Clustering & PCA Assignment

By.

Sharath Kumar T S

Problem Statement

- 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. It runs a lot of operational projects from time to time along with advocacy drives to raise awareness as well as for funding purposes.
- After the recent funding programmes, they 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. The significant issues that come while making this decision are mostly related to choosing the countries that are in the direst need of aid.
- And this is where you come in as a data analyst. Your job is to categorize the countries using some socio-economic and health factors that determine the overall development of the country. Then you need to suggest the countries which the CEO needs to focus on the most. The datasets containing those socio-economic factors and the corresponding data dictionary are provided below.

Analysis Approach

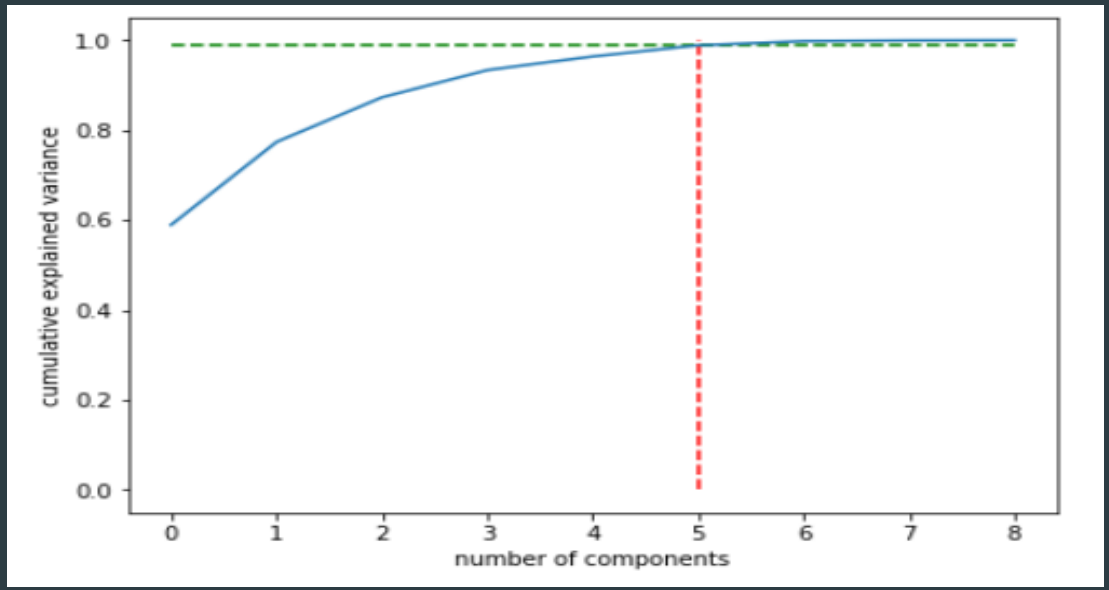
As the data here given is a unsupervised data, we had used the clustering approach to cluster the countries into different clusters, based mainly on their income, child mort rate and per capital gdp.

Here the data consists of 167 countries which includes developed, developing and under developed countries.

Here first we check weather we have any null values in the dataset, and here from the notebook we can find that there is no null values in the dataset.so the dataset is fine and have all the values, next we check for outlier detection by using a box plot, there are some outliers in the dataset we need to treat the outliers.

In the next step we'll check the correlation matrix on a heat map and then we scale the dataset.

