Clustering & PCA Assignment

•Present the overall approach of the analysis in a presentation Mention the problem statement and the analysis approach.
•Explain the results of Principal Component Analysis and Clustering briefly.
•Include visualisations and summarise the most important results in the presentation.
•Make sure that you mention the final list of countries here (Don't just mention the cluster id or cluster name here. Mention the names of all the countries.)

Problem Statement:

- **HELP International, an international humanitarian NGO raised the funds and want to invest it for overall development of the countries that are in the direst need of aid.**
- **we need to categorise the countries using some socio-economic and health factors that determine the overall development of the country.**

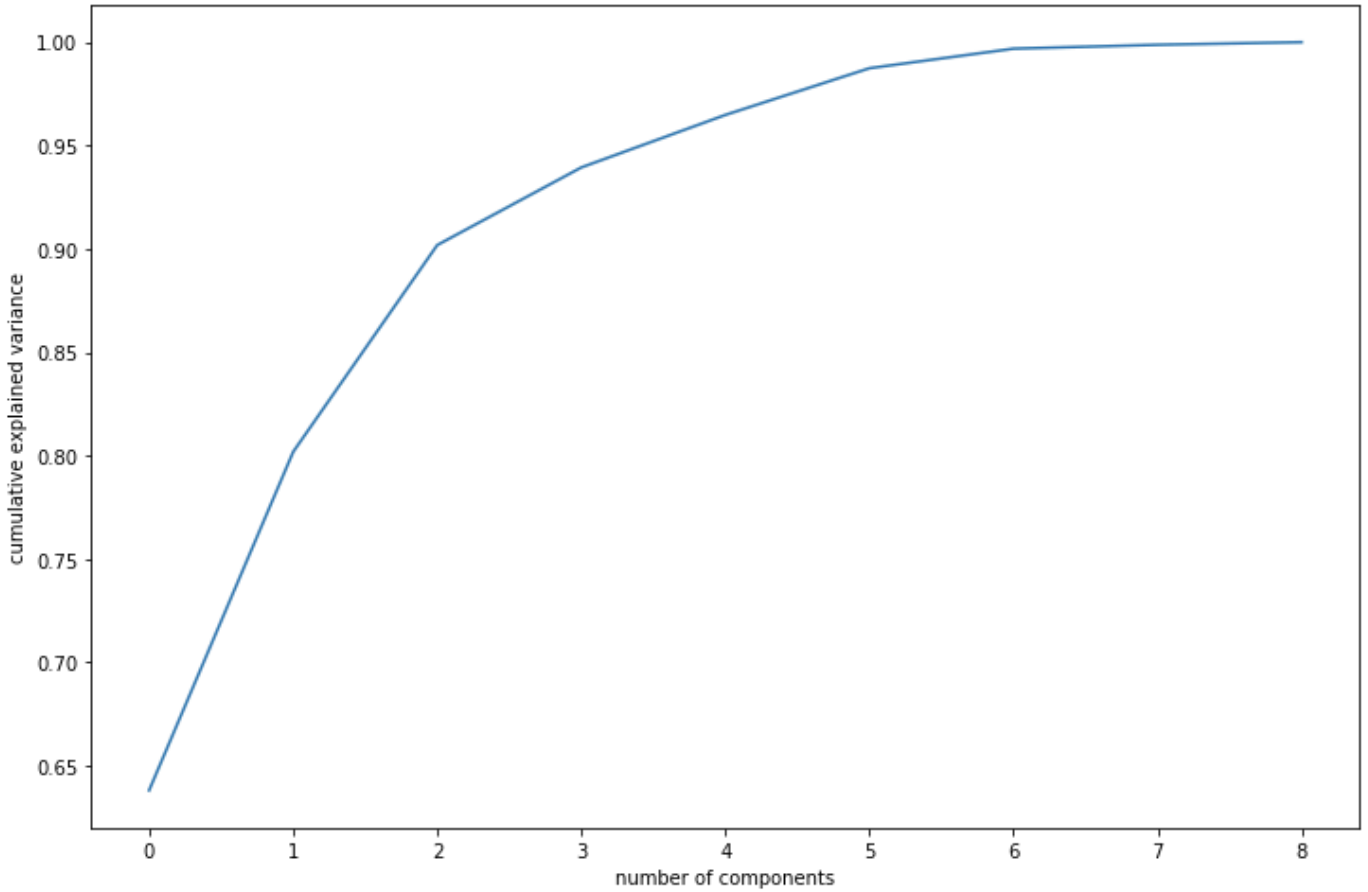
Analysis Approach

- EDA on the dataset. Read the data, find missing values and do the Outlier Analysis
- Perform PCA on the dataset and obtain the new dataset with the Principal Components. Choose the appropriate number of components k and perform the clustering activity on this new dataset, i.e. the PCA modified dataset with the k components
- Try both K-means and Hierarchical clustering (both single and complete linkage) on this dataset to create the clusters.
- Analyze the clusters and identify the ones which are in dire need of aid by comparing how these three variables - [gdpp, child_mort and income] vary for each cluster of countries to recognize and differentiate the clusters of developed countries from the clusters of under-developed countries.
- perform visualizations on the clusters that have been formed.
- The final list of countries depends on the number of components that are chosen and the number of clusters that are finally formed.

Principal Component Analysis

- Scale the data using StandardScaler
- Perform the PCA on Scaled data
- Check for explained_variance_ratio_ and plot the scree plot to find the number of PCA components, which explains the 96% of the variance. 5 PCA components are observed and incremental PCA is performed.
- From the scatter plot of PC1 and PC2, it is observed that Child_mort is more high on PC2 where as Import , export, income and GDPP are more high on PC1

