ANALYSIS & RECOMMENDATION OF COUNTRIES

For HELP INTERNATIONAL

By Abhishek Sinha

Abstract

Objective

We, HELP International, 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. We runs a lot of operational projects from time to time along with advocacy drives to raise awareness as well as for funding purposes.

Problem Statement

One common problem which we face every time is choosing the countries that are in the direst need of aid.

During the recent funding programmes, we have been able to raise around $ 10 million and now we want to utilise this funding in best possible way i.e. helping the countries who needs aid.

As a data analyst team, we have done extensive analysis and came up with recommendations on countries which need support & aid and where these funding can be utilised

Analysis Methodology - Flowchart

Data Collection and Understanding

- Importing data
- Understanding the data
  - Checking rows & columns
  - Checking NULL
  - Numerical values analysis
- Converting required columns to actual values

Outliers analysis and treatment

- Checking Outliers using Boxplots
- Removing outliers
  - Removing the countries which have high ‘gdpp' as those are developed countries with strong economy (> 95 percentile)

Visualising the data

- Checking Correlation between all the numerical features using:
  - Heatmaps
  - Pairplots

Hopkins Statistics

- To check if the provided data has tendency to form clusters, if so, how strong?

Scaling the data

- Standardising all the numerical features for unbiased clustering.

K-Means Clustering

- Identify ‘K' using SSD and Silhouette score
- Forming K clusters and assigning labels
- Analysing the countries assigned to each cluster
- Identifying the countries which requires aid

Analysis Methodology - Flowchart

Hierarchical Clustering

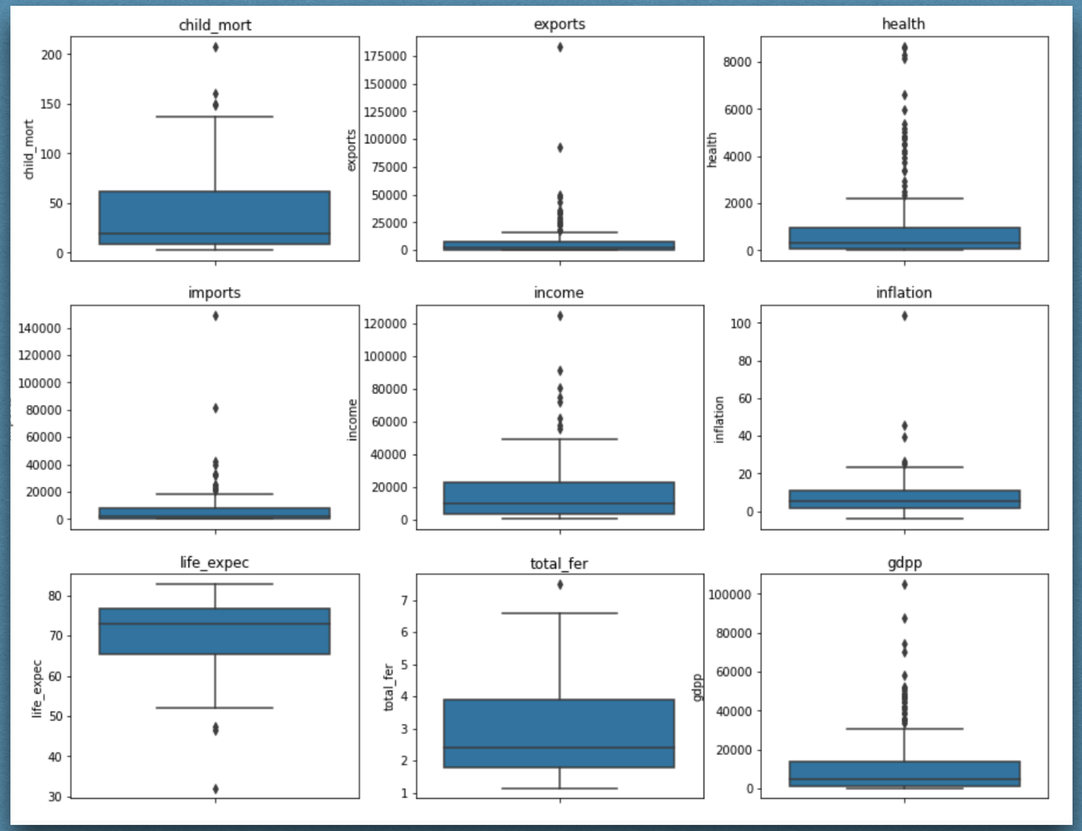
- Identify ’n’ via Dendogram
- Forming ’n’ clusters
- Analysing the countries assigned to each cluster
- Identifying the countries which requires aid

Final Recommendation of Countries

- Recommend Top 5 countries which are identified by both K-means and Hierarchical clustering, which requires aid.

