Life Expectancy Analysis

Samuela Abigail Mathew 71762108039
Nivetha S S 71762108030
Haripriya V 71762108011
Pavithra G 71762108031

Department of Artificial Intelligence and Data Science

Coimbatore Institute of Technology

Abstract- This is a report holding studies on factors affecting a person's life expectancy worldwide and related findings.

Keywords- life expectancy, income composition, adult mortality, schooling, BMI

I. Introduction

Life expectancy refers to the average number of years a person who has reached a certain age can expect to live based on actuarial data. It is a reliable snapshot of population health and mortality in a given country, territory, or geographic area. The analysis was done using the Life Expectancy dataset of WHO to study the immunization factors, mortality factors, economic factors, social factors and other health related factors which affect life expectancy of a person with the aim of finding the predicting factor contributing to lower value of life expectancy and help in suggesting areas of countries to which importance in improving life expectancy of its population should be given.

The dataset has the following attributes-

Attribute name	Datatype
Country	Categorical (Nominal)
Year	Categorical (Ordinal)
Status	Categorical (Nominal)
Life expectancy (age)	Numerical (Continuous)
Adult mortality (between 15 and 60 years per 1000 population for both sexes)	Numerical (Discrete)
Infant death (per 1000 population)	Numerical (Discrete)
Alcohol (per capita consumption in litres)	Numerical (Continuous)
Percentage expenditure (on health as % of Gross Domestic Product per capita)	Percentage
Hepatitis B (immunization coverage among 1-year-olds)	Percentage
Measles (per 1000 population)	Numerical (Discrete)
BMI (average of entire population)	Numerical (Continuous)
Under-five deaths(per 1000 population)	Numerical (Discrete)
Polio (immunization coverage among 1-year-olds)	Percentage
Total expenditure (general government expenditure on health)	Percentage
Diphtheria (immunization coverage among 1-year-olds)	Percentage
HIV/AIDS (deaths per 1000 live births in 0-4 age)	Percentage
GDP (USD)	Numerical (Continuous)
Population	Numerical (Discrete)
Thinness 1-19 years	Percentage
Thinness 5-9 years	Percentage
Income composition of resources	Range (0 to 1)
Schooling (years)	Numerical (Continuous)

The dataset has 22 attributes (columns) and 2938 rows where null values in dataset were replaced using interpolation. No rows or columns were dropped as there were not significant number of null values, with number of null values being highest in Population, GDP, and Hepatitis B columns. The analysis was performed on whole dataset (global analysis) as well as by continent-wise splitting (continent-wise analysis). As Status is a categorical variable, it was converted into numerical variable where 0 means developed and 1 means developing.

II. Global Analysis

There are totally 193 countries in this dataset with over 80% being developing countries and the rest being developed countries with totally 32 developed countries, with a mean life expectancy of 69.2 years. It contains data of 15 years from the year 2000-2015 for each country in the dataset. The following were observed over those 15 years from the dataset-

Over 55% countries have life expectancy greater than 70 years, 4% have between 20-50 years, and remaining 40% have between 50-70 years with the global average lifespan being 69 (see fig 2.1).

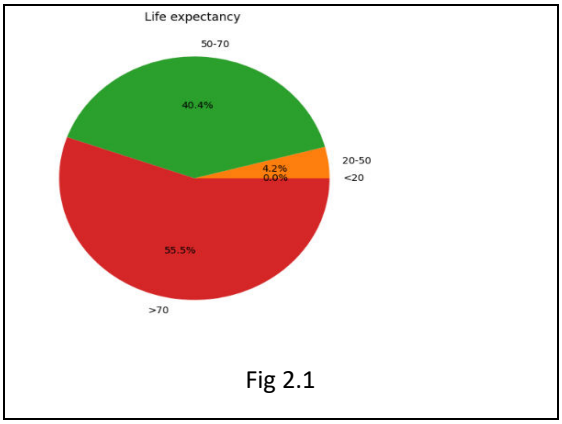

Fig 2.1

The global minimum and maximum life expectancy is 36 years (Haiti) and 89 years (Belgium, Finland, France, France, Germany, Italy, New Zealand, Norway, Portugal, Spain, Sweden) respectively. From box plot of year vs life expectancy, middle 50% of life expectancy lies between 60 and 75 years of age. The median life expectancy is 72 years and the mode of that is 73 years, with the interquartile range value being 12.5 years, which is a low value suggesting that data points show minimum deviation from median and thus less dispersion/ variance for life expectancy.

