

**PROJECT REPORT**

ARTIFICIAL

INTELLIGENCE

-Taimoor Manzoor (SP20-BSCS-0021)-(BM)

# **TITLE: Life Expectancy (WHO)**

**DOMAIN:**

In the past although there have been a lot of research happen in the past on factors affecting life expectancy in which we see different factor of life like disease, income tax etc. It was found that most of the countries do not focus on the life expectancy Also, some of the past research was done considering different model like multiple linear regression, lasso, ridge etc. based on data set of two or three year for all the countries. Hence, this gives motivation to resolve both the factors stated previously by formulating a regression model based on different model while the dataset is considering is around 2000 to 2015 and here is consider different parameter like health, environment, economic etc. This dataset is release by the World Health Organization which is very beneficial for the research and analysis

**DATASET LINK:**

https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

**DATASET DESCRIPTION:**

**Data contains 2938 Observation and 22 Features.**

Features

* Country: Country
* Year: Year
* Status: Developed or Developing status
* Life expectancy: Life Expectancy in age
* Adult Mortality: Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)
* infant deaths: Number of Infant Deaths per 1000 population
* Alcohol: 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: Hepatitis B (HepB) immunization coverage among 1-year-olds (%)
* Measles: 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: Polio (Pol3) immunization coverage among 1-year-olds (%)
* Total expenditure: General government expenditure on health as a percentage of total government expenditure (%)
* Diphtheria : Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)
* HIV/AIDS : Deaths per 1 000 live births HIV/AIDS (0-4 years)
* GDP:Gross Domestic Product per capita (in USD)
* Population: Population of the country
* 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(indexrangingfrom 0 to 1)
* Schooling:Number of years of Schooling(years)
* *Target is\** (target)

**MODELS APPLIED AND METHODOLOGY:**

We use preprocessing for improve the data for example extract the outlier and also fill the missing value and Exploratory data analysis also do visualization for better understanding the dataset and use machine learning model.

* Lasso
* Ridge
* Decision Tree Regression

# **RESULTS AND COMPARISION**

The Decision tree regression is better than both other models (ridge and lasso) because lasso r2\_score value is 0.779 and ridge r2\_score value is 0.777 and Decision tree regression r2\_score value is 0.9 which is greater both of other

