CLUSTERING & PCA ASSIGNMENT

SUBMISSION

Submitted By

Richa Goel

Socio-Economic Analysis of Countries

Objective:

- The objective is to identify and categorise the countries who are poor and are in direst need of aid using some socio-economic and health factors that determine the overall development of the country.

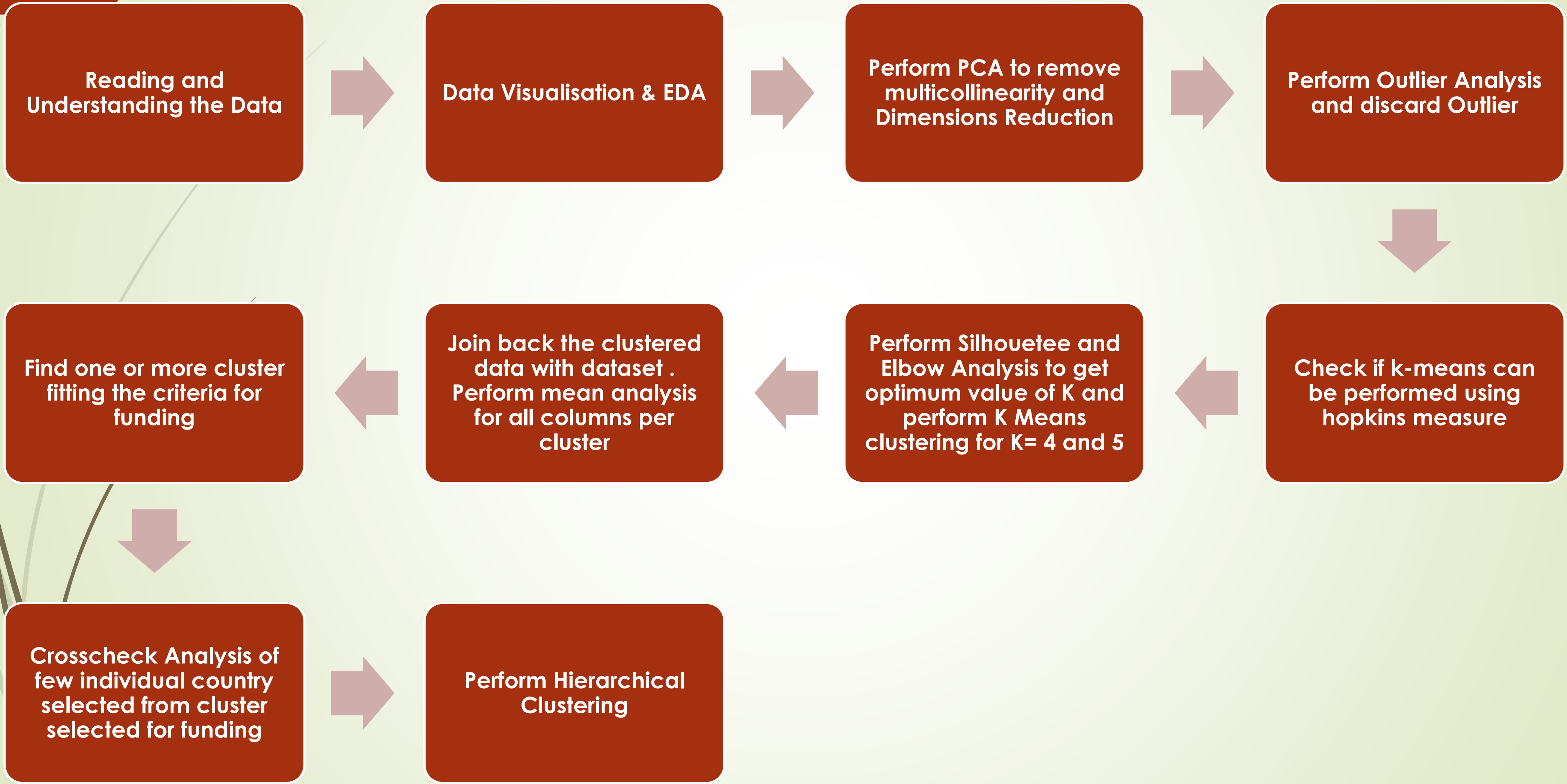
Abstract:

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 project that included a lot of awareness drives and 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.

Our main task is to cluster the countries by the factors mentioned above and then present our solution and recommendations to the CEO .

<Problem solving methodology>



Reading and Understanding the Data

Data Visualisation & EDA

Perform PCA to remove multicollinearity and Dimensions Reduction

Perform Outlier Analysis and discard Outlier

Check if k-means can be performed using hopkins measure

Perform Silhouetee and Elbow Analysis to get optimum value of K and perform K Means clustering for K= 4 and 5

Join back the clustered data with dataset . Perform mean analysis for all columns per cluster

Find one or more cluster fitting the criteria for funding

Crosscheck Analysis of few individual country selected from cluster selected for funding

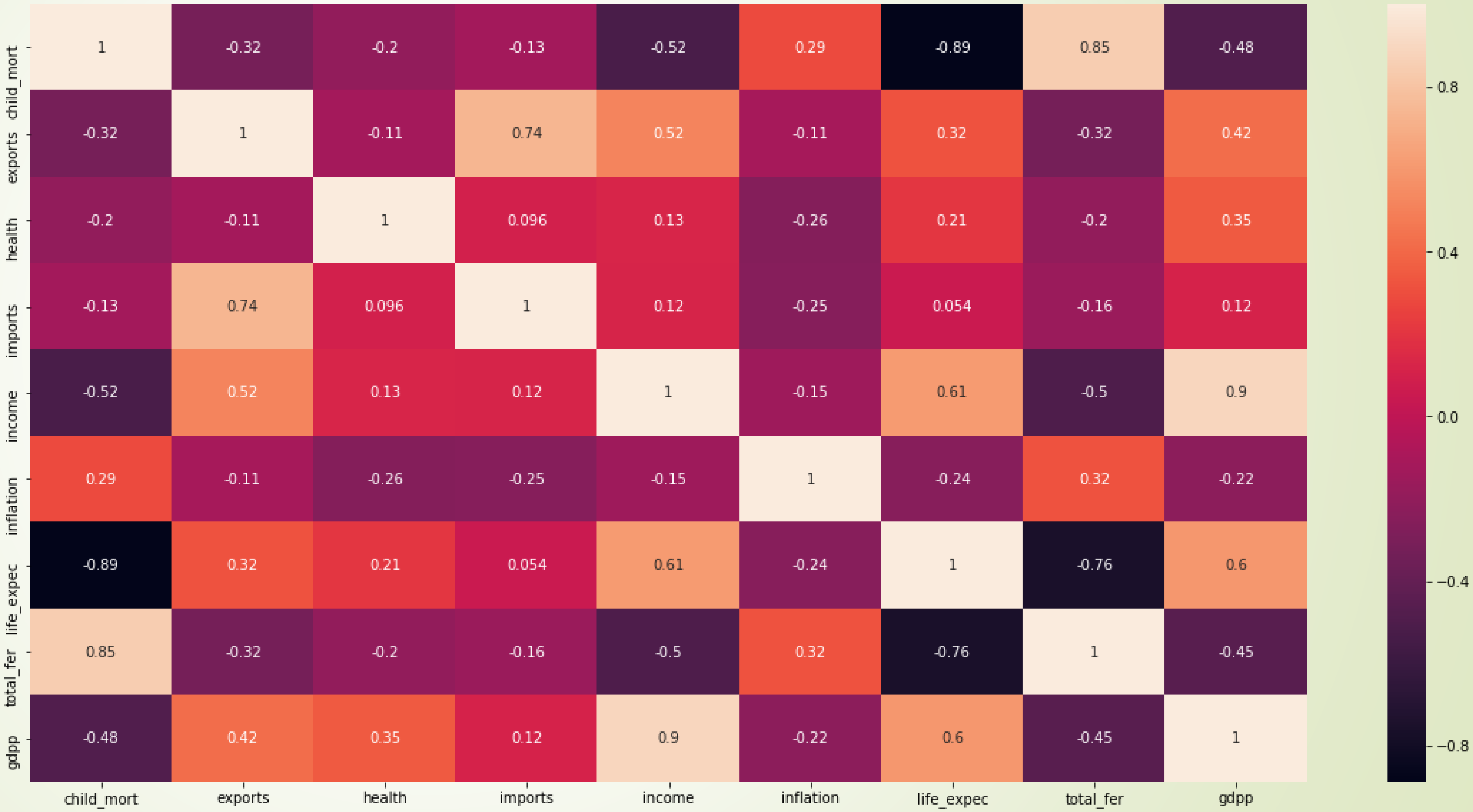
Perform Hierarchical Clustering

Principal Component Analysis

There are high multicollinearity

between variables.

We performed PCA to reduce multicollinearity between variables.

