# US Air Pollution Forecasting and Analysis Dashboard

Student Name(s): [Your Name]

Course/Batch and Institute: B.Sc. (Hons.) Data Science, Batch of 2022–2025, [Your Institute Name]

Project Guide / Mentor Name: [Guide/Mentor Name]

Period of Internship: 14th Jan 2025 – 30th April 2025

Report submitted to: IDEAS – Institute of Data Engineering, Analytics and Science Foundation, ISI Kolkata

## Abstract

This project presents a Streamlit-based interactive dashboard for analyzing and forecasting air pollution trends across various U.S. cities using real-world data from the U.S. Environmental Protection Agency (EPA). Users can explore historical pollutant data (NO₂, SO₂, CO, O₃), apply smoothing, and generate forecasts using two predictive models: Facebook Prophet and Linear Regression. The project also evaluates and compares model performance using R² scores and visualizations. This tool empowers users with data-driven insights into environmental quality trends, aiding both public awareness and policy planning.

## Introduction

Air pollution is a growing global concern with direct impacts on health and climate. This project seeks to analyze historical pollution data from U.S. cities and forecast pollutant concentrations using machine learning techniques. Technologies used include Python, Streamlit (for web dashboard development), Plotly (for data visualization), and Facebook Prophet and Scikit-learn (for forecasting models). The EPA’s dataset provides measurements from 2000 to 2016, enabling in-depth temporal analysis. The purpose is to create an accessible tool for stakeholders to visualize trends, explore statistical patterns, and forecast future pollution levels, ultimately supporting informed decision-making.

## Project Objective

* Build an interactive dashboard for air pollution analysis using Streamlit.
* Enable users to filter data by state, city, and pollutant.
* Implement smoothed trend visualization using rolling averages.
* Perform forecasting using both Prophet and Linear Regression models.
* Compare model performance using R² scores and visualizations.

## Methodology

1. Data Loading and Cleaning  
- EPA dataset (`pollution\_us\_2000\_2016.csv`) loaded via pandas.  
- Date columns parsed, null values dropped.  
  
2. Dashboard Development  
- Built using Streamlit with custom CSS for UI enhancement.  
- Sidebar filters for State, City, and Pollutant selection.  
- Display of raw data and summary statistics using st.expander().  
  
3. Visualization  
- Plotly used to create interactive time series plots.  
- Optionally apply a 30-day rolling average to smooth trends.  
  
4. Forecasting Models  
Prophet:  
- Uses date (`ds`) and pollutant concentration (`y`).  
- Automatically detects seasonality and trend.  
Linear Regression:  
- Transforms date to integer days from baseline.  
- Fit model using sklearn.linear\_model.LinearRegression.  
- Forecasts generated for next 365 days.  
  
5. Model Evaluation  
- Linear Regression performance evaluated on training data.  
- R² and MSE metrics displayed for model selection.  
  
6. Forecast Comparison  
- Option to plot both model forecasts on the same graph.  
- Users can download forecasts as CSV files.  
  
7. Tools Used  
- Languages: Python  
- Libraries: Streamlit, pandas, plotly, prophet, sklearn, numpy  
- Development Platform: Local Python environment  
- Visualization: Plotly Express for dynamic plotting

## Data Analysis and Results

Descriptive Analysis  
- Summary statistics (mean, std, min, max) displayed for each pollutant.  
- Raw data preview helps verify data structure and outliers.  
  
Trend Analysis  
- Time-series plots show patterns in pollutant levels.  
- Rolling average reveals long-term trends.  
  
Forecasting Models  
1. Prophet  
- Automatically captures yearly, weekly seasonality.  
- Generates lower and upper bounds for predictions.  
2. Linear Regression  
- Simpler model based on temporal linear trends.  
- Produces deterministic output with high interpretability.  
  
Model Comparison  
- R² scores evaluated for Linear Regression.  
- Visual comparison using Plotly chart with colored lines for each model.  
- Forecasts saved as downloadable .csv files for further analysis.

## Conclusion

This project successfully developed an interactive dashboard for U.S. air pollution analysis and forecasting. It integrates data preprocessing, visualization, and model-driven prediction, offering both accuracy and usability. The Prophet model outperformed linear regression in most cases due to its robust handling of seasonality. Future work can include real-time data integration, expansion to other countries, and deployment on the cloud for public access.

## APPENDICES

Appendix A – References

- U.S. Environmental Protection Agency Dataset (via Kaggle)

- Facebook Prophet Documentation

- Scikit-learn API Reference

- Streamlit and Plotly official docs

Appendix B – Code Extract

(See attached file: d.py)

Appendix C – Document Link

A copy of this report and project code is available on GitHub:  
👉 [GitHub/Google Drive Link – to be added]