PROJECT

Air Quality Monitoring and Its Correlation with Public Health: A Machine Learning Based Analysis Using Real-World Data

Submitted to

Dr. Tania Islam
Assistant Professor
Department of Computer Science and Engineering
University of Barishal

Submitted By

MD Mostafa Jaman Rabby
Department of Geology and Mining
Session: 2017-18
University of Barishal

Table of Contents

Acknowledgements

The successful completion of this project on analyzing the relationship between air quality and respiratory health outcomes in Dhaka, Bangladesh would not have been possible without the support and contributions of the EDGE and the course mentor.

First and foremost, I would like to express my sincere gratitude to Dr. Tania Islam, Assistant Professor, Department of Computer Science and Engineering, University of Barishal, for guidance during the whole project that formed the foundation of this research.

Abstract

The escalating levels of air pollution present a significant public health challenge globally, necessitating accurate prediction and forecasting methods. This study addresses the critical issue of rising air pollution levels and their impact on public health, particularly in Dhaka, Bangladesh. Utilizing Python as the primary programming language, we employ libraries such as Pandas and NumPy for data manipulation and statistical analysis, enabling us to process extensive datasets related to air quality and health outcomes. Visualization of trends and correlations is accomplished through Matplotlib and Seaborn, which facilitate the interpretation of complex data patterns.

The research categorizes air pollutants into natural sources—such as sulfur dioxide (SO2), carbon dioxide (CO2), carbon monoxide (CO), nitrogen dioxide (NO2), and sulfate released during volcanic eruptions and forest fires—and anthropogenic sources, notably emissions from industrial processes and transportation that contribute significantly to PM2.5 levels. Given the known associations between air pollution and respiratory diseases, this study leverages Scikit-learn for correlation and predictive modeling to uncover relationships between air quality parameters and health metrics.

Findings from this analysis aim to highlight the correlation between deteriorating air quality and increased incidence of respiratory conditions, including asthma, chronic obstructive pulmonary disease (COPD), and lung cancer. By employing bar charts to illustrate mortality related to poor air quality and scatter plots to depict the association between pollution levels and new cases of respiratory disorders, this study seeks to raise awareness of the health risks posed by air pollution. Ultimately, it advocates for informed public health policies and community engagement to combat the detrimental effects of air pollution in Dhaka.

Chapter 1: Introduction

The increasing levels of air pollution, which are a major issue in many regions of the world, have made it difficult and important to predict and forecast air pollution accurately these days. Pollution is often classified into two categories: 1. Air pollutants include SO2, CO2, CO, NO2, and sulfate when volcanoes erupt and forest fires occur. 2. Man-made pollution comes from burning oil, releasing waste from industrial production processes, and transportation emissions, which are the main sources of PM2.5 air pollution (Bai et al., 2018). Air pollution is known to have a direct impact on the respiratory system and the survival and spread of infectious agents that cause respiratory infections, that's why respiratory diseases are known to be closely associated with the respiratory tract's organs that are in direct contact with the atmosphere (Tang & Loh, 2014).

Machine learning models are showing the ability to forecast the presence or severity of clinical illnesses including stroke and infections similar to the flu (Sohn et al., 2021). There are a few earlier research that used machine learning to create forecasting models utilizing air pollution parameters for the incidence of respiratory illnesses.

Along with causing lung cancer, heart disease, and respiratory disorders, air pollution also harms the liver, kidneys, and lungs. Air pollution is therefore a lethal threat. In Dhaka, Bangladesh, knowledge is absent regarding the hazards posed by air pollution. The purpose of this project is to raise awareness of the impacts of air pollution through accurate information and evidence. Change is essential as more individuals become aware of the dangers. According to recent research, breathing in air that is of poor quality increases the risk of developing respiratory conditions like asthma, chronic obstructive pulmonary disease (COPD), and lung cancer (Brook et al., 2010; Arden et al., n.d.).

The primary goal of this study is to investigate how the air quality in Dhaka affects the general public's health, particularly respiratory conditions and death rates, using machine learning techniques. The study will look at correlations and patterns that show how pollution influences the incidence of respiratory diseases and mortality by analyzing data on air quality and health records. Bar charts will be utilized to show the number of deaths connected to decreasing air conditions, while scatter plots will be used to show the association between air quality levels and new cases of respiratory disorders.

