Analysis of Factors Influencing Fine Dust

By SeonYoung Jhang

Project Context

- **What is Fine Dust?**

Fine dust is particulate matter that can be found in the air that is incredibly small – containing air pollutants such as sulfur dioxide, nitrogen oxides, lead, ozone, carbon monoxide, etc.

These pollutants are emitted from sources like automobiles, factories, and cooking processes, and consist of fine particles with a diameter of 10 ㎛ or less, which can linger in the air for an extended period

- **The Impact of Fine Dust on Our Lives:**

\- <u>Health</u>
Prolonged exposure to fine dust can adversely affect human health.

\- <u>Environment</u>
When fine dust accumulates in vinyl greenhouses, it can lead to reduced sunlight and disruption of photosynthesis in crops, contributing to soil degradation

\- <u>Economy</u>
Economic losses occur due to the negative impact of fine dust, causing economic downturns and affecting industries sensitive to fine dust, such as semiconductors and displays.
(estimated economic losses amount to approximately KRW 4.23 trillion annually – Hyundai Economic Research Institute)

Project Context

In the Ministry of Environment's report on domestic environmental trends, the causes of fine dust are analyzed as follows:

1) Domestic Factors

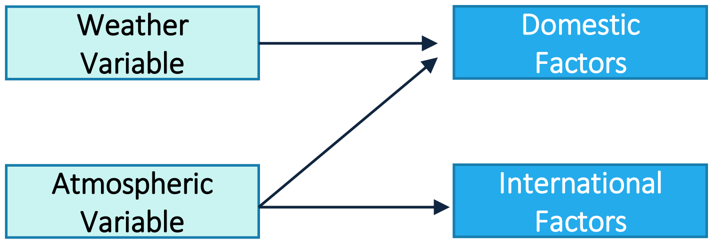
The proportion of domestically generated fine dust is 50-70%

: 51% of fine dust in Seoul being domestically produced.

: 68.2% are particles released through the combustion of fossil fuels.

2) International Factors

Approximately 43% of domestic fine dust is attributed to factors such as emissions from factories and vehicle exhaust in China, as well as desert dust.



Other related Information

1) Direction of the Wind

When winds from the west blow, the concentration of fine dust is high

2) Speed of the Wind

In the wind speed range of 0-6 m/s, the highest concentration occurs during calm conditions and the concentration decreases as wind speed increases. Beyond 6 m/s, the concentration increases with the rising wind speed.

3) Weather

On days with yellow dust or foggy weather, the concentration of fine dust tends to be high

4) Humidity

When the humidity is between 60% and 90%, the fine dust concentration is elevated. High humidity promotes the formation of secondary particulate matter in the atmosphere, such as sulfuric and nitric acid salts.

Project Objective

The background information indicated that various factors influence the concentration of fine dust.
Therefore, there is a need to verify whether the given fine dust-related data aligns with these background information based on the understanding the fine dust.

Given the multifaceted negative impacts of fine dust, the goal is to better identify factors influencing fine dust and to develop a prediction model of its occurrence amount

Project Process

1. Data Collection

2. Data Quality

- Null Values
- Outliers
- Derived Variables

3. EDA

- Data Visualization

4. Analysis

- Correlation
- 2-Sample t-test

5. Modelling

- Multiple Regression
- Decision Tree
- Random Forest
- Gradient Boosting

6. Model Evaluation

7. Conclusion

Data Collection

- Jul 1st, 2019 ~ June 30th, 2020
- Target Variable: PM10
- Independent Variable: O3, NO2, CO, SO2, TEMP, RAIN, WIND, WIND_DIR, HUMIDITY, ATM_PRESS, SNOW, CLOUD

2) 파일 불러오기

```
df_raw = pd.read_csv("AIR_POLLUTION.csv", encoding = 'cp949', parse_dates=["MeasDate"])
```

