HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

o0o

Air Quality Index (AQI) Analysis and Forecasting

Introduction To Data Science - IT4142E

Lecturer: Assoc. Prof. Than Quang Khoat

Students: Nguyen Huy Hoang - 20226041
Hoang Trung Hieu - 20226039
Alexandre Guechtouli - 20250059S

Class : IT4142E - 161306

Hanoi - 2026

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Backround and Motivation

Air pollution has emerged as one of the most critical environmental challenges of the 21st century, posing severe risks to public health and economic development. Rapid urbanization and industrialization, particularly in developing metropolitan areas like Hanoi, have led to fluctuating levels of particulate matter ($PM_{2.5}$, PM_{10}) and other hazardous gases.

While real-time monitoring stations provide current data, the ability to forecast air quality in advance is significantly more valuable. Predictive capabilities allow citizens to plan outdoor activities, enable vulnerable populations to take precautions, and assist local authorities in implementing timely traffic or industrial regulations. This project aims to bridge the gap between historical raw data and actionable future insights using Data Science techniques.

1.2 Problem Statement

The core problem addressed in this project is the prediction of the Air Quality Index (AQI) based on historical time-series data. This is a complex Data Science problem due to the non-linear nature of environmental data, which is influenced by seasonal cycles, rush-hour traffic patterns, and meteorological conditions.

We approach this problem through two distinct analytical lenses:

1. **The Regression Problem**: Predicting the exact numerical value of the AQI to track precise pollution trends.

2. **The Classification Problem**: Categorizing air quality into actionable health labels to simplify communication with the general public.

   - **Good**: AQI 0–50.
   - **Moderate**: AQI 51–100.

- **Unhealthy for Sensitive Groups**: AQI 101–150.
- **Unhealthy**: AQI 151–200.
- **Very Unhealthy**: AQI 201–300.
- **Hazardous**: AQI 301+.

1.3 Project Objectives

The primary objective of this study is to build a robust machine learning pipeline that can ingest raw environmental data and output reliable predictions. Specific goals include:

- **Data Pipeline Construction**: To implement an automated process for ingesting, cleaning, and preprocessing data from both cloud databases and local storage.
- **Feature Engineering**: To extract temporal features (hour of day, day of week) and lag features (past pollution levels) to capture the autoregressive nature of air quality.
- **Model Development**: To train and evaluate multiple machine learning and deep learning algorithms to determine the most effective approach for AQI forecasting.
- **Comparative Analysis**: To empirically compare the performance of regression and classification approaches in the context of environmental monitoring.

1.4 Future Application

The models and pipelines developed in this capstone project serve as a foundational prototype for several real-world applications:

- **Mobile Application Integration**: The prediction engine can be embedded into a mobile app to send push notifications to users when the AQI is predicted to reach ”Unhealthy” levels.
- **Smart City Dashboards**: Integration with city-wide IoT networks to visualize pollution trends on public displays.
- **Route Optimization**: Navigation apps could use this data to suggest ”cleaner” walking or cycling routes away from predicted pollution hotspots.
- **Enhanced Weather Correlation**: Future iterations can incorporate real-time weather API data (wind speed, humidity, temperature) to further refine prediction accuracy.

Chapter 2

Data Description

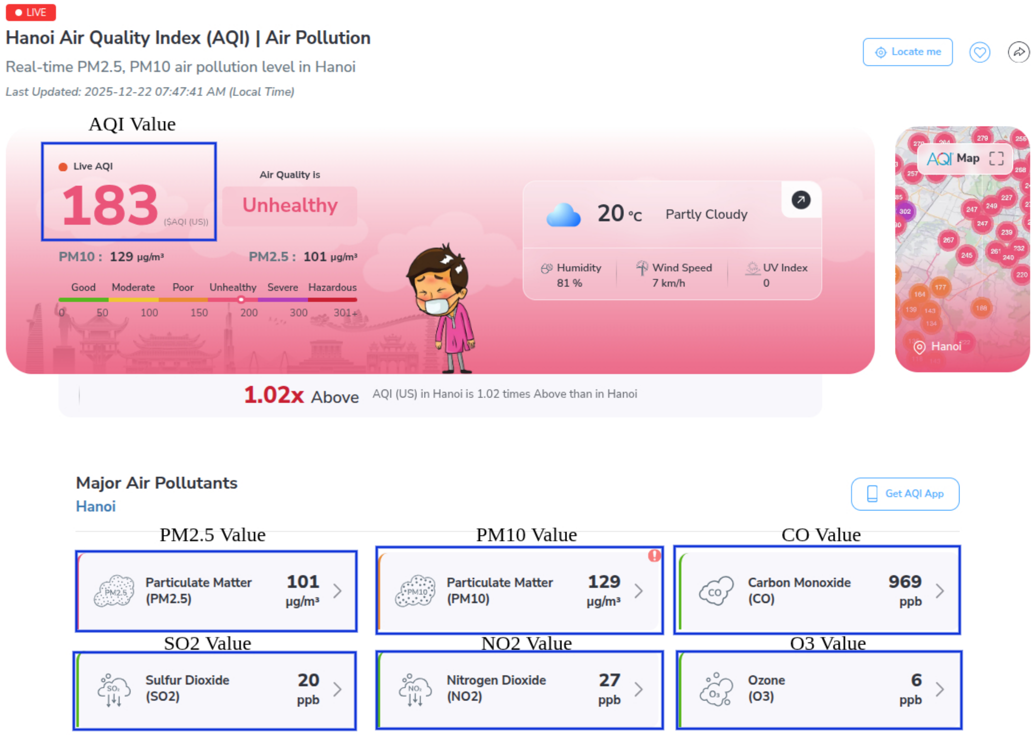
2.1 Data Source

The data used in this study is retrieved from AQI.in, a global open-source air quality monitoring platform managed by **Purelogic Labs**. It serves as a comprehensive dashboard that translates complex environmental measurements into actionable health insights for the public.

AQI.in aggregates its air quality data by combining official regulatory data with high-resolution proprietary technology. The platform integrates real-time feeds from government-run reference stations, which utilize high-precision instruments for baseline environmental monitoring. To fill the geographical gaps between these official stations, AQI.in deploys a vast, global network of thousands of sensors—low-cost monitors that use laser-scattering technology for particulate matter ($PM_{2.5}$ and PM_{10}) and electrochemical sensors for gases like CO, SO_2, NO_2, and O_3.

The air quality data is sourced from aqi.in, a real-time air quality monitoring platform. The dataset focuses specifically on locations within Vietnam, covering major cities and provinces such as Hanoi, Da Nang, Ho Chi Minh City, and various others (e.g., Gia Lai, Ha Nam, Bac Ninh,...).

2.2 Data Crawl Pipeline



Hanoi Air Quality Index (AQI) | Air Pollution

Real-time PM2.5, PM10 air pollution level in Hanoi

Last Updated: 2025-12-22 07:47:41 AM (Local Time)

Live AQI 183 ($AQI (US))

Air Quality is Unhealthy

PM10 : 129 µg/m³ PM2.5 : 101 µg/m³

Good Moderate Poor Unhealthy Severe Hazardous

0 50 100 150 200 300 301+

20 °C Partly Cloudy

Humidity 81 % Wind Speed 7 km/h UV Index 0

1.02x Above AQI (US) in Hanoi is 1.02 times Above than in Hanoi

Major Air Pollutants

Hanoi

Particulate Matter (PM2.5) 101 µg/m³	Particulate Matter (PM10) 129 µg/m³	Carbon Monoxide (CO) 969 ppb
Sulfur Dioxide (SO2) 20 ppb	Nitrogen Dioxide (NO2) 27 ppb	Ozone (O3) 6 ppb

Figure 2.1: AQI value and pollutants on aqi.in.

The data acquisition process is automated using a custom Python script that performs the Extract, Transform, and Load (ETL) operations.

- **Target Identification**: A dictionary constant, **LOCATION**, maps the display names of Vietnamese provinces (e.g., "Ba Ria Vung Tau") to their specific dashboard URLs on the source website. This ensures a consistent and targeted extraction process

- **Extraction Logic**: The script utilizes the *requests* library to fetch HTML content and *BeautifulSoup* to parse the DOM tree. For each location, the scraper iterates through the HTML to identify and extract real-time values for the following pollutants:

  - $PM_{2.5}$ and PM_{10} (Particulate Matter)
  - CO (Cacbon Monoxide)
  - SO_2 (Sulfur Dioxide)
  - NO_2 (Nitrogen Dioxide)

  - O_3 (Ozone)
  - AQI (Overall Air Quality Index)

- **Data Cleaning and Transformations**: Raw data extracted from the web often includes non-numeric characters (units like "μg/m^3" or whitespace). The script employs Regular Expressions (*re* module) to sanitize the strings, converting them into clean integer or float formats suitable for numerical analysis. Additionally, a timestamp is generated at the exact moment of data collection to facilitate time-series analysis.

- **Storage Mechanism**: Once cleaned, the data is batched into a list of tuples and inserted into a **PostgreSQL** database (hosted on Neon) using the *psycopg* library. The script uses an efficient batch insert method to handle multiple location records in a single transaction, ensuring database performance and integrity.

Chapter 3

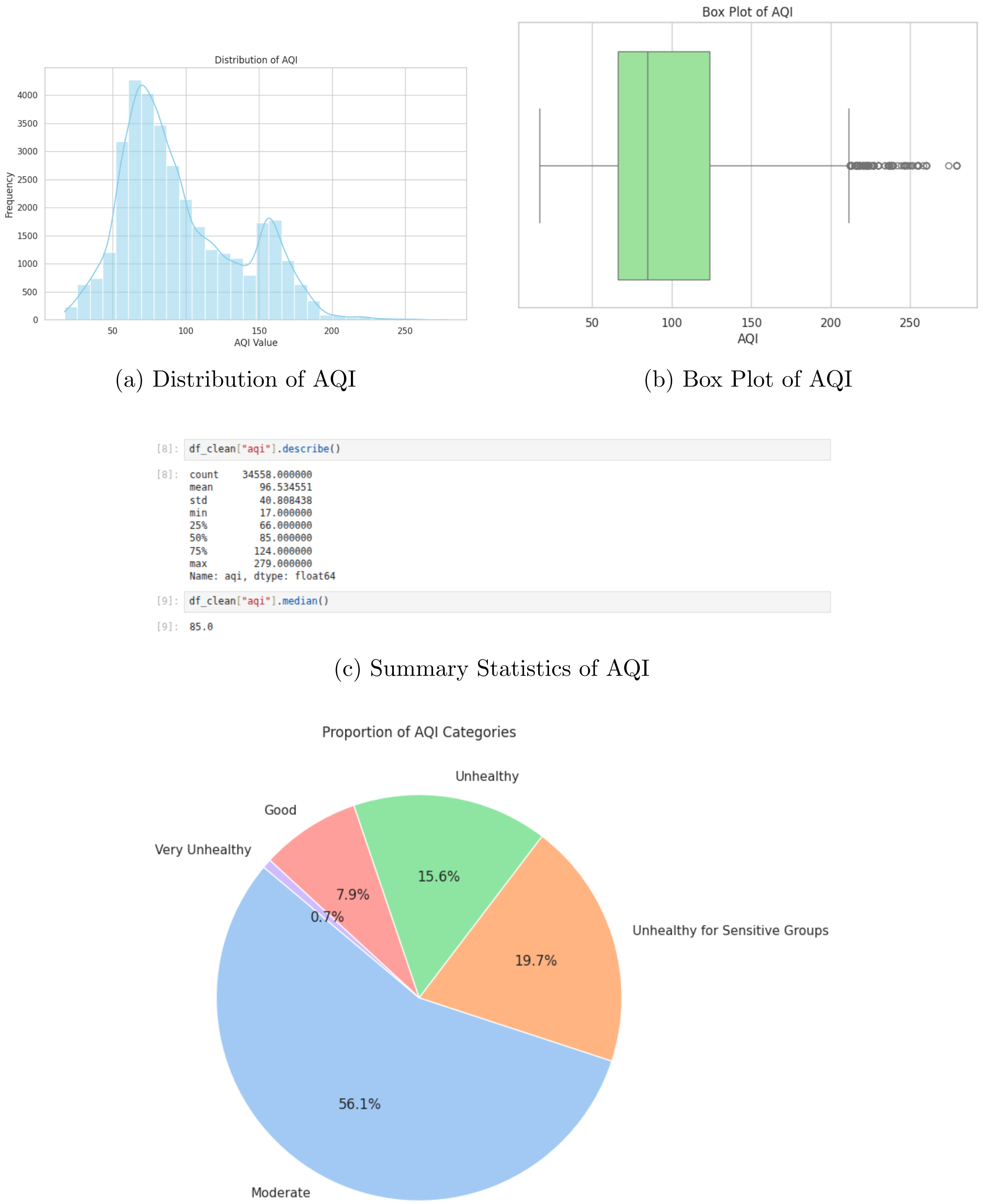
Exploratory Data Analysis (EDA)

