Hierarchical Clustering: HELP International

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.

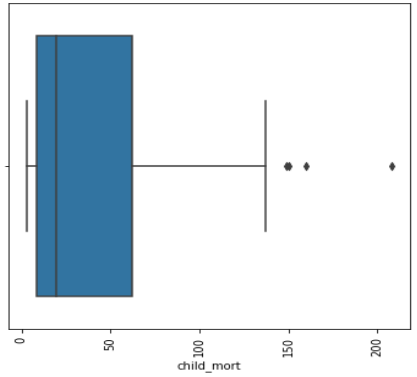
Objective

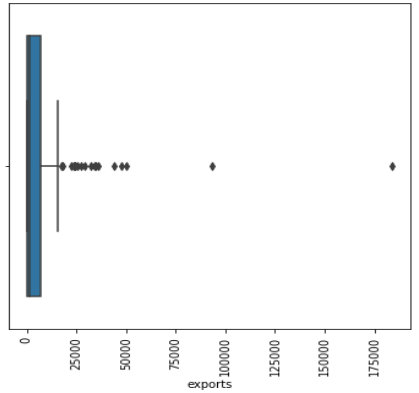
- To mention the countries that are in direst need of aid on the basis of socio-economic and health factors determining the overall development of the country.
- Suggesting the countries to CEO at least 5 countries which are in direst need of aid from the analysis work that you perform.

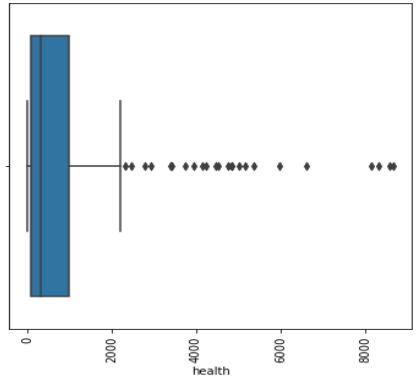
Analysis Approach

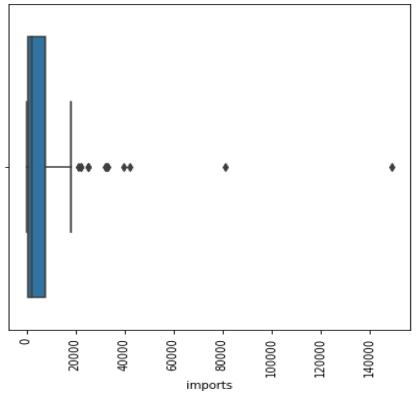
- **Data Reading and Data Cleaning:** Importing the data and reading it for its shape info, checking the data incase it has any duplicates, null values that needs cleaning. Some columns were in %of gddp form so converted them into absolute values.
- **Visualizing Data:** Some outliers were found after visualizing the data that needed treatment.
- **Handling Outliers:** There were outliers in almost every column. Some outliers like in gdpp column for example, have outliers on high end of spectrum which we can remove safely because high gdpp countries won't need urgent aid. For columns such as 'child_mort', 'inflation', 'total_fer', we removed lower range outliers only (lower capping). For rest of the columns, we removed upper range outliers only (upper capping).

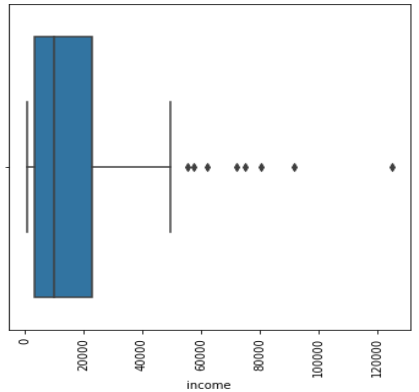
- **Handling Outliers:** There were outliers in almost every column. Some outliers like in gdpp column for example, have outliers on high end of spectrum which we can remove safely because high gdpp countries won't need urgent aid. For columns such as 'child_mort', 'inflation', 'total_fer', we removed lower range outliers only (lower capping). For rest of the columns, we removed upper range outliers only (upper capping).