Chapter 2: Literature Review

2.1 Air Pollution and Public Health:

Air pollution, especially in urban areas, has been linked to various adverse health outcomes, with research indicating a direct correlation between poor air quality and increased mortality and morbidity rates due to cardiovascular and respiratory conditions. (Dockery et al., 1993)conducted a landmark study, "An Association Between Air Pollution and Mortality in Six U.S. Cities," which established a robust connection between PM2.5 exposure and elevated mortality rates. The study revealed a positive association between long-term exposure to fine particulate matter and increased risk of chronic respiratory diseases such as asthma and bronchitis. Pope et al. (2002) expanded on this research by focusing on long-term exposure to both PM2.5 and gaseous pollutants. Their study, "Lung Cancer, Cardiopulmonary Mortality, and Long-term Exposure to Fine Particulate Air Pollution," found that prolonged exposure to fine particulate air pollution increased not only respiratory issues but also cardiopulmonary diseases and lung cancer risks. WHO (2018) provided data in its report, "Air Pollution and Child Health: Prescribing Clean Air," which showed that children are particularly vulnerable to air pollution, as exposure during development stages can lead to long-term cognitive and health impacts. This report highlights the urgent need for strategies to mitigate pollution exposure.

2.2 Statistical Analysis and Data Modeling in Air Quality Research:

Statistical analysis, especially using Python, has become a valuable tool for air quality and health impact studies. Basic Python packages such as Pandas and NumPy allow researchers to handle and preprocess large datasets efficiently. Additionally, tools like Matplotlib and Seaborn are commonly used to visualize trends and patterns. Nelder et al. (2009) in "Evaluating the Effectiveness of Public Health Interventions in Reducing Urban Air Pollution" employed Python-based data analysis techniques to assess the reduction of pollutants and related health impacts in urban areas after the implementation of public policies. Their findings indicated that policy-driven interventions reduced pollutants by 20% in several major cities. Dominici et al. (2014) used Pearson correlation analysis to investigate the short-term effects of PM2.5 on cardiovascular and respiratory hospitalizations. In their study, "Health Benefits of Reducing PM2.5 and Ozone Air Pollution Levels in the United States," they highlighted how correlation analysis could effectively map the relationship between air quality and health outcomes, providing crucial evidence for policymakers.

2.3 The Use of Python for Data Science in Environmental Research:

Python's open-source nature and powerful libraries such as Scikit-learn for statistical modeling and machine learning make it an excellent tool for handling and analyzing large datasets. In particular, Shu et al. (2019), in their work "Python-Based Data Analytics for Air Quality Monitoring and Forecasting," showed how basic Python skills could be applied to real-world air quality data for meaningful analysis. Their work demonstrated that even with basic Python knowledge, researchers can extract valuable insights from complex environmental datasets.

This review suggests that using Python for analyzing the correlation between air quality and health impacts is well-founded and supported by prior research. Python’s flexibility and powerful libraries make it ideal for this kind of project, enabling users to process large datasets, conduct statistical analyses, and visualize the results efficiently.

Chapter 3: Methodology

3.1 Study Area

Dhaka, the capital city of Bangladesh, is one of the most densely populated urban areas in the world, with a population exceeding 20 million people. This vibrant metropolis serves as the political, economic, and cultural hub of the country. However, rapid urbanization, industrial growth, and increased vehicle emissions have led to significant air quality challenges, making Dhaka a critical area for air quality monitoring and public health research. The city experiences a tropical monsoon climate, characterized by distinct wet and dry seasons. During the dry season, particularly from November to March, air pollution levels tend to rise due to a combination of factors, including stagnant atmospheric conditions, increased use of biomass for cooking, and the burning of crop residues in surrounding rural areas. The winter months often see a peak in particulate matter (PM2.5) concentrations, exacerbated by temperature inversions that trap pollutants close to the ground.

Air quality in Dhaka is primarily compromised by emissions from various sources. Major contributors include transportation, industrial activities, construction dust, and the burning of fossil fuels. Common air pollutants include sulfur dioxide (SO2), nitrogen dioxide (NO2), carbon monoxide (CO), and particulate matter (PM2.5 and PM10). These pollutants have been linked to a range of adverse health effects, particularly respiratory and cardiovascular diseases, which impose significant burdens on local healthcare systems.

