INT375

PROJECT REPORT

(Project Semester January–April 2025)

POLLUTION REPORT OF DIFFERENT INDIAN STATES

Submitted by:

Name: Rahul

Registration No: 12325234

Programme and Section: B.Tech CSE K23EU

Course Code: INT375

Department of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab (India)

April 2025

Under the Guidance of

Tanima Thakur
Discipline of CSE/IT
Lovely School of Computer Science and Engineering
Lovely Professional University, Phagwara

CERTIFICATE

This is to certify that Rahul Kumar Bala , bearing Registration No. 12325234, has successfully completed the project work entitled “POLLUTION REPORT OF DIFFERENT INDIAN STATES”
as part of the course INT375 during the project semester January–April 2025, under my supervision and guidance.To the best of my knowledge, the present work is the result of the student’s original research, development, and effort.

Signature and Name of the Supervisor
Tanima Thakur

DECLARATION

I, Rahul Kumar Bala , a student of B.Tech under the CSE/IT Discipline at Lovely Professional University, Punjab, hereby declare that the project work entitled "POLLUTION REPORT OF DIFFERENT INDIAN STATES " submitted in partial fulfillment of the course INT375, is the result of my own intensive work. The content in this report is original, genuine, and has not been copied from any unauthorized source. All efforts and data analysis have been conducted with sincerity and academic integrity.

Date: 11-04-2025

Signature: _____

RegNo.: 12325234

Name of the Student Rahul Kumar Bala

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to all those who supported me throughout the completion of this project titled

“POLLUTION REPORT OF DIFFERENT INDIAN STATES”.

First and foremost, I would like to thank **Tanima Thakur**, my respected project supervisor, for her invaluable guidance, support, and encouragement. Her expert advice, timely feedback, and motivation helped me stay focused and complete this project efficiently.

I would also like to extend my appreciation to the **faculty and staff of the Discipline of CSE/IT, Lovely Professional University**, for creating a learning environment that inspired this research.

A special thanks to my peers, friends, and family for their constant moral support and encouragement throughout the project.

Lastly, I would like to acknowledge the use of the dataset provided by **data,gov.in** which served as the foundation for the analysis conducted in this report.

Rahul Kumar Bala

Registration No.: 12325234

Pollution Level Analysis Report – India

1. Introduction

India, a rapidly developing country, faces significant environmental challenges, one of which is the high levels of air pollution in its cities. Pollution levels in India have been rising due to increased industrial activity, vehicular emissions, and other anthropogenic factors. This report delves into the pollution levels across various states of India, analyzing trends, geographical distribution, and the overall state of air quality using data collected from multiple regions.

Purpose:

The primary purpose of this analysis is to understand the distribution and trends of pollution levels across India. By examining various pollution metrics (e.g., average, minimum, and maximum pollutant levels), this report aims to highlight regions with the highest pollution levels and provide insights into how these levels have evolved over time.

Dataset:

The dataset used for this analysis consists of pollution data from multiple states and cities in India. It includes measurements of various air quality metrics such as particulate matter (PM10, PM2.5), nitrogen oxides (NO2), sulfur dioxide (SO2), and carbon monoxide (CO). The dataset also includes metadata such as geographic coordinates, city names, states, and timestamps of the observations.

2. Data Cleaning and Preparation

Data cleaning is a crucial step in any data analysis process. In this case, the raw data needed some preprocessing before any meaningful analysis could be performed.

Steps Taken:

Handling Missing Data:

Several columns in the dataset had missing values. Specifically, pollutant readings such as pollutant_min, pollutant_max, and pollutant_avg had missing values. These were handled by:

Filling missing pollutant_min and pollutant_max values with the pollutant_avg value where applicable.

For rows where pollutant_avg was missing, the average of pollutant_min and pollutant_max was used as an approximation.

Time Conversion:
The column last_update was initially in string format. It was converted to a datetime object for time-based analysis.

Handling Duplicates:
Any duplicate rows in the dataset were removed to ensure the data was unique and accurate.

Creating Additional Columns:

A new column year_month was created to group the data by month and year. This was useful for observing trends over time.

A categorical column pollutant_level was introduced to classify pollution levels into categories such as “Good”, “Moderate”, and “Hazardous” based on predefined bins.

3. Descriptive Analysis and Data Distribution

Before diving into the visualizations, it is important to understand the distribution of the data, which will provide context to the findings.

Pollutant Level Distribution:
The dataset was classified into various pollution levels. The distribution of these categories can help identify the extent of pollution in the regions analyzed. Here's a summary of the pollution levels:

Good: Air quality is considered satisfactory, and air pollution poses little or no risk.

Moderate: Air quality is acceptable; however, there may be some risk to sensitive people.

Unhealthy for Sensitive Groups: The air quality may be a concern for sensitive individuals such as those with pre-existing health conditions.

Unhealthy: The air quality may pose a health risk to the general population.

Very Unhealthy: Air quality is considered dangerous for everyone.

Hazardous: Air quality is extremely dangerous for everyone.

4. Trend of Pollution Levels Over Time

One of the most important aspects of pollution data is identifying trends over time. Trends can reveal whether pollution levels are improving or worsening, and they can highlight seasonal fluctuations or the impact of policy changes.

The analysis of pollution levels over time showed notable trends, especially during the winter months when pollution levels in urban areas tend to rise due to factors such as vehicular emissions and burning of crop residue.

Insights from the Time Series Data:
Seasonality: Certain months of the year consistently show higher levels of pollution, particularly during the winter (October to February).

Long-term Trends: Over the past few years, there is an observed increase in pollution levels in major urban centers.

5. Geographic Distribution of Pollution

India's geography plays a significant role in its pollution distribution. Urban areas like Delhi, Mumbai, and Bangalore experience much higher pollution levels than rural areas due to industrial activities, transportation, and population density.

Key Findings:
Urban Areas: Major cities, particularly Delhi, have significantly higher pollution levels. This can be attributed to high vehicular emissions, industrial emissions, and a large population.

Rural Areas: While rural areas tend to have lower pollution levels on average, areas near industrial zones or where agricultural burning is common can experience spikes in pollution.

An interactive map was created to visualize this geographic distribution. The map allows us to zoom in on various regions and identify pollution hotspots across the country.

6. Pollution Comparison Between States

India's states show vast variation in pollution levels. States like Delhi, Uttar Pradesh, and Bihar have significantly higher pollution levels compared to states like Kerala and Goa.

A bar plot was used to visualize the average pollution levels across different states. The states with the highest pollution levels were Delhi, Uttar Pradesh, Maharashtra, and West Bengal.

Insights:
Top Polluted States: Delhi has the highest average pollution level, followed by states like Uttar Pradesh and Haryana. These regions have high levels of industrial activity and traffic.

Least Polluted States: States such as Sikkim, Goa, and Kerala exhibit relatively lower levels of pollution.

7. Pollution Distribution by Time of Day

Pollution levels also vary depending on the time of day. The highest levels of pollution are typically observed during the morning and evening rush hours, when vehicular traffic is at its peak.

Using box plots, the distribution of pollution levels by hour of the day was analyzed, showing the following:

Peak Hours: The highest levels of pollution were found between 8 AM and 9 AM, and again between 6 PM and 7 PM.

