Clustering & PCA Assignment

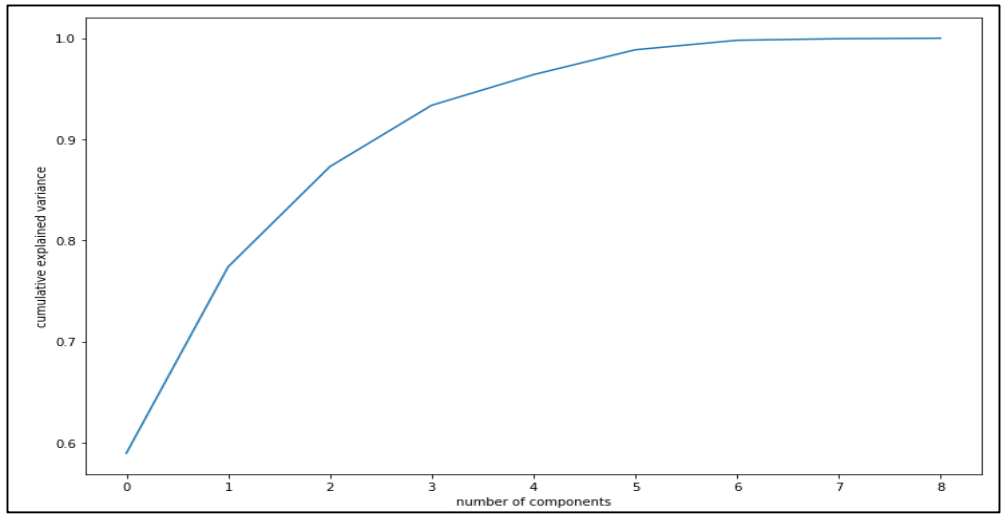
Problem Statement:

'HELP' 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. 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

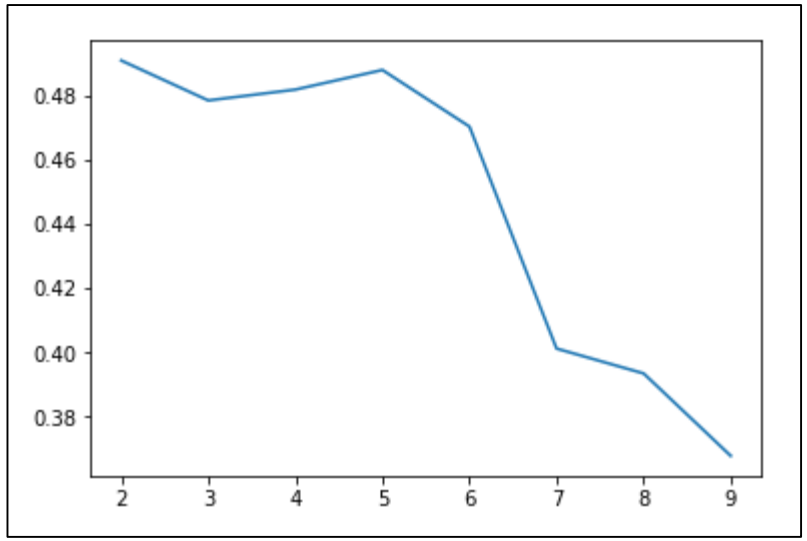
Analysis Approach:

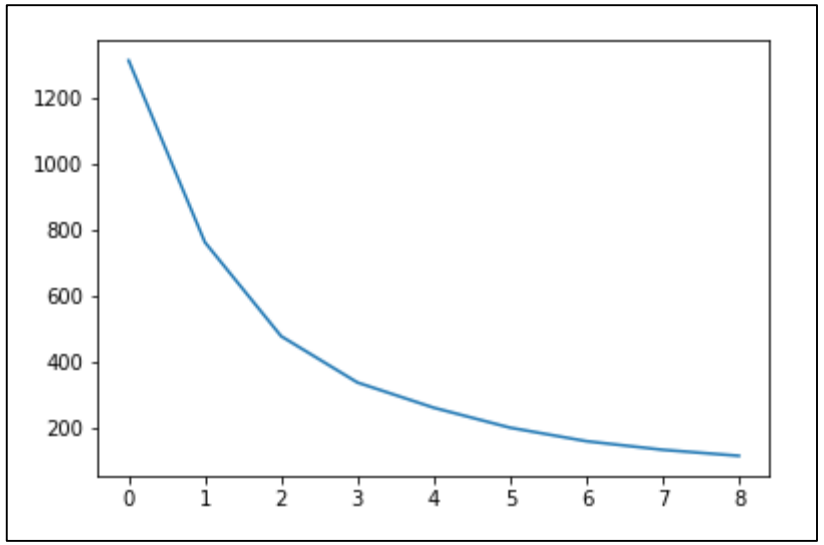
1. Read the given data into Python file and do necessary data cleaning and prepare the right data for model building.
2. Perform scaling so that all variables are brought to same range for further steps
3. By performing PCA, reduce the dimensions of data without loosing any information. Basis the variance ratio decide on number of components
4. Basis ***Hopkins score*** decide if the data frame can be clustered.
5. Perform both K means Clustering and hierarchal clustering for the PCA data set. (Decide on 1 clustering approach basis results)
6. After clusters are identified and tagged to main data set, visualize the data, remove outliers (if necessary) and understand which countries would need aid

Clustering & PCA Assignment



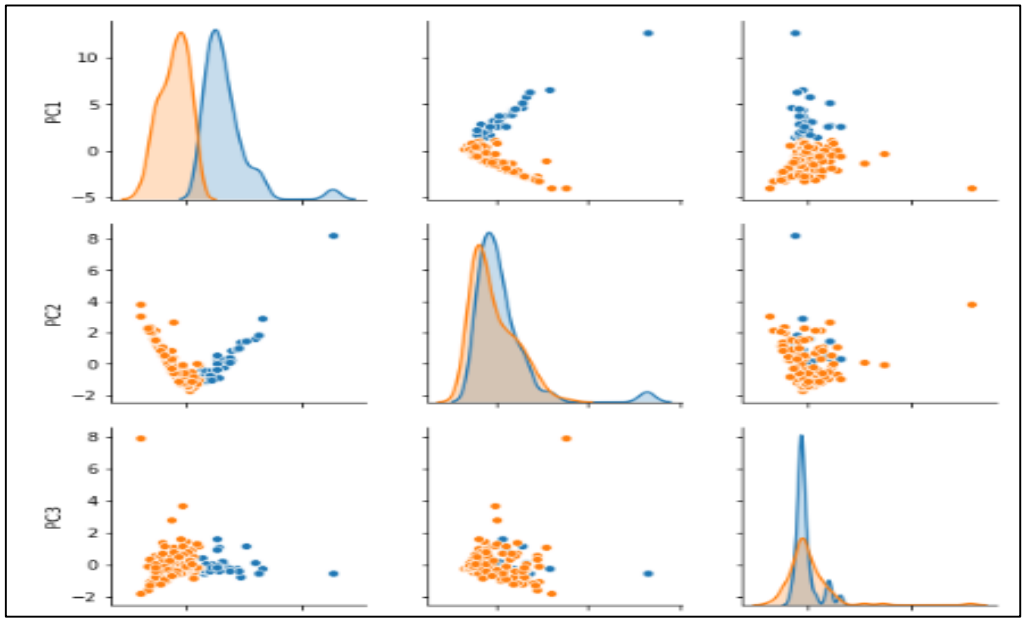
Since around 93% of data is being explained with 3 components, going ahead with 3 PCA variables



