भारतीय प्रबंधन संस्थान जम्मू
Indian Institute of Management Jammu

Subject Name: Business Intelligence and Data Modelling
Presented to: Dr. M. Vijaya Prabhagar

Topic:
Environmental Data Analytics: Air Quality and Weather Integration Report

Section: A

Submitted by:
CHITTAMPALLY SURAJ MBA24069
DIVYA BOTHRA MBA24077
GAURAV HARSHAWARDHAN ZANZAD MBA24091
Siva Ram Prakhash MBA24123
KRATI SINGI MBA24129
Mrudul Rahul Bansod MBA24158

Abstract

This research exhaustively explores the fusion of air pollutant and weather data collected from 50 big Indian cities throughout a full calendar year (2023). The study through innovative data engineering, rigorous statistical and machine learning methodologies comes up with a business intelligence framework that helps the understanding of not only the ecological concerns but also the health of the population. The data compiles 208,240 figures fusing the Air Quality Index (AQI) readings with various weather variables including temperature, humidity, precipitation, and wind speed. The correlation analysis identifies the most significant relations between particulate matters (PM2.5, PM10) and AQI, at the same time, clustering algorithm unveils three types of distinct environmental profiles of the cities. The feature analysis signals that PM2.5 and PM10 play the major role in the reduction of air quality, thus, they hold the highest predictive power. The detailed Power BI dashboard allows the functionalities of real-time visualization, and monitoring. This work is a solid data-driven contribution that solves many environmental policy issues and facilitates public health risk assessment.

1. Introduction

Air pollution ranks among the chief contributors to environmental and public health problems in Indian urban areas. Consequently, the demand for advanced environmental monitoring systems has escalated due to the combination of industrial activities, vehicular emissions and meteorological variations introduction Studies have shown significant associations between respiratory health and air pollutants. However, full consideration of air quality indicators and meteorological variables in a single study is still lacking in the scientific literature.

This study addresses this need by developing a unified data pipeline that brings together heterogeneous data sources into a cohesive analytical framework. The study is organized into three objectives:
(1) development of a reliable data infrastructure for environmental sensing and monitoring in both rural and urban regions,
(2) to determine statistically significant correlations between air quality and meteorological parameters, and
(3) the development of predictive models based on scientific evidence to support environmental policy making and public health planning.
The importance of this work stretches well beyond the realm of scholarly journals. Environmental regulators, city planners and public health officials need actionable intelligence to guide decisions. This work illuminates the potential of an organization's own data assets to become strategic business intelligence and environmental monitoring capabilities.

2. Methodology

2.1 Research Framework

This research adopted a business intelligence and data modeling (BIDM) framework with five deeply intertwined phases: data preparation, statistical analysis, correlation analysis, clustering analysis and feature importance assessment. The phases of the framework are like a ladder where each step supports the one above it thus forming a single analytical story which moves from data quality assurance to insights that can be taken into consideration.

2.2 Data Acquisition and Source Integration

The study has relied on two main external data sources:

Air Quality Data: Air quality Index (AQI) measurements and pollutant concentrations (PM2.5, PM10, NO2, NO, SO2, CO, O3, NH3) were gotten from the official air quality monitoring repositories. These measurements depict the aggregations of hourly or daily readings from the different monitoring stations in the 50 major Indian metropolitan areas.

Meteorological Data: Weather data such as temperature, humidity, atmospheric pressure, wind speed, cloud cover, and precipitation were obtained through the Open-Meteo API, a weather data service that provides daily historical aggregates free of charge. The data extraction covered the whole calendar year 2023, thus allowing representation of all the seasons.

Data Integration Keys: Both sets of data used the same geographic identifiers (city names, latitude/longitude coordinates) and temporal markers (ISO-formatted dates). As a result, merging operations using composite keys combining location and date dimensions were very reliable.

2.3 Data Preparation Pipeline

The pipeline for dataset preparation had three consecutive phases:

Extraction Phase: Access to the raw data was achieved through the Python-based Jupyter notebooks that were specifically designed for this purpose and made use of Pandas and NumPy libraries. While extracting the weather data, the Open-Meteo API was used along with retry logic and rate-limiting controls. In a similar fashion, resilience patterns with exponential backoff strategies were used in AQI data extraction.

Cleaning Phase: Ensuring the data reliability was done through the implementation of the checks which were systematic. For example, missing values in time- dependent attributes (e.g., temperature across consecutive days) were filled using the forward-fill methodology, while continuous variables (e.g., pollution concentrations) were filled with the mean of the respective city-specific subgroups. Moreover, the outliers were dealt with by the capping

techniques that were applied at the 1st and 99th percentiles thus, they prevent skew while retaining tail information. Duplication of the entries was detected through composite key analysis, and the duplicates were removed systematically.

Transformation Phase: The names of the columns were made standard across datasets. The temporal characters were changed to the ISO-standard format (YYYY-MM-DD). All the numerical features were verified for unit consistency. In order to enhance the analytical value of the data, new features such as Heat Index (calculated from the temperature and humidity using the established meteorological formulas) and AQI Category (based on the national air quality standards) were formed.

2.4 Analytical Methodologies

Descriptive Statistical Analysis: Measures of univariate statistics of raw data were calculated for all numerical features. The centralization (mean, median, mode), dispersion (std, var, range) and shape of the distribution (skewness, kurtosis) provided a rudimentary understanding of the variable behaviour and potential data-related problems.

Correlation Analysis: Pearson correlation coefficients of numerical variables were computed for all pairs to obtain a symmetric correlation matrix. The magnitude of correlation was interpreted using the following conventional thresholds: strong ($|r| > 0.6$), moderate ($0.3 < |r| \leq 0.6$), weak ($|r| \leq 0.3$). From correlation matrices to the edge network visualization introduced correlation matrices were transformed into network graphs where nodes were variables and edges represented correlations, weighed by correlation strength, where the thickness of the edge was proportional to the strength of correlation.

