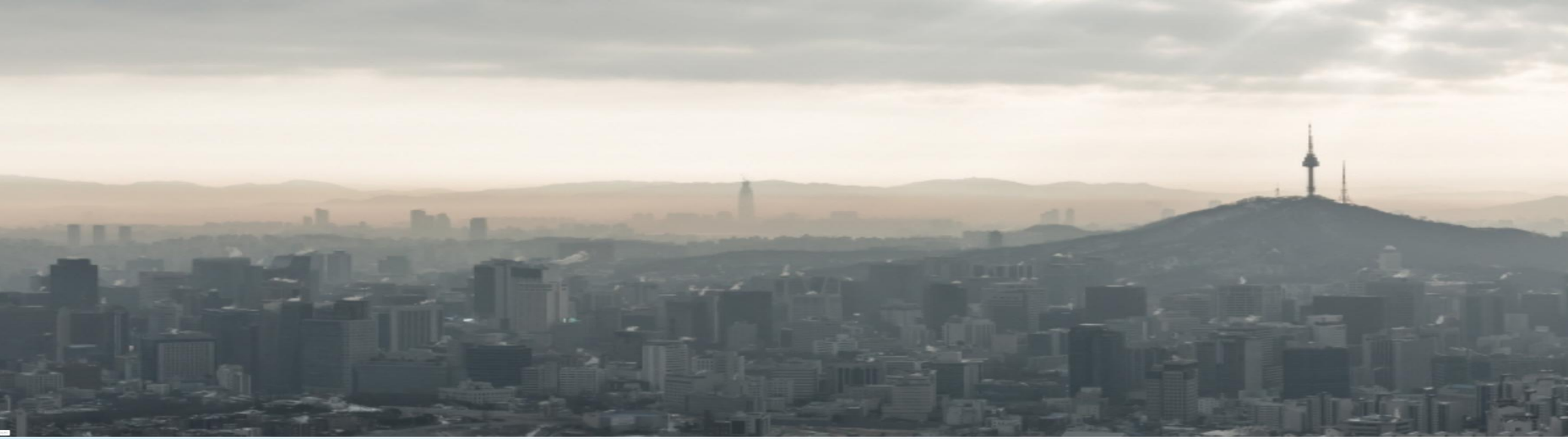


# 미세먼지 유발 영향인자 확인 및 개선방안 제시

# 분석 프로세스



| 프로세스 | 주요 내용 |
| --- | --- |
| 과제 정의 | • 미세먼지 영향 인자 분석과제 수행 이유<br>• 과제 분석 목표 |
| 분석 계획 | • 가설 설정<br>• 데이터 품질 확인<br>• 변수간 상관관계 분석 및 모델 수립 |
| 분석항목 정의 | • 변수 설명 및 형태 확인<br>• 변수 선택 |
| 데이터 현황 | • 결측치 확인 및 처리<br>• 이상치 확인 및 처리<br>• 요약통계량 확인 및 표준화 필요성 검토 |
| 모델링 | • 다중 선형회귀 모델링<br>• 의사결정나무 모델링<br>• 랜덤 포레스트 모델링<br>• 그래디언트 부스팅 모델링 |
| 예측결과 확인 | • 수립한 가설에 대한 결과 확인<br>• 최종 의견 정리 |
| 결론 및 대안제시 | • 미세먼지 발생량 감소 방안 제시<br>• 실습에 대한 고찰 |

# 과제 정의

## 미세먼지 발생량을 예측해야 하는 이유


- 2017년 **OECD**가 발표한 미세먼지(PM2.5) 농도 통계에서 한국이 25.14㎍/㎥으로 **회원국들 중 농도가 가장 높은 수준**인 것으로 조사
- OECD 회원국들의 평균으로 집계된 12.5㎍/㎥의 두 배가 넘는 수치이며 실제로 OECD에서는 한국의 대기오염도를 측정하며 이 결과가 **국민들의 건강과 직결될 것으로 예상**
- 2016년 OECD 보고서에서도 **한국 대기오염 심각성을 경고**
- 제대로 대처 못하는 경우 40년 뒤 **대기오염으로 인한 조기 사망률이 가장 높은 OECD 회원국으로 한국이 전망**

미세먼지 농도는 국민의 **건강**과 직결된 문제
따라서 해당 과제를 통해서 **어떤 인자가 미세먼지 농도에 영향을 주는지 파악**하여 미세먼지 농도를 감소시킬 수 있는 방안을 찾고자 함

# 과제 정의

## 분석 방향 및 목표설정

대기오염, 기상 상황, 계절 등이 미세먼지 발생량에 영향을 미침

미세먼지 발생에 영향을 미치는 요인을 분석하고 영향인자를 선정한 후 대응 방안을 수립함

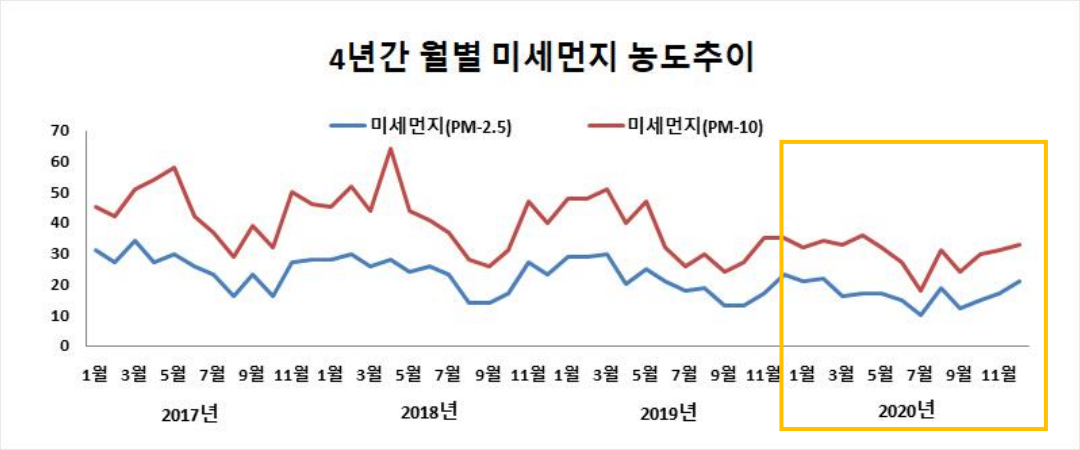
- 미세먼지 발생량과 영향인자들의 예상 관계

| 목표변수 | 대기오염 | | | 기상정보 | | | 고려대상 | |
|---|---|---|---|---|---|---|---|---|
| 미세먼지 발생량 | 이산화탄소 농도 | 아황산가스 농도 | 이산화질소 농도 | 기온 | 기압 | 습도 | 오존 농도 | 계절성 |
| 높음 ↕ 낮음 | 높음 ↕ 낮음 | 높음 ↕ 낮음 | 높음 ↕ 낮음 | 낮음 ↕ 높음 | 높음 ↕ 낮음 | 낮음 ↕ 높음 | 관련성 유무 확인 필요 | 겨울 ↕ 여름 |

# 분석 계획

## 가설 설정

**가설1. 대기오염이 미세먼지 발생량의 가장 주요한 영향인자이다.**




4년간 월별 미세먼지, 초미세먼지 농도
(2017~2020)

