Clustering and PCA Assignment Part - I

- Student Name: Hafizullah Mahmudi
- Student Id: APFE19M00734
- IIIT Role Number: DDS1950112
- Kabul, Afghanistan

- 23rd February, 2020

- Institute: upGrad/IIIT-B

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.

Analysis Approach

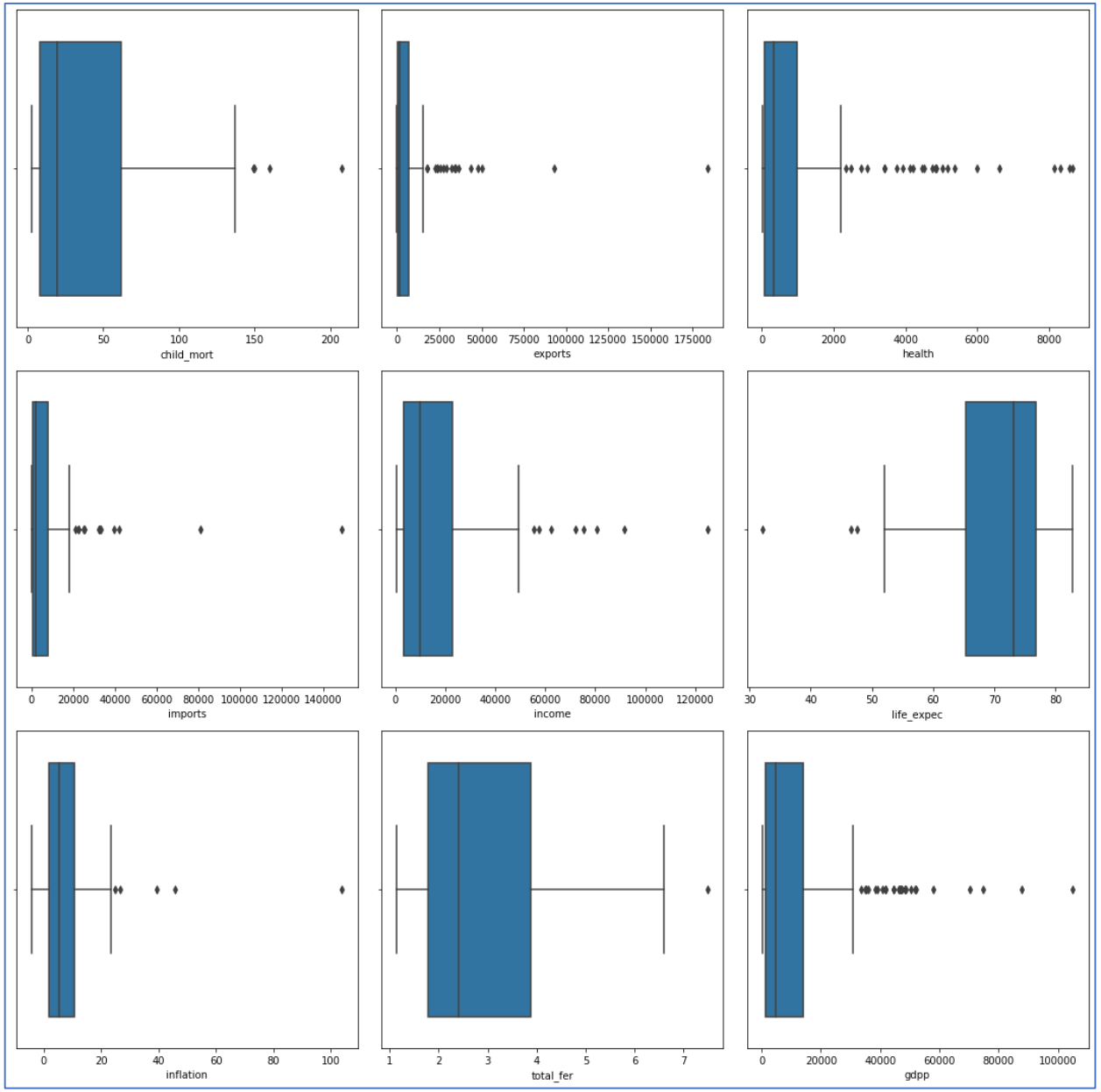
At first the data was reviewed, null values checked and then outliers were identified and replaced with cap values.

