Team Details

Team Name: Team Synergy
Team Members:

1. Rishitha Rani Pakam – Team Lead
   Email ID: rpaka1@unh.newhaven.edu
2. Narasimha Reddy Padire
   Email ID: npadi1@unh.newhaven.edu
3. Lakshmi Reddy Bhavanam
   Email ID: lbhav2@unh.newhaven.edu

Dataset Title: Life Expectancy (World Health Organization) 2024

Source: https://www.kaggle.com/datasets/sonialikhan/life-expectancy-who-2024

Selected dataset contains 22 attributes and 2938 records. It contains a wide range of variables across 193 countries, including health, economic, and social indicators. Attributes in the dataset are Country, Year, Status, Life expectancy, Adult Mortality, infant deaths, Alcohol percentage, expenditure, Hepatitis B, Measles, BMI, under-five deaths, Polio, Total expenditure, Diphtheria, HIV/AIDS, GDP, Population, thinness 1-19 years, thinness 5-9 years, Income composition of resources, Schooling. The dataset contains both numerical and categorical data. There are 2 attributes of object type, 11 decimal and 9 integer types.

Research Question

What are the most significant factors predicting life expectancy, and how do health, economic, and social variables interact to influence life expectancy in the top two developed countries and the top two developing countries?

Exploration Techniques Used:

Univariate:

df.describe() – Count, Min, Max, Mean, Median, Quartiles, Standard Deviation
Boxplot (for outlier detection)
Boxplot (winsorization)
Histograms

Bivariate:

Correlation matrix
Correlation heatmap
Line chart
Scatter plot
Bar chart

Visualizations:

Line chart
Scatter plot
Bar chart

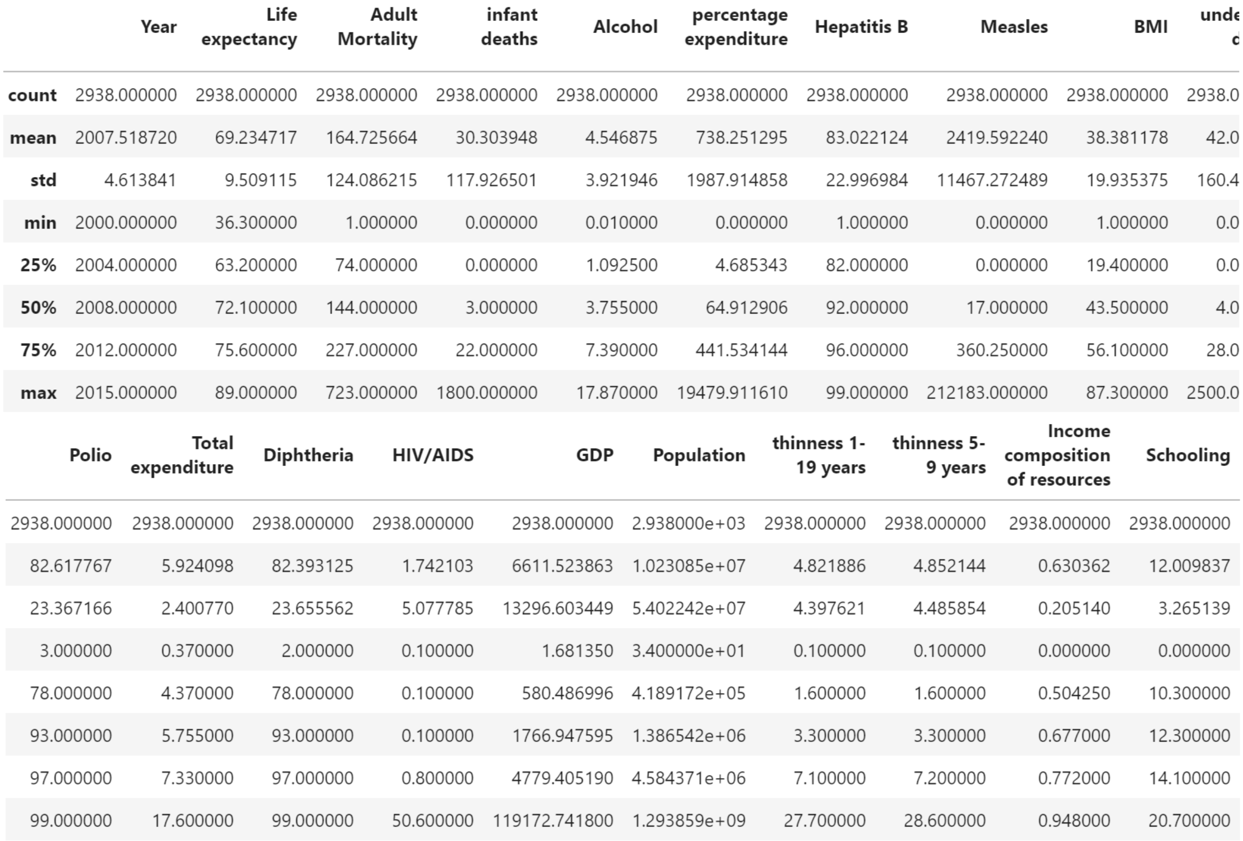
1. Missing Values
2. Outliers

Handling Missing Values:

We checked for missing values and observed that 14 columns contain missing values. We filled those missing values with median of each column. We used median because it is not affected by extreme values, which provides a stable and accurate fill for missing data.

Findings based on Data Exploration:

Table: Statistical Analysis



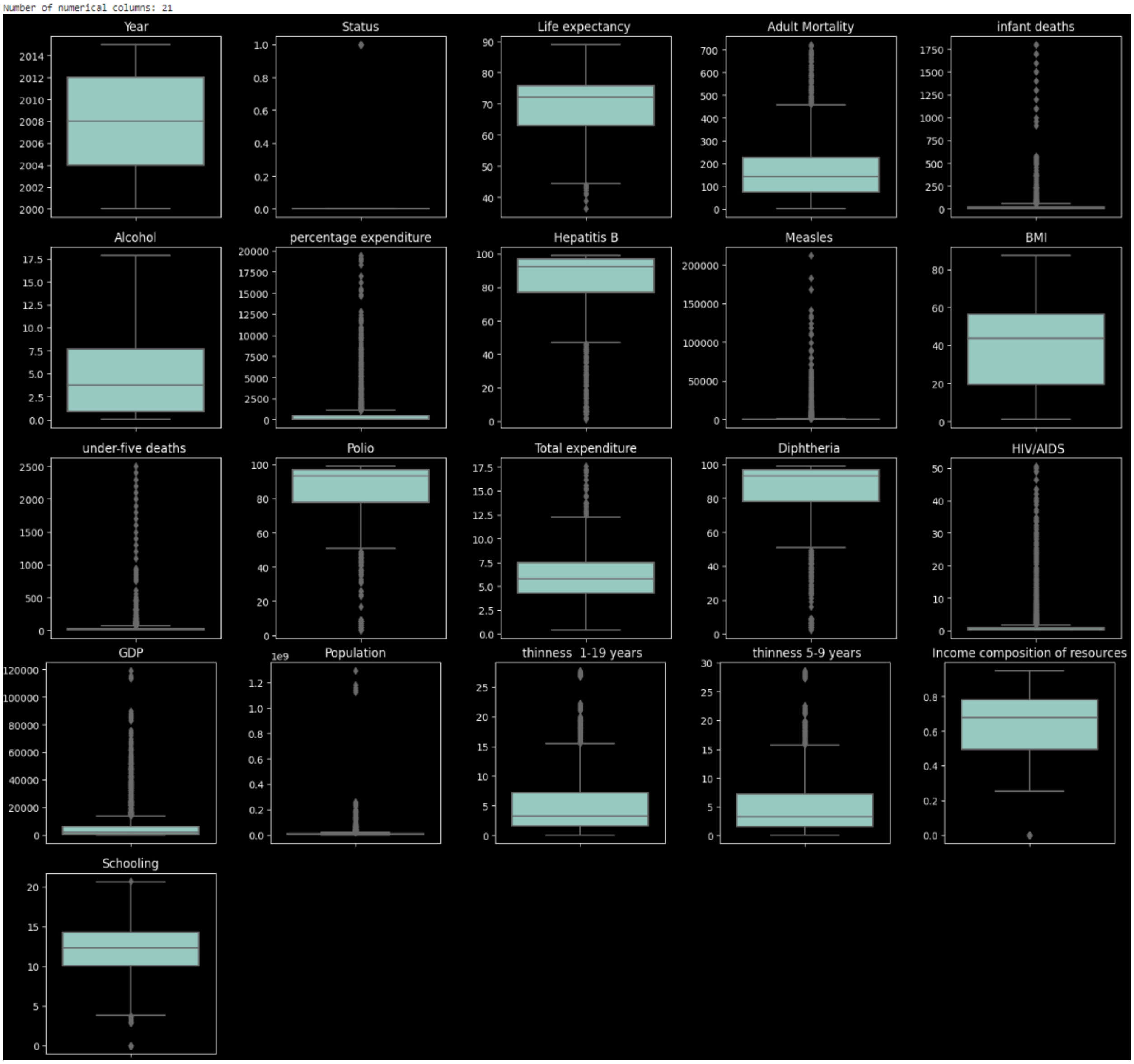
	Year	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	BMI	unde c
count	2938.000000	2938.000000	2938.000000	2938.000000	2938.000000	2938.000000	2938.000000	2938.000000	2938.000000	2938.0
mean	2007.518720	69.234717	164.725664	30.303948	4.546875	738.251295	83.022124	2419.592240	38.381178	42.0
std	4.613841	9.509115	124.086215	117.926501	3.921946	1987.914858	22.996984	11467.272489	19.935375	160.4
min	2000.000000	36.300000	1.000000	0.000000	0.010000	0.000000	1.000000	0.000000	1.000000	0.0
25%	2004.000000	63.200000	74.000000	0.000000	1.092500	4.685343	82.000000	0.000000	19.400000	0.0
50%	2008.000000	72.100000	144.000000	3.000000	3.755000	64.912906	92.000000	17.000000	43.500000	4.0
75%	2012.000000	75.600000	227.000000	22.000000	7.390000	441.534144	96.000000	360.250000	56.100000	28.0
max	2015.000000	89.000000	723.000000	1800.000000	17.870000	19479.911610	99.000000	212183.000000	87.300000	2500.0

