Clustering of countries

By: Pramodini V. Nayak

Abstract

Objective:

We, HELP International humanitarian NGO, committed to fight poverty and provide the people of backward countries with basic amenities and relief during the time of disasters and natural calamities. We run a lot of operational projects from time to time, along with advocacy, drives to raise awareness as well as for funding purposes.

Problem statement:

During the recent funding programmes, we have been able to raise around $ 10 million. As an analyst, we have to come up with the countries list that are in the direst need of aid.

Analysis methodology

Data collection and cleaning

- Import the data
- Identifying the data quality issues and clean the data

Outlier analysis and removal

- Removing the outlier where ever required as per understanding the problem statement.

Visualizing the data

- Visualizing few original data variables to look for any pattern or correlation.

Scaling the data

- Standardizing all the continuous variables.

PCA on the data

- To derive principal components
- To check the variance ratios
- Screeplot - plotting the cumulative variance against the number of components
- Going ahead and doing dimensionality reduction using incremental PCA
- Reducing the correlation to almost zero

Hopkins Statistics

- To check if data has tendency to form clusters

Analysis methodology Cont…

K means clustering

- Identify the ‘k’ by silhouette analysis and sum of squared distances graph.
- Forming n – clusters on PCA modified data
- Visualizing the clusters with various variables
- Analyzing the clusters
- Identifying the countries which requires aid.

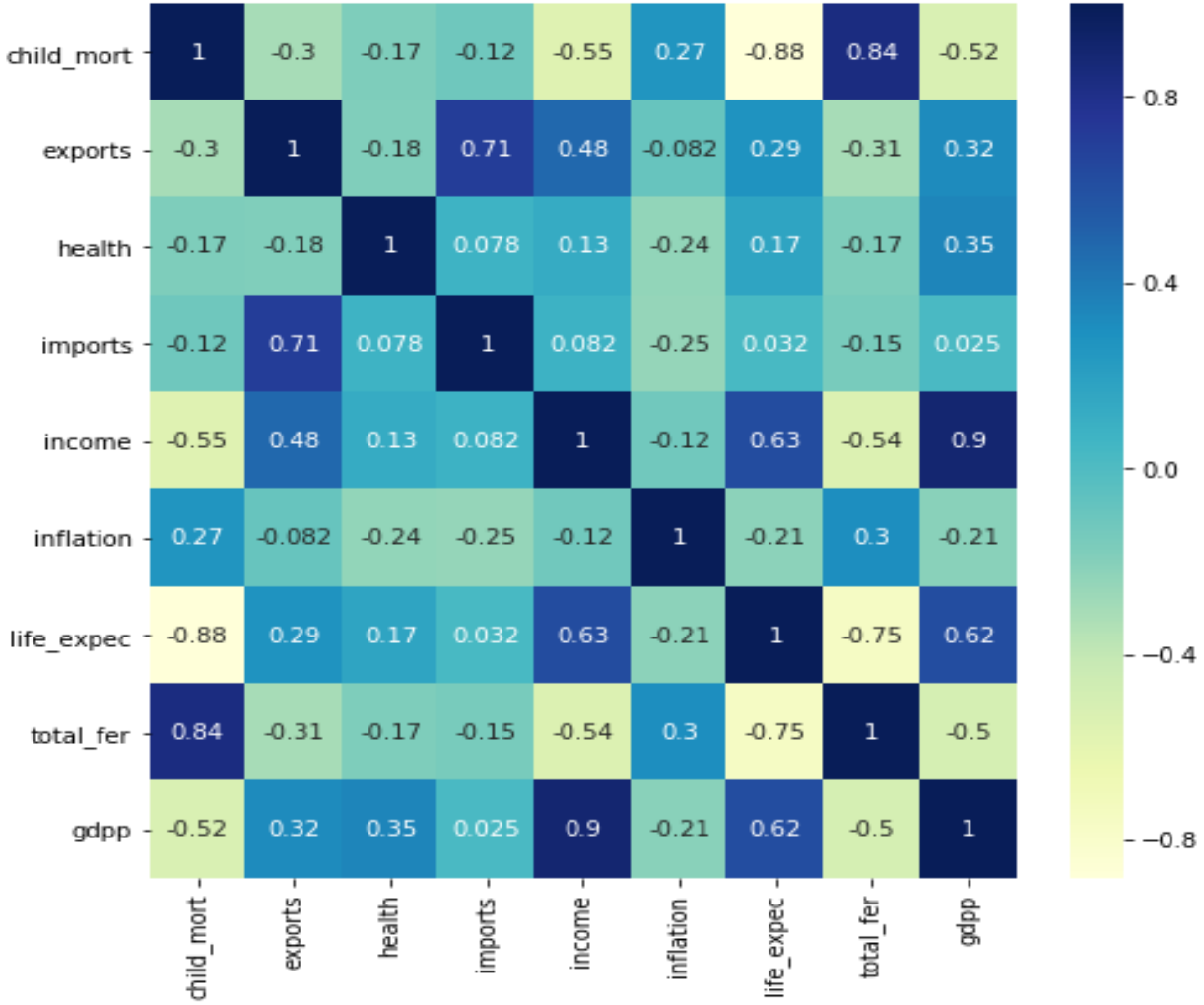
Hierarchical Clustering

- Identify the ‘n’ via dendrogram.
- Forming n – clusters on PCA modified data
- Visualizing the clusters with various variables
- Analyzing the clusters
- Identifying the countries which requires aid.

