SINGAPORE INSTITUTE OF TECHNOLOGY

The Exploration of PM2.5 changes over the years???

Ng Wei Herng, Timothy Zoe Delaya, Clarence Agcanas, Yeo Song Chen, Lee Ru Yuan (Group 6)

I. Original Data Visualization in News Media

The visualization titled “Comparing PM2.5 Concentrations in Capital Cities” created by Pallavi Rao (2023) and published on “The Visual Capitalist”, presents a snapshot of PM2.5 air pollution levels in various capital cities around the world for the year 2022. PM2.5 refers to particulate matter that is less than 2.5 micrometers in diameter, which is small enough to penetrate the lungs and enter the bloodstream, posing significant health risks.

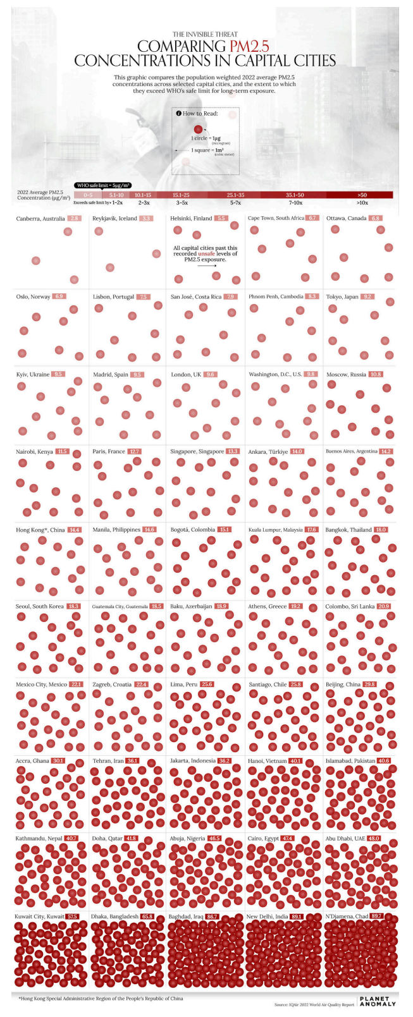

Figure 1: Visualized: Air Quality and Pollution in 50 Capital Cities (IQAir 2022 World Air Quality Report)

This visualization uses a series of red circles to represent the PM2.5 concentrations in each capital city. The number of circles corresponds to the level of PM2.5 concentration that exceeds the World Health Organization’s (WHO) safe limit for PM2.5, which is 5 µg/m³. Any value above this indicates a higher risk for adverse health effects.

i. Issues with the Original Visualization

While the original visualization provides a clear depiction of air pollution in popular cities, it does not show additional factors that might contribute to the pollution levels, such as population size and GDP. These factors could offer insights into the underlying causes of the pollution levels observed in the visualization. The original visualization is not easy to grasp at a glance due to the use of multiple red circles, which can be overwhelming and confusing for viewers. This complexity makes it difficult to quickly understand the relative pollution levels between different cities.

II. Dataset

The original data set used for the visualization was sourced from the IQAir World Air Quality Report (2022). While IQAir provides an API for users to obtain their air quality data, the API only gives real-time data instead of data in a time series ranging from the past to recent years.
As such, the data set we have chosen to use comes from the World Health Organization (WHO) which provides data on air quality for various countries from a wider year range. The data set contains information on PM2.5 concentrations for different countries and years and is also more precise as WHO has a 60% inclusion requirement whereby the recorded data require annual data availability of at least 60% of the total number of hours in a year to be included. Alongside this data set, we have also chosen to use 3 additional data set for the purposes of enhancing the visualization, as well as to improve on the data engineering and data cleaning aspect of the WHO data set.

1. Country Codes (2024)
2. Population Data (2024)
3. GDP Data (2024)

i. Data cleaning and preparation

We began by reading in the main dataset, which contains PM2.5 concentration data. We filtered this dataset to include only the years from 2017 to 2019 to focus on the most recent data.

