Clustering & PCA

Ravi Edla

Info

Objective:

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.

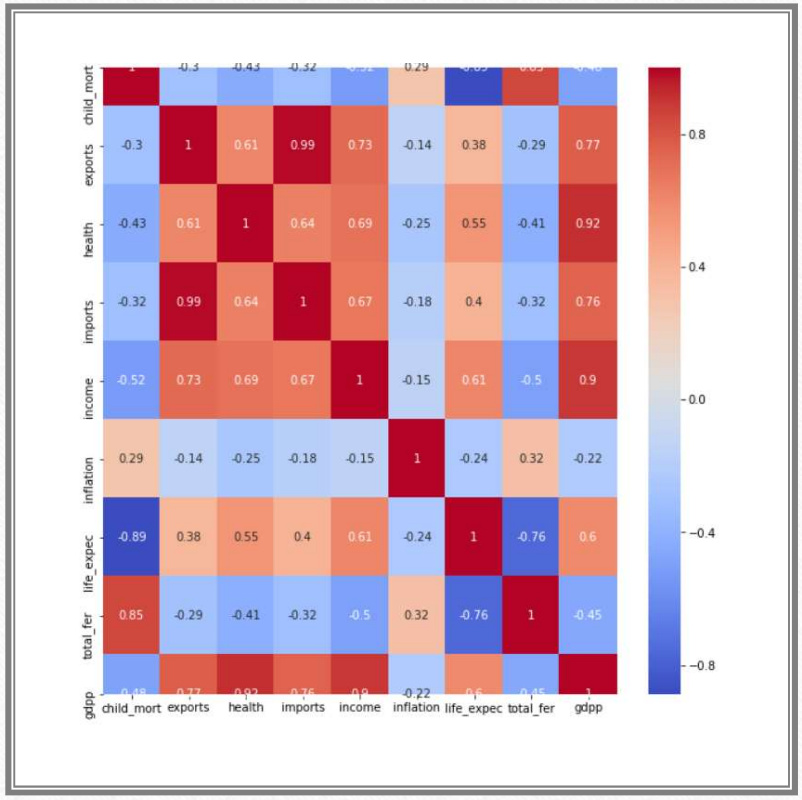
After the recent funding programs, HELP NGO has been able to raise around $ 10 million. The objective is to categorize the countries using some socio-economic and health factors that determine the overall development of the country. Then we need to suggest about the countries which the CEO needs to focus on the most.

Problem Statement:

As a data analyst, our job is to categorize the countries using some socio-economic and health factors that determine the overall development of the country. Then we need to suggest the countries list that are in the direst need of aid to the CEO who needs to focus on the most.

Analysis Methodology

- Data Collection and Cleaning
- Outlier Analysis and Removals
- Visualizing the Data
- Scaling the Data
- Principal Component Analysis on the Data
- Hopkins Statistics
- K means Clustering
- Hierarchical Clustering
- Evaluating and Decision Making



