Exploring the Relationship between AQI and Weather in DANANG

Hoàng Phi Hùng - hunghpde180523@fpt.edu.vn
Nguyễn Ngọc Bảo Quang - lucasnnnguyen1302@gmail.com
Đỗ Tuấn Minh - tuanminhdo1203@gmail.com
Phan Công Tâm - phancongtam190305@gmail.com

7 November 2024

Abstract

This project investigates the relationship between the Air Quality Index (AQI) and weather conditions in Danang City, aiming to understand how weather factors influence air quality. We collected AQI and weather data from various sources and analyzed trends and correlations across several key pollutants, including $PM_{2.5}$, SO_2, PM_{10}, O_3, NO_2, and CO. These pollutants play significant roles in air quality, with $PM_{2.5}$ being the primary contributor to pollution in the region. $PM_{2.5}$ is especially affected by low wind speeds, high humidity, and minimal rainfall, which lead to its accumulation in the atmosphere. SO_2 showed weak correlations with weather factors, while PM_{10} exhibited similar patterns to $PM_{2.5}$, with lower wind speeds and high humidity supporting its build-up. O_3 concentrations were highest during warmer months, and NO_2 increased with low wind speed, low rainfall, and high temperatures. CO levels tended to rise in colder, humid, low-wind conditions. While weather factors did not show strong linear correlations with all pollutants, seasonal patterns indicated indirect effects on air quality. A predictive model for $PM_{2.5}$, using all weather variables, achieved an MAE of 13.65369, MSE of 426.227, and R^2 of 0.4733, suggesting moderate predictive capability. Further research is needed to better understand the full range of factors affecting AQI in Danang.

Keywords: $PM_{2.5}$, weather, predictive model, AQI, Danang, Air Pollution, PM2.5, Weather Correlation, AQI, Machine Learning, Environmental Data, Pollution Prediction, Data Analysis

1 Introduction

Air quality is crucial for both public health and environmental sustainability, with the Air Quality Index (AQI) serving as a key measure to assess pollution levels and their impact on communities. In Danang City, air pollution has become a growing concern, especially with rapid urbanization and increasing industrial activities. The main pollutants, including $PM_{2.5}$, PM_{10} (fine particulate matter), NO_2, SO_2, CO and O_3, can severely affect health, leading to respiratory and cardiovascular diseases.

Danang is located in a coastal area with a tropical monsoon climate, characterized by high humidity and distinct wet and dry seasons. The city experiences heavy rainfall from September to December, which can help wash away some pollutants, but it also faces hot and dry conditions from June to August that can exacerbate air pollution levels. Additionally, the city's topography, which includes mountains and hills surrounding the urban area, can influence air circulation patterns, leading to the accumulation of pollutants in certain areas, particularly in valleys.

Weather factors such as temperature, humidity, wind speed, and rainfall can significantly influence the dispersion and accumulation of these pollutants. For instance, under low wind conditions and high humidity, fine particulate matter tends to accumulate more, resulting in increased AQI levels and heightened health risks. However, the extent and nature of weather's impact on different pollutants remain complex and require further investigation.

This study aims to explore the relationship between AQI and weather conditions in Danang, shedding light on how specific weather patterns may influence air quality. The findings from this research could support policy planning, raise public awareness, and promote measures to mitigate pollution impacts, contributing to a healthier and more sustainable urban environment.

2 Method

2.1 Data Collection

The data collection process was a crucial component of this project, as accurate and comprehensive data are essential for analyzing the relationship between the Air Quality Index (AQI) and weather conditions in Danang City. We utilized two primary tools for this purpose: Selenium and various APIs.

2.1.1 Tool for Crawling

Selenium was used to automate the extraction of data from web pages, especially from dynamic websites that require user interaction to display content. This tool allowed us to gather real-time AQI data. In addition to Selenium, we also leveraged several weather APIs to obtain structured weather data, which facilitated efficient data retrieval and integration into our analysis framework.

