PCA + Clustering
(HELP International NGO Fund Investment)
Assignment

Created by : Hardik Kaneriya

Abstract

- Problem statement
- Business and data understanding
- Problem solving approach
- PCA(principal component analysis)
- K-means clustering
- Hierarchical clustering
- Both Analysis conlusion (Final list of countries)

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.

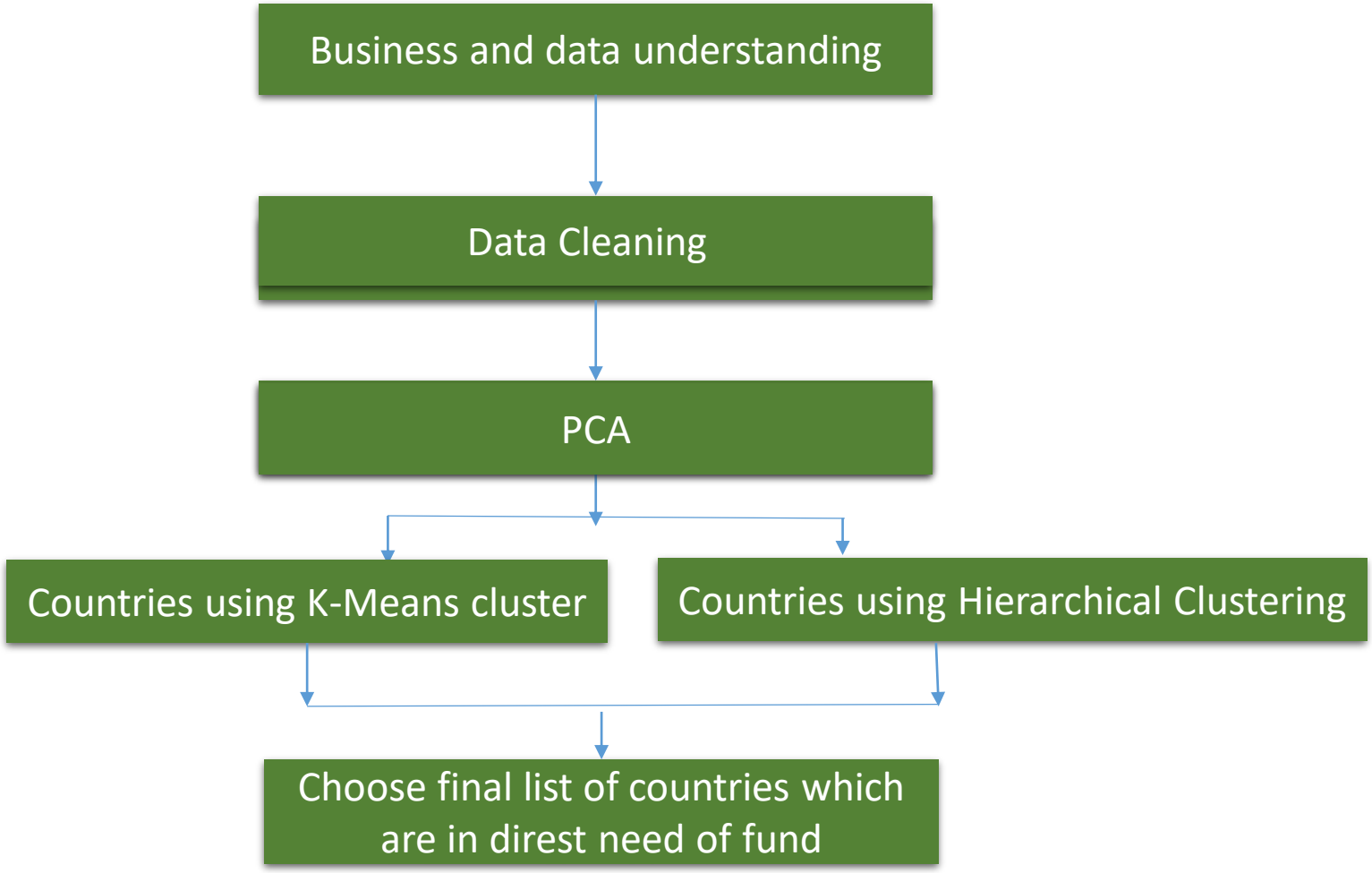
- Categorise the countries using some socio-economic and health factors that determine the overall development of the country. Then suggest the countries which the CEO needs to focus on the most.

