Introduction

Research Question

"What are key health and economic factors influencing life expectancy in developed and developing countries?"

Research Gap Analysis

There are many studies analyzing the effect of economic factors (GDP) or health factors (mortality rates, vaccination rates) in isolation on life expectancy but lack of research analyzing the impact of multiple factors. The aim is to fill this gap by doing so.

Importance of the study

Life expectancy is a key indicator of overall well being of any country. This reflects the quality of healthcare system, technological advancements, economical prosperity and the quality of the education system. By identifying the drivers of the life expectancy countries can take informed decisions to improve public, allocate resources effectively and reduce preventable deaths.

Dataset overview

Link for the dataset used: https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

The 'life-expectancy' data set by WHO provides information on global life expectancy and related economic and health factors.

Time period: Multiple years (making it suitable for a time series analysis)

Geographical coverage: data from over 190 countries.

Data Transformation

Variable identification

Dependent variable:

- **Life expectancy**: Average life expectancy in age

Independent variables:

- **Adult Mortality**: Probability of dying between 15 and 60 years per 1000 population.
- **Infant Deaths**: Number of infant deaths per 1000 live births
- **Hepatitis B Coverage** : Vaccination rates for Hepatitis B.
- **Polio and Diphtheria Vaccination Rates** (%): Coverage rates for these essential vaccines.
- **HIV/AIDS** : Death rate per 1,000 population due to HIV/AIDS.
- **Alcohol Consumption**: Per capita alcohol consumption (liters per person per year).
- **GDP**: Gross Domestic Product per capita in USD.
- **Population**: Total population of each country.
- **Schooling**: Average number of years of schooling per person.
- **Development Status**: Categorical variable classifying countries as "Developed" or "Developing."
- **Percentage Expenditure**: Proportion of government expenditure related to healthcare.
- **Income Composition of Resources**: Human development index (HDI) measure accounting for income.

Evaluation of Data Tidiness

- The structure of the dataset was already tidy (Each column was a variable and each row was an observation) but there were missing values present in crucial fields such as Life expectancy, Population, GDP as well as in many other columns used for the analysis.
- Columns such as ‘thinnesss..5.9.years’ and ‘thinness..1.9.years’ were found irrelevant for the analysis.
- Duplicates were checked and confirmed absent.
- There were potential derived variables that could be used further analysis.

Transformations that were made to the dataset

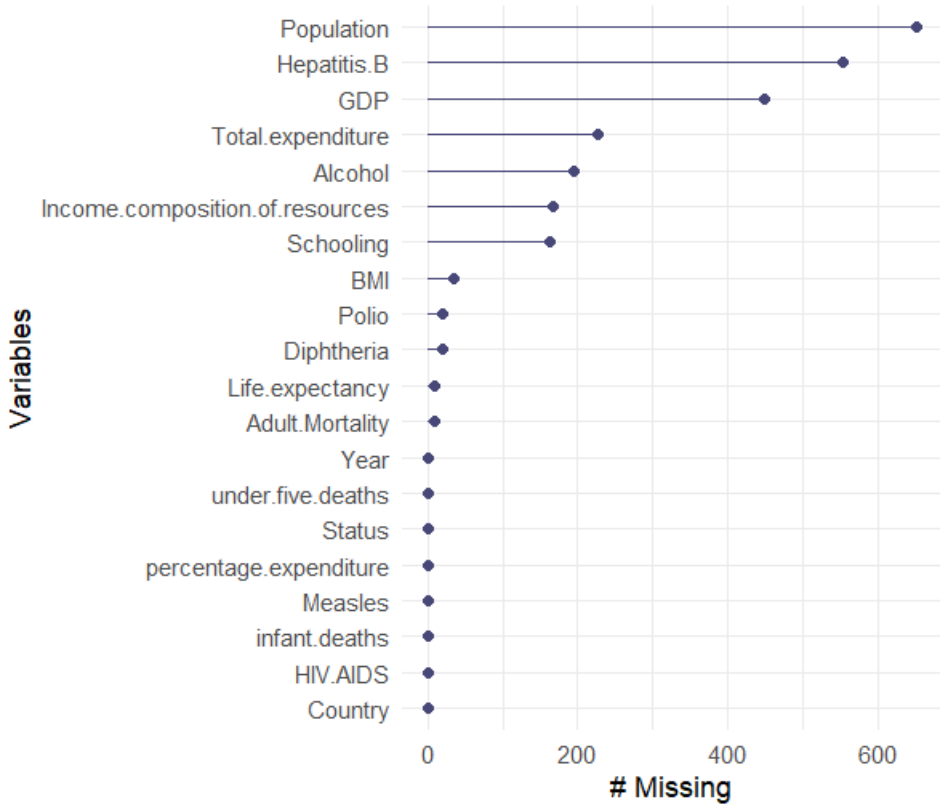
First I dropped the ‘thinnesss..5.9.years’ and ‘thinness..1.9.years’ columns which were irrelevant.

```
# Load the dataset
health_data <- read.csv('Life Expectancy Data.csv')

# Check for column names
print("Column Names:")
print(names(health_data))  # Verify exact column names

# Step 2: Drop unnecessary columns
health_data <- health_data %>%
  select(-c(`thinness.5.9.years`, `thinness..1.19.years`))  # Drop columns
```

Then I checked for missing values for each column and visualize the results by using gg_miss_var() function which is a part of "naniar" library.

```
# Step 3: Check for missing and duplicated rows
total_rows <- nrow(health_data)
missing_counts <- colSums(is.na(health_data))
missing_summary <- data.frame(
  Column = names(health_data),
  Missing_Count = missing_counts,
  Missing_Percentage = (missing_counts / total_rows) * 100
)
print("Missing Values Summary:")
print(missing_summary)

# Visualize missing values per variable
gg_miss_var(health_data)
```

The results were as bellow



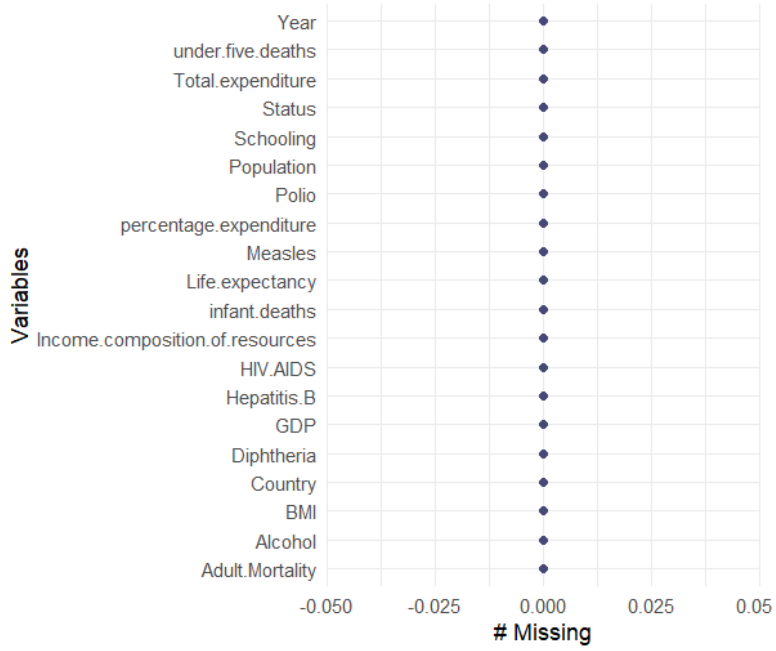
Then I replaced the missing values using applicable methods for each column.

As the missing values were high in some key variables I decided to group them by the 'country' name or by the 'status' which were 'developed' and 'developing' to improving the quality and reliability of the analysis.

```
# Step 4: Replace missing values for Population
health_data <- health_data %>%
  group_by(Country) %>%
  mutate(
    Population = ifelse(
      is.na(Population),
      ifelse(all(is.na(Population)), median(health_data$Population, na.rm = TRUE), median(Population, na.rm = TRUE)),
      Population
    )
  ) %>%
  ungroup()
```

```r
# Step 6: Replace missing values for Hepatitis.B
health_data <- health_data %>%
  group_by(Status) %>%
  mutate(
    Hepatitis.B = ifelse(is.na(Hepatitis.B), mean(Hepatitis.B, na.rm = TRUE), Hepatitis.B)
  ) %>%
  ungroup()

# Step 7: Replace missing values for Life.expectancy
health_data <- health_data %>%
  group_by(Status) %>%
  mutate(
    Life.expectancy = ifelse(
      is.na(Life.expectancy),
      ifelse(all(is.na(Life.expectancy)), mean(health_data$Life.expectancy, na.rm = TRUE), mean(Life.expectancy, na.rm
      Life.expectancy
    )
  ) %>%
  ungroup()
```

After following above method for all the columns I confirmed all the null values were replaced by using a visualization.



