PCA and clustering Assignment

By Kanika Khattar Ahuja

UpGrad iit-b ज्ञानमुत्तमम्

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. 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.
- You need to categorise 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.

Analysis Approach [1/2]

- Step 1: Read and understand the data
- Step 2: Clean and visualize the data
- Step 3: Prepare the data for modeling
  - Step 3.1: Standardizing data
  - Step 3.2: Perform PCA and select the number of components
  - Step 3.3: Performing PCA with selected components
- Step 4: Modelling with k-Means clustering
  - Step 4.1 Checking data compatibility with K-Means
  - Step 4.2 Find out what should be the optimal number of clusters using SSD and silhouette score
  - Step 4.3 Performing K-Means with chosen number of clusters

Analysis Approach [2/2]

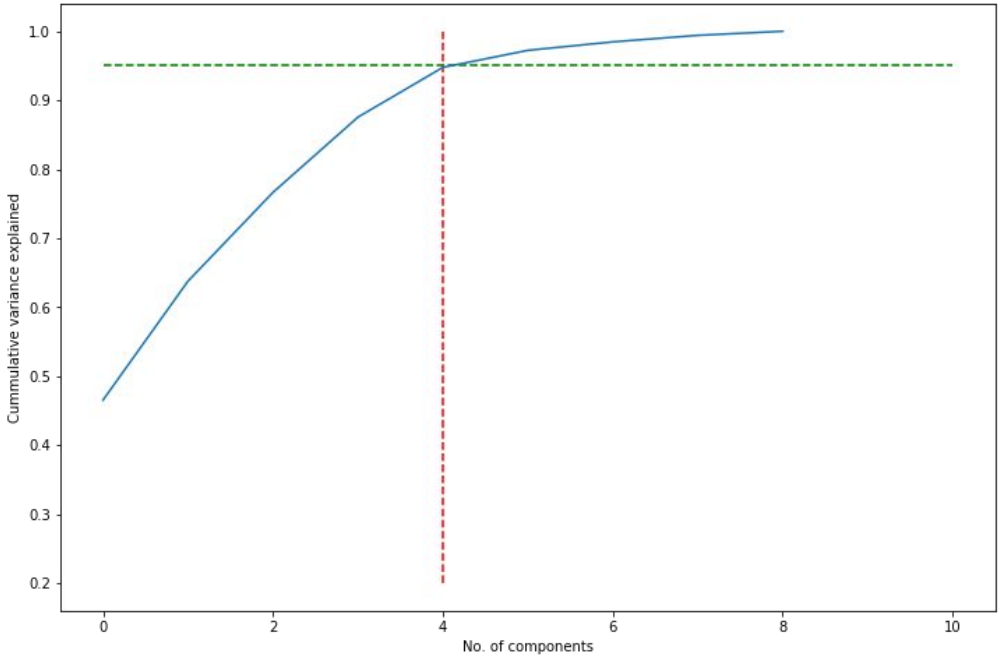
- Step 5: Cluster Profiling
  - Step 5.1 Analysis based on the cluster ids
  - Step 5.2 Analysis of a particular cluster
- Step 6: Modeling with Heirarchical Clustering
  - Step 6.1 Analysis based on the cluster ids
  - Step 6.2 Analysis of a particular cluster
- Final Analysis

Principal Component Analysis - Need of PCA

- Initially all the variables are equally contributing towards explaining the variance. If we drop any feature, we will be losing informations. Therefore, we will be using PCA to reduce the dimensions before applying any model.
- We need to apply KMeans model and for any distance based algorithm there is a drawback.
- As we increase the number of dimensions the data points start looking equidistant from the cluster center.
- In such a case, K-Means clustering will fail to assign the data to its nearest cluster. Therefore, we will be using PCA to reduce the dimensions.