Business and data understanding

Column name and their definition is as follows:

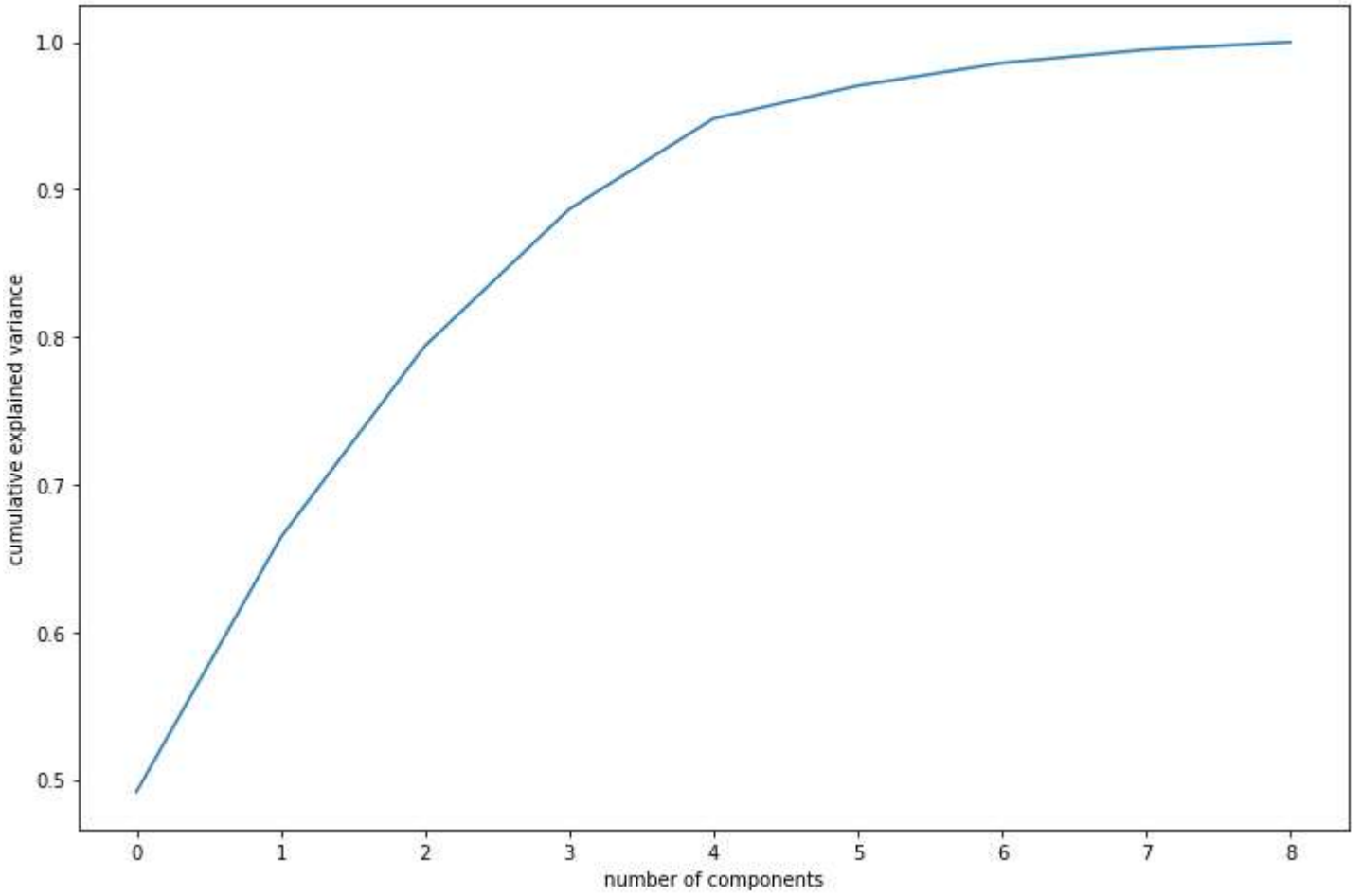
- country: Name of the country
- child_mort: Death of children under 5 years of age per 1000 live births
- exports: Exports of goods and services. Given as %age of the Total GDP
- health: Total health spending as %age of Total GDP
- Income: Net income per person
- Inflation: The measurement of the annual growth rate of the Total GDP
- life_expec: The average number of years a new born child would live if the current mortality patterns are to remain the same
- total_fer: The number of children that would be born to each woman if the current age-fertility rates remain the same.
- gdpp: The GDP per capita. Calculated as the Total GDP divided by the total population.

