Aadith Joseph Mathew

3 MSc DS B

23122101

Analysing Factors Affecting Life Expectancy: Multiple Linear Regression Approach

Introduction:

Life expectancy is a crucial indicator of a nation's overall health and well-being. This study aims to analyse the factors affecting life expectancy by considering immunization, human development index, and other relevant variables. The motivations for this research are twofold: first, previous studies have not adequately accounted for the impact of immunization and human development index, and second, past research has primarily relied on single-year data and multiple linear regression models. By employing a mixed-effects model and multiple linear regression analysis on data spanning from 2000 to 2015, this study seeks to provide a more comprehensive understanding of the determinants of life expectancy across countries.

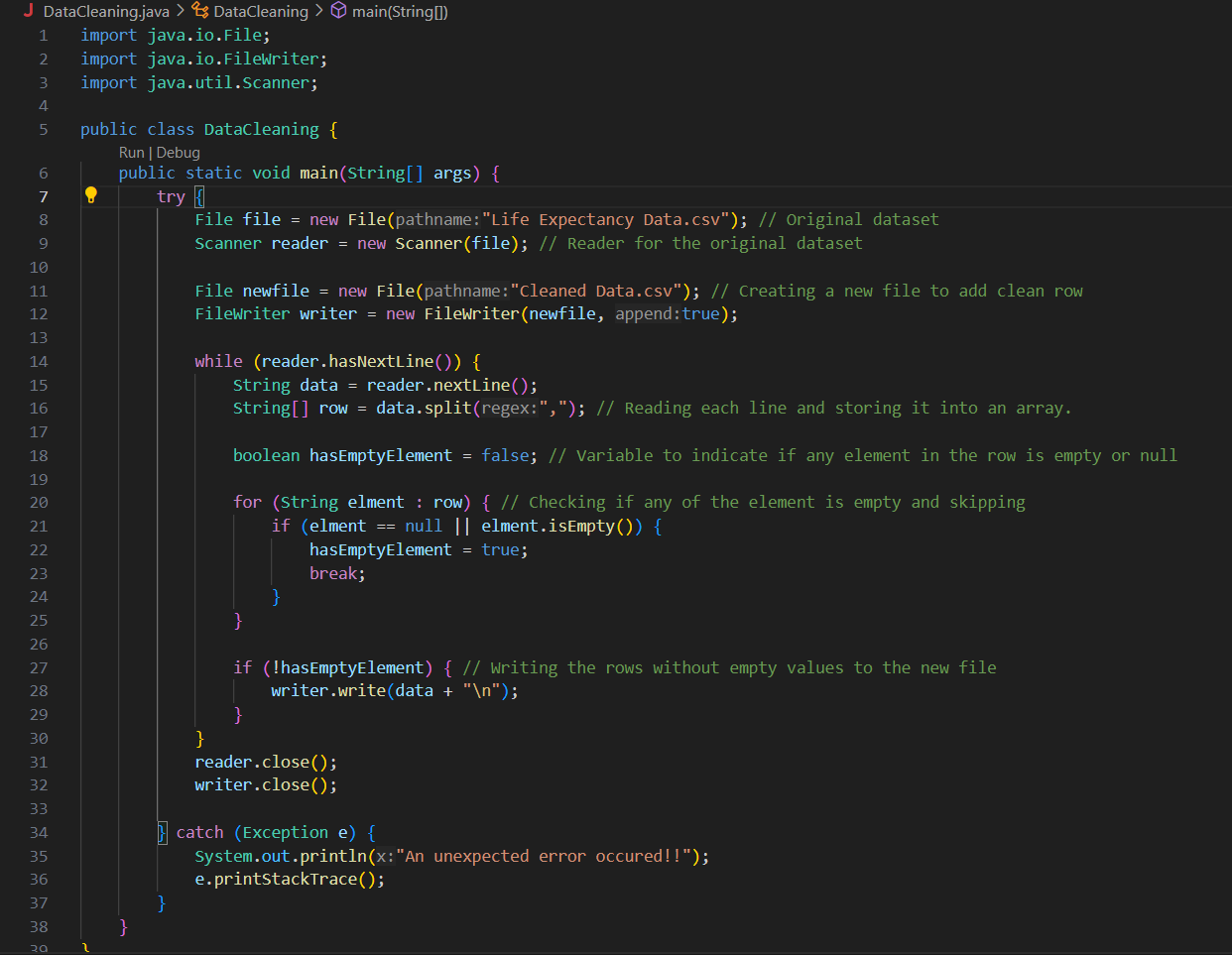
Dataset Description:

The dataset used in this study is obtained from the Global Health Observatory (GHO) data repository under the World Health Organization (WHO) and the United Nations website. It comprises data on life expectancy, health factors, and economic indicators for 193 countries from 2000 to 2015.

The dataset includes 22 columns and 2938 rows, representing 20 predictor variables. These variables are categorized into four broad groups: immunization-related factors, mortality factors, economic factors, and social factors. The dataset has undergone preprocessing, including handling missing values and excluding countries with significant data gaps.

Data Cleaning:

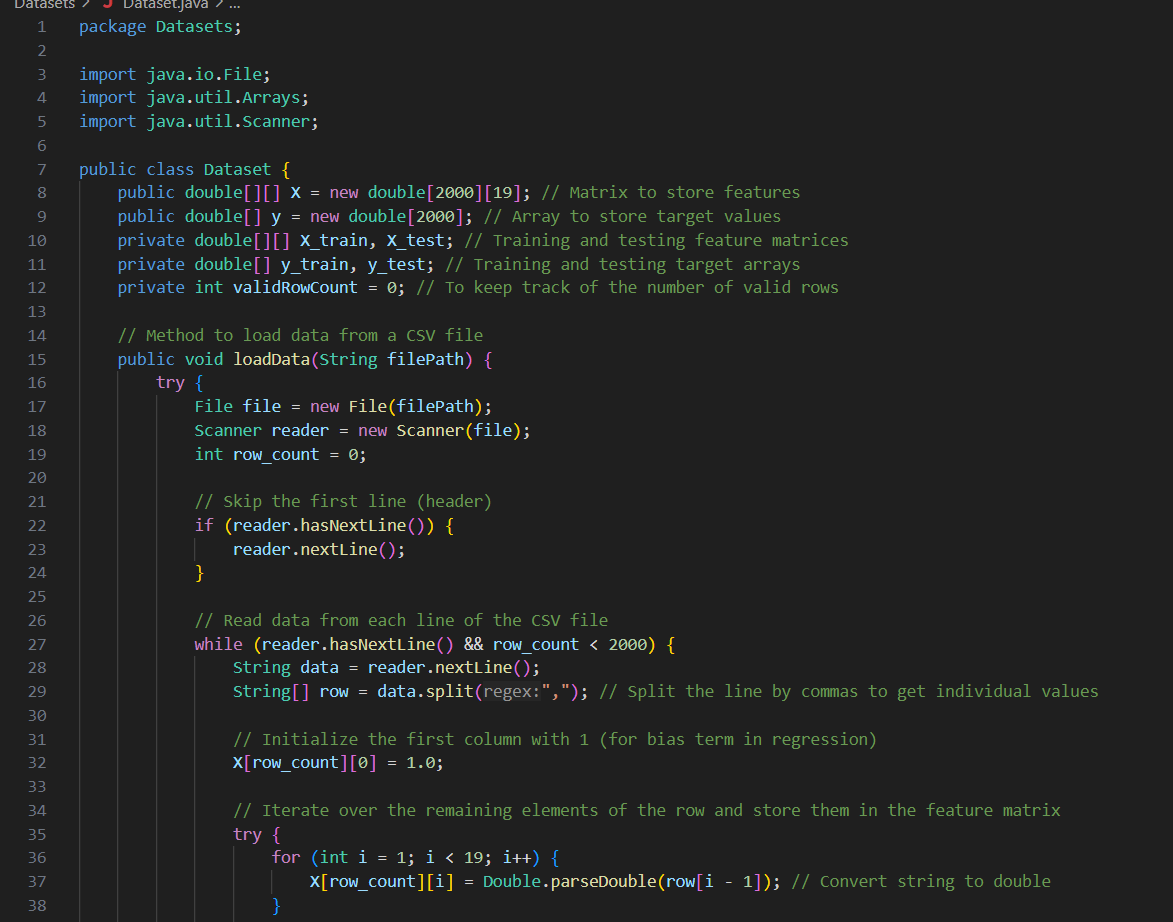
This section covers the data cleaning process, which is an essential step in preparing the dataset for analysis. The rows with missing values are removed in this section of code. Cleaned data is written into another file and the same file is used for model building and evaluation.

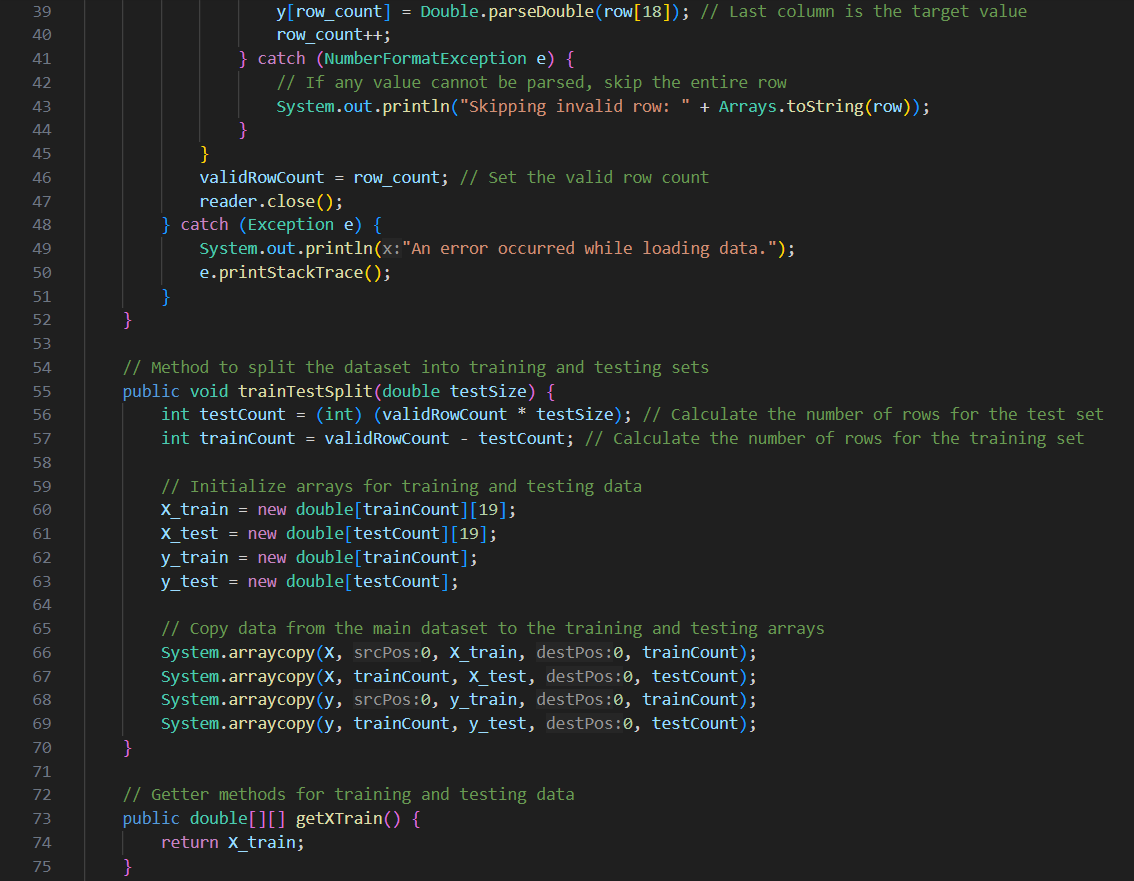


Dataset Loading:

These code snippets demonstrate the process of loading the dataset into the analysis environment. The dataset was loaded from ‘Cleaned Data.csv’ file, and the necessary libraries were imported for data manipulation and analysis.

Data is loaded into a 2D array of size 2000,19 where each column represents a feature in the dataset. Calculations are performed on this loaded array.