2.1.2 Problems

During the data collection phase, we faced several challenges. First, the selected weather API had limitations regarding costs, which raised concerns for large-scale data retrieval. Additionally, the main website for AQI crawling (`https://aqicn.org/city/da-nang/vn/`) lacked sufficient information, particularly in terms of ozone (O_3) levels. Compounding these issues, another potential data source, the local meteorological website (`https://cem.gov.vn/`), was found to be unresponsive for crawling due to significant delay and unfriendly performance issues.

2.1.3 Solutions

To mitigate these challenges, we decided to reach out to the weather API provider via email to explore potential solutions or alternatives for accessing the necessary data. Furthermore, to enhance our AQI data collection, we determined that it would be essential to crawl an alternative website, `https://air.plumelabs.com/`, to ensure we obtained comprehensive air quality data for our analysis.

2.2 Data Storage

Data were stored in a relational database designed to handle time-series data efficiently. The schema was optimized for quick retrieval and processing.

2.3 Data Pre-processing

Main Data Issues

- **Data Gaps:** Significant gaps between 2018-2020.
- **Missing Data:** Many missing values, especially for certain pollutants.

Two Stages:

- **Pre-COVID (before 2018):**
  - Complete $PM_{2.5}/PM_{10}$ data.
  - Missing SO_2, CO, O_3, NO_2.
  - Many outliers present.
- **Post-COVID (after 2020):**
  - Complete SO_2, CO, O_3, NO_2 data.
  - Missing $PM_{2.5}/PM_{10}$.
  - Missing O_3 data after 2024.

Solutions:

- **Option 1:** Merging both stages—**rejected** due to data quality and differences in conditions.
- **Option 2 (The selected option):** Focus on post-COVID (2020-2024) for more complete and relevant data.

Handling Missing Data

Filling Missing Data for 2020 and 2022:

- **2020:** Based on a report indicating a 12.8% drop in PM (both $PM_{2.5}$andPM_{10}) compared to 2019, imputed data using adjusted 2014 data, which is highly correlated to 2019 (increased by change percentage).

$$\text{Change percentage} = \left(\frac{PM_{2019} - PM_{2014}}{PM_{2014}} \right) \times 100\% = 72.5\%$$

- **2022:** Imputed missing data by averaging values from corresponding days in available years.

Solution for Missing O_3 Data: Used Selenium to scrape data from websites.

Handling Small Gaps:

- **Linear Imputation:** Applied to PM data due to temporal continuity.
- **Nearest-Neighbor Imputation:** Used for NO_2, SO_2, O_3, and CO due to abrupt changes in these datasets.

2.4 Data Analysis

In this study, we utilized various tools to analyze and visualize the data. Tools such as `Matplotlib` and `Seaborn` were used for specific purposes: line plots to illustrate trends and scatter plots to analyze correlations. These visualizations helped us derive actionable insights.

To support our analysis, we used the Avien database, which provides air quality data from various monitoring stations. This database allows us to access real-time and historical air quality indicators, providing an overall view of the air quality situation. Utilizing Avien helps us ensure the accuracy and completeness of the data, thereby enhancing the reliability of our analyses and conclusions.

2.4.1 Visualization Techniques

- **Line Plots:** We employed line plots to illustrate the temporal trends of AQI over the studied periods. These plots enabled us to observe fluctuations in air quality and identify any seasonal patterns.
- **Scatter Plots:** Scatter plots were used to analyze the relationships between different air pollutants and the overall AQI. This helped us to visually assess correlations and determine which pollutants had the most significant impact on air quality.
- **Bar Charts:** Bar charts were utilized to examine the distribution of AQI values across different time frames. They provided insights into variability and outliers in the data, helping us understand the range of air quality levels.
- **Heatmaps:** To analyze the correlations between various pollutants, we used heatmaps. These visualizations allowed us to quickly identify which pollutants are positively or negatively correlated with each other, providing a clearer picture of air quality dynamics.

2.4.2 Statistical Analysis

In addition to visualizations, we performed statistical analyses to quantify relationships and trends:

- **Correlation Coefficients:** We calculated Pearson correlation coefficients to quantify the strength and direction of the relationships between AQI and individual pollutants, aiding in identifying significant predictors of air quality.
- **Linear Regression Analysis:** Linear regression analysis showed a relationship between AQI and air pollutants. The regression coefficients show the change in AQI with variations in pollutants.

