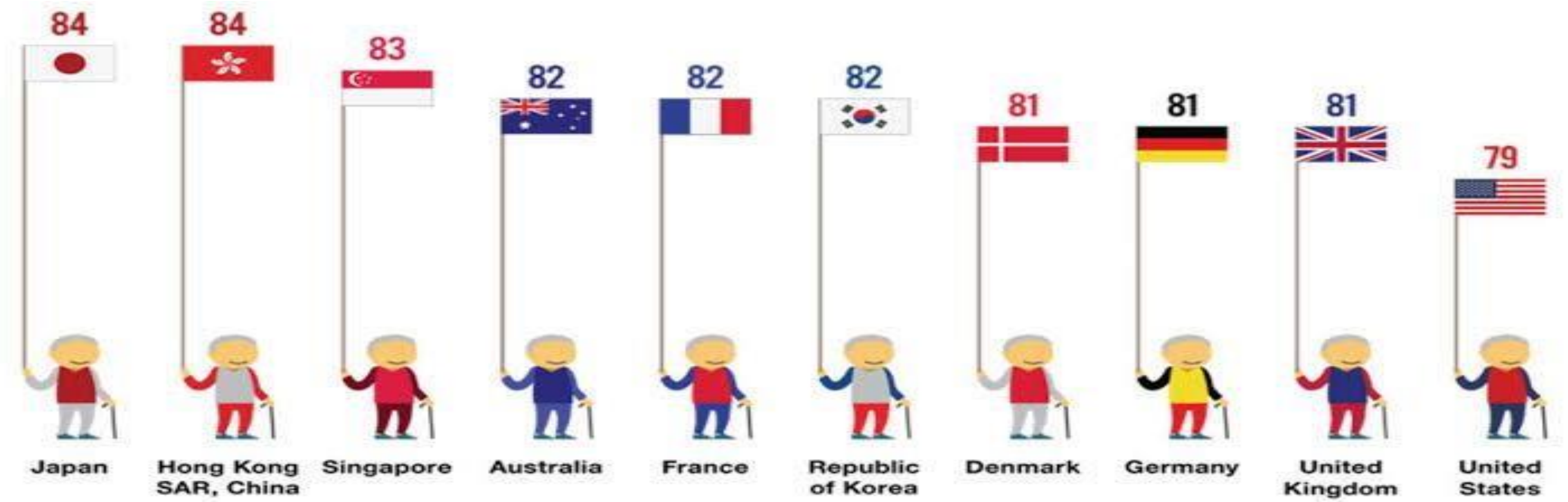
Life Expectancy
Analysis



By Akshay Pokale

About Dataset

- Life Expectancy means The number of years a person can expect to live.
- The Global Health Observatory (GHO) data repository under World Health Organization (WHO) keeps track of the health status as well as many other related factors for all countries.
- The dataset related to life expectancy, health factors for 193 countries has been collected from the same WHO data repository website and its corresponding economic data was collected from United Nation website.
- Among all categories of health-related factors only those critical factors were chosen which are more representative.
- It has been observed that in the past 15 years, there has been a huge development in health sector resulting in improvement of human mortality rates especially in the developing nations in comparison to the past 30 years. Therefore, in this project we have considered data from year 2000-2015 for 193 countries for further analysis.

Index

- 1) Introduction to dataset
- 2) Work Flow
- 3) EDA
- 4) Implementation of ML algorithms
- 5) Performance Parameters
- 6) Conclusion

