HELP INTERNATIONAL NGO CLUSTERING ASSIGNMENT

By-Ashmita Sarkar

OVERVIEW

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. With 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 for the countries that are in the direst need of aid.

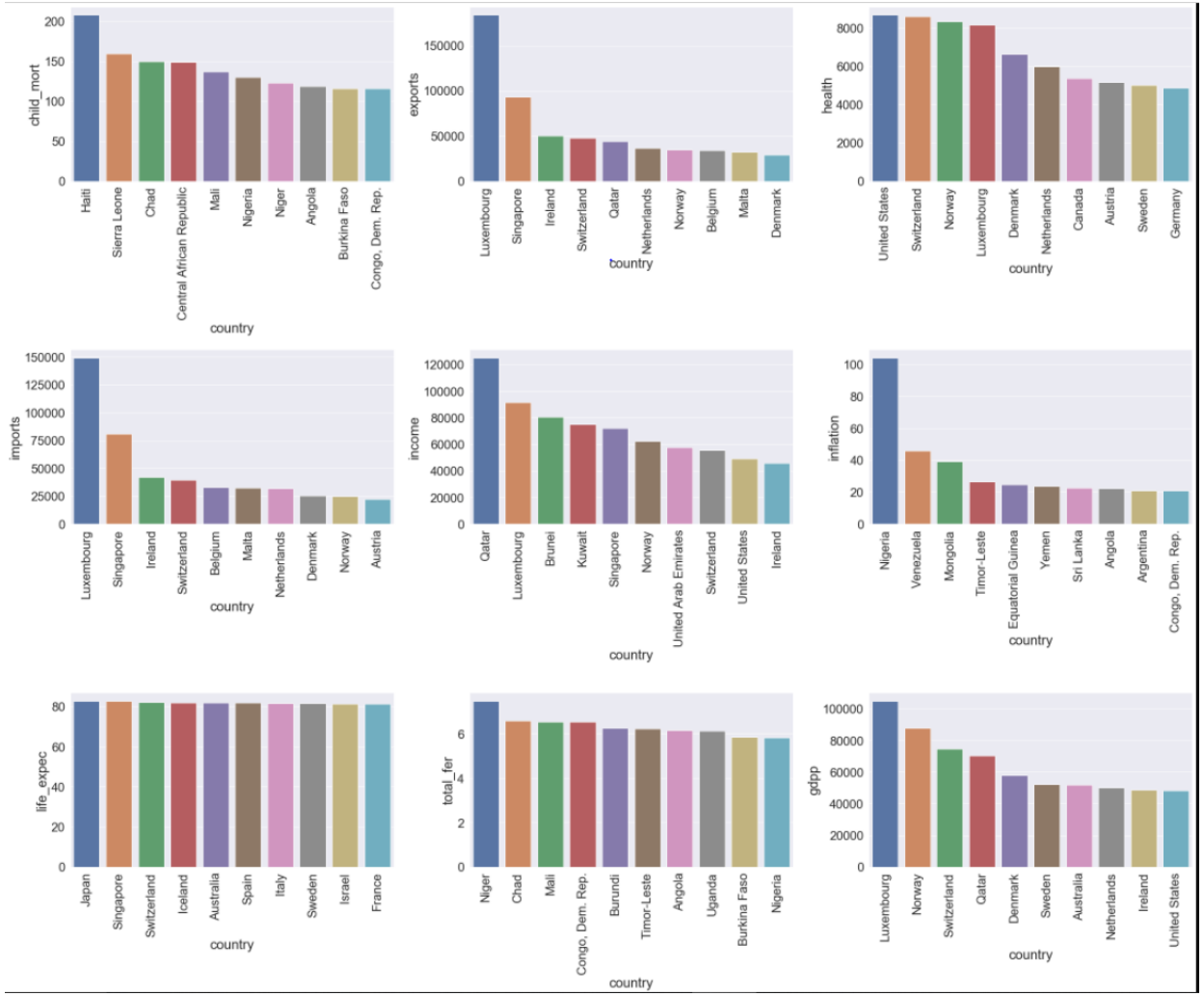
OBJECTIVE OF ASSIGNMENT

As a data analyst we need to Ccategorize the countries using some socio-economic and health factors that determine the overall development of the country. Then you need to suggest the countries in the direst need of aid which the CEO needs to focus on the most.

EXPLORATORY DATA ANALYSIS

BARPLOT

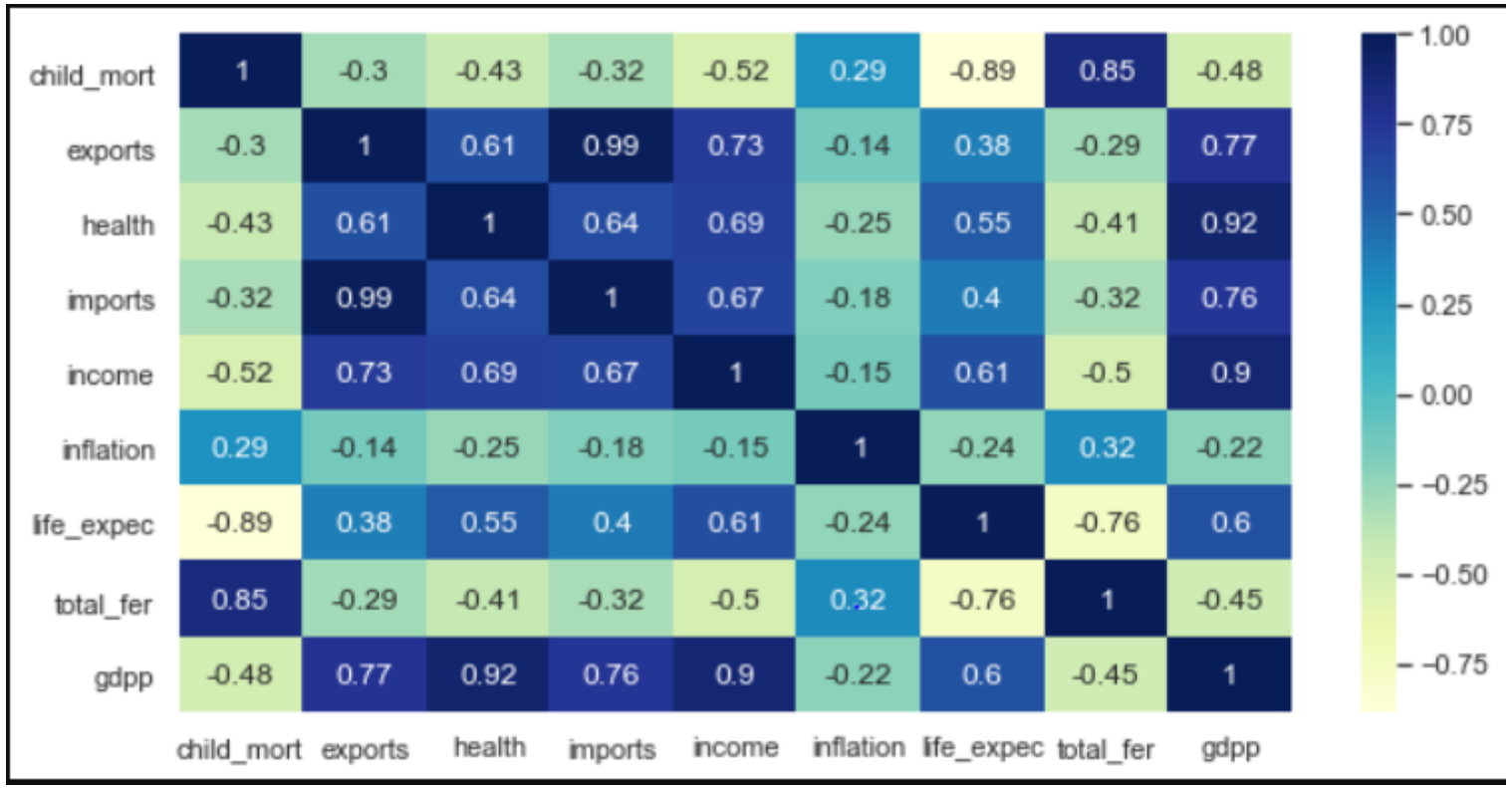
- Haiti has the highest child mortality rate.
- Luxembourg has the highest number of exports, imports and highest GDPP as well.
- Niger has the highest fertility rate
- Nigeria has the most inflation.
- Income of Qatar is the highest.
- United States spend most on Health on GDPP per capita