2.4.3 Insights Derived

The visualizations and statistical analyses provided actionable insights into air quality management. Key findings include:

- Identification of critical periods with elevated AQI levels, which may inform policy decisions for air quality improvement.
- Recognition of pollutants that contribute most significantly to AQI variations, guiding targeted interventions.
- Trends indicating the effectiveness of existing air quality regulations and potential areas for improvement.

2.5 Modeling

In this study, we utilize a multiple linear regression model to predict $PM_{2.5}$concentration based on various independent variables, including temperature (°C), humidity (%), air pressure (inches), visibility (km), wind speed (km/h), daily rainfall (mm), and air pollution indicators such as PM_{10}, SO_2, NO_2, O_3, and CO, all measured in AQI. The seasonal variable, categorized into spring, summer, autumn, and winter, is encoded using One-Hot Encoding to create binary variables representing each season.

The multiple linear regression model can be expressed by the following equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon \tag{1}$$

Where:

- Y is the dependent variable ($PM_{2.5}$ concentration).
- β_0 is the y-intercept (the expected value of Y when all X variables are zero).
- $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients representing the impact of each independent variable $X_1, X_2, \ldots, X_n$ on Y.
- ϵ is the error term, accounting for the variability in Y not explained by the independent variables.

The dataset consists of a total of 1,728 observations, which we divide into two sets: 80% for the training set and 20% for the testing set. To ensure that all input variables are scaled appropriately, we apply the Standard Scaler method for normalization. The Standard Scaler standardizes features by removing the mean and scaling to unit variance, which is calculated as follows:

$$X' = \frac{X - \mu}{\sigma} \tag{2}$$

Where:

- X' is the standardized value.
- X is the original value.
- μ is the mean of the feature.
- σ is the standard deviation of the feature.

This normalization enhances the model's performance by ensuring that each feature contributes equally to the distance computations.

For model evaluation, we employ three metrics:

1. **Mean Absolute Error (MAE)**:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3}$$

   Where:

   - y_i is the actual value.
   - $\hat{y}_i$ is the predicted value.
   - n is the number of observations.

   Meaning: MAE measures the average magnitude of errors in a set of predictions, without considering their direction. It provides a linear score that represents the average error in the same units as the target variable.

2. **Mean Squared Error (MSE)**:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4}$$

   Meaning: MSE measures the average of the squares of the errors—that is, the average squared difference between actual and predicted values. It emphasizes larger errors due to the squaring operation, making it useful for identifying significant deviations.

3. **R-squared ($\mathbf{R^2}$)**:

$$R^2 = 1 - \frac{\text{SS}_{\text{RES}}}{\text{SS}_{\text{TOT}}} \tag{5}$$

   Where:

- $\text{SS}_{\text{RES}} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the sum of squares of residuals.
- $\text{SS}_{\text{TOT}} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the total sum of squares.
- $\bar{y}$ is the mean of the actual values.

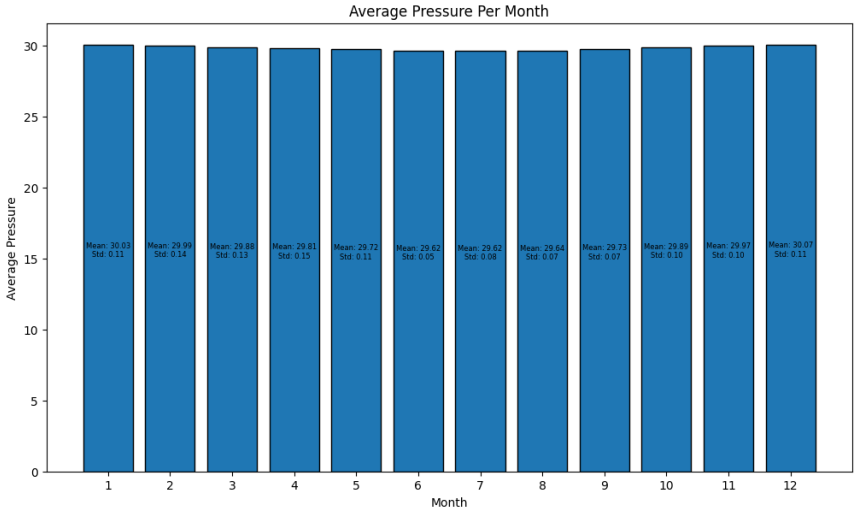
Meaning: $\mathbf{R^2}$ indicates the proportion of variance in the dependent variable that can be explained by the independent variables. An R^2 value close to 1 suggests that a large proportion of the variance is accounted for by the model, while a value close to 0 indicates that the model does not explain much of the variability in the data.

