PREDICTING LIFE EXPECTANCY

Shubradip Ghosh

SKILLVERTEX Data science Jan_batch

Many studies on the determinants influencing a nation's life expectancy have been conducted in the past, accounting for demographic characteristics, income distribution, and death rates. It was discovered that the human development index and the impact of vaccination had not previously been taken into consideration. I'll expose you to a Python-based data science project on life expectancy.

Introduction

According to the statistical average, a person's life expectancy is the number of years they can anticipate to live. It depends on the region's geographical setting. The average life expectancy in the world before modernisation was about 30 years. The beginning of the 19th century saw an improvement in life expectancy, but only in certain countries, while it remained low in the rest of the world.

This shows that health standards are not the same all over the world. In the 20th century, this global inequality is reduced and similarly, life expectancy is approaching 70 to 75 years and similarly no country in the world today has a low life expectancy than countries with high life expectancy in 1800.

Overview

Previously there were many studies on linear regression model to predict life expectancy however in most of them, affect of immunization and human development index was not taken into account. This project aims to build a regression model to predict life expectancy and investigate which factors affect life expectancy in the world with data of immunisation and human development included. A country can use this model to predict life expectancy and determine the predicting factor which is contributing to lower value of life expectancy. This will help in suggesting a country which area should be given importance in order to efficiently improve the life expectancy of its population.This model will be devrived from WHO life expectancy with data from every countries between 2000-2015 using statsmodel library in Python. From our model we found 9 significant factors contributing to life expectancy and the model is able to predict life expectancy to high level of accuracy.

Business Problem

Life Expectancy is affected by various factors. WHO wishes to predict life expectancy and determine which factors has significant impact. From this project, WHO would be able to give a country its life expectancy and suggestions on which factor to focus on to improve their life expectancy.

Questions to consider:

Does various predicting factors which has been chosen initially really affect the Life expectancy? What are the predicting variables actually affecting the life expectancy What is the impact of Immunization coverage on life Expectancy? Do densely populated countries tend to have lower life expectancy? What is the impact of schooling on the lifespan of humans?

Abstract

Based on a publicly available WHO dataset, this research analyses the variables that affect life expectancy. Data was gathered during the years of 2000 and 2015. A country's development status (developed versus developing), GDP, population, schooling years, alcohol use, BMI, government health spending, health spending per unit of GDP, various immunisation coverage, thinness disease, measles cases, HIV/AIDS deaths, and the mortality rate of adults, children, and infants were among the factors that were examined.

Data were carefully examined (horizontally and vertically), cleansed, and changed during the processing stage. Using the Bagged-trees technique, missing values were imputed. Box and whisker plots, histograms, and multiple factor analyses (MFA) were used in exploratory data analysis (EDA) to explore and mine the trends within the data. MFA is a method of unsupervised machine learning.

The results of a multiple linear regression model that passed the assumption tests indicated that education (Coeff. Est: 1.15), total government health spending (Coeff. Est: 0.08), BMI (0.03), GDP (Coeff. Est: 0.00004), and diphtheria and polio vaccinations (Coeff. Est: 0.03) and polio vaccinations (Coeff. Est: 0.02) vaccinations are positively correlated significant variables (p Similar findings are supported by a partial study of the longitudinal multilevel modeling's entire model, which also found that if a country had a lower beginning life expectancy in the year 2000. Between 2000 and 2015, it would have a faster rate of life expectancy improvement (intercept-slope corr = -0.55). Between 2000 to 2015, it suggests an improvement in life expectancy in developing nations that were linked to lower life expectancy. The whole model also comes to the conclusion that the average human life expectancy rises year by 0.25 years, with a degree of confidence ranging from 0.16 years to 0.34 years (p 0.001).

The factors that WHO uses to calculate the Life Expectancy of a country as the data is provided by WHO

No.	Variable	Description
1	Country	Country
2	Year	Year
3	Status	Developed or Developing status
4	Life expectancy	Life Expectancy in age
5	Adult Mortality	Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)
6	infant deaths	Number of Infant Deaths per 1000 population
7	Alcohol	Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)
8	percentage expenditure	Expenditure on health as a percentage of Gross Domestic Product per capita(%)
9	Hepatitis B	Hepatitis B (HepB) immunization coverage among 1-year-olds (%)
10	Measles	Measles - number of reported cases per 1000 population
11	BMI	Average Body Mass Index of entire population
12	under-five deaths	Number of under-five deaths per 1000 population
13	Polio	Polio (Pol3) immunization coverage among 1-year-olds (%)
14	Total expenditure	General government expenditure on health as a percentage of total government expenditure (%)
15	Diphtheria	Diphtheria tetanus toxoid and pertussis

		(DTP3) immunization coverage among 1-year-olds (%)
16	HIV/AIDS	Deaths per 1 000 live births HIV/AIDS (0-4 years)
17	GDP	Gross Domestic Product per capita (in USD)
18	Population	Population of the country
19	thinness 10-19 years	Prevalence of thinness among children and adolescents for Age 10 to 19 (%)
20	thinness 5-9 years	Prevalence of thinness among children for Age 5 to 9(%)
21	Income composition of resources	Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
22	Schooling	Number of years of Schooling(years)

Data summary

Name life

- Number of rows 2938
- Number of columns 22

Column type frequency:

- character 2
- numeric 20

—

- Group variables None

	Country	Status
count	2938	2938
unique	193	2
top	Afghanistan	Developing
freq	16	2426

If we just look at the numeric columns central tendencies:

	count	mean	std	min	25%	50%	
Year	2938.0	2.007519e+03	4.613841e+00	2000.00000	2004.000000	2.008000e+03	2.012000
Life expectancy	2928.0	6.922493e+01	9.523867e+00	36.30000	63.100000	7.210000e+01	7.570000
Adult Mortality	2928.0	1.647964e+02	1.242921e+02	1.00000	74.000000	1.440000e+02	2.280000
infant deaths	2938.0	3.030395e+01	1.179265e+02	0.00000	0.000000	3.000000e+00	2.200000
Alcohol	2744.0	4.602861e+00	4.052413e+00	0.01000	0.877500	3.755000e+00	7.702500
percentage expenditure	2938.0	7.382513e+02	1.987915e+03	0.00000	4.685343	6.491291e+01	4.415341
Hepatitis B	2385.0	8.094046e+01	2.507002e+01	1.00000	77.000000	9.200000e+01	9.700000
Measles	2938.0	2.419592e+03	1.146727e+04	0.00000	0.000000	1.700000e+01	3.602500
BMI	2904.0	3.832125e+01	2.004403e+01	1.00000	19.300000	4.350000e+01	5.620000
under-five deaths	2938.0	4.203574e+01	1.604455e+02	0.00000	0.000000	4.000000e+00	2.800000
Polio	2919.0	8.255019e+01	2.342805e+01	3.00000	78.000000	9.300000e+01	9.700000
Total expenditure	2712.0	5.938190e+00	2.498320e+00	0.37000	4.260000	5.755000e+00	7.492500
Diphtheria	2919.0	8.232408e+01	2.371691e+01	2.00000	78.000000	9.300000e+01	9.700000
HIV/AIDS	2938.0	1.742103e+00	5.077785e+00	0.10000	0.100000	1.000000e-01	8.000000
GDP	2490.0	7.483158e+03	1.427017e+04	1.68135	463.935626	1.766948e+03	5.910806
Population	2286.0	1.275338e+07	6.101210e+07	34.00000	195793.250000	1.386542e+06	7.420359
thinness 1-19 years	2904.0	4.839704e+00	4.420195e+00	0.10000	1.600000	3.300000e+00	7.200000
thinness 5-9 years	2904.0	4.870317e+00	4.508882e+00	0.10000	1.500000	3.300000e+00	7.200000
Income composition of resources	2771.0	6.275511e-01	2.109036e-01	0.00000	0.493000	6.770000e-01	7.790000
Schooling	2775.0	1.199279e+01	3.358920e+00	0.00000	10.100000	1.230000e+01	1.430000

