CLUSTERING & PCA

HELP INTERNATIONAL - CLUSTERING OF COUNTRIES

ANMOL PARIDA

PROBLEM STATEMENT
HELP INTERNATIONAL

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 program, 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.

DATA EXPLORATION

'Country-data.csv' contains all the information of the countries.

Fields in the Data set: country, child_mort, exports, health, imports, income, inflation, life_expec, total_fer, gdpp

We will be analyzing our results the variables: **child_mort, income, gdpp**

DATA CLEANING AND MANIPULATION
COUNTRY-DATA.CSV

Observation

- None of the rows have missing values
- All variables are int64/float64 excluding country, no categorical variables
- No Data Preparation or Dummy variable creation is required for the columns in this dataset

SCALING AND STANDARDIZATION

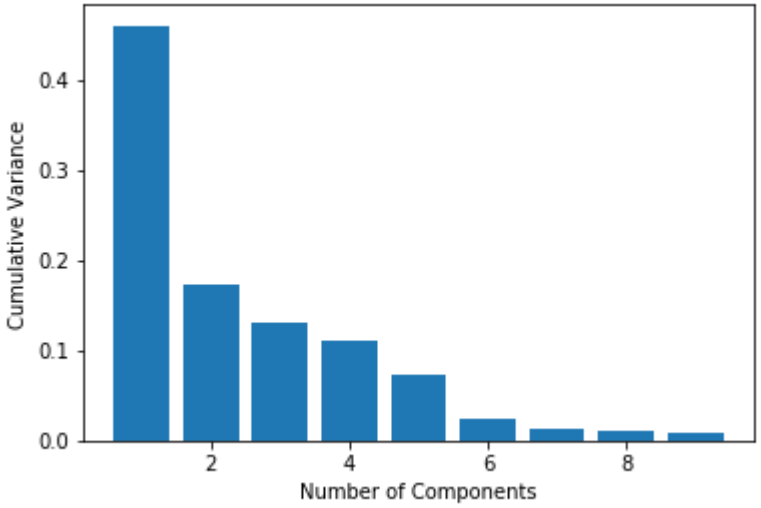
- We preformed the data description to understand the Data

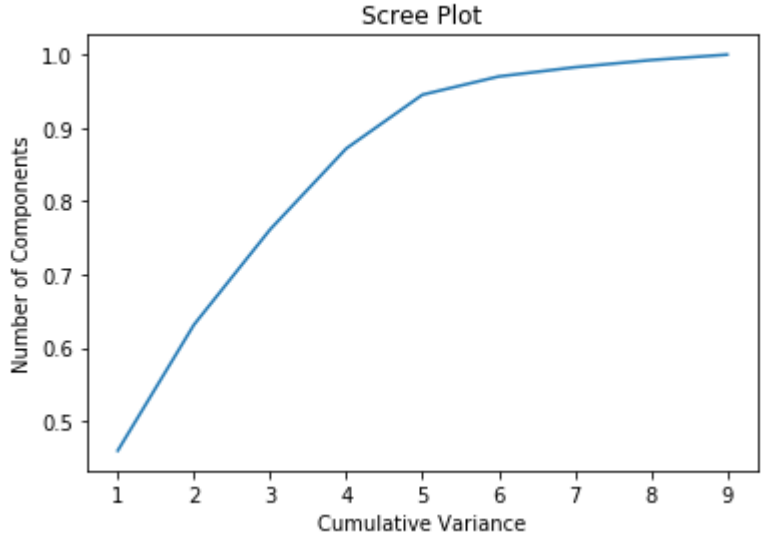
	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
count	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000
mean	38.270060	41.108976	6.815689	46.890215	17144.688623	7.781832	70.555689	2.947964	12964.155689
std	40.328931	27.412010	2.746837	24.209589	19278.067698	10.570704	8.893172	1.513848	18328.704809
min	2.600000	0.109000	1.810000	0.065900	609.000000	-4.210000	32.100000	1.150000	231.000000
25%	8.250000	23.800000	4.920000	30.200000	3355.000000	1.810000	65.300000	1.795000	1330.000000
50%	19.300000	35.000000	6.320000	43.300000	9960.000000	5.390000	73.100000	2.410000	4660.000000
75%	62.100000	51.350000	8.600000	58.750000	22800.000000	10.750000	76.800000	3.880000	14050.000000
90%	100.220000	70.800000	10.940000	75.420000	41220.000000	16.640000	80.400000	5.322000	41840.000000
95%	116.000000	80.570000	11.570000	81.140000	48290.000000	20.870000	81.400000	5.861000	48610.000000
99%	153.400000	160.480000	13.474000	146.080000	84374.000000	41.478000	82.370000	6.563600	79088.000000
max	208.000000	200.000000	17.900000	174.000000	125000.000000	104.000000	82.800000	7.490000	105000.000000

- Data was scaled and standardized before performing the PCA.