A. Outlier Detection

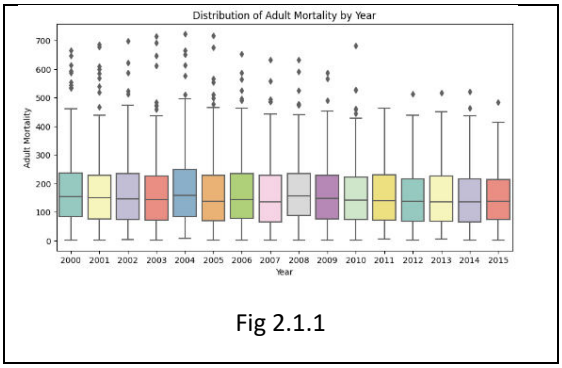

Fig 2.1.1

Outliers were detected in large numbers in all attribute columns (fig 2.1.1) except alcohol and BMI which had no outliers.

B. Correlated Attributes

A correlation matrix (heatmap) was created to identify factors which affected life expectancy throughout the globe (fig 2.2.1). From the correlation map we can see that life expectancy has strong positive correlation (directly proportional to) with BMI, Polio, Diphtheria, GDP, income composition of resources, and schooling with correlation being the strongest with BMI, income composition of resources, and schooling. Life expectancy has strong negative correlation (inversely proportional to) with adult mortality, HIV/AIDS , status and thinness (both age categories) with correlation being strongest for adult mortality.

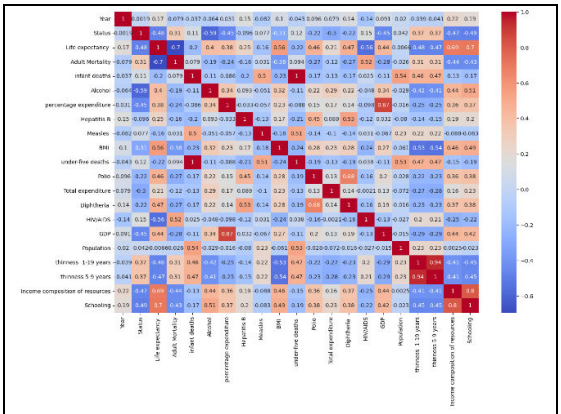

Fig 2.2.1

It was also found that a strong positive correlation exists between percentage expenditure and GDP, thinness in 5-9 years and thinness in 1-19 years, and income composition of resources and schooling. The correlation between infant deaths and under-five deaths is 1, having a linear relationship on applying linear regression.

A strong negative correlation exists between status and alcohol, and BMI and thinness (both categories). On applying logistic regression to status and alcohol, we get a sigmoid curve which indicates probability of status being developed decreasing with decrease in alcohol consumption (fig 2.2.2). On plotting distribution curves of attributes with respect to life expectancy, most of them were right skewed bell-shaped distributions with the rest being bimodal or left skewed bell-shaped distribution.

