
AIR QUALITY ANALYSIS IN CALGARY

Member Name	Section
ENGI TAKLA (30018332)	L02
SHAHDAD VAHDATI DANESHMAND (30113452)	L02
JARON BARNETT (30121169)	L02
HARNOOR SINGH (30087486)	L01

SENG 550: SCALABLE DATA ANALYTICS

INSTUCTORS SARAH SHAH & ARMIN ZIRAK

DECEMBER 19, 2024

TABLE OF CONTENTS

<i>TABLE OF TABLES</i>	3
<i>TABLE OF FIGURES</i>	4
<i>PREAMBLE</i>	5
<i>ABSTRACT</i>	7
<i>INTRODUCTION</i>	8
<i>METHODOLOGY</i>	11
<i>RESULTS</i>	14
<i>REFERNCES</i>	29

TABLE OF TABLES

TABLE 1: MEMBER SIGNATURES 6

TABLE 2. LIST OF 17 PARAMETERS FOUND WITHIN THE DATASET..... 12

TABLE OF FIGURES

FIGURE 1: HOURLY VARIATIONS OF PARAMETERS SHARING THE PPM UNIT	15
FIGURE 2: HOURLY VARIATION FOR WIND SPEED	15
FIGURE 3: HOURLY VARIATION FOR WIND DIRECTION.....	16
FIGURE 4: HOURLY VARIATION FOR RELATIVE HUMIDITY	16
FIGURE 5: HOURLY VARIATION FOR WIND DIRECTION.....	17
FIGURE 6: MONTHLY AVERAGES FOR PARAMETERS SHARING THE PPM UNIT.....	18
FIGURE 7: MONTHLY AVERAGE CO LEVELS	18
FIGURE 8: MONTHLY AVERAGES FOR WIND SPEED	19
FIGURE 9: MONTHLY AVERAGES FOR PARTICULAR MATTER	19
FIGURE 10: MONTHLY AVERAGES FOR WIND DIRECTION.....	20
FIGURE 11: MONTHLY AVERAGES FOR RELATIVE HUMIDITY	20
FIGURE 12: ANNUAL AVERAGES FOR PARAMETERS SHARING THE PPM UNIT	21
FIGURE 13: ANNUAL AVERAGES FOR WIND SPEED	22
FIGURE 14: ANNUAL AVERAGES FOR WIND DIRECTION	22
FIGURE 15: ANNUAL AVERAGES FOR RELATIVE HUMIDITY	23
FIGURE 16: ANNUAL AVERAGES FOR PARTICULAR MATTER.....	23
FIGURE 17: WIND DIRECTION AVERAGES BY STATION	24
FIGURE 18: WIND SPEED AVERAGE BY STATION	25
FIGURE 19: PPM PARAMETERS BY STATION.....	25
FIGURE 20: RELATIVE HUMIDITY AVERAGES BY STATION.....	26
FIGURE 21: PARTICULATE MATTER AVERAGES BY STATION.....	26

PREAMBLE

CONTRIBUTIONS

Below is a breakdown of the contributions made by individual team members, along with an estimate of their total contribution percentage:

- **Shahdad (25%):**
 - Defined the project scope and problem statement.
 - Provided an overview of the dataset and experimental setup.
 - Created a comprehensive boilerplate for the code
 - Documented the methodology section.
- **Engi (25%):**
 - Performed exploratory data analysis, focusing on temporal and spatial trends.
 - Documented the Abstract, data analysis questions, and conclusion sections
 - Edit and format final report.
- **Harnoor (25%):**
 - Identified gaps in existing research and highlighted the specific contributions of the project.
 - Reviewed related work and summarized previous approaches in air quality analysis.
 - Summarized the proposed outcomes and key findings of the project as well as limitations.
- **Jaron (25%):**
 - Summarized the proposed outcomes and key findings of the project.
 - Interpreted results from the exploratory analysis and documented conclusions.
 - Wrote the experimentation factors and experiment process sections.

DECLARATION

We, the undersigned, hereby declare that the above statement of contributions and estimate of total contribution percentages is accurate and true.

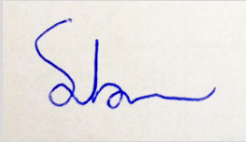

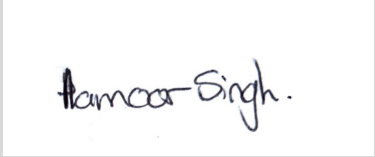

Member Name	Signature
Shahdad Vahdati Daneshmand	
Engi Takla	
Harnoor Singh	
Jaron Barnett	

TABLE 1: MEMBER SIGNATURES

LINK TO REPOSITORY

You can find our source code for our Google Colab Notebook here:

<https://colab.research.google.com/drive/1rlPPrPhhL819y2aabLDeYd4Jku9CdZP4?usp=sharing>

You can also find a public repository to our GitHub repository here:

<https://github.com/Shahdad20/SENG550-DataAnalytics>

ABSTRACT

This project focuses on analyzing Calgary's air quality using a near real-time dataset provided by the City of Calgary. The dataset includes measurements from three monitoring stations, capturing pollutant-specific data, spatial information, and timestamps. The goal of the analysis was to explore temporal trends, spatial variations, and pollutant-specific patterns to better understand Calgary's air quality and its potential implications for public health and environmental policy. The analysis employed a straightforward methodology, including basic data cleaning such as replacing missing values with 0 and trimming whitespace for consistency. Data exploration revealed distinct temporal trends, including peak pollution periods and variations by time of day or season. Spatial analysis, based on station-specific data, identified areas with consistently higher or lower pollutant levels, providing localized insights. Pollutant-specific patterns were also examined to determine their contribution to air quality concerns. While the dataset provided valuable insights, its limitation to three stations out of many across Calgary restricts the generalizability of the findings. Despite this, the analysis highlights key patterns and risks, offering actionable information for public health advisories.

INTRODUCTION

THE PROBLEM

The problem we selected involves analyzing air quality in Calgary to understand trends and variations in pollutant levels using historical data. By examining pollutants such as particulate matter, nitrogen dioxide, and ozone, patterns influenced by factors like traffic, industrial activity, and seasonal changes can be uncovered. This analysis is significant for both public health and policymaking. Poor air quality is linked to respiratory issues, cardiovascular diseases, and increased mortality rates, making it crucial to provide residents, especially vulnerable groups like children and the elderly, with the knowledge needed to take preventive measures. Additionally, reliable air quality data aids city planners in implementing strategies to reduce emissions, such as improving traffic flow or regulating industrial activities, thereby contributing to a sustainable urban environment.

DATASET DESCRIPTION

The dataset used for this analysis is the City of Calgary's air quality dataset. It includes key columns such as *ReadingDate*, which records the date and time of each measurement, and *Station Name*, identifying the specific monitoring station where the reading was taken. The *Parameter* column specifies the type of pollutant or environmental measurement (e.g., PM2.5, NO2), while the *Value* column provides the measured concentration of the parameter. Additionally, the dataset includes *Latitude* and *Longitude* to indicate the geographic coordinates of each monitoring station. With its near real-time measurements and spatial and temporal attributes, this dataset enables comprehensive analyses of variations in air quality across different locations and time periods.

This project aims to leverage historical data to provide a better understanding of carbon monoxide and offering insights into air quality trends and potential risks. The dataset can be found [here](#) and contains near real-time air quality measurements. It is accessible via API or exportable as a CSV file, with CSV being the preferred option for its straightforward integration into analysis pipelines.

IMPORTANCE

Air quality has a significant impact on public health and urban sustainability in Calgary. Exposure to pollutants can worsen respiratory conditions such as asthma and bronchitis, with seasonal variations like increased particulate matter during wildfire seasons posing additional risks. From an urban planning perspective, Calgary's growing population and industrial zones contribute to localized air quality issues, making it essential to design cleaner transportation systems and implement effective industrial regulations. Seasonal factors, including weather patterns like Chinooks and wildfire smoke, also play a major role in influencing pollutant levels. Understanding these dynamics is crucial for promoting long-term environmental sustainability. This analysis aims to support data-driven interventions that improve Calgary's air quality, contributing to a healthier and more livable city.

LITERATURE REVIEW

Many previous studies in the field of air quality analysis have focused on the use of real-time monitoring to inform public health advisories during high-risk events such as wildfires or extreme weather conditions. These studies have primarily conducted exploratory analyses of pollutant trends in urban areas, often examining pollutant levels on a general scale without delving into granular or

localized trends. While some work has included limited predictive modeling for localized air quality forecasting, much of the focus has been on evaluating air quality against environmental standards and promoting public awareness. Research on pollutants such as Carbon Monoxide (CO) and Fine Particulate Matter (PM_{2.5}) has been prevalent; however, these studies typically lack a connection to localized contexts like Calgary.

Despite these contributions, prior work has notable limitations. There is a lack of granularity in exploring seasonal and station-specific variations in pollutant levels, which is crucial for targeted interventions. Additionally, there has been minimal focus on providing actionable outcomes that could guide policymakers or public health officials. Furthermore, many studies have failed to employ visualizations that simplify complex pollutant patterns, making it challenging for non-technical audiences to engage with the findings. These gaps highlight the need for more detailed, localized, and accessible air quality analyses.

GAPS AND CONTRIBUTIONS

Prior analyses of Calgary's air quality have significant gaps that this project aims to address. One notable gap is the lack of detailed temporal and spatial analysis. Previous studies have often overlooked seasonal patterns, such as PM_{2.5} spikes during wildfire season, or failed to examine station-specific variations. There has been limited exploration of pollutant variations across time scales, such as monthly, hourly, and seasonal trends, which are crucial for understanding air quality dynamics. Another gap lies in the need for actionable insights that can directly benefit policymakers and residents. Few studies provide practical recommendations for urban planners or public health officials to mitigate air quality risks. Additionally, there has been limited emphasis on using visualizations to simplify complex trends. This has resulted in insufficient identification of high-risk periods or areas, making it difficult for non-technical audiences to engage with and act on the findings.

This project contributes to closing these gaps through a focused exploration of Calgary's unique air quality patterns. In the temporal analysis, month-by-month insights are provided for key pollutants such as CO and NO₂, highlighting seasonal peaks like wildfire activity in the summer and CO increases in the winter. The spatial analysis identifies high-risk areas, including Central Inglewood, which consistently reports elevated levels of PM_{2.5} and NO₂, as well as the industrial influences in Southeast Calgary. Cleaner areas like Varsity are also highlighted as benchmarks for better air quality.

