

**TASK**

**Exploratory Data Analysis on the WORLD HAPPINESS REPORT FOR 2015 AND 2019**

[](http://www.hyperiondev.com/portal/)

**Introduction**

**THE WORLD HAPPINESS REPORT EDA (EXPLORATORY DATA ANALYSIS)**

We in this world we often say that people are at peace when they are happy. This is because when people are all happy the mind can do things at maximum potential. However, when happiness in people is not there it affects how they live, how they interact with others in their community, psychologically and physically. In this project, we are going to look at data about world happiness report conducted in 2015 and 2019.

This survey looks at national happiness levels across countries as a measure of national progress. The data used is from the Gallup world poll. The answers are based on answers to the main life evaluation asked in the poll. The columns following the happiness score estimate the extent to which each of the 6 factors below contribute to making life evaluations higher in each country than they are in dystopia, a hypothetical country that has values equal to the world's lowest national averages for each of the six factors. They have no impact on the total score reported for each country, but they do explain why some countries rank higher than others.

#### The six characteristics

1) GDP per capita

2) family quality of citizens

3) health of citizens

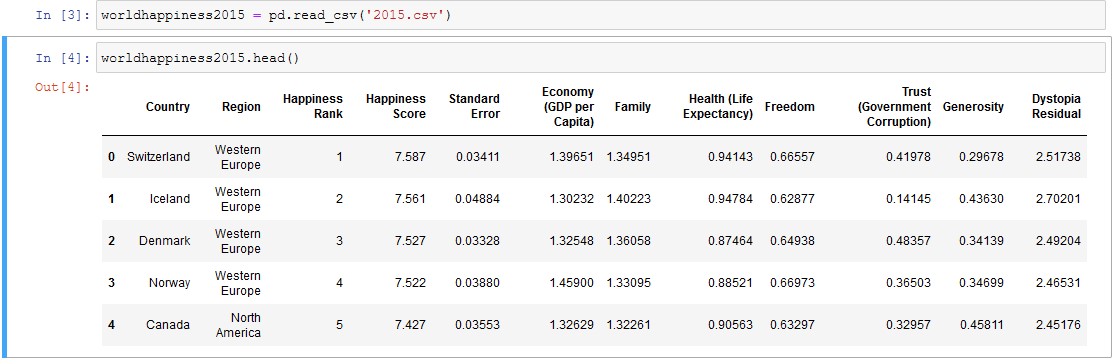
4) freedom of citizens

5) generosity of citizens

6) trust in government

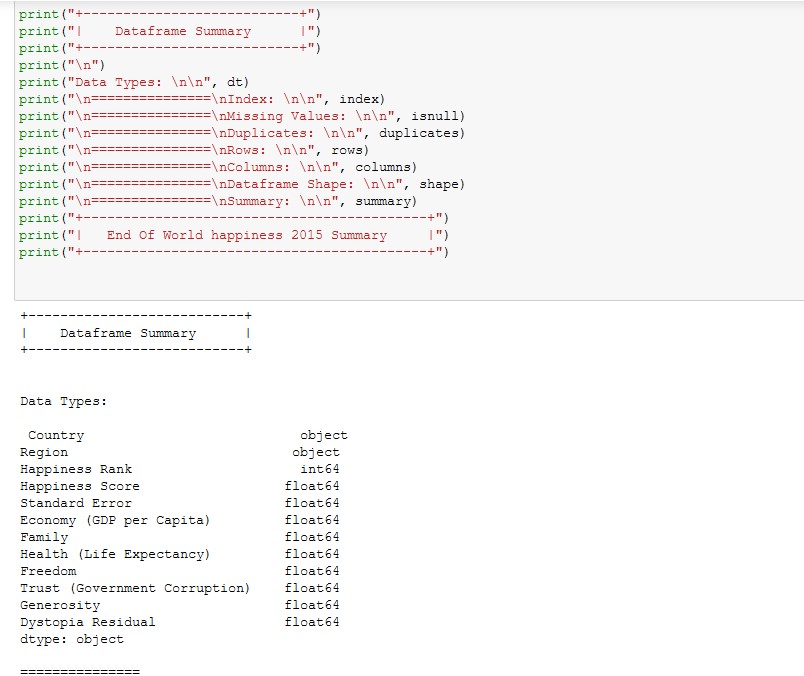
**DATA PREPARATION AND CLEANING**

* **We load the 2015 and 2018 data and view the first 5 rows for each report**

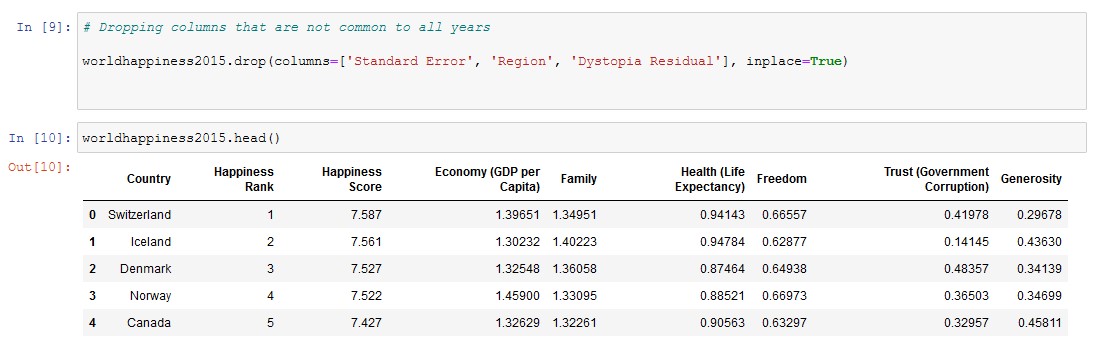
****

****

* **We create a short report of the data which looks at the index, summary, shape, number of columns and rows, and finally display it so that we can use it later for data cleaning**

****

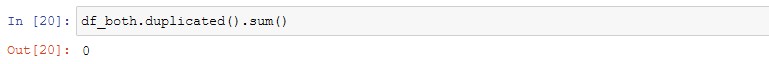
* **It is evident that in 2015 we have a column called Dystopia residual, standard error and region which are not there in 2019. Because of this uncommon of variables, we are going to drop them for uniform data frame. We then look at the data again**

****

* **The data aligns with the 2019 one however the naming is different so we make these changes and rename the columns so that they match. This is to prepare for merging of the two-separate data. The column called YEAR is added to distinguish between data from 2015 and one from 2019. The following is the merged data; the column year is visible at the end.**

****

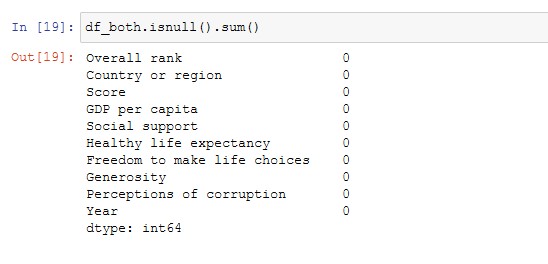
* **We checked for duplicates in our data, there was none**