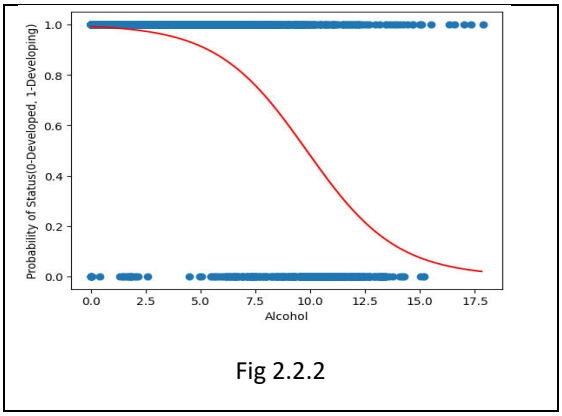

Fig 2.2.2

C. Exploring Strongly Correlated Attributes

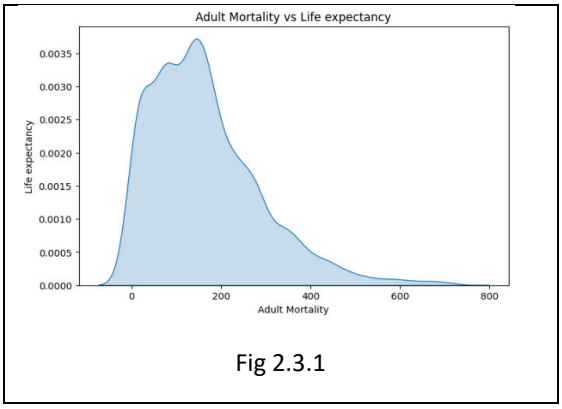

Fig 2.3.1

Adult mortality which has strongest negative correlation with life expectancy has a right skewed bell-shaped distribution (fig 2.3.1). Most of the countries with higher life expectancy have adult mortality less than 200. Now we'll explore strong positively correlated attributes-

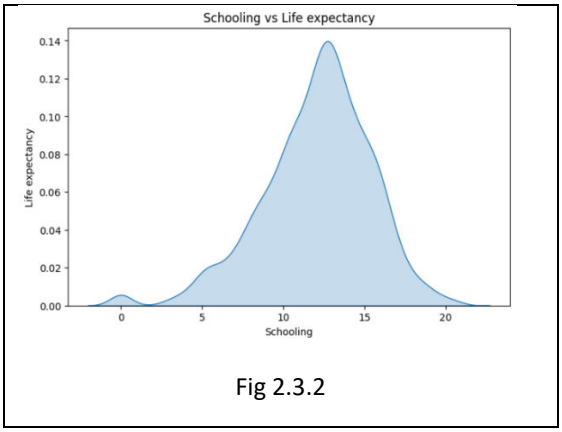

Fig 2.3.2

It was noted that schooling in countries of higher life expectancy is more than 12 and that of lower expectancy countries is less than 12 years of schooling. Thus, it indicates that education plays a crucial role in health and well-being, leading to longer life expectancies.

The bimodal distribution of BMI (fig 2.3.3) may indicate health disparities such as presence of two distinct subpopulations with different health characteristics or lifestyle factors like different lifestyle patterns or socioeconomic factors associated with BMI and life expectancy. It could indicate the presence of distinct groups with different dietary habits, physical activity levels, or access to healthcare, which in turn influences BMI and overall health outcomes.

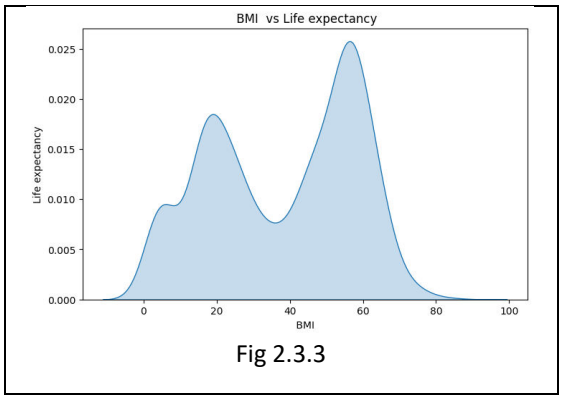

Fig 2.3.3

Income composition of resources has a trimodal distribution (fig 2.3.4) which may suggest the existence of three distinct subpopulations within the data, each with its own unique characteristics. These subpopulations could be related to different demographic groups, geographical regions, or other factors that influence income composition of resources.

