Welcome to clustering Model

Categorising The Countries Based On Their Poverty Line For Funding As Well As For Other Facilities.

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 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.

Data exploration

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	Afghanistan	90.2	10.0	7.58	44.9	1610	9.44	56.2	5.82	553
1	Albania	16.6	28.0	6.55	48.6	9930	4.49	76.3	1.65	4090
2	Algeria	27.3	38.4	4.17	31.4	12900	16.10	76.5	2.89	4460
3	Angola	119.0	62.3	2.85	42.9	5900	22.40	60.1	6.16	3530
4	Antigua and Barbuda	10.3	45.5	6.03	58.9	19100	1.44	76.8	2.13	12200

- This is our dataset containing all the details of countries relevant for categorising.
- We need to identify the top 5 countries who facing the poverty in its extreme.

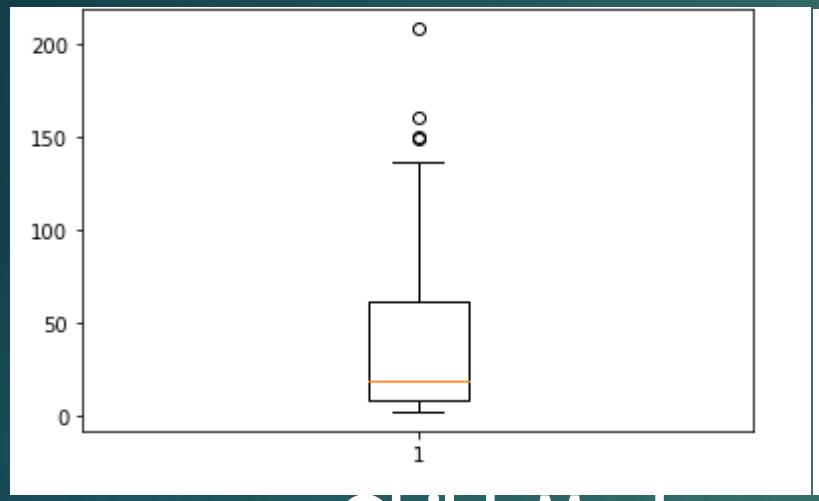
Data Handling and Cleaning

- These is most important part in the data analysis. We need make sure that there are no missing values or incorrect data types before we proceed to the analysis stage. These can achieved by following ways:
- For Missing Values: Some common techniques to treat this issue are
  - Dropping the rows containing the missing values more then 30%.
  - Imputing the missing values
  - Keep the missing values if they don't affect the analysis

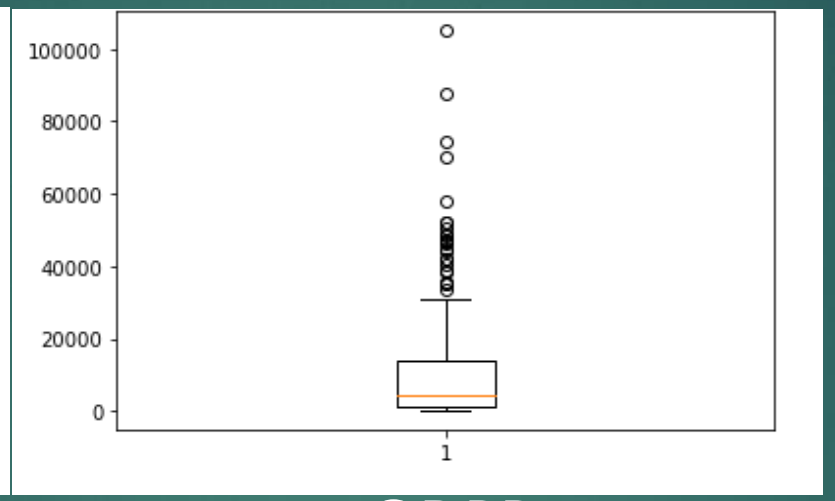
Incorrect Data Types:

  - Clean certain values
  - Clean and convert an entire column

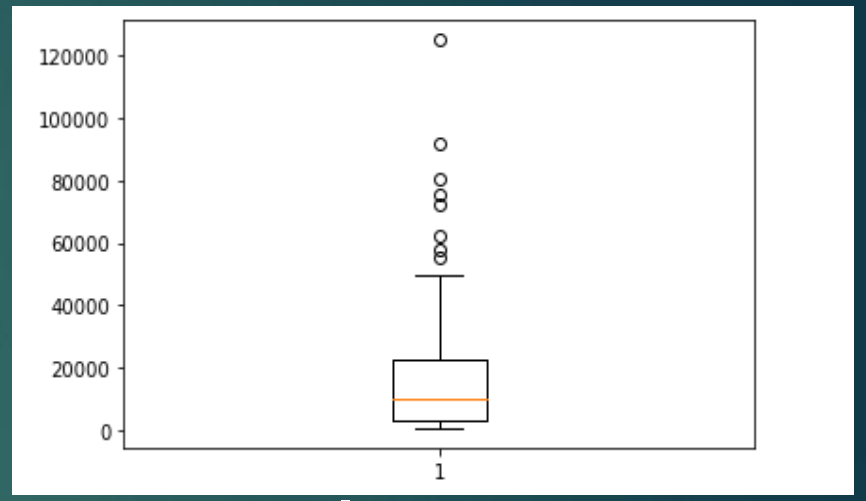
Outlier treatment



Child_Mort



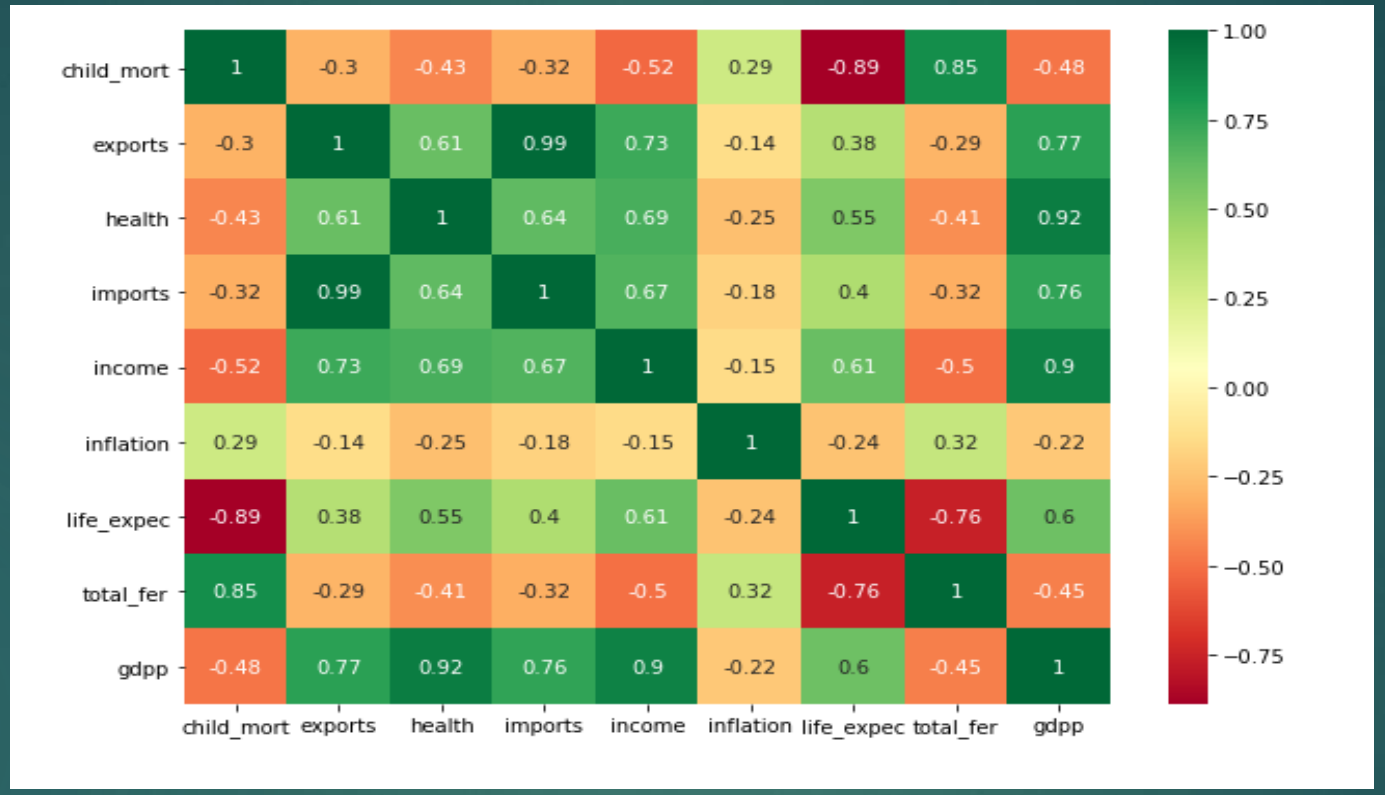
GDPP



Income

- We have taken **child_mort**, **Gdpp**, and **Income** as our main features for cluster profiling.
- In our case, we also have outliers for these features. So, its important to handle it carefully.
- For Child_mort variable, we shouldn't remove outliers, because it plays an vital role for clustering