Introduction to dataset

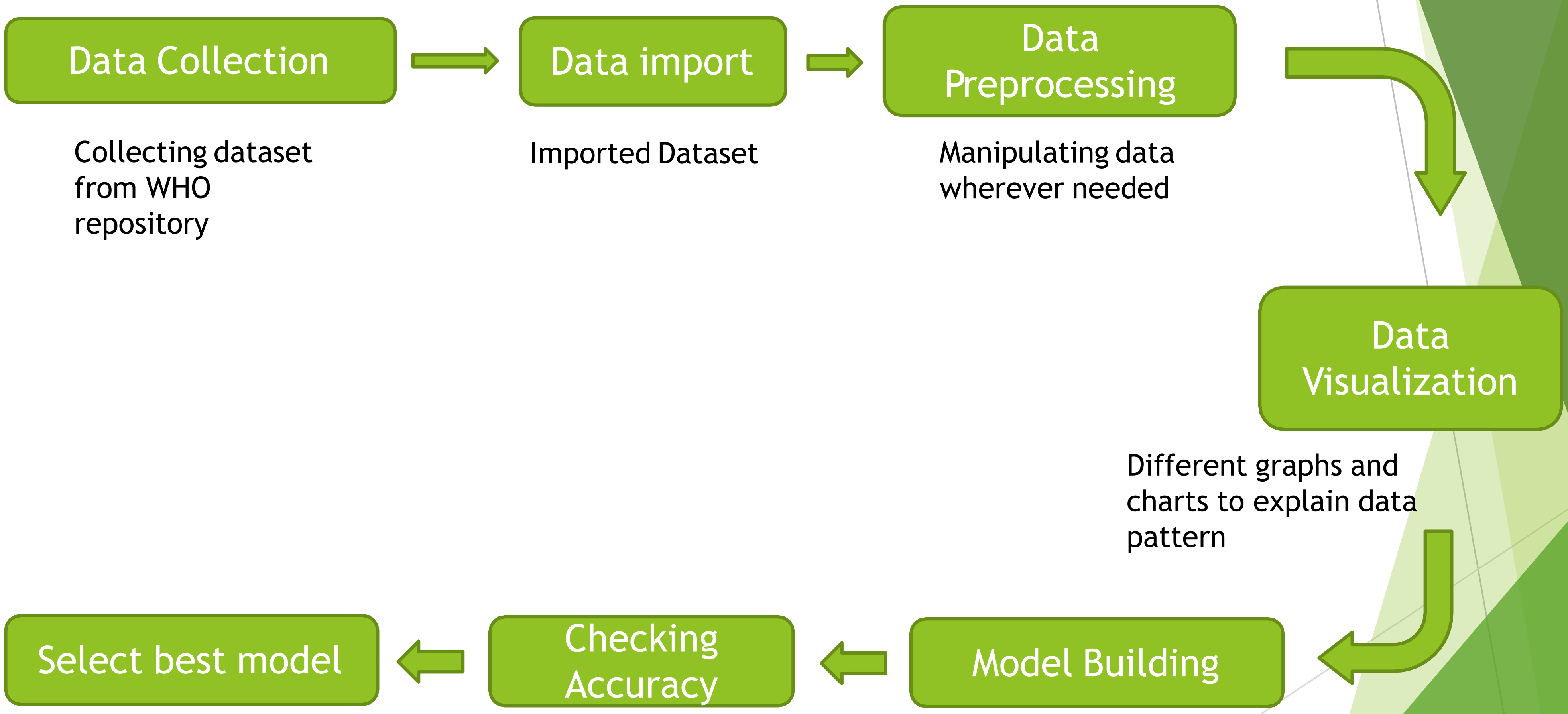
Data points 58760

Shape 2938, 20

Columns as Follows :

- Schooling
- Status
- Income composition of resources
- 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
- Life expectancy is target column