****

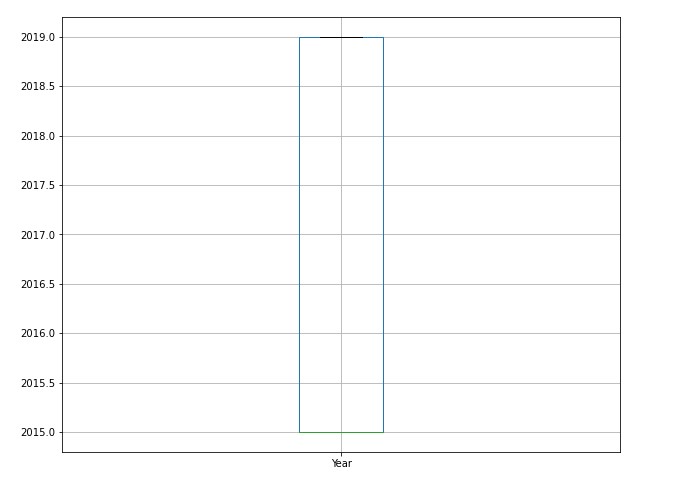
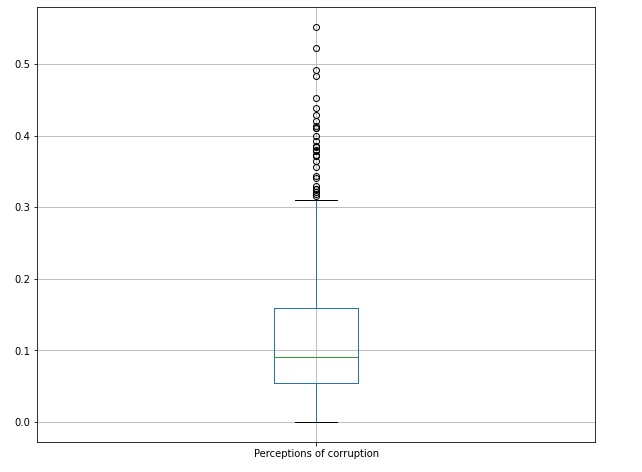
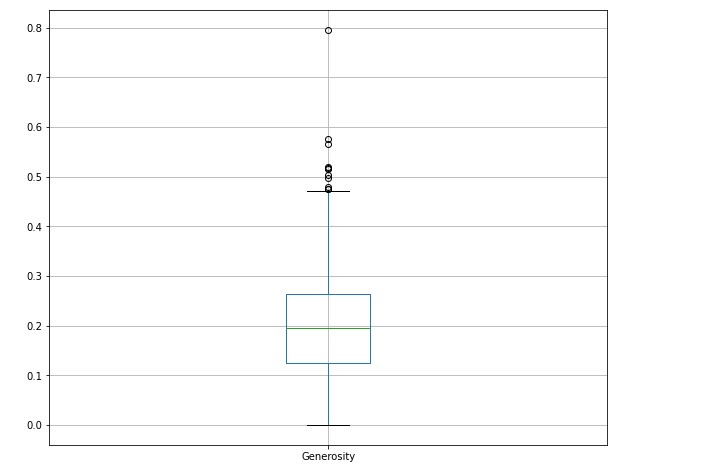
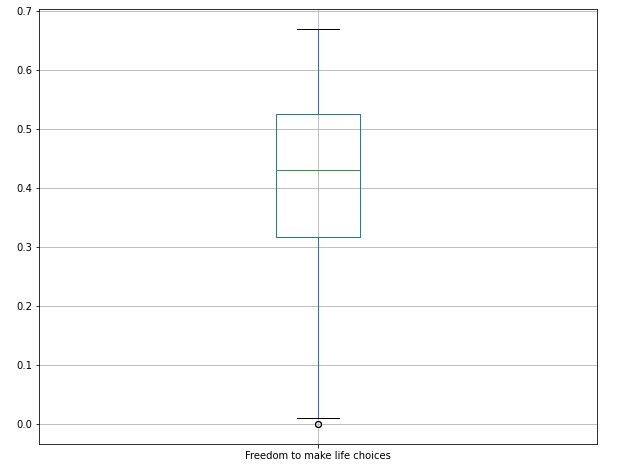
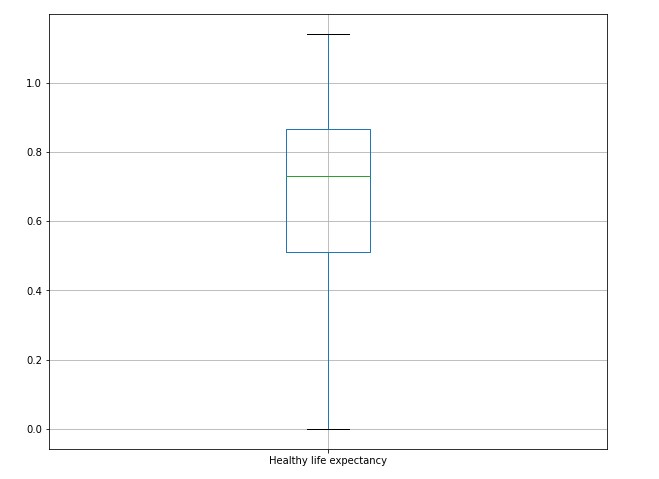
