Air Quality Analysis Report: Delhi (2021–2024)

Prepared By: VIDHYA V

Tools Used: SQL, Python (Pandas, Matplotlib, Seaborn), Power BI

Dataset: Delhi Air Quality Dataset (2021–2024)

Project Type: End-to-End Data Analytics Project

Year: 2025

EXECUTIVE SUMMARY

This report presents a multi-dimensional analysis of Delhi's air quality across four years (2021–2024). The study combines SQL queries, Power BI dashboards, and Python visualizations to identify pollutant behavior, seasonal variations, and weekly patterns. Findings highlight that PM10 and PM2.5 are the primary contributors to poor AQI, winter months show the worst pollution levels, and air quality improves during monsoon. These insights can support policymakers in designing effective pollution control measures.

INTRODUCTION

Air pollution remains one of the major environmental challenges in India. The aim of this analysis is to understand pollutant trends, seasonal patterns, and daily variations in AQI using multiple analytical approaches. SQL was used for data exploration, Power BI for dashboard insights, and Python for advanced visualization. This report combines all three to provide a holistic view of air

DATA OVERVIEW

How The Data Was Collected

- This project uses the **Delhi Air Quality Dataset** from Kaggle, which includes daily pollutant measurements (NO_2, CO, Ozone, PM_{10}, $PM_{2.5}$) and AQI values for Delhi.
  - Kaggle dataset link:

Kunsh Bhatia. (2024). *Delhi Air Quality Dataset*. Kaggle.
https://www.kaggle.com/datasets/kunshbhatia/delhi-air-quality-dataset?select=final_dataset.csv

Features Identified for Analysis

The dataset contains several critical features that influence air quality patterns across Delhi. Key variables include major pollutants such as PM2.5, PM10, NO_2, SO_2, CO, and Ozone, which directly affect AQI. Temporal variables such as Year, Month, Date, and Season help in identifying seasonal and long-term pollution trends. Additionally, derived features like AQI Category, Weekday/Weekend classification, and Day Type (Holiday vs Working Day) were created to study daily behavioral patterns. These features together enable comprehensive analysis of pollutant behavior and its impact on overall AQI.

SQL Data EXPLORATION

SQL 1. Top pollutants contributing to high AQI

```
 SELECT
    'PM10' AS pollutant, AVG(PM10) AS avg_level
FROM final_dataset
WHERE AQI > 100
UNION ALL
SELECT
    'PM2.5', AVG(PM2_5)
FROM final_dataset
WHERE AQI > 100
UNION ALL
SELECT
    'NO2', AVG(NO2)
FROM final_dataset
WHERE AQI > 100
UNION ALL
SELECT
    'SO2', AVG(SO2)
FROM final_dataset
WHERE AQI > 100
UNION ALL
SELECT
    'CO', AVG(CO)
FROM final_dataset
WHERE AQI > 100
UNION ALL
SELECT
    'O3', AVG(Ozone)
FROM final_dataset
WHERE AQI > 100
ORDER BY avg_level DESC;
```

100 % | 1 | 0

Results | Messages

	pollutant	avg_level
1	PM10	259.0390464469
2	PM2.5	107.437602893365
3	NO2	39.8294663420917
4	O3	36.5185826533944
5	SO2	20.3322572027083
6	CO	1.13719159959182

Note: This query identifies which pollutants contribute the most on highly polluted days (AQI > 100). It helps reveal the primary drivers of poor air quality.

SQL 2. Monthly AQI trend

```
select Month_name,avg(aqi) as avg_AQI,season
from final_dataset
group by month_name ,season order by avg_AQI desc

-----------------------------------------------------------------------------------------------
```

100 % ✖ 1 ⚠ 0

Results | Messages

	Month_name	avg_AQI	season
1	November	342	Post-Monsoon
2	January	305	Winter
3	December	297	Winter
4	February	239	Winter
5	April	222	Summer
6	March	200	Summer
7	May	199	Summer
8	October	191	Post-Monsoon
9	June	163	Monsoon
10	July	90	Monsoon
11	August	89	Monsoon
12	September	87	Post-Monsoon

Note: Shows how air quality varies month-wise across the year. Helps detect seasonal pollution patterns and identify high-risk months.

SQL 3. Top 10 AQI days in winter

```
select top 10
full_date , AQI,season
from final_dataset
where season='winter'
order by AQI desc
```

100 % ✖ 1 ⚠ 0

Results | Messages

	full_date	AQI	season
1	2023-01-07	500	Winter
2	2023-01-08	500	Winter
3	2021-12-26	490	Winter
4	2021-01-02	482	Winter
5	2021-12-24	473	Winter
6	2021-12-23	468	Winter
7	2021-12-25	467	Winter
8	2022-01-20	465	Winter
9	2021-01-01	462	Winter
10	2024-12-19	460	Winter

Note: Lists the most 10 polluted days during the winter season. These insights highlight winter as a critical period for severe air pollution.