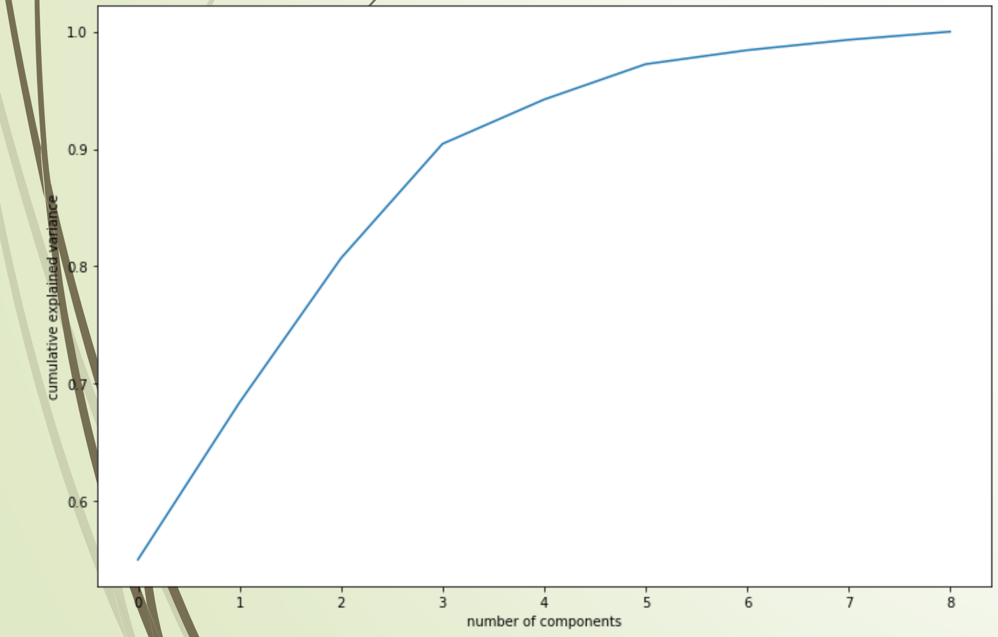


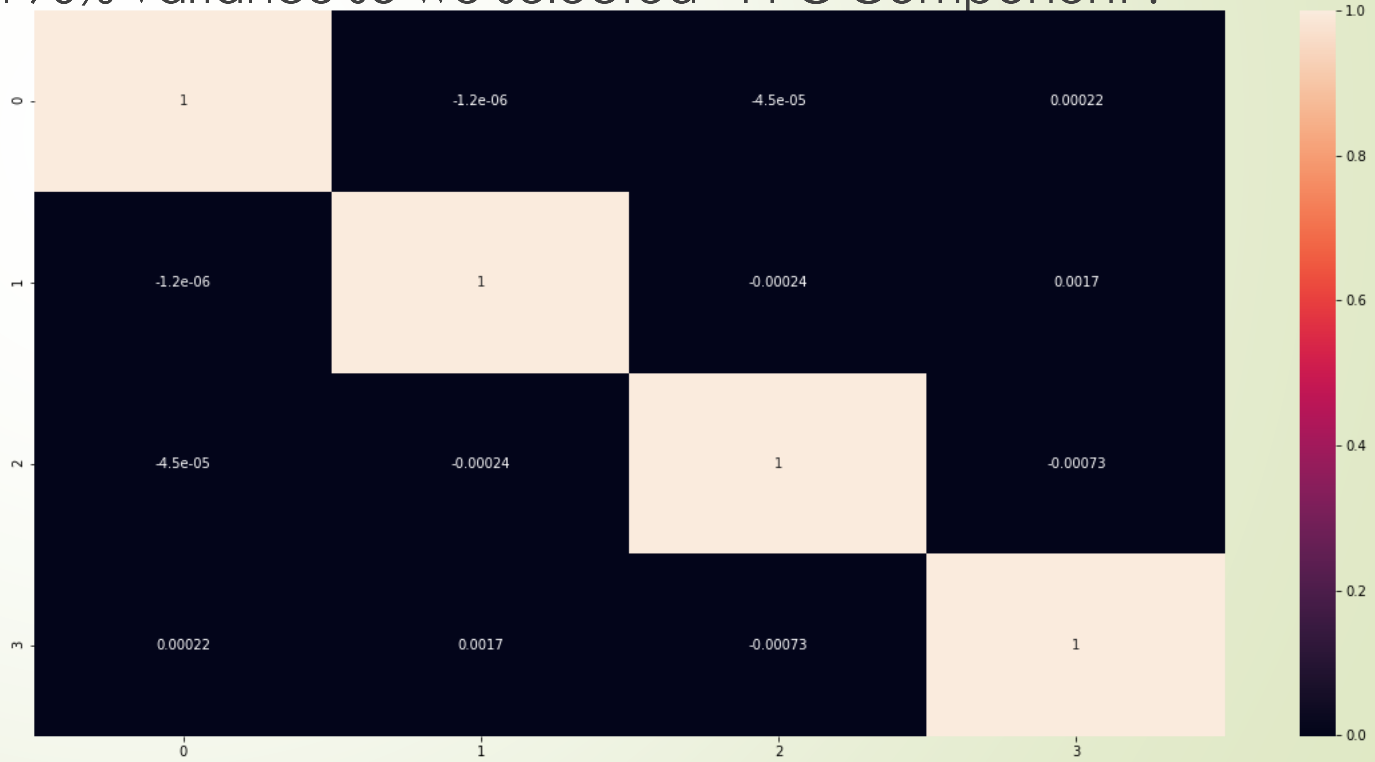
	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
child_mort	1	-0.32	-0.2	-0.13	-0.52	0.29	-0.89	0.85	-0.48
exports	-0.32	1	-0.11	0.74	0.52	-0.11	0.32	-0.32	0.42
health	-0.2	-0.11	1	0.096	0.13	-0.26	0.21	-0.2	0.35
imports	-0.13	0.74	0.096	1	0.12	-0.25	0.054	-0.16	0.12
income	-0.52	0.52	0.13	0.12	1	-0.15	0.61	-0.5	0.9
inflation	0.29	-0.11	-0.26	-0.25	-0.15	1	-0.24	0.32	-0.22
life_expec	-0.89	0.32	0.21	0.054	0.61	-0.24	1	-0.76	0.6
total_fer	0.85	-0.32	-0.2	-0.16	-0.5	0.32	-0.76	1	-0.45
gdpp	-0.48	0.42	0.35	0.12	0.9	-0.22	0.6	-0.45	1

<PCA Results>

After PCA , correlation between PCA variable was almost 0.

PC=4 was able to explain 90% variance so we selected 4 PC Component .

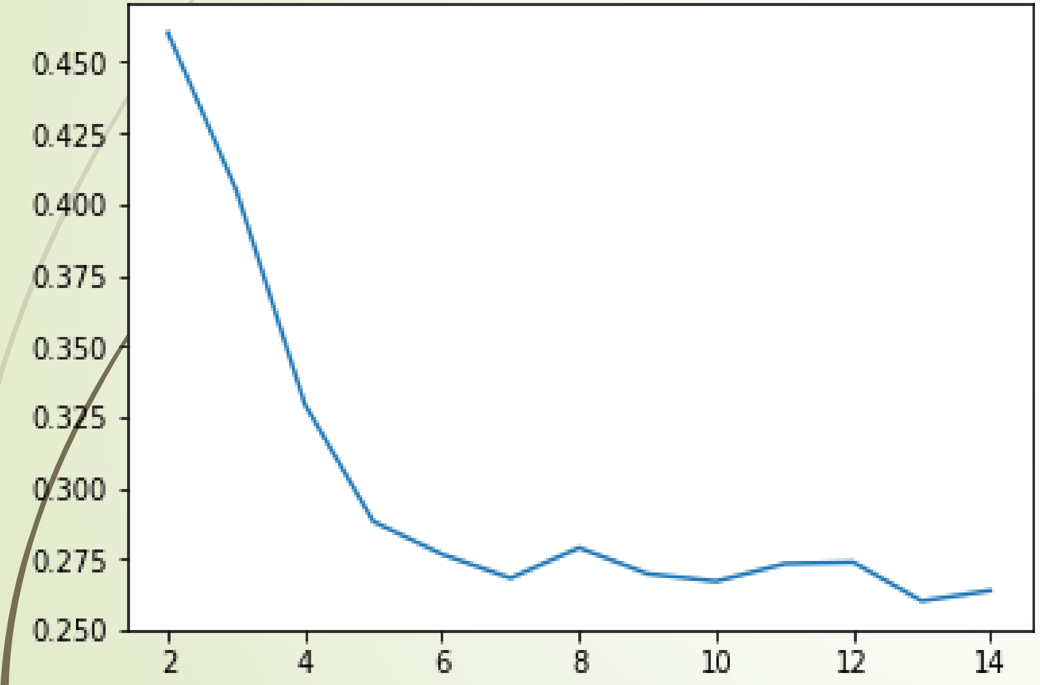




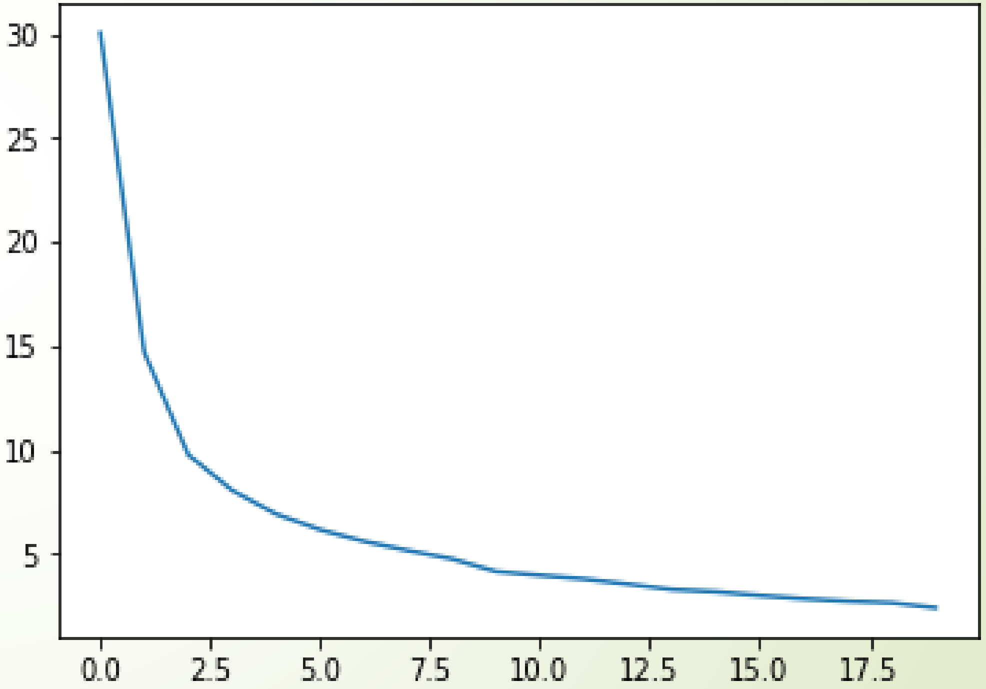
	0	1	2	3
0	1	-1.2e-06	-4.5e-05	0.00022
1	-1.2e-06	1	-0.00024	0.0017
2	-4.5e-05	-0.00024	1	-0.00073
3	0.00022	0.0017	-0.00073	1

