Analyzing Global Air Pollution by Key Pollutants

Student Name	Student Number	Student Mail ID
Rupam Misra	A00052919	rupam.misra2@mail.dcu.ie
Achal Nanjundamurthy	A00050840	achal.nanjundamurthy2@mail.dcu.ie

Abstract

This study provides a detailed analysis of air pollution across countries, aiming to uncover unique "pollution signatures" of PM2.5, PM10, and NO2. The research addresses how these pollutants influence overall pollution levels in highly affected nations and how their composition varies geographically. We combined three datasets: a 23,463 row global air pollution dataset, a 10,000 row global air quality dataset, and a 32,191 row WHO air quality database. Using these sources, we calculated a composite pollution index in micrograms per cubic meter, defined as the sum of PM2.5, PM10, and NO2, and determined the percentage contribution of each pollutant at the country level.. The analysis shows that South Asian and Middle Eastern countries, face the highest pollution levels, with Pakistan reaching approximately 430 µg/m³. Particulate matter dominates in these regions, with PM2.5 and PM10 together contributing 70% to 90% of total pollution, while NO2 accounts for 10% to 22%. PM10 prevails in dusty and industrial regions, PM2.5 dominates urbanized nations, and NO2 marks areas with high traffic intensity, highlighting the importance of targeting particulate pollutants in environmental and health policies.

1. Datasets

1.1 Datasets Overview and Characteristic

Dataset & Source	Rows & Columns	Key Attributes	Description	Size	Data Types Present
23k Global Air Pollution Dataset (Dataset 1) from Kaggle [1]	23,463 rows and 12 columns	Country, City, AQI Value, AQI Category, CO AQI Value, CO AQI Category, Ozone AQI Value, Ozone AQI Category, NO2 AQI Value, NO2 AQI Category, PM2.5 AQI Value, PM2.5 AQI Category	City-level AQI components and pollutant measurements (µg/m³). Provides CO, Ozone, NO2, PM2.5 values for multiple cities.	10.89 MB	Objects (Country, City, AQI Categories), Int64 (AQI)

10k Global Air Quality Dataset (Dataset 2) from Kaggle [2]	10,000 rows and 12 columns	City, Country, Date, PM2.5, PM10, NO2, SO2, CO, O3, Temperature, Humidity, Wind Speed	Daily city-level concentrations of pollutants and weather variables (PM2.5, PM10, NO2, SO2, CO, O3, plus temperature, humidity, wind speed). CSV file.	2.54 MB	Objects (City, Country, Date), Float64 (pollutants and weather variables)
WHO Air Quality Database 2022 (Dataset 3) from WHO Database [3]	32,191 rows and 15 columns	WHO Region, ISO3, WHO Country Name, City or Locality, Measurement Year, PM2.5 (µg/m³), PM10 (µg/m³), NO2 (µg/m³), PM2.5 temporal coverage (%), PM10 temporal coverage (%)	Long-term mean concentrations of key pollutants by city. Original Excel converted to CSV. Includes temporal coverage and metadata.	14.79 MB	Objects (WHO Region, ISO3, Country Name, City, Reference, type of monitoring stations), Float64/Int64 (pollutant values, coverage, Year, Version, Status)

1.2 Chosen aspect of Big Data - Variety

In this particular project, the main focus of Big Data was on the aspect of data variety, as opposed to data volume or velocity. Although the data had a size of 23,463, 10,000, and 32,191 rows, this was not a problem that required distributed processing or storage capabilities. Moreover, since the data was static, there was no need to stream or update the data.

The most difficult aspect was neither the size of the data nor the velocity of the data. It was the variability of the data from the different sources, and the fact that the data had to be consolidated. This is because each of the different datasets offered different pieces of information.

- **Dataset 1** (23k Global Air Pollution) provided city-level AQI values for PM2.5, NO2, O3, and CO but did **not include PM10**. AQI values for PM2.5 and NO2 were converted to µg/m³ using standard conversion functions.

