Journal of Informatics and Web Engineering

Vol. 5 No. 1 (February 2026) eISSN: 2821-370X

Efficiency of Neo4j in Designing and Analysing Graph Database for Air Quality Analysis of Indian Metro Cities

Jiten Chavda[1*], Kishan Bharvad[2], Rishap Parmar[3], Nidhi Arora[4]

[1,2*]Department of Computer Science, Gujarat University, Navarangpura Ahmedabad-380009, Gujarat, India

[3]School of Emerging Science and Technology (Department of AIML), Gujarat University, Navarangpura Ahmedabad-380009, Gujarat, India

[4]Advicon Solutions Pvt Ltd, 1201, The Capital, 2, Science City Rd, near Shell Petrol Pump, Sola, Ahmedabad, Gujarat 380060, India

**corresponding author: (chavdajiten00@gmail.com; ORCiD: 0009-0005-6338-873X)*

Abstract - The analysis of the Air Quality Index (AQI) is currently a popular subject in the area of sustainable research, as it is crucial for investigating and analysing the effects of air pollutants on human health in urban environments. It has been identified over the last decade that airborne pollution has become a critical issue and will remain an important concern in India in the coming years. In recent years, a variety of models and algorithms utilizing big data techniques have been developed for the analysis of air quality data. In this paper, we suggest monitoring and feature analysis of air quality data using a graph database. The research aims to analyse the annual and seasonal variations of AQI over a 10-year period between 2015–2024 from daily averaged concentration data of key air pollutants for 5 metro cities of India. The trends shown by all the cities have been compared to understand the seasonal variations in the average Air quality index. The variations of Average AQI in different severity classes in the cities also provide in-depth analysis of the trends. The findings from this analysis yield highly valuable information to assist in air pollution control, consequently leading to substantial societal and technical impacts. Finally, we offer a perspective on the future of air quality analysis, presenting some promising and challenging concepts. The results of this study can promote a more effectual environment monitoring system to detect drastic or unusual changes in atmosphere through the use of modern technologies.

Keywords—Air Quality Index, Air Quality Analysis, Air Quality Monitoring, Data Analysis Neo4j, Graph Database, Indian Cities.

Received: 30 May 2025; Accepted: 27 July 2025; Published: 16 February 2026

1. INTRODUCTION

Air quality analysis has matured highly in the last decade. In the view of the speedy economic growth and a lot of incidents around air pollution are noticed in several prominent and densely populated cities across India. Due to this

MULTIMEDIA UNIVERSITY

Journal of Informatics and Web Engineering
https://doi.org/10.33093/jiwe.2026.5.1.2
© Universiti Telekom Sdn Bhd.
Published by MMU Press. URL: https://journals.mmupress.com/jiwe

reason, air pollution has become a pressing issue in recent years. Major contributors to air pollution include the transportation industry, industrial expansion, and the burning of fossil fuels, including coal and gas. Emissions of greenhouse gases lead to global warming, which in turn drives climate change. Developing countries like India are facing a lot of challenges related to air pollution and its bad impacts not only on public health, but also on the environment.

Air pollutants are largely found in the form of particulate matter and gaseous pollutants which have different levels and impacts on the environment. Some of the adverse effects are formed by meteorological changes like rainfall intensity, relative humidity, solar radiation while some others are because of wind direction, wind speed, and temperature [1]. These pollutants are responsible in determining the Air Quality Index (AQI) for a particular region or city. These air pollutants not only pollute air nut also bring several side effects like depletion of ozone layer, global warming, heating up of Earth's surface by rising average temperature, climate change, and acid rain [2], [3].

In both developing and developed countries, the challenges held by air pollution have given rise to more public consciousness regarding the quality of air they breathe in. Air pollution has significantly harmed the development of many nations affecting the physical and mental health of individuals. Hence, many sectors have moved to pay more attention to the AQI, which is utilized to assess air quality. Monitoring and forecasting of air pollution is also taken on priority by government and other regulatory agencies in view of its ill-effects on public health. A wide variety of air pollutants that affect human health are Carbon Monoxide (CO), Sulphur Dioxide (SO2), Nitrogen Dioxide (NO2), Ozone (O3) and Particulate Matter like Respirable Suspended Particulate Matter (RSPM) and Suspended Particulate Matter (SPM). If any of these pollutants are found in higher concentration in air, it can be dangerous to life giving rise to many health problems ranging from respiratory issues to headaches, and dizziness and in extreme cases it can even result in heart attacks [4]. So, it is important for individuals to understand the adverse effects of poor air quality on their health. The AQI helps define a rating scale or index for reporting the daily cumulative effects of air pollutants recorded from sites under monitoring.

The AQI is considered to be a standard indicator to express the degree of pollution in air and its possible implications for public health. It presents a comprehensive value by combining several pollutants providing a clear understanding of air quality conditions in a given period. In India, AQI is defined by the Central Pollution Control Board (CPCB) [5] which considers significant pollutants like PM2.5, PM10, NO2, SO2, CO, O3, NH3 and BTX. The resulting AQI is measured on a scale from 0 to 500, with higher values reflecting poorer air quality and heightened health risks. The AQI in India is divided into following categories:

- **Good (0–50)**: Minimal impact.
- **Satisfactory (51–100)**: Minor breathing discomfort to sensitive people.
- **Moderate (101–200)**: Breathing discomfort to people with lung/heart diseases.
- **Poor (201–300)**: Breathing discomfort to most people on prolonged exposure.
- **Very Poor (301–400)**: Respiratory illness on prolonged exposure.
- **Severe (401–500)**: Serious health effects, even on healthy people.

The AQI is derived from the sub-index values of individual pollutants, where the overall AQI represents the highest sub-index among all pollutants, provided that at least three pollutants are evaluated (including either PM2.5 or PM10). Each pollutant's sub-index is computed using a piecewise linear interpolation formula that takes into account its concentration and specific breakpoints. The formula for the sub-index (I_p) of a pollutant P is in Equation (1).

$$I_p = \left[\frac{(I_{Hi}-I_{Lo})}{(C_{Hi}-C_{Lo})} \times \left(C_p - C_{L0}\right)\right] + I_{Lo} \quad (1)$$

Where,

I_p: Sub-index for pollutant
C_p: Measured concentration of pollutant P
I_{Hi}: AQI value corresponding to the upper breakpoint

I_{Lo}: AQI value corresponding to the lower breakpoint
C_{Hi}: Concentration breakpoint ≥ C_p
C_{Lo}: Concentration breakpoint ≤ C_p

The overall AQI is calculated as in Equation (2).

$$AQI = \max\left(I_{PM2.5}, I_{PM10}, I_{NO_2}, I_{SO_2}, I_{CO}, I_{O_3}, I_{NH_3}, I_{Pb}\right) \tag{2}$$

According to the regulations by the environmental protection authorities, an AQI value between 0 to 50 corresponds to an AQI level of one, indicating excellent air quality. An AQI value ranging from 51 to 100 is classified as AQI level of two, which reflects good air quality. If the AQI value falls between 101 and 150, then it is in AQI level of three, indicating mild pollution. An AQI value between the range 151 to 200 is categorized as AQI level four, signifying medium pollution. Lastly, an AQI value found within the range of 201 to 300 corresponds to an AQI level of five, which denotes heavy pollution. Values above 300 results in an AQI level of six, indicating very serious pollution. The growing severity of air pollution is attracting increasing attention from researchers and regulatory authorities, particularly in relation to public health and the prevention of pollution incidents. These initiatives have caused the collection of a good amount of historical data on air quality monitoring. The large-scale drawbacks of air pollution open a lot of options for researchers to study air quality as a subject of study to understand its challenges and bring a variety of solutions to effectively manage air pollution.

The AQI established by the CPCB for India is determined by the levels of eight primary air pollutants that have a considerable impact on air quality and public health in urban settings [5]. These pollutants are aggregated to produce a singular value known as the AQI. The specifics of the pollutants are as follows:

i. Particulate Matter (PM2.5): These are fine particles measuring ≤ 2.5 micrometres in diameter, capable of penetrating deeply into the lungs and bloodstream, leading to respiratory and cardiovascular complications.

ii. Particulate Matter (PM10): These are larger particles measuring ≤ 10 micrometres, which can adversely affect respiratory health and contribute to haze formation.

iii. NO2: This gas is produced from vehicle emissions and industrial operations and is associated with respiratory issues.

iv. SO2: Emitted from industrial activities and power generation, this gas contributes to acid rain and can irritate the respiratory system.

v. CO: This is a colourless and odourless gas resulting from incomplete combustion (such as from vehicles and biomass burning), which hampers oxygen transport in the body.

vi. O3: This secondary pollutant is generated through chemical reactions in the presence of sunlight and can impair lung function at elevated concentrations.

vii. Ammonia (NH3): Released from agricultural practices, waste management, and certain industrial processes, NH3 plays a role in the formation of particulates.

Benzene, Toluene, and Xylene (BTS) are classified as Volatile Organic Compounds (VOCs), collectively referred to as BTX that are significant air pollutants due to their toxicity and their involvement in atmospheric chemistry.

In the context of AQI analysis, sustainability refers to the ability to preserve and enhance air quality in coming time, which is vital for promoting environmental health, human well-being, and economic stability, to match the principles of sustainable development aiming to balance all the three aspects of sustainability i.e. environmental, social, and economic aspirations. Figure 1 depicts the sustainability concerns of AQI analysis, illustrating the three interconnected dimensions environmental, social, and economic sustainability that must be balanced for long-term development.