Low Pollution Hours: Pollution levels are lower during late-night and early morning hours when traffic is reduced.

8. Pollution Levels and Geographic Hotspots

An interactive scatter plot was created to visualize the pollution levels across various geographic coordinates. This scatter plot helped identify areas with both high pollution and population density.

Geographic Hotspots:

Delhi: As expected, Delhi is the most polluted city, with multiple areas exhibiting hazardous levels of pollution.

Other Cities: Mumbai, Lucknow, and Kolkata also stand out as regions with particularly high pollution levels.

9. Key Insights and Observations

From the visualizations and analysis, several key insights were drawn:

Air Pollution Is a Major Concern in Urban Areas: Cities like Delhi and Mumbai have among the highest pollution levels, largely due to traffic congestion, industrial activity, and burning of crop residue.

Seasonal Variations: Pollution levels tend to peak during the winter months, particularly due to meteorological conditions such as low wind speeds and temperature inversions, which trap pollutants near the surface.

Disparities Between States: Pollution levels vary significantly between states, with northern and industrialized regions showing higher levels of pollution than southern and rural states.

Time-of-Day Impact: Pollution levels show clear patterns related to the time of day, with the highest levels during rush hours.

10. Conclusion and Recommendations

In conclusion, air pollution is a critical issue affecting the health of millions of people across India. Urban areas face the highest levels of pollution, with Delhi standing as the most polluted region in the country.

Recommendations:
Government Action: Stronger environmental policies are needed to regulate industrial emissions and vehicular pollution.

Public Awareness: Awareness programs focusing on the health risks of pollution and the importance of reducing emissions could help improve air quality.

Sustainable Urban Development: Cities should focus on sustainable urban planning, including better public transportation, green spaces, and stricter emission controls.

PYTHON CODE USED FOR DATA ANALYSIS

(POLLUTION REPORT OF DIFFERENT INDIAN STATES)

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from datetime import datetime
import plotly.express as px
import warnings

warnings.filterwarnings('ignore')

# Set style for beautiful visualizations
plt.style.use('seaborn-v0_8-darkgrid')
sns.set_palette("husl")

# Load the data
# Assuming the data is in a CSV file named 'pollution_data.csv'
df = pd.read_csv('C:/Users/Admin/OneDrive/Desktop/3b01bcb8-0b14-4abf-b6f2-
c1bfd384ba69.csv')

# Data Cleaning
def clean_data(df):
    # Convert last_update to datetime
    df['last_update'] = pd.to_datetime(df['last_update'])

    # Handle missing values
    df['pollutant_min'] = df['pollutant_min'].fillna(df['pollutant_avg'])
```

```
    df['pollutant_max'] = df['pollutant_max'].fillna(df['pollutant_avg'])
    df['pollutant_avg'] = df['pollutant_avg'].fillna((df['pollutant_min'] + df['pollutant_max']) / 2)

    # Drop any remaining rows with missing critical data
    df = df.dropna(subset=['pollutant_id', 'pollutant_avg'])

    # Create a new column for year-month for time-based analysis
    df['year_month'] = df['last_update'].dt.to_period('M')

    # Create categorical bins for pollutant levels
    bins = [0, 50, 100, 150, 200, 300, 500, np.inf]
    labels = ['Good', 'Moderate', 'Unhealthy(Sensitive)', 'Unhealthy',
              'Very Unhealthy', 'Hazardous', 'Extremely Hazardous']
    df['pollutant_level'] = pd.cut(df['pollutant_avg'], bins=bins, labels=labels)

    return df

df = clean_data(df)

# Visualization 1: Pollutant Distribution by Level (Pie Chart)
plt.figure(figsize=(10, 8))
df['pollutant_level'].value_counts().plot.pie(autopct='%1.1f%%',
                                               explode=[0.05]*len(df['pollutant_level'].unique()),
                                               shadow=True)
plt.title('Distribution of Pollution Levels', fontsize=16)
plt.ylabel('')
plt.tight_layout()
plt.savefig('pollution_levels_pie.png')
plt.show()
```

```
# Visualization 2: Trend of Average Pollution Over Time (Line Plot)
plt.figure(figsize=(14, 7))
df.groupby('year_month')['pollutant_avg'].mean().plot(kind='line', marker='o', linewidth=2)
plt.title('Trend of Average Pollution Levels Over Time', fontsize=16)
plt.xlabel('Year-Month', fontsize=12)
plt.ylabel('Average Pollution Level', fontsize=12)
plt.xticks(rotation=45)
plt.grid(True, alpha=0.3)
plt.tight_layout()
plt.savefig('pollution_trend_line.png')
plt.show()

# Visualization 3: Pollution by Country (Bar Plot)
plt.figure(figsize=(12, 8))
country_avg = df.groupby('country')['pollutant_avg'].mean().sort_values(ascending=False)
sns.barplot(x=country_avg.values, y=country_avg.index, palette='viridis')
plt.title('Average Pollution Levels by Country', fontsize=16)
plt.xlabel('Average Pollution Level', fontsize=12)
plt.ylabel('Country', fontsize=12)
plt.tight_layout()
plt.savefig('pollution_by_country_bar.png')
plt.show()

# Visualization 4: Boxplot of Pollution by State (Top 10 States)
plt.figure(figsize=(14, 8))
top_states = df['state'].value_counts().nlargest(10).index
sns.boxplot(data=df[df['state'].isin(top_states)],
            x='state', y='pollutant_avg',
            palette='magma')
plt.title('Pollution Distribution in Top 10 States', fontsize=16)
```

```
plt.xlabel('State', fontsize=12)
plt.ylabel('Pollution Level', fontsize=12)
plt.xticks(rotation=45)
plt.tight_layout()
plt.savefig('pollution_by_state_box.png')
plt.show()

# Visualization 5: Heatmap of Correlation Between Variables
plt.figure(figsize=(10, 8))
numeric_df = df.select_dtypes(include=[np.number])
sns.heatmap(numeric_df.corr(), annot=True, cmap='coolwarm', center=0)
plt.title('Correlation Heatmap of Numerical Variables', fontsize=16)
plt.tight_layout()
plt.savefig('correlation_heatmap.png')
plt.show()

# Visualization 6: Scatter Plot of Min vs Max Pollution
plt.figure(figsize=(10, 8))
sns.scatterplot(data=df, x='pollutant_min', y='pollutant_max',
                hue='pollutant_level', palette='Set2', s=100)
plt.title('Minimum vs Maximum Pollution Levels', fontsize=16)
plt.xlabel('Minimum Pollution Level', fontsize=12)
plt.ylabel('Maximum Pollution Level', fontsize=12)
plt.legend(title='Pollution Level')
plt.tight_layout()
plt.savefig('min_max_scatter.png')
plt.show()

# Visualization 7: Violin Plot of Pollution Distribution by Level
plt.figure(figsize=(12, 8))
```

```
sns.violinplot(data=df, x='pollutant_level', y='pollutant_avg',
               palette='Spectral', inner='quartile')
plt.title('Distribution of Pollution Levels', fontsize=16)
plt.xlabel('Pollution Level Category', fontsize=12)
plt.ylabel('Pollution Level', fontsize=12)
plt.xticks(rotation=45)
plt.tight_layout()
plt.savefig('pollution_violin.png')
plt.show()