HEATMAP

We can see high correlation between:

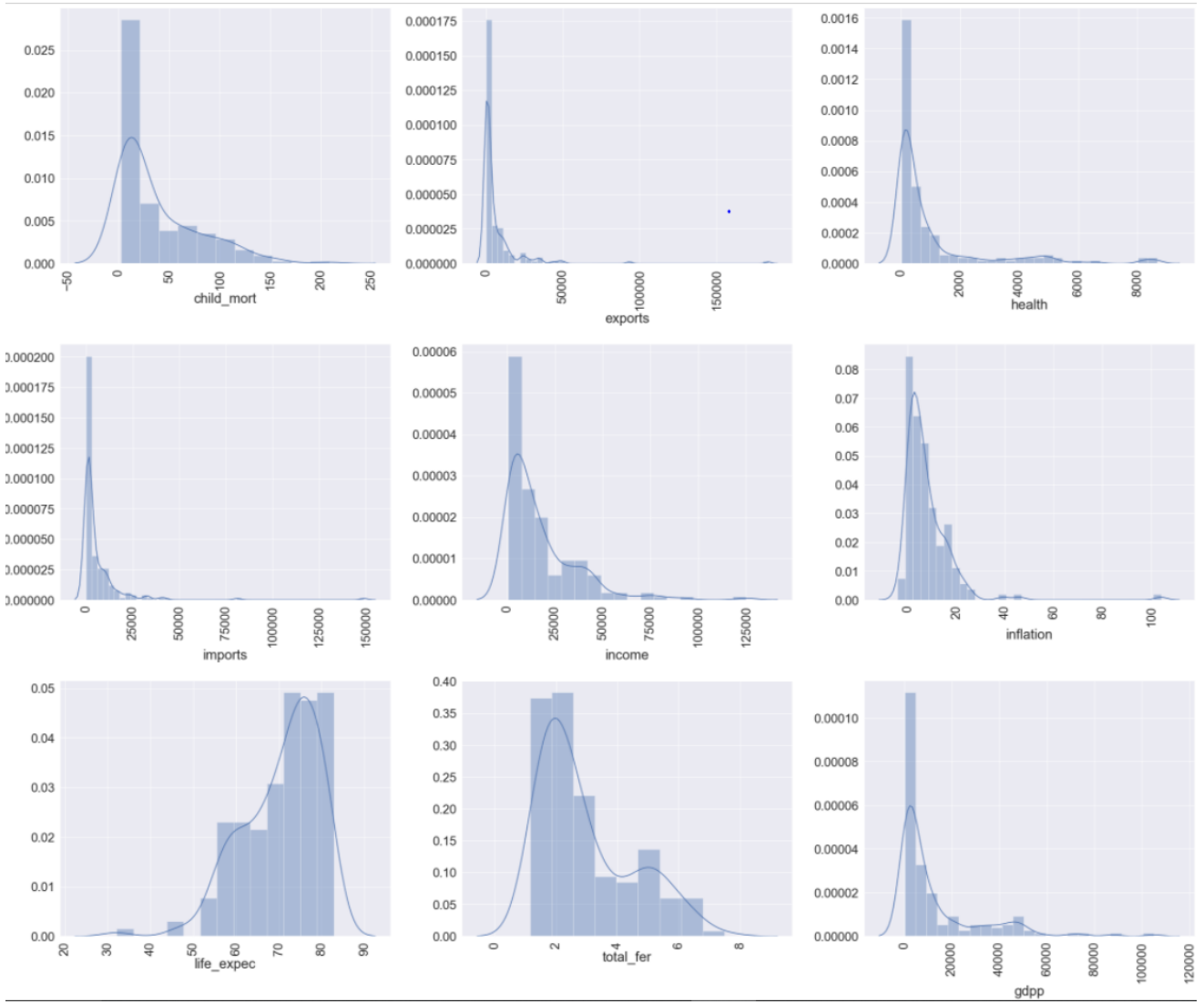
- Imports and Exports are most highly positively correlated with correlation of 0.99
- GDPP and Health are positively correlated with correlation of 0.92
- GDPP and Income are positively correlated with correlation of 0.9
- CHILD_MORTALITY and LIFE_EXPENTENCY are highly negatively correlated with correlation of -0.89



