# Project Report

Project Title: Data Analysis and Forecasting Project

1. Introduction  
This project focuses on analyzing a given dataset to extract meaningful insights through data processing, visualization, and predictive modeling. The objective is to understand patterns in the data and evaluate different forecasting models.

2. Objectives  
- To preprocess and clean the dataset for accurate analysis.  
- To perform exploratory data analysis (EDA) to identify trends and anomalies.  
- To implement and compare different forecasting models.  
- To evaluate model performance using statistical error metrics.  
- To analyze environmental factors affecting air pollution levels.

3. Methodology  
1. Data Collection: The dataset was imported from a CSV file.  
2. Data Preprocessing: Missing values were handled, and irrelevant data was removed.  
3. Exploratory Data Analysis (EDA):Various statistical measures were computed to understand the dataset structure.  
4. Model Implementation: Three different models—ARIMA, XGBoost, and LSTM—were implemented for forecasting.  
5. Model Evaluation: The models were assessed based on Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).  
6. Environmental Factor Analysis: The relationships between air pollutants and meteorological factors were examined.

4. Key Findings

Model Performance Evaluation:  
The performance of the forecasting models was evaluated as follows:

|  |  |  |
| --- | --- | --- |
| Model | MAE | RMSE |
| ARIMA | 0.1107 | 0.1550 |
| XGBoost | 0.1307 | 0.1474 |
| LSTM | 0.0488 | 0.0640 |

- The LSTM model demonstrated the best performance with the lowest MAE (0.0488) and RMSE (0.0640), indicating better accuracy.  
- ARIMA performed moderately, showing slightly higher errors compared to LSTM.  
- XGBoost had the highest MAE but a relatively competitive RMSE, indicating that it struggled more with absolute deviations.

Environmental Factor Analysis:  
1. Strong Correlations Among Air Pollutants  
- CO, NO₂, SO₂, and PM2.5show strong positive correlations (0.55 - 0.73), suggesting common sources such as vehicle emissions and industrial pollution.  
- CO & NO₂ correlation (0.62):Both are byproducts of combustion engines.

2. Temperature vs. Air Quality  
- Ozone (O₃) & Temperature (0.30 🔺):Higher temperatures correlate with increased ozone due to photochemical reactions.  
- NO₂ & Temperature (-0.40 🔻):Higher temperatures reduce NO₂ as dispersion increases.

3.Humidity & Air Quality  
- Ozone & Humidity (-0.46 🔻):High humidity reduces ozone levels by absorbing radiation.  
- Humidity & PM2.5 (-0.20 🔻):Moist conditions may trap pollutants, but rain can clear the air.

4. Wind Speed & Pollution  
- Wind Speed & PM2.5 (-0.12 🔻):Wind disperses pollutants, reducing concentrations.

5. Precipitation (Rain) & Air Quality  
- Precipitation & PM2.5 (-0.21 🔻):Rain helps clean the air by reducing PM2.5 levels.  
- Precipitation & Humidity (0.37 🔺):More rain leads to increased humidity.

5. Conclusion  
The analysis provided a comprehensive understanding of the dataset and its forecasting potential. The evaluation of different models showed that deep learning (LSTM) outperformed traditional statistical (ARIMA) and machine learning (XGBoost) models in accuracy. The environmental analysis revealed key factors affecting air pollution, offering valuable insights for pollution management and mitigation strategies.

6. Recommendations

Insights & Recommendations  
1. Hot Weather Increases Ozone:  
 - Implement ozone warnings during heatwaves.  
 - Promote reduced vehicle emissions in high-temperature seasons.  
2. Rain Helps Improve Air Quality:  
 - Cities should leverage rainfall forecasting to predict natural pollution reductions.  
3. Wind Disperses Pollution, But Not Always Enough:  
 - Low-wind days lead to high pollution accumulation.  
 - Urban areas should incorporate wind circulation strategies (e.g., green belts, open spaces).  
4. High Humidity Can Reduce Ozone:  
 - In humid regions, ozone pollution may be less severe.