WORK FLOW



Data Collection

Collecting dataset from WHO repository

Data import

Imported Dataset

Data Preprocessing

Manipulating data wherever needed

Data Visualization

Different graphs and charts to explain data pattern

Model Building

Build 3 logistic regression model

Checking Accuracy

Comparing accuracies of all the models

Select best model

Deciding the best model based on accuracies

EDA

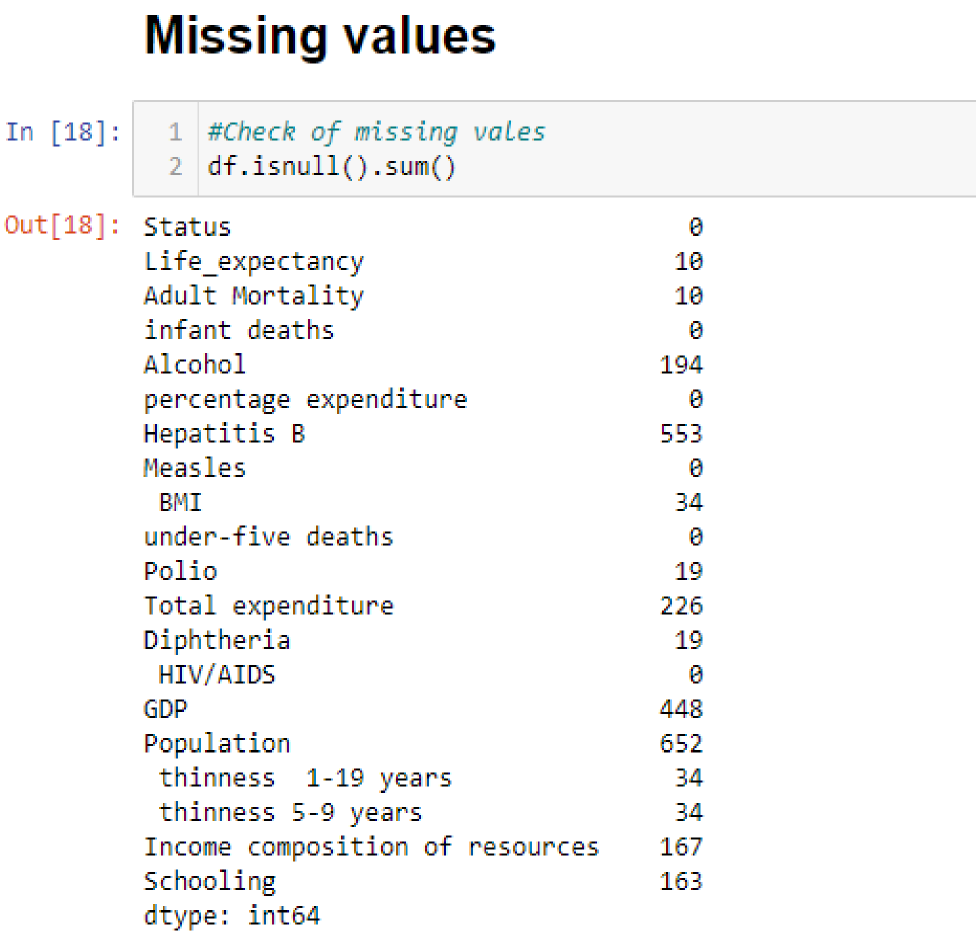
Challenges in dataset

- Missing values
- Outliers
- Categorical data

  I) Label Binarize

  II) Standard Scalar
- Feature selection

Missing Value Treatment



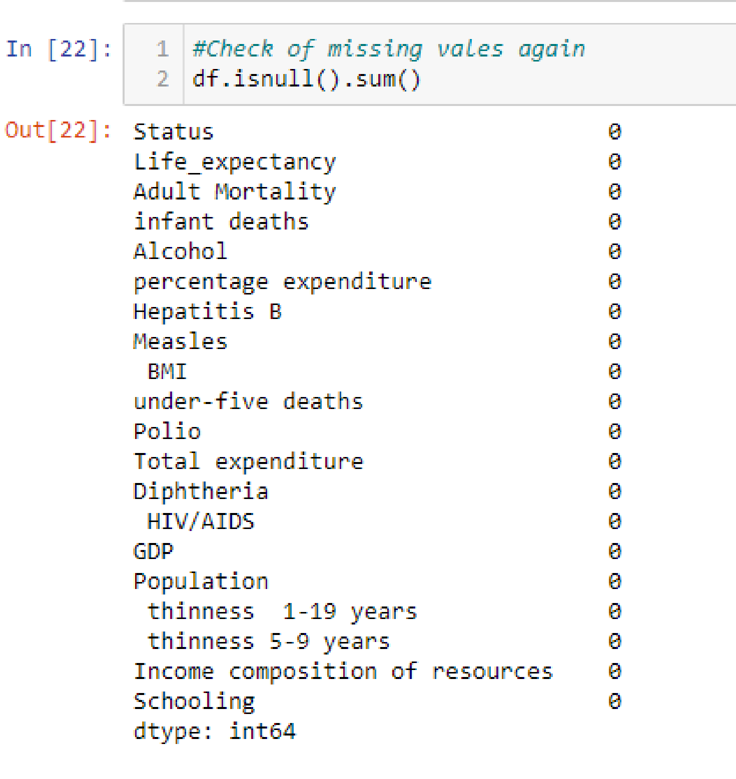
Missing values

In [18]:

```
#Check of missing vales
df.isnull().sum()
```

Out[18]:

```
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
```



In [22]:

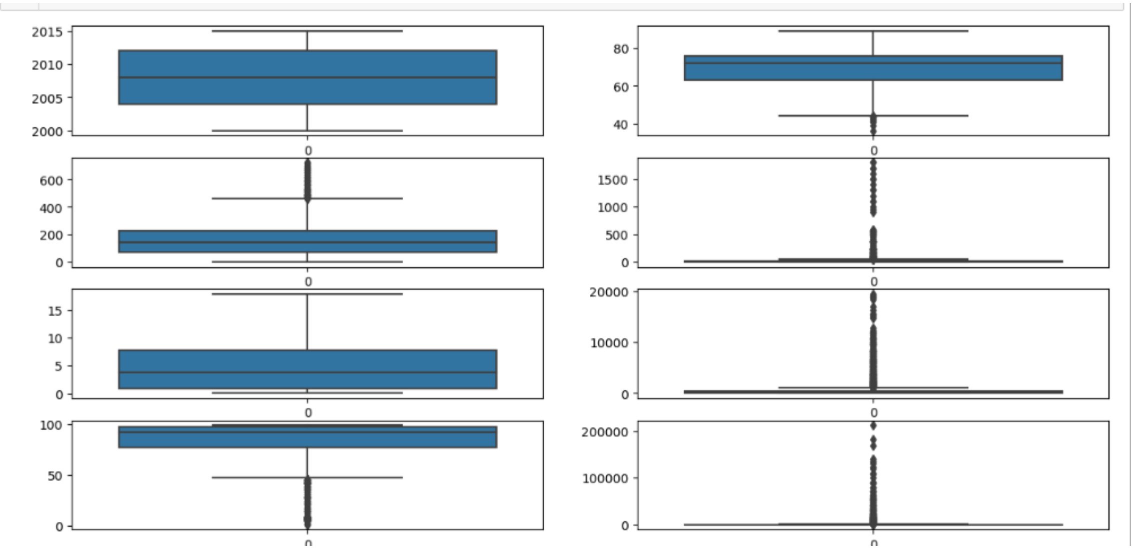
```
#Check of missing vales again
df.isnull().sum()
```