PRINCIPAL COMPONENT ANALYSIS (PCA)

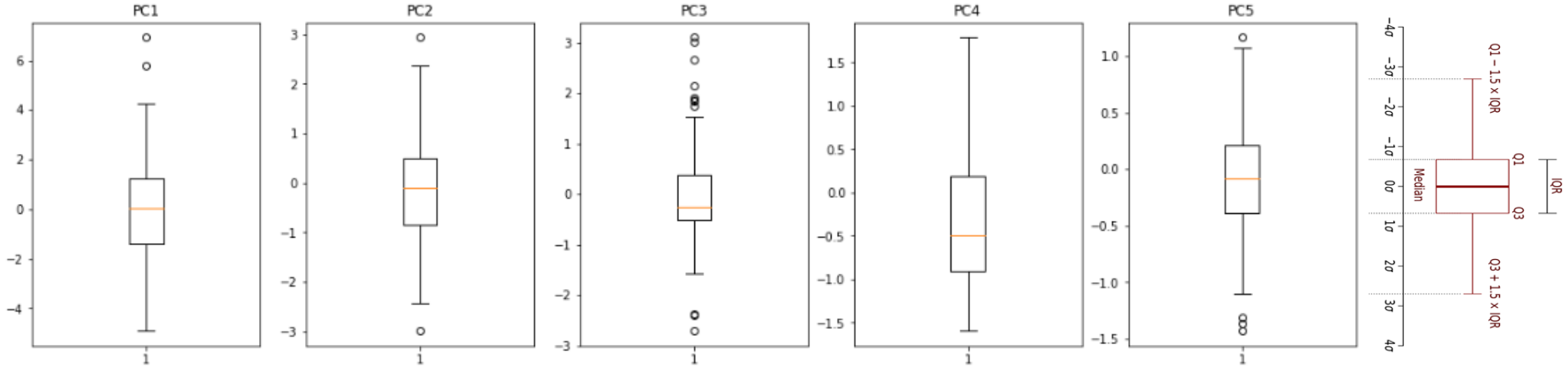
- For PCA we used sklearn.decomposition import PCA
- We transformed and fit the data to perform PCA
- We used pca.components_ and pca.explained_variance_ratio_ to find out the principal components
- 5 Principal Components were identified from the Graphs





OUTLIER ANALYSIS AND TREATMENT
DATA PREPARATION FOR CLUSTERING

- Outlier treatment was done after the Scaling and PCA for all 5 principal components.



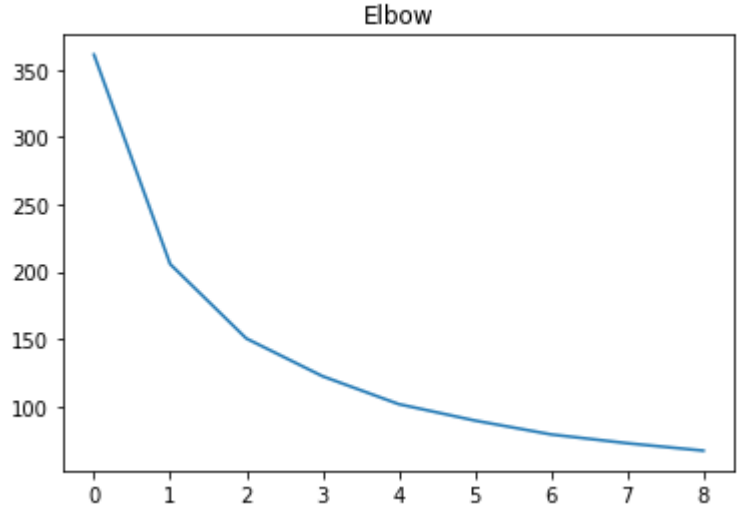
CLUSTERING
HOPKINS STATISTIC

- The Hopkins is a way of measuring the cluster tendency of a data set.
- It belongs to the family of sparse sampling tests. It acts as a statistical hypothesis test where the null hypothesis is that the data is generated by a Poisson point process and are thus uniformly randomly distributed.
- A value close to 1 tends to indicate the data is highly clustered, random data will tend to result in values around 0.5, and uniformly distributed data will tend to result in values close to 0.

