LIFE EXPECTANCY:

DOES IMMUNIZATION MATTER?

Mireille P. Feudjio T.

Outline

- Context
- Data Wrangling
- Exploratory data analysis
- Modelling
- Conclusion

Context

- Past 15 years, huge development of health section - reduction human mortality
- Developing countries (30 years)
- Assessment of population health – life expectancy
- Vaccination – improving life expectancy
- Pandemic
- Impact of vaccination on life expectany?
- Does immunization matter?
- Factors contribution in life expectancy ?

- The dataset (life expectancy, health factors for 193 countries) 2000-2015.

within a month, the present project assessed the contribution of features on life expectancy with a special focus on immunization factors, and develop a regression model to predict life expectancy.

Constraints : The dataset has important missing values to handle .

Who care? Governments are concerned about the health of the population and are constantly looking for ways to improve life expectancy.

Data acquision and wrangling (1)

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

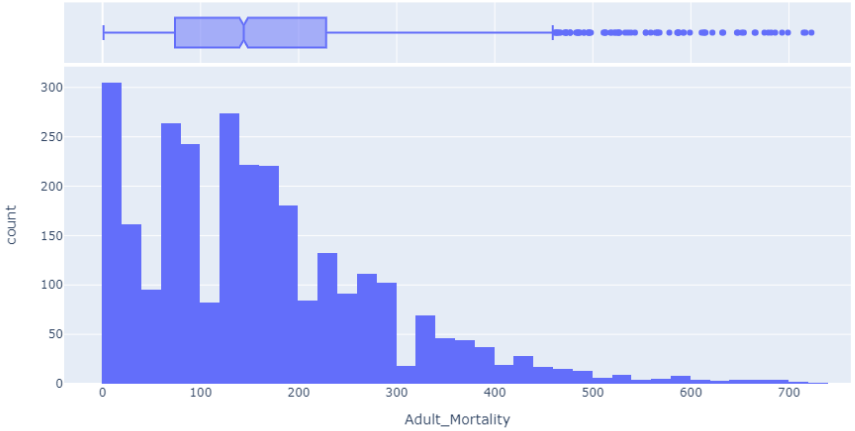
Data wrangling:

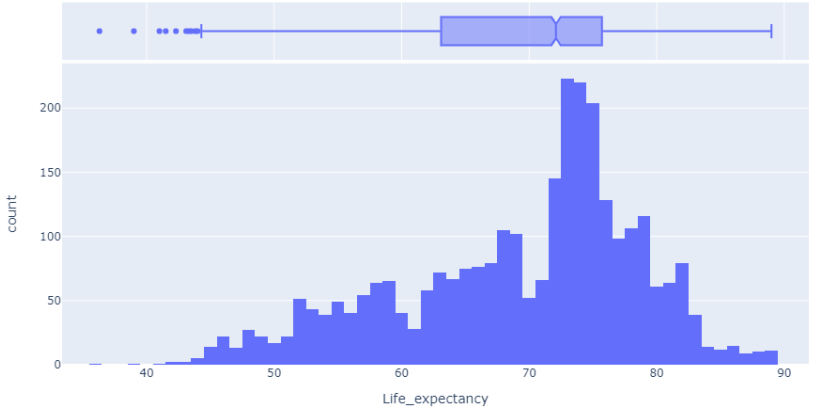
- Renaming column
- Replacing uncommon value with NAN
- Check for data type
- Check for duplicate
- Analysis missing value
- Analyzing outliers

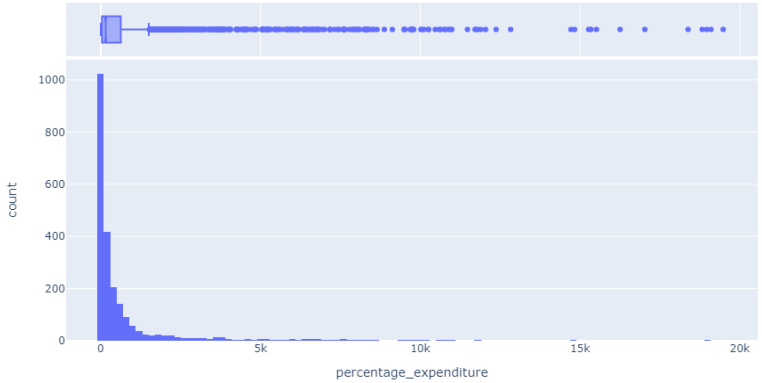
Mortality factors	Social factors
• Adult Mortality Rates of both sexes • Number of Infant Deaths • Number of under-five deaths • Deaths births HIV/AIDS (0-4 years) • Measles - number of reported cases	• Alcohol, recorded per capita (15+) consumption • Number of years of Schooling(years) • Average Body Mass Index • Prevalence of thinness -Age 10 to 19 (%) • Prevalence of thinness - Age 5 to 9(%)
Immunization factors	**Economic factors**
• Hepatitis B (HepB) immunization coverage • Polio (Pol3) immunization coverage • Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage	• Gross Domestic Product (in USD) • General government expenditure on health as a percentage of total government expenditure (%) • Expenditure on health as a percentage of Gross Domestic Product per capita (%) • Human Development Index in terms of income composition of resources (index ranging from 0 to 1)

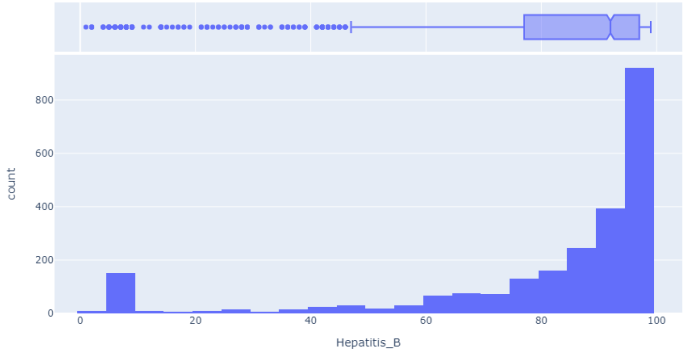
Data wrangling (2)

Outliers visualization and treatment





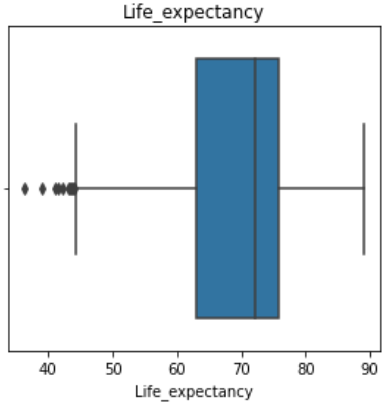


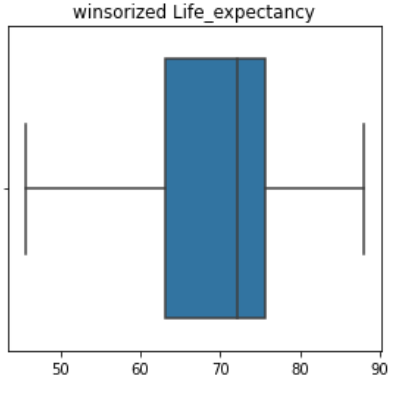


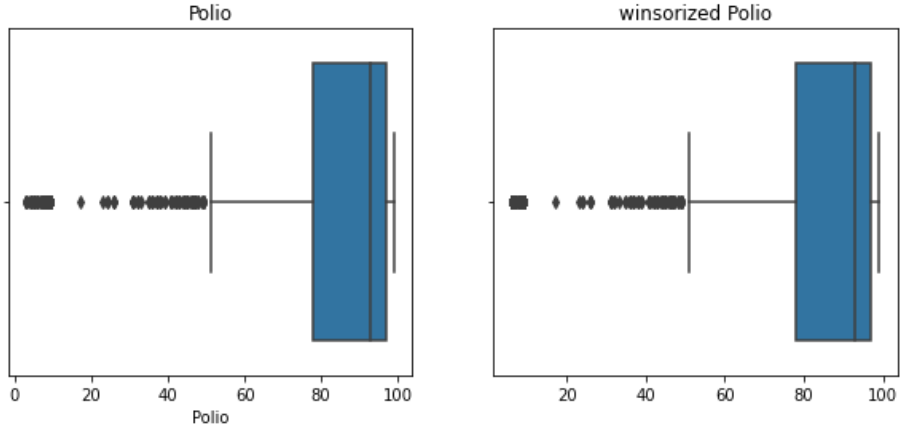
Data wrangling (3)