Polio	Total expenditure	Diphtheria	HIV/AIDS	GDP	Population	thinness 1-19 years	thinness 5-9 years	Income composition of resources	Schooling
2938.000000	2938.000000	2938.000000	2938.000000	2938.000000	2.938000e+03	2938.000000	2938.000000	2938.000000	2938.000000
82.617767	5.924098	82.393125	1.742103	6611.523863	1.023085e+07	4.821886	4.852144	0.630362	12.009837
23.367166	2.400770	23.655562	5.077785	13296.603449	5.402242e+07	4.397621	4.485854	0.205140	3.265139
3.000000	0.370000	2.000000	0.100000	1.681350	3.400000e+01	0.100000	0.100000	0.000000	0.000000
78.000000	4.370000	78.000000	0.100000	580.486996	4.189172e+05	1.600000	1.600000	0.504250	10.300000
93.000000	5.755000	93.000000	0.100000	1766.947595	1.386542e+06	3.300000	3.300000	0.677000	12.300000
97.000000	7.330000	97.000000	0.800000	4779.405190	4.584371e+06	7.100000	7.200000	0.772000	14.100000
99.000000	17.600000	99.000000	50.600000	119172.741800	1.293859e+09	27.700000	28.600000	0.948000	20.700000

The dataset includes important statistics that summarize the information. The mean shows the average value, like a life expectancy of 69.2 years, while the median indicates the middle value, such as 12.3 years of schooling. The standard deviation reveals how much the values vary from the average. The minimum and maximum values show the range, for example, GDP ranging from 1,681.35 to 119,172.74. Percentiles help us understand how values are distributed. Overall, these statistics provide a clear view of health, education, and economic differences among countries.

