Big Data Analysis Project- Assignment 1
Part B

Impact of Climate Variability
on
Urban Air Pollution:
A Big Data Analysis of Sydney (2015–2025)

Utsav Punia : a1956304

The University of Adelaide

4533_COMP_SCI_7209 : Big Data Project

Table of Contents

1. Main Research Question:

"How has climate variability influenced air pollution levels in Sydney over the past decade?"

Supporting Questions(RSQs: Research Supporting Questions):

- **What are the key meteorological factors** (such as temperature, humidity, and wind) that contribute to variations in the concentrations of pollutants like PM2.5 and NO_2 in Sydney?
- **Are there identifiable trends in urban air quality that align with recurring climate patterns**, such as seasonal cycles or extreme weather events?
- **How do urbanization and climatic variability combining influence** average pollution levels across different regions of Sydney?
- **Can predictive models be developed using historical meteorological and air quality data to anticipate pollution spikes?**

2. Dataset Description

2.1. Overview of Datasets

This project integrates two primary datasets obtained through API::

- **Air Quality Monitoring Data**, which includes hourly pollutant concentration levels recorded at monitoring stations across New South Wales (NSW), including the Sydney region.
- **Meteorological Data**, which contains hourly environmental variables such as temperature, humidity, wind speed, wind direction, and sigma theta.

Together, these datasets provide a comprehensive basis for analyzing the relationship between climate variability and pollution trends over time.

2.2. Data Source and Structure

The datasets were accessed using the **NSW Government's Air Quality API**, which provides structured environmental observations from its ambient monitoring network. A custom POST request was designed in Python to collect data across selected parameters and monitoring stations. The retrieved data, initially in JSON format, was then normalized and exported to CSV for further analysis.

Source and Documentation: The datasets were accessed using the NSW Government's Air Quality API (NSW Government, 2024a), which provides structured environmental observations. Full documentation is available in the API User Guide (NSW Government, 2024b).

Key characteristics of the retrieved data:

- **Data Format:** Data retrieved in JSON format Converted to CSV.
- **Temporal coverage:** Hourly data from June 1 ,2015 to June 1,2025.

- **Spatial coverage**: Data obtained from five monitoring sites distributed across urban regions of Sydney, covering eastern, western, and northwestern zones.
- **Fields included:**

Table A: Data Fields

Field Name	**Description**
Site_Id	Unique identifier for the monitoring site
Date	Date of observation
Hour	Hour of observation (1 to 24)
HourDescription	Time range for the hour (e.g., "12 am - 1 am")
Value	Measured value of the parameter
AirQualityCategory	Category assigned based on pollutant thresholds
DeterminingPollutant	Pollutant used to determine air quality category
Parameter.ParameterCode	Code representing the measured parameter (e.g., PM2.5
Parameter.ParameterDescription	Full name of the parameter
Parameter.Units	Abbreviation of the unit of measurement
Parameter.UnitsDescription	Full description of the unit of measurement
Parameter.Category	Type of parameter (e.g., Averages)
Parameter.SubCategory	Subtype within category (e.g., Hourly)
Parameter.SubCategory	Frequency of measurement (e.g., Hourly average)

- **Selected Parameters**

  The variables extracted for analysis include both pollutants and weather-related measurements which were collected under Parameter.ParameterCode :

Table B: Air Pollutants Monitored in the Study

Parameter	**Description**
PM10	Particulate Matter ≤10 micrometres in diameter, affects respiratory health
PM2.5	Fine Particulate Matter ≤2.5 micrometres, penetrates deeper into lungs and bloodstream
NO2	Nitrogen Dioxide, a traffic-related pollutant harmful to lungs
CO	Carbon Monoxide, a gas that reduces oxygen delivery in the body
OZONE	Ground-level Ozone, formed by chemical reactions involving NO_2 and sunlight; irritates airways

Table C: Meteorological Variables Used in the Analysis

Parameter	**Description**
TEMP	Air temperature (°C)
HUMID	Relative humidity (%)
WSP	Wind speed (m/s)
WDR	Wind speed (m/s)
SD1	Sigma Theta – Standard deviation of wind direction (°)

- **Monitoring Sites Selected**

  The following five monitoring sites in Sydney were selected for this study:

  Table D: Selected Monitoring Sites in Sydney

Site Name	Site ID	Region
Rozelle	39	Sydney East
Chullora	222	Sydney East
Parramatta North	919	Sydney North-west
Campbelltown West	2560	Sydney South-west
Liverpool	107	Sydney South-west

  These stations were chosen because they span across geographically diverse urban regions of Sydney including eastern, western, and southwestern corridors offering a balanced spatial distribution. More importantly, these locations report a wide range of both meteorological and pollutant variables (including PM2.5, NO_2, CO, and wind metrics) consistently and reliably, making them ideal for comprehensive temporal and spatial analysis. Site metadata and data availability were confirmed via the **NSW Air Quality API** and its official documentation (NSW DCCEEW, 2023).

2.3. Data Cleaning and Preprocessing

The initial cleaning and transformation of the dataset were done during Assignment 1A using a sample file. In this part, the full dataset (2015–2025) was retrieved and similar steps were repeated to maintain consistency across all sites and parameters.

Key preprocessing steps included:

- **Datetime Parsing:** Converted date and hour columns into a proper Timestamp field using ParsedDate and HourDescription, ensuring hourly granularity.
- **Pivoting:** Transformed the dataset into a long format indexed by Site_Id and Timestamp, with each parameter (like PM2.5, NO_2, TEMP, etc.) as columns.
- **Forward-Fill Imputation:** Applied forward-fill (ffill) within each site group to handle missing values over time in a sensible way for time-series data.
- **Variable Filtering:** Kept only relevant pollutant and meteorological features (PM2.5, PM10, NO_2, CO, OZONE, TEMP, HUMID, WSP, WDR, SD1).
- **Final Dataset:** The cleaned and aligned dataset was saved as full_cleaned.csv for all further visualisation and modelling.

3. Clustering and Pattern Detection

This section looks at patterns in air pollution using clustering, correlation, and regression techniques. The aim is to understand how weather conditions affect pollution levels and whether we can find groups of similar environmental conditions. All findings are related back to the research questions.

3.1. Correlation Analysis

To explore how meteorological variables affect pollution levels, we calculated the Pearson correlation matrix between pollutants (PM2.5, PM10, NO_2, CO, OZONE) and weather-related features (TEMP, HUMID, WSP, WDR, SD1).

