Life Expectancy Analysis

Context:

Although there have been a lot of studies undertaken in the past on factors affecting life expectancy, considering demographic variables, income composition, and mortality rates. It was found that effect of immunization and human development index was not taken into account in the past. Also, some of the past research was done considering multiple linear regression based on a dataset of one year for all the countries. Hence, this gives motivation to resolve both the factors stated previously by formulating a regression model based on ma ixed effects model and multiple linear regression while considering data from the period 2000 to 2015 for all the countries. Important immunizations like Hepatitis B, Polio, and Diphtheria will also be considered. In a nutshell, this study will focus on immunization factors, mortality factors, economic factors, social factor,s and other health-related factors as well. Since the observations in this dataset are based on different countries, it will be easier for a country to determine the predicting factor that contributes to a lower value of life expectancy. This will help in suggesting a country which area should be given importance in order to efficiently improve the life expectancy of its population.

Content:

The project relies on the accuracy of data. The Global Health Observatory (GHO) data repository under the World Health Organization (WHO) keeps track of the health status as well as many other related factors for all countries. The data sets are made available to the public for the purpose of health data analysis. The dataset related to life expectancy, health factors for 193 countries has been collected from the same WHO data repository website, and its corresponding economic data was collected from the United Nations website. Among all categories of health-related factors, only those critical factors were chosen that are more representative. It has been observed that in the past 15 years, there has been a huge development in hthe ealth sector resulting in improvement of human mortality rates, especially in the developing nations, in comparison to the past 30 years. Therefore, in this projec,t we have considered data from years 2000-2015 for 193 countries for further analysis. The individual data files have been merged together into a single dataset. On initial visual inspection of the data showed some missing values. As the datasets were from WHO, we found no evident errors. Missing data was handled in R software by using Mthe issmap command. The result indicated that most of the missing data was for population, Hepatitis B, and GDP. The missing data were from less-known countries like Vanuatu, Tonga, Togo, Cabo Verde, etc. Finding all data for these countries was difficul,t and hence, it was decided that we would exclude these countries from the final model dataset. The final merged file(final dataset) consists of 22 Columns and 2938 rows, which meant 20 predicting variables. All predicting variables were then divided into several broad categories: Immunization-related factors, Mortality factors, Economic factors, and Social factors.

Acknowledgements:

The data was collected from the WHO and the United Nations websites with the help of Deeksha Russell and Duan Wang.

Inspiration:

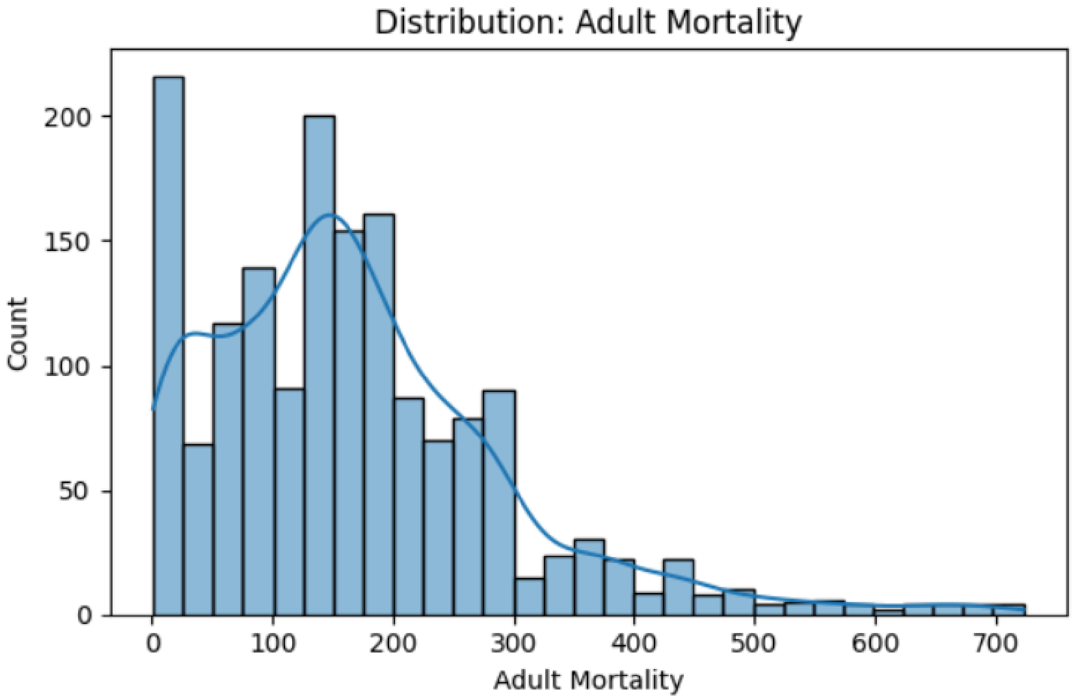
The dataset aims to answer the following key questions:

1. Do the various predicting factors that have been chosen initially really affect the Life expectancy? What are the predicting variables actually affecting life expectancy?
2. Should a country having a lower life expectancy value(<65) increase its healthcare expenditure in order to improve its average lifespan?
3. How do Infant and Adult mortality rates affect life expectancy?
4. Does Life Expectancy havea positive or negative correlation with eating habits, lifestyle, exercise, smoking, drinking alcohol, etc.
5. What is the impact of schooling on the lifespan of humans?
6. Does Life Expectancy have a positive or negative relationship with drinking alcohol?
7. Do densely populated countries tend to have lower life expectancy?
8. What is the impact of Immunization coverage on life Expectancy?

Analyzing life expectancy data involves examining various factors that may affect life expectancy rates in different countries. For a finance analyst project, we can approach this analysis by looking at economic indicators, healthcare spending, and other socio-economic factors that might influence life expectancy. Here, I'll outline a basic life expectancy analysis project using Python, incorporating data collection, cleaning, and EDA (Exploratory Data Analysis), and some basic statistical analysis.

1) Exploratory Data Analysis (Observations)

1.1 Distribution of Adult Mortality:



This picture displays the distribution of Adult Mortality across countries.

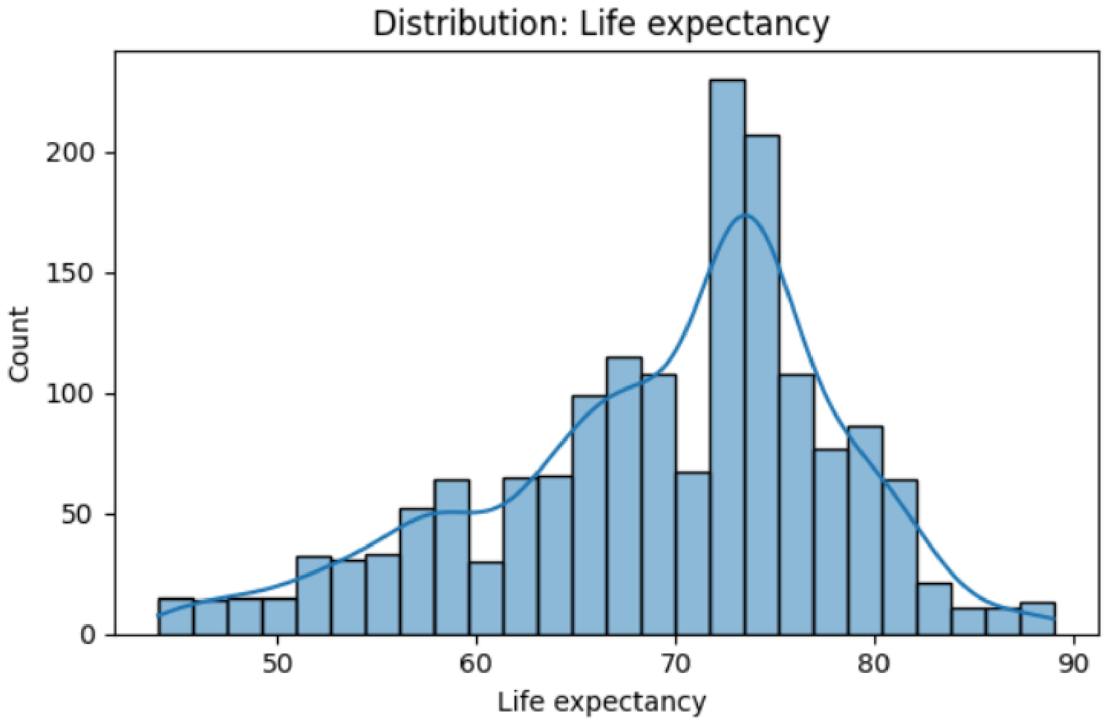
Observations:

- The distribution is right-skewed, with most countries having adult mortality rates below 300.
- A prominent peak appears between 50 and 150, followed by a steady decline.
- A small number of countries have very high adult mortality, exceeding 600.

Insight:

Most countries exhibit relatively low adult mortality, though the long tail indicates that certain nations still face significant public health challenges.

1.2 Life Expectancy:



This picture displays the distribution of Life Expectancy across countries.

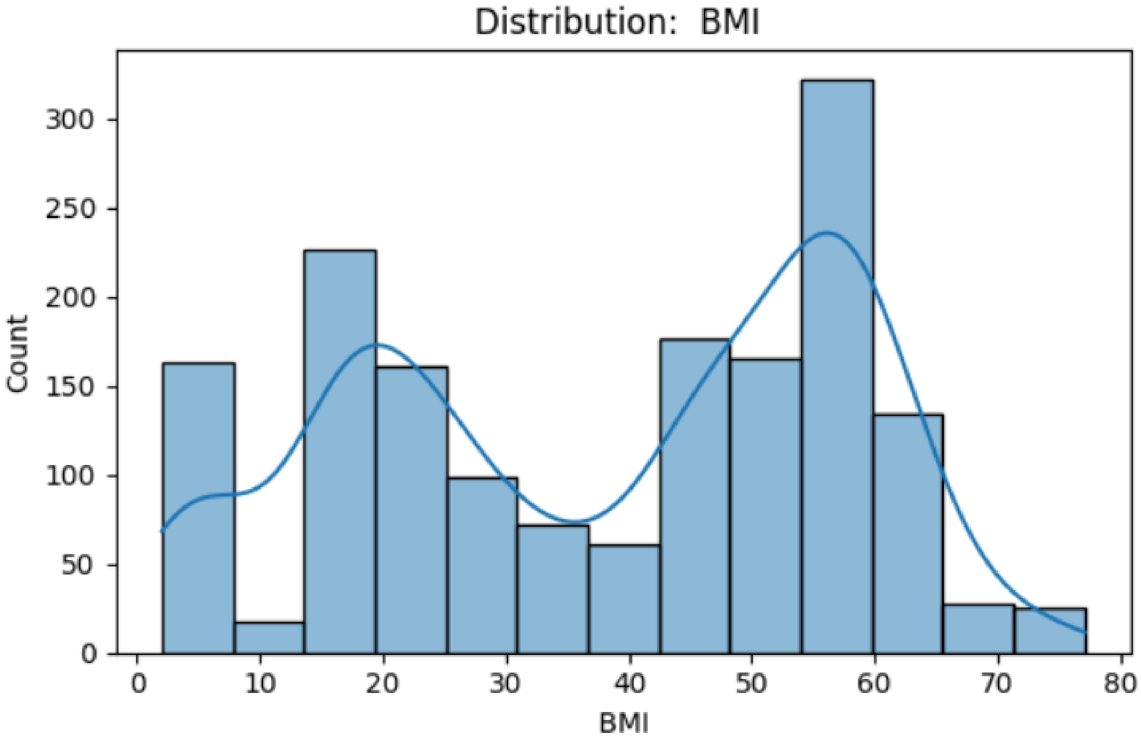
Observations:

- The data is nearly normally distributed, centering around 70–75 years.
- Most countries fall within the 60 to 80-year life expectancy range.
- Very few countries lie at the extremes (below 50 or above 85 years).