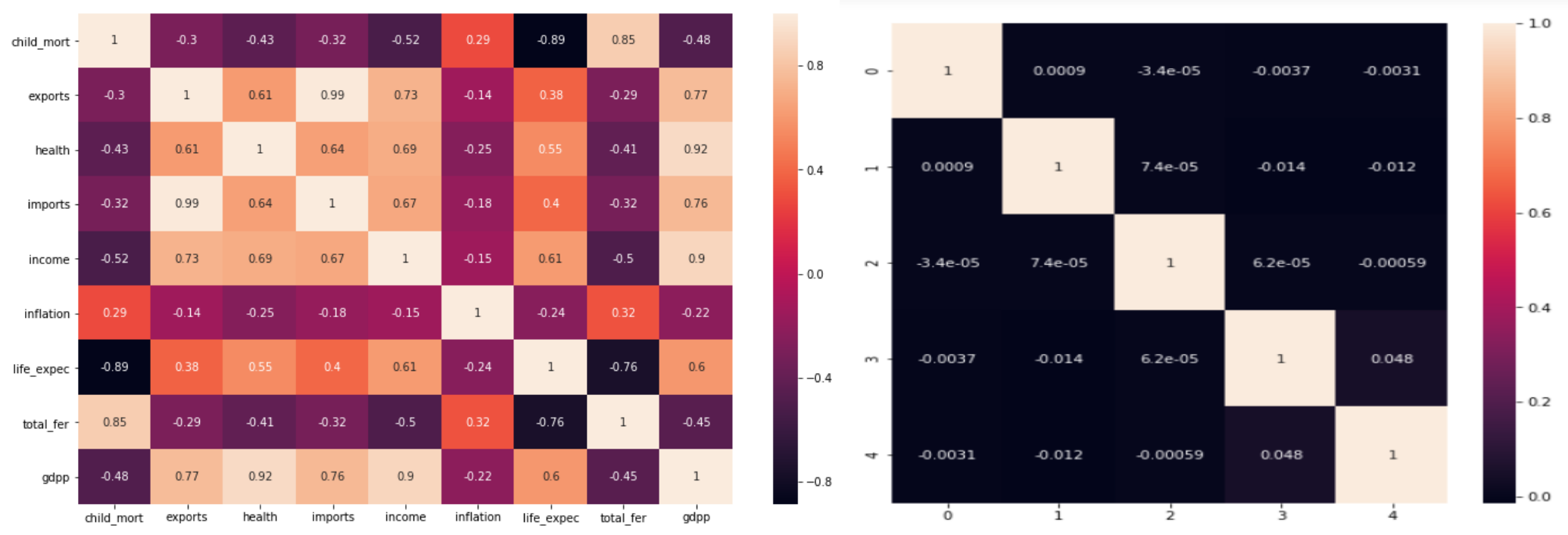
Performing the PCA on the dataset and plotting the cumulative variance against the number of components, we get the below graph.



From the above graph we can see that about 99% of data is being explained by the first 5 components. So we go ahead and perform the dimensionality reduction.

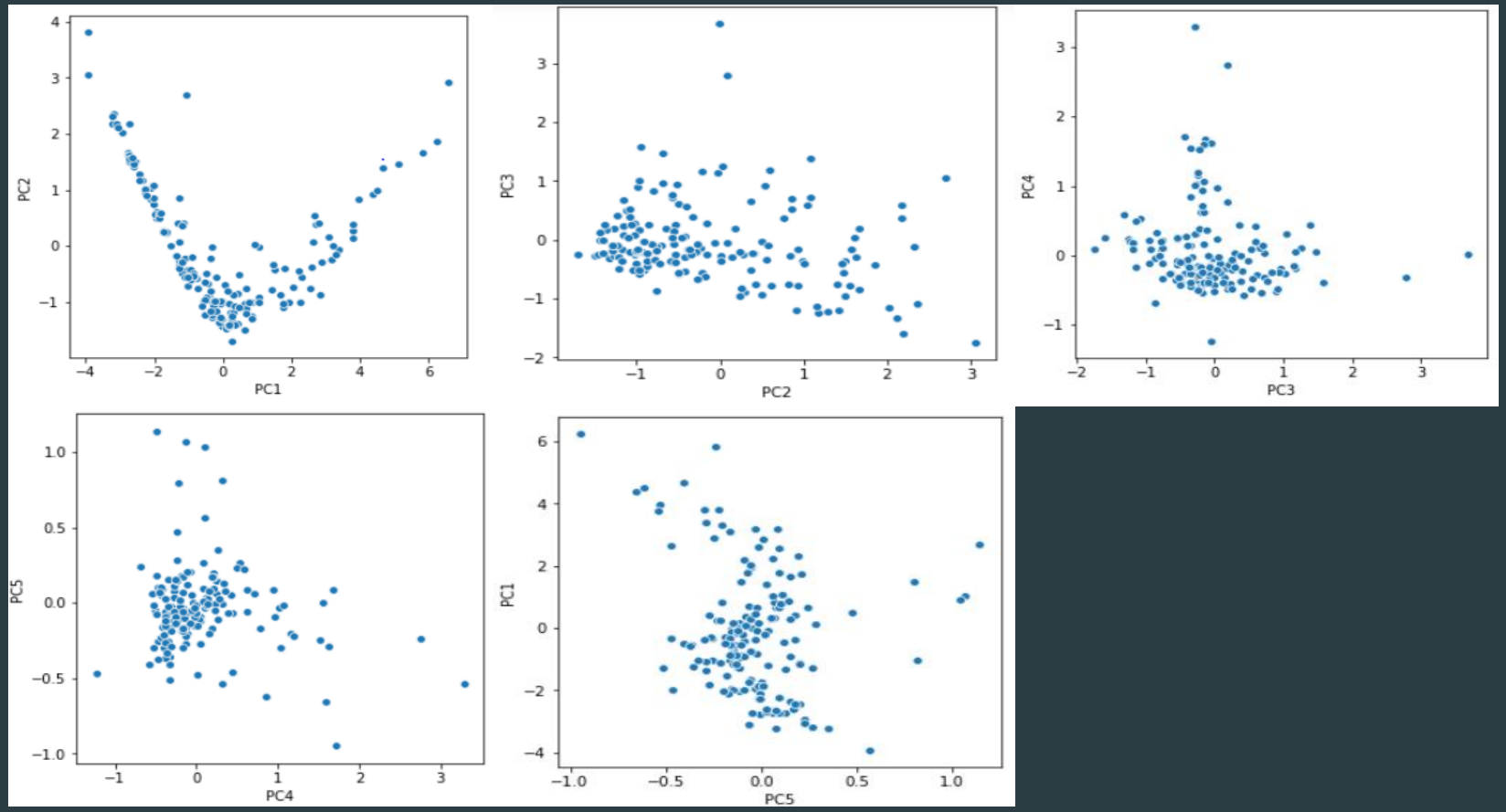
After considering 5 PC’s we need to we need to check the correlation matrix, here we are plotting a heat map for that.

