PCA & Clustering Assignment

Rahul Kumar

rahul.cs068@gmail.com

+91-9538109454

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.

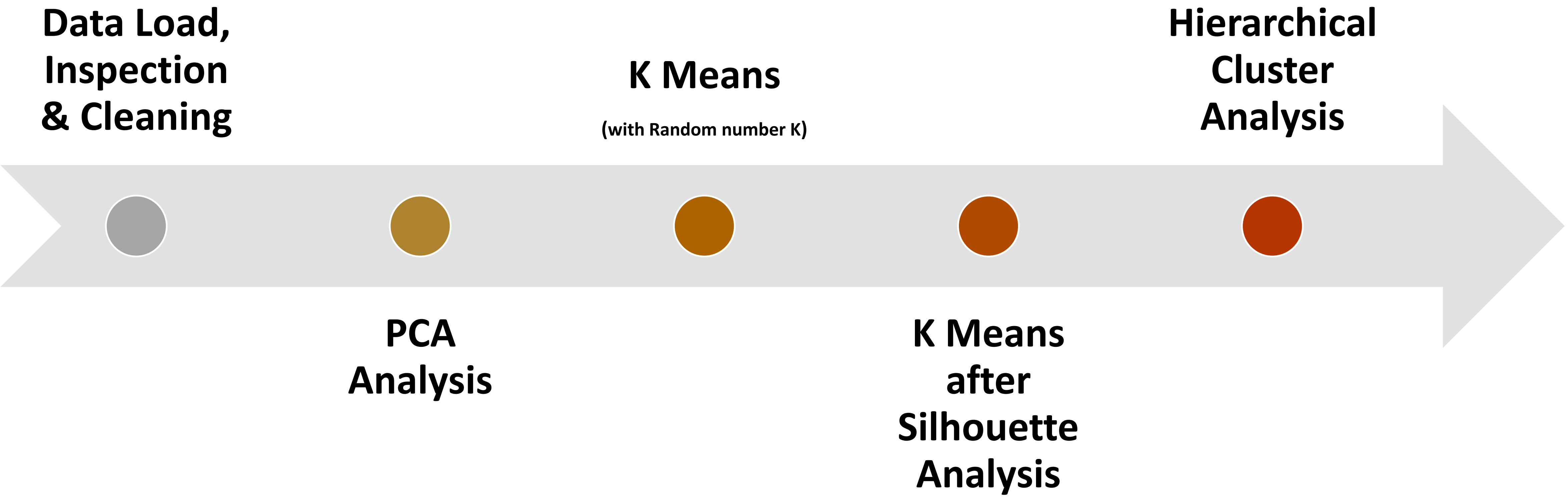
After the recent project that included a lot of awareness drives and funding programmes, they have been able to raise around $ 10 million. Now the CEO of the NGO needs to decide how to use this money strategically and effectively. The significant issues that come while making this decision are mostly 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.

Objectives

Your main task is to cluster the countries by the factors mentioned above and then present your solution and recommendations to the CEO using a PPT. You are also supposed to use dimensionality reduction using PCA to get the visualisations of the clusters in a 2-D form.

Problem solving methodology



Data Load, Inspection & Cleaning

PCA Analysis

K Means

(with Random number K)

K Means after Silhouette Analysis

Hierarchical Cluster Analysis

Principal Component Analysis (PCA)

- Before we do clustering, it is required to do PCA on the data to see whether the co-related factors can be removed and use only non-co-related data for the analysis. Fig 1 shows the co-relation between the PCA Components and they are not co-related.
- Number of factors given in the problem are 9 and we could reduce this to 4 using PCA as these 4 PCA's explains 87% of the entire data.
  - Child_mort
  - Exports
  - Health
  - Imports
- This PCA data is used in K means and Hierarchical Clustering Algorithm.
- IQR method is used to remove outliers from the PCA dataset.
- HopKins Analysis is used to check whether the clustering is required for the data are not and we found that HopKins value is greater than 0.7 and data require clustering.

