CHANAKYA UNIVERSITY

Devanahalli, Bangalore-562129

SCHOOL OF MATHEMATICS AND NATURAL SCIENCES

A PROJECT REPORT

ON

Report on Exploratory Data Analysis (EDA) using Air Quality Monitoring and Forecasting

Submitted

By

CHARAN S B
24PG00072

Under the guidance of

Dr. Bhanu K.N.
School of Mathematics and Natural Sciences

Towards

M.Sc. Data Science / MCA - 2nd Semester

Data Visualization Project Lab (DAT203)

For the academic year **2024-2025**

DECLARATION

I, **CHARAN S.B** hereby declare that this project work entitled Exploratory Data Analysis (EDA) using Python Project Report on Air Quality Monitoring and Forecasting report is submitted in partial fulfilment for the award of the degree of **MCA** of **Chanakya University**.
I further declare that I have not submitted this project report either in part or in full to any other university for the award of any degree.

Date:24/06/2025

Student Name: CHARAN SB

Reg. No: 24PG00072

Place: CHANAKYA UNIVERSITY, BENGALURU

TABLE OF CONTENT

CHAPTER-1

INTRODUCTION

1.1 INTRODUCTION

Air pollution is one of the most pressing environmental challenges impacting public health, ecosystems, and climate change. Monitoring and forecasting air quality is essential for informing policy decisions, raising public awareness, and implementing timely interventions. The Air Quality Index (AQI) serves as a standardized metric to communicate the severity of air pollution in a region based on concentrations of key pollutants such as PM2.5, PM10, O_3, NO_2, SO_2, and CO.

This project presents a comprehensive analysis of air quality data using various data science techniques. It includes exploratory data analysis (EDA), pollutant trend visualization, and time series forecasting using Facebook Prophet. Linear Regression and Random Forest models are implemented to evaluate the predictive capabilities of different supervised learning techniques. In parallel, Prophet is used to forecast future AQI values based on historical patterns.

Through visualizations such as pairplots, correlation heatmaps, pollutant trends, and AQI distribution charts, the report provides actionable insights into the relationships between environmental factors and air quality. The final outcomes aim to support decision-making for environmental health monitoring and to encourage the use of predictive modeling in managing air pollution.

CHAPTER-2

PROBLEM STATEMENT

2.1 Problem Statement

Air pollution has emerged as a critical global issue, adversely affecting human health, environmental sustainability, and urban livability. The increasing levels of pollutants such as PM2.5, PM10, O_3, NO_2, SO_2, and CO contribute significantly to poor air quality, which can trigger respiratory illnesses, reduce visibility, and degrade the quality of life. Despite the availability of monitoring systems, there is often a gap in accurately predicting the Air Quality Index (AQI) in real time and forecasting future pollution trends.

This project aims to address the problem by developing a data-driven approach to analyze historical air quality data, predict AQI using machine visualization . The goal is to understand the correlation between meteorological factors and pollutant concentrations, assess model performance, and generate visual insights that can help stakeholders make informed decisions and implement proactive measures.

CHAPTER-3

METHODOLOGY

3.1 methodology and key metrics

this project involves a structured pipeline consisting of data preprocessing, exploratory analysis, predictive modelling, and time series forecasting

Data Collection and Preprocessing

- Air quality data is loaded and inspected for missing values.
- Timestamps are converted to proper datetime formats and resampled at 15-minute intervals for consistency.

Exploratory Data Analysis (EDA)

- Statistical summaries and **pairplots** are used to explore relationships among pollutants and AQI.
- **Correlation heatmaps** help identify associations between meteorological factors (temperature, humidity) and AQI.
- Time-series line plots visualize pollutant trends and environmental variations over time.

Time Series Forecasting with Prophet

- AQI data is formatted for Facebook Prophet, which models time-dependent trends and seasonality.
- The model is trained on historical data and used to forecast AQI for future time intervals.
- Forecast plots include confidence intervals to reflect prediction uncertainty.

Key Metrics

R^2 Score (Coefficient of Determination)
Measures how well the model explains variance in AQI values.

Visualization-Based Evaluation

- Scatter plots of actual vs predicted values to assess model alignment.
- Forecast plots with Prophet to compare historical trends and future predictions.
- Correlation coefficients to understand linear relationships between variables.