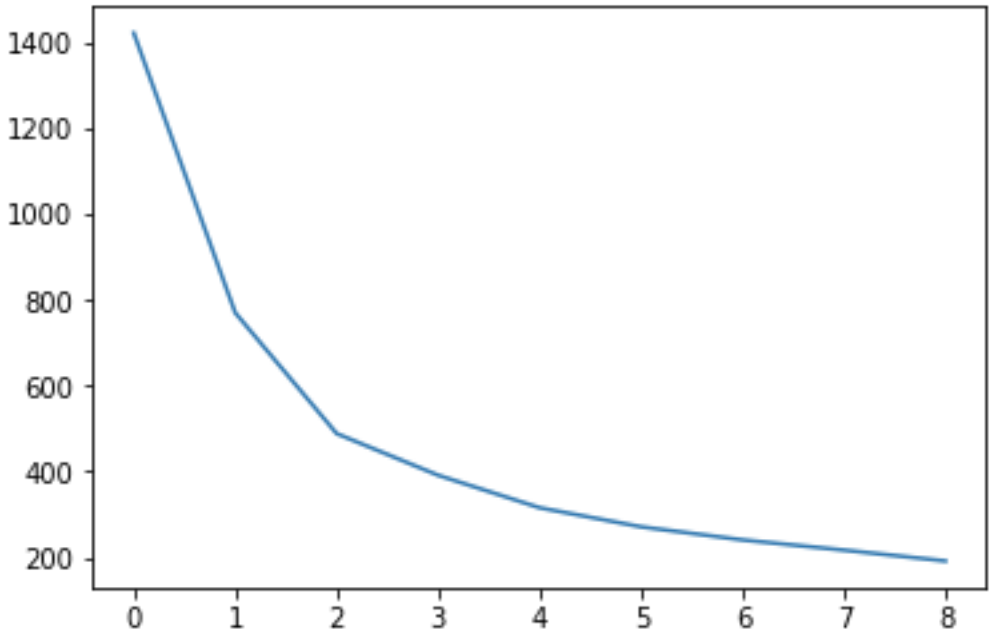
PCA: scree plot



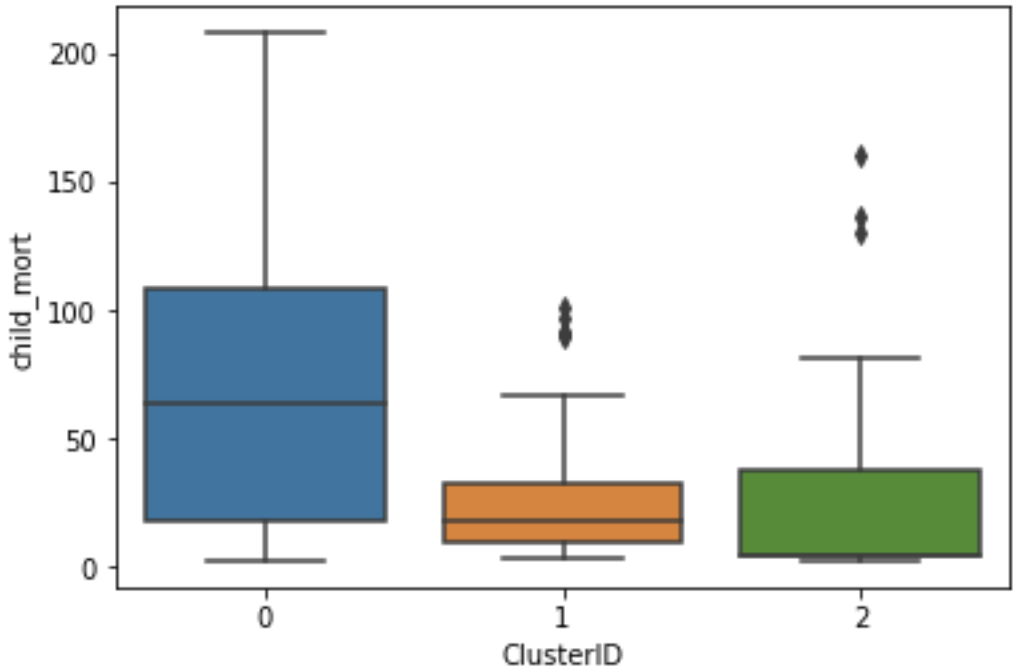
K Means Clustering

- Hopkins statistic :- it tells us that whether we can perform K means clustering on the given data set or not. As we saw that Hopkins score is 0.88. which is a good number. We can perform the K means to form the clusters.
- silhouette analysis and Elbow curve methods are used to find the number of clusters. We observed that there should be 3 clusters as it make the business sense also. We can have 3 clusters of underdeveloped , developed and developing countries.
- Cluster 0 47 underdeveloped countries
- Cluster 1 87 developing countries
- Cluster 2 27 developed countries

