**Problem Definition**

This is a dataset of World Happiness Report. This report gives the happiness level of various countries. This is the report of 2017, the first report was published in 2012, the second in 2013, the third in 2015 and the fourth in 2016. This report ranks 155 countries by their happiness levels. The happiness score column is estimated by six factors –Economy, Family, Health, Freedom, Trust, Generosity and it also evaluates the life in each country with Dystopia Rsidual. Dystopia is a hypothetical country that has values equal to the world’s lowest national averages for each of the six factors. I have done this project to make a predicting model that predicts the Happiness score by considering the above six factors. This model will help to predict the Happiness score exactly as the original dataset.

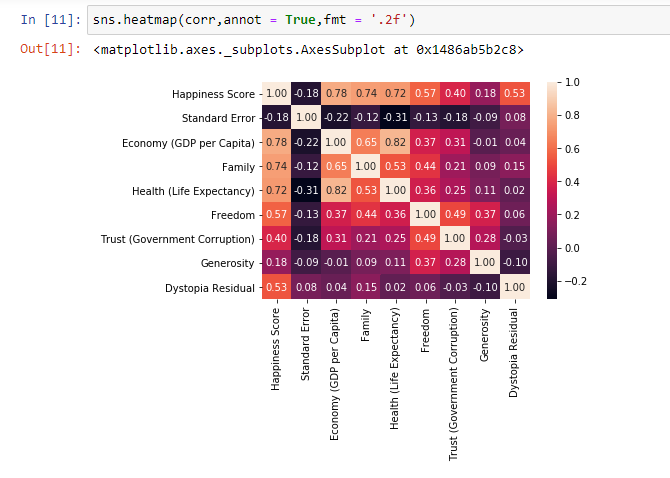
**Data Analysis**

The dataset was collected from github repository. The dataset contains 158 instances and 12 attributes. The dataset was in the .csv format. I have chosen jupyter note book to do the entire project. After loading the data to jupyter note book I dropped three columns ('Country', 'Region', 'Happiness Rank') which are irrelevant for predicting the happiness score. All the columns are of float type and there are no missing values found in the data

**EDA Concluding Remarks**

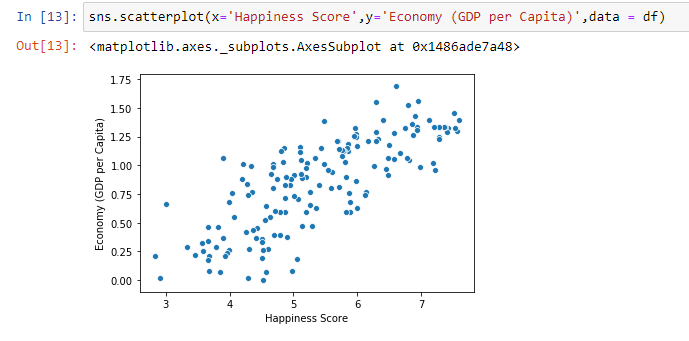
In this dataset happiness score is determined using the six factors (Economy, Family, Health, Freedom, Trust, Generosity and Dystopia Residual). Dystopia is an imaginary county made to have a benchmark against which all countries can be favourably compared in terms of each of the key variables. Residual is differed for each country, reflecting the extend to which the six variables either – over or under – explain average 2014-2016 life evaluations.

After dropping the irrelevant columns, checked for the missing values in the data using isnull() function and there is no missing values in the data. From the describe function we can understand that the min happiness score is 2.83 and the highest is 7.58. When I check the 75th percentile and max it is clear that in some columns outliers are present. Then I find the correlation of these columns with the corr() function. And then plot it in the heat map using seaborn visualization.

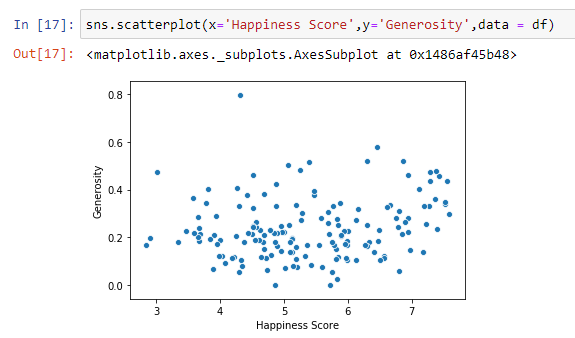
*This pic below shows the correlation diagram of the dataset*

From the heatmap it is clear that Economy, Family, Health and Freedom have high correlation with the Happiness Score. Generosity have the least correlation among them. And the Standard Error have the negative correlation with the Happiness Score.