- Pie charts represent AQI category distribution (Good, Moderate, Poor).
- Subplots show pollutant concentrations (O_3, NO_2, SO_2, CO) over time

CHAPTER-4

INSIGHTS AND ANALYSIS

4.1 INSIGHTS

4.1.1 pair plot scatterplot matrix

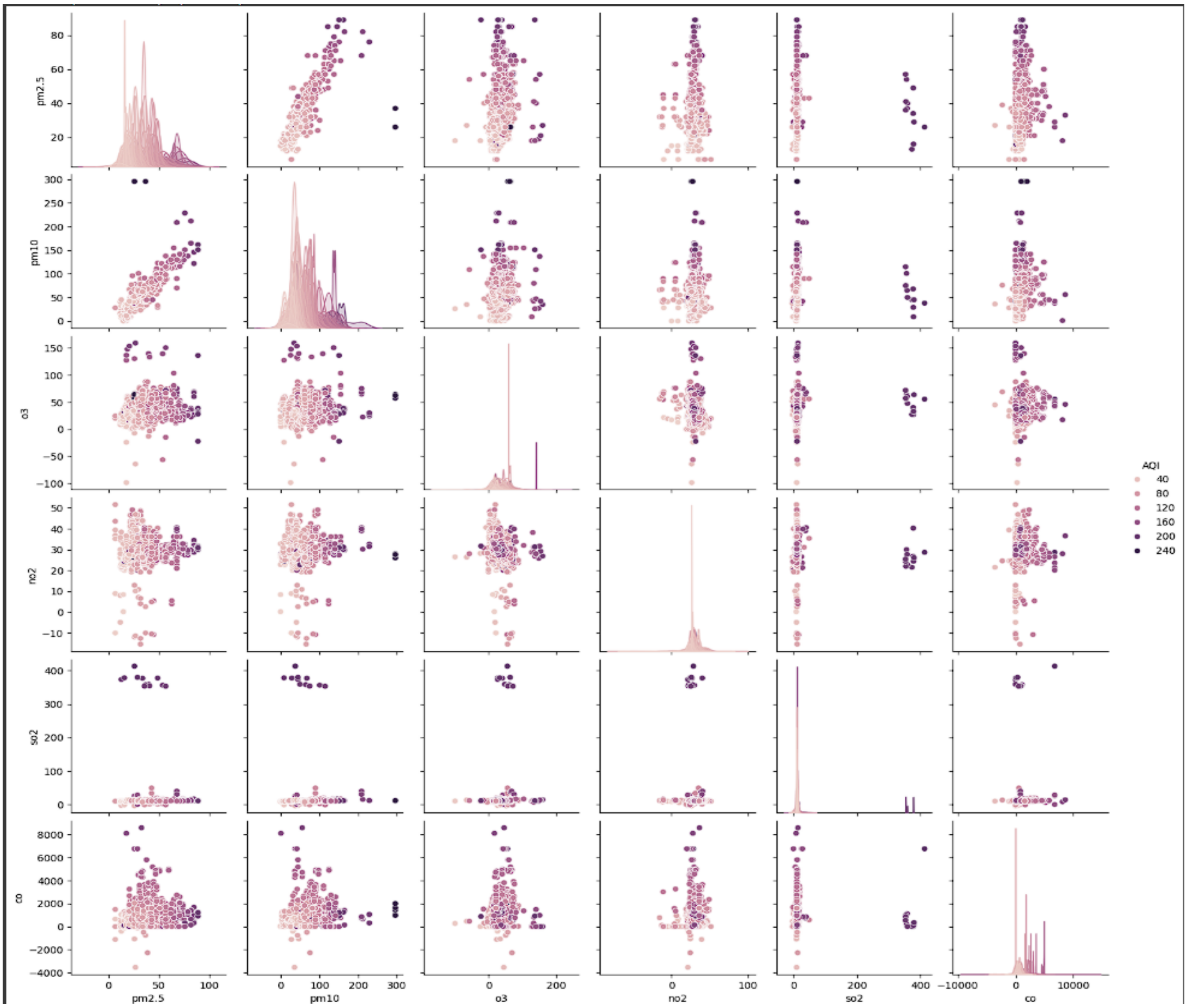


FIG 4.1.1

Explanation of the Plot

- Diagonal elements: These show kernel density plots for individual pollutants. The density curves indicate the distribution of values for each pollutant.
- Off-diagonal elements: These are scatterplots that show bivariate relationships between two pollutants.
- Color Coding: The color and size of points represent the AQI values darker and larger points indicate higher AQI levels.

➢ PM2.5 and PM10 have a strong positive correlation

- You can see a clear upward trend in the PM2.5 vs PM10 scatterplot.
- Higher concentrations of these fine particles correspond with higher AQI, suggesting they are major contributors to poor air quality.