After outlier treatment, data were normalized to unify the scale of the numbers

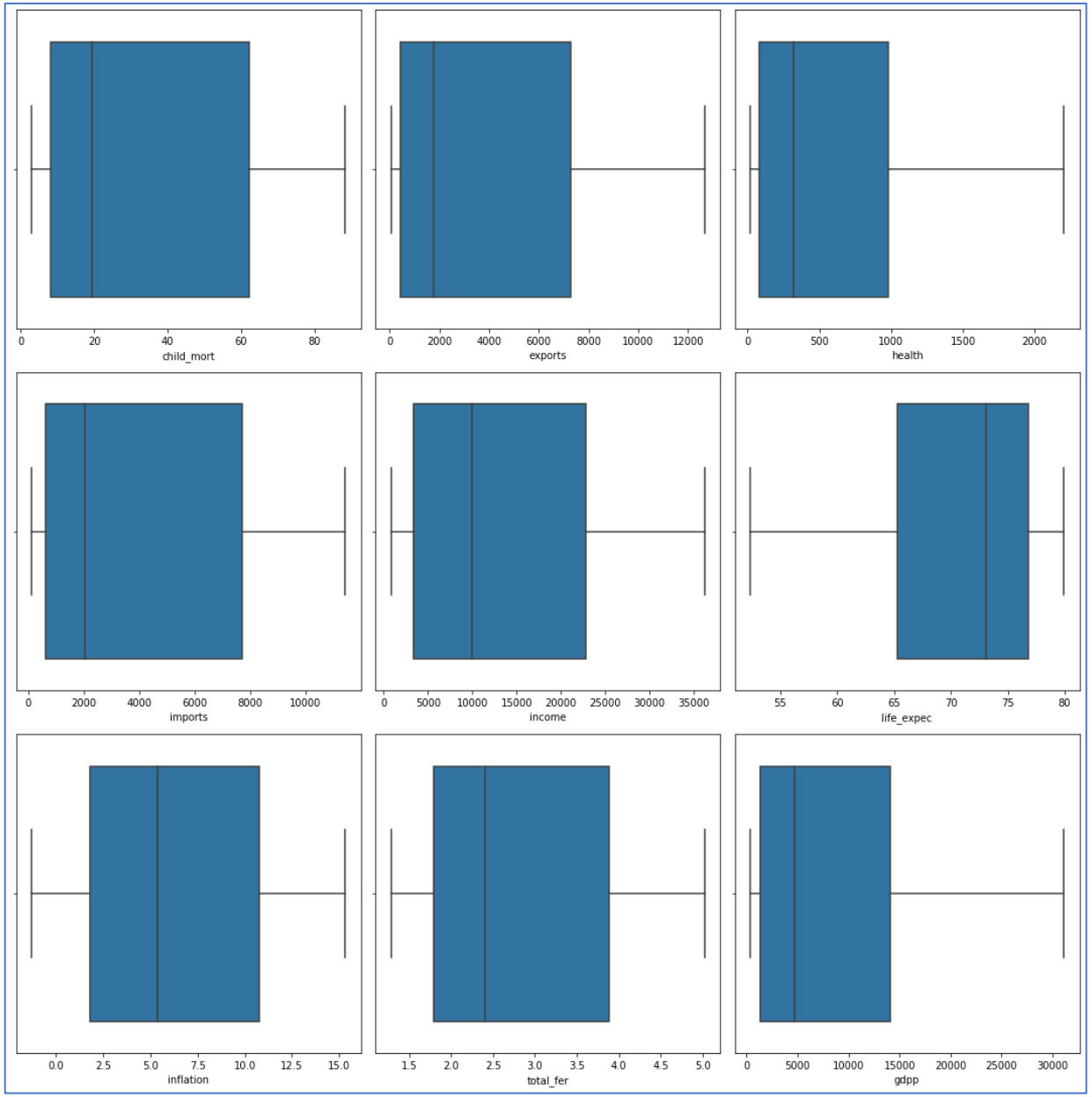
Then PCA applied and the result of PCA were used with two cluster analysis methods (K-Means and Hierarchical Clustering)

Outlier Treatment

Before treatment



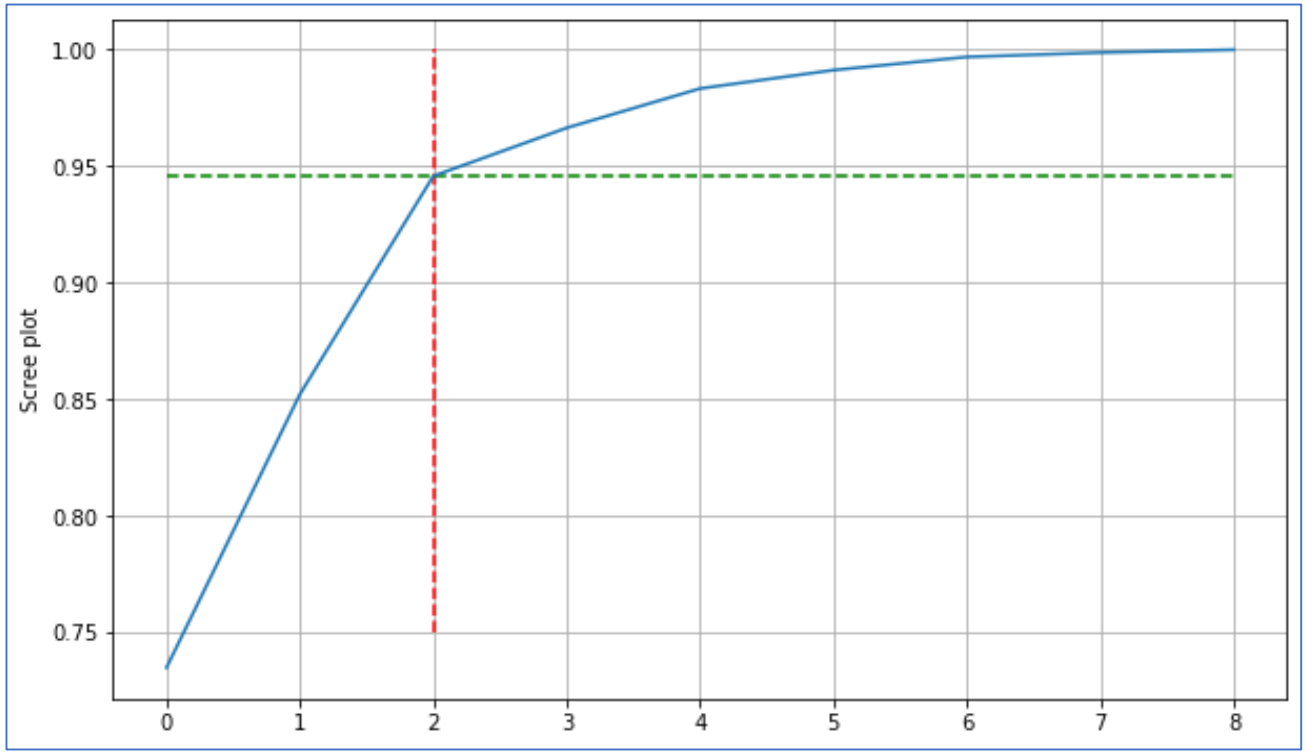
After treatment



Applying Principle Component Analysis (PCA)