Additionally, the project enhances public understanding by presenting actionable insights through intuitive visualizations such as line plots. These visualizations support data-driven recommendations for public health advisories and urban planning. By combining temporal and spatial analysis with accessible visual tools, this project provides a comprehensive framework to address Calgary's air quality challenges effectively.

PROPOSAL

We propose to develop a comprehensive data analysis that focuses on understanding the changes in air quality in the city of Calgary. Leveraging historical air quality data from the city, we aim to identify seasonal trends, daily variations, and station-specific patterns for a variety of parameters.

Our analysis will also investigate discrepancies across monitoring stations, and explore correlations between pollution levels and external factors, such as traffic or weather conditions.

Prior to analyzing the dataset, we expected to observe trends influenced by seasonal and regional factors. Air quality was assumed to be worse during Winter. Lower temperatures result in denser air which traps pollutants closer to the ground and there would be a spike in emissions as heating is used more liberally to stay warm. We also predicted that there would be a spike in pollution during the wildfire season (May-September) especially if we experienced a dry and windy Summer. On a day-to-day basis, we expect that conditions will worsen during high traffic times, namely morning and evening rush hour and during extreme weather events. Indicators of poor air quality should be more apparent in areas near industrial zones, major roadways and densely populated neighborhoods.

Our main Data Analysis questions are the following:

1. How do the pollutant levels vary by hour?
2. How do the pollutant levels vary by month?
3. How do the pollutant levels vary annually?
4. How do the pollutant levels vary from station to station?
5. What months/seasons have higher pollutant levels?
6. Are there times of the day where pollutant levels are higher?

METHODOLOGY

DATA FEATURES

The dataset used in this project contains several key attributes that enable a comprehensive analysis of air quality in Calgary. The temporal data, represented by the ReadingDate column, provides timestamps for each measurement. This allows for the identification of trends over time, such as hourly, daily, or seasonal variations in pollutant levels. By leveraging this temporal information, we can analyze patterns like peak pollution periods or changes across different seasons.

For the spatial aspect, the analysis focused on the monitoring stations listed in the Station Name column, with each station being uniquely associated with specific Latitude and Longitude coordinates. These coordinates provide a fixed geographical location for each station, enabling a precise mapping of pollutant levels. This information allowed us to compare air quality across three locations in Calgary (Varsity, Central Inglewood, Southeast) and identify potential hotspots or areas with better air quality.

Another important pair of columns are the Parameters and Value columns, which detail the specific pollutants measured (e.g., CO, Wind Direction, etc.) and their corresponding concentrations. These attributes enable pollutant-specific analyses, allowing us to examine trends and relationships for individual pollutants and assess their contribution to the overall air quality.

The data cleaning process involved handling missing values, and trimming whitespace. Missing values were replaced with '0' to handle gaps in the measurements, and unnecessary whitespace in text fields, such as Station Names and Parameters, was trimmed to ensure consistency in data formatting. These steps helped prepare the dataset for analysis while maintaining its integrity.

EXPERIMENT SETUP

The experimental setup involves a structured process to analyze Calgary's air quality data effectively. The first step is **data preparation**, which includes loading the dataset into Google Colab and performing data cleaning using PySpark to handle large-scale data efficiently. This involves addressing missing or inconsistent values in critical columns (e.g., Value) and ensuring data types and formats are consistent, such as converting dates to a uniform datetime format. Next is **data exploration and visualization**, where Pandas and Matplotlib are used for initial exploratory data analysis (EDA). Pollutant levels are plotted over time to identify seasonal trends or anomalies, while spatial distributions are visualized using scatter plots or heatmaps. Seaborn is utilized for more sophisticated statistical visualizations, such as examining correlations between parameters. The final step is **data splitting**, dividing the dataset into multiple parts for individual analyses.

The tools and libraries used include **Google Colab**, a cloud-based environment that ensures accessibility and scalability, **PySpark** for efficient processing of large datasets, and **Pandas**, **Matplotlib**, and **Seaborn** for creating graphs and visualizations to understand trends and correlations. By combining PySpark's big data processing capabilities with the granular analysis and visualization tools provided by Pandas and Matplotlib, this setup balances scalability with detailed analytical insights.

EXPERIMENTATION FACTORS

Generally, the factors that we most wanted to focus on were temporal groupings and spatial analysis. We used various time frames to look at the variations in parameters to get an idea of what kind of things would influence air quality. Hourly trends would be more useful for finding links to things such as the day-to-day actions of people such as commuting to work. Monthly trends would be more useful for showing the impact of the seasons and everything that comes with them from temperature, humidity to precipitation and more. Yearly trends could allow us to find correlations to other world events and track patterns over a longer period of time. Spatial analysis would be based on the locations of the stations. Considerations could then include the effect of proximity to industrial sectors, high traffic roadways, high population neighbourhoods and more.

We ultimately chose to not pursue a machine learning component for this project. Instead, we were hoping to hone in more on a statistical and graphical analysis of the selected data.

EXPERIMENT PROCESS

The analysis began by setting up the computational environment using PySpark, pandas, Matplotlib and Seaborn in Google Colab. The dataset, a CSV file, was sourced from the City of Calgary. Preprocessing steps included data cleaning, where null values were replaced with 0, and white spaces were trimmed. Although we initially considered removing duplicate entries, this step was discarded due to unexpected behavior in the data. The cleaned dataset was then inspected, starting with the identification of all parameters, resulting in a list of 17; see table 2.

parameter	units
Air Quality Health Index	NA
Hydrogen Sulphide	ppm
Ozone	ppm
Relative Humidity	%
Total Hydrocarbons	ppm
Nitrogen Dioxide	ppm
Outdoor Temperature	°C
Wind Speed	kph
Non-methane Hydrocarbons	ppm
Carbon Monoxide	ppm
Nitric Oxide	ppm
Fine Particulate Matter	µg/m3
Sulphur Dioxide	ppm
Inhalable Particulate Matter	µg/m3
Wind Direction	°
Methane	ppm
Total Oxides of Nitrogen	ppm

TABLE 2. List of 17 parameters found within the dataset

Monthly averages for each parameter were calculated and plotted according to their respective units of measurement and annual averages were determined for the dataset's coverage period (2021–2024). Further analysis focused on identifying seasonal variations in wind direction and relative

humidity as well as peak values for each parameter by hour to detect daily spikes. We also examined variations between monitoring stations to assess potential geographical influences on air quality. We also began a more thorough exploration into Carbon Monoxide data, tracking the average monthly values from September 2021 through December 2024.

RESULTS

KEY RESULTS (EDA)

Since no machine learning was used, we focused on data analysis.

The analysis covered all major pollutants, including PM2.5, CO, NO2, ozone, and relative humidity, across all available monitoring stations. Temporal trends were thoroughly examined, including hourly, monthly, and seasonal variations, to reveal patterns such as the summer peaks in pm2.5 during wildfire season and the winter spikes in co due to increased heating. Spatial trends were also analyzed to identify high-risk areas such as Central Inglewood, which reported elevated pollutant levels, and relatively clean areas such as Varsity.

Visualizations were designed to communicate findings and address the research questions effectively. Line charts depict temporal trends, such as monthly and seasonal variations in PM2.5 and CO levels. Visuals illustrate spatial patterns across monitoring stations, highlighting pollutant hotspots like Southeast Calgary. Seasonal trends for parameters like Wind Direction and Relative Humidity were presented through clear and concise line charts. Each visualization included detailed titles, labeled axes, and legends to enhance clarity. Units of measurement, such as $\mu\text{g}/\text{m}^3$ for PM2.5 or ppm for CO, were explicitly stated. Intuitive color schemes were used to distinguish pollutants, ensuring that charts were accessible and easy to interpret for both technical and non-technical audiences.

1. How do the pollutant levels vary by hour?

Temporal analysis revealed clear hourly patterns. Please see Figures 1 – 5.

Carbon Monoxide (CO) levels show significant peaks during the morning (7–9 AM) and evening (5–7 PM) rush hours, likely due to increased vehicle emissions. Please see Figure 1.

Methane (CH4) Displays relatively stable levels throughout the day, with slight increases during late afternoon (4–6 PM), likely due to industrial activity or natural gas emissions. Please see Figure 1.

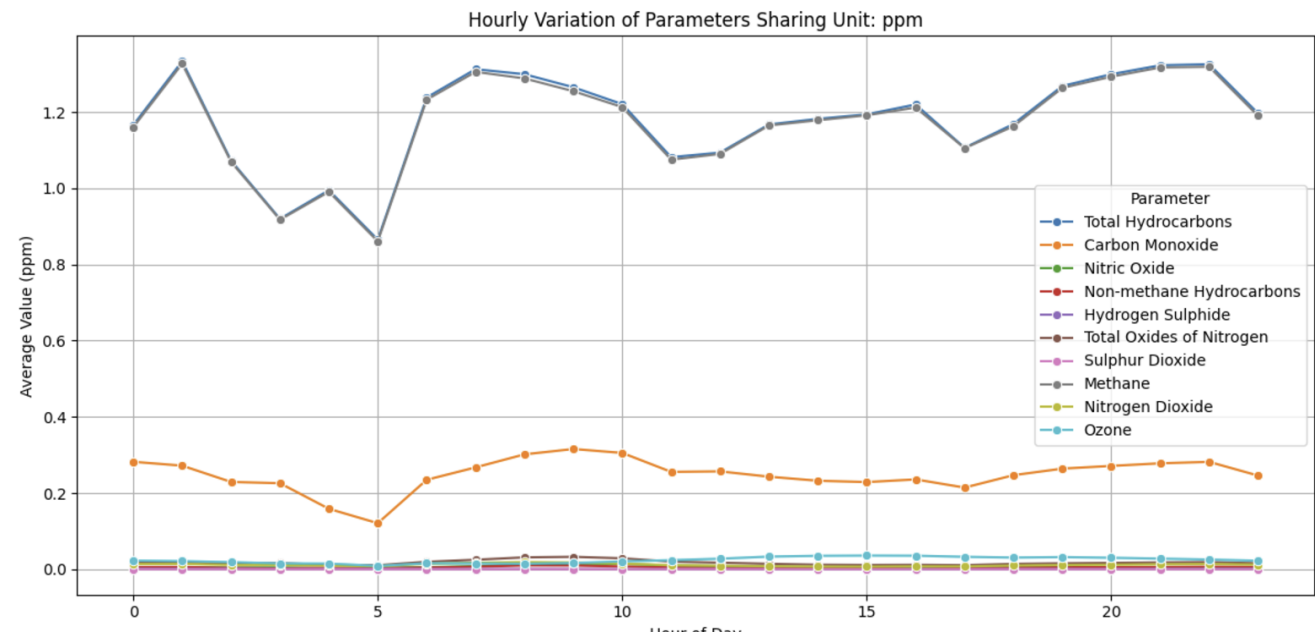


FIGURE 1: HOURLY VARIATIONS OF PARAMETERS SHARING THE PPM UNIT

Wind Speed peaks during mid-afternoon (2–4 PM), correlating with increased atmospheric mixing. This impacts pollutant dispersion and leads to lower concentrations for many ground-level pollutants during this period. Please see Figure 2.

