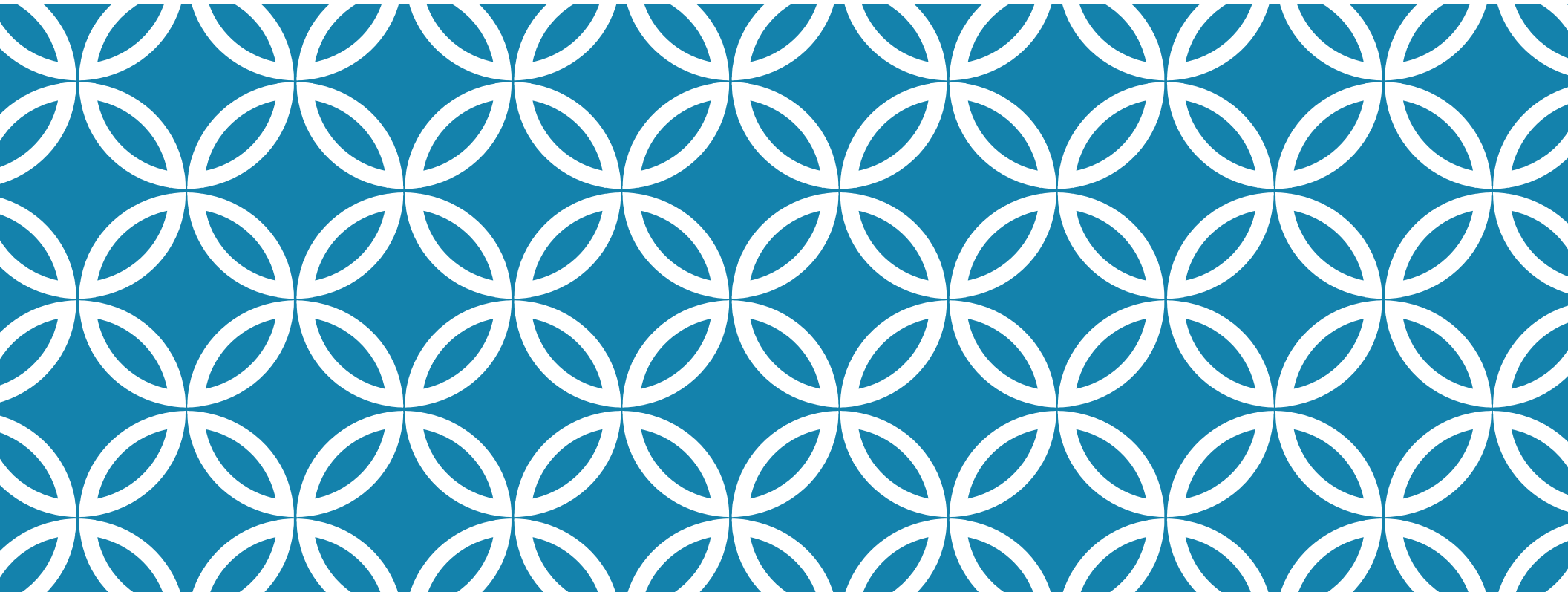


UNSUPERVISED LEARNING-CLUSTRING

PROBLEM STATEMENT

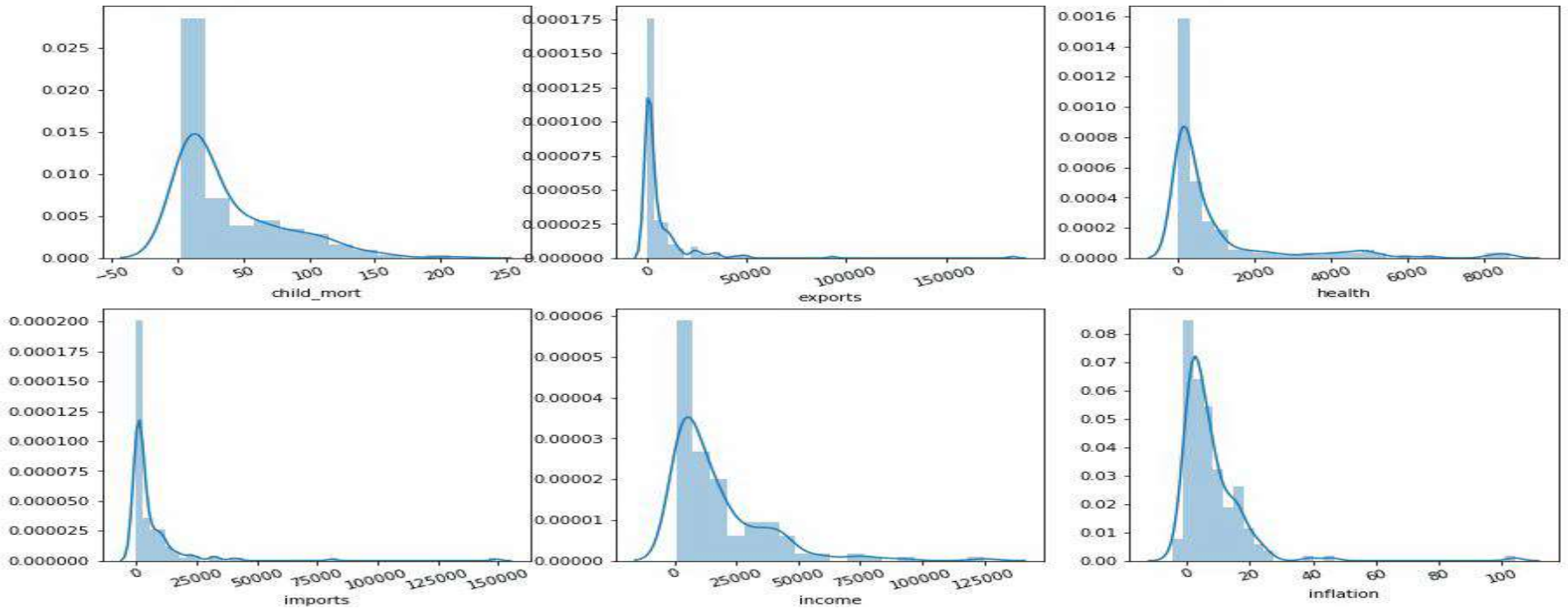
HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities. It runs a lot of operational projects from time to time along with advocacy drives to raise awareness as well as for funding purposes.

