VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

EXPLORATORY DATA ANALYSIS ON FACTORS INFLUENCING LIFE EXPECTANCY

PROJECT REPORT

CSE5007: EXPLORATORY DATA ANALYSIS – EMBEDDED PROJECT

Under the guidance of

Dr. Rushi Kumar B
Associate Professor Sr., SAS,
VIT-Vellore

Submitted by:

Melisa Jiji (21MDT0004)
Srishti Todi (21MDT0005)
Divy Anjali (21MDT0065)

MSc. Data Science students, Department of Mathematics, SAS
Vellore Institute of Technology-Vellore, India

ACKNOWLEDGEMENT

We would like to express our special thanks and gratitude to our professor Dr. Rushi Kumar B for his constant help and support which led us to complete the project on time and which gave us the golden opportunity to do this wonderful project. We would also like to thank our friends for helping us learn new skills to make our project better and also the team members who gave their all in finalizing this project within the limited time frame.

CONTENT

INTRODUCTION

Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

EDA is primarily used to see what data can reveal beyond the formal modeling or hypothesis testing task and provides a better understanding of data set variables and the relationships between them. It can also help determine if the statistical techniques you are considering for data analysis are appropriate. Originally developed by American mathematician John Tukey in the 1970s, EDA techniques continue to be a widely used method in the data discovery process today.

OBJECTIVE

To analyse the predicting variables which are actually affecting life expectancy.

Our major objective here is to study the important factors affecting the life expectancies of various countries and get insights from the data.

MOTIVATION

Although there have been a lot of studies in the past regarding the factors that affect life expectancy and various comparisons have also been made; it was found that no one has compared the life expectancy amongst the continents. We create a new feature called continents and segregate the country continent-wise for this analysis.

It was also found that the effect of immunization and the human development index was not taken into account in the past. Thus, we take into account the effect of important immunizations like Hepatitis B, Polio, and Diphtheria.

BACKGROUND

Life expectancy is a statistical measure of the average time an organism is expected to live, based on the year of its birth, its current age, and other demographic factors like sex. To assess the quality of additional years of life, 'healthy life expectancy has been calculated for the last 30 years. Since 2001, the World Health Organization has published statistics called Healthy life expectancy (HALE), defined as the average number of years that a person can expect to live in "full health" excluding the years lived in less than full health due to disease and/or injury.

Disparities in life expectancy are often cited as demonstrating the need for better medical care or increased social support. A strongly associated indirect measure is income inequality. There are great variations in life expectancy between different parts of the world, mostly caused by differences in public health, medical care, and diet.

LITERATURE SURVEY

The data-set related to life expectancy, and health factors for 193 countries have been collected from the WHO data repository website and its corresponding economic data was collected from the United Nation website. Some of the past research was done considering multiple linear regression based on data set of one year for all the countries. Hence, this gives the motivation to resolve both the factors stated previously by formulating a regression model based on mixed effects model and multiple linear regression while considering data from a period of 2000 to 2015 for all the countries.

In 2022, Naor and Eden used Exploratory Data Analysis on the dataset to answer various questions. They also used three algorithms, namely, K-nearest Neighbours Regression, Linear Regression, and Decision tree Regression algorithm to predict the life expectancy where the Decision tree algorithm gave the highest R-squared score of 0.8898.

In 2021, Zelalum Getahun used libraries like Matplotlib, Seaborn, and Pandas to perform exploratory data analysis and make insightful inferences based on it, leaving scope for exploration on recent data, i.e., 2016-2021 data, creating a gap to make comparisons with previous data.

PROJECT DESCRIPTION

Life expectancy is the key metric for assessing population health. The project focuses on all the countries to determine the predicting factor which is contributing to the lower value of life expectancy. This will help in suggesting countries, which area should be given importance in order to efficiently improve the life expectancy of their population.

It relies on Exploratory Data Analysis for the accuracy of the life expectancy data for further study, which includes techniques like data extraction, data cleaning, and selecting major features from the data. The missing values are imputed wherein with a large dataset, missingness of up to 40% may even be acceptable.

Among all categories of health-related factors, those critical factors which are more representative of chosen affecting life expectancy will be segregated using feature selection. The resultant is visualized using various visualization techniques, depicting scatteredness, correlation, etc. among each other.

GOALS

- Visualizing the effects of various factors on life expectancy.
- Analysis of continent-wide life expectancy.
- To determine the predicting factor which leads to lower life expectancy.
- Identifying the main factors positively affecting life expectancy.
- To find whether densely populated countries tend to have a lower life expectancy.

RESULTS AND DISCUSSION

Using Juptyer Notebook for performing EDA.

- *Understanding the Data*

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

import warnings
warnings.filterwarnings("ignore")
```

`Country` : 193 unique values
`Status` : Developed or Developing status
`Life Expectancy` in age
`Adult Mortality Rates` of both sexes (probability of dying between 15 and 60 years per 1000 population)
`Number of Infant Deaths` per 1000 population
`infant deaths` : Number of Infant Deaths per 1000 population
`Alcohol` , recorded per capita (15+) consumption (in litres of pure alcohol)
`percentage expenditure` : Expenditure on health as a percentage of Gross Domestic Product per capita(%)
`Hepatitis B` (HepB) immunization coverage among 1-year-olds (%)
`Measles` : number of reported cases per 1000 population
`BMI` : Average Body Mass Index of entire population
`under-five deaths` : Number of under-five deaths per 1000 population
`Polio` (Pol3) immunization coverage among 1-year-olds (%)
`Total expenditure` : General government expenditure on health as a percentage of total government expenditure (%)
`Diphtheria` tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)
`HIV/AIDS` : Deaths per 1000 live births HIV/AIDS (0-4 years)
`GDP` : Gross Domestic Product per capita (in USD)
`thinness 1-19 years` : Prevalence of thinness among children and adolescents for Age 10 to 19 (%)
`thinness 5-9 years` : Prevalence of thinness among children for Age 5 to 9(%)
`Income composition of resources` : Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
`Schooling` : Number of years of Schooling(years)

In [4]:
```
data = pd.read_csv("life_data.csv")
data.head()
```

Out[4]:

	Country	Continent	Year	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	...	Polio	Total expenditure	Diphtheria	HIV/AIDS	G
0	Afghanistan	Asia	2015	Developing	65.0	263.0	62	0.01	71.279624	65.0	...	6.0	8.16	65.0	0.1	584.259
1	Afghanistan	Asia	2014	Developing	59.9	271.0	64	0.01	73.523582	62.0	...	58.0	8.18	62.0	0.1	612.696
2	Afghanistan	Asia	2013	Developing	59.9	268.0	66	0.01	73.219243	64.0	...	62.0	8.13	64.0	0.1	631.744
3	Afghanistan	Asia	2012	Developing	59.5	272.0	69	0.01	78.184215	67.0	...	67.0	8.52	67.0	0.1	669.959
4	Afghanistan	Asia	2011	Developing	59.2	275.0	71	0.01	7.097109	68.0	...	68.0	7.87	68.0	0.1	63.537

5 rows × 23 columns

In [5]:
```
data.columns
```

Out[5]:
```
Index(['Country', 'Continent', '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'],
      dtype='object')
```

In [6]:
```
data.shape
```

Out[6]: (2938, 23)

```
In [5]: data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2938 entries, 0 to 2937
Data columns (total 23 columns):
 #   Column                           Non-Null Count  Dtype
---  ------                           --------------  -----
 0   Country                          2938 non-null   object
 1   Continent                        2938 non-null   object
 2   Year                             2938 non-null   int64
 3   Status                           2938 non-null   object
 4   Life expectancy                  2928 non-null   float64
 5   Adult Mortality                  2928 non-null   float64
 6   infant deaths                    2938 non-null   int64
 7   Alcohol                          2744 non-null   float64
 8   percentage expenditure           2938 non-null   float64
 9   Hepatitis B                      2385 non-null   float64
 10  Measles                          2938 non-null   int64
 11   BMI                             2904 non-null   float64
 12  under-five deaths                2938 non-null   int64
 13  Polio                            2919 non-null   float64
 14  Total expenditure                2712 non-null   float64
 15  Diphtheria                       2919 non-null   float64
 16   HIV/AIDS                        2938 non-null   float64
 17  GDP                              2490 non-null   float64
 18  Population                       2286 non-null   float64
 19   thinness  1-19 years            2904 non-null   float64
 20   thinness 5-9 years             2904 non-null   float64
 21  Income composition of resources  2771 non-null   float64
 22  Schooling                        2775 non-null   float64
dtypes: float64(16), int64(4), object(3)
memory usage: 528.0+ KB
```

```
In [6]: data.nunique()
```

```
Out[6]: Country                          193
Continent                          6
Year                              16
Status                             2
Life expectancy                  362
Adult Mortality                  425
infant deaths                    209
Alcohol                         1076
percentage expenditure          2328
Hepatitis B                       87
Measles                          958
 BMI                             608
under-five deaths                252
Polio                             73
Total expenditure                818
Diphtheria                        81
 HIV/AIDS                        200
GDP                             2490
Population                      2278
 thinness  1-19 years            200
 thinness 5-9 years             207
Income composition of resources  625
Schooling                        173
dtype: int64
```

- *Data Cleaning: Null Value Correction and Imputations*

```
In [7]: data.isnull().sum()
```

```
Out[7]: Country                          0
Continent                        0
Year                             0
Status                           0
Life expectancy                 10
Adult Mortality                 10
infant deaths                    0
Alcohol                        194
percentage expenditure           0
Hepatitis B                    553
Measles                          0
 BMI                            34
under-five deaths                0
Polio                           19
Total expenditure              226
Diphtheria                      19
 HIV/AIDS                        0
GDP                            448
Population                     652
 thinness  1-19 years           34
 thinness 5-9 years             34
Income composition of resources  167
Schooling                      163
dtype: int64
```

```
In [8]: # Column-wise null percentages

round(data.isnull().sum()*100/data.isnull().sum().sum(),2)
```

```
Out[8]: Country                          0.00
Continent                        0.00
Year                             0.00
Status                           0.00
Life expectancy                  0.39
Adult Mortality                  0.39
infant deaths                    0.00
Alcohol                          7.57
percentage expenditure           0.00
Hepatitis B                     21.58
Measles                          0.00
 BMI                             1.33
under-five deaths                0.00
Polio                            0.74
Total expenditure                8.82
Diphtheria                       0.74
 HIV/AIDS                        0.00
GDP                             17.48
Population                      25.44
 thinness  1-19 years            1.33
 thinness 5-9 years              1.33
Income composition of resources  6.52
Schooling                        6.36
dtype: float64
```

GDP, Hepatitis B and Population covers majority of the NaN values.

```
In [6]: #understanding the data
df.describe()
```

Out[6]:

	Year	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	BMI	under-five deaths	Polio	exp
count	2938.000000	2928.000000	2928.000000	2938.000000	2744.000000	2938.000000	2385.000000	2938.000000	2904.000000	2938.000000	2919.000000	27
mean	2007.518720	69.224932	164.796448	30.303948	4.602861	738.251295	80.940461	2419.592240	38.321247	42.035739	82.550188	
std	4.613841	9.523867	124.292079	117.926501	4.052413	1987.914858	25.070016	11467.272489	20.044034	160.445548	23.428046	
min	2000.000000	36.300000	1.000000	0.000000	0.010000	0.000000	1.000000	0.000000	1.000000	0.000000	3.000000	
25%	2004.000000	63.100000	74.000000	0.000000	0.877500	4.685343	77.000000	0.000000	19.300000	0.000000	78.000000	
50%	2008.000000	72.100000	144.000000	3.000000	3.755000	64.912906	92.000000	17.000000	43.500000	4.000000	93.000000	
75%	2012.000000	75.700000	228.000000	22.000000	7.702500	441.534144	97.000000	360.250000	56.200000	28.000000	97.000000	
max	2015.000000	89.000000	723.000000	1800.000000	17.870000	19479.911610	99.000000	212183.000000	87.300000	2500.000000	99.000000	

Here, we notice a few anomalies:

1. Adult mortality of 1 doesn't make sense.

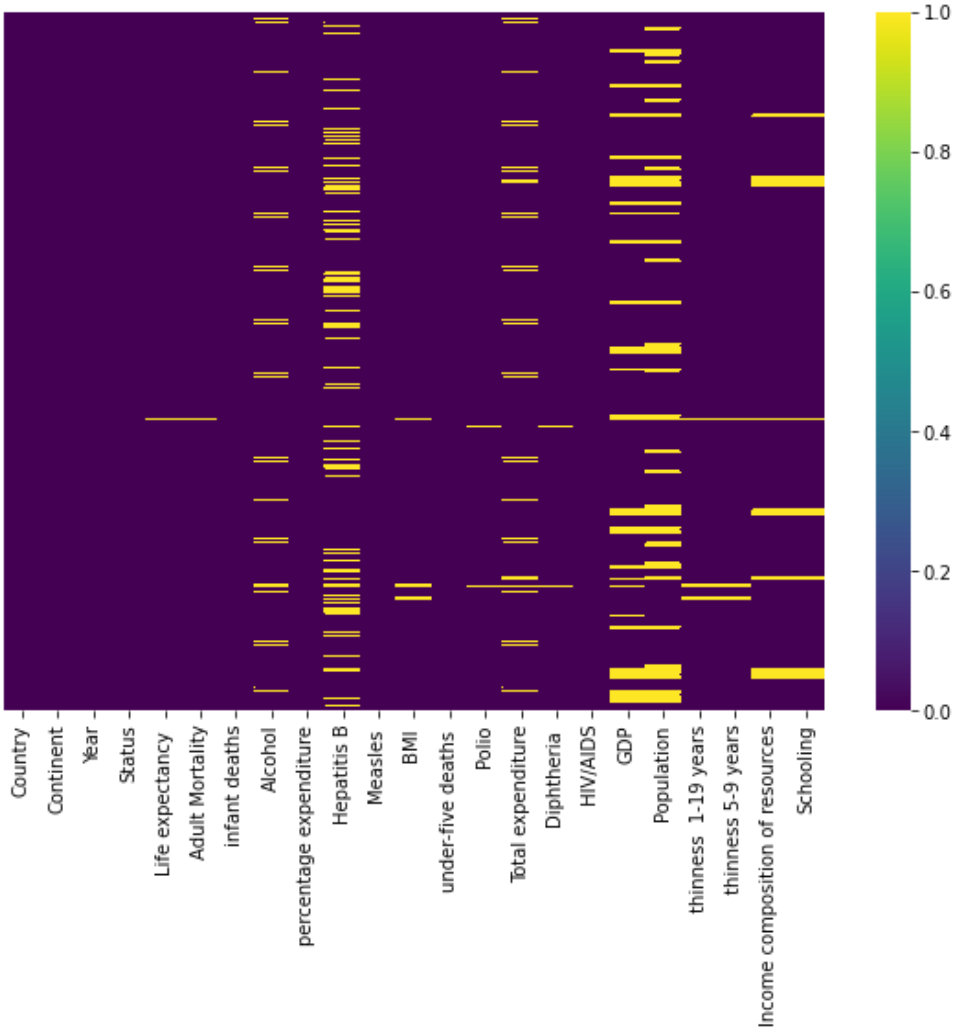
2. Infant deaths and Under-five deaths cannot be 0. That is logically impossible.

3. BMI should be between 110 and 50 as lower than 10 suggests excessive under-weight and above 50 suggests extreme obessity. Thus, minimum value of 1 and maximum value of 87 suggests anomalies in the data.

- *Null Values Heatmap*