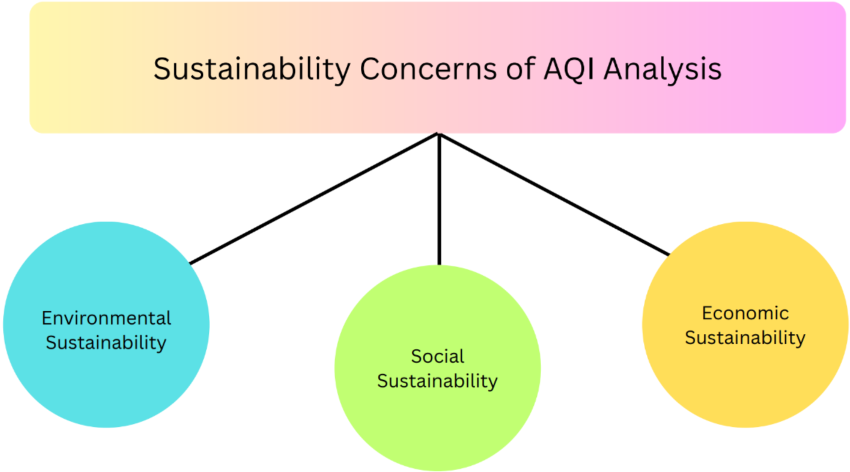

Sustainability Concerns of AQI Analysis
Environmental Sustainability
Social Sustainability
Economic Sustainability

Figure 1. Sustainability Concerns of AQI Analysis

AQI has been an important topic of discussion for many due to its sustainability concerns. The analysis of AQI, particularly when enhanced through the use of graph databases, aids in sustainability in many ways. As far as Environmental Sustainability is concerned, AQI analysis can track the key pollutants. As poor air quality harms ecosystems, AQI analysis helps prioritize pollution control, protecting biodiversity and natural resources in and around metro cities. Social Sustainability is connected to public Health Protection where AQI analysis can inform communities about health risks associated with poor air quality to raise awareness, promote sustainable lifestyles and support green policies. On Economic Sustainability, AQI analysis can guide cost-effective opportunities towards sustainable city planning by saving a lot of economic resources and also improving economic resilience by attracting investment to cleaner cities. The value of AQI analysis is evident in its role as a fundamental tool for understanding, communicating about, and addressing air pollution in India's metropolitan cities. Here are the key reasons that illustrate the importance of AQI analysis:

i. Visualizing AQI data through graphs facilitates the effective communication of urgent air quality updates to the public.

ii. The AQI simplifies intricate pollutant data into a single, comprehensible value (ranging from 0 to 500) with classifications that span from Good to Severe. This enables residents to take essential precautions, such as refraining from outdoor activities on days when the AQI is Poor.

iii. Data from AQI analysis can aid policy-making bodies and regulators in tracking compliance with National Air Quality Standards, ensuring that industries and cities minimize emissions.

iv. Insights gained from AQI trends can prioritize investments in public transportation or cycling infrastructure to mitigate NO2 and CO levels.

AQI analysis highlights pollution hotspots, allowing authorities and communities to advocate for equitable improvements in air quality.

2. LITERATURE REVIEW

Air quality analysis is found in literature with lot of significant advances as far as monitoring infrastructure, predictive modelling, and the interpretation of air quality is concerned. This is also because of the expansion of air quality monitoring networks and their wide documentation. Studies from India shows that the number of cities reporting AQI have seen significant increase from 22 to 271, with a greater number of monitoring stations growing from 31 to 469 between years 2015 and 2023. This indicates a substantial improvement in spatial coverage and data availability [6], [7], [8]. Although more data has been collected from these monitoring stations, challenges persist in ensuring monitoring specifically in regions with a wide variety of pollution sources. The standardization of AQI and development of methodologies around AQI measurement and analysis play a crucial role in public health communication and policy formulation. In India, a national AQI system in 2018 was adopted which allowed for daily bulletins and more effective health advisories to public [9], [10], [11]. The AQI framework has also raised awareness and guided regulatory actions by categorizing air quality based on concentrations of all constituent pollutants. Research has shown that programs like the National Clean Air Programme (NCAP) have led to significant improvements in air quality, with some cities even reporting reductions in PM levels by more than 50%. [12], [13], [14].

The existing studies on air quality analysis primarily emphasize predictive behaviour over descriptive behaviour of air pollutants. Lot of different models have been utilized for predicting air pollutants. Literature shows numerous statistical models facilitating the prediction of pollutant concentrations [15], [16], [17]. Other studies indicate researchers' interest in Gaussian dispersion models for air quality predictions in many air pollution studies. Although dispersion models possess some physical foundations, detailed knowledge about the sources of pollution and other relevant parameters is still insufficient [18]. Ameer et al. proposed a real-time air pollution monitoring model by integrating Internet of Things (IoT) sensors along with machine learning algorithms [19]. Their model utilizes multiple regression techniques to predict air pollution. Similarly, Ojagh developed an IoT-based air pollution monitoring system to evaluate the air quality in some part of Canada. The system used an edge and cloud-based mixed prediction model, as opposed to conventional methods [20].

Predictive analysis has become a crucial area of innovation in air quality research. Traditional statistical approaches like multiple linear regression are supported by advanced machine learning models and artificial intelligence techniques. Recent studies have given emphasis on the effectiveness of complex models such as random forests, neural networks, and Large Language Models (LLMs) in forecasting AQI for better accuracy. Research by Zhu et al. (2021) demonstrates that random forests and deep neural networks models significantly perform better than traditional regression models in predicting PM2.5 concentrations [21]. Similarly, the use of LLMs in air quality prediction has shown outstanding accuracy with some of the studies showing R^2 values reaching as high as 0.99 with low mean squared errors. The integration of predictive models in real-time monitoring system has made it possible to obtain more timely and accurate air quality forecasts for supporting better public health and taking the right policy decisions. There are some adaptive monitoring systems also which combine meteorological data with the concentration of pollutants to improve spatial mapping and reduce power consumption [22]. These advancements have been especially advantageous for those regions where the monitoring infrastructure is limited.

As per the study by Prakash [23], India ranks as the world’s leading country for emitting Sulfur Dioxide. This is primarily because of coal-fired power generation and its usage in industrial activities. The same applies for NO2 levels which are seen to continuously rise. These pollutants cause acid rain and at the same time influence atmospheric composition. They reduce both the quantity and quality of sunlight reaching the Earth’s surface which makes it inadequate for living [24]. Also, the vehicular emissions in the form of CO, Nitrogen Oxides, Hydrocarbons, Sulfur Compounds, Lead, and SPM also degrade the quality of air making it inappropriate for breathing generating serious risks to both human health and ecosystems [25]. The combined effect of air pollution, climatic change, and emerging geoengineering interventions causing altered solar radiation patterns affecting crop development and yield in unpredictable ways.

Air pollution is considered to one of the bigger challenges looking from the viewpoints of both; environmental sustainability and the wellness of living beings especially in urban areas [26]. This is because cities experience a variety of pollution including vehicle emissions, industrial waste etc. leading to accumulation of harmful pollutants [27]. A major concern specifically points to PM2.5, a fine particulate matter whose tiny structure makes it easy to enter the circulatory system through lung tissues, heightening the risk of respiratory and cardiovascular diseases [28] [29]. These challenges not only require innovative tools, but also real-time monitoring technologies supported by predictive models built on machine learning techniques to address the issue. This can help in taking data-driven decisions and improving urban air quality management [30]. Hence, the past decade has seen good progress in terms of air quality analysis, due to the expansion of monitoring networks, the standardization of AQI frameworks, and the adoption of advanced predictive models. However, still there exist a lot of challenges in ensuring comprehensive coverage for air quality monitoring and bringing improvements in polluted areas. It is required by the research community to invest more time in monitoring infrastructure and developing more sophisticated models for addressing these challenges for improving air quality management in place.

3. RESEARCH METHODOLOGY

3.1 Graph Databases

Graph databases are a special type of NoSQL knowledge databases useful for storing and organizing data in the form of collection of nodes and edges represented by a graph. In contrast to the traditional relational databases (RDBMS), which use tables, rows, and columns, the graph databases are focused on interconnection of data points for efficiently querying complex relationships. Graph databases are more appropriate in scenarios where relationships among the data points are as important as the data itself. There are several common applications to use a graph database, some of which can be, in social networks where modelling relationships like friends, followers, or interactions is meaningful or in geospatial models to store relationships between routes or proximity in navigation systems for two or more stations. Some of the benefits of graph databases over the traditional RDBMS are shown in Table 1.

Table 1: Benefits of Graph Databases Over Traditional RDBMS

Aspect	Graph Databases	Traditional RDBMS
Efficient Relationship Queries	Optimized for traversing relationships; fast and intuitive queries like "friends of friends" even on large datasets.	Requires complex JOINs; slower and resource-intensive for deeply nested relationships.
Flexible Schema	Schema-less or flexible; easy to add new nodes, relationships, or properties.	Rigid schema; predefined tables/columns make changes time-consuming and costly.
Intuitive Data Modelling	Naturally represents real-world relationships (e.g., social networks, supply chains).	Forces relationships into tabular forms, less intuitive for complex data.
Scalability for Connected Data	Handles interconnected data efficiently; traversal performance remains stable as data grows.	Performance degrades with increased JOIN complexity or data volume.
Real-Time Insights	Enables real-time relationship analysis; ideal for fraud detection, recommendations, etc.	Slower for relationship-heavy queries; may require batch processing or heavy indexing.
Simplified Querying	Uses Cypher, Gremlin, etc.; designed for easy and direct graph traversal.	SQL is powerful but can be verbose and complex for multi-relationship queries.
Handling Sparse Data	Efficiently manages heterogeneous/sparse data; supports varying properties per node or edge.	Sparse data leads to inefficiencies (e.g., NULLs) and harder-to-manage queries.