SQL 4. Average AQI on weekdays vs weekends/holidays

```sql
select daytype, AVG(AQI) as Avg_AQI
from final_dataset group by DayType order by Avg_AQI desc
```

100 % 1 0

Results Messages

	daytype	Avg_AQI
1	Working Day	202
2	Holiday	202

Note: Compares pollution levels between working days and holidays. Helps analyze the impact of traffic and human activities on air quality.

SQL 5. Count of days per AQI category

```sql
select AQI_Category,
count(*) as total_days
from final_dataset
group by AQI_Category
order by total_days desc
```

Results Messages

	AQI_Category	total_days
1	Poor	463
2	Very Poor	384
3	Moderate	267
4	Severe	231
5	Hazardous	65
6	Good	51

Note: Shows how many days fall into each AQI category (Good, Moderate, Poor, etc.). Useful for presenting overall air quality distribution.

SQL 6. Average PM10 & PM2.5 by season

```sql
select AVG(PM10) as Avg_PM10,
AVG(PM2_5) as Avg_PM2_5,Season
from final_dataset
group by Season
order by Avg_PM10 desc
```

100 % | ✖ 1 | ⚠ 0

Results | Messages

	Avg_PM10	Avg_PM2_5	Season
1	292.615512869034	133.565429433891	Winter
2	229.459834593993	89.0862362987393	Post-Monsoon
3	228.666060188542	91.9097827880279	Summer
4	123.672961864782	49.3323098358941	Monsoon

Note: Displays how major particulate pollutants (PM10 & PM2.5) vary across seasons. Highlights seasons with the highest particulate concentration.

SQL 7. Total AQI TREND OVER TIME

```
select YEAR,sum(AQI) as Total_AQI
from final_dataset
group by Year order by Total_AQI desc
```

Results | Messages

	YEAR	Total_AQI
1	2021	78772
2	2022	76116
3	2024	71535
4	2023	69007

Note: Summarizes yearly AQI totals to show long-term pollution trends. Helps identify whether overall air quality is improving or worsening year-to-year.

SQL FINDINGS

SQL 1:

Finding: PM10 and PM2.5 have the highest values on high AQI days.
Insight: Particulate matter is the main contributor to poor air quality.

SQL 2:

Finding: AQI peaks in winter months (Dec–Feb).
Insight: Seasonal patterns affect pollution; winter requires more monitoring.

SQL 3:

Finding: Top 10 most polluted days occur in December and January, sometimes "Very Unhealthy."
Insight: Critical period for public health alerts.

SQL 6:

Finding: The average PM10 and PM2.5 concentrations are highest in the winter season and lowest in the monsoon season.

Insight: Air pollution peaks during winter, likely due to lower dispersion and increased emissions, indicating the need for targeted pollution control measures during colder months.

SQL 7:

Finding: The total AQI has been decreasing from 2021 (78,772) to 2023 (69,007), with a slight increase in 2024 (71,535).

Insight: Overall air quality is showing a gradual improvement over the years, suggesting that pollution control measures may be having a positive effect, though vigilance is needed to sustain this trend.

PYTHON ANALYSIS

Python cleaning

Python (Pandas) was used to clean and prepare the dataset for analysis.
The following cleaning steps were performed:

1.Loaded the dataset using Pandas

2.Checked for missing values (nulls)

3.Handled missing data by replacing or removing null values

4.Fixed data types

- Converted Date column to datetime
- Converted numeric columns from object/string to float

5.Created new columns

- Year, Month, Day, Weekday, Day type Season

6.Removed unwanted columns and duplicates

- After cleaning and preparing the dataset, several visualizations were generated to explore AQI patterns, seasonal variations, pollutant impact, and weekday-month interactions.

