iiit-b
ज्ञानमुत्तमम्
International
Institute of Information
Technology Bangalore

UpGrad

PCA & Clustering Assignment

NIKUNJ PATEL

ABSTRACT

To categorise the countries using socio-economic and health factors that determine the overall development of the country.

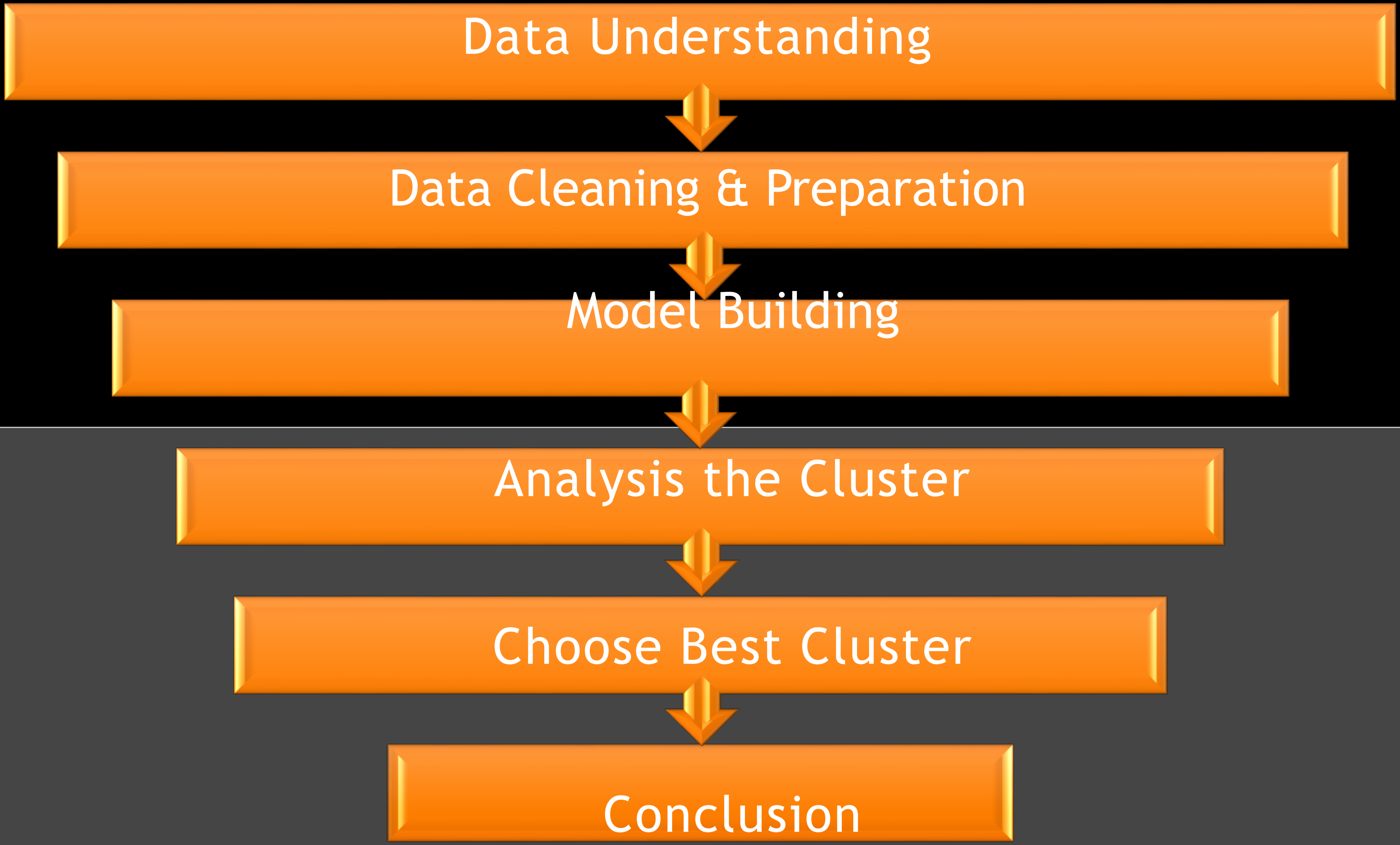
About organization:

HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities.

PROBLEM STATMENT

HELP International have been able to raise around $ 10 million. Now the CEO of the NGO needs to decide how to use this money strategically and effectively. So, CEO has to make decision to choose the countries that are in the direst need of aid. Hence, My job as a Data analyst is to categorise the countries using some socio-economic and health factors that determine the overall development of the country. Then I need to suggest the countries which the CEO needs to focus on the most.

