HELP International NGO

Assignment on PCA and Clustering

Anupam Kumar Singh

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 programs, 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 categorize 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.

Data Preparation

The data provided contains parameters for list of countries defining:

Income,

GDP

Health Expenditure,

Child Mortality Rate,

Life Expectancy, Inflation,

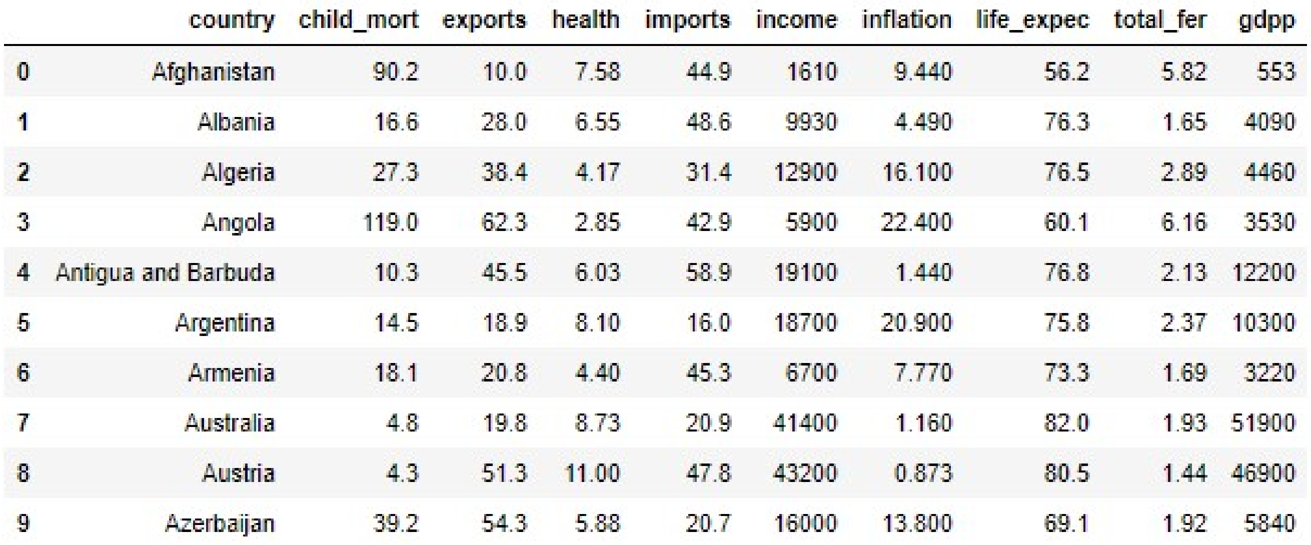
Total Fertility,

Imports and

Exports.

Some of the parameters – Health, Imports, Exports and Inflation were expressed as % of the Total GDPP.

These parameters were converted to the absolute actual values.