K – Means Cluster

As per Silhouette Analysis and SSD Graph Optimum value for K was between 4 to 5



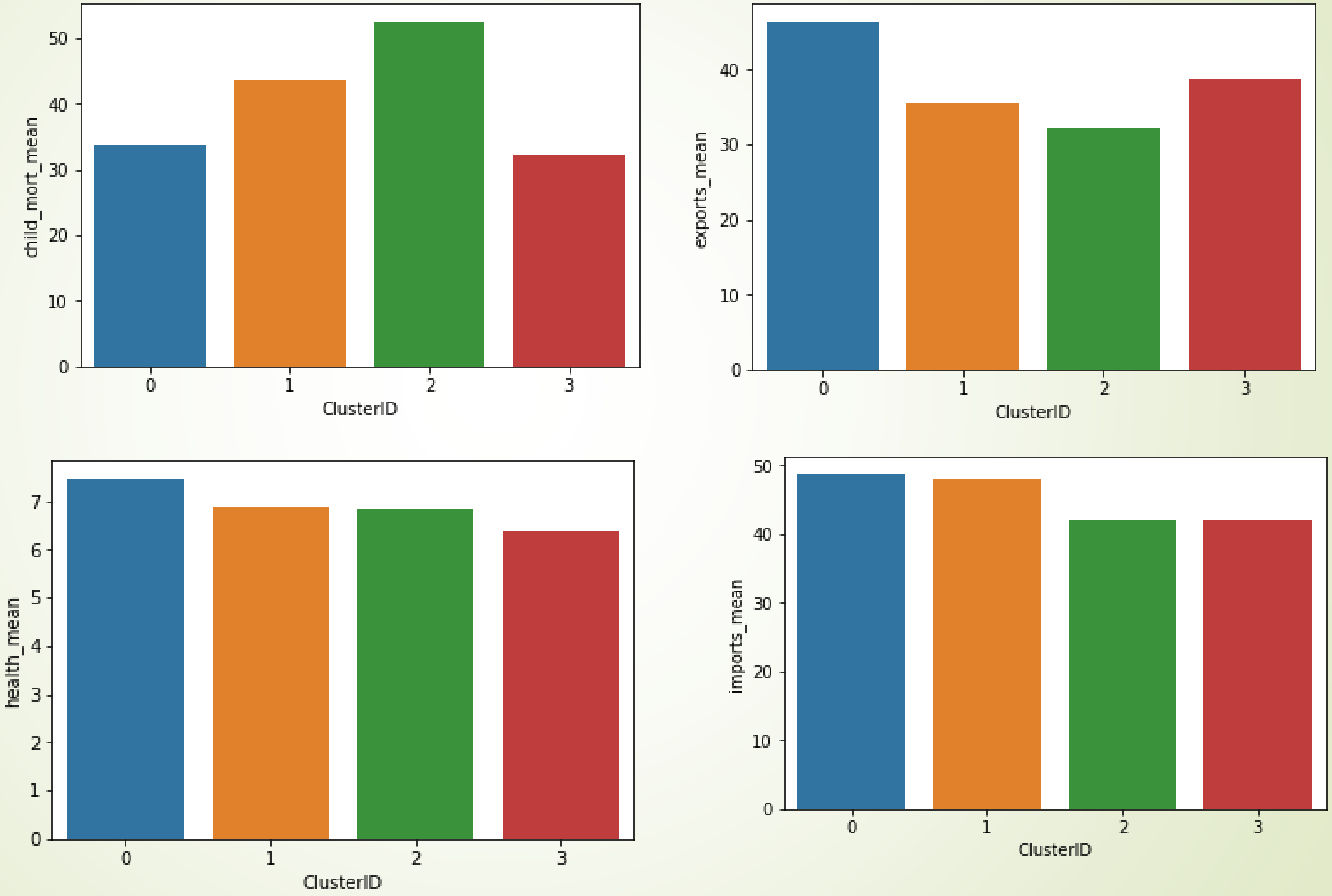
Silhouette Analysis



Shortest Sum of Distance

<Results>

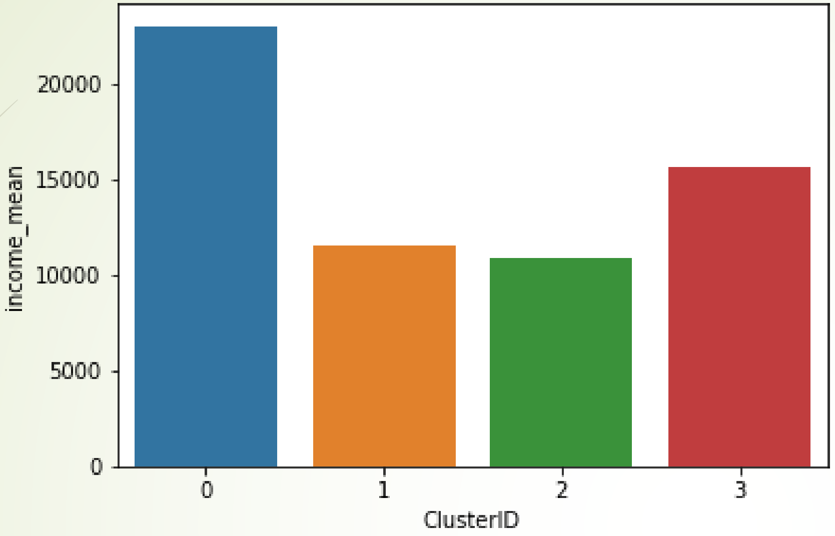
We created Cluster for Value K=4 and total 4 clusters. We see that cluster 2 has the highest child mortality rate and least exports.

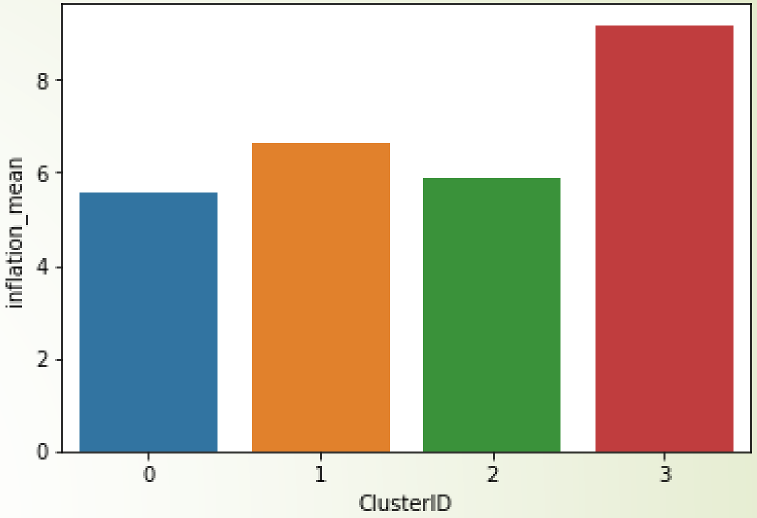


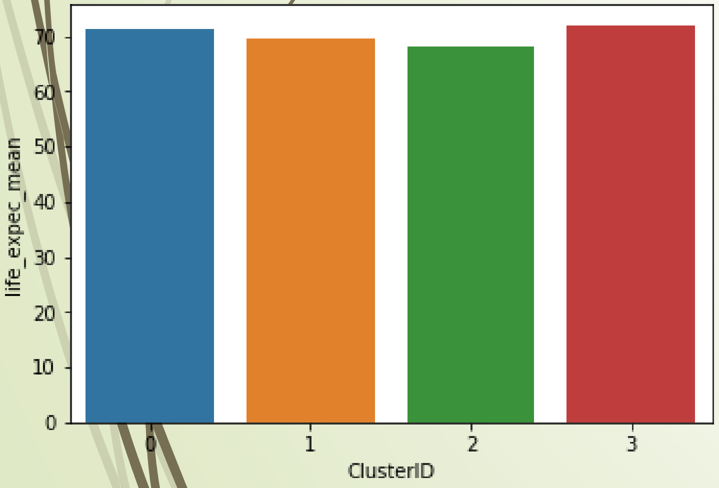
<Results>

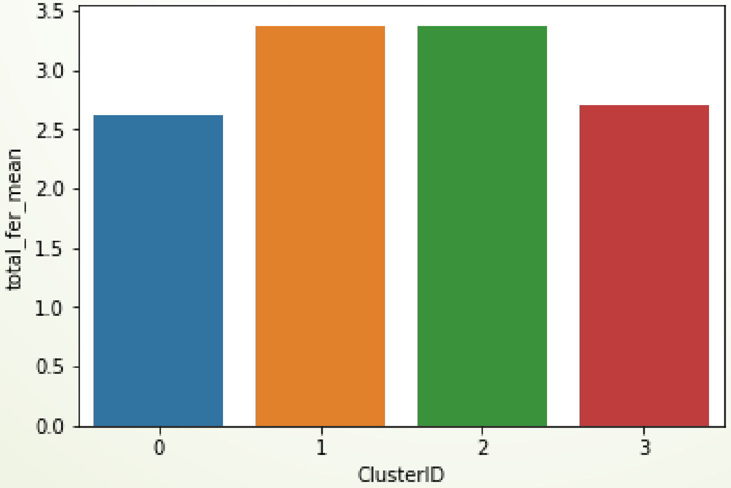
Cluster 2 has lowest income and lowest GDP.