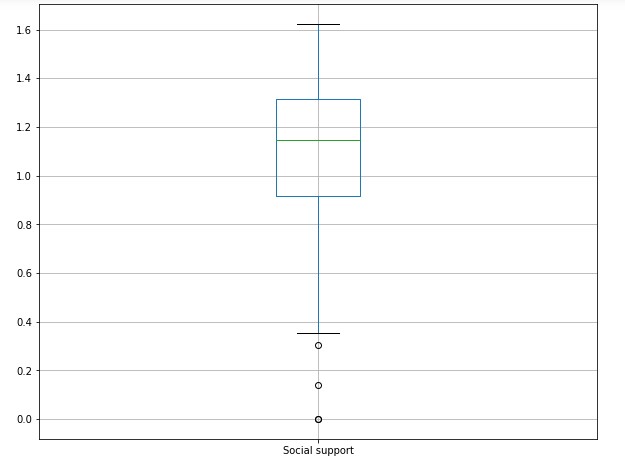
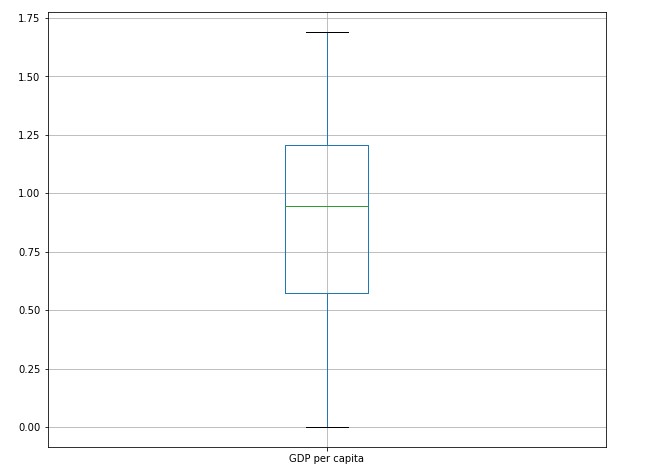
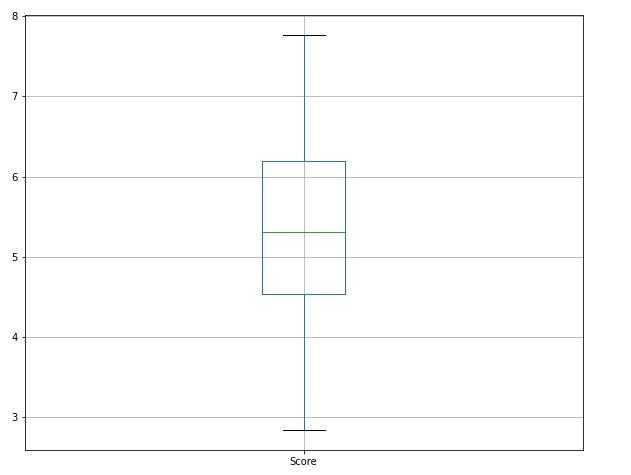
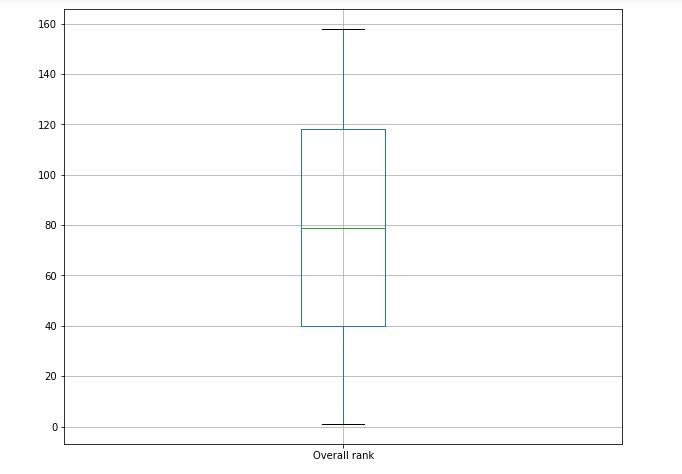
MISSING DATA