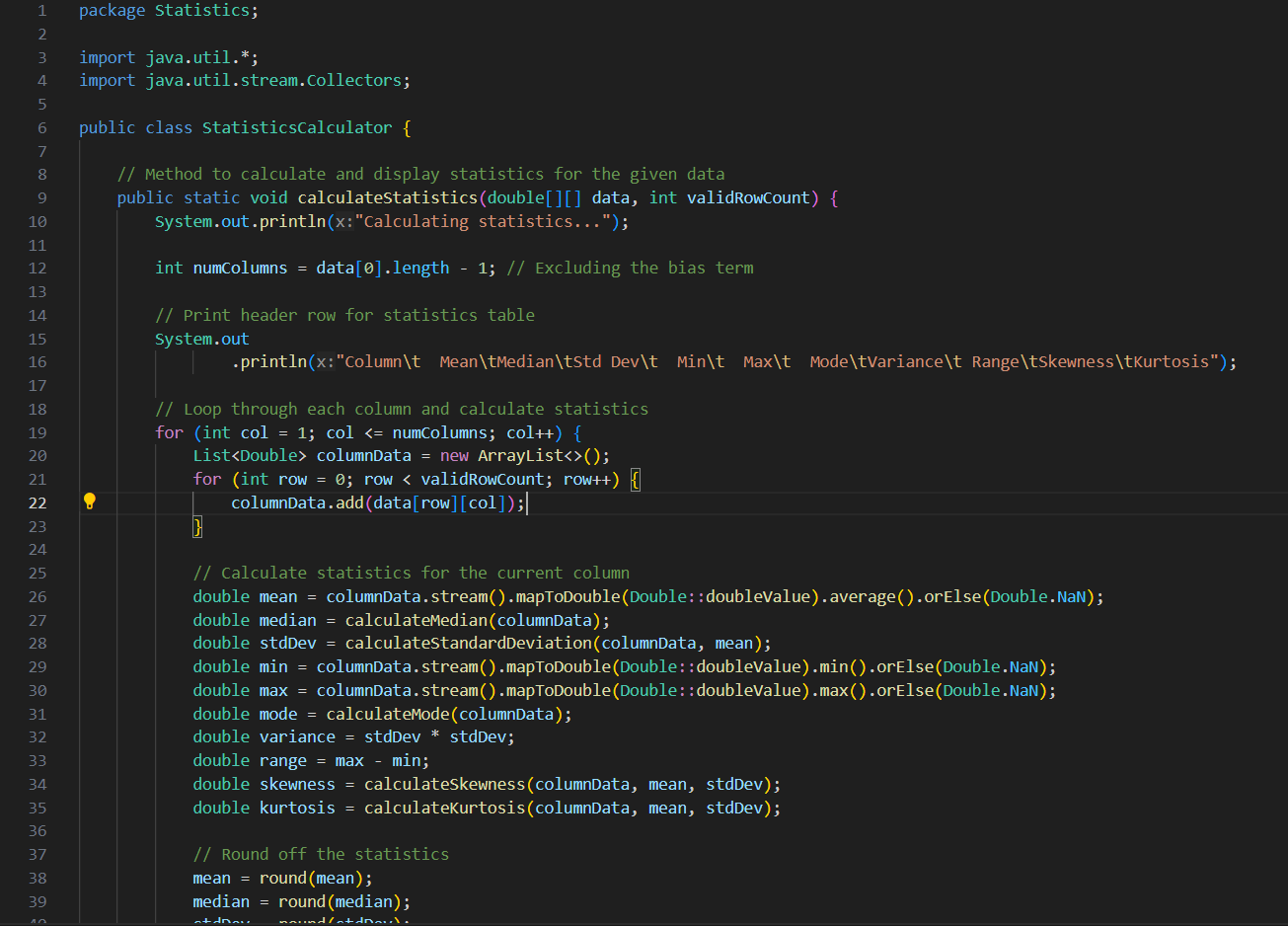


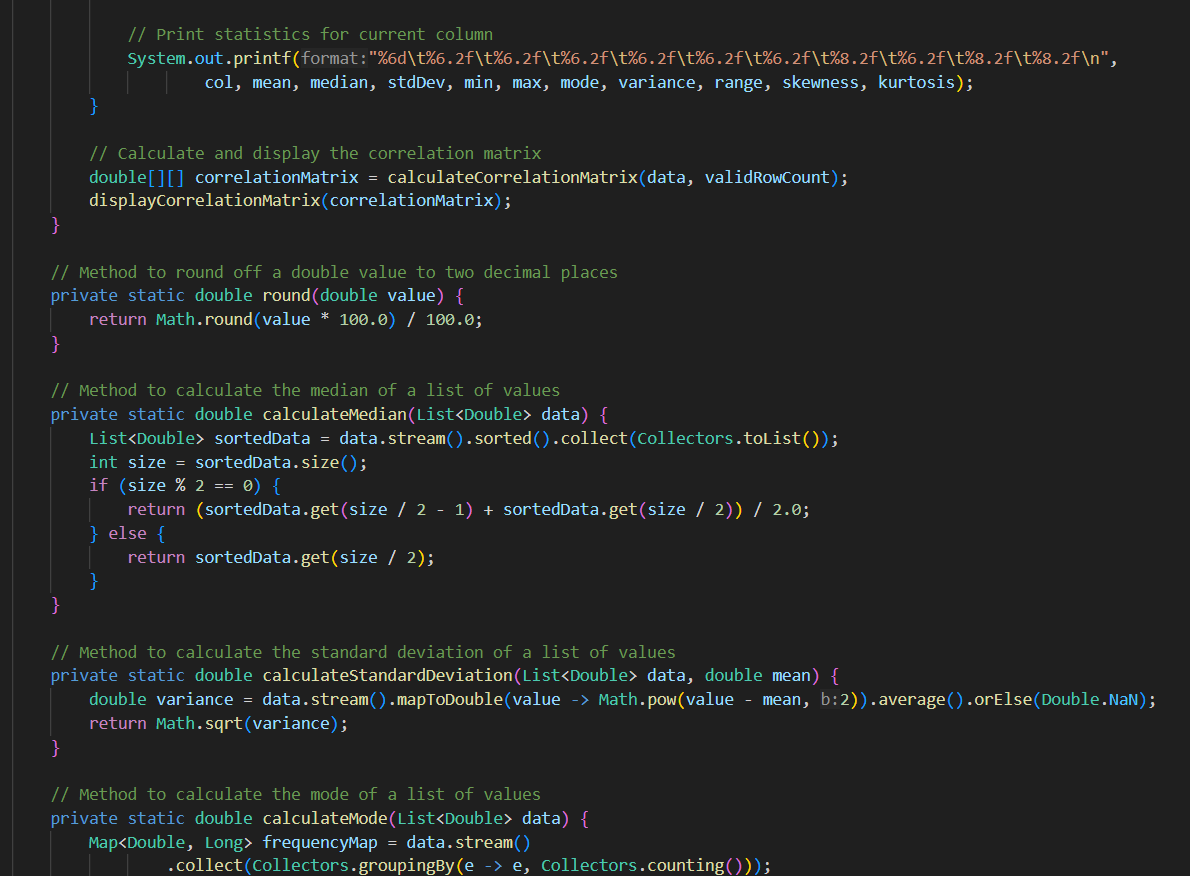
Dataset Statistics:

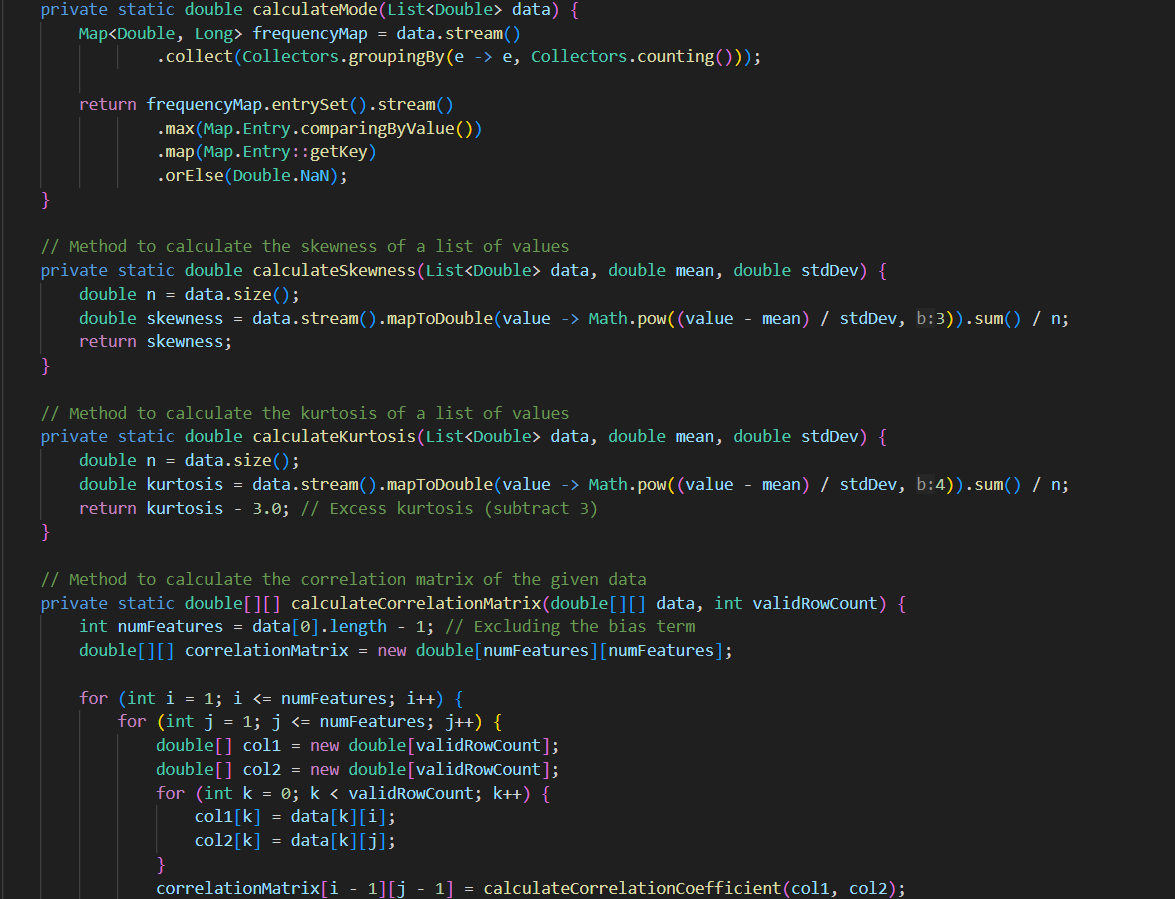
The descriptive statistics section provides an overview of the dataset, including summaries of the predictor variables and their distributions. The statistics include:

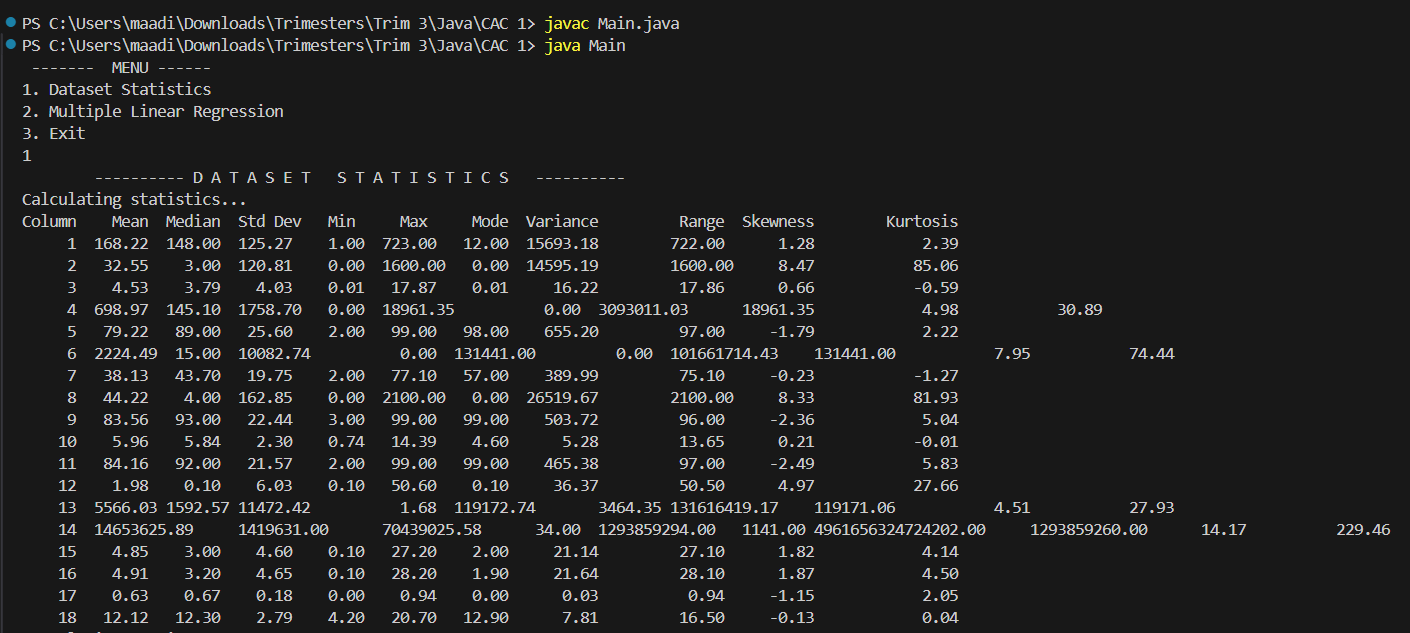
* Mean
* Median
* Standard Deviation
* Minimum
* Maximum
* Mode
* Variance
* Range
* Skewness
* Kurtosis

The correlation matrix is also displayed below.