Data type and structure

There are 2938 rows of observations and 22 columns of variables

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2938 entries, 0 to 2937
Data columns (total 22 columns):
 #   Column                          Non-Null Count  Dtype  
---  ------                          --------------  -----  
 0   Country                         2938 non-null   object 
 1   Year                            2938 non-null   int64  
 2   Status                          2938 non-null   object 
 3   Life expectancy                 2928 non-null   float64
 4   Adult Mortality                 2928 non-null   float64
 5   infant deaths                   2938 non-null   int64  
 6   Alcohol                         2744 non-null   float64
 7   percentage expenditure          2938 non-null   float64
 8   Hepatitis B                     2385 non-null   float64
 9   Measles                         2938 non-null   int64  
 10   BMI                            2904 non-null   float64
 11  under-five deaths               2938 non-null   int64  
 12  Polio                           2919 non-null   float64
 13  Total expenditure               2712 non-null   float64
 14  Diphtheria                      2919 non-null   float64
 15   HIV/AIDS                       2938 non-null   float64
 16  GDP                             2490 non-null   float64
 17  Population                      2286 non-null   float64
 18   thinness  1-19 years           2904 non-null   float64
 19   thinness 5-9 years             2904 non-null   float64
 20  Income composition of resources 2771 non-null   float64
 21  Schooling                       2775 non-null   float64
dtypes: float64(16), int64(4), object(2)
memory usage: 505.1+ KB
```

Data Pre-processing:

Vertical NA Check (Column)

There are 2 character variables “Country” and “Status”, and the rest of the variables are numerical.

This situation becomes difficult to take the proceedings straight into machine learning. Before we feed the data, we would like to convert the categorical feature in to numeric by one-hot-encoding method and then we will pass the data into algorithm for easy computation.

Since there are 158 countries, which is very high and often difficult in computation.

So, we chose to drop the country column from the dataset after performing the EDA.

Variable type: character

variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
Country	0	1	4	52	0	193	0
Status	0	1	9	10	0	2	0

