Project Report: Life Expectancy Analysis

1. Project Overview

This project aims to analyze various factors influencing life expectancy and build a predictive model to estimate it. We will explore relationships between different socio-economic and health-related features with life expectancy using a comprehensive dataset.

2. Dataset Preparation and Exploration

The dataset, after being read into a Pandas DataFrame (df), has a shape of (2938, 22), indicating 2,938 rows (observations) and 22 columns (features).

The initial check of data types revealed a mix of numerical (float64, int64) and categorical (object) columns. Several key predictor columns contained non-null counts lower than 2938, indicating the presence of missing values.

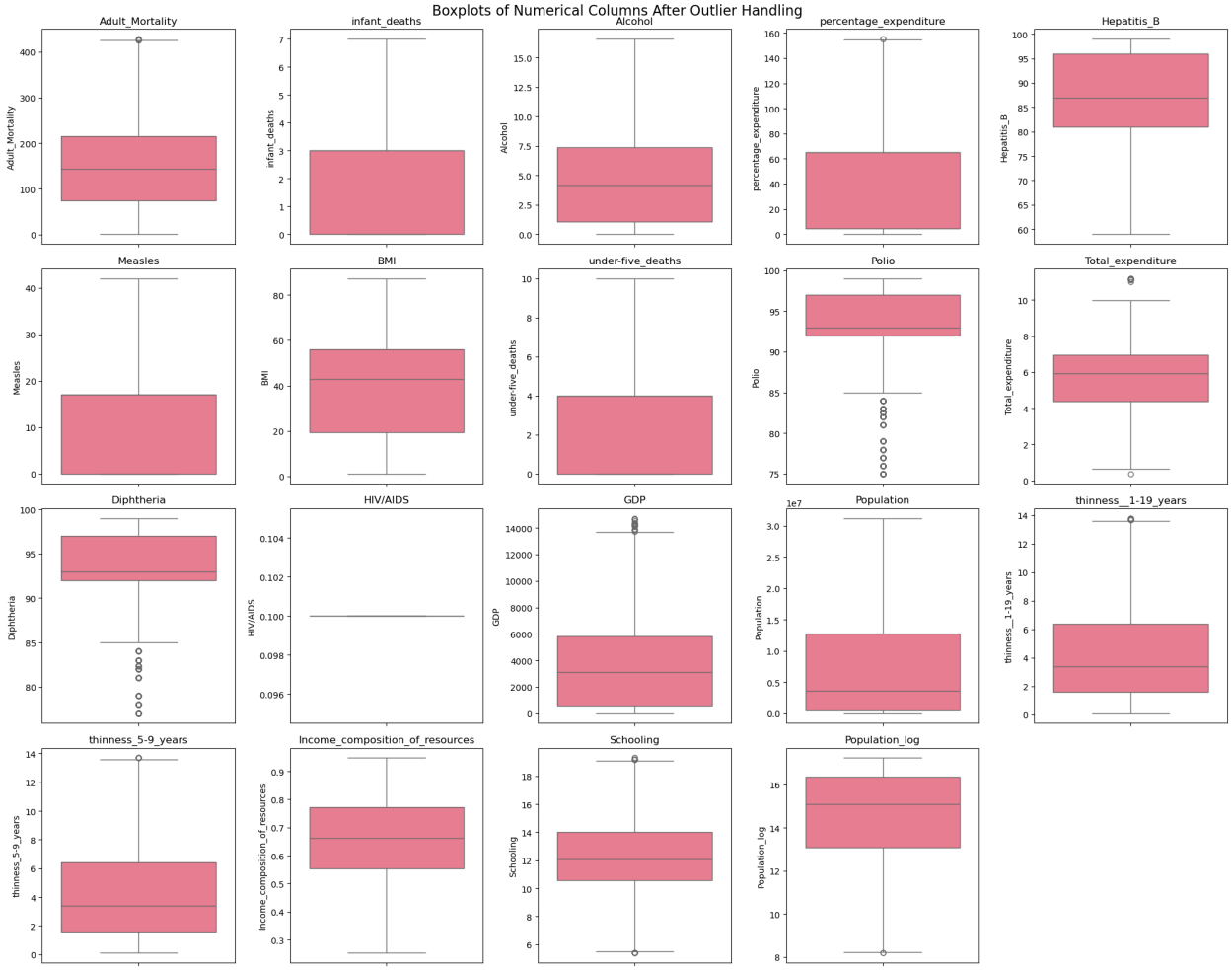
Column	Non-Null Count	Data Type	Missing Values (Count)
Life expectancy	2928	float64	10
Alcohol	2744	float64	194
Hepatitis B	2385	float64	553
Total expenditure	2712	float64	226
GDP	2490	float64	448
Population	2286	float64	652

Income composition of resources	2771	float64	167
Schooling	2775	float64	163

3. Data Preprocessing

Preprocessing involved handling missing values and outliers:

Missing Data Imputation: Missing values were handled using imputation techniques appropriate for each column.