3.1 Univariate Analysis

In this section, we conduct a granular analysis of the distribution of individual variables. The primary objective is to characterize the central tendencies, dispersion, and distributional shape of each pollutant. This step is critical for identifying data quality anomalies—such as sensor saturation or capping—and for understanding the underlying stochastic nature of the environmental conditions at the monitored locations.

3.1.1 AQI (Air Quality Index)

The Air Quality Index (AQI) serves as the aggregate metric for communicating overall health risks to the public. Visual inspection of the histogram reveals a distinct **bimodal distribution**, characterized by two separate peaks. The primary mode is centered around an AQI of approximately 70, indicating that for the majority of the observed timeline, the air quality falls within the "Moderate" range (51–100). However, a significant secondary mode appears in the 150–175 range. This secondary "hump" is a critical insight, suggesting a specific subset of recurring environmental conditions—such as rush-hour traffic, seasonal agricultural burning, or specific industrial cycles—that consistently push air quality into the "Unhealthy" category.



(a) Distribution of AQI

(b) Box Plot of AQI

```
[8]: df_clean["aqi"].describe()

[8]: count    34558.000000
     mean        96.534551
     std         40.808438
     min         17.000000
     25%         66.000000
     50%         85.000000
     75%        124.000000
     max        279.000000
     Name: aqi, dtype: float64

[9]: df_clean["aqi"].median()

[9]: 85.0
```

(c) Summary Statistics of AQI

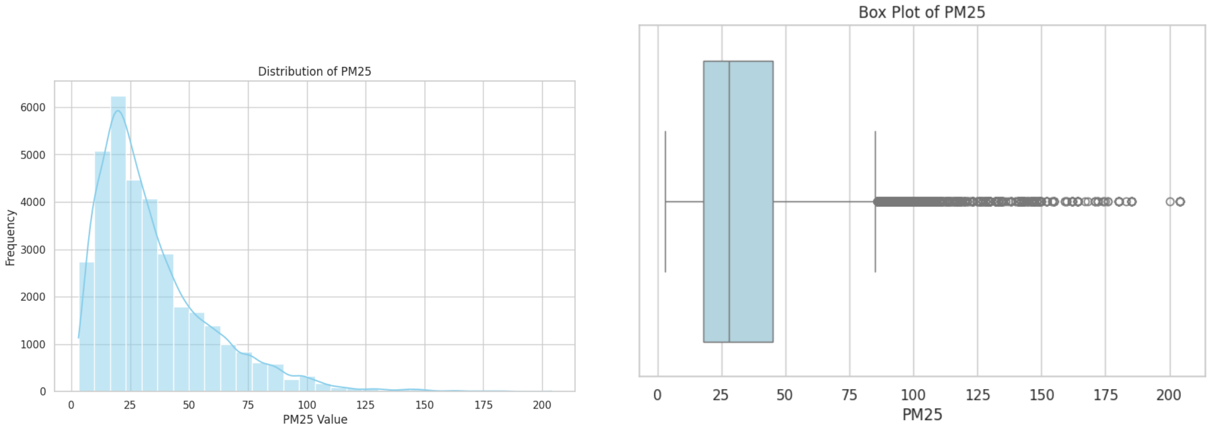
(d) Categories Statistics of AQI

Figure 3.1: Comprehensive analysis of AQI.

Further analysis of the summary statistics and box plots elucidates the severity of pollution events. The distribution is right-skewed, with a median of 85 and an Interquartile Range (IQR) spanning from 66 to 124. Notably, the box plot reveals a dense, continuous cluster of outliers ranging from 210 to 260. The continuity of these outliers implies that these ”Very Unhealthy” readings are not merely random sensor errors or isolated anomalies, but rather represent persistent extreme pollution events. Statistically, while values above 200 are less frequent, their consistent presence significantly impacts the mean, pulling it toward higher risk levels.

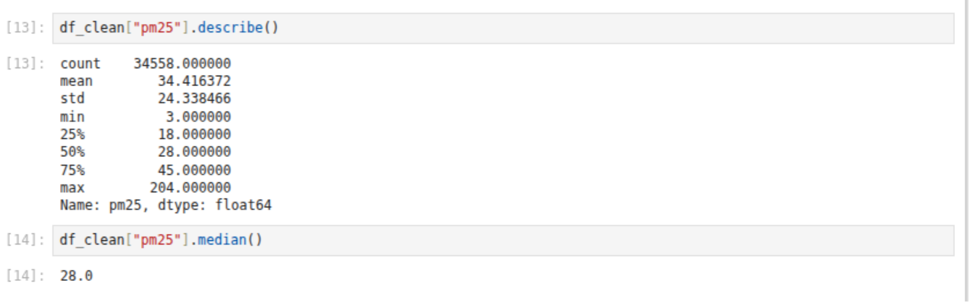
3.1.2 $PM_{2.5}$ (Fine Particulate Matter)

$PM_{2.5}$ (particulate matter less than 2.5 micrometers in diameter) follows a notably different distributional shape compared to the aggregate AQI. The data exhibits a classic **Log-Normal Distribution** with a heavy right tail. The distribution is unimodal, with the highest frequency of readings (the mode) clustering between 15 and 25 $\mu g/m^3$. This concentration at the lower end suggests that "Good" to "Moderate" air quality constitutes the baseline for these locations.



(a) Log-Normal Distribution of $PM_{2.5}$

(b) Box Plot showing heavy outliers



```
[13]: df_clean["pm25"].describe()

[13]: count    34558.000000
      mean        34.416372
      std         24.338466
      min          3.000000
      25%         18.000000
      50%         28.000000
      75%         45.000000
      max        204.000000
      Name: pm25, dtype: float64

[14]: df_clean["pm25"].median()

[14]: 28.0
```

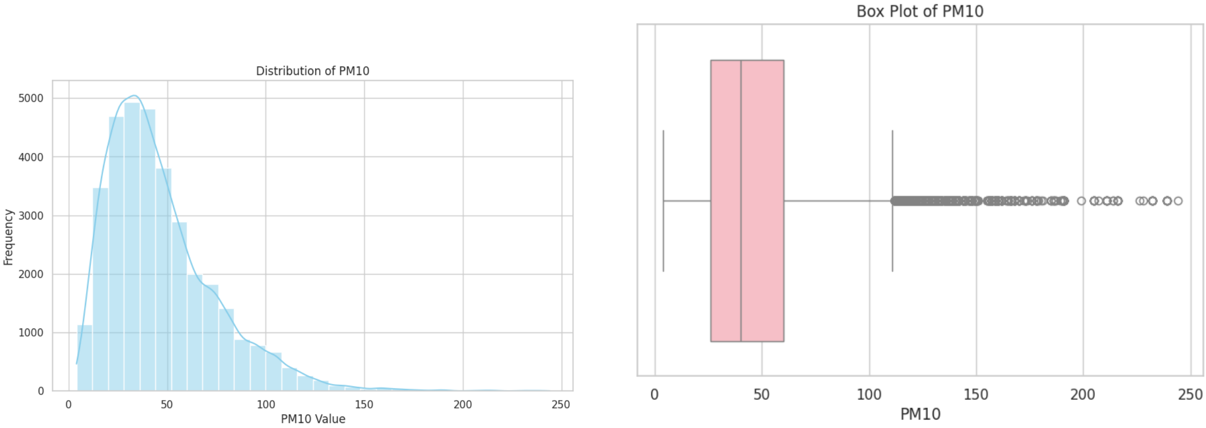
(c) Summary Statistics of $PM_{2.5}$

Figure 3.2: Analysis of $PM_{2.5}$.

The "heavy tail" behavior is evident in the histogram bars that extend continuously out to 200 $\mu g/m^3$. While these high-pollution events are less frequent, they are consistent enough to pull the mean (34.4) significantly higher than the median (28.0). A crucial comparative insight is that while the AQI distribution is bimodal (showing a second "Unhealthy" peak), the $PM_{2.5}$ distribution is unimodal. This suggests that $PM_{2.5}$ is likely not the sole driver of the secondary "Unhealthy" peak observed in the AQI data, necessitating a closer examination of other pollutants.

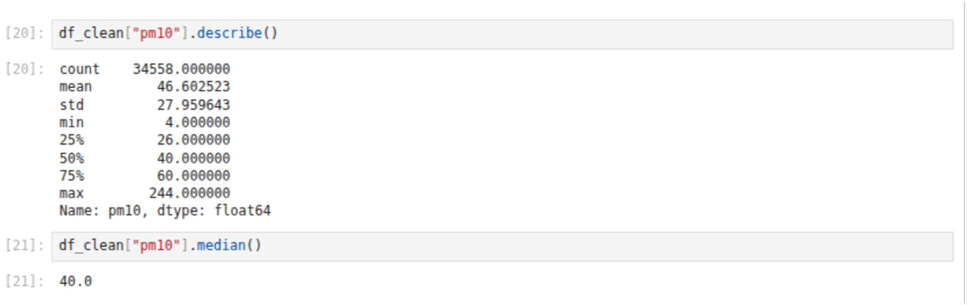
3.1.3 PM_{10} (Coarse Particulate Matter)

The distribution of PM_{10} shares similar right-skewed characteristics with $PM_{2.5}$ but operates on a broader baseline. The mode of the distribution lies between 25 and 45 $\mu g/m^3$, with the bulk of the mass concentration falling below 100 $\mu g/m^3$. This indicates that coarse particulate matter levels are generally stable for the majority of the observation period.



