DA 2015 - Statistics Laboratory II - Group Project – Report

Group Details
Group number: 05

- Manoj Gamage - 24ada002
- Gayan Ranasingha - 24ada066
- Chaminda Kaluarachchi - 24ada067

Title of the Project:

Impact of Air Pollutions on Human Respiratory Diseases

---

1. Introduction

The dataset you provided is a compilation of health burdens attributable to Air Pollution from the State of Global Air (SOGA). This data is part of a comprehensive effort to quantify the health impact of air pollution across different countries and over time, typically using the methodology of the Global Burden of Disease (GBD) study.

1.1. Key Data Characteristics

- Time Period: The data spreads from 1990 to 2023.
- Geographical Scope: It covers a total of 169 unique countries.
- Exposure/Risk Factor: The data is specifically focused on the health burden attributed to Air Pollution and PM 2.5.
- Cause of Death: The health outcomes considered are for all causes of death related to air pollution exposure and PM 2.5.
- Demographics: The estimates are aggregated for both sexes and for all ages.
- Core Measures: The dataset includes two main measures of health burden, with their associated metrics:
  1. Death (Measure): The Number (Metric) of premature deaths attributable to air pollution.
  2. DALY (Disability-Adjusted Life Years) (Measure): The Number (Metric) of DALYs attributable to air pollution. DALYs represent the sum of years of life lost due to premature mortality and years lived with disability.

1.2. Key Variables

Column Name	Description	Data Type
Country	The specific country for the measurement.	Categorical
Year	The year the measurement was recorded.	Numerical
Measure	The type of health burden: 'death' or 'daly'.	Categorical

Burden Mean	The central estimate (mean) of the health burden (number of deaths or DALYs).	Numerical
Burden Upper	The upper bound of the 95% uncertainty interval for the burden estimate.	Numerical
Burden Lower	The lower bound of the 95% uncertainty interval for the burden estimate.	Numerical
REI Name	The risk exposure/environmental factor, which is consistently 'Air pollution'.	Categorical

2. Descriptive and Exploratory Analysis

2.1. Summary Statistics:

Calculate the mean, median, standard deviation, minimum, and maximum for Burden Mean (for both 'death' and 'DALY') and 25th – 50th – 75th quantiles to understand the overall magnitude and variability of the burden across all countries and years.

```
# 01. Summary Statistics

print(" Summary Statistics of Burden Mean (by Measure - Death and DALY) ")
burden_stats = df.groupby('Measure')['Burden Mean'].describe().transpose()
print(burden_stats)
```

```
...  Summary Statistics of Burden Mean (by Measure - Death and DALY)
    Measure          daly         death
    count    8.112000e+03  1.140200e+04
    mean     7.774689e+05  3.059001e+04
    std      3.802926e+06  1.641227e+05
    min      4.000000e+00  1.000000e+00
    25%      1.720000e+04  5.582500e+02
    50%      7.930000e+04  3.100000e+03
    75%      3.587250e+05  1.400000e+04
    max      6.301000e+07  2.333000e+06
```

2.2. Time-Series Trends:

Analyze the trend of the air pollution and PM 2.5 burden over time (1990 to 2023) by aggregating the Burden Mean across all countries for both Deaths and DALYs

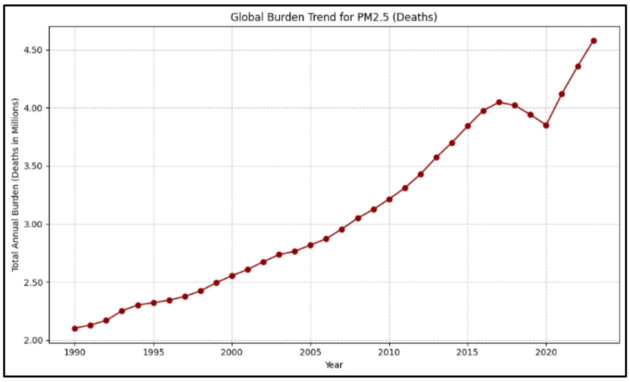
```python
# List of burden columns to clean (as previously identified)
burden_cols_with_spaces = ['  Burden Mean  ', '  Burden Mean Rounded  ', '  Burden Upper  ', '  Burden Lower  ']
cleaned_burden_cols = ['Burden Mean', 'Burden Mean Rounded', 'Burden Upper', 'Burden Lower']

# Clean and convert columns to numeric
for col in burden_cols_with_spaces:
    new_col_name = col.strip()
    df.rename(columns={col: new_col_name}, inplace=True)

    # Remove commas and use errors='coerce' to turn non-numeric values into NaN
    df[new_col_name] = df[new_col_name].astype(str).str.replace(',', '', regex=False)
    df[new_col_name] = pd.to_numeric(df[new_col_name], errors='coerce')

# Remove rows with NaN in the specified columns (which were none in the previous check, but good practice)
df_clean = df.dropna(subset=cleaned_burden_cols)

# 2. Calculate Global Trends
global_trend = df_clean.groupby(['Year', 'Pollutant Name', 'Measure'])['Burden Mean'].sum()

# 3. Define Plotting Parameters
pollutants = ['PM2.5', 'Air Pollution']
measures = ['death', 'daly']
titles = {
    'death': 'Deaths',
    'daly': 'DALYs (Disability-Adjusted Life Years)'
}
colors = {
    'PM2.5': 'darkred',
    'Air Pollution': 'darkblue'
}
```