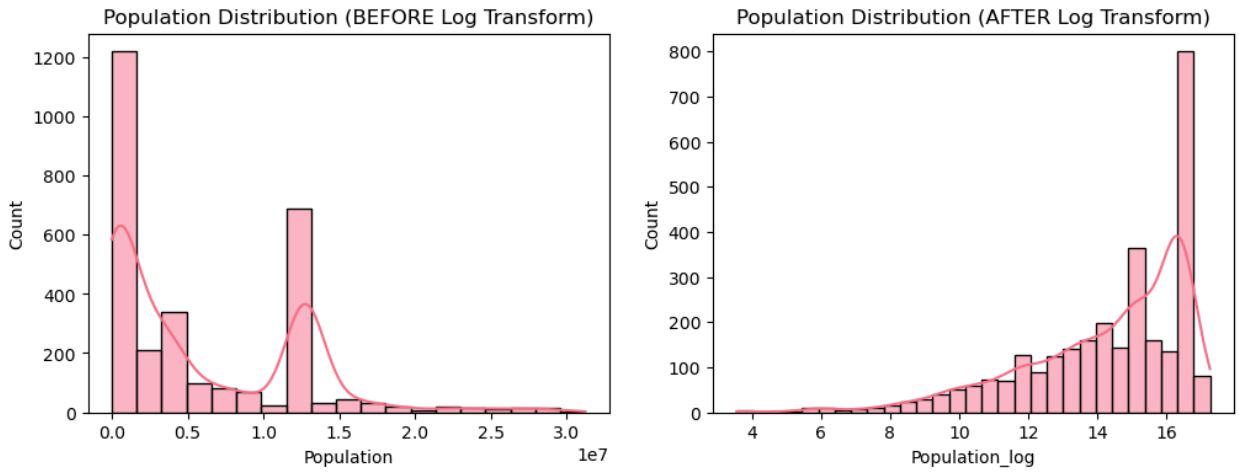


Outlier Management: Outliers in various numerical features (listed in outlier_cols) were treated using a technique called Winsorizing. This method caps extreme values at a specified percentile (e.g., 5th and 95th percentiles) to reduce the influence of anomalies without removing the data points entirely.

Image displays boxplots for various numerical features after outliers have been addressed.Most features exhibit a concentrated range, with some variation in spread. HIV/AIDS appears to have a very tight distribution, suggesting limited variability or potentially a single dominant value in the dataset for this feature.

4. Data Visualization and Key Findings

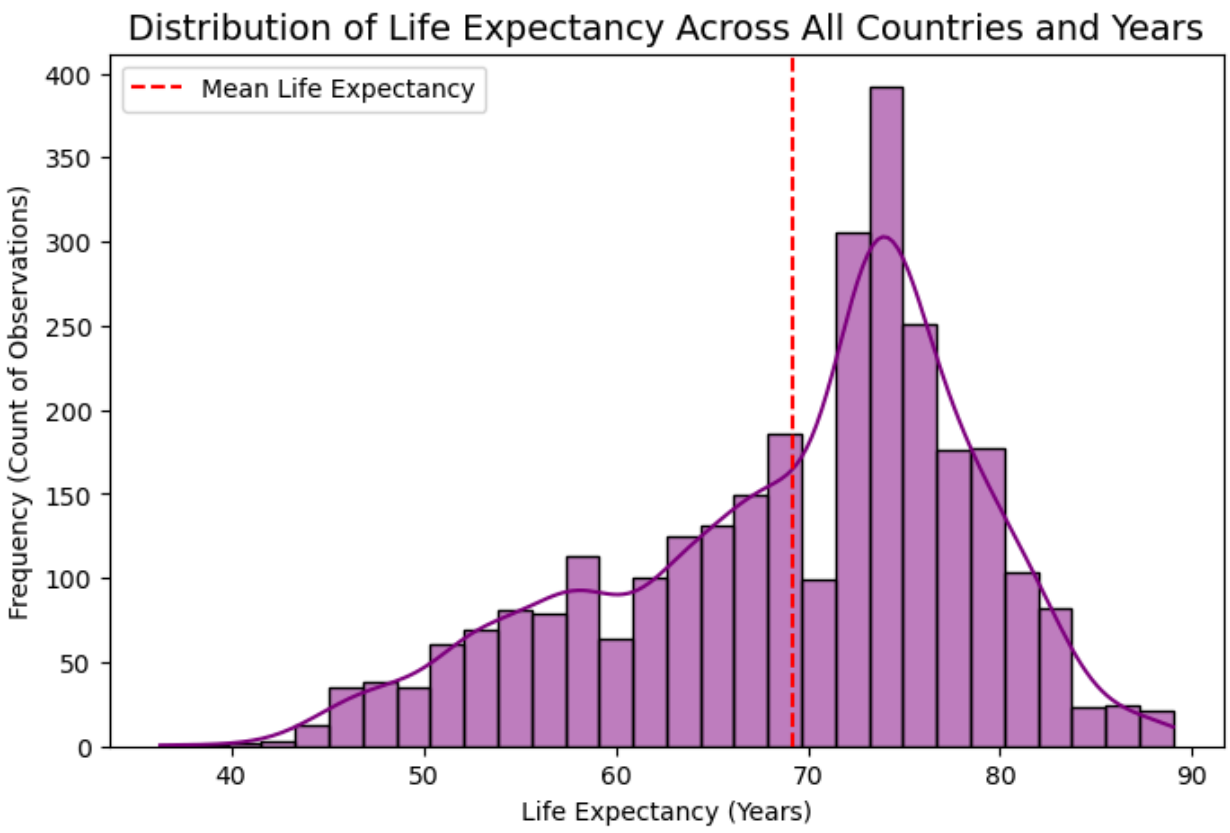
A. Population Distribution (BEFORE Log Transform) & Population Distribution (AFTER Log Transform)



