUTO ARIMA)
3. 모델 학습 : ARIMA모델에 데이터 학습
4. 모델 평가 : AIC / BIC / RMSE 등을 통해 평가 (AIC / BIC / RMSE 모두 값이 작을수록 좋은 모델)
5. 예측

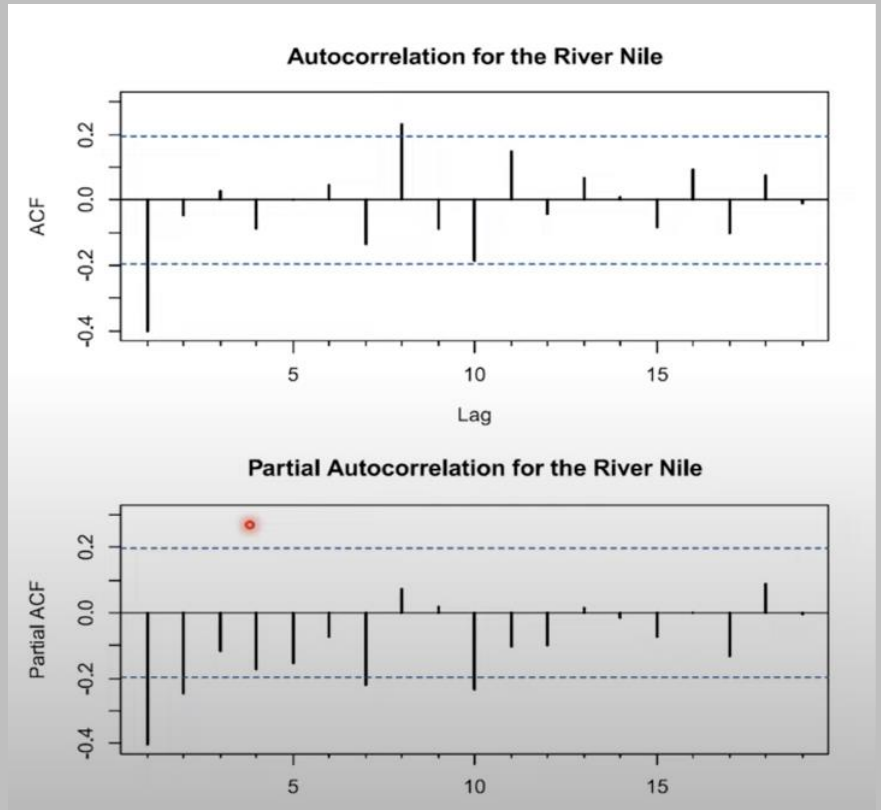
모델링(ARIMA)