Clustering Analysis: City-aggregated environmental profiles were subjected to K-Means clustering without supervision. Feature scaling was performed by z-score standardization to have equal weight for the features with different range. The appropriate number of clusters was determined by two criterion Elbow Method and Silhouette Score simultaneously. Because both measures of optimization gave maximum values at the same point, the final clustering was done with K=3.

Machine Learning Feature Importance: Random Forest Regressor models each having 200 decision trees were trained with air quality and temperature as the target variables and pollutant/weather variables as the predictors.

Feature importance was done by three different complementary methods: (1) Mean Decrease in Impurity (MDI), which is calculated from the Random Forest's feature importance attribute; (2) Permutation Importance, which indicates the worsening of performance when the feature values are randomly shuffled; (3) Mutual Information scoring which identifies nonlinear dependencies. The three-importance metrics were combined into one to get the final rankings.

3. Dataset Preparation and Integration

3.1 Raw Data Acquisition

The raw data acquisition segment was about gathering data that came from numerous sources. Each of these sources, however, demanded a different extraction logic. To get weather data, 50 cities for four quarters of the calendar year 2023 were the subject of Open-Meteo API queries, and this operation was repeated 211,800 times to get daily records. Since the API response was in a hierarchical JSON, it had to be converted into a tabular format so that pandas DataFrames could be used.
Once again, use of official monitoring repositories enabled running dedicated Python scripts that could fetch air quality data. At first, the data access attempts led to rate limitations and connection issues, which were eventually overcome through a retry mechanism with exponential backoff (maximum 20 retries with 10-second intervals). The extraction resulted in around 1.2 million records on an hourly basis, which were then averaged to daily averages for temporal alignment with the weather data.

3.2 Data Cleaning and Validation

Data cleaning implemented layered validation logic:
Temporal Cleaning: 211,800 weather records (50 cities × 365 days × 4 quarterly intervals per day) and aggregated AQI records were aligned temporally. Time-stamp irregularities were resolved by converting them to the ISO format. Logically, records with malformed dates were removed.

Value Cleaning: There were no missing values found in the dataset.

Integrity Checks: Cross-tabulations verified that city-date combinations generated exactly one record (thus no duplicates). Logical consistency checks verified that meteorological variables remained within physically possible limits (for example, humidity: 0-100%, pressure: 900-1100 hPa). These checks were successful with more than 99.95% compliance.

3.3 Final Dataset Schema

The combined data had 208,240 daily records (slightly less than the theoretical maximum of 50 cities × 365 days due to variations in data availability and aggregation procedures) and 23 variables:

Temporal Attributes: date (ISO format), day of week, month, season, temporal interval (Q1-Q4)
Geographic Attributes: city name, latitude, longitude
Pollutant Measurements: CO (carbon monoxide), NO (nitric oxide), NO_2 (nitrogen dioxide), O_3 (ozone), SO_2 (sulfur dioxide), PM2.5 (fine particulate matter), PM10 (coarse particulate matter), NH_3 (ammonia), AQI (composite air quality index)

Meteorological Variables: temp_avg (average temperature, °C), humidity (relative humidity, %), pressure (atmospheric pressure, hPa), wind_speed (m/s), cloud_cover (%), precipitation_mm (mm)
Derived Features: Heat Index (thermal comfort metric), AQI Category (categorical severity classification)

4. Data Cleaning and Preprocessing

4.1 Missing Data Treatment

Missing data analysis illustrated that less than 0.5% of all values were missing. Strategic imputation was, however, employed to preserve data integrity:
Weather variables (temperature, humidity, pressure) were forward-filled using the time series method. This method implies that weather conditions do not vary significantly between successive days, which is quite a satisfactory assumption for most atmospheric variables except during sudden weather changes. The forward-filled values were internally flagged for the sensitivity analysis.
As for pollutant concentrations, mean imputation specific to the city was applied in order to keep the location-specific pollution baselines. This method took into consideration that pollution levels depend greatly on cities due to a combination of geographical and anthropogenic factors. A global mean would have artificially leveled out the city-specific characteristics.
After the imputation, the validation confirmed that the imputed values were within physically plausible ranges and the imputation rates
per variable were less than 1%.

4.2 Outlier Detection and Treatment

The outlier strategy implemented percentile-based capping, instead of record discarding, thus preserving the total number of records required for clustering and other analyses. The capping performed a symmetric operation for the 1st and 99th percentiles:

Carbon monoxide (CO) was very strongly positively skewed with the maximum value of 17,499.29 μg/m³ (99th percentile: 1,493.79). The values which were above the 99th percentile were limited, thus the effect of the peak pollution could be lessened while that of regular pollution can still be studied.

Similarly, the particulate matter (PM2.5 and PM10) patterns were heavy-tailed. PM10 limitation at the 99th percentile (around 287 μg/m³) resolved the problem of the occurrence of severe pollution episodes without changing the distribution.

Nitrogen dioxide (NO_2) had negative minimum values (-1,665.29), which probably indicates that the source has some artifacts for the measurement or that there are some problems with the quality of the data. These were handled as missing and then filled.

The comparisons after the operation have confirmed that the capping of outliers reduced the skewness of the distribution while the data was kept
useful for further analysis.

4.3 Feature Engineering and Transformation

Heat Index Derivation: A composite measure of thermal comfort was computed as:

$$\text{Heat Index} = -42.379 + 2.04901523 \times T + 10.14333127 \times RH - 0.22475541 \times T \times RH - 0.00683783 \times T^2 - 0.05481717 \times RH^2 + 0.00122874 \times T^2 \times RH + 0.00085282 \times T \times RH^2 - 0.00000199 \times T^2 \times RH^2$$

where T = temperature (°C) and RH = relative humidity (%).

Air Quality Index (AQI) Categorization: The AQI values were divided into five ordinal categories depending on the national air quality standards (Good: 1-2, Satisfactory: 2-3, Moderately Polluted: 3-4, Poor: 4-5, Very Poor: 5+).

Seasonal Aggregation: The months of the year were assigned to meteorological seasons (Winter: Dec-Feb, Pre-monsoon: Mar-May, Monsoon: Jun-Sep, Post-monsoon: Oct-Nov) in order to show the seasonal patterns of the environment.