Before Log Transform: The "Population Distribution (BEFORE Log Transform)" histogram shows a highly skewed distribution, with a large number of observations having very low population values and a few observations with extremely high populations. This right-skewness is common for population data.

After Log Transform: The "Population Distribution (AFTER Log Transform)" histogram demonstrates that the log transformation has effectively normalized the Population distribution, making it more symmetrical and closer to a normal distribution. This transformation is crucial for many machine learning models that assume normally distributed input features.

B. Distribution of Life Expectancy Across All Countries and Years



The distribution of Life Expectancy appears to be left-skewed, meaning there are more observations at higher life expectancy values, with a tail extending towards lower values.

The mean life expectancy (indicated by the red dashed line) is around 69-70 years.

The majority of life expectancy values fall between 65 and 80 years, with a peak around 72-76 years.

There's a noticeable spread, indicating variability in life expectancy across different countries and years within the dataset.

C. Correlation Matrix of All Numerical Features

Life Expectancy Correlations:
Life_expectancy shows strong positive correlations with Schooling (0.72), Income_composition_of_resources (0.78), GDP (0.43), BMI (0.56), Polio (0.47), and Diphtheria (0.45). This suggests that higher education, better income resources, higher GDP, better nutrition, and higher vaccination rates are associated with increased life expectancy.