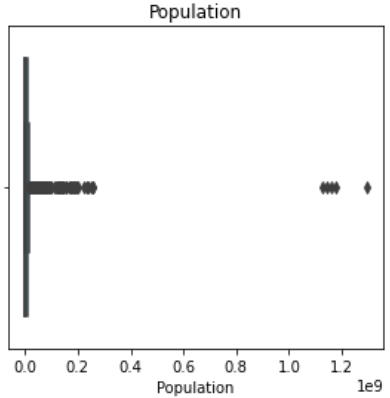
Outliers visualization and treatment

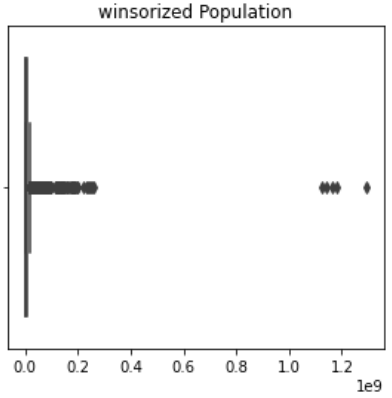
Winsorize method to treat Outliers







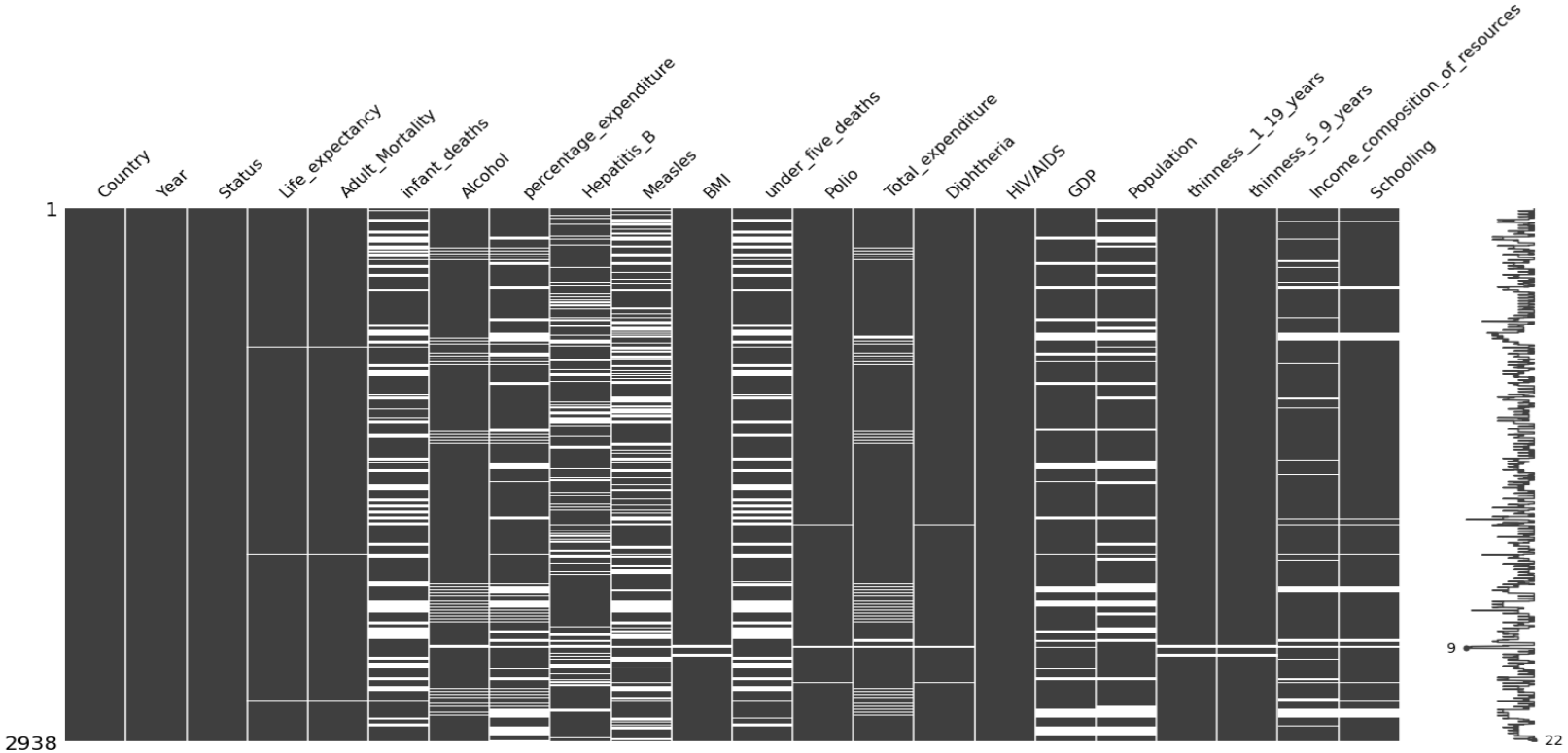




Box plot before and after the 98% winsorize application (1% top and 1% bottom)

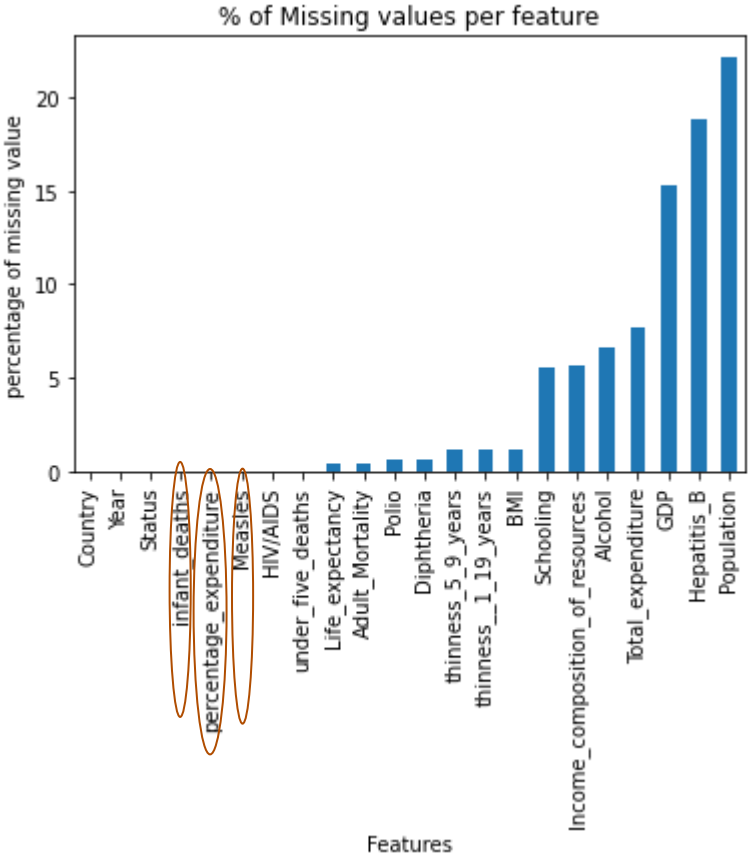
Data wrangling (4)

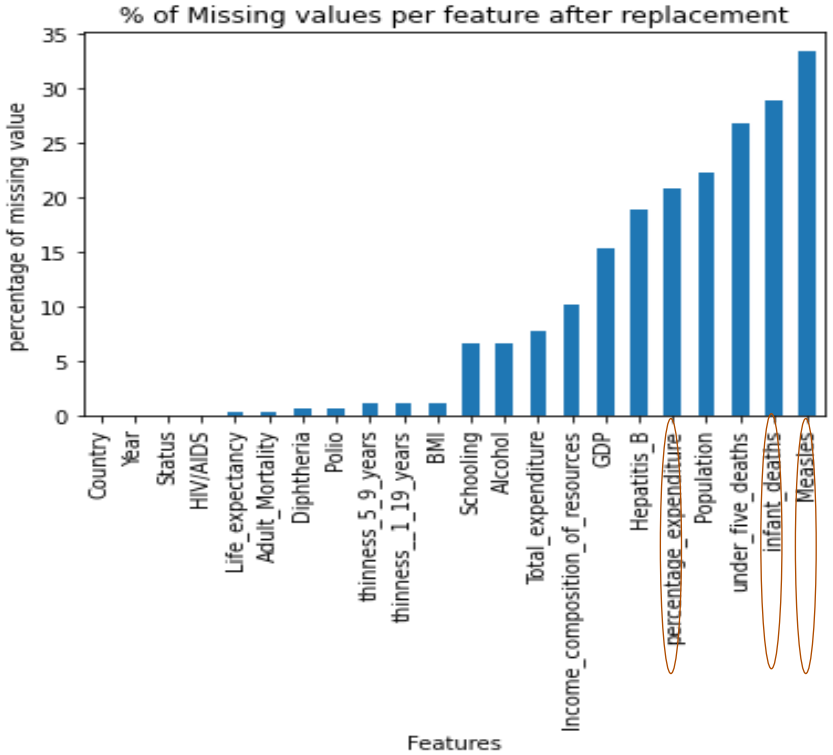
Assessing and treating missing value



Data wrangling (5)

Assessing and treating missing value





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

Data wrangling (6)

