**MACHINE LEARNING ON WORLD HAPPINESS REPORT**

**Date:- 24-Feb-21 Author- Vikas Ojha**

**Problem Definition**

Context:

The happiness report is the survey of global happiness index. The first report was published in 2012. In 2017, a report was released at the United Nations at the event celebrating International Day of happiness on 20th March which contains 155 countries by their happiness level. The report continues to gain global recognition as government organization and civil society increasingly used happiness indicators to inform their policy making decisions.

Content:

The happiness scores and ranking use data from Gallup World Poll. The questions are based on the evaluation questions asked in the poll. These questions are known as Cantrial ladder, ask the respondents were asked to give the rating between 0 to 10 to give their own current lives on this scale, 0 being the lowest and 10 is the highest.

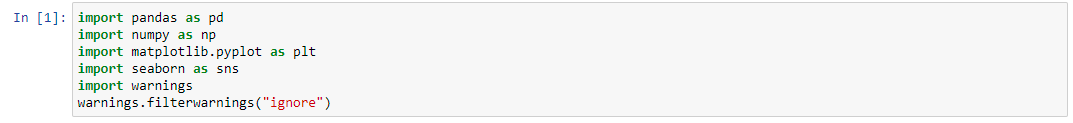
This happiness score as calculated on the six factors namely - economic production, social support, life expectancy, freedom, absence of corruption and generosity. I have chosen Happiness Report dataset of year 2015 as it contains data of 158 countries.

In this project I have to check the happiness status of different countries. Which countries rank the highest in overall happiness and each of the six factors contributing to happiness?

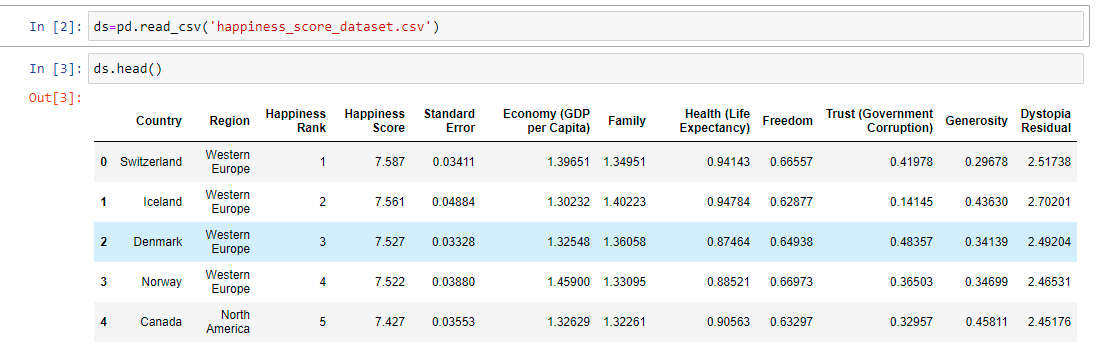
**Data Analysis**

Before I start the Analysis of data, I need to import the supporting libraries such as pandas, numpy, matplotlib, seaborn, etc. These libraries are essentially required to perform task on our dataset.

1. *Support Libraries:*



1. *Loading Dataset:*

**

The problem statement says I have to analyze do you happiness score of different countries so happiness score column.

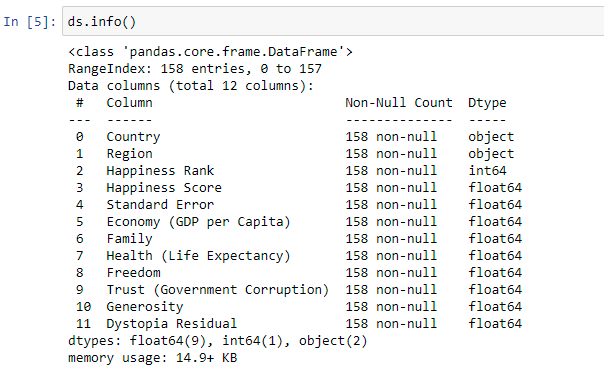
Dataset contains 12 columns namely Country, Region, Happiness Rank, Happiness Score, Standard Error, Economy, Family, Health, Freedom, Trust, Generosity and Dystopia Residual.

