Data-Driven Insights on Life Expectancy:

Exploratory Data Analysis and Predictive Modeling

2000-2015

Problem Statement:

The analysis should aim to answer questions such as:

- Are there anomalies in the dataset that need to be addressed?
- Are there any significant differences in life expectancy across countries or regions?
- Are there any trends or patterns in life expectancy over time?
- What are the factors influencing life expectancy?
- Can a predictive model (Random Forest) be developed to accurately predict life expectancy?

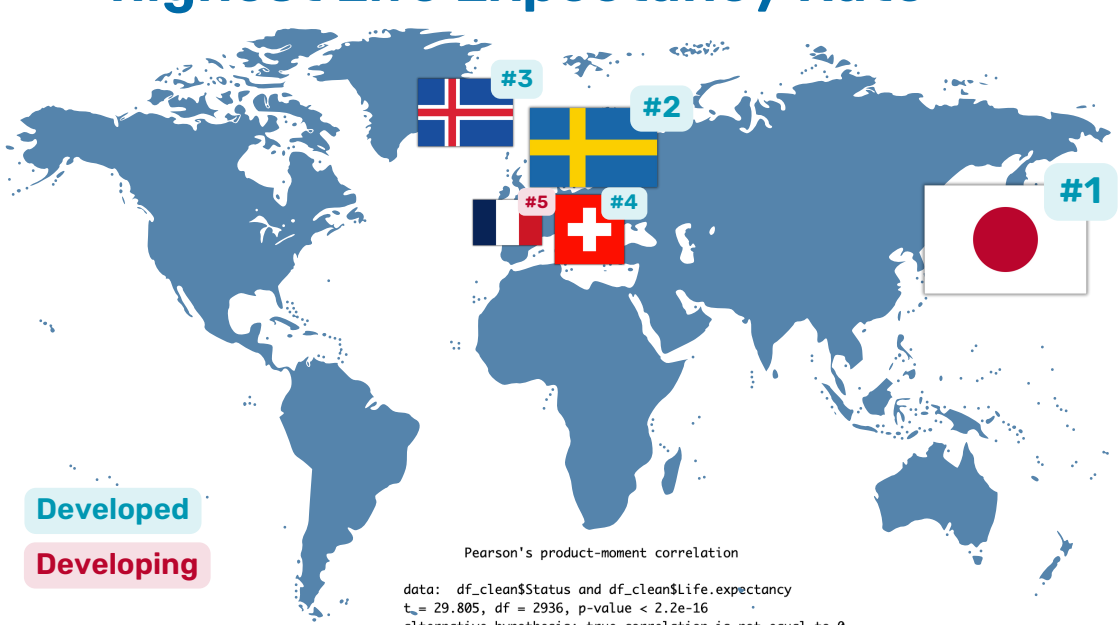
Objectives:

The objective of this task is to perform EDA on the Life Expectancy dataset that can provide meaningful insights and visualizations so that we can create a predictive model to predict Life Expectancy.

Data Descriptions:

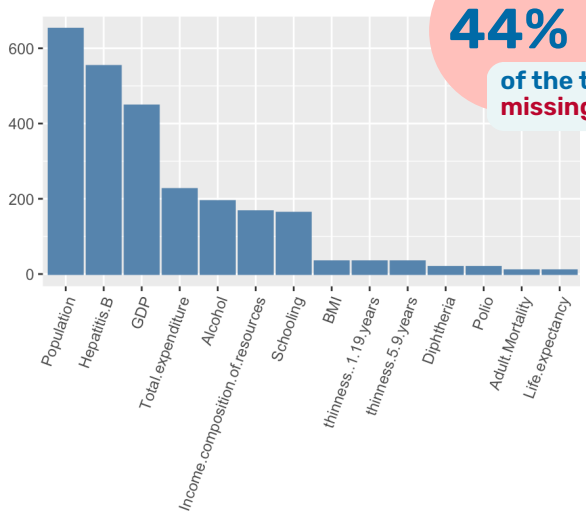
The dataset is from the following link: https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who. This dataset contains information about various features that may affect the life expectancy of individuals in different countries. All features was divided into several broad categories, such as Immunization related factors, Mortality factors, Economical factors and Social factors.