- **Dataset 2** (10k Global Air Quality) included daily city-level measurements of PM2.5, PM10, NO2, and other pollutants, along with weather data. These values were already in µg/m³ and were used as-is.

- **Dataset 3** (WHO Air Quality Database 2022) provided long-term city-level pollutant averages, temporal coverage information, and metadata. It was used to validate and fill remaining missing values.

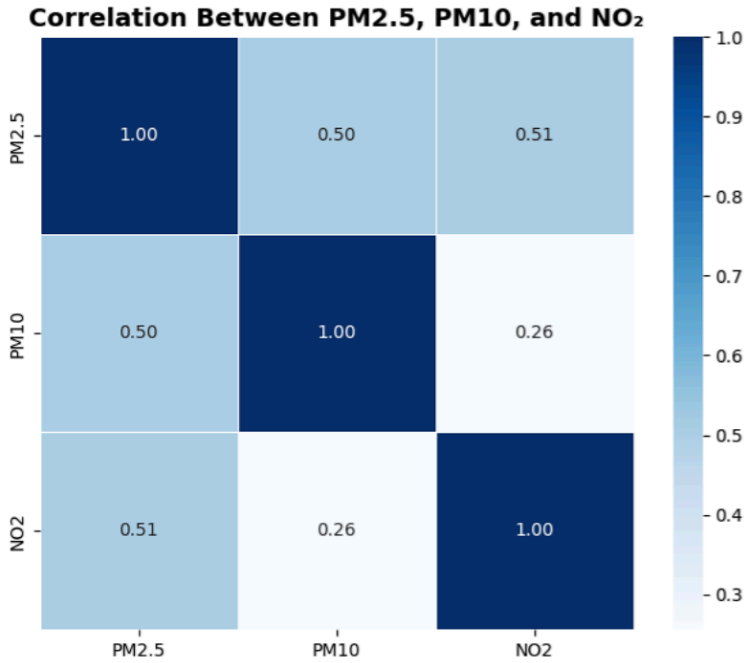
Integrating the three datasets required careful column standardization, handling of missing values, and alignment of units to build a consistent country-level view of pollution. This process reflects the variety aspect of Big Data, as it involves merging sources with different formats, attributes, and structures. The final combined dataset brought together multiple pollutants and locations, enabling robust analysis across more than 65,000 rows while preserving data quality..

2. Data Exploration, Processing, Cleaning, and Integration

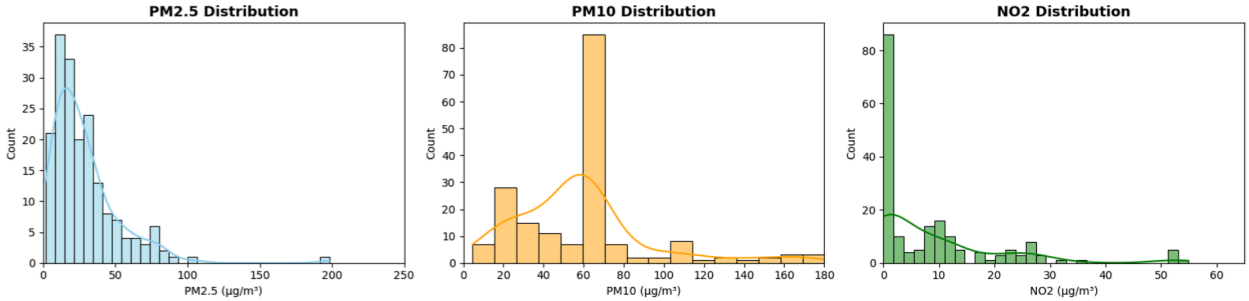
2.1 Data Exploration

The first stage involved reviewing each dataset to understand its size, structure, and completeness. Dataset 1 contained 23,463 city-level records for AQI components including CO, ozone, NO2, and PM2.5, but lacked both data and any method for determining PM10. Dataset 2 provided 10,000 city-level records with concentration data for all pollutants, including PM10. Dataset 3, sourced from the World Health Organization, held 32,191 long-term averages for PM2.5, PM10, and NO2 but was highly incomplete, particularly for the particulate measures. This review involved checking initial rows, assessing data types, computing descriptive statistics, and identifying missing values, which showed that each dataset had different gaps and could not be relied on individually. Variable selection therefore focused on Country, City, PM2.5, PM10, and NO2 because they appeared consistently across all datasets and were essential for cross-country pollution comparisons.

