**Data Science Term Project**

PROJECT REPORT

*Submitted by*

**GANTI SURYA SAI ASHWIN BL.EN. U4CSE19043**

**K. CHAITANYA VADIRAJ BL.EN. U4CSE19058**

**KISHAN MANIYAR BL.EN. U4CSE19065**

***for the course***

***19CSE304 - Foundations of Data Science***



***Evaluated by:***

***Dr Kumaran***

**AMRITA SCHOOL OF ENGINEERING, BANGALORE**

**AMRITA VISHWA VIDHYAPEETHAM**

**BANGALORE - 560 035**

**Introduction**

Our purpose behind choosing the Happiness index dataset is to analyze various continents and their countries across the world. The inferences we obtained directed us to the regions where the people were satisfied to a great extent with their life in terms of social support, freedom in making life choices, generosity, life ladder, health life expectancy, perception of corruption. This is also reflected in the GDP per capita of the regions.

**Why did we choose these attributes?**

Every human seeks happiness in his/her life. We all expect our lives to work in a certain way and that satisfaction leads us to happiness. Our attributes focus on these basic expectations which when fulfilled lead to positiveness in humans

**Country names**: List of countries across the world whose data was collected.

The data consists of 149 countries recorded.

**Regional indicator**: The region to which the country belongs. There are 10 regions to which these countries are distributed.

**GDP per capita**: a country's GDP divided by its total population.

**Life ladder**: The person is asked to imagine his life in terms of a ladder and is asked to verify the satisfaction with his/her life.

**Social Support**: How well people around support each other during a crisis. This definitely builds confidence in a person, when he has people to support him.

**Generosity**: Willingness to share knowledge and other beneficial things to uplift society.

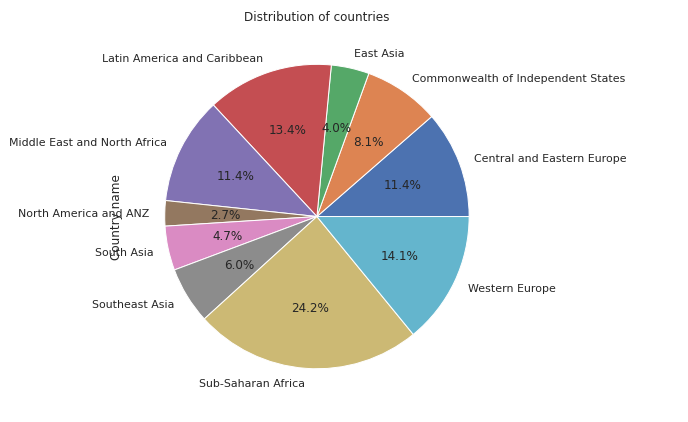
**Freedom of life choices**: How free do people feel in taking independent decisions for their life. This depicts the open-mindedness of the society

**Health life expectancy**: The average working age of a person. This describes motivation and purpose in life

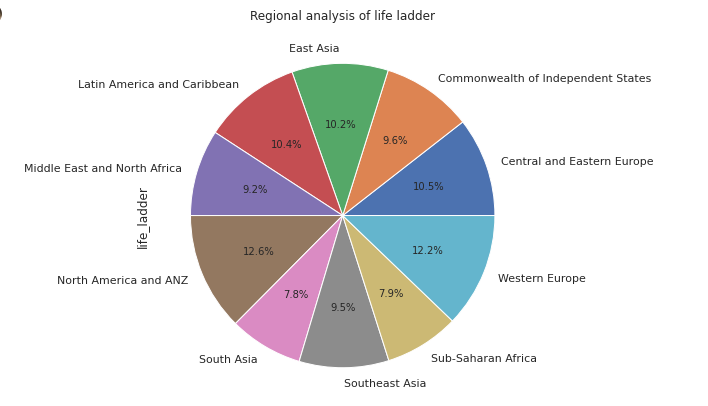
**Perception of Corruption**: Rate of exploitation by private sectors and government on the public

* We observe that  Sub-Saharan Africa has the highest number of

countries (24.2%) followed Western Europe, Latin America, and the Caribbean are having the next set of highest country counts (total-27.5%).

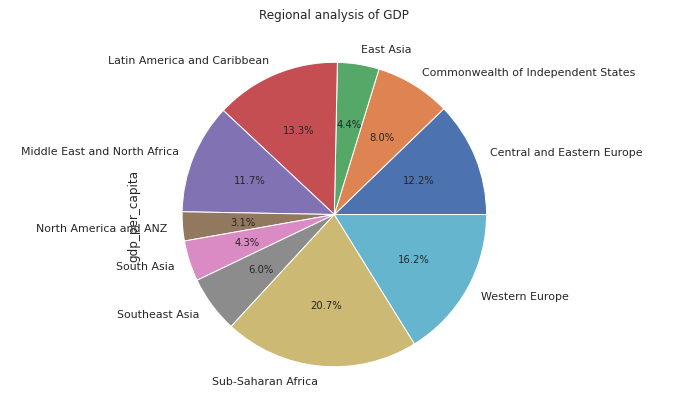


* Life ladder is a measure of the satisfaction of a person (on a scale of 1-10) The pie chart gives an understanding of how regions contribute to this measure throughout the world.