All methodologies and computations are implemented using the Scikit-learn (sklearn) library in Python, ensuring robust implementation and performance evaluation.

3 Result - Discussion

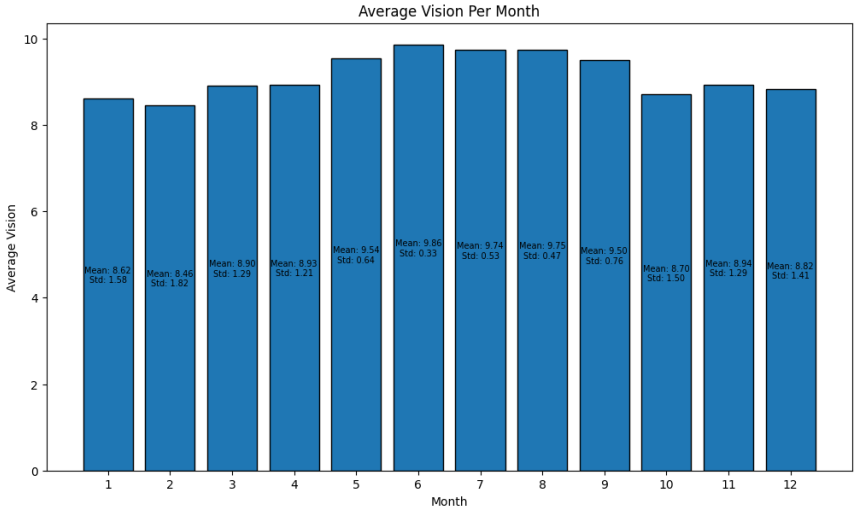
1. **Air Pressure**

   Analysis shows that atmospheric pressure remains stable across years, months, and seasons, with an average range from 29 to 30 inches and a standard deviation between 0.05 and 0.2. This indicates that pressure does not vary significantly over time and is not heavily influenced by weather factors or other atmospheric fluctuations. This stability suggests that atmospheric pressure is minimally impacted by short-term climate changes and other environmental factors.



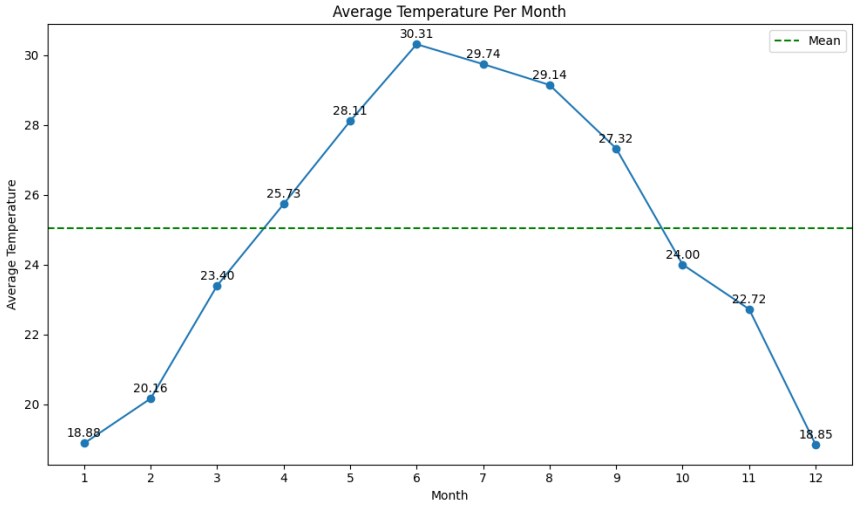
2. **Visibility**

   Visibility remains relatively stable across time periods, with only minor seasonal differences - specifically, visibility decreases by about 1 km in winter compared to other seasons. The average visibility is around 9 km. Overall, visibility tends to be stable and is not significantly influenced by air quality factors. However, there are still slight variations between seasons