ARIMA모델 – ACF / PACF

ACF는 q값을 FACF는 p값을 결정한다

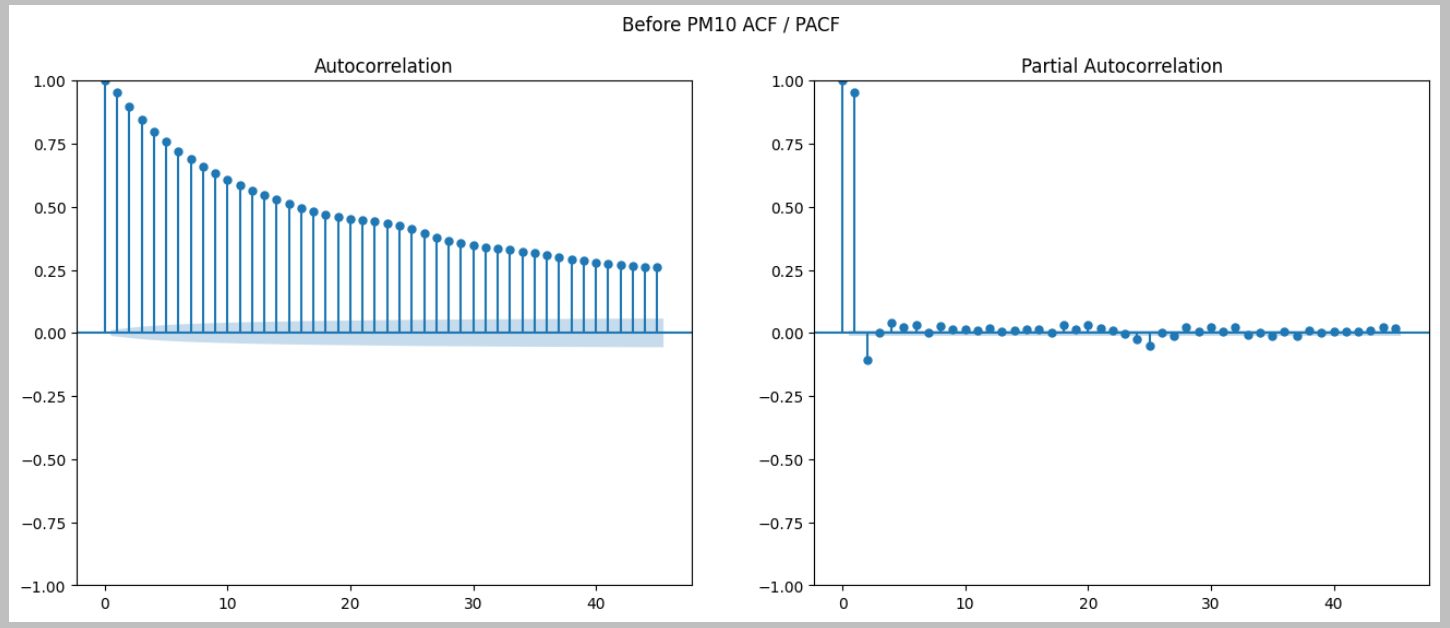
ACF(AutoCorrelation Function) : 자기상관 함수
- ARIMA(p, d, q) 중 **q값을 결정**하는 함수
- 현재 데이터가 과거 데이터에 얼마나 의존하는가
- 시차에(lag) 따른 상관 계수를 그린 그래프를 보고 적절한 값을 판단한다.
- 특정 시차 사이의 모든 시차 값을 토대로 상관계수를 구함
- 특정 시차에서 급격하게 0에 가까워지면 그 값을 채용

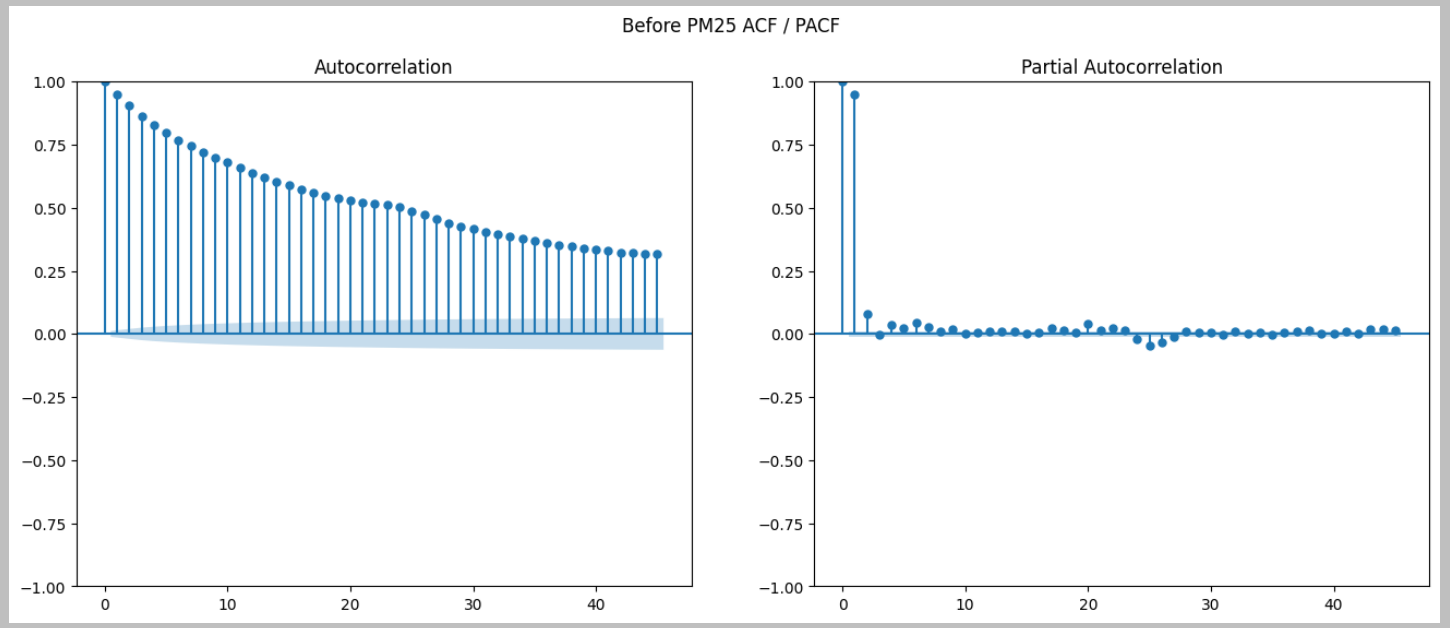
PACF(Patial AutoCorrelation Function) : 부분자기상관 함수
- ARIMA(p, d, q) 중 **p값을 결정**하는 함수
- 시차에(lag) 따른 상관 계수를 그린 그래프를 보고 적절한 값을 판단한다.
- 특정 시차 값과 현재 값 단 두가지만의 상관관계만 구함
- 특정 시차에서 급격하게 0에 가까워지면 그 값을 채용