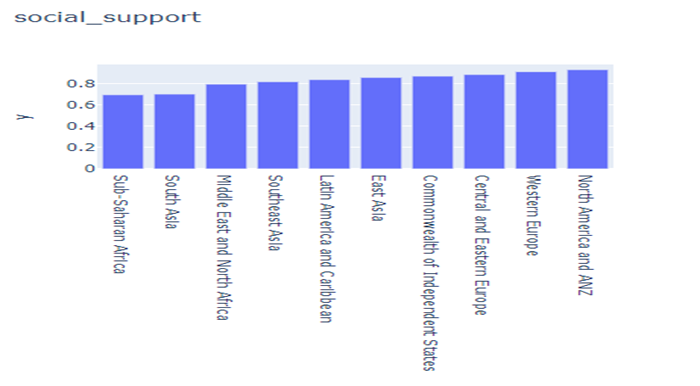


* Gross domestic product (GDP) is the standard measure of the value-added created through the production of goods and services in a country during a certain period

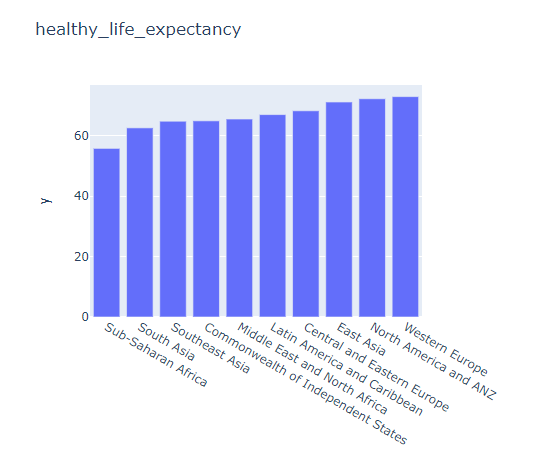
From the below Pie chart, Sub-Saharan Africa and Western Europe contribute to gross GDP per capita with around 37%. From previous studies, it is observed that these regions have a 62% higher country count compared to other regions in the world. Also, the cumulative life ladder shares up to 20% in the world.



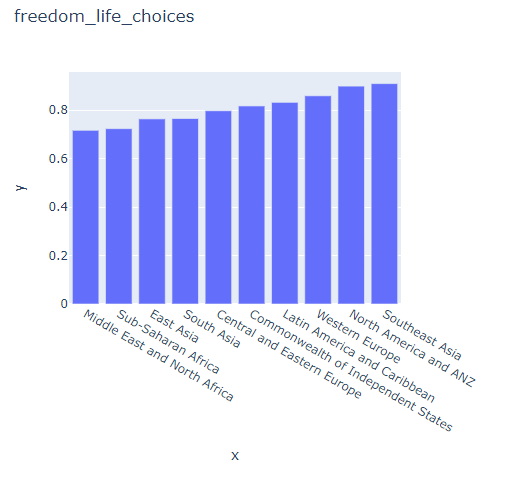
* **Social Support:**
  + The people of North America and ANZ, Western Europe feel the most supportive socially.
  + Sub-Saharan Africa and South Asia feel the least supported among the public.



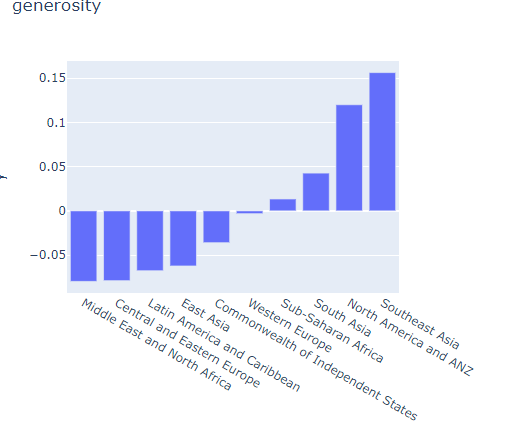
* **Healthy Life Expectancy:**
  + The people of North America and ANZ, Western Europe remain healthier for a longer duration in their life. As a result, their total life ladder contributes to 25% of the world. Also, on studying the correlation plot, we infer that the correlation between the two features is around 0.9 which explains the fact higher the working years of a person the more satisfied he is.
  + Sub-Saharan Africa has the least health life expectancy.



