
LIFE EXPECTANCY

DOES IMMUNIZATION MATTER?

Mireille P. Feudjio T.
Springboard Fellow

Outline

- **Context and Problem statement**
- **Data Wrangling**
- **Exploratory data analysis**
- **Modelling**
- **Conclusion**

Context and Problem statement

Contexte

- In view of the current pandemic, vaccination does not seem to find the consent of some people in the world. A look at the impact of vaccination on life expectancy could be important to highlight.
- **So, does immunization matter?**
- The dataset (life expectancy, health factors for 193 countries) 2000-2015.
- The predicting variables were then divided into several broad categories:
 - Immunization related factors,
 - Mortality factors,
 - Economical factors,
 - and Social factors.

The present project assessed the contribution and the relationship of each feature on life expectancy with a special focus on immunization factors, and develop a model to predict life expectancy.

Scope of solution space: The model development should take into account all the features with special attention on immunisation factors

Constraints : The dataset has important missing values to handle (table 1). This could impact the model depending on the imputation technique chosen.

Stakeholders to provide key insight: SpringBoard Mentors

Key data sources [data](#)

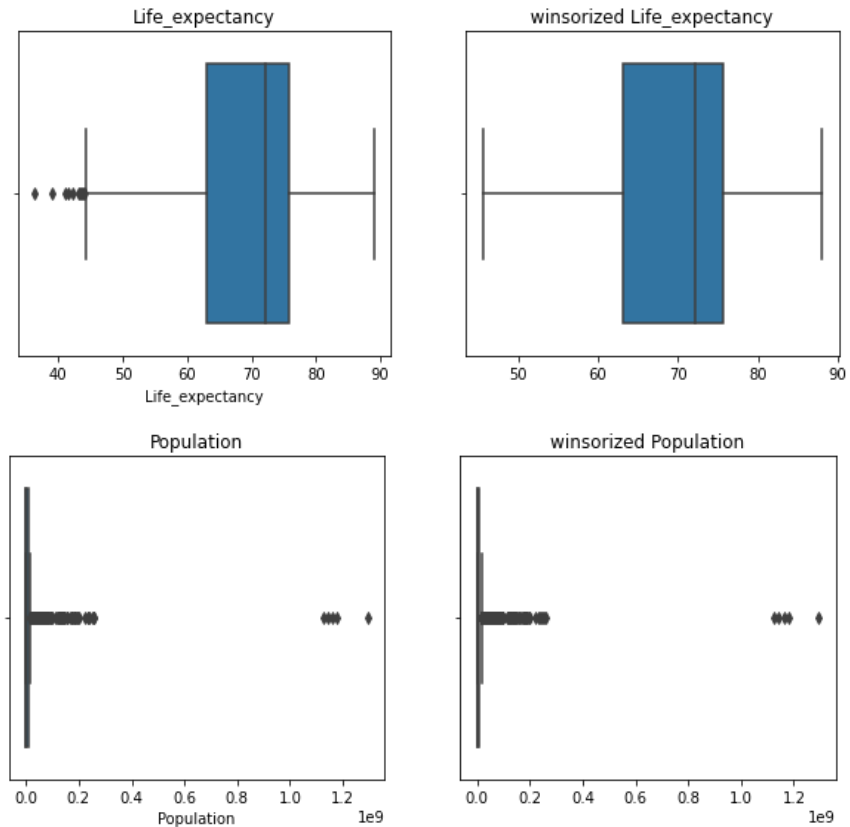
Data Wrangling

Outliers visualization and treatment

Winsorize method to treat Outliers

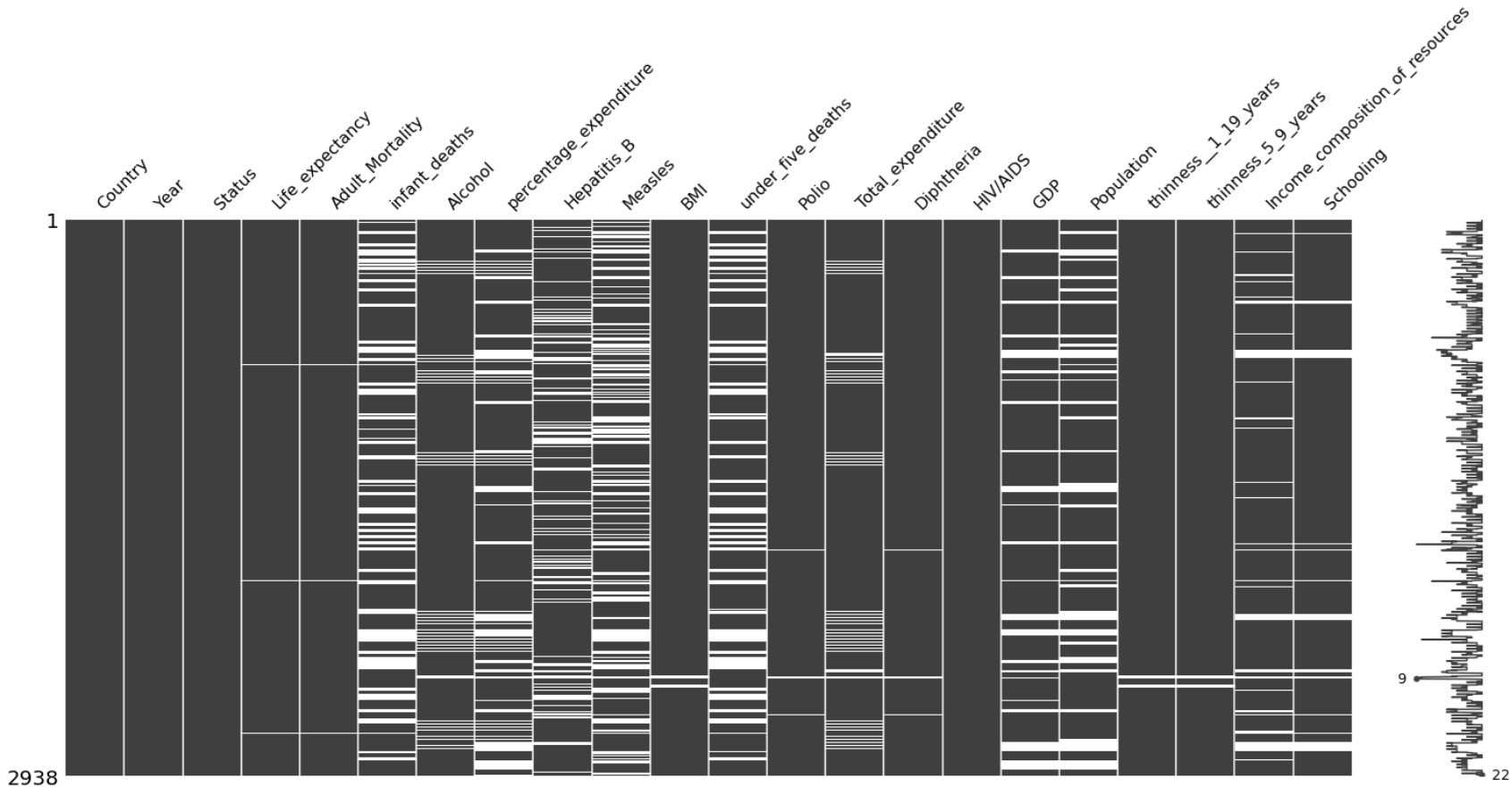
Dataset has 2938 observations and 22 columns (21 are independent variables)

Predicting variables were then divided into several broad categories: Immunization related factors, Mortality factors, Economical factors, and Social factors.