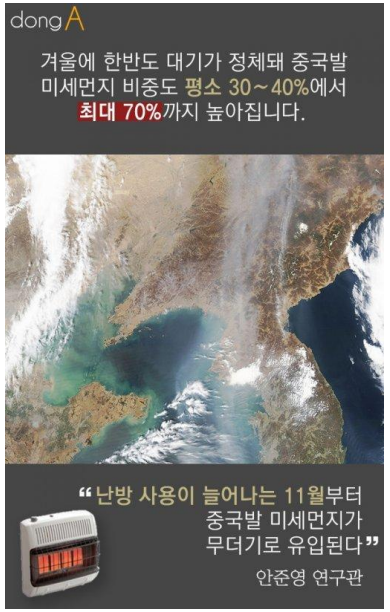

2020년, 코로나 19로 인한 대기오염 감소

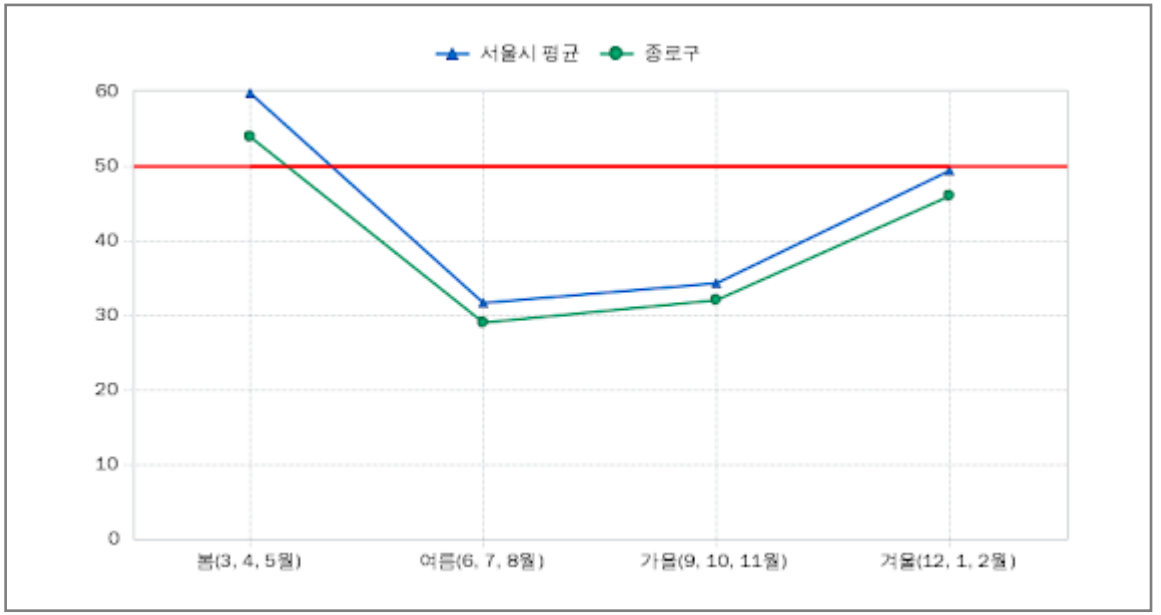
2020년은 코로나19로 인해 전세계적으로 **대기오염이 현저히 감소**함

2020년의 미세먼지 농도가 이전보다 상당히 감소한 것으로 보아
**대기오염이 미세먼지 발생량의 가장 주요한 영향인자일 것으로 예측**

가설 설정

## 가설2. 봄, 겨울철 미세먼지 발생량이 증가한다.



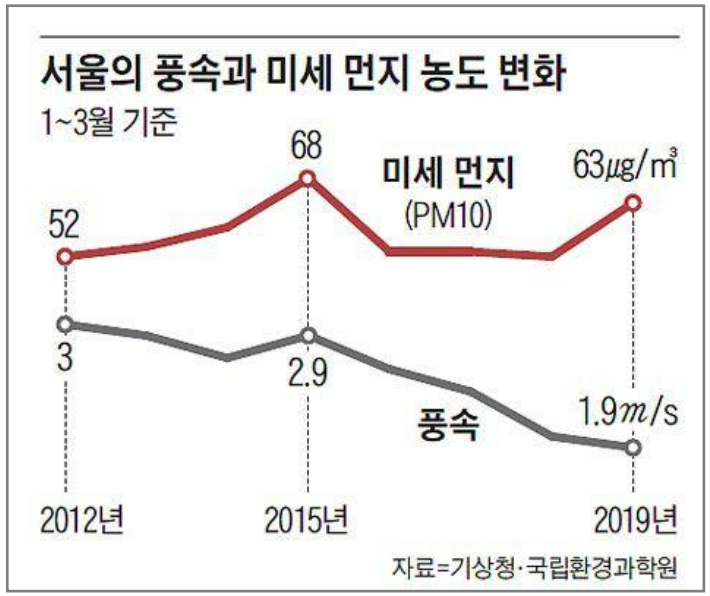



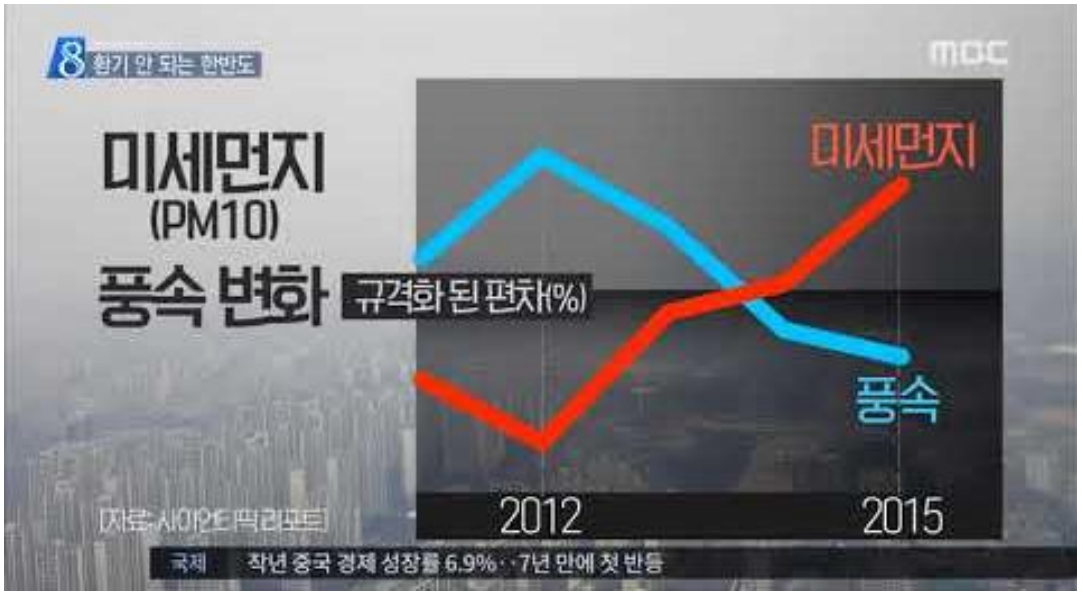

전반적으로 봄, 겨울철 미세먼지 발생량이 증가하는 것을 볼 수 있음

**계절은 미세먼지 발생량의 영향인자일 것으로 예상**됨

가설 설정

**가설3. 풍속이 약할수록 미세먼지 발생량은 증가한다.**




풍속이 약할수록 미세먼지 발생량이 증가함

풍속과 미세먼지 발생량 사이에 **음의 상관관계가 있을 것으로 예상**됨

# 분석 계획

데이터 분석 계획

## 데이터 품질 확인

- 결측치가 확인될 경우, 결측치가 해당 설명변수의 20% 이상 경우 해당 변수는 분석에서 제외
- 이상치가 확인될 경우, 최대한 제거하지 않고 분석함
- Scale의 차이가 크다면 정규화 진행

## 상관관계 분석

- 미세먼지 발생량과 대기오염 인자, 기상 상황, 계절별 상관관계가 있는지 분석
- 변수 간의 상관성 분석
- 다중공선성이 확인될 경우 모델 선정에 유의

## 모델 수립

- 회귀분석, Decision Tree, Random Forest, Gradient Boosting 모델을 각각 생성하여 목표변수 설명력을 확인함
- 최적의 모델 선택

# 분석 계획

데이터 분석 과정

# 분석항목 정의

## 변수 설명 및 형태

| 변수 | 변수 설명 | 변수 역할 | 변수 형태 |
|---|---|---|---|
| **MesDate** | 측정일자 | 제외 | 연속형 |
| **PM10** | 미세먼지 10㎍/㎥ | 목표변수 | 연속형 |
| **03** | 오존 농도 | 설명변수 | 연속형 |
| **NO2** | 이산화질소 농도 | 설명변수 | 연속형 |
| **CO** | 일산화탄소 농도 | 설명변수 | 연속형 |
| **SO2** | 아황산가스 농도 | 설명변수 | 연속형 |
| **TEMP** | 기온(℃) | 설명변수 | 연속형 |
| **RAIN** | 강수량(mm) | 설명변수 | 연속형 |
| **WIND** | 풍속(m/s) | 설명변수 | 연속형 |
| **WIND_DIR** | 풍향(16방위) | 설명변수 | 연속형 |
| **HUMIDITY** | 습도(%) | 설명변수 | 연속형 |
| **ATM_PRESS** | 현지기압(hPa) | 설명변수 | 연속형 |
| **SNOW** | 적설(cm) | 설명변수 | 연속형 |
| **CLOUD** | 전운량(10분위) | 설명변수 | 연속형 |

