

Analysis of Factors influencing Life Expectancy

Badam Jwala Sri Hari

Intern-CHUBB

Abstract:

Life expectancy at birth is defined as how long, on average, a new-born can expect to live, if current death rates do not change. However, the actual age-specific death rate of any particular birth cohort cannot be known in advance. In the past two decades Life Expectancy is increased by 6 years which was influenced by a range of factors like Immunization factors, Mortality factors, Economic factors, social factors. Analysis on factors which are influencing Life span of people is essential because from past data it will be clear which factors influencing in increases life span and which factors are influencing reduce in Life Expectancy. In this report Life expectancy of world is analysed from the year of 2000-2015.

Dataset Source:

Life Expectancy WHO is taken from [Kaggle](#) and helps in predicting life expectancy with the help of numerous factors for a period of 15 years. The Global Health Observatory (GHO) data repository under 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 data-set related to life expectancy, health factors for 193 countries have been collected from the same WHO data repository website and its corresponding economic data was collected from the United Nation website. 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 the effect of immunization and the human development index were not considered in the past. This dataset is to be used for Machine Learning and Data Visualization purposes.

About the data: Each row in the data consists of information about the Country, Continent, Year, Status, Life expectancy, Adult Mortality, infant deaths, Alcohol, GDP in addition to other input parameters.

Immunization factors:

- Hepatitis B: Immunization coverage against hepatitis B (HepB) among one year old children (In percentage).
- Polio: Polio immunization coverage (Pol3) among one-year-old children (%).
- Diphtheria: Immunization coverage against diphtheria and pertussis tetanus (DTP3) among children aged 1 year.

Mortality factors

- Infant deaths: Infant deaths per 1000 population.
- Under-five deaths: Deaths of children under five years of age per 1000 population.
- HIV / AIDS: Mortality per 1,000 live births HIV / AIDS (0-4 years).
- Adult Mortality: Adult mortality rates for both sexes (probability of dying between the ages of 15 and 60 per 1000 population).

Economic factors:

- percentage expenditure: Health care expenditure as a percentage of gross domestic product per capita (%).
- Total expenditure: Total government spending on health as a percentage of total government spending (%).
- GDP: Gross Domestic Product per capita (in US dollars).
- Income composition of resources: Human Development Index in terms of income structure of resources (index from 0 to 1).

Social factors:

- BMI: Average body mass index of the entire population.
- Thinness 1-19 years: Prevalence of thinness among children and adolescents aged 10 to 19 years (%).
- Thinness 5-9 years: Prevalence of thinness among children aged 5 to 9 (%).
- Alcohol: Accounting for alcohol consumption per capita (15+) (in liters of pure alcohol).
- Population: Population of the country.
- Schooling: Number of years of study (years).

Approach:

- The various python libraries such as Numpy, pandas, Matplotlib are used for the purpose of mathematical calculations, extraction of data and visualization, respectively.
- Some other packages are used like sklearn which helps to find relationship between factors.
- The extracted dataset was found to have null values which are then effectively handled to achieve accurate analysis rather than removing them from the dataset.

➤ Data Pre-processing

1. Handling the Missing value:

Data sets consists of 2938 rows and 22 columns with 193 countries data with the missing value which should be handled.

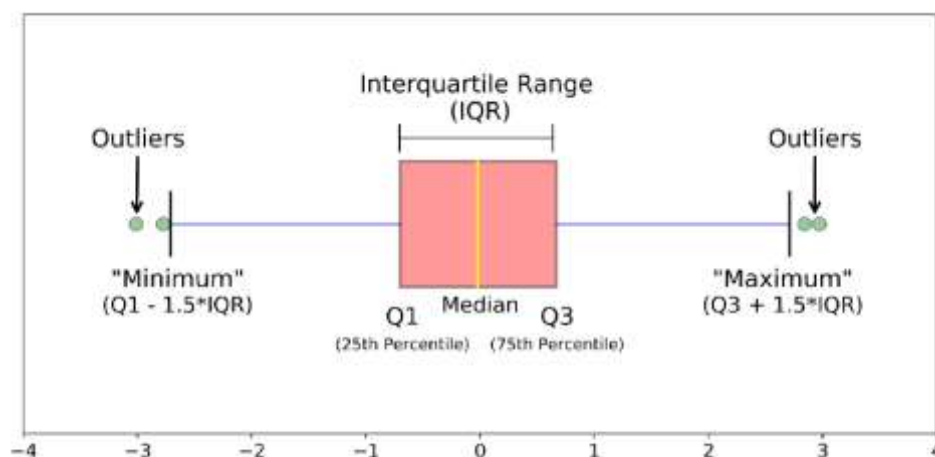
	Features	Missing_Values	Missing_Percent
3	Life expectancy	10	0.340
4	Adult Mortality	10	0.340
6	Alcohol	194	6.600
8	Hepatitis B	553	18.820
10	BMI	34	1.160
12	Polio	19	0.650
13	Total expenditure	226	7.690
14	Diphtheria	19	0.650
16	GDP	448	15.250
17	Population	652	22.190
18	thinness 1-19 years	34	1.160
19	thinness 5-9 years	34	1.160
20	Income composition of resources	167	5.680
21	Schooling	163	5.550

Missing value can be replaced with Mean, Mode, Median based on the Column. Here in all features missing values are replaced with the mean of that column.

	Features	Missing_Values	Missing_Percent
0	Country	0	0.000
1	Year	0	0.000
2	Status	0	0.000
3	Life_expectancy	0	0.000
4	Adult_Mortality	0	0.000
5	infant_deaths	0	0.000
6	Alcohol	0	0.000
7	percentage_expenditure	0	0.000
8	Hepatitis_B	0	0.000
9	Measles	0	0.000
10	BMI	0	0.000
11	under_five_deaths	0	0.000
12	Polio	0	0.000
13	Total_expenditure	0	0.000
14	Diphtheria	0	0.000
15	HIV/AIDS	0	0.000
16	GDP	0	0.000
17	Population	0	0.000
18	thinness	0	0.000
19	thinness_5-9_years	0	0.000
20	Income_composition_of_resources	0	0.000
21	Schooling	0	0.000

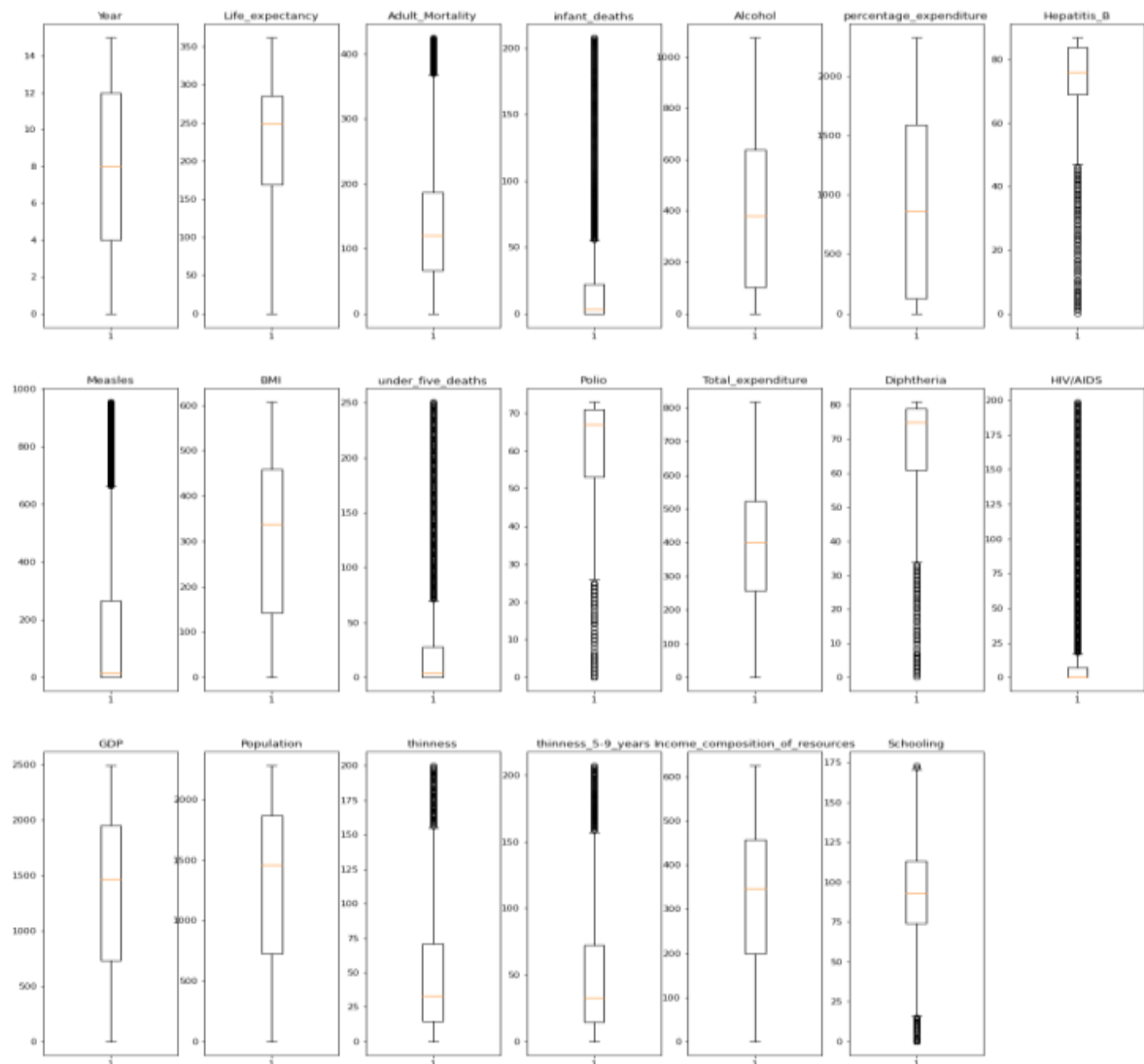
2. Detection and removal of Outliers:

Outliers are unusual values in your dataset, and they can distort statistical analyses and violate their assumptions. Outliers increase the variability in your data, which decreases statistical power. Consequently, excluding outliers can cause your results to become statistically significant. Outliers can be identified by using Boxplots



- **Median (Q2/50th Percentile):** the middle value of the dataset.
- **First quartile (Q1/25th Percentile):** the middle number between the smallest number (not the “minimum”) and the median of the dataset.
- **Third quartile (Q3/75th Percentile):** the middle value between the median and the highest value (not the “maximum”) of the dataset.
- **Interquartile range (IQR):** 25th to the 75th percentile. Difference between First Quartile and Third Quartile.
- **Maximum:** $Q3 + 1.5 \cdot IQR$
- **Minimum:** $Q1 - 1.5 \cdot IQR$
- The Blue lines which connect minimum and maximum with IQR are called Whiskers.
- Outliers are shown with green circles which are greater than maximum and minimum value which were found with IQR.

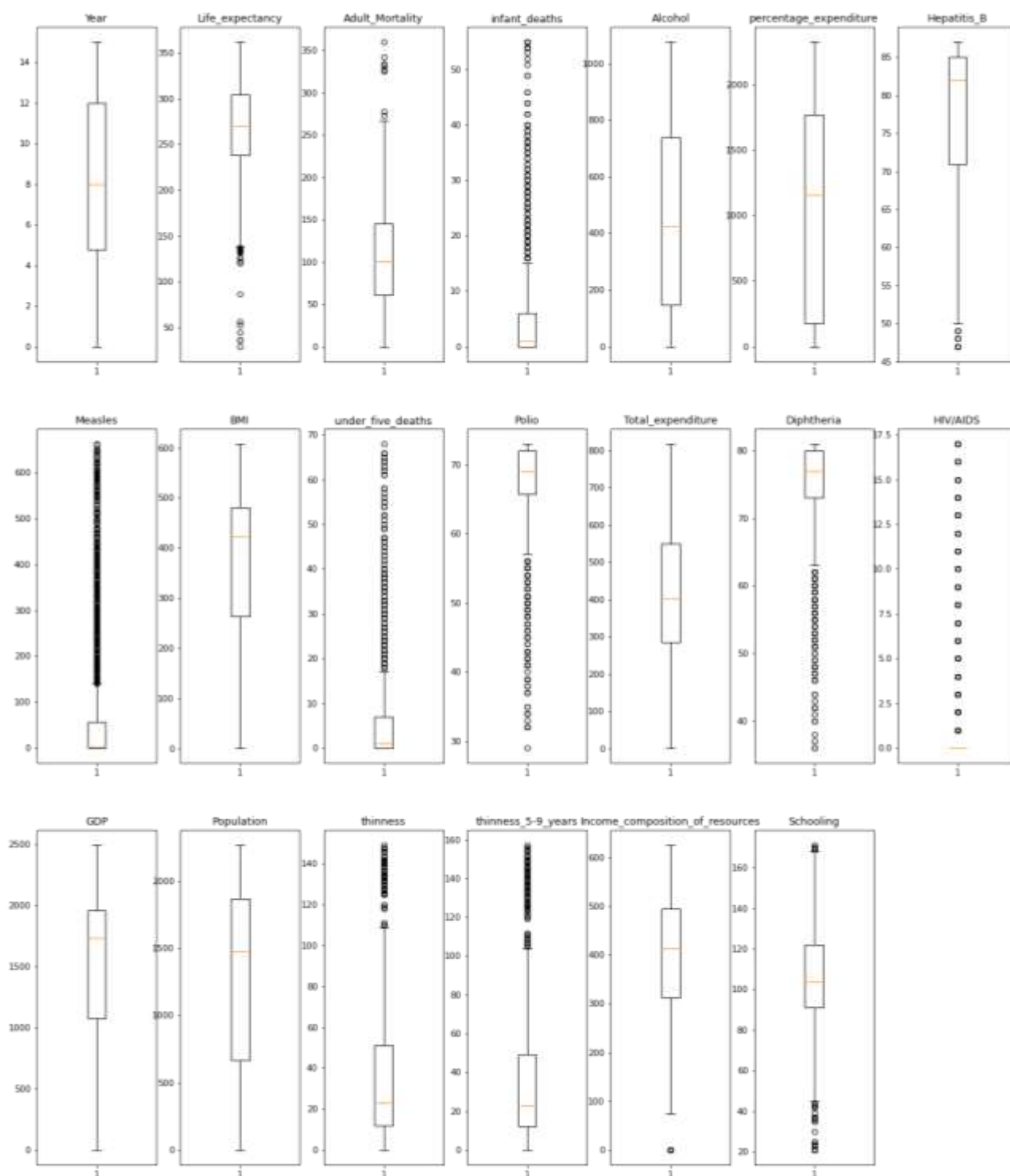
Boxplot of given numeric data is as shown below:



Boxplot of numeric data before removing outliers

From the above boxplot, it is clear there are outliers in adult mortality, Infant deaths, Hepatitis B, Measles, Under-five deaths, Polio, Diphtheria, HIV, thinness, Schooling.