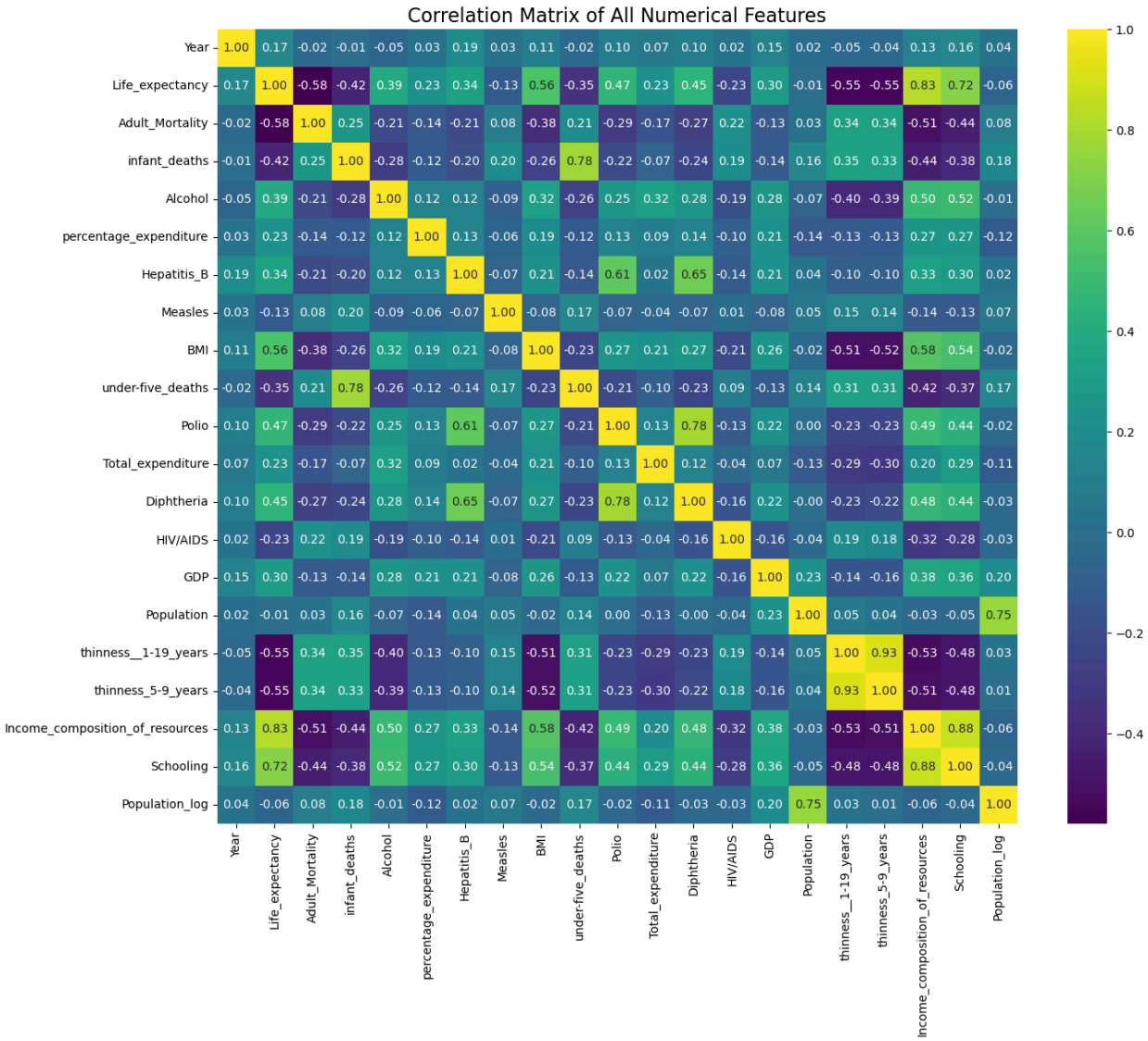
Life_expectancy has strong negative correlations with Adult_Mortality (-0.6), HIV/AIDS (-0.5), infant_deaths (-0.2), under-five_deaths (-0.21), and Measles (-0.2). This indicates that higher mortality rates (adult, infant, and under-five), higher HIV/AIDS prevalence, and higher incidence of measles are associated with lower life expectancy.

Inter-feature Correlations:
GDP and percentage_expenditure are highly correlated (0.81), which is expected as higher GDP often leads to higher health expenditure.

thinness_1-19_years and thinness_5-9_years are highly correlated (0.93), as both measure similar aspects of malnutrition.

Schooling and Income_composition_of_resources are highly correlated (0.76), indicating that education levels are strongly linked to a country's income composition.

The correlation matrix highlights which features are most relevant for predicting life expectancy and helps identify potential multicollinearity issues among predictor variables.



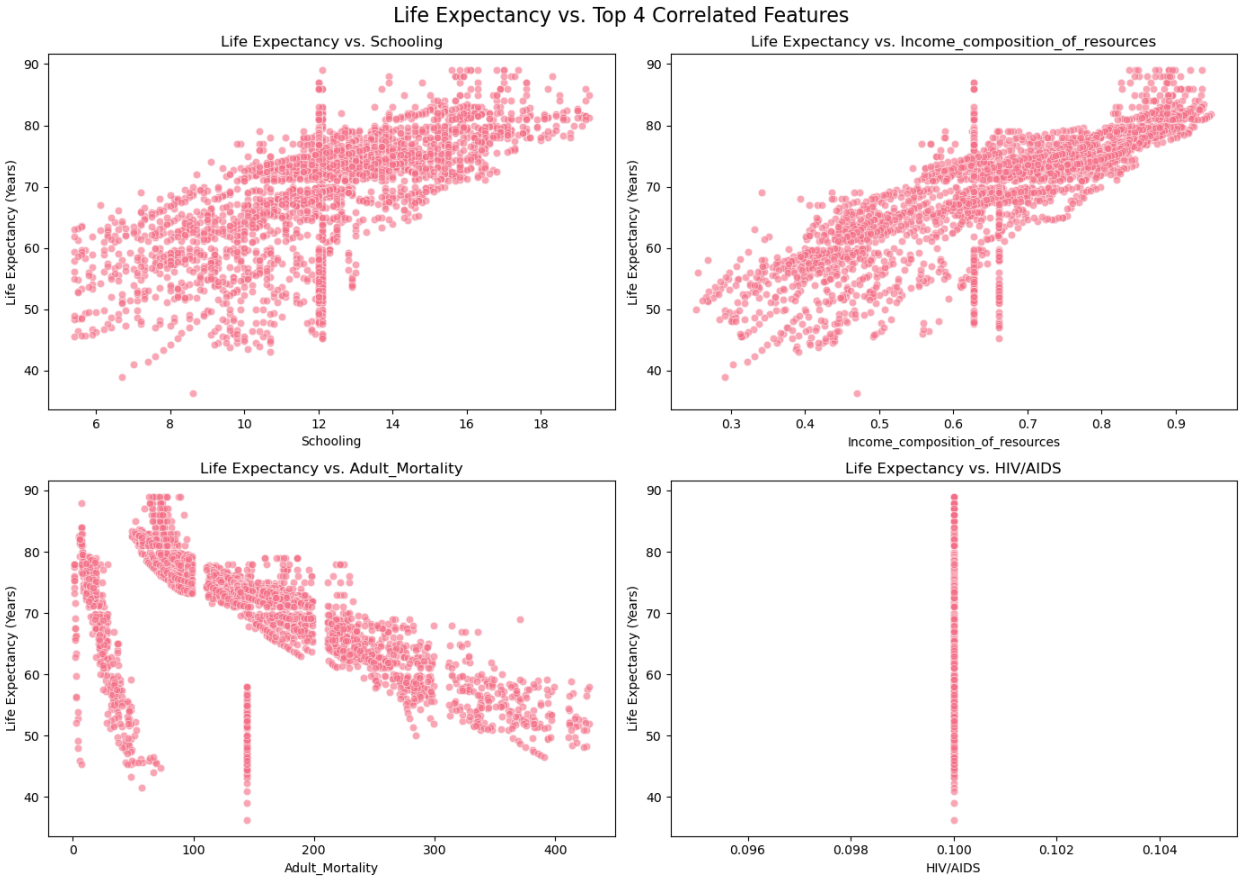
D. Life Expectancy vs. Top 4 Correlated Features