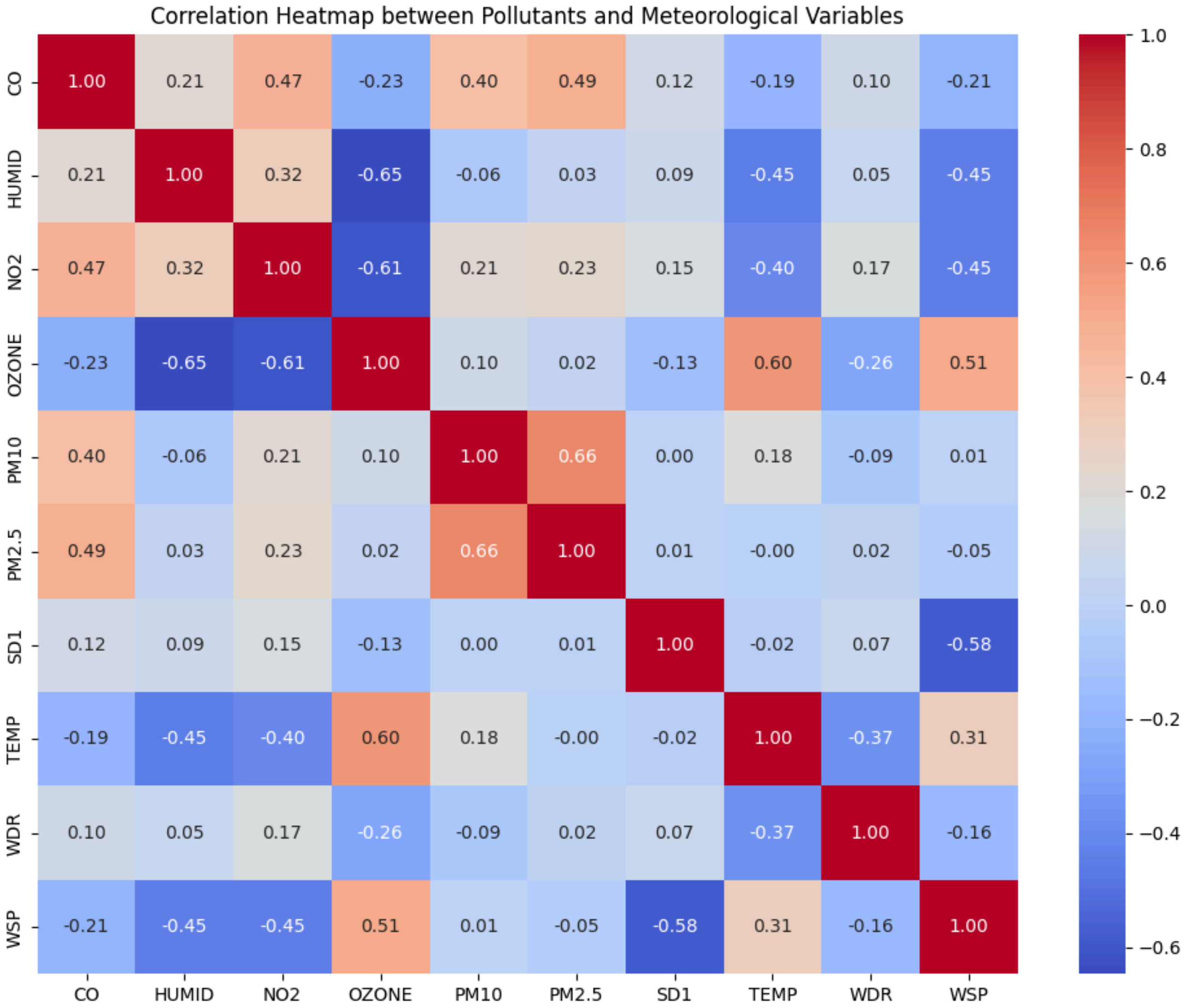


Fig. 1 : Correlation Heatmap between Pollutants and Meteorological Variables

Correlation Heatmap Analysis:

- **OZONE** shows a **strong positive correlation with temperature** (0.60) and **moderate positive correlation with wind speed** (0.51). It also shows a **strong negative correlation with humidity (-0.65)** and NO_2 (-0.61), suggesting that ozone levels rise on hot, dry, and sunny days when NO_2 is low.

- **PM2.5** and **PM10** both have a **moderate positive correlation with each other** (0.66), and **PM2.5** shows a moderate positive correlation with **CO** (0.49), pointing to shared sources like combustion or traffic. Both PM pollutants have **weak or negligible correlation with weather variables**, though PM10 is slightly negatively correlated with wind speed.
- **NO_2** is **positively related to CO (0.47)** and humidity (0.32), which suggests a common source such as vehicle emissions. It has a **negative relationship with temperature (-0.40)**, possibly due to reduced traffic or better dispersion on warmer days.
- **CO** is weakly related to most meteorological variables, but has a **moderate positive correlation with PM2.5 and PM10**, supporting the idea of local emission sources.

These correlations support **RSQ1**, which investigates how meteorological conditions influence pollution levels. For example, we observe that **windy and humid conditions reduce ozone and particulates**, while **temperature increases ozone** but decreases NO_2.

3.2. Clustering with KMeans and t-SNE

To find groups of similar air quality conditions, KMeans clustering algorithm was used on the standardised pollution and weather data. Different values of k were experimented and k = 3 was chosen for clustering.

Cluster count k = 3 was selected after testing multiple configurations, including k = 4 (see Appendix, Fig. 1). The 3-cluster setting gave clearer visual separation in the t-SNE projection and more interpretable groupings based on environmental conditions. In contrast, the fourth cluster in k = 4 led to overlapping and less distinct groups, indicating over-segmentation.

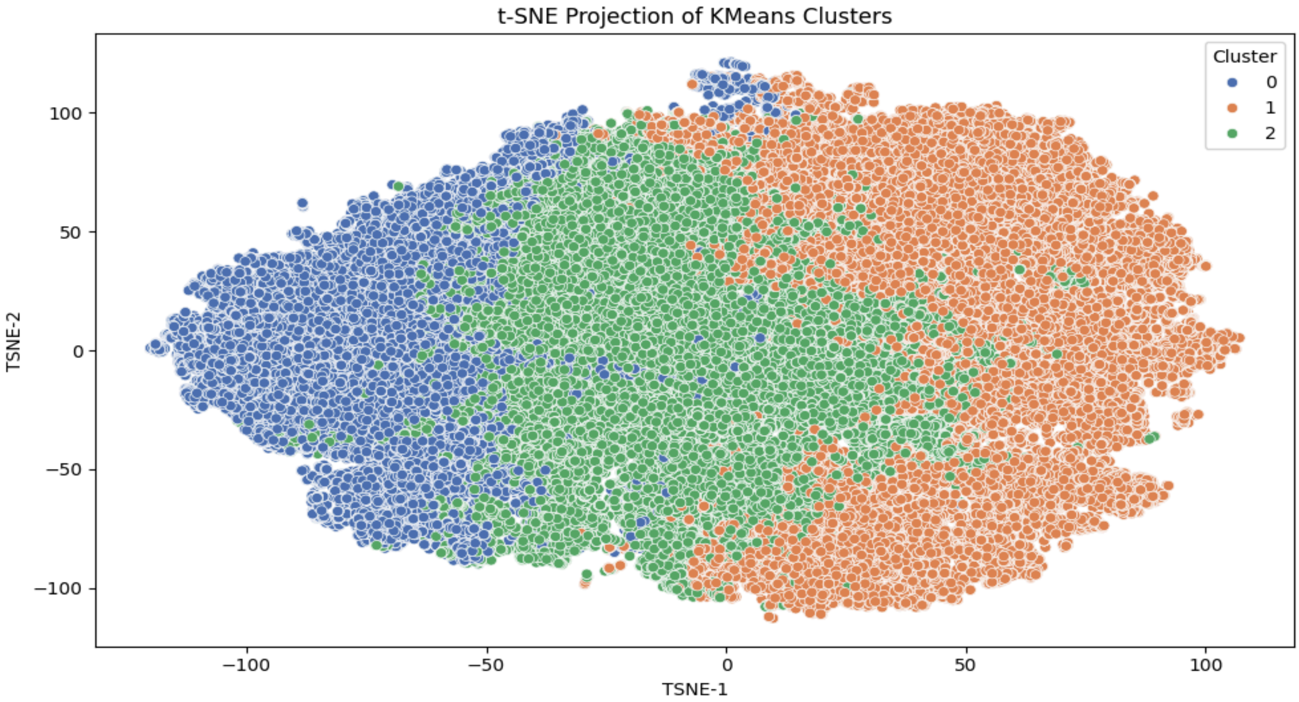


Fig. 2 : T-SNE Projection of KMeans Clusters

To visualise the clusters, we applied **t-SNE** to reduce the high-dimensional data to 2D. Each point represents one hourly observation, and nearby points have similar weather and pollution conditions. The colours show which cluster they belong to:

- **Cluster 0 (Blue):** This group is characterised by higher wind speeds and lower levels of PM2.5 and PM10. These conditions help disperse pollutants more effectively, leading to cleaner air overall. The cleaner conditions in this cluster are likely due to stronger air movement.
- **Cluster 1 (Orange):** This cluster shows higher PM2.5 levels, along with higher humidity and lower wind speeds. The combination of humid and still air can trap pollutants near the ground, causing more buildup of particulate matter.
- **Cluster 2 (Green):** This cluster is characterised by high temperature and ozone levels, and low humidity. These hot and dry conditions favour ozone formation through sunlight. Even though it's hot, PM levels are not very high, possibly due to better vertical air mixing or fewer emissions at those times.

3.3. Regression: Predicting PM2.5 from Weather

To understand how weather affects pollution levels, separate linear regression models were trained for each of the major pollutants :PM2.5, PM10, NO_2, CO, and OZONE,using meteorological inputs such as temperature, humidity, wind speed, wind direction, and sigma theta (SD1).

Each model's performance was evaluated using **R^2 Score** and **RMSE** (Root Mean Squared Error).These metrics help us assess how well weather conditions can explain variation in pollutant levels.

Table D: Regression Results

Pollutant	**R^2 Score**	**RMSE**
PM2.5	0.004	10.923
PM10	0.037	15.118
NO_2	0.282	0.655
CO	0.061	0.219
OZONE	0.582	0.803

These results show clear differences in how well each pollutant can be predicted:

- **OZONE** had the highest R^2 score (0.582), indicating that it is strongly influenced by weather, especially temperature and sunlight.
- **NO_2** also had moderate predictability ($R^2 = 0.282$), likely due to its links to wind patterns and stability conditions.
- **CO** and **PM2.5** had very low R^2 scores (0.061 and 0.004 respectively), suggesting they are influenced more by local emissions and non-weather factors.
- **PM10** had the weakest overall fit ($R^2 = 0.037$), indicating high variability and complex sources.

