BUNCH OF THOUGHTS ON LIFE EXPECTANCY

A STATISTICAL REPORT ON LONGEVITY

Shaan Kohli	260684930
Venkatesh Chandra	260886983

CONTENTS

TABLES

FIGURES

ACRONYMS AND ABBREVIATIONS

BMI	Body Mass Index
DTP3	Diphtheria tetanus toxoid and pertussis
GDP	Gross Domestic Product
GHO	Global Health Observatory
HDI	Human Development Index
PCA	Principal Component Analysis
RSS	Residual Sum of Squares
WHO	World Health Organization

1. Introduction

1.1 Context

How well does your country take care of you? Does living in country X mean that I will live longer? These are highly philosophical questions that some take for granted while for others it remains a serious concern. We are Santa's data scientists and, in the spirit of Christmas, will be trying to give countries some insight into life expectancy because what's nicer than knowing how long you/ your people are likely to live?

1.2 Objectives

From a data scientist's point of view, there are several questions to investigate, such as:

- What factors affect life expectancy?
- Does government expenditure on health care play a role in a longer/shorter life expectancy?
  - How can the government improve life expectancy?
- Do infant and adult mortality rates affect life expectancy?

To ensure our analysis is accurate, data from The Global Health Observatory (GHO) was used from the World Health Organization (WHO) data repository which keeps track of the health status as well as many other related factors for all countries. This project focusses on factors such as immunization, mortality, economic, social and other health-related factors to develop both supervised (Regression Trees) and unsupervised (Clustering/Principal Component Analysis (PCA)) models in order to understand factors affecting life expectancy. This will help a country spend resources on areas which actually improve the life expectancy of the population.

2. Data Description

The dataset used in the project was downloaded in .xlsx format containing approximately 3000 rows of data of 193 countries. For each country, the dataset contained information on life expectancy, economic and health factors for these countries across 16 years.

2.1 Understanding the Predictors

Table 1 shown below briefly describe the predictors.

Table 1: Factors in the Global Health Observatory dataset

Predictor	Description
Status	Indicates if a country is Developed or Developing
Life expectancy	Life expectancy in years
Adult mortality	Mortality rate (probability of dying between 15 and 60 per 1000 population)
Infant deaths	Number of infant deaths per 1000 population
Alcohol consumption	Per capita alcohol consumption (15+) per 1000 population

Percentage expenditure on health	Expenditure on health as a percentage of Gross Domestic Product (GDP) per capita (%)
Hepatitis B	HepB immunization coverage among 1-year-olds (%)
Measles	Number of reported cases of measles per 1000 population
Body Mass Index	Average Body Mass Index (BMI) of entire population
Under five deaths	Number of under five deaths per 1000 population
Polio immunization coverage	Pol3 immunization coverage among 1-year-olds (%)
Total expenditure on health	General government expenditure on health as a percentage of total government expenditure (%)
Diphtheria immunization coverage	Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)
HIV/AIDS	Deaths per 1000 live births HIV/AIDS (0-4 years)
Gross Domestic Product	GDP per capita in USD
Population	Population of the country
Thinness among 10-19 years	Prevalence of thinness among children and adolescents for Age 10 to 19
Thinness among 5-9 years	Prevalence of thinness among children and adolescents for Age 5 to 19
Income Composition of Resources / Human Development Index	Human Development Index (HDI) in terms of income composition of resources (ranges from 0 to 1)
Schooling	Number of years of Schooling (in years)

2.2 Data Pre-processing

The distribution, relevance and quality of each predictor were studied and the relationship between the variables was examined to understand the data points. While the distribution helps to identify patterns and relationships among variables, the relevancy and quality aim to evaluate if including a predictor adds any value to the statistical models built. The correlation heat map (Fig. 1), and pairs panels plot (Fig. 2) helped identify the collinearity between predictors.

The following predictors were removed from the dataset as a result of the data cleaning steps:

- **Hepatitis B**: The Hepatitis B vaccine is provided as a combined vaccine with diphtheria, hence the predictor did not add significant value
- **Measles per capita**: This predictor contains some values which are outside the domain
- **Infant deaths**: This predictor contains some values which are outside the domain
- **Percentage expenditure on health as a percentage of GDP**: This predictor is collinear with GDP (Fig. 2)
- **Under-five deaths**: This predictor contains some values which are outside the domain
- **Thinness among age 5-9**: This predictor is collinear with Thinness among age 10-19 (Fig. 2)
- **Schooling**: This predictor is collinear with the HDI (Fig. 2)

The following data points were removed from the dataset as a result of data quality checks:

- **2015 data**: Data for the year 2015 was not up to date and not available for most of the countries (183 rows removed).
- 41 countries did not have population data. 608 rows removed
- **Data for alcohol consumption** was not available for South Sudan. 15 rows removed
- **GDP data for Syrian Arabic Republic** was not available. 15 rows removed
- **Data for India** was removed because the data points are outliers. India is a special case as it follows an unusual pattern with respect to life expectancy, economic and health factors. 15 rows removed
- **Life expectancy data for Palau and Tuvalu** was not available. 2 rows removed

Chapter 2: Data Description

After performing the data cleaning steps, the analytical dataset had 2100 rows of data for 140 countries and 19 features. The list below has information regarding the distribution and statistics of the features included in the analytical dataset, which is also shown in Fig. 3.

- **Adult mortality**: The distribution exhibits some skewness towards the right. The median adult mortality per 1000 individuals is 147.
- **Alcohol consumption**: The distribution exhibits some skewness towards the right. The median value of alcohol consumption is 4.13 litres.
- **Polio immunization coverage**: The distribution is skewed to the left. The median value of polio immunization is 92%.
- **Total expenditure on health**: The distribution is normal with the median value equal to 5.9%.
- **Diphtheria immunization coverage**: The distribution is skewed to the left. The median value of diphtheria immunization is 92%.
- **HIV/AIDS**: HIV/AIDS is a controlled disease (concentrated around the median values) with the median value being equal to 0.1%.
- **GDP**: The distribution is right-skewed which means the data is biased by the presence of developing countries. The median GDP per capita is 1473 US dollars.
- **Population**: The distribution is right-skewed with the median value equal to 1.3 million.
- **Thinness among 10-19 years**: The distribution is partially right-skewed with the median value equal to 2.8%.
- **HDI**: The distribution is normal. The median value of HDI is 0.66.