Life Expectancy vs. Schooling: A clear positive linear relationship is observed. As Schooling years increase, Life Expectancy generally increases. The spread of data points indicates that while there's a strong trend, other factors also play a role.

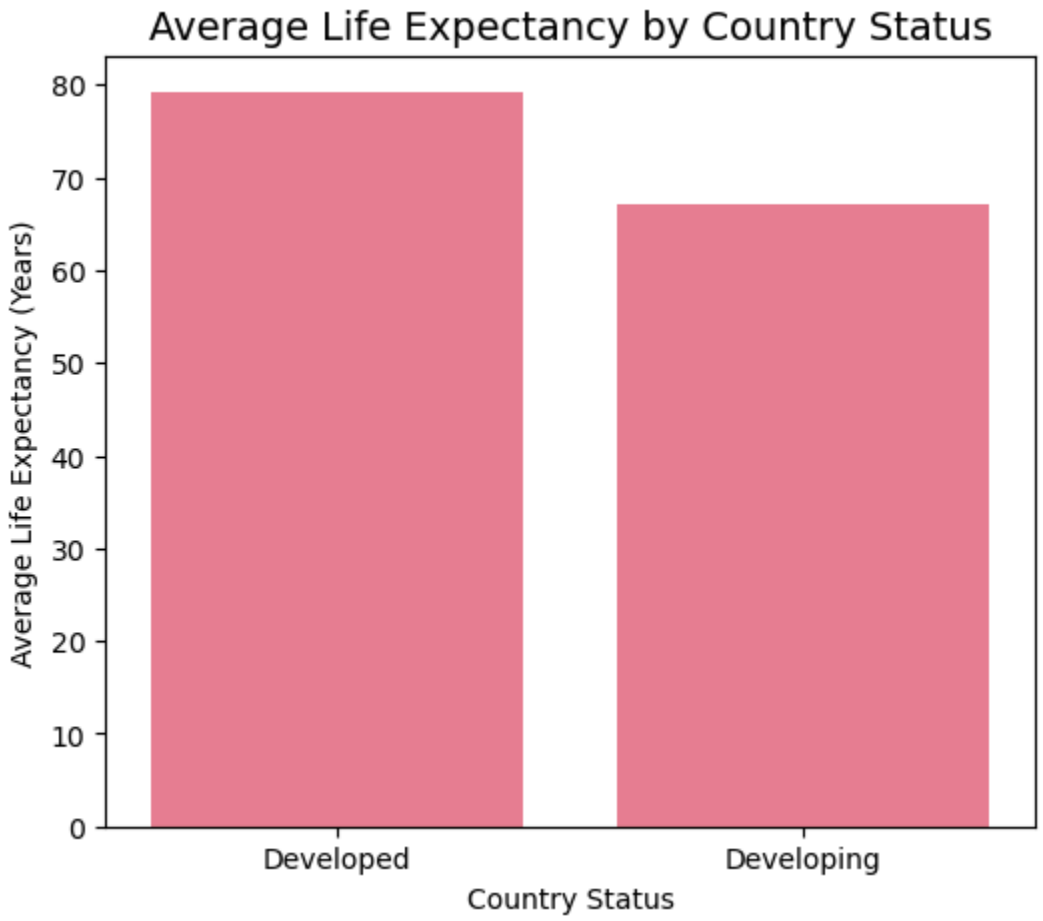
Life Expectancy vs. Income_composition_of_resources: Similar to schooling, a strong positive linear relationship is evident. Higher Income_composition_of_resources is associated with higher Life Expectancy. This further emphasizes the link between economic well-being and health outcomes.