Following are the details of each columns: -

|  |  |  |
| --- | --- | --- |
| ***Sr No.*** | ***Column*** | ***Description*** |
| *1* | ***Country*** | *Name of Countries. Dataset contains 158 different countries.* |
| *2* | ***Region*** | *The Area in which country belongs to* |
| *3* | ***Happiness Rank*** | *Rank of country based on Happiness Score.* |
| *4* | ***Happiness Score*** | *Rating between 0 to 10 given by people in evaluation questions where 10 is the highest.* |
| *5* | ***Standard Error*** | *The Standard error of happiness score.* |
| *6* | ***Economy*** | *The extent to which GDP contributes to Happiness Score.* |
| *7* | ***Family*** | *The extent to which family contributes to Happiness Score.* |
| *8* | ***Health*** | *The extent to which Life Expectancy contributes to Happiness Score.* |
| *9* | ***Freedom*** | *The extent to which Freedom contributes to Happiness Score.* |
| *10* | ***Trust*** | *The extent to which Perception of corruption contributes to Happiness Score.* |
| *11* | ***Generosity*** | *The extent to which Generosity contributes to Happiness Score.* |
| *12* | **Dystopia Residual** | *The extent to which Dystopia Residuals contributes to Happiness Score.* |

In this project we will use different exploratory data analysis techniques. Before I proceed further, I need to understand more about dataset. Let’s understand the variables to which we are playing with.

1. *Dataset Types:*



There are 3 different datatypes, namely object, int64 and float64. Country and Region columns are of categorical type which we need to convert afterwards using Label Encoder as machine learning model cannot understand categorical variables.

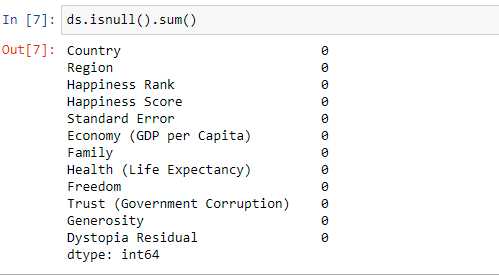
As our target variable is continuous in nature so we will use regression model.

We will perform the following steps further in this dataset: -

1. Finding missing values in dataset, if any
2. Univariate and Bi Variate Analysis
3. Summary Statistics
4. Correlation Check
5. Skewness check
6. Label Encoding of categorical data
7. Checking and removing outliers, if any
8. Defining x and y variables
9. Removing skewness from x variable
10. Apply standard scaler
11. Apply different Algorithms on model
12. Hyper Parameter tuning using GridsearchCV
13. Saving the best model

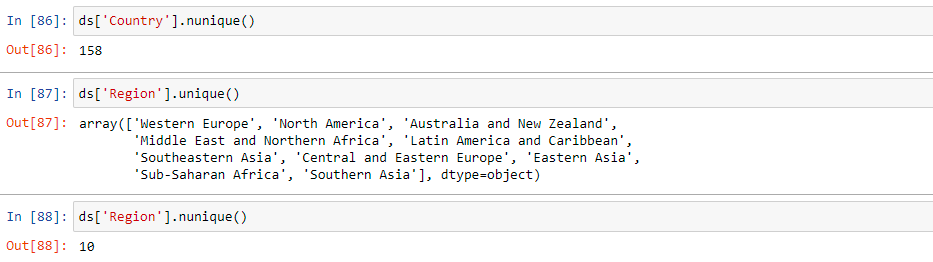
**Exploratory Data Analysis**

As discussed above, now I will check the missing data in Dataset.



Luckily, we don’t have any missing values in dataset.

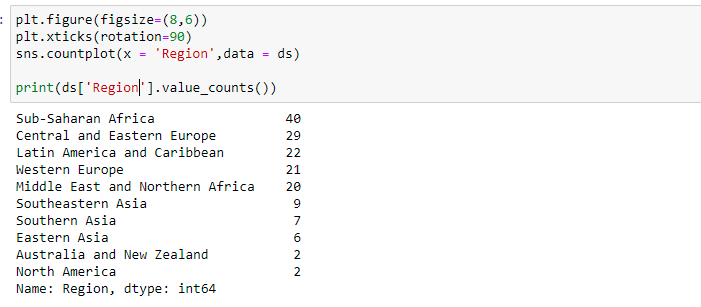
Let’s check the unique values in country and region:

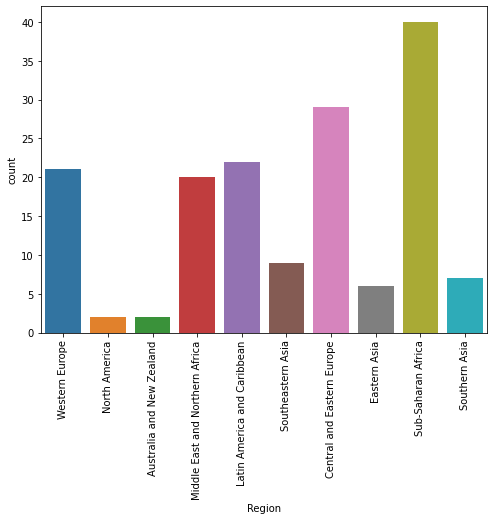


It was found that there are 158 countries from 10 different region namely *Western Europe, North America, Australia and New Zealand, Middle East and Northern Africa, Latin America and Caribbean, Southeastern Asia, Central and Eastern Europe, Eastern Asia, Sub-Saharan Africa, Southern Asia*

**Univariate Analysis**

We have plotted graph for Region column. Following are the insights from it:

****



We’ve plotted frequency distribution of Region column of dataset. I have found from the distribution that Sub-Saharan Africa area has maximum countries in it. It has 40 countries followed by Central and Eastern Europe and area North America and Australia and New Zealand has min countries.