By using IQR method Outliers are removed. Finding First and third Quartile in every numeric column of data and minimum, maximum values are calculated with quartile and IQR values. The values which are less than minimum and values which are greater than maximum are removed.



Boxplot of numeric data after removing outlier

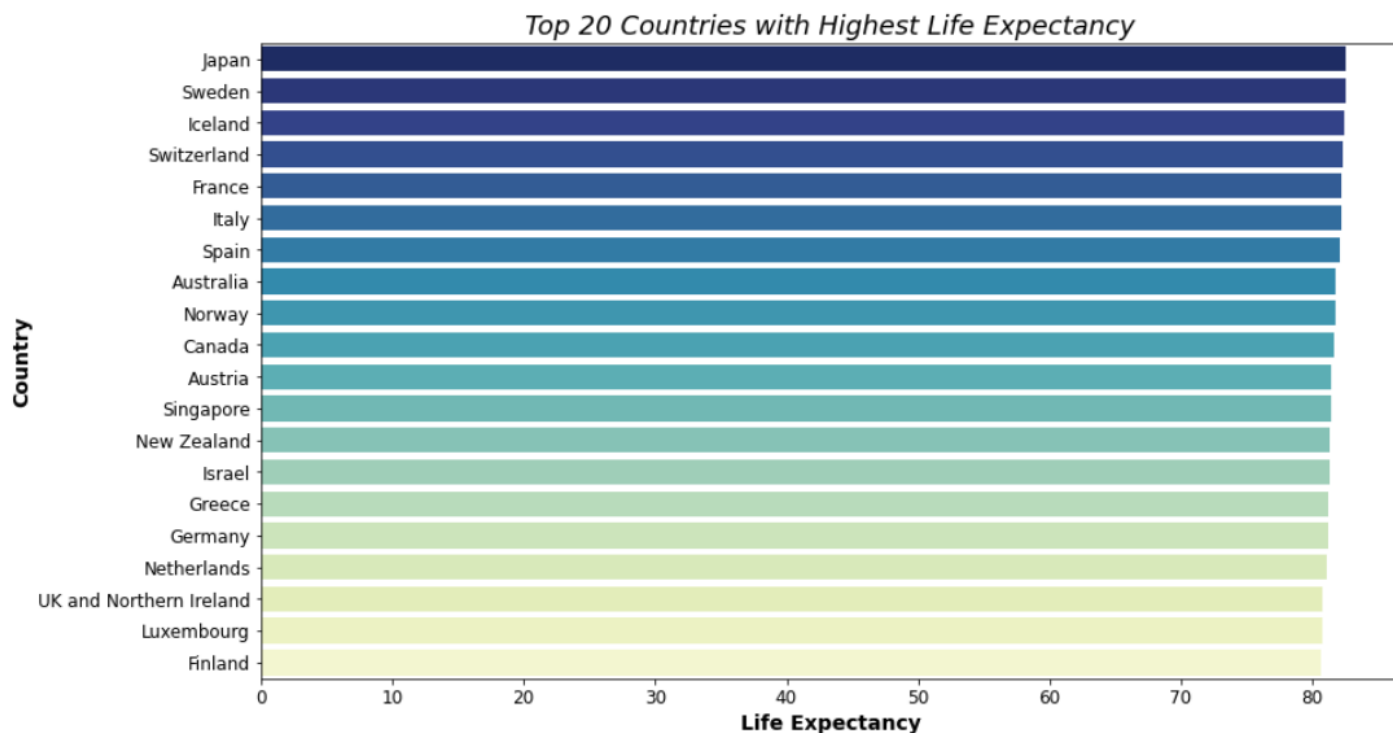
From the above graph is it evident that the density of outliers is reduced after applying the IQR method. In some cases, completely removing is not possible because the maximum of data points is lying below the minimum and above the maximum.

Analysis and Visualization:

1. Top twenty and last twenty countries with Life Expectancy:

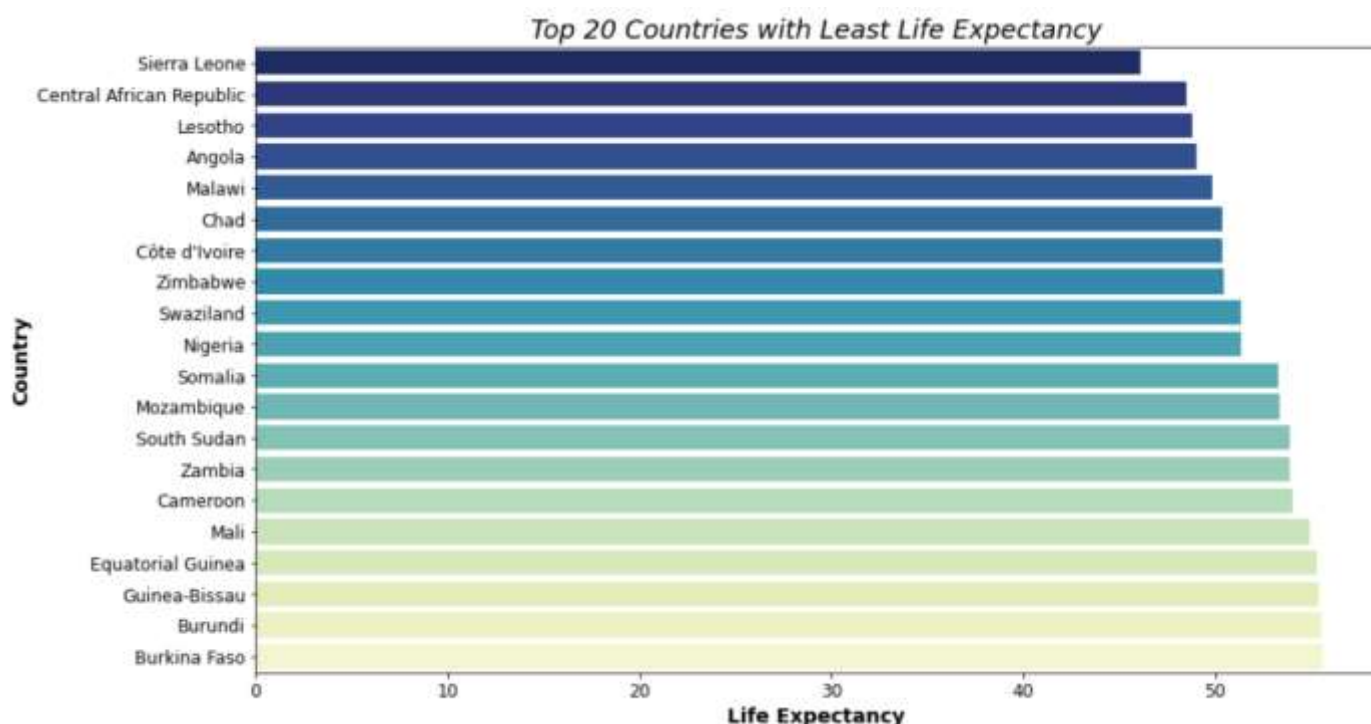
Purpose: To know which countries are in top and last in Average Life expectancy in span of 15 years.

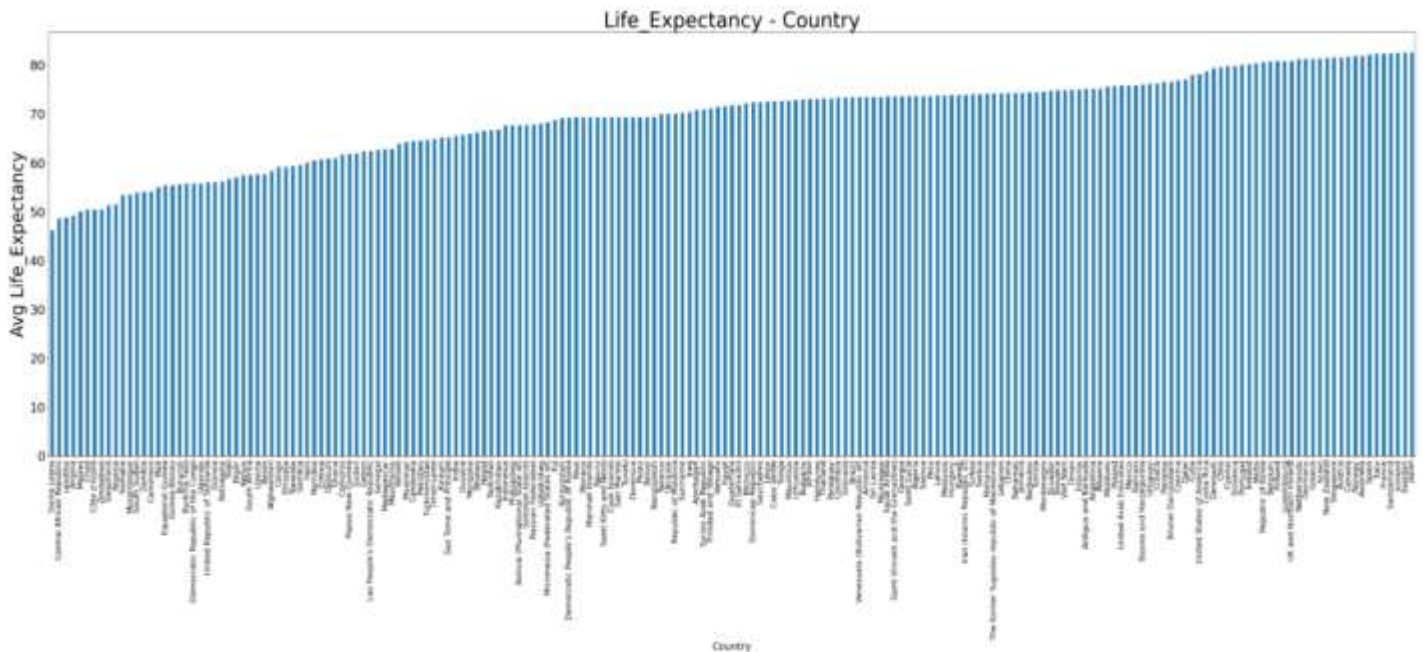
Top 20 Countries with Most Life Expectancy		Top 20 Countries with Least Life Expectancy	
Country		Country	
Japan	82.537	Sierra Leone	46.112
Sweden	82.519	Central African Republic	48.513
Iceland	82.444	Lesotho	48.781
Switzerland	82.331	Angola	49.019
France	82.219	Malawi	49.894
Italy	82.188	Chad	50.388
Spain	82.069	Côte d'Ivoire	50.388
Australia	81.813	Zimbabwe	50.487
Norway	81.794	Swaziland	51.325
Canada	81.688	Nigeria	51.356
Austria	81.481	Somalia	53.319
Singapore	81.475	Mozambique	53.394
New Zealand	81.337	South Sudan	53.875
Israel	81.300	Zambia	53.906
Greece	81.219	Cameroon	54.019
Germany	81.175	Mali	54.938
Netherlands	81.131	Equatorial Guinea	55.312
UK and Northern Ireland	80.794	Guinea-Bissau	55.369
Luxembourg	80.781	Burundi	55.537
Finland	80.713	Burkina Faso	55.644
Name: Life_expectancy, dtype: float64		Name: Life_expectancy, dtype: float64	



Inference:

From the above graph, it is clear that Japan is in lead in Life expectancy because of food habits and Expenditure on health but all the twenty countries have an average Life Expectancy above 80 years.





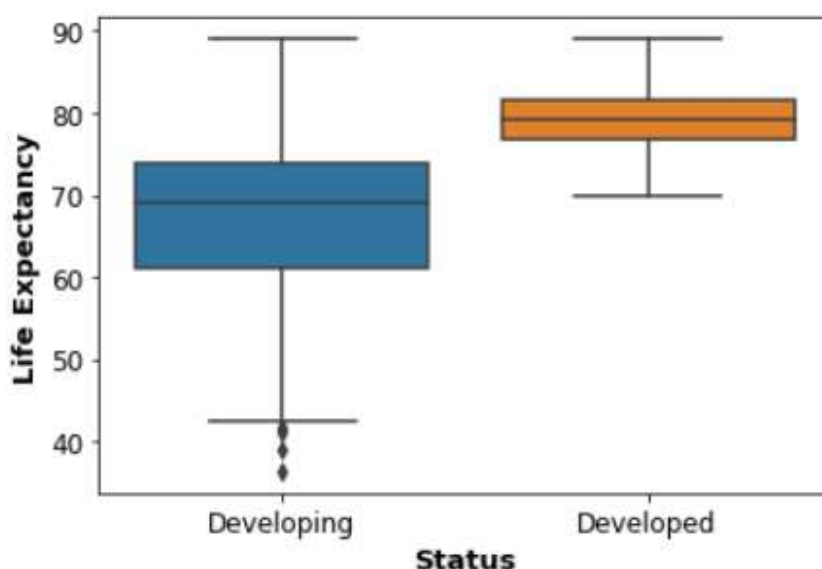
Inference:

It is evident that from the above graph Sierra Leone has less Life expectancy because the hospital faculties are not at all good only 355 trained nurses over the population of 5.5 million. Sierra Leone is the country which is suffering from running water infants are dying because of proper hygiene.

