Clustering & PCA Assignment

Submitted by :

Deepa Rajesh

Help International NGO Case Study

Problem Statement :

HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries. After the recent funding programmes, they have been able to raise around $ 10 million.

The NGO needs to decide how to use this money strategically and effectively by choosing the countries that are in the direst need of aid.

Business Objective :

The given data needs to be analyzed such that the countries are categorized based on some socio-economic and health factors that determine the overall development of the country

Strategy:

- Source the data for analysis
- Clean and Prepare the data.
- Identify the principal components.
- Identify the clusters and visualize the clusters using principal components
- Identify the clusters that need more attention
- Visualize the clusters for few parameters
- Identify 5 countries which are in dire need of aid

Problem solving methodology

Data Sourcing , Cleaning and Preparation

- Read the Data from Source
- Convert data into clean format suitable for analysis
- Remove duplicate data
- Outlier Treatment
- Feature Standardisation.

Principal Component Analysis

- Identify the principal components using available techniques.
- Determine that there is very little or no correlation amongst the principle components

Clustering

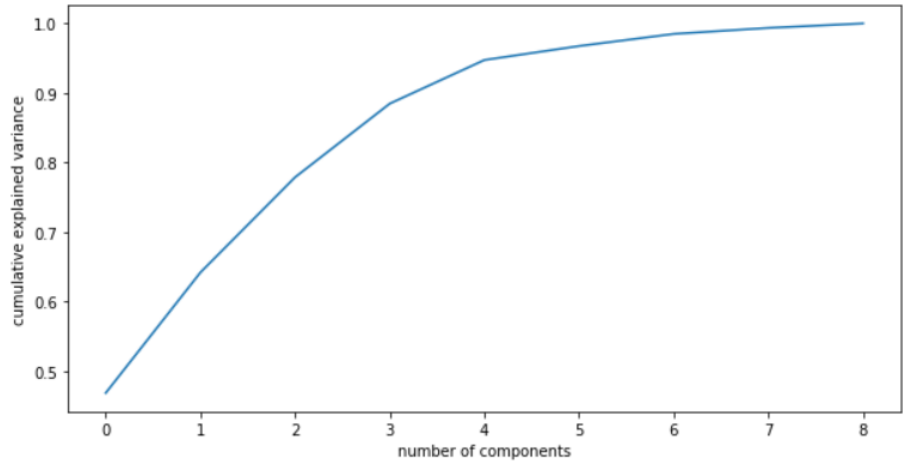
- Determine how well the data can be clustered using the Hopkins Statistics technique
- Apply clustering using the following :
  - K Means
  - Hierarchical
- Determine number of clusters using the techniques available.
- Analyze the clusters formed

Result

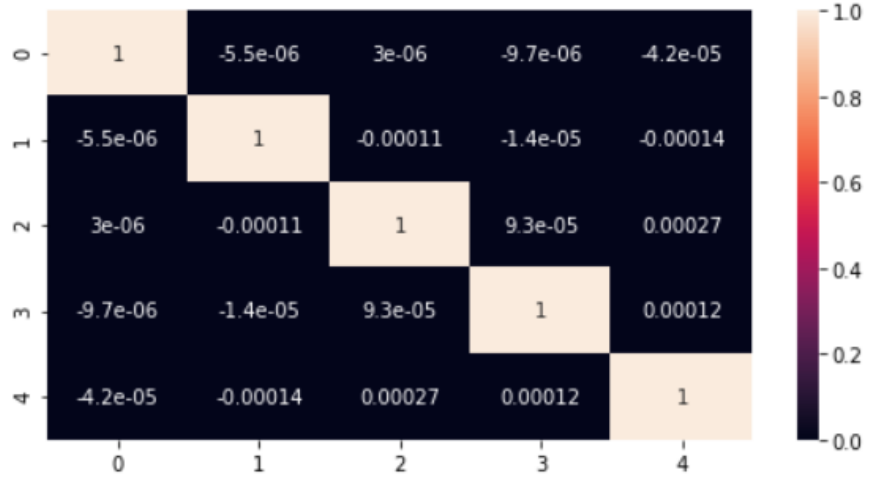
- Identify how the original variables vary for each cluster of countries.
- Visualization of clusters for principal components and also for original variables.
- Identify list of 5 countries that are in dire need of aid