Data visualization

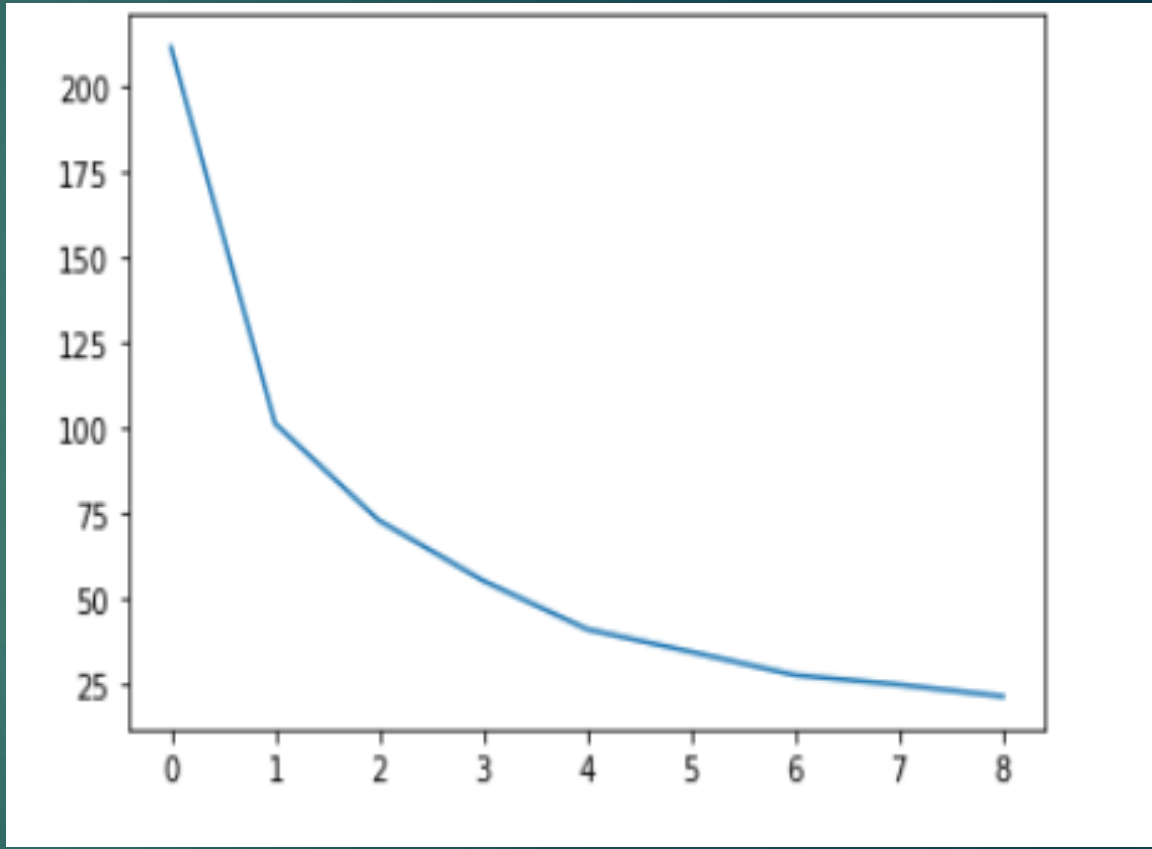


- From the plot above its clearly impling that, if the Gdpp and Exports are improving, then automatically the health improves and Child death will decrease simultaneously.

Clustering

Kmeans Clustering

- Here for 2 clusters, it shows better performance of cluster. But it's a bad idea to prefer 2 as K value.
- And for 3, the slope is not as good for the clustering profile.
- So, we have taken 4 as our K value, because it implies good for clustering due to better slope.

