DATA ANALYTICS

Final Review

DataStorm

PROBLEM STATEMENT

Analysing various effect of numerous factors like health, demographic and other environmental factors like economic status of country on Life Expectancy.

IMPORTANCE

Life expectancy is a common performance measure which is used to compare not only the living conditions of individuals, but also the health facilities and economic conditions of countries.

This study is aimed at investigating various determinants of life expectancy, and providing a comparative analysis between countries, so as to aid policymakers in allocating the countries' resources appropriately, and improve overall standard of living.

DATASET

The dataset for our project has been majorly collected from Kaggle, containing life expectancy of 193 countries across the globe collected from WHO data repository with immunization, mortality, economic and social factors under consideration.
https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who
(Dataset-1)
This was also augmented by another source,
https://data.worldbank.org/indicator/SH.XPD.CHEX.GD.ZS (Dataset-2)

PLAN AND PROCESS

WORKFLOW

- Analyse the data and obtain the important features that majorly affect LE (feature selection in general)
- Build appropriate models to predict the life expectancy like Multiple Linear Regression (MLR), XGBoost, Random Forest etc.
- Perform status wise analysis(developed,developing), and compare the factors which majorly affect each category.
- Perform country wise analysis of the top and bottom countries.

EDA

DATASET DESCRIPTION

- Year ranging from 2000-2015
- Country (193)
- Population - Population of the country
- Life Expectancy -the average period that a person may expect to live(age)
- Status - Developed / Developing
- Alcohol - recorded per capita (15+) consumption (in litres of pure alcohol)
- 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
- Under five Deaths- Number of under-five deaths per 1000 population
- Hepatitis B - Hepatitis B (HepB) immunization coverage among 1-year-olds (%)
- Measles - number of reported cases per 1000 population
- Polio - Polio (Pol3) immunization coverage among 1-year-olds (%)

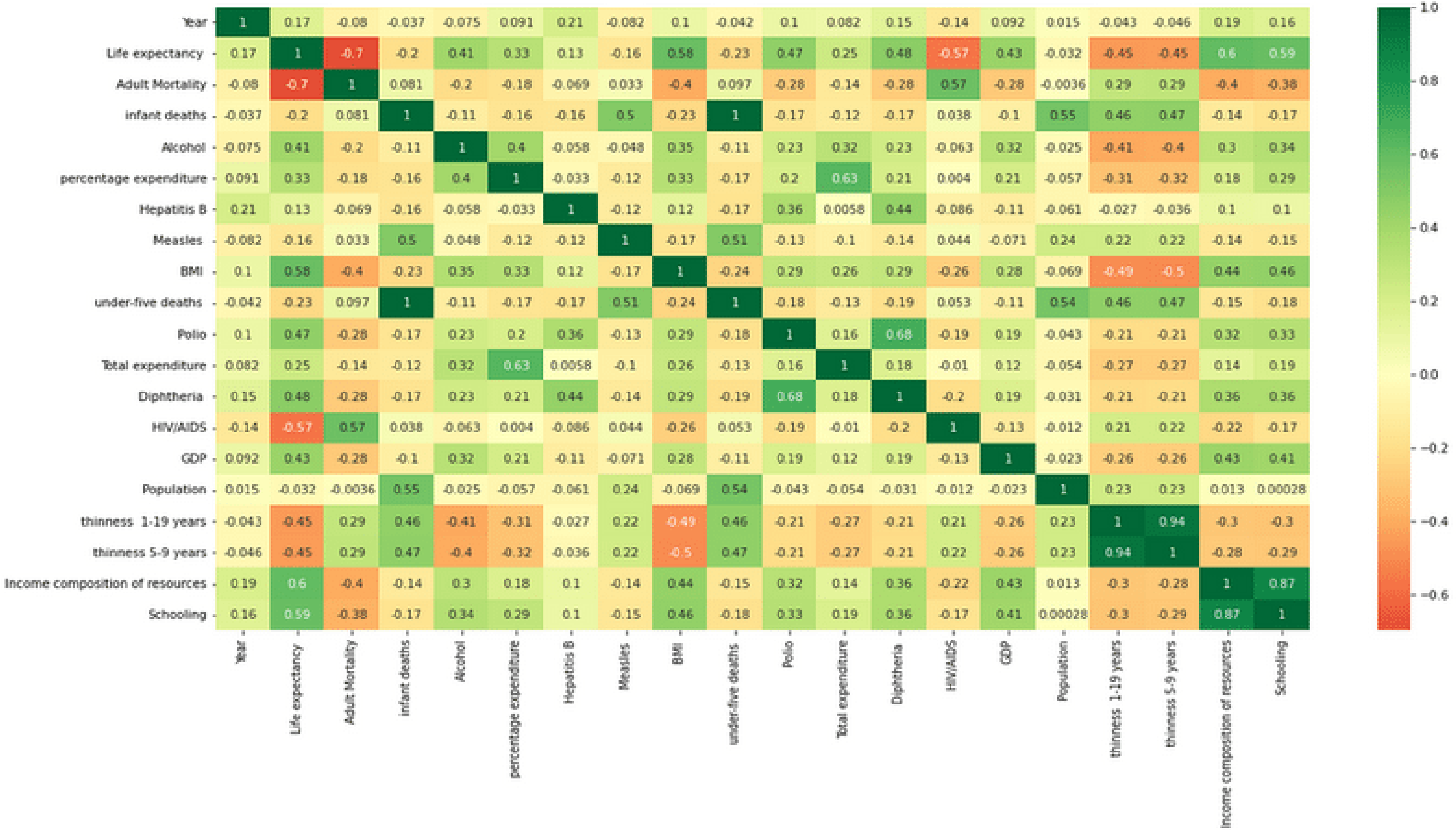
DATASET DESCRIPTION

- 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)
- Thinness 1-19 years - Prevalence of thinness among children and adolescents for Age 10 to 19 (%)
- Thinness 1-5 years - Prevalence of thinness among children for Age 5 to 9(%)
- BMI - Average Body Mass Index of entire population
- GDP - Gross Domestic Product per capita (in USD)
- Total Expenditure - General government expenditure on health as a percentage of total government expenditure (%)
- Percentage Expenditure - Expenditure on health as a percentage of Gross Domestic Product per capita(%)
- 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)

MISSING VALUES AND IMPUTATION

- The missing values in various columns were identified and were replaced with appropriate imputations.
- There were few (10) missing values in the life expectancy (dependent variable) column, due to which the corresponding rows were dropped.
- As the dataset contains the data ordered from 2015 to 2000 for each country, the missing values were replaced using the method of forward and backward filling country wise.
- For countries for which the entire value set of an attribute was missing, zero imputation was done.
- Since the percentage expenditure attribute had inaccurate and inconsistent data in the Kaggle dataset, this column was replaced by the values from Dataset-2 for each country and year respectively.