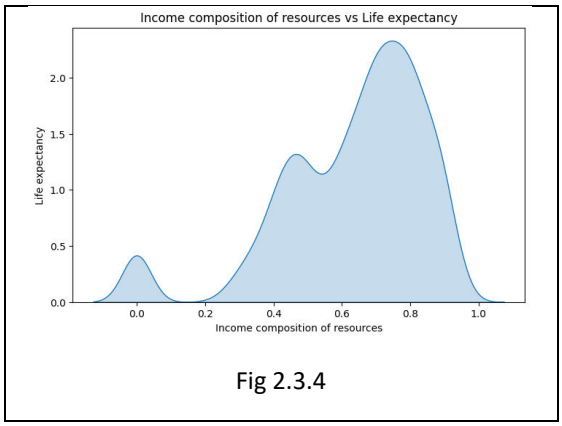

Fig 2.3.4

III. Continent-wise Analysis

Life expectancy can vary significantly across different continents due to variations in healthcare systems, socio-economic conditions, lifestyle factors, and environmental factors. By analysing life expectancy continent-wise, we can identify regions or continents with higher or lower life expectancies, which can shed light on the factors influencing these differences such as access to education, income levels, disease prevalence, and

socio-economic determinants of health. It also enables comparisons between different regions and serves as a benchmark for assessing the performance of countries within a specific continent and gain insights into best practices, areas where interventions have been effective, and help identify areas where further efforts are needed to improve life expectancy. This analysis is performed on all continents except Antarctica due to unavailability of data, with the remaining 6 continents being Asia, Africa, Oceania, Europe, North America, and South America. These are the observations for each of the analysed 6 continents over the 15 years from 2000 to 2015-

A. Africa

Africa has 54 countries of which all are still developing. The mean life expectancy of Africa is 58.6 years with minimum and maximum being 39 (Sierra Leone) and 79 years (Egypt) respectively. The median life expectancy is 57.8 years and mode of that is 55 years. The IQR value is 10.325 years, thus indicating less variance/ dispersion in life expectancy. The middle 50% of life expectancy lies between 49-58 years in the 2000s and increased to 53-65 years from 2009. Overall, the middle 50% lies between 52.8-63 years.

Life expectancy has strong negative correlation with adult mortality and HIV/AIDS, and has positive correlation with BMI, Polio, schooling, Diphtheria, and income composition of resources where correlation with income composition of resources and BMI is the strongest. The correlation between under-five deaths and infant deaths is 1, and there is strong positive correlation between thinness (5-9 years) and thinness (1-19 years), income composition of resources and schooling (all 3 of which are linear on applying linear regression), and Diphtheria and Polio having a linear relationship between them to some extent.

Algeria, Tunisia, Seychelles, Libya and Mauritius are the countries with life expectancy greater than 70 years throughout 15 years followed closely by countries Cabo Verde, Egypt, and Morocco. When compared with year-wise heatmap of other attributes, these countries had distinguishing trends of low adult mortality and high BMI compared to other countries. Life expectancy was less than 50 years in countries like Sierra Leone and Central African Republic (fig 3.1.1), but no distinguishing trends were found. From time series plot of the 15 years, we can see that life expectancy is slowly increasing over the years.

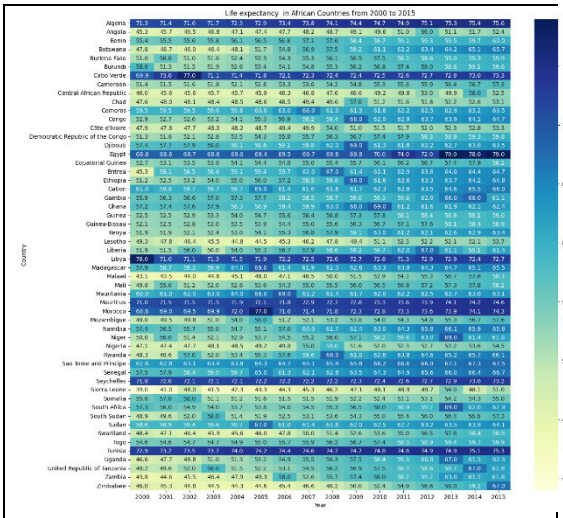

Fig 3.1.1

B. Asia

Asia has 47 countries of which 3 are developed (Cyprus, Japan, Singapore) and rest are developing countries. The mean life expectancy of Asia is 71.2 years with minimum and maximum being 54.8 (Afghanistan) and 87 years (Republic of Korea, Singapore) respectively. The median life expectancy is 72.55 years and mode of that is 74.5 years. The IQR value is 8.225 years. The middle 50% of life expectancy lies between 66.6-75 years.