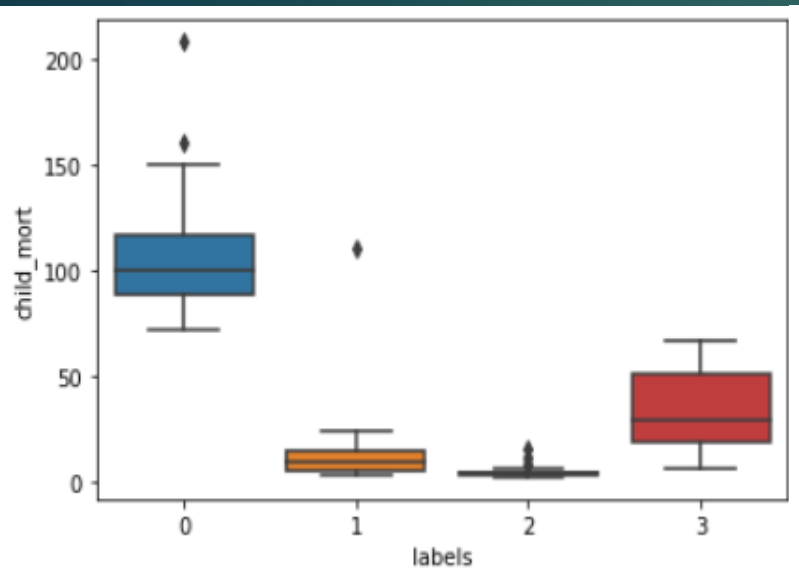


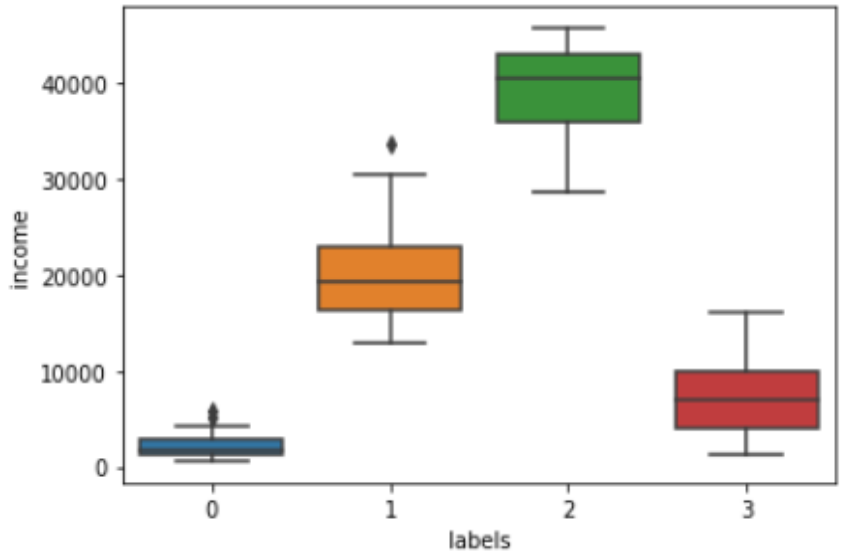
	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	labels
0	Afghanistan	90.2	55.30	41.9174	248.297	1610	9.44	56.2	5.82	553	0
1	Albania	16.6	1145.20	267.8950	1987.740	9930	4.49	76.3	1.65	4090	3
2	Algeria	27.3	1712.64	185.9820	1400.440	12900	16.10	76.5	2.89	4460	3
3	Angola	119.0	2199.19	100.6050	1514.370	5900	22.40	60.1	6.16	3530	0
4	Antigua and Barbuda	10.3	5551.00	735.6600	7185.800	19100	1.44	76.8	2.13	12200	1
...	...	...	...	...	...	...	...	...	...	...	...
162	Vanuatu	29.2	1384.02	155.9250	1565.190	2950	2.62	63.0	3.50	2970	3
163	Venezuela	17.1	3847.50	662.8500	2376.000	16500	45.90	75.4	2.47	13500	1
164	Vietnam	23.3	943.20	89.6040	1050.620	4490	12.10	73.1	1.95	1310	3
165	Yemen	56.3	393.00	67.8580	450.640	4480	23.60	67.5	4.67	1310	3
166	Zambia	83.1	540.20	85.9940	451.140	3280	14.00	52.0	5.40	1460	0

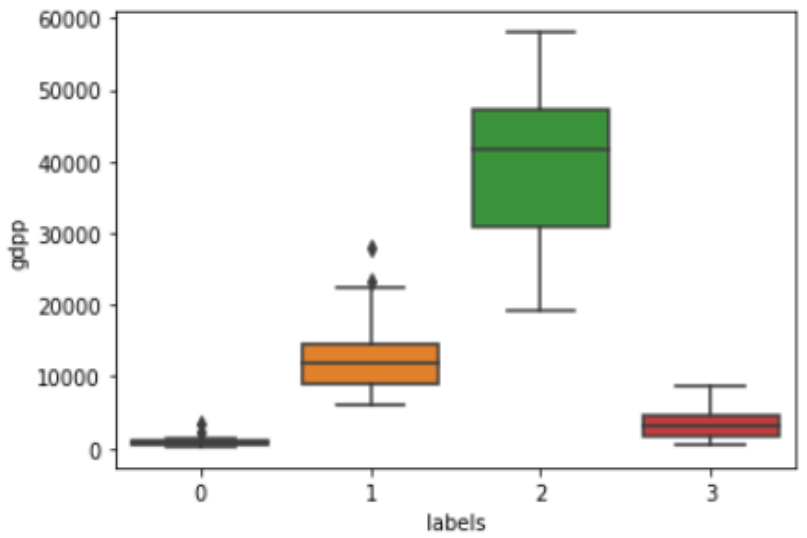
158 rows × 11 columns

- Thus we have labelled our dataset with 4 clusters (0,1,2,3)
- Now we can visualize it separetly to recognize the relavant cluster