Decision Making

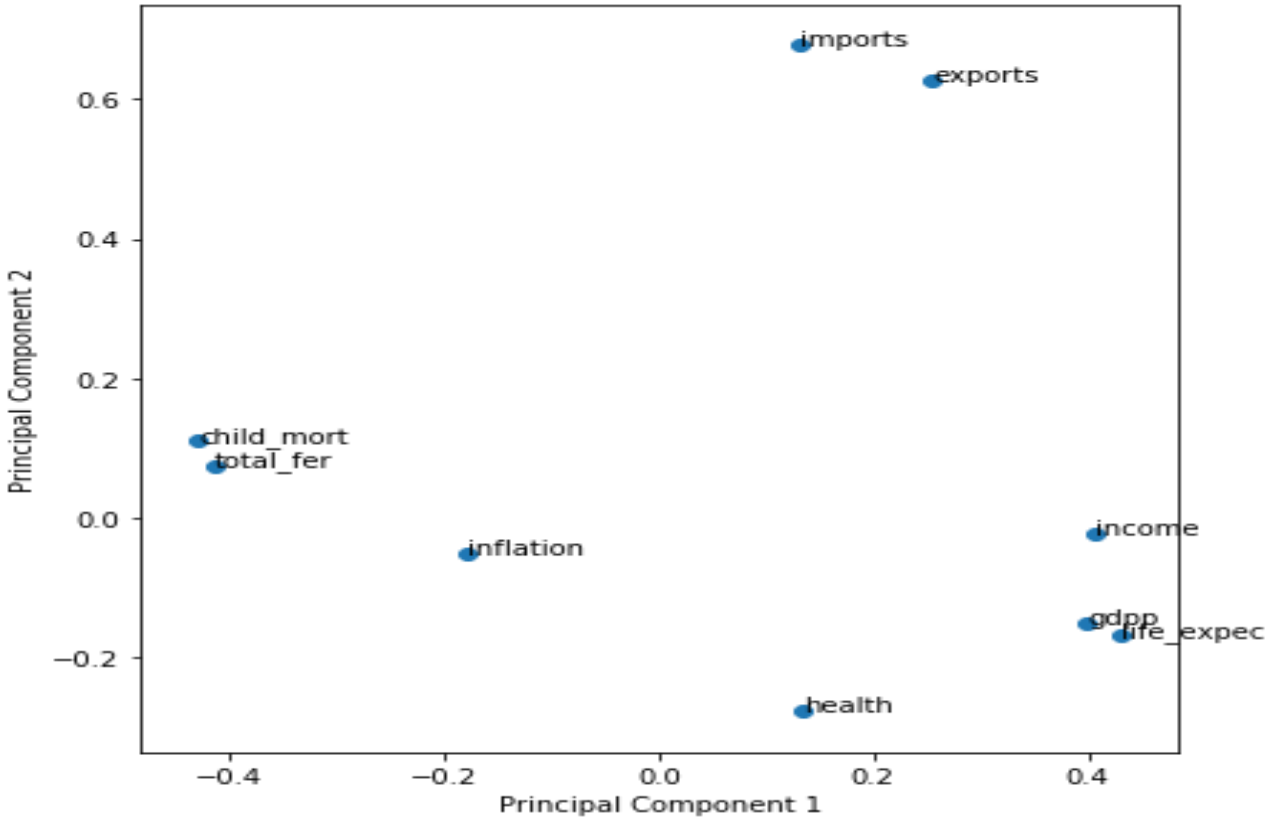
- Identifying the countries which requires aid by analyzing both K-means and Hierarchical Clustering results.

Correlation in the data:



- After data cleaning , we removed outlier from gdpp column because the country with high gdpp would not require any aid as there are already doing good.
- We did standardized scaling to standardize all parameters on cleaned, outlier removed data.
- Looking at the heatmap, we see that few variables like (total fertility, child mortality) , (income , gdpp) and (imports and exports) have high correlation.