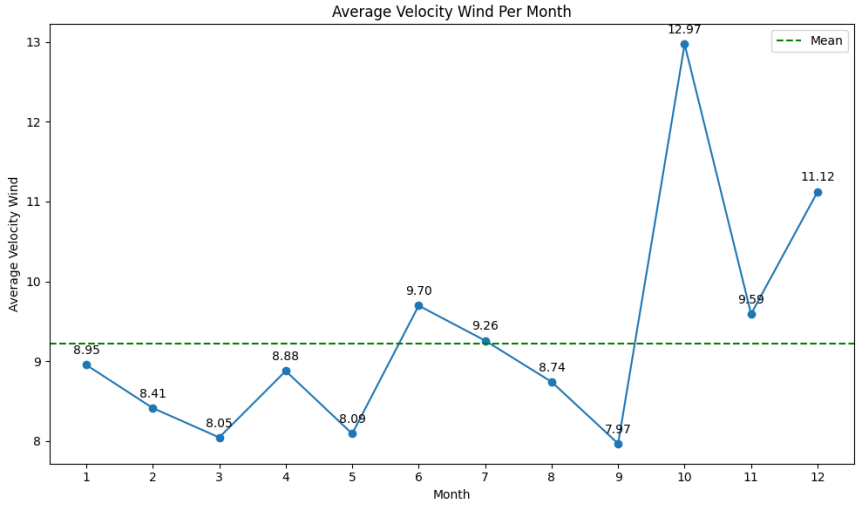
3. **Temperature**

Temperatures are quite similar throughout the year. Spring and autumn are quite cool, only about 25 degrees. Summer is quite hot, the average temperature can reach 30.31 degrees in June. Winter is the coldest season with the lowest average temperature in December and January, only 18 degrees.



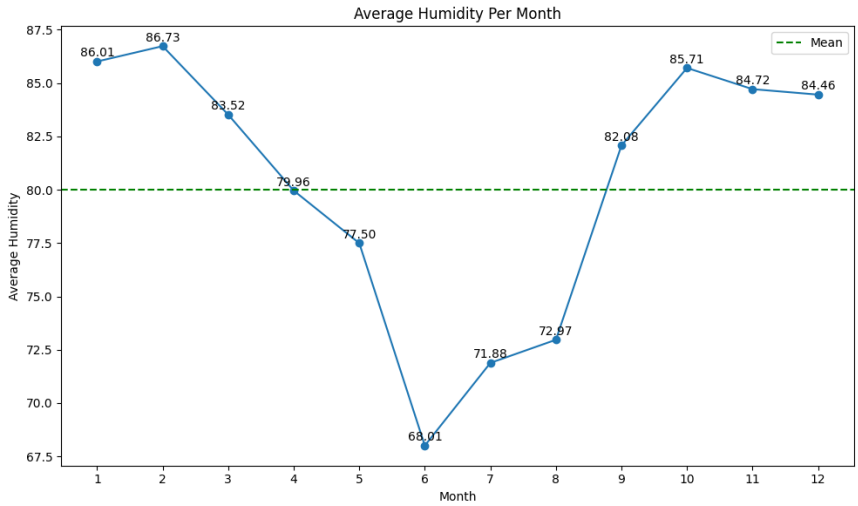
4. **Wind speed**

Wind speed is quite stable from January to September. But from October, wind speed increases quite strongly and at the end of the year wind speed is unstable but fluctuates quite a lot.