```
In [9]: # taking a look at the spread of the null values on a heatmap for clear visualisation

        plt.figure(figsize=(10,7))
        sns.heatmap(data.isnull(),yticklabels=False,cbar=True,cmap="viridis")
        plt.show()
```



◦ Treating the null values

1. Since GDP and Population almost have null values in the same row together (mostly in developing countries), and do not account for any major effect on the Life Expectancy, we drop the columns altogether.

2. Treating the null values of 'Hepatitis B', 'Alcohol', 'Total expenditure' column using mean imputation for each country.

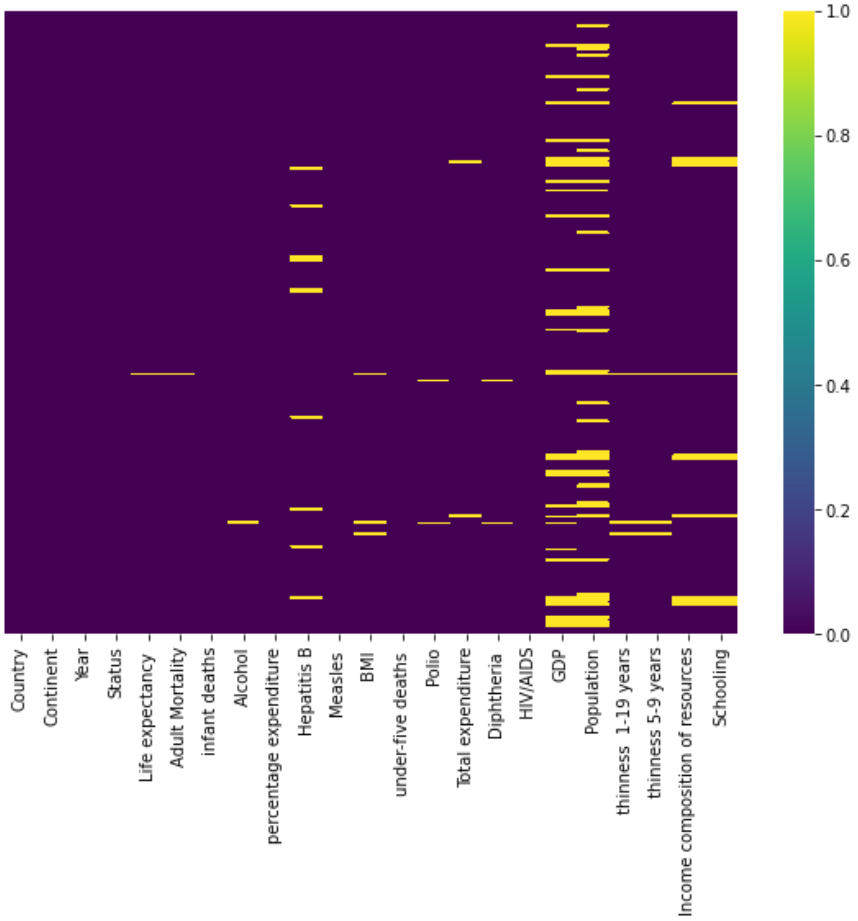
In [10]:
```
data.isna().sum()
```

Out[10]:
```
Country                            0
Continent                          0
Year                               0
Status                             0
Life expectancy                   10
Adult Mortality                   10
infant deaths                      0
Alcohol                          194
percentage expenditure             0
Hepatitis B                      553
Measles                            0
 BMI                              34
under-five deaths                  0
Polio                             19
Total expenditure                226
Diphtheria                        19
 HIV/AIDS                          0
GDP                              448
Population                       652
 thinness  1-19 years             34
 thinness 5-9 years               34
Income composition of resources  167
Schooling                        163
dtype: int64
```

In [7]:
```
for i, col in enumerate(['Hepatitis B', "Alcohol","Total expenditure"], start=1):
    data[col] = data[col].fillna(data.groupby('Country')[col].transform('mean'))
    data[col] = data[col].round(2)
```

In [12]:
```
# taking a look at the spread of the null values on a heatmap for clear visualisation

plt.figure(figsize=(10,7))
sns.heatmap(data.isnull(),yticklabels=False,cbar=True,cmap="viridis")
plt.show()
```



Imputations

- There are some data values in the dataset which do not make any sense in real life.
- Treating those data values we get -

In [8]:
```
# Replacing Adult mortality rates lower than the 5th percentile with mean value

mort = np.percentile(data["Adult Mortality"].dropna(),5)
data["Adult Mortality"] = data.apply(lambda x: np.nan if x["Adult Mortality"] < mort else x["Adult Mortality"], axis=1)
data["Adult Mortality"] = data["Adult Mortality"].fillna(data.groupby('Country')["Adult Mortality"].transform('mean'))
data["Adult Mortality"] = data["Adult Mortality"].round(3)
```

In [9]:
```
# Wensorise the invalid BMI(less than 10 and more than 50)

data[" BMI "] =data.apply(lambda x : 10 if (x[" BMI "] <10) else (50 if x[" BMI "] > 50 else x[" BMI "]), axis=1)
```

In [10]:
```
# Replacing 'Income composition of resources', 'Schooling', 'infant deaths', 'under-five deaths ' of 0 with mean value

for i, col in enumerate(['Income composition of resources', "Schooling","infant deaths","under-five deaths "], start=1):
    data[col] = data[col].replace(0,np.nan)
    data[col] = data[col].fillna(data.groupby('Country')[col].transform('mean'))
    data[col] = data[col].round(3)
data.iloc[10:25]
```

Out[10]:

	Country	Continent	Year	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	...	Polio	Total expenditure	Diphtheria	HIV/AIDS	
10	Afghanistan	Asia	2005	Developing	57.3	291.000	85.0	0.02	1.388648	66.0	...	58.0	8.70	58.0	0.1	25.2
11	Afghanistan	Asia	2004	Developing	57.0	293.000	87.0	0.02	15.296066	67.0	...	5.0	8.79	5.0	0.1	219.1
12	Afghanistan	Asia	2003	Developing	56.7	295.000	87.0	0.01	11.089053	65.0	...	41.0	8.82	41.0	0.1	198.7
13	Afghanistan	Asia	2002	Developing	56.2	286.800	88.0	0.01	16.887351	64.0	...	36.0	7.76	36.0	0.1	187.8
14	Afghanistan	Asia	2001	Developing	55.3	316.000	88.0	0.01	10.574728	63.0	...	35.0	7.80	33.0	0.1	117.4
15	Afghanistan	Asia	2000	Developing	54.8	321.000	88.0	0.01	10.424960	62.0	...	24.0	8.20	24.0	0.1	114.5
16	Albania	Europe	2015	Developing	77.8	74.000	1.0	4.60	364.975229	99.0	...	99.0	6.00	99.0	0.1	3954.2
17	Albania	Europe	2014	Developing	77.5	57.667	1.0	4.51	428.749067	98.0	...	98.0	5.88	98.0	0.1	4575.7
18	Albania	Europe	2013	Developing	77.2	84.000	1.0	4.76	430.876979	99.0	...	99.0	5.66	99.0	0.1	4414.7
19	Albania	Europe	2012	Developing	76.9	86.000	1.0	5.14	412.443356	99.0	...	99.0	5.59	99.0	0.1	4247.6
20	Albania	Europe	2011	Developing	76.6	88.000	1.0	5.37	437.062100	99.0	...	99.0	5.71	99.0	0.1	4437.1
21	Albania	Europe	2010	Developing	76.2	91.000	1.0	5.28	41.822757	99.0	...	99.0	5.34	99.0	0.1	494.3
22	Albania	Europe	2009	Developing	76.1	91.000	1.0	5.79	348.055952	98.0	...	98.0	5.79	98.0	0.1	4114.1
23	Albania	Europe	2008	Developing	75.3	57.667	1.0	5.61	36.622068	99.0	...	99.0	5.87	99.0	0.1	437.5
24	Albania	Europe	2007	Developing	75.9	57.667	1.0	5.58	32.246552	98.0	...	99.0	6.10	98.0	0.1	363.1

15 rows × 23 columns

In [16]:
```
data = data.interpolate(method = 'linear', limit_direction = 'forward')
data.iloc[60:68]
```

Out[16]:

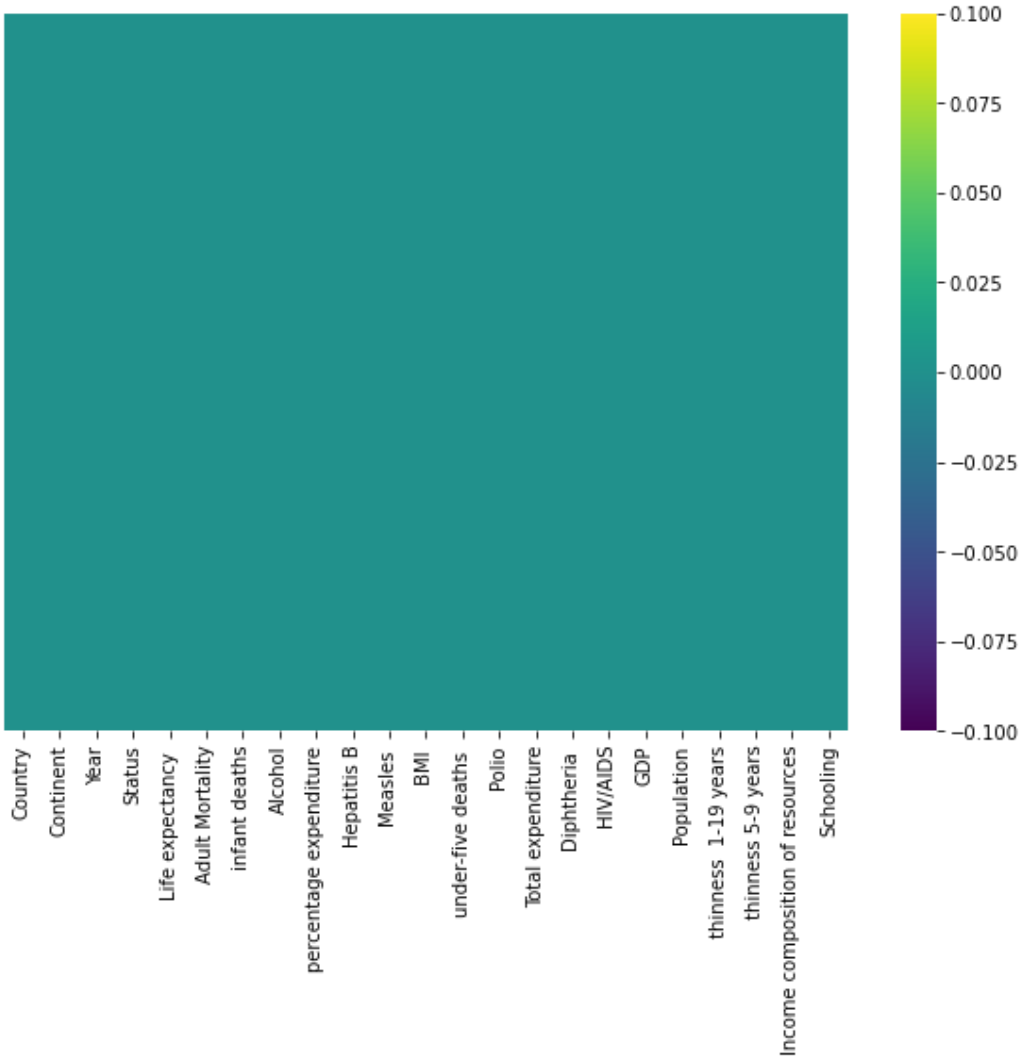
	Country	Continent	Year	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	...	Polio	Total expenditure	Diphtheria	HIV/AIDS	
60	Angola	Africa	2003	Developing	46.8	388.0	95.000000	3.49	35.933491	70.22	...	4.0	4.41	4.0	2.4	779
61	Angola	Africa	2002	Developing	46.5	391.0	96.000000	2.82	24.037942	70.22	...	37.0	3.63	41.0	2.3	711
62	Angola	Africa	2001	Developing	45.7	44.0	97.000000	2.58	30.359936	70.22	...	41.0	5.38	38.0	2.1	526
63	Angola	Africa	2000	Developing	45.3	48.0	97.000000	1.85	15.881493	70.22	...	3.0	2.79	28.0	2.0	555
64	Antigua and Barbuda	North America	2015	Developing	76.4	13.0	91.764706	7.95	0.000000	99.00	...	86.0	4.79	99.0	0.2	13566
65	Antigua and Barbuda	North America	2014	Developing	76.2	131.0	86.529412	8.56	2422.999774	99.00	...	96.0	5.54	99.0	0.2	12888
66	Antigua and Barbuda	North America	2013	Developing	76.1	133.0	81.294118	8.58	1991.430372	99.00	...	98.0	5.33	99.0	0.2	12224
67	Antigua and Barbuda	North America	2012	Developing	75.9	134.0	76.058824	8.18	2156.229842	98.00	...	97.0	5.39	98.0	0.2	12565

8 rows × 23 columns

- *Cleaned Data*

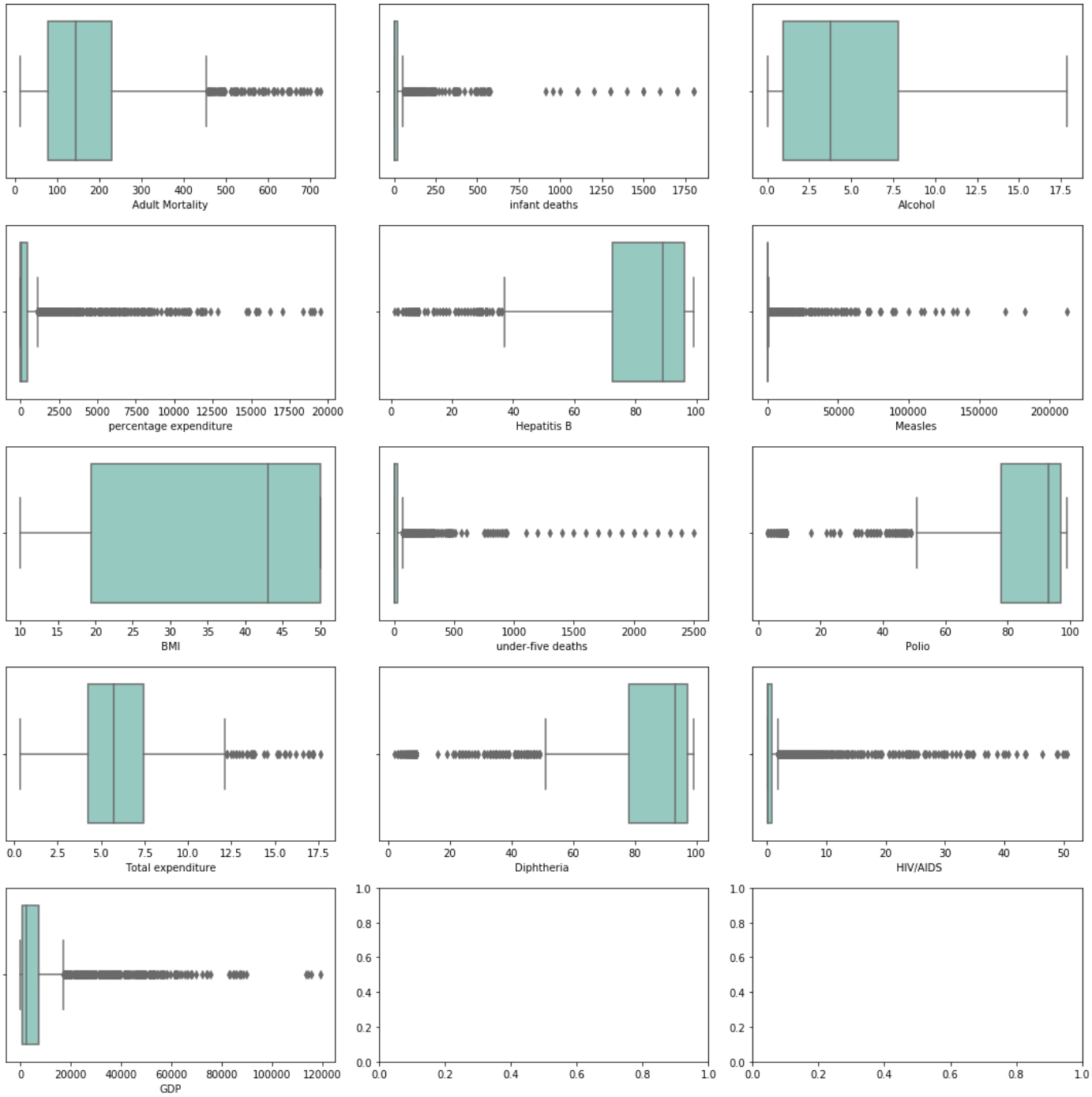
In [17]:
```
# Final presentation of cleaned data