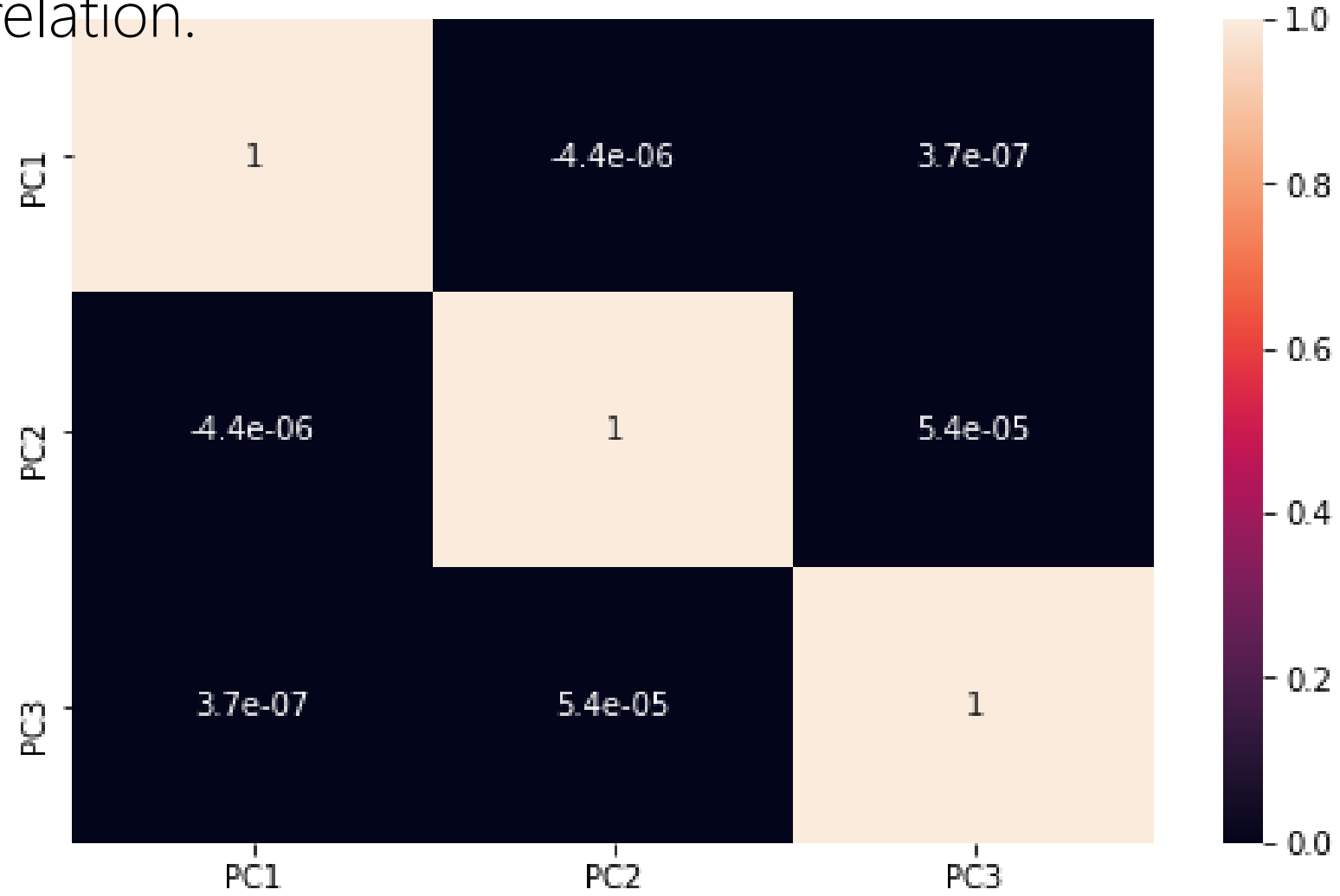
The scaled data was used for PCA. The PCA helps to find most optimum components while removing less significant variables.

The scree plot analysis shown here suggests 3 component for PCA analysis.
Selecting 3 components will explain about 94.57% of variance.



Correlation Matrix of the PCA Components

As seen from the below graph, there is nor relation between components. All components are around zero which is an indication of no correlation.