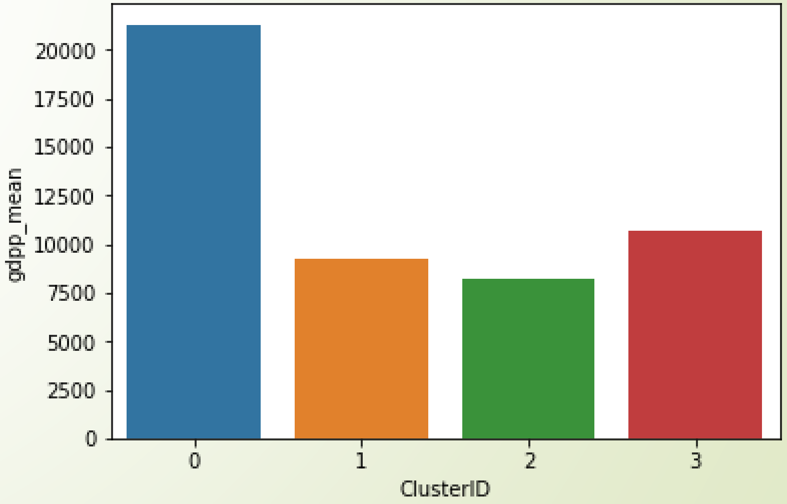
Thus cluster 2 comes to be the group of countries which needs aid











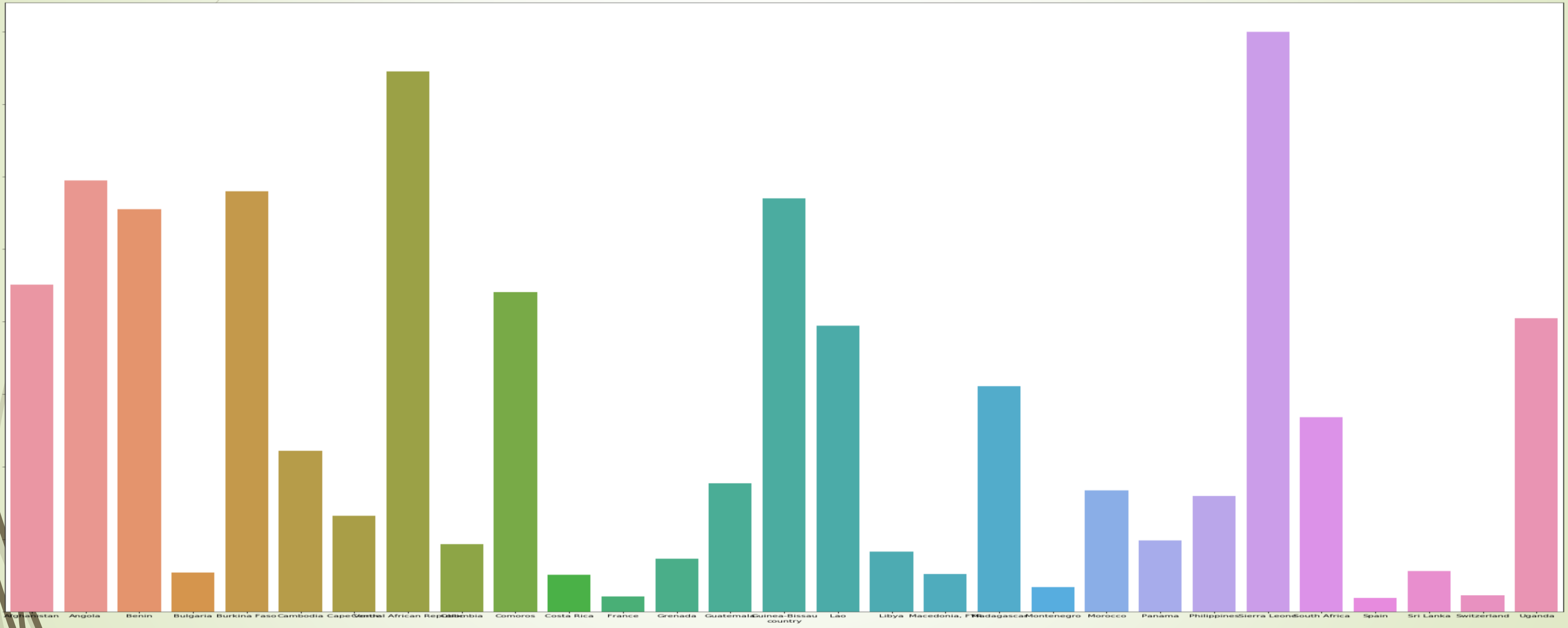
Results

List of Cluster 2 countries
We have few countries like Spain, France and Switzerland, which are actually rich countries but are appearing in cluster 2.

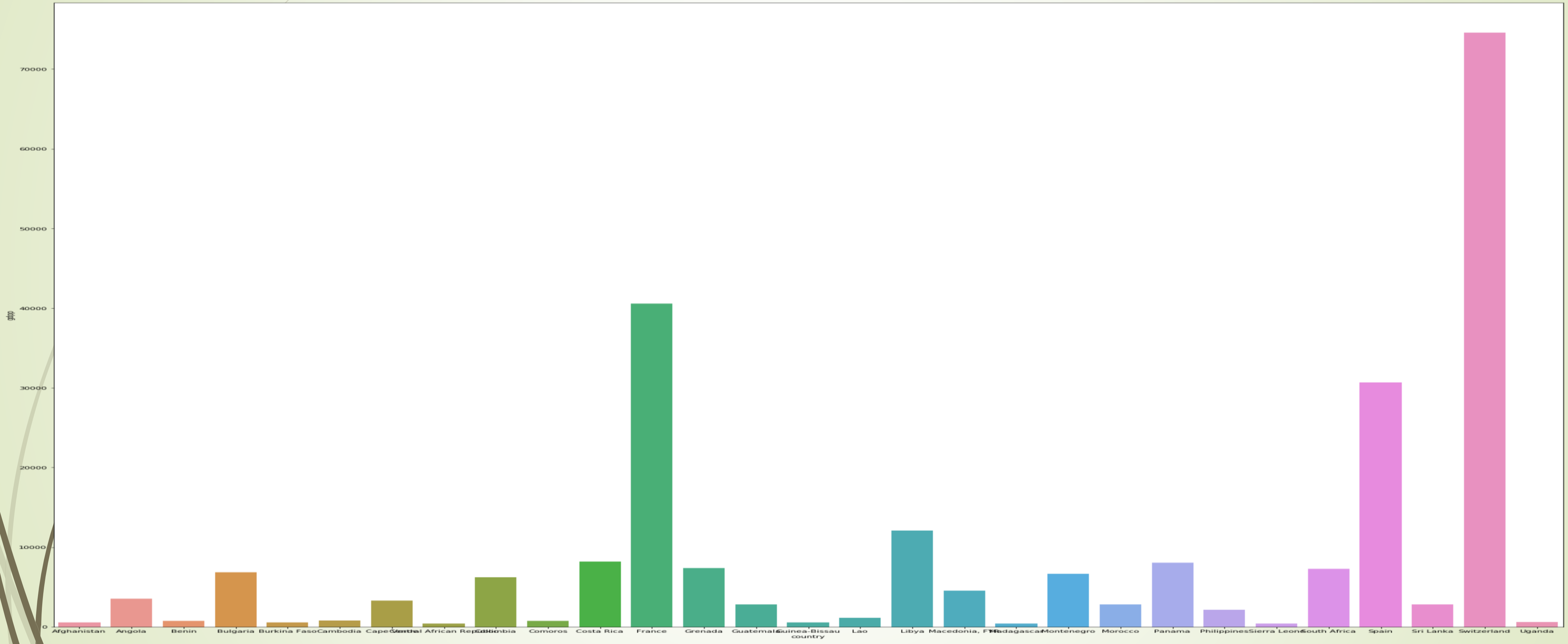
So we are deciding to go for 5 Clusters to see if we can overcome these exceptions.

country	income	gdpp	child_mort
Afghanistan	1610	553	90.2
Angola	5900	3530	119
Benin	1820	758	111
Bulgaria	15300	6840	10.8
Burkina Faso	1430	575	116
Cambodia	2520	786	44.4
Cape Verde	5830	3310	26.5
Central African Republic	888	446	149
Colombia	10900	6250	18.6
Comoros	1410	769	88.2
Costa Rica	13000	8200	10.2
France	36900	40600	4.2
Grenada	11200	7370	14.6
Guatemala	6710	2830	35.4
Guinea-Bissau	1390	547	114
Lao	3980	1140	78.9
Libya	29600	12100	16.6
Macedonia, FYR	11400	4540	10.4
Madagascar	1390	413	62.2
Montenegro	14000	6680	6.8
Morocco	6440	2830	33.5
Panama	15400	8080	19.7
Philippines	5600	2130	31.9
Sierra Leone	1220	399	160
South Africa	12000	7280	53.7
Spain	32500	30700	3.8
Sri Lanka	8560	2810	11.2
Switzerland	55500	74600	4.5
Uganda	1540	595	81