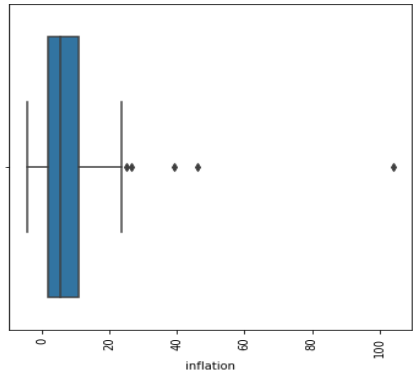


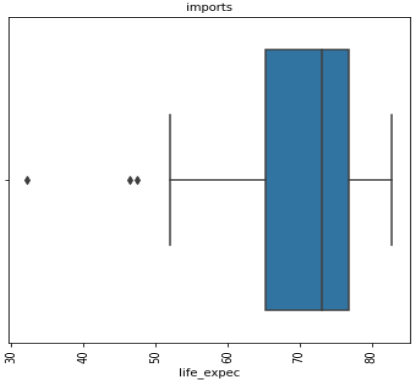


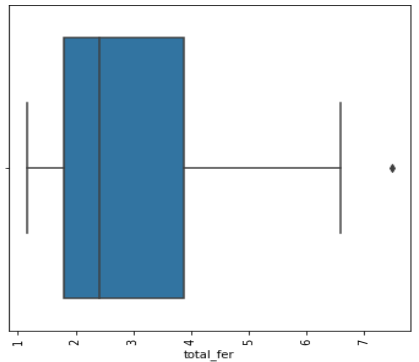


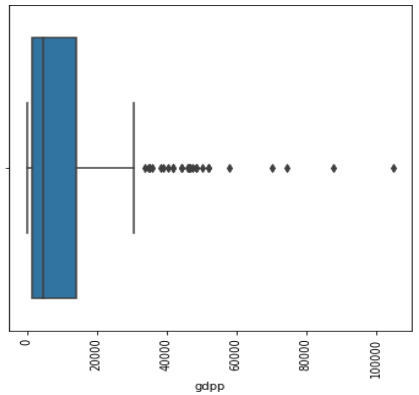




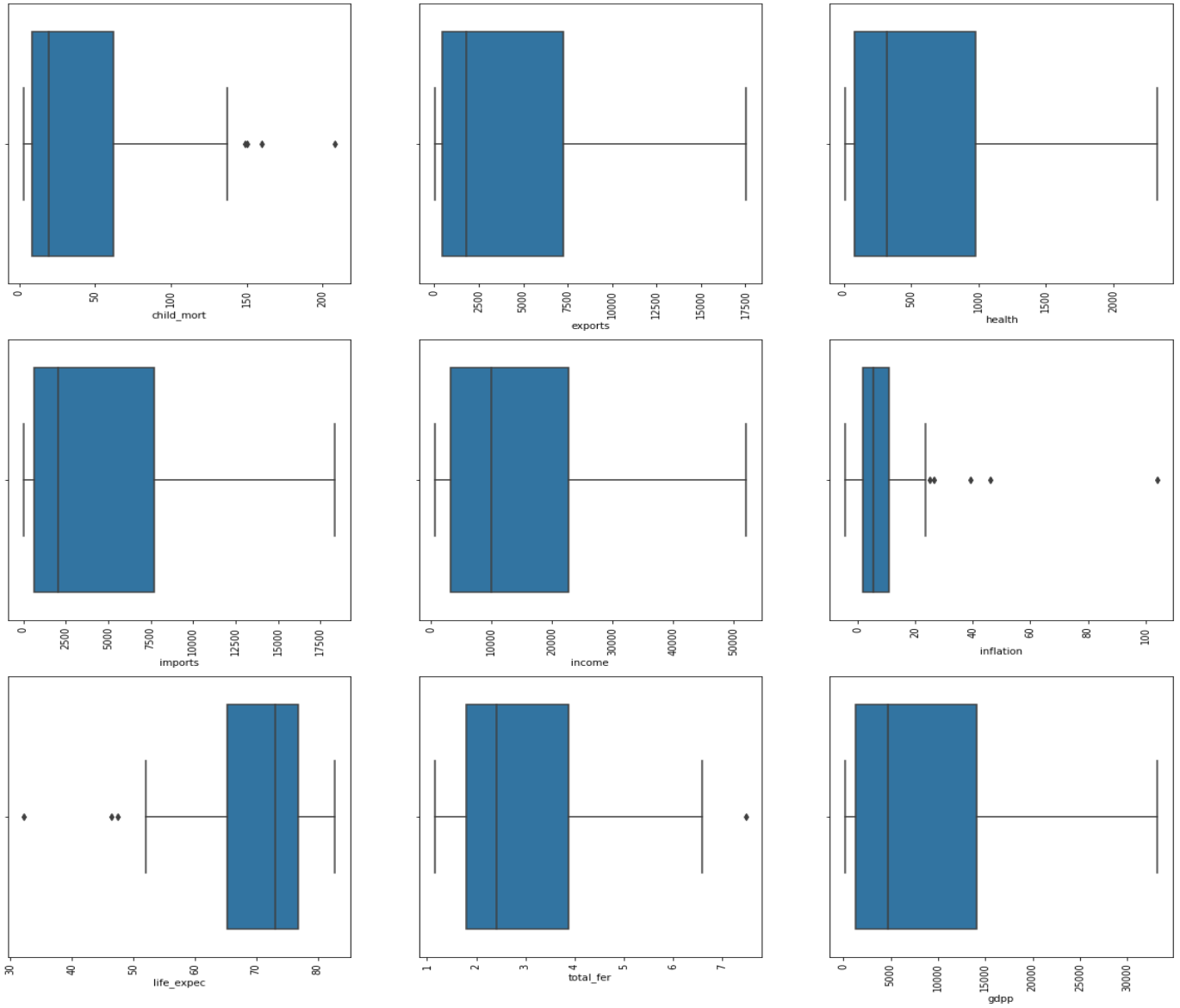






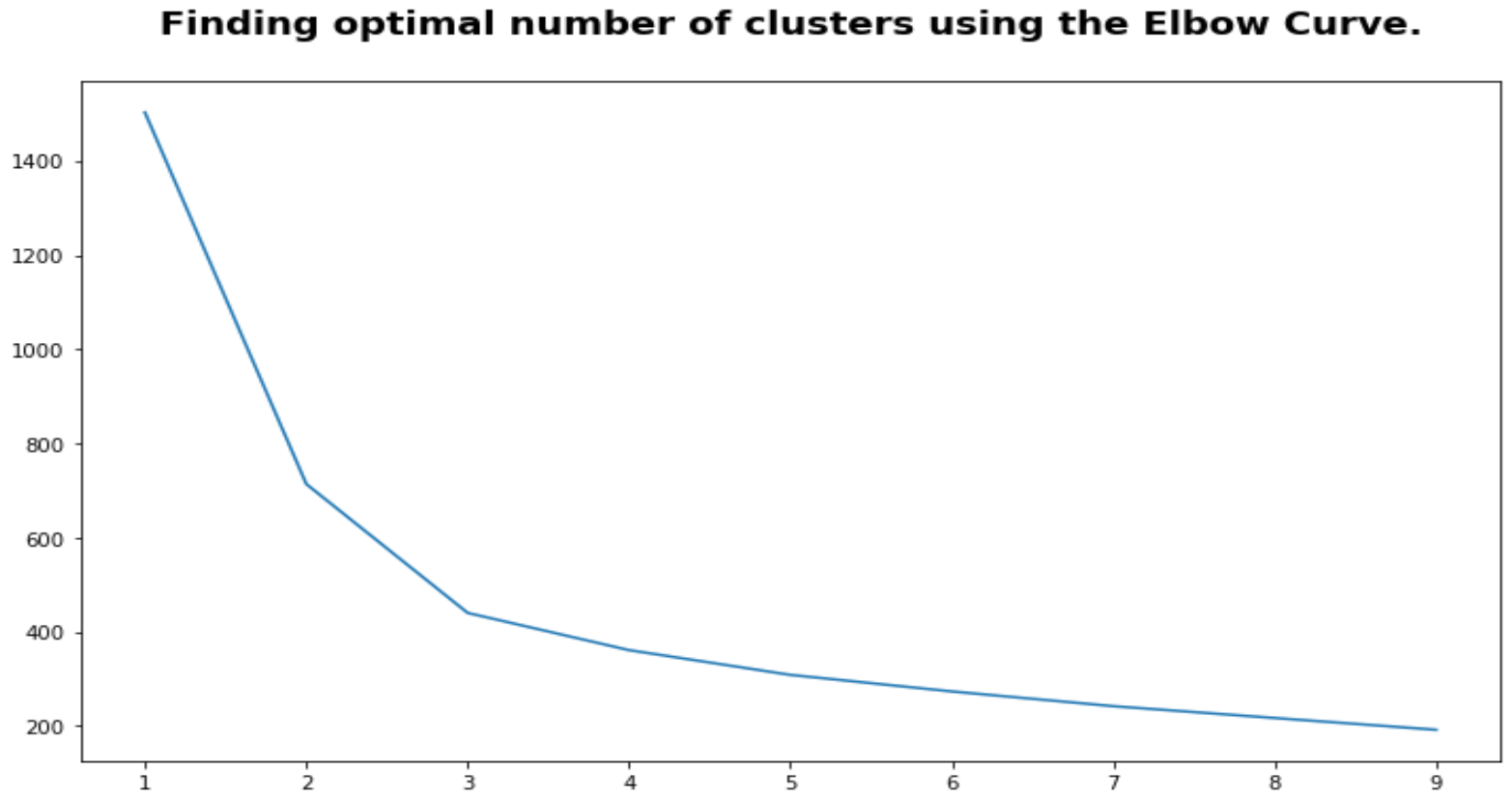