Categorise 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

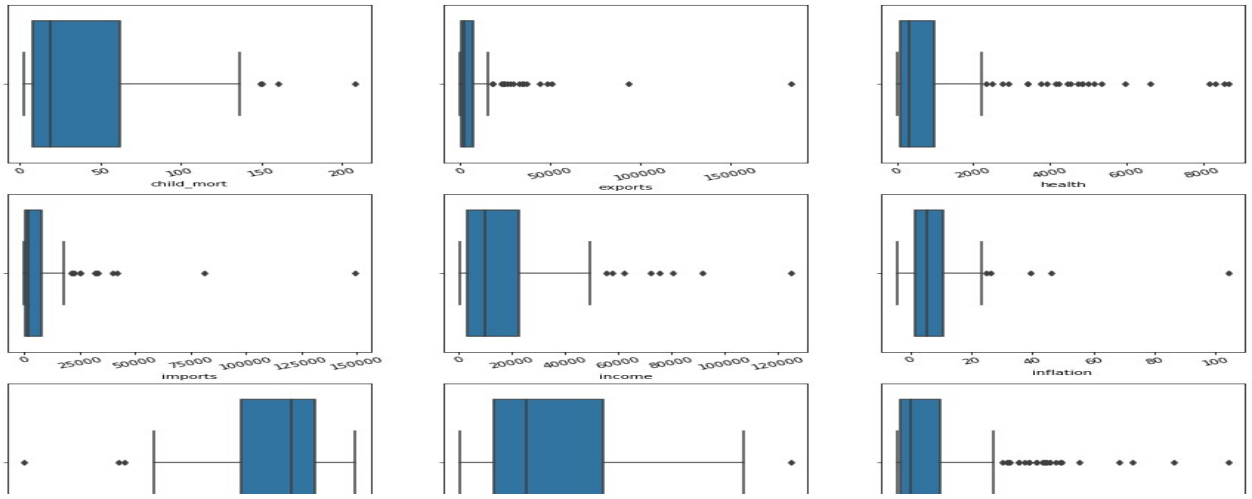
- Done EDA Analysis on data set
- Converted exports,health and imports from %gdpp into actual values
- Checked the distribution of values of each column. They are not normally distributed hence it is viable for Clustering

