Clustering of countries

Siddharth Babu

Objective:

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

Outlier analysis and removal

Visualizing the data

Scaling the data

PCA on the data

Hopkins Statistics

K means clustering

Hierarchical Clustering

Decision Making

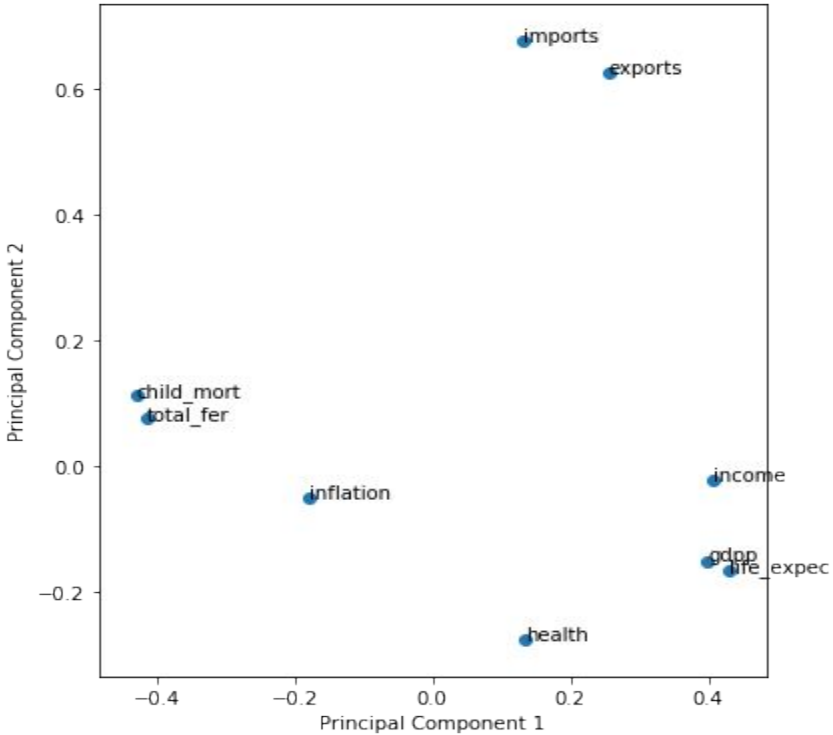
Correlation in data :

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.

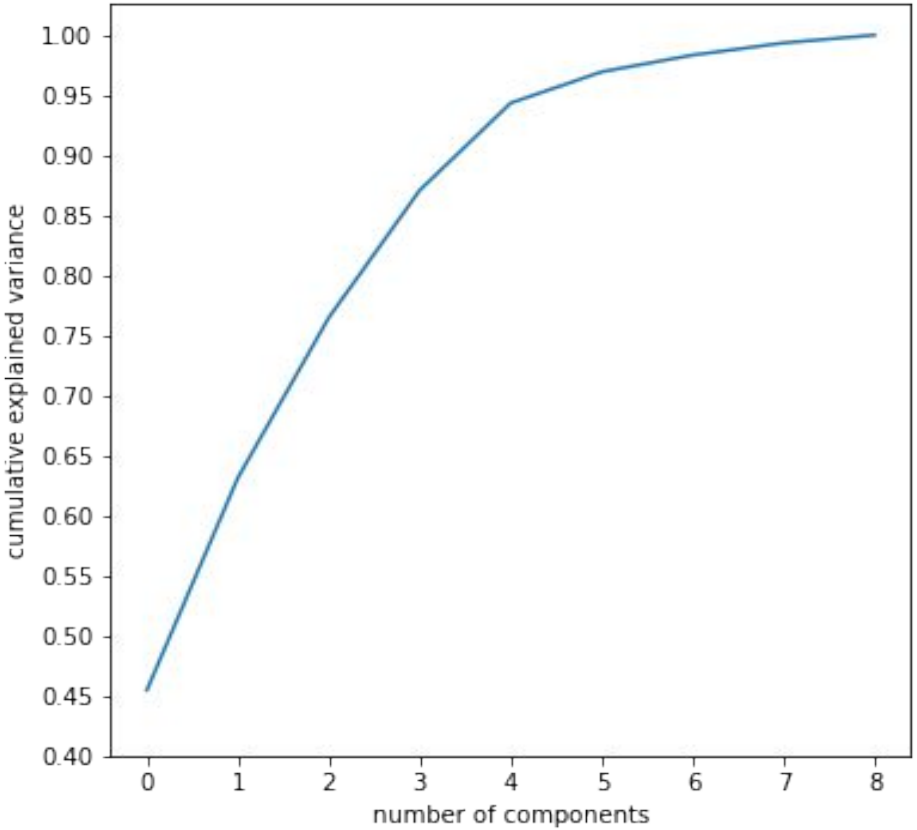
We did standardized scaling to standardize all parameters on cleaned, outlier removed data.