Outliers are found in each column.

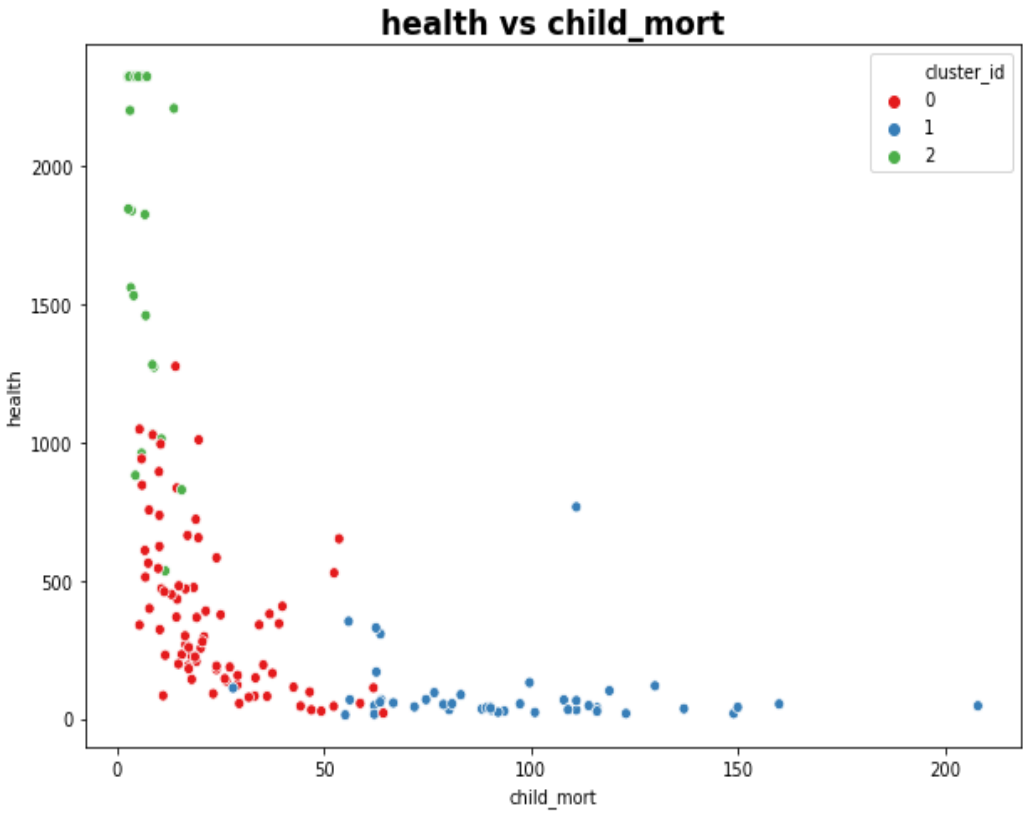


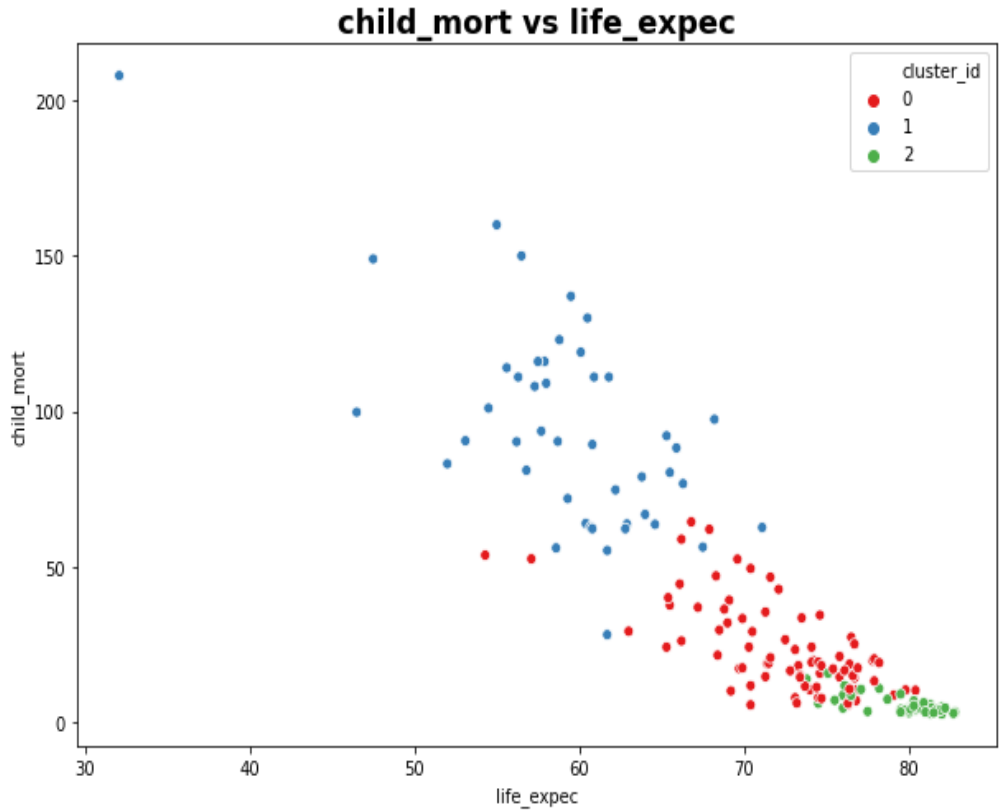
We can see, capping has been done successfully for respective columns and the outliers have been handled.

