Air Quality in Korea: Analyzing Daily Trends in Temperature, Precipitation, and PM2.5

Zachary Chase Bonin

Western Governors University

Table of Contents

A. Project Highlights

South Korea has faced an air pollution crisis for decades. Despite air quality improving over time, the issue continues to result in health problems for its nearly 52 million people, which increase healthcare costs, hurt worker productivity, and disrupt entire sectors of the economy. These pollutants originate domestically, from fossil fuel combustion, and internationally, with a significant portion of fine dust pollution from China. The primary pollutant is PM2.5, particulate matter less than 2.5 micrometers in diameter and often more hazardous to human health due to the ease with which it enters the lungs. My project addresses the research question: how do daily temperature and precipitation patterns influence daily and yearly trends in PM2.5 levels across South Korea?

Included in the scope of this project are data import, data cleaning, statistical testing, data visualization, and data modeling. Not included are model deployment, real-time data ingestion, explicit policy recommendations, cost/benefit analysis, and a long-term forecasting system.

Following the CRISP-DM methodology, this project utilizes Python via Jupyter Notebook. Specifically, it uses correlation tests via Python's .corr function, generates visualizations using matplotlib, and creates an XGBoost model using functions from scikit-learn. Model evaluation metrics include RMSE and R^2, and feature importance is considered for each model iteration.

B. Project Execution

The project was executed over two months. The goals of the project were the following:

- Goal 1: To prepare several clean datasets for EDA and modeling use.
    - Objective 1.1: Import data using Pandas and/or an API.
    - Objective 1.2: Clean and tidy data, including the management of missing values and proper joins.
- Goal 2: To use basic statistical methods and visualizations to gain a preliminary understanding of the data.
    - Objective 2.1: Calculate summary statistics.
    - Objective 2.2: Generate correlation matrix and additional correlation coefficients as needed.
    - Objective 2.3: Visualize PM2.5 and temperature/precipitation trends across each of the three regions in South Korea.
- Goal 3: Fit a model to the data and understand feature importance.
    - Objective 3.1: Train a basic XGBoost model.
    - Objective 3.2: Add additional features and lag to improve model performance, as well as tune hyperparameters.
    - Objective 3.3: Analyze feature importance charts to understand which features are of most value to the model.

This project follows the CRISP-DM methodology, excluding the model deployment phase. The business understanding phase contained research into the air quality crisis in South Korea, including the origins of the PM2.5 pollutants and the degree to which they affect the Korean population. Within this phase, I also determined project goals and scope. The data understanding phase included data import via Pandas and/or an API and basic methods in

Python that reveal the shape and quality of the data. The data preparation phase included all work performed directly on the dataset with the intent of modifying it for EDA and modeling use — re-encoding variables, imputing missing values, rearranging columns, etc. After additional EDA, the data modeling phase included creating a base model and refitting with different features and hyperparameters, alongside generating feature importance charts and employing k-fold cross-validation. Lastly, the model evaluation phase included assessments of each model - their RMSE and R^2 - in addition to understanding how each feature contributed to each model.

The project spanned between one and two months and was split into roughly 10-day phases for data import/understanding, EDA, modeling, and report write-ups. Compared to Task 2, there were no significant variances. The modeling process took slightly longer than expected but did not delay the project.

C. Data Collection Process

The first dataset was downloaded from Kaggle as a CSV file and loaded into Python using Pandas. The second dataset was accessed via Python's request library to retrieve data from an API in JSON format. Both methods worked according to plan, and no additional code was needed. No obstacles were encountered.

These datasets did not present data governance concerns. All data is intended for public access and does not contain sensitive or personally identifiable information. Data was imported securely and all modifications made were properly documented.

C.1 Advantages and Limitations of Data Set

How the data was recorded with respect to time presents both an advantage and a disadvantage. In both datasets, pollutant levels and climate readings were recorded once daily at a consistent time. This allowed for simple and efficient visualizations that display annual trends in precipitation, temperature, and PM2.5 levels. However, since climate and pollutant metrics can fluctuate drastically by the hour, the lack of hourly data hindered the modeling stage by limiting the models' abilities to capture short-term variations in the features and target variable.

