Exploratory Data Analysis on Global TB Burden Using Healthcare Dataset

Presented by: Akshith Chidurala
Intern – Rhives Technologies
Week 3 Project: Data Analysis using Python

Contents

1. Project Introduction
2. Problem Statement
3. Dataset Description
4. Data Preprocessing
5. Data Analysis
6. Visualizations & Findings
7. Dashboard
8. Insights
9. Conclusion

Project Introduction

- This project analyzes the global burden of Tuberculosis (TB) using real healthcare data.
- The goal is to understand patterns in deaths, incidence rates, regional variations, and risk factors.
- Python is used to perform:
- Data cleaning
- EDA
- Visualization
- Insights generation

Problem Statement

- TB remains one of the world’s leading infectious causes of death.
- Understanding the patterns of TB deaths and incidence can help governments and health organizations take better action.
- The dataset contains multiple indicators related to TB, and analyzing them helps identify:
- High-burden regions
- Key risk metrics
- Trends over time

Dataset Description

Dataset Name: Rhives Healthcare Dataset
Total Rows: 5120
Total Columns: 47
Key Columns:

- Country / Region
- Year
- Estimated TB Deaths
- Population
- TB Incidence per 100,000
- Case Detection Rate
- TB Prevalence
- Mortality Estimates

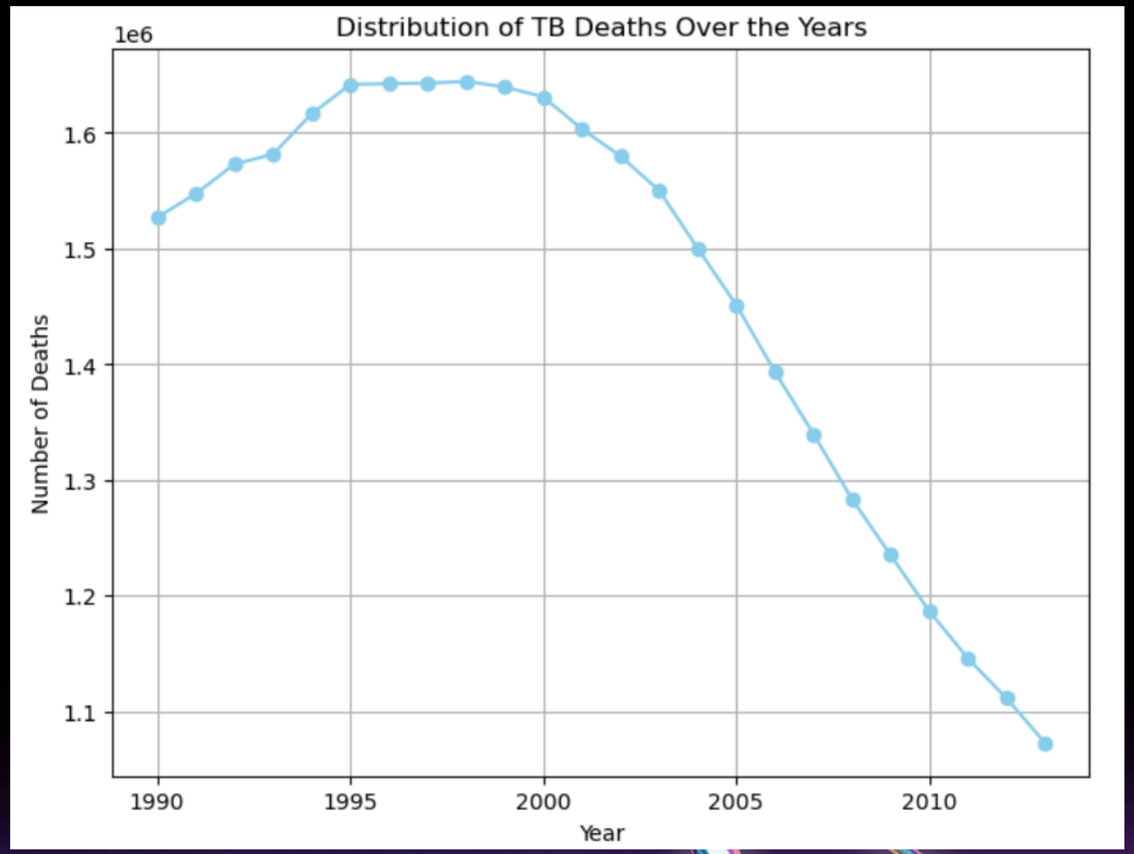
Data Preprocessing

- Loaded dataset using Pandas
- Checked data types
- Handled missing values
- Verified no duplicate rows
- Converted important columns to numeric
- Created new metric: Deaths per 100k population