ACF 값이 시차 1 이후 감소 / PACF 값이 시차 2 이후 감소 / P = 2 Q = 1로

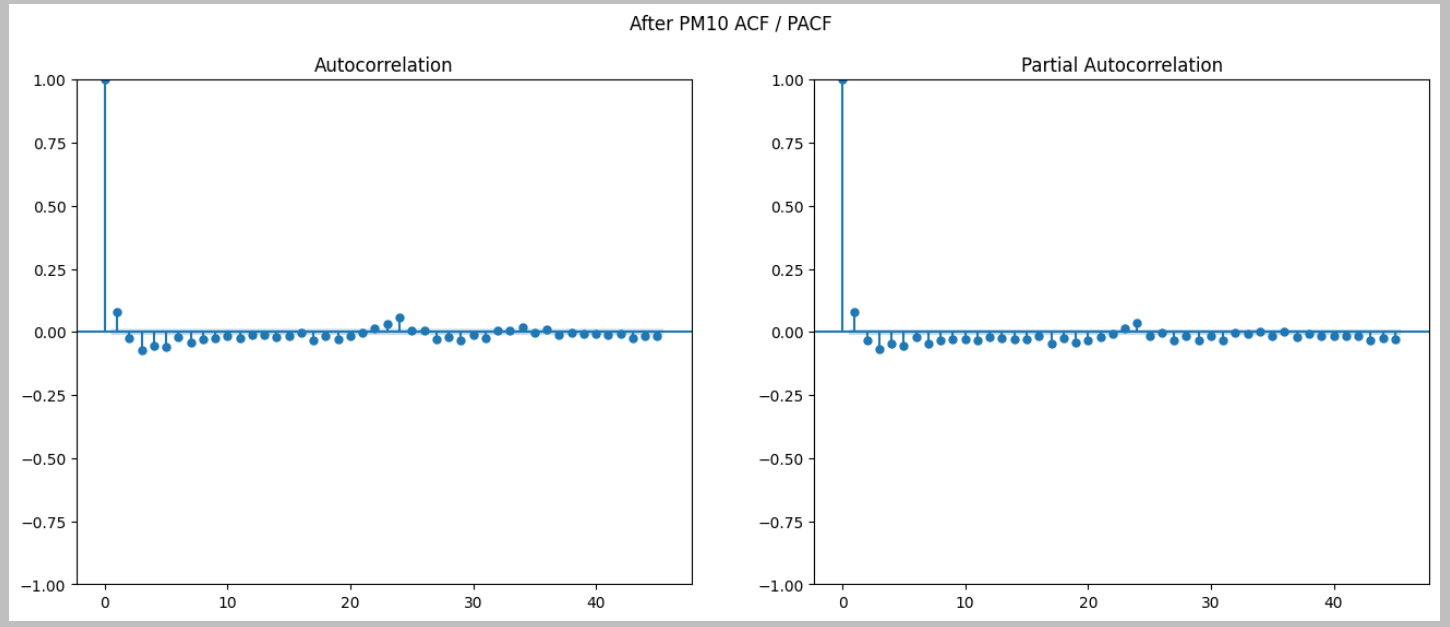
모델링(ARIMA)