2. Analysis between Developed and Developing Countries:

Purpose: To know Life expectancy in Developed and Developing countries.

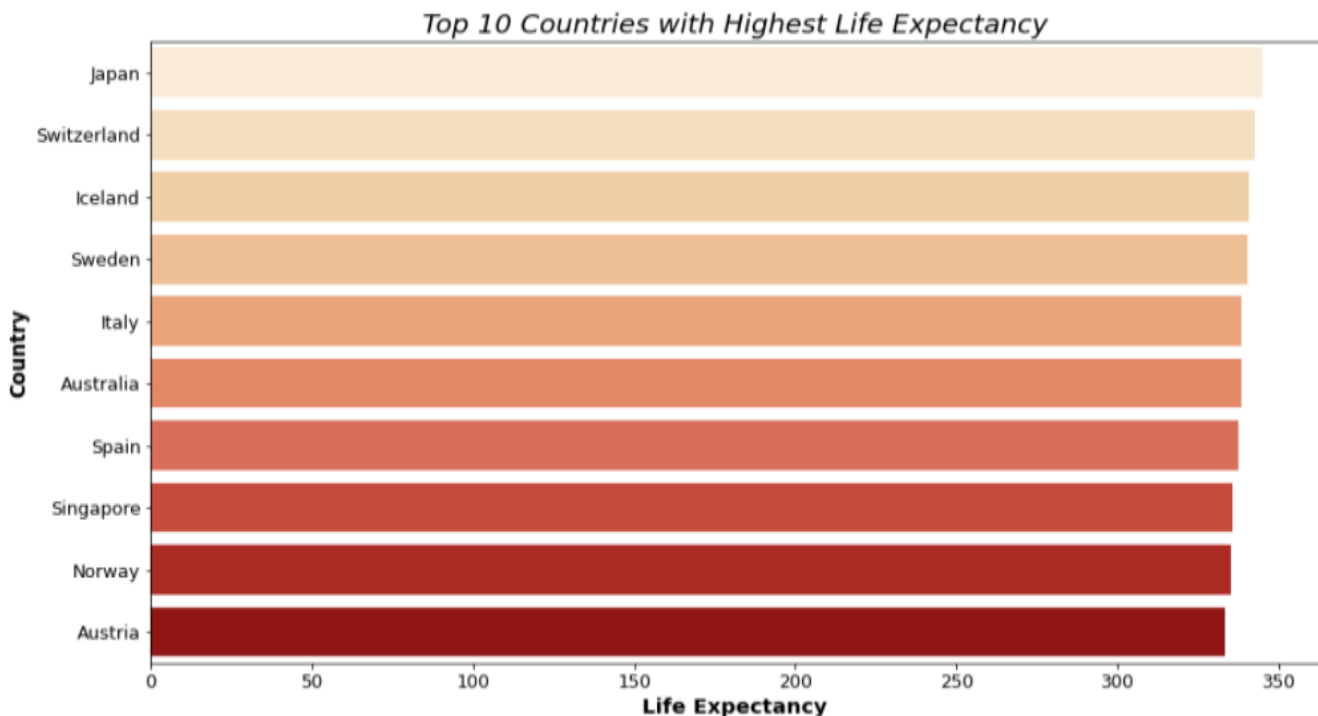
There are 32 Developed countries and 161 Developing countries the Life Expectancy in developed countries is 79 and developing countries is 67 because developed countries are able to invest more on health care.



Above Graph shows in both developing and developed countries maximum value is 89 and minimum value in developing countries is 38, In developed countries it is 70 where this is Maximum, Minimum values are calculated by using Inter quartile range.

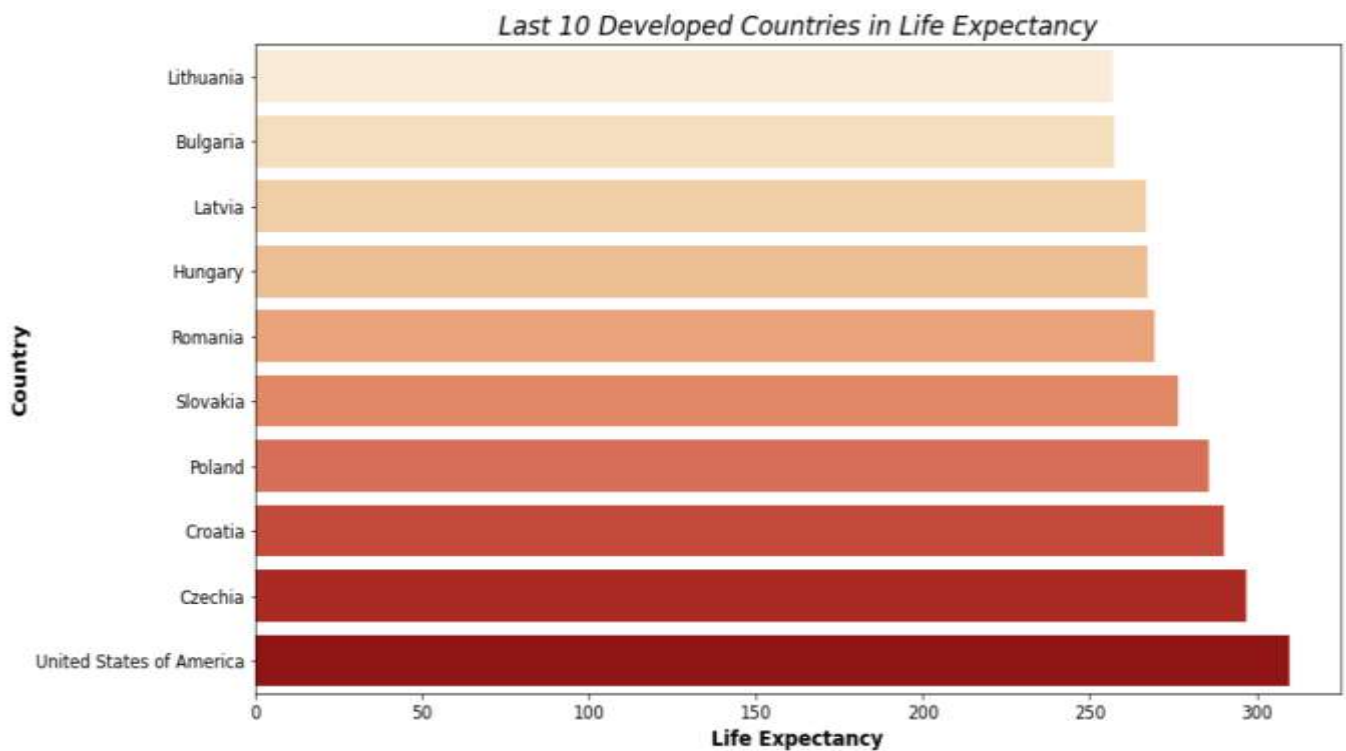
➤ **Top 10 and Last 10 in Developed countries on bases with Life Expectancy:**

Top 10 Developed Countries Life Expectancy		Last 10 Developed Countries in Life Expectancy	
Country		Country	
Japan	82.537	Lithuania	72.806
Sweden	82.519	Bulgaria	72.850
Iceland	82.444	Latvia	73.731
Switzerland	82.331	Hungary	73.825
Italy	82.188	Romania	74.050
Spain	82.069	Slovakia	74.750
Australia	81.813	Poland	75.650
Norway	81.794	Croatia	76.119
Austria	81.481	Czechia	76.769
Singapore	81.475	United States of America	78.063
Name: Life_expectancy, dtype: float64		Name: Life_expectancy, dtype: float64	



Inference:

From the above graph, it is clear that Japan is in lead in Life expectancy in developed countries as Japan because the obesity rate in Japan is low, the food diet is unique all over the country and consumption of seafood is more.

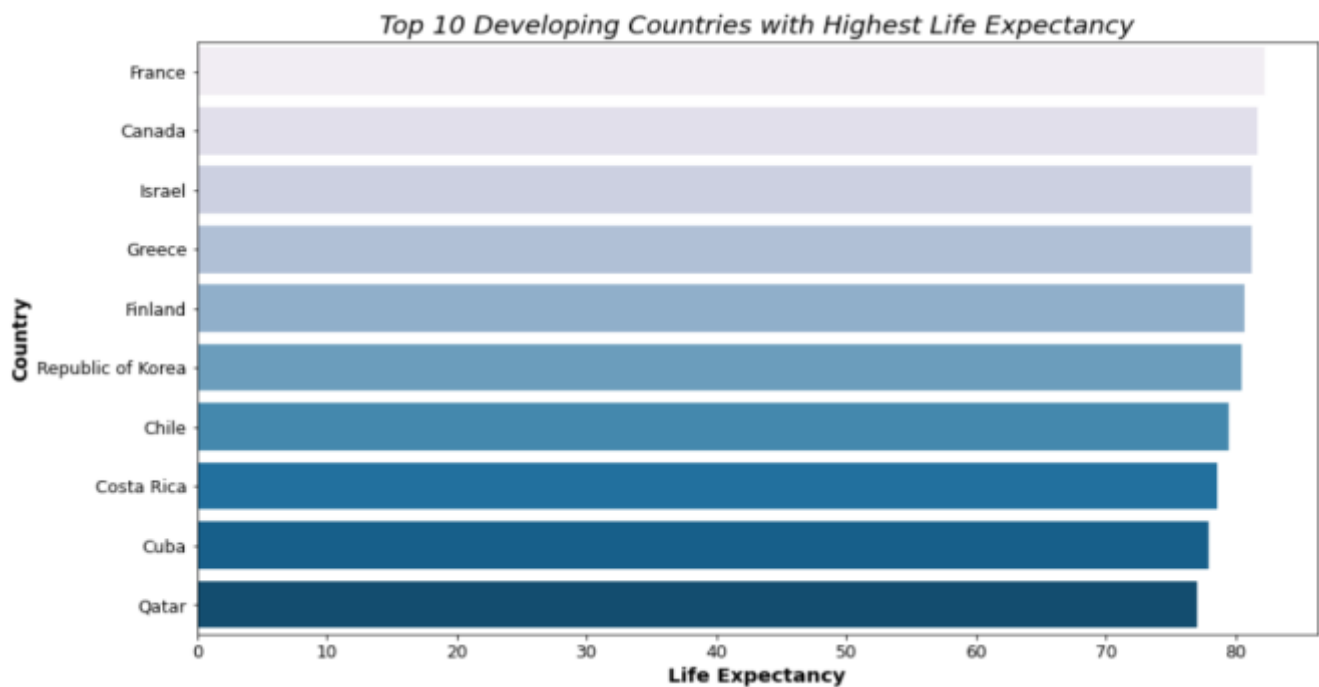


Inference:

In Developed countries, Lithuania is a country which stood last in life expectancy of 72 years. The richest United states of America with Life expectancy of 78 years. From graph it is also clear that in developed countries least life expectancy is 72 years.

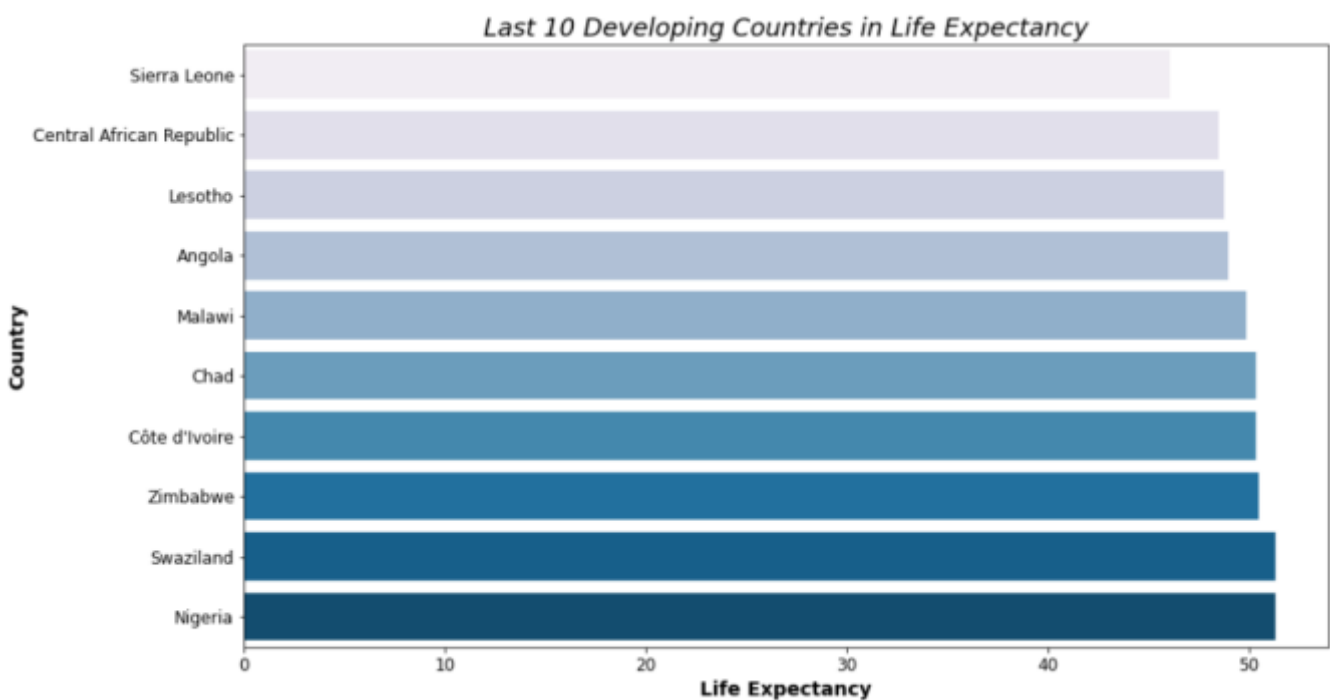
➤ **Top 10 and Last 10 in Developing countries on bases with Life Expectancy:**

