**INT375: DATA SCIENCE TOOLBOX: PYTHON PROGRAMMING**

**PROJECT REPORT**

(Project Semester January- April 2025)

*Analysis and Visualization of Air Quality Data using Python*

Submitted by

Nandhana cs

Registration No: 12223505

Programme and Section: B-Tech (CSE), K23EH

Course Code: INT375

Under the Guidance of

**Madhu Bala**

**Discipline of CSE/IT**

**Lovely School of Computer Science**

**Lovely Professional University, Phagwara**

# Certificate

This is to certify that Nandhana cs, bearing Registration No.12223505, has completed the INT375 minor project titled “Analysis and Visualization of Air Quality Data using Python” under my guidance and supervision. The work presented in this report is original and reflects her own efforts and study.

Madhu Bala

School of Computer Science Engineering

Lovely Professional University Phagwara, Punjab.

Date: 12th April, 2025

# Declaration

I, Nandhana cs, student of B.Tech K23EH under the CSE Discipline at Lovely Professional University, Punjab, declare that all the information furnished in this project report is based on my own work. I affirm that no part of this report has been copied from any source without proper citation.

Date: 12-04-2025 Signature: Nandhana

# Acknowledgement

I am deeply thankful to my project guide, Madhu bala, for her constant support, guidance, and encouragement throughout this project. Her insightful feedback helped me improve my analysis and presentation. I also extend my gratitude to my classmates for their valuable suggestions and to my family for their unwavering support and motivation.

# Table of Contents

1. Introduction

2. Source of Dataset

3. EDA Process

4. Analysis on Dataset

4.1 Pollutant Trends Over Time

4.2 Distribution of Pollutants

4.3 Missing Data and Outlier Detection

4.4 Correlation Analysis

4.5 Normality Testing

4.6 Linear Regression Modeling

5. Conclusion

6. Future Scope

7. References

# 1. Introduction

Air pollution poses one of the greatest environmental risks to public health. Monitoring and analyzing air quality data can help identify pollution patterns, sources, and the effectiveness of regulatory policies. In this project, I performed a comprehensive analysis of air quality indicators using Python programming tools. The objective was to explore and visualize trends in pollutants such as PM2.5, NO2, SO2, and O3 across time and locations.

Using Python libraries like Pandas, NumPy, Matplotlib, Seaborn, and Scikit-learn, I applied data wrangling, visualization, and basic machine learning techniques to uncover insights and relationships within the dataset.

# 2. Source of Dataset

The dataset used in this project is titled "Air Quality Data" and contains measurements of key air pollutants recorded across various regions.

**Dataset Title:** Air Quality Data  
**Source:** Data.gov  
**Link:** <https://catalog.data.gov/dataset/air-quality>

It was provided for academic use and includes the following fields: Unique ID, Indicator ID, Name, Measure, Measure Info, Geo Type Name, Geo Join ID, Geo Place Name, Time Period, Start\_Date, and Data Value.

# 3. EDA Process

Exploratory Data Analysis is the first step in understanding the structure and hidden patterns in the dataset. Below is a summary of the steps taken:

1. Data Loading: Dataset was imported using Pandas.
2. Preview & Inspection: df.head(), df.info(), and df.describe() were used.
3. Column Cleanup: Column names were standardized for easier access.
4. Missing Value Handling: Missing numeric data in 'Data Value' was filled using column mean.
5. Type Conversion: 'Start\_Date' column was converted to datetime format.
6. Pollutant Selection: Rows with pollutant names in the 'Name' column were filtered for focused analysis on PM2.5, NO2, SO2, and O3.

# 4. Analysis on Dataset

## 4.1 Pollutant Trends Over Time

Time-series plots were generated using the 'Start\_Date' and 'Data Value' columns, filtered by 'Name' for each pollutant. This helped visualize how concentrations of pollutants changed over time.

## 4.2 Distribution of Pollutants

Histograms and boxplots were created using 'Data Value', grouped by the pollutant names in the 'Name' column. These visualizations helped detect skewed distributions and outliers.

## 4.3 Normality Testing

Q-Q plots and Shapiro-Wilk tests were applied to the 'Data Value' column of each pollutant type to test for normality.

## 4.4 Missing Data and Outlier Detection

Missing 'Data Value' entries were imputed using column mean. Boxplots further revealed outliers in pollutants like SO2.

## 4.5 Linear Regression Modeling

A simple linear regression model was built to predict PM2.5 'Data Value' using PM10 (if available), using Scikit-learn. The 'Name' column was used to filter required indicators, and results were visualized using regplots.

## 4.6 Correlation Analysis

The dataset was pivoted such that each pollutant became a column and 'Start\_Date' was used as the index. Correlation between pollutants (based on their 'Data Value') was visualized using heatmaps.

# 5. Conclusion

This project explored air quality data using Python-based data science tools. The analysis revealed patterns in pollutant levels, relationships between variables, and opportunities for further modeling. The project demonstrated the practical application of data wrangling, visualization, and basic statistics to a real-world environmental issue.

# 6. Future Scope

* Time Series Forecasting: Apply ARIMA/LSTM to forecast future pollutant levels.
* Geo-Spatial Analysis: Use 'Geo Place Name' and 'Geo Join ID' to map pollution trends geographically.
* Health Impact Studies: Relate 'Data Value' to public health metrics and hospitalization records.

# 7. References

* Air Quality Dataset (2025)
* Python for Data Science by Reema Thareja
* Seaborn & Matplotlib Documentation
* Scikit-learn User Guide