Assessing and treating missing value

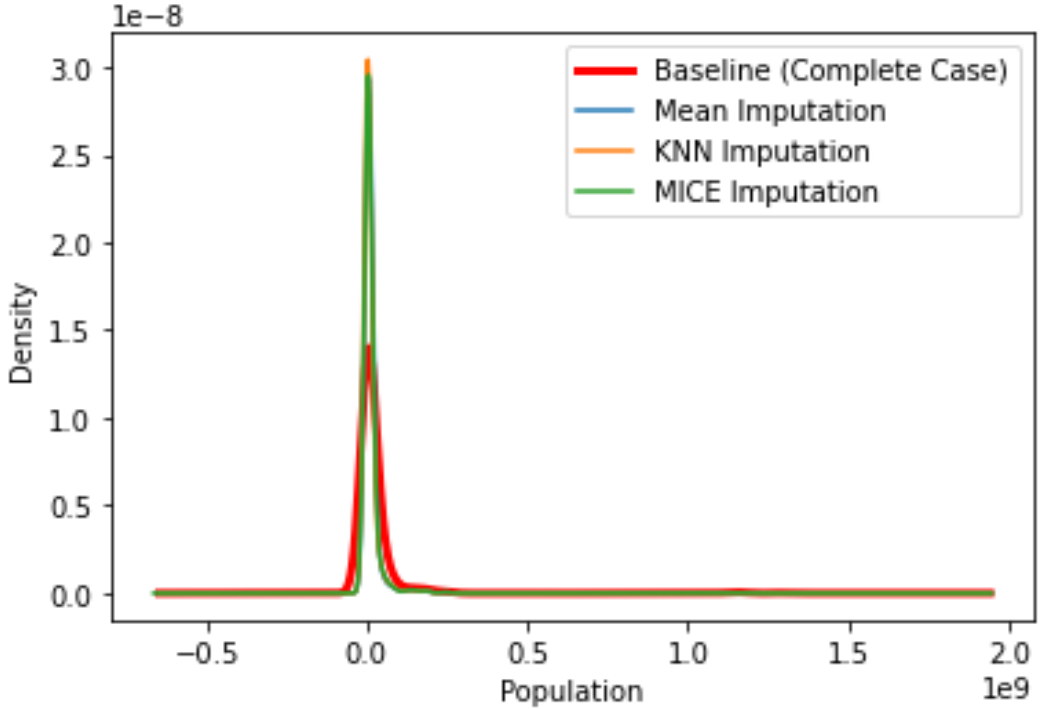
- loss of **69%**
- Mean, mode, constant and median imputation.
- KNN imputation and , MICE imputation .
- Evaluation of the imputation methods.

Type of imputation	Adj. R-squared
Complete Case	0.899322
Mean Imputation	0.837261
KNN Imputation	0.857469
Mode Imputation	0.835953
Constant Imputation	0.665765
MICE Imputation	0.900985

Data wrangling (7)

Assessing and treating missing value

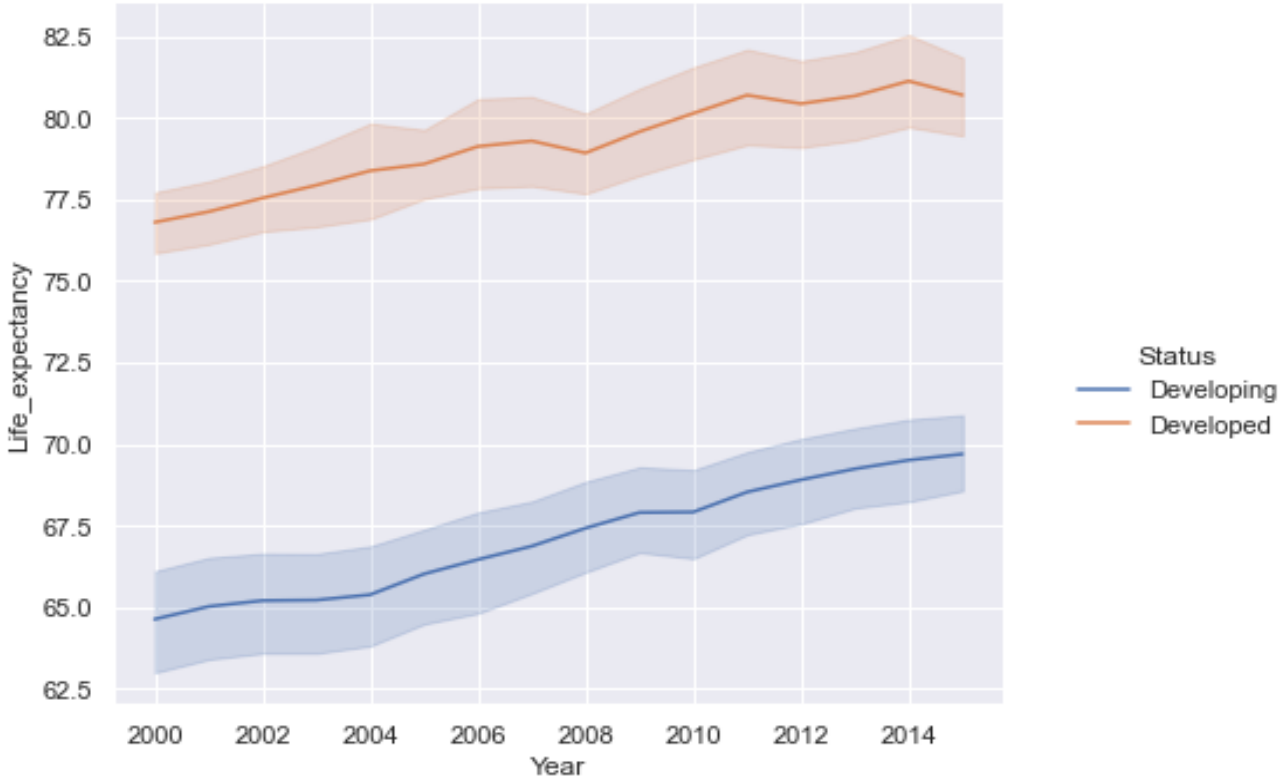
The best imputation technique is:
MICE Imputation



Visualizing Mean, Mice and KNN imputation on population feature

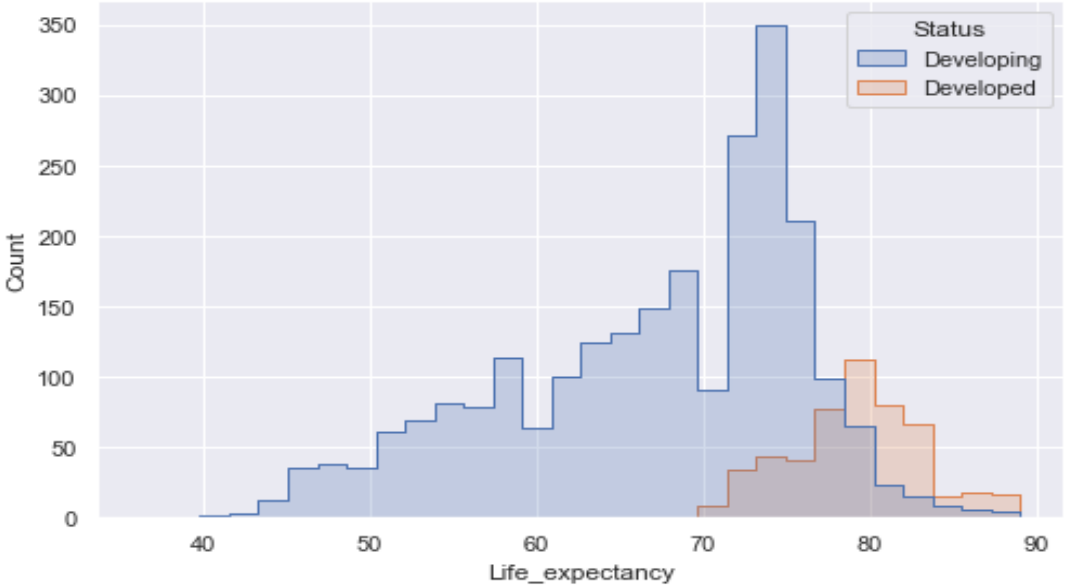
Exploratory Data Analysis (1)

What is the trend of life expectancy?