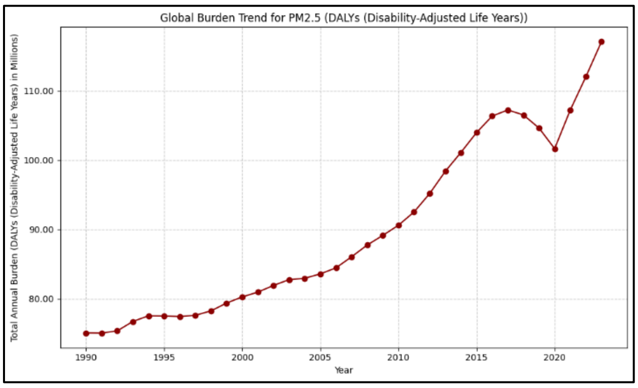
```python
# 4. Generate Plots ---
plot_filenames = []
summary_data = {}

def plot_trend(pollutant, measure):
    """Filters data and plots the trend for a specific pollutant and measure."""

    # Filter the series
    series = global_trend.loc[:, pollutant, measure]

    # Check if data exists for the combination
    if series.empty:
        print(f"No data found for {pollutant} - {measure}.")
        return

    # Store first and last year data for summary
    summary_data[f'{pollutant} - {measure}'] = {
        '1990': series.iloc[0],
        '2023': series.iloc[-1]
    }

    plt.figure(figsize=(10, 6))

    # Plot the series
    plt.plot(series.index, series.values,
             label=f'Global {titles[measure]} Burden',
             color=colors[pollutant],
             marker='o',
             linestyle='-')

    # Formatting
    title = f'Global Burden Trend for {pollutant} ({titles[measure]})'
    plt.title(title)
    plt.xlabel('Year')
    plt.ylabel(f'Total Annual Burden ({titles[measure]} in Millions)')

    # Convert y-axis labels to millions
    formatter = plt.FuncFormatter(lambda x, pos: f'{x/1e6:.2f}')
    plt.gca().yaxis.set_major_formatter(formatter)

    plt.grid(True, linestyle='--', alpha=0.7)
    plt.tight_layout()

    # Save file
    filename = f'global_burden_trend_{pollutant.replace(" ", "_")}_{measure}.png'
    plt.savefig(filename)
    plot_filenames.append(filename)

# Execute plotting for all four combinations
for pollutant in pollutants:
    for measure in measures:
        plot_trend(pollutant, measure)

print("Generated Plots:")
for filename in plot_filenames:
    print(filename)

print("\nSummary of Trends (Burden Mean in Millions):")
for key, values in summary_data.items():
    print(f"--- {key} ---")
    # Convert to millions for display
    start_burden = values['1990'] / 1e6 if not pd.isna(values['1990']) else 'N/A'
    end_burden = values['2023'] / 1e6 if not pd.isna(values['2023']) else 'N/A'

    print(f"1990 Burden: {start_burden:.2f}M")
    print(f"2023 Burden: {end_burden:.2f}M")
```

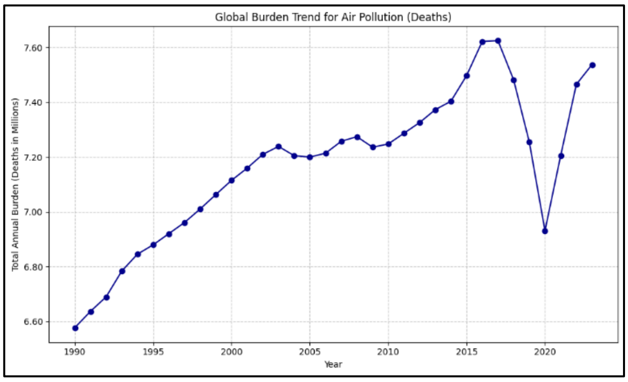
Generated Plots:

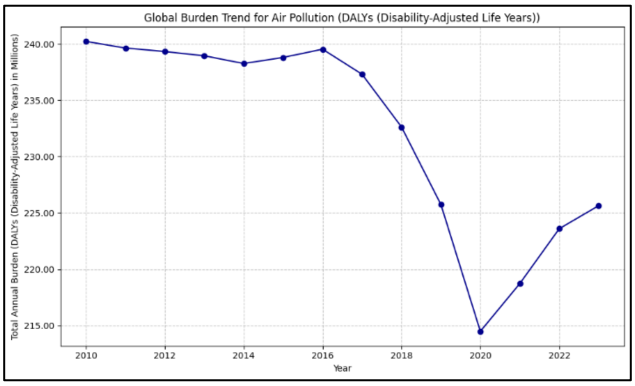
- global_burden_trend_PM2.5_death
- global_burden_trend_PM2.5_daly
- global_burden_trend_Air_Pollution_death
- global_burden_trend_Air_Pollution_daly

```
Generated Plots:
global_burden_trend_PM2.5_death.png
global_burden_trend_PM2.5_daly.png
global_burden_trend_Air_Pollution_death.png
global_burden_trend_Air_Pollution_daly.png

Summary of Trends (Burden Mean in Millions):
--- PM2.5 - death ---
1990 Burden: 2.10M
2023 Burden: 4.58M
--- PM2.5 - daly ---
1990 Burden: 75.05M
2023 Burden: 117.12M
--- Air Pollution - death ---
1990 Burden: 6.58M
2023 Burden: 7.54M
--- Air Pollution - daly ---
1990 Burden: 240.22M
2023 Burden: 225.64M
```









2.3.Geographical Distribution:

Identify the top 10 countries with the highest and lowest average burdens (death and DALYs) over the specific time period. Herewith we have checked the 2023 data.