Top 10 Developing Countries Life Expectancy		Last 10 Developing Countries in Life Expectancy	
Country	Life Expectancy	Country	Life Expectancy
France	82.219	Sierra Leone	46.112
Canada	81.688	Central African Republic	48.513
Israel	81.300	Lesotho	48.781
Greece	81.219	Angola	49.019
Finland	80.713	Malawi	49.894
Republic of Korea	80.487	Chad	50.388
Chile	79.450	Côte d'Ivoire	50.388
Costa Rica	78.594	Zimbabwe	50.487
Cuba	77.975	Swaziland	51.325
Qatar	77.031	Nigeria	51.356
Name: Life_expectancy, dtype: float64		Name: Life_expectancy, dtype: float64	



Inference:

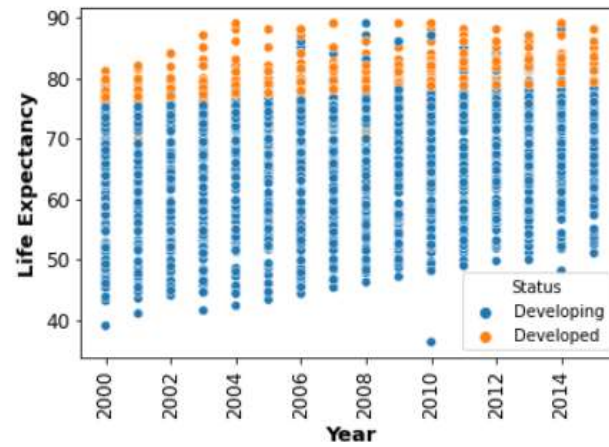
Graph speaks that in Developing countries France is a country which stood top in life expectancy of 82 years and Qatar is in 10th place in developing countries out of 161 countries with Life expectancy of 77 years.



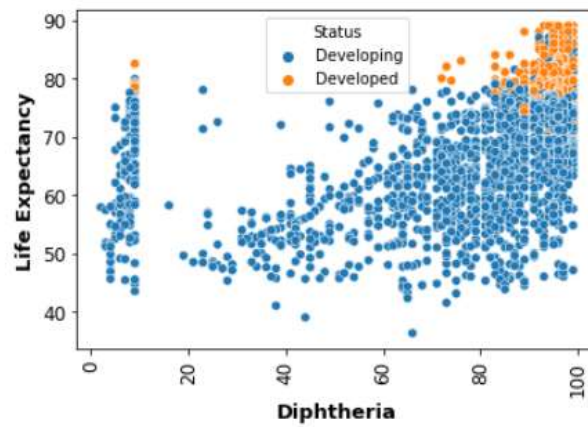
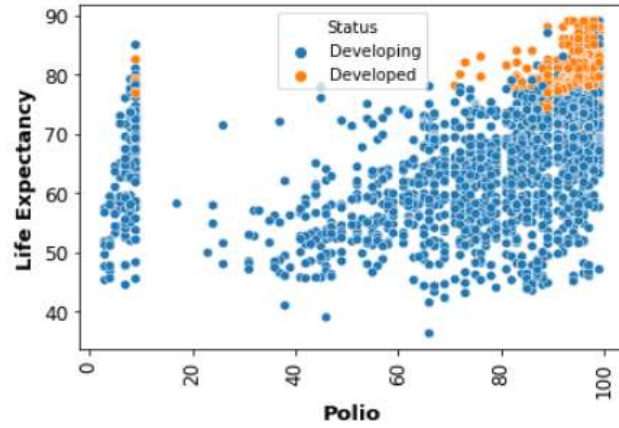
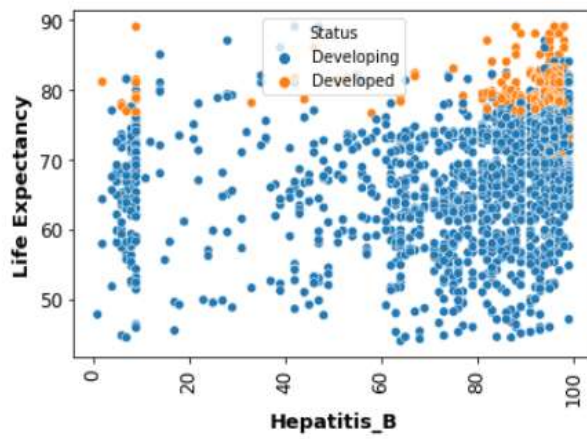
Inference:

The above Graph says that in Developing countries Sierra Leone is a country which stood last in life expectancy of 46 years and Nigeria is in 10th place from last in developing countries out of 161 countries with Life expectancy of 51 years.

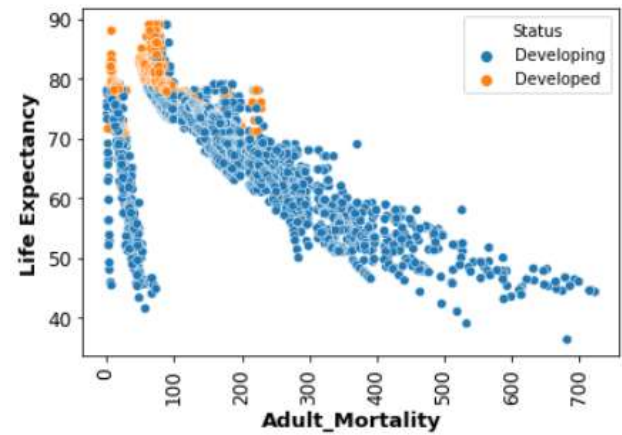
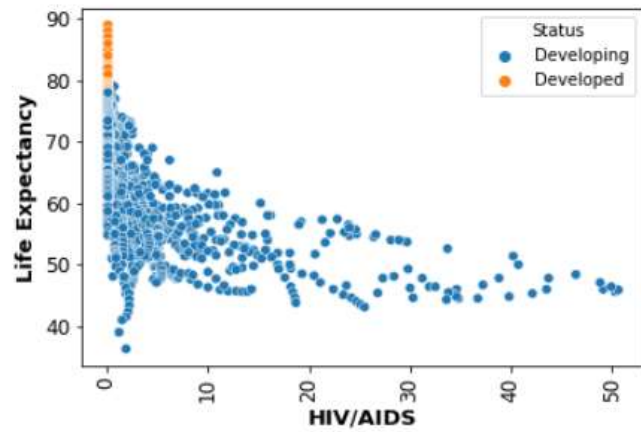
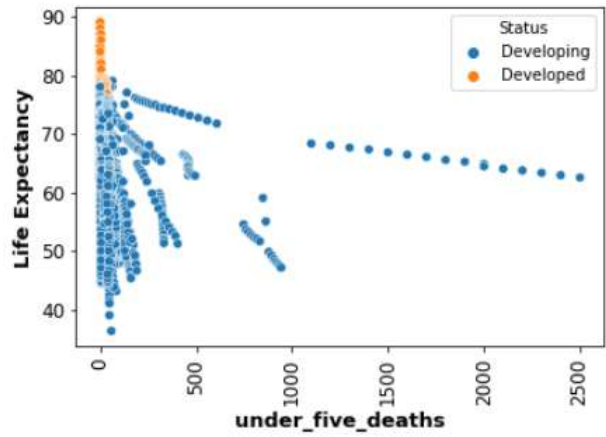
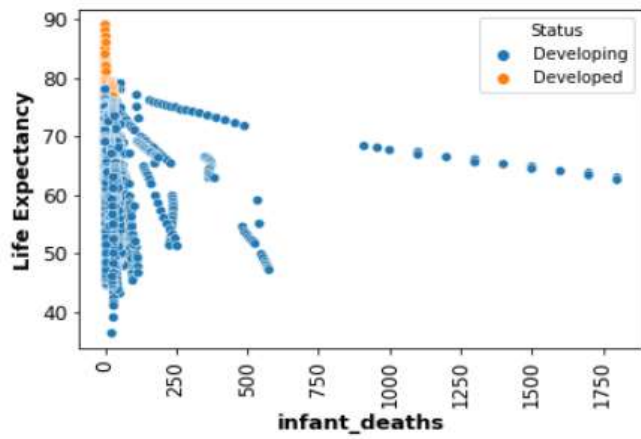
➤ Life Expectancy Vs features



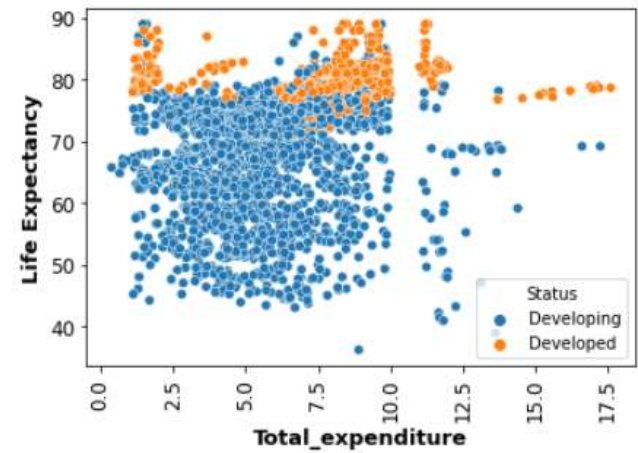
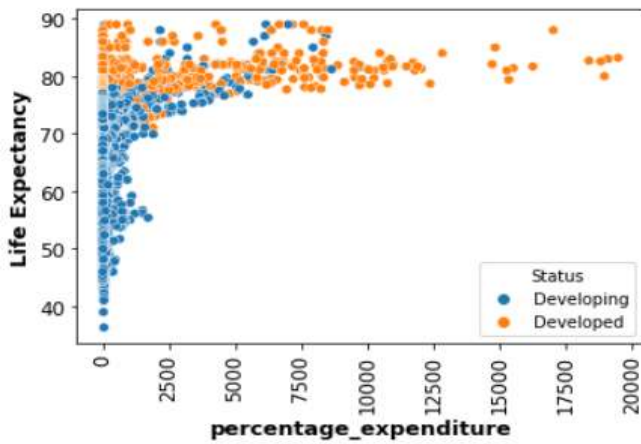
Immunization factors:

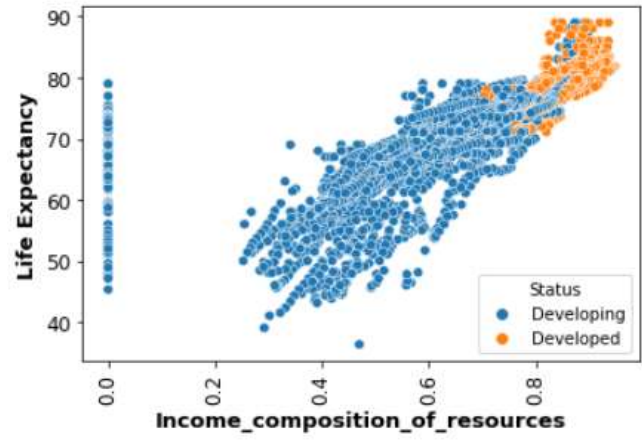
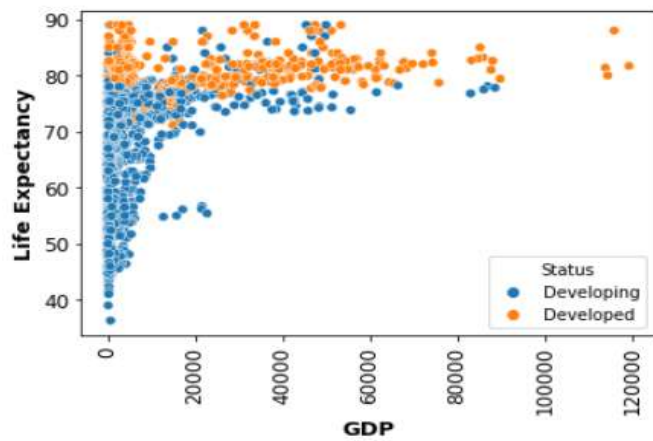