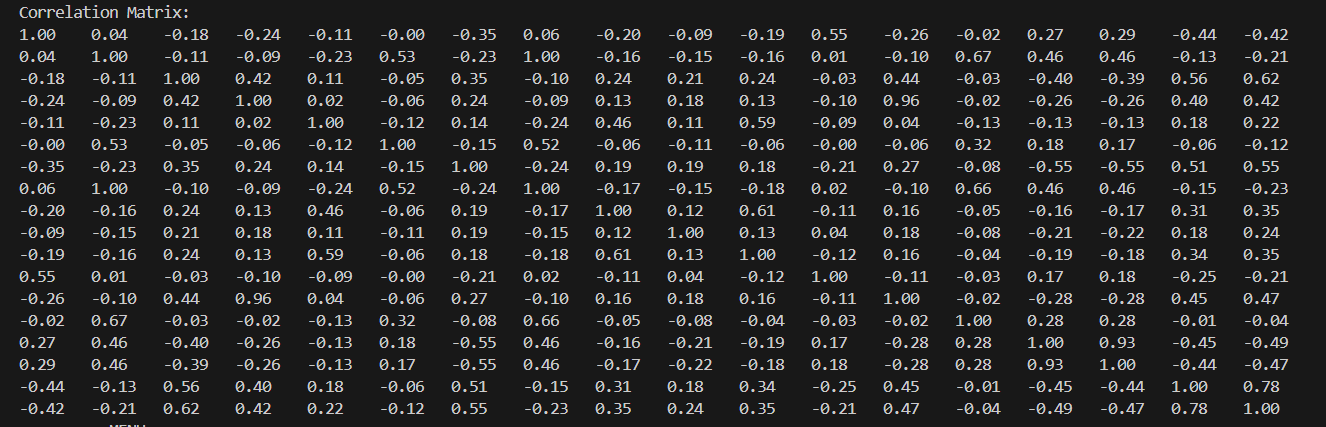










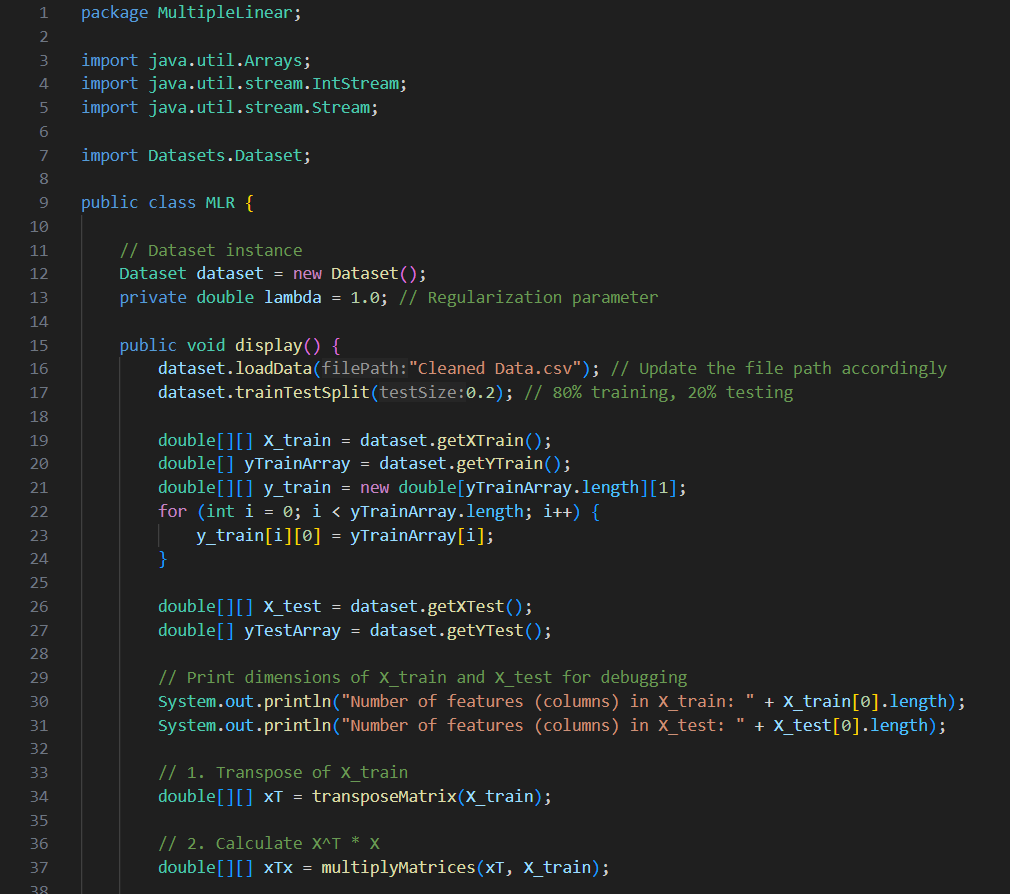


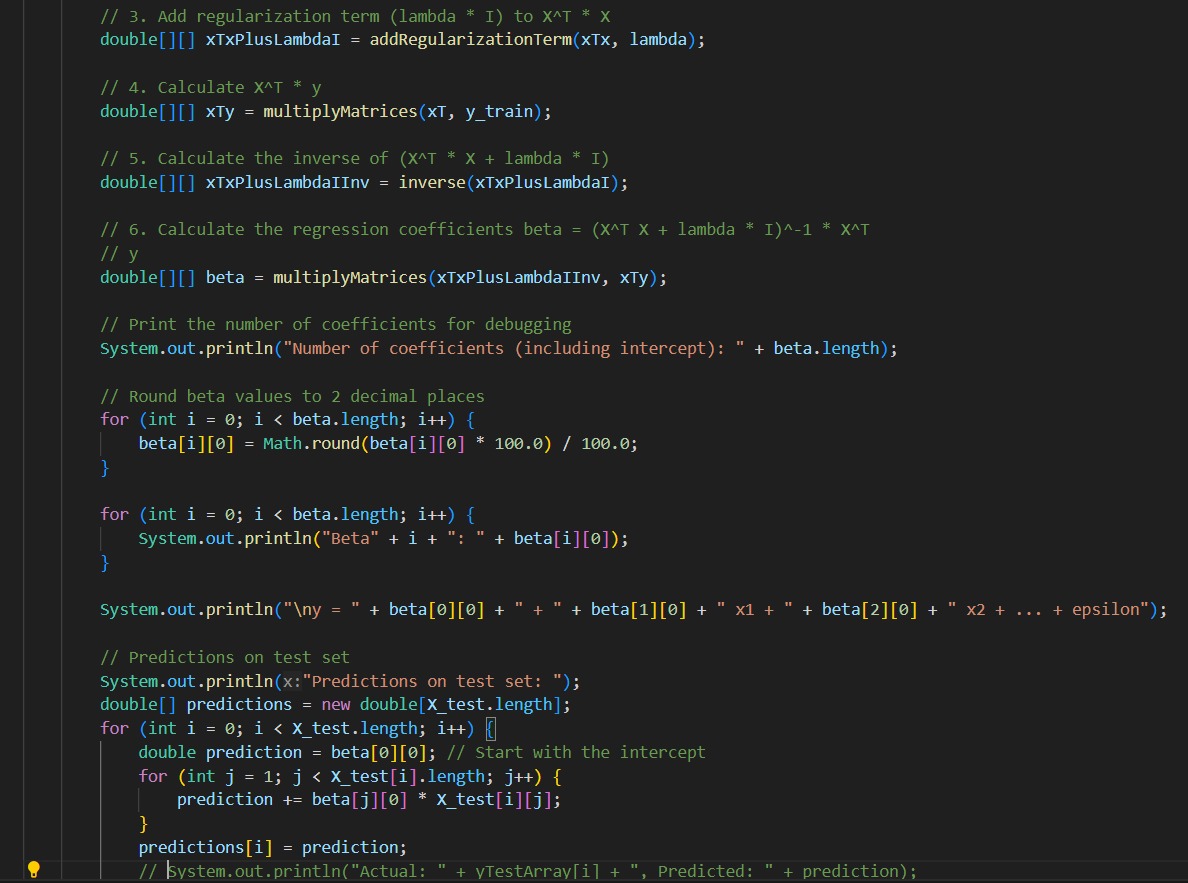
Model Building:

This section focuses on building the multiple linear regression model. The code splits the dataset into training and testing sets, which is a common practice in machine learning to evaluate the model's performance on unseen data.

