FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY

BIT34503 DATA SCIENCE

ANALYZING AIR QUALITY AND HEALTH IMPACT IN INDIA

GROUP MEMBERS	:	NURKHAIRINA BALQIS BINTI MOHAMMAD JOE (AI220206)
		NURUL JANNAH BINTI KAMARUL ZAMAN (AI220147)
		NURULAINA NISA BINTI SAHANUAN (AI220179)
		TUAN KHALIDAH SYAZWANA BINTI TUAN MOHD KASMAWI (AI220118)
		TUAN NUR RIFAQIAH BIN TUAN HANIZI (AI220040)
SECTION	:	05
LECTURER	:	DR. NORHAMREEZA BINTI ABDUL HAMID

TABLE OF CONTENT

1.0 Introduction

Air pollution is one of the growing concerns in India, with ever-increasing levels of pollutants affecting public health and quality of life. This project focuses on examining air quality and its health impacts using advanced data-driven analytics techniques. These include descriptive, diagnostic, predictive, and prescriptive analytics, each offering a different lens through which to analyze and interpret data. These approaches will be harnessed in an effort to present a comprehensive understanding of the complex relationship between air pollution and its health consequences.

This project focuses on the analysis of air pollution trends, their health and economic impacts, and mitigation strategies. Particular attention is given to the most affected areas and sensitive populations. Findings are therefore not just insight-based but actually targeted toward those who require them most.

The methodology applies descriptive analytics to analyze the trends of pollution, diagnostic analytics to find out the causes and health impacts, predictive analytics to forecast future scenarios, and prescriptive analytics to propose effective mitigation strategies. This holistic approach ensures that the project will provide actionable insights to address air pollution and its health impacts comprehensively.

2.0 Problem Statement

Air pollution in India has reached alarming levels and seriously affects the health of its people. With highly concentrated PM2.5, PM10, and NO2, respiratory illnesses have increased, while cardiovascular diseases and life expectancy have been reduced significantly. This is a crisis that needs urgent attention, keeping in view the lives of millions of citizens, particularly those in urban areas where pollution levels are high.

Even though air quality monitoring data are available, integrated analysis that links the pollution trend to health outcomes is still a gap. Most of the current studies focus either on air quality data or on health impacts separately, hence leaving a big gap in how these factors interlink. This, therefore, calls for an approach that will synthesize data across domains comprehensively.

Currently, policymakers and healthcare professionals do not have sufficient insights to address the twin challenges of air pollution and its health consequences. In view of the absence of definite evidence and lack of relevant recommendations, efforts at mitigating the impacts of air pollution are fragmented and scanty. This presupposes an urgent need for data-driven solutions that can inform effective policy making and healthcare interventions.

3.0 Objective

Every problem requires a systematic approach to find practical solutions. This project focuses on using data-driven analytics to uncover trends, predict future impacts, and recommend strategies for improvement. By addressing specific goals, it aims to contribute to the fight against air pollution and its health effects.

1. To identify patterns by analyzing air quality data, and assess its impact on public health metrics like respiratory diseases and mortality rates.
2. To investigate the relationship between air pollution levels and health outcomes in order to provide evidence-based findings.
3. To help policymakers design effective interventions by offering data-driven insights and recommendations for air quality improvement and public health protection.

4.0 Technical Tools

In order to analyze the PM 2.5 levels and their health impacts in India, the project incorporates technical tools and techniques for data processing, visualization, and analytical insights. The emphasis is on using Power BI and Data Analysis Expressions (DAX) to provide an in-depth understanding of air quality trends and their implications.

4.1 Software and Tools

The main tool utilized in this project for data analysis and visualization is Power BI. Because of its strong capabilities, interactive dashboards and reports may be created. The analysis of PM2.5 data patterns and their health correlations requires sophisticated computations and aggregations, which are made possible by the use of Data Analysis Expressions (DAX), which deepens the analysis.