Then I created Mortality Ratio, Vaccination_Coverage_Index and Health GDP Percentage columns that were used for the further analysis using mutate().

```r
# Calculate Vaccination Coverage Index
data <- health_data %>%
  mutate(Vaccination_Coverage_Index = rowMeans(select(., Hepatitis.B, Polio, Diphtheria), na.rm = TRUE))

# Calculate Health GDP Percentage
data <- data %>%
  mutate(Health_GDP_Percentage = (Total.expenditure / GDP) * 100)

#Calculate Mortality Ratio
data <- data %>%
  mutate(Mortality_Ratio = infant.deaths / Adult.Mortality)
```

At the end I checked whether there are any duplicate rows , which I did not found any.

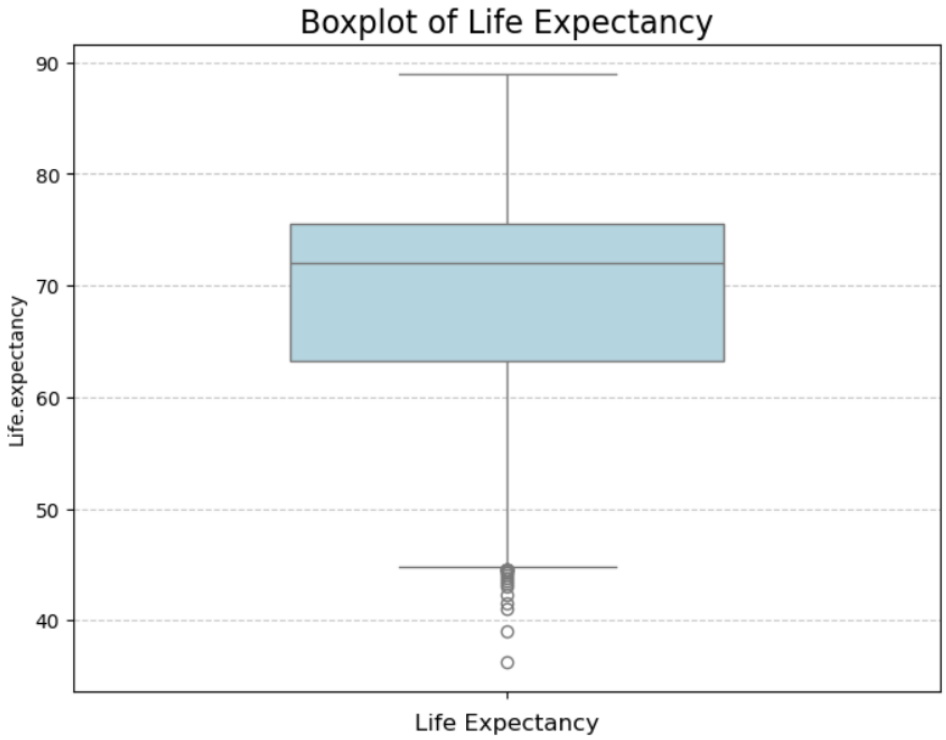
```r
# Check for duplicate rows
cat("Number of duplicated rows:", sum(duplicated(health_data)), "\n")
```

Finally saved it to a new csv file to be used for the exploratory analysis.

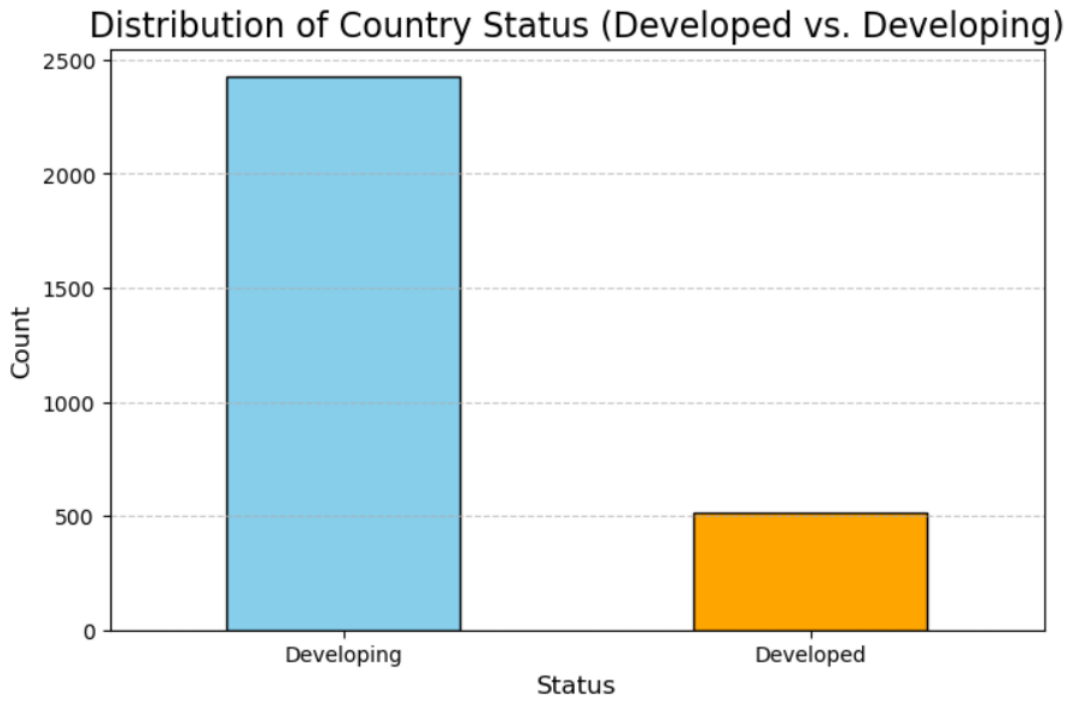
```r
# Step 18: Save the filtered dataset
write.csv(data, "Modified_Health_Data_final2.csv", row.names = FALSE)
```