3.4. Visualising Model Performance

To check how well the regression models worked, we plotted the **actual pollutant values** (x-axis) against the **predicted values** (y-axis) for each pollutant. A red dashed line shows where perfect predictions would fall. The closer the points are to this line, the better the model is at predicting.

Key observations from the plots:

- **OZONE**: The points are fairly close to the line, matching the relatively high R^2 score (0.582). The model captured the general trend well.

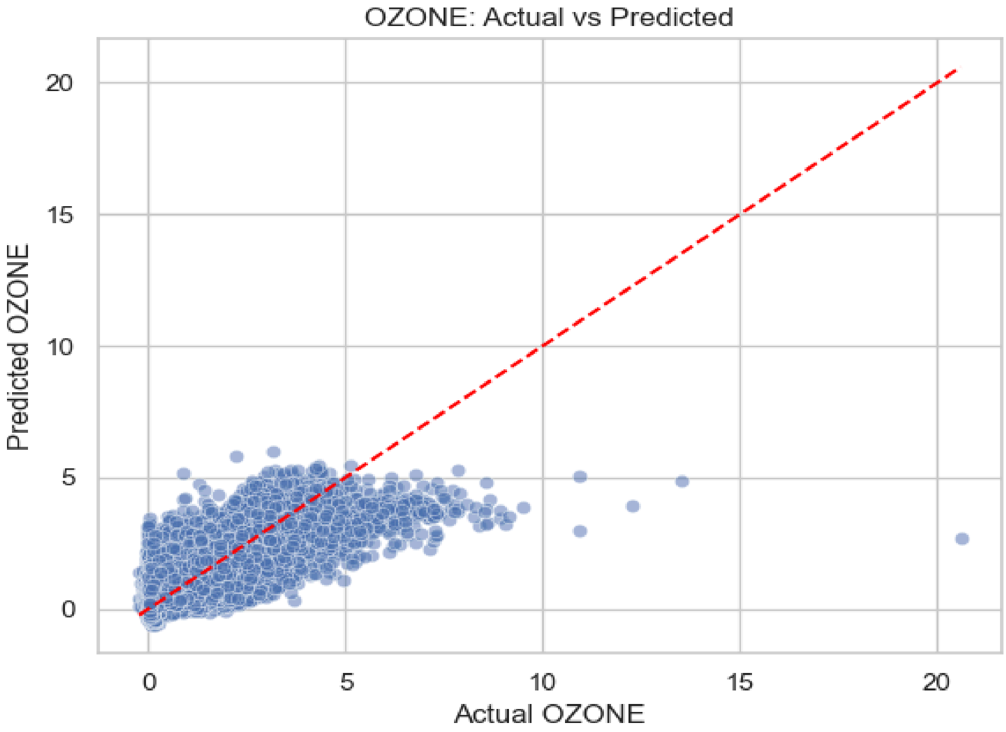


Fig. 3: Actual vs. Predicted OZONE

- **NO_2**: Predictions also follow the correct pattern, though slightly more spread out. The model still showed reasonable accuracy.

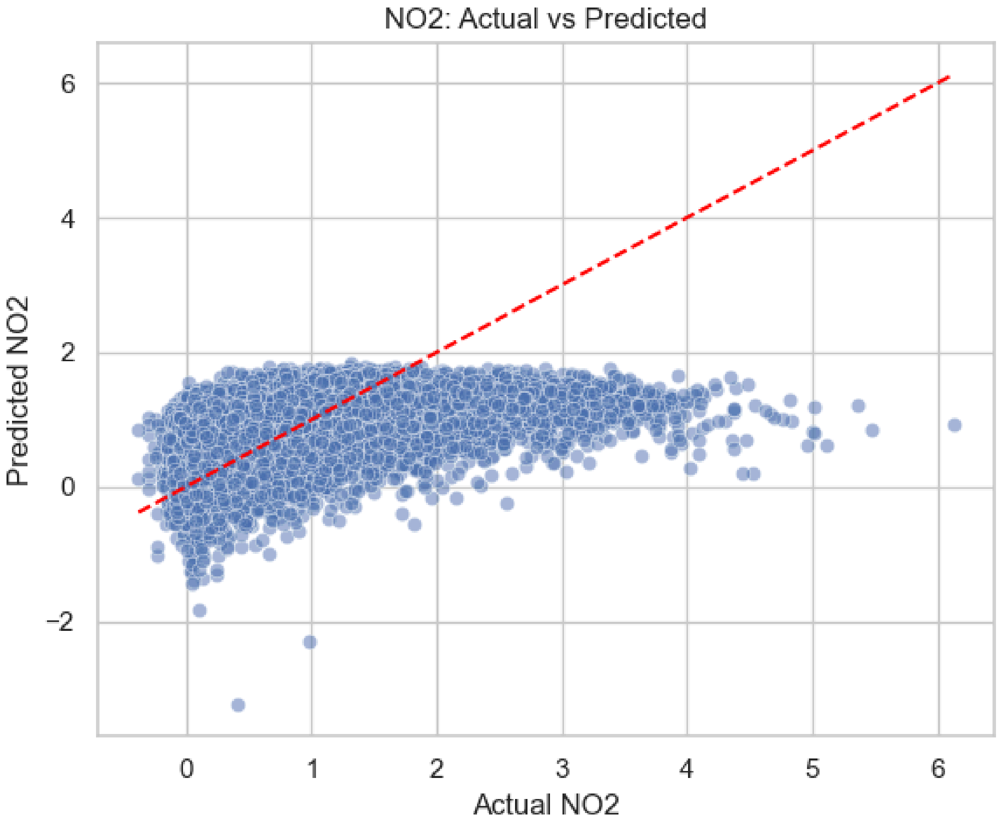


Fig. 4: Actual vs. Predicted NO_2

- **PM2.5 and PM10**: These plots show low and scattered predictions. The models were not able to track the higher observed values, resulting in poor fit.