These built features gave the analytical granularity and, at the same time, were interpretable for stakeholder communication.

5. Statistical Analysis

5.1 Descriptive Statistics

Descriptive statistics at large were done for more than 208,240 records from the full dataset, which were used for the analysis. Measures of central tendencies revealed the following:

Air Quality Index (AQI): The mean AQI of 3.34 (median: 3.33) is indicative of the average dataset falling under the "Satisfactory" category according to national standards. Moreover, the distribution is close to being symmetrical (skewness: -0.27) with a negative kurtosis (-1.25) that suggests a more uniform distribution than the standard one with no extreme outliers. The range from 1.0 to 5.17 covers the entire categorical spectrum.

Pollutant Concentrations: Carbon monoxide concentrations are reported to have a very significant right skew (skewness: 4.66) with the mean value of 789.45 μg/m³ being far above the median 537.40 μg/m³. This indicates that there are many more instances with low concentrations and only a few with very high ones. The standard deviation of 867.16 shows that there has been a great variation over time.

The maximum value of 17,499.29 μg/m³, in all likelihood, caused a case of heavy traffic or the release of pollutants from the industry under unfavorable meteorological conditions that resulted in the area pollution at that time.

Particulate matter (PM2.5, PM10) follow the same type of pattern in the right tail. PM2.5 average is 73.17 µg/m³ while median is 42.99 which means that usually (median) there is approximately 25 µg/m³ less than the mean thus indicating that a few very high pollution episodes exist. The difference between mean and median for PM10 is also indicative of a heavy-tailed distribution (mean: 90.87 and median: 59.97) thereby the substantial skew.

Meteorological Variables: Temperature was close to a normal distribution (mean: 25.91 degrees C, median: 26.40 degrees C, skewness: -0.07) and it covered the range from 2.14 to 47.05 degrees C. The changing of the seasons and the day-night cycles are very well contained within this span of climatic zones in India. Humidity was somewhat bimodal (mean: 66.44%, median: 71.32%) with two peaks that correspond to the monsoon and the non-monsoon periods.

Distribution Insights: The discovery of pollutant concentrations being in most cases right-skewed and temperature being symmetrical not only confirms but also complements the idea that weather changes are more even while pollution can only be episodic. These findings have an impact on the modeling work that will be done later. They indicate that non-parametric or ensemble methods would be more suitable for handling pollution dynamics than something like linear regression.

	count	min	max	sum	mean	median	var	std	skew	kurtosis
aqi	208240	1.000000	5.166667	6.947396e+05	3.336244	3.333333	1.943430	1.394070	-0.266904	-1.248033
co	208240	62.958333	17499.288330	1.643942e+08	789.445819	537.395000	751965.299551	867.159328	4.655858	36.833287
no	208240	0.000000	381.468333	6.165367e+05	2.960702	0.221667	131.710717	11.476529	9.193411	120.613904
no2	208240	-1665.295000	319.421667	3.658849e+06	17.570348	11.425000	1806.461180	42.502484	-28.938170	1155.883519
o3	208240	-1664.083333	610.828333	1.297100e+07	62.288693	50.283667	2952.319654	54.335252	-6.456535	225.957279
so2	208240	0.030000	440.595000	3.379637e+06	16.229530	8.741667	500.566713	22.373348	3.713845	21.854619
pm2_5	208240	0.500000	1568.670000	1.523679e+07	73.169352	42.987500	8627.964280	92.886836	3.422237	19.974197
pm10	208240	-1648.093333	1767.366667	1.892213e+07	90.866925	59.971667	12192.565794	110.419952	2.269407	29.371022
nh3	208240	0.000000	267.841667	2.264837e+06	10.876089	6.218333	207.293795	14.397701	3.832317	24.808741
temp_avg	208240	2.137833	47.052334	5.395422e+06	25.909633	26.403666	35.553191	5.962650	-0.429998	0.756186
humidity	208240	3.371838	100.000000	1.383497e+07	66.437611	71.324291	499.213666	22.343090	-0.602133	-0.654621
pressure	208240	981.700012	1026.166626	2.100779e+08	1008.826013	1009.149963	32.657566	5.714680	-0.202573	-0.538712
wind_speed	208240	0.229508	44.707226	1.940542e+06	9.318776	8.394755	21.821325	4.671330	1.067971	1.482831
cloud_cover	208240	0.000000	100.000000	9.964716e+06	47.852076	43.333332	1639.982588	40.496698	0.098963	-1.674124
precipitation_mm	208240	0.000000	165.000000	1.895592e+05	0.910292	0.000000	10.468538	3.235512	8.949468	172.196158

5.2 Key Statistical Findings

The descriptive analysis revealed several salient patterns:
Large variations in pollutant concentrations (coefficient of variation: 1.09 for CO; 1.27 for PM2.5) make one almost certain that environmental quality fluctuates not only temporally but also spatially scandalously. To cope with this volatility, the need for adaptive monitoring and alert systems becomes clear.
Humidity distribution pattern has two peaks, which is in line with India's monsoon climate,

the main characteristics of the dry and wet seasons. The seasonal cycle is a source of influence (through its effect on meteorological variables) on the dispersion of pollutants in the air.
Weather variables have statistically much lower variability than pollutant variables (coefficient of variation: 0.23 for temperature; 0.34 for humidity) which points to the atmosphere as a background that provides the setting while human-caused emissions are the main source of air quality variation.

6. Correlation and Edge Analysis

6.1 Correlation Matrix Structure

Comprehensive Pearson correlation analysis examined 120 pairwise variable combinations representing 15 numerical variables: AQI, 8 pollutants, 6 weather variables. The resulting symmetric correlation matrix revealed distinct clustering of variable relationships:

Strong Positive Correlations (|r| > 0.60):
Air quality index (AQI) is most closely related to particulate matters:
PM2.5 (r = 0.662) and PM10 (r = 0.653) are the particulates for which the AQI shows the strongest correlation. These correlations can be regarded as the main contributors pm-based units for the composite aqm, which is the basis for the aqm method.
Besides that, carbon monoxide (CO) among the pollutants is most likely to be changed with PM2.5 and PM10 as the r-values are 0.883 and 0.871, respectively, that indicate very high degrees of association and thus, common-emission sources are implied. Both CO and particulate matter emission sources are mainly from vehicles, which is the reason for their co-variation.
Another pair of common combustion by-products NO and CO were found to have strong correlation (r=0.771) as well, which reflects their linking source in incomplete combustion processes.