# ANY MISSING DATA? HOW DID YOU HANDLE IT

* After checking for missing values from our merged data we found that there is none. So, no step was further taken in this regard.



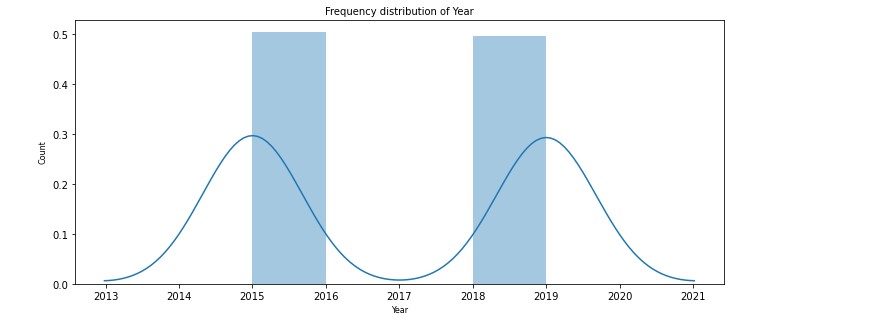
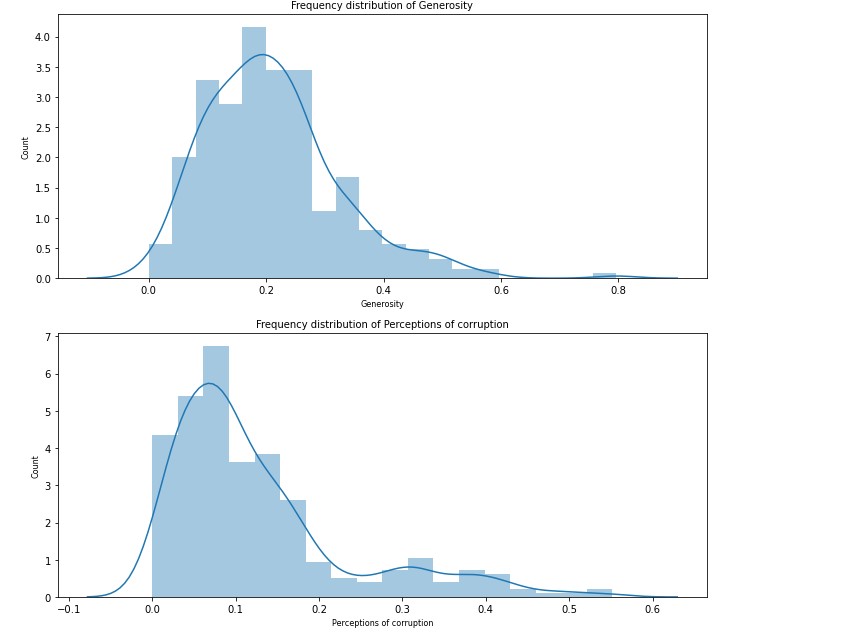
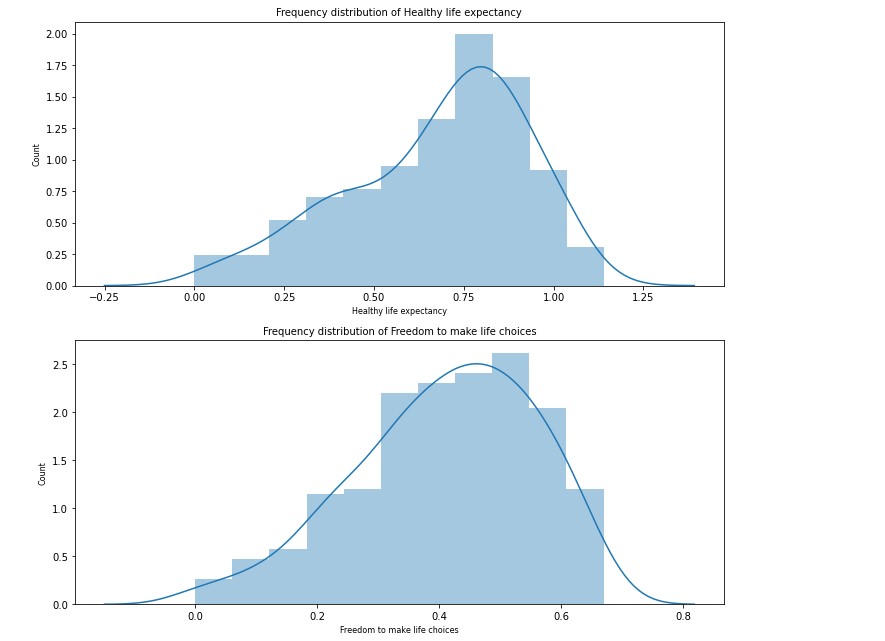
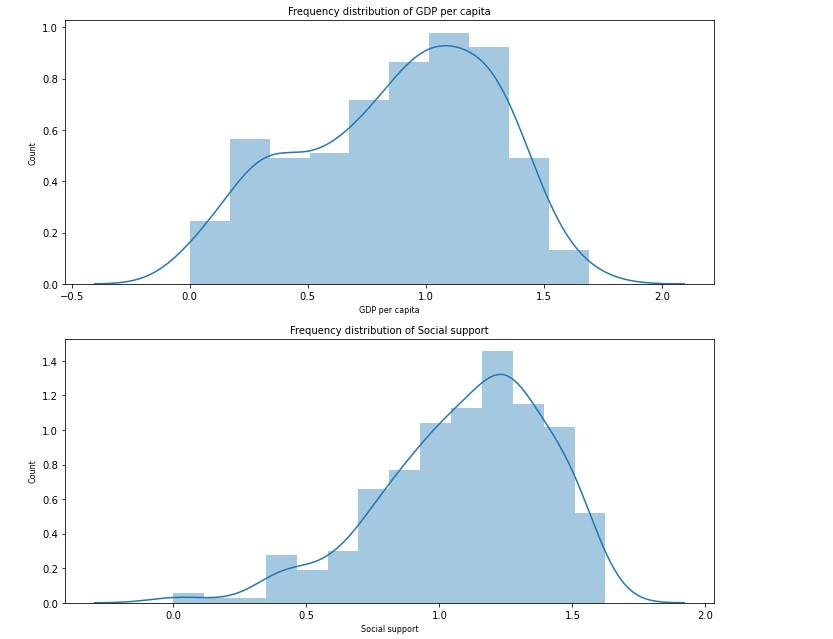
DATA STORIES AND VISUALIZATIONS

