Term Project Report

on

**DEATHS CAUSED BY AIR POLLUTION**

A yellow and purple seal with text

Description automatically generated

**Team Members:**

Siri Sushma Penumatsa

Vaagdevi Veeramachaneni

Suguna Sri Nandini Gogusetti

**CPS585 Applied Data Engineering**

(4/22/2024)

**Section 1: OVERVIEW**

In this project, we are utilizing datasets on global air pollution from Kaggle and deaths due to air pollution from open weather API to gain insights into various factors. The datasets encompass country names, air quality index (AQI) values, carbon monoxide, ozone, as well as statistics on deaths caused by air pollution, including those due to indoor air pollution and ozone pollution.

We have taken the Open Weather API for getting the data for three different datasets namely, Historical Data, Current Data and Forecast Data. The historical data consists of deaths caused in different years from 2021, 2022, 2023, like outdoor air pollution deaths, indoor air pollution deaths and more. The historical data is taken from Kaggle, and the other two datasets have been taken from API’s. We have an open weather API, extracting current data from atmosphere like AQI values and several other chemical values. We also have forecasted data obtained from the API for which we have created average values of all the columns.

We have merged the Kaggle dataset by extracting the respective data like chemicals and AQI values and merged those values together to form a complete dataset consisting of years and their concurrent values.

**Section 2: HIGH LEVEL DESIGN**

A diagram of information storage

Description automatically generated

Our project involves the following high-level design components:  
  
**Data Ingestion:**

* The system starts by ingesting raw data from various sources such as CSV files, APIs, or databases.
* This raw data includes information about air pollution deaths, geographical coordinates of countries, and historical air quality data.

**Data Transformation:**

* Once the raw data is ingested, it needs to be transformed into a format suitable for storage and visualization.
* Transformation involves tasks like data cleaning, normalization, merging datasets, and adding derived features.
* For example, you may need to join air pollution death data with geographical coordinates to obtain latitude and longitude information for each country.

**Data Storage:**

* The transformed data is then stored in a SQLite database for efficient retrieval and management.
* SQLite is a lightweight, file-based database that is suitable for small to medium-sized datasets and provides good performance for read-heavy workloads.
* We can organize the database tables to reflect the structure of our data and optimize queries for visualization.

**Visualization:**

* Tableau is used for visualizing the data stored in the SQLite database.
* Tableau provides a user-friendly interface for creating interactive and insightful visualizations such as maps, charts, and dashboards.
* Visualizations can be customized to display various metrics related to air pollution deaths, indoor and outdoor air quality, trends over time, geographical distribution, and comparisons between different countries or regions.

**Section 3: IMPLEMENTATION AND RESULTS**

**3.1 Data Collection and Preprocessing**  
We started by collecting data on air pollution-related deaths from the provided CSV file named "Air Pollution Deaths.csv". The data included information on the entities (countries), years, and air pollution-related deaths. We preprocessed the data by renaming the "Entity" column to "countries" for better clarity and consistency. We also extracted unique years and countries from the dataset to facilitate further analysis.

**3.2 Geolocation and API Integration**  
We utilized the GeoPy library to obtain latitude and longitude coordinates for each country in the dataset. These coordinates were then used to fetch air pollution data using the OpenWeatherMap API. We implemented functions to calculate Unix timestamps for the start and end of each year and applied them to create new columns in the DataFrame. Additionally, we fetched air pollution data for each country by making API requests based on their coordinates and time intervals.

**3.3 Data Analysis and Visualization**  
We performed correlation analysis to explore the relationships between various air pollutants (e.g., CO, NO, NO2, SO2, PM2.5, PM10) and air pollution-related deaths. The correlations were computed using the Pandas library, and the results were interpreted to understand the impact of different pollutants on mortality rates due to air pollution.

A screenshot of a computer

Description automatically generated

When we examined various gases and air pollution-related deaths, we found some interesting connections. Carbon monoxide (CO), nitric oxide (NO), and nitrogen dioxide (NO2) seem to have a link with more deaths from air pollution, with correlation coefficients of approximately 0.09, -0.06, and -0.15, respectively. The connection is weaker with NO NO2 and CO. However, sulfur dioxide (SO2) doesn't show as strong a connection, with a correlation coefficient of approximately -0.08. Overall, it looks like higher levels of CO, NO, and NO2 does not mean more deaths from air pollution.