- Hopkins Statistic for our analysis: 0.648080923499628

K-MEANS CLUSTERING
SILHOUETTE AVG SCORE & ELBOW CURVE

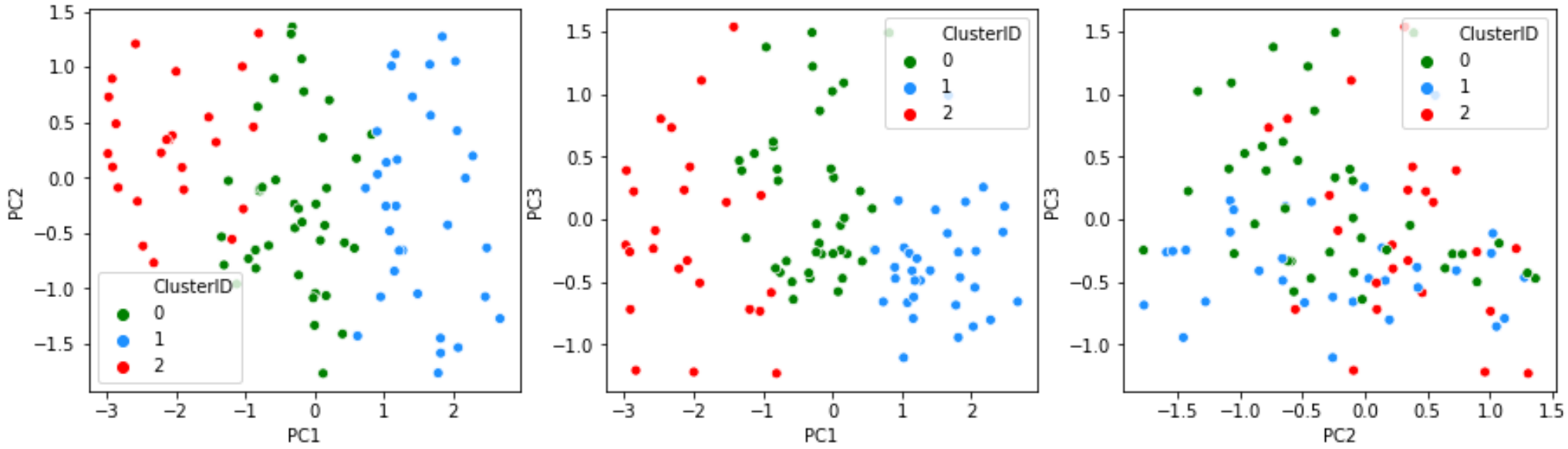
- For K-Means clustering we need the number of clusters to be identified.
- For this SILHOUETTE_AVG_SCORE was calculated and plotted on the curve.
- Elbow curve was obtained and analysed for number of clusters.
- Number of Clusters selected: 3