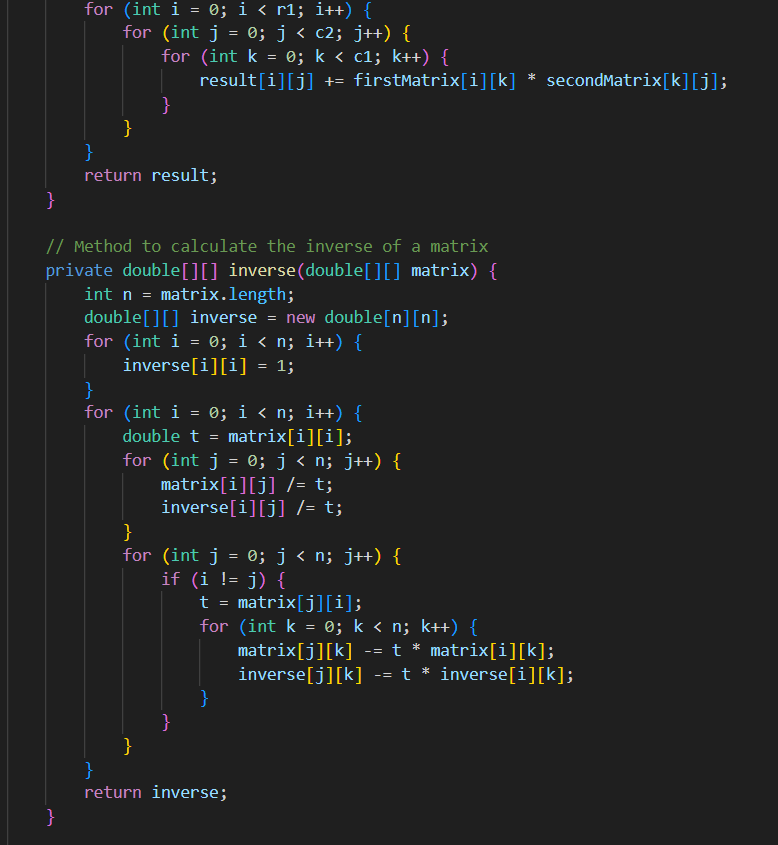
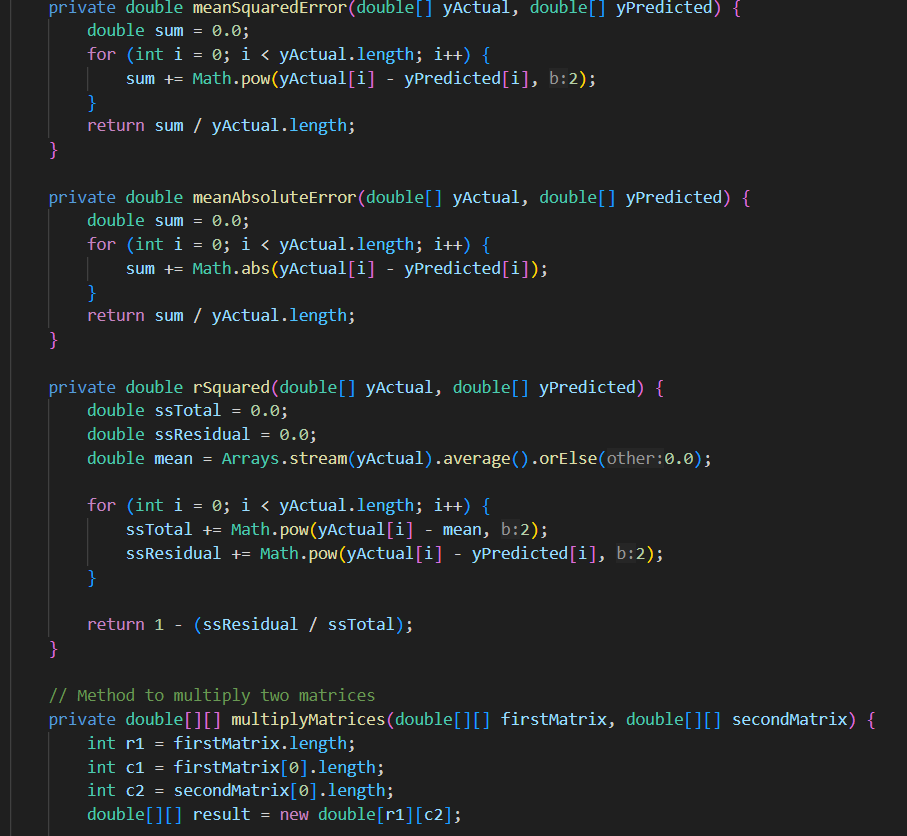
The next step is feature scaling, where the features are scaled or normalized using the Z-score conversion.

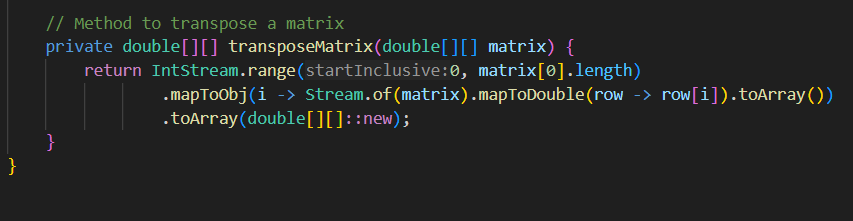
Once the features are selected, the multiple linear regression model is trained on the training data. The code shows the usage of ordinary least squares (OLS) or regularization methods (e.g., Ridge, Lasso) to prevent overfitting.

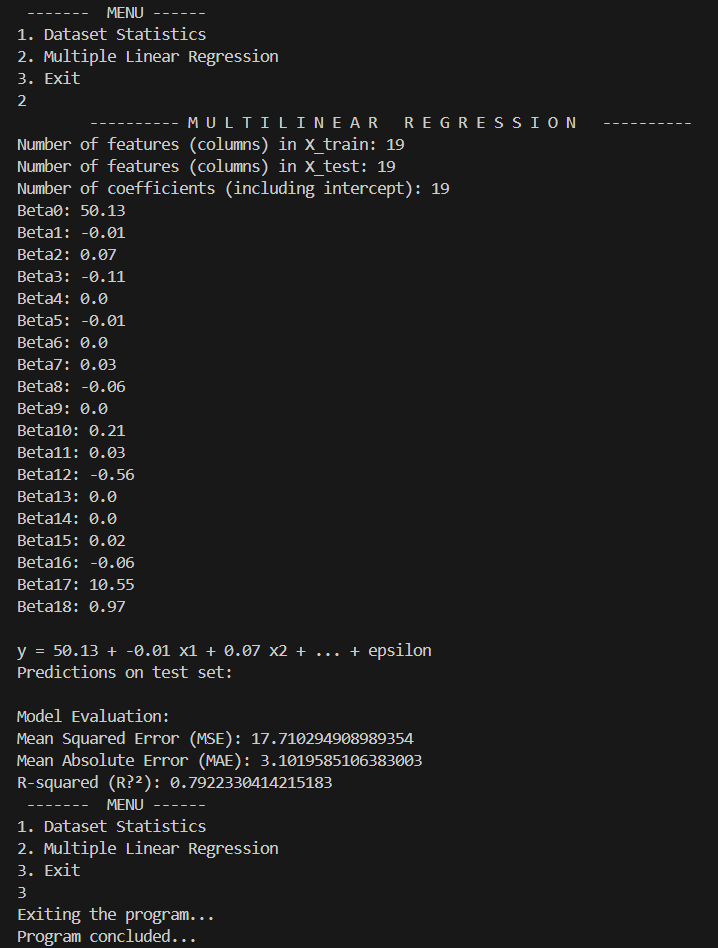












After training the model, it is evaluated using various metrics to assess its performance. The code calculates metrics such as mean squared error (MSE), root mean squared error (RMSE), and R-squared. These metrics provide insights into the model's accuracy, predictive power, and goodness of fit.