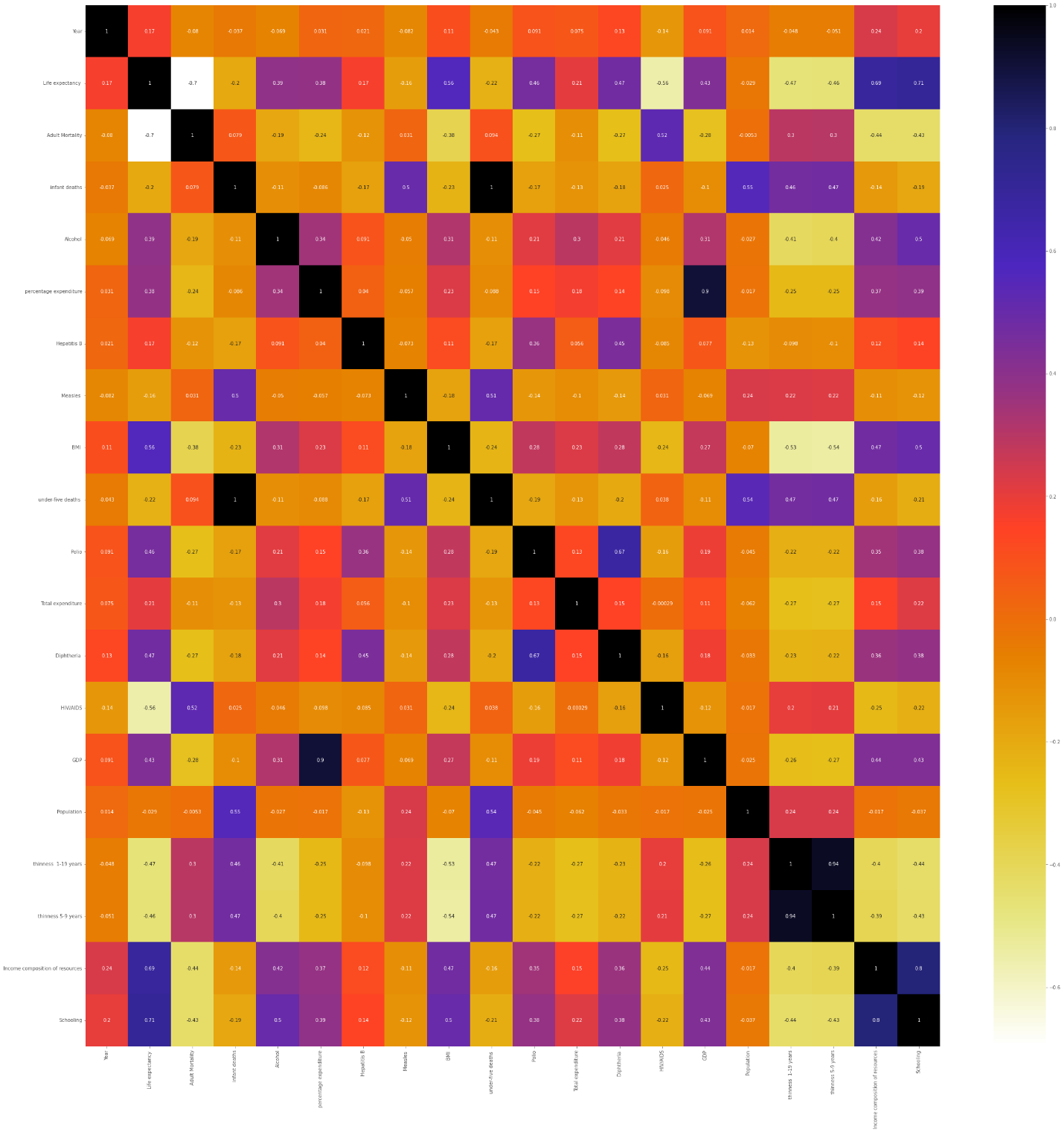
**CODE:**

**https://colab.research.google.com/drive/1hbw7qDlYrVcWJrQI35ycj9wCjLxJM6Wq#scrollTo=defc3cdf**

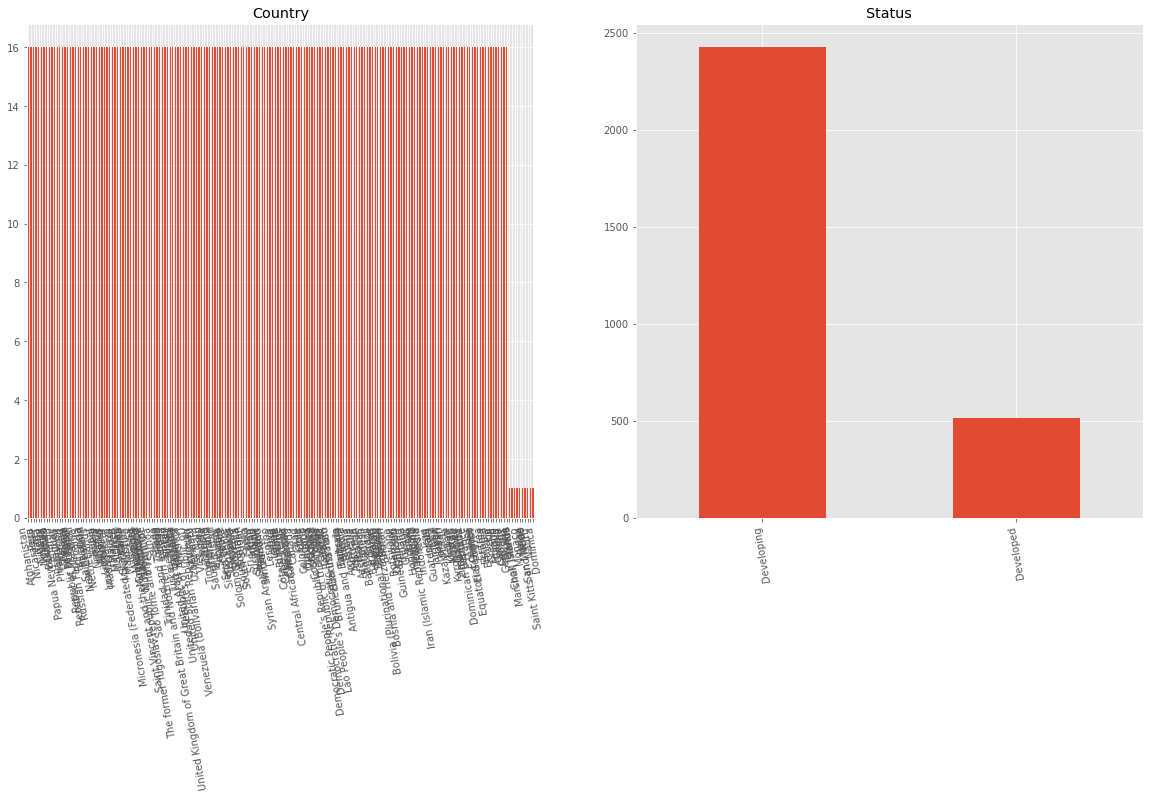


**VISUALIZATION:**

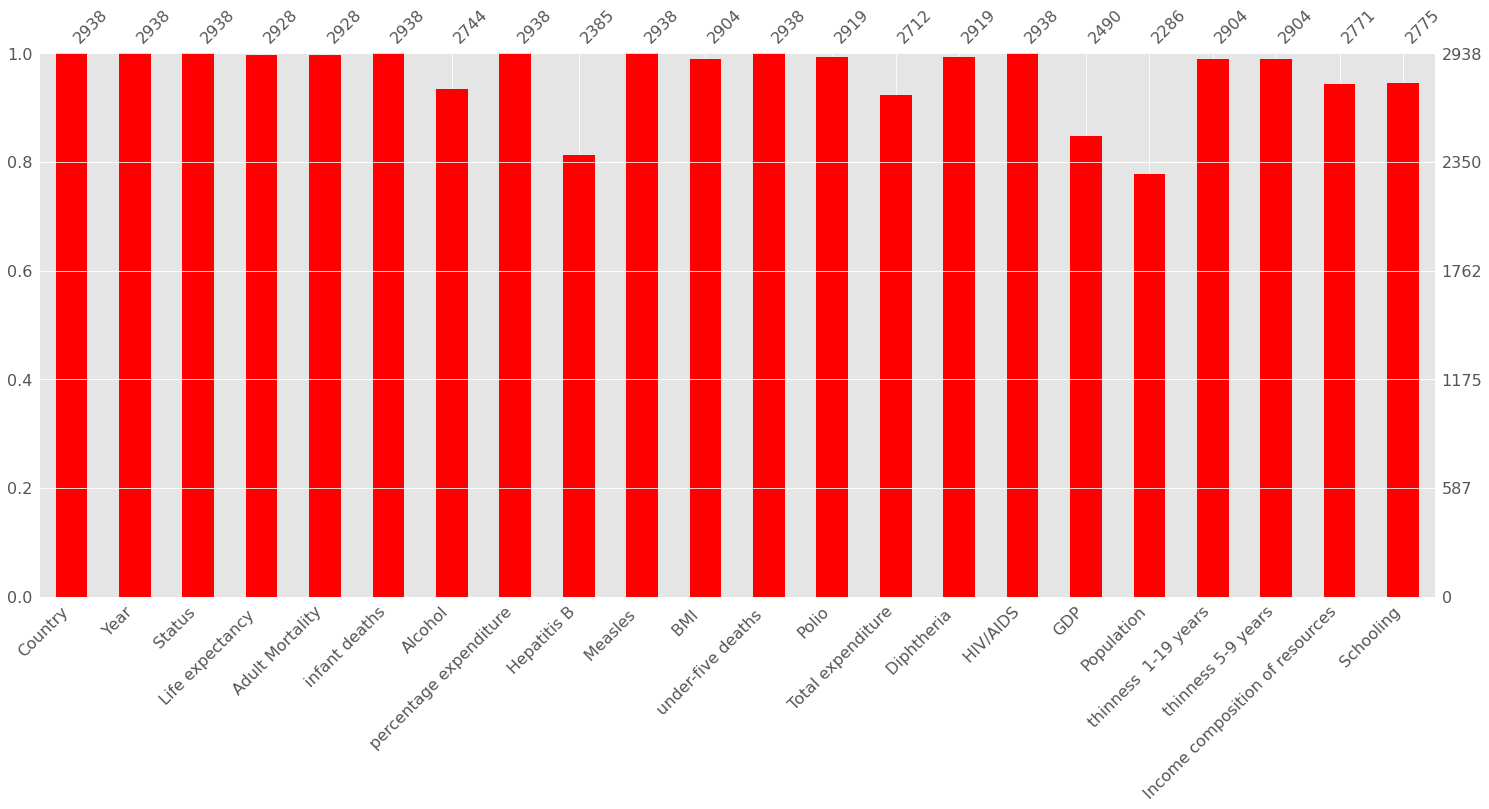
1. Heatmap



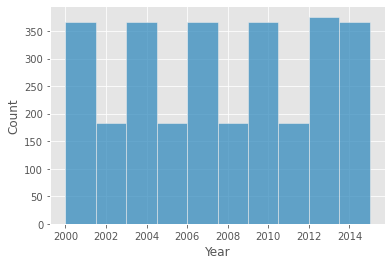
1. Bar Plot

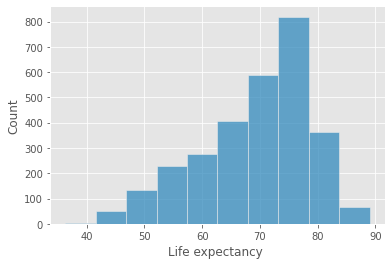


