Clustering and PCA Assignment

-By Riddhi Agarwal

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 programs, 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.

What need to be done

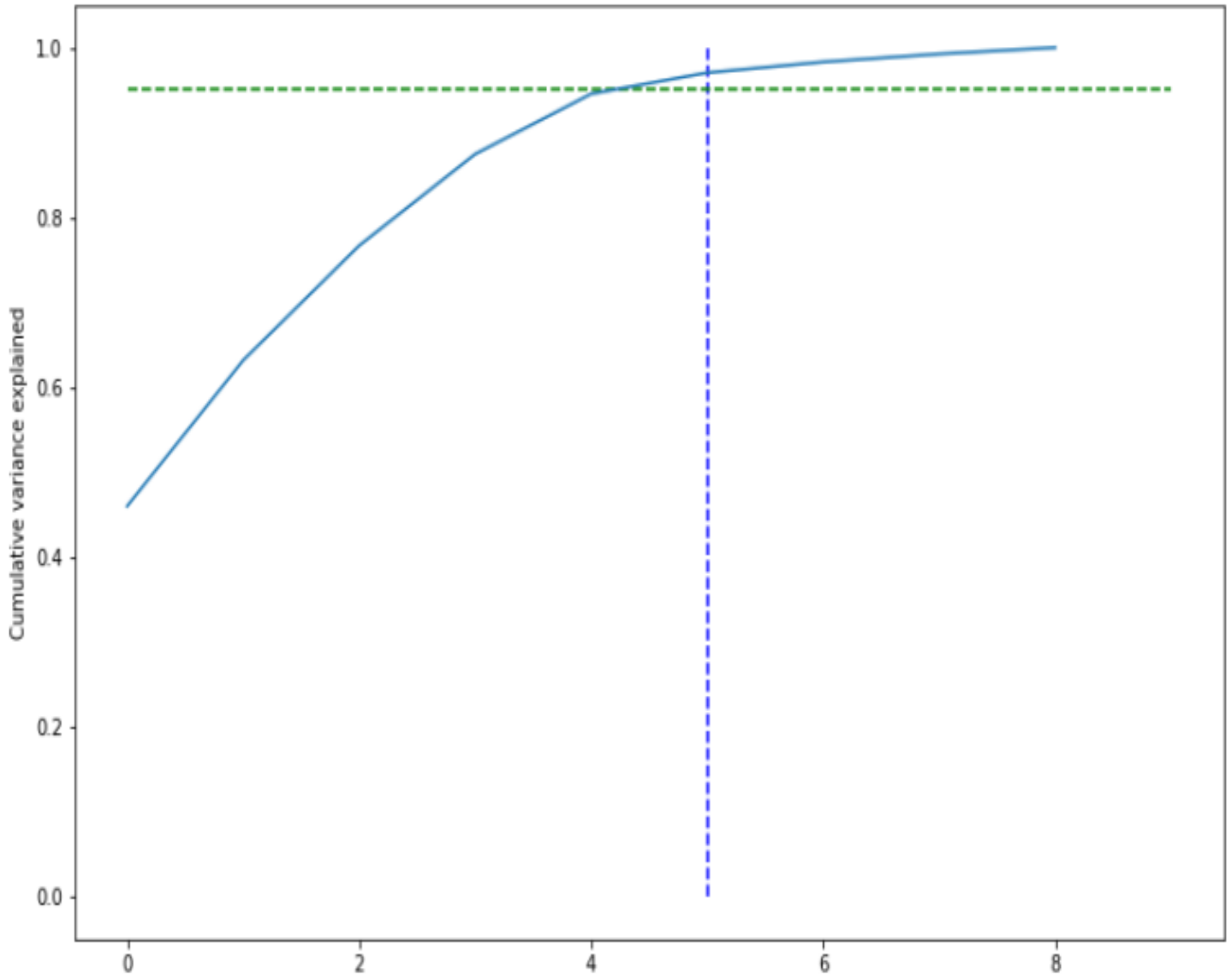
- To categorize the countries using some socio-economic and health factors that determine the overall development of the country. Then you need to suggest the countries which the CEO needs to focus on the most.