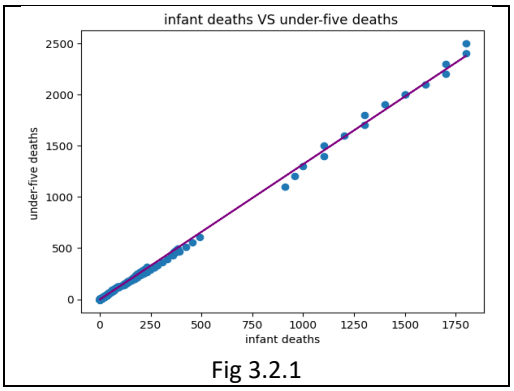

Fig 3.2.1

Life expectancy has strong negative correlation with adult mortality, status, and thinness (both categories) where correlation with adult mortality is strongest, and strong positive correlation with income composition of resources and schooling. The correlation between under-five deaths and infant deaths is 1, and there is strong positive correlation between percentage expenditure and GDP, thinness (5-9 years) and thinness (1-19 years), schooling and income composition of

resources (all 3 are linearly related on applying linear regression) (fig 3.2.1), and percentage expenditure and status (has logistic relationship with sigmoid curve on applying logistic regression). On applying logistic regression to life expectancy vs status, we infer that probability of life expectancy being greater than 75 years is high when status of country is developed.

Over the 15 years, Israel and Japan have mostly maintained greater than 80 years life expectancy, followed by Singapore, Republic of Korea, and Cyprus. They are identified by their distinct trend of low adult mortality. Countries having relatively least life expectancy are Afghanistan and Lao People's Democratic Republic in range 54-65 years, identified by their distinct trends of relatively highest adult mortality and relatively highest percentage of thinness (both categories). On comparing time series plot, it was found that life expectancy is almost constant but slowly increasing over the years for all countries.

C. Oceania

Oceania has 16 countries of which 2 are developed- Australia, New Zealand. Country with lowest life expectancy of 58.9 years is Papua New Guinea, and New Zealand has highest life expectancy of 89 years. The mean life expectancy is 71 years. The analysis results of Oceania may be heavily biased due to the low number of countries and huge missing data for many years for some countries due to not being given in the dataset.

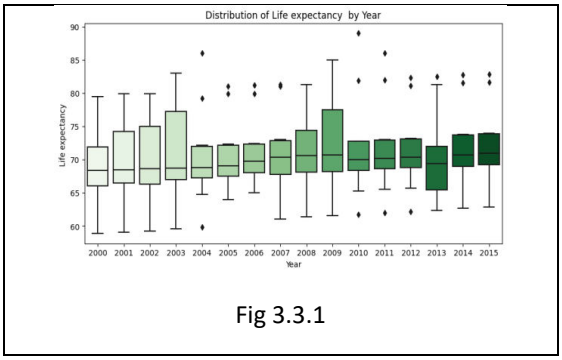

Fig 3.3.1

Median life expectancy is 69.4 years and mode is 68 years, while IQR value is 6.325. The middle 50% of life expectancy lies between 67-73 years although it changes abruptly in years like 2003, 2009, and 2013(fig 3.3.1).

Life expectancy has strong negative correlation with status, adult mortality, HIV/AIDS, under-five deaths, and infant deaths where correlation with status is strongest. Life expectancy has strong positive correlation with alcohol, schooling, income composition of resources, GDP, and percentage expenditure where correlation with alcohol and schooling is strongest.