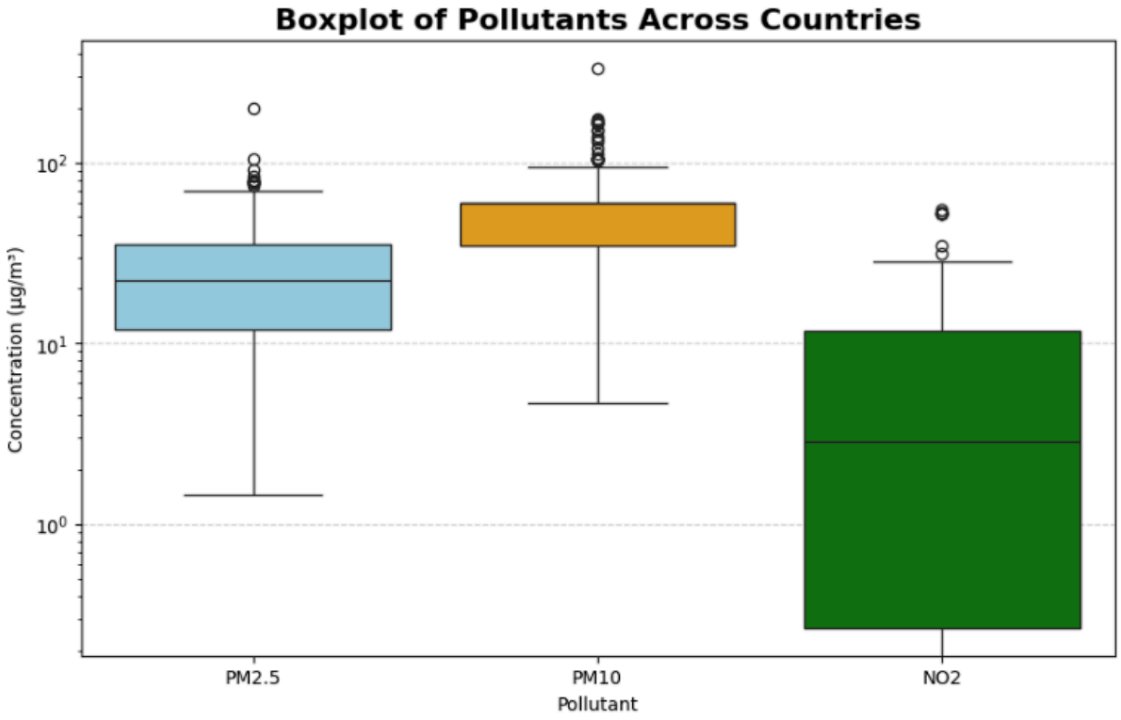
During the exploratory stage, visualisations were used to observe pollutant behaviour across regions before any formal analysis. A correlation heatmap was created to examine relationships among PM2.5, PM10, and NO2, showing that PM2.5 is moderately correlated with both PM10 (0.50) and NO2 (0.51), while PM10 and NO2 have a weaker link (0.26). These early patterns, along with observed variability and outliers, guided later processing and shaped the direction of the full analysis.



We then explored how each pollutant is distributed by plotting **histograms** for PM2.5, PM10, and NO2. The results showed clear right-skewed patterns, with most values at lower levels and a small number of locations showing very high concentrations. This confirmed the presence of extreme pollution events and highlighted the need to account for outliers in later analysis.