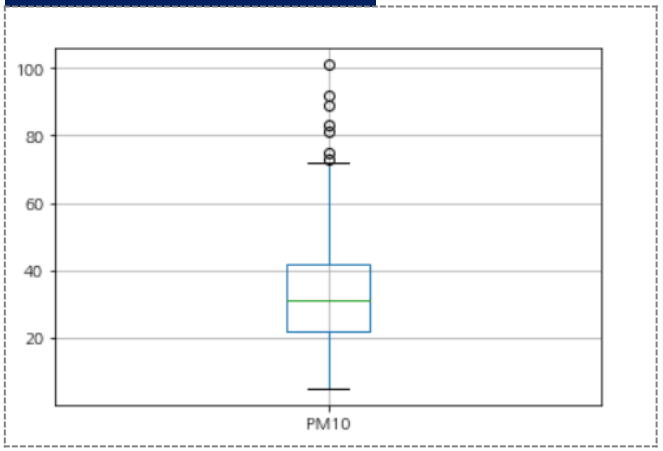
# 데이터 현황

- 결측치 확인 및 처리
  - 결측치 확인 결과 PM10, 03, NO2, SO2는 각각 1개의 결측치가 발생하였으며 해당 변수들 모두 시계열 데이터가 아니므로 평균치로 대체함
  - CO는 366개 중 55개의 결측치가 발생하였는데 앞서 세웠던 계획에 비추어 결측치가 해당 변수의 20% 이하이므로 제거하지 않고 변수들의 평균치로 대체

| | |
|---|---|
| PM10 | 1 |
| 03 | 1 |
| NO2 | 1 |
| CO | 55 |
| SO2 | 1 |

- 이상치 확인 및 처리
  - 목표변수 PM10에서 일부 이상치가 발견됨

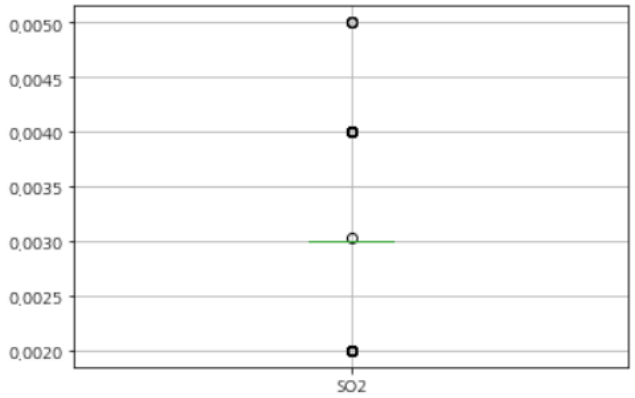
목표변수

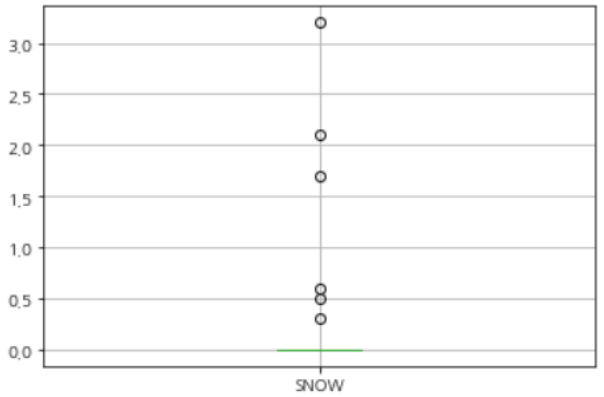


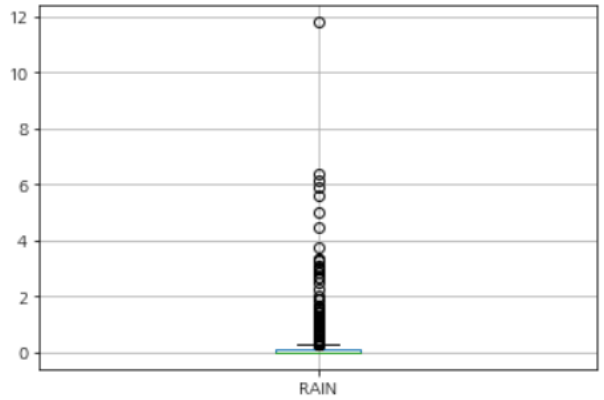
# 데이터 현황

- 이상치 확인 및 처리
  - 설명변수 중 7개의 변수에 이상치가 존재하며 SNOW, RAIN 변수에서 이상치가 다수 발견됨
  - 해당 변수들을 살펴본 결과 기상 상황과 계절에 따라 달라지는 값으로 목표변수에 유의한 값으로 작용할 수 있기에 제거하지 않음

설명변수



# 데이터 현황

- 요약통계량 확인
  - 요약통계량과 변동계수 확인결과 변수의 표준화가 필요하다고 판단

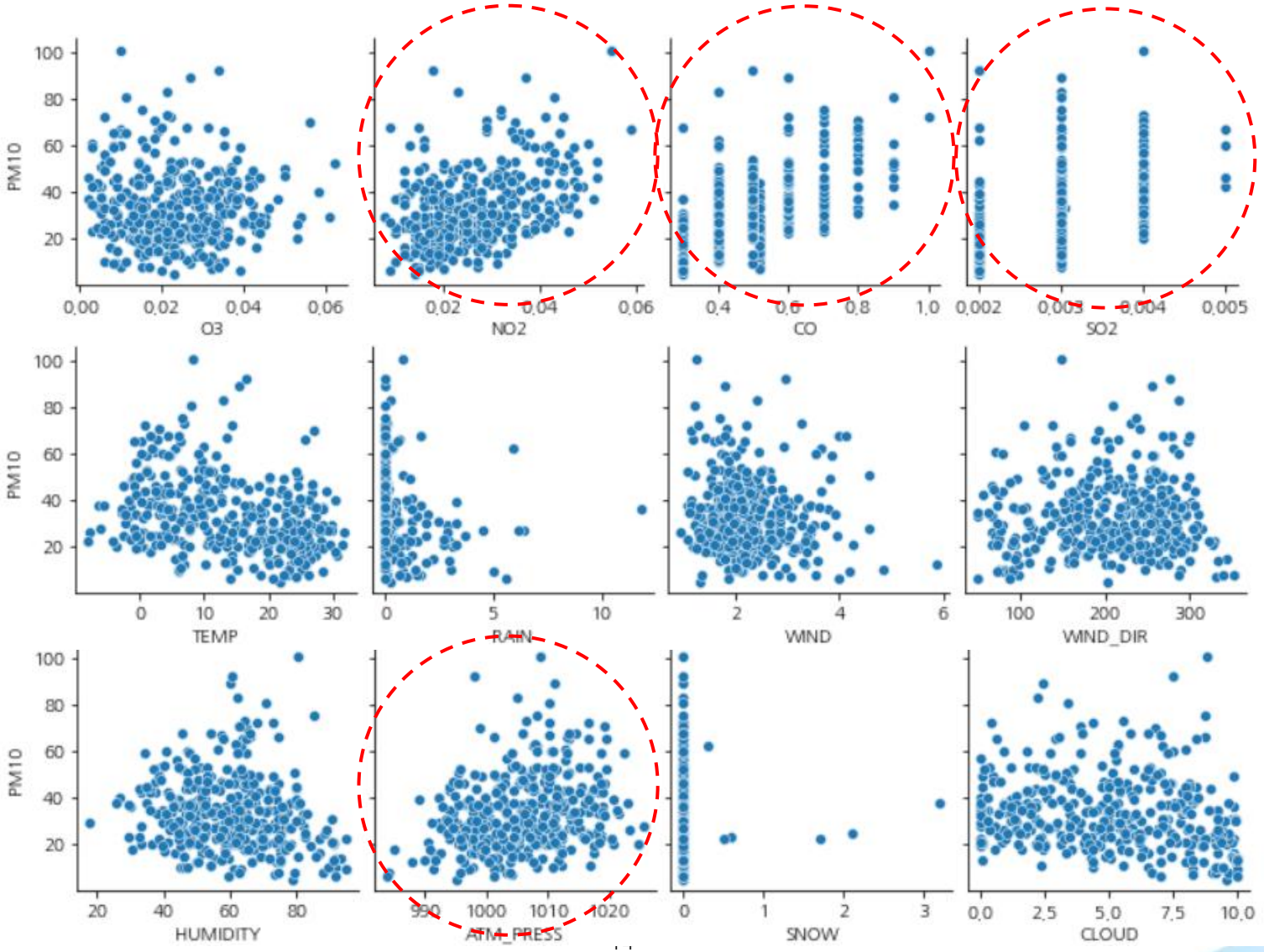
| | PM10 | O3 | NO2 | CO | SO2 | TEMP | RAIN | WIND | WIND_DIR | HUMIDITY | ATM_PRESS | SNOW |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| count | 366.000000 | 366.00000 | 366.000000 | 366.000000 | 366.000000 | 366.000000 | 366.000000 | 366.000000 | 366.000000 | 366.000000 | 366.000000 | 366.000000 |
| mean | 33.421918 | 0.02360 | 0.026814 | 0.517042 | 0.003033 | 13.863798 | 0.381639 | 2.225301 | 209.450820 | 60.295082 | 1005.848907 | 0.022951 |
| std | 15.916135 | 0.01188 | 0.010257 | 0.140836 | 0.000632 | 9.830280 | 1.122127 | 0.723171 | 70.735018 | 14.534983 | 8.126823 | 0.222361 |
| min | 5.000000 | 0.00200 | 0.008000 | 0.300000 | 0.002000 | -7.950000 | 0.000000 | 0.940000 | 50.000000 | 17.900000 | 983.800000 | 0.000000 |
| 25% | 22.000000 | 0.01425 | 0.019000 | 0.400000 | 0.003000 | 5.492500 | 0.000000 | 1.722500 | 160.250000 | 49.650000 | 999.400000 | 0.000000 |
| 50% | 31.000000 | 0.02300 | 0.025000 | 0.500000 | 0.003000 | 14.000000 | 0.000000 | 2.095000 | 221.000000 | 61.050000 | 1006.450000 | 0.000000 |
| 75% | 42.000000 | 0.03200 | 0.033750 | 0.600000 | 0.003000 | 23.070000 | 0.115000 | 2.620000 | 266.000000 | 69.950000 | 1011.575000 | 0.000000 |
| max | 101.000000 | 0.06200 | 0.059000 | 1.000000 | 0.005000 | 31.720000 | 11.800000 | 5.880000 | 351.000000 | 95.000000 | 1025.500000 | 3.200000 |

```
PM10        0.476218
O3          0.503378
NO2         0.382514
CO          0.272388
SO2         0.208251
TEMP        0.709061
RAIN        2.940282
WIND        0.324977
WIND_DIR    0.337717
HUMIDITY    0.241064
ATM_PRESS   0.008080
SNOW        9.688587
CLOUD       0.589674
dtype: float64
```

