**PROJECT REPORT**

**Topic: Global Tuberculosis Statistical Report**

**Course:** **MFE-II**

**Group Members:**

* Sanya Goel - 220321
* Jasmine Sindhu - 220327
* Vidhika Mangla - 220628
* Rohan Dhingra - 220330

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| Serial No. | Content | Page No. |
| 1. | Introduction | 3 |
| 2. | Methodology | 4 |
| 3. | Data Dictionary | 5 |
|  | Analysis and Results |  |
| 4. | Descriptive Analysis | 6 |
| 5. | Correlation Analysis | 9 |
| 6. | Scatter Analysis | 10 |
| 7. | Time Series Analysis | 12 |
| 8. | Predictive Analysis | 15 |
| 9. | Conclusion | 17 |
| 10. | References | 18 |
| 11. | Contribution | 19 |

**INTRODUCTION**

Tuberculosis (TB) is a communicable disease caused by the bacterium Mycobacterium tuberculosis. Despite having the potential to harm the brain, spine, and kidneys, tuberculosis is a disease that mostly affects the lungs. When an infected person speaks, sneezes, or coughs, TB is spread through the air. People with compromised immune systems, such as those with HIV, are more likely to develop active TB. Despite being treatable, TB nevertheless poses a serious threat to global public health. With an anticipated 1.5 million fatalities from a single infectious agent in 2018, TB is one of the top 10 global causes of mortality. Additionally, TB is a significant public health issue, particularly in low- and middle-income nations. The Global Tuberculosis Statistical Report contains information on the prevalence, incidence, mortality, and other factors relating to TB globally. This statistical report study intends to investigate the effects of TB, particularly death rates, in many countries with and without HIV. The study will examine TB trends and patterns in various locations and continent wise using data from the years 2007 to 2014. To better comprehend the effects of this illness on a worldwide scale, a correlation between death rates, including and excluding HIV, and TB prevalence has to be established. This analysis can aid in the development of policies and initiatives that aim to lower the incidence and mortality rates of TB around the world.

**METHODOLOGY**

Finding the appropriate data source is essential for any examination of a statistical report. We considered a number of websites and choices in choosing a good data set for our research of statistics on tuberculosis around the world. After filtering through different datasets, we ultimately chose to use the information from:

<https://www.kaggle.com/datasets/henajose/determine-the-pattern-of-tuberculosis-spread/code>

for a number of reasons.

Firstly, the website's user-friendly layout makes it simple for us to browse the data and extract the details we need for our study. The content is grouped properly, with distinct subheadings and categories that make it simple to find a particular piece of knowledge.

The website provided a wealth of information on tuberculosis, including prevalence, incidence, mortality, and other relevant characteristics. This makes sure we have a large collection of data to use for our analysis. This included data for over 100 countries, which makes our study vaster.

Overall, it satisfied our criteria for a thorough data source for our examination of statistics on tuberculosis around the world. It is the perfect option for our analysis because of its user-friendly interface, and has large range of data.

**DATA DICTIONARY**

We have provided a data dictionary to help you understand the information presented in our dataset. It contains explanations of all the terms used, making it easier for you to follow our analysis and draw your own conclusions:

1. **NODEXHIV**: Number of deaths due to tuberculosis, excluding HIV.
2. **NODEXHIVSTART**: Minimum number of deaths due to tuberculosis, excluding HIV (start range).
3. **NODEXHIVEND**: Maximum number of deaths due to tuberculosis, excluding HIV (end range).
4. **PREV**: Number of prevalent tuberculosis cases.
5. **PREVSTART**: Minimum number of prevalent tuberculosis cases.
6. **PREVEND**: Maximum number of prevalent tuberculosis cases.
7. **NODINHIV**: Deaths due to tuberculosis among HIV-negative people (per 100000 population).
8. **NODINHIVSTART**: Minimum number of deaths due to tuberculosis among HIV-negative people (per 100000 population).
9. **NODINHIVEND**: Maximum number of deaths due to tuberculosis among HIV-negative people (per 100000 population).
10. **PREVPERPOP**: Prevalence of tuberculosis per 100000 population.
11. **PREVPERPOPSTART**: Minimum prevalence of tuberculosis per 100000 population.
12. **PREVPERPOPEND**: Maximum prevalence of tuberculosis per 100000 population.

**ANALYSIS AND RESULTS**

**-> ANALYSIS OF PREVALENCE OF TUBERCULOSIS (HIV- NEGATIVE AND HIV-POSITIVE) USING HISTOGRAMS:**

1. Method Description:

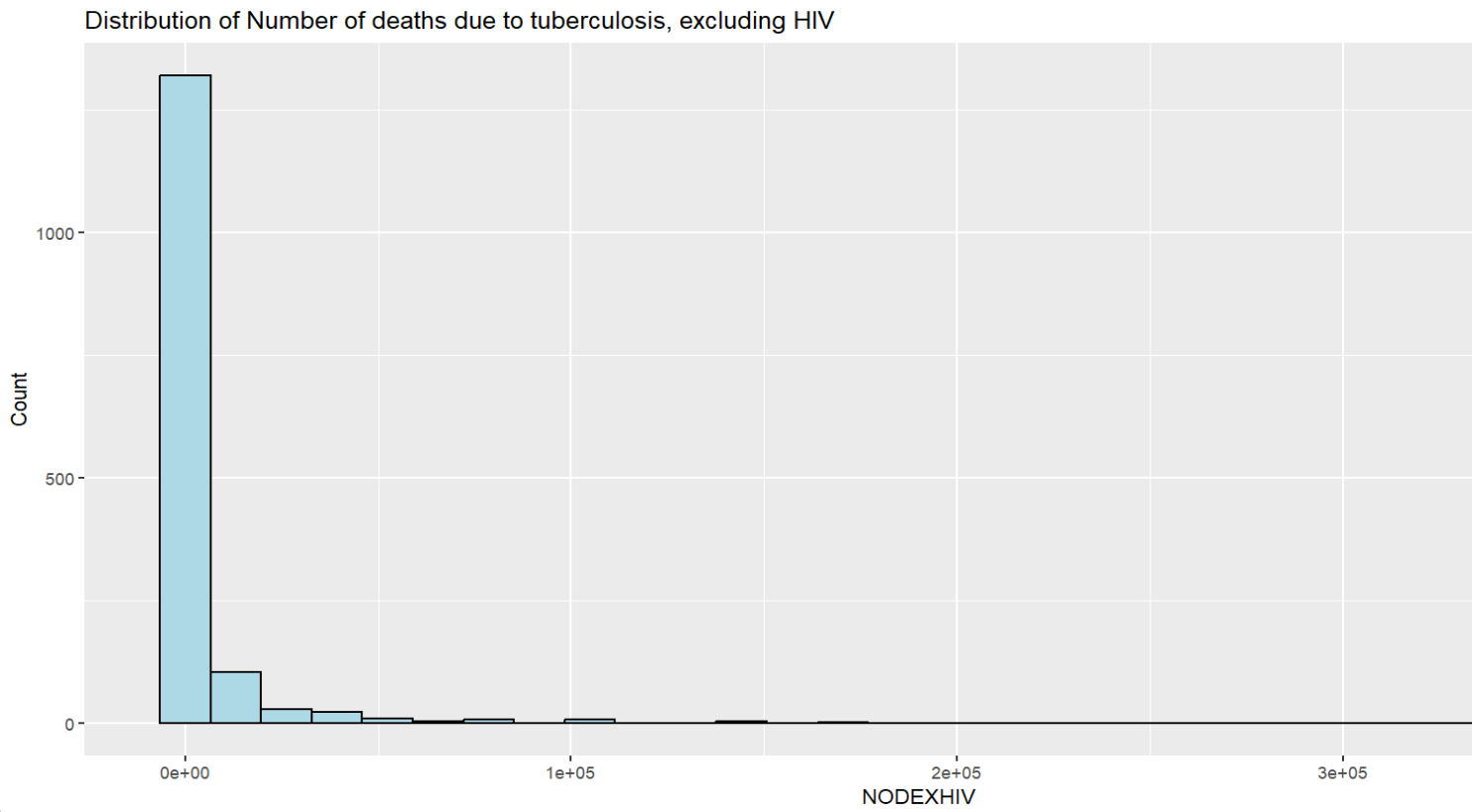
* The plot shows the distribution based off of each parameter (no of deaths, cases inclusive of and exclusive of HIV, prevalence, etc). Each bin on the x-axis represents a range of values of the variable, and the height of each bar represents the count or frequency of observations within that bin.

