

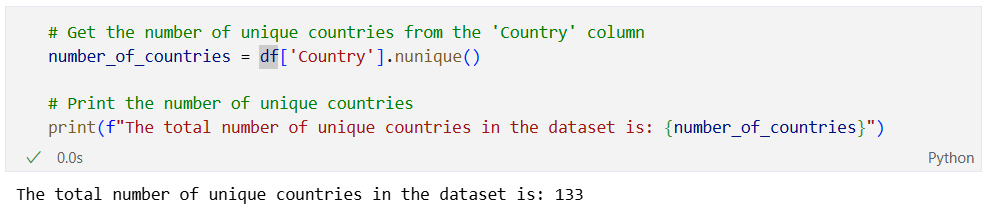
**Final Report: Analysis of Global Life Expectancy Data**

**Author:** **GROUP D (EVE)**

|  |  |
| --- | --- |
| NAKALEMA JULIAN | 23/U/13388/EVE |
| KAYIWA RAHIM | 23/U/09499/EVE |
| NAMEMBWA SHERRY | 23/U/14586/EVE |
| NSUBUGA IBRAHIM | 23/U/24436/EVE |

### Introduction to the Dataset

The dataset under analysis contains life expectancy data for 133 countries.



The data spans from the year 2000 to 2015 and incorporates a variety of factors believed to influence the life expectancy of a nation's population.

The dataset comprises 22 columns (features) and 1,649 rows (observations).



### Key Objectives

The primary goal of this analysis is to identify the key drivers of life expectancy across the globe. The specific objectives are:

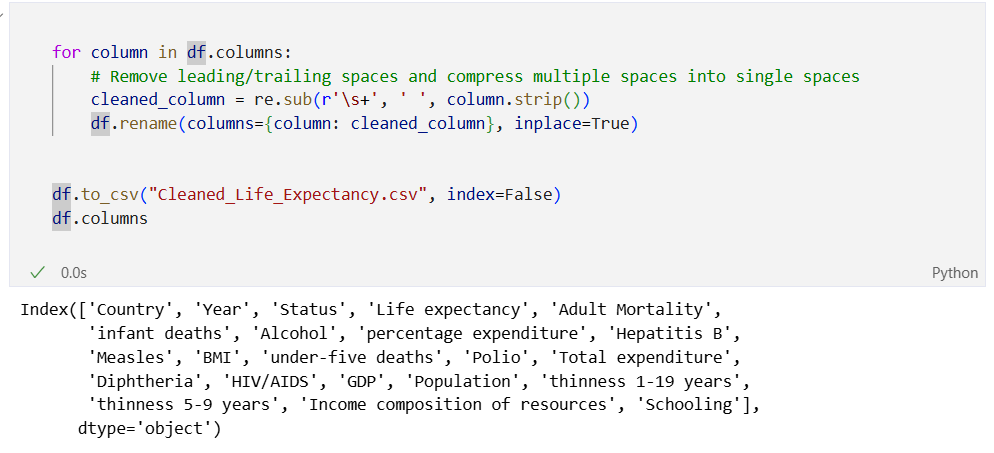
1. **To identify the most determinative social, economic, and health factors to life expectancy.**
2. To find how the status of a specific country affects its life expectancy.
3. To find out the countries that have the highest and lowest life expectancy
4. To find how schooling influences life expectancy
5. To find how HIV/AIDS influence life expectancy.
6. To test how well the model performs using regression metrics.

## Task 1: Exploring the dataset and identifying missing values

### Header Normalization

The original document had malformed column names and so cleaning was performed. This involved a two-step process:

Removing any leading or trailing whitespace characters, and then, replacing any sequence of one or more internal whitespace characters with a single space.



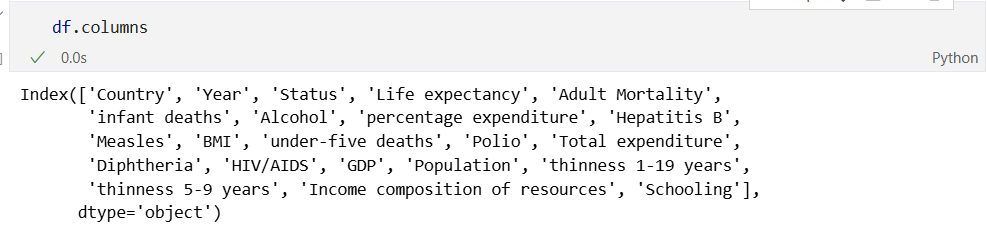
This improves the usability and accessibility of the Data Frame columns, making them easier to reference in analysis.

The features can be broadly categorized into:

* **Demographic:** Country, Year, Status, Population.
* **Health Indicators:** Life expectancy, Adult Mortality, infant deaths, under-five deaths, BMI, thinness 1-19 years, thinness 5-9 years.
* **Healthcare Factors:** Polio, Diphtheria, Hepatitis B, Total expenditure (on health).
* **Disease prevalence:** Measles, HIV/AIDS
* **Socioeconomic Factors:** GDP, Schooling, Income composition of resources
* **Lifestyle Factors:** Alcohol

### Feature Descriptions

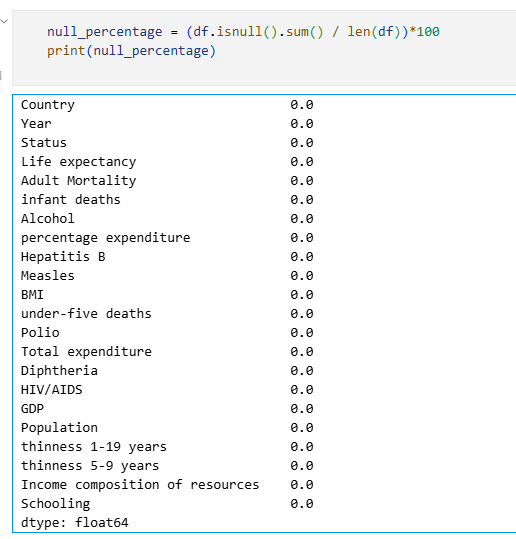
This section provides detailed descriptions of each feature within the dataset. For certain features whose original labels were ambiguous, their precise representation has been clarified through a thorough analysis of their correlation with 'Life expectancy', the primary outcome variable in the correlation heatmap plot.