D. Data Extraction and Preparation

The first dataset was extracted from Kaggle as a CSV file and read into Python via the Pandas read_csv() method. The section taken from the first dataset was mostly complete. One issue was 17 missing values in one column among 3,777 rows. These values were inputted as 0, although PM2.5 values of 0 are not possible, so these values are likely either missing or negligible. Either way, they will not hinder the data analysis. In addition, the 'date' column had to be re-encoded as a DateTime object, and the columns were rearranged for ease. Otherwise, the data was of high quality.

The second dataset was extracted from Open-Meteo using their Weather API. The code to do so utilized requests_cache and retry libraries, as well as Pandas to create DataFrames for each of the three locations. Once imported, the dataset was complete and formatted correctly. Once the time component of the DateTime object was removed, it contained no unnecessary information. As such, it presented no obstacles and required minimal wrangling.

E. Data Analysis Process

E.1 Data Analysis Methods

In addition to calculating basic summary statistics, correlation tests were performed to calculate the correlation coefficient between two variables. Specifically, they estimated the correlation between temperature and PM2.5, and precipitation and PM2.5 across three regions in South Korea. The purpose was to get a rudimentary idea of the relationships between the two features and the target variable prior to modeling. This informed what modeling decisions were made.

In the modeling phase, the XGBoost algorithm, an ensemble method based on gradient boosting, was employed via the scikit-learn library. Due to its high accuracy and ability to capture complex patterns, it is a suitable choice for modeling weather data.

E.2 Advantages and Limitations of Tools and Techniques

Correlation coefficients allow for the simple assessment of relationships between variables with little computational work. However, they are limited to measuring linear relationships; therefore, correlation tests alone are not sufficient in analyzing complex data. The tool used to perform these tests, Python's corr() method, cleanly accesses the two columns and efficiently calculates the correlation coefficients. One potential limitation of corr() is that it only supports three types of correlation; however, this is sufficient for the purposes of this project.

The XGBoost algorithm is an excellent general-purpose model with high accuracy, high efficiency, and built-in measures against overfitting. However, it depends on properly

inputted hyperparameters; without them, the RMSE may suffer significantly. XGBRegressor() from the XGBoost Python package is used to generate the XGBoost models. One advantage is that XGBRegressor() automatically includes L1 and L2 regularization to protect against overfitting. However, one limitation is that it is computationally expensive and uses up a significant amount of memory.

E.3 Application of Analytical Methods

Correlation tests were performed with Python's .corr() method, inputting the two variables and using a loop to repeat the process for all three regions in South Korea. The loc[] method was used to extract the relevant correlation coefficients instead of the 2x2 correlation matrix. Moreover, the shift() method was used to reposition data points and calculate correlation coefficients with a lag implemented. These correlation tests assume linearity, which is not verified to be the case in our data. However, this limitation was acknowledged in the code report, and no definitive conclusions are drawn from these tests.

The XGBoost model was constructed using the scikit-learn library, using data from the Gwanak dataset limited to 2015-2019. A function was constructed to train each model and implement cross-validation by year, whereby each year acts as the test set, representing 4-fold cross-validation. Hyperparameters were adjusted to achieve the lowest RMSE, and aggregate lists of RMSE and R^2 values are computed. Ultimately, several models were constructed, each with different features, although the hyperparameters were tuned to be the same for all models. Features included temperature, precipitation, and their lagged versions (including lagged PM2.5). These lagged features were also set up with the shift() function, whereby

temporary datasets were created to train lagged models and intermediate NA values created in the establishment of these datasets were dropped. Lastly, feature importance charts were generated via the xgb.plot_importance() function.

Although XGBoost can handle missing values, any missing values were previously imputed. The only requirement that XGBoost enforces is numeric data, which is what is contained in our dataset. A sufficient quantity of high-quality data is provided, in addition, so that scaling or other transformations to our data were not necessary.

F. Data Analysis Results

F.1 Statistical Significance

The XGBoost model (supervised regression) is an ensemble method that works via gradient boosting where parallel decision trees are constructed slowly according to gradient descent optimization. Moreover, it has built-in regularization and highly efficient processing. The model was evaluated with Root Mean Squared Error (RMSE) and R^2. RMSE is convenient due to its interpretability; it measures error in the units of the target variable, so the efficacy of the model can be easily understood. R^2 is used to understand how much variability in the target variable is explained by the features used.