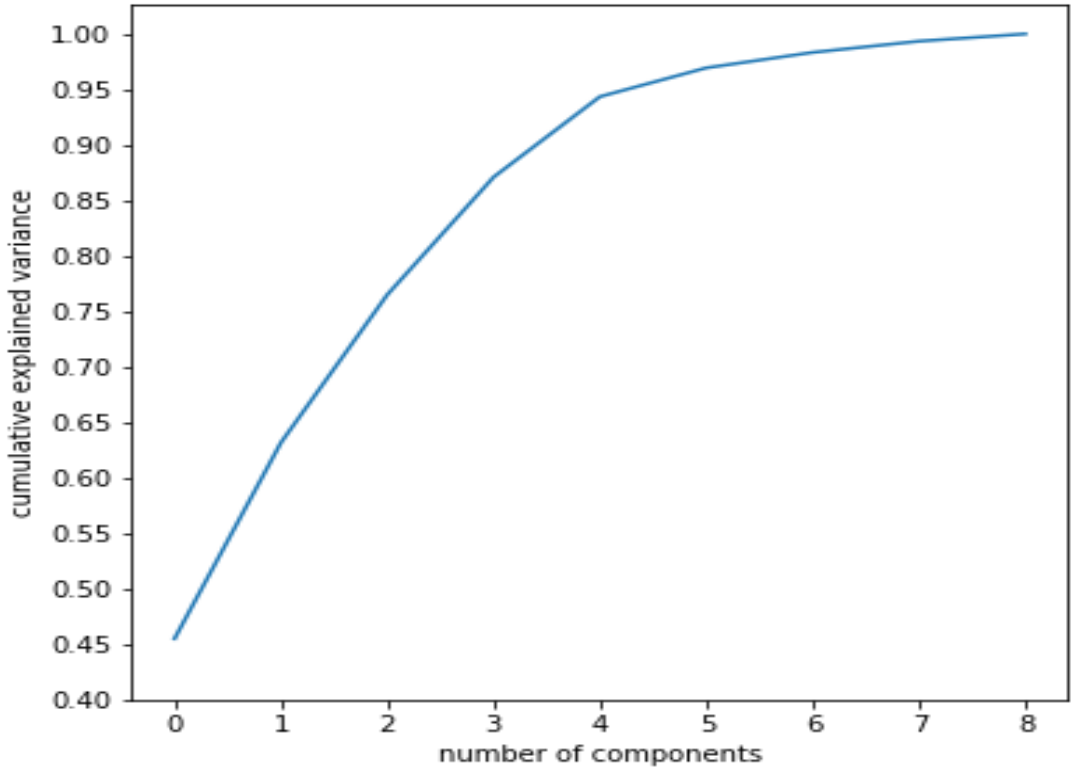
Principal Component Analysis



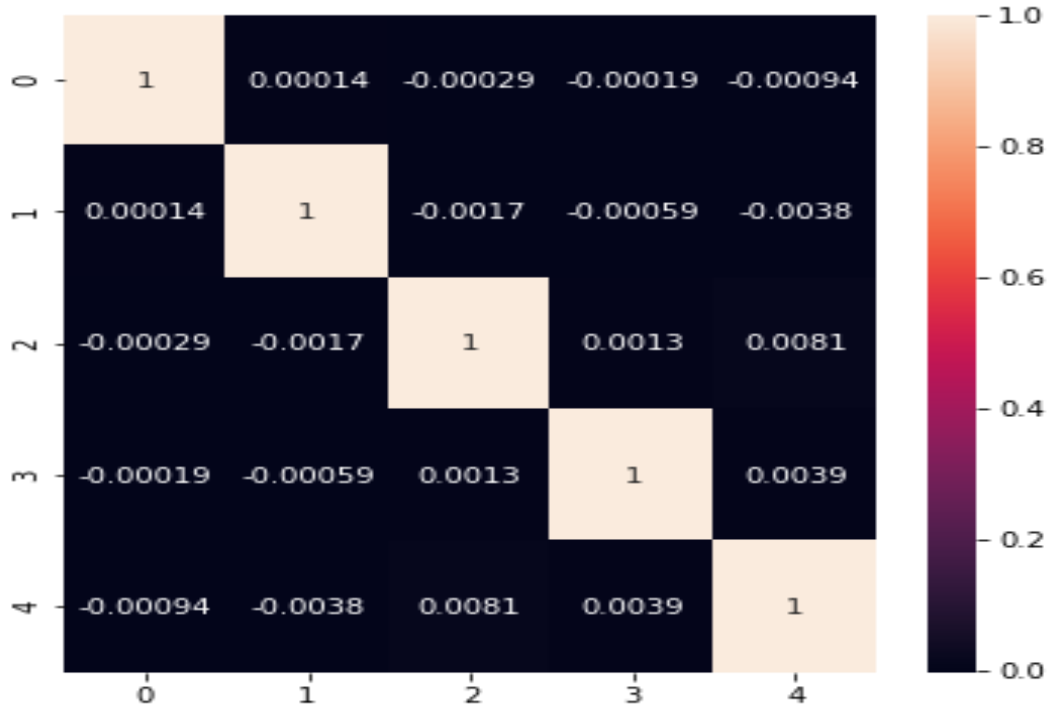
Visualising the features by loaded along PC1 and PC2

We see that features like gdpp, life expectancy and income are along the direction of PC1 and other features like total fertility and child mortality are along PC2 direction.

Principal Component Analysis

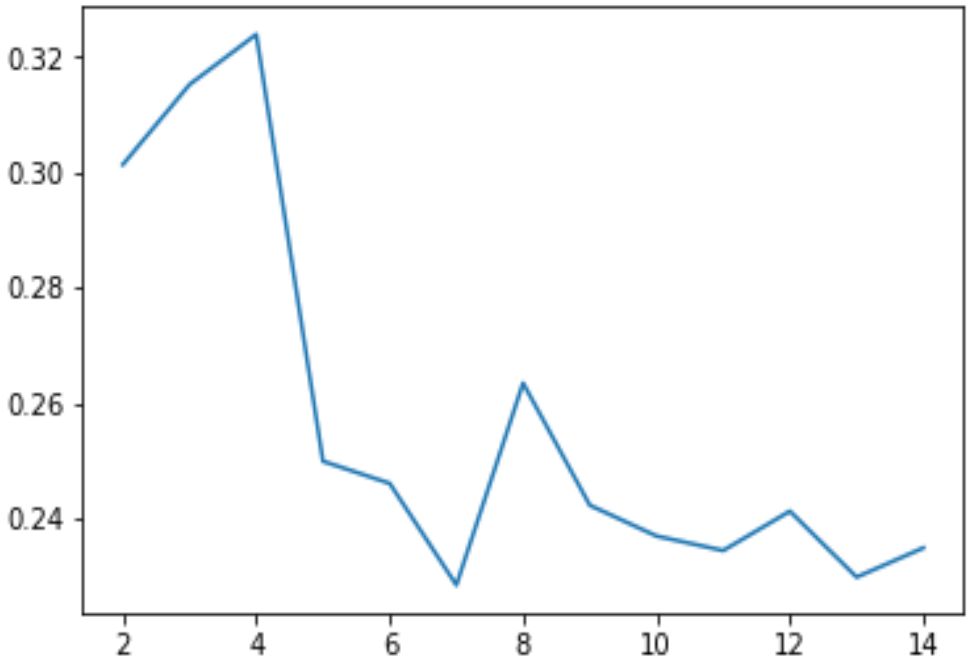


Looking at the **screeplot** , we see that 95% of variance in the data is explained by 5 components.