3. Model Selection and Methodology

3.1 Unsupervised Machine Learning

K-means clustering is a type of unsupervised machine learning algorithm which helps identify hidden patterns in data by grouping data points based on similarity. The motivation behind using this algorithm for the dataset was to group the countries across their years to determine the effect that a government's expenditure on health and economy has on life expectancy. This would also help determine the cause behind an increase in life expectancy of a country across different years.

3.2 Principal Component Analysis

The plan was to understand the underlying structure of the data. PCA is one of the most powerful tools that help understand the variability in the data. It also helps in measuring the data in terms of its principal components. PCA on the dataset is done to determine the factors that lead to a higher life expectancy of a country and identify those which have a negative effect.

3.3 Regression Tree

A regression tree is a supervised machine learning method that predicts the outcome, in our case life expectancy, based on the input of various variables. The regression tree gives the ability to not only find the relationship between life expectancy, and health and economic factors, but also the level of importance of each predictor. It provides a good intuition as to which is the leading contributor to both high and low life expectancy. A regression tree was run using life expectancy as the target variable and using adult mortality, alcohol, BMI, polio, total expenditure, diphtheria, HIV/AIDS, GDP, population, thinness from age 10 to 19 and income composition of resources. The tree will be optimized for the ideal cp value (i.e. the value at which the tree should stop trying to split when splitting improves Residual Sum of Squares (RSS) by said cp value).

4. Results

4.1 Interpretation of Principal Components

The PCA variability plot (Fig. 4) gives an idea of the variability explained by the principal components. The increase in explanation of variability is gradual as the number of components increases. The cumulative variability plot follows a logarithmic distribution. About 40% of the variability can be explained by the first principal component and a total of 6 components are required to capture 80% variation (Fig. 5 – Fig. 8).

The first principal component divides the data points on life expectancy, HDI, the prevalence of thinness among 10-19 years and adult mortality rates. It explains around 40% of the variation in the data. The second principal component tells that a further 15% variability can be explained by the number of deaths due to HIV-AIDS and government expenditure on health. The third component tells us that polio and diphtheria immunization coverage help explain another 15% variation in the data. Th population is the next predictor which helps in explaining ~12% of the variation and is a part of the fourth principal component.

From the PCA plot (Fig. 9), it can be inferred that high government spending on health as a percentage of GDP, good coverage of Diphtheria and Polio, higher HDI and high per capita GDP are the factors which drive life expectancy to increase. On the other hand, a high number of HIV-AIDS related deaths (per 1000 births), higher probability of deaths of people between 15- and 60-years age (per 1000 population) and prevalence of thinness among children and adolescents of age 10 -19 years have a negative effect on life expectancy.

4.2 Insights from Clusters Coupled with Principal Component Analysis

From the PCA plot (Fig. 9) and the Cluster Information Table (Fig. 10), we got some interesting insights, which are described below.

- **Cluster 1: Developed Countries:** Cluster 1 is composed of the best performing years of countries over the last 15 years. It is found that the average life expectancy is high (~78 years), and HDI has a higher value. The population is controlled with low cases of HIV-AIDS related deaths. They also have good coverage of immunization of polio and diphtheria and less prevalence of thinness among children aged between 10 - 19 years. During these "best performing" years of respective countries, per capita GDP was high with high expenditure on health. Alcohol consumption is the highest across these years. HDI is quite low for these cases (0.83 on an average)
- **Cluster 2: Developing Countries:** Cluster 2 comprises the developing years of the respective countries in the cluster. The average life expectancy is 71 years. The population of the countries is high, and the GDP is low, which is further worsened by a lower expenditure on health resources. Although the diseases are controlled which is evident from a good coverage of immunization for polio and diphtheria, and low cases of HIV-AIDS related deaths, the prevalence of thinness among children aged between 10-19 years and the adult mortality rate is high. During these years, the countries might not have spent on the overall health of their citizens.
- **Cluster 3: Underdeveloped Countries:** Cluster 3 includes the worst performing years of countries in terms of life expectancy. On average, life expectancy is just 57 years. The

countries during these years have failed to boost the economy of the nation and the health of their citizens. Low GDP coupled with low government expenditure on health has resulted in a high number of cases of HIV-AIDS related deaths, low coverage of immunization for polio and diphtheria, and high adult mortality rate. HDI is quite low for these cases (0.4 on an average) adult mortality rate is high. During these years, the countries might not have spent on the overall health of their citizens.

The curious case of Alcohol consumption

It is interesting to note that during the best performing years of countries, alcohol consumption per capita is quite high as compared to the other cases. Alcohol is either banned or not so popular in most of the countries who are a part of cluster 3.

At first glance, it is tempting to jump into conclusions that alcohol consumption increases life expectancy. On the other hand, there are numerous studies done on the effects of alcohol on health. Therefore, implying causation from correlation is dangerous. Usually, the developed countries with a high life expectancy consume more alcohol. This is driven by factors such as better quality of life which is closely associated with higher consumption of liquors.

4.3 Regression Tree Analysis

After running the analysis (Fig. 11), it was found that the ideal threshold for tree split was 0.004.

Worst Case Scenario

The most significant predictor for life expectancy was the income composition of resources. This means that if the relative income share of each income source for a country (expressed as a percentage of the aggregate total income of that group or area) is below 57%, the less likely an individual is expected to live in that country. If a said country also has more than 1.6 births per 1000 related to HIV/AIDS, then individuals are likely to have an even lower life expectancy (second strongest predictor for lower life expectancy). Given these conditions, if said country also has the number of adult deaths (between 15 and 60 years) above 493 per 1000 people, the life expectancy of the said country will be lower.

Chapter 4: Results

Best Case Scenario

If the relative income share of each income source for a country (expressed as a percentage of the aggregate total income of that group or area) is above 57%, the more likely an individual is expected to live in that country. If a said country has a relative income share of each income source to be greater than 80%, then expected life expectancy is greater. Finally, if the country does not have more than 106 adult deaths (aged between 15-60), the expected life expectancy for that country is even higher.

5. Conclusions and Recommendations

From our understanding, we can say with confidence that the world is not perfect. However, this is not a surprise. Classification approaches such as PCA and K-means clustering provided consistent results in that most countries are grouped similarly. Countries with high government spending on health as a percentage of GDP, good coverage of Diphtheria and Polio, higher HDI and high per capita GDP are usually grouped together while those with a high number of HIV-AIDS related deaths (per 1000 births), higher probability of adult deaths and prevalence of thinness among children and adolescents (between 10 -19 years of gage) are usually grouped together.