**Mean of Happiness Score**

Now I’ve plotted the frequency distribution of Happiness Score column. Below are the findings from plot.

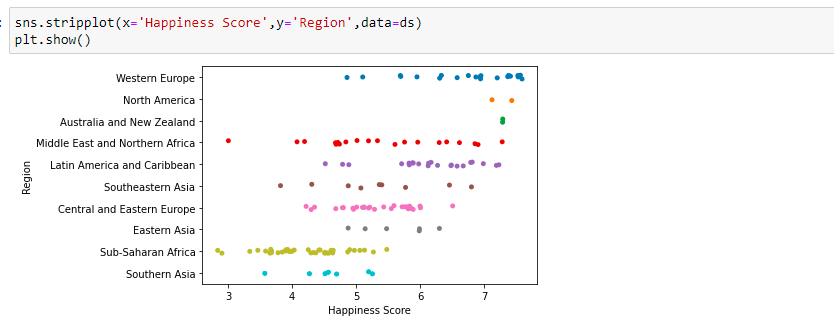
****

It was found from the frequency plot that the mean of Happiness score is 5.4.

**BIVARIATE ANALYSIS**

**Region with Highest Happiness Score**

Now let’s see which Region has the maximum happiness score. I’ll plot strip plot to observe.

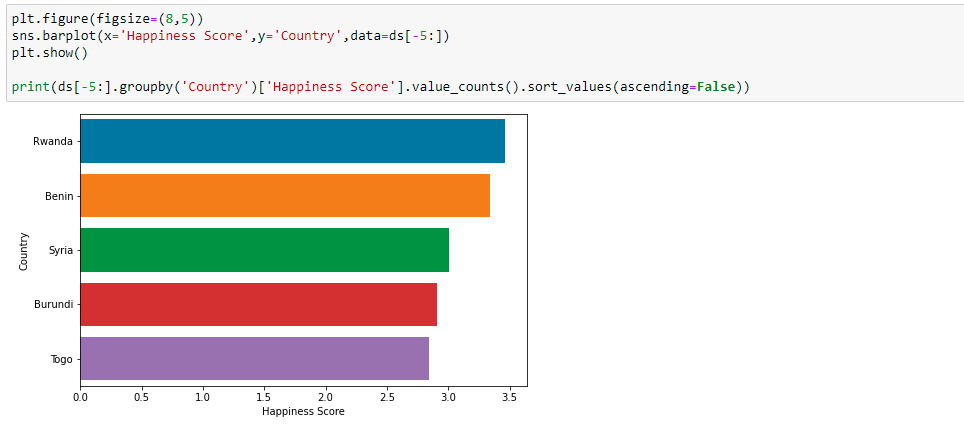


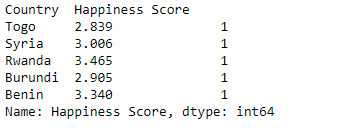
I’ve found an interesting fact that Western Europe has the maximum happiness and Southern Asia has the lowest score.

**Top 5 countries with highest Happiness Score**



We have plotted the bar plot to identify the top 5 countries which has maximum Happiness Score. It was observed from the plot that Switzerland has the maximum Happiness Score.

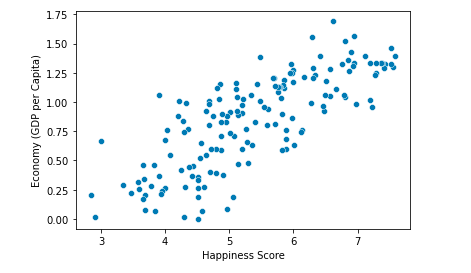
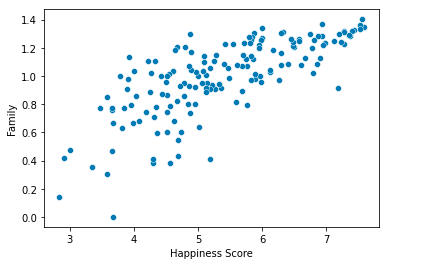
**Country with Lowest Happiness Score**

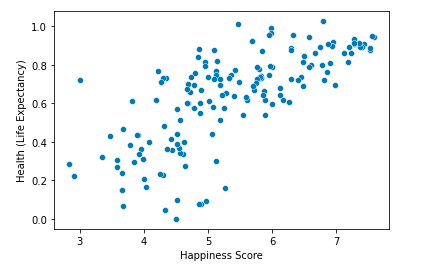
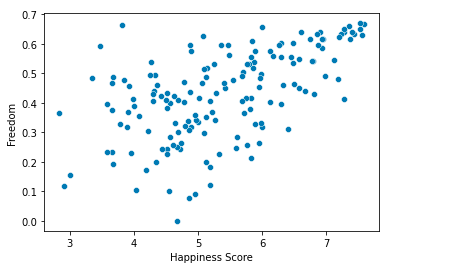