# Visualization 8: Stacked Bar Chart of Pollution Levels by Country
plt.figure(figsize=(14, 8))
pd.crosstab(df['country'], df['pollutant_level'], normalize='index').plot.bar(stacked=True,
                                                                               figsize=(14,8),
                                                                               colormap='tab20')
plt.title('Pollution Level Composition by Country', fontsize=16)
plt.xlabel('Country', fontsize=12)
plt.ylabel('Proportion', fontsize=12)
plt.legend(title='Pollution Level', bbox_to_anchor=(1.05, 1))
plt.xticks(rotation=45)
plt.tight_layout()
plt.savefig('pollution_stacked_bar.png')
plt.show()

# Visualization 9: Geographic Distribution of Pollution (using Plotly)
fig = px.scatter_geo(df,
                     lat='latitude',
                     lon='longitude',
                     color='pollutant_avg',
                     hover_name='city',
```

```
                scope='world',
                color_continuous_scale=px.colors.sequential.Plasma,
                title='Geographic Distribution of Pollution Levels')
fig.update_layout(height=600, width=1000)
fig.write_image("geographic_distribution.png")
fig.show()

# Visualization 10: Time Series Decomposition of Pollution Levels
from statsmodels.tsa.seasonal import seasonal_decompose

# Create a time series dataframe
ts_df = df.set_index('last_update').resample('D')['pollutant_avg'].mean().fillna(method='ffill')

# Perform decomposition
result = seasonal_decompose(ts_df, model='additive', period=30)

plt.figure(figsize=(14, 10))

plt.subplot(4, 1, 1)
plt.plot(result.observed)
plt.title('Observed', fontsize=12)

plt.subplot(4, 1, 2)
plt.plot(result.trend)
plt.title('Trend', fontsize=12)

plt.subplot(4, 1, 3)
plt.plot(result.seasonal)
plt.title('Seasonal', fontsize=12)
```

```
plt.subplot(4, 1, 4)
plt.plot(result.resid)
plt.title('Residual', fontsize=12)

plt.suptitle('Time Series Decomposition of Pollution Levels', fontsize=16)
plt.tight_layout()
plt.savefig('time_series_decomposition.png')
plt.show()

# Visualization 11: Pollution Distribution by Hour of Day
df['hour'] = df['last_update'].dt.hour
plt.figure(figsize=(12, 6))
sns.boxplot(data=df, x='hour', y='pollutant_avg', palette='cool')
plt.title('Pollution Levels by Hour of Day', fontsize=16)
plt.xlabel('Hour of Day', fontsize=12)
plt.ylabel('Pollution Level', fontsize=12)
plt.tight_layout()
plt.savefig('pollution_by_hour.png')
plt.show()

# Visualization 12: Pairplot of Numerical Variables
plt.figure(figsize=(12, 10))
sns.pairplot(df[['pollutant_min', 'pollutant_max', 'pollutant_avg', 'latitude', 'longitude']],
             diag_kind='kde',
             plot_kws={'alpha': 0.6})
plt.suptitle('Pairplot of Numerical Variables', y=1.02, fontsize=16)
plt.tight_layout()
plt.savefig('numerical_pairplot.png')
plt.show()
```

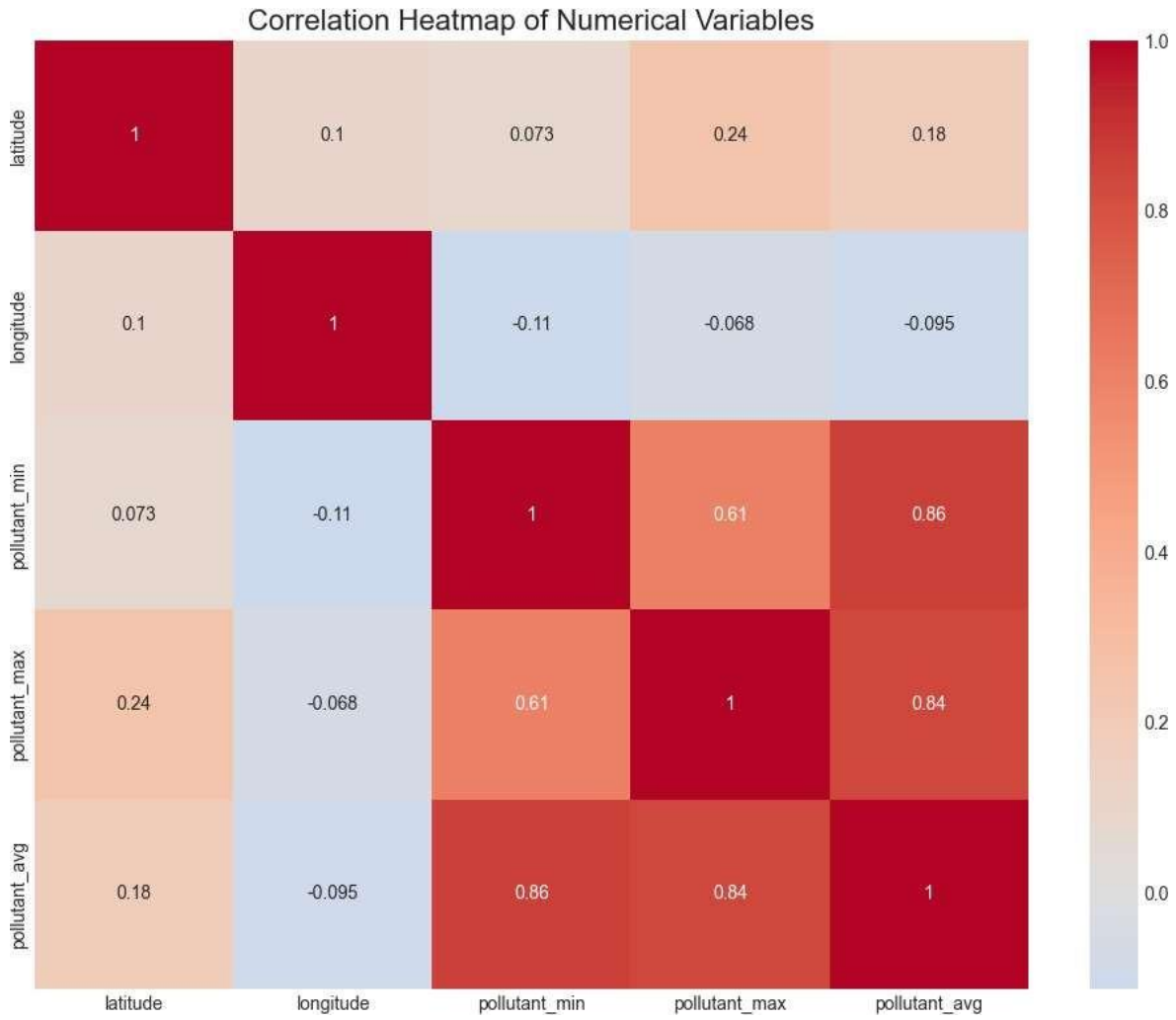
```
# Visualization 13: Pollution Level by Month (Heatmap)
df['month'] = df['last_update'].dt.month_name()
month_order = ['January', 'February', 'March', 'April', 'May', 'June',
               'July', 'August', 'September', 'October', 'November', 'December']
pivot_table = df.pivot_table(values='pollutant_avg',
                             index='country',
                             columns='month',
                             aggfunc='mean')

plt.figure(figsize=(14, 10))
sns.heatmap(pivot_table[month_order], cmap='YlOrRd', annot=True, fmt=".1f")
plt.title('Average Pollution Levels by Country and Month', fontsize=16)
plt.xlabel('Month', fontsize=12)
plt.ylabel('Country', fontsize=12)
plt.tight_layout()
plt.savefig('pollution_monthly_heatmap.png')
plt.show()