Exploratory Data Analysis

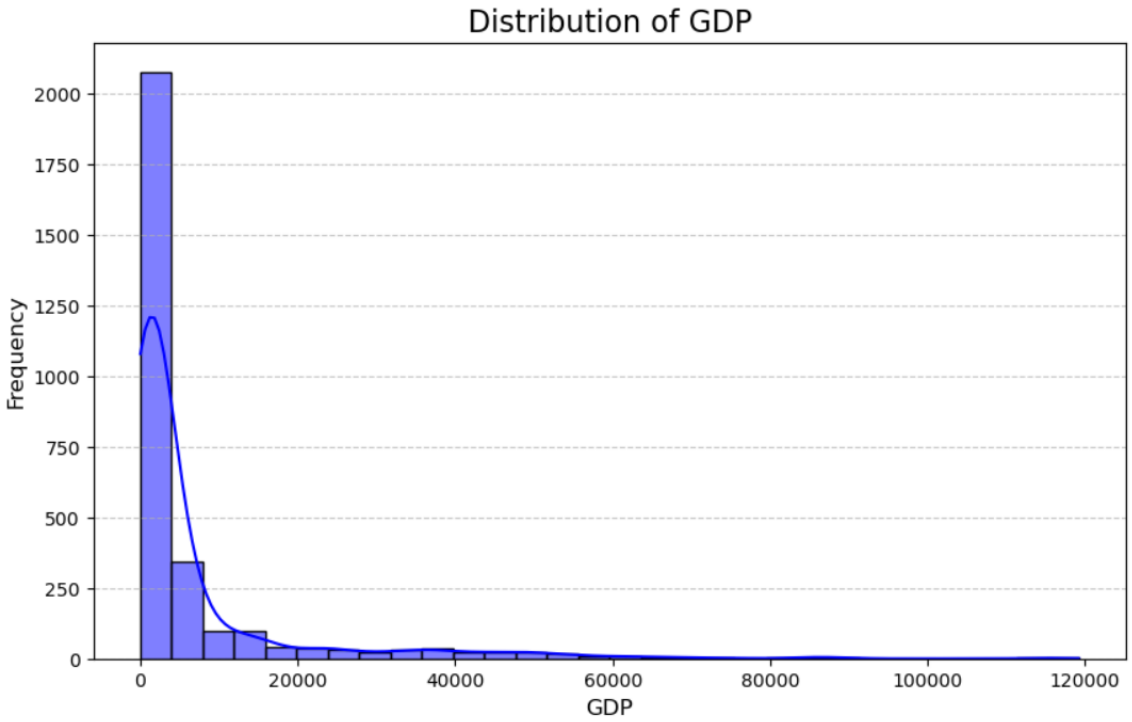
Univariate Analysis



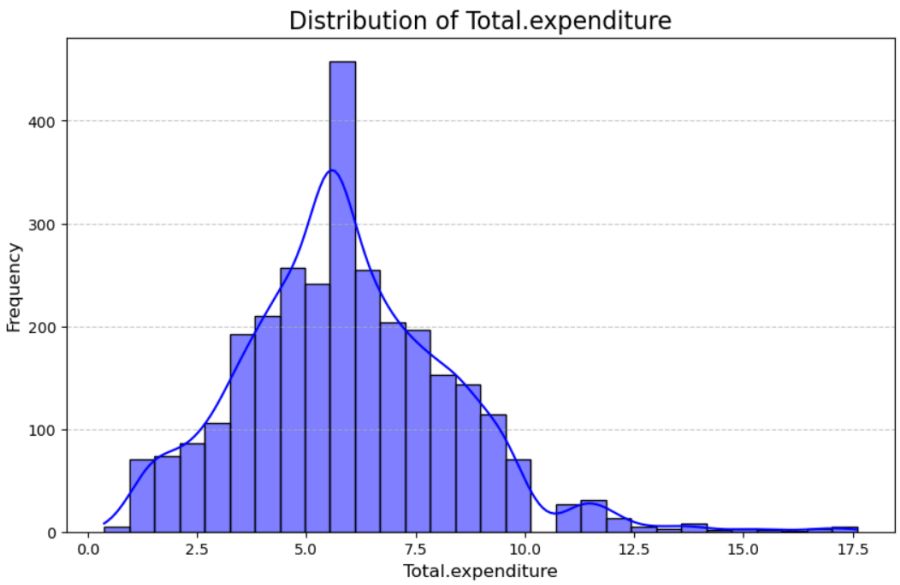
Majority of countries have a life expectancy between 60 and 80 years. There are few where it is less than 50.



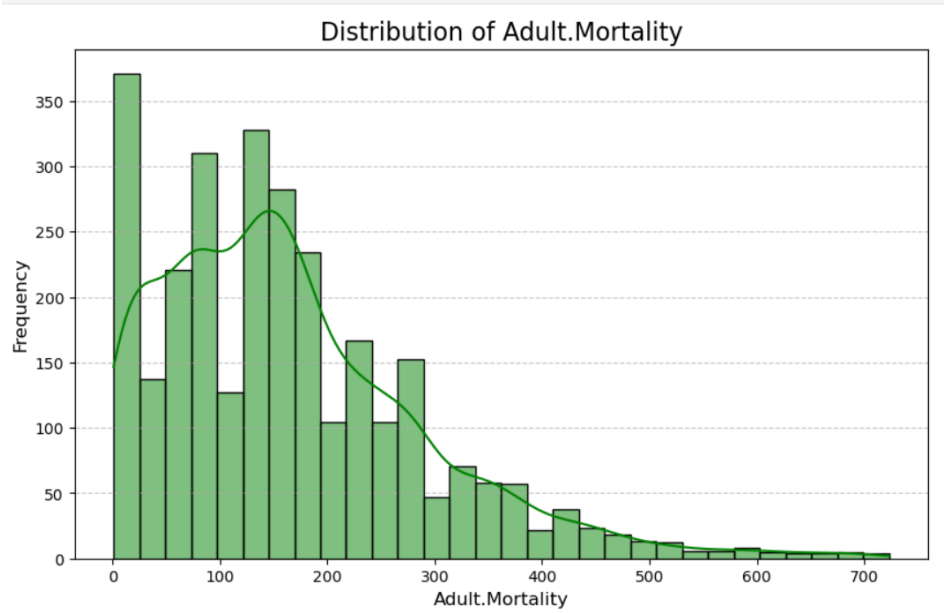
This shows that the developing counties represents the majority in the dataset.



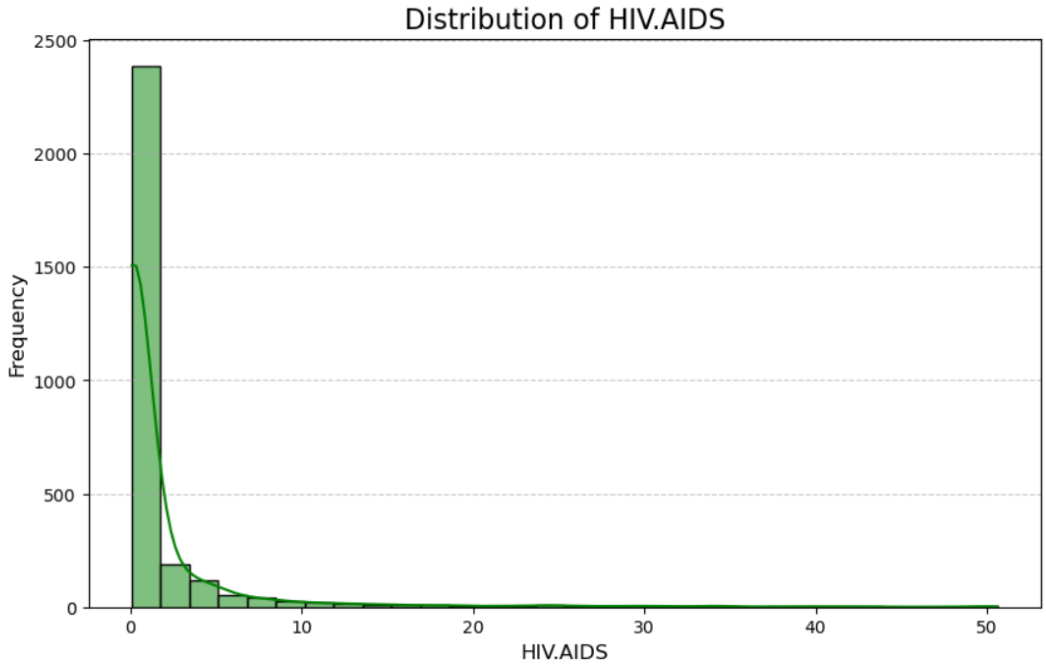
Suggests that the more countries have less GDPs. As majority belong to ‘developing’ category.



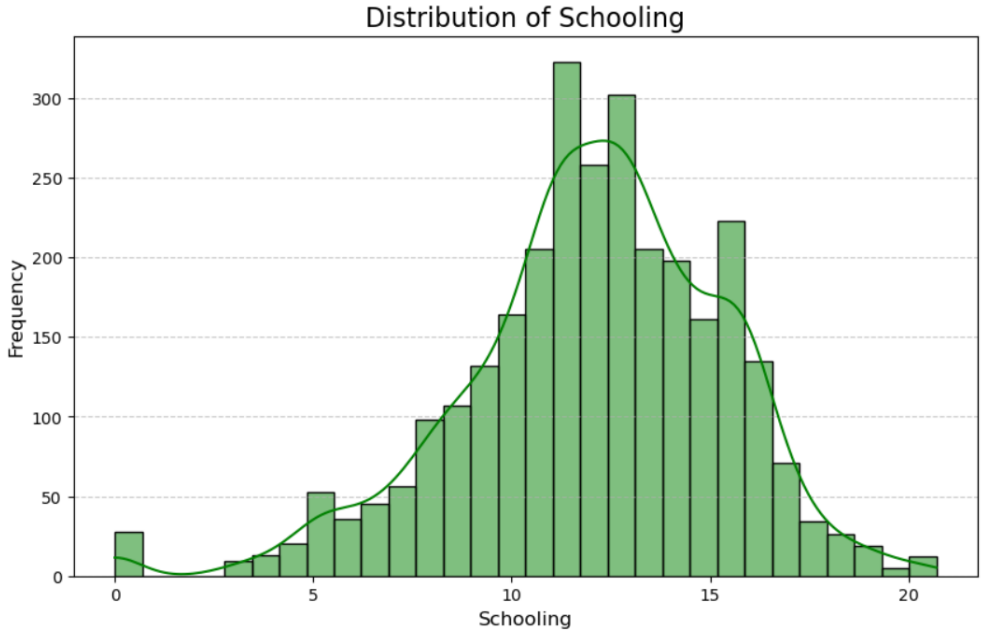
This suggest that the majority of countries spend between 5% and 10% of GDP on health.