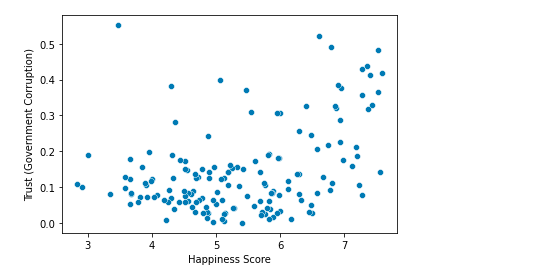
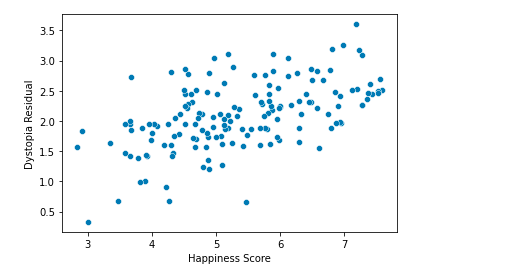
Observations:

We have plotted a graph to look which country has the lowest happiness score. It was observed that Rwanda, Benin, Syria, Burundi and Togo have the lowest Happiness Score. The Happiness Score of Togo is 2.84 which is the lowest among all countries.

Now let’s have a look which factors are affecting the happiness score.



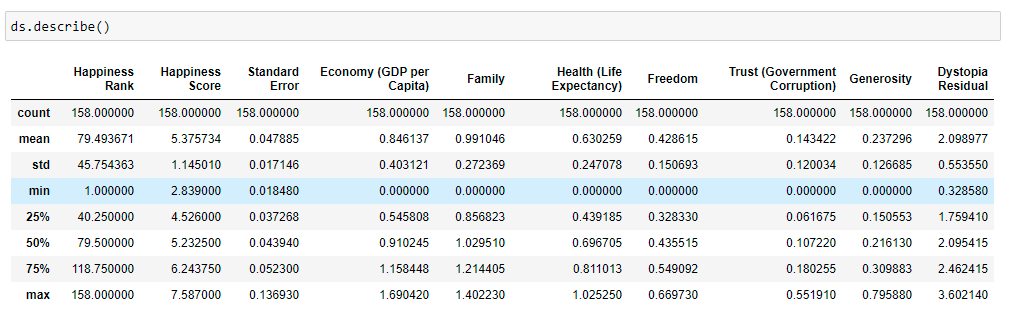




I’ve plotted the scatterplot to find which factors are affecting the Happiness Score. It was found that all the factors are positively correlated with the Happiness Score.

1. GDP-Happiness Score – It can be seen from plot that GDP is playing very important role in building happiness score. As the GDP of country is increasing, the score of happiness is also increasing.
2. Family-Happiness Score – From plot it was observed that Family is playing very important role in Happiness Score. The people living with family are happier than the people living alone.
3. Health-Happiness Score – It can be seen that health is positively correlated with score. In the country where the health is good, more are the people are happier.
4. Freedom-Happiness Score - It can be seen from the plot that in country where there is more freedom given by government, more are their people are.
5. Govt. Corruption – Happiness Score – In country where the corruption is more, their people are less happy.

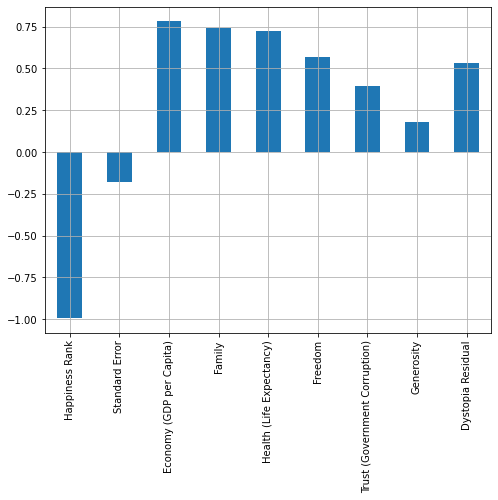
**Summary Statistics**

Let’s look at summary statistics of dataset and draw some observation:

Observations: -

1. As the dataset contains 158 countries, so maximum happiness rank is 158.
2. Maximum Happiness Score is 7.58 and Minimum is 2.84.
3. Highest GDP is 1.69 and lowest is 0.
4. Standard Error, Family,Trust, Generosity and Dystropia Residual have outliers present. Observation drawn from 75 Quantile and Max value.