Snapshort of Pandas & Other Library Package -Profiling Reports

[14]
11s

```
from google.colab import files
uploaded = files.upload()

import pandas as pd
import matplotlib.pyplot as plt

df = pd.read_csv('final_dataset.csv')
df.head()
```

Choose Files final_dataset.csv
final_dataset.csv(text/csv) - 78368 bytes, last modified: 20/11/2025 - 100% done
Saving final_dataset.csv to final_dataset (1).csv

	Date	Month	Year	Holidays_Count	Days	PM2.5	PM10	NO2	SO2	CO	Ozone	AQI
0	1	1	2021	0	5	408.80	442.42	160.61	12.95	2.77	43.19	462
1	2	1	2021	0	6	404.04	561.95	52.85	5.18	2.60	16.43	482
2	3	1	2021	1	7	225.07	239.04	170.95	10.93	1.40	44.29	263
3	4	1	2021	0	1	89.55	132.08	153.98	10.42	1.01	49.19	207
4	5	1	2021	0	2	54.06	55.54	122.66	9.70	0.64	48.88	149

Figure 1: Loading the Dataset in Pandas Dataframe

```
yearly = df.groupby('Year')['AQI'].mean()

plt.figure(figsize=(3,4))
plt.plot(yearly, marker='o')
plt.title("AQI Trend Over Years")
plt.xlabel("Year")
plt.ylabel("Average AQI")
plt.grid(True)
plt.show()
```

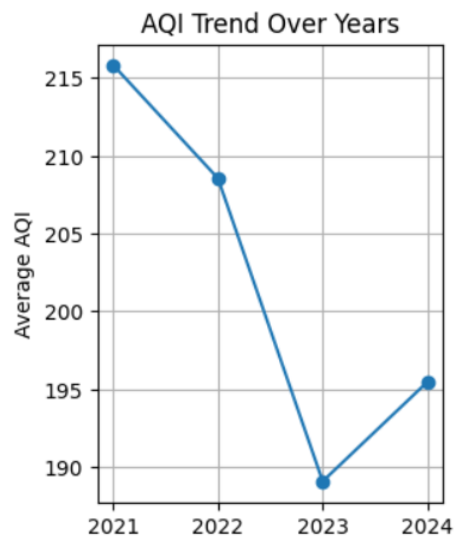


Figure 2: Line plot Showing AQI Trend over Years

```
seasonal = df.groupby('Season')['AQI'].mean()
plt.figure(figsize=(2,4))
seasonal.plot(kind='bar')
plt.title("Seasonal AQI Levels")
plt.ylabel("Average AQI")
plt.show()
```

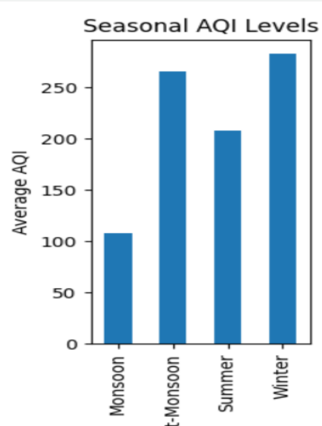


Figure 3:Column chart Showing Seasonal AQI levels

```
[ ]
pollutants = ['PM10', 'PM2.5', 'NO2', 'SO2', 'Ozone', 'CO']
pollutant_avg = df[pollutants].mean()
plt.figure(figsize=(3,4))
pollutant_avg.plot(kind='bar')
plt.title("Average Pollutant Levels")
plt.xlabel("Pollutants")
plt.ylabel("Average Value")
plt.show()
```

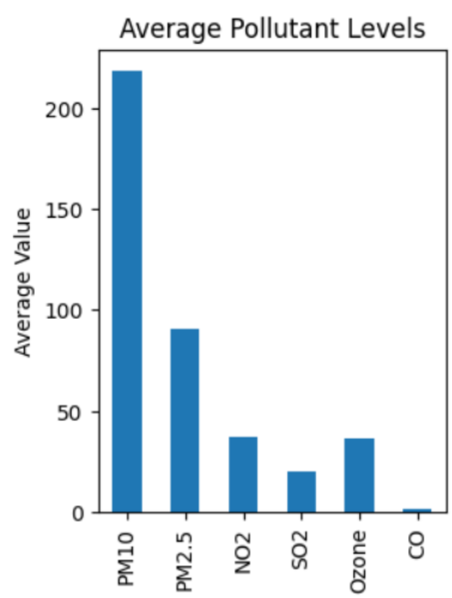


Figure 4 : Column chart showing Most Pollutant items

```
[ ]
weekday_avg = df.groupby('Weekday')['AQI'].mean().sort_values(ascending=False)
print(weekday_avg)
```

```
Weekday
Tue    204.440191
Fri    203.870813
Wed    203.504808
Thu    203.250000
Sun    201.665072
Sat    200.593301
Mon    198.162679
Name: AQI, dtype: float64
```

Figure 5: Most pollutant affect weekdays

[47] ✓ 0s

```
heatmap_data = df.pivot_table(
    values='AQI',
    index='Weekday',
    columns='Month',
    aggfunc='mean'
)
```