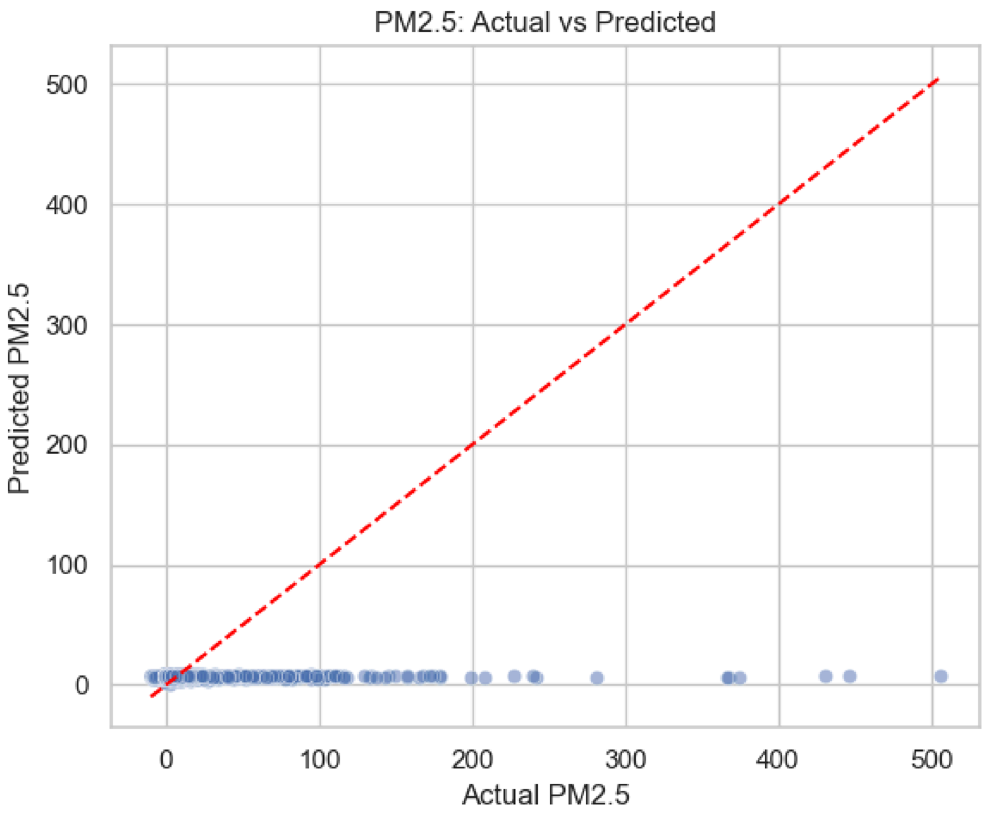


Fig 5: Actual vs. Predicted PM2.5

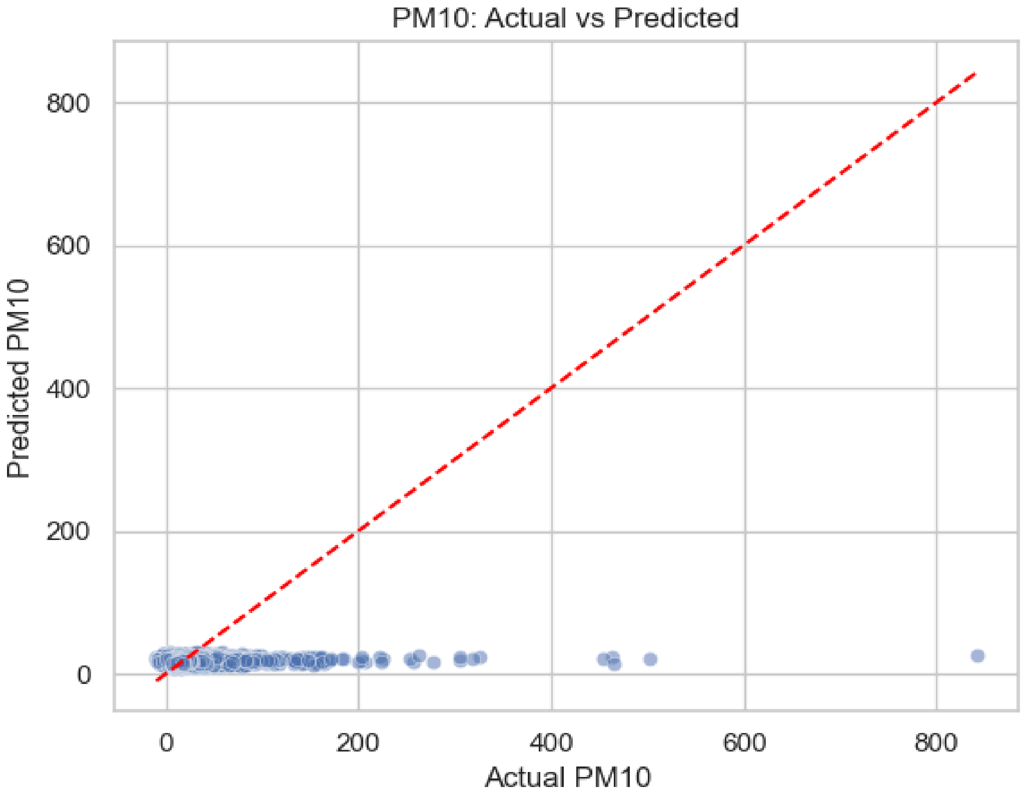


Fig. 6: Actual vs. Predicted PM10

- **CO**: Predictions were low overall and failed to capture higher actual values, indicating underestimation at upper ranges.

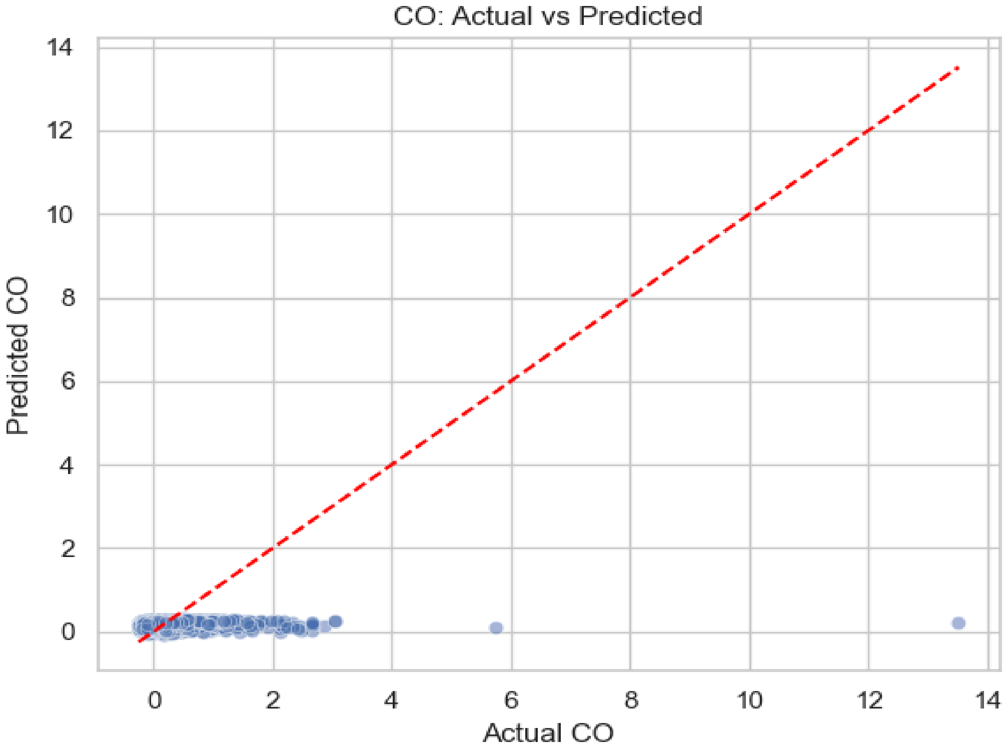


Figure 7: Actual vs. Predicted CO

These visual comparisons confirm the quantitative results which indicated that OZONE and NO₂ are better predicted from weather features, while PM2.5, PM10, and CO are more difficult to model, possibly due to local sources or sudden events not captured in the dataset.

Note: *The limited performance of the linear regression model,particularly for PM2.5 and PM10 indicates that more complex, non-linear relationships may exist in the dataset. In future stages of the project, advanced models such as **Random Forests**, **Gradient Boosting**, or **Neural Networks** could be explored to better capture these complex relationships with high accuracy.*

4. Feature Relationships and Addressing RSQs

This section brings together key patterns observed in the analysis and explains how they relate to our main research questions (RSQs).

4.1. Correlation Between Weather and Pollution

In the earlier correlation heatmap (Section 3.1), some clear patterns were observed:

- **OZONE** shows a strong positive correlation with temperature and a strong negative correlation with humidity and NO_2, suggesting that **high ozone levels are associated with warm, dry, and sunny conditions.**
- **PM2.5** and **PM10** both show moderate negative correlation with wind speed, indicating that particulate matter tends to build up when the air is calm.
- **NO_2** and **CO** are positively correlated with each other and also with particulate matter, likely due to common sources like vehicle emissions.

These correlations support **RSQ1** by showing how weather affects pollution, and also help with **RSQ2**, which asks whether weather features can be used to predict pollution levels.

4.2. Clustering Patterns

The clustering results (Section 3.2) demonstrate that pollution patterns are not random but emerge from distinct combinations of weather conditions. These groupings reflect meaningful environmental states, confirming that certain meteorological profiles tend to produce specific pollution behaviours. This directly supports **RSQ3**, whether air quality can be grouped based on shared weather-pollution characteristics.

4.3. Regression Insights

Regression results (Section 3.3) show that pollutants differ in how well they respond to weather-based prediction. OZONE and NO_2 had stronger predictive performance, while PM2.5, PM10, and CO were less responsive, likely due to additional local factors. This provides clear evidence for **RSQ2**, confirming that meteorological variables can explain part of the variation in pollution but not equally across all pollutants.

4.4. Seasonal and Temporal Trends

Time-series plots (Section 3.4) showed clear seasonal patterns. OZONE levels were higher during summer months, likely due to stronger sunlight and higher temperatures that promote its formation. In contrast, PM2.5 levels were higher in winter, possibly due to stagnant air and emissions from heating. These trends help answer **RSQ4** by confirming that pollution varies with seasonal and temporal changes in weather.

5. Data Visualization

To highlight major trends in pollution and weather, four focused visualisations were created. These were selected based on their ability to clearly show patterns relevant to the research questions (**RSQs**) and to reinforce findings from earlier sections, such as clustering, regression, and correlation analysis.

- **Average PM2.5 Over Time:** This time-series plot shows how PM2.5 levels change throughout the year. Higher values appear during colder months, likely due to calm weather and heating activity. A sharp spike during late 2019 to early 2020 aligns with known bushfire events (NSW BUSH FIRE HISTORY, 2025).This pattern supports **RSQ1** (seasonal variation) and is consistent with results from the **correlation heatmap** and **temporal trends** sections.