Data Analysis

- Performed Exploratory Data Analysis (EDA) to understand global TB patterns and disease burden.
- Conducted univariate analysis to study individual metrics such as TB deaths, incidence per 100k, and regional distribution.
- Applied bivariate analysis to examine relationships like population vs deaths and region-wise variations in mortality.
- Used multivariate analysis to explore combined relationships between incidence, deaths per 100k, and region.
- Identified high-burden countries and regions with significantly higher mortality and incidence rates.
- Detected strong correlations between incidence and mortality, highlighting critical risk patterns.
- Extracted meaningful insights to support data-driven public health decisions.

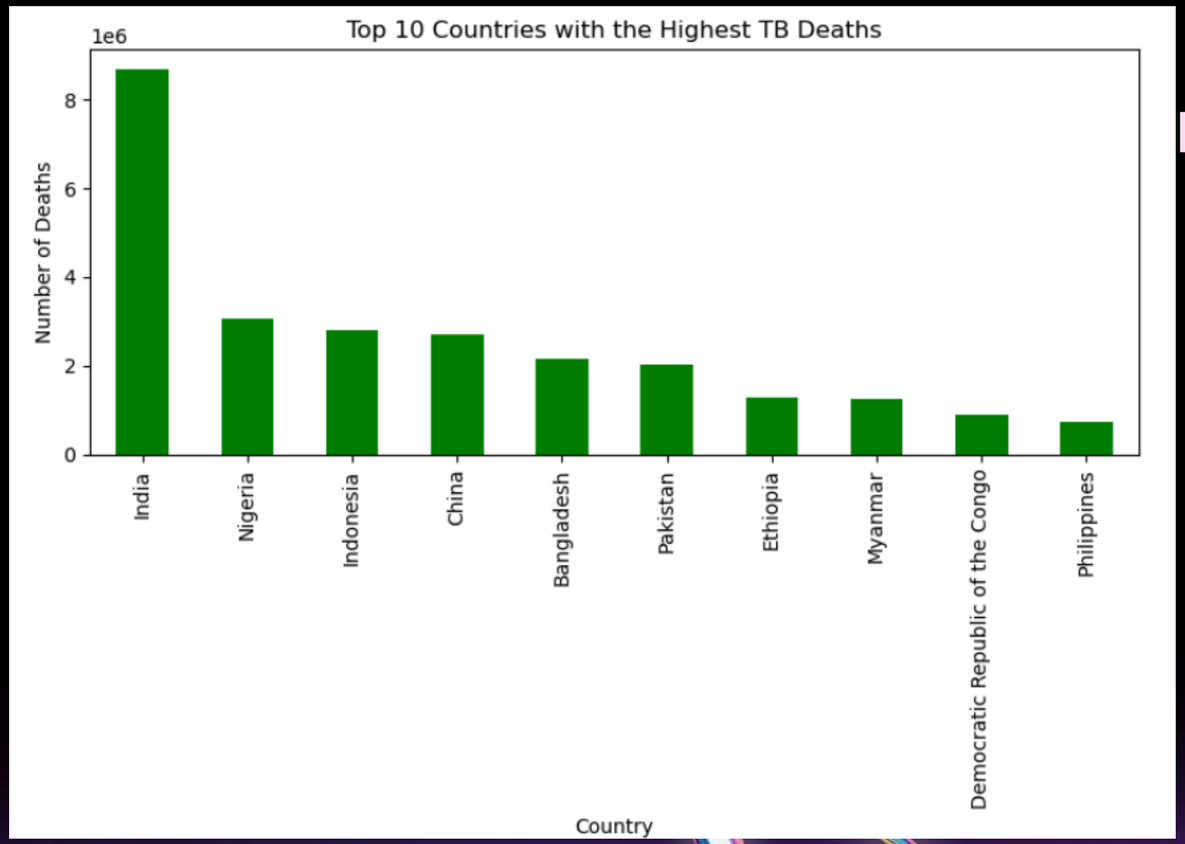
Distribution of Deaths Over the Years



Findings:

- TB deaths fluctuate across years.
- Some years show significant increases, indicating periods of higher disease activity.