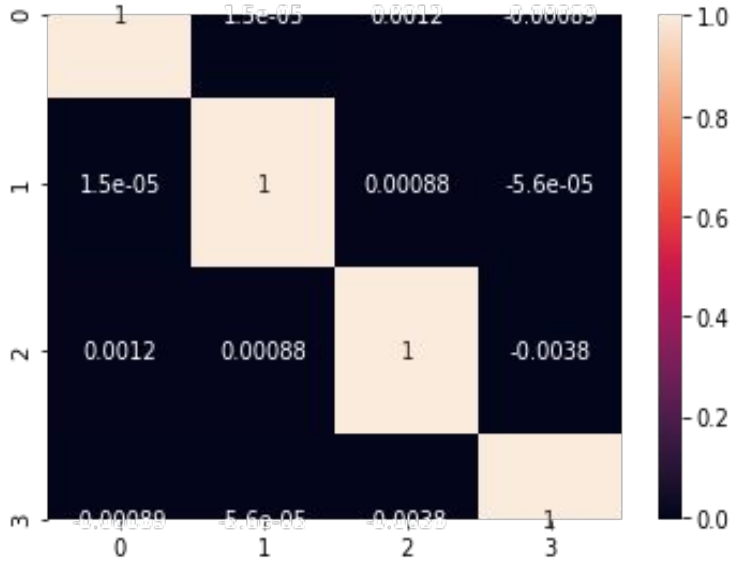
PCA - Selecting the number of Components

The scree plots shows that close to 95% of the variance is explained by 4 Principal Components. Therefore, we took 4 PCs and performed PCA.



PCA - Transformed Data

- The variances explained by PCs were as follows.
  - PC1 53.1843
  - PC2 19.7041
  - PC3 14.6490
  - PC4 12.4627
- In the transformed data the maximum variance is explained by PC1 and all components seems not to be corelated with any other components.

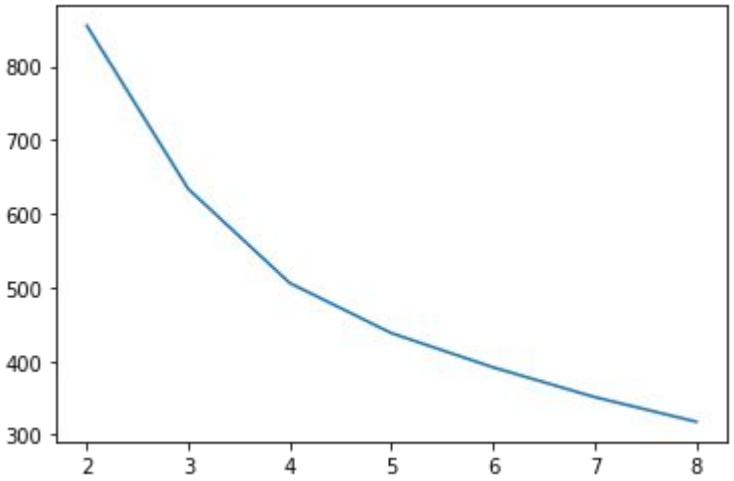


Heat map of corelations among PCs in the tranformed data

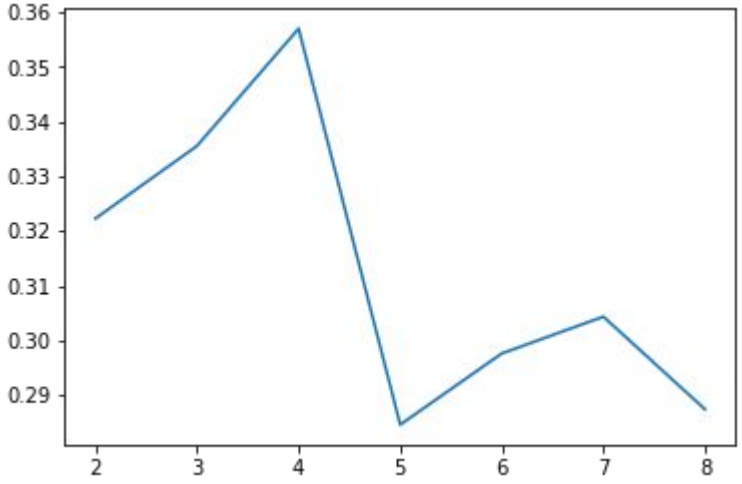
Clustering - Checking the compatibility

- Hopkins statistic was used to check the compatibility.
- On multiple runs the score was between 76 to 85
- The data was compatible for performing KMeans clustering

KMeans Clustering - Number of clusters



The above plot with elbow curve we can see that 3 or 4 clusters should be enough to create clusters of countries.