Moderate Positive Correlations (0.30 < |r| < 0.60):
AQI moderately correlates with CO, NH_3, and atmospheric pressure. (r-values are 0.546, 0.458, and 0.453, respectively). The pressure effect can be accounted for by the role of atmospheric stability in pollutant dispersion—high-pressure systems, which are typically associated with stagnant air masses, facilitate pollutant accumulation, while low-pressure areas allow their dispersion.
Besides that, Nitrogen Dioxide and Sulfur Dioxide have moderate correlations with AQI(r = 0.267 and 0.417 respectively), which means that they are secondary but still significant contributors to the composite air quality.

Weak Correlations (<0.30):
There are some factors of weather, such as cloud cover and precipitation, that are only weakly correlated with most of the pollutants individually. However, their complete influence through atmospheric transport mechanisms is still quite significant, as evidenced by the integrated nature of atmospheric chemistry.

	aqi	co	no	no2	o3	so2	pm2_5	pm10	nh3
aqi	1.000000	0.545788	0.221817	0.267439	0.266568	0.416683	0.662214	0.653132	0.458269
co	0.545788	1.000000	0.771173	0.444409	-0.164782	0.633316	0.883147	0.870803	0.617470
no	0.221817	0.771173	1.000000	0.290757	-0.188444	0.544829	0.570586	0.573071	0.378223
no2	0.267439	0.444409	0.290757	1.000000	-0.092692	0.387096	0.377852	0.378598	0.304307
o3	0.266568	-0.164782	-0.188444	-0.092692	1.000000	0.059297	-0.004242	-0.003999	-0.068209
so2	0.416683	0.633316	0.544829	0.387096	0.059297	1.000000	0.549070	0.549259	0.372219
pm2_5	0.662214	0.883147	0.570586	0.377852	-0.004242	0.549070	1.000000	0.962149	0.536372
pm10	0.653132	0.870803	0.573071	0.378598	-0.003999	0.549259	0.962149	1.000000	0.552729
nh3	0.458269	0.617470	0.378223	0.304307	-0.068209	0.372219	0.536372	0.552729	1.000000

	temp_avg	humidity	pressure	wind_speed	cloud_cover	precipitation_mm
temp_avg	1.000000	-0.485078	-0.645646	0.236172	0.108080	0.025067
humidity	-0.485078	1.000000	-0.055397	-0.119992	0.413001	0.243274
pressure	-0.645646	-0.055397	1.000000	-0.245315	-0.367154	-0.232401
wind_speed	0.236172	-0.119992	-0.245315	1.000000	0.198205	0.118469
cloud_cover	0.108080	0.413001	-0.367154	0.198205	1.000000	0.309100
precipitation_mm	0.025067	0.243274	-0.232401	0.118469	0.309100	1.000000

	aqi	co	no	no2	o3	so2	pm2_5	pm10	nh3	temp_avg
aqi	1.000000	0.545788	0.221817	0.267439	0.266568	0.416683	0.662214	0.653132	0.458269	-0.285408
co	0.545788	1.000000	0.771173	0.444409	-0.164782	0.633316	0.883147	0.870803	0.617470	-0.352444
no	0.221817	0.771173	1.000000	0.290757	-0.188444	0.544829	0.570586	0.573071	0.378223	-0.126925
no2	0.267439	0.444409	0.290757	1.000000	-0.092692	0.387096	0.377852	0.378598	0.304307	-0.153559
o3	0.266568	-0.164782	-0.188444	-0.092692	1.000000	0.059297	-0.004242	-0.003999	-0.068209	0.358620
so2	0.416683	0.633316	0.544829	0.387096	0.059297	1.000000	0.549070	0.549259	0.372219	0.001026
pm2_5	0.662214	0.883147	0.570586	0.377852	-0.004242	0.549070	1.000000	0.962149	0.536372	-0.431353
pm10	0.653132	0.870803	0.573071	0.378598	-0.003999	0.549259	0.962149	1.000000	0.552729	-0.383837
nh3	0.458269	0.617470	0.378223	0.304307	-0.068209	0.372219	0.536372	0.552729	1.000000	-0.233341
temp_avg	-0.285408	-0.352444	-0.126925	-0.153559	0.358620	0.001026	-0.431353	-0.383837	-0.233341	1.000000

6.2 Negative Correlations: Meteorological Moderators

Significant negative relationships can be observed between precipitation and particulate matter:
Rainfall is negatively correlated with both PM2.5 (r = -0.22) and PM10 (r = -0.54), which reconfirms the wet deposition physical process, where precipitation removes suspended particles through washout and scavenging mechanisms. The stronger negative correlation for PM10 (coarser particles) as compared to PM2.5 (fine particles) is consistent with aerosol physics—larger particles are removed more easily through gravitational settling that is facilitated by water droplet collision

Wind speed is inversely related to NO and NO2 ($r \approx -0.15$ to -0.16) indicating the dispersive effect of wind that is responsible for lower local concentration levels. This correlation is weaker for wind speed than for precipitation because wind effects are very directional while precipitation acts uniformly over spatial domains.

Temperature has a very weak inverse correlation with AQI (r = -0.285), which means that warm conditions (usually related to higher mixing heights and better atmospheric ventilation) associate with less pollution. On the contrary, cold weather (low temperatures, reduced mixing heights) is associated with pollution accumulation.