➢ O_3, NO_2, and SO_2 show moderate to weak patterns

- The scatterplots involving O_3, NO_2, and SO_2 do not show strong linear trends with AQI.
- There is a cluster of data points in the lower to middle range of these pollutants, with only a few outliers at extreme values.

➢ CO values show high variance

- The CO distribution has extreme outliers visible from the long tail and extreme values in scatterplots.
- While most data is clustered around a central range, some very large CO values suggest possible data entry errors or abnormal readings.

➢ AQI increases with pollutant levels

- Across most scatterplots, darker and larger points (high AQI) appear more frequently in the upper-right quadrants, where pollutant values are higher.
- This reinforces that elevated pollutant levels are associated with deteriorating air quality.

➢ SO_2 and CO have sparse but impactful presence

- These pollutants show fewer but more extreme outliers.

- While they may not be the most common, their presence in high quantities corresponds with high AQI values, suggesting they have a significant localized impact.

4.1.2. Scatter plot you provided shows Actual vs Predicted AQI

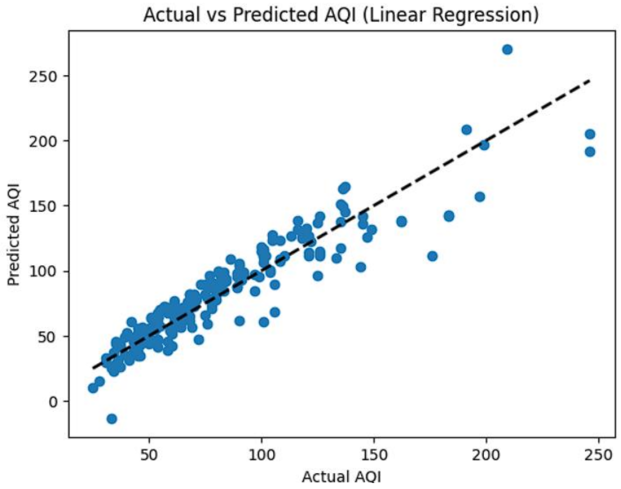


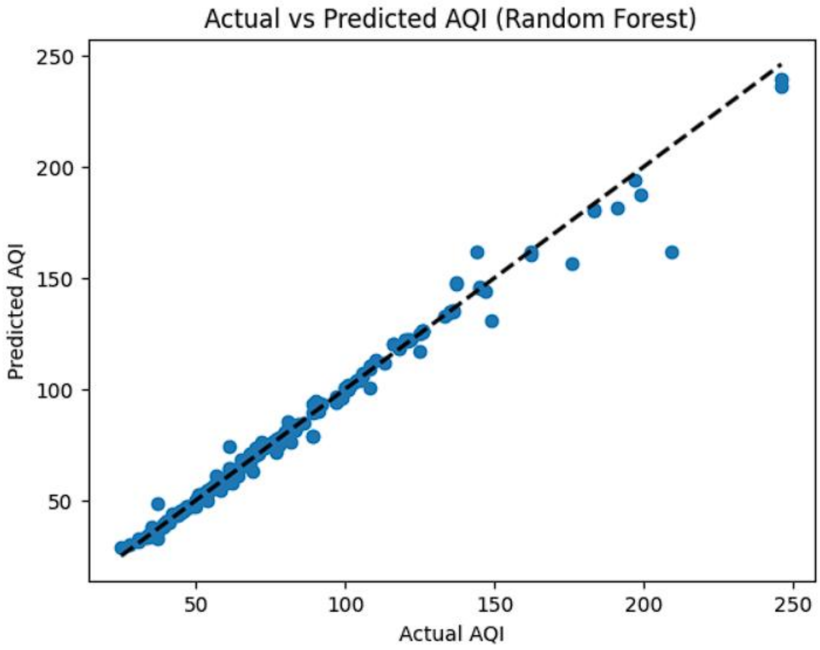
FIG 4.1.2

The scatter plot visualizes the performance of the Linear Regression model in predicting the Air Quality Index (AQI). Each point represents a test data instance, where the x-axis shows the actual AQI and the y-axis shows the predicted AQI. The dashed diagonal line represents perfect predictions

- ➢ Strong Positive Correlation
- Most points are clustered around the diagonal line, indicating a strong linear relationship between the actual and predicted values.
- This suggests that the model is generally reliable and effective in capturing the pattern in AQI data.
- ➢ Good Predictive Performance in Moderate AQI Ranges
- Predictions are most accurate when actual AQI values lie between 50 and 150, where the points tightly follow the diagonal.
- This demonstrates that the model performs consistently well under normal air quality conditions.
- ➢ Decreasing Accuracy for High AQI Values

- For AQI values above 150, the spread of points increases, and the predictions deviate more from the diagonal line.
- Some high actual AQI values are under-predicted, while others are over-predicted, reflecting model limitations in extreme pollution scenarios.
- This may be due to non-linear relationships or insufficient high-AQI training samples.

