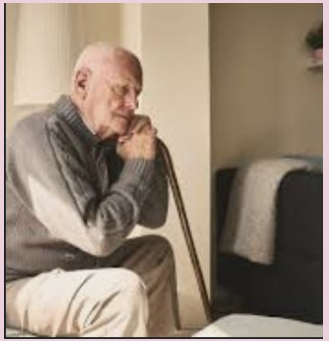
Predicting Life Expectancy With EDA & Machine Learning

Capstone project: PCDSI0121



Prepared by:
Sarala Devi

Scope

- Information on life expectancy dataset
- Challenges faced in dataset
- Approach taken to overcome challenges
- Problem statement
- EDA(Analysis) on dataset
- Machine learning models used for prediction
- Conclusion
- Resources
- End

Information on dataset

Name of dataset: Life Expectancy Data (WHO).csv

Link to dataset: https://www.kaggle.com/code/mrinath/starter-life-expectancy-who-9536c272-3/data

Records: Year 2000 - 2015

	Country	Year	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	...	Polio	Total expenditure	Diphtheria	HIV/AIDS	GDP	Population	thinness 1-19 years
0	Afghanistan	2015	Developing	65.0	263.0	62	0.01	71.279624	65.0	1154	...	6.0	8.16	65.0	0.1	584.259210	33736494.0	17.2
1	Afghanistan	2014	Developing	59.9	271.0	64	0.01	73.523582	62.0	492	...	58.0	8.18	62.0	0.1	612.696514	327582.0	17.5
2	Afghanistan	2013	Developing	59.9	268.0	66	0.01	73.219243	64.0	430	...	62.0	8.13	64.0	0.1	631.744976	31731688.0	17.7
3	Afghanistan	2012	Developing	59.5	272.0	69	0.01	78.184215	67.0	2787	...	67.0	8.52	67.0	0.1	669.959000	3696958.0	17.9
4	Afghanistan	2011	Developing	59.2	275.0	71	0.01	7.097109	68.0	3013	...	68.0	7.87	68.0	0.1	63.537231	2978599.0	18.2
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2933	Zimbabwe	2004	Developing	44.3	723.0	27	4.36	0.000000	68.0	31	...	67.0	7.13	65.0	33.6	454.366654	12777511.0	9.4
2934	Zimbabwe	2003	Developing	44.5	715.0	26	4.06	0.000000	7.0	998	...	7.0	6.52	68.0	36.7	453.351155	12633897.0	9.8
2935	Zimbabwe	2002	Developing	44.8	73.0	25	4.43	0.000000	73.0	304	...	73.0	6.53	71.0	39.8	57.348340	125525.0	1.2
2936	Zimbabwe	2001	Developing	45.3	686.0	25	1.72	0.000000	76.0	529	...	76.0	6.16	75.0	42.1	548.587312	12366165.0	1.6
2937	Zimbabwe	2000	Developing	46.0	665.0	24	1.68	0.000000	79.0	1483	...	78.0	7.10	78.0	43.5	547.358878	12222251.0	11.0

2938 rows × 22 columns

Rows: 2,938
Columns: 22

```
 #   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
```

Challenges faced in dataset

Country	Year	Status	Life expecti	Adult Mort	infant deat
Afghanista	2015	Developing	65	263	62
Afghanista	2014	Developing	59.9	271	64
Afghanista	2013	Developing	59.9	268	66
Afghanista	2012	Developing	59.5	272	69
Afghanista	2011	Developing	59.2	275	71
Afghanista	2010	Developing	58.8	279	74
Afghanista	2009	Developing	58.6	281	77
Afghanista	2008	Developing	58.1	287	80
Afghanista	2007	Developing	57.5	295	82
Afghanista	2006	Developing	57.3	295	84

```
Country                            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
```

There were several missing data values in dataset.

Not able to clearly see how each feature affected life expectancy data.

Not all data values were numerical.

Some of the data were in random sequence.

Approach taken to overcome challenges

- ➔ Use EDA to analyse and clean data, drop the data that has missing values.
- ➔ Use EDA to analyse and use scatter plot graphs, histogram and correlation matrix to show how each feature affected the life expectancy data.
- ➔ Use Label Encoding and Normalising for non-numerical data.
- ➔ Use Machine learning models such as Linear regression, Polynomial regression, Decision tree regression and Random forest regression for prediction in data.

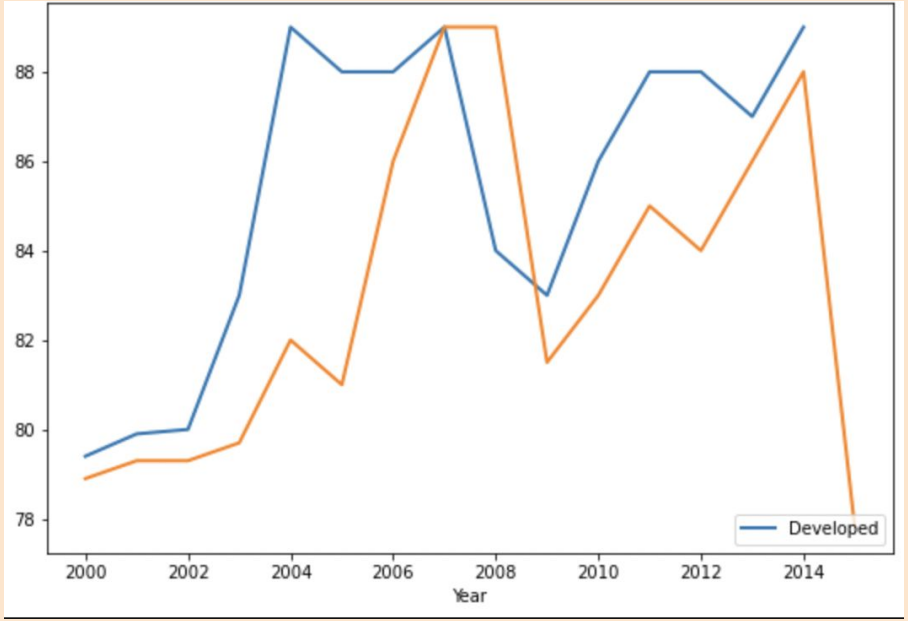
Problem Statement

Predicting Life Expectancy

- ★ Life expectancy means how long can a person live up to. It can be dependent upon his/her age, health, schooling, lifestyles and many more.
- ★ Aim of this project is to predict life expectancy using “Life Expectancy Data(WHO)” dataset.
- ★ By performing the use of Exploratory data analysis (EDA) and Machine learning models to predict life expectancy.