Analysis Approach

- First performed the PCA on data, out of 9 principal components selected only 5 components as they were explaining the 95% variance in the data. Then performed k means and hierarchical clustering, in both the models divided the data into three clusters so as to classify countries as under-developed, developing and developed.
- Hierarchical clustering clustered most of the countries in single cluster so results were not much reliable. K means cluster model produced better results. Visualized the clusters of k means based on gdpp, child mortality and income. Then selected that cluster in which income and gdpp were minimum and child mortality was maximum.
- This cluster countries to be presented to NGO which are in direst need of aid.

Principal Component Analysis

- Converted the data into it's principal components.
- As there were 9 columns so 9 principal components were formed.
- 5 components were explaining 95% variance of the data so only 5 components were used for model building and clustering.

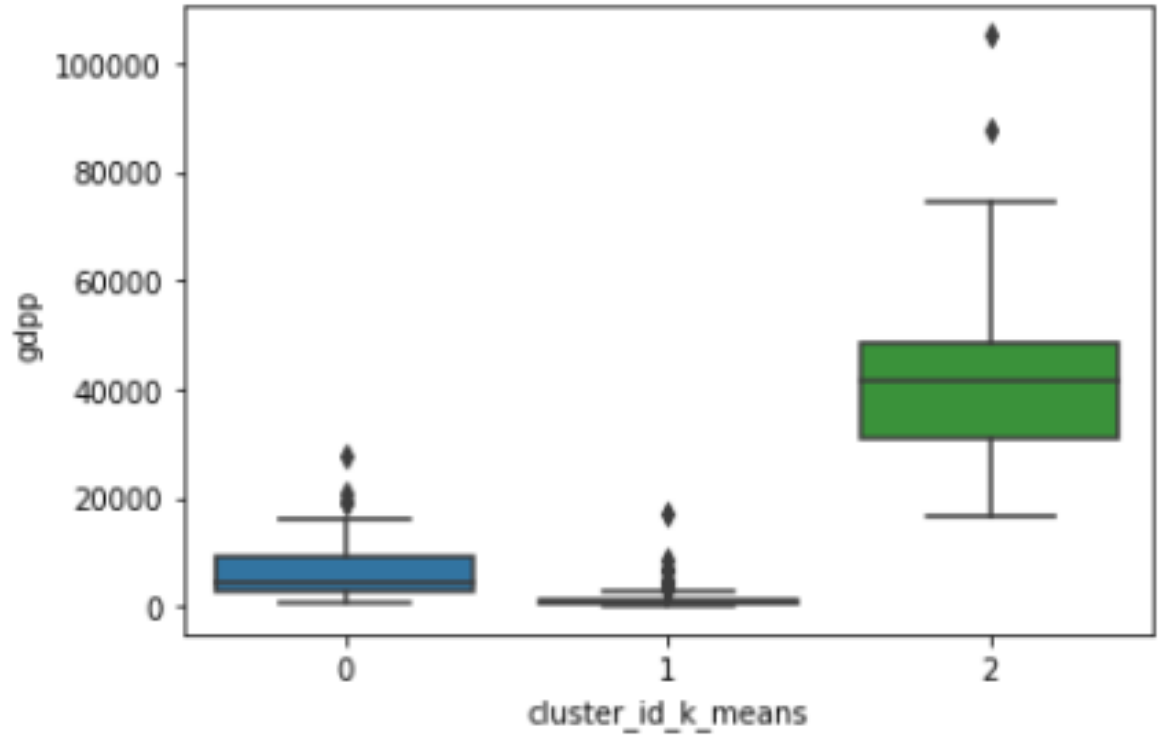


K-Means Clustering

Divided the data into 3 clusters so as to divide the countries into 3 groups- under developed, developing and developed countries.