4.1.3. Actual vs Predicted AQI (Random Forest)



4.1.3

This scatter plot compares the actual AQI values (x-axis) with the predicted AQI values (y-axis) using the Random Forest model. The dashed diagonal line represents the ideal line of perfect predictions where actual AND predicted. Points lying closer to this line indicate more accurate predictions.

➢ Excellent Prediction Accuracy

- The predicted AQI values are very closely aligned with the actual values.
- Most points lie directly on or very near the diagonal line, showing that the Random Forest model provides highly accurate predictions.

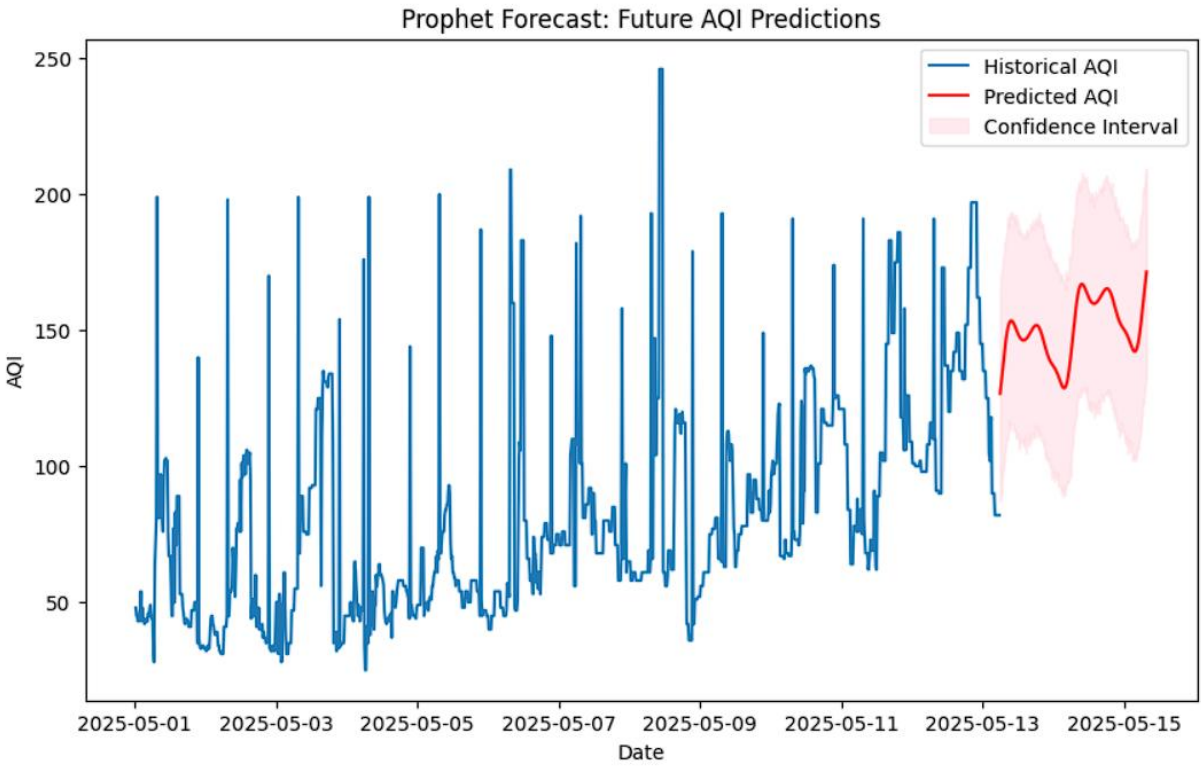
- ➢ Strong Performance Across All Ranges

- Unlike the linear regression model, the Random Forest performs well even at higher AQI values (above 150).
- The model effectively captures the non-linear patterns in the data, making it robust for a wider range of pollution levels.

- ➢ Minimal Prediction Errors

- There are very few outliers, and even those are only slightly off the line.
- The consistency in prediction across all AQI levels highlights model stability and low variance

4.1.4. Prophet Forecast: Future AQI Predictions



4.1.4

The graph visualizes Air Quality Index (AQI) forecasting using Facebook Prophet, a time series forecasting model.