* **Country:** The geographical entity for which the data is recorded.
* **Year:** The calendar year of the observation, ranging from 2000 to 2015.
* **Status:** Categorizes the country's development level as either 'Developed' or 'Developing'.
* **Life expectancy:** The average number of years a newborn infant is expected to live.
* **Adult Mortality:** The probability of an individual dying between the ages of 15 and 60, expressed per 1000 people.
* **infant deaths:** The absolute number of infants who died before reaching their first birthday in a given year for that country.
* **Alcohol:** The recorded per capita consumption of pure alcohol (in litres) for individuals aged 15 and older.
* **Hepatitis B:** The percentage of 1-year-old children immunized with the Hepatitis B vaccine, indicating public health program coverage.
* **Measles:** The absolute number of reported measles cases in that country for that year, reflecting the prevalence of this infectious disease.
* **BMI:** The average (mean) Body Mass Index for the entire population of the country, serving as a national-level indicator of nutritional status.
* **under-five deaths:** The absolute number of children who died before reaching their fifth birthday in a given year for that country.
* **Polio:** The percentage of 1-year-old children who have received the Polio immunization.
* **Total expenditure:** The general government expenditure on health, expressed as a percentage of the total government expenditure across all sectors.
* **Diphtheria:** The percentage of 1-year-old children who have received the Diphtheria, Tetanus Toxoid, and Pertussis (DTP3) immunization.
* **HIV/AIDS:** The number of deaths per 1000 live births attributed to HIV/AIDS in children under the age of 5, measuring the epidemic's mortality impact on the very young.
* **GDP:** Gross Domestic Product per capita in US Dollars ($), representing the average economic output per person and serving as a standard measure of a nation's wealth.
* **Population:** The total population of the country for the given year.
* **thinness 1-19 years:** The prevalence (percentage) of thinness (being underweight) among children and adolescents aged 10 to 19 years.
* **thinness 5-9 years:** The prevalence (percentage) of thinness (being underweight) among children aged 5 to 9 years.
* **Income composition of resources:** A component of the Human Development Index (HDI) that reflects the standard of living, calculated using gross national income per capita and scaled from 0 to 1.
* **Schooling:** The average number of years of education received by the population of that country.

### Identify Missing and Anomalous Data

An introductory analysis was conducted to assess the presence of missing values in the dataset. While no explicit missing values (e.g., NaN, None, or empty cells) were found.



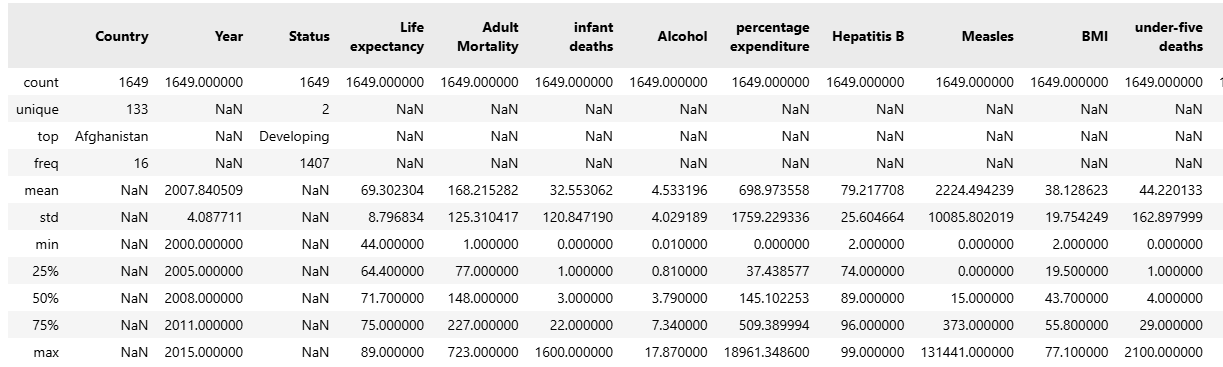
Further inspection revealed certain variables exhibiting characteristics typically associated with data quality issues most notably, extreme outliers likely introduced through placeholder values or data entry errors.

To support these findings, we examined the output from the statistical summary.

Our investigation focused on identifying features with suspicious distributions using the following methods:

* **Percentile Range Comparison**: We compared the minimum values with the 25th percentile (Q1) and the maximum values with the 75th percentile (Q3) for each feature. Significant jumps between Q3 and the maximum value suggested the presence of high-end outliers, while large gaps between Q1 and the minimum value indicated possible low-end outliers.

Features flagged through this approach included:  
**Adult Mortality**, **infant deaths**, **percentage expenditure**, **Measles**, **under-five deaths**, **Polio**, **Diphtheria**, **HIV/AIDS**, **Population**, **thinness 1–19 years**, and **thinness 5–9 years**.

****

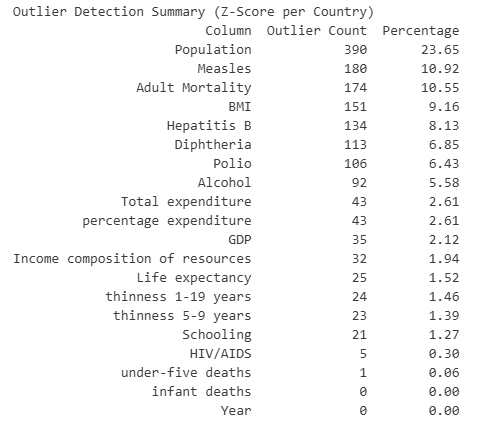
### Advanced Outlier Detection

Recognizing that the earlier exploratory methods based on summary statistics and distribution comparisons were heuristic and may not consistently yield precise results, we proceeded to apply a more robust statistical technique for identifying outliers:

#### ****1. Z-Score Method****

The Z-score method calculates how many Median Absolute Deviations (**MAD**) a data point lies from the median. Values with an absolute Z-score greater than 3.5 were classified as outliers. This approach enabled a more objective, quantitative detection of anomalous data across all numerical features.