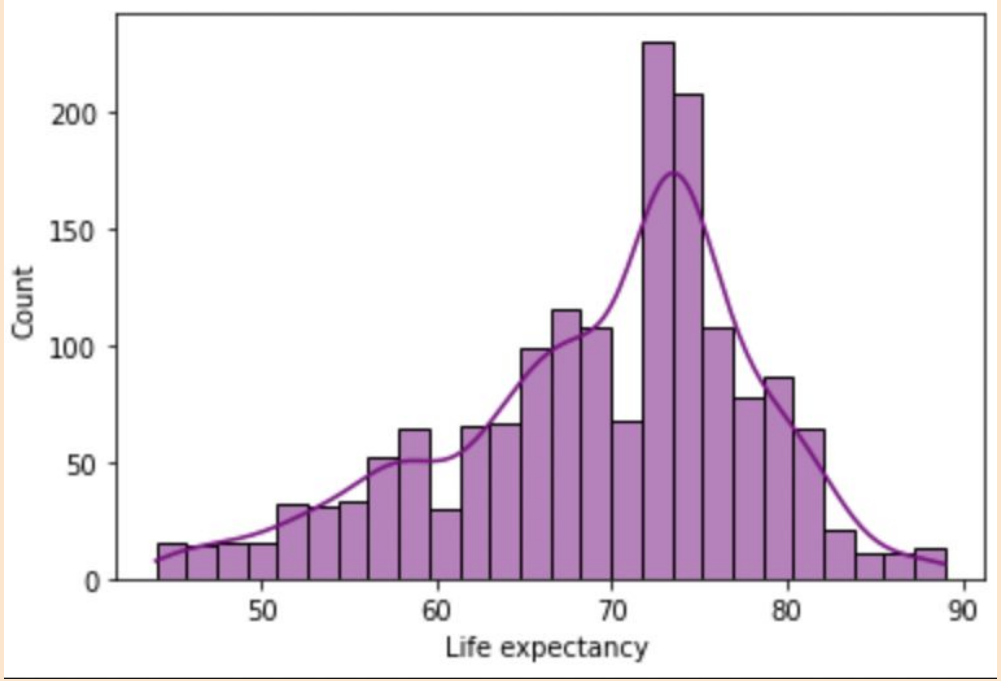
Life expectancy based on country status

Developed vs Developing



It is observed that Developed status had higher life expectancy compared to Developing status.

Distribution graph on Life Expectancy

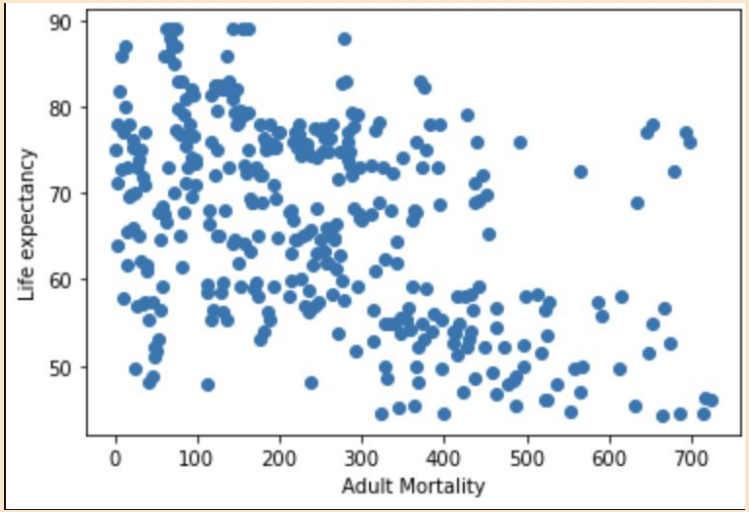


It is observed the life expectancy commonly ranges from 45 years to 90 years.

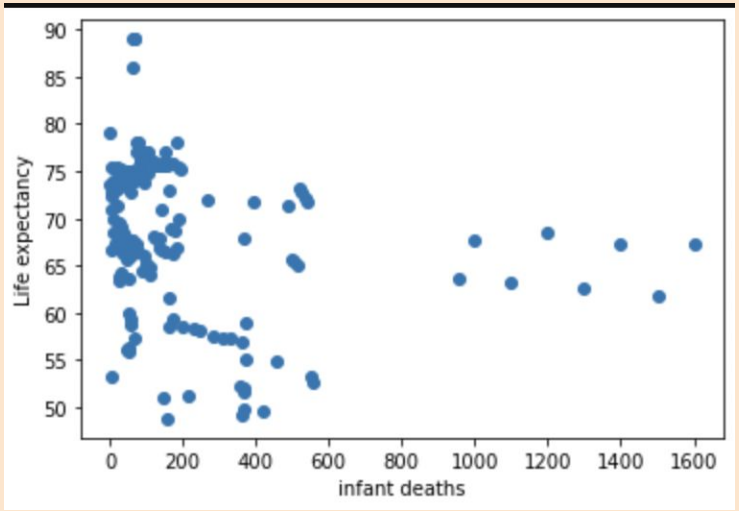
Scatter plot graph on Life Expectancy affected by Adult Mortality and Infant deaths

It is observed that Adult Mortality affect life expectancy more than infant deaths. Adult mortality also known as deaths caused by specific or general reasons.