Exploratory Data Analysis (2)

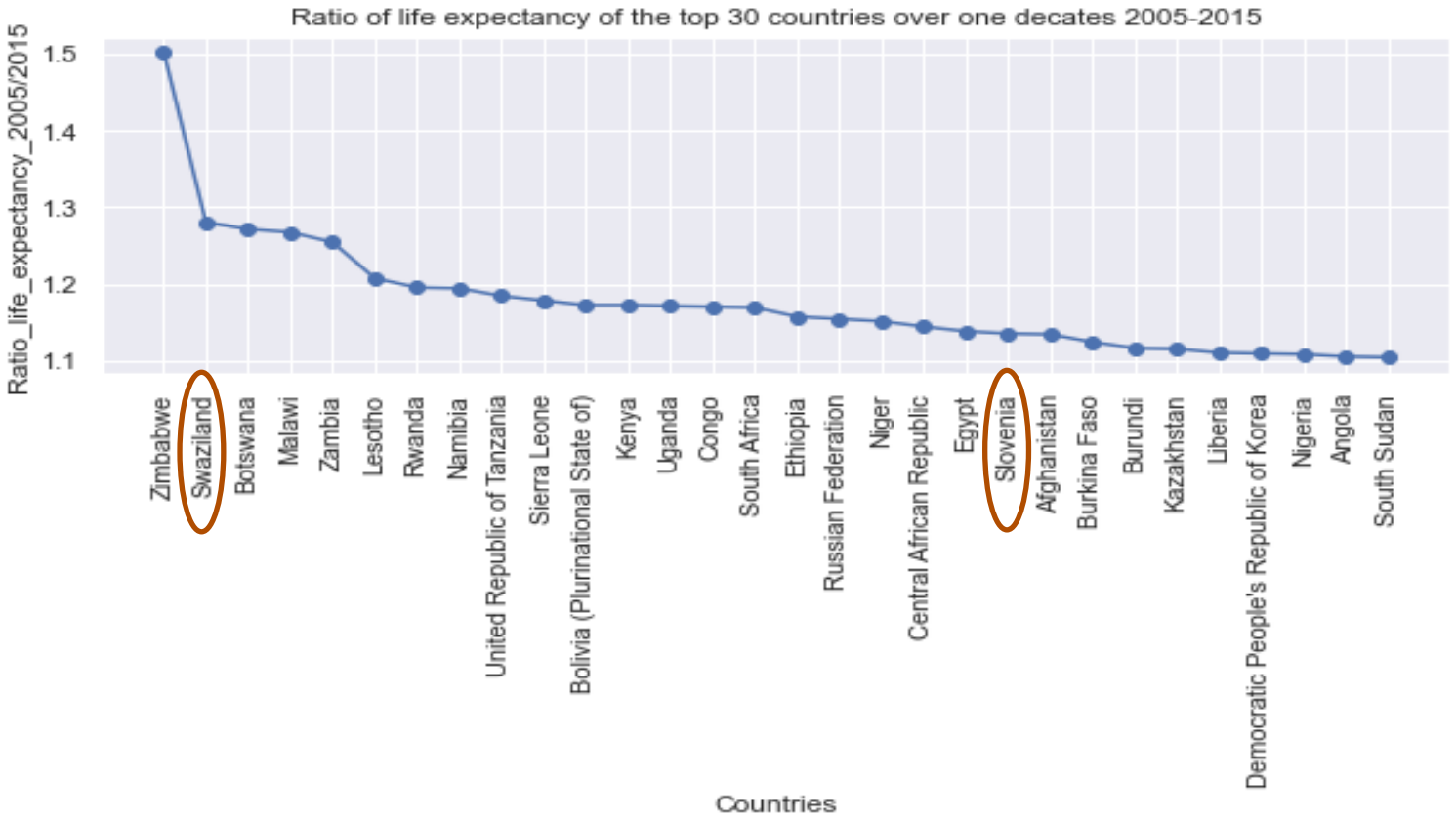
How is the distribution of life expectancy look like?



Countries types	count	mean	std	min	max
Developed	512.00	79.20	3.93	69.90	89.00
Developing	2416.00	67.11	9.00	36.30	89.00

Exploratory Data Analysis (3)

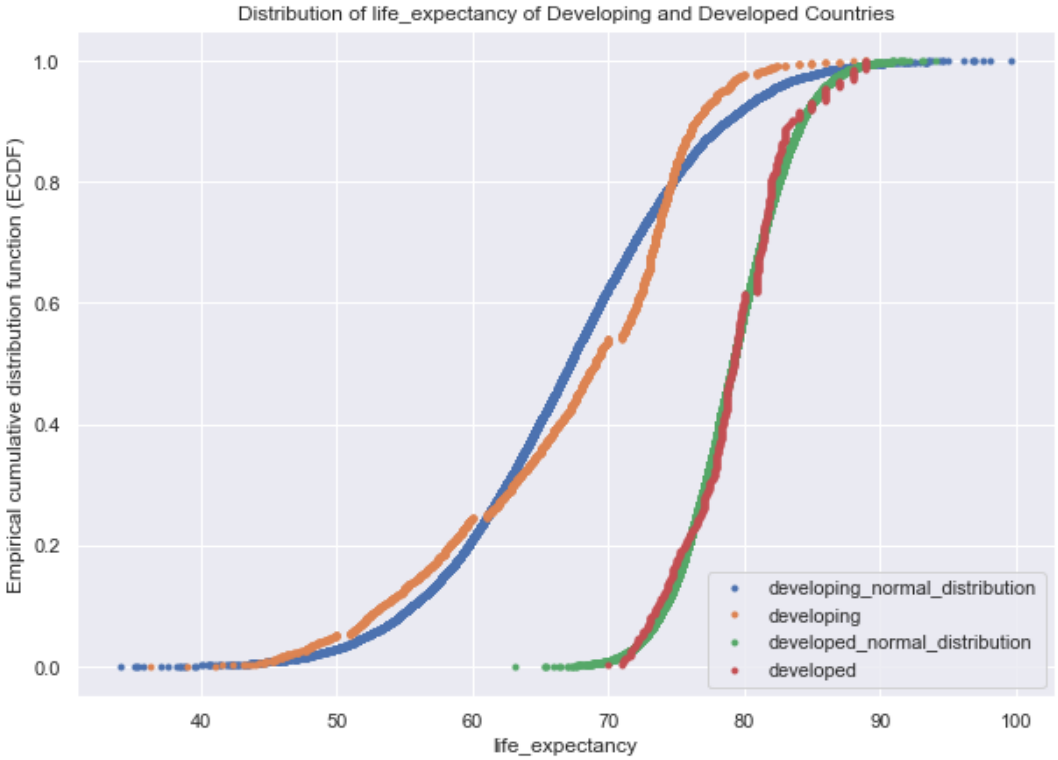
How was life expectancy over one decade (2005 to 2015) ?



Exploratory Data Analysis (4)

Statistical Analysis

What is the confidence interval of life expectancy at 95% ?



Apply the Central limit theoreme

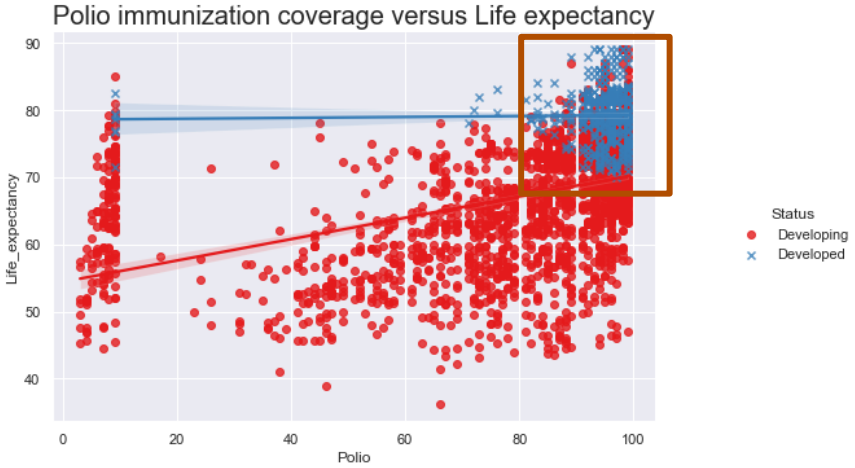
Confidence Interval of life expectancy at 95%

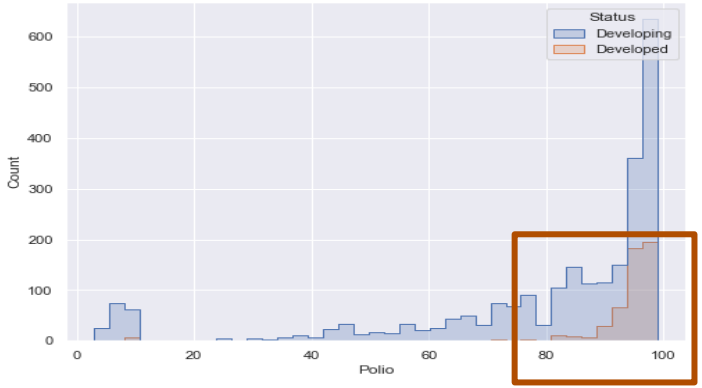
*CI developing countries : [64.834, 68.364]

