Global Air Quality Analysis Report

1. Problem Definition

Introduction

Air pollution is a major global concern affecting millions of people. High levels of pollutants such as PM2.5, NO2, and SO2 contribute to respiratory diseases, climate change, and reduced quality of life.

This study aims to analyze global air pollution trends using historical data, identify the most and least polluted regions, and predict future pollution levels using statistical forecasting techniques. By leveraging data science methodologies, this analysis provides insights into air pollution variations over time and across different locations, which can help policymakers, environmentalists, and researchers make data-driven decisions.

Objectives

- Analyze global air pollution trends using historical data.
- Identify the most polluted and cleanest regions worldwide.
- Predict future pollution levels using ARIMA and Facebook Prophet forecasting models.
- Provide insights that could inform environmental policies and interventions.

Research Questions

- How does air pollution vary across different regions?
- What are the trends in major pollutants (PM2.5, NO2, CO, SO2, etc.)?
- Which countries experience the highest and lowest pollution levels?
- Can we predict future air quality trends?

2. **Methodology**

2.1 Data Collection

- Source: The dataset used in this study contains historical air pollution levels from multiple countries and is sourced from global environmental monitoring agencies.

- **Parameters:**

  - `Location` – Country or region name.

  - `Period` – Year of the recorded data.

  - `Pollution Value` – Average pollution level (PM2.5 concentration in µg/m³).

  - `Pollution Low` – Lower bound pollution level.

  - `Pollution High` – Upper bound pollution level.

  - `Indicator` – Air pollution measurement type.

- Storage Format: The dataset is stored in a structured CSV file (`global_air_quality_cleaned.csv`).

2.2 Data Preprocessing

- Handling Missing Values:

  - Checked for missing pollution values and filled or removed them where necessary.

- Datetime Conversion:

  - Ensured that the period (year) was correctly formatted for time-series analysis.

- Standardized Column Names

  - Renamed columns for better readability and analysis consistency.

2.3 Exploratory Data Analysis (EDA)

EDA was conducted to understand the dataset's characteristics, distribution, and trends:

- Statistical Summary

  - Computed descriptive statistics (mean, median, standard deviation, min, max values).

- **Visualizations:**
  - Box plots** to analyze pollution distributions across locations.
  - Line plots to track trends over time.
  - Correlation heatmaps to identify relationships between pollutants.
  - Bar charts for country-wise and continent-wise comparisons.

2.4 Predictive Modeling (ARIMA & Facebook Prophet Forecasting)

To predict future air pollution levels, we used two forecasting models:

ARIMA Model (AutoRegressive Integrated Moving Average)

- Captures linear trends in pollution levels.
- Suitable for time-series forecasting where past values influence future trends.
- Predicts a gradual decrease in pollution from 2020-2030.

Facebook Prophet Model

- Designed for handling seasonality and trend shifts.
- Provides uncertainty estimates (confidence intervals) for predictions.
- Captures periodic fluctuations in pollution levels.

Both models were trained on historical data (2010-2019) and used to forecast pollution trends for 2020-2030.

3. Analysis & Results

3.1 Most Polluted Regions

- Countries with the highest pollution levels:
  - Afghanistan -(Highest PM2.5 concentration)
  - Tajikistan

  - Kuwait
  - India

- Factors contributing to high pollution:
  - Industrial emissions
  - Rapid urbanization
  - Traffic congestion
  - Limited air quality regulations

3.2 Cleanest Regions

- The cleanest regions identified include:
  - Bahamas
  - Finland
  - Iceland
  - Canada

- These countries have:
  - Strict environmental regulations
  - Lower industrial emissions
  - Higher investments in renewable energy

3.3 Pollution Trends Over Time

- Pollution levels peaked around 2014 before showing a slight improvement.
- South Asia and the Middle East have the highest pollution levels, while Scandinavian countries and island nations remain the least polluted.
- Seasonal variations observed due to factors such as winter pollution buildup and industrial cycles.

3.4 Forecast for Future Pollution Trends (ARIMA & Facebook Prophet)

- India’s pollution is expected to decline slightly but remains high.

- Finland’s pollution remains low and continues to improve.

- Global air pollution shows a slight downward trend, suggesting successful environmental efforts.

- ARIMA vs. Prophet Comparison: Prophet provides a more dynamic trend analysis with seasonal adjustments, while ARIMA focuses on linear projections.

3.5 Key Visualizations

-Boxplots highlight extreme pollution levels in certain regions.

- Line charts track pollution variations over time.

- Heatmaps illustrate the correlation between different pollutants.

- ARIMA & Prophet forecasts** predict future pollution trends up to 2030.

4. Conclusion

4.1 Summary of Findings

- Certain regions experience extreme air pollution levels due to industrialization, vehicle emissions, and lack of regulations.

- Countries with strict air quality laws and renewable energy sources maintain cleaner air.

- Forecasting models suggest that pollution levels may gradually decline in some areas but will remain a concern in others.

4.2 Limitations

- The dataset does not include real-time pollution updates or external factors such as weather conditions and policy changes.

- Some regions have limited data availability**, making trend predictions less precise.

- The forecasting models do not account for sudden environmental or political changes (e.g., COVID-19 lockdown effects).

4.3 Future Work

- Expanding the dataset with real-time air pollution APIs.
- Incorporating additional factors such as population density, economic growth, and energy consumption.
- Refining predictions using deep learning models like LSTMs for better accuracy.
- Developing an interactive dashboard for real-time pollution monitoring using `Streamlit` or `Dash`.

5. Deployment Recommendations

- Google Colab: Share and run the analysis interactively.
- GitHub Repository: Host code and allow collaboration.

Appendix

Technologies Used

- Python (Pandas, Matplotlib, Seaborn, Statsmodels)
- ARIMA & Facebook Prophet Time-Series Forecasting
- Jupyter Notebook / Google Colab for interactive analysis

References

- Global Air Quality Dataset
- World Health Organization Air Pollution Reports

Author: Ritalee Monde