## DATA SCIIENCE USING PTHON PROJECT REPORT

(Project Semester January-April 2025)

# *Air Quality Data Cleaning and Exploratory Data Analysis Using Python*

### Submitted by

Yaswanth Kumar Mallela

Registration no: 12315067

Section: K23FD

Course Code: INT375

**Under the Guidance of** Dr. Baljinder Kaur

(27952)

### Discipline of CSE/IT

**Lovely School of Computer Science Engineering Lovely Professional University, Phagwara**

**CERTIFICATE**

This is to certify that **Yaswanth Kumar Mallela** bearing Registration no. **12315067** has completed **INT375** project titled, **“Air Quality Data Cleaning and Exploratory Data Analysis Using Python”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

GitHub Link: <https://github.com/Yash864512/Python_Project>

### Dr. Baljinder Kaur Assistant Professor

**School of Computer Science Engineering**

Lovely Professional University Phagwara, Punjab.

Date: 12th April, 2025

**DECLARATION**

I, Yaswanth Kumar Mallela, student of B. tech under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

GitHub Link: <https://github.com/Yash864512/Python_Project>

Date: 12th April, 2025 Signature

Registration No.12315067 Yaswanth Kumar Mallela

**Acknowledgment**

I would like to express my deepest gratitude to Prof Dr. Baljinder Kaur for his exceptional mentorship and unwavering support throughout the duration of this project. His vast knowledge in the fields of data science and machine learning, combined with his patient and thoughtful guidance, played a pivotal role in the successful completion of this work. His insightful suggestions and feedback consistently challenged me to think critically and improve the quality of my research. I am also grateful for the learning environment he fostered, which encouraged exploration and innovation.

In addition, I sincerely thank my peers and classmates for their helpful discussions, encouragement, and collaborative spirit during this project. Their input provided fresh perspectives that contributed meaningfully to the final outcome. I am also thankful to the open-source community for providing the tools, libraries, and resources that made the implementation of this project possible. Lastly, I acknowledge the dataset contributors for making this analysis feasible.

# Introduction

The purpose of this project is to analyze air quality data through a systematic process of data cleaning, transformation, and visualization. The dataset used, *Air\_Quality.csv*, contains environmental measurements related to air quality indicators across various geographical locations and timeframes.

Data cleaning plays a crucial role in ensuring the accuracy and reliability of any analysis. This project begins by identifying and handling missing or inconsistent values in key columns such as 'Data Value', 'Geo Place Name', and 'Measure Info'. Numeric conversions and imputations were carefully applied to maintain data integrity.

Following the cleaning process, exploratory data analysis (EDA) was conducted to uncover trends, distributions, and relationships within the data. Various visualizations—such as histograms, boxplots, violin plots, and heatmaps—were utilized to highlight insights like the distribution of data values, comparisons across different measures, and temporal patterns.

This comprehensive approach not only enhances data quality but also enables meaningful interpretation of air quality trends, which can inform environmental policy, health research, and community awareness initiatives.

# Dataset Description

The dataset used in this project, titled **"Air\_Quality.csv"**, contains measurements related to air quality indicators collected across different geographic regions. It includes both quantitative and categorical variables that help in understanding the environmental conditions and pollutant levels over time.

**Key Attributes:**

* **Measure**: Represents the type of air quality indicator measured, such as PM2.5, Ozone concentration, etc.
* **Data Value**: A numeric value indicating the measurement of the air quality indicator. This is the primary variable used in analysis.
* **Geo Place Name**: The specific location (e.g., city, borough) where the measurement was recorded.
* **Geo Type Name**: The classification of the location, such as borough or city.
* **Measure Info**: Additional context or explanation about the type of measurement or its source.
* **Start\_Date** (if present): The date when the measurement was taken, allowing for temporal trend analysis.

**Dataset Characteristics:**

* The dataset contains both numerical and categorical data.
* Missing values were present in several columns, especially in 'Data Value', 'Geo Place Name', and 'Measure Info'.
* After cleaning, the dataset was saved as **Air\_Quality\_Cleaned.csv**, and a final transformed version was saved as **Air\_Quality\_Transformed.csv**.

This dataset enables exploration of how air quality varies across different regions and timeframes, providing valuable insights into environmental health patterns.

