**World Happiness Report**

**Problem Definition**

It is required to build a linear regression model that can be used to predict the Happiness Score of countries through the given values of several attributes such as Economy, Health, Trust, Freedom, Region, Family, and Generosity.

**Data Analysis**

The World Happiness Report is a survey that is used to determine the measurement of happiness throughout the various countries of the world and then rank them according to their Happiness Score. The World Happiness Report of year 2015 is used in which 158 countries were ranked based on their respective Happiness Score. Other than UN, several other countries have started understanding the importance of this report and have started mending their policies to improve the sectors that have a major impact on the happiness of the citizens such as economy, corruption, health etc. to increase the happiness score in the coming future and ensure a better lifestyle and well-being for the people.

The data from Gallup World Poll is used to determine the scores that are based on answers to the question asked in the survey in which scores from 0 to 10 are given based on the current life quality. Dystopia, an imaginary country with values equal to the world’s lowest national averages meaning a country with the world’s lowest incomes, lowest life expectancy, lowest generosity, most corruption, least freedom and least social support, is used as a benchmark to calculate the happiness score of each country.

The 9 columns in the dataset are as follows:

1. Country – There are 158 countries that have been ranked based on their Happiness score.

2. Region – There are total 10 different regions that are used for grouping the countries.

3. Happiness Rank – The rank of each country based on its Happiness Score.

4. Happiness Score – The six factors that are GDP per Capita, Family, Life Expectancy, Freedom, Generosity, and Trust (Government Corruption) when added up gives the happiness score.

5. Standard Error – deviation in the Happiness Score of each country.

6. Economy (GDP per Capita) - The extent to which GDP contributes to the calculation of the Happiness Score.

7. Family - This is having someone to count on in times of trouble and is the national average of the binary responses (either 0 or 1) to the question “If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?”

8. Health (Life Expectancy) - The time series of healthy life expectancy at birth are calculated by the authors based on data from the World Health Organization (WHO), the World Development Indicators (WDI), and statistics published in journal articles.

9. Freedom - Freedom to make life choices is the national average of responses to the question “Are you satisfied or dissatisfied with your freedom to choose what you do with your life”

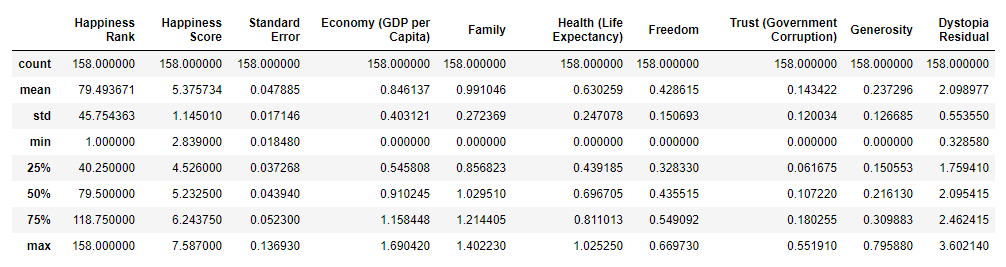
10. Trust (Government Corruption) - The measure is the national average of the survey responses to two questions in the Gallup World Poll: “Is corruption widespread throughout the government or not” and “Is corruption widespread within businesses or not?” The overall perception is just the average of the two responses. The corruption perception at the national level is just the average response of the overall perception at the individual level.

11. Generosity - This is the residual of regressing national average of response to the question “Have you donated money to a charity in the past month?” on GDP per capita.

12. Dystopia Residual - The residuals, or unexplained components, differ for each country, and these residuals have an average value of approximately zero over the whole set of countries.

**EDA Concluding Remarks**

The dataset consists of total 12 columns and 158 rows with no null values present in it.