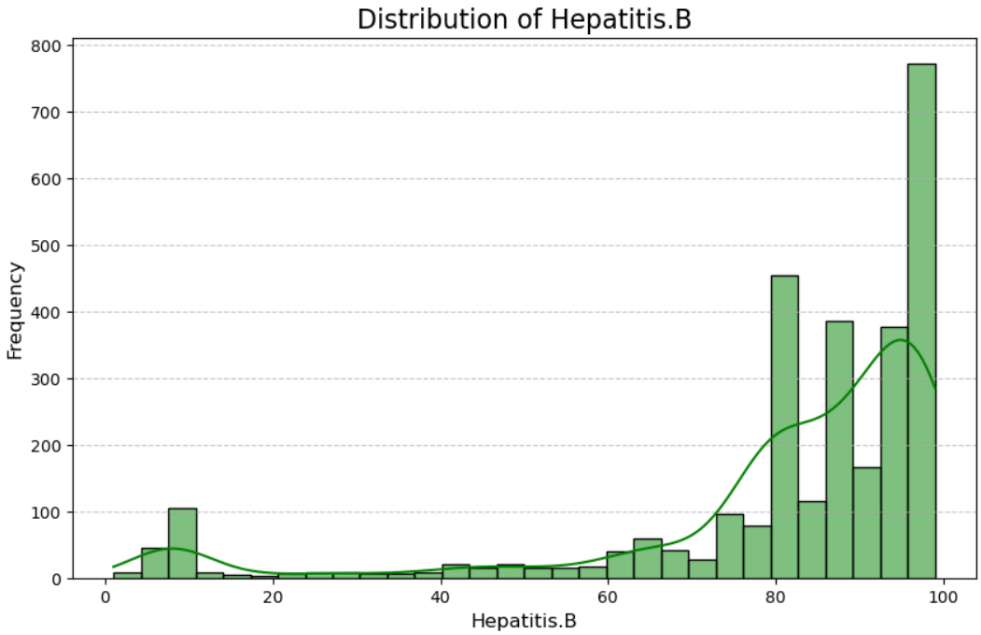
Many countries shows an adult mortality rates concentrated between 100 and 300 deaths per 1000 population while few shows significantly high rates, probably due to health challenges or some crisis.



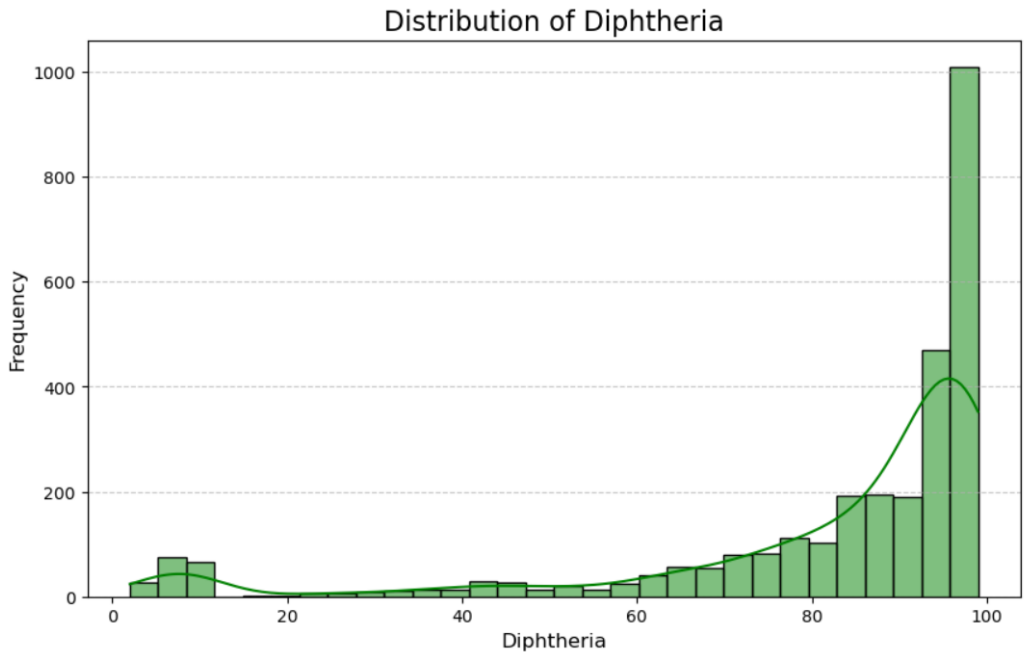
Majority of countries have a very low HIV/AIDS prevalence, with most of values near or below 1 death per 1000 population.



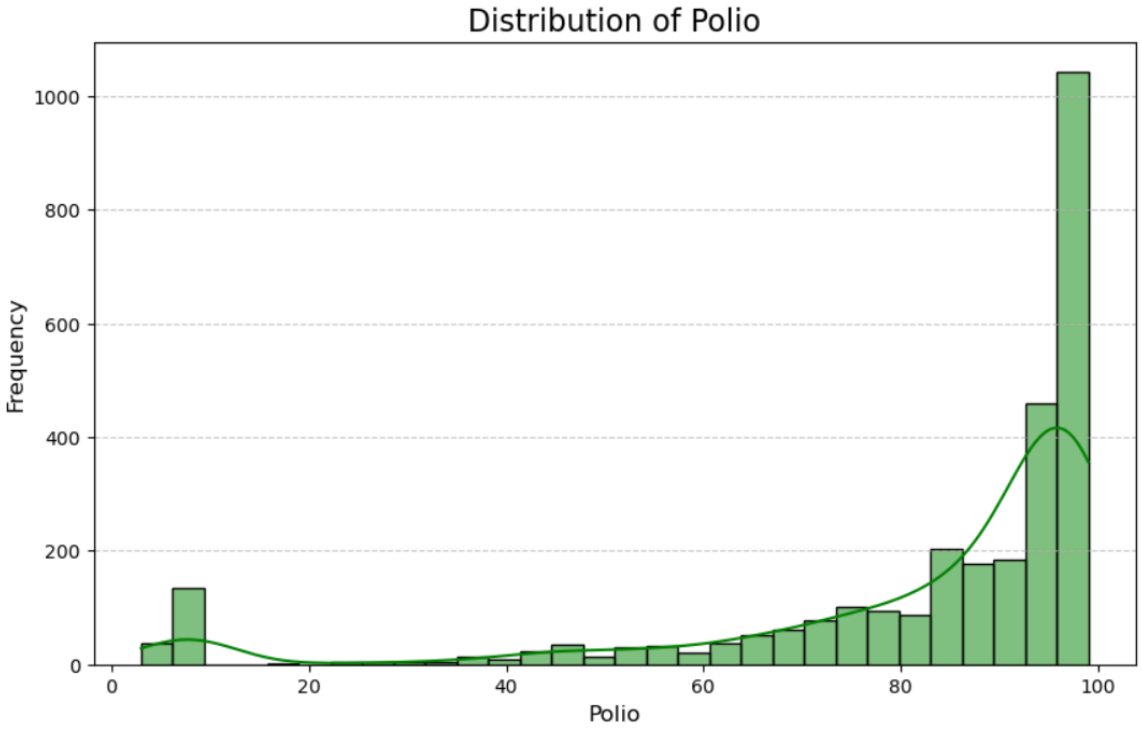
Majority of countries have an average of schooling years between 8 and 15, which is a typical range globally.



For Hepatitis B vaccination, most countries seems to have a wide vaccination rates between 60 percent and 100 percent.