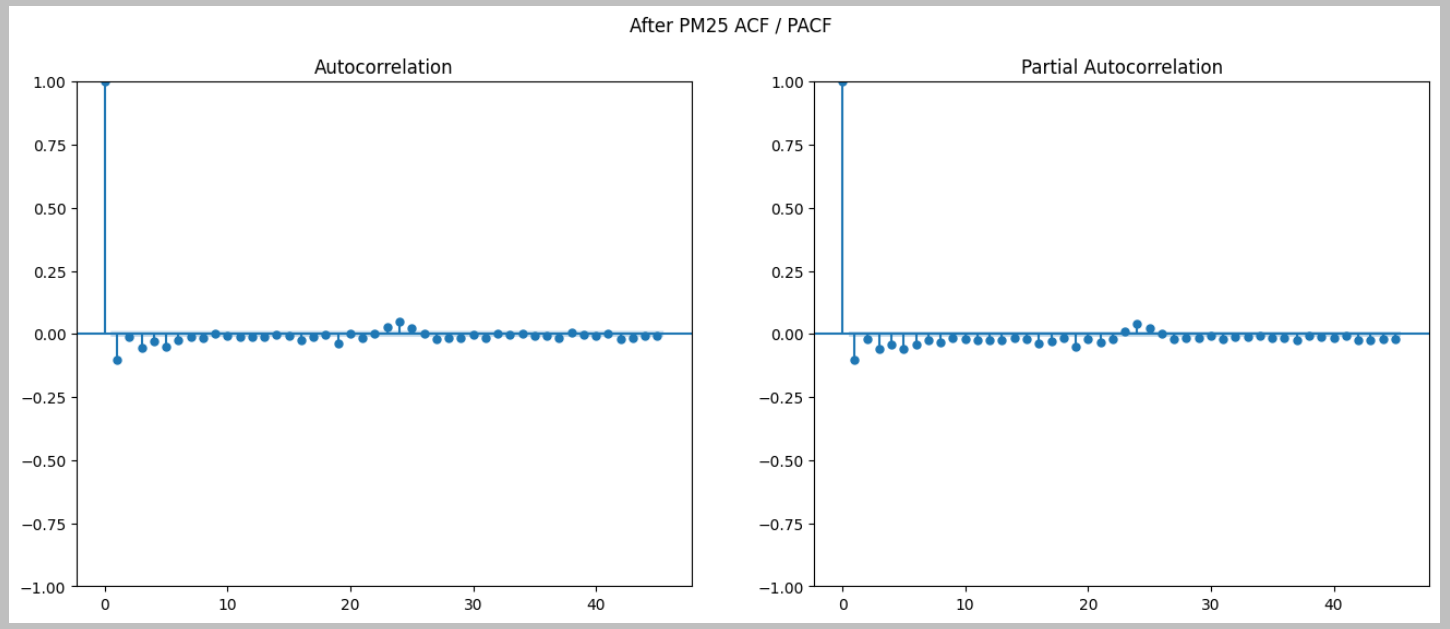




차분 전 ACF / PACF

모델링(ARIMA)





1차 차분 후 ACF / PACF
ARIMA(1, 1, 1) 채용

모델링(ARIMA)

```
model = ARIMA(train_data['PM10'], order = (1, 1, 1))
model_fit = model.fit()
model_fit.summary()
```