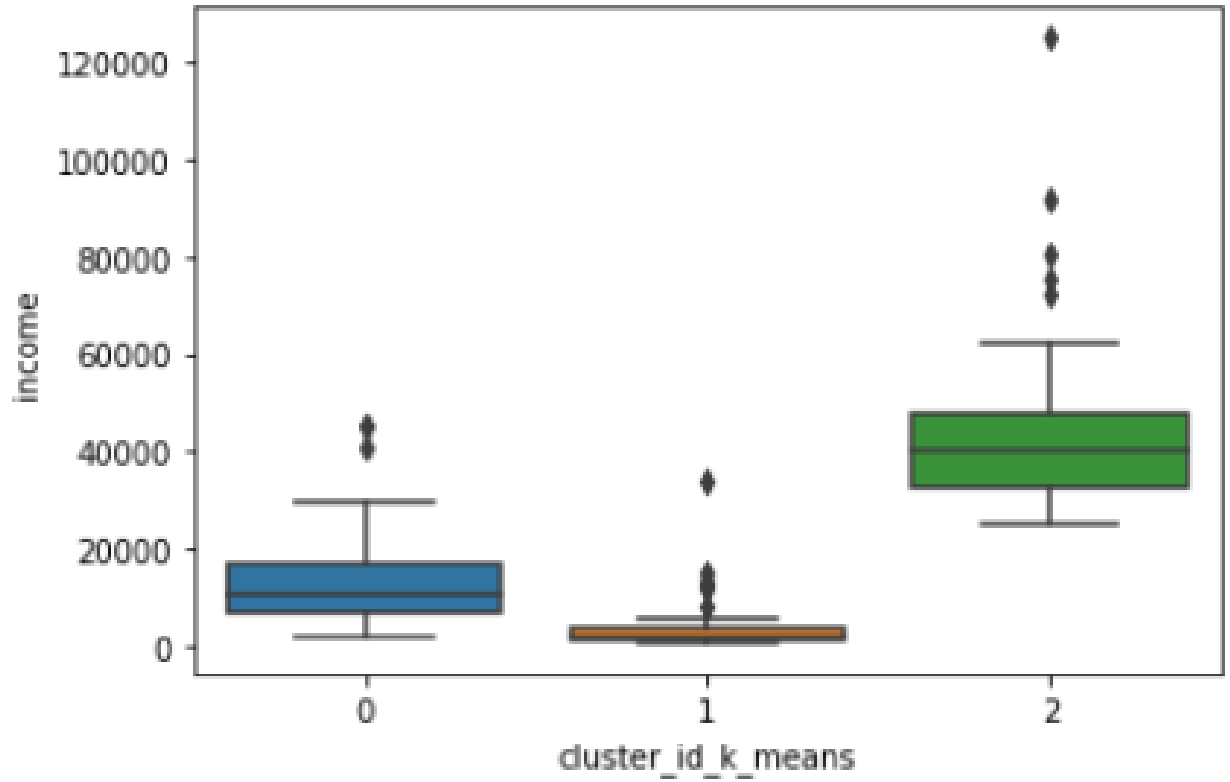
Assigned the cluster labels to the original data and visualized it’s results based on gdpp, income and child mortality.