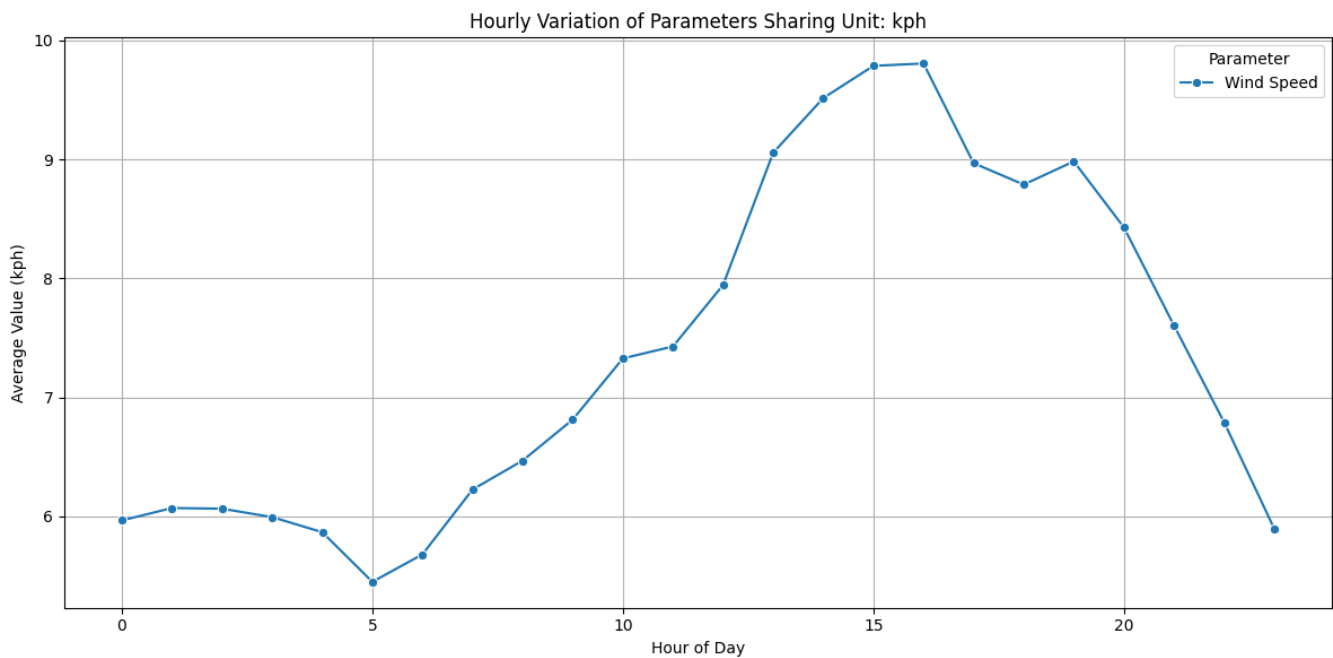


FIGURE 2: HOURLY VARIATION FOR WIND SPEED

These hourly variations emphasize the importance of controlling traffic emissions during peak hours to reduce pollutant levels. Public health advisories can target vulnerable populations (e.g., children, elderly) to avoid outdoor activities during high-pollution hours.