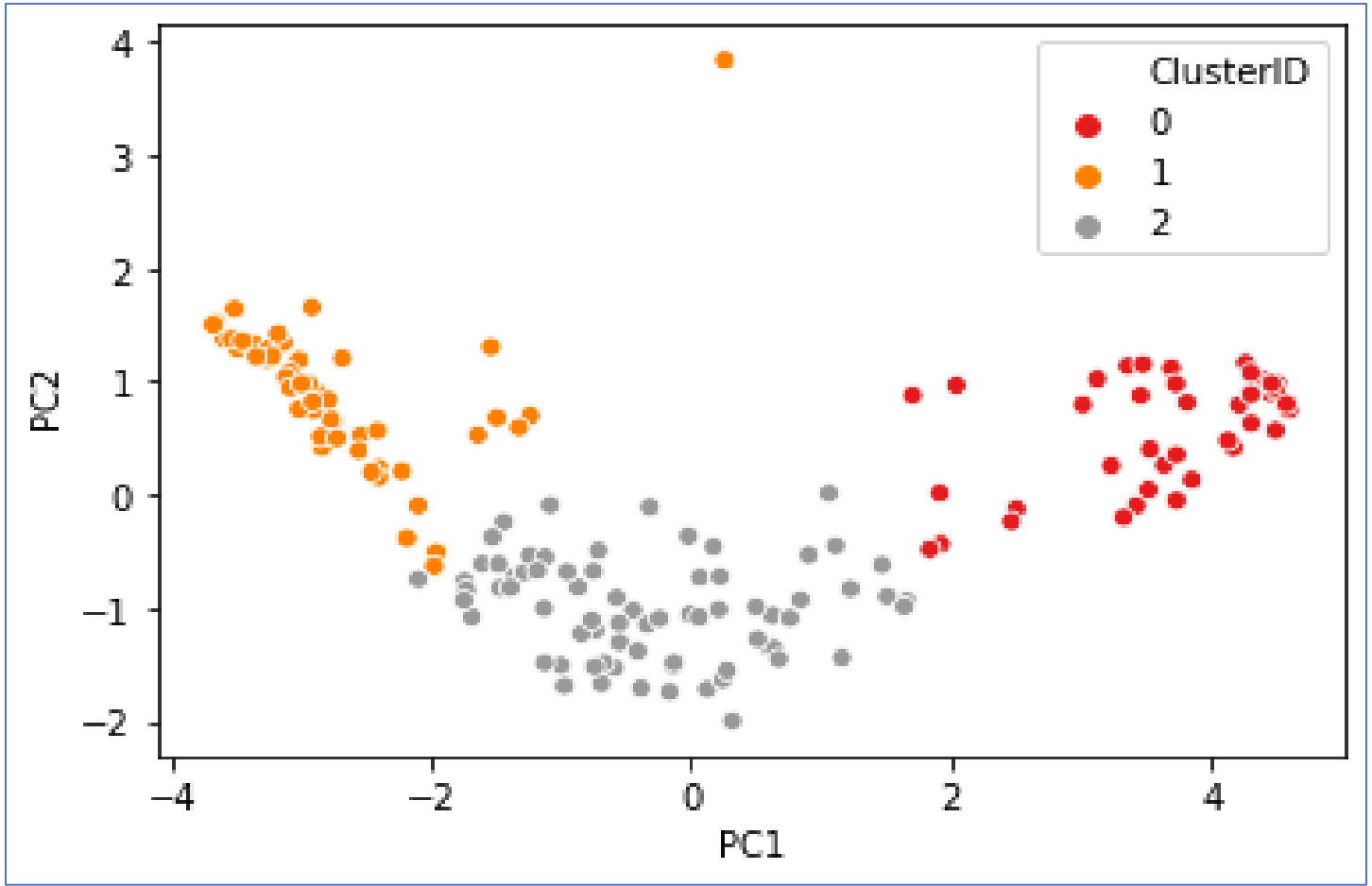
Clustering

The clustering (unsupervised clustering) K-Means and Hierarchical clustering helps us to group countries that are close to each other in terms of socio – economic profile of the given data.

