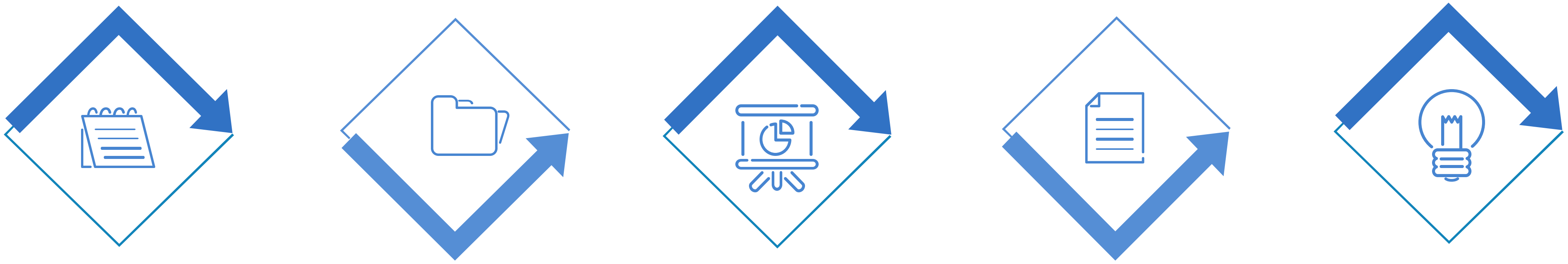
Analysis of Factors Influencing Fine Dust

SeonYoung Jhang

Table of Contents

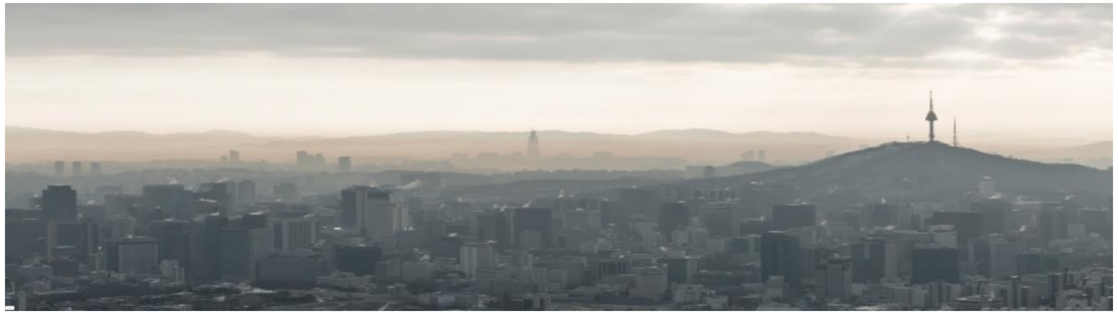


Project Overview

- **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

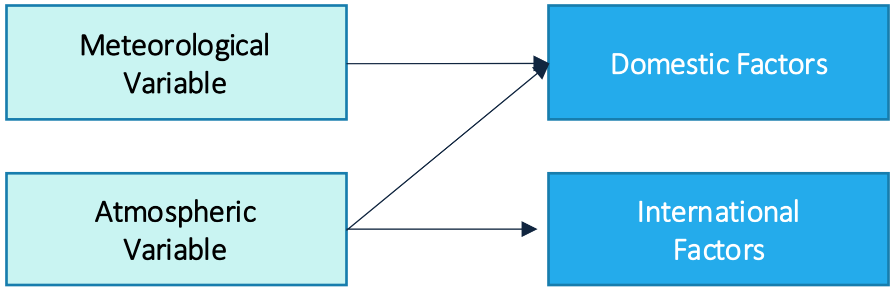
\- <u>Economy</u>
Economic losses occur due to the negative impact of fine dust, causing economic downturns and affecting industries sensitive to fine dust, such as semiconductors and displays. *(estimated economic losses amount to approximately KRW 4.23 trillion annually – Hyundai Economic Research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

Project Overview

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

<u>1) Domestic Factors</u>
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.

<u>2) International Factors</u>
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

<u>1) Direction of the Wind</u>
When winds from the west blow, the concentration of fine dust is high

<u>2) Speed of the Wind</u>
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.

<u>3) Weather</u>
On days with yellow dust or foggy weather, the concentration of fine dust tends to be high

<u>4) Humidity</u>
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

Data Preparation

- 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

```python
df = pd.read_csv("AIR_POLLUTION.csv")
df.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

Data shape: (365, 13)