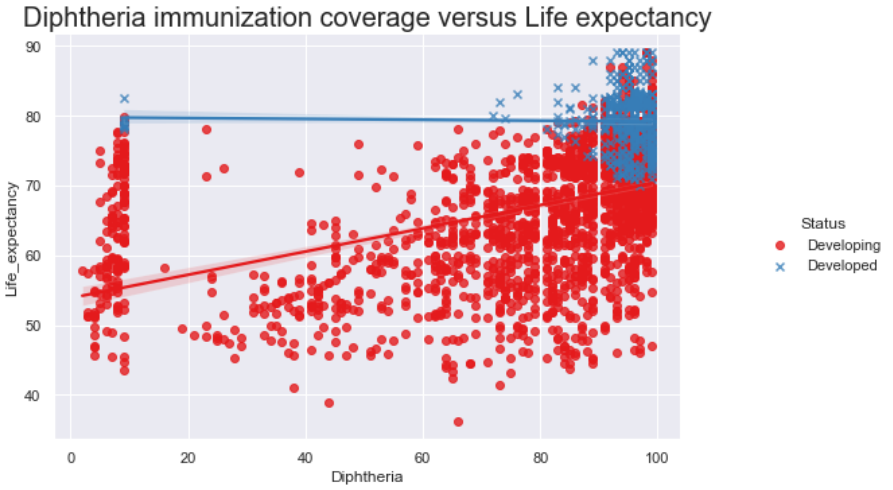
*CI developed countries : [77.911, 79.451]

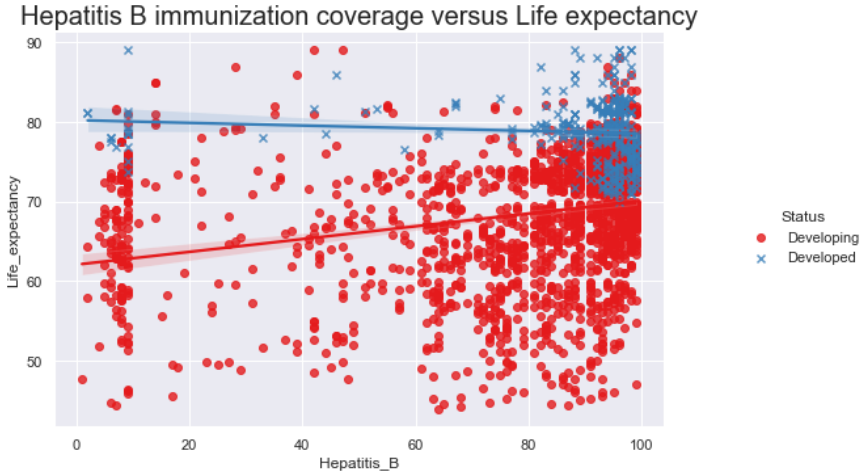
Exploratory Data Analysis (5)

Immunization and life expectancy



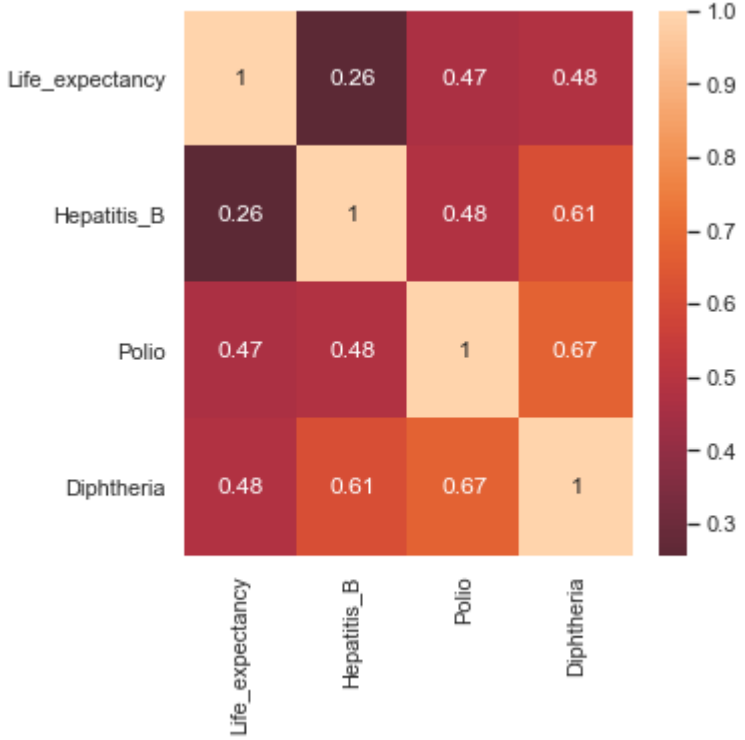


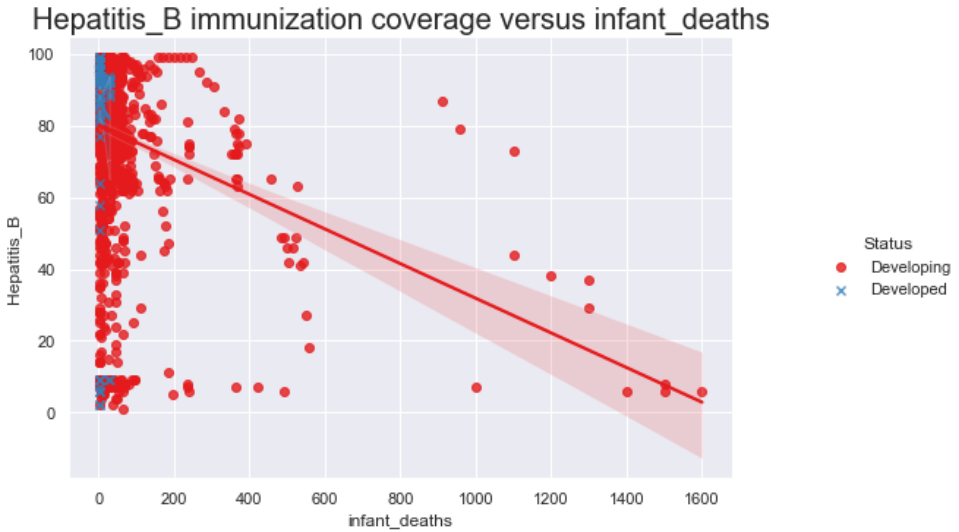


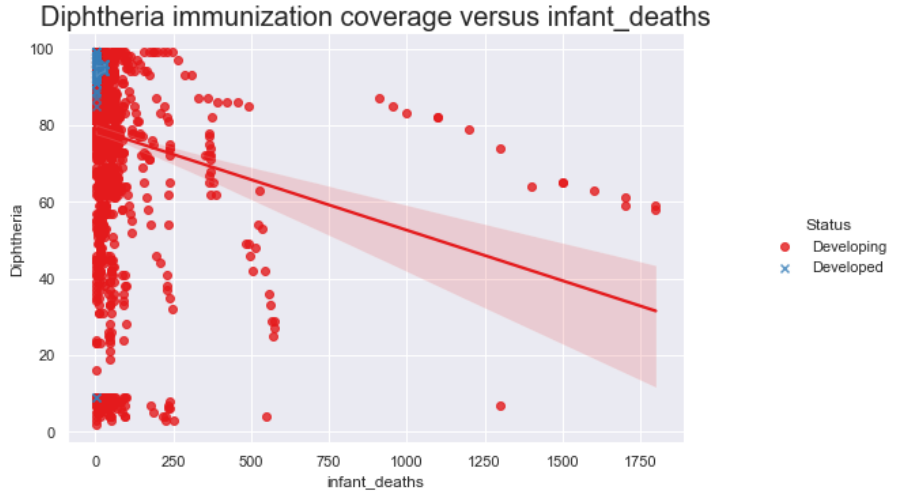


Exploratory Data Analysis (6)