Correlation between infant deaths and under-five deaths, and thinness (5-9 years) and thinness (1-19 years) is 1. There is strong positive correlation between HIV/AIDS and infant deaths, HIV/AIDS and under-five deaths, GDP and percentage expenditure (all of which have linear relationship), alcohol and schooling, and income composition of resources and schooling (both of which form clusters/subgroups). There is strong negative correlation between alcohol and status, schooling and status, GDP and status, and percentage expenditure and status, all of which have a sigmoid curve on plotting.

Over the 15 years, Australia and New Zealand have maintained life expectancy mostly in range 78-83 years or above, identified by their distinct trends of low adult mortality, very high alcohol consumption, high percentage expenditure, and high GDP. Whereas Papua New Guinea has relatively low life expectancy of 58-63 years, identified by it's distinct trends of very high adult mortality, very high infant deaths, very high under-five deaths, and very high HIV/AIDS.

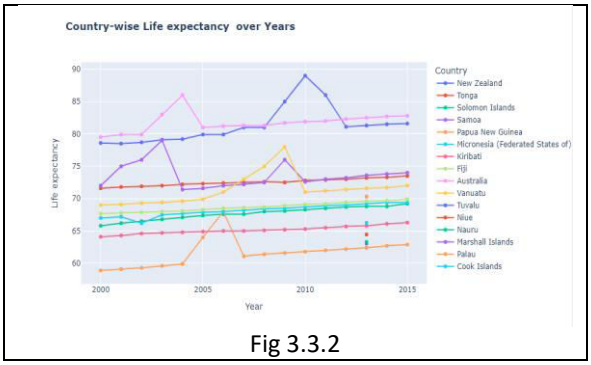

Fig 3.3.2

On comparing time series plot, life expectancy is almost constant (with peaks before 2010 for some countries) but slowly increasing over the years for all countries (fig 3.3.2).

D. Europe

Europe has 41 countries of which 26 are developed countries, thus having highest proportion of developed countries among all continents and only country to have more developed countries than developing countries (fig 3.4.1). Russian Federation has the lowest life expectancy of 64.6 years, and countries Belgium,

Finland, France, Germany, Italy, Norway, Portugal, Spain, and Sweden have highest life expectancy of 89 years. It was found that data for Monaco and San Marino were not given in dataset for all years except 2013. The mean life expectancy is 77.4 , and median, mode, and IQR value of life expectancy is 77.8 years, 78 years, and 6.9 respectively.

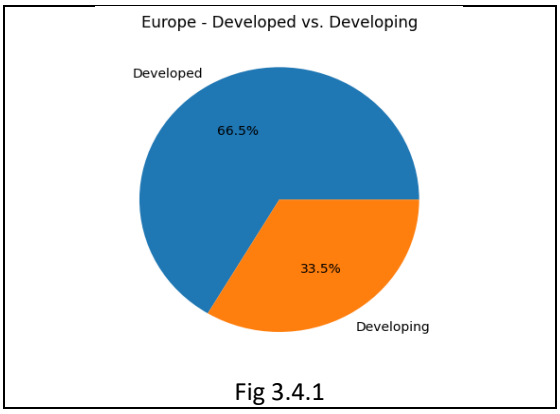

Fig 3.4.1

Life expectancy has strong positive correlation with percentage expenditure, GDP, income composition of resources, and schooling, and strong negative correlation with adult mortality and thinness (both categories) where correlation with thinness (both categories) is strongest. Correlation between infant deaths and under-five deaths, and thinness (5-9 years) and thinness (1-19 years) is 0.99, and there is strong positive correlation between GDP and percentage expenditure. Income composition of resources has strong negative correlation with thinness (both categories). All of these have linear relationship between them.

Over the 15 years, all developed countries of Europe have maintained life expectancy greater than 78 years, especially Finland, Germany, and Greece. They are identified by their trends of high percentage expenditure, high GDP, and low thinness (both categories). Russian Federation, Ukraine, Belarus, and Republic of Moldova have relatively less life expectancy in range 65-73 years, identified by their trends of high adult mortality and high thinness (both categories). On comparing time series plot, it was found that life expectancy was mostly constant and slowly increasing, but for developed countries and some countries with low life expectancy it was zig-zag (fig 3.4.2).

