Title: Air Quality Monitoring and Predictive Categorization in India: A PySpark-Based Exploratory Analysis and Classification Framework

Author: NARRA LOKESH REDDY

Affiliation: MALLAREDDY UNIVERSITY

Date: 23th October 2025

ABSTRACT

This study presents an exploratory data analysis (EDA) and a rule-based classification framework for real-time Air Quality Index (AQI) data sourced from various locations in India. Leveraging **PySpark (Apache Spark with Python)**, the project efficiently processes a large-scale public dataset to extract valuable insights on six key pollutants. Initial data cleaning involved handling inconsistent string values and removing records with missing pollutant averages. The analysis revealed significant variability in pollution across states and cities, identifying **Himachal Pradesh** (Avg. 44.71) as the state with the highest reported average pollution and **PM10** (Avg. 55.75) as the most concentrated pollutant. The ultimate outcome is a structured approach to categorize air quality into **Good, Moderate, Poor, and Severe** indices based on standard pollutant thresholds, laying a foundation for future machine learning-driven AQI prediction models.

1. INTRODUCTION

Air pollution remains a critical public health and environmental challenge globally, particularly in rapidly developing nations like India. The increasing concentrations of criteria pollutants, such as Particulate Matter ($PM_{2.5}$ and $PM_{10}$), pose severe risks to respiratory and cardiovascular health. Effective environmental governance requires timely, accurate, and scalable analysis of continuous air quality monitoring data.

This paper addresses the analytical challenge posed by the sheer volume and velocity of real-time air quality data by implementing a solution on the **PySpark** platform. The study utilizes a publicly available dataset to perform comprehensive data processing and **Exploratory Data Analysis (EDA)**. The core contribution is the establishment of a robust, foundational framework for automatically classifying air quality data into recognized **AQI categories** based on calculated pollutant averages.

2. METHODOLOGY

2.1 Data Acquisition and Environment

1. **Source**: Real-Time Air Quality Index for various Indian locations (Data.gov.in).
2. **Platform**: PySpark running within a Jupyter Notebook environment.
3. **Libraries**: Primary libraries included SparkSession for environment management and pyspark.sql.functions for optimized column operations (e.g., avg, round, when).

2.2 Data Preprocessing and Cleaning

The raw dataset was loaded and inspected, revealing several quality issues, notably inconsistent representations of missing values:

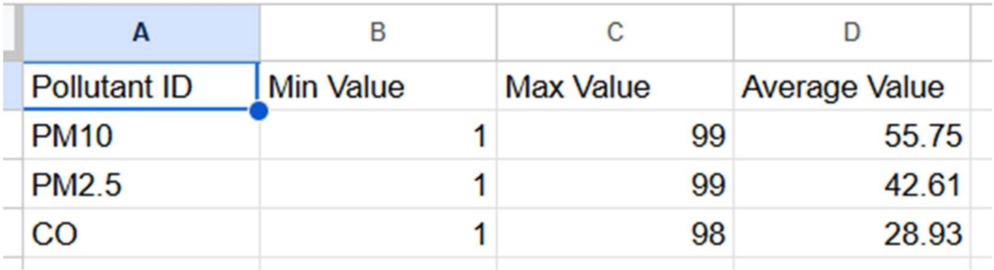
1. **Handling Missing Data**: The string "NA" was uniformly replaced with the Python-equivalent None to enable proper null handling.
2. **Data Integrity**: Rows where the primary variable of interest, pollutant_avg, contained a missing value were dropped, resulting in the clean DataFrame, df_clean. This step ensured the reliability of subsequent calculations.
3. **Monitoring Scale**: The cleaning process confirmed data consistency across **470 unique monitoring stations**.