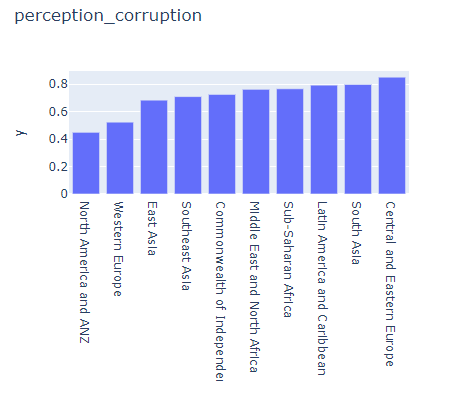
* **Freedom of life choices:**
* The people of North America and ANZ and Southeast Asia are more independent in making their life decisions which further reflects in the region’s confidence and social as well as financial success.
* Sub-Saharan Africa, the Middle East, and North Africa do not experience the same freedom of independent decision-making.



* **Generosity:**
  + Southeast Asia, North America, and ANZ dedicate their time and money to add value to the needful people.
  + Southeast Asia, North America, and ANZ dedicate their time and money to add value to the needful people.
  + The generosity measure of the Middle East and North Africa, Central and Eastern Europe are the lowest. This explains that the willingness to add value to society in any possible form is not encouraged in these regions

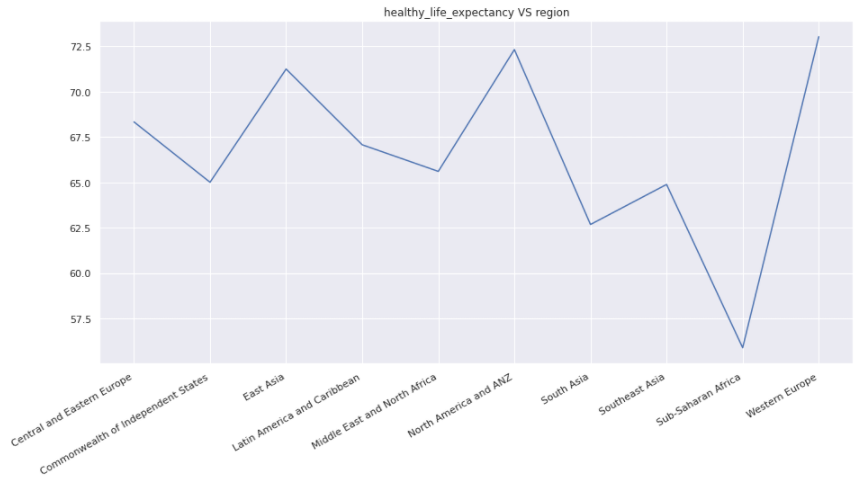


* **Perception of Corruption:**
  + As per the data The Central and Eastern Europe, South Asia are the most corrupted in terms of politics as well as private exploitation.
  + Whereas regions like North America and ANZ, Western Europe have controlled their corruption levels which will definitely reflect in freedom and growth of the people.

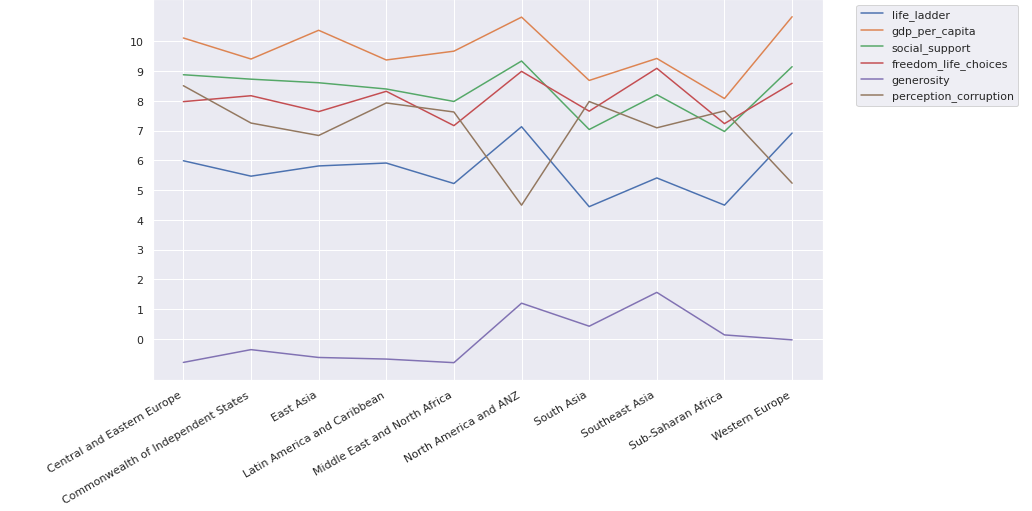


* **Scaling**

We scale certain features in our data set for more inferrable data visualization. Social\_support, Freedom\_life\_choices, generosity, perception\_corruption are the features scaled by 10 units. We have then plotted a line graph health life expectancy VS Region in order to see the scaled values across different regions



