CLUSTERING OF COUNTRIES

Neha Shukla

BUSINESS UNDERSTANDING

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.

And this is where you come in as a data analyst. Your job is 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.

Data Understanding

The country data consists of following columns:

Column Name	Description
country	Name of the country
child_mort	Death of children under 5 years of age per 1000 live births
exports	Exports of goods and services per capita. Given as %age of the GDP per capita
health	Total health spending per capita. Given as %age of GDP per capita
imports	Imports of goods and services per capita. Given as %age of the GDP per capita
Income	Net income per person
Inflation	The measurement of the annual growth rate of the Total GDP
life_expec	The average number of years a new born child would live if the current mortality patterns are to remain the same
total_fer	The number of children that would be born to each woman if the current age-fertility rates remain the same.
gdpp	The GDP per capita. Calculated as the Total GDP divided by the total population.

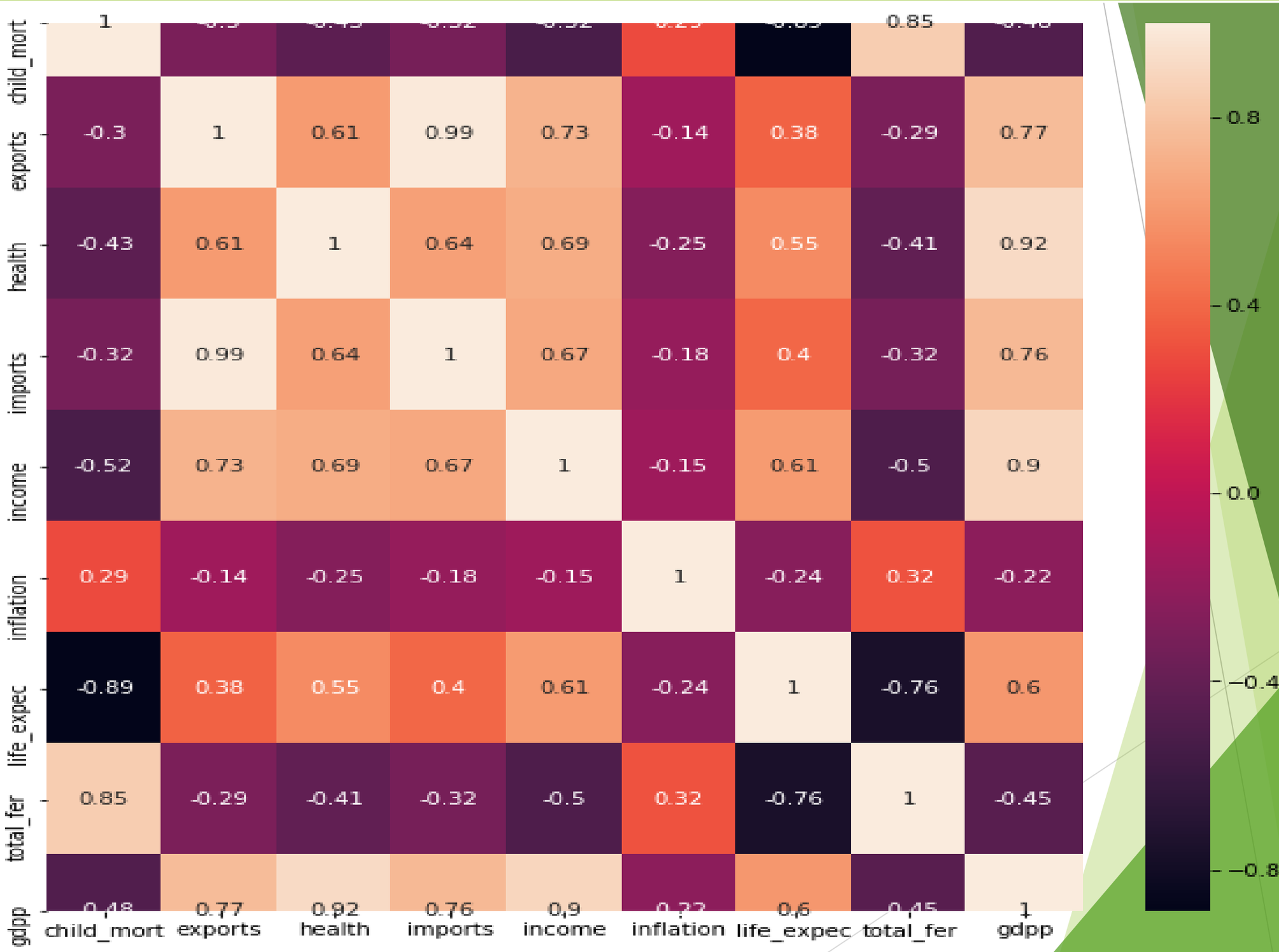
Data Cleaning

Data set neither has any missing values nor any inconsistent datatype.

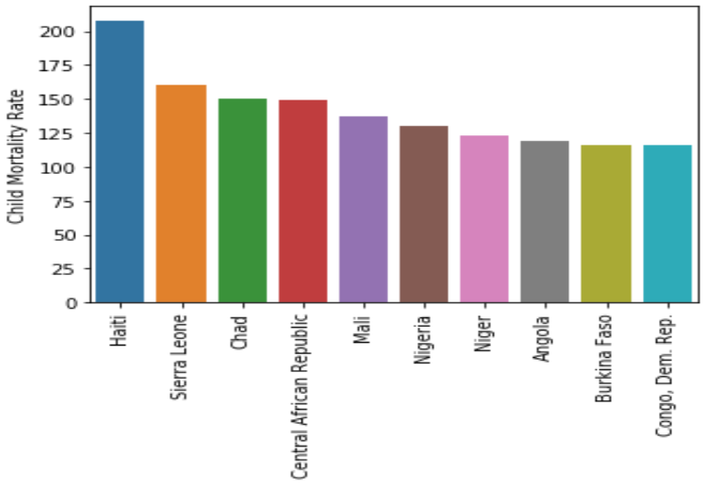
There is no duplicate values provided in dataset.

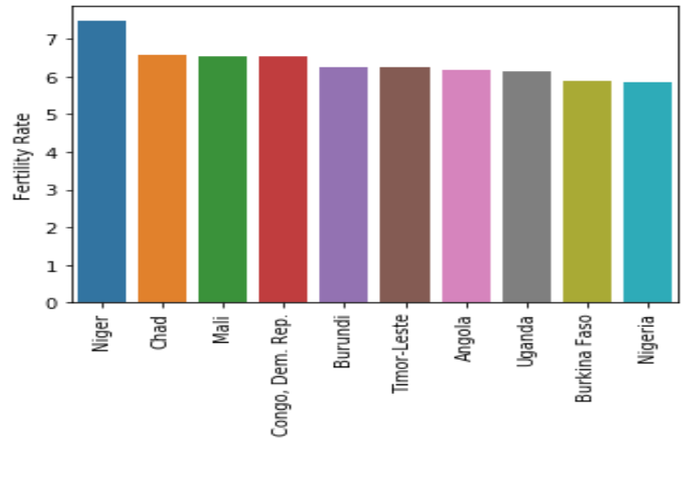
We have converted imports, exports and health spending from percentage values to actual values of their GDP per capita. Because the percentage value doesn't give a clear picture of that country.

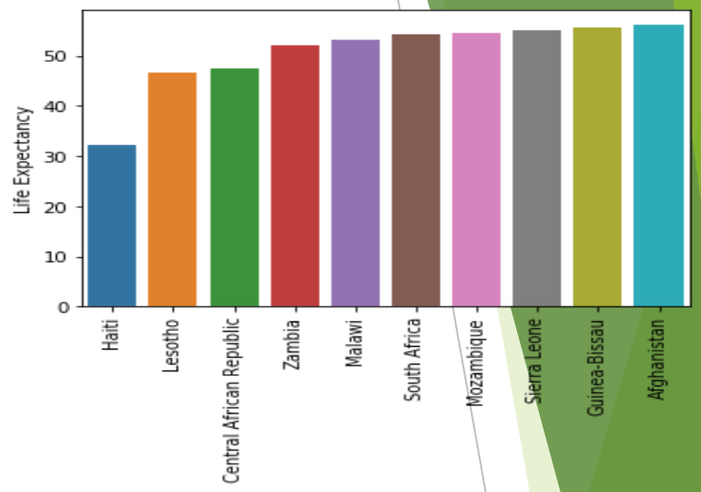
Correlation between variables in dataset is shown by below Heatmap.