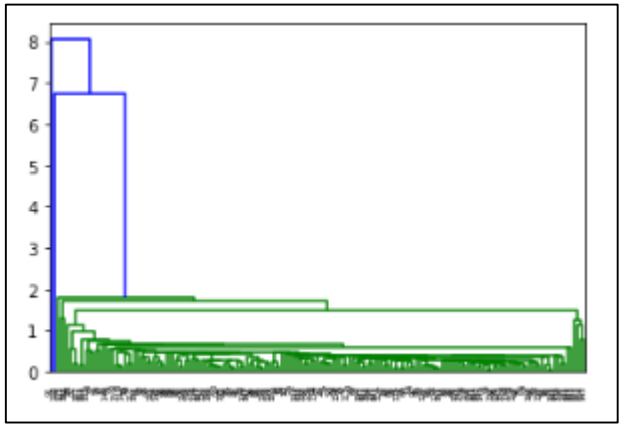


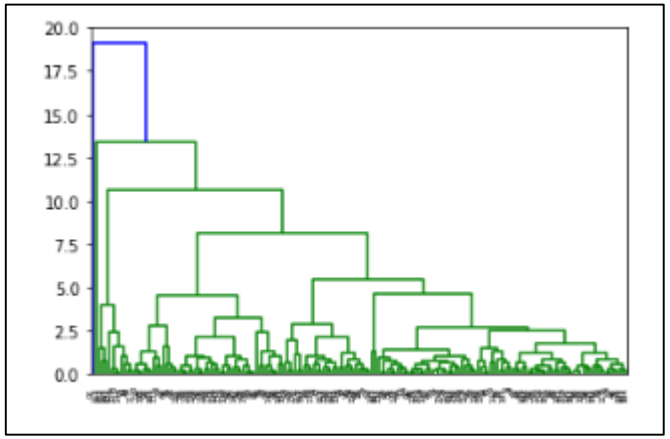
From the silhouette score analysis we find that 2 seems to be a good number of clusters for K means algorithm

Clustering & PCA Assignment



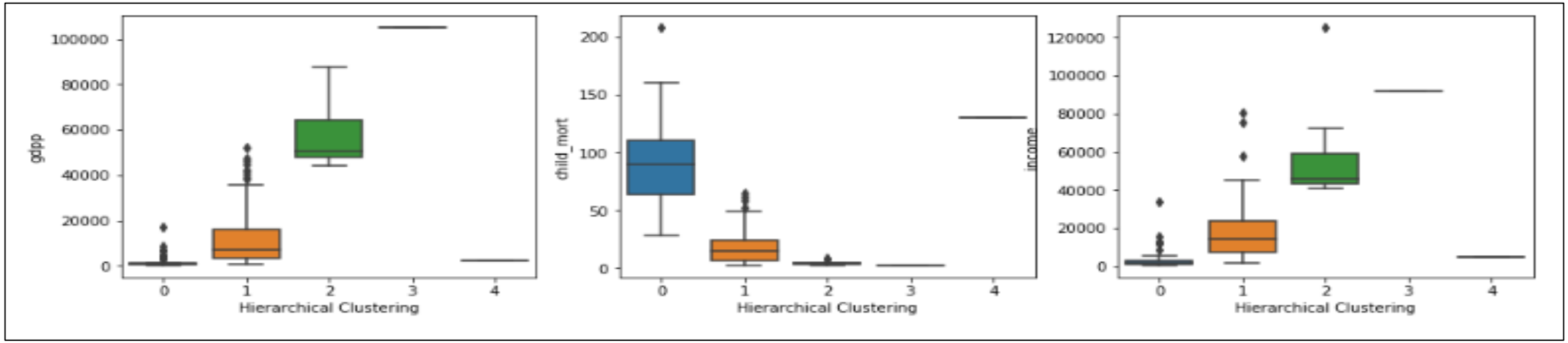
Basis K means 2 clusters were created and are visualized as seen against all 3 PCA components





Basis Hierarchal clustering, 5 clusters were formed and decided to go with 5 clusters