3.2 Neo4j Graph Database

Neo4j stands out as a leading graph database platform crafted to store, manage, and query data that is highly interconnected, utilizing a graph structure consisting of nodes, edges, and properties. It is particularly adept at managing complex relationships, making it an excellent choice for applications like social networks, fraud detection, recommendation systems, and knowledge graphs. Figure 2 shows a comparative view of popular graph database platforms, where Neo4j is recognized as one of the most widely adopted solutions among others such as TigerGraph, Amazon Neptune, ArangoDB, and JanusGraph. The platform uses Cypher query language for intuitive and efficient navigation of data within the graph structure and for providing real-time insights especially in large datasets which otherwise becomes challenging. Its schema-flexible design includes dynamic data models, and it comes with robust features like Atomicity, Consistency, Isolation, Durability (ACID) compliance, high availability, and scalability by means of clustering. Neo4j is a choice across various industries for offering enterprise-grade support. It also allows integration with several third-party tools like GraphQL and Apache Spark, and with all major cloud platforms such as AWS, Azure, and Google Cloud.

Neo4j
Amazon Neptune
Dgraph
AllegroGraph
Stardog
ArangoDB
JanusGraph
RedisGraph
TigerGraph
OrientDB
Memgraph
GraphDB

Figure 2. Some Well-Known Graph Databases

NeoDash is an open-source, web-based dashboard builder designed especially for Neo4j. It helps in interactively exploring and visualizing graph data. It allows creating customized dashboards to display charts, tables, maps, and graph visualizations without requiring any knowledge of extensive coding. NeoDash connects to a Neo4j graph database, enabling developers to write Cypher queries to create shareable visualizations for real-time data exploration. These visualizations are useful for business analysts, data scientists, and developers for conveying data insights to non-technical users. NeoDash also has features like parameterized queries, drill-down options in charts, and exportable reports. These features make NeoDash a powerful tool for transforming raw Neo4j data into visual narratives for taking actions by stakeholders.

3.3 AQI Analysis

Although there are a variety of traditional and graph databases as pointed out in the last section, some of the promising reasons for using Neo4j in this study are because of its features over other databases. Neo4j offers an effective approach compared to MongoDB and Dgraph for handling highly connected data. It supports the property graph model which is known for its rich set of features available through the Graph Data Science (GDS) library. It can also be integrated seamlessly with tools based on machine learning and data analytics. MongoDB is known as a powerful NoSQL database with a flexible document structure, but it does not support graph operations, making it less effective when deep relationship analysis is required. On the other hand, Dgraph allows for a graph-based design with GraphQL, but it has limited built-in tools for machine learning and data visualization. Neo4j is favourite choice of graph databases among researchers and developers due to its intuitive interface, strong community support and effective analytics environment for diverse use cases including but not limited to fraud analysis, descriptive and predictive modelling, and real-time graph analytics.

The dataset contains air quality data measured across five major Indian cities viz. Delhi, Mumbai, Chennai, Kolkata, and Bangalore observed from multiple monitoring stations for the period of January 1, 2015, to December 31, 2024. The data comprises values of pollutants like NO, NO2, NOx, NH3, CO, SO2, O3, and PM2.5, PM10 with BTX measured on a daily basis. The AQI is calculated using the formula stated above, and each record is classified into its corresponding AQI Bucket. Each record is identified by a combination of city, date of observation, and station under observation for analysing air quality trends. The dataset contains 36,531 records and is largely well-structured requiring minimal data preprocessing with replacing the missing values with their mode in the respective columns.

In order to utilize the relational abilities of Neo4j, the dataset is transformed into a graph database. The following steps are used for creating the graph database:

- **Schema Design**: The graph schema is designed to model relationships between entities. Various nodes are created to represent parameters like city, station, date, and its associated readings. To connect cities to their monitoring stations and link stations to their air quality reading, appropriate relationships are defined. Properties of the reading nodes are shown in the form of levels of pollutants, their AQI value, and the corresponding AQI bucket.
- **Data Ingestion**: The Neo4j database create nodes for each unique city, station, and reading, and establish relationships based on the data. Cypher queries are used to batch-insert data efficiently and to ensure future scalability.
- **Index Creation**: Indexes are created on frequently asked queries for properties to optimize the query performance.
- **Validation**: The graph structure is validated by running sample queries to validate the right establishment of all nodes and relationships to accurately reflect the data values.

The graph database's structure is visually represented using Neo4j's visualization tools. Each city node in Figure 3 is linked to multiple station nodes, which in turn connect to numerous reading nodes which store pollutant levels and AQI as properties, enabling efficient querying of air quality trends.

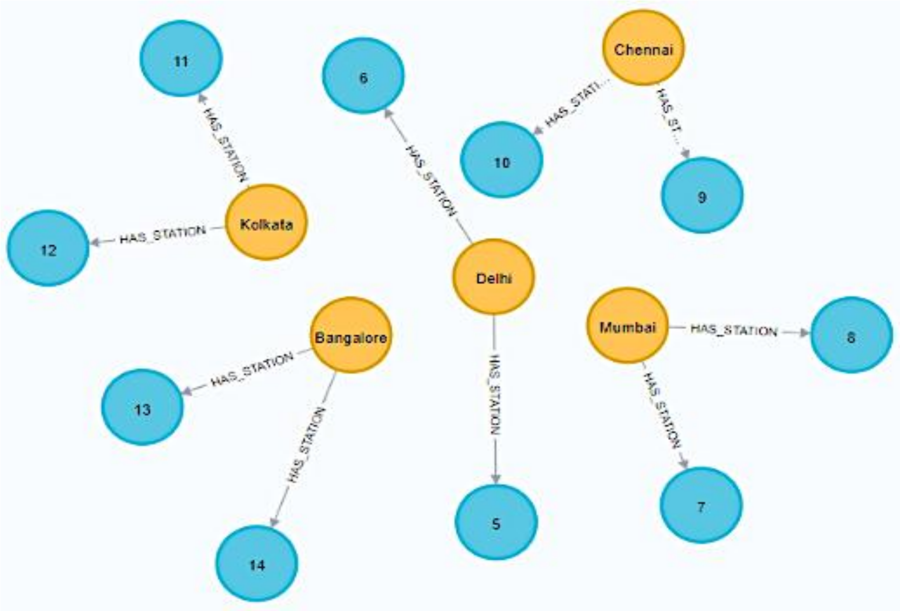


Figure 3. Graph Data Structure Representation in Neo4j

Figure 4 illustrates the schema, depicting nodes for City, Station, and Reading, connected by directed relationships. These visualizations are generated using Neo4j Browser, which provides an intuitive interface for exploring the graph structure.

Cypher queries are written to perform data analysis on the dataset. As the charts generated by Neo4j are not attractive from the viewpoint of users, the same is also done using Google Looker Studio for clear and interactive data representation. The data analysis enables dynamic trend analysis and effective comparison across cities and time.

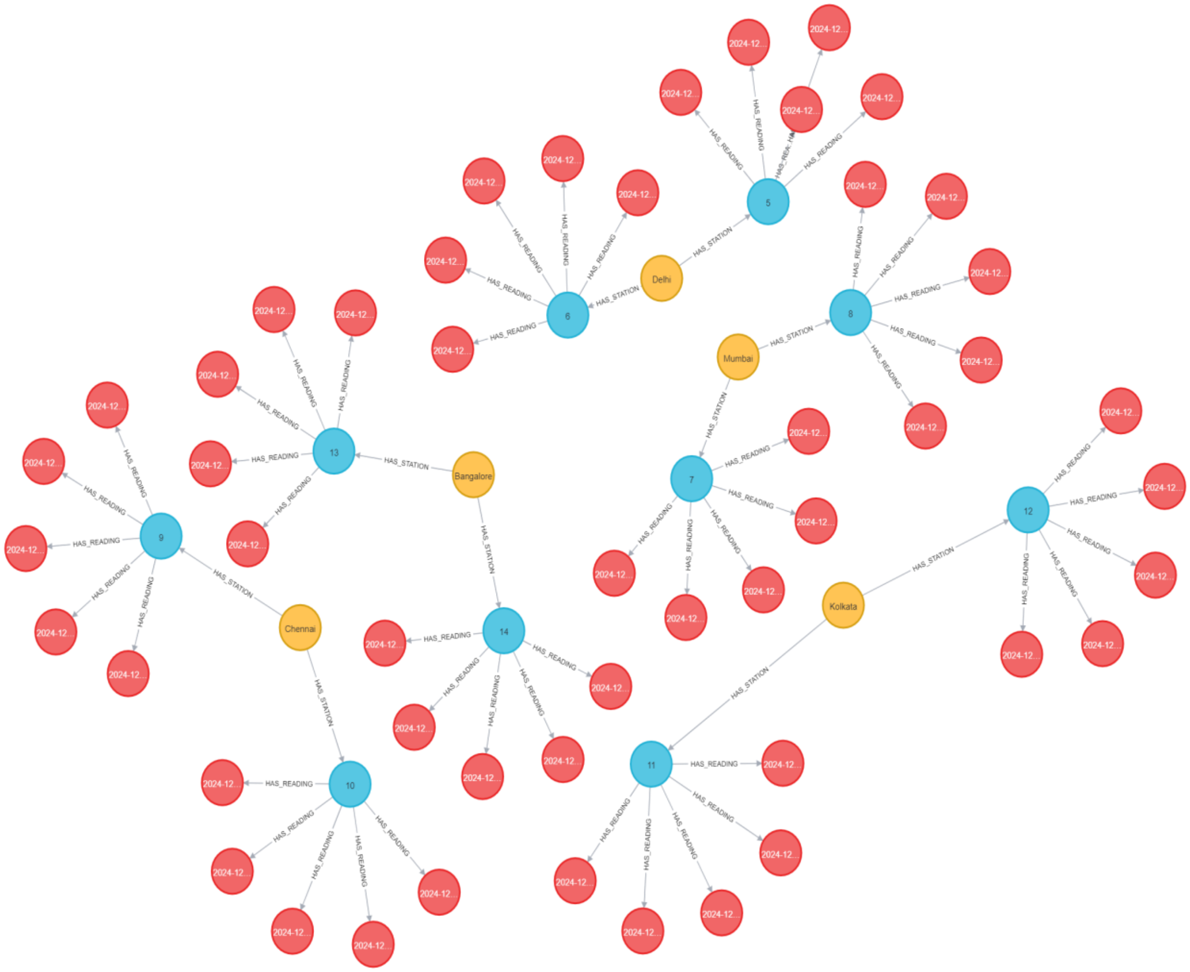
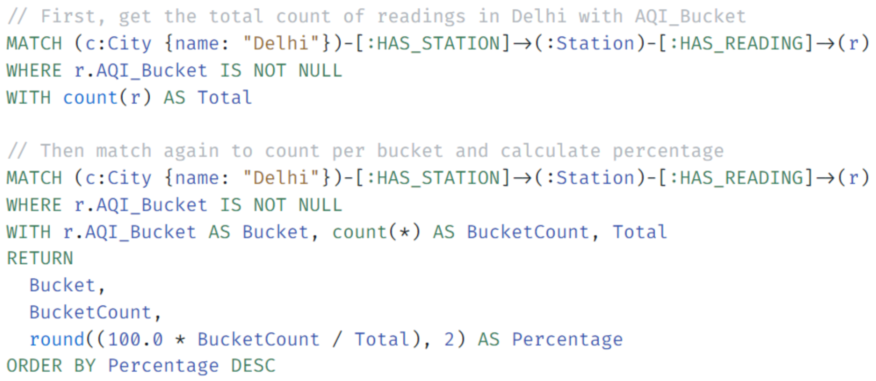