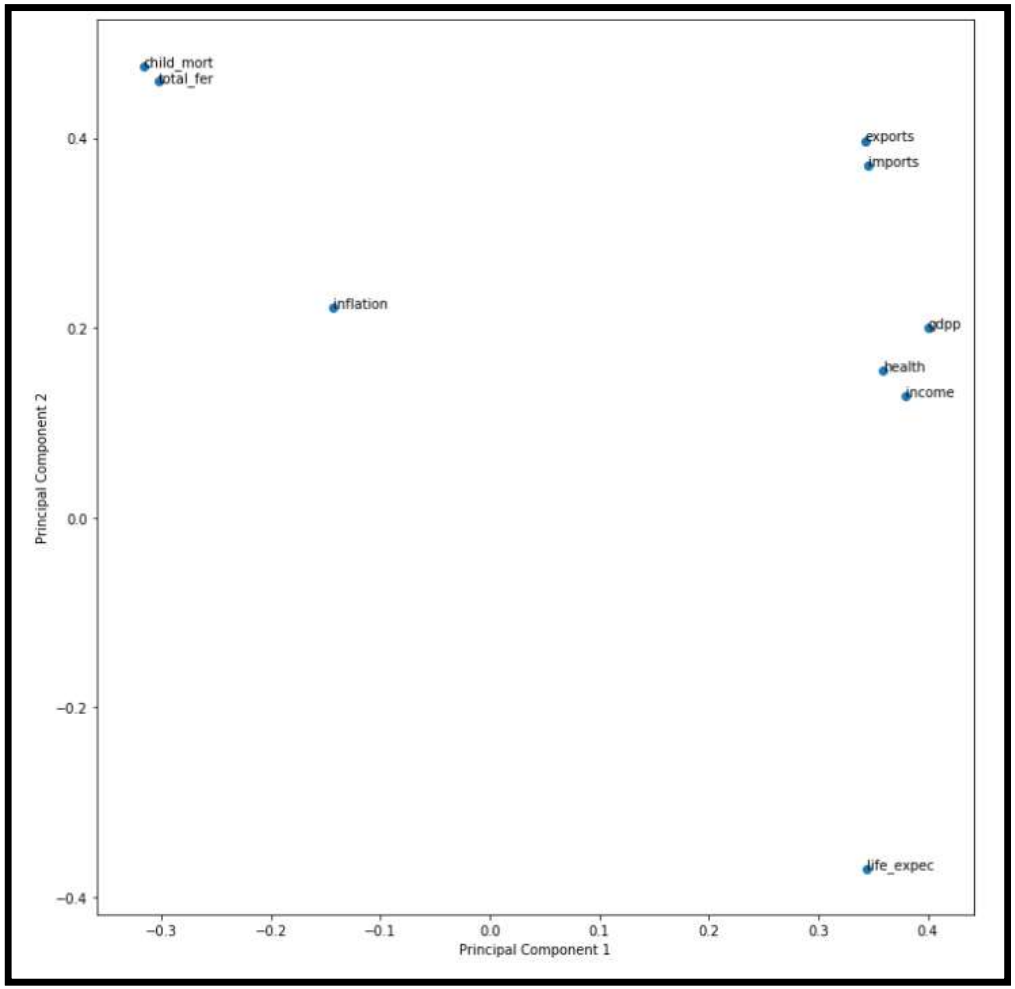
Data Correlation

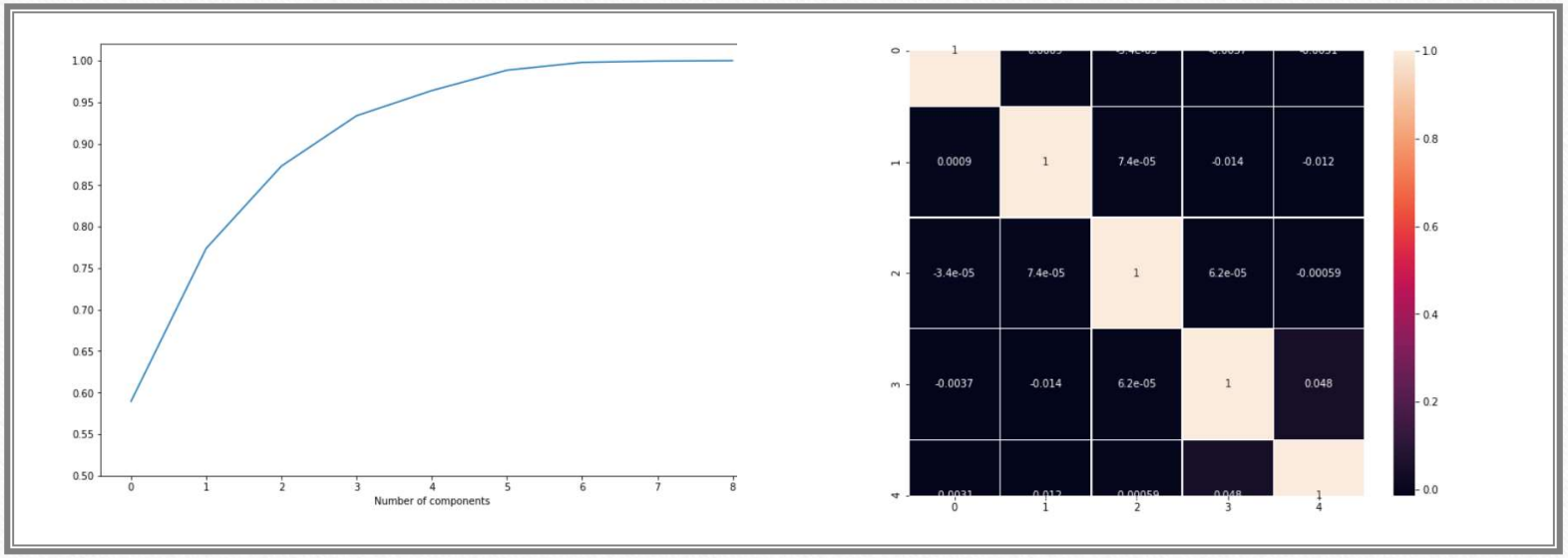
- From the heat map, we can notice that there is a high correlation between variables like (total_fertility, child_mortality), (imports, exports) and (gdpp, income).
- After cleaning the data, we removed the gdpp column as the country with high gdpp doesn't require aid as they are already doing good.
- Standardized scaling to standardize all parameters on cleaned, outlier removed data.