3) 데이터 확인하기

```
df_raw.head()
```

	MeasDate	PM10	O3	NO2	CO	SO2	TEMP	RAIN	WIND	WIND_DIR	HUMIDITY	ATM_PRESS	SNOW	CLOUD
0	2019-07-01	29.0	0.054	0.021	0.5	0.003	24.03	0.0	2.30	249	63.2	995.1	0.0	5.70
1	2019-07-02	26.0	0.053	0.020	0.5	0.003	24.29	0.0	2.26	265	63.2	998.6	0.0	3.83
2	2019-07-03	30.0	0.042	0.023	0.4	0.003	24.18	0.0	1.79	280	65.3	998.3	0.0	6.29
3	2019-07-04	28.0	0.034	0.026	0.4	0.003	25.35	0.0	2.04	263	58.6	996.6	0.0	2.54
4	2019-07-05	29.0	0.045	0.035	0.5	0.003	27.30	0.0	1.45	175	45.5	993.5	0.0	3.92

```
df_raw.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 366 entries, 0 to 365
Data columns (total 14 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   MeasDate   366 non-null    datetime64[ns]
 1   PM10       365 non-null    float64
 2   O3         365 non-null    float64
 3   NO2        365 non-null    float64
 4   CO         311 non-null    float64
 5   SO2        365 non-null    float64
 6   TEMP       366 non-null    float64
 7   RAIN       366 non-null    float64
 8   WIND       366 non-null    float64
 9   WIND_DIR   366 non-null    int64
 10  HUMIDITY   366 non-null    float64
 11  ATM_PRESS  366 non-null    float64
 12  SNOW       366 non-null    float64
 13  CLOUD      366 non-null    float64
dtypes: datetime64[ns](1), float64(12), int64(1)
memory usage: 40.2 KB
```

Data Cleansing

- Null Values

Variable	# of Null
PM10	1
O3	1
NO2	55
CO	1
SO2	1

- Multiple missing values were identified in 2020-05-02
→ Removal

- For NO2, input the average value by season

- Outliers

Using user-defined function

Detecting outliers for each variable

```python
def outlier_iqr(data, column):
    # lower, upper 글로벌 변수 선언하기
    global lower, upper

    #4 분위수 기준 정하기
    q25, q75 = np.quantile(data[column], 0.25), np.quantile(data[column], 0.75)

    # IQR 계산하기
    iqr = q75 - q25

    # outlier cutoff 계산하기
    cut_off = iqr * 1.5

    # lower와 upper bound 값 구하기
    lower, upper = q25 - cut_off, q75 + cut_off

    print('IQR은', iqr.round(3), '이다.')
    print('lower bound 값은', lower.round(3), '이다.')
    print('upper bound 값은', upper.round(3), '이다.')

    # 1사 분위와 4사 분위에 속해있는 데이터 각각 저장하기
    data1 = data[data[column]>upper]
    data2 = data[data[column]<lower]

    # 이상치 총 개수 구하기
    return print('총 이상치 개수는', data1.shape[0] + data2.shape[0], '이다.')
```

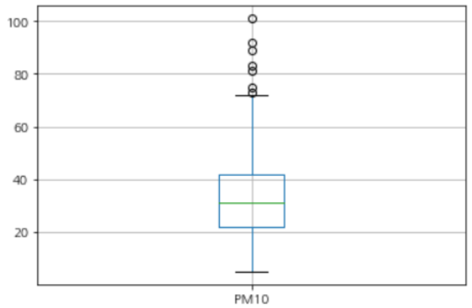
PM10

```python
outlier_iqr(df_raw, 'PM10')
```

```
IQR은 20.0 이다.
lower bound 값은 -8.0 이다.
upper bound 값은 72.0 이다.
총 이상치 개수는 7 이다.
```

```python
df_raw.boxplot(column="PM10")
```

```
<AxesSubplot:>
```