Insight:

Global life expectancy tends to cluster in the 70s, reflecting improving living conditions and healthcare access across the majority of nations.

1.3 BMI:



This picture displays the distribution of average Body Mass Index (BMI) across countries.

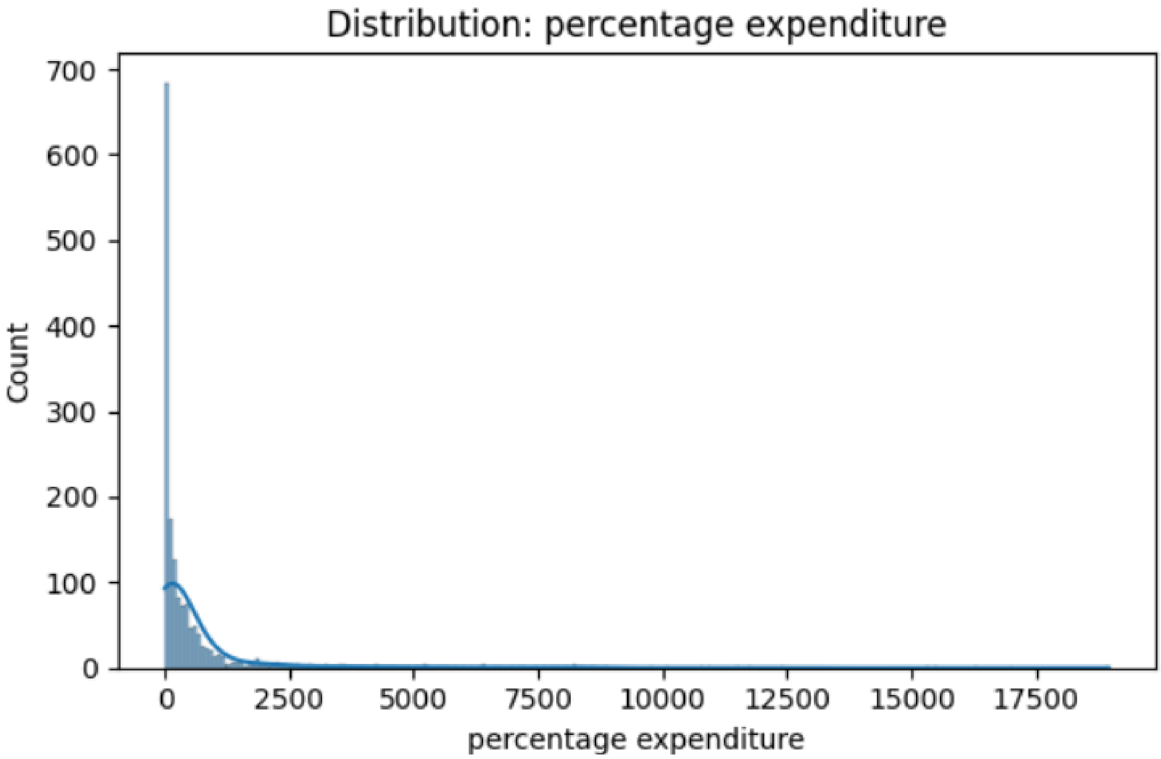
Observations:

- The distribution appears bimodal, with peaks around 20 and again around 55.
- There is wide variation in BMI levels across countries.
- Some countries show notably high BMI averages, possibly above 60.

Insight:

BMI patterns suggest a divide between undernourished and overnourished populations, likely reflecting disparities between developing and developed nations.

1.4 Percentage Expenditure:



This picture displays the distribution of percentage health expenditure (as part of GDP) across countries.

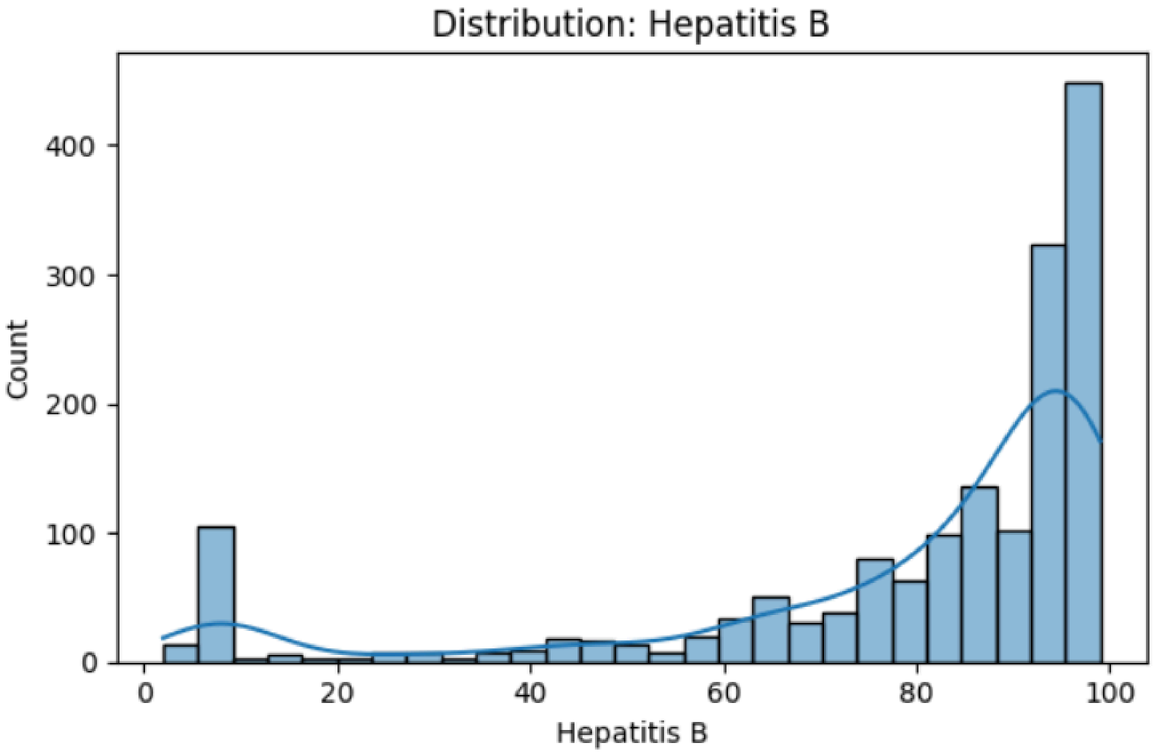
Observations:

- The distribution is highly right-skewed, with most countries spending under 2000%.
- A few outliers report extremely high expenditures, exceeding 10,000%.
- The majority of values cluster close to the lower end of the axis.

Insight:

Health expenditure levels vary widely, with a small number of countries showing unusually high spending, possibly due to specific economic or health crises.

1.5 Hepatitis B:



This picture displays the distribution of Hepatitis B immunization coverage (% of 1-year-olds) across countries.

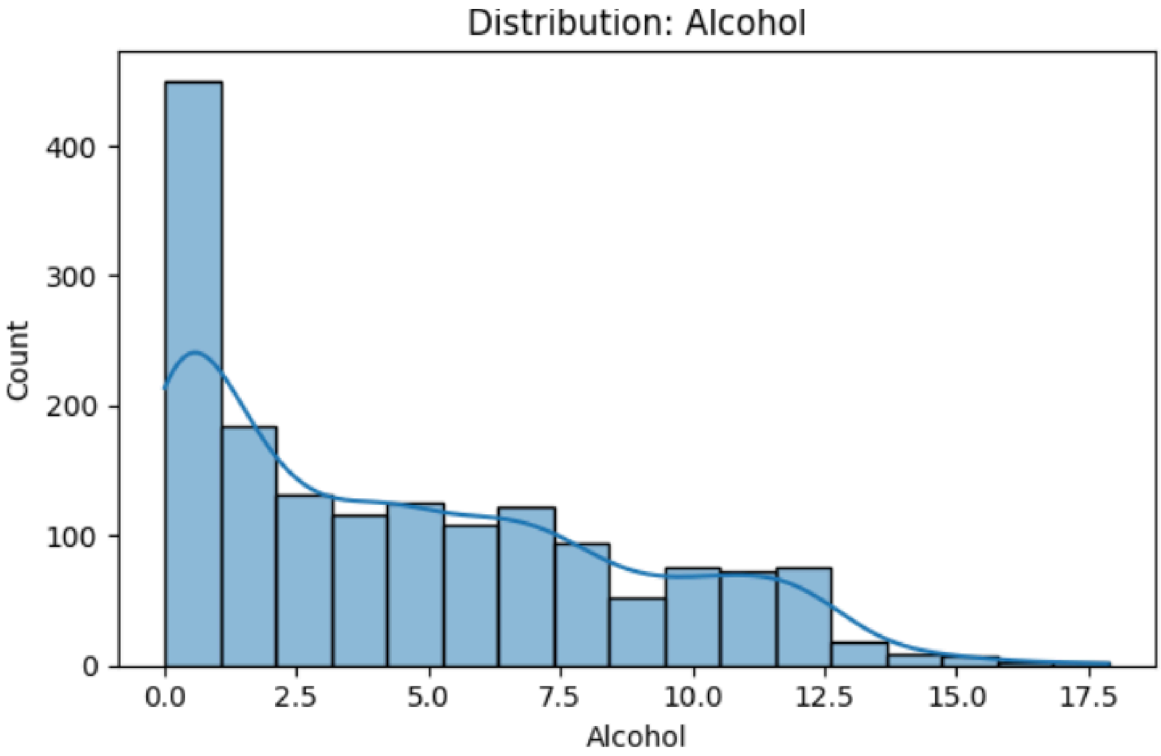
Observations:

- Most countries report high immunization coverage, particularly above 80%.
- A sharp peak occurs at 100%, indicating full coverage in many countries.
- A smaller group of countries shows significantly lower vaccination rates.

Insight:

While global Hepatitis B immunization rates are strong overall, a small number of countries still lag behind in ensuring universal vaccine coverage.

1.6 Alcohol:



This picture displays the distribution of Alcohol consumption per capita (in liters) across countries.