Using the clustering analysis and the results of our PCA analysis, predictors in the aforementioned groups can be categorized as a developed, developing or, underdeveloped country. On a year to year basis, countries considered as developing can be described as either emerging into a developed country or, sadly, the opposite (Fig. 12). In 2014, countries such as Russia and Trinidad and Tobago have been able to improve their status by developing factors such as lower HIV/AIDS rates, lower thinness between the years of 10-19, greater income composition of resources and an increase in total health expenditure by the government (Fig. 12).

5.1 Our Take on Life Expectancy

Life expectancy is not something that should be taken lightly. It is quite surprising to see that some countries do not have basic care about prevention against preventable diseases. In an era where food wastage is at an all-time high, life expectancy should not decrease due to famine (thinness). However, there is optimism, some countries have been able to manage resources and provide greater care to their people.

5.2 What should the Government do with these Results?

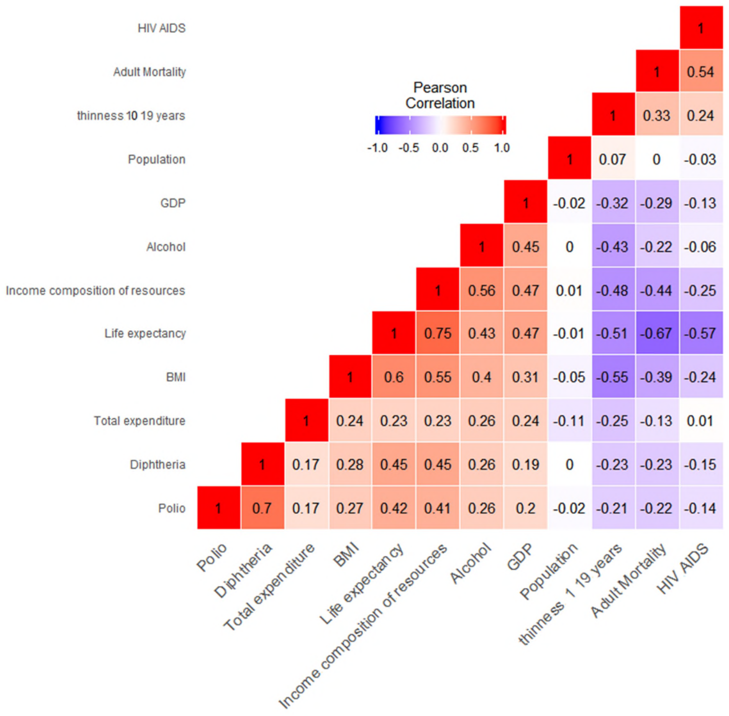
Governments can use these insights to develop policies and regulations to better allocate resources towards helping people improve quality of life through healthcare. While some countries are not in a position to improve due to a lack of resources, countries should look to create awareness with these results. Awareness, in the age of social, can lead to public funding, thereby providing resources to increase basic health care in undeveloped countries.

5.3 Parting Notes and Next Steps

So, what is the best present Santa can give? It's the insights, the holy grail. As Santa's data scientists, we believe that the results and analysis provided above can incentivize countries to look into their respective government's healthcare system. Information is endless and we should not be waiting for Christmas once again. Santa has provided the first steps and it's our responsibility to work collectively to improve. In the future, it might be noteworthy to see if there is data regarding the impact of social media on government funding and operations. Is social media an adequate outlet when voicing issues with government legislation and policies? Are the voices of young individuals heard? Are governments making sure that problems tackled the right way?

6. Appendix

6.1 Correlation Heat Map



	Polio	Diphtheria	Total expenditure	BMI	Life expectancy	Income composition of resources	Alcohol	GDP	Population	thinness 1 19 years	Adult Mortality	HIV AIDS
HIV AIDS												1
Adult Mortality											1	0.54
thinness 10 19 years										1	0.33	0.24
Population									1	0.07	0	-0.03
GDP								1	-0.02	-0.32	-0.29	-0.13
Alcohol							1	0.45	0	-0.43	-0.22	-0.06
Income composition of resources						1	0.56	0.47	0.01	-0.48	-0.44	-0.25
Life expectancy					1	0.75	0.43	0.47	-0.01	-0.51	-0.67	-0.57
BMI				1	0.6	0.55	0.4	0.31	-0.05	-0.55	-0.39	-0.24
Total expenditure			1	0.24	0.23	0.23	0.26	0.24	-0.11	-0.25	-0.13	0.01
Diphtheria		1	0.17	0.28	0.45	0.45	0.26	0.19	0	-0.23	-0.23	-0.15
Polio	1	0.7	0.17	0.27	0.42	0.41	0.26	0.2	-0.02	-0.21	-0.22	-0.14