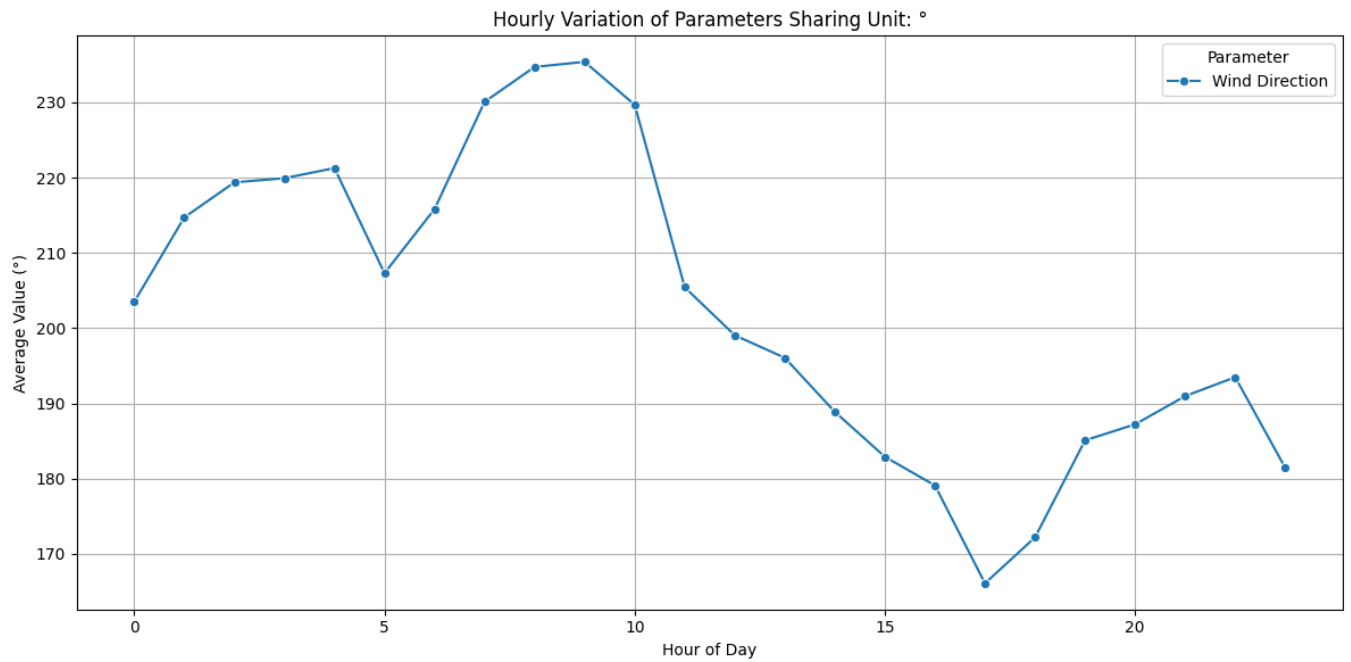


FIGURE 3: HOURLY VARIATION FOR WIND DIRECTION

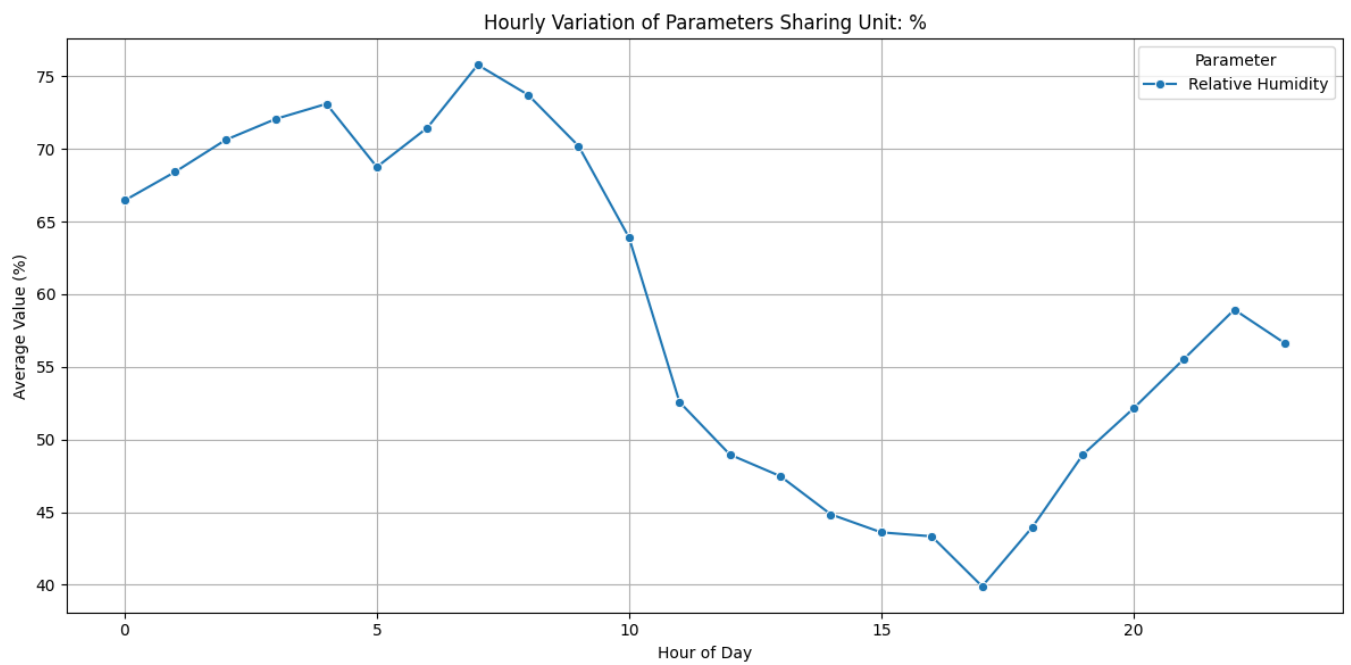


FIGURE 4: HOURLY VARIATION FOR RELATIVE HUMIDITY

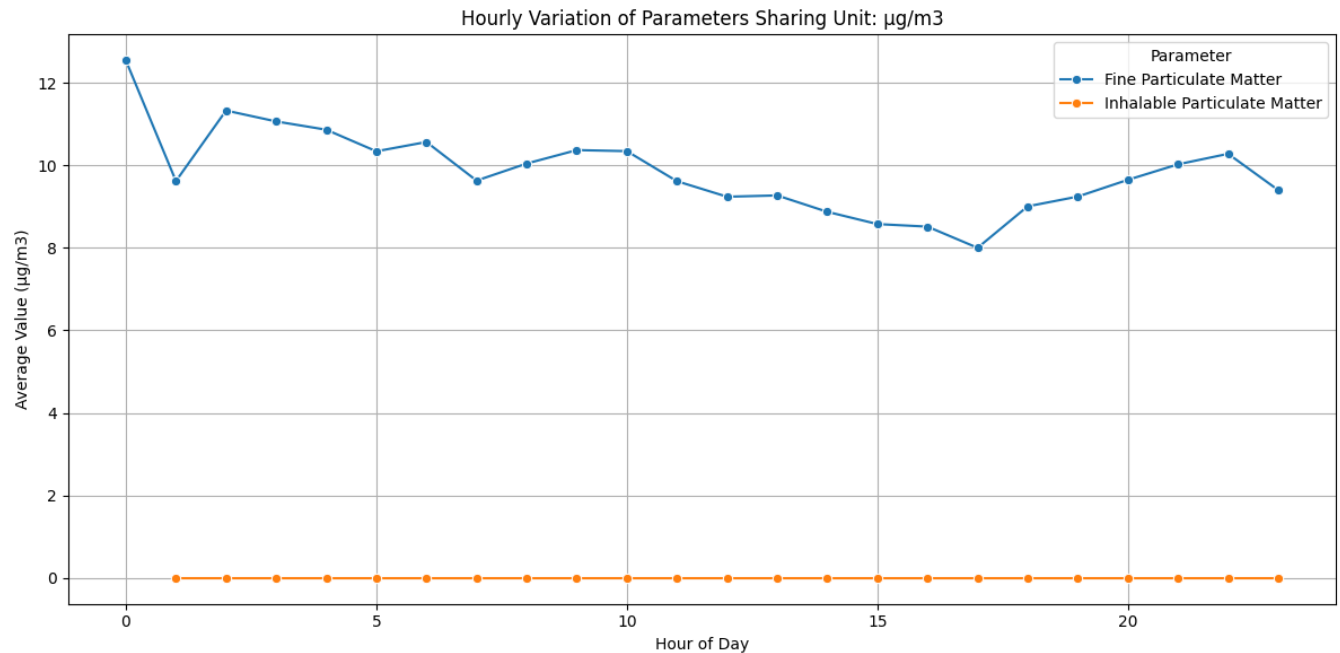


FIGURE 5: HOURLY VARIATION FOR WIND DIRECTION

2. How do the pollutant levels vary by month?

Monthly trends provide a clear understanding of seasonal variations. Please see Figures 6 – 11.

CO levels are highest in December and January, during the colder months, due to increased heating and lower atmospheric dispersion.

Ozone concentrations peak in May–June, driven by higher sunlight intensity, which enhances photochemical reactions.

Methane (CH_4): Consistently elevated in winter months (December–February), reflecting increased heating and industrial emissions.

Hydrocarbons: Peak during summer (July–August), possibly due to photochemical reactions and emissions from volatile organic compounds (VOCs).

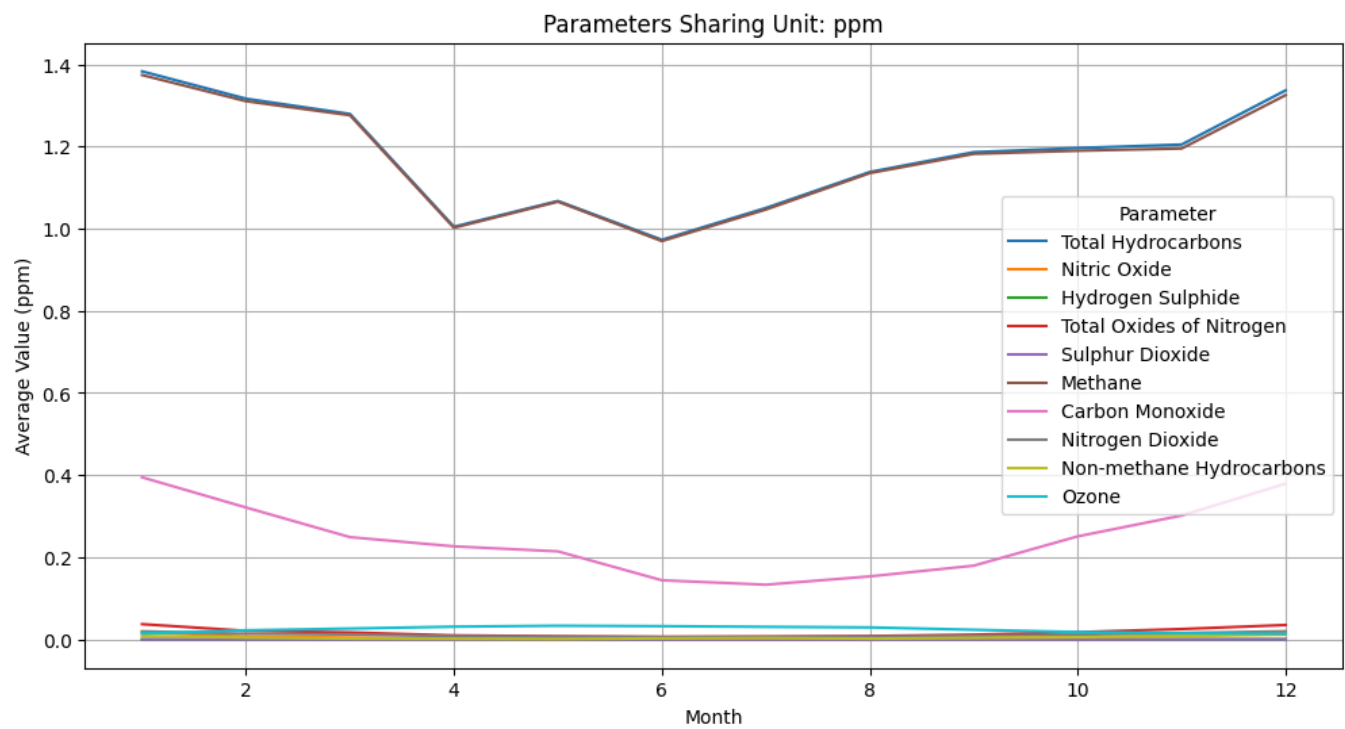


FIGURE 6: MONTHLY AVERAGES FOR PARAMETERS SHARING THE PPM UNIT

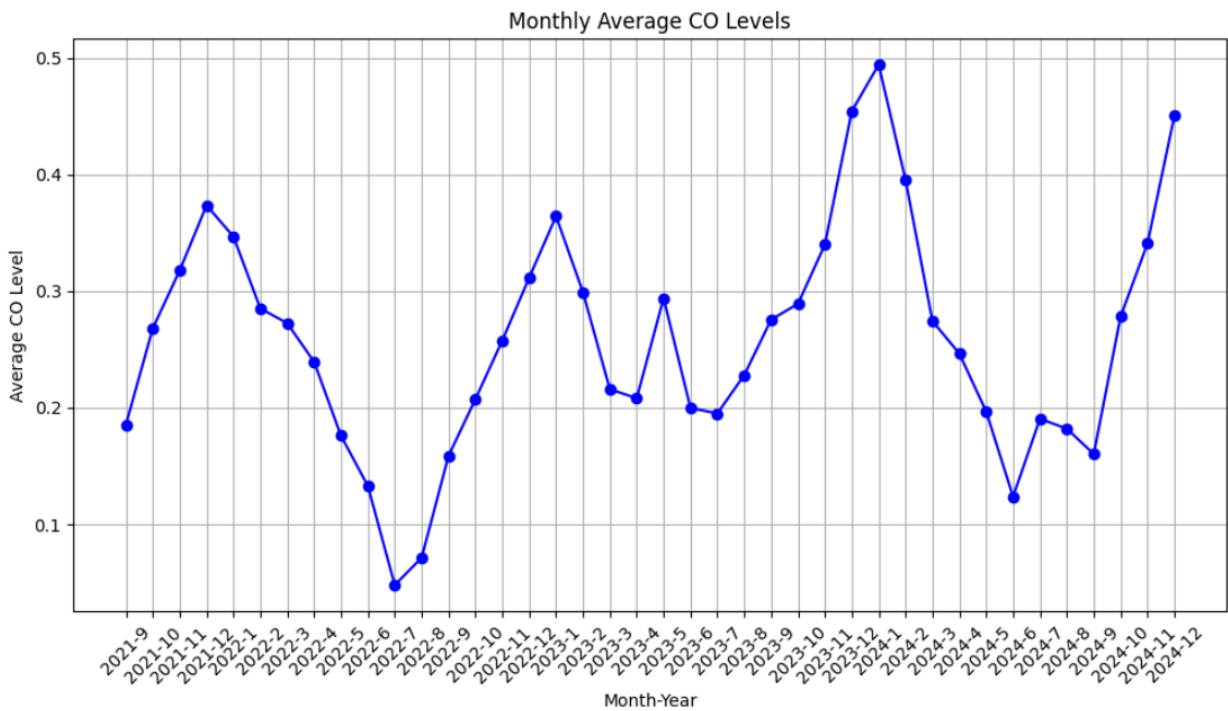


FIGURE 7: MONTHLY AVERAGE CO LEVELS

Wind Speed: Lowest in winter, reducing atmospheric dispersion and contributing to pollutant buildup

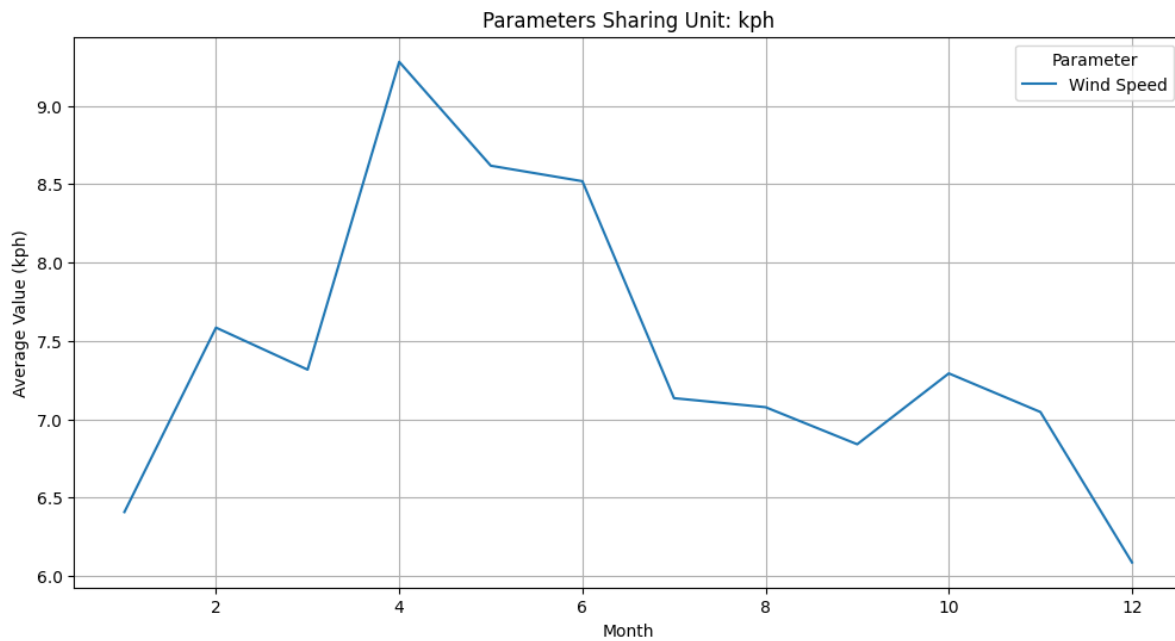


FIGURE 8: MONTHLY AVERAGES FOR WIND SPEED

PM2.5 levels peak during July and August, correlating with wildfire activity in the region.