Mortality factors

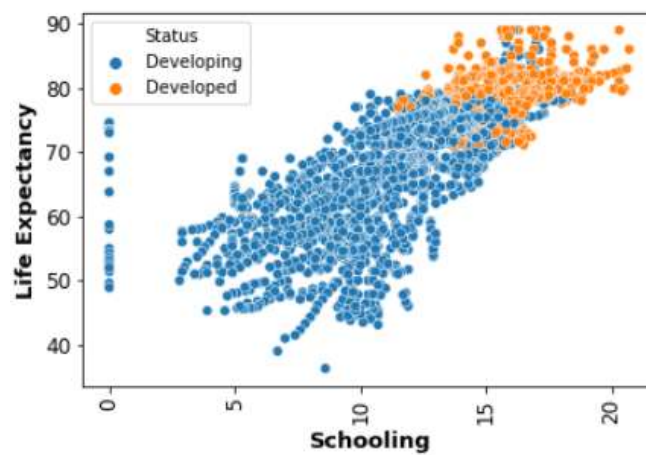
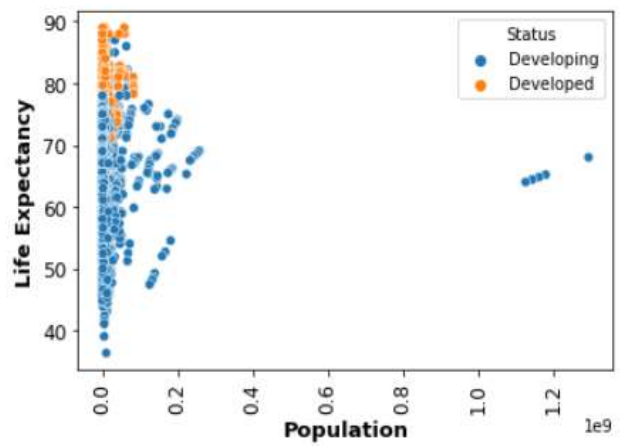
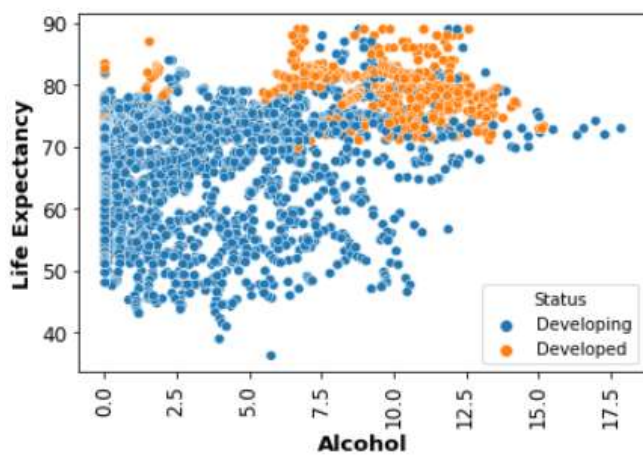
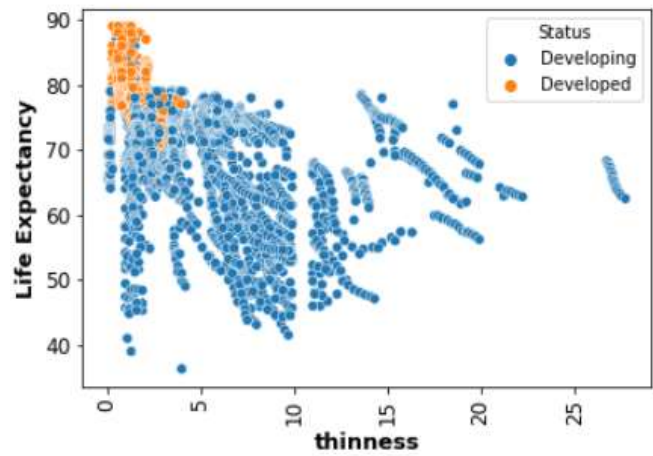
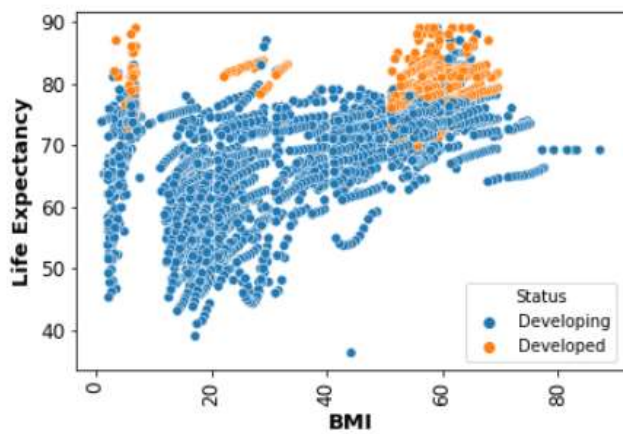


Economic factors:

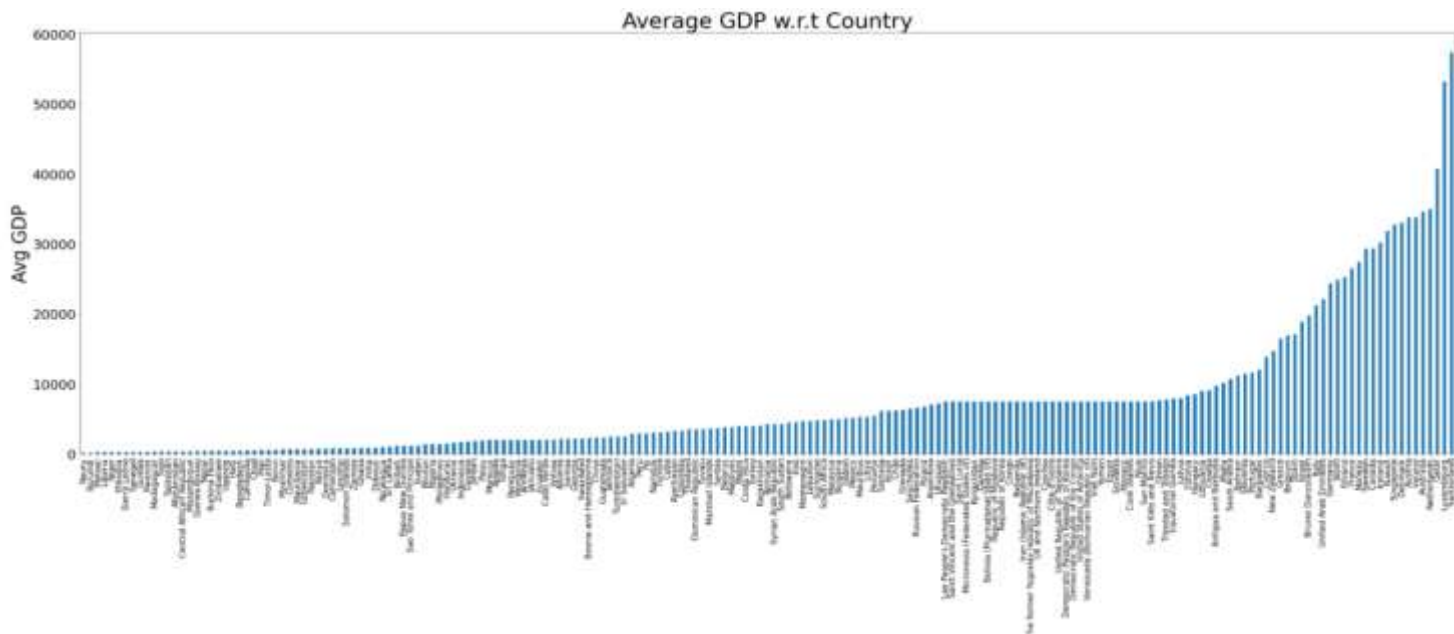




Social factors:



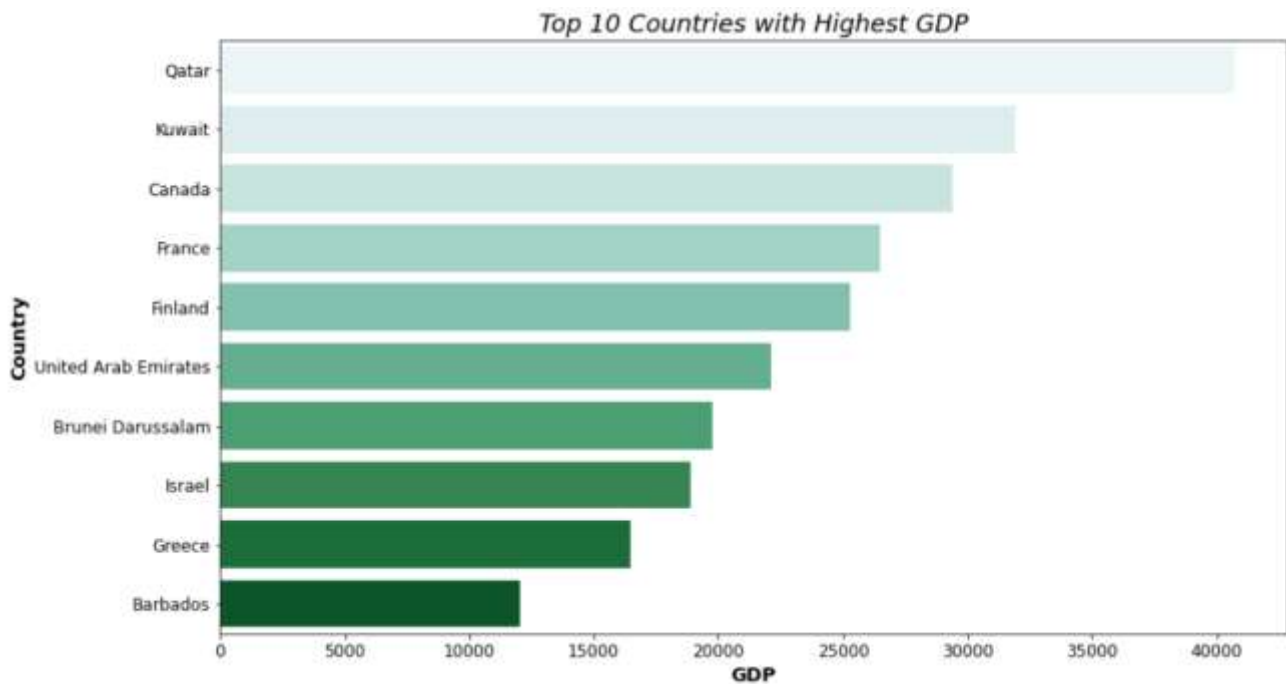
3. GDP Vs Life Expectancy:



Top 10 and Last 10 countries based on GDP:

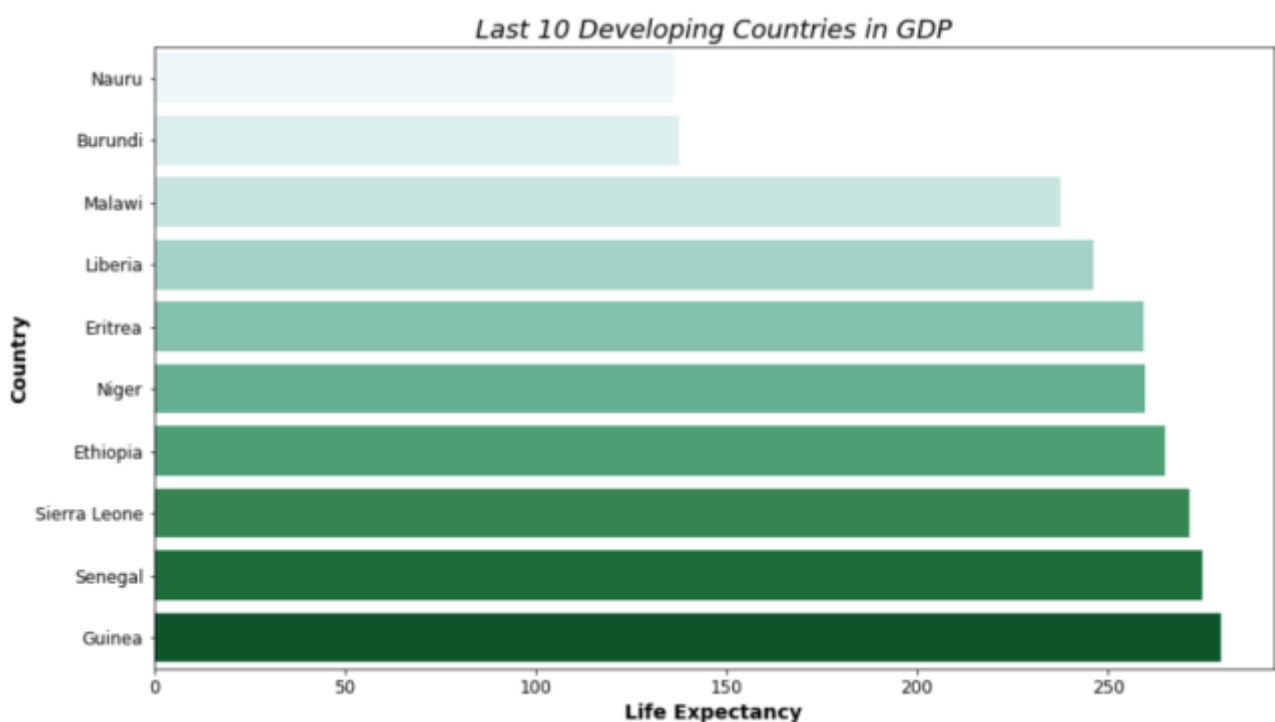
```
=====
Top 10 Countries with Highest GDP
=====
Country
Qatar                40748.444
Kuwait               31914.378
Canada               29382.908
France               26465.551
Finland              25268.650
United Arab Emirates 22110.367
Brunei Darussalam    19744.808
Israel               18860.476
Greece               16454.236
Barbados             12017.099
Name: GDP, dtype: float64
```

```
=====
Last 10 Countries in GDP
=====
Country
Nauru                136.183
Burundi              137.815
Malawi               237.504
Liberia              246.282
Eritrea              259.395
Niger                259.782
Ethiopia             264.971
Sierra Leone        271.506
Senegal              274.611
Guinea               279.465
Name: GDP, dtype: float64
```

Inference:

From above graph it is clear that Top GDP countries are Qatar and Kuwait because these both countries mainly depend on Crude oil and gas production. 10th highest GDP country is Barbados with GDP of 12017 US dollar.

