Project 1: Life Expectancy (WHO)

Sadaf Shahbaz

Agnesa Percuku Gerxhaliu

1. Introduction to Data Set

2. Exploratory Data analysis

2.1.Has life expectancy improved over the years ?

2.2.Does population have an impact on long life?

2.3.Are Wealthy countries being more likely to live longer?

2.4. Should countries with lower life expectancy increase their percentage expenditure on Health?

2.5.What is main reason of very low LE in Africa?

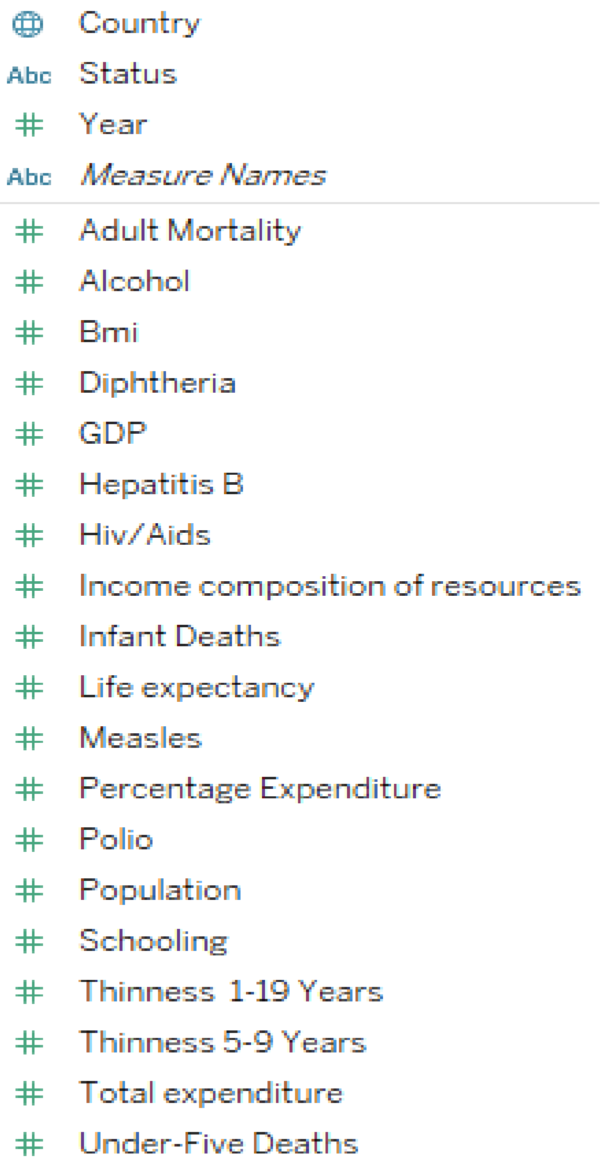
2.6. Do education helps make people improving LE?