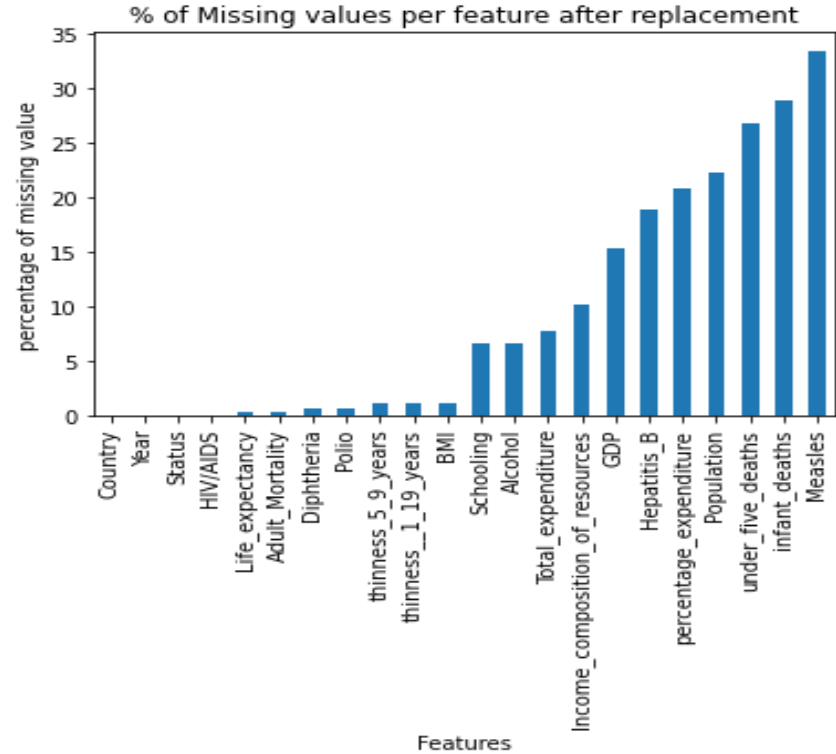
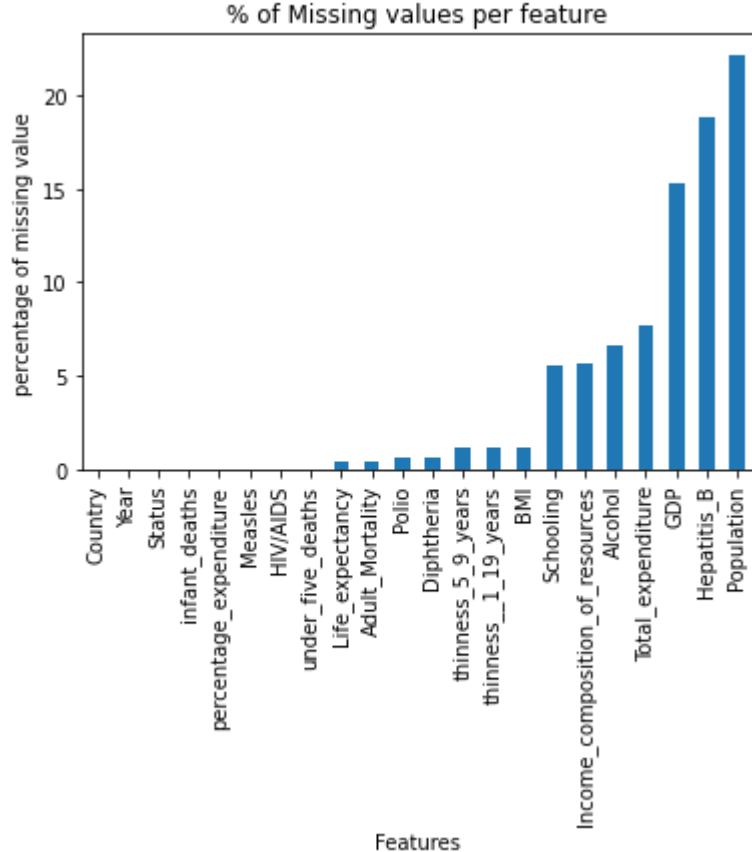


Box plot before and after winsorize application (0.01, 0.02)