```python
df.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    object
 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: float64(12), int64(1), object(1)
memory usage: 40.2+ KB
```

Data Preparation

Target Variable

Variable	Details
PM10	Particulate Matter 10㎍/㎥

Atmospheric Variable

Variable	Details
O3	Ozone Concentration
NO2	Nitrogen Dioxide Concentration
CO	Nitric Oxide Concentration
SO2	Sulfur Dioxide Concentration

Meteorological Variable

Variable	Details
TEMP	Temperature (°C)
RAIN	Precipitation (mm)
WIND	Wind Speed (m/s)
WIND_DIR	Wind Direction (16Cardinal Directions)
HUMIDITY	Humidity(%)
ATM_PRESS	Atmospheric Pressure(hPa)
SNOW	Snowfall(cm)
CLOUD	Cloud Cover (in tenths)
Season	Four Seasons

Datetime Variable

Variable	Details
MeasDate	Measured Date

Derived Variable

Variable	Details
Season	Four Seasons

- Winter: Dec – Feb
- Summer: Jun – Aug
- Spring: Mar – May
- Fall: Sep – Nov

Data Cleansing

- Null Values

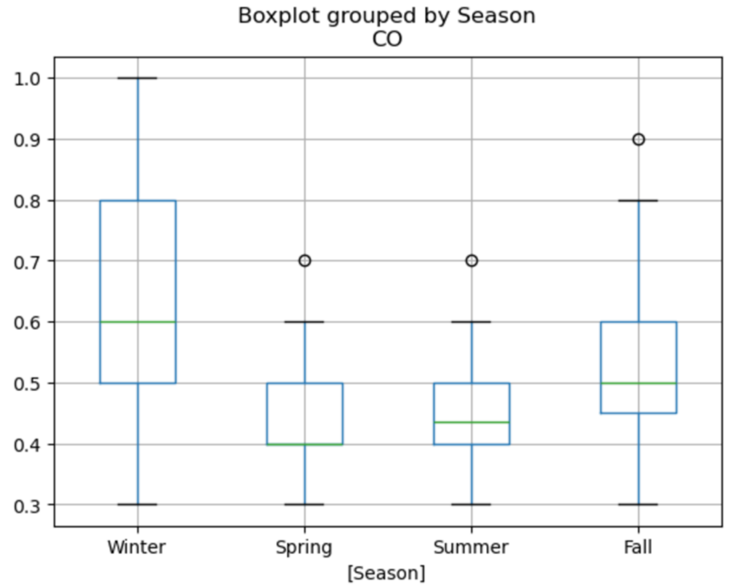
Variable	# of Null
PM10	1
O3	1
NO2	1
CO	55
SO2	1

- Multiple missing values were identified in 2020-05-02
  → replaced by the average value by season

- For CO, input the average value by season

```
df.boxplot(column="CO", by = ["Season"])
# Customize x-axis labels
plt.xticks([1, 2, 3, 4], ["Winter", "Spring", "Summer", "Fall"])
```

```
([<matplotlib.axis.XTick at 0x16aa633e0>,
  <matplotlib.axis.XTick at 0x16aaa3620>,
  <matplotlib.axis.XTick at 0x16aaa1010>,
  <matplotlib.axis.XTick at 0x16aaa14f0>],
 [Text(1, 0, 'Winter'),
  Text(2, 0, 'Spring'),
  Text(3, 0, 'Summer'),
  Text(4, 0, 'Fall')])
```



The combination of the boxplot and ANOVA results confirms that `CO` levels differ significantly across seasons. The low p-value (< 0.05) provides strong statistical evidence for this conclusion, and the high F-statistic further supports the existence of meaningful differences between seasonal groups.

Data Cleansing

- Outliers

Using user-defined function

Detecting outliers for each variable

```
def outlier_iqr(data, column):
    # declaring lower, upper as a global varialbe
    global lower, upper

    q25, q75 = np.quantile(data[column], 0.25), np.quantile(data[column], 0.75)

    # calculating IQR
    iqr = q75 - q25

    # calculating the outlier cutoff
    cut_off = iqr * 1.5

    # lower & upper bound
    lower, upper = q25 - cut_off, q75 + cut_off

    print('IQR is', iqr.round(3))
    print('lower bound is', lower.round(3))
    print('upper bound is', upper.round(3))

    data1 = data[data[column]>upper]
    data2 = data[data[column]<lower]

    return print('Total outliers are', data1.shape[0] + data2.shape[0])
```

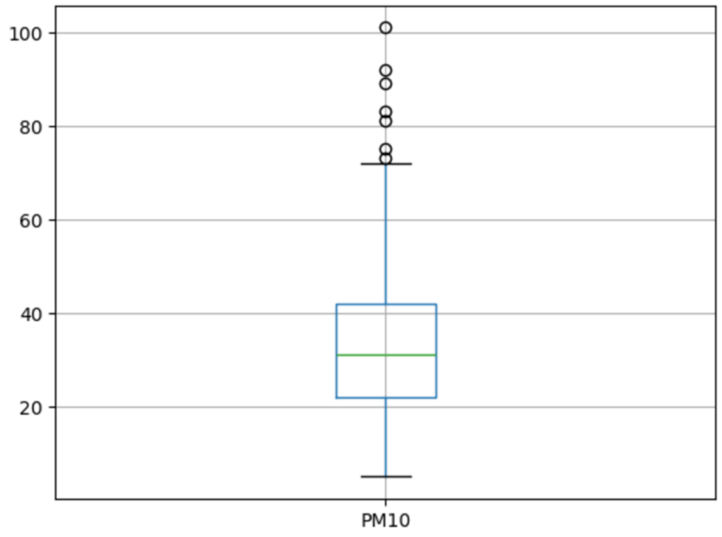
example