plt.figure(figsize=(10,7))
sns.heatmap(data.isnull(),yticklabels=False,cbar=True,cmap="viridis")
plt.show()
```



- *Analysis of outliers*

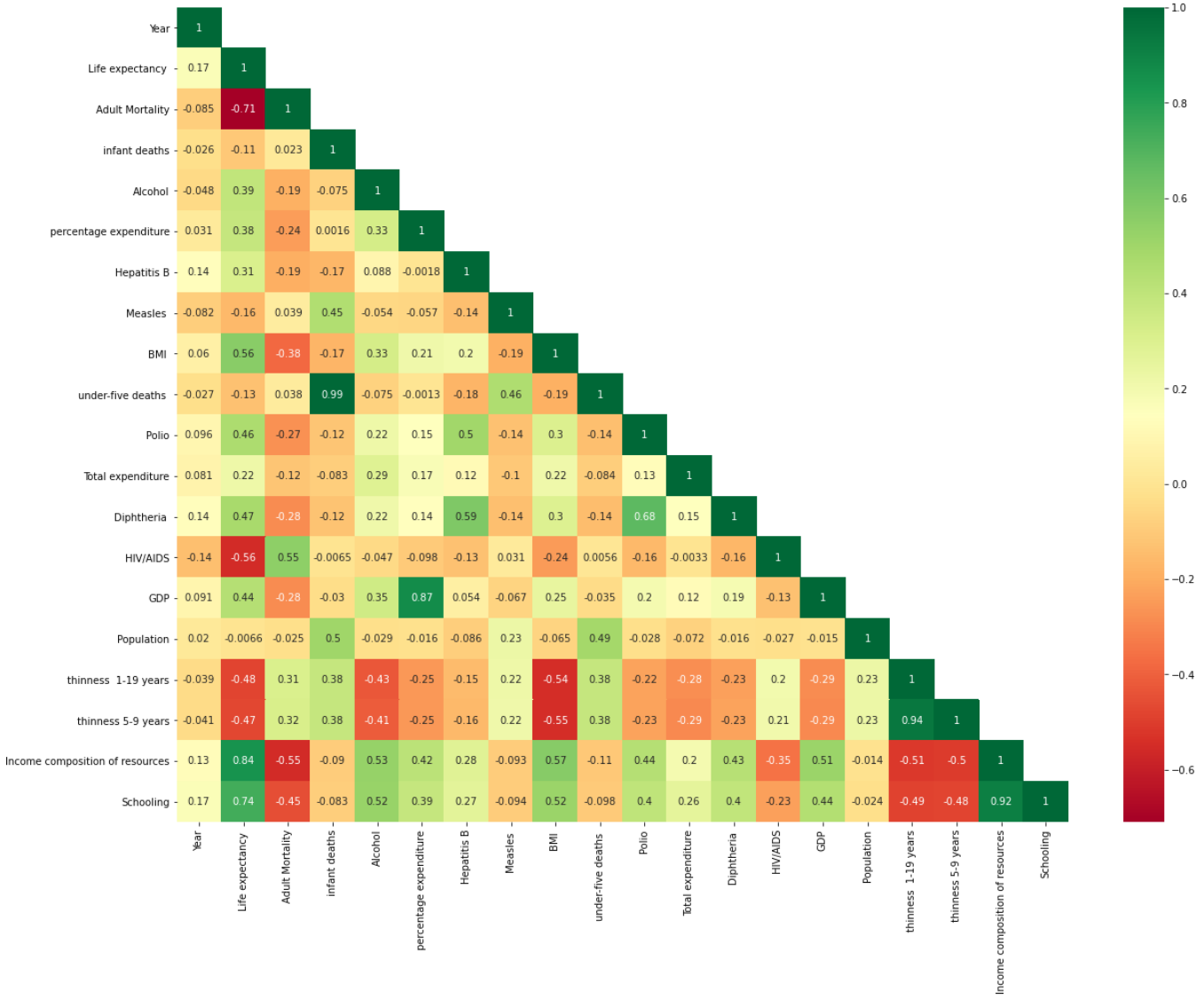
```
rows=5
cols=3
fig,ax=plt.subplots(nrows=rows,ncols=cols,figsize=(15,15))
col=data.columns
index=5
for i in range(rows):
    for j in range(cols):
        sns.boxplot(data[col[index]],ax=ax[i][j], palette="Set3")
        index=index+1
        if index>=18:
            break
plt.tight_layout()
```



We can observe that there is a huge amount of outliers in "infant deaths", "percentage expenditure", "Hepatitis B" "Measles", "under-five deaths", "HIV/AIDS", and "Population"
1. Looking at measle values concentrated at 0, the original data for some specific countries are absent and thus assigned as 0.
2. HIV/AIDS values are mostly concentrated around 0 to 1 in developed countries and range up to 10 in developing countries.
Due to the absence of data in some columns for specific countries, there are different ranges of outliers (considering all the countries). Hence, the outliers cause no harm to the data.

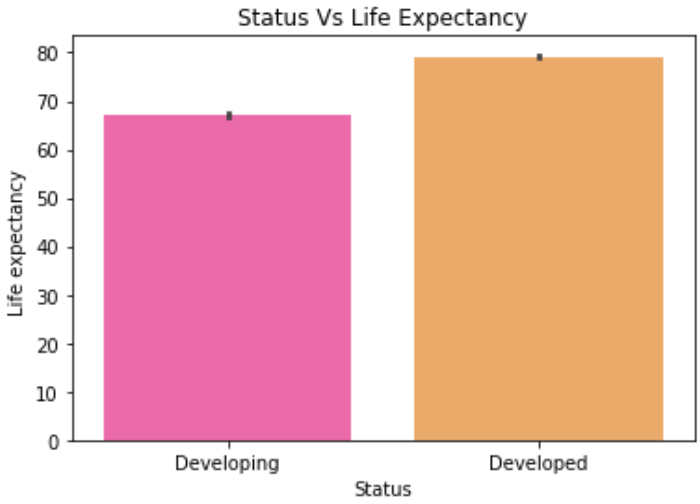
- *Correlation Matrix*