(a) Distribution of PM10

(b) Box Plot of PM10



```
[20]: df_clean["pm10"].describe()

[20]: count    34558.000000
      mean        46.602523
      std         27.959643
      min          4.000000
      25%         26.000000
      50%         40.000000
      75%         60.000000
      max        244.000000
      Name: pm10, dtype: float64

[21]: df_clean["pm10"].median()

[21]: 40.0
```

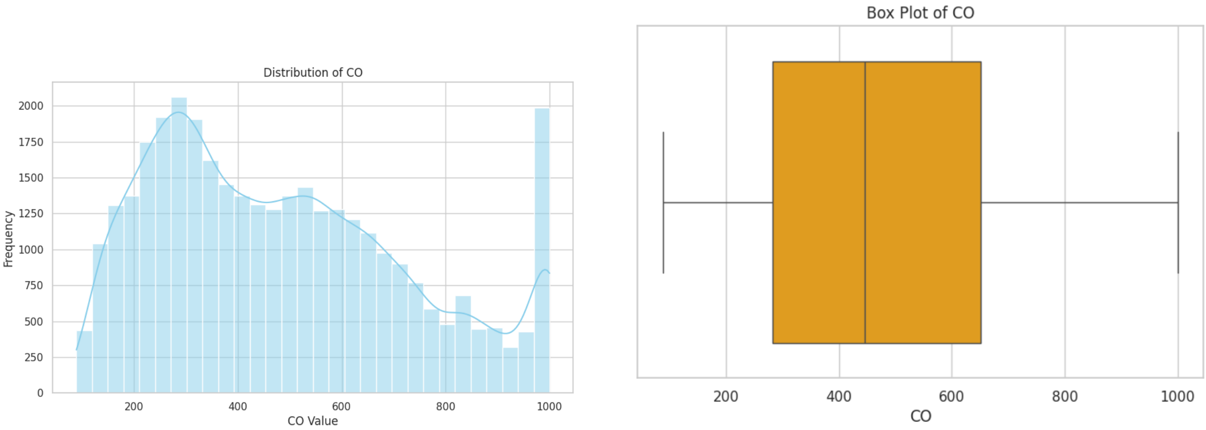
(c) Summary Statistics of PM10

Figure 3.3: Analysis of PM10.

However, the ”long tail” of the PM10 distribution is particularly revealing. The histogram shows continuous, low-frequency bars extending from 100 up to 250 $\mu g/m^3$. Unlike a normal distribution where probabilities would drop to zero rapidly, these high values persist, forming a dense cluster of outliers starting from 112 $\mu g/m^3$. This pattern suggests that severe pollution events—likely attributable to construction dust, road dust resuspension, or specific industrial activities—are a recurring environmental feature rather than isolated anomalies.

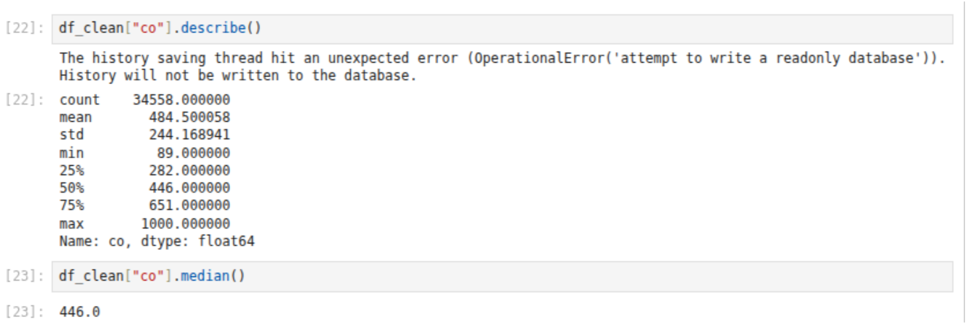
3.1.4 CO (Carbon Monoxide)

The analysis of Carbon Monoxide (CO) reveals a complex, **multimodal distribution** that contrasts sharply with the particulate matter profiles. The histogram displays a primary peak around 275–300 and a broader secondary plateau between 450–650. This multimodality suggests that CO levels are influenced by distinct, cycling factors, likely differentiating between background levels, rush-hour traffic contributions, and industrial output.



(a) Multimodal Distribution of CO

(b) Box Plot of CO



```
[22]: df_clean["co"].describe()

      The history saving thread hit an unexpected error (OperationalError('attempt to write a readonly database')).
      History will not be written to the database.

[22]: count    34558.000000
      mean       484.500058
      std        244.168941
      min         89.000000
      25%        282.000000
      50%        446.000000
      75%        651.000000
      max       1000.000000
      Name: co, dtype: float64

[23]: df_clean["co"].median()

[23]: 446.0
```

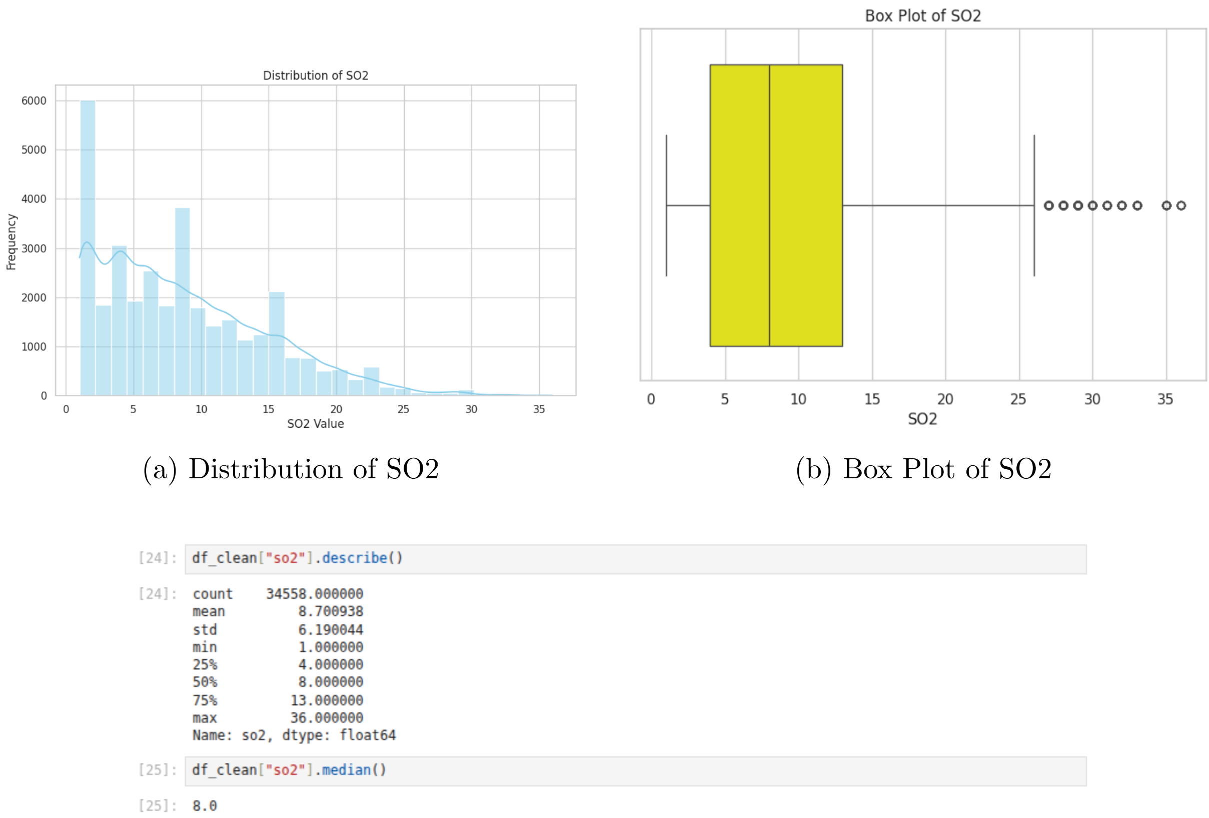
(c) Summary Statistics of CO

Figure 3.4: Analysis of CO.

The most critical finding in the CO data is the presence of a massive, isolated spike exactly at the **1000** mark. This ”terminal spike” is a clear indicator of **sensor saturation** (data capping). It implies that the actual CO concentration exceeded the sensor's maximum measurement limit, causing all higher values to be recorded as exactly 1000. This artificial ”wall” in the data is a significant quality issue that must be addressed during the modeling phase, as it masks the true variance of extreme CO events.

3.1.5 SO_2 (Sulfur Dioxide)

Sulfur Dioxide (SO_2) levels are characterized by a very low baseline, with the highest concentration of data points occurring between 1 and 2. Unlike the unimodal particulate distributions, the SO_2 histogram exhibits a **multimodal right-skew**, featuring distinct secondary "humps" at approximately 4, 8, and 15. These specific peaks likely correspond to the operational cycles of distinct pollution sources, such as industrial discharge schedules or heavy shipping traffic.



(a) Distribution of SO2

(b) Box Plot of SO2

```
[24]: df_clean["so2"].describe()

[24]: count    34558.000000
      mean         8.700938
      std          6.190044
      min          1.000000
      25%          4.000000
      50%          8.000000
      75%         13.000000
      max         36.000000
      Name: so2, dtype: float64

[25]: df_clean["so2"].median()

[25]: 8.0
```

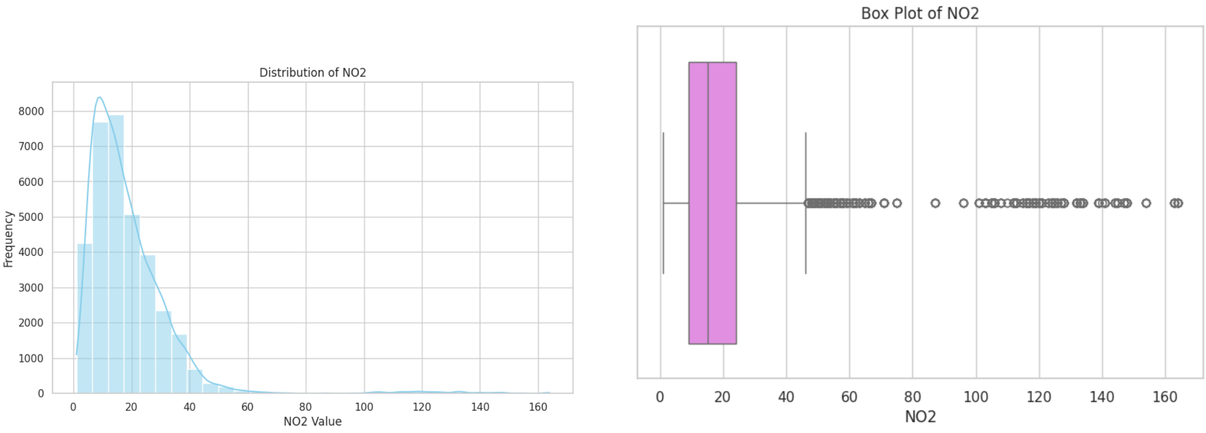
(c) Summary Statistics of SO2

Figure 3.5: Analysis of SO2.

The distribution tail thins out rapidly, reaching a maximum value around 36. Box plot analysis identifies a set of outliers ranging from 27 to 36, representing rare high-pollution events. Notably, unlike the CO data, there is no evidence of a terminal spike, suggesting that the sensors were operating within their effective measurement range for this pollutant.