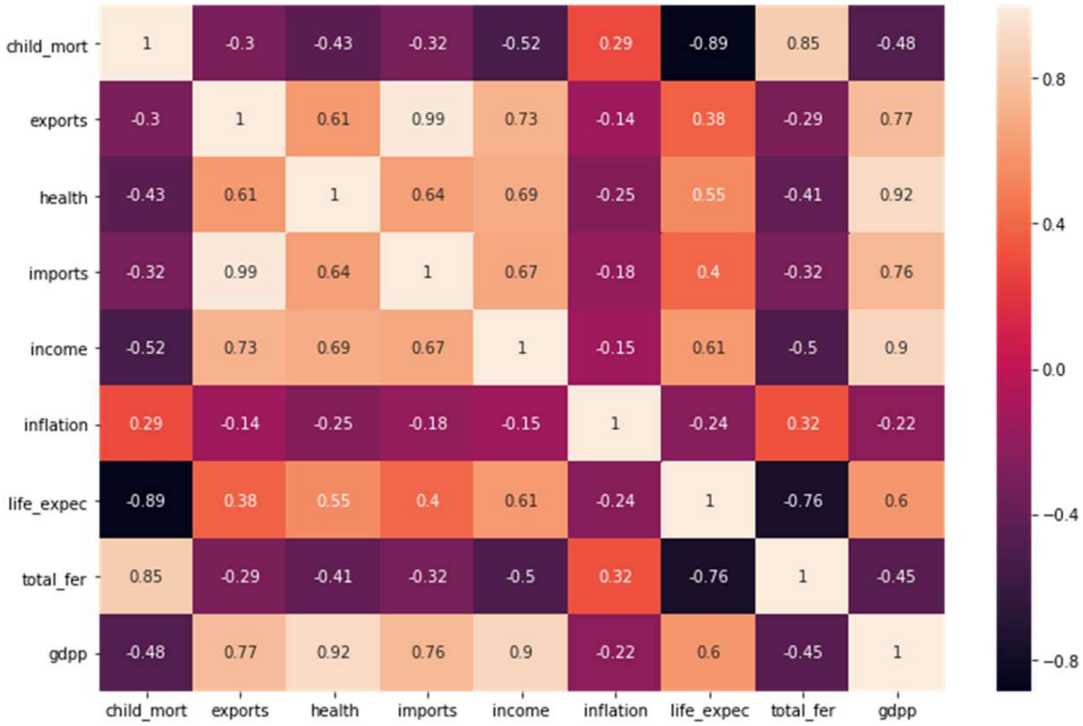


	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	Afghanistan	90.2	10.0	7.58	44.9	1610	9.440	56.2	5.82	553
1	Albania	16.6	28.0	6.55	48.6	9930	4.490	76.3	1.65	4090
2	Algeria	27.3	38.4	4.17	31.4	12900	16.100	76.5	2.89	4460
3	Angola	119.0	62.3	2.85	42.9	5900	22.400	60.1	6.16	3530
4	Antigua and Barbuda	10.3	45.5	6.03	58.9	19100	1.440	76.8	2.13	12200
5	Argentina	14.5	18.9	8.10	16.0	18700	20.900	75.8	2.37	10300
6	Armenia	18.1	20.8	4.40	45.3	6700	7.770	73.3	1.69	3220
7	Australia	4.8	19.8	8.73	20.9	41400	1.160	82.0	1.93	51900
8	Austria	4.3	51.3	11.00	47.8	43200	0.873	80.5	1.44	46900
9	Azerbaijan	39.2	54.3	5.88	20.7	16000	13.800	69.1	1.92	5840

PCA Analysis

The correlation Matrix on the original Data set reveals that there is quite a correlation between the parameters which was needed to be addressed.

Before PCA: *Image-*

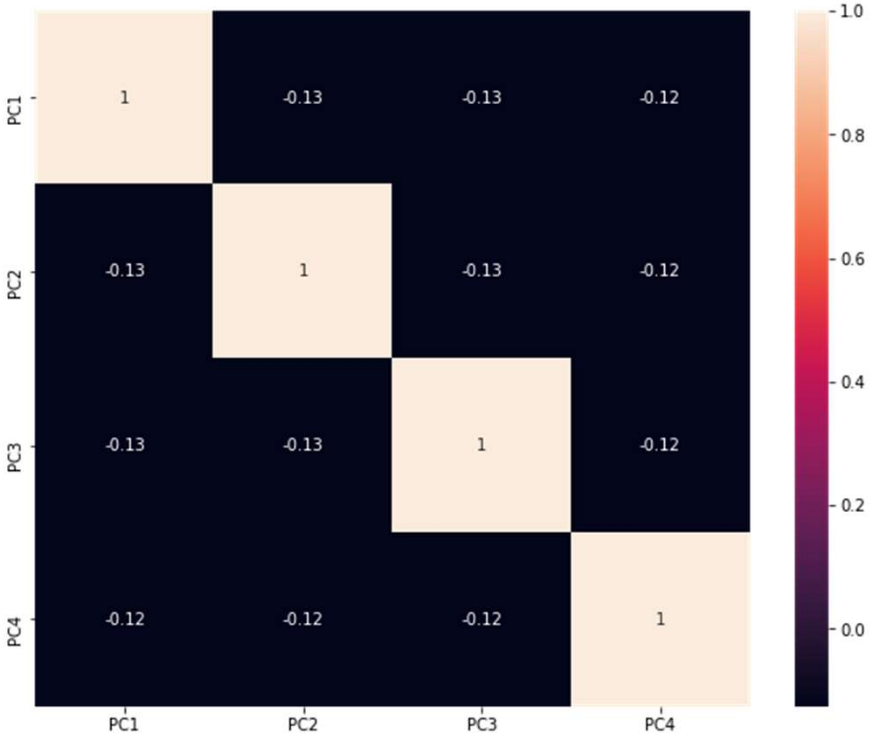


	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
child_mort	1	-0.3	-0.43	-0.32	-0.52	0.29	-0.89	0.85	-0.48
exports	-0.3	1	0.61	0.99	0.73	-0.14	0.38	-0.29	0.77
health	-0.43	0.61	1	0.64	0.69	-0.25	0.55	-0.41	0.92
imports	-0.32	0.99	0.64	1	0.67	-0.18	0.4	-0.32	0.76
income	-0.52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

PCA Analysis

The PCA process was done to reduce the data dimensionality and to take of the correlation