```
In [27]: plt.figure(figsize=(20,15))
         mask = np.triu(np.ones_like(data.corr(), dtype=bool), k=1)
         sns.heatmap(data.corr(), annot=True, cmap='RdYlGn', mask=mask);
```

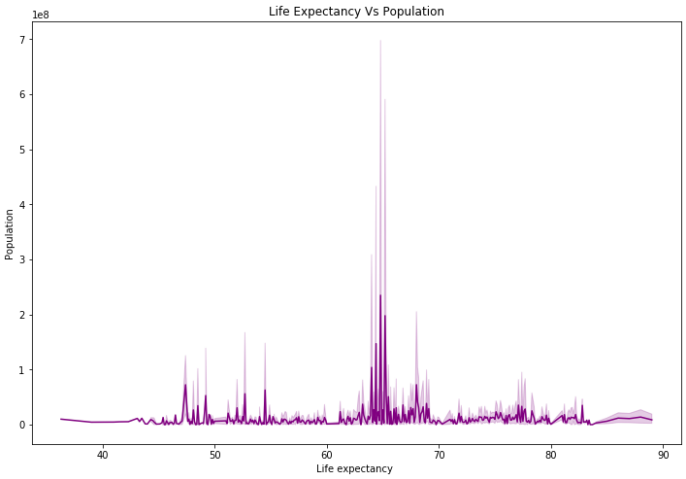


We see that life expectancy is negatively correlated to HIV/AIDS, under-five deaths, measles, population, adult mortality, infant deaths, and thinness. It is strongly positively correlated to income composition of resources, schooling, GDP, percentage expenditure and on immunizations.

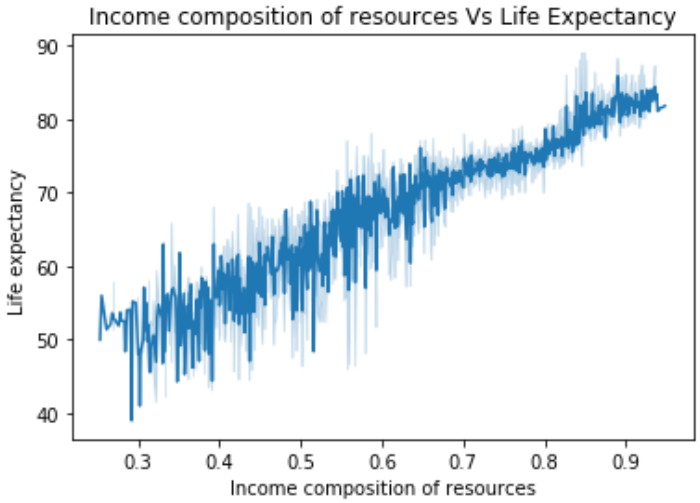
- *Data Visualization*



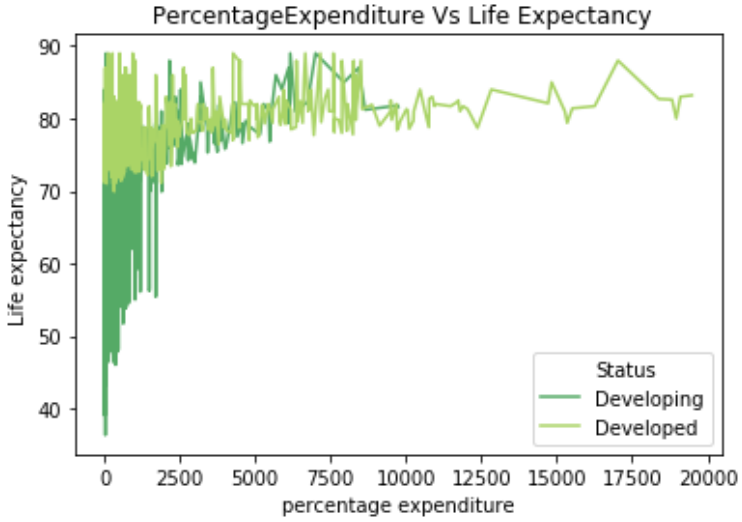
The life expectancy in developed countries is more than in developing countries.



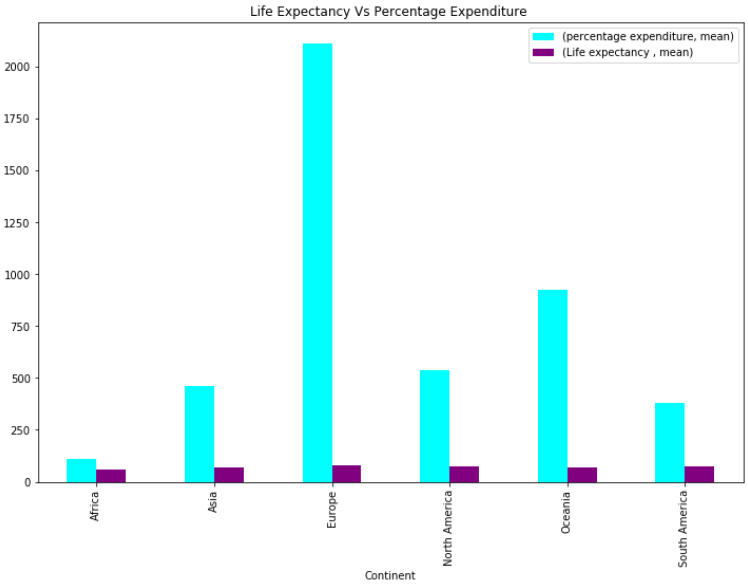
The densely populated countries have a life expectancy up to the age group 60-70 years.



The income compositions of resources have a linear trend with life expectancy.

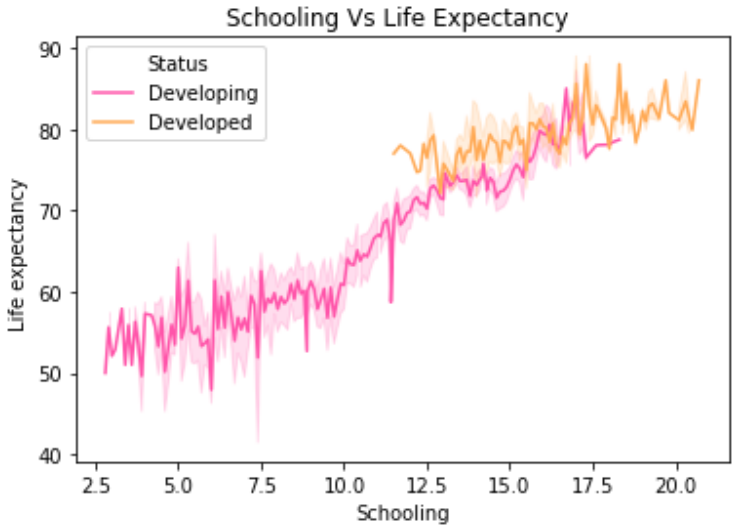


The life expectancy of developed countries is higher regardless of the expenditure amount and the maximum expenditure extends up to 20,000.
Whereas the life expectancy in developing countries considerably increases *with an increase in expenditure and extends mostly up to 10,000.*



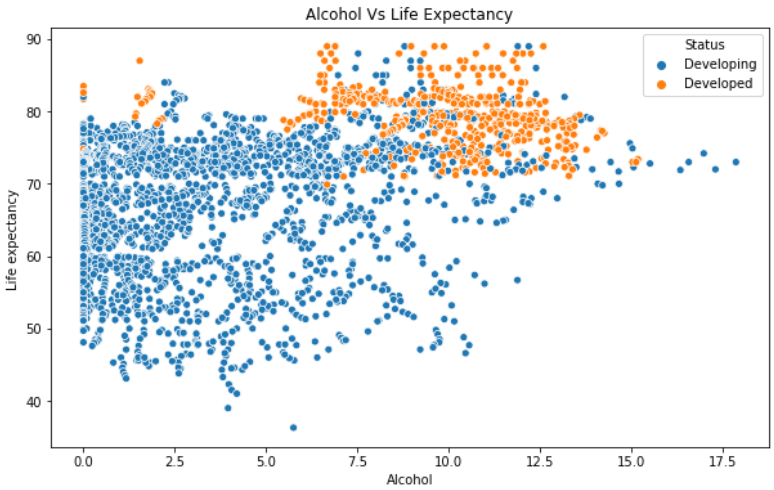
Although the life expectancy of Asia Europe and America are approximately the same, the expenditure in Europe is drastically higher than in any other continent.

Also, Africa has the lowest Expenditure value as compared to the other countries.



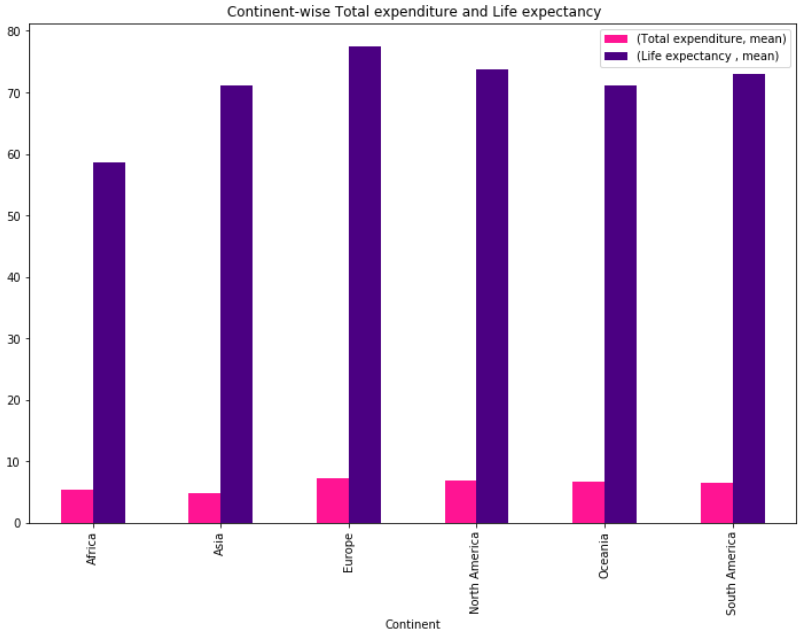
The Schooling period is higher in developed countries where the life expectancy takes only values above 70.

Developing countries have a maximum to a minimum range of schooling intervals and a linear growth in life expectancy rate.

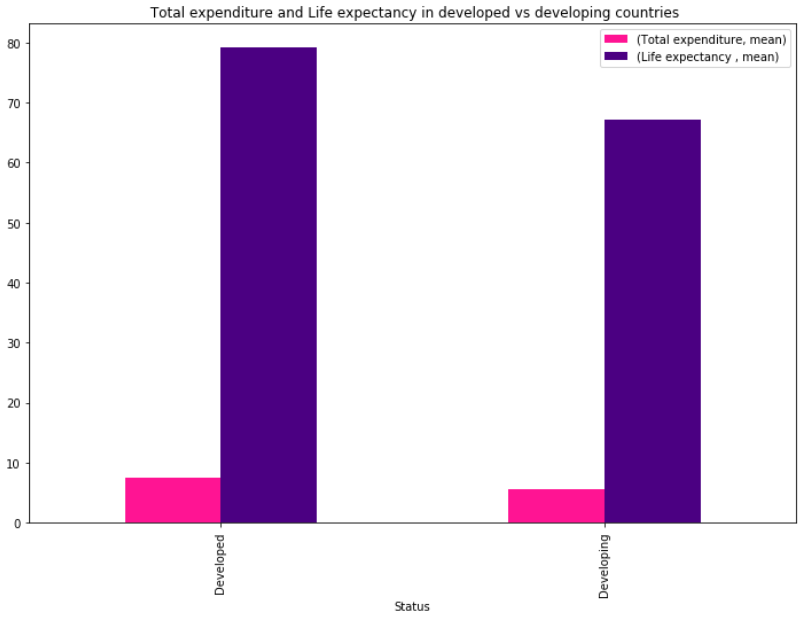


Alcohol consumption in developing countries is higher than in developed countries.

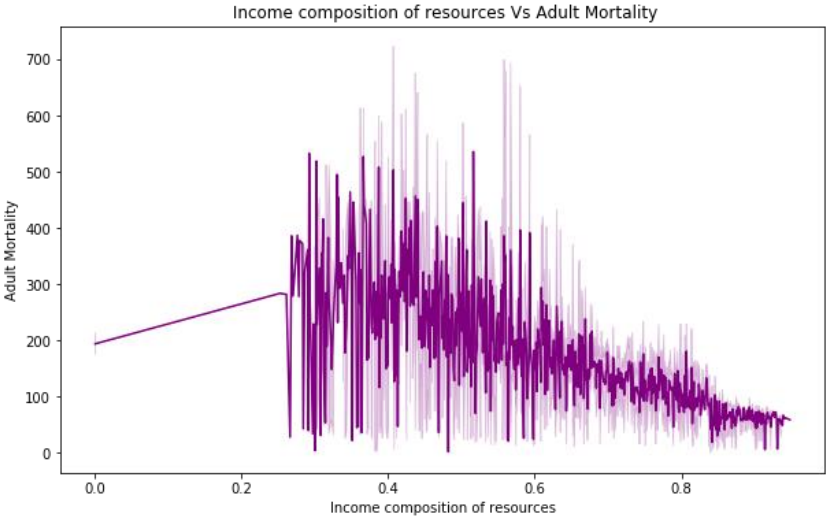
The life expectancy of the people in developing countries is moderately lower than people living in developed countries.



