‘Help International’ NGO Data Analysis with Clustering & PCA

- **Presented By Bhaswati Paul**

Present the overall approach of the analysis:

NGO ables to an awareness drives and funding programmes, they have been able to raise around $ 10 million.

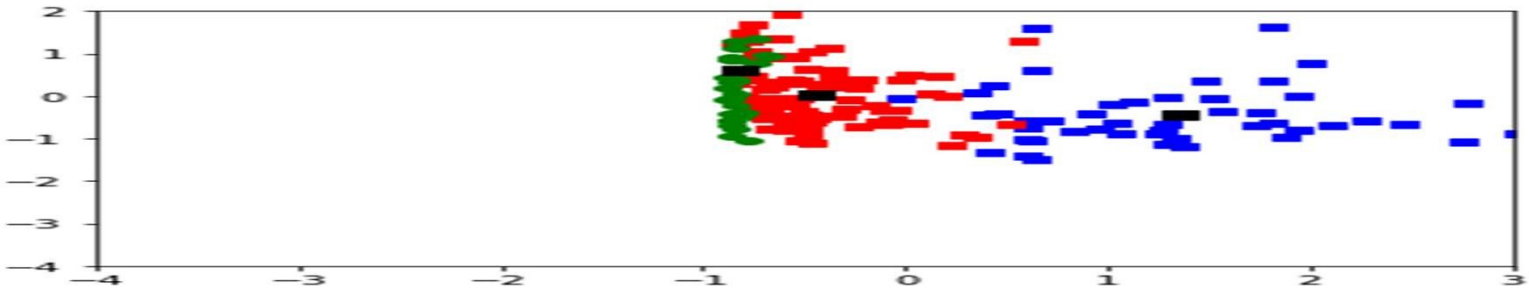
NGO needs to decide how to use this money strategically and effectively. The significant issues that come while making this decision are mostly related to choosing the countries that are in the direst need of aid.

Our job is to categorise the countries using some socio-economic and health factors that determine the overall development of the country.

Based on analysis CEO should focuses the countries where financial helps & awareness camp would require.

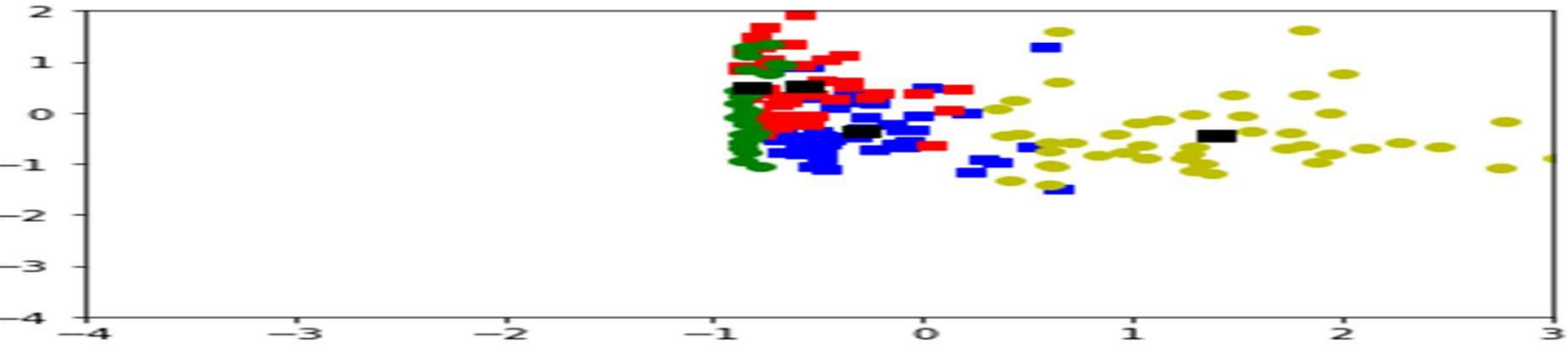
- The data set contains 167 rows corresponding to countries and nine factors, child_mort, Exports, Health, Imports, Income, Inflation, life_expec, total_fer,Economy in GDP per capital. Country is categorical variable whereas all others are numerical.
- Here we conduct a k-means clustering and PCA.
- For ease of analysis, we have chosen only the numerical variables because we cannot do clustering or PCA with categorical variables. The features or variables chosen are 'child_mort', 'Exports', 'Health', 'Imports', 'Income,' 'Inflation', 'life_expec', 'total_fer','Economy in GDP per capital'.
- **Technical Approaches**:
- **Standardize the data**
- we have selected the data we need to standardize the data. We will standardize the data by defining a function for standardization.
- ## Standardize the variables. def standardization_f(x): x_bar = np.mean(x) s = np.std(x) x_z = (x - x_bar) / s return(x_z) print(x.corr()) x_z = x.apply(standardization_f, broadcast = True) x_z.head()