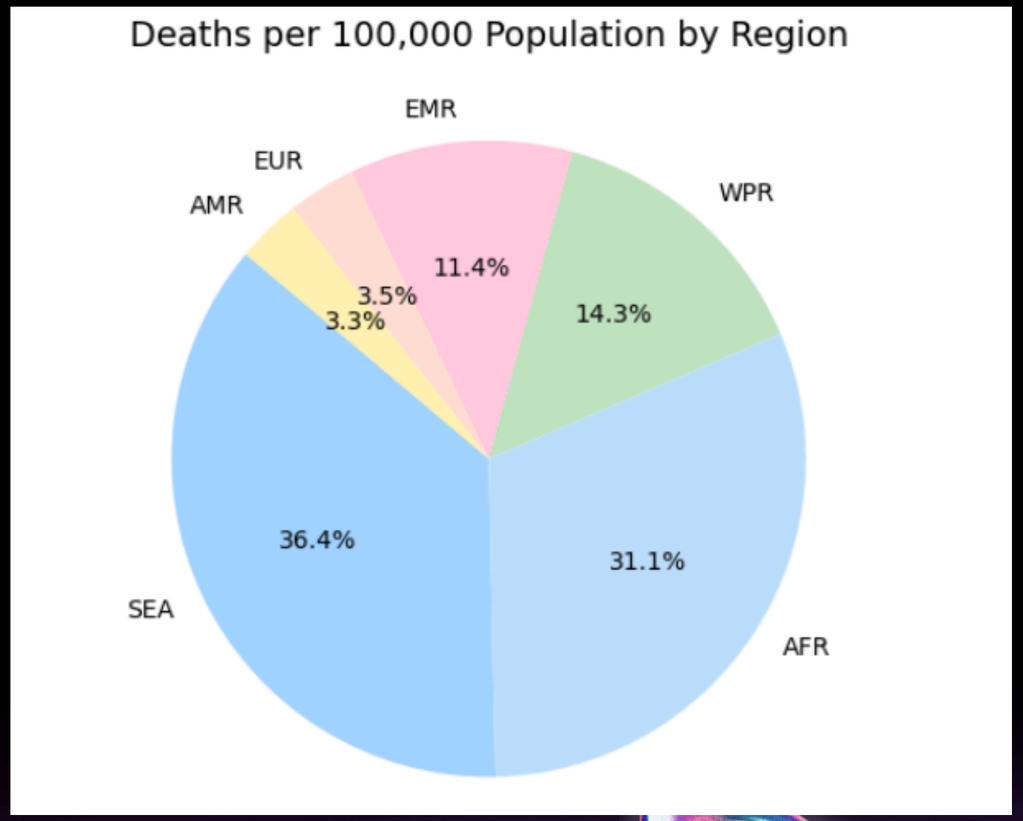
Top 10 Countries with the Highest TB Death Rates



Finding:

- A few countries dominate the global TB burden.
- These countries need prioritized interventions.

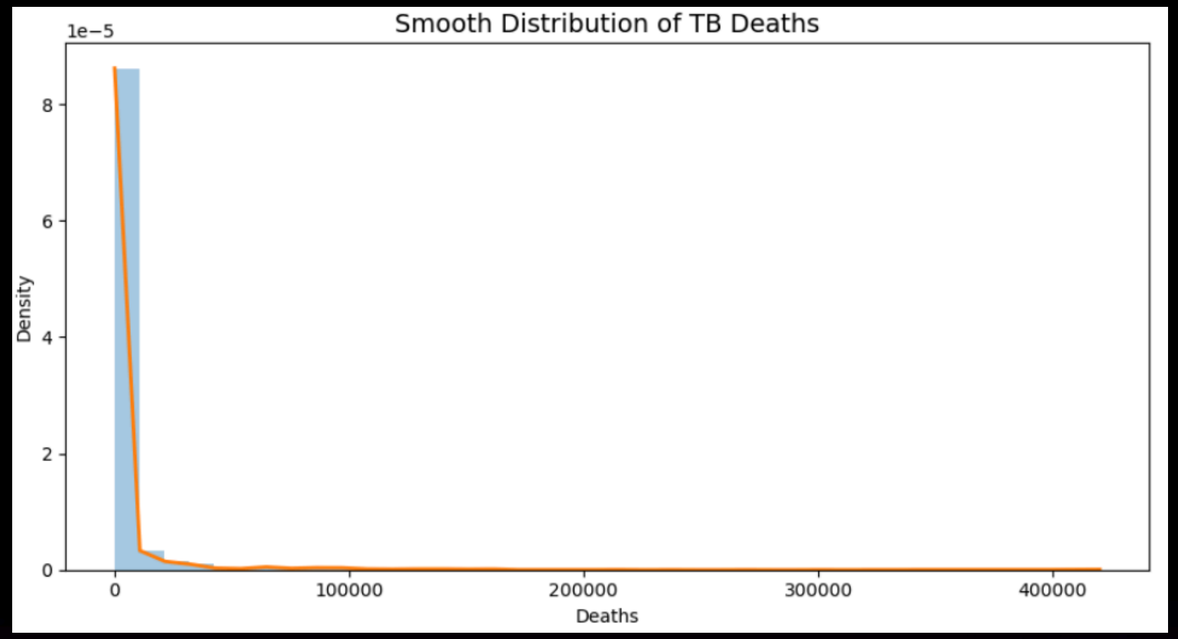
Deaths per 100k by Region



Finding:

