STATISTICAL ANALYSIS TO UNDERSTAND LIFE EXPECTANCY

Gabriel Ravi[1] **Abdallah Zaher**[1] **Cristina Maria**[1] **Nicolae Radu**[1]

[1]NOVA Information Management School [2]Universidade Nova de Lisboa

Introduction

Life Expectancy is growing up for the last decades but there are some countries that grows slower.
This study is going to analyse 17 sustainable variables for 195 countries to understand which actions and cares can prolong the Life Expectancy in the World.

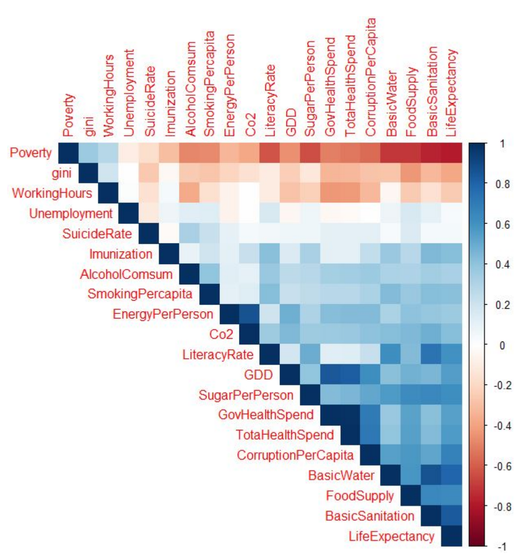
Data Exploration

The data was obtained from Open Access DataWorldBank: https://data.worldbank.org/ and hosted on: https://github.com/GRaviSantos79/StatsForDS.
The MicroWorldBankdata Library is a collection of data sets from the World Bank and other international, regional and national organizations. These data sets contains information about the most important indicators of sustainability of the world.
This analysis consists of a longitudinal collection of 20 indicators of sustainability about 195 countries and response variables, Life Expectancy.

BasicWater	% of Sugar/person
Co2	co2 emissions
GINI	Dist. of whealth
BasicSanitation	% Sanitation Available
Energy	% of Energy/person
Sugar	% of Water/person
GDD	Growing Degree Day
Smoking	% pop that smokes
Imunization	% pop that Immunized
Suicide	% of Suicide
Working hours	Working-Hours/person
Literacy	Literacy/person
Corruption	% of corruption
GovHealthy	Gov Spend with healthy
FoodSupply	% of Food/person
Unemployment	% of Unemployment
TotalHealthy	Pop Spend with healthy
Alcohol	% of Alcohol/person

]

Here you can check and analyse the descriptive statistics of each variables. Almost every variable contain a different scale, so a standardization (log transformation) is necessary for the following analysis.



Basic sanitation, food supply, basic water, corruption and GINI (negative correlation) are the index that has more correlation with life expectancy.

Methodology

Converting LifeExpectancy to a logarithm scale because there are non-linear relationships in the data set, the chances of producing errors may also be skewed negatively.
There are 2% of missing data, if we delete the countries that has NaN values, the data set will lose its quality. The missing values were filled by the Median because it's the technique that don't interfer on the results

```
# Analysing and transforming the main variables
summary(data1$LifeExpectancy)
y = data1$LifeExpectancy
ly = log(y)
# Checking and treating the NA values for the training set
apply(Z, 2, function(X) sum(is.na(X)) )
X = ifelse(is.na(X), median(X, na.rm = TRUE, X))
```

And it is important to check if any predictors are correlated with each other to avoid Multicolinearity and guarantee that the model will not suffer of instability.

```
# MultiColinearity for the numerical variables
COR <- cor(data1)
Multt = findCorrelation(COR, cutoff = 0.75)
data1 = data1[,-Multt]
```

Implementation

After the guarantee that there is a good training set and applying Shapiro to check the normality, the Stepwise Regressions techniques (Backward, Foward and both) will be applied. After that, it will be applied on the final variables a Principal Components analysis to give the most important variables that affects Life Expectancy.

	R	**Nº Vars**
Backward	0.82	4
Foward	0.81	4
Stepwise	0.82	3

After the analysis, considering 95% of confidence, the main variables are Poverty, BasicSanitation and Basic Water. And the final model is:

LifeExpec = 4.2 - 0.03 Poverty + 0.04 Sanitation +0.03BasicWater

And after de PCA, this study will consider two components mainly composed by the variables described above and those componentes explain 78.3% of the Total Variance.

Conclusions

After the analysis and the PCA, we can say that the most important variables to increase Life Expectancy are Poverty, BasicSanitation and Basic Water; in other words, these variables contains the bigger influence when we want to increase Life Expectancy.
For example, in the final mode, when a government increase 3% on the BasicWater for the population, the Life Expectancy will increase 1%.