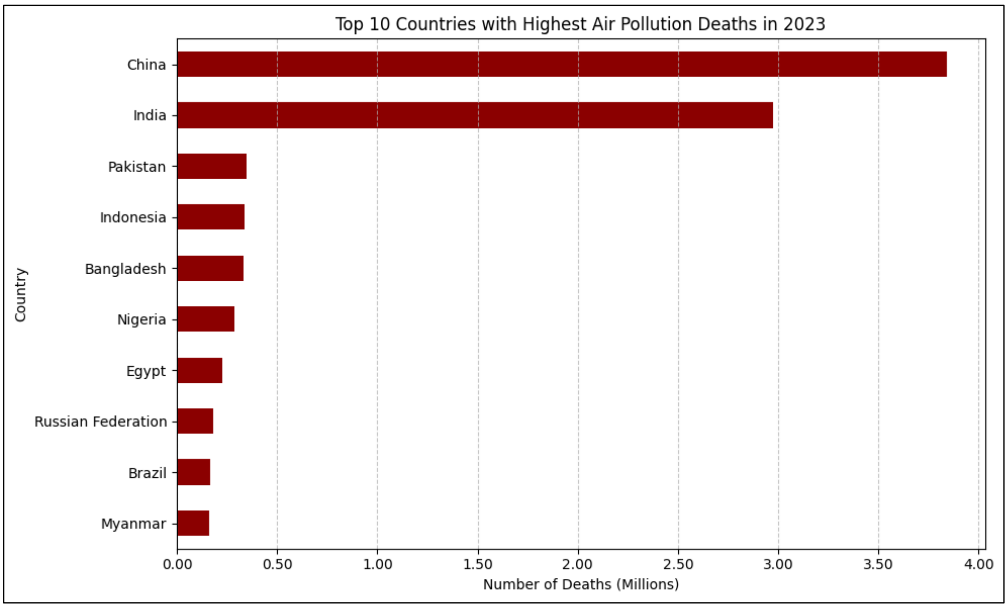
```python
#03. Geographical Distribution Analysis (Year 2023)
df_2023 = df[df['Year'] == 2023]
country_burden_2023 = df_2023.groupby(['Country', 'Measure'])['Burden Mean'].sum().unstack()
print(country_burden_2023.head())
```

```python
# Top 10 Countries by Death Burden in 2023
top_10_deaths = country_burden_2023.sort_values(by='death', ascending=False).head(10)['death']
print("\n_Top 10 Countries by Air Pollution Deaths in 2023_")
print(top_10_deaths)

# Top 10 Countries by DALY Burden in 2023
top_10_DALYs = country_burden_2023.sort_values(by='daly', ascending=False).head(10)['daly']
print("\n_Top 10 Countries by Air Pollution DALYs in 2023_")
print(top_10_DALYs)
```

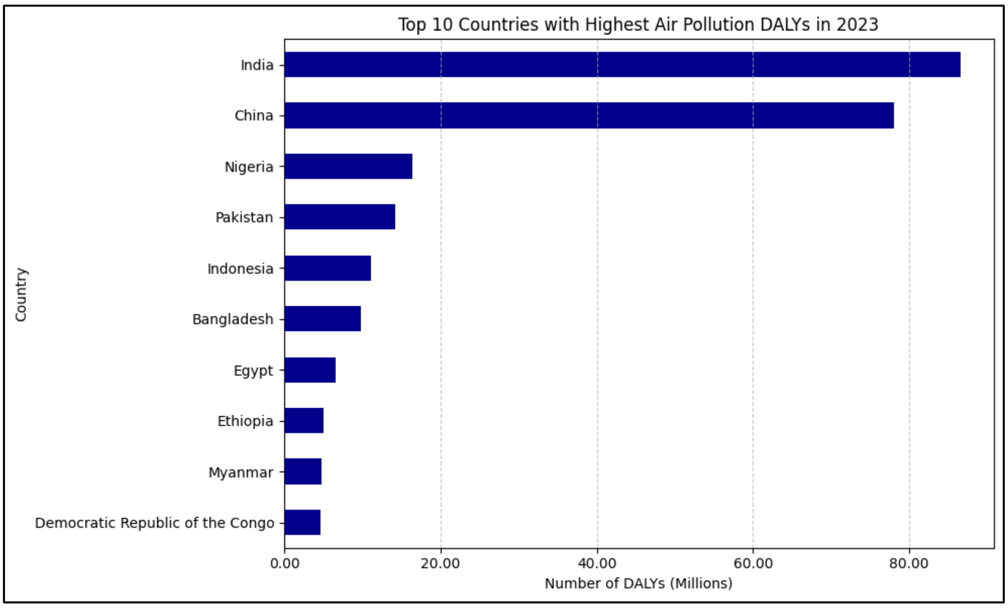
```
_Top 10 Countries by Air Pollution Deaths in 2023_
Country
China                 3841000.0
India                 2975700.0
Pakistan               347000.0
Indonesia              339200.0
Bangladesh             334600.0
Nigeria                287400.0
Egypt                  224400.0
Russian Federation     181400.0
Brazil                 165400.0
Myanmar                162400.0
Name: death, dtype: float64

_Top 10 Countries by Air Pollution DALYs in 2023_
Country
India                                86530000.0
China                                78060000.0
Nigeria                              16354000.0
Pakistan                             14271000.0
Indonesia                            11098000.0
Bangladesh                            9866000.0
Egypt                                 6621000.0
Ethiopia                              5037200.0
Myanmar                               4816000.0
Democratic Republic of the Congo      4649100.0
Name: daly, dtype: float64
```