HEATMAP REGARDING CORRELATION AMONGST FEATURES

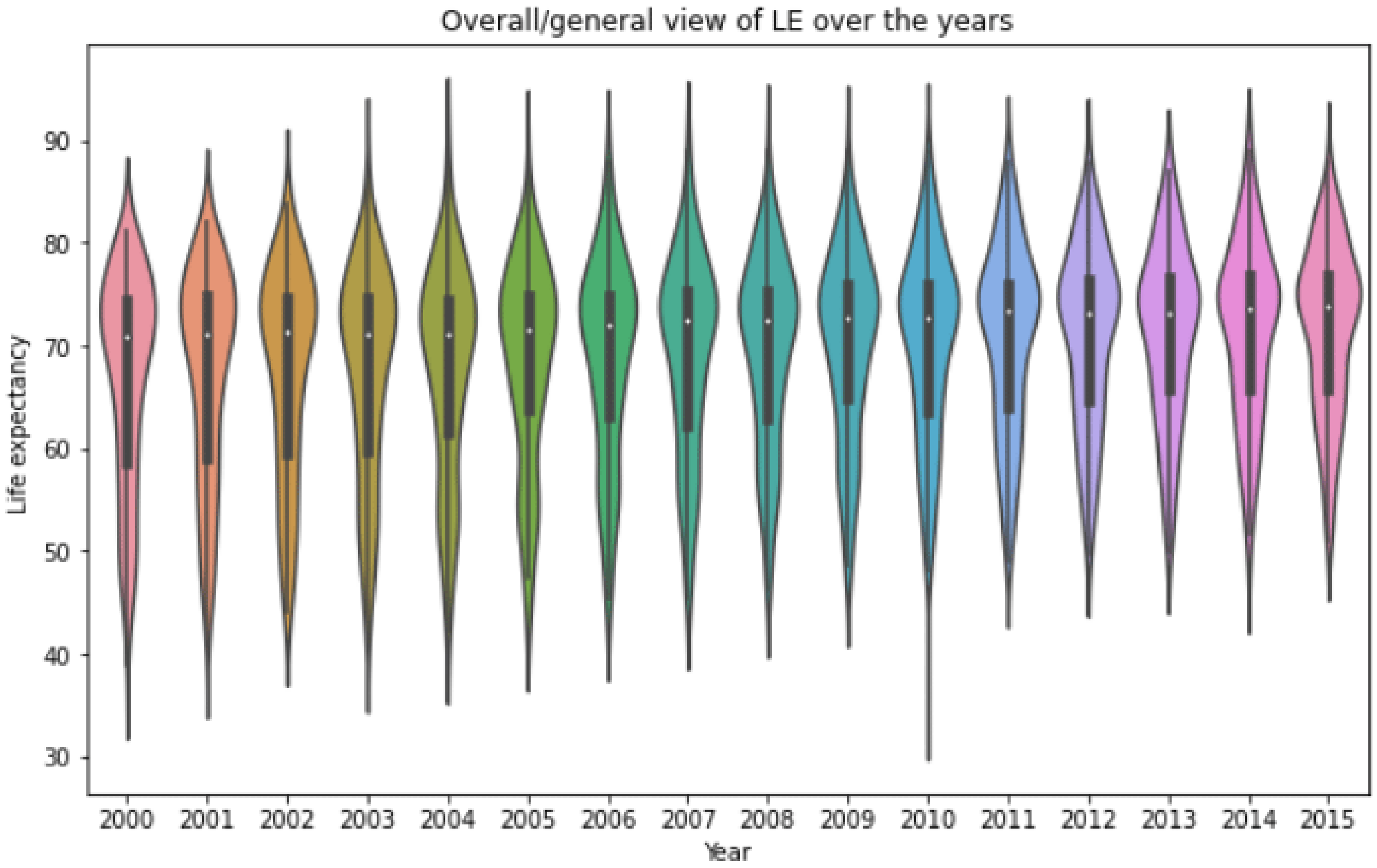


	Year	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	BMI	under-five deaths	Polio	Total expenditure	Diphtheria	HIV/AIDS	GDP	Population	thinness 1-19 years	thinness 5-9 years	Income composition of resources	Schooling
Year	1	0.17	-0.08	-0.037	-0.075	0.091	0.21	-0.082	0.1	-0.042	0.1	0.082	0.15	-0.14	0.092	0.015	-0.043	-0.046	0.19	0.16
Life expectancy	0.17	1	-0.7	-0.2	0.41	0.33	0.13	-0.16	0.58	-0.23	0.47	0.25	0.48	-0.57	0.43	-0.032	-0.45	-0.45	0.6	0.59
Adult Mortality	-0.08	-0.7	1	0.081	-0.2	-0.18	-0.069	0.033	-0.4	0.097	-0.28	-0.14	-0.28	0.57	-0.28	-0.0036	0.29	0.29	-0.4	-0.38
infant deaths	-0.037	-0.2	0.081	1	-0.11	-0.16	-0.16	0.5	-0.23	1	-0.17	-0.12	-0.17	0.038	-0.1	0.55	0.46	0.47	-0.14	-0.17
Alcohol	-0.075	0.41	-0.2	-0.11	1	0.4	-0.058	-0.048	0.35	-0.11	0.23	0.32	0.23	-0.063	0.32	-0.025	-0.41	-0.4	0.3	0.34
percentage expenditure	0.091	0.33	-0.18	-0.16	0.4	1	-0.033	-0.12	0.33	-0.17	0.2	0.63	0.21	0.004	0.21	-0.057	-0.31	-0.32	0.18	0.29
Hepatitis B	0.21	0.13	-0.069	-0.16	-0.058	-0.033	1	-0.12	0.12	-0.17	0.36	0.0058	0.44	-0.086	-0.11	-0.061	-0.027	-0.036	0.1	0.1
Measles	-0.082	-0.16	0.033	0.5	-0.048	-0.12	-0.12	1	-0.17	0.51	-0.13	-0.1	-0.14	0.044	-0.071	0.24	0.22	0.22	-0.14	-0.15
BMI	0.1	0.58	-0.4	-0.23	0.35	0.33	0.12	-0.17	1	-0.24	0.29	0.26	0.29	-0.26	0.28	-0.069	-0.49	-0.5	0.44	0.46
under-five deaths	-0.042	-0.23	0.097	1	-0.11	-0.17	-0.17	0.51	-0.24	1	-0.18	-0.13	-0.19	0.053	-0.11	0.54	0.46	0.47	-0.15	-0.18
Polio	0.1	0.47	-0.28	-0.17	0.23	0.2	0.36	-0.13	0.29	-0.18	1	0.16	0.68	-0.19	0.19	-0.043	-0.21	-0.21	0.32	0.33
Total expenditure	0.082	0.25	-0.14	-0.12	0.32	0.63	0.0058	-0.1	0.26	-0.13	0.16	1	0.18	-0.01	0.12	-0.054	-0.27	-0.27	0.14	0.19
Diphtheria	0.15	0.48	-0.28	-0.17	0.23	0.21	0.44	-0.14	0.29	-0.19	0.68	0.18	1	-0.2	0.19	-0.031	-0.21	-0.21	0.36	0.36
HIV/AIDS	-0.14	-0.57	0.57	0.038	-0.063	0.004	-0.086	0.044	-0.26	0.053	-0.19	-0.01	-0.2	1	-0.13	-0.012	0.21	0.22	-0.22	-0.17
GDP	0.092	0.43	-0.28	-0.1	0.32	0.21	-0.11	-0.071	0.28	-0.11	0.19	0.12	0.19	-0.13	1	-0.023	-0.26	-0.26	0.43	0.41
Population	0.015	-0.032	-0.0036	0.55	-0.025	-0.057	-0.061	0.24	-0.069	0.54	-0.043	-0.054	-0.031	-0.012	-0.023	1	0.23	0.23	0.013	0.00028
thinness 1-19 years	-0.043	-0.45	0.29	0.46	-0.41	-0.31	-0.027	0.22	-0.49	0.46	-0.21	-0.27	-0.21	0.21	-0.26	0.23	1	0.94	-0.3	-0.3
thinness 5-9 years	-0.046	-0.45	0.29	0.47	-0.4	-0.32	-0.036	0.22	-0.5	0.47	-0.21	-0.27	-0.21	0.22	-0.26	0.23	0.94	1	-0.28	-0.29
Income composition of resources	0.19	0.6	-0.4	-0.14	0.3	0.18	0.1	-0.14	0.44	-0.15	0.32	0.14	0.36	-0.22	0.43	0.013	-0.3	-0.28	1	0.87
Schooling	0.16	0.59	-0.38	-0.17	0.34	0.29	0.1	-0.15	0.46	-0.18	0.33	0.19	0.36	-0.17	0.41	0.00028	-0.3	-0.29	0.87	1

CORRELATION ANALYSIS AND REDUNDANT ATTRIBUTES

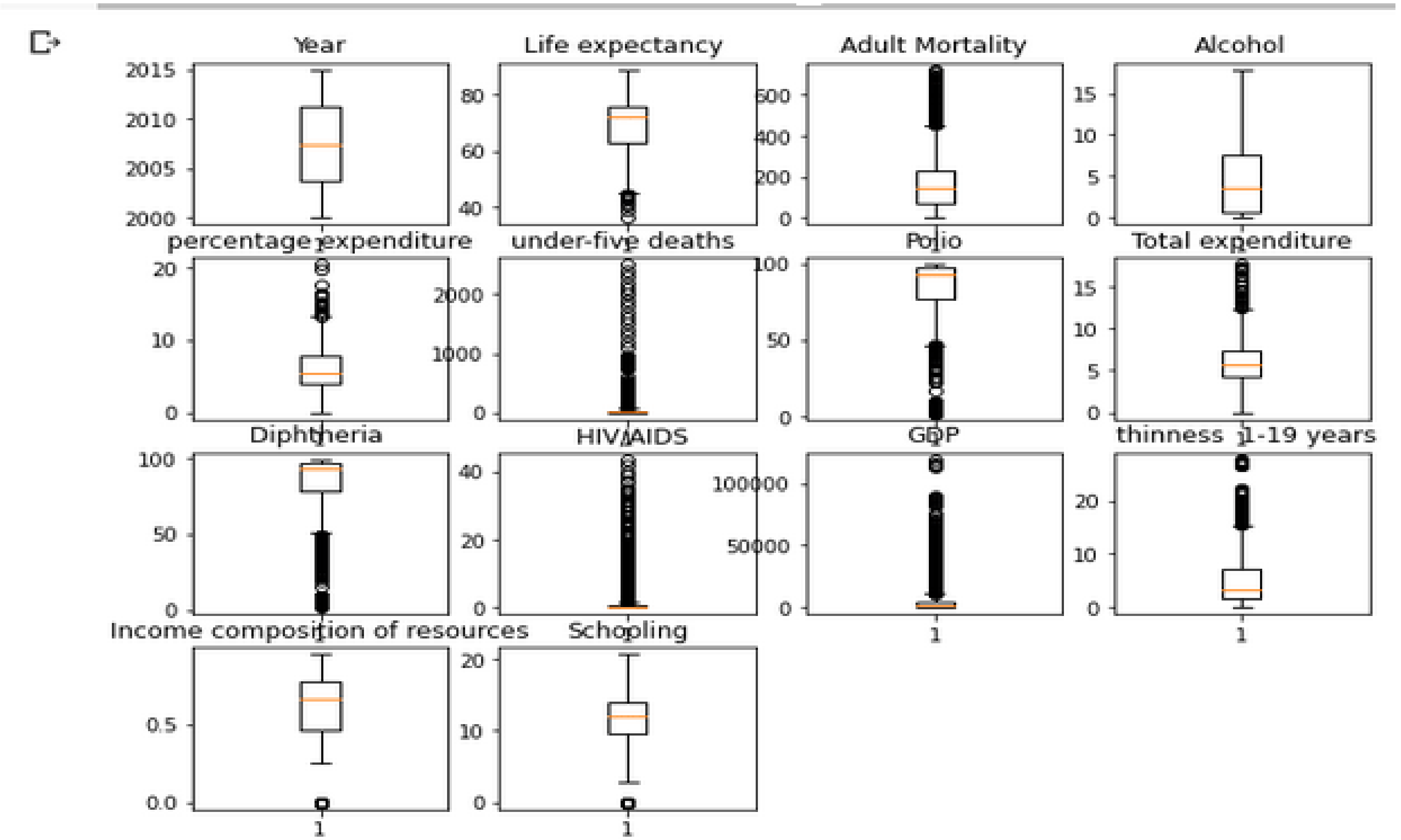
- Life expectancy has barely any correlation to the population, but is highly correlated with adult mortality, BMI, HIV and income composition of resources.
- There are some redundant features which can be avoided for further analysis.
- Correlation between under-5 deaths and infant deaths is 1 (very high correlation), leading to the conclusion that **infant deaths** can be dropped.
- The LE of a country depends on the adequate distribution of quality resources irrespective of population, which has low correlation (r= -0.032) with LE in the heatmap, owing to which **population** has been dropped.
- Thinness 1-19 is a superset of **thinness 5-9 years**, and thus the latter can be dropped.