Figure 4. Detailed Graph Data Structure Representation in Neo4j

The composite visual as shown in Figure 5(a) and Figure 5(b) compares the percentage share of AQI in various categories like Good, Satisfactory, Moderate, Poor, Very Poor, and Severe across five major Indian cities. As seen clearly, Delhi exhibits the highest concentration of unhealthy air, with a significant share falling under the Severe, Very Poor, and Poor categories. Kolkata shows slightly better air quality but remains concerning. Mumbai presents a more balanced distribution, though poor air persists. Bangalore performs marginally better than other metros yet still reflect unhealthy levels. Chennai displays a relatively even spread across AQI categories, with notable presence in the Very Poor and Severe segments.



```
// First, get the total count of readings in Delhi with AQI_Bucket
MATCH (c:City {name: "Delhi"})-[:HAS_STATION]->(:Station)-[:HAS_READING]->(r)
WHERE r.AQI_Bucket IS NOT NULL
WITH count(r) AS Total

// Then match again to count per bucket and calculate percentage
MATCH (c:City {name: "Delhi"})-[:HAS_STATION]->(:Station)-[:HAS_READING]->(r)
WHERE r.AQI_Bucket IS NOT NULL
WITH r.AQI_Bucket AS Bucket, count(*) AS BucketCount, Total
RETURN
  Bucket,
  BucketCount,
  round((100.0 * BucketCount / Total), 2) AS Percentage
ORDER BY Percentage DESC
```

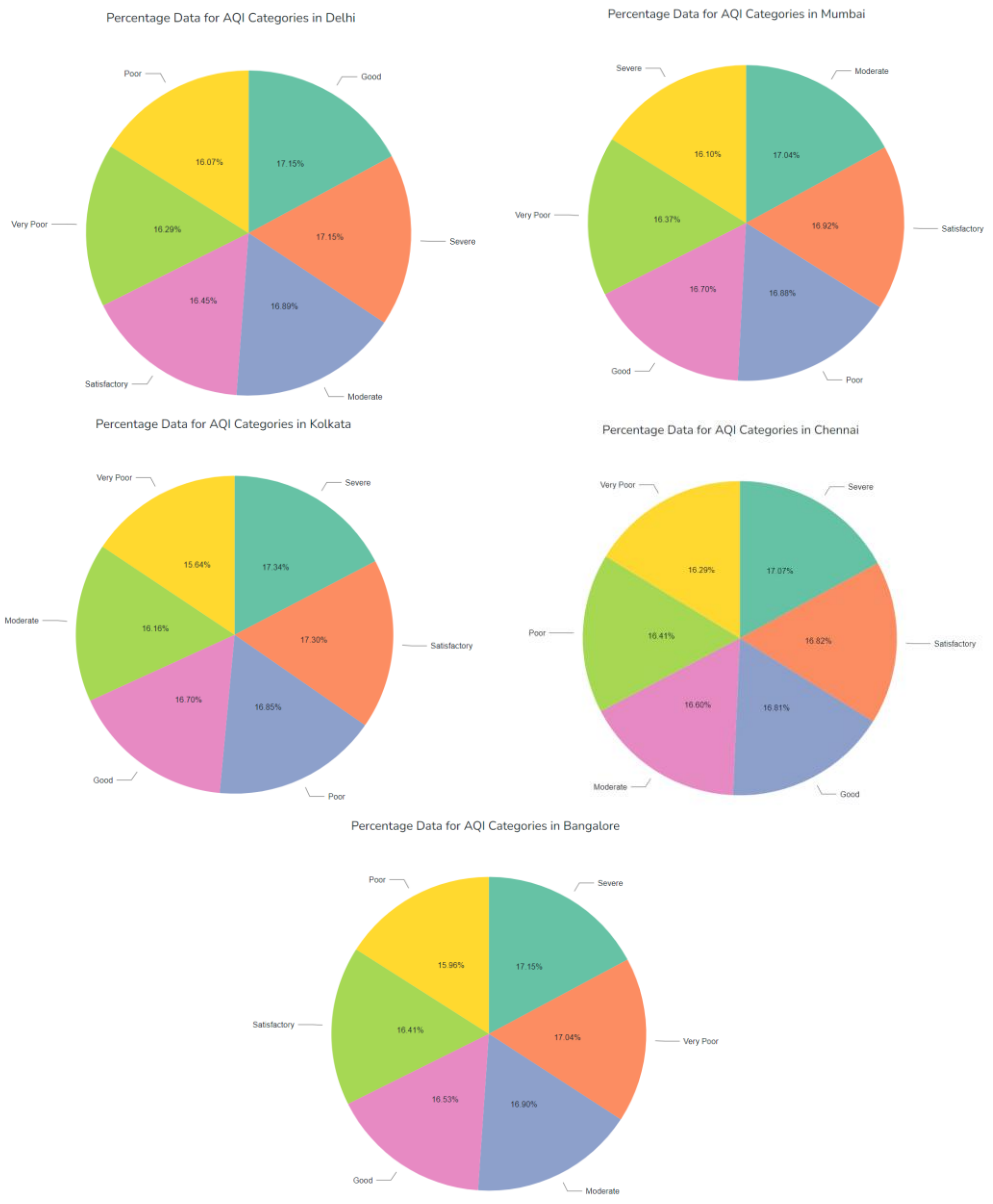


Figure 5(a). Queries and Distribution of AQI Categories Across Metro Cities Neo4j

Figure 5(a) and Figure 5(b) show the condition of all the cities averaging over various AQI categories. This figure also shows Delhi as the most polluted city, consistently dominated by the Very Poor and Severe buckets, while other cities show mixed AQI average. Policies can be enforced to reduce emissions from vehicles and industrial areas in Delhi, especially during high-risk periods.

A further comparative analysis is done for all the cities on average AQI in year 2024. Figure 6 shows that Delhi leads with the highest average AQI, while Chennai records the lowest, suggesting relatively better air quality among all the cities. The government must urgently implement both short-term like banning firecrackers during festivals and long-term strategies in Delhi such as mass transit improvements and industrial zoning reforms.