Problem Solving Approach



Dimension reduction (PCA)

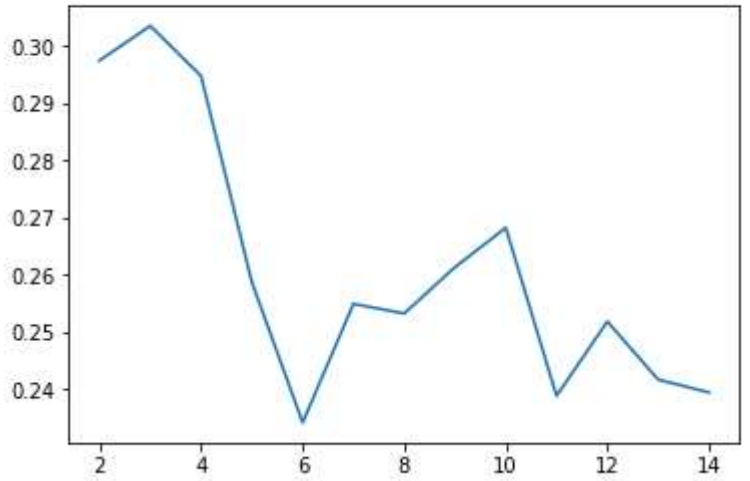
- We performed PCA and we were able to find the 5 principal component which were able to explain almost 95% of the total data without losing any information and droping any original variable.



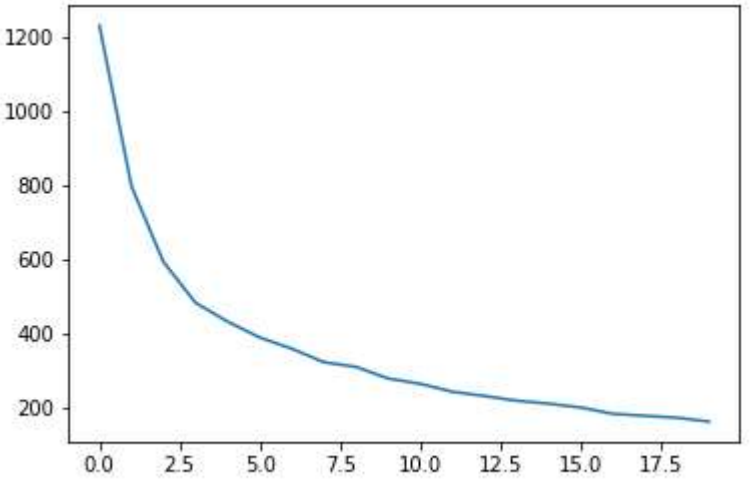
K – Means clustering

- After getting the PC's we created dataframe and performed k-means clustering on that.
- With the help of silhouette_score and sum of squared distance analysis we found that we can built 3 clusters for the countries

```
[<matplotlib.lines.Line2D at 0x26863b3550>]
```