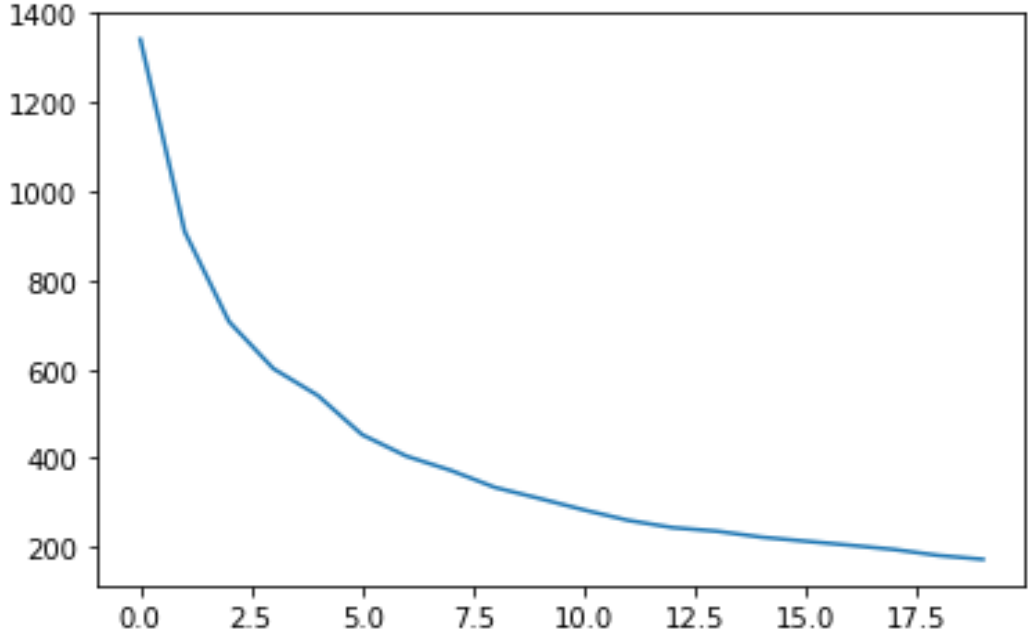


After doing dimensionality reduction via incremental PCA by taking 5 components, we see that the correlation in the data has almost reduced to zero.

