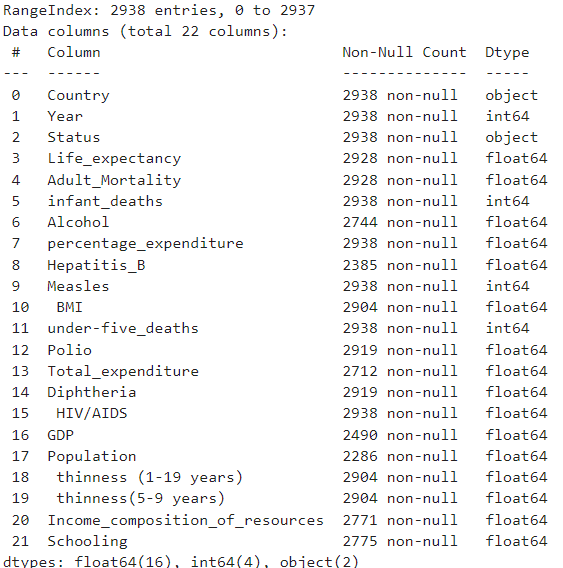
Summary of the Regression dataset   
– Life Expectancy Prediction

**About the data:**

Link: [Life expectancy Regression with ANN](https://www.kaggle.com/code/ranasabrii/life-expectancy-regression-with-ann/input)  
Data Info:  
 

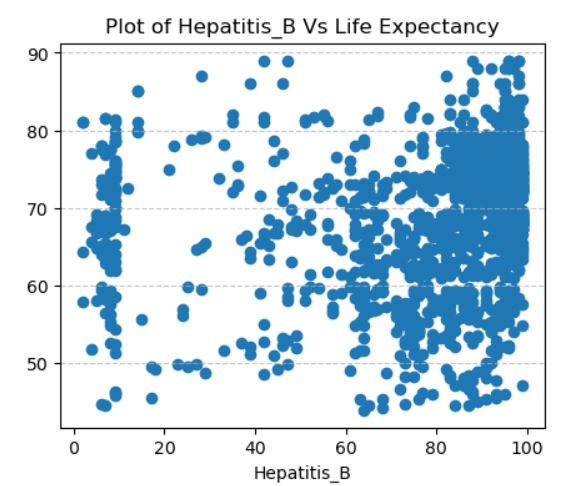
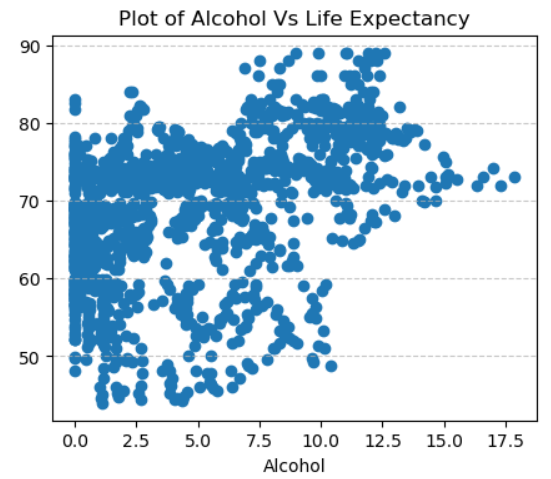
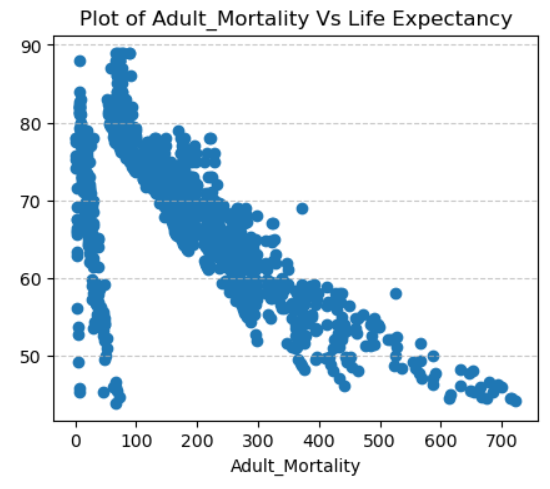
Columns:  
1) Country: Name of the country (has 193 different countries)  
2) Year: Tells the year the data if taken from.  
3) Status: Developed or Developing status.  
4) Life Expectancy: Average life expectancy of the country.  
5) Adult Mortality: Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years   
per 1000 population).  
6) Infant Mortality: Number of Infant Deaths per 1000 population.  
7) Alcohol: Alcohol consumption (in litres of pure alcohol) per capita.  
8) Percentage Expenditure: Expenditure on health as a percentage of Gross Domestic Product per capita (%).  
9) Hepatatis B: Percentage of Hepatitis B immunization coverage among 1-year-olds.  
10) Measles:  Number of reported cases of Measles 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: Percentage of 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: Percentage of Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds.  
16) HIV/AIDS: Deaths per 1000 live births HIV/AIDS (0-4 years).