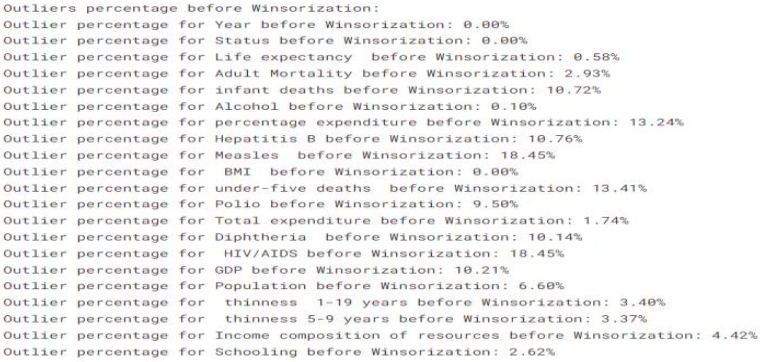
Identifying and Handling Outliers:

The next crucial step in the analysis of life expectancy data is to identify and handle any potential outliers. Outliers are extreme values that deviate significantly from the rest of the data and can potentially skew the results of the analysis. Below are the results.



```
Number of numerical columns: 21
```

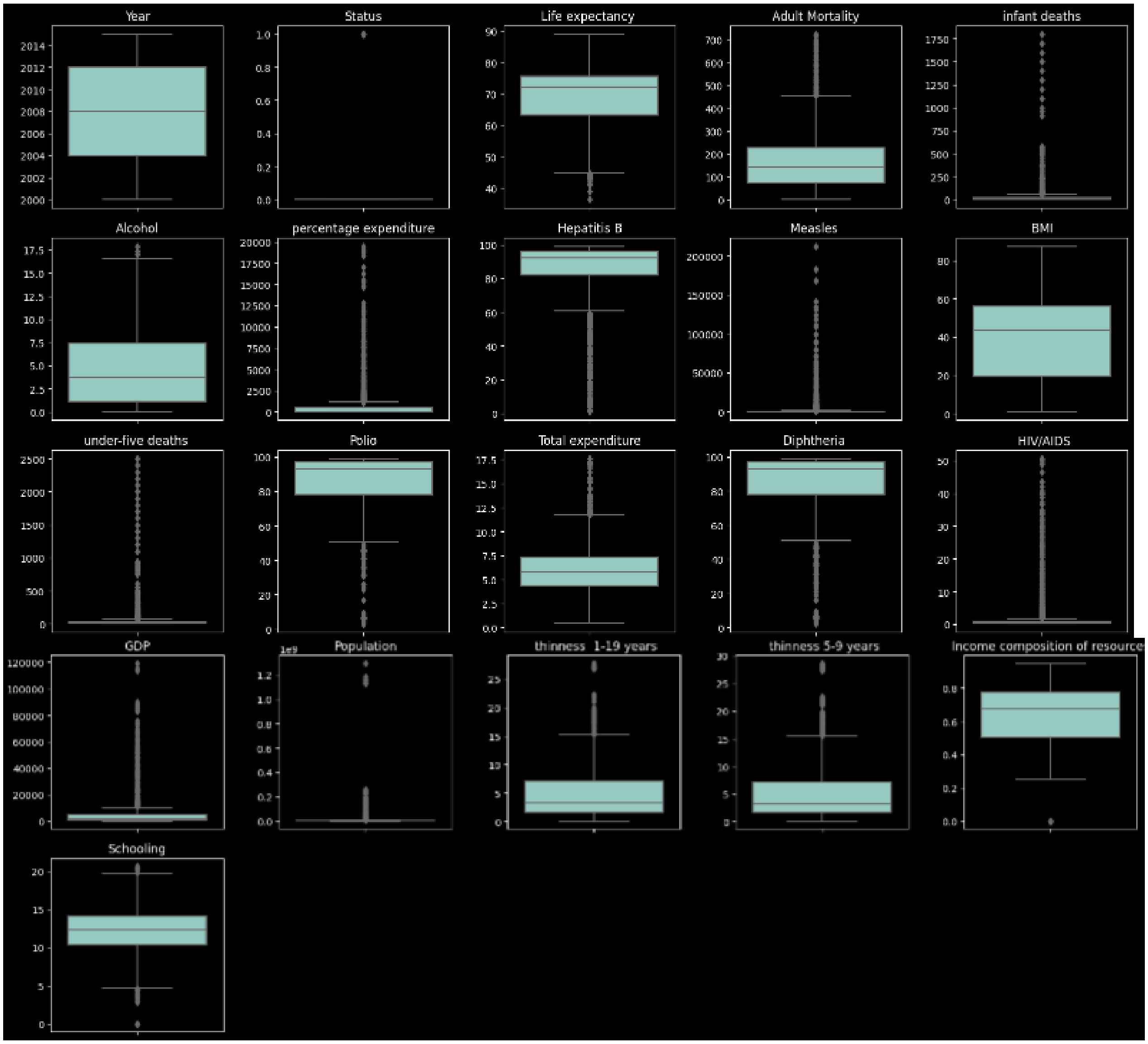
We calculated the outlier percentage for each column to guide effective treatment and ensure data integrity.



```
Outliers percentage before Winsorization:
Outlier percentage for Year before Winsorization: 0.00%
Outlier percentage for Status before Winsorization: 0.00%
Outlier percentage for Life expectancy  before Winsorization: 0.58%
Outlier percentage for Adult Mortality before Winsorization: 2.93%
Outlier percentage for infant deaths before Winsorization: 10.72%
Outlier percentage for Alcohol before Winsorization: 0.10%
Outlier percentage for percentage expenditure before Winsorization: 13.24%
Outlier percentage for Hepatitis B before Winsorization: 10.76%
Outlier percentage for Measles  before Winsorization: 18.45%
Outlier percentage for  BMI  before Winsorization: 0.00%
Outlier percentage for under-five deaths  before Winsorization: 13.41%
Outlier percentage for Polio before Winsorization: 9.50%
Outlier percentage for Total expenditure before Winsorization: 1.74%
Outlier percentage for Diphtheria  before Winsorization: 10.14%
Outlier percentage for  HIV/AIDS before Winsorization: 18.45%
Outlier percentage for GDP before Winsorization: 10.21%
Outlier percentage for Population before Winsorization: 6.60%
Outlier percentage for  thinness  1-19 years before Winsorization: 3.40%
Outlier percentage for  thinness 5-9 years before Winsorization: 3.37%
Outlier percentage for Income composition of resources before Winsorization: 4.42%
Outlier percentage for Schooling before Winsorization: 2.62%
```