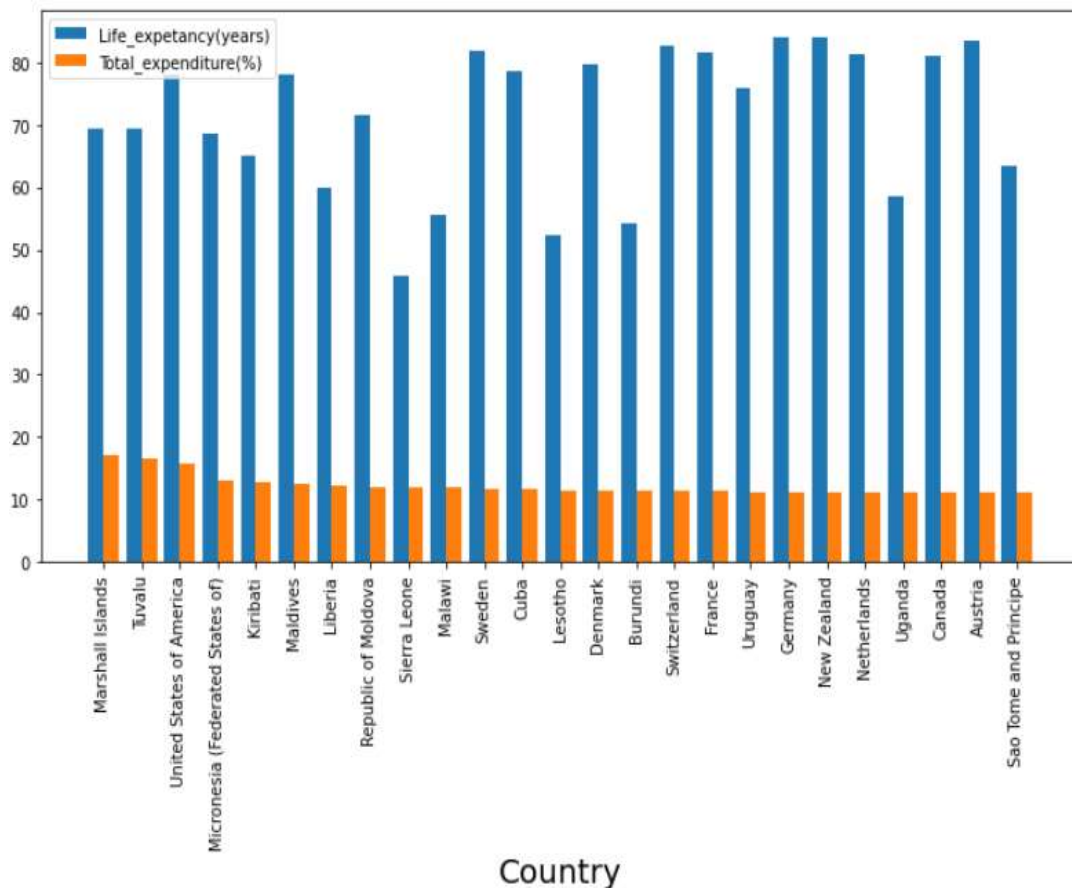


Inference:

Graph speaks that in Nauru is a country which stood last in GDP with 136 US dollar. In last ten countries the Average GDP is 237 US dollars. Most are countries in last 10 are from either Africa or an island.

4. Countries which are invest more than 10% of their income in Health care:

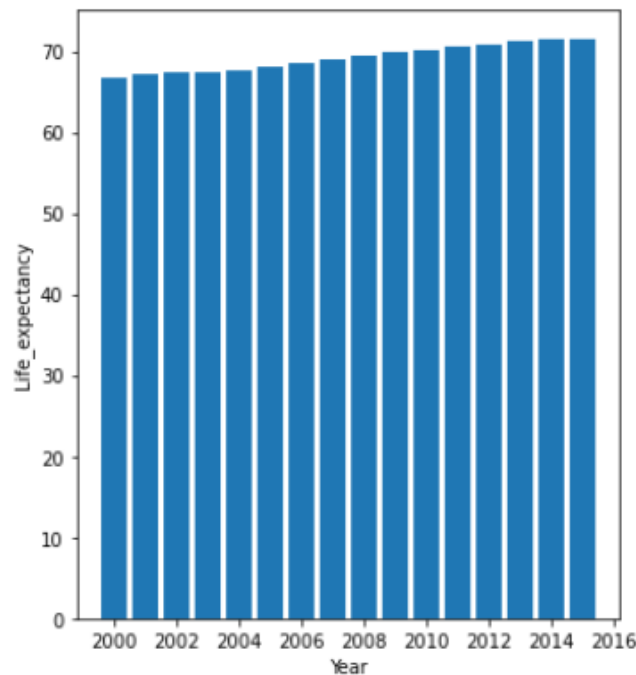
There are 25 countries which are investing more than 10 percentage of income in health care. The average Life expectancy is 71 years. The below graph shows the Life expectancy with percentage investment



Inference: Marshall Islands are investing in health care 17.2 percentage of income in health care which has life expectancy of 70. There are countries which are investing good percentage in health but still has less Life expectancy like Sierra Leone, Malawi, Lesotho, Burundi cause can be like this their income can be low or lack basic necessities like Running water.

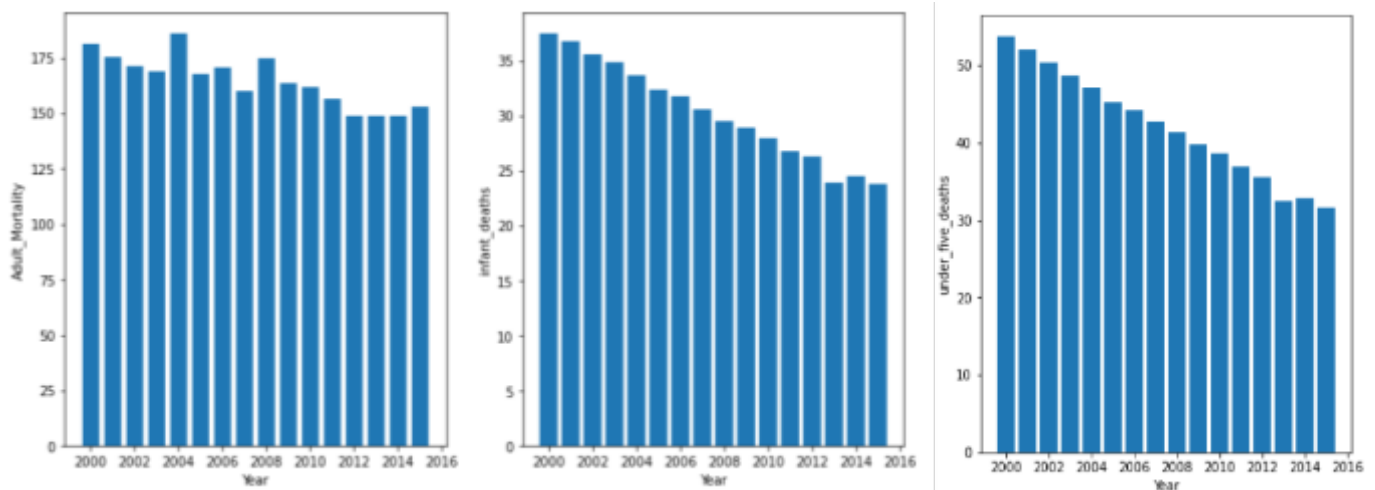
5. How features are changing based on Year:

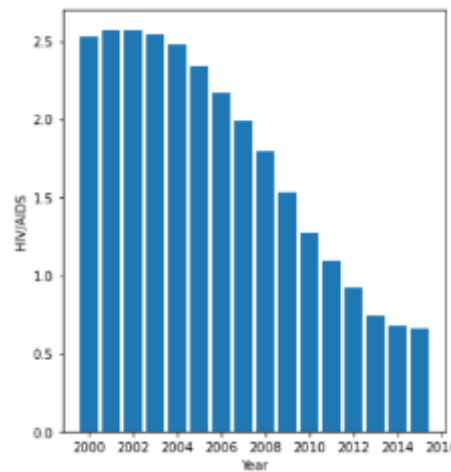
Purpose: Analysing feature based on year says that how the features are changed and how there should be changed.



It is clear that Average Life expectancy of world is gradually increasing because of factors like Immunization factors, Mortality factors, Economic factors, social factors.

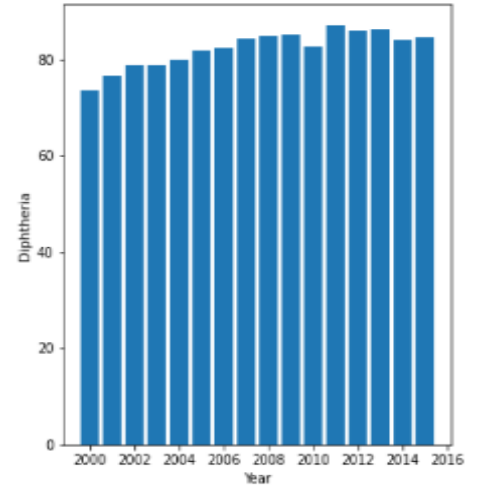
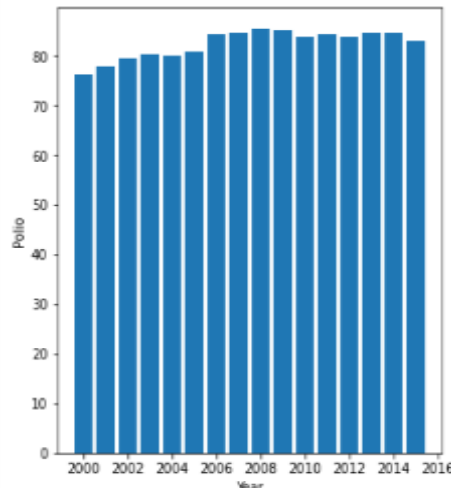
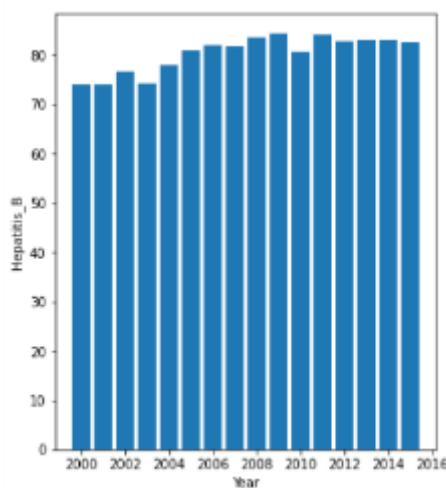
Mortality factors:





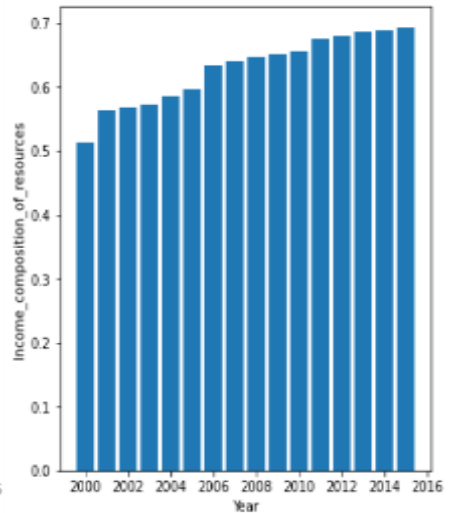
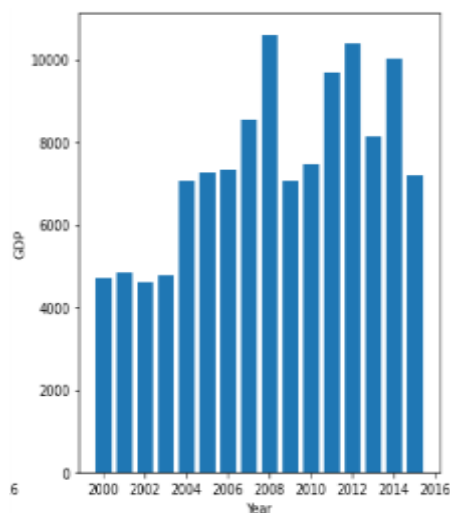
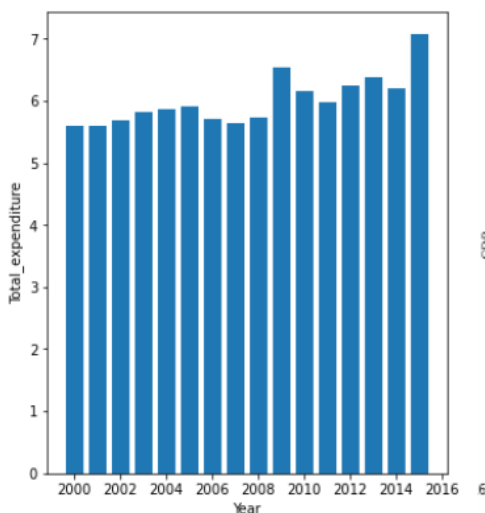
It is evident that morality Factors like Adult mortality, Infant deaths, Under-five deaths, HIV/AIDS are decreasing graphs. As year passing morality factors are decreasing which indicates in increases in Life Expectancy.