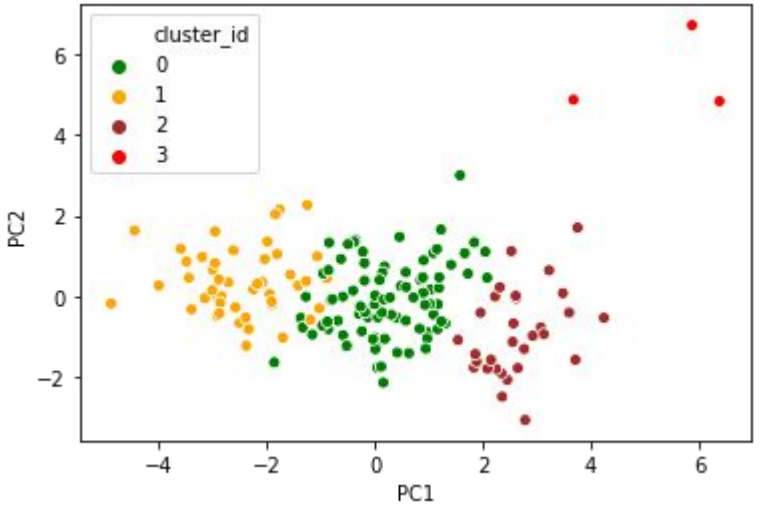
Considering k=4 for the final model as silhouette score is max for k=4

For n_clusters=4, the silhouette score is 0.35703213439911113

KMeans Clustering - Performance

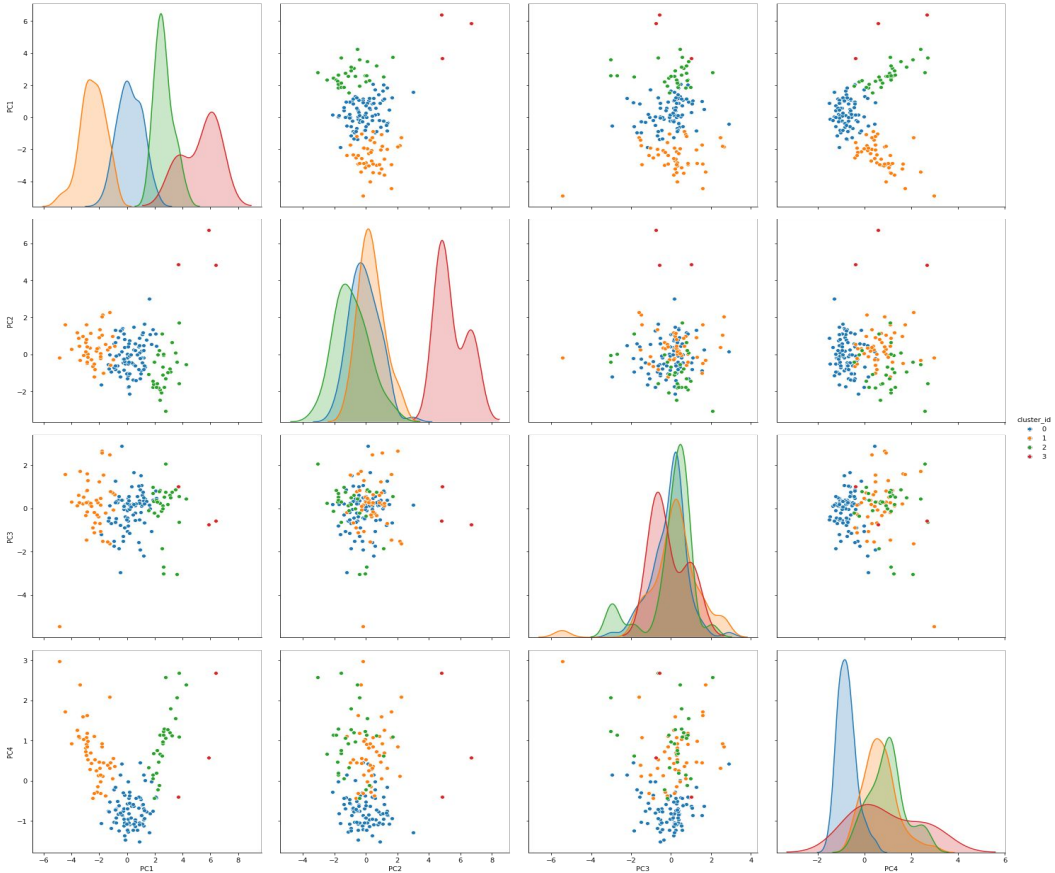
After Performing KMeans Clustering the countries were clustered as follows