Our two final models have R^2 values of 0.121 and 0.425. Although this indicates weak predictive strength, that our values are positive indicates some predictive power is held in temperature, precipitation, and lagged PM2.5 as features. This affirms my hypothesis. However, the feature importance charts indicate that temperature, on the whole, was utilized more than precipitation in predicting PM2.5 values in the test set, which goes against my

hypothesis. Moreover, the feature importance charts revealed that the three most important features are PM2.5, temperature, and precipitation, each lagged by one day.

F.2 Practical Significance

The practical significance of the model can be assessed via its RMSE and R^2. With R^2 values of 0.121 and 0.425, the two final models are limited in their applications. Model 1 can be used to model PM2.5 long-term, but its weak predictive strength prevents much practical use. However, it reveals a relationship between lagged temperature/precipitation and PM2.5, specifically that 1-day lagged temperature is the most significant predictor, aside from lagged PM2.5, and that temperature features, even with lags, are typically more indicative of PM2.5 values than precipitation features. (Basic correlation tests hinted at this, but not as conclusively.) Model 2 may have more practical use due to its stronger R^2, which indicates the supremacy of lagged PM2.5 as a predictor. However, because it requires PM2.5 values from earlier in the week, it can only be used for short-term forecasting, perhaps by a meteorological association, such as modeling the following day's PM2.5.

F.3 Overall Success

Despite not resulting in the creation of strong ($R^2 >= 0.8$) models, the project is successful. To begin, data was successfully wrangled and visualized, and through the modeling process, significant insight was uncovered regarding how each feature and their lagged versions play into the prediction of the target variable. What’s more, these findings can form the base for

future research that aims to forecast PM2.5 based on not only temperature and precipitation, but also wind speed, humidity, and atmospheric pressure, among other factors.

G. Conclusion

G.1 Summary of Conclusions

Over a seven-year period, the temperatures across three regions of South Korea ranged from -16.2°C to 31.6°C (3°F to 89°F), with a mean of around 12.5°C (54.5°F). This suggests a fairly mild climate with possible cold snaps. Precipitation varies more widely, from 0 inches to a whopping 110 inches in a single day. However, on average, 75% of days experience no more than 1.3 inches of precipitation, indicating that extreme rainfall in one day is not common. Our outlier values are likely linked to intense summer monsoons.

PM2.5 values have the greatest variation, ranging from 0 to 195 μg/m3. The average PM2.5 level is 70.9 μg/m3 with 75% of days recording levels of 88 μg/m3 or less. This is considered "moderate," which may cause acute health problems in those with severe respiratory sensitivities. While this may not sound severe, chronic exposure, compounded with "moderate" to "unhealthy" values of other pollutants like PM10 and NO2, can create long-term health issues in large sectors of the population.

There is an inverse trend between temperature and PM2.5 levels, with temperature peaking in the summer (July-August) and PM2.5 peaking in the winter (December-February). This suggests that warmer temperatures have a positive effect on air quality. However, there

appears to be more variation in PM2.5 within months than variation between consecutive months. This variation is more pronounced in Gwanak-gu and Hyeoksin-dong, regions found in the west, where air quality tends to be slightly worse. In contrast, the far-east Cheongnim-dong receives less fine dust pollution from China, resulting in more stable air quality. Unlike temperature, precipitation does not follow seasonal trends as clearly, making it more difficult to assess its relationship with PM2.5 solely through visualizations. That said, the months with the most frequent and extreme precipitation activity (July through September) correlate with lower levels of PM2.5. In addition, there are several days around March 1st, April 1st, and mid-May when a sharp decline in PM2.5 immediately follows a day with heavy precipitation. This follows what we understand about precipitation capturing particulate matter and falling to the ground.

Our correlation coefficients do not indicate any particularly strong relationships. However, they suggest that temperature has a weak inverse effect on PM2.5, and that precipitation has a very weak inverse effect on PM2.5. We expected PM2.5 values to decrease as temperature and precipitation increase, but not necessarily for temperature to have a stronger effect on PM2.5 than precipitation does.

The variations in correlation coefficients between regions suggest that areas farther from China exhibit a stronger relationship between temperature, precipitation, and PM2.5 levels. With Gwanak-gu being the closest to China and Cheongnim-dong the farthest, there is a logical reason why our coefficients are stronger in regions farther from China. As previously suggested, temperature and precipitation are greater factors in determining air quality in these regions due to less pollution from China. Specifically, Cheongnim-dong, which receives the least fine dust from

China, shows the strongest negative correlations. In contrast, Gwanak-gu, more exposed to external pollution sources, displays weaker correlations, suggesting that pollutants from overseas have a greater impact on air quality.