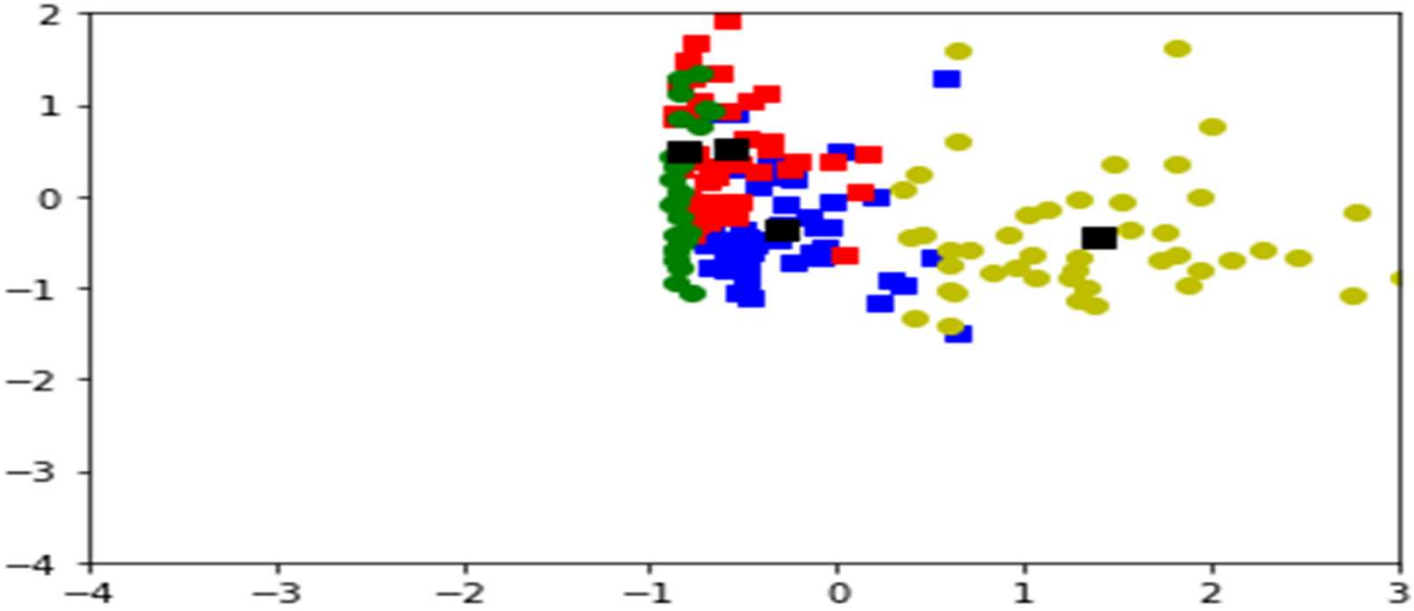
- In k-means clustering, we come up with number of clusters we would like to see. Then after plotting a scatter plot, we see if the number of clusters are appropriate or not.
- Here, we will start by asking the algorithm to assign three clusters and see how it fits the data.



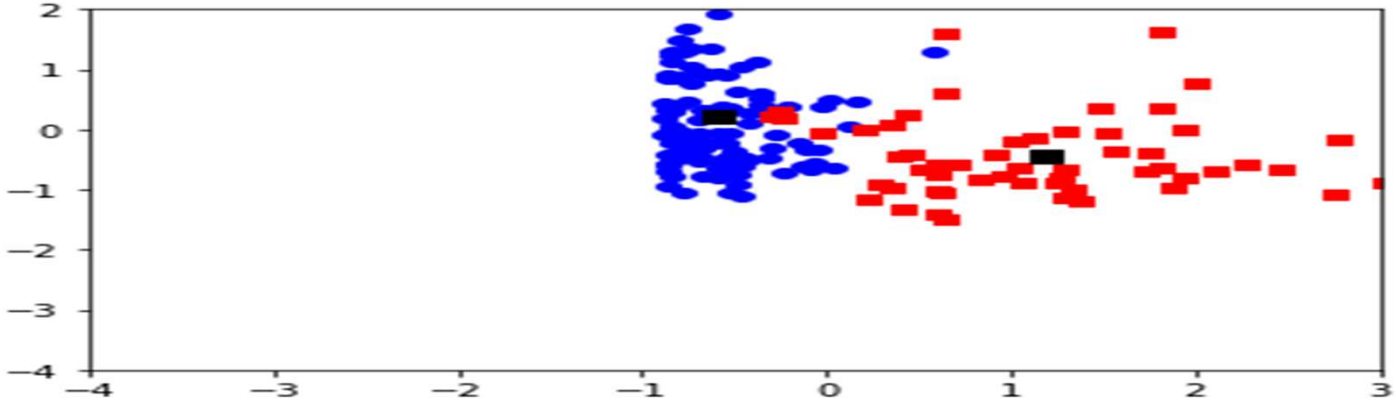
Each color represents a cluster. We can see more or less clean separation in blue and red but slightly more overlap in red and green.

- Now lets try asking the algorithm for four clusters and observe it in a scatterplot.
- kmeans clustering for 4 clusters.





Four clusters look worse than three. Now, lets try two clusters only.



We conducted K-means clustering using K = 2, 3, and 4. We think K = 3 gives the best discrimination based on scatter plots observed above.

Principal Component Analysis:

Principal component analysis is a unsupervised machine learning algorithm which reduces the dimensionality of the data by eigenvalue and eigenvector decomposition of the covariance matrix.

- **First start by computing the co-variance matrix of the standardized data.**
- **Then, calculate the eigenvalues and eigenvectors of covariance matrix.**
- ## Compute eigenvalues and eigenvectors of covariance matrix. eigen_RR = np.linalg.eig(RR) print(eigen_RR[0]) print(eigen_RR[1])

In this output, the first row is the eigenvalues for the 9 variables we have used in this dataset. The subsequent 9 rows are the eigenvectors in 9 directions.

The number of principal components is equal to the number of variables we have in the dataset. But, not all of the principal components are informative. To see, how many of them are informative or relevant, we calculate the proportion of variance explained by each PC.

Now, lets see what proportion of variance is explained by the Principal components in this analysis.

Proportion of variance explained by each PC. print(eigen_RR[0] / sum(eigen_RR[0]))

#We see that first PC explains 46% of the variance and second PC explains 17% of the variance while the third explains only 13% of the variance. So, based on this we will select only the first two PCs.