For the GDP data, we read in the dataset and skipped the first four rows, which likely contain metadata. We transformed the data from wide to long format, converting year columns into key-value pairs, and renamed columns for clarity. Missing GDP values were imputed within each country group using previous years’ data to ensure completeness. Unnecessary columns (indicator_name and indicator_code) were dropped, and the data was filtered to include only the years 2017 to 2019. The cleaned data was then saved to a new CSV file.
For the population data, we read in the cleaned population dataset and imputed missing population values within each country group using previous years’ data. The data was filtered to include only the years 2017 to 2019, and the cleaned data was saved to a new CSV file.
Finally, we merged the GDP and population data with the main PM2.5 dataset on the matching country names and years. Any remaining rows with null values were dropped to ensure our final dataset was complete. The final merged dataset was saved to a new CSV file for further analysis.

III. Improved Visualisation

To improve upon the original visualization, decided that we would create three visualizations to provide a more comprehensive and insightful analysis of PM2.5 concentrations globally. These visualizations include:

- **Improved Visual Appeal**: All of our visualizations are designed to be visually appealing and easy to understand, making it easier for the audience to interpret the data.
- **Additional Data**: We incorporated population and GDP data to provide context and insights into the factors contributing to air pollution levels in different regions. This additional information enhances the audience’s understanding of the underlying causes of pollution.
- **Interactive Elements**: All of our visualizations are interactive, allowing users to explore the data further by hovering over data points or selecting specific countries to view detailed information.

i. Choropleth Map

The choropleth map displays PM2.5 concentrations in countries globally, with darker colors representing higher PM2.5 levels. This visualization allows for quick and clear comparison of air quality across countries. Unfortunately, due to the lack of data of air pollution, most countries are greyed out. However, the countries with data are shown in the map below.