K-means clustering



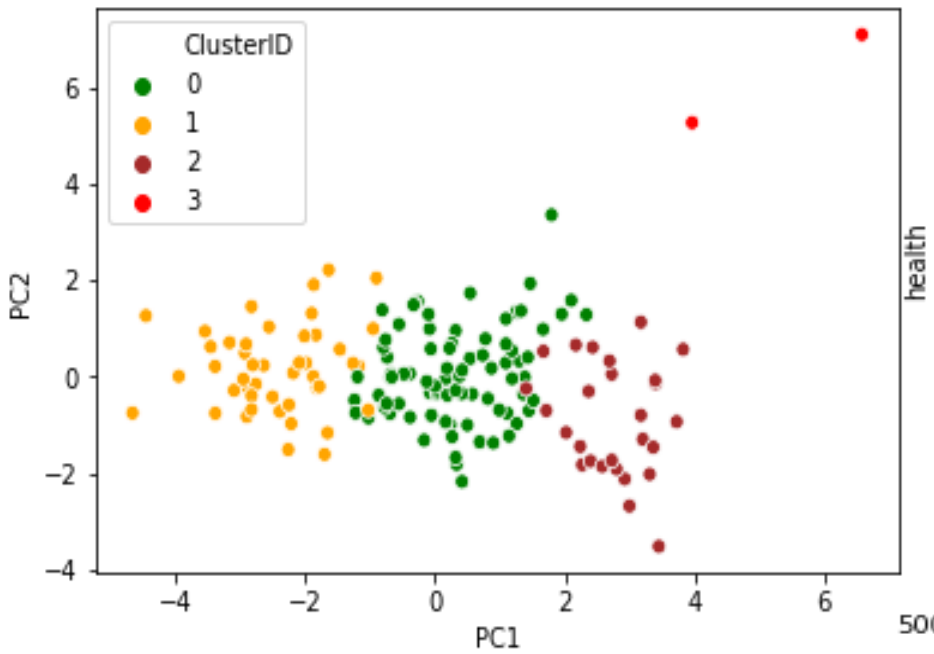
Silhouette Analysis



Sum of Squared Distances

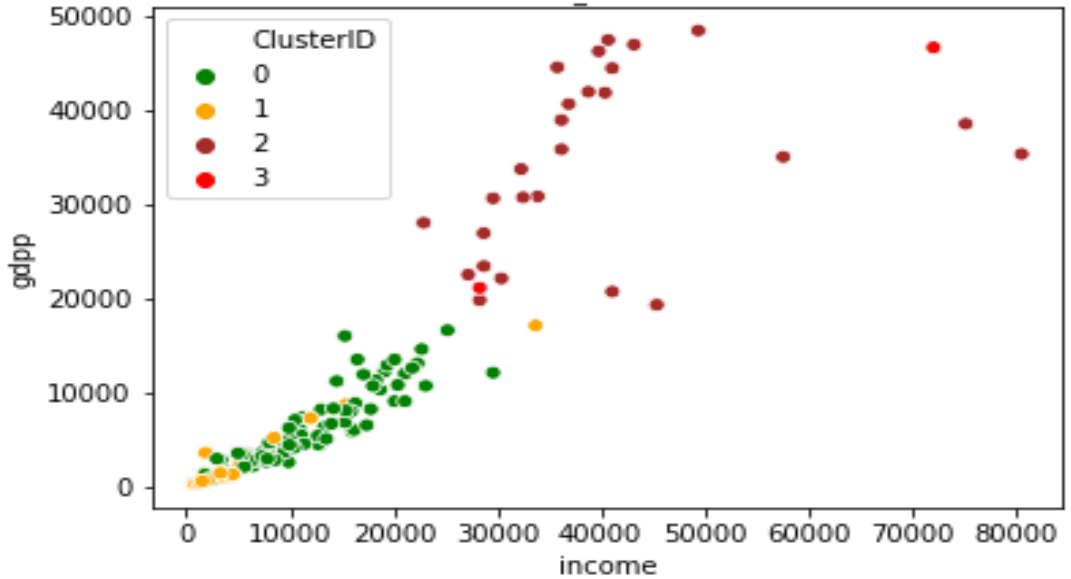
By looking silhouette analysis, we see the highest peak is at k =4 and in sum of squared distances graph , we see that the elbow is in the range of 3 to 5 , so we are going ahead with k as 4.

K-means clustering