Principal Component Analysis

The graph below helps to determine that out of 9 original components, approximately 5 components can be used for analysis:



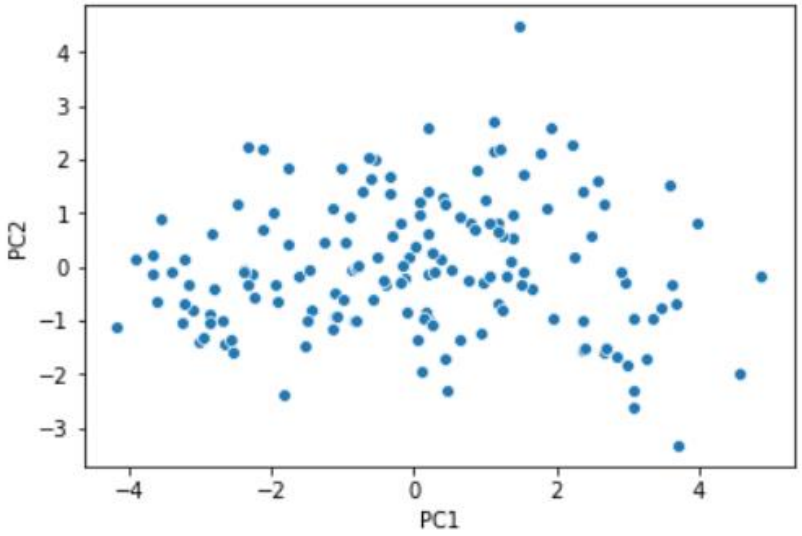
Correlation between the 5 components is almost 0



Following are the 5 principal components that would be used in analysis since they contribute to 95% variance

child_mort	0.47
exports	0.17
health	0.14
imports	0.11
income	0.06

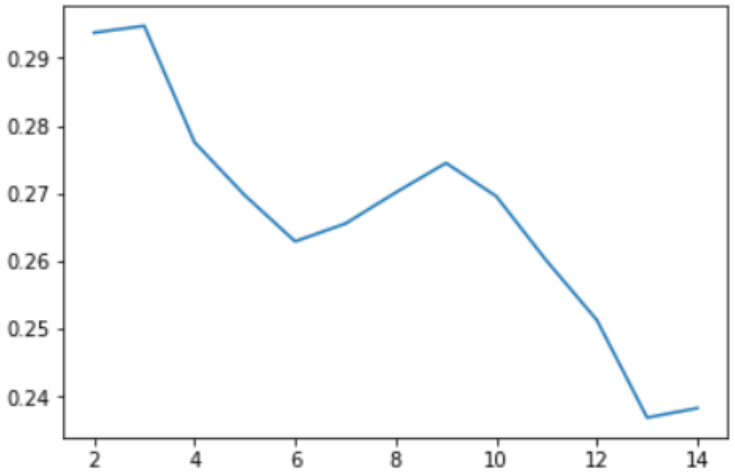
At this stage the scatter plot does not display any meaningful cluster. However, a closer look at the screen below helps in understanding that atleast a minimum of 3 clusters would be needed