<Results – Plot of Cluster 2 countries against child mortality>

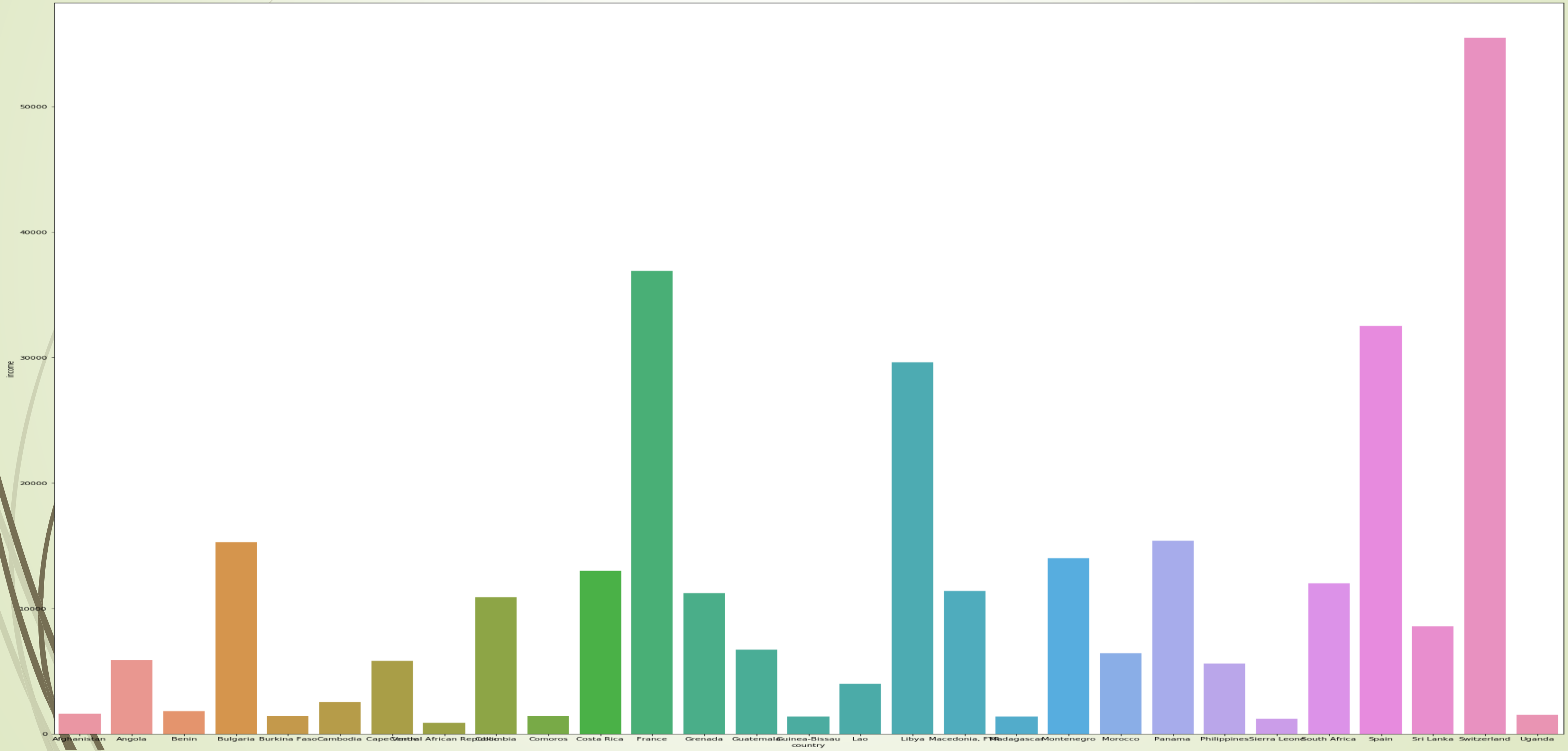

child_mort
0
20
40
60
80
100
120
140
160
Afghanistan
Angola
Benin
Bulgaria
Burkina Faso
Cambodia
Cape Verde
Central African Republic
Colombia
Comoros
Costa Rica
France
Grenada
Guatemala
Guinea-Bissau
Lao
Libya
Macedonia, FYR
Madagascar
Montenegro
Morocco
Panama
Philippines
Sierra Leone
South Africa
Spain
Sri Lanka
Switzerland
Uganda
country

<Results – Plot of Cluster 2 countries against gdpp>



gdpp
70000
60000
50000
40000
30000
20000
10000
0
Afghanistan
Angola
Benin
Bulgaria
Burkina Faso
Cambodia
Cape Verde
Central African Republic
Colombia
Comoros
Costa Rica
France
Grenada
Guatemala
Guinea-Bissau
Lao
Libya
Macedonia, FYR
Madagascar
Montenegro
Morocco
Panama
Philippines
Sierra Leone
South Africa
Spain
Sri Lanka
Switzerland
Uganda
country

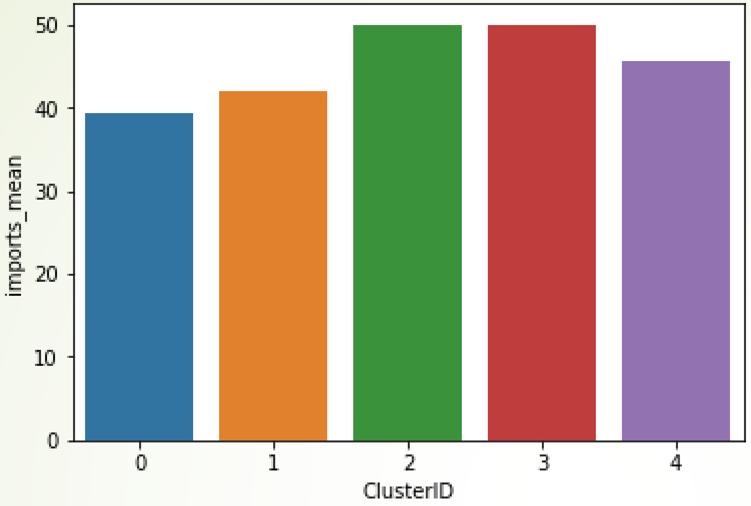
<Results – Plot of Cluster 2 countries against income>

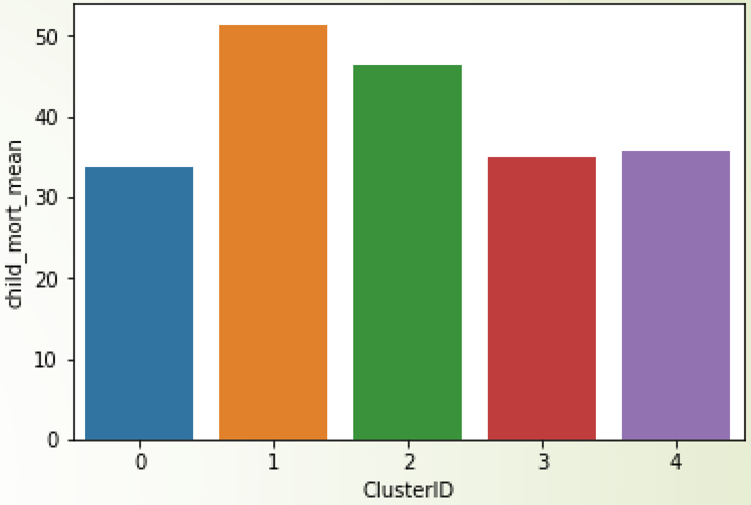


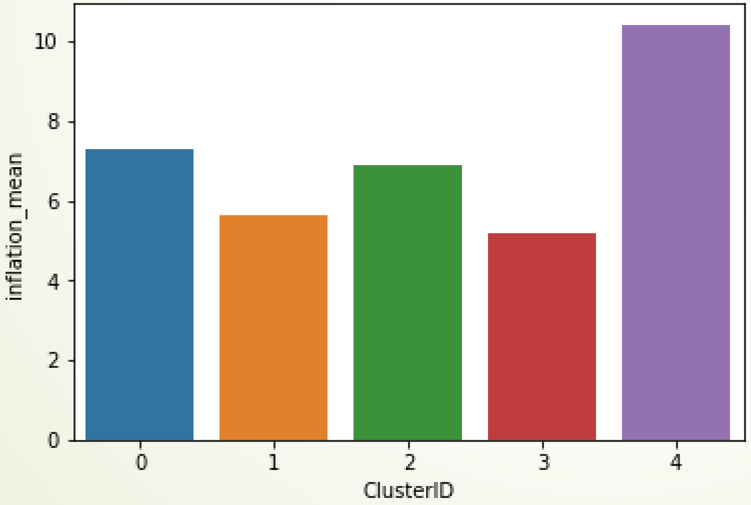
income
50000
40000
30000
20000
10000
0
Afghanistan
Angola
Benin
Bulgaria
Burkina Faso
Cambodia
Cape Verde
Central African Republic
Colombia
Comoros
Costa Rica
France
Grenada
Guatemala
Guinea-Bissau
Lao
Libya
Macedonia, FYR
Madagascar
Montenegro
Morocco
Panama
Philippines
Sierra Leone
South Africa
Spain
Sri Lanka
Switzerland
Uganda
country

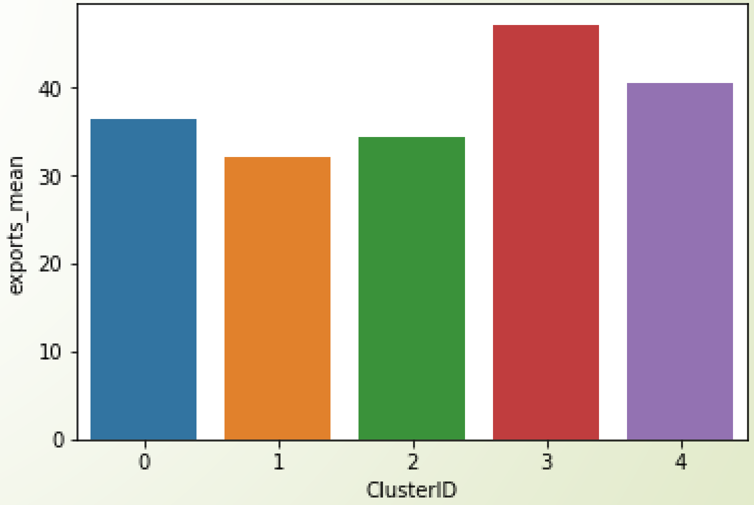
<Results>

We created Cluster for Value K=5 and total 5 clusters. We see that cluster 2 has the highest child mortality rate and least exports.