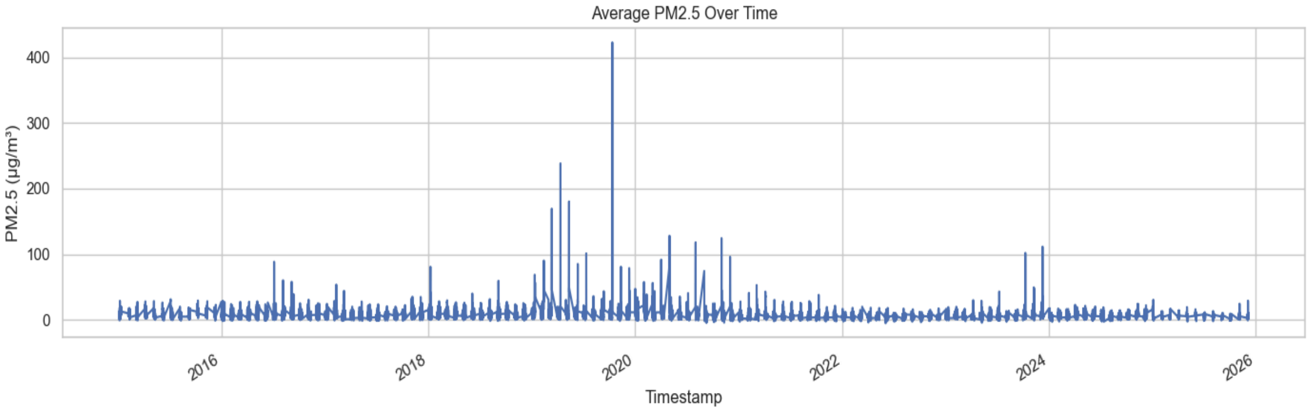


Fig 8: Average PM2.5 Over Time

- **Average OZONE Over Time:** Ground-level ozone tends to increase during warmer months and decrease during cooler ones. This behaviour reflects its photochemical formation, which depends on sunlight and temperature.The trend aligns with RSQ3 and matches the regression results, where OZONE was found to be the most predictable pollutant from weather features.

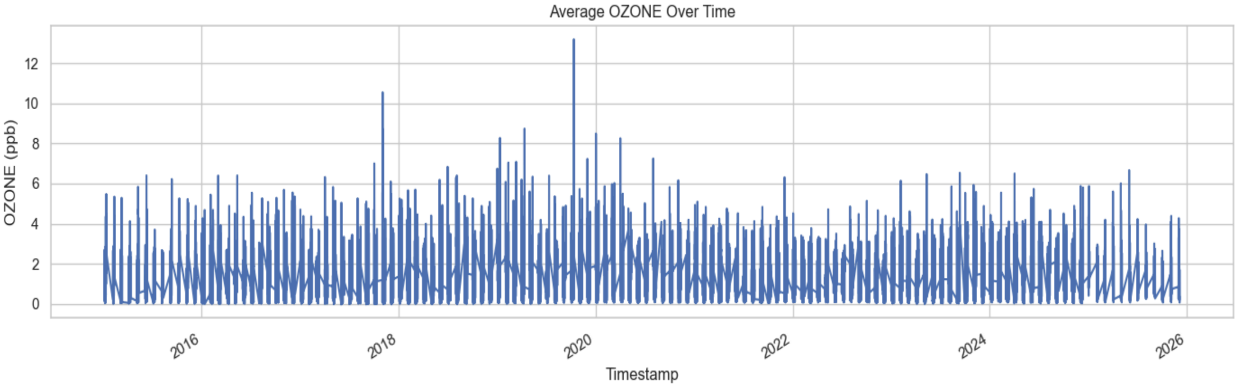


Fig. 9: Average Ozone Over Time

- **PM10 Distribution by Site:** The boxplot in Figure 10 shows how PM10 levels vary across different monitoring sites in Sydney. Notably, Liverpool (Site 107) and Parramatta North (Site 919) show higher PM10 readings and more frequent extreme values (outliers), suggesting greater particulate pollution. These areas may experience more traffic or nearby industrial activity, which can lead to increased dust and emissions. In contrast, Rozelle (Site 39) and Campbelltown West (Site 2560) show lower and more stable PM10 levels, possibly reflecting their more residential or less congested environments.These differences support **RSQ2**, which investigates how pollution levels vary across locations. They also align with earlier clustering results, where site-level characteristics contributed to the formation of distinct environmental clusters.

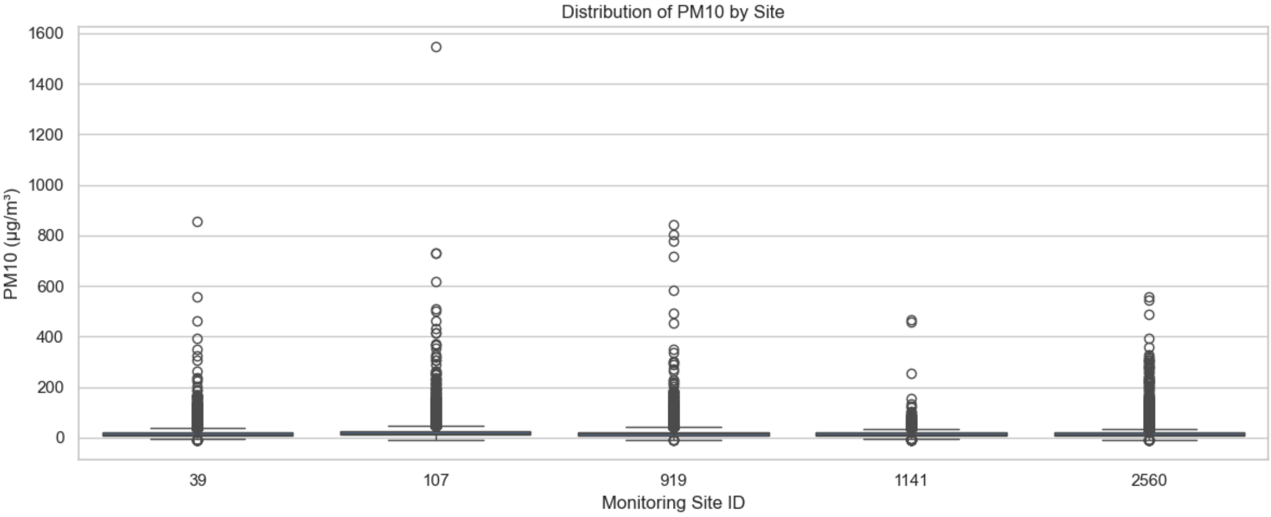
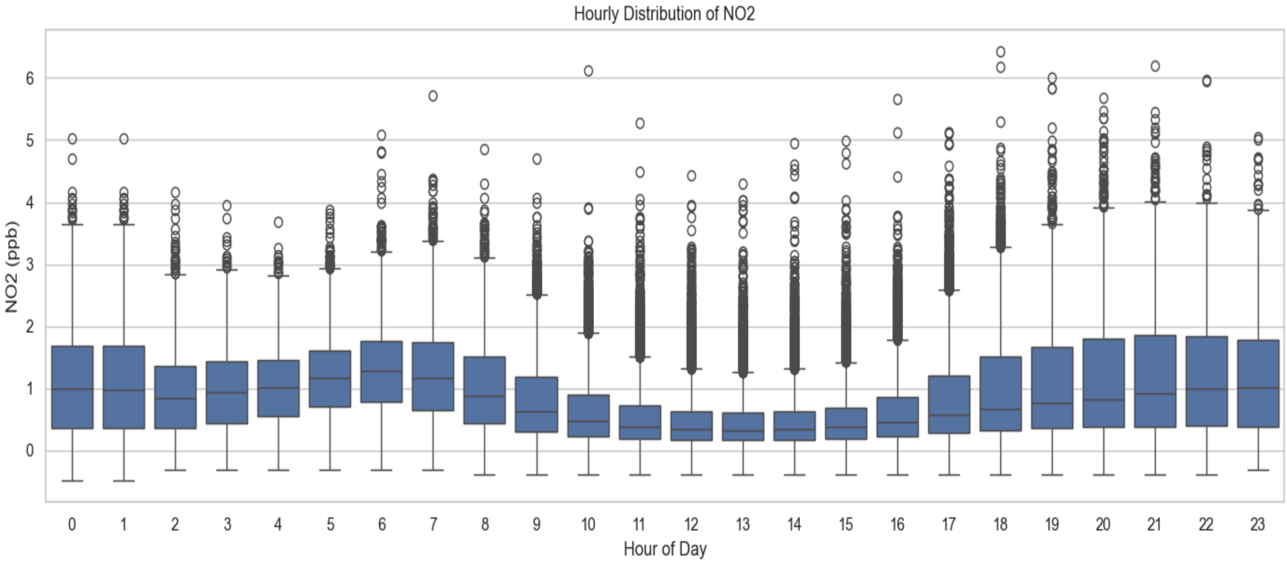


Fig. 9: Average Ozone Over Time

- **NO₂ Distribution by Hour of Day:** This hourly boxplot shows how **NO₂ (Nitrogen Dioxide)** levels vary throughout the day across all monitoring sites. The plot reveals two clear peaks, Morning peak (around 7–9 AM) and Evening peak (around 6–9 PM). These spikes likely reflect rush hour traffic, where vehicle emissions are at their highest. Between 10 AM and 4 PM, NO₂ levels drop significantly, possibly due to reduced traffic and increased atmospheric dispersion under sunlight.