Adult mortality



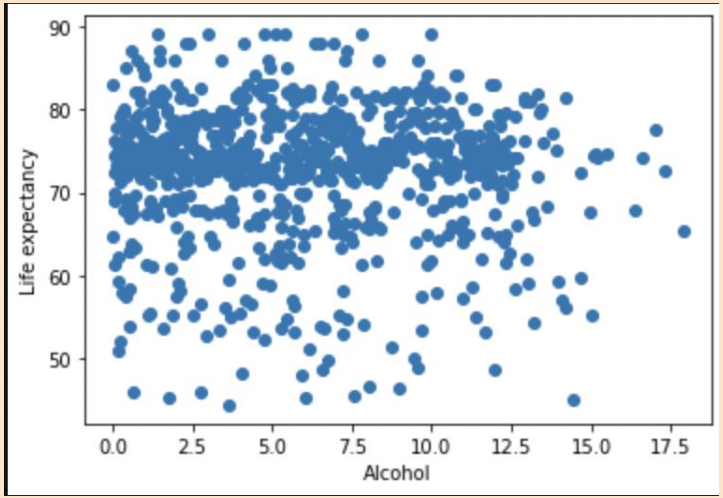
Infant deaths



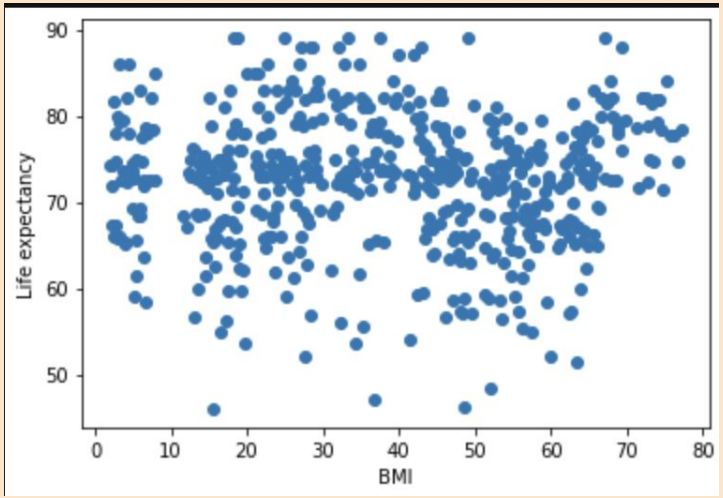
Scatter plot graph on Life Expectancy affected by Alcohol and BMI

It is observed that Alcohol affect life expectancy more than BMI.

Alcohol



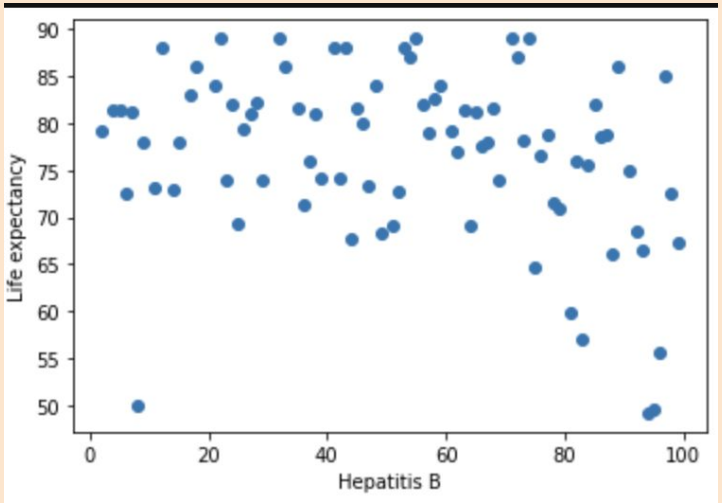
BMI



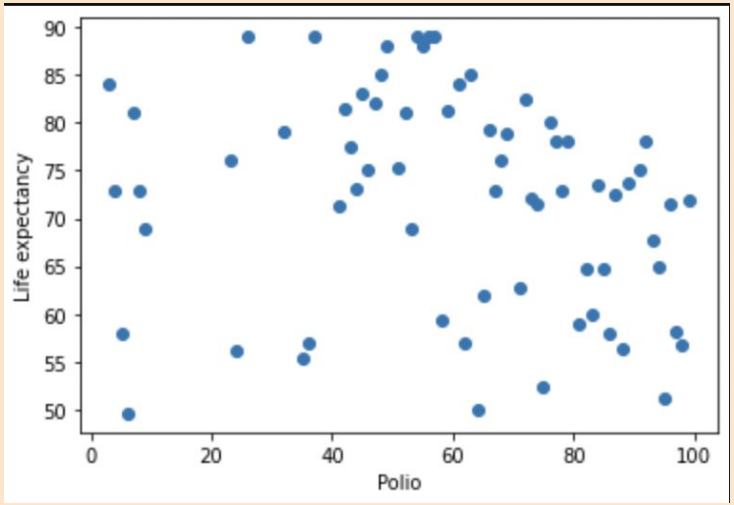
Scatter plot graph on Life Expectancy affected by Polio and Hepatitis B

It is observed that Hepatitis B affect life expectancy more than Polio. However, there is not much vast difference.