DISTPLOT

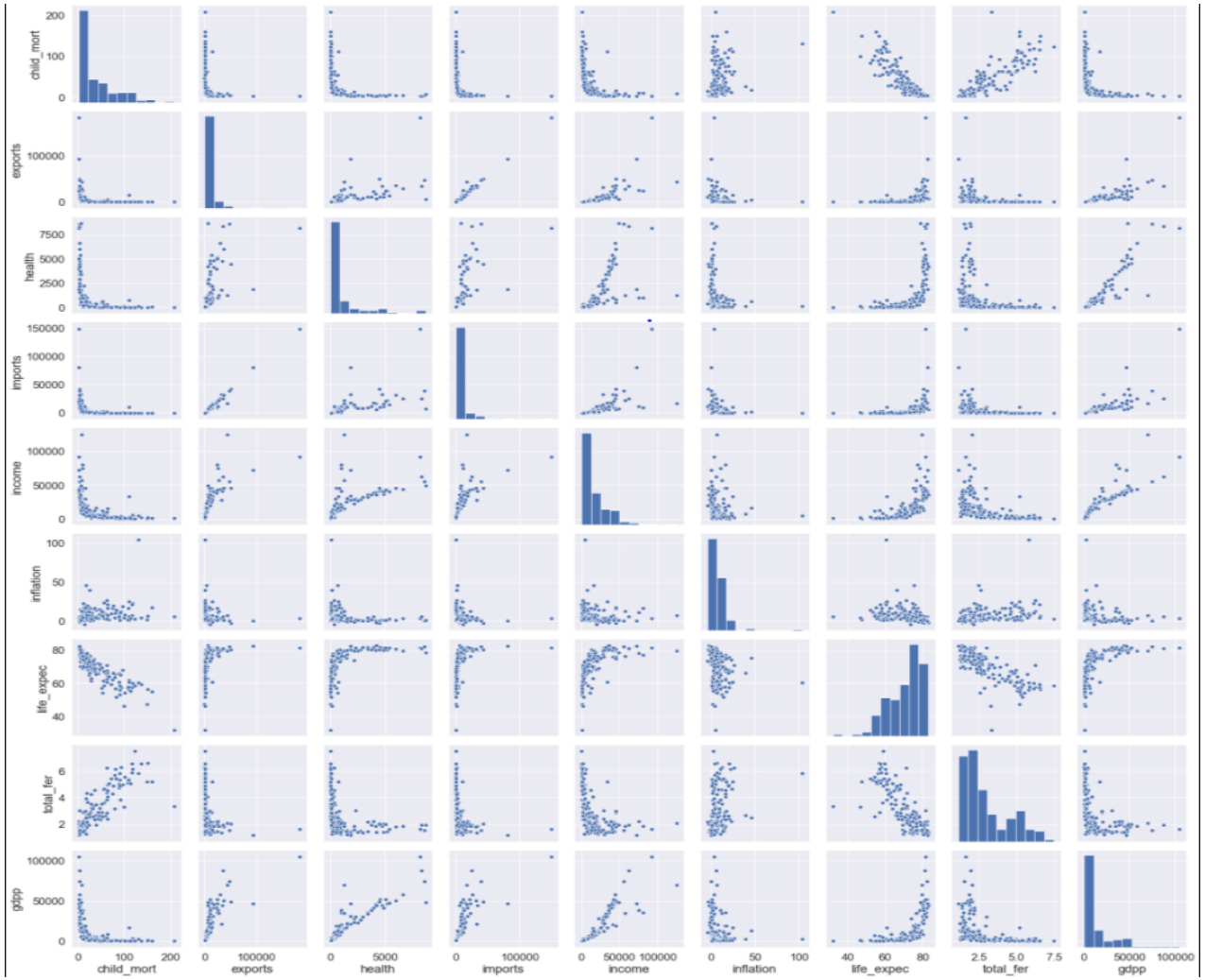
From the Distplot we can see that:

- child_mort,exports and health
  - They are normally distributed with some skewness.
  - The values belong to the specific range.
  - These columns does not have any internal groupings.
- life_expec, total_fer, income
  - They are binomially distributed
  - Showing signs of having internal groupings in the data.
- All the columns are useful to perform Clustering



PAIRPLOT

- We can see that health and income are linearlly related to gdpp
- child_mort and total_fer linearlly related
- life_expec and child_mort are inversly related.

