

Data:

We got the data from site (www.kaggle.com)

And we will use two data:

1)world happiness report in 2021

(rows=149, columns=20)

```
columns=(['Country name', 'Regional indicator', 'Ladder score',  
         'Standard error of ladder score', 'upperwhisker', 'lowerwhisker',  
         'Logged GDP per capita', 'Social support', 'Healthy life expectancy',  
         'Freedom to make life choices', 'Generosity',  
         'Perceptions of corruption', 'Ladder score in Dystopia',  
         'Explained by: Log GDP per capita', 'Explained by: Social support',  
         'Explained by: Healthy life expectancy',  
         'Explained by: Freedom to make life choices',  
         'Explained by: Generosity', 'Explained by: Perceptions of corruption',  
         'Dystopia + residual'])
```

2)world happiness report

(rows=1949, columns=11)

```
columns=(['Country name', 'year', 'Life Ladder', 'Log GDP per capita',  
         'Social support', 'Healthy life expectancy at birth',  
         'Freedom to make life choices', 'Generosity',  
         'Perceptions of corruption', 'Positive affect', 'Negative affect'])
```

What we are going to do:**Problem statements:**

- Cleaning data
- Merge two datasets
- How does each strong happiness factor correlate with one another in 2021?
- What is mean of corruption in all countries in Regional indicator in 2021?
- What the happiest and saddest countries (top 10 and bottom 10 in 2021)?
- How is the Healthy life expectancy factor in the happiest and saddest countries in 2021?

.Models

Important features can be extracted through feature importance ability in random forest regression and for the beginning step all the features are inputted

The regression model is possible with Ladder score as the predictor variable

Regression models like random forest linear regression and other regression models will be tested and the best model will be selected

Method:

Initially, the dataset is divided into training and testing. Then, the testing set will be used for evaluation RMSE or other techniques will be used to quality the accuracy of the model and to choose the best fit model.

Tools:

Pandas, numpy, matplotlib, seaborn.

Done by: Hana Alamri & Hind Alasker