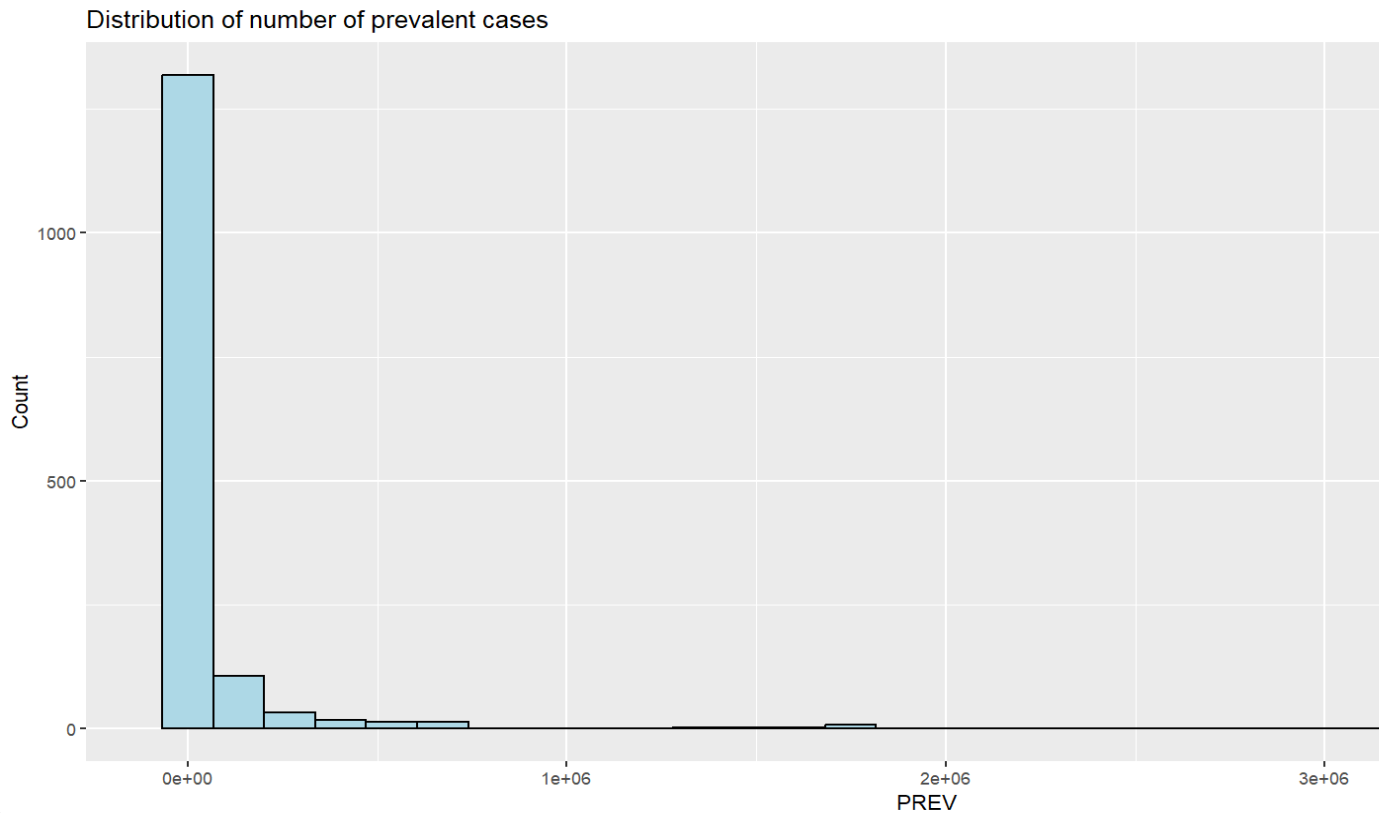
1. PROCESS:

* Collecting the data to determine the range and number of bins: By looking at the range of scores and deciding the number of bins needed to divide the range into.
* Calculating the bin width: By dividing the range of scores by the number of bins to determine the width of each bin. In this case, bin width is 30.
* Creating the histogram: By plotting the histogram using the chosen number of bins and bin width. The x-axis represents the range of scores, divided into bins. The y-axis represents the frequency or count of scores falling into each bin.
* Interpreting the histogram: The histogram visually displays how the exam scores are distributed across different score ranges. It provides insights into the shape of the distribution, the presence of any outliers, and the concentration of scores in specific ranges.