We created a boxplot of pollutant levels across countries to examine variability and outliers. Using a log scale highlighted differences, showing PM2.5 and PM10 are right-skewed, while NO2 remains low except in some urban or industrial areas. Though country labels are omitted, the plot clearly illustrates the spread and extremes, complementing the earlier distribution and correlation analyses.



2.2 Processing

Column names were harmonised to standardise schema (e.g., PM2.5 AQI Value to PM2.5, WHO Country Name to Country). Only required attributes were selected:

- **Dataset 1** had Country, City, PM2.5, NO2, CO
- **Dataset 2** had Country, City, PM2.5, PM10, NO2
- **Dataset 3** had Country, City, PM2.5, PM10, NO2

City level pollutants were aggregated at the **country level** using the mean to reduce fluctuations and obtain a representative value per nation. Total concentration per country was computed as **PM2.5 + PM10 + NO2**, to create a composite pollution index. We acknowledge that this calculation involves overlapping mass, but it was maintained to ensure all collected pollutant readings contributed to the overall magnitude ranking and percentage visualization.

2.3 Cleaning

Cleaning addressed missing values and ensured consistency. Datasets 1 and 2 had minimal missing entries, while Dataset 3 contained thousands of missing values for PM2.5, PM10, and NO2. Missing values were filled using **country level mean imputation**, preserving regional characteristics. Irrelevant columns, such as AQI categories and metadata about monitoring stations, were removed to simplify the dataset. Units were standardized, as Dataset 1 used AQI while Datasets 2 and 3 were already in **µg/m³**. The final merged dataset contained **zero missing values**, **185 countries**, and consistent numeric types for all pollutants.

2.4 Integration

Integration resulted in the merging of the three data sets to form a harmonized country data set. Data sets 1 and 2 were merged through the operation of an outer join, then finally data set 3. The columns that resulted from multiple data sets, hence duplicates, had their averages calculated for the resultant pollutant concentration. The concentration of PM10 was derived from data sets 2 and 3, as this information was absent from data set 1. The concentration of total pollutants, as well as the percentage contribution of NO2,PM2.5, and PM10, was obtained.Verification through sample checks confirmed **accurate merging, clean values, and readiness for visualisation**.

2.5 Preparing the datasets for visualization

- Column names were standardized across datasets (e.g., PM2.5 AQI Value to PM2.5).
- Pollutant columns and the Country attribute were selected: **NO2, PM2.5, PM10, Country**.
- City-level data was aggregated to **country level averages** for simplified comparison.
- Missing values were filled using **country level mean imputation** to avoid gaps in visualisation.
- Total pollution per country and percentage contributions for each pollutant were computed for **stacked bar chart representation**.
- Irrelevant metadata and auxiliary columns were removed to focus on key pollutants.

2.6 Attribute selection and data subset

- **Pollutants Chosen:** We focused on **PM2.5, PM10, and NO2** because these pollutants are widely monitored, have significant health impacts [4], and appear consistently across all datasets. **PM2.5** (particles ≤ 2.5 µm) is especially harmful since it can penetrate deep into the lungs and enter the bloodstream, increasing the risk of cancer [5]. **PM10** (particles ≤ 10 µm) affects the upper respiratory system and contributes to haze and reduced visibility, while **NO2**, a gas produced mainly from vehicles and industrial activities, signals urban and industrial pollution intensity. We

did not include **CO**, even though it is toxic, because it is short-lived in the atmosphere, highly variable locally, and less consistently monitored across countries. Choosing PM2.5, PM10, and NO2 lets us capture both particulate and gaseous pollution sources.

- **Aggregation Level:** Data was aggregated at the **country level** to provide a clear, high-level overview of air pollution, rather than city-specific variations. This approach reduces noise from local fluctuations, allows consistent comparison across nations, and is suitable for visualisation in a **single explanatory chart**. Country-level aggregation also aligns with public health reporting standards and global policy discussions.