3. RESULTS AND ANALYSIS (EDA)

3.1 Pollutant Concentration and Frequency

Analysis of the cleaned data provided insights into the ubiquity and intensity of monitored pollutants:

1. **Most Frequently Monitored Pollutants (Top 3)**:
    - **CO** (Carbon Monoxide): 445 records
    - **OZONE**: 436 records
    - **NO2** (Nitrogen Dioxide): 431 records
2. **Pollutant Concentration Ranking (Average Value)**:
    - The most concentrated pollutant by average value was **PM_{10}** (55.75), followed closely by 42.61.This highlights particulate matter as a significant problem across the monitored regions.



A	B	C	D
Pollutant ID	Min Value	Max Value	Average Value
PM10	1	99	55.75
PM2.5	1	99	42.61
CO	1	98	28.93

3.2 Geographical Hotspot Analysis

The project grouped the data by geography to pinpoint areas experiencing the highest pollution burden:

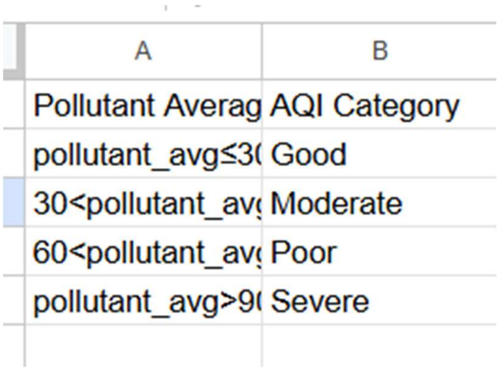
1. **Top 3 States by Average Pollutant Level**:
    - **Himachal Pradesh**: 44.71
    - **Jharkhand**: 40.0
    - **Delhi**: 33.92
2. **Top 3 Cities by Average Pollution**:
    - **Chikkaballapur**: 88.5
    - **Gummidipoondi**: 67.6
    - **Mandikhera**: 56.75

4. AQI CLASSIFICATION FRAMEWORK

A rule-based classification model was implemented to instantly translate the raw quantitative pollutant averages into a qualitative **Air Quality Index (AQI)** category. This framework serves as a critical first step towards automated reporting and prediction.

4.1 Classification Rules

The categorical assignment was based on the numerical threshold of the pollutant_avg:



A	B
Pollutant Averag	AQI Category
pollutant_avg≤3	Good
30<pollutant_av	Moderate
60<pollutant_av	Poor
pollutant_avg>9	Severe

4.2 Identification of Critical Records

To validate the need for immediate intervention, a filter identified all individual records where the pollutant_avg exceeded 50, placing them in the 'Moderate', 'Poor', or 'Severe' categories.

1. **Example Critical Pollutants (>50 Avg.)**:
    - Samastipur, **PM2.5**: 83 (Categorized as **Poor**)
    - Rajgir, **PM10**: 87 (Categorized as **Poor**)
    - Vijayawada, **PM2.5**: 56 (Categorized as **Moderate**)

5. CONCLUSION AND FUTURE SCOPE

5.1 Conclusion

This research successfully demonstrated the application of **PySpark** for large-scale **Air Quality data analysis**. The project effectively addressed data ingestion, cleaning, and complex aggregations to highlight key pollutants, most notably $PM_{10}$ and $PM_{2.5}$, and significant geographical disparities in air quality. The implemented rule-based AQI categorization framework offers a foundational classification capability essential for real-time environmental monitoring.

5.2 Future Work

1. **Machine Learning Integration**: The next phase will involve utilizing the categorized data for supervised machine learning (Classification models) to **predict** the AQI category based on pollutant inputs, moving beyond descriptive analysis to predictive modeling.
2. **Advanced Feature Engineering**: Incorporating temporal features (seasonal trends from last_update data) and geospatial features (clustering latitude/longitude data) to improve model accuracy and enhance predictive power.
3. **Data Visualization**: Develop a centralized dashboard to visually represent the spatial distribution of the generated AQI categories and highlight temporal trends.

6. REFERENCES

1. Project Data Source: Real-time Air Quality Index data.
2. PySpark Documentation: Apache Spark and PySpark libraries for distributed processing.
3. PySpark Output Logs: Dataframe schemas, aggregations, and counts.
4. Academic Literature on AQI Categorization (for methodology rationale).