# REPORT

## Analysis of TB and TB+HIV Deaths Across Regions and Years

Shweta Ajay Shinde

Masters in Data Analytics, San Jose State University

Data 230: Data Visualization

Instructor: Venkata Duvvuri

1st Sept 2024

# Analysis of TB and TB+HIV Deaths Across Regions and Years

## Executive Summary

This report analyzes regional trends in tuberculosis (TB) and TB-related deaths combined with HIV across various areas over time. Visualizations such as charts illustrate the impact of these diseases on different populations. Key findings reveal significant regional differences and trends in TB and HIV-related mortality. These insights are essential for targeting public health interventions and resource allocation.

## Introduction

This report looks at how tuberculosis (TB) and deaths from TB combined with HIV vary across different regions over time. We used charts to make the data easier to understand and to spot patterns and trends. By analyzing this information, we can see how TB and HIV affect different populations. These findings are important for improving public health strategies and making sure resources are used where they are needed most.

## Methodology

Data was extracted from the TB_Burden_Country.csv file, which includes detailed records on TB and HIV-related deaths. The analysis focused on the following columns like:
Country or Territory, Region,
Estimated Total Population,
Method for Prevalence Estimates,
Estimated TB Deaths (excluding HIV) etc.
This data was used to identify trends and regional impacts.

## Summary of Visualization

### 1.Bar Plot: Total population by region:

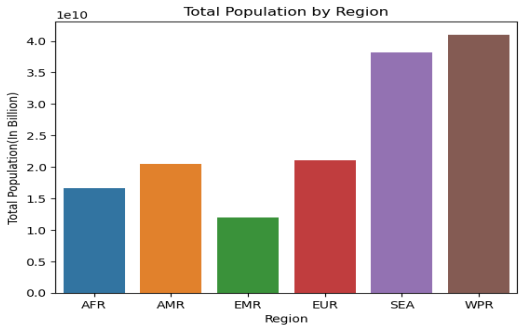
**Description**: This chart highlights the total population of every region. This will help us analyze further how many are affected by TB.

**Insights:**

- Region WPR has the highest population distribution around 40 billion.
- EMR has the lowest population distribution at around 11 billion.

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Load the Excel file
TB = pd.read_csv(r'F:\TB_Burden_Country.csv')

# Query the data group by 'Region' and sum the 'Estimated total population number'
grouped_data = TB.groupby(['Region'])['Estimated total population number'].sum().reset_index()
print(grouped_data)

sns.barplot(x='Region', y='Estimated total population number' ,data=grouped_data)

# Add labels and title
plt.xlabel('Region')
plt.ylabel('Total Population(In Billion)')
plt.title('Total Population by Region')
# Show the plot
plt.show()
```

```
  Region  Estimated total population number
0    AFR                         16622102766
1    AMR                         20467595830
2    EMR                         11936276692
3    EUR                         21032782394
4    SEA                         38194876821
5    WPR                         41029528965
```



**Method:**
First, we group the data by region and calculate the sum of the population for each region. Then, using the new data variable we plot these sums on a bar graph using sns.barplot to visualize the population distribution across regions. On x axis we plot population and on y the regions.

### 2.Line Plot: Total Death in every region due to TB and TB+HIV

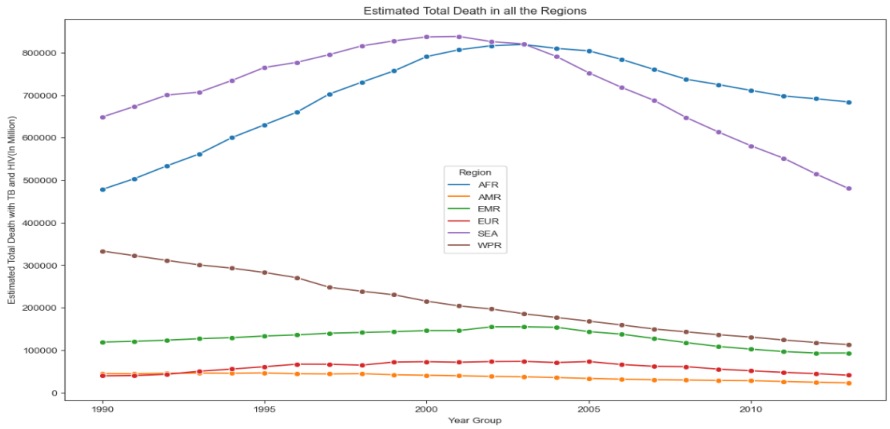
**Description:** This plot shows which countries bear the highest burden of death due to TB and TB+HIV cases.

**Insights:**

- AFR and SEA have the highest number of deaths.
- WPR deaths were high in the year 1990, later it gradually decreased till the year 2013.