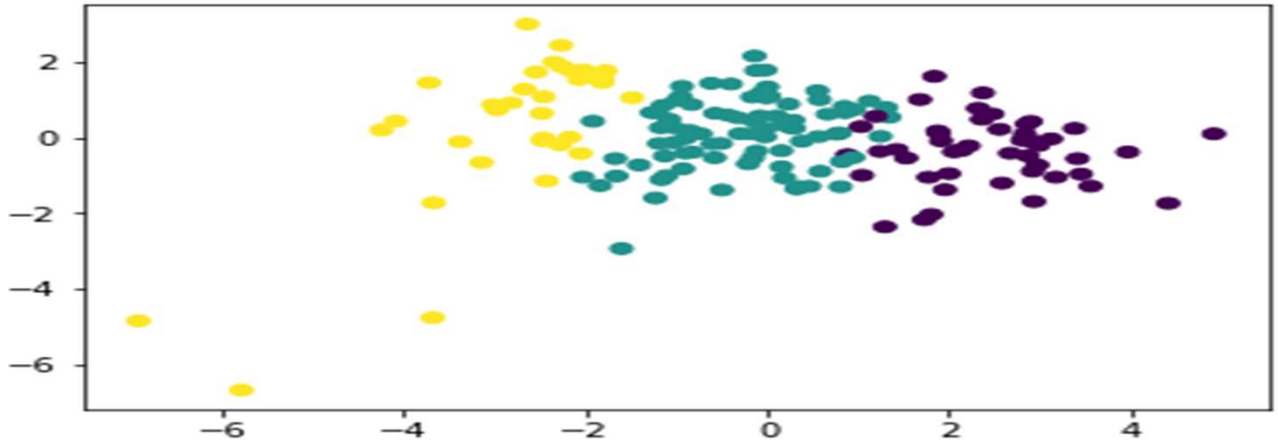
#Now, lets calculate the PC values for the first two PCs. We can do that by matrix multiplying the standardized data with the first two eigenvectors.

Compute first 2 PCs by matrix multiplying the standardized data array

by the first 2 columns of the array of eigenvectors.

if the PCA algorithm will be able to separate the clusters, We are adding clusters from (k =3) to the original dataframe. Then we will plot the scatterplot of PC1 vs PC2 to see if the three clusters separate as they should.

The cluster membership is added as a separate column “cluster”. Now, we will plot the scatterplot of PC1 vs PC2 and color code them by cluster membership.
Scatterplot of first 2 PCs, color coded by cluster membership. plt.scatter(PCs[:, 0], PCs[:, 1],c= x_z['cluster’])



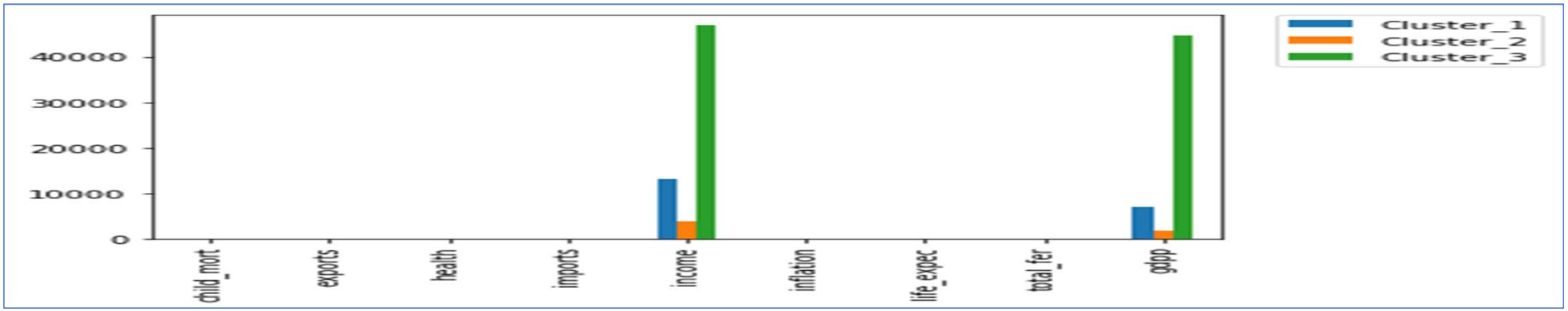
The PCA clearly separates the three clusters that we found via the k-means clustering (as shown in the fig by different colors).
Now, lets look at the characteristics of each cluster by calculating means of each clusters.

	Cluster_1	Cluster_2	Cluster_3
child_mort	20.897674	92.366667	4.857576
exports	41.576512	28.546229	58.163636
health	6.274419	6.296458	8.981515
imports	48.402326	41.443040	50.872727
income	13100.581395	3937.770833	46893.939394
inflation	7.471174	11.915938	2.578182
life_expec	73.037209	59.345833	80.393939
total_fer	2.281860	4.953958	1.766061
gdpp	7014.813953	1902.916667	44557.575758
cluster_ID	0.000000	1.000000	2.000000

#In the above table, we have tabulated three clusters that we identified by k-means clustering with their means for respective variables.

#From the table, looking at the mean values for each features, we can say that Cluster 2 are the most poor countries in-term of income,GDPP,child mortality followed by cluster 1 and cluster 3. Cluster3 are the most richest counties based on all metrics. Cluster 1 are far better on child_mort, export, Health, income, and inflation than Cluster 2 but cluster 2 is slightly better in 'total_fer' factor.

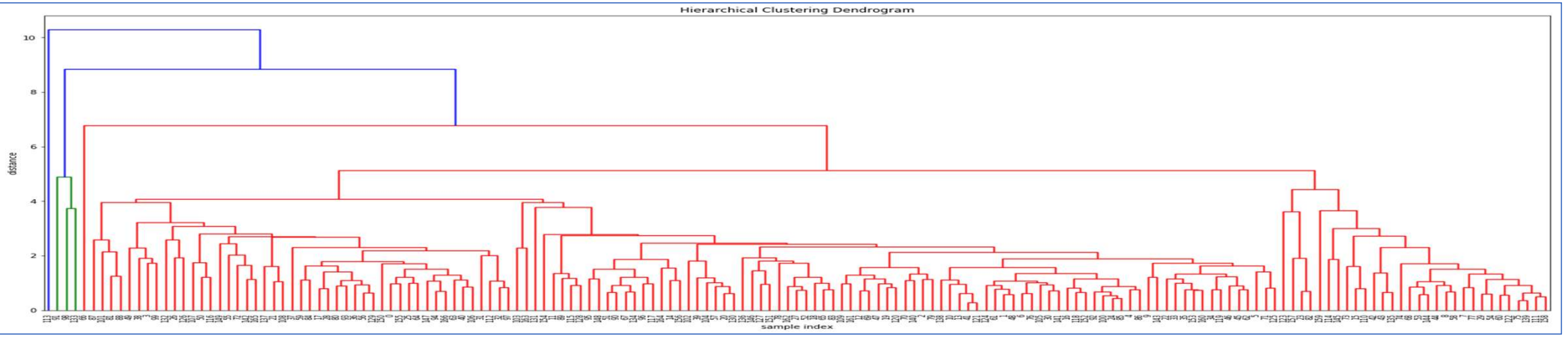
- Bar Graph represents above scenario quite well in to the next slide.