Clustering ID and gdpp



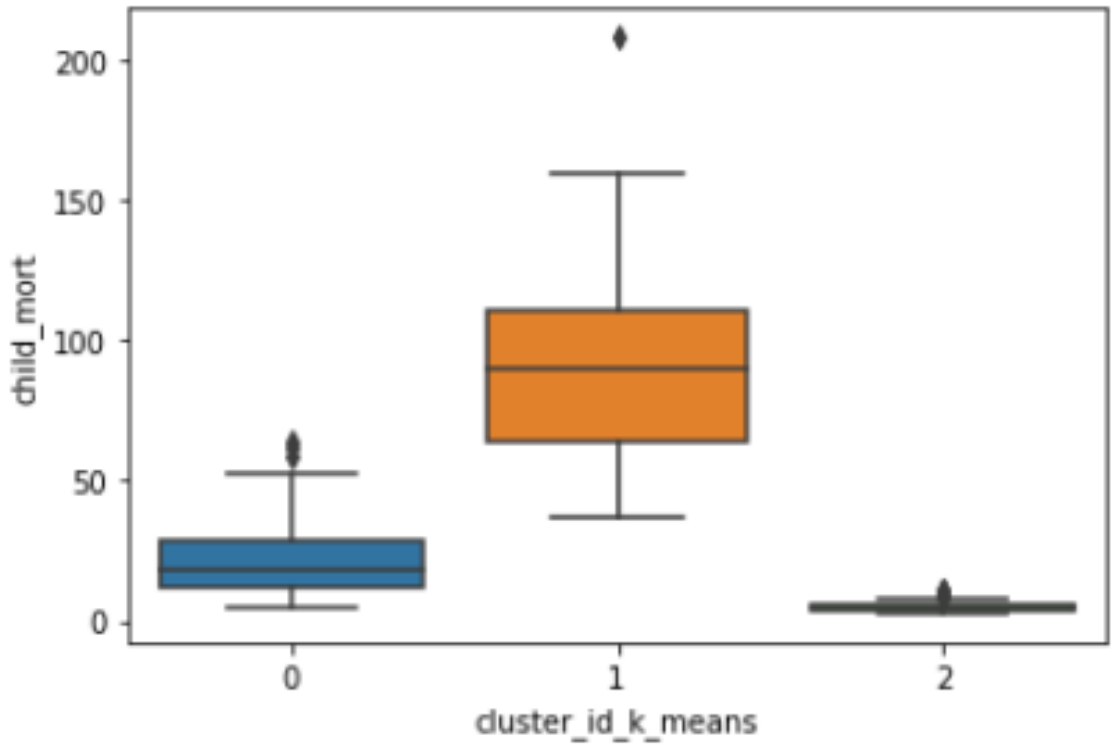
It can be visualized from the box plot that cluster 1 has minimum gdpp

Clustering ID and income



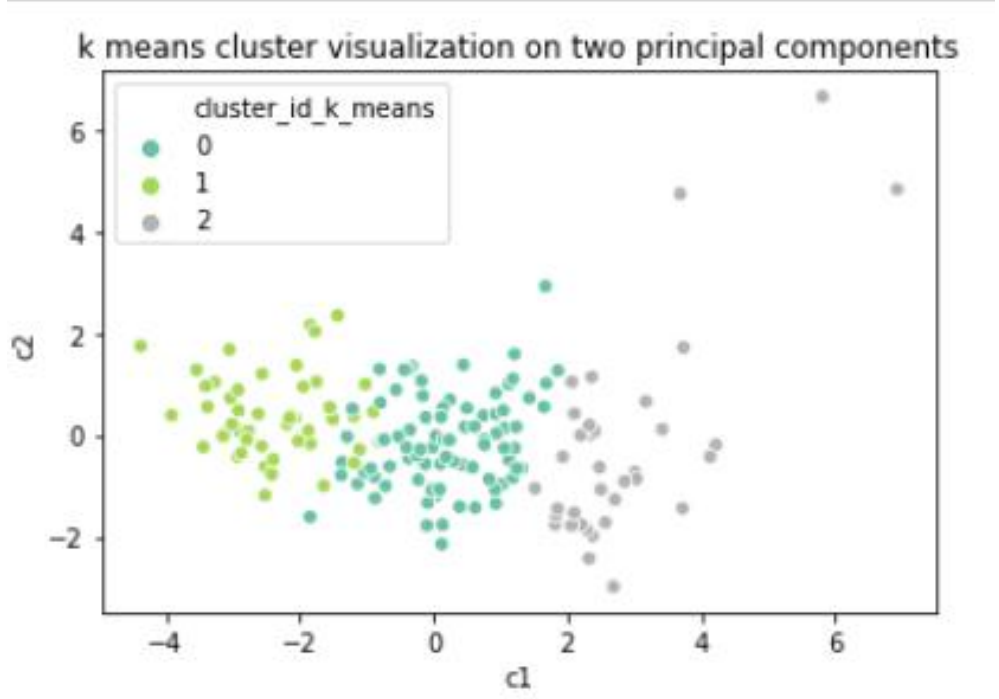
It can be visualized from the box plot that cluster 1 has minimum income

Clustering ID and child mortality



It can be visualized from the box plot that cluster 1 has maximum child mortality.