K-MEANS CLUSTERING

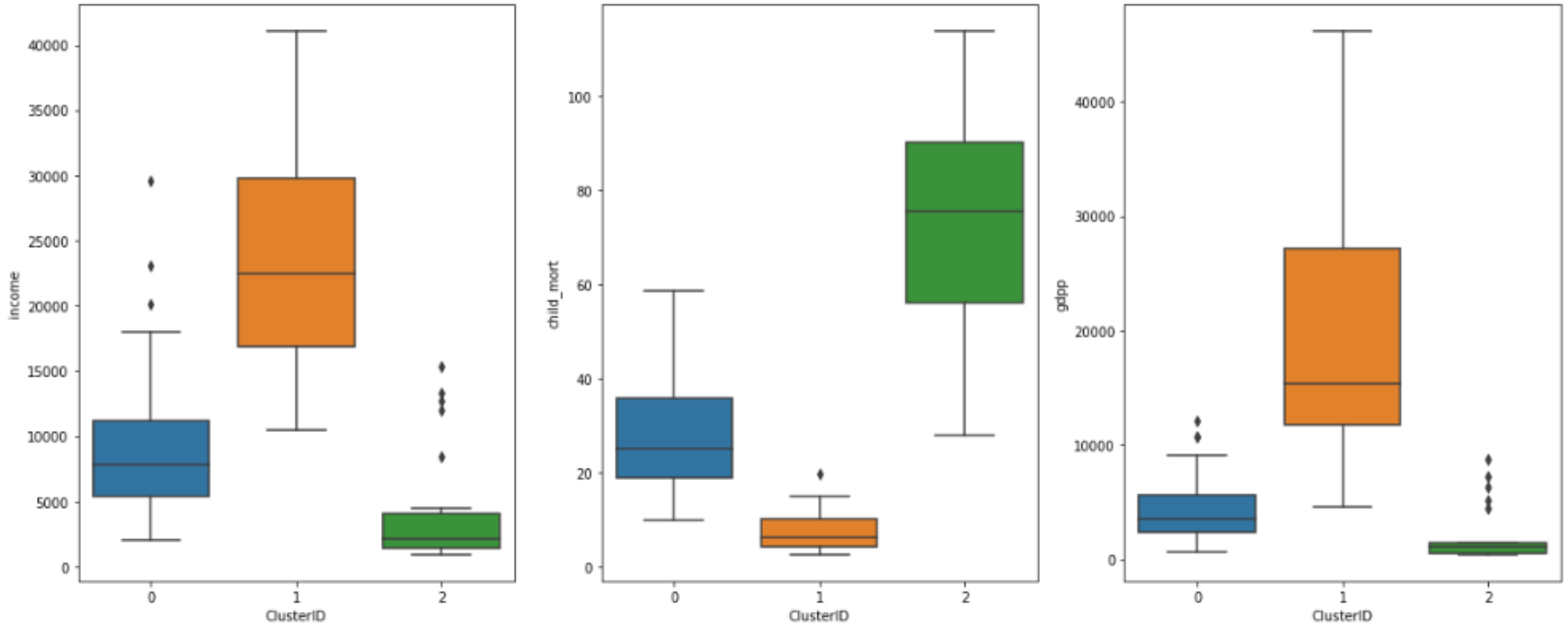
CLUSTER PROFILING - 3 CLUSTERS

Visualising the 3 clusters for for First 3 Components

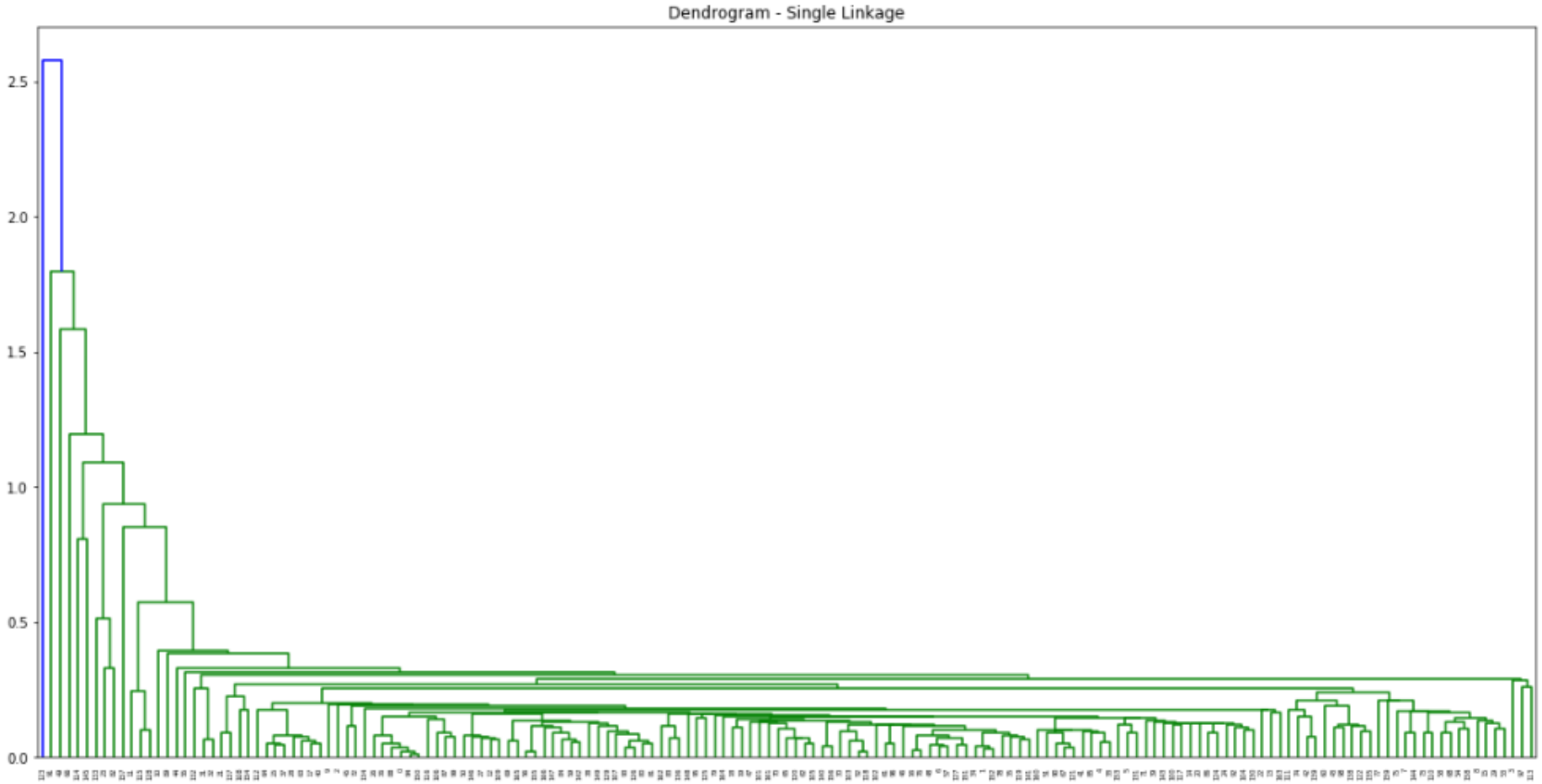


K-MEANS CLUSTERING
POST ANALYSIS AND RECOMENDATION

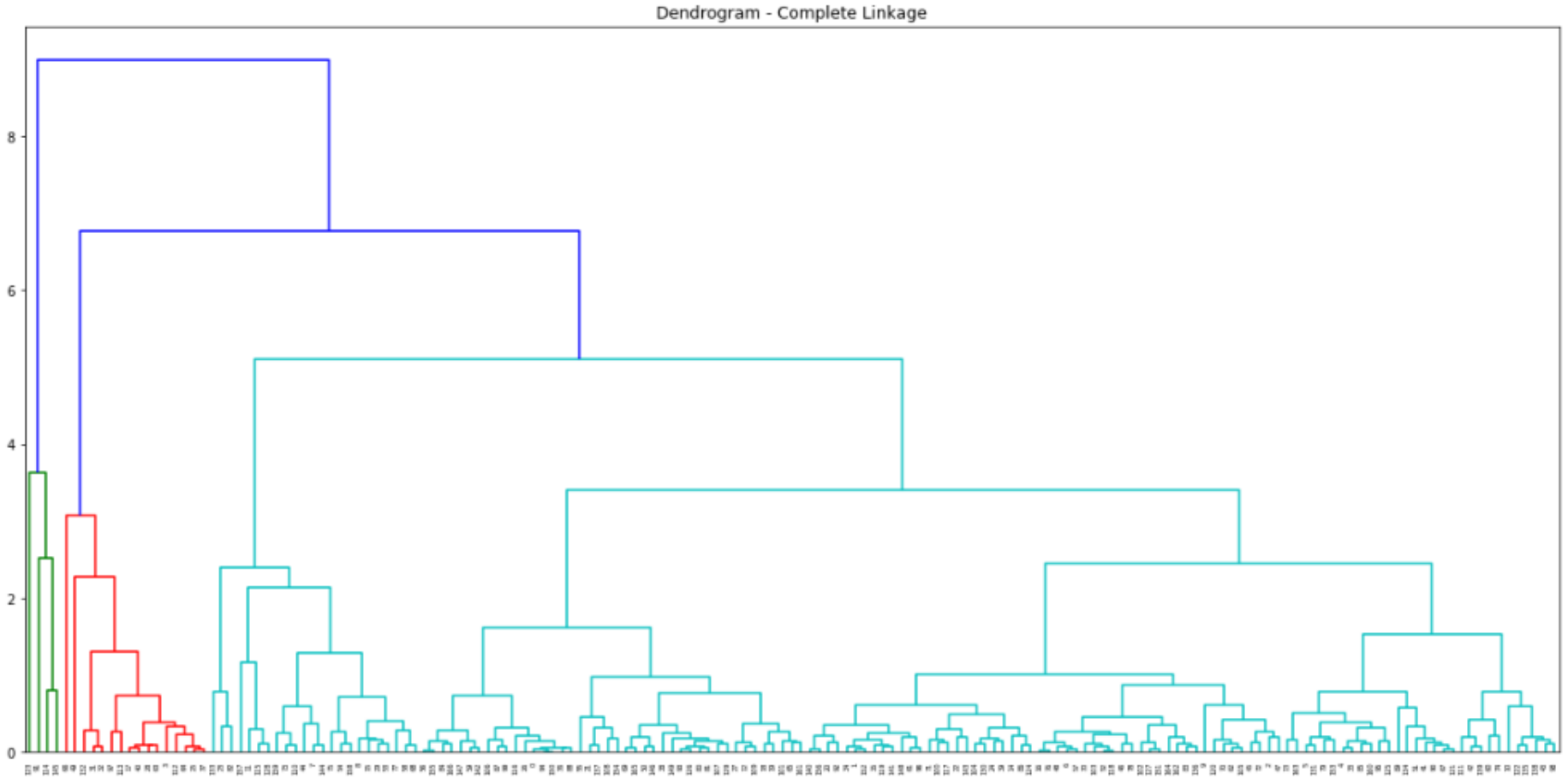
- Countries in Cluster 2 have an average child mortality 90, 90/1000 child die before the age of 5
- Countries in Cluster 2 have an average gdpp of 553 which is very low compared to clusters 0 & 1
- Countries in Cluster 2 have an average income of 1600 which is very low compared to clusters 0 & 1