**Correlation Check**

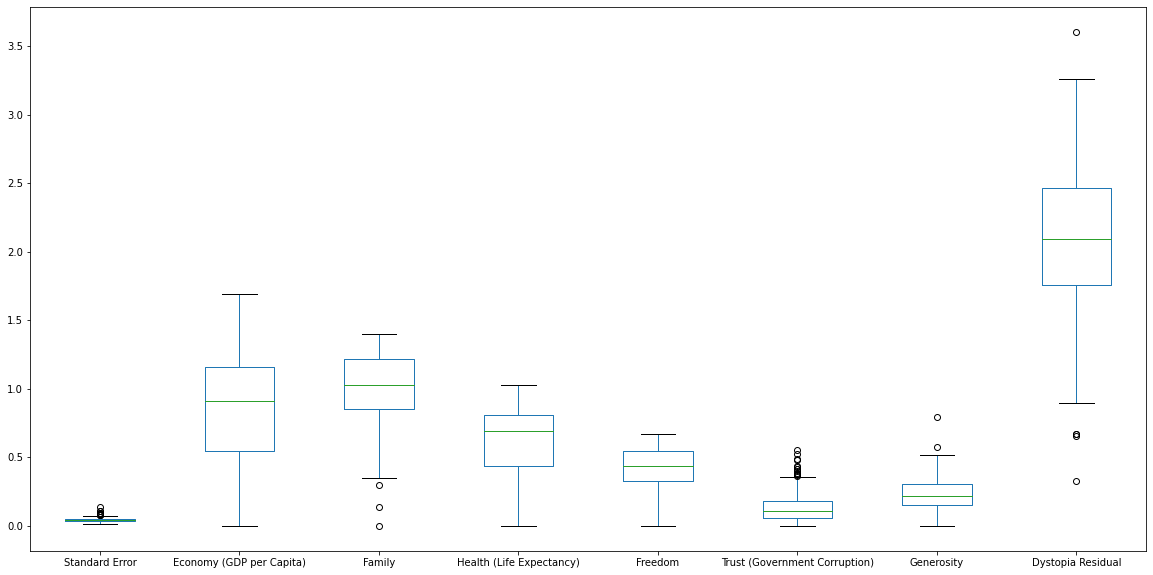
Now let’s find Correlation of different attributes with Target and draw some observations. Our Happiness score is the Target.

Observations: -

Economy, Family, Health, Freedom, Trust, Generosity, Dystropia Residual are highly correlated with target variable while Standard Error is negatively correlated with target.

**Checking Outliers**

Before proceeding to model building let’s check outliers in dataset. I’ll plot boxplot to determine outliers.

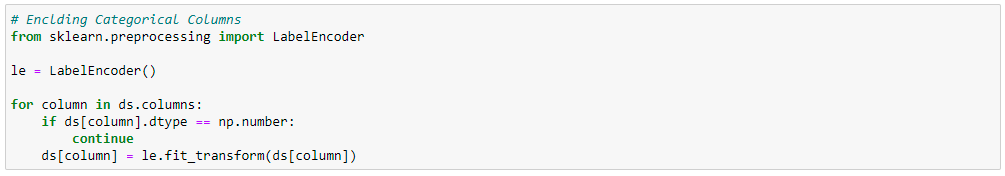


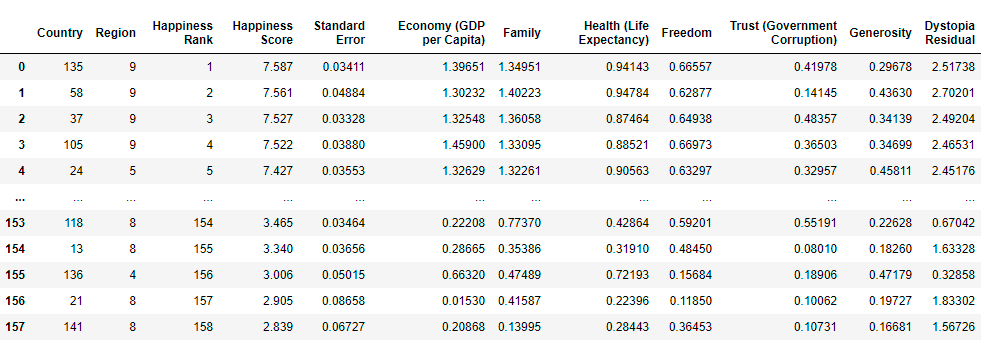
Observations:

It was observed that in Standard Error, Family, Trust, Generosity and Dystropia Residual have few outliers present. The same was observed from summary statistics. I need to check whether I need to keep the outliers or remove it.

**Label Encoding**

As seen in dataset types, column Country and Region is in categorical type, we need to convert it using Label Encoder. Let’s convert it to numerical form. I’ve passed loop method to Encode categorical data.