We discovered a significant amount of missing data in numerous variables from the variables "n missing" and "complete rate".

Moreover, the mean and standard deviation are calculated.

The fact that no variable has more than 40% missing data is advantageous because I will follow the 60% rule and consider deleting a variable if its column has a completion rate below 0.6 (i.e., has 0.4 or 40% missing data).

So to take care of those missing values we chose to drop the columns which has more than 40% of missing values

for better accuracy of the models.

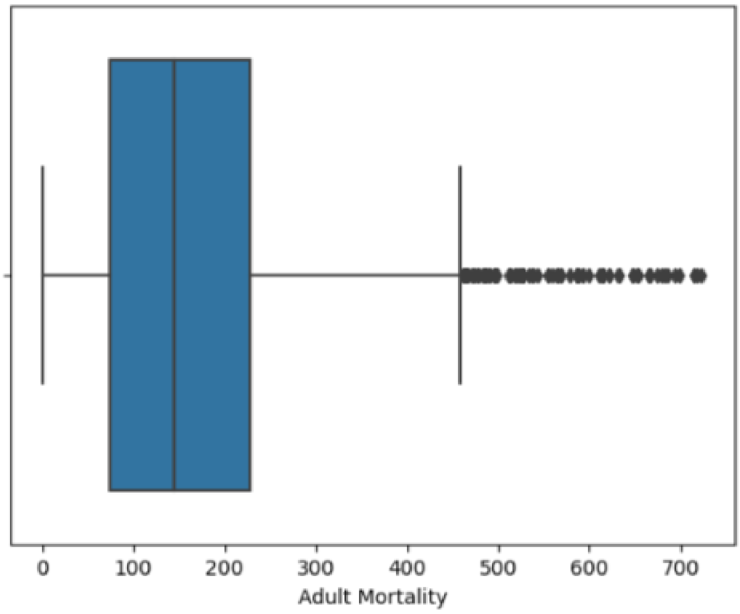
Performed one hot encoding:

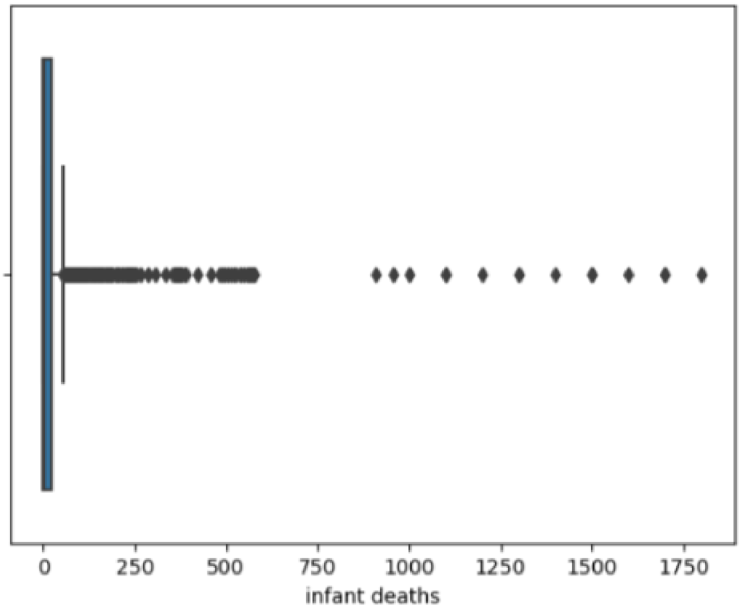
Column status: we have labelled '0' for developing & '1' for developed countries.

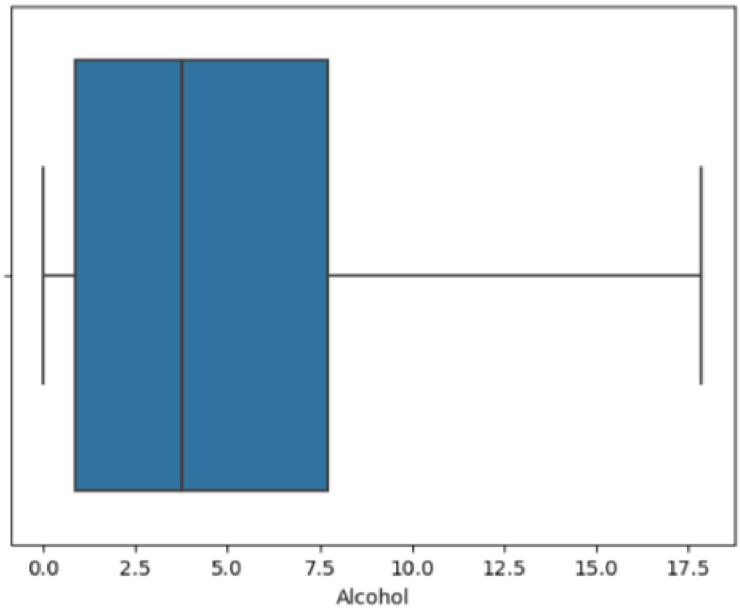
Then we have divided the dataset in to depended variable:Y: 'Life _Expectancy' & all other independent variables into X dataset.

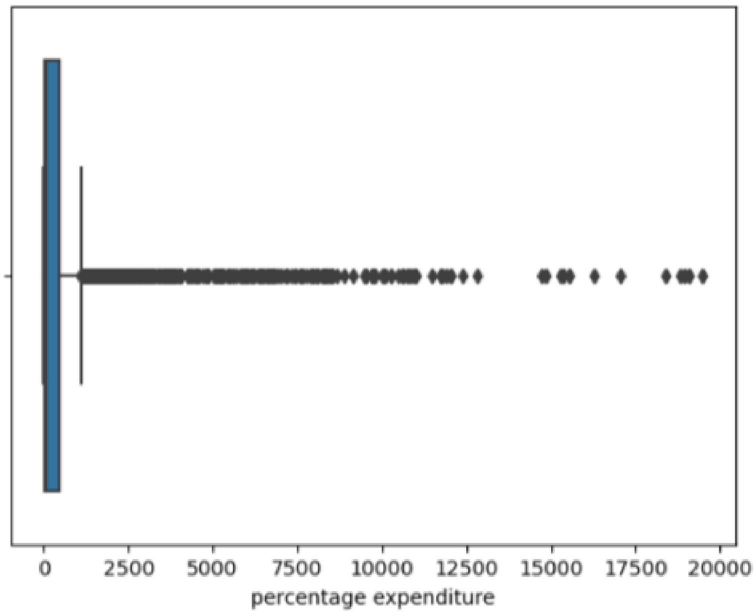
We have imputed the missing values of Y with one of the suitable central tendencies i.e., mean here and X with median values, so that the models doesn't get effected by their imputation.

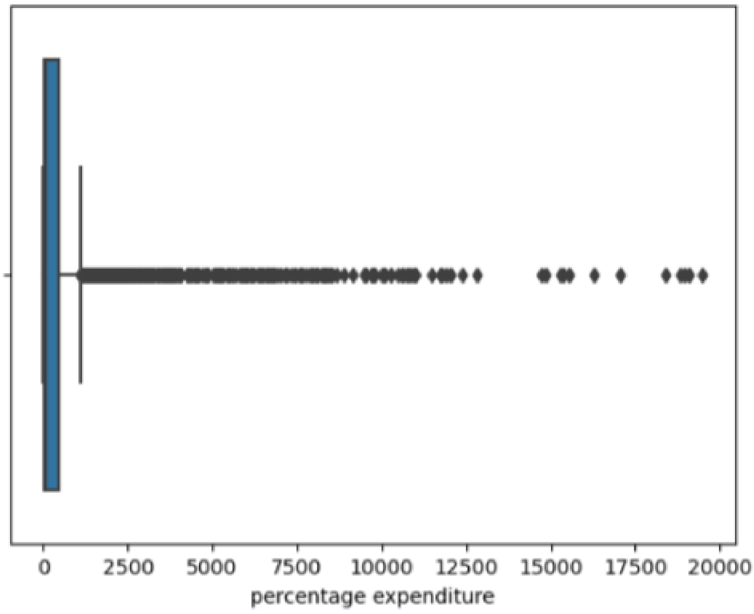
Exploratory Data Analysis:

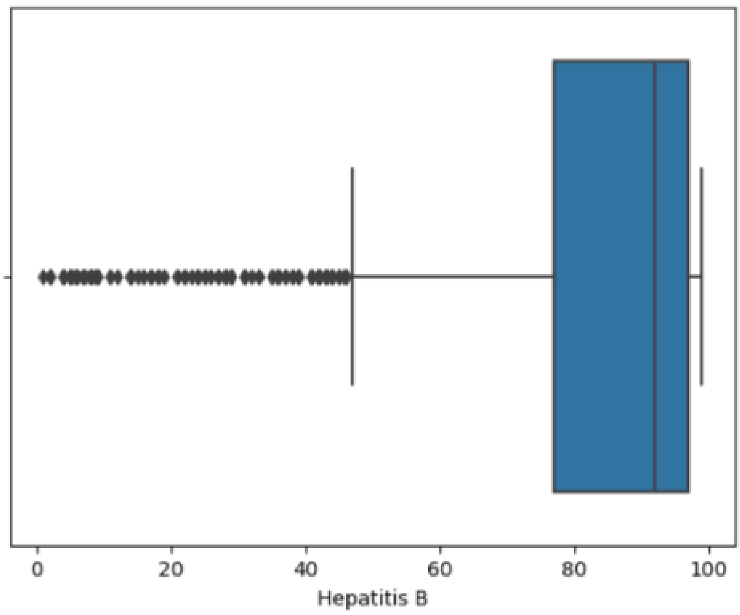

0
100
200
300
400
500
600
700
Adult Mortality

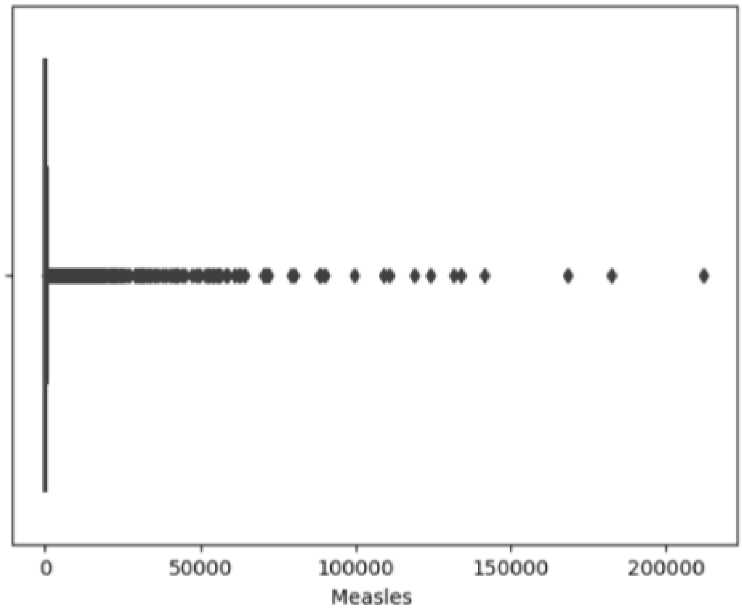

0
250
500
750
1000
1250
1500
1750
infant deaths

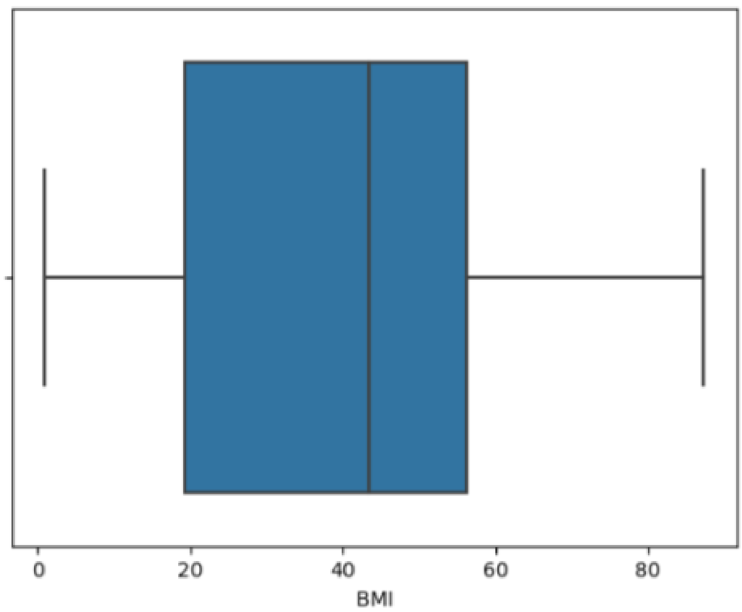

0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Alcohol

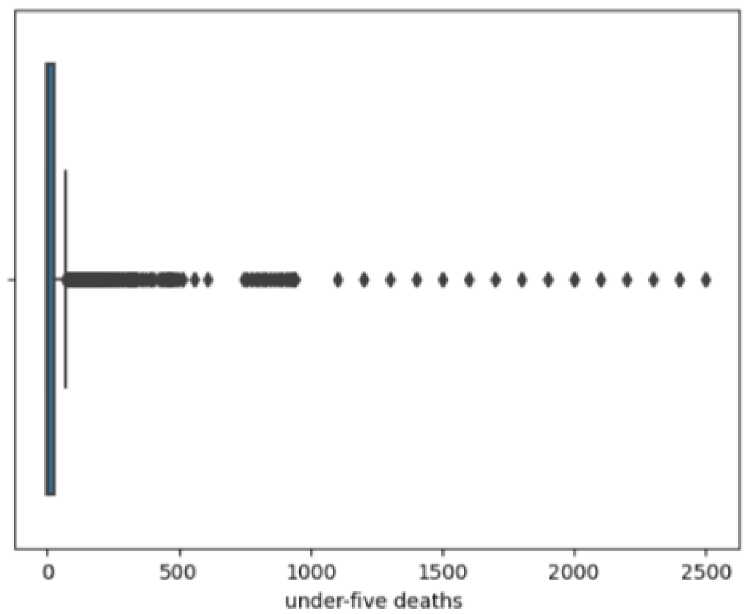

0
2500
5000
7500
10000
12500
15000
17500
20000
percentage expenditure

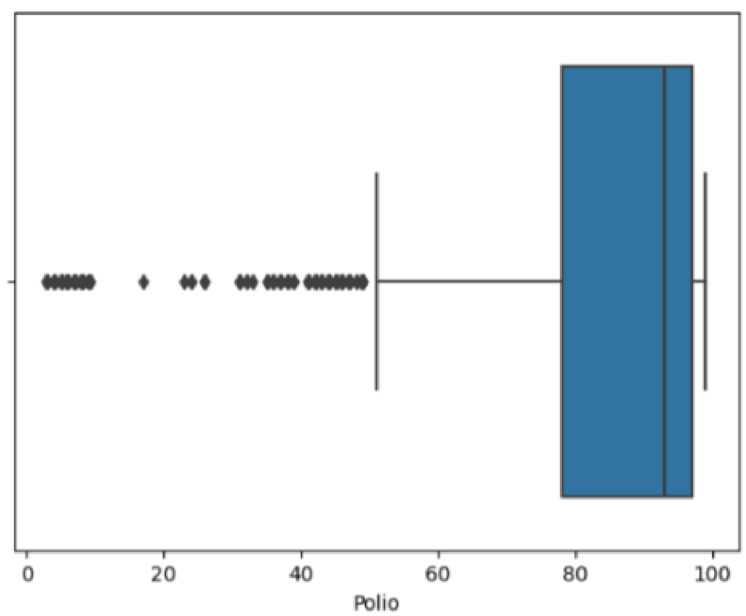

0
2500
5000
7500
10000
12500
15000
17500
20000
percentage expenditure

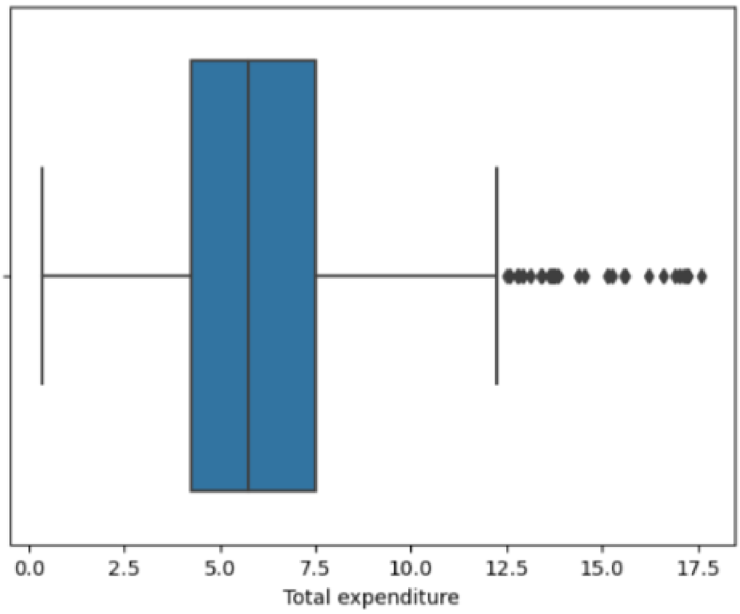

0
20
40
60
80
100
Hepatitis B

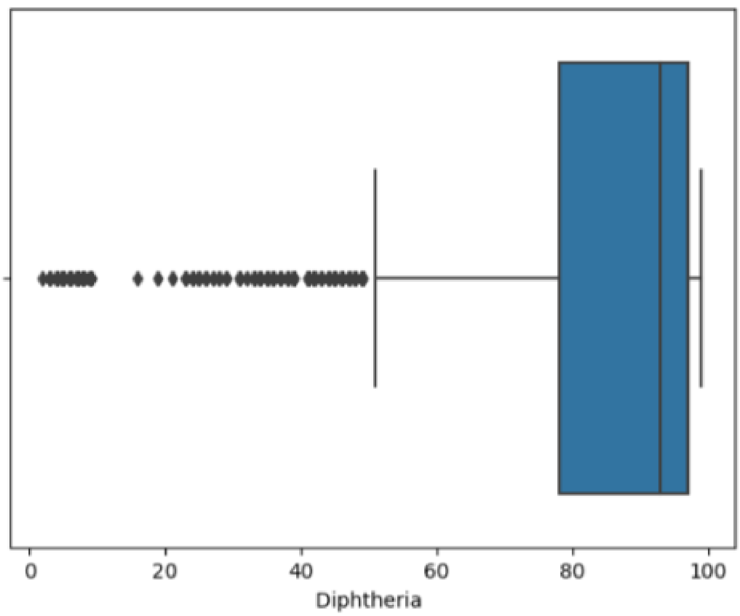

0
50000
100000
150000
200000
Measles

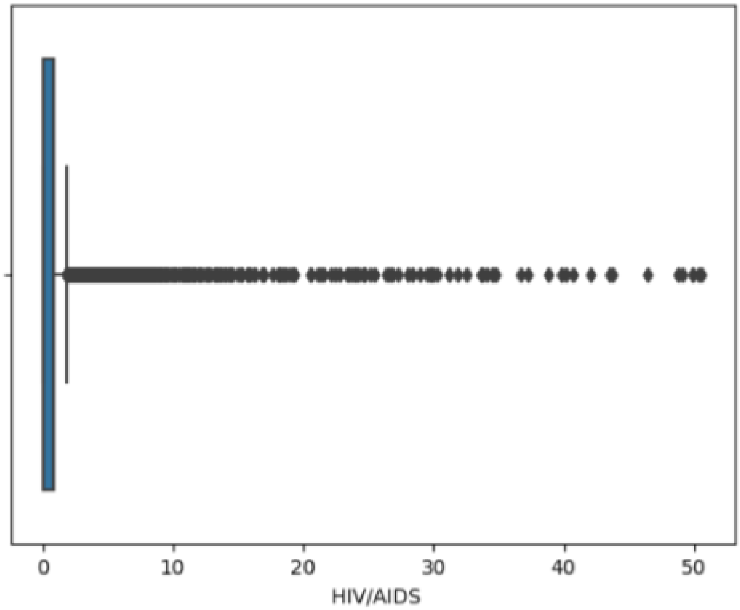

0
20
40
60
80
BMI

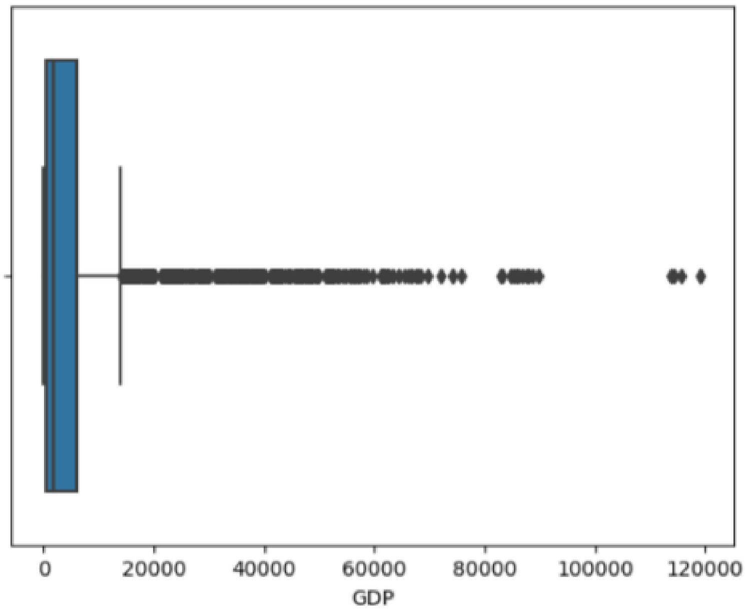

0
500
1000
1500
2000
2500
under-five deaths

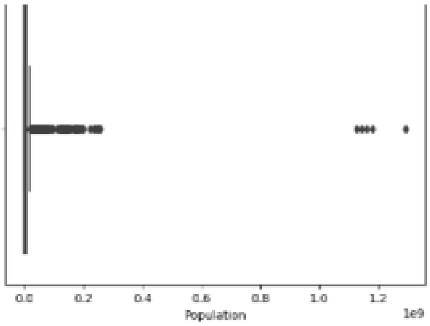

0
20
40
60
80
100
Polio

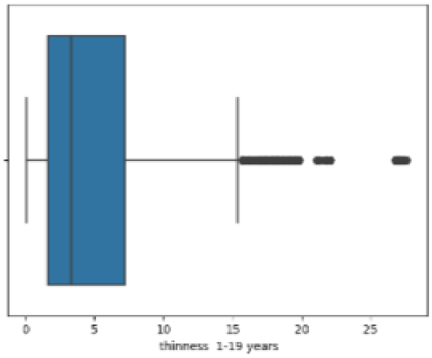

0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Total expenditure

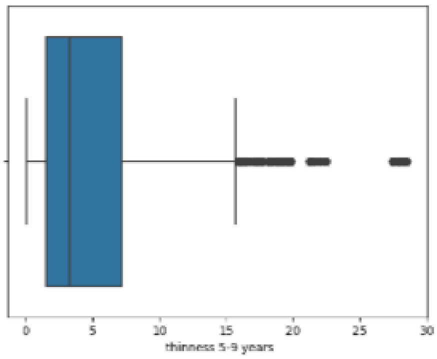

0
20
40
60
80
100
Diphtheria

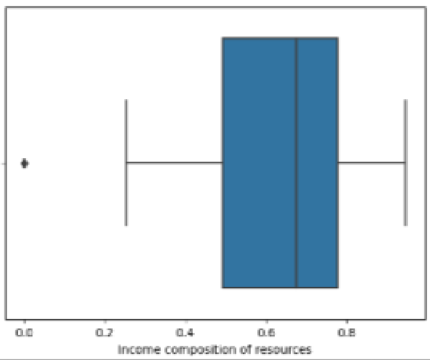

0
10
20
30
40
50
HIV/AIDS

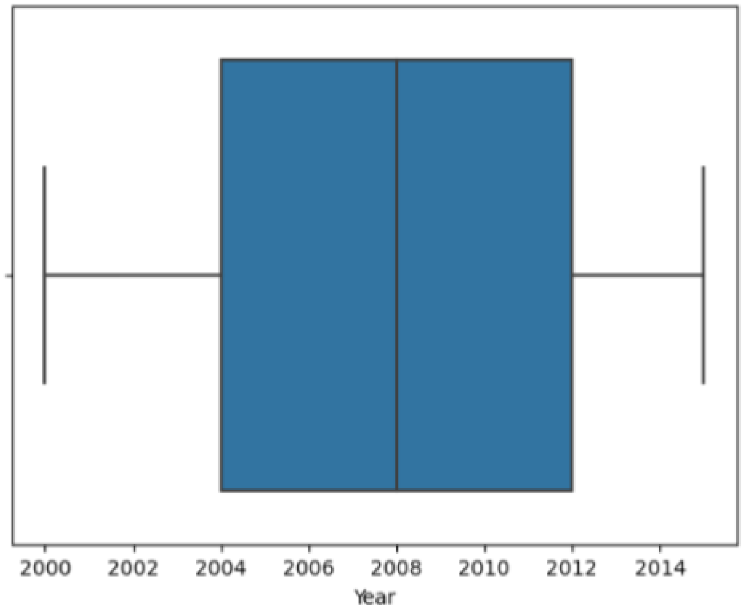

0
20000
40000
60000
80000
100000
120000
GDP

