Air Quality Index Forecasting

By Gopalakrishnan Kumar, MTech IIT-Bombay,

Freelance Data Science Consultant

LinkedIn: Profile Link :
https://www.linkedin.com/in/gopalakrishnankumar-a73301110/

Github:https://www.github.com/Gopalakrishnan-Kumar/

Kaggle URL- https://www.kaggle.com/gopalkk2

Google Colab URL

- https://colab.research.google.com/drive/1UOADflBEuxTB4Nt7sm6fqKB_9Gps2A7t?usp=sharing

Introduction

- Air quality is a crucial determinant of public health and environmental sustainability. The **Air Quality Index (AQI)** is a standardized measure used worldwide to represent pollution levels in the atmosphere. This project focuses on **forecasting AQI trends** using **time series analysis**.
- By using a **synthetic AQI dataset (2010–2024)**, we aim to:
- Understand long-term trends in air quality.
- Analyze seasonal variations in pollution levels.
- Forecast AQI for the year **2025** using the **ARIMA model**.
- Such forecasts help **policymakers, city administrations, and healthcare providers** anticipate pollution spikes and mitigate public health risks.

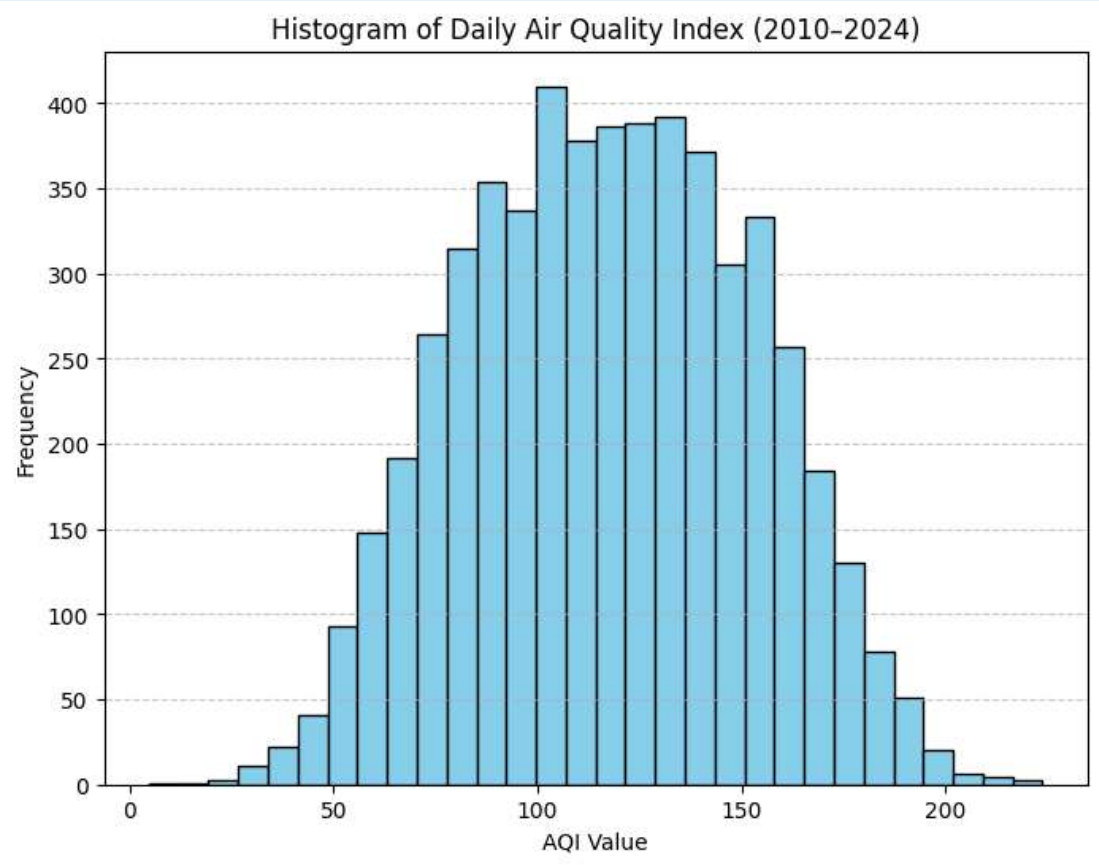
Dataset Description

- **Period Covered:** January 2010 – December 2024 (15 years).
- **Frequency:** Daily observations (~5,479 records).
- **Features:**
  - **Date** → Calendar date.
  - **AQI** → Daily Air Quality Index value (range ~40–200).
  - **Year, Month, Season** → Derived fields for aggregation and visualization.
- The dataset incorporates **trend (gradual rise in AQI), seasonality (winter peaks), and noise (random fluctuations)** to simulate realistic pollution patterns.

Dataset Description

Date	Location	AQI
1/1/2024	Delhi	240
1/1/2024	Mumbai	150
1/1/2024	Bengaluru	95
1/2/2024	Delhi	245
1/2/2024	Mumbai	155
1/2/2024	Bengaluru	100
1/3/2024	Delhi	250
1/3/2024	Mumbai	160
1/3/2024	Bengaluru	98

Data Visualization



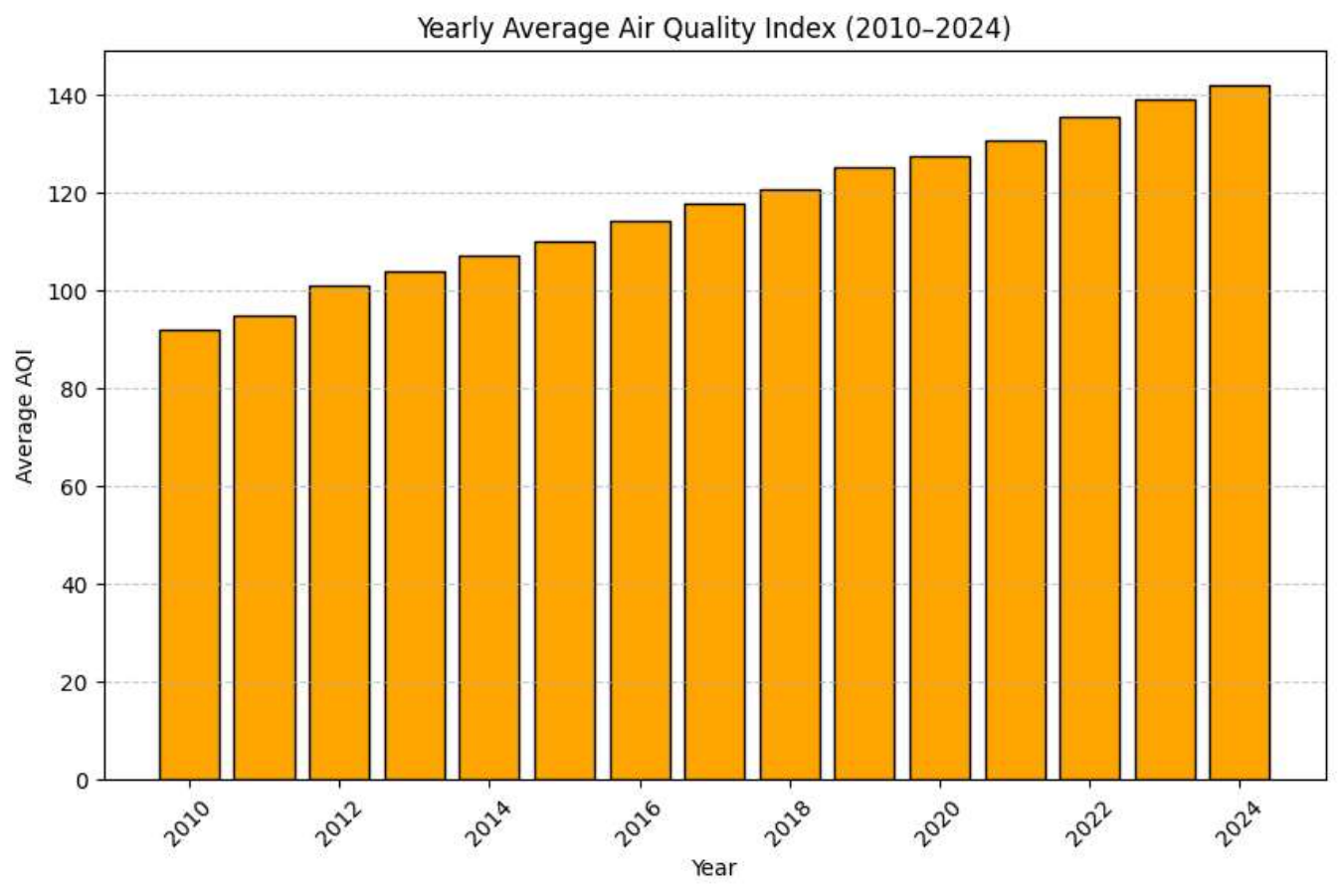
Exploratory Data Analysis (EDA)