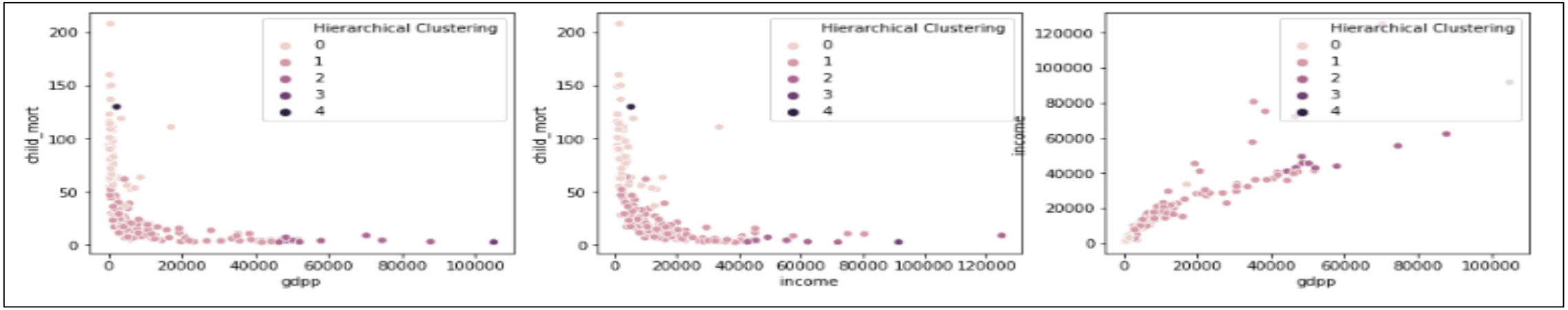
Clustering & PCA Assignment



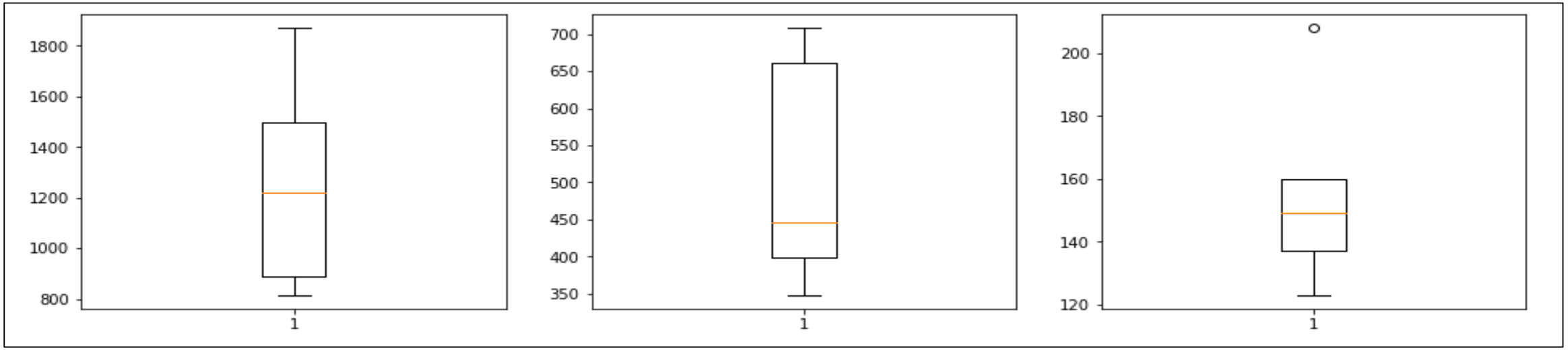
Basis Hierarchal clustering, 5 clusters were formed and can be clearly seen that Cluster 0 has low income and low gdpp when compared with countries from other clusters.

So decided to deep dive into cluster 0

Clustering & PCA Assignment



Clearly, Countries under cluster 0 need a helping hand. Deep dive into cluster 0 and removed few outliers. Below graph shows the same for 3 main variables



Clustering & PCA Assignment

	country	income	gdpp	child_mort
31	Central African Republic	888	446	149.0
66	Haiti	1500	662	208.0
97	Mali	1870	708	137.0
112	Niger	814	348	123.0
132	Sierra Leone	1220	399	160.0

Countries: Central African Republic, Haiti, Mali, Niger & Sierra Leone are the ones whose income & gdpp is very less compared to others and also high on child Mort. Hence suggesting these countries to be given a helping hand by CEO of the NGO.