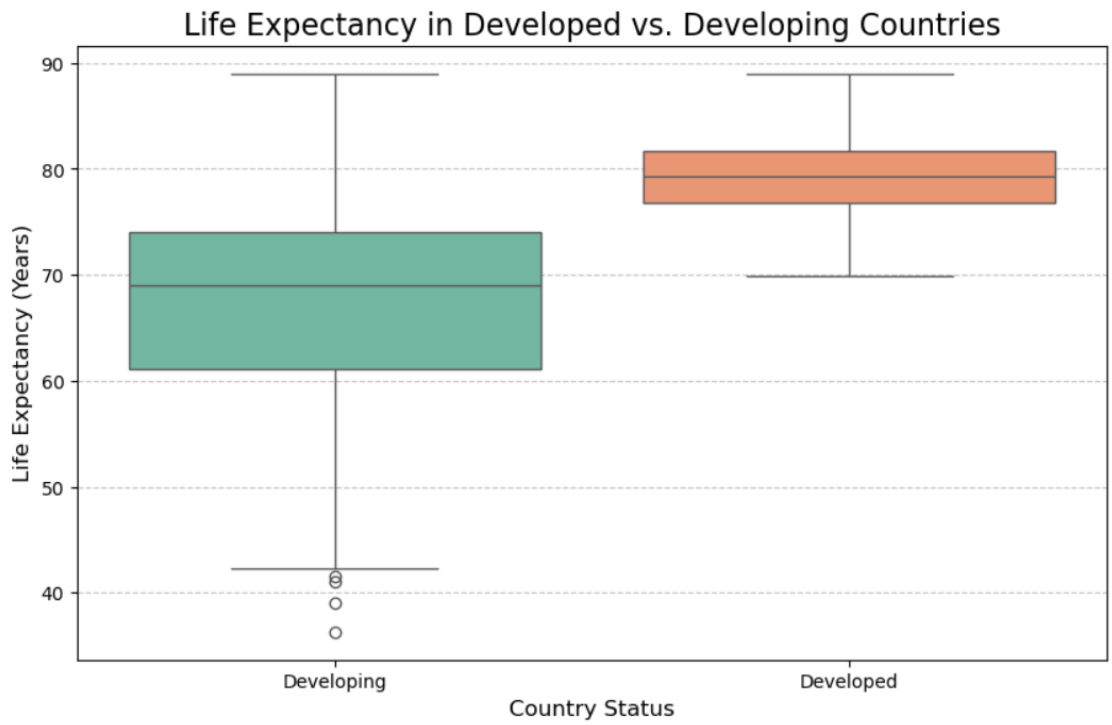


In case of Diphtheria it seems most countries have a vaccination rate of around 80 percent to 100. Which is a good sign.

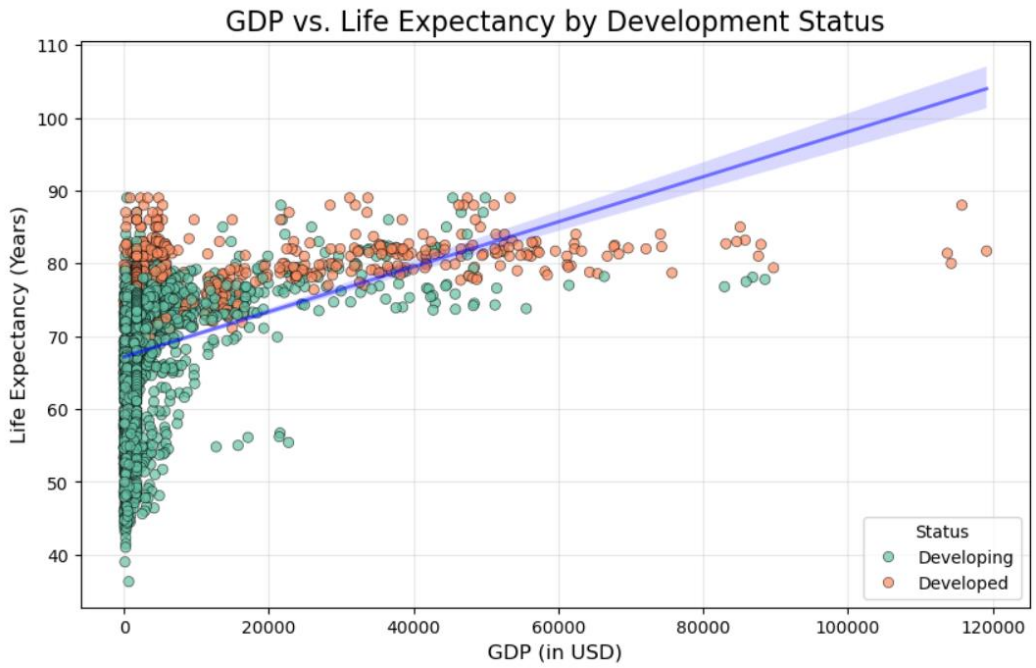


For Polio vaccination majority of countries shows a vaccination rate of more than 80.

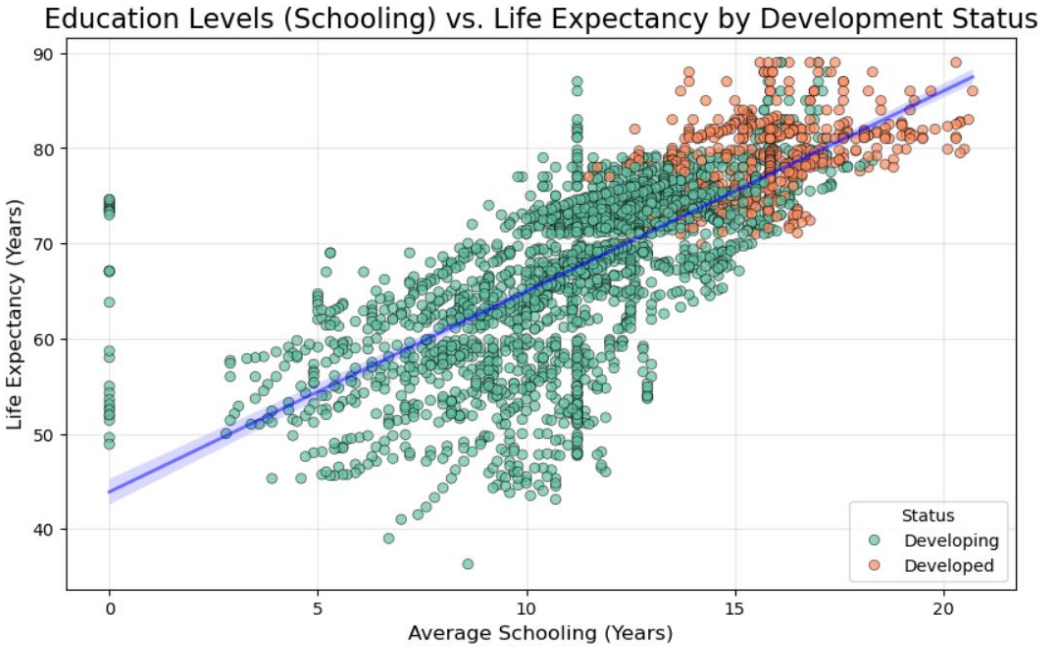
Multivariate Analysis



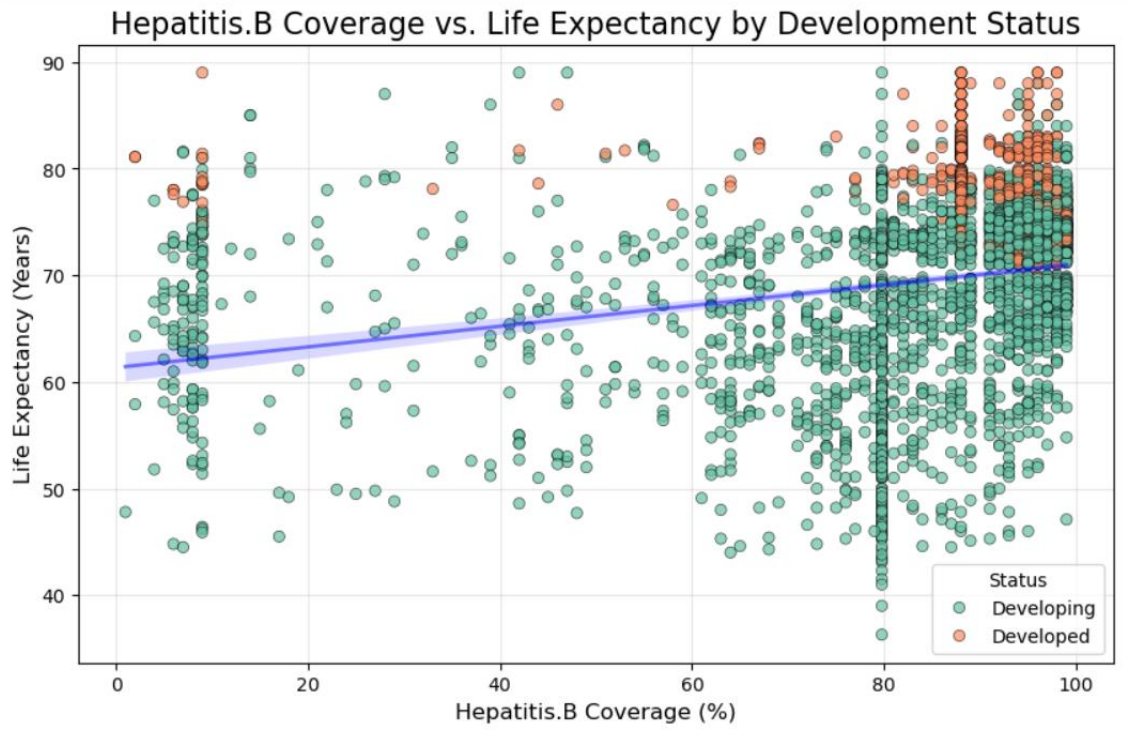
For developed countries we can see there is a higher median for life expectancy which is above around 75 indicating better healthcare and living conditions. While for developing the median is around 68.