PROBLEM SOLVING METHODOLOGY



Data Understanding

Data Cleaning & Preparation

Model Building

Analysis the Cluster

Choose Best Cluster

Conclusion

Analysis Approch

Step 1 : Import libraries and load the data

Step 2 : Inspect the Dataset

Step 3 : Data Preparation

3.1 : Scaling the Data

3.2 : Perform PCA (We choose 5 components Because most of 98% information fall in 5 components

3.3 : Checking Outlier Based on PCA New Dataset

Step 4 : Perform K-means Clustering

4.1 : Perform the Hopekins Anakysis

4.2 : Finding the optimal number of Cluster using Two method

4.2.1 : Using SSD

4.2.2 ; Using Elow Curve (Based on Result we decide to choose k = 3)

4.3 : Perform Silhouette Analysis

Step 5 : Merge the Original Dats with Cluster Data

Step 6 : Analysis of K-Means

6.1 : Find the mean and concate the result

6.2 : Create Binning to find out Developed ,Developing,Poor Country

6.3 : Based on Binning We find the Poor country

Analysis Approch

Step 7 : Perform Hierarchical Clustering

- 7.1 : Perform Single & Complete Linkage
- 7.2 : Cut-Tree Based on Dendograms
- 7.3 Analysis the Hierarchical clustering

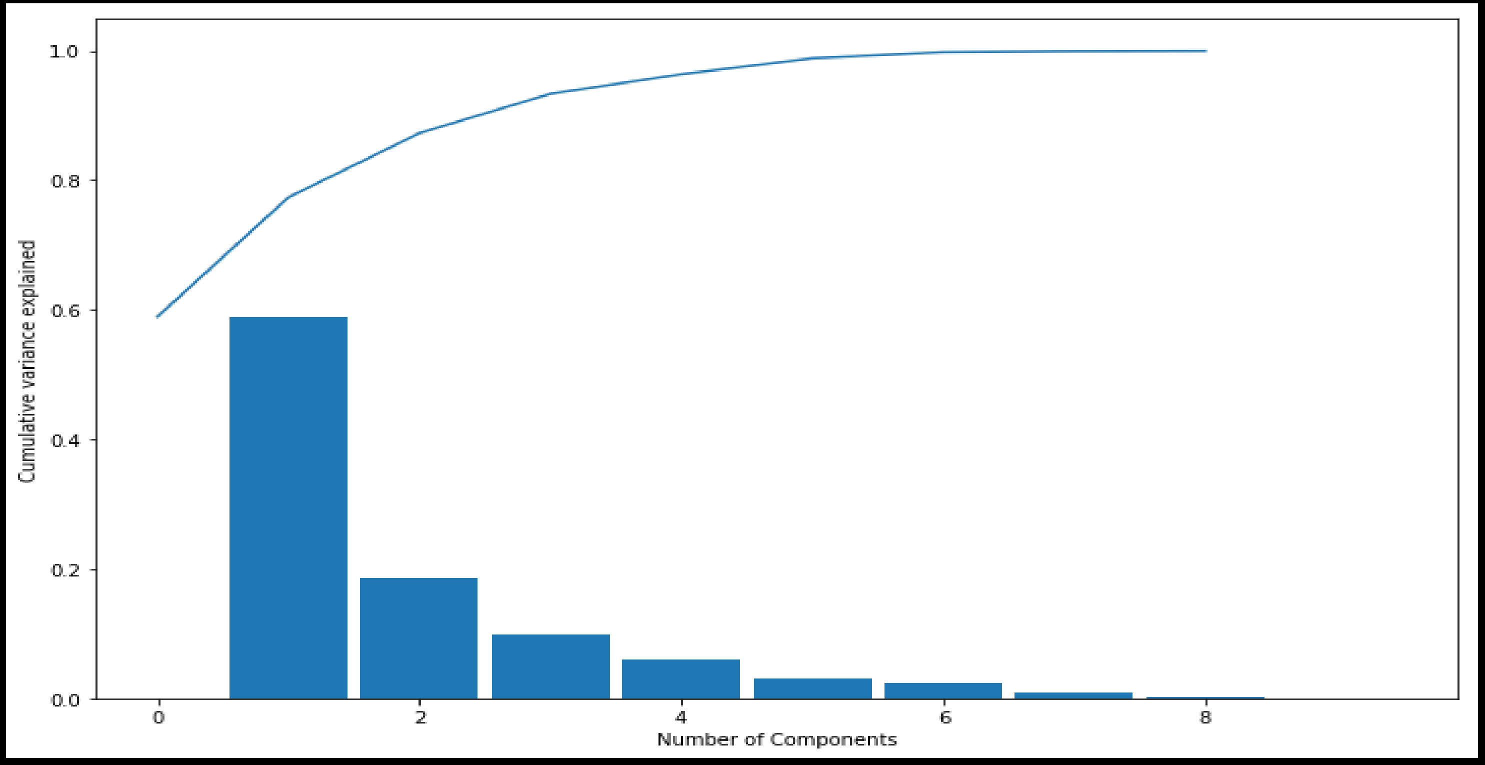
Step 8 : Conclusion :

Results of Principal Component Analysis

PCA will reduce the number of features from the number of Country Dataset, then numbr of principal components I have chosen.

