HELP International NGO

PCA & Clustering Analysis

Submitted by-

Gaurav Soni

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. The significant issues that come while making this decision are mostly related to choosing the countries that are in the direst need of aid.

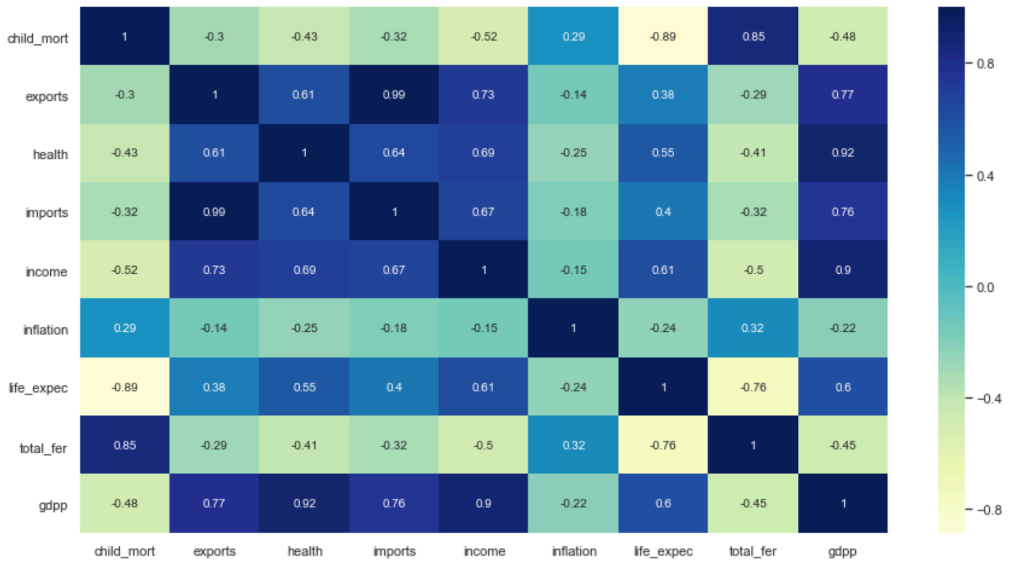
- The analysis is to categorise the countries using some socio-economic and health factors that determine the overall development of the country. Then you need to suggest the countries which the CEO needs to focus on the most. The datasets containing those socio-economic factors and the corresponding data dictionary are provided below.

Business & Data Understanding

- Total Corpus available for investment is $**10** Million
- Data available for around **167** countries
- Key Attributes of the Datasets are
  - **Country** - Name of the Country
  - **Child_Mor** - 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
  - **Imports** - Imports of goods and services. Given as %age of the 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.

Business & Data Understanding

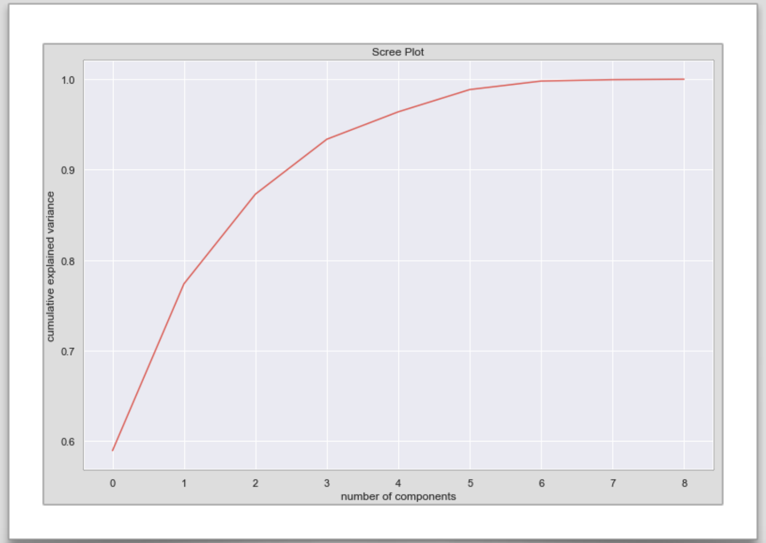
- Since the Attributes exports, health and imports are given in percentage % of the GDPP , we converted them into actual values for the ease of the analysis
- Outliers may affect the analysis and few of the algorithms are sensitive to the outliers , so initially we removed the outliers ,but later we realised that 25% of the data is removed and hence the dataset became skewed .Hence we have taken decision to keep the outliers as it is in the data.
- The data is really good with no nulls and NaN values which makes analysis more easier.
- No duplicate countries are found in the dataset.
- Few Attributes are highly correlated with each other which does make sense like child_mort is highly correlated with Total_Fer