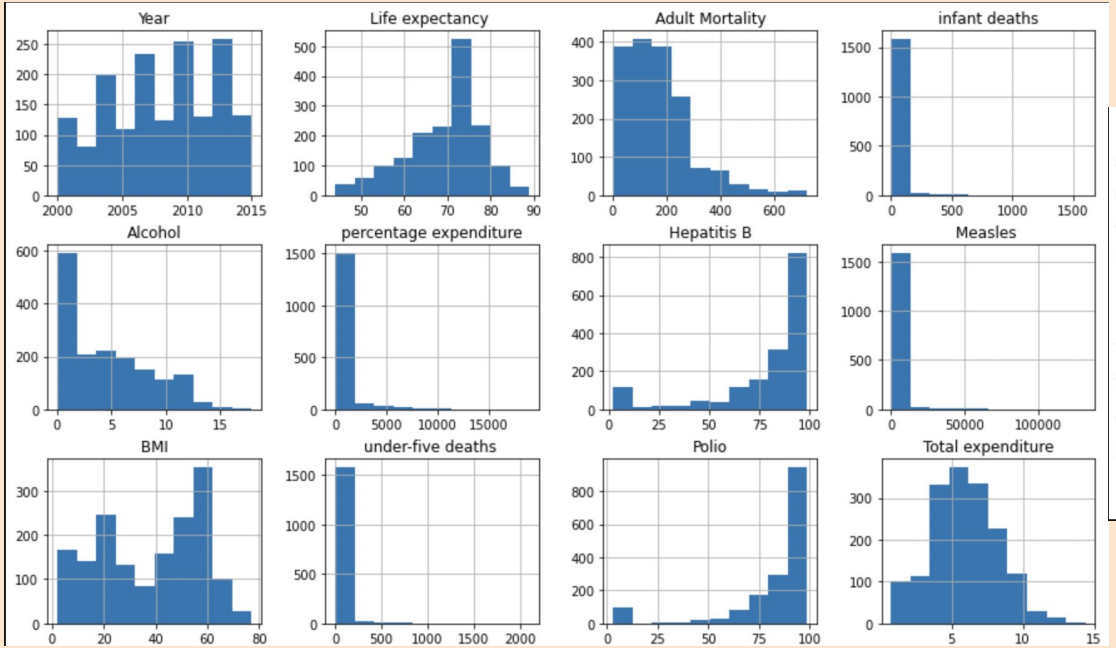
Hepatitis B

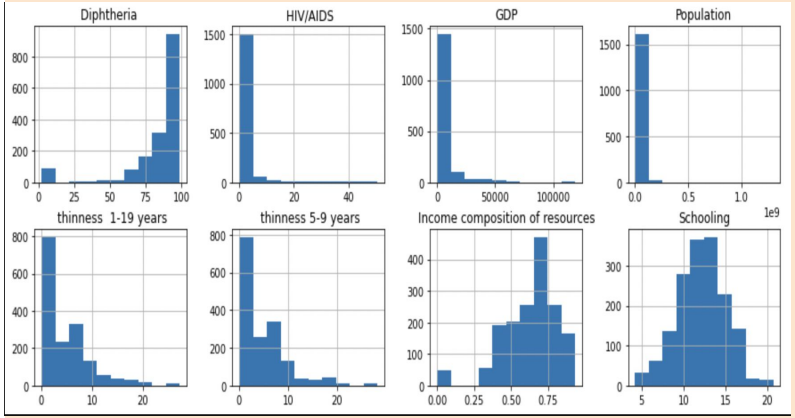


Polio

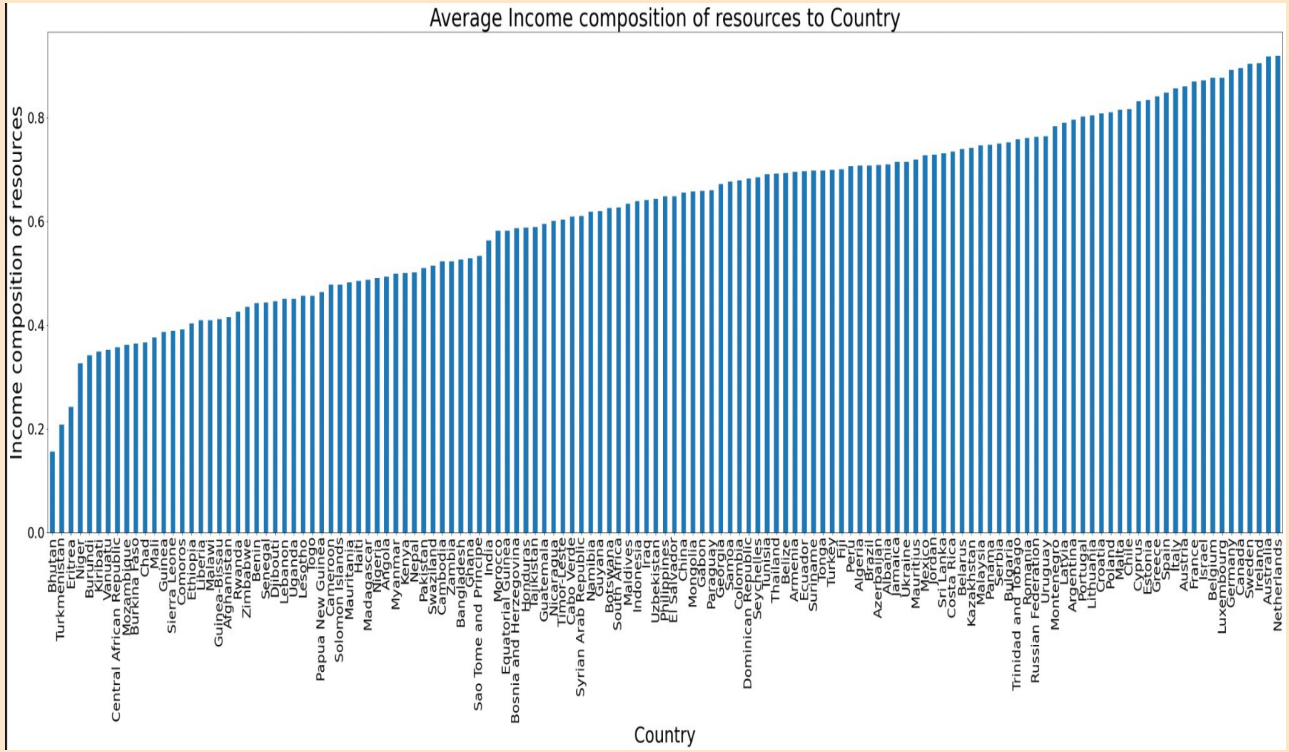


Overview of histogram graphs showing all features



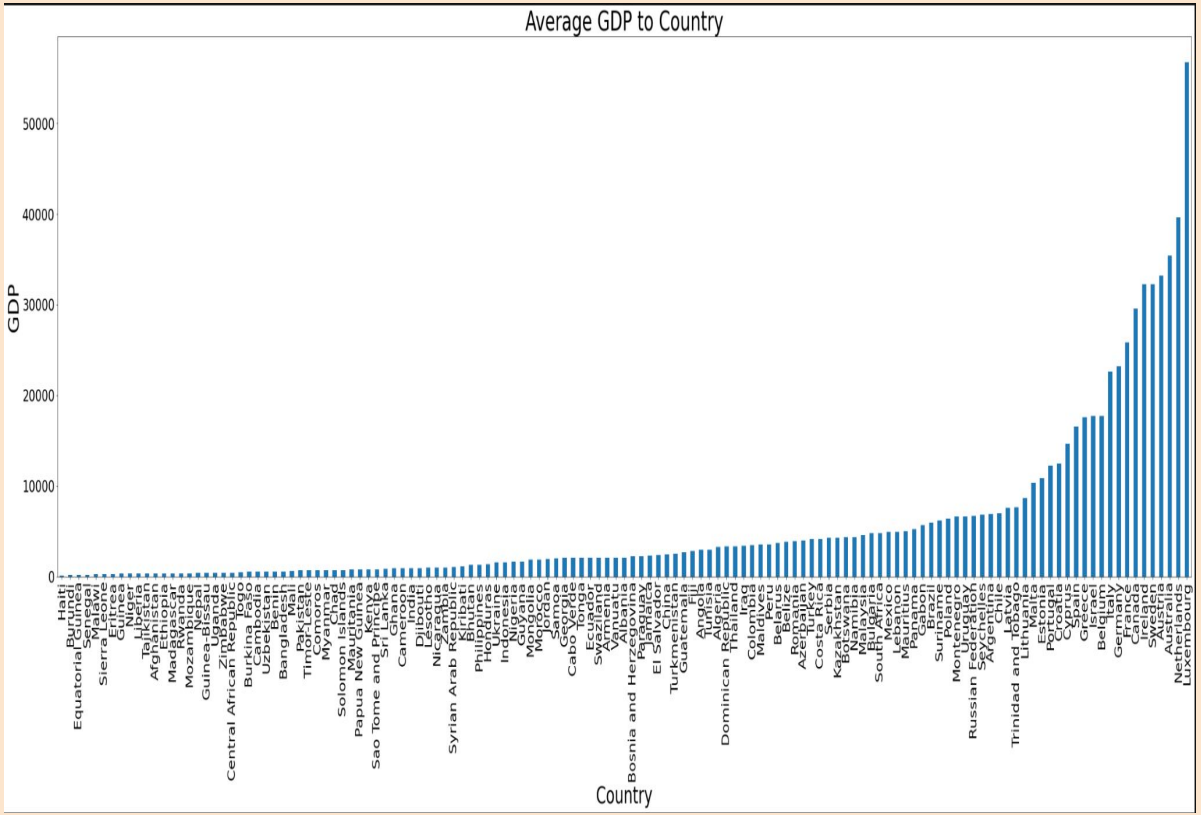


Plot graph for Income composition of resources in each country