Immunization factors:



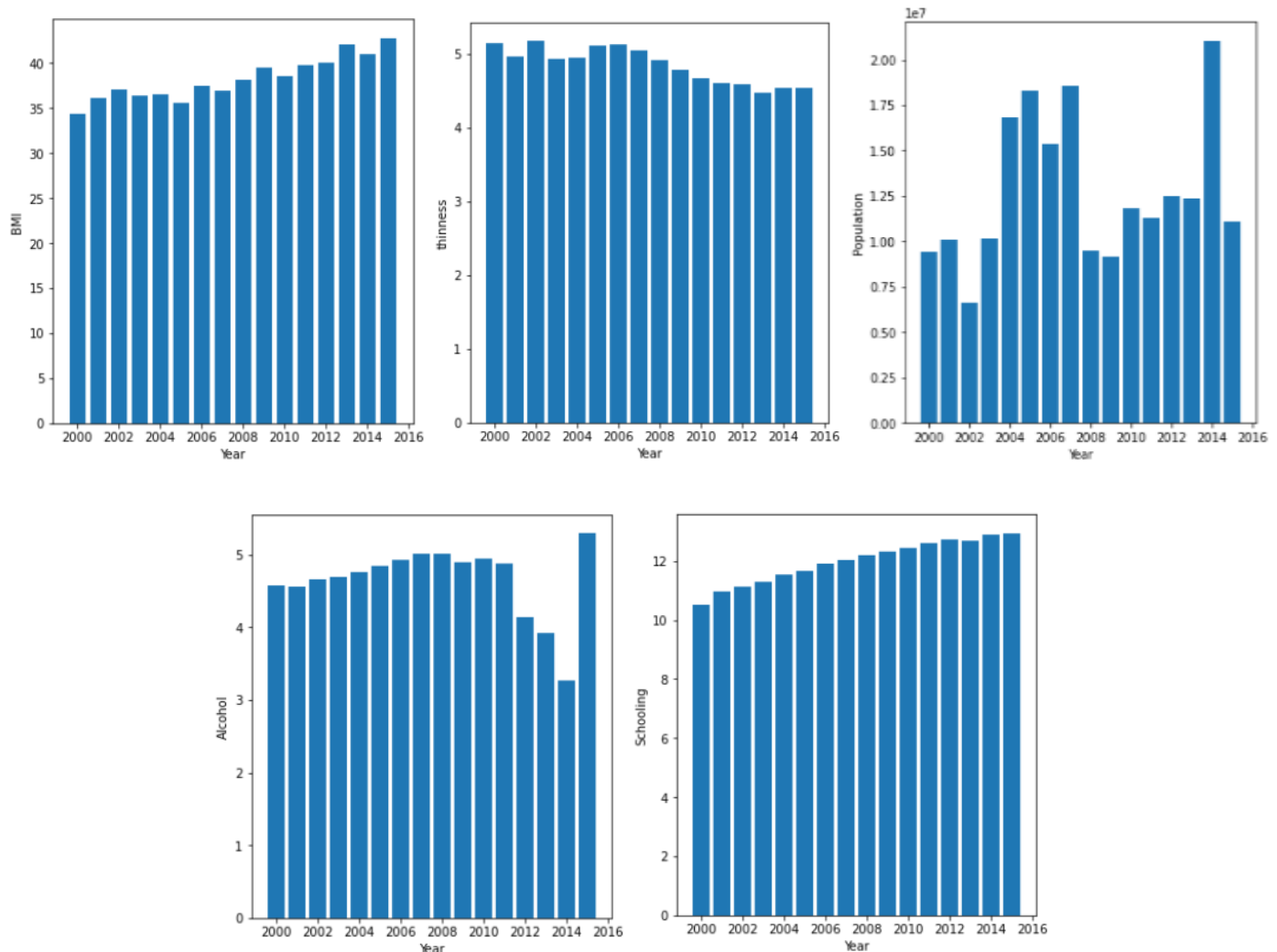
Immunization factors graphs are increasing because as the year are going people are getting vaccinated for respective Disease which effecting increases in Life Expectancy.

Economic factors:



In Economic factors Total expenditure in health care is increasing year by year. In GDP graph increasing is not that accurate but GDP bar compared to 2000 it was increased by 1/3rd in year 2016. Income composition of resources graph is also increasing as the investment towards health is increased Life expectancy is also increased.

Social factors:



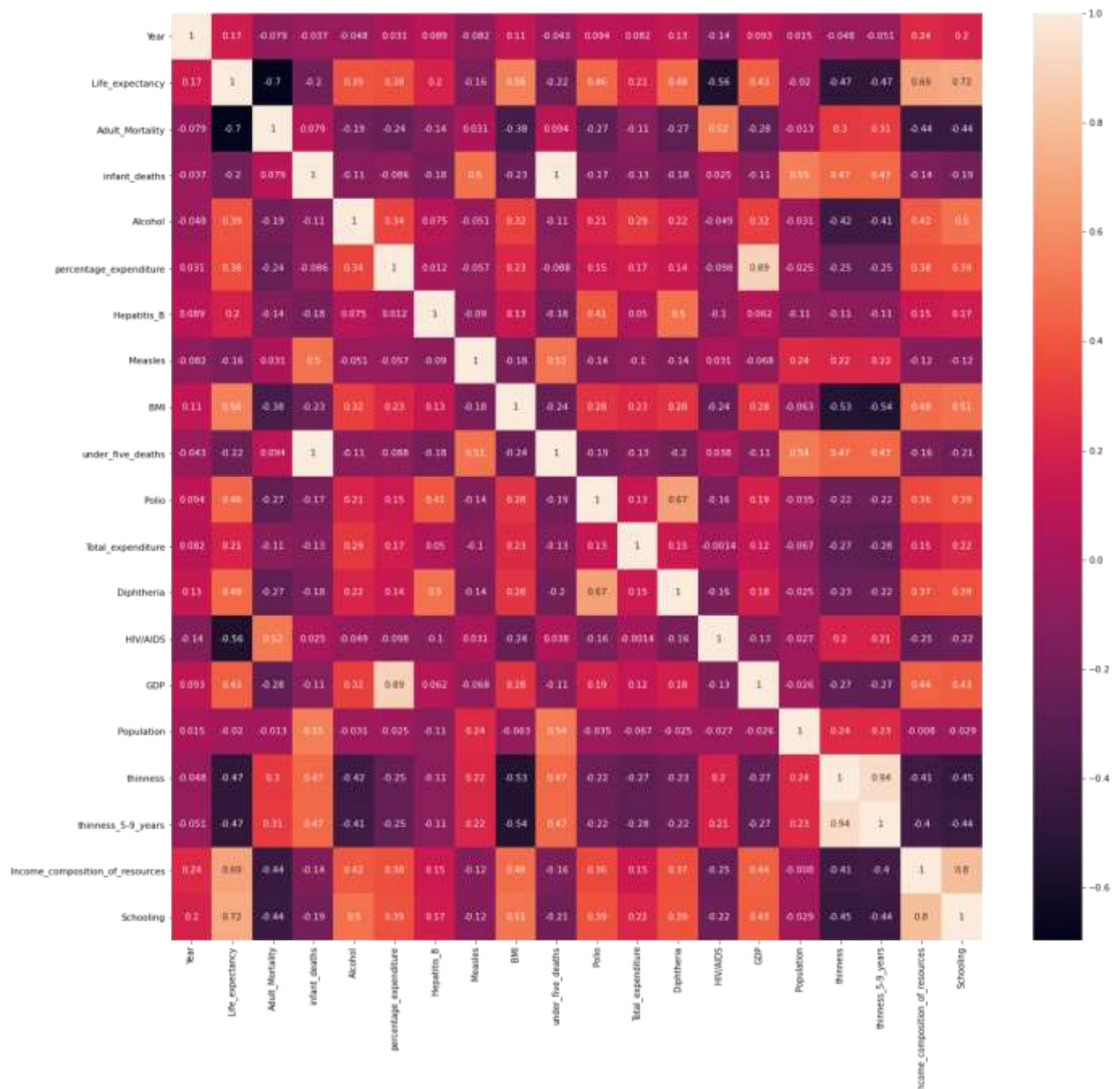
BMI is increases graph which leads to decrease in Life expectancy but BMI is not affecting much on Life expectancy. Thinness, alcohol graphs is decreasing which affects increase in Life expectancy. Number of schooling years graph is increasing which leads to increase in Life expectancy.

Inference:

Increase in Life expectancy is affected by Decreases in Morality factors, Increases in Immunization factors, Increases in Economic factors, in social factors it is dependent on factor decrease in use alcohol, increase in schooling year, Decreases in thinness.

6. Correlation between features:

Heatmap is the best way to represent the correlation between the features



Inference:

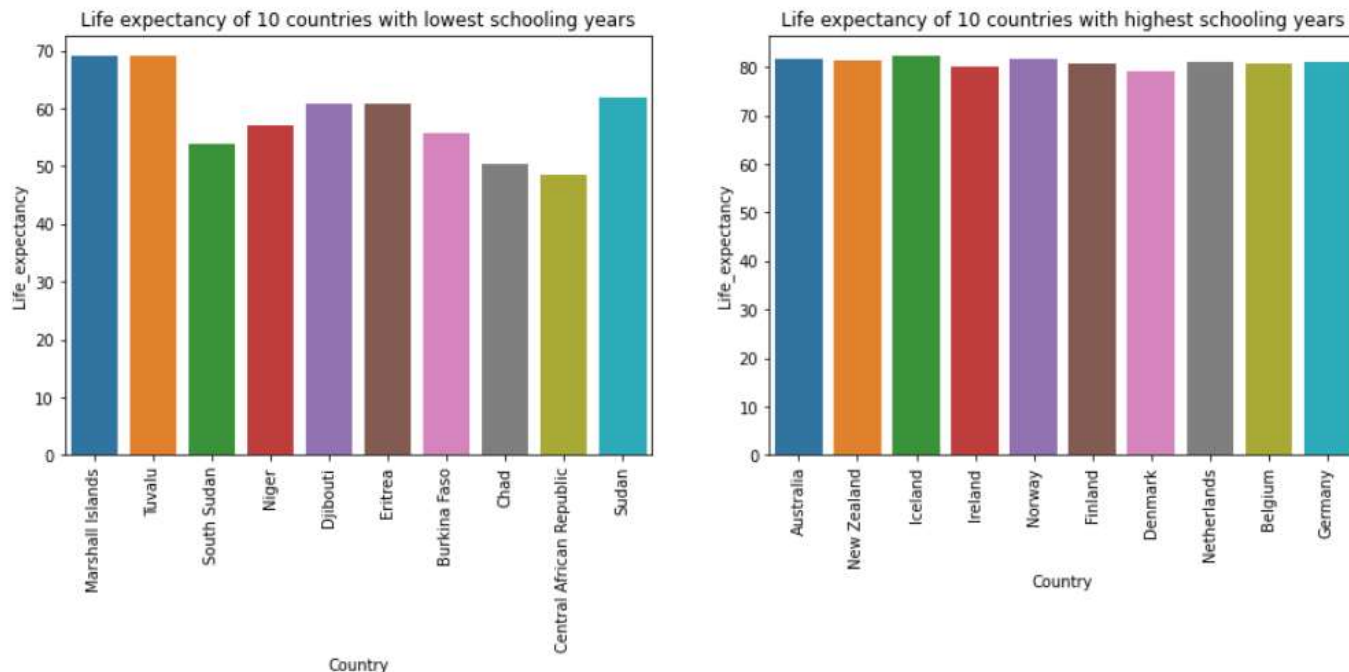
- Life expectancy is more correlated with Schooling and Income composition of resources with 0.72 and 0.69 respectively.
- The following features are highly correlated:
 - i. Under 5 deaths with Infant Deaths
 - ii. Thinness 5-9 with Thinness 1-19
- If a Feature is having correlation as 1 with any other feature, then that feature should be drop. It a kind of redundancy elimination.

7. Life Expectancy Vs highly correlated with Life Expectancy:

Purpose: To prove Correlation between Life expectancy and individual factor.

Life Expectancy Vs Schooling:

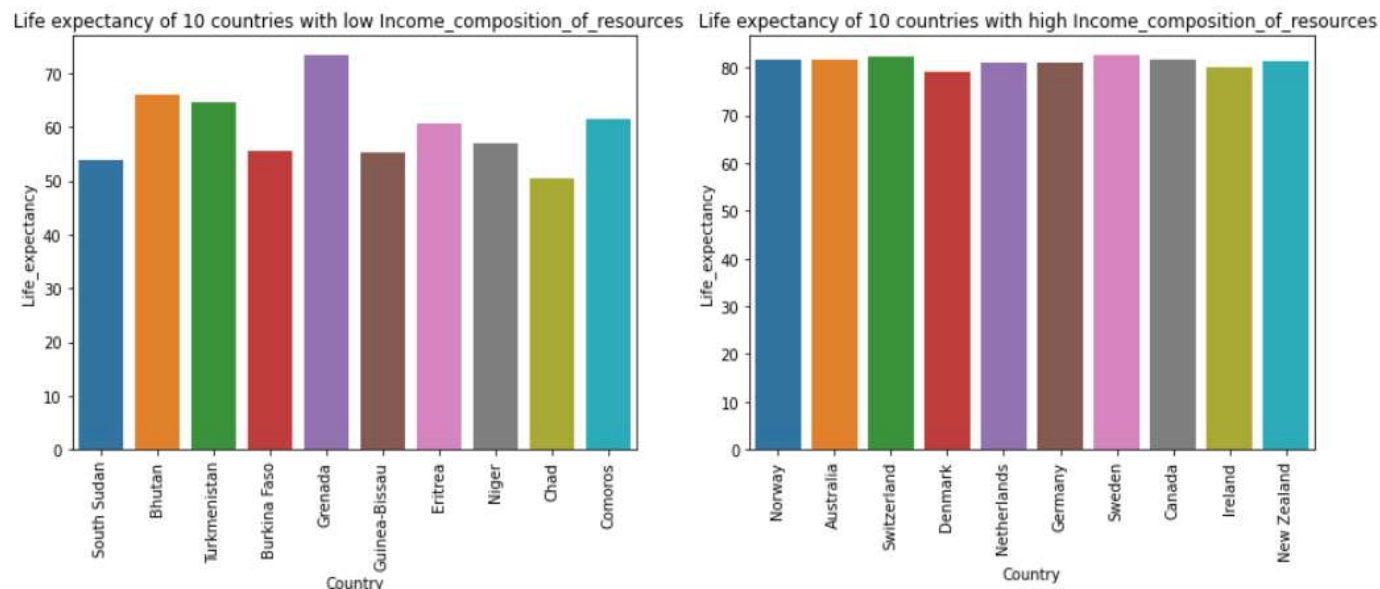
Life expectancy is more correlated with number of schooling years with 0.72. It can be demonstrate using below graph.