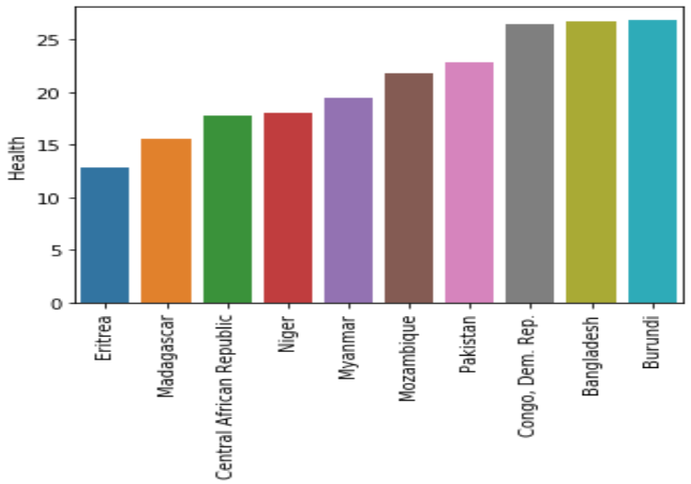


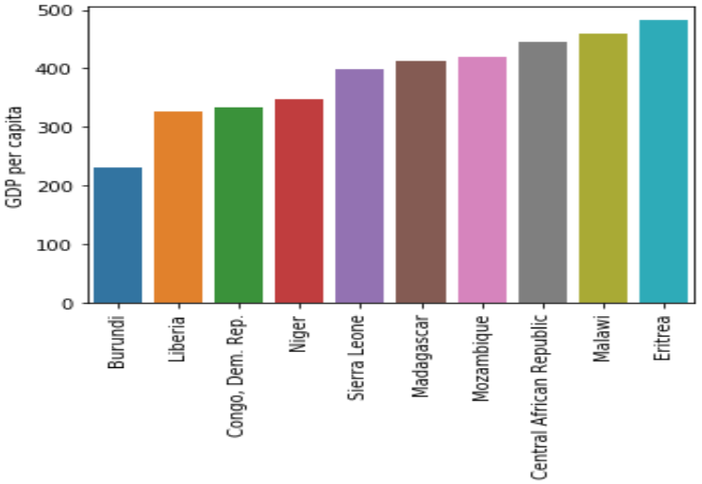
Exploratory Data Analysis

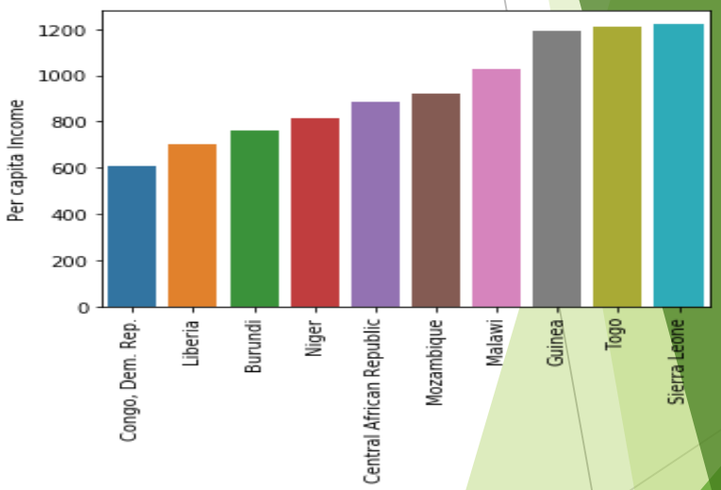


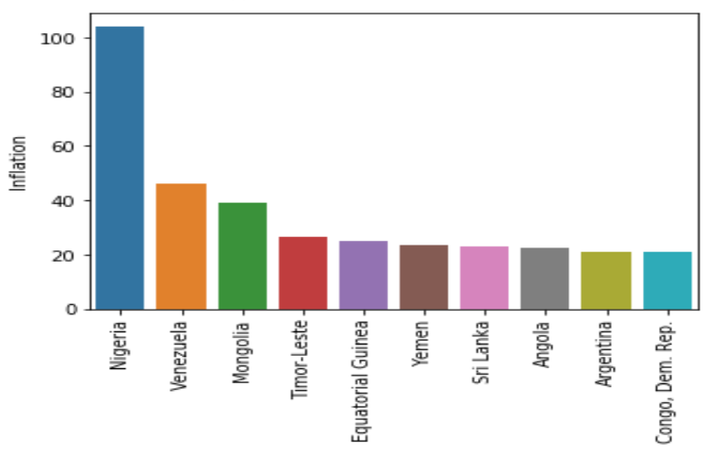


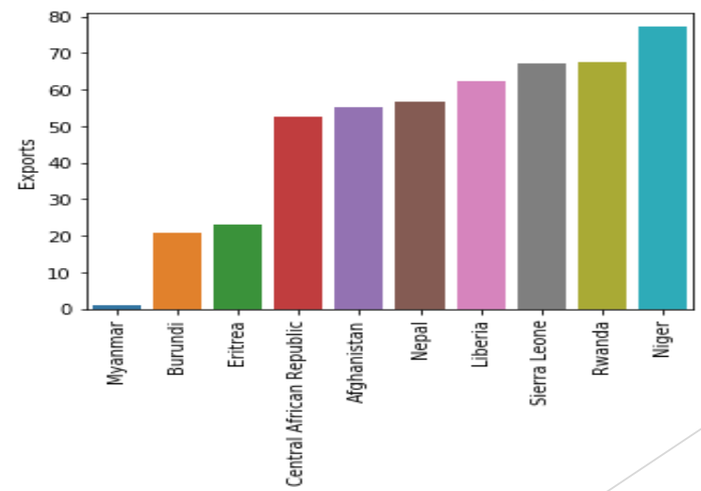


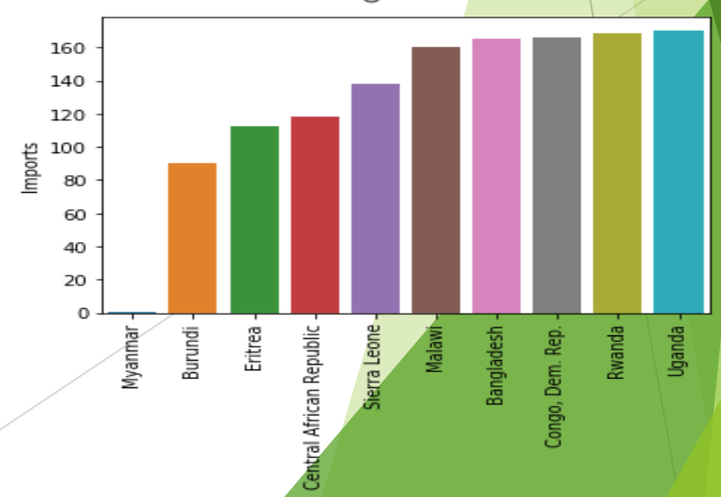




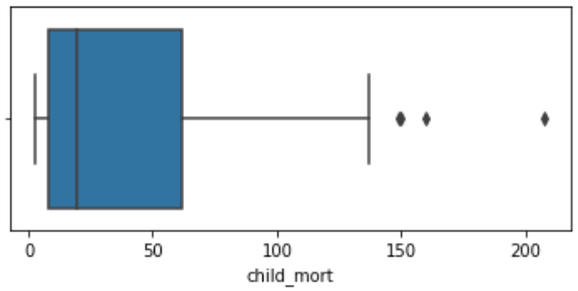


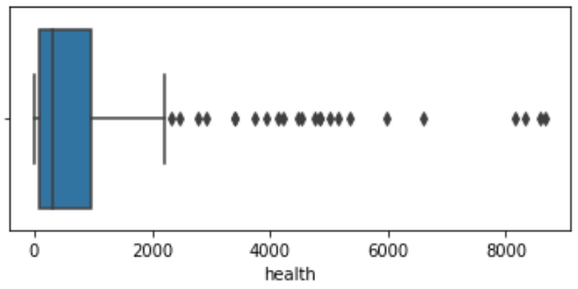


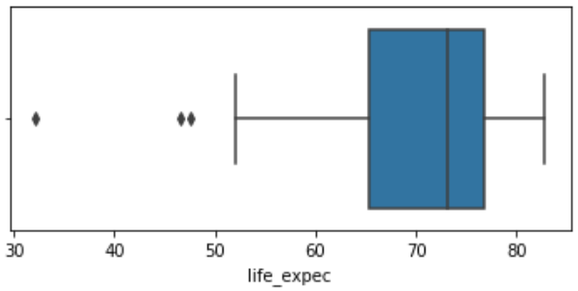


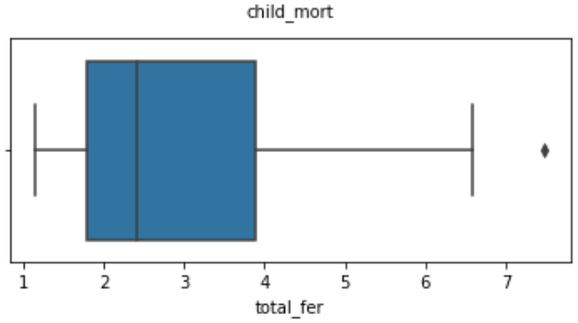


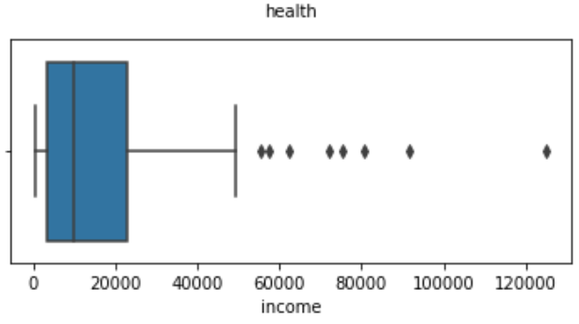
OUTLIER ANALYSIS

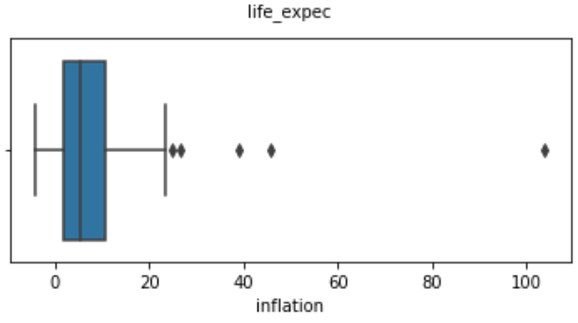


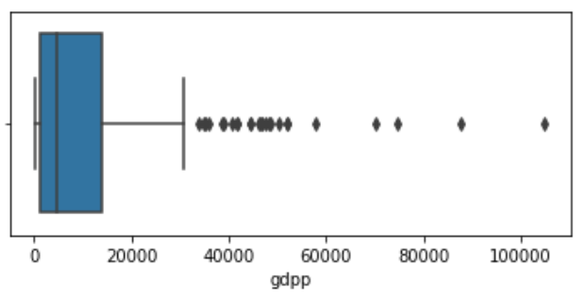