The table below summarizes the number and percentage of outliers detected per feature using the Z-score method:



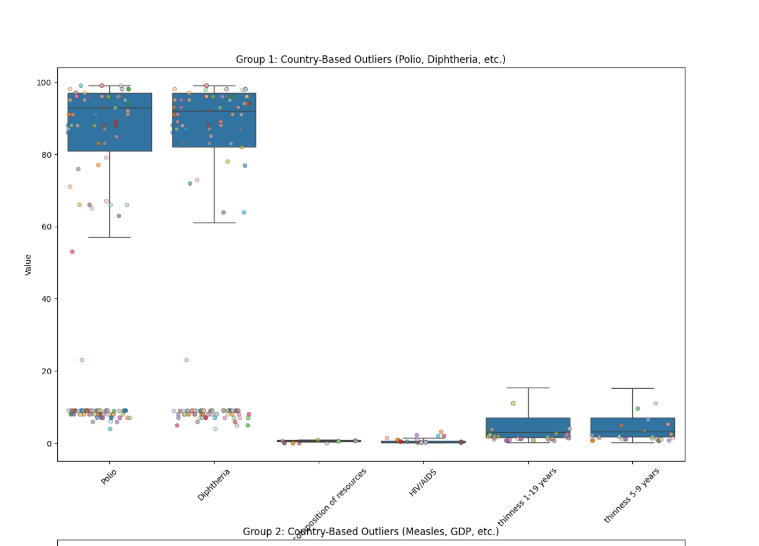
To ensure data quality while avoiding over-cleaning, we set a threshold of **2% maximum acceptable outlier rate per feature.** Features exceeding this limit were marked for further treatment using capping, transformation, or imputation techniques. Conversely, features below the threshold were considered statistically sound and retained without modification.

#### ****2. Box Plot (Visual Inspection)****

To supplement statistical techniques, we also employed **box plots** as a visual method for identifying outliers in key features. Box plots offer a way to observe the distribution of data and highlight potential anomalies using the interquartile range (IQR).

**Each box plot displays:**

* The **median** (Q2),
* The **interquartile range** (IQR: Q1 to Q3),
* And **whiskers** that extend to 1.5 × IQR beyond Q1 and Q3.
* Data points lying beyond these whiskers are plotted individually as **potential outliers**.



## Task 2: Strategy for Handling Missing Data

The dataset did not contain explicit missing values (e.g, NaN, null, or blank fields). However, through both statistical summaries and visual inspection, we identified **suspicious values** and **outliers** likely resulting from placeholder values, data entry errors, or structural anomalies across countries.

To address these inconsistencies, we adopted a **multi-strategy approach**, customized per feature group based on their statistical characteristics:

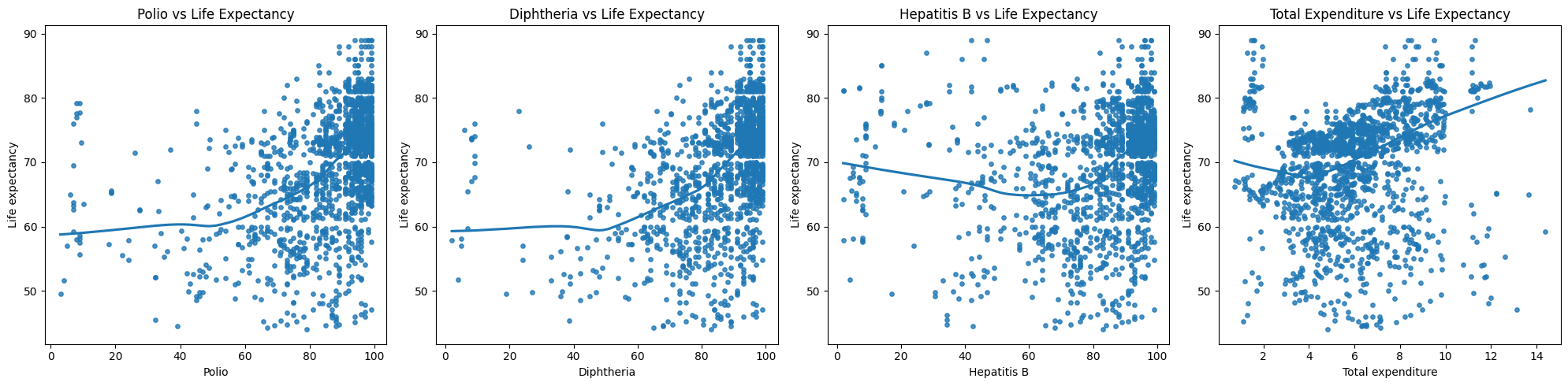
#### ****Log Transformation****

**Applied to:** Measles, Population, Alcohol, percentage expenditure. This method was chosen for features that are highly skewed.



#### ****IQR Capping****

**Applied to:** Polio, Diphtheria, Hepatitis B, Total expenditure. This method is ideal for features that are expected to be relatively stable or bounded within a specific range for a given country, but may contain occasional anomalies due to reporting errors or administrative issues.



#### ****Percentile Capping****

**Applied to:** Adult Mortality. This method was chosen because Adult Mortality is highly skewed and not bounded by a natural limit like a percentage. The distribution can be heavily influenced by catastrophic events like war, famine or pandemics which create legitimate but extreme values.

#### ****Custom GDP Handling****

**Applied to:** GDP. A single method is insufficient for GDP due to two distinct challenges: extreme variance between countries (inter-country) and potential for anomalies within a country's timeline (intra-country). A hybrid approach was necessary.

1. **IQR Capping (Per-Country):** This first step addresses intra-country anomalies. A sudden spike or dip in a country's GDP that is inconsistent with its own trend is likely a data entry error. The per-country IQR method is perfect for catching these.
2. **Domain-Specific Lower Bound:** The statistical lower bound from the IQR method is often negative, which is economically impossible. We enforced a minimum plausible value (e.g., $100 per capita) to catch any erroneously low values.
3. **Nullify and Impute:** Outliers were first set to NaN (null) to completely remove their influence. Then, these nulls were imputed using the country's own (now cleaned) mean GDP. This prevents an outlier from contaminating its own replacement value and provides a stable, contextually relevant estimate.