# Visualization 14: Radar Chart of Pollution Metrics by Country
from math import pi

# Select top 5 countries
top_countries = df['country'].value_counts().nlargest(5).index.tolist()
radar_df = df[df['country'].isin(top_countries)].groupby('country')[['pollutant_min',
'pollutant_avg', 'pollutant_max']].mean()

# Number of variables
categories = radar_df.columns.tolist()
N = len(categories)
```

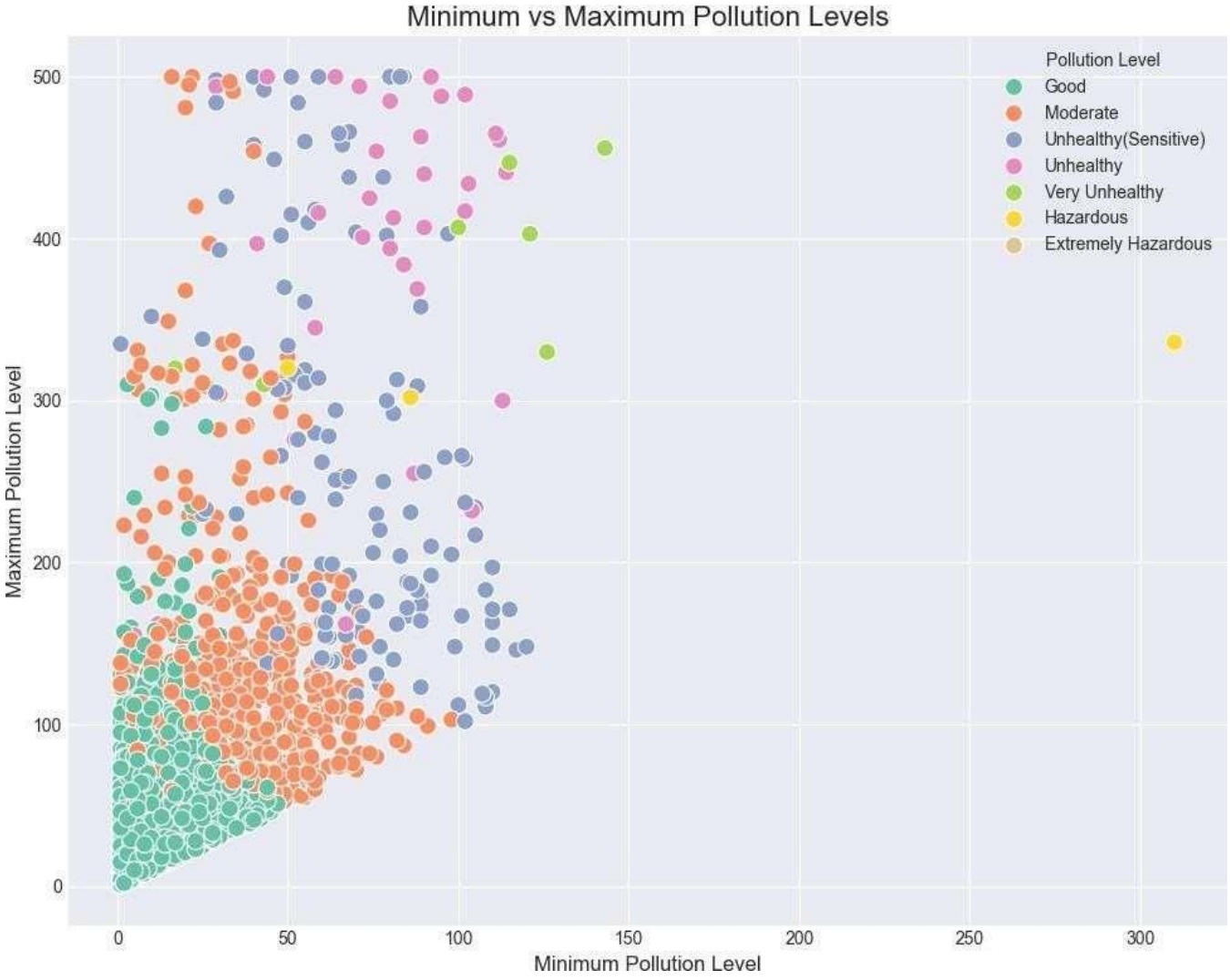
```
# Create radar chart
angles = [n / float(N) * 2 * pi for n in range(N)]
angles += angles[:1]

plt.figure(figsize=(10, 10))
ax = plt.subplot(111, polar=True)
ax.set_theta_offset(pi / 2)
ax.set_theta_direction(-1)
plt.xticks(angles[:-1], categories)

# Draw ylabels
ax.set_rlabel_position(0)
plt.yticks(color="grey", size=10)

# Plot each country
for idx, country in enumerate(radar_df.index):
    values = radar_df.loc[country].values.flatten().tolist()
    values += values[:1]
    ax.plot(angles, values, linewidth=2, linestyle='solid', label=country)
    ax.fill(angles, values, alpha=0.25)