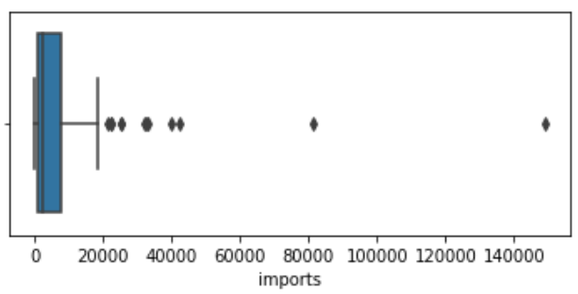


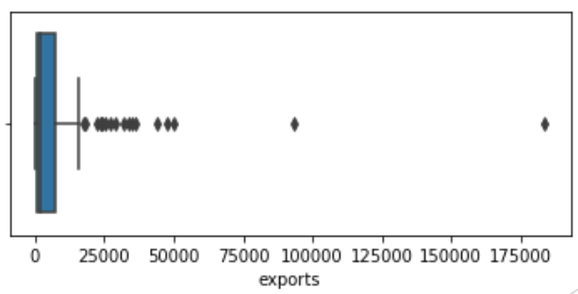






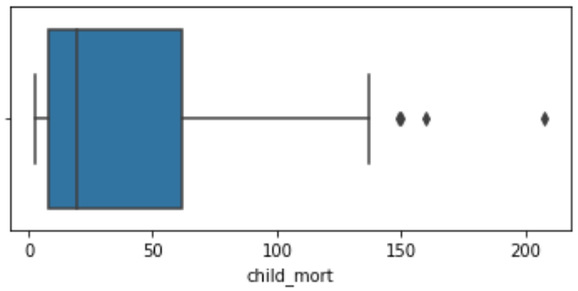


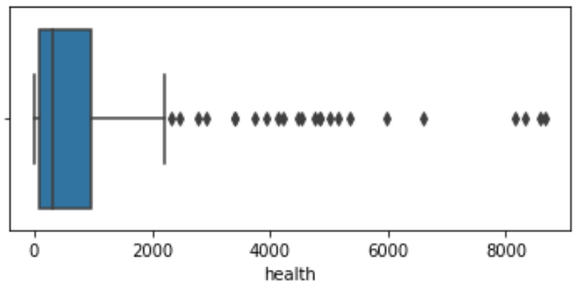


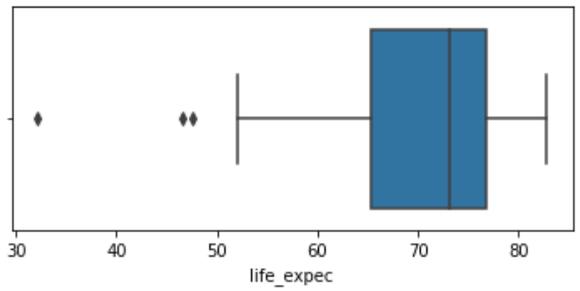


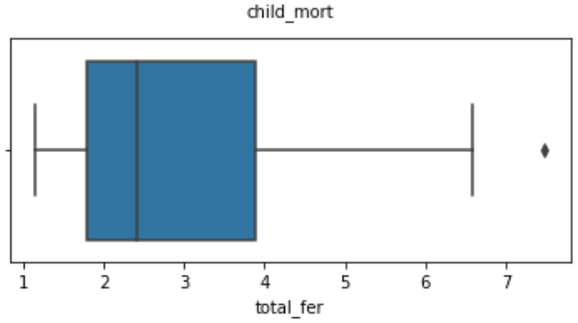
OUTLIER TREATMENT

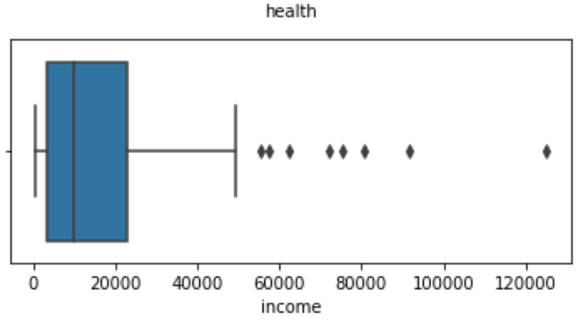
As we can see there are a number of outliers in the data. We need to identify backward countries based on socio economic and health factors.
We will cap the outliers to values accordingly for analysis. The capping boundary used is 0.01 and 0.99