  This pattern strongly supports **RSQ4**, which focuses on daily pollution cycles and their links to human activities. It also aligns with known environmental behaviour, where NO₂ commonly emitted from vehicles is typically higher during busy commuting hours.



6. Problem Refinement

The initial **RQs** focused on understanding how meteorological features influence pollution and whether these relationships can be used to cluster or predict air quality outcomes. Based on the results, more refined set of sub-question is developed:

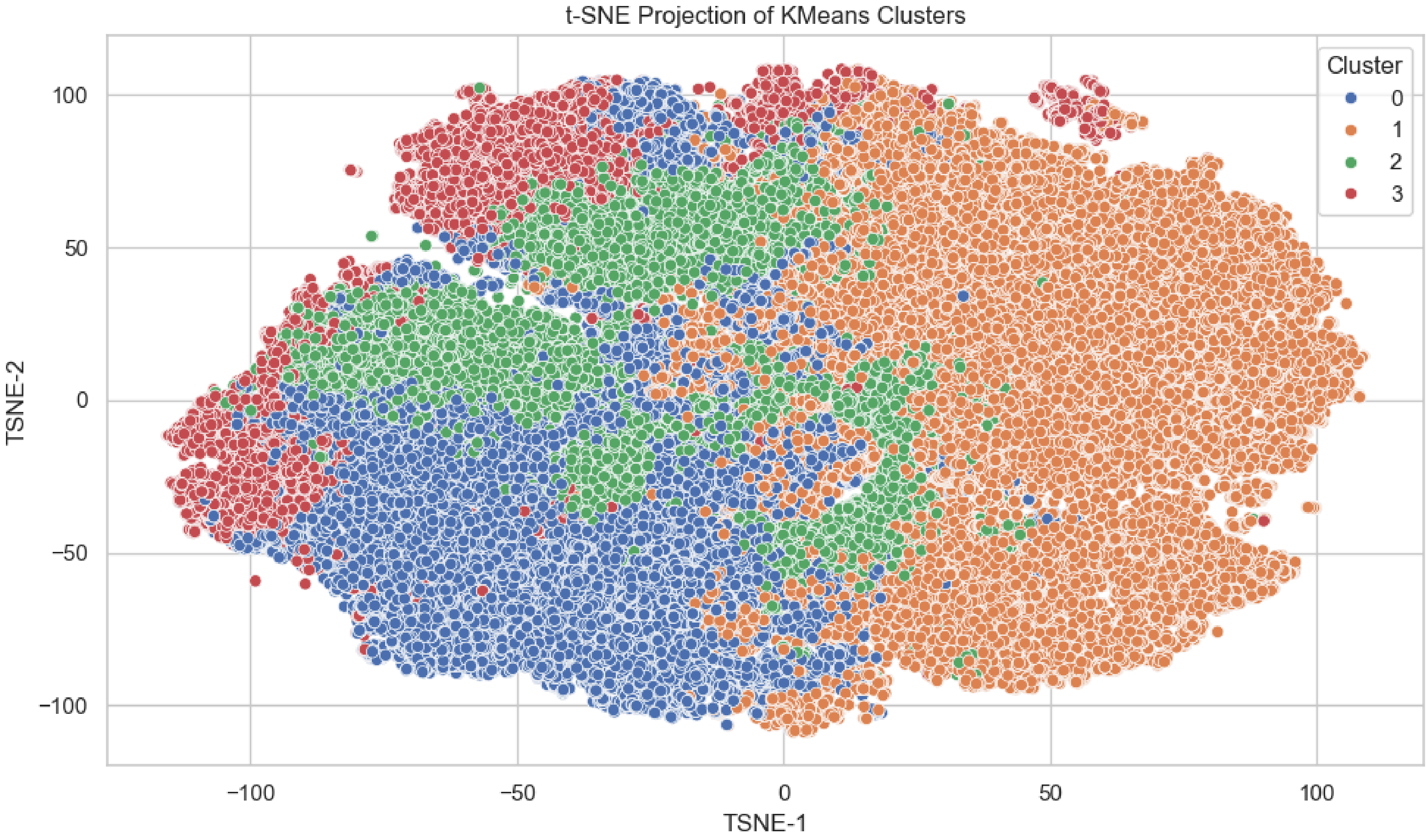
- Can **extreme pollution events** (e.g., high PM2.5 spikes) be attributed to **non-meteorological events**, such as bushfires or localized emissions?
- Are there **distinct temporal signatures** (e.g., weekday vs weekend patterns) that can further improve prediction models for pollutants like NO_2?
- Can we develop a **site-specific early warning system** using site-wise thresholds learned from the cluster analysis and hourly trends?
- To what extent can historical meteorological data be used to train pollutant-specific models for reliably forecasting pollution spikes, and how do their performance limitations differ across pollutants such as PM2.5, NO_2, and OZONE?

These refined sub-questions emerge from the observed limitations in regression prediction performance (especially for PM2.5 and CO) and the distinct seasonal/temporal and cluster-based patterns revealed in the data.

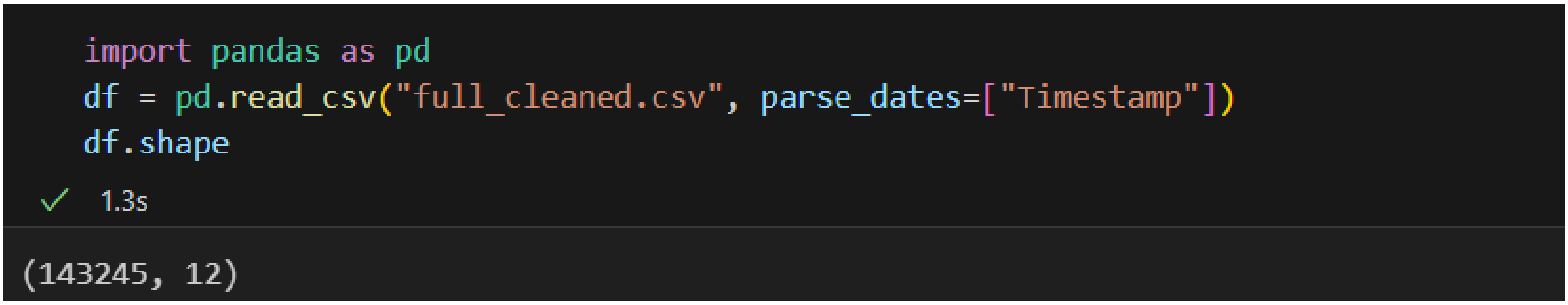
7. References

1. Bureau of Meteorology (2024a). *Climate trends and extreme weather in Australia*. Available at: https://www.bom.gov.au/climate/change/ [Accessed 12 June 2025].
2. Bureau **of Meteorology (2024b).** *Climate history.* Available at: http://www.bom.gov.au/climate/history/ [Accessed 12 June 2025].
3. NSW Department of Climate Change, Energy, the Environment and Water (2023). *Air Quality Monitoring Network Overview*. Available at: https://www.environment.nsw.gov.au/ [Accessed 12 June 2025].
4. NSW Government (2024a). *Air Quality API.* Available at: https://www.airquality.nsw.gov.au/air-quality-data-services/air-quality-api [Accessed 12 June 2025].
5. NSW Government (2024b). *Air Quality API – Application Programming Interface User Guide*. Available at: https://www.environment.nsw.gov.au/sites/default/files/air-quality-application-programming-interface-user-guide-210346.pdf [Accessed 12 June 2025].
6. OECD(2024). *Air pollution*. Available at: https://www.oecd.org/en/topics/sub-issues/air-pollution.html [Accessed 12 June 2025].
7. NSW BUSH FIRE HISTORY, 2025 (2025). *NSW Bushfire History Viewer*. Available at:https://unsw-au.maps.arcgis.com/apps/MapSeries/index.html?appid=80816503d949422ca5725507c714b429 [Accessed 4 July 2025].