Principal Component Analysis

- From the adjacent PCA, we see that the features gdpp, life expectancy, Income etc; are more towards PC1.
- Features like child mortality and total fertility are towards PC2.

Visualizing features along PC1 and PC2



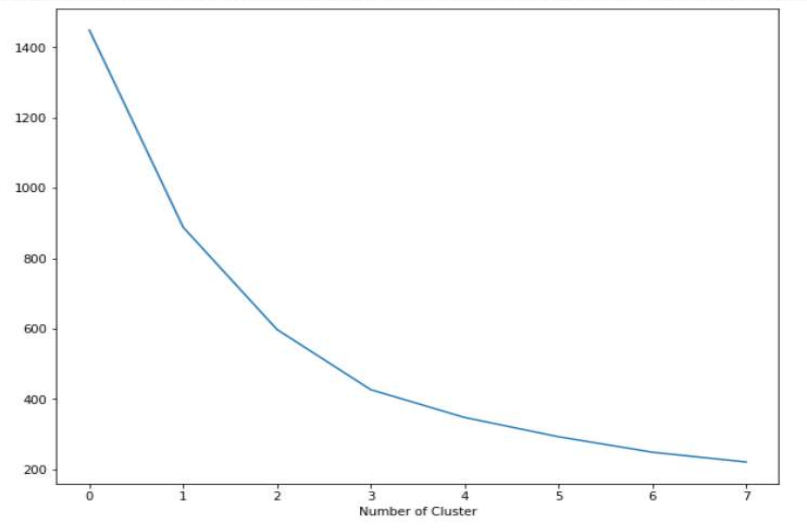


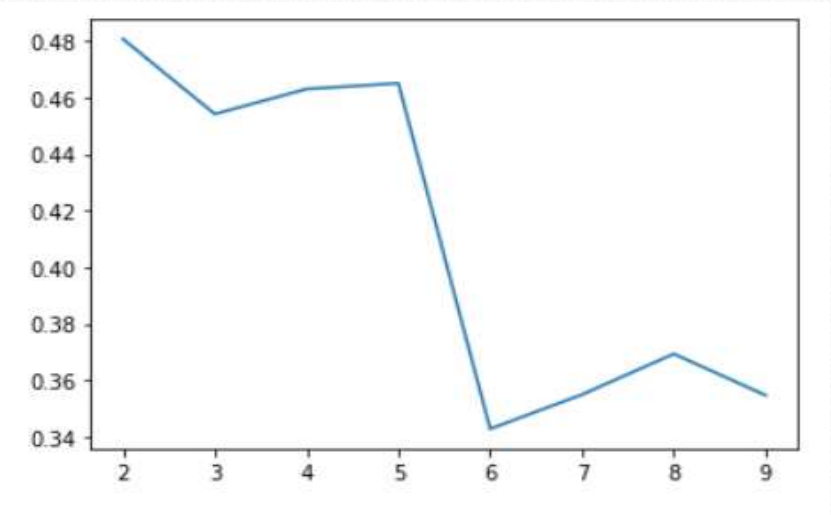
Principal Component Analysis

- Around 95% of the information is being explained by 5 components.
- From above heat map, we can see that no components are correlated with each other.

Clustering K-means

There is a distinct bend at around 3 clusters. Hence it seems a good K to choose for performing K means using K=3.