After PCA: *Image-*

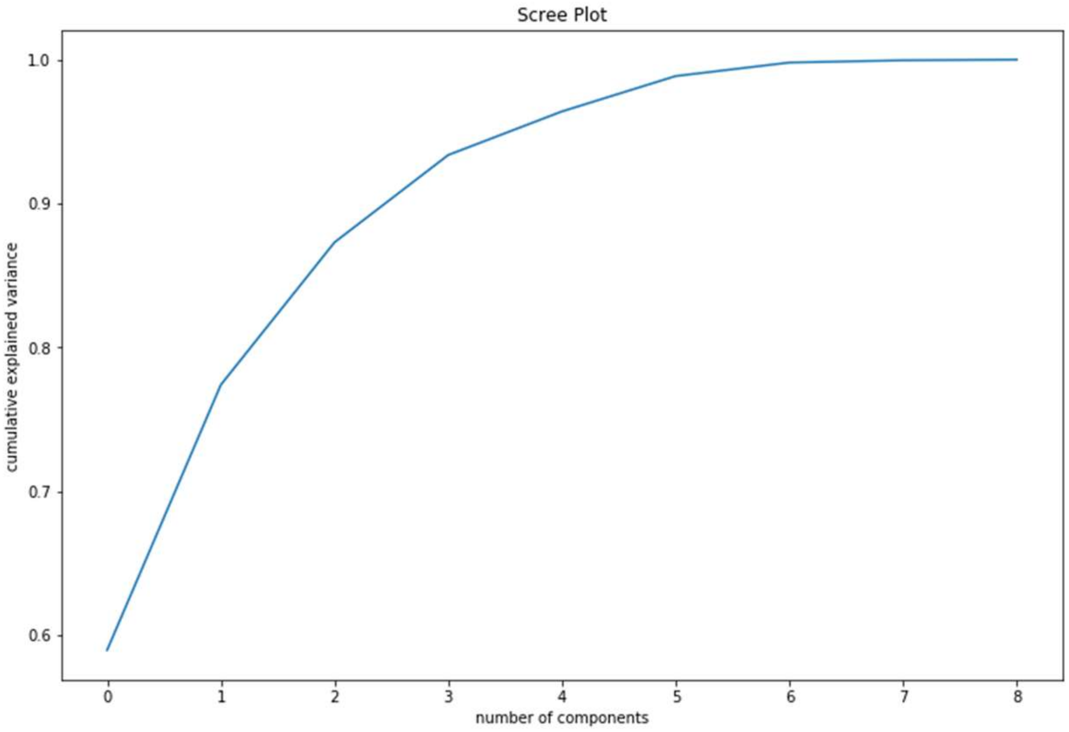


PCA Analysis

The Scree Plot was made and on analysing we can see the with just 4 components we are able to explain more than 90% of the variance.

Hence 4 Principal components were taken into consideration.

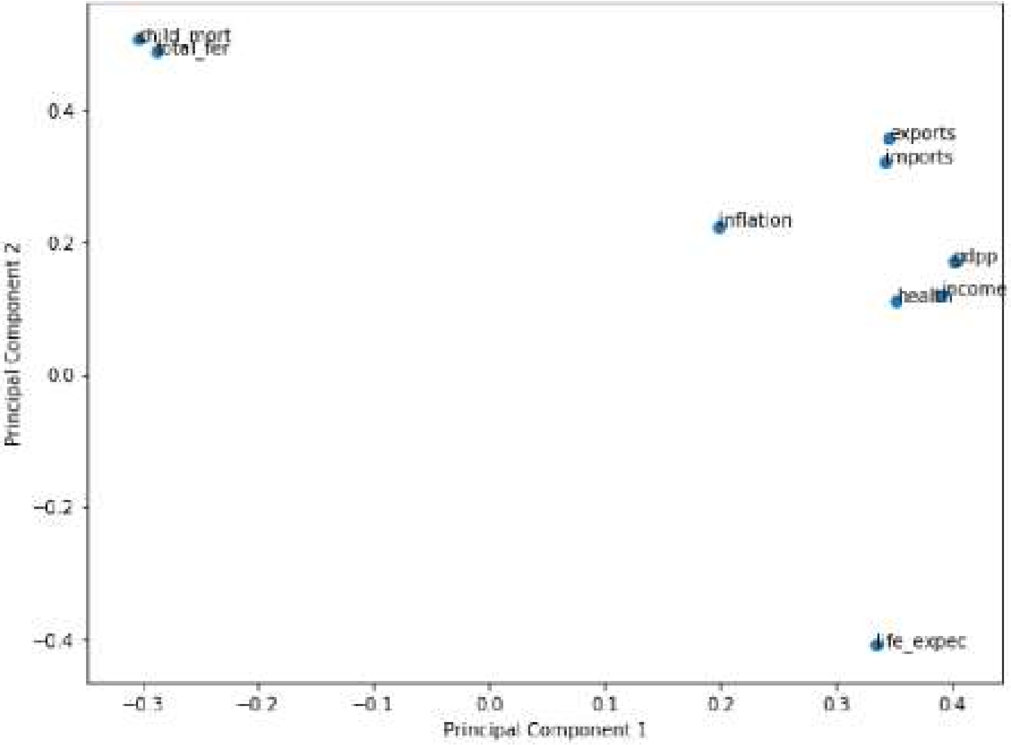
Scree Plot: *Image-*



PCA Analysis

The Data transformed into PCAs are plotted against to analyse the pattern.

The features depicting the high PC1 and PC2 are selected for further analysis.