3. Predicting life expectancy

4.Logistic Regression to predict status

5. Conclusions and Recommendations

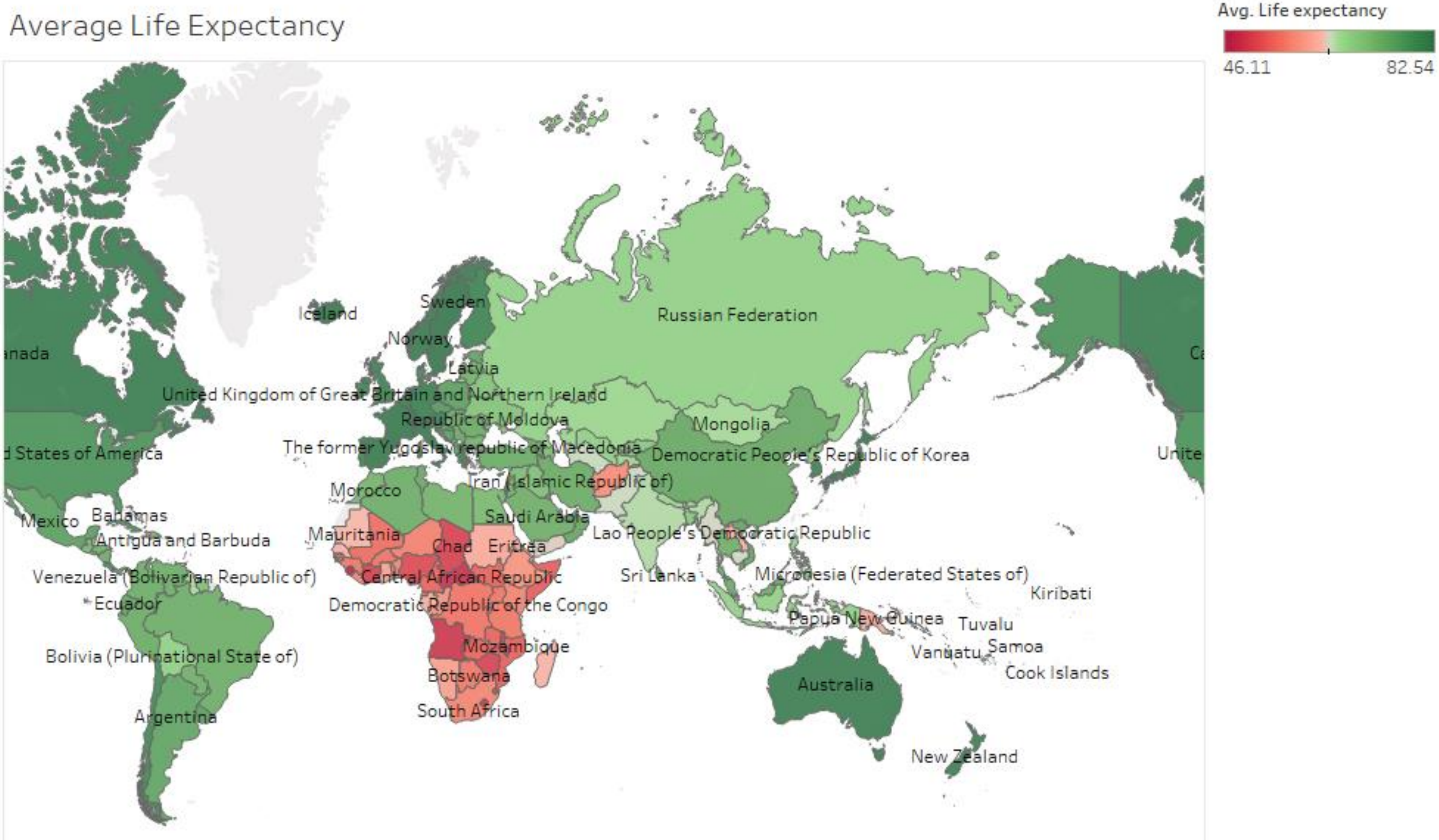
Life Expectancy dataset

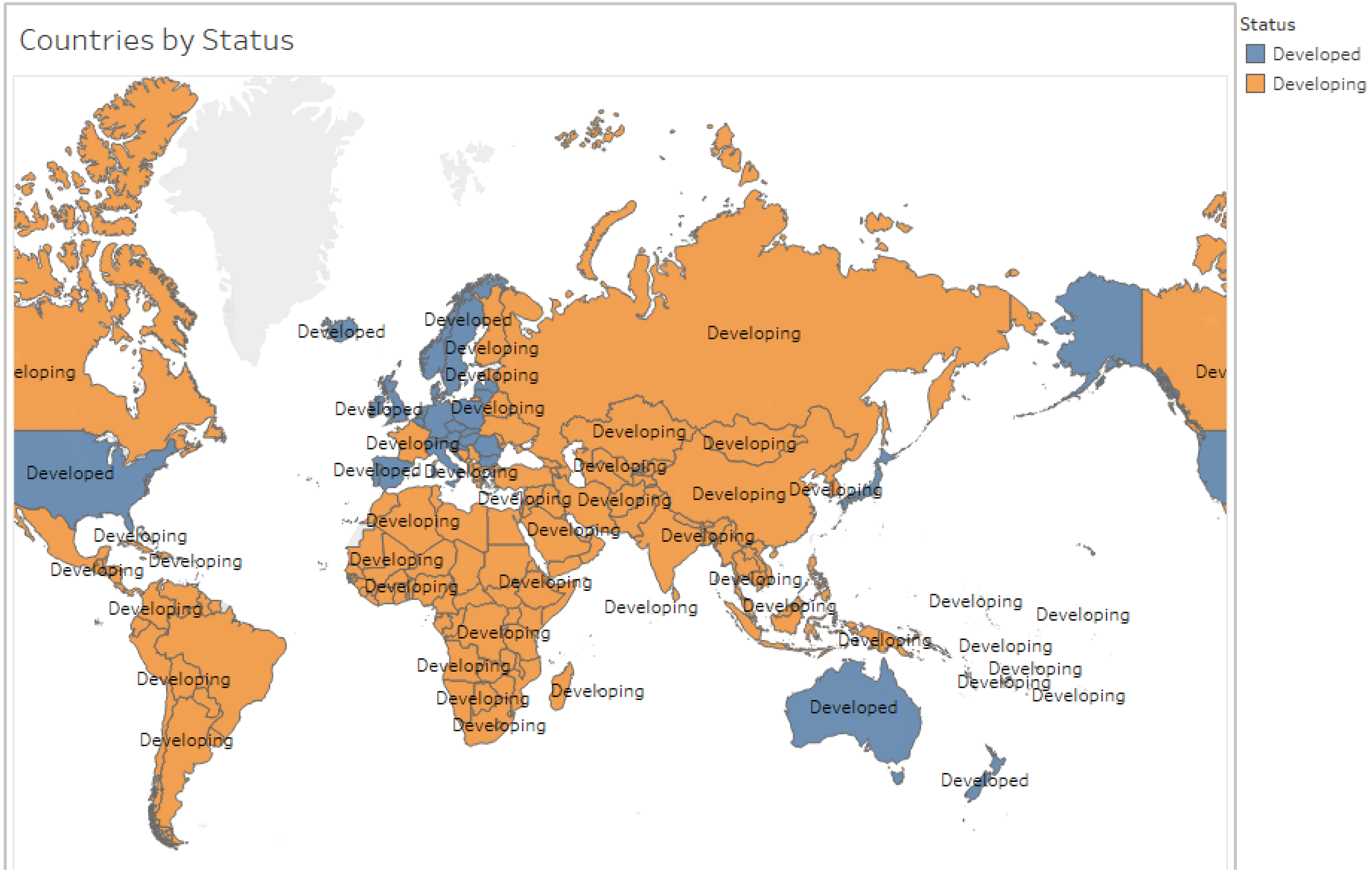


🌐 Country
Abc Status
Year
Abc *Measure Names*

Adult Mortality
Alcohol
Bmi
Diphtheria
GDP
Hepatitis B
Hiv/Aids
Income composition of resources
Infant Deaths
Life expectancy
Measles
Percentage Expenditure
Polio
Population
Schooling
Thinness 1-19 Years
Thinness 5-9 Years
Total expenditure
Under-Five Deaths

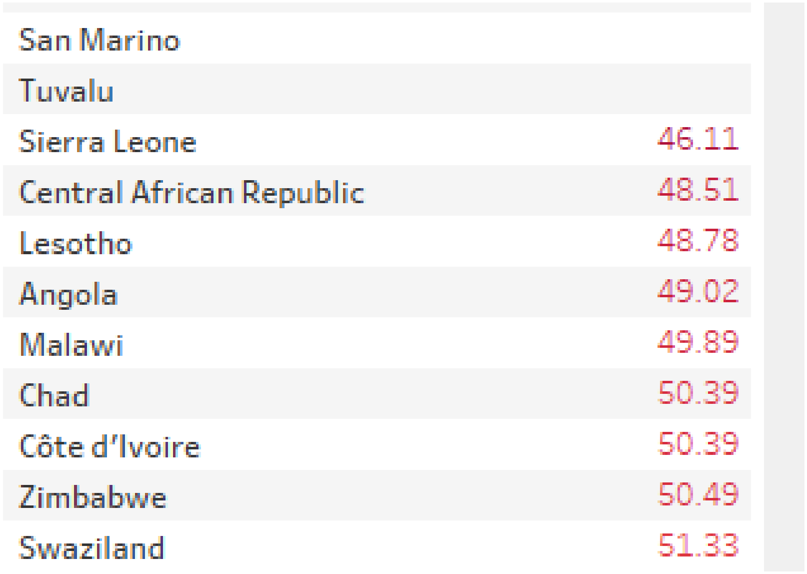
Average Life Expectancy



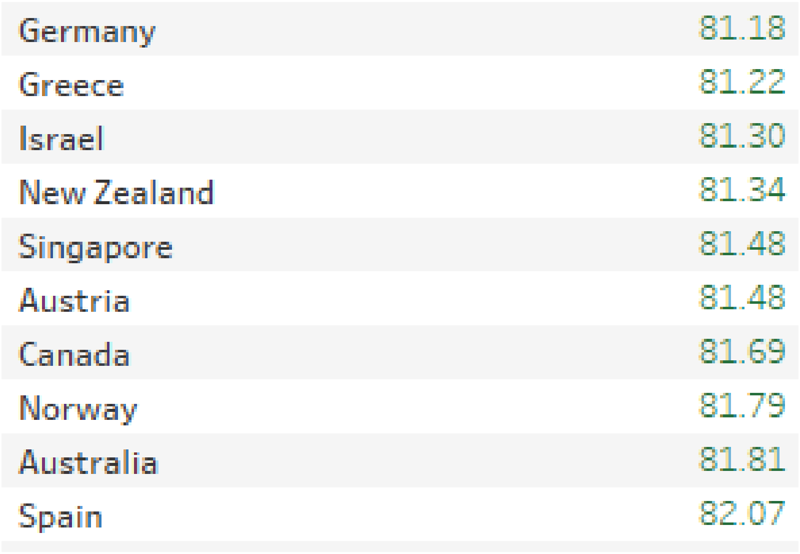


Countries by Status

Countrie with lowest and highest life avg. life expectancies

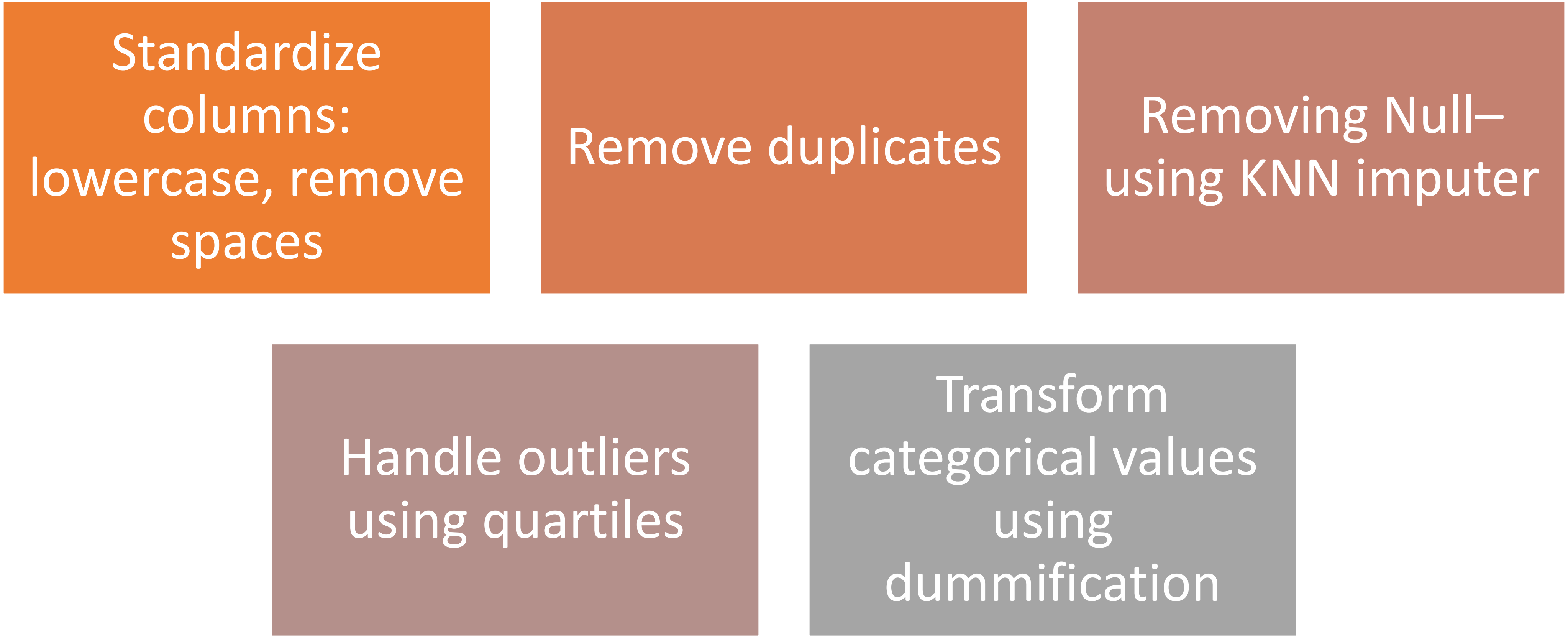


San Marino	
Tuvalu	
Sierra Leone	46.11
Central African Republic	48.51
Lesotho	48.78
Angola	49.02
Malawi	49.89
Chad	50.39
Côte d'Ivoire	50.39
Zimbabwe	50.49
Swaziland	51.33