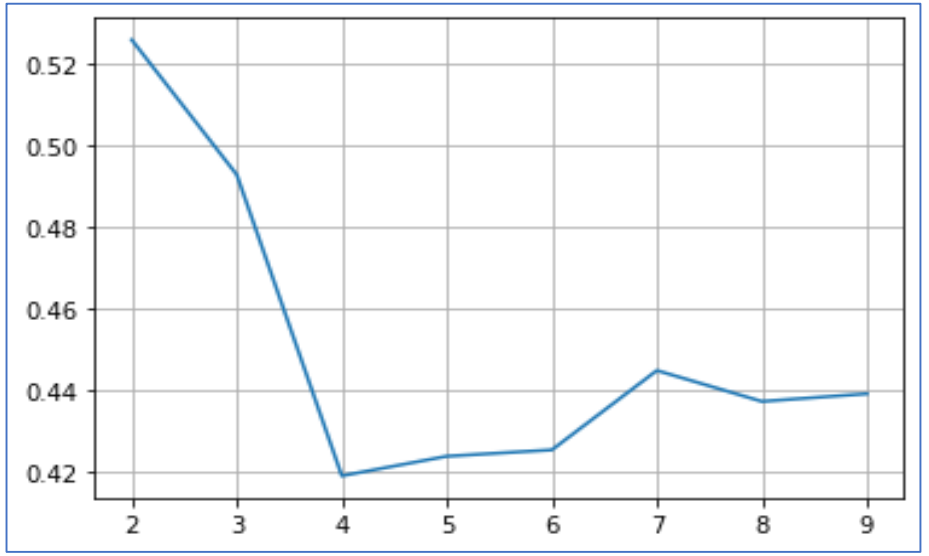
K-Means Clustering

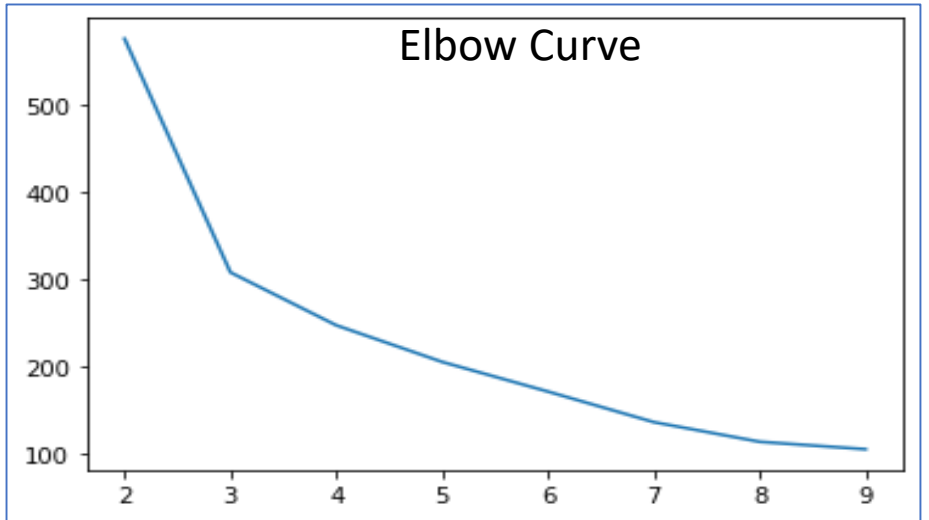
Both Silhouette and Elbow cure plots indicates a good fit dividing the countries into 3 clusters