Below we can see correlation between the variables before and after PCA



From the above graph we can see that now the correlation between the variables is almost zero.

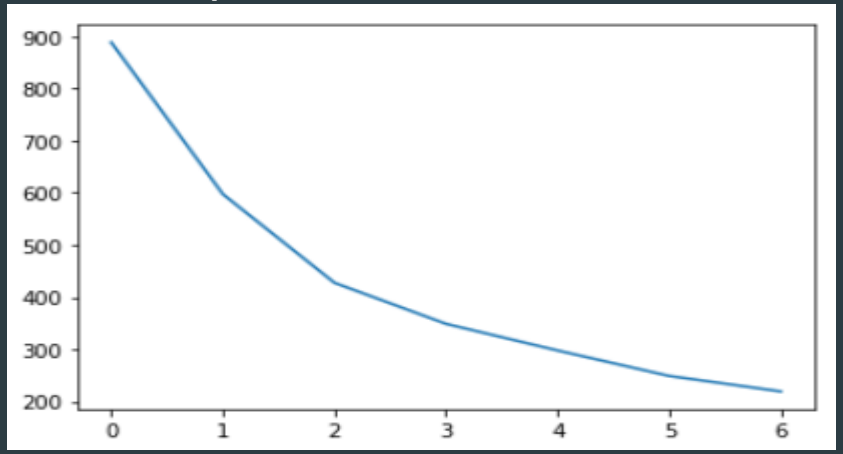
- After performing the PCA we check the data by plotting a scatter plot for the principal components



Next we perform the Hopkins statistic to find the cluster tendency.

Here we are getting a value which is more than 0.5, which implies the dataset good for clustering,

Now we perform the Elbow curve analysis to find the value of k.



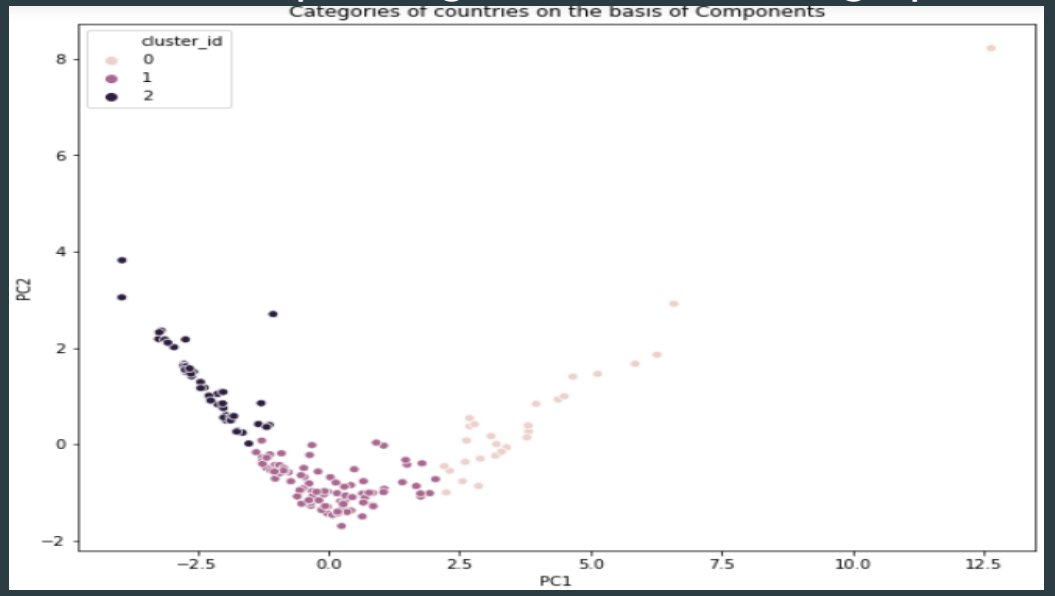
The we perform the silhouette analysis also and we finally consider 3 and the k value.