- Cluster id = 0 87 countries
- Cluster id = 1 47 countries
- Cluster id = 2 30 countries
- Cluster id = 3 3 countries



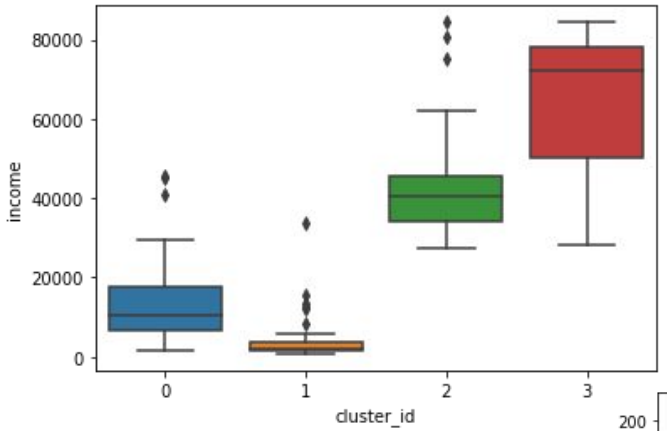
Visualization of our clusters based on PC1 and PC2

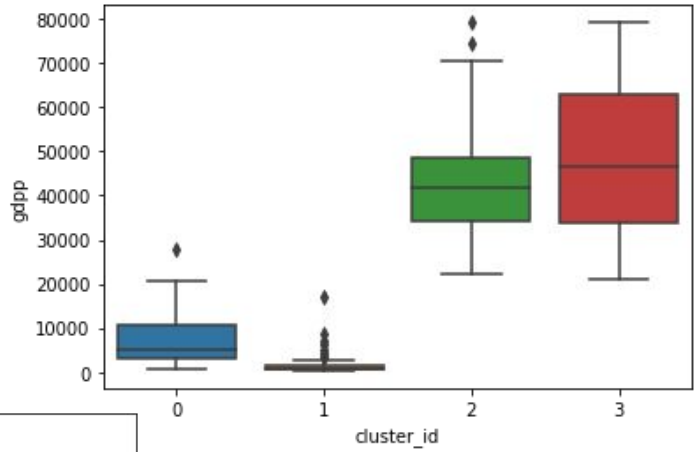
KMeans Clustering - Performance

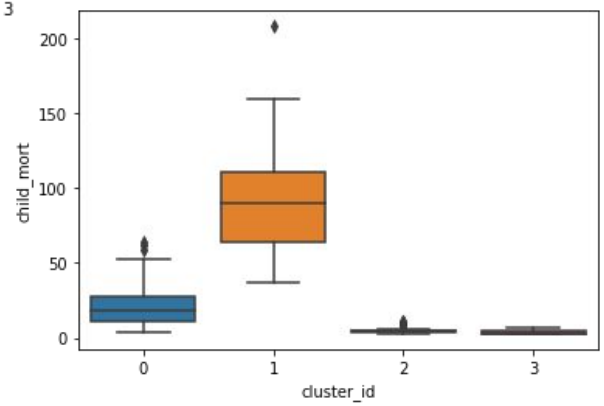


The 4 clusters are clearly visible and are separated mainly on the basis of PC1 as PC1 captures the maximum variance in the data.