4.2 Dataset

Four datasets are being used for this project regarding the PM2.5 concentration readings gathered from several Indian cities that make up the dataset. Reliable websites that offer precise and consistent air quality measurements, like Kaggle and GitHub, are examples of data sources. Government health surveys and reliable international organizations like the WHO provide further statistics on health implications, such as hospital admissions and the incidence of respiratory diseases. Descriptive and diagnostic analyses are based on these datasets.

4.3 Programming Language

Although Python and other programming languages are not used in this project, DAX is essential to the transformation and analysis of data in Power BI. Calculations for measures like yearly PM2.5 averages, annual variations, and correlations between PM2.5 levels and health effect indicators are made easier using DAX formulas.

4.4 Output

Interactive Power BI dashboards displaying the summary of the Air Quality Index (AQI) , city-by-city heatmaps, and trends in PM2.5 levels are among the project's deliverables. DAX-powered computations yield important insights, such as pinpointing cities with consistently elevated PM2.5 values, correlating the PM2.5 levels with the health impact, and forecasting the PM2.5 levels. The results show the cities with high PM2.5 levels, the average PM2.5 levels in 10 years, and also offer practical suggestions for reducing air pollution and its negative health effects.

5.0 Solution Architect

This project's solution architecture centers on an end-to-end procedure intended to efficiently extract, process, and evaluate PM2.5 data. The elements in this section include data cleaning and transformation, descriptive analytics, diagnostic analytics, predictive analytics, and prescriptive analytics.

5.1 Data Cleaning and Transformation

The first step in cleaning data in Power BI is to load the datasets into the application. When dealing with missing or null values in the Power Query Editor, tools like Transform > Replace Values or the Fill Down/Up feature are used to replace them with the mode or placeholders for categorical data and the mean or median for numerical data. By using filters or conditions to eliminate extreme values and showing data distributions like scatter plots or histograms, outliers are eliminated. Lastly, columns are renamed appropriately and units along with the data format are standardized to improve data consistency.

Using the Merge Queries option, we append datasets for data transformation, including PM2.5 and From Date values. Additionally, we utilize Group By or Pivot/Unpivot Columns to modify data and aggregate it by qualities, such as PM2.5, to calculate the average. The data is verified after it has been produced by looking at transformation previews and using visualizations such as scatter plots or line charts to see any irregularities. The append query is applied after the data has been cleaned and transformed, and then the cleansed dataset is loaded into Power BI for additional analysis and visualization. After cleaning, the data is exported for other usage.

5.2 Descriptive Analytics

The purpose of Descriptive analytics was to summarize past data , Based on Roy, D., Srivastava, (2022) this analytics technique was used as data mining to get insights on what has happened in the past to provide insights, which summarizes distributions, trends, and patterns . It offers information on pollution levels such as trends in PM2.5, health outcomes such as hospitalizations for cardiovascular and respiratory disorders, and regional inequalities in the context of air quality and health impact assessments. Visualizations of the relationship between air pollution and health effects are made possible by tools that were used, which is Power BI, which also aids stakeholders and policymakers in seeing patterns, setting priorities

for actions, and efficiently allocating funds. In order to lessen the detrimental consequences of air pollution on public health, this analysis is essential for increasing awareness and promoting evidence-based decision-making.