Top 5 out of 193 countries with the highest Life Expectancy Rate




```
        Pearson's product-moment correlation

data:  df_clean$Status and df_clean$Life.expectancy
t = 29.805, df = 2936, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4537077 0.5092489
sample estimates:
      cor
0.4819623
```

Number of Null Values in Each Variables



44%

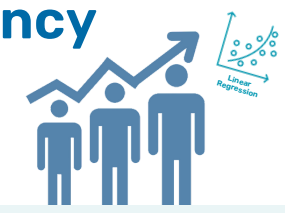
of the total data is **missing** on the data

*it is quite a lot, but it is understandable as finding data for lesser-known countries such as Vanuatu, Tonga, Togo, Cabo Verde, etc. is quite difficult.

*To address this anomaly, we inputed the missing value of each variable with its means value.

Where of the top 5 countries, most are occupied by developed countries compared to developing countries

Life Expectancy Over Time



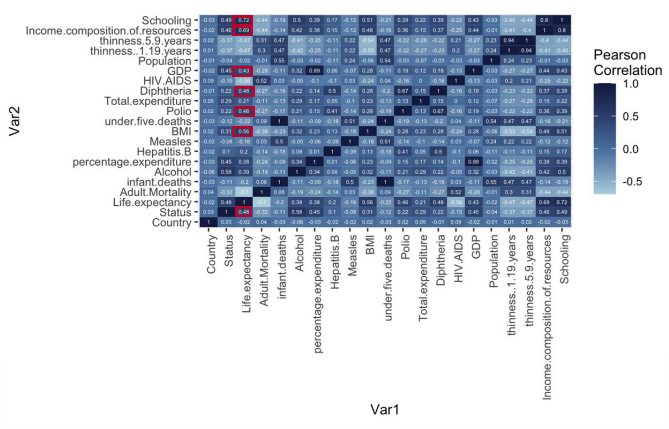
From 2000-2015 there was a significantly increase in life expectancy every year*

*based on the linear regression model that has been done

```
Call:
lm(formula = Life.expectancy ~ Year, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-33.792  -6.442   2.833   6.258  21.005

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -632.47819   75.24081  -8.406   <2e-16 ***
Year           0.34954    0.03748   9.326   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.371 on 2936 degrees of freedom
Multiple R-squared:  0.02877,   Adjusted R-squared:  0.02844
F-statistic: 86.98 on 1 and 2936 DF,  p-value: < 2.2e-16
```

Correlations of Each Variables*

*Calculated with Pearson Correlation



As per the output, we can see that the variables "Year" and "Life.expectancy" **have a significant correlation**, and the variable "Year" significantly affects the variable "Life.expectancy". This can be interpreted from the **p-value** given in the output, which is **"<2e-16"**, which means it is far below the significance level of 0.05.

In addition, we can also see the regression coefficient for the variable **"Year" which is 0.34954**. This coefficient indicates that every one unit increase in the "Year" variable (i.e., one year), is expected **to increase the life expectancy value by 0.34954**.

Predictive Modeling*

*Calculated with predictive modeling randomForest supervised machine learning

```
Call:
 randomForest(formula = f, data = train)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 7

          Mean of squared residuals: 0.001175487
                    % Var explained: 96.44
```


The **Mean Squared Error (MSE)** obtained from the predictive modeling is **approximately 0.0009394606**

```
[1] 0.0009394606
```

This model has **good accuracy with the points mostly close to the trend line** (blue line)

Another plot of predictive modeling

It can be seen that the "HIV.AIDS" feature has a very high **"IncNodePurity" value, which is 21.1817738**. This indicates that the **"HIV.AIDS" feature is very important in predicting the value** of the target or dependent variable.

Scan This QR Code to See Interactive Visualization for Life Expectancy Dataset

Powered by Tableau

Alyza Rahima Pramudya
2502032125
Computer Science
BINUS University
alyza.pramudya@binus.ac.id

Faishal Kamil
2502001063
Computer Science
BINUS University
faishal.kamil@binus.ac.id

Shafa Amira Qonitatin
2502009173
Computer Science
BINUS University
shafa.qonitatin@binus.ac.id