Germany	81.18
Greece	81.22
Israel	81.30
New Zealand	81.34
Singapore	81.48
Austria	81.48
Canada	81.69
Norway	81.79
Australia	81.81
Spain	82.07

Data Preprocessing



Standardize columns: lowercase, remove spaces

Remove duplicates

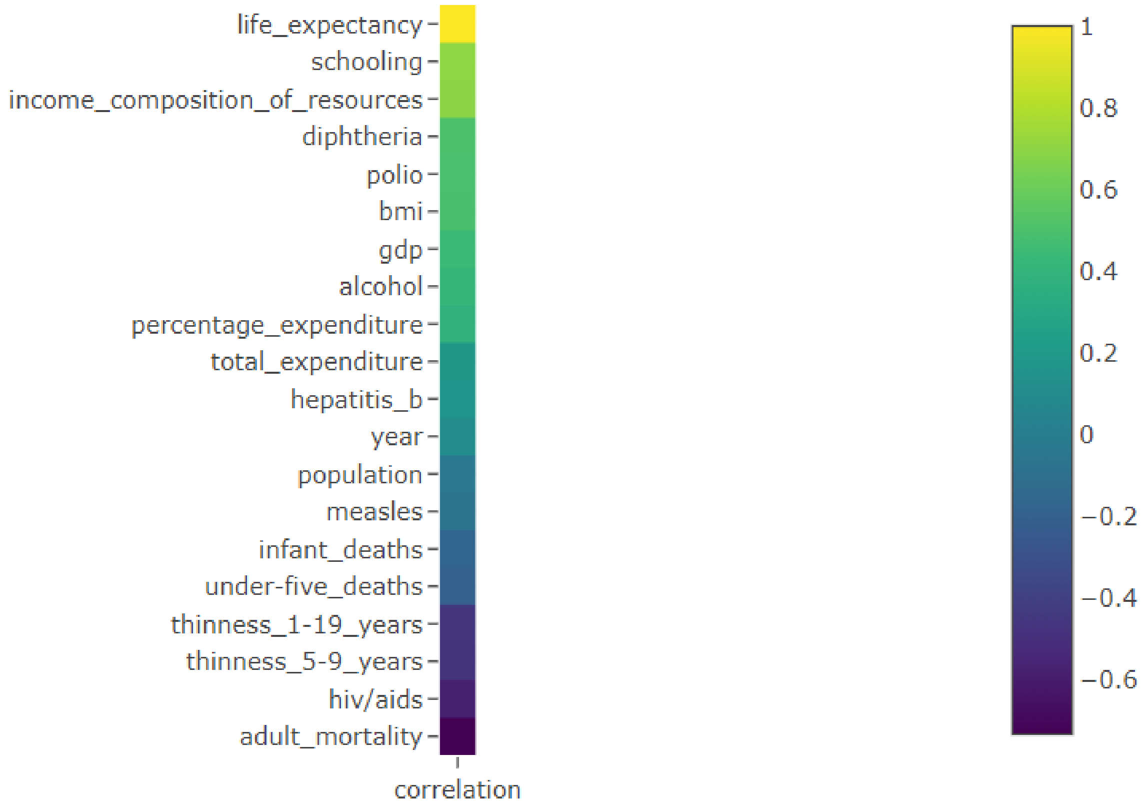
Removing Null– using KNN imputer

Handle outliers using quartiles

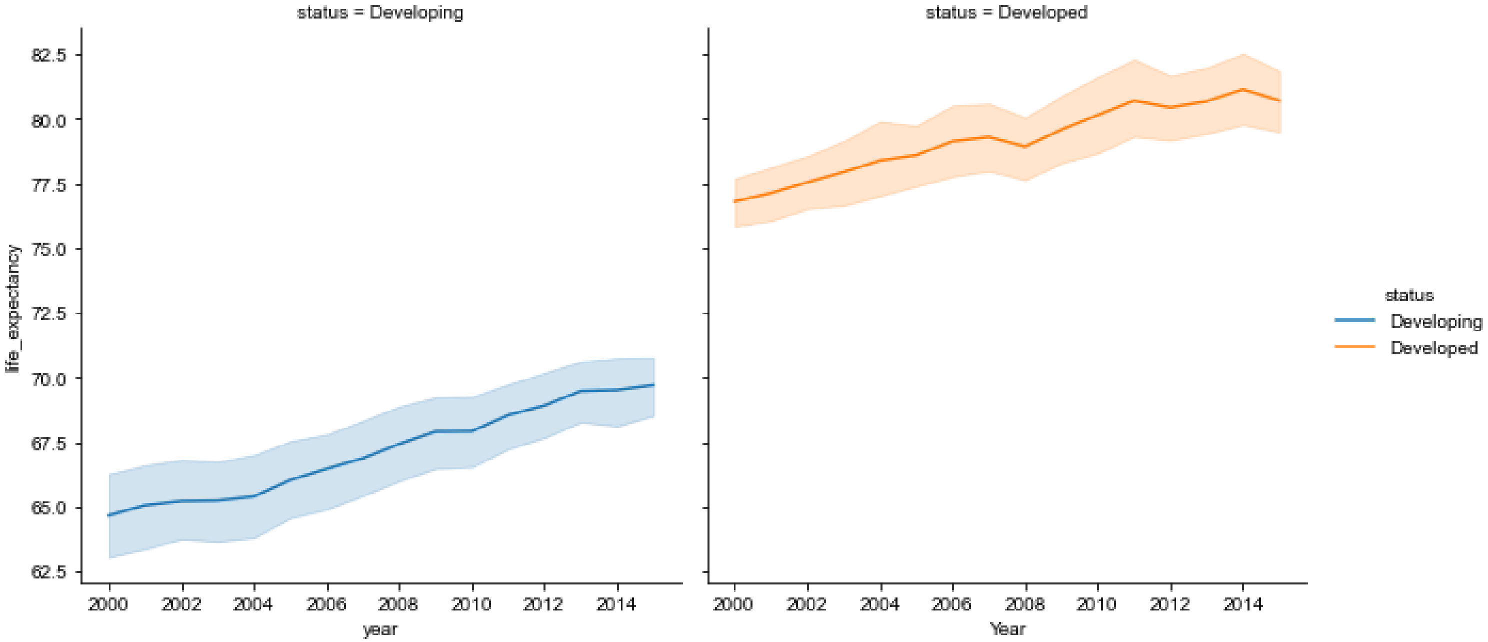
Transform categorical values using dummification

Correlation Coefficient

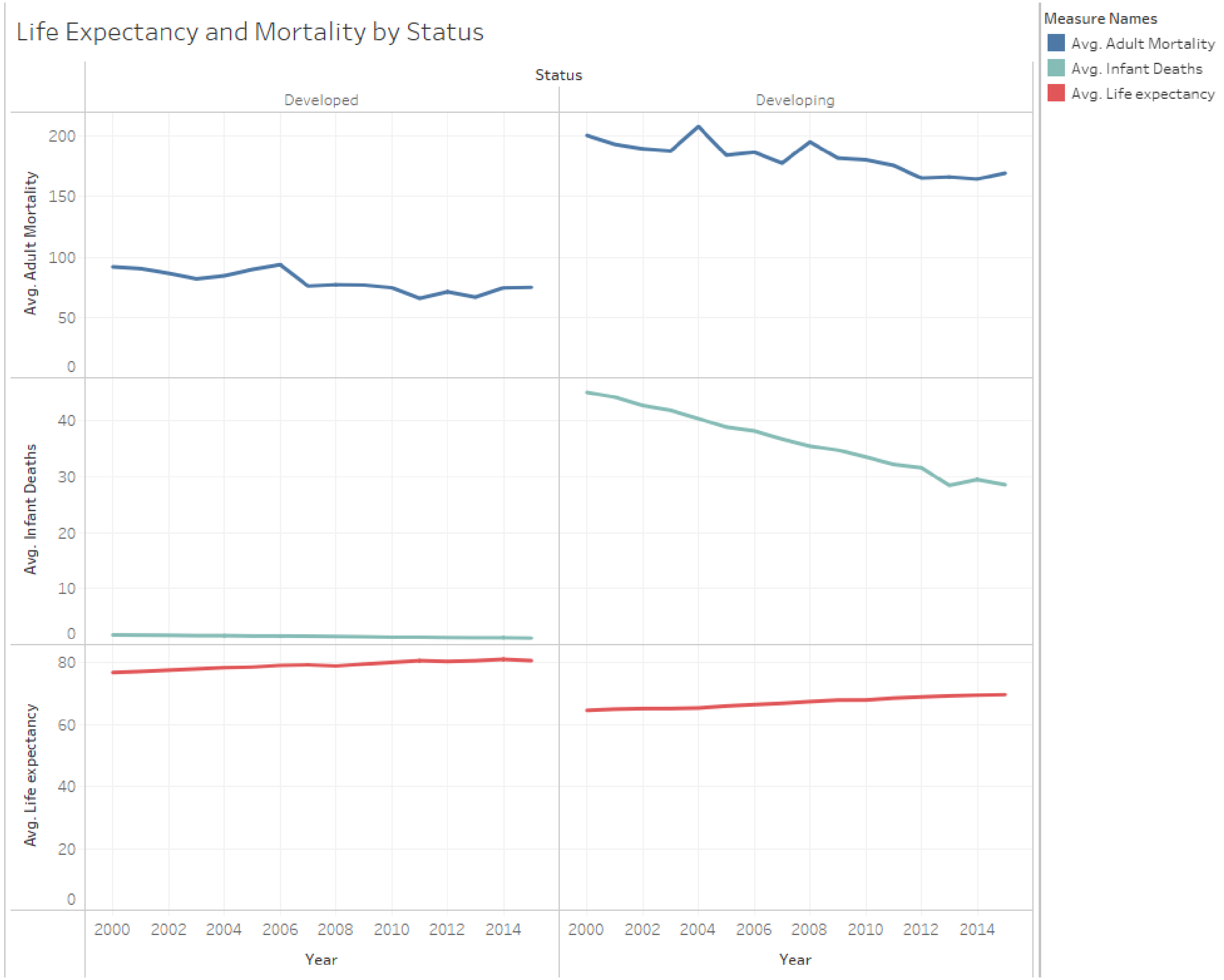
Except Diphtheria, polio, alcohol and hepatitis B are coefficients are explainable.