Analysis Methodology - Visualisation

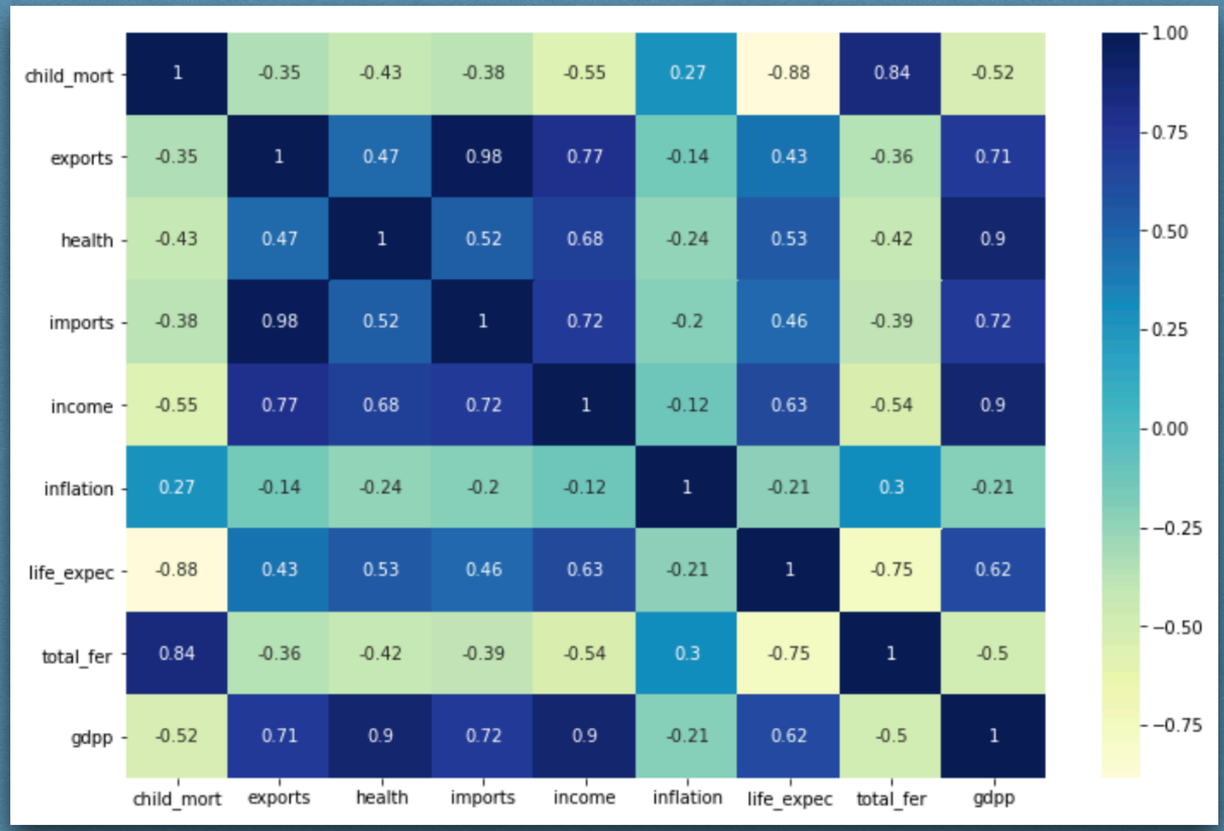
Outliers Analysis



We can see that there are Outliers present in almost all of the features.

But chose the feature, where outlier removal doesn't affect the overall objective of the solution i.e. finding the countries which needs aid. So, naturally the countries which have high `**gdpp**` may not need aid and support as they are highly developed countries with standard infrastructure in place to cope up with any natural calamities and disaster.