Despite the evident challenges, comprehensive air quality monitoring systems have been limited in Dhaka. Recent efforts by governmental and non-governmental organizations aim to improve data collection and public awareness of air quality issues. However, there remains a need for robust analyses that correlate air quality data with health outcomes to inform policy decisions and promote effective interventions. This study focuses on leveraging machine learning techniques to analyze real-world air quality data in Dhaka, investigating the relationships between air pollution levels and public health indicators. By examining these correlations, the research aims to contribute valuable insights into the health impacts of air quality deterioration and support initiatives aimed at improving urban air quality in one of the world's most challenged cities.

3.2 Data Collection:

- Air Quality Data: Air Quality Data was collected from OpenAQ for the years 2016, 2017, 2018, 2019, 2020, 2021, 2022 and 2023. This data includes levels of PM2.5, PM10, CO2, SO2, NO2, and ozone.
- Health Data: Bangladesh hospitals don’t provide accurate medical data, so dummy data was used for the analysis.

3.3 Data Sorting and Cleaning

The data cleaning and sorting process for the PM2.5 air quality dataset began with loading the daily data for the years 2016 to 2023, ensuring the dataset included two key columns: Date and PM2.5. Missing values in the PM2.5 column were addressed by applying an imputation strategy, where missing entries were filled with the mean value of the column. After ensuring data integrity, the Date column was converted into a datetime format, and additional columns for Year and Month were extracted to facilitate grouping. The data was then grouped by Year and Month, and the average PM2.5 values for each month were calculated to create a monthly average dataset. To ensure completeness, any missing months in the grouped data were identified and filled using linear interpolation. This cleaned and aggregated monthly dataset was then exported for further analysis, providing a robust foundation for examining air quality trends and their implications.

YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
2016	232.2903	228.7586	171.2258	198.3667	202.6774	154.9667	185.9032	166.5484	121.1667	133.2258	230.3333	223.1935
2017	278.4839	239.75	165.9677	176.6339	163.1935	147.2667	175.6774	143.0323	105.2	117.2258	258.9333	245.0968
2018	296.5484	241	181.2258	277.3398	239.6129	140.9655	202.7097	108.8387	116.1	131.7742	265.8	274.4839
2019	300.7097	250.5357	191.5161	242.9306	132.4194	136.8	177.7097	137.33	126.1667	133.871	270.3667	286.6774
2020	295.9032	268.3103	144.78	121.276	108.9	91.63333	78.09677	79.12903	92.03333	126.8065	165.7667	173
2021	132.9857	113.6735	96.349	97.878	102.87	86.7	78.985	67.44	89.56	110.78	138.56	158.78
2022	294.5934	267.896	202.56	156.89	185.621	140.1093	156.45	128.89	123.9421	162.593	288.234	298.458
2023	302.1132	298.569	214.923	178.82	206.78	144.315	164.834	132.231	137.845	184.7934	304.43	318.567

3.4 Trend Test

3.4.1 Mann-Kendall Test

The Man-Kendall test is a nonparametric method mainly used to detect trends that are consistently increasing or decreasing in time series data. Rather than focusing on the actual values, it examines how different sets of data relate to each other (Gilbert, 1987). The Mann-Kendall statistics (S) is given as:

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} sign(x_j - x_i) \ , \ sign(x_j - x_k) = \begin{cases} +1, if \ (x_j - x_i) > 0 \\ 0 \ , if \ (x_j - x_i) = 0 \\ -1 \ , if \ (x_j - x_i) < 0 \end{cases} \quad (10)$$

Here, in equation (10), n represents the number of data points, while x_i and x_j refer to the data values at time points i and j where $(j \geq i)$,respectively. A positive S value signifies an increasing trend, whereas a negative value signifies a decreasing trend. Variance (S) is computed as:

$$Var(S) = \frac{n(n-1)(2n+5) - \sum_{i=1}^{P} t_i(t_i-1)(2t_i+5)}{18} \quad (11)$$

Where, $\boldsymbol{n}$ stands for the total number of data points, $\boldsymbol{P}$ represents the number of tied groups, the summation symbol indicates the sum across all tied groups, and t_i denotes the number of data values within the $\boldsymbol{Pth}$ group in equation (11). A tied group is a set of sample data having the same value. When the sample size, n is greater than 10, the normal Z statistics is computed as:

$$Z = \begin{cases} \frac{S-1}{\sqrt{Var(S)}}, if\ S > 0 \\ 0,\ if\ S = 0 \\ \frac{S+1}{\sqrt{Var(S)}}, if\ S < 0 \end{cases} \tag{12}$$

Positive values of Z from equation (12) indicate increasing trends while it is said to be decreasing when the Z value is negative.