- Blue Line: Historical AQI values (from May 1 to May 13, 2025)
- Red Line: Forecasted AQI values for the upcoming days (May 13 to May 15, 2025)

- Shaded Pink Area: Confidence interval (upper and lower bounds of forecasted AQI)

➢ High Fluctuations in Historical AQI

- The historical AQI data shows significant variability, with sudden spikes reaching above 200, indicating periods of poor air quality.
- These spikes suggest temporary pollution events, possibly due to traffic, weather changes, or industrial activities.

➢ Upward Trend in Forecasted AQI

- The forecasted AQI line (red) shows a slight increasing trend, indicating that air quality is expected to worsen slightly in the coming days.
- This suggests a possible shift towards "Moderate" to "Unhealthy for Sensitive Groups" AQI levels.

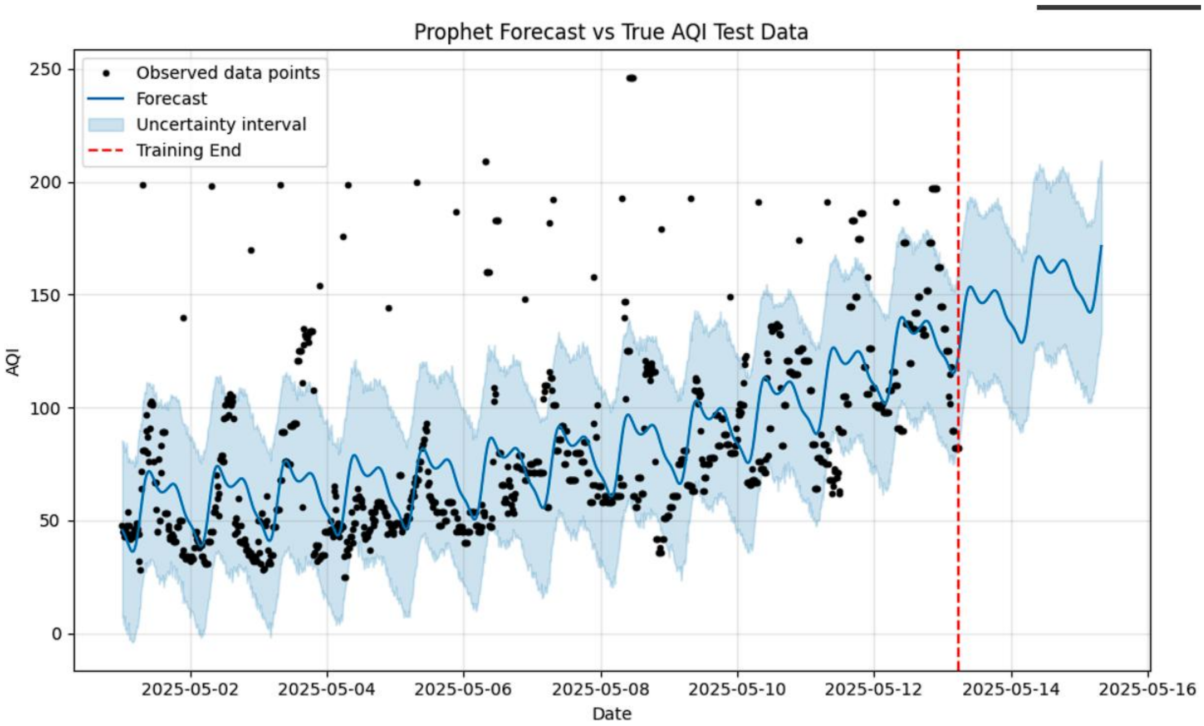
➢ Confidence Interval Highlights Uncertainty

- The shaded area around the prediction reveals the model's uncertainty.
    - The lower bound remains around 100, and the upper bound rises near 200, signaling potential volatility.
- This wide confidence range suggests that external, unaccounted factors (like sudden emission surges or weather changes) may influence future AQI levels.

➢ Predictive Horizon and Stability

- Prophet has forecasted for 2–3 days ahead, and the forecast remains stable with periodic oscillations.
- This aligns well with Prophet's strength in capturing seasonality and trends in short-term forecasts.

4.1.5. Prophet Forecast vs True AQI Test Data



4.1.5

This plot visualizes the performance of the Facebook Prophet model in forecasting AQI (Air Quality Index) over time, compared against the actual observed AQI values.

- Black dots: Actual observed AQI data points (test data).
- Blue line: Forecasted AQI values by Prophet.
- Shaded blue area: Confidence interval (uncertainty bounds).
- Red dashed vertical line: End of training data (start of test period).