Scatter plot of PC1 , PC2 for various clusters. We see the formation of the cluster.

Scatter plot of health spending , child mortality for various clusters. We see that for cluster 1, the health spending as % of gdp is lower and at the same time child mortality is very high.

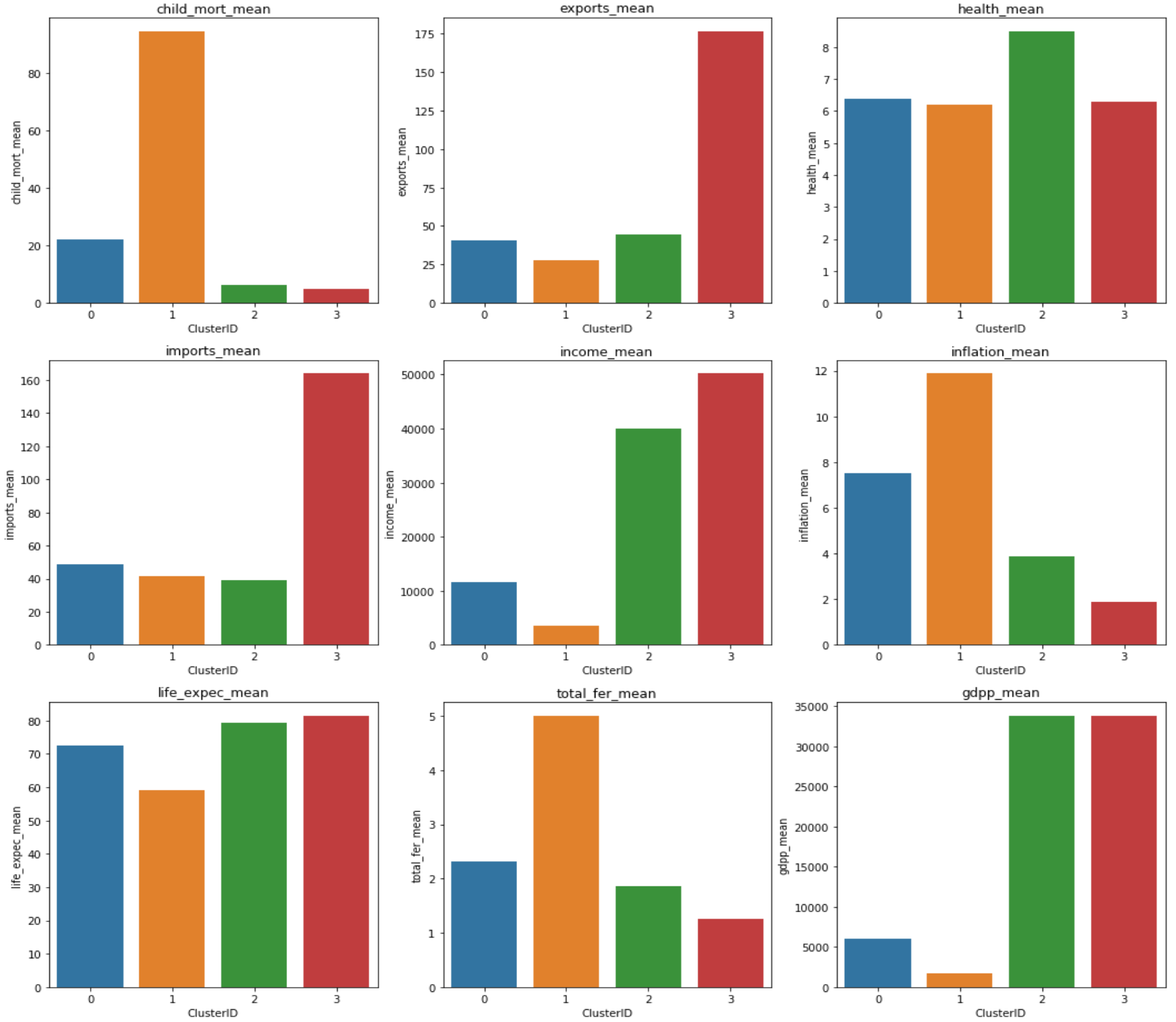


Scatter plot of gdpp , income for various clusters. We see that for cluster 1 , both gdpp and net income per person are very low.