K-Means Clustering Result in 3 clusters



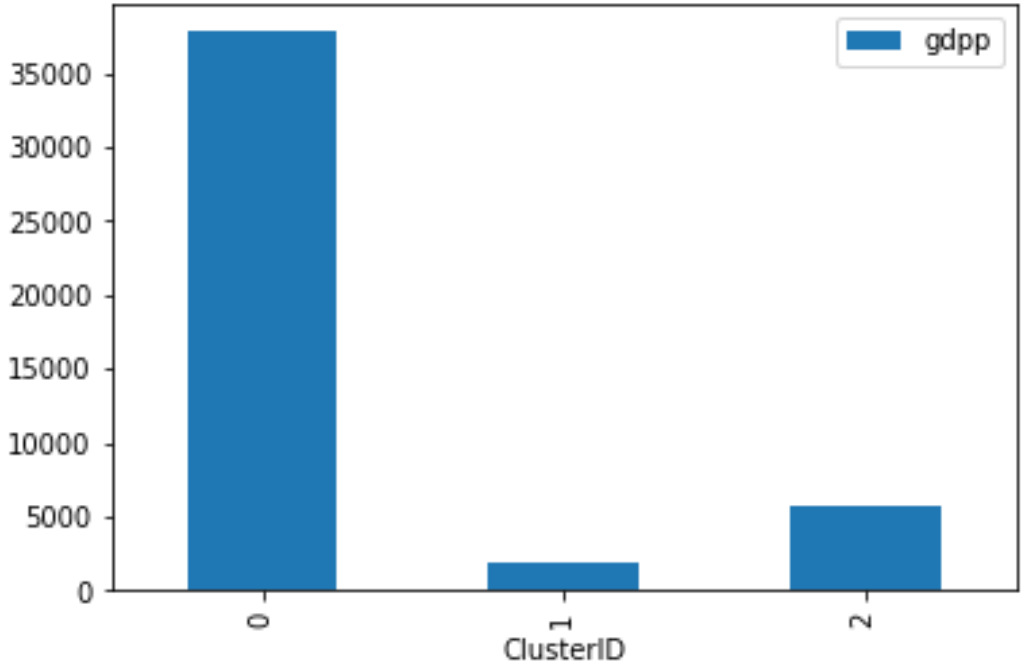
Silhouette

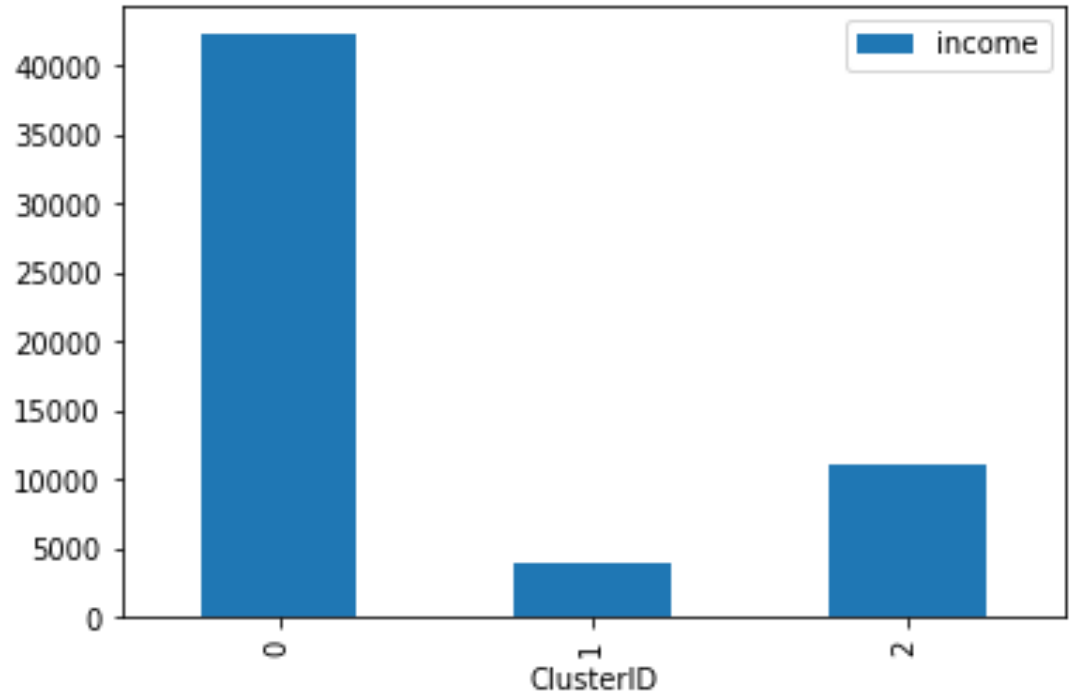


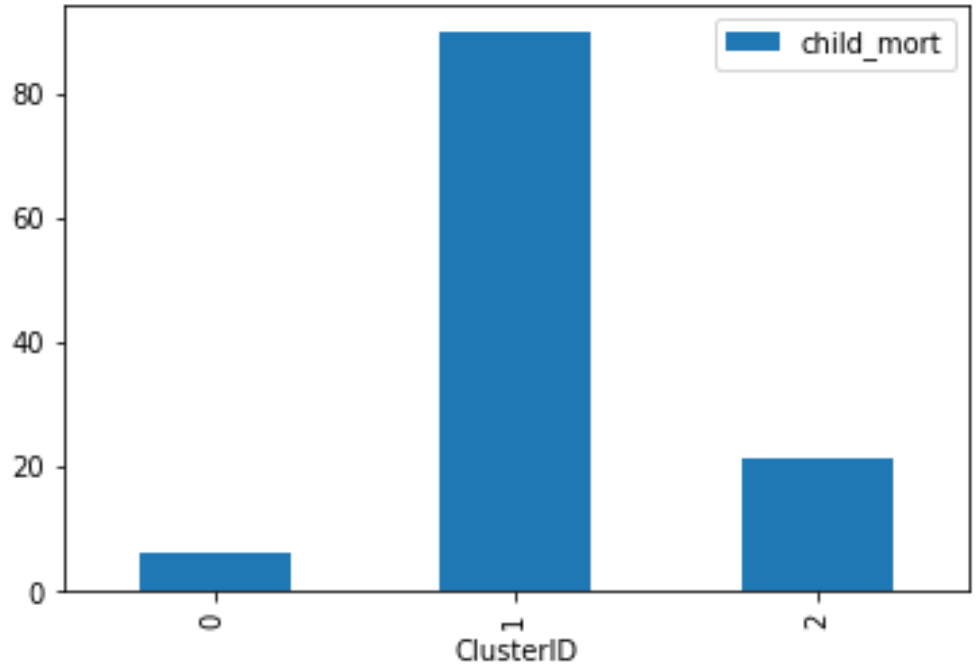


Analysis of K-Means

When analyzing the average per cluster for the following profile of each country (gdpp, child_mort, income).
The clusters 1 shows countries with low income. Cluster 2 with middle income and cluster 0 with high income.