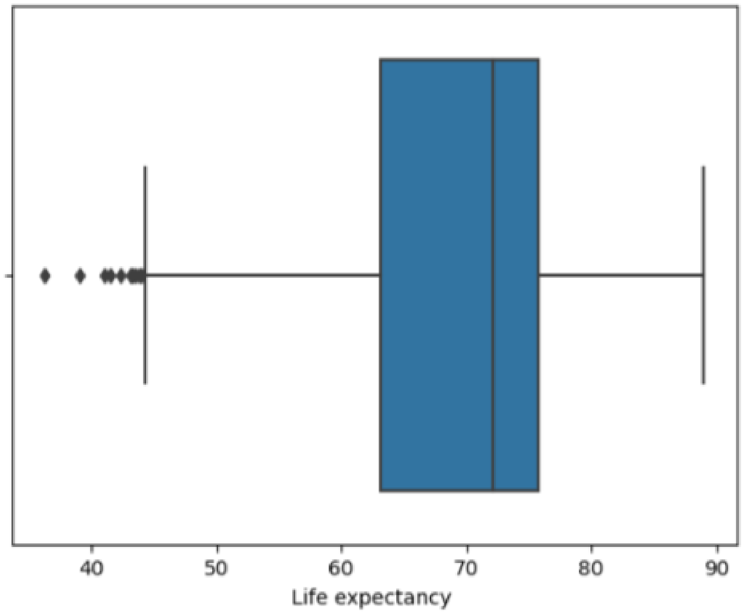

0.0
0.2
0.4
0.6
0.8
1.0
1.2
1e9
Population


0
5
10
15
20
25
thinness 1-19 years


0
5
10
15
20
25
30
thinness 5-9 years


0.0
0.2
0.4
0.6
0.8
Income composition of resources



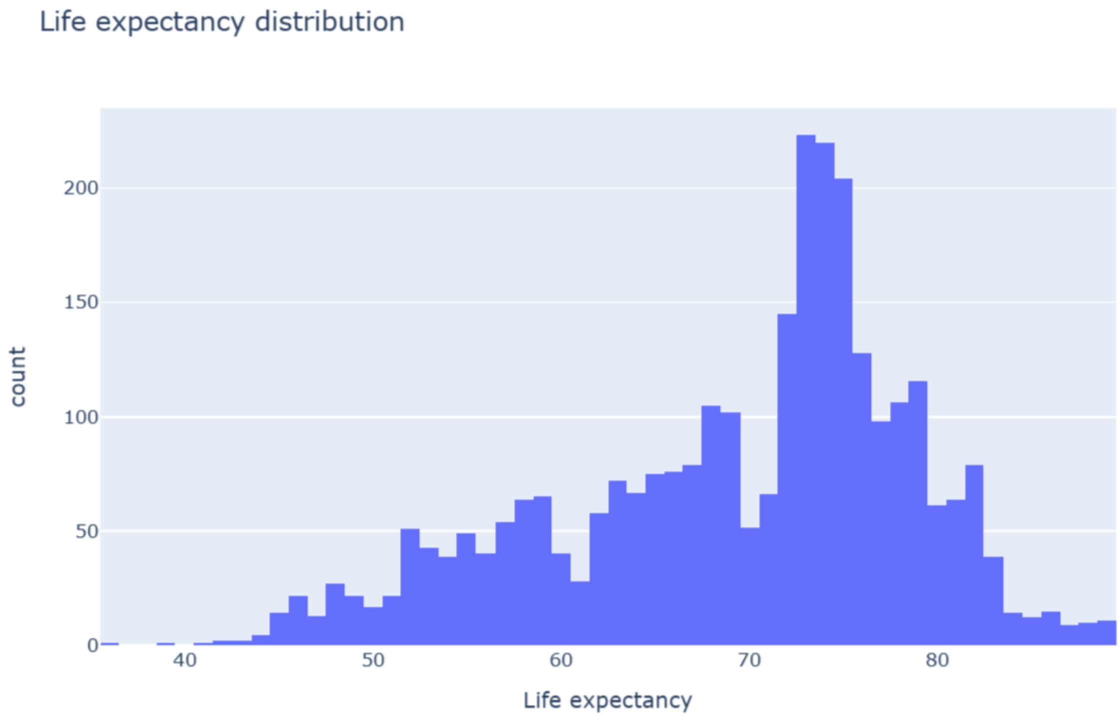


Infant_Deaths represents several infant deaths per 1,000 population. That is why the number beyond 1000 is unrealistic. We will therefore remove them as outliers. The same is true for measles and deaths under five, as both are a number per 1,000 population.

As we can see, some countries spend up to 20,000% of their GDP on health. Most countries spend less than 2,500% of their GDP on health. Since the values are very important in the Expenditure_Percentage, GDP, and Population columns, it is better to take a logarithmic value or use winsorization if necessary.

The BMI values are very unrealistic because the value plus 40 is considered extreme obesity. The median is over 40 and some countries have an average of around 60 which is not possible. We can delete this whole column.

we have used winsorization method from sklearn package to get the mean value which helped us for the direct imputation of extreme values.



Maximum life expectancy is about 72.2-74.6 years.

Country wise life expectancy report:

	Country	Life expectancy
84	Japan	82.53750
165	Sweden	82.51875
75	Iceland	82.44375
166	Switzerland	82.33125
60	France	82.21875
82	Italy	82.18750
160	Spain	82.06875
7	Australia	81.81250
125	Norway	81.79375
30	Canada	81.68750

All the developed countries having a good life expectancy rate where Japan is among the leading country with max life expectancy rate.

Top 10 country with low life expectancy rate:

	Country	Life expectancy
152	Sierra Leone	46.11250
31	Central African Republic	48.51250
94	Lesotho	48.78125
3	Angola	49.01875
100	Malawi	49.89375
32	Chad	50.38750
44	Côte d'Ivoire	50.38750
192	Zimbabwe	50.48750
164	Swaziland	51.32500
123	Nigeria	51.35625

Status of Life expectancy rate among the countries on the basis of percapita income:

	Status	Life expectancy
0	Developed	79.197852
1	Developing	67.111465

Developed countries have high percapita income and standard of living ,thus increases the life expectancy.

	Country	GDP
166	Switzerland	57362.874601
98	Luxembourg	53257.012741
136	Qatar	40748.444104
119	Netherlands	34964.719797
7	Australia	34637.565047
80	Ireland	33835.272005
8	Austria	33827.476309
47	Denmark	33067.407916
153	Singapore	32790.105907
89	Kuwait	31914.378339