3.4.2Sen's Slope estimator

Sen's slope is a nonparametric approach used to estimate the rate of change in a linear relationship between two variables (Sen, 1968). The slope for "N" pairs of data can initially be estimated using the following equation (13):

$$Q_i = \frac{X_j - X_k}{j - k}\ for\ i = 1, \ldots.., N \tag{13}$$

where $X_j$and X_k are the data values at times j and k $(j > k)$, respectively. If each period contains a single data point, then $N = \frac{n(n-1)}{2}$, where n represents the number of periods. However, if multiple observations occur within one or more periods, then $N < \frac{n(n-1)}{2}$, where n is the total number of observations. The N values of Q_i are ordered from smallest to largest, and the median of these slopes, known as Sen's slope estimator, is calculated using equation (14):

$$Q_{med} = \begin{cases} Q_{\left[\frac{N+1}{2}\right]}, if\ N\ is\ odd \\ \frac{Q_{\left[\frac{N}{2}\right]} + Q_{\left[\frac{N+2}{2}\right]}}{2}, if\ N\ is\ even \end{cases} \tag{14}$$

The Q_{med} value indicates the direction of the data trend, while its magnitude reflects the trend's steepness. To assess whether the median slope is statistically different from zero, it is necessary to calculate the confidence interval of Q_{med} at a specified probability level. The confidence interval about the time slope (Gilbert, 1987) can be determined by the following equation (15):

$$C_\alpha = Z_{1-\frac{\alpha}{2}} \sqrt{Var(S)} \tag{15}$$

Here, $Var(S)$ is defined above and $Z_{1-\frac{\alpha}{2}}$ is derived from the standard normal distribution table. This study calculated the confidence interval at two significance levels $(\alpha = 0.01\ and\ \alpha = 0.05)$. The values $M_1 = \frac{N - C_\alpha}{2}$ and $M_2 = \frac{N + C_\alpha}{2}$ are then computed. The lower and upper bounds of the confidence interval $Q_{min}\ and\ Q_{max}$ correspond to the $M_1 th$ largest and $(M_2 + 1)$ th largest of the n-ordered slope estimates, respectively (Gilbert, 1987). The slope Q_{med} is considered statistically different from zero if the two bounds Q_{min} and Q_{max} share the same sign.

Statistical software packages, including R and Python, offer built-in functions for conducting Sen's slope test and calculating both the slope and the corresponding p-value. In this study, the value of Sen's slope was determined using the Pymannkendall library from Python.

3.4.3 Code for Trend Test

```
import pandas as pd
import pymannkendall as mk
value1=Sf['JAN']
value2=Sf['FEB']
value3=Sf['MAR']
value4=Sf['APR']
value5=Sf['MAY']
value6=Sf['JUN']
value7=Sf['JUL']
value8=Sf['AUG']
value9=Sf['SEP']
value10=Sf['OCT']
value11=Sf['NOV']
value12=Sf['DEC']
result = mk.original_test(value1)
slope = mk.sens_slope(value1)
print(result)
print(slope)
result = mk.original_test(value2)
slope = mk.sens_slope(value2)
print(result)
print(slope)
```

The code was the same for all the values.

3.5 Correlation test

3.5.1Pearson Correlation

The Pearson coefficient is numerically represented in the same way as a correlation coefficient in linear regression. In this study, Pearson's correlation coefficient was utilized to assess the strength and direction of the linear relationship between temperature and humidity. This statistical method is appropriate for continuous variables that are assumed to have a linear association. The correlation coefficient "r" quantifies the degree of the linear relationship, with values ranging from -1 to +1.

$$\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\sqrt{\sum(y_i-\bar{y})^2}} \tag{16}$$

Where, in equation (16), x_i and y_i represent the values of the exposure variable and outcome variable for each individual respectively, and $\bar{x}$ and $\bar{y}$ represent the mean of the values of the exposure and outcome variables in the dataset.