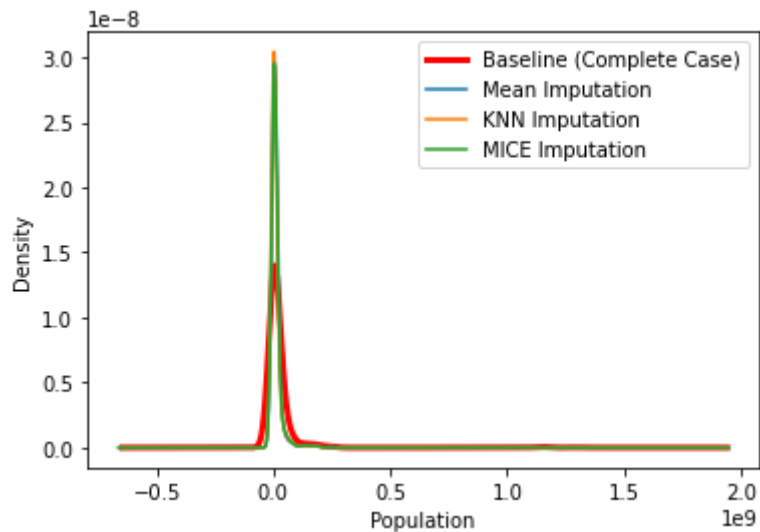
Assessing and treating Missing Value



Assessing and treating Missing Value



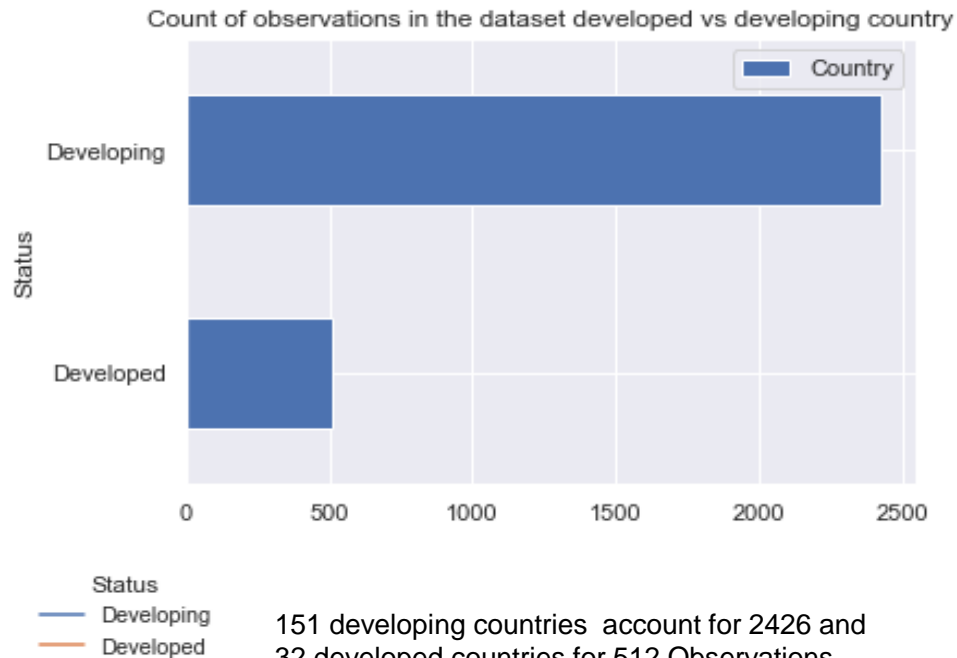
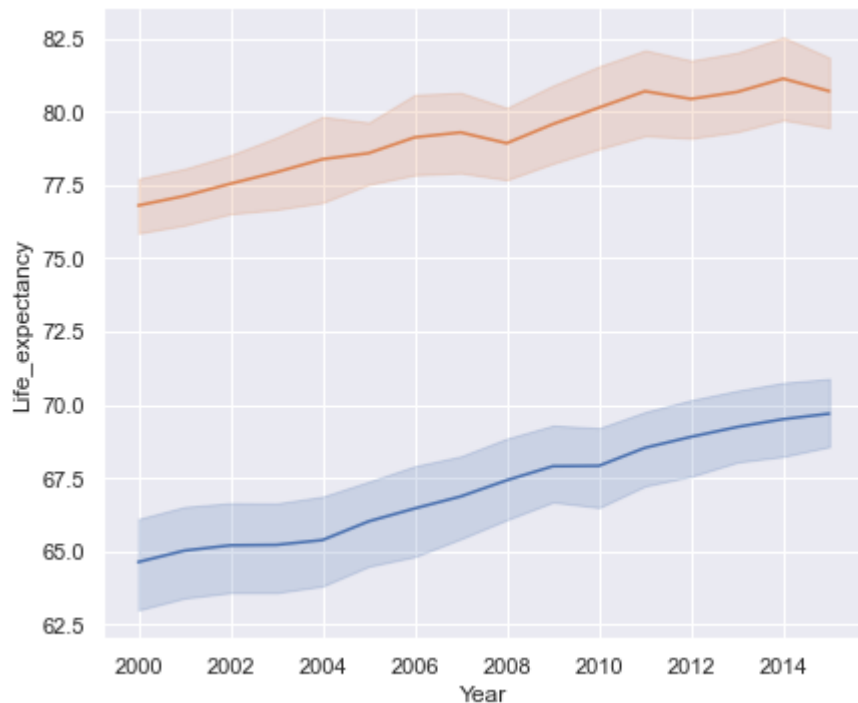
Original state of data with missing value and after the replacement of the uncommon type with NAN



The best imputation technique is: MICE Imputation

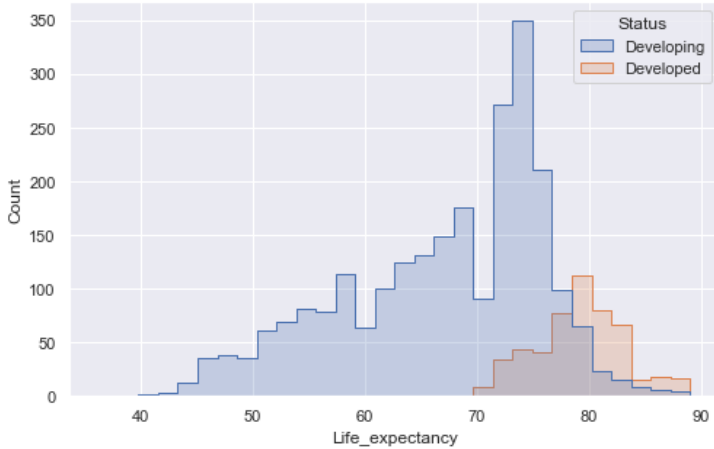
Exploratory Data Analysis

What is the trend of life expectancy?



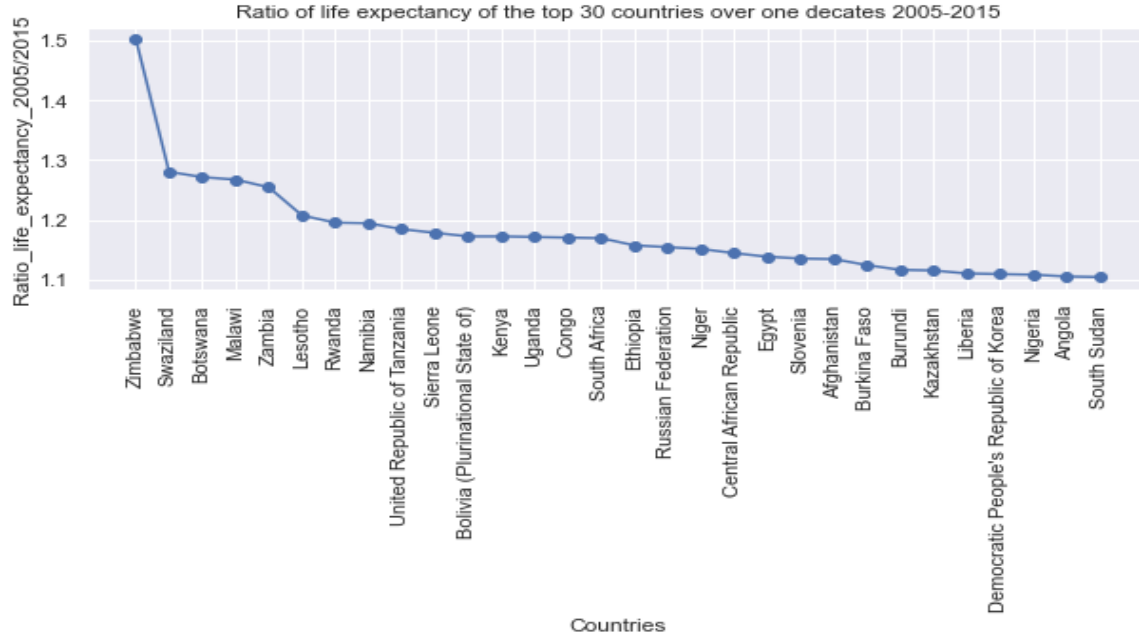
Exploratory Data Analysis

How does the distribution of life expectancy look like?



Min in developed countries is 69 /developing countries is 39.
Maxi in developed countries is 89 , the same in developing countries.