```
outlier_iqr(df, 'PM10')
```

```
IQR is 20.0
lower bound is -8.0
upper bound is 72.0
Total outliers are 7
```

```
df.boxplot('PM10')
```

```
<Axes: >
```

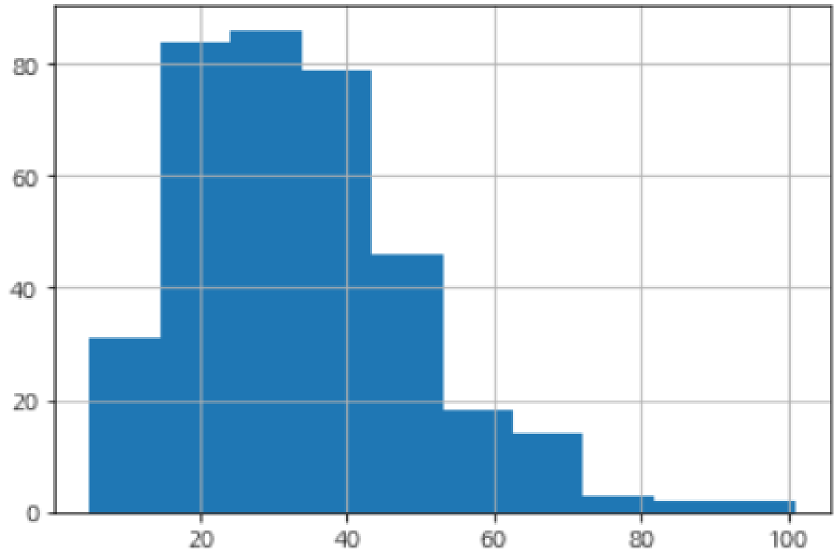


However, no separate handling for outliers is done

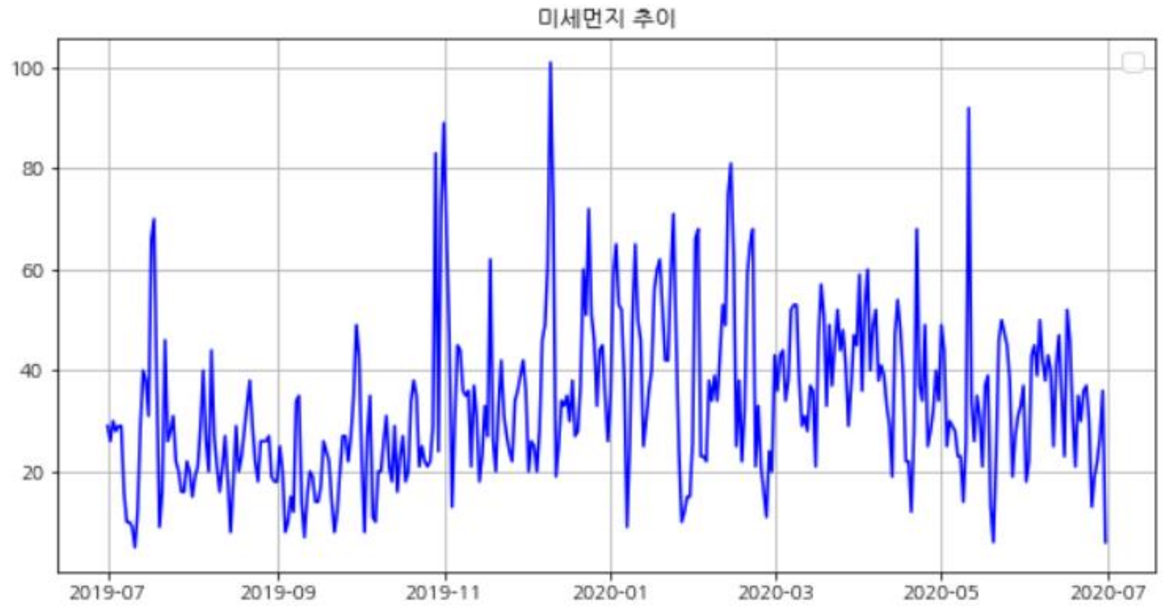
EDA

1) Target Variable (PM10)

Checking the distribution through a Histogram



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



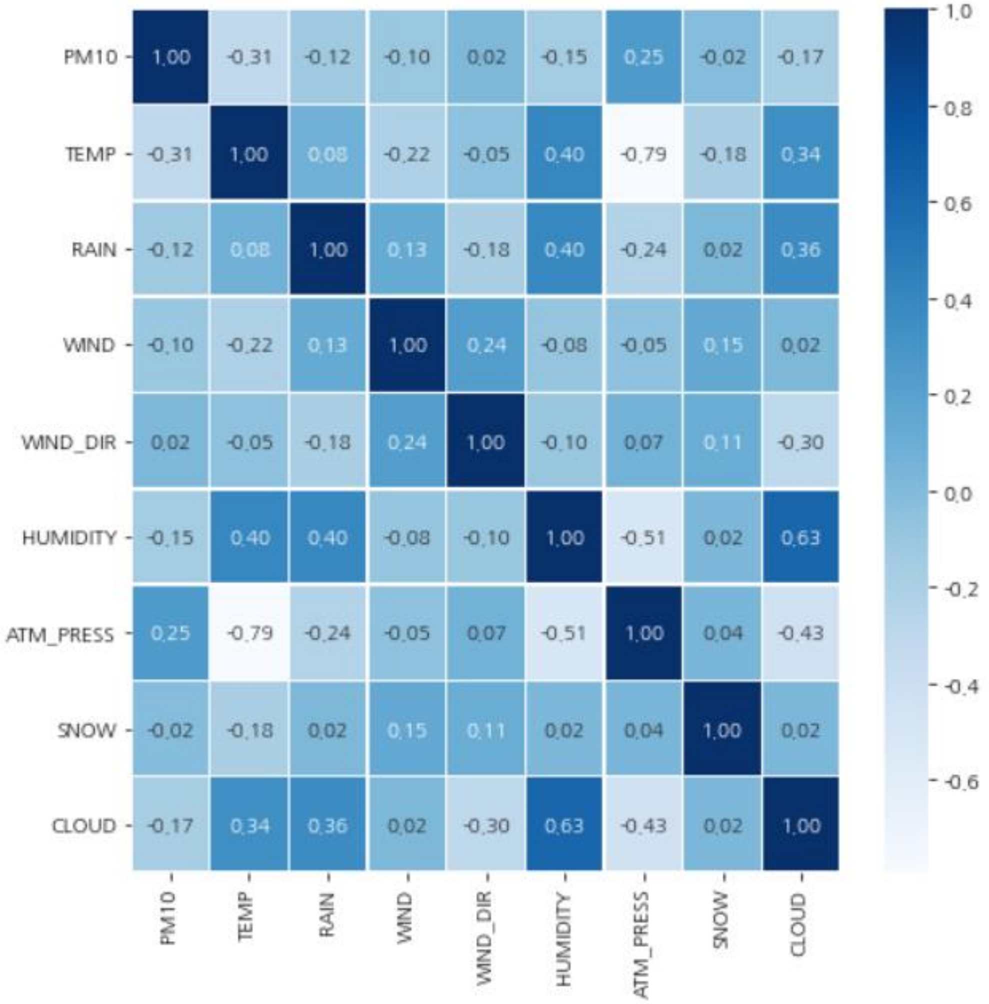
EDA

2) Independent Variables

The given X variables were grouped based on two main criteria

<u>1) Meterological Variables</u>

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



→ There is no significant correlation with PM10

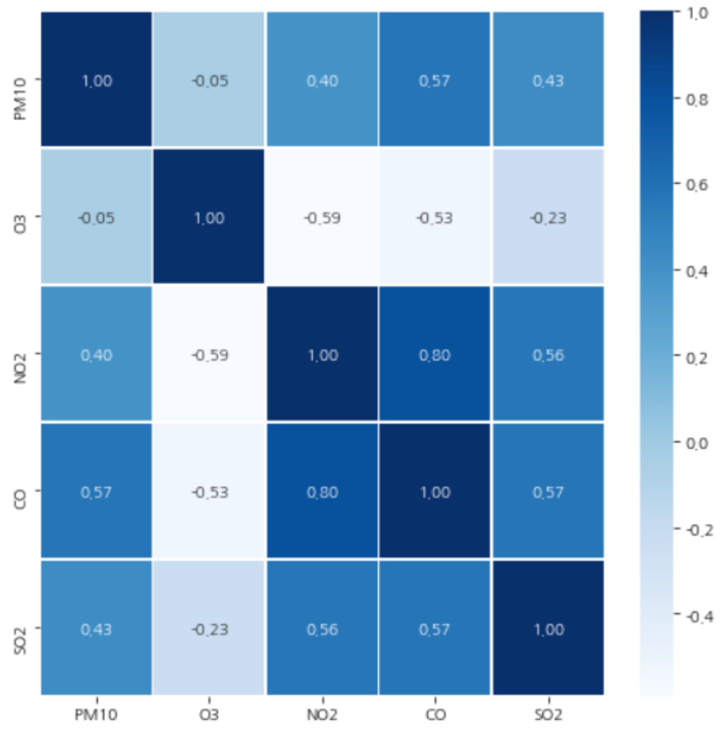
Data Visualization

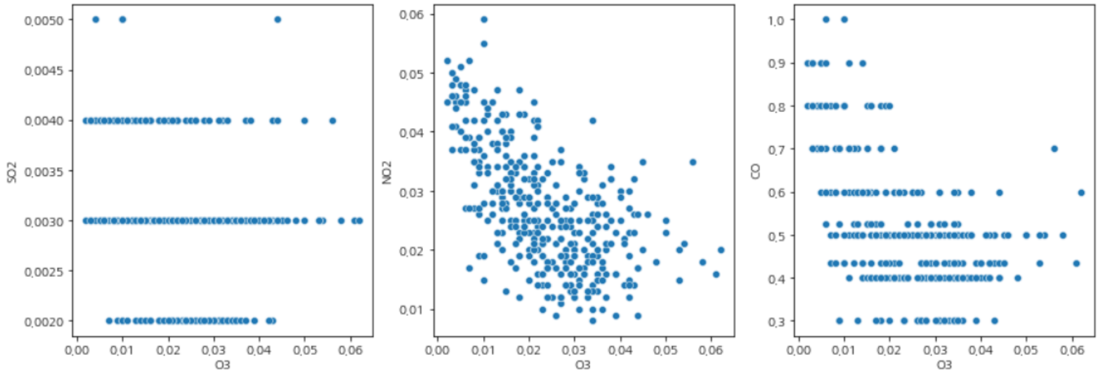
2) Independent Variables

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.
However, it 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 with PM10

	Correlation	P-value
NO2	0.396	0.000
CO	0.573	0.000
SO2	0.428	0.000
O3	-0.051	0.326

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

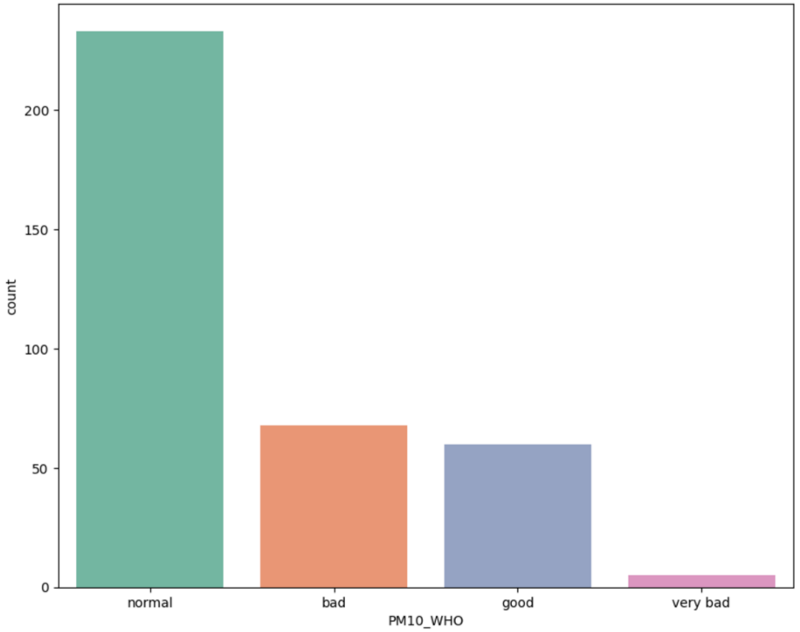
EDA

3) Derived Variables

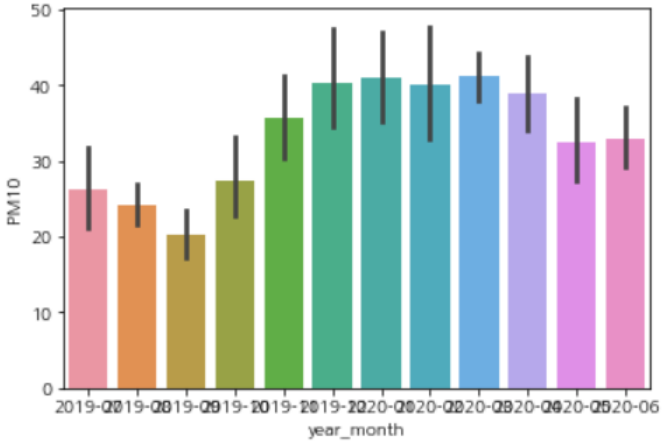
<u>PM10_WHO</u>