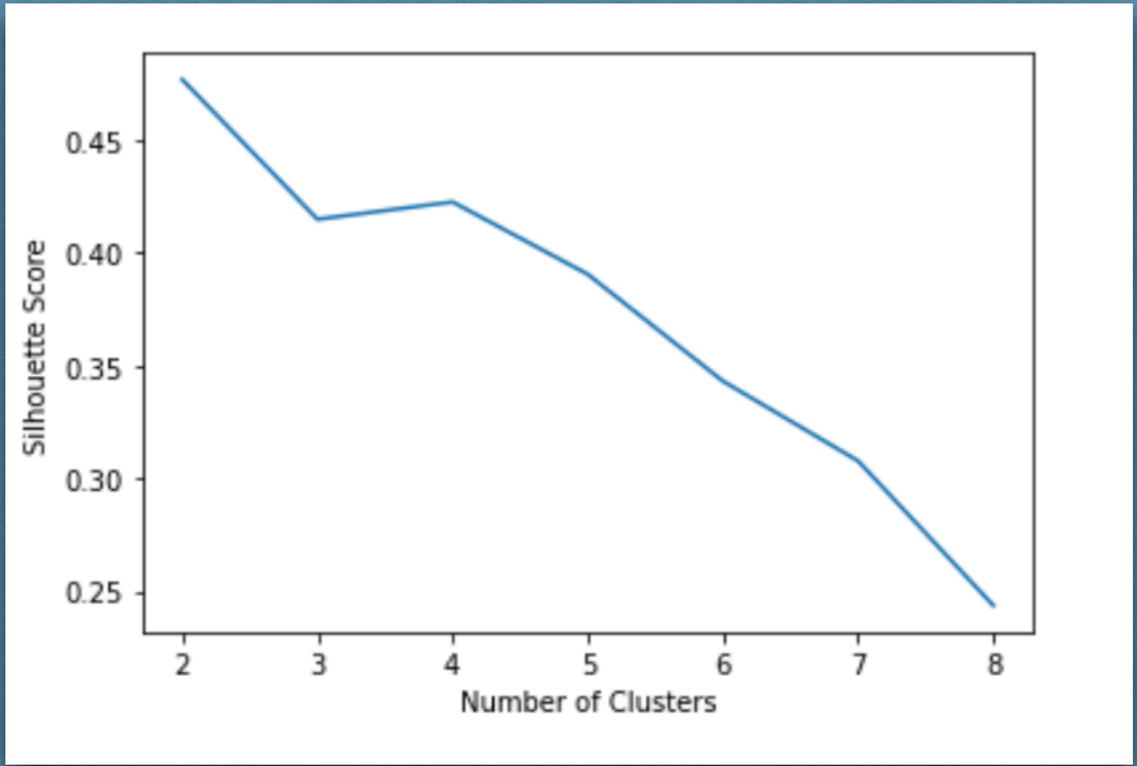
Correlation Check



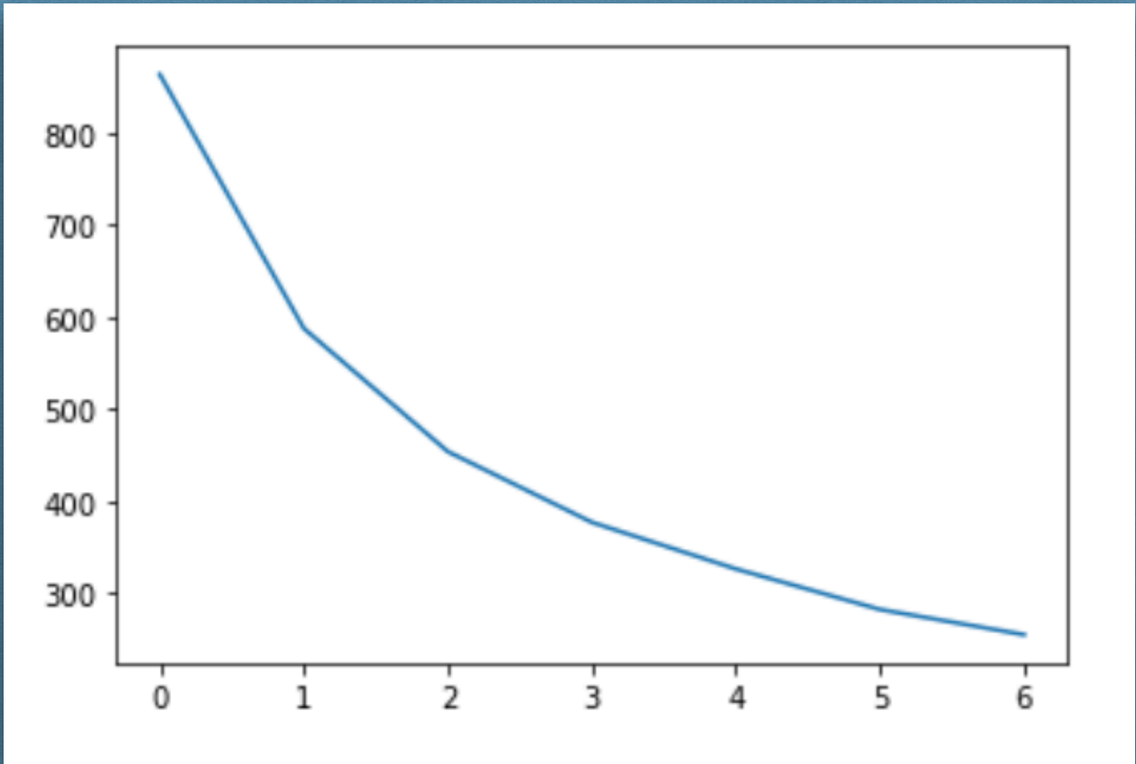
- After Cleaning data i.e. removing the outliers, we plotted the Heatmap to check the correlation among numerical features.
- Looking at the Heatmap we can see -
  - Income & gdpp has high correlation
  - child_mort and life_expec has high but negative correlation.
  - child_mort and total_fer has high correlation, which means countries which has high fertility rate have high child mortality as well.