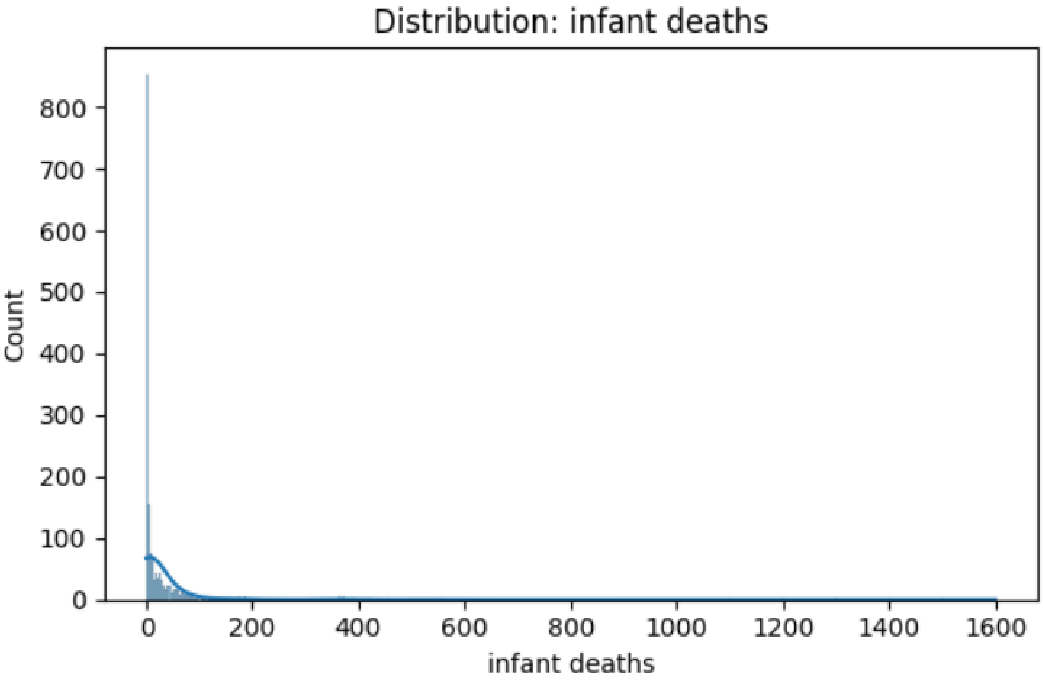
Observations:

- The data is right-skewed, with the majority of countries consuming fewer than 5 liters per capita.
- A gradual decline follows, with some countries reaching up to 17 liters.
- The KDE curve shows a sharp drop-off after initial peak consumption levels.

Insight:

Most countries maintain low to moderate alcohol consumption, though a few outliers drive the global average upward due to significantly higher intake.

1.7 Infant Deaths:



This picture displays the distribution of Infant Deaths (per 1000 live births) across countries.

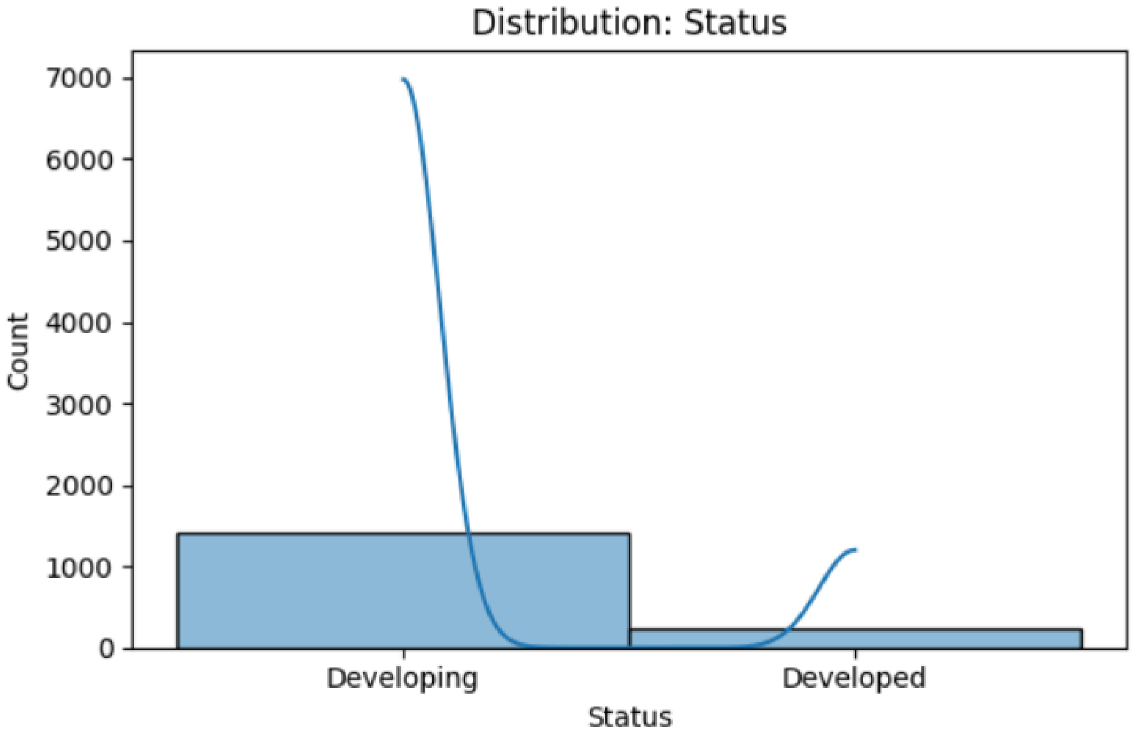
Observations:

- The distribution is extremely right-skewed, with most countries reporting fewer than 100 infant deaths.
- A long tail stretches out to over 1500 deaths in some countries.
- The KDE indicates a steep decline in frequency as mortality increases.

Insight:

While many countries have successfully reduced infant deaths, others still face serious infant health issues, highlighting ongoing inequalities in child care access.

1.8 Status:



This picture displays the distribution of country classification as Developing or Developed.

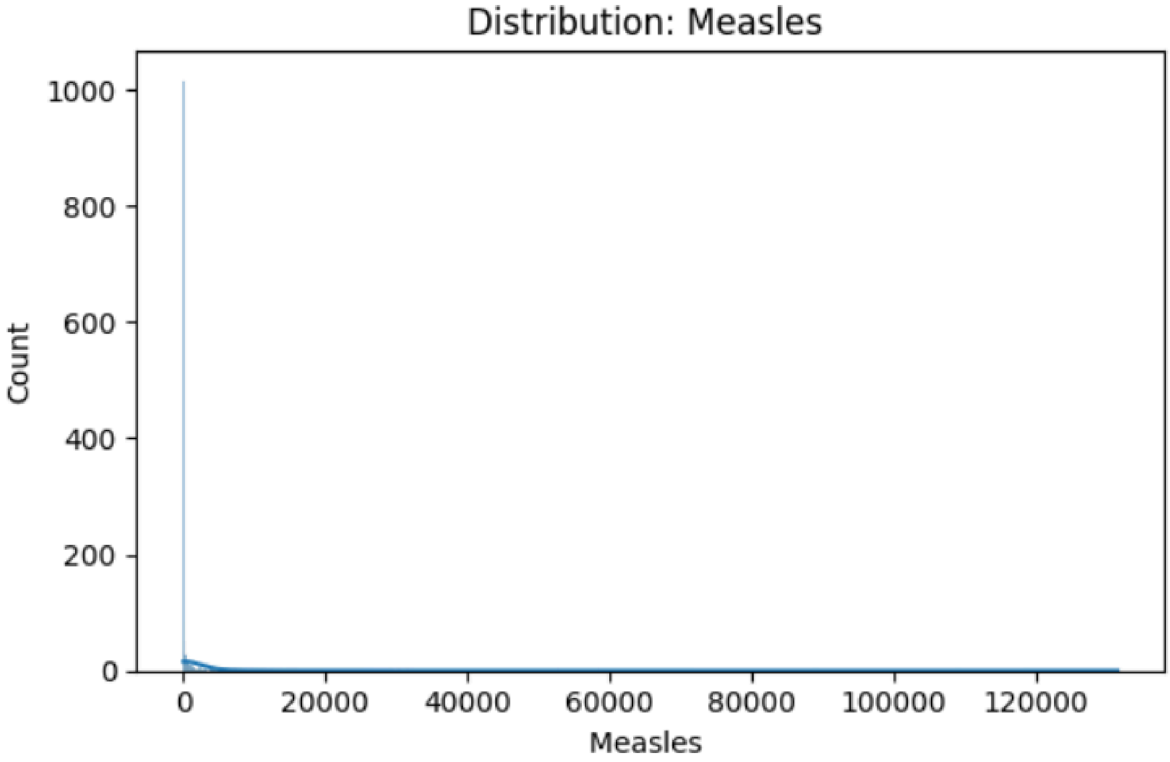
Observations:

- A large majority of the data represents developing countries.
- Developed countries form a smaller portion of the dataset.
- KDE is not meaningful due to the categorical nature of the variable.

Insight:

The dataset is heavily skewed toward developing nations, which may influence the overall analysis of global health trends.

1.9 Measles:



This picture displays the distribution of reported Measles cases across countries.

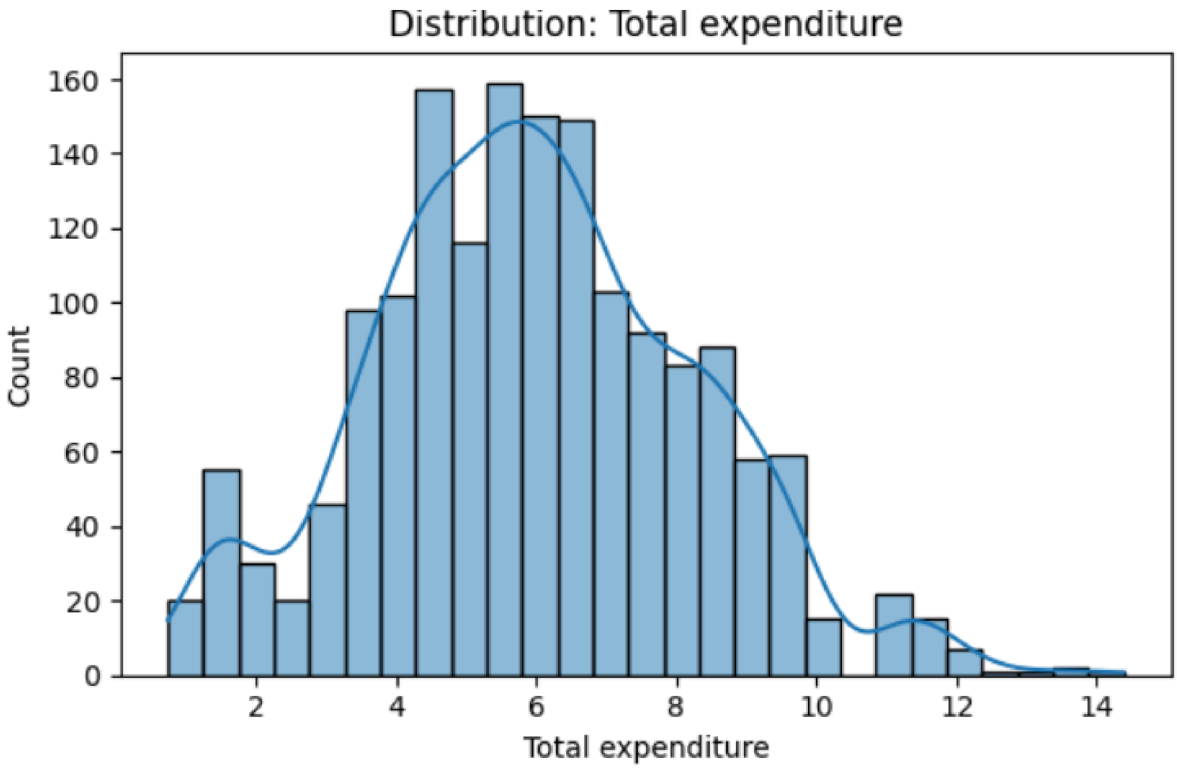
Observations:

- The distribution is highly right-skewed, with most countries reporting low measles incidence.
- Some extreme outliers report over 100,000 cases.
- The vast majority of countries fall well below 10,000 cases.