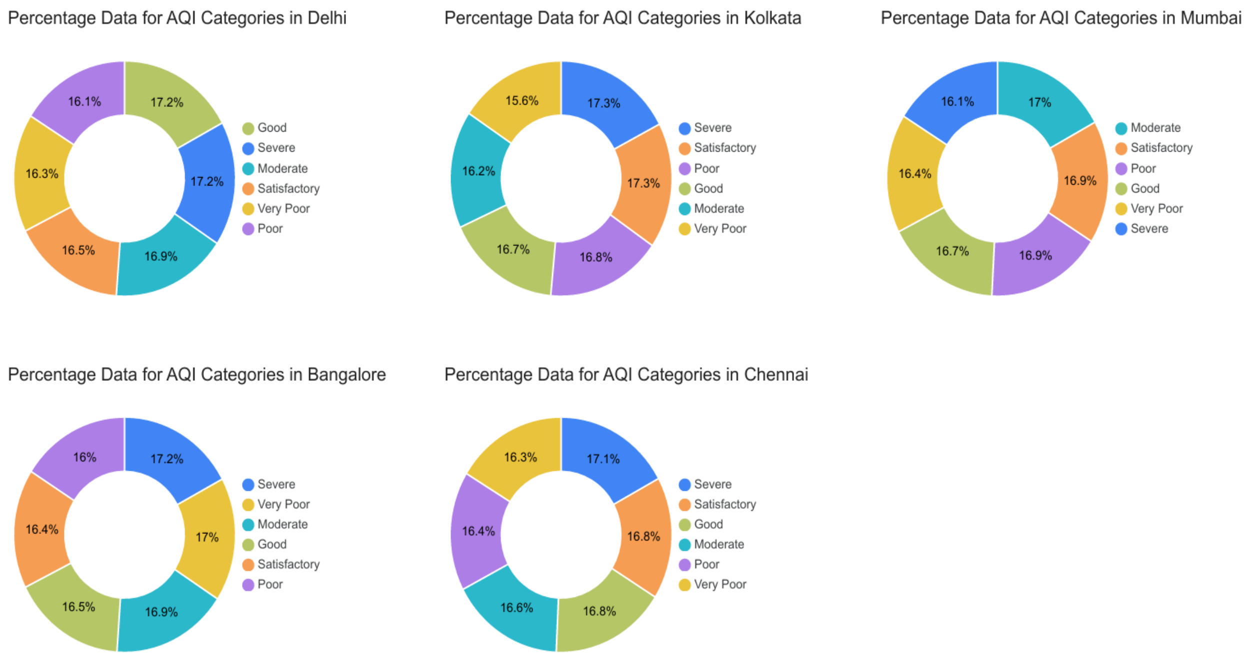


Figure 5(b). Distribution of AQI Categories Across Metro Cities (Looker)

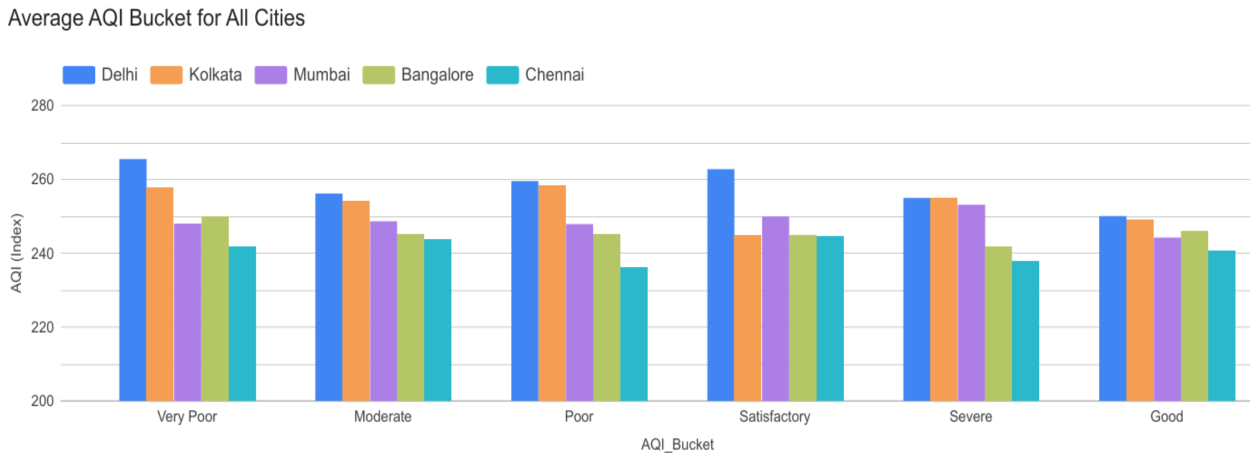
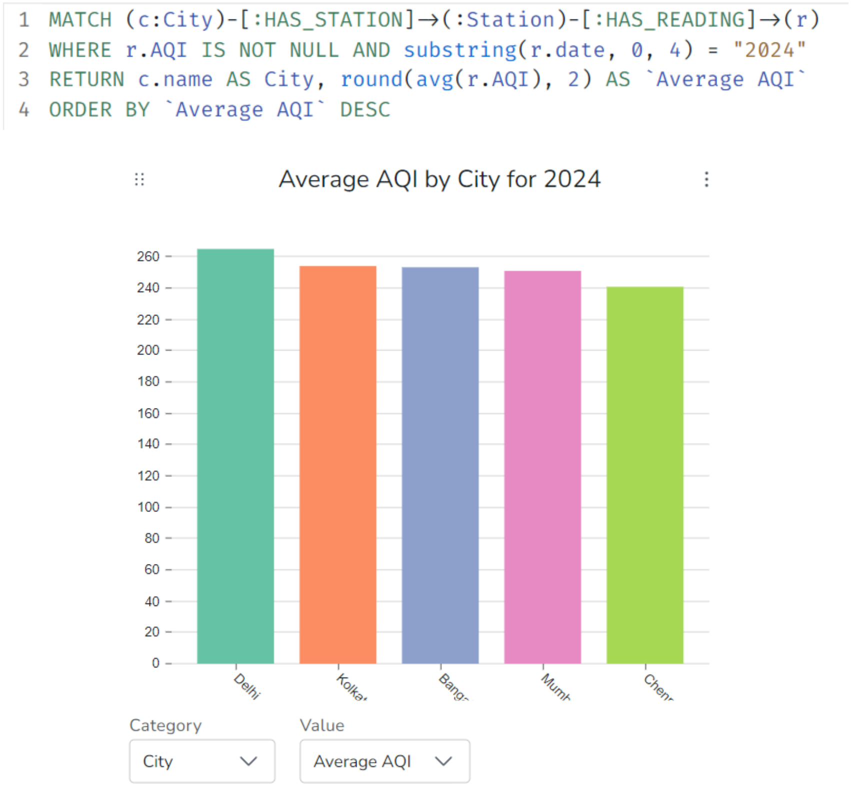


Figure 6. Average AQI Bucket for All Cities

In order to do analysis of Delhi city, which is of major concern due to rising pollution, the average AQI changes in Delhi from 2015 to 2024 is plotted in Figure 7(a) and Figure 7(b). While there are some fluctuations, the overall AQI remains in the unhealthy range throughout the decade. This long-term trend suggests a need for consistent policy enforcement and year-on-year environmental performance reviews. The government should also invest in public awareness and alternative fuel infrastructure. Even with some minor fluctuations, the AQI remains poor. This calls for a multi-year roadmap from the government involving continuous monitoring, clean energy incentives, and pollution-specific action plans each year.



```
MATCH (c:City)-[:HAS_STATION]→(:Station)-[:HAS_READING]→(r)
WHERE r.AQI IS NOT NULL AND substring(r.date, 0, 4) = "2024"
RETURN c.name AS City, round(avg(r.AQI), 2) AS `Average AQI`
ORDER BY `Average AQI` DESC
```

Category: City
Value: Average AQI

Figure 7(a). Queries and Average AQI by City for 2024 (Neo4j)

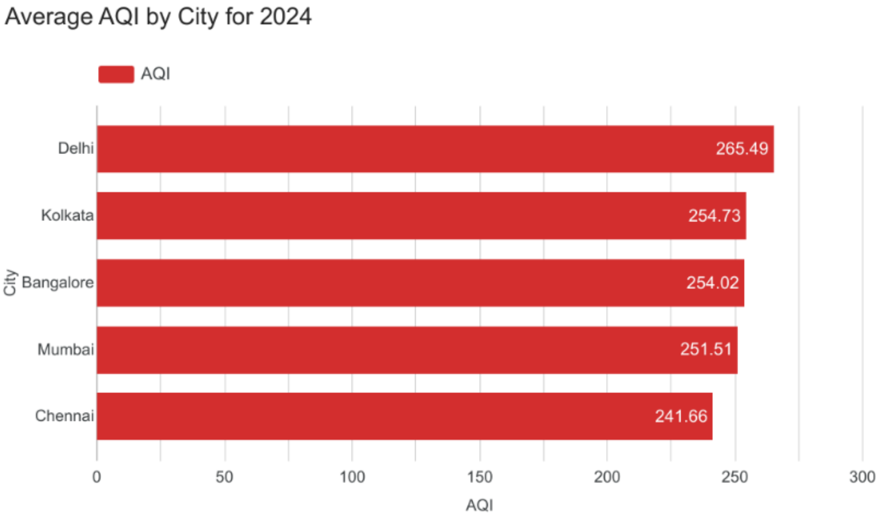
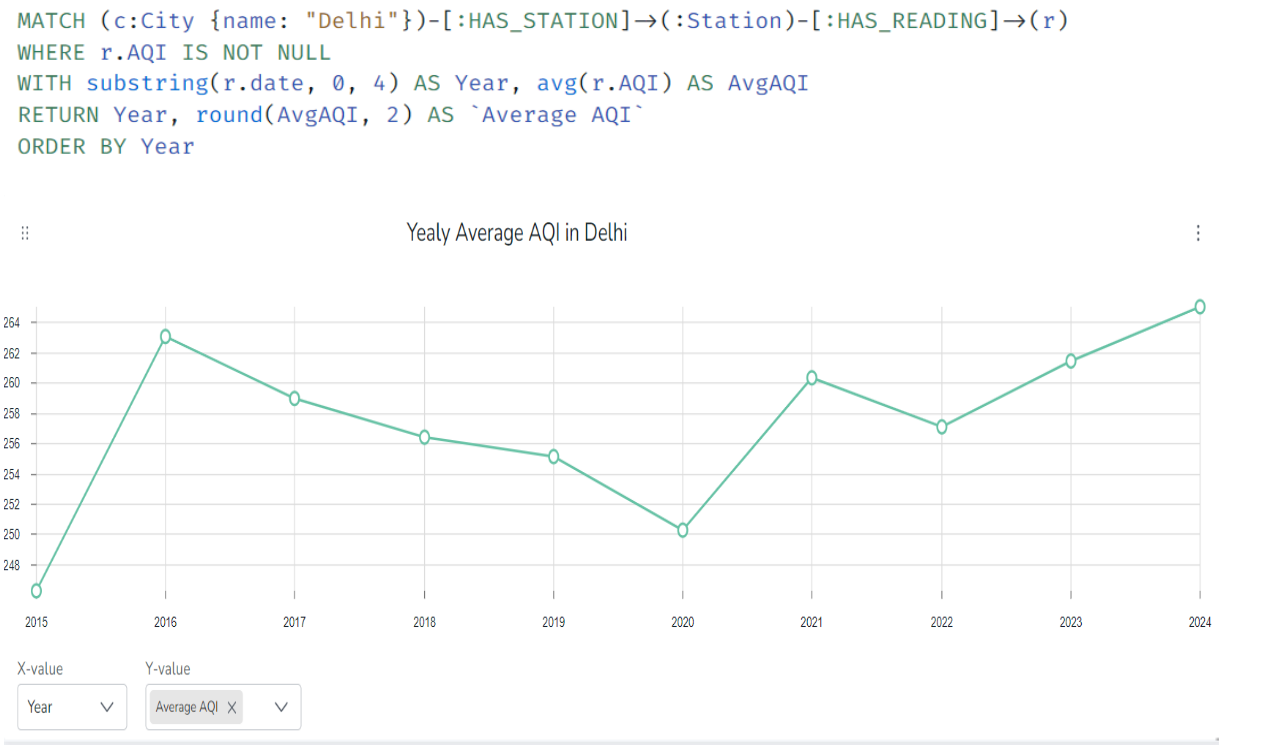