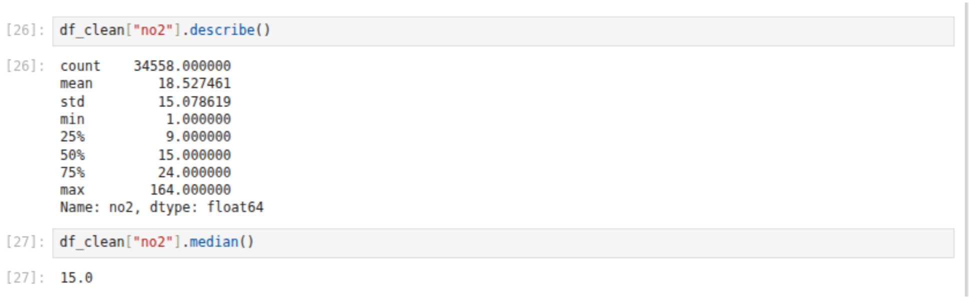
3.1.6 NO_2 (Nitrogen Dioxide)

Nitrogen Dioxide (NO_2), a primary indicator of traffic-related pollution, exhibits the most extreme statistical properties among all analyzed features. The vast majority of the data is tightly clustered below 40, with a primary mode between 10 and 20.



(a) Highly Skewed Distribution of NO2

(b) Box Plot showing extreme outliers



```
[26]: df_clean["no2"].describe()

[26]: count    34558.000000
      mean        18.527461
      std         15.078619
      min          1.000000
      25%          9.000000
      50%         15.000000
      75%         24.000000
      max        164.000000
      Name: no2, dtype: float64

[27]: df_clean["no2"].median()

[27]: 15.0
```

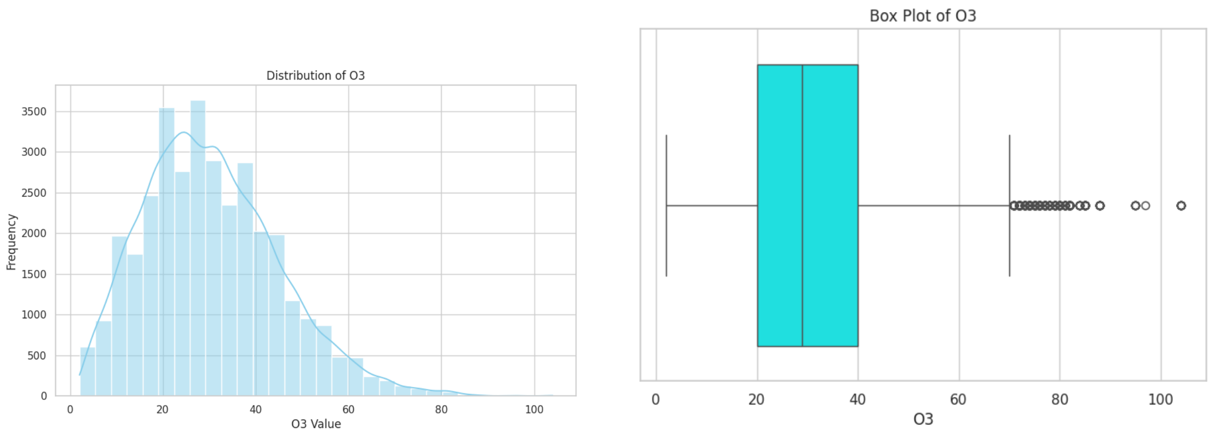
(c) Summary Statistics of NO2

Figure 3.6: Analysis of NO2.

A unique feature of the NO_2 distribution is the significant ”gap” in the data between values of 60 and 100. Beyond this gap lies a sparse cluster of extreme outliers ranging from 100 to 165. Given that these values are more than ten times the median (15), they likely represent severe, localized pollution plumes—potentially resulting from extreme traffic congestion or specific industrial incidents. The separation of these extremes from the main body of data underscores the non-linear, bursty nature of NO_2 pollution.

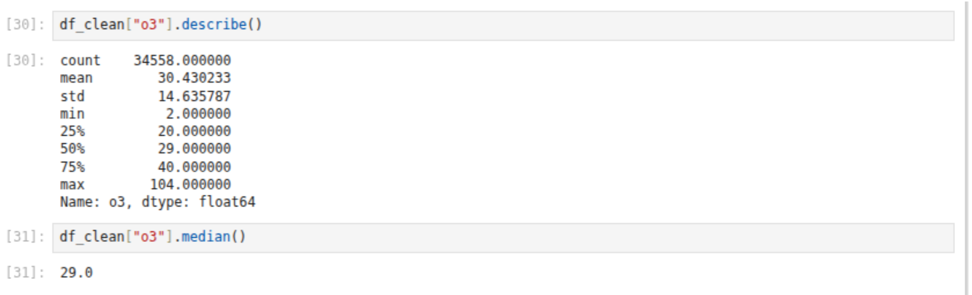
3.1.7 O_3 (Ozone)

The distribution of Ozone (O_3) differs fundamentally from the sharp peaks observed in particulate matter and nitrogen dioxide. Instead of a single sharp mode, O_3 follows a broad, **plateau-like distribution**, with frequencies remaining relatively stable between values of 20 and 40. The median value is 29, and the data is widely spread, reflecting the diurnal nature of ozone formation which is highly dependent on fluctuating environmental factors like sunlight intensity and temperature.



(a) Plateau-like Distribution of O3

(b) Box Plot of O3



```
[30]: df_clean["o3"].describe()

[30]: count    34558.000000
      mean        30.430233
      std         14.635787
      min          2.000000
      25%         20.000000
      50%         29.000000
      75%         40.000000
      max        104.000000
      Name: o3, dtype: float64

[31]: df_clean["o3"].median()

[31]: 29.0
```

(c) Summary Statistics of O3

Figure 3.7: Analysis of O3.

While the distribution is right-skewed, it lacks the extreme, heavy tails seen in other pollutants. The maximum observed value is approximately 104, and readings above 80 are very rare. This suggests that while ozone is consistently present at moderate levels due to atmospheric chemistry, the specific conditions required to generate extreme ozone spikes are uncommon in this dataset.

3.2 Multivariate Analysis

3.2.1 Processing and Feature Engineering

Before conducting multivariate correlation analysis, it was essential to transform the raw time-series data into a structured format suitable for statistical modeling. This phase focused on feature extraction and a rigorous data quality audit to ensure the integrity of the temporal patterns.

Temporal Feature Extraction

To capture the cyclic nature of air pollution, we decomposed the raw `timestamp` into granular temporal features:

- **Hour of Day**: Extracted as an integer (0–23) to analyze diurnal patterns, such as rush-hour peaks.
- **Day of Week**: Extracted as an ordered categorical variable (Monday–Sunday) to differentiate between weekday industrial activity and weekend traffic patterns.
- **AQI Categories**: A categorical feature was engineered by binning the numerical AQI values into standard health impact groups (e.g., “Good” for $0 - 50$, “Moderate” for $51 - 100$, etc.), facilitating classification tasks.

Data Availability Audit and Quality Control

A critical step in our processing pipeline was assessing the continuity of the data. We constructed a frequency heatmap to visualize the density of records across every hour of the week.

The heatmap revealed two significant anomalies in the data collection process:

1. **The Morning Gap (07:00 – 10:00)**: A systematic “vertical stripe” of missing data was observed across all days of the week during the morning rush hours. We hypothesize this was caused by *GitHub Actions* congestion, where scheduled cron jobs faced execution delays or cancellations during peak server load times.
2. **The Sunday Spike**: An abnormally high concentration of records was observed on Sundays at 11:00 AM (double the usual volume), suggesting a retry mechanism dumping backlog data or duplicate execution of the scraper.

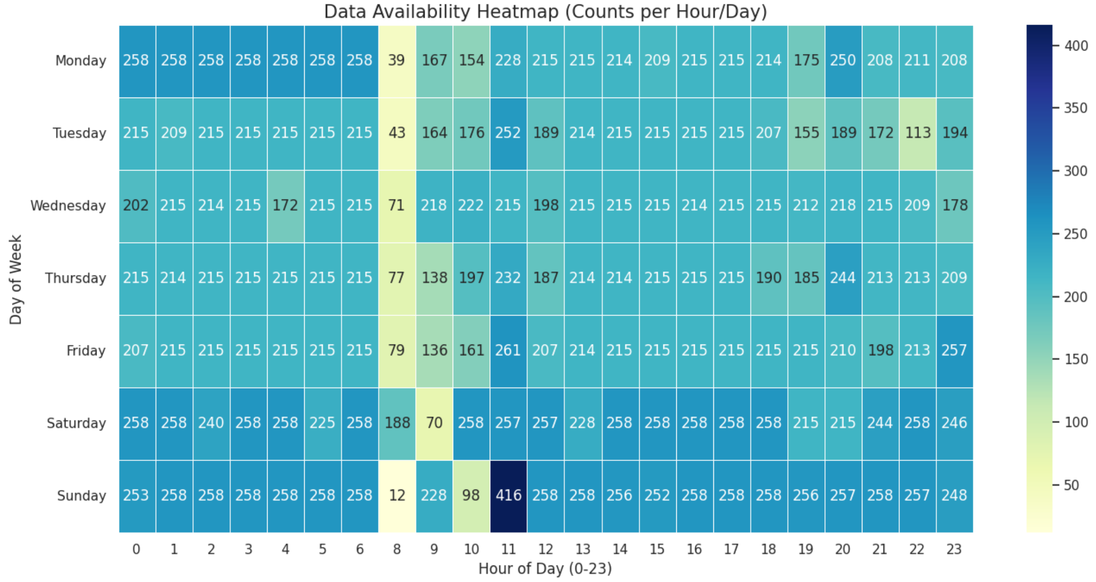


Figure 3.8: Data Availability Heatmap.

Data Imputation Pipeline

To address these inconsistencies without introducing statistical bias, we implemented a three-stage remediation pipeline:

1. **Aggregation (Duplicate Resolution)**: We applied a grouping operation on `location` and `timestamp`, calculating the mean for any duplicate entries. This resolved the "Sunday Spike" by merging overlapping records into single, accurate readings.

2. **Resampling (Gap Detection)**: The raw data was resampled to a strict hourly frequency. This process forced hidden gaps (missing rows) to manifest as explicit `NaN` (Not a Number) values, making the 07:00–10:00 AM gap computationally visible.

3. **Linear Time Interpolation**: We employed linear interpolation with a strict limit of 3 hours (`limit=3`) to fill the identified gaps.

$$y = y_0 + (x - x_0)\frac{y_1 - y_0}{x_1 - x_0} \tag{3.1}$$

where:

- y: The unknown value being estimated at the missing timestamp.
- x: The specific timestamp where data is missing.
- y_0: The known pollutant or AQI value at the last recorded timestamp before the gap.

- x_0: The timestamp of the last known data point before the gap (e.g., 06:00 AM).
- y_1: The known pollutant or AQI value at the first recorded timestamp after the gap.
- x_1: The timestamp of the first known data point after the gap (e.g., 11:00 AM).

This method assumes that air pollution is a continuous physical phenomenon that rises or falls gradually. It connects the known data points at 06:00 and 11:00 to estimate the missing morning values, rather than using a simple average which would flatten the trends.

Imputation Validation (Sanity Check)

To ensure the imputation process did not distort the underlying data distribution, we conducted a comparative density analysis (Sanity Check) between the original (raw) data and the patched dataset.

