**Project 3: World Happiness Report**

**Dataset Description:**

We have two datasets for the World Happiness Report for the years 2015 and 2016. These were collected from Kaggle Website. The data was essentially obtained from Gallup World Poll. Survey Responses to various factors contributing to the happiness of a country are collected. These factors are measured in a scale of 1-10.

The various factors include :

* **GDP**: GDP per capita indicates the average purchasing power parity of each country. This data was collected from the World Development Indicators(WDI) as released by World Bank in August 2016.
* **Family**: National average of responses to ‘whether you have a friend or family to help you when in trouble’. It explains about the general Social support that is prevalent to individuals in a country.
* **Health(life expectancy):** This factor explains about the average life expectancy of individuals of a nation, based on the data collected from World Health Organization(WHO).
* **Freedom :** Freedom to make life choices i.e., satisfaction level of your freedom to do something in your life. This factor is measured based on responses from people of each country.
* **Trust(Corruption):** Score based on average responses to whether corruption exists in the government and businesses of a country is recorded.
* **Generosity:** Average responses regarding money donation to a charity is recorded as per GDP per capita.
* **Dystopia:** Hypothetical country(lowest national average for six factors)

**Goals**:

* Visualize the data- World Happiness data distribution for the years 2015 and 2016.
* Identify correlation between the six factors contributing to the Happiness Score.
* Predict the happiness score for the year 2017 based on previous years and also plot the distribution
* Gather Insights based on the predicted values.

**Analysis:**

* **Collection of latitude, longitude information - Web scraping:**

We wanted to plot the World Happiness scores for each country on a world map. Hence, we needed latitude and longitude information of each country. To obtain this, we used *Beautiful Soup* module to do web scraping and collect the respective latitude and longitude values for the respective countries.

Attached below is the code the scrape the latitude and longitude values:



After scraping the data, we created a new excel sheet to record the *Country Name, Happiness Score, Latitude and Longitude* information for the years 2015 and 2016 as attached below:



* **Generating world map – Geo visualization(Basemap):**

We have used panda’s data frame to read the above excel sheets. We have also made use of Python’s Basemap module to generate the world map and therefore plot the Happiness Score indicator on countries based on the scale below.

Scale: Eyeballing the happiness score of countries, we have come up with a criterion that classifies a country as below.

Green dots indicate the countries that have *High* Happiness scores, Blue dots indicate the countries that have *Moderate* Happiness scores while Red dots indicate the countries that have *Low* Happiness scores.