plt.title('Pollution Metrics Comparison by Country', size=16, y=1.1)
plt.legend(loc='upper right', bbox_to_anchor=(1.3, 1.1))
plt.tight_layout()
plt.savefig('pollution_radar_chart.png')
plt.show()
```

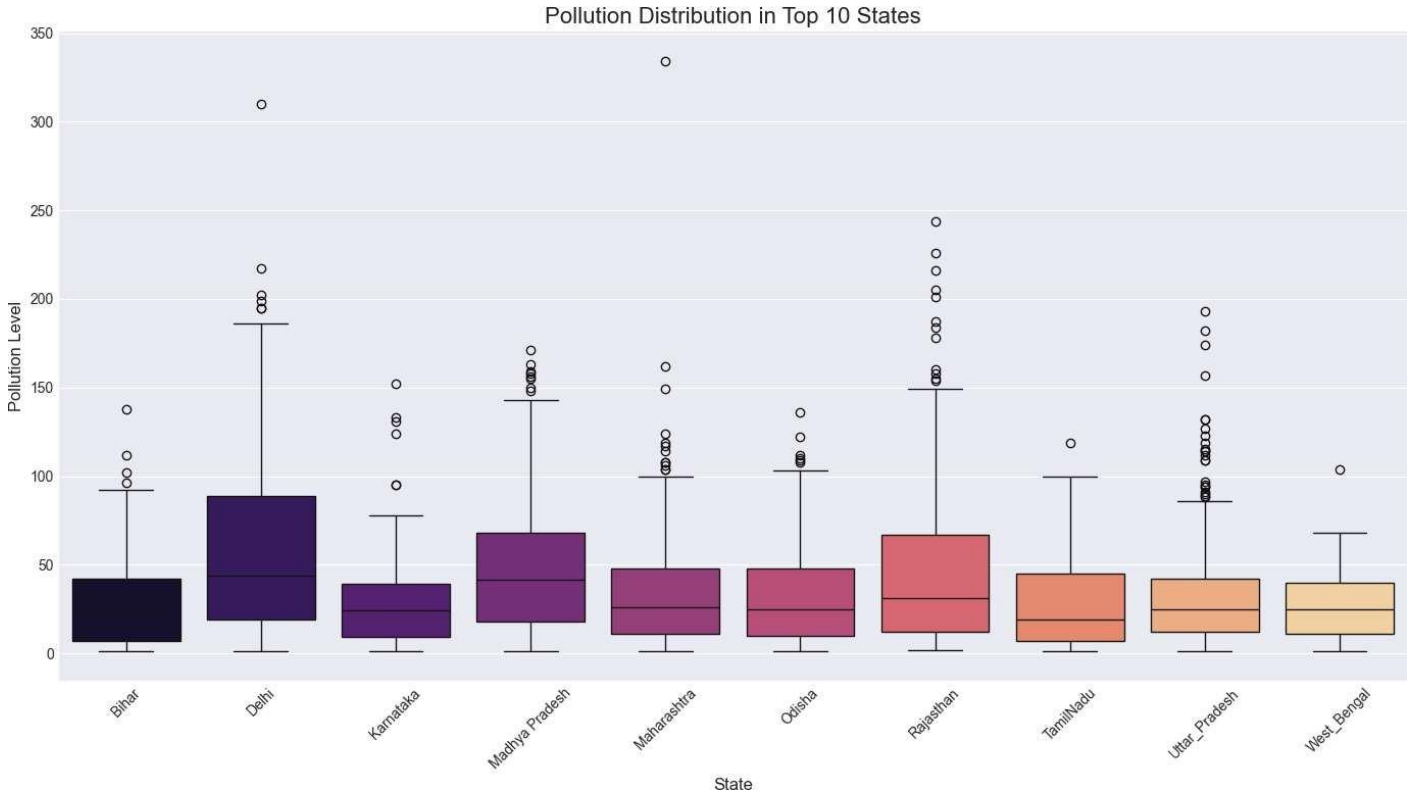
LINK OF LINKEDIN:

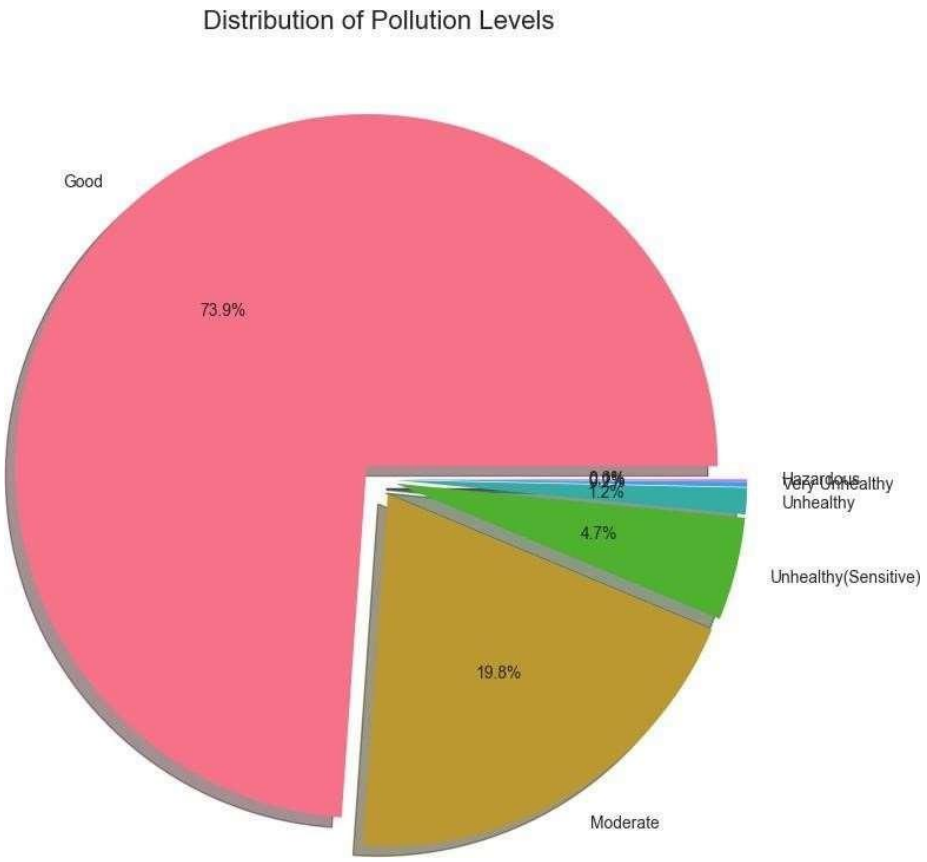
https://www.linkedin.com/in/rahulkumar00/

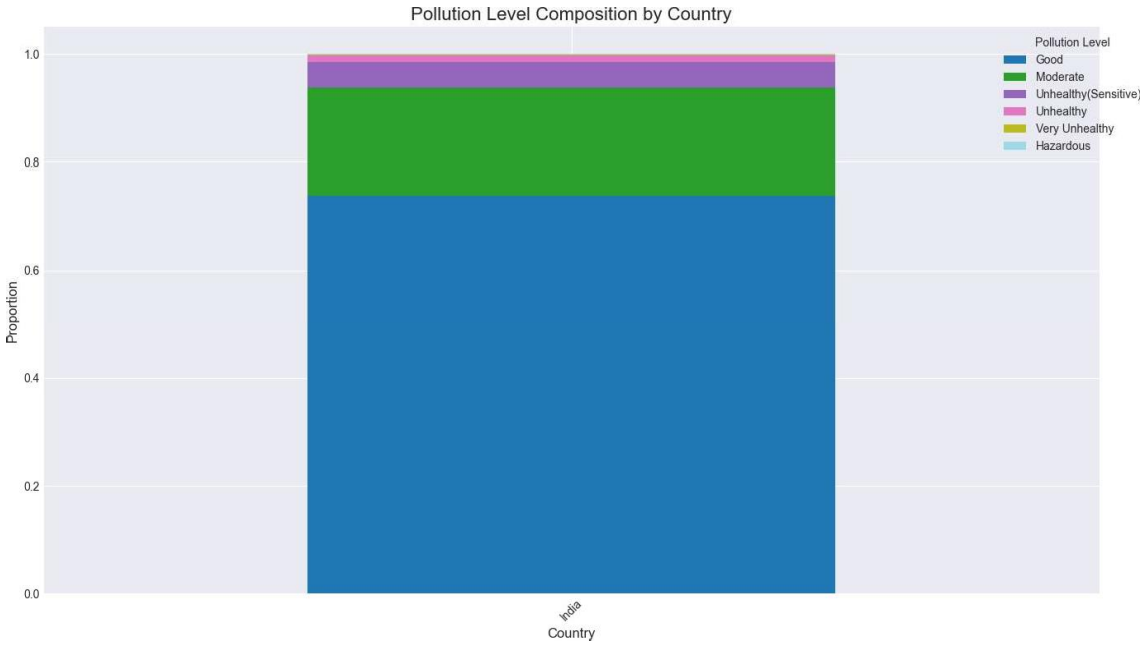
DIFFERENT CHARTS AND VISUALIZATION

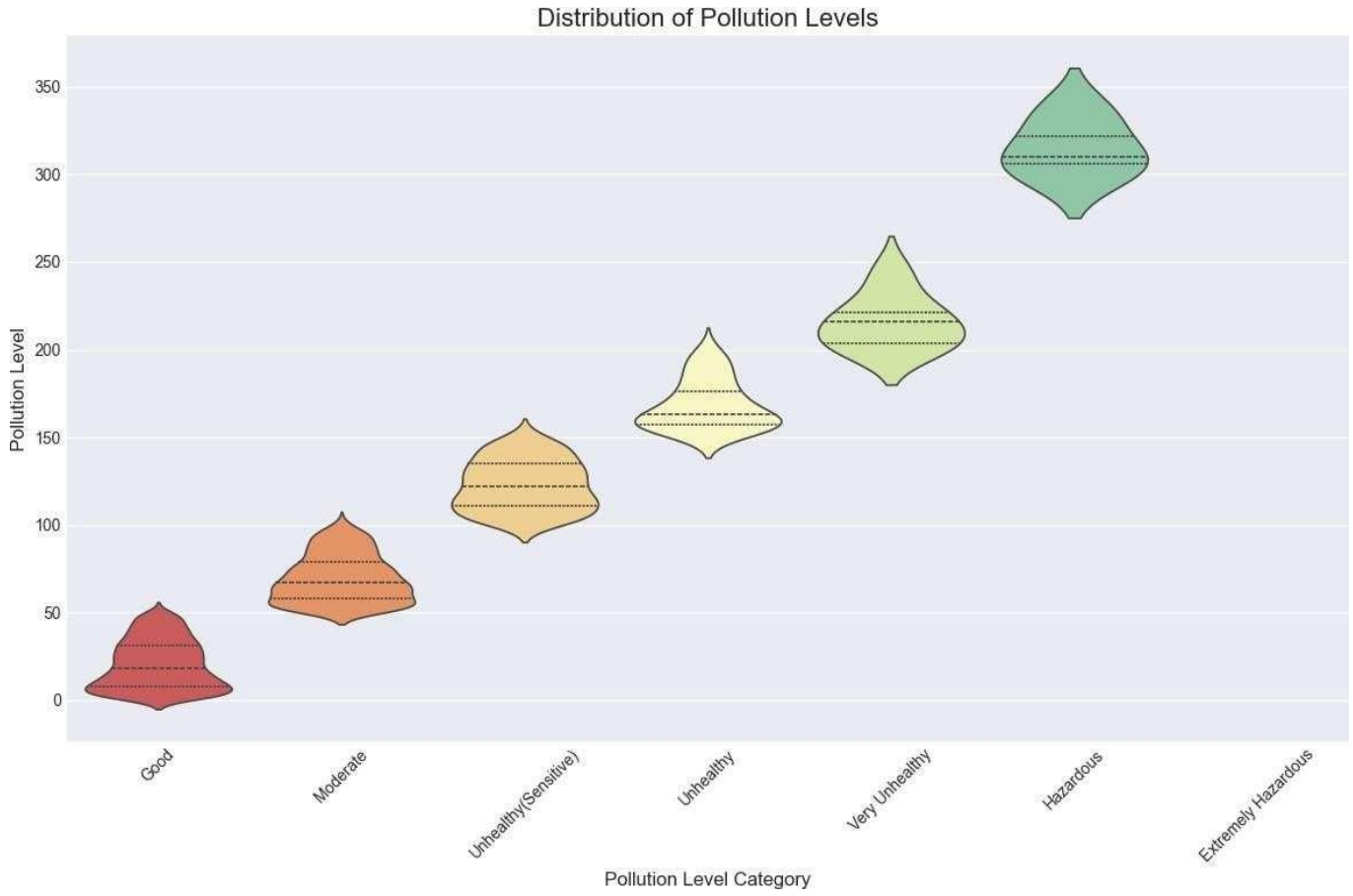



Minimum vs Maximum Pollution Levels
Pollution Level
Good
Moderate
Unhealthy(Sensitive)
Unhealthy
Very Unhealthy
Hazardous