```
grouped_data2 = TB.groupby(['Region','Year'])['Total_Death_TB_and_HIV'].sum().reset_index()
print(grouped_data2)

plt.figure(figsize=(14, 8))
sns.lineplot(data=grouped_data2, x='Year', y='Total_Death_TB_and_HIV', hue='Region', marker='o')

# Set plot labels and title
plt.title('Estimated Total Death in all the Regions')
plt.xlabel('Year Group')
plt.ylabel('Estimated Total Death with TB and HIV(In Million)')

# Display the plot
plt.show()
```

**Method:**
First, we created a new column in the file, 'Total_Death_TB_and_HIV', by summing the (column R) and (column X). We then grouped the data by Region and Year, calculated the sum of deaths for each group, and plotted a line graph using sns.lineplot. On x axis we plot deaths and on y the year. The hue are regions with different colors for better understanding.




### 3.Scatter Plot: Method to derive prevalence estimates

**Description**: This plot helps to understand the Method to derive prevalence estimates adopted by different regions.

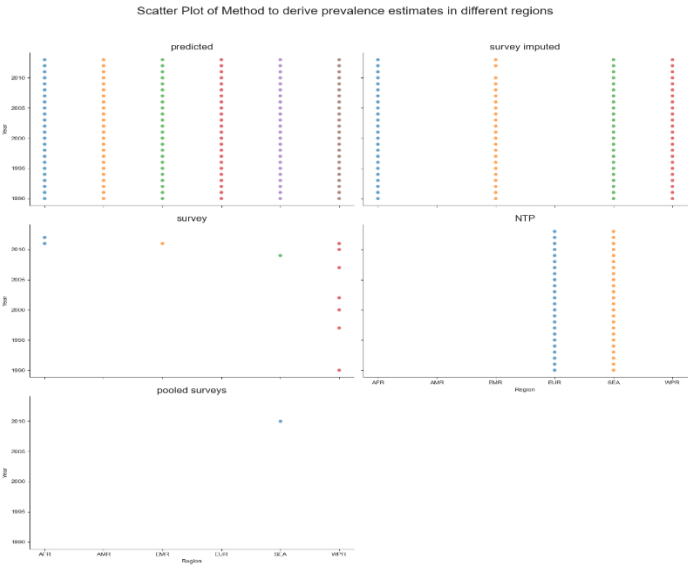
**Insights:**

- Most of the regions have used the ‘predicted’ method all these years.

**Analysis of TB and TB+HIV Deaths Across Regions and Years**

- The ‘pooled survey’ is used in just SEA region in the year 2012.