```
Top 10 Countries with Lowest GDP
```

	Country	GDP
117	Nauru	136.183210
26	Burundi	137.815321
100	Malawi	237.504042
95	Liberia	246.281748
55	Eritrea	259.395356
122	Niger	259.782441
57	Ethiopia	264.970950
152	Sierra Leone	271.505561
149	Senegal	274.611166
69	Guinea	279.464798

Its necessary to look a

	Country	HIV/AIDS
164	Swaziland	32.94375
192	Zimbabwe	23.26250
94	Lesotho	22.96875
158	South Africa	18.49375
100	Malawi	16.68125
21	Botswana	16.52500
116	Namibia	13.64375
191	Zambia	11.93125
114	Mozambique	11.38750
31	Central African Republic	8.98125

```
Top 10 countries with with high average Body Mass Index of entire populati
```

	Country	BMI
117	Nauru	87.30000
128	Palau	83.30000
38	Cook Islands	82.80000
105	Marshall Islands	81.60000
178	Tuvalu	79.30000
124	Niue	77.30000
88	Kiribati	69.43125
104	Malta	66.18125
136	Qatar	65.65000
109	Micronesia (Federated States of)	65.15000

Top 10 countries with low average Body Mass Index of entire population¶

In [20]:

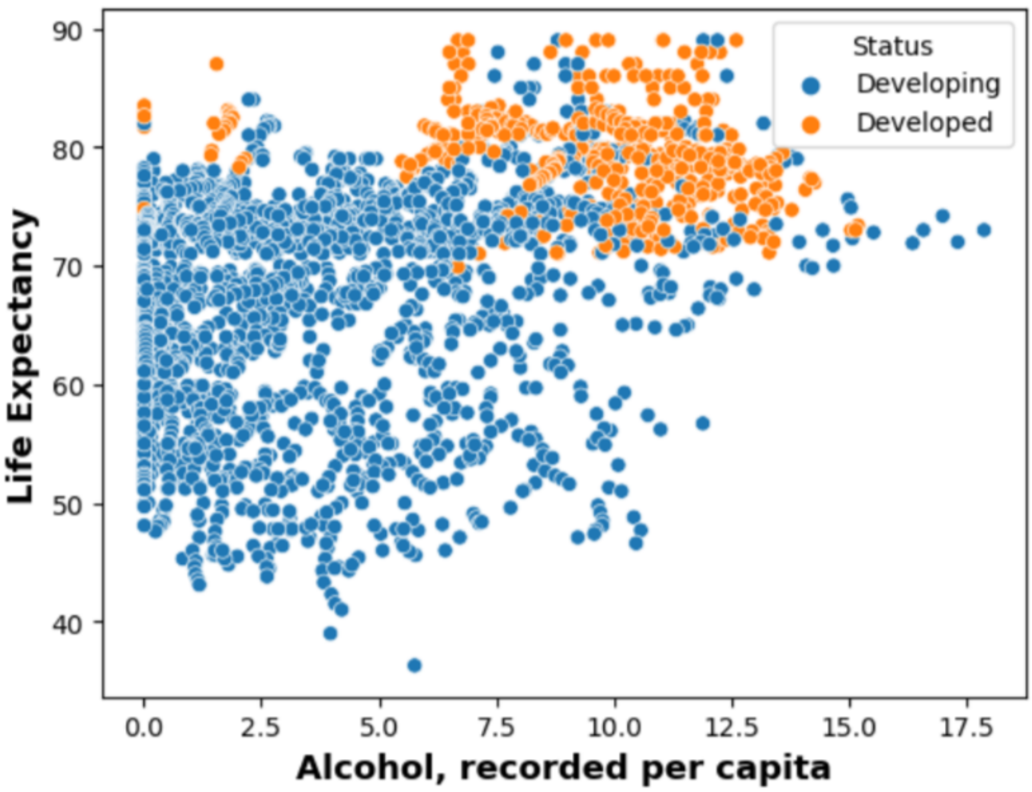
	Country	BMI
117	Nauru	87.30000
128	Palau	83.30000
38	Cook Islands	82.80000
105	Marshall Islands	81.60000
178	Tuvalu	79.30000
124	Niue	77.30000
88	Kiribati	69.43125
104	Malta	66.18125
136	Qatar	65.65000
109	Micronesia (Federated States of)	65.15000

Top 10 countries with low average Body Mass Index

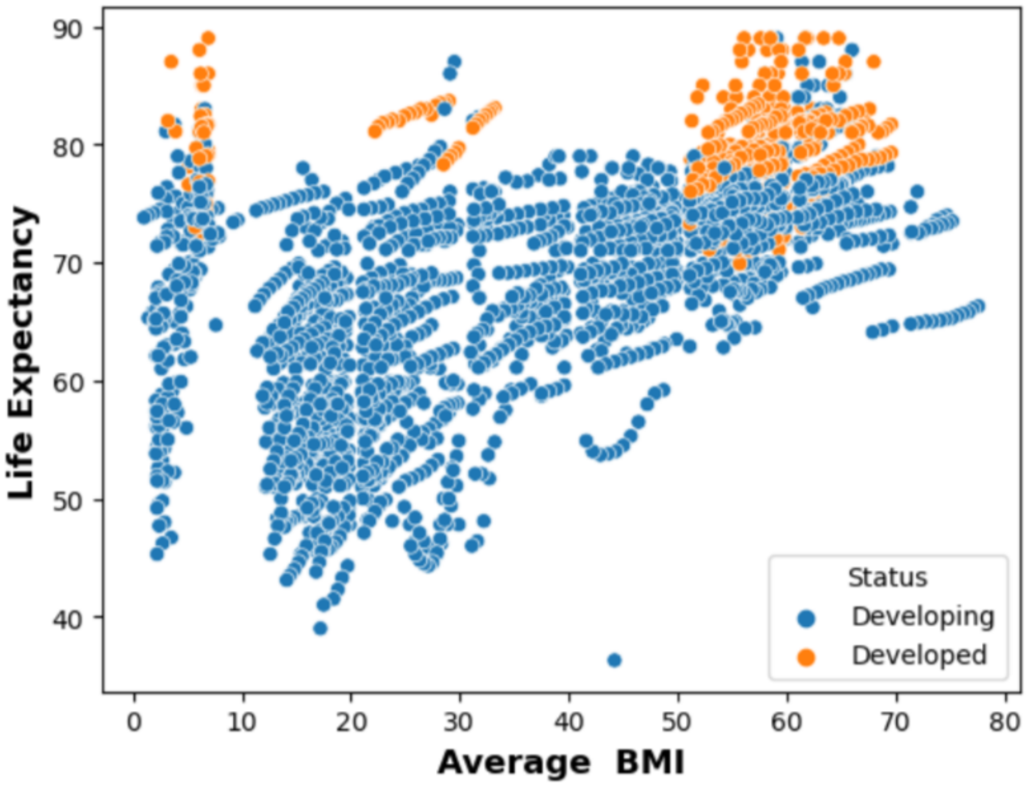
	Country	BMI
142	Saint Kitts and Nevis	5.20000
189	Viet Nam	11.18750
12	Bangladesh	12.87500
91	Lao People's Democratic Republic	14.36250
171	Timor-Leste	14.55000
141	Rwanda	14.75000