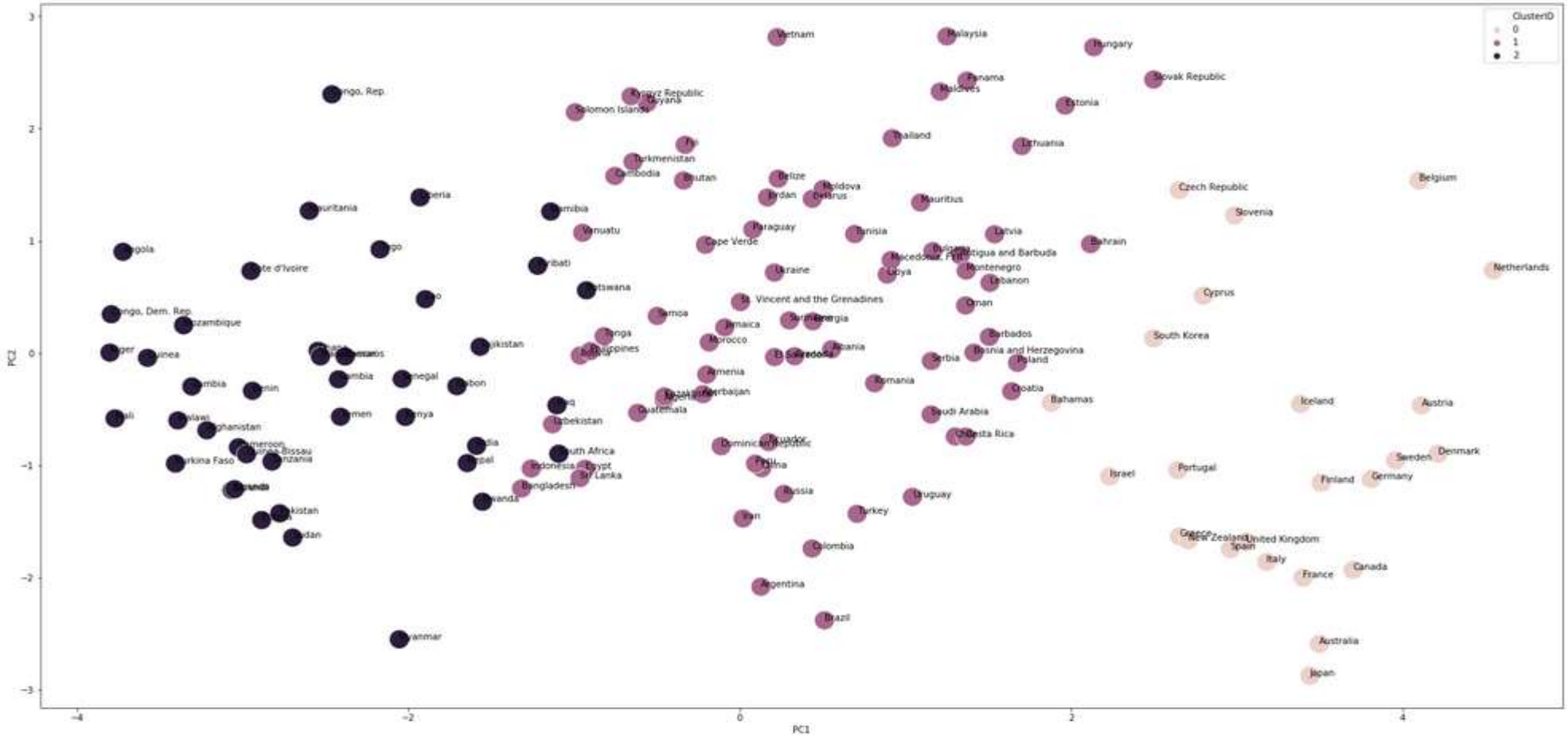


Silhouette score



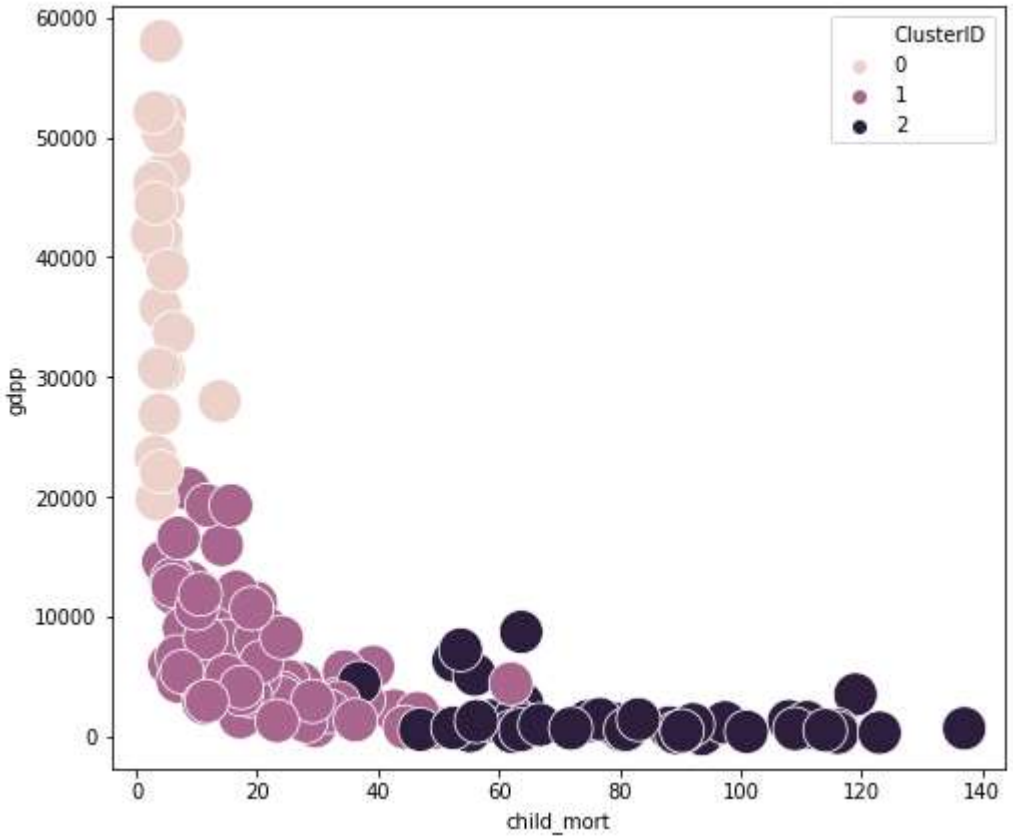
Elbow Curve

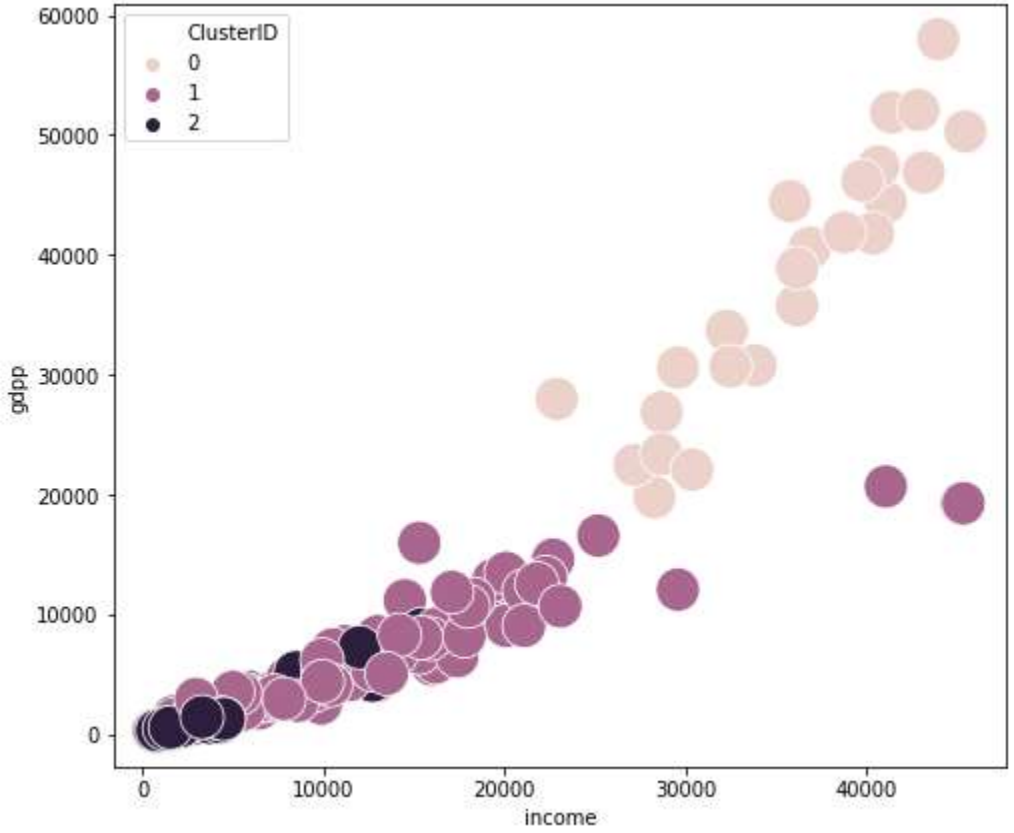
Countries falling under 3 clusters



Above plot clearly shows three clusters with different color. In these graph poor countries are spread left side of the graph

Relation between two original variables





Most Poor countries as per K-means

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
26	Burundi	93.6	8.92	11.60	39.2	764	12.3	57.7	6.26	231
37	Congo, Dem. Rep.	116.0	41.10	7.91	49.6	609	20.8	57.5	6.54	334
63	Guinea	109.0	30.30	4.93	43.2	1190	16.1	58.0	5.34	648
94	Malawi	90.5	22.80	6.59	34.9	1030	12.1	53.1	5.31	459
132	Sierra Leone	160.0	16.80	13.10	34.5	1220	17.2	55.0	5.20	399
166	Zambia	83.1	37.00	5.89	30.9	3280	14.0	52.0	5.40	1460