Figure 1: Correlation Matrix of Variables

6.2 Collinearity Test – Pairs Panels

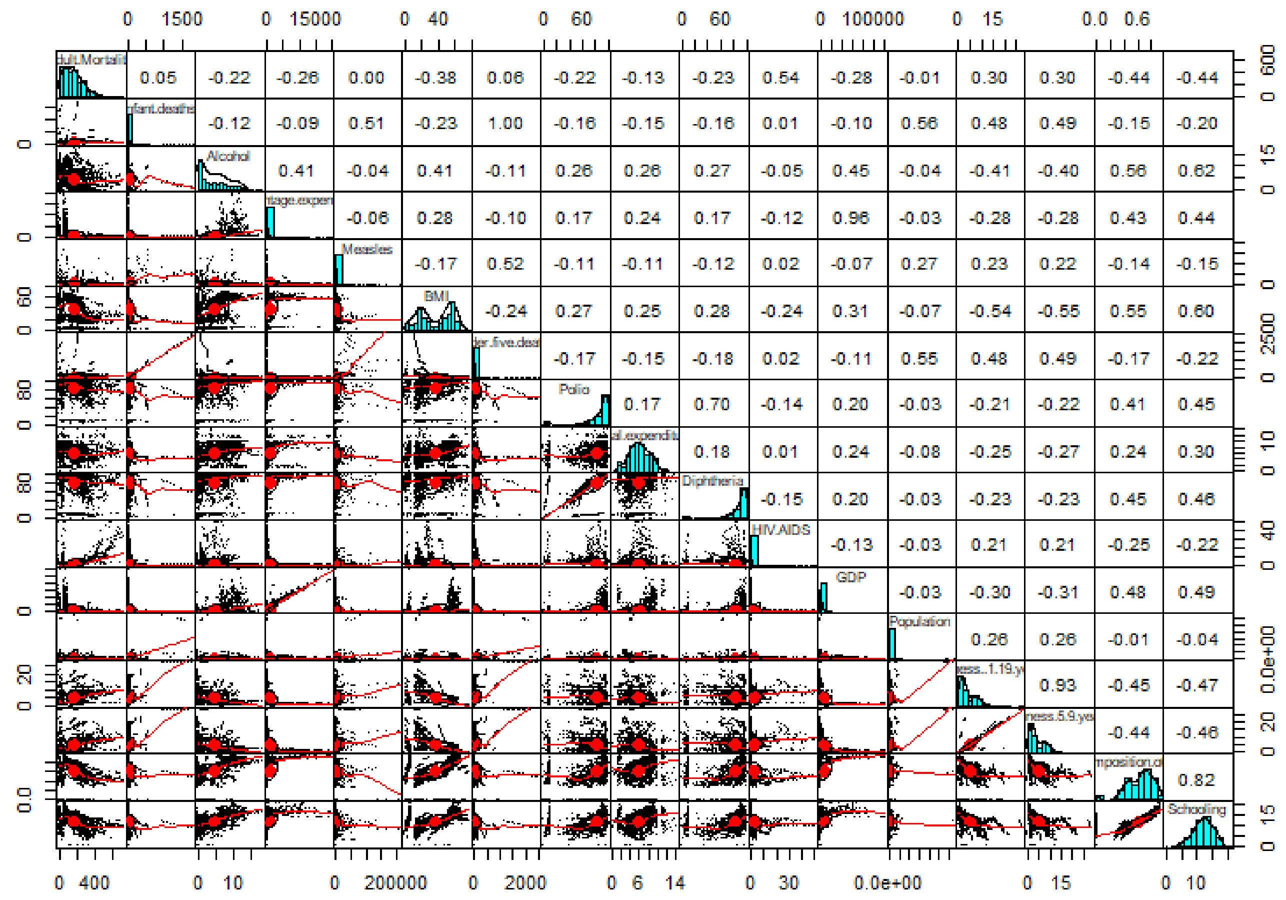
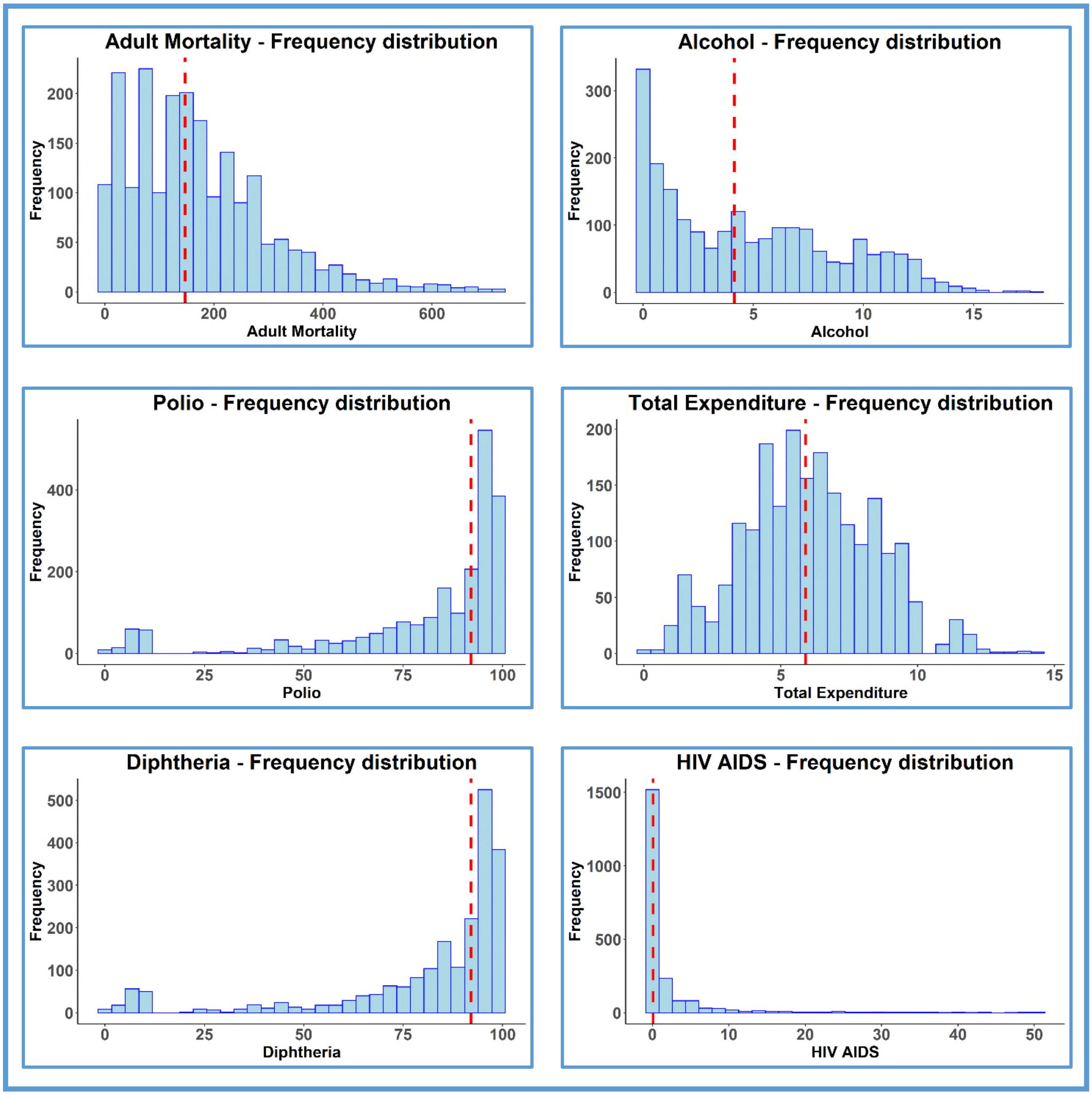