Cluster visualizing

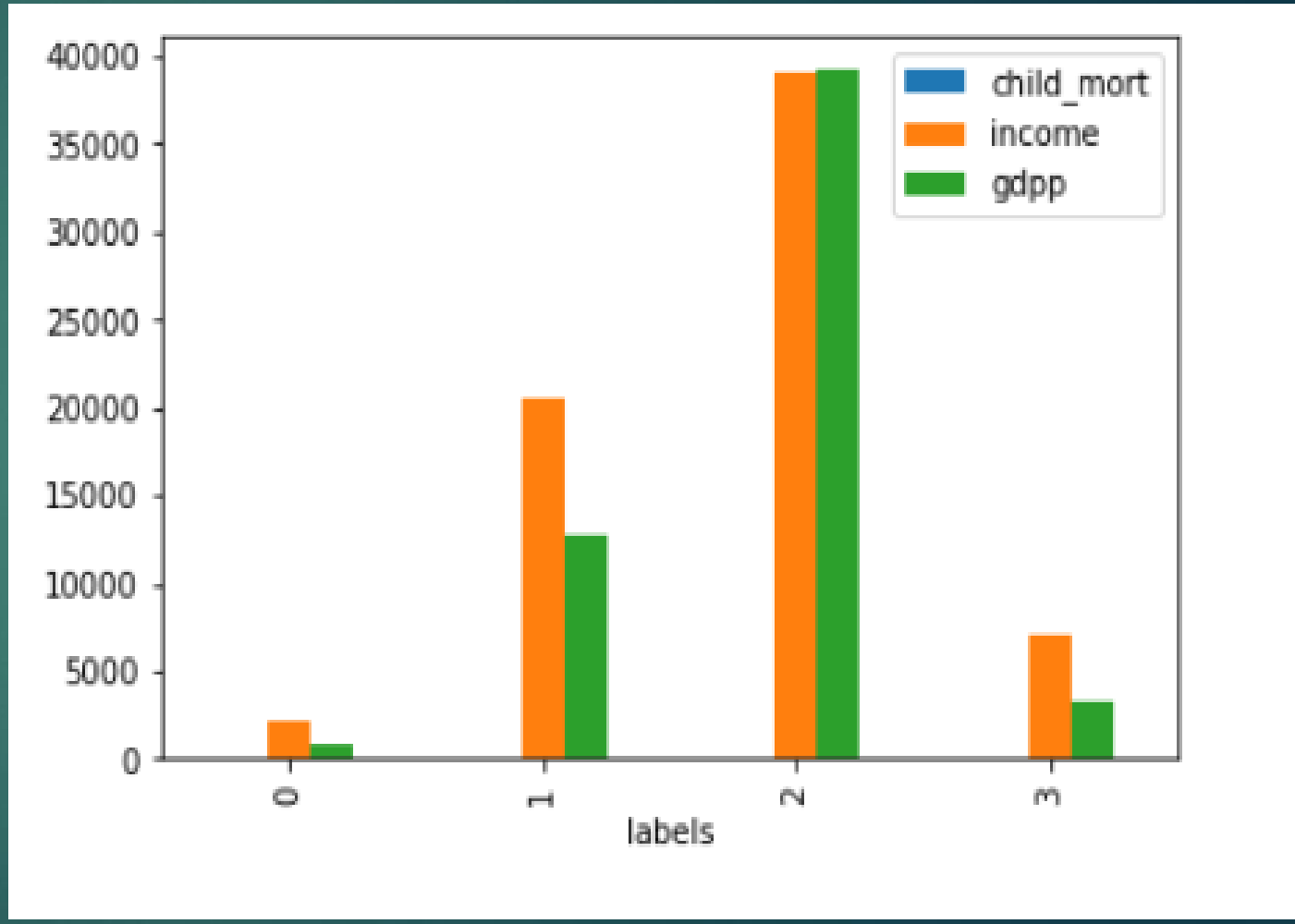






- ✓ From the boxplot for the main 3 features, we can infer that cluster 0 is our target segment to be focused
- ✓ We need the countries those having high child death, low Gdpp and income.
- ✓ So, the cluster 0 is highly satisfying our profile
- ✓ And countries under cluster 2 is a better developing countries compared to others.