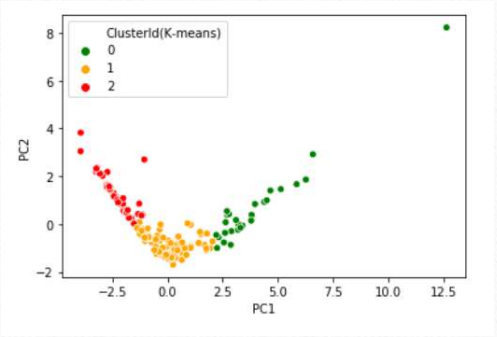


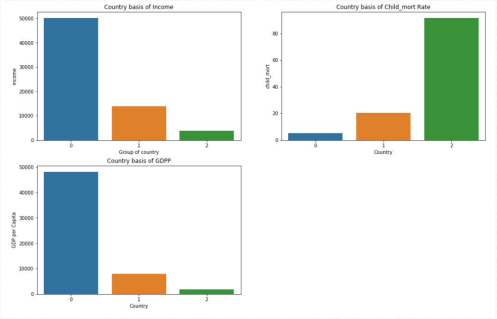


Clustering K-means

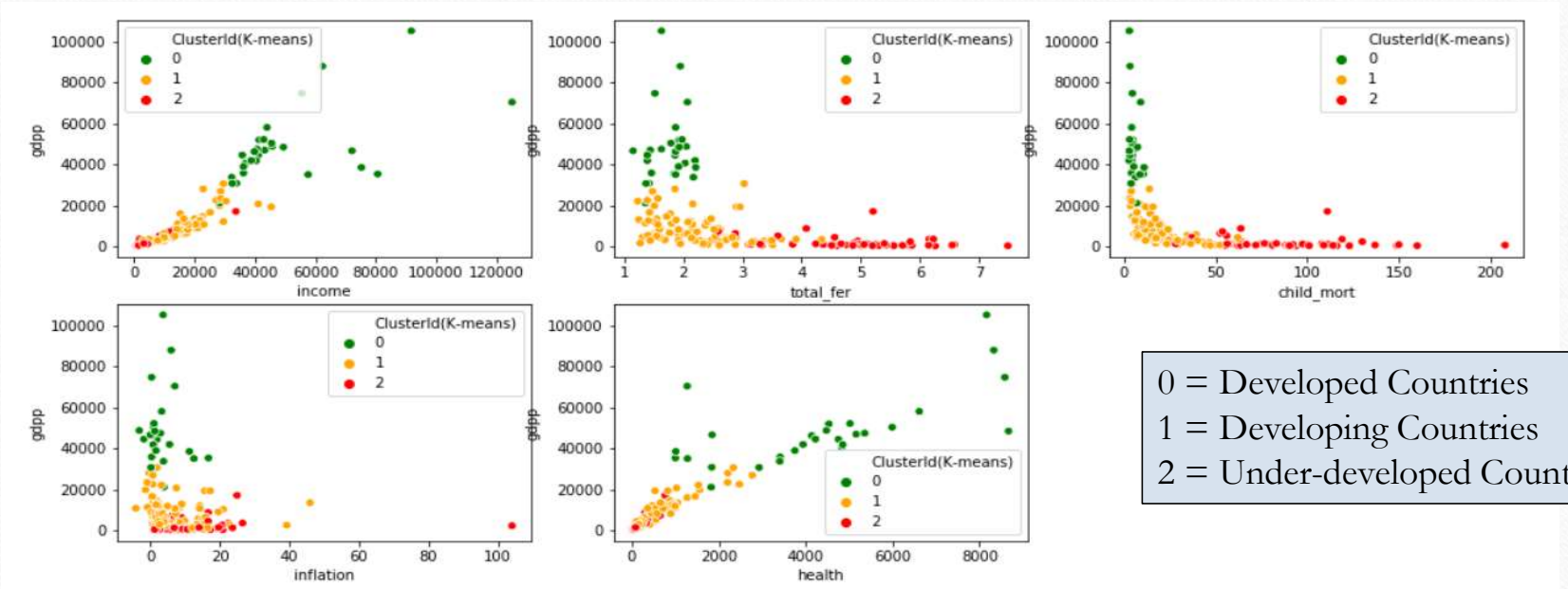
From the bar plots, we come to the conclusions that:

- Bar chart shows, all the developed countries are having high income per person, developing countries are having average income per person and poor countries are having the least income per person.
- All the developed countries are having low number in deaths of children, developing countries having average deaths and poor countries are having the most death rates.
- Developed countries are having high GDPP per capita, Developed countries are having Average GDPP per capita and the Poor countries are having low GDPP per capita.