VISUALIZATION- DISTRIBUTION OF VALUES

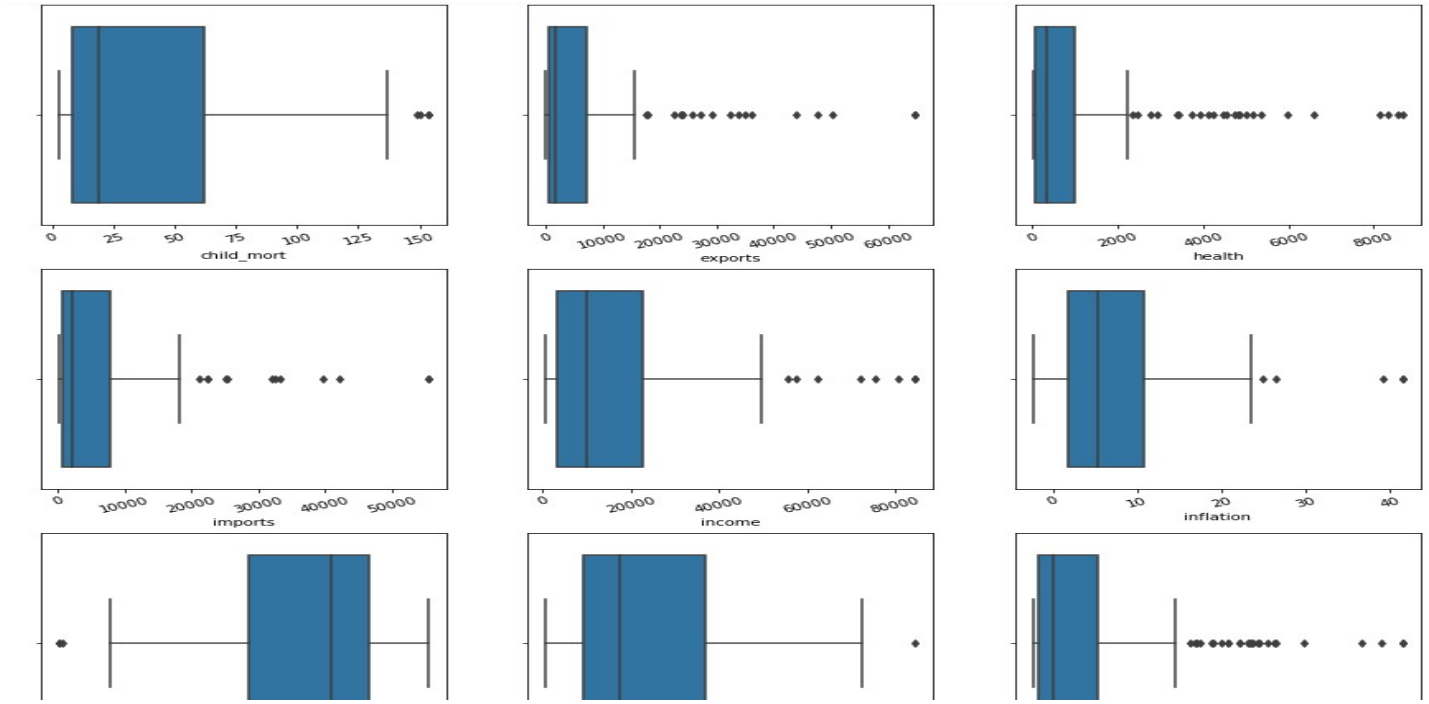


OUTLIER TREATMENT

Outliers has been addressed using capping technique. The values below 1% and above 99% values replaced with 1 and 99% values respectively



AFTER OUTLIER TREATMENT



HOPKINS STATISTICS

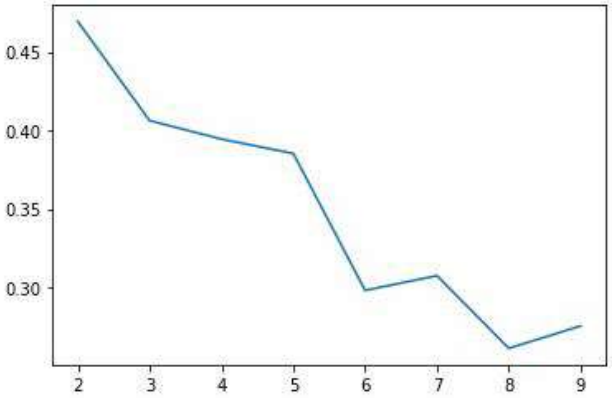
- Hopkins statistic is to identify how viable the data set for clustering
- We have got 0.96 Hopkin value for given data set which means that 96% values in data set can be differentiate with each other. Hence this data set is feasible for clustering.