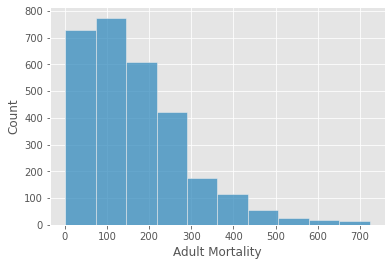
#Null Values

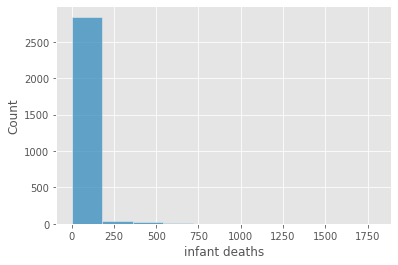


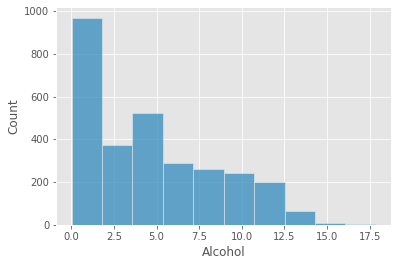
1. Histogram Plot

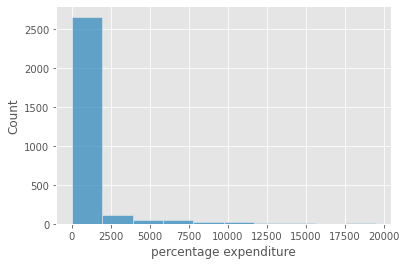


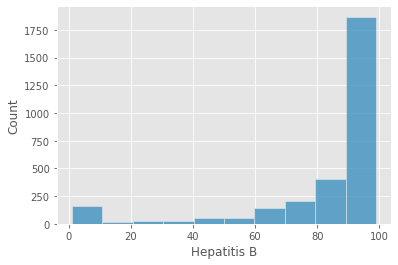


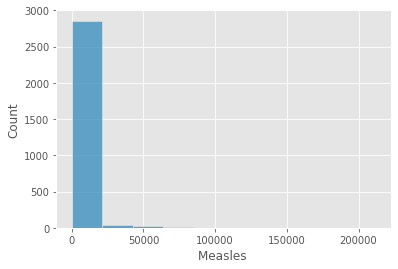


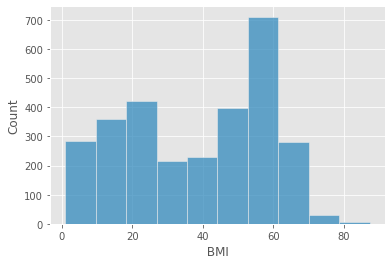


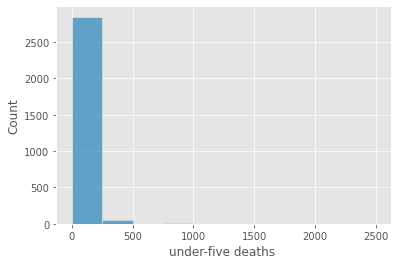