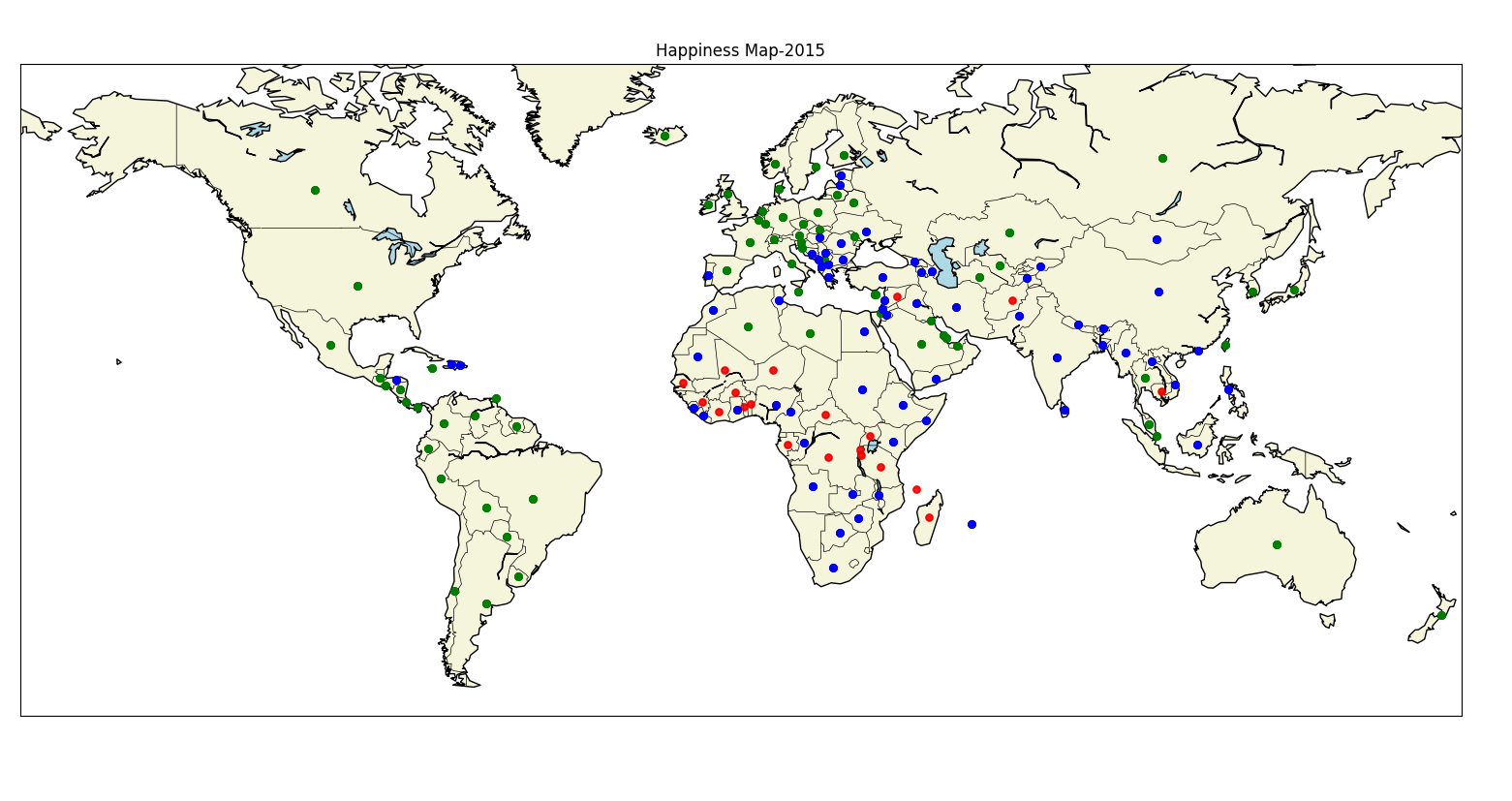


Please find the code attached below to generate these world maps plotted based on Happiness scores.