Plot the total burden for a specific recent year (e.g: 2023) across the group of countries which were selected as ‘Top 10 Countries by DALY Burden in 2023’ and ‘Top 10 Countries by deaths Burden in 2023’ for comparison (e.g., a bar chart). Same analysis can be done for the PM2.5 also.

```python
# Plotting the Top 10 Countries (Deaths)

plt.figure(figsize=(10, 6))
top_10_deaths.sort_values(ascending=True).plot(kind='barh', color='darkred')
plt.title('Top 10 Countries with Highest Air Pollution Deaths in 2023')
plt.xlabel('Number of Deaths (Millions)')
plt.ylabel('Country')
formatter = plt.FuncFormatter(lambda x, pos: f'{x/1e6:.2f}')
plt.gca().xaxis.set_major_formatter(formatter)
plt.grid(axis='x', linestyle='--', alpha=0.7)
plt.tight_layout()
plt.savefig('top_10_deaths_2023.png')
```




```python
# Plotting the Top 10 Countries (DALYs)

plt.figure(figsize=(10, 6))
top_10_dalys.sort_values(ascending=True).plot(kind='barh', color='darkblue')
plt.title('Top 10 Countries with Highest Air Pollution DALYs in 2023')
plt.xlabel('Number of DALYs (Millions)')
plt.ylabel('Country')
formatter = plt.FuncFormatter(lambda x, pos: f'{x/1e6:.2f}')
plt.gca().xaxis.set_major_formatter(formatter)
plt.grid(axis='x', linestyle='--', alpha=0.7)
plt.tight_layout()
plt.savefig('top_10_dalys_2023.png')
```



3. Comparative Analysis

3.1. DALY-to-Death Ratio:

Calculate, analyze, and interpret the ratio of DALYs to Deaths for all countries and years. This highlights where air pollution or PM2.5 is contributing to more premature mortality and/or disability. Herewith we have done the analysis for ‘Air Pollution’ from 2010 to 2023.

The DALY-to-Death Ratio is calculated as:

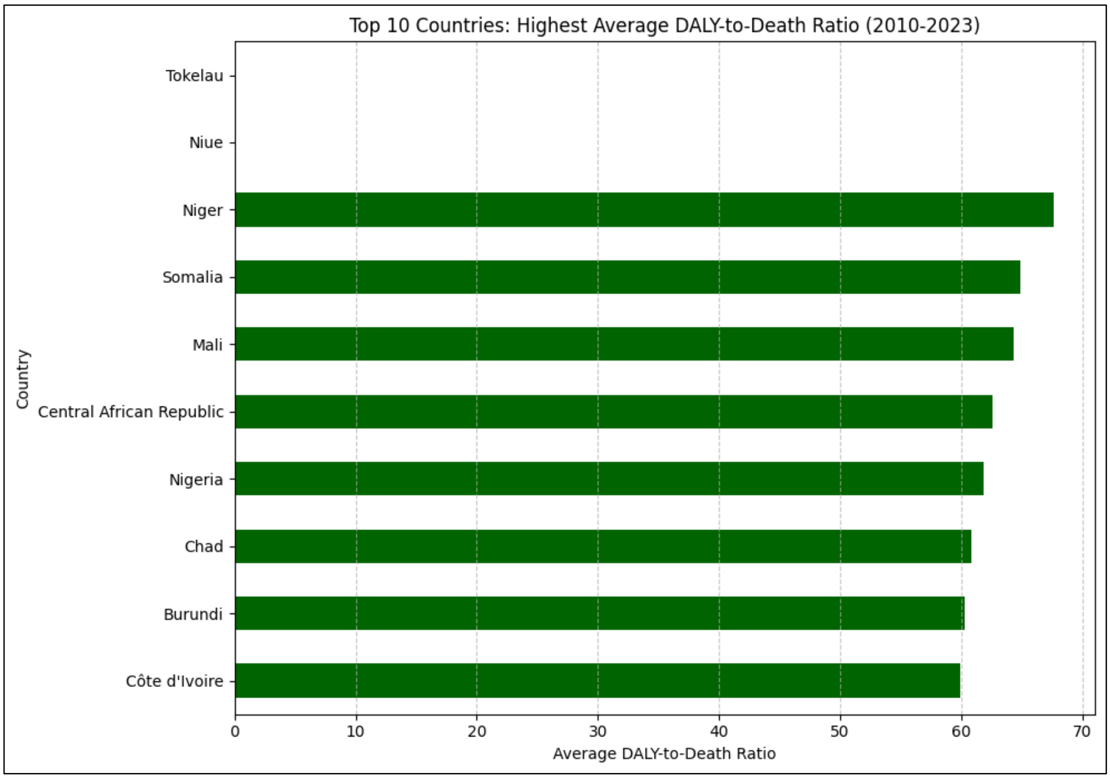
Ratio = (Burden DALYs / Burden Deaths)

Interpretation:

- This ratio shows the average number of healthy years lost for each death caused by air pollution or PM2.5.
- A high ratio means the disease is affecting people at younger ages or causing long-term illness. This usually happens in poorer areas where many people are exposed to air pollution.
- A low ratio means most of the health impact is happening among older people.