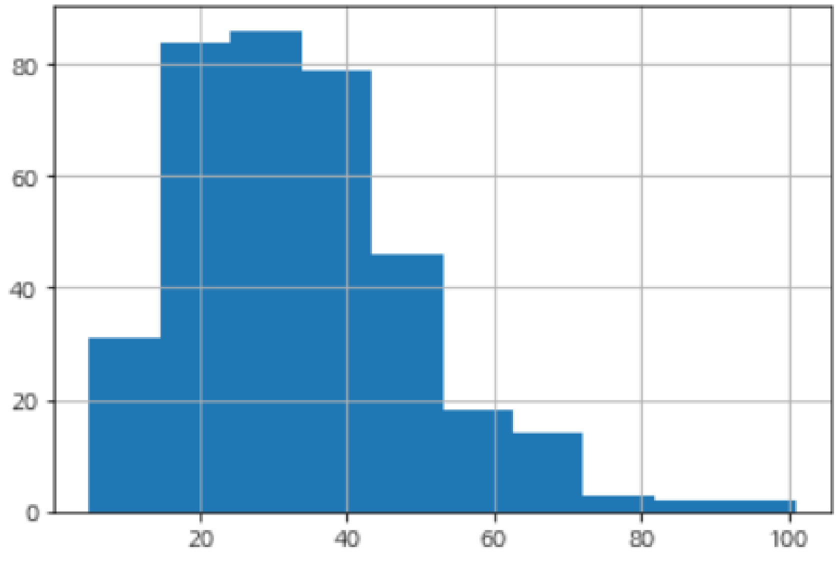


However, no separate handling for outliers is done

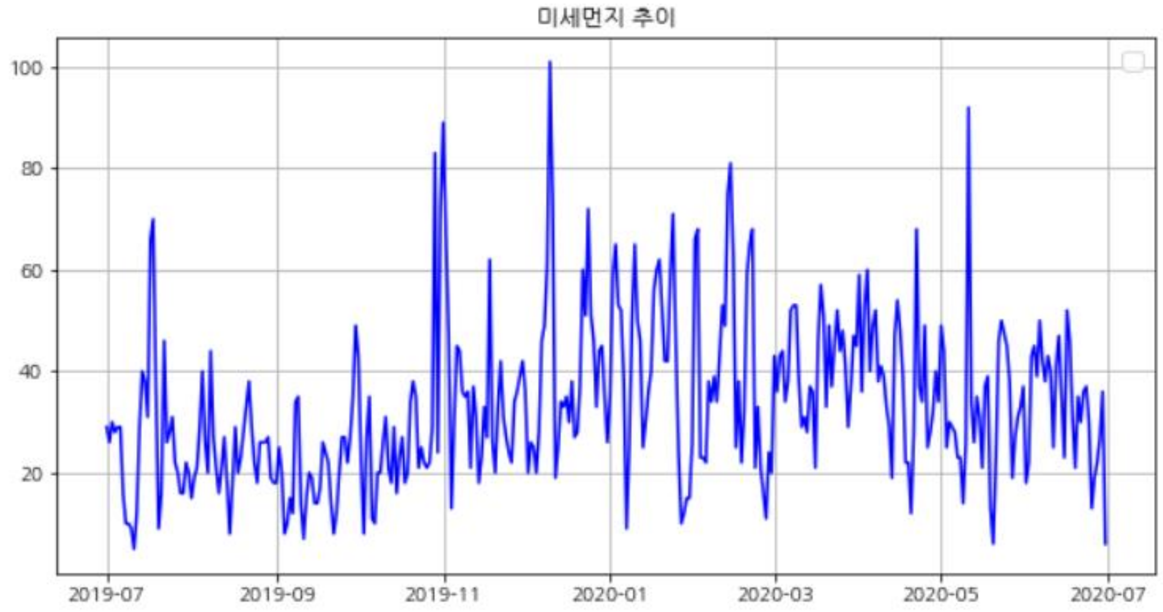
Data Visualization

1) Target Variable (PM10)

Checking the distribution through a Histogram



Verifying the trend of fine dust over time using a Line Graph.