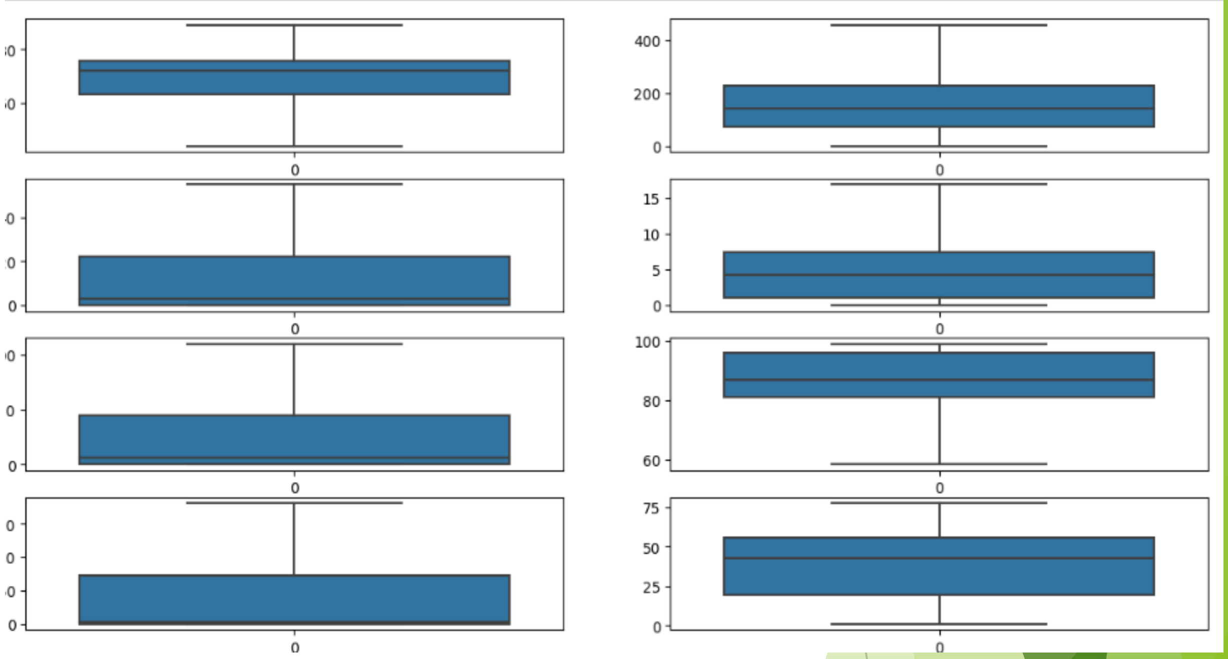
Out[22]:

```
Status                             0
Life_expectancy                    0
Adult Mortality                    0
infant deaths                      0
Alcohol                            0
percentage expenditure             0
Hepatitis B                        0
Measles                            0
 BMI                               0
under-five deaths                  0
Polio                              0
Total expenditure                  0
Diphtheria                         0
 HIV/AIDS                          0
GDP                                0
Population                         0
 thinness  1-19 years              0
 thinness 5-9 years                0
Income composition of resources    0
Schooling                          0
dtype: int64
```

- Life Expectancy is our Target Column so we drop records wherever missing value
- Number datatype variable convert into mean of same column