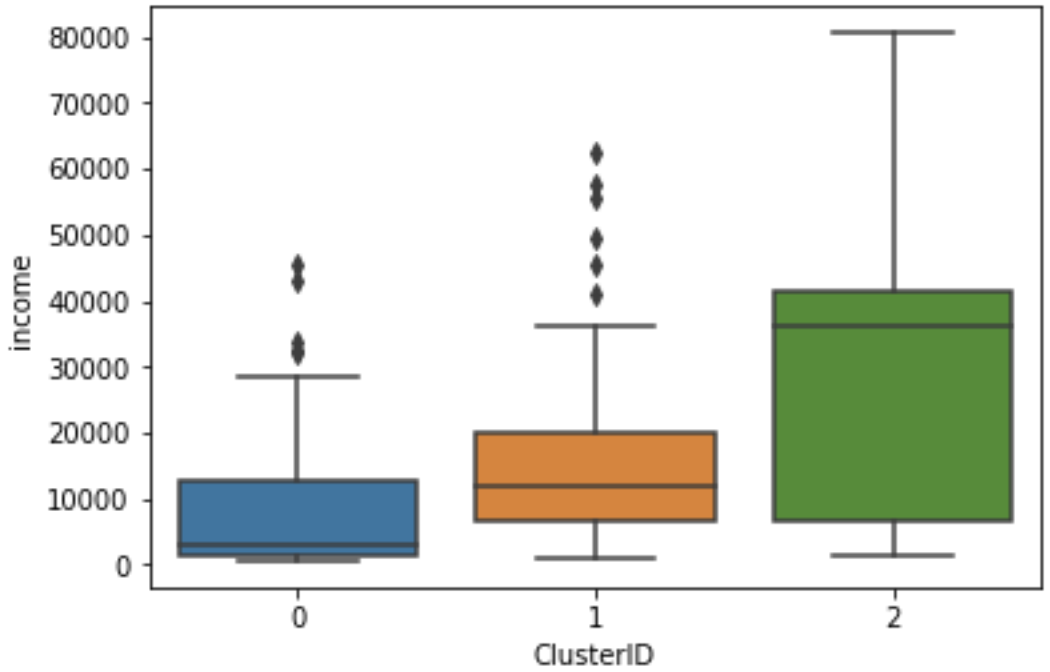
Elbow curve: 3 clusters



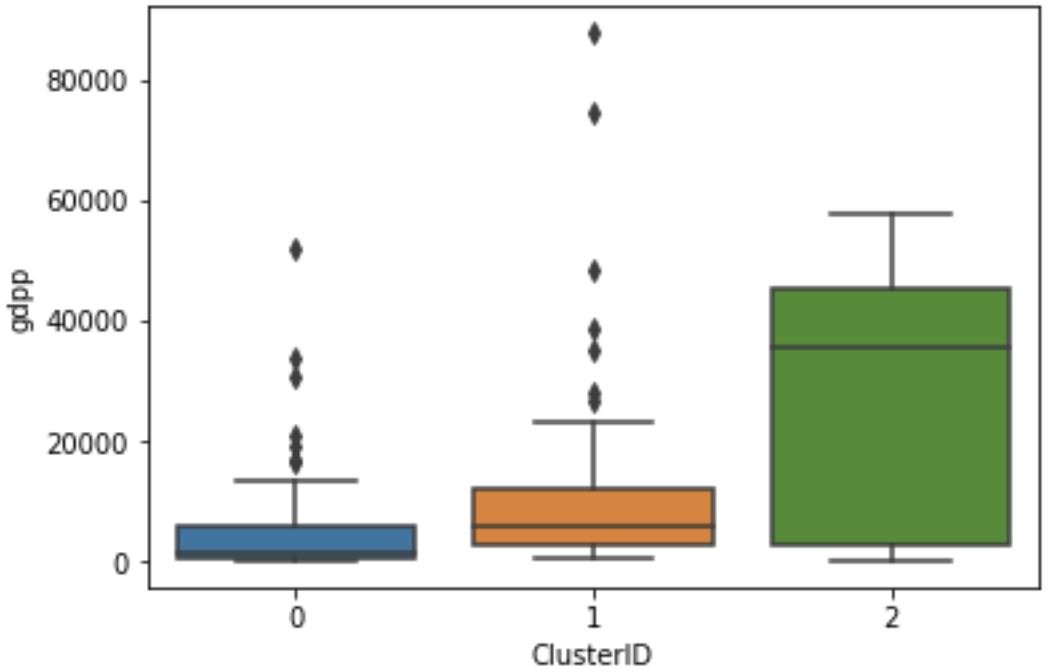
Cluster 0 has High Child Mort



Cluster 2 has High Income
Cluster 0 has low Income



Cluster 2 has High gdpp
Cluster 0 has low gdpp



list of 10 underdeveloped countries using K mean clustering