Clusters visualization



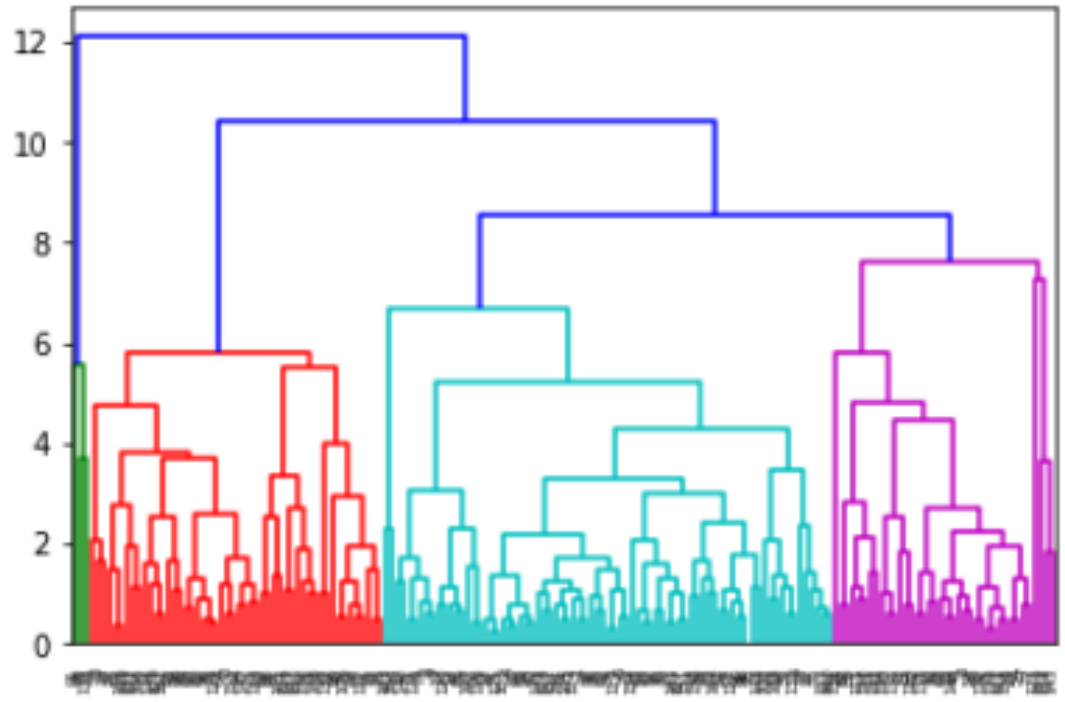
Hierarchical Clustering

Similar to k-means clustering divided the data into 3 clusters so as to divide the countries into 3 groups- under developed, developing and developed countries.