[48] ✓ 0s

```
# full avg per weekday
weekday_avg = df.groupby('Weekday')['AQI'].mean()

# add as a new column
heatmap_data['Total_Avg'] = weekday_avg
```

[49] ✓ 0s

```
order = ['Tue', 'Fri', 'Wed', 'Thu', 'Sun', 'Sat', 'Mon']
heatmap_data = heatmap_data.reindex(order)
```

[50] ✓ 1s

```
plt.figure(figsize=(14,6))
sns.heatmap(heatmap_data, annot=True, fmt=".1f", cmap="YlOrRd")
plt.title("Average AQI Heatmap (Weekday vs Month + Total Average)")
plt.xlabel("Month")
plt.ylabel("Weekday")
plt.show()
```

...

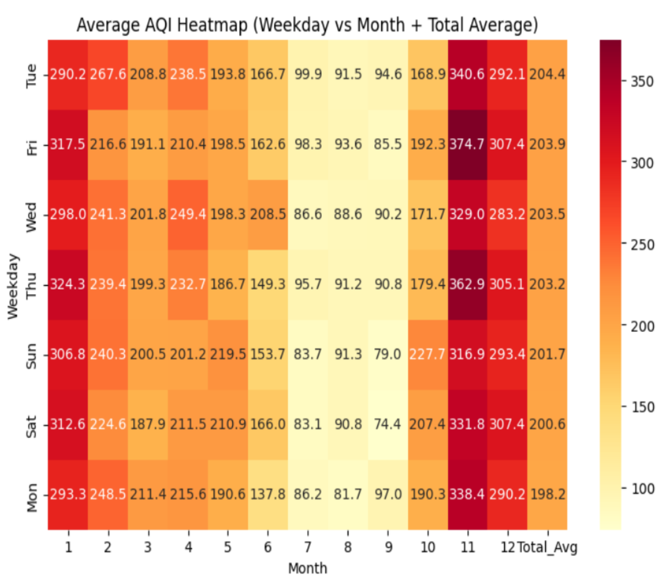


Figure 6 : Heatmap showing Avg AQI weekday at month

PYTHON-BASED DATA VISUALIZATION AND INSIGHTS

AQI Trend Over Years

- The yearly AQI trend visualization shows how air quality changes over different years, helping identify improvement or deterioration. It highlights long-term pollution Behavior influenced by seasonal and environmental factors.

Seasonal AQI Levels

- The seasonal AQI analysis compares pollution across various seasons and reveals clear seasonal impact on air quality. Winter shows the highest AQI, while monsoon demonstrates significant improvement due to rainfall.

