CLUSTERING AND PCA ASSIGNMENT
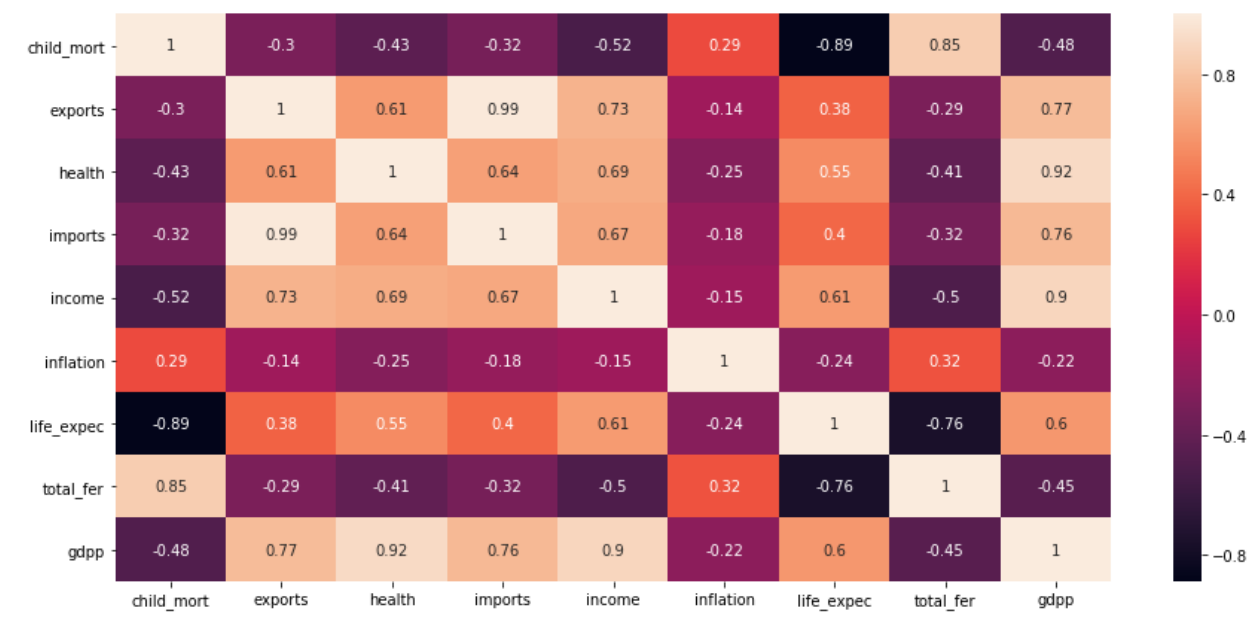
Submitted by : ANUSHKUMAR K

Problem Statement:

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 programmes, 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.
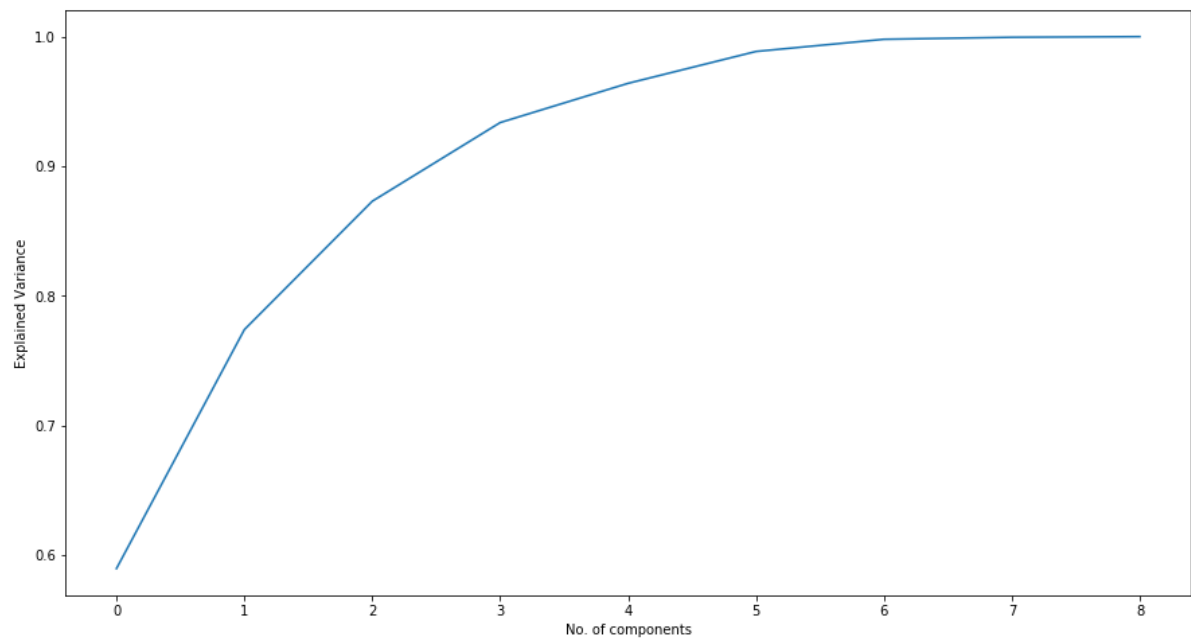
And this is where you come in as a data analyst. Your job is to categorise the countries using some socio-economic and health factors that determine the overall development of the country. Then you need to suggest the countries which the CEO needs to focus on the most. The datasets containing those socio-economic factors and the corresponding data dictionary are provided below.
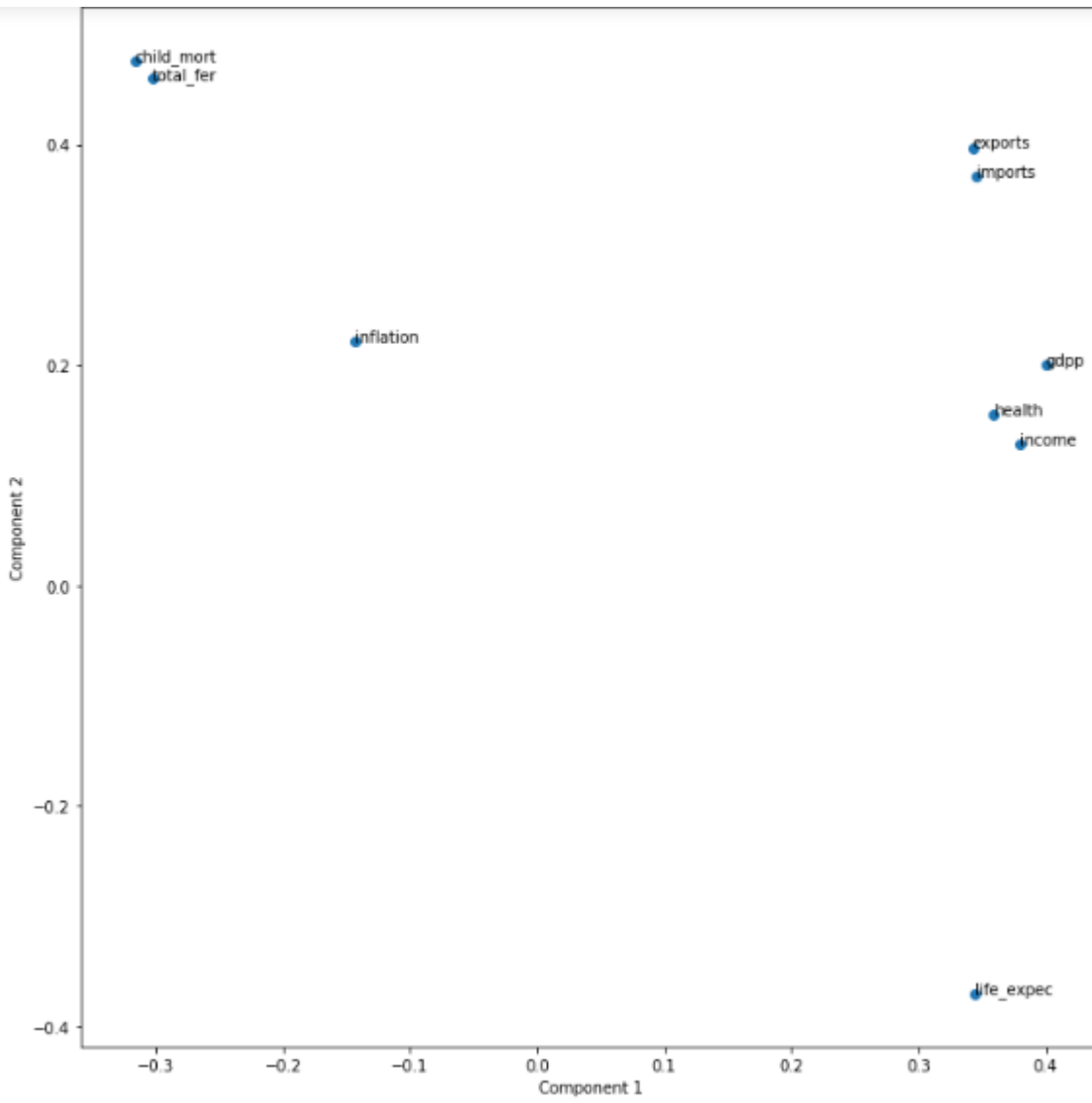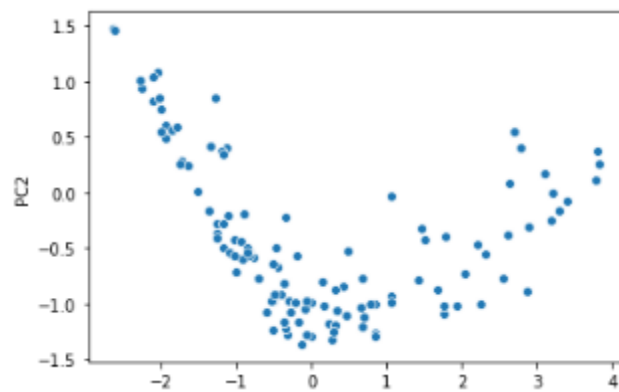
Correlation of the columns in the dataset:

	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
child_mort	1	-0.3	-0.43	-0.32	-0.52	0.29	-0.89	0.85	-0.48
exports	-0.3	1	0.61	0.99	0.73	-0.14	0.38	-0.29	0.77
health	-0.43	0.61	1	0.64	0.69	-0.25	0.55	-0.41	0.92
imports	-0.32	0.99	0.64	1	0.67	-0.18	0.4	-0.32	0.76
income	-0.52	0.73	0.69	0.67	1	-0.15	0.61	-0.5	0.9
inflation	0.29	-0.14	-0.25	-0.18	-0.15	1	-0.24	0.32	-0.22
life_expec	-0.89	0.38	0.55	0.4	0.61	-0.24	1	-0.76	0.6
total_fer	0.85	-0.29	-0.41	-0.32	-0.5	0.32	-0.76	1	-0.45
gdpp	-0.48	0.77	0.92	0.76	0.9	-0.22	0.6	-0.45	1

From the above, we could see that many variables such as income, child_mort,total_fert are correlated
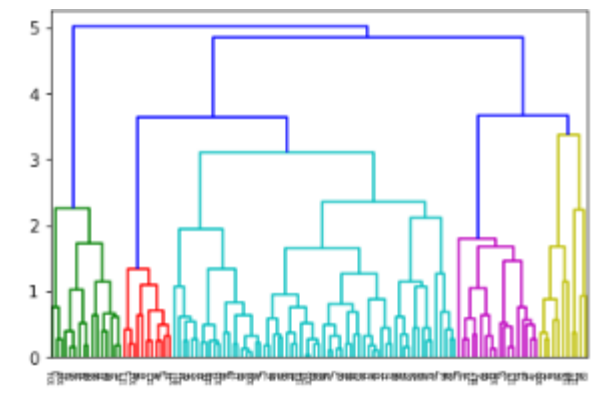
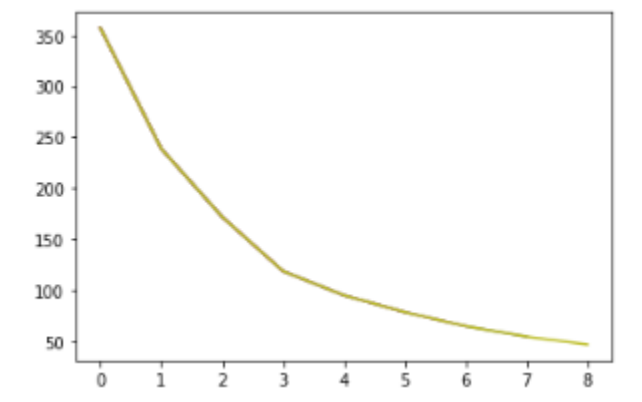## Cumulative Variance against the components:

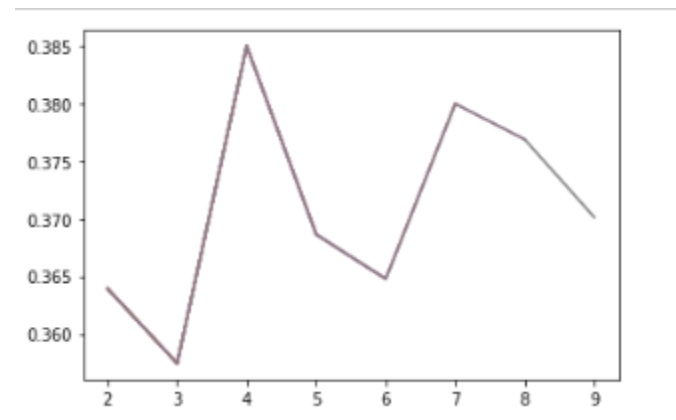**Spread of the dataset:**

Heirarchical Clustering:

Clusters are widely spread and the dendogram has split the clusters with cut tree

KMeans clustering:

From the above we can see that the clusters from 2 to 4 can be good clusters. We observe that the child_mort, inflation, gdpp, income.

Major focus should be on the below countries.

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	Afghanistan	90.2	55.3000	41.9174	248.297	1610	9.440	56.2	5.82	553
17	Benin	111.0	180.4040	31.0780	281.976	1820	0.885	61.8	5.36	758
25	Burkina Faso	116.0	110.4000	38.7550	170.200	1430	6.810	57.9	5.87	575
26	Burundi	93.6	20.6052	26.7960	90.552	764	12.300	57.7	6.26	231
28	Cameroon	108.0	290.8200	67.2030	353.700	2660	1.910	57.3	5.11	1310
31	Central African Republic	149.0	52.6280	17.7508	118.190	888	2.010	47.5	5.21	446
32	Chad	150.0	330.0960	40.6341	390.195	1930	6.390	56.5	6.59	897
36	Comoros	88.2	126.8850	34.6819	397.573	1410	3.870	65.9	4.75	769
37	Congo, Dem. Rep.	116.0	137.2740	26.4194	165.664	609	20.800	57.5	6.54	334
40	Cote d'Ivoire	111.0	617.3200	64.6600	528.260	2690	5.390	56.3	5.27	1220
56	Gambia	80.3	133.7560	31.9778	239.974	1660	4.300	65.5	5.71	562
63	Guinea	109.0	196.3440	31.9464	279.936	1190	16.100	58.0	5.34	648
64	Guinea-Bissau	114.0	81.5030	46.4950	192.544	1390	2.970	55.6	5.05	547
66	Haiti	208.0	101.2860	45.7442	428.314	1500	5.450	32.1	3.33	662
87	Lesotho	99.7	460.9800	129.8700	1181.700	2380	4.150	46.5	3.30	1170
88	Liberia	89.3	62.4570	38.5860	302.802	700	5.470	60.8	5.02	327
94	Malawi	90.5	104.6520	30.2481	160.191	1030	12.100	53.1	5.31	459
97	Mali	137.0	161.4240	35.2584	248.508	1870	4.370	59.5	6.55	708
106	Mozambique	101.0	131.9850	21.8299	193.578	918	7.640	54.5	5.56	419
112	Niger	123.0	77.2560	17.9568	170.868	814	2.550	58.8	7.49	348
132	Sierra Leone	160.0	67.0320	52.2690	137.655	1220	17.200	55.0	5.20	399
150	Togo	90.3	196.1760	37.3320	279.624	1210	1.180	58.7	4.87	488
155	Uganda	81.0	101.7450	53.6095	170.170	1540	10.600	56.8	6.15	595