```
# Group data
grouped_data3 = TB.groupby(['Region', 'Year', 'Method to derive prevalence estimates'])['Total_Death_TB_and_HIV)'].sum().rese

# Create a FacetGrid with larger plots and more informative details
g = sb.FacetGrid(
    grouped_data3,
    col="Method to derive prevalence estimates",
    col_wrap=2,  # Adjust number of columns per row; change as needed
    height=5,    # Height of each facet in inches
    aspect=1.5   # Aspect ratio (width / height)
)

# Map scatter plot to each facet
g.map_dataframe(
    sb.scatterplot,  # Use seaborn's scatterplot to handle color mapping
    x="Region",  # X-axis
    y="Year",                       # Y-axis
    hue="Region",                   # Color by Region
    alpha=0.7,                      # Transparency for better visibility of overlapping points
    edgecolor='w'                   # White edge color for better contrast
)

# Add titles and labels to the facets
g.set_axis_labels("Region", "Year")
g.set_titles(col_template="{col_name}", size=16)  # Set titles for each subplot

# Improve layout and spacing
plt.subplots_adjust(top=0.9)
g.fig.suptitle("Scatter Plot of Method to derive prevalence estimates in different regions", fontsize=20)

# Show the plot
plt.show()
```




**Method:**

In this plot we first create a group of 'Region', 'Year', 'Method to derive prevalence estimates' to create different scatter plot with sb.scatterplot for all the methods used in different regions. On x axis we plot region and on y the year.

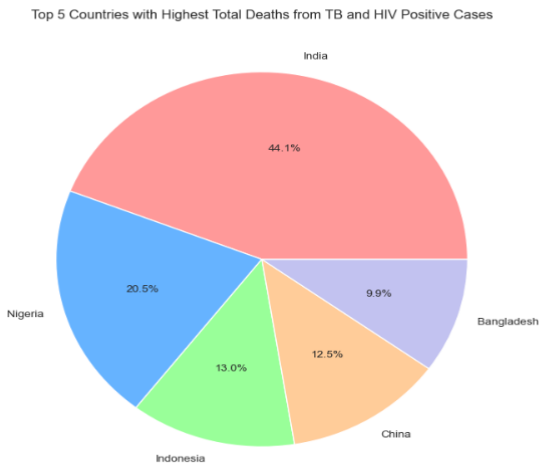
**4.Pie Chart: Top 5 countries with maximum number of deaths**

**Description**: This chart offers insights into the relative contribution of 5 countries to the total number of deaths. It shows how much a high these countries have contributed to the total number of deaths, helping decision-makers see where they should focus their efforts.

**Insights:**

- 'India' has the highest contribution to TB and TB+HIV deaths, followed by Nigeria, Indonesia, China, and Bangladesh.

```
import pandas as pd
import sqlite3
# Load CSV data into a pandas DataFrame
csv_file_path = r'F:\TB_Burden_Country.csv'
df = pd.read_csv(csv_file_path)

# Create an in-memory SQLite database
conn = sqlite3.connect(':memory:')
df.to_sql('TB_Burden_Country', conn, index=False, if_exists='replace')

# Define SQL query to find the top 5 countries with the highest number of deaths
query = """
SELECT
    Country,
    SUM(Total_Death_TB_and_HIV) AS Total_Deaths
FROM
    TB_Burden_Country
GROUP BY
    Country
ORDER BY
    Total_Deaths DESC
LIMIT 5;
"""
# Execute the SQL query and fetch the results
top5_countries_df = pd.read_sql_query(query, conn)