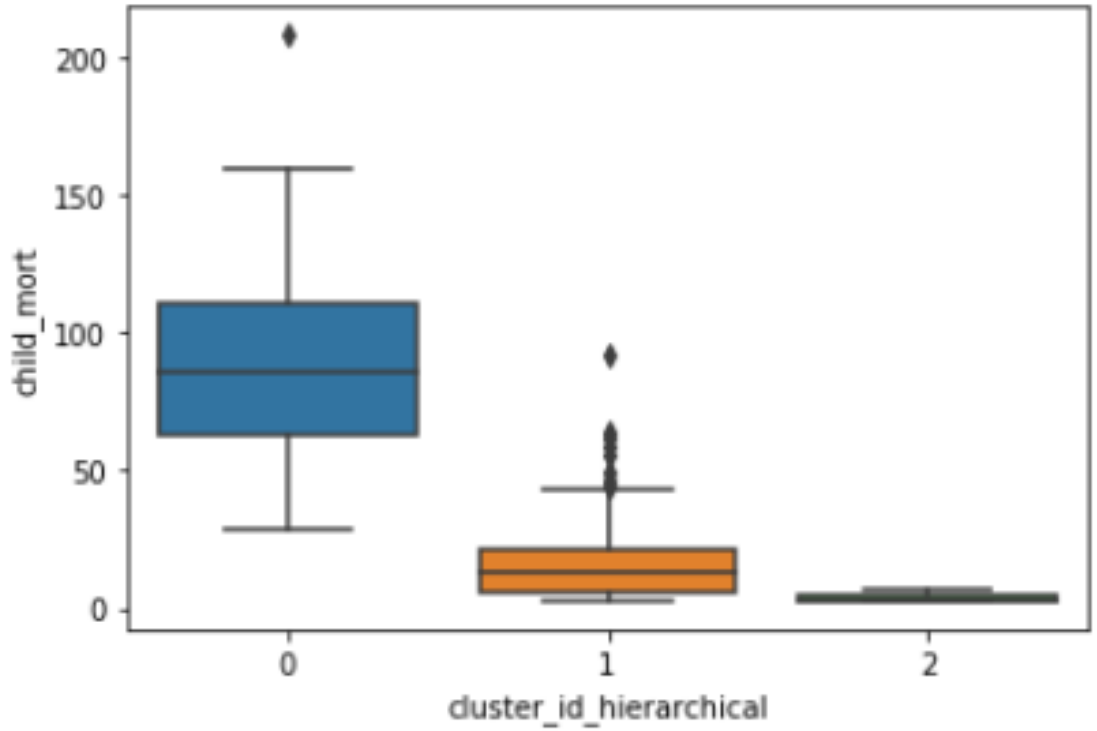
Assigned the cluster labels to the original data and visualized it's results based on gdpp, income and child mortality.

Performed both single linkage and complete clustering, single linkage clustering was giving very cluttered results so moved forward with complete clustering

