Help International - NGO

Clustering Of Countries

Problem Statement

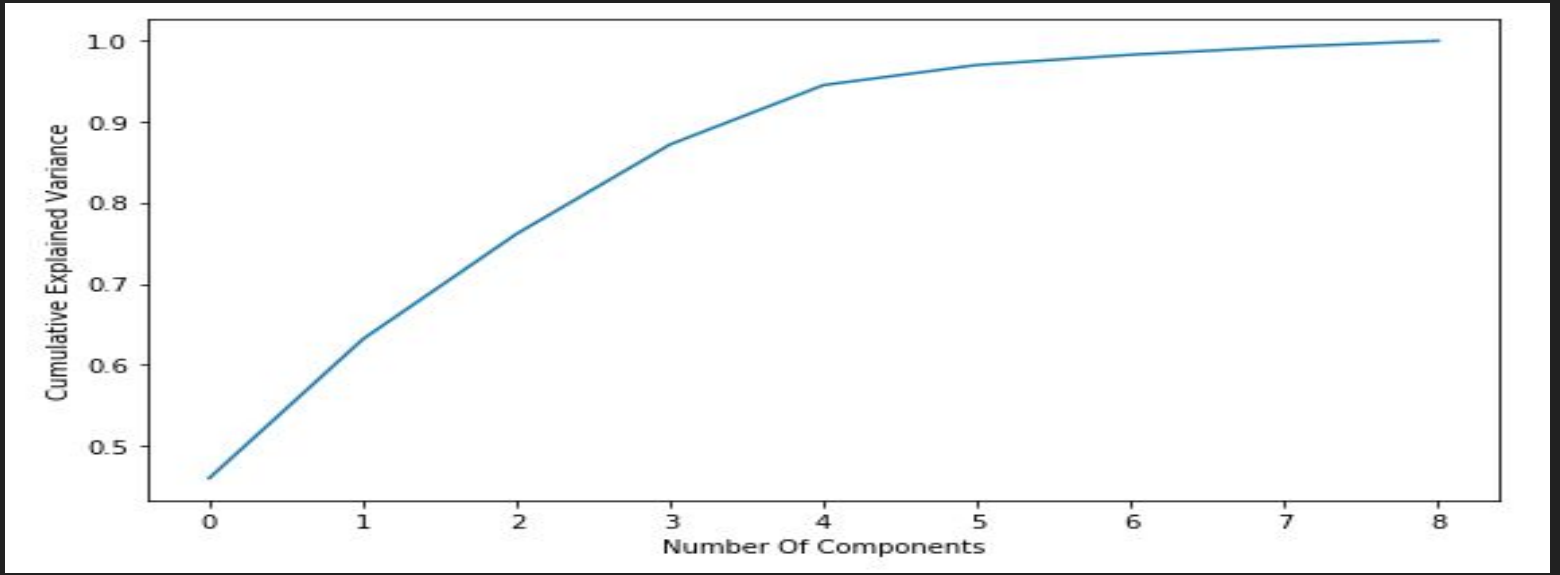
Select list of countries that are in dire need of aid due to their weak socio-economic conditions.