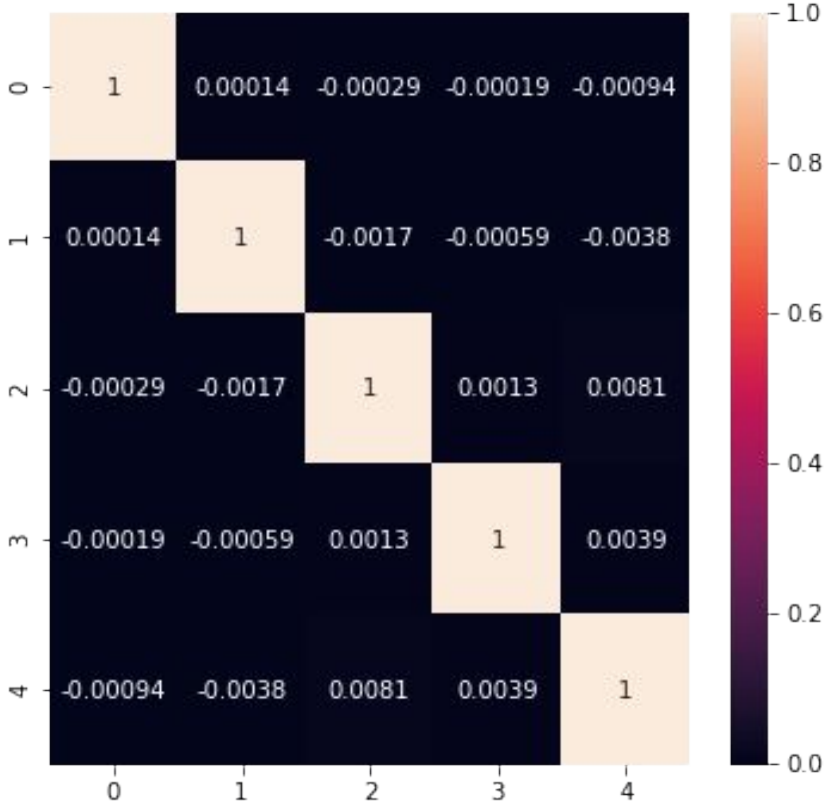
Looking at the heatmap, we see that few variables like (total fertility, child mortality) , (income , gdpp) and (imports and exports) have high correlation.