Complete Linkage dendogram

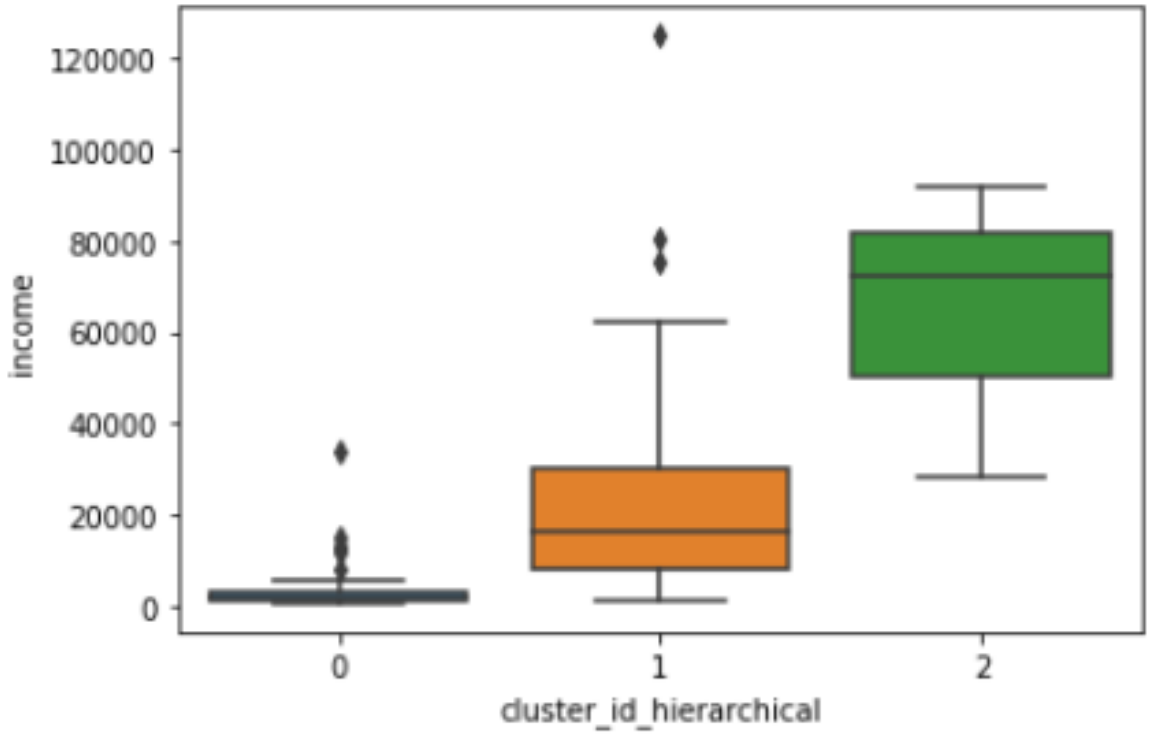


Clustering ID and child mortality



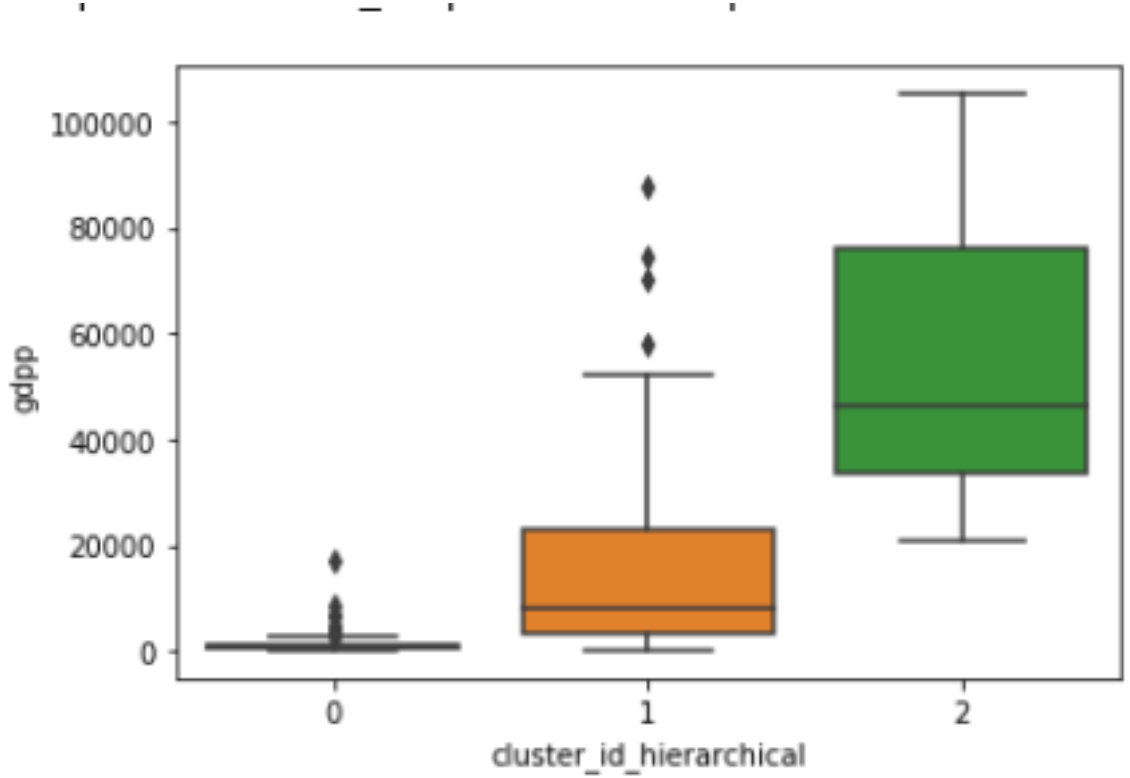
It can be visualized from the box plot that cluster 0 has maximum child mortality