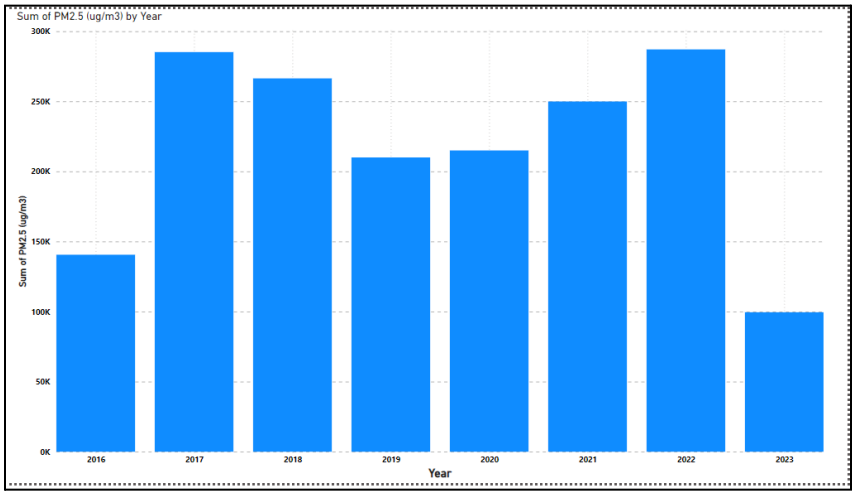


Figure 5.2.1 : Annual Trends of Key Pollutants (PM2.5) from 2016-2023

A bar chart in Figure 5.2.1 is called the Annual Trends of Key Pollutants (PM2.5) from 2016-2023. Shows the annual total of PM2.5 levels from 2016 to 2023, expressed in micrograms per cubic meter (μg/m³). An essential tool for descriptive analytics, this graphic enables observers to identify trends and variations in air quality across time. Interestingly, it shows that PM2.5 concentrations peaked in 2017 and 2022, which correspond to times of increased pollution, and then significantly decreased in 2023. Such knowledge is essential for evaluating general trends in air quality as well as the efficacy of pollution management strategies put in place over time. In the end, this research helps stakeholders understand the problems caused by air pollution and the steps taken to address them.

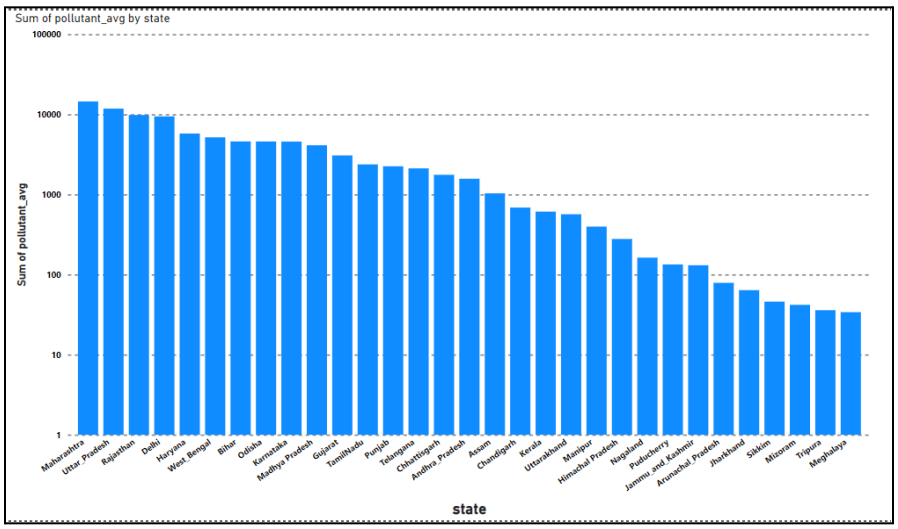


Figure 5.2.2: Graph Bar State-wise Pollution Levels

The State-wise Pollution Levels map ranks Indian states according to their overall average pollutant levels as shown in Figure 5.2.2, offering a comparative examination of pollution levels across states. Significant regional differences in air quality are highlighted by this descriptive analytics graphic, with the most polluted states being Rajasthan, Uttar Pradesh, and Maharashtra. The chart provides a clear picture of which areas are dealing with the worst air quality issues by rating the states. Policymakers may use this information to properly allocate resources and prioritize measures to mitigate pollution in the highest-ranking states, which makes it very useful. By doing this, the analysis backs specific tactics meant to enhance public health and air quality nationwide.