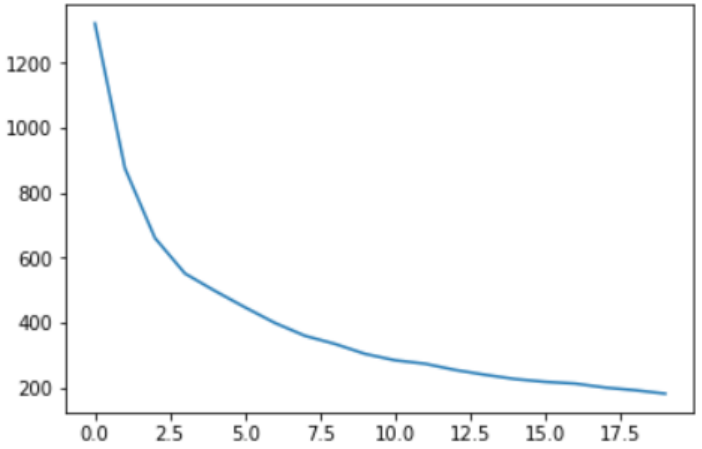
Clustering and Visualisation of Clusters for Principal Components

- The Silhouette Analysis and Elbow Curve Method graphs below both indicate an existence of approximately 3 clusters.
- The Principle components clusters scatter plot below of the first 2 principle components show a clear existence of 3 clusters

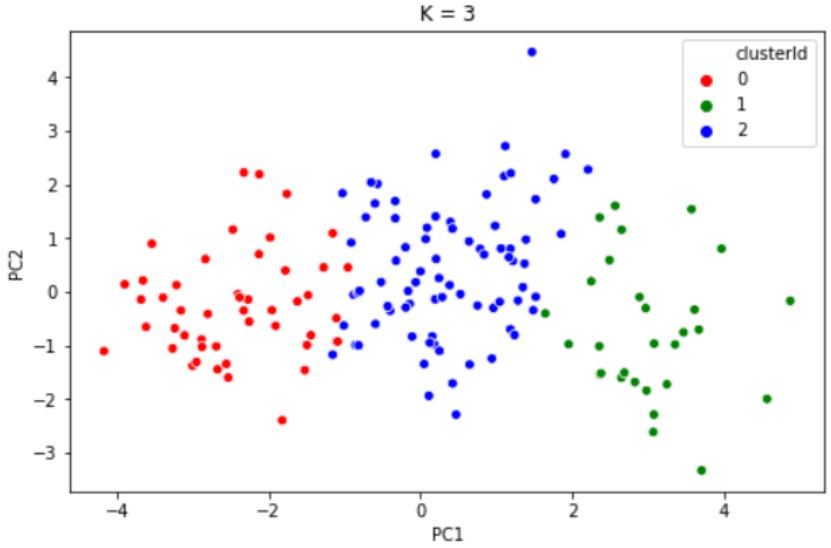
Silhouette Analysis



Elbow Curve Method

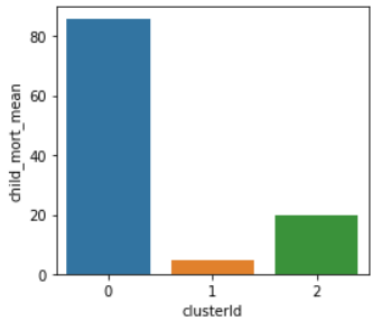


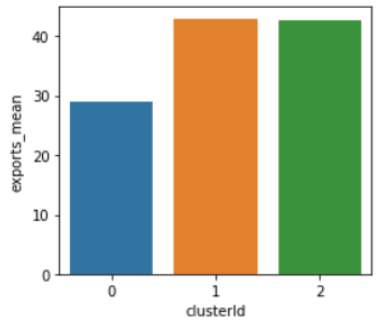
Principal Components Clusters

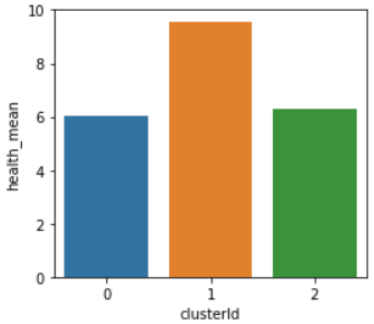


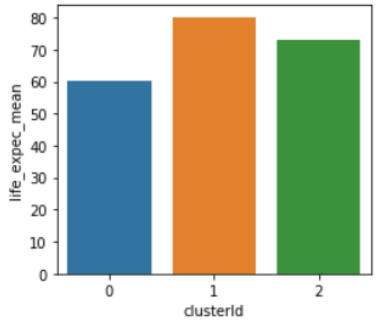
Analysis on how clusters vary for different variables