Now we perform the k-means clustering and assign the cluster number to the respective cluster points. From the below diagram we can find that distribution of countries with respect to the

1	91
2	48
0	28

cluster.

Now we are plotting the PC1 V/s PC2 graph along with the classification.

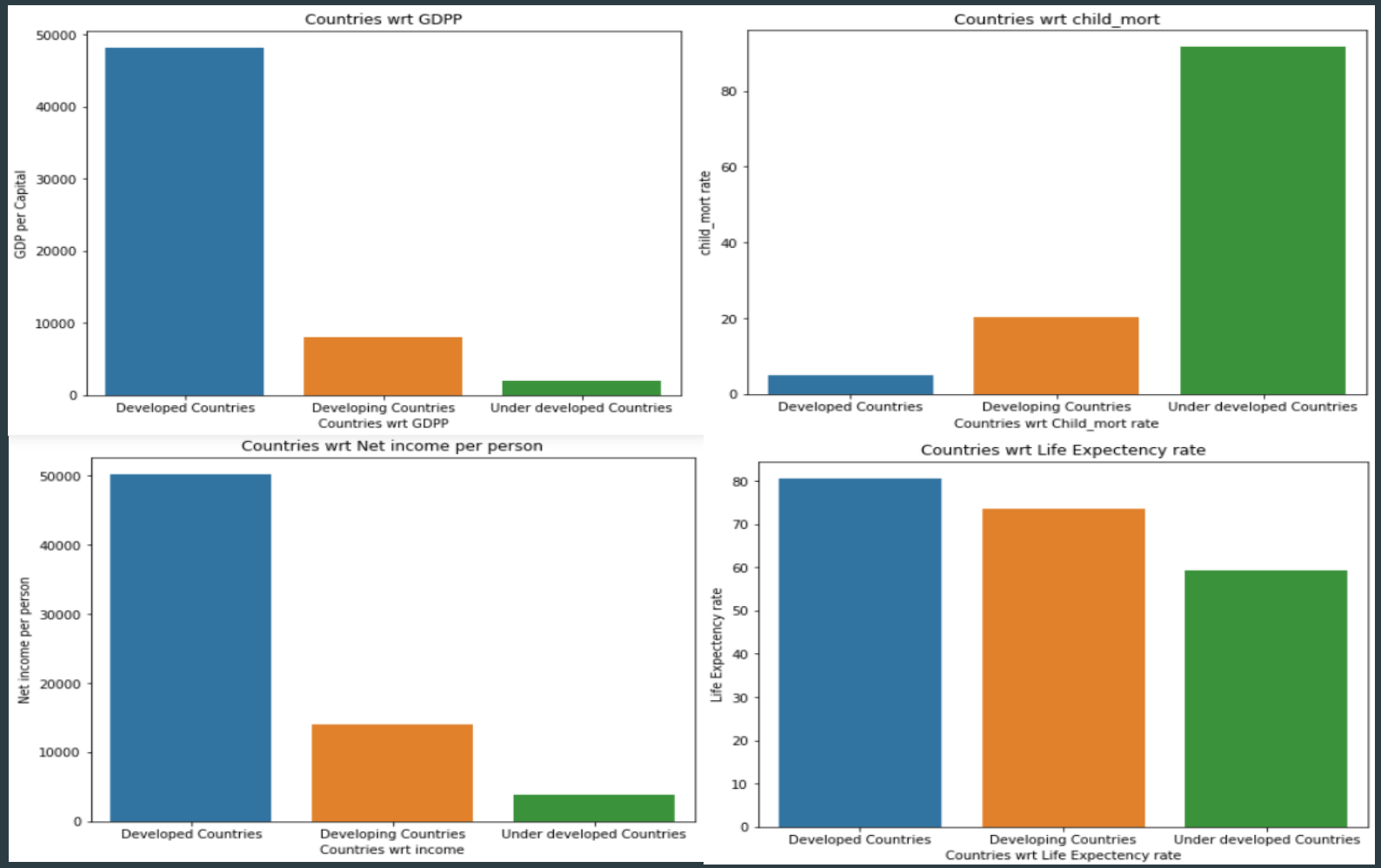


Now we are merging the original dataset with the PC's along with the cluster id, and then we are removing the Principal components.