<- 변동계수

# 탐색적 분석

- 그래프 분석(목표변수 PM10과 설명변수 간 산점도)
  - NO2, CO, SO2, ATM_PRESS 변수가 영향인자로 예상됨



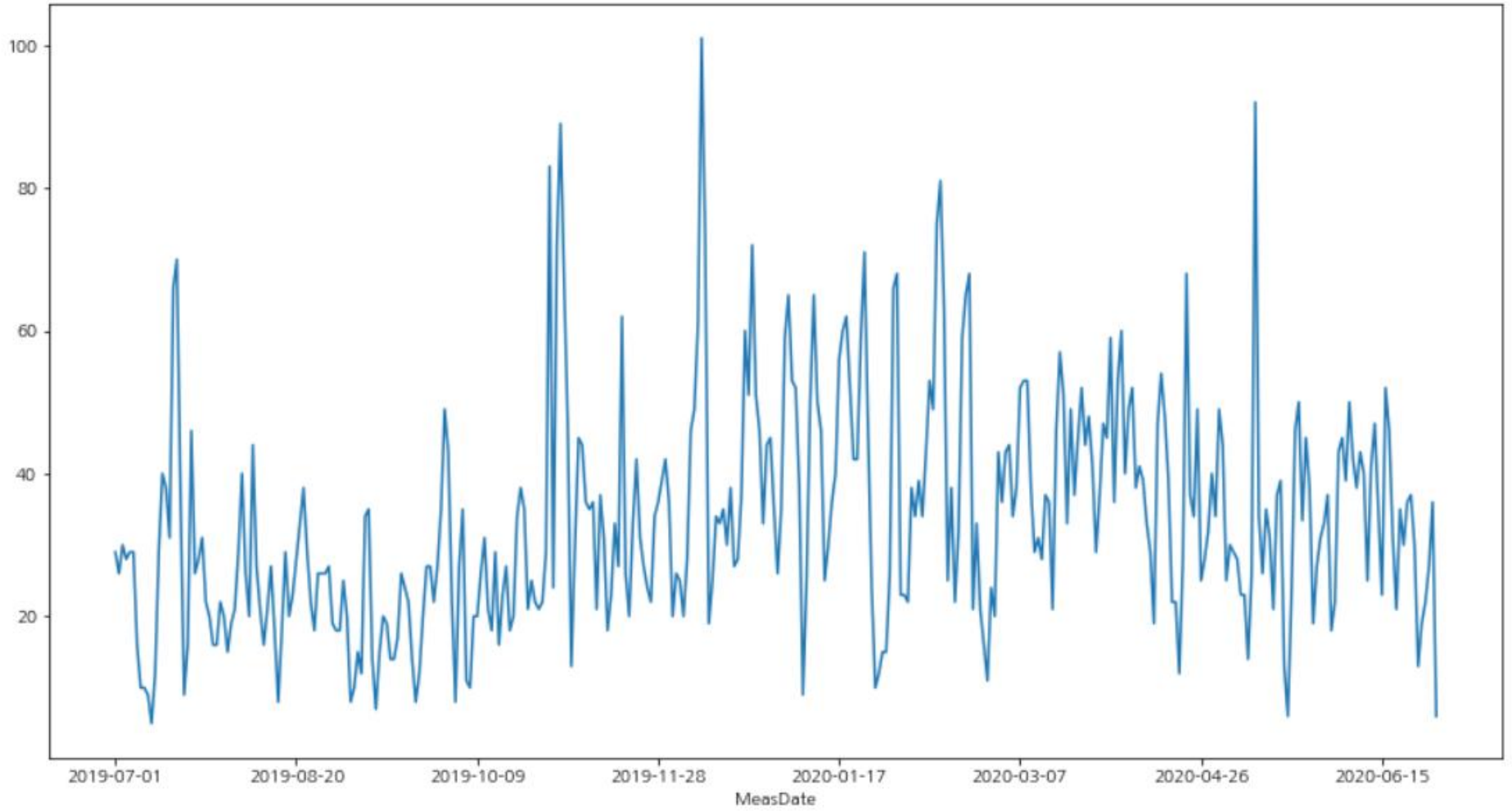
# 탐색적 분석

- 상관관계 분석
  - PM10과 CO, SO2 변수가 선형적 관계가 있는 것으로 판단됨
  - CO와 NO2를 비롯한 다수의 설명변수 간의 선형관계가 확인되어 다중공선성 확인 및 처리가 필요함

| | PM10 | O3 | NO2 | CO | SO2 | TEMP | RAIN | WIND | WIND_DIR | HUMIDITY | ATM_PRESS | SNOW | CLOUD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PM10 | 1.000 | -0.052 | 0.396 | 0.548 | 0.429 | -0.310 | -0.121 | -0.100 | 0.020 | -0.149 | 0.253 | -0.020 | -0.172 |
| O3 | -0.052 | 1.000 | -0.592 | -0.509 | -0.234 | 0.516 | -0.104 | 0.165 | 0.269 | -0.038 | -0.534 | 0.004 | -0.119 |
| NO2 | 0.396 | -0.592 | 1.000 | 0.786 | 0.563 | -0.237 | 0.029 | -0.536 | -0.408 | -0.065 | 0.420 | -0.121 | 0.017 |
| CO | 0.548 | -0.509 | 0.786 | 1.000 | 0.559 | -0.340 | 0.037 | -0.412 | -0.319 | 0.057 | 0.385 | -0.056 | 0.037 |
| SO2 | 0.429 | -0.234 | 0.563 | 0.559 | 1.000 | -0.274 | -0.129 | -0.253 | -0.093 | -0.301 | 0.334 | -0.103 | -0.191 |
| TEMP | -0.310 | 0.516 | -0.237 | -0.340 | -0.274 | 1.000 | 0.078 | -0.215 | -0.050 | 0.404 | -0.792 | -0.185 | 0.342 |
| RAIN | -0.121 | -0.104 | 0.029 | 0.037 | -0.129 | 0.078 | 1.000 | 0.128 | -0.181 | 0.399 | -0.237 | 0.019 | 0.360 |
| WIND | -0.100 | 0.165 | -0.536 | -0.412 | -0.253 | -0.215 | 0.128 | 1.000 | 0.236 | -0.080 | -0.056 | 0.145 | 0.019 |
| WIND_DIR | 0.020 | 0.269 | -0.408 | -0.319 | -0.093 | -0.050 | -0.181 | 0.236 | 1.000 | -0.096 | 0.066 | 0.108 | -0.294 |
| HUMIDITY | -0.149 | -0.038 | -0.065 | 0.057 | -0.301 | 0.404 | 0.399 | -0.080 | -0.096 | 1.000 | -0.512 | 0.021 | 0.629 |
| ATM_PRESS | 0.253 | -0.534 | 0.420 | 0.385 | 0.334 | -0.792 | -0.237 | -0.056 | 0.066 | -0.512 | 1.000 | 0.040 | -0.431 |
| SNOW | -0.020 | 0.004 | -0.121 | -0.056 | -0.103 | -0.185 | 0.019 | 0.145 | 0.108 | 0.021 | 0.040 | 1.000 | 0.021 |
| CLOUD | -0.172 | -0.119 | 0.017 | 0.037 | -0.191 | 0.342 | 0.360 | 0.019 | -0.294 | 0.629 | -0.431 | 0.021 | 1.000 |

# 탐색적 분석

- 그래프 분석(목표변수 PM10과 측정일자 간 Trend Chart)
  - 측정일자에 따른 감소 구간과 증가 구간 등 추세선이 보임
  - 계절이라는 범주형 파생변수를 생성하여 계절과 미세먼지 발생량 간의 관계를 살펴봄