1. RESULT:

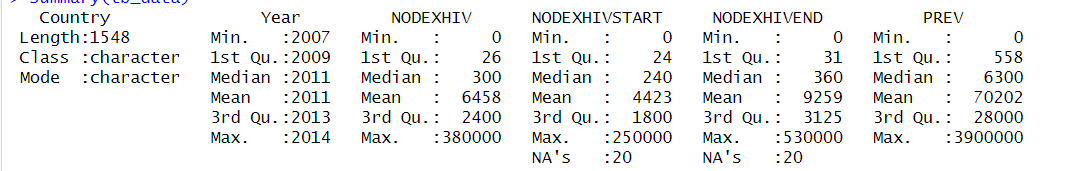
Among the global population overall, crude estimates from 2007-2014 were:

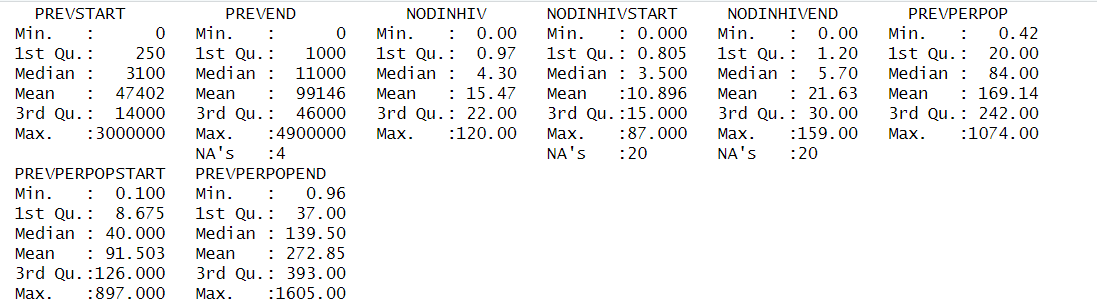


****

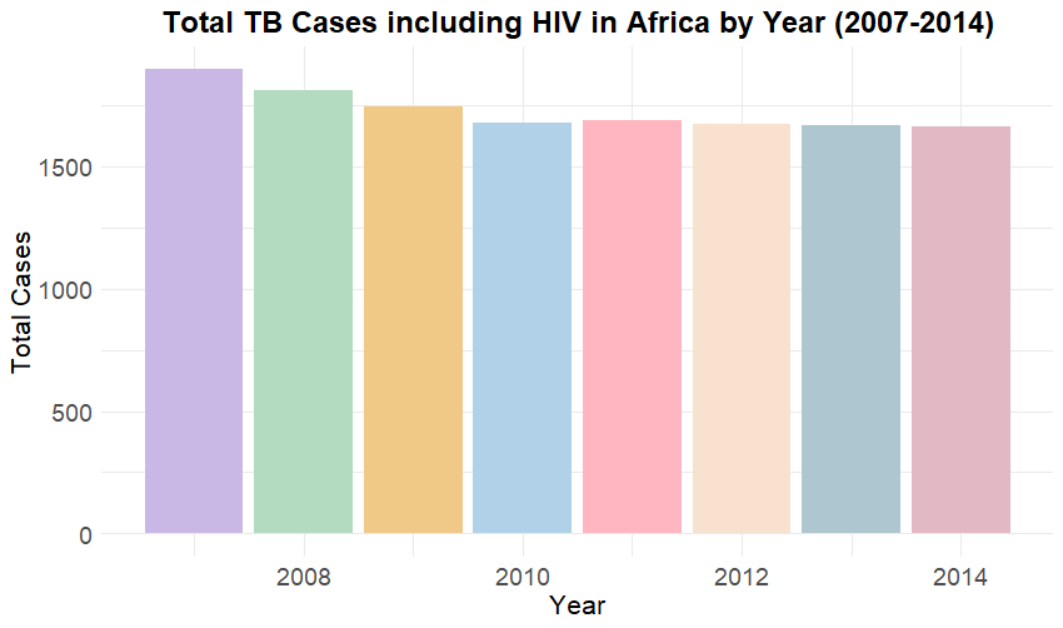
**-> DESCRIPTIVE ANALYSIS:**

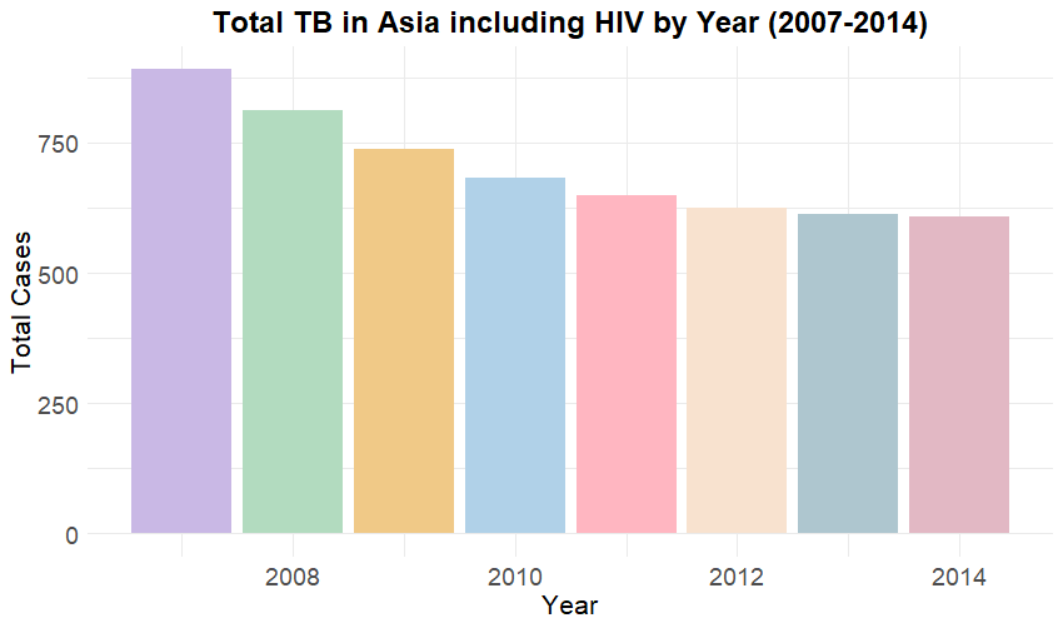
We conducted a descriptive analysis of our TB dataset using various statistical measures and visualization techniques. We first calculated the mean, standard deviation, minimum, and maximum values of the variables in the dataset. In addition, we created sub datasets for all continents by grouping over 100 countries in these continents to delve deeper into our analysis.