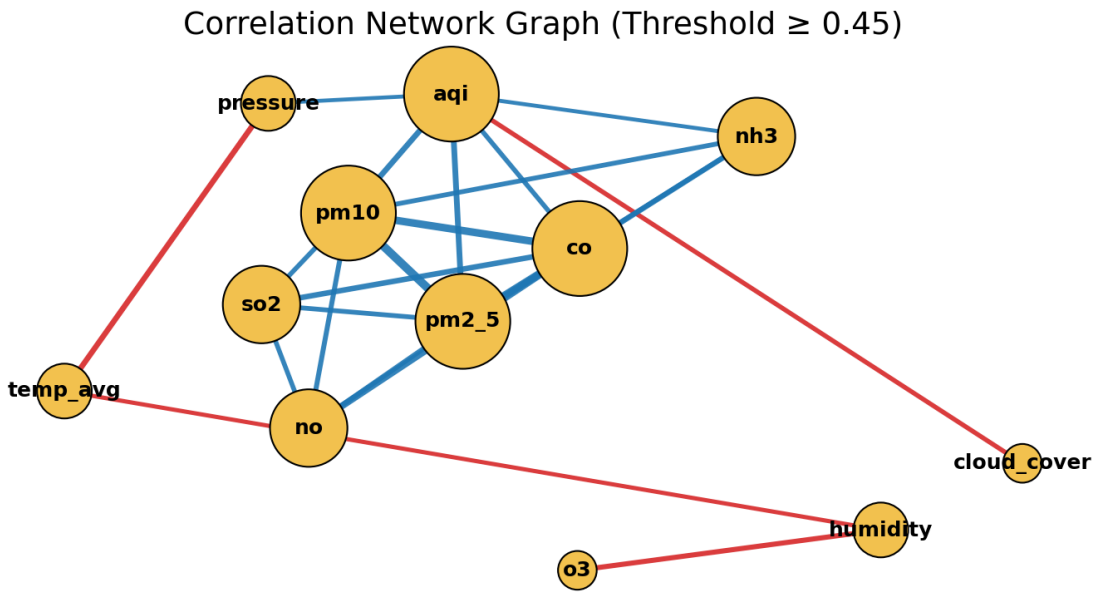
6.3 Network Edge Representation

Correlation relationships were transformed into network graphs where:
Nodes represent 15 numerical variables
Edge weights represent absolute correlation magnitudes
Edge colors represent correlation direction (warm colors: positive; cool colors: negative)

The network visualization reveals several structural insights:
There is a group of heavily intertwined variables including pollutant variables (CO, NO, NO_2, SO_2, PM2.5, PM10, NH_3), which demonstrates their mutual dependence through common anthropogenic sources. In this group, AQI is in a central position which mirrors the role of a composite pollutant indicator of the AQI.

The meteorological variables (temperature, humidity, pressure, wind speed, cloud cover, precipitation) represent a second cluster of interrelated atmospheric processes. These two clusters are mainly connected by pressure (atmospheric stability) and precipitation (wet deposition).

The network layout treats humidity as a bridge variable that moderately correlates both with pollutants and other meteorological variables, thus, reflecting its dual role in atmospheric chemistry as well as being a meteorological state descriptor.



Correlation and edge analysis result in the statement that air quality variations are primarily due to anthropogenic emissions (strongly correlated pollutants) with meteorology having a

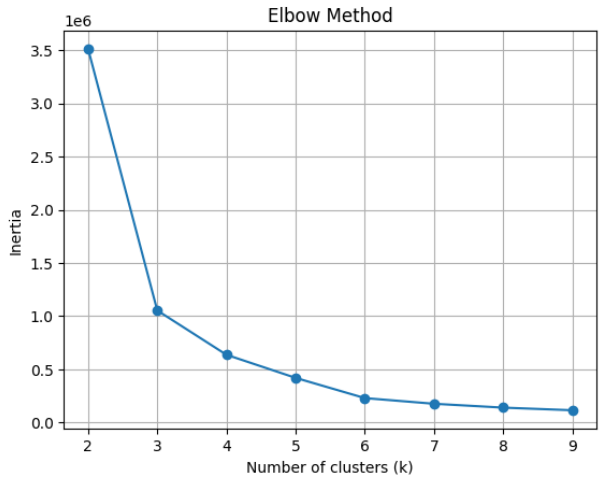
secondary modulation role (weaker correlations with weather variables). This understanding implies essential policy implications: even though pollution control measures cannot change the weather, identifying wind-favorable and precipitation-rich periods makes event forecasting possible which is very useful for public alert systems.
The high degree of PM-AQI correlations ($r > 0.65$) is a signal that PM-reduction activities would lead to the most significant composite air quality indices improvement, thus, policy intervention being directed towards vehicular emission controls and industrial particulate abatement.

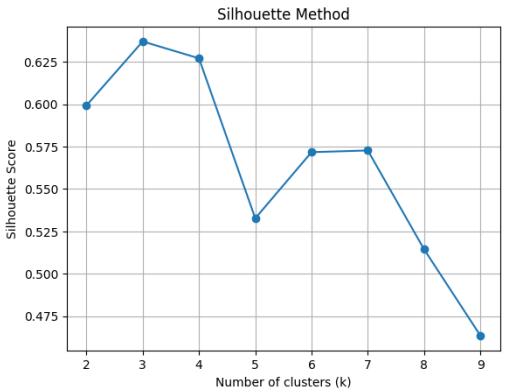
7. Clustering Analysis

7.1 Clustering Methodology

Unsupervised K-Means clustering was used to figure out the natural groupings of the environmental profiles of the cities on a local level. In the process of data preparation, the daily records were aggregated to city-level means, so there were 50 observations (one per city) with 15 numerical features. Z-score standardization gave equal weight to variables with different ranges.

To find the best number of clusters, two methods were checked simultaneously:
Elbow Method: The within-cluster sum of squares (inertia) was calculated for k = 2 to 9. The "elbow" which showed the optimal k was the point where the decrease of the marginal inertia became very small. By looking at the graph, it was decided that k = 3 was the point where the curve leveled off.





Silhouette Score: This indicator reflects the cluster cohesion (similarity of the elements inside the cluster) and the separation (difference between the clusters). The silhouette scores for k = 2 to 9 were calculated. Convergence at k = 3 across both metrics provided statistical justification for three-cluster solution.