Average Pollutant Levels

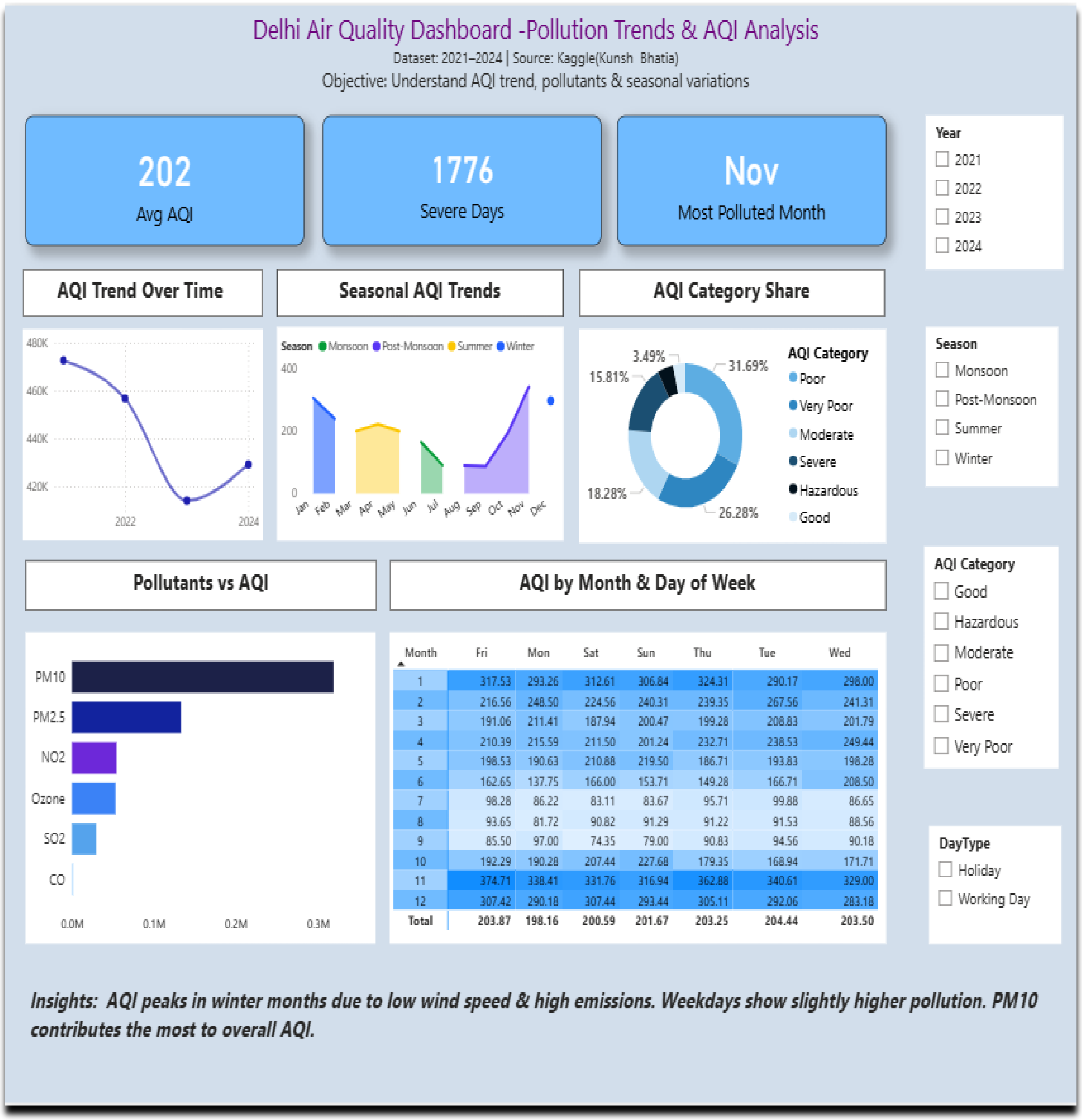
- This visualization compares the average concentration of key pollutants to determine which contribute most to poor air quality. PM10 and PM2.5 emerge as the dominant pollutants requiring priority attention.

Weekday–Month AQI Heatmap

- The heatmap displays AQI variation across weekdays and months, showing how pollution fluctuates with human activity and weather patterns. It identifies high-AQI combinations, particularly during winter and busy weekdays

VISUALIZATION GENERATED USING POWER BI

Power BI visualizations were developed to validate and compare the findings obtained through SQL, Python libraries analysis.



Dashboard Overview

The dashboard contains:

- Donut chart for AQI Category distribution
- Line/Wave chart for seasonal patterns
- Column chart for pollutant comparison
- Heatmap for Month vs Weekday AQI

Data Cleaning Steps

- Handled missing values
- Converted data types (date, numeric)
- Added AQI Category column, Season, Week Days , Day type using DAX
- Unpivoted pollutant columns for analysis

Data Model

- A single-table model was used with calculated columns & DAX measures.
- Relationships were not required as the dataset was flat.

POWER BI DASHBOARD INSIGHTS & FINDINGS

1. PM10 & PM2.5 are the highest contributors to poor AQI

These two pollutants show consistently elevated values, indicating they are the primary drivers of degraded air quality in Delhi.

2. AQI is significantly worse during winter months

Pollution levels spike between November–January due to weather conditions, low wind speed, and seasonal activities (e.g., crop burning, festivals).

3. Most days fall under the "Moderate" AQI category

The dominant AQI range is Moderate, indicating persistent but manageable pollution with occasional spikes.

4. Seasonal trend suggests clear variation in pollutant behavior

Winter: High particulate matter

Summer: Higher ozone

Monsoon: AQI improves due to rainfall

5. NO_2 levels remain consistently high in traffic-heavy periods

Suggests vehicular emissions are a major contributor.

6. Sudden AQI spikes correlate with festival periods (Diwali)

Short bursts of severe pollution are visible around festival dates.

CONCLUSION

Overall, the analysis confirms that particulate pollutants, especially PM10 and PM2.5, play a dominant role in Delhi’s air quality deterioration. Seasonal and weekday patterns reveal predictable pollution spikes, especially during winter and busy workdays. The combined SQL, Power BI, and Python methods provide a strong foundation for data-driven environmental decision-making.

REFERENCES

- Bhatia, K. (2024). *Delhi Air Quality Dataset*. Kaggle.
- Central Pollution Control Board (CPCB). *National Air Quality Standards*.
- Python Software Foundation. *Pandas Documentation*.
- Microsoft. *Power BI Documentation*.