Analysis Approach

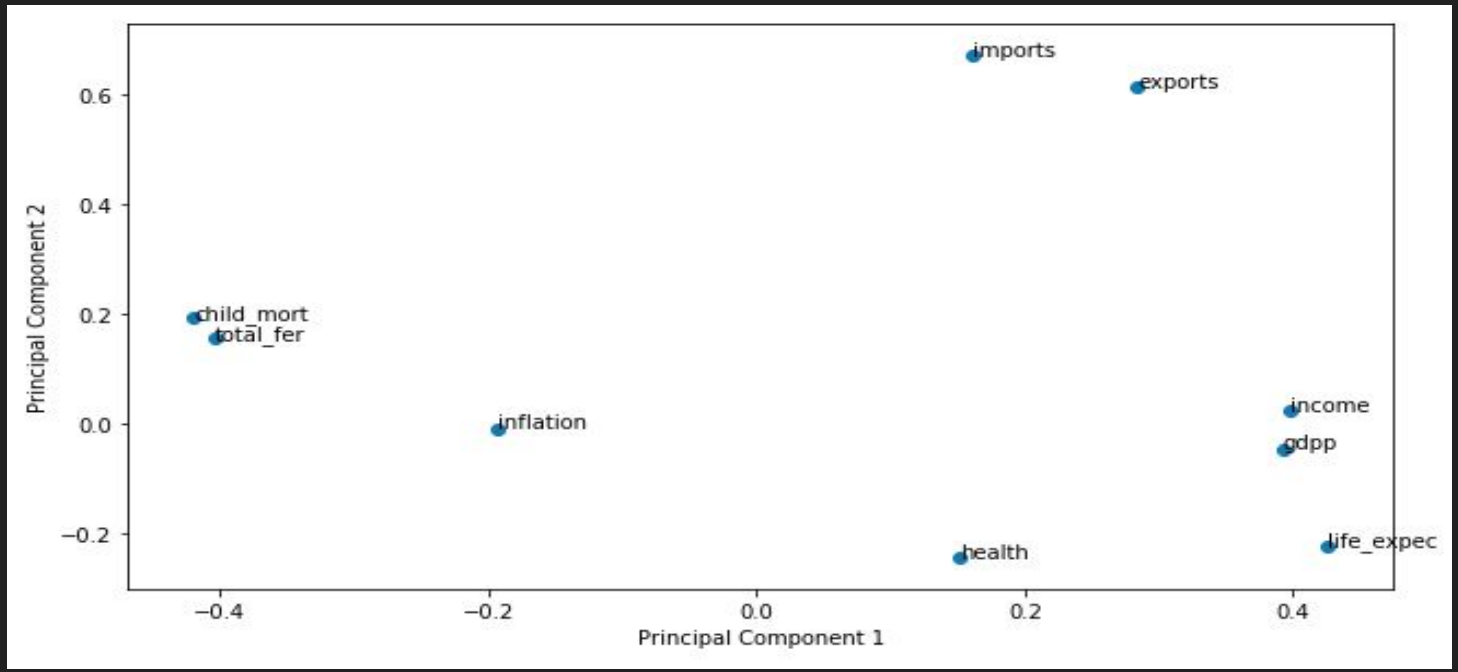
- Selecting most important principal components from the dataset that explains maximum variability between features and can help achieve at least 94-95% of variation with just 4-5 features. This approach is called Principal Component Analysis(PCA).
- Run the dataset through Hopkins method to find out if Clustering is possible or not. If the result is more than 70-75% then we will continue with clustering.
- We will perform clustering on the PCA dataset to form different clusters of countries that can help us narrow down to the ones that are under-developed and need help from the NGO.

Principal Component Analysis (PCA)

- Based on PCA operation performed on the standardize dataset, we use following Scree plot to determine the number of components needed. We can see that about 5 principal components explain close to 95% of the variance which is good enough for further analysis.