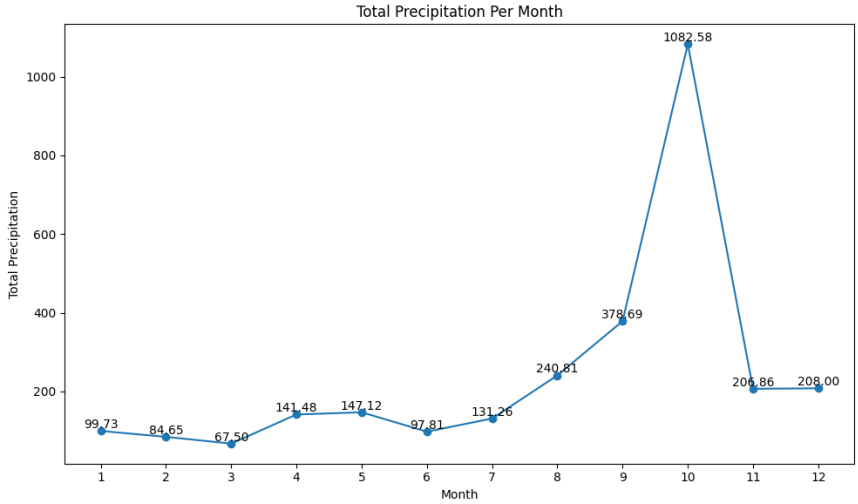
5. **Humidity**

Humidity varies quite a lot between seasons. Humidity is very low in summer, especially the lowest in June, only about 68%. Humidity is high in autumn and winter, more than 84%. The difference in humidity between summer and winter is quite large.



6. **Total Precipitation**

The average total rainfall for the whole year is very high, 2886.5mm. Rainfall is mainly concentrated in the last months of the year, highest in October, accounting for more than 50% of the total annual rainfall.



7. **The Impact of Weather on the AQI of CO**

Although the statistical correlations between the AQI of CO and the weather factors — temperature, rainfall, humidity, and wind speed—are generally low, the graphical analysis reveals distinct patterns. The AQI of CO tends to be higher in colder months and decreases when temperatures rise. Additionally, the AQI of CO is elevated during dry periods with minimal rainfall and drops during heavy rainfall. There is a noticeable relationship between humidity and the AQI of CO, with both increasing in the early and late months of the year, and decreasing during the summer when humidity levels are lower. Wind speed also shows an influence, as strong winds tend to reduce the AQI of CO, while weaker winds allow for higher AQI levels. These findings suggest that while there is no clear linear relationship, weather factors do play a role in influencing the AQI of CO.

8. - **The Impact of Weather on the AQI of NO_2**

Although the correlations between NO_2 and weather factors—temperature, rainfall, humidity, and wind speed—are generally low and indicate no clear linear relationship, the trend charts suggest that these factors still influence NO_2 levels indirectly. Higher temperatures tend to correspond with higher NO_2, and low rainfall in early months is associated with higher NO_2 levels, while increased rainfall in later months leads to a decrease in NO_2. Similarly, lower humidity appears to coincide with higher NO_2, particularly during summer months, while increased humidity leads to a reduction in NO_2. Wind speed also plays a role, as lower wind speeds from January to September are associated with rising NO_2 levels, while higher wind speeds in October lead to a sharp decrease in NO_2. These findings suggest that seasonal changes in weather factors can indirectly impact NO_2 levels throughout the year.

9. - **The Impact of Weather on the AQI of O_3**

   The distribution indicates that the AQI of O_3 in the "Good" air quality zone (0–50 µg/m³) is significantly more prevalent compared to higher concentration zones. This suggests that the majority of recorded AQI values for O_3 remain within acceptable air quality standards. Winter and Autumn: The AQI of O_3 is predominantly within low levels (0–50 µg/m³), indicating generally favorable air quality during these seasons. Spring and Summer: The AQI of O_3 increases, with a higher density of levels exceeding 50 µg/m³, particularly in summer. This seasonal variation suggests that O_3 pollution tends to rise during warmer months.