K-means clustering

As per our K- means clusters- Cluster 1 is area of concern due to :

- Low gdpp
- Low income
- High child mortality
- High inflation
- High total fertility

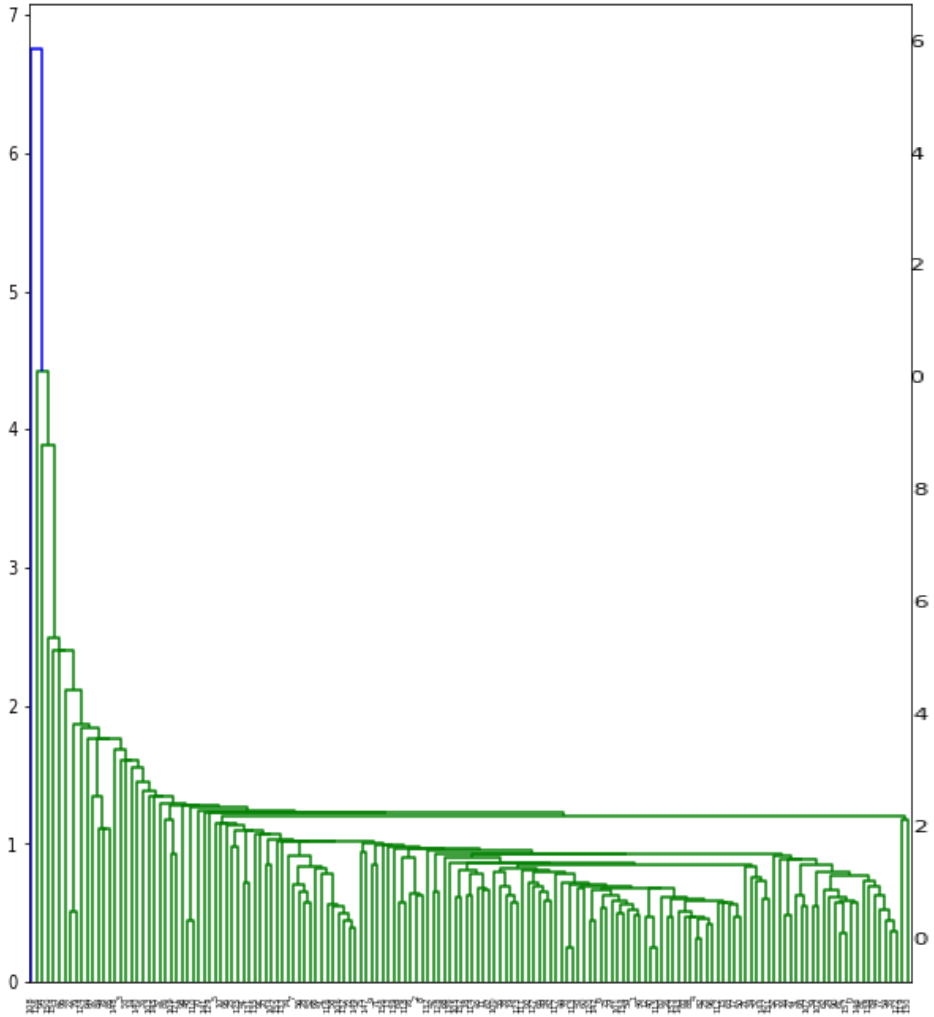


K-means clustering

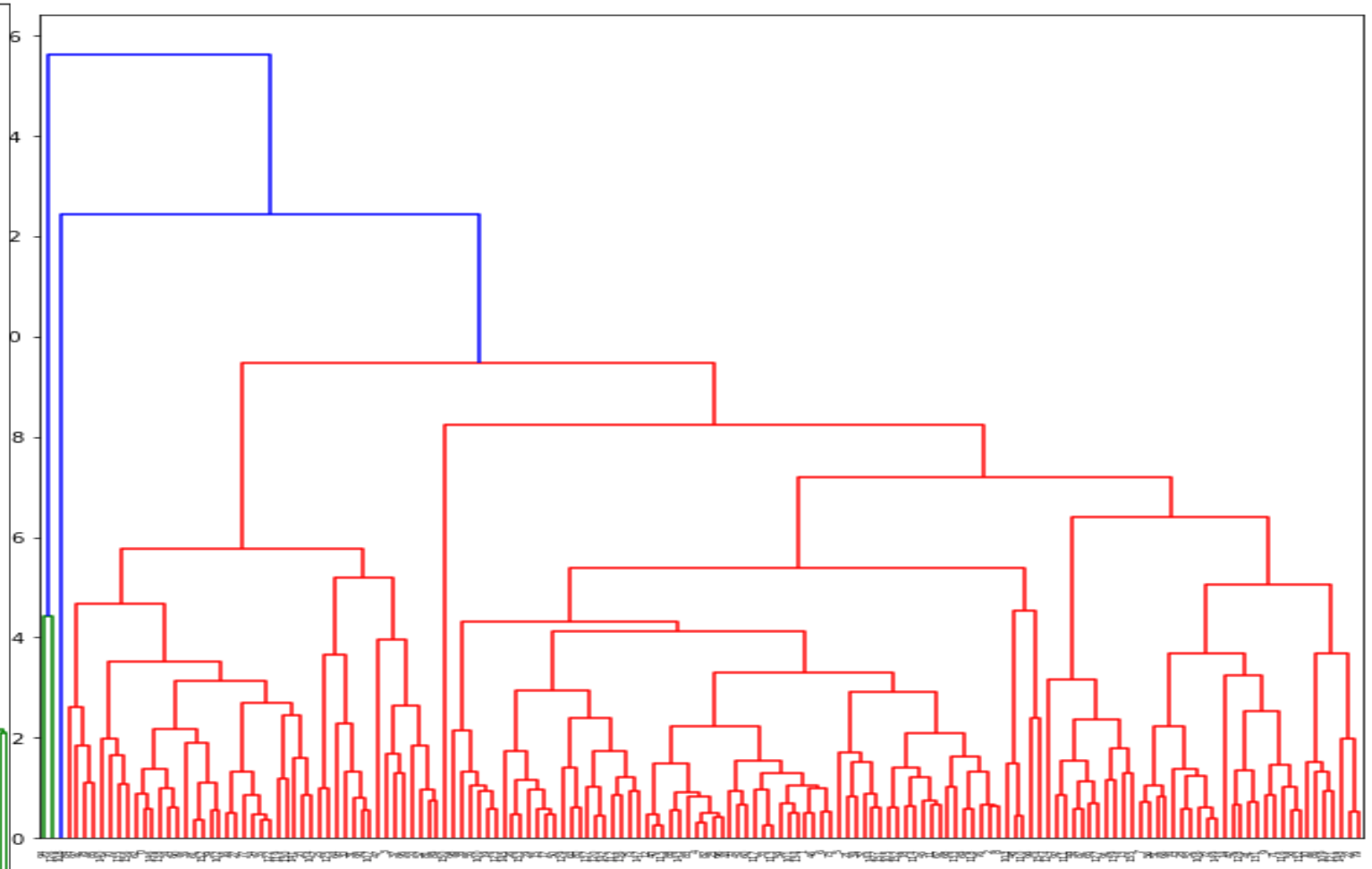
10 countries under cluster 1 are:

country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	ClusterID
Burundi	93.6	8.92	11.60	39.2	764	12.30	57.7	6.26	231	1
Liberia	89.3	19.10	11.80	92.6	700	5.47	60.8	5.02	327	1
Congo, Dem. Rep.	116.0	41.10	7.91	49.6	609	20.80	57.5	6.54	334	1
Niger	123.0	22.20	5.16	49.1	814	2.55	58.8	7.49	348	1
Sierra Leone	160.0	16.80	13.10	34.5	1220	17.20	55.0	5.20	399	1
Madagascar	62.2	25.00	3.77	43.0	1390	8.79	60.8	4.60	413	1
Mozambique	101.0	31.50	5.21	46.2	918	7.64	54.5	5.56	419	1
Central African Republic	149.0	11.80	3.98	26.5	888	2.01	47.5	5.21	446	1
Malawi	90.5	22.80	6.59	34.9	1030	12.10	53.1	5.31	459	1
Eritrea	55.2	4.79	2.66	23.3	1420	11.60	61.7	4.61	482	1