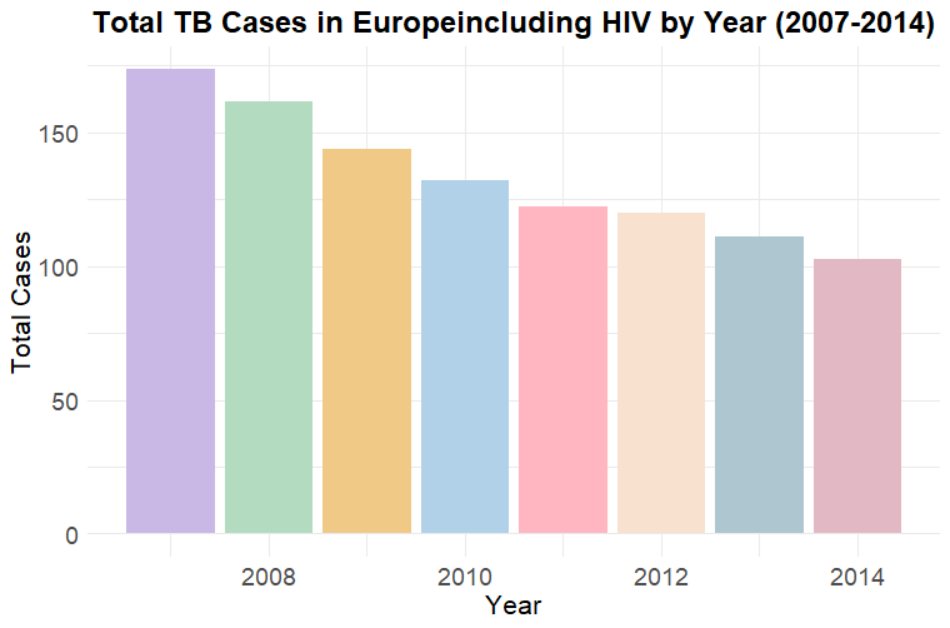


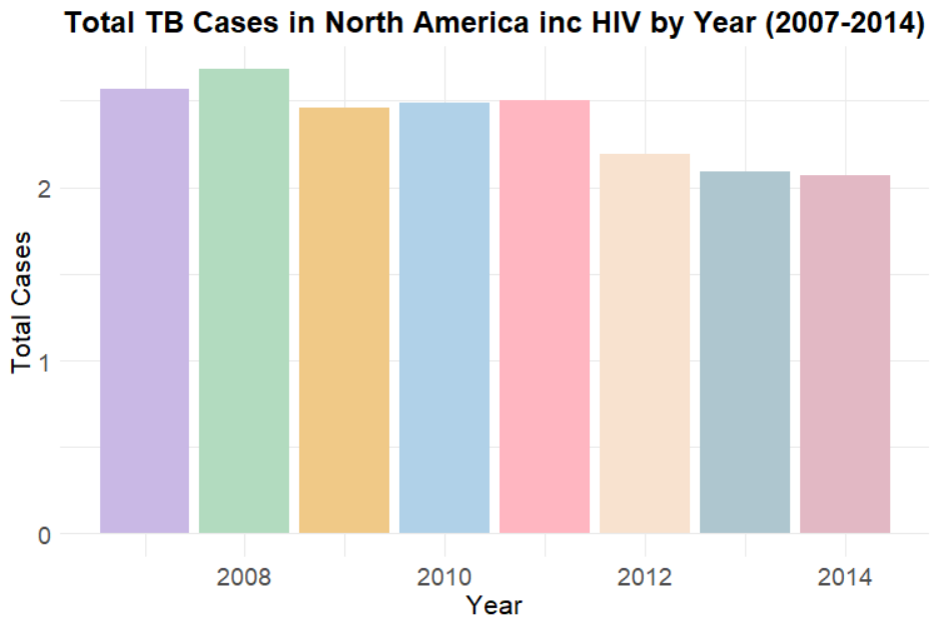


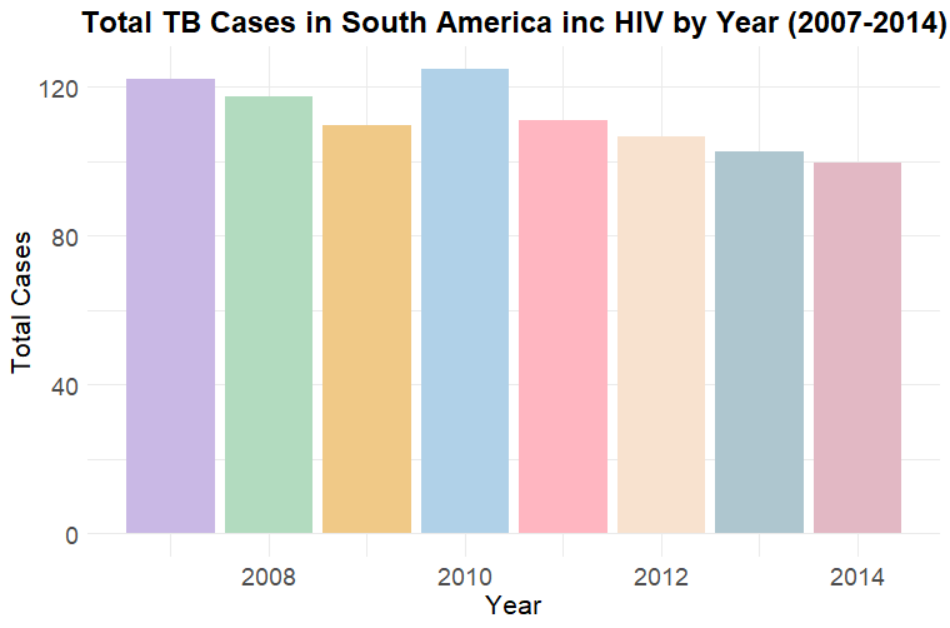
To visualize the data, we plotted a histogram of the total number of TB cases including HIV over the years (2007-2014) for individual continents. Our observations revealed that Africa did not show a drastic decrease in TB cases over the years. Rather, the number of cases remained constant. For continents such as Europe and Asia, we observed a significant decrease in TB cases over the years. However, for North America, South America, and Oceania, there was no uniform pattern that we were able to identify. Instead, the number of cases fluctuated between high and low over the years.