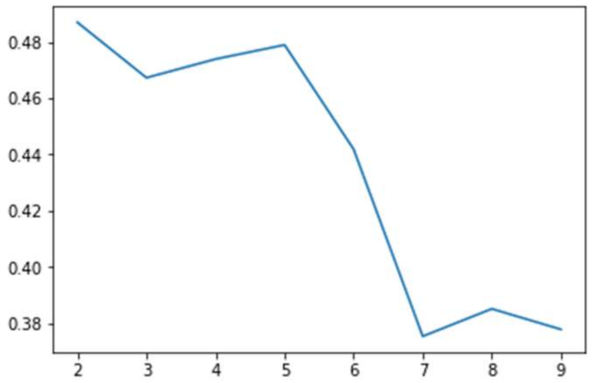
PCA1/PCA2: *Image-*

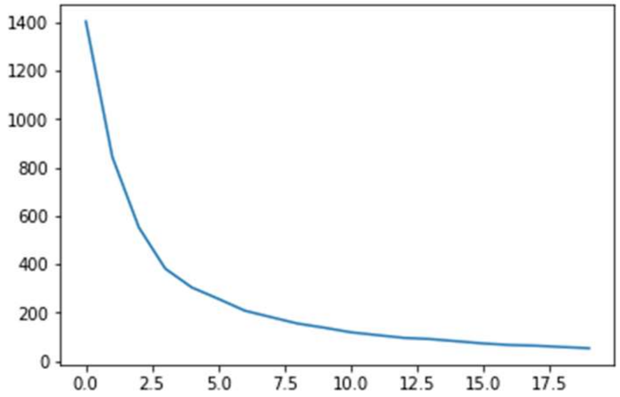
Clustering Analysis

Silhouette Score and Sum of Squared Distances were calculated and plotted to figure out the number of optimal clusters

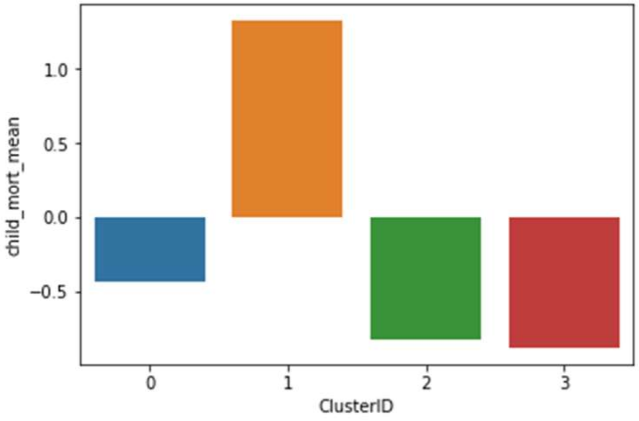
K=4, was considered and clustering was performed

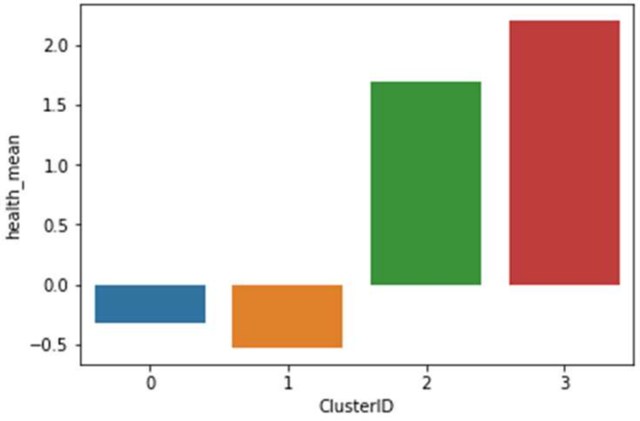
After clustering, the mean values for each clusters were calculated and grouped. The plot was also made.