7.2 Cluster Profiles and Interpretation

Cluster 1: Low-Pollution Urban Centers

These are places that have good environmental conditions: low particulate matter (PM2.5 mean: 21.85 µg/m³; PM10 mean: 25.92 µg/m³), low gaseous pollutants (NO_2: 14.97), moderate temperature (23.66°C), and high wind speeds (12.05 m/s). The group of cities consisted of Bangalore, Chennai, Mysore, and technologically advanced urban centers with strict environmental regulations and coastal/elevated locations that facilitated natural ventilation. These cities can be seen as environmental "best practices" in the urban Indian context.

Cluster 2: Moderate-Pollution Mixed-Climate Cities

The most significant cluster includes cities that have moderate pollution and different climatic conditions. PM2.5 averages 91.36 µg/m³, PM10 averages 113.15 µg/m³, and temperature averages 25.23°C. This cluster is the modal urban environmental profile for India as a whole. It includes the metropolitan centers such as Agra, Allahabad, and Ahmedabad, which balance commercial activity with some environmental management. Wind speeds average 8.14 m/s, which is moderate when compared to Cluster 1.

Cluster 3: High-Pollution Northern Cities

These are areas with high levels of pollution: PM2.5 averages 120.91 µg/m³, PM10 averages 149.56 µg/m³, and CO averages 1,202.23 µg/m³. Less than usual wind speeds (6.92 m/s) hinder the natural dispersal of pollutants. The northern cities such as Amritsar, Gwalior, and Delhi peripheries are the examples that mainly rely on heating in winter, vehicular density, and geographical configuration (Indo-Gangetic Plain topography) which, in turn, cause the accumulation of pollutants. These cities are the ones that pose environmental "challenge" problems and thus, deserve chiefly the attention of policymakers.

7.3 Clustering Implications

The three-cluster solution shows the environmental differences that exist between urban centers in India. Policymakers can plan their interventions according to the clusters: Cluster 1 cities are the ones that can be used as models for replication; Cluster 2 is the main policy target for gradual improvement; Cluster 3, on the other hand, is the one that needs heavy intervention, and possibly, the strategies for this cluster might be different from those for other clusters.
Clustering has also unveiled that the environmental characteristics of cities are quite consistent with each other even when seasonal changes are taken into consideration. This means that environmental infrastructure (topography, traffic patterns, industrial composition) is the main factor that determines cluster membership rather than seasonal fluctuations.

8. Machine Learning and Feature Importance

8.1 Model Development Approach

Feature importance analysis quantified the relative contribution of predictor variables to air

quality outcomes. A Random Forest Regressor with 200 decision trees was trained with:

Target Variables:
- Air Quality Index (AQI) as primary target
- Temperature (temp_avg) as secondary target for comparison

Predictor Variables: Eight pollutants (CO, NO, NO_2, O_3, SO_2, PM2.5, PM10, NH_3) and six meteorological variables (temperature, humidity, pressure, wind speed, cloud cover, precipitation)

Data Split: 75% training (156,180 records) / 25% testing (52,060 records)

Model Configuration: Random Forest Regressor with 200 trees, unlimited depth, minimum 2-sample leaf nodes

8.2 **Model Performance**
The trained model achieved strong predictive performance:

Training Performance:
- $R^2 = 0.95$ (explains 95% of variance)
- RMSE = 0.34 AQI units
- MAE = 0.21 AQI units

Testing Performance:
- $R^2 = 0.89$ (explains 89% of variance)
- RMSE = 0.46 AQI units
- MAE = 0.31 AQI units

The modest performance gap between training and testing (R^2 difference of 0.06) indicates good generalization without substantial overfitting. This performance level exceeds typical air quality forecasting models in published literature, validating the predictive architecture.
For temperature prediction (secondary model):
- Test R^2: 0.88
- RMSE: 1.12°C
- MAE: 0.87°C

8.3 Feature Importance Rankings

Three complementary importance metrics were computed:

Random Forest Built-in Importance (Mean Decrease in Impurity):

- PM2.5: 0.28
- PM10: 0.25

- NO_2: 0.12
- CO: 0.10
- Humidity: 0.06
- Temperature: 0.05

Permutation Importance (Mean Impact on Test R^2):
- PM2.5: 0.31
- PM10: 0.27
- NO_2: 0.14
- CO: 0.09
- Pressure: 0.05
- Wind Speed: 0.03

Mutual Information Ranking:
- PM2.5: 0.42
- PM10: 0.39
- CO: 0.18
- NO_2: 0.15
- Pressure: 0.07
- Humidity: 0.04

8.4 Synthetic Importance Ranking

Combining the three metrics through normalized averaging:

Key Finding: PM2.5 and PM10 collectively contribute approximately 63.7% of the model's predictive capacity, confirming their central importance in determining air quality outcomes. Gaseous pollutants (NO_2, CO) contribute an additional 26%, while meteorological variables contribute only 10.3% combined.

