Clustering and PCA

GEETAKRISHNASAI GUNAPATI

Agenda

Objective:

The objective of this assignment is to identify the countries which require immediate assistance because of their socio-economic factors so that HELP foundation can use their funds to the best effect

Methodology:

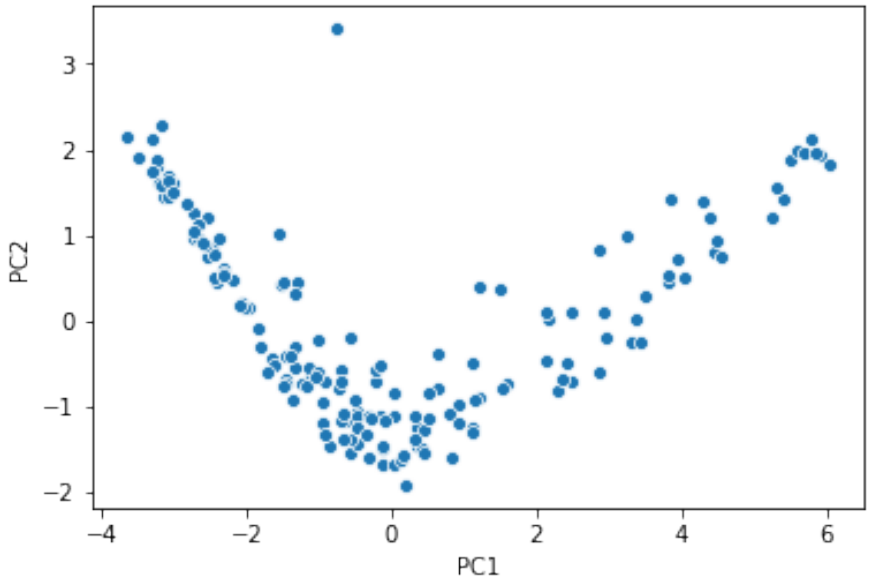
- Data understanding
- Perform PCA and identify the number of components
- Perform clustering using Kmeans and Hierarchical Clustering techniques
- Identify the cluster and the countries that require help

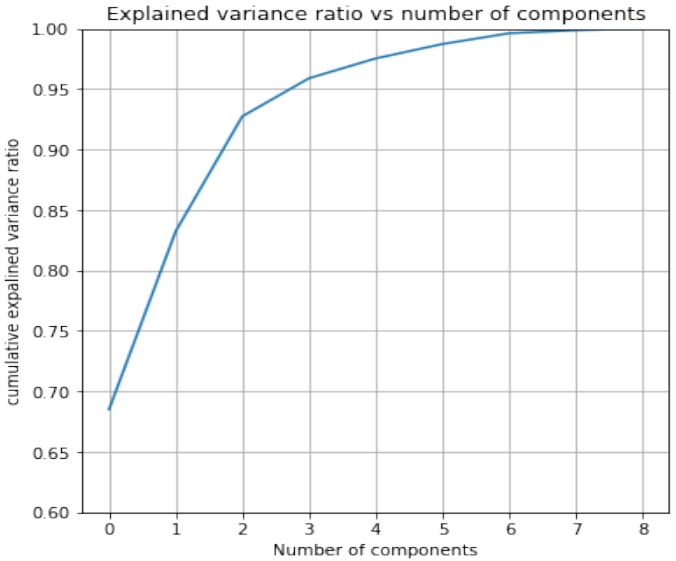
Data Understanding and Preparation

- Data has 9 socio-economic factors of 167 countries
- The columns health , exports and imports are given as percentage of GDP, they were converted into absolute values by multiplying with gdpp
- The outliers were treated by imputing the values, all the values which are lower than 0.05 percentile value and greater then 0.95 percentile were capped to those corresponding percentile values
- This will handle the issue of outliers and doesnt effect the PCA much because of size of data being very small
- All the numeric columns were standardized using standardscaler before performing PCA

Principal Component Analysis

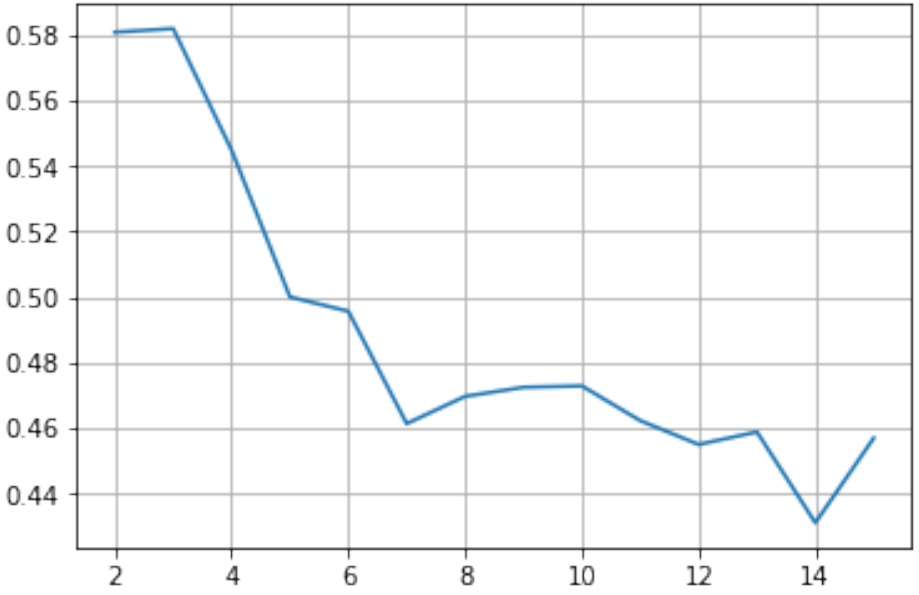
- PCA helps in removing multicollinearity and also in dimensionality reduction
- Scree curve plotted showed that 93% of the variance is explained using 2 PCs
- Hence dimensionality of the problem is reduced from 9 to 2
- The below curve shows the distribution of data in PCs plane and we seen pattern emerging