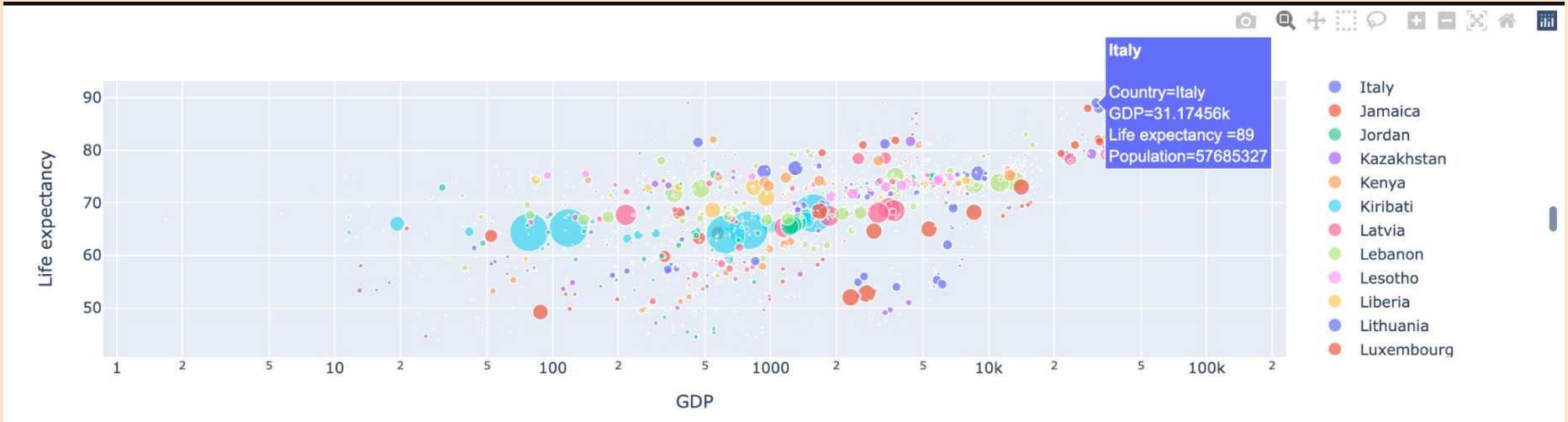
It is observed that countries such as Netherlands, Australia and Ireland have highest income composition resources.

Plot graph for Average GDP in each country



It is observed that countries such as Luxembourg, Netherlands, Australia and Ireland have highest average GDP (gross domestic product).