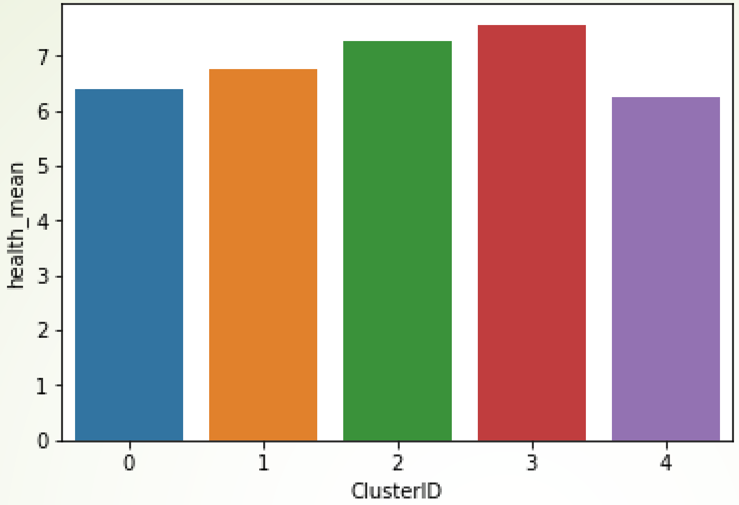


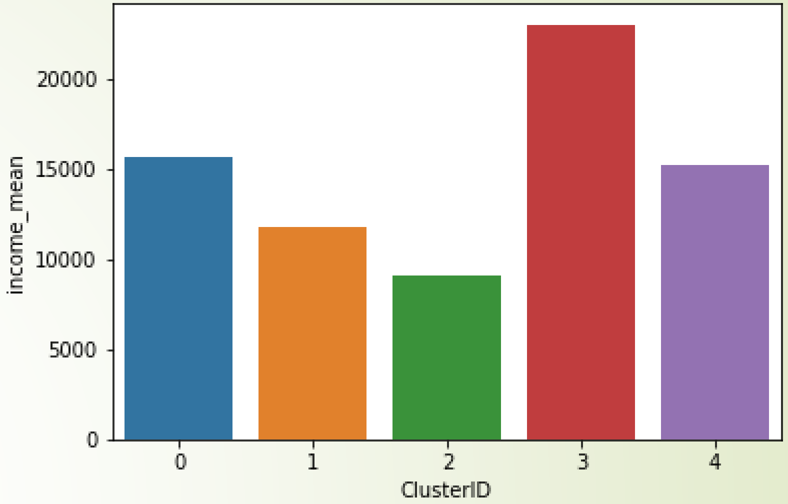


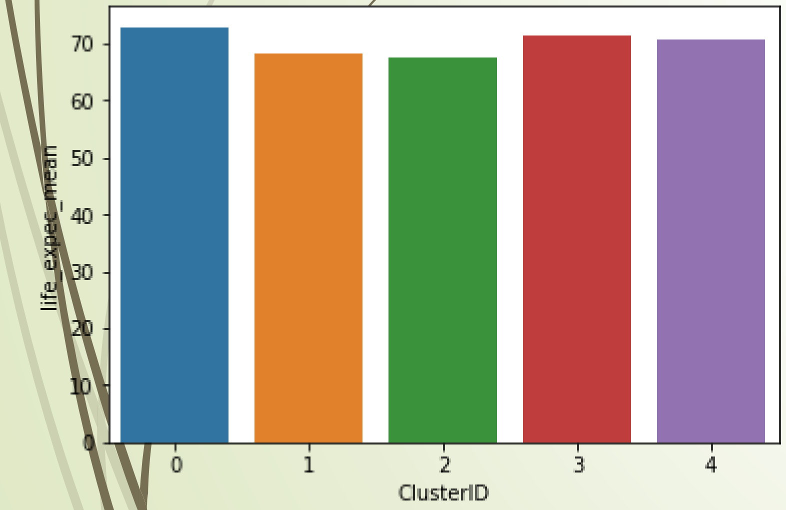
<Results>

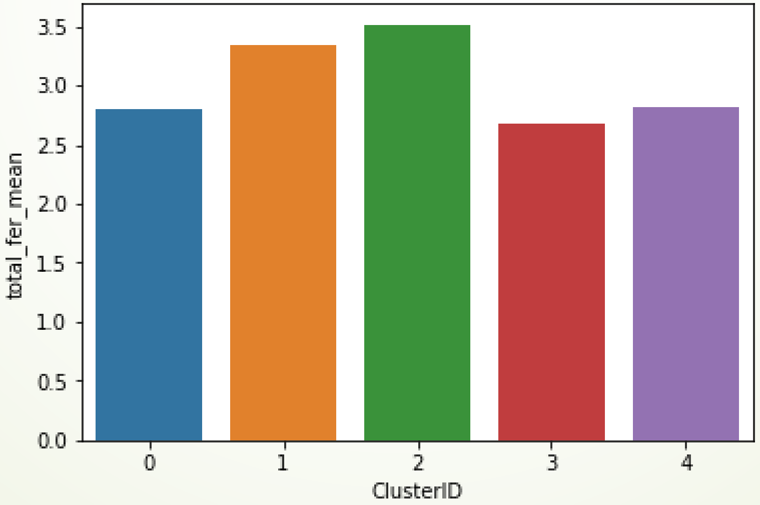
Cluster 2 has lowest income and lowest GDP.

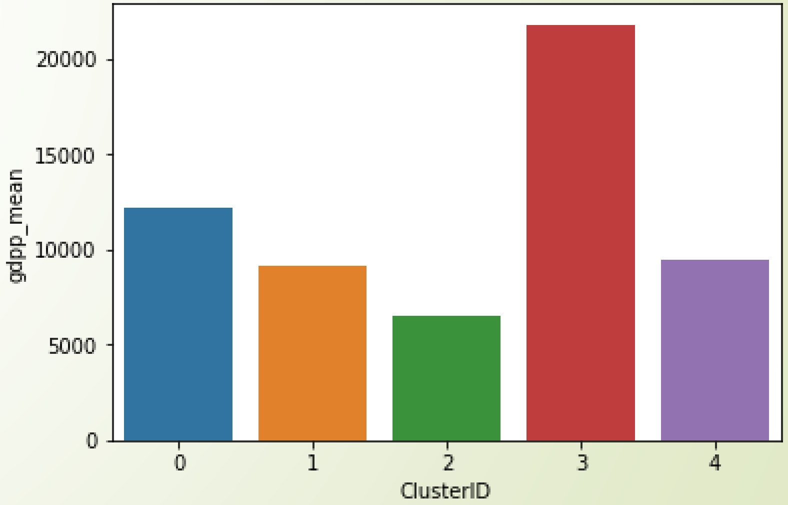
Thus cluster 2 comes to be the group of countries which needs aid