Figure 7(b). Average AQI by City for 2024 (Looker)

Looking closely, Figure 8(a) and Figure 8(b) show the monthly AQI trend in Delhi during 2024, with severe spikes in winter months due to stubble burning and unfavourable weather. To combat this, the government should implement crop residue management programs in nearby states and explore artificial rain or smog towers as seasonal mitigation techniques.



```
MATCH (c:City {name: "Delhi"})-[:HAS_STATION]->(:Station)-[:HAS_READING]->(r)
WHERE r.AQI IS NOT NULL
WITH substring(r.date, 0, 4) AS Year, avg(r.AQI) AS AvgAQI
RETURN Year, round(AvgAQI, 2) AS `Average AQI`
ORDER BY Year
```

X-value: Year

Y-value: Average AQI

Figure 8(a). Queries and Yealy Average AQI in Delhi (Neo4j)

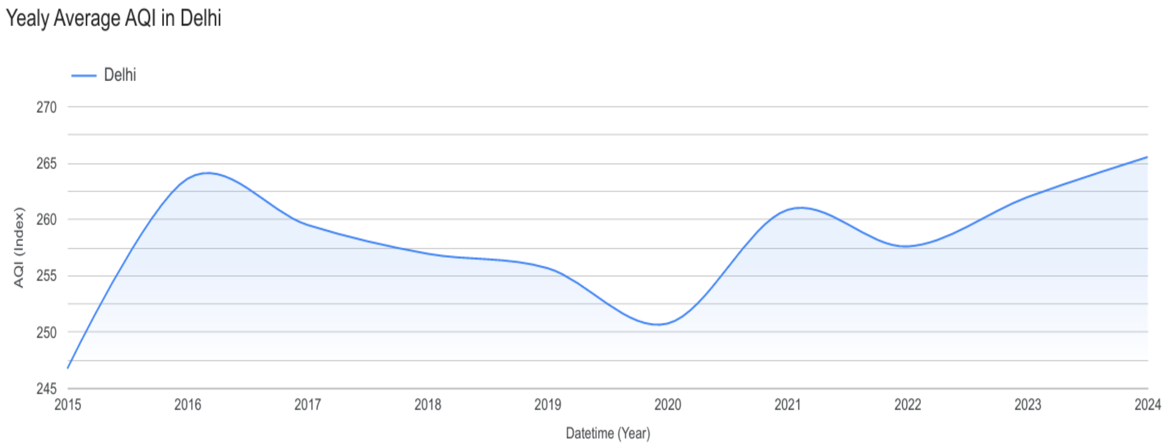
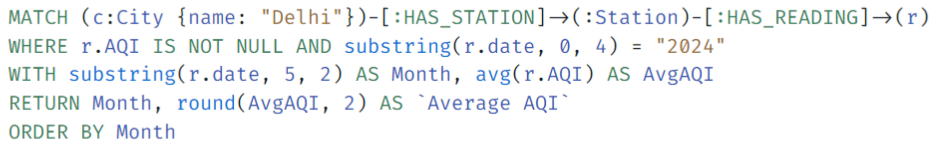
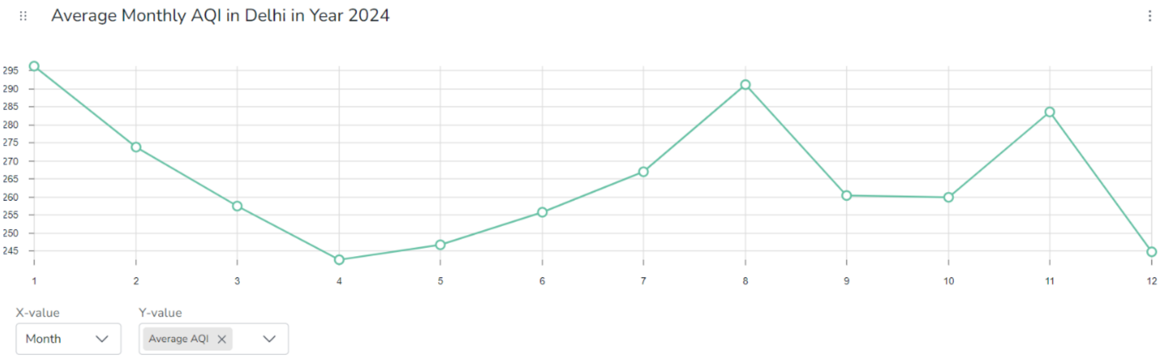


Figure 8(b). Yealy Average AQI Over Time in Delhi (Looker)

The chart in Figure 9(a) and Figure 9(b) outline monthly PM2.5 concentrations in Delhi over a 10-year period. The high and fluctuating PM2.5 levels reflect a chronic pollution issue. The government should focus on particulate control through improved traffic regulation, construction of dust management, and stricter industrial norms.



```
MATCH (c:City {name: "Delhi"})-[:HAS_STATION]->(:Station)-[:HAS_READING]->(r)
WHERE r.AQI IS NOT NULL AND substring(r.date, 0, 4) = "2024"
WITH substring(r.date, 5, 2) AS Month, avg(r.AQI) AS AvgAQI
RETURN Month, round(AvgAQI, 2) AS `Average AQI`
ORDER BY Month
```



X-value: Month

Y-value: Average AQI

Figure 9(a). Queries and Average Monthly AQI in Delhi in Year 2024 (Neo4j)

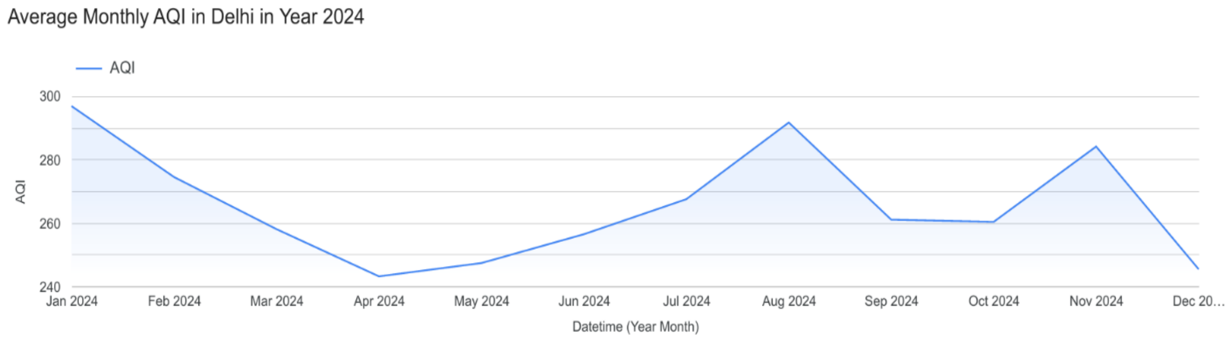


Figure 9(b). Average Monthly AQI in Delhi in Year 2024 (Looker)

The chart in Figure 10 compares monthly averages of O_3 and PM2.5, both of which follow distinct seasonal patterns. Peaks in different months suggest varying sources and behaviour. The government must deploy sensor-based tracking and adjust policy responses seasonally—for example, managing O_3 during hot months and PM2.5 in winter.