# Define a list of distinct colors
colors = ['#ff9999','#66b3ff','#99ff99','#ffcc99','#c2c2f0']

# Plotting the pie chart
plt.figure(figsize=(8, 8))
plt.pie(
    top5_countries_df['Total_Deaths'],
    labels=top5_countries_df['Country'],
    autopct='%1.1f%%',
    colors=colors
)
plt.title('Top 5 Countries with Highest Total Deaths from TB and HIV Positive Cases')
```



**Method:**
First, we connect to the file using an SQL query to retrieve the top 5 countries with the highest number of deaths. We then use these records to create a pie chart using plt.pie that illustrates the contribution of each of these top 5 countries to the total death toll.

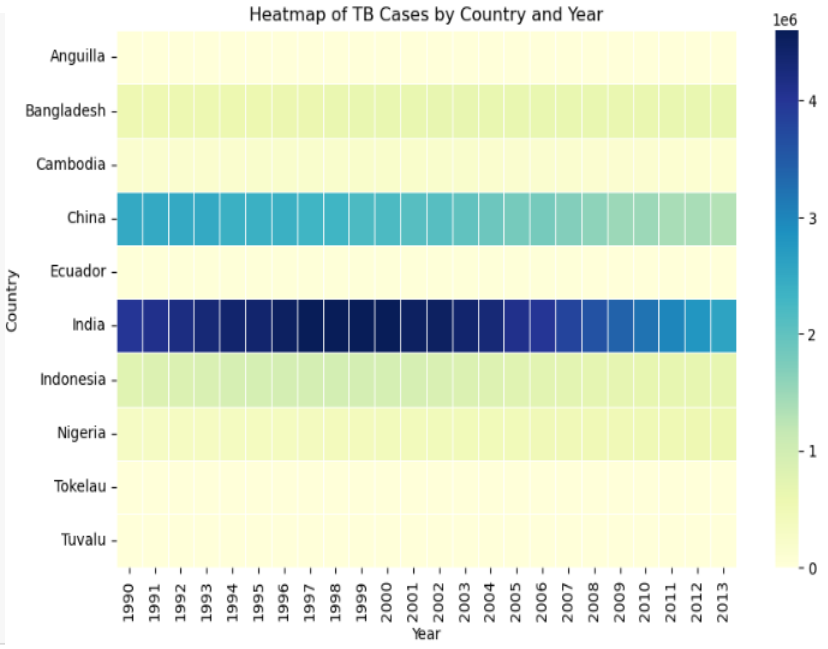
### 5.Heatmap: TB Cases by Country

**Description**: The heatmap visualizes the density of TB cases in a few countries, using color intensity to represent values.

**Insights:**

- India and China have maximum cases of TB.
- Tuvalu and Anguilla have very less or no cases of TB

```
# Select 10 random countries
# Drop duplicates to ensure unique country names if necessary
unique_countries = df['Country'].drop_duplicates()
random_countries = unique_countries.sample(n=10, random_state=1).tolist()

print("Selected Countries:", random_countries)

# Filter the data for the selected countries
filtered_df = df[df['Country'].isin(random_countries)]

# Pivot the data for the heatmap
# Assuming columns 'Year' and 'Estimated prevalence of TB (all forms)'
pivot_table = filtered_df.pivot_table(
    index='Country',
    columns='Year',
    values='Estimated prevalence of TB (all forms)'
)