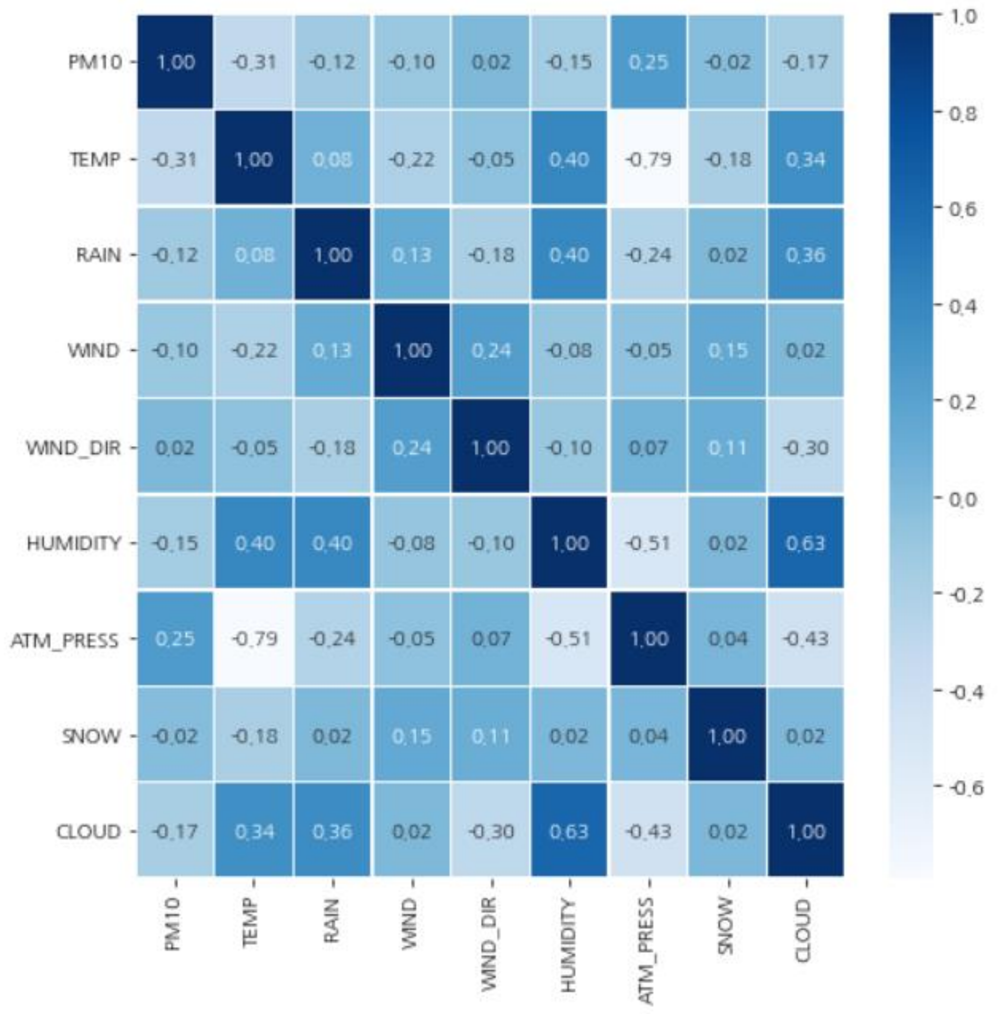
Data Visualization

2) Independent Variables

The given X variables were grouped based on two main criteria

<u>1) Weather Variable</u>

: TEMP, RAIN, WIND, WIND_DIR, HUMIDITY, ATM_PRESS, SNOW, CLOUD



→ There is no significant correlation with PM10

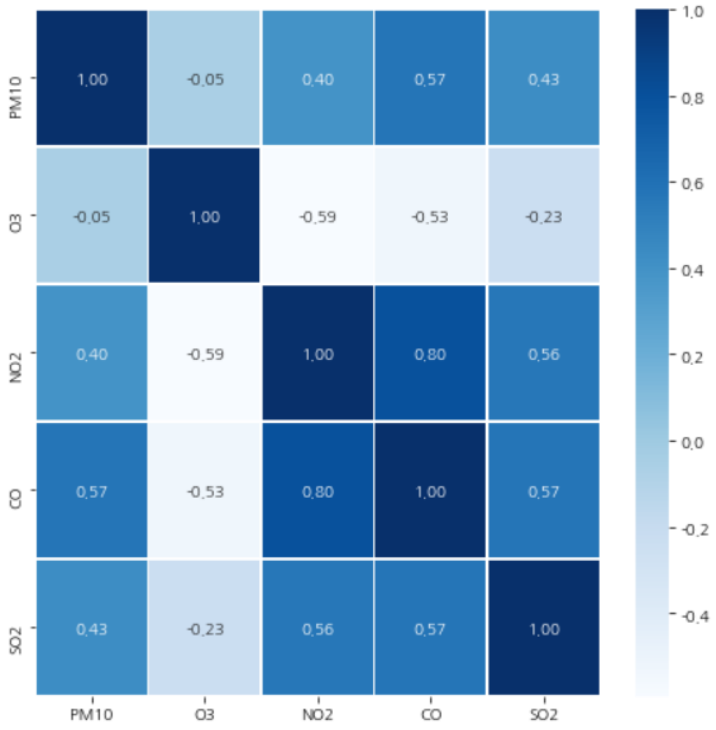
Data Visualization

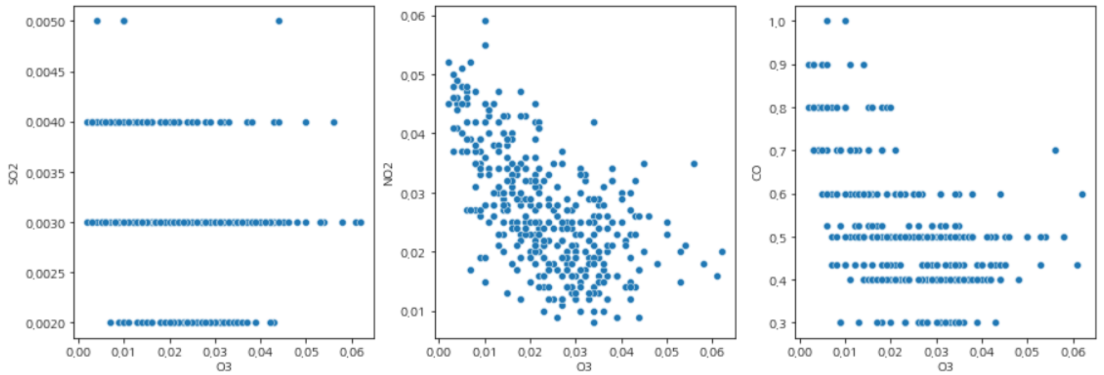
2) Independent Variables

The given X variables were grouped based on two main criteria

<u>2) Atmospheric Variable</u>

: O3, NO2, CO, SO2





→ O3: Not directly influenced by PM10.

BUT, shows a negative correlation with SO2, NO2, and CO as observed in the scatter plots.

In other words, while O3 may not have a direct impact, but it is indirectly associated

<Correlation Analysis>
PM10과 NO2/CO/SO2

```
Correlation Analysis
corr:0.396
p-value:0.000
```

PM10과 NO2는 약한 상관성이 있다고 할 수 있다. (H0 기각)

```
Correlation Analysis
corr:0.573
p-value:0.000
```

PM10과 CO는 상관성이 있다고 할 수 있다. (H0 기각)