- Certain regions show higher standardized death rates.
- Indicates unequal healthcare access and disease control.

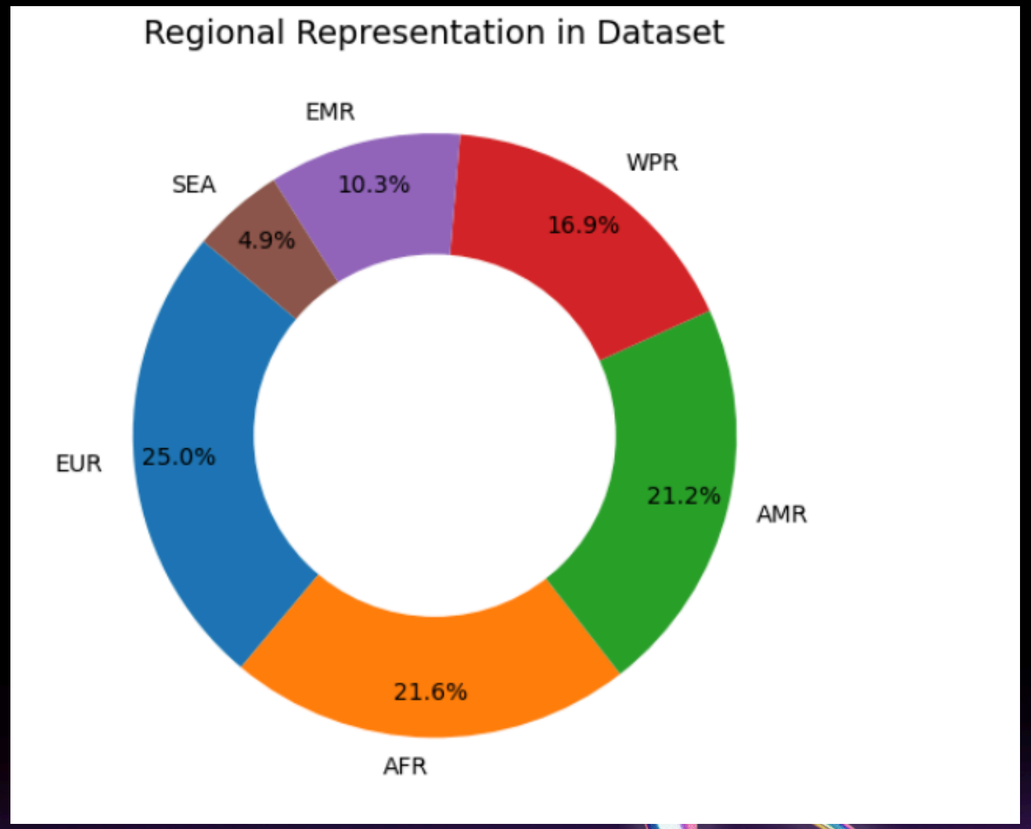
Univariate Plot 1: Distribution of TB Deaths



Findings:

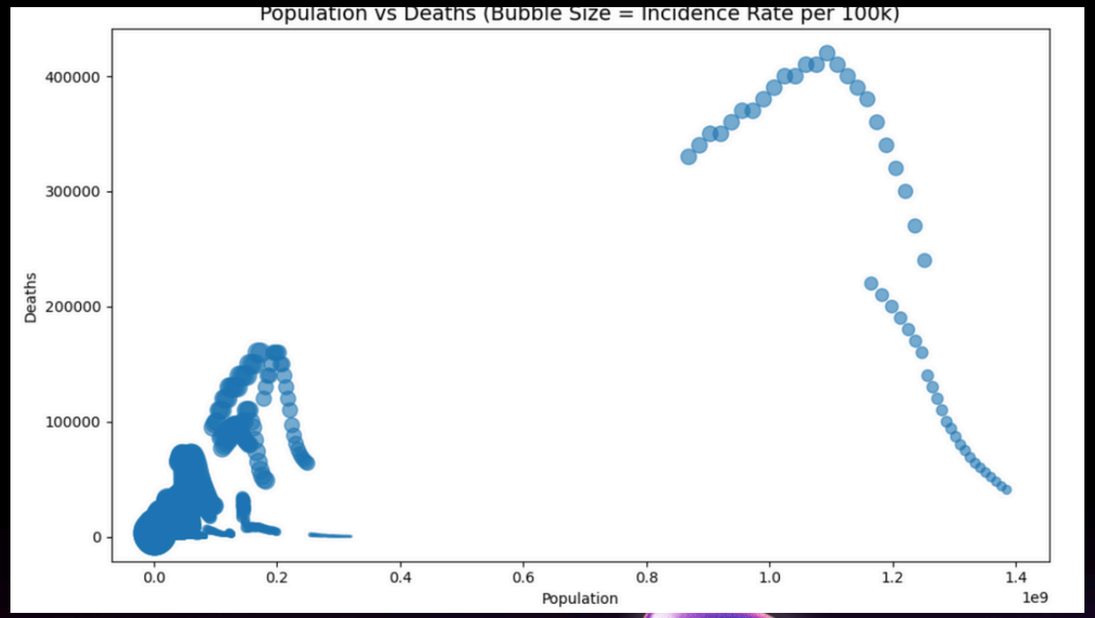
TB deaths show a strong right-skew, meaning only a few countries have very high deaths while most countries report low-to-moderate TB deaths.

Regional Representation (Donut Chart)



Finding:
The dataset is unevenly distributed across regions, with some regions contributing significantly more records than others.

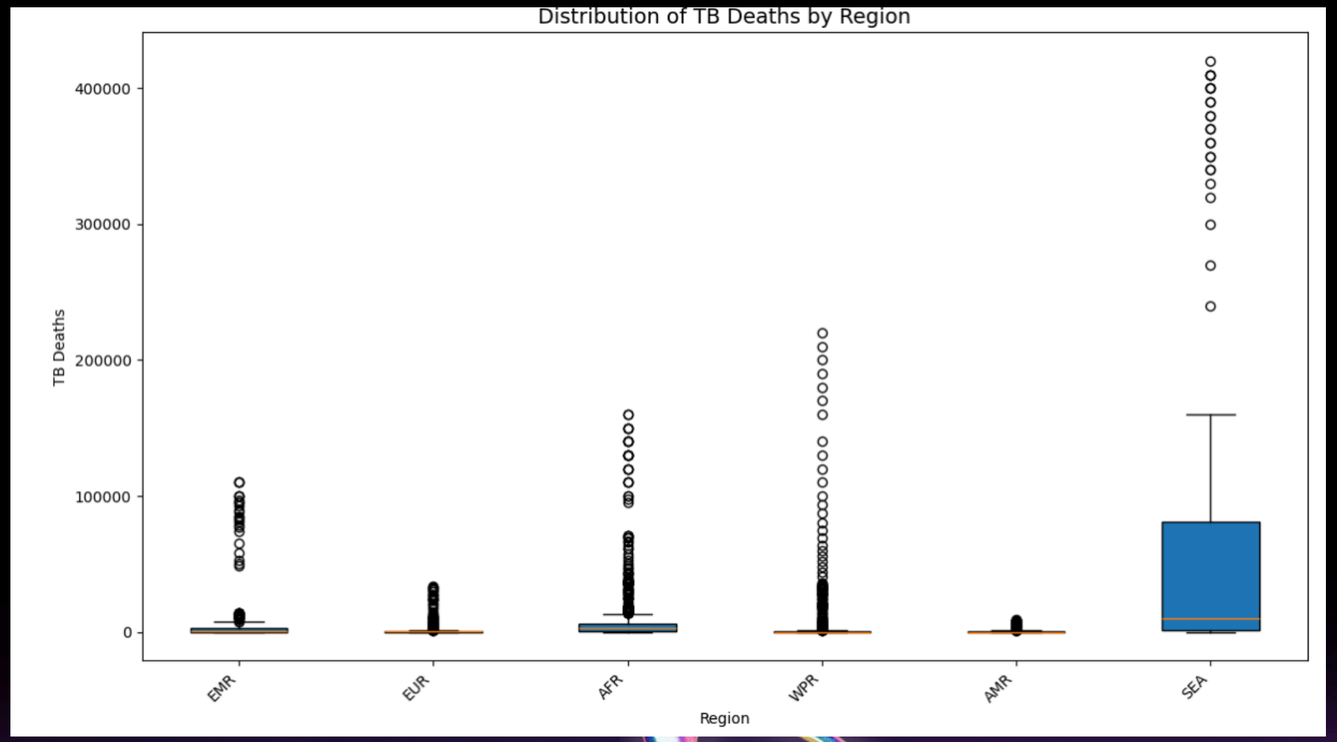
Population vs Deaths (Bubble Plot)



Finding:

Countries with larger populations generally show higher total TB deaths, but the bubble sizes reveal that incidence per 100k varies independently of population size. Some moderately populated countries have very high incidence, contributing to larger TB burdens.