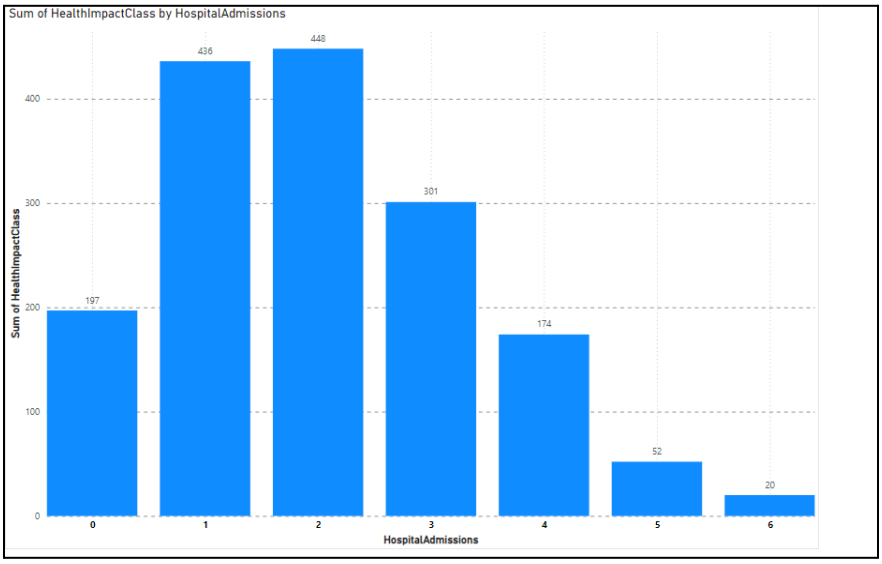


Figure 5.2.3 : Graph Bar Hospital Admissions by Health Impact Class

The graphic shown in Figure 5.2.3 is a bar chart that shows the overall number of hospital admissions by various health effect classes, which range from 1 to 5. Descriptive analytics are facilitated by this figure, which makes it evident how India's air quality impacts health outcomes. It is noteworthy that Health Impact Class 2 has the most admissions (448), followed by Class 3 (301), while Classes 4 and 5 have far lower numbers. This distribution emphasizes how urgently focused public health initiatives are needed to address the health effects of air pollution, especially in the higher-impact groups. All things considered, the analysis offers insightful information that can guide the distribution of resources and tactics meant to reduce the health hazards connected to poor air quality.

5.3 Diagnostic Analytics

Diagnostic analytics involves investigating data to uncover the reasons behind specific trends or patterns. In this study, the analysis used Dataset 1, which contains health impact information, and Dataset 2, which captures PM2.5 levels across various cities, to explore the relationship between air pollution and its effects on health.

A heatmap was created to identify cities that have consistently high levels of PM2.5 as shown in Figure 5.3.1. First, the data was grouped by city and date through DAX calculations, then visualized on a grid showing months on the X-axis and city names on the Y-axis. The intensity of color used on the heat map showed the average of PM2.5 levels, where dark color indicated higher pollution. This visualization showed an overview of cities that had poor air quality throughout the year.