Figure 2: Results of Pairs Panels showing Collinearity among Predictors

6.3 Frequency Distribution Plots for Predictors



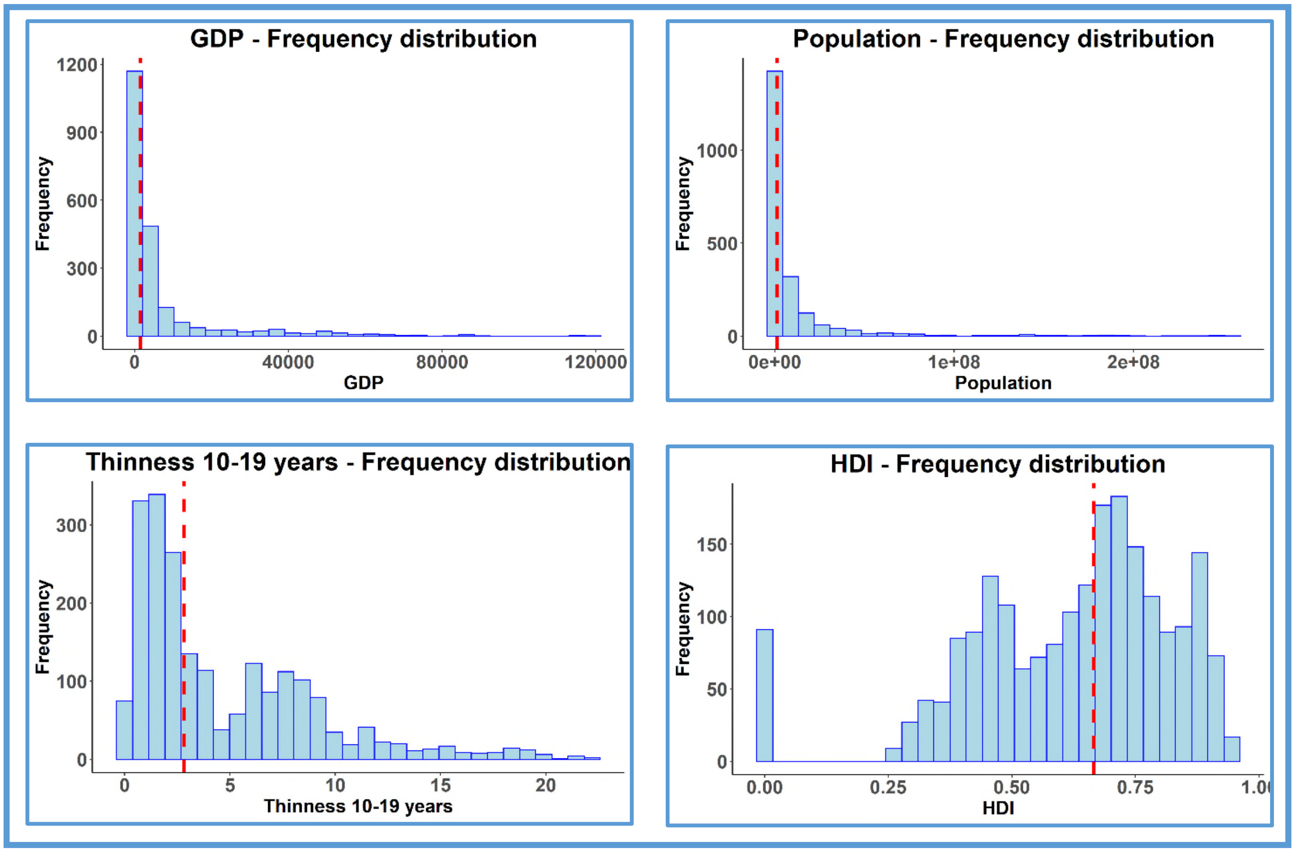


Figure 3: Frequency Distribution Plots for Predictors

6.4 Principal Component Analysis Variability Plots

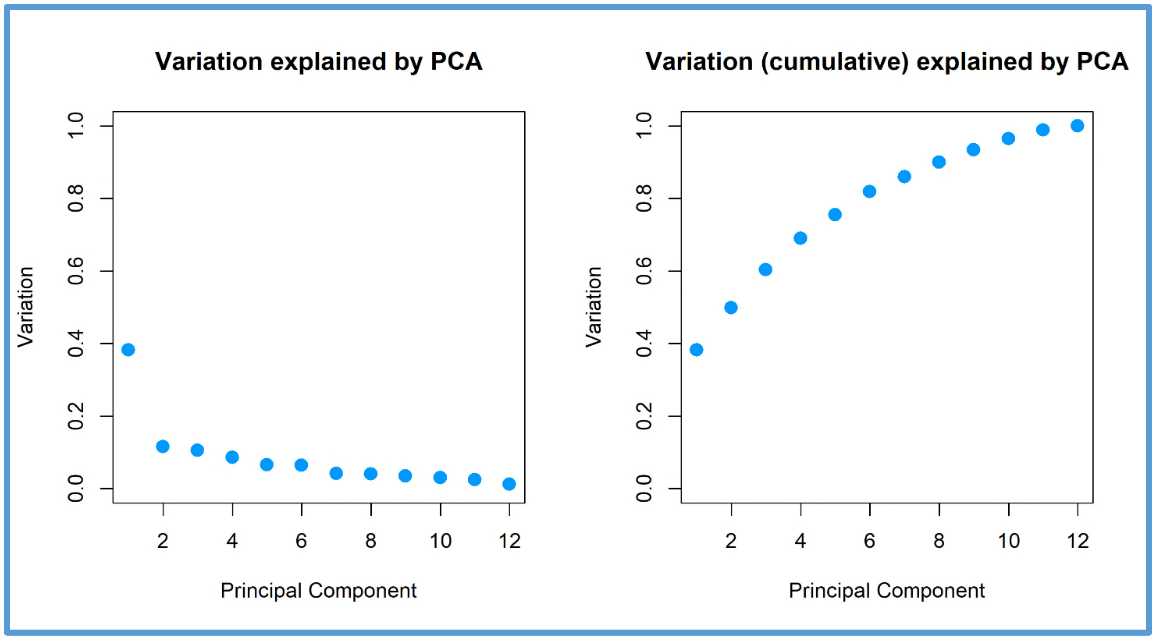


Figure 4: Principal Component Analysis Variability Plot describing Percentage Variation Explained