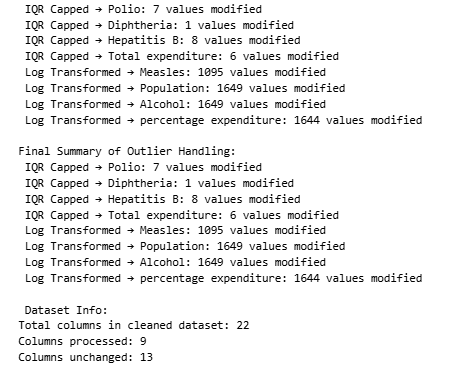
#### ****Custom BMI Handling****

**Applied to:** BMI. The primary issue with BMI is not statistical skew but biological implausibility. A country's average BMI cannot be, for example, 2 or 70. This requires a domain-knowledge approach followed by a sophisticated imputation method.

1. **Domain-Based Filtering:** We first identified and nullified any BMI values outside a biologically plausible range for a national average (e.g., 15 to 40). This is more direct and accurate than relying on statistical methods.
2. **Multivariate Imputation (MICE):** Simple mean or median imputation for BMI would be inaccurate because BMI is strongly correlated with other features like thinness, Schooling, income composition of resources and Life expectancy. **Multiple Imputation by Chained Equations (MICE)** is a superior method that leverages these relationships. It builds a model to predict the missing BMI values based on the other columns, resulting in a much more accurate and realistic imputation that preserves the overall structure of the dataset.

## Task 3: Implementation of Data Handling and evaluation of its impact

Our cleaning strategy focused on robust handling of **missing values** and **outliers**, tailored per feature based on its distribution and sensitivity. The entire process was conducted **on a per-country basis** to ensure national statistical integrity and minimize distortions caused by global-scale thresholds.



### Evaluation of Data Cleaning Impact

Following the systematic cleaning and transformation of the dataset, we assess below the **quantitative and qualitative impacts** of the procedures applied, particularly focusing on **missing value imputation, outlier detection and correction** and **feature transformation**. Each step aimed to enhance the integrity, quality, and reliability of the dataset for subsequent analysis and modelling.

### ****Outlier Detection and Correction Summary****

|  |  |  |  |
| --- | --- | --- | --- |
| **Method** | **Feature(s) Handled** | **Modified Values** | **Notes** |
| **Percentile Capping** | Adult Mortality | 238 | Controlled extreme skew at high end |
| **IQR Capping** | Polio, Diphtheria, Hepatitis B, Total expenditure | 130, 122, 156, 60 | Adjusted moderate outliers while preserving distribution |
| **Log Transformation** | Measles, Population, Alcohol, percentage expenditure | 1095, 1649, 1649, 1644 | Compressed large variance and skew |
| **Custom Handling (GDP)** | GDP | 133 (outliers detected and treated) | Combined IQR detection, nullification, and country-wise mean imputation |
| **Custom Handling (BMI)** | BMI | 1092 values invalidated and imputed | Range filtering + MICE Imputation based on Life expectancy |

### ****Specific Impacts****

#### ****Adult Mortality (Percentile Capping)****

* 238 values adjusted.
* Controlled extreme outliers likely due to conflict or misreporting.
* Helped stabilize skewed mortality trends across countries.

#### ****IQR-Capped Features****

* Capping Polio, Diphtheria, Hepatitis B, and expenditure fields reduced volatility from abrupt reporting variations.
* Changes were **modest** (between 60–156 values per feature), indicating mostly consistent data with a few outliers.
* Preserved the structure while controlling spread.

#### ****Log-Transformed Features****

* Extremely skewed features like Population and Measles were normalized.

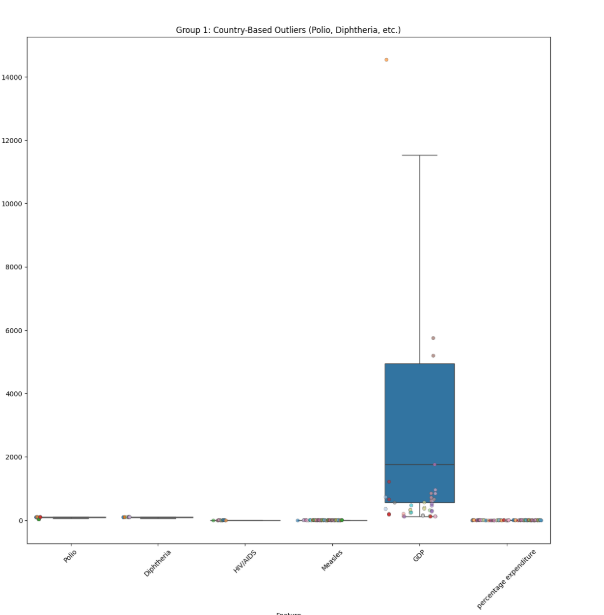


#### ****Custom GDP Handling****

* **133 outliers** identified and replaced using a per-country approach.
* Ensured that exceptionally low or inflated GDP values did not distort national economic profiles.
* Final GDP imputation used:
  + **Country mean** for contextual consistency.
  + **Global mean fallback** only if a country had no valid GDP entries.

#### ****BMI Handling via MICE****

* **1092 invalid BMI values** identified (<15 or >40).
* Replaced with NaN and imputed using **Multiple Imputation by Chained Equations (MICE)** based on Life expectancy.
* Enhanced realism of BMI entries, aligning with biological plausibility.