After Perform the PCA on the country Dataset and making the BAR PLOT with Scree PLOT **Around 98% of the information is being explained through 5 components.**

So We will Analysis further using the PC1,PC2,PC3,PC4,PC5

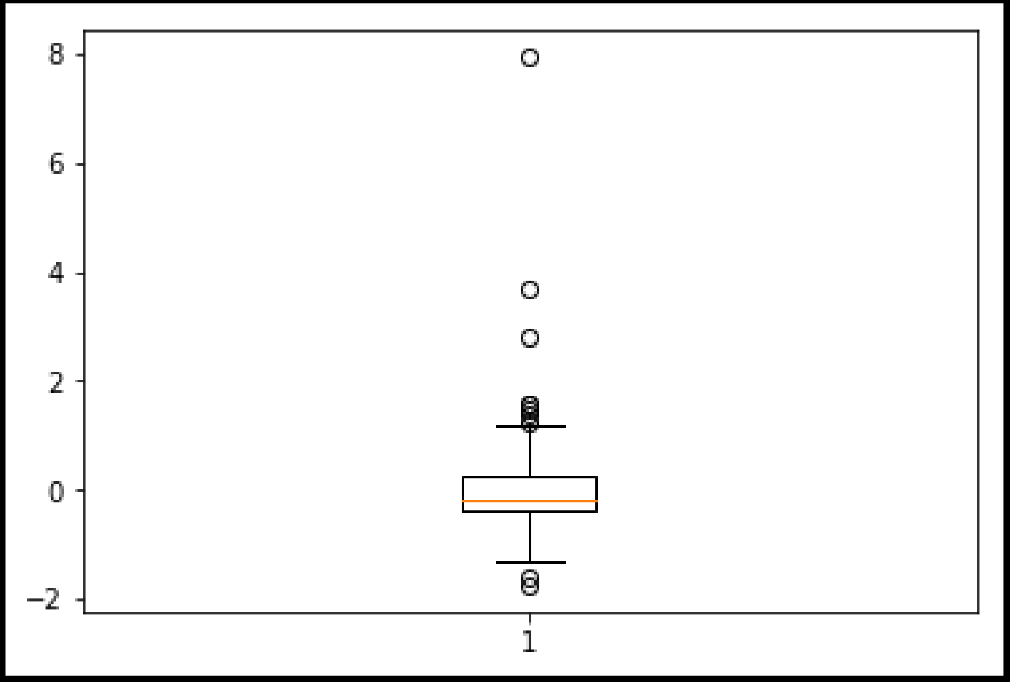


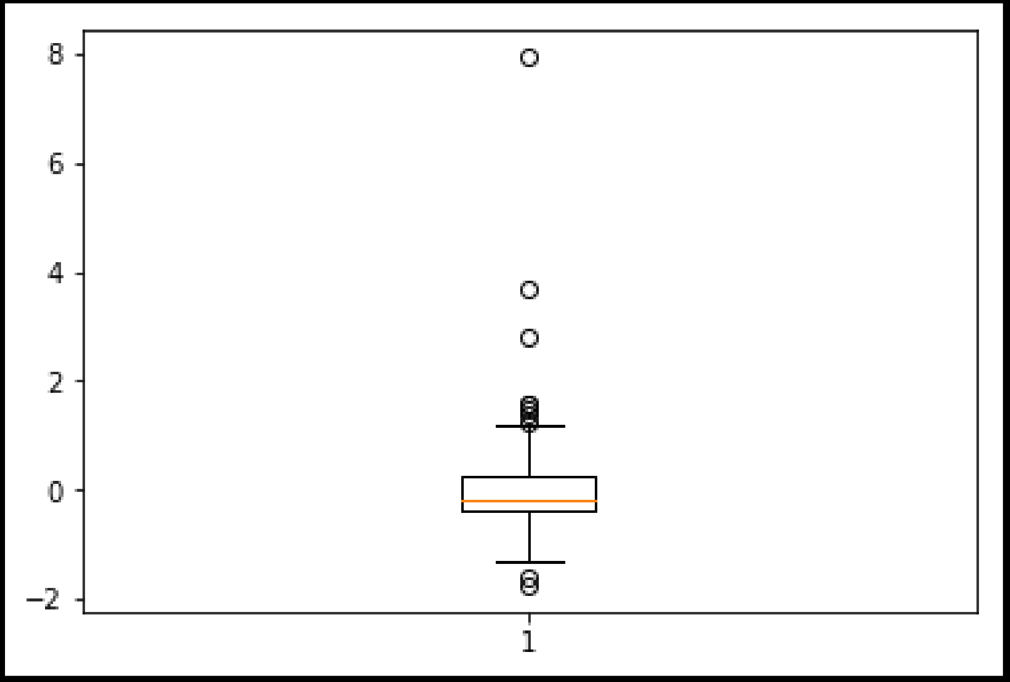
Outlier Tretment

After Perform the PCA and choose 5 Components We will check the outlier in the New PCA dataset.