10. - **The Impact of Weather on the AQI of PM_{10}**

    PM_{10} levels exhibit distinct seasonal variation, with peaks at the beginning and end of each year, particularly in April, and lower levels during the mid-year months. The majority of PM_{10} concentrations fall within the "Good" air quality category, with an average AQI of 20. Over the years, there has been a gradual decrease in PM_{10} levels, suggesting an overall improvement in air quality. Humidity shows a minimal linear correlation with PM_{10}, but higher humidity levels tend to coincide with higher PM_{10} concentrations, suggesting that humidity might contribute to particle accumulation. Finally, while there is no direct correlation between PM_{10} and precipitation, low rainfall (below 10 mm) is often associated with elevated PM_{10} levels, implying that limited rainfall may contribute to higher pollutant concentrations.

11. - **The Impact of Weather on the AQI of $PM_{2.5}$**

    $PM_{2.5}$ levels peak at the beginning and end of the year, particularly in April, with lower levels during mid-year. Most $PM_{2.5}$ concentrations fall within the "Good" air quality category, with an average AQI of 36, and there has been a gradual decrease in $PM_{2.5}$ levels over recent years, indicating an overall improvement in air quality. Temperature, wind speed, and precipitation show minimal correlation with $PM_{2.5}$, though low wind speeds tend to be associated with higher $PM_{2.5}$ levels. Higher humidity levels are often linked to higher $PM_{2.5}$ concentrations, suggesting a role in particle accumulation. Additionally, low precipitation (below 10 mm) is frequently associated with elevated $PM_{2.5}$ levels, implying that limited rainfall may allow higher pollutant concentrations.

12. **The Impact of Weather on the AQI of SO_2**

    Precipitation is not a strong predictor of SO_2 levels, with most data points showing low precipitation (below 100 mm) and SO_2 concentrations ranging from 0 to 50, suggesting an unreliable relationship. There is a slight upward trend in SO_2 levels with increasing temperature, but the relationship is weak due to data dispersion. Wind speed shows a weak negative correlation with SO_2, with low wind speeds associated with higher and more variable SO_2 levels, while higher wind speeds help stabilize SO_2 at lower levels, indicating better pollutant dispersion. SO_2 levels tend to decrease slightly with increasing humidity, with higher humidity (above 80%) corresponding to lower SO_2 values, although the correlation is weak.