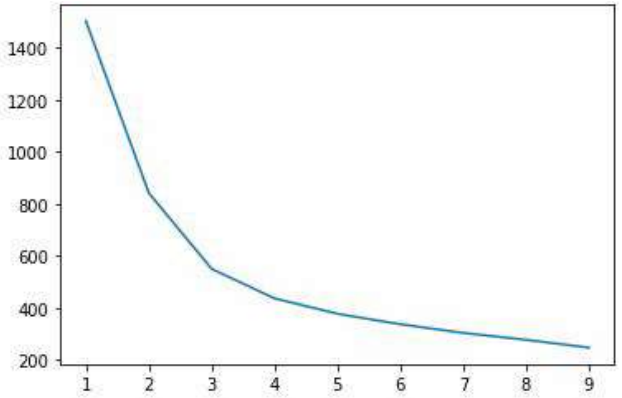
```
hopkins(df.drop(['ID','country'], axis = 1))
```

```
0.9639425593061799
```

K-MEANS ALGORITHM

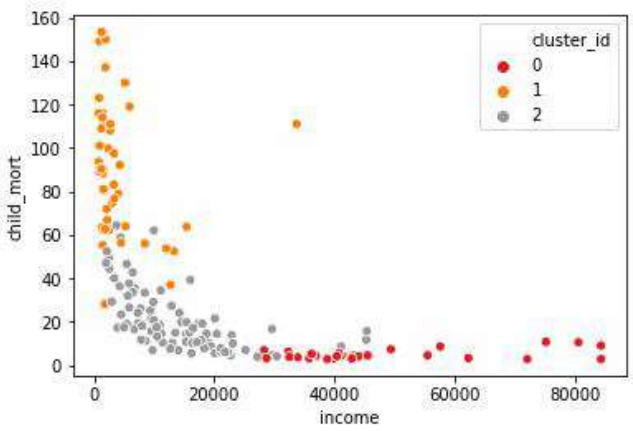
- Before applying K-means algorithm we have scaled the data using Standard Scaler
- Used silhouette score and elbow curve to identify no.of clusters initially i.e 3.