1. Checking for outliers in our data

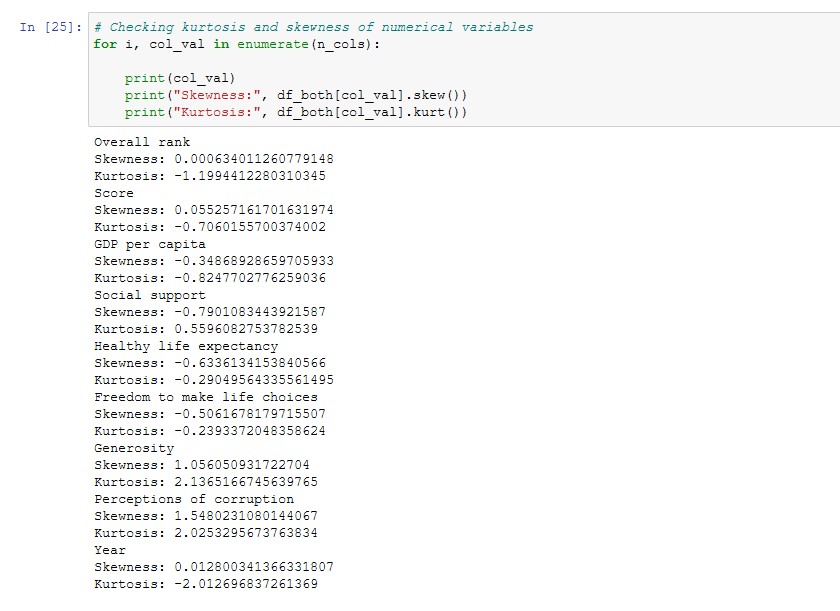
* There are columns that are having some outliers. This means that we need to scale this data if we are going to use it for modelling. We will first use the describe function to look at our statistical description then look at numerical variable distribution which will be graphical.