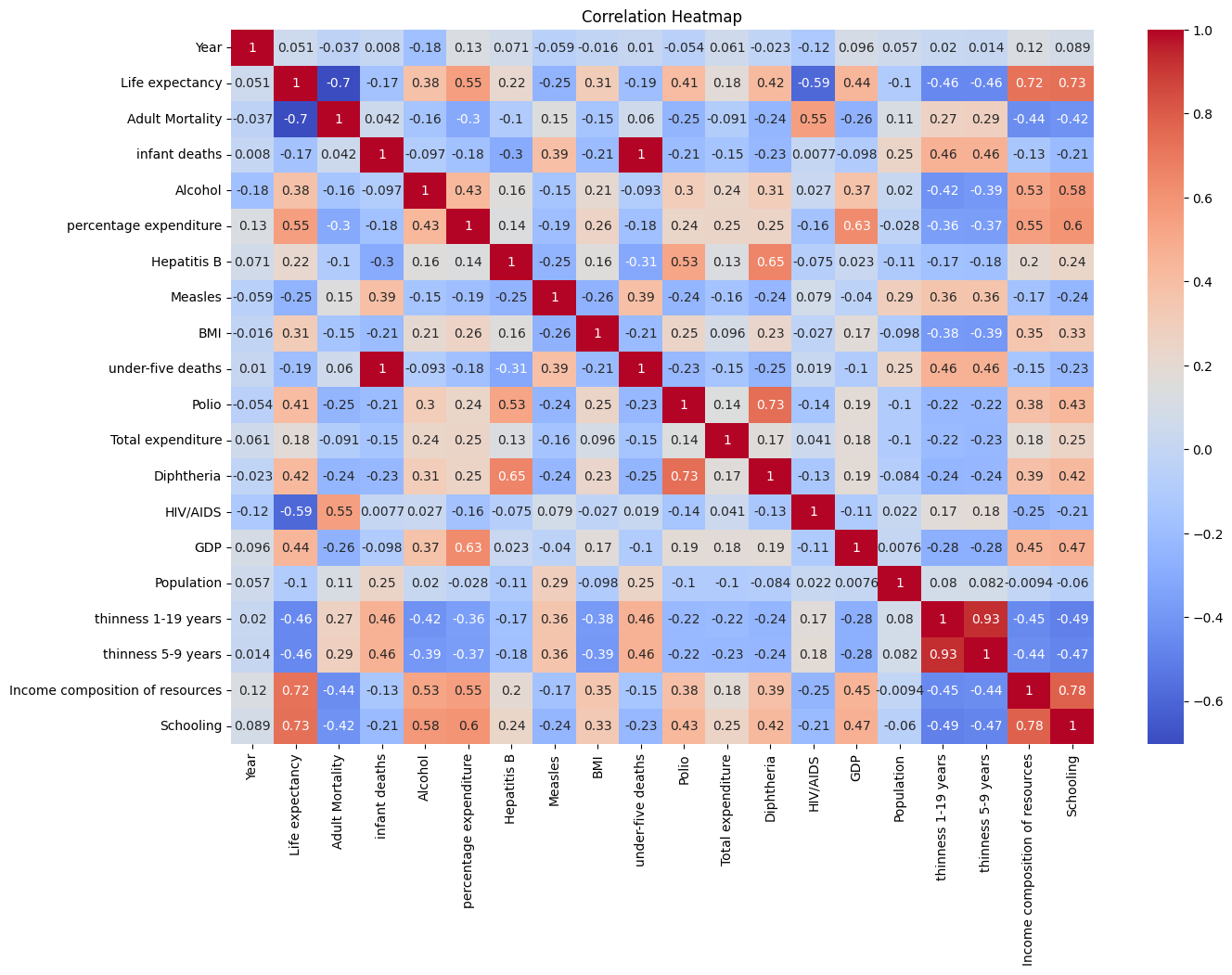


## Task 4: Exploring the dataset and identifying potential features.

In this task, we explored the dataset to identify a set of features that are most relevant and non-redundant for predicting **Life Expectancy**.

### ****Objective 1: To identify the most determinative social, economic, and health factors to life expectancy.****

We explored the data set so as to identify the key determinants of life expectancy. We used a correlation heatmap as shown below.



Based on our findings, we considered factors whose correlation magnitude with respect to life expectancy was greater than 0.5.

The health indicators were Adult mortality and HIV/AIDS. The economic factors were Percentage expenditure and income composition. The social factor was schooling.

A key challenge in feature selection is addressing **multicollinearity** a situation where independent variables are highly correlated, potentially distorting model interpretation and degrading performance.

To address this, I used the **Variance Inflation Factor (VIF)** technique to quantify the multicollinearity among the numerical features.

### ****Methodology****

1. **Selection of Numerical Features**  
   I selected all numerical features (float and int types) from the cleaned dataset, excluding the target variable 'Life expectancy'.
2. **Computation of VIF**  
   I computed the Variance Inflation Factor (VIF) for each feature using the variance\_inflation\_factor function. VIF measures how much the variance of a feature is inflated due to its correlation with other features.
3. **Interpretation of VIF**
   * **High VIF (typically > 10)**: Indicates that the feature is highly correlated with other features and may be redundant.
   * **Low VIF (< 5)**: Suggests that the feature is likely independent and adds unique information to the model.
4. **Output Summary**  
   The VIF results were printed in a descending order to highlight the most collinear features. Notably, features such as:
   * 'under-five deaths' and 'infant deaths' had extremely high VIF values (~200),
   * 'Diphtheria', 'Schooling', and 'Polio' also showed high multicollinearity,
   * Features like 'GDP', 'BMI', and 'HIV/AIDS' had relatively low VIF values, suggesting they offer independent predictive information.
5. **Feature Selection**  
   Based on the VIF scores, we manually selected features with lower multicollinearity for modelling. These include: 'Adult Mortality', 'Alcohol', 'Total expenditure', 'Measles', 'HIV/AIDS', 'GDP', 'BMI'

## Task 5: Implementation of Feature Engineering

To enhance the predictive power of our model, several new features were engineered by combining existing variables. The goal was to create more nuanced indicators that capture complex interactions between health, economic, and social factors.