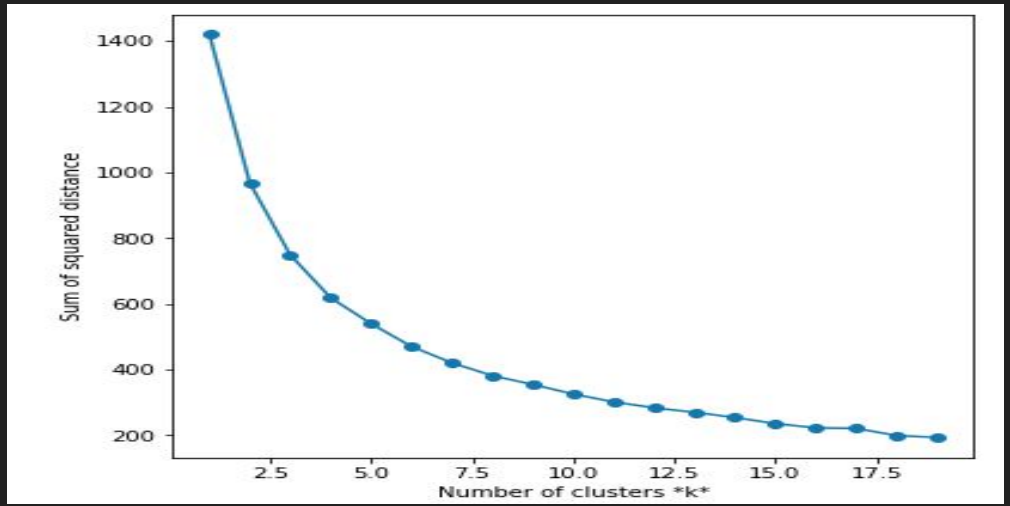


- From the below scatter plot we can see how various features are being explained by just two principal components, with features child_mort and gdpp are furthest from each other tells us they have maximum variance.

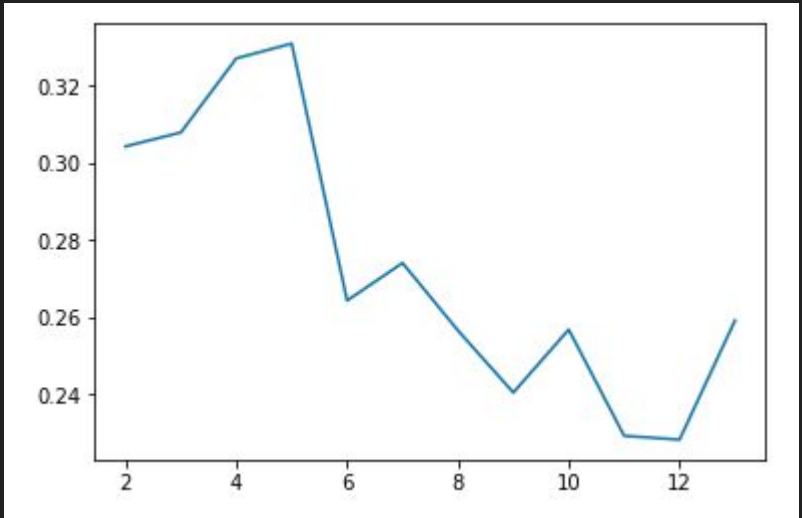


Clustering

- Based on K-Means and Hierarchical clustering performed on the dataset, we saw 5 and 6 clusters forming via each clustering approach respectively.
- We plotted Sum of squared distances between each cluster centroids to finalize 5 as our final number of clusters for K-Means.