Principal component analysis

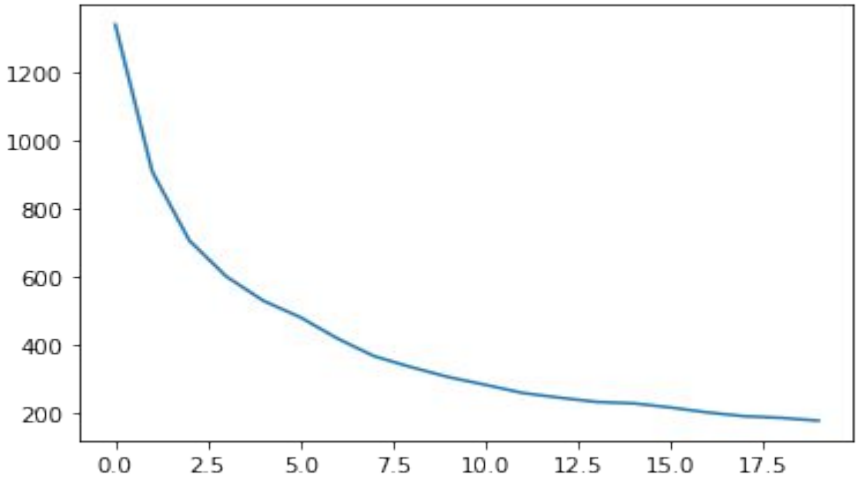


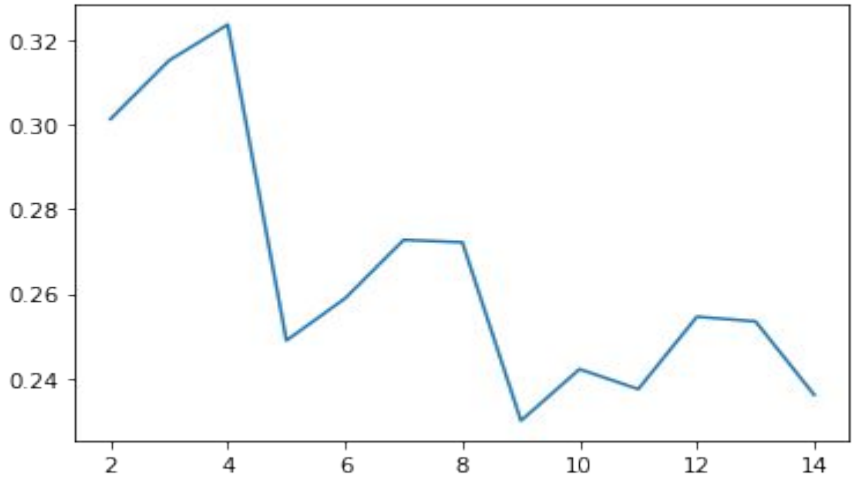
We see that features like gdpp, life expectancy and income are along the direction of PC1 and other features like total fertility and child mortality are along PC2 direction.


cumulative explained variance
number of components
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0
1
2
3
4
5
6
7
8

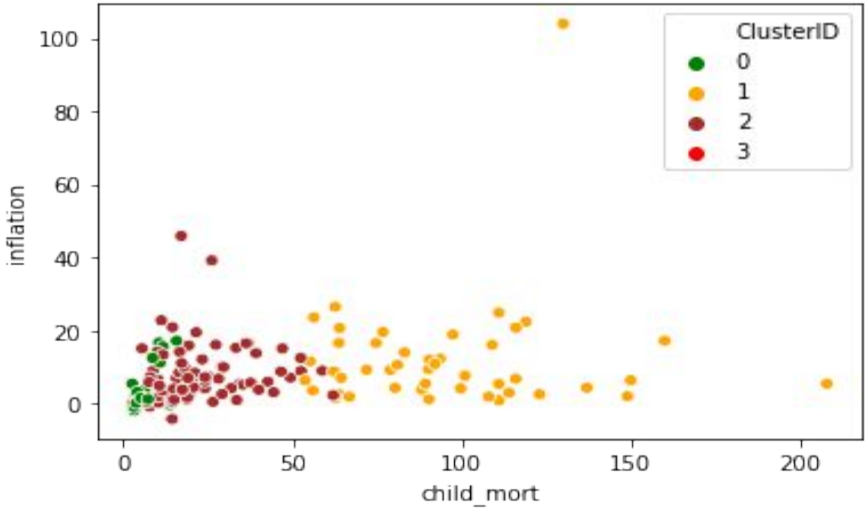

1 0.00014 -0.00029 -0.00019 -0.00094
0.00014 1 -0.0017 -0.00059 -0.0038
-0.00029 -0.0017 1 0.0013 0.0081
-0.00019 -0.00059 0.0013 1 0.0039
-0.00094 -0.0038 0.0081 0.0039 1
0
1
2
3
4
1.0
0.8
0.6
0.4
0.2
0.0