A screenshot of a computer

Description automatically generated

When we checked the air for tiny particles like PM2.5 (Particulate Matter ≤ 2.5 micrometers) and air pollution-related deaths, with a coefficient of about 0.274, indicating a moderate relationship. Similarly, PM10 (Particulate Matter ≤ 10 micrometers) also shows a strong positive correlation, approximately 0.248. These findings suggest that higher levels of both PM2.5 and PM10 are linked to increased air pollution-related deaths.

**3.4 Database Storage**  
To ensure data persistence and accessibility, we stored the processed data frames into an SQLite database named "data.db". The historical data, current data, and forecasted data were stored in separate tables within the database, enabling easy retrieval and manipulation for future analyses.  
  
**3.5 Tableau Visualization**

**Historical Dataset:**

**a. Bar Chart of Deaths By Each Pollution Type**

**A screenshot of a graph

Description automatically generated**

We looked at how different types of pollution affect the number of deaths. In the historical data, we found that Total Indoor Deaths, Total Outdoor Deaths, and Total Outdoor Ozone Deaths all had the same number of deaths, which was 40,407. This means that each type of pollution seemed to cause a similar number of deaths in the past.

**b. Number of Deaths in Each Country**

**A screenshot of a computer screen

Description automatically generated**

Our visualization maps out air pollution-related deaths by country, with darker regions indicating higher mortality rates. This helps identify areas most affected by air pollution fatalities, aiding in targeted interventions and policy decisions.

**c. Time Series Analysis of AQI**

**A screenshot of a graph

Description automatically generated**

Our bar graph depicts a decrease in the Air Quality Index (AQI) over time. This downward trend indicates an improvement in air quality, suggesting potential success in pollution control measures or environmental interventions. This analysis provides valuable insights into the effectiveness of air quality management efforts and highlights areas of progress in mitigating pollution levels.

**d. Top 10 most polluted Countries**

**A screenshot of a computer

Description automatically generated**

These countries exhibit the highest levels of pollution, as indicated by their Air Quality Index (AQI) values. Addressing pollution in these regions is critical for protecting public health and preserving the environment.

**e. Average AQI across different Locations**

**A map of the world

Description automatically generated**

We computed the mean Air Quality Index (AQI) across different areas to assess overall air pollution levels. This metric offers a concise overview of pollution trends across diverse locations, aiding in understanding and addressing environmental challenges effectively.

**f. SUM of Pollutants**

**A screenshot of a computer

Description automatically generated**

This metric represents the sum of different pollutants present in the air. It provides a concise summary of the variety and extent of pollutants in each environment, offering insight into potential health risks and environmental concerns. Here, Co seems to be the highest.

**Current Data**

**a.** **Create histograms to visualize the frequency distribution of AQI values**

**A screenshot of a graph

Description automatically generated**

To visualize the frequency distribution of AQI values in the current dataset, histograms can be created by plotting AQI values on the x-axis and their corresponding frequencies on the y-axis. Adjustments such as selecting appropriate bin widths or ranges ensure clarity in the visualization. Analyzing peaks, clusters, and patterns in the histogram provides insights into the concentration levels of air pollutants present in the dataset, aiding in understanding and addressing environmental concerns effectively.

**b.** **Average AQI across different Locations**

**A map of the world

Description automatically generated**

Calculating the average Air Quality Index (AQI) across different locations provides a comprehensive understanding of air pollution levels. This metric helps gauge overall pollution trends and assess the effectiveness of pollution control measures. By averaging AQI values from diverse geographic locations, we gain insights into the broader environmental impact and potential health risks associated with air pollution.

**c. SUM of Pollutants**

**A screenshot of a computer

Description automatically generated**

This metric represents the total count of different pollutants present in a given environment. It offers a concise overview of the overall pollution level, providing insight into the extent and variety of pollutants affecting the area.

**Forecast Data**

**a.Average AQI across different locations**

**A map of the world

Description automatically generated**

This metric calculates the mean Air Quality Index (AQI) across various geographic locations, offering a concise overview of pollution levels. It provides insights into overall air quality trends and helps assess environmental health risks across different regions.

**b. SUM of Pollutants**

**A screenshot of a computer

Description automatically generated**

This metric represents the total count of various pollutants present in a specified environment. It offers a comprehensive snapshot of pollution levels, indicating the combined impact of multiple pollutants on air quality or environmental health.