- **Scaling Data:** Standardizing all the continuous variables.
- **Hopkins Test:** Hopkins statistics is used for checking the cluster tendency of the dataset. We got Hopkins score of 0.876, which seems to be a good one for further clustering.
- **K-means Clustering:** Identifying the “k” through silhouette analysis and elbow curve. Then forming the cluster on scaled data, then adding the cluster id on original data for better interpretation of data. And visualizing the clusters.

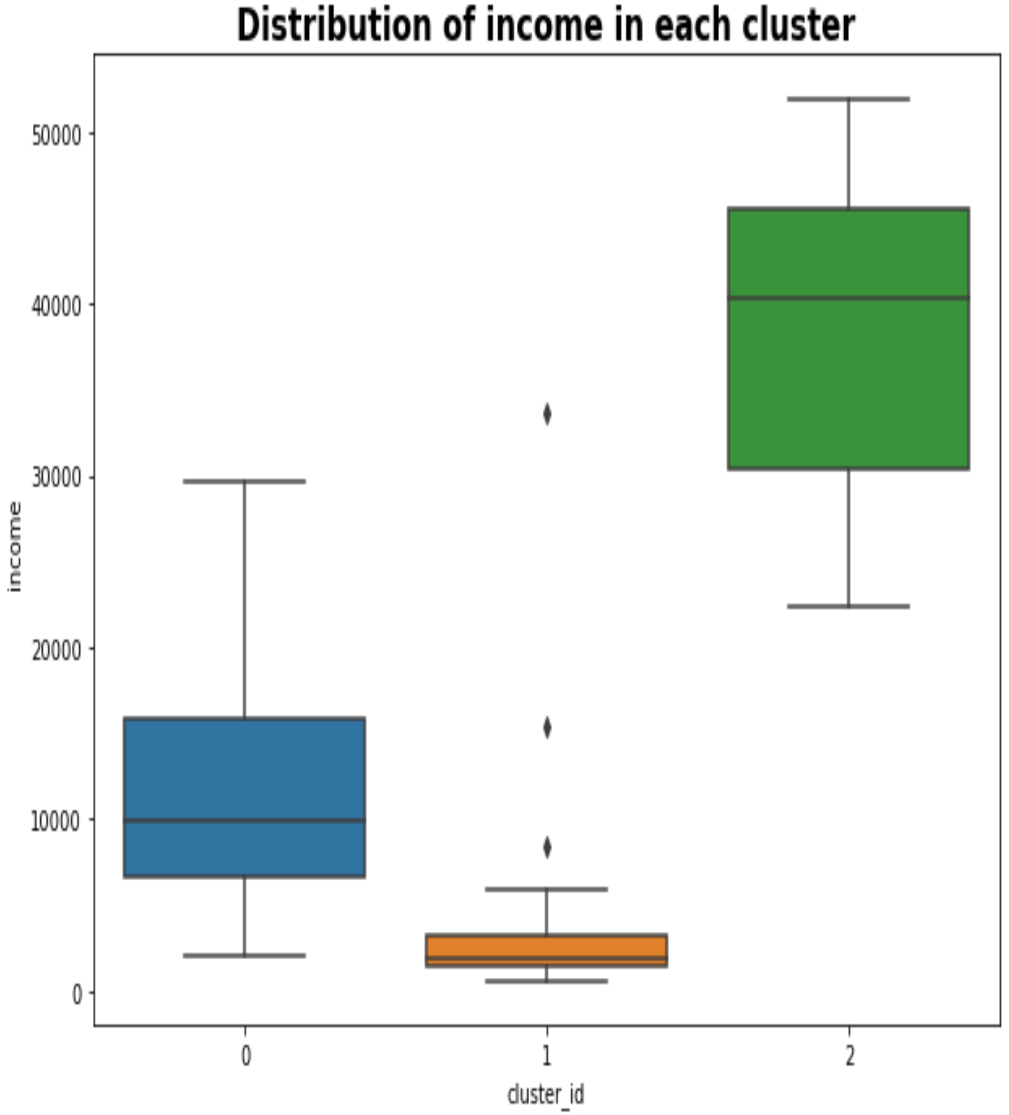


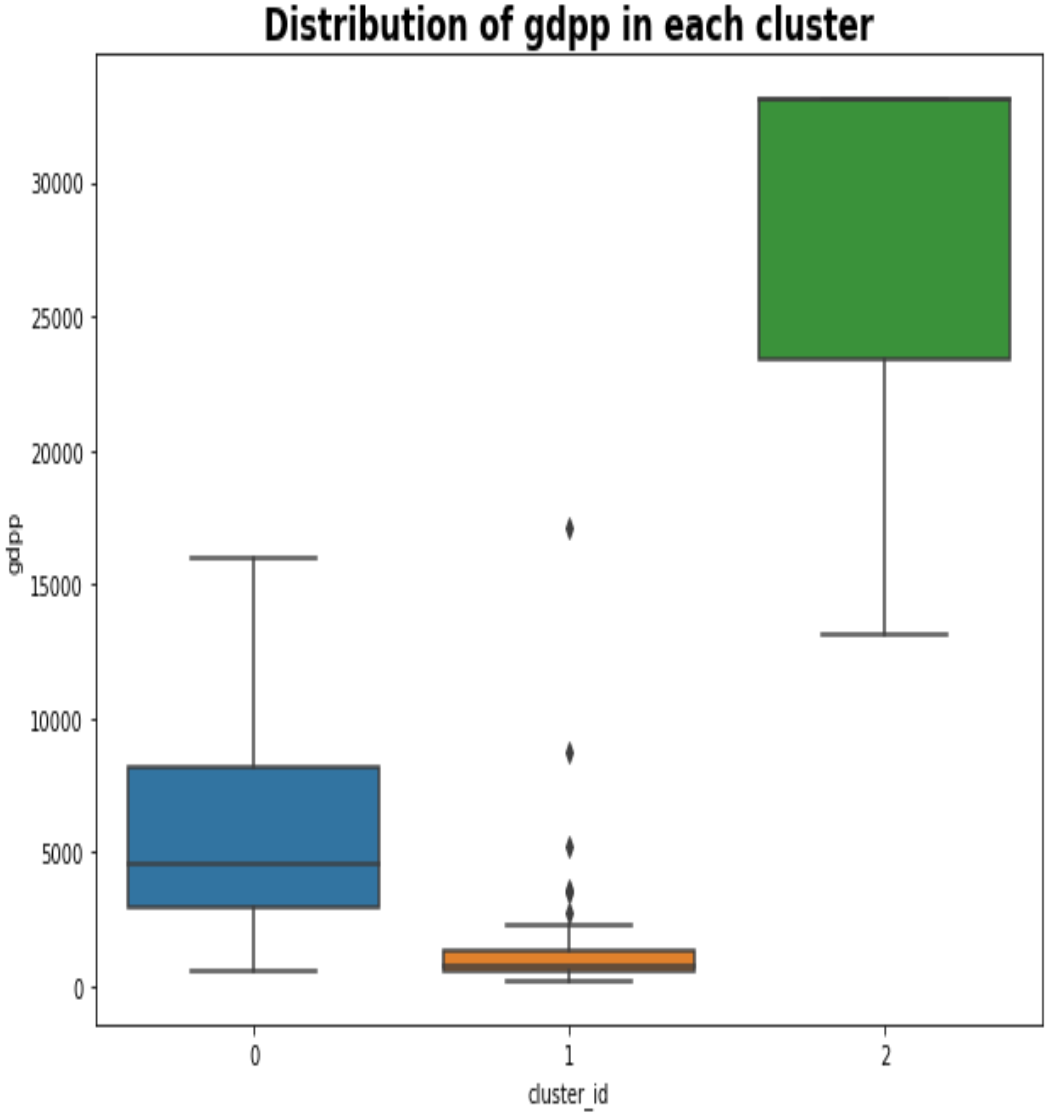
Looking at the elbow curve it looks good to proceed with either 4 or 5 clusters.

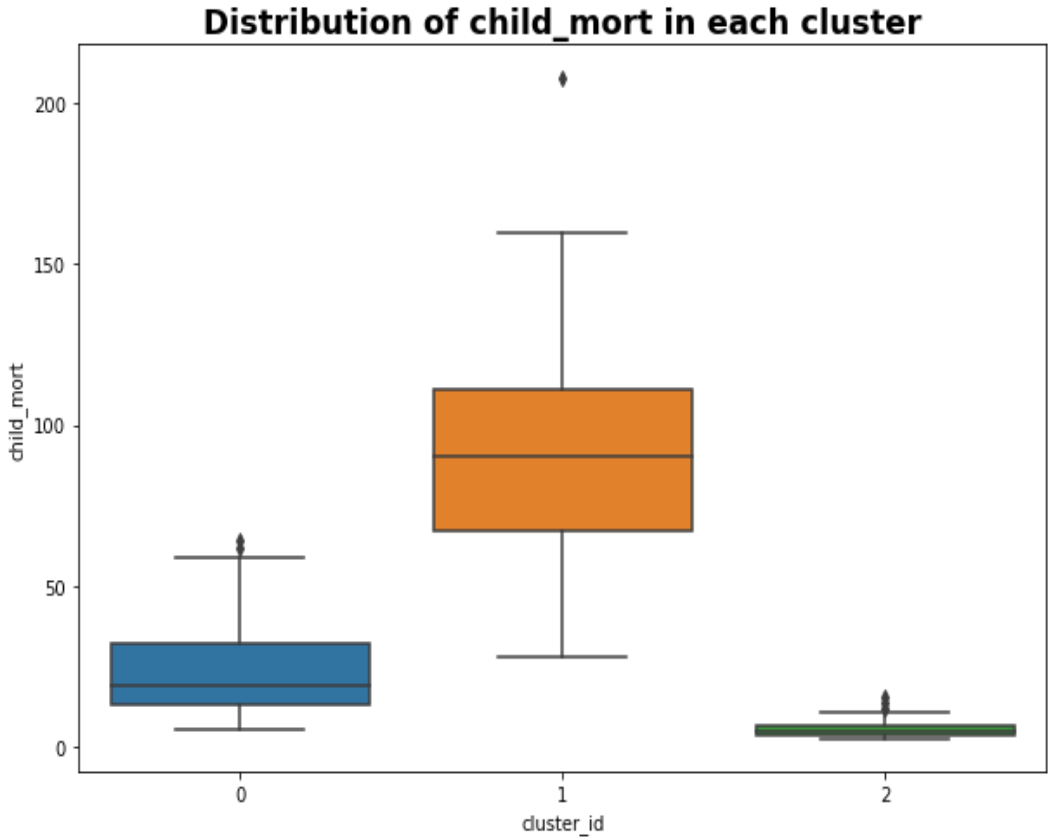




We can infer from graphs that in cluster 1 health expenditure is low and child mortality is quite high and also, life expectancy is very low.