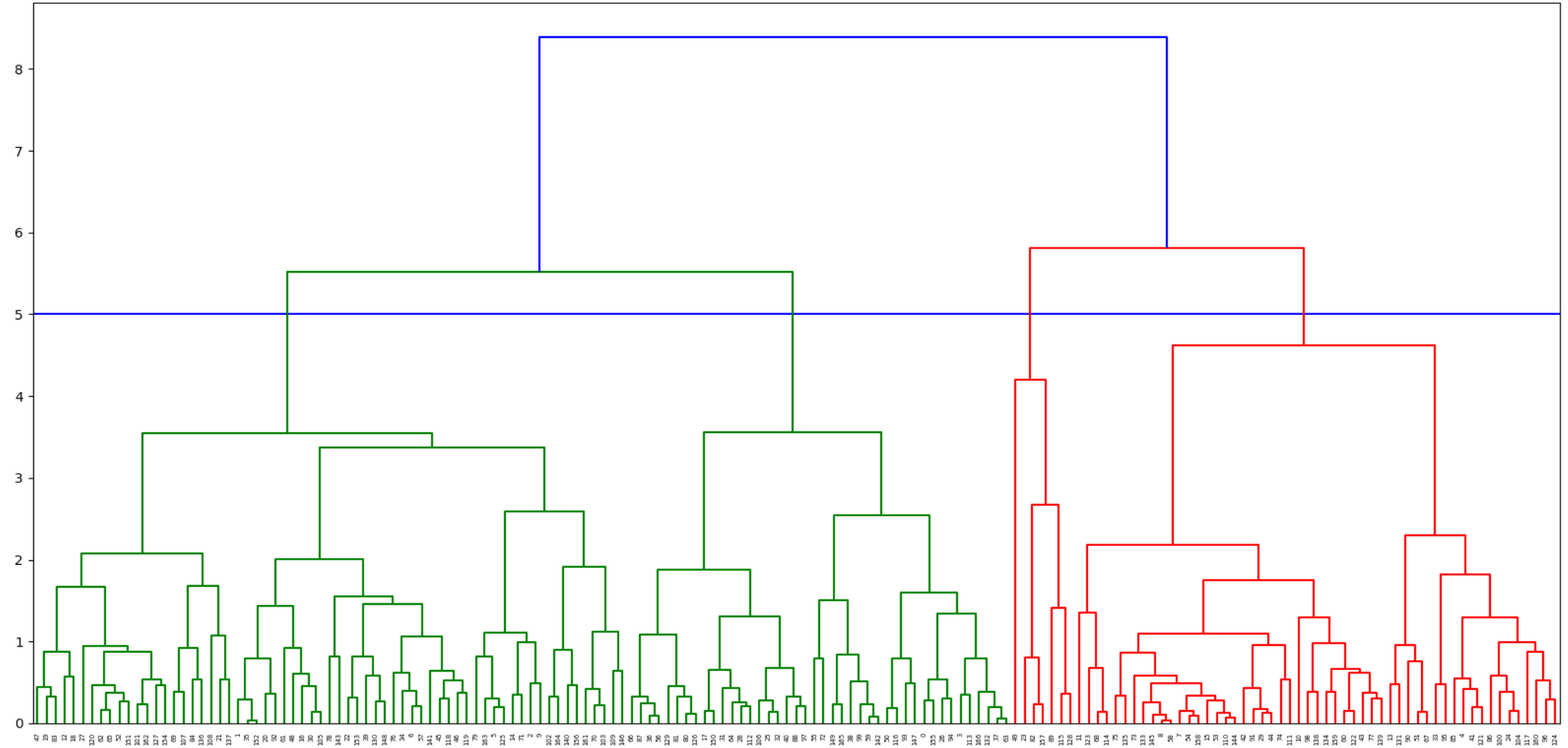






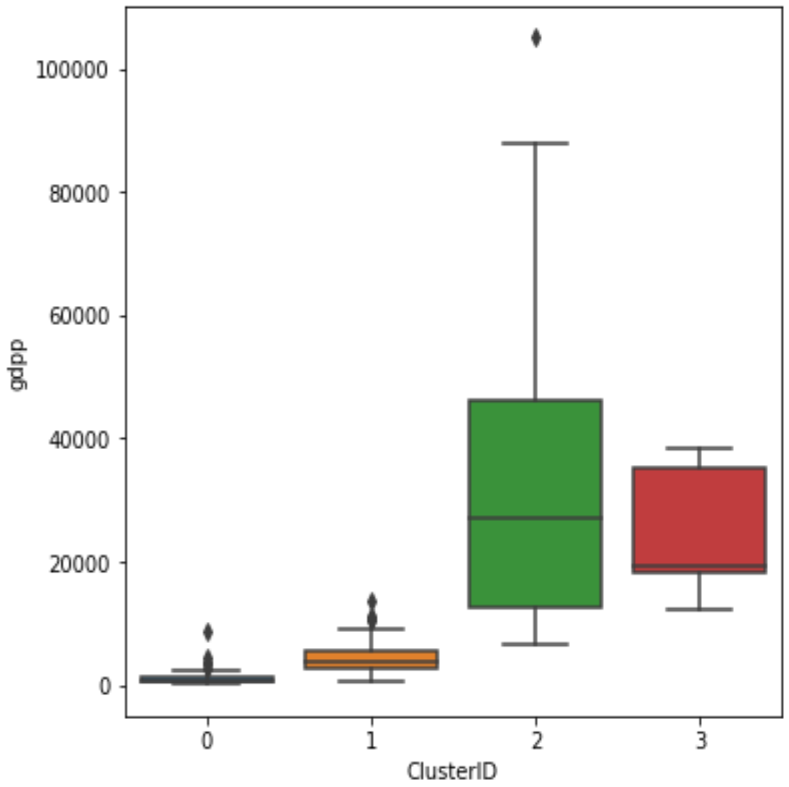
Hierarchical clustering

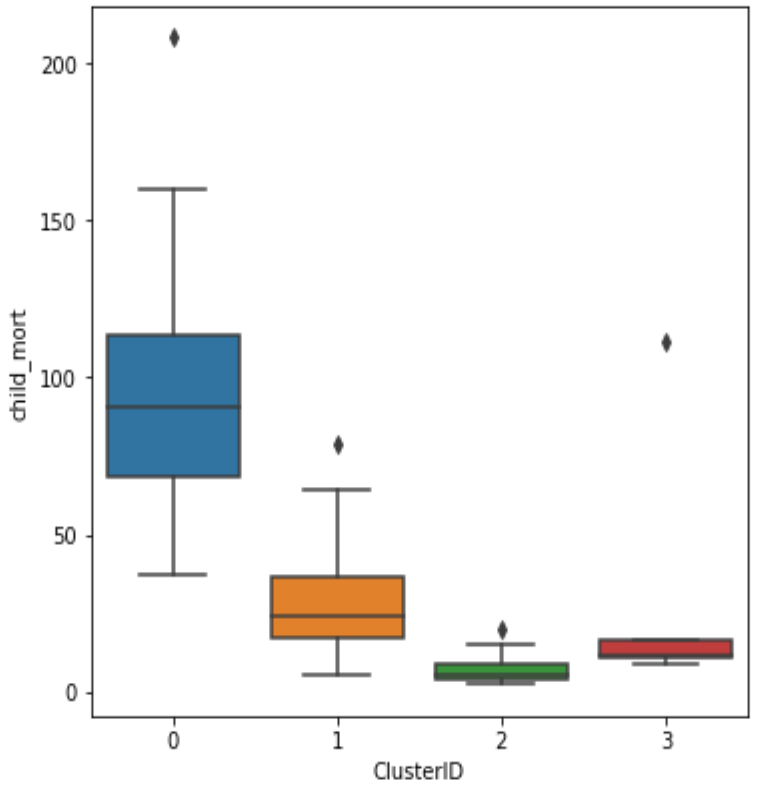
Looking the following dendrogram, a cut of 4 clusters is a good choice. This divides the countries into 4 categories, low-income, lower-middle, upper-middle income and upper income countries.