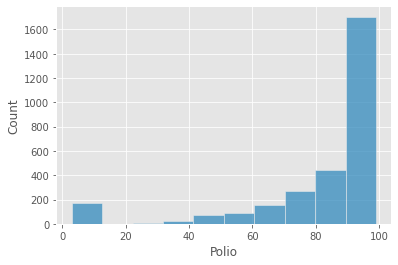


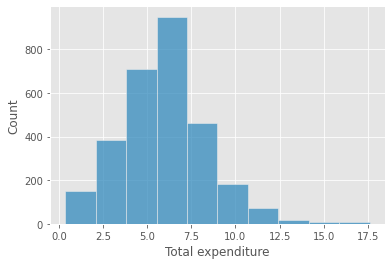


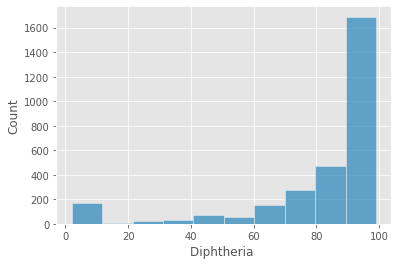


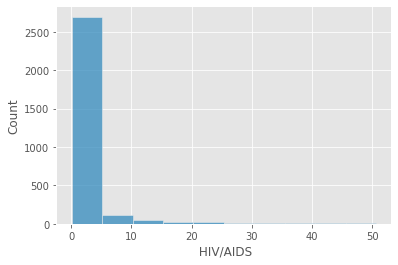


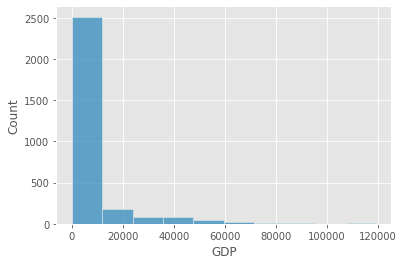


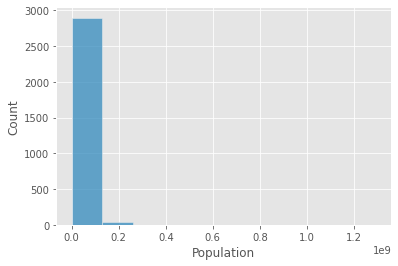


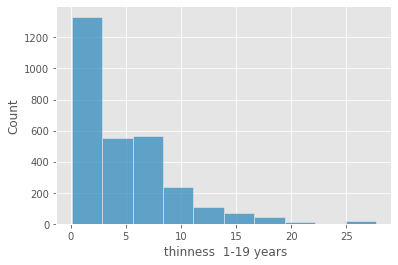


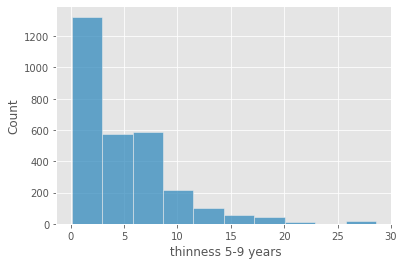