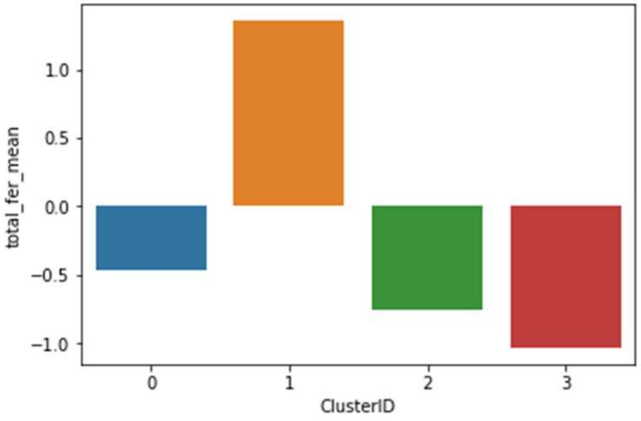


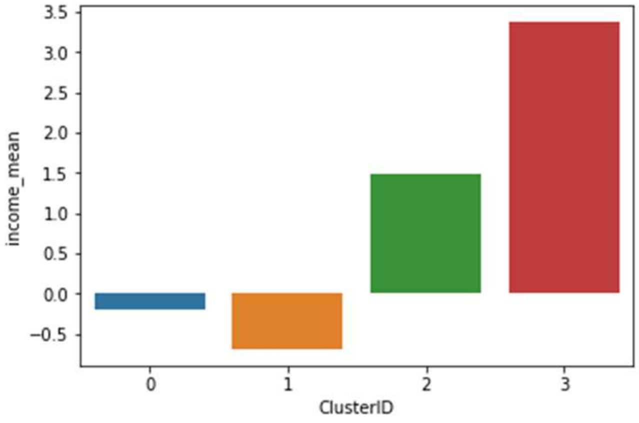


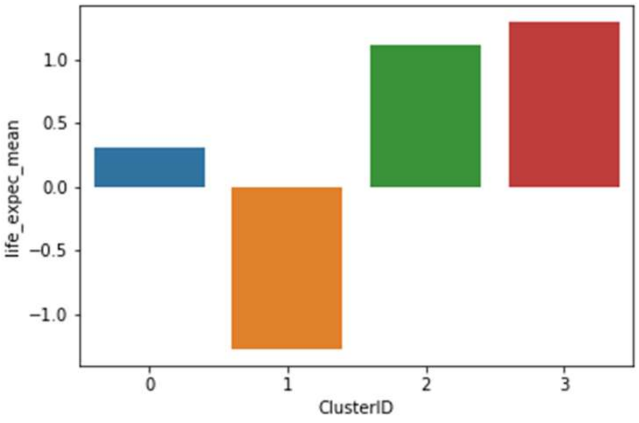
Clustering Analysis

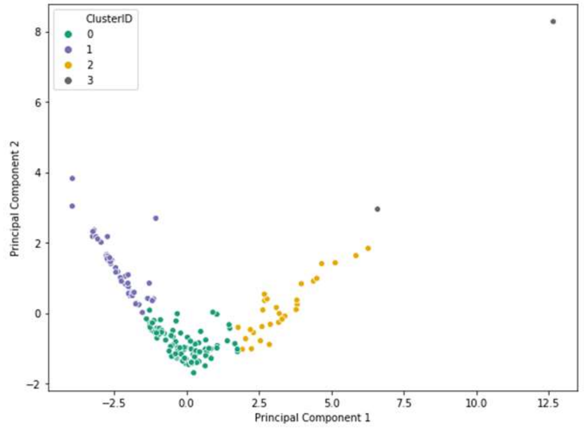








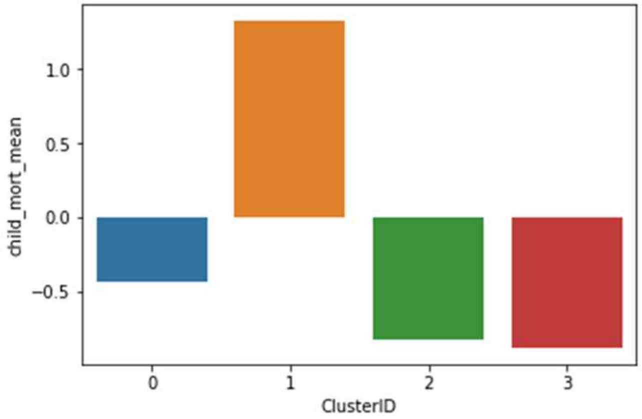


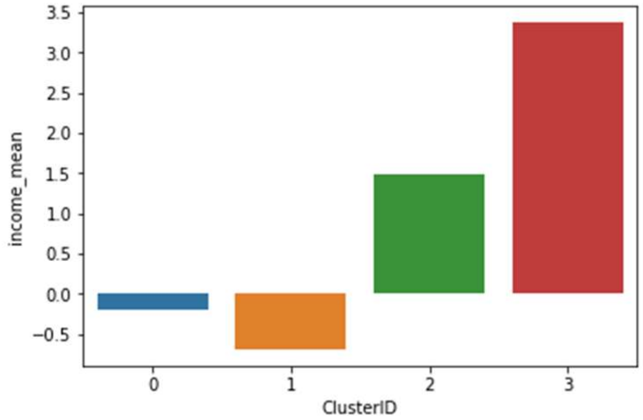


Clustering Analysis

Upon analysis we can see that Cluster 1 happens to be high on Child Mortality rate and Total fertility. Yet they are very low in Income, Health Expenditure.

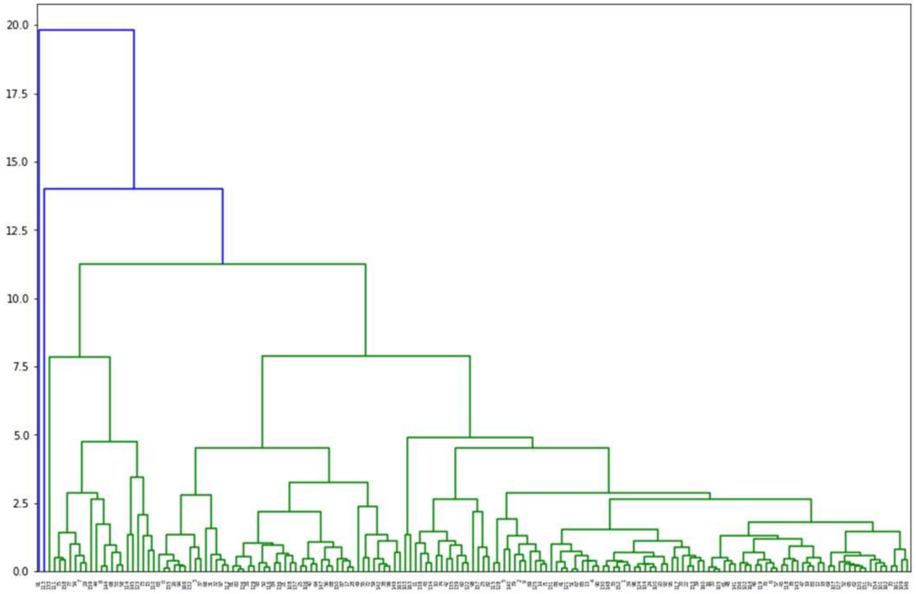
So the cluster we will be targeting is 1