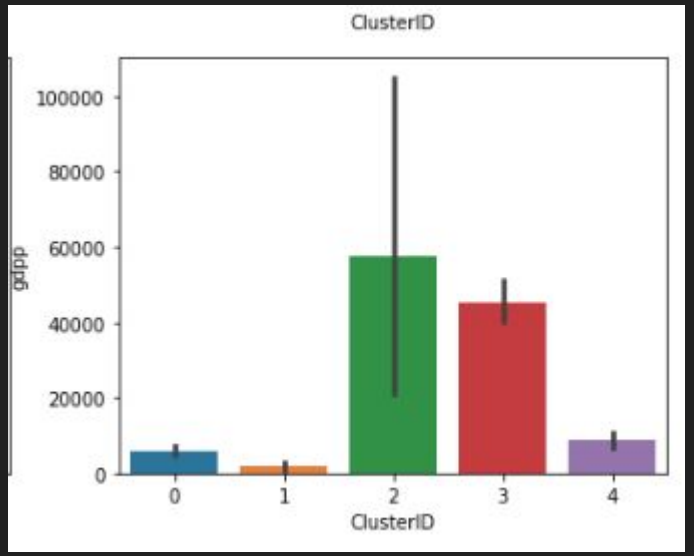


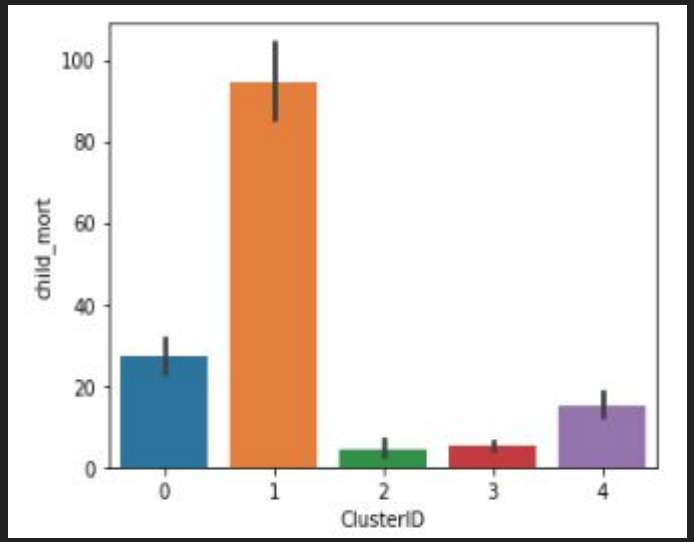
- Based in Silhouette analysis and the plot below, we can clearly see 5 as maximum numbers of clusters.
- Hence we proceeded with 5 clusters in Kmeans clustering algorithm.



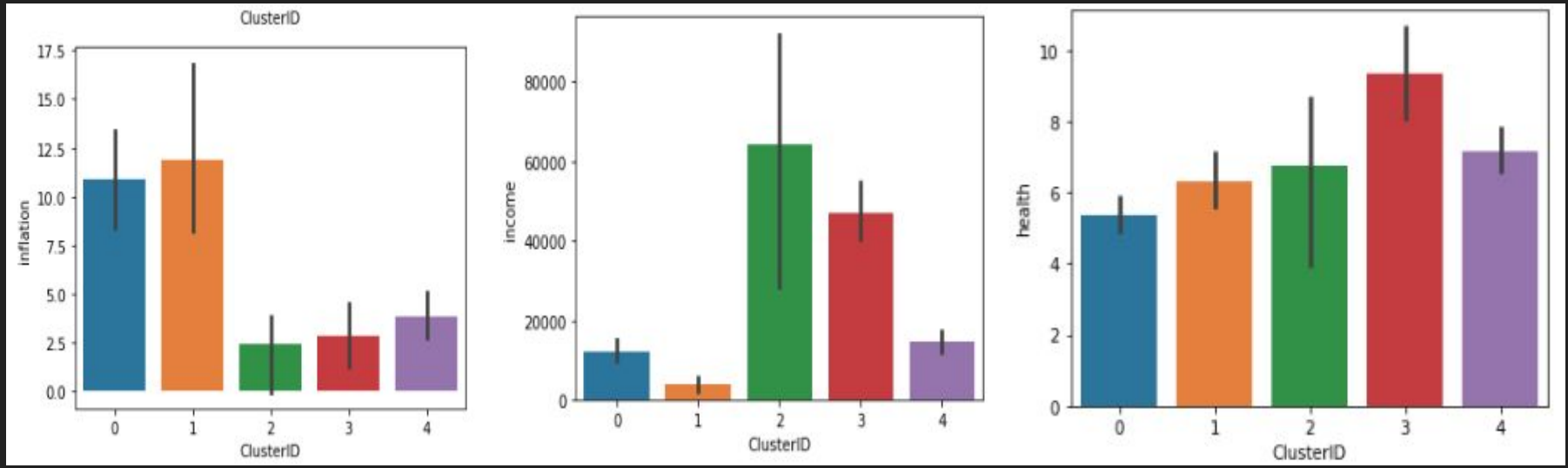
Clustering Outcome

- From the below two bar plot its visible that country GDP and income is inversely correlated to child mortality.