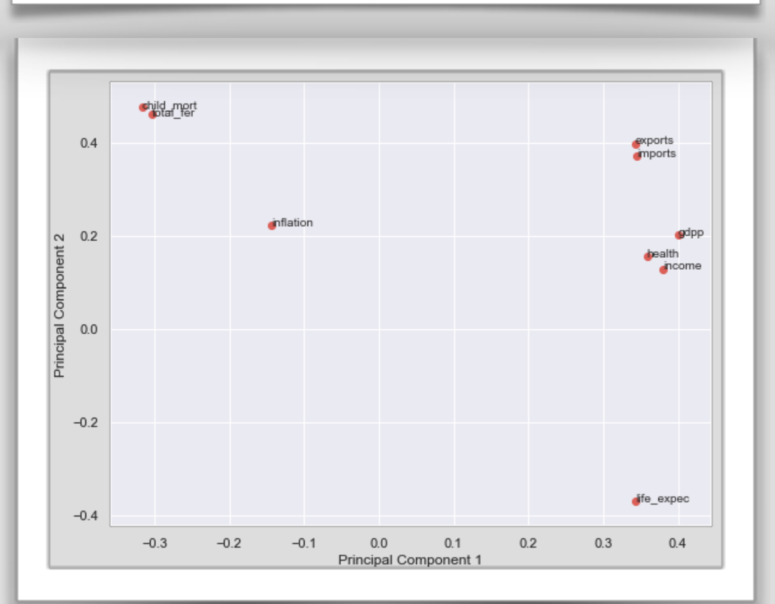


	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
child_mort	1	-0.3	-0.43	-0.32	-0.52	0.29	-0.89	0.85	-0.48
exports	-0.3	1	0.61	0.99	0.73	-0.14	0.38	-0.29	0.77
health	-0.43	0.61	1	0.64	0.69	-0.25	0.55	-0.41	0.92
imports	-0.32	0.99	0.64	1	0.67	-0.18	0.4	-0.32	0.76
income	-0.52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

Principle Component Analysis

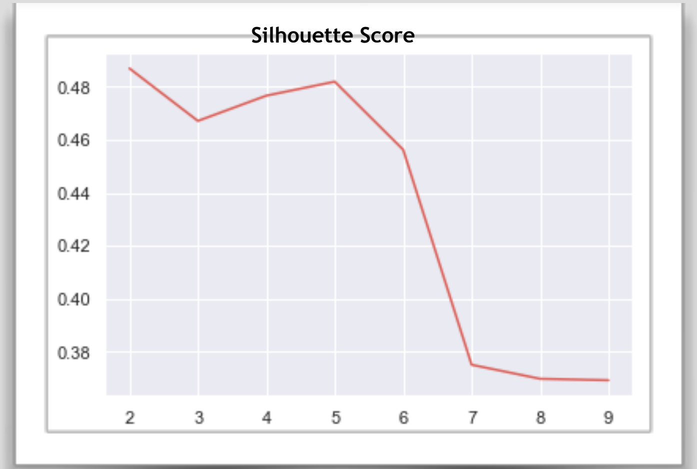
- The attributes are highly correlated and hence dimensionality reduction is required on the dataset ,so we used Principle Component Analysis before going ahead with the Clustering.
- On Plotting the Scree Plot, Cumulative Variance against the number of components ,we came To the conclusion that 4 PCE can cover almost 93% of the variance ,Please check the graph.
- After done with the PCE ,PC1 have income, gdpp, export, imports , life expectancy and health whereas PC2 points to child_mort and total_fertility
- Now on this reduced set we will be performing clustering .
  - K- Mean
  - Hierarchical Clustering