```
Correlation Analysis
corr:0.429
p-value:0.000
```

PM10과 SO2는 상관성이 있다고 할 수 있다. (H0 기각)

< Correlation Analysis >
PM10과 O3

```
Correlation Analysis
corr:-0.052
p-value:0.324
```

PM10과 O3은 상관성이 있다고 할 수 없다. (H0 채택)

→ There is a correlation between PM10 and NO2, CO, and SO2.

Data Visualization

3) Derived Variables

- PM10

: Categorized according to WHO standards into Good/Normal/Bad/Very Bad

- WIND_DIR

: Detailed categorization of wind direction into 8 compass directions.

- WIND

: Detailed categorization based on wind speed.

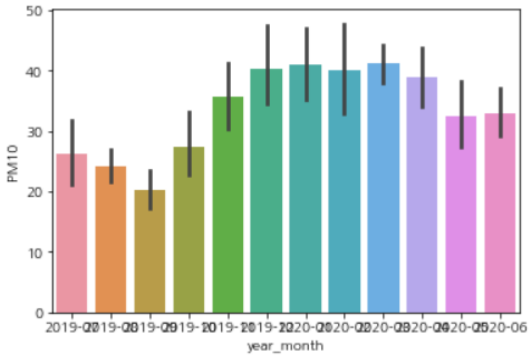
- Seasons

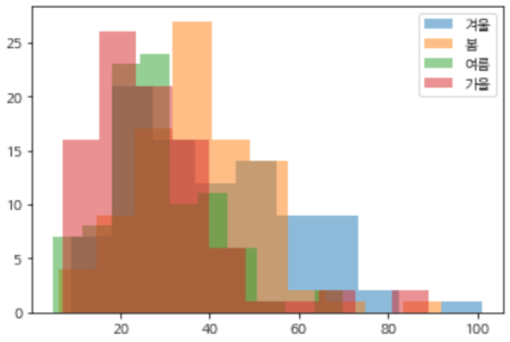
- Rain/Snow

<Reference>
WHO Standard

미세먼지농도 (μg/㎥, 일평균)	좋음	보통	나쁨	매우나쁨
PM10	0~20	21~45	46~75	76이상

1) Seasons

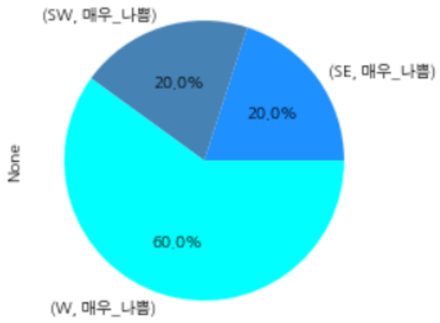


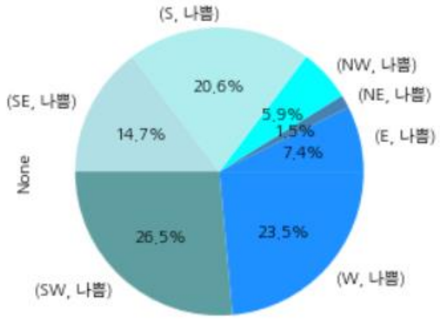


→ PM10 exhibiting varying trends across different seasons

→ Particularly high levels of PM10 are observed during the winter.