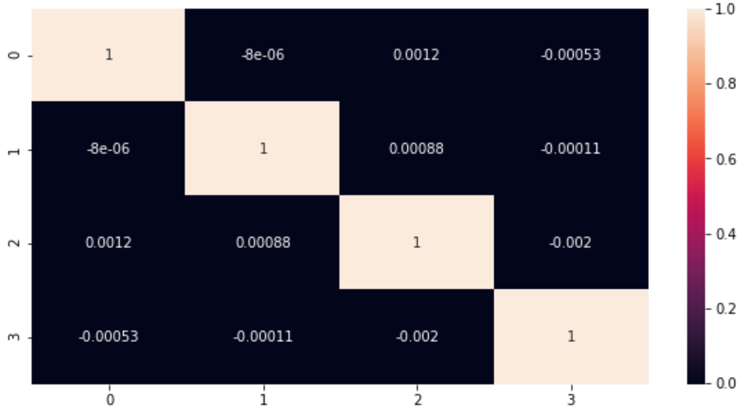


Fig 1

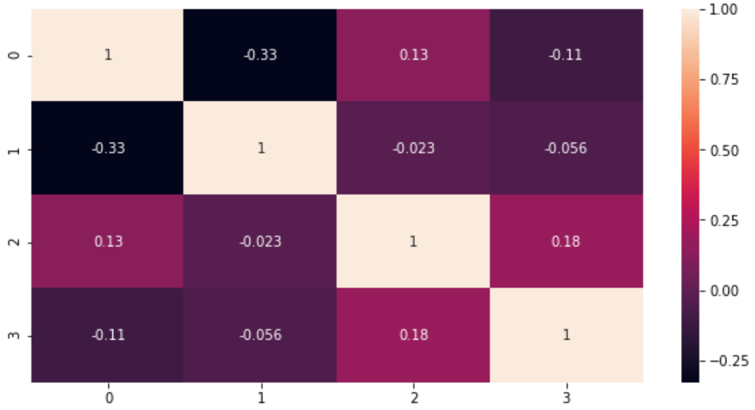


Fig 2

K-means clustering with random K

- We choose a random K to do clustering of the data. We use the libraries form sklearn to categorize the countries.
- Randomly 4 is chosen for K and countries are clustered among these 4 categories. Refer to Fig 3
- To verify whether the chosen K is right or not we used Silhouette analysis which gave optimal k as 3. Refer to Fig 4