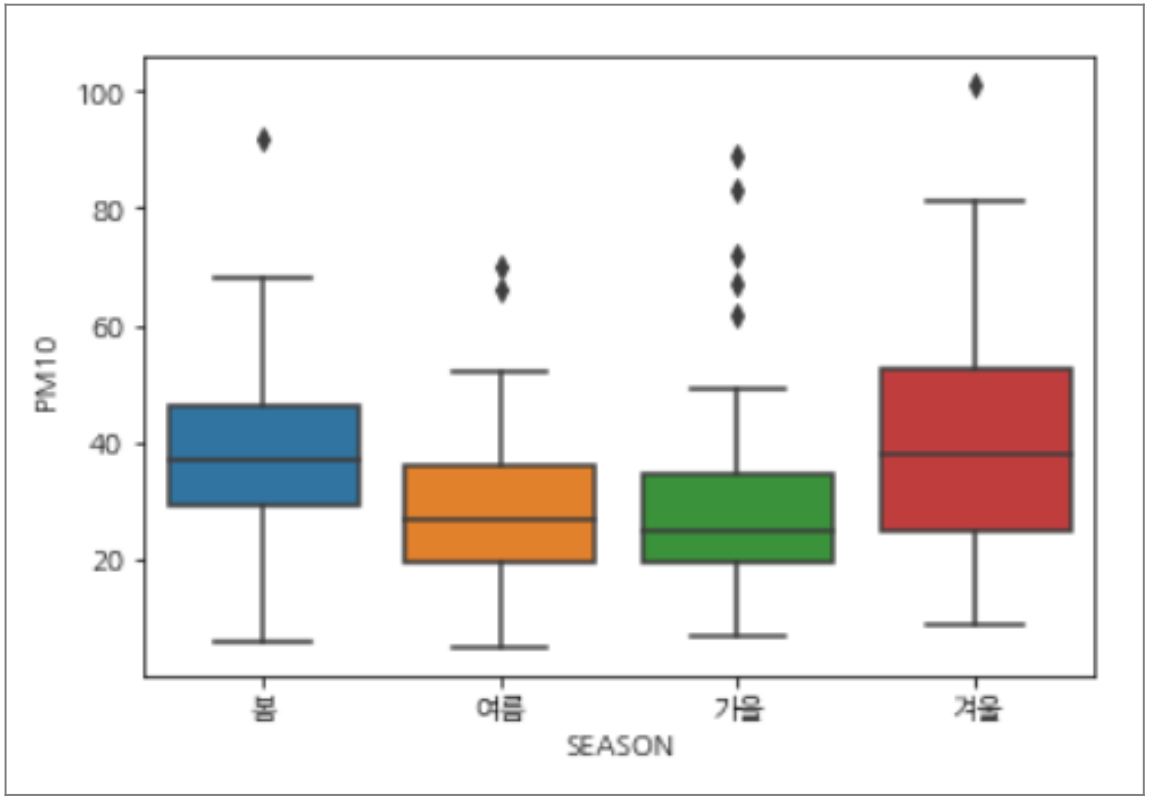


- 파생변수 계절과 미세먼지 발생량의 Box Plot 확인
  - Box Plot으로 계절 간 미세먼지 발생량의 차이를 봤을 때, 봄, 겨울철에 상대적으로 미세먼지 발생량이 높은 것으로 보임

```
1 df_raw['SEASON'].unique()
```

```
array(['여름', '가을', '겨울', '봄'], dtype=object)
```

* 일반적인 계절분류에 따라 봄(3, 4, 5월), 여름(6, 7, 8월), 가을(9, 10, 11월), 겨울(12, 1, 2월)로 분류함



# 탐색적 분석

- 계절별 미세먼지 발생량의 차이 여부 검정
  - 앞선 Box Plot을 통해 육안으로 확인한 차이의 유의성을 검정하기 위해 ANOVA를 실행
  - 유의수준 5%에서 검정 결과 p-value가 0으로 계절별 미세먼지 발생량의 차이가 있다고 할 수 있음.

```
One-way ANOVA
F검정통계량 : 17.838
p-value : 0.0
```

- 그래프 분석결과 종합
  - PM10(미세먼지 발생량)의 영향인자 : CO, NO2
  - PM10(미세먼지 발생량)과 계절 간의 상관관계 확인 : 봄, 겨울철 미세먼지 발생량 증가

# 모델링

- 미세먼지 발생량 예측 및 영향인자 확인
  - 모든 설명변수에 대해 VIF가 10이하이므로 설명변수간 독립이 확인

| | variable | VIF |
|---|---|---|
| 11 | SNOW | 1.112 |
| 6 | RAIN | 1.348 |
| 8 | WIND_DIR | 1.501 |
| 4 | SO2 | 1.941 |
| 7 | WIND | 1.943 |
| 12 | CLOUD | 2.093 |
| 9 | HUMIDITY | 2.659 |
| 1 | O3 | 2.691 |
| 3 | CO | 3.577 |
| 5 | TEMP | 4.498 |
| 10 | ATM_PRESS | 4.639 |
| 2 | NO2 | 4.937 |
| 0 | const | 73600.335 |

# 모델링

- 1. 다중 선형회귀 분석
  - 모든 설명변수에 대한 다중 선형 회귀모델 결과
  - F검정 결과 p-value 값이 유의수준 0.05보다 작으므로 회귀모델로서 유의하며 52.7%의 설명력을 가지는데 기상 현상 예측이라는 상황에 비춰 봤을 때 변수를 조절하더라도 더 높은 설명력을 기대하기는 어려움
  - 회귀모델의 기본 가정인 잔차의 정규성과 독립성이 위배되는 것으로 보임

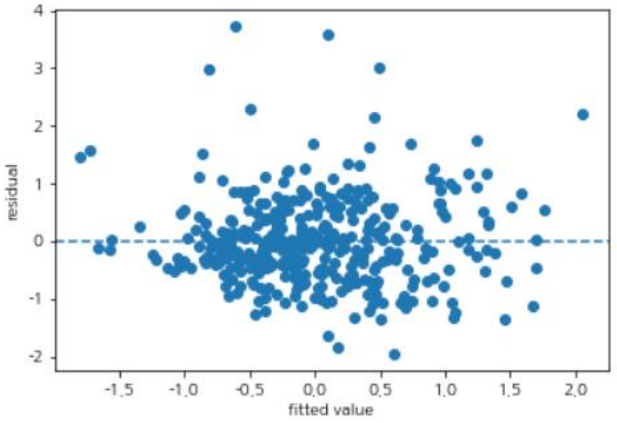
OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | PM10 | R-squared: | 0.548 |
| Model: | OLS | Adj. R-squared: | 0.527 |
| Method: | Least Squares | F-statistic: | 26.43 |
| Date: | Fri, 05 Mar 2021 | Prob (F-statistic): | 1.23e-50 |
| Time: | 12:38:48 | Log-Likelihood: | -1386.4 |
| No. Observations: | 366 | AIC: | 2807. |
| Df Residuals: | 349 | BIC: | 2873. |
| Df Model: | 16 | | |
| Covariance Type: | nonrobust | | |

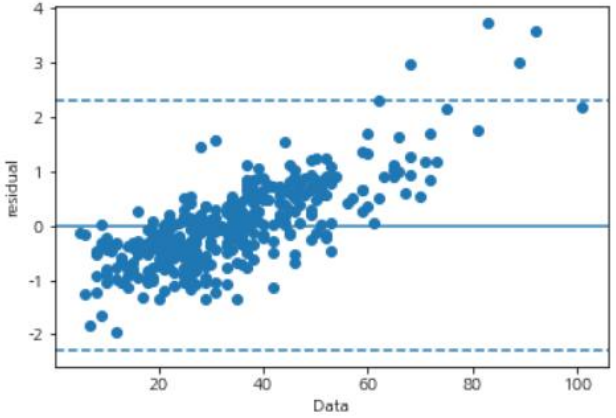
| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 1.5300 | 175.416 | 0.009 | 0.993 | -343.476 | 346.536 |
| SEASON[T.겨울] | 4.7615 | 2.774 | 1.717 | 0.087 | -0.694 | 10.217 |
| SEASON[T.봄] | 11.7779 | 2.447 | 4.814 | 0.000 | 6.966 | 16.590 |
| SEASON[T.여름] | 6.2809 | 2.607 | 2.409 | 0.017 | 1.153 | 11.409 |
| O3 | 457.0338 | 83.267 | 5.489 | 0.000 | 293.266 | 620.802 |
| NO2 | 558.7148 | 132.486 | 4.217 | 0.000 | 298.143 | 819.287 |
| CO | 62.5162 | 8.028 | 7.787 | 0.000 | 46.726 | 78.306 |
| SO2 | 1342.9917 | 1285.042 | 1.045 | 0.297 | -1184.408 | 3870.392 |
| TEMP | -0.2826 | 0.162 | -1.746 | 0.082 | -0.601 | 0.036 |
| RAIN | -0.9162 | 0.602 | -1.521 | 0.129 | -2.101 | 0.268 |
| WIND | 2.9843 | 1.183 | 2.522 | 0.012 | 0.657 | 5.311 |
| WIND_DIR | 0.0487 | 0.011 | 4.619 | 0.000 | 0.028 | 0.069 |
| HUMIDITY | 0.0393 | 0.065 | 0.602 | 0.548 | -0.089 | 0.168 |
| ATM_PRESS | -0.0492 | 0.171 | -0.288 | 0.774 | -0.385 | 0.287 |
| SNOW | -0.9422 | 2.775 | -0.340 | 0.734 | -6.400 | 4.516 |
| CLOUD | -0.0767 | 0.283 | -0.272 | 0.786 | -0.632 | 0.479 |
| Month | -0.1609 | 0.227 | -0.708 | 0.480 | -0.608 | 0.286 |

| | | | |
|---|---|---|---|
| Omnibus: | 149.235 | Durbin-Watson: | 1.329 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 696.837 |
| Skew: | 1.709 | Prob(JB): | 4.83e-152 |
| Kurtosis: | 8.832 | Cond. No. | 2.31e+06 |