- ✓It also implies to choose 0 as our target cluster.
- ✓So, lets identify the top 5 countries to be hired for funding



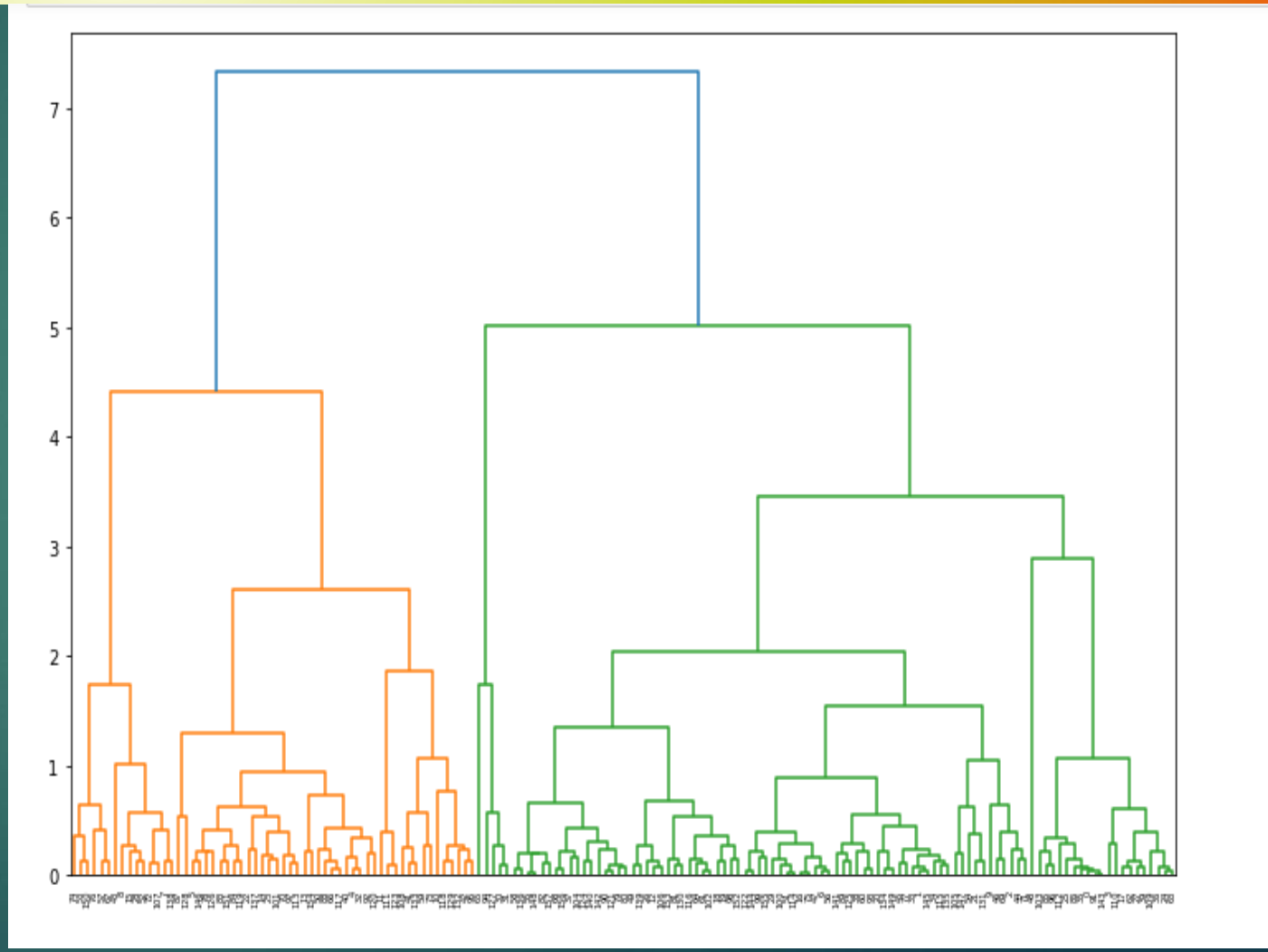
Top 5 countries

- ✓These are the most critical countries been identified as facing the poverty in its extreme level
- ✓So, they need the facilities as soon as possible to overcome their life