: Categorized according to WHO standards into

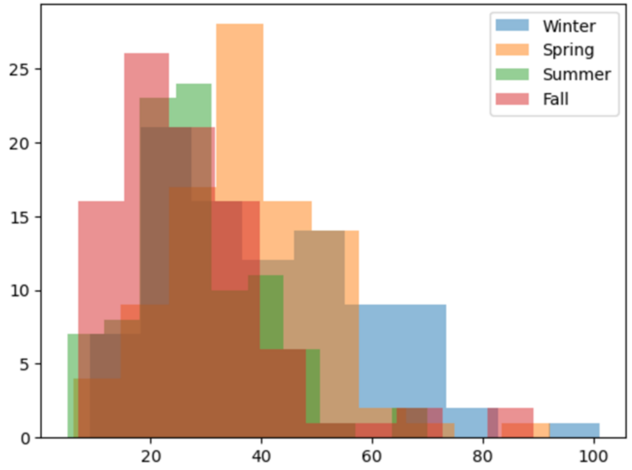
Good/Normal/Bad/Very Bad



<u>Season</u>



→ PM10 exhibiting varying trends across different seasons

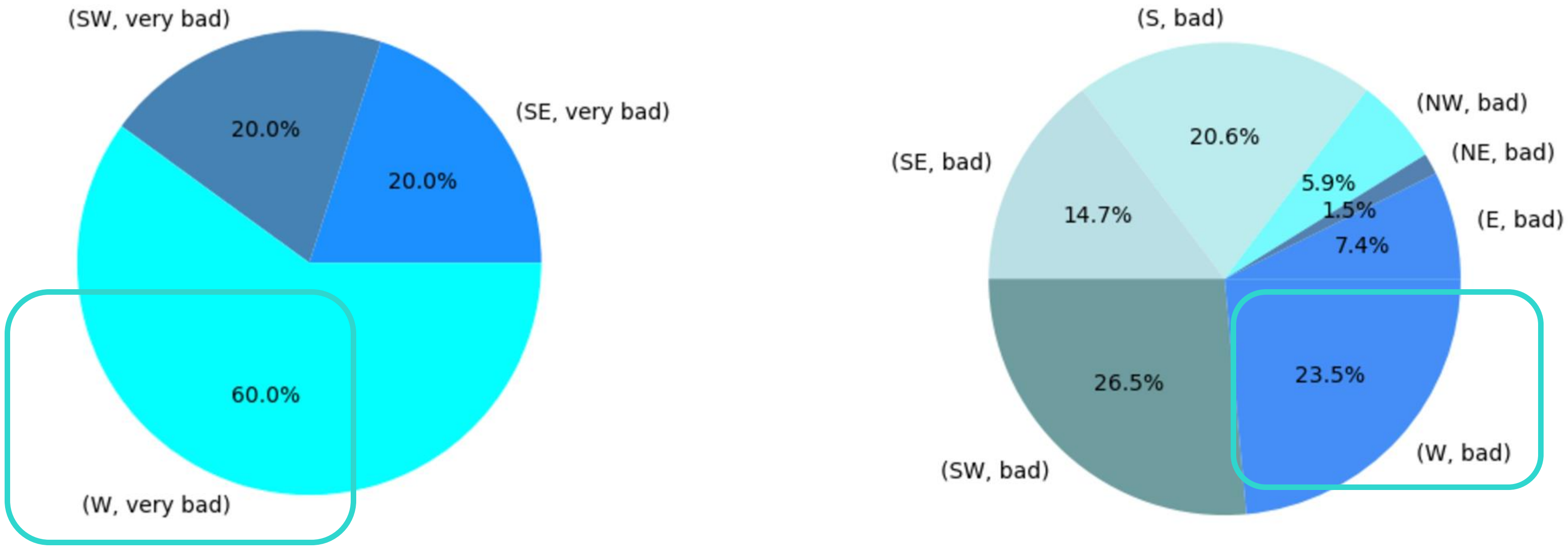


→ Particularly high levels of PM10 are observed during the winter.

EDA

3) Derived Variables

- <u>Wind Direction</u>



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

Modeling

Multiple Linear Regression

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   PM10   R-squared:                       0.534
Model:                            OLS   Adj. R-squared:                  0.521
Method:                 Least Squares   F-statistic:                     40.67
Date:                Tue, 31 Dec 2024   Prob (F-statistic):           5.13e-53
Time:                        15:34:20   Log-Likelihood:                -1392.0
No. Observations:                 366   AIC:                             2806.
Df Residuals:                     355   BIC:                             2849.
Df Model:                          10
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept  -9404.8628   3244.763     -2.898      0.004   -1.58e+04   -3023.488
O3           438.4451     76.253      5.750      0.000     288.481     588.409
NO2          508.1143    128.661      3.949      0.000     255.080     761.148
CO            65.1312      7.656      8.507      0.000      50.073      80.189
WIND           2.4794      1.035      2.396      0.017       0.444       4.514
WIND_DIR       0.0452      0.010      4.459      0.000       0.025       0.065
HUMIDITY      -0.0355      0.060     -0.589      0.556      -0.154       0.083
ATM_PRESS     -0.1633      0.127     -1.289      0.198      -0.412       0.086
CLOUD         -0.3464      0.275     -1.259      0.209      -0.888       0.195
Year           4.7229      1.597      2.957      0.003       1.582       7.864
Season        -2.4299      0.778     -3.122      0.002      -3.961      -0.899
==============================================================================
Omnibus:                      136.307   Durbin-Watson:                   1.271
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              582.246
Skew:                           1.573   Prob(JB):                    3.69e-127
Kurtosis:                       8.318   Cond. No.                     1.28e+07
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.28e+07. This might indicate that there are
strong multicollinearity or other numerical problems.
```