Hierarchical Clustering

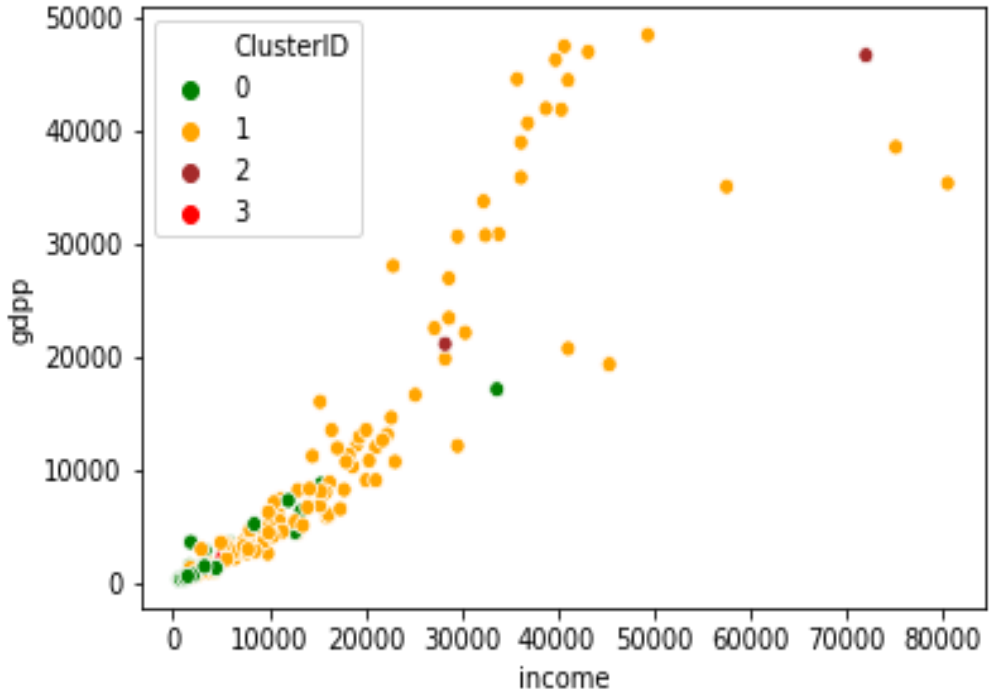


Single method hierarchical clustering

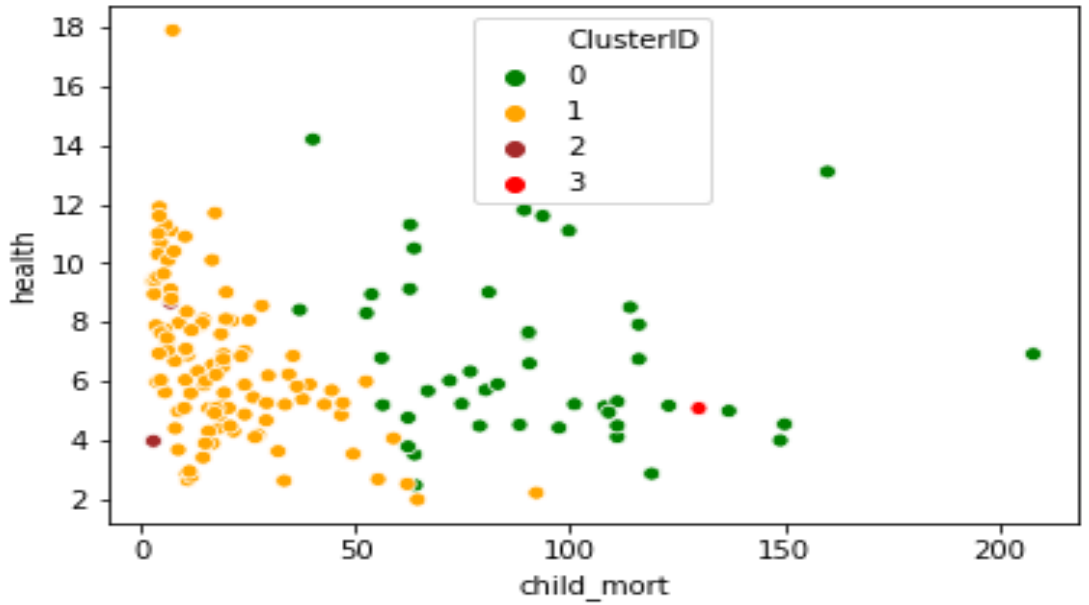


We are going for **Complete method hierarchical clustering** as single method clustering is not clear. By looking at this dendogram taking n-clusters as 4.

Hierarchical Clustering



Scatter plot of gdpp , income for various clusters. We see that for cluster 0 , both gdpp and net income per person are very low.



Scatter plot of health spending , child mortality for various clusters. We see that for cluster 0, the health spending as % of gdp of few countries is lower and for those countries -the child mortality is very high.

Hierarchical Clustering

As per our Hierarchical clusters- Cluster 0 is area of concern due to :

- Low gdpp
- Low income
- High child mortality
- High inflation
- High total fertility

10 countries under cluster 0 are:

country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	ClusterID
Burundi	93.6	8.92	11.60	39.2	764	12.30	57.7	6.26	231	0
Liberia	89.3	19.10	11.80	92.6	700	5.47	60.8	5.02	327	0
Congo, Dem. Rep.	116.0	41.10	7.91	49.6	609	20.80	57.5	6.54	334	0
Niger	123.0	22.20	5.16	49.1	814	2.55	58.8	7.49	348	0
Sierra Leone	160.0	16.80	13.10	34.5	1220	17.20	55.0	5.20	399	0
Madagascar	62.2	25.00	3.77	43.0	1390	8.79	60.8	4.60	413	0
Mozambique	101.0	31.50	5.21	46.2	918	7.64	54.5	5.56	419	0
Central African Republic	149.0	11.80	3.98	26.5	888	2.01	47.5	5.21	446	0
Malawi	90.5	22.80	6.59	34.9	1030	12.10	53.1	5.31	459	0
Togo	90.3	40.20	7.65	57.3	1210	1.18	58.7	4.87	488	0

Summary

As by both K means and Hierarchical clustering method - we have got same countries which requires aid.

The following are the countries which are in direst need of aid by considering socio – economic factor into consideration:

country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
Burundi	93.6	8.92	11.60	39.2	764	12.30	57.7	6.26	231
Liberia	89.3	19.10	11.80	92.6	700	5.47	60.8	5.02	327
Congo, Dem. Rep.	116.0	41.10	7.91	49.6	609	20.80	57.5	6.54	334
Niger	123.0	22.20	5.16	49.1	814	2.55	58.8	7.49	348
Sierra Leone	160.0	16.80	13.10	34.5	1220	17.20	55.0	5.20	399
Madagascar	62.2	25.00	3.77	43.0	1390	8.79	60.8	4.60	413
Mozambique	101.0	31.50	5.21	46.2	918	7.64	54.5	5.56	419
Central African Republic	149.0	11.80	3.98	26.5	888	2.01	47.5	5.21	446
Malawi	90.5	22.80	6.59	34.9	1030	12.10	53.1	5.31	459