HIERARCHICAL CLUSTERING
SINGLE LINKAGE



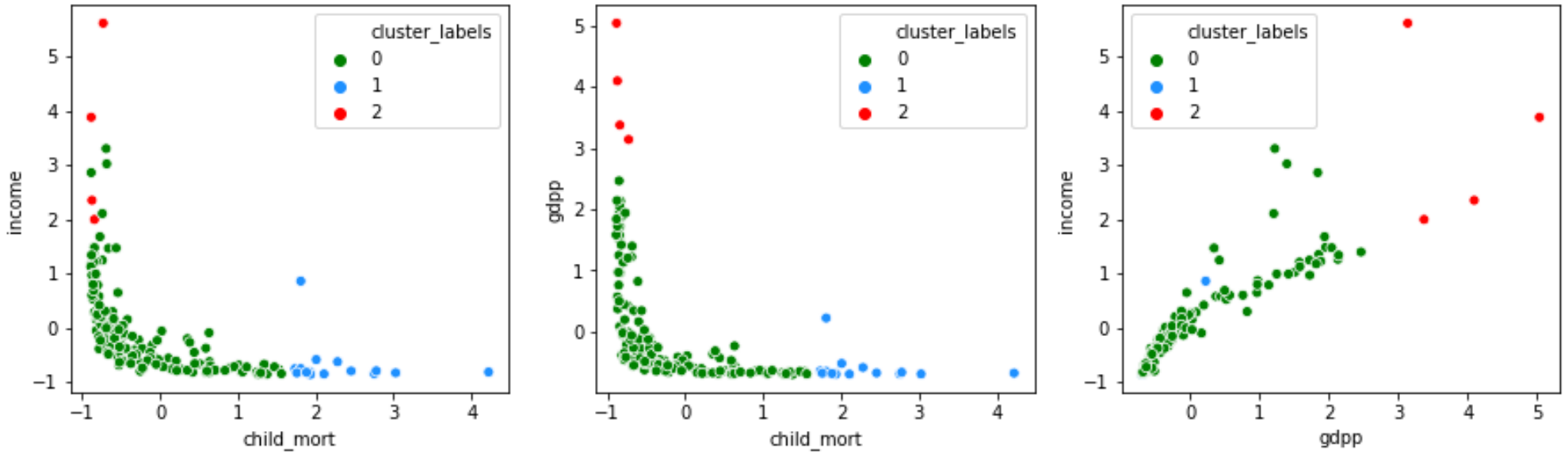
HIERARCHICAL CLUSTERING
COMPLETE LINKAGE