Clustering ID and income



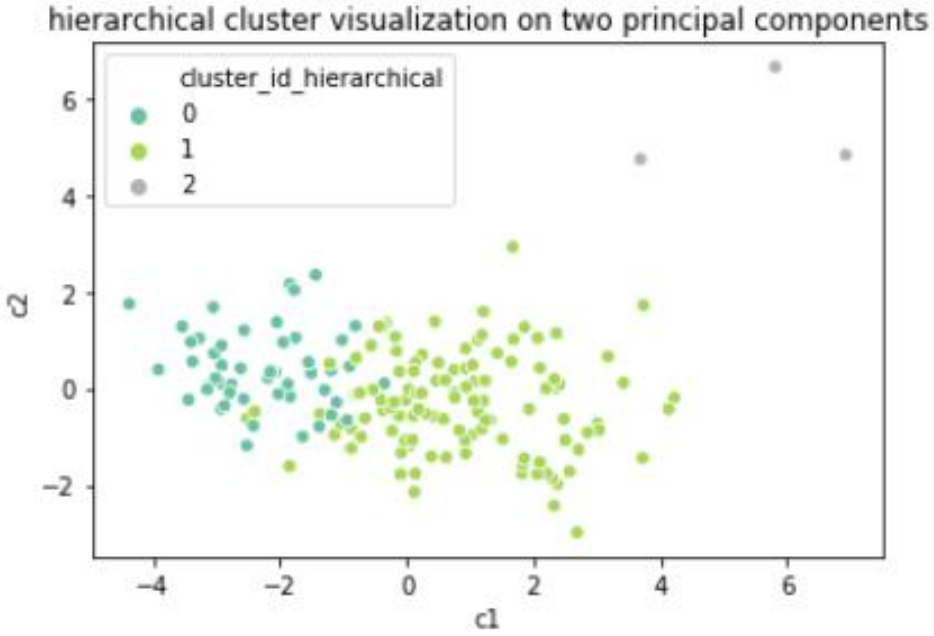
It can be visualized from the box plot that cluster 0 has minimum income

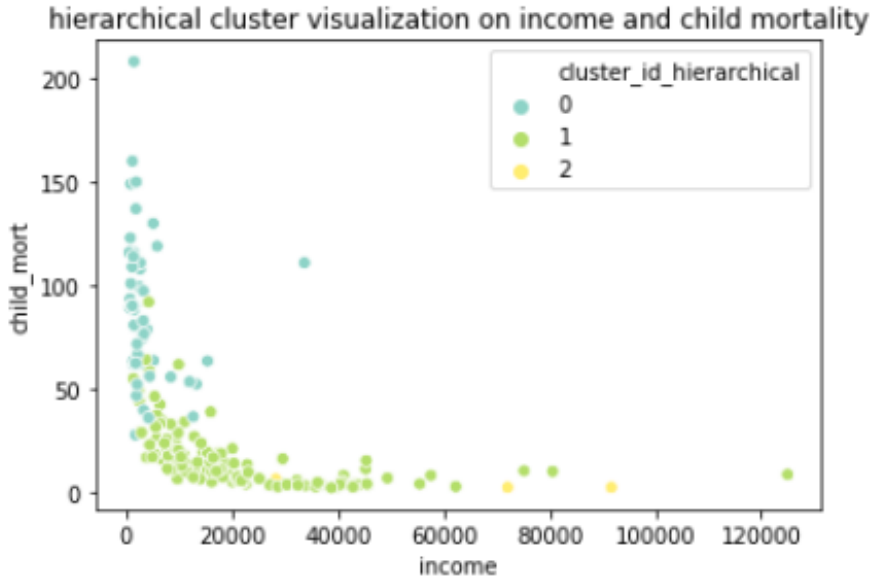
Clustering ID and gdpp



It can be visualized from the box plot that cluster 1 has minimum gdpp

Clusters visualization





Result Analysis

k means model seems to be more accurate having all most equal number of countries in each clusters.

Hierarchical clustering has only 3 countries as developed countries even united states and many other developed coming under developing country according to hierarchical clustering model while k means seems to giving correct result so going forward with k means clustering model for final list of countries.

hierarchical_count

cluster_id_hierarchical	
0	50
1	114
2	3

k_means_count

cluster_id_k_means	
0	85
1	47
2	35

Final List Of Countries where money should be spent

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

According to k means cluster, cluster 1 has minimum income, minimum gdpp and highest child mortality countries. So, selecting cluster 1 countries for final list.