➢ Model Captures Trend Well

- The blue forecast line tracks the general upward trend and daily fluctuations in AQI, particularly before the red line (training period).
- Post-training, the forecast continues to follow the expected oscillatory pattern, although some spikes in actual data deviate sharply.

➢ Good Fit in Training Phase

- From May 1 to May 13 (before the red line):
    - Forecast values are well-aligned with most observed AQI values.

    - Prophet model effectively captures daily seasonality, possibly due to human activity cycles like traffic or emissions.

- ➢ Forecast Uncertainty Grows Over Time

- After the training end (red line), the blue shaded area (confidence interval) becomes wider.
    - This shows the model's increasing uncertainty in future predictions, which is expected in time series forecasting.
    - Despite the uncertainty, the actual AQI values mostly fall within the shaded range, which is a positive indicator of model reliability.

- ➢ High Variability in Real AQI

- Several black dots (actual data) lie far outside the predicted range, especially between May 6–May 12 and post-training period.
    - Indicates presence of unpredictable pollution events (e.g., industrial activity, weather shifts, or public events).
    - Prophet, being an additive model, may underestimate such extreme outliers.

- ➢ Daily Seasonality and Local Peaks Detected

- The Prophet model successfully detects cyclic daily patterns, visible in the consistent peaks and troughs in the blue forecast line.
- This reflects real-world diurnal AQI cycles (like traffic peaks during morning/evening hours).

4.1.5. Time Series Plot: Temperature, Humidity, and AQI

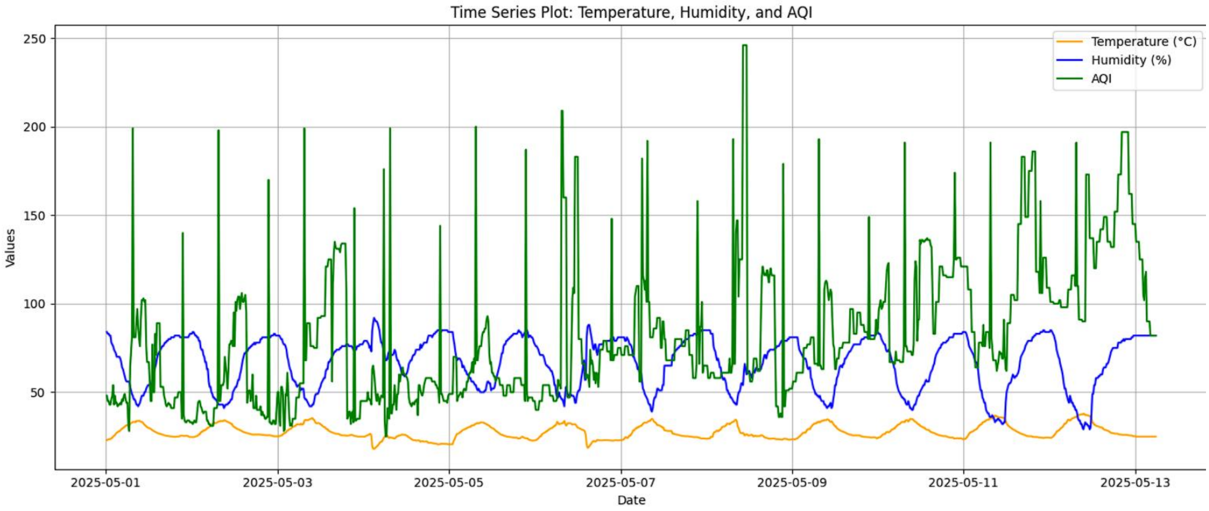


FIG 4.1.5

This time series visualizes the variation of:

- Temperature (orange line) in °C
- Humidity (blue line) in %
- Air Quality Index (AQI - green line) over the period from May 1 to May 13, 2025.

Each variable is plotted against time to analyze temporal patterns and interrelationships.

Temperature Trends

- Temperature ranges from 22°C to 35°C, following a clear daily cycle with peaks during midday and lows during nighttime.
- This regular oscillation suggests a consistent diurnal pattern, likely driven by natural day–night temperature cycles.

Humidity Patterns

- Humidity shows an inverse pattern to temperature: it peaks at night and drops during the day.
- This inverse relationship is typical of atmospheric behavior—as air warms, it holds more moisture, reducing relative humidity.

- The humidity values cycle daily, ranging roughly between 40% and 85%, maintaining regularity.