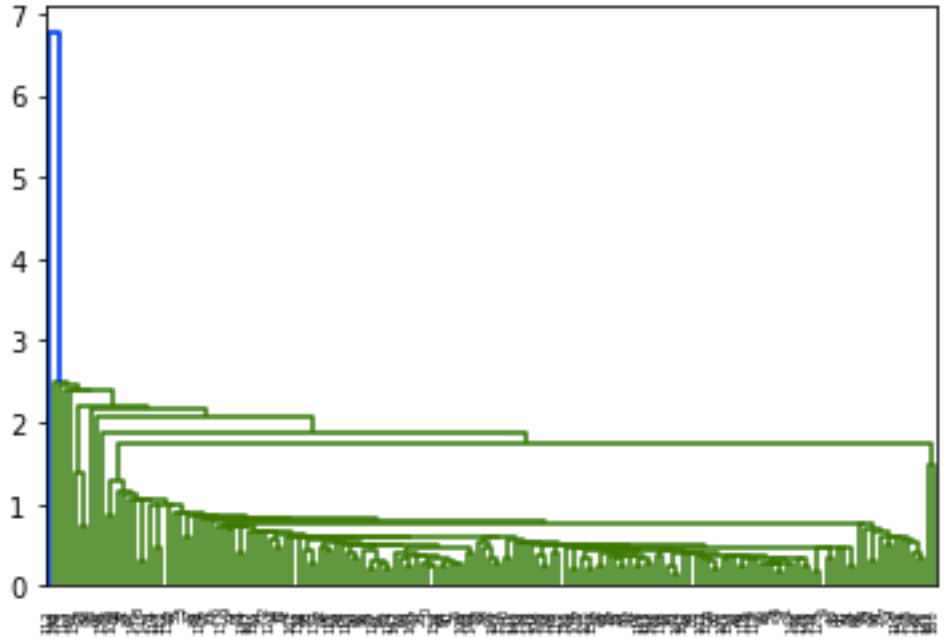
- Found the list of 10 countries which has avg income and GDP less then the cluster 0 Average.
- Congo, Dem. Rep.
- Niger
- Sierra Leone
- Central African Republic
- Guinea-Bissau
- Guinea
- Haiti
- Burkina Faso
- Mali
- Benin