99	Madagascar	14.76875
76	India	14.79375
57	Ethiopia	14.80000
55	Eritrea	15.15625

This indicates developing countries are having low BMI than the global average BMI, which would results in low life expectancy rate in this regions.

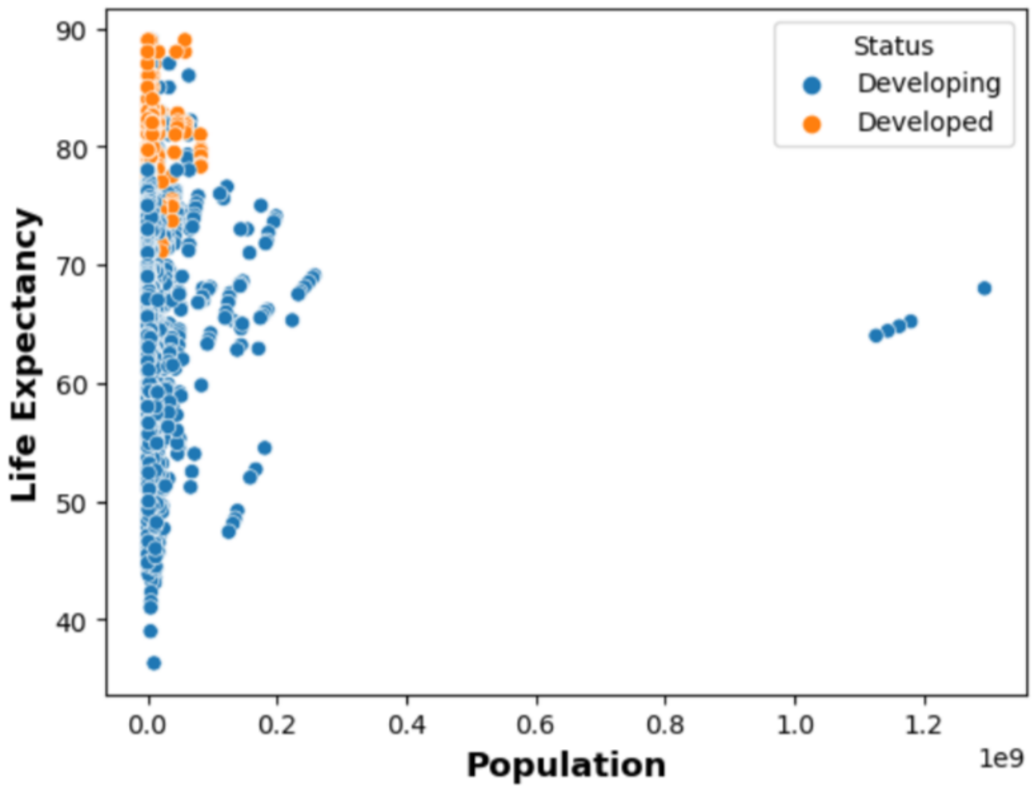
Alcohol, recorded per capita vs Life Expectancy based *on status*



There are more alcohol consumers in developing countries than the developed nations which gives a clear view on the life expectancy rate among the status of those countries.



This tells us clearly that the low BMI leads to reduced life expectancy.

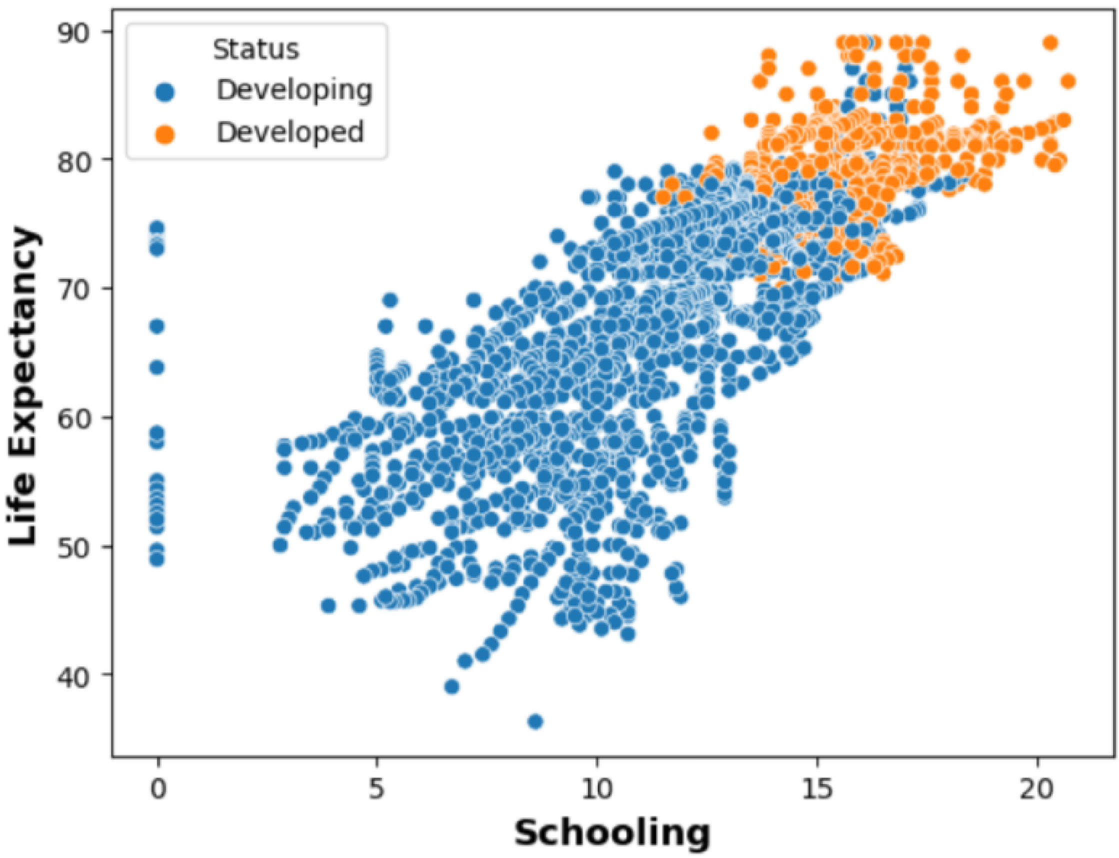


Population density is one of the main factor to compare life expectancy between the nations.

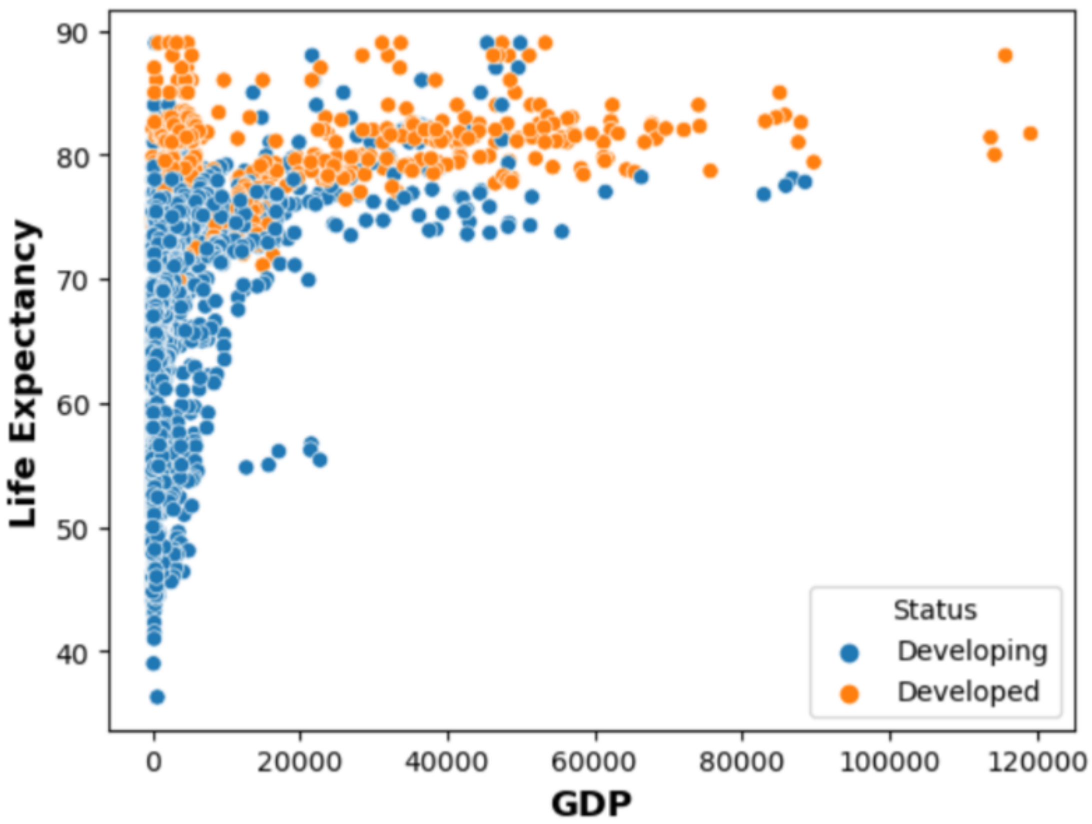
Lower the population higher the life expectancy and higher the population lower the life expectancy.This is also saying the standardity of living among the people over the globe.

Developing countries are economically less stable and hence impacting on literacy rate with low life expectancy. Which we can

see from the next plot.


Status
Developing
Developed
Life Expectancy
90
80
70
60
50
40
0
5
10
15
20
Schooling

schooling increases the life expectancy of an individual.



we know GDP has great impact on SDGs, which further leads to impact on our economy and health .

Here we can infer that the developed countries has more GDP with higher life expectancy rate unlike developing countries with

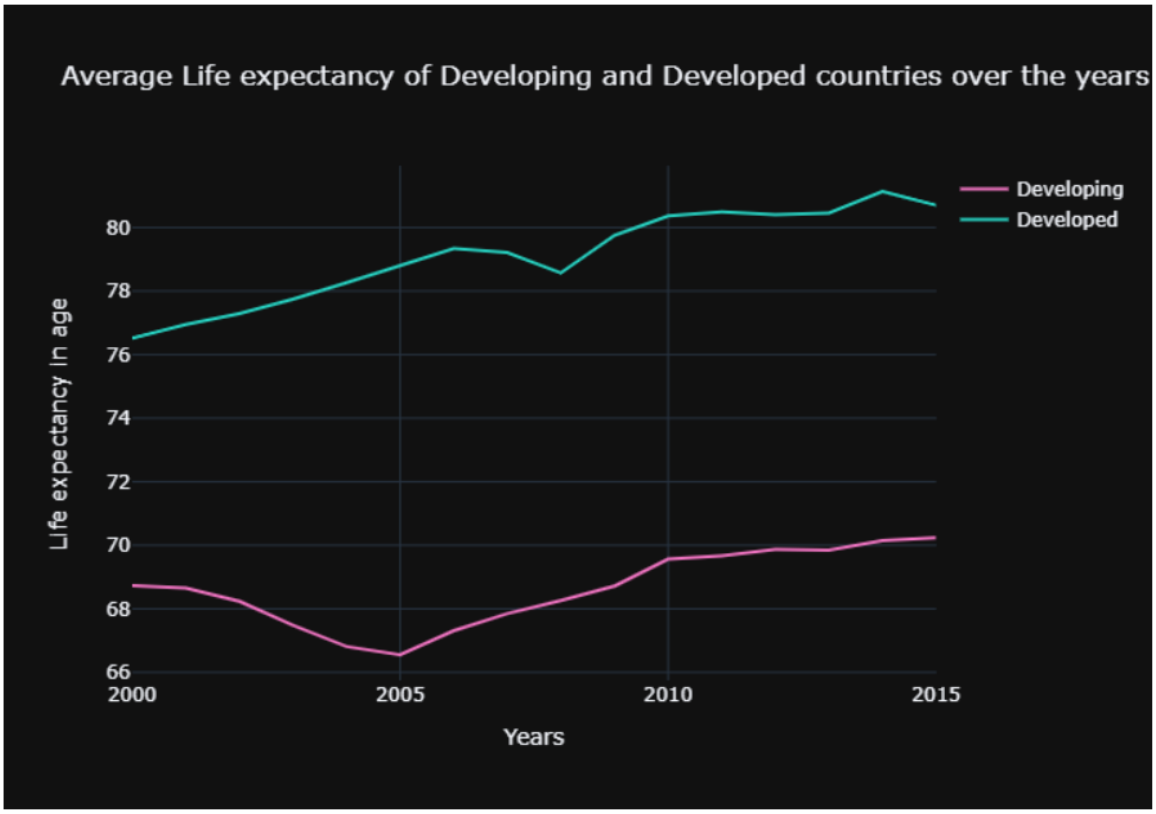
low GDP and lower life expectancy rate.