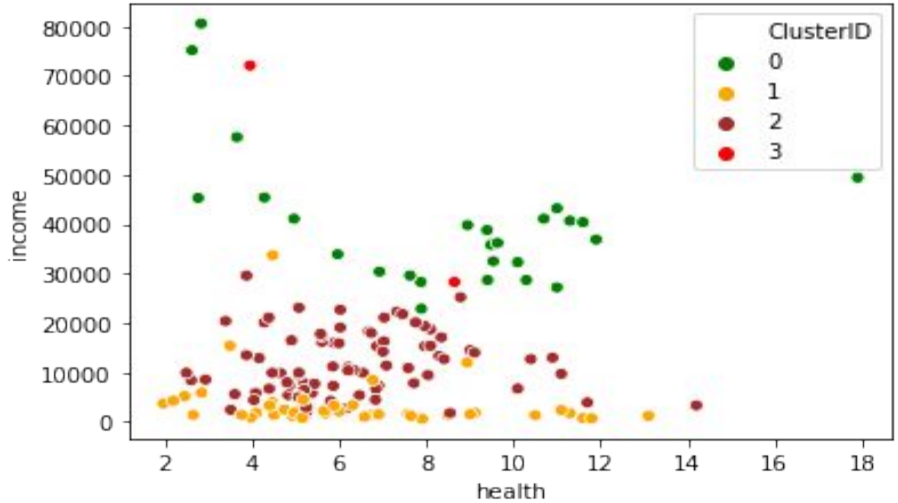
K-means clustering

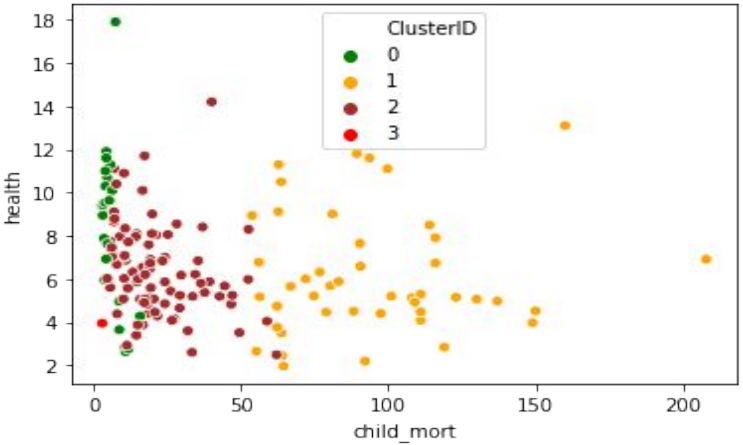


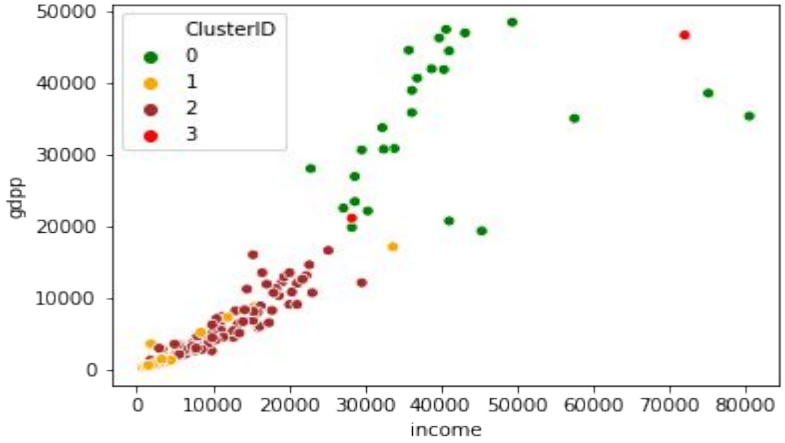


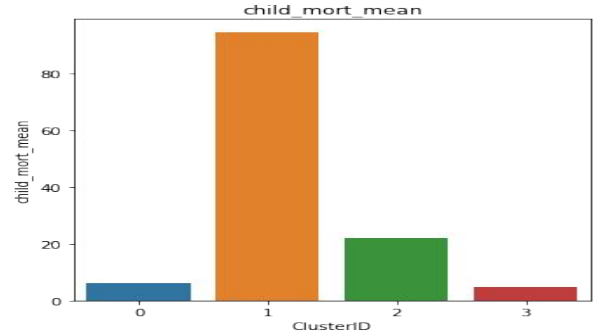
By looking silhouette analysis, we see the highest peak is at k =4 and in sum of squared distances graph , we see that the elbow is in the range of 3 to 5 , so let us take k as 4.