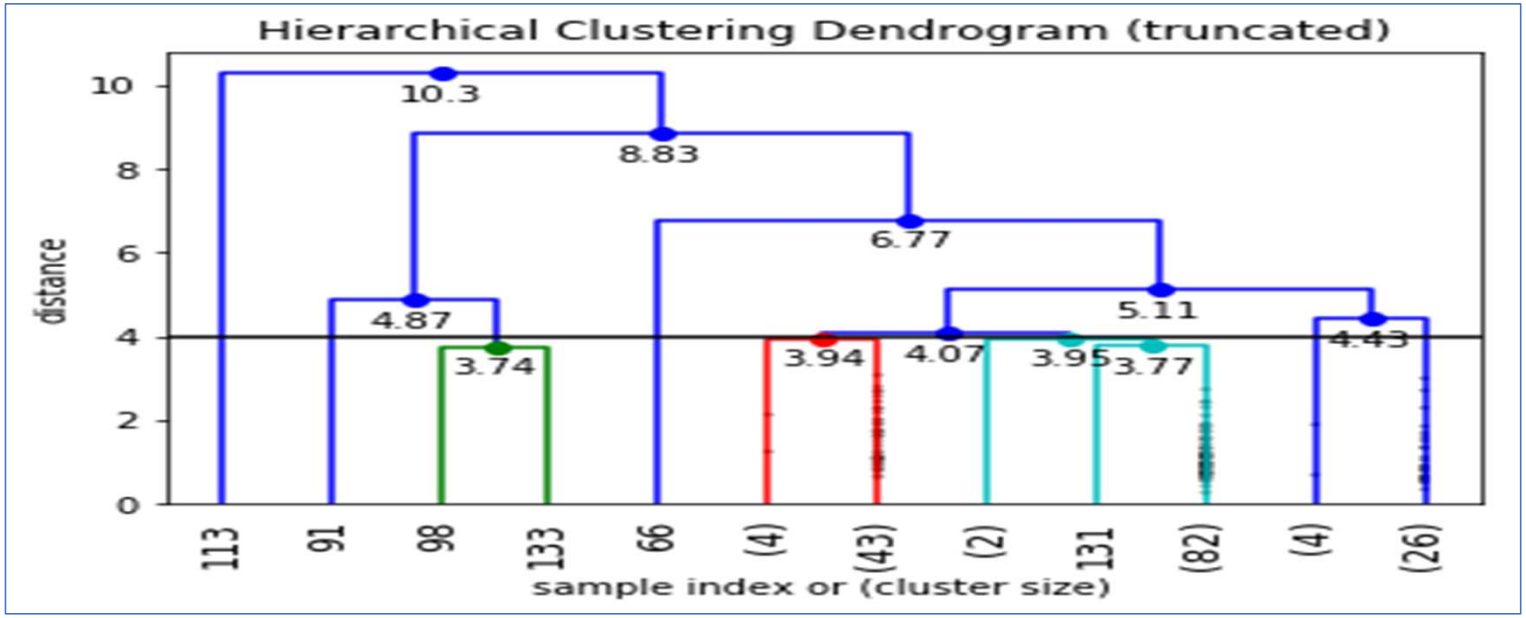


Hierarchical Clustering:

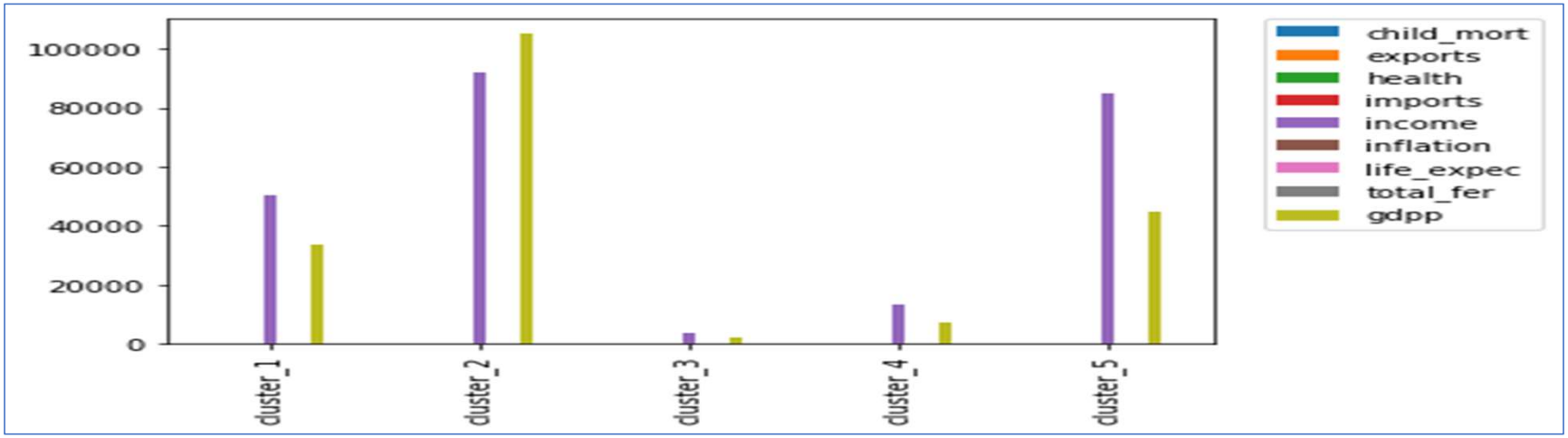
Cophenetic Correlation Coefficient gives the correlation between actual pairwise distance of all samples to those implied by hierarchical clustering.