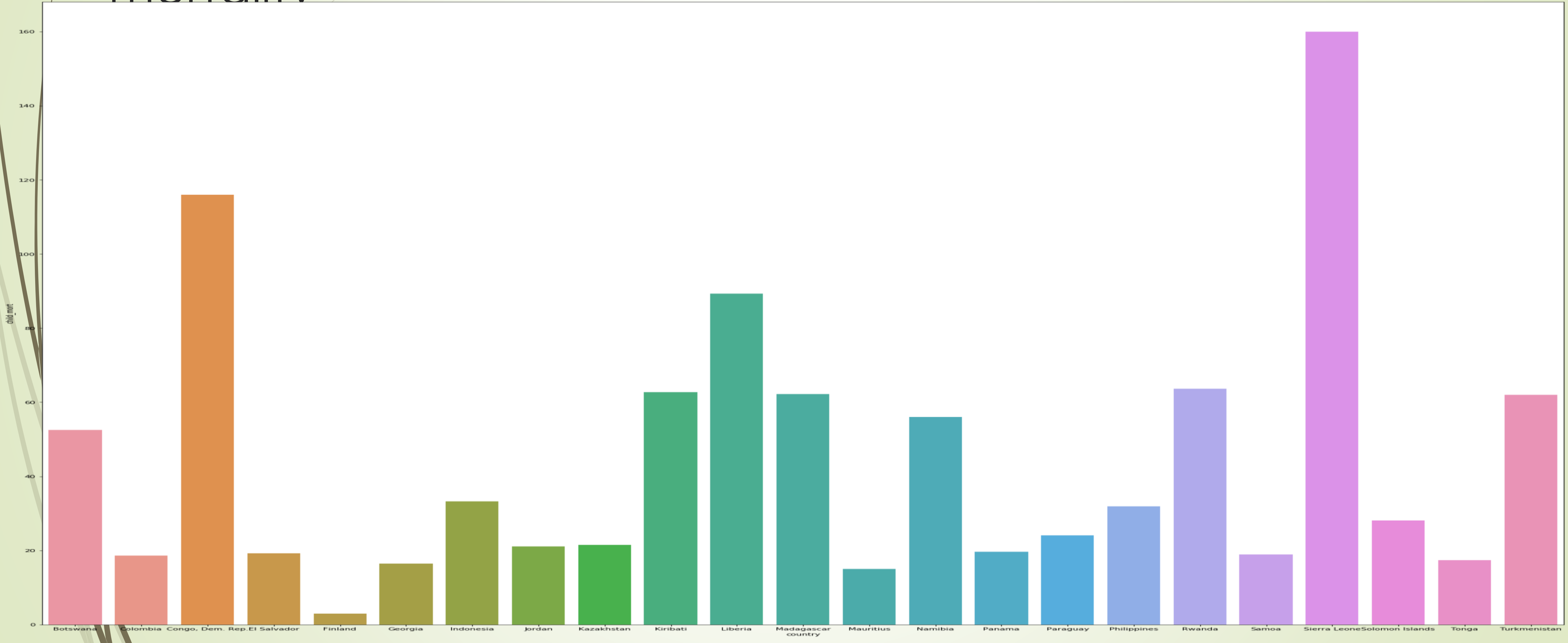
Results

List of Cluster 2 countries

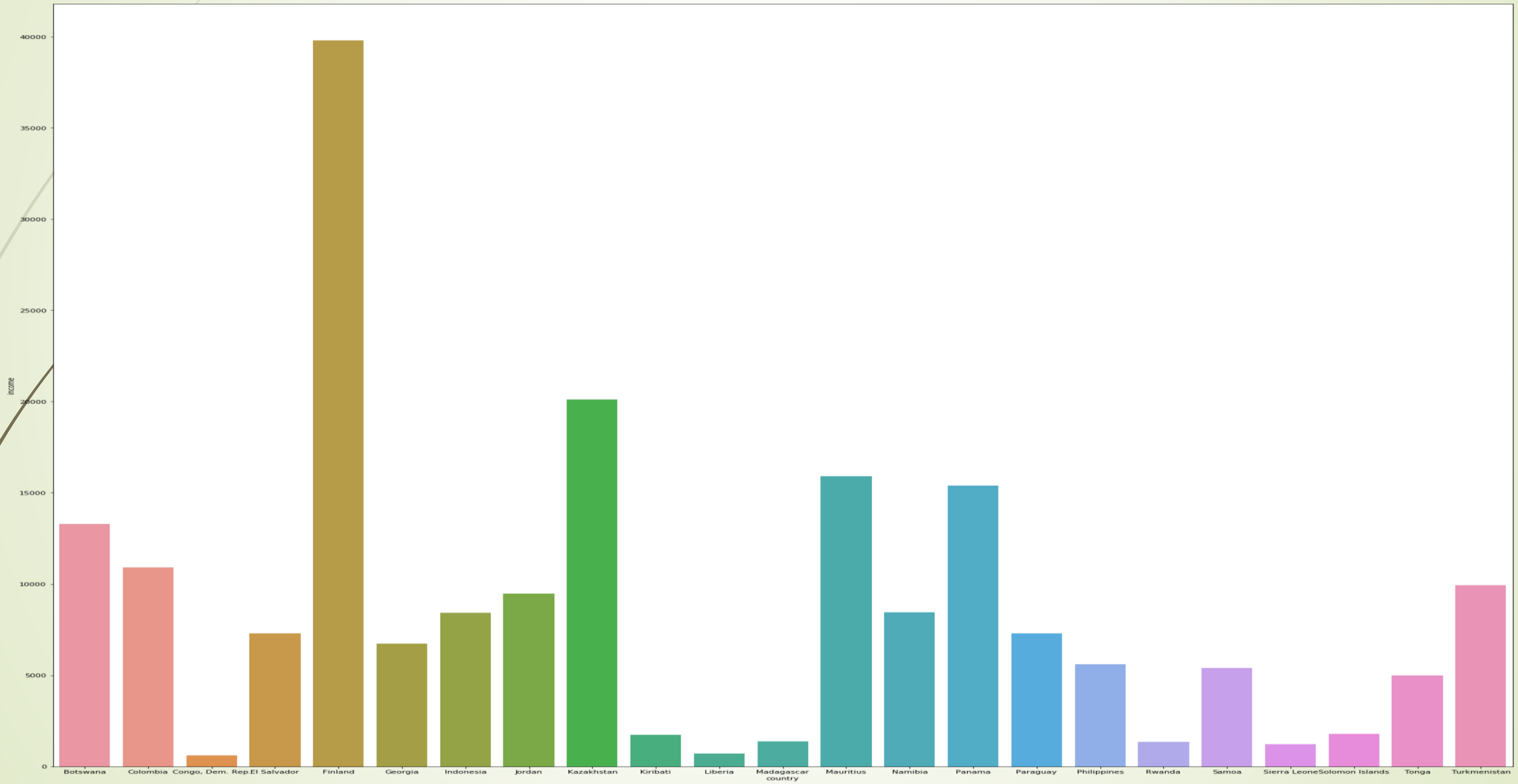
So we are deciding to go for 2 Clusters to see if we can overcome these exceptions.

country	income	gdpp	child_mort
Botswana	13300	6350	52.5
Colombia	10900	6250	18.6
Congo, Dem. Rep.	609	334	116
El Salvador	7300	2990	19.2
Finland	39800	46200	3
Georgia	6730	2960	16.5
Indonesia	8430	3110	33.3
Jordan	9470	3680	21.1
Kazakhstan	20100	9070	21.5
Kiribati	1730	1490	62.7
Liberia	700	327	89.3
Madagascar	1390	413	62.2
Mauritius	15900	8000	15
Namibia	8460	5190	56
Panama	15400	8080	19.7
Paraguay	7290	3230	24.1
Philippines	5600	2130	31.9
Rwanda	1350	563	63.6
Samoa	5400	3450	18.9
Sierra Leone	1220	399	160
Solomon Islands	1780	1290	28.1
Tonga	4980	3550	17.4
Turkmenistan	9940	4440	62

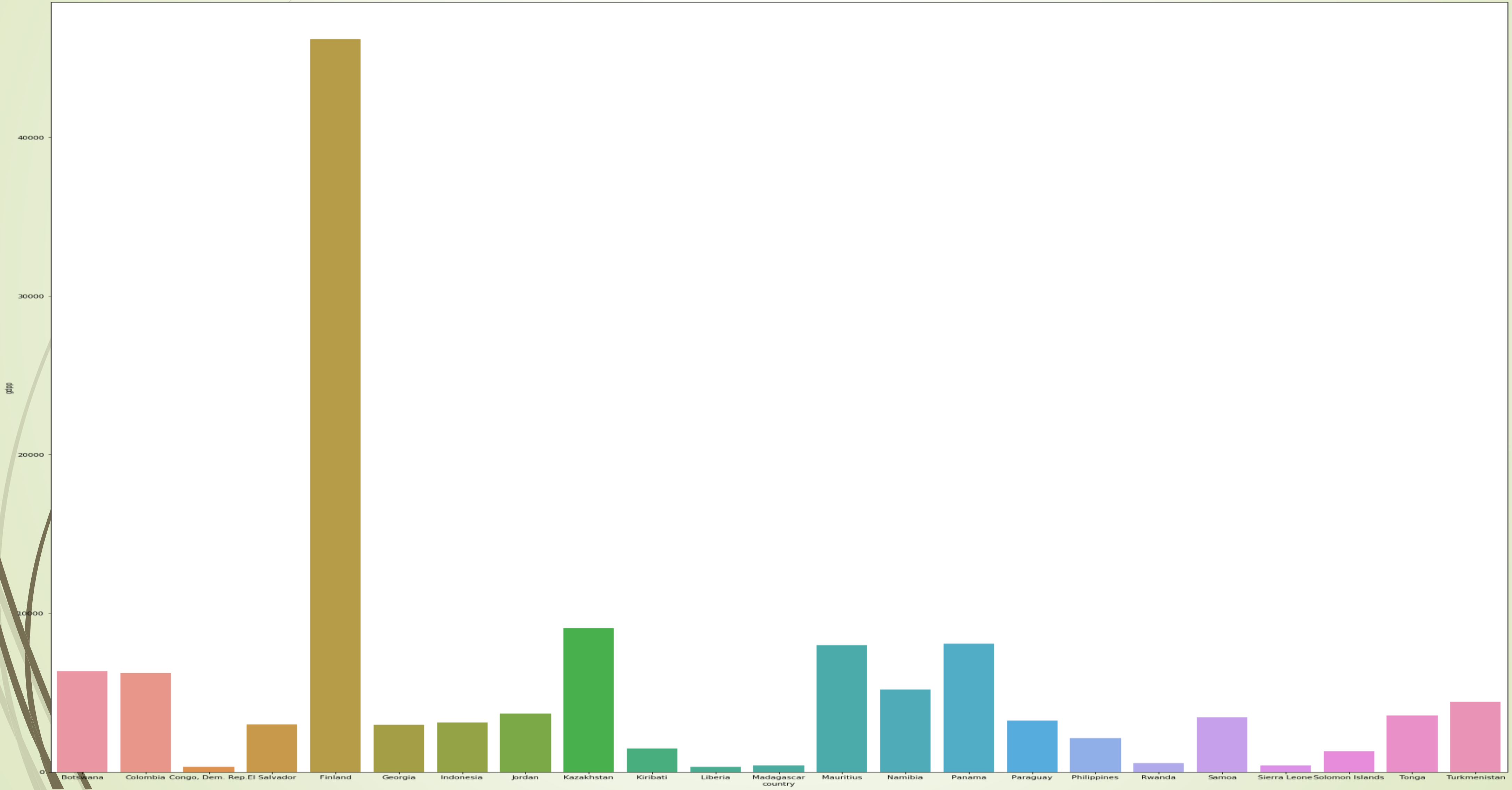
<Results – Plot of Cluster 2 countries against child mortality>


160
140
120
100
80
60
40
20
0
child_mort
Botswana
Colombia
Congo, Dem. Rep.
El Salvador
Finland
Georgia
Indonesia
Jordan
Kazakhstan
Kiribati
Liberia
Madagascar
Mauritius
Namibia
Panama
Paraguay
Philippines
Rwanda
Samoa
Sierra Leone
Solomon Islands
Tonga
Turkmenistan
country