3.2.Top 10 countries with highest DALY-to-Death Ratio (Air Pollution)

```
Top 10 Countries with the Highest Average DALY-to-Death Ratio (Air Pollution, 2010-2023):
Country
Tokelau                          inf
Niue                             inf
Niger                      67.646374
Somalia                    64.852985
Mali                       64.328860
Central African Republic   62.583337
Nigeria                    61.814936
Chad                       60.859722
Burundi                    60.240356
Côte d'Ivoire              59.937167
Name: DALY_to_Death_Ratio, dtype: float64

Top 10 Countries with the Lowest Average DALY-to-Death Ratio (Air Pollution, 2010-2023):
Country
Hungary      20.021734
Ukraine      19.845494
Romania      19.823569
Slovenia     19.708175
Serbia       19.630644
Estonia      19.416699
Czechia      19.225106
Croatia      19.059623
Latvia       18.785038
Lithuania    18.456782
Name: DALY_to_Death_Ratio, dtype: float64
```



Key Insights:

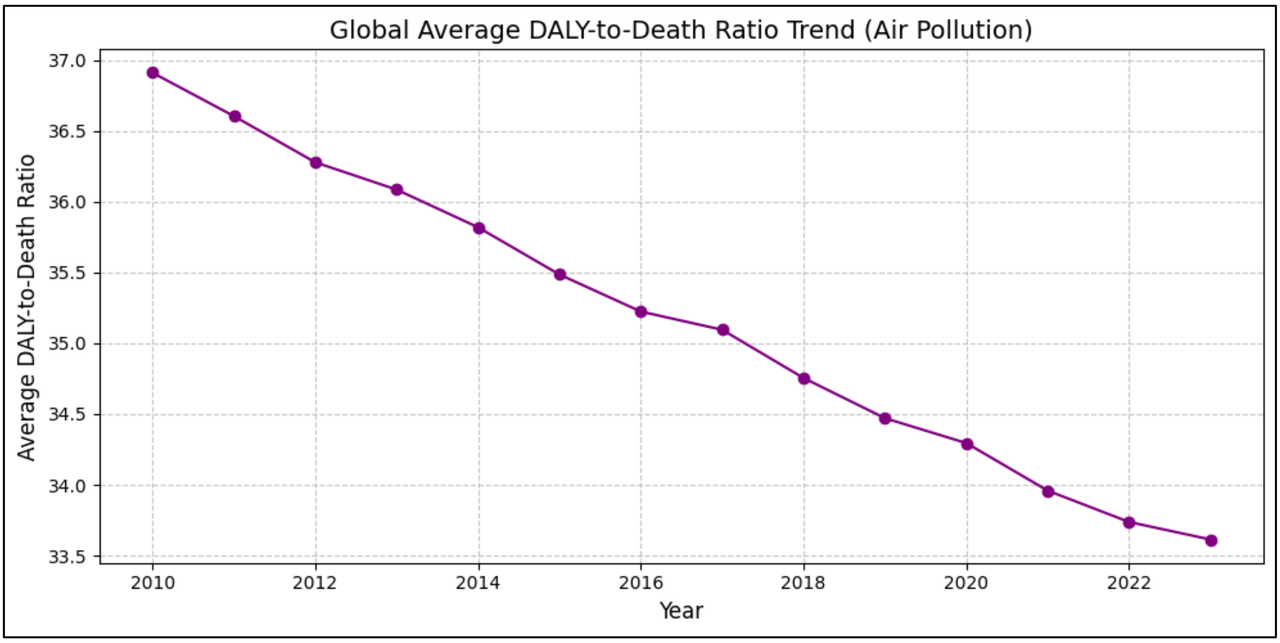
The top 10 countries are all in Sub-Saharan Africa. This is an important finding because it shows that air pollution is causing serious health problems and early deaths, especially among young people in these countries.

3.3. Global Average DALY-to-Death Ratio Trend (Air Pollution)

The chart shows a clear pattern:

- Trend: From 2010 to 2023, the global average DALY-to-Death Ratio has steadily gone down.
- Decline: It dropped from about 36.9 in 2010 to around 33.6 in 2023.

This steady decrease suggests that, over time, air pollution is affecting slightly older age groups more, or that long-term illness (YLD) is making up a smaller share of the total health impact compared to years of life lost (YLL).

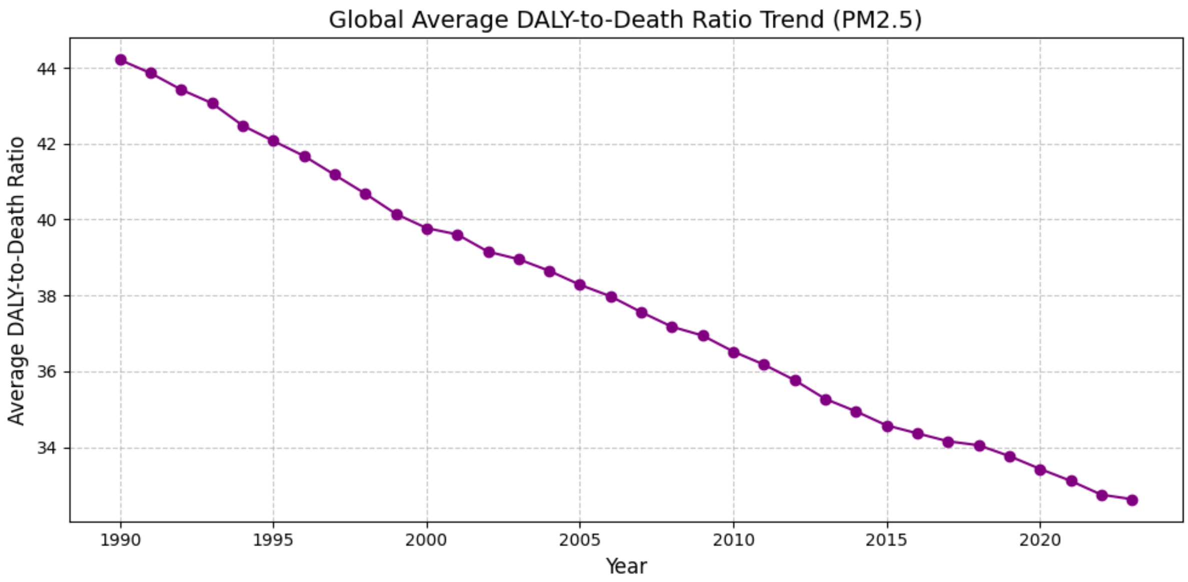



```
Global Average DALY-to-Death Ratio Trend Data:
Year
2010    36.910833
2011    36.604182
2012    36.277974
2013    36.084933
2014    35.818948
2015    35.485434
2016    35.223867
2017    35.095137
2018    34.754411
2019    34.470628
2020    34.294890
2021    33.958823
2022    33.737349
2023    33.613834
```