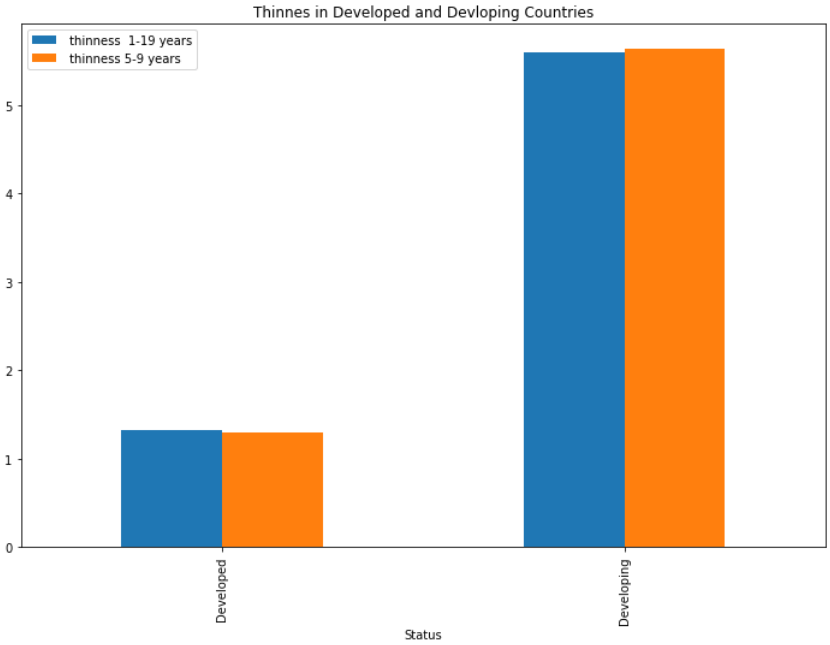
Total expenditure for Europe is the highest and so is its life expectancy. Total expenditure for Asia is the lowest but the life expectancy in Africa is the lowest.



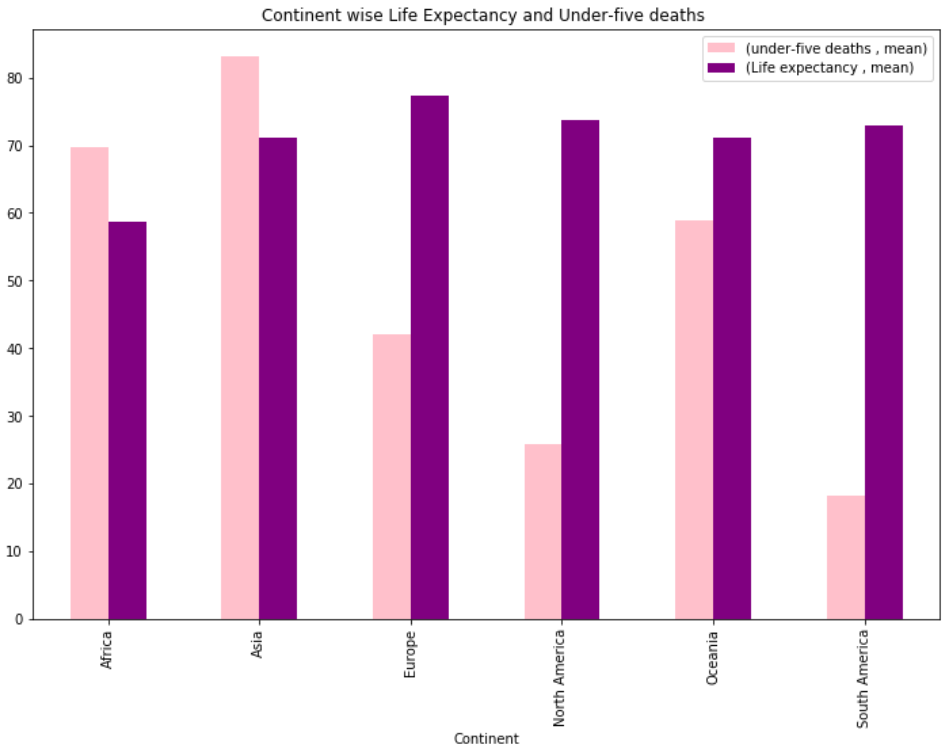
Total expenditure for developed countries is higher. Thus, their life expectancy is also higher compared to developing countries.



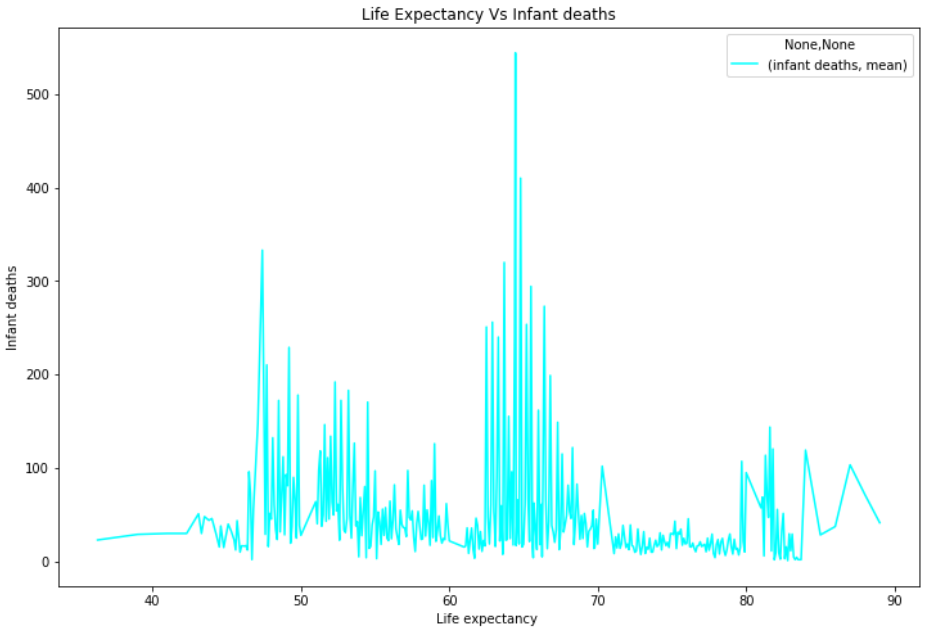
We see that as income composition increases, Adult Mortality decreases considerably which is consistent with our previous result that income composition is directly related to Life Expectancy.



The prevalence of thinness among children and adolescents in developed countries is much lower than in developing countries which can be due to the low lifestyle in these countries.

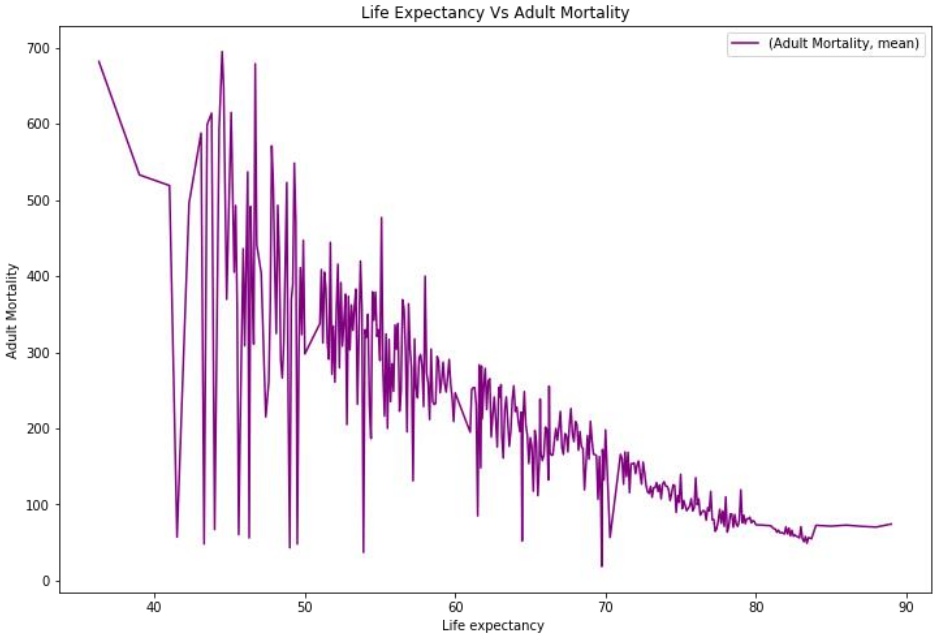


South America has the least 'under-five deaths' and a considerably high average life expectancy. Whereas Asia and Africa have higher mortality than the life expectancy.



Countries with a life expectancy between 62-68 years have the highest infant deaths.

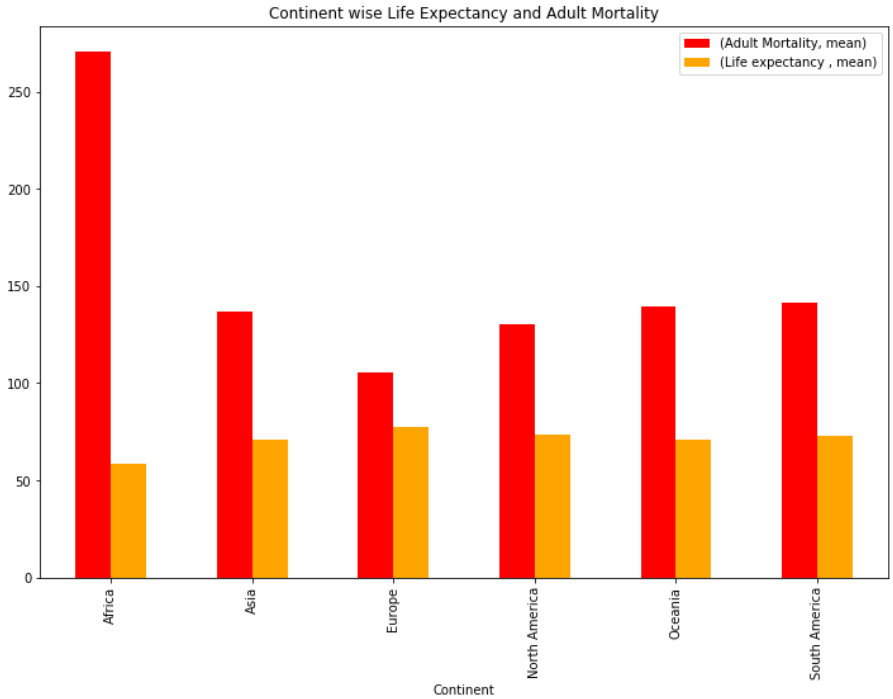
Countries with a life expectancy, of 45-58 years also have noticeably high infant deaths.



The adult mortality rate for the life expectancy of different countries has a very high fluctuation but in a decreasing pattern.

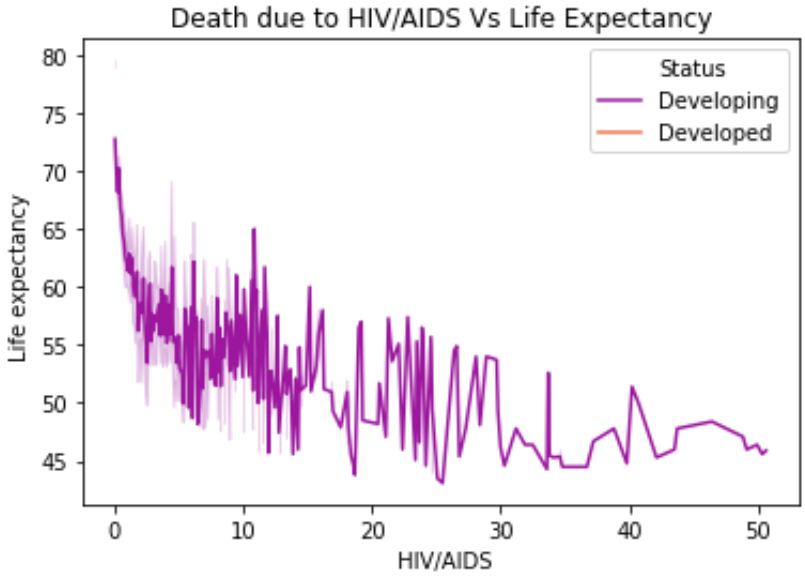
In countries with a life expectancy between 40-50 years, there is both a high and low adult mortality.

The adult mortality rate is low in countries with high life expectancy.

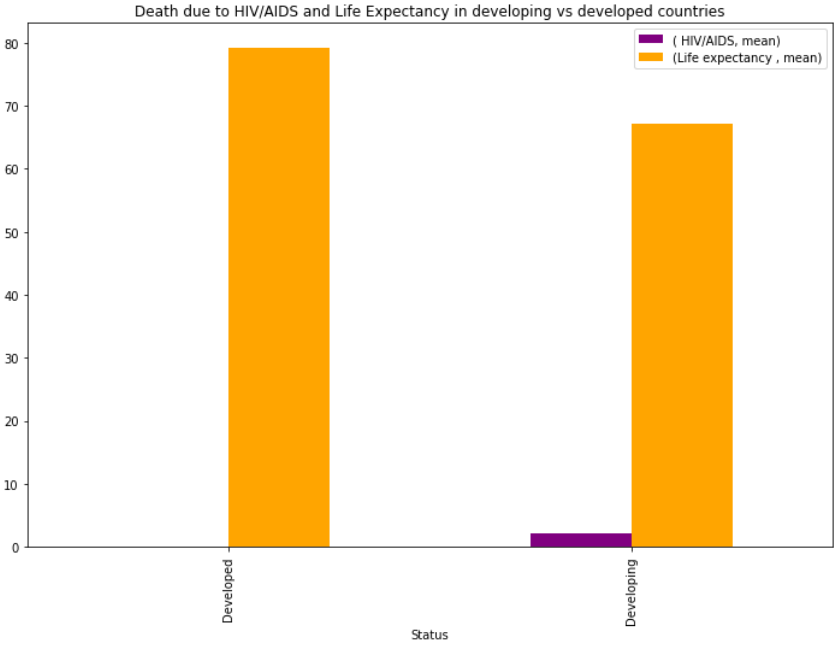


Africa the has highest Adult mortality and lowest life expectancy.

Europe has the lowest adult mortality and highest life expectancy.

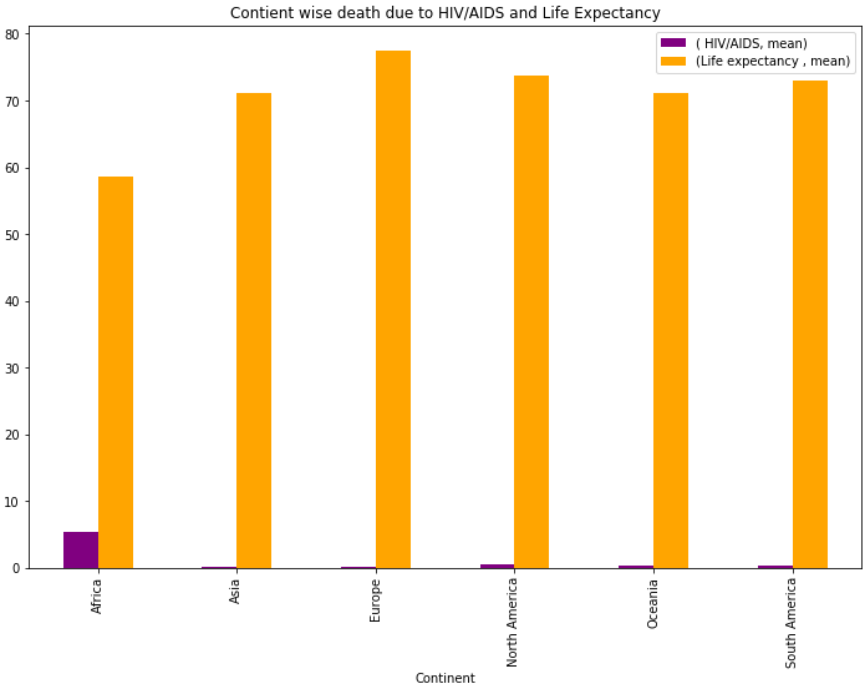


Countries with higher life expectancy have lower death rates due to HIV/AIDS.



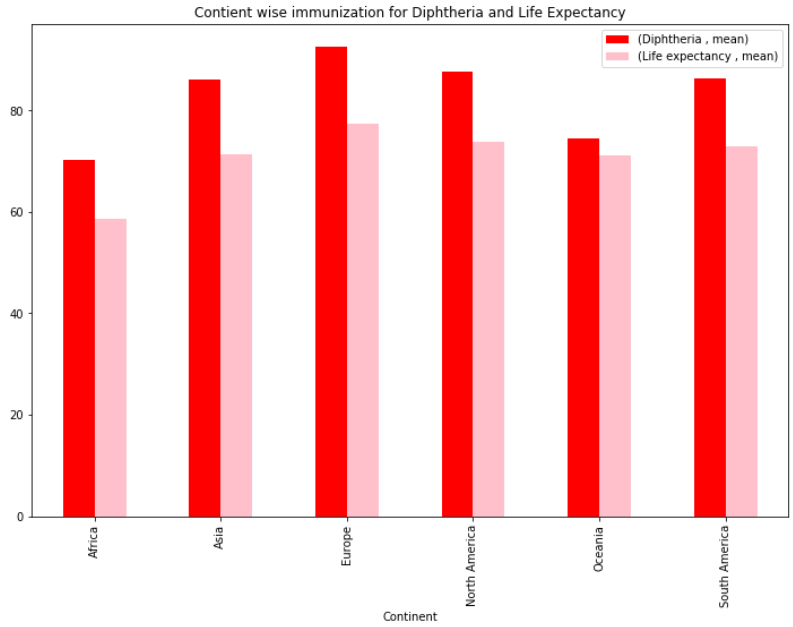
Developing countries have deaths due to HIV/AIDS.

Whereas Developed Countries have no evident deaths due to HIV/AIDS.



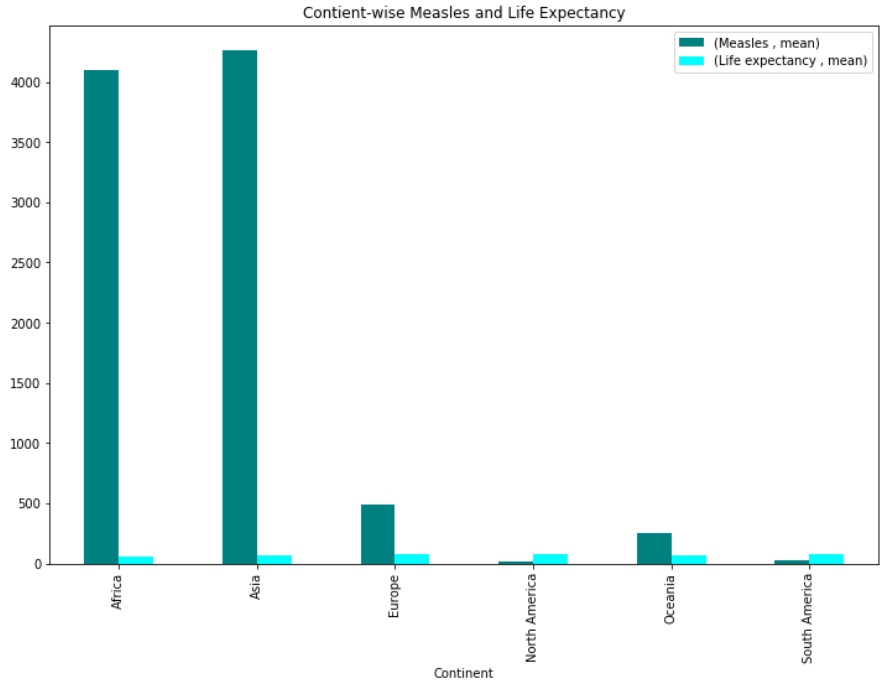