# Create the heatmap
plt.figure(figsize=(10, 6))
sb.heatmap(
    pivot_table,
    #annot=True,        # Annotate cells with the numeric value
    cmap='YlGnBu',     # Color map
    linewidths=.5,     # Width of lines separating cells
    fmt='.0f'          # Format for annotation
)
plt.title('Heatmap of TB casesCountry and Year')
plt.xlabel('Year')
plt.ylabel('Country')
plt.show()
```




**Method:**
First, we write a code to randomly select 10 countries from the dataset. Then, we create a heatmap with sb.heatmap using the selected countries, years, and the number of TB cases reported in those countries. On x axis we plot countries and on y the years.

### 6.Boxplot: TB Cases by Country

**Description**: The box plot visualizes the distribution of the estimated prevalence of TB (all forms) across selected countries, highlighting central tendencies and variability.

**Insights:**

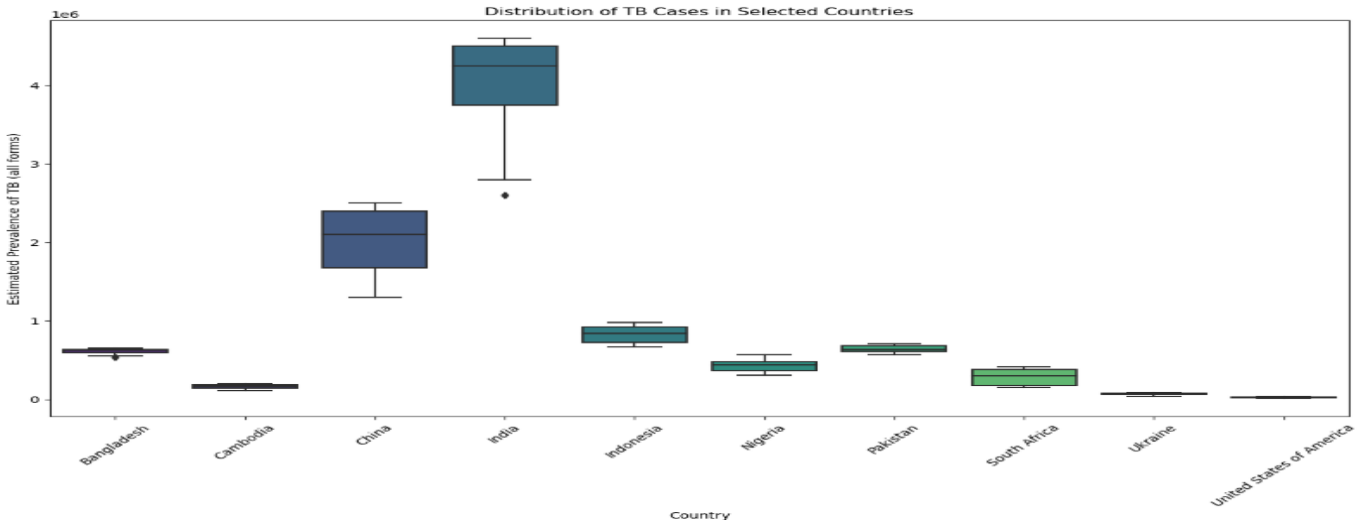
- India and China have maximum cases of TB.
- United States and Ukraine has very a smaller number of cases.

**Analysis of TB and TB+HIV Deaths Across Regions and Years**

```
# Filter for selected countries
selected_countries = [
    'India', 'Nigeria', 'Indonesia', 'China', 'Bangladesh',
    'Cambodia', 'United States of America', 'Ukraine', 'Pakistan', 'South Africa'
]
filtered_df = df[df['Country'].isin(selected_countries)]

# Create a box plot
plt.figure(figsize=(15, 8))
sb.boxplot(data=filtered_df, x='Country', y='Estimated prevalence of TB (all forms)', palette='viridis')

# Add titles and labels for clarity
plt.title('Distribution of TB Cases in Selected Countries')
plt.xlabel('Country')
plt.ylabel('Number of Cases')

# Show the plot
plt.xticks(rotation=45)  # Rotate x-axis labels for better readability if necessary
plt.show()
```

## Method:

First, we list a selection of countries to plot their data, and subsequently, we use the same set of data to generate a box plot with sb.boxplot for comparative analysis. On x axis we plot countries and on y the number of cases.




## Conclusion

The visualizations provide a comprehensive view of the global TB situation.

- **India** has the highest burden of TB cases and deaths.
- Regions like **South Asia** and **Sub-Saharan Africa** show significant numbers in both cases and deaths.
- Trends over time indicate increasing death rates in certain regions. These insights are crucial for targeting public health interventions and allocating resources effectively.
- Most of the regions have used the '**predicted**' method all these years.

## Recommendation

- **Focus on High-Burden Regions:** Target interventions in India, South Asia, and Sub-Saharan Africa.
- **Improve Data Accuracy:** Enhance surveillance to better track and address rising death rates.
- **Strengthen Health Programs:** Boost TB control efforts in Africa and Southeast Asia.
- **Use Predictive Models:** Continue refining predictions to stay ahead of trends.
- **Encourage Regional Collaboration:** Share resources and strategies across affected regions.

## Reference

Waskom, M. (2024). *Seaborn examples*. Seaborn. Retrieved from https://seaborn.pydata.org/examples/index.html

San José State University. (n.d.). Datafile. Retrieved, from https://sjsu.instructure.com/courses/1595119/files/78129236?wrap=1