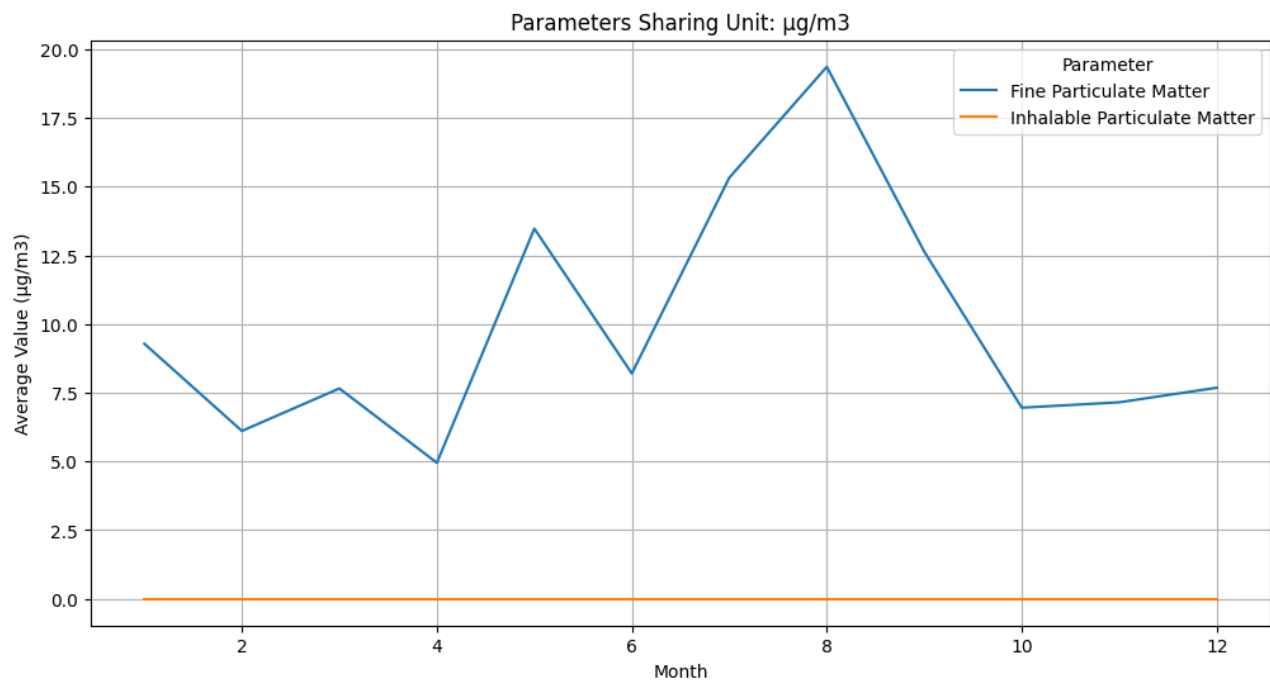


FIGURE 9: MONTHLY AVERAGES FOR PARTICULAR MATTER

These monthly patterns provide insights for seasonal planning. During wildfire seasons, additional monitoring and air quality advisories are critical. Winter months require proactive measures, such as improving heating efficiency and managing emissions from residential and industrial sources.

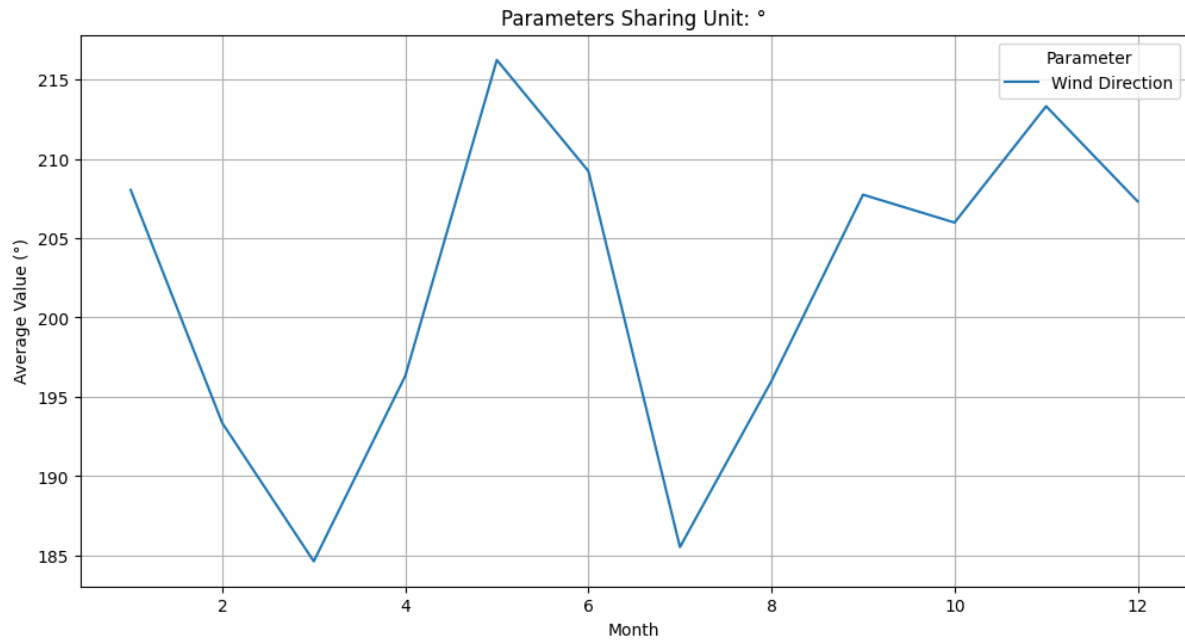


FIGURE 10: MONTHLY AVERAGES FOR WIND DIRECTION

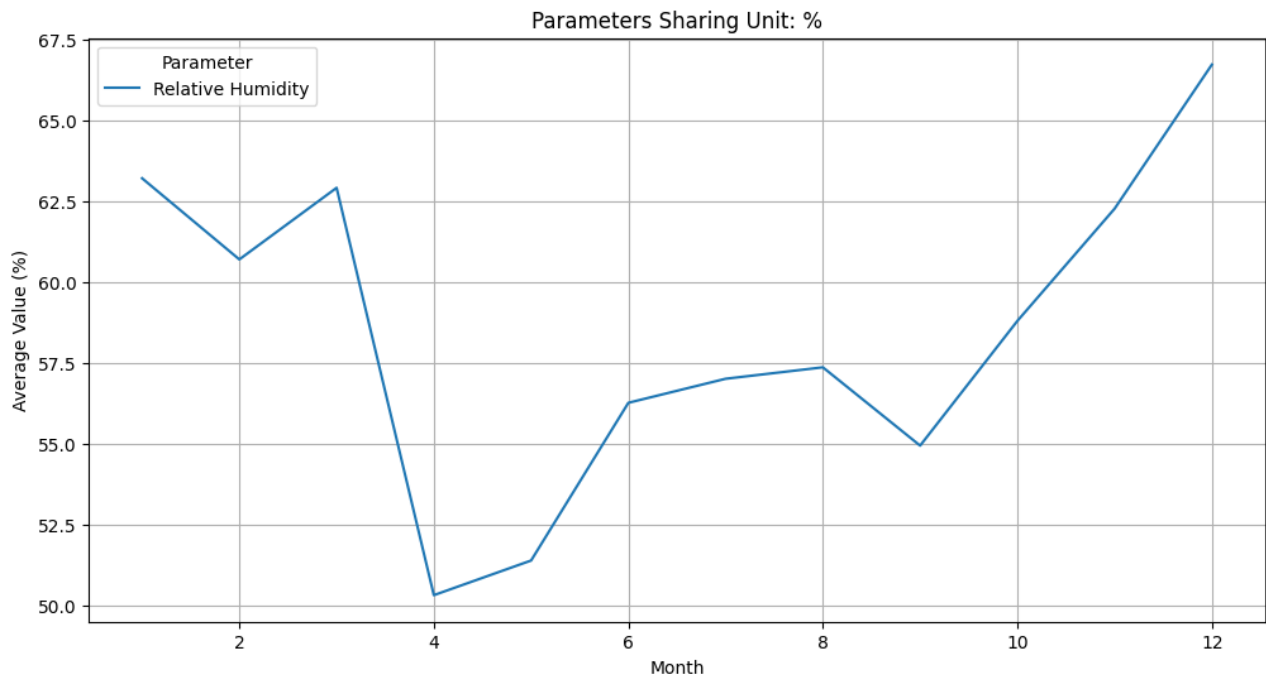


FIGURE 11: MONTHLY AVERAGES FOR RELATIVE HUMIDITY

3. How do the pollutant levels vary annually?

Annual variations show trends in pollutant levels over multiple years. Please see Figures 12 – 16.

Methane: Stable levels with minor fluctuations depending on industrial activity and natural gas usage. Please see Figure 12.

Hydrocarbons: Show variability tied to seasonal factors like increased photochemical activity in summer. Please see Figure 12.