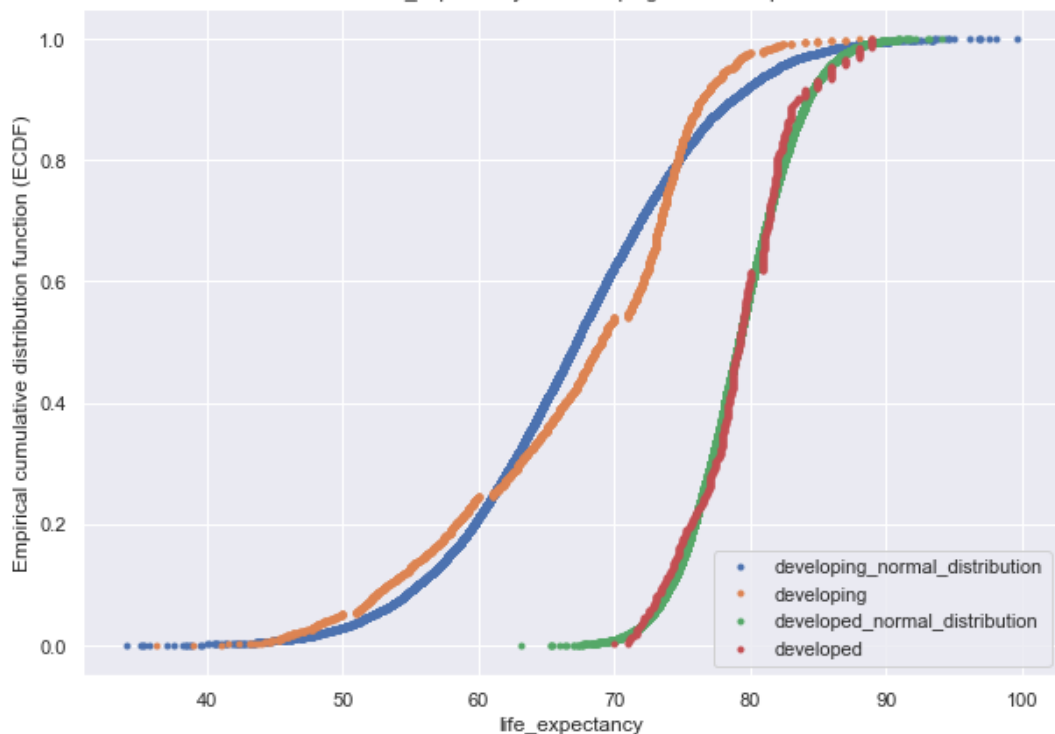
How was the life expectancy over one decade (2005 to 2015) ?



Exploratory Data Analysis

Statistical Analysis

Distribution of life_expectancy of Developing and Developed Countries



$p = 3.4988e-34$

The null hypothesis can be rejected

Confidence Interval of Life expectancy at 95%

*CI developing countries : [64.834, 68.364]

*CI developed countries : [77.911, 79.451]

Exploratory Data Analysis

Statistical Analysis

1- state the hypothesis

Null Hypothesis: the average mean of life expectancy from developed countries is always greater than the one of developing Countries

H_0 : mean avg of LE _developed = mean avg of LE _developing

H_1 : mean avg of LE _developed \neq mean avg of LE _developing

2- state the significance level (here we set the threshold for the test)

$\alpha = 0.05$ or 5% $z = 1.96$ for one tail, and $z = 1.64$ for two tail

3- identify the test statistic

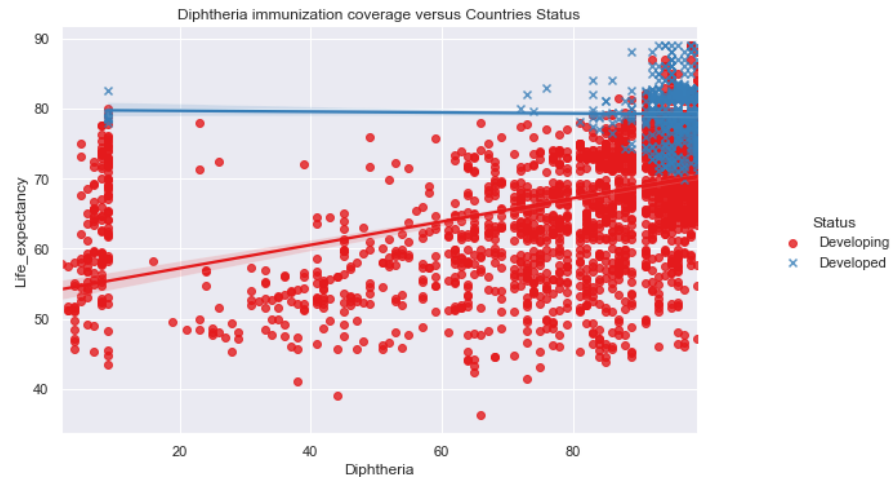
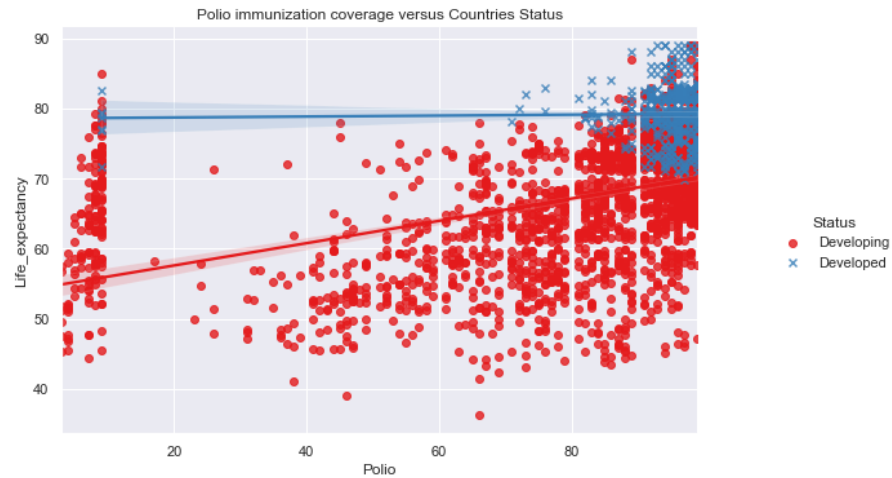
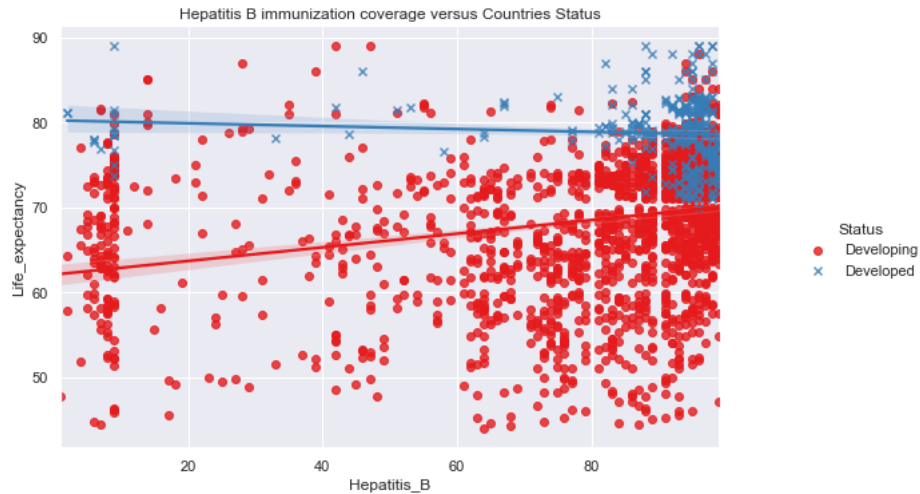
we conduct a Z test for 2 independants samples,

4- Conclusion:

From This result, we reject the null hypothesis, we found that there is a significant difference between the mean average life expectancy of developed countries to that of developing countries.