1) As per result there is some Outlier in the Data Set and We did not treat them Because of there is no large number of outlier and if we treat the outlier we loose the information which is required for analysis in clustering.

2) So we decide to carry outlier and analysis the clustering.Below is the image which show the BoxPlot with outlier.





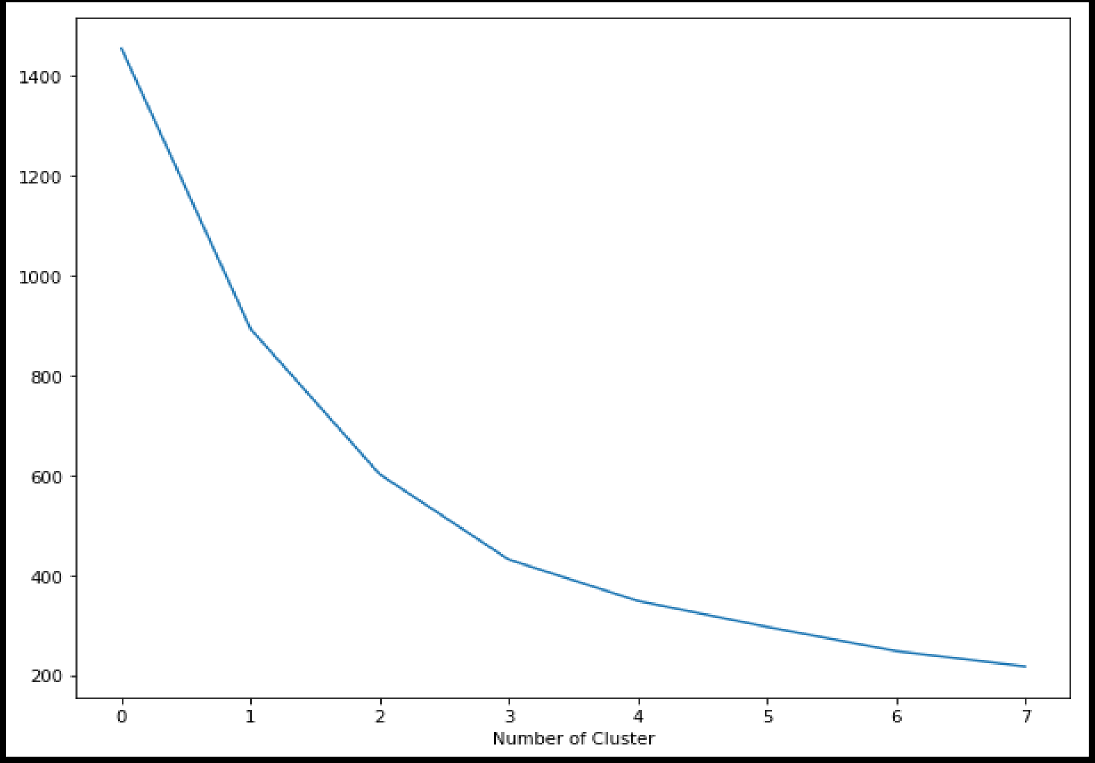
Cluster Analysis

1) As we perform the K-means Clustering on the PCA dataset first we perform the HOPKINS STATISTIC and We get 0.92.

2) Then we find the optimal number of curve using the elow cure and we decide to choose k =3 because 95% information fall in k =3

3) Then We also check the Silhouette Analysis

4) And then we will fit the Cluster id to PCA & observe most of Data fall in Cluster 1.