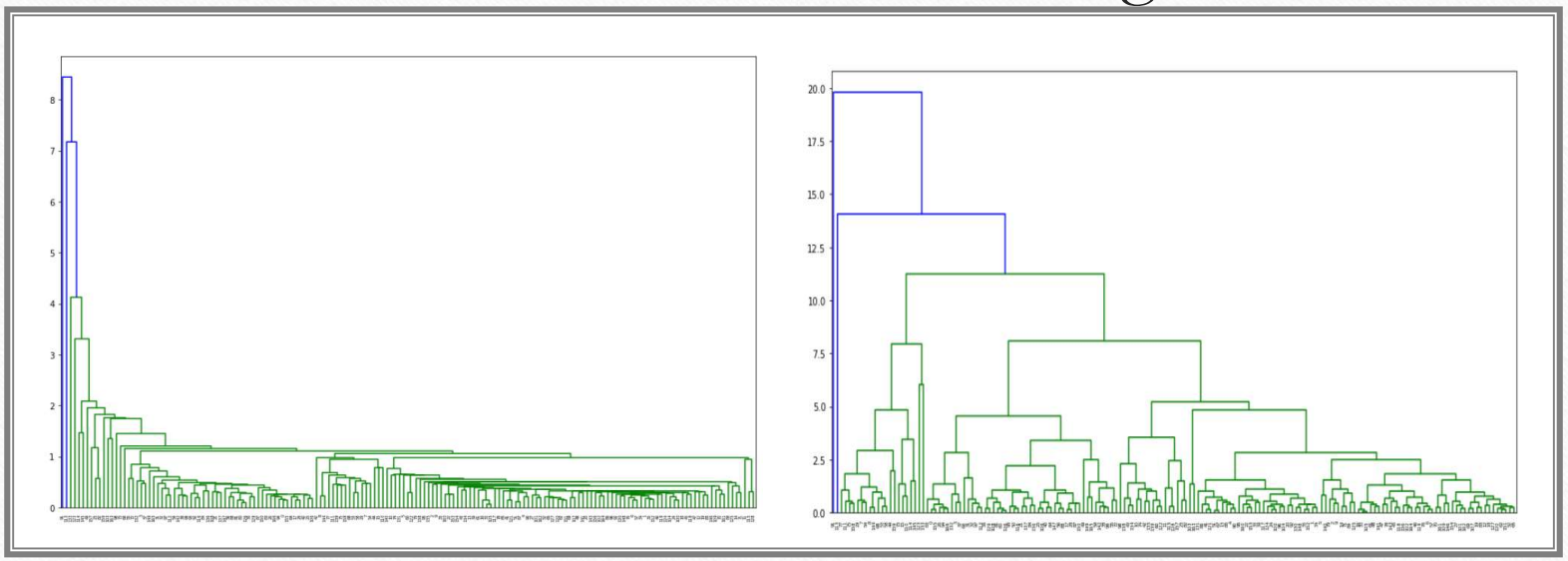


Clustering K-means



0 = Developed Countries
1 = Developing Countries
2 = Under-developed Countries

Hierarchical Clustering



Single Linkage hierarchical clustering

Complete method hierarchical clustering

By looking at these Dendograms, we are taking n-clusters as 3

Hierarchical Clustering

Hence as per the Hierarchical clustering, cluster 0 is the our main area of concern due to:

- Lowest gdpp
- Lowest income
- Highest child mortality
- Highest inflation
- Highest total fertility

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	ClusterId(K-means)	Cluster_lables(H)
17	Benin	111.0	180.4040	31.0780	281.976	1820.0	0.885	61.8	5.36	758.0	2	0
25	Burkina Faso	116.0	110.4000	38.7550	170.200	1430.0	6.810	57.9	5.87	575.0	2	0
26	Burundi	93.6	20.6052	26.7960	90.552	764.0	12.300	57.7	6.26	231.0	2	0
28	Cameroon	108.0	290.8200	67.2030	353.700	2660.0	1.910	57.3	5.11	1310.0	2	0
31	Central African Republic	149.0	52.6280	17.7508	118.190	888.0	2.010	47.5	5.21	446.0	2	0
32	Chad	150.0	330.0960	40.6341	390.195	1930.0	6.390	56.5	6.59	897.0	2	0
37	Congo, Dem. Rep.	116.0	137.2740	26.4194	165.664	609.0	20.800	57.5	6.54	334.0	2	0
40	Cote d'Ivoire	111.0	617.3200	64.6600	528.260	2690.0	5.390	56.3	5.27	1220.0	2	0
63	Guinea	109.0	196.3440	31.9464	279.936	1190.0	16.100	58.0	5.34	648.0	2	0
64	Guinea-Bissau	114.0	81.5030	46.4950	192.544	1390.0	2.970	55.6	5.05	547.0	2	0

CONCLUSION

From both the analysis, K means and Hierarchical clustering, we found that in both the analysis, the top 5 countries are same. So we are considering the analysis found in the final outcome by K means.

Hence, based on our analysis works, below are the top 5 countries which are in direst need of aid.

- Benin
- Bukrina Faso
- Burundi
- Cameroon
- Central African Republic

Reference links

- Quick commands help:
  - www.hackerrank.com
  - www.geeksforgeeks.org
  - https:/pandas.pydata.org/pandas-docs
  - https:/stackoverflow.com/