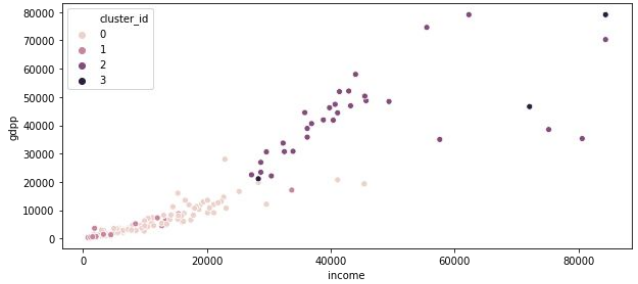
KMeans Clustering - Profiling

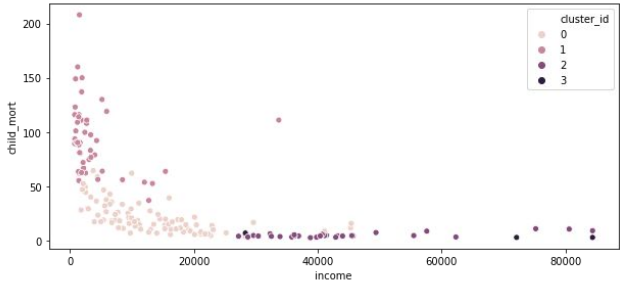


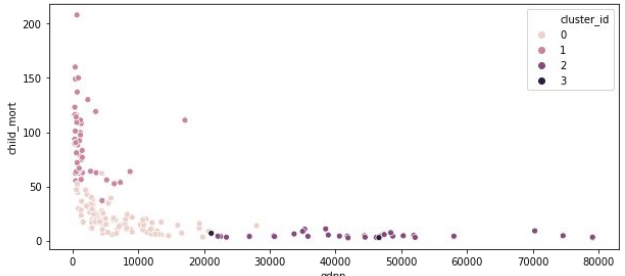




KMeans Clustering - Profiling







- Countries with high income usually have high gdpp.
- Lower income and lower gdpp countries have very high child mortality rate
- The clusters are clearly visible in these plots

KMeans Clustering - Profiling

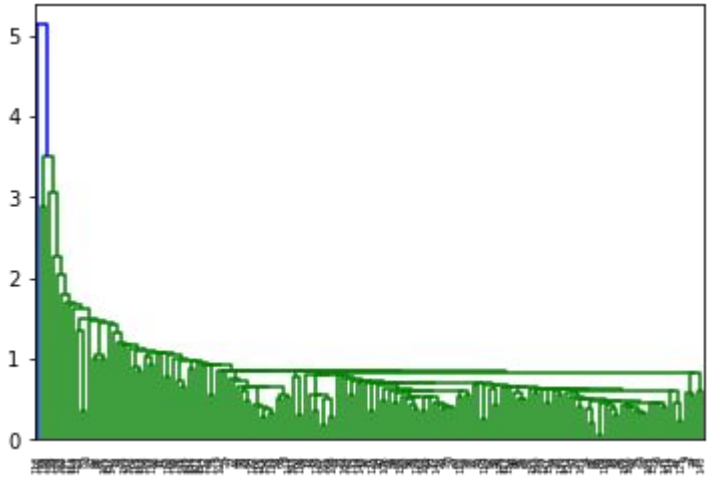
- There are just 3 countries in cluster 3 and that seems to have very high income and very high gdpp with very low mortality rate
- Countries in cluster 2 have high gdpp, high income and low mortality rate.
- Countries in cluster 0 have low income and low gdpp and a compaitively higher child mortality rate than countries in cluster 2 and 3
- Countries in cluster 1 have a very low income and very low gdpp. The child mortality rate seems to be very high.
- Looking at the above analysis, we see that countries in **cluster 1** are in need of financial aid,