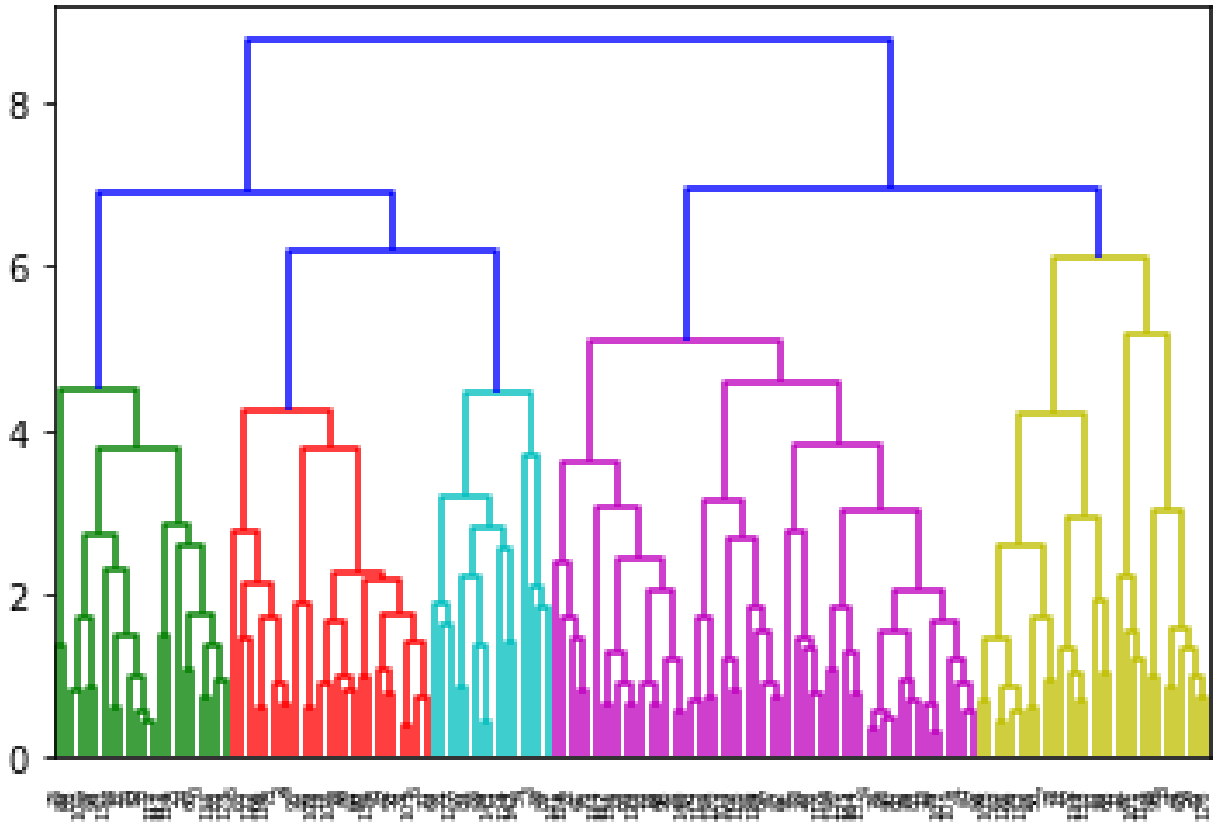
Conclusion for K-means clustering :

we found the top 6 poor countries which are in dier need of funds from the NGO

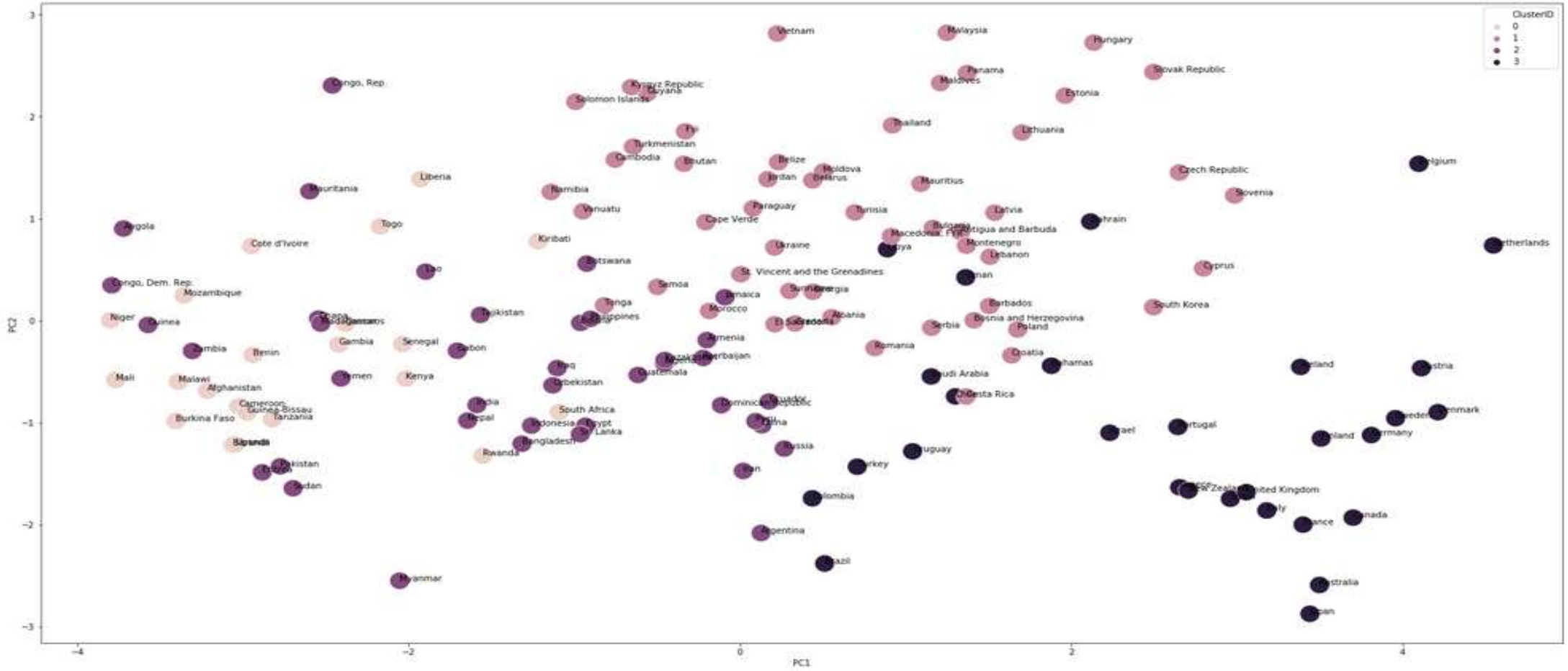
those countries : Burundi, Congo Dem. Rep., Guinea, Malawi, Sierra Leone, Zambia