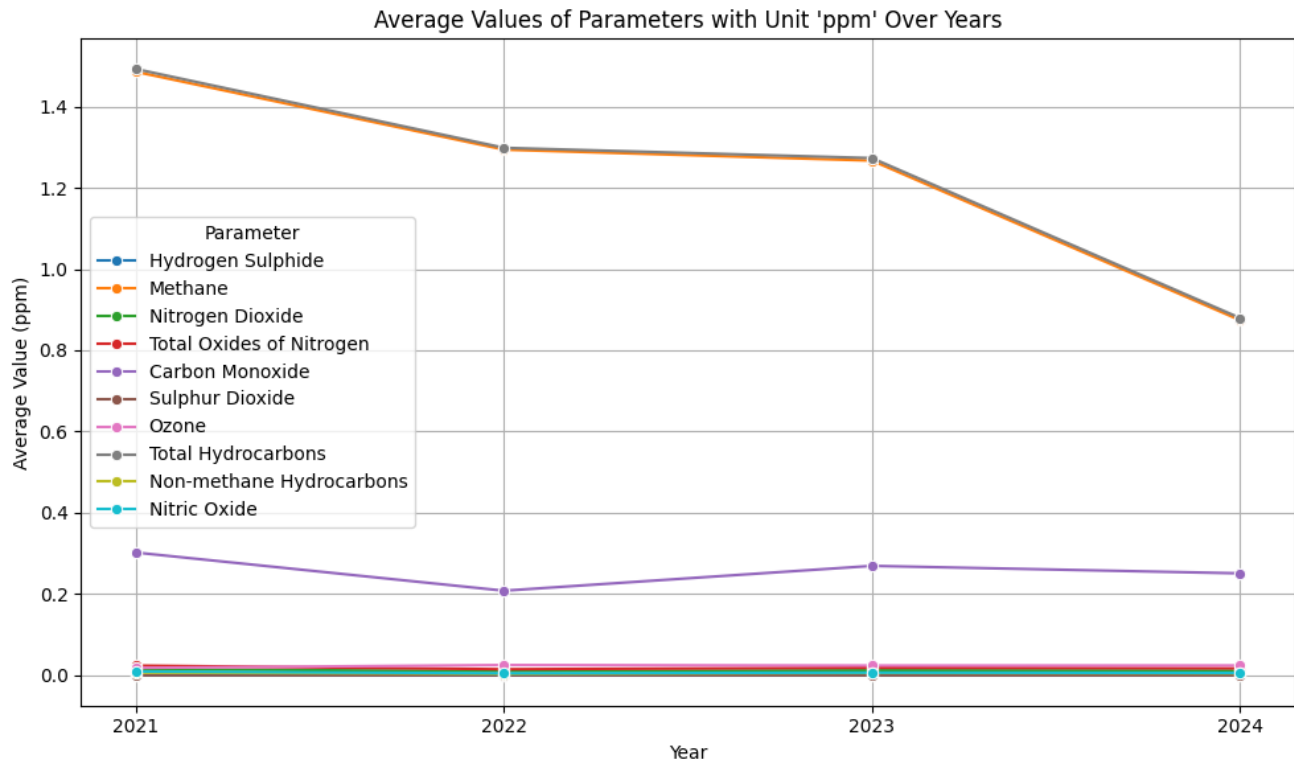


FIGURE 12: ANNUAL AVERAGES FOR PARAMETERS SHARING THE PPM UNIT

Wind Speed: Annual patterns remain consistent, with seasonal fluctuations driving pollutant dispersal trends. Please see Figure 14.

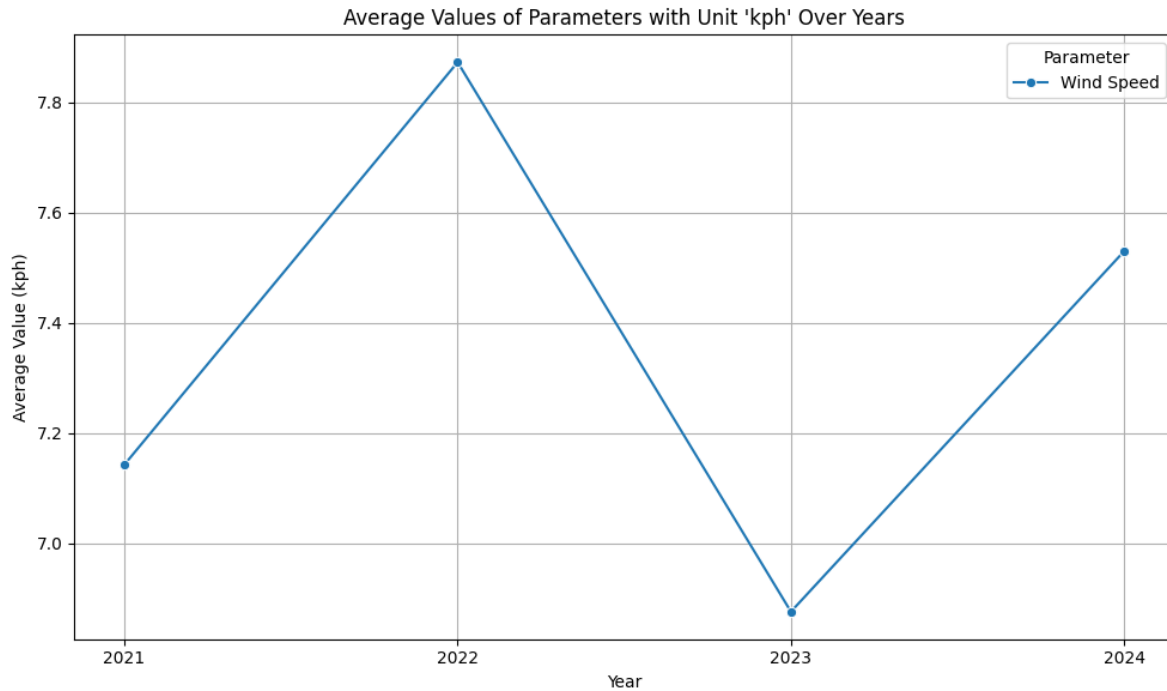


FIGURE 13: ANNUAL AVERAGES FOR WIND SPEED

A couple of things that we found were notable and interesting was the dip in the wind direction and speed in the year 2023, and humidity levels in 2022. We also noticed a spike in fine particulate matter in 2023.

Annual trends help assess the long-term effectiveness of policies like emission regulations. This insight can inform future urban planning and environmental policy decisions, ensuring sustainable air quality improvements.