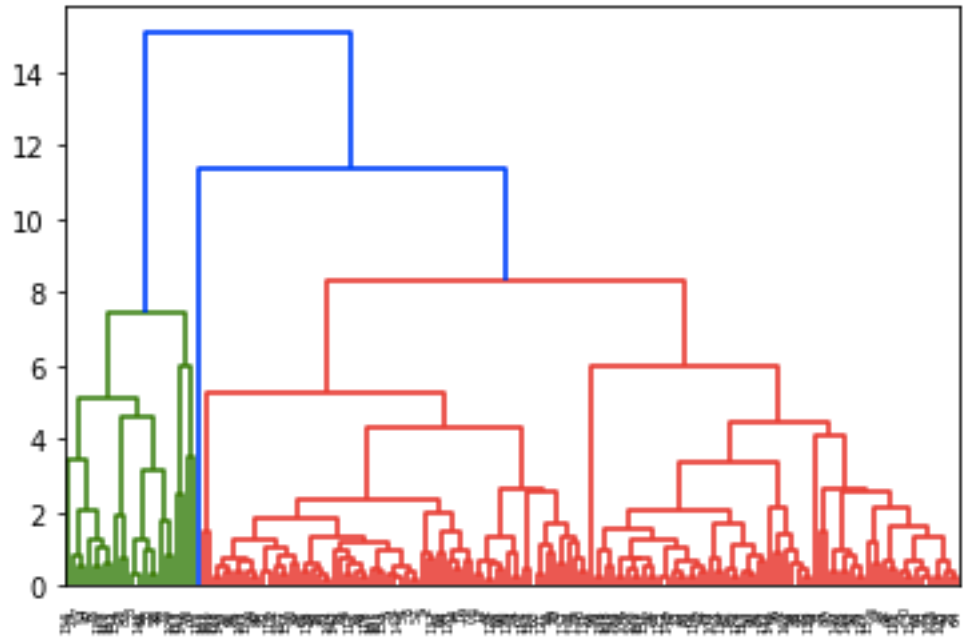
Analysis Summary : K mean

- - Cluster 0 - Underdeveloped countries:- Low GDPP,Low Income, Low ratio of Import and export, Less Expenditure on health Hence low Life expectancy and High Child Mort rate
- - Cluster 1 - Developing countries :- Avarage GDPP, Average ratio of Import and export, Avg. Expenditure on health Hence good Life expectancy and Avg Child Mort rate
- - Cluster 2 - Developed Countries:- High GDPP,High Income, High ratio of Import and export, High Expenditure on health Hence High Life expectancy and low Child Mort rate