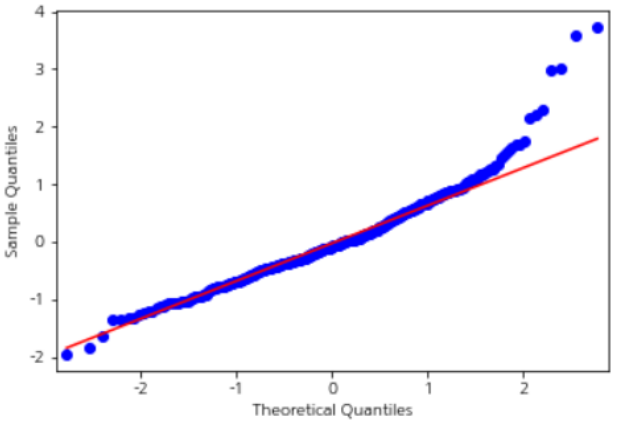
- 1. 다중 선형회귀 분석
  - 잔차분석결과 등분산성, 독립성, 정규성 모두 위배되는 것으로 보임




**등분산성**



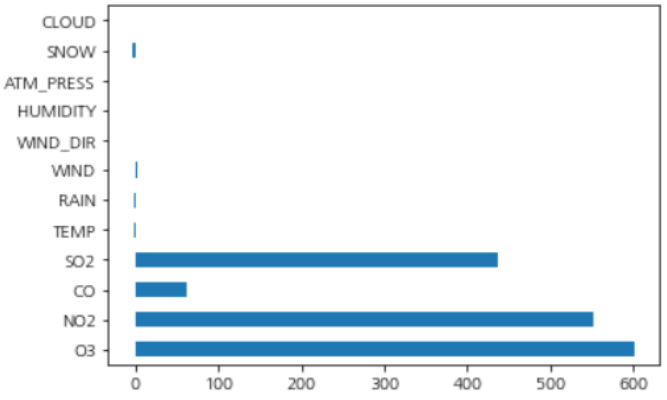

**독립성**



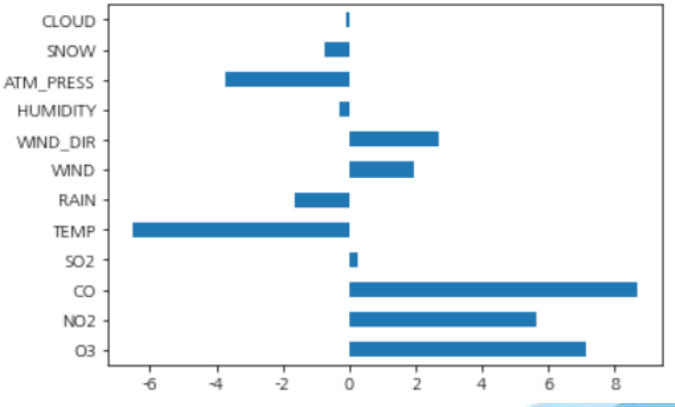

**정규성**

변수 중요도

- 변수 별 단위차이가 커 표준화하기 전 변수 중요도와 표준화 후 변수 중요도의 차이가 큼
- 표준화 전과 후 가장 차이가 많이 난 변수는 SO2와 CO이며 NO2와 O3는 중요한 영향인자로 판단됨




표준화



# 모델링

## 1. 다중 선형회귀분석

- 표준화한 모든 변수들로 다중 선형 회귀모델을 생성한 결과 수정결정계수는 52.7%가 나옴
- 표준화한 변수들의 변수 중요도 상위 5개로 다중 선형회귀 모델을 생성한 결과 수정결정계수는 45%가 나옴

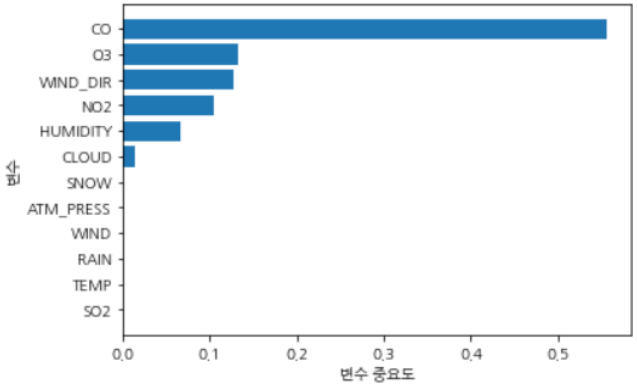
## 최종 회귀식 도출결과

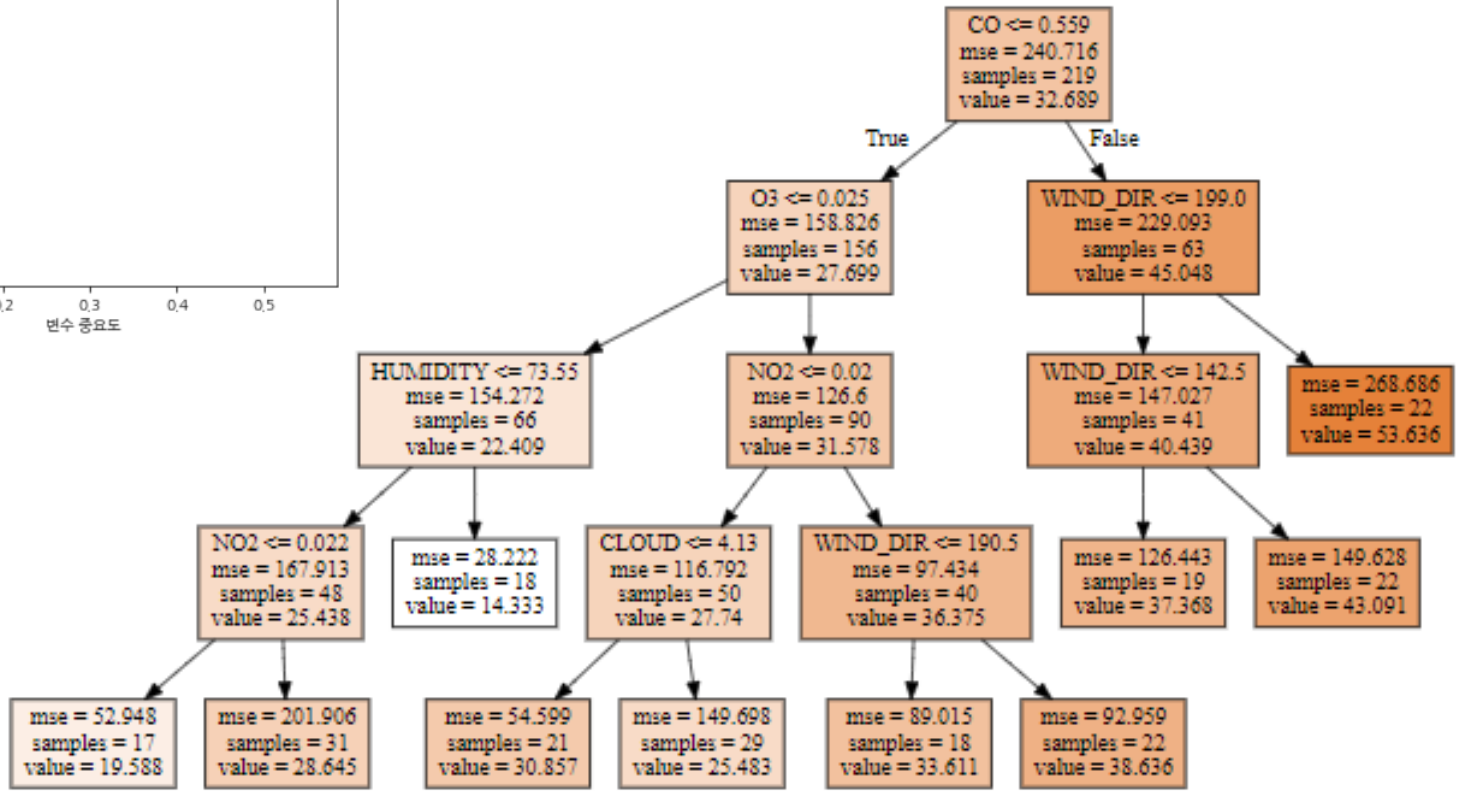
- 표준화 전 모든 변수로 도출한 회귀식
- PM10 = 433.3944 + 600.8265*O3 + 552.1572*NO2 + 61.6154*CO + 436.2575*SO2 – 0.6635*TEMP – 1.4370*RAIN + 2.6891*WIND + 0.0382*WIND_DIR – 0.0211*HUMIDITY –0.4622*ATM_PRESS – 3.2925* SNOW –0.0293*CLOUD