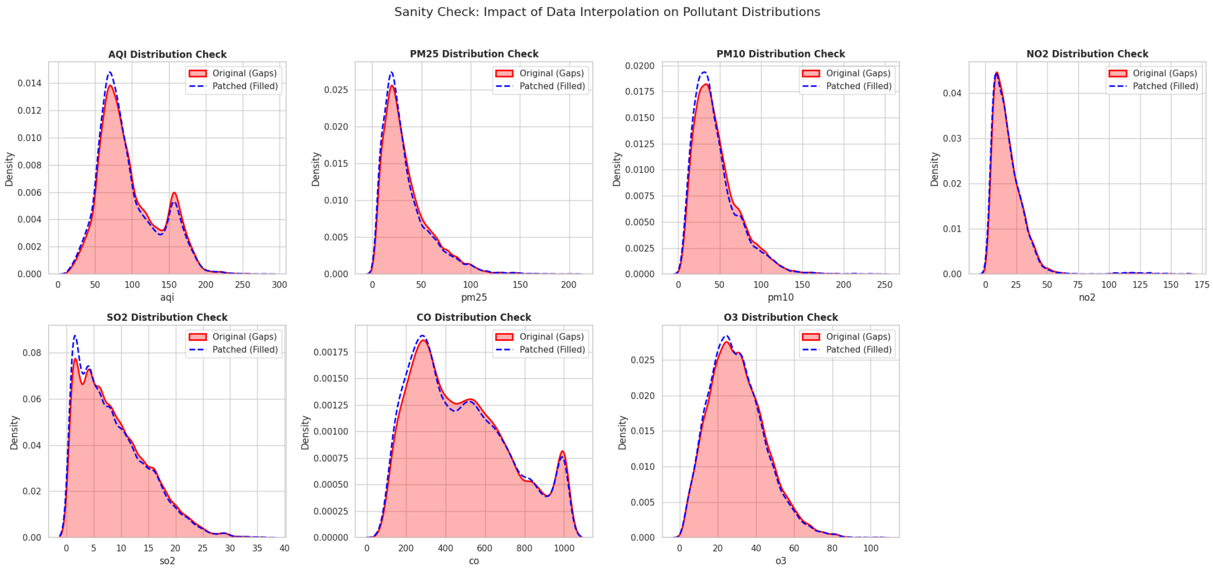


Figure 3.9: Comparison of probability density functions before (Red) and after (Blue) imputation.

The analysis confirmed the robustness of our method:

- **Preservation of Bimodality**: The patched AQI distribution (Blue Line) retained the distinct bimodal shape (peaks at ≈ 70 and ≈ 160) observed in the raw data. This confirms that filling the missing morning hours did not blur the distinction between “Clean” and “Polluted” days.

- **Reinforced Central Tendency**: For pollutants like $PM_{2.5}$ and SO_2, the patched data showed a slightly higher density at the central peaks, indicating that the missing morning values largely followed the standard distribution of the dataset.
- **Absence of Artifacts**: The smooth continuity of the patched lines confirms that no artificial spikes or “walls” of data were introduced, validating the use of time-based interpolation over simple mean imputation.

3.2.2 Correlation and Feature Interaction Analysis

To understand the interdependencies between different pollutants and their collective impact on the Air Quality Index (AQI), we conducted a multivariate analysis using a Correlation Matrix and Pair Plot (Scatter Matrix). The visual analysis reveals distinct structural relationships that are critical for understanding the composition of air pollution in the monitored regions.

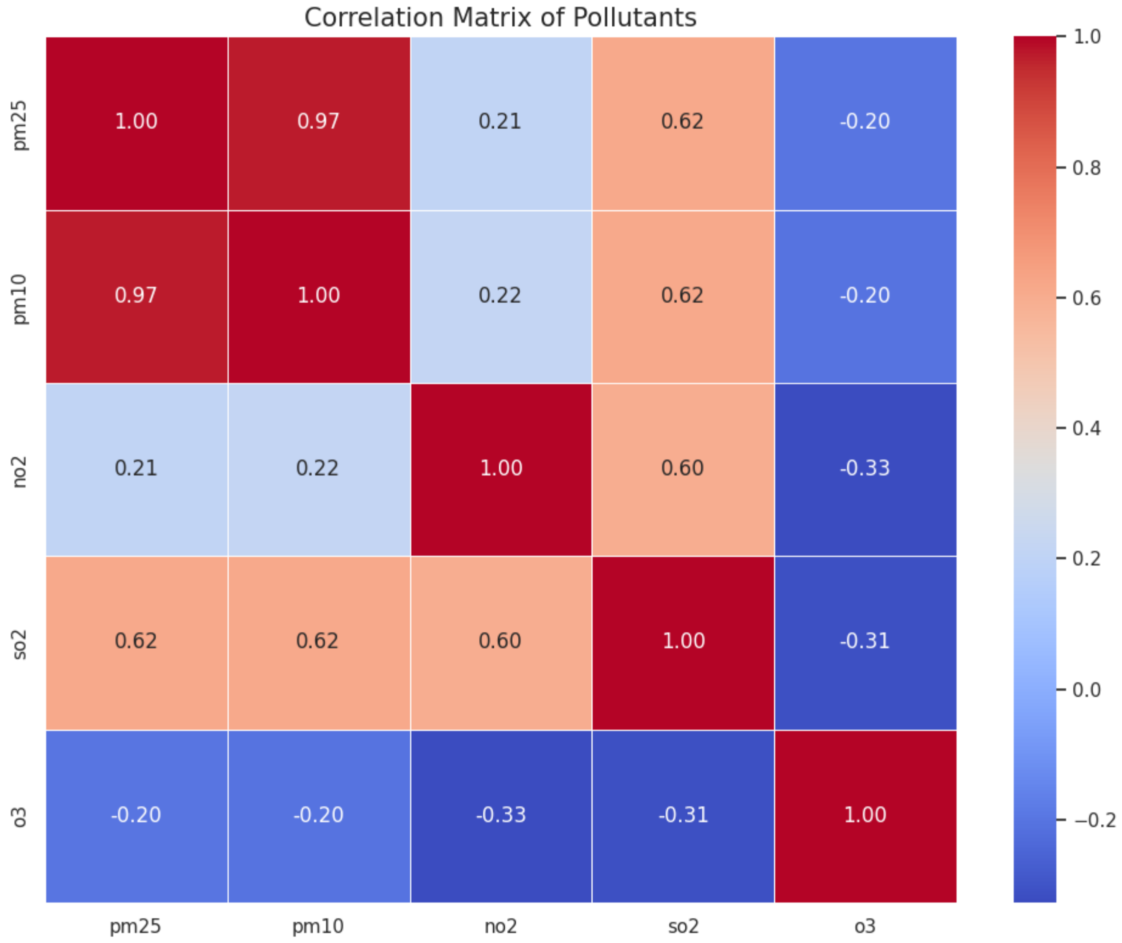


Figure 3.10: Correlation matrix of pollutants.

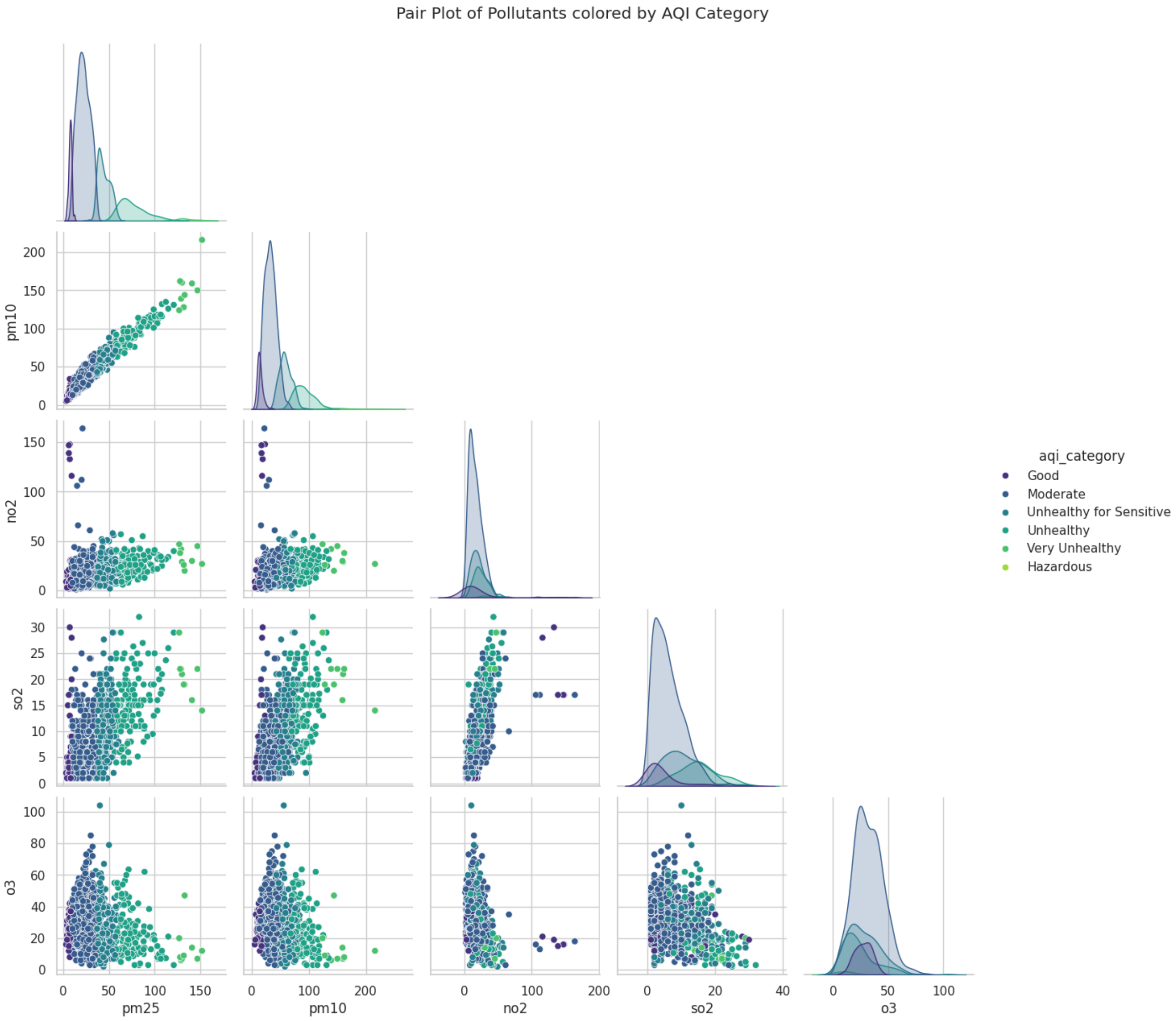


Figure 3.11: Pair Plot of pollutants colored by AQI Category.

The most significant insight derived from this analysis is the dominance of Particulate Matter ($PM_{2.5}$ and PM_{10}) as the primary drivers of AQI. As illustrated in the scatter plots of $PM_{2.5}$ versus AQI, there is a nearly perfect linear correlation, where data points form a tight diagonal line. This indicates that in the monitored locations, the overall health index is almost entirely determined by particulate matter concentrations.

Furthermore, the analysis reveals a high degree of multicollinearity between $PM_{2.5}$ and PM_{10}. The relationship between these two variables is strictly linear, suggesting they provide redundant information regarding pollution trends. In contrast, gaseous pollutants such as NO_2, SO_2, and O_3 exhibit a much weaker correlation with the overall AQI. The scatter plots for these gases appear as diffuse ”clouds” rather than linear trends, implying that high AQI events (indicated by yellow and red hues) can occur even when gaseous pollutant levels are relatively low.