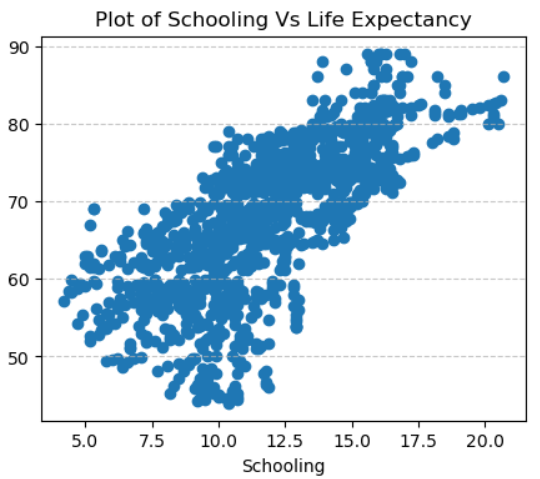
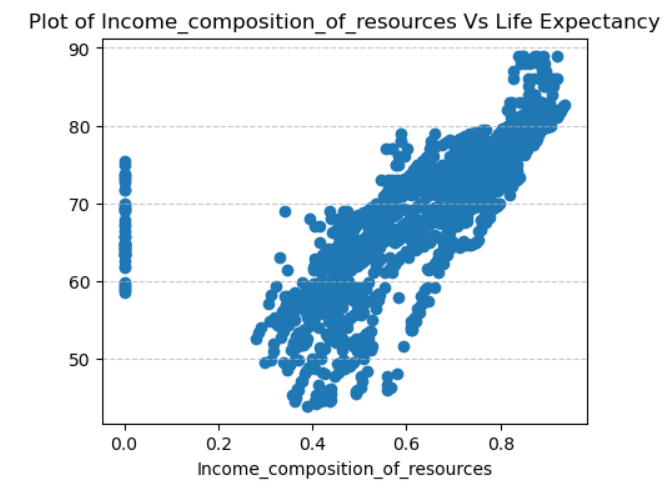
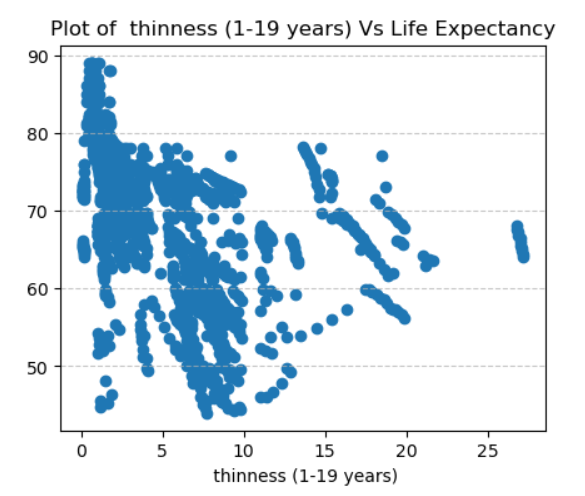
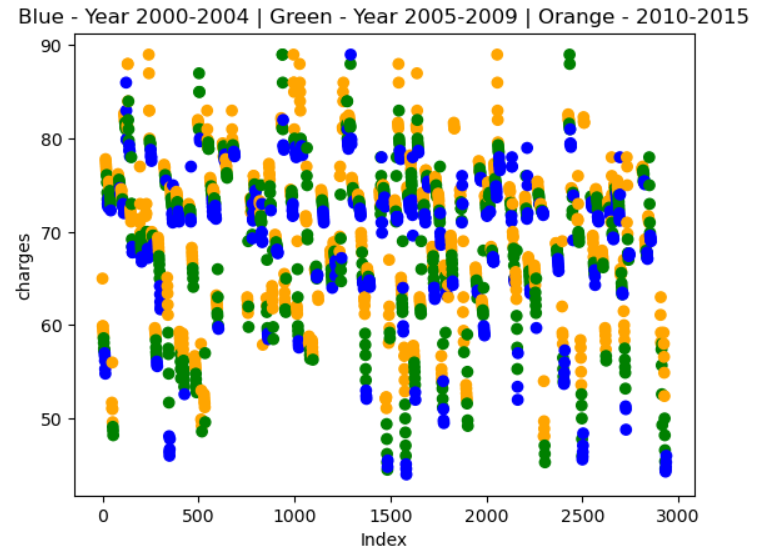
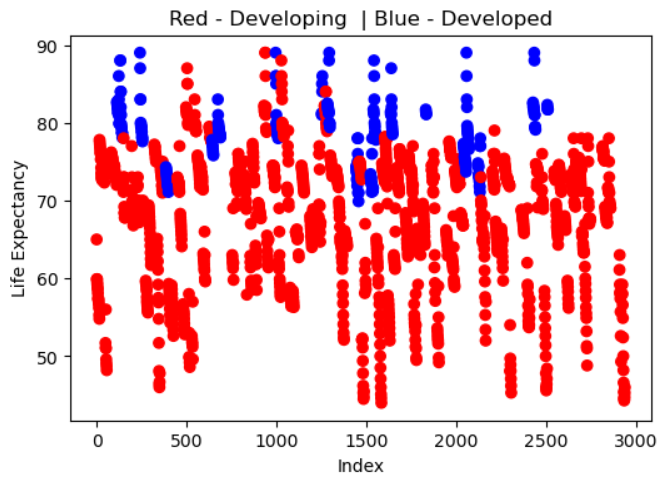
17) GDP: Gross Domestic Product per capita (in USD).  
18) Population: Population of the country.  
19) Thinness (1-19 Years): Percentage of prevalence of thinness among children and adolescents for Age 10 to 19.  
20) Thinness (5-9 Years): Percentage of prevalence of thinness among children for Age 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).