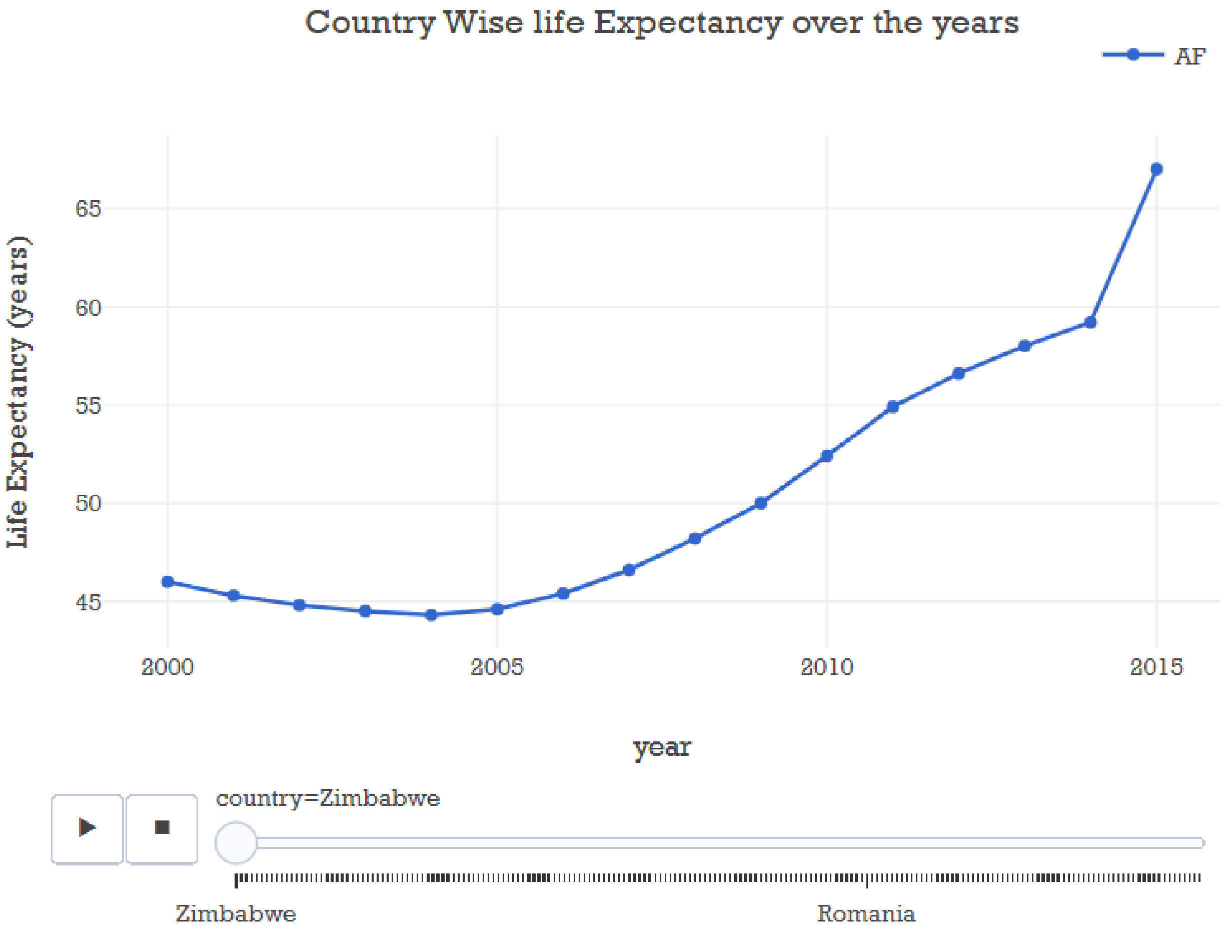


Has life expectancy improved over the years ?

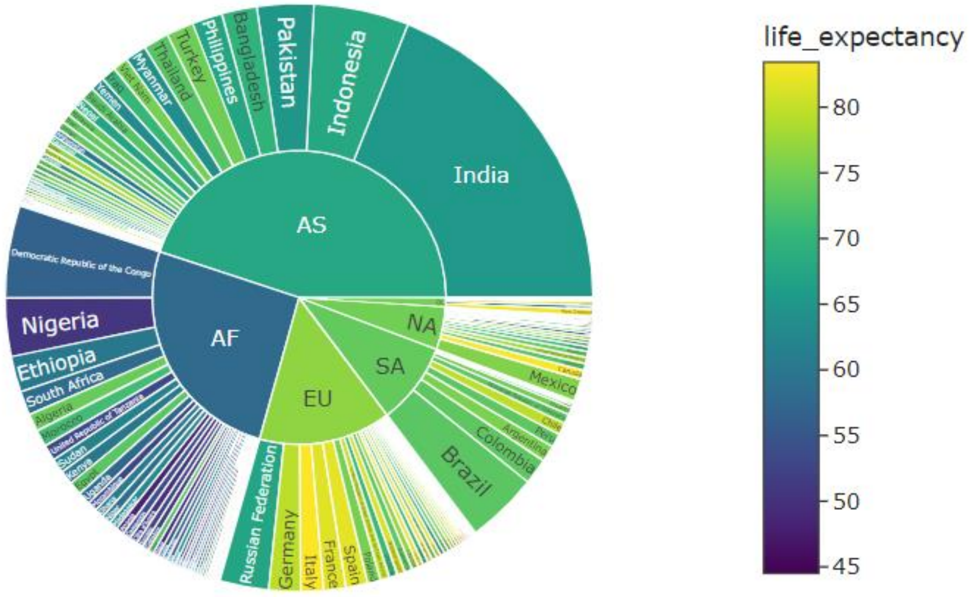


- **Developing countries**- higher adult and infant deaths and lower life expectancy compared to developed countries.
- But we see that with time, the deaths are going down and life expectancy slowly going up.