INFERENCES FROM EDA



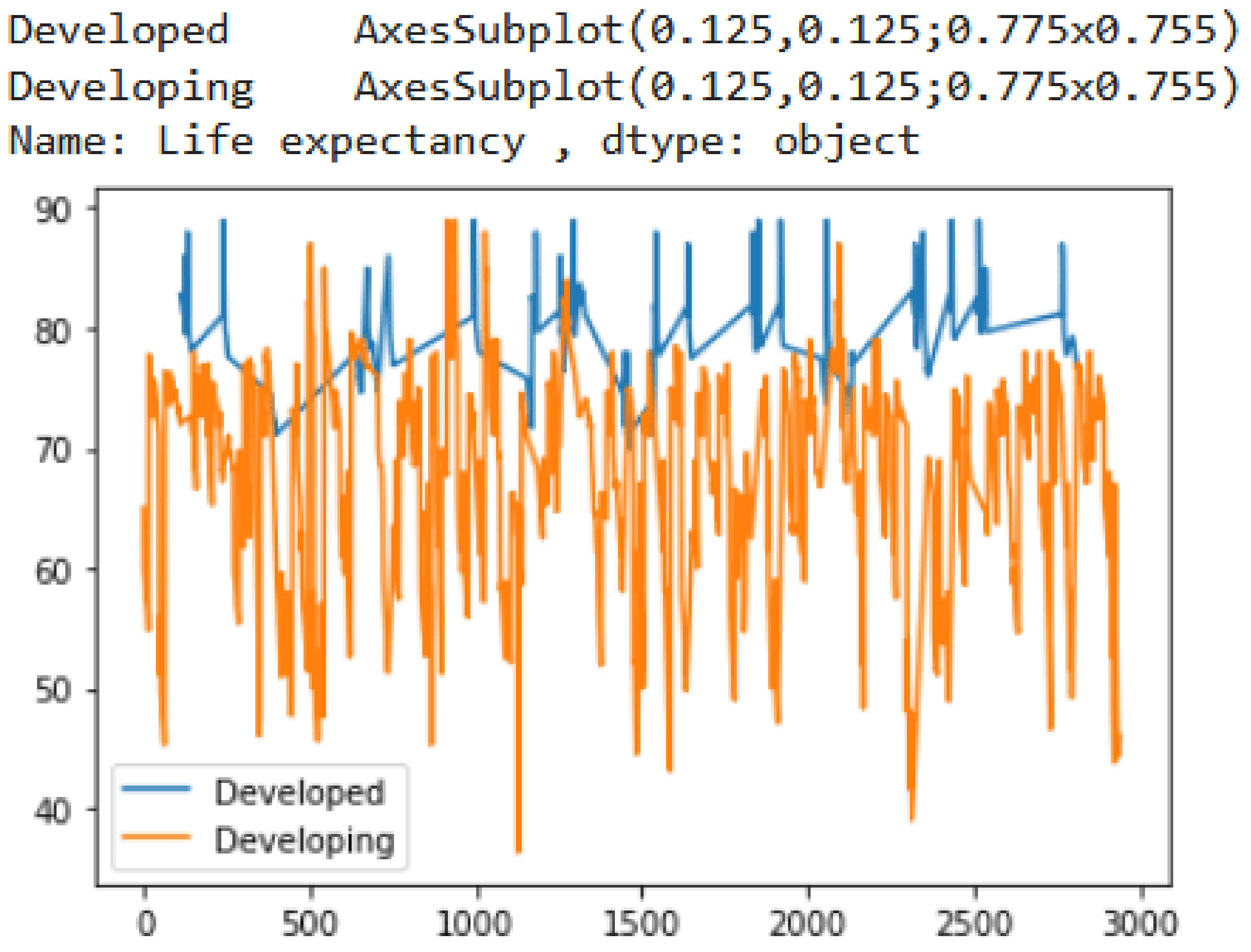
The same inference can be also viewed and inferred from the violin plot, its observed that median of life expectancy has slightly increased over the years and is more concentrated around 70.

Outlier Analysis:



The outlier plots show that the columns like HIV/AIDS, GDP etc have outliers. We decided not to drop the outliers as their presence might be due to the variation of count in each country based on their population.

Visualizing life expectancy for developing and developed countries



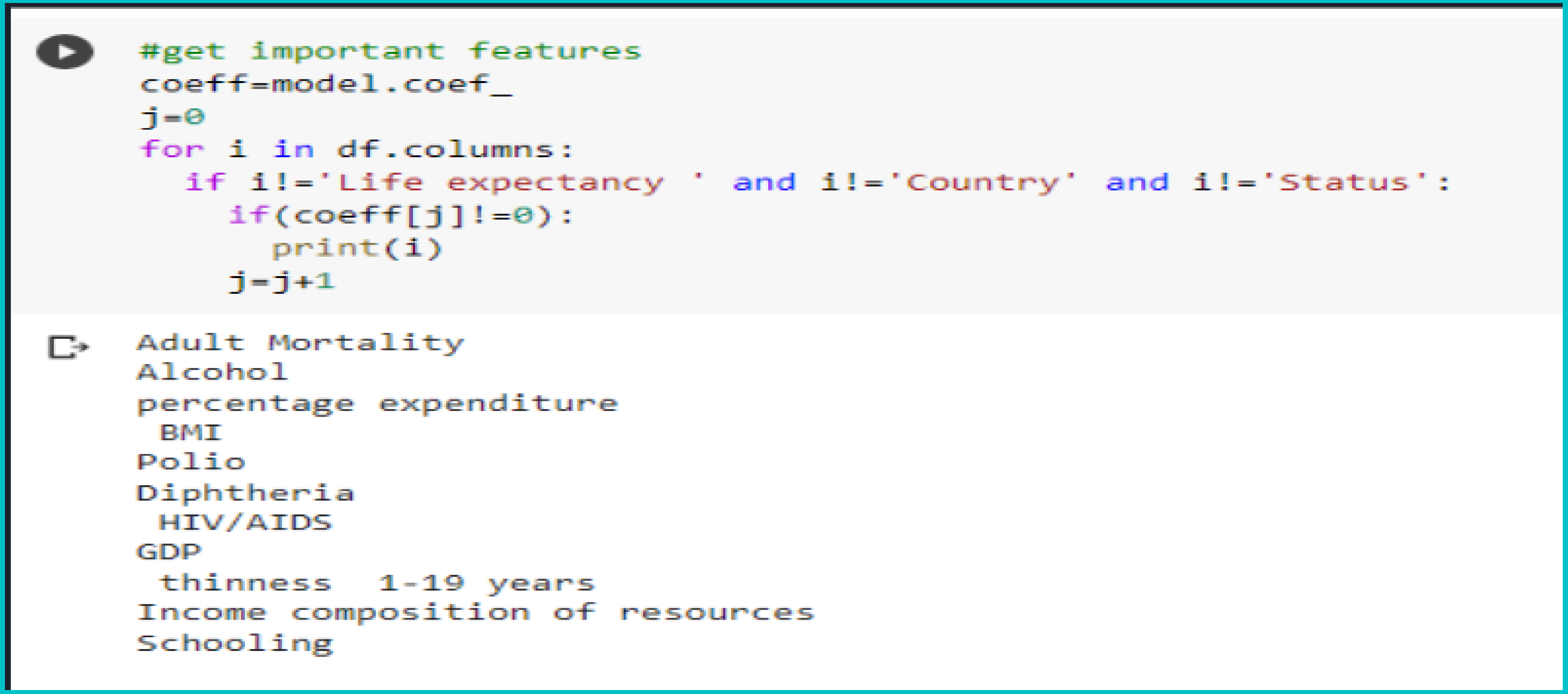
```
Developed     AxesSubplot(0.125,0.125;0.775x0.755)
Developing    AxesSubplot(0.125,0.125;0.775x0.755)
Name: Life expectancy , dtype: object
```

Life expectancy of developed countries takes a higher stand over developing countries.

FEATURE SELECTION & MODEL BUILDING

Lasso Regression:

We tried lasso regression for feature extraction where unimportant features were recognized, whose lasso coefficients were zeroes. Other features like year, under-5 deaths, hepatitis B, measles, total expenditure were deemed to be less prominent.
This model resulted in an r2 score of 0.7442, and an RMSE of 4.867



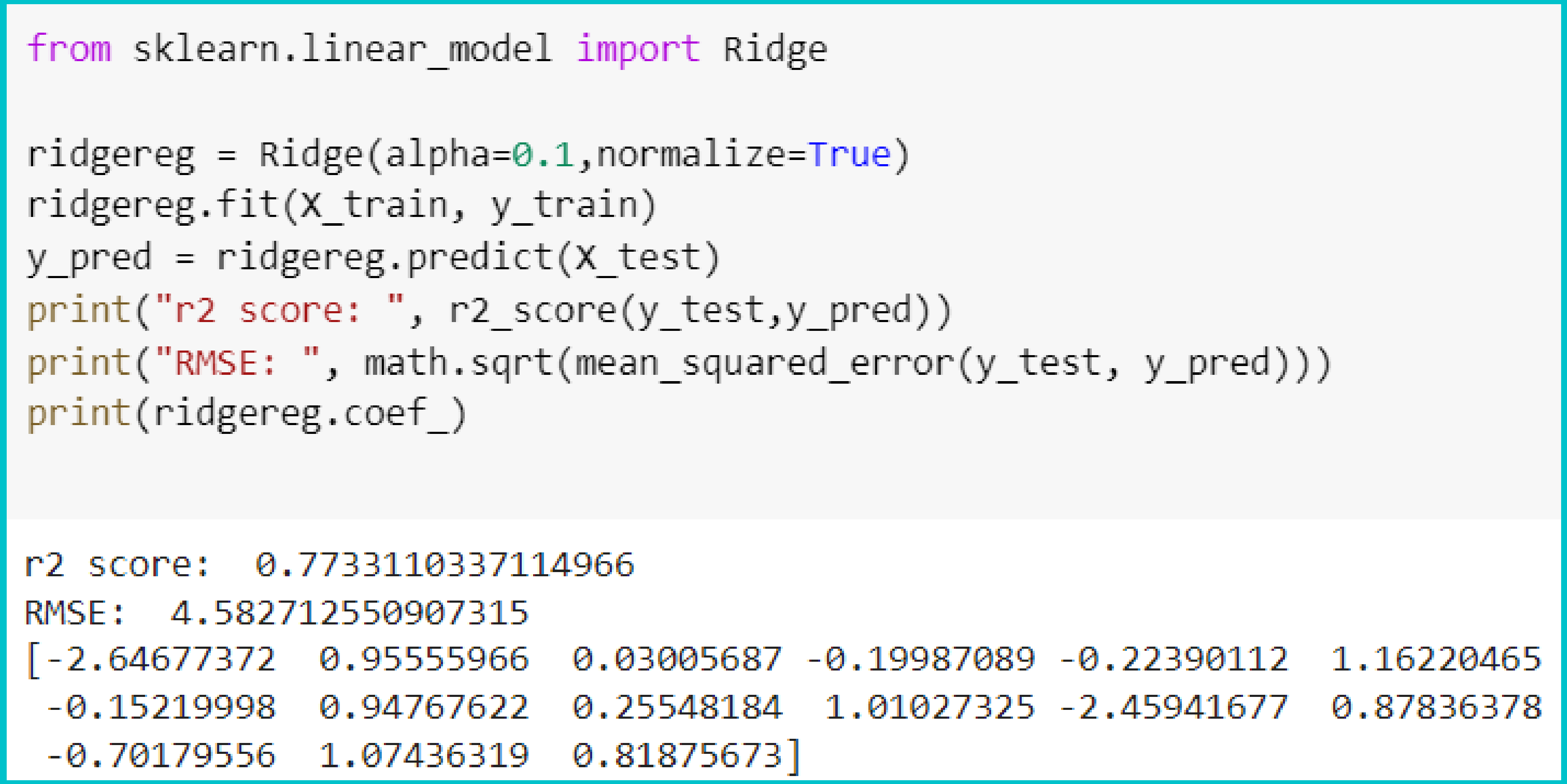
```
#get important features
coeff=model.coef_
j=0
for i in df.columns:
  if i!='Life expectancy ' and i!='Country' and i!='Status':
    if(coeff[j]!=0):
      print(i)
    j=j+1
```

```
Adult Mortality
Alcohol
percentage expenditure
 BMI
Polio
Diphtheria
 HIV/AIDS
GDP
 thinness  1-19 years
Income composition of resources
Schooling
```