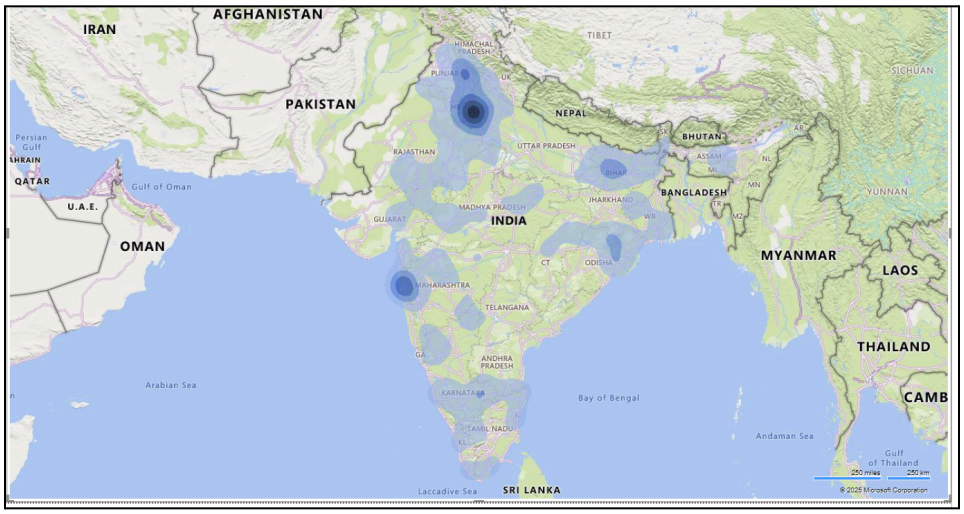


Figure 5.3.1: Heatmap of PM2.5 Levels by City and Date.

Next, a scatter plot was done to analyze PM2.5 levels versus health impacts as shown in Figure 5.3.2. The X-axis was the average PM2.5 level of each city, and the Health Impact Scores from Dataset 1 were on the Y-axis. Each dot on the scatter plot represented one city.

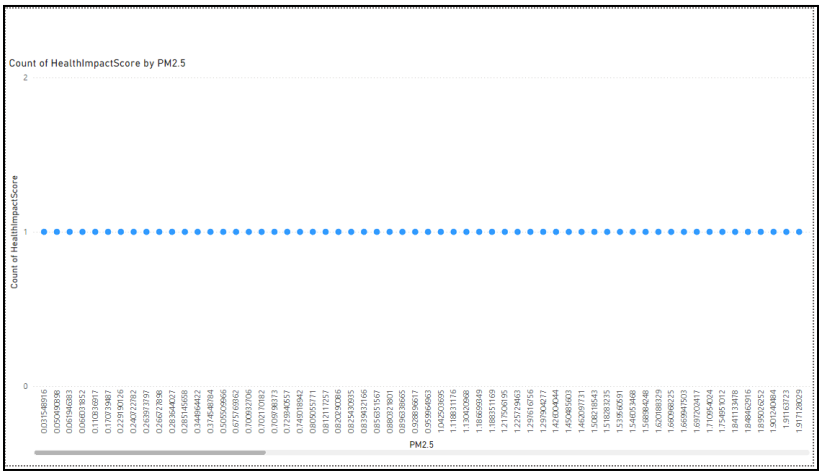


Figure 5.3.2: Scatter Plot of PM2.5 Levels and Health Impacts.

5.4 Predictive Analytics

In predictive analytics, Dataset 3 is being used to calculate and display the forecast average of PM2.5 in 10 years. Firstly, the table for average PM2.5 per year is created by using DAX query and formulas as shown in Figure 5.4.1. The data in the table created is used to generate the line graph Year as the X-axis and Average_PM25 as the Y-axis. Once the line graph is shown, the forecast setting is turned on to show the predicted future value of the average PM2.5 in 10 years. The data of the forecast value is then exported for better vision. Figure 5.4.2 shows the visualization of the forecast average PM2.5 for 10 years.

```
1 Average_PM25_Per_Year =
2 SUMMARIZE(
3     'DATASET 3',
4     'DATASET 3'[Year],          -- Group by Year
5     "Average_PM25", AVERAGE('DATASET 3'[PM2.5 (ug/m3)])  -- Calculate average PM2.5
6 )
7
```

Year	Average_PM25
2016	31.92
2017	32.55
2018	30.42
2019	23.97
2020	24.57
2021	28.67
2022	69.99
2023	53.42
2024	51.68