Obervation of K-means Clustering

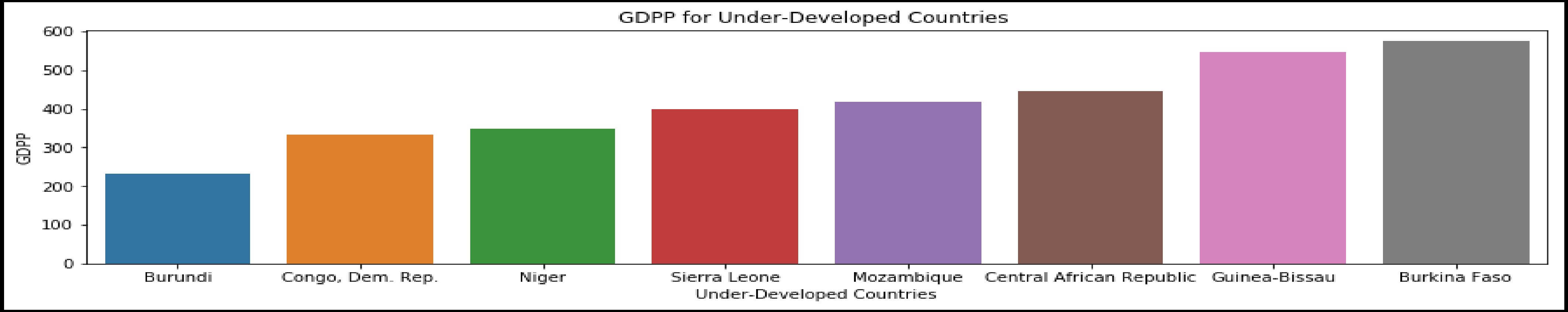
Analyse the clusters by comparing how these three variables - [gdpp, child_mort and income] vary for each cluster of countries to recognise and differentiate the clusters of developed countries from the clusters of under-developed countries.

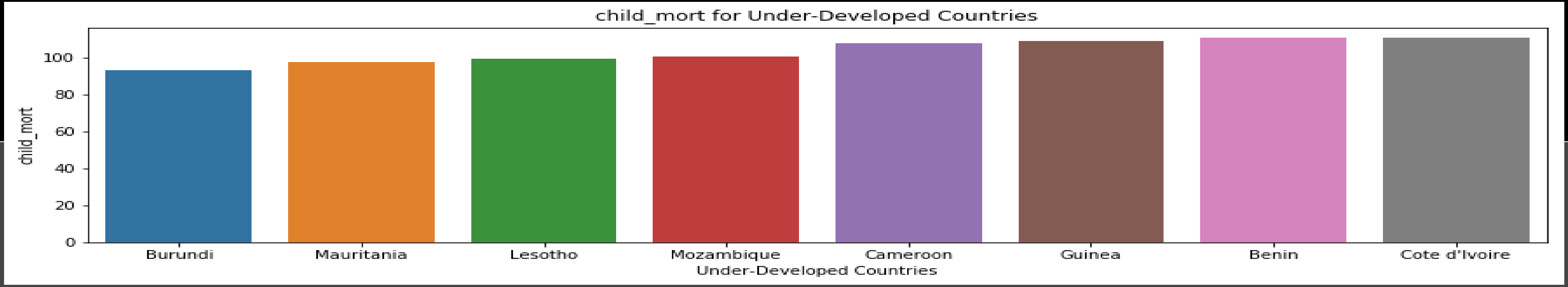
0 for 'Developed Countries'
1 for 'Developing Countries'
2 for 'Under-developed Countries'

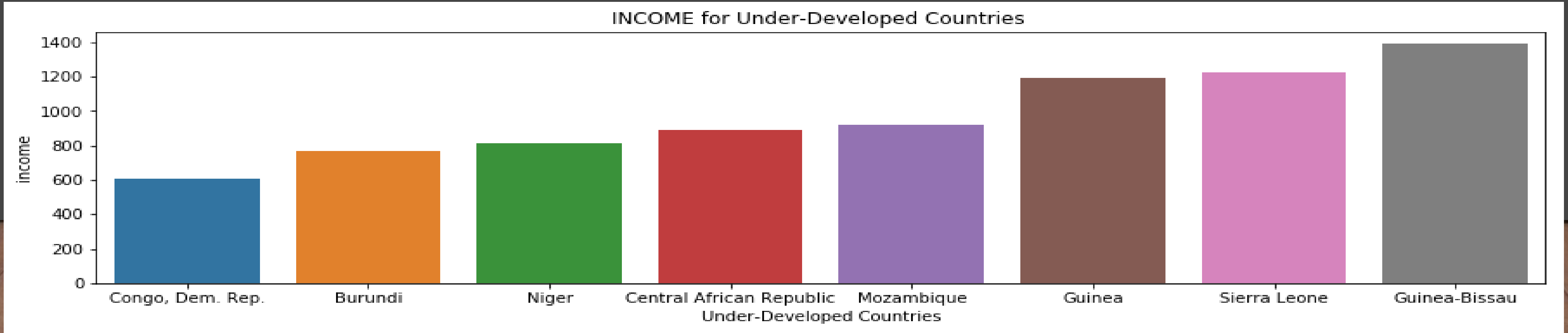
As wee see on plot Under Developed Country Based on GDPP , INCOME , CHILD_MORT IS