Hierarchical Clustering

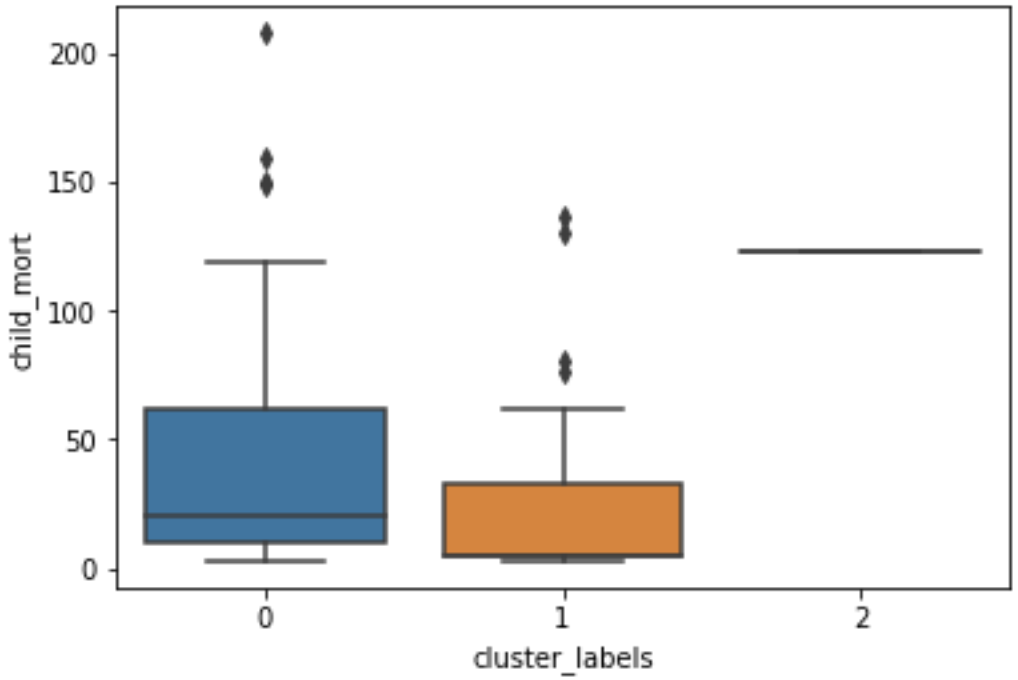
single linkage and complete linkage



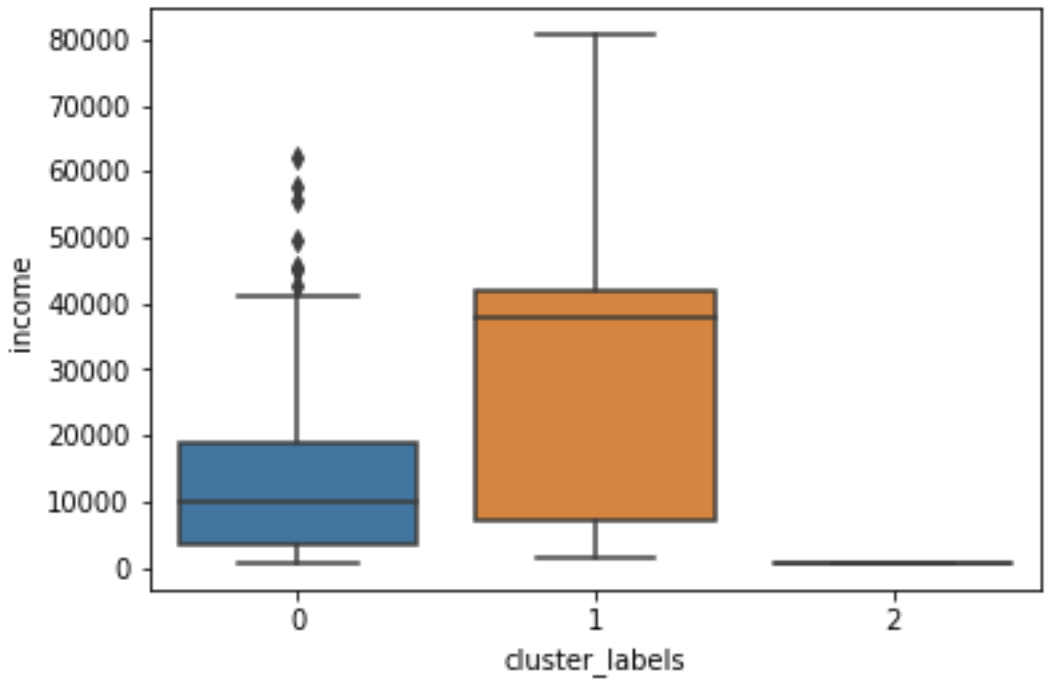


- Performed Hierarchical Clustering on existing PCA dataset
- Formed 3 cluster : cluster 0 139 countries cluster 124 countries cluster 2 1 countries
- cluster 0 – developing
- cluster 1 - developed
- cluster 2 - underdeveloped