Immunization and life expectancy , and mortality factors

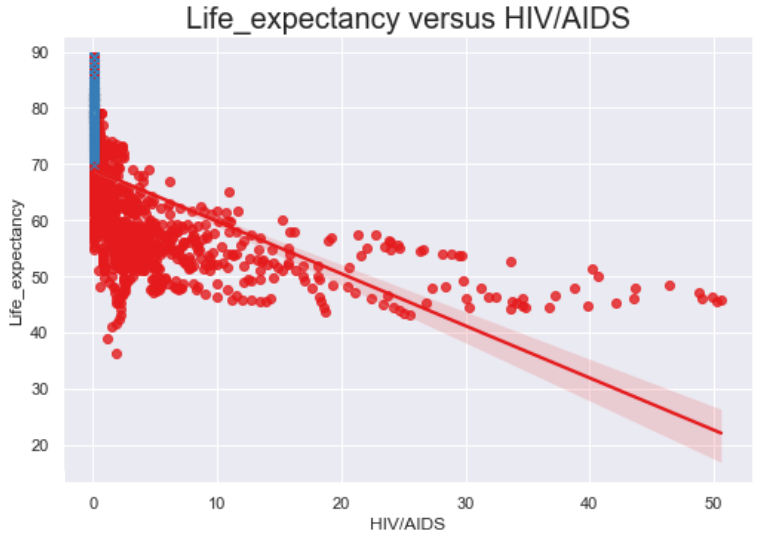




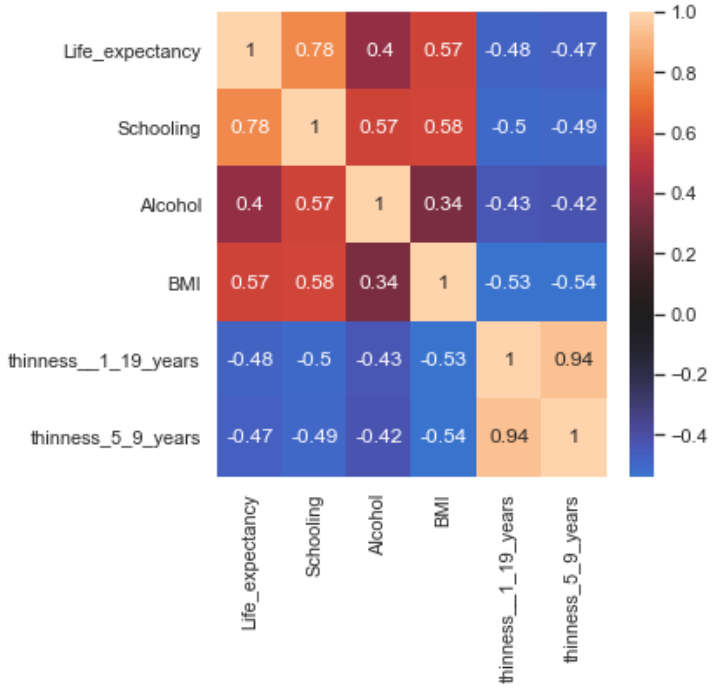


Exploratory Data Analysis (7)

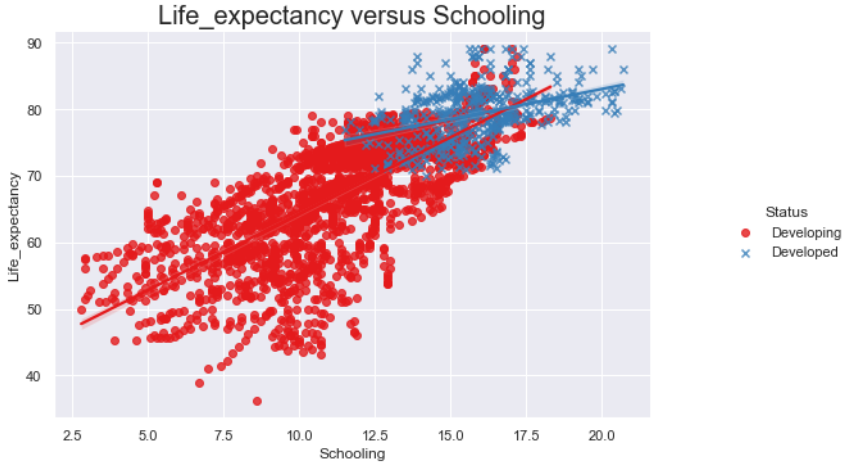
Mortality_factors and Life_expectancy

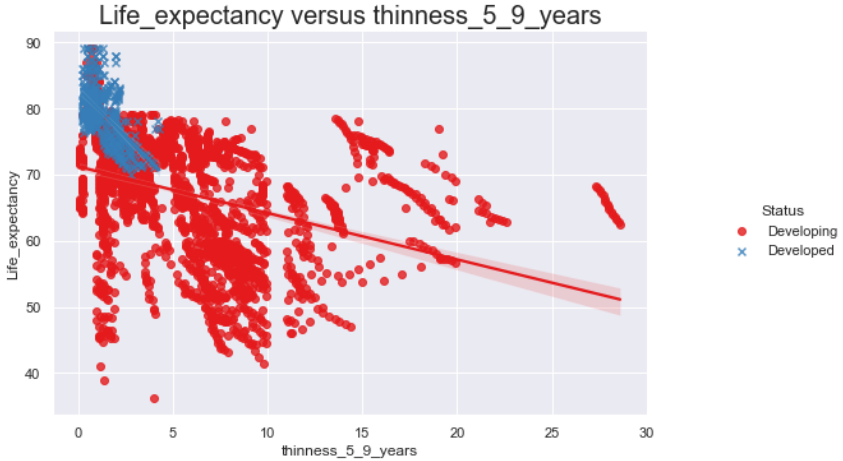


Exploratory Data Analysis (8)



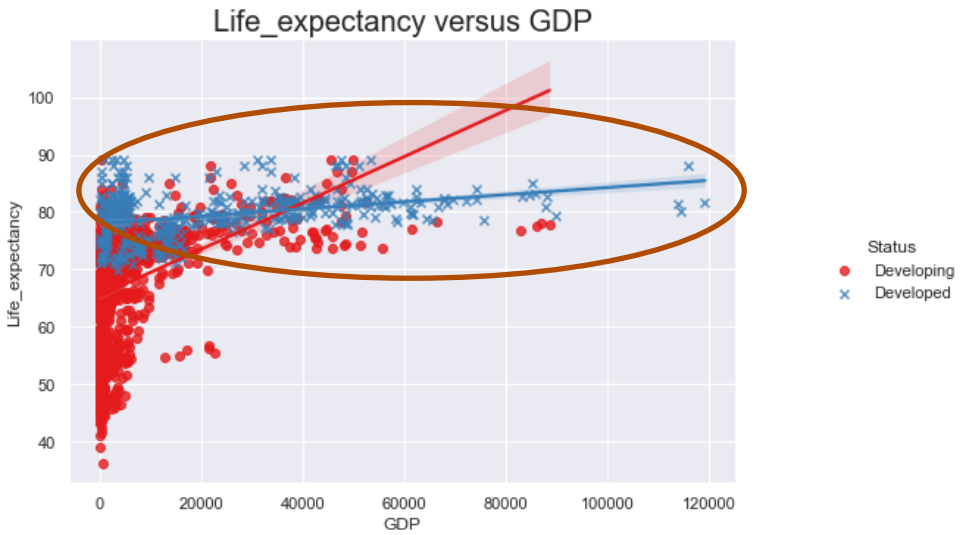
Social Factors and life_expectancy

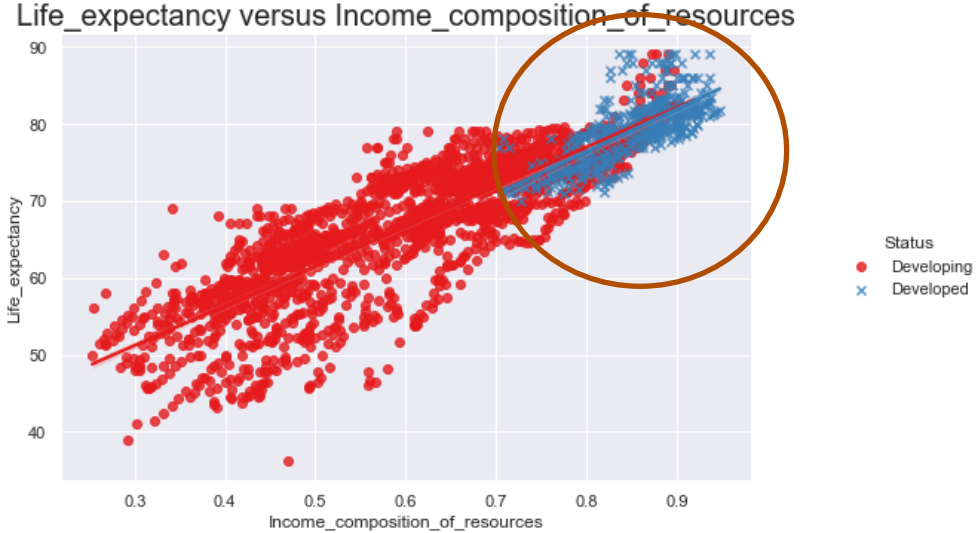




Exploratory Data Analysis (9)