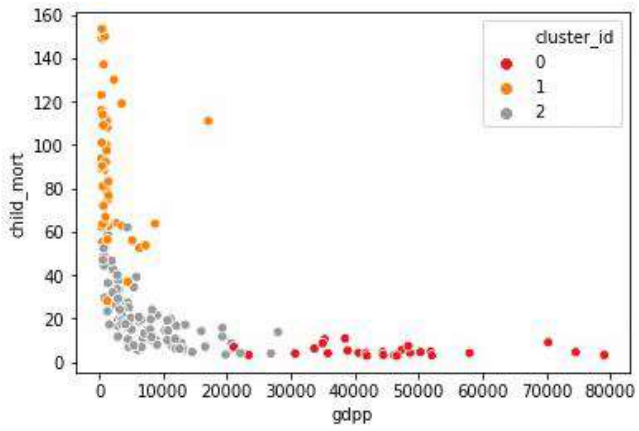


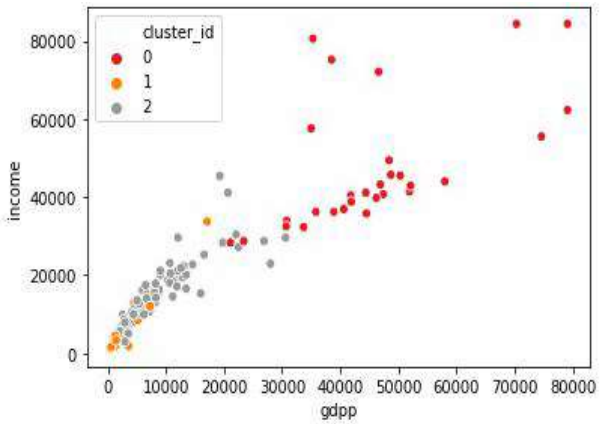


K-MEANS ALGORITHM

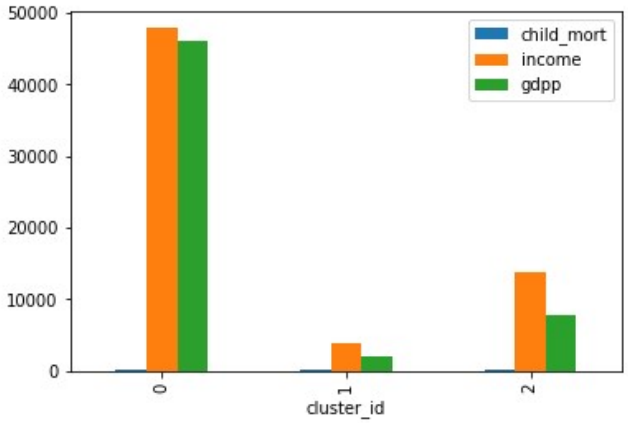
- Build Model with K-Means Algorithm and divide all datapoints into 3 clusters.
- Visualize clusters for the columns gdpp , income and child mortality
- Based on that we identified cluster 1 is has data points with low gdpp and income and high child mortality







CLUSTER PROFILING



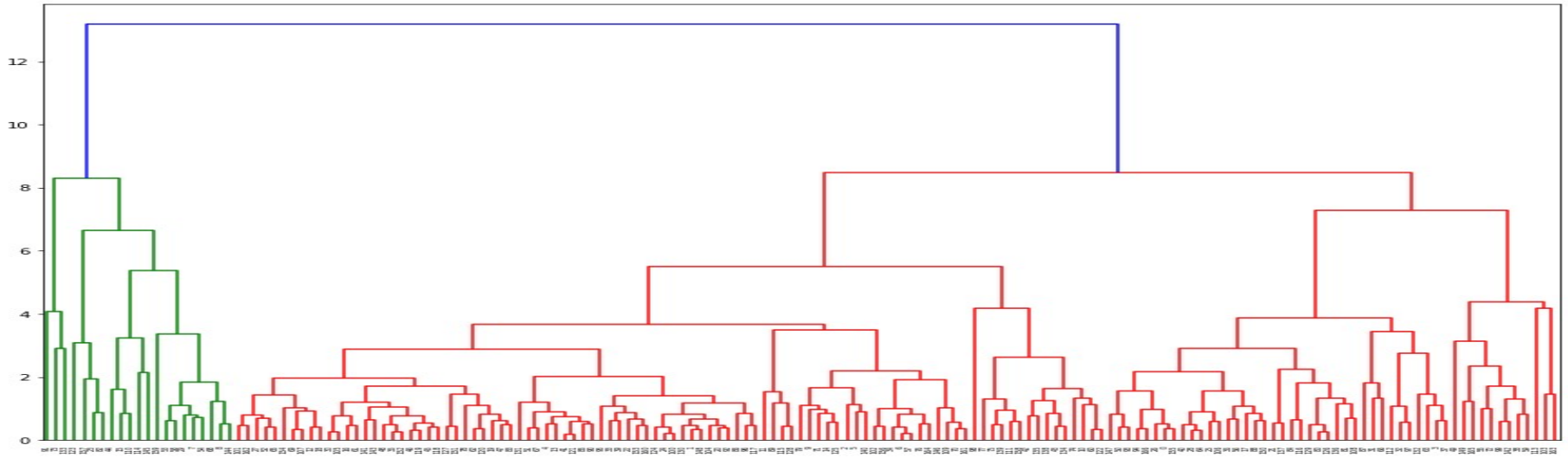
TOP 5 COUNTRIES TO BE CONSIDERED

- Below are the top five countries to be considered for funding using K-Means Clustering as these countries have low gdpp and income and high child mortality

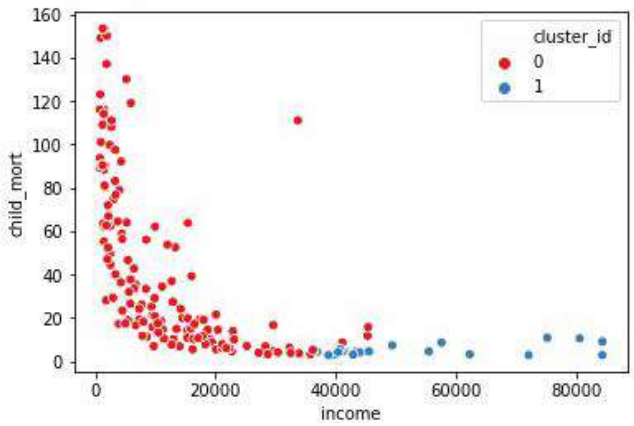
	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	ID	cluster_id
132	Sierra Leone	153.4	67.032	52.2690	137.655	1220.0	17.20	55.00	5.20	399.0	232	1
66	Haiti	153.4	101.286	45.7442	428.314	1500.0	5.45	47.16	3.33	662.0	166	1
32	Chad	150.0	330.096	40.6341	390.195	1930.0	6.39	56.50	6.59	897.0	132	1
31	Central African Republic	149.0	52.628	17.7508	118.190	888.0	2.01	47.50	5.21	446.0	131	1
97	Mali	137.0	161.424	35.2584	248.508	1870.0	4.37	59.50	6.55	708.0	197	1