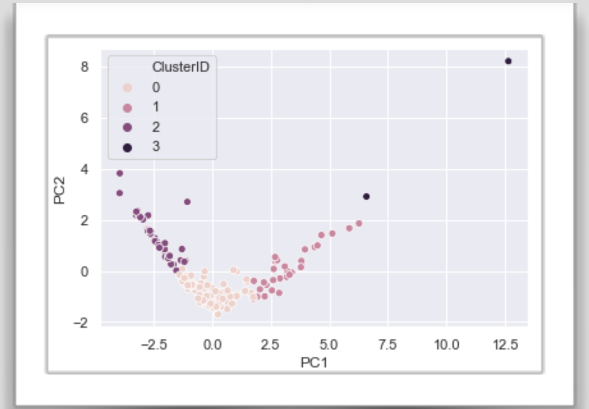


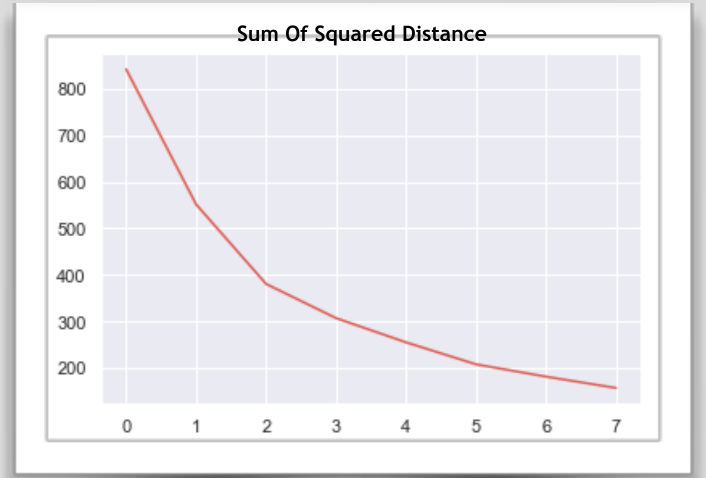


Clustering - (K-means)

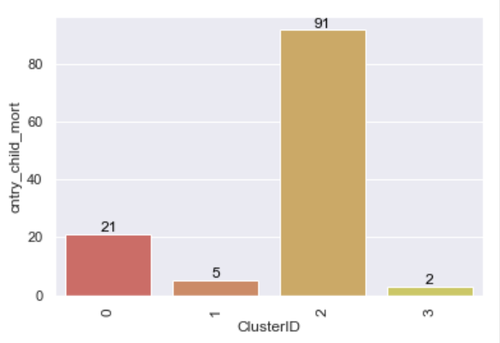
- The first step of clustering is to find out the number of clusters ,hence Silhouette Score and Sum of Squared Distance is been calculated
- It is found with the Silhouette , the K =5 or 4 is good to start.
- We have done clustering with both K=4 or 5 ,but K=4 gives us better result.
- Merged the Clustered data with the original dataset and tried to find the pattern using the plots.