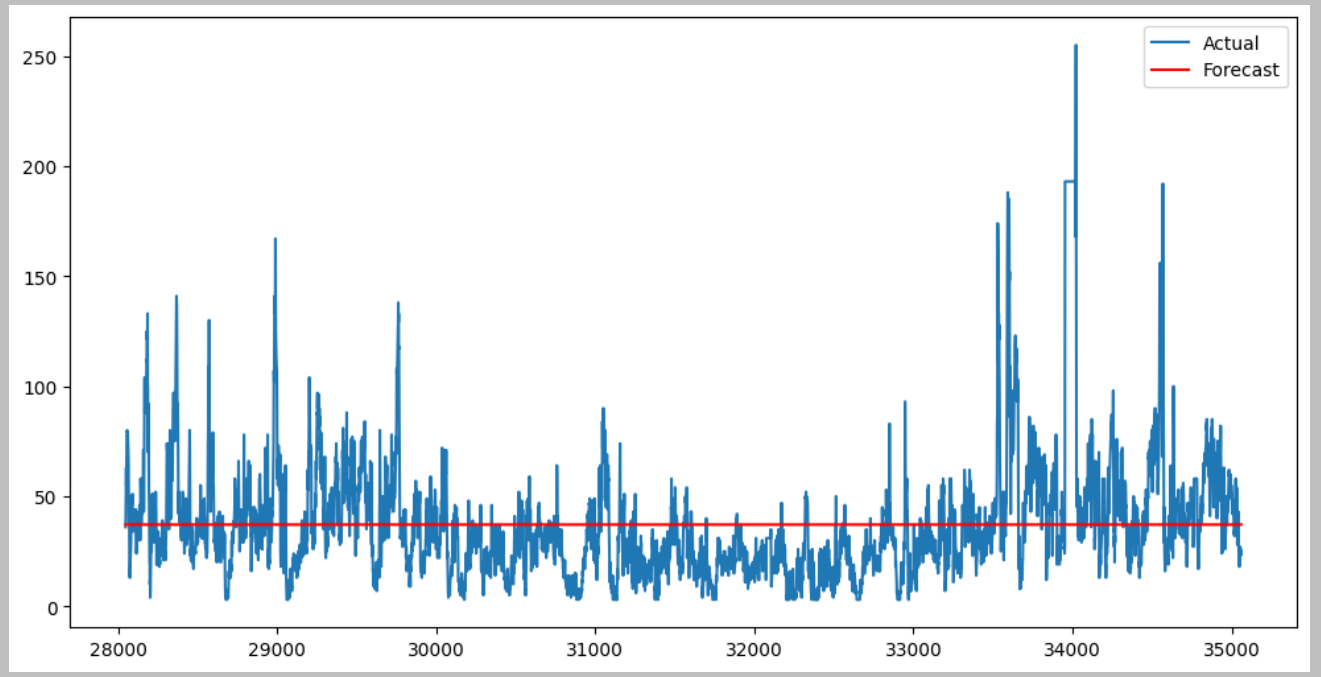
SARIMAX Results

Dep. Variable:	PM10	**No. Observations:**	28048
Model:	ARIMA(1, 1, 1)	**Log Likelihood**	-103180.612
Date:	Thu, 19 Sep 2024	**AIC**	206367.225
Time:	07:00:01	**BIC**	206391.950
Sample:	0	**HQIC**	206375.183
	- 28048		

Covariance Type: opg

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.0886	0.027	-3.245	0.001	-0.142	-0.035
ma.L1	0.1997	0.028	7.158	0.000	0.145	0.254
sigma2	91.8050	0.281	327.197	0.000	91.255	92.355

Ljung-Box (L1) (Q):	0.03	**Jarque-Bera (JB):**	458026.26
Prob(Q):	0.86	**Prob(JB):**	0.00
Heteroskedasticity (H):	0.68	**Skew:**	0.58
Prob(H) (two-sided):	0.00	**Kurtosis:**	22.76



AIC : 206367.225....
BIC : 206391.950....

```
mse = mean_squared_error(test_data['PM10'], pred)
print('MSE: ', mse)
```

```
MSE:  770.9995826762112
```

MSE : 770.99958...