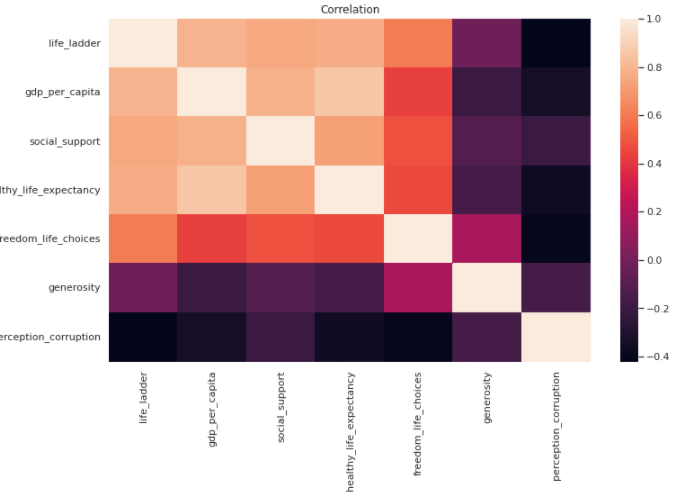
We then added multiple line graphs with scaled values across the region which On analyzing, we can observe Life ladder, GDP per capita, social support, generosity, and freedom of life choices go hand in hand i.e. the values are highly correlated. The reason behind the correlation lies in the fact that these features measure positive temperaments in the population. The satisfaction of people (life ladder) living in the region is an outcome of how socially supported and independent they are. Also higher the healthy life expectancy, the more they have been healthy towards their life which will reflect in the GDP of the country as well as the life ladder. The generosity .i.e the kindness and willingness measure also reflects in the GDP measures. Also, if the GDP is high, there is a higher chance of people being generous as they have money to give and open-mindedness. We see a huge dip in the perception of corruption line for ANZ, as it is a negative measure, the less corrupt the region is, the higher the positive numbers.



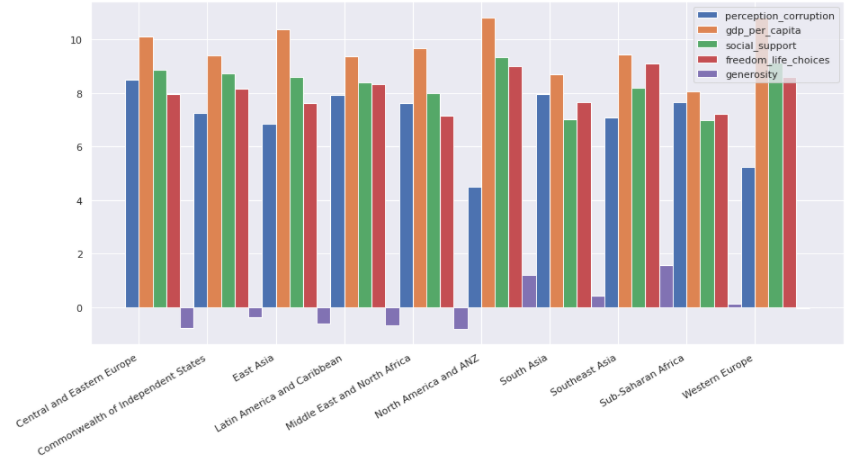
* **Correlation :**

In this case, we can observe a negative correlation which means that in the relationship between two variables, higher values of one variable tend to be associated with lower values of the other.

On analyzing the heat map when the perception of corruption is in between 0 - -0.4 the other parameters are lowering which indicates lesser satisfaction(happiness) of the countries. When perception corruption reduces the related parameters see an upward growth, thus there is an increase in positiveness and satisfaction in the country.