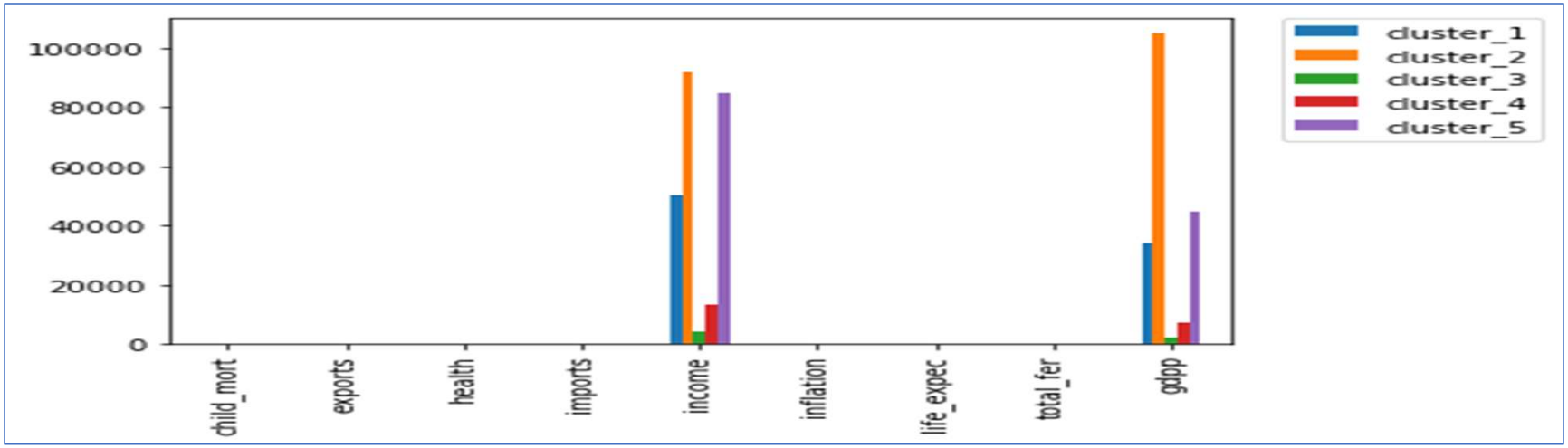
The closer the value is to 1, the better the clustering preserves the original distances. Here, we are calculating that using cophenet() function.





	cluster_1	cluster_2	cluster_3	cluster_4	cluster_5
child_mort	4.800	2.80	87.991489	20.672941	9.7250
exports	176.500	175.00	28.789766	41.789176	68.5250
health	6.305	7.77	6.477660	6.181176	2.7350
imports	164.000	142.00	42.301402	48.018824	36.4500
income	50200.000	91700.00	3951.127660	13215.411765	84600.0000
inflation	1.892	3.62	9.921596	7.514365	11.8450
life_expec	81.500	81.30	60.029787	73.127059	77.8250
total_fer	1.255	1.63	4.937872	2.268000	1.9975
gdpp	33850.000	105000.00	1940.595745	7063.694118	44775.0000
cluster_ID	2.000	2.00	0.978723	0.000000	2.0000
cluster_id	1.000	2.00	3.000000	4.000000	5.0000


100000
80000
60000
40000
20000
0
cluster_1
cluster_2
cluster_3
cluster_4
cluster_5
child_mort
exports
health
imports
income
inflation
life_expec
total_fer
gdpp


100000
80000
60000
40000
20000
0
child_mort
exports
health
imports
income
inflation
life_expec
total_fer
gdpp
cluster_1
cluster_2
cluster_3
cluster_4
cluster_5

Conclusions based on hierarchical clustering

#Cluster 3 are mostly poor countries with low GDP, low income & low life_expec & health & high child_mortal factor.

but they are happy in terms of total_fer factor.

#Cluster4 are also poor but not as poor as cluster 3. They have better income and GDP & life_expec & health & low child mortality than cluster 3.

#Cluster 1 are the better countries with higher GDP, higher income,life_expec & lower child_mort than cluster3 & cluster4.

Counties.

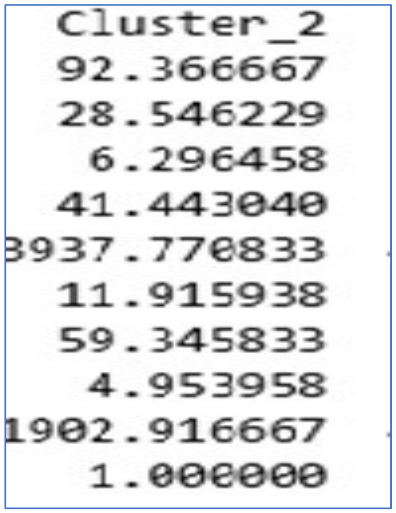
#Cluster 5 are better countries with higher GDP, higher income,life_expec,low child_mort, than cluster4,cluster3 and cluster1.

#Cluster 2 are the best countries with higher GDP, higher income,life_expec,lowest child_mort.It is good in respect of all factors.

- **In hierarchical clustering Cluster 3 contains most poor counties list. Name of the countries name are almost same as K mean clustering.**
- **In hierarchical clustering Cluster 4 contains developing counties list. Name of the countries name are same as K mean clustering.**
- **Cluster 3 are most focusing countries as per Income,GDP,Health,Life_Expect,Child_mortality factors.**
- **Fund distribution and awareness campaign should be focussing on these countries.**
- **Cluster5 countries as below:**
- **#United Arab Emirates#Kuwait#Brunei#Qatar**
- **Cluster1 countries as below:**
- **#Czech Republic#Malta#Singapore**
- **Cluster2 countries:Luxembourg**
- **Cluster5,cluster1,cluster2 all are developed countries.**

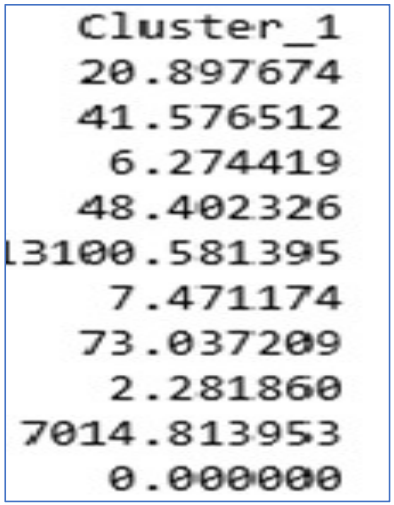
- **Name of countries best on K mean Clustering in next slide**