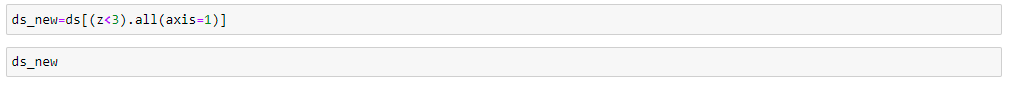
****

****After Encoding our dataset looks like below. All the columns are in numeric form. It is now ready for preparing machine learning models.

**Outliers Removal Techniques**

****

****

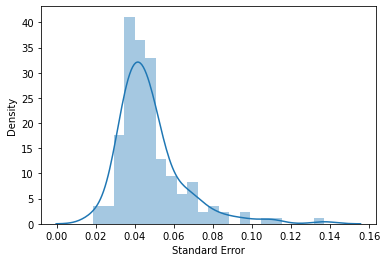
****

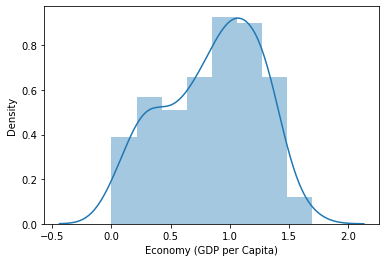
Observsations:

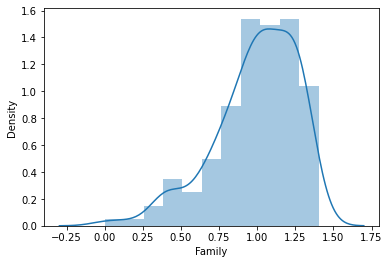
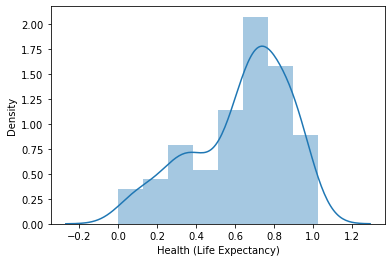
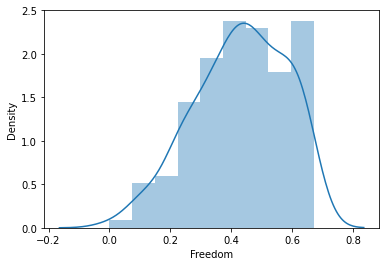
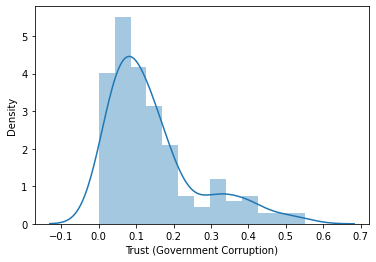
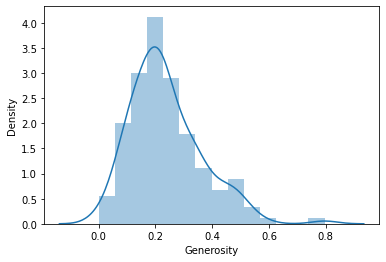
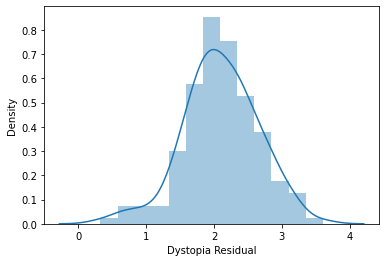
I’ve chosen zscore method for the removal of skewness. For that I need to load zscore from scipy.stats. Taking threshold limit = 3, it was found that approx 5% of data is lost. I have chosen to remove the outliers.

**Skewness Check**

After removal of outliers, let’s plot histogram to check the skewness of data.







Observation:

1. From the distribution plot it can be observed that Economy, Family, Health, Freedom and Dystropia Residual are left skewed.
2. Standard error, trust and Generosity are highly skewed.
3. Need to perform skewness removal techniques to remove the skewness.
4. I’ll use power-transform method for removal of skewness

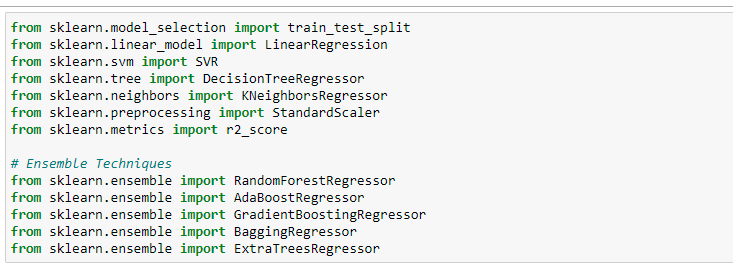
Note: - I will use power transform method on all columns except target column i.e. Happiness Score.