K-Means Clustering

Silhouette Analysis

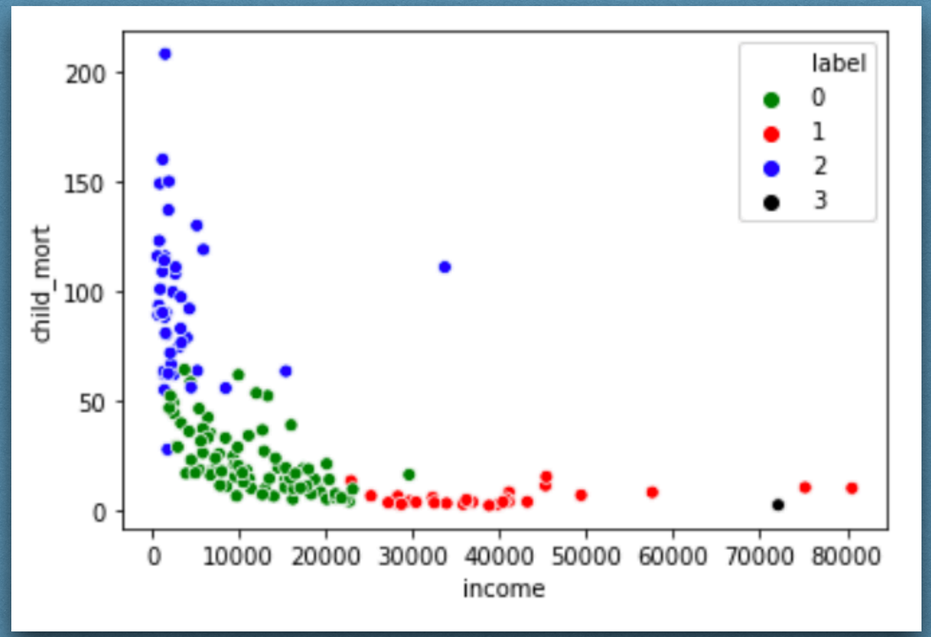


SSD / Elbow Curve Analysis

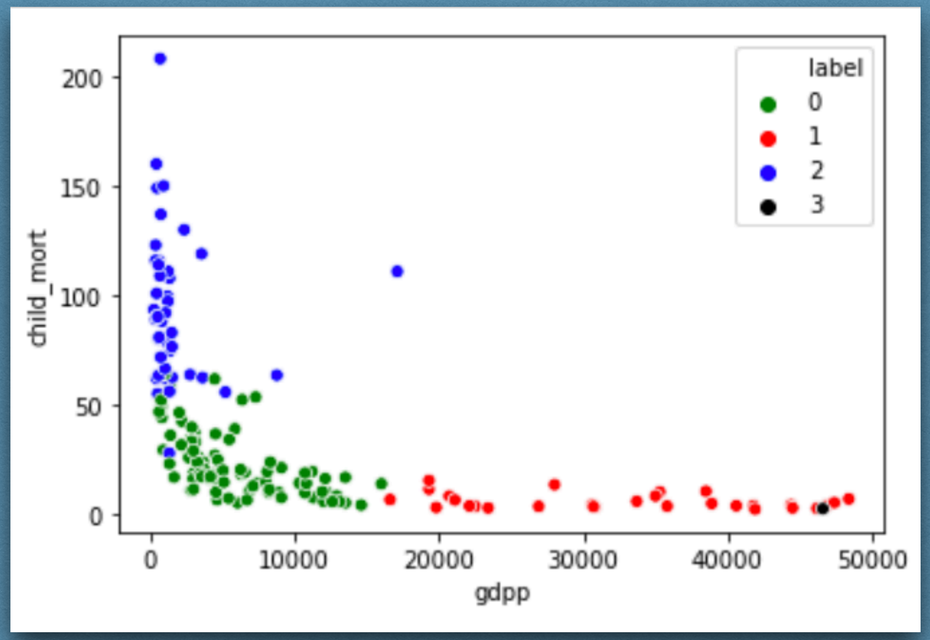


By looking above we can see that we have high value of Silhouette score at 4 & 5 and Elbow curve shows steep slope between 3 and 5, so we take 4 as optimum value (**K=4)**

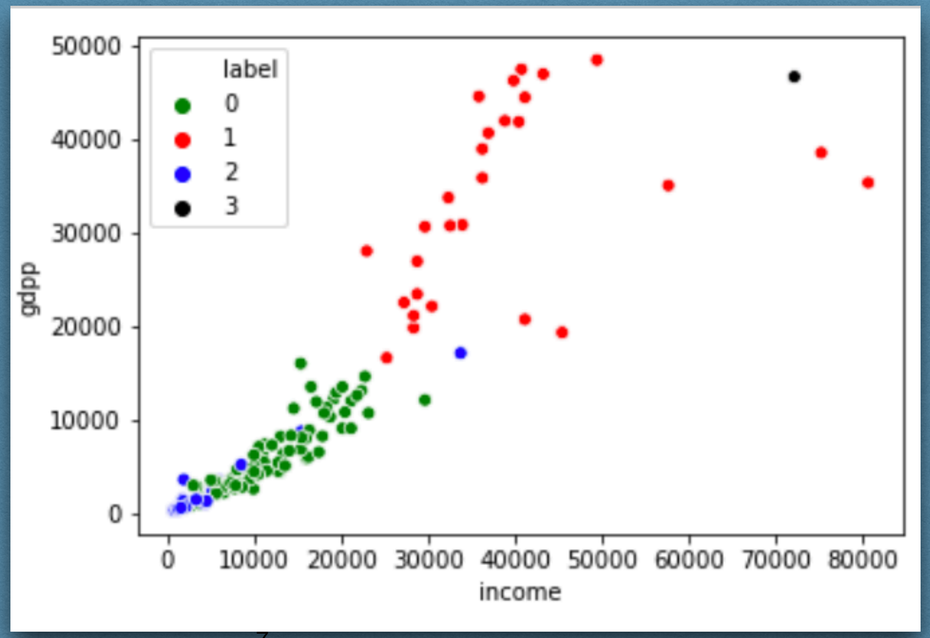
K-Means Clustering



We can see that countries which are grouped into **Cluster 2** have ***high child_mort*** and ***low income***



We can see that countries which are grouped into **Cluster 2** have ***high child_mort*** and ***low gdpp***