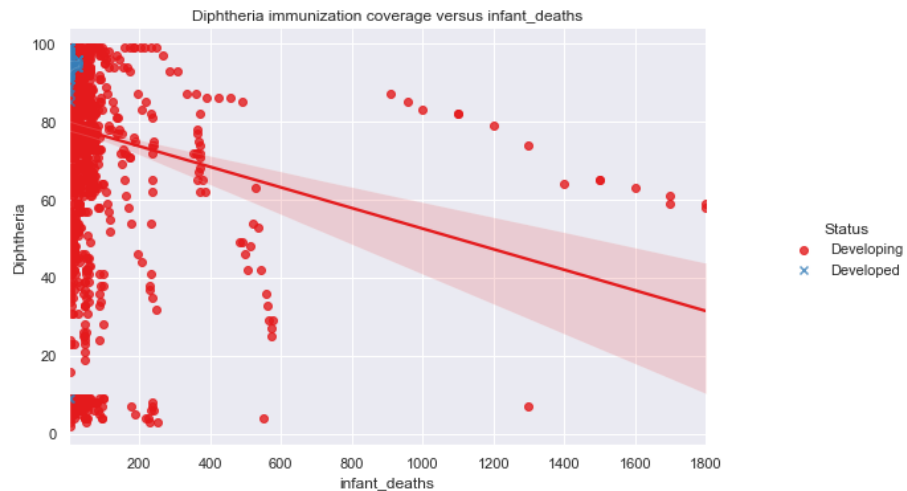
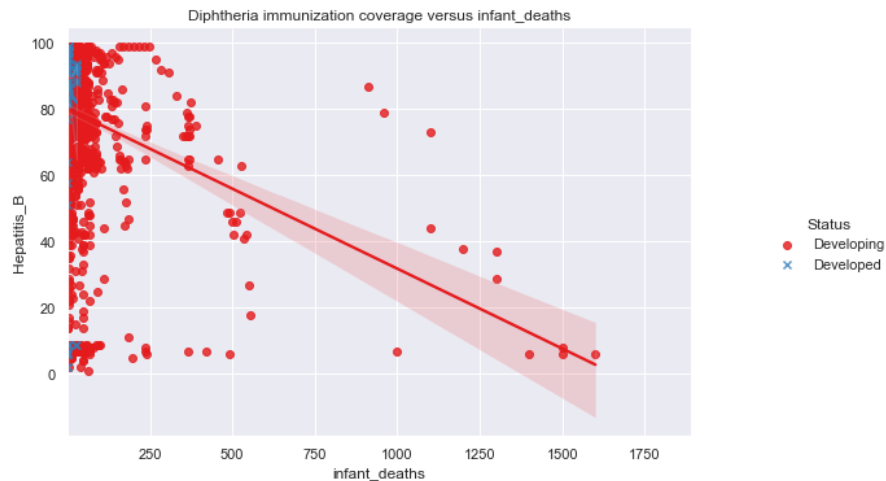
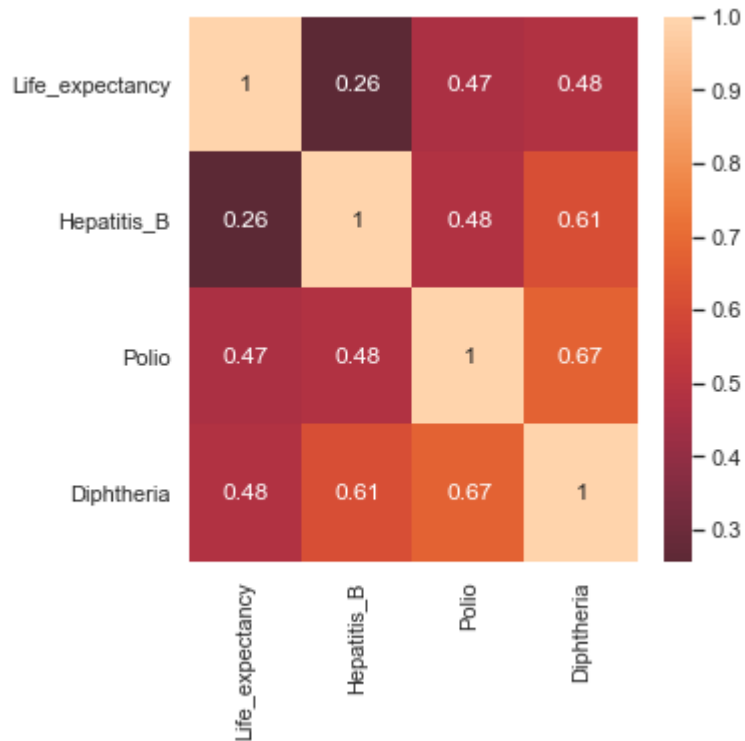
Exploratory Data Analysis

Immunization and life expectancy

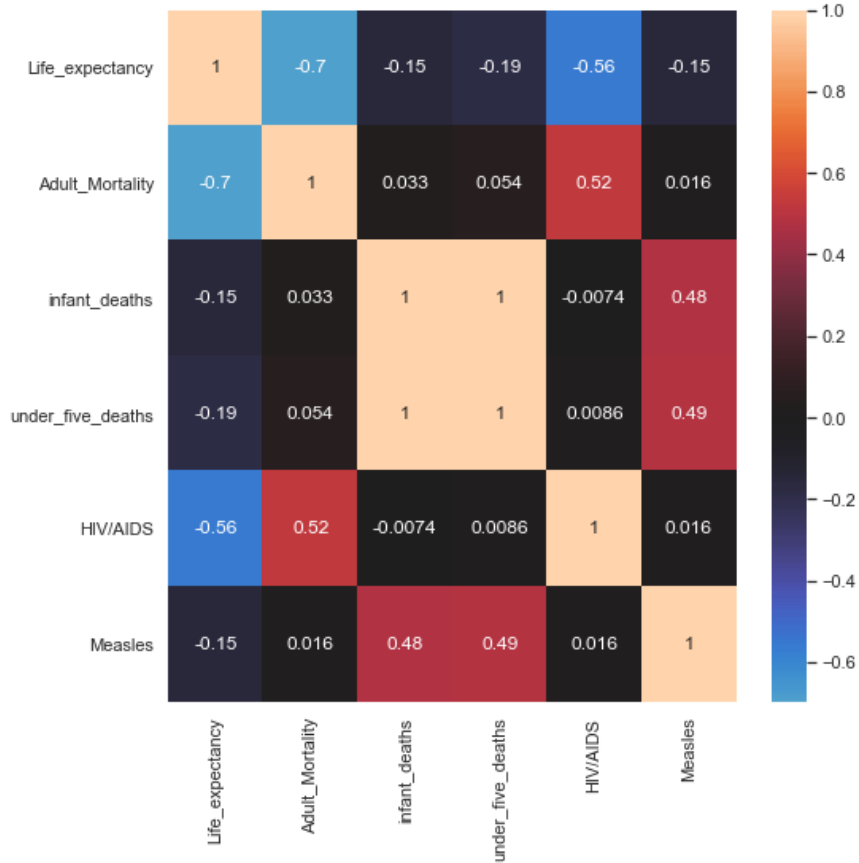


Exploratory Data Analysis

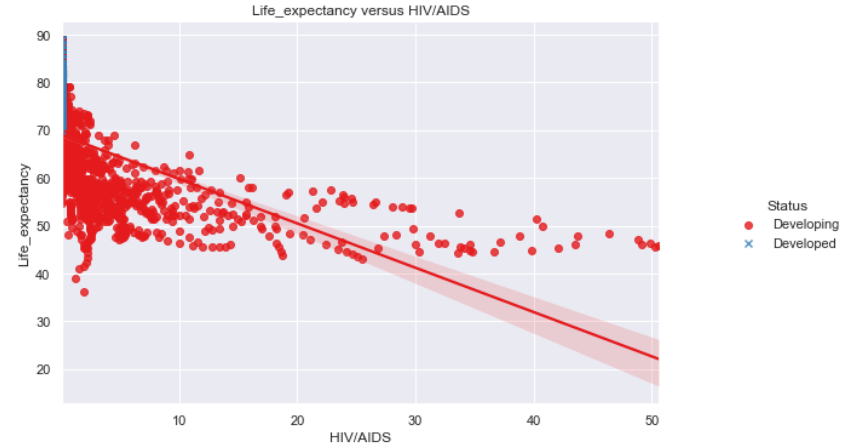
Immunization and life expectancy , and mortality factors