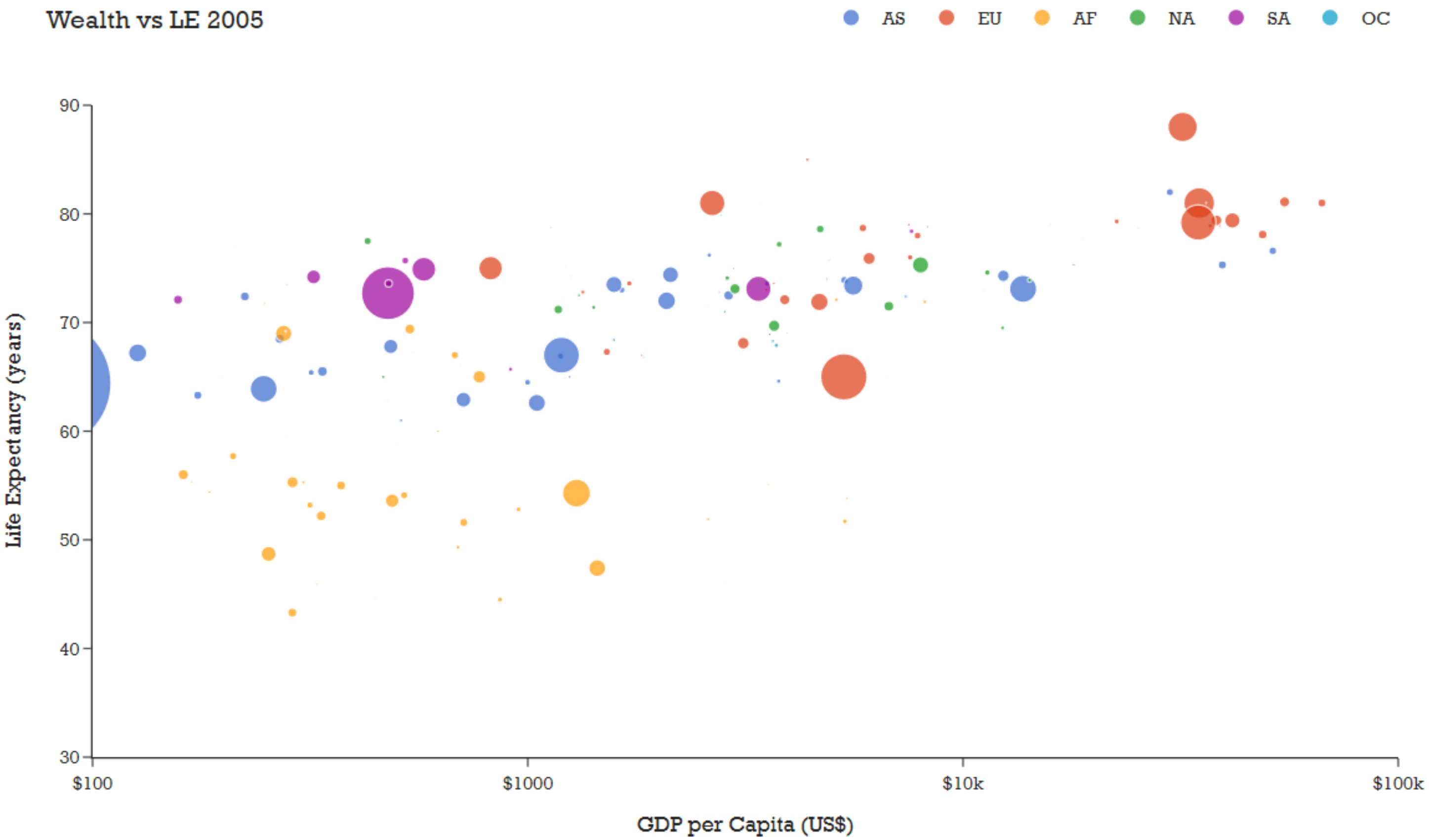



Country Wise life Expectancy over the years
AF
Life Expectancy (years)
65
60
55
50
45
2000
2005
2010
2015
year
country=Zimbabwe
Zimbabwe
Romania

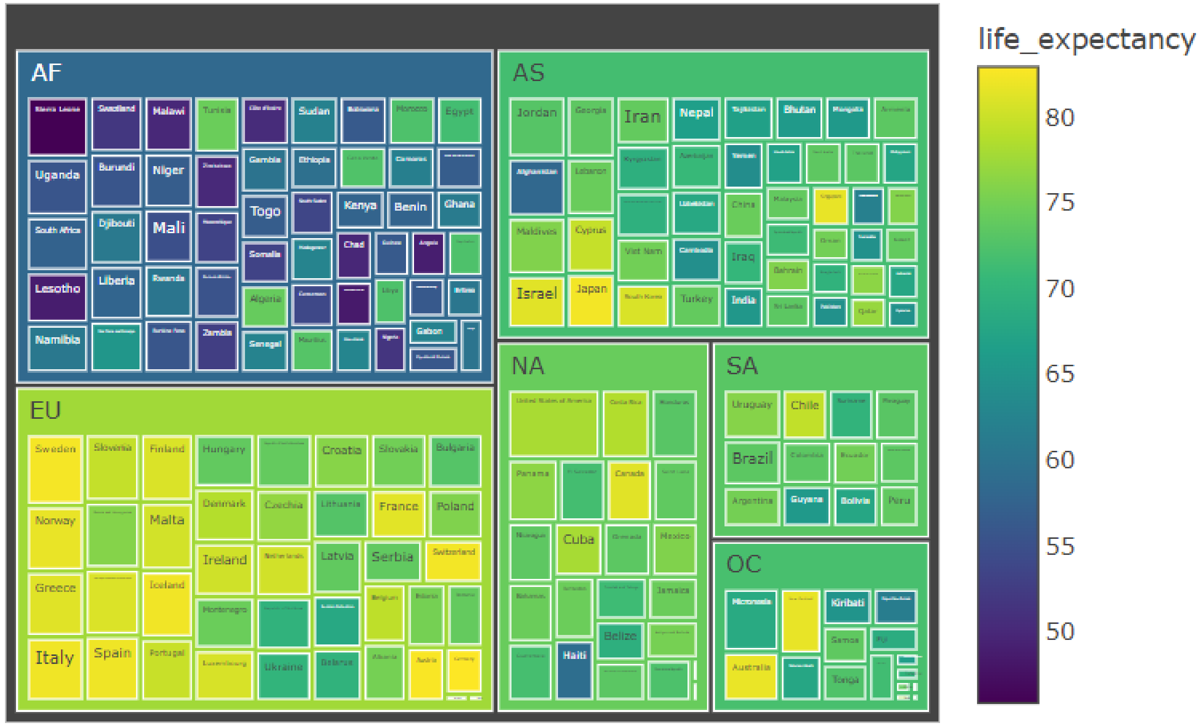
Does population have an impact on long life?



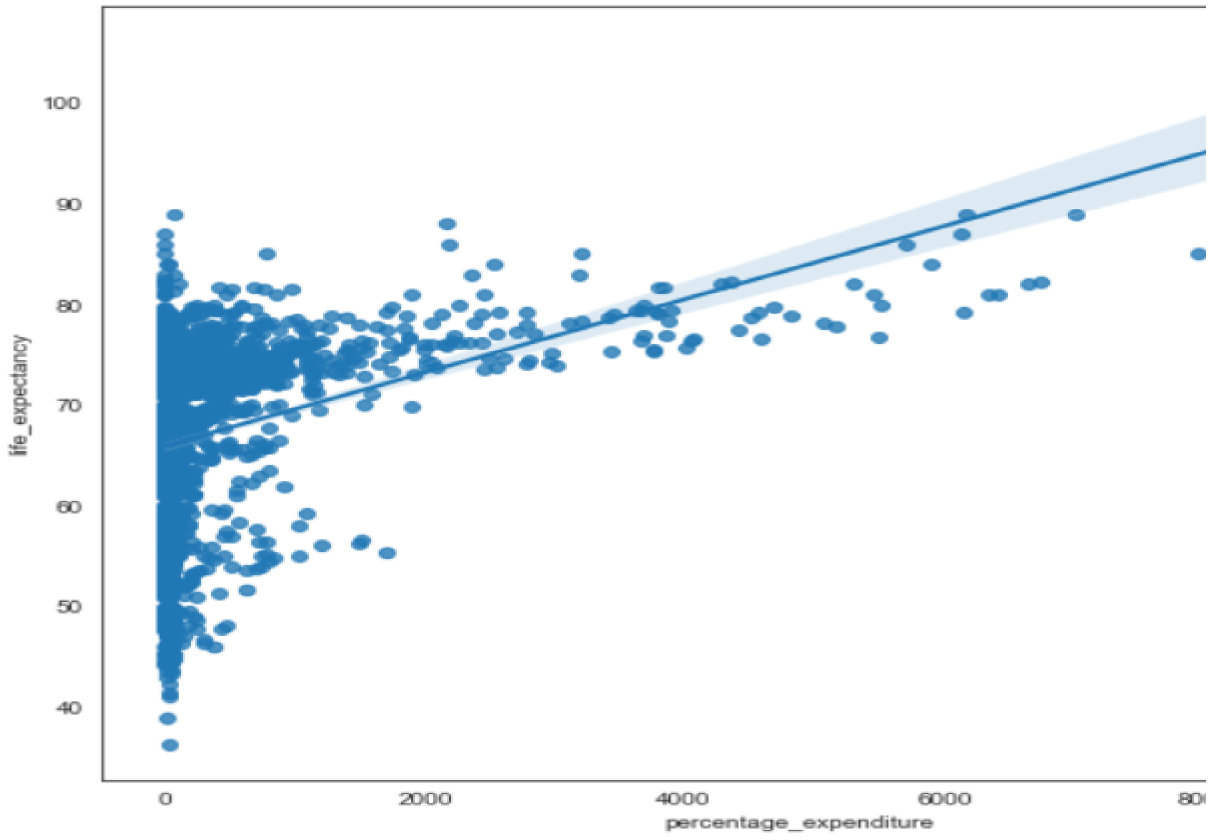
Are Wealthy countries being more likely to live longer?