Africa has a high death rate due to HIV/AIDS and a low life expectancy.

Europe has a low death rate due to HIV/AIDS and a high life expectancy.



Africa has the lowest immunization for Diphtheria and a low life expectancy.

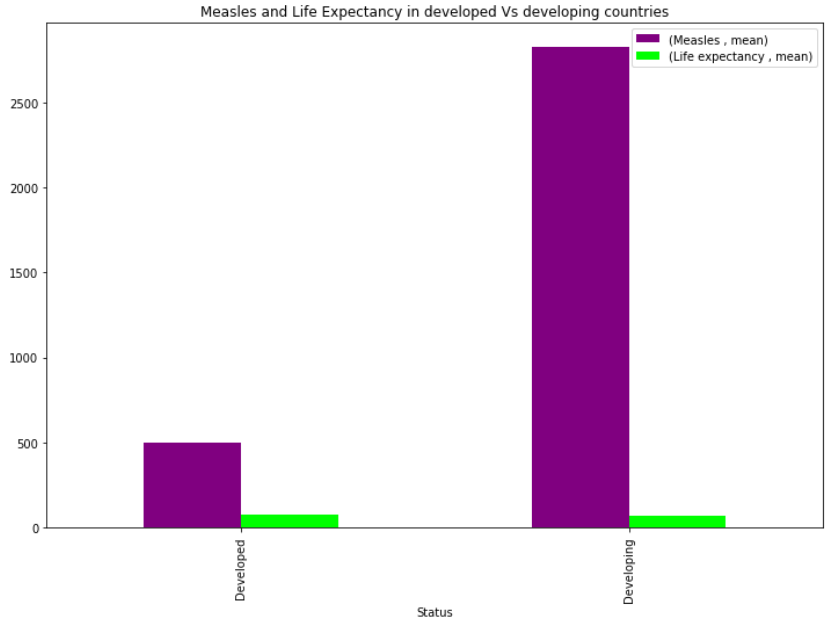
Europe has the highest immunization for Diphtheria and a high life expectancy.



The mean value of Measles for Asia is the highest and for North America is the lowest.

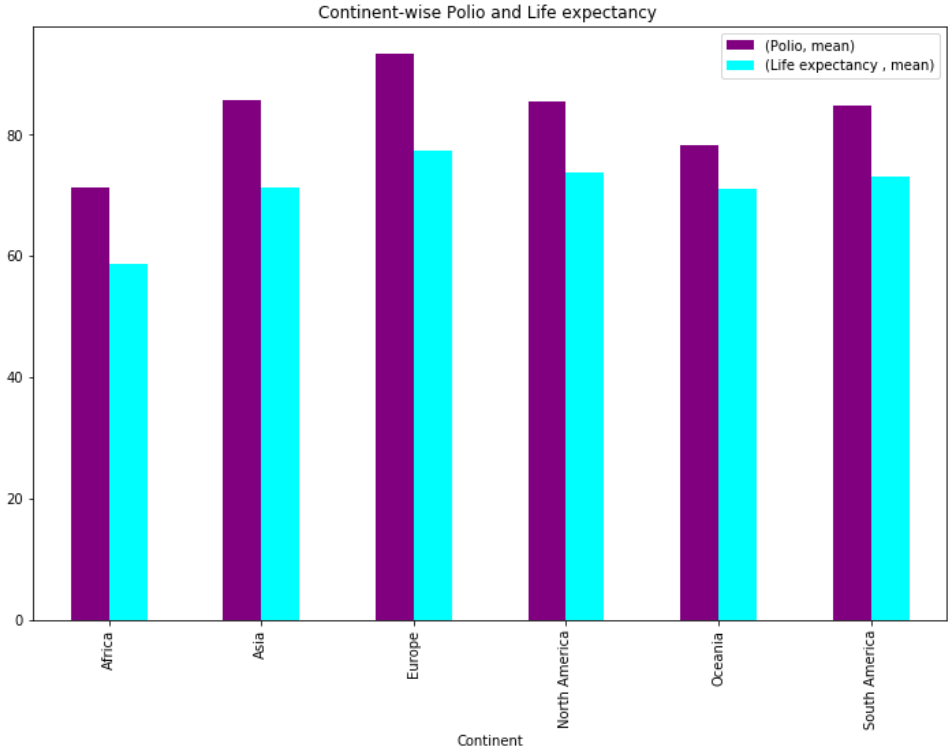
However, the life expectancy in Europe and North America is the highest, and it's roughly the same for Asia and Africa.

It shows that there is not a very strong relationship between Measles and Life Expectancy when compared continent-wise.



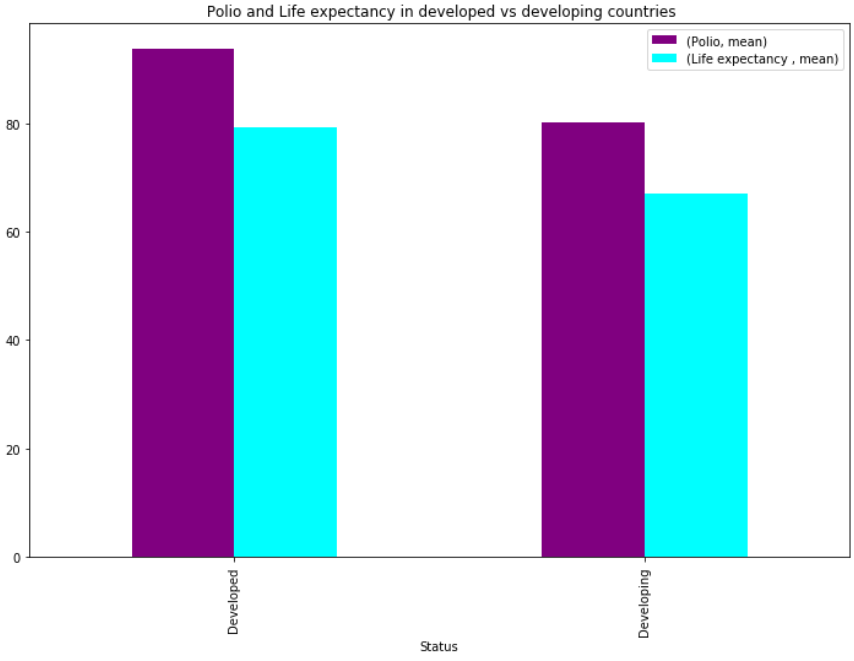
Mean value of Measles for developing countries is quite higher than that of developed countries and thus, its life expectancy is considerably lower.

This inverse relation can also be accurately verified by our heatmap which suggests a negative correlation between measles and life expectancy.



Mean Polio (immunization) for Europe is the highest and so is its life expectancy. Mean Polio for Oceania is the lowest but the life expectancy of Africa is the lowest.

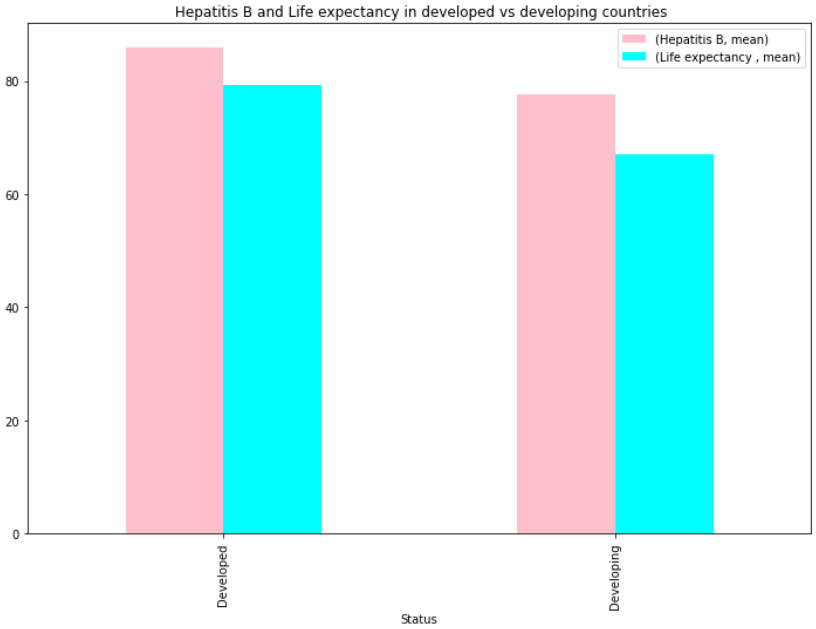
The reason for the lowest value of Polio for Oceania could be its low population.



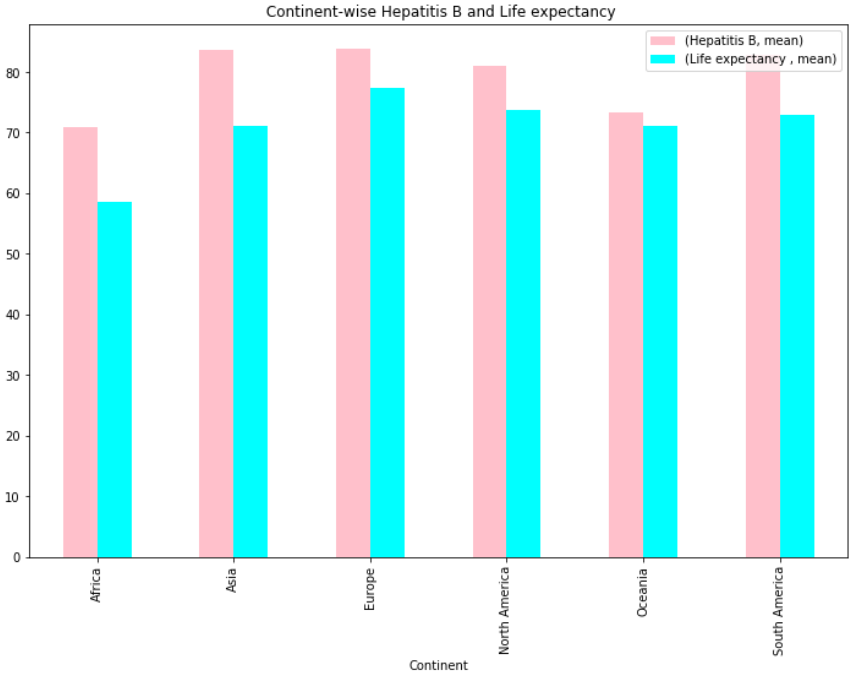
Mean Polio (immunization) for developed countries is higher than that of developing countries and so is its life expectancy.

We see that as polio (which is an immunization) increases, Adult Mortality decreases considerably.

This shows that immunization is directly related to Life Expectancy.



Mean Hepatitis B (immunization) for developed countries is higher than that of developing countries and so is its life expectancy.



Mean Hepatitis B (immunization) for Europe is the highest and so is its life expectancy. Mean Hepatitis B for Oceania is the lowest but the life expectancy of Africa is the lowest.

The reason for the lowest value of Hepatitis B for Oceania could be its low population