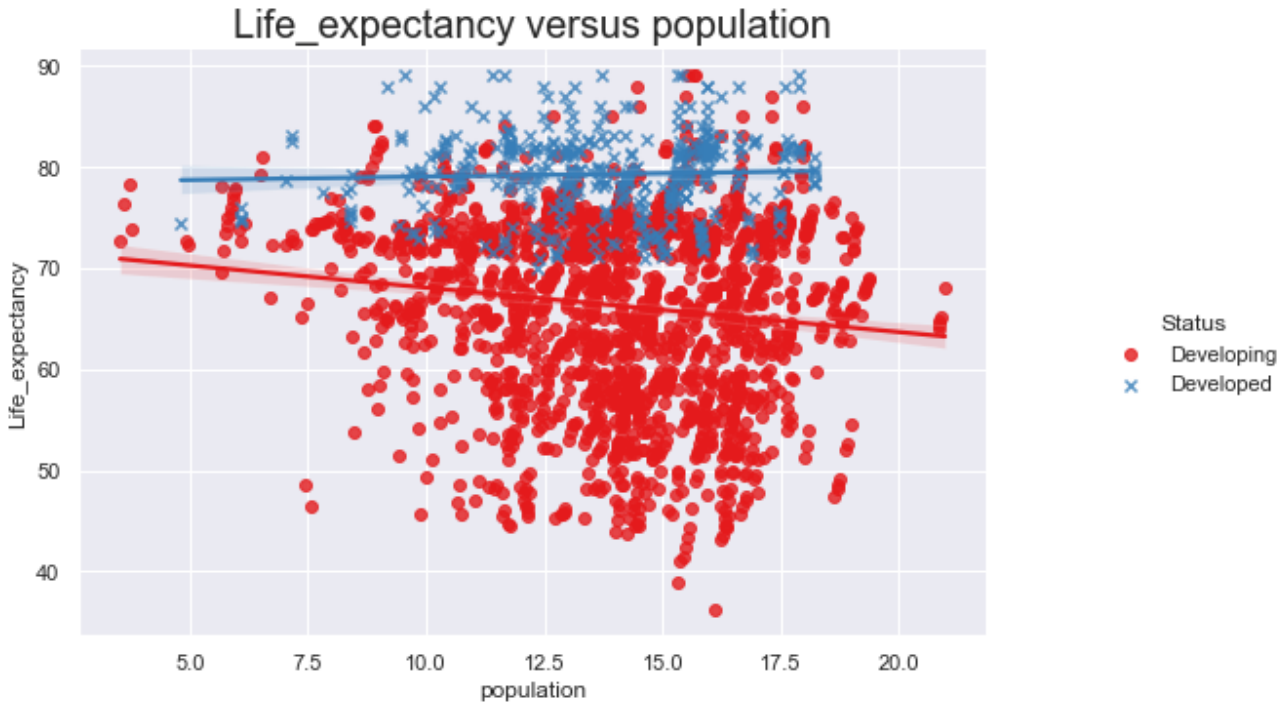
Economical_factors and Life_expectancy





Exploratory Data Analysis (10)

Population and life expectancy



Modeling (1)

- Label Encoder of categerical variable
- Split data into test set and train set (30% , 70%)
- Imputing missing value (Mice imputation)
- Scaling the dataset
- Model specifications
- GridSearchCV , Random Search CV
- Train and evaluate the models, models selection

Modelling (2)

Linear Models

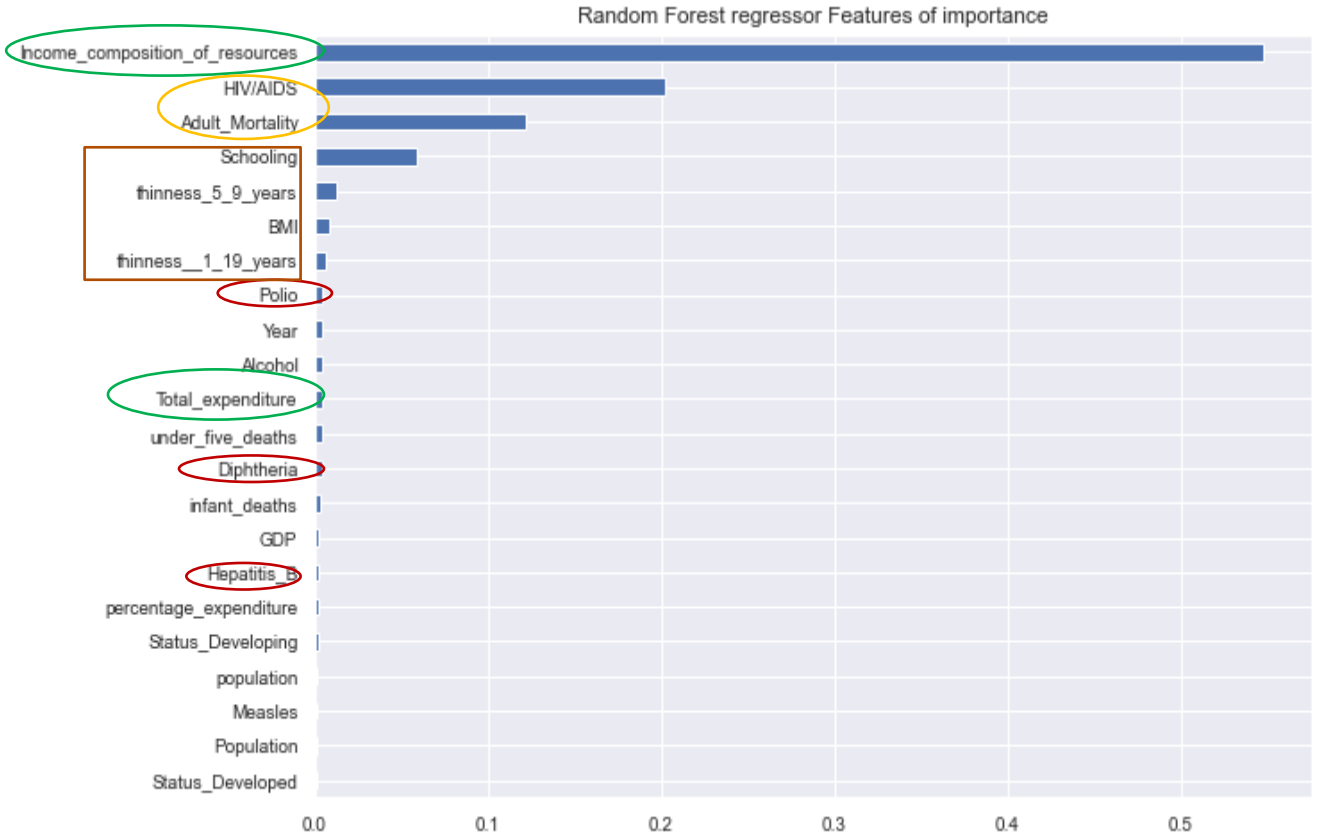
Model	Model definition
linear_reg	Pipeline(steps=[('iterativeimputer', IterativeImputer()), ('standardscaler', StandardScaler()), ('selectkbest', **SelectKBest(k=22**, score_func=<function f_regression at 0x000002477C893670>)), ('linearregression', LinearRegression())])
linear_reg2	Pipeline(steps=[('iterativeimputer', IterativeImputer()), ('standardscaler', StandardScaler()), ('pca', PCA(n_components=22)), ('linearregression', LinearRegression())])
ridge_reg	Pipeline(steps=[('iterativeimputer', IterativeImputer()), ('standardscaler', StandardScaler()), ('ridge', Ridge(alpha=0.5))])
Elastic_net	ElasticNet(alpha=0.01)

Modelling (5)

