**CRISP-DM Business and Data Understanding Report: Analysis of Air Quality Data in India**

Table of Contents

[1. Introduction: 3](#_Toc142363675)

[2. Business Understanding: 3](#_Toc142363676)

[3. Data Understanding: 3](#_Toc142363677)

[4. Insights from Data Understanding: 4](#_Toc142363678)

[5. Conclusion and Next Steps: 4](#_Toc142363679)

[6. Reference 4](#_Toc142363680)

# **Introduction:**

Data mining projects can now be organised with the help of the Cross-Industry Standard Process for Data Mining (CRISP-DM). This report examines the application of the first two steps of the CRISP-DM methodology to the air quality data in India: Business Understanding as well as Data Understanding.

# **Business Understanding:**

* 1. ***Problem Definition:*** Many cities in India regularly have air pollution levels that are well above what is considered safe by the World Health Organisation. Air pollution is bad for people's health, the environment, as well as the bottom line, so it's important to do some serious research to figure out what can be done about it.
  2. ***Objectives:***
* Identify the main sources of pollution in urban India.
* Learn the patterns of pollution over time.
* Call attention to areas with chronically hazardous air quality.
* Facilitate the development of predictive models and suggestions.
  1. ***Stakeholders:*** Those most affected by emissions are the general public, government agencies, policymakers, as well as major polluting industries (Ziv, and Parmet, 2022).

# **Data Understanding:**

* 1. ***Data Source:*** The "Air Quality Data in India" dataset was obtained from Kaggle. During certain time periods, it details the levels of various air pollutants in a number of Indian cities.
  2. ***Initial Data Exploration:*** When first inspecting the data, you'll notice that it's split up into several columns, each of which represents a different type of pollutant. The data in every row represents averages from different cities' monitoring stations over the course of a day.
  3. ***Data Quality:***
* **Missing Values:** Some days as well as time intervals lack data for certain pollutants. These voids must be filled by either imputed data or by excluding the affected records.
* **Consistency:** Some cities take readings more frequently than others, so it's important to compare how often data is recorded between locations.
* **Outliers:** Outliers have been detected in early visualisations, which may be the result of instrument malfunctions or particularly severe instances of pollution.
  1. ***Data Distribution and Trends:***
* **Temporal Trends:** An initial time series plot indicates that pollution levels fluctuate throughout the year, with some pollutants reaching their peaks at certain times of the year (Nodeh, et. Al., 2020).
* **Spatial Distribution:** There is clear evidence that pollution levels are consistently higher in some cities than in others, particularly those in the northern belt.
* **Correlations:** There appears to be a link between certain pollutants, which could indicate shared origins as well as interdependent atmospheric reactions.

# **Insights from Data Understanding:**

* 1. ***Pollutant Levels:*** Many cities have dangerously high concentrations of fine particulate matters (PM2.5 and PM10). The ability of these pollutants to reach deep into the lungs is of particular concern.
  2. ***Temporal Patterns:*** Stubble burning, decreased wind speeds, as well as increased vehicle emissions may all contribute to a worsening of air quality during the winter months in North India.
  3. ***Regional Disparities:*** Consistently high pollution levels in cities like Delhi, Kanpur, as well as Varanasi highlight the need for region-specific interventions.

# **Conclusion and Next Steps:**

The project now has a well-defined course thanks to the work done in the Business as well as Data Understanding phases. Pollutant levels, especially PM2.5 and PM10, in many Indian cities have reached crisis proportions, highlighting the urgency of the situation. Despite its limitations, the dataset provides a complete picture of the air quality situation in India.

# **Reference**

Ziv, B. and Parmet, Y., 2022. Improving nonconformity responsibility decisions: a semi-automated model based on CRISP-DM. *International Journal of System Assurance Engineering and Management*, *13*(2), pp.657-667.

Nodeh, M.J., Calp, M.H. and Şahin, İ., 2020. Analyzing and processing of supplier database based on the cross-industry standard process for data mining (CRISP-DM) algorithm. In *Artificial Intelligence and Applied Mathematics in Engineering Problems: Proceedings of the International Conference on Artificial Intelligence and Applied Mathematics in Engineering (ICAIAME 2019)* (pp. 544-558). Springer International Publishing.