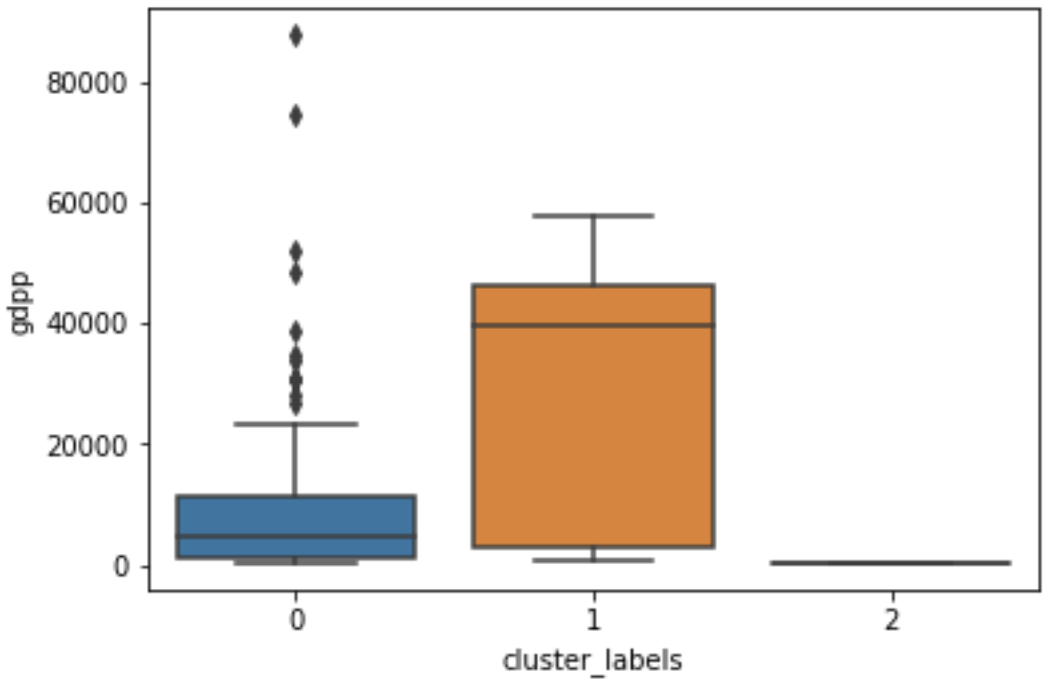
Cluster 0 has High Child Mort



Cluster 1 has High Income



Cluster 1 has High gdpp



Analysis Summary: Hierarchical Clustering

- - Cluster 0 - Developing countries:- Avarage GDPP, Average ratio of Import and export, Avg. Expenditure on health Hence good Life expectancy and Avg Child Mort rate
- - Cluster 1 - Developed countries :- High GDPP,High Income, High ratio of Import and export, High Expenditure on health Hence High Life expectancy and low Child Mort rate
- - Cluster 2 - Underdeveloped Countries:- Low GDPP,Low Income, Low ratio of Import and export, Less Expenditure on health Hence low Life expectancy and High Child Mort rate

list of 21 underdeveloped countries

- To find the list of countries which are in diar need of help based on GDPP child mort and income . we can choose cluster means as cut offs and find the final list of countries from the original dataset.
- Burundi
- Liberia
- Congo,Dem. Rep.
- Niger
- Sierra Leone
- Mozambique
- Central African Republic
- Malawi
- Togo
- Guinea-Bissau

list of 21 underdeveloped countries(cont..)

- Afghanistan
- Burkina Faso
- Guinea
- Haiti
- Mali
- Benin
- Comoros
- Chad
- Lesotho
- Cote d'Ivoire
- Cameroon

Final Summary

- From the above 2 clustering methods , below are the 5 countries(common countries among the list of top 10 in the increasing order of GDP) which are in diar need of help
- - Congo, Dem. Rep.
- - Niger
- - Sierra Leone
- - Central African Republic
- - Guinea-Bissau