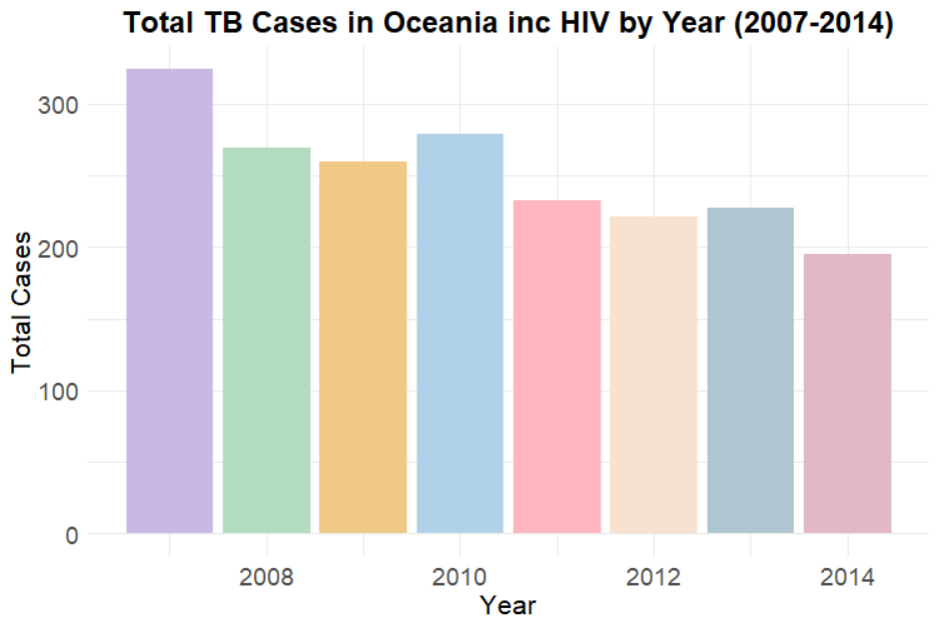






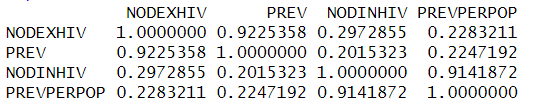






Overall, our analysis provided valuable insights into the trend of TB cases over the years and highlighted the differences in the pattern of TB cases among various continents.

**-> CORRELATION ANALYSIS:**



Based on the correlation matrix**,** we can see that there is a strong positive correlation between NODEXHIV and PREV (correlation coefficient = 0.92), indicating that as the number of TB deaths including HIV increases, so does the prevalence of HIV/TB co-infection.

There is also a moderate positive correlation between NODEXHIV and NODINHIV (correlation coefficient = 0.30), suggesting that as the number of TB deaths including HIV increases, so does the number of TB deaths in general.

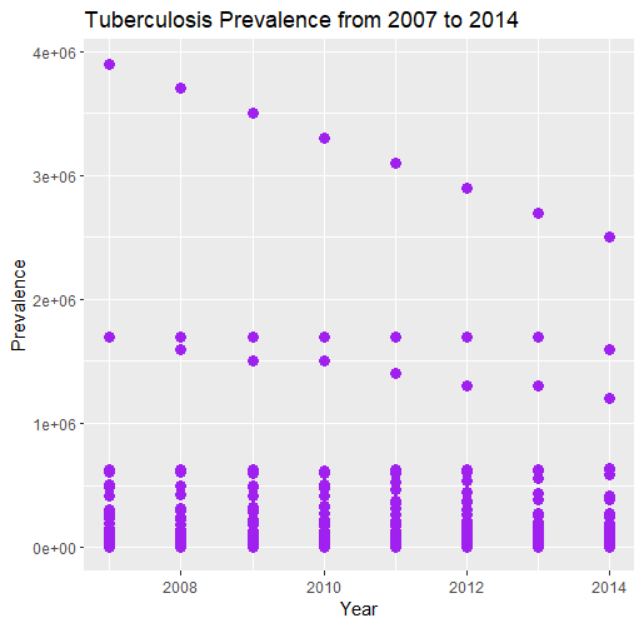
Additionally, there is a strong positive correlation between NODINHIV and PREVPERPOP (correlation coefficient = 0.91), indicating that as the number of TB deaths in general increases, so does the prevalence of TB in the population.

Overall, these correlations suggest that HIV/TB co-infection is strongly related to TB death, and both are related to the overall burden of TB in the population.

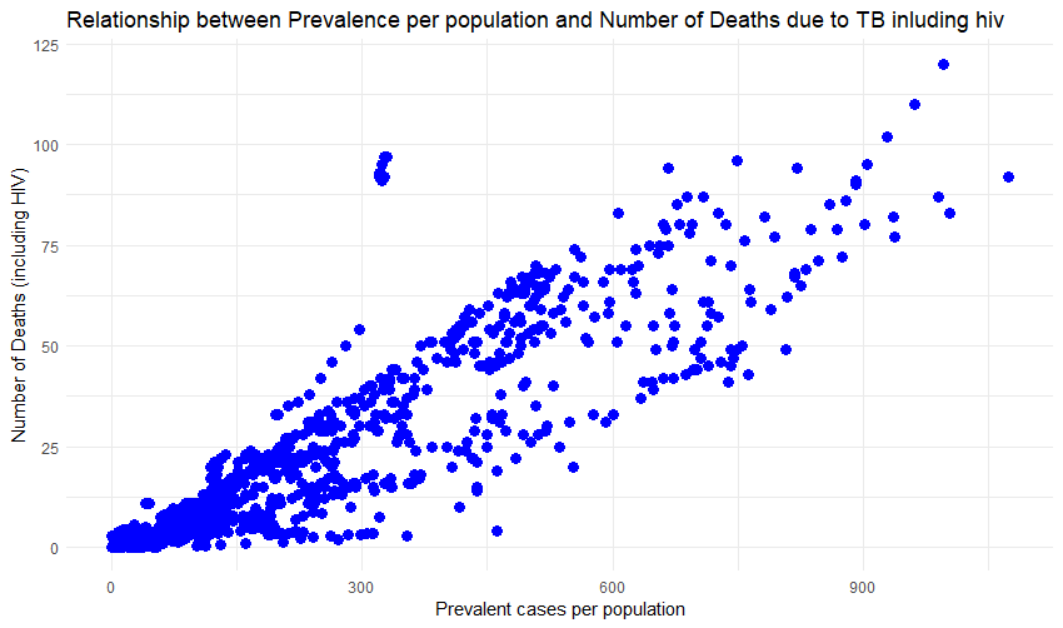
This correlation analysis helps us in determining which features should be selected for predictive analysis, and which are not that statistically significant.