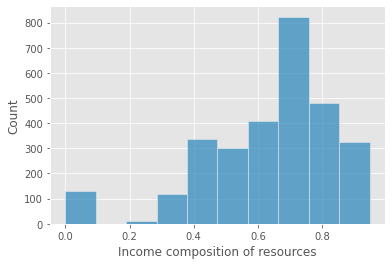


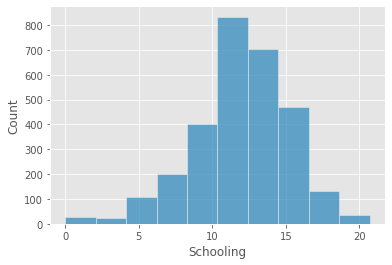




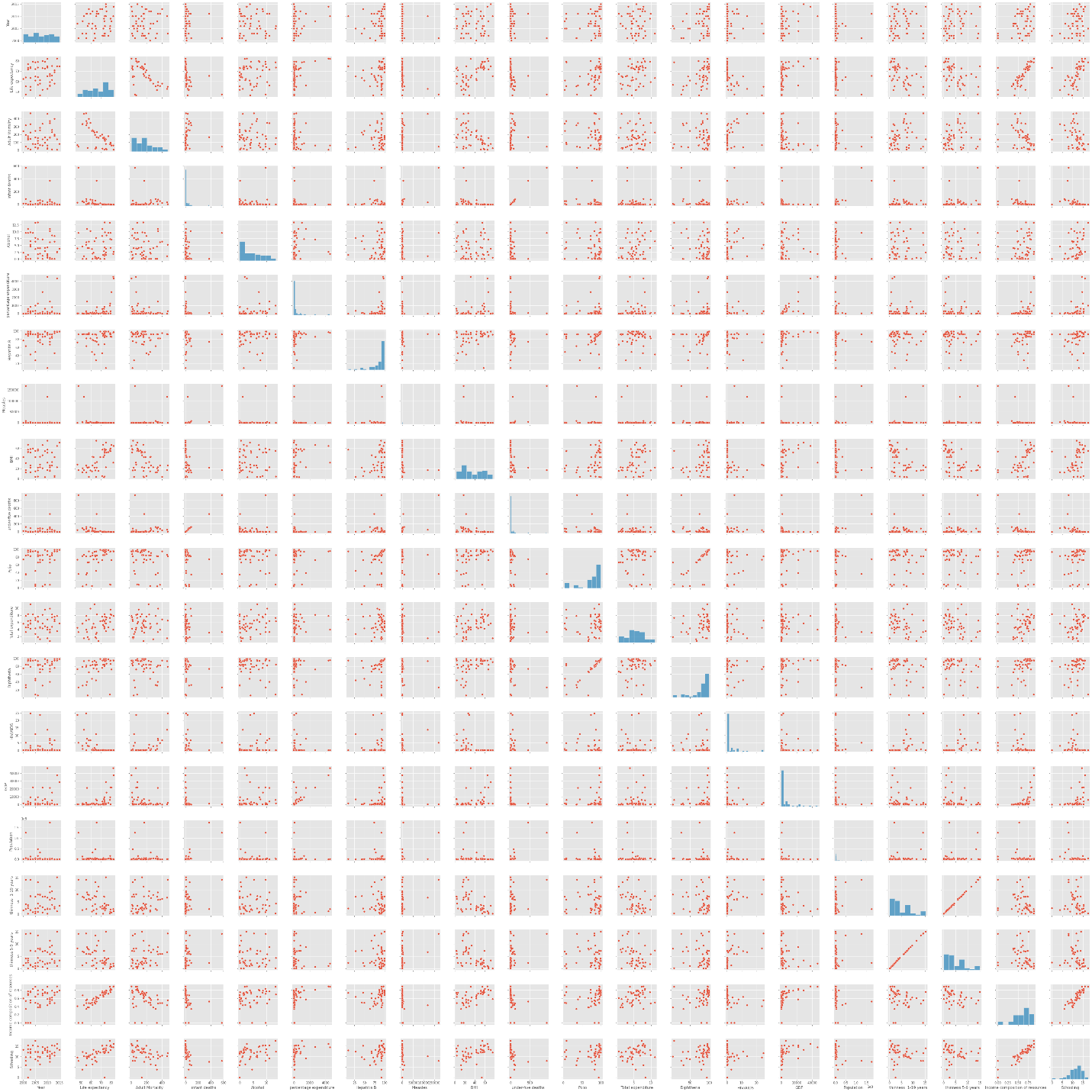




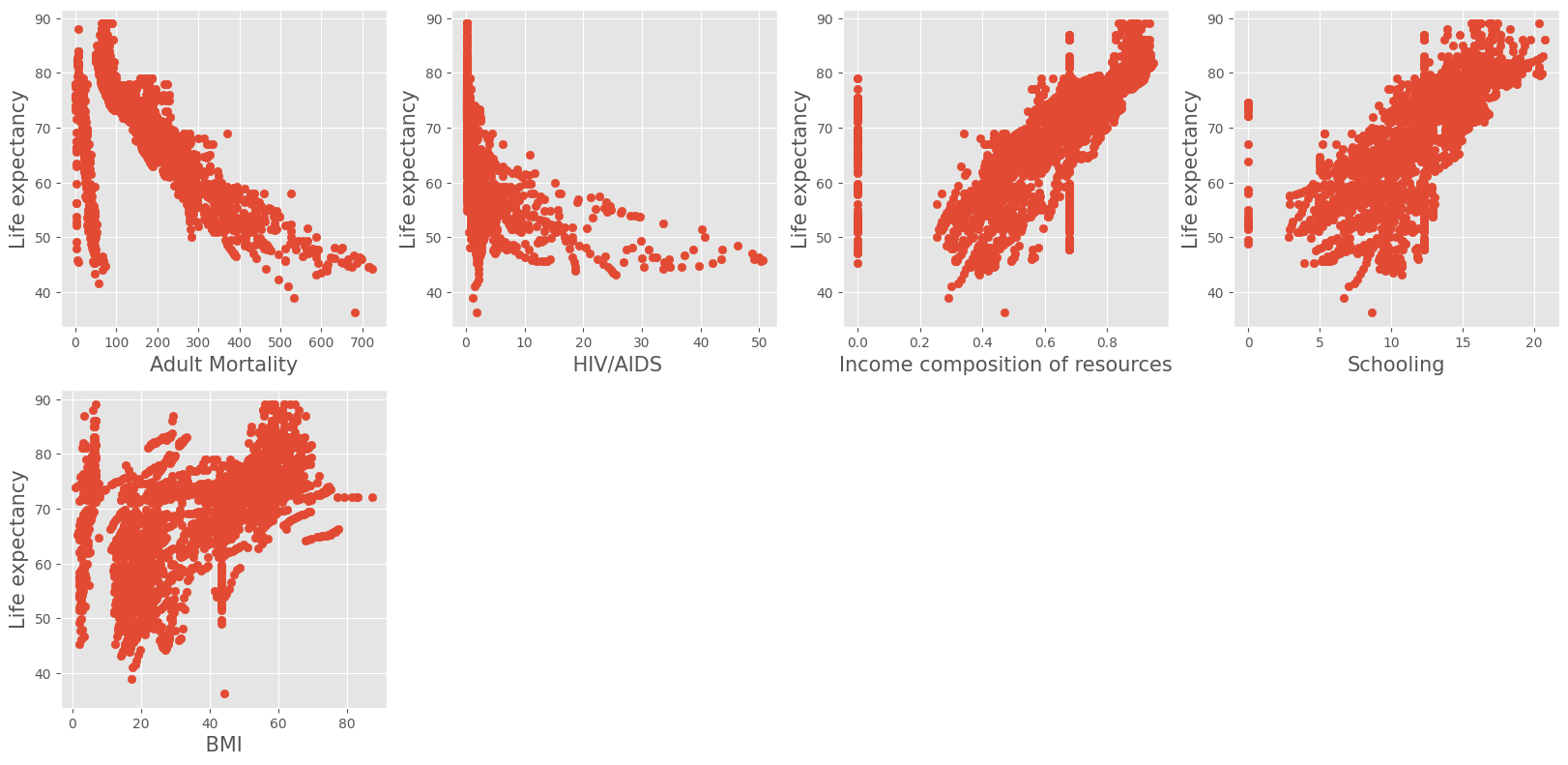




1. Pair Plot



1. Scatter Plot



# **FUTURE WORK**

# In this particular study we are using Lasso and Ridge Regression and Decision Tree Regression to predict the data.

In future, we can enhance our research through various formats we will be doing analysis in formats such as SVM, and Random Forest.

**APPLICATION**

From the data analysis we have come to conclude that health diseases differ in different areas and we can use this data to predict how each disease has effect on the life expectancy of someone and also we can take step to improve the life quality of people. If we take chunk of the dataset we can use it for improve the life expectancy of our country.

**GAP REPORT:**

The gap report is that model is accuracy is little less because I want the accuracy of the greater than 0.95 so in the future I will apply different machine learning for improving accuracy of the model.