Even with a lag implemented, our highest correlation between temperature and PM2.5 is only ~-0.44 (2-day lag). This suggests that temperature may impact air quality two days later, but that the difference between the 0-day and 2-day correlation is only 0.05 on average, and our sample size is only ~1000, so we cannot state this conclusively. On the other hand, precipitation appears to have an even weaker correlation with PM2.5 when a lag is included, suggesting it has a more rapid effect on PM2.5. These trends hold across the three regions, with the temperature-PM2.5 correlation peaking at a 2-day lag. At the same time, the correlation between precipitation and PM2.5 either drops or remains roughly constant with increasing lag. This suggests that the effects of lag on climate and PM2.5 are consistent across regions despite the relationship between climate and PM2.5 varying across regions.

After the modeling phase, we are essentially left with two models:

1. A model using only lagged temperature and precipitation ($R^2 = 0.121$)
2. A model incorporating lagged temperature, precipitation, and PM2.5 ($R^2 = 0.425$)

Model 1 can theoretically be used to predict PM2.5 values long-term, but its weak accuracy prevents much practical usage. However, it suggests a relationship between lagged temperature/precipitation and PM2.5, specifically that 1-day lagged temperature is the most useful predictor, aside from lagged PM2.5, and that temperature features, even with lags, typically indicate PM2.5 values more accurately than precipitation features. (Basic correlation

tests hinted at this, but not as conclusively.) Model 2 may have more practical use due to its stronger accuracy, which indicates the supremacy of lagged PM2.5 as a predictor. However, because it requires PM2.5 values from earlier in the week, it can only be used for short-term forecasting, such as modeling the following day's PM2.5.

G.2 Effective Storytelling

The first visualization depicts a correlation matrix of all six pollutants, color-coded according to the strength of each correlation pair. This matrix is generated first because it gives an overview of all pollutants rather than our pollutant of focus, PM2.5. As such, it serves as an introductory point from which to jump into deeper analyses and visualizations of our chosen pollutant. It also highlights how high concentrations of one pollutant can correlate with that of another, supporting the idea that investigating one pollutant also sheds light on how climate interacts with other pollutants. The following six visualizations visualized PM2.5 and temperature/precipitation trends across each of the three regions in South Korea. These are essential to understand how our three variables of interest fluctuate every year. The plots feature dual-axes to depict as much information as possible in each plot without overwhelming the viewer and are color-coded for further ease of viewing.

G.3 Recommended Courses of Action

Further research could investigate the relationship between our features and target variables across different regions in South Korea, as the eastern region of Cheongnim-dong was previously observed to have stronger relationships between climate and PM2.5 than the western Gwanak-gu. This would allow meteorological organizations and/or government agencies to tailor

their approach to pollutant control based on the geographical location of each city in question. It should also incorporate more variables that impact air quality, such as wind speed, humidity, and atmospheric pressure. Although the goal of this project was to investigate temperature and precipitation, more advanced models could be constructed with these previously mentioned features. It would also be able to generalize across regions more efficiently.

H. Panopto Presentation

Link:

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=b017934e-9e83-45f2-9120-b2a301219ee0

References

Allabakash, S., Lim, S., Chong, K.-S., & Yamada, T. J. (2022). Particulate Matter Concentrations over South Korea: Impact of Meteorology and Other Pollutants. *Remote Sensing*, *14*(19), 4849. https://doi.org/10.3390/rs14194849

Kim, B. Y., Cha, J. W., Chang, K. H., & Lee, C. (2022). Estimation of the visibility in Seoul, South Korea, based on particulate matter and weather data, using a machine-learning algorithm. *Aerosol and Air Quality Research, 22*(220125). https://doi.org/10.4209/aaqr.220125

Koo, J. H., Kim, J., Lee, Y. G., & et al. (2020). The implication of the air quality pattern in South Korea after the COVID-19 outbreak. *Scientific Reports, 10*, 22462. https://doi.org/10.1038/s41598-020-80429-4

Appendix A

Data Sources

The Kaggle data can be found here: https://www.kaggle.com/datasets/calebreigada/south-korean-pollution. And the Open-Meteo data can be found here: https://open-meteo.com/en/docs.