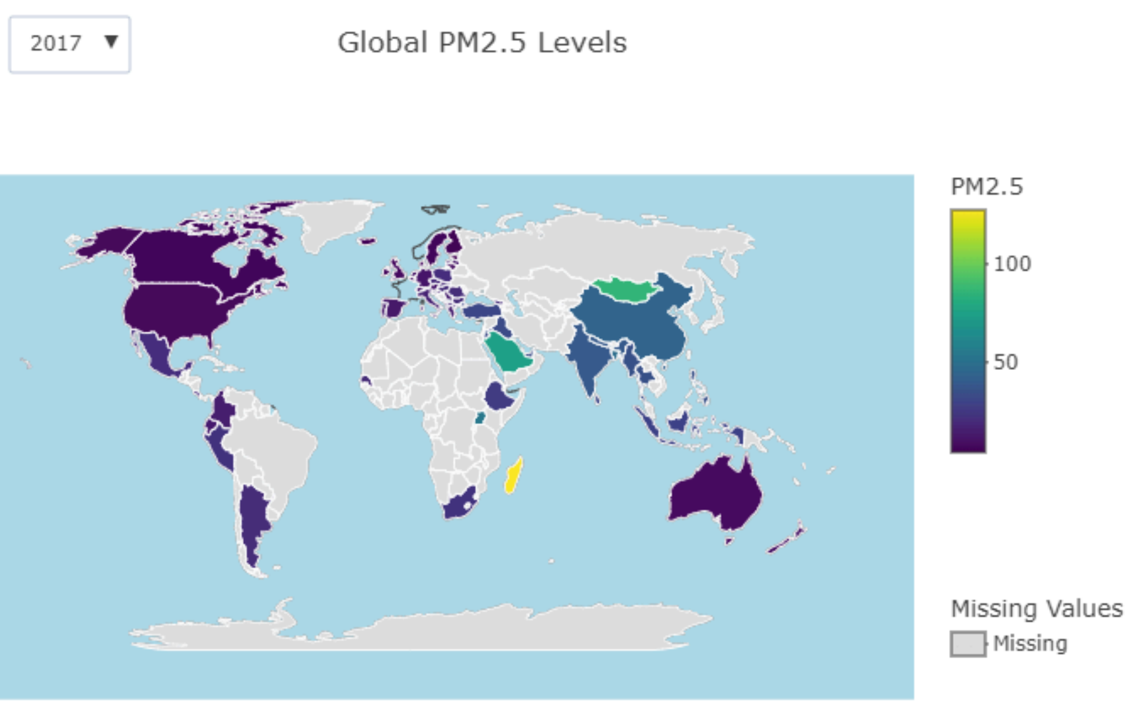

Figure 2: Choropleth Map showing the Global PM2.5 Levels

ii. Scatter Plot, PM2.5 vs GDP

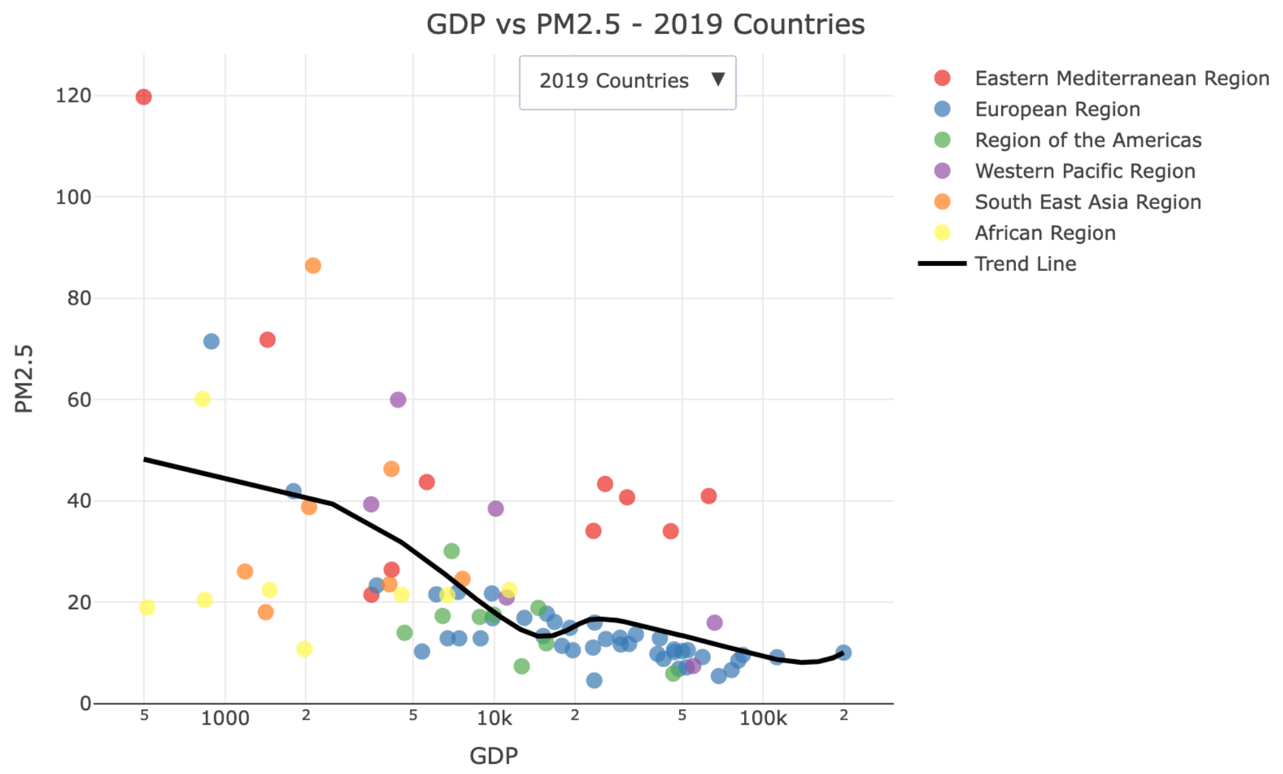

Figure 3: Global PM.2 and GDP

- Averaging of Cities to Get Country Air Pollution

To derive country-level air pollution data from city-level data, we aggregate the city-level PM2.5 concentrations by taking the average for each country. This process involves grouping the data by country and year, and then computing the mean PM2.5 concentration for each group. By doing this, we obtain a representative value of PM2.5 concentration for each country, which can then be used for further analysis.

- Analysis using Trend Line

The trend line provides a visual representation of the general pattern or relationship between GDP and PM2.5 concentrations. This can help to identify whether higher GDP is associated with higher or lower levels of air pollution. We use LOESS Method because it does not assume a specific functional form (like linear regression) and can adapt to the underlying data structure, providing a more flexible and accurate representation of the relationship.