Distribution of income in each cluster
income
50000
40000
30000
20000
10000
0
0
1
2
cluster_id


Distribution of gdpp in each cluster
gdpp
30000
25000
20000
15000
10000
5000
0
0
1
2
cluster_id



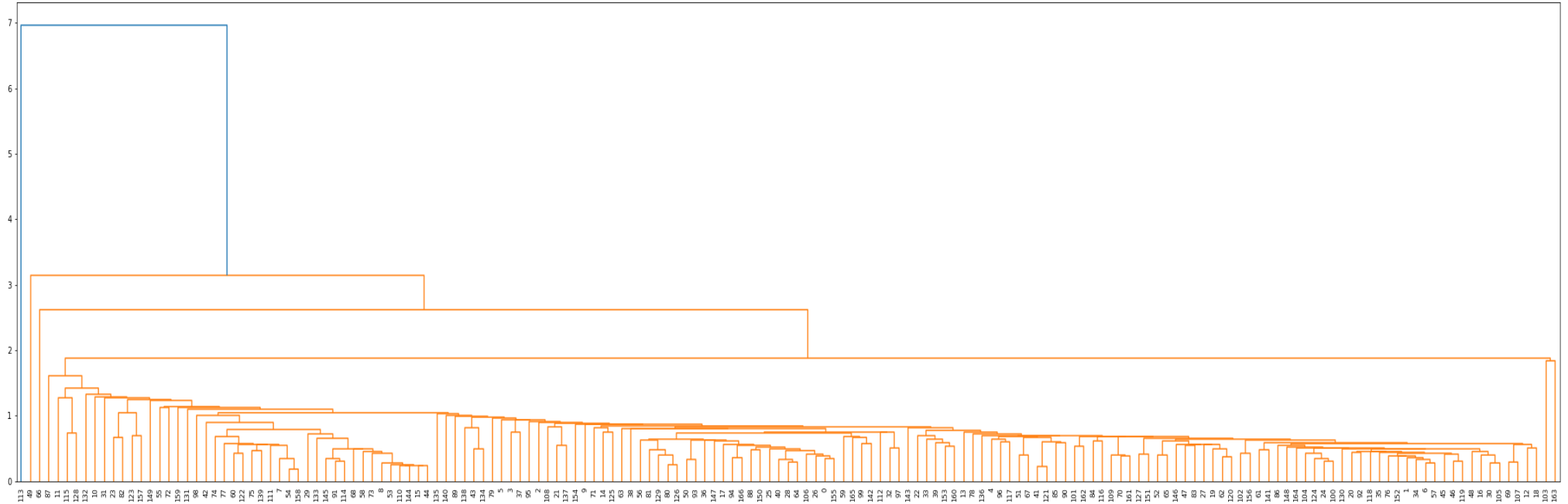
	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_id
0	Haiti	208.0	101.286	45.7442	428.314	1500.0	5.45	32.1	3.33	662.0	1
1	Sierra Leone	160.0	67.032	52.2690	137.655	1220.0	17.20	55.0	5.20	399.0	1
2	Chad	150.0	330.096	40.6341	390.195	1930.0	6.39	56.5	6.59	897.0	1
3	Central African Republic	149.0	52.628	17.7508	118.190	888.0	2.01	47.5	5.21	446.0	1
4	Mali	137.0	161.424	35.2584	248.508	1870.0	4.37	59.5	6.55	708.0	1
5	Nigeria	130.0	589.490	118.1310	405.420	5150.0	104.00	60.5	5.84	2330.0	1
6	Niger	123.0	77.256	17.9568	170.868	814.0	2.55	58.8	7.49	348.0	1
7	Angola	119.0	2199.190	100.6050	1514.370	5900.0	22.40	60.1	6.16	3530.0	1
8	Congo, Dem. Rep.	116.0	137.274	26.4194	165.664	609.0	20.80	57.5	6.54	334.0	1
9	Burkina Faso	116.0	110.400	38.7550	170.200	1430.0	6.81	57.9	5.87	575.0	1

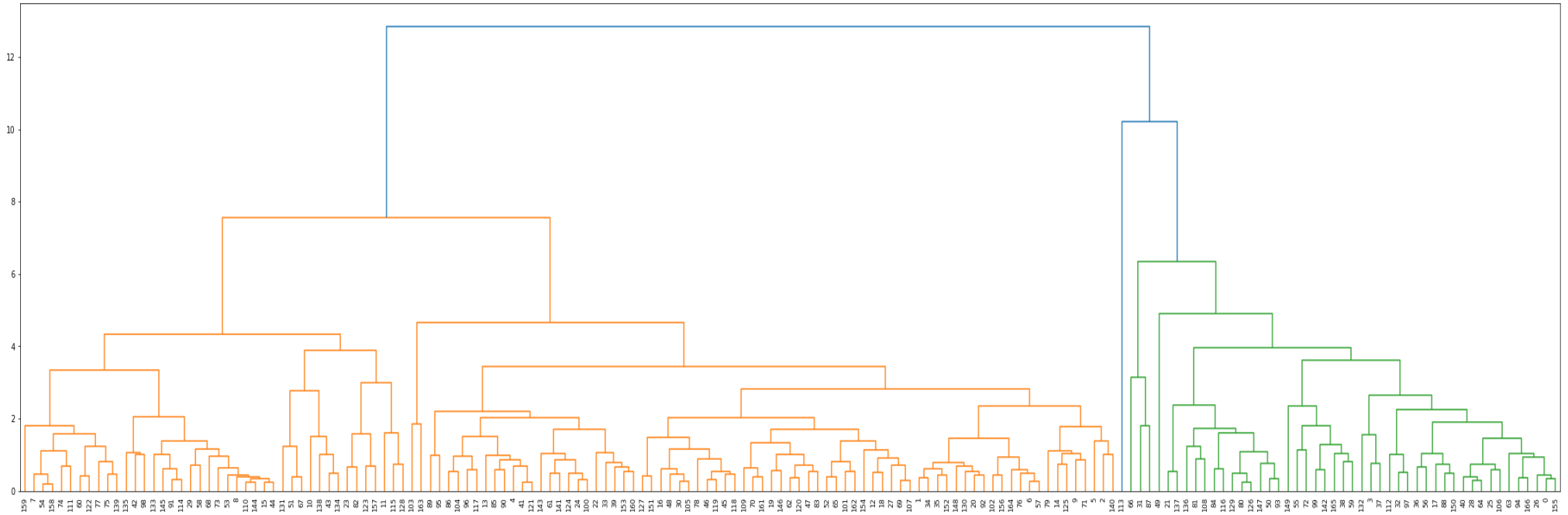
Cluster id 1 has very low income and gdpp and very high child mortality rate this cluster will be our focus.