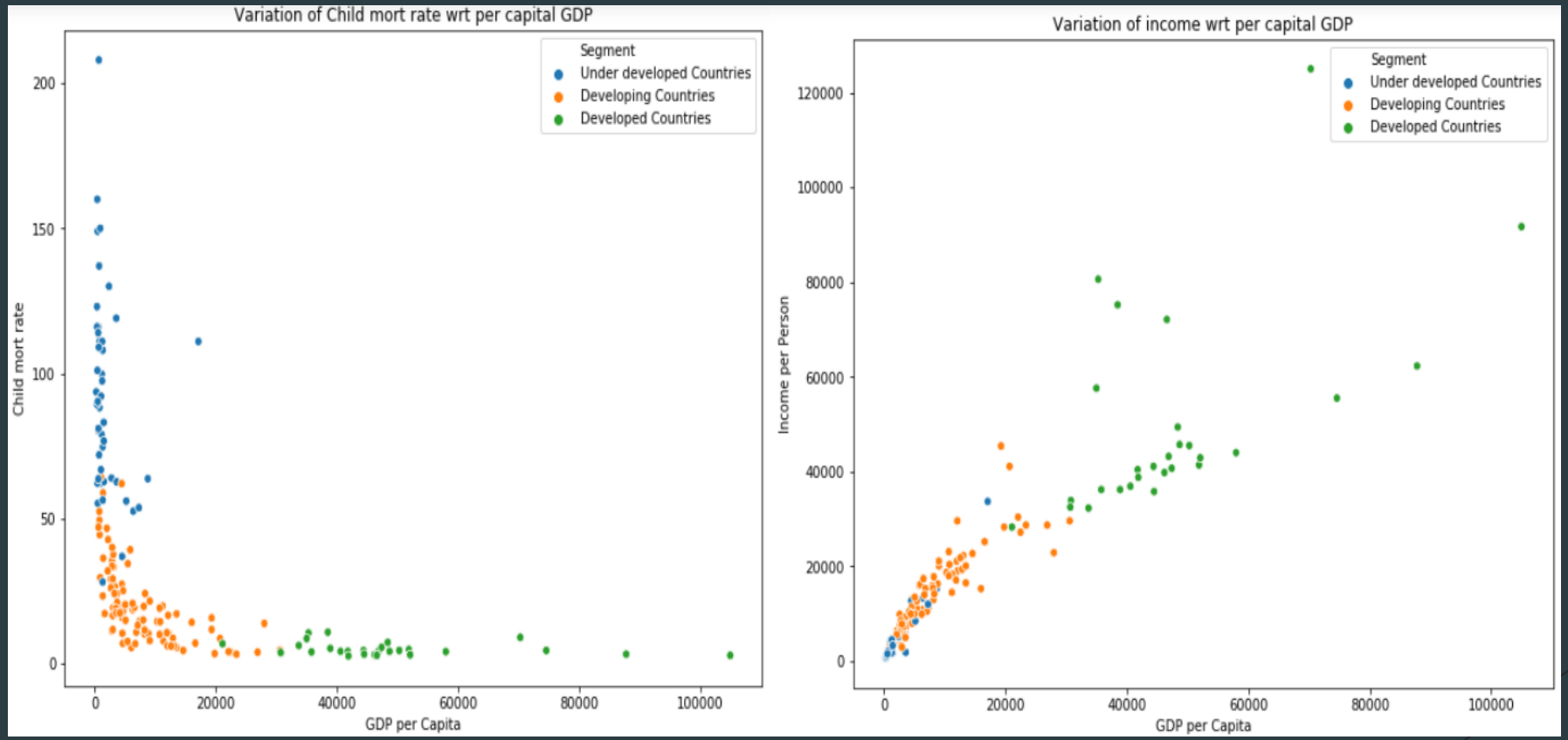
Now we find the mean of all the columns and then we prepare a separate table with only mean values wrt each clusters.

We rename the cluster ids to names, as the countries with highest GDPP will be developed country, the countries with medium level GDPP called developing countries and the countries with low GDPP comes under developed countries.