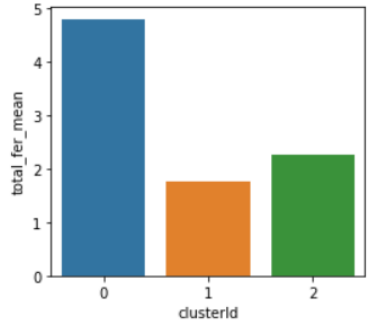
- In the below plot, cluster 0 represents underdeveloped countries. cluster 1 represents developed countries and cluster 2 represents developing countries
- For developed countries, exports, spending on health, income per person, life expectancy and gdpp are high as compared to that of underdeveloped countries.
- For developed countries, child mortality rate, imports, inflation and total fertility is low when compared to underdeveloped countries.
- Hence it can be determined that for underdeveloped countries more attention needs to be given for child mortality, income, increase in gdpp and more spending on health. This would in turn decrease child mortality and also help in increasing life expectancy. Also enhancing the increase in income per person would enable the population to spend money on their health when required.

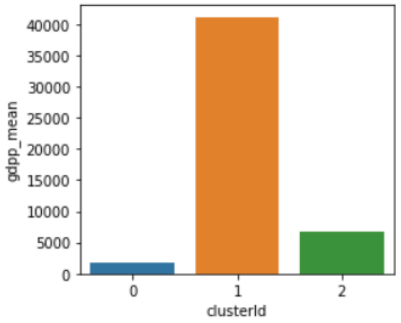


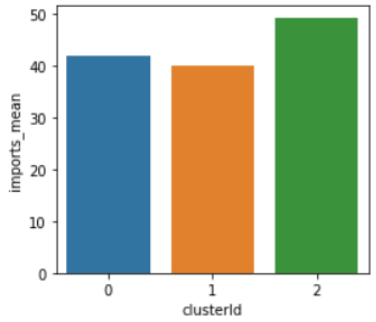


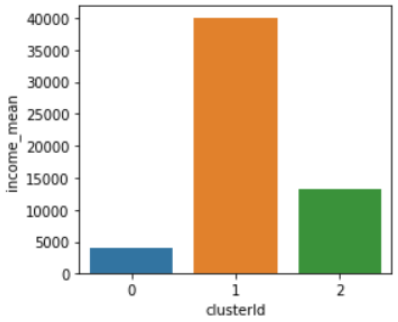


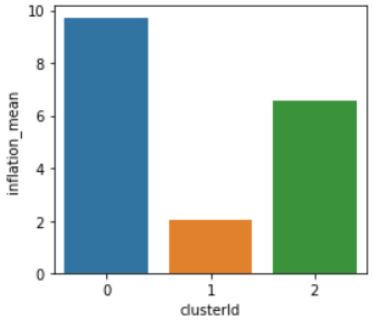










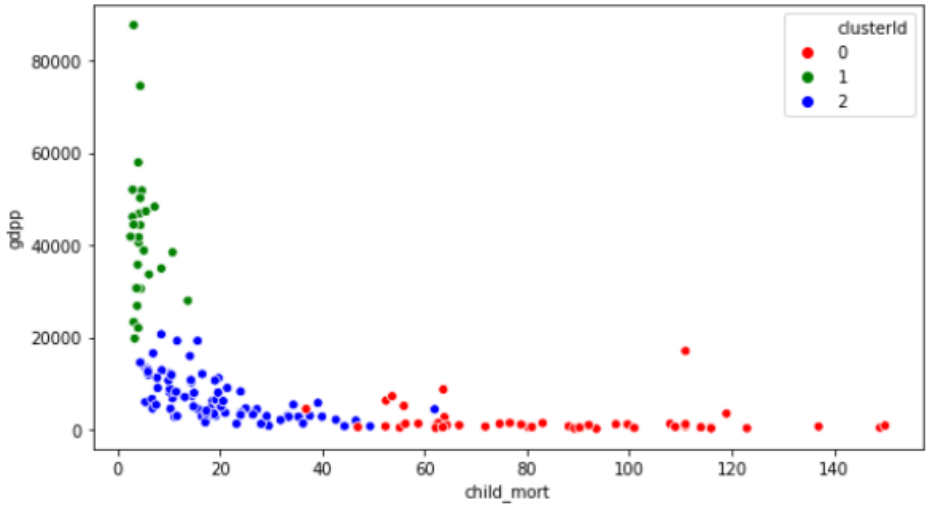


Cluster 0 → Underdeveloped Countries
Cluster 1 → Developed Countries
Cluster 2 → Developing Countries

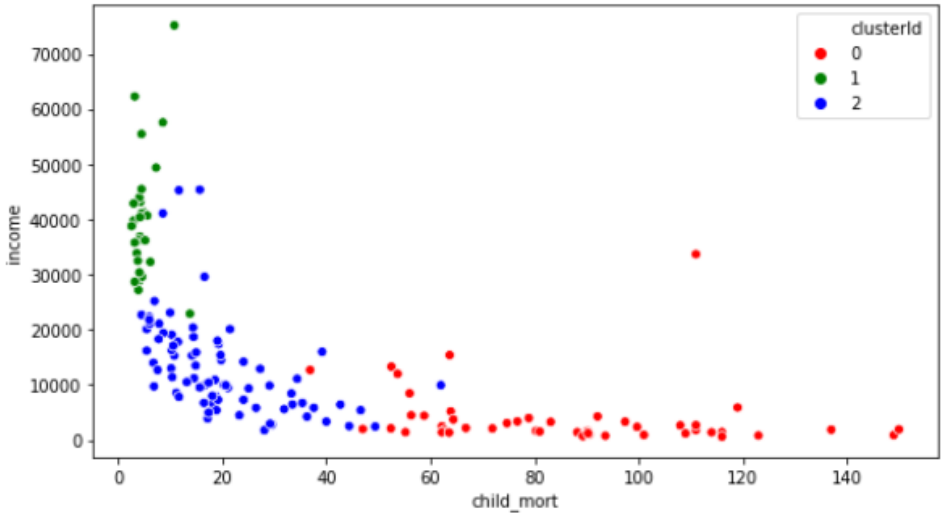
Visualisation of Clusters for parameters that need more attention

- Inferences drawn from the graphs below :