8. Appendix



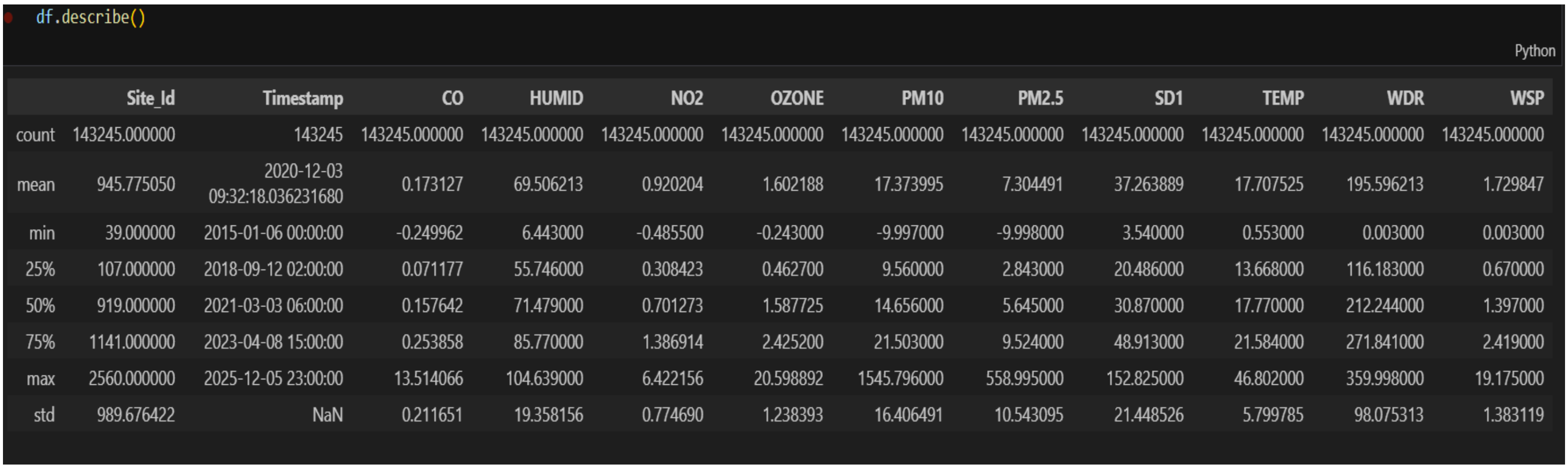
Appendix Fig: 1 Clustering with K=4



```
import pandas as pd
df = pd.read_csv("full_cleaned.csv", parse_dates=["Timestamp"])
df.shape
```

```
(143245, 12)
```

Appendix Fig: 2 Shape of the Collected Dataset

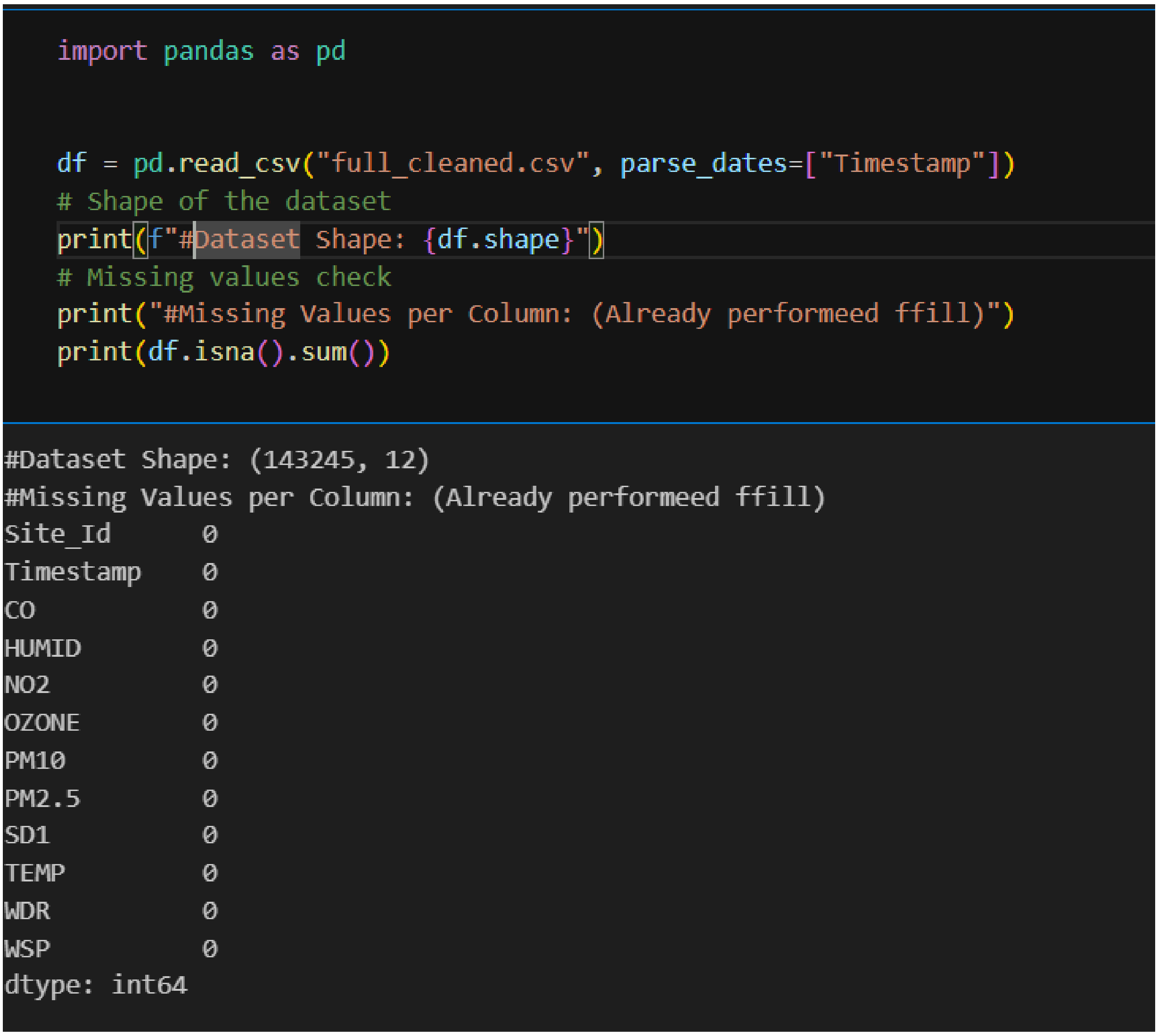


```
df.describe()
```

Python

	Site_Id	Timestamp	CO	HUMID	NO2	OZONE	PM10	PM2.5	SD1	TEMP	WDR	WSP
count	143245.000000	143245	143245.000000	143245.000000	143245.000000	143245.000000	143245.000000	143245.000000	143245.000000	143245.000000	143245.000000	143245.000000
mean	945.775050	2020-12-03 09:32:18.036231680	0.173127	69.506213	0.920204	1.602188	17.373995	7.304491	37.263889	17.707525	195.596213	1.729847
min	39.000000	2015-01-06 00:00:00	-0.249962	6.443000	-0.485500	-0.243000	-9.997000	-9.998000	3.540000	0.553000	0.003000	0.003000
25%	107.000000	2018-09-12 02:00:00	0.071177	55.746000	0.308423	0.462700	9.560000	2.843000	20.486000	13.668000	116.183000	0.670000
50%	919.000000	2021-03-03 06:00:00	0.157642	71.479000	0.701273	1.587725	14.656000	5.645000	30.870000	17.770000	212.244000	1.397000
75%	1141.000000	2023-04-08 15:00:00	0.253858	85.770000	1.386914	2.425200	21.503000	9.524000	48.913000	21.584000	271.841000	2.419000
max	2560.000000	2025-12-05 23:00:00	13.514066	104.639000	6.422156	20.598892	1545.796000	558.995000	152.825000	46.802000	359.998000	19.175000
std	989.676422	NaN	0.211651	19.358156	0.774690	1.238393	16.406491	10.543095	21.448526	5.799785	98.075313	1.383119