Clustering Hierarchical

In Hierarchical Clustering we plotted the dendrogram to analyse the number of clusters we would opt for which was considered as 4.

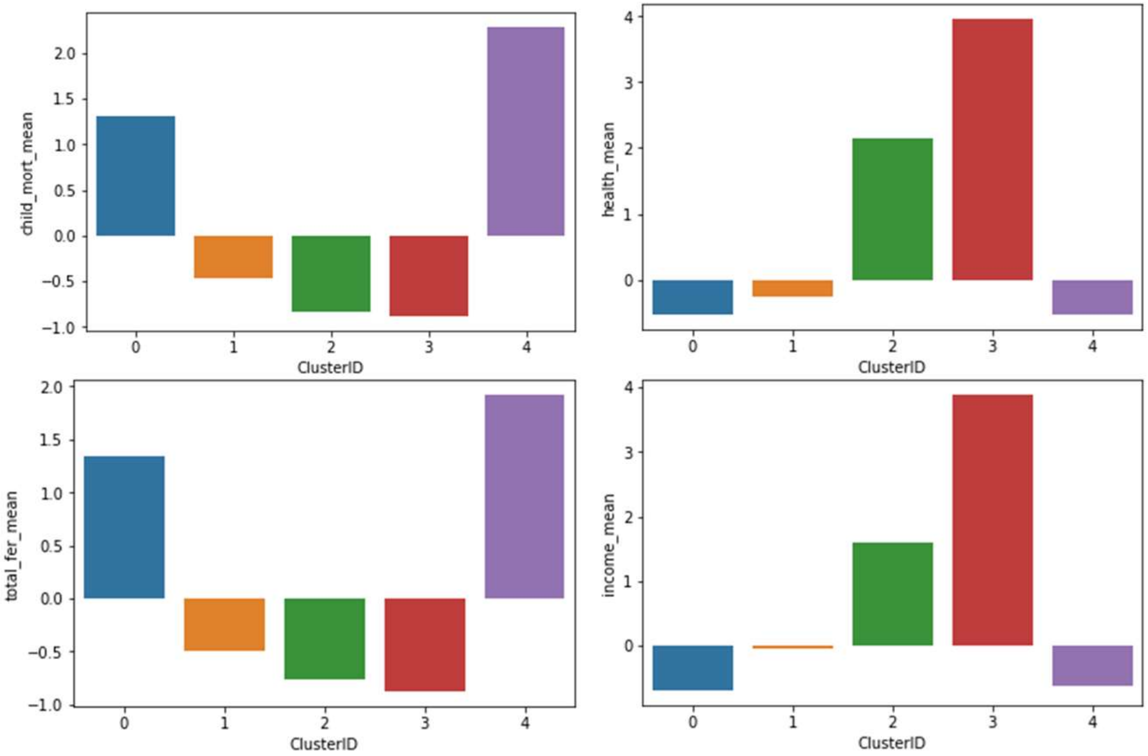


Clustering Hierarchical

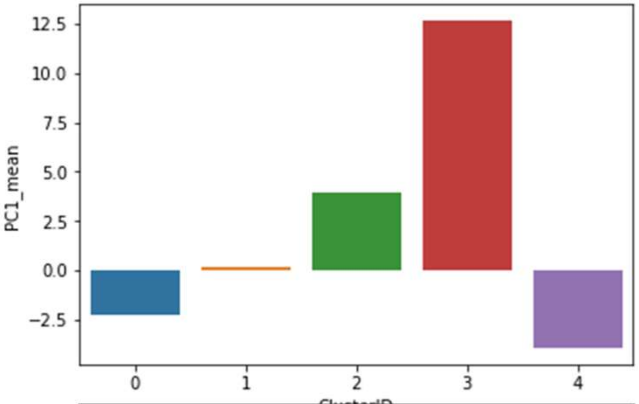
Upon Clustering the data, cluster 3 had high child mortality rate and total fertility as well, the other features such as income, health expenditure were low.

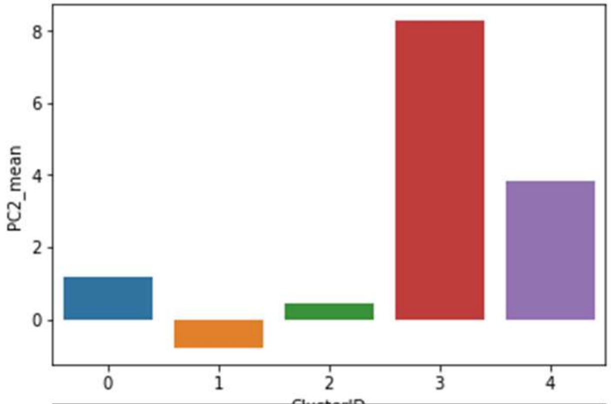
Upon further checking, only one country was clustered and that happened to be Nigeria.