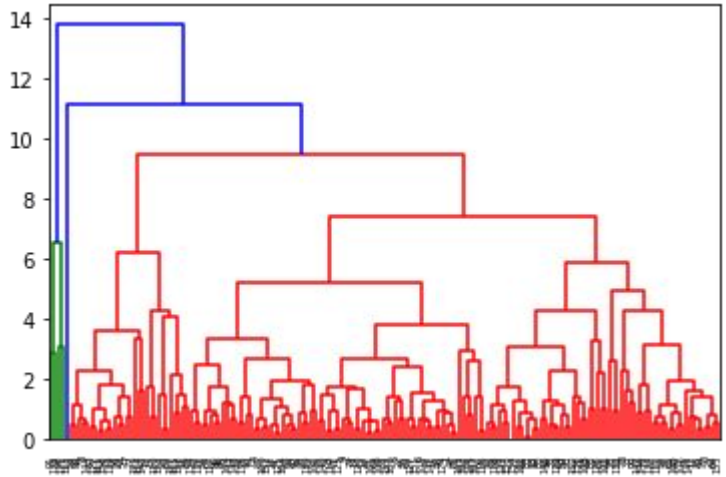
KMeans Clustering - Analysis

- The countries were sorted based on the socio-economic factors. The sort order is selected based on the importance of the factor.
- There were total 47 countries out of which following top 10 were selected.
  - **Liberia - country with minimum GDPP and income**
  - **Burundi**
  - **Congo, Dem. Rep. - country with minimum income**
  - **Niger**
  - **Sierra Leone**
  - **Madagascar**
  - **Mozambique**
  - **Central African Republic**
  - **Malawi**
  - **Eritrea**

country	gdpp	income	child_mort
Liberia	331.62	742.24	89.3
Burundi	331.62	764.00	93.6
Congo, Dem. Rep.	334.00	742.24	116.0
Niger	348.00	814.00	123.0
Sierra Leone	399.00	1220.00	160.0
Madagascar	413.00	1390.00	62.2
Mozambique	419.00	918.00	101.0
Central African Republic	446.00	888.00	149.0
Malawi	459.00	1030.00	90.5
Eritrea	482.00	1420.00	55.2