Top 10 countries obtained from K-Means Models are:

- Haiti
- Sierra Leone
- Chad
- Central African Republic
- Mali
- Nigeria
- Niger
- Angola
- Congo, Dem. Rep. and Burkina Faso

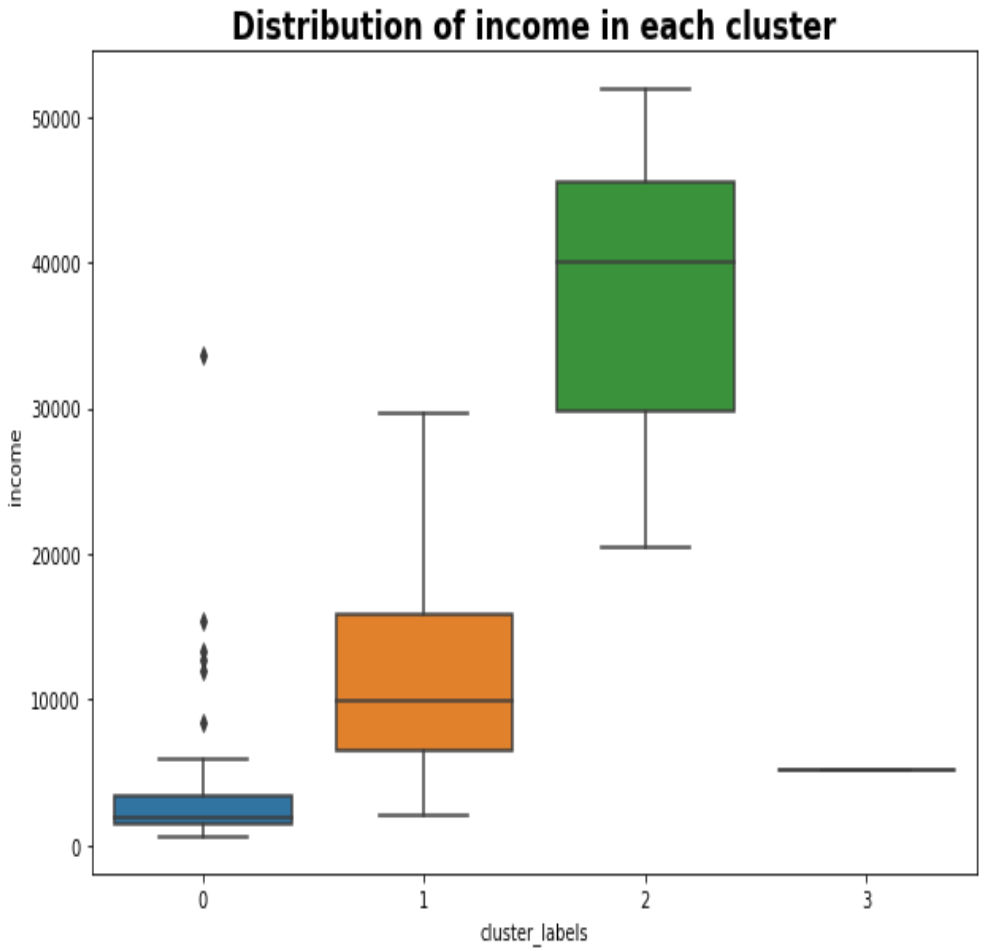
- **Hierarchical Clustering:** Identifying optimal number for k by analyzing dendrogram. Then forming the cluster on scaled data and adding the cluster label to original data for better interpretation. Visualization of clusters was also done.