Outliers Treatment

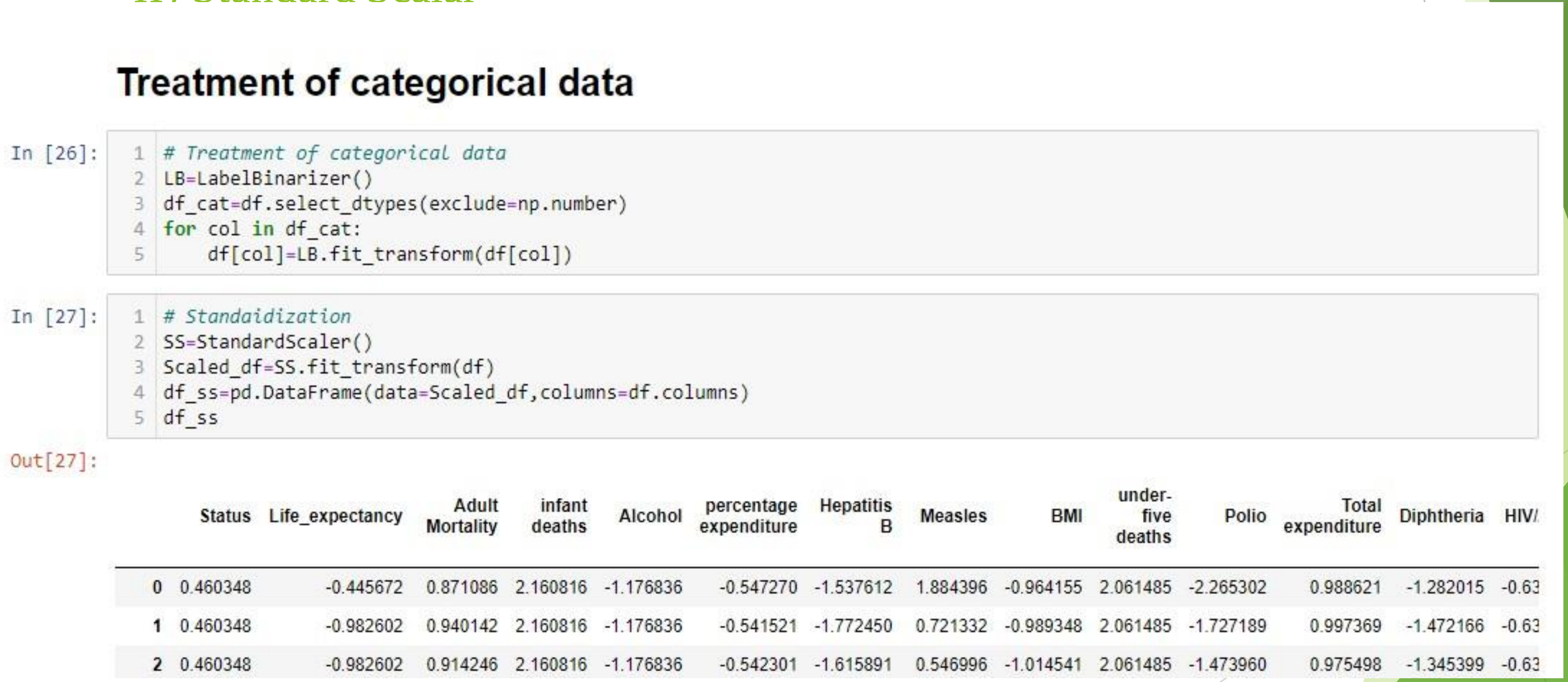




Using Winsorizing Technique we remove Outlier

Categorical data Treatment

I) Label Binarize
II) Standard Scalar



Treatment of categorical data

In [26]:

```
# Treatment of categorical data
LB=LabelBinarizer()
df_cat=df.select_dtypes(exclude=np.number)
for col in df_cat:
    df[col]=LB.fit_transform(df[col])
```

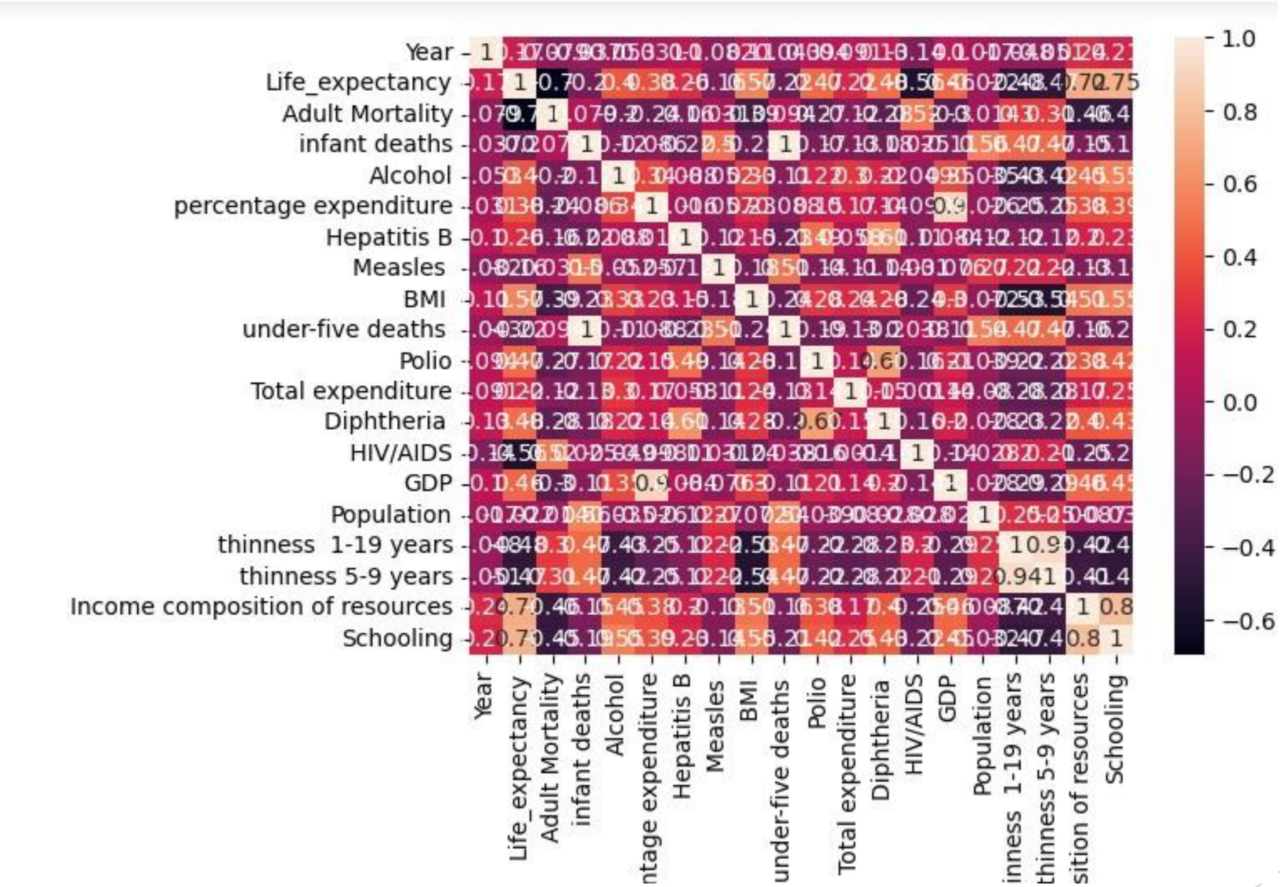
In [27]:

```
# Standaidization
SS=StandardScaler()
Scaled_df=SS.fit_transform(df)
df_ss=pd.DataFrame(data=Scaled_df,columns=df.columns)
df_ss
```