- 표준화 후 모든 변수로 도출한 회귀식
- PM10 = 33.4219 + 7.1279*O3 + 5.6555*NO2 + 8.6658*CO + 0.2752*SO2 – 6.5130*TEMP – 1.6103*RAIN + 1.9420*WIND + 2.7017*WIND_DIR - 0.3064*HUMIDITY – 3.7510*ATM_PRESS - 0.7311* SNOW – 0.0873*CLOUD

- 표준화한 변수 중 상위 5개 변수로 도출한 회귀식
- PM10 = 33.4219 + 8.2098*O3 + 3.7738*NO2 + 8. 2786*CO – 6. 9032*TEMP – 31.8342*ATM_PRESS

- 2. 의사결정나무 분석
  - 최적의 파라미터: min_samples_leaf = 17, max_depth=9, min_samples_split=10
  - Training data에 대한 설명력: 46.1%, Test data에 대한 설명력 14.6%
  - 변수 중요도 순위: CO > 03 > WIND_DIR > NO2

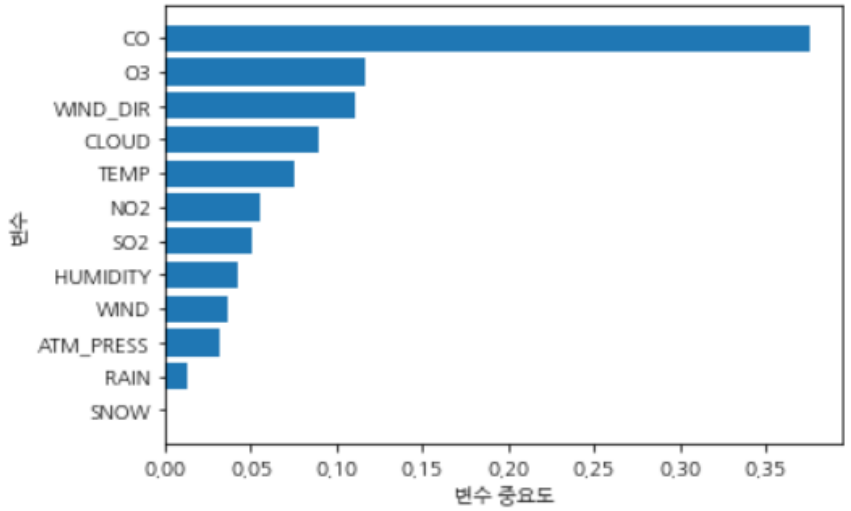


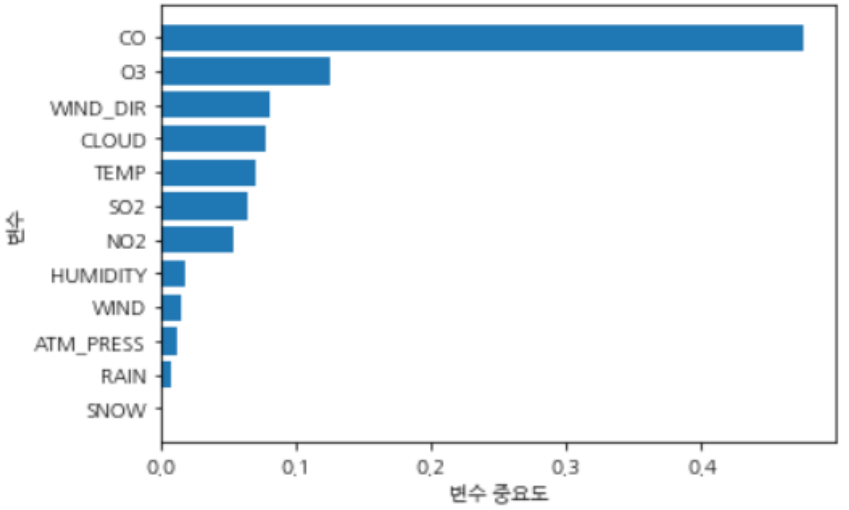
- 3. 랜덤 포레스트 분석
  - 최적의 파라미터: min_samples_leaf = 3, min_samples_split = 4, max_depth = 7, n_estimators = 100
  - Training data에 대한 설명력: 81.0%, Test data에 대한 설명력 39.3%
  - 변수 중요도 순위: CO > 03 > WIND_DIR > CLOUD > TEMP

- 4. 그래디언트 부스팅 분석
  - 최적의 파라미터: min_samples_leaf = 4, min_samples_split = 22, max_depth = 1, learning_rate = 0.3 , n_estimators = 100
  - Training data에 대한 설명력: 71.5%, Test data에 대한 설명력 38.6%
  - 변수 중요도 순위: CO > 03 > WIND_DIR > CLOUD >TEMP

**랜덤 포레스트 변수 중요도**

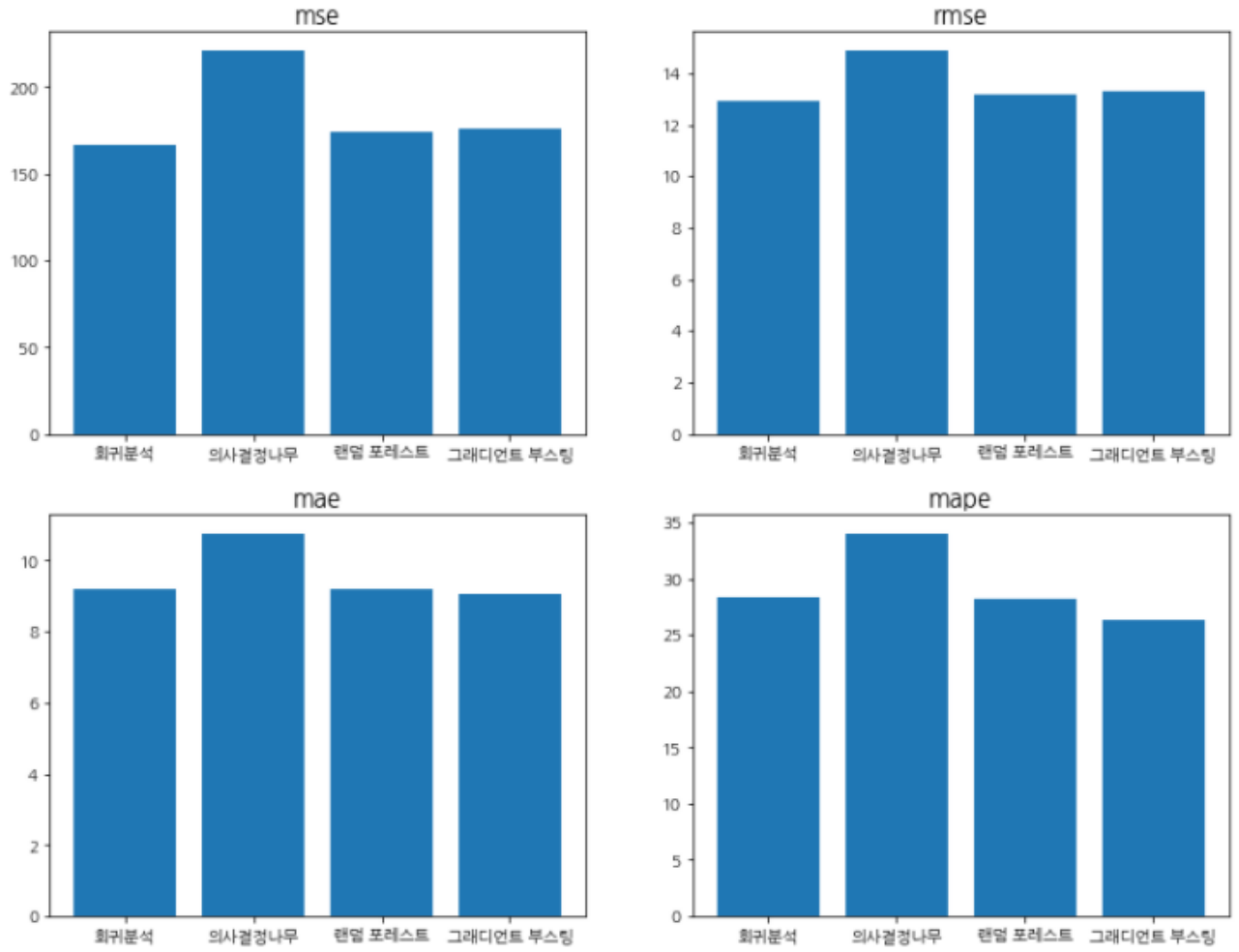



**그래디언트 부스팅 변수 중요도**



# 모델링 & 요약

- 5. 모델 평가
  - 전반적으로 의사결정나무의 error가 가장 크고 나머지 모델들은 비슷한 것으로 보임
  - 앞선 test score에서 의사결정나무의 score가 가장 좋지 않았는데 그 결과와 일치하는 것으로 보임



# 모델링 & 요약

- 3가지 가설에 대한 결과

**가설1. 대기오염이 미세먼지 발생량의 가장 주요한 영향인자이다.**

- 미세먼지 발생량의 주요 영향 인자 5가지(CO, NO2, O3, WIND_DIR, CLOUD ) 중 3가지가 대기오염과 관련된 인자이다.
- 따라서 해당 가설은 맞다고 보여진다.

