# Phase-2 Data Science Project Report

## 1. Project Title & Objective

**Title:** Analysis of Air Quality Index (AQI) Data
**Objective:** To analyze AQI levels across cities and identify trends and the most polluted areas using data visualization and statistical techniques.

## 2. Dataset Overview

The dataset contains AQI data for 5,377 city-month combinations. It includes monthly AQI values and an average AQI value per city.

Sample Data (First 5 rows):

| | rank | city | avg | jan | feb | mar | apr | may | jun | jul | aug | sep | oct | nov | dec |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | Begusarai, India | 223 | 413 | 337 | 250 | 258 | 209 | 205 | 131 | 115 | 100 | 114 | 298 | 249 |
| 1 | 2 | Patna, India | 212 | 354 | 297 | 225 | 230 | 169 | 183 | 82 | 100 | 84 | 136 | 402 | 277 |
| 2 | 3 | Saharsa, India | 207 | 418 | 344 | 238 | 220 | 167 | 149 | 85 | 93 | 91 | 110 | 282 | 292 |
| 3 | 4 | New Delhi, India | 205 | 325 | 244 | 167 | 181 | 175 | 124 | 70 | 110 | 91 | 210 | 405 | 352 |
| 4 | 5 | Noida, India | 201 | 304 | 212 | 154 | 187 | 176 | 129 | 70 | 125 | 118 | 237 | 367 | 338 |

Statistical Summary:

| | rank | avg |
|---|---|---|
| count | 5377.00000 | 5377.000000 |
| mean | 2689.00000 | 32.171657 |
| std | 1552.35053 | 27.075191 |
| min | 1.00000 | 1.000000 |
| 25% | 1345.00000 | 16.000000 |
| 50% | 2689.00000 | 26.000000 |
| 75% | 4033.00000 | 37.000000 |
| max | 5377.00000 | 223.000000 |

Missing Value Check:

| | |
|---|---|
| rank | 0 |
| city | 0 |
| avg | 0 |
| jan | 0 |
| feb | 0 |
| mar | 0 |
| apr | 0 |
| may | 0 |

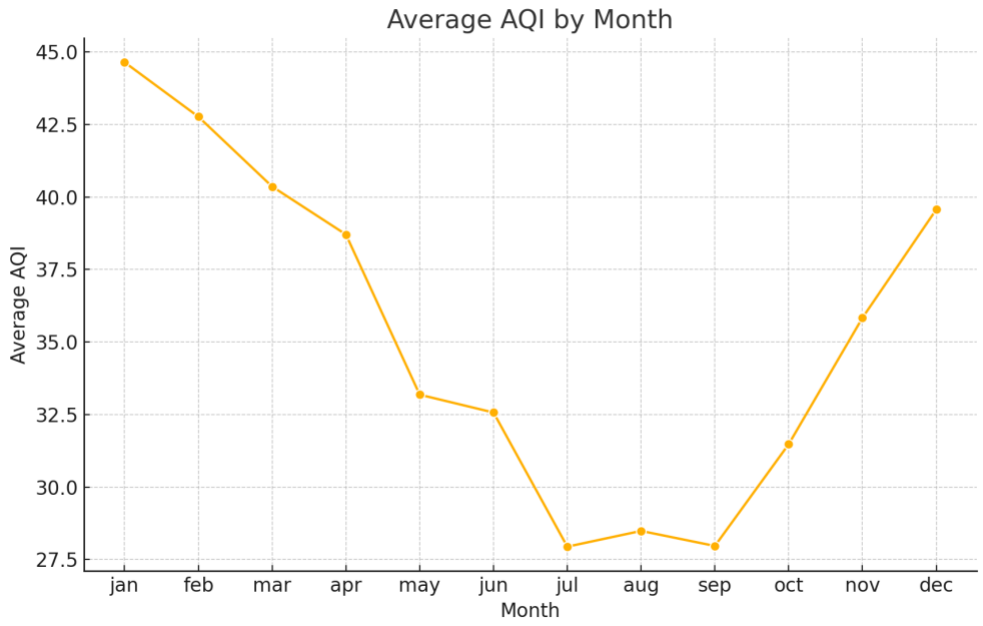
jun 0
jul 0
aug 0
sep 0
oct 0
nov 0
dec 0

## 3. Data Preprocessing

The monthly AQI values were initially of type 'object' and have been converted to numeric values. Any non-numeric values were coerced into NaNs. No missing values were found, so no imputation was needed.

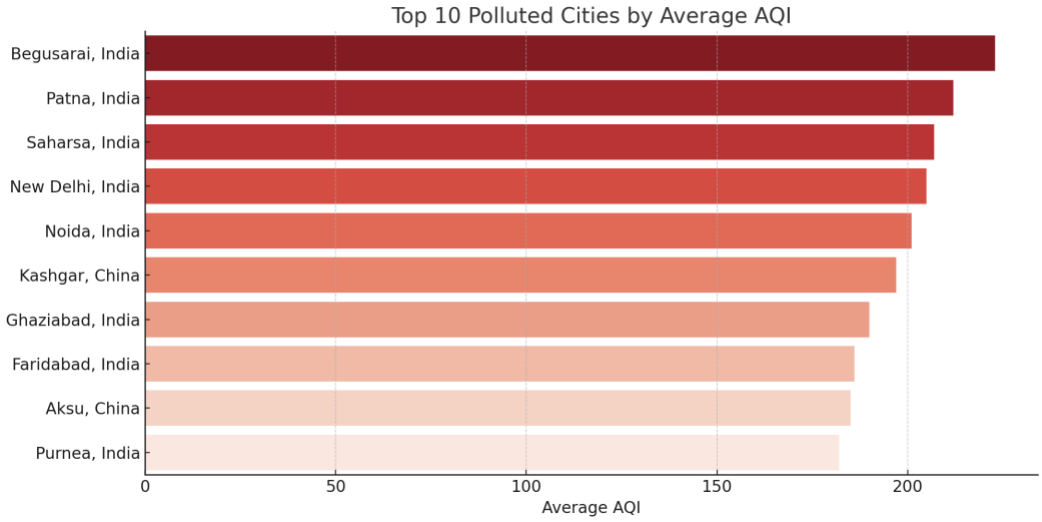
## 4. Exploratory Data Analysis

Below is the average AQI observed for each month across all cities:




The line graph above shows higher AQI values during winter months (Jan, Nov, Dec), indicating seasonal pollution.

The following chart displays the top 10 cities with the worst average AQI:




Cities like Begusarai, Patna, and Saharsa consistently show high pollution levels.

## 5. Code with Explanation

The project used Python libraries such as pandas, seaborn, and matplotlib for data analysis and visualization. Here's a breakdown of major code sections:

1. **Loading the Dataset:** Used `pd.read_csv()` to load the AQI data.
2. **Preprocessing:** Converted monthly columns to numeric using `pd.to_numeric()`.
3. **EDA:** Used `df.describe()` and `.isnull().sum()` to summarize and check for missing data.
4. **Visualization:** Plotted monthly AQI trends and top polluted cities using seaborn and matplotlib.

## 6. Insights & Conclusion

From the AQI data, we observed clear seasonal patterns with worse air quality during the winter months. Additionally, several cities in Bihar and Delhi NCR rank highest in pollution levels, necessitating targeted environmental policies.