Pairs of Columns with high corelation:  
-> Infant deaths & under five deaths: 0.997  
-> Percentage expenditure & GDP: 0.959  
-> thinness (1-19 years) & thinness (5-9 years): 0.928

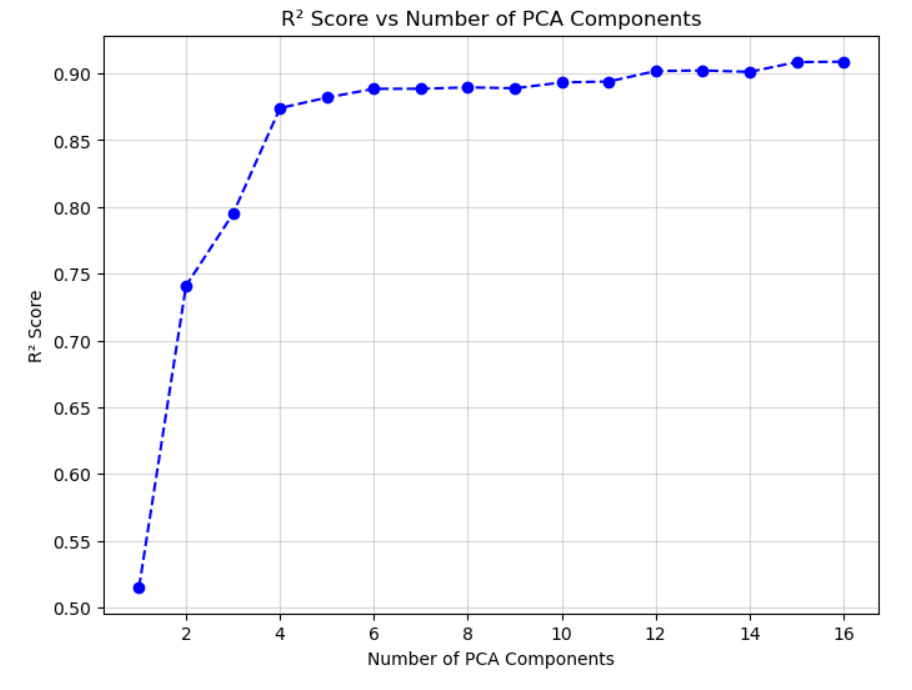
**So, we can delete - infant deaths, percentage expenditure, thinness (5-9 years)**

**Plotting the data:**

1) Various continuous data field plotted against Life Expectancy:  


  
2) For visualising Categorical data fields:  


**So, we can drop year and country as the they evenly spread for all categories and not providing much information.**

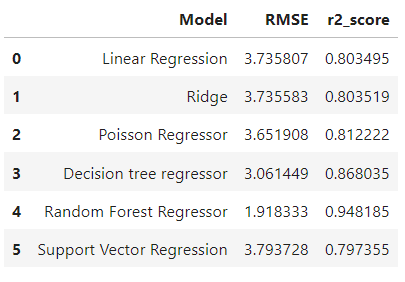
**Trying PCA and BERUTA for further Pre-processing:**1) Plot of R2\_score and n-components of PCA**:  
**

**PCA can’t be applied as it is only decreasing the r2 score**

2) Selected Features after BERUTA:

|  |  |  |
| --- | --- | --- |
| Adult Mortality | under five deaths | GDP |
| Alcohol | Polio | Population |
| Measles | Total expenditure | thinness (1-19 years) |
| BMI | HIV/AIDS | Income composition of resources |
| Schooling |  |  |

**Model Selection**

Root-mean square error and R2 score of different models when applied on the data: ****

Hyper parameter tuned Random Forest and Decision Tree for more generalised model:

|  |  |  |
| --- | --- | --- |
|  | Random Forest | Decision Tree |
| RMSE | 1.92 | 2.86 |
| R2 Score | 0.95 | 0.88 |

Overall Result:  
Got by using BERUTA and then Random Forest Regressor to get least RMSE of 1.92 and maximum R2 score of 0.95.