<Results – Plot of Cluster 2 countries against income>

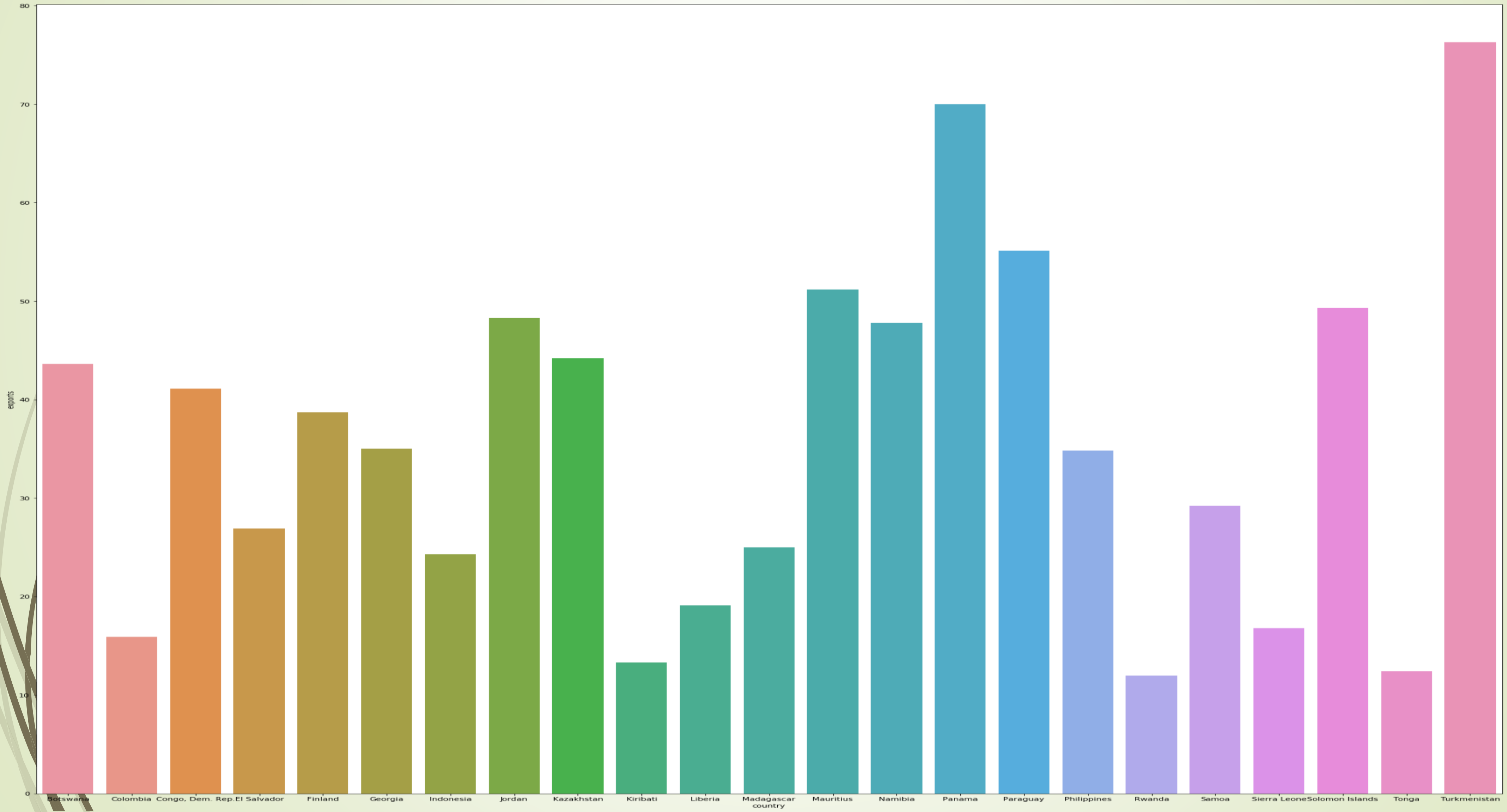

40000
35000
30000
25000
20000
15000
10000
5000
0
income
Botswana
Colombia
Congo, Dem. Rep.
El Salvador
Finland
Georgia
Indonesia
Jordan
Kazakhstan
Kiribati
Liberia
Madagascar
Mauritius
Namibia
Panama
Paraguay
Philippines
Rwanda
Samoa
Sierra Leone
Solomon Islands
Tonga
Turkmenistan
country

<Results – Plot of Cluster 2 countries against gdpp>



gdpp
0
10000
20000
30000
40000
Botswana
Colombia
Congo, Dem. Rep.
El Salvador
Finland
Georgia
Indonesia
Jordan
Kazakhstan
Kiribati
Liberia
Madagascar
Mauritius
Namibia
Panama
Paraguay
Philippines
Rwanda
Samoa
Sierra Leone
Solomon Islands
Tonga
Turkmenistan
country

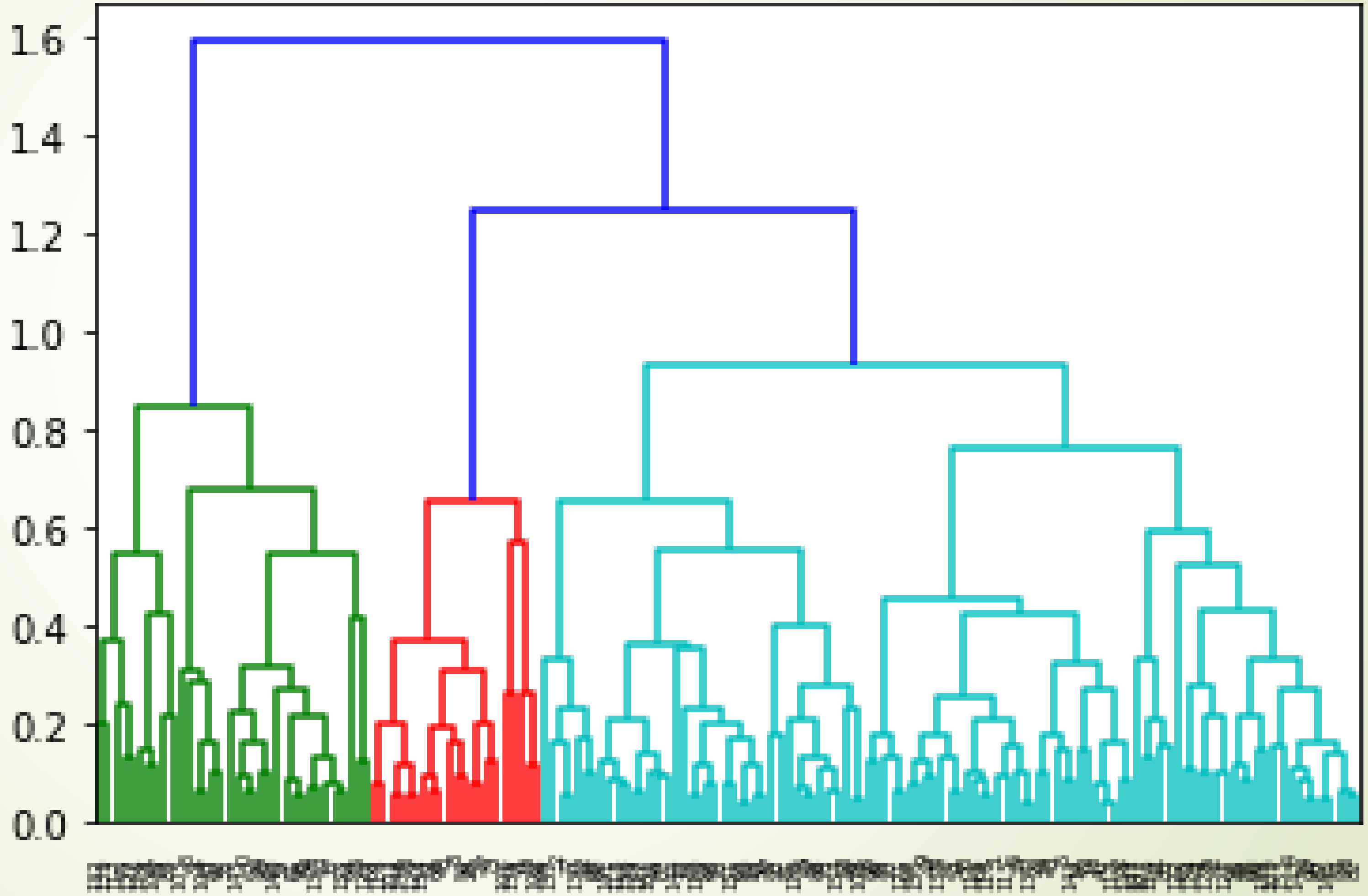
<Results – Plot of Cluster 2 countries against exports>



exports
0
10
20
30
40
50
60
70
80
Botswana
Colombia
Congo, Dem. Rep.
El Salvador
Finland
Georgia
Indonesia
Jordan
Kazakhstan
Kiribati
Liberia
Madagascar
Mauritius
Namibia
Panama
Paraguay
Philippines
Rwanda
Samoa
Sierra Leone
Solomon Islands
Tonga
Turkmenistan
country

Hierarchical Clustering - Dendogram

This also suggest optimum clustering as 5. Here after performing clustering and mean analysis we get cluster 0 as countries to be provided as aid.



Conclusion

After analysis of Socio-Economic and Health data for various countries we came up with Five Clusters. Out of these five clusters, Cluster Zero are group of countries which are underdeveloped and are in dire need of Financial Aid.

These countries are as follows:
Botswana, Colombia, Congo, Dem. Rep., El Salvador, Finland, Georgia, Indonesia, Jordan, Kazakhstan, Kiribati, Liberia, Madagascar, Mauritius, Namibia, Panama, Paraguay, Philippines, Rwanda, Samoa, Sierra Leone, Solomon Islands, Tonga, Turkmenistan

There can be a few exception in the list of countries as these are clusters so individual Socio-economic data also needs to be considered before providing financial aid.