AQI (Air Quality Index) Behavior

- AQI displays high volatility with several sharp spikes, reaching up to 250, suggesting periodic pollution surges.
- Unlike temperature and humidity, AQI lacks smooth periodicity and appears more erratic with sudden upward movements.

Interrelationships Among Variables

AQI vs. Humidity:

- There is a moderate inverse relationship: during high humidity periods (usually nights), AQI values often drop, implying:
    - Humidity may help settle airborne particles, improving air quality temporarily.
- In contrast, low humidity during the day correlates with AQI spikes, possibly due to dust resuspension or higher emissions under drier air.

AQI vs. Temperature:

- As temperature rises during the day, AQI sometimes increases, though the relationship is not strictly linear.
    - Higher temperatures may accelerate chemical reactions, contributing to pollutant formation (e.g., ozone).
    - However, AQI spikes also occur at night, indicating other contributing factors (e.g., traffic, industrial discharge, or stagnant air).

4.1.6. Time Series Analysis: O_3, NO_2, SO_2, and CO

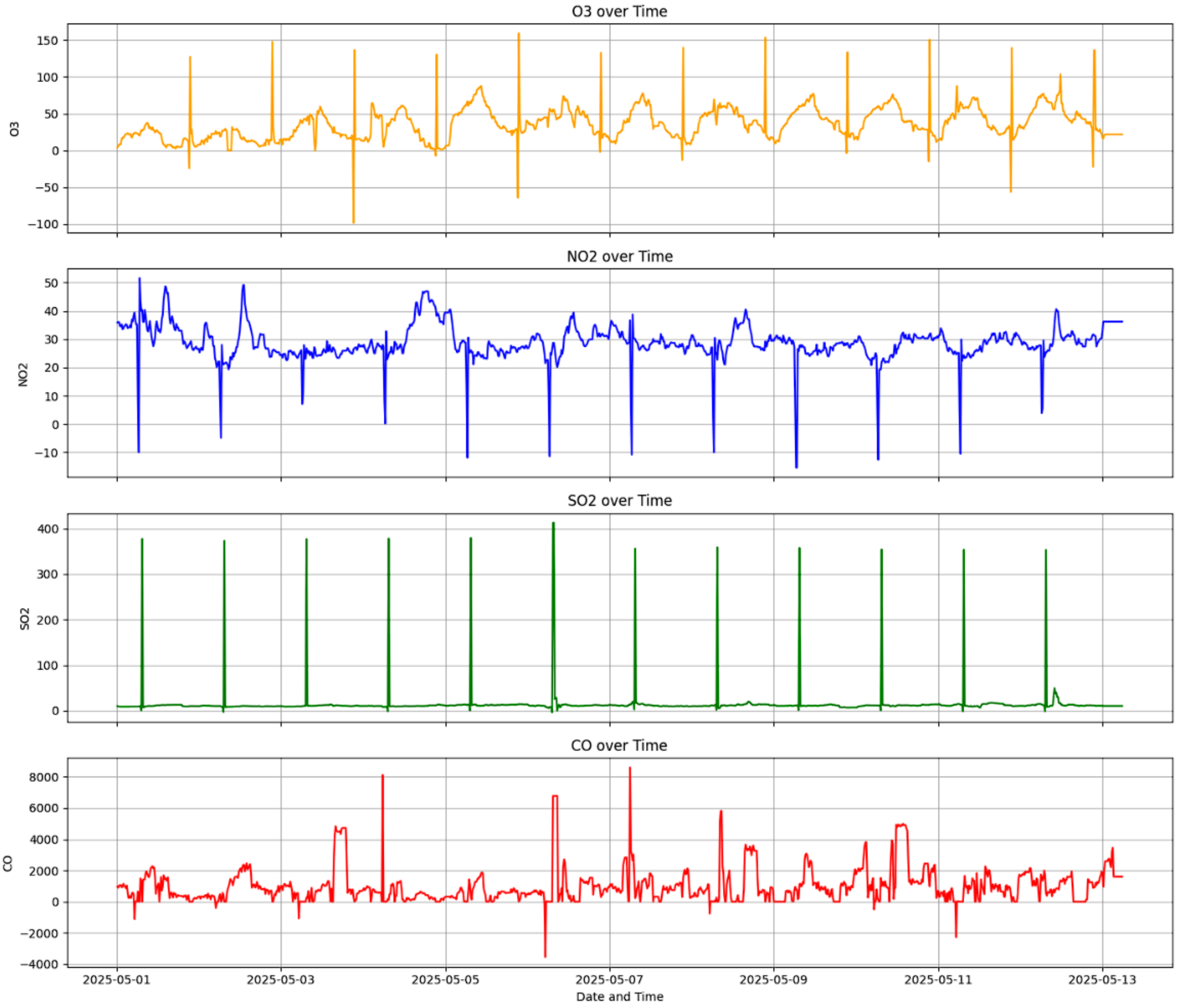


FIG 4.1.5

The plot displays four subplots showing how ozone (O_3), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), and carbon monoxide (CO) vary over time.