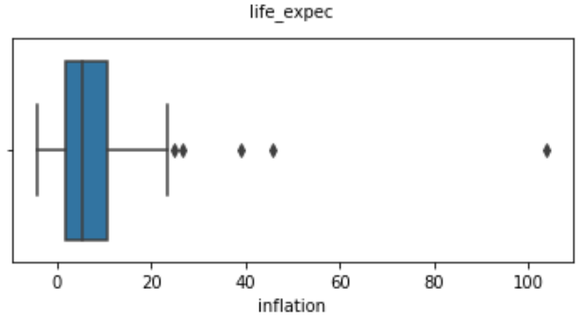


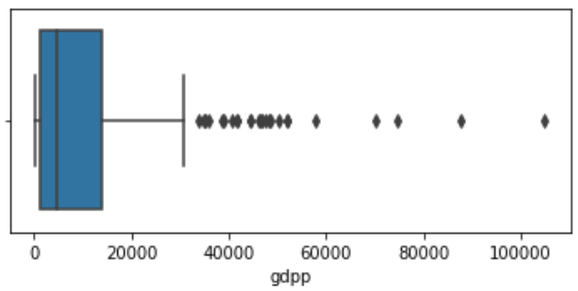


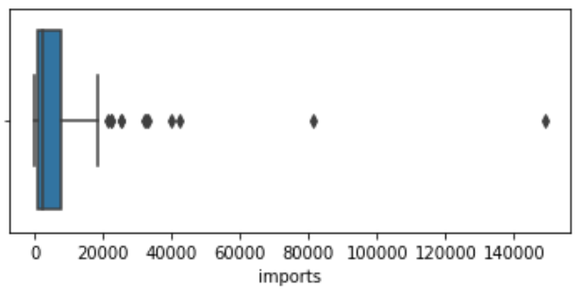


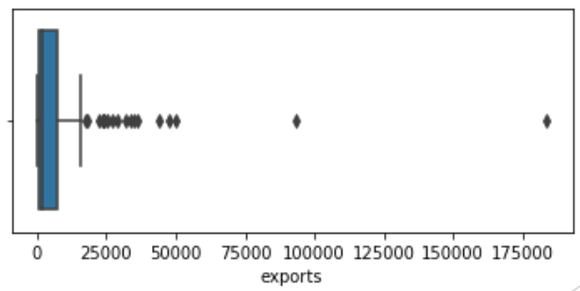












HOPKIN'S CHECK

The Hopkins statistic, is a statistic which gives a value which indicates the cluster tendency, in other words: how well the data can be clustered.

- If the value is between {0.01, ...,0.3}, the data is regularly spaced.
- If the value is around 0.5, it is random.
- If the value is between {0.7, ..., 0.99}, it has a high tendency to cluster.

Here, the Hopkin's value came out to be 0.95 which means that dataset has a tendency to cluster

SCALING OF THE DATA

Scaling is a step of Data Pre Processing which is applied to independent variables or features of data. It basically helps to normalize the data within a particular range. Sometimes, it also helps in speeding up the calculations in an algorithm. It becomes an essential step before clustering as Euclidean distance is very sensitive to the changes in the differences

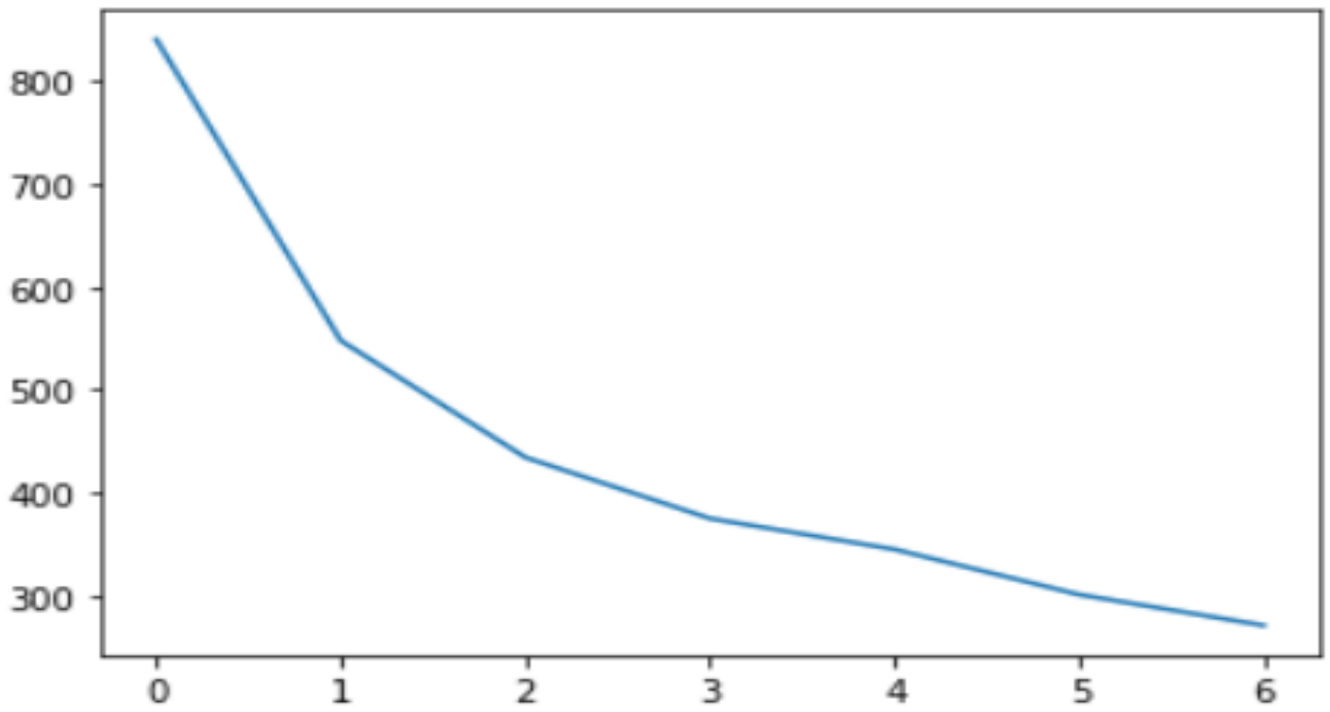
Data head after performing the scaling:

	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	1.344012	-0.569638	-0.566983	-0.598844	-0.851772	0.263649	-1.693799	1.926928	-0.702314
1	-0.547543	-0.473873	-0.440417	-0.413679	-0.387025	-0.375251	0.663053	-0.865911	-0.498775
2	-0.272548	-0.424015	-0.486295	-0.476198	-0.221124	1.123260	0.686504	-0.035427	-0.477483
3	2.084186	-0.381264	-0.534113	-0.464070	-0.612136	1.936405	-1.236499	2.154642	-0.531000
4	-0.709457	-0.086754	-0.178431	0.139659	0.125202	-0.768917	0.721681	-0.544433	-0.032079

Finding the Optimal Number of Clusters

Method 1:**Elbow-Curve/SSD**

A fundamental step for any unsupervised algorithm is to determine the optimal number of clusters into which the data may be clustered. The Elbow Method is one of the most popular methods to determine this optimal value of k.