VARIABLE DISTRIBUTION



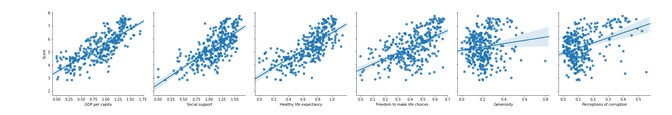
WE WILL LOOK AT THE SKEWNESS AND KURTOIS OF OUR VARIABLES



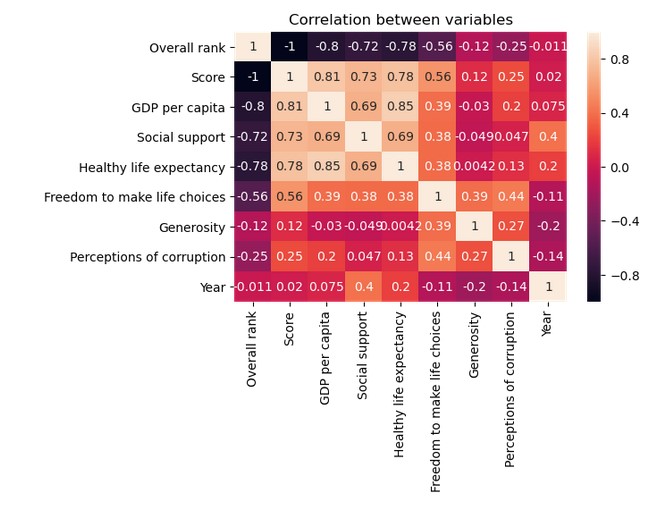
**OBSERVATION**

1. **SKEWNESS**
2. Most of the data distributions are slightly skewed, except for generosity and perceptions of corruption which are positively skewed.
3. **KURTOIS**
4. The distribution for all the variables are platkurtic as they have low kurtois. This is especially apparent for overall rank, and GPD per capita.

* We plot all features against happiness score



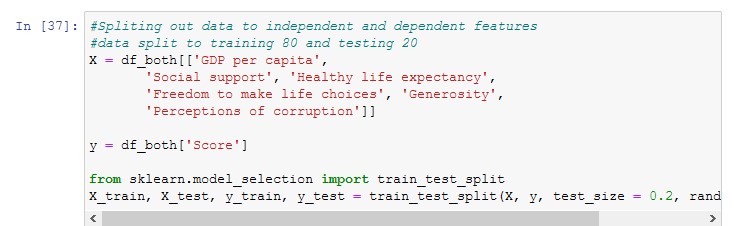
* Here we see the generosity and perception of corruption having weakest correlation as compared to GPD per capita and Healthy life expectancy which are having the strongest correlation. Is it also evident that all the features seem to have linear relationship with happiness score.



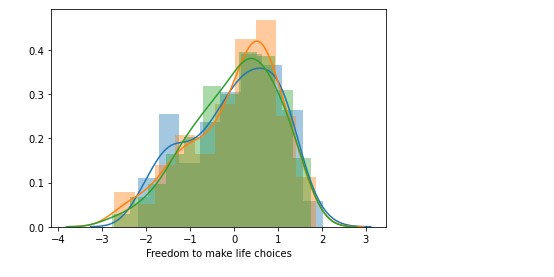
* Looking at the above diagram we can see that score here is highly correlated to GPD per capita and healthy life expectancy and weakly correlated to generosity and perceptions of corruption.