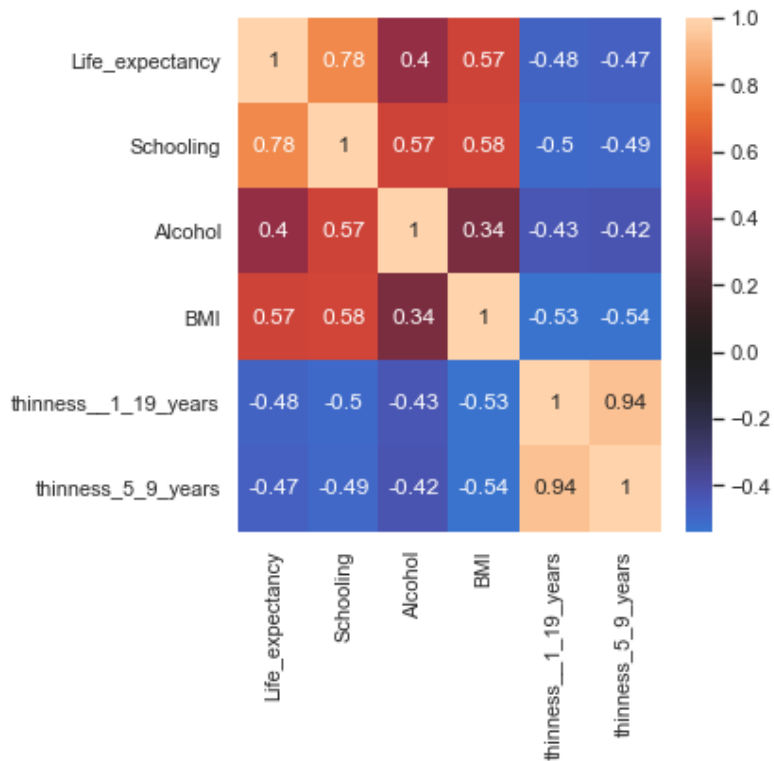
Exploratory Data Analysis



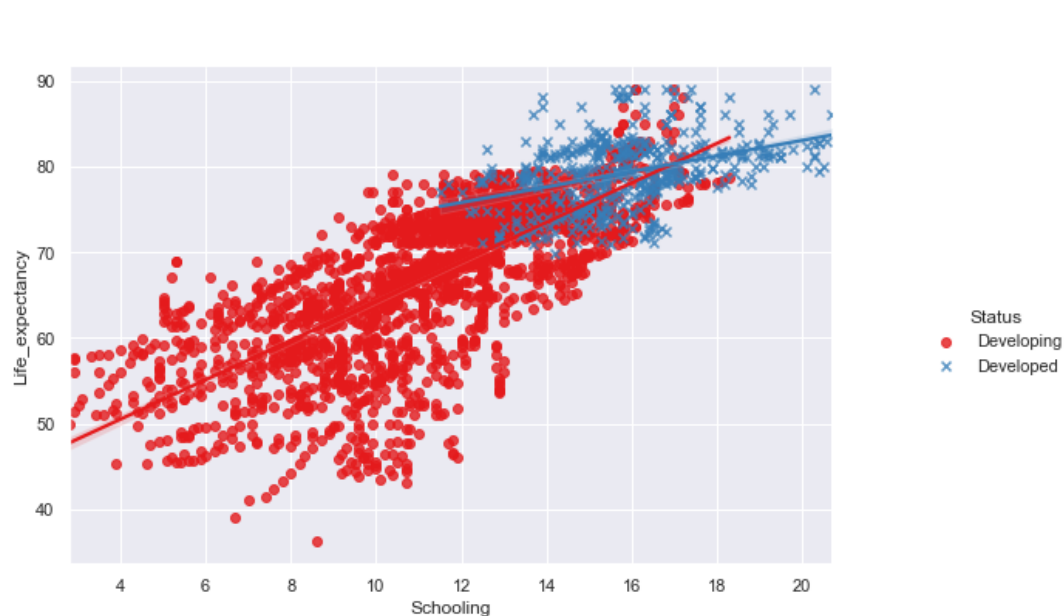
Mortality_factors and Life_expectancy



Exploratory Data Analysis

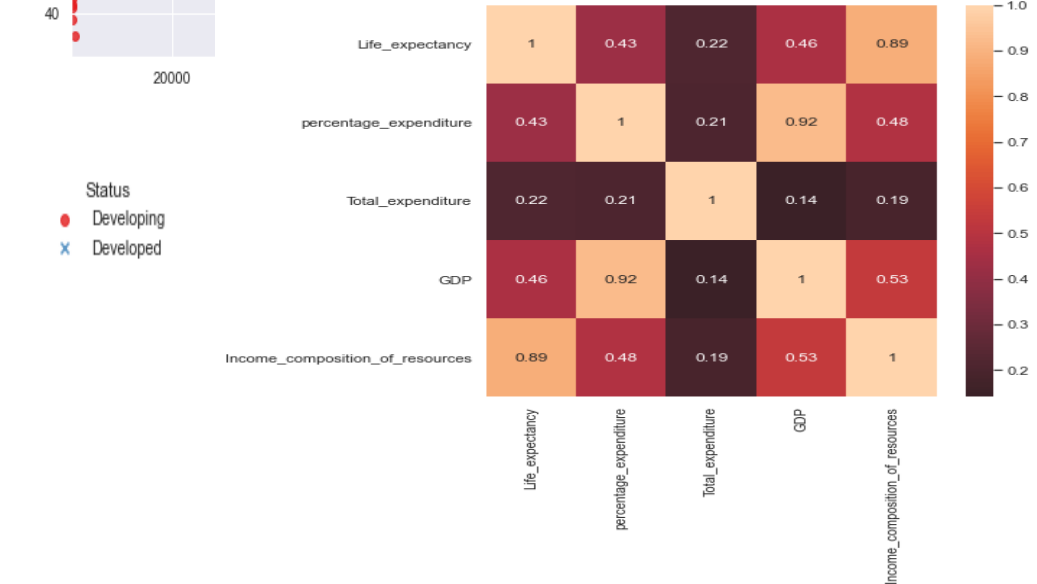
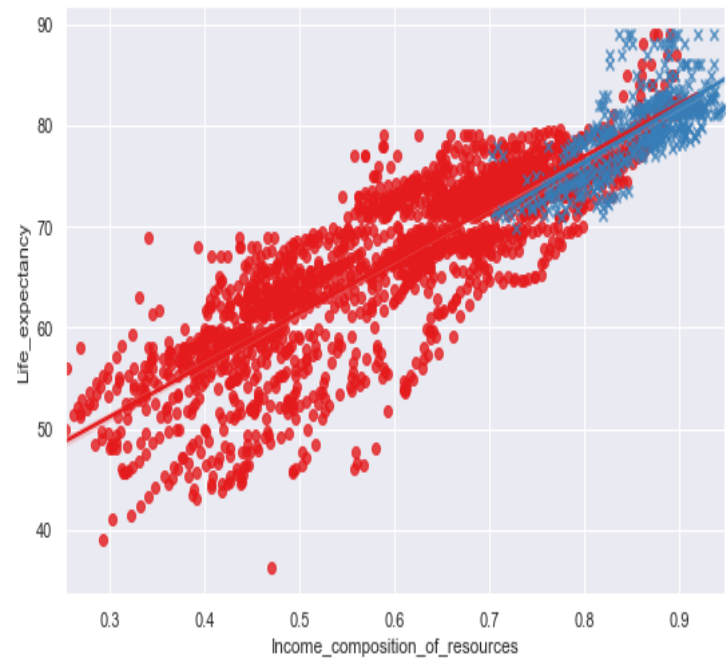