**-> SCATTER ANALYSIS:**

We created a scatter plot to visualize the tuberculosis prevalence from 2007 to 2014, and the plot helped us to identify some interesting trends. We observed that some countries had much higher prevalence rates than others, which suggests that targeted resources and interventions may be needed in these regions to control the spread of the disease. Additionally, the scatter plot revealed that there was a significant drop in tuberculosis prevalence overall which could be attributed to improvements in screening, diagnosis, and treatment programs. These variations could be due to a range of factors, such as changes in healthcare policies and infrastructure, as well as improvements in disease surveillance.



We created another scatter plot to test the relation between prevalence per pop and HIV included TB death cases.



In summary, our visual analysis confirms the positive correlation between prevalence per population and number of deaths due to TB including HIV. The scatter plot shows a clear linear pattern, indicating that a linear regression model can be a good fit for these variables. This information could be useful in predicting the number of deaths due to TB including HIV for a given prevalence per population value or vice versa. Overall, this analysis reinforces the importance of addressing the prevalence of TB in populations to reduce the number of deaths caused by this disease.

**-> TIME SERIES ANALYSIS:**

During our analysis, we decided to delve deeper into the prevalence category by creating interactive line plots for all continents. These plots allowed us to observe the overall prevalence trends for all countries in a particular continent with different colours.

1. Interactive Line Plots:

* Interactive line plots were created to visualize the prevalence trends for all countries in each continent. Different colors were used to differentiate between countries.
* The line plots allowed for a comprehensive observation of the overall prevalence trends across countries within a continent.

1. Method:

* Data Preparation:

The prevalence data for TB cases in each country within the continents was collected and organized.

The data was structured in a format suitable for creating interactive line plots.

* Interactive Line Plot Creation:

Interactive line plots were generated using appropriate visualization tools.

Each line plot represented the prevalence trend for a specific country, with different colors used to distinguish between countries within a continent.

1. Process:

* Data Analysis:

The prevalence trends in each continent were analyzed by examining the interactive line plots.

Key observations regarding the prevalence trends of different countries within each continent were recorded.

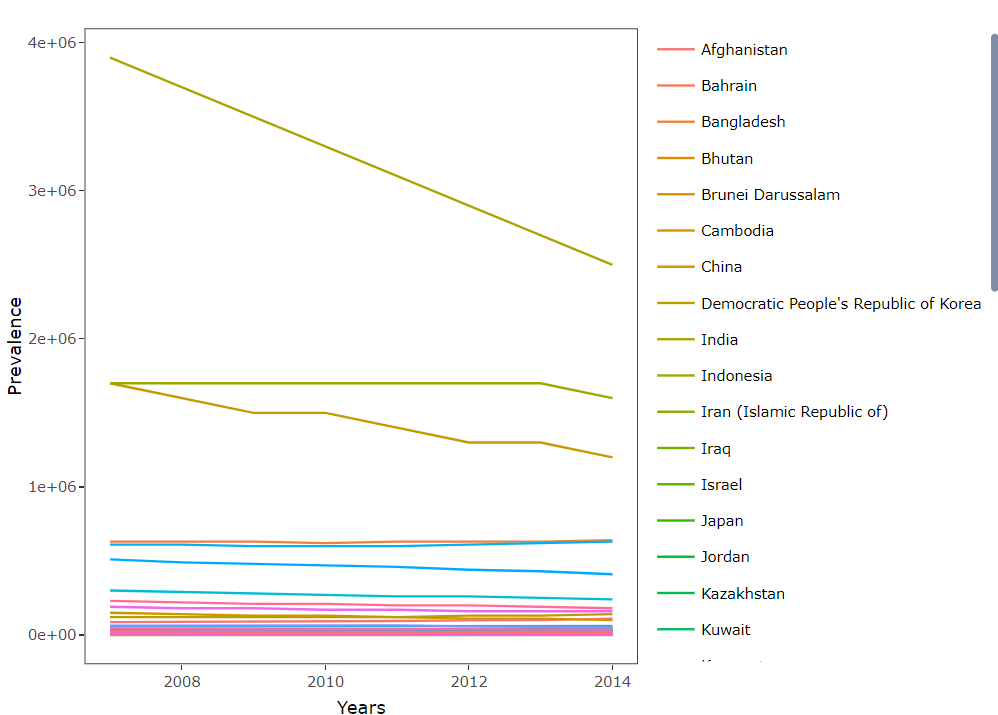
* Comparison and Interpretation:

A comparison was made between countries within each continent to identify prominent trends.

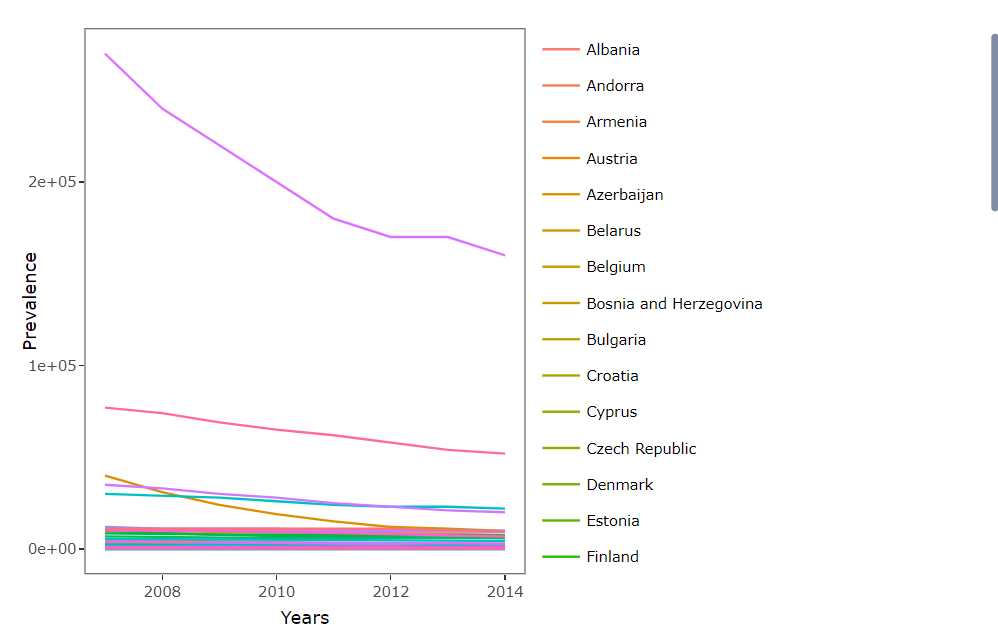
The observations from the line plots were interpreted to understand the variations in TB prevalence among different countries and continents.

* Conclusion: The analysis of TB prevalence trends revealed the following findings.