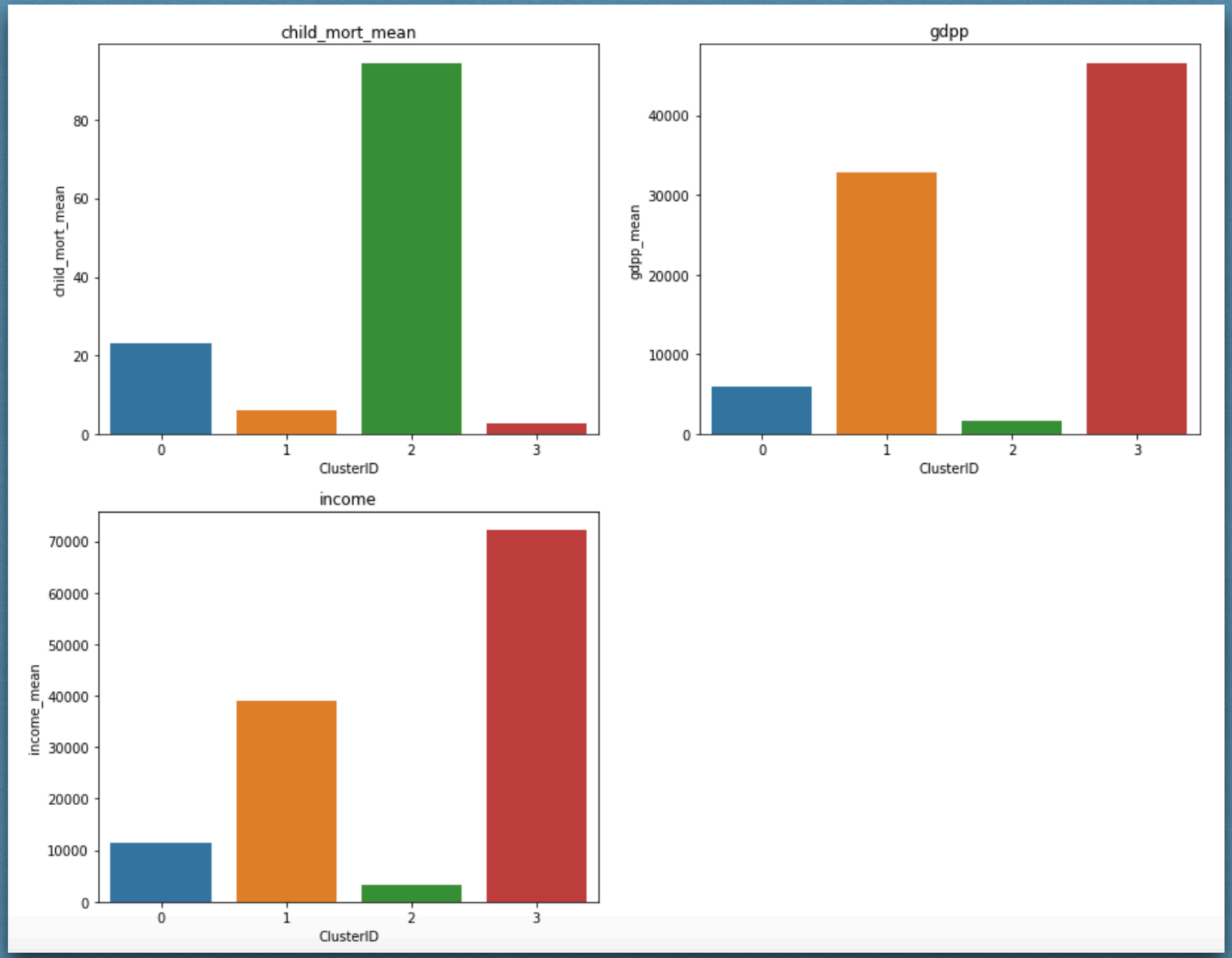


We can see that countries which are grouped into **Cluster 2** have ***low income*** and ***low gdpp***

K-Means Clustering

As per our K-Means Clustering, ***Cluster - 2*** contains group of countries which are in dire need of aid as they have -

- ***High*** Child Mortality rate
- ***Low*** GDP
- ***Low*** Income



K-Means Clustering

After Cluster profiling and result analysis, we concluded that all the countries belonging to ***Cluster ID - 2*** are in Dire need of aid.

We sorted the country in Cluster ID 2 based on the ***low*** **gdpp**, ***low*** **income** and ***high*** **child_mort** rate and below are list of Top 5 countries:

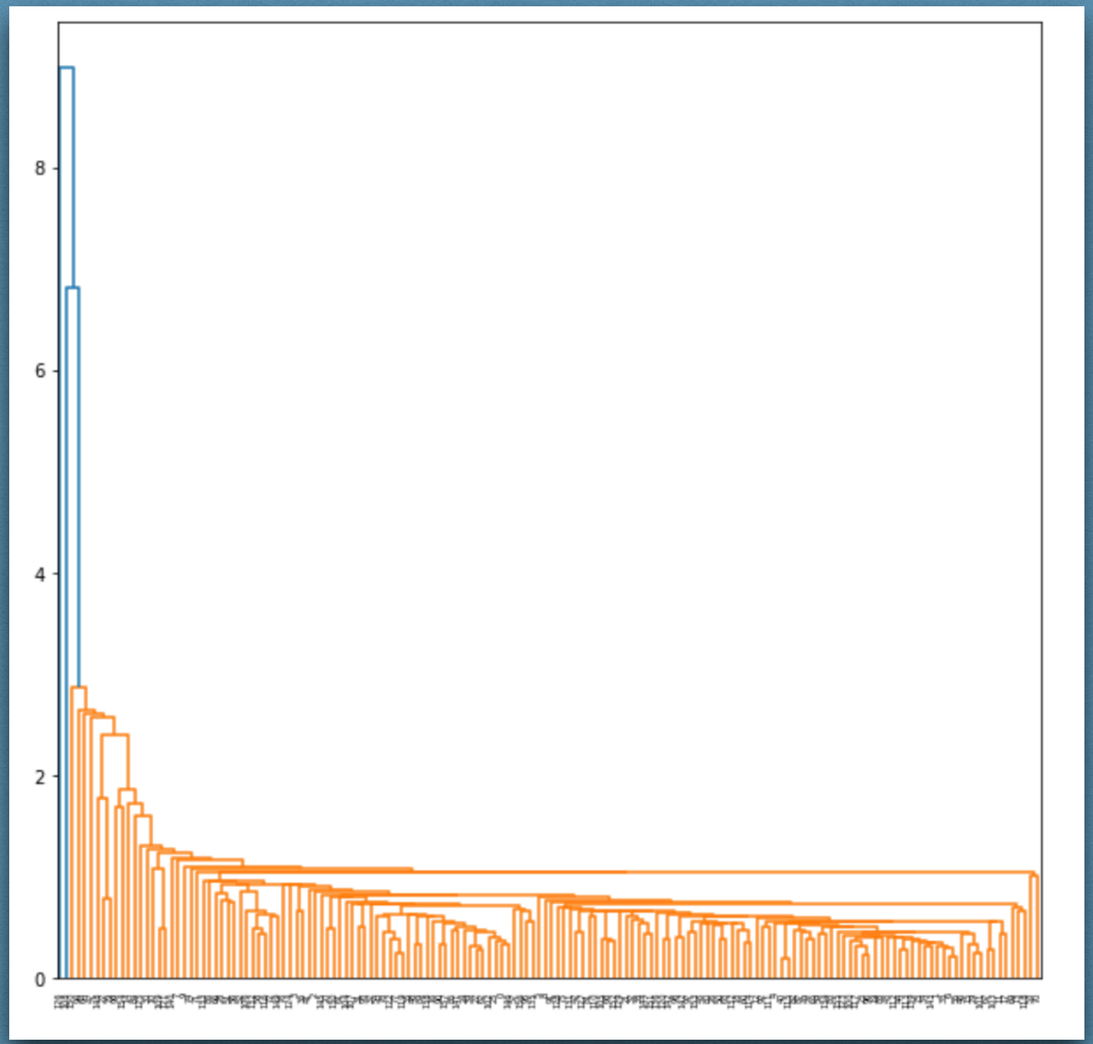
	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	label
25	Burundi	93.6	20.6052	26.7960	90.552	764	12.30	57.7	6.26	231	2
85	Liberia	89.3	62.4570	38.5860	302.802	700	5.47	60.8	5.02	327	2
36	Congo, Dem. Rep.	116.0	137.2740	26.4194	165.664	609	20.80	57.5	6.54	334	2
107	Niger	123.0	77.2560	17.9568	170.868	814	2.55	58.8	7.49	348	2
125	Sierra Leone	160.0	67.0320	52.2690	137.655	1220	17.20	55.0	5.20	399	2