HIERARCHICAL CLUSTERING

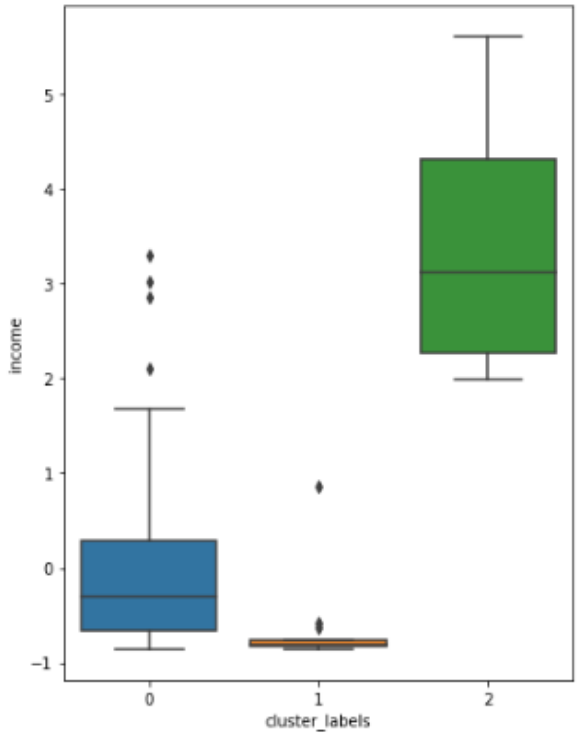
CLUSTER PROFILING - 3 CLUSTERS

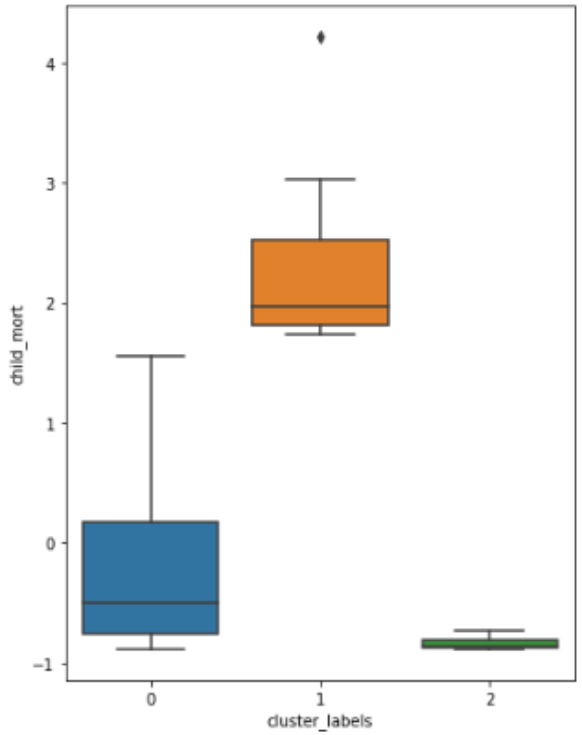
Visualizing the 3 clusters for First 3 Components

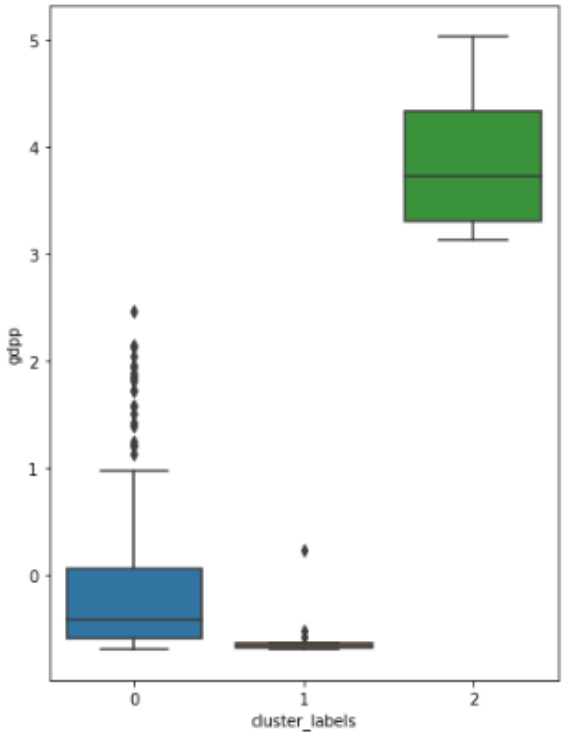


HIERARCHICAL CLUSTERING
POST ANALYSIS AND RECOMENDATION

- Countries in Cluster 2 have good income, gdpp and low child mortality.
- Countries in Cluster 1 have income almost similar to Cluster 0 but with high child mortality.
- In my analysis countries from Cluster 1 should be considered.