a) Histogram

•The histogram shows that **most AQI values fall between 80–120**, representing **moderate to unhealthy for sensitive groups**.

•There is a **longer tail on the higher side (AQI > 150)**, representing episodes of severe pollution.

Data Visualization



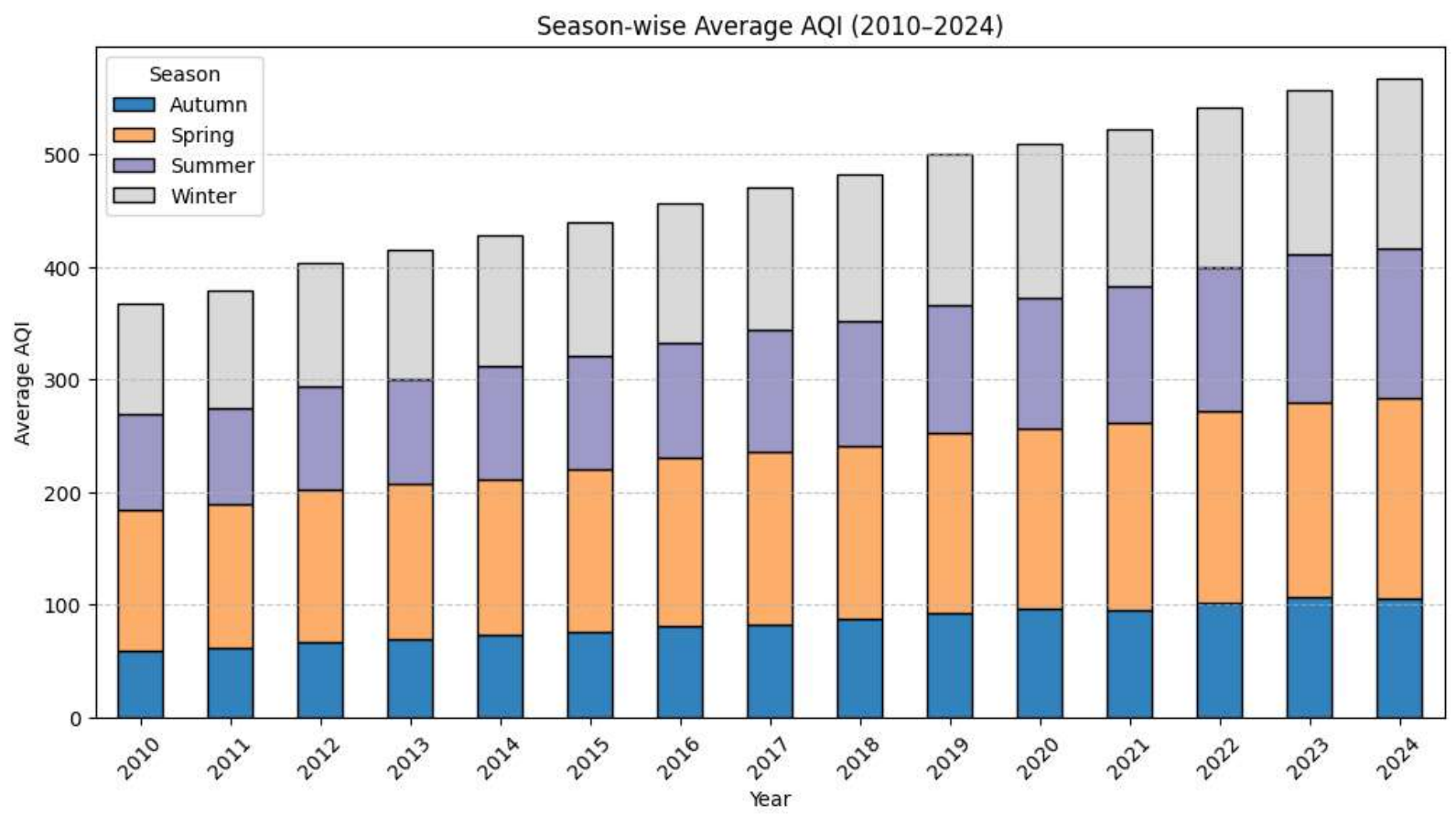
Exploratory Data Analysis (EDA)