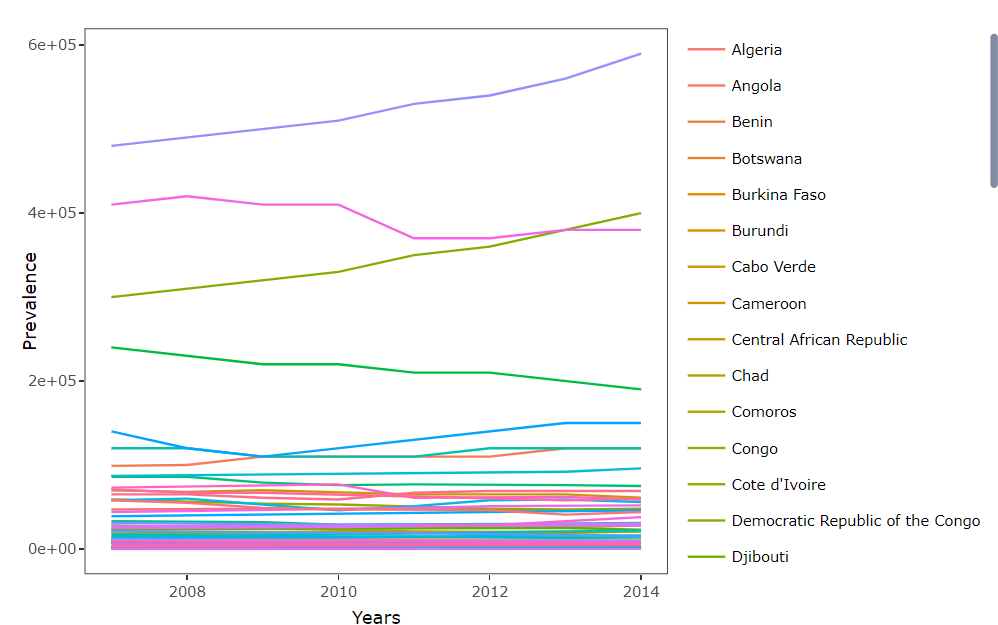
1. Prevalence Trends by Continent:



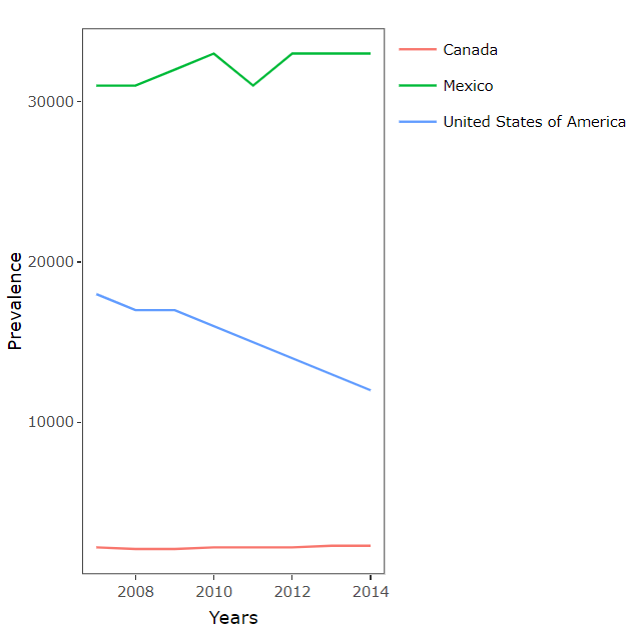
* ASIA: In Asia, we observed that India had the highest prevalence of TB cases compared to other countries in the continent. Moreover, the decrease in prevalent TB cases in India was significantly high. We also noted a similar pattern in China, although the scale was relatively lower than India.



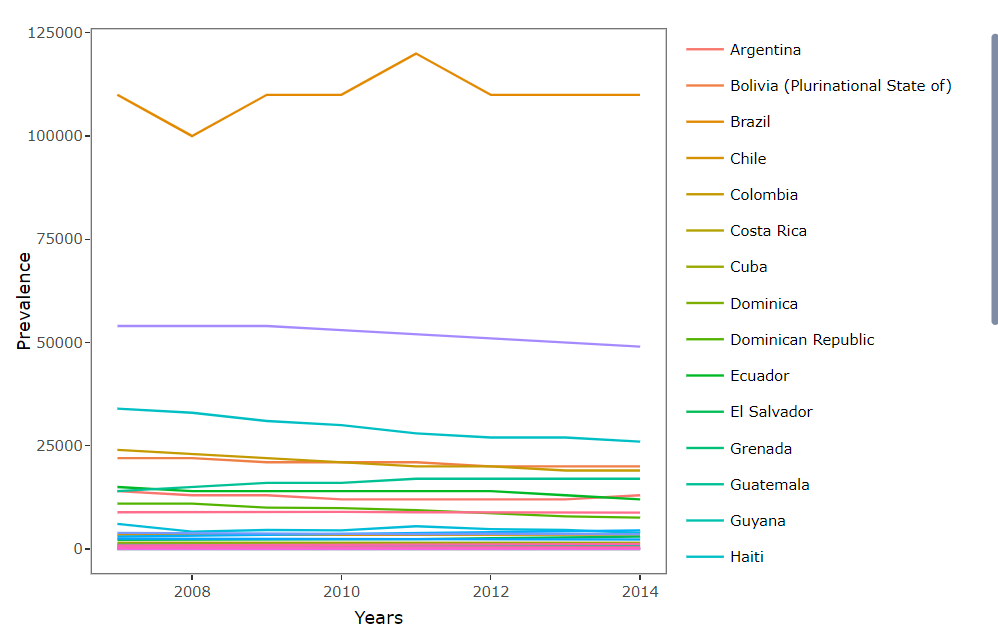
* EUROPE: For Europe, we found that the Russian Federation and Ukraine showed different prevalence trends compared to the rest of the countries. These countries exhibited a significant decrease in TB prevalence.