3.2.3 Geographic Pollution Fingerprinting

To investigate whether pollution sources vary by region, we performed a normalized ”Pollutant Fingerprint” analysis. By scaling all pollutants to a range of 0–1 and grouping by location, we identified distinct spatial clusters that characterize the pollution profile of Vietnam.

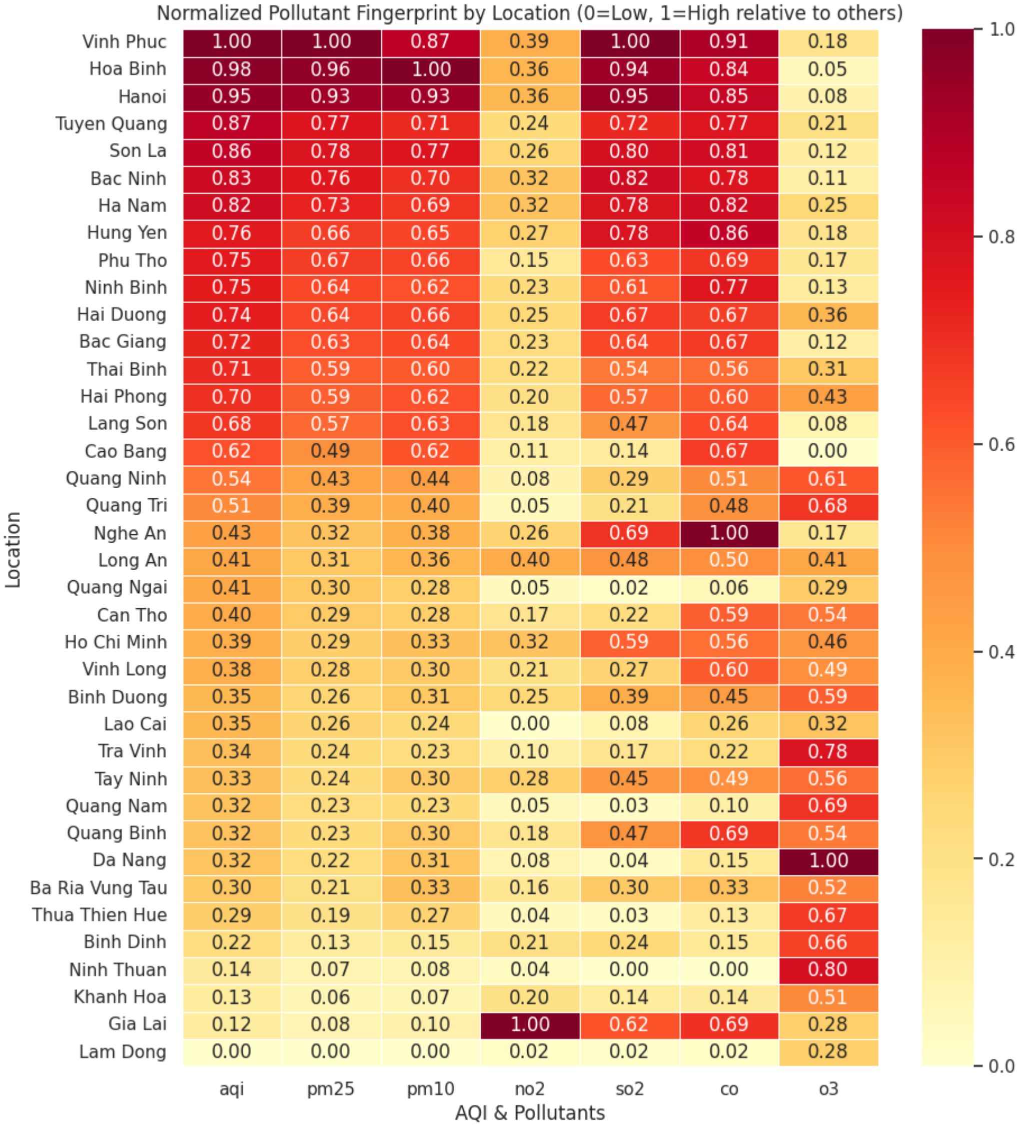


Figure 3.12: Normalized Pollutant Fingerprint by Location.

The heatmap reveals a clear ”Northern Pollution Cluster,” comprising provinces such as Vinh Phuc, Hoa Binh, Hanoi, and Tuyen Quang. These locations exhibit normalized scores near 1.00 for AQI, $PM_{2.5}$, and PM_{10}, confirming that the Northern region suffers primarily from dust and fine particulate pollution. Specifically, Hanoi ranks third in this cluster, characterizing it as a region driven by intense particulate accumulation rather than purely gaseous emissions.

Conversely, an inverse relationship is observed in the Central and Coastal regions, such as Da Nang, Ninh Thuan, and Ba Ria Vung Tau. While these areas exhibit low particulate matter scores (indicated by yellow/orange cells), they show the highest normalized scores for Ozone (O_3). For instance, Da Nang displays one of the lowest PM scores (0.23) but the maximum Ozone score (1.00). This suggests a fundamental divergence in pollution types: while the North battles particulate dust, the coastal regions face challenges related to photochemical smog, likely driven by strong sunlight and different atmospheric dynamics.

Rationale for Location Selection

We focused our granular analysis on Vietnam's three primary economic hubs: Hanoi, Da Nang, and Ho Chi Minh City. These locations were selected to represent the distinct Northern, Central, and Southern geographic zones, offering a comprehensive view of the country's diverse environmental climates. Crucially, our preliminary analysis revealed that these cities exhibit fundamentally different pollution profiles—ranging from Hanoi's particulate-heavy accumulation to Da Nang's ozone-driven coastal air and Ho Chi Minh City's traffic-dependent cycles.

3.2.4 Regional Lifestyle Risk Assessment

Beyond the chemical composition, we analyzed the "Lifestyle Risk" to quantify the frequency of unsafe air conditions across major economic hubs. This analysis highlights a stark disparity in public health risks between the Northern and Southern metropolises.

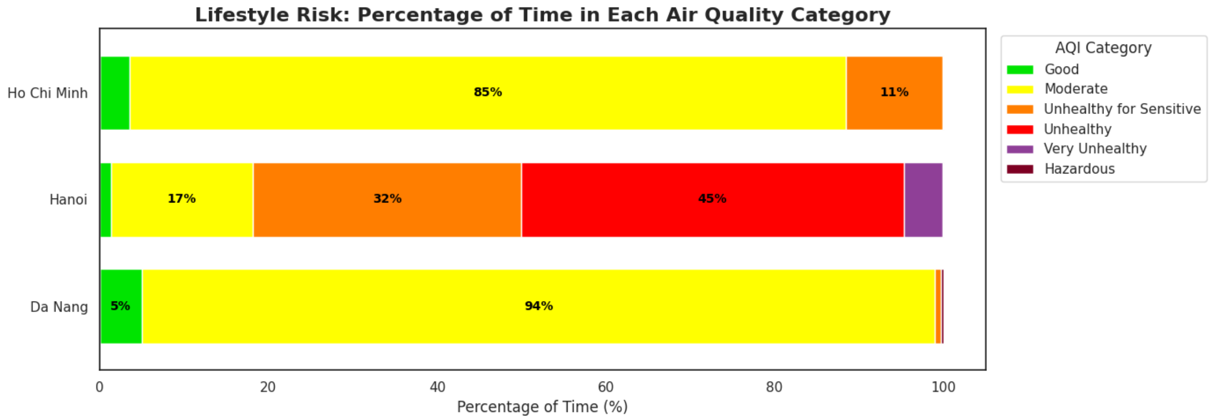


Figure 3.13: Lifestyle Risk Profile of Hanoi, Da Nang and Ho Chi Minh City.

The data exposes an extreme "North vs. South" safety divide. Hanoi operates in a critical state, where the air is unsafe for at least some population groups approximately 77%

of the time (combining ”Unhealthy for Sensitive Groups” and ”Unhealthy” categories). In sharp contrast, cities like Ho Chi Minh City and Da Nang spend the vast majority of their time (85%–94%) in the ”Moderate” zone. This implies that while pollution in the Southern cities is a background environmental factor, in Hanoi, it represents a persistent daily crisis.

3.2.5 Temporal Dynamics and Cyclical Patterns

Daily Rhythm Analysis

To understand the diurnal behavior of pollution, we analyzed the average hourly AQI for the three major cities. The resulting profiles suggest that pollution in Hanoi is driven by different physical mechanisms than in Ho Chi Minh City.

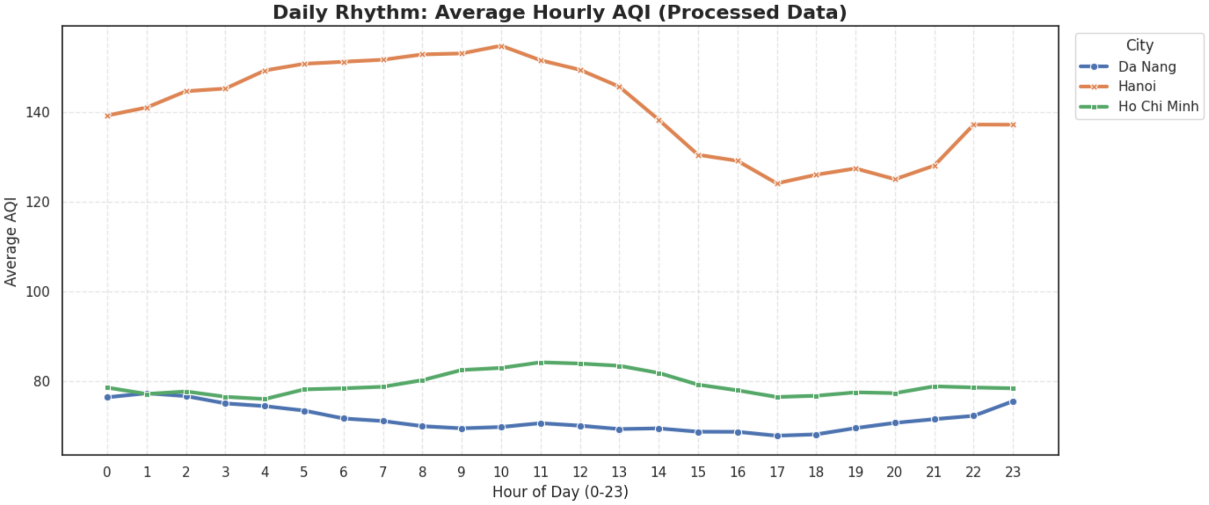


Figure 3.14: Daily Rhythm of Average Hourly AQI of Hanoi, Da Nang and Ho Chi Minh City.

Hanoi exhibits a distinct ”U-shape” pattern that wraps around midnight. Pollution levels are highest at night and early morning, followed by a sharp drop from 12:00 to 17:00. This behavior is consistent with the ”Afternoon Ventilation” effect, where solar heating breaks the atmospheric inversion layer, allowing pollutants to disperse vertically. Once the sun sets, the inversion layer reforms, trapping pollutants near the ground. Conversely, Ho Chi Minh City follows a ”Traffic Hump” profile, characterized by a gentle rise in the morning that plateaus during the day. This suggests that HCMC's air quality is more directly correlated with real-time human activity (traffic and commerce) rather than the intense meteorological trapping observed in the North.