Figure 5.4.1: Forecasting Average PM2.5 for 10 Years

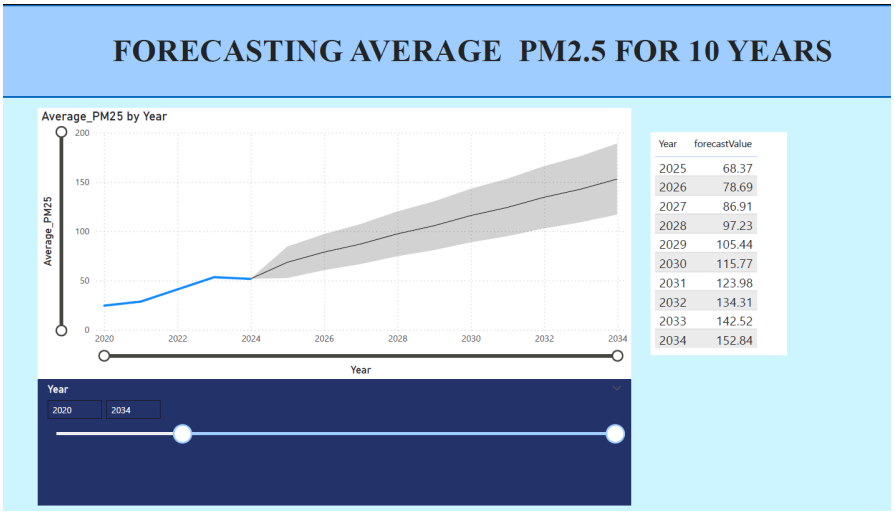


Figure 5.4.2: Forecasting Average PM2.5 for 10 Years

Based on the result of the forecast average PM2.5 in 10 years, the concentration levels of

PM2.5 are increasing year by year. The current value of PM2.5, which is in 2024, is only 57.68, meanwhile, the predicted value in 2025 increases by 16.69. In 10 years, the predicted value will be the highest, which is 152.84.

5.5 Prescriptive Analytics

In prescriptive analytics, the Dataset 3 is being used to provide actionable insights into the impact of various policies on PM2.5 levels. Prescriptive analytics goes beyond descriptive and predictive analytics by not only identifying what has happened and what might happen but also recommending actions to achieve desired outcomes. Firstly, the measures for Transport Policy Impact and Industrial Policy Impact are created using DAX query and formulas as shown in Figure 5.5.1. The data in the measure created is used to generate the Policy Impact Cards.

```
1 Transport_Impact =
2 VAR ReductionRate = 0.25
3 RETURN
4     [Average_PM25] * ReductionRate
```

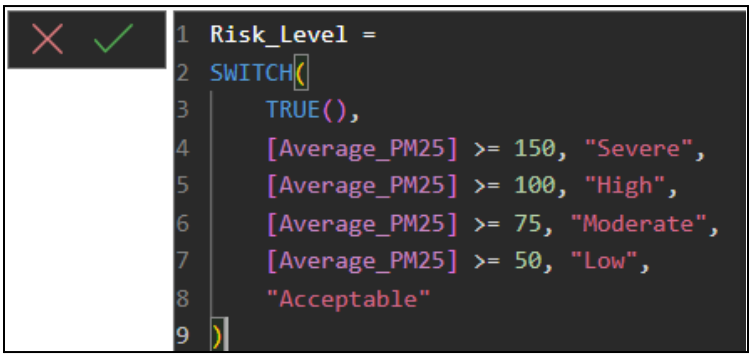
```
1 Industrial_Impact =
2 VAR ReductionRate = 0.30
3 RETURN
4     [Average_PM25] * ReductionRate
```

Figure 5.5.1: DAX Query measure of Transport and Industrial Impact

The Policy Impact Cards section of the dashboard includes three key metrics; Transport Policy Impact, Industrial Policy Impact and Average PM2.5. Transport Policy Impact visualizes the impact of transportation policies on PM2.5 levels. By analyzing historical data, we can identify trends and correlations between transportation policies and air quality improvements. Furthermore, the Industrial Policy Impact card highlights the effect of industrial policies on PM2.5 levels. It helps us understand the effectiveness of regulations and initiatives aimed at reducing industrial emissions. Lastly, the Average PM2.5 card provides an overall view of the average PM2.5 levels, serving as a benchmark to measure the success of various policies. Figure 5.5.2 shows the visualization of the Policy Impact Cards.