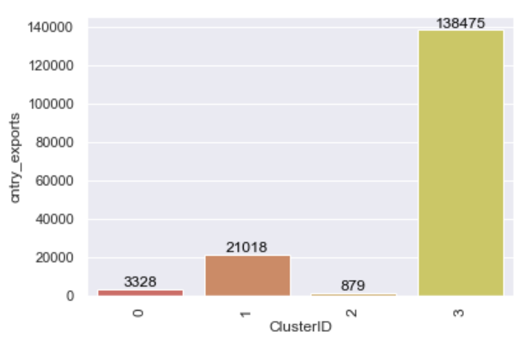


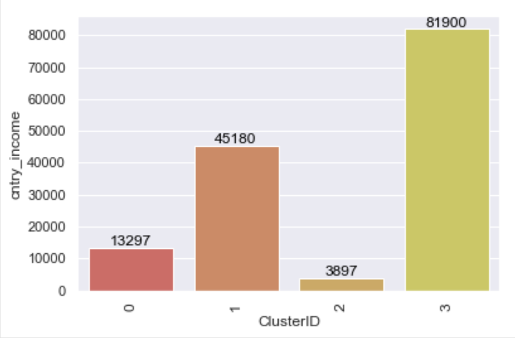


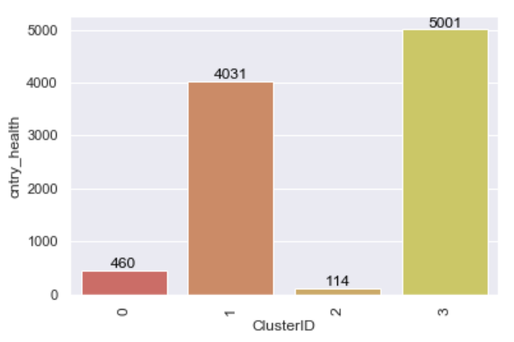


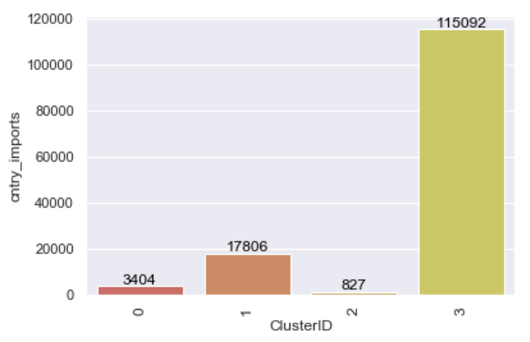
Plots for K-means Clustering

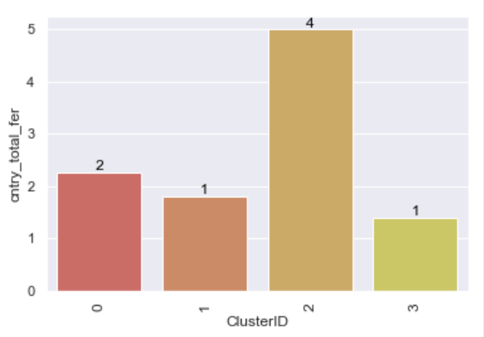


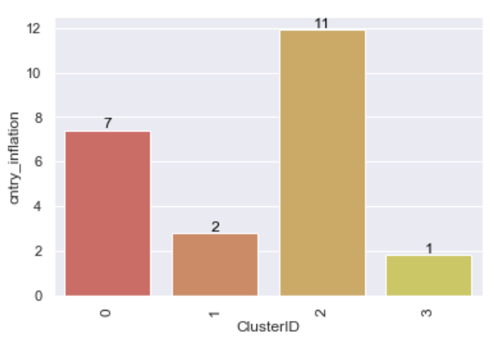


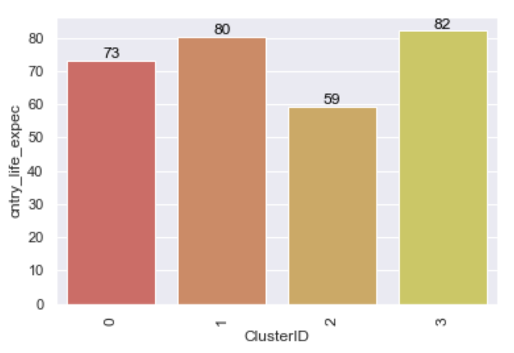


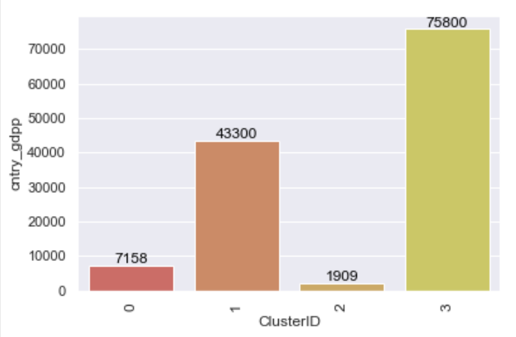












Cluster Analysis K-means

- Looking to the plots in the previous slide
  - Cluster 2 has the highest Child Mortality Rate ,whereas Cluster 3 has the lowest.
  - Cluster 2 has the lowest Exports for goods and services whereas Cluster 3 has the highest.
  - Cluster 2 countries spend less in the Health ,whereas Cluster 3 the highest.
  - Cluster 2 has the lowest imports whereas 3 has the highest
  - Cluster 2 has the lowest income per capita whereas 3 has the highest
  - Cluster 2 has the highest inflation whereas 3 has the lowest
  - Cluster 2 has the lowest life expectancy whereas 3 has the most
  - Cluster 2 has the highest Total Fertility where as 3 and 1 has the lowest
  - Cluster 2 has the lowest GDPP ,whereas 3 has the highest
  - Total Countries in the Cluster 2 is 48.
  - The conclusion from the above stats is Cluster 2 countries are the top contender for the Aids from HELP International as per K-means algo.

Clustering Hierarchical

- We created the Dendrogram and cut at an to obtain 4 clusters.
- Attach the result with the original data that contains the PCE’s
- Plots has been created based on the results
- The result from the Hierarchical is almost the same as the K- means clustering ,hence took the intersect of the countries from the K-means dataset and the Hierarchical cluster with the list of countries which needs aid

Disadvantage of PCA