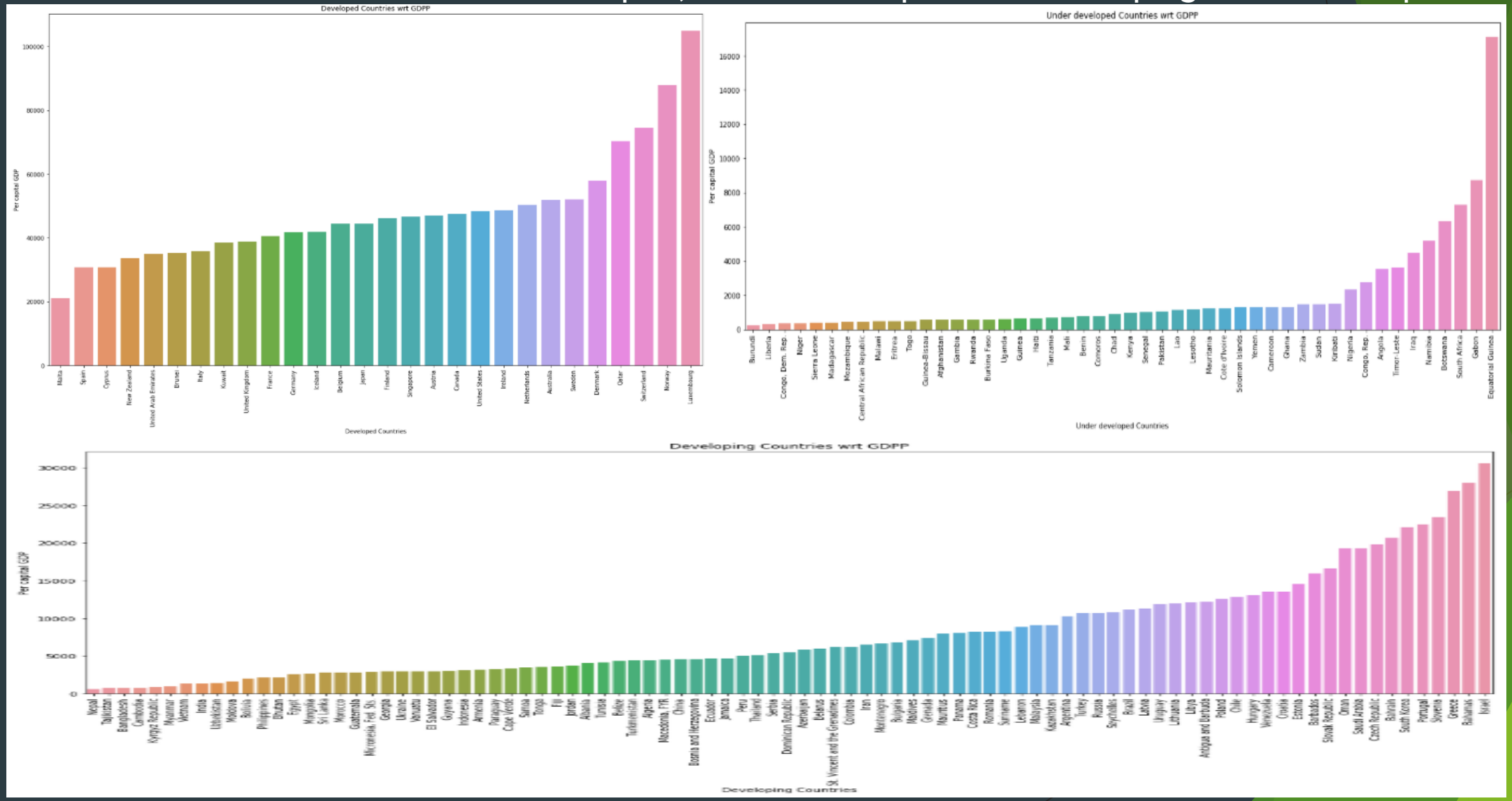
- Plotting some of the graphs. The graphs shows the achievement of each cluster wrt GDPP, Childmort, Income & Life Expectancy



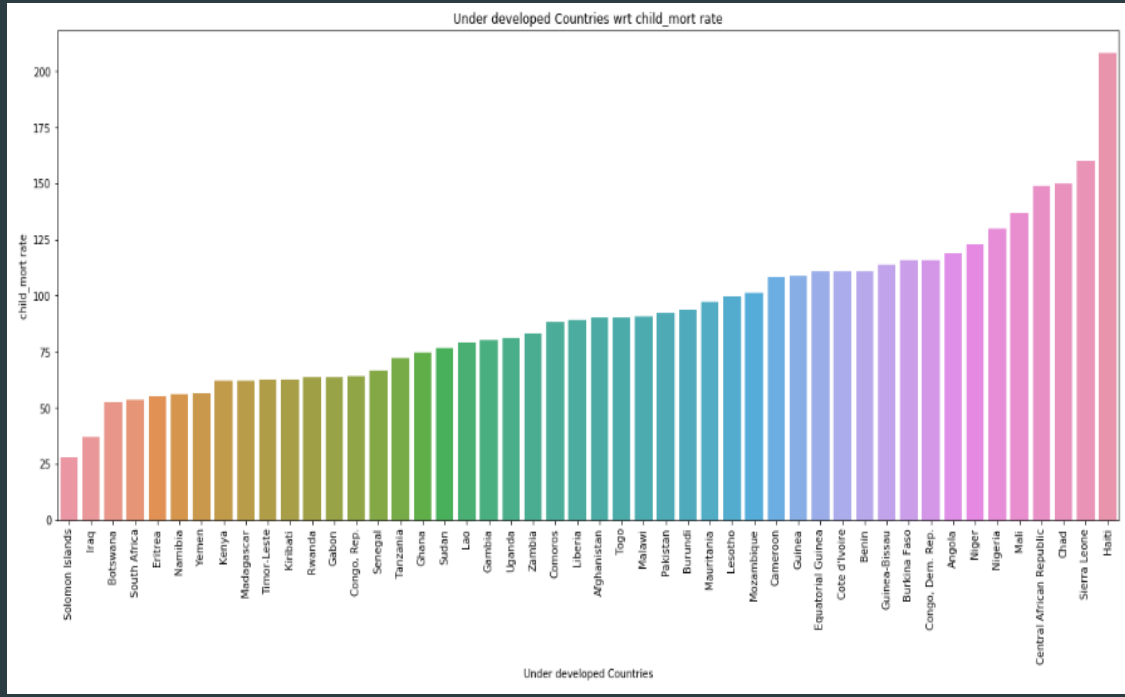
- Plotting the variation of GDPP wrt child mort rate on a scatter graph. And another graph with GDPP wrt income .