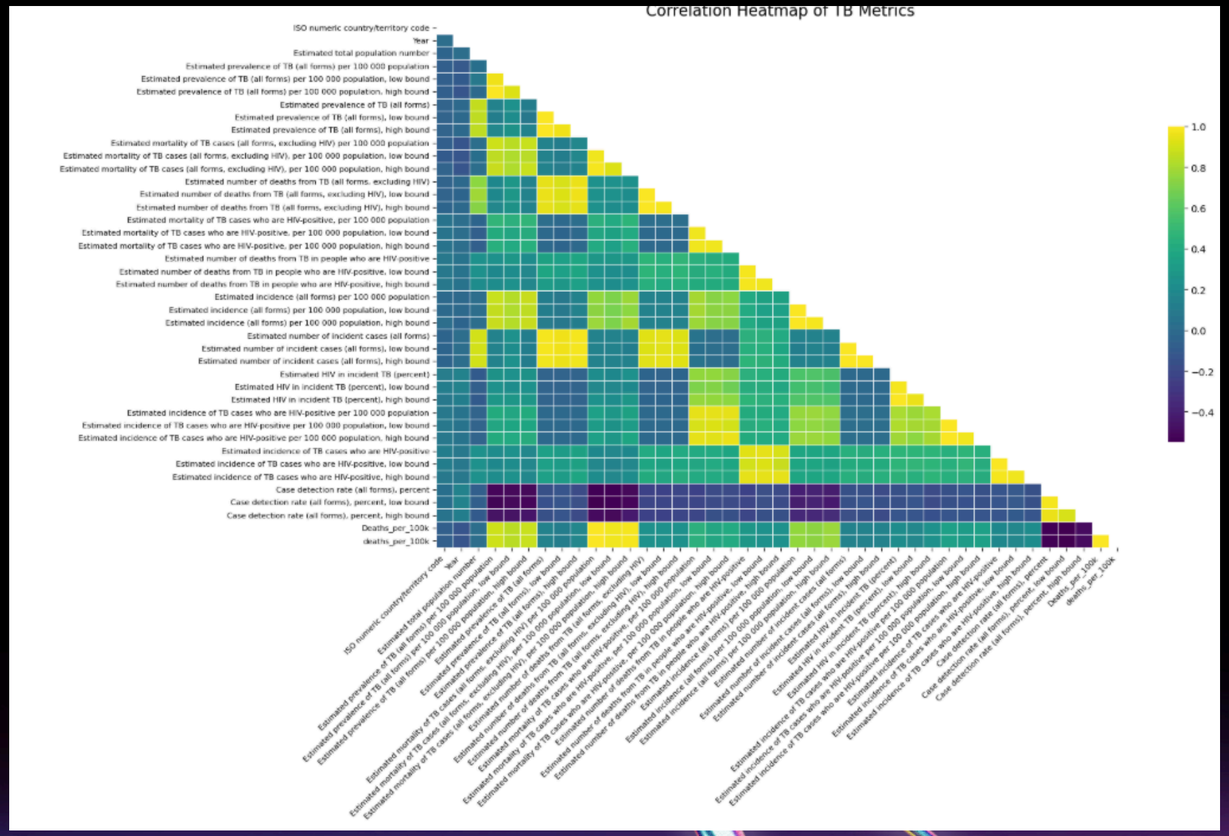
Boxplot – TB Deaths by Region



Finding:

TB deaths vary widely across regions, with some regions showing a larger spread and higher median deaths. This variation highlights regional inequality in TB control, healthcare access, and outbreak management.