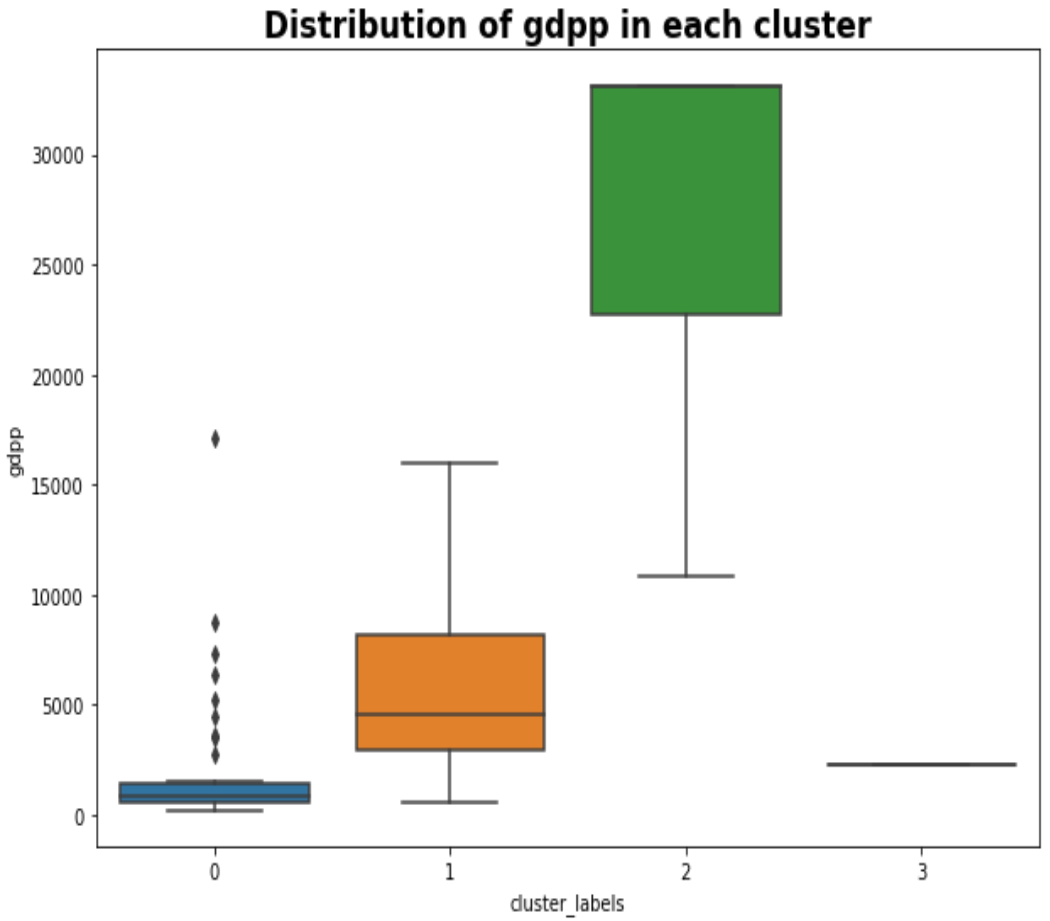


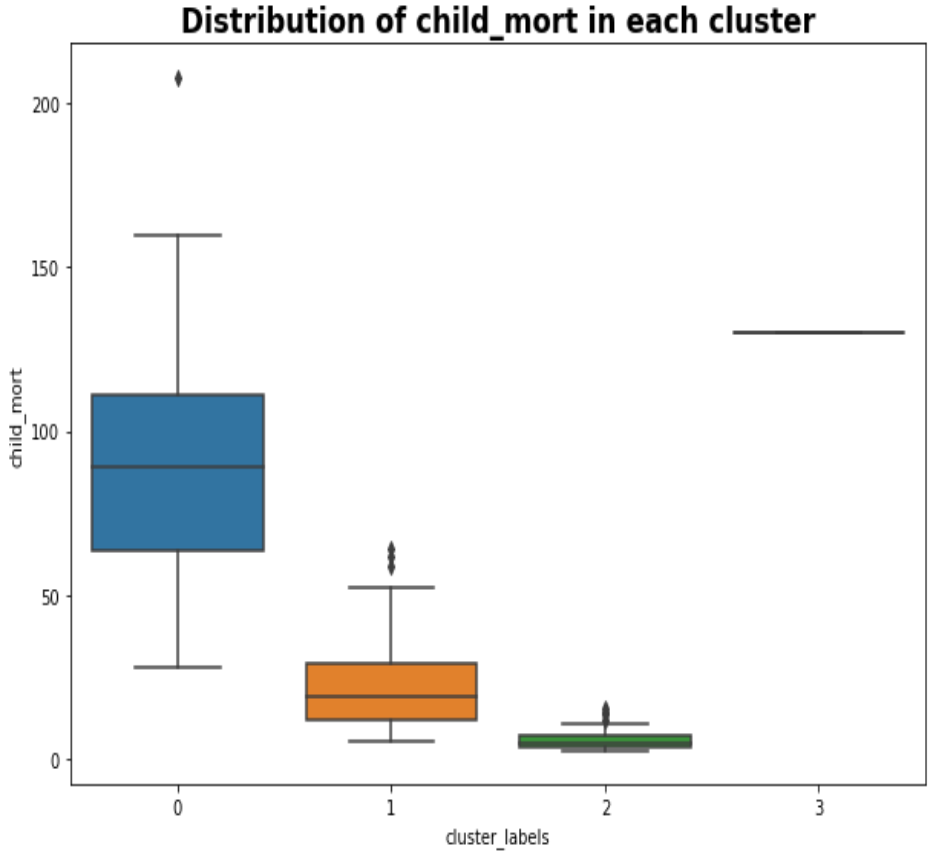


We will go with the complete linkage hierarchical clustering because single linkage is more complex and cluster formation is not good in single linkage.

Looking at dendrogram of hierarchical clustering there seem to be 3 clusters.


Distribution of income in each cluster
income
cluster_labels
0
10000
20000
30000
40000
50000
0
1
2
3


Distribution of gdpp in each cluster
gdpp
cluster_labels
0
5000
10000
15000
20000
25000
30000
0
1
2
3



	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_labels
0	Haiti	208.0	101.286	45.7442	428.314	1500.0	5.45	32.1	3.33	662.0	0
1	Sierra Leone	160.0	67.032	52.2690	137.655	1220.0	17.20	55.0	5.20	399.0	0
2	Chad	150.0	330.096	40.6341	390.195	1930.0	6.39	56.5	6.59	897.0	0
3	Central African Republic	149.0	52.628	17.7508	118.190	888.0	2.01	47.5	5.21	446.0	0
4	Mali	137.0	161.424	35.2584	248.508	1870.0	4.37	59.5	6.55	708.0	0
5	Niger	123.0	77.256	17.9568	170.868	814.0	2.55	58.8	7.49	348.0	0
6	Angola	119.0	2199.190	100.6050	1514.370	5900.0	22.40	60.1	6.16	3530.0	0
7	Congo, Dem. Rep.	116.0	137.274	26.4194	165.664	609.0	20.80	57.5	6.54	334.0	0
8	Burkina Faso	116.0	110.400	38.7550	170.200	1430.0	6.81	57.9	5.87	575.0	0
9	Guinea-Bissau	114.0	81.503	46.4950	192.544	1390.0	2.97	55.6	5.05	547.0	0

Countries in cluster 0 have low income, low gdpp and high child_mort rate. Country in cluster 3 appears to be in severe need of aid.

Top 10 countries obtained from Hierarchical Clustering Models are:

- Haiti
- Sierra Leone
- Chad
- Central African Republic
- Mali
- Nigeria
- Niger
- Angola
- Congo, Dem. Rep.
- Burkina Faso

- **Decision Making:** Successfully identified the top 10 countries by analyzing both model which are in direst need of aid.