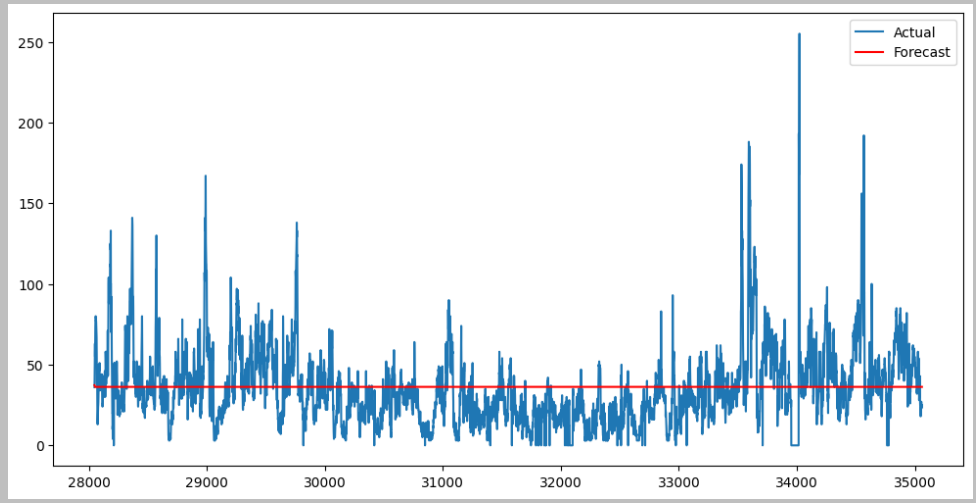
예측 성능이 매우 떨어짐

모델링(ARIMA)

Auto ARIMA 사용

```
1 stepwise_fit = auto_arima(train_data['PM10'], trace = True, suppress_warnings = True)
2
3 stepwise_fit.summary()
```

```
Performing stepwise search to minimize aic
 ARIMA(2,1,2)(0,0,0)[0] intercept   : AIC=inf, Time=46.23 sec
 ARIMA(0,1,0)(0,0,0)[0] intercept   : AIC=208186.218, Time=1.40 sec
 ARIMA(1,1,0)(0,0,0)[0] intercept   : AIC=208006.009, Time=1.61 sec
 ARIMA(0,1,1)(0,0,0)[0] intercept   : AIC=207997.228, Time=4.13 sec
 ARIMA(0,1,0)(0,0,0)[0]             : AIC=208184.219, Time=0.42 sec
 ARIMA(1,1,1)(0,0,0)[0] intercept   : AIC=207996.783, Time=8.23 sec
 ARIMA(2,1,1)(0,0,0)[0] intercept   : AIC=inf, Time=64.79 sec
 ARIMA(1,1,2)(0,0,0)[0] intercept   : AIC=inf, Time=46.81 sec
 ARIMA(0,1,2)(0,0,0)[0] intercept   : AIC=207992.736, Time=5.64 sec
 ARIMA(0,1,3)(0,0,0)[0] intercept   : AIC=207864.373, Time=7.40 sec
```



Best model: ARIMA(0,1,5)(0,0,0)[0]
Total fit time: 504.190 seconds

SARIMAX Results

Dep. Variable:	y	**No. Observations:**	28048
Model:	SARIMAX(0, 1, 5)	**Log Likelihood**	-103812.722
Date:	Tue, 22 Oct 2024	**AIC**	207637.443
Time:	11:07:07	**BIC**	207686.893
Sample:	0 - 28048	**HQIC**	207653.361

Covariance Type: opg

	coef	std err	z	P>|z|	[0.025	0.975]
ma.L1	0.0690	0.002	30.801	0.000	0.065	0.073
ma.L2	-0.0380	0.003	-12.044	0.000	-0.044	-0.032
ma.L3	-0.0874	0.003	-29.471	0.000	-0.093	-0.082
ma.L4	-0.0729	0.003	-23.691	0.000	-0.079	-0.067
ma.L5	-0.0719	0.003	-22.087	0.000	-0.078	-0.066
sigma2	96.0386	0.274	351.131	0.000	95.502	96.575

Ljung-Box (L1) (Q):	0.14	**Jarque-Bera (JB):**	553935.31
Prob(Q):	0.71	**Prob(JB):**	0.00
Heteroskedasticity (H):	0.71	**Skew:**	0.78
Prob(H) (two-sided):	0.00	**Kurtosis:**	24.72

자동으로 최적의(p,d,q) 값을 찾아줌

AIC : 207637.443
BIC : 207686.893...