- R-squared: The model can explain 56.4%
- Prob(F-statistics): 5.13e-53, considered to be statistically significant

Final Model Regression Equation

Y_hat = -9404.8628 + 438.4451 O3 + 508.1143 NO2 + 65.1312 CO – 2.4794 WIND + 0.0452 WIND_DIR – 0.0355 HUMIDITY – 0.1633 ATM_PRESS -0.3464 CLOUD + 4.7229 Year -2.4299 Season

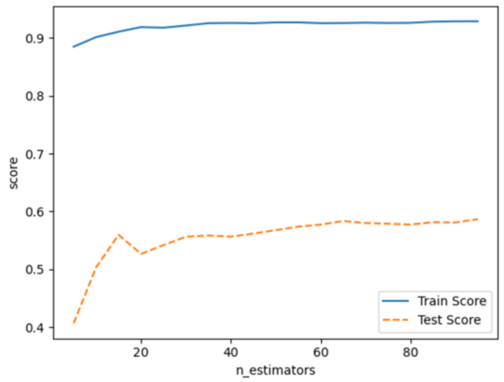
MAPE: 25.28

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

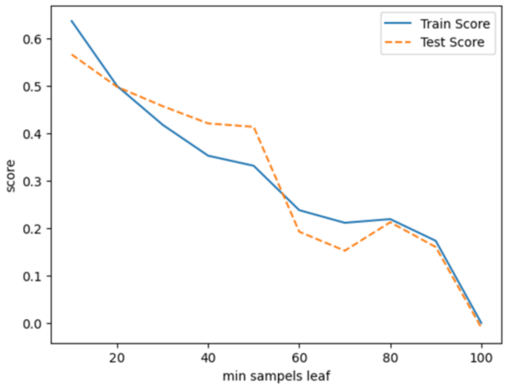
Modeling

Random Forest

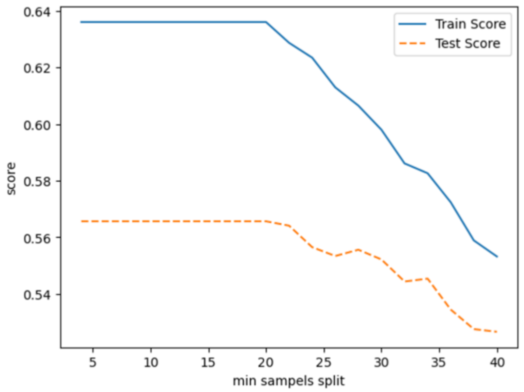
- Split train and validation data into 20:80