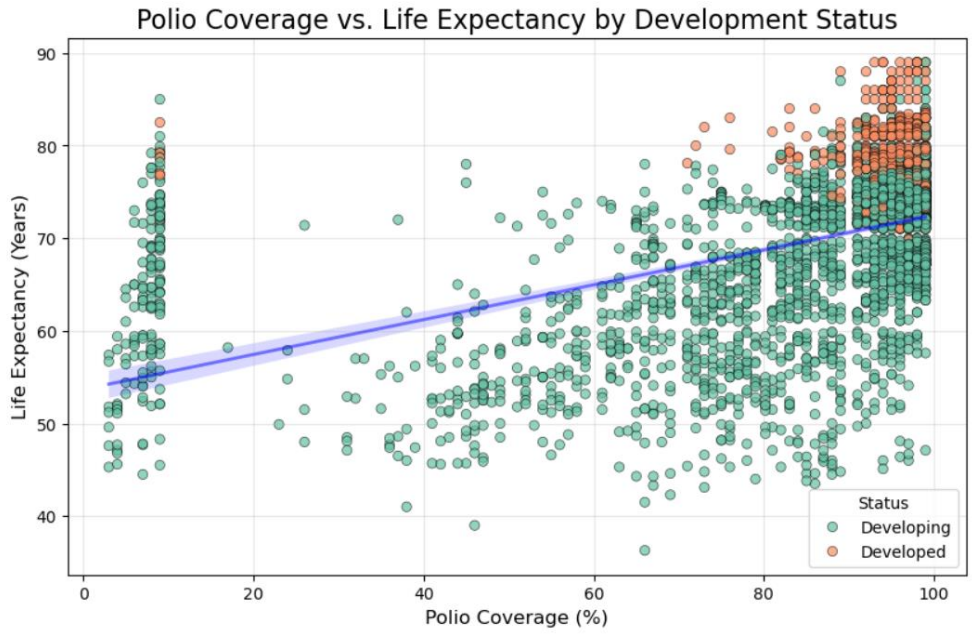
As we can see there is a clear positive relationship between GDP and life expectancy. Higher GDP is related with higher life expectancy, reflecting the role of economic resources here. In case of developed countries there is a higher life expectancy of 75 – 90 years with a higher GDP of above 20000 USD for majority of cases. While for developing countries life expectancy is from 40 – 80 with an average GPA less than 10000 USD.



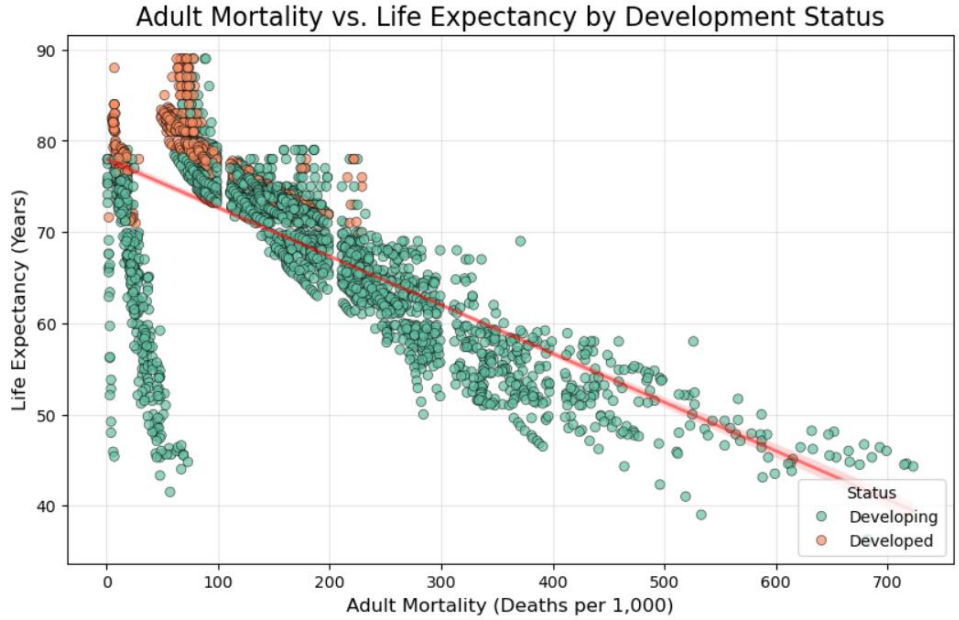
As we can clearly see there is a positive coloration between life expectancy and average schooling years meaning that as the education levels increase life expectancy will also tend to increase. For developed countries we can see that the average schooling years are more than 15 years for majority and in case of developing countries there is a wider range of schooling years and low life expectancy.



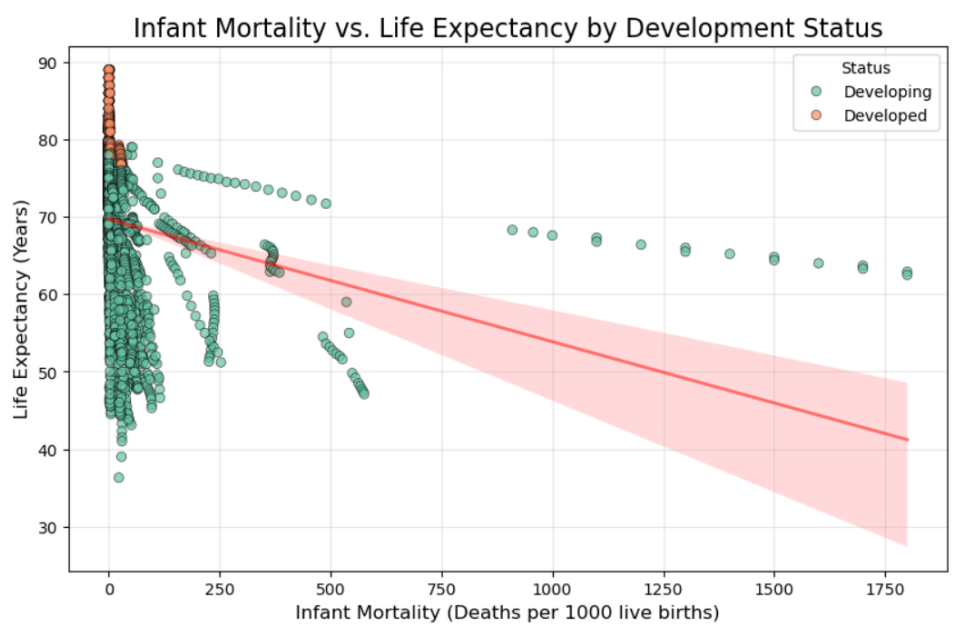
In case of Hepatitis.B coverage there is a positive coloration. Meaning that increased Hepatitis.B vaccination rates align with higher life expectancy.



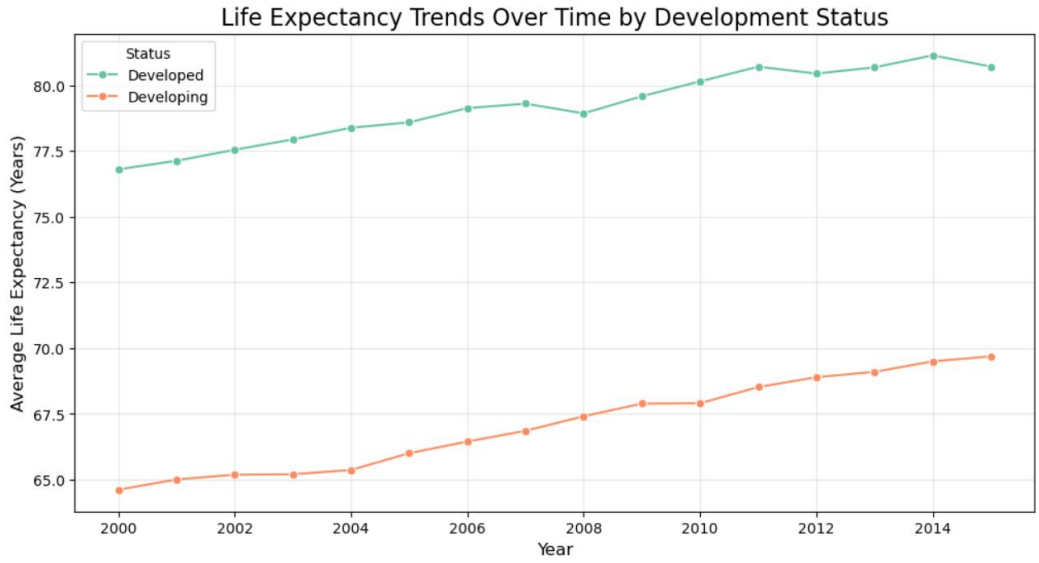
Just like for Hepatitis.B we can see a positive coloration for Polio vaccination indicating that increased polio vaccination rates comes with higher life expectancy.