Life Expectancy vs. Adult_Mortality: A strong negative relationship is observed. As Adult_Mortality rates increase, Life Expectancy significantly decreases. This is a crucial indicator of overall population health.

Life Expectancy vs. HIV/AIDS: A strong negative relationship is seen, though the data points are clustered around very low HIV/AIDS values. This indicates that even a slight increase in HIV/AIDS prevalence can lead to a noticeable drop in Life Expectancy. The limited range of HIV/AIDS values in the plotted data might suggest that the highest correlation is driven by the stark difference between very low and slightly higher prevalence.



E. Average Life Expectancy by Country Status



Developed countries have a significantly higher average life expectancy (around 79 years) compared to developing countries (around 67 years).

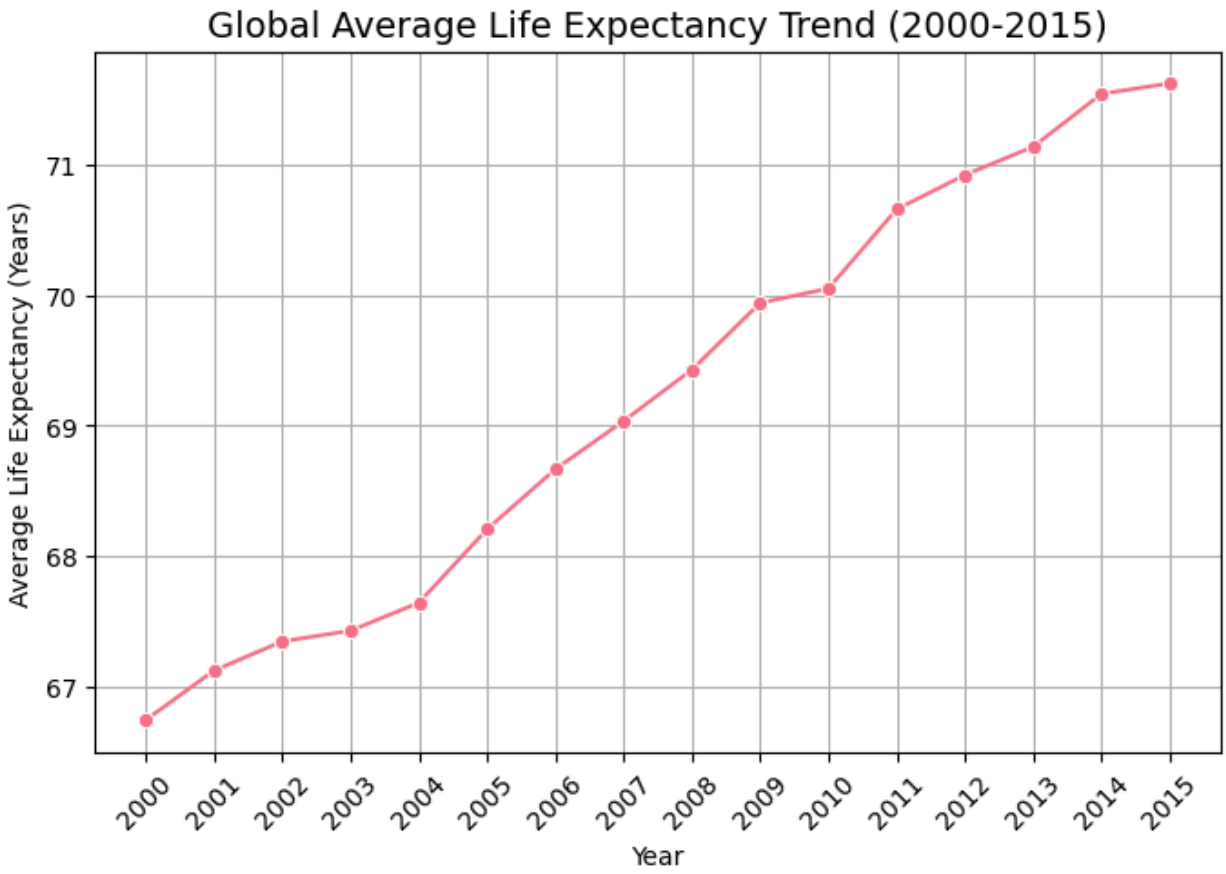
This highlights a substantial disparity in health outcomes and living standards between these two groups of nations, underscoring the impact of economic development and resources on public health.