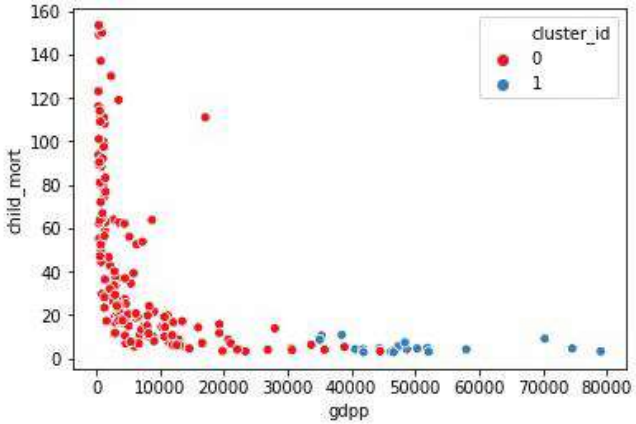
HIERARCHICAL CLUSTERING

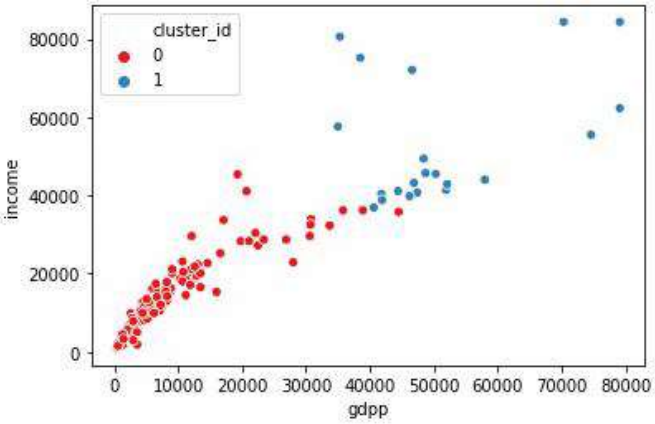
- Find the no.of clusters using dendrograms and identified that data points can be grouped into two clusters (Tried 4 clusters but when profiled we found 2 clusters are enough)



VISUALIZING CLUSTERS

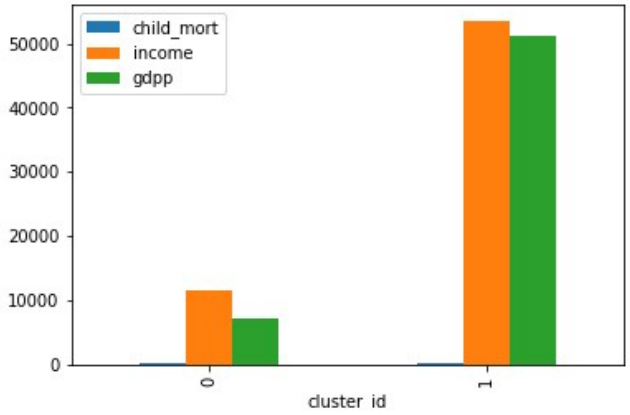






CLUSTER PROFILING –HIERARCHICAL CLUSTRING

- Using cluster profiling we identified cluster 0 represents countries with low income,gdpp and high child mortality



FINAL RESULTS

- Below are the top five countries with low income , gdpp and high child mortality using hierarchical clustering can be considered for funding.

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	ID	cluster_id
132	Sierra Leone	153.4	67.032	52.2690	137.655	1220.0	17.20	55.00	5.20	399.0	232	0
66	Haiti	153.4	101.286	45.7442	428.314	1500.0	5.45	47.16	3.33	662.0	166	0
32	Chad	150.0	330.096	40.6341	390.195	1930.0	6.39	56.50	6.59	897.0	132	0
31	Central African Republic	149.0	52.628	17.7508	118.190	888.0	2.01	47.50	5.21	446.0	131	0
97	Mali	137.0	161.424	35.2584	248.508	1870.0	4.37	59.50	6.55	708.0	197	0