Should countries with lower life expectancy increase their percentage expenditure on Health?



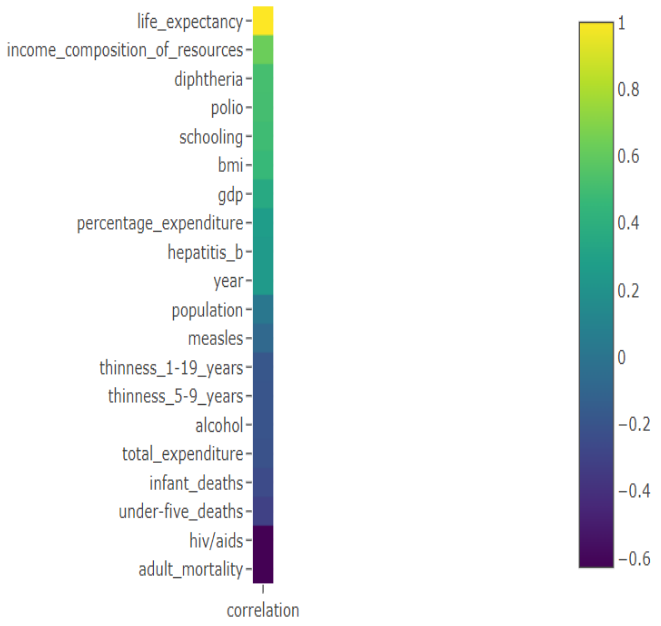
Regression plot (developing countries)



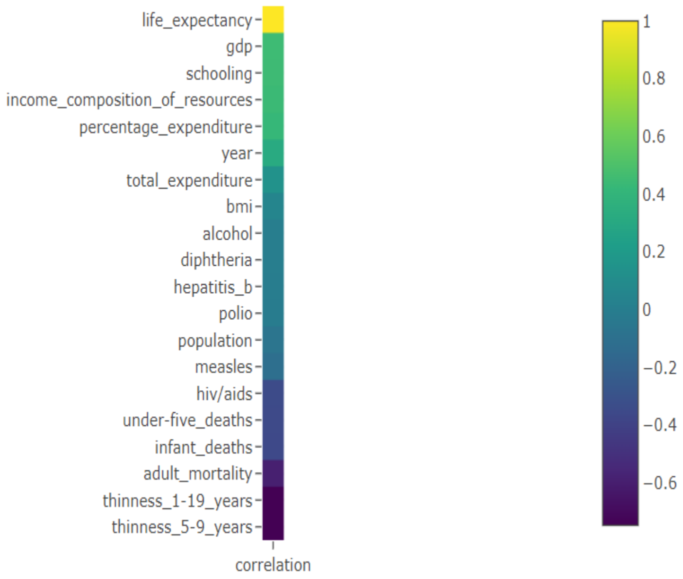
- Narrow confidence interval
- Most of obs. are around fitted line.

What is main reason of very low LE in Africa?

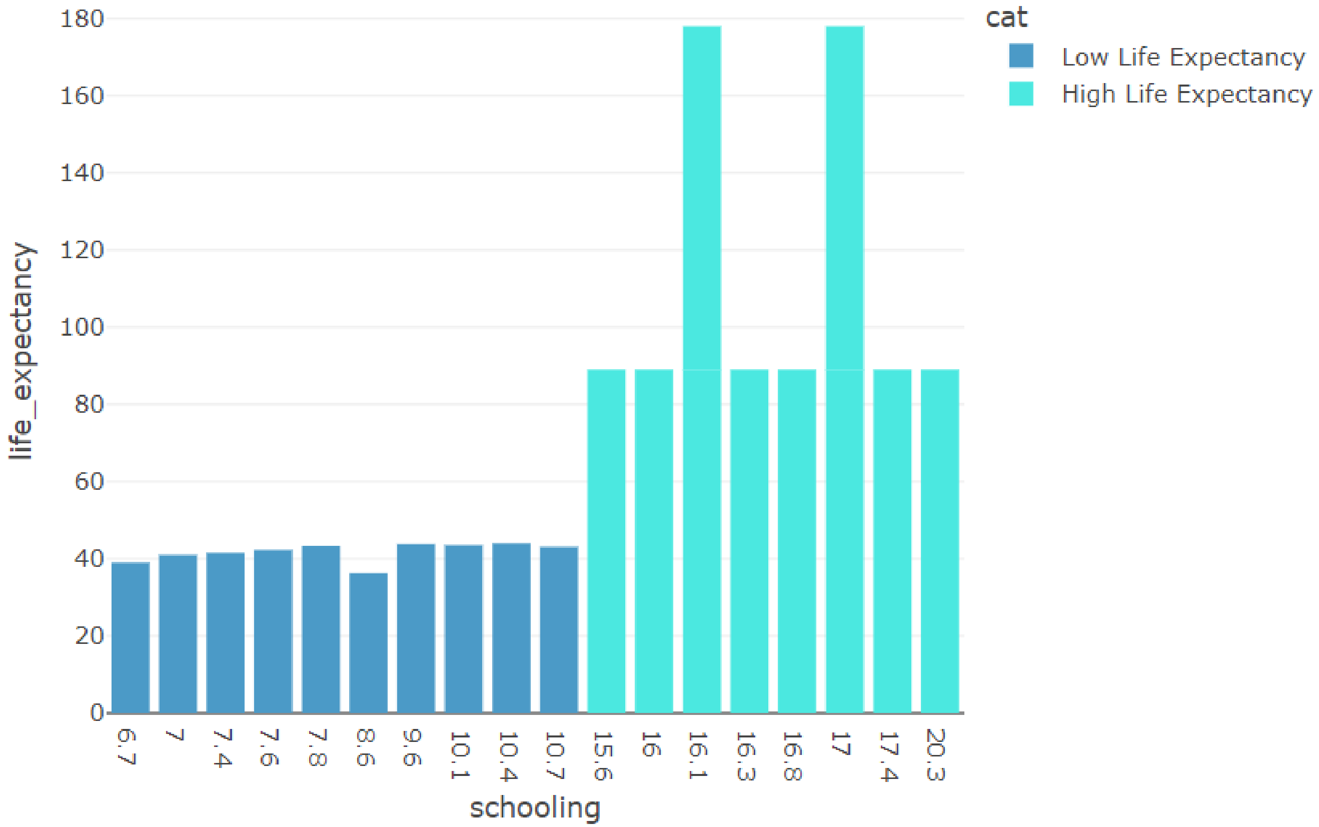
Correlation coefficients (Africa)



Correlation Coefficients (Europe)



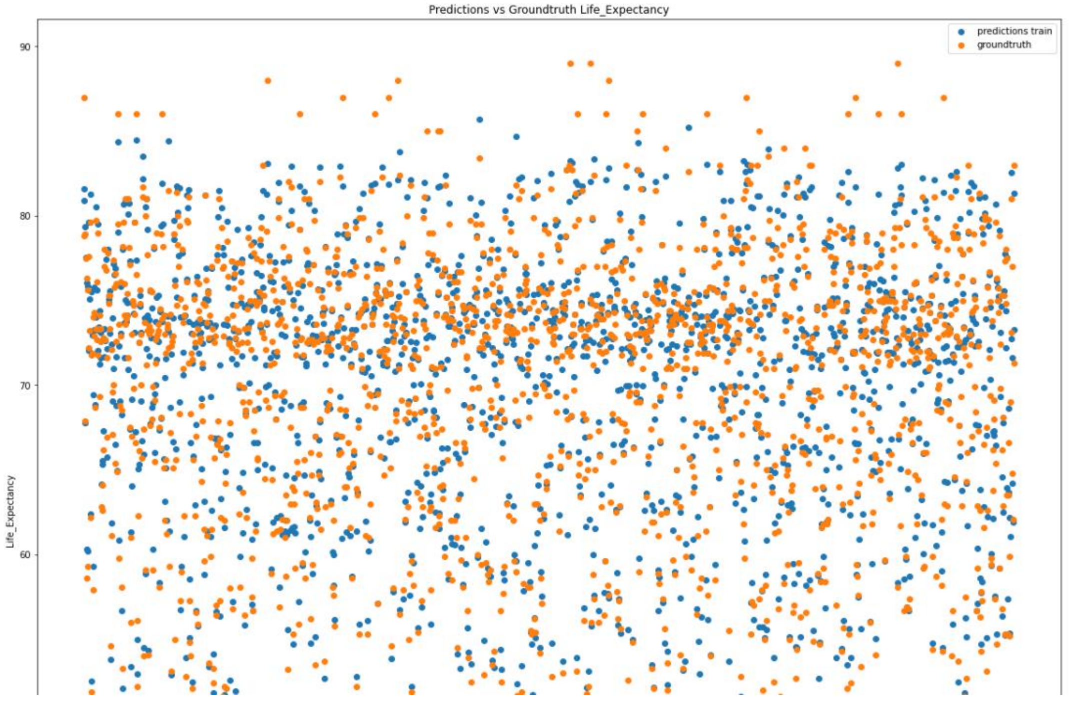
Do education helps make people improving LE?