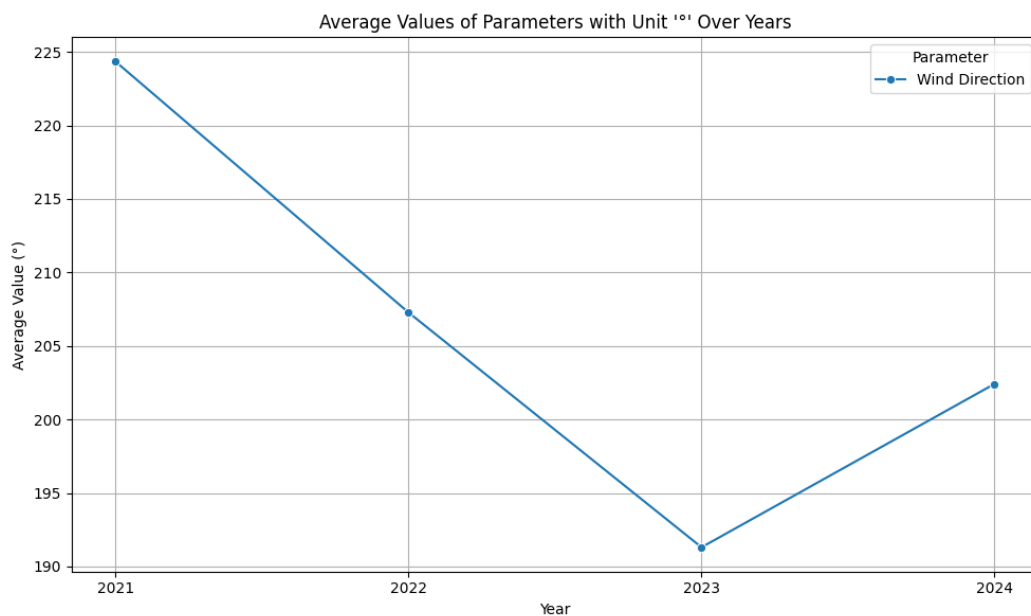


FIGURE 14: ANNUAL AVERAGES FOR WIND DIRECTION

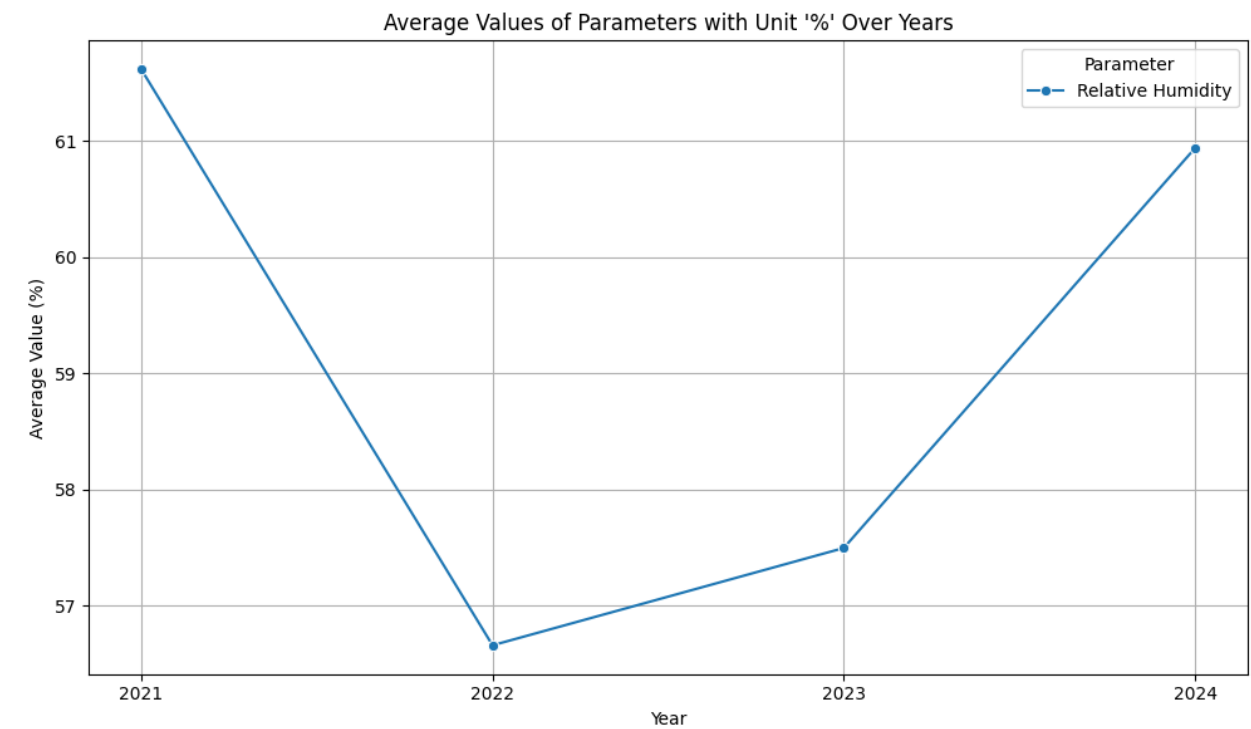


FIGURE 15: ANNUAL AVERAGES FOR RELATIVE HUMIDITY

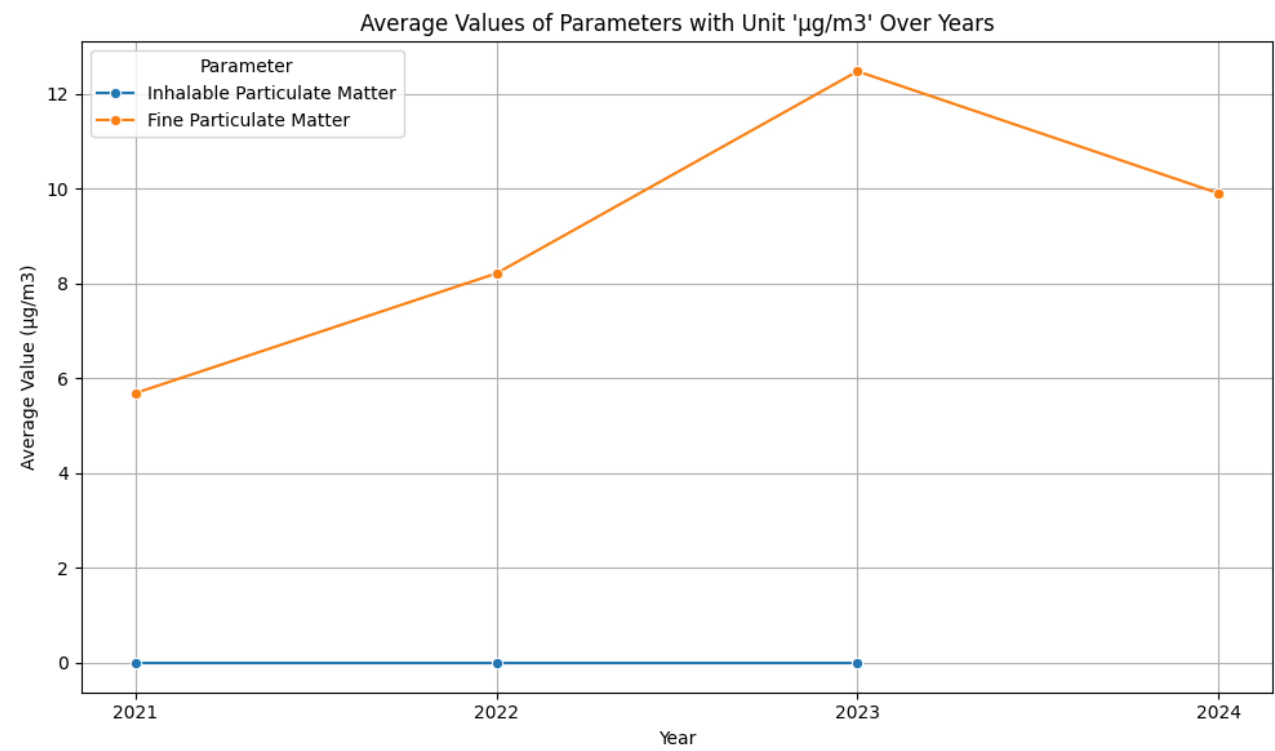


FIGURE 16: ANNUAL AVERAGES FOR PARTICULAR MATTER

4. How do the pollutant levels vary from station to station?

Central Inglewood:

- Reports higher levels of Relative Humidity, which contributes to pollutant persistence, and moderate levels of Hydrocarbons.
- Wind speeds are lower, limiting pollutant dispersion and potentially exacerbating pollution-related health risks.

Southeast Calgary:

- Elevated levels of Methane (CH₄) and Hydrocarbons indicate significant industrial emissions in the area.
- Wind Speed is moderate, which limits the effective dispersion of pollutants.
- Higher levels of pollutants like hydrocarbons correlate with nearby industrial zones and natural gas emissions.

Varsity:

- Consistently lower levels of Methane, Hydrocarbons, and other pollutants highlight better air quality in this residential area.
- Higher wind speeds in comparison to other locations aid in pollutant dispersion.

Please see Figures 17 – 21.