#### Engineered Features:

1. **Immunization Score**
   * **Formula:** (Polio + Diphtheria + Hepatitis B) / 3
   * **Rationale:** Instead of treating each vaccination type separately, this feature creates a single, composite score representing the overall strength of a country's core immunization program for children. A higher score indicates a more robust and comprehensive public health system.
2. **HIV Impact Score**
   * **Formula:** HIV/AIDS \* Adult Mortality
   * **Rationale:** This interaction term aims to capture the synergistic negative effect of HIV/AIDS and general adult mortality. A high value suggests that HIV/AIDS is a significant contributor to the country's overall adult death rate, likely having a profound negative impact on life expectancy.
3. **Health Investment**
   * **Formula:** percentage expenditure / (GDP + 1e-5)
   * **Rationale:** This creates a ratio representing health spending relative to the nation's wealth (GDP per capita). It normalizes the absolute expenditure, providing a better measure of a country's priority towards health. The small value 1e-5 is added to prevent division by zero.
4. **Mortality Rate Ratio**
   * **Formula:** Adult Mortality / (Life expectancy + 1e-5)
   * **Rationale:** This feature explores the relationship between how long people live and the rate at which adults die prematurely. A high ratio might indicate that even though life expectancy is at a certain level, the risk of dying as an adult is disproportionately high.
5. **Resource Efficiency**
   * **Formula:** (Income composition of resources \* Schooling) / (Alcohol + 1e-5)
   * **Rationale:** This complex feature attempts to model how effectively a country translates its human capital (income and education) into positive outcomes, offset by a negative lifestyle factor (alcohol consumption). It hypothesizes that countries with high income and schooling but low alcohol consumption are more "efficient" at promoting a long life.
6. **Health Spending Per Capita**
   * **Formula:** (Total expenditure \* GDP) / (Population + 1e-5)
   * **Rationale:** This calculates a more accurate estimate of the actual health spending per person by relating the total government health expenditure percentage to the country's GDP and population size.
7. **BMI Deviation**
   * **Formula:** |BMI - 22.5|
   * **Rationale:** Assuming an "ideal" average BMI is around 22.5, this feature measures the absolute deviation from that ideal. It treats being significantly underweight and significantly overweight as equally negative deviations from a healthy national average.
8. **Thinness Ratio**
   * **Formula:** (thinness 1-19 years) / (thinness 5-9 years + 1e-5)
   * **Rationale:** This ratio compares the prevalence of thinness in adolescents to that in younger children. It could provide insights into whether nutritional problems are worsening or improving as a cohort ages.
9. **Healthcare Access Score**
   * **Formula:** Immunization Score \* Health Investment
   * **Rationale:** This combines the measure of immunization coverage with the relative investment in health. A high score suggests that a country not only invests in health but also has the infrastructure to deliver essential services like vaccinations effectively.
10. **GDP per Health $**
    * **Formula:** GDP / (Total expenditure + 1e-5)
    * **Rationale:** This feature represents how much economic output (GDP) a country generates for every "unit" of its government budget spent on health.
11. **Child Vulnerability Index**
    * **Formula:** (0.6 \* HIV/AIDS) + (0.4 \* (under-five deaths + infant deaths))
    * **Rationale:** This is a weighted index that combines the mortality risk from HIV/AIDS with general infant and under-five mortality. It creates a single, powerful indicator of the overall health challenges facing a nation's youngest and most vulnerable population.

## Task 6: Evaluation of New Features' Impact

To rigorously evaluate the impact of the newly engineered features, a comparative analysis was conducted. Two separate models were trained and tested using 5-fold cross-validation to ensure the results were stable and reliable.

1. **Baseline Model:** This model was trained using only the original, selected features from the dataset.
2. **Engineered Model:** This model was trained using both the original features and the 11 new features created in Task 5.

The performance of each model was measured by its **R-squared (R²)** score, which indicates the proportion of the variance in life expectancy that the model can explain.

#### Results:



* **R² Score (Baseline Model):** **0.6907**
  + This indicates that the model using only the original features could explain approximately 69.1% of the variation in life expectancy.
* **R² Score (Engineered Model):** **0.8478**
  + This indicates that the model enhanced with the new features could explain approximately 84.8% of the variation in life expectancy.

#### Conclusion:

The evaluation demonstrates a significant improvement in model performance. The R² score increased by over 15 percentage points (from ~69% to ~85%). This magnificent gain is direct evidence that the feature engineering process was highly successful. The new features effectively captured complex, non-linear relationships and interactions in the data that the original features alone could not, leading to a much more accurate and insightful predictive model.

## Task 7: Identifying Key Variables for Visualization

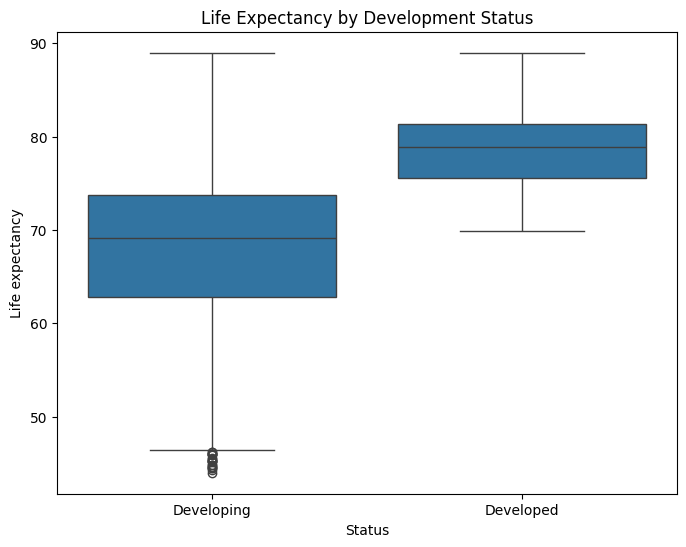
* The key visualizations were chosen due to their high correlation with Life expectancy

**Key Variables Identified:**

* **Schooling** – **Positive correlation**  
  Countries with higher average years of schooling tend to have higher life expectancy, indicating the importance of education in promoting health.
* **Income Composition of Resources** – **Positive correlation**  
  A higher index (indicating more equitable income distribution and access to resources) is associated with longer life expectancy.
* **Percentage Expenditure** – **Positive correlation**  
  Greater health expenditure as a percentage of GDP often results in better healthcare infrastructure and outcomes.
* **HIV/AIDS** – **Strong negative correlation**  
  High prevalence of HIV/AIDS significantly reduces life expectancy, especially in low-resource settings.
* **Adult Mortality** – **Strong negative correlation**  
  Higher adult mortality rates directly reflect reduced life expectancy due to disease burden and inadequate healthcare.