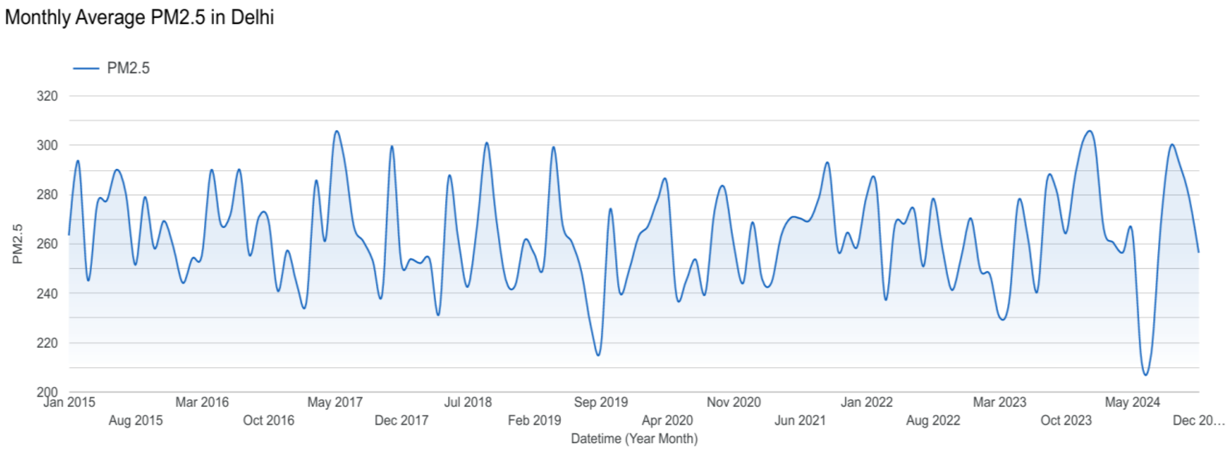


Figure 10. Monthly Average PM2.5 in Delhi

Figure 11 shows the average concentrations of BTX, with Toluene being the most prevalent—likely from vehicles and industries. To address this, stricter emission norms for industries and real-time pollution reporting systems in high-traffic zones can help reduce BTX exposure.

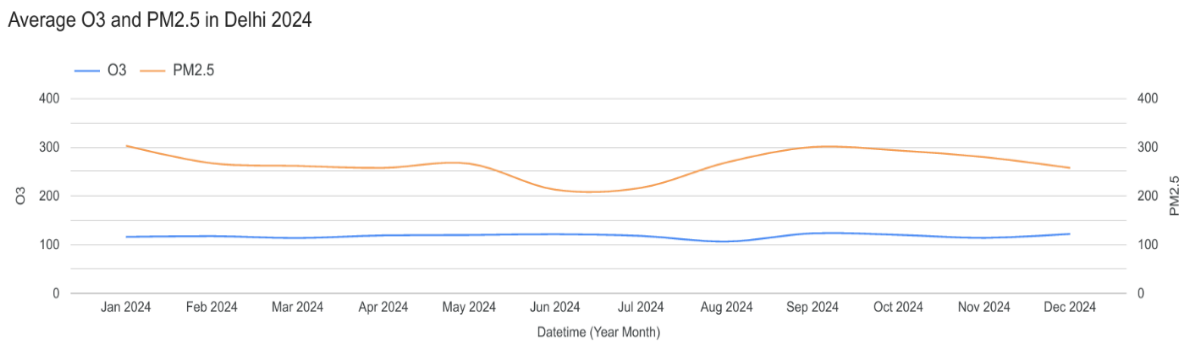


Figure 11. Average O3 and PM2.5 in Delhi 2024

Figure 12 shows the average BTX in Delhi for 2024.

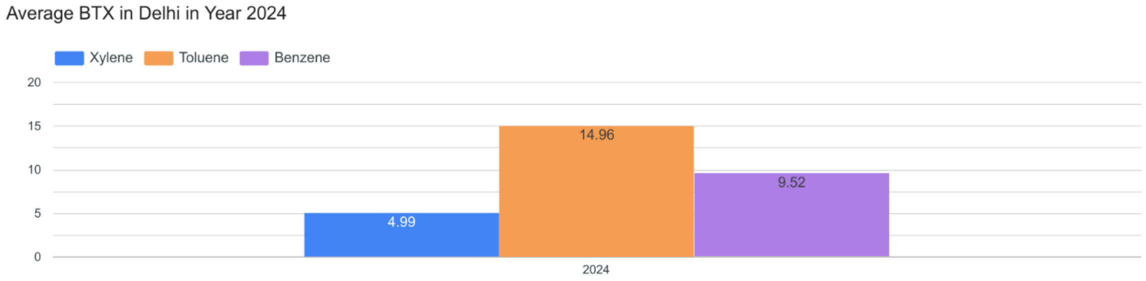


Figure 12. Average BTX in Delhi for Year 2024

4. RESULTS AND DISCUSSIONS

Based on the air quality data from 2015 to 2024, Delhi stands out as the most polluted city among the five major metros, followed by Delhi, Kolkata, Mumbai, Bangalore, and Chennai. The comprehensive data shows that Delhi has the highest average AQI of 257.88, closely followed by Kolkata (252.86), Mumbai (248.42) and Bangalore (245.19), while Chennai recorded the lowest at 240.52. Among all cities, Delhi in the year 2024 stands out as the period with the highest average AQI and PM2.5 levels, indicating severe air quality deterioration. The monthly AQI and PM2.5 levels in Delhi during 2024 show multiple peaks, particularly in winter months, suggesting worsening pollution likely due to seasonal factors and increasing emissions. Based on AQI bucket distributions and yearly pollutant trends, the overall ranking from most to least polluted cities are Delhi followed by Kolkata then Mumbai followed by Bangalore and lastly Chennai. The highest pollution levels were observed in Delhi in the year 2024, with an average AQI of 265.49 and PM2.5 levels frequently exceeding 300, especially in the winter months. The year 2024 also showed the highest concentrations of harmful gases like Benzene, Toluene, and Xylene. Throughout the dataset, Delhi consistently recorded the worst air quality across multiple pollutants and AQI buckets, with a significant portion of data falling into the Severe and Very Poor categories. In contrast, Chennai maintained the lowest pollution levels, making it the least polluted city. Overall, the data highlights a clear pollution severity ranking from Delhi (most polluted) to Chennai (least polluted), emphasizing the critical need for pollution control interventions, especially in the national capital. The phenomenon of climate change has a prolonged effect on atmospheric temperatures and weather patterns, whether they are caused by human intervention or natural processes. The AQI simplifies complex pollutant data into a single easily interpretable value which ranges from 0 to 500 with classifications varying between Good to Severe. This enables people to take necessary precautions like avoiding going outdoors on those days when the AQI is known to be Poor.

5. CONCLUSION

In Air is a fundamental element for the survival of all living entities on Earth. Existence on Earth without air is impossible. Factors like population growth, fossil fuel combustion, industrial activities, poor agricultural practices and vehicle emissions have contributed to decrease in air quality. The common air pollutants resulting from these activities, such as particulate matter 2.5 and 10 (PM2.5 and PM10), Carbon dioxide (CO2), SO2, Nitrous Oxide (NOx), and NO2, are analysed to evaluate their impact on the Air Quality Index. The contributions and adverse effects of various pollutants on the AQI have been studied and analysed over a duration of 10 years for 5 major cities of India. The air quality data across major Indian cities from 2015 to 2024 reveals a consistently alarming trend, with Delhi being the most polluted city throughout the decade. AQI data analysis visualized through graphs can make it easier to communicate urgent air quality updates to the public enabling rapid responses to pollution spikes. A poor air quality in India can contribute to global pollution leading to ongoing climate challenges. In this view, a detailed AQI analysis done on specific stations of major cities of India has revealed the pollution prone cities, allowing authorities and communities to give more attention for equitable air quality improvements.

Neo4j is found to be powerful tool for managing large dataset in graph structure and for querying the dataset rich in relationship. In the context of air quality data, which consists of elements like pollutants, their geographic locations, and the time of pollution are closely linked. Neo4j's native graph structure makes it a good choice for efficient data storage and quick data retrieval for real-time querying of interconnected information. As Neo4j is not primarily designed for data analysis and data visualization, we have utilized its real strength of performing fast and intuitive queries to surpass the capabilities of relational databases. In this view, the data analysis is carried out in Looker studio for better understanding and descriptive analysis of data for decision-makers. The combination of Neo4j and Looker is useful for scenarios where it is required to discover hidden patterns and identify irregularities from the dataset and for generating meaningful insights from complex and large-scale datasets. The research demonstrates the power of graph databases in uncovering complex air quality patterns and data visualization tools in drawing insights from the data. This work supports environmental science by presenting a scalable and well-informed approach to air quality monitoring.

ACKNOWLEDGEMENT

The authors would like to thank the anonymous reviewers for the suggestions to improve the paper.

FUNDING STATEMENT

The authors received no funding from any party for the research and publication of this article.

AUTHOR CONTRIBUTIONS

Jiten Chavda: Data Curation, Writing, Original Draft Preparation;
Kishan Bharvad: Data Analysis, Data Interpretation, Writing;
Rishap Parmar: Data Querying, Methodology, Validation;
Nidhi Arora: Conceptualization, Project Administration, Supervision, Writing – Review & Editing.

CONFLICT OF INTERESTS

No conflict of interests were disclosed.

ETHICS STATEMENTS

Our publication ethics follow The Committee of Publication Ethics (COPE) guideline. https://publicationethics.org/.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] C. Barbu, N. Tomus, A. Radu, M. Zlagnean, and D. Banu, “Comparative leaching tests of gold from unroasted and roasted pyrite using microwave radiation”, *Revista De Chimie*, vol. 71, no. 10, pp. 38-49, 2020, doi: 10.37358/rc.20.10.8348.

[2] K. Balakrishnan, S. Dey, T. Gupta, R. Dhaliwal, M. Brauer, A. Cohen et al., “The impact of air pollution on deaths, disease burden, and life expectancy across the states of India: The global burden of disease study 2017", *The Lancet Planetary Health*, vol. 3, no. 1, pp. e26-e39, 2019, doi: 10.1016/s2542-5196(18)30261-4.