6.5 Principal Component Analysis Interpretation Plots

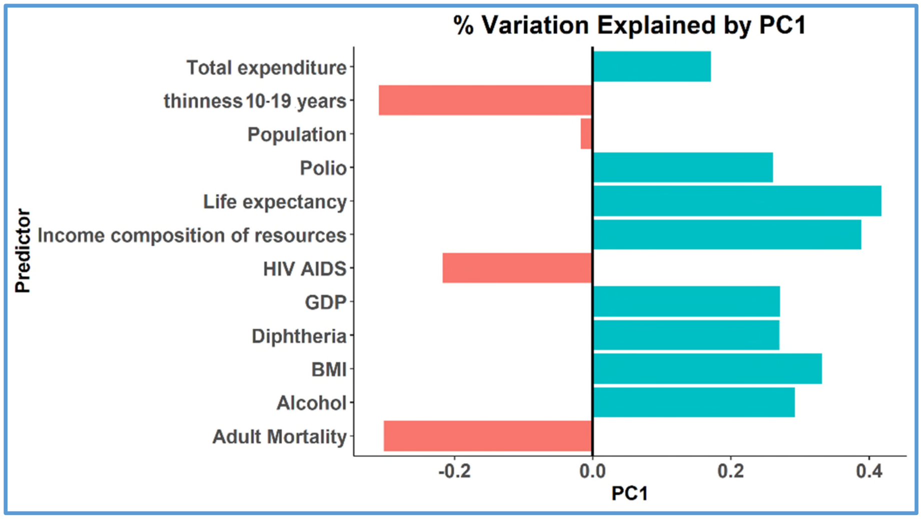


Figure 5: Percentage Variation explained by Principal Component 1

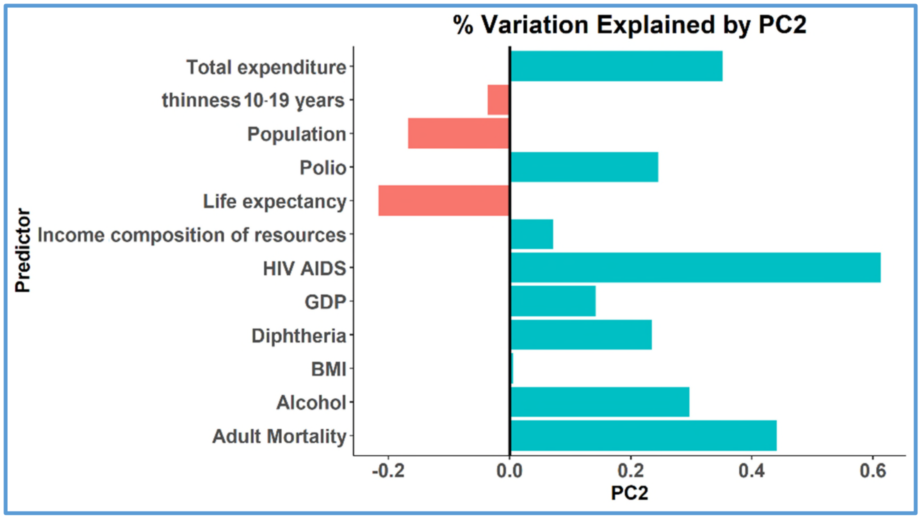


Figure 6: Percentage Variation explained by Principal Component 2

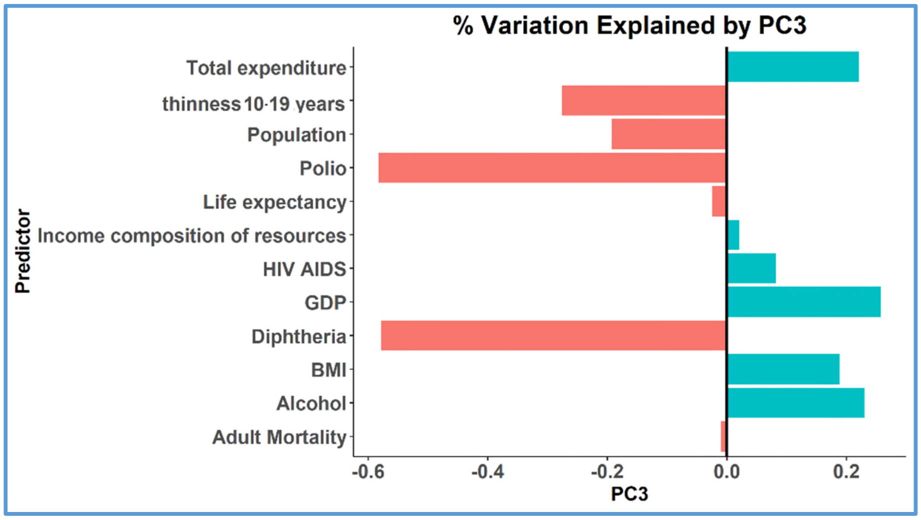


Figure 7: Percentage Variation explained by Principal Component 3

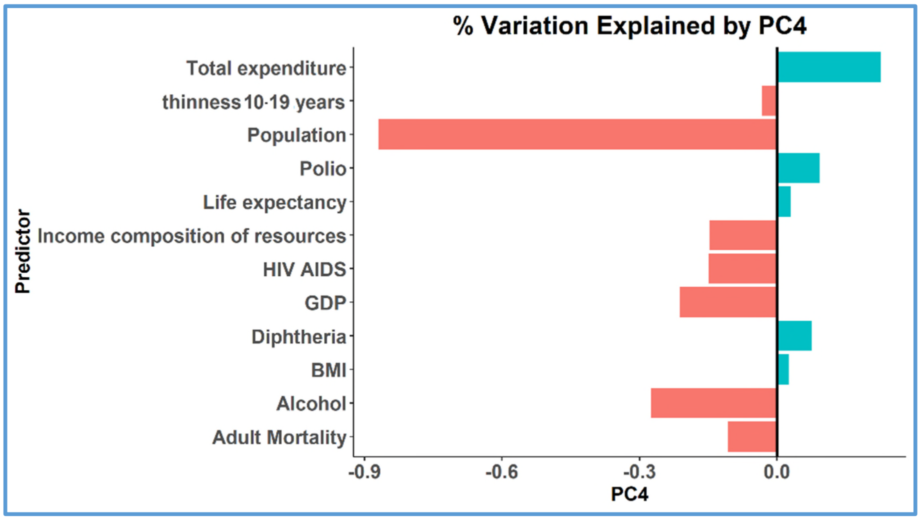


Figure 8: Percentage Variation explained by Principal Component 4

6.6 Principal Component Analysis Two-Dimensional Plot

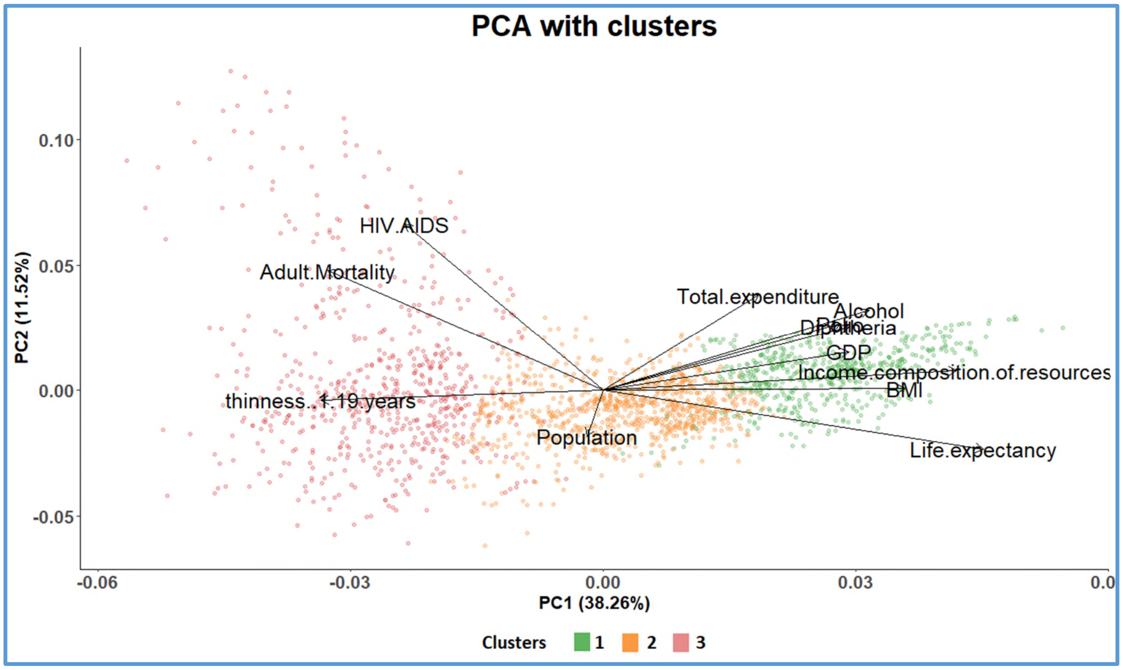