1) Burundi
2) Congo , Dem,REP
3) Niger
4) Sierra Leone
5. Central African Republic
6. Mozambique
7. Guinea-Bissau
8. Burkina Faso

PLOT FOR GDPP , INCOME,CHILD_MORT FOR UNDER-DEVELOPED COUNTRY

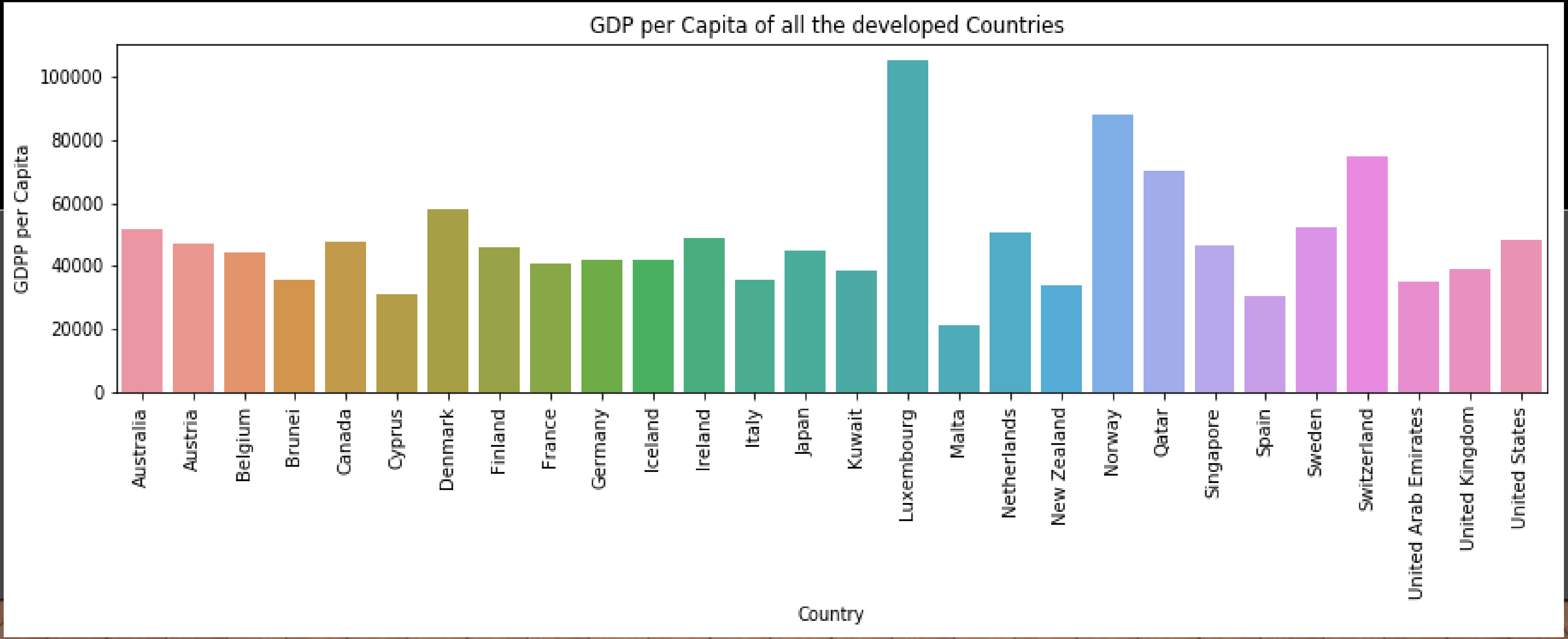






Developed , Developing , Poor Countris Plot based on GDPP

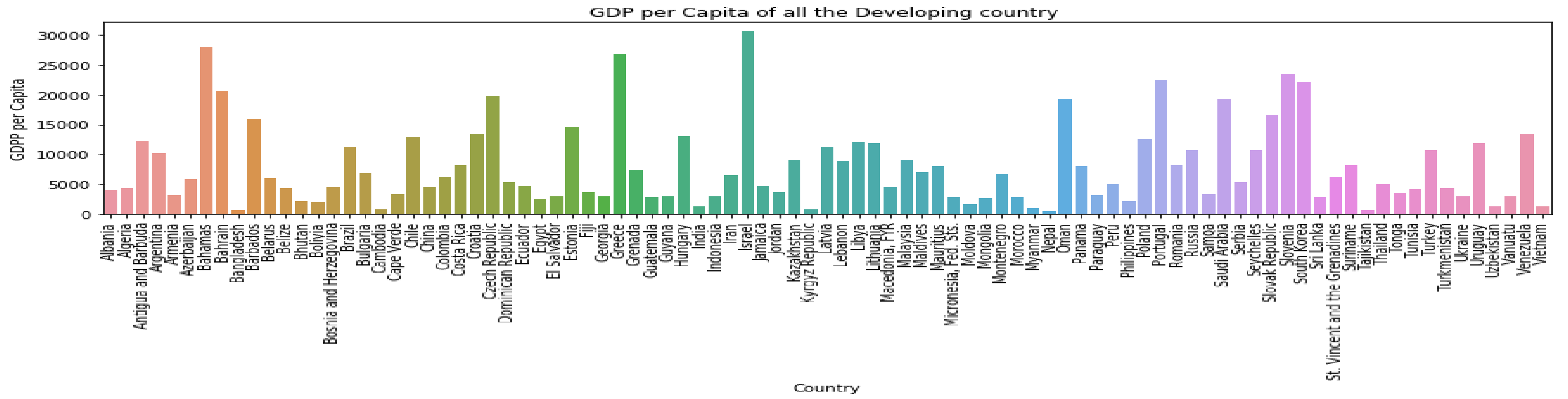