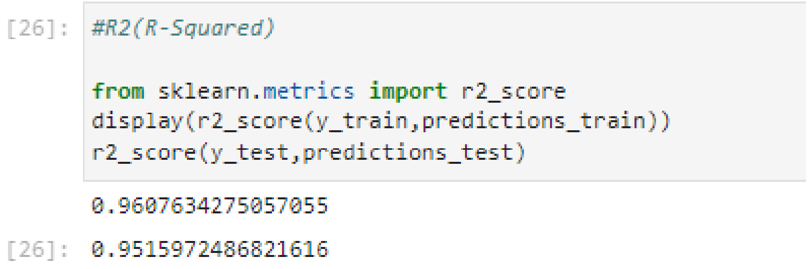


Linear Regression

- The linear regression model seems to be
- a good model for predicting Life Expectancy.

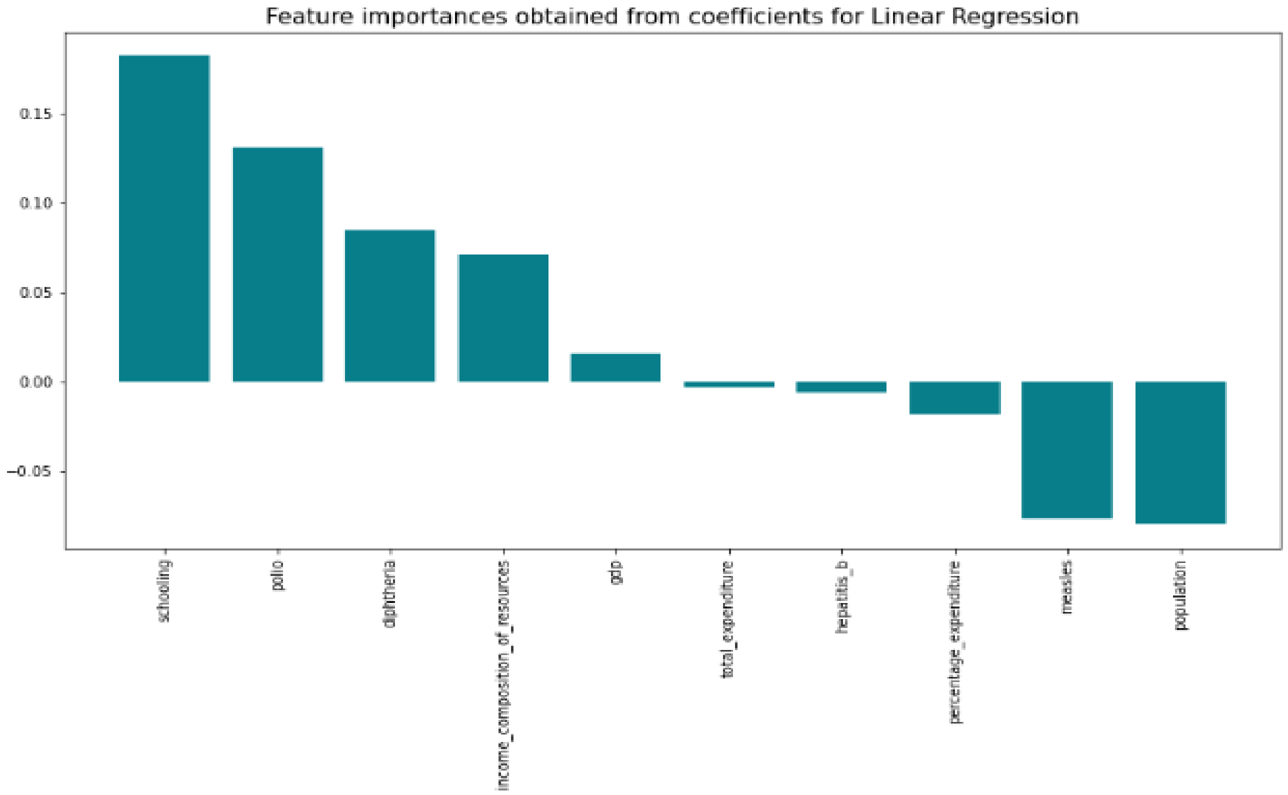
- R2 score in training is 0.96 while in testing 0.95





```
[26]: #R2(R-Squared)

      from sklearn.metrics import r2_score
      display(r2_score(y_train,predictions_train))
      r2_score(y_test,predictions_test)

      0.9607634275057055

[26]: 0.9515972486821616
```

Feature Importance



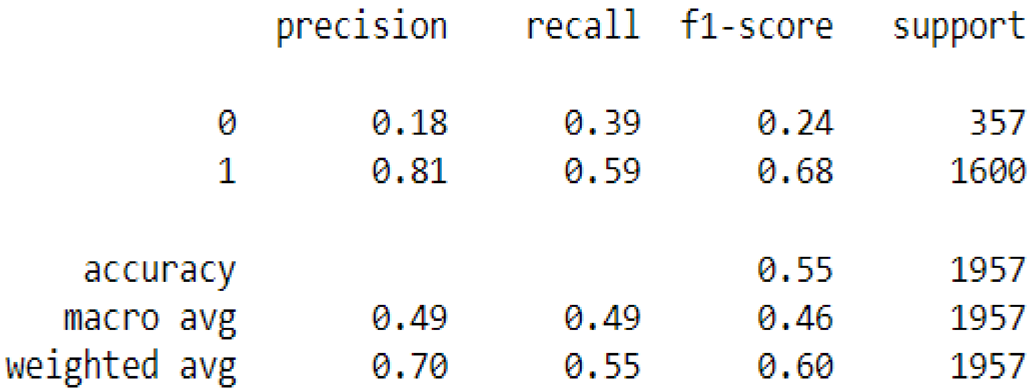
Logistic Regression

Target Variable = Status (Developed = 0, Developing = 1)

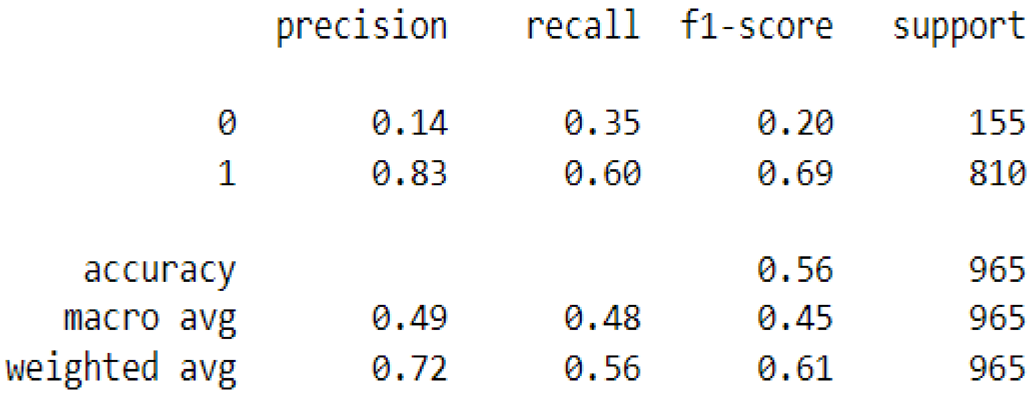
Explanatory variables = life_expectancy adult_mortality, infant_deaths, alcohol, percentage_expenditure, hepatitis_b, diphtheria, hiv/aids, gdp,population, thinness-1-9 years, thinness 5-9 years, income composition of resources, schooling, measles, under_five_deaths, polio, total_expenditure

Classification Report

accuracy on train set:
0.5528870720490546
accuracy on test set:
0.5575129533678757



```
              precision    recall  f1-score   support

           0       0.18      0.39      0.24       357
           1       0.81      0.59      0.68      1600

    accuracy                           0.55      1957
   macro avg       0.49      0.49      0.46      1957
weighted avg       0.70      0.55      0.60      1957
```



```
              precision    recall  f1-score   support

           0       0.14      0.35      0.20       155
           1       0.83      0.60      0.69       810

    accuracy                           0.56       965
   macro avg       0.49      0.48      0.45       965
weighted avg       0.72      0.56      0.61       965
```

Precision – What percent of your predictions were correct?

`Precision = TP/(TP + FP)`

Recall – What percent of the positive cases did you catch?