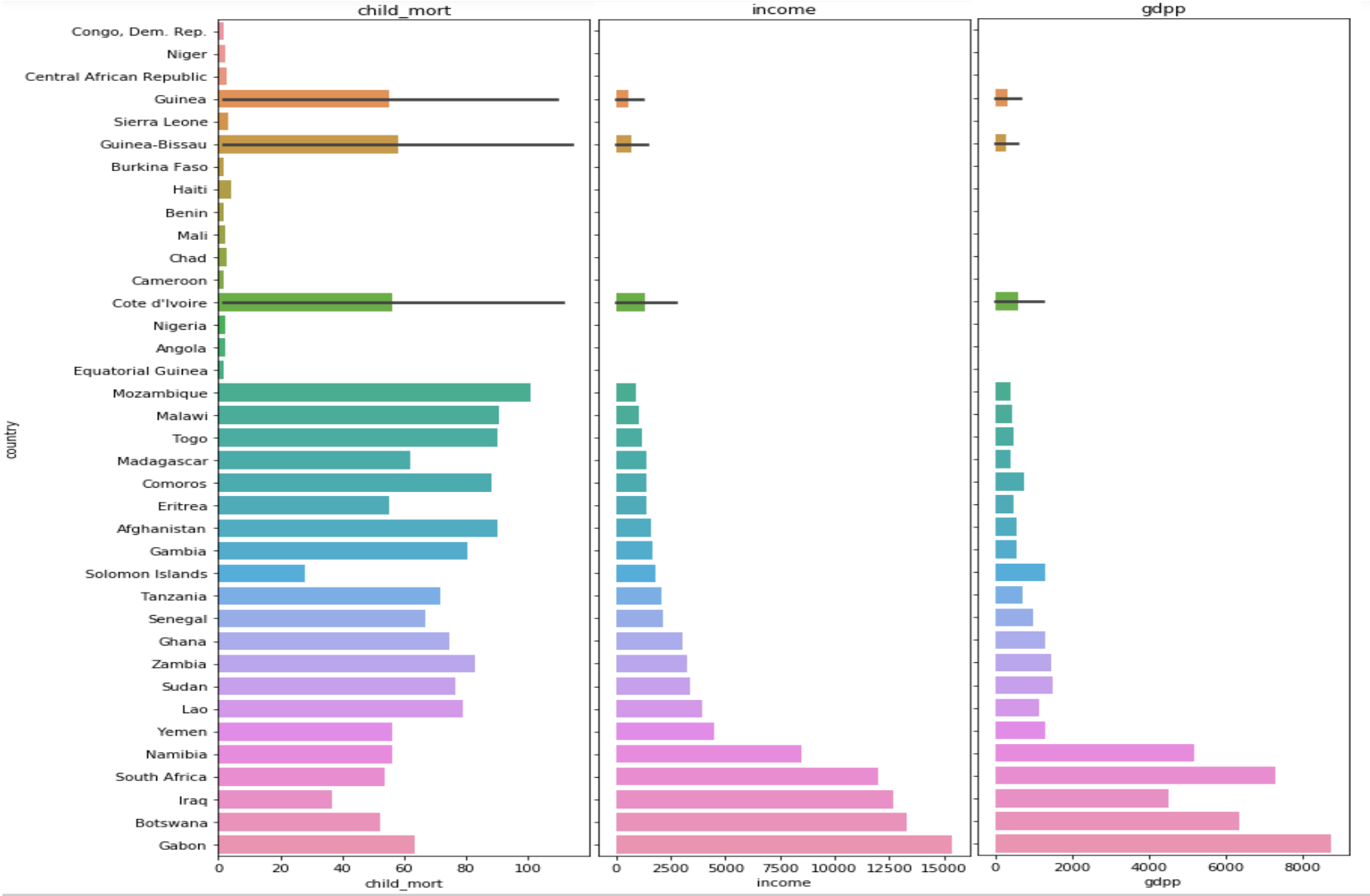






COMBINED ANALYSIS

K-MEANS & HIERARCHICAL



RECOMENDATION
COMBINED ANALYSIS

From the analysis Countries which need Attention in specific areas

- Child Mortality : Mozambique, Malawi, Togo, Guinea, Comoros
- Income: Congo, Dem. Rep., Niger, Central African Republic, Sierra Leone, Burkina Faso, Haiti
- GDPP: Congo, Dem. Rep., Niger, Central African Republic, Sierra Leone, Burkina Faso, Haiti

20 Countries having overall low Child Mortality, income and GDPP.

Congo Dem. Rep., Niger, Central African, Guinea, Sierra Leone, Guinea-Bissau, Burkina Faso, Haiti, Benin, Mali, Chad, Cameroon, Cote d'Ivoire, Nigeria, Angola, Equatorial Guinea, Mozambique, Malawi, Guinea