- **The Composite Pollution Index:** The total pollution concentration for each country was computed as a composite index, specifically the sum of PM2.5, PM10, and NO2. While PM2.5 is scientifically a fraction of PM10, this additive approach was chosen to ensure all three monitored pollutant metrics contributed distinctly to the ranking and visualization. This approach facilitates ranking of countries by overall measured burden but should be interpreted as a composite index of collected data rather than a true measure of total physical airborne mass.

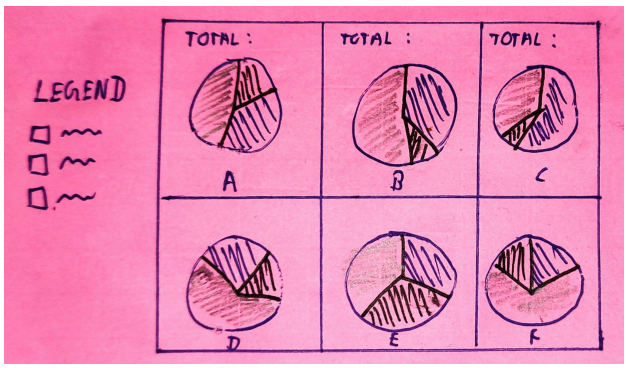
- **Percentage Contribution for top 20 countries:** The chart highlights the 20 most polluted countries and shows the percentage each pollutant contributes to their overall pollution. This makes it easy to compare how polluted each country is and what is driving that pollution, since PM10 and PM2.5 usually dominate while NO_2 forms a smaller share. Together, these patterns reveal clear regional differences in air quality.

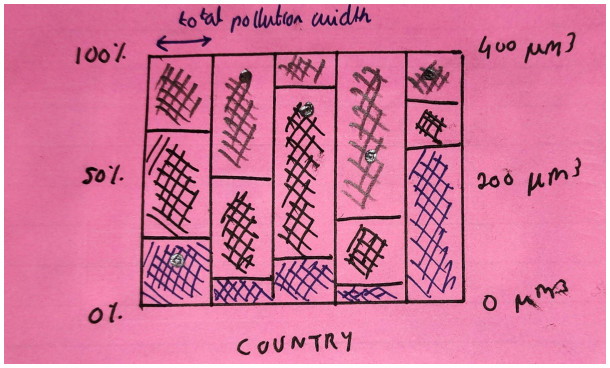
This data selection ensures the final chart is meaningful. Showing both total pollution and pollutant shares, highlights pollution severity and dominant pollutants, offering a view of global air quality trends.

3. Final Visualisation

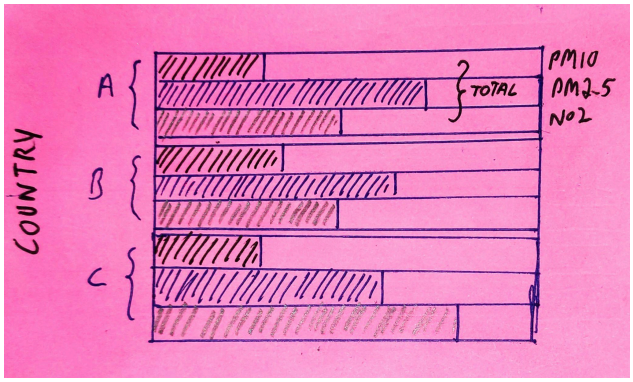
3.1 Initial Sketches

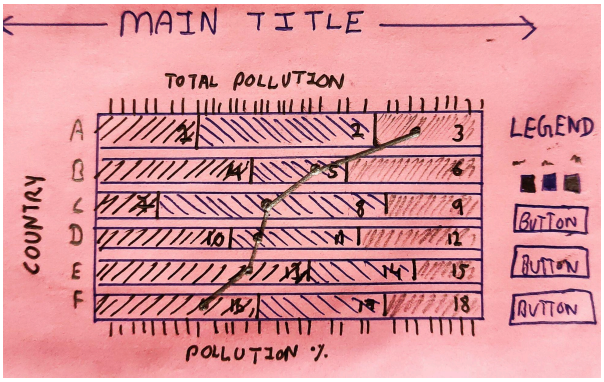
I first considered small pie charts because they seemed visually neat for showing each country's pollutant breakdown, but I rejected them once I realised they make cross-country comparisons and total pollution representation difficult. I then explored a mosaic plot, which initially felt efficient since it combined total pollution and pollutant composition in one view. However, I abandoned this idea as well, because the varying widths and stacked segments make accurate comparisons nearly impossible.