**Mean Square Error : 17.71**

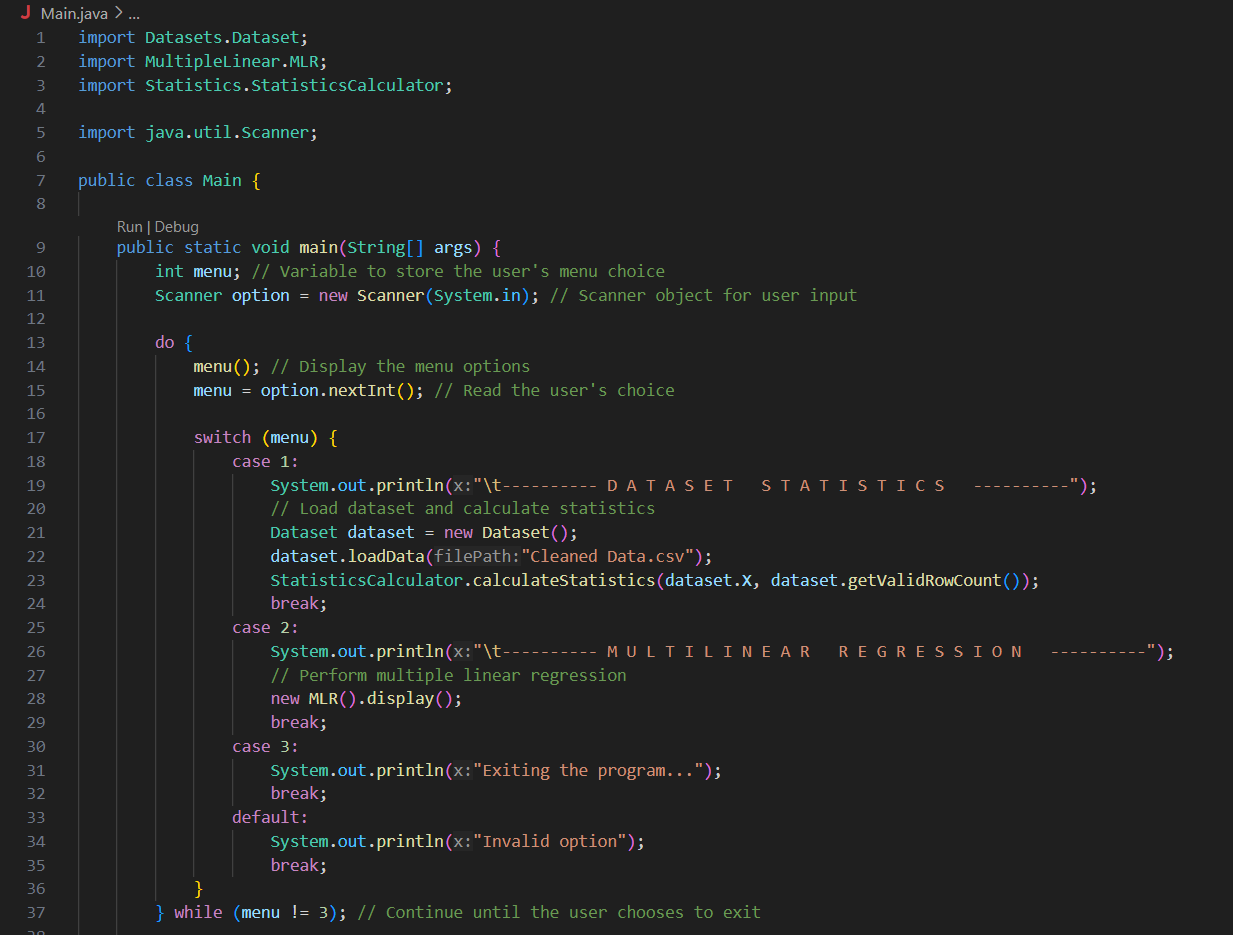
**Mean Absolute Error : 3.10**

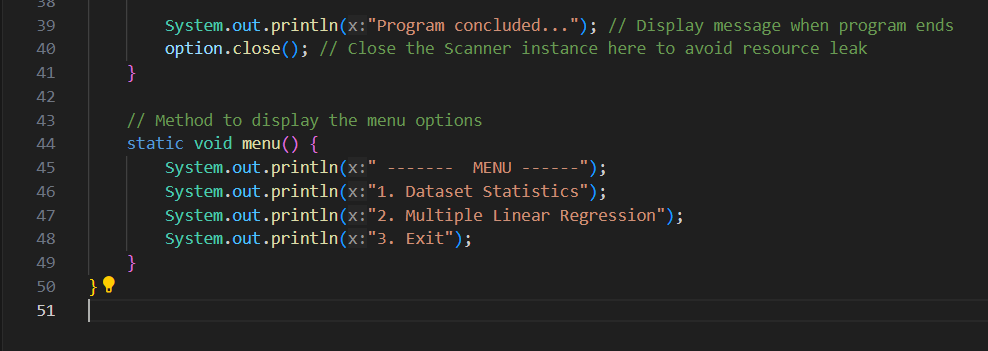
**R-squared : 0.7922**

Main Class

This section contains the main program or script that orchestrates the entire analysis pipeline. It imports the necessary modules or functions and executes the data loading, cleaning, exploratory data analysis, model building, and model evaluation steps in a structured manner.

The main class is responsible for tying all the components together and ensuring that the analysis workflow runs smoothly from start to finish.





Conclusion:

The analysis of factors affecting life expectancy across countries using a mixed-effects model and multiple linear regression has provided valuable insights into the determinants of this crucial health indicator. By considering immunization rates, human development index, and various other predictor variables, this study offers a comprehensive understanding of the complex interplay between different factors and their impact on life expectancy.

The multiple linear regression model employed in this study provided a robust framework for quantifying the relationships between these various factors and life expectancy. The model evaluation metrics, such as mean squared error, root mean squared error, and R-squared, indicated a good fit and predictive power, lending credibility to the findings.

It is important to note that while this study offers valuable insights, there are limitations and opportunities for further research. Future studies could explore additional factors, such as environmental conditions, healthcare infrastructure, and cultural influences, which may also impact life expectancy. Additionally, longitudinal analyses and more granular data at the regional or local levels could provide a deeper understanding of the dynamics within countries.