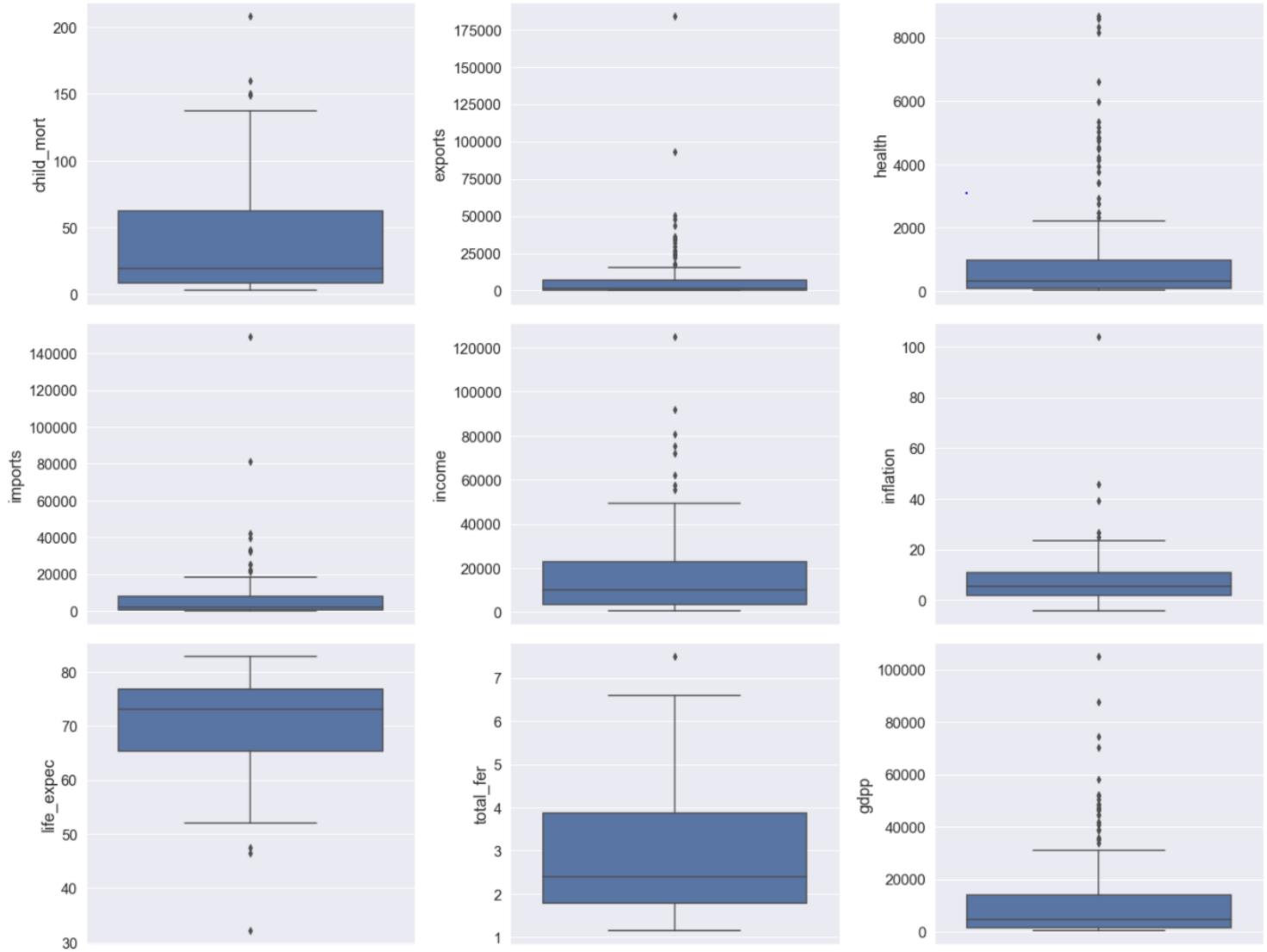


OUTLIER TREATMENT

Here we can see the boxplots of the columns before outlier treatment.

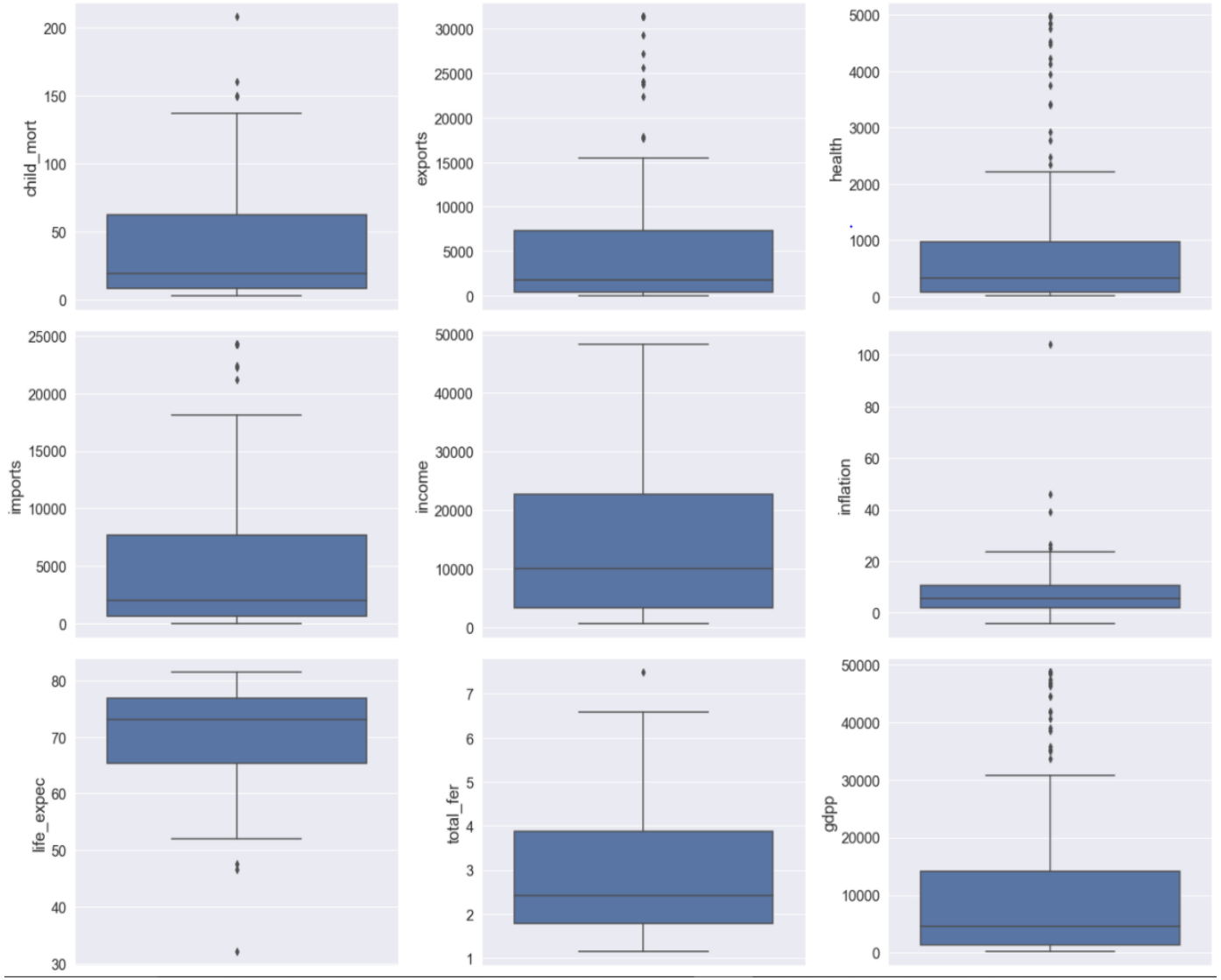
Outlier treatment has been done for the following columns:

- Exports
- Health
- Imports
- Income
- life_expec
- Gdpp



Here we can see the boxplots columns after outlier treatment.

We perform the hard capping as the treatment, as they were having outliers on the higher end of the spectrum.
These values can be capped as we need to find the countries who are in direst need of the aid.
So, accordingly we capped the capped the values at **0.95 quantile**.