```
Life expectancy                    1.000000
Schooling                          0.747556
Income composition of resources    0.724790
 BMI                               0.565697
Diphtheria                         0.478427
Polio                              0.464486
GDP                                0.461126
Alcohol                            0.403077
percentage expenditure             0.381418
Hepatitis B                        0.255452
Total expenditure                  0.217304
Year                               0.170819
Population                        -0.021600
Measles                           -0.157767
infant deaths                     -0.196769
Name: Life expectancy , dtype: float64
```

Findings:

- **Schooling increases the life expectancy of an individual by 74.7%**
- **Income composition of resources of a country increases the life expectancy by 72.4%**
- **A BMI is directly impacting on life expectancy of an individual by 56.5% .**
- **Several disease prevalence in an area or country is directly impacting on life periods.**
- **Adult mortality is positively correlated with HIV/AIDS and negatively correlated with education and the distribution of resource income.**
- **Baby deaths and Under five deaths are significantly positively correlated.**
- **Alcohol and education work well together.**
- **The mix of resource income, GDP, and life expectancy are all positively correlated with percentage expenditure.**
- **Polio and diphtheria have a substantial positive association with hepatitis B.**
- **Moreover, there is a direct correlation between polio and life expectancy, diphtheria, and hepatitis B.**
- **Polio and life expectancy have a strong favourable association with diphtheria.**
- **The heat map shows that Life expectancy is positively correlated with GDP, diphtheria, polio, education, resource income composition, and percentage spending. With respect to Adult Mortality, Thinness 1-19 Years, Thinness 5-9 Years, and Life Expectancy, there is a negative correlation with Under five deaths, Infant deaths, and HIV/AIDS.**

To complete the life expectancy analysis task, let's examine them in greater detail:



We can see from the two graphs above that developed countries have more life expectancy than in developing countries.

Since our dependent variable is life_expectancy which is numerical . Therefore we will perform regression models for this problem.

We will perform

Before we proceed to the Regression models, let us consider some scales of measurement for the comparison of the model.

SCALES OF MEASURE :

The following metrics are considered for the analysis of the model.

Regression model evaluation metrics

The MSE, MAE, RMSE, and R-Squared metrics are mainly used to evaluate the prediction error rates and model performance in regression analysis.

- **MAE** (Mean absolute error) represents the difference between the original and predicted values extracted by averaged the absolute difference over the data set.
- **MSE** (Mean Squared Error) represents the difference between the original and predicted values extracted by squared the average difference over the data set.
- **RMSE** (Root Mean Squared Error) is the error rate by the square root of MSE.
- **R-squared** (Coefficient of determination) represents the coefficient of how well the values fit compared to the original values. The value from 0 to 1 interpreted as percentages. The higher the value is, the better the model is.

The above metrics can be expressed,

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}|$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2}$$

$$R^2 = 1 - \frac{\sum (y_i - \hat{y})^2}{\sum (y_i - \bar{y})^2}$$

Where,

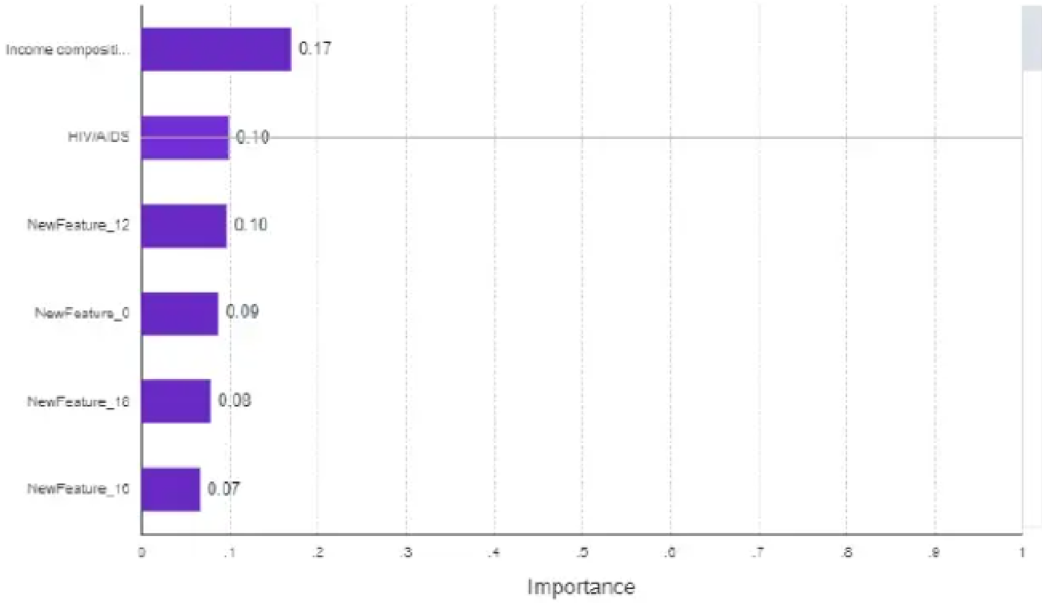
$\hat{y}$ $- predicted\ value\ of\ y$

$\bar{y} - mean\ value\ of\ y$

Lets consider the significant variable which are really associated with the dependent variable:

So to find that we do feature importance:

Feature Importance:



1.Simple Linear Regression:

Test results

- **MAE : 2.7304834**
- **MSE : 12.46**
- **RMSE : 3.530076**
- **R-squared : 0.855782**

Which is a slightly overfitted model with 82% of accuracy r-square value.