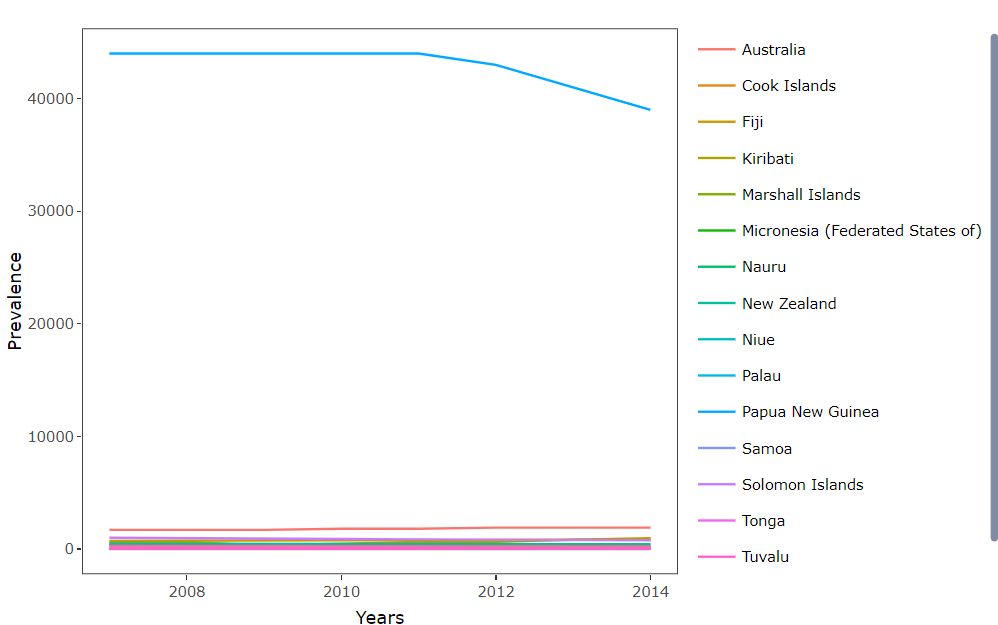
* AFRICA: In Africa, we observed a shocking increase in TB prevalence in Nigeria and the Democratic Republic of The Congo. Overall, compared to other continents, Africa showed a higher increase in TB prevalence.



* NORTH AMAERICA: Moving on to North America, we observed a fluctuating trend in Mexico, a significant decrease in the USA, and a constant prevalence trend in Canada.



* SOUTH AMERICA: For South America, almost all countries showed similar constant prevalence trends, except Brazil, which exhibited high fluctuations in their TB prevalence cases, later settling into a constant trend.



* OCEANIA: Lastly, in Oceania, Papua New Guinea showed a constant prevalence, which later deteriorated, while the other countries showed almost no change in their TB prevalence at all.

**-> PREDICTIVE ANALYSIS:**

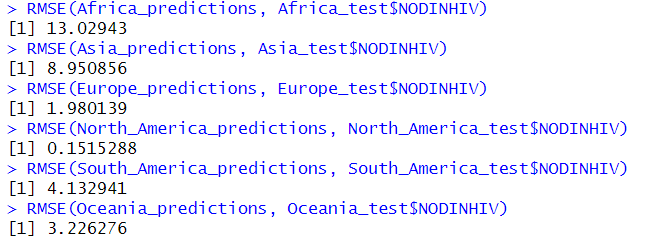
* 1. Analysis:

The goal of the analysis was to create a model that accurately predicts the number of deaths due to TB cases including HIV. The predictors used in the model were the prevalence and prevalence per population since these two features are statistically significant as observed from the correlation analysis

* 1. Success Rate:

The success rate was evaluated by calculating the Root Mean Squared Error (RMSE) for the predictions made by the models. The RMSE provides a measure of the average prediction error.

* 1. Methods:
* Correlation analysis: This analysis revealed that the prevalence and prevalence per population were statistically significant predictors for TB cases including HIV.
* Regression models: Regression analysis was performed on the testing data for all continents to predict the number of deaths. Separate models were created for each continent, and a global model was also developed using the entire dataset.
* RMSE calculation: The RMSE values were computed for each model to assess their accuracy in predicting the number of TB cases with HIV.
* After applying regression on our testing data for all continents, we calculated the accuracy score for individual models for different continents and then for the whole dataset as a whole. We are trying to see if splitting the data continent wise increased our chances of getting a more accurate death predicting model or not.



* 1. Analysis:
* North America: The North American model had the lowest error with an RMSE of 0.1515, indicating high accuracy in predicting TB cases with HIV in North America.
* Europe: The European model had a relatively low error with an RMSE of 1.9801, suggesting good predictive performance.
* Africa: The African model had the highest error with an RMSE of 13.0294, indicating lower accuracy in predicting TB cases with HIV in Africa.
* Other continents: Asia had an RMSE of 8.9509, South America had an RMSE of 4.1329, and Oceania had an RMSE of 3.2263. These continents had moderate errors compared to North America and Europe.
* Global model: The global model had an RMSE of 0.3006352, indicating reasonably good overall performance. However, its RMSE value was relatively higher compared to the individual continent models, suggesting that it may be less accurate in predicting TB prevalence for specific regions.



* 1. Conclusions:
* The analysis suggests that splitting the data continent-wise improved the accuracy of the predictive models for certain regions. The North American and European models performed well, while the African model showed lower accuracy. The other continents had moderate accuracy. The global model performed reasonably well overall but was less accurate than the individual continent models for specific regions
* From correlation analysis we observed that for TB cases including HIV, prevalence and prevalence per population were statistically significant, we will use these 2 features as the predictors, and we tried to create a model that can predict number of deaths accurately.

**CONCLUSION**

In conclusion, our analysis provided valuable insights into the trend of tuberculosis (TB) cases over the years and highlighted the differences in the pattern of TB cases among various continents. We conducted a descriptive analysis of our TB dataset using various statistical measures and visualization techniques. We also performed a correlation analysis, which suggests that HIV/TB co-infection is strongly related to TB death, and both are related to the overall burden of TB in the population. Our visual analysis confirmed the positive correlation between prevalence per population and the number of deaths due to TB, including HIV. Furthermore, during our time series analysis, we observed variations in TB prevalence among different continents and even countries within a continent. Finally, we used predictive analysis to create a model that can predict the number of TB deaths accurately, and we found that the North American model had the lowest error, indicating that it was the most accurate in predicting the number of TB cases with HIV in North America. Overall, our analysis could be useful in developing targeted resources and interventions to control the spread of TB and reduce the number of deaths caused by this disease.

**REFERENCES**

We have taken the help of the sample report for majority of this report, the in-depth study of certain topics was assisted with the help of the sites provided below:

1. <https://ggplot2.tidyverse.org/index.html>
2. <https://rdrr.io/cran/caret/man/createDataPartition.html>
3. <https://statisticsglobe.com/correlation-matrix-in-r>
4. <https://www.geeksforgeeks.org/scatter-plots-in-r-language/>

**CONTRIBUTION**

* **Sanya Goel**: Coding and Analysis
* **Jasmine Sindhu**: Analysis and Presentation
* **Vidhika Mangla**: Data Collection and Detailed Analysis
* **Rohan Dhingra**: Presentation and Compiling