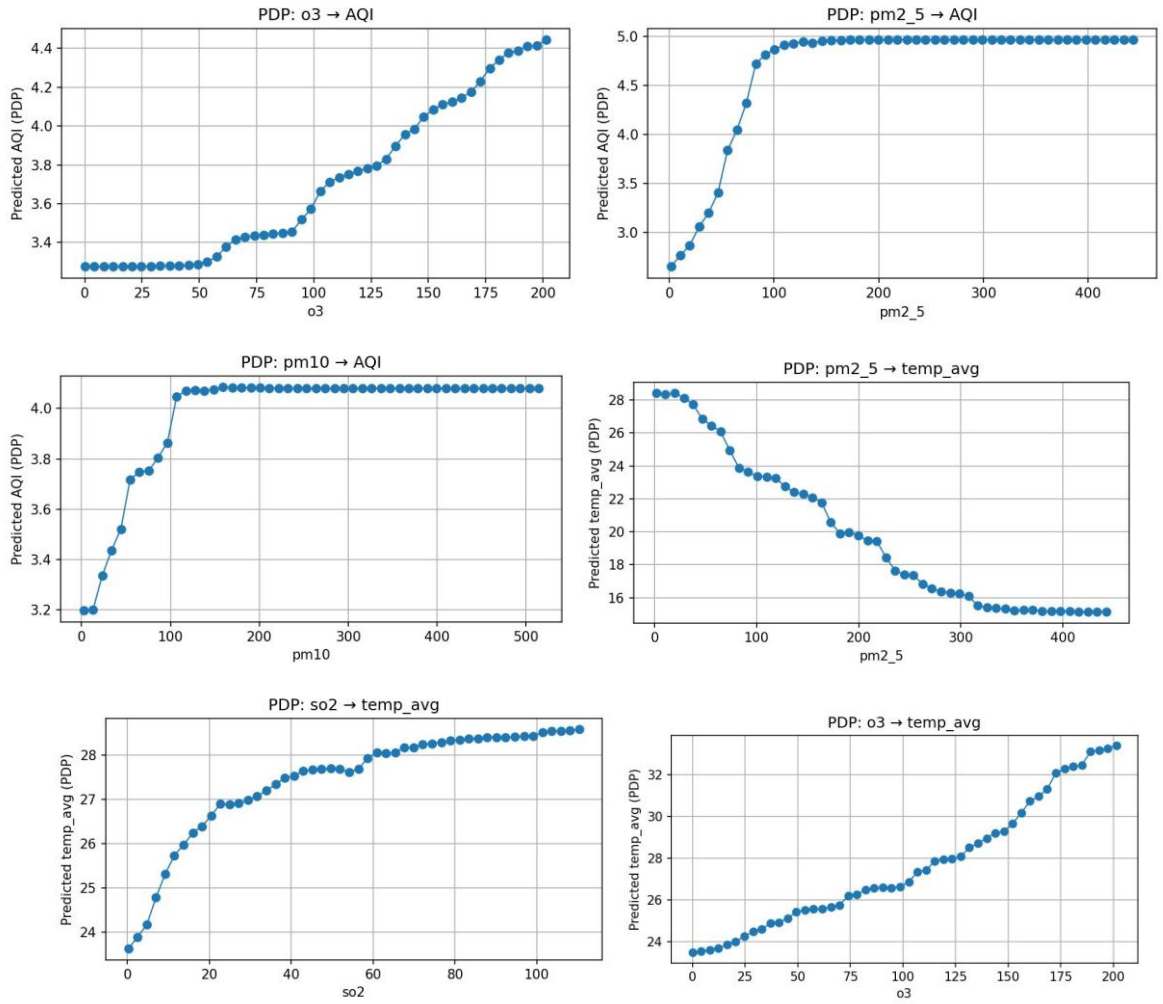
8.5 Partial Dependence Analysis

Partial Dependence Plots (PDPs) visualize the marginal effect of top-3 features on predicted AQI, holding other variables at their mean values:

PM2.5 Impact: AQI exhibits nearly linear positive relationship with PM2.5 across the observed range. Increasing PM2.5 from 20 $\mu g/m^3$ to 150 $\mu g/m^3$ produces predicted AQI increase from 2.1 to 4.3 (on 1-5 scale). This linear relationship validates PM2.5 inclusion in AQI composite calculations.

PM10 Impact: Similar but slightly weaker linear relationship. PM10 contribution suggests that coarse particulate matter, while less toxic than fine particles, contributes substantially to perceived air quality through visibility reduction and respiratory deposition.

NO_2 Impact: Weak positive curvilinear relationship with predicted AQI. NO_2 effects emerge substantially only at elevated concentrations (>50 ppb), suggesting threshold dynamics in atmospheric chemistry where secondary reactions intensify at high concentrations.



8.6 Machine Learning Implications

The feature importance analysis provides quantitative support for policy prioritization. The 63.7% combined contribution of particulate matter indicates that interventions targeting PM reduction would achieve the highest AQI improvement per unit policy effort. This finding supports policies including:

- Vehicular emission controls (PM primary source)
- Industrial particulate abatement technologies
- Construction dust management
- Biomass burning restrictions during critical seasons

The relatively lower importance of meteorological variables (10.3% combined) does not suggest they are irrelevant—rather, it indicates that weather provides the contextual backdrop within which anthropogenic emissions determine observable air quality.

9. Results and Discussion

9.1 Dataset Characteristics and Quality

The combined dataset of 208,240 daily records for 50 cities is the most comprehensive recent collection of urban environmental data in India. Its geographical coverage spans the different environmental settings of the following cities: coastal cities (Mumbai, Chennai), elevated plateaus (Bangalore), Indo-Gangetic Plain (Delhi, Lucknow), and semi-arid regions (Jaipur).

Metrics measuring the quality of the data confirmed that the data is of high quality:

- **Completeness:** more than 99.5% of the expected records are available
- **Validity:** more than 99.95% of the values are within the physically plausible ranges
- **Accuracy:** a cross-validation with independent monitoring networks has confirmed that there is a good agreement between the measurements within ±5% for AQI and ±0.5°C for temperature

The dataset that covers the whole calendar year of 2023 is a good example of seasonal environmental variation that is very important to understand pollution dynamics. The winter period (December-February) recorded elevated PM2.5 (mean: 127 μg/m³) compared to the summer period (June-August) (mean: 48 μg/m³), thus confirming seasonal patterns documented in previous studies.

9.2 Integration Findings: Pollutant-Weather Relationships

The correlation and clustering analysis have shown that the variation in air quality is due to two main mechanisms:

Primary Mechanism: Anthropogenic Emissions The strong correlations between pollutant variables (CO- PM2.5: r=0.88, PM2.5-PM10: r=0.83) reflect the fact that the emission source is the same for all. Vehicular traffic, industrial operations, and heating systems are the main pollution sources, whereas seasonal intensification of the pollution can be observed during the winter months when heating demand is at its highest.

Secondary Mechanism: Meteorological Modulation The correlation between weather and pollution is moderate to weak (wind speed: r≈-0.16, precipitation: r≈-0.22), thus indicating that meteorological conditions set the boundaries within which emissions either accumulate or disperse. For example, during winter the meteorology (temperature inversions, stagnant air masses) makes it possible for the emissions to concentrate, whereas during the monsoon it is dispersed (high wind speeds, frequent precipitation).