Out[27]:

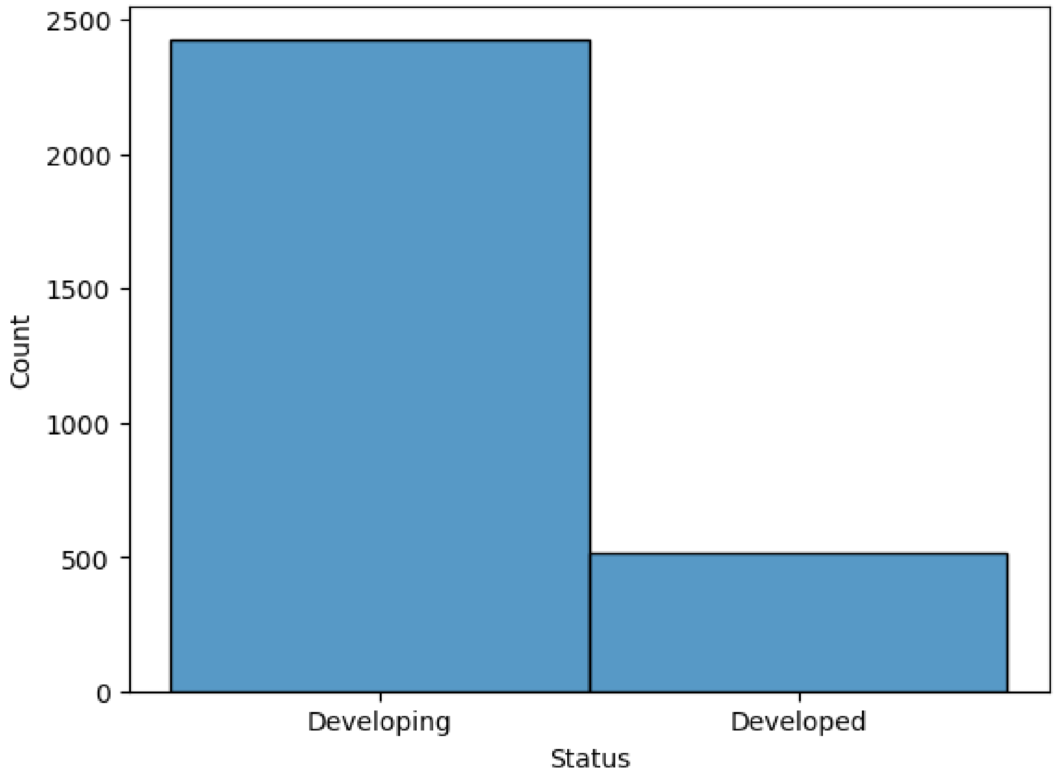
	Status	Life_expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	BMI	under-five deaths	Polio	Total expenditure	Diphtheria	HIV/
0	0.460348	-0.445672	0.871086	2.160816	-1.176836	-0.547270	-1.537612	1.884396	-0.964155	2.061485	-2.265302	0.988621	-1.282015	-0.63
1	0.460348	-0.982602	0.940142	2.160816	-1.176836	-0.541521	-1.772450	0.721332	-0.989348	2.061485	-1.727189	0.997369	-1.472166	-0.63
2	0.460348	-0.982602	0.914246	2.160816	-1.176836	-0.542301	-1.615891	0.546996	-1.014541	2.061485	-1.473960	0.975498	-1.345399	-0.63

Correlations



The target col has highest correlation with schooling and income composition.