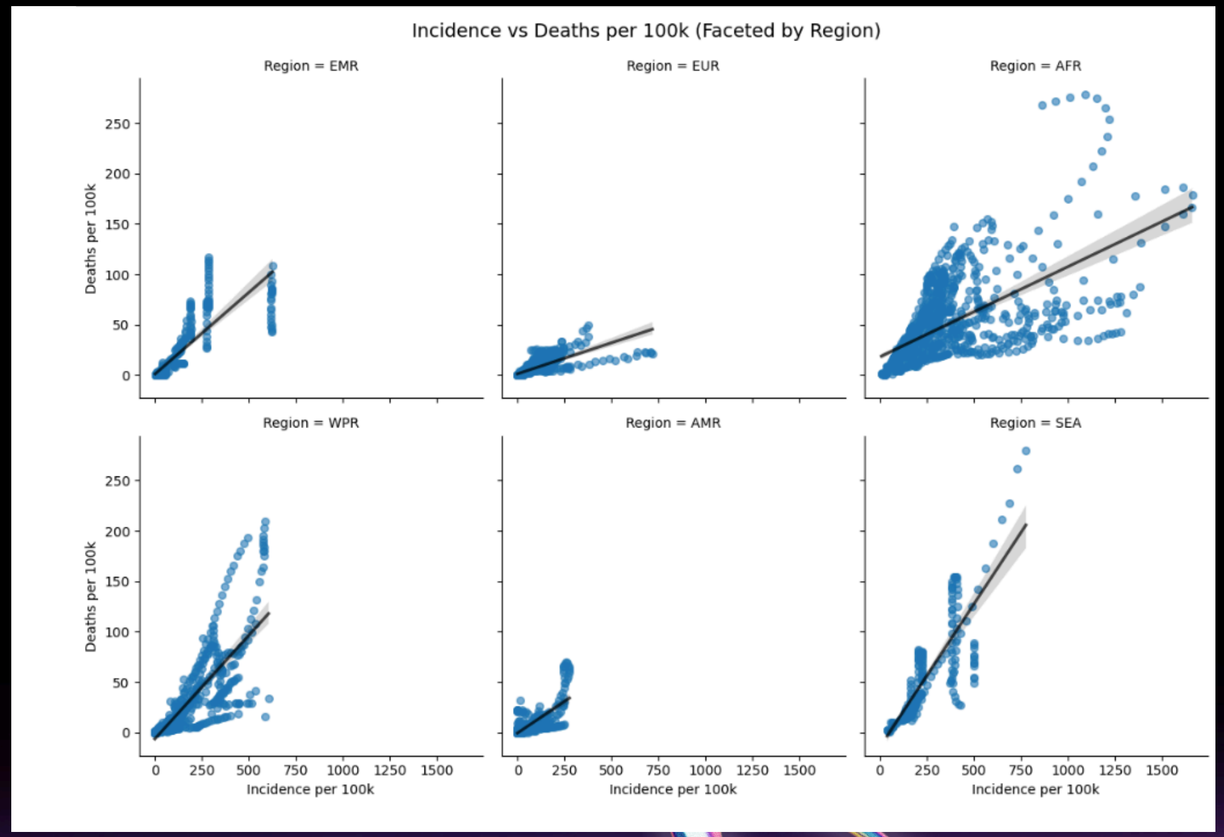
Correlation Heatmap of TB Metrics



Finding:

The heatmap shows that TB incidence per 100k and deaths per 100k are strongly correlated, confirming incidence as the strongest predictor of TB mortality. Several other demographic variables show weak or moderate correlations, indicating that mortality is primarily driven by infection intensity rather than general population metrics.