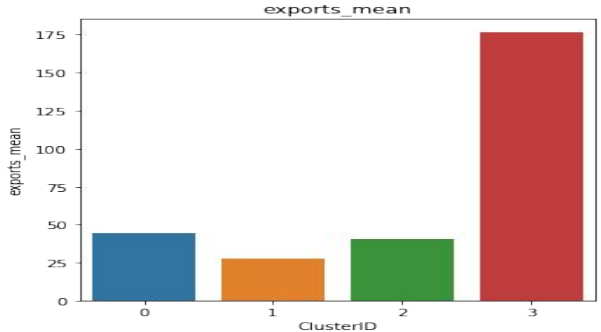

ClusterID
0
1
2
3
inflation
child_mort
0
20
40
60
80
100
0
50
100
150
200

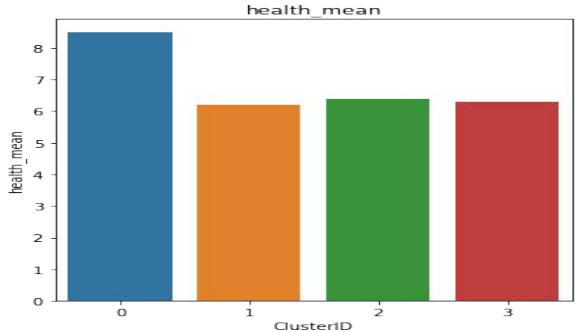

ClusterID
0
1
2
3
income
health
0
10000
20000
30000
40000
50000
60000
70000
80000
2
4
6
8
10
12
14
16
18

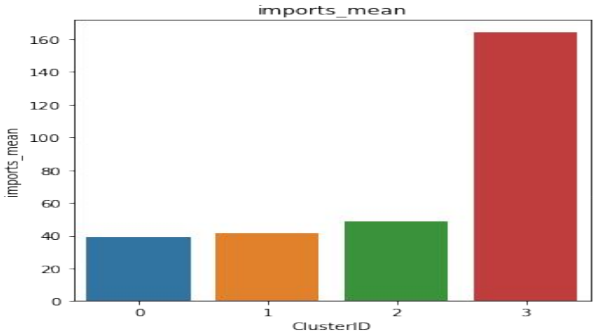

ClusterID
0
1
2
3
health
child_mort
2
4
6
8
10
12
14
16
18
0
50
100
150
200

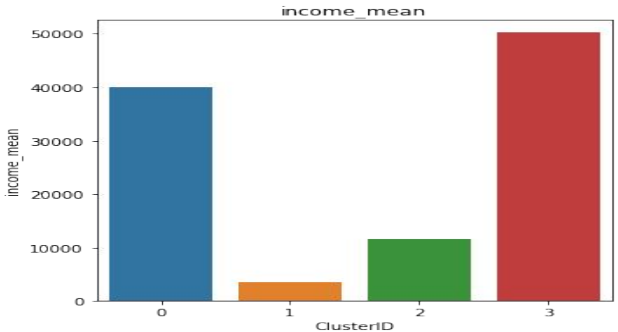

ClusterID
0
1
2
3
gdpp
income
0
10000
20000
30000
40000
50000
0
10000
20000
30000
40000
50000
60000
70000
80000