Insight:

Measles outbreaks are rare in most countries but persist in certain areas, signaling gaps in vaccination or outbreak management.

1.10 Total Expenditure:



This picture displays the distribution of Total Expenditure.

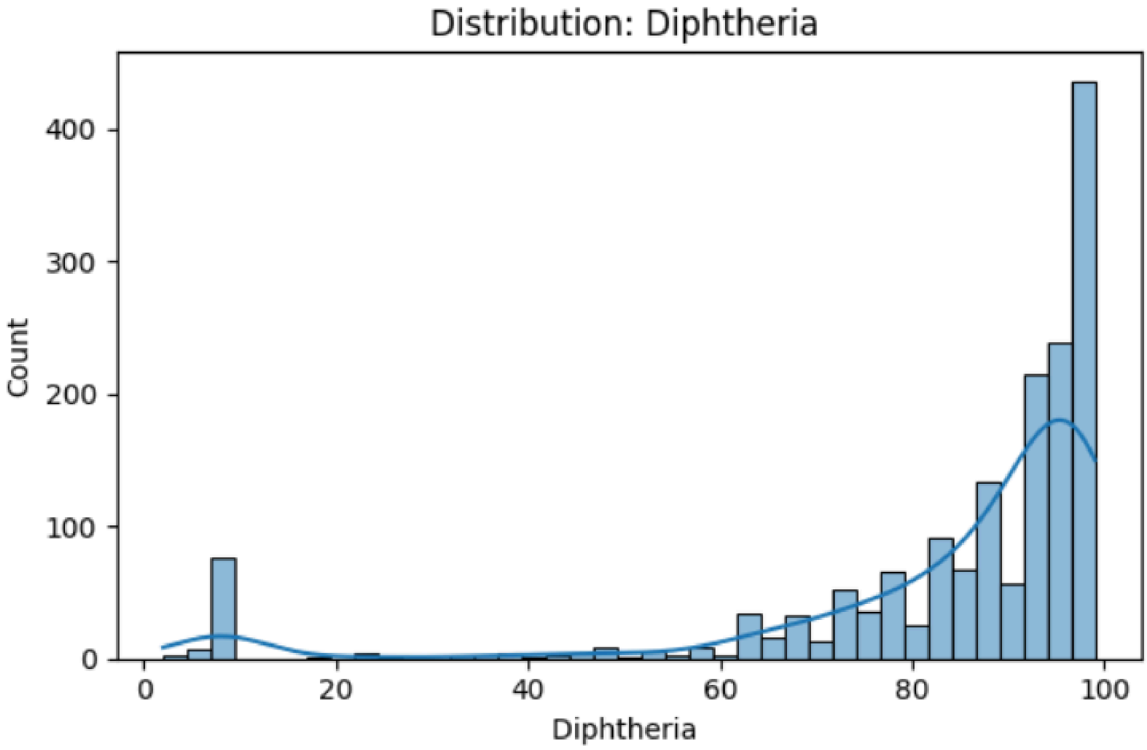
Observations:

- The distribution is roughly bell-shaped, resembling a normal distribution.
- Most values fall between 4 and 8, with a peak around 5 to 6.
- Few data points lie outside the 2 to 10 range, indicating less variation in extreme values.

Insight:

Total expenditure tends to cluster around a central value, suggesting that spending is relatively consistent across most observations, with limited outliers.

1.11 Diphtheria vaccination rates:



This picture displays the distribution of Diphtheria vaccination rates.

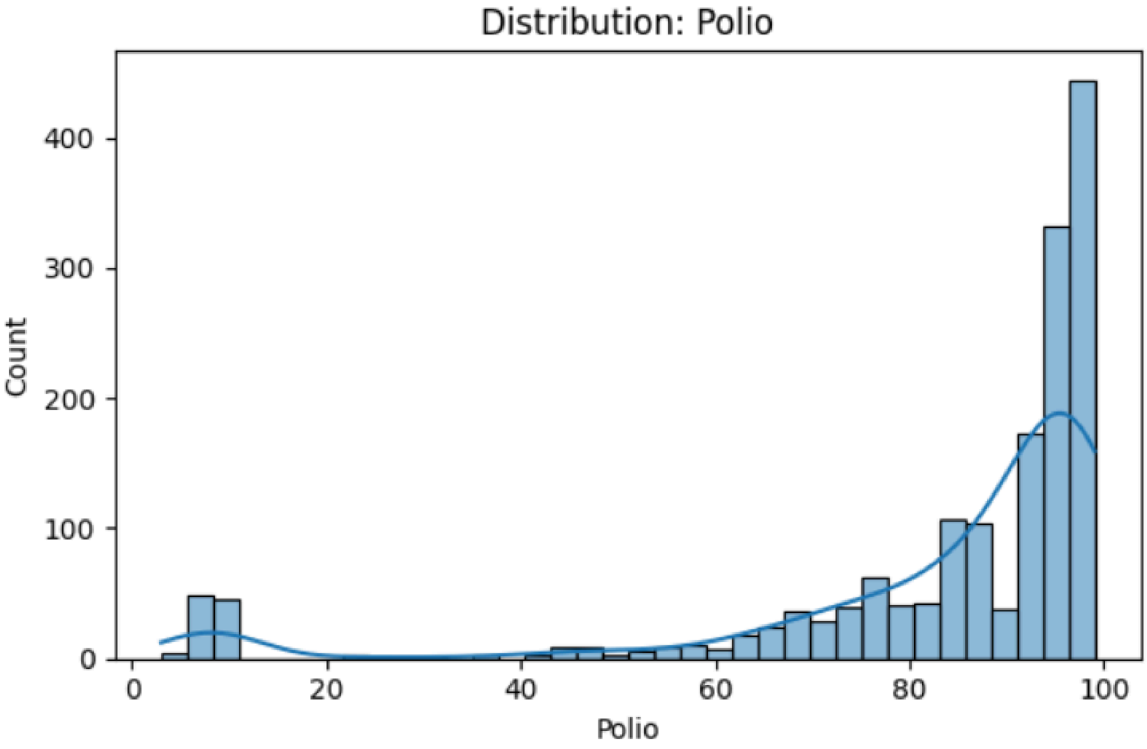
Observations:

- The distribution is heavily right-skewed, with most countries having high vaccination rates (above 80%).
- A significant peak exists near the 100% mark.
- There are some outliers with low vaccination rates, but they are few.

Insight:

Most populations have achieved high diphtheria vaccination coverage, indicating effective public health campaigns, though a few countries may need targeted interventions.

1.12 Polio vaccination rates:



This picture displays the distribution of Polio vaccination rates.

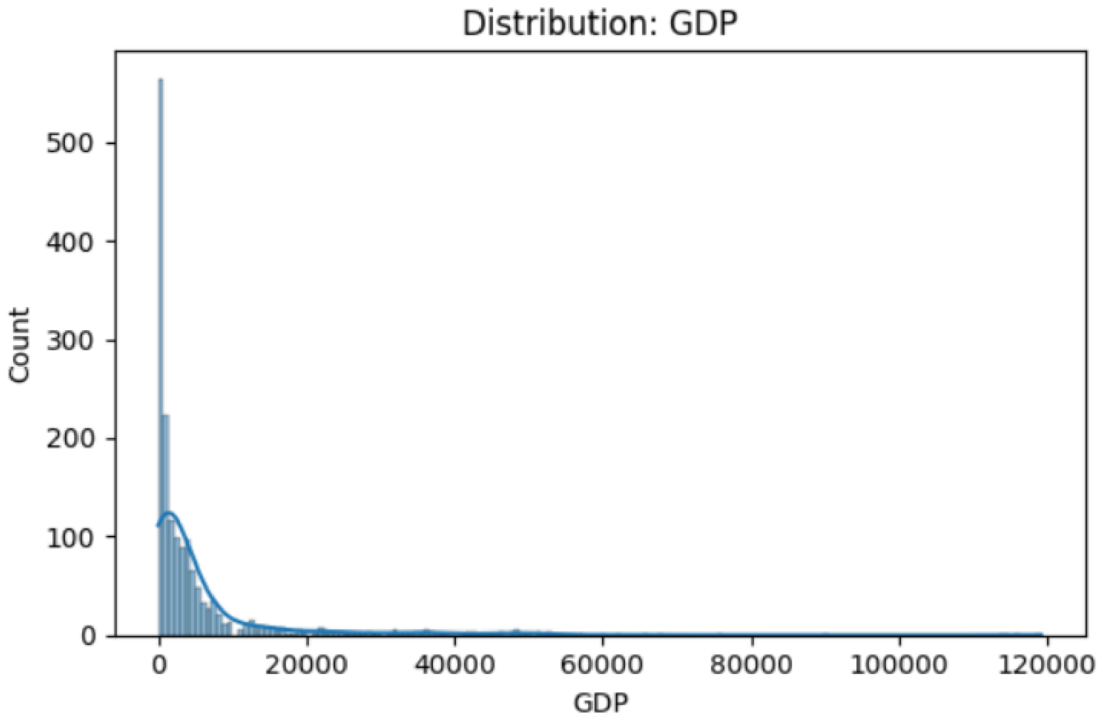
Observations:

- The distribution is similar to diphtheria: right-skewed with a strong peak near 100%.
- A few countries show significantly lower vaccination rates, with a long tail toward lower values.
- Most data is concentrated between 80% and 100%.

Insight:

Global polio vaccination efforts appear widely successful, though disparities remain in some regions that could benefit from enhanced outreach and access.

1.13 GDP:



This picture displays the distribution of GDP.