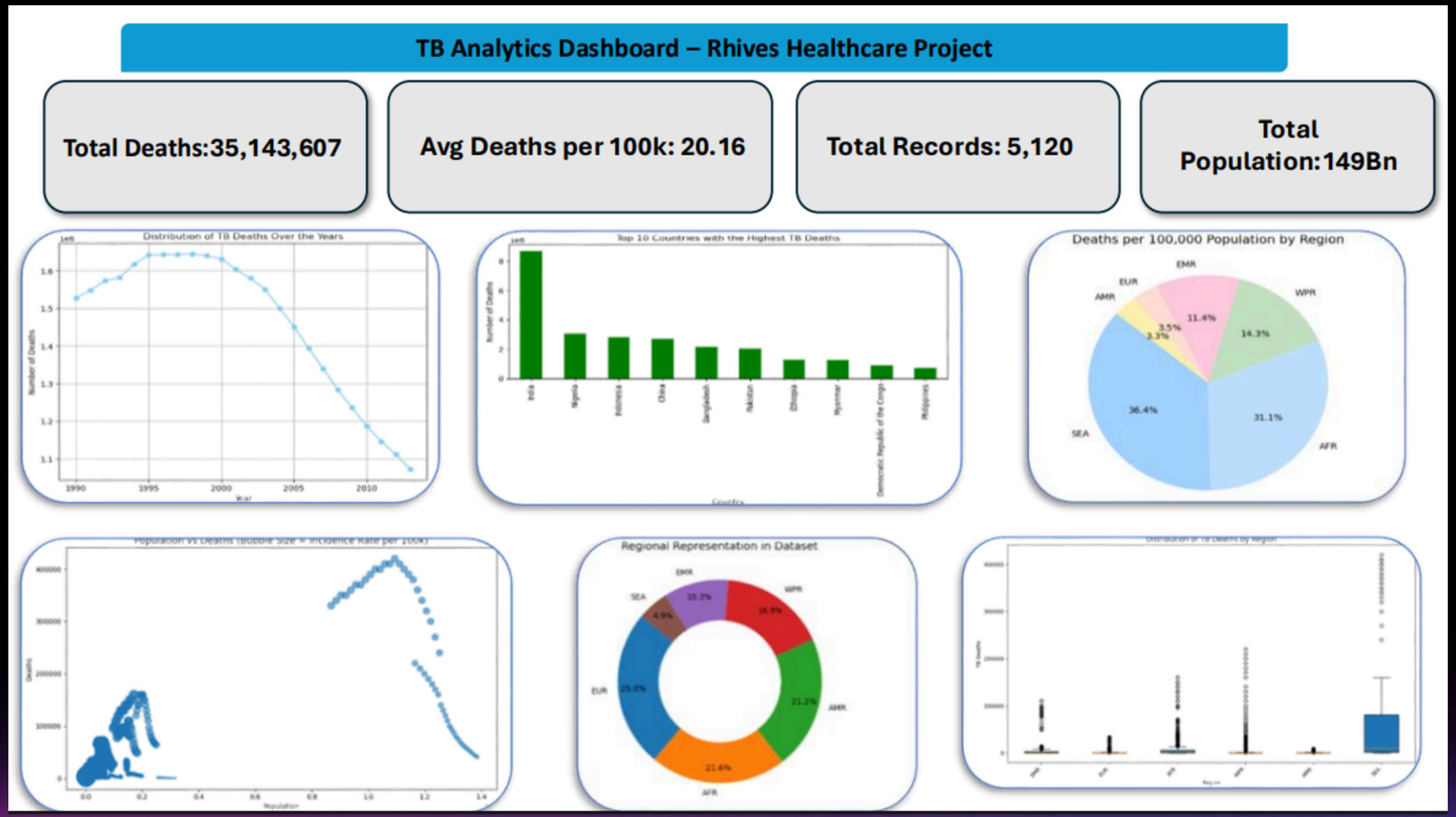
LMPlot – Incidence vs Deaths per 100k



Finding:

Each region displays a different relationship between incidence and mortality. Some regions show steeper slopes, meaning deaths rise quickly with increasing incidence, while others show flatter trends, suggesting better TB control or healthcare systems. This highlights clear regional differences in mortality response to TB incidence.

Dashboard


TB Analytics Dashboard – Rhives Healthcare Project
Total Deaths:35,143,607
Avg Deaths per 100k: 20.16
Total Records: 5,120
Total Population:149Bn
Distribution of TB Deaths Over the Years
Top 10 Countries with the Highest TB Deaths
Deaths per 100,000 Population by Region
EMR
EUR
AMR
WPR
SEA
AFR
11.4%
14.3%
3.5%
3.3%
36.4%
31.1%
Regional Representation in Dataset
10.3%
18.9%
21.2%
21.6%
25.6%
4.9%

Insights

1. Death Trend
TB deaths fluctuate over the years, showing periods of increase and decline.
2. High-Burden Countries
A few countries consistently record very high TB death rates due to higher incidence and weaker healthcare systems.
3. Regional Patterns
Some regions show higher deaths per 100k population, indicating uneven disease burden.
4. Population vs Deaths
Larger populations often show higher total deaths, but not always higher mortality rates.
5. Distribution Insights
TB death data is right-skewed: most countries have low deaths, while a few have extremely high numbers.
6. Multivariate Analysis
Incidence per 100k strongly relates to deaths per 100k.
Regions show different patterns, indicating varied health infrastructure and detection rates.

.

Conclusion

- TB continues to be a major global health concern with a highly uneven burden across countries and regions.
- A small number of high-burden countries contribute to most TB deaths, highlighting the need for focused interventions.
- Strong correlation between incidence per 100k and deaths per 100k shows that early detection and timely treatment are essential.
- Regional variations reveal differences in healthcare access, control programs, and disease management effectiveness.
- Insights from the analysis can guide policymakers to strengthen TB control strategies and allocate resources more effectively.


Thank you