Poor Countries	
Congo, Dem. Rep.	Chad
Liberia	Tanzania
Burundi	Senegal
Niger	Lesotho
Central African Republic	Kenya
Mozambique	Cameroon
Malawi	Cote d'Ivoire
Guinea	Ghana
Togo	Zambia
Sierra Leone	Mauritania
Rwanda	Sudan
Guinea-Bissau	Lao
Madagascar	Pakistan
Comoros	Yemen
Eritrea	Nigeria
Burkina Faso	Congo, Rep.
Haiti	Namibia
Uganda	South Africa
Afghanistan	Iraq



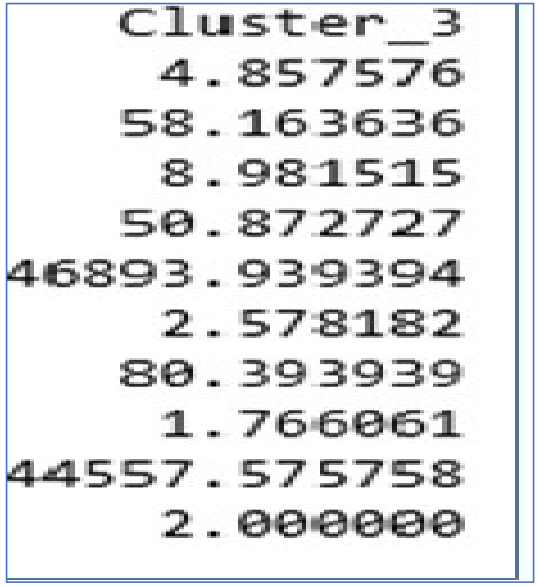
- **Poor Countries List**

countries 2				
Solomon Islands	Philippines	Ecuador	Serbia	Iran
Nepal	Cape Verde	Jordan	Algeria	Romania
Tajikistan	Guyana	China	Costa Rica	Turkey
Bangladesh	Bhutan	Bosnia and Herzegovina	Thailand	Argentina
Cambodia	Morocco	Egypt	Montenegro	Antigua and Barbuda
Kyrgyz Republic	Armenia	St. Vincent and the Gren	Suriname	Chile
Vanuatu	Guatemala	Albania	Brazil	Croatia
Micronesia, Fed. Sts.	Georgia	Turkmenistan	Barbados	Kazakhstan
Myanmar	Paraguay	Peru	Bulgaria	Lithuania
Moldova	El Salvador	Tunisia	Gabon	Malaysia
Uzbekistan	Fiji	Maldives	Panama	Poland
India	Mongolia	Colombia	Mauritius	Hungary
Vietnam	Ukraine	Dominican Republic	Azerbaijan	Estonia
Tonga	Belize	Grenada	Belarus	Bahamas
Samoa	Jamaica	Macedonia, FYR	Lebanon	Russia
Bolivia	Indonesia	South Africa	Venezuela	Israel
Philippines	Sri Lanka	Iraq	Uruguay	Libya



- **Developing countries**

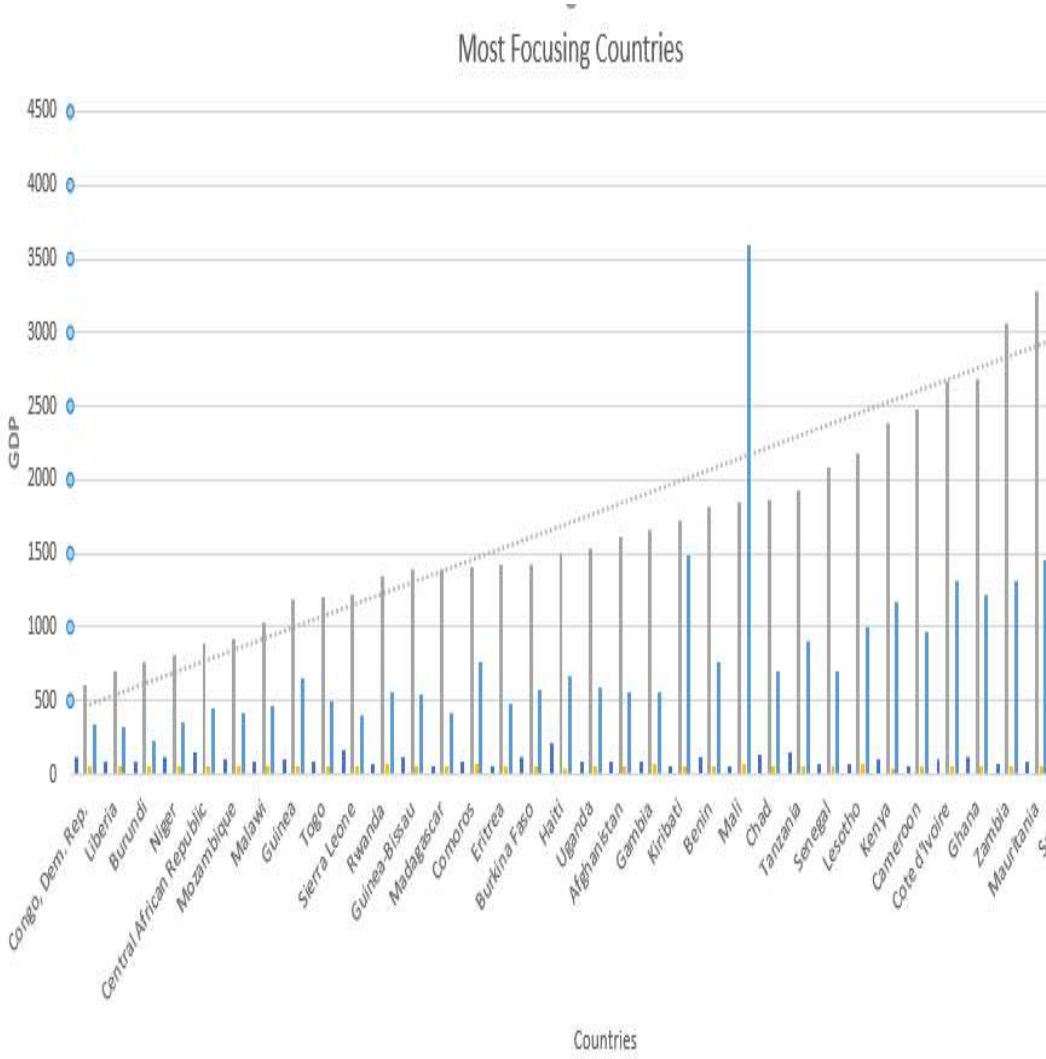
Best Countries		
Slovak Republic	Sweden	France
Portugal	Austria	Iceland
Czech Republic	Denmark	Finland
Malta	Netherlands	Germany
Greece	Ireland	Canada
Slovenia	United States	Bahrain
Israel	Switzerland	Belgium
South Korea	United Arab Emir	Australia
New Zealand	Norway	
Spain	Singapore	
Cyprus	Kuwait	
Japan	Brunei	
Italy	Luxembourg	
United Kingdom	Qatar	