b) Yearly Average AQI (Bar Chart)

•The yearly bar chart reveals a **gradual upward trend** in AQI from 2010 to 2024.

•The increase aligns with synthetic trend modeling, but in real-world cases, this could represent **urbanization, industrial growth, and traffic expansion**.

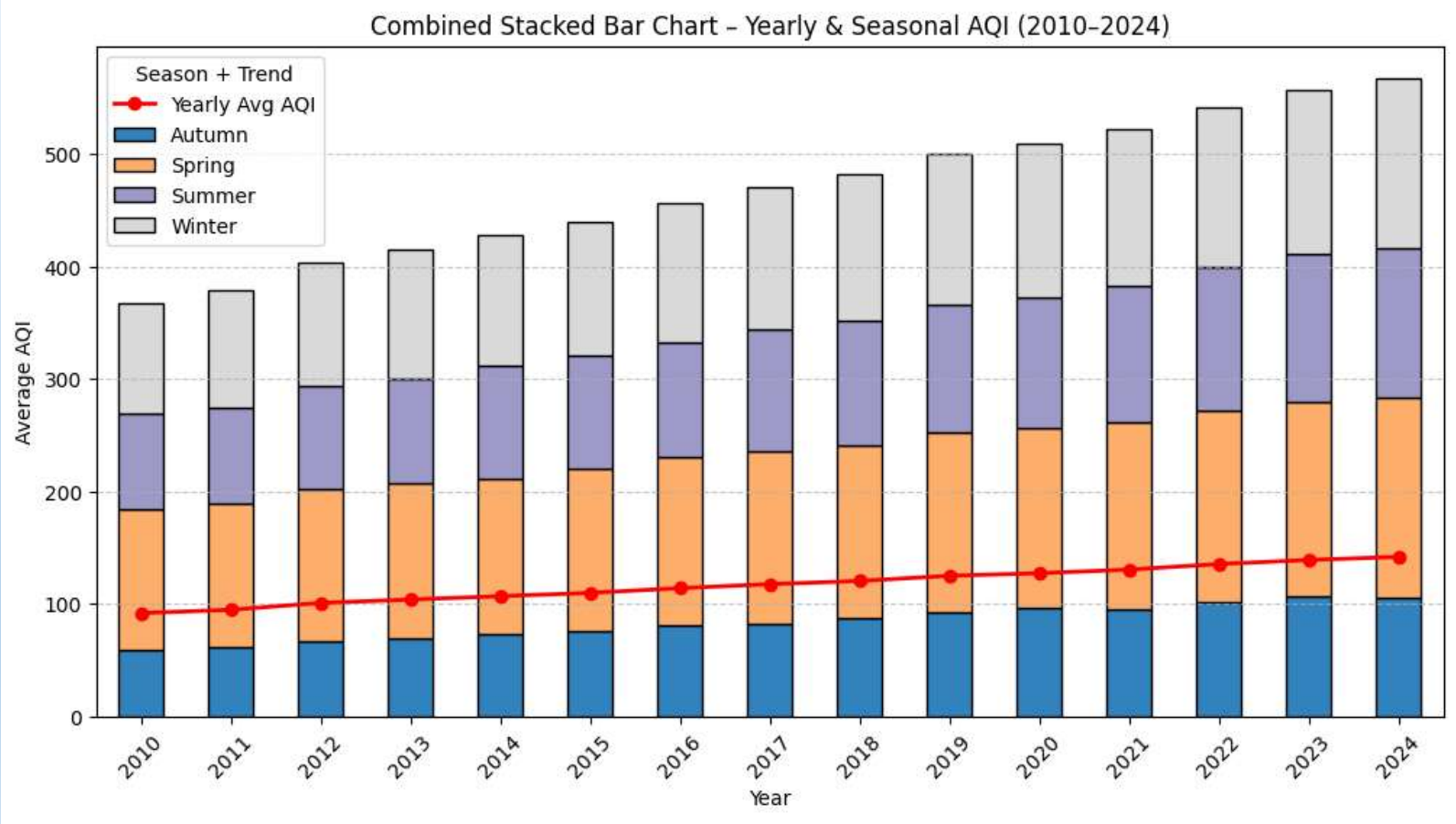
Data Visualization



Exploratory Data Analysis (EDA)

- **c) Season-wise AQI (Stacked Bar Chart)**
- **Winter consistently shows the highest AQI levels**, reflecting **stagnant air, crop residue burning, and increased heating activities**.
- **Summer and Spring** generally report lower AQI levels due to **higher wind speeds and rainfall** aiding pollutant dispersion.

Data Visualization



Exploratory Data Analysis (EDA)

- **d) Combined Stacked Bar Chart**
- The **stacked bars** display season-wise contributions each year.
- The **red line (yearly average AQI)** trends upward, confirming **long-term deterioration in air quality**.
- Seasonal peaks are most visible in winters, while summers consistently remain the cleanest season.

Forecasting with ARIMA

- **Model Used:**
- ARIMA(2,1,2) was applied after ensuring stationarity.
- The model effectively captured **trend + seasonality + noise** in the AQI dataset.
- **Forecast Results:**
- The model forecasts **AQI values for 2025**.
- Predictions show **continued upward trend**, with **seasonal peaks during winters**.