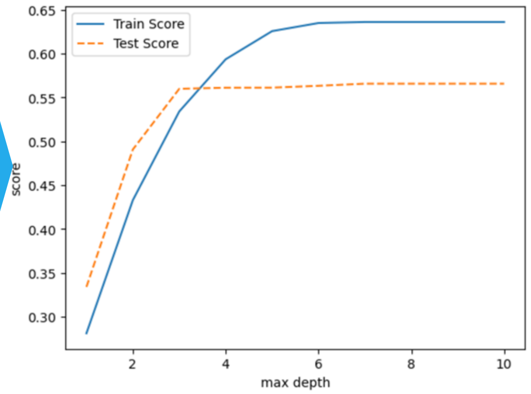
# of trees



# of leaves



# of splits



Max depth

Final Random Forest Regressor Model

- N_estimator: 15
- Leaf: 10
- Split : 20
- Max depth: 4

```
rf_final = RandomForestRegressor(random_state=1234, n_estimators=15, min_samples_leaf=10, min_samples_split=20, max_depth=4)
rf_final.fit(df_train_x, df_train_y)

print("Score on training set:{:.3f}".format(rf_final.score(df_train_x, df_train_y)))
print("Score on test set:{:.3f}".format(rf_final.score(df_test_x, df_test_y)))
```

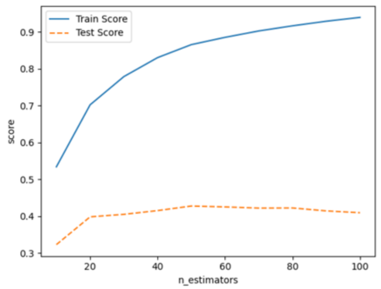
```
Score on training set:0.594
Score on test set:0.561
```

With this model, the score on the test set was <u>0.561</u>

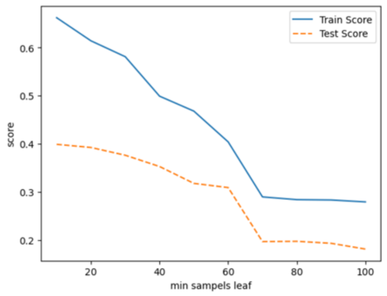
Modeling

Gradient Boosting

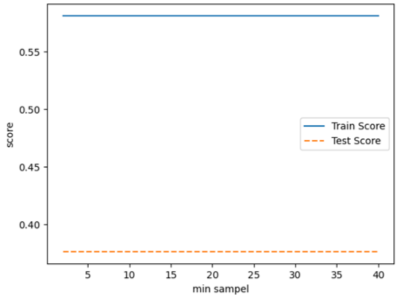
- Split train and validation data into 20:80