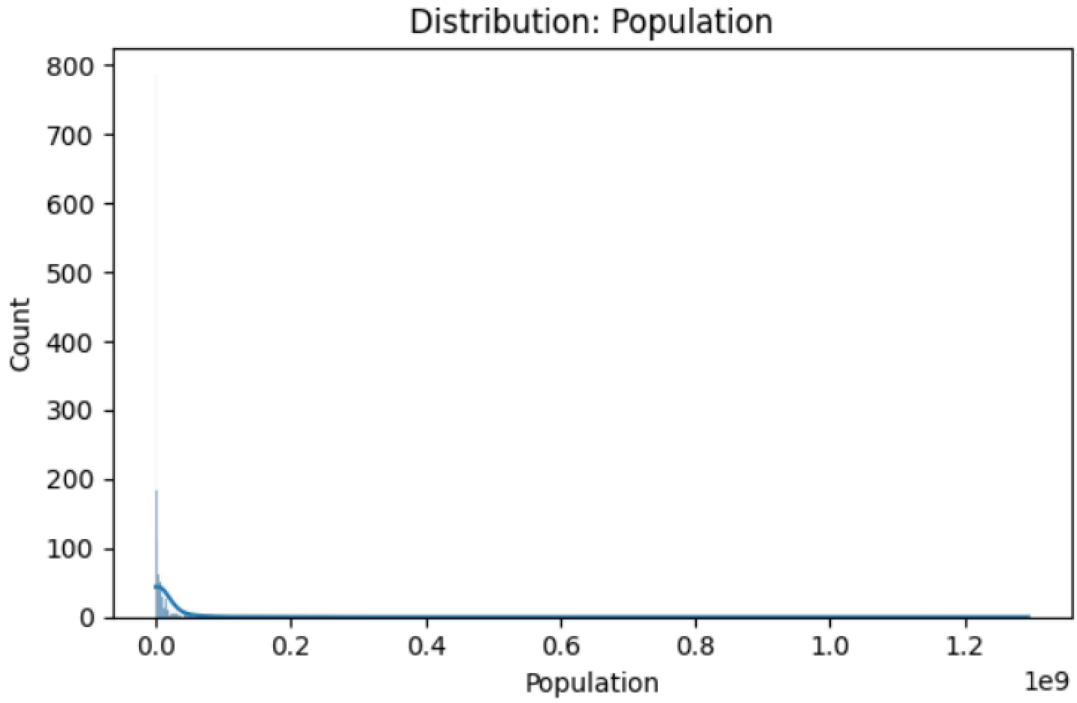
Observations:

- The distribution is extremely right-skewed with a very high peak near the lower end.
- Most countries have GDP values under 20,000, but a long tail extends to over 100,000.
- Very few countries have exceptionally high GDPs.

Insight:

Global GDP distribution is highly uneven, with most nations having relatively low GDPs and a small number of highly wealthy countries skewing the upper range.

1.14 Population:



This picture displays the distribution of Population.

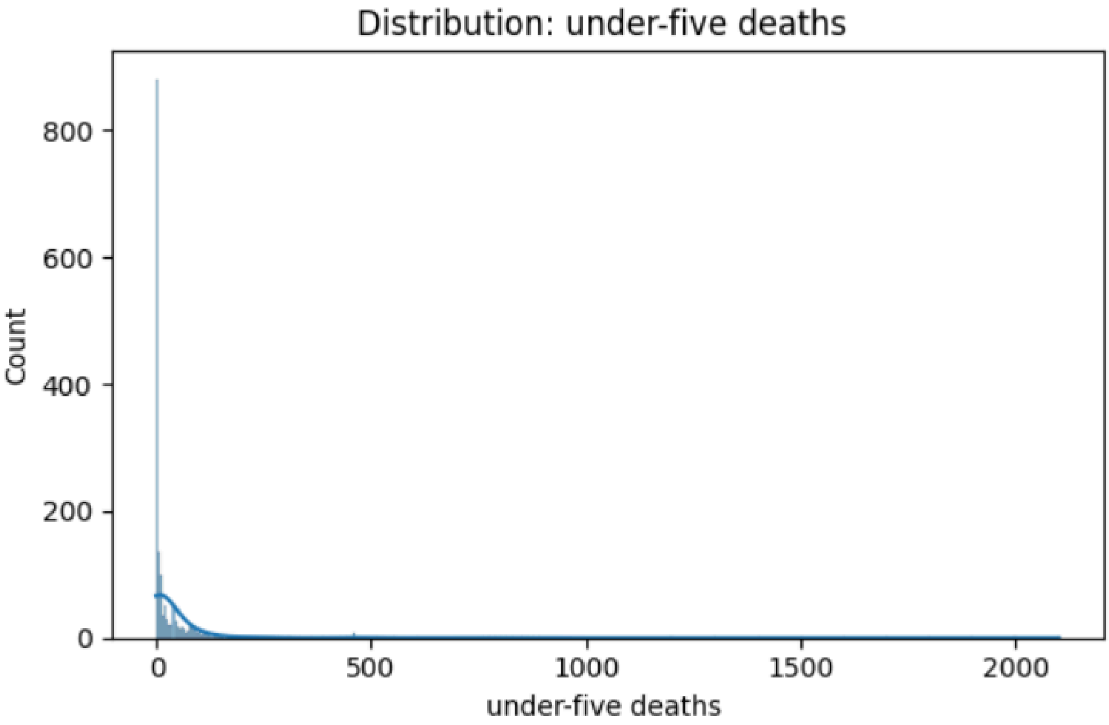
Observations:

- The data is heavily skewed to the right.
- Most countries have relatively low population values, but a few have populations in the hundreds of millions to billions.
- There is a sharp peak at the lower end.

Insight:

A small number of countries account for a large portion of the world’s population, while most have significantly smaller populations.

1.15 Under-Five Deaths:



This picture displays the distribution of Under-Five Deaths.

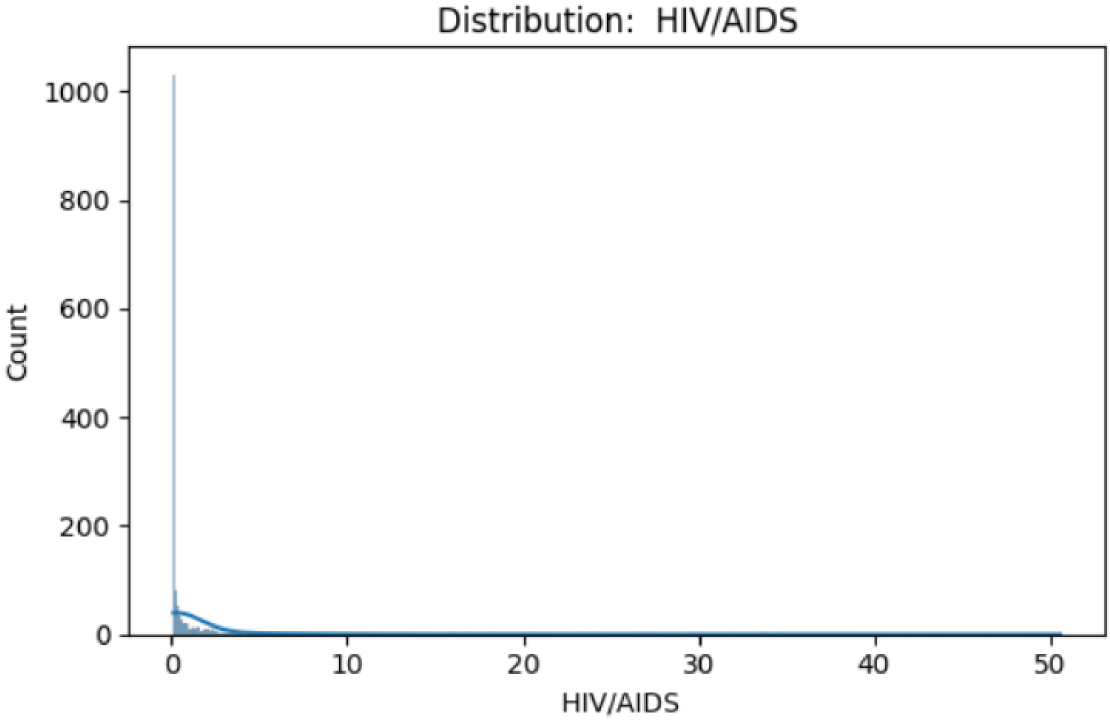
Observations:

- The distribution is heavily right-skewed.
- Most countries report low under-five death counts, but a small number have very high counts.
- A sharp peak is visible near the lower values, indicating fewer deaths in the majority of cases.

Insight:

Under-five mortality is concentrated in a limited number of countries, suggesting localized health crises or disparities in child healthcare systems.

1.16 HIV/AIDS death rates:



This picture displays the distribution of HIV/AIDS death rates.

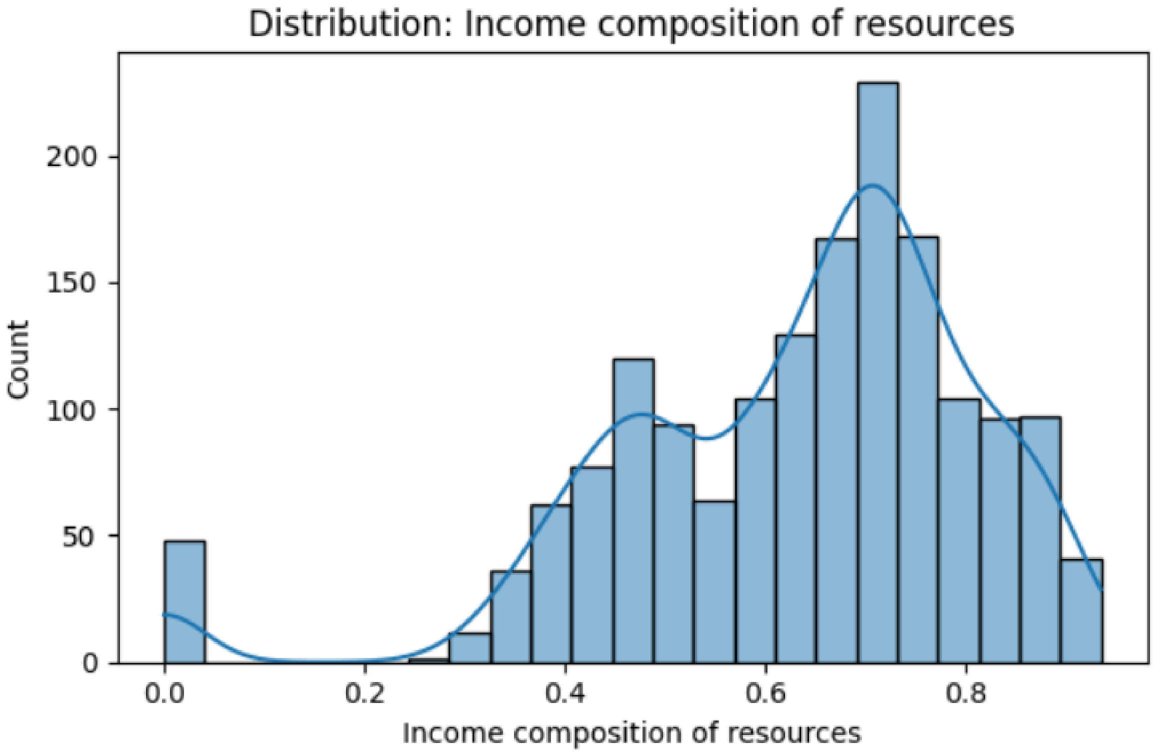
Observations:

- The distribution is extremely right-skewed with a peak close to zero.
- Most countries report very low or no HIV/AIDS deaths, while a few show significantly higher numbers.
- The long tail indicates rare but severe impacts in some regions.