  - In the graph Child Mortality Vs gdpp, it can be noticed that the developed countries are more aligned to high gdpp and under-developed countries are more aligned towards high child mortality, thereby indicating that under-developed countries have more child mortality rate.
  - In the graph Child Mortality Vs income, it can be noticed that the developed countries are more aligned to high income and the under-developed countries again are more aligned towards high child mortality
- Hence in under-developed countries, child mortality is a greater cause for concern and thus more attention needs to be given on that.

Child Mortality Vs gdpp



Child Mortality Vs Income



Cluster 0 → Underdeveloped Countries

Cluster 1 → Developed Countries

Cluster 2 → Developing Countries

Conclusion

Based on high child mortality rate, low gdpp and low income of countries population, following are the 5 under – developed countries that need more attention :

- Haiti
- Sierra Leone
- Chad
- Central African Republic
- Mali

country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
Haiti	208.0	15.3	6.91	64.7	1500	5.45	32.1	3.33	662
Sierra Leone	160.0	16.8	13.10	34.5	1220	17.20	55.0	5.20	399
Chad	150.0	36.8	4.53	43.5	1930	6.39	56.5	6.59	897
Central African Republic	149.0	11.8	3.98	26.5	888	2.01	47.5	5.21	446
Mali	137.0	22.8	4.98	35.1	1870	4.37	59.5	6.55	708

Hence the recommendation would be to focus more on the above countries as they are in dire need of aid.