Figure 5.5.2: Policy Impact Cards based on the PM 2.5 Levels

Besides, it can also generate health risk levels from dataset 3. Figure 5.5.3 shows that the DAX Query and formulas are used to create the measure. The data were analyzed by predictive analytics. It will be used to do the Health Risk Matrix throughout the year.



```
Risk_Level =
SWITCH(
    TRUE(),
    [Average_PM25] >= 150, "Severe",
    [Average_PM25] >= 100, "High",
    [Average_PM25] >= 75, "Moderate",
    [Average_PM25] >= 50, "Low",
    "Acceptable"
)
```

Figure 5.5.3: DAX Query and formulas to measure health level risk based on Average PM 2.5 Levels

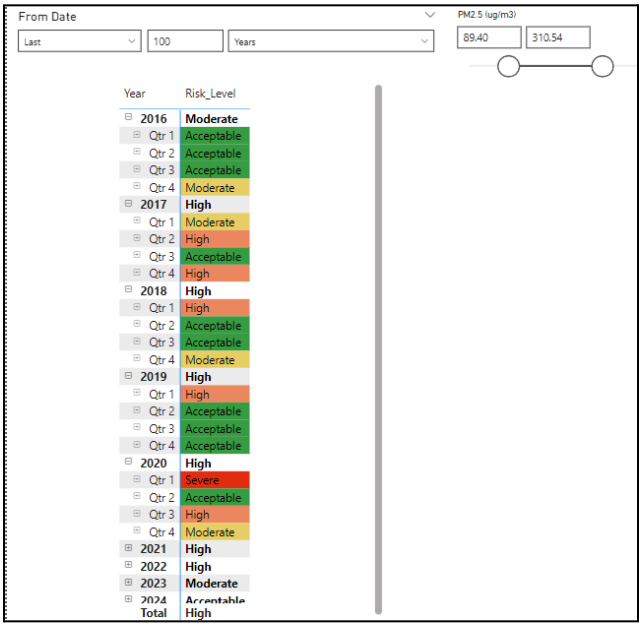


Figure 5.5.4: Health Risk Matrix throughout the year

The Health Risk Matrix is a crucial component of our dashboard, providing a comprehensive view of the health risks associated with PM2.5 levels over time as shown in Figure 5.5.4. This visual representation helps to understand the impact of air quality on public health and make informed decisions to mitigate risks.

The Matrix visual in Power BI is used to display the 'From Date' (by Year) in the Rows and 'Health_Risk_Level' in the Values. By organizing the data in this manner, it can easily identify trends and patterns in health risks over different time periods.To enhance the readability and interpretability of the Health Risk Matrix, we apply conditional formatting with color scales. This allows us to visually distinguish between different levels of health risk, making it easier to identify areas of concern.

6.0 Outcome And Insights

In the analysis of the data presented in our dashboard, we have derived several valuable insights and actionable recommendations. By examining the impacts of various policies on PM2.5 levels, we can identify effective measures and areas that require further improvement. This section highlights the key findings from our analysis and provides recommendations to enhance air quality through targeted policies and interventions.

We can gain significant insights and recommendations by evaluating the data offered in the Policy Cards. If the Transport Policy Impact card shows a considerable reduction in PM2.5 levels, it suggests that the policies are effective, and we urge that they be continued or enhanced to improve air quality even further. Conversely, if the Industrial Policy influence card reveals a moderate or low influence, it indicates that industrial policies should be reconsidered or enhanced. We advocate undertaking a thorough review of industrial emissions and enacting stronger laws if needed. Furthermore, the Average PM2.5 card provides a brief overview of air quality. If the average PM2.5 levels remain high despite the current policies, it indicates the need for a comprehensive review of all policies and possibly the introduction of new measures to address other sources of pollution.