VISUALIZING THE LIFE EXPECTANCY CONTINENT WISE

In [21]:
```
import geopandas as gpd
import shapefile as shp
import plotly.express as px
```

In [6]:
```
world = gpd.read_file(gpd.datasets.get_path('naturalearth_lowres'))
```

In [22]:
```
world.head()
```

Out[22]:

	pop_est	continent	name	iso_a3	gdp_md_est	geometry
0	920938	Oceania	Fiji	FJI	8374.0	MULTIPOLYGON (((180.00000 -16.06713, 180.00000...
1	53950935	Africa	Tanzania	TZA	150600.0	POLYGON ((33.90371 -0.95000, 34.07262 -1.05982...
2	603253	Africa	W. Sahara	ESH	906.5	POLYGON ((-8.66559 27.65643, -8.66512 27.58948...
3	35623680	North America	Canada	CAN	1674000.0	MULTIPOLYGON (((-122.84000 49.00000, -122.9742...
4	326625791	North America	United States of America	USA	18560000.0	MULTIPOLYGON (((-122.84000 49.00000, -120.0000...

In [16]:
```
country=world.merge(data, how="left", left_index=True, right_index=True)
country.head()
```

Out[16]:

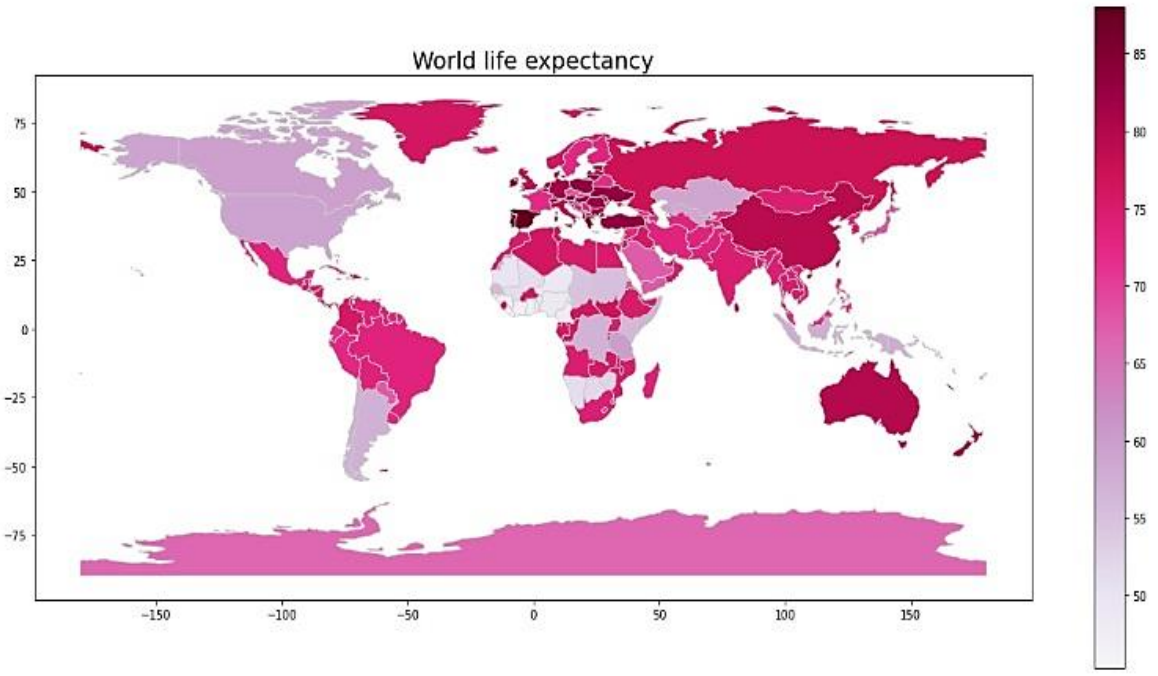
	pop_est	continent	name	iso_a3	gdp_md_est	geometry	Country	Continent	Year	Status	...	Polio	Total expenditure	Diphtheria	HIV/AIDS
0	920938	Oceania	Fiji	FJI	8374.0	MULTIPOLYGON (((180.00000 -16.06713, 180.00000...	Afghanistan	Asia	2015	Developing	...	6.0	8.16	65.0	0.1
1	53950935	Africa	Tanzania	TZA	150600.0	POLYGON ((33.90371 -0.95000, 34.07262 -1.05982...	Afghanistan	Asia	2014	Developing	...	58.0	8.18	62.0	0.1
2	603253	Africa	W. Sahara	ESH	906.5	POLYGON ((-8.66559 27.65643, -8.66512 27.58948...	Afghanistan	Asia	2013	Developing	...	62.0	8.13	64.0	0.1
3	35623680	North America	Canada	CAN	1674000.0	MULTIPOLYGON (((-122.84000 49.00000, -122.9742...	Afghanistan	Asia	2012	Developing	...	67.0	8.52	67.0	0.1
4	326625791	North America	United States of America	USA	18560000.0	MULTIPOLYGON (((-122.84000 49.00000, -120.0000...	Afghanistan	Asia	2011	Developing	...	68.0	7.87	68.0	0.1

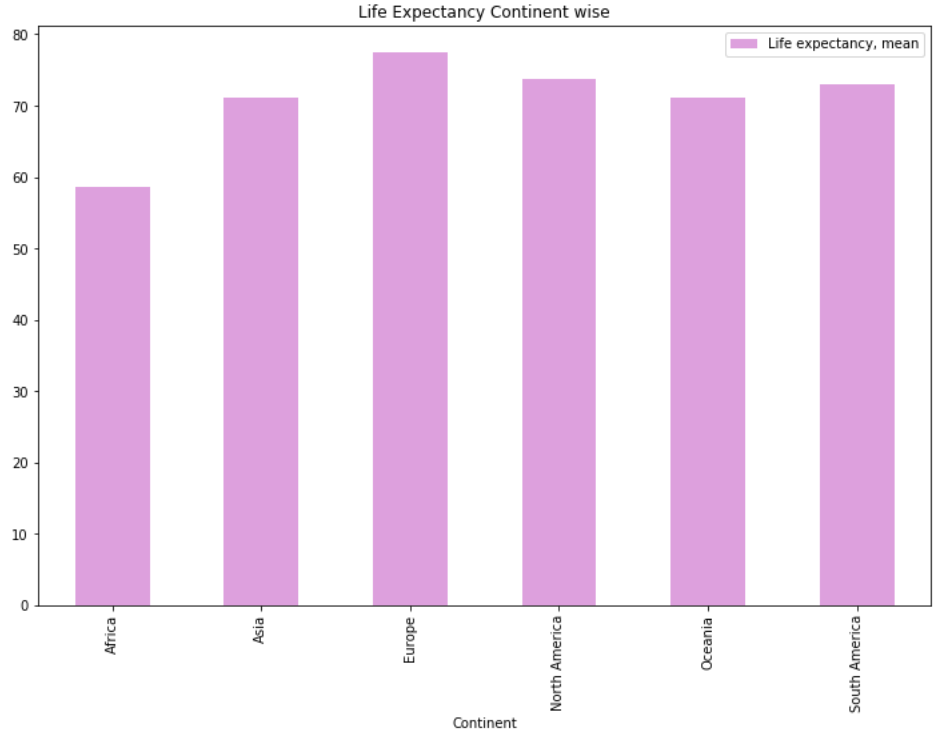
5 rows × 29 columns

In [20]:
```
#create figure and axis for the matplotlib
fig, ax=plt.subplots(1, figsize=(20,10))
ax.axis("off")
ax.set_title('World life expectancy', fontdict={'fontsize': '20', 'fontweight' : '3'})