Following is the description of the dataset:

1. Happiness Score:

* Maximum value = 7.587
* Minimum value = 2.839
* Mean value = 5.375

1. Standard Error:

* Maximum value = 0.136
* Minimum value = 0.017
* Mean value = 0.047

1. Economy (GDP per capita):

* Maximum value = 1.69
* Minimum value = 0.0
* Mean value = 0.846

1. Family:

* Maximum value = 1.40
* Minimum value = 0.0
* Mean value = 0.99

1. Health (Life Expectancy):

* Maximum value = 1.025
* Minimum value = 0.0
* Mean value = 0.0.63

1. Freedom:

* Maximum value = 0.669
* Minimum value = 0.0
* Mean value = 0.428

1. Trust (Government Corruption):

* Maximum value = 0.55
* Minimum value = 0.0
* Mean value = 0.14

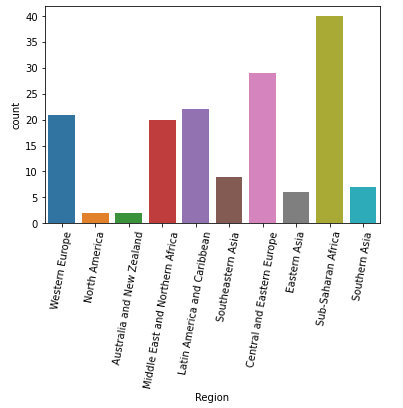
1. Generosity:

* Maximum value = 0.795
* Minimum value = 0.0
* Mean value = 0.0.237

1. Dystopia Residual:

* Maximum value = 3.60
* Minimum value = 0.328
* Mean value = 2.098

Univariate Analysis:



The maximum countries are from the region of Sub-Saharan Africa with a total number 40 countries.

Sub-Saharan Africa - 40

Central and Eastern Europe - 29

Latin America and Caribbean - 22

Western Europe - 21

Middle East and Northern Africa - 20

Southeastern Asia - 9

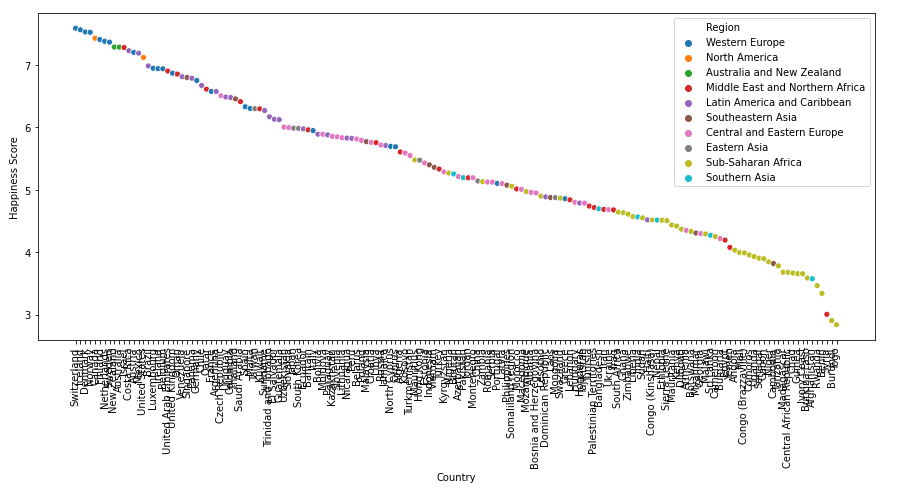
Southern Asia - 7

Eastern Asia - 6

North America - 2

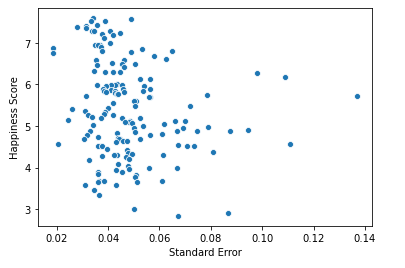
Australia and New Zealand - 2

Bivariate Analysis:

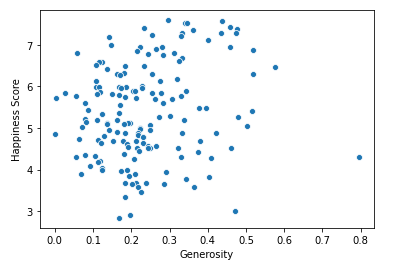


The above analysis shows that most of the countries of Western Europe region, North America, Australia, and New Zealand have the highest Happiness Score whereas the countries belonging to the Sub-Saharan Africa region have the lowest Happiness Score.