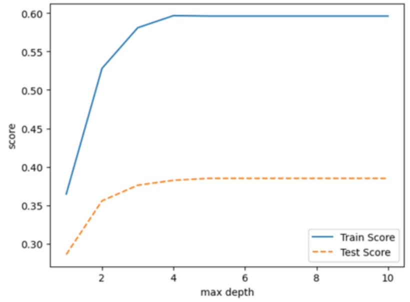
# of trees



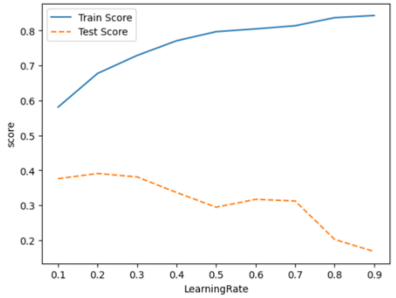
# of leaves



# of splits



Max depth



Learning rate

Final Gradient Boosting Model

- N_estimator: 20
- Leaf: 30
- Split : -
- Max depth: 3
- Learning rate: 0.2

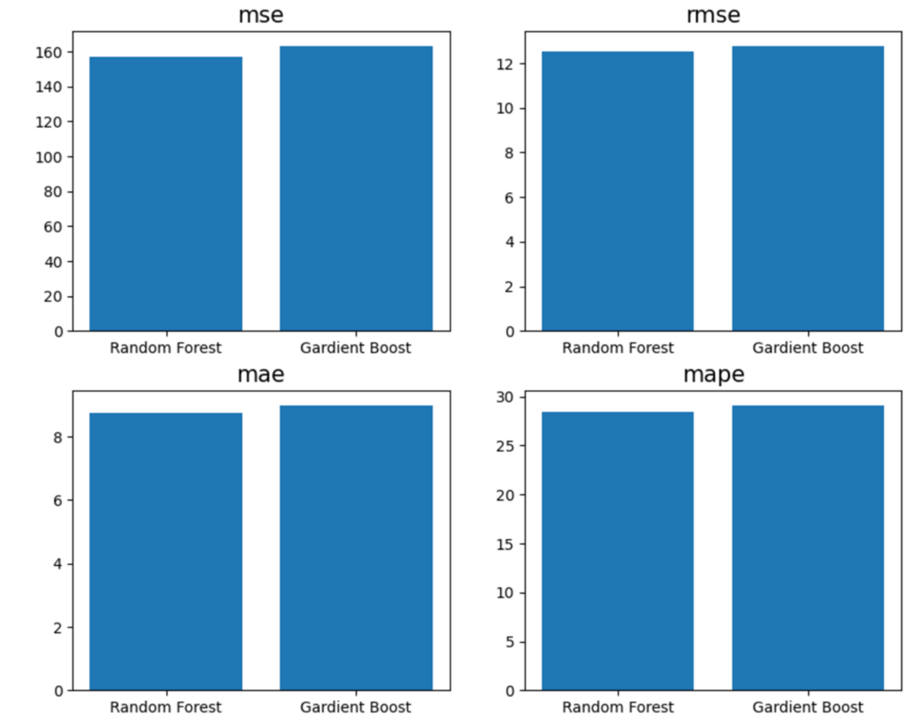
```
gb_final = GradientBoostingRegressor(random_state=1234, n_estimators=20, min_samples_leaf=30, max_depth=3,learning_rate=0.2)
gb_final.fit(df_train_x, df_train_y)

print("Score on training set:{:.3f}".format(gb_final.score(df_train_x, df_train_y)))
print("Score on test set:{:.3f}".format(gb_final.score(df_test_x, df_test_y)))
```

```
Score on training set:0.677
Score on test set:0.391
```

With this model, the score on the test set was <u>0.391</u>

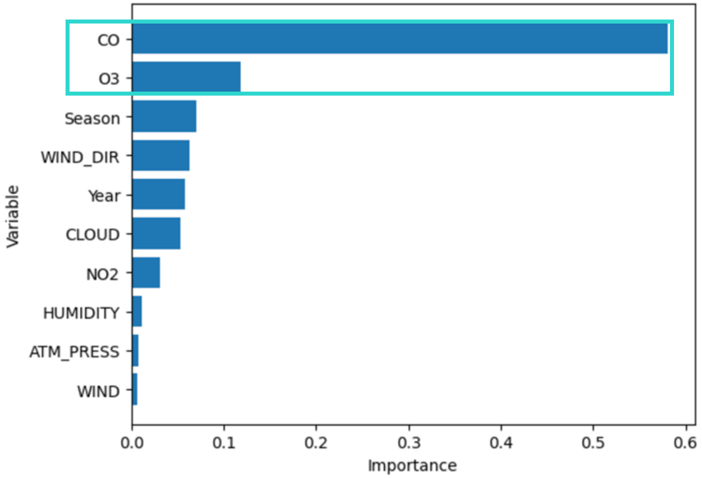
Model Evaluation

Random Forest model is better for this case since it has higher test score rate and low MSE, RMSE, MAE, MAPE values.

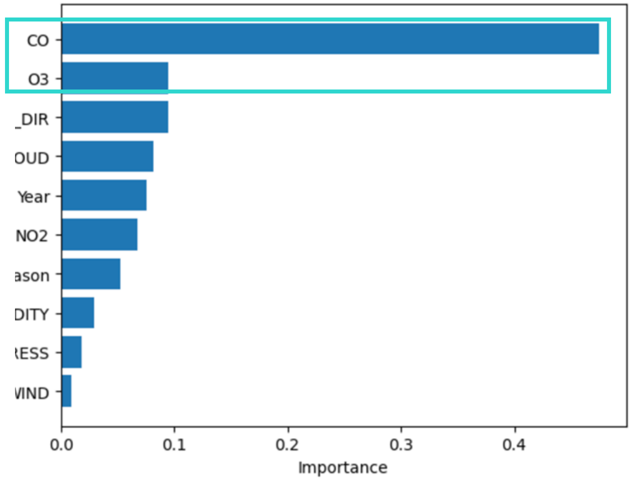


Feature Importance

Random Forest



Gradient Boosting



Both in the Random Forest model and Gradient Boost model have CO and O3 as the highest feature Importance.

Conclusion

- Analysis Result

  - 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'.

- Prediction Model

  - Used Multiple Regression Analysis, Random Forest, Gradient Boosting to devise the prediction model

  - Random Forest showed comparatively high accuracy.

  - Shared Key Influential Variables : CO, O3

Appendix

#	Variable	Details	Type	Reason for excluding	EDA			Modelling		
					Graph	Statistic Analysis	Correlation Analysis	Regression	RF	GB
1	MeasDate	Measured Date	Continuous	-						
2	PM10	Particulate Matter 10㎍/㎥	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
5	CO	Nitric Oxide Concentration	Continuous		heatmap, box plot, scatterplot		O	O	O	O
6	SO2	Sulfur Dioxide Concentration	Continuous		heatmap, box plot, scatterplot		O		O	O
7	TEMP	Temperature (°C)	Continuous		heatmap		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
10	WIND_DIR	Wind Direction (16Cardinal Directions)	Continuous	*Converted to Categorical	heatmap, box plot, pie chart		O	O	O	O
11	HUMIDITY	Humidity(%)	Continuous		heatmap				O	O
12	ATM_PRESS	Atmospheric Pressure(hPa)	Continuous	*Converted to Categorical	heatmap			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


Thank You