## Task 9: Interpreting visualizations to uncover patterns and insights in the data.

### Objective 2: To find how the status of a specific country affects its life expectancy.



#### Insights

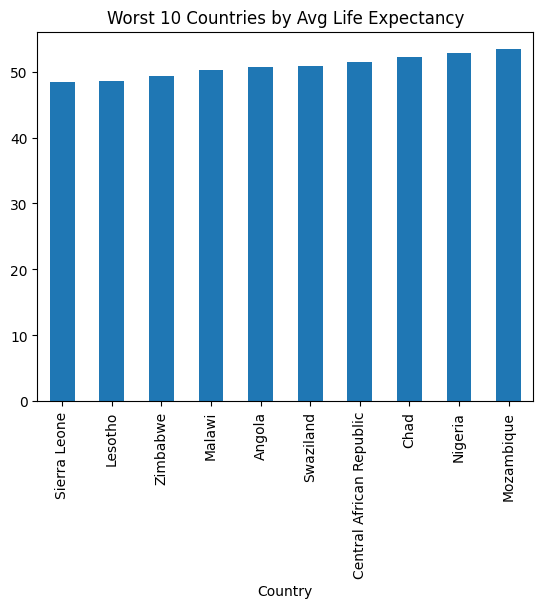
* **Significant Life Expectancy Gap:** Developed countries exhibit substantially higher life expectancy (~80 years) compared to developing countries (~60 years) which highlights significant socioeconomic inequalities in healthcare access, nutrition, sanitation and disease prevention.

#### **Extreme Outliers in Developing Countries**: Several developing countries show catastrophic life expectancy below 50 years (visible as outlier dots), indicating severe health system failures requiring emergency intervention.

#### Recommendations

* **Prioritize Basic Health Infrastructure:** Focus on interventions that move countries from the bottom to middle of the developing country range - clean water, vaccination programs, maternal care, and infectious disease control.
* **Study High-Performing Developing Countries:** Identify developing countries in the upper quartile and systematically copy their health policies, infrastructure investments and service delivery models.

### Objective 3: To find out the countries that have the highest and lowest life expectancy



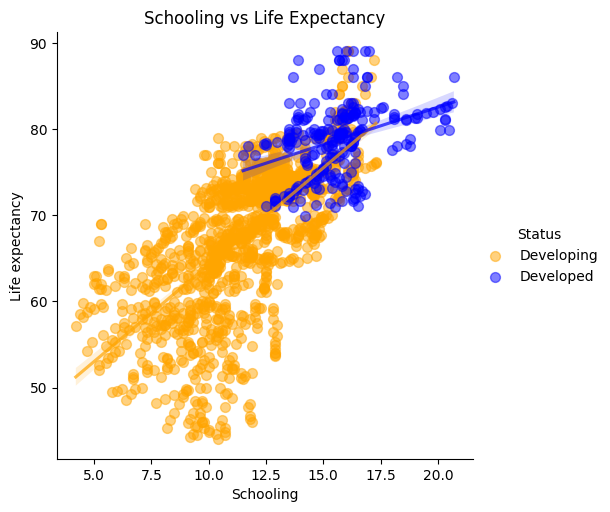
#### Insights

* Extreme Life Expectancy Gap: There's a devastating 35+ year gap between the highest (Ireland ~85 years) and lowest (Sierra Leone ~48 years) life expectancy countries, representing nearly half a human lifetime.
* Geographic Clustering: The worst-performing countries are predominantly African nations, while top performers are developed nations (mostly European countries indicating regional challenges.

#### Recommendations

* **Benchmark Against Best Performers**: Identify the top 3 countries with similar economic conditions and systematically adopt their health policies, infrastructure approaches, and service delivery models.
* **Implement Universal Healthcare**: Establish free basic healthcare coverage immediately, starting with maternal care, child immunizations, and treatment of common infectious diseases.

### Objective 4: To find how schooling influences life expectancy



#### Insights

* **Strong positive correlation**: The plot shows a clear positive relationship between years of schooling and life expectancy. Each additional year of education appears to correlate with meaningful gains in population health outcomes.
* **Development status Divide**: There's a distinct clustering pattern where developed countries with both high schooling (13-20+ years) and high life expectancy (70-88 years), while developing countries (orange) show much greater variation and generally lower values on both metrics.

#### Recommendations

* **Invest in Education Infrastructure through** building more schools, especially in rural and underserved areas.
* Ensure free or affordable access to primary and secondary education for all children.
* **Promote Health Education** through including **health and hygiene** topics in school curriculums and empowering people through community health workshops for adults with limited schooling.

### Objective 5: To find how HIV/AIDS influence life expectancy.



#### Insights

Strong Inverse Relationship: The plot reveals a clear negative correlation between HIV/AIDS prevalence and life expectancy. Countries with higher HIV/AIDS rates experience dramatically reduced life expectancy, with some nations dropping below 30-40 years compared to the 70+ years seen in low-prevalence countries.

The developed countries cluster have low HIV/AIDS rates and higher life expectancy while developing countries show much wider variation and include high HIV/AIDS rates and lower life expectancy.

#### Recommendations

* Prevention as Priority: Invest heavily in prevention programs, as the data shows even moderate increases in HIV/AIDS prevalence can significantly impact population health outcomes i.e. high life expectancy. Prevention is far more cost-effective than treatment at scale.
* Health System Strengthening: Countries with high HIV/AIDS burden need improved health system support
* Regional Cooperation: Given that developing countries bear the highest burden, regional partnerships and international aid coordination are essential for effective response.

## Task 10: Data Splitting and Model Training

To build and evaluate a predictive model, the dataset was systematically prepared, split and used to train several machine learning algorithms.

#### 1. Feature and Target Selection

* **Features (X):** The set of predictor variables was defined. This included all the original cleaned features plus the newly engineered features from Task 5.
* **Target (y):** The Life expectancy column was designated as the target variable, which the models aim to predict.

#### 2. Data Splitting

* **Method:** The dataset was split into two distinct sets using the train\_test\_split function.
  + **Training Set (80%):** This portion of the data was used to teach the machine learning models the underlying patterns connecting the features to life expectancy.
  + **Testing Set (20%):** This portion was held back and used for the final evaluation. The model does not see this data during training, making it a fair and unbiased test of the model's performance on new, unseen data.
* **Reproducibility:** A random state of 42 was used to ensure that the split is the same every time the code is run, making the results reproducible.

#### 3. Model Training and Initial Evaluation

* **Objective:** To identify the best-performing algorithm for this specific prediction task, a suite of diverse regression models was selected for training and comparison.
* **Models Tested:**
  + Linear Regression
  + Decision Tree
  + Random Forest
  + Gradient Boosting
  + K-Nearest Neighbors
  + XGBoost
* **Process:** Each model was trained on the training set (X\_train, y\_train). Subsequently, its performance was evaluated on both the training set (to check for overfitting) and the testing set (to assess generalization ability). Performance was measured using the R-squared (R²) metric and the best performing model was XGBoost with 0.9999 for Train R2 and 0.9559 for Test R2

## Task 11: Cross Validation and Model Evaluation

To obtain a more robust and reliable estimate of each model's performance, 5-fold cross-validation was performed on the entire dataset. This technique provides a better measure of how the model is likely to perform on new unseen data.

#### 1. Methodology

* **Process:** For each model, the data was divided into 5 "folds" or subsets. The model was trained on 4 of the folds and evaluated on the 5th (the hold-out fold). This process was repeated 5 times, with each fold serving as the hold-out set once.
* **Evaluation Metrics:**
  + **R-squared (R²):** Measures the proportion of variance in life expectancy that the model can predict. A higher value is better, with 1.0 being a perfect score.
  + **Mean Absolute Error (MAE):** Measures the average absolute difference between the model's predictions and the actual values. A lower value is better, representing a smaller average prediction error in years.
  + **Standard Deviation (Std Dev):** The standard deviation of the R² scores across the 5 folds. A low value indicates that the model's performance is stable and consistent.

#### 2. Cross-Validation Results

The table below summarizes the mean performance of each model across the 5 folds:

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **CV R² (Mean)** | **Std Dev** | **MAE (Mean)** |
| Linear Regression | 0.8470 | 0.0571 | 2.3210 |
| Decision Tree | 0.8453 | 0.0499 | 2.3785 |
| Random Forest | 0.9187 | 0.0179 | 1.7443 |
| Gradient Boosting | 0.9240 | 0.0096 | 1.7656 |
| K-Nearest Neighbors | 0.3167 | 0.0933 | 5.3738 |
| XGBoost | 0.9152 | 0.0187 | 1.8067 |

#### 3. Interpretation and Model Selection

* **Top Performers:** The cross-validation results confirm that the ensemble models are superior. **Gradient Boosting** emerges as the top-performing model with the highest mean R² (0.9240) and an extremely low standard deviation (0.0096), indicating very stable and reliable performance.
* **Random Forest** and **XGBoost** are also excellent, with R² scores around 0.92 and low MAE values.
* **Lowest Error:** The **Random Forest** model achieved the lowest Mean Absolute Error (1.7443 years), meaning its predictions were, on average, the closest to the true life expectancy.
* **Poor Performers:** K-Nearest Neighbors performed very poorly, with a low R² and a very high MAE, confirming it is not a suitable model for this dataset.
* **Conclusion:** Based on its combination of high accuracy (R²) and exceptional stability (low Std Dev), **Gradient Boosting is selected as the best overall model for this task.** Random Forest is a very close second, offering slightly lower average error.

## ****Task 12: Conclusions, Actionable Insights, Recommendations, and Summary****

### ****1. Key Objectives of the Analysis****

The primary objective of this life expectancy analysis was to **identify the key determinants of life expectancy across countries** using statistical techniques and machine learning models. This involved:

* Cleaning and preparing the dataset.
* Engineering meaningful features to enhance predictive power.
* Building and evaluating regression models.
* Interpreting patterns and insights for actionable policy and health-related recommendations.

### ****2. Key Findings****

* **High Predictive Performance Achieved**: Gradient Boosting and Random Forest models achieved R² scores above 0.92 on cross-validation, confirming that the engineered features meaningfully improved the model's ability to predict life expectancy.
* **Feature Engineering Was Effective**: Engineered features like Health Investment, Immunization Score, Healthcare Access Score, and Mortality Rate Ratio captured complex relationships not present in the raw data and boosted performance.
* **Robustness Confirmed via Cross-Validation**: Gradient Boosting had the most consistent and reliable performance (lowest standard deviation in R²), indicating strong generalization ability.

### ****3. Actionable Insights****

From both model importance and domain logic, the following insights were drawn:

|  |  |  |
| --- | --- | --- |
| **Insight** | **Description** | **Implication** |
| **Healthcare Investment** | Higher percentage expenditure and GDP correlate with higher life expectancy | Countries should allocate more resources towards health infrastructure |
| **Immunization Programs** | Higher Immunization Score (Polio, Hepatitis B, Diphtheria) linked to better outcomes | Improve immunization coverage, especially in developing countries |
| **Education and Income** | Higher Schooling and Income Composition directly impact longevity | Education and income equality programs should be prioritized |
| **HIV and Child Mortality** | HIV/AIDS and under-five deaths are strong negative predictors | Strengthen maternal and reproductive health programs |

### ****4. Recommendations****

Based on the data and modelling results, the following recommendations can be made for governments, health organizations, and policy makers:

1. **Invest Heavily in Healthcare Systems**  
   Increase national health expenditure, especially in low-income countries, to improve accessibility and quality of care.
2. **Expand Immunization Programs**  
   Ensure full vaccination coverage, especially for preventable diseases like Hepatitis B, Polio, and Diphtheria.
3. **Promote Educational Equity**  
   Support universal primary and secondary education as a long-term investment in public health outcomes.
4. **Target High-Risk Areas**  
   Direct interventions to reduce infant mortality and control HIV prevalence in vulnerable populations.
5. **Use Predictive Modelling for Early Policy Planning**  
   Deploy machine learning models like Gradient Boosting to forecast life expectancy trends and evaluate potential policy interventions.
6. **Lay Groundwork for Monitoring & Evaluation**

**Based on the measurable outcomes,** the government should devise means on how these outcomes can be tracked over time to evaluate the success of interventions.

1. **Facilitation of Comparative Benchmarking**

Enable comparison between countries or regions to assess where a nation stands and what strategies might be adopted from better-performing counterparts.

1. **Drive Targeted Public Health Campaigns**

* Inform design of programs focused on education, disease prevention, and access to healthcare, especially for at-risk populations.

### ****5. Summary****

This project explored a comprehensive life expectancy dataset across countries, conducted extensive feature engineering, and built predictive models to understand and forecast life expectancy outcomes.

* The best-performing model (Gradient Boosting) achieved an R² of **0.9240**, with low error and high consistency.
* The analysis revealed **socioeconomic status, healthcare spending, immunization,** and **education** as key drivers of life expectancy.
* The project culminates in **practical, data-driven recommendations** that can guide health and development policies at national and international levels.