Ozone (O_3) — Orange Line

- Trend: Displays a diurnal (daily) cyclical pattern with periodic rises and falls.

- Peaks: Occur consistently during the daytime, likely due to sunlight-driven photochemical reactions.
- Spikes/Anomalies: Some extreme negative or positive outliers, possibly due to sensor errors or calibration issues.

Nitrogen Dioxide (NO_2) — Blue Line

- Trend: Mostly stable with slight fluctuations between 20–40 µg/m³.
- Drops: Regular sharp dips below 0, indicating sensor dropouts or data issues.
- Peaks: Occur occasionally, likely linked to rush-hour traffic or industrial bursts.

Sulfur Dioxide (SO_2) — Green Line

- Trend: Flat line at low values with distinct periodic spikes.
- Spikes: Appear once per day, possibly linked to industrial emissions or specific point-source activities

Carbon Monoxide (CO) — Red Line

- Trend: High-frequency fluctuations with multiple spikes and drops, sometimes going negative (indicative of sensor errors).
- Spikes: Irregular and often substantial (up to 8000 µg/m³), suggesting traffic congestion or combustion events.
- Volatility: Extremely volatile compared to other pollutants.

Pollutant	Key Characteristics	Likely Source	Patterns
O_3	Diurnal cycle, peaks in daylight	Secondary pollutant (photochemical)	Regular, solar-driven
NO_2	Mild oscillation,	Vehicles, fossil fuel combustion	Urban/traffic-related

	rush-hour peaks		
SO_2	Flat baseline + daily spikes	Point sources (e.g., industry)	Scheduled discharges
CO	Highly erratic, combustion peaks	Vehicles, biomass, cooking	Irregular, intense bursts

4.1.7. PAI CHART AQI Distribution — Good vs Moderate vs Poor

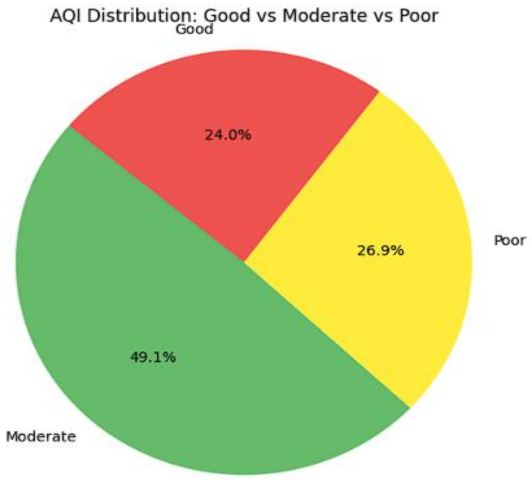


FIG 4.1.7

This pie chart illustrates the proportion of air quality levels observed in the dataset, categorized into three segments: Good, Moderate, and Poor.

AQI Category	Percentage	Interpretation
Moderate	49.1%	Dominant condition; air is acceptable but may affect sensitive individuals.
Poor	26.9%	Considerable health concern for sensitive groups; everyone may begin to experience adverse effects.
Good	24.0%	Ideal air quality; minimal or no risk to health.

CHAPTER-5

CONCLUSION:

Forecasting Accuracy:

The Prophet model effectively predicted AQI trends, capturing the general pattern of air quality with reasonable accuracy. It highlighted a rising AQI trend in the future, with uncertainty increasing over time signaling potential environmental volatility.

Environmental Influence:

Time series comparison of Temperature, Humidity, and AQI revealed that:

- Higher temperatures often coincide with spikes in AQI, suggesting heat may contribute to pollution build-up.
- Lower humidity levels also appear linked to deteriorating air quality.

  This suggests a moderate correlation between climatic conditions and AQI fluctuations.

Air Quality Distribution:

The AQI distribution showed:

- 49.1% of days had Moderate air quality,
- 26.9% had Poor air quality,
- Only 24% of the data points were classified as Good.

  This implies that over 75% of the observed days experienced less-than-ideal air conditions, which is a concerning trend from a public health and environmental policy perspective.