2) Wind Direction



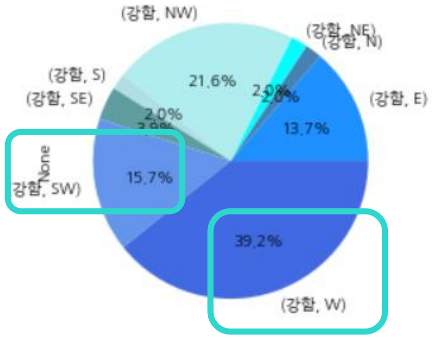


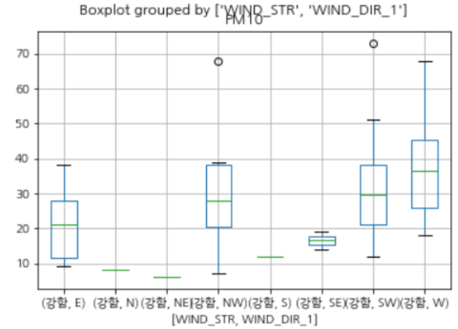
In cases where PM10 is categorized as 'Very Bad' or 'Bad', W (west) and SW (southwest) directions account for more than 50%.

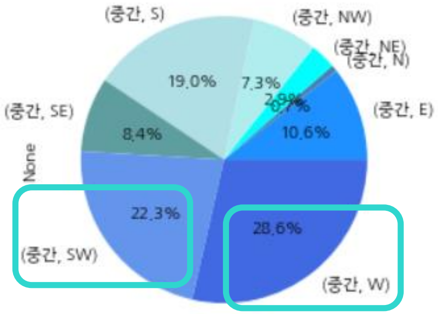
Data Visualization

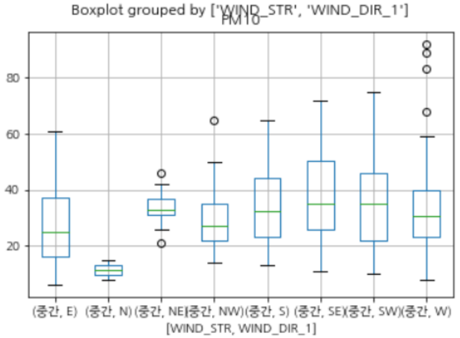
3) Derived Variables

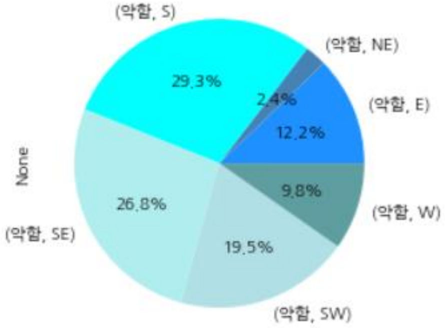
3) Wind Speed

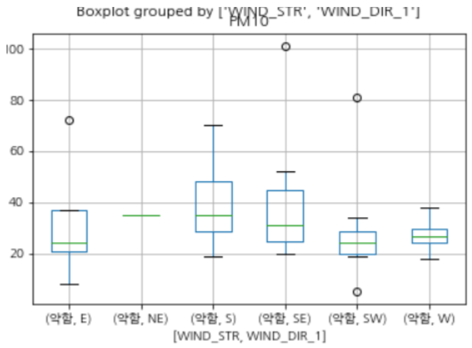













```
corr_ana(df_raw_h4['WIND'], df_raw_h4["PM10"])
```

```
Correlation Analysis
corr:-0.033
p-value:0.773
```

[Hypothesis] If the wind speed is high and from the west, PM10 levels are expected to be high.

→ According to the correlation analysis, there is no correlation between wind speed and PM10.

<u>BUT</u> Referring to the Boxplot, it can be observed that when the wind speed is high/medium, and the wind direction is from the W (west) or SW (southwest), PM10 levels were high or exhibited outliers towards the higher values.