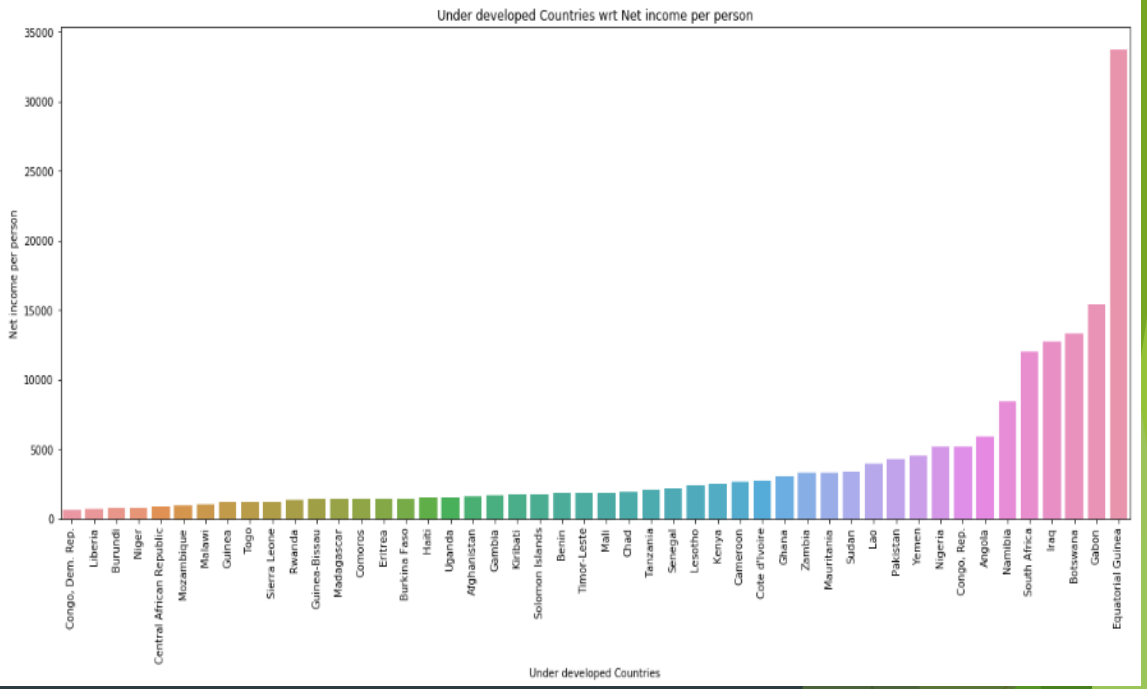


Next we create separate 3 dataframes with each segment value. The below shown graph gives us the countries which come under Developed, under-developed and developing countries respectively.



- The below graph shows the variation of child mort rate and income for the under-developed countries.





Final list of countries in need for aid from k-means clustering algorithm.

Now we will take top 10 countries with child mort rate, last 10 countries with gdpp and income, we create 3 dataframes and then we'll intersect all these 3 dataframes wrt country names to find the country names with less gdpp & income and more child mort rate. Finally we get 8 countries with need for aid as quick as possible from k-means clustering method.