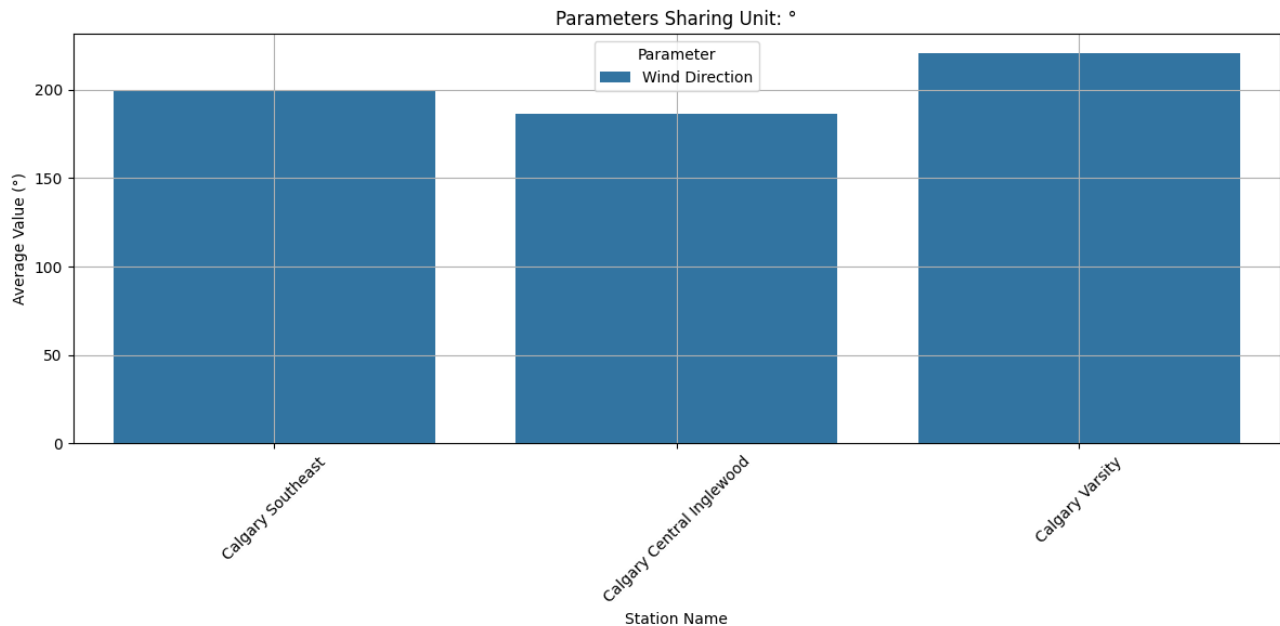


FIGURE 17: WIND DIRECTION AVERAGES BY STATION

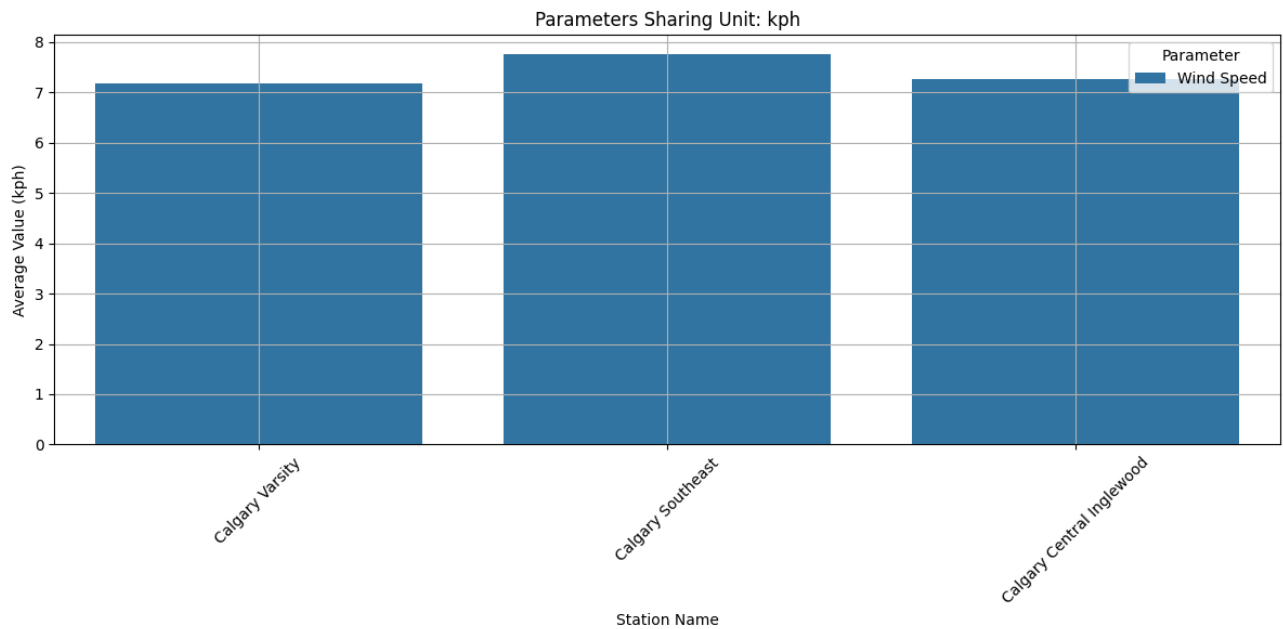


FIGURE 18: WIND SPEED AVERAGE BY STATION

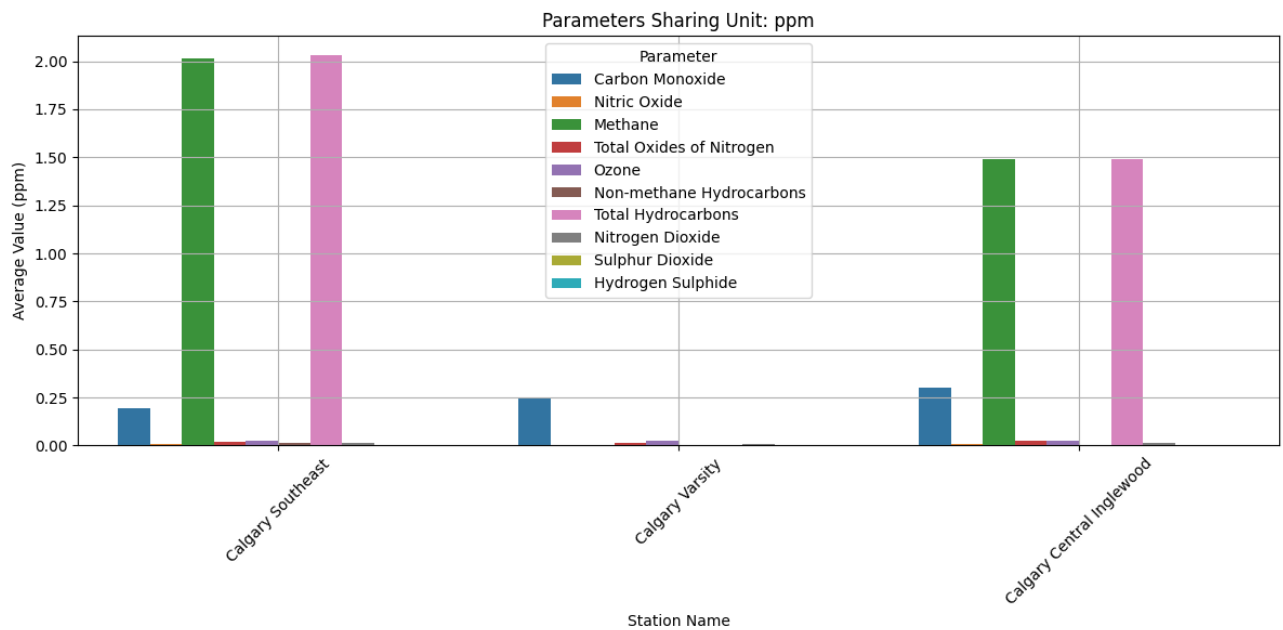


FIGURE 19: PPM PARAMETERS BY STATION

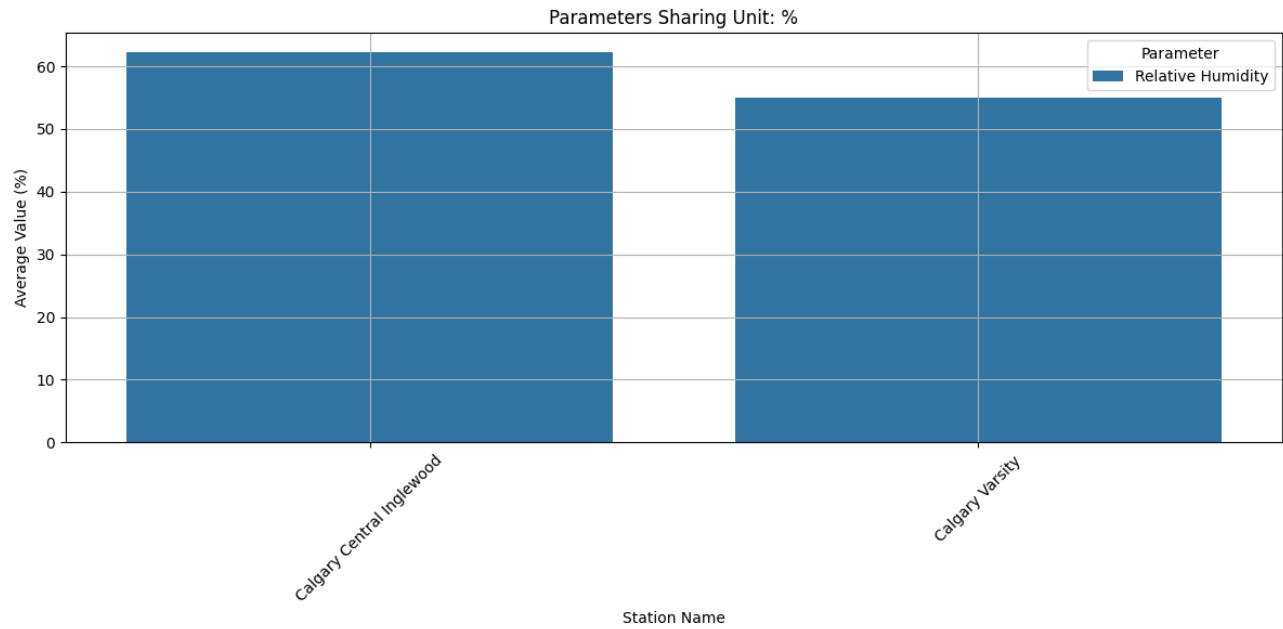


FIGURE 20: RELATIVE HUMIDITY AVERAGES BY STATION

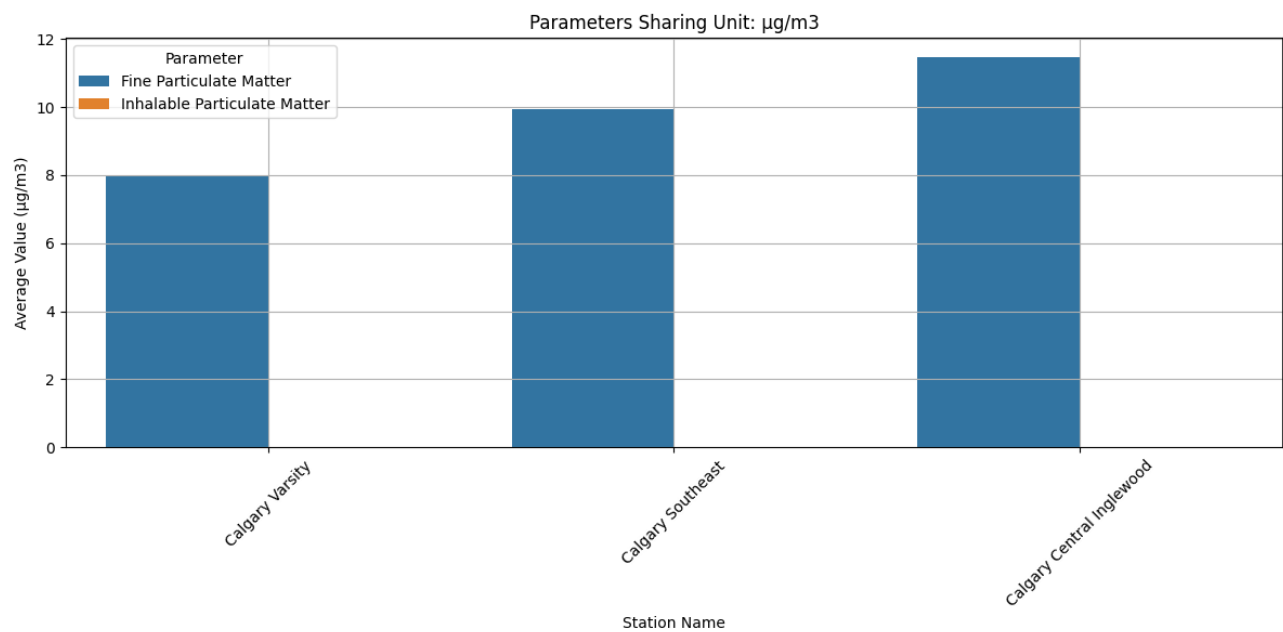


FIGURE 21: PARTICULATE MATTER AVERAGES BY STATION

CORRELATIONS

Some correlations we noticed were the following:

PM2.5 and NO2: A moderate positive correlation was observed, especially during winter months when heating and vehicular emissions contribute to both pollutants. This relationship is strongest in areas like Central Inglewood, where traffic and industrial activities are prominent.

Methane (CH₄) and Hydrocarbons: A positive correlation was observed between methane and hydrocarbons, particularly in industrial zones like Southeast Calgary. This suggests that emissions from similar sources, such as natural gas usage and volatile organic compounds (VOCs), contribute to both pollutants. The strength of this relationship is especially pronounced during winter months when heating and industrial activities increase.

Wind Speed and Methane (CH₄): A negative correlation exists between wind speed and methane concentrations. Higher wind speeds during the afternoon (2–4 PM) contribute to better dispersion, leading to lower methane levels. Conversely, during periods of low wind, methane concentrations remain elevated, particularly in areas with industrial activity.

Relative Humidity and Hydrocarbons: Higher relative humidity correlates with increased hydrocarbon concentrations, particularly in the late fall and winter months. Humidity likely enhances the persistence of hydrocarbons in the atmosphere, contributing to their accumulation.

Hydrocarbons and Wind Direction: Hydrocarbon levels show moderate correlations with specific wind directions, which may align with prevailing winds transporting industrial emissions into urban areas.

DISCUSSION AND LIMITATIONS

The Value column, containing pollutant measurements, had missing entries that were replaced with 0 to maintain dataset completeness. Some pollutants, such as Methane or Sulphur Dioxide, had sparse data compared to others like PM2.5 or CO, limiting their analysis.

The impact of missing data may bias temporal and spatial trends, especially for pollutants with limited coverage.

Some limitations in generalizing findings without advanced predictive models include:

- **Temporal Limitations:** While seasonal and hourly trends provide valuable insights, they are based solely on historical data. Without advanced predictive models, it is challenging to forecast pollutant levels under changing conditions, such as increased industrial activity or shifting weather patterns.
- **External Factors:** The analysis does not integrate external datasets, such as traffic volume, weather data (e.g., temperature, wind speed), or wildfire occurrences, which could provide a more holistic view of the factors influencing air quality. Predictive models could incorporate such variables to simulate future scenarios or evaluate the effectiveness of proposed interventions.

- **Rare Events:** Outliers, such as unexpected spikes in pollutant levels, are identified but not explained in detail. Predictive models, such as time-series forecasting, could help contextualize these anomalies or assess their recurrence probability.

CONCLUSION

In conclusion, this project analyzed Calgary's air quality using a dataset containing temporal, spatial, and pollutant-specific measurements from three monitoring stations across the city; Central Inglewood, Varsity, and Southeast. While the dataset provided valuable insights into temporal trends, spatial variations, and pollutant-specific patterns, it is important to note that the analysis was limited to these three stations, out of many more available throughout Calgary. This limitation means that the findings may not fully capture the city's overall air quality conditions. Data cleaning, such as replacing missing values with 0 and trimming whitespace, ensured a consistent and usable dataset for exploration. Despite its limitations, the analysis highlighted key areas and times of concern for air quality, with implications for public health, urban planning, and environmental policy. Future work could expand the scope by including data from additional stations and integrating external datasets, such as weather or traffic data, to provide a more comprehensive understanding of Calgary's air quality. This project underscores the value of leveraging local data to address environmental challenges and inform actionable solutions.

REFERENCES

Alberta Environmental Monitoring, E. and R. A. (2024, December 15). *Air Quality Data (near real-time): Open Calgary*. data.calgary.ca. https://data.calgary.ca/Environment/Air-Quality-Data-near-real-time-/g9s5-qhu5/about_data