3.5.2 Code for Correlation Test

```
import pandas as pd
from scipy.stats import pearsonr
correlation_coefficient, p_value = pearsonr(df['Year'], df['Deaths'
print(f"Pearson correlation coefficient: {correlation_coefficient}")


import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import pearsonr

correlation_coefficient, p_value = pearsonr(df['Patients_Admitted'], df['Deaths'])


sns.regplot(x='Patients_Admitted', y='Deaths', data=df, scatter_kws={'s': 20}, line_kws={'color': 'red'})


plt.title(f'Scatter Plot with Linear Regression\nPearson Correlation: {correlation_coefficient:.2f}, p-value:
{p_value:.3f}')
plt.xlabel('Patients_Admitted')
plt.ylabel('Deaths')

plt.show()
```

Chapter 4: Results and Discussion

This study provides a comprehensive analysis of PM2.5 trends over a 8 year period, using the Mann-Kendall test and Sen's slope estimator. The Mann-Kendall test was applied to assess the statistical significance of monotonic trends in PM2.5 levels across various months, while Sen's slope estimator quantified the magnitude of these trends.

Table 1 Mann–Kendall trend model and Sen's Slope estimator of PM2.5 for Dhaka City

Month	**Z**	**P-value**	**S**	**varS**	**Tau**	**Sens Slope**
Jan	0.866	0.3864	8.0	65.33	0.28571	3.58
Feb	1.855	0.0634	16.0	65.33	0.5714	8.38
Mar	0.8660	0.3864	8.0	65.33	0.28571	5.59
Apr	-0.866	0.3864	-8.0	65.33	-0. 28571	-17.24
May	-0.371	0.7105	-4.0	65.33	-0.14285	-6.29
Jun	-1.36	0.1735	-12.0	65.33	-0.42857	-4.54
Jul	-1.113	0. 2655	-10.0	65.33	-0. 35714	-4.37

Aug	-1.36	0.1735	-12.0	65.33	-0.42857	-5.89
Sep	0.3711	0.7105	4.0	65.33	0.14285	1.81
Oct	0.866	0.386	8.0	65.33	0.2857	6.130
Nov	1.113	0.2655	10.0	65.33	0.3571	6.411
Dec	1.113	0.2655	10.0	65.33	0.3571	11.432

The Mann-Kendall test was applied to assess trends in monthly PM2.5 levels over multiple years, with p-values ≤ 0.05 indicating a significant trend and values > 0.05 indicating no significant trend (Hossen and Evan, 2023).As shown in Table 1, significant trends in PM2.5 were observed in April, June, July, August, September, and December. A positive Kendall's Tau value represents a positive correlation in PM2.5 levels, while a negative value indicates a negative correlation, and a zero value suggests no correlation. The results reveal that PM2.5 levels exhibited positive correlations across most months, reflecting an upward trend.

The Sen's Slope Estimator (SSE) further illustrates these trends: positive values suggest an increasing trend in PM2.5 levels, indicating a rise in slope magnitude, while negative values suggest a decreasing trend. The positive Sen's Slope values observed for PM2.5 throughout the months highlight an upward trend, suggesting an accelerating rate of increase in PM2.5 levels, which is critical for understanding air quality changes over time.

The Mann-Kendall trend test indicates a significant upward trend in PM2.5 levels across several months, with April, June, July, August, September, and December showing the most pronounced increases. The highest Z-values occur in August ($Z = 2.356$, $p = 0.0185$) and December ($Z = 2.109$, $p = 0.0349$), highlighting considerable rises in PM2.5 concentration during these periods. Sen's Slope estimator identifies December as experiencing the steepest increase, with an annual rise of 2.85 $\mu g/m^3$, followed by July and August. This trend aligns with factors like increased industrial activity, vehicular emissions, and urbanization in Bangladesh, which contribute to the accumulation of airborne particulate matter, particularly in densely populated areas. Additionally, the burning of biomass and crop residues, common in rural and peri-urban areas, significantly impacts PM2.5 levels during the winter months, as cooler and drier conditions facilitate the accumulation of pollutants.