Hierarchical Clustering - Dendrograms



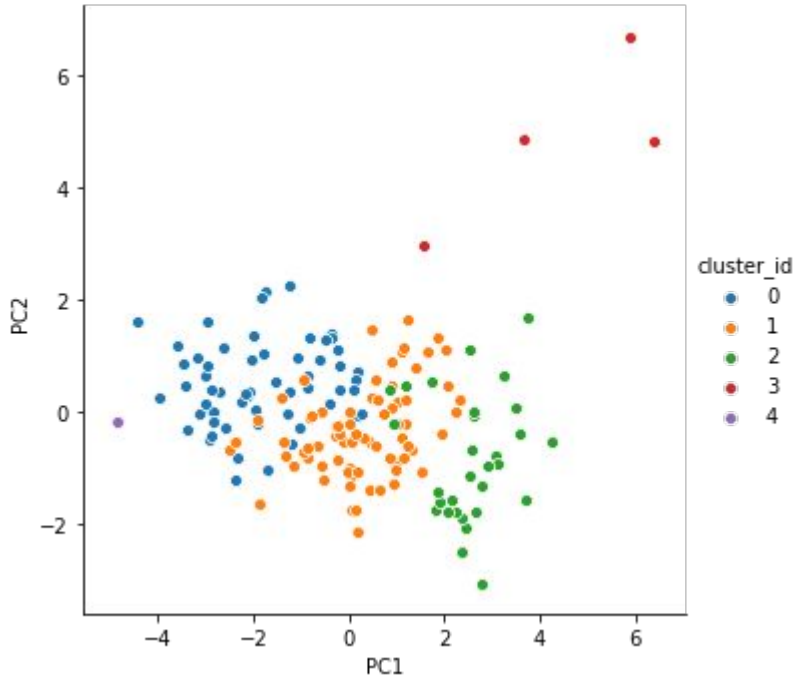


Hierarchical Clustering - Profiling

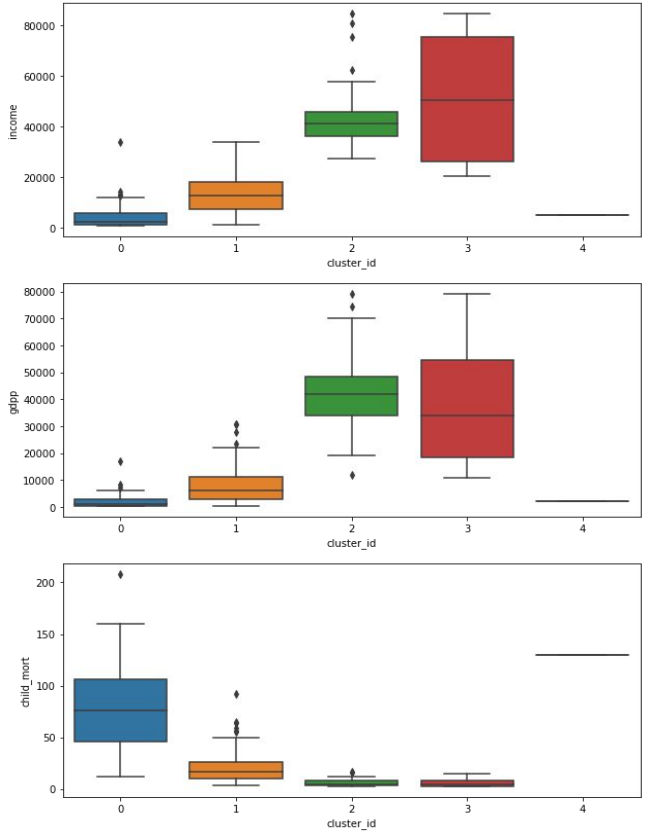
5 Clusters were made

Clusters formed are somewhat similar with KMeans but there is a slight overlap between the clusters

- 1 74 countries
- 0 58 countries
- 2 30 countries
- 3 4 countries
- 4 1 country



Hierarchical Clustering - Profiling



- There are 58 countries in cluster 0 with low income and low gdpp and high child mortality rate.
- Also one cluster i.e cluster 4 has only one country with low gdpp, low income and high mortality rate.
- There are just 4 countries in cluster 3 with very high income and very high gdpp
- There are 30 countries in cluster 2 which is also have high income and high gdpp and low mortality rate.

Hierarchical Clustering - Analysis

country	gdpp	income	child_mort
Liberia	331.62	742.24	89.3
Burundi	331.62	764.00	93.6
Congo, Dem. Rep.	334.00	742.24	116.0
Niger	348.00	814.00	123.0
Sierra Leone	399.00	1220.00	160.0
Madagascar	413.00	1390.00	62.2
Mozambique	419.00	918.00	101.0
Central African Republic	446.00	888.00	149.0
Malawi	459.00	1030.00	90.5
Togo	488.00	1210.00	90.3

- The countries were sorted based on the socio-economic factors. The sort order is selected based on the importance of the factor.
- The top 10 countries which need the aide are shown.

Final Analysis

Though the cluster are formed differently with KMeans and Hierarchical Clustering, the top 9 countries which seems to be in direst need of aid are the same. The CEO needs to focus on the following 9 countries.

- **Liberia - country with minimum GDPP and income**
- **Burundi**
- **Congo, Dem. Rep. - country with minimum income**
- **Niger**
- **Sierra Leone**
- **Madagascar**
- **Mozambique**
- **Central African Republic**
- **Malawi**