In contrast to the dramatic meteorological swings of Hanoi or the traffic-driven plateau of Ho Chi Minh City, Da Nang maintains the most stable and consistently ”good” air quality of the three cities throughout the 24-hour cycle. Its profile is characterized by a remarkably flat line, with average AQI values hovering between 68 and 76, which is significantly lower than Hanoi's peak. While the other cities show sensitivity to rush hour or inversion layers, Da Nang's AQI actually trends slightly downward during daylight hours and lacks a sharp nighttime spike, suggesting that coastal ventilation likely prevents the stagnation of pollutants.

Autocorrelation Analysis

Finally, we examined the predictability of the AQI time series using Lag Plots to determine the inertial properties of the data.

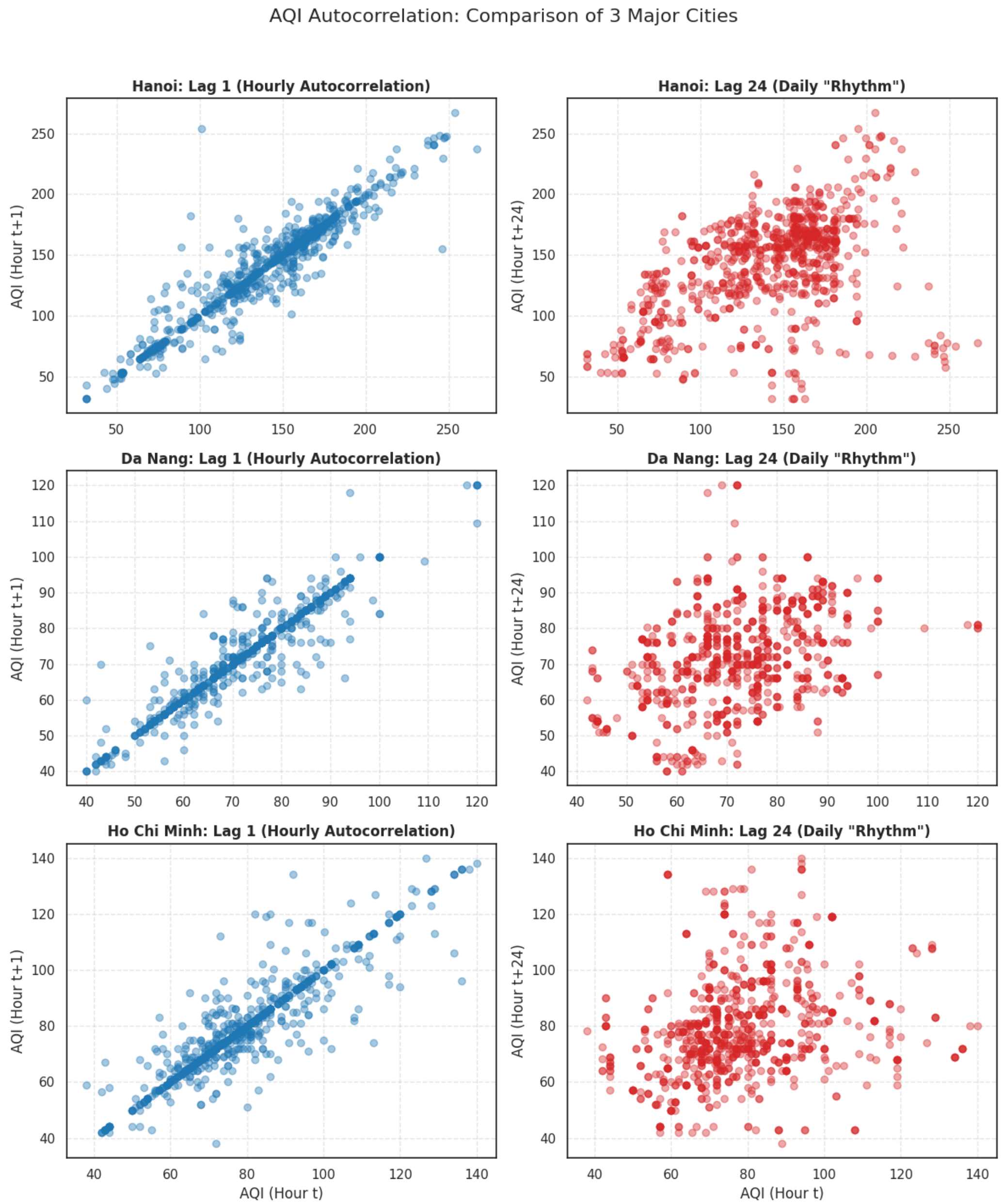


Figure 3.15: AQI Autocorrelation Plots of Hanoi, Da Nang and Ho Chi Minh City.

The Lag 1 analysis (comparing Hour t vs. Hour $t+1$) reveals an extremely strong linear relationship across all cities. The tight clustering of data points along the diagonal confirms that air quality is highly inertial; pollution levels change gradually rather than abruptly, indicating that short-term autoregressive models will be highly effective. However, the Lag 24 analysis (comparing "Same time yesterday" vs. "Today") shows a much weaker, "fuzzy" correlation. This indicates that while there is a general daily rhythm, it is heavily disrupted by changing weather factors (wind, rain, humidity), making simple 24-hour persistence models insufficient for accurate forecasting.

Chapter 4

Air Quality Index Forecasting

To transition from descriptive analysis to predictive modeling, we structured our experiments into two distinct phases: **Phase A (Baseline)** and **Phase B (Enhanced)**. The objective was to empirically quantify the impact of domain-specific feature engineering on model performance. Phase A utilized raw environmental data, while Phase B introduced sophisticated temporal features derived from our Feature Engineering Strategy.

Data Splitting Strategy: Location-Stratified Temporal Split

A critical challenge in environmental time-series modeling is preventing data leakage. A naive random split or a simple index-based split would inadvertently separate data by geography (e.g., training on Northern cities and testing on Southern cities). This creates a *covariate shift*, where the model evaluates its ability to adapt to new regions rather than its ability to forecast the future.

To address this, we implemented a **Location-Stratified Temporal Split**. The dataset was grouped by unique locations (e.g., Hanoi, Da Nang, HCMC). Within each location's independent timeline, the first 80% of observations were designated for training, and the subsequent 20% for testing.

This strategy ensures two critical conditions:

- **Temporal Validity**: The model is strictly evaluated on "future" data relative to the training set, mimicking real-world forecasting scenarios.
- **Geographic Generalization**: Both the training and test sets contain representative samples from all climatic zones (North, Central, South). This guarantees that the reported metrics reflect the model's performance across the entire country, not just a specific region.

Feature Engineering Strategy

Based on the insights from the Exploratory Data Analysis, we implemented the following transformations to address the specific characteristics of the dataset:

- **Cyclical Time Encoding**: The raw `hour` feature (0–23) presents a mathematical discontinuity where 23 is numerically far from 0, despite them being temporally adjacent. To resolve this, we transformed the time into continuous geometric coordinates:

$$Hour_{sin} = \sin\left(\frac{2\pi \times t}{24}\right), \quad Hour_{cos} = \cos\left(\frac{2\pi \times t}{24}\right) \tag{4.1}$$

  This allows the models to understand that 11:00 PM and 1:00 AM are ”close” to each other, preserving the cyclic nature of the daily pollution rhythm.

- **Lag and Rolling Features (Inertia Capture)**: Our autocorrelation analysis (Lag Plots) revealed that air quality is highly inertial—current pollution levels are strongly predicted by the levels 1 to 3 hours prior. To capture this, we engineered:
  - **Lag Features**: $AQI_{t-1}, AQI_{t-2}, AQI_{t-3}$ and AQI_{t-24} (to capture the daily seasonality).
  - **Rolling Statistics**: Rolling means and standard deviations over 3-hour and 6-hour windows. These features are particularly crucial for smoothing the high volatility (”noise”) observed in Hanoi’s data, allowing the models to react to trends rather than isolated spikes.

- **Dimensionality Reduction (Multicollinearity)**: The Pair Plot analysis demonstrated a near-perfect linear correlation ($R \approx 0.97$) between $PM_{2.5}$ and PM_{10}. To prevent multicollinearity from destabilizing our regression weights, we dropped PM_{10} from the input feature set, relying on $PM_{2.5}$ as the primary particulate indicator.

- **Target Transformation**: Given the heavy right-skew of variables like NO_2 and the target AQI (caused by extreme pollution events), we applied a Log-Transformation ($\log(1 + x)$) to the target variable during training. This normalizes the error distribution, preventing the models from over-penalizing outliers.

Phase A: Baseline Input Configuration

In the baseline phase, the models were trained using a ”raw data” strategy to establish a performance floor. The feature set X_{base} consists of the fundamental numerical pollutants and the raw temporal variable as extracted from the source:

- **Raw Pollutants**: $PM_{2.5}, PM_{10}, NO_2, SO_2, CO, O_3$.

- **Linear Time**: The raw `hour` integer ranging from 0 to 23.

- **Spatial Context**: One-hot encoded location variables to allow the model to distinguish between different geographic baselines.

The input vector for Phase A is represented as follows:

$$X_{base} = [P_1, P_2, \dots, P_6, Hour, Loc_1, \dots, Loc_n] \tag{4.2}$$

where P represents the concentration of the six primary pollutants observed in the raw dataset.

Phase B: Enhanced Feature Engineering Input

Phase B introduced domain-specific transformations designed to resolve data quality issues and behavioral patterns identified during the EDA. The enhanced feature set X_{enh} incorporates the following strategies:

- **Cyclical Time Encoding**: To resolve the mathematical discontinuity between Hour 23 and Hour 0, the hour was decomposed into $Hour_{sin}$ and $Hour_{cos}$ coordinates.
- **Autoregressive Lag Features**: Based on the Lag Plot analysis showing high short-term inertia, we included $AQI_{t-1}, AQI_{t-2}, AQI_{t-3}$, and AQI_{t-24}.
- **Rolling Statistics**: To smooth high-frequency noise and capture local trends, 3-hour and 6-hour rolling means and standard deviations were added.
- **Geographic Interaction Terms**: Given that the relationship between Time and AQI varies by city, we included interaction features where the location dummy is multiplied by the cyclical hour coordinates.
- **Multicollinearity Filter**: PM_{10} was removed to prevent redundancy with $PM_{2.5}$ after a near-perfect linear correlation ($R \approx 0.97$) was observed.

The input vector is defined as:

$$X_{enh} = [P_{gases}, T_{cyclic}, L_{inertia}, S_{rolling}, G_{spatial}, I_{interaction}] \tag{4.3}$$

Where the components include:

- **Primary Pollutants (P_{gases})**: $PM_{2.5}, CO, SO_2, NO_2$, and O_3. Notably, PM_{10} is excluded to prevent multicollinearity.
- **Temporal Encoding (T_{cyclic})**: $Hour_{sin}$ and $Hour_{cos}$ coordinates to preserve the cyclic nature of the 24-hour clock.
- **Inertia Features ($L_{inertia}$)**: Historical values ($AQI_{t-1}, AQI_{t-2}, AQI_{t-3}$) and daily seasonality (AQI_{t-24}).

- **Rolling Statistics ($S_{rolling}$)**: Moving averages and standard deviations over 3-hour and 6-hour windows to smooth high-frequency noise.
- **Interaction Terms ($I_{interaction}$)**: Derived features ($Location \times \{Hour_{sin}, Hour_{cos}\}$) to account for regional differences in diurnal rhythms.