# Source of Dataset

The dataset utilized for this project was sourced from [**data.gov.in**](https://data.gov.in), the official open data portal of the Government of India. This platform provides access to a wide range of datasets across multiple domains such as health, education, environment, agriculture, and urban development. It aims to promote transparency, encourage data-driven decision-making, and support research and innovation by making public data freely available to citizens, researchers, and developers.

The specific dataset used in this project pertains to **Air Quality Measurements** collected from various geographical locations across India. It was published by a government agency responsible for environmental monitoring and reporting (such as the Central Pollution Control Board or a state-level agency). The dataset includes detailed information on air quality indicators like particulate matter (PM), ozone levels, and other pollutants, along with metadata such as the location of measurement, date of collection, and type of geographic unit (e.g., city, district, borough).

This dataset was selected for its credibility, nationwide coverage, and the relevance of the variables it contains for environmental analysis. Its availability in a structured and machine-readable format made it suitable for data analysis and visualization tasks. The insights derived from this dataset can contribute to a better understanding of pollution patterns, support public awareness campaigns, and inform environmental policy-making.

# Exploratory Data Analysis (EDA) Process

Exploratory Data Analysis (EDA) was conducted to gain a deeper understanding of the air quality dataset, identify patterns, detect anomalies, and uncover relationships among variables. The process involved both summary statistics and various forms of visual analysis.

**1. Overview and Structure**

The EDA process began by examining the structure and content of the dataset using methods such as .info() and .describe(). This provided:

* The total number of rows and columns
* Data types of each column
* Counts of non-null entries
* Basic statistical summaries (mean, median, standard deviation, min, max, etc.)

Additionally, the number of unique values in each categorical column (like Measure, Geo Place Name, and Geo Type Name) was assessed to understand the scope and variety of observations.

**2. Handling Missing and Inconsistent Data**

Before performing deep analysis, missing values were addressed:

* The Data Value column (a key numerical metric) was converted to numeric type and missing values were imputed using the column’s mean.
* Categorical columns like Geo Place Name and Measure Info were filled using the mode or a placeholder ("Unknown").

**3. Statistical Grouping and Aggregations**

Various aggregations were applied to better understand trends:

* **Average Data Value by Measure**: Identified which air quality measures had the highest average pollution levels.
* **Maximum Data Value by Geo Place Name**: Highlighted the most polluted areas.
* **Year-wise trends**: If the Start\_Date column was present, data was grouped by year to detect temporal trends in air quality.

**4. Correlation Analysis**

A correlation matrix was computed to examine relationships between numerical variables. A heatmap visualization helped reveal any strong linear relationships or multicollinearity within the dataset.

**5. Categorical Combination Analysis**

Frequent combinations of Measure and Geo Type Name were explored to understand which metrics were most commonly recorded in which geographic formats.

**6. Visualizations**

Several plots were used to support the above analyses:

* **Histogram & KDE Plot**: Showed the distribution of air quality values.
* **Boxplots & Violin Plots**: Visualized the spread and density of values across different measures.
* **Barplots**: Displayed top geographic locations and most frequent categories.
* **Countplots**: Highlighted the distribution of geographic types.
* **Time Series Plot**: Tracked changes in average pollution levels over time.
* **Pie Chart**: Illustrated the proportion of each geographic type in the dataset.
* **Heatmap**: Offered visual insight into correlations between numerical columns.

The EDA process provided valuable context and patterns in the dataset, setting the stage for further interpretation, modeling, or policy recommendations based on air quality trends in different regions and timeframes.

# Analysis on Dataset

The analysis phase focused on deriving insights from the cleaned air quality dataset by examining trends, distributions, and key indicators. The goal was to identify high-risk areas, compare pollutant types, and understand how air quality varied across geography and time.

**1. Distribution of Air Quality Values**

A histogram with a kernel density estimate (KDE) was used to visualize the distribution of the Data Value column. The distribution was moderately skewed, indicating that while most air quality readings were within a normal range, there were some high pollution events acting as outliers.

**2. Comparison by Measure Type**

Boxplots and violin plots were generated to compare the spread and concentration of Data Value across different air quality measures. This helped to:

* Identify which pollutants had the highest variance
* Highlight measures that regularly recorded higher values
* Detect outliers within each pollutant category

The bar plot showing **average data values by measure** further confirmed which pollutants posed greater risk on average.