We have used Winsorization. This process limits extreme values by capping data at specific percentiles to reduce the effect of outliers. In this project, applying winsorization had minimal impact, showing that the data was already well-distributed with few significant outliers.

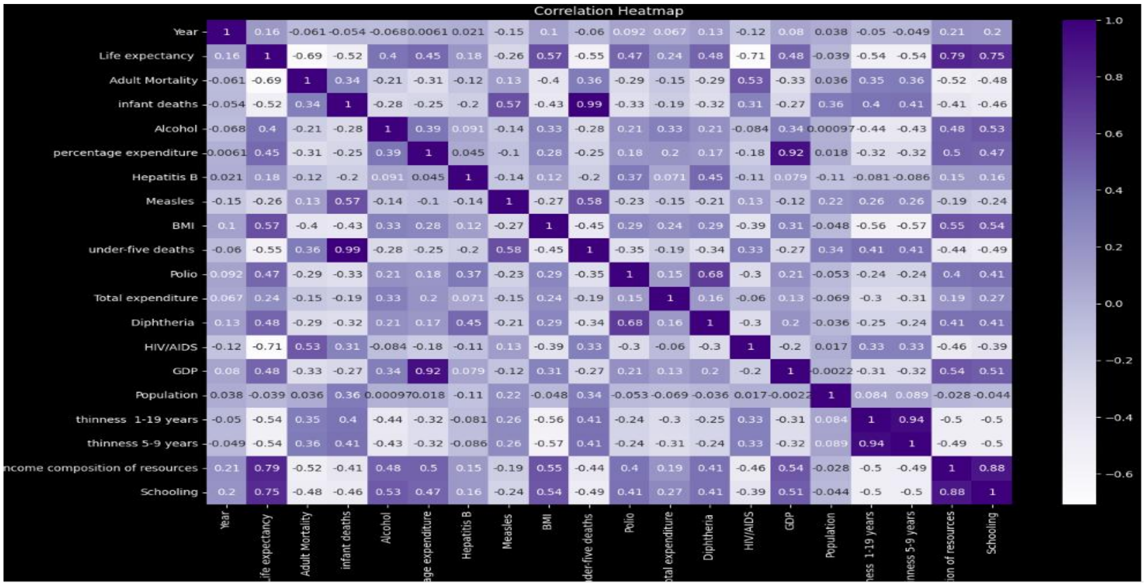
```
Number of numerical columns: 21
```



Before applying visualization techniques, we addressed the outliers in the dataset. Upon applying winsorization, it was observed that the y-axis scale did not change significantly. This indicates that the winsorization process did not drastically alter the extreme values, suggesting that the original dataset was already well-distributed without severe outliers impacting the overall range. As a result, the distribution of data points remained stable before and after winsorization.

These preprocessing steps ensured that the data is more reliable and suitable for further analysis.Now that the data has been cleaned and outliers addressed, we will employ various visualization techniques to explore relationships between variables and identify patterns. These visualizations will provide insights into the distribution of key variables and illustrate correlations, enhancing our understanding of the factors affecting life expectancy.

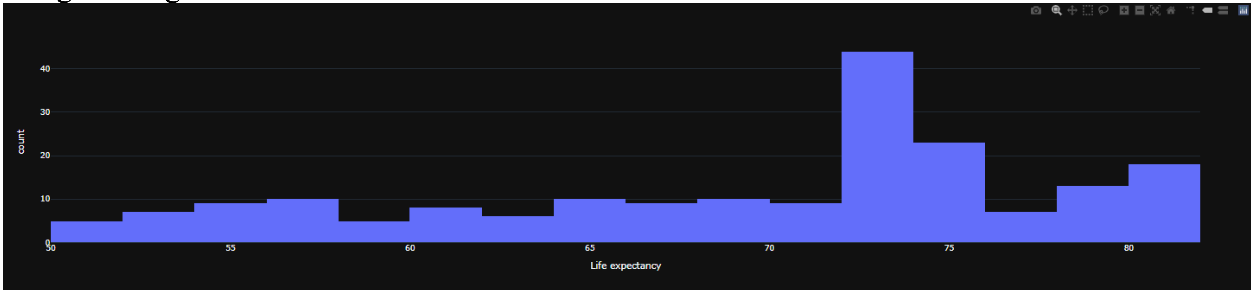
Correlation Heat Map:



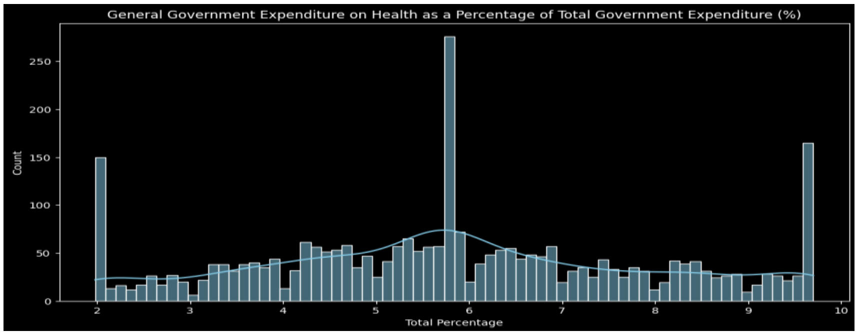
From the above Heat Map, we can say that features like infant deaths, percentage expenditure, under-five deaths, GDP, thinness 1-19 years and thinness 5-9 years are highly correlated.

Histograms:

Next, let's examine the distribution of life expectancy, the primary variable in this dataset, using a histogram.