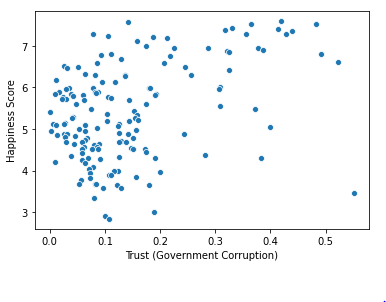
Scatter plots:



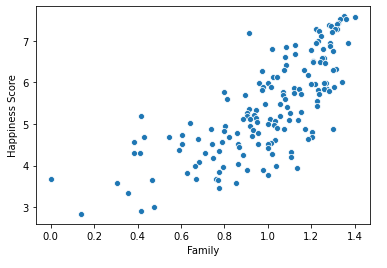
* The happiness score is concentrated between the standard error of 0.3-0.5 and does not show any significant linear correlation.



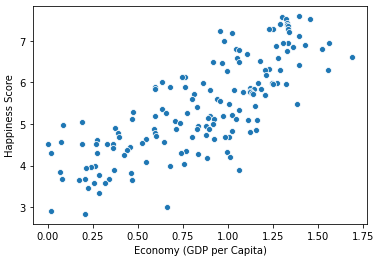
* Generosity does not have any significant linear correlation with happiness score. The values are mostly dispersed between 0-0.4 with few outliers.



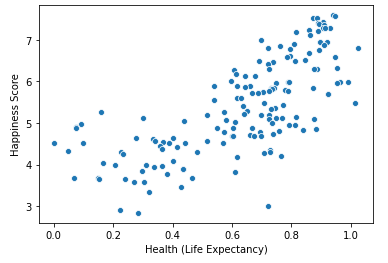
* The happiness score can be seen increasing with the increase in the value of trust which shows that they have a slight positive correlation.



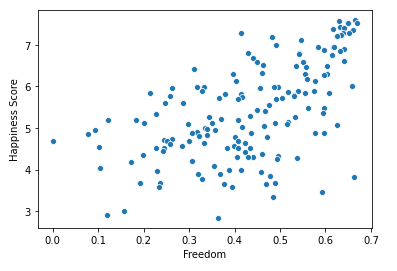
* Family shows a very high positive correlation with happiness score as the data is very less dispersed.



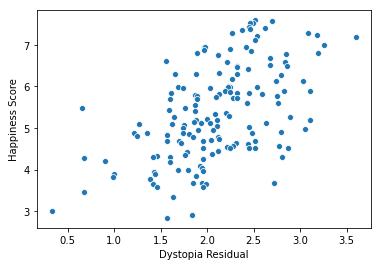
* Economy also shows a high positive correlation with the happiness score.



* Health also has a high positive correlation with happiness score.

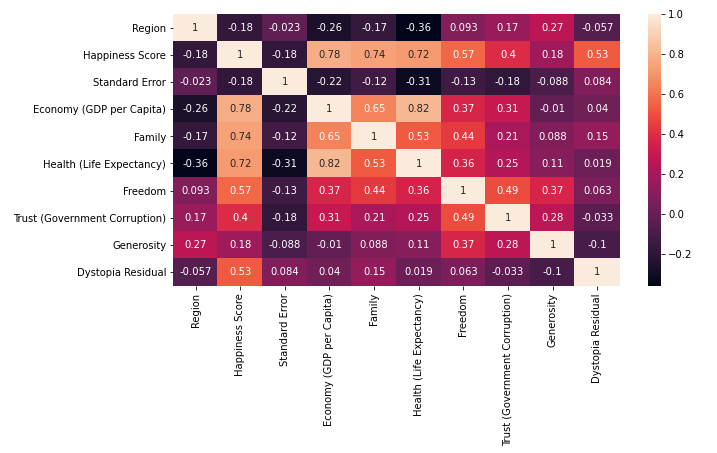


* Freedom has a good positive correlation with happiness score with slightly dispersed data.