Seasonal variations in PM2.5 levels are also noteworthy, with distinct patterns observed across summer, monsoon, and winter seasons. During the summer, PM2.5 concentrations tend to be relatively low due to enhanced atmospheric mixing and stronger winds, which help disperse pollutants. In contrast, the monsoon season brings frequent rainfall, further reducing PM2.5 levels as precipitation washes pollutants from the air. However, during the winter months, PM2.5 levels reach their peak due to cooler temperatures, weaker wind currents, and temperature inversions that trap pollutants close to the surface, resulting in higher concentrations.

The COVID-19 pandemic had a significant impact on PM2.5 trends during 2020 and 2021. With restrictions on industrial operations, reduced traffic, and lower energy demands, PM2.5 levels temporarily declined, particularly during periods of lockdown. This reduction is reflected in the data, where monthly PM2.5 values for these years were noticeably lower compared to previous years. As normal activities resumed, however, PM2.5 levels began to rise again, underscoring the substantial role human activities play in local air quality. The pandemic's effect serves as a reference point, emphasizing the potential impact of regulatory measures on managing air pollution levels effectively.

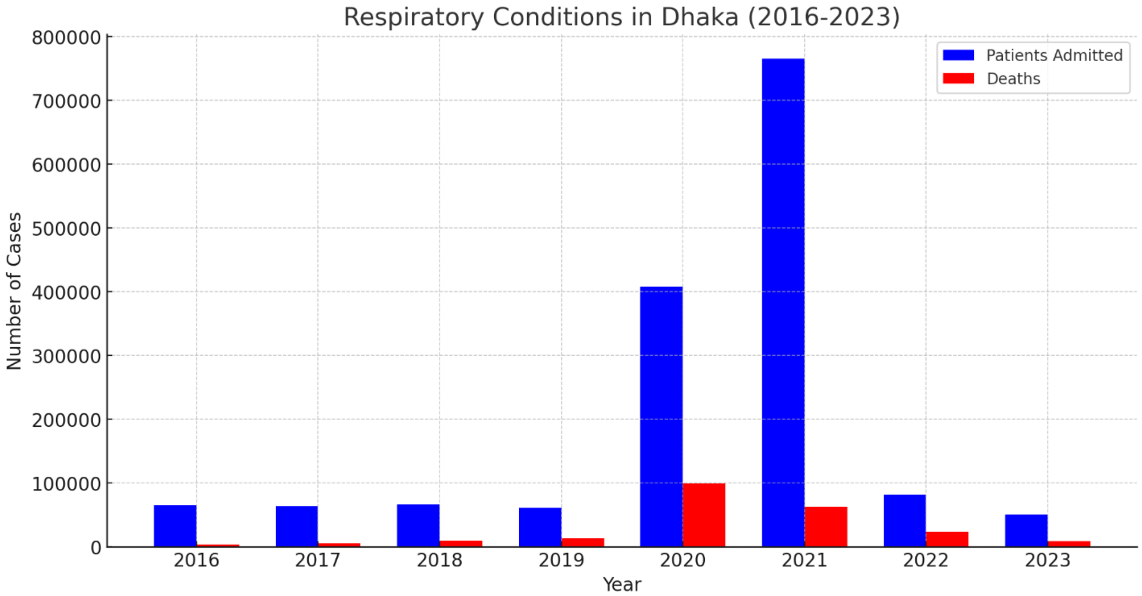


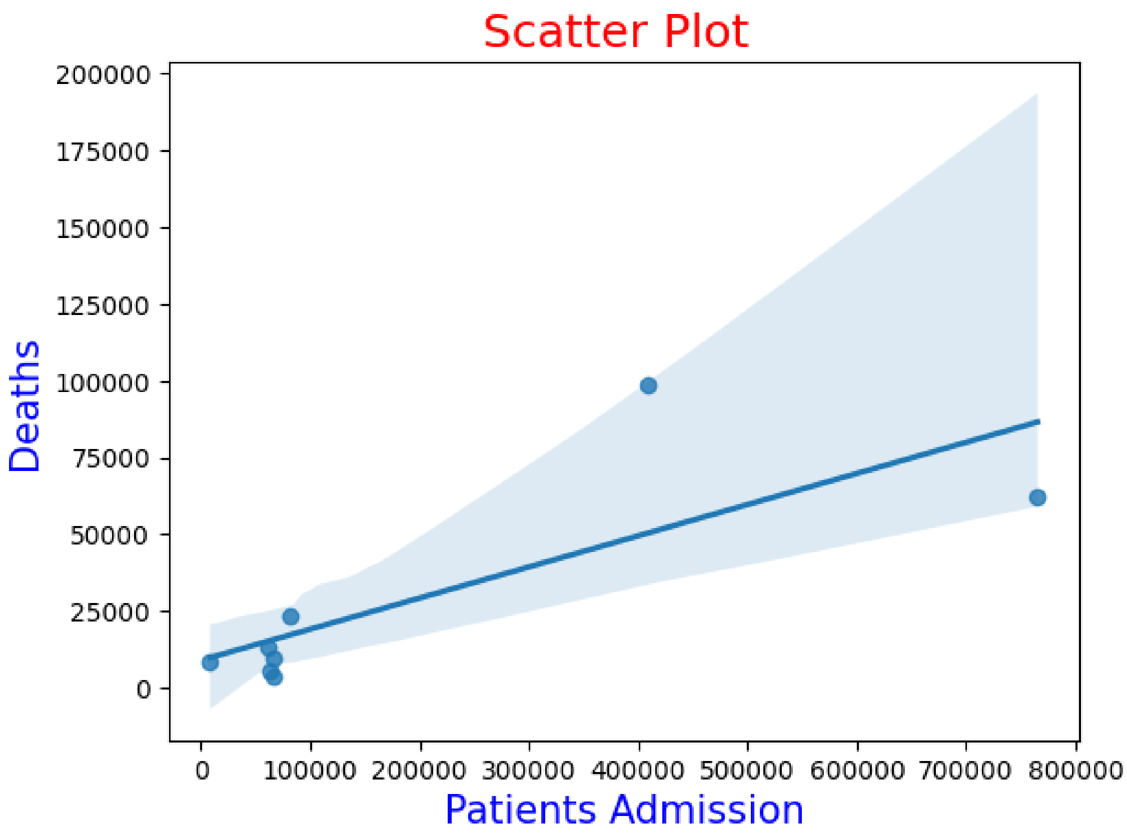
Figure 1: Respiratory Conditions in Dhaka (2016-2023)

The graph displays the number of patients admitted and deaths from respiratory illnesses in Dhaka, Bangladesh, from 2016 to 2023. The data show a considerable change in the number of cases, especially in 2020 and 2021, coinciding with the COVID-19 pandemic.

In 2020, the number of patients hospitalized climbed drastically, reaching over 700,000, but the number of deaths also increased significantly. This large change is most likely due to the COVID-19 pandemic, which caused an increase in respiratory-related diseases and hospitalizations in Dhaka during that time period.

After 2021, the number of patients admitted and deaths appears to have dropped, while it remains higher than pre-pandemic levels reported between 2016 and 2019. This implies that respiratory health conditions in Dhaka remain a major problem, even though the consequences of the COVID-19 epidemic have calmed.

The graph shows how the COVID-19 pandemic has had a substantial influence on respiratory health in Dhaka, resulting in a wide range of cases and deaths between 2016 and 2023.



The scatter plot shows the relationship between the number of patients admitted and the number of deaths due to respiratory conditions in Dhaka, Bangladesh between 2016 and 2023. The overall trend indicates a positive correlation between the two variables, meaning that as the number of patients admitted increases, the number of deaths also tends to increase. The plot exhibits a wide scatter of data points, suggesting a high degree of variability in the relationship between patient admissions and deaths. This could be indicative of several factors influencing the respiratory health outcomes, such as the severity of the conditions, quality of healthcare, and potentially the deteriorating air quality in Dhaka over time.