3.4. Global Average DALY-to-Death Ratio Trend (PM2.5)

The chart shows a clear pattern:

- Trend: From 1990 to 2023, the global average DALY-to-Death Ratio has steadily gone down.
- Decline: It dropped from about 44.21 in 1990 to around 32.63 in 2023.




```
Global Average DALY-to-Death Ratio Trend Data:
Year
1990    44.207754
1991    43.857930
1992    43.421899
1993    43.066339
1994    42.478104
1995    42.072030
1996    41.676333
1997    41.179378
1998    40.682634
1999    40.148109
2000    39.775834
2001    39.608394
2002    39.155676
2003    38.955386
2004    38.649284
2005    38.282379
2006    37.977525
2007    37.561486
2008    37.176777
2009    36.944732
2010    36.524440
2011    36.180339
2012    35.770940
...
2020    33.432752
2021    33.118000
2022    32.753976
2023    32.634478
```

Interpretation:

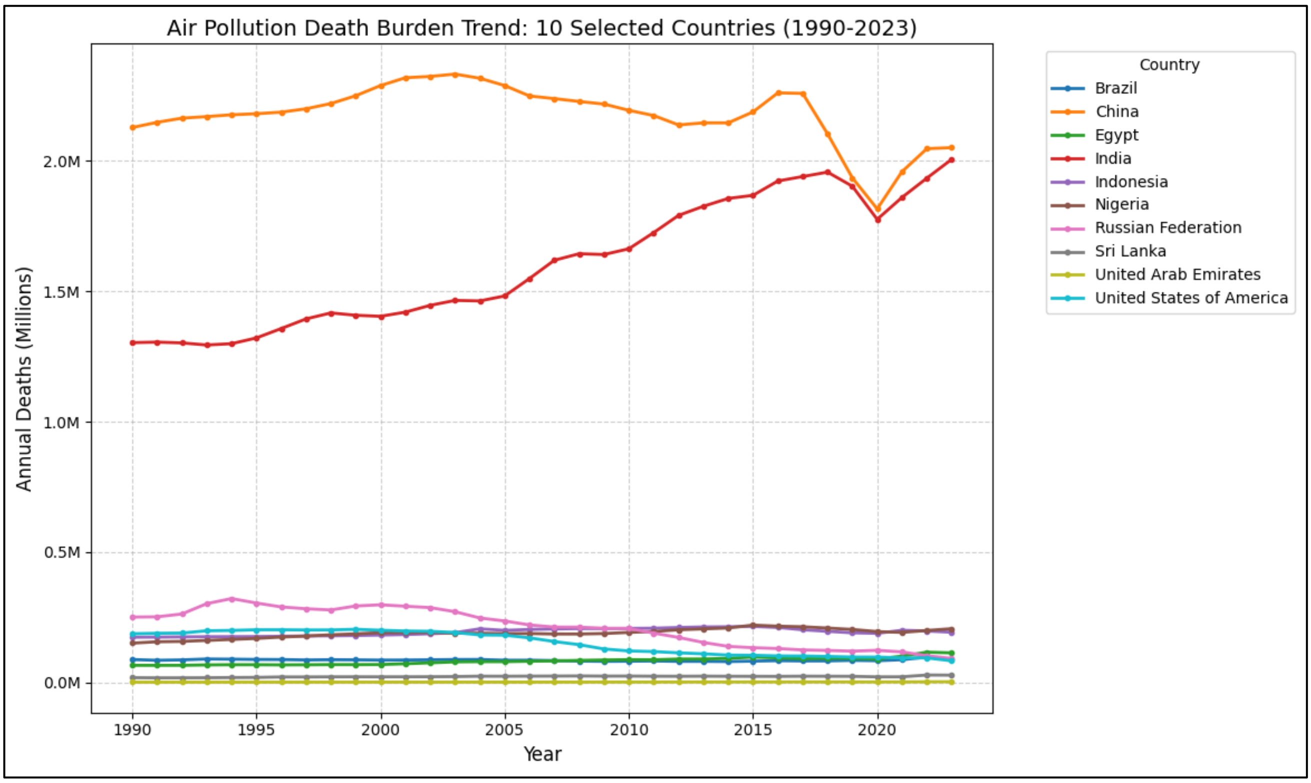
- The decline in the DALY-to-Death Ratio highlights the good news. People are living longer and deaths linked to PM2.5 are happening in older age categories. This shows progress in overall health and life expectancy.
- But it is not all positive. The total number of deaths from PM2.5 is still increasing, meaning the problem hasn't gone away.
- In simple terms PM2.5 is still taking millions of lives, but now it is affecting people at older ages rather than declining lives short at younger categories.