<Selection of Influential Factors for Prediction Model>

NO2 / CO / SO2 / Atmospheric Pressure / Temperature / Humidity / Season / Wind Direction / Wind Speed

Prediction Model

Multiple Linear Regression

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   PM10   R-squared:                       0.489
Model:                            OLS   Adj. R-squared:                  0.480
Method:                 Least Squares   F-statistic:                     57.09
Date:                Mon, 08 Nov 2021   Prob (F-statistic):           2.57e-49
Time:                        01:30:52   Log-Likelihood:                -1405.5
No. Observations:                 365   AIC:                             2825.
Df Residuals:                     358   BIC:                             2852.
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    324.1706    140.082      2.314      0.021      48.683     599.658
O3           631.5493     72.518      8.709      0.000     488.935     774.164
NO2          405.6936    120.883      3.356      0.001     167.964     643.423
CO            64.4865      7.629      8.453      0.000      49.484      79.489
TEMP          -0.6233      0.115     -5.406      0.000      -0.850      -0.397
WIND_DIR       0.0356      0.010      3.595      0.000       0.016       0.055
ATM_PRESS     -0.3461      0.139     -2.489      0.013      -0.620      -0.073
==============================================================================
Omnibus:                       96.804   Durbin-Watson:                   1.165
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              265.903
Skew:                           1.242   Prob(JB):                     1.82e-58
Kurtosis:                       6.364   Cond. No.                     2.66e+05
==============================================================================
```

```
MAPE(df_test_y, reg_y_pred)    31.146956791735402
```

- R-squared: 48.6%의 설명력

- Prob(F-statistics): 2.57e-49

Statistically significant

Final Model Regression Equation

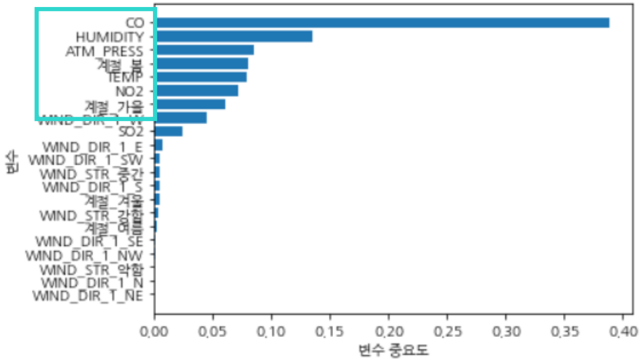
$$\hat{Y} = 324.1706 + 631.5493\,O3 + 405.6936\,NO2 + 64.4865\,CO - 0.6233\,TEMP - 0.3451\,ATM\_PRESS$$

MAPE: 31.13

Judged as a highly reasonable prediction as it is below 50. (Tofallis, 2016)

Random Forest

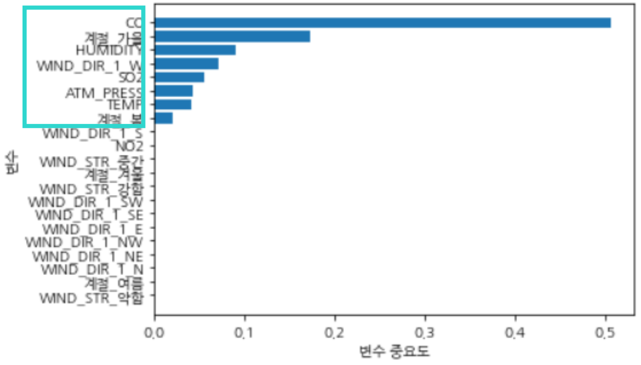
```
Score on training set:0.698
Score on test set:0.447
```



- N_estimator: 50
- Leaf: 5
- Split : 10
- Max depth: 8

Decision Tree

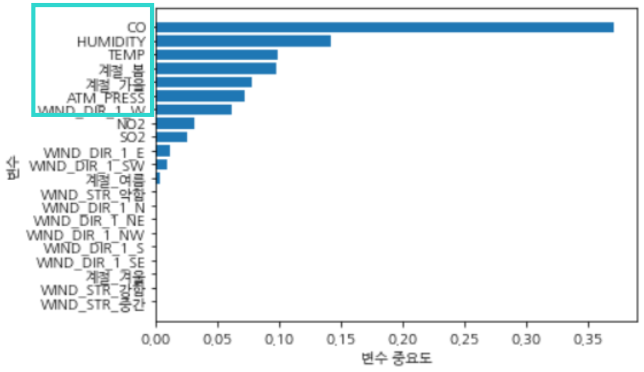
```
Score on training set:0.454
Score on test set:0.222
```



- Leaf : 10
- Split : 22
- Max depth: 4

Gradient Boosting