From above list - 5 countries which are in dire need of Aid

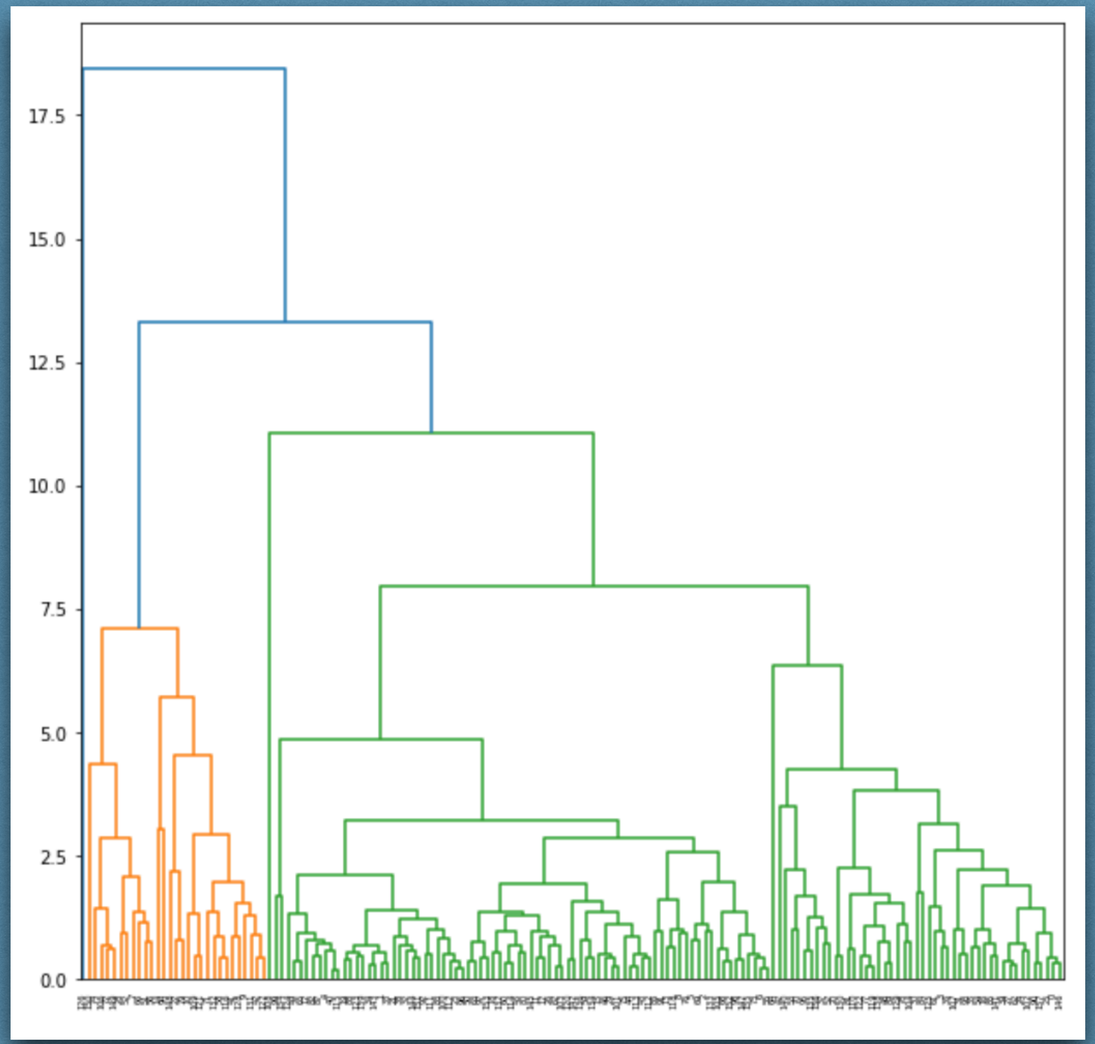
1. Burundi
2. Liberia
3. Congo, Dem. Rep
4. Niger
5. Sierra Leone