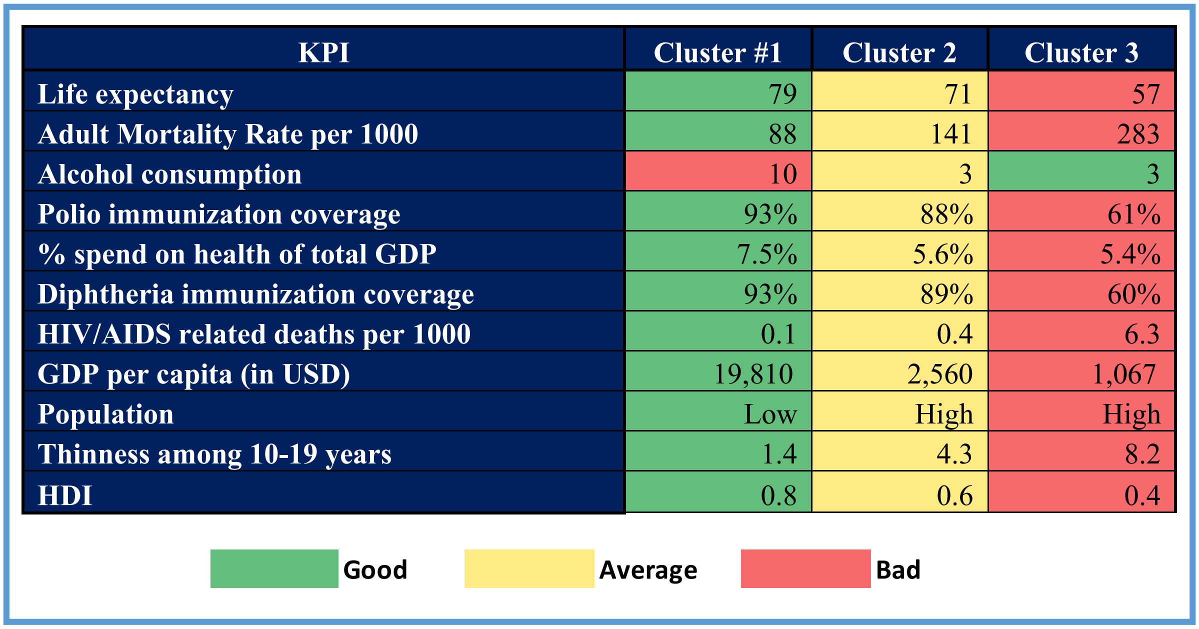


Figure 9: PCA Two-Dimensional representation of observations colored by Clusters

6.7 Cluster Information Chart



KPI	Cluster #1	Cluster 2	Cluster 3
Life expectancy	79	71	57
Adult Mortality Rate per 1000	88	141	283
Alcohol consumption	10	3	3
Polio immunization coverage	93%	88%	61%
% spend on health of total GDP	7.5%	5.6%	5.4%
Diphtheria immunization coverage	93%	89%	60%
HIV/AIDS related deaths per 1000	0.1	0.4	6.3
GDP per capita (in USD)	19,810	2,560	1,067
Population	Low	High	High
Thinness among 10-19 years	1.4	4.3	8.2
HDI	0.8	0.6	0.4