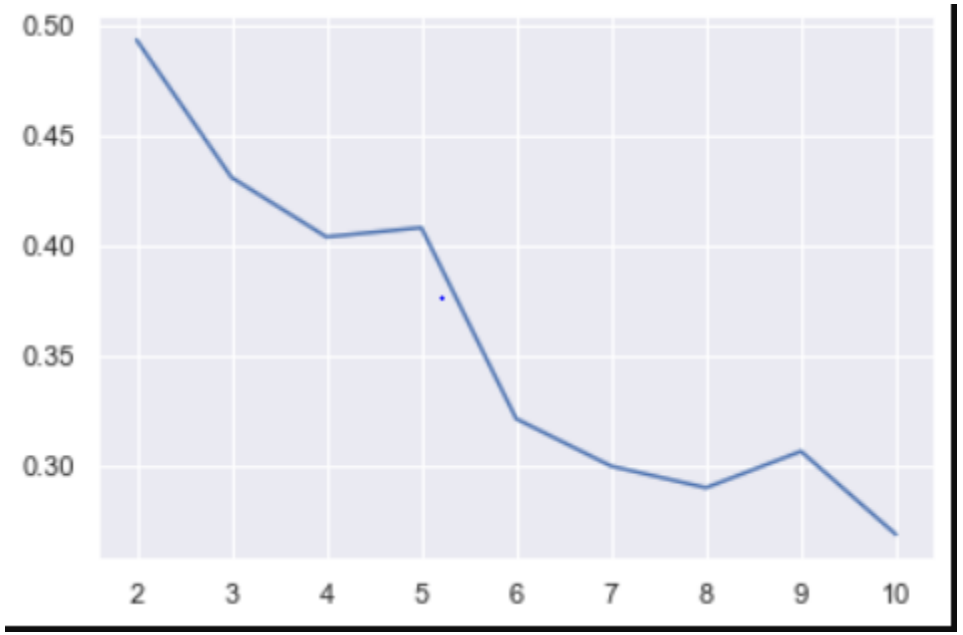
Find best val of k : Silhouette score and SSD(elbow)

SILHOUETTE SCORE

- Silhouette score for a set of sample data points is used to measure how dense and well-separated the clusters are.
- The silhouette score falls within the range [-1, 1].
- The silhouette score of 1 means that the clusters are very dense and nicely separated.
- The score of 0 means that clusters are overlapping.

Here, I have calculated the Silhouette score for a range of 2 to 10 clusters.

	0	1
0	2	0.493390
1	3	0.430945
2	4	0.403899
3	5	0.408123
4	6	0.321320
5	7	0.299814
6	8	0.289984
7	9	0.306547
8	10	0.268786



ELBOW CURVE

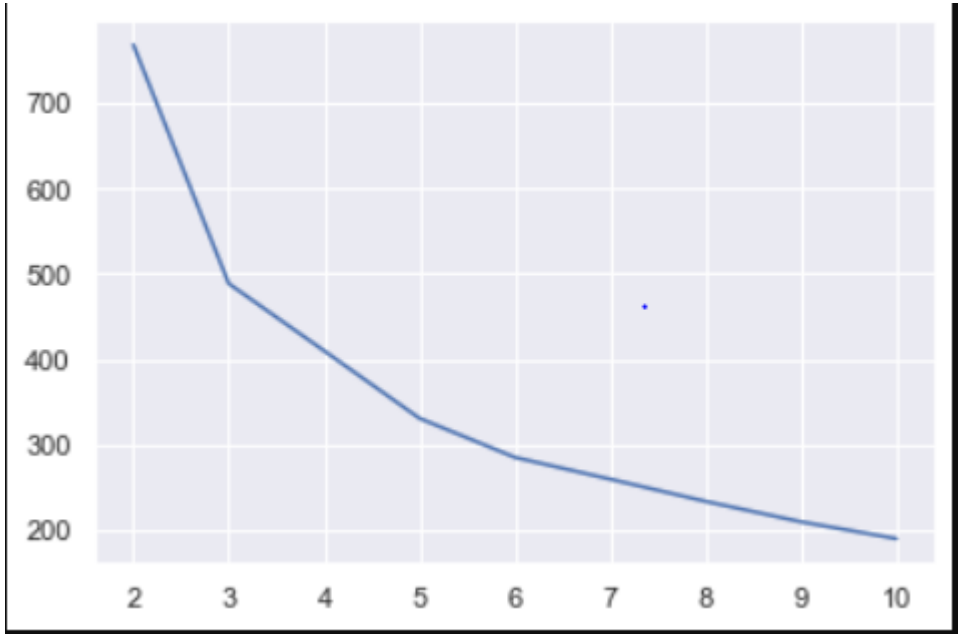
The Elbow method is a very popular technique and the idea is to run k-means clustering for a range of clusters *k* (let's say from 1 to 10) .

For each value, we are calculating the sum of squared distances from each point to its assigned centre(distortions).

When the distortions are plotted and the plot looks like an arm then the "elbow"(the point of inflection on the curve) is the best value of *k*.

Here, I have plotted the elbow curve for and seeing the plot, the optimum value of k should be 3

	0	1
0	2	768.031545
1	3	488.363553
2	4	409.655221
3	5	330.557874
4	6	284.790698
5	7	259.260592
6	8	233.231394
7	9	209.313138
8	10	189.618493