On top of that, several critical insights and recommendations can be derived from studying the data offered in the Health Risk Matrix. The color-coded matrix allows us to immediately identify periods with significant health hazards, motivating us to look into the reasons for these high-risk periods and implement targeted steps to lower PM2.5 levels during these times. We may also evaluate the success of policies by comparing health risk levels before and after they are implemented. If particular policies are shown to be helpful in decreasing health hazards, we urge that they be continued or expanded. Furthermore, the matrix identifies patterns in health concerns, allowing us to take preemptive measures to protect public health. For instance, if we observe a seasonal increase in PM2.5 levels, we can implement temporary restrictions or advisories to minimize exposure during these periods.

7.0 Conclusion

This project provides a comprehensive analysis of air pollution trends and their health impacts in India using a data-driven approach. By employing descriptive analytics, the study highlighted historical patterns in PM2.5 levels and their regional disparities, offering a clear picture of the most affected states and cities. Diagnostic analytics identified key contributors to pollution, such as seasonal and geographical factors, and revealed the strong correlation between PM2.5 levels and increased hospital admissions for respiratory and cardiovascular diseases. These insights emphasize the critical need for targeted interventions to address the adverse effects of air pollution on public health.

Predictive analytics forecasted an alarming rise in PM2.5 levels over the next decade, highlighting the urgency for effective and sustainable mitigation strategies. Prescriptive analytics offered actionable insights by evaluating the impact of transport and industrial policies on pollution levels and providing recommendations to improve their effectiveness. The health risk matrix further underlined the importance of timely policy implementation and public health measures, especially during high-risk periods, to minimize the health burden associated with poor air quality. These findings equip policymakers and healthcare professionals with the tools and evidence needed to implement data-driven solutions.

Overall, the project demonstrates the value of integrating advanced analytics into environmental and public health planning. The use of Power BI and DAX allowed for in-depth analysis and visualization, making complex data accessible and actionable. By combining technical rigor with practical recommendations, this study serves as a critical resource for stakeholders aiming to improve air quality, protect public health, and create a sustainable future.

References

1. Balakrishnan, K., Dey, S., Gupta, T., Dhaliwal, R. S., Brauer, M., Cohen, A. J., ... & Dandona, L. (2019). The impact of air pollution on deaths, disease burden, and life expectancy across the states of India: the Global Burden of Disease Study 2017. The Lancet Planetary Health, 3(1), e26-e39.

2. El Kharoua, R. (2023). Air Quality and Health Impact Dataset. Kaggle. Retrieved January 6, 2025, from https://www.kaggle.com/datasets/rabieelkharoua/air-quality-and-health-impact-dataset

3. Gordon, T., Balakrishnan, K., Dey, S., Rajagopalan, S., Thornburg, J., Thurston, G., ... & Nadadur, S. (2018). Air pollution health research priorities for India: Perspectives of the Indo-US Communities of Researchers. Environment international, 119, 100.

4. KSTWKIV. (2023). Real-Time Air Quality in India Analytics with Python. GitHub. Retrieved January 6, 2025, from https://github.com/kstwkiv/real-time-air-quality-in-India-Analytics-with-Python/blob/main/real%20time%20air%20quality%20dataset.csv

5. Pandey, A., Brauer, M., Cropper, M. L., Balakrishnan, K., Mathur, P., Dey, S., ... & Dandona, L. (2021). Health and economic impact of air pollution in the states of India: the Global Burden of Disease Study 2019. The Lancet Planetary Health, 5(1), e25-e38.

6. Roy, D., Srivastava, R., Jat, M., Karaca, M.S. (2022). A Complete Overview of Analytics Techniques: Descriptive, Predictive, and Prescriptive. Decision Intelligence Analytics and the Implementation of Strategic Business Management. https://doi.org/10.1007/978-3-030-82763-2_2