EDA Country Development

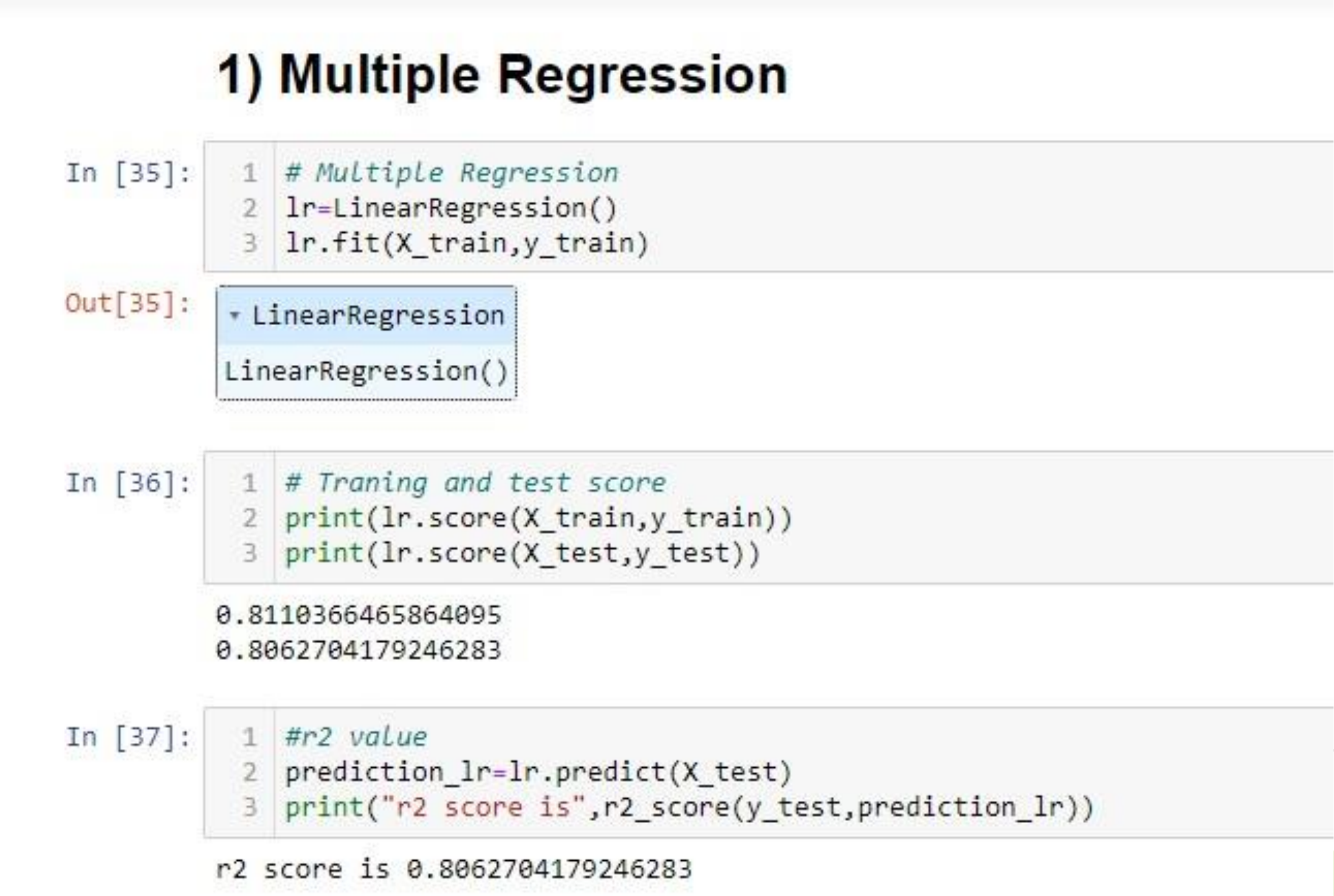


To check which status of country is more

Implementation of ML algorithm

- Multiple Linear Regression
- Random Forest Regression

Multiple Linear Regression



1) Multiple Regression

```
In [35]:
# Multiple Regression
lr=LinearRegression()
lr.fit(X_train,y_train)
```

```
Out[35]:
▾ LinearRegression
LinearRegression()
```