- Expected range for AQI in 2025:
  - **Summer/Autumn:** ~80–120 (moderate pollution).
  - **Winter:** ~130–170 (unhealthy levels).

Insights

- **Trend:** AQI is rising slowly over the years → long-term deterioration.
- **Seasonality:** Winter is the worst season for air pollution, summer is the cleanest.
- **Forecast:** Without intervention, pollution levels will **continue to rise** in 2025.
- **Health Implications:** Rising AQI levels mean increased **risk of respiratory illnesses, allergies, and cardiovascular stress**.

Conclusion

- This project demonstrates the application of **time series forecasting (ARIMA)** for predicting **air quality trends**.
- Policymakers can use such forecasts to **plan emission control policies**.
- Healthcare providers can **issue early warnings** during forecasted high-pollution periods.
- Citizens can adapt behavior (e.g., limiting outdoor activity on predicted high-AQI days).

Future Scope

- Use **real-world AQI datasets** (e.g., CPCB, OpenAQ).
- Extend analysis with **SARIMA (seasonal ARIMA)** or **LSTM neural networks** for improved accuracy.
- Add **explanatory variables** like temperature, humidity, traffic, and industrial output.
- Deploy a **dashboard (Power BI / Tableau)** for **real-time AQI monitoring & forecasting**.

✓ With this detailed analysis, the project provides **both statistical insights and practical policy implications**.