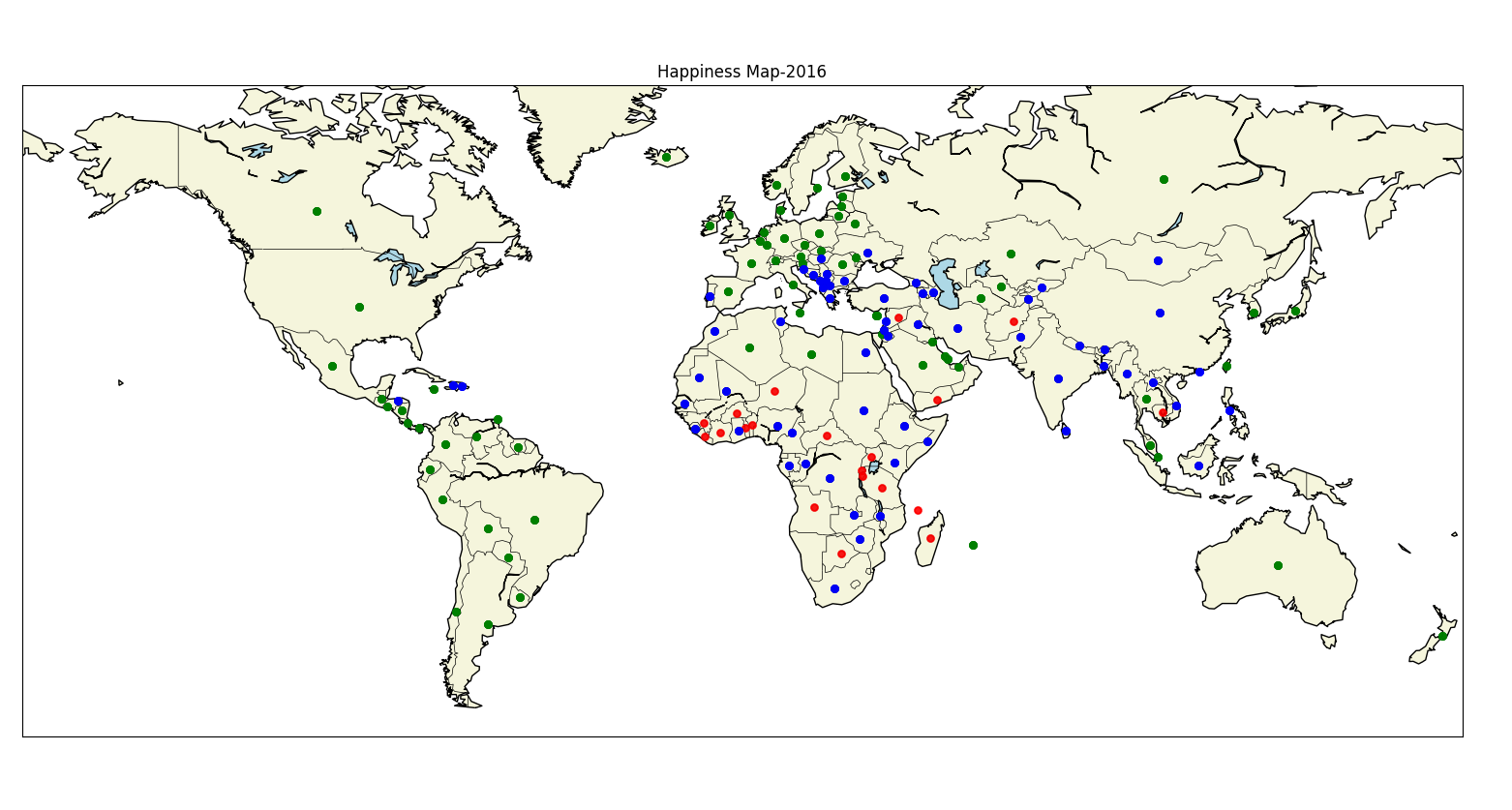


Hence, below are the plots for the years 2015 and 2016.