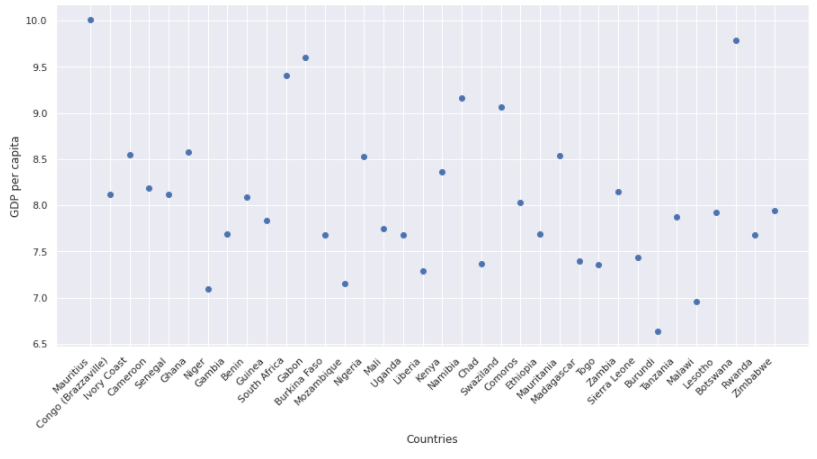
* Checking for which region has the highest mean in terms of all the parameters and mean is calculated within each region for each parameter and bar plot has been plotted.

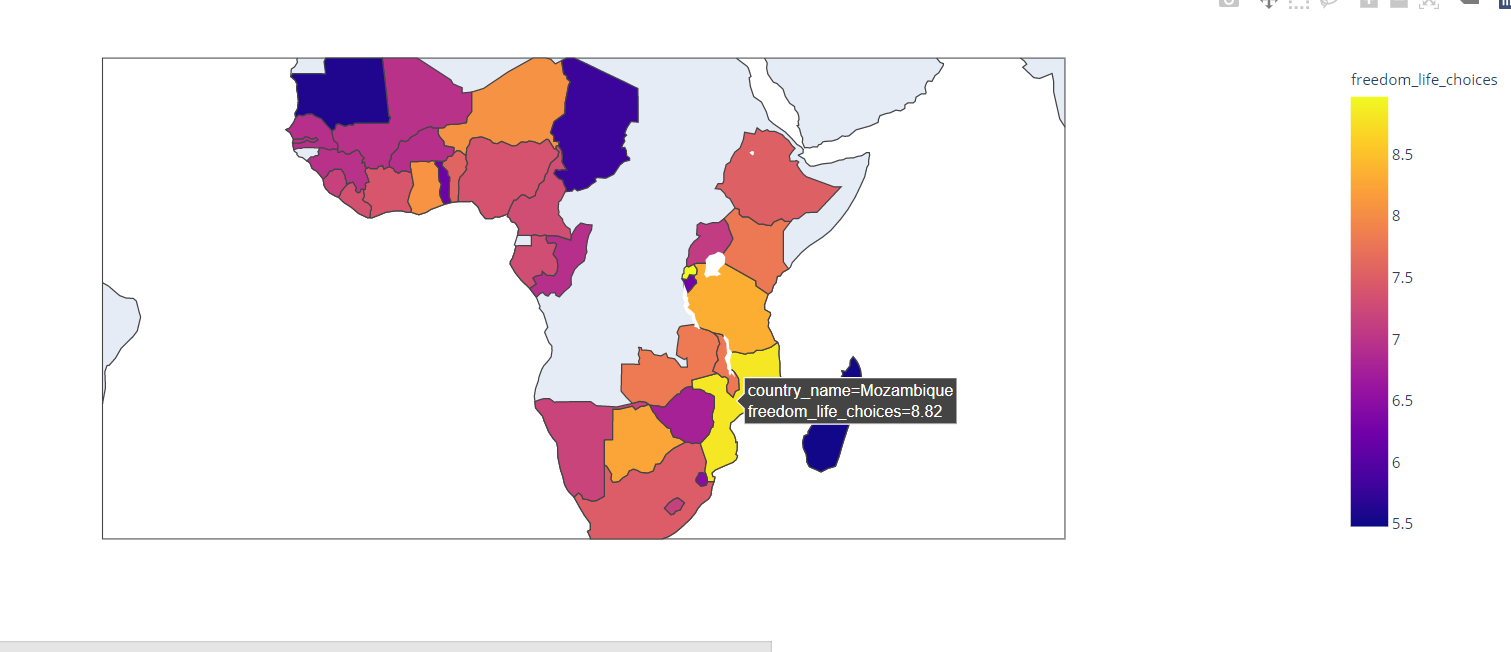


As we can see from the above plot we can conclude that the ANZ is the best performing region in terms of the overall high mean value and that Sub Saharan Africa and Western-Europe underperforms so we analyze the region and infer.

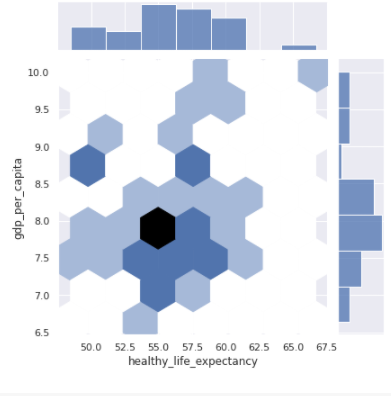
Here we are considering Sub-Saharan And African regions. We can see from the previous plots that the Sub-Saharan And African region is underperforming. So, now we are trying to analyze each country and its behavior with other attributes.