Ridge Regression:

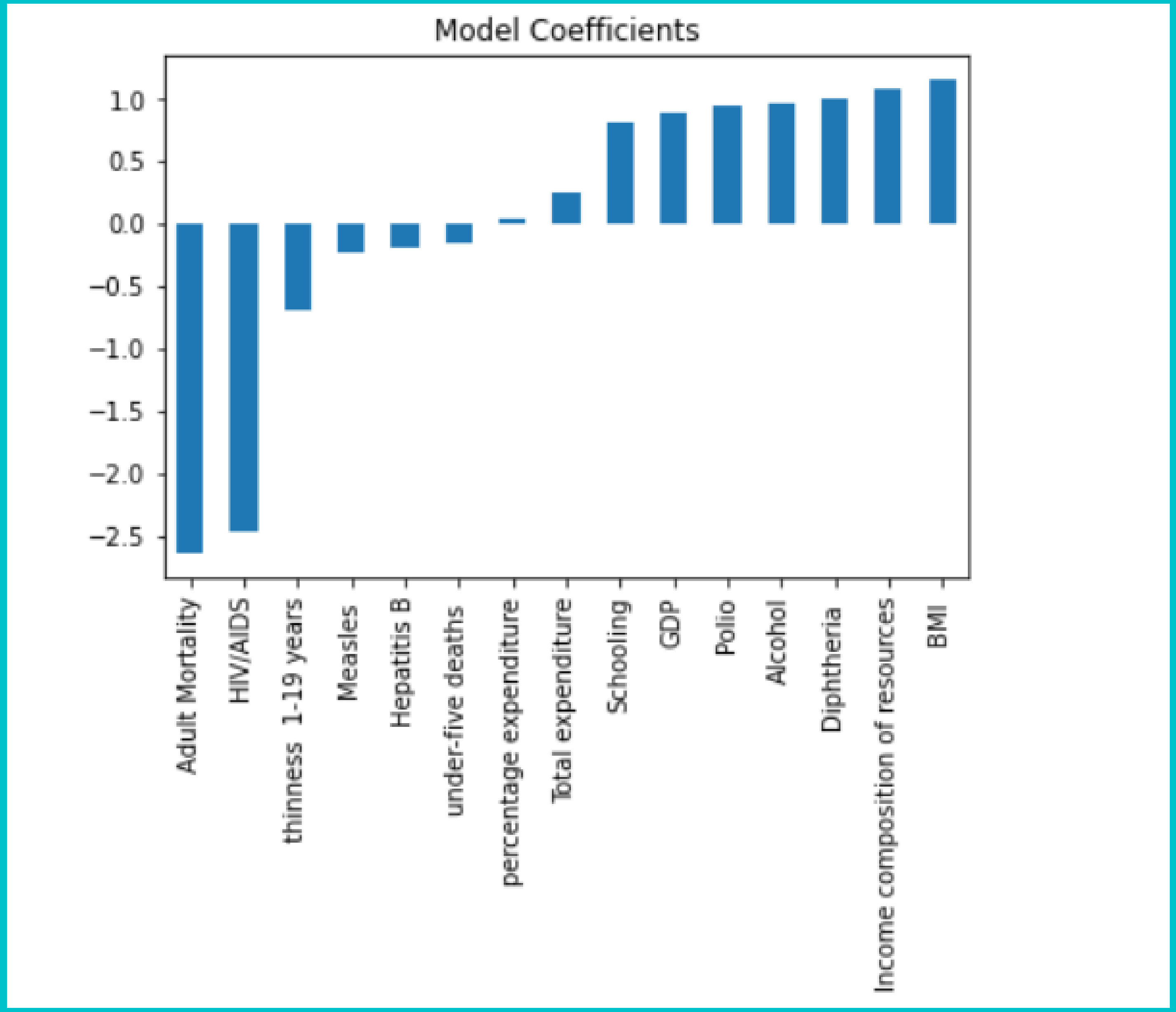
The relevant features considering a minimum threshold of coefficient value of +0.5 and -0.5, are Adult Mortality, Income composition of resources, HIV/AIDS, BMI, thinness 1-19 years, Diphtheria, Polio, Alcohol, GDP, Schooling and percentage expenditure.

The observed r2 score and RMSE are 0.773 and 4.583 respectively.

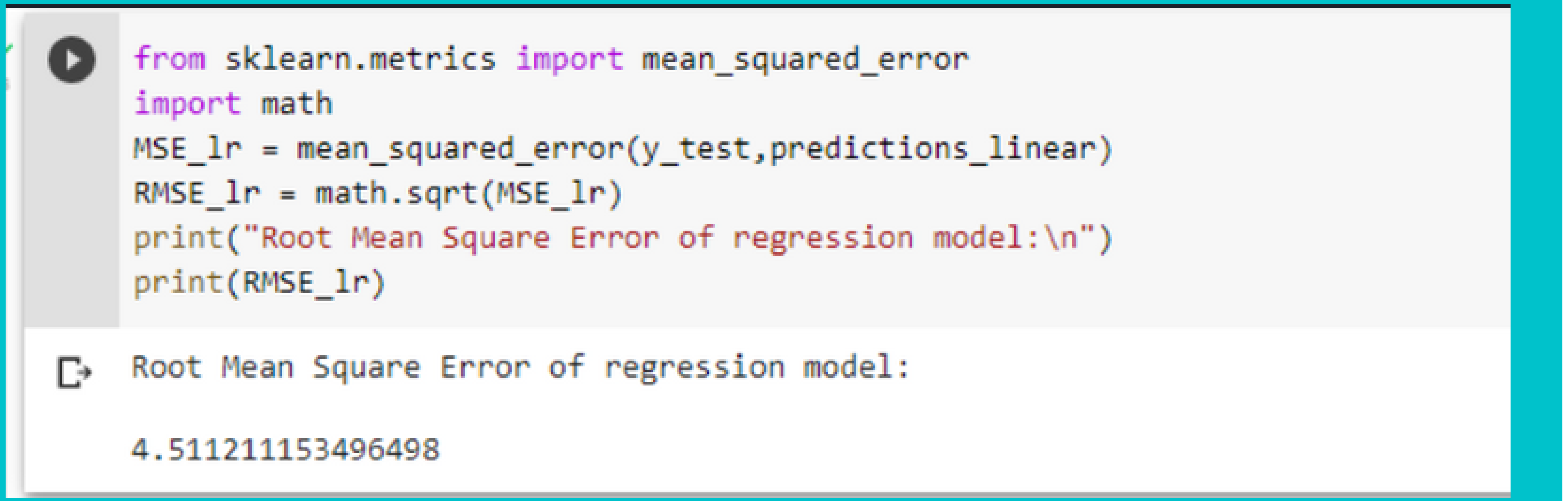


```
from sklearn.linear_model import Ridge

ridgereg = Ridge(alpha=0.1,normalize=True)
ridgereg.fit(X_train, y_train)
y_pred = ridgereg.predict(X_test)
print("r2 score: ", r2_score(y_test,y_pred))
print("RMSE: ", math.sqrt(mean_squared_error(y_test, y_pred)))
print(ridgereg.coef_)
```

```
r2 score:  0.7733110337114966
RMSE:  4.582712550907315
[-2.64677372  0.95555966  0.03005687 -0.19987089 -0.22390112  1.16220465
 -0.15219998  0.94767622  0.25548184  1.01027325 -2.45941677  0.87836378
 -0.70179556  1.07436319  0.81875673]
```



Multiple Linear Regression:



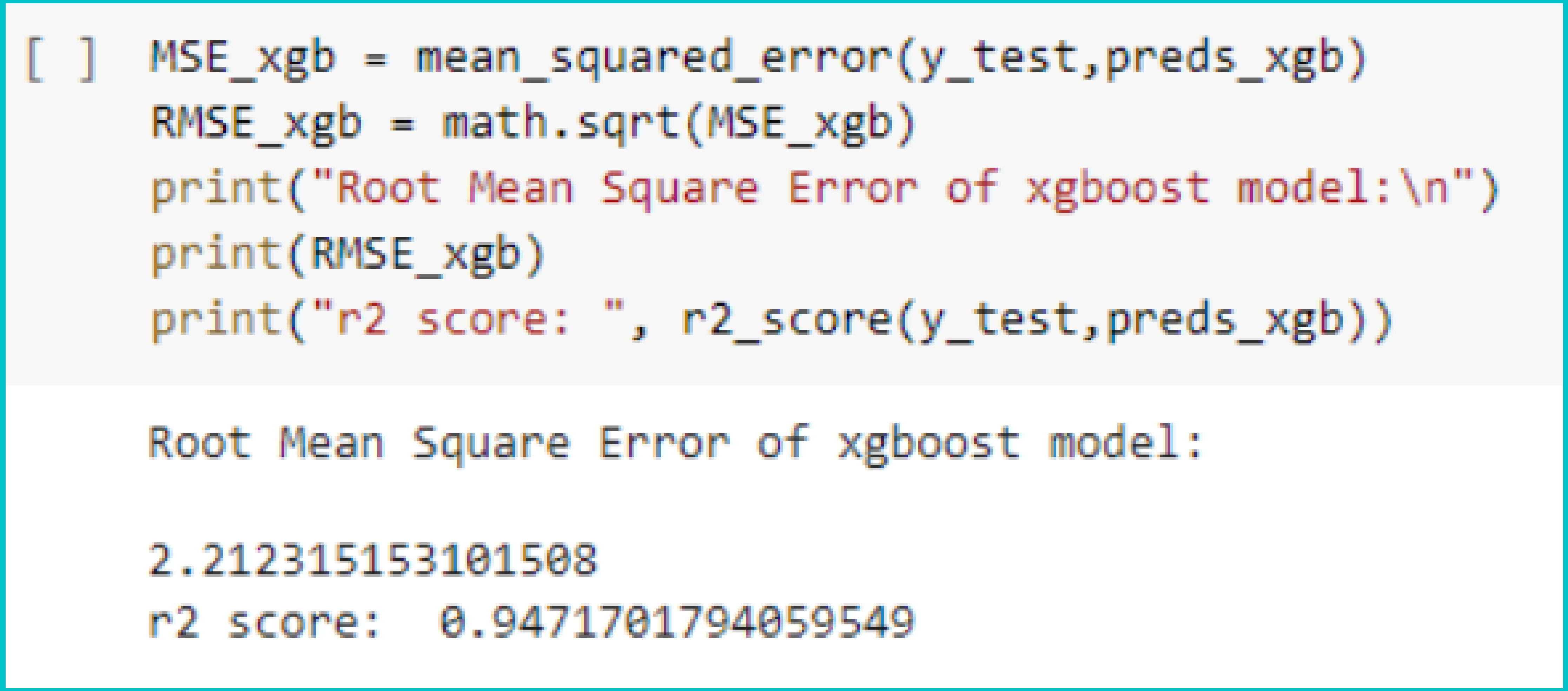
```
from sklearn.metrics import mean_squared_error
import math
MSE_lr = mean_squared_error(y_test,predictions_linear)
RMSE_lr = math.sqrt(MSE_lr)
print("Root Mean Square Error of regression model:\n")
print(RMSE_lr)
```

```
Root Mean Square Error of regression model:

4.511211153496498
```

- Manipulation of categorical variables:
- Binary encoders were used for the country names.
- One-hot encoding was used for the status of the countries as there are only 2 classes.
- The RMSE of the fitted model was found to be 4.511 and the r2 score 0.7803.

XGBoost:

On training the data on XGBoost, we received an RMSE of 2.213

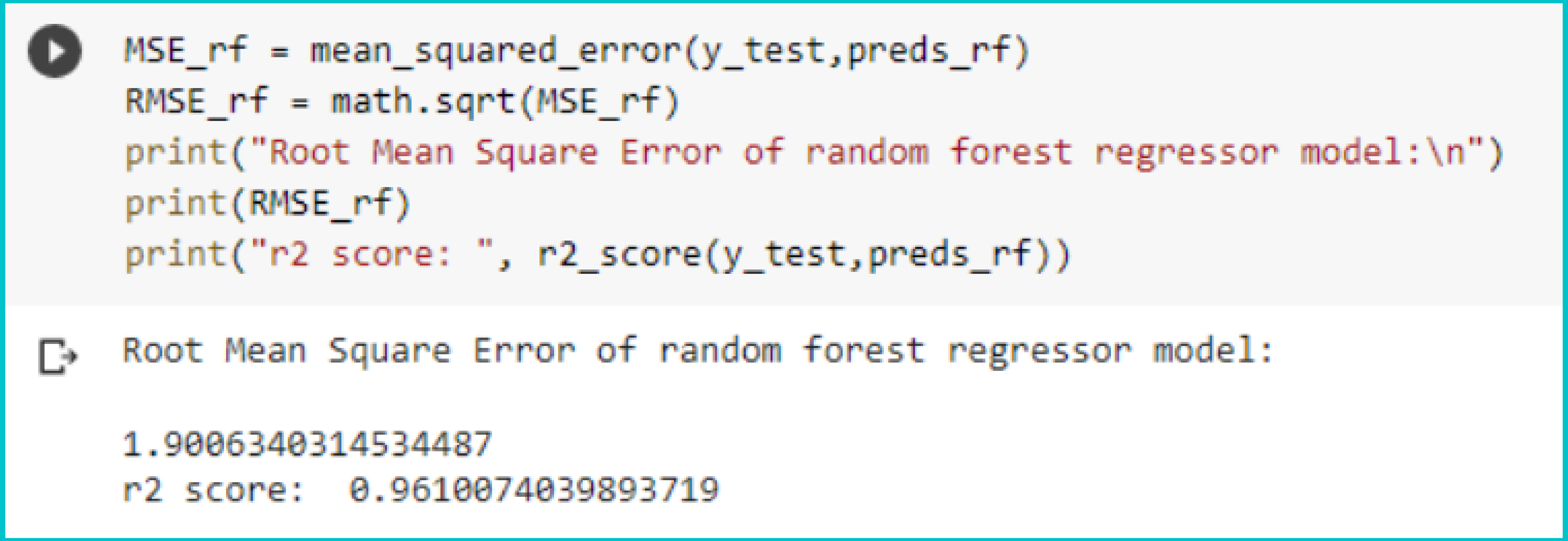


```
[ ]  MSE_xgb = mean_squared_error(y_test,preds_xgb)
     RMSE_xgb = math.sqrt(MSE_xgb)
     print("Root Mean Square Error of xgboost model:\n")
     print(RMSE_xgb)
     print("r2 score: ", r2_score(y_test,preds_xgb))
```

```
Root Mean Square Error of xgboost model:

2.212315153101508
r2 score:  0.9471701794059549
```

Random Forest:

This model is being used as it reduces error when compared to single regressor. The observed results were:



```
MSE_rf = mean_squared_error(y_test,preds_rf)
RMSE_rf = math.sqrt(MSE_rf)
print("Root Mean Square Error of random forest regressor model:\n")
print(RMSE_rf)
print("r2 score: ", r2_score(y_test,preds_rf))
```

```
Root Mean Square Error of random forest regressor model:

1.9006340314534487
r2 score:  0.9610074039893719
```

ANALYSIS ON STATUS DIFFERENTIATION

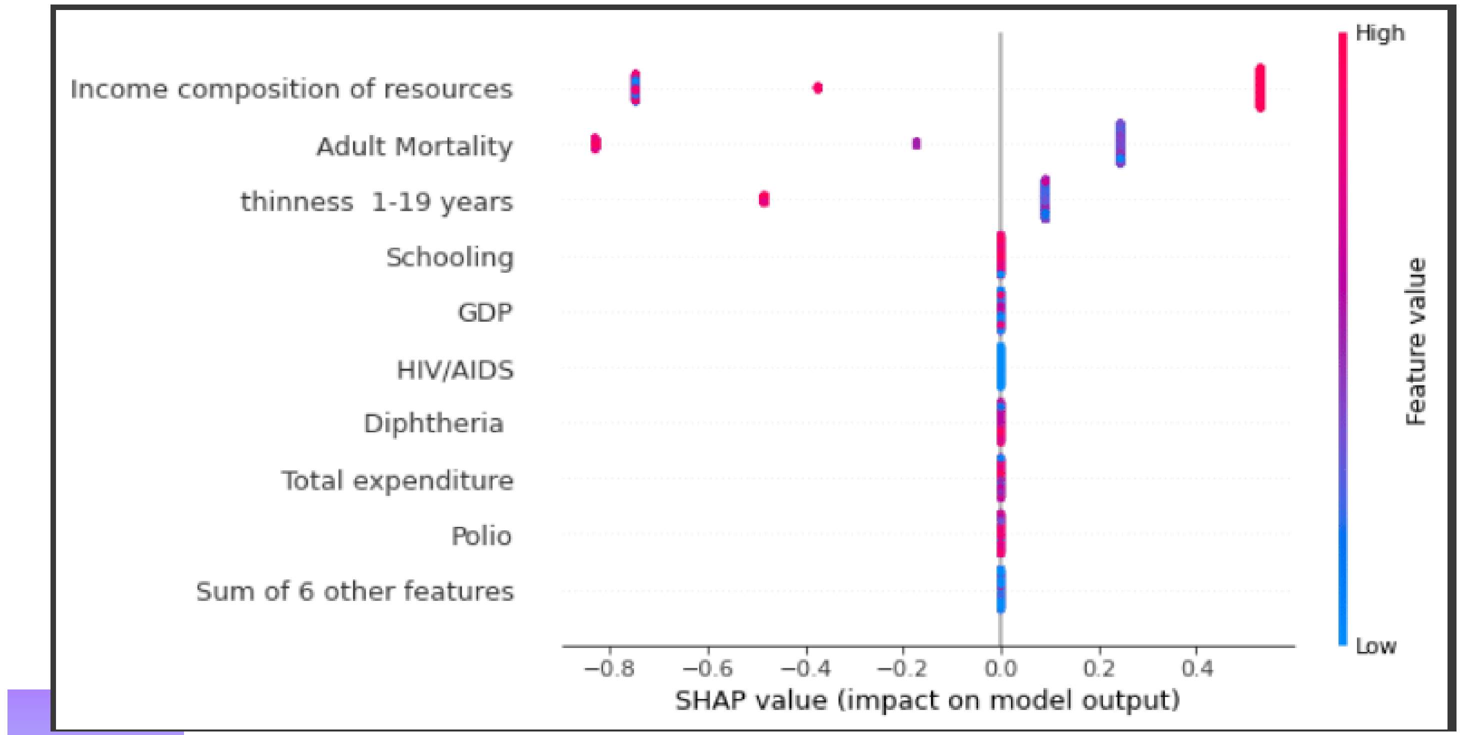
Factors affecting different categories of nations

The SHAP values for data from developed and devloping nations were examined to get the following features:

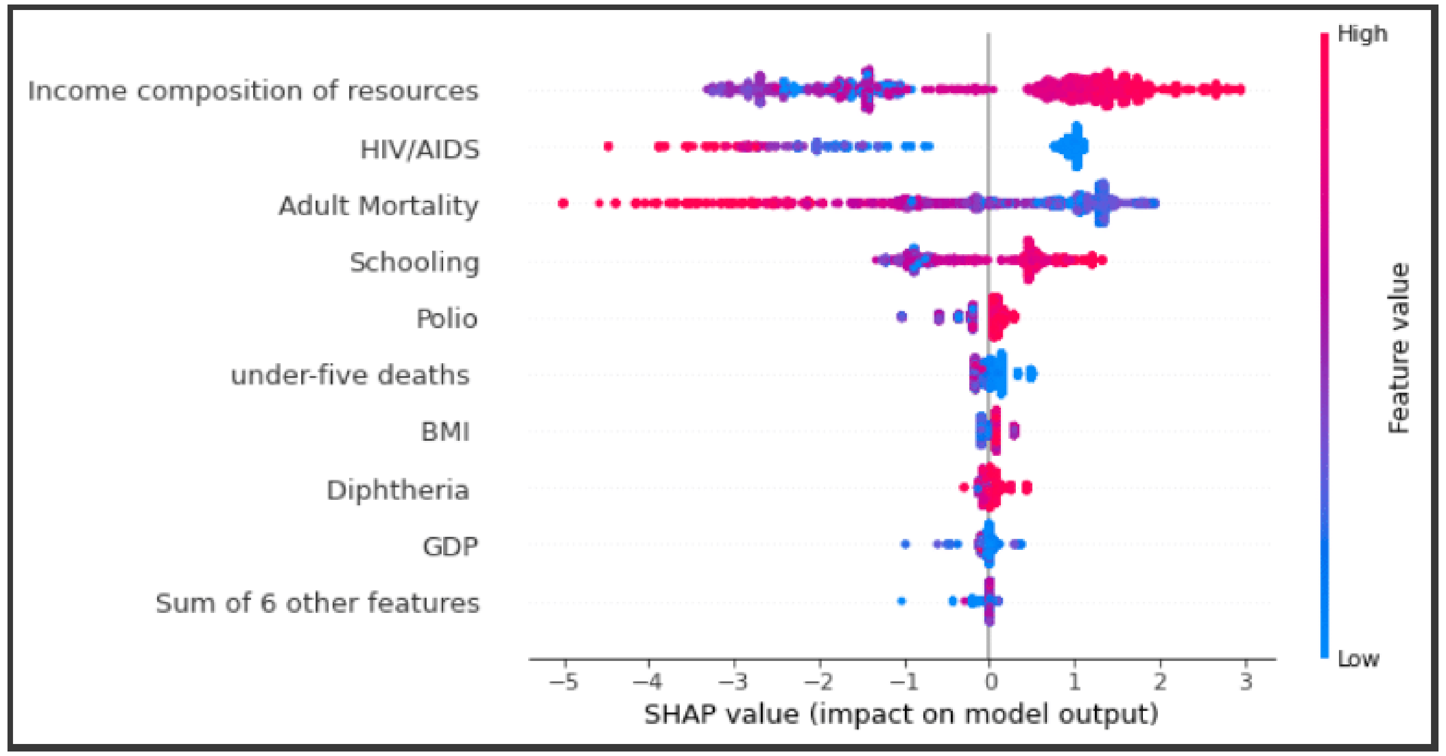
Developed - Income composition of resources, Adult Mortality, thinness 1-19 years

Developing - Income composition of resources, HIV/AIDS, Schooling, Adult Mortality, BMI, Polio, GDP

Developed Nations



Developing Nations

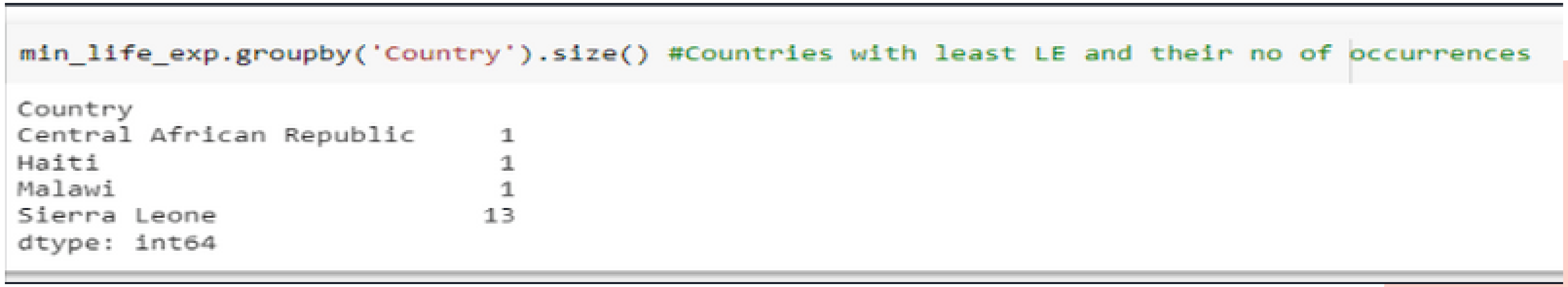


Factors affecting different categories of nations

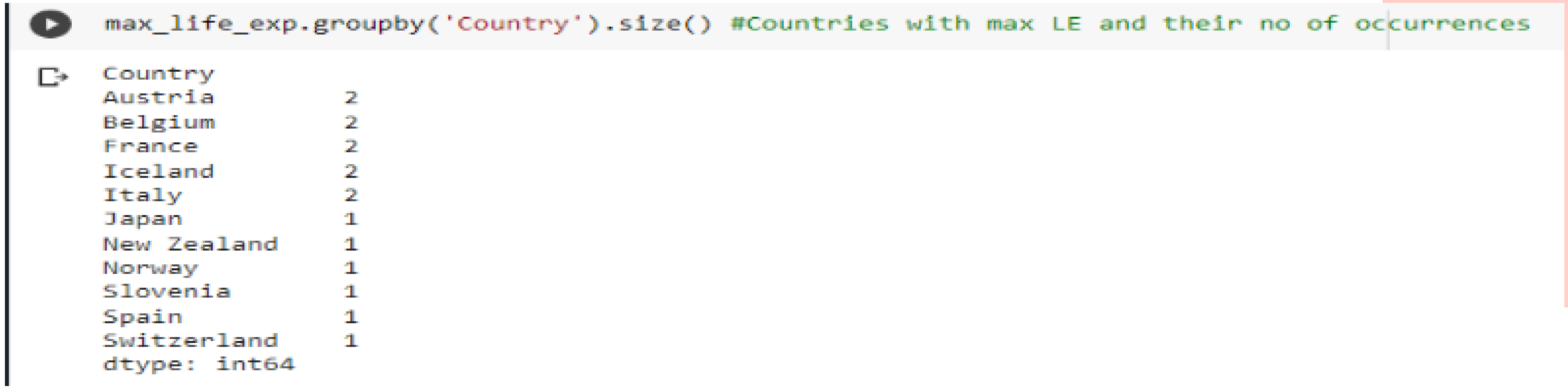
Random Forest was trained on the 2 groups of features pertaining to developed and developing nations.

Data	RMSE	r2-score
Developed	1.852	0.762
Developing	1.939	0.955

LE GROWTH RATE OF COUNTRIES



```
min_life_exp.groupby('Country').size() #Countries with least LE and their no of occurrences

Country
Central African Republic     1
Haiti                        1
Malawi                       1
Sierra Leone                13
dtype: int64
```



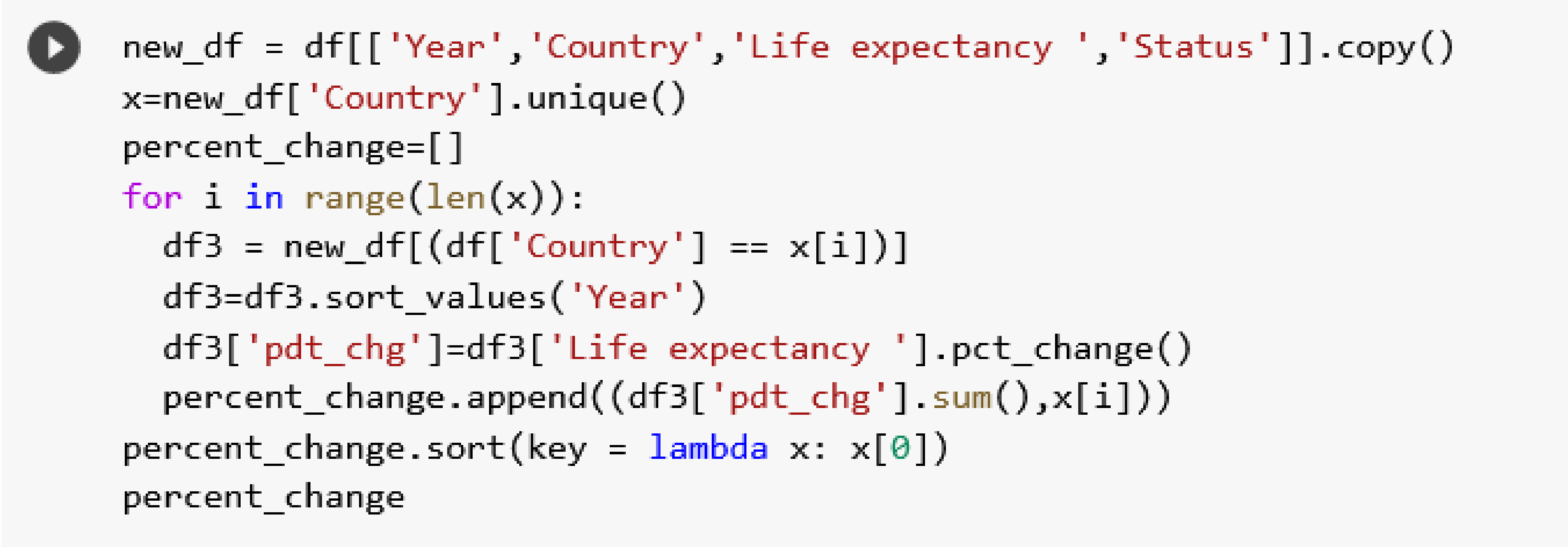
```
max_life_exp.groupby('Country').size() #Countries with max LE and their no of occurrences

Country
Austria        2
Belgium        2
France         2
Iceland        2
Italy          2
Japan          1
New Zealand    1
Norway         1
Slovenia       1
Spain          1
Switzerland    1
dtype: int64
```

It was observed, from the years in the range 2000-2015, countries like Austria, Belgium, France which had high quality of life had higher life expectancy whereas countries like Sierra Leone and Haiti, where AIDS, malnutrition, civil strife, and other factors like the Ebola virus outbreak had taken a tremendous toll on human life, had the least life expectancy.