Social Factors and life_expectancy



Exploratory Data Analysis

Economical_factors and Life_expectancy



Exploratory Data Analysis

Population and life expectancy



Preprocessing the data

Label Encoder of categorical variable with one hot_encoder

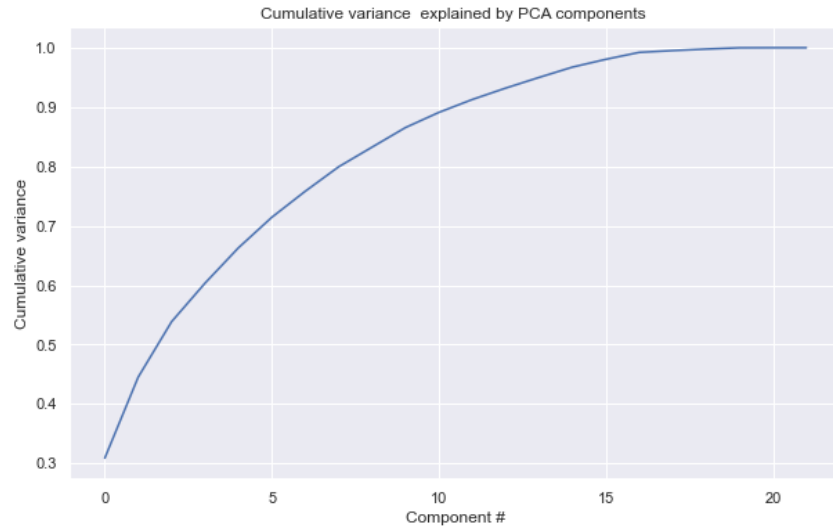
Imputing missing value (Mice imputation)

Divide in test set and train set (30% , 70%)

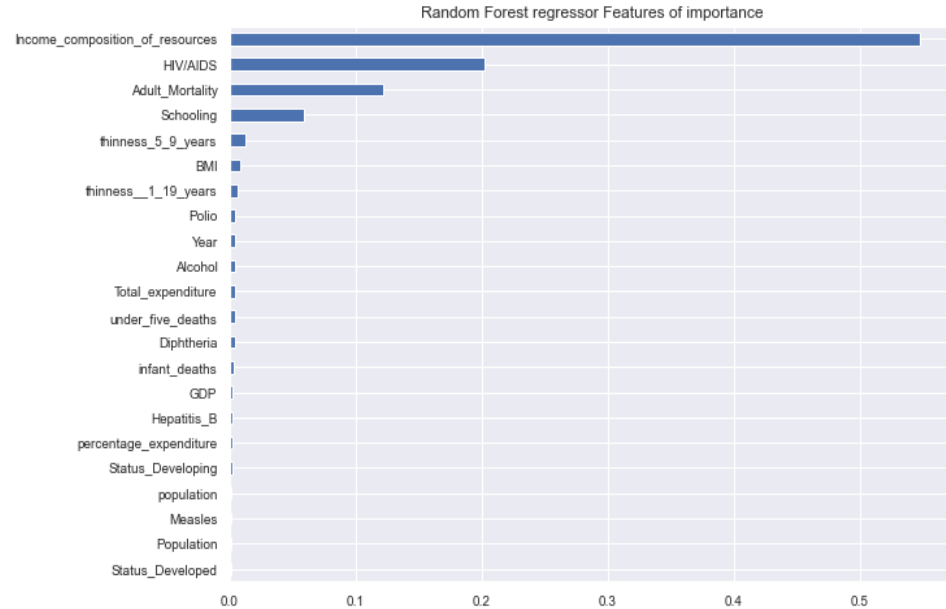
Scaling the dataset

PCA transformation

Preprocessing the data



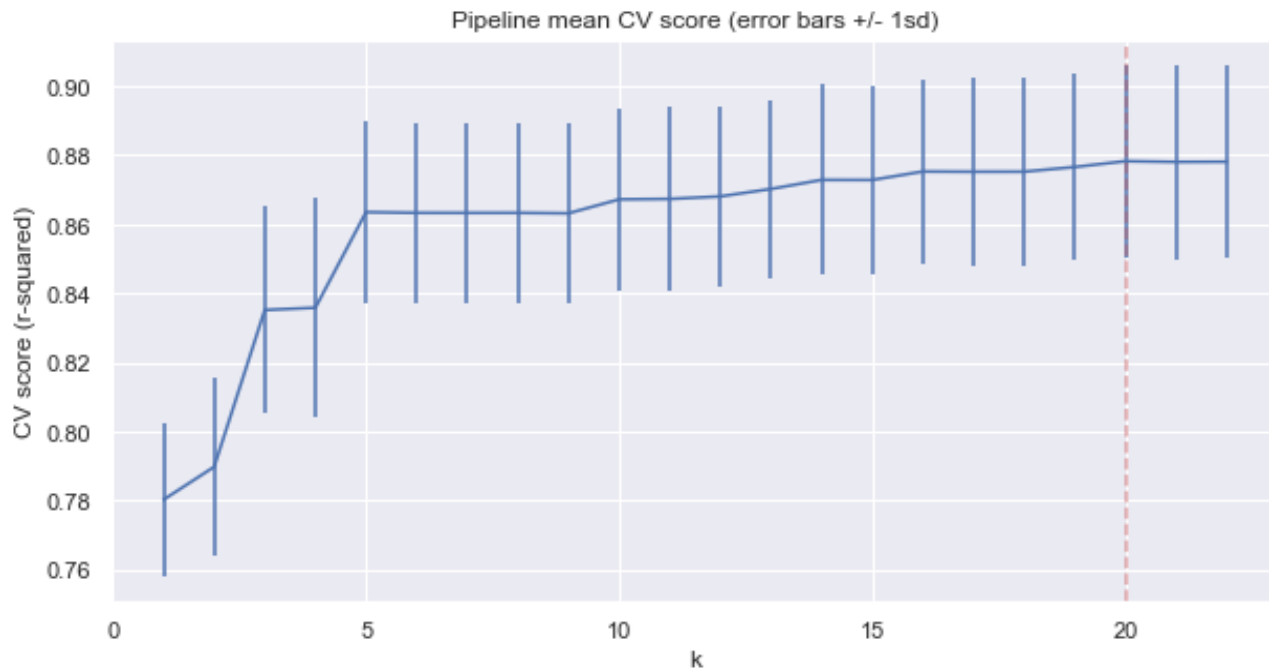
Note: The first five components seem to account for over 75% of the variance, and 10 components is 92% of the variance