예측 성능 또한 매우 떨어지는 모습을 보임

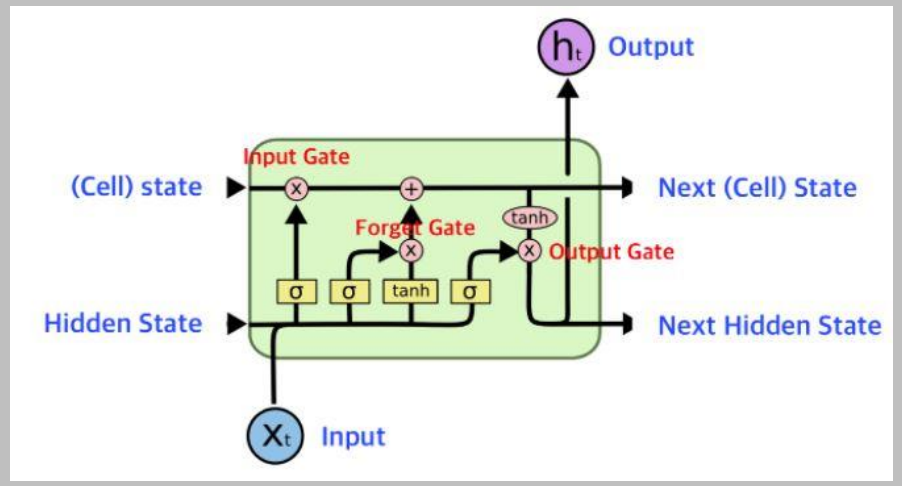
모델링(LSTM)

LSTM모델 (Long Short-Term Memory)

RNN 모델의 한 종류로 시계열, 문장 데이터와 같은 **연속적 데이터를 처리하는데 효과적인 모델**

기존 RNN 모델의 **장기 의존성 문제**를 해결하기 위해 설계

장기 의존성 문제 : 은닉층을 통해 과거 정보를 전달하지만 시간이 길어질수록 과거 정보가 소실된다!



이를 방지하기 위해 **4가지 게이트** 추가

Input Gate : 새로운 **데이터에서 무엇을 기억**할지 결정

Forget Gate: **기존 기억에서 무엇을 잃을지** 결정

Cell State: 입력 게이트와 망각 게이트를 거친 데이터를 조합하여 **장기 기억으로 저장**

Output Gate: 입력 게이트를 통해 들어온 정보 중 **중요한 정보만 단기기억**으로 만들어 다음 입력 게이트에 넘겨줌

모델링(LSTM)

LSTM모델 (Long Short-Term Memory) 사용법

적절하게 시퀀스를 나눠주고....

```
def sequenceCut(data, len_step):
    sequence = [] # 한 덩어리의 시퀀스가 들어가는 리스트
    labels = [] # 한 덩어리의 시퀀스의 정답이 들어가는 리스트
    for i in range(len(data) - len_step):
        sequence.append(data[i : i + len_step]) # i가 0이라면 train_data[0 : 5]까지의 값이 시퀀스 덩어리로 묶인다
        labels.append(data[i + len_step])# i가 0이라면 train_data[6]의 값이 정답지로 들어간다.

    return np.array(sequence), np.array(labels) # array 배열로 바로 만들어 바로 lstm 모델에 들어갈 수 있게 함
```

```
len_step = 5

X, y = sequenceCut(scaled_train, len_step)
```

적절하게 모델을 만들어 주면 된다.

```
# PM10 LSTM

model = Sequential()
model.add(LSTM(64, activation = 'tanh', return_sequences = True, input_shape = (len_step, 1)))
model.add(Dropout(0.2))
model.add(LSTM(32, return_sequences = False))
model.add(Dropout(0.2))
model.add(Dense(25))
model.add(Dense(y_train.shape[1]))
```