Percentage change in LE for all countries

The following piece of code if for finding the percentage change (rather growth) in LE of all the countries in our dataset.

It was essential to check on with the countries status and growth with respect to LE inorder to draw conclusion if the countries require help and support in any of the categories- be it healthcare or economic stability.

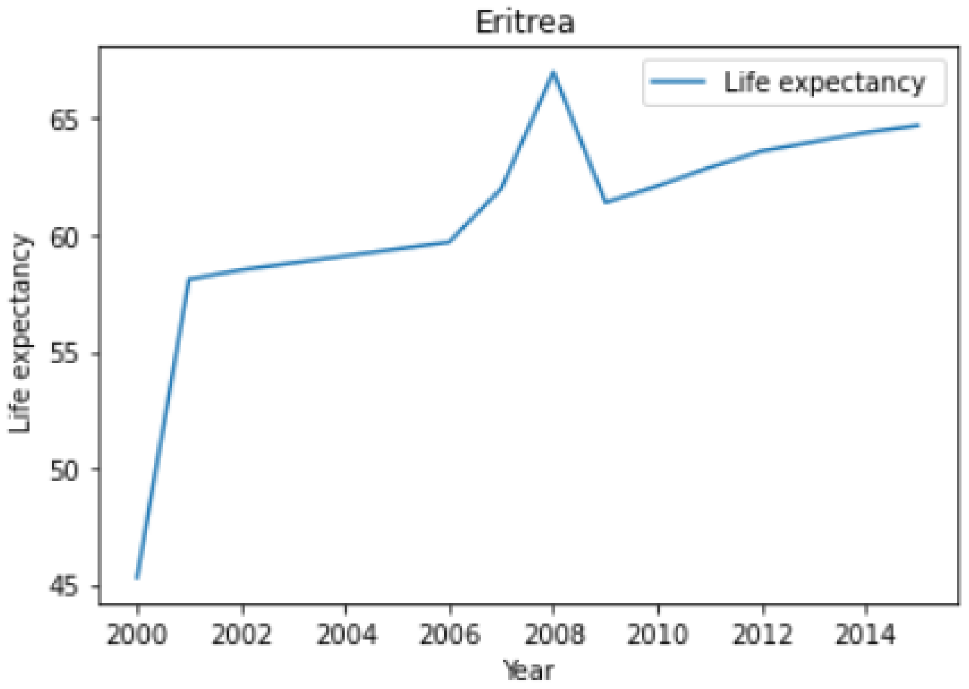


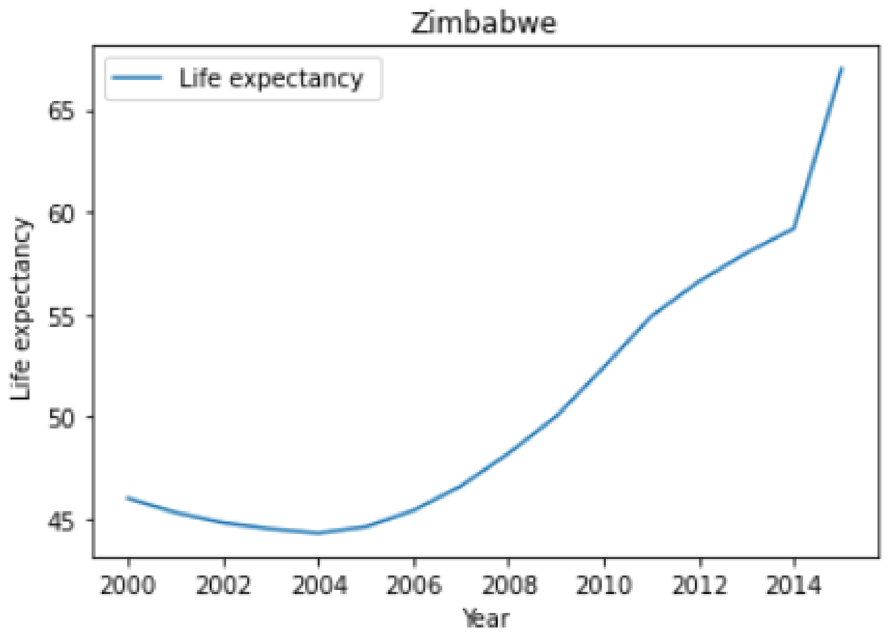
```
new_df = df[['Year','Country','Life expectancy ','Status']].copy()
x=new_df['Country'].unique()
percent_change=[]
for i in range(len(x)):
  df3 = new_df[(df['Country'] == x[i])]
  df3=df3.sort_values('Year')
  df3['pdt_chg']=df3['Life expectancy '].pct_change()
  percent_change.append((df3['pdt_chg'].sum(),x[i]))
percent_change.sort(key = lambda x: x[0])
percent_change
```

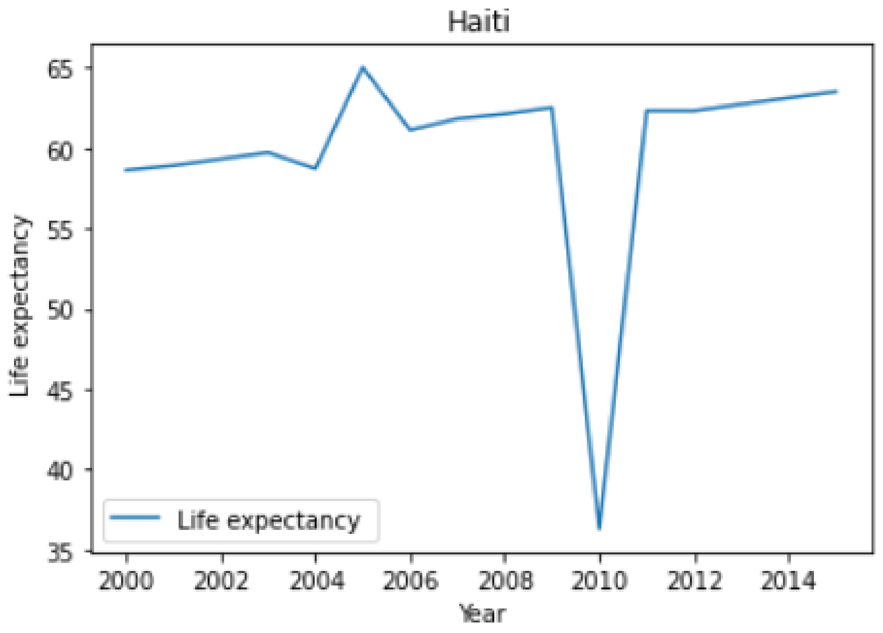
Countries with Maximum % Change or growth

The figures below shows the LE plot in the years 2000-2015.Countries Eritrea, Zimbabwe and Haiti were recognized as top 3 countries with maximum growth in LE.

All these countries are "developing" and it is great to observe that these countries are striving harder to uplift the livability of population with an average Life expectancy across these countries being 55.