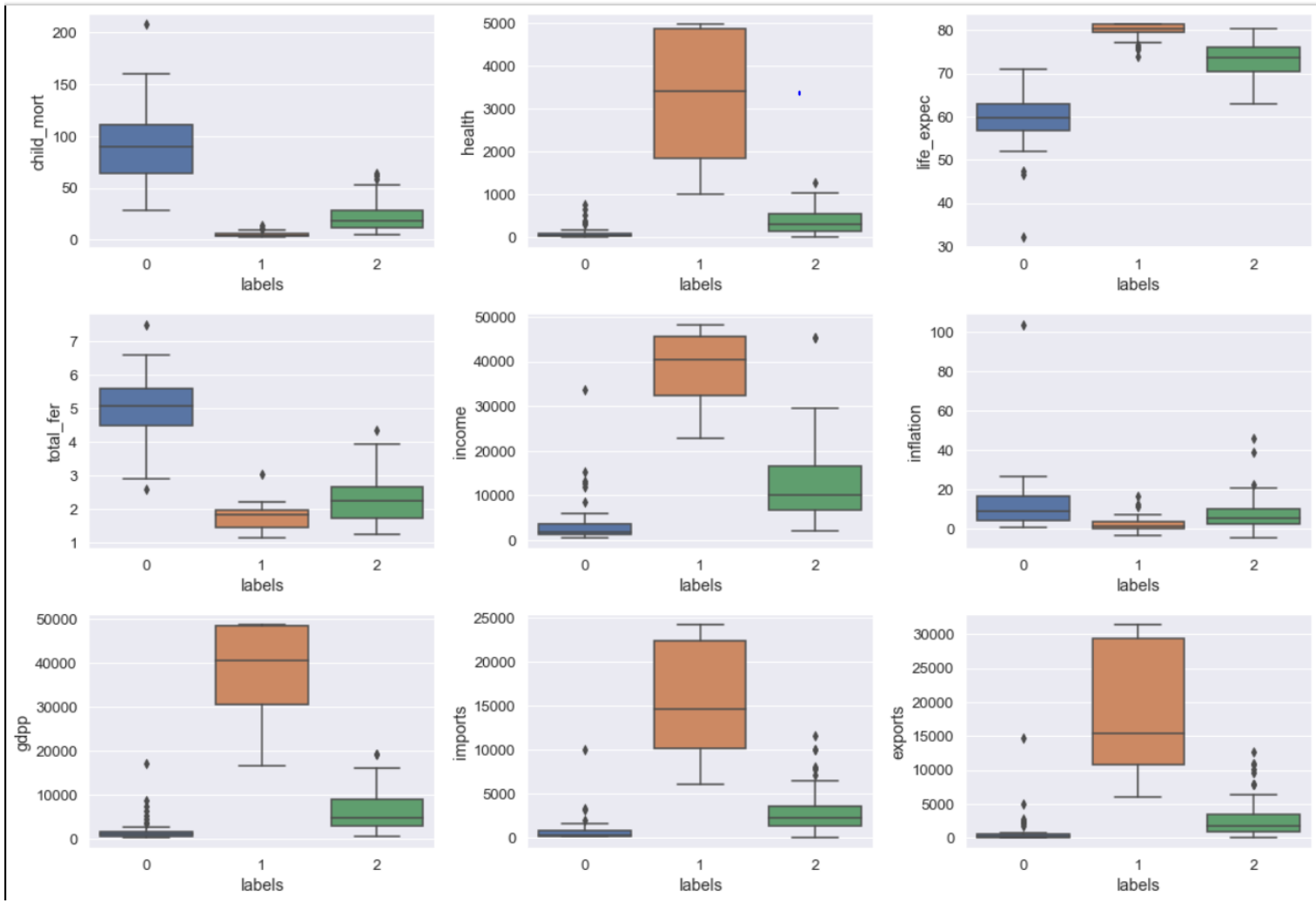
Visualization

WITH BOXPLOT

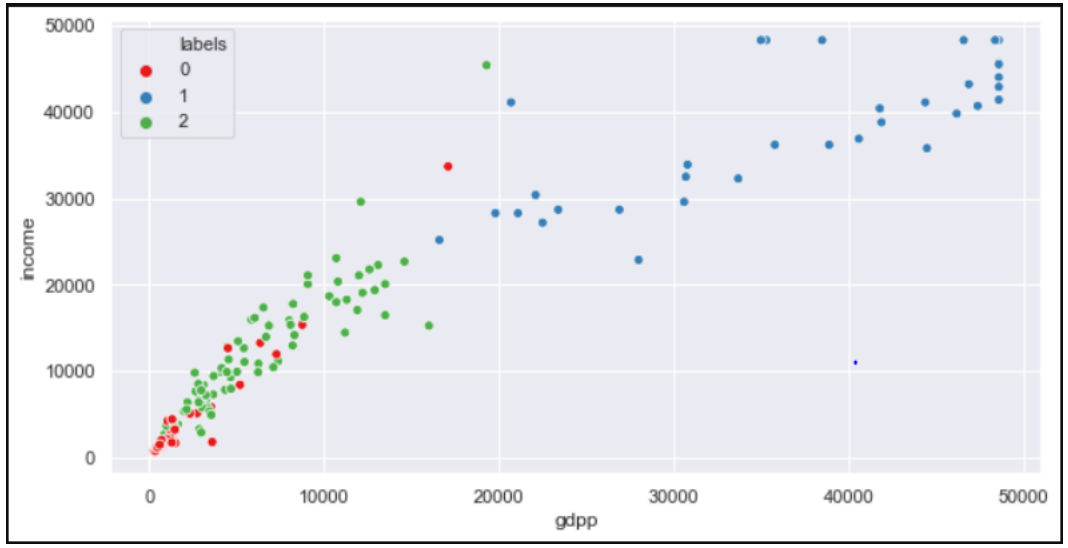
We can straightaway see that Cluster 1 countries have:

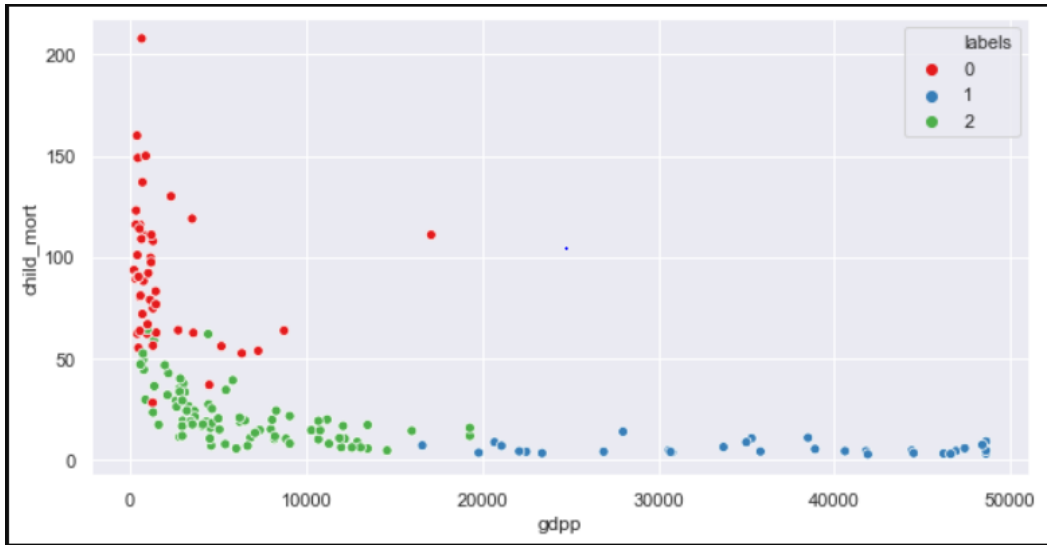
a. high child mortality rate.
b. low health.
c. low life expectancy.
d. low income rates.
e. high inflation
f. low GDPP
g. lowest imports and exports