Then, I considered a horizontal grouped bar chart, where each country would have three separate bars side by side for the pollutants. This layout made the graph very long vertically and limited how many countries I could display, while also complicating total pollution calculations. I realised that a **horizontal stacked bar chart looked good**, where each country has a single bar with the pollutants stacked horizontally. This compact design allowed more countries to fit, made adding a total pollution line cleaner, and simplified sorting and comparison.





3.2 Choice of Chart Type

The main goal of our research was to explore how PM2.5, PM10, and NO2 contribute to overall air pollution in the countries most affected. We wanted to present not only the absolute pollution levels in micrograms per cubic meter but also the share of each pollutant to make cross country comparisons clear.

After looking at different visualization options as seen in the sketches above, we found that a horizontal stacked bar chart with a total pollution line on top seemed to work best. In this chart, each bar represents the total pollution for a country, while the segments show how much each pollutant contributes. This design clearly shows both the overall pollution severity and which pollutants are dominant, giving each country a distinct pollution profile.

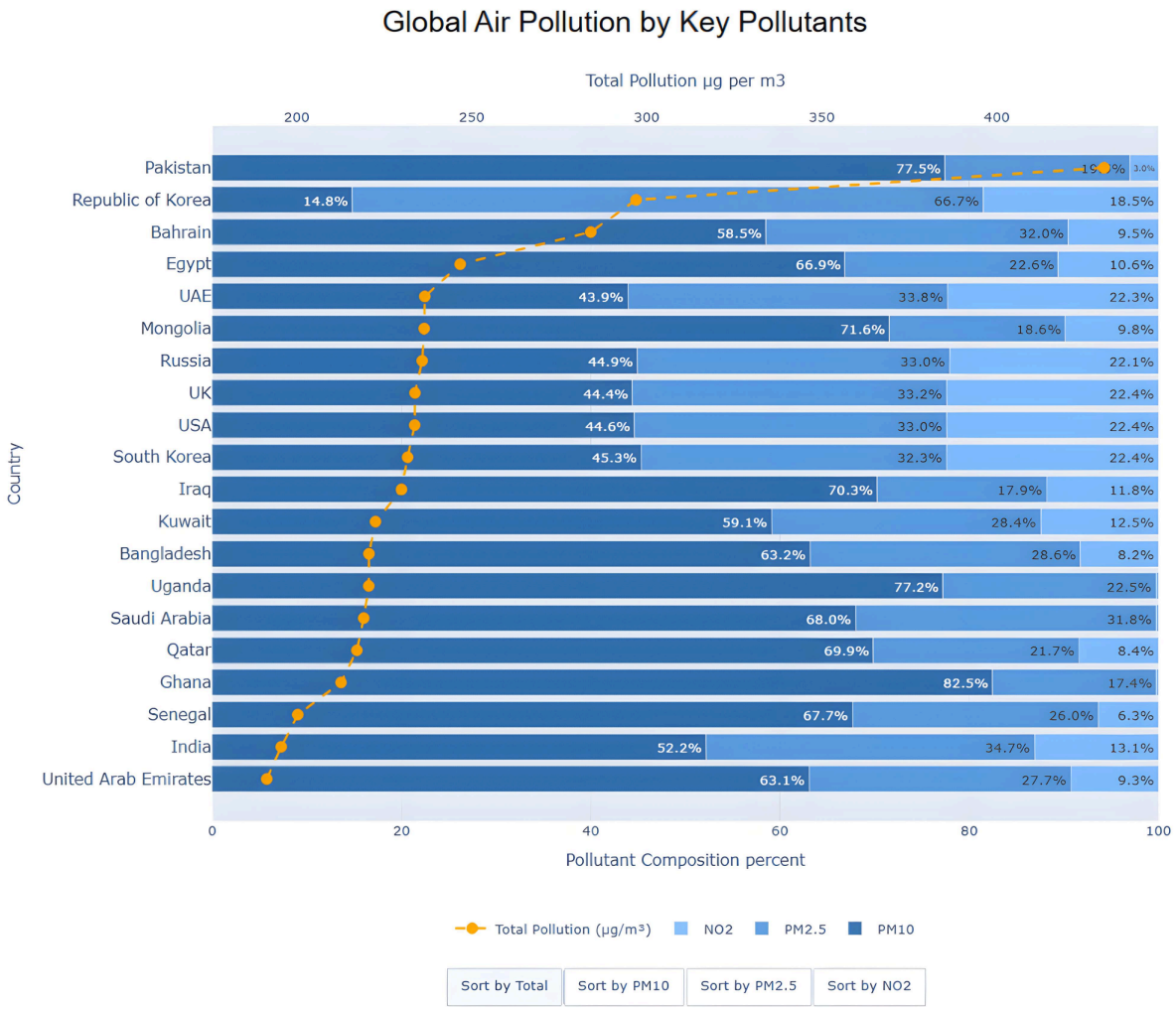
3.3 Chart Type Exploration

During the design process, we explored multiple chart types:

- **Vertical bar charts:** Show total pollution but fail to convey pollutant composition.
- **Grouped bar charts:** Allow comparison of pollutants, but become cluttered with countries.
- **Mosaic Charts:** Varying widths and segments make accurate comparisons nearly impossible.
- **Pie charts:** Only suitable for single countries, impractical for cross-country comparison.
- **Line and scatter plots:** Useful for correlations but not for discrete country level compositions.