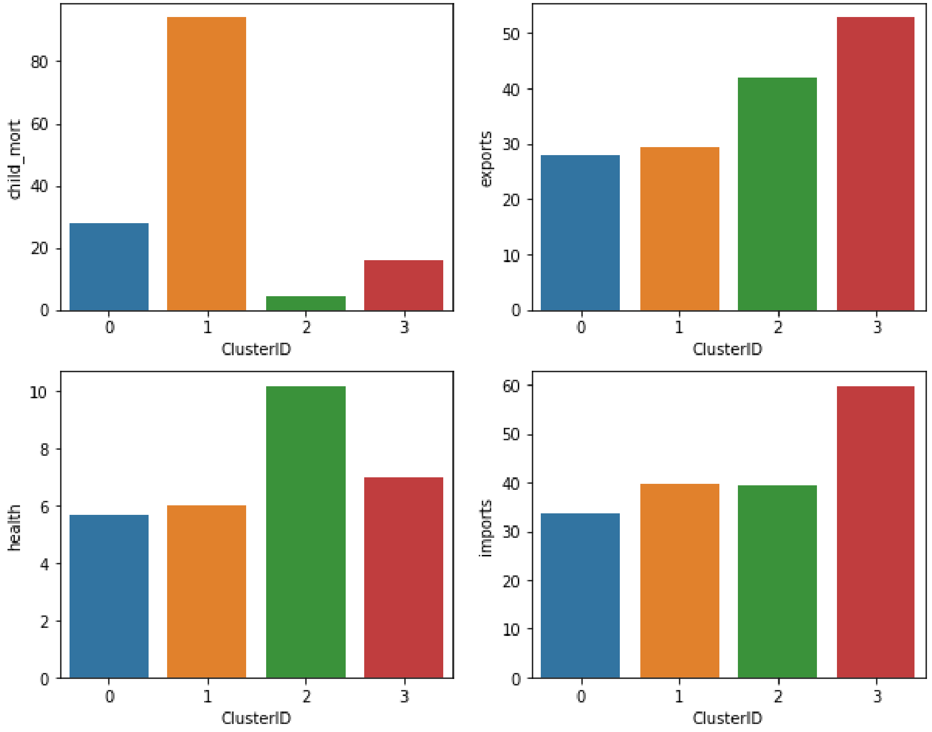


Fig 3

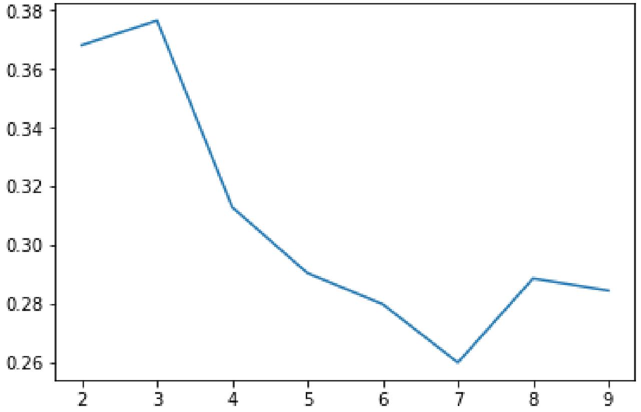
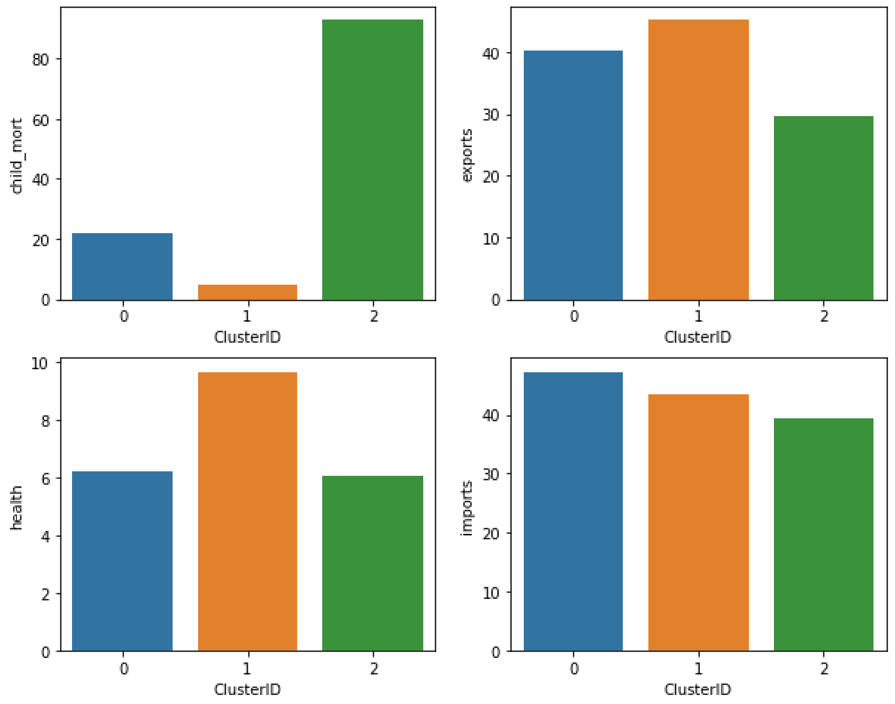


Fig 4

K-means clustering with K=3

- We choose a K = 3 from Silhouette Analyss and do the clustering of the data.
- Countries are categorized as 3 clusters.

Cluster	Number of Countries	Description
2	28	Socio-Economically Poor
0	78	Socio-Economically Average
1	43	Socio-Economically Strong



Countries – Socio Economical Poor and need help

Botswana	Iraq	Sudan
Burkina Faso	Kenya	Tanzania
Burundi	Lao	Timor-Leste
Cameroon	Madagascar	Togo
Central African Republic	Malawi	Uganda
Chad	Mali	Yemen
Comoros	Mauritania	Zambia
Congo Dem. Rep.	Mozambique	
Congo Rep.	Namibia	
"Cote dIvoire"	Niger	
Equatorial Guinea	Pakistan	
Eritrea	Rwanda	
Gabon	Senegal	
Gambia	Sierra Leone	
Ghana	South Africa	

Countries – Socio Economical Average and need small help

Australia	Israel
Austria	Italy
Bahamas	Japan
Belgium	Netherlands
Canada	New Zealand
Cyprus	Norway
Czech Republic	Portugal
Denmark	Slovak Republic
Finland	Slovenia
France	South Korea
Germany	Spain
Greece	Sweden
Iceland	Switzerland
Ireland	United Kingdom

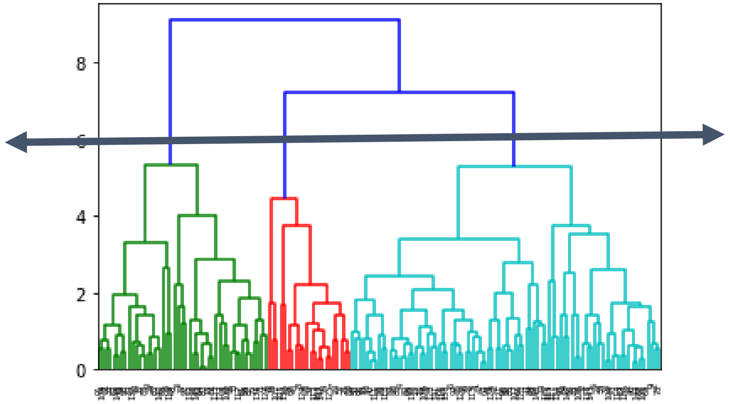
Countries – Socio Economical Strong, NO help req.