🟩 Good 🟨 Average 🟥 Bad

Figure 10: Clusters and their Average Values for Metrics

6.8 Regression Tree Chart

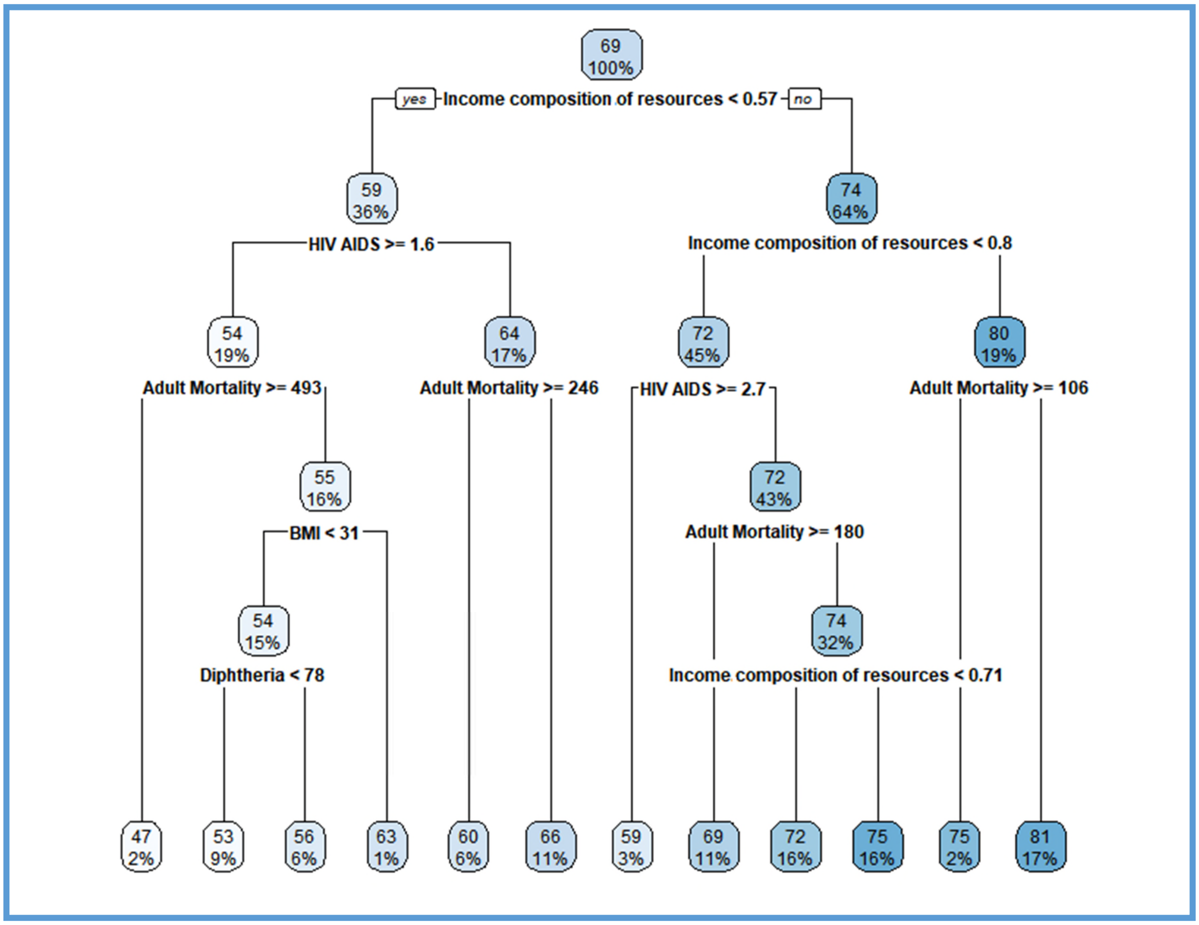


Figure 11: Regression Tree Output showing Splits based on Purity Maximization

6.9 One World View (2014)

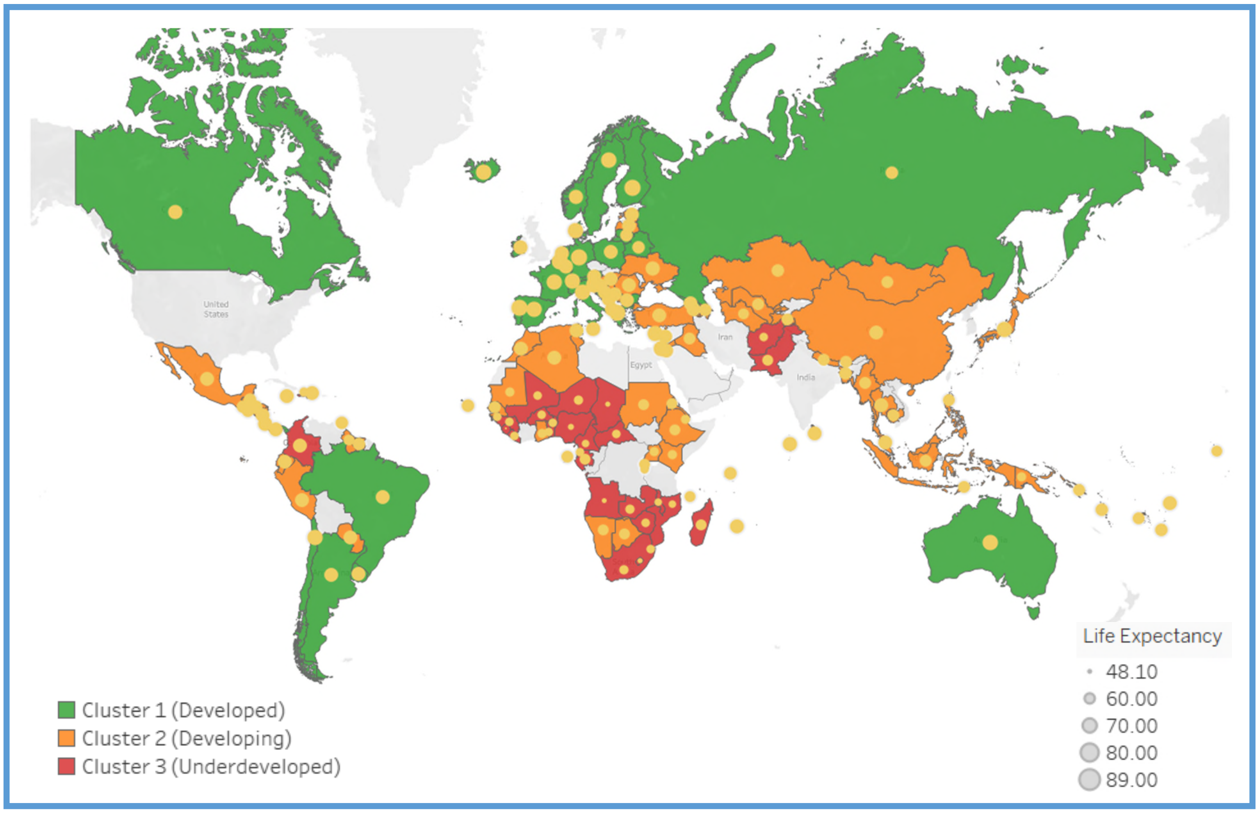


Figure 12: Representation of Countries colored by their Clusters; size of dots represents Life Expectancy