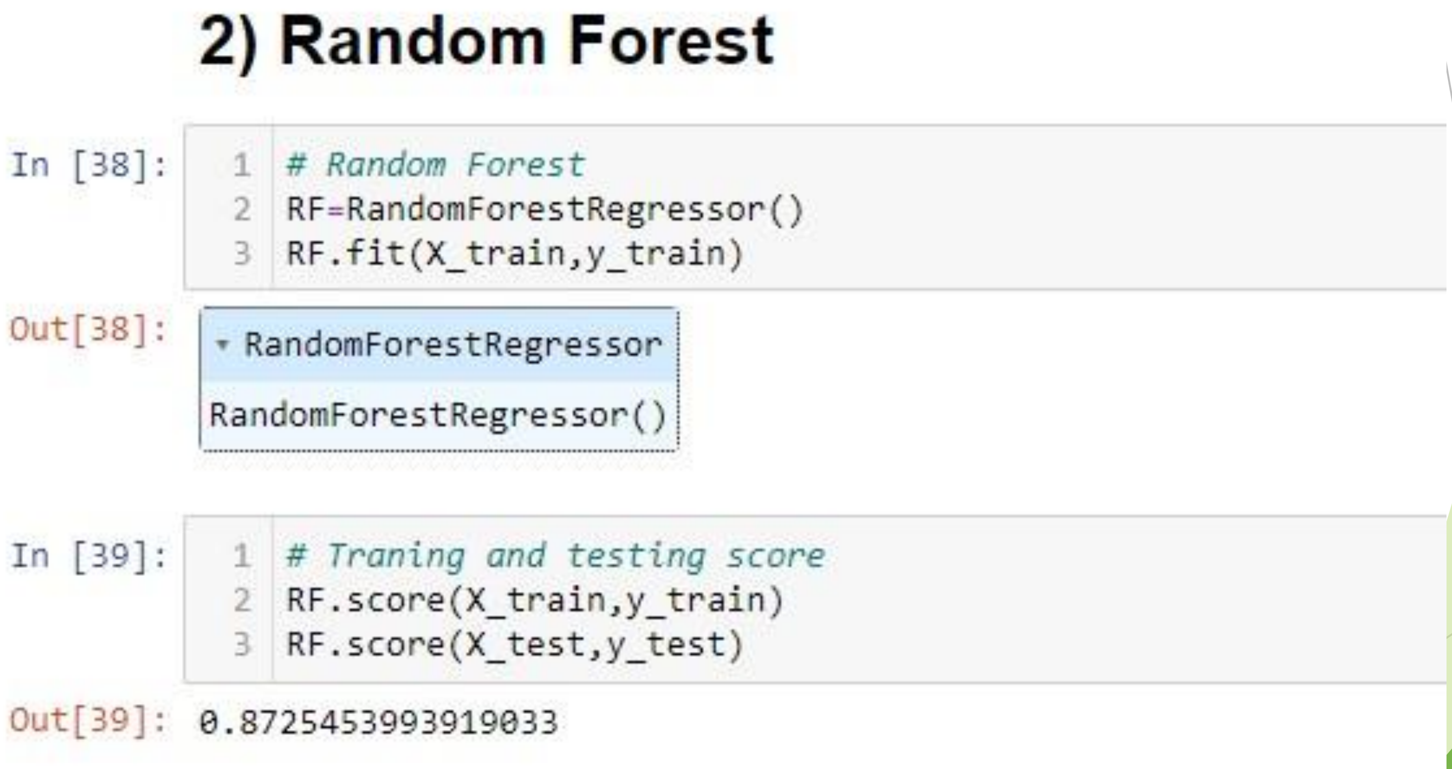
```
In [36]:
# Traning and test score
print(lr.score(X_train,y_train))
print(lr.score(X_test,y_test))
```

```
0.8110366465864095
0.8062704179246283
```

```
In [37]:
#r2 value
prediction_lr=lr.predict(X_test)
print("r2 score is",r2_score(y_test,prediction_lr))
```

```
r2 score is 0.8062704179246283
```

Random Forest Regression



2) Random Forest

```
In [38]:  # Random Forest
          RF=RandomForestRegressor()
          RF.fit(X_train,y_train)

Out[38]:  ▾ RandomForestRegressor
          RandomForestRegressor()
```

```
In [39]:  # Traning and testing score
          RF.score(X_train,y_train)
          RF.score(X_test,y_test)

Out[39]:  0.8725453993919033
```

Conclusion

- After doing some analysis on this dataset, we can conclude that Schooling is the most important variable in life expectancy.
- Schooling Improved Knowledge and Health Literacy also improve Income and because of income lifestyle Improved Healthcare Access, Overall Quality of Life.
- The model was developed using a variety of machine learning algorithms, including Multiple Linear Regression, Random Forest.
- The best performing algorithm was the random forest algorithm, which achieved and accuracy of 87.25 %.
- The Life Expectancy dataset provides valuable insights into the life dependent Factors and how we can grow. Gains in life expectancy at birth can be attributed to a number of factors, including rising living standards, improved lifestyle and better education, as well as greater access to quality health services.