**Building Machine Learning Models**

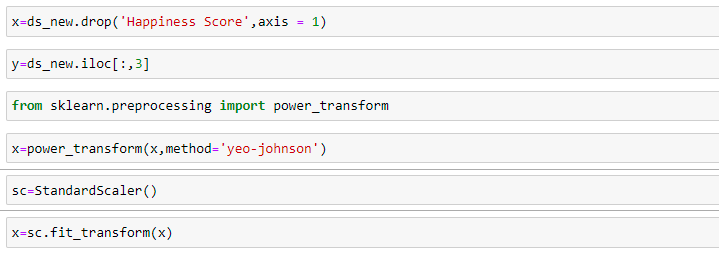
As Happiness Score is our target variable and we can see that the values in this column is continuous in nature so we will use Linear Regression Model in our Machine Learning process. We will try different models of regression like Linear Regression, Decision Tree Regressor, KNeighbors Regressor, Lasso, Ridge,etc. We will also use Ensemble techniques like Random Forest Regressor, Ada Boost Regressor, Gradiant Boosting Regressor, Bagging Regressor and Extra Tree Regressor.

Before we proceed to build the model using different algorithms, I need to remove the skewness from dataset. As discussed above I’ll use power transform method to remove the skewness. Before proceeding we need to import machine learning models from sklearn.

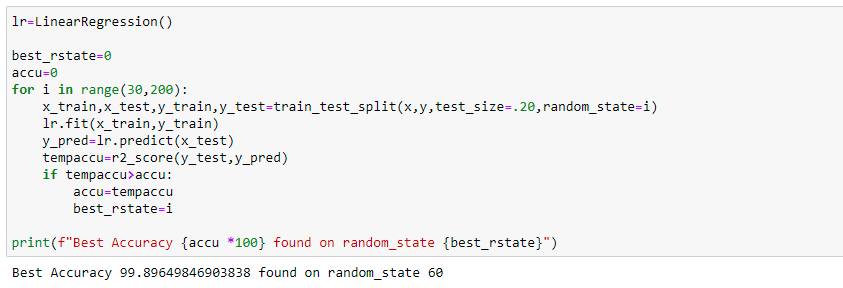
**Importing Machine Learning Algorithms**



Let’s define x and y variables where y is the target variable. As we have defined x and y. Now let’s we will remove skewness from x as column. We will use ‘yeo-johnson’ method for skewness removal as this method works perfect on positive and negative skewness. After that standard scaler will be used to bring the dataset in a similar scaled format.



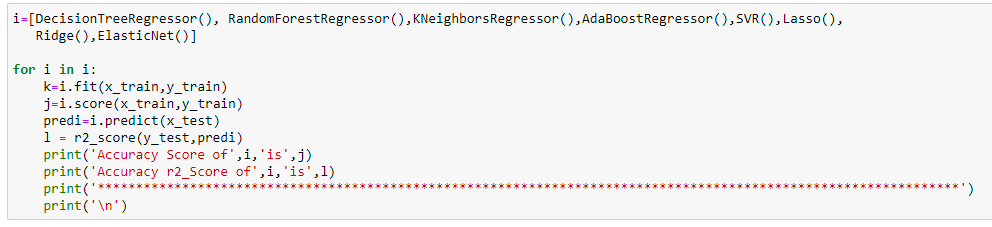
Now we will pass multiple algorithms in a loop to check which model works perfectly on our dataset. Initially we will find the best random state and r2 score. Next step will be to perform cross validation score to control overfitting/underfitting. Then perform Hyper Parameter tuning using GridSearchCV to check if we can improve our score. Our next step will be to verify the Mean and Standard Deviation of the model. The model whose score, r2 Score in maximum and Standard Deviation is lower will be our final model and we can save the model for production.



The accuracy of 99.90% was achieved with random\_state of 60. Let’s split the data into train and test data. 20% of data will be our training data and rest will be for testing.

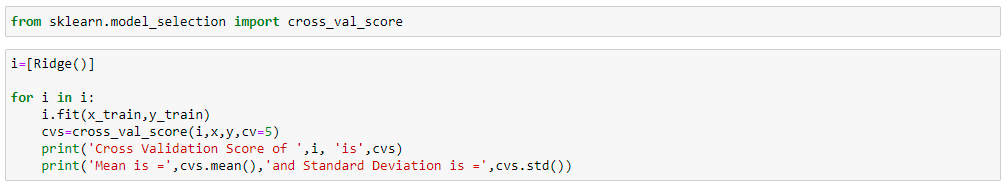


Now let’s check which model performs the best on our dataset.



It was observed that Ridge Regression is giving the best score and r2 score. The score of 99.54% and r2 score of 99.90% was achieved. So, we can say that Ridge model is the best for our dataset.

Now let’s cross validation score of best model.