PLOT 1 : we can see all the Developed countries like Luxembourg, Australia,Norway,Qater etc.

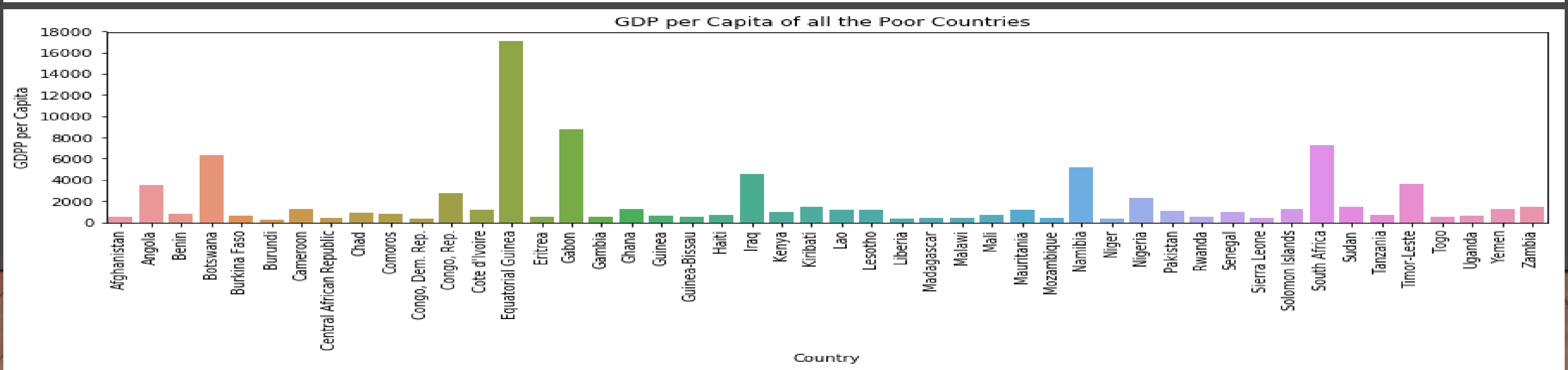


Developed , Developing , Poor Countris Plot based on GDPP

PLOT 2 : we can see all the Developing countries like India,Isreal,Iran, Albania, etc.

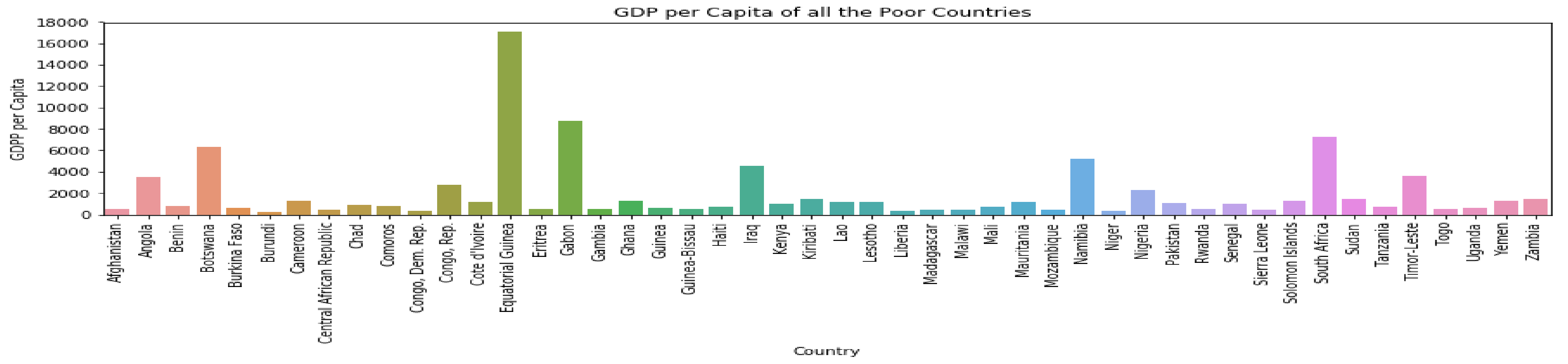
PLOT 3 : we can see all the Poor countries like Bruundi,Nigar,Afghanisthan etc.

