# Report

## About data:

WHO supplies information on tuberculosis (TB) cases and estimates the TB burden for various countries. The data, along with the corresponding dictionary, were acquired from the WHO website: https://www.who.int/teams/global-tuberculosis-programme/data#csv_files

The relevant columns in the dataset encompass the year, country, g_whoregion, e_inc_100k, e_inc_100k_lo, e_inc_100k_hi, c_newinc, signifying the year, country code, WHO region, infection estimates per 100,000 population per year, lower bounds of infection estimates per 100,000 population per year, upper bounds of infection estimates per 100,000 population per year, and new or relapsed TB cases reported in that particular year, respectively.

## Data visualization:

Plot 1: This GIF presents a dynamic 2D map illustrating the global distribution of TB infection. The map is color-coded to represent varying levels of TB infection rates across different regions and countries. The color gradient ranges from low to high infection rates, allowing for a quick visual understanding of the prevalence of TB. With the year, the scale of colors on the map and the guide change, reflecting the temporal changes in TB infection rates. A global decrease in TB infection with the year is observed. This dynamic representation provides a comprehensive overview of how TB infection is distributed geographically and how it evolves over time.

Plot 2: This illustrates the TB infections and the annual GDP expenditure for a user-selected country or WHO region. The blue line denotes the estimated infection rates per 100,000 population per year, while the shaded region signifies the margin of error associated with the estimation method. Simultaneously, the red line depicts the GDP per capita in international dollars, with data sourced from the World Bank. The right axis provides the GDP scale. When the user selects 'Group' as 'WHO Region,' the cumulative data for all countries within that specific region is aggregated and displayed. This aggregation process is applied to both infection cases and GDP. Instances where there is a gap in GDP signify that no information is available. Hovering over the plot reveals the corresponding infection and GDP values, distinguished by their respective colors.

Plot 3: This Bar plot illustrates the estimated infection rates per 100,000 population by year for all the countries present in the WHO region.

## Time series analysis:

Auto-correlation (ACF) and partial auto-correlation (PACF) are statistical methods employed in time series analysis to grasp the patterns and relationships inherent in a sequence of observations. ACF quantifies the correlation between a time series and its lagged values (previous observations), while PACF gauges this correlation by eliminating the influences of intermediate lags. These plots play a crucial role in time series modeling for identifying the order of autoregressive (AR) or moving average (MA) models.

Plot 4: ACF and PACF plot. The horizontal axis represents the lag, and the vertical axis represents the correlation coefficient. Significant spikes in the ACF and PACF plots can indicate important patterns in the time series.

Plot 5: We implemented an auto ARIMA model (package forecasts in R) that evaluates several models with different degrees and returns the best model based on AICc. We used estimated infection rates per 100,000 population as observation (black lines and dots) and predicted the next 5 years. Model fitting are represented in blue dots and predicted values and corresponding error is shown in blue and shaded region.

Plot 6: Similar to plot 5 where we implemented an auto ARIMA model. We used new and relapsed cases as observations (black lines and dots) and predicted the next 5 years. Model fitting are represented in blue dots and predicted values and corresponding error is shown in blue and shaded region.