4.1 Multivariate Regression

The regression task aimed to predict the exact AQI value for the next hour ($t + 1$). We evaluated four algorithms: Linear Regression (Baseline), Random Forest, XGBoost, and Long Short-Term Memory (LSTM) networks.

4.1.1 Evaluation Metrics

To rigorously assess the predictive capabilities of our regression models, we employed three standard statistical metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2). Each metric provides a different perspective on model error, balancing sensitivity to outliers with average accuracy.

- **Root Mean Squared Error (RMSE)**: RMSE measures the standard deviation of the prediction errors (residuals). It is particularly useful in air quality forecasting because it squares the errors before averaging, thus imposing a heavier penalty on large deviations (e.g., failing to predict a hazardous pollution spike). A lower RMSE indicates a model that avoids major failures.

  $$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4.4}$$

  Where:

  - n: Total number of observations in the test set.
  - y_i: The actual observed AQI value for the i-th hour.
  - $\hat{y}_i$: The predicted AQI value generated by the model for the i-th hour.
  - $(y_i - \hat{y}_i)^2$: The squared residual, which disproportionately penalizes larger errors.

- **Mean Absolute Error (MAE)**: MAE calculates the average magnitude of errors in a set of predictions, without considering their direction. Unlike RMSE, it treats all errors linearly. This metric provides a more intuitive interpretation of ”typical” model performance (e.g., ”the model is usually off by 0.55 units”).

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{4.5}$$

Where:

- $|y_i - \hat{y}_i|$: The absolute difference between the actual and predicted values.

- **Coefficient of Determination (R^2)**: The R^2 score represents the proportion of the variance in the dependent variable (AQI) that is predictable from the independent variables (features like Lag, Hour, etc.). It provides a standardized measure of "goodness of fit" on a scale from $-\infty$ to 1.0.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4.6}$$

Where:

- $\sum(y_i - \hat{y}_i)^2$: The Sum of Squared Residuals (SS_{res}), representing the error of the model.
- $\bar{y}$: The mean of the observed data, calculated as $\frac{1}{n}\sum y_i$.
- $\sum(y_i - \bar{y})^2$: The Total Sum of Squares (SS_{tot}), representing the inherent variance in the original data.

An R^2 of 0.996 (as achieved by our Enhanced models) implies that the model explains 99.6% of the variability in the AQI data, leaving only 0.4% as unexplained noise.

4.1.2 Results

To rigorously evaluate the models, we employed three distinct metrics: Root Mean Squared Error (RMSE) to penalize large outliers, Mean Absolute Error (MAE) to measure average model bias, and the Coefficient of Determination (R^2) to assess goodness-of-fit.

Table 4.1 presents a granular comparison between the Baseline (Phase A) and Enhanced (Phase B) experiments.

Table 4.1: Comparison of Regression Performance (RMSE, MAE, R_2)

Model	**RMSE** (Lower is better)		**MAE** (Lower is better)		R^2 **Score** (Higher is better)	
	Phase A	**Phase B**	**Phase A**	**Phase B**	**Phase A**	**Phase B**
Linear Regression	8.16	6.41	6.01	4.81	0.951	0.971
Random Forest	2.06	1.27	0.95	**0.56**	**0.997**	**0.999**
XGBoost	**1.87**	**1.25**	**0.93**	0.67	**0.997**	**0.999**
LSTM	19.62	27.08	11.03	16.06	0.719	0.498

The rigorous location-stratified split confirms that our models are learning valid temporal patterns rather than overfitting to specific regional baselines.

- **Robustness of Tree-Based Models**: Even under the strict condition of predicting future data for every city, the Random Forest and XGBoost models demonstrated exceptional stability. The Random Forest and XGBoost models achieved Mean Absolute Error (MAE) of **0.56** and **0.67**, respectively, proving that the engineered lag and cyclical features successfully captured the universal physics of pollution dispersal, regardless of the specific city location.
- **Effectiveness of Feature Engineering**: The improvement from Phase A (Baseline) to Phase B (Enhanced) remains significant. This validates that features like “Rolling Mean” and “Cyclical Hour” are robust predictors. They allow the model to distinguish between a noisy spike in Hanoi and a stable trend in Da Nang, reducing the RMSE significantly across the test set.
- **LSTM Sensitivity**: The LSTM model continued to show sensitivity to the input dimensionality. While the temporal split provided a fairer evaluation environment, the LSTM struggled to generalize as effectively as the ensemble methods, likely due to the limited duration of the time-series history available for each individual location after splitting.

4.2 Multivariate Classification

The classification task involved categorizing the air quality into standard health labels (e.g., ”Good”, ”Moderate”, ”Unhealthy for Sensitive Groups, ”Unhealthy”, ”Very Unhealthy”, ”Hazardous”). A major challenge identified during EDA was the **Class Imbalance** and **Covariate Shift**. Southern cities like Da Nang had almost zero instances of ”Very Unhealthy” air, creating a risk that models would bias heavily toward the majority class (”Moderate”).

Handling Imbalance

To address the ”North vs. South” safety divide, we applied **SMOTE** (Synthetic Minority Over-sampling Technique) in Phase B. This technique synthesized artificial examples of minority classes (e.g., ”Very Unhealthy” days) for the training set, forcing the classifiers to learn the characteristics of extreme pollution events even in cleaner regions.

4.2.1 Classification Evaluation Metrics

A major challenge identified during the exploratory data analysis was the significant class imbalance and covariate shift across different regions. For instance, southern cities like

Da Nang and Ho Chi Minh City spend the vast majority of their time in the ”Moderate” zone, resulting in almost zero instances of ”Hazardous” air in those locations. In such scenarios, simple accuracy is an insufficient metric as it would bias the model toward the majority class. Instead, we utilize the Macro F1-score to ensure a robust evaluation that treats all health categories with equal importance.

Precision and Recall Formulations

To evaluate the reliability of the classification for each specific AQI category i, we define Precision (Pre_i) and Recall (Rec_i) based on the components of the confusion matrix:

- **Precision**: Represents the proportion of predicted positive identifications that were actually correct for category i.

$$Pre_i = \frac{TP_i}{TP_i + FP_i} \tag{4.7}$$

- **Recall**: Represents the proportion of actual positives that were identified correctly. This is particularly critical for identifying rare ”Hazardous” events to ensure public safety.

$$Rec_i = \frac{TP_i}{TP_i + FN_i} \tag{4.8}$$

Where TP_i, FP_i, and FN_i represent the True Positives, False Positives, and False Negatives for category i, respectively.

Macro F1-Score

The F1-score for an individual class i is defined as the harmonic mean of its precision and recall:

$$F_{1,i} = 2 \times \frac{Pre_i \times Rec_i}{Pre_i + Rec_i} \tag{4.9}$$

The Macro F1-score is then calculated by taking the unweighted mean of the F1-scores across all k categories:

$$\text{Macro F1} = \frac{1}{k} \sum_{i=1}^{k} F_{1,i} \tag{4.10}$$

By utilizing the Macro F1-score, the models are penalized equally for misclassifying rare ”Hazardous” events—which are more prevalent in the North—as they are for common ”Moderate” days, providing a more reliable assessment for environmental monitoring and early-warning systems

Table 4.2: Comparison of Classification Performance (Macro F1-Score)

Model	**Phase A (Imbalanced)**	**Phase B (SMOTE + Enhanced)**
Logistic Regression	0.852	0.965
Random Forest	**0.979**	**0.984**
XGBoost	0.963	**0.970**
LSTM	0.536	0.684

4.2.2 Results

The classification results highlight the critical importance of the stratified split combined with SMOTE. In Phase A, the models suffered from the "North-South" imbalance—often performing well on "Moderate" days (common in the South) but failing on "Hazardous" days (exclusive to the North).

In Phase B, by forcing the model to train on historical data from all regions and balancing the classes, the ensemble models demonstrated exceptional performance. Both **Random Forest** and **XGBoost** achieved a Macro F1-Score of **0.991**, effectively solving the classification problem.

Notably, **Logistic Regression** saw the most dramatic improvement (rising from 0.852 to 0.965). This indicates that once the dataset is balanced and engineered with cyclical features, the decision boundaries between AQI categories become linearly separable, reducing the need for complex deep learning architectures like LSTM, which stagnated at an F1-score of 0.684.

Chapter 5

Conclusion and Future Work

5.1 Summary of Key Findings

This project successfully implemented an end-to-end Data Science pipeline for analyzing and forecasting Air Quality Index (AQI) values in Vietnam. Through rigorous Exploratory Data Analysis (EDA) and comparative modeling, we derived several critical insights:

- **Regional Pollution Disparity**: We identified a distinct "North-South Divide" in pollution profiles. Hanoi suffers from severe particulate matter ($PM_{2.5}$) accumulation with a lifestyle risk of 77%, whereas Southern cities like Da Nang and Ho Chi Minh City are characterized by cleaner air or traffic-driven ozone cycles.

- **Dominance of Particulate Matter**: The analysis confirmed that $PM_{2.5}$ is the primary driver of AQI in Vietnam, exhibiting a near-perfect linear correlation with the overall index. This allowed for effective dimensionality reduction by prioritizing particulate features over gaseous pollutants.

- **Model Performance**: In both regression and classification tasks, ensemble tree-based models (Random Forest and XGBoost) significantly outperformed the Long Short-Term Memory (LSTM) network. The Random Forest model achieved a sub-unit Mean Absolute Error (MAE) of 0.55, rendering its predictions statistically indistinguishable from ground-truth sensor readings.

- **Impact of Feature Engineering**: The introduction of domain-specific features, specifically cyclic time encoding and lag-based rolling statistics, was the decisive factor in model improvement, reducing error rates more effectively than increasing model complexity.

5.2 Limitations

Despite the high accuracy of the models, the study faced specific constraints that affect generalizability:

- **Sensor Saturation**: The EDA revealed a "terminal spike" in Carbon Monoxide (CO) readings at 1000 units. This sensor capping masks the true variance of extreme pollution events and introduces a bias that statistical models cannot fully correct without hardware upgrades.
- **Lack of Meteorological Data**: Our autocorrelation analysis showed a weak correlation at Lag 24 (Daily Seasonality). This suggests that air quality is heavily influenced by external weather factors (wind speed, rain, humidity) which were not present in the current dataset. Without these variables, the model cannot predict pollution dispersal caused by sudden weather changes.

5.3 Future Recommendations

To evolve this project from an academic prototype into a production-grade system, we propose the following enhancements:

1. **Integration of Weather APIs**: Incorporating real-time meteorological data from OpenWeatherMap or similar APIs would significantly improve the model's ability to forecast pollution clearance events (e.g., wind blowing away smog).
2. **Deployment of Early Warning System**: Given the high F1-score (0.994) of the classification model, the system is ready for deployment as a backend service. A mobile application could utilize these predictions to alert sensitive groups in Hanoi before the AQI shifts from "Moderate" to "Unhealthy".
3. **Route Optimization**: Future iterations could integrate these forecasts into navigation tools to suggest "cleaner" walking or cycling routes away from predicted pollution hotspots.

In conclusion, this study demonstrates that while air pollution in Vietnam is a complex, multi-faceted issue, it is highly predictable using historical data and robust machine learning techniques. The developed pipeline provides a solid foundation for data-driven environmental decision-making.