Hierarchical Clustering

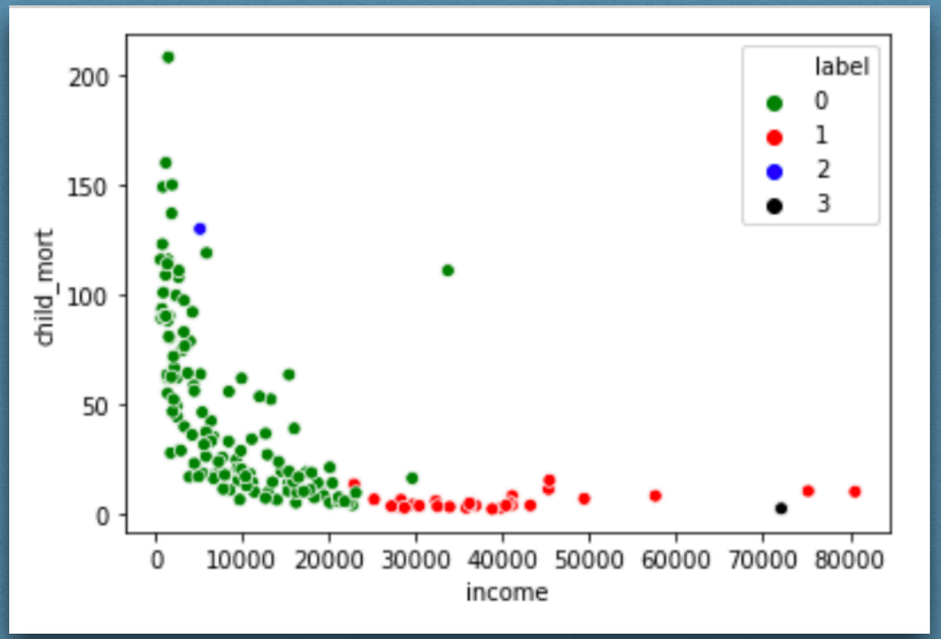
'Single' Method Hierarchical Clustering:
By look at below Dendogram, we can clearly see, single linkage doesn't produce a good enough result for us to analyse the clusters



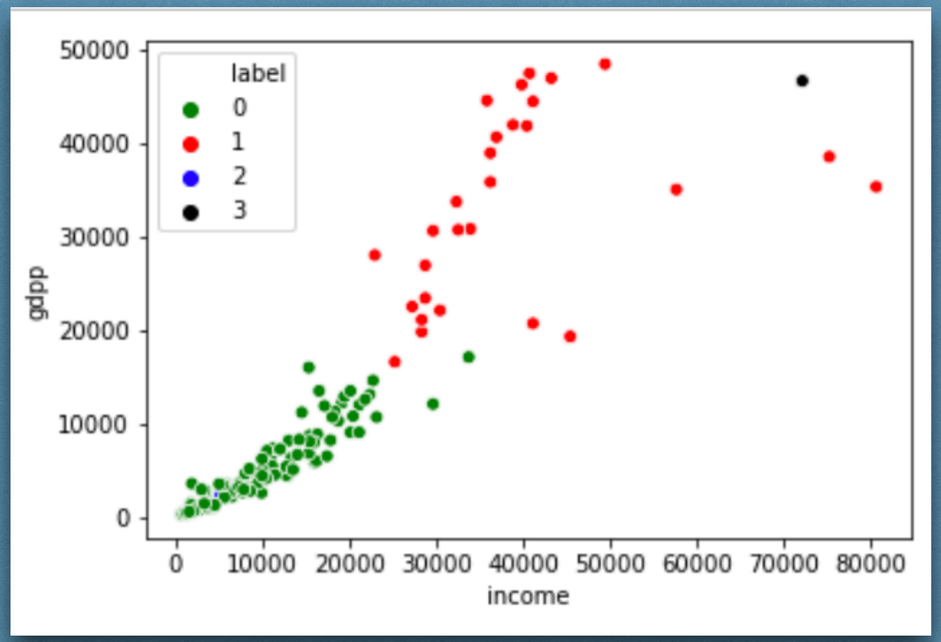
'Complete' Method Hierarchical Clustering:
By look at below **Dendogram**, we can see that ***n between 2.5 and 5 s***eems like right choice, so lets choose ***n=4***



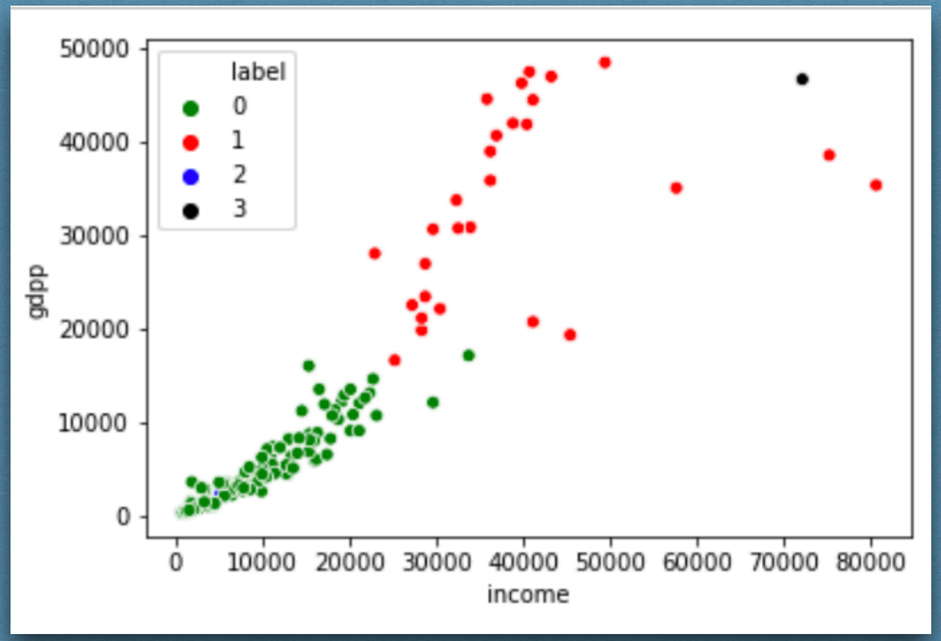
Hierarchical Clustering



We can see that countries which are grouped into **Cluster 0** have ***high child_mort*** and ***low income***



We can see that countries which are grouped into **Cluster 0** have ***high child_mort*** and ***low gdpp***