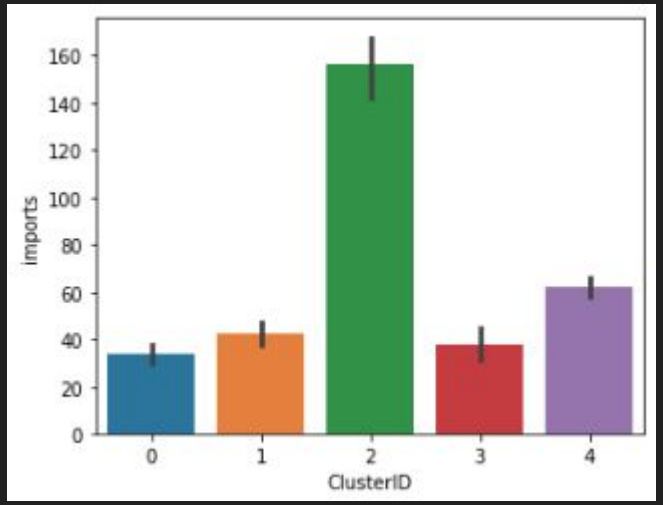


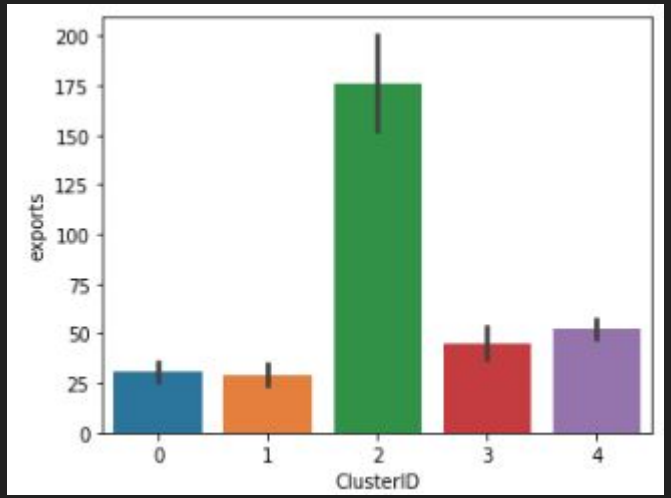


- Inflation and income or GDPP is also, inversely correlated. However, income and GDPP are directly correlated.
- Health is positively correlated with GDP or Income.



- Countries also have high export and import, if they are more developed , i.e, they have high GDPP and Per capita income. That is represented by cluster id 2.





Summary

- Based on various analysis performed on the data, we concluded that countries that have high child mortality rate and low income/gdpp should be considered for assistance.
- We also considered the health variable for each countries and as it turns out countries with low income/gdpp are also the countries that has poor overall health and thus these countries also have low life expectancy.
- We also see pattern with child mortality and total fertility count. Underdeveloped countries mostly have high total fertility and high child mortality.
- We considered cluster 1 to be most suitable for help.

- Here are the numbers that has been kept as a cutoff based on various clusters formed on the dataset to come up with the final list of countries that are in dire need of aid.
  - GDPP less than **1793.34**
  - Child_mort more than 96
  - Life_expec less than 59.01
  - Total_fer more than 5.06
  - Health less than 6.34
- Based on the above criterias, final list of countries that should be given assistance are:
  - Cameroon
  - Central African Republic
  - Chad
  - Cote d'Ivoire
  - Guinea
  - Mozambique
  - Niger

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	ClusterID
28	Cameroon	108.0	22.2	5.13	27.0	2660	1.91	57.3	5.11	1310	1
31	Central African Republic	149.0	11.8	3.98	26.5	888	2.01	47.5	5.21	446	1
32	Chad	150.0	36.8	4.53	43.5	1930	6.39	56.5	6.59	897	1
40	Cote d'Ivoire	111.0	50.6	5.30	43.3	2690	5.39	56.3	5.27	1220	1
63	Guinea	109.0	30.3	4.93	43.2	1190	16.10	58.0	5.34	648	1
106	Mozambique	101.0	31.5	5.21	46.2	918	7.64	54.5	5.56	419	1
112	Niger	123.0	22.2	5.16	49.1	814	2.55	58.8	7.49	348	1