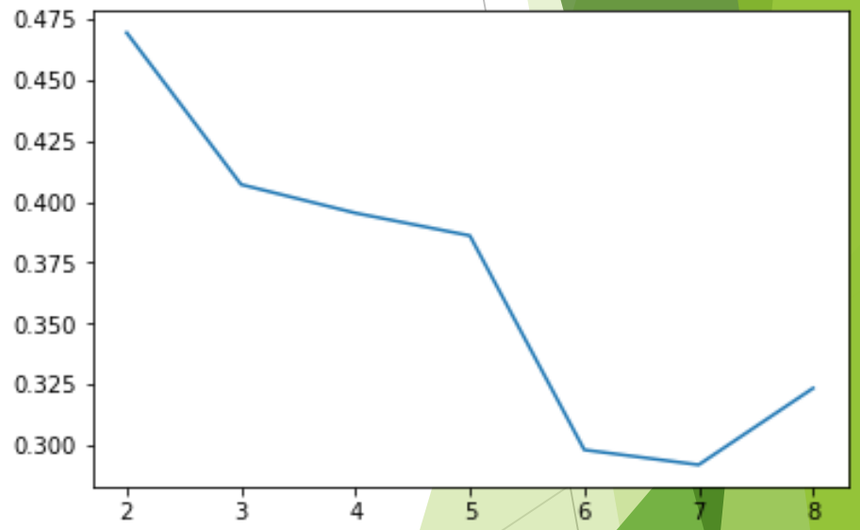


Unfortunately, we do not always have such clearly clustered data. So we use Silhouette method to determine number of clusters.

Method 2: **Silhouette Analysis**

- The value of the silhouette score range lies between -1 to 1.
- A score closer to 1 indicates that the data point is very similar to other data points in the cluster,
- A score closer to -1 indicates that the data point is not similar to the data points in its cluster.

```
For n_clusters=2, the silhouette score is 0.469399802877881 13
For n_clusters=3, the silhouette score is 0.40708993455880504
For n_clusters=4, the silhouette score is 0.3953914230955 1445
For n_clusters=5, the silhouette score is 0.3864288935632213
For n_clusters=6, the silhouette score is 0.30280954105276137
For n_clusters=7, the silhouette score is 0.3142190342453547
For n_clusters=8, the silhouette score is 0.2858078357857632
```



The silhouette score reaches a peak at around 3 clusters indicating that it might be the ideal number of clusters. (k=3)

K Means clustering

Using the previous methods of finding the number of clusters as 3, K-Means clustering is performed. There is a label assigned to each cluster id column. The data head is shown below.

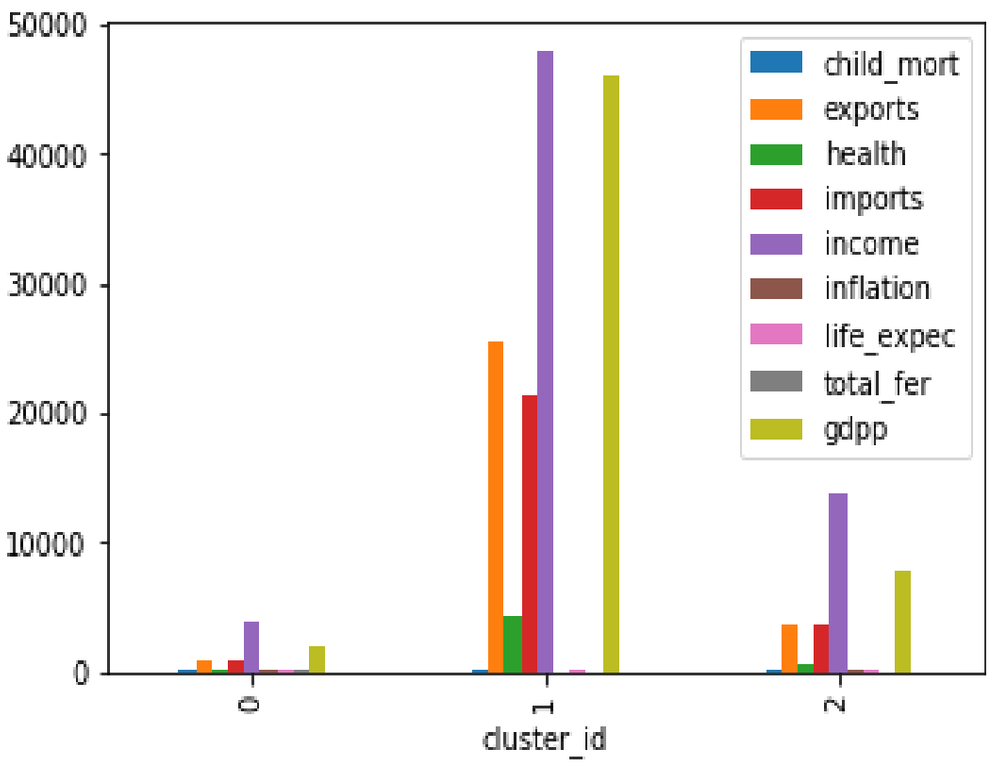
	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_id
0	Afghanistan	90.2	55.30	41.9174	248.297	1610.0	9.44	56.2	5.82	553.0	0
1	Albania	16.6	1145.20	267.8950	1987.740	9930.0	4.49	76.3	1.65	4090.0	2
2	Algeria	27.3	1712.64	185.9820	1400.440	12900.0	16.10	76.5	2.89	4460.0	2
3	Angola	119.0	2199.19	100.6050	1514.370	5900.0	22.40	60.1	6.16	3530.0	0
4	Antigua and Barbuda	10.3	5551.00	735.6600	7185.800	19100.0	1.44	76.8	2.13	12200.0	2

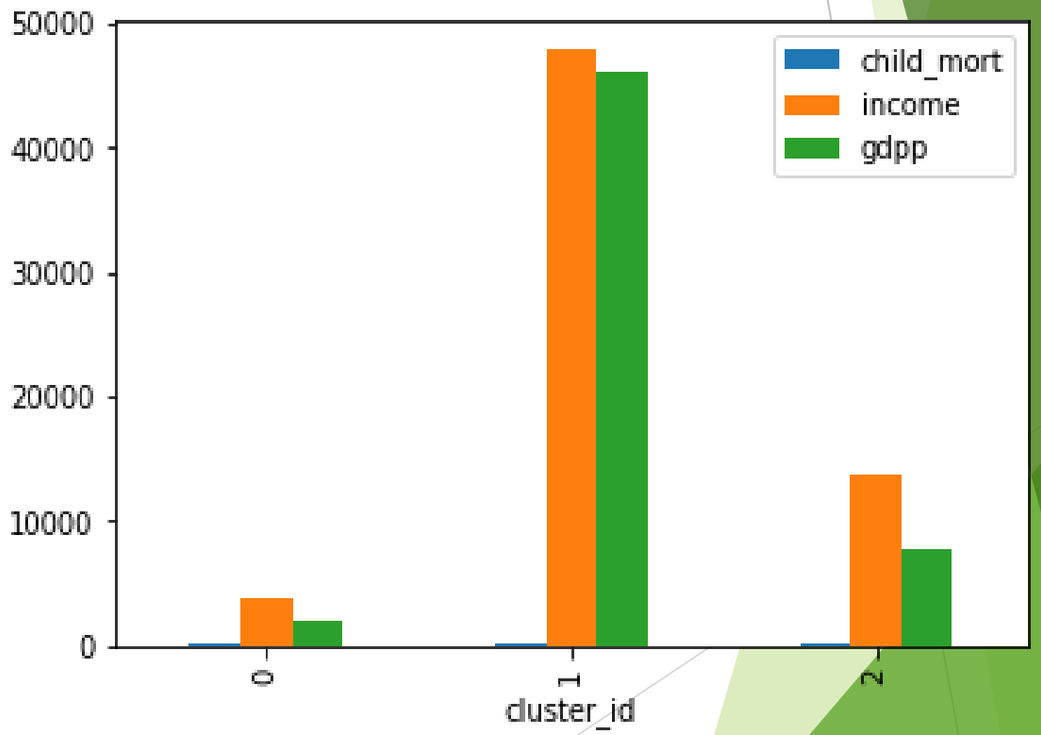
The value count of each cluster is shown below.

Cluster ID	Value Counts
1	90
0	48
2	29

CLUSTER PROFILING

cluster_id	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	90.335417	879.097657	114.939003	827.327888	3901.010000	10.608604	59.567083	4.972233	1911.400833
1	4.989655	25405.359310	4239.330028	21316.695862	47784.413793	2.906731	80.453103	1.757352	46068.137931
2	20.547778	3477.250726	528.894338	3589.291996	13804.333333	7.131624	73.393333	2.242591	7808.577778





Child Mortality is highest for Cluster 0, These clusters need some aid.
Income and Gdpp are measures of development. Higher the per capita income and gdpp better is the country's development. Income per capita and gdpp seems lowest for countries in clusters 0. Hence, these countries need some help.