Modelling

```
pipe = make_pipeline(  
    IterativeImputer(),  
    StandardScaler(),  
    SelectKBest(f_regression),  
    LinearRegression())
```

```
{'selectkbest__k': 20}
```



The above suggests a good value for k is 20

Modeling

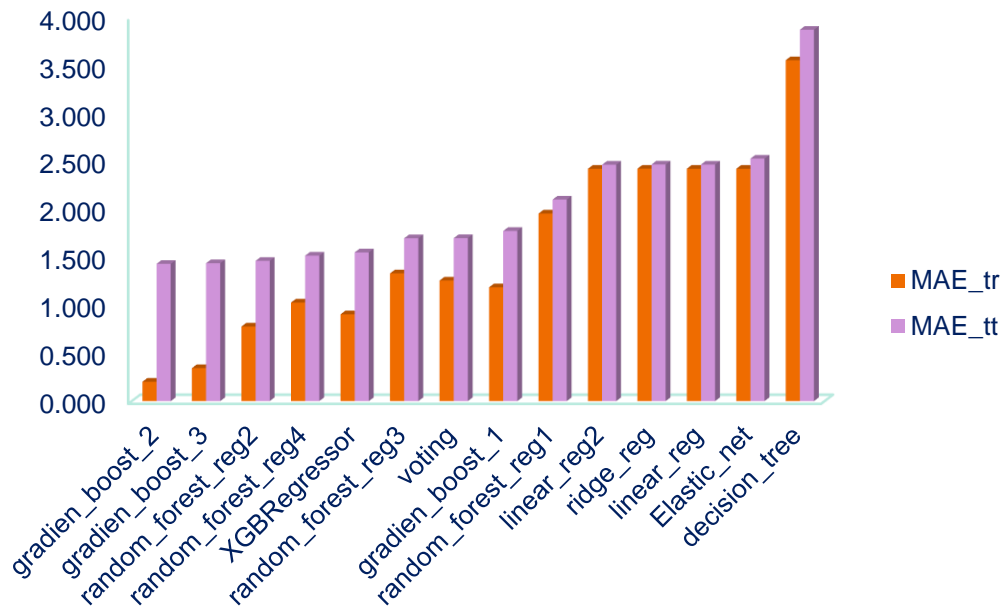
model	model_definition
linear_reg	<pre> Pipeline(steps=[('iterativeimputer', IterativeImputer()), ('standardscaler', StandardScaler()), ('selectkbest', SelectKBest(k=19, score_func=<function f_regression at 0x00000025E60390C10>)), ('linearregression', LinearRegression()))]) </pre>
linear_reg2	<pre> Pipeline(steps=[('iterativeimputer', IterativeImputer()), ('standardscaler', StandardScaler()), ('pca', PCA(n_components=19)), ('linearregression', LinearRegression()))]) </pre>
ridge_reg	<pre> Pipeline(steps=[('iterativeimputer', IterativeImputer()), ('standardscaler', StandardScaler()), ('ridge', Ridge(alpha=0.5))]) </pre>
Elastic_net	<pre> ElasticNet(alpha=0.0001, l1_ratio=0.4) </pre>
decision_tree	<pre> DecisionTreeRegressor(max_depth=4, max_features=0.2, min_samples_leaf=0.1, random_state=1) </pre>

model	model_definition
random_forest_reg1	<pre> RandomForestRegressor(max_depth=4, max_features=0.4, n_estimators=200, n_jobs=-1, random_state=1) </pre>
random_forest_reg2	<pre> RandomForestRegressor(max_depth=10, max_features=0.4, n_estimators=200, n_jobs=-1, random_state=1) </pre>
random_forest_reg3	<pre> RandomForestRegressor(max_depth=7, max_features=0.3, n_jobs=-1, random_state=1) </pre>
random_forest_reg4	<pre> RandomForestRegressor(max_depth=8, max_features=0.6, n_estimators=200, random_state=1) </pre>
gradien_boost_1	<pre> GradientBoostingRegressor(n_estimators=150, random_state=1) </pre>
gradien_boost_2	<pre> GradientBoostingRegressor(learning_rate=0.08249999999999999, max_depth=10, max_features=0.6000000000000001, min_samples_leaf=8, min_samples_split=10, n_estimators=118) </pre>
gradien_boost_3	<pre> GradientBoostingRegressor(learning_rate=0.08249999999999999, max_depth=7, max_features=0.8, min_samples_leaf=4, min_samples_split=12, n_estimators=150, random_state=1) </pre>
XGBRegressor	<pre> XGBRegressor(base_score=0.5, booster='gbtree', colsample_bylevel=1, colsample_bynode=1, colsample_bytree=1, enable_categorical=False, gamma=0, gpu_id=-1, importance_type=None, interaction_constraints="", learning_rate=0.04, max_delta_step=0, max_depth=5, min_child_weight=1, missing=nan, monotone_constraints=()), n_estimators=200, n_jobs=-1, num_parallel_tree=1, predictor='auto', random_state=0, reg_alpha=0, reg_lambda=1, scale_pos_weight=1, subsample=1, tree_method='exact', validate_parameters=1, verbosity=None) </pre>
voting	<pre> VotingRegressor(estimators=[('gb', GradientBoostingRegressor(random_state=47)), ('rf', RandomForestRegressor(random_state=47)), ('lr', LinearRegression())]) </pre>

Modeling

R square score in train set and test set

Model Evaluation MAE of tran/test



model	R2_tr	R2_tt
gradien_boost_2	0.999	0.950
gradien_boost_3	0.997	0.947
random_forest_re g2	0.986	0.945
random_forest_re g4	0.976	0.942
XGBRegressor	0.982	0.941
random_forest_re g3	0.962	0.931
voting	0.965	0.930
gradien_boost_1	0.970	0.929
random_forest_re g1	0.919	0.900
linear_reg2	0.882	0.868
linear_reg	0.882	0.868
ridge_reg	0.882	0.868
Elastic_net	0.882	0.863
decision_tree	0.743	0.682

Conclusion

- Life expectancy has increased over years in both developed and developing countries
- The mean average of the life expectancy of developed countries is generally higher compared to that of developing countries
- However, the ratio of LE over the decade of 2005 to 2015 showed that life expectancy in developing countries has greatly increased.
- It has been highlighted that immunization has impacted the improvement of life expectancy in a developing country, as well as the reduction in infant deaths.
- The analysis revealed that economic factors play an important role in the system, it is why countries with higher income resources and GDP tend to have high life expectancy even if the population is big. In developing countries, an increase in the population tends to impact negatively life expectancy.
- Many (14) regression models have been developed to predict expectancy, the chosen one is Gradient boost with MAE of 0.202 on train set and 1.431 on the test set, R square is 0.94 on the test set.