After testing different options, **we finally selected a horizontal stacked bar chart** with interactive sorting buttons and an overlaid total line. This design supports full country names on the y-axis, shows relative contributions of each pollutant while retaining the total pollution magnitude, highlights overall severity with a dashed total line, and allows interactive exploration by total or individual pollutants without clutter. The approach balances clarity, and storytelling, making it ideal for comparing both the magnitude and composition of air pollution across countries.

3.4 Our Chart



Final visualisation focuses on the **top 20 countries** by total pollution, emphasizing critical contributors.

- Bars are divided into **PM10, PM2.5, and NO2**, with **percentages labeled only if >1%** to reduce clutter.
- **Total pollution** is highlighted using **orange markers with a dashed line** on a secondary x-axis for prominence without overcrowding the chart.

The chart uses a monochromatic blue palette for particulate matter, with PM10 in dark blue, PM2.5 in medium blue, and a lighter shade for NO2 to convey environmental context while minimizing visual strain. Total pollution is emphasized with orange markers and a dashed line on a secondary axis. Horizontal bars allow long country names and easy segment comparisons, while focusing on the top 20 countries highlights the most critical information without overwhelming viewers. Interactive features, such as hovering to reveal exact values and buttons to sort by total or individual pollutants, enable dynamic exploration. This visualisation **effectively combines clarity, detail, and interactivity**, helping viewers quickly identify the most polluted countries and dominant pollutants, while allowing deeper insights through interactive exploration.

4. Conclusion

4.1 Tools and libraries used

Jupyter Notebook was used to filter, preprocess, and clean the datasets, as well as to perform calculations, aggregations, and generate intermediate visualizations. Libraries such as **pandas, numpy, matplotlib, seaborn, and plotly** were used for data manipulation, plotting, and interactive charting. Since my teammate and I live nearby, all work was done together in person, so no collaborative platforms like Google Colab were needed. The final visualisation charts were prepared directly in the Jupyter Notebook.