country.plot(column="Life expectancy ", cmap="PuRd", linewidth=0.8, ax=ax, edgecolor="0.8", legend=True)
```

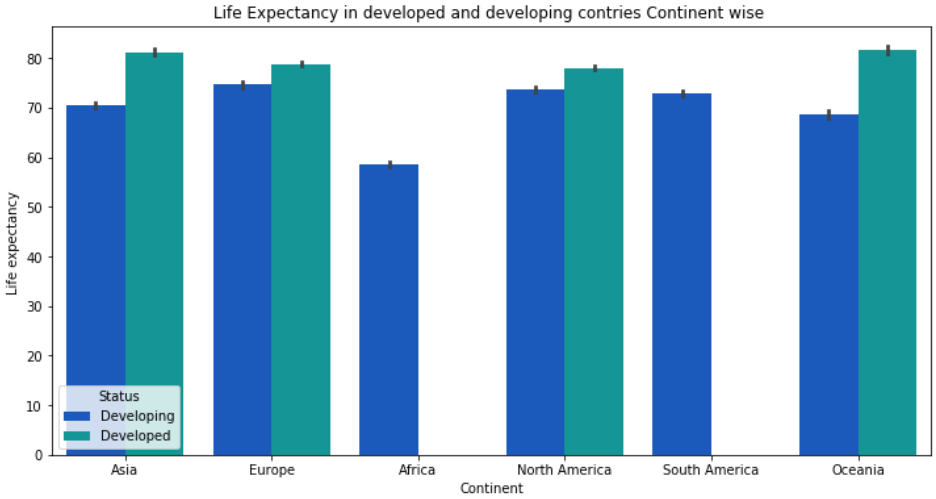
Out[20]: <AxesSubplot:title={'center':'World life expectancy'}>



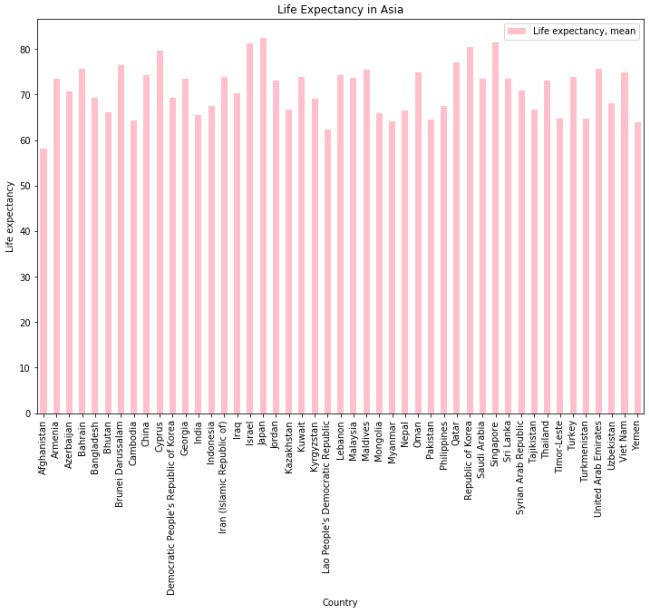


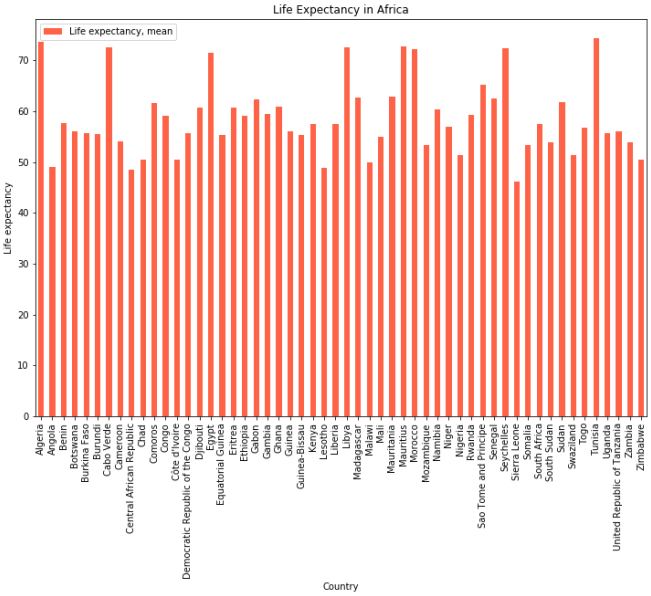
The average life expectancy is approximately ranged 70% - 80% with the highest in Europe followed by America and Asia.

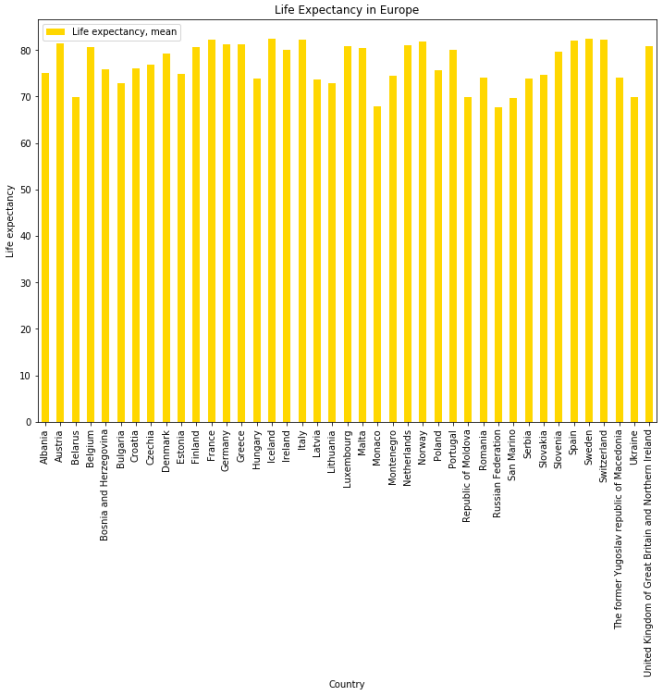
The least is in Africa due to the presence of higher developing countries.

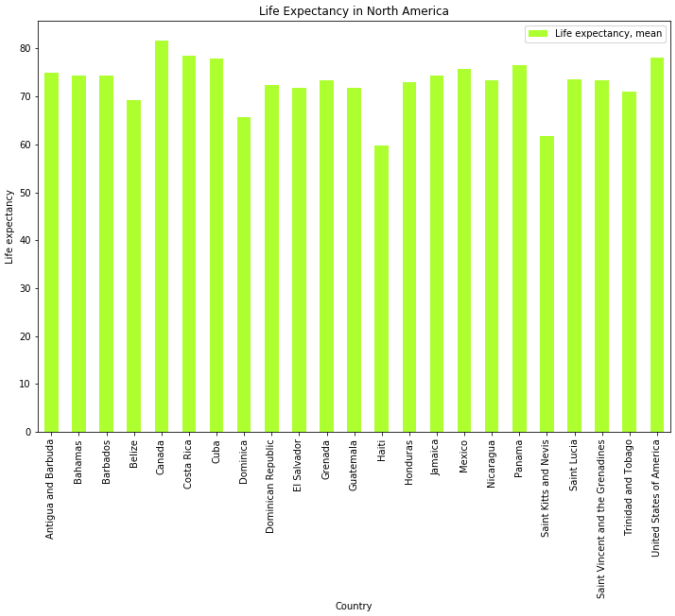


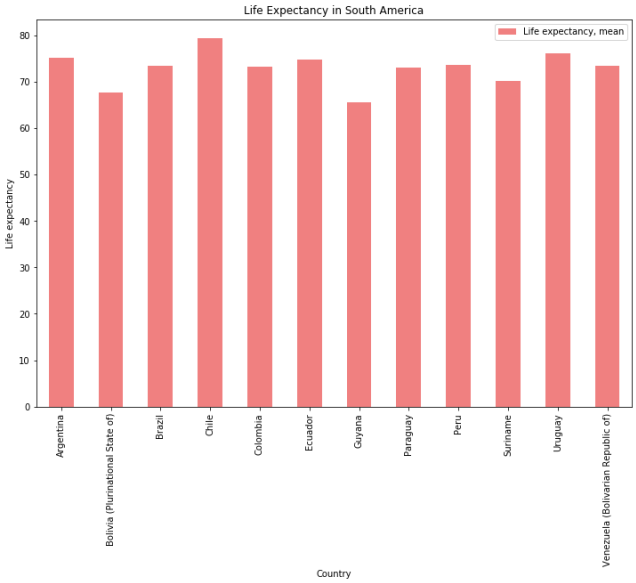
Africa and South America have only developing countries and thus account for the overall life expectancy rate. Developed countries have higher life expectancy than developing countries on average.

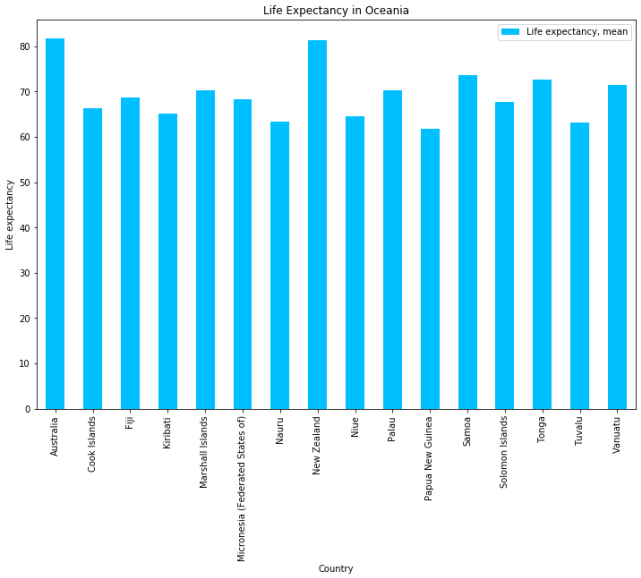

Life Expectancy in Asia
Life expectancy, mean
Life expectancy
Country
Afghanistan
Armenia
Azerbaijan
Bahrain
Bangladesh
Bhutan
Brunei Darussalam
Cambodia
China
Cyprus
Democratic People's Republic of Korea
Georgia
India
Indonesia
Iran (Islamic Republic of)
Iraq
Israel
Japan
Jordan
Kazakhstan
Kuwait
Kyrgyzstan
Lao People's Democratic Republic
Lebanon
Malaysia
Maldives
Mongolia
Myanmar
Nepal
Oman
Pakistan
Philippines
Qatar
Republic of Korea
Saudi Arabia
Singapore
Sri Lanka
Syrian Arab Republic
Tajikistan
Thailand
Timor-Leste
Turkey
Turkmenistan
United Arab Emirates
Uzbekistan
Viet Nam
Yemen


Life Expectancy in Africa
Life expectancy, mean
Life expectancy
Country
Algeria
Angola
Benin
Botswana
Burkina Faso
Burundi
Cabo Verde
Cameroon
Central African Republic
Chad
Comoros
Congo
Côte d'Ivoire
Democratic Republic of the Congo
Djibouti
Egypt
Equatorial Guinea
Eritrea
Ethiopia
Gabon
Gambia
Ghana
Guinea
Guinea-Bissau
Kenya
Lesotho
Liberia
Libya
Madagascar
Malawi
Mali
Mauritania
Mauritius
Morocco
Mozambique
Namibia
Niger
Nigeria
Rwanda
Sao Tome and Principe
Senegal
Seychelles
Sierra Leone
Somalia
South Africa
South Sudan
Sudan
Swaziland
Togo
Tunisia
Uganda
United Republic of Tanzania
Zambia
Zimbabwe


Life Expectancy in Europe
Life expectancy, mean
Life expectancy
Country
Albania
Austria
Belarus
Belgium
Bosnia and Herzegovina
Bulgaria
Croatia
Czechia
Denmark
Estonia
Finland
France
Germany
Greece
Hungary
Iceland
Ireland
Italy
Latvia
Lithuania
Luxembourg
Malta
Monaco
Montenegro
Netherlands
Norway
Poland
Portugal
Republic of Moldova
Romania
Russian Federation
San Marino
Serbia
Slovakia
Slovenia
Spain
Sweden
Switzerland
The former Yugoslav republic of Macedonia
Ukraine
United Kingdom of Great Britain and Northern Ireland


Life Expectancy in North America
Life expectancy, mean
Life expectancy
Country
Antigua and Barbuda
Bahamas
Barbados
Belize
Canada
Costa Rica
Cuba
Dominica
Dominican Republic
El Salvador
Grenada
Guatemala
Haiti
Honduras
Jamaica
Mexico
Nicaragua
Panama
Saint Kitts and Nevis
Saint Lucia
Saint Vincent and the Grenadines
Trinidad and Tobago
United States of America

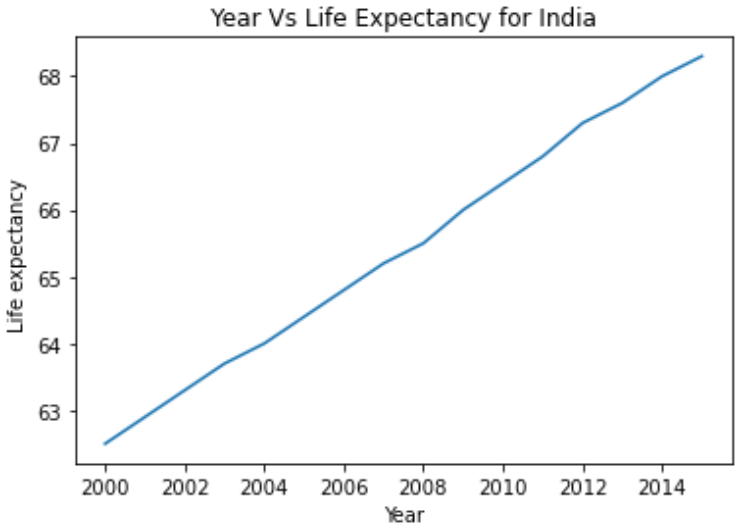

Life Expectancy in South America
Life expectancy, mean
Life expectancy
Country
Argentina
Bolivia (Plurinational State of)
Brazil
Chile
Colombia
Ecuador
Guyana
Paraguay
Peru
Suriname
Uruguay
Venezuela (Bolivarian Republic of)


Life Expectancy in Oceania
Life expectancy, mean
Life expectancy
Country
Australia
Cook Islands
Fiji
Kiribati
Marshall Islands
Micronesia (Federated States of)
Nauru
New Zealand
Niue
Palau
Papua New Guinea
Samoa
Solomon Islands
Tonga
Tuvalu
Vanuatu

Insights from the above graphs

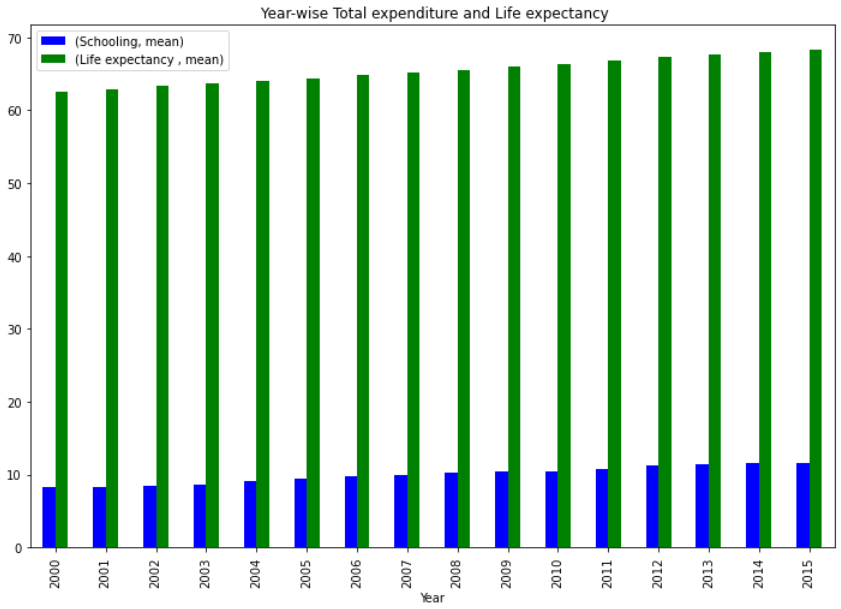
- *Japan and Singapore have the highest life expectancy in Asia.*
- *Afghanistan has the lowest life expectancy in Asia.*
- *Tunisia and Algeria have the highest life expectancy in Africa.*
- *Sierra Leone has the lowest life expectancy in Africa.*
- *Iceland and Sweden have the highest life expectancy in Europe.*
- *Russian Federation has the lowest life expectancy in Europe.*
- *Canada has the highest life expectancy in North America.*
- *Haiti has the lowest life expectancy in North America.*
- *Chile has the highest life expectancy in South America.*
- *Guyana has the lowest life expectancy in South America.*
- *Australia has the highest life expectancy in Oceania, followed by New Zealand.*
- *Papua New Guinea has the lowest life expectancy in Oceania.*

INDIA

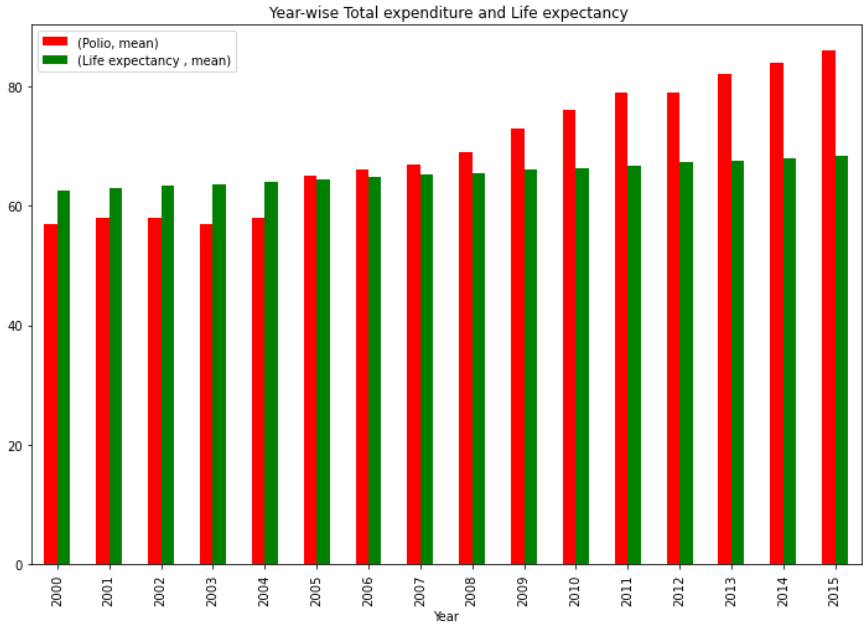


We see that over the years life expectancy has increased in India. Though, after 2015, it saw a stunted growth in life expectancy.

Adult Mortality has decreased over the years with an increase in life expectancy. Though, we see that adult mortality for the years 2006,2007,2008 is quite low which could be an anomaly.



Schooling has a direct effect on life expectancy as life expectancy increases in consistence to schooling.



As immunization for Polio has increased over the course of 15 years, life expectancy shows improvement, though it is not that considerable.

LIMITATIONS

Continent Antarctica is not included in this study, due to the unavailability of data. But it won't be a major issue because there is a very less constant population there.

The primary dataset used in this project has many missing values and anomalies, which are mostly taken care of during data cleaning. But some factors/ countries have some major missing values which were hard to impute and can cause bias in the conclusions.

Since the data is huge and many factors were involved, the analysis of the factors in the time frame from 2000 to 2015 could be only carried out.

CONCLUSION

- The life expectancy in developed countries is more than in developing countries.
- The peak of the population ranges from 60-to 70 years of life expectancy. The population has almost no correlation with life expectancy with a value of -0.0016.
- The Schooling period is higher in developed countries where the life expectancy takes only values above 70.
- The income compositions of resources have a linear trend with life expectancy, also as income composition increases, Adult Mortality decreases considerably.
- The life expectancy of developed countries is higher regardless of the expenditure amount, Whereas the life expectancy in developing countries
- Africa the has highest adult mortality, the lowest immunization for Diphtheria, and the lowest life expectancy.
- Europe has the lowest adult mortality, the highest mean Polio (immunization), the highest immunization for Diphtheria, and the highest life expectancy.
- As polio (which is an immunization) increases, Adult Mortality decreases considerably. This shows that immunization is directly related to Life Expectancy.
- Also, there is a negative correlation between measles and life expectancy.
- Mean Hepatitis B (immunization) for developed countries is higher than that of developing countries and so is its life expectancy.

REFERENCES

- https://www.kaggle.com/kumarajarshi/life-expectancy-who
- https://apps.who.int/gho/data/node.main.686?lang=en
- https://apps.who.int/iris/bitstream/handle/10665/112738/9789240692671_eng.pdf
- https://jovian.ai/zelalemgetahun9374/project-life-expectancy-exploratory-data-analysis
- https://www.kaggle.com/code/aishwaryakshirsagar/96-r2-score-and-eda
- https://www.ibm.com/in-en/cloud/learn/exploratory-data-analysis
- https://en.wikipedia.org/wiki/Exploratory_data_analysis
- https://pestleanalysis.com/outlier-analysis/