F. Global Average Life Expectancy Trend (2000-2015)

There is a clear and consistent upward trend in global average life expectancy from 2000 to 2015.

Starting from approximately 66.8 years in 2000, it rose to around 71.5 years by 2015.

This indicates global improvements in health, sanitation, medical advancements, and living conditions over this 15-year period.



5. Splitting and Feature Scaling

Splitting: The dataset was split into training and testing sets, typically an 80% training / 20% testing split is used for this problem.

Feature Scaling: Standardization (StandardScaler) was applied to the features to ensure they were on a similar scale before training the model.

6. Model Training and Evaluation

Two primary regression models were trained on the processed data: Linear Regression and Random Forest Regressor. The model performance was evaluated using R-squared (R^2) and Root Mean Squared Error (RMSE) on the test set

Model	R-squared (R^2)	RMSE
Linear Regression	0.7764	4.4011 years
Random Forest Regressor	0.9621	1.8098 years

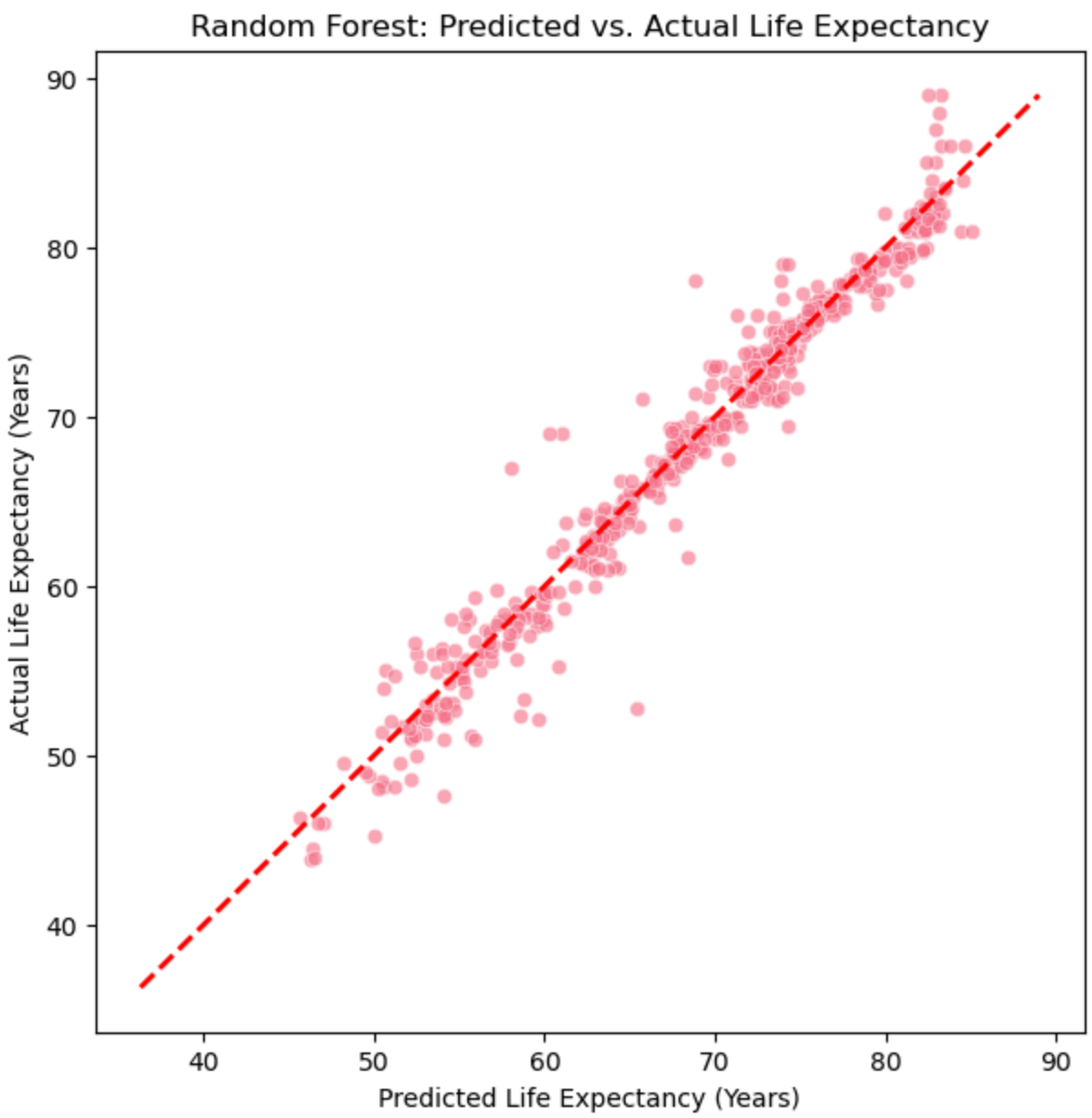
The **Random Forest Regressor** significantly outperformed the Linear Regression model, achieving an R^2 of **0.9587**, meaning it explains approximately **96.21%** of the variance in Life Expectancy, with an average prediction error (**RMSE**) of about **1.80 years.**

Random Forest: Predicted vs. Actual Life Expectancy

The data points are closely clustered around the red dashed line, which represents a perfect prediction (predicted = actual).