```
                    Mixed Linear Model Regression Results
=============================================================================================
Model:                        MixedLM            Dependent Variable:           Life_Expectancy
No. Observations:             2864               Method:                       REML
No. Groups:                   179                Scale:                        2.8842
Min. group size:              16                 Log-Likelihood:               -7072.2854
Max. group size:              16                 Converged:                    No
Mean group size:              16.0
---------------------------------------------------------------------------------------------
                                                  Coef.  Std.Err.    z    P>|z|  [0.025 0.975]
---------------------------------------------------------------------------------------------
Intercept                                         79.581      1.574 50.569 0.000  76.496 82.665
Adult_Mortality_scaled                            -0.051      0.024 -2.153 0.031  -0.097 -0.005
Alcohol                                           -0.015      0.100 -0.147 0.883  -0.211  0.182
Polio_scaled                                       0.169      0.034  4.914 0.000   0.102  0.237
hivaids                                           -1.659
BMI                                                0.025      0.052  0.491 0.623  -0.076  0.126
thinness_1to19_years                              -0.616      0.199 -3.100 0.002  -1.006 -0.227
Developing                                       -11.848      1.684 -7.036 0.000 -15.148 -8.547
Group Var                                          4.665  67179.833
Group x Adult_Mortality_scaled Cov                -0.060      0.350
Adult_Mortality_scaled Var                         0.025      0.003
Group x Alcohol Cov                                0.068      0.724
Adult_Mortality_scaled x Alcohol Cov               0.012
Alcohol Var                                        1.038      0.087
Group x Polio_scaled Cov                          -0.017      0.286
Adult_Mortality_scaled x Polio_scaled Cov          0.009
Alcohol x Polio_scaled Cov                        -0.239
Polio_scaled Var                                   0.217
Group x hivaids Cov                                0.354 335899.162
Adult_Mortality_scaled x hivaids Cov              -0.007      0.016
Alcohol x hivaids Cov                              0.020
Polio_scaled x hivaids Cov                        -0.004
hivaids Var                                        3.475      0.351
Group x BMI Cov                                   -0.022      0.469
Adult_Mortality_scaled x BMI Cov                   0.000
Alcohol x BMI Cov                                 -0.013
Polio_scaled x BMI Cov                             0.011      0.017
hivaids x BMI Cov                                  0.023
BMI Var                                            0.465
Group x thinness_1to19_years Cov                  -0.120      1.534
Adult_Mortality_scaled x thinness_1to19_years Cov  0.000
Alcohol x thinness_1to19_years Cov                -0.047
Polio_scaled x thinness_1to19_years Cov            0.137
hivaids x thinness_1to19_years Cov                -0.280
BMI x thinness_1to19_years Cov                     0.024
thinness_1to19_years Var                           3.000
Group x Developing Cov                             0.948
Adult_Mortality_scaled x Developing Cov           -0.010      0.355
Alcohol x Developing Cov                           0.062      0.697
Polio_scaled x Developing Cov                     -0.014      0.312
hivaids x Developing Cov                           0.373 335899.162
BMI x Developing Cov                               0.003      0.485
thinness_1to19_years x Developing Cov             -0.146      1.577
Developing Var                                     4.792
=============================================================================================
```

We will take all the significant variables atlast to predict with the best model.

2.Decision Tree Model:

- **MAE : 1.47212121**
- **MSE : 6.915393**
- **RMSE : 2.6297**
- **R-squared : 0.9199674166**

Which is highly overfitted model with 91% of accuracy in the test and 100 % accuracy in the train data.

3.Random Forest:

- **MAE : 1.128212127**
- **MSE : 3.4514**
- **RMSE : 1.82897**
- **R-squared : 0.9584**

- Which is giving 96% of accuracy, Compared to other model score RF is giving an amazing result. Although the model is slightly overfitted i.e., giving 99% accuracy in train data and 95.84% in test data. We can choose to optimise the model by various optimization technique.

4.Support Vector Model:

This model is performing good results with accuracy 86.6% in the train and 85.7% in the validating set with less than 1% slightly overfitting .

5.Ada Boost:

- **This model is a perfect model with similar test and train accuracy result which proves to be a good model for predicting life expectancy.**

- **90% accuracy**

6.Grad Boost:

This model is slightly overfitted model with 96% accuracy on train data and 94% in test data.

7.Lasso: The model performs good with slightly lower accuracy.

8.Ridge:

- **This model performs good in predicting life expectancy with 82% in train data and 81% in test set.**
- **Which is slightly overfitting .**

	mean_absolute_error	mean_squared_error	train_accuracy	test_accuracy
Linear Reg	3.065368	4.134512	0.822087	0.813881
Ridge	3.061431	4.137167	0.821895	0.813642
Lasso	3.430167	4.639142	0.775612	0.765675
SVR	2.447159	3.618697	0.866374	0.857424
Decision Tree	1.608050	2.711103	1.000000	0.919973
RF	1.211522	1.952591	0.994181	0.958489
Ada Boost	2.345264	2.996107	0.908574	0.902263
Grad Boost	1.634460	2.271066	0.963189	0.943843

Conclusion:

The projected value of the coefficient, according to our final model summary, shows how significantly each component affects life expectancy. Nine important variables that determine life expectancy were discovered. They include adult mortality, BMI, polio, diphtheria, HIV/AIDS, GDP, and thinness 1 19yrs. Since HIV/AIDS has the highest predicted coefficient value, it has the most adverse impact on life expectancy. Moreover, we discovered that immunisation against diphtheria and polio does increase life expectancy. Population has no bearing on life expectancy because it was excluded during the stepwise selection process, has a p-value of 0.972, and has a weakly correlated with life expectancy. According to our research, the number of years spent in school does have a positive link and is a major determinant**.**

According to this paradigm, nations with high HIV/AIDS rates should concentrate on reducing them, boost their polio and diphtheria immunisation rates, and make educational investments. These nine factors impacting life expectancy should be taken into consideration by nations.

Limitations:

This data collection only contains information from 2000 to 2015; a more recent dataset might give a more accurate picture of the current life expectancy.

A non-linear model would be a better fit for this dataset since our linear regression model did not meet the homoceskedasticity requirements.