Conclusion:

We observe that **Child mortality, Income, Inflation and GDP per capita** are good predictors for the development of a country. We have found that countries in Cluster 0 need the most aid as the Child mortality rate is highest, income and GDP per capita are the lowest among all the clusters.

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_id
88	Liberia	89.3	62.457000	38.5860	302.80200	742.24	5.47	60.8	5.0200	331.62	0
37	Congo, Dem. Rep.	116.0	137.274000	26.4194	165.66400	742.24	20.80	57.5	6.5400	334.00	0
26	Burundi	93.6	22.243716	26.7960	104.90964	764.00	12.30	57.7	6.2600	331.62	0
112	Niger	123.0	77.256000	17.9568	170.86800	814.00	2.55	58.8	6.5636	348.00	0
31	Central African Republic	149.0	52.628000	17.7508	118.19000	888.00	2.01	47.5	5.2100	446.00	0
106	Mozambique	101.0	131.985000	21.8299	193.57800	918.00	7.64	54.5	5.5600	419.00	0
94	Malawi	90.5	104.652000	30.2481	160.19100	1030.00	12.10	53.1	5.3100	459.00	0
63	Guinea	109.0	196.344000	31.9464	279.93600	1190.00	16.10	58.0	5.3400	648.00	0
150	Togo	90.3	196.176000	37.3320	279.62400	1210.00	1.18	58.7	4.8700	488.00	0
132	Sierra Leone	153.4	67.032000	52.2690	137.65500	1220.00	17.20	55.0	5.2000	399.00	0

Hierarchical Clustering

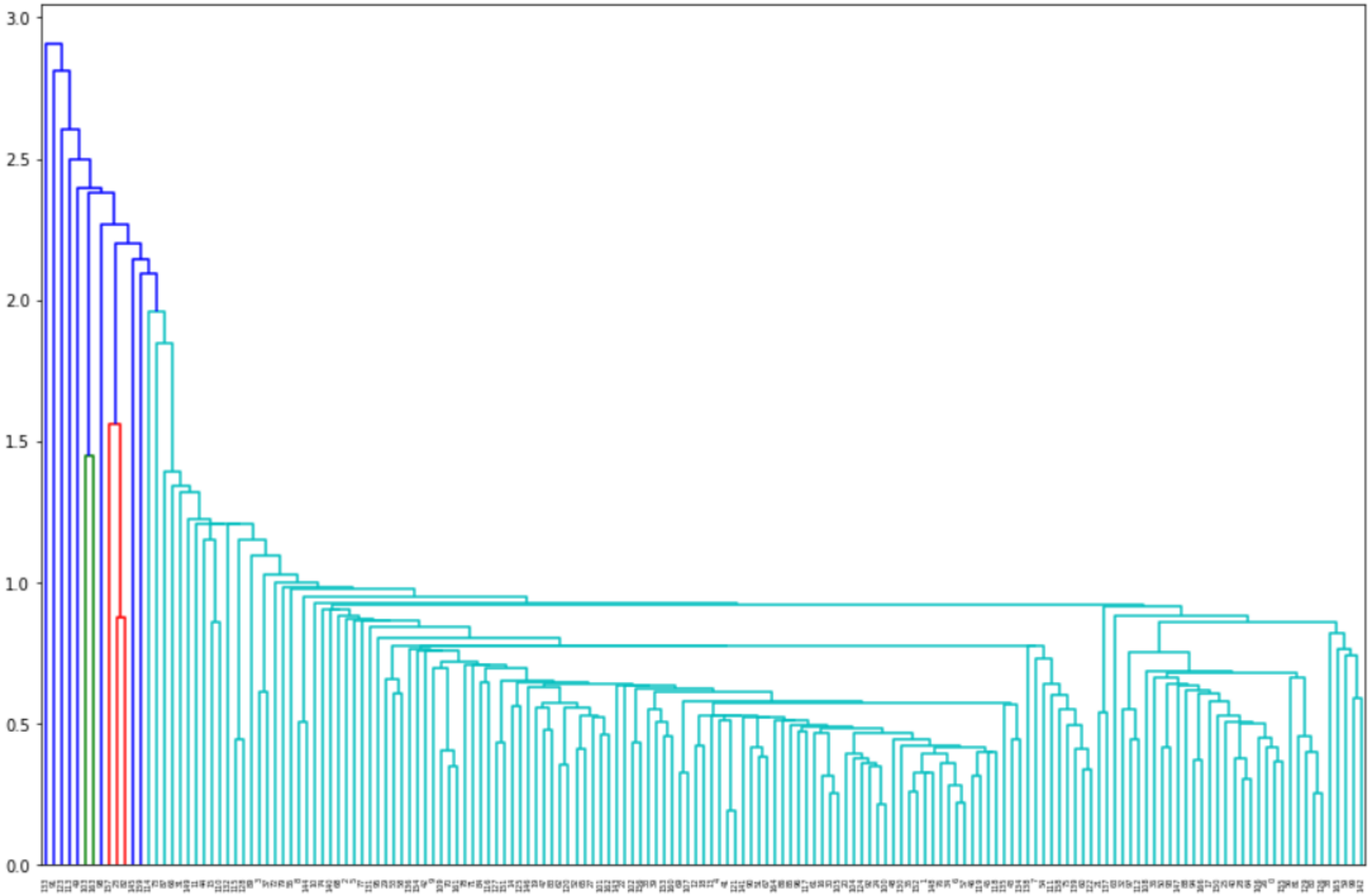
Hierarchical Clustering has one advantage over K-means Clustering which is that we don't have to select the initial number of clusters before performing clustering.

It has a different concept of linkage through which it performs the clustering operations. There are two types of Linkage:

1. Single Linkage
2. Complete Linkage

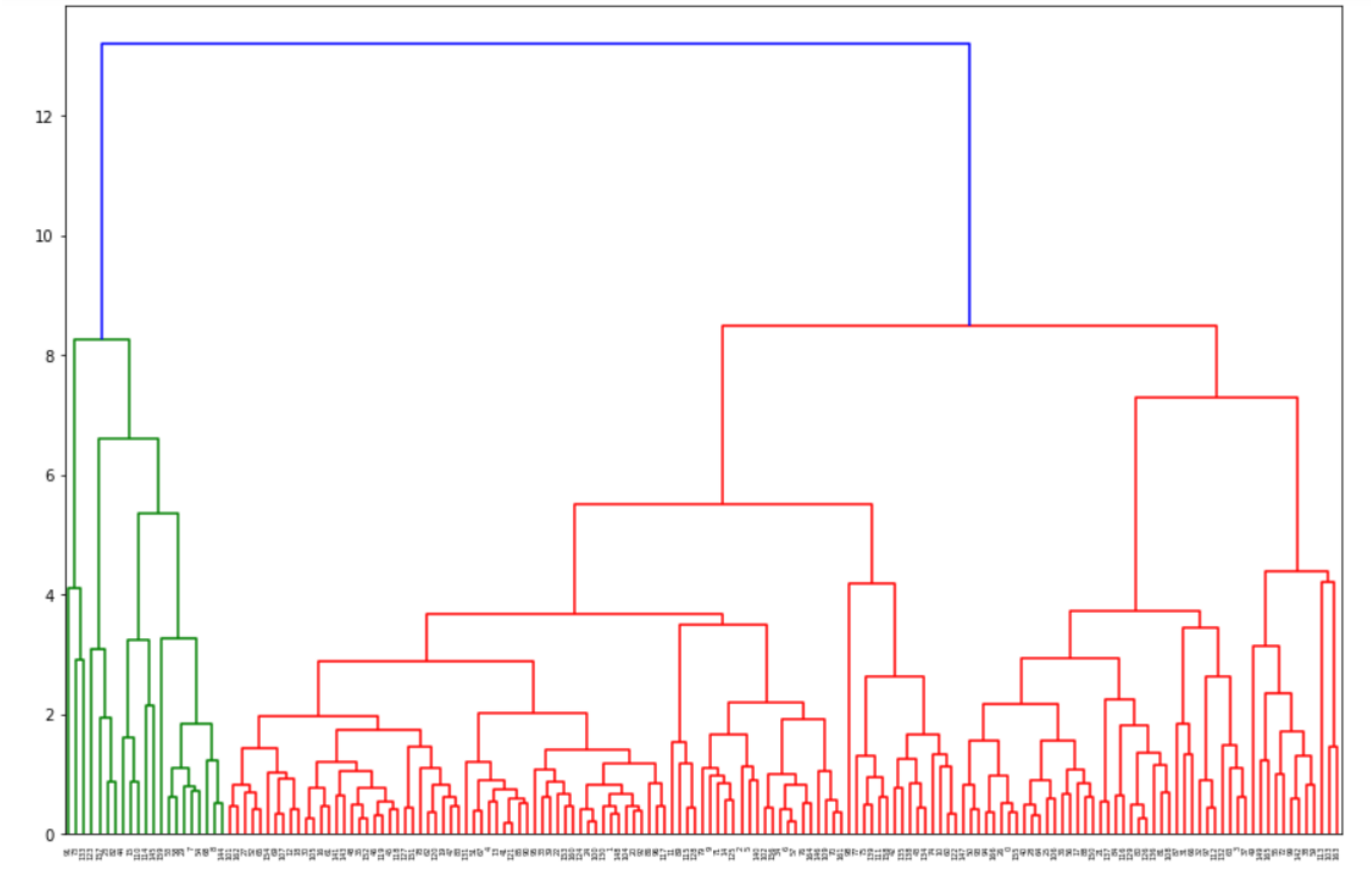
Lets try both the methods on our country data and see if the results are good enough.