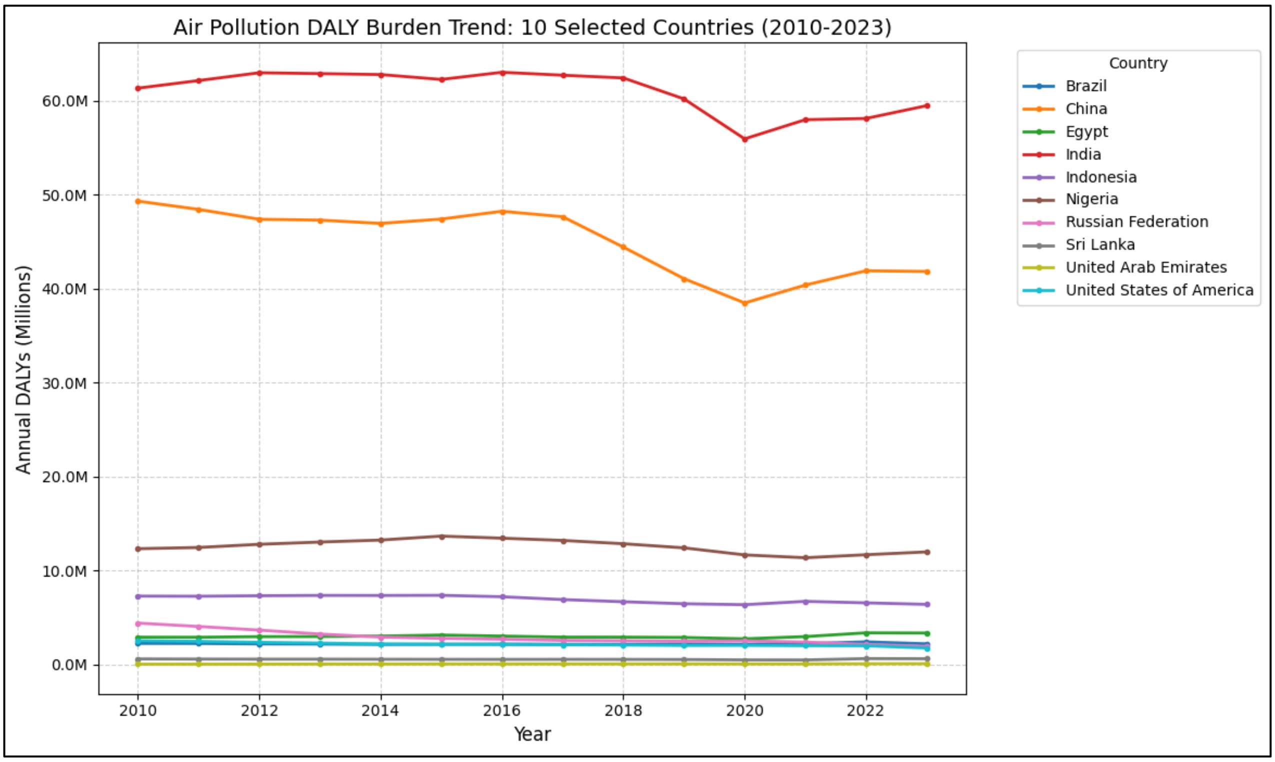
3.2. Country-Specific Trends:

Ten of each diverse group of countries were selected and generated two charts comparing their time-series trends for the Death and DALY burden attributed to Air Pollution and PM 2.5 both.

The countries selected are: Brazil, China, Egypt, India, Indonesia, Nigeria, Russian Federation, Sri Lanka, United Arab Emirates, United States of America.