Appendix Fig: 3 Dataset Information

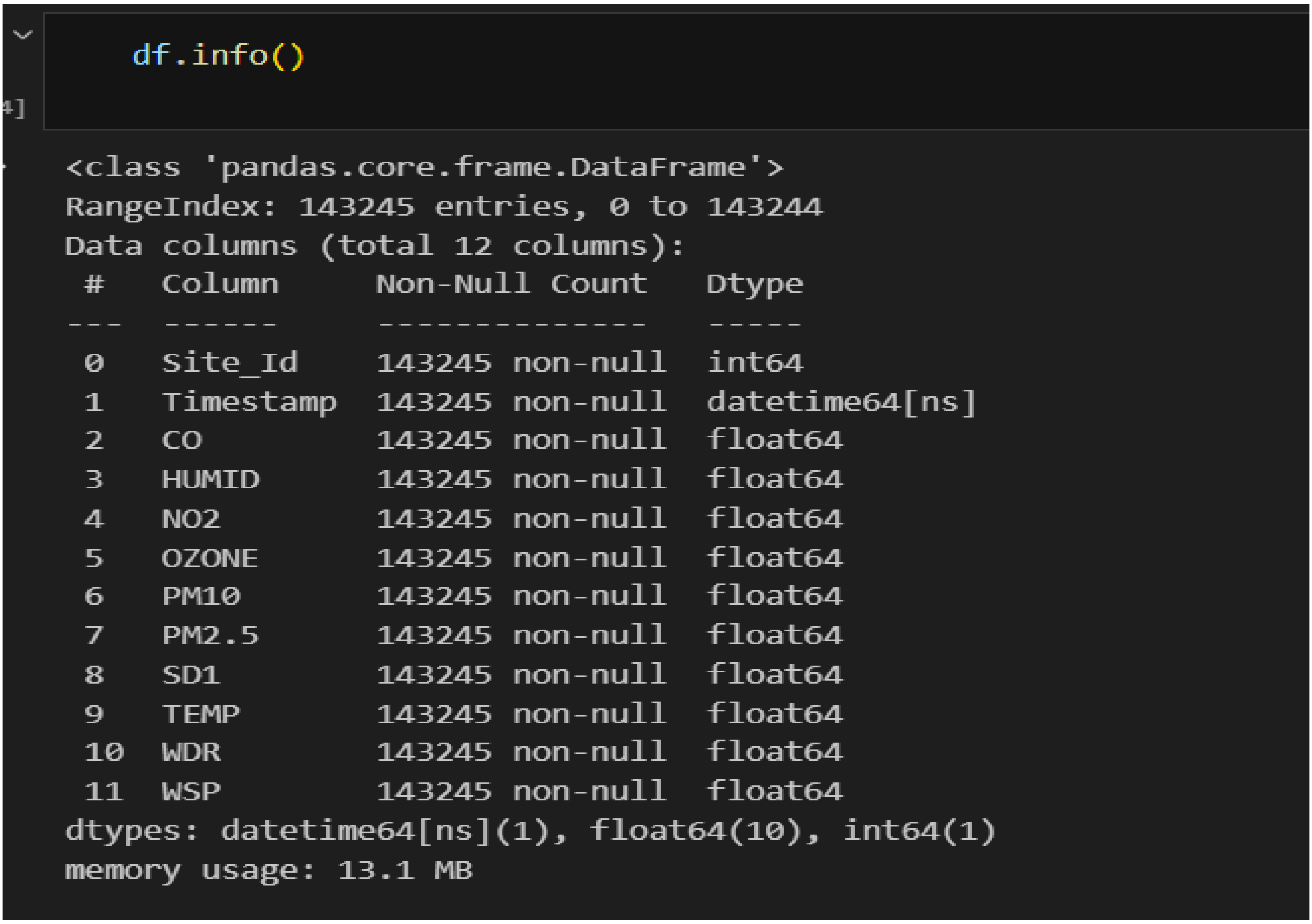


```
import pandas as pd


df = pd.read_csv("full_cleaned.csv", parse_dates=["Timestamp"])
# Shape of the dataset
print(f"#Dataset Shape: {df.shape}")
# Missing values check
print("#Missing Values per Column: (Already performeed ffill)")
print(df.isna().sum())
```

```
#Dataset Shape: (143245, 12)
#Missing Values per Column: (Already performeed ffill)
Site_Id      0
Timestamp    0
CO           0
HUMID        0
NO2          0
OZONE        0
PM10         0
PM2.5        0
SD1          0
TEMP         0
WDR          0
WSP          0
dtype: int64
```

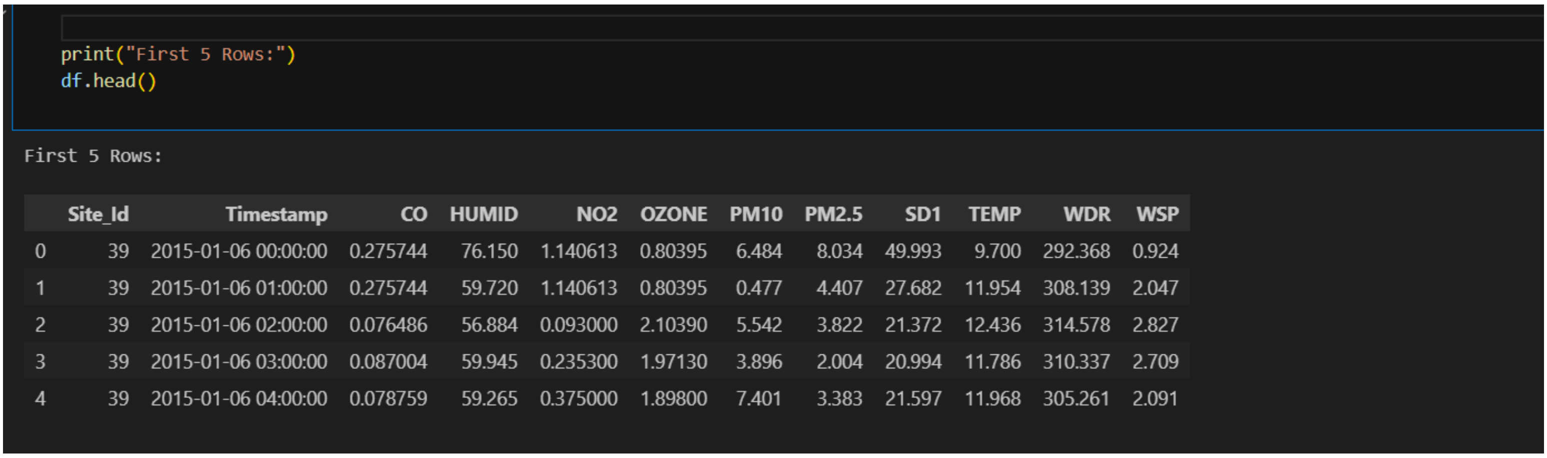
Appendix Fig: 4 Dataset Summary



```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 143245 entries, 0 to 143244
Data columns (total 12 columns):
 #   Column     Non-Null Count   Dtype
---  ------     --------------   -----
 0   Site_Id    143245 non-null  int64
 1   Timestamp  143245 non-null  datetime64[ns]
 2   CO         143245 non-null  float64
 3   HUMID      143245 non-null  float64
 4   NO2        143245 non-null  float64
 5   OZONE      143245 non-null  float64
 6   PM10       143245 non-null  float64
 7   PM2.5      143245 non-null  float64
 8   SD1        143245 non-null  float64
 9   TEMP       143245 non-null  float64
 10  WDR        143245 non-null  float64
 11  WSP        143245 non-null  float64
dtypes: datetime64[ns](1), float64(10), int64(1)
memory usage: 13.1 MB
```

Appendix Fig: 5 Dataset Information



```
print("First 5 Rows:")
df.head()
```

```
First 5 Rows:
```

	Site_Id	Timestamp	CO	HUMID	NO2	OZONE	PM10	PM2.5	SD1	TEMP	WDR	WSP
0	39	2015-01-06 00:00:00	0.275744	76.150	1.140613	0.80395	6.484	8.034	49.993	9.700	292.368	0.924
1	39	2015-01-06 01:00:00	0.275744	59.720	1.140613	0.80395	0.477	4.407	27.682	11.954	308.139	2.047
2	39	2015-01-06 02:00:00	0.076486	56.884	0.093000	2.10390	5.542	3.822	21.372	12.436	314.578	2.827
3	39	2015-01-06 03:00:00	0.087004	59.945	0.235300	1.97130	3.896	2.004	20.994	11.786	310.337	2.709
4	39	2015-01-06 04:00:00	0.078759	59.265	0.375000	1.89800	7.401	3.383	21.597	11.968	305.261	2.091

Appendix Fig: 6 Displaying Few lines of the Dataset