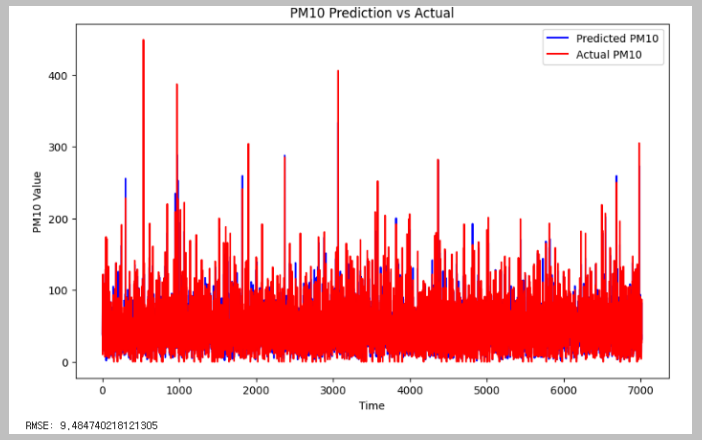
모델링(LSTM) – 결과 분석

LSTM모델 (Long Short-Term Memory) RMSE 값

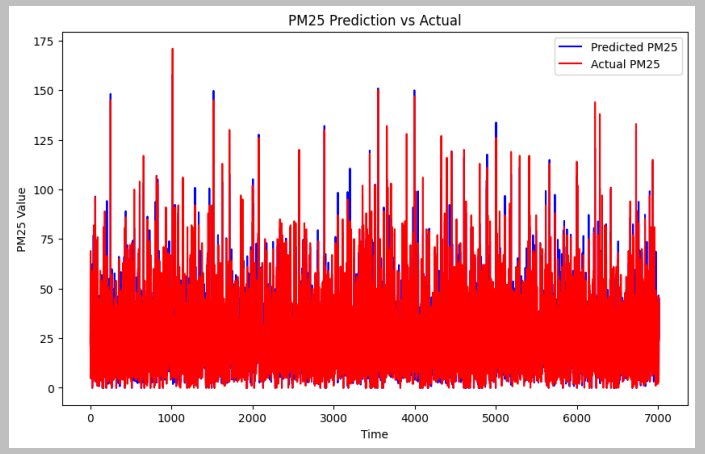
11개구 RMSE값

지역	PM10	PM2.5
강동구	9.48	5.66
강남구	8.59	5.83
동작구	7.09	4.59
광진구	7.45	5.09
강서구	9.89	5.64
마포구	7.34	5.15
서초구	12.03	5.86
성동구	10.93	6.27
송파구	9.09	5.97
영등포구	10.10	6.38
용산구	7.34	5.06

강동구 모델 평가



PM10 RMSE : 9.48

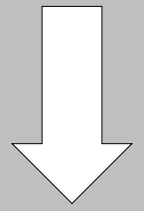


PM25 RMSE : 5.66

실제 데이터 예측

해당 모델을 활용한 강동구 미세먼지 수치는...?

19	202412301300	강동구	60	36
20	202412301400	강동구	59	34
21	202412301500	강동구	70	42
22	202412301600	강동구	68	38
23	202412301700	강동구	69	39
24	202412301800	강동구	65	41



	pred_pm10	pred_pm25
0	35.91074	52.039394

한계

1. 데이터 전처리의 단순화
   결측치를 **단순 평균으로 대체**했기 때문에 **계절적 요인과 같은 특성이 반영이 안될 가능성**이 크다.

2. 다양한 요소 간 상관관계 분석의 부족
   PM10과 Pm2.5의 연관성을 풍향과 강수량 등 **기후적 요인에서만 찾아 다른 요인들이 누락됐을 가능성**이 있다.

3. 다양한 요소 간 상관관계 분석의 부족
   PM10과 Pm2.5의 연관성을 풍향과 강수량 등 **기후적 요인에서만 찾아 다른 요인들이 누락됐을 가능성**이 있다.

4. 새로운 모델 활용(예측분류모델)
   이용자는 **구체적인 숫자보단** 미세먼지 분류표 상 **분류가 (좋음 보통 나쁨 매우나쁨) 중요한 정보**일 수 있다.