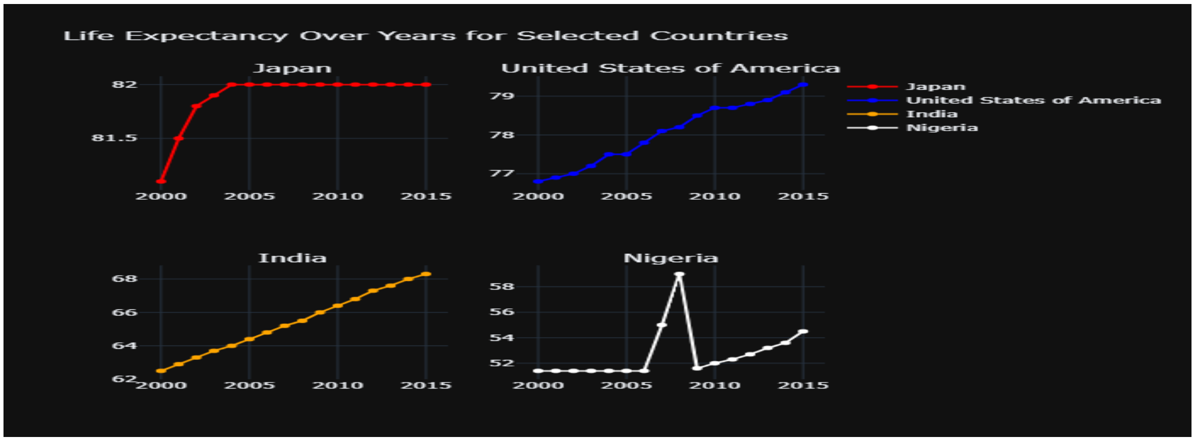


The above histogram illustrates the distribution of life expectancy among countries, revealing that most countries fall between 70 and 75 years. However, some countries have significantly higher or lower life expectancies, indicating considerable global variation likely influenced by factors like healthcare systems, economic development, and lifestyle. The histogram might be slightly skewed to the right, indicating a few countries with exceptionally high life expectancies.



The above histogram shows that most countries allocate 5-6% of government expenditure to health, with the distribution skewed to the right, indicating a few countries with much higher spending. The range spans from around 2% to 10%, with a potential outlier at 10%. The density curve confirms the right-skewed shape, highlighting that while most countries spend a modest portion on health, some allocate significantly more.

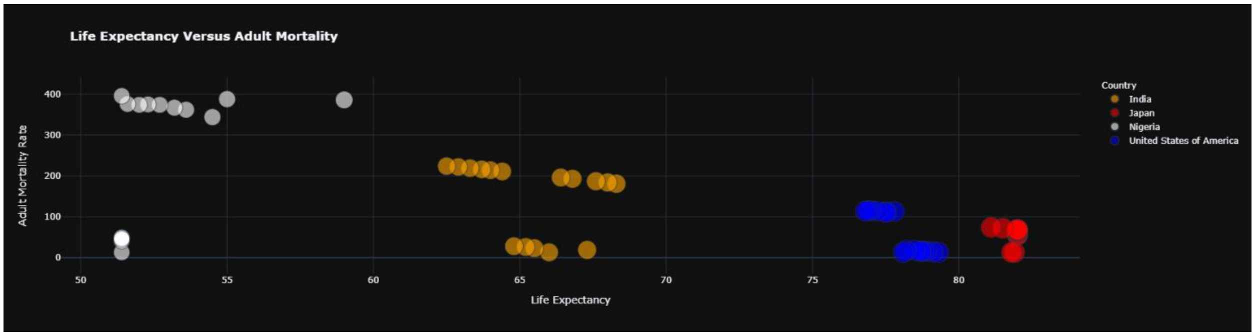
Line chart:



From the results, we observed that, Japan shows steady increase in life expectancy over time, reaching one of the highest levels among the countries. United States of America shows gradual increase in life expectancy, but with a slower pace compared to Japan. When we look at India, it shows Significant improvement in life expectancy, particularly in recent years, but still below the other countries (Japan and USA). Nigeria shows fluctuations in life expectancy, with a notable decline around 2005 followed by a gradual increase.

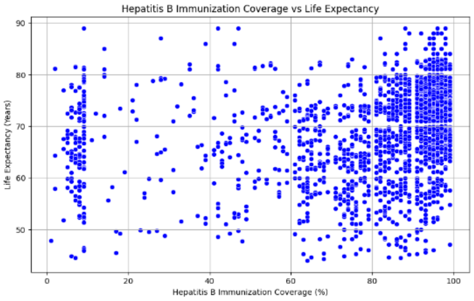
Scatter Plot:

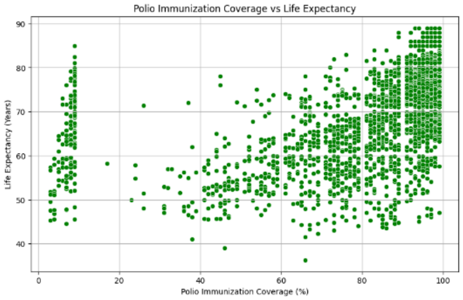
From the below figure we observe that, Japan has highest life expectancy and lowest adult mortality rate. United States of America has Slightly lower life expectancy than Japan, but similar adult mortality rate. India has Lower life expectancy and higher adult mortality rate compared to developed countries. Nigeria has Lowest life expectancy and highest adult mortality rate among the four.