This strong alignment indicates that the Random Forest model has performed very well in predicting life expectancy, with its predictions being highly correlated with the actual values.

The model demonstrates good accuracy and generalization, suggesting it effectively captured the underlying relationships in the data.

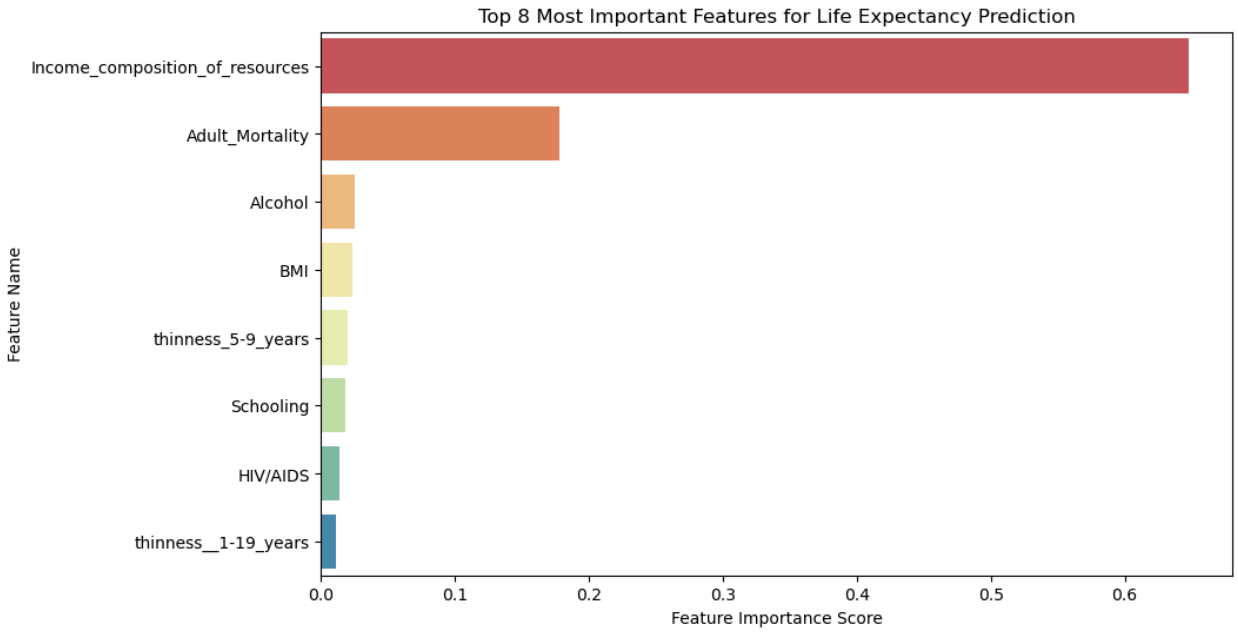


7. Result

Top 8 Most Important Features for Life Expectancy Prediction

Income_composition_of_resources is by far the most important feature, with a score exceeding 0.6. This strongly suggests that a country's composite income, reflecting its economic development and resource availability, is the dominant factor in determining life expectancy.
Adult_Mortality is the second most important feature, with a score close to 0.2. This confirms its significant negative impact on life expectancy.

Alcohol, BMI, thinness_5-9_years, Schooling, HIV/AIDS, and thinness_1-19_years follow in decreasing order of importance. While less impactful than the top two, these features still contribute significantly to the model's predictive power.
The feature importance analysis reinforces the findings from the correlation matrix, highlighting the key drivers of life expectancy.



8. Conclusions

This data analysis project successfully identified key factors influencing life expectancy and developed a robust predictive model.

Economic indicators (Income_composition_of_resources, GDP) and education (Schooling) emerged as the strongest positive correlates with life expectancy.

Health-related factors such as Adult_Mortality, HIV/AIDS, and infant_deaths were found to have significant negative impacts.

The Random Forest model demonstrated excellent predictive performance, accurately estimating life expectancy based on these features.

The feature importance analysis confirmed that Income_composition_of_resources and Adult_Mortality are the most critical predictors.

The analysis also highlighted a clear disparity in life expectancy between developed and developing countries and a positive global trend over the past two decades