**가설2. 봄, 겨울철 미세먼지 발생량이 증가한다.**

- 계절이라는 파생변수를 생성하여 Box Plot으로 봄, 겨울철에 상대적으로 미세먼지 발생량이 높은 것을 확인
- 이 차이가 다른 계절에 비해 유의한 차이인지 검정하기 위해 ANOVA를 진행 -> 유의한 차이임이 밝혀짐
- 따라서 해당 가설 역시 옳다고 판단된다.

**가설3. 풍속이 약할수록 미세먼지 발생량은 증가한다.**

- 산점도와 상관계수로부터 풍속과 미세먼지 발생량이 음의 상관성이 있다는 것을 파악하였다.
- 하지만 상관계수를 유의수준 0.05에서 관찰했을 때 그 값이 유의미하다고 판단하기에 애매함이 있다.
- 따라서 해당 가설은 어느 정도는 맞다고 볼 수 있지만 더 많은 데이터로 다시 살펴보아야 할 필요성이 있다.

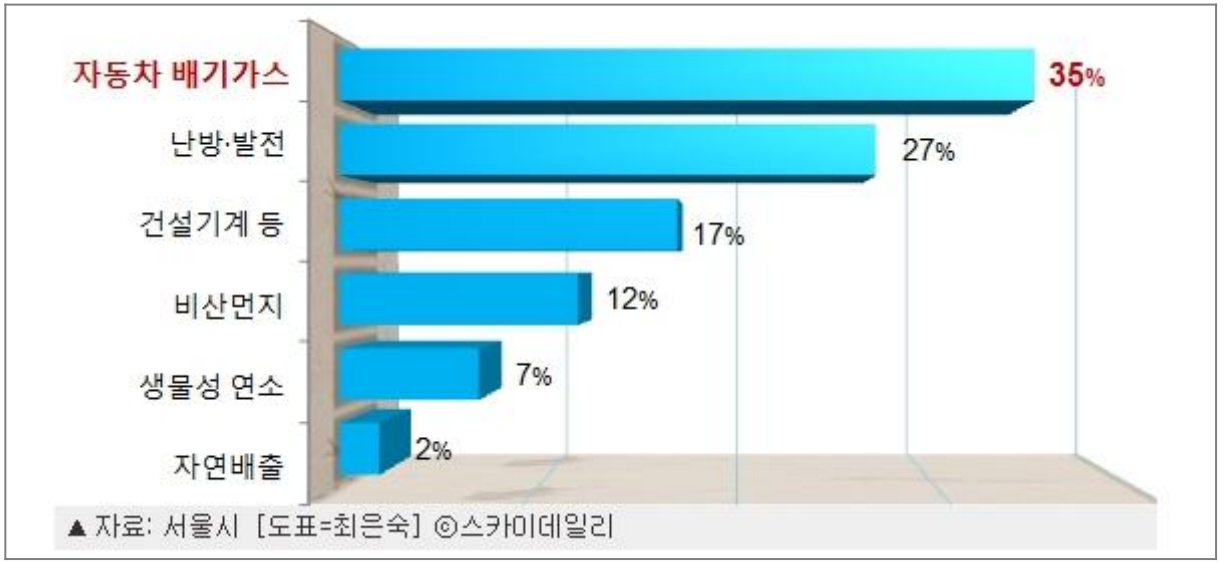
```
Correlation Analysis
미세먼지 발생량과 풍속 상관계수: -0.100
p-value: 0.055
```

# 모델링 & 요약

- 최종 의견
  - 주어진 데이터로 미세먼지 발생량과 설명변수들(대기오염, 기상상황, 계절) 간의 상관성을 알아보고 해당 변수들로 미세먼지 발생량을 예측하는 분석을 진행함
  - 모델 생성 결과 설명변수와 목표변수 간의 상관성은 알아낼 수 있었지만 주어진 데이터로 **미세먼지 발생량을 예측하기는 어렵다고 판단**됨
  - 4가지 모델(다중선형회귀, 의사결정나무, 랜덤포레스트, 그래디언트 부스팅)에 대해 모델링 해보았지만 **미세먼지 발생량 예측에 대해 유의한 설명을 하는 모델을 찾지 못함**
  - 그래프 분석과 모델을 통해 종합한 미세먼지 발생량에 대한 대표적인 **영향인자로는 CO, NO2, O3, WIND_DIR, CLOUD** 가 있음
  - 미세먼지 발생량을 예측하는 모델을 찾지는 못했지만 **초기에 설정했던 과제 수행 목적인 미세먼지 발생량에 대한 영향인자를 찾았다는 점에서 소기의 목적은 달성**했다고 볼 수 있음
  - 또한 3가지 **가설에 대한 검증 역시 원했던 방향으로 진행**되었고 두 가지 가설은 참으로 나머지 한 가지는 더 많은 데이터로 재검정이 필요할 것으로 판단
  - **결과적으로, 미세먼지 발생량을 감소시키기 위해서는 대기오염 인자에 대한 컨트롤이 필요하며 앞서 모델링 한 4가지 모델들 모두 현업에서 미세먼지 발생량 예측에 적용하기는 어려울 것으로 판단됨. 또한 더 정확한 예측을 위해서는 한반도 외 전체적인 기압의 흐름, 중국발 미세먼지의 영향 등 더 많은 인자들을 활용하여 복합적인 설명이 필요함**

# 모델링 & 요약

- 대안 제시
  - 미세먼지 발생의 **주 영향인자는 대기오염 인자**들이며 이 대기오염은 다시 6가지 주요 원인들에 의해 발생함




**자동차 배기가스와 난방 및 발전이 대기오염의 50% 이상을 차지하므로 이 요인들을 해결한다면 미세먼지 발생량 개선에도 연쇄적으로 영향을 줄 것으로 기대됨**

- 특히 차량 배기가스의 경우, 전기차, 수소차 보급이 증가되면 미세먼지 개선 효과가 높을 것으로 예상됨
- 따라서 **배기량이 많은 화물차에 대한 규제와 미래차 보조금 지급 확대 및 검토가 필요**
- 난방 및 발전의 경우 화학 에너지 사용에 의한 대기오염이 심한 것으로 파악되며 **신재생 에너지에 대한 투자**가 확대되어야 함

실습을 진행하기 전에는 주어진 데이터들로부터 미세먼지 발생량을 예측할 수 있을 것으로 예상했는데 첫 번째 모델을 돌려본 후 test score를 보고 제 생각이 틀렸다는 것을 알았습니다.
모델을 잘못 생성한 것은 아닌지 이상치를 제거하고 정규화를 진행했을 때와 아닌 모델을 비교해보고, 변수 개수도 조절해보았지만 test score가 0.5를 넘는 모델이 단 하나도 나오지 않아서 고민을 많이 했습니다.
고민 끝에 기상이라는 변수가 큰 데이터로 어떤 특정한 목표 값을 예측하기는 힘들다는 생각이 들었고 그것을 예측하는 모델이 선형 모델이면 더욱 어려울 것이라고 판단했습니다.

실제로 모델링을 마친 후 2021/3/4 대구에서 발생한 미세먼지 발생량과 다중 선형회귀모델로부터 도출한 식으로 미세먼지 발생량을 계산해서 비교해본 결과 실제 PM10은 43, 예측 모델로 나온 값은 85.59로 약 2배 가까이 차이가 나 해당 모델을 실제로 사용하기는 어렵다고 판단했습니다

그 후로는 분석 방향을 미세먼지 발생량 예측보다는 영향인자 판단과 가설에 대한 검정으로 잡았고 분석을 진행하면서 가설에 대한 유의미한 값들을 발견 할 수 있었습니다.

CO 변수는 모든 모델에서 중요도가 높은 변수로 나왔는데 이 결과가 결측치 55개를 평균으로 채웠기 때문인지 실제로 CO변수가 영향인자인지 판단할 수 없어서 이 부분에 대한 공부가 더 필요할 것 같습니다.