Extremely Hazardous
Maximum Pollution Level
Minimum Pollution Level
0
100
200
300
400
500
0
50
100
150
200
250
300

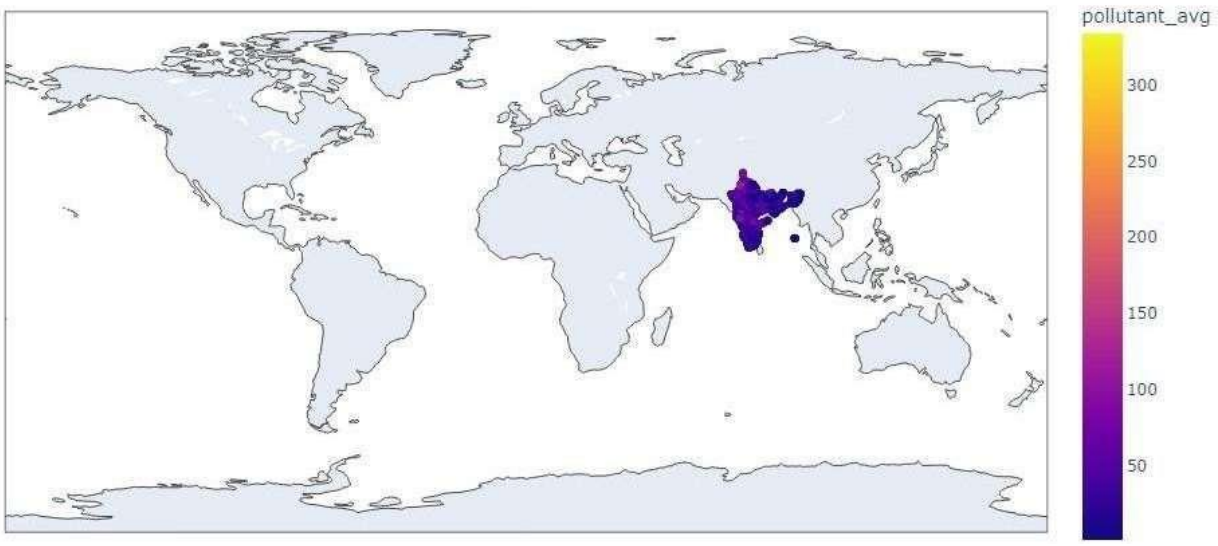

Pollution Distribution in Top 10 States
Pollution Level
350
300
250
200
150
100
50
0
Bihar
Delhi
Karnataka
Madhya Pradesh
Maharashtra
Odisha
Rajasthan
TamilNadu
Uttar_Pradesh
West_Bengal
State


Distribution of Pollution Levels
Good
73.9%
Hazardous
Very Unhealthy
1.2%
Unhealthy
4.7%
Unhealthy(Sensitive)
19.8%
Moderate

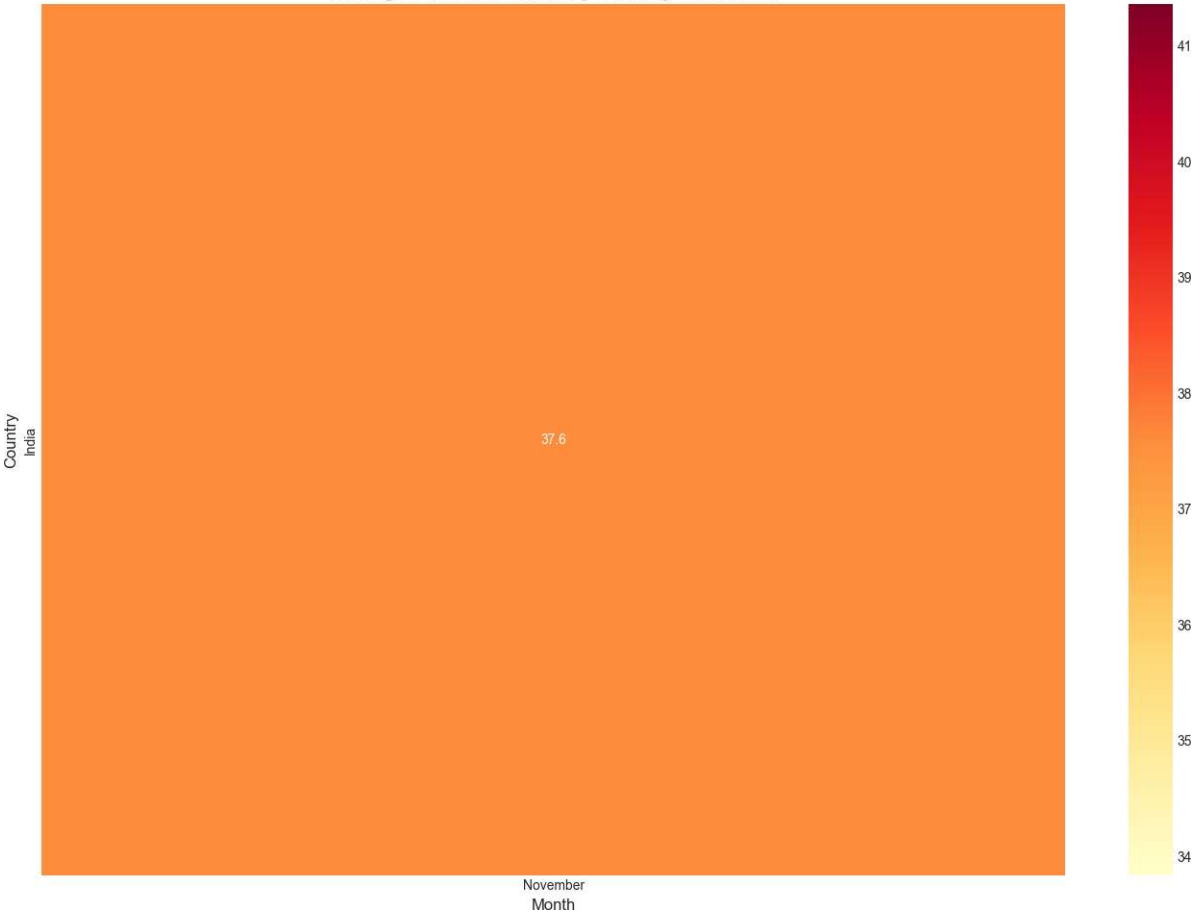

Pollution Level Composition by Country
Pollution Level
Good
Moderate