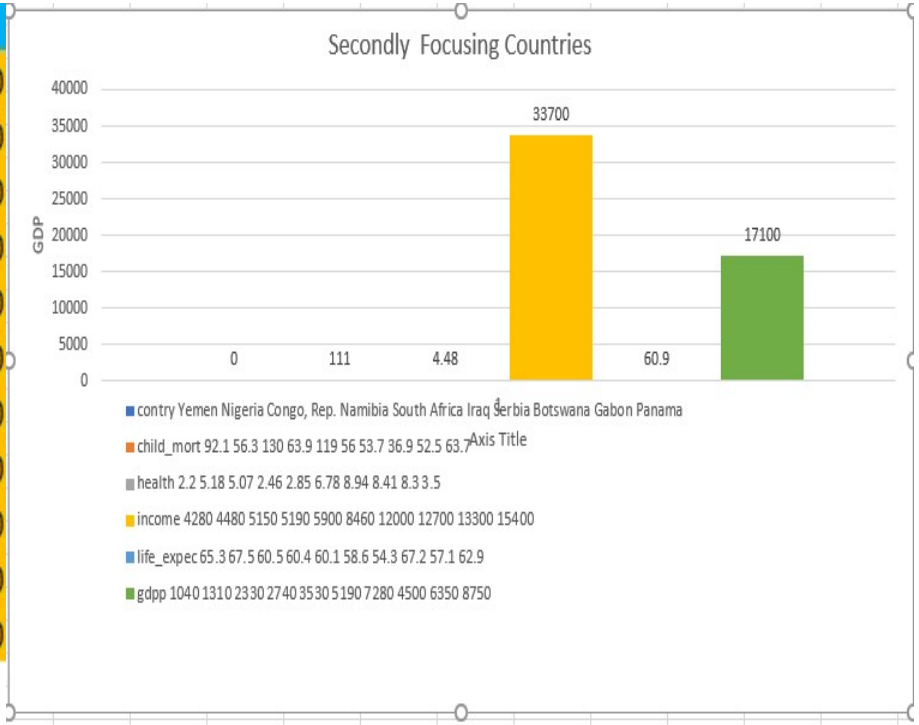
- **Richest Countries**

Highly Focusing countries where major funds & awareness requires

contry	child_mor	exports	health	imports	income	inflation	life_expe	total_fer	gdpp
Congo, De	116	41.1	7.91	49.6	609	20.8	57.5	6.54	334
Liberia	89.3	19.1	11.8	92.6	700	5.47	60.8	5.02	327
Burundi	93.6	8.92	11.6	39.2	764	12.3	57.7	6.26	231
Niger	123	22.2	5.16	49.1	814	2.55	58.8	7.49	348
Central Af	149	11.8	3.98	26.5	888	2.01	47.5	5.21	446
Mozambic	101	31.5	5.21	46.2	918	7.64	54.5	5.56	419
Malawi	90.5	22.8	6.59	34.9	1030	12.1	53.1	5.31	459
Guinea	109	30.3	4.93	43.2	1190	16.1	58	5.34	648
Togo	90.3	40.2	7.65	57.3	1210	1.18	58.7	4.87	488
Sierra Leo	160	16.8	13.1	34.5	1220	17.2	55	5.2	399
Rwanda	63.6	12	10.5	30	1350	2.61	64.6	4.51	563
Guinea-Bi	114	14.9	8.5	35.2	1390	2.97	55.6	5.05	547
Madagasc	62.2	25	3.77	43	1390	8.79	60.8	4.6	413
Comoros	88.2	16.5	4.51	51.7	1410	3.87	65.9	4.75	769
Eritrea	55.2	4.79	2.66	23.3	1420	11.6	61.7	4.61	482
Burkina Fa	116	19.2	6.74	29.6	1430	6.81	57.9	5.87	575
Haiti	208	15.3	6.91	64.7	1500	5.45	32.1	3.33	662
Uganda	81	17.1	9.01	28.6	1540	10.6	56.8	6.15	595
Afghanist	90.2	10	7.58	44.9	1610	9.44	56.2	5.82	553
Gambia	80.3	23.8	5.69	42.7	1660	4.3	65.5	5.71	562
Kiribati	62.7	13.3	11.3	79.9	1730	1.52	60.7	3.84	1490
Benin	111	23.8	4.1	37.2	1820	0.885	61.8	5.36	758
Mali	62.6	2.2	9.12	27.8	1850	26.5	71.1	6.23	3600
Chad	137	22.8	4.98	35.1	1870	4.37	59.5	6.55	708
Tanzania	150	36.8	4.53	43.5	1930	6.39	56.5	6.59	897
Senegal	71.9	18.7	6.01	29.1	2090	9.25	59.3	5.43	702
Lesotho	66.8	24.9	5.66	40.3	2180	1.85	64	5.06	1000
Kenya	99.7	39.4	11.1	101	2380	4.15	46.5	3.3	1170
Cameroor	62.2	20.7	4.75	33.6	2480	2.09	62.8	4.37	967
Cote d'Ivo	108	22.2	5.13	27	2660	1.91	57.3	5.11	1310
Ghana	111	50.6	5.3	43.3	2690	5.39	56.3	5.27	1220
Zambia	74.7	29.5	5.22	45.9	3060	16.6	62.2	4.27	1310
Mauritani	83.1	37	5.89	30.9	3280	14	52	5.4	1460
Sudan	97.4	50.7	4.41	61.2	3320	18.9	68.2	4.98	1200