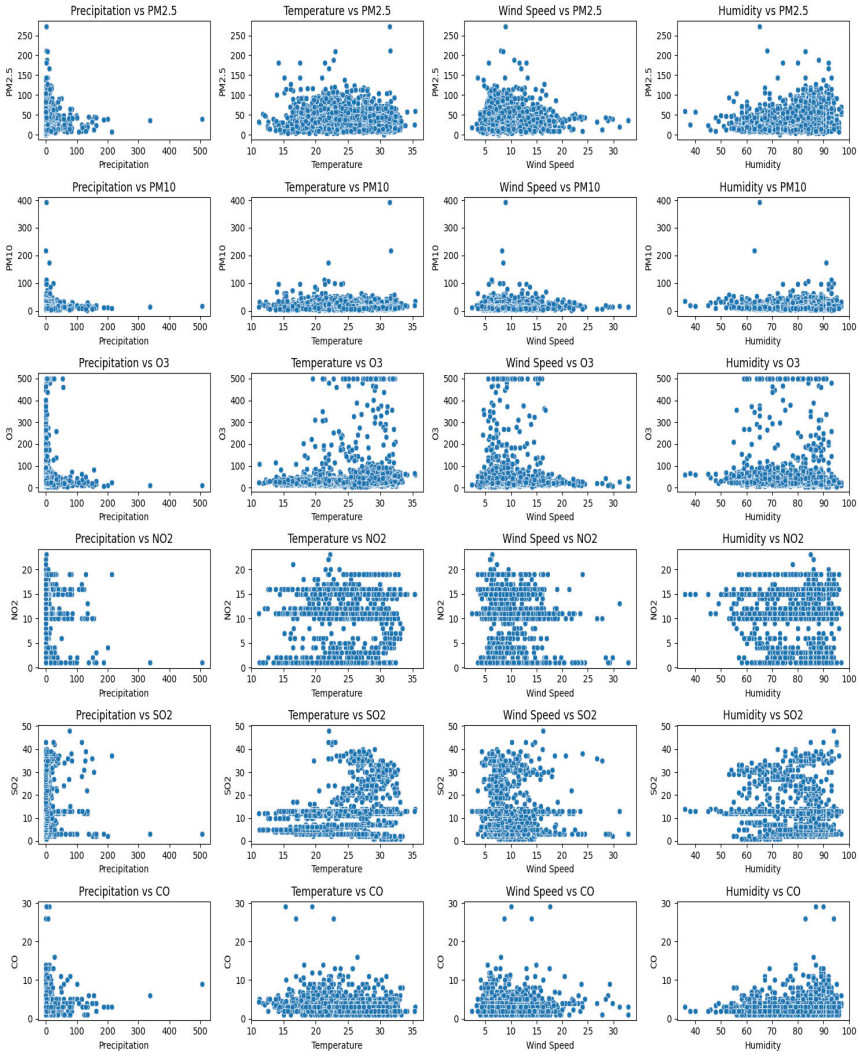


Figure 1: Correlations between weather factors and AQI

13. **Model Result**

The model for predicting the $PM_{2.5}$ index based on features such as temperature, humidity, air pressure, visibility, wind speed, daily rainfall, and air pollution indicators like PM_{10}, SO_2, NO_2, O_3, and CO produced the following results:

- Mean Absolute Error (MAE): 13.65369
- Mean Squared Error (MSE): 426.22731
- R-squared (R^2): 0.47433

With a MAE of 13.65369, the model was able to predict $PM_{2.5}$ levels with a reasonable degree of accuracy, although there are significant errors in some cases. The MSE of 426.22731 indicates that there is a substantial average deviation between the predicted and actual values. The R^2

value of 0.47433 shows that the model explains approximately 47% of the variance in the $PM_{2.5}$ data. This suggests that the model still has room for improvement in predicting $PM_{2.5}$ more accurately.

4 Conclusion

The findings of this study demonstrate a discernible correlation between the weather and climate of Danang city and the air quality, specifically the AQI. By analyzing weather patterns and air quality data from 2020 to early 2024, we identified that Danang's climate can be divided into two distinct seasons—dry and wet—each characterized by noticeable differences in weather patterns. The air quality indicators, such as PM2.5 and NO2, were found to be relatively influenced by weather conditions, while other air quality indices exhibited a weaker connection. These results underscore the significant relationship between these two factors.

While our research on the correlation between weather and air quality in Danang may not be groundbreaking on a global scale, it represents a novel approach within the local context, where such studies are rare. Our project has unveiled the relationship between air quality and weather in Danang, but the linear regression models, statistical parameters, and correlation methods we employed revealed relatively low correlation coefficients. This limitation can be attributed to incomplete AQI data, as well as the lack of comprehensive information regarding industrial development, traffic volume, and construction activities. As a result, this study serves as an important first step, contributing valuable insights for future research endeavors that can refine the models and improve the accuracy of solutions to these challenges.

In conclusion, this study provides a crucial perspective on the relationship between weather and air quality in Danang. It lays a solid foundation for future generations of research, offering the potential to inform more effective policies and strategies that raise public awareness about pollution levels and promote proactive health measures based on known environmental factors.

References

- Ngo Ha (2021), Báo cáo đầu tiên về hiện trạng bụi $PM_{2.5}$ toàn quốc. https://khoahocphattrien.vn/anh-clip/infographic-bao-cao-dau-tien-ve-hien-trang-bui-pm-25-toan-quoc/20211201041523825p1c936.htm
- Aneesh Mathew,P R Gokul,Padala Raja Shekar,Hazem Ghassan Abdo,& Ahmed Abdullah Al Dughairi (2023), Air quality analysis and PM2.5 modelling using machine learning techniques: A study of Hyderabad city in India. https://www.tandfonline.com/doi/full/10.1080/23311916.2023.2243743#d1e446
- Nilesh N. Maltare, Safvan Vahora https://www.sciencedirect.com/science/article/pii/S277250812300011X