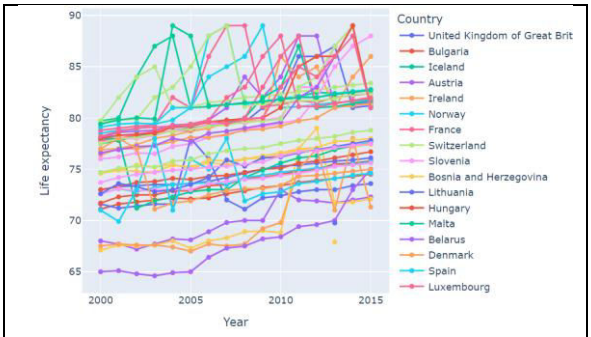

Fig 3.4.2

E. North America

It has 23 countries and the only developed country is United States of America. Canada has the highest life expectancy of 87 years and Haiti has lowest life expectancy of 36.3 years. The mean life expectancy is 73.7 years, and the median, mode, and IQR value of life expectancy is 74 years, 75 years, and 4.474 respectively. The middle 50% of life expectancy lies in range 72-76.3 years. It was also found that data for Dominica and Saint Kitts and Nevis wasn't given in dataset except for year 2013.

Life expectancy has strong positive correlation with GDP and schooling, and strong negative correlation with HIV/AIDS, thinness (both categories) and adult mortality where correlation with HIV/AIDS is strongest. Correlation between thinness (5-9 years) and thinness (1-19 years) is 1, and that between infant deaths and under-five deaths is 0.99 (both have linear relationship). Strong positive correlation exists between GDP and percentage expenditure, having a linear relationship to some extent. Strong negative correlation exists between total expenditure and status, indicating that when total expenditure is high, probability of country being developed is high.

Over the 15 years, it was found that all countries except Haiti had high life expectancy of mostly in range 70-75 years, especially Canada, Costa Rica, Cuba, and USA which had life expectancy greater than 76 years. Haiti has trend of high HIV/AIDS. No distinguishing trends were found for countries with high life expectancy. On comparing time series, for life expectancy the plot lines were almost constant except for Haiti.

F. South America

It has 12 countries were all are developing, where lowest and highest life expectancy is 62.6 years (Bolivia) and 85 years (Chile) respectively. The mean, median, and mode of life expectancy is 73, 73.65, and 73.6 years respectively. The IQR value 3.825, the lowest value among all continents. The middle 50% of life expectancy lies in range 71.5-75.3 years.

Life expectancy has strong positive correlation with income composition of resources and schooling, and strong negative correlation with thinness (both categories), HIV/AIDS, and adult mortality where correlation with thinness is strongest. Correlation between infant deaths and under-five deaths, and thinness (5-9 years) and thinness (1-19 years) is 1 (linear relationship). Strong positive correlation exists between GDP and percentage expenditure, and HIV/AIDS and thinness (both categories). Strong negative correlation exists between schooling and thinness (both categories).

Over the 15 years, Chile has maintained highest life expectancy in range 77-85 years. Bolivia, Guyana, and Suriname have relatively less life expectancy in range 62-71 years, having distinct trends of high adult mortality and high HIV/AIDS. On comparing time series plot of life expectancy, we found it is slowly increasing.