In countries which have more no of schooling years has the Average Life expectancy at most 80 years and the graph is also consistent but in the countries which have a smaller number of schooling years has less Life expectancy. It is clear from the graph the maximum Life expectancy in a smaller number of schooling years countries is 69.

Life Expectancy Vs Income composition of resources:

Life expectancy is more correlated with Income composition of resources with 0.69. It can be demonstrate using below graph.

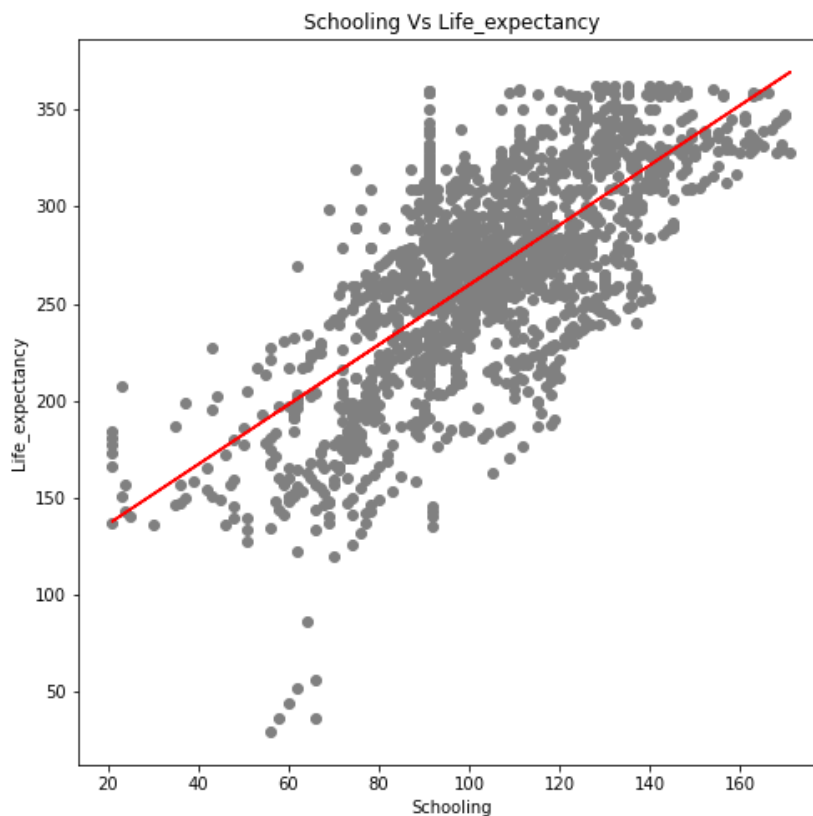


The Countries which have a high-income composition of resources have a good Average in Life Expectancy compared to countries that have a low-Income composition of resources because the country's productivity mainly depends on the Income composition of resources.

Regression Analysis:

1. Relationship between Life expectancy and Schooling:

Purpose: Analysing the Relationship between Life expectancy and schooling

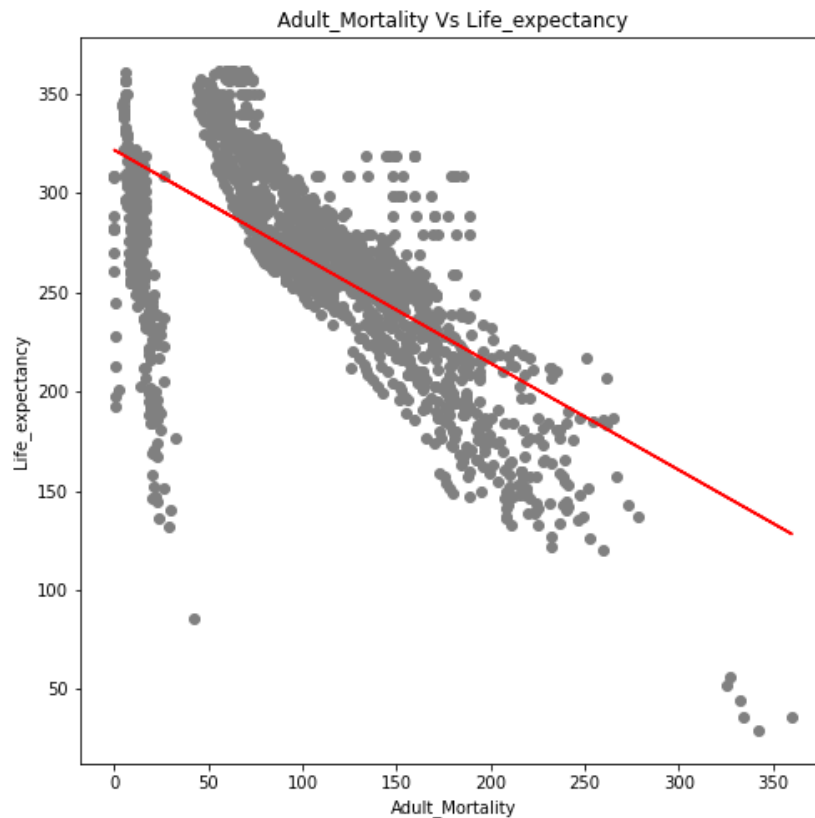


Points are scattered on the graph as the number of points is more curve can't be fitted. The Line can be fitted. So, the Linear Regression technique can be to analyse the relationship between Life expectancy and Schooling. Taking Independent variable as Schooling and dependent as Life expectancy. As a result, the slope is 1.544 and Bias is 105.17. So, the Regression line is $y = 1.544 X + 105.17$. The increase in value of Schooling with respect to Life expectancy is times 1.544 with the change of 105.17.

Inference: Positive correlation is existing between schooling and Life expectancy as the value of schooling increases the life expectancy is increasing.

2. Relationship between Life expectancy and Adult Mortality:

Purpose: Analysing the Relationship between Life expectancy and Adult Mortality

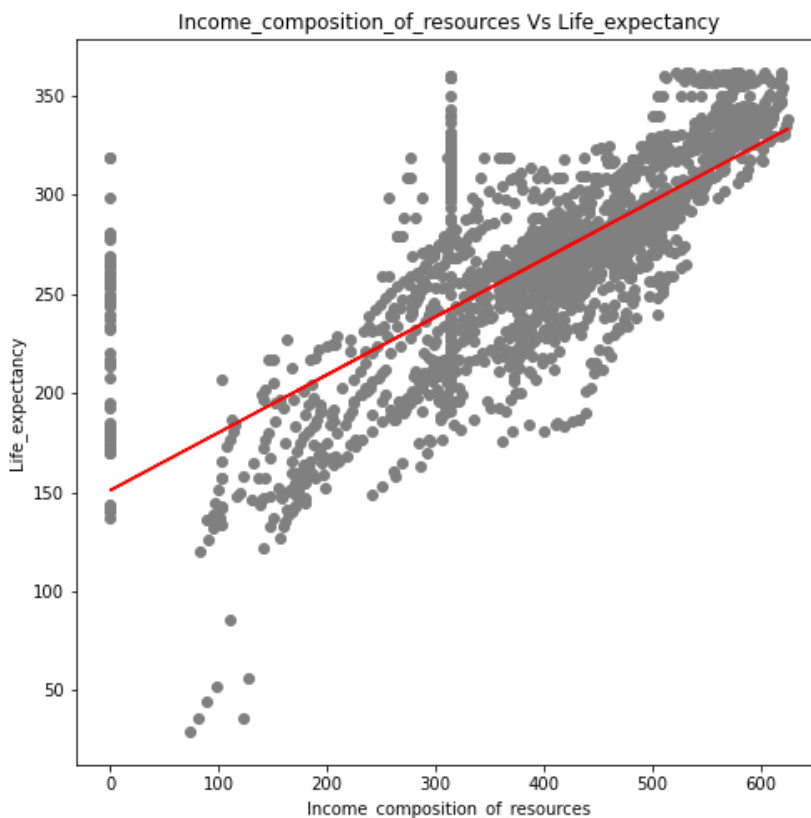


Points are scattered on the graph as the number of points is more curve can't be fitted. The Line can be fitted. So, the Linear Regression technique can be to analyse the relationship between Life expectancy and Adult Mortality. Taking Independent variable as Adult Mortality and dependent as Life expectancy. As a result, the slope is -0.538 and Bias is 321.94 . So, the Regression line is $y = -0.538 X + 321.94$. The negative value of slope indicates that the right-side angle of the line with the x-axis is more than 90 degrees. The increase in value of Adult Mortality with respect to Life expectancy is times -0.538 with the change of 321.94 .

Inference: Negative correlation is existing between Adult Mortality and Life expectancy as the value of Adult Mortality increases the life expectancy is decreases.

3. Relationship between Life expectancy and Income composition of resources:

Purpose: Analysing the Relationship between Life expectancy and Income composition of resources.

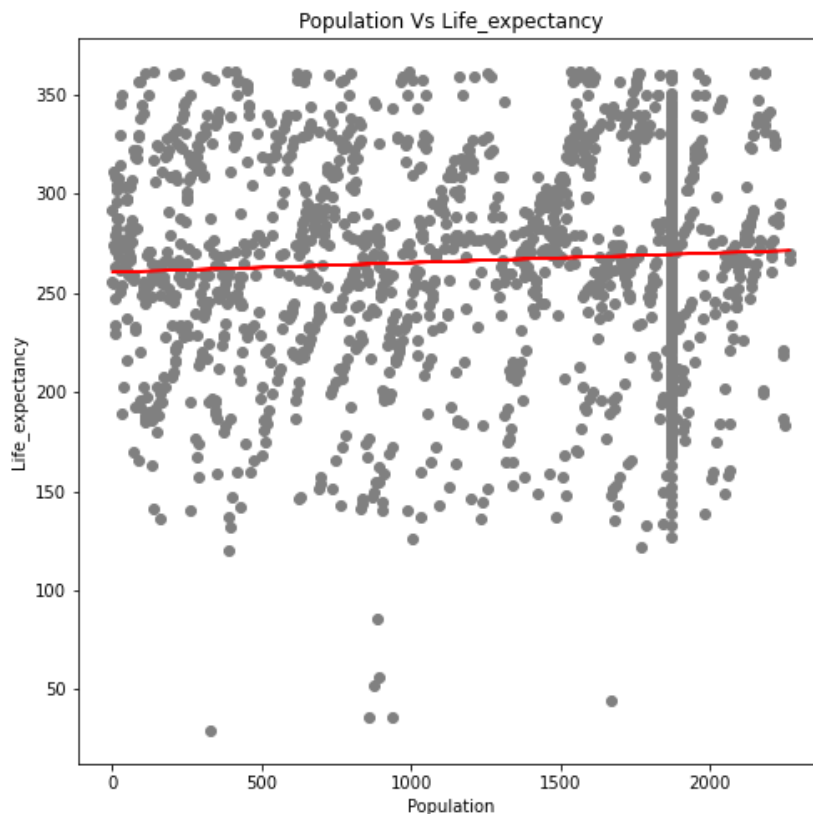


Points are scattered on the graph as the number of points is more curve can't be fitted. The Line can be fitted. So, the Linear Regression technique can be to analyse the relationship between Life expectancy and Income composition of resources. Taking Independent variable as Income composition of resources and dependent as Life expectancy. As a result, the slope is 0.291 and Bias is 150.9. So, the Regression line is $y = 0.291 X + 150.9$. The increase in value of Income composition of resources with respect to Life expectancy is times 0.291 with the change of 150.9.

Inference: Positive correlation is existing between the Income composition of resources and Life expectancy as the value of Income composition of resources increases the life expectancy is increasing.

4. Relationship between Life expectancy and Schooling:

Purpose: Analysing the Relationship between Life expectancy and Population.



Points are scattered on the graph as the number of points is more curve can't be fitted. The Line can be fitted. So, the Linear Regression technique can be to analyse the relationship between Life expectancy and Population. Taking Independent variable as Population and dependent as Life expectancy. As a result, the slope is 0.0048 and Bias is 260.56. So, the Regression line is $y = 0.0048 X + 260.56$. The increase in value of Population with respect to Life expectancy is times 0.0048 with the change of 260.56.

Inference: As the slope value is near to zero means Regression line is almost parallel to x – axis which indicates no correlation is exists between Population and Life expectancy as the value of Population increases the life expectancy is increasing.

Conclusion:

- Life expectancy in last 6 year is increased to many like Decreases in Morality factors, Increases in Immunization factors, Increases in Economic factors; in social factors it is dependent on factor decrease in use alcohol, increase in schooling year, Decreases in thinness.
- In developed countries the life expectancy is more compared to developing countries as the developed country has a greater number of schooling years and there can invest more in medical and health care.
- Income composition of resources indicates both the country development and Average Life expectancy of people in country.