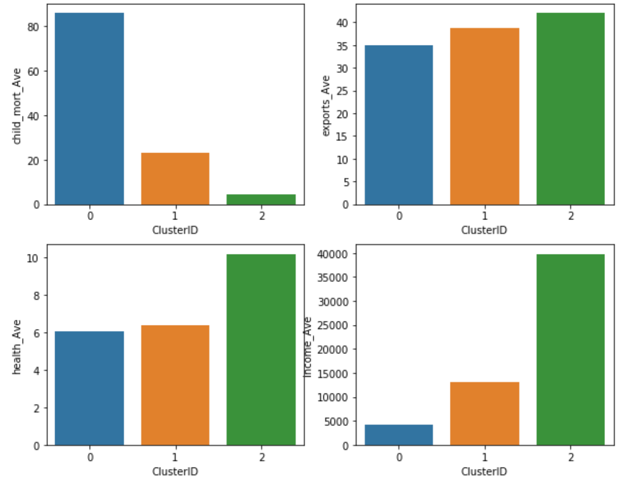
- Albania
- Algeria
- Antigua and Barbuda
- Argentina
- Armenia
- Azerbaijan
- Bahrain
- Bangladesh
- Barbados
- Belarus
- Belize
- Bhutan
- Bolivia
- Bosnia and Herzegovina
- Brazil
- Bulgaria
- Cambodia
- Cape Verde
- Chile
- China
- Colombia
- Costa Rica
- Croatia
- Dominican Republic
- Ecuador
-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

Hierarchical Clustering

- Dendrogram with method = ‘complete’ is used build the tree and we used Divisive clustering to cut the tree.
- Using HC method also the countries are clustered.

Cluster	Number of Countries	Description
2	22	Socio-Economically Poor
1	82	Socio-Economically Average
0	45	Socio-Economically Strong





Countries – Socio Economical Poor and need help

- Afghanistan
- Angola
- Benin
- Bhutan
- Botswana
- Burkina Faso
- Burundi
- Cambodia
- Cameroon
- Central African Republic
- Chad
- Comoros
- Congo Dem. Rep.
- Congo Rep.
- "Cote dIvoire"
- Equatorial Guinea
- Fiji
- Gambia
- Ghana
- Guinea
- Guinea-Bissau
- Guyana
- Haiti
- Iraq
- Kenya
- Kyrgyz Republic
- Lao
- Madagascar
- Malawi
- Mali
- Mauritania
- Mozambique
- Namibia
- Niger
- Senegal
- Sierra Leone
- Solomon Islands
- South Africa
- Tajikistan
- Tanzania
- Togo
- Turkmenistan
- Uganda
- Vanuatu
- Zambia

Countries – Socio Economical Average and need small help

- Albania
- Algeria
- Antigua and Barbuda
- Argentina
- Armenia
- Azerbaijan
- Bahamas
- Bahrain
- Bangladesh
- Barbados
- Belarus
- Belize
- Bolivia
- Bosnia and Herzegovina
- Brazil
- Bulgaria
- Cape Verde
- Chile
- China
- Colombia
- Costa Rica
- Croatia
- Cyprus
- Czech Republic
- Dominican Republic
- Ecuador
- Egypt
- El Salvador
- Eritrea
- Estonia
- Gabon
- Georgia
- Grenada
- Guatemala
- Hungary
- India
- Indonesia
- Iran
- Jamaica
- Jordan
- Kazakhstan
- Latvia
- Lebanon
- Libya
- Lithuania
- Macedonia FYR
- Malaysia
- Maldives
- Mauritius
- Moldova
- Mongolia
- Montenegro
- Morocco
- Myanmar
- Nepal
- Pakistan
- Panama
- Paraguay
- Peru
- Philippines
- Poland
- Romania
- Russia
- Rwanda
- Samoa
- Serbia
- Slovak Republic
- Slovenia
- South Korea
- St. Vincent and the Grenadines
- Sudan
- Suriname
- Thailand
- Timor-Leste
- Tonga
- Tunisia
- Turkey
- Ukraine
- Uruguay
- Uzbekistan
- Vietnam
- Yemen

Countries – Socio Economical Strong, NO help reqr.

Australia	Israel
Austria	Italy
Belgium	Japan
Canada	Netherlands
Denmark	New Zealand
Finland	Norway
France	Portugal
Germany	Spain
Greece	Sweden
Iceland	Switzerland
Ireland	United Kingdom

Thank You