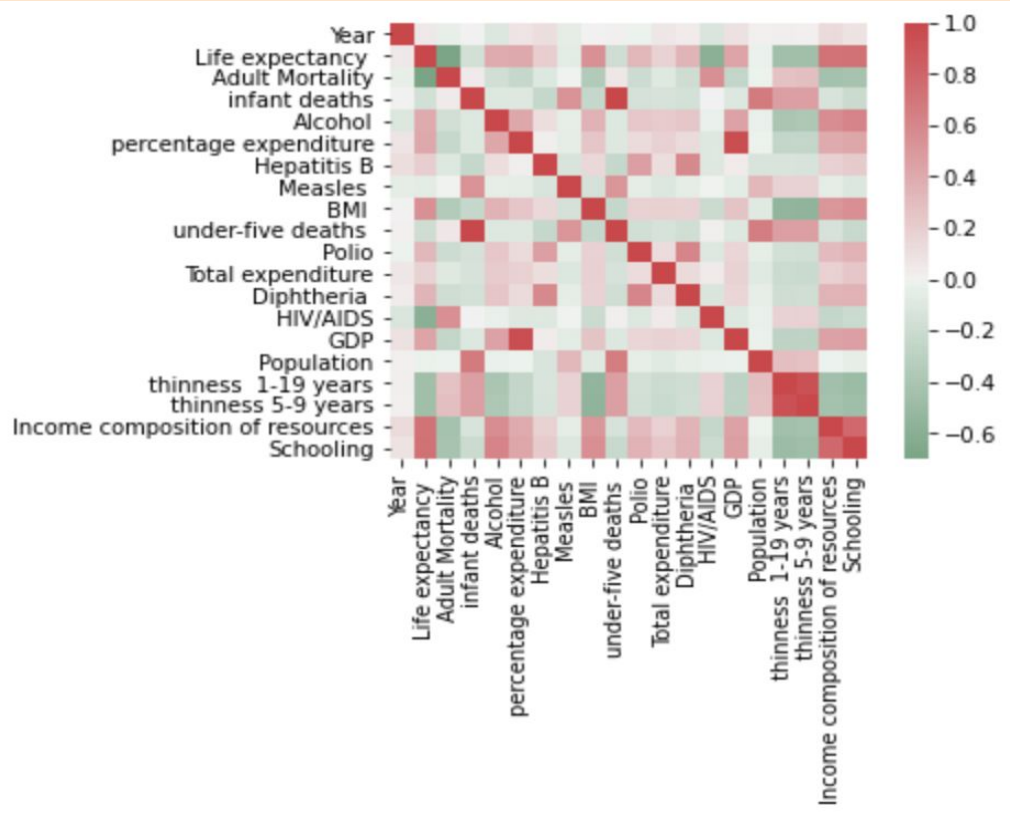
Plotly bubble chart for life expectancy and GDP in each country



E.g Show Italy

It is observed that countries such as Italy, France, Spain, Netherlands, Australia and Ireland have highest GDP and highest life expectancy. The bubble size indicates the population size in each country.

Correlation heatmap of all features affecting life expectancy



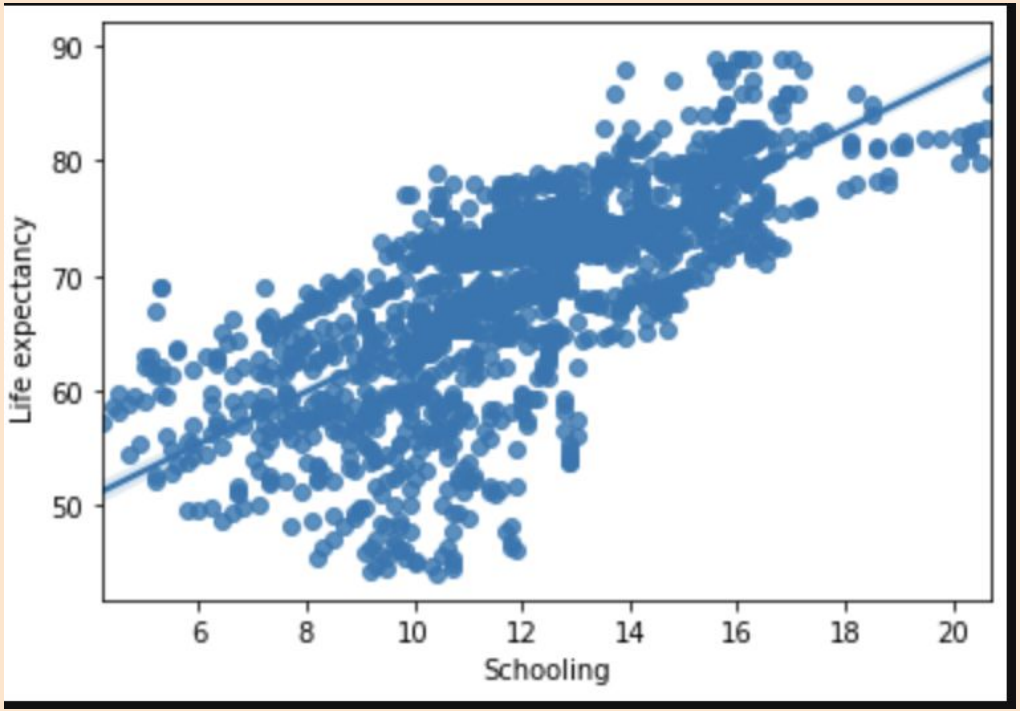
It is observed that **schooling** has highest positive correlation with life expectancy. Perhaps, this shows that schooling is very important to cultivate proper education to lead safe and healthy lifestyles which may help to increase life expectancy rate.