[3] I. Manisalidis, E. Stavropoulou, A. Stavropoulos, and E. Bezirtzoglou, “Environmental and health impacts of air pollution: A review”, *Frontiers in Public Health*, vol. 8, 2020, doi: 10.3389/fpubh.2020.00014.

[4] N. Kunzli, R. Kaiser, S. Medina, M. Studnicka, O. Chanel, P. Filliger et al., “Public-health impact of outdoor and traffic-related air pollution: A European assessment”, *The Lancet*, vol. 356, no. 9232, pp. 795-801, 2000, doi: 10.1016/s0140-6736(00)02653-2.

[5] Central Pollution Control Board, *National Air Quality Index*, Ministry of Environment, Forest and Climate Change, Government of India, 2014. [Online]. Available: https://cpcb.nic.in/national-air-quality-index/.

[6] D. Sharma, and D. Mauzerall, “Analysis of air pollution data in India between 2015 and 2019”, *Aerosol and Air Quality Research*, vol. 22, no. 2, pp. 210204, 2022, doi: 10.4209/aaqr.210204.

[7] M. Sharma and O. Dikshit, “Comprehensive study on air pollution and greenhouse gases (GHGs) in Delhi,” Govt. of NCT Delhi and DPCC Delhi, 2016. [Online]. Available: http://environment.delhigovt.nic.in.

[8] A. Kumar, R. Singh, and S. Gupta, "Advances in air quality monitoring networks in India: Trends and challenges," J*ournal of Environmental Monitoring*, vol. 25, no. 3, pp. 123–135, 2023.

[9] P. Pant, R. Lal, S. Guttikunda, A. Russell, A. Nagpure, A. Ramaswami et al., “Monitoring particulate matter in India: Recent trends and future outlook”, *Air Quality, Atmosphere & Health*, vol. 12, no. 1, pp. 45-58, 2018, doi: 10.1007/s11869-018-0629-6.

[10] S. Guttikunda, and G. Calori, “A GIS based emissions inventory at 1 km x 1 km spatial resolution for air pollution analysis in Delhi, India”, *Atmospheric Environment*, vol. 67, pp. 101-111, 2013, doi: 10.1016/j.atmosenv.2012.10.040.

[11] P. Sharma and V. Pandey, "Air quality index and public health: Progress and challenges in India," *Environmental Science and Pollution Research*, vol. 31, no. 5, pp. 789–802, 2024.

[12] S. K. Guttikunda, S. K. Dammalapati, G. Pant, and A. Upadhya, "Assessing air quality during India's National Clean Air Programme (NCAP): 2019–2023," *Science of the Total Environment*, vol. 917, pp. 170408, 2023, doi: 10.1016/j.scitotenv.2024.170408.

[13] K. Maji, A. Namdeo, and L. Bramwell, “Driving factors behind the continuous increase of long-term PM2.5-attributable health burden in India using the high-resolution global datasets from 2001 to 2020”, *Science of the Total Environment*, vol. 866, pp. 161435, 2023, doi: 10.1016/j.scitotenv.2023.161435.

[14] S. N. Tripathi, A. Sharma, and V. Singh, “Assessing the impact of the National Clean Air Programme in Uttar Pradesh's non-attainment cities: A prophet model time series analysis,” *Environmental Science and Pollution Research*, vol. 31, no. 10, pp. 1456–1470, 2024.

[15] G. Finzi, and G. Tebaldi, “A mathematical model for air pollution forecast and alarm in an urban area", *Atmospheric Environment (1967)*, vol. 16, no. 9, pp. 2055-2059, 1982, doi: 10.1016/0004-6981(82)90276-1.

[16] I. Ziomas, D. Melas, C. Zerefos, A. Bais, and A. Paliatsos, “Forecasting peak pollutant levels from meteorological variables”, *Atmospheric Environment*, vol. 29, no. 24, pp. 3703-3711, 1995, doi: 10.1016/1352-2310(95)00131-h.

[17] G. N. Polydoras, J. S. Anagnostopoulos, and G. C. Bergeles, “Deposition of particles in a divergent–convergent street canyon using the Dippade model,” *Environmental Monitoring and Assessment.*, vol. 52, no. 3, pp. 335–346, 1998, doi: 10.1023/A:1005822600549.

[18] A. Chelani, C. Rao, K. Phadke, and M. Hasan, “Prediction of Sulphur Dioxide concentration using rtificial neural networks”, *Environmental Modelling & Software*, vol. 17, no. 2, pp. 159-166, 2002, doi: 10.1016/s1364-8152(01)00061-5.

[19] S. Ameer et al., “comparative analysis of machine learning techniques for predicting air quality in smart cities,” in *IEEE Access*, vol. 7, pp. 128325-128338, 2019, doi: 10.1109/ACCESS.2019.2925082.

[20] S. Ojagh, F. Cauteruccio, G. Terracina, and S. Liang, “Enhanced air quality prediction by edge-based spatiotemporal data preprocessing”, *Computers & Electrical Engineering*, vol. 96, pp. 107572, 2021, doi: 10.1016/j.compeleceng.2021.107572.

[21] Z. Zhu, Y. Qiao, Q. Liu, L. Cong-hua, E. Dang, W. Fu et al., “The impact of meteorological conditions on air quality index under different urbanization gradients: A case from Taipei”, *Environment, Development and Sustainability*, vol. 23, no. 3, pp. 3994-4010, 2020, doi: 10.1007/s10668-020-00753-7.

[22] A. -S. Chowdhury, M. S. Uddin, M. R. Tanjim, F. Noor, and R. M. Rahman, “Application of data mining techniques on air pollution of Dhaka city,” *2020 IEEE 10th International Conference on Intelligent Systems (IS)*, Varna, Bulgaria, pp. 562-567, 2020, doi: 10.1109/IS48319.2020.9200125.

[23] J. Prakash, S. Agrawal, and M. Agrawal, “Global trends of acidity in rainfall and its impact on plants and soil”, *Journal of Soil Science and Plant Nutrition*, vol. 23, no. 1, pp. 398-419, 2022, doi: 10.1007/s42729-022-01051-z.

[24] J. Proctor, “Atmospheric opacity has a nonlinear effect on global crop yields”, *Nature Food*, vol. 2, no. 3, pp. 166-173, 2021, doi: 10.1038/s43016-021-00240-w.

[25] U. Goel, S. Sathyan, D. N. A. Siddiqui, and P. Sachan, "Vehicular pollution, their effect on human health and mitigation measures," *The International Journal of Creative Research Thoughts (IJCRT)*, vol. 6, no. 1, pp. 73–77, Mar. 2018, [Online]. Available: https://www.ijcrt.org/papers/IJCRT184140.pdf.

[26] C. Pande, R. Latha, and M. Satyanarayana, “Evaluation of machine learning and deep learning models for daily air quality index prediction in Delhi City, India”, *Environmental Monitoring and Assessment*, vol. 196, no. 12, 2024, doi: 10.1007/s10661-024-13351-1.

[27] K. Rajesh, and S. Kumar, “Deep reinforcement learning for urban air quality management: multi-objective optimization of pollution mitigation booth placement in Metropolitan environments”, *IEEE Access,* vol. 13, pp. 146503-146526, 2025, doi: 10.1109/access.2025.3599541.

[28] P. Patel, and A. Kumar, “Comparative study of air quality index of two Metropolitan cities (Lucknow and Kanpur) in year-2023”, *World Journal of Advanced Research and Reviews*, vol. 22, no. 2, pp. 1637-1651, 2024, doi: 10.30574/wjarr.2024.22.2.1590.

[29] S. Banerjee, U. Basu, and B. Basu, “An entropy based comparative study of regional and seasonal distributions of particulate matter in Indian cities,” *Environmental Science*, Feb. 12, 2025, doi: 10.48550/arXiv.2502.08491.

[30] G. Bhatia, “Predictive urban air quality monitoring for healthier cities”, *South Eastern European Journal of Public Health*, pp. 1627-1634, 2024, doi: 10.70135/seejph.vi.2166.

BIOGRAPHIES OF AUTHORS

	Jiten Chavda is an independent researcher with a solid foundation in statistics. He analyses data using Python, SQL, and Looker Studio to uncover valuable insights that support informed decision-making. He can be contacted at chavdajiten00@gmail.com.
	Kishan Bharvad is a dedicated data analyst who enjoys turning raw data into useful insights. Using tools like Python and graph databases such as Neo4j, he creates clear reports and visualizations that help teams make better decisions. With a sharp eye for detail, he spots patterns, tracks performance, and finds new opportunities across different industries. Always eager to learn, he keeps improving his skills in analytics, data tools, and graph technology to stay up to date in today’s data-driven world. He can be contacted at kishanbharvad4221@gmail.com.
	Rishap Parmar is an individual researcher specializing in data analysis. With a strong foundation in statistics and programming, he transforms complex datasets into clear, actionable insights. Passionate about uncovering patterns and trends, he leverages tools like Python, R, and SQL to explore data from diverse domains. His work supports decision-making processes across industries, from finance to healthcare. Rishap is committed to continuous learning, staying current with emerging techniques in machine learning and data visualization to elevate his analytical impact. He can be contacted at rishapparmar360@gmail.com.
	Dr. Nidhi Arora is a fervent researcher in the field of Artificial Intelligence and Data Analytics. She has excellent proficiency in Predictive Analytics, Machine Learning algorithms and custom algorithms development around AI. She has published many research papers in Scopus indexed journals and book chapters with well-known publishers. She can be contacted at nidhi.fst.1070@gmail.com.