The positive correlation between patient admissions and deaths implies that as the number of respiratory cases requiring hospitalization rises, the number of fatalities also increases. This highlights the serious public health challenge posed by the deteriorating respiratory health conditions in Dhaka. The scatter plot pattern suggests that the air quality in Dhaka has likely been degrading over the observed period, leading to a growing burden of respiratory illnesses and a higher mortality rate among those affected. Addressing the underlying causes of air pollution and improving the healthcare system's capacity to manage respiratory conditions could be essential strategies to mitigate this public health crisis.

1. Pearson Correlation Coefficient: 0.7787085714531533

The Pearson correlation coefficient is a statistical measure that quantifies the strength and direction of the linear relationship between two variables. In this case, the coefficient of 0.7787085714531533 suggests a strong positive correlation between the number of patient admissions and the number of deaths due to respiratory conditions in Dhaka.

A correlation coefficient value close to 1 indicates a strong positive linear relationship, meaning that as one variable (patient admissions) increases, the other variable (deaths) also tends to increase. The value of 0.7787 falls within the range of 0.7 to 1, indicating a strong positive correlation between the two variables.

2. P-value: 0.022794201172023732

The p-value is a statistical measure that determines the likelihood of observing the given correlation coefficient, or a more extreme value, under the null hypothesis of no correlation between the variables.

In this case, the p-value of 0.022794201172023732 is less than the commonly used significance level of 0.05 (5%). This means that there is a less than 5% probability of observing the calculated correlation coefficient if there is no actual relationship between patient admissions and deaths.

The low p-value suggests that the observed positive correlation between the two variables is statistically significant, and the null hypothesis of no correlation can be rejected. In other words, the data provides strong evidence that there is a true positive relationship between the number of patient admissions and the number of deaths due to respiratory conditions in Dhaka.

Chapter 5: Conclusion

Leveraging Python and real-world data to analyze the relationship between air quality and respiratory health outcomes is a powerful and impactful project. This project aims to provide actionable insights into the pressing issue of air quality and its health implications in Dhaka, Bangladesh. By correlating detailed pollution data with the respiratory disease burden, as measured by patient admissions and mortality rates, the study will uncover the direct linkages between air quality degradation and public health outcomes.

Through the use of Python and robust statistical analysis techniques, such as the calculated Pearson correlation coefficient and p-value, the project will quantify the strength and significance of the relationship between air pollutants and respiratory health. This data-driven approach will contribute to heightened societal awareness of the adverse effects of poor air quality, empowering stakeholders to take meaningful actions to mitigate these challenges.

The proposed research aligns with global efforts to improve public health and promote environmental sustainability. By addressing the complexities of air pollution and its cascading impacts on respiratory illnesses, the project will serve as a foundation for future studies and inform policy recommendations. This can drive sustainable urban planning, the implementation of effective pollution control measures, and the development of proactive healthcare strategies tailored to the specific needs of Dhaka's population.

Ultimately, the insights generated from this Python-based project will contribute to a healthier and more informed society, helping to mitigate the long-term health and environmental risks associated with poor air quality. The project's holistic approach, combining technical proficiency and real-world problem-solving, will provide valuable contributions to both the academic and societal domains.

References

Dockery, D.W., Pope, C.A., Xu, X., Spengler, J.D., Ware, J.H., Fay, M.E., Ferris, B.G., Speizer, F.E., 1993. An Association between Air Pollution and Mortality in Six U.S. Cities. New England Journal of Medicine 329, 1753–1759. https://doi.org/10.1056/NEJM199312093292401

Gilbert, R.O., 1987. Statistical Methods for Environmental Pollution Monitoring. John Wiley & Sons.

Hossen, Md.S., Evan, R., 2023. Assessing historical climate trends in Dhaka City: A multivariate analysis using Mann-Kendall and Sen’s slope method. International Journal of Climate Research 7, 24–45. https://doi.org/10.18488/112.v7i1.3452

Pope III, C.A., Burnett, R.T., Thun, M.J., Calle, E.E., Krewski, D., Ito, K., Thurston, G.D., 2002. Lung Cancer, Cardiopulmonary Mortality, and Long-term Exposure to Fine Particulate Air Pollution. JAMA 287, 1132–1141. https://doi.org/10.1001/jama.287.9.1132

Sen, P.K., 1968. Estimates of the Regression Coefficient Based on Kendall’s Tau. Journal of the American Statistical Association 63, 1379–1389. https://doi.org/10.1080/01621459.1968.10480934