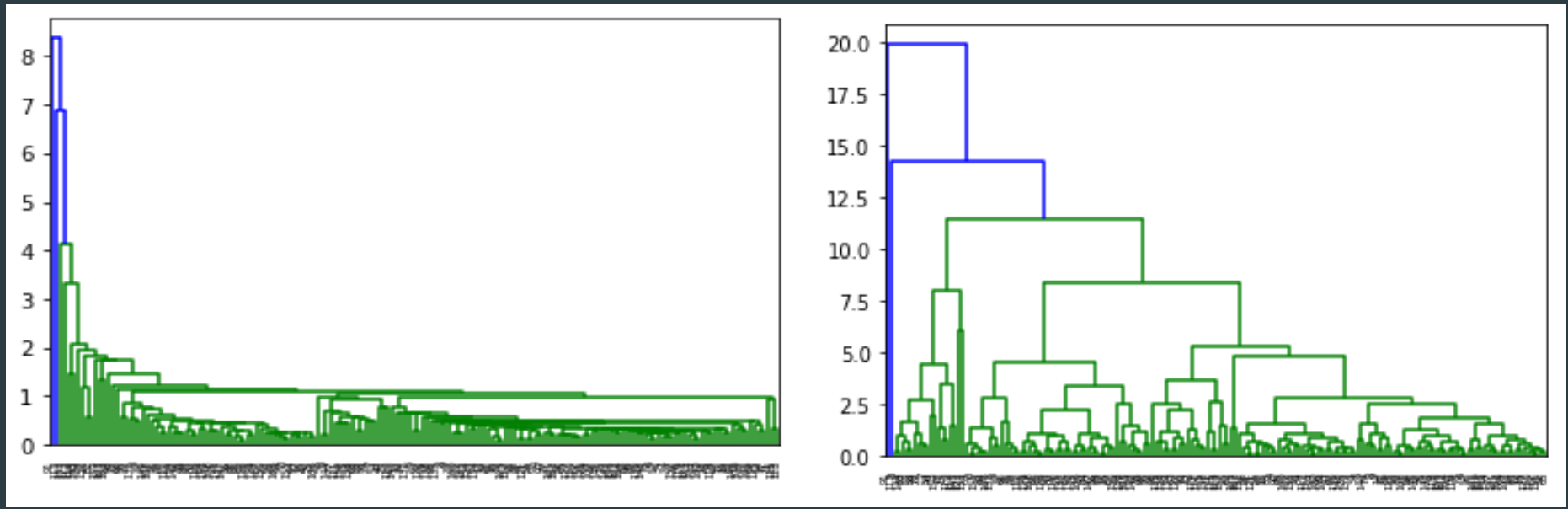
The countries are:

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

Hierarchical clustering

Now we are performing the Hierarchical clustering to find the countries in need for aid.

First we perform the Hierarchical clustering with single linkage and complete linkage, the plots are shown below respectively.



Now we cut the dendrogram with k = 3

Next we calculate the clusters which they belong to and add the cluster labels in the dataframe with principal components, next we merge this new dataframe with the original dataframe and the we remove the principal components.

Now again here we calculate the mean for each columns for the under developed countries, and then we add all the dataframe to a single dataframe and check the mean for each columns for the under-developed category,

In the mean while we changed the cluster labels with the country names.

Now we filter the original dataframe with respect to mean values which we found using heirarchial clustering method. Finally we got 6 countries that come under this category, which are in eager need of aid. The countries are:

1. Central African Republic
2. Chad
3. Haiti
4. Mali
5. Nigeria
6. Sierra Leone

Conclusion

- By using the k means and hierarchical clustering techniques we had successfully clustered the data into 3 clusters.
- The clusters are made wrt GDPP, hence the developed countries will have more GDPP, and as GDPP is linked with income, child mort rate, and other features, The developed countries will have very less child mort rate and high income per person.
- In the developing country the GDPP is in the mid range, and its growing, GDPP is less than the developing country nut greater than under developed country.
- In the under developed countries the GDPP is very low, which implies the income per person is low and the child mort rate is high, hence the under developed countries are the one for which the aid is required, as we can't provide the aid to all the under developed countries, hence we take the top and least 10 from each column and find the countries which are dearest need for aid.

Recommendations

- From the above analysis it implies that the GDPP is directly related to growth of the country, the country with least GDPP have high child mort rate,low income, low life expectancy.
- WRT GDPP Burundi, Congo, Niger, etc. are considered as the poorest countries in the world,
- Haiti, Central African Republic, Sierra Leone, chad, Nigeria are the countries with highest chikd mort rate
- Congo.Dem.Rep., Liberia, Burundi, Niger & Central African Republic are the countries with least net income per person.
- From the bar chart which we draw for the under developed countries wrt child mort rate and life expectancy rate, we can consider top nations which are need for aid: they are

1. Central African Republic
2. Haiti
3. Sierra Leone
4. Chad
5. Nigeria
6. Lesotho