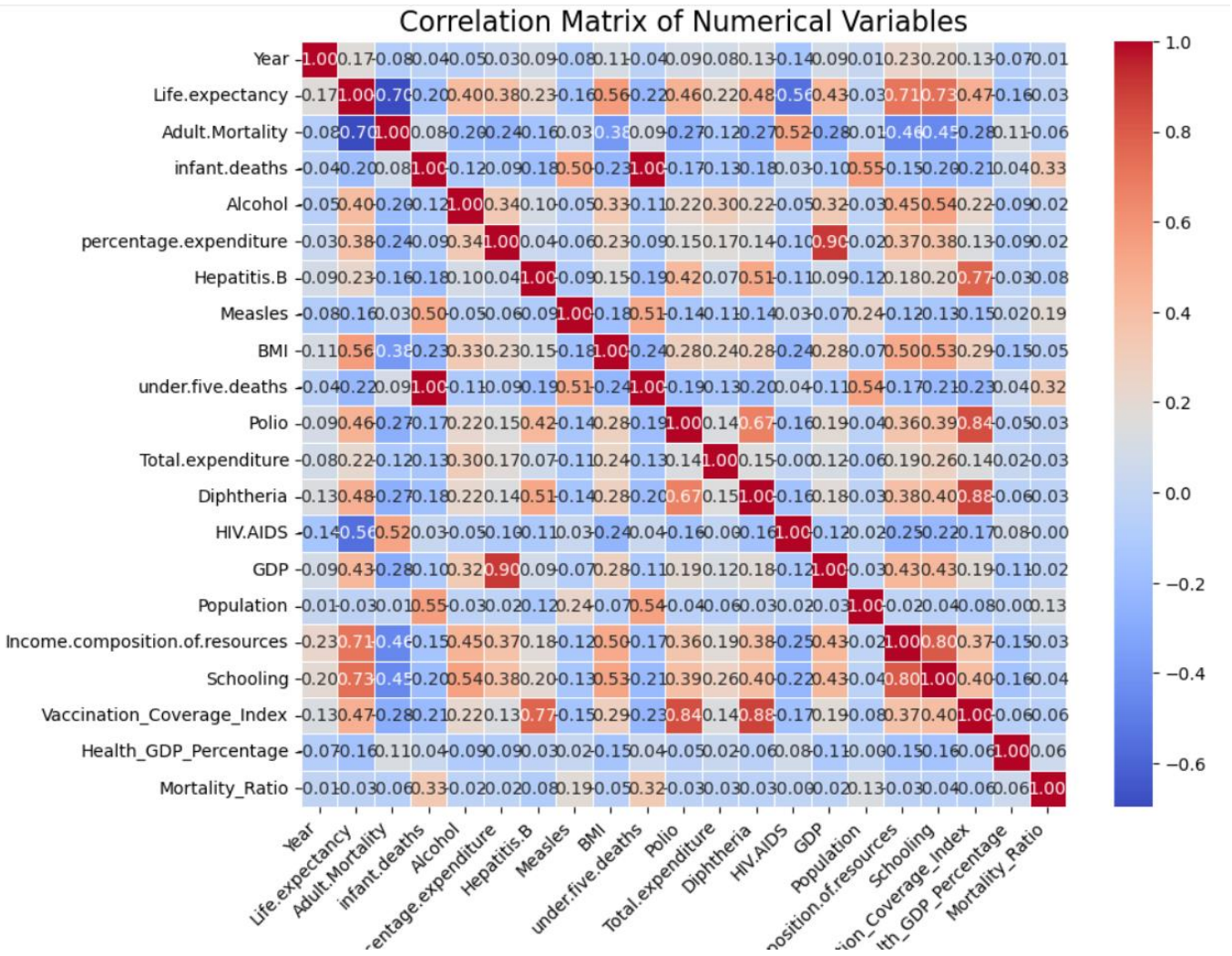
There is a clear inverse relationship as the adult mortality increase, life expectancy goes down.



Just like for adult mortality, there is a strong negative relationship for life expectancy and infant mortality.



Can see significant improvements in life expectancy through years for both.



By using the above correlation matrix we can understand the linear relationships between variables and understand patterns.

Data storytelling

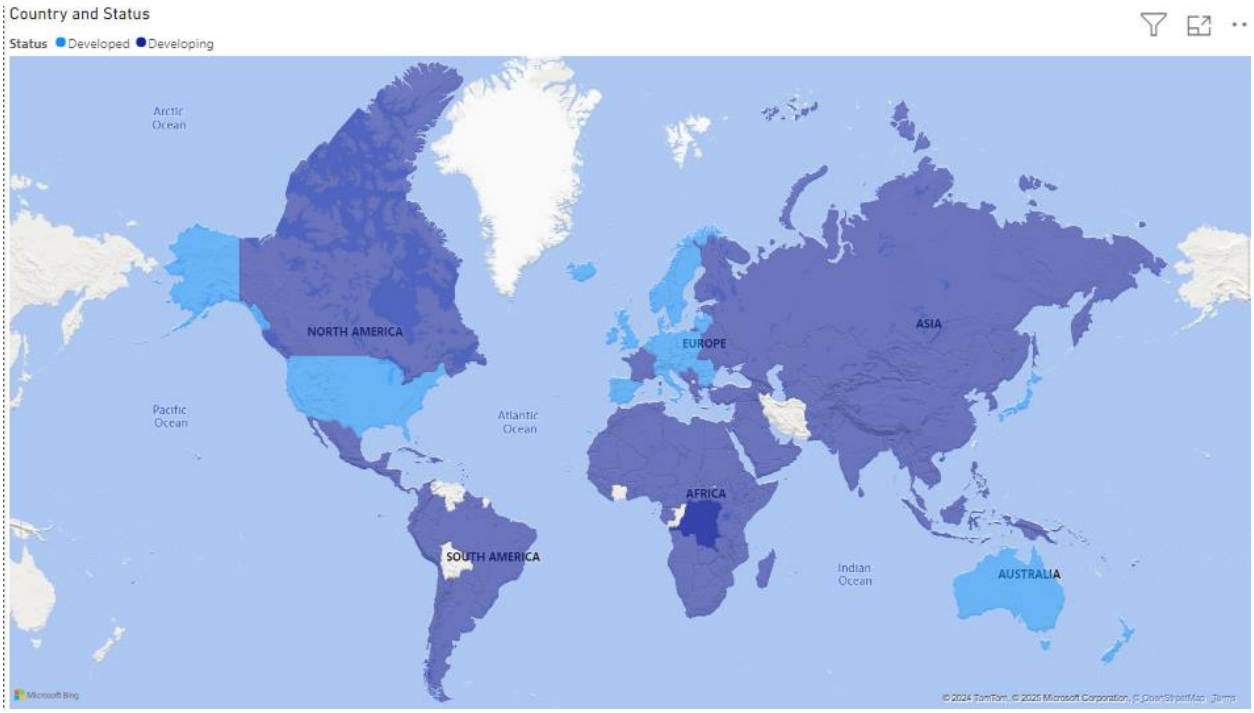
Introduction: wealth is health?

Life expectancy is not just a number, it is way beyond that. It is a reflection of economic stability, advancements of healthcare system and quality of life in a country. Yet, why some countries struggle with shorter life span despite advances in modern technology and medicine? Do developed (wealthy) nations always have better life expectancy than the rest? Or do the developing countries outperform expectations?

Understanding what influence the life expectancy in nations requires us to dive deeper into health and economic factors. Today we aim to reveal them by exploring data of various developed and developing countries.

Chapter 1: A Glimpse of Life Expectancy Across the Globe

I think first it is better to take a peek into the development state of the word countries before diving into the life expectancy factors.



In the given map the light blue area represents the developed nations and the dark blue area for still developing ones. As we can see it is clear that the majority of the word countries are still under developing state except for counties like United States, United Kingdom, Australia and some parts of Europe.

Now we will see the distribution of life expectancy throughout the globe.

By using the map above now it is clear that developed countries I mentioned before have better life expectancy compared to the rest, as all of them shown in yellow which is the highest.