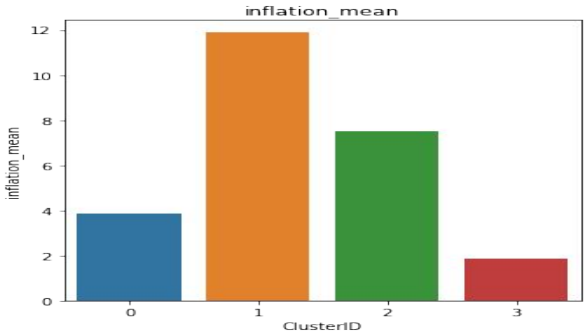

child_mort_mean
child_mort_mean
0
20
40
60
80
0
1
2
3
ClusterID

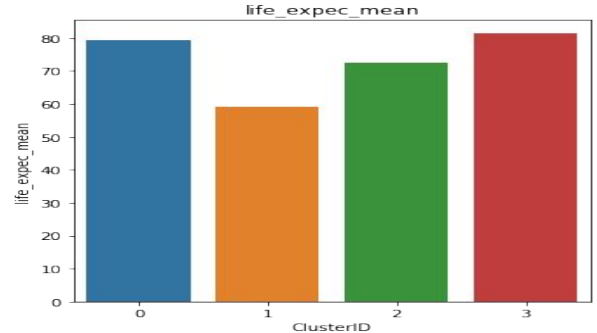

exports_mean
exports_mean
0
25
50
75
100
125
150
175
0
1
2
3
ClusterID

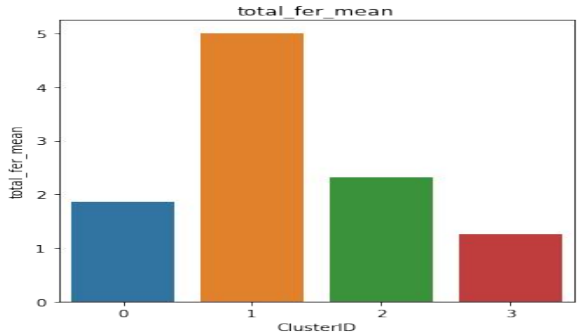

health_mean
health_mean
0
1
2
3
4
5
6
7
8
0
1
2
3
ClusterID

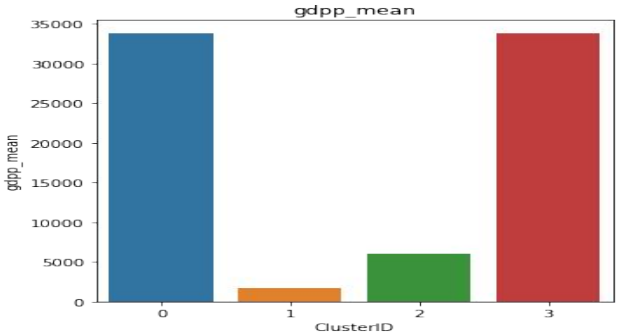

imports_mean
imports_mean
0
20
40
60
80
100
120
140
160
0
1
2
3
ClusterID


income_mean
income_mean
0
10000
20000
30000