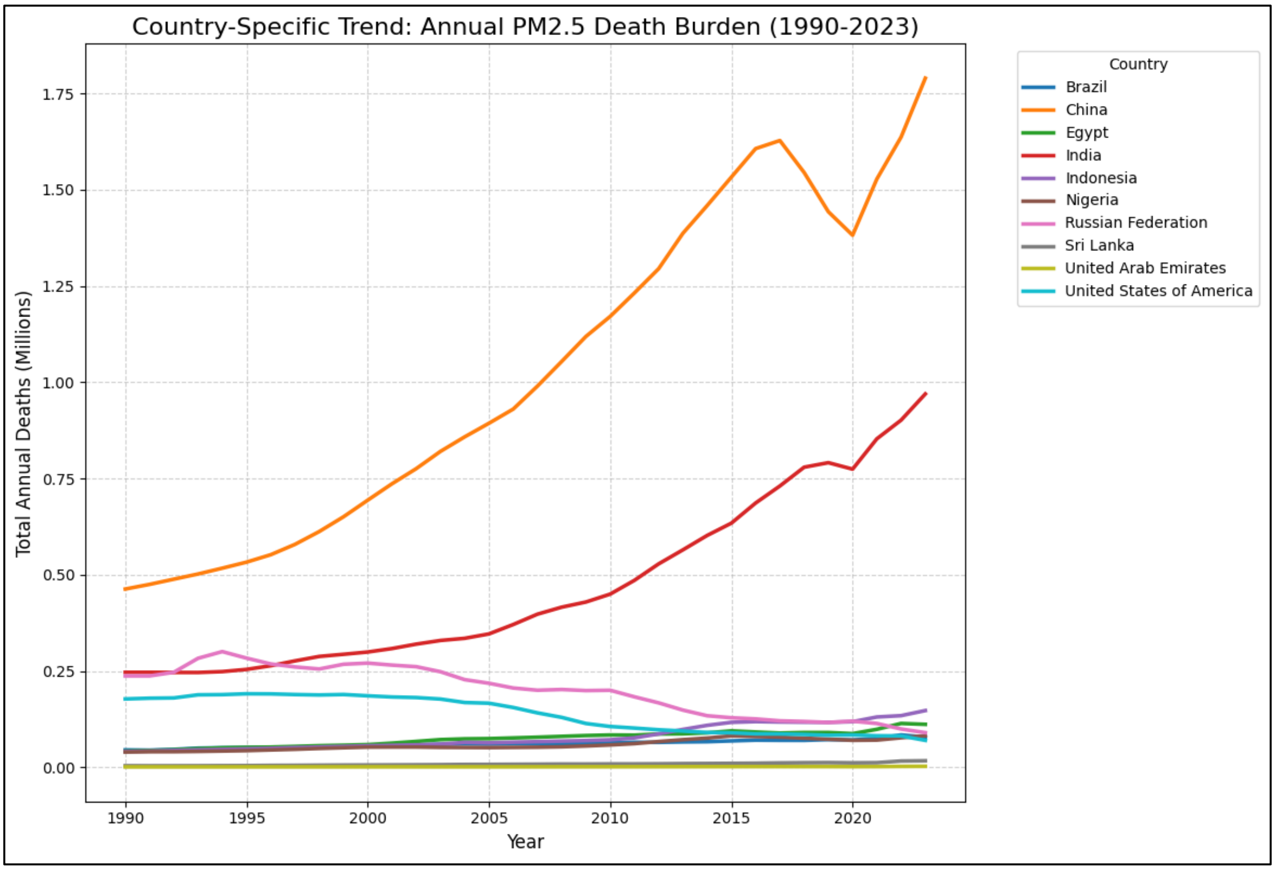
```
Country Death Burden Trend Data (1990 & 2023 comparison):
Country   Brazil      China     Egypt      India  Indonesia   Nigeria  Russian Federation  Sri Lanka  United Arab Emirates  United States of America
Year
1990     87500.0  2128000.0   65400.0  1303000.0   173100.0  150400.0            250400.0    18000.0                 673.0                  186400.0
2023     87400.0  2051000.0  113100.0  2006000.0   192100.0  205800.0             91700.0    27600.0                2300.0                   82500.0

Country DALY Burden Trend Data (2010 & 2023 comparison):
Country     Brazil       China      Egypt       India  Indonesia     Nigeria  Russian Federation  Sri Lanka  United Arab Emirates  United States of America
Year
2010     2269000.0  49310000.0  2931000.0  61320000.0  7301000.0  12340000.0           4430000.0   617400.0               58500.0                 2550000.0
2023     2224000.0  41830000.0  3373000.0  59460000.0  6421000.0  12000000.0           1952000.0   639200.0              109900.0                 1777000.0
```





Country-Specific PM2.5 Death Trend Data (Sample):

Country Year	Brazil	China	Egypt	India	Indonesia	Nigeria	Russian Federation	Sri Lanka	United Arab Emirates	United States of America
1990	44800.0	462800.0	41500.0	246500.0	40600.0	39300.0	237100.0	3740.0	673.0	177300.0
2000	55000.0	693700.0	58600.0	299200.0	55500.0	52200.0	270400.0	5680.0	868.0	185400.0
2010	63200.0	1171000.0	83700.0	449600.0	71100.0	57800.0	199800.0	8400.0	1140.0	105700.0
2023	78000.0	1790000.0	111300.0	969700.0	147100.0	81600.0	89700.0	16700.0	2280.0	69700.0



Key Observations and Comparison:

Country	1990 PM2.5 Deaths	2023 PM2.5 Deaths	Trend Summary
China	462,800	1,790,000	Dramatic Increase. PM2.5 is the dominant driver of China's air pollution deaths. It increased by nearly 400%, despite a decline in *total* air pollution deaths.
India	246,500	969,700	Substantial Increase. The PM2.5 death burden has increased almost 4 times by demonstrating the growing crisis of outdoor air quality.
Nigeria	39,300	81,600	Steady Growth. The burden more than doubled, reflecting ongoing urbanization and increasing outdoor pollution exposure.
USA	177,300	69,700	Significant Decline. PM2.5 deaths have been reduced by over 60% showing the success of continuous regulatory measures in high-income nations.
Pakistan	41,500	99,100	Steep Increase. The burden more than doubled which highlights an emerging major health challenge in this region.

<u>Overall Interpretation:</u>

- PM2.5 as the Main Problem:

In countries with heavy air pollution like China and India total deaths from air pollution have started to stabilize or even decline a bit because indoor pollution is being reduced. However, deaths caused specifically by PM2.5 (particulate matters in the air) have increased sharply. This shows that outdoor air pollution from PM 2.5 is now the biggest and fastest growing threat of air pollution deaths in these countries.

- Different Trends Across the World:

While the United States has seen a major reduction in deaths linked to air pollution countries such as China and India are experiencing steep increases. This difference clearly shows the large gap in air quality and health outcomes between high income and low income regions of the world.

4. <u>Interpretations & Insights</u>

- PM2.5 is the dominant threat: Especially in rapidly urbanizing countries like China and India.
- Health burden is shifting from younger to older populations globally, but Sub-Saharan Africa still sees high early-life impact.
- Policy implications:
  - ❖ High-income countries show that regulation works (e.g., USA).
  - ❖ Low- and middle-income countries need urgent interventions to reduce rising PM 2.5 exposure.