Use of label encoding and normalising data

For those non numerical data

	Year	Status	Life_expectancy_	Adult_Mortality	infant_deaths	Alcohol	percentage_expenditure	Hepatitis_B	Measles_	_BMI_	...	Polio	Total_expenditure	Diphtheria_	_HIV/AID
0	2015	0	0.730337	0.363762	0.038750	0.000560	0.003759	0.656566	0.008780	0.247730	...	0.060606	0.567060	0.656566	0.00197
1	2014	0	0.673034	0.374827	0.040000	0.000560	0.003878	0.626263	0.003743	0.241245	...	0.585859	0.568450	0.626263	0.00197
2	2013	0	0.673034	0.370678	0.041250	0.000560	0.003861	0.646465	0.003271	0.234760	...	0.626263	0.564976	0.646465	0.00197
3	2012	0	0.668539	0.376210	0.043125	0.000560	0.004123	0.676768	0.021203	0.228275	...	0.676768	0.592078	0.676768	0.00197
4	2011	0	0.665169	0.380360	0.044375	0.000560	0.000374	0.686869	0.022923	0.223087	...	0.686869	0.546908	0.686869	0.00197
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
2933	2004	0	0.497753	1.000000	0.016875	0.243984	0.000000	0.686869	0.000236	0.351492	...	0.676768	0.495483	0.656566	0.66403
2934	2003	0	0.500000	0.988935	0.016250	0.227196	0.000000	0.070707	0.007593	0.346304	...	0.070707	0.453092	0.686869	0.72529
2935	2002	0	0.503371	0.100968	0.015625	0.247902	0.000000	0.737374	0.002313	0.341115	...	0.737374	0.453787	0.717172	0.7865
2936	2001	0	0.508989	0.948824	0.015625	0.096251	0.000000	0.767677	0.004025	0.335927	...	0.767677	0.428075	0.757576	0.8320
2937	2000	0	0.516854	0.919779	0.015000	0.094012	0.000000	0.797980	0.011283	0.330739	...	0.787879	0.493398	0.787879	0.85968

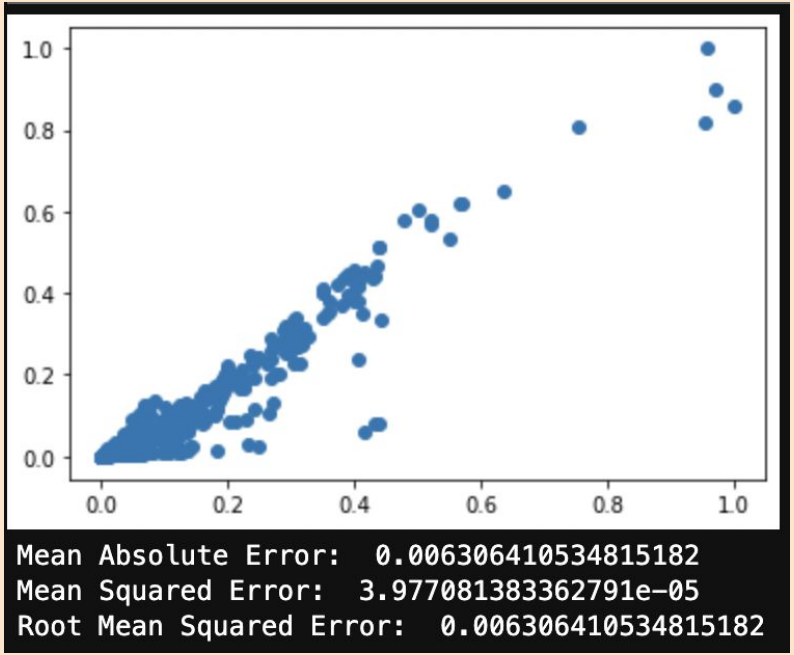
Regression plot of schooling with life expectancy



It is observed that schooling has a very close-knitted relationship with life expectancy.

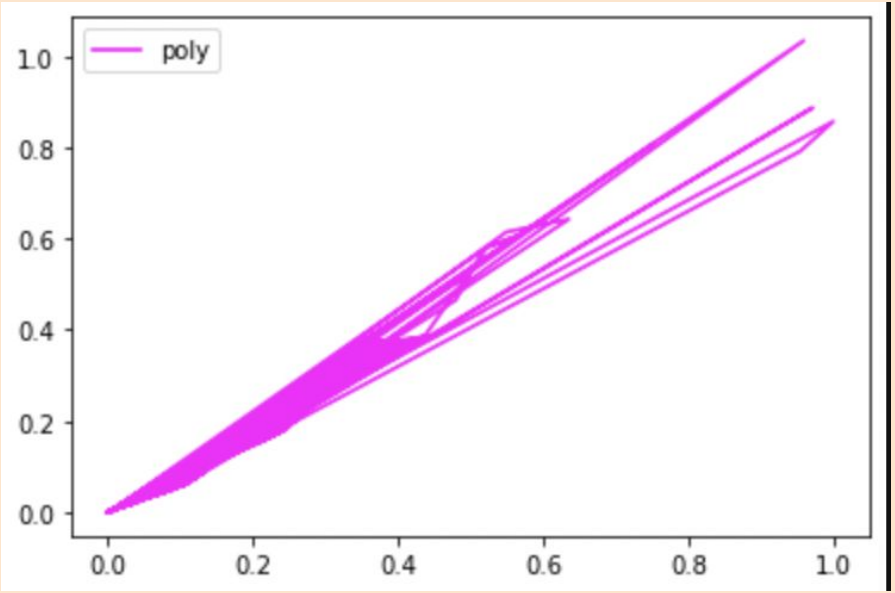
Linear Regression model using GDP, year, life expectancy and percentage expenditure