IV. Hypothesis Testing

Hypothesis testing was done using random sampling and t-test on each continent as well as using whole dataset (global). Some questions answered using hypothesis testing are-

- Should a country having a lower life expectancy value (<65) increase its healthcare expenditure in order to improve its average lifespan? -Yes
- How does infant death and adult mortality rates affect life expectancy? -As adult mortality and infant death rates increase, life expectancy decreases.
- How does alcohol consumption impact life expectancy? -For those with moderate alcohol consumption the life expectancy is higher than for those who drink occasionally or heavily.
- What is the impact of schooling and income composition of resources on the lifespan of humans? When years of schooling is and income composition of resources is more, life expectancy is more.
- Do densely populated countries tend to have lower life expectancy? -No, population density has no significant impact on life expectancy.
- Does life expectancy vary significantly between continents? -Yes (tested using one-way ANOVA)

V. Summary

Life expectancy is slowly increasing all over the world over the years, with Europe having the highest average life expectancy and Africa having the lowest average life expectancy (fig 5.1). Globally, the maximum life expectancy is 89 years with an average of 69.2 years. Europe is the continent with highest number of developed countries, whereas Africa and South America have no developed countries, and developed countries have more life expectancy than developing countries. In all continents, correlation between infant deaths and under-five deaths is 1 or 0.99, and a strong positive correlation exists between thinness (5-9 years) and thinness (1-19 years), and schooling and income composition of resources. In many continents, strong positive correlation exists between GDP and percentage expenditure also.

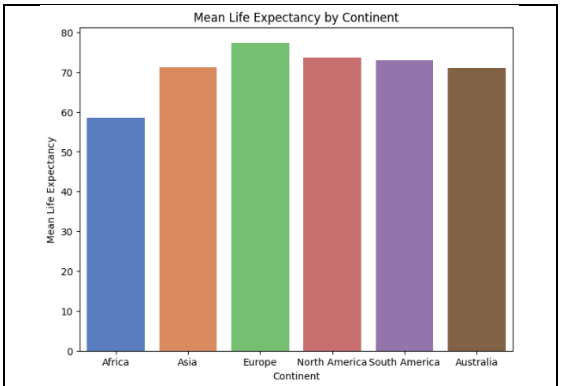

Fig 5.1

Life expectancy varies significantly continent-wise, with each continent affected by different factors-

- Africa- adult mortality, HIV/AIDS, BMI and income composition of resources
- Asia- adult mortality, income composition of resources, and schooling
- Oceania- status, alcohol, and schooling
- Europe- thinness (both categories), percentage expenditure, GDP, income composition of resources, and schooling
- North America- GDP, HIV/AIDS, and schooling

- South America- thinness (both categories), income composition of resources, and schooling

VI. Challenges and Interesting Findings

- Increase in alcohol consumption increases life expectancy, which is conflicting as alcohol consumption increases chances of stroke and other health issues, which should mean life expectancy is supposed to decrease.
- Due to huge number of outliers in most of the attributes, it is possible that the analysis is biased.
- It's possible that sample consisting of data for each country from a specific year can be taken as a representative of population since correlation of year with other attributes is close to 0.
- Africa is the only continent where countries with less than 50 years of life expectancy exist.

VII. Conclusion

It was found that factors contributing to low life expectancy are developing status, high rates of adult mortality, HIV/AIDS, and thinness (both categories), less years of schooling, low Polio and Diphtheria immunization rates, low income composition of resources, BMI, and GDP. Their role as predictors was confirmed by hypothesis testing as they had significant impact on life expectancy. Using these factors as predictors, a random forest regressor was made to predict life expectancy by training on this dataset. The accuracy was calculated using R-square score which was found to be approximately 0.985. The countries whose life expectancies were predicted to be less than 65 years are mostly from Africa, followed by Asia (fig 7.1).

```
Total countries with life expectancy less than 65 years:  71
Africa :  46
Asia :  16
Oceania :  5
Europe :  1
North America :  2
South America :  1
```

Fig 7.1

So, a total of 71 countries need to improve their life expectancy, especially African countries since it is the only continent where countries with less than 50 years of life expectancy exist. So in African countries, BMI and income composition of resources has to be increased and it should reduce it's adult mortality and HIV/AIDS rates to improve life expectancy. Some low life expectancy African countries are Sierra Leone, Zimbabwe, Zambia, and Central African Republic. Similarly, for Asia we have to reduce adult mortality rate, and increase income composition of resources and number of years of schooling to improve life expectancy is low life expectancy countries like Lao People's Democratic Republic, Afghanistan, Cambodia, and Timor-Leste.