Single Linkage:



Single Linkage do not give clear cluster formation so we have to try complete linkage in the next step.

Complete Linkage:



Now we see some good amount of clusters getting formed.

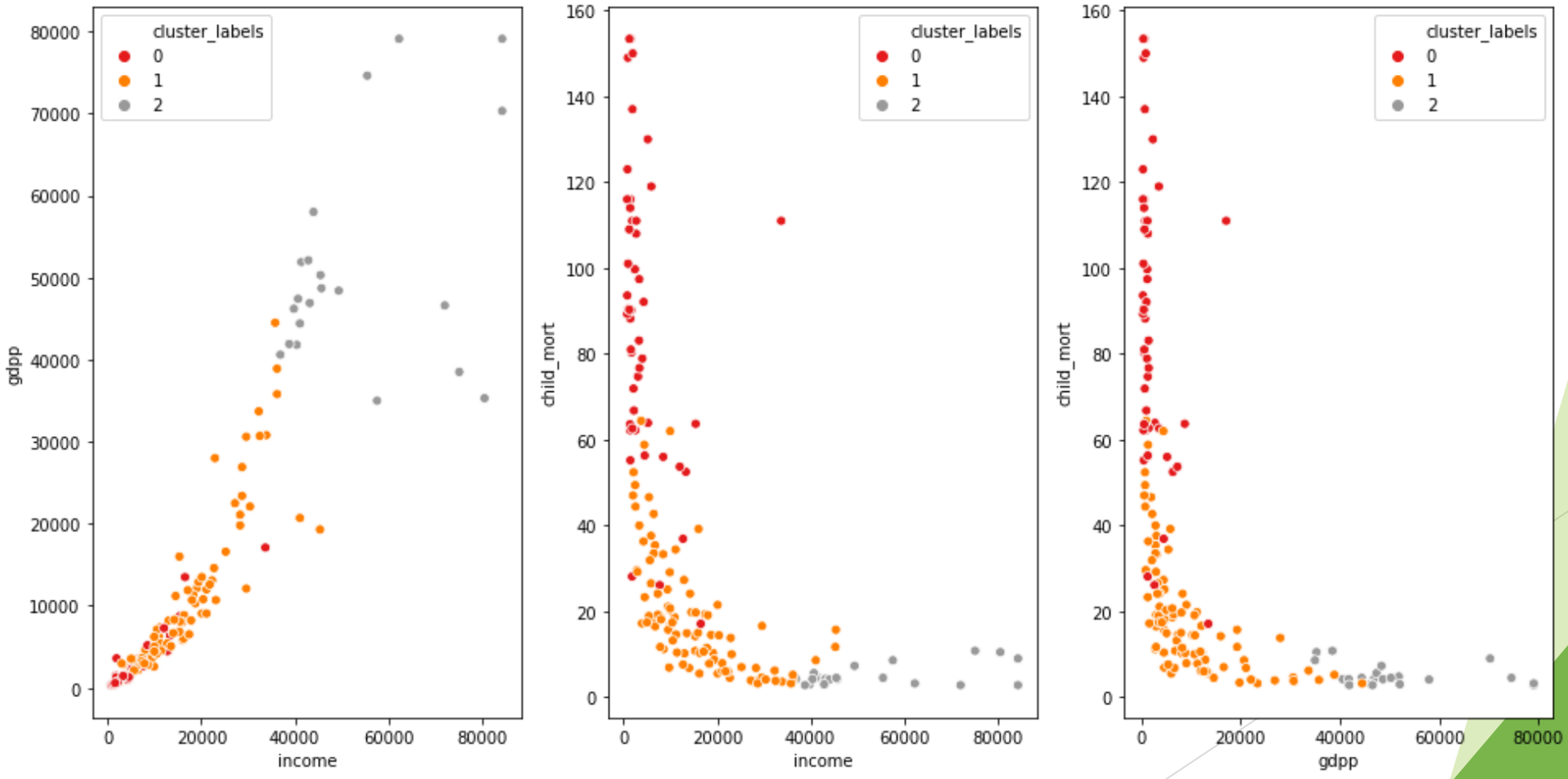
Now if we cut the tree at 5 clusters and look at our data head after assigning the cluster ids.
The results are as such:

Cluster ID	Value Counts
1	96
0	50
2	21

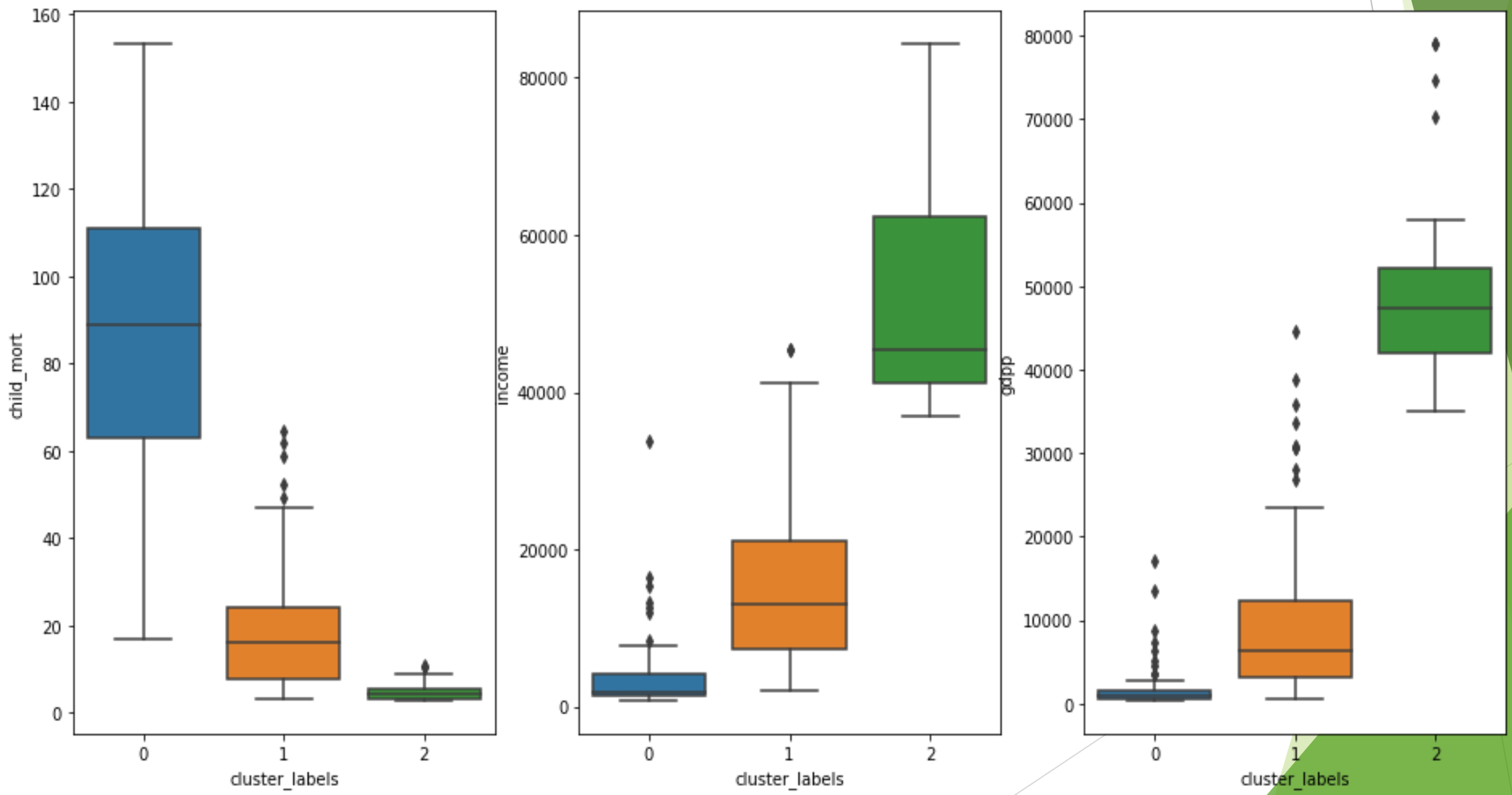
Hierarchical clustering Data head

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_id	cluster_labels
0	Afghanistan	90.2	55.30	41.9174	248.297	1610.0	9.44	56.2	5.82	553.0	2	0
1	Albania	16.6	1145.20	267.8950	1987.740	9930.0	4.49	76.3	1.65	4090.0	0	1
2	Algeria	27.3	1712.64	185.9820	1400.440	12900.0	16.10	76.5	2.89	4460.0	0	1
3	Angola	119.0	2199.19	100.6050	1514.370	5900.0	22.40	60.1	6.16	3530.0	2	0
4	Antigua and Barbuda	10.3	5551.00	735.6600	7185.800	19100.0	1.44	76.8	2.13	12200.0	0	1

There is an improper distribution of countries in each cluster. This can be shown in below scatter plot made after hierarchical clustering

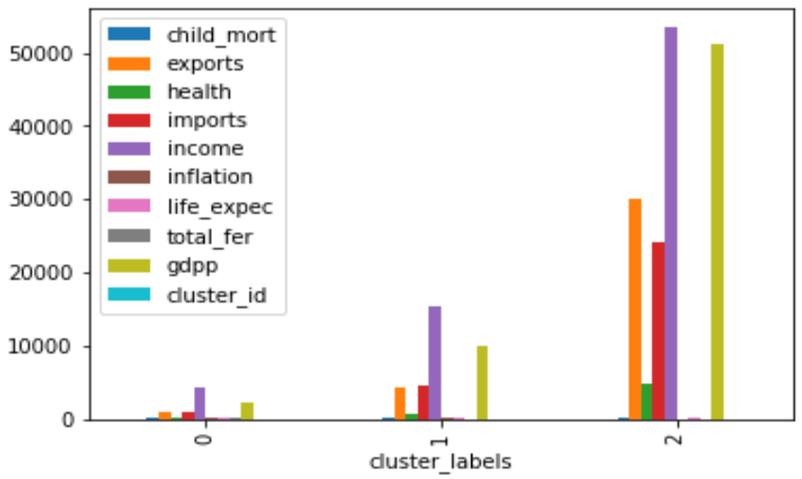


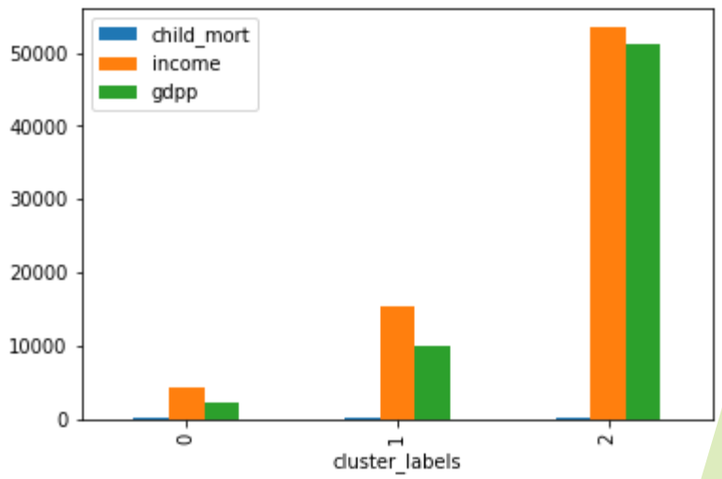
Below boxplot gives has certain inferences.
Cluster 0 has highest rate of child mortality and lowest gdpp, so it is in highest need of aid
Cluster 2 has highest income and gdpp but lowest child mortality rate, so it has lowest need of id



CLUSTER PROFILING

cluster_labels	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_id
0	87.586000	945.634750	126.481642	871.805773	4229.169600	11.797820	60.016400	4.875544	2157.944800	1.920000
1	19.188542	4326.711618	733.089171	4474.111767	15438.333333	5.936460	74.069479	2.181075	9849.187500	0.083333
2	5.176190	29964.696190	4731.309086	24182.246667	53421.333333	3.598248	80.298571	1.823962	51289.333333	1.000000





Child Mortality is highest for Cluster 0, so it needs some aid.
Income and Gdpp are measures of development. Higher the per capita income and gdpp better is the country's development. Income per capita and gdpp seems lowest for countries in cluster 0. Hence, these countries need some help.

FINAL ANALYSIS

It has been observed that clusters formed from both K-means and Hierarchical clustering and are not identical.

In this solution K-means and Hierarchical don't produce identical insights. It would be perfectly fine if both provide identical insights in any other case

The clusters formed in K-means is better as compared to Hierarchical as it seems to be more precise.

We observed that the clustering on the other clusters are not so prominent as it was in K-means

So, let's will proceed with the clusters formed by K-means

List of top 10 Countries which need help on the basis of health and socio-economic factors are in Cluster 0 formed by K means clustering

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_id
88	Liberia	89.3	62.457000	38.5860	302.80200	742.24	5.47	60.8	5.0200	331.62	0
37	Congo, Dem. Rep.	116.0	137.274000	26.4194	165.66400	742.24	20.80	57.5	6.5400	334.00	0
26	Burundi	93.6	22.243716	26.7960	104.90964	764.00	12.30	57.7	6.2600	331.62	0
112	Niger	123.0	77.256000	17.9568	170.86800	814.00	2.55	58.8	6.5636	348.00	0
31	Central African Republic	149.0	52.628000	17.7508	118.19000	888.00	2.01	47.5	5.2100	446.00	0
106	Mozambique	101.0	131.985000	21.8299	193.57800	918.00	7.64	54.5	5.5600	419.00	0
94	Malawi	90.5	104.652000	30.2481	160.19100	1030.00	12.10	53.1	5.3100	459.00	0
63	Guinea	109.0	196.344000	31.9464	279.93600	1190.00	16.10	58.0	5.3400	648.00	0
150	Togo	90.3	196.176000	37.3320	279.62400	1210.00	1.18	58.7	4.8700	488.00	0
132	Sierra Leone	153.4	67.032000	52.2690	137.65500	1220.00	17.20	55.0	5.2000	399.00	0

Thank You…