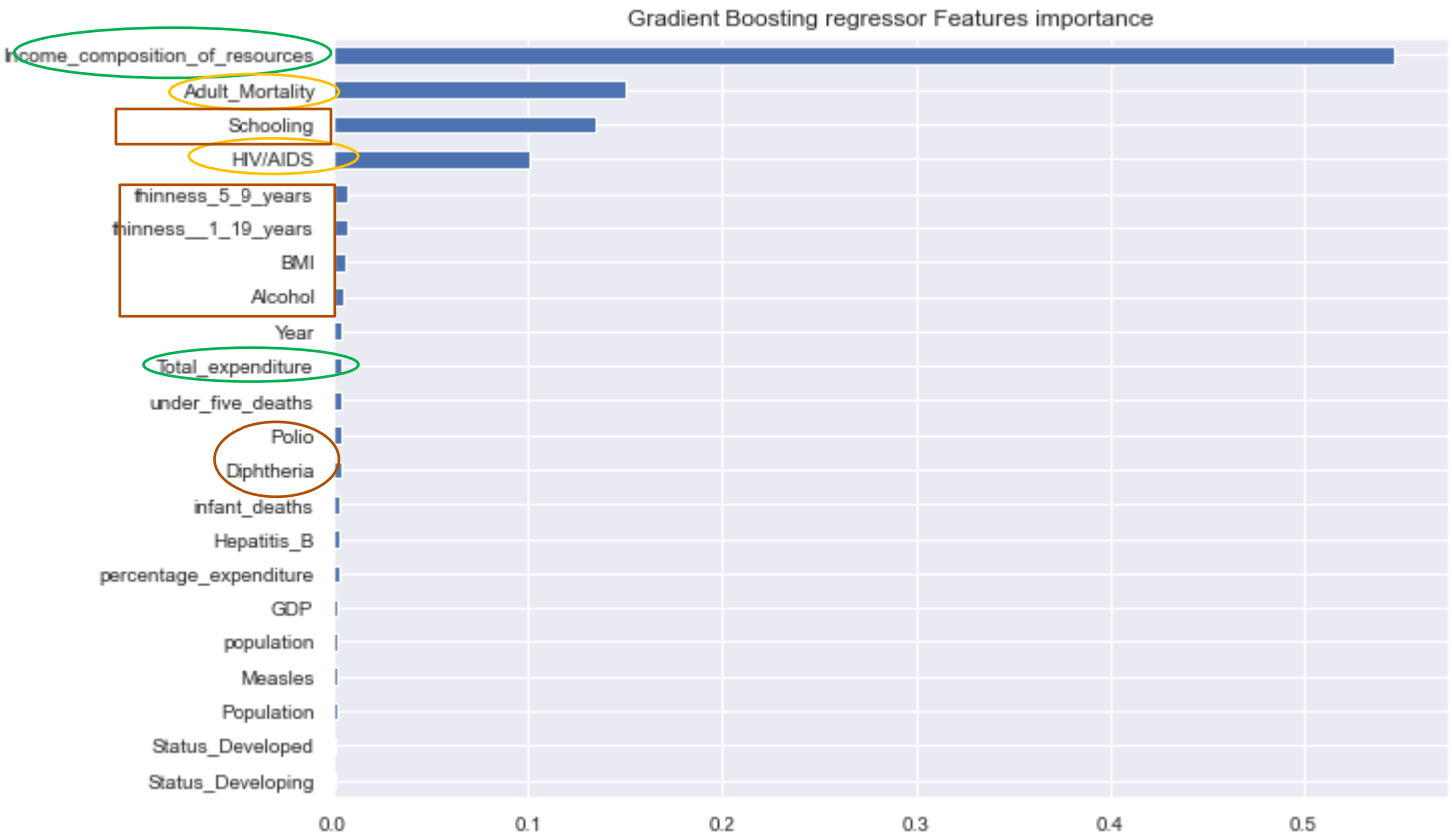
Tree based models

Model	Model definition
decision tree	DecisionTreeRegressor(max_depth=4, max_features=0.2, min_samples_leaf=0.1, random_state=1)
random_forest_reg1	RandomForestRegressor(max_depth=4, max_features=0.4, n_estimators=200, n_jobs=-1, random_state=1)
random_forest_reg2	RandomForestRegressor(max_depth=10, max_features=0.4, n_estimators=200, n_jobs=-1, random_state=1)
random_forest_reg3	RandomForestRegressor(max_depth=7, max_features=0.3, n_jobs=-1, random_state=1)
random_forest_reg4	RandomForestRegressor(max_depth=8, max_features=0.6, n_estimators=200, random_state=1)
gradien_boost_1	GradientBoostingRegressor(n_estimators=150, random_state=1)
gradien_boost_2	GradientBoostingRegressor(learning_rate=0.082, max_depth=10, max_features=0.60, min_samples_leaf=8, min_samples_split=10, n_estimators=118)
gradien_boost_3	GradientBoostingRegressor(learning_rate=0.082, max_depth=10, max_features=0.60, min_samples_leaf=8, min_samples_split=6, n_estimators=150, random_state=1)
voting	VotingRegressor(estimators=[('gb', GradientBoostingRegressor(random_state=47)), ('rf', RandomForestRegressor(random_state=47)), ('lr', LinearRegression())])
XGBRegressor	XGBRegressor(learning_rate=0.04, max_depth=5, n_estimators=200, n_jobs=-1,random_state=0)

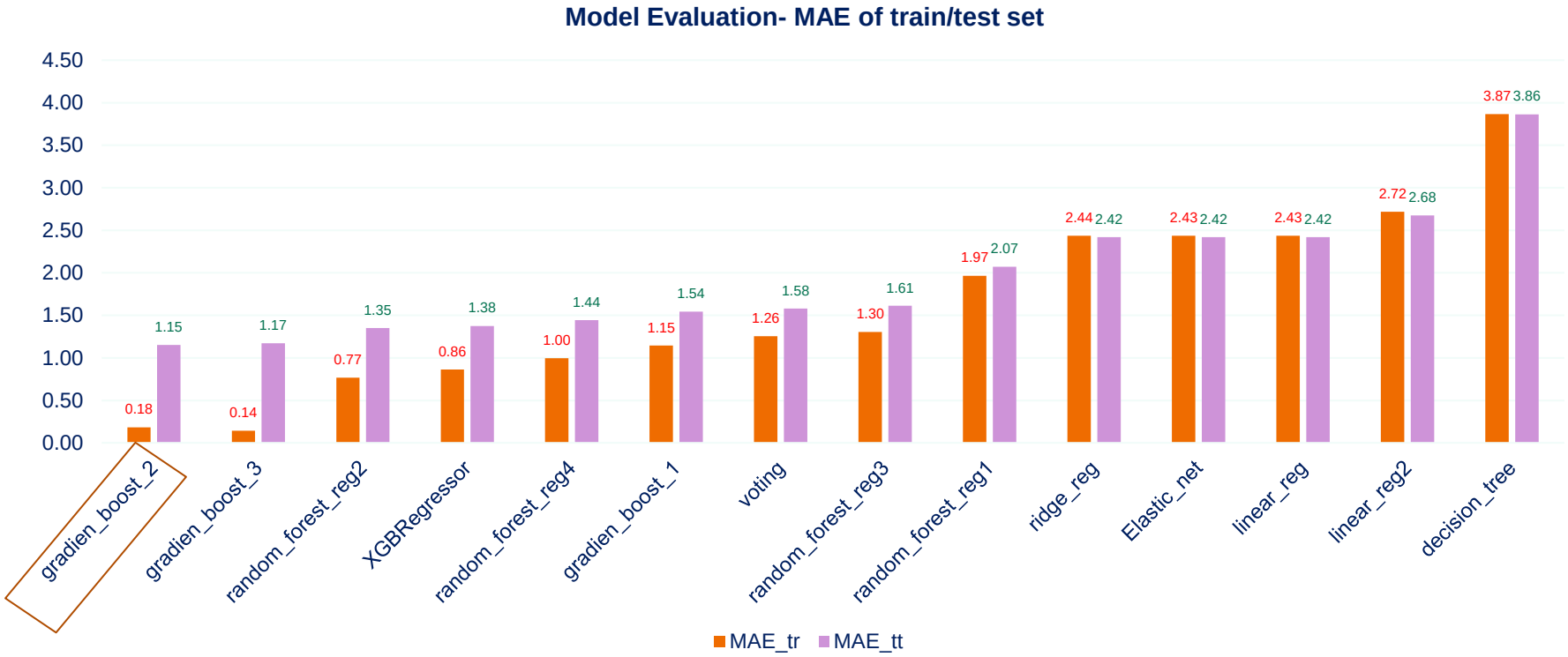
Modelling (6)


Random Forest regressor Features of importance
Income_composition_of_resources
HIV/AIDS
Adult_Mortality
Schooling
thinness_5_9_years
BMI
thinness__1_19_years
Polio
Year
Alcohol
Total_expenditure
under_five_deaths
Diphtheria
infant_deaths
GDP
Hepatitis_B
percentage_expenditure
Status_Developing
population
Measles
Population
Status_Developed
0.0
0.1
0.2
0.3
0.4
0.5

Modelling (7)


Gradient Boosting regressor Features importance
Income_composition_of_resources
Adult_Mortality
Schooling
HIV/AIDS
thinness_5_9_years
thinness__1_19_years
BMI
Alcohol
Year
Total_expenditure
under_five_deaths
Polio
Diphtheria
infant_deaths
Hepatitis_B
percentage_expenditure
GDP
population
Measles
Population
Status_Developed
Status_Developing
0.0
0.1
0.2
0.3
0.4
0.5

Modeling (8)


Model Evaluation- MAE of train/test set
4.50
4.00
3.50
3.00
2.50
2.00
1.50
1.00
0.50
0.00
0.18
1.15
0.14
1.17
0.77
1.35
0.86
1.38
1.00
1.44
1.15
1.54
1.26
1.58
1.30
1.61
1.97
2.07
2.44
2.42
2.43
2.42
2.43
2.42
2.72
2.68
3.87
3.86
gradien_boost_2
gradien_boost_3
random_forest_reg2
XGBRegressor
random_forest_reg4
gradien_boost_1
voting
random_forest_reg3
random_forest_reg1
ridge_reg
Elastic_net
linear_reg
linear_reg2
decision_tree
MAE_tr
MAE_tt

Conclusion (1)

- Life expectancy has increased over years.
- The mean average of the life expectancy in developed countries.
- Life expectancy ratio over one decade 2005-2015.
- Immunization has impacted the improvement of life expectancy and the reduction of infant deaths.
- Feature of importance reveals that immunization features has a very low contribution to the model .

Conclusion (2)

- Economic factors and mortality factors played 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 negatively impact life expectancy.

- (14) regression models have been developed to predict life expectancy.
- the chosen one is Gradient boost with **MAE of 0.18** on train set and **1.15 on the test** set.

Conclusion (3)

To increase life expectancy, the government must improve policies related to the human development index (income and resource composition) and social factors .

Although immunization factors are not the most important, they help reduce the effect of mortality factors on life expectancy.

Thus, promoting immunization will indirectly play a big role in life expectancy.

Thank you