contry	child_mor	exports	health	imports	income	inflation	life_expe	total_fer	gdpp
Yemen	92.1	13.5	2.2	19.4	4280	10.9	65.3	3.85	1040
Nigeria	56.3	30	5.18	34.4	4480	23.6	67.5	4.67	1310
Congo, Re	130	25.3	5.07	17.4	5150	104	60.5	5.84	2330
Namibia	63.9	85.1	2.46	54.7	5190	20.7	60.4	4.95	2740
South Afri	119	62.3	2.85	42.9	5900	22.4	60.1	6.16	3530
Iraq	56	47.8	6.78	60.7	8460	3.56	58.6	3.6	5190
Serbia	53.7	28.6	8.94	27.4	12000	6.35	54.3	2.59	7280
Botswana	36.9	39.4	8.41	34.1	12700	16.6	67.2	4.56	4500
Gabon	52.5	43.6	8.3	51.3	13300	8.92	57.1	2.88	6350
Panama	63.7	57.7	3.5	18.9	15400	16.6	62.9	4.08	8750
Equatorial	111	85.8	4.48	58.9	33700	24.9	60.9	5.21	17100



#Above countries come under cluster 2 that means, it comes under poor countries list but their avg. income ,GDP are better, But Child mortality rate is quite high than other countries in poor countries list. So need proper awareness camp, medical attention to reduce child motility.

#Arrangement of Purified drinking water which will reduce diarrhea, de-hydration, infections and can reduces child mortality rate.

Problem & Resolution:
#Most of these countries ‘ve less no of Industries.
#Political in-stability, corruptions.
#Less Rain fall over the years.
#Less scope of Agricultures and industries.
#Less Education & Medical awareness.
#Less Government awareness on Health & over all developments.
#Lack of Proper Drinking- Water is also major issues which causes a lot of child mortalities, health problems.
#Lack of Proper Sanitation facilities.
#As a result they ‘ve low avg income,health,low GDP,high child mortality rate.

Resolution:
#Making of Pure Drinking Water Plant may reduce child mortality, Dehydration, diarrhea related deaths. It also reduces a lot of contaminated daises & health problems.
#If educated to people with Rain water harvesting and purification of water by using simple processes may reduce child motility rate.
#Proper Medication & build up hygienic sanitation system helps in this regard a lot.
#Proper Education and awareness Programme also helpful to over come the situation.
#Awareness of recycling water and utilised natural resources also improve the situation.
#From the Analysis it ‘s observed
#Major fund should be distributed to the Poor countries & some of developing countries based on their Income,GDP,Health,Child_mortality, life_expec factors.

Final List of Countries:

- **Congo, Dem. Rep.**
- **Liberia**
- **Burundi**
- **Niger**
- **Central African Republic**
- **Mozambique**
- **Malawi**
- **Guinea**
- **Togo**
- **Sierra Leone**
- **Rwanda**
- **Guinea-Bissau**
- **Madagascar**
- **Comoros**
- **Eritrea**
- **Burkina Faso**
- **Haiti**
- **Uganda**
- **Afghanistan**
- **Gambia**
- **Kiribati**
- **Benin**
- **Mali**
- **Chad**
- **Tanzania**
- **Senegal**
- **Lesotho**
- **Kenya**
- **Cameroon**
- **India**
- **Uzbekistan**
- **Cote d'Ivoire**
- **Ghana**
- **Zambia**
- **Mauritania**
- **Sudan**
- **Lao**
- **Pakistan**
- **Solomon Islands**
- **Nepal**
- **Tajikistan**
- **Bangladesh**
- **Cambodia**
- **Kyrgyz Republic**
- **Vanuatu**
- **Myanmar**

Analysis countries more granular form from final list

Major fund needed:

1st Level focusing countries

- Congo, Dem. Rep.
- Liberia
- Burundi
- Niger
- Central African Republic
- Mozambique
- Malawi
- Guinea
- Togo
- Sierra Leone
- Rwanda
- Guinea-Bissau
- Madagascar
- Comoros
- Eritrea
- Burkina Faso
- Haiti
- Uganda
- Afghanistan
- Gambia
- Kiribati
- Benin
- Mali
- Chad
- Tanzania

2nd level focusing countries

- Senegal
- Lesotho
- Kenya
- Cameroon
- Cote d'Ivoire
- Ghana
- Zambia
- Mauritania
- Sudan
- Lao
- Pakistan

3rd level focusing countries

- Solomon Islands
- Nepal
- Tajikistan
- Bangladesh
- Cambodia
- Kyrgyz Republic
- Vanuatu
- Myanmar
- Moldova
- Uzbekistan
- India

#Medical & Hygiene, Pure Drinking -water, Proper Sanitation awareness camp also required here so child mortality should be reduced along with fund distribution.

Country Name Analysis Method:

From the analysis purpose we used combination of following important factors.

- Income
- GDP
- Health
- Life Expect
- Child Mortality
- inflation

We use K-mean cluster values as cut off values

- Income(609 to <=3942)
- GDP(231 to <=1922)
- Health(2.66 to <=6)
- Llife_expec(32.2 to <=59)
- Inflation(.885 to <=12)

- From the analysis, we get some important factors(Income,GDP,Health,Life Expect,Child Mortality,inflation) based on which we can distribute the fund properly.
- On the analysis we observe some of countries belong to poor countries list but they ‘ve high income, GDP rate but also high child mortality rate. In these countries we require proper awareness programme so that they can avoid high child mortality.