Hierarchical clustering

- After getting the PC's we created dataframe and performed hierarchical clustering on that.
- With the help of dendgram we decided to form 4 cluster for to find the most poor countries

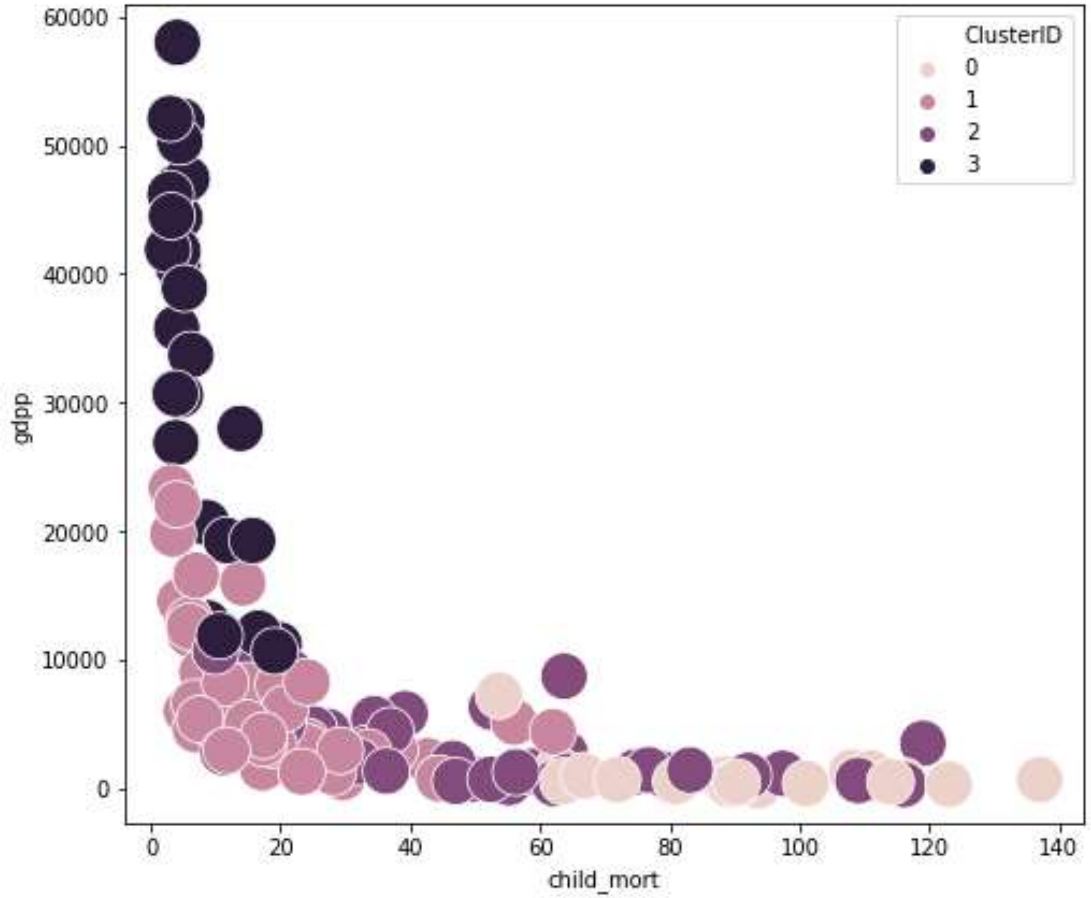


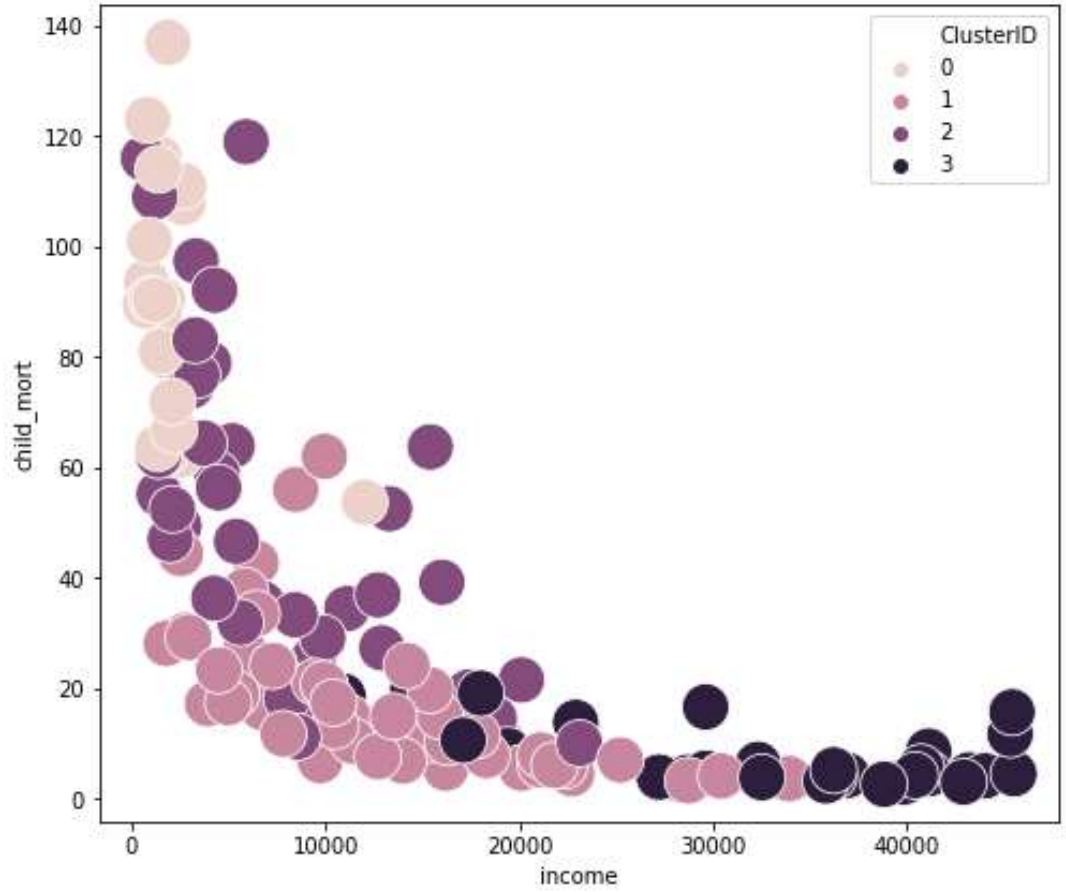
Countries falling under 4 clusters



Above plot clearly shows Four clusters with different color. In these graph poor countries are in the cluster 2 which are spread across the middle of the plot

Relation between two original variables





Most Poor countries as per Hierarchical

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
25	Burkina Faso	116.0	19.20	6.74	29.6	1430	6.81	57.9	5.87	575
26	Burundi	93.6	8.92	11.60	39.2	764	12.30	57.7	6.26	231
32	Chad	150.0	36.80	4.53	43.5	1930	6.39	56.5	6.59	897
37	Congo, Dem. Rep.	116.0	41.10	7.91	49.6	609	20.80	57.5	6.54	334
63	Guinea	109.0	30.30	4.93	43.2	1190	16.10	58.0	5.34	648
106	Mozambique	101.0	31.50	5.21	46.2	918	7.64	54.5	5.56	419
112	Niger	123.0	22.20	5.16	49.1	814	2.55	58.8	7.49	348

Conclusion for Hierarchical clustering :

we found the top 6 poor countries which are in dier need of funds from the NGO

Those countries are : Burkina Faso, Burundi, Chad, Congo, Dem. Rep.,Guinea, Mozambique,Niger

Both Analysis Conclusion (Final list of countries)

- After plotting the mean values for each original values based on cluster we found that there is a huge gape between developed countries and underdeveloped countries for gdpp,income,child_mort and life_expec.

By combining both two analysis results we can focus on below countries for fund which are in direst need of aid from NGO like HELP.

- Burundi
- Burkina Faso
- Congo Dem. Rep.
- Guinea
- Chad
- Malawi
- Sierra Leone
- Mozambique
- Zambia
- Niger

Thank you