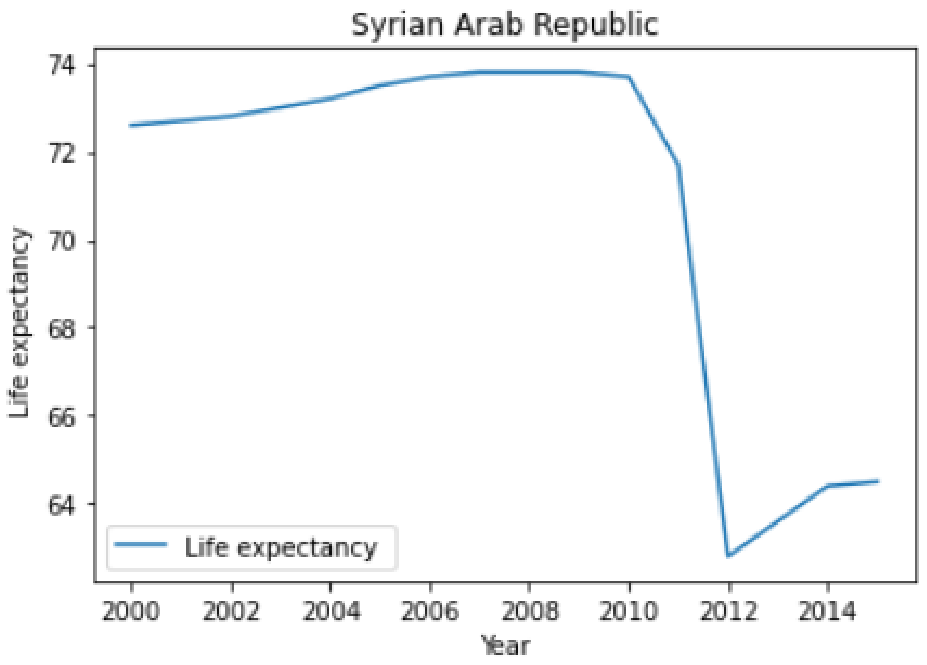


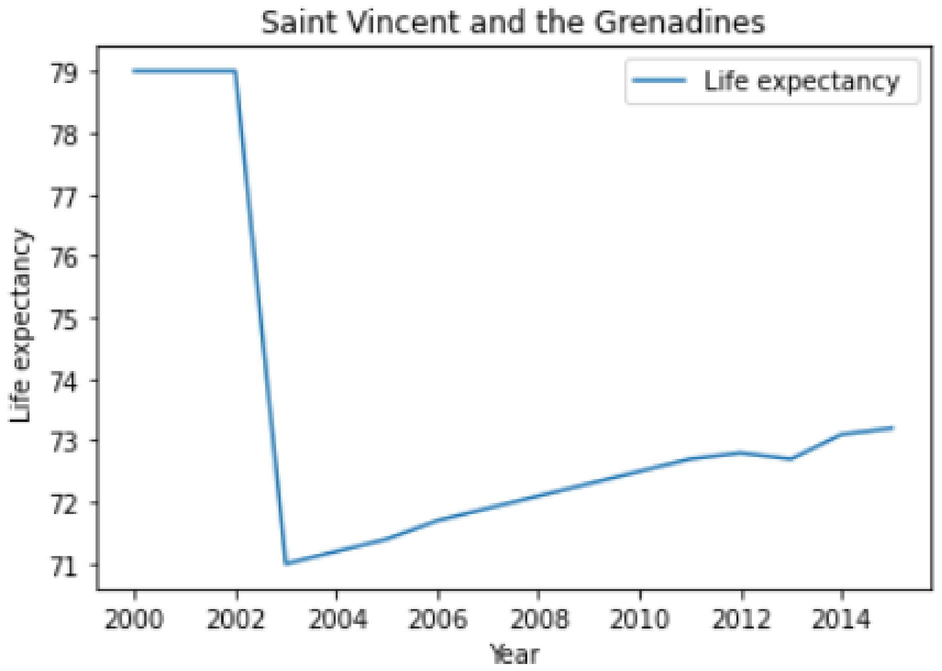


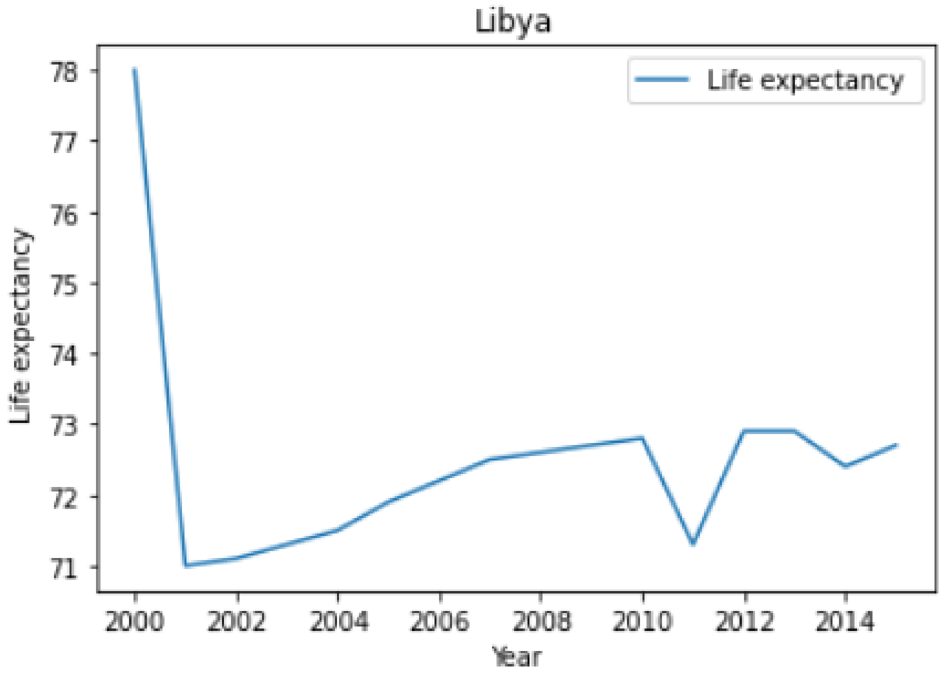


Countries with Minimum % Change or growth

The countries Syrian Arab Republic,Saint-Vincent and the Grenadines and Libya had a sudden dip in their their LE due to socio-economic-political factors and since these were developing countries, recovering from the fall was considered as a huge challenge.



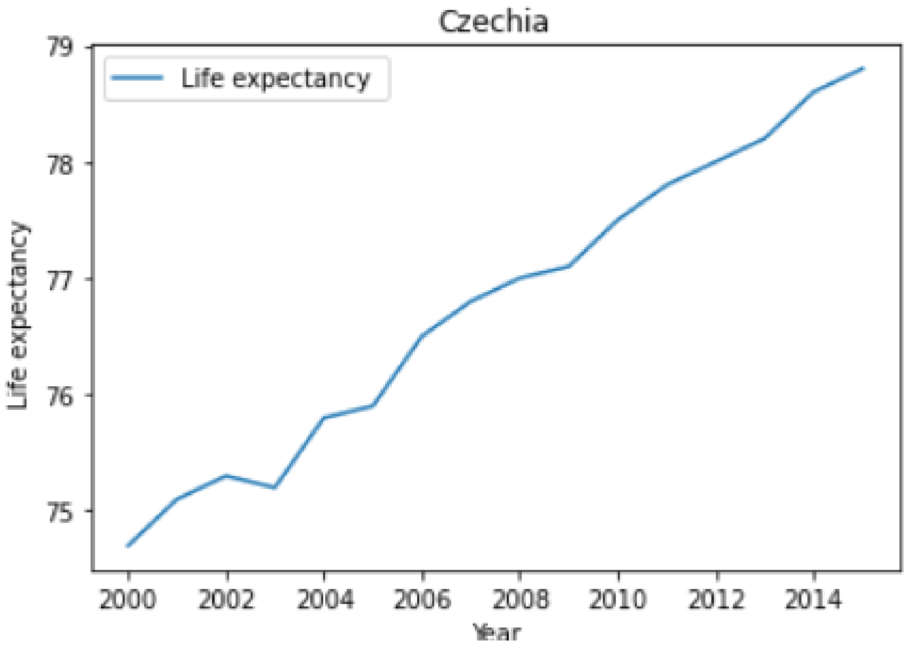


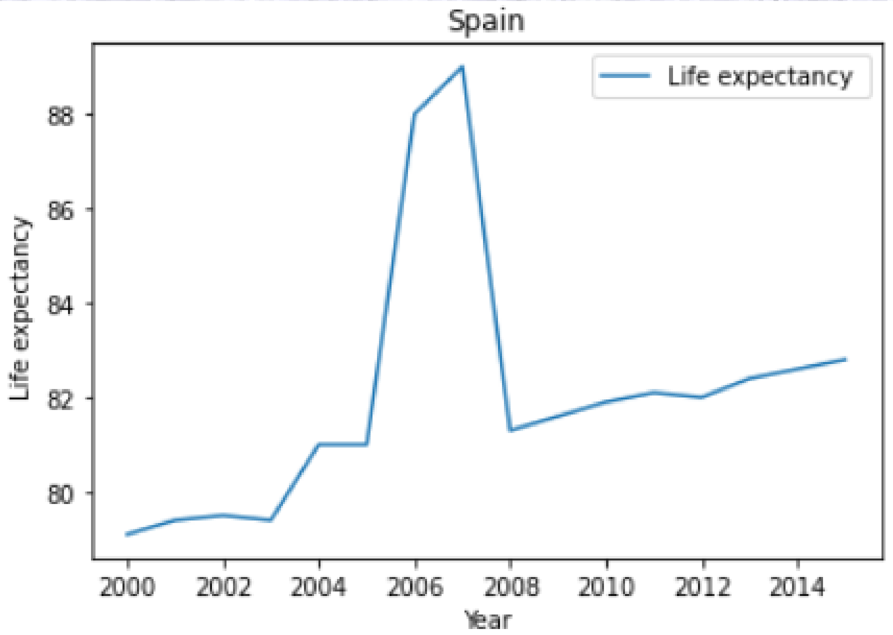


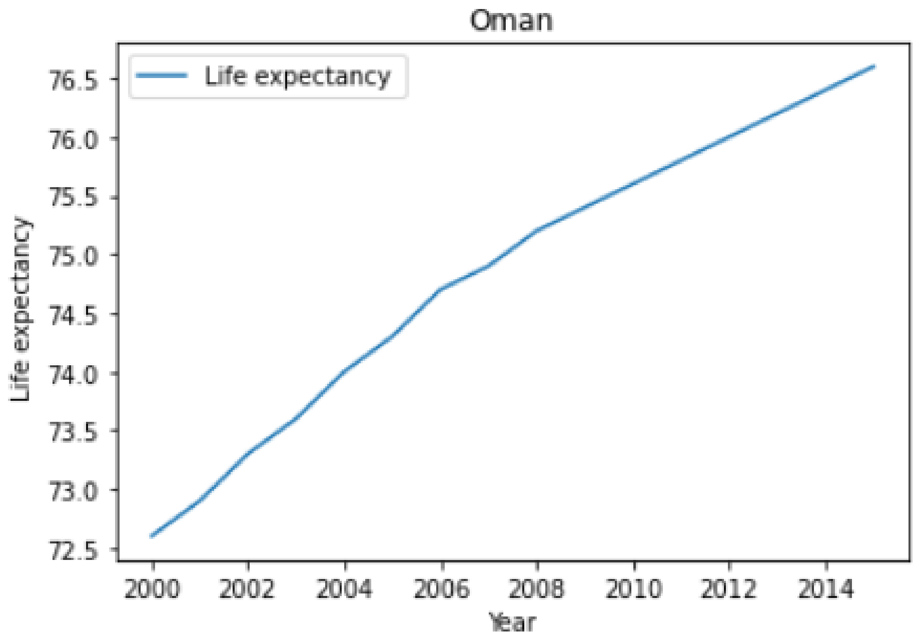
Countries with Average % Change or growth

The countries Czechia, Spain and Oman were recognised to have an average increase-decrease pattern in their percentage change in their LE probably due to the fact that their LE had reached higher values and thus saturated.

The main cause of Spain's irregular pattern can be accounted to the housing bubble and the accompanying unsustainable high GDP growth rate.







CREDITS

CONTRIBUTIONS

Spoorthi KS	Analysis of factors influencing LE in developed and developing countries using SHAP values, modeled random forest for the same,eda,preprocessing
Thrupthi N	Modeled MLR,XGBoost,RandomForest regressor,feature selection using lasso regression,eda,preprocessing
Sahana Rao	Country-wise analysis,comparing the growth of life expectancy amongst countries across years,eda,preprocessing

TEAM DataStorm

SPOORTHI K S
PES1UG20CS433

THRUPTHI N
PES1UG20CS470

SAHANA RAO
PES1UG20CS588