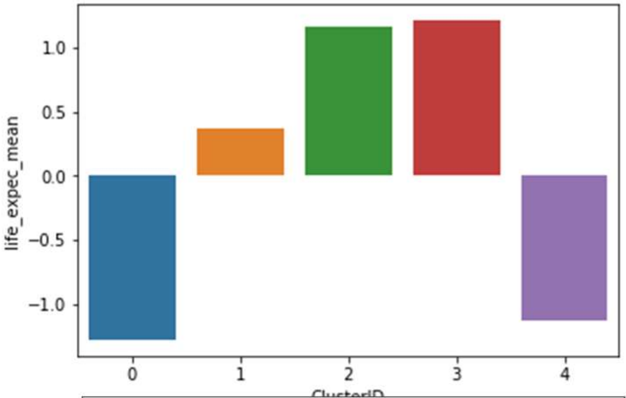
Further, if we check the cluster 0, there are quite few countries which follow the same pattern as Nigeria and are also in dire need of aid.

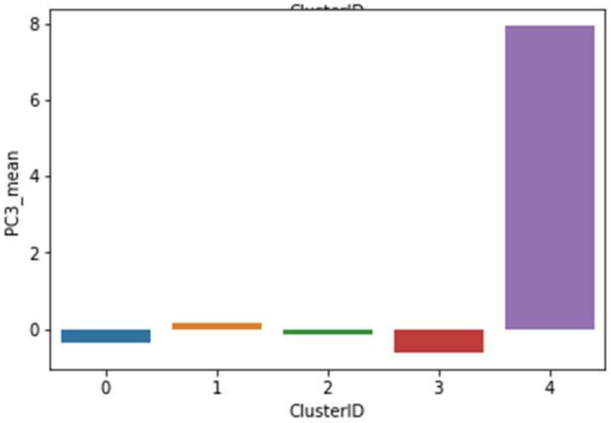


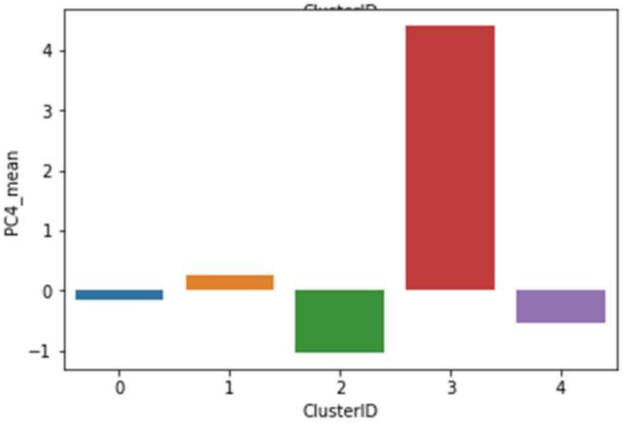
Clustering Hierarchical

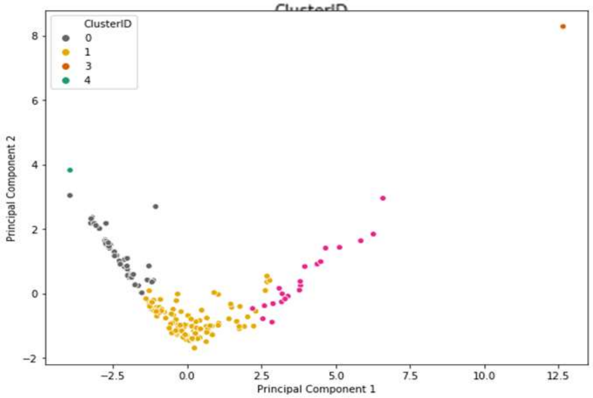












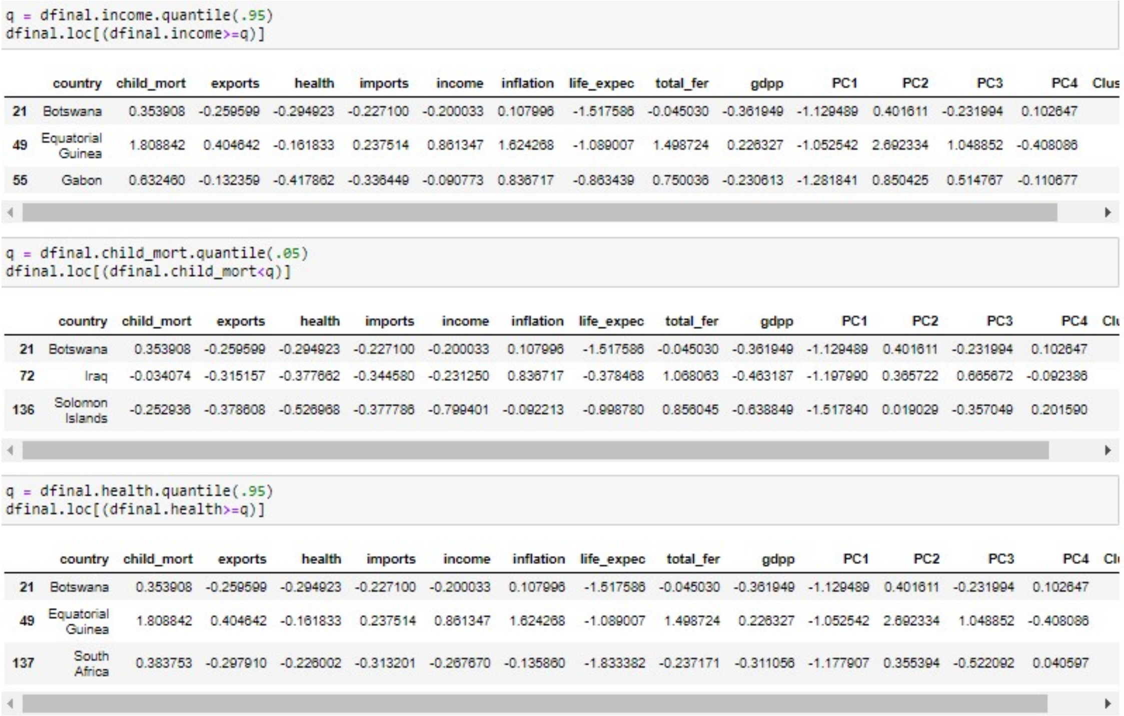
List of Countries:

The List of countries which are direst need of aid upon analysis are as listed beside.

- Afghanistan
- Angola
- Benin
- Botswana
- Burkina Faso
- Burundi
- Cameroon
- Central African Republic
- Chad
- Comoros
- Congo, Dem. Rep.
- Congo, Rep.
- Cote d'Ivoire
- Equatorial Guinea
- Eritrea
- Gabon
- Gambia
- Ghana
- Guinea
- Guinea-Bissau
- Haiti
- Iraq
- Kenya
- Kiribati
- Lao
- Lesotho
- Liberia
- Madagascar
- Malawi
- Mali
- Mauritania
- Mozambique
- Namibia
- Niger
- Pakistan
- Rwanda
- Senegal
- Sierra Leone
- Solomon Islands
- South Africa
- Sudan
- Tanzania
- Timor-Leste
- Togo
- Uganda
- Yemen

List of Countries:

Also, upon checking further, we can see **Botswana** can be considered a little less priority as the country has less child mortality and is better in health expenditure and income compared to the rest.



```
q = dfinal.income.quantile(.95)
dfinal.loc[(dfinal.income>=q)]
```

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	PC1	PC2	PC3	PC4	Clus
21	Botswana	0.353908	-0.259599	-0.294923	-0.227100	-0.200033	0.107996	-1.517586	-0.045030	-0.361949	-1.129489	0.401611	-0.231994	0.102647	