40000
50000
0
1
2
3
ClusterID

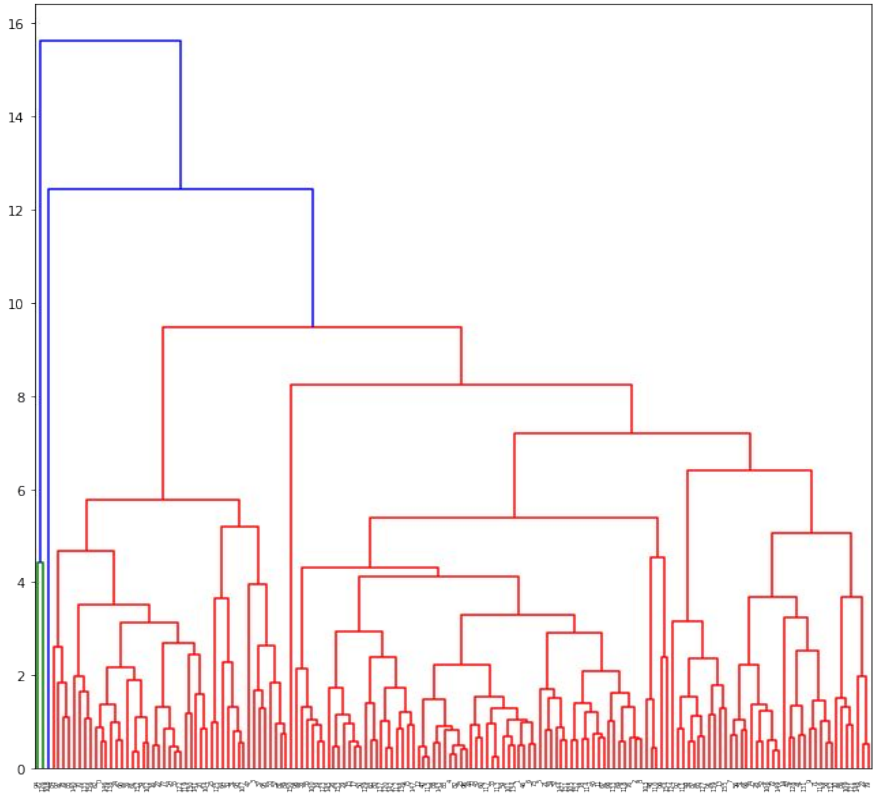

inflation_mean
inflation_mean
0
2
4
6
8
10
12
0
1
2
3
ClusterID

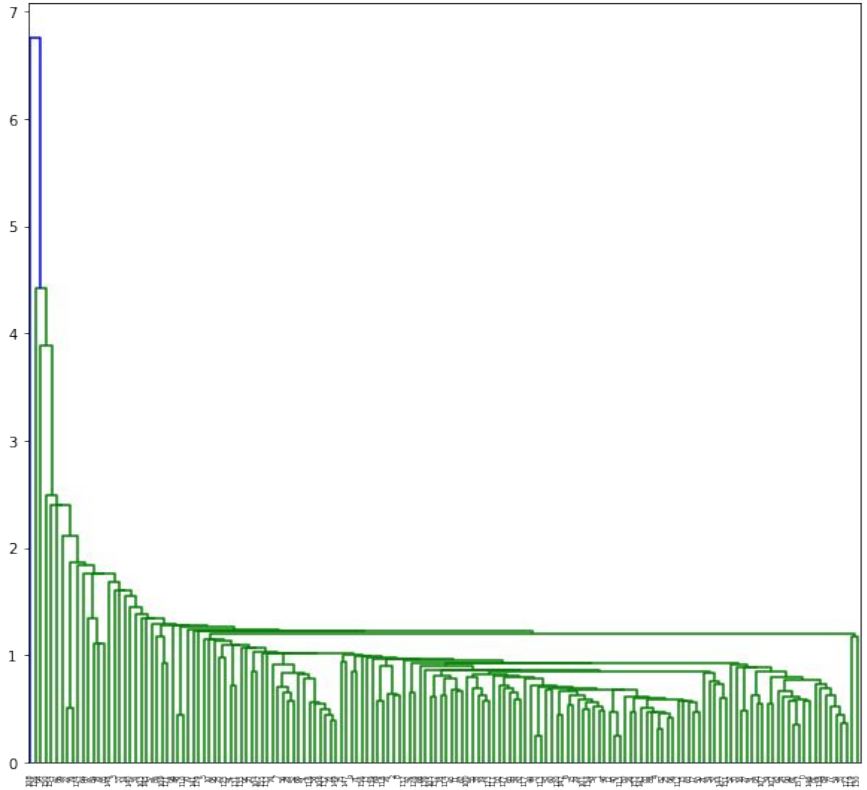

life_expec_mean
life_expec_mean
0
10
20
30
40
50
60
70
80
0
1
2
3
ClusterID


total_fer_mean
total_fer_mean
0
1
2
3
4
5
0
1
2
3
ClusterID


gdpp_mean
gdpp_mean
0
5000
10000
15000
20000
25000
30000
35000
0
1
2
3
ClusterID

Hierarchical clustering





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