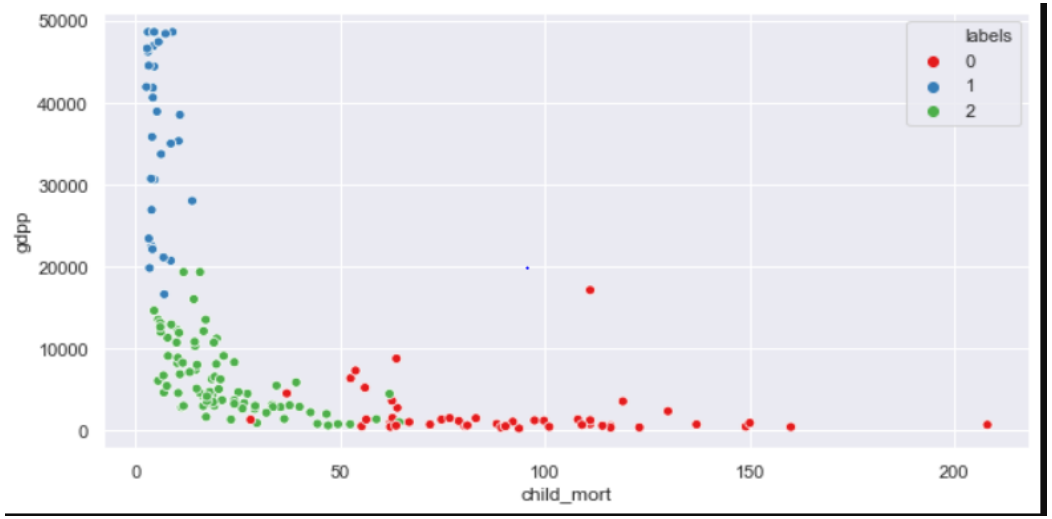
Hence, we can say that **CLUSTER 0** countries are in dire need of development as they have high child mortality rate, lowest GDPP and low income.



WITH SCATTER PLOT





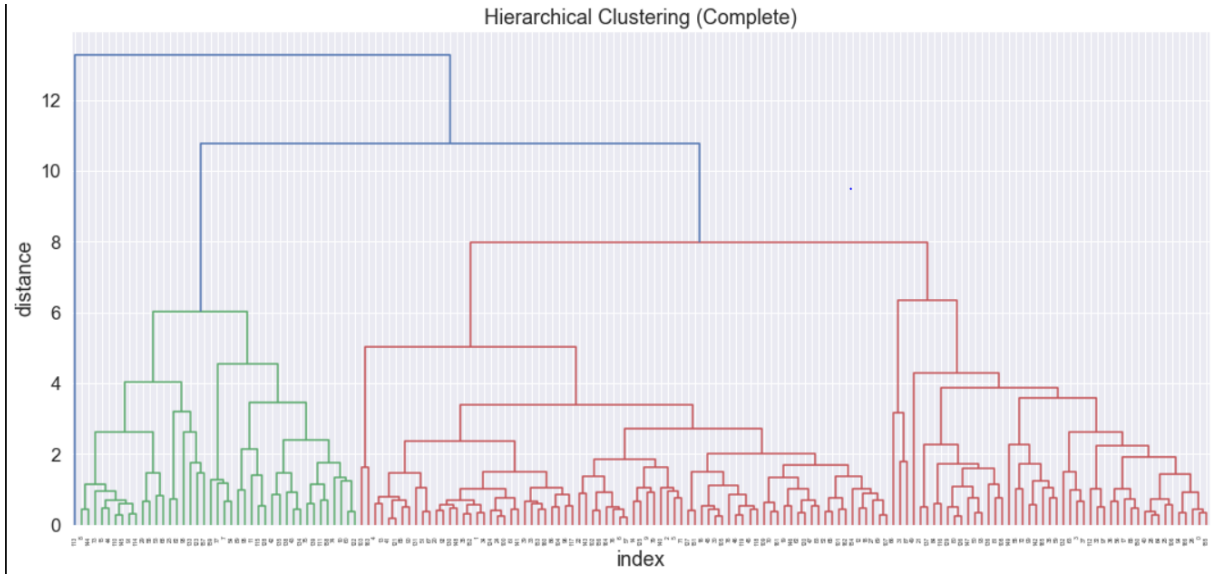


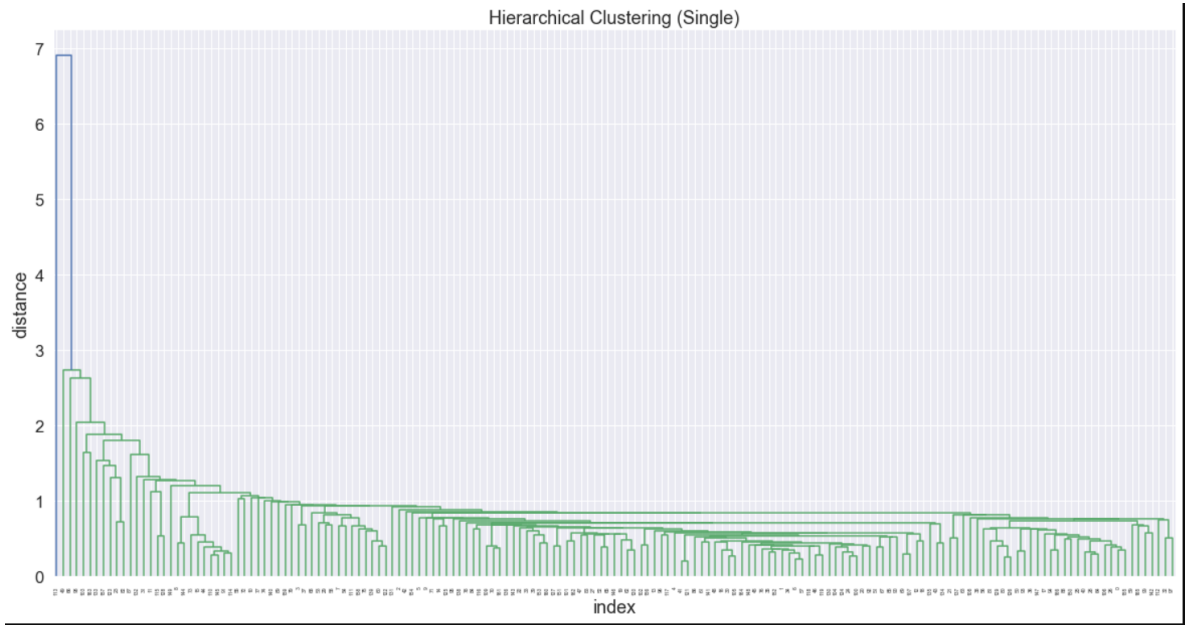
After applying kmeans algorithm with number of clusters=3, we get these graphs showing 3 distinct clusters