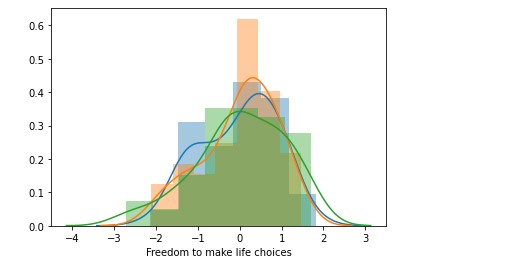
**MODELING**

* We drop unnecessary columns such as overall rank, country or region and year
* We use standard scaler that scale the data to a mean of 0 and standard deviation of 1.
* We split the data into test and training with test size of 20

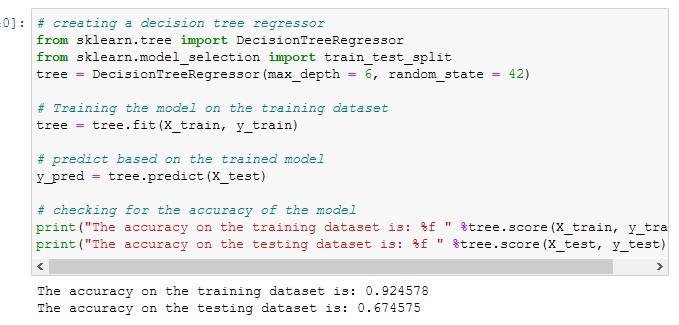
****

* **We visualize the distribution of train and test data after scaling**

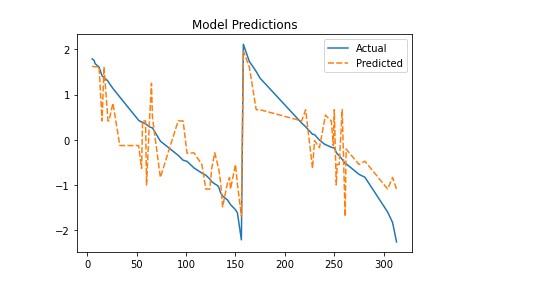
****

****

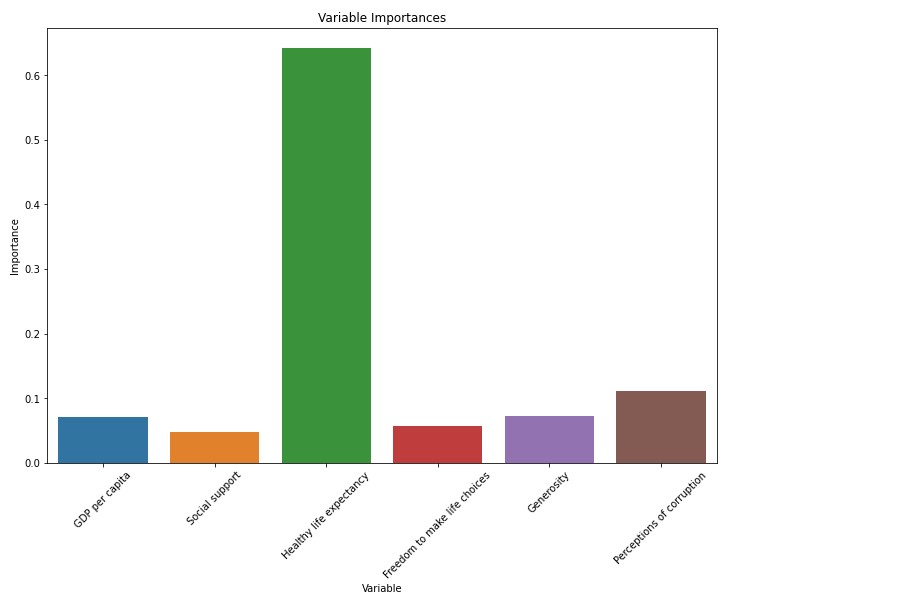
* **The distributions are similar now and closer to normal distribution.**
* **We generate a decision tree and fit it. We then predict based on the X\_test**
* **We output the accuracy on the training and test dataset.**

****

* **We compare the actual and predicted values**

****

* **We can see here that the model prediction accuracy is on point.**
* **We generate variable importance and plot it**

****

**Conclusion**

* **From our observation, we can see that the healthy life expectancy is the most important feature followed by perceptions of corruption and GPD per capita.**
* **We then say that for people to remain happy each national must be able to ensure that people’s lives come first. If people are healthy and can trust the government in dealing with corruption, then they remain happier**
* **The GDP per capita has some effect on people’s happiness but not so much.**

**THIS REPORT WAS WRITTEN BY : Suprice Sitoe**