From the below plot we can see that the GDP of Sub Saharan Africa is high but in the previous statement, we have said it underperforms because GDP is not the only parameter but we also have to see the amount of social support, freedom, and how generous people are to one another which increases the morale which results in satisfying the individual.



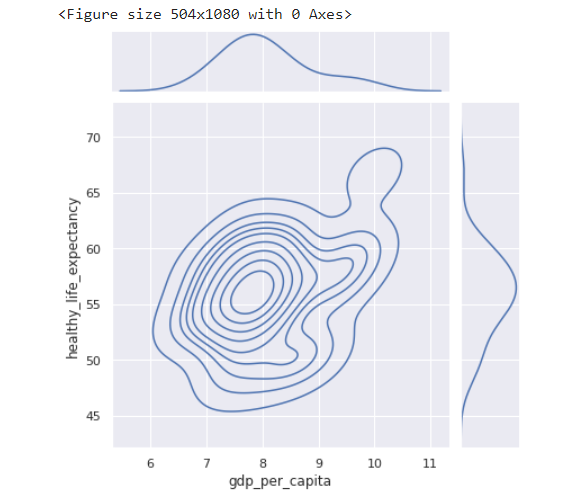


The hex bin plot depicts the region where most data points are crowded. In this plot, the average healthy life expectancy is 55. Most countries who lie in this region have a GDP of 8, less than the world average GDP of 9.4, which explains why life expectancy is expected to perform better in these countries. Hence these countries have to work on maintaining the health of people and keep them motivated to work for a longer span

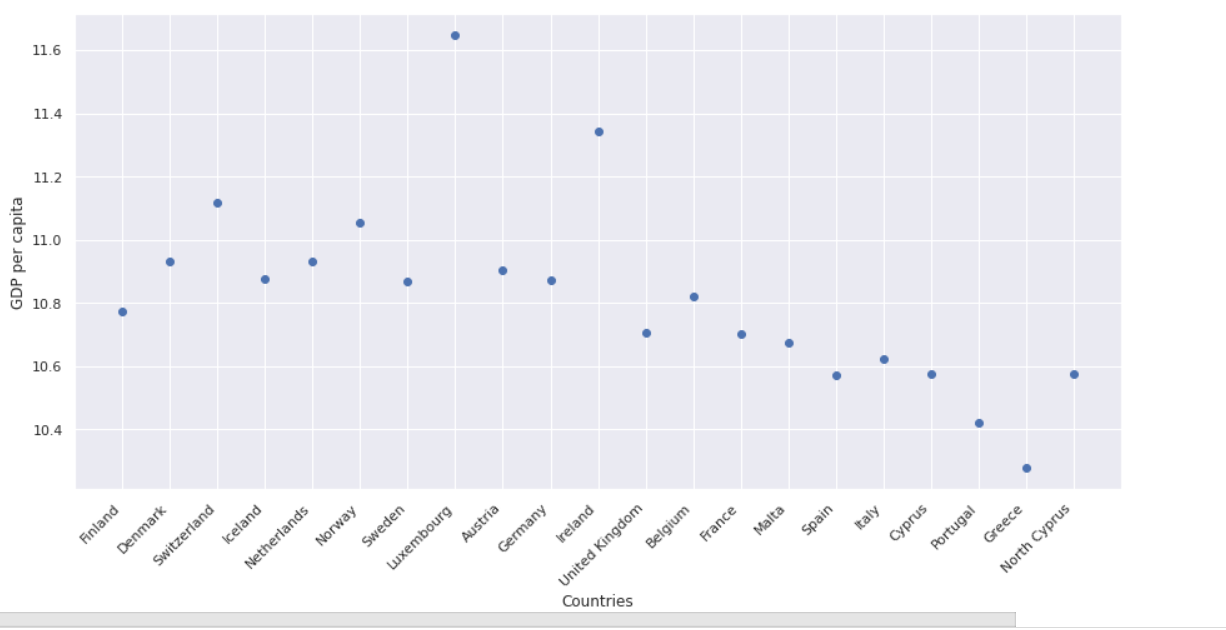


The KDE plot density plot explains the spread of data as well as the outliers. The plot depicts that as healthy life expectancy gets higher, the GDP of the country also improves. There is a high correlation between these attributes.

The highest is 10.5(high performance) GDP as the healthy life expectancy is around 67.