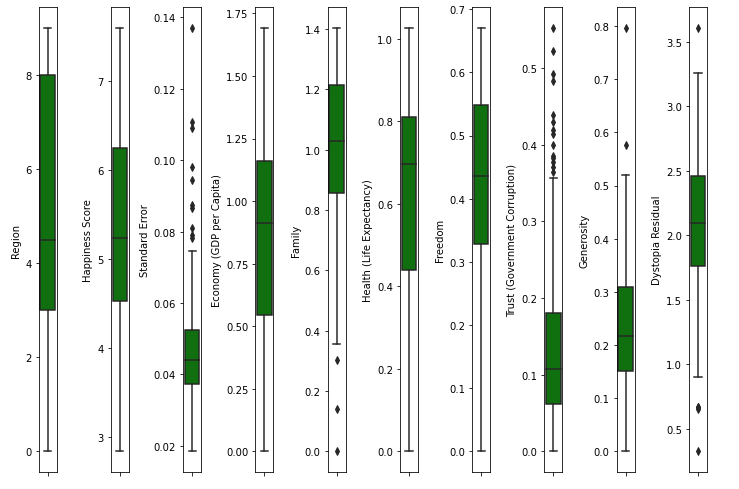
* Dystopia residual also has a good positive correlation with happiness score with slightly dispersed data.

The above observation can be verified using the following correlation heatmap:

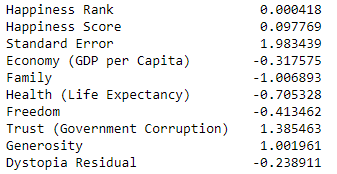


* The correlation heatmap shows that Economy, Family, Health, Freedom, Trust, and Dystopia residual have high positive correlation with Happiness Score.
* Region and Standard error have low negative correlation with Happiness Score.

Checking for outliers using Boxplots:



Skewness:



* Standard error, Family, Health, Trust, and Generosity have high skewness in the data.

**Pre-processing Pipeline**

1. The first and the foremost step of pre-processing is dropping the irrelevant columns from the dataset. These include ‘Country’ and ‘Happiness Rank’ as these columns contains all unique values throughout the 158 columns which are irrelevant to the prediction of the Happiness score.

2. The next step is to encode the categorical data into numerical data. The column ‘Region’ consists of 10 sub-categories that are encoded using Label Encoder into number from 0-9.

3. After all the columns are converted into numerical data type, the z-score method is used to remove the outliers present in the dataset which removes all the data that does not lie in the range of z-score of -3 to +3. The z-score method removes total 9 rows containing outliers after which 149 rows and 10 columns are left in the dataset.

4. In the next step of pre-processing, the target variable i.e. ‘Happiness Score’ is assigned to variable ‘y’ and the rest of the dataset is assigned to variable ‘x’.

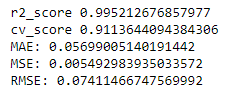
5. One of the most important steps of pre-processing the data is to remove the skewness from the data. To remove skewness, the method of ‘yeo-johnson’ of power transformation is used.

**Building Machine Learning Models**

As our target variable is of continuous type, it is required to use the linear regression algorithms to build our model. The most commonly used algorithm is the LinearRegression from sklearn linear model. To determine the accuracy of the model, scored such as r2 –score, mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE) are used.

1. First a loop is used to determine the best random state for the train\_test split using the LinearRegression algorithm. In our case, at random state of 595 we get the highest training and test scores for a test size of 20%.

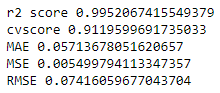
2. The r2-score obtained for the LinearRegression algorithm is 0.995 and the cross-validation score is 0.911. The high value of cross-validation score suggests that there is no over-fitting or under-fitting of the data.



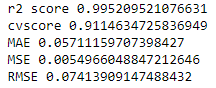
3. To further ensure that there is no over-fitting or under-fitting of the data, Regularization methods that are Lasso and Ridge regression are used.

4. To determine the best parameters for the Lasso and Ridge regression models, the best\_params\_ method of GridSearchCV is used which gives out the most suited parameters for each model.

5. Following are the scores obtained using Lasso regression algorithm:



6. Following are the scores obtained using Ridge regression algorithm:



7. After the fitting of model using various algorithms, it is required to save the best fitted model which in our case is the Ridge regression model as it has the highest score and lowest error along with no over-fitting or under-fitting due to Regularization. The ‘pickle’ method is used to save the best model.

**Concluding Remarks**

1. North America, Australia & New Zealand and most of the countries of Western Europe region have the highest Happiness Score in the world, whereas South Asia and sub-Saharan Africa have the least Happiness Score in the world.
2. The six most important features for determining the Happiness Score of a country are family, generosity, freedom, trust, health, economy.
3. A linear regression model is built using the Regularization and hypertuning methods.