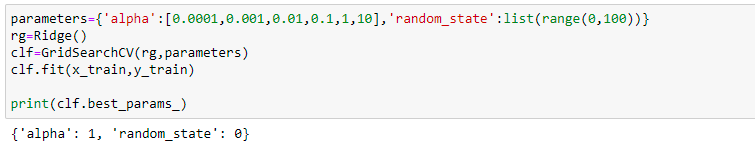
We have got the below result after performing cross validation score: -

Cross Validation Score of Ridge () is [0.98284843 0.87967837 0.90273381 0.85153872 0.76702902]

**Mean** is = 0.8767656714504051 and **Standard Deviation** is = 0.07016396412743313

It can be seen that our model ridge regression is also giving good accuracy in cross validation score.

Now let’s perform Hyper-Parameter Tuning using GridSearchCV to check overfittng or underfitting of model. First find out the best parameters for hyper parameter tuning.

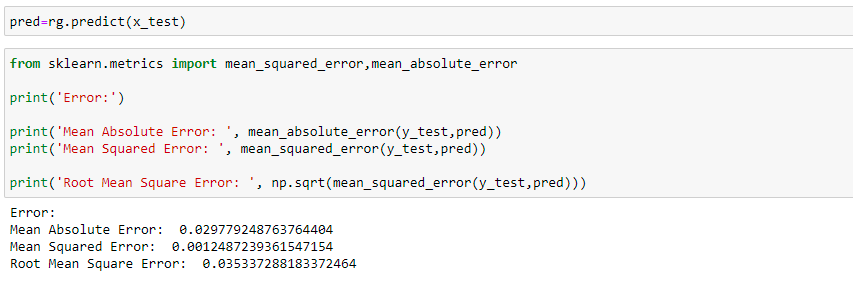


We have found the best parameters for Grid Search CV to be alpha = 1 and random state = 0. Let’s fit this parameter in model to find out the score of our model.



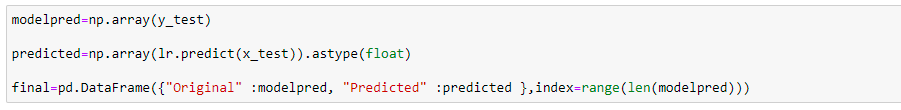
We have achieved a score of 99.54 % even after hyper parameter tuning. Our score has decreased little bit which means we have prevented the model from overfitting.

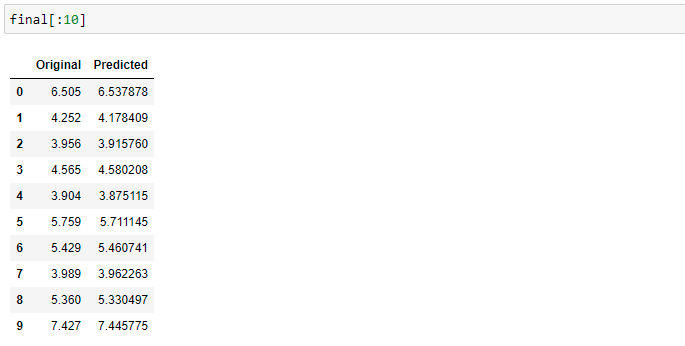
Let’s find out Mean Absolute Error, Mean Squared Error, Root Mean Square Error of our final model.



We can see that the regression loss in errors is very less to we can say our model is good.

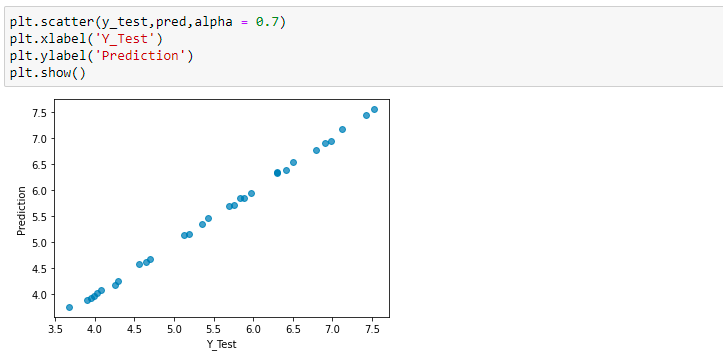
Our model is ready and we are ready to check the original and predicted values given by our model.





We can see that the predicted score by our model is very close to the original score.

Let’s plot it on a graph to analyze the predicted and compare with the original value.

****

We can see that the results of prediction are very close to y\_test which means our model has given good accuracy.

**Conclusion:**

Our objective was to find out top five countries which has highest Happiness Score. We have seen in the analysis that Switzerland on top amongst 158 countries in Happiness Ranking with score of 7.58. Second most happy country is Iceland with score of 7.56.

Also, we have seen that the six factors economic production, social support, life expectancy, freedom, absence of corruption and generosity affects the Happiness Score directly. So, we can say that all the factors play vital role in increasing or decreasing the Happiness Score directly. There are much more data to explore in dataset.