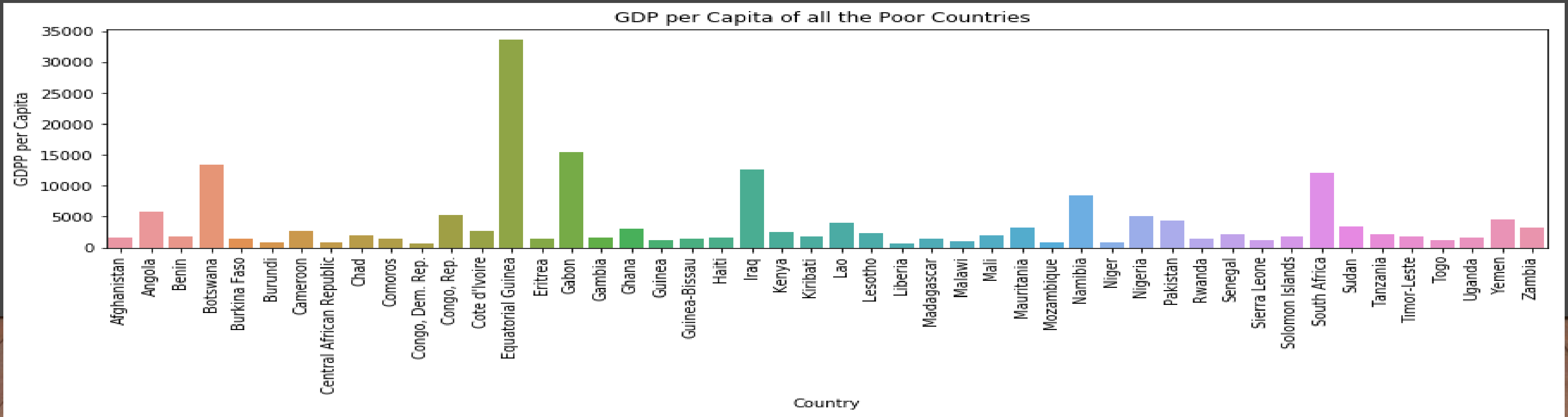




Under- Developed Country based on GDPP,Income,child_mort

PLOT 1 : we can see all the Poor countries like Bruundi,Nigar,Afghanisthan etc.





Hierarchical Clustering

I also Perform the Hierarchical Clustering on the Data I also perform the Single linkage and complete linkage on the PCA components and besed on that I cut the tree on the level of 3 because most of data is fall under 3 level.

And I also perform the Outlier on the data ans I also oberve the data and I see there is same countries fall in poor countries so both the algorithm gives the same result.

As per My Understanding and Knowledge, Based on insights drawn out of the data , The following Conclusion can be made to the Country who need Help

After comparing both K-means and Heirarchical clustering method. I am going with the K-means outcomes as the plots are clearly visible. As in both the methods, the top 8 under-developed countries are similar. I am considering the result of k-means outcome.

After grouping all the countries into 3 groups by using some socio-economic and health factors, we can determine the overall development of the country.

Here, the countries are categorised into list of developed countries, developeing countries and under-developed countries.

In Developed countries, we can see the GDP per capita and income is high where as Death of children under 5 years of age per 1000 live births i.e. child-mort is very low, which is expected.

In Developing countries and Under-developed countries, the GDP per capita and income are low and child-mort is high. Specifically, for under-developed countries, the death rate of children is very high.

Conclusion or Recomendetions

From bar chats, we can clearly see the socio-economic and heath situation of the under developed countries. In countries like Haiti, Sierra Leone,Chad, etc., the death rate of children under 5 years of age per 1000 (child-mort) is high.

In countries like Burundi, Congo, Niger, etc., GDP per capita is very low. So, in those countries, the income per person is also low. So, these countries are considered as poor contries.

Finally, as per categories of the countries, top 8 under-developed countries which are in direst need of aid are as below:

1) Burundi
2) Congo, Dem. Rep.
3) Niger
4) Sierra Leone
5) Central African Republic
6) Mozambique
7) Guinea-Bissau
8) Burkina Faso

THANK YOU