HIERARCHICAL CLUSTERING

Hierarchical Clustering can be done using two ways :

- Single linkage: It is defined as the shortest distance between any 2 points in the clusters.
- Complete linkage: It is defined as the maximum distance between any 2 points in the clusters



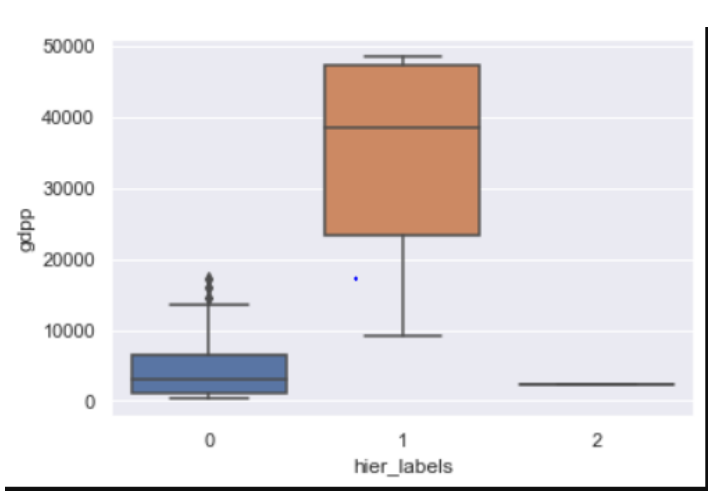


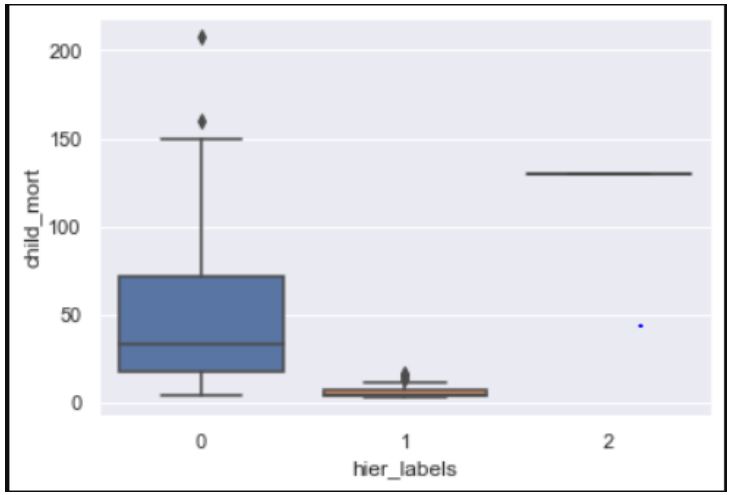
From the 2 dendrograms we can see that the complete linkage has a better cluster formation.

Hence, going ahead we will use complete linkage output for our analysis and we will use number of clusters=3

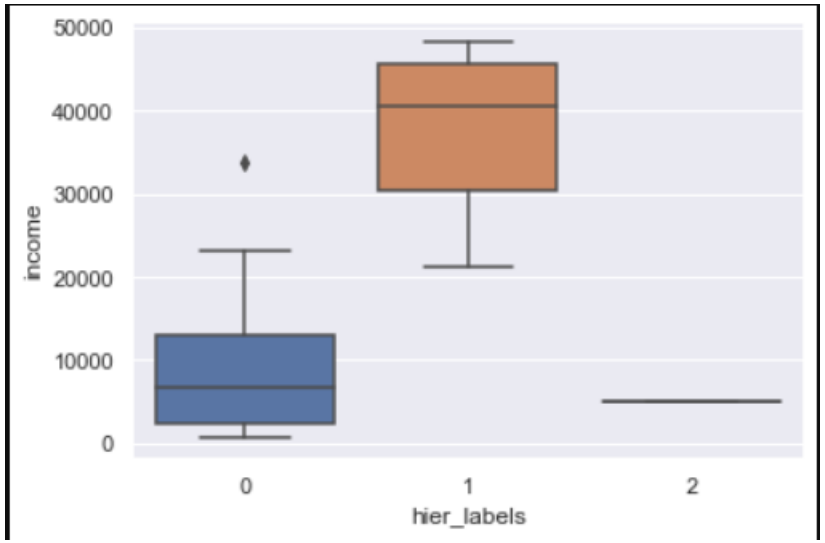
Visualization

WITH BOXPLOT





From the above boxplots we can see that the **cluster 0** has lowest income, lowest GDPP and highest child mortality rate.



CLUSTER PROFILING : GDPP, CHILD_MORT, INCOME

After performing cluster profiling on both datasets(one from Kmeans clustering and one from Hierarchical clustering) by checking how the columns **GDPP**, **CHILD_MORT** and **INCOME** vary for each cluster, we have reached to the conclusion that the countries in direst need of aid are:

- Burundi
- Liberia
- Congo, Dem. Rep.
- Niger
- Sierra Leone

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	labels	hier_labels
26	Burundi	93.6	20.6052	26.7960	90.552	764.0	12.30	57.7	6.26	231	0	0
88	Liberia	89.3	62.4570	38.5860	302.802	700.0	5.47	60.8	5.02	327	0	0
37	Congo, Dem. Rep.	116.0	137.2740	26.4194	165.664	609.0	20.80	57.5	6.54	334	0	0
112	Niger	123.0	77.2560	17.9568	170.868	814.0	2.55	58.8	7.49	348	0	0
132	Sierra Leone	160.0	67.0320	52.2690	137.655	1220.0	17.20	55.0	5.20	399	0	0
93	Madagascar	62.2	103.2500	15.5701	177.590	1390.0	8.79	60.8	4.60	413	0	0
106	Mozambique	101.0	131.9850	21.8299	193.578	918.0	7.64	54.5	5.56	419	0	0
31	Central African Republic	149.0	52.6280	17.7508	118.190	888.0	2.01	47.5	5.21	446	0	0
94	Malawi	90.5	104.6520	30.2481	160.191	1030.0	12.10	53.1	5.31	459	0	0
50	Eritrea	55.2	23.0878	12.8212	112.306	1420.0	11.60	61.7	4.61	482	0	0

Thank You