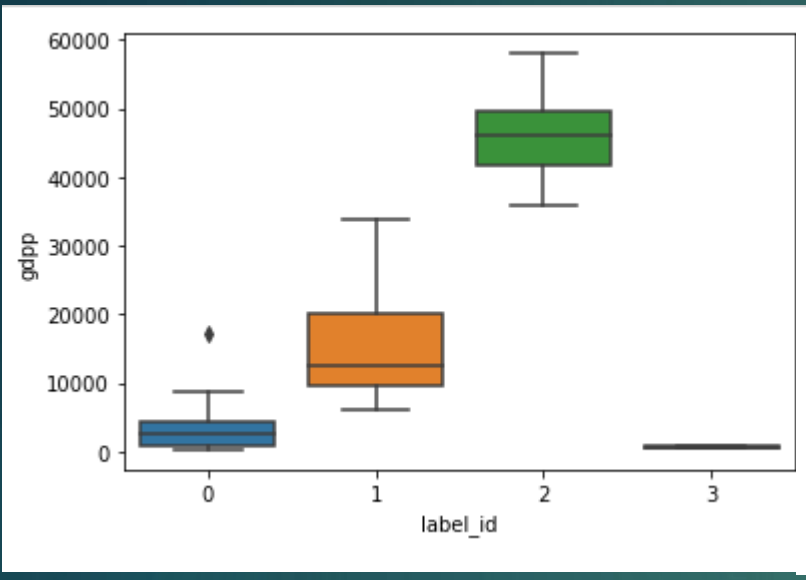
	country
0	Burundi
1	Liberia
2	Congo, Dem. Rep.
3	Niger
4	Sierra Leone

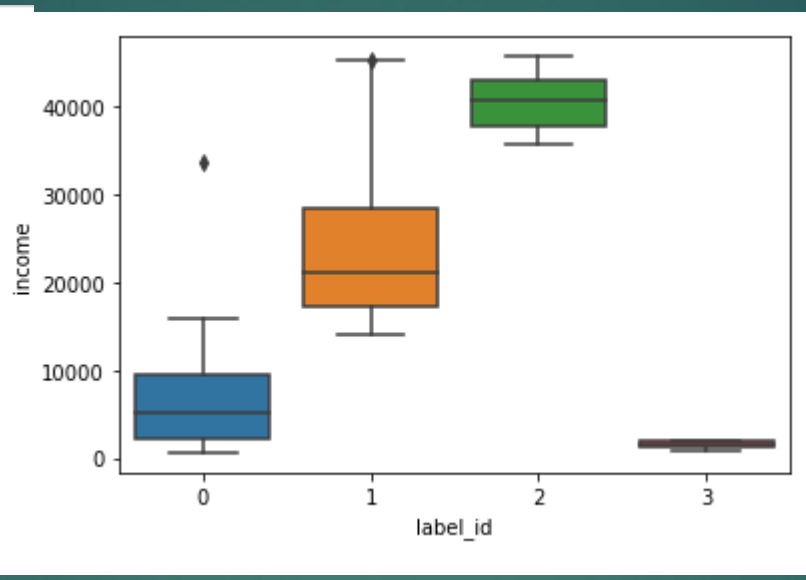
Hierarchical clustering

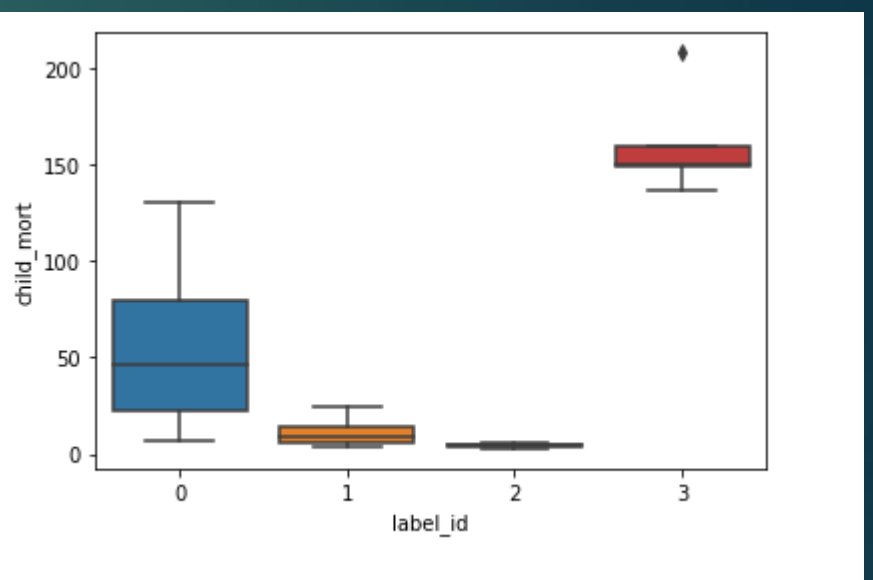
- It also showing to prefer 4 as our cluster.
- So, we can cut on the y axis 4 for 4 cluster