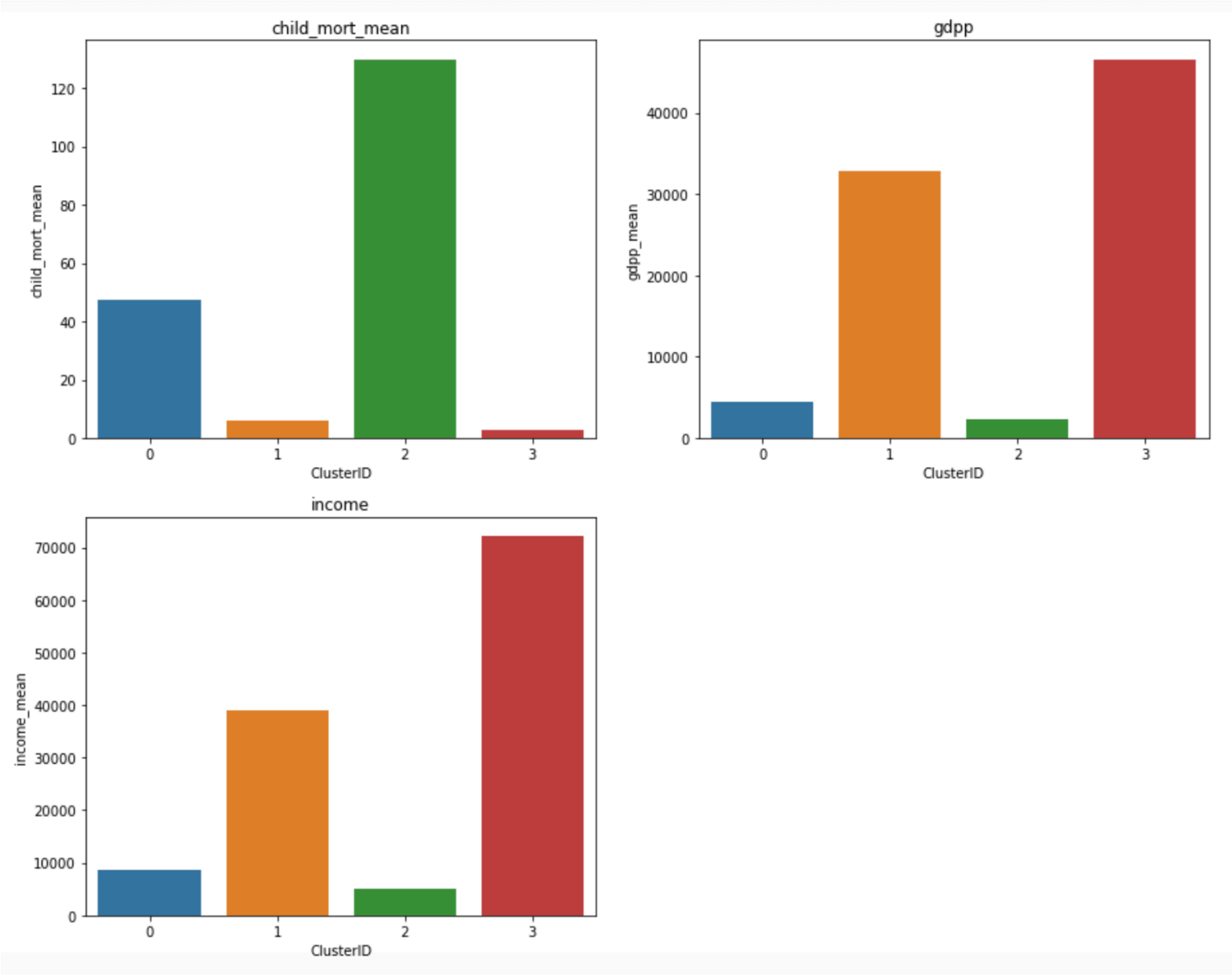
We can see that countries which are grouped into **Cluster 0** have ***low income*** and ***low gdpp***

Hierarchical Clustering

As per our K-Means Clustering, ***Cluster - 0*** contains group of countries which are in dire need of aid as they have -

- ***High*** Child Mortality rate
- ***Low*** GDP
- ***Low*** Income

***We can ignore Cluster 2 for child_mort as it has it has only 1 country assigned to it.*



Hierarchical Clustering

After Cluster profiling and result analysis, we concluded that all the countries belonging to ***Cluster ID - 0*** are in Dire need of aid.

We sorted the country in Cluster ID 0 based on the ***low*** **gdpp**, ***low*** **income** and ***high*** **child_mort** rate and below are list of Top 5 countries:

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	label
25	Burundi	93.6	20.6052	26.7960	90.552	764	12.30	57.7	6.26	231	0
85	Liberia	89.3	62.4570	38.5860	302.802	700	5.47	60.8	5.02	327	0
36	Congo, Dem. Rep.	116.0	137.2740	26.4194	165.664	609	20.80	57.5	6.54	334	0
107	Niger	123.0	77.2560	17.9568	170.868	814	2.55	58.8	7.49	348	0
125	Sierra Leone	160.0	67.0320	52.2690	137.655	1220	17.20	55.0	5.20	399	0

From above list - 5 countries which are in dire need of Aid

1. Burundi
2. Liberia
3. Congo, Dem. Rep
4. Niger
5. Sierra Leone

Summary

We can see that both **K-Means** and **Hierarchical clustering** method has provided same recommendation of Countries which are in dire need of aid and support.

Below are the List of Countries from *original dataset*, which are recommended by both clustering methods who needs our support -

country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
Burundi	93.6	8.92	11.60	39.2	764	12.30	57.7	6.26	231
Liberia	89.3	19.10	11.80	92.6	700	5.47	60.8	5.02	327
Congo, Dem. Rep.	116.0	41.10	7.91	49.6	609	20.80	57.5	6.54	334
Niger	123.0	22.20	5.16	49.1	814	2.55	58.8	7.49	348
Sierra Leone	160.0	16.80	13.10	34.5	1220	17.20	55.0	5.20	399