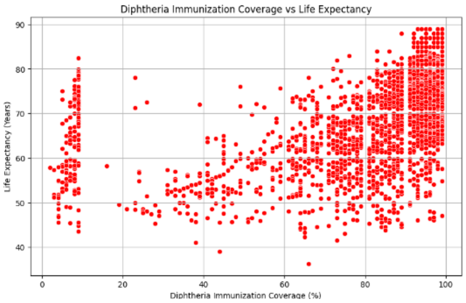


In the below Scatter plot, X-Axis (Life Expectancy), this axis represents the life expectancy of individuals in years. Y-Axis (Adult Mortality), this axis represents the adult mortality rate, which typically indicates the number of deaths per 1,000 adults per year.

Impact of Immunization Coverage on Life Expectancy:

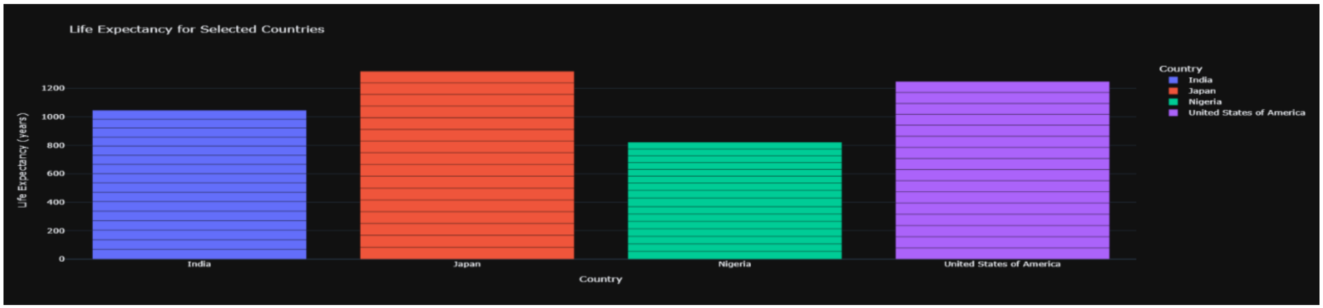






The scatter plots illustrate strong positive correlations between immunization coverage for Hepatitis B, Polio, and Diphtheria and life expectancy. Countries with higher immunization rates generally have longer life expectancies, with most countries showing coverage rates between 80% and 100%. While lower immunization rates correspond to lower life expectancies, the variation in the data suggests that other factors also influence health outcomes. Overall, these findings highlight the importance of immunization in promoting public health and improving life expectancy.

Bar Chart:



The bar chart compares life expectancy in Japan, the United States of America, India, and Nigeria. Japan has the highest life expectancy, reflecting its advanced healthcare and living standards. The United States follows closely, but figures can vary widely within the country. In contrast, India and Nigeria have lower life expectancies, highlighting challenges in healthcare access and living conditions. This visualization underscores the significant disparities between developed and developing nations in public health outcomes.

Conclusion:

Missing values were filled with the median to ensure stability, and outliers were managed using winsorization, maintaining a stable data distribution. Significant correlations were found between infant deaths, healthcare expenditure, and life expectancy, highlighting the impact of health policies. Japan exhibited the highest life expectancy, while India and Nigeria faced challenges due to lower healthcare access and socioeconomic factors. Strong positive correlations between immunization coverage and life expectancy emphasized the critical role of vaccination in public health. The analysis what we have done underscores the importance of healthcare systems and economic factors in influencing life expectancy, revealing significant disparities between developed and developing nations.

Github Repository Address: https://github.com/rishitharani24/Data-Mining---Team-Synergy