Unhealthy(Sensitive)
Unhealthy
Very Unhealthy
Hazardous
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
India
Country

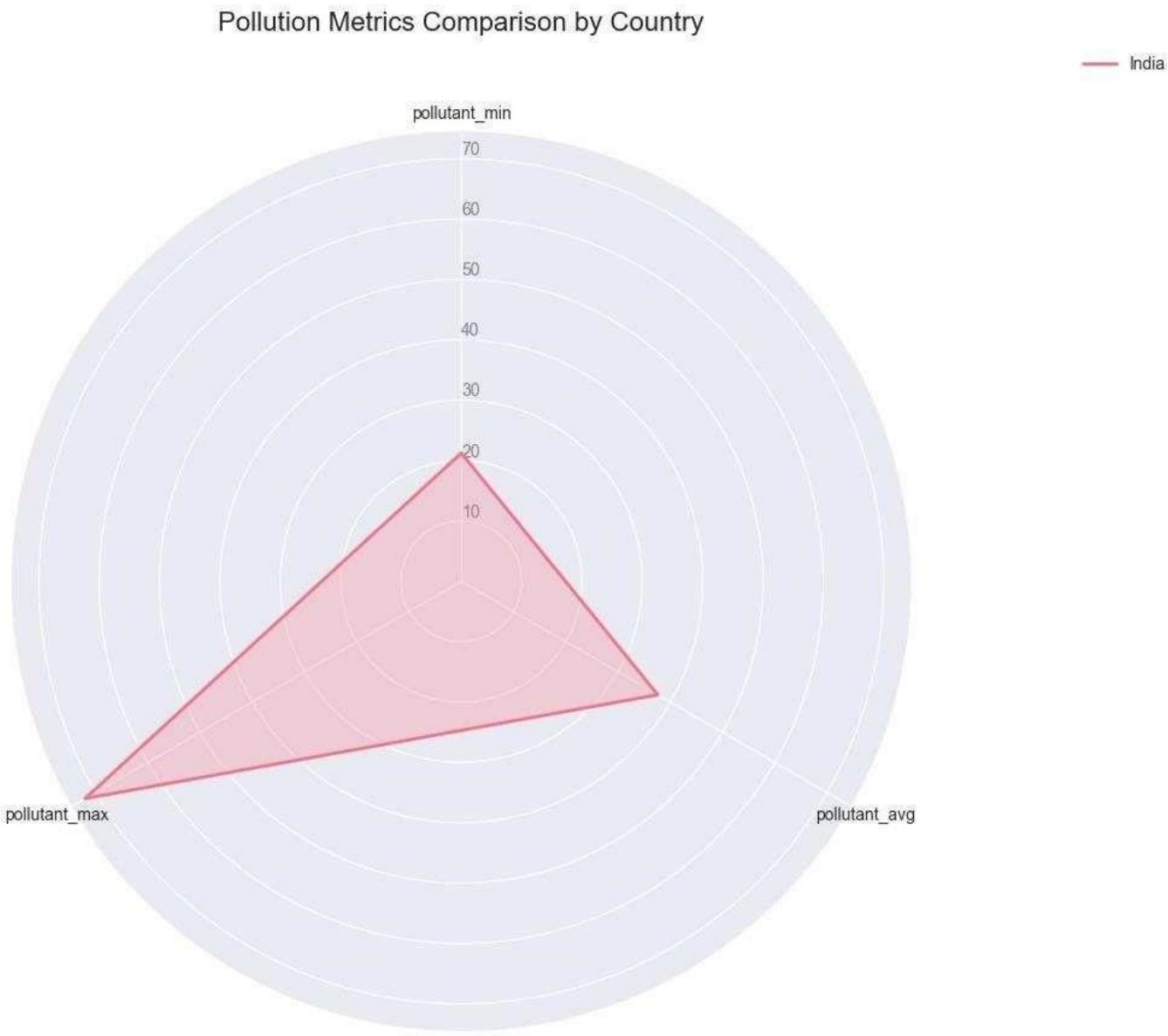

Distribution of Pollution Levels
Pollution Level
350
300
250
200
150
100
50
0
Good
Moderate
Unhealthy(Sensitive)
Unhealthy
Very Unhealthy
Hazardous
Extremely Hazardous
Pollution Level Category

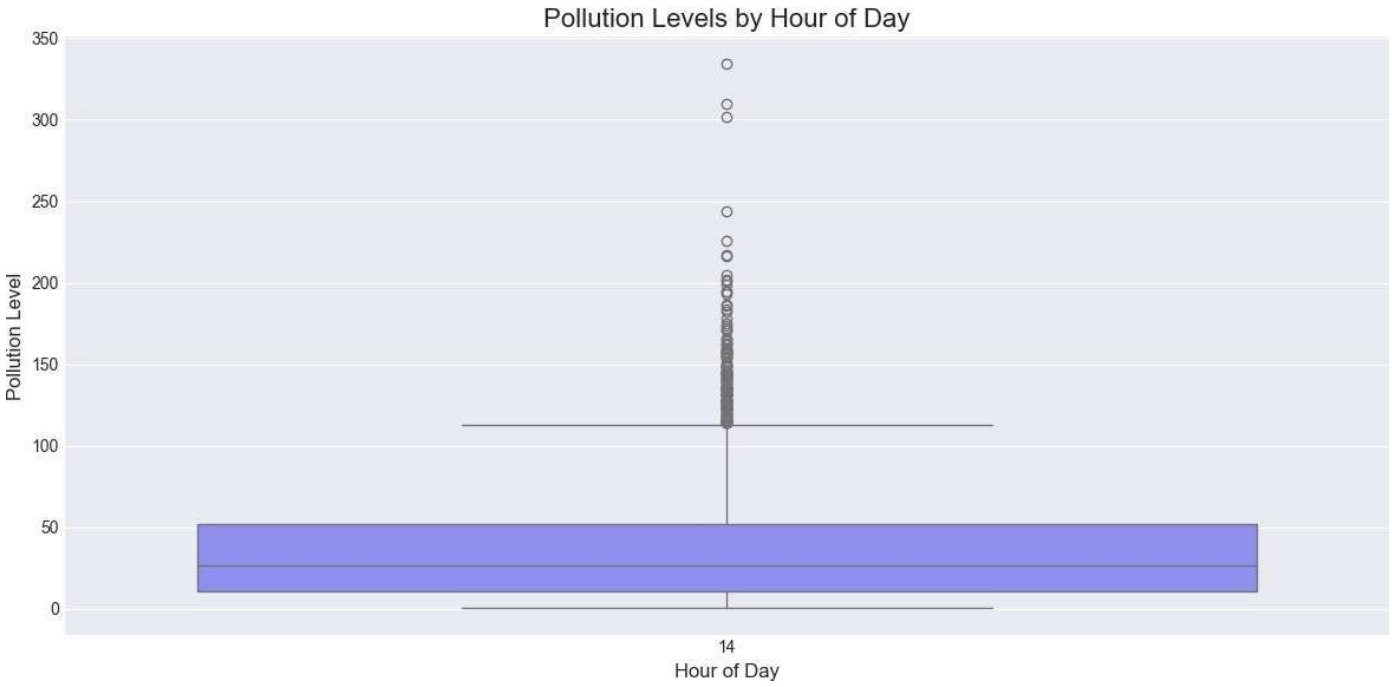
Geographic Distribution of Pollution Levels

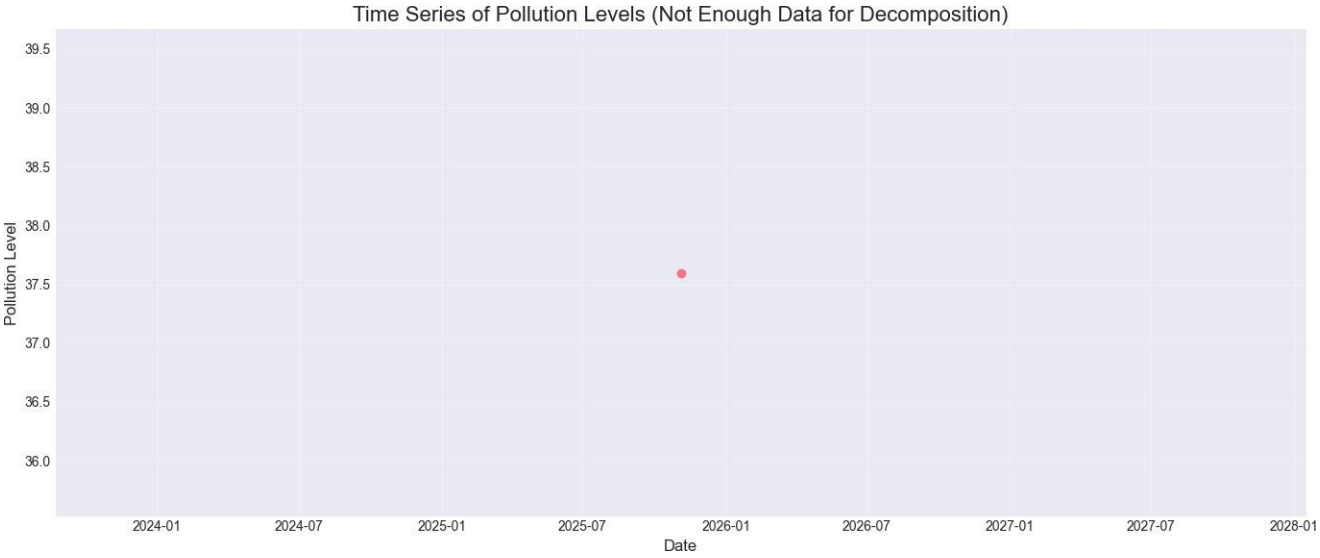


Average Pollution Levels by Country and Month




Pollution Metrics Comparison by Country
India
pollutant_min
70
60
50
40
30
20
10
pollutant_max
pollutant_avg


Pollution Levels by Hour of Day
Pollution Level
350
300
250
200
150
100
50
0
14
Hour of Day


Time Series of Pollution Levels (Not Enough Data for Decomposition)
Pollution Level
39.5
39.0
38.5
38.0
37.5
37.0
36.5
36.0
2024-01
2024-07
2025-01
2025-07
2026-01
2026-07
2027-01
2027-07
2028-01
Date

REFERENCES

1. Central Pollution Control Board (CPCB) – Real-Time Air Quality Data.
2. CPCB – National Air Quality Index (NAQI)
3. Statista – PM2.5 Concentrations by State
4. Swachh Survekshan – Sanitation and Cleanliness Data