```
Score on training set:0.721
Score on test set:0.450
```



- N_estimator: 70
- Leaf: 22
- Max depth: 3
- Learning rate: 0.1

Conclusion

- <u>Analysis Result</u>

  - PM10 levels have a weak correlation with weather-related variables and a strong correlation with atmospheric-related variables.

  - Seasons play a significant role in relation to PM10, especially showing high PM10 in winter.

  - Based on the 8-direction wind criteria, fine dust concentrations are particularly high in cases of west and southwest winds, categorized as 'Bad' or 'Very Bad'.

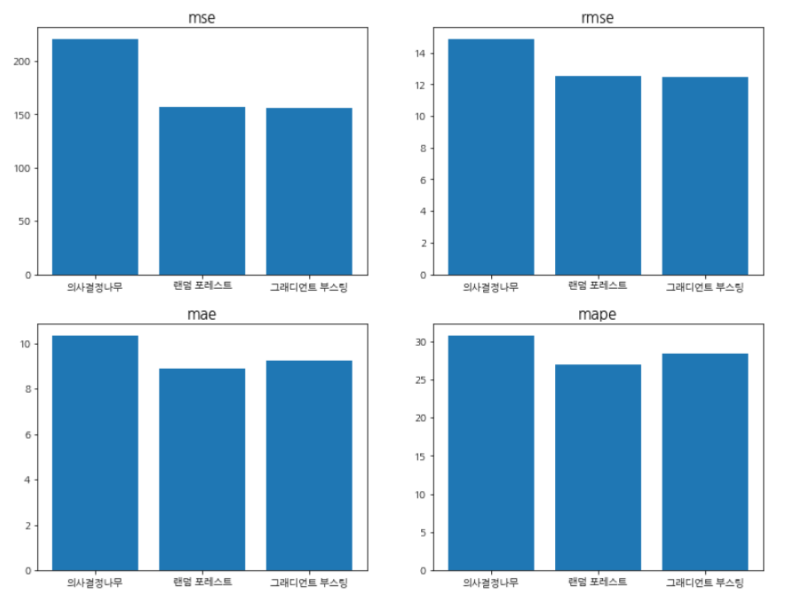
- <u>Prediction Model</u>

  - Used Multiple Regression Analysis / Decision Tree / Random Forest / Gradient Boosting to devise the prediction model

  - Random Forest and Gradient Boosting showed comparatively high accuracy.

  - Shared Key Influential Variables

  : CO, HUMIDITY, ATM_PRESS, Season.



Appendix

#	Variable	Details	Type	Reason for excluding	EDA			Modelling			
					Graph	Statistic Analysis	Correlation Analysis	Regression	DT	RF	GB
1	MeasDate	Date	Continuous	-							
2	PM10	Particulate Matter $10\mu g/m^3$	Continuous		line, histogram						
3	O3	Ozone Concentration	Continuous	Indirect Influence	heatmap, box plot, scatterplot		O	O			
4	NO2	Nitrogen Dioxide Concentration	Continuous		heatmap, box plot, scatterplot		O	O	O	O	O
5	CO	Nitric Oxide Concentration	Continuous		heatmap, box plot, scatterplot		O	O	O	O	O
6	SO2	Sulfur Dioxide Concentration	Continuous		heatmap, box plot, scatterplot		O		O	O	O
7	TEMP	Temperature (°C)	Continuous		heatmap		O	O	O	O	O
8	RAIN	Precipitation (mm)	Continuous	Low Correlation	heatmap	2 sample t-test					
9	WIND	Wind Speed (m/s)	Continuous	*Converted to Categorical	heatmap, box plot, pie chart		O		O	O	O
10	WIND_DIR	Wind Direction (16Cardinal Directions)	Continuous	*Converted to Categorical	heatmap, box plot, pie chart		O	O	O	O	O
11	HUMIDITY	Humidity(%)	Continuous		heatmap				O	O	O
12	ATM_PRESS	Atmospheric Pressure(hPa)	Continuous	*Converted to Categorical	heatmap			O	O	O	O
13	SNOW	Snowfall(cm)	Continuous	Low Correlation	heatmap	2 sample t-test					
14	CLOUD	Cloud Cover (in tenths)	Continuous	Low Correlation	heatmap						
15	**Season**	Four Seasons	Discrete		heatmap, box plot, histogram, bar plot	2 sample t-test	O		O	O	O


Thank You