- **Independent variables become less interpretable:** After implementing PCA on the dataset, the original features will turn into Principal Components. Principal Components are the linear combination of your original features. Principal Components are not as readable and interpretable as original features.
- **Data standardization is must before PCA:** We must standardize your data before implementing PCA, otherwise PCA will not be able to find the optimal Principal Components.
- **Information Loss:** Although Principal Components try to cover maximum variance among the features in a dataset, if we don't select the number of Principal Components with care, it may miss some information as compared to the original list of features.

List of Countries shortlisted after Clustering for AID

1. Afghanistan
2. Angola
3. Benin
4. Botswana
5. Burkina Faso
6. Burundi
7. Cameroon
8. Central African Republic
9. Chad
10. Comoros
11. Congo, Dem. Rep.
12. Congo, Rep.
13. Cote d'Ivoire
14. Equatorial Guinea
15. Eritrea
16. Gabon
17. Gambia
18. Ghana
19. Guinea
20. Guinea-Bissau
21. Haiti
22. Iraq
23. Kenya
24. Kiribati
25. Lao
26. Lesotho
27. Liberia
28. Madagascar
29. Malawi
30. Mali
31. Mauritania
32. Mozambique
33. Namibia
34. Niger
35. Nigeria
36. Pakistan
37. Rwanda
38. Senegal
39. Sierra Leone
40. Solomon Islands
41. South Africa
42. Sudan
43. Tanzania
44. Timor-Leste
45. Togo
46. Uganda
47. Yemen
48. Zambia

List of 5 countries in direst need

From our Analysis we can conclude that below features are important for any country growth

1. Income
2. Life Expectancy
3. GDPP

Based on this we comes to the conclusion that below countries require immediate Aids for the survival

1. Congo, Dem.Rep.
2. Liberia
3. Burundi
4. Niger
5. Central Africa Republic