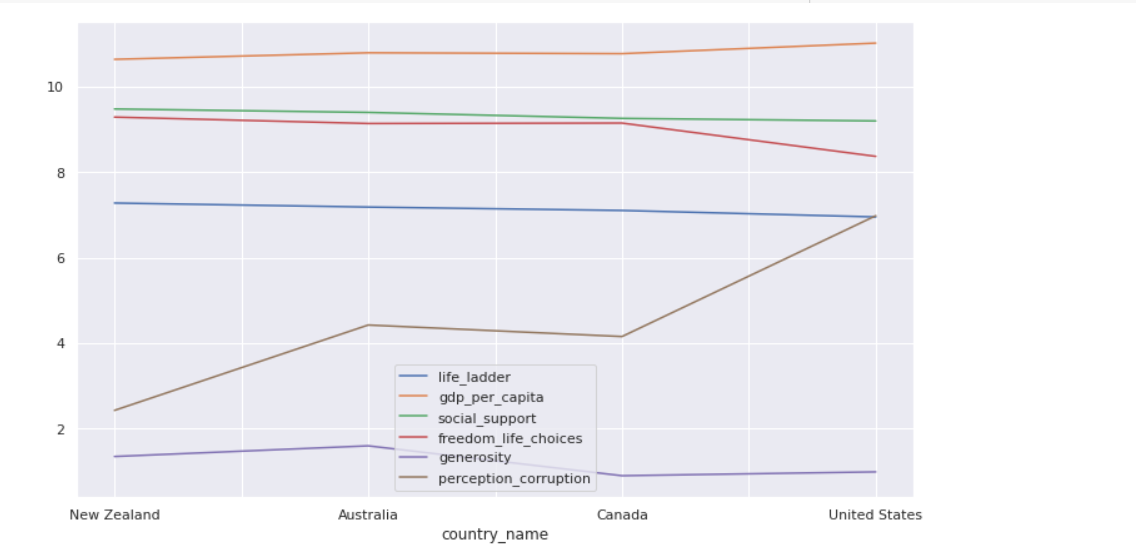


Here we are considering Western Europe region. We can see from the previous plots that the Western Europe region is underperforming. So, now we are trying to analyze each country and its behavior with other attributes.



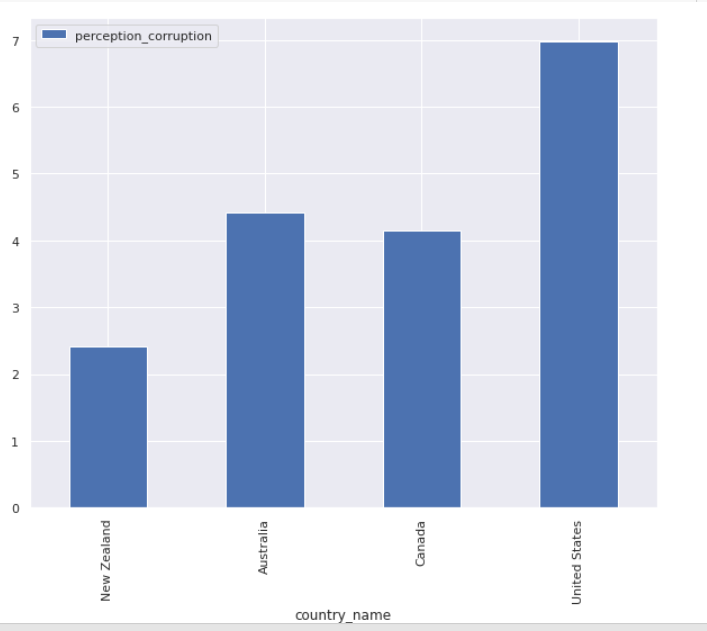
Here from the Western Europe region, we have plotted a scatter plot between the GDP of each country and the Name of the countries. From the scatter plot we can see that Luxembourg has the highest GDP value which means that this particular country is well developed and followed by Ireland.

Here we are considering American and ANZ regions. We can see from the previous plots that the American and ANZ region has a higher performance. So, now we are trying to analyze each country and its behavior with other attributes.



As we have explained before, the line plot here as well explains the positive correlation of the attributes specific to the countries of ANZ. As we see all have performed very well.

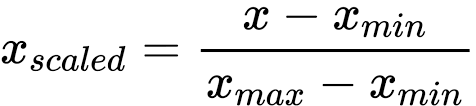
In the North America and ANZ region, we have only had 4 countries and from the bar graph we can infer that New Zealand has the least corruption rate and the United States has the highest Corruption rate.



**Machine Learning Models:**

Country\_mean attribute gives us a measure of the mean of all the attributes (on scaling using min-max scaler) per country. The higher the mean, the better performing the country is.