`Recall = TP/(TP+FN)`

F1 score – What percent of positive predictions were correct?

Problem in Model

Imbalanced Data

No. of observation for developed class = 512

No. of observations for developing class = 2410

Need to balance data

Resampling the data

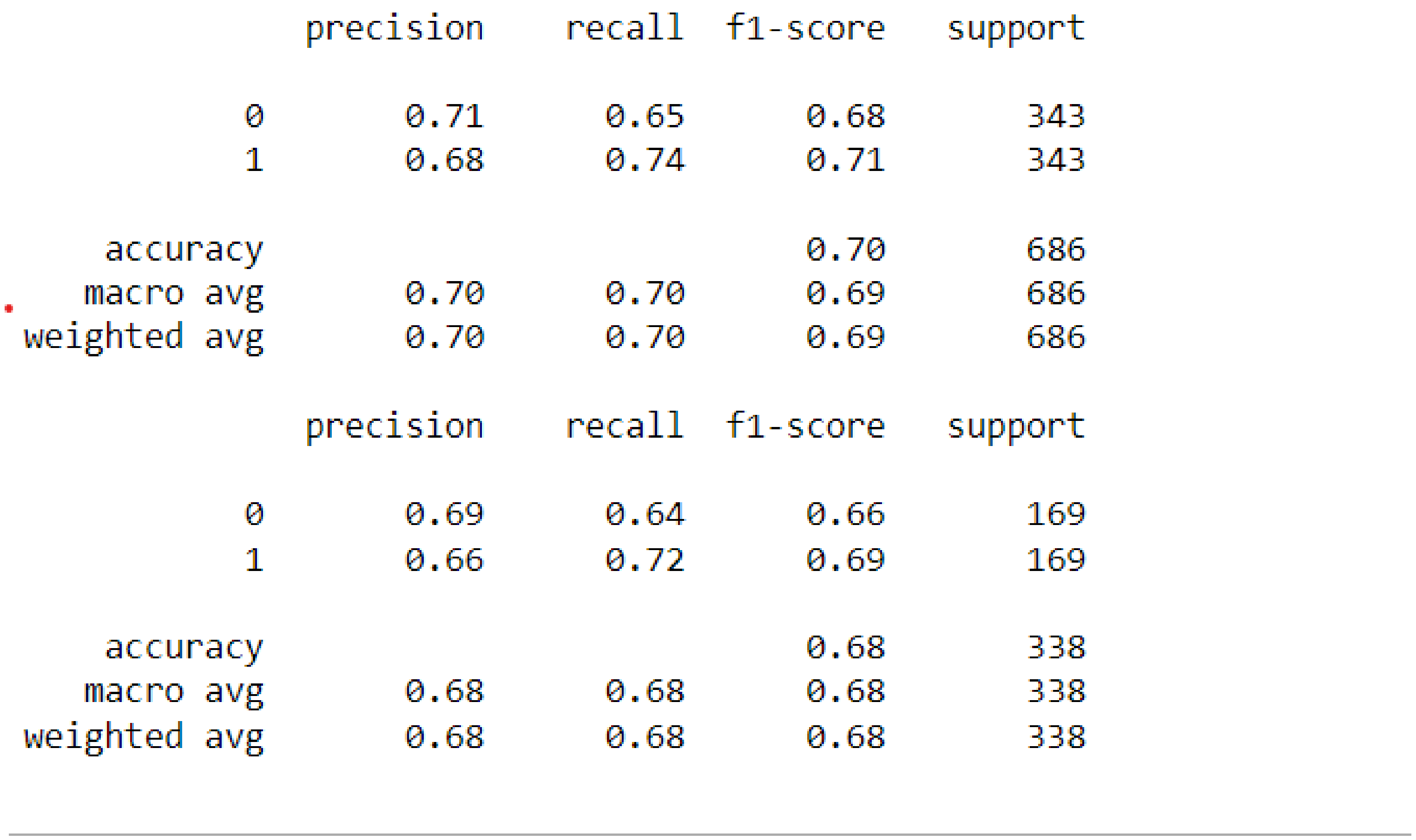
1.Downsampling

2. Upsampling

Classification Report after Down sampling

0= developed = 512

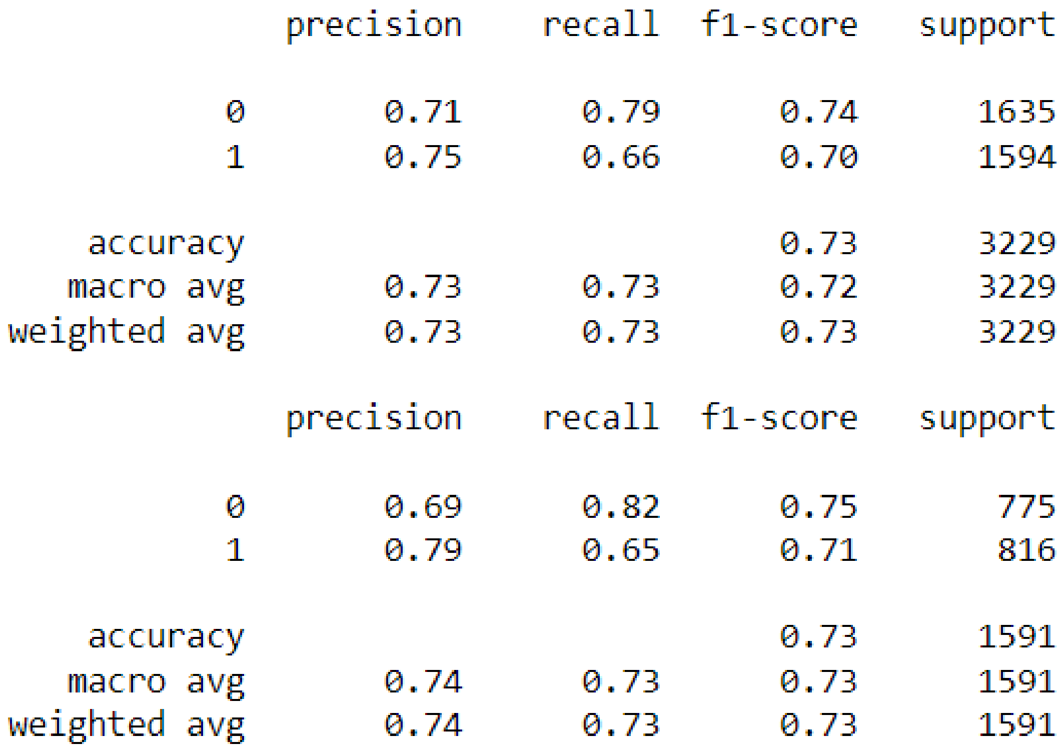
1= developing = 512



```
              precision    recall  f1-score   support

           0       0.71      0.65      0.68       343
           1       0.68      0.74      0.71       343

    accuracy                           0.70       686
   macro avg       0.70      0.70      0.69       686
weighted avg       0.70      0.70      0.69       686
              precision    recall  f1-score   support

           0       0.69      0.64      0.66       169
           1       0.66      0.72      0.69       169

    accuracy                           0.68       338
   macro avg       0.68      0.68      0.68       338
weighted avg       0.68      0.68      0.68       338
```

Classification Report after Up sampling

developed = 2410
developing = 2410



```
              precision    recall  f1-score   support

           0       0.71      0.79      0.74      1635
           1       0.75      0.66      0.70      1594

    accuracy                           0.73      3229
   macro avg       0.73      0.73      0.72      3229
weighted avg       0.73      0.73      0.73      3229
              precision    recall  f1-score   support

           0       0.69      0.82      0.75       775
           1       0.79      0.65      0.71       816

    accuracy                           0.73      1591
   macro avg       0.74      0.73      0.73      1591
weighted avg       0.74      0.73      0.73      1591
```

Feature Importance

- **1. Alcohol**
- **2. Life expectancy**
- **3. Income composition of resources**
- **4. Thinness 1-9 years**
- **5 Adult Mortality**
- **6. gdp**
- **7.bmi**
- **8. percentage expenditures**
- **9. polio**
- **10. total expenditures**
- **11. diphtheria**
- **12. HIV/Aids**
- **13. Hepatitis b**
- **14. under five deaths**
- **15. infant deaths**
- **16. Measles**
- **17. population**

Conclusion

- EDA shows:
- LE is getting better over the years in both developing and developed countries.
- In order to improve LE govts needs to spend more money to increase percentage expenditures on health, education and controlling diseases.
- Analysis also reveals some abnormal positive relationships e.g. +ive relationship of alcohol, dipherthia, etc. needs further investigation.
- Even with abnormalities in dataset model predicted very well life expectancy. Needs further investigation , what are the most important features.
- Model with oversampling technique performed best. Next step could be to select features by using ANOVA and rerun the model.