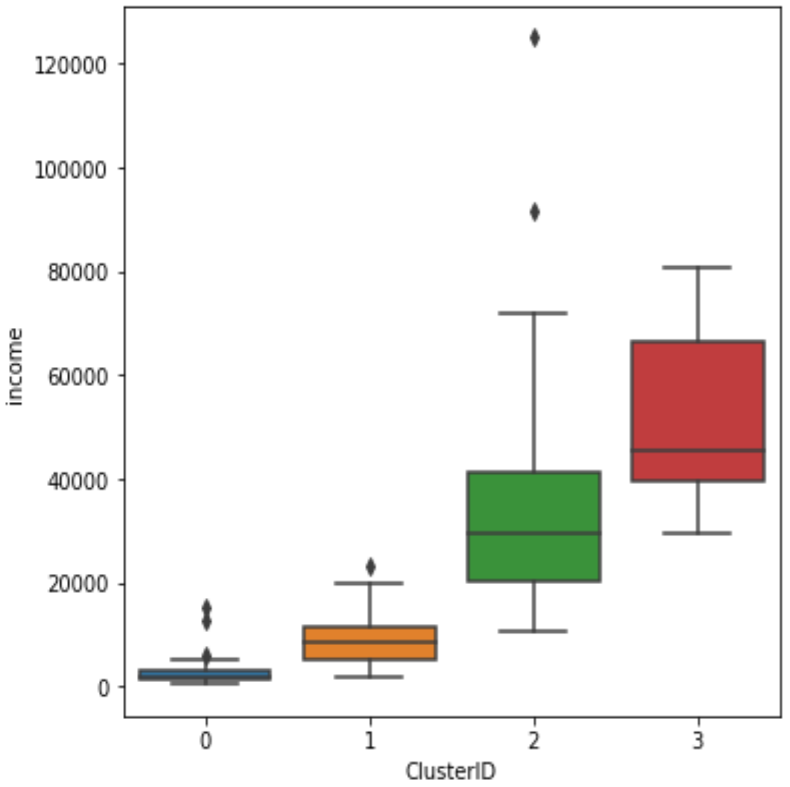


Analysis of Hierarchical clustering

When analyzing the average per cluster for the following profile of each country (gdpp, child_mort, income).
The clusters 0 shows countries with low income and lower-middle income.







Recommendation (option 1)

- According to World Bank, https://datahelpdesk.worldbank.org/knowledgebase/articles/906519
- Countries with GNI or income of per capita between $1,026 and $3,995 is considered as low-middle income countries and countries with lower GNI of $1,025 is considered as low-income countries.
- These countries have a very high percentage of population under poverty line

Assuming our data set reflects that latest data then

- These lists out those countries that have average income of $2722, high child mortality rate of 79 and low gdpp of $1274. So the following countries will fall under above average values.

These countries are:

'Afghanistan' 'Angola' 'Benin' 'Burkina Faso' 'Burundi' 'Cameroon' 'Central African Republic' 'Chad' 'Comoros' 'Congo, Dem. Rep.' 'Congo, Rep.' "Cote d'Ivoire" 'Eritrea' 'Gabon' 'Gambia' 'Ghana' 'Guinea' 'Guinea-Bissau' 'Haiti' 'Iraq' 'Kenya' 'Kiribati' 'Lesotho' 'Liberia' 'Madagascar' 'Malawi' 'Mali' 'Mauritania' 'Mozambique' 'Niger' 'Nigeria' 'Pakistan' 'Rwanda' 'Senegal' 'Sierra Leone' 'Sudan' 'Tanzania' 'Timor-Leste' 'Togo' 'Uganda' 'Yemen' 'Zambia'

Recommendation (Option 2) – to be cautious

But still if a prioritization is needed or the budget is limited or want to be more strict then it is we can filter down the above countries who are below average income of $2722, gdpp of $1274 and child mortality rate above average of 79 children

These countries are:
There are 22 countries that falls in this category.

'Afghanistan' 'Benin' 'Burkina Faso' 'Burundi' 'Central African Republic' 'Chad' 'Comoros' 'Congo, Dem. Rep.' "Cote d'Ivoire" 'Gambia' 'Guinea' 'Guinea-Bissau' 'Haiti' 'Lesotho' 'Liberia' 'Malawi' 'Mali' 'Mozambique' 'Niger' 'Sierra Leone' 'Togo' 'Uganda'

Recommendation (Option 3)

On the right we have list of countries in descending order based on their GDPP, income and topset child mortality.

The top 5 direst countries are listed on top and other countries are also listed for reference if needed to allocate or provide fun.

no	country	gdpp	income	child_mort
1	Burundi	231	764	93.6
2	Liberia	327	700	89.3
3	Congo, Dem. Rep.	334	609	116
4	Niger	348	814	123
5	Sierra Leone	399	1220	160
6	Mozambique	419	918	101
7	Central African Republic	446	888	149
8	Malawi	459	1030	90.5
9	Togo	488	1210	90.3
10	Guinea-Bissau	547	1390	114
11	Afghanistan	553	1610	90.2
12	Gambia	562	1660	80.3
13	Burkina Faso	575	1430	116
14	Uganda	595	1540	81
15	Guinea	648	1190	109
16	Haiti	662	1500	208
17	Mali	708	1870	137
18	Benin	758	1820	111
19	Comoros	769	1410	88.2
20	Chad	897	1930	150
21	Lesotho	1170	2380	99.7
22	Cote d'Ivoire	1220	2690	111

The End