**Section 4: CONCLUSION**

In summary, our project embarked on a comprehensive exploration of global air pollution, drawing insights from datasets sourced from Kaggle and the OpenWeatherMap API. Through diligent data collection, preprocessing, and transformation, we prepared the groundwork for our analysis, ensuring the accuracy and reliability of our findings. Leveraging the capabilities of SQLite for data storage and Tableau for visualization, we navigated through the complexities of the data landscape, uncovering compelling trends and correlations.  
  
Our analysis revealed a notable decrease in air pollution levels over time, reflecting the effectiveness of pollution control measures and environmental policies. This encouraging trend underscores the importance of continued investment in sustainable practices and regulatory frameworks aimed at curbing pollution and protecting public health. As we move forward, our project serves as a testament to the power of data-driven insights in guiding informed decision-making and fostering a healthier, more sustainable future for generations to come.

**APPENDIX A: WHAT I DID & LEARNED**  
  
**Member 1:** Vaagdevi Veeramachaneni

**Contribution to the Project:** I took charge of drafting the initial outline of our report and played a significant role in extracting and cleaning up the data for global air pollution and deaths due to air pollution datasets. My primary focus was ensuring the data's quality and relevance to our analysis. Collaborating closely with my teammates, I helped aggregate and standardize the data for further processing.

**Key Learnings:** Being involved in this project taught me a lot about data extraction techniques and the importance of thorough data cleaning. I gained valuable experience in navigating APIs and handling large datasets efficiently. One challenge I faced was dealing with inconsistencies in data formats from different sources, which underscored the significance of meticulous data preprocessing. Additionally, drafting the report enhanced my communication skills and ability to articulate complex findings.

**Liked/Disliked:** I enjoyed the hands-on experience of working with real-world data and solving practical problems using data engineering principles. While the data cleaning process was time-consuming, I recognized its necessity in ensuring the quality of our analysis.

**Member 2:** Siri Sushma Penumatsa

**Contribution to the Project:** My main contributions to the project involved data analysis, visualization, and drafting sections of the report related to these aspects. I collaborated with team members to identify meaningful insights from our datasets and used Tableau to create interactive visualizations that effectively communicated our findings.

**Key Learnings:** Participating in this project deepened my understanding of data analysis techniques and proficiency in data visualization tools like Tableau. Uncovering correlations between air pollution levels and mortality rates was particularly interesting, highlighting the power of visualization in revealing underlying patterns. Drafting sections of the report also improved my ability to interpret and present complex data clearly and concisely.

**Liked/Disliked:** I found the creative aspect of designing visualizations to be fulfilling, as it made complex information more accessible and actionable. However, selecting the most appropriate visualization techniques could be challenging at times.

**Member 3:** Suguna Sri Nandini Gogusetti

**Contribution to the Project:** My main responsibilities centered around database management and storage. I set up the SQLite database and ensured the smooth storage and retrieval of our transformed data. Additionally, I contributed to drafting sections of the report related to database management and optimization.

**Key Learnings:** This project provided me with hands-on experience in database management and SQL query optimization. Designing and implementing a relational database schema that met our data requirements effectively was a valuable learning experience. Optimizing database performance to handle large volumes of data efficiently posed an interesting challenge, requiring fine-tuning of indexing and query execution strategies. Drafting sections of the report also honed my ability to articulate technical concepts clearly and document implementation details effectively.

**Liked/Disliked:** I enjoyed the technical aspects of database management and found it satisfying to design and implement an efficient database schema. However, troubleshooting database issues, especially with complex queries and performance bottlenecks, could be challenging. Overall, this project provided valuable learning opportunities and enhanced my skills in data management and database optimization.

**APPENDIX B: REFERENCES**

1. OpenWeatherMap. OpenWeatherMap Air Pollution API. <https://openweathermap.org/api/air-pollution>
2. Akshat0giri. (n.d.). Deaths caused by the air pollution dataset. Kaggle. Retrieved from <https://www.kaggle.com/datasets/akshat0giri/death-due-to-air-pollution-199020>
3. Akshat0giri. Deaths Due to Air Pollution (1990-2017). Kaggle. <https://www.kaggle.com/datasets/akshat0giri/death-due-to-air-pollution-19902017>
4. Python Software Foundation. (n.d.). sqlite3 — DB-API 2.0 interface for SQLite databases. Retrieved from <https://docs.python.org/3/library/sqlite3.html>