Then I plot scatter plots to check the relation between all the columns with Happiness Score, to know how it is affecting the happiness score. First, I plot Dystopia Residual with Happiness Score. From it is found that Dystopia happiness and the residual values or unexplained value increases with the Happiness Score and it is positively correlated. Similarly, I have plotted Economy, Health, Freedom and they all are seems to that the value increases with the Happiness Score. While we scatterplot Trust against Happiness Score it was seen that Happiness Score is increasing after 6.5 score only.

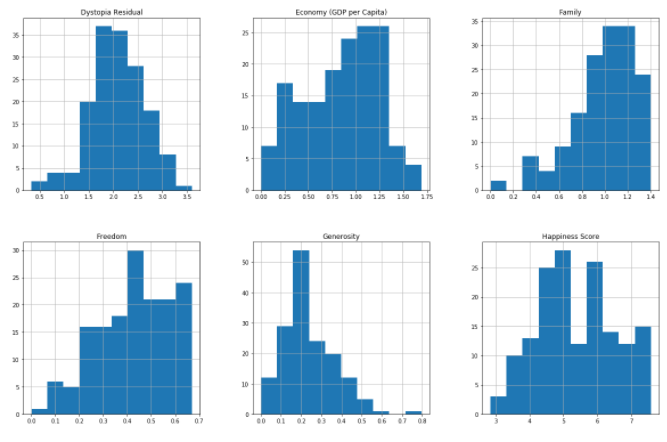


*The above chart shows the positive correlation of Economy with Happiness Score*



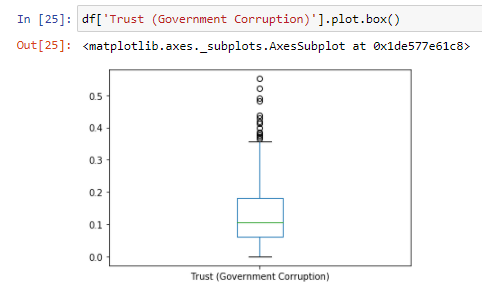
*Above chart shows the least correlation of Generosity which have least correlation with the Happiness Score*

The data is then checked for skewness by using skew() function and found that there are some skewness present in the dataset. So to see the distribution of data of each columns I have plotted all columns in the histogram. From that it is clear that Standard Error, Trust, Generosity are right skewed data and Economy, Health, Freedom, Dystopia Residual are right skewed data.



*The above pic shows distribution of data of various columns*

Then after I have plotted all the columns in boxplot using pandas plot.box() function. Then I identify that there are some outliers present in certain columns such as Family, Trust, Standard Error, Generosity, Dystopia Residual. For removing the outliers I have imported zscore from scipy.stats library. Then the threshold is set to 3. If the zscore value is greater than 3 it means there are outliers. So we have printed all the columns with the outliers and make a new dataset by removing all those outlier column as df\_new. The new dataset have 149 columns, that is 9 columns have eliminated. The data is then split into x and y in which x consists of independent features other than the target ,which is ‘Happiness Score’ column and y consist of ‘Happiness Score’ column.



*The pic above shows the outliers that is present in the Trust column*

Now after removing all the outliers I have imported power\_transform from sklearn.preprocessing library. Then power\_transform is done in x variable by using ‘yeo-johnson’ as the parameter and it is saved to x variable. Then all the data was scaled between 0 and 1 by using MinMaxScaler from sklearn.preprocessing library and created an instance for MinMaxScaler. Then the data is scaled by putting the x data in fit\_transform and saved to x variable.

**Building Machine Learning Models**

Now after all the EDA process have done, the data is then split into training part and testing part. This is done by importing train\_test\_split from the sklearn.model\_selection library. In that x, y, test\_size are given as parameters. Hence finding the happiness score is a regression type dataset I used LinearRegression algorithm and DecisionTreeRegressor

First I have imported LinearRegression from sklearn.linear\_model library.Then I make instance for LinearRegression() as ‘lr’ after that fit the x\_train and y\_train into it and then predict the model by using lr and print the score (with x-train, y\_train as parameters). Then I have imported another metrics named DecisionTreeRegressor from sklearn.tree library. After that I have make an instance for DecisionTreeRegressor() as ‘tree’. And then train the model by using fit method with instance. Then test by predicting the model by giving the x\_test values and calculate the score by give x\_train and y\_train as parameters

**Concluding Remarks**

The best model I choose for this dataset is DecisionTreeRegressor because it giving score of 100 percentage. The other model which I have tested is only giving 99 percentage. So I choose the DecisionTree and save it in a ‘Happiness\_score\_tested.pkl‘ file.