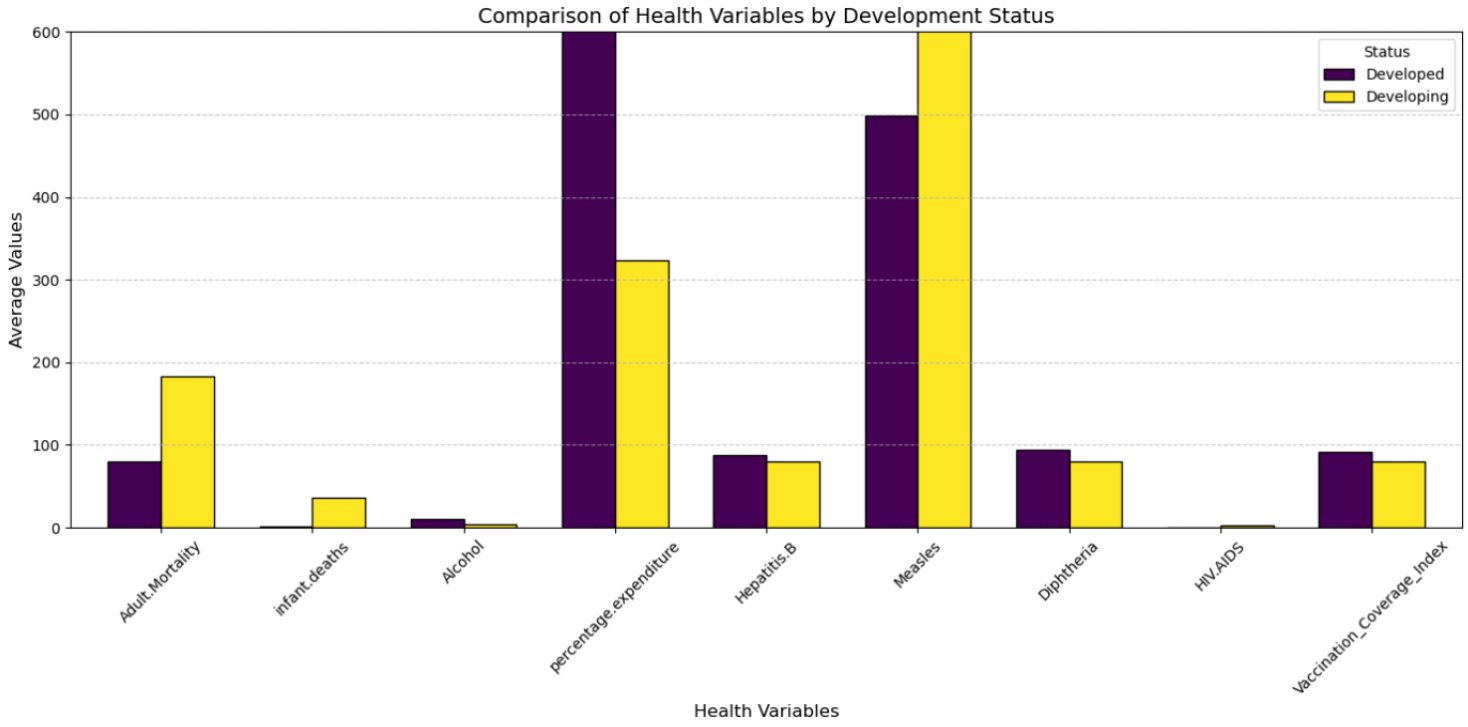
But the question is ‘why’?

Why people in those counties get to live a longer life than the rest of the word?

Now we shall dive into the factors.

Chapter 2: The pulse of health

In this chapter we will examine the core health factors in developed and undeveloped countries so we could try finding the reasons for life expectancy difference between nations.



We will start this by discussing the percentage expenditure, which I feel like the most important. in here what it meant by this is how much portion of a country's total expenditure allocated to health. For an example high expenditure means that a significant amount of government's total budget is dedicated to health services, such as hospitals, disease prevention and vaccination programs. As we could clearly see developed countries dedicates a higher percentage of their total expenditure for health.

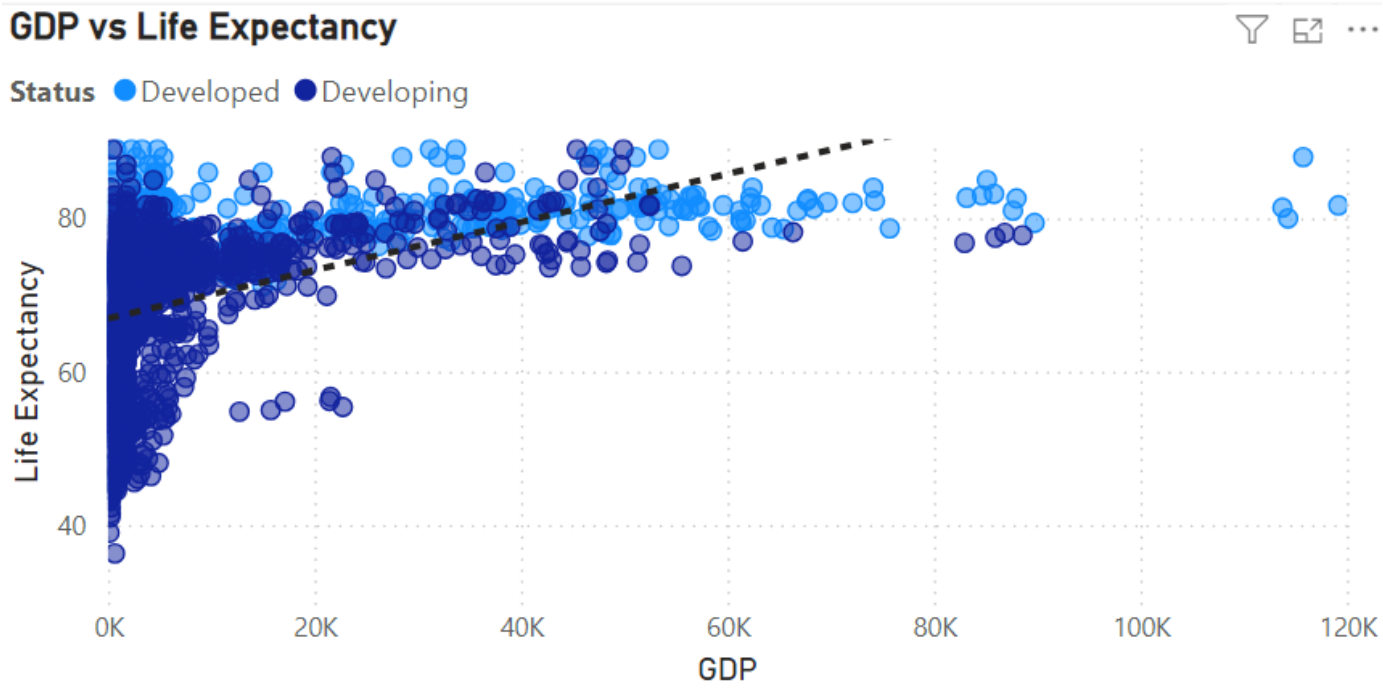
The Vaccination coverage Index which is like the shield against preventable diseases, shows the average vaccination rates for Polio, Hepatitis B and Diphtheria also plays a huge role in life expectancy. The data shows a significant gap for this between developed and developing countries. Showing us that developed countries contributes higher when it comes to decease prevention among their populations.

Alcohol consumption seems to be notably higher in developed nations. Which raise concerns of lifestyle diseases such as liver issues and cardiovascular problems. For developing nations lower alcohol consumption could be due to they being less affordable and restrictions on access.

Adult mortality and infant deaths rates are significant markers for a country. Developing countries with their advanced medical infrastructure and medical systems seems to have significant lower rates for both of these. In other hand developing countries suffer from the curse of higher mortality rates.

Chapter 3: Economic Determinants of Longevity

Longevity of a country can be called as a fundamental measure for nation's health, which highly connected with factors such as GDP.



The above scatterplot shows a positive coloration between GDP and Life Expectancy. This could be due to with a higher GDP a country have much more resources to put towards healthcare and education. These investments will provide a healthier life for its citizens.

Here developing countries holds a higher GDP levels with consistently high life expectancy which is mostly above 75. Highlighting stability of the health care systems.

Some developing countries with a lower GDP exceed the life expectancy expectations indicating outliers. This could be due to effective public health policies or cultural factors effecting longevity.

Conclusion: Lessons from the Data

The data teaches us improving health in a global level is a multifaced challenge which required tailored solutions. Economic growth is not a sole answer here, governments should prioritize efficient resources allocation, improve healthcare and education to improve health in all the development levels.