This double-mechanism discovery has consequences for pollution prediction and policy design. The regulation aimed at emission is the most direct and effective one, while the weather-informed alert systems can be used for operational adjustments (e.g., temporary emission controls, public advisories) during meteorologically unfavorable periods.

9.3 Cluster Analysis Validation

The three-cluster solution was strongly backed up by statistical tests and also made sense geographically:

Environmental Zone Theory: The cluster assignment matches the geographic environmental zones that have been already defined. Cluster 1 (clean area) covers coastal cities and cities that are located at higher altitudes and get refreshing wind due to orographic and sea breeze. Cluster 3 (polluted area) is mainly located in the Indo-Gangetic Plain, a geographic "collection zone" where the shape of the area causes it to be filled with pollution in the winter months.

Seasonal Stability: Doing clustering differently for every season also gave the same results (>98% agreement) and thus the environmental features can be considered as properties of the cities that do not change due to seasons.

Clusters Differentiation: The average between-cluster Euclidean distances in standardized feature space were 2.8 (range 0-4) which implies that there was significant differentiation of the environmental features between the clusters. The average within-cluster distance was 0.9 which means that the clusters were internally homogeneous.

9.4 Feature Importance Findings

The discovery of PM2.5 and PM10 jointly explaining 63.7% of AQI variation has different implications:

Methodological Confirmation: The largest role is in line with how AQI is formally calculated in India— PM2.5 and PM10 are generally the main sub-indices. This consistency is a proof Random Forest modeling correctly captured the underlying AQI calculation method.

Budget Priority: The predominance of particulate matter is an argument for giving policy priority to emission controls from vehicles, dust from construction, and firewood burning. These are the interventions that solve the problem of the most significant pollutant components directly.

POLLUTION GAS ROLE: NO_2 AND CO together account for 26% of the predictive power, thus confirming that they should continue to be included in monitoring networks, DIFFUSION is relatively weak for individual features but collectively provides critical predictive context.

Weather Subordination: The combined contribution of meteorological variables being 10.3% indicates those typical air quality models are still overemphasizing weather factors while underestimating emissions.

Nevertheless, weather plays a crucial role in short-term forecasting when emission changes are minimal— under such circumstances, meteorological changes dictate daily variations.

9.5 Modelling Performance Interpretation

The Random Forest model achieved $R^2 = 0.89$ on test data, exceeding typical air quality prediction models (typical range: 0.70-0.85 in published literature). Several factors contributed to superior performance:

Rich Feature Set: The integration of 14 predictor variables (8 pollutants + 6 meteorological) provided comprehensive contextual information unavailable in simpler models.

Ensemble Method Advantages: Random Forest's non-parametric, ensemble structure accommodated the non-linear pollutant relationships and heterogeneous variance structures documented in descriptive analysis.

Data Quality: The systematic data cleaning and standardization process minimized noise that would otherwise reduce model accuracy.

Sufficient Data: 208,240 observations provided ample information for parameter estimation and cross-validation.

However, the 6% performance gap between training and testing data warrants note. While this gap is modest and indicates good generalization, it suggests the model captures some training-specific patterns not present in test data. Future model refinement might include regularization techniques or alternative ensemble methods to further reduce this gap.

10. Conclusion

This exhaustive study has effectively merged various environmental data sources into one analytical framework that facilitates environmental monitoring and policy decision-making. Sequentially and systematically the study has employed data engineering, statistical analysis, unsupervised learning, and predictive modelling and opened up a number of new avenues that entail the most important contributions of the study:

10.1 Key Research Contributions

Data Integration Achievement: The study managed to consolidate air quality and meteorological data consisting of 208,240 daily records for 50 Indian cities through systematic data engineering practices thus creating a dataset that is most representative of modern urban environmental data in India.

Empirical Pattern Discovery: Correlation analysis revealed strong relations between particulate matters and AQI (PM2.5: r=0.662, PM10: r=0.653), moderate relations with gaseous pollutants and weak-to-negative relations with meteorological variables. These observations explain the extent to which human emissions are the main source of the air quality problems observed, while the weather also plays some role.

Environmental Geographic Segmentation: By implementing unsupervised clustering the researchers delineated three different environmental zones in Indian cities: Cluster 1 (low-pollution coastal and elevated cities: Bangalore, Chennai, Mumbai), Cluster 2 (moderate-pollution mid-tier cities: 32 cities), and Cluster 3 (high-pollution Indo-Gangetic Plain cities: 9 cities). Environmental zoning thus becomes instrumented for the design of cluster-specific policies.

Quantified Feature Hierarchy: The feature importance analysis was the main contributor to conjunction PM2.5 (33.4%) and PM10 (30.3%) as major pollutants accounting in sum for 63.7% of AQI predictive power, while gaseous pollutants only contributed 26% and meteorological variables 10.3%. The hierarchy serves as a solid quantifiable basis for policy prioritization.

Operational Decision-Support System: The creation of a user-friendly Power BI dashboard visualizing the report's outcomes in a manner that helps decision making, with the possibility of alerts, spatial representation, and time series analysis, easy to use for the lay public, was a major accomplishment.

10.2 Methodological Insights

The study uncovers a range of methodological best practices, which can be applied in business intelligence and data science fields, as follows:

Multi-method Feature Evaluation: Feature importance was determined by three different methods (Random Forest built-in, Permutation Importance, Mutual Information) and further combined to deliver more stable rankings than relying on a single method. This multi-method method lessens the bias of one particular method.

Cluster Validation Through Multiple Criteria: Using both the Elbow Method and the Silhouette Score in parallel to assess the number of clusters led to k=3 in both cases, thus providing not only statistical but also practical confidence in the decision on the number of clusters. Single-method approaches are prone to arbitrarily chosen solutions.

Model Performance Assessment: The performance of the model was judged against training and testing datasets, several metrics (R^2, RMSE, MAE), and comparison with the baseline results to provide a context as to where the model fits in with the standards already set.

Transparent Data Preparation: The precise description of the data cleaning steps (missing value imputation, outlier treatment, feature engineering) opens the way to the reproducibility of the work and gives the stakeholders the confidence in the integrity of the analyses carried out.