Insight:

HIV/AIDS is no longer a widespread global killer, but still poses a major health challenge in select countries, highlighting the need for targeted disease control programs.

1.17 Income Composition of Resources:



This picture displays the distribution of Income Composition of Resources.

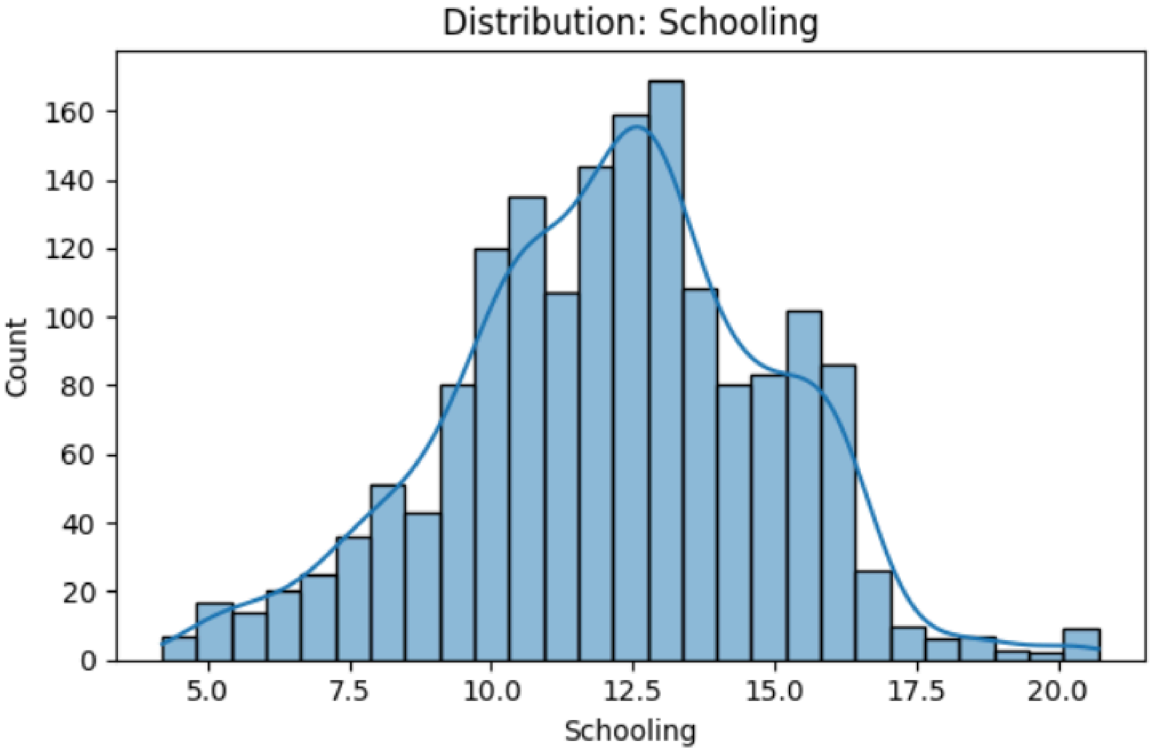
Observations:

- The distribution is moderately right-skewed, with most values concentrated between 0.4 and 0.8.
- A peak appears around 0.7, suggesting many observations have high income resource composition.
- A small spike at 0 indicates that some countries may have extremely poor income resource access.

Insight:

While most regions have moderate to high income composition of resources, some still face severe income inequality or lack access to income-generating resources, highlighting potential targets for economic aid or reform.

1.18 Schooling:



This picture displays the distribution of Schooling.

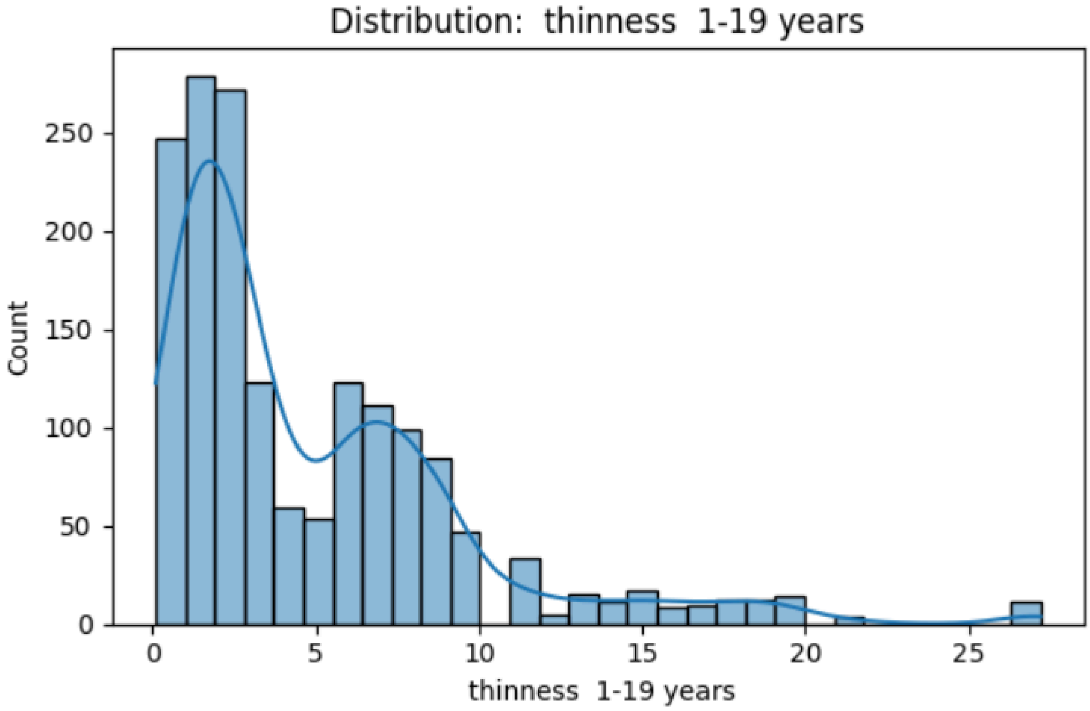
Observations:

- The data forms a bell-shaped curve, resembling a normal distribution.
- Most observations fall between 10 and 14 years of schooling, with a peak around 12–13 years.
- Very few countries have average schooling below 6 or above 18 years.

Insight:

Globally, average years of schooling are centered around secondary education, indicating progress in education systems but also potential for improvement in countries on the lower end of the spectrum.

1.19 Distribution of Thinness (1–19 years):



This picture displays the distribution of Thinness (1–19 years).

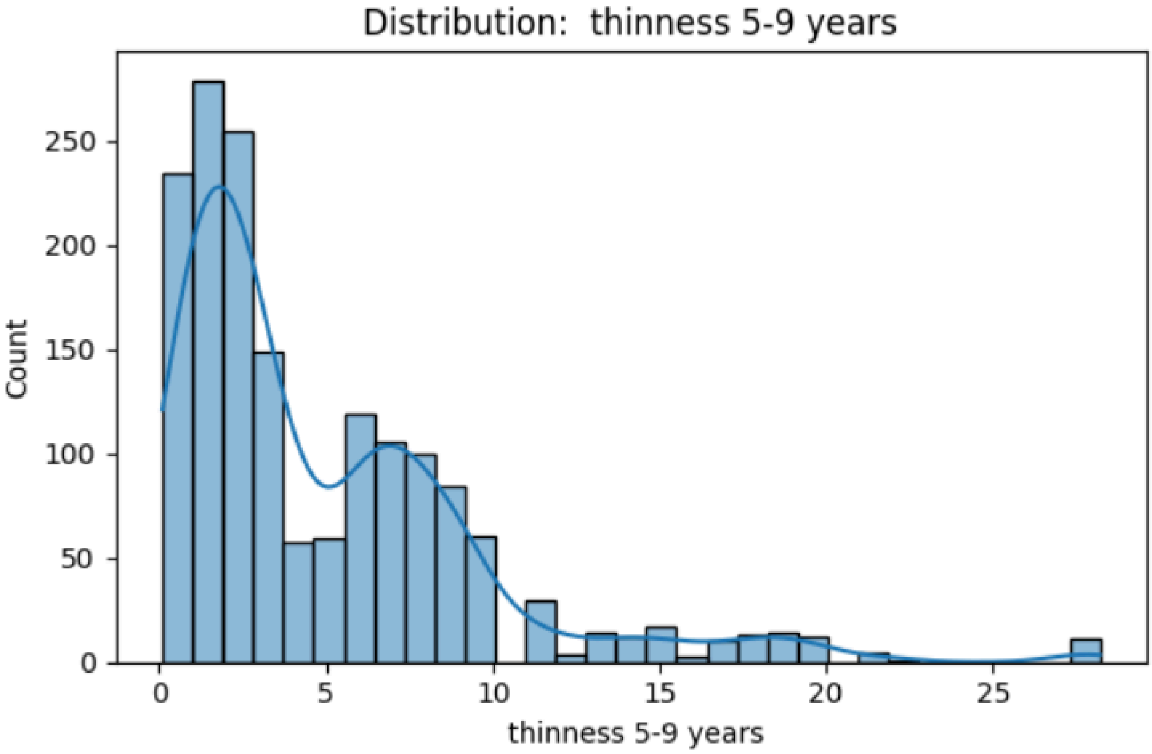
Observations:

- The distribution is highly right-skewed, with the majority of values between 0 and 5.
- A large concentration occurs close to 0–2%, indicating low thinness prevalence in most cases.
- A long tail shows that some populations have significantly higher thinness rates, up to 25%.

Insight:

Most regions report low levels of thinness in the 1–19 age group, but a subset still suffers from high rates of malnutrition or undernourishment among youth, requiring targeted nutritional interventions.

1.20 Distribution of Thinness (5–9 years):



This picture displays the distribution of Thinness (5–9 years).