49	Equatorial Guinea	1.808842	0.404642	-0.161833	0.237514	0.861347	1.624268	-1.089007	1.498724	0.226327	-1.052542	2.692334	1.048852	-0.408086	
55	Gabon	0.632460	-0.132359	-0.417862	-0.336449	-0.090773	0.836717	-0.863439	0.750036	-0.230613	-1.281841	0.850425	0.514767	-0.110677	

```
q = dfinal.child_mort.quantile(.05)
dfinal.loc[(dfinal.child_mort<q)]
```

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	PC1	PC2	PC3	PC4	Cl
21	Botswana	0.353908	-0.259599	-0.294923	-0.227100	-0.200033	0.107996	-1.517586	-0.045030	-0.361949	-1.129489	0.401611	-0.231994	0.102647	
72	Iraq	-0.034074	-0.315157	-0.377662	-0.344580	-0.231250	0.836717	-0.378468	1.068063	-0.463187	-1.197990	0.365722	0.665672	-0.092386	
136	Solomon Islands	-0.252936	-0.378608	-0.526968	-0.377786	-0.799401	-0.092213	-0.998780	0.856045	-0.638849	-1.517840	0.019029	-0.357049	0.201590	

```
q = dfinal.health.quantile(.95)
dfinal.loc[(dfinal.health>=q)]
```

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	PC1	PC2	PC3	PC4	Cl
21	Botswana	0.353908	-0.259599	-0.294923	-0.227100	-0.200033	0.107996	-1.517586	-0.045030	-0.361949	-1.129489	0.401611	-0.231994	0.102647	
49	Equatorial Guinea	1.808842	0.404642	-0.161833	0.237514	0.861347	1.624268	-1.089007	1.498724	0.226327	-1.052542	2.692334	1.048852	-0.408086	
137	South Africa	0.383753	-0.297910	-0.226002	-0.313201	-0.267670	-0.135860	-1.833382	-0.237171	-0.311056	-1.177907	0.355394	-0.522092	0.040597	