Cluster visualizing

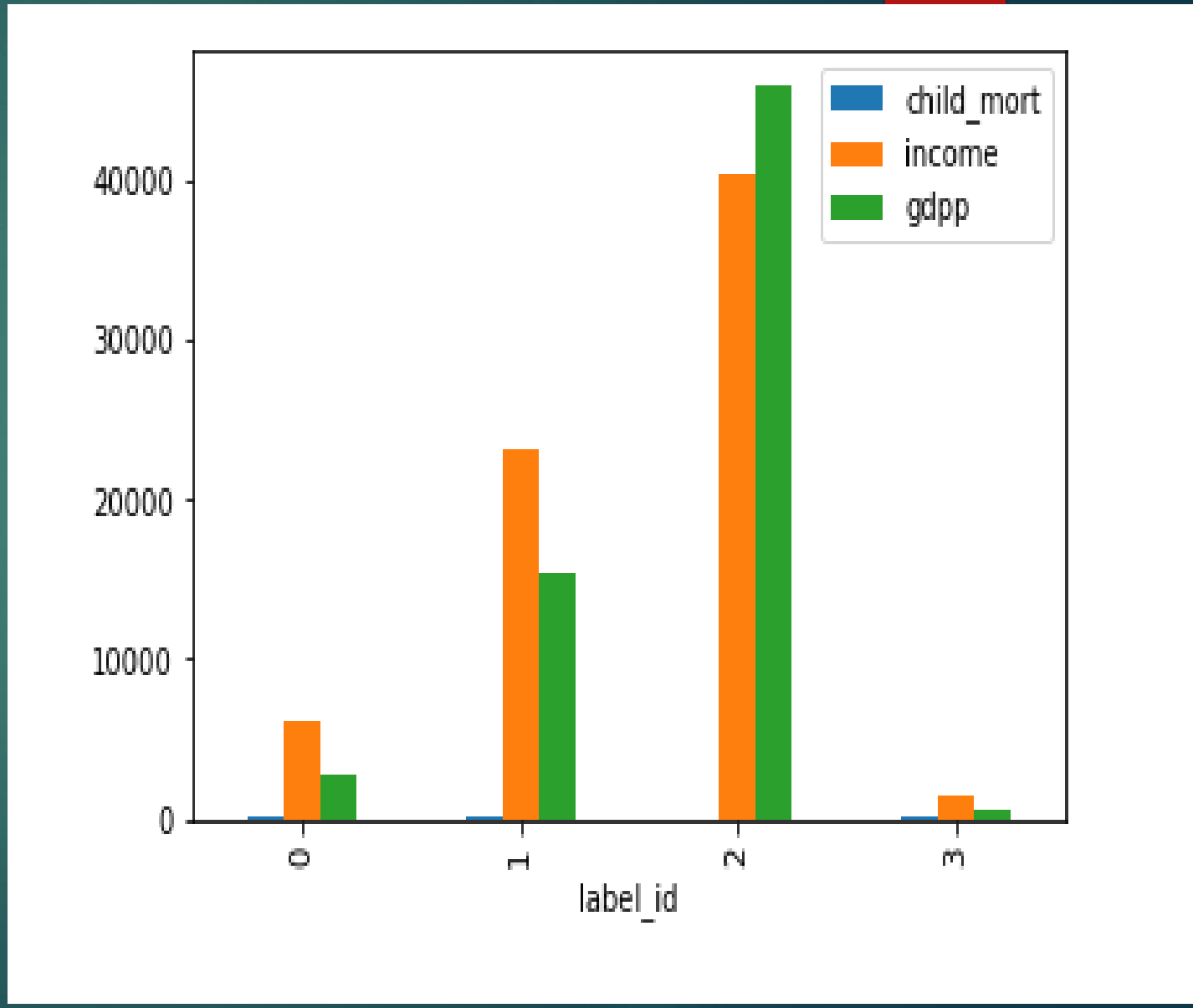






- ✓ From the boxplot for the main 3 features, we can infer that cluster 3 is our target segment to be focused
- ✓ We need the countries those having high child death, low Gdpp and income.
- ✓ So, the cluster 3 is highly satisfying our profile
- ✓ And countries under cluster 2 is a better developing countries compared to others.

- ✓It also implies to choose 3 as our target cluster.
- ✓So, lets identify the top 5 countries to be hired for funding



Top 5 countries

- ✓These are the most critical countries been identified as facing the poverty in its extreme level
- ✓So, they need the facilities as soon as possible to overcome their life

]:

	country
0	Sierra Leone
1	Central African Republic
2	Haiti
3	Mali
4	Chad

Thus we have identified the top 5 countries to be more focused.

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

Under Kmeans Clustering

]:

	country
0	Sierra Leone
1	Central African Republic
2	Haiti
3	Mali
4	Chad

Under Hierarchical clustering

Thank you