**3. Geographic Analysis**

The analysis of Geo Place Name revealed:

* The **top 10 most frequently recorded locations**, useful for identifying hotspots of consistent monitoring or higher pollution concern.
* The **maximum recorded data value by location**, showing which cities or districts experienced extreme pollution levels.

The **countplot of Geo Type Name** indicated whether the data was mostly collected from cities, districts, boroughs, or other administrative zones.

A **pie chart** illustrated the proportional distribution of these geographic types in the dataset.

**4. Temporal Trends**

If a Start\_Date column was present and successfully converted to datetime:

* A **time series plot** showed how average pollution levels changed over time.
* Grouping data by **year** revealed any long-term trends or seasonal fluctuations in air quality, such as increases during certain months or years.

**5. Correlation Analysis**

A correlation matrix heatmap was generated to explore relationships between numerical variables. While the dataset primarily contained one main numeric variable (Data Value), this analysis is useful when multiple quantitative indicators are available, as it can guide multivariate modeling or dimensionality reduction.

**6. Measure vs Geo Type Combinations**

The dataset was also grouped by combinations of Measure and Geo Type Name to understand:

* Which pollutants were most commonly tracked in which types of locations
* Data collection biases or gaps in geographical coverage for certain indicators

This analysis provides useful insights for both environmental researchers and policy planners regarding monitoring practices.

# Conclusion

This project successfully demonstrated the application of data cleaning and exploratory data analysis techniques on a real-world environmental dataset focused on air quality. The data, sourced from [data.gov.in](https://data.gov.in), provided valuable insights into the pollution levels across various regions and time periods in India.

Through systematic cleaning, the dataset was transformed into a more reliable and analysis-ready format. Missing values were appropriately handled, and key variables were formatted for consistency and accuracy.

The exploratory analysis revealed notable trends, including:

* Significant variation in air quality across different geographic areas.
* Certain pollutants showing consistently higher average values, indicating potentially greater health risks.
* Temporal patterns that may align with seasonal or industrial activity.
* Geographic areas with frequent high readings, highlighting potential pollution hotspots.

Visualizations such as histograms, boxplots, bar graphs, and time series plots made these patterns clearer and more interpretable. The analysis also demonstrated the importance of continuous and structured data collection for effective environmental monitoring.

Overall, this project not only provided insights into air quality patterns but also underscored the power of data-driven approaches in supporting public health, urban planning, and policy decisions aimed at improving environmental quality.

1. Future Scope

While this project provided a comprehensive exploratory analysis of air quality data, there are several avenues for extending the work to generate deeper insights and support more impactful decision-making:

**1. Integration with Weather and Traffic Data**

Air quality is influenced by multiple external factors such as temperature, humidity, wind speed, and vehicular emissions. Combining this dataset with meteorological and traffic data could help establish causal relationships and improve the understanding of pollution sources.

**2. Predictive Modeling**

With a clean and well-structured dataset, machine learning models could be developed to:

* Forecast air quality levels based on historical trends
* Predict high-pollution events
* Identify anomalies or sudden spikes in pollutant levels

These models can aid in proactive policy decisions and public health advisories.

**3. Geographic Visualization (GIS)**

Implementing spatial mapping using Geographic Information Systems (GIS) can provide a visual representation of pollution levels across different regions. This would make the data more intuitive and helpful for location-specific analysis and planning.

**4. Real-Time Data Analysis**

Extending the system to handle real-time data feeds from air quality monitoring stations can enable live dashboards and alerts, which are crucial for timely public health interventions.

**5. Deeper Temporal Analysis**

A more granular time-series analysis—such as daily, weekly, or seasonal trends—could help in identifying recurring patterns, festival-related spikes (e.g., Diwali), or long-term improvements due to policy changes.

**6. Public Health Impact Assessment**

Linking air quality data with health outcome datasets (e.g., respiratory illnesses or hospital admissions) can quantify the impact of pollution on public health, supporting stronger environmental regulation and awareness campaigns.

By expanding the scope in these directions, this project can evolve from basic analysis to a more advanced, decision-support system that benefits environmental authorities, researchers, and the general public.

# Snapshots:

1. GitHub Link: <https://github.com/Yash864512/Python_Project>

# 

# 

# 

# 

# 