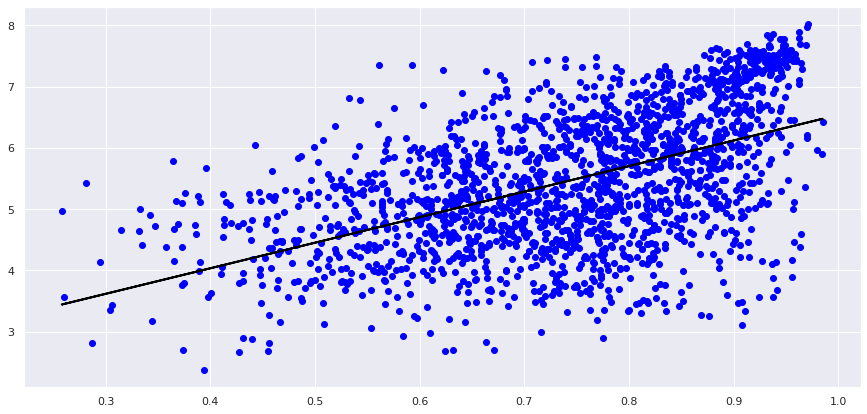
MinMaxScaler(): This estimator scales and translates each feature individually such that it is in the given range on the training set, By-default between zero and one.

****

1. Linear Regression:

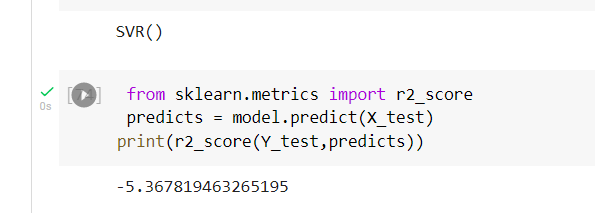
Linear regression is a machine learning model to estimate the relationships between a dependent variable and one or more independent variables.

We choose this model to train our complete dataset and predict a regression line that indicates the world average performance. It predicts the growth in character and the satisfaction of people around the world. The regression line is drawn across a scatter plot of life ladder v/s freedom of life choices estimates a positive yet gradual movement of improvement in the way people think, in the coming future.

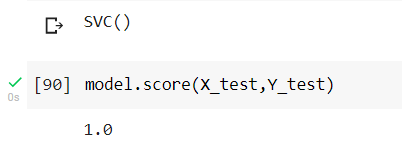


1. **Support Vector Machines :**

Support Vector Machine is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. SVM chooses the extreme points/vectors that help in creating the hyperplane. Our dataset has continuous values so we really can’t use classification. So we prefer using a regressor which works well with regression problems.



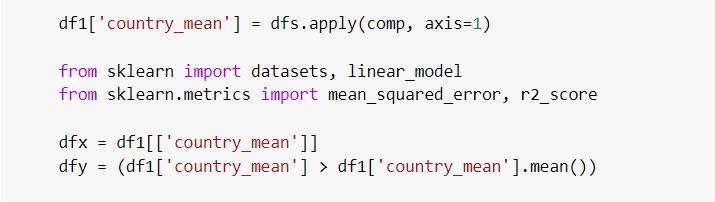
For the regressor when we try to print the r2\_score value we get a negative value. This tells us that using an SVM is not the correct model for our dataset.



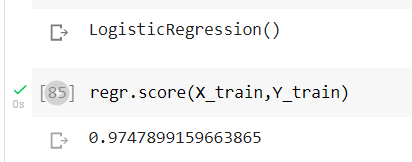
Even when we convert our dataset into a binary classification problem we get an overfitting problem. A model is said to be overfitted when we train it with a lot of continuous data. When a model gets trained with so much data, it starts learning from the noise and inaccurate data in our data set.

1. **Logistic Regression:**

Logistic Regression is a Machine Learning algorithm that is used for classification problems, it is a predictive analysis algorithm and based on the concept of probability. The major issue with our data values was that they were all continuous and they didn’t meet the classification criteria. So we first had to get the data as a classification problem. For doing this, we first added another feature to our dataset which includes the values of the mean of all other features in the dataset. We named this feature as country\_mean. To counter this in the y value of our model we added a binary classification condition.



If the value of the country\_mean is a positive high value we can conclude by saying the countries are much happier than the other countries which have negative low values.



On training our dataset, we get an accuracy of 97.4 %.

**Conclusion:**

Our analysis brings us to the conclusion that there are various attributes that contribute to the happiness of a countries people that ranges from

1. How the person thinks about himself/herself
2. How society treats him/her and social exploitation.
3. How the government works towards the upliftment of peoples’ way of life

All our attributes assess these points precisely. Hence directly indicating happiness measures of various parts of the world as well as a wholesome picture of the future.