**2015:**



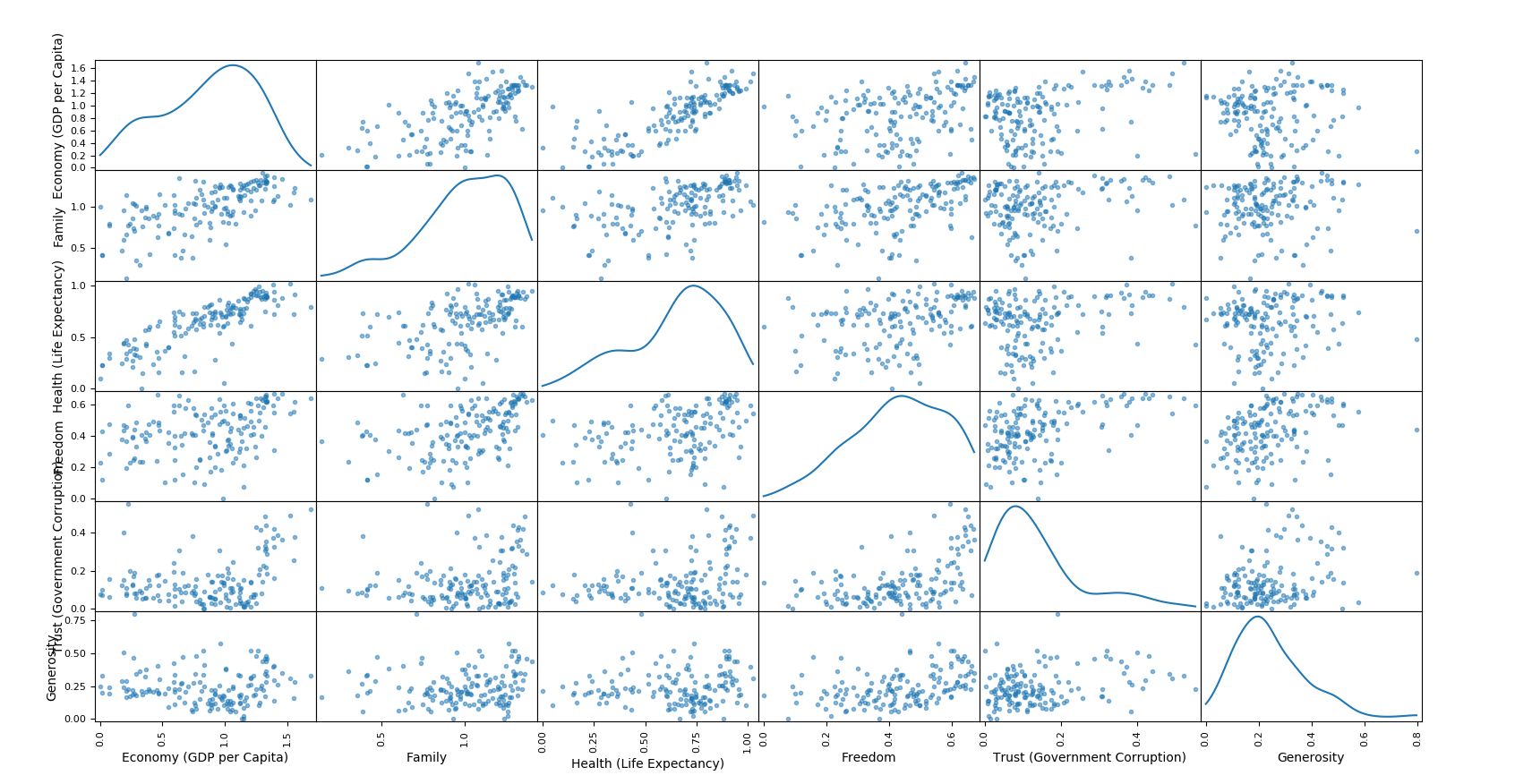
**2016:**

****

**Inferences based on 2015 and 2016 plots:**

* It is observed that most countries in both North America and South America have higher happiness scores. Also, most countries in Europe, Australia exhibit higher happiness scores.
* Countries in Asia and South-Asia Pacific regions exhibit moderate happiness scores
* Most countries in Africa show Low happiness scores. It has also a considerable number of countries with moderate happiness scores.
* **Correlation among the factors:**

1. We have considered all the 6 factors that include GDP, Family, Health(life expectancy), Freedom, Trust(Corruption) and Generosity.
2. We are checking if there are any correlation between these six factors by implementing Panda’s Scatter Matrix plotting module. Below is the scatter matrix plot.



1. The six factors in the regression are analyzed for multicollinearity and from the scatter matrix, we could not find any strong correlation between the 6 factors.
2. However, it is observed that there is a mild positive correlation between GDP and Health and an even lesser positive correlation between GDP and Family.

Also attached below is the code to generate the above Correlation Scatter matrix:



* **Prediction Model – Scikit learn(Regression):**

From the correlation plot, we were unable to find strong correlations between any of the factors. Hence, we were considering regression approaches to predict the happiness score of the year 2017 and then plot in on the world map to gain insights about the happiness scores of different countries.

Since we observed that there is no multicollinearity between any of the factors, we concluded that Linear regression would be the ideal fit for our prediction.

The average values of each of these six factors were considered for predicting the happiness score of the year 2017. It was observed that the prediction results were quite close to the 2016 data.

Attached below is the code to predict the Happiness score for 2017:

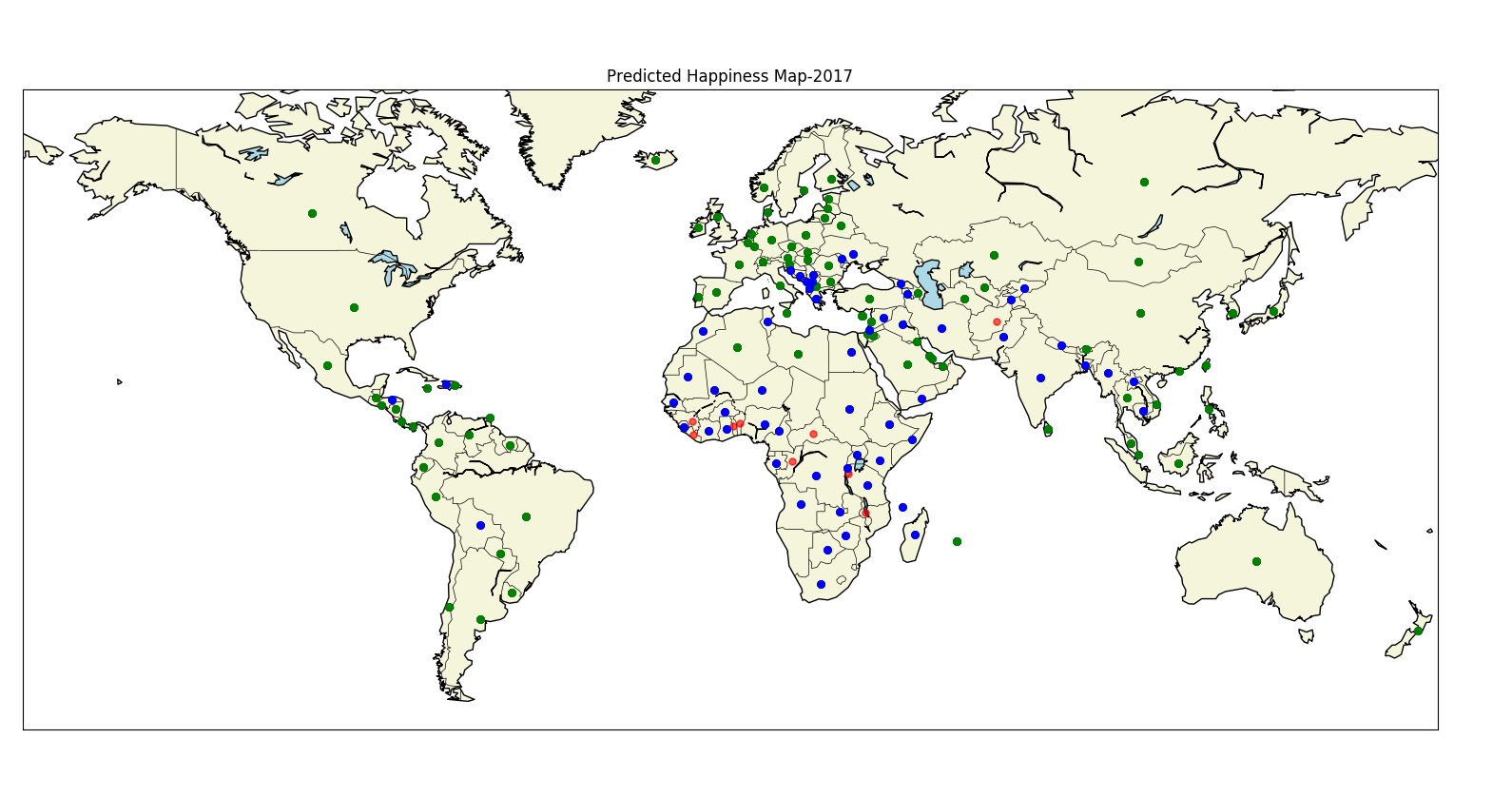


Below is the excel sheet attached containing Country Name, Latitude, Longitude and Happiness Score data for the year 2017.



Again, we have used Panda’s data frame to read the above excel sheet and therefore plot the Happiness Score indicators on countries using Python’s *Basemap* module.

Here is the plotted map based on Predicted World Happiness scores for the year **2017:**



From the graph, it was observed that certain countries such as Mexico in North America, Niger and some parts of Congo of African continent has reduced in happiness scores while countries like Bangladesh, Sri Lanka in the Indian subcontinent have moved up the ladder showing some increase in happiness scores.

**References:**

* <https://discuss.analyticsvidhya.com/t/comparison-between-ridge-linear-and-lasso-regression/8213/3>
* [http://stackoverflow.com](http://stackoverflow.com/)
* <https://www.kaggle.com/unsdsn/world-happiness>
* <https://developers.google.com/public-data/docs/canonical/countries_csv>