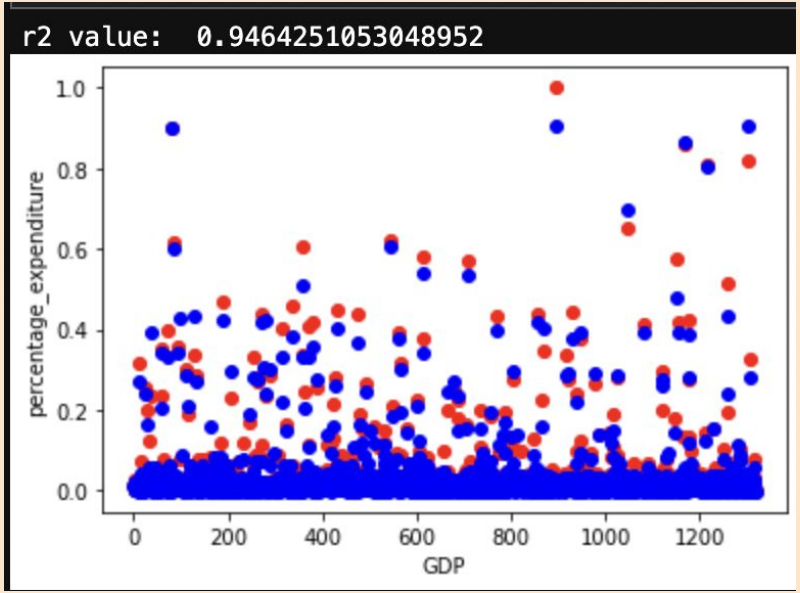
Description	Value
Linear regression coefficient	0.96
Mean absolute error	0.0063
Mean squared error	3.977
Root Mean squared error	0.0063



It is observed that the coefficient value closer to 1, indicate a more optimal prediction.

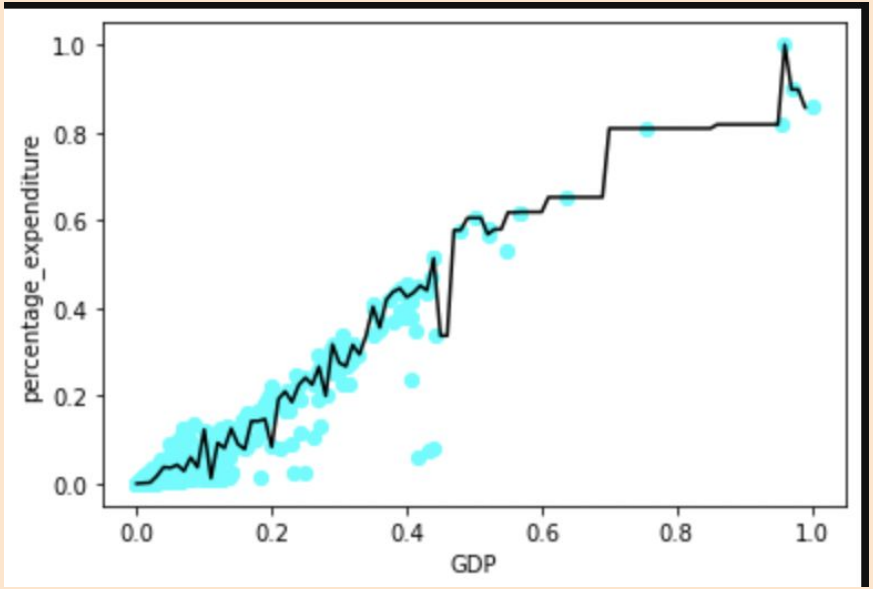
Polynomial Regression model using GDP and percentage expenditure





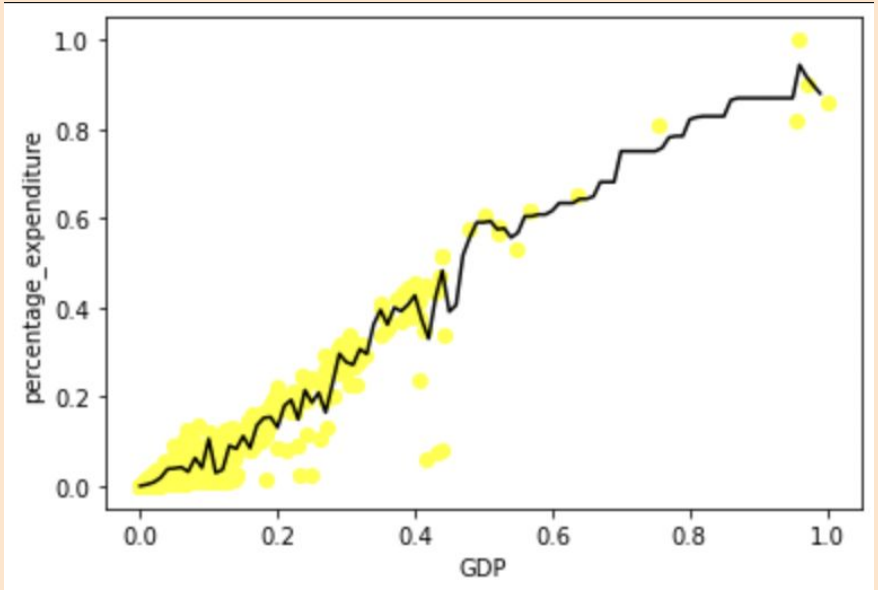
Description	Value
r2	0.95

Decision Tree Regression model using GDP and percentage expenditure



Description	Value
Prediction	0.86

Random Forest Regression model using GDP and percentage expenditure



Description	Value
Prediction	0.88

Conclusion

To conclude, it has been observed from this project, how the different features affect the life expectancy. From the correlation heatmap, it was seen that schooling has the highest correlation with life expectancy. Then, from the machine learning models applied, it was seen that Linear Regression provided a more accurate prediction value that was closer to 1.

However, certain features were not included in the dataset such as environmental factors that may affect life expectancy. Thus, the overall analysis in this project may not be wholesome to make the best accurate prediction yet.

Resources

- https://www.kaggle.com/code/mrinath/starter-life-expectancy-who-9536c272-3/data
- Medium.com
- Towardsdatascience.com
- Kaggle.com
- Stackoverflow,com
- github.com

End