4.2 What the Data Reveals: Linking our results to Real-World Trends

In this project, we analyzed global air pollution to identify "pollution signatures" of PM10, PM2.5, and NO2 and their impact on total pollution across countries. We combined a 23,463-row global air pollution dataset, a 10,000-row global air quality dataset, and a 32,191-row WHO air quality database to calculate a total pollution index (μg/m³) and each pollutant's share by country. South Asian and Middle Eastern nations, including Pakistan, Bangladesh, Egypt, Iraq, Kuwait, and Saudi Arabia, showed the highest pollution, with Pakistan reaching about 430 μg/m³. Particulate matter dominates in these regions, contributing 70 to 90% of total pollution, while NO2 accounts for 10 to 22%. PM10 is the main pollutant in dusty or industrial countries such as Pakistan, Uganda, Mongolia, Egypt, Iraq, Ghana, and Qatar, driven by dust storms, desertification, construction, industrial activity, and weaker environmental regulations [6].

PM2.5 dominates urbanized and industrialized nations like the Republic of Korea and Russia, primarily due to combustion, traffic, power plants, and regional transport. NO2 is most significant in traffic-heavy countries, contributing around 22%, and shows a weaker correlation with PM10 (0.26), while PM2.5 is moderately correlated with PM10 (0.50) and NO2 (0.51). The countries with the highest total pollution are Pakistan (430.6 μg/m³), Republic of Korea (296.9), Bahrain (284), Egypt (246.7), and UAE (236.6), while China does not appear in the top 20 because city-level averaging reduces national totals [7]. These patterns indicate that PM10 dominates dusty or industrial regions, PM2.5 prevails in urbanized areas [8], and NO2 marks traffic-dense locations. Cleaning, aggregating, and visualizing these datasets provides actionable insights for policymakers, linking pollution to geography, urbanization, and industrial activity

4.3 Collaboration

The work was evenly divided between us. Rupam handled filtering and preprocessing of the datasets, while Achal focused on data cleaning and integration. Both of us collaborated closely on every decision regarding chart design, including layout, color scheme, and visual presentation, ensuring that the final visualisation was clear, accurate, and easy to interpret. We contributed equally throughout the project, from data preparation to the creation of the final results.

References

[1] Hasibil Muzdadid. (2023). *Global Air Pollution Dataset*. Kaggle. Available at: https://www.kaggle.com/datasets/hasibalmuzdadid/global-air-pollution-dataset

[2] WAQI. (2023). *Global Air Quality Dataset*. Kaggle. Available at: www.kaggle.com/datasets/waqi786/global-air-quality-dataset

[3] World Health Organization. (2022). *WHO Air Quality Database*. Available at: https://www.who.int/data/gho/data/themes/air-pollution/who-air-quality-database/2022

[4] Orellano et al. (2020). Short-term exposure to particulate matter (PM10 and PM2.5), nitrogen dioxide (NO2), and ozone (O3) and all-cause and cause-specific mortality: Systematic review and meta-analysis. *Environment International, 142*, 105876. doi: 10.1016/j.envint.2020.105876. Available at: https://pubmed.ncbi.nlm.nih.gov/32590284/

[5] European Environment Agency (EEA). (2025). *How air pollution affects our health*. Available at: https://www.eea.europa.eu/en/topics/in-depth/air-pollution/eow-it-affects-our-health

[6] Hussein et al. (2020). *Particulate Matter Concentrations in a Middle Eastern City – An Insight to Sand and Dust Storm Episodes.* Aerosol and Air Quality Research, 20(12), 2780-2792. Available at: https://aaqr.org/articles/aaqr-20-05-oa-0195

[7] Yang, L et al. (2023). *Quantifying the Spatiotemporal Heterogeneity of PM2.5 Pollution and Its Determinants in 273 Cities in China*. International Journal of Environmental Research and Public Health, 20(2), 1183. Available at: https://doi.org/10.3390/ijerph20021183

[8] Proestakis, E. et al. (2025). *Atmospheric dust and air quality over large-cities and megacities of the world.* Atmospheric Chemistry and Physics, 25(21), 14777–14823. https://doi.org/10.5194/acp-25-14777-2025 acp.copernicus.org+1