**Introduction**

1. **Background:**

The world happiness report gives us a clear understanding of rankings of different countries based on happiness quotient of people. It is based on several factors which vary from country to country and depending on how they vary a score is given to each country. Highest scorer is the happiest country. These scores change from year to year based on different attributes like GDP etc.

1. **Problem:**

The data collected in this regard will help us find the score of additionally added countries based on the previous sets of data. This in turn would produce an overall rank and score of new countries. Apart from this, data can also be used to determine what factors play an important role in determining the happiness quotient of an individual country and the project aims to predict how countries can improve their happiness quotient.

1. **Interest:**

Now understanding why looking at such a data set is important. For any organization, no matter how small or big, it is important to keep a check on the mental state of the people working there. Only then, the organization grows altogether. Making sure working people are happy lets organizations work more efficiently.

**Data**

1. **Data Sources:**

The main source of this data set has been Kaggle where individual data sets of different years is available. For this project I’m considering the 2019 year’s data.

1. **Data Insights:**

The data consists of seven different variables each of which are numerical values. The numerical values range from 0 and go all the way above. Based on the values of independent feature sets, a score is derived for each country which decides the rank of the country. There are 155 countries data that is available for modelling this project’s model. The independent feature sets namely GDP, social support, healthy life expectancy, freedom to make life choices, generosity, perceptions of corruption are used for finding the dependent variable which is the score. It is also clearly visible from the table that Country or Region column is filled with country names which is non numeric value. Nevertheless since, we are not utilizing this column for our findings it can remain as it is.

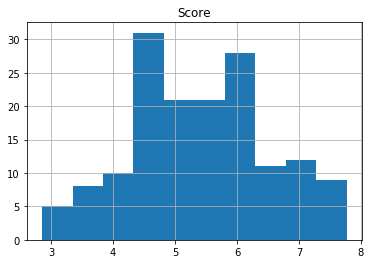
1. **Approach:**

The idea is to utilize 30% of this data to train the model and use the remaining 70% to see how accurate the model is working. Also, when a new country is added to the data we can determine its rank based on the working model available with us.

**Data Cleaning**

The data is cleaned by transforming all the available data into the format required for operations. Generally for all machine learning activities, we use numeric variables (float). The necessary columns required for operating on this data are separated by creating a different data frame.

If we look at the data available to us, we have Score which is the determining variable and the rest like GDP, Social Support, Life Expectancy, Freedom, Generosity, Perceptions, are variables which decide the score based on weightage factor. For this very purpose all the variables are separated into a different data frame mdf, where the required variables are used for training and testing the given data set.

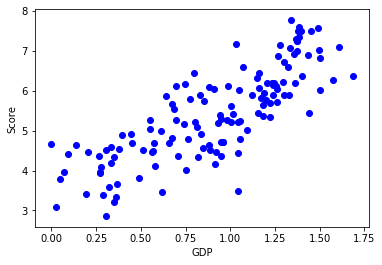


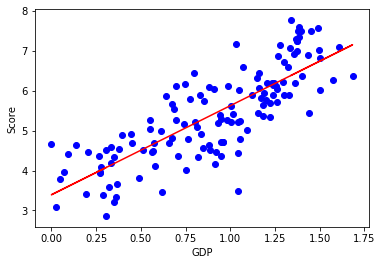
Here we can see histogram of country and score. Since, country has an input variable of string, the histogram just shows the variance between different countries at different number points.

### Modelling

Once the given data is all cleaned and filtered through, the next step is to train and test the data. This is done using various methods available based on the data and the type of prediction required. Here we are doing a linear regression on the model to see how different attributes play a role in changing the overall score of the country. Below we can see the dependency of score on GDP. Here the independent variable GDP varies from country to country. But from the scatter plot below we can see a correlation between GDP and score. When the independent variable GDP is increasing there is a rise in score thereby leading to positive correlation.

Below is a scatter plot which shows the correlation between GDP and score. A positive correlation can clearly be seen since as GDP increases the score of the country automatically goes up. Thereby an attempt has been made to create a model which could give a clear picture of their correlation.



Below is a scatter plot which clearly shows a linear line going through the points. But, we can clearly see, though the relationship between them looks linear, but the line stays way far apart from the points.

The distance from the line to the points seem very far off which clearly states why the negative value of r square.

### Evaluation and Conclusion

An attempt has been made to understand training and testing of the happiness dataset. There are various possible models available to perform the evaluation of such models, but linear regression gives us a better clarity on correlations, which is why it was chosen. Since the available data set was quite small with very less values the model might not have given the perfect evaluations but there is a future prospectus to it, where we can determine the overall score of a new country when given its other independent variable values.