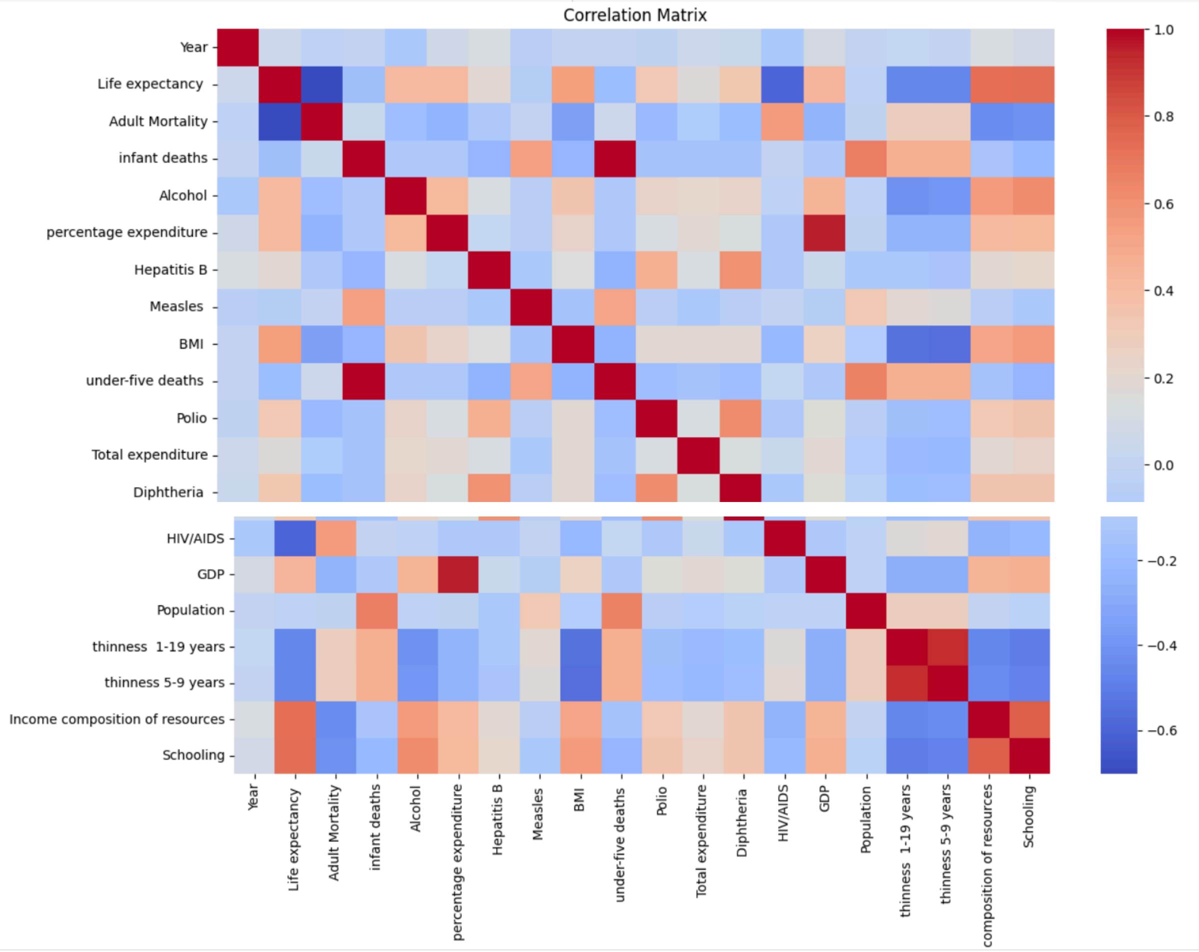
Observations:

- The shape is very similar to the previous plot, being strongly right-skewed.
- The highest counts occur between 0 and 3%, indicating low thinness prevalence in most children aged 5–9.
- A small number of countries have alarmingly high thinness rates, above 15–20%.

Insight:

Thinness among children aged 5–9 is generally under control globally, but certain regions still exhibit serious undernutrition challenges in this age group, signaling a need for focused public health programs.

2) Correlation Matrix of various health, economic, and demographic variables



This picture displays the Correlation Matrix of various health, economic, and demographic variables.

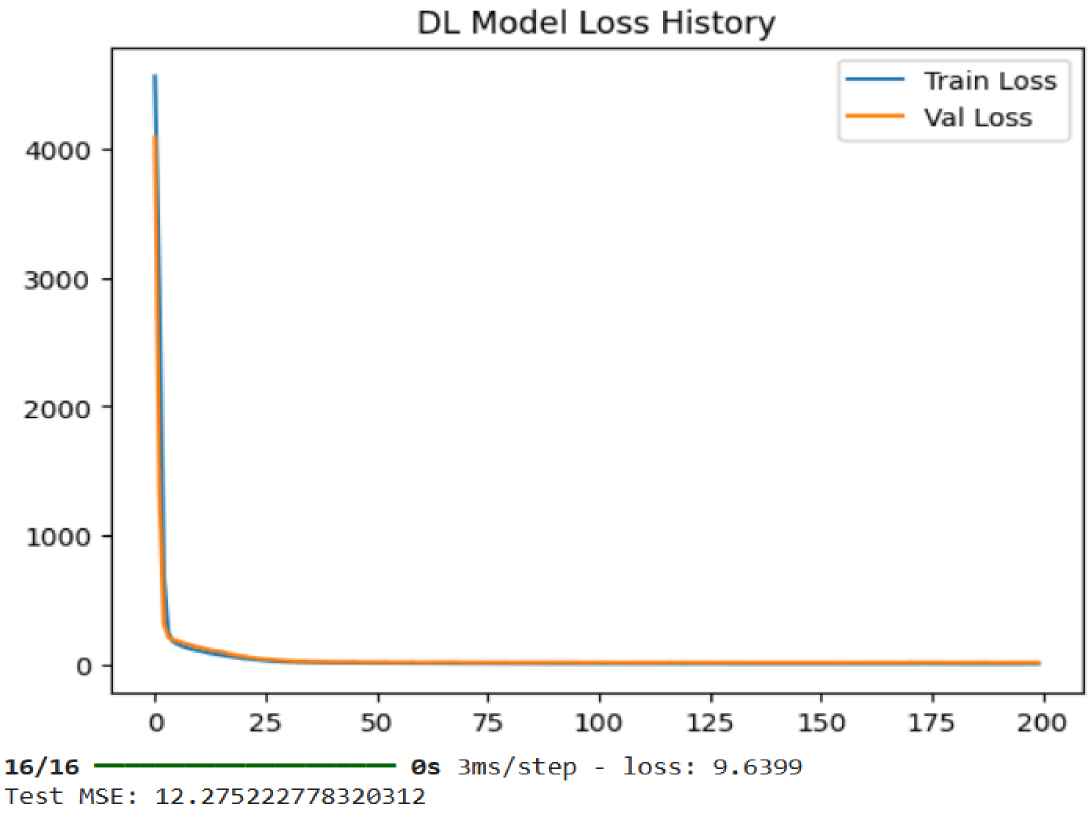
Observations:

- Life expectancy shows strong positive correlation with schooling, income composition of resources, and total expenditure, and a negative correlation with HIV/AIDS, infant deaths, and adult mortality.
- Infant deaths and under-five deaths are highly positively correlated, indicating overlapping health outcomes.
- HIV/AIDS is negatively correlated with life expectancy, schooling, and GDP, suggesting regions with high HIV prevalence tend to have poorer outcomes.
- GDP positively correlates with schooling and income composition, and negatively with thinness and mortality indicators.
- Thinness (1–19 and 5–9 years) shows strong negative correlation with income composition of resources and schooling, indicating undernutrition is more common in resource-poor regions.

Insight:

Life expectancy and general health outcomes are positively influenced by factors such as education, economic prosperity, and resource availability. Conversely, high mortality, malnutrition, and disease rates are strongly tied to poverty and poor access to healthcare and education. These insights underscore the interdependence of socio-economic and health indicators in global development.

3) Deep Learning Model Performance and Loss Analysis



```
16/16 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - loss: 9.6399
Test MSE: 12.275222778320312
```

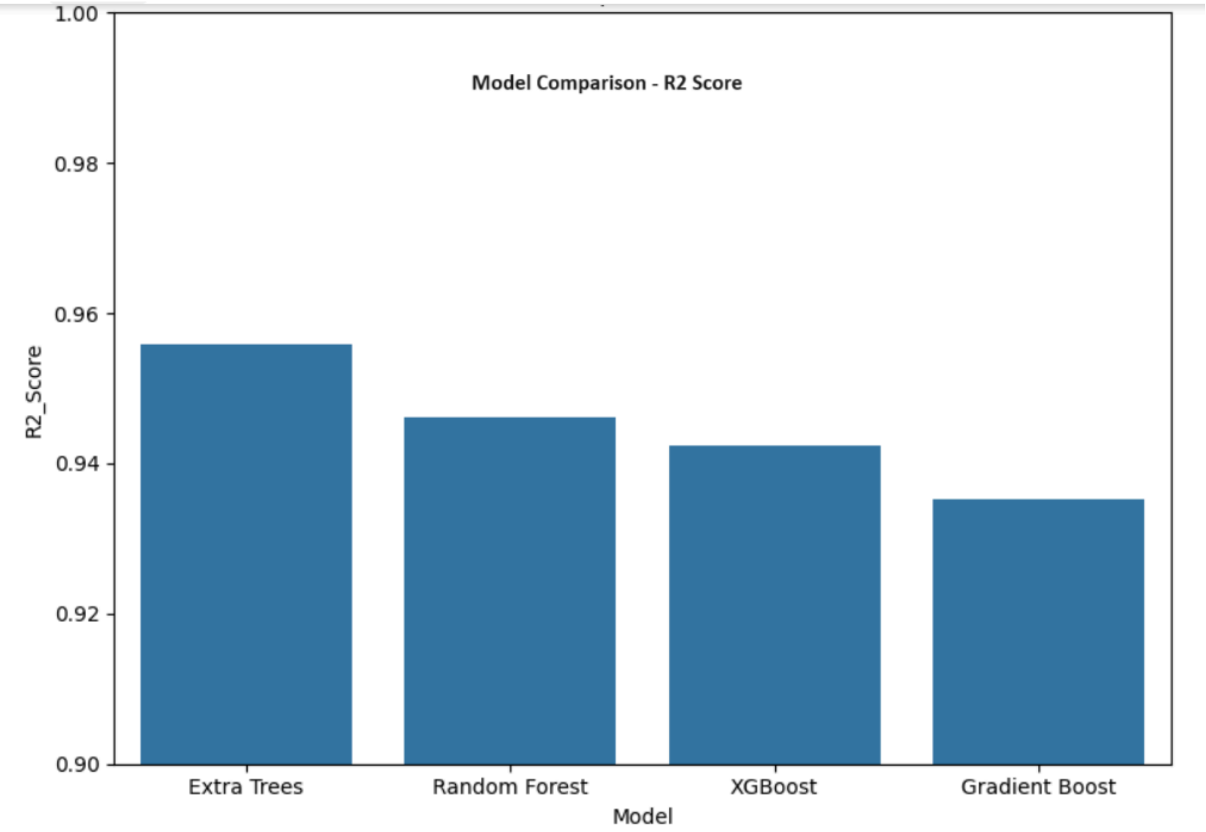
Explanation:

The graph tracks the training and validation loss (error) over 200 epochs for a deep learning regression model. The goal is to minimize these losses to improve predictive accuracy.

Insight:

- Both **training and validation loss decrease rapidly** and plateau at similar levels, indicating effective learning without overfitting.
- The final **test MSE is 12.28**, confirming the model performs well on unseen data.
- This suggests that deep learning, when properly tuned, can be a robust alternative to ensemble tree-based models.

4) Comparison of Different Machine Learning Models

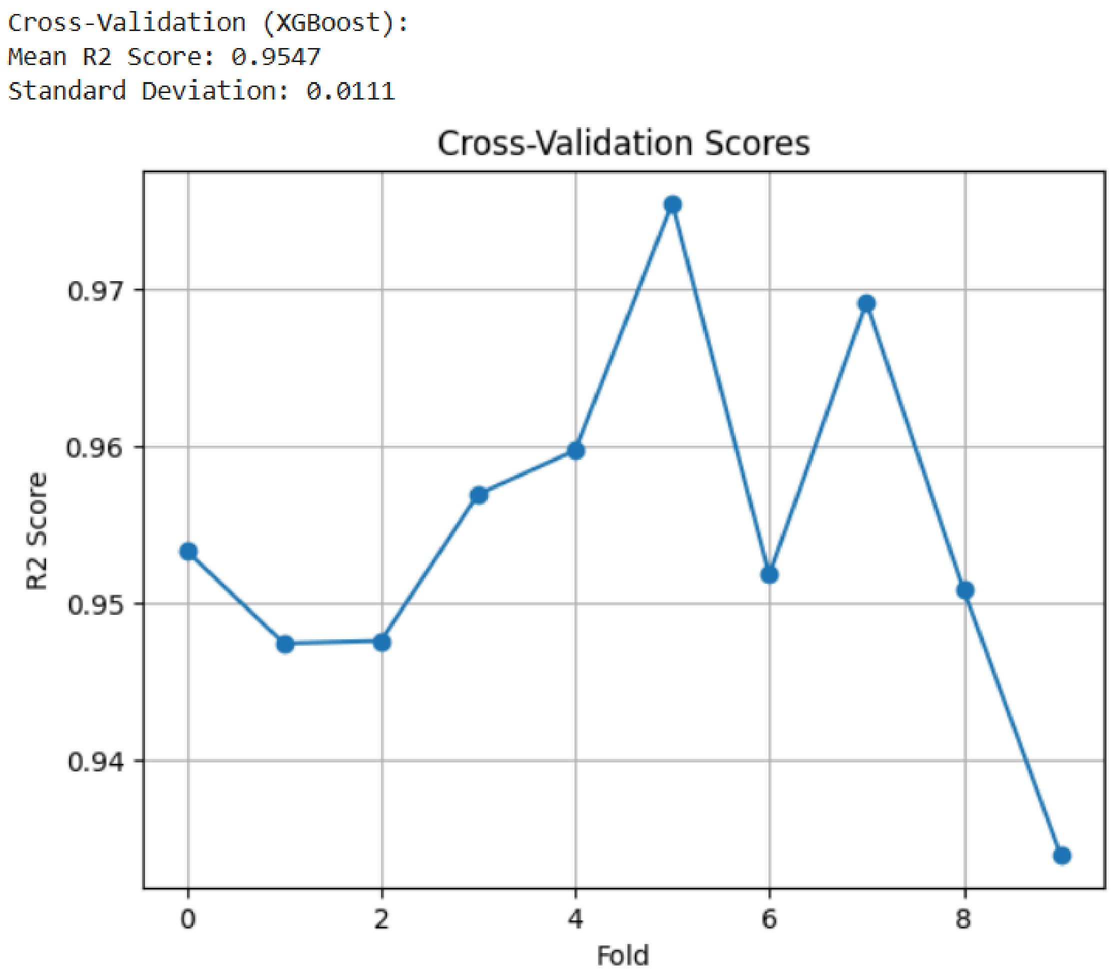


Explanation:
This bar chart compares the performance of four different machine learning models using their R² scores on the test dataset.

Insight:

- **Extra Trees** is the best-performing model with the highest R², followed closely by **Random Forest**, **XGBoost**, and **Gradient Boost**.
- All models demonstrate strong performance (**R² > 0.93**), indicating the feature set and data quality are highly predictive.
- This comparison helps select the best model for production use or further fine-tuning.

5) XGBoost Model Evaluation via Cross-Validation



```
Cross-Validation (XGBoost):
Mean R2 Score: 0.9547
Standard Deviation: 0.0111
```

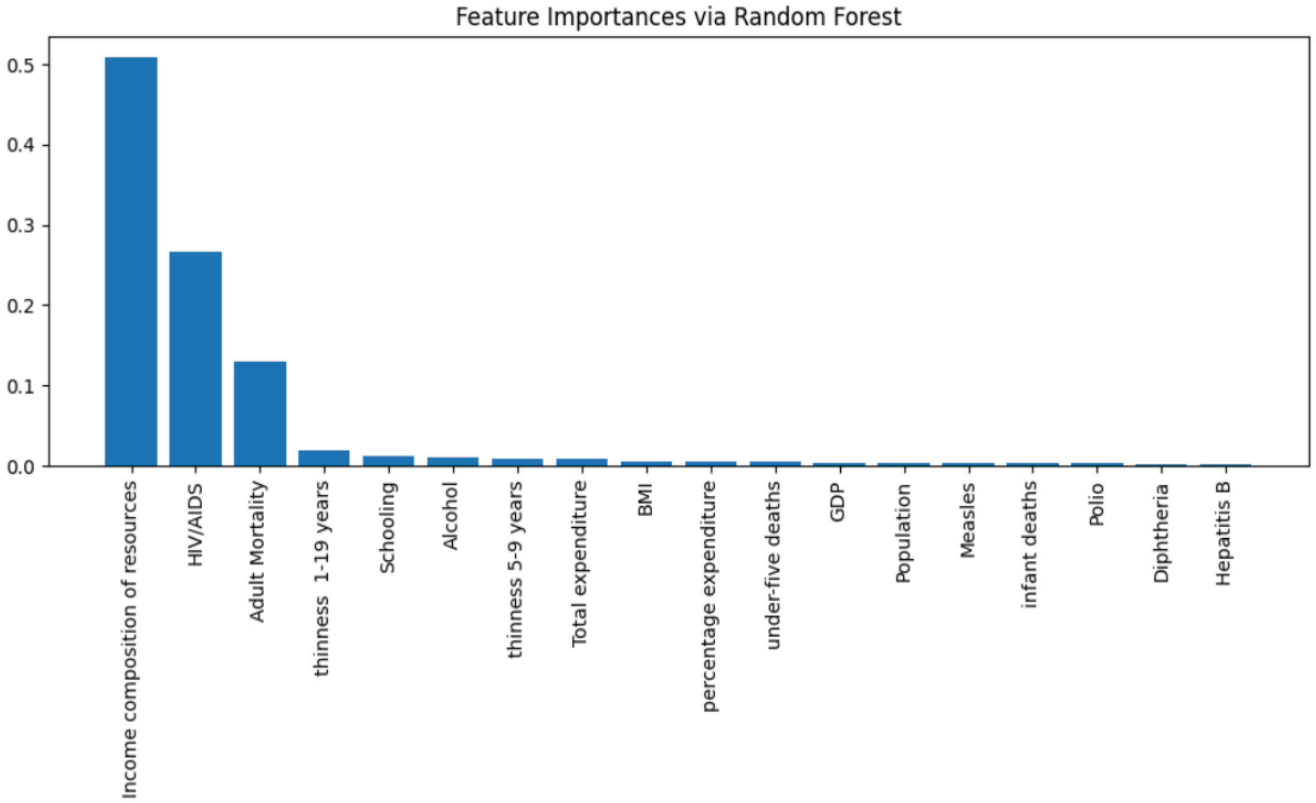
Explanation:

This line plot shows the results of 10-fold cross-validation using the XGBoost algorithm. Each point represents the R² score (a measure of prediction accuracy) for one fold.

Insight:

- The model achieves a high **mean R² score of 0.9547**, indicating it explains approximately 95.5% of the variance in life expectancy.
- The **low standard deviation (0.0111)** suggests the model is consistent and reliable across different data splits.
- This strong and stable performance makes XGBoost a suitable model for life expectancy prediction.

6) Identifying Key Predictors Using Random Forest



Explanation:
The bar chart displays feature importance values derived from a Random Forest regression model. Each bar represents how much a particular feature contributes to predicting life expectancy.

Insight:

- **Income composition of resources** is by far the most influential predictor, suggesting that better access to financial and material resources strongly correlates with longer life expectancy.

- Other significant features include **HIV/AIDS prevalence** and **Adult Mortality**, indicating the impact of infectious diseases and early death on public health.

- Features like **Schooling** and **Thinness (1-19 years)** have minimal impact according to the model, though they may still be socially important.

Suggested Actions for Policymakers and Health Authorities

1. Increase Investment in Education:

- **Rationale:** Strong positive correlation between *schooling years* and life expectancy.
- **Action:** Implement universal primary and secondary education, especially in low-income countries.

2. Strengthen Immunization Programs:

- **Rationale:** High immunization coverage for Hepatitis B, Polio, and Diphtheria aligns with higher life expectancy.
- **Action:** Focus on increasing coverage in underperforming regions through outreach and incentives.

3. Address Mortality Factors:

- **Rationale:** High adult and infant mortality rates negatively affect life expectancy.
- **Action:** Improve maternal health, neonatal care, and adult health services, particularly in developing nations.

4. Target HIV/AIDS Control Programs:

- **Rationale:** HIV/AIDS shows a strong negative correlation with life expectancy.
- **Action:** Increase access to antiretroviral therapy, awareness, and testing, especially in high-prevalence regions.

5. Improve Nutrition and Tackle Thinness

- **Rationale:** Thinness (especially in youth) correlates with low income and poor health outcomes.
- **Action:** Launch school meal programs, child nutrition plans, and poverty alleviation efforts.

6. Prioritize Healthcare Spending Strategically

- **Rationale:** Countries with higher total and percentage healthcare expenditure tend to have longer life expectancies.

- **Action:** Ensure healthcare budgets are efficiently allocated toward preventive and primary care services.

7. Monitor and Reduce Alcohol Abuse

- **Rationale:** Excessive alcohol consumption negatively affects health and life expectancy.

- **Action:** Regulate alcohol access, raise public awareness, and support rehabilitation initiatives.

8. Focus on Economic Upliftment

- **Rationale:** Income composition of resources is the most influential predictor in life expectancy.

- **Action:** Promote employment, entrepreneurship, social welfare, and equitable access to economic resources.

Conclusion:

This comprehensive life expectancy analysis underscores the **multifactorial nature** of public health outcomes. It reveals that **education, immunization, healthcare expenditure, economic equity, and mortality control** are key levers for improving longevity. Countries, especially developing ones, must adopt an **integrated approach**, combining health, education, and economic policies, to foster a healthier, longer-living population.

The insights from advanced machine learning models, including Random Forest and XGBoost, confirm that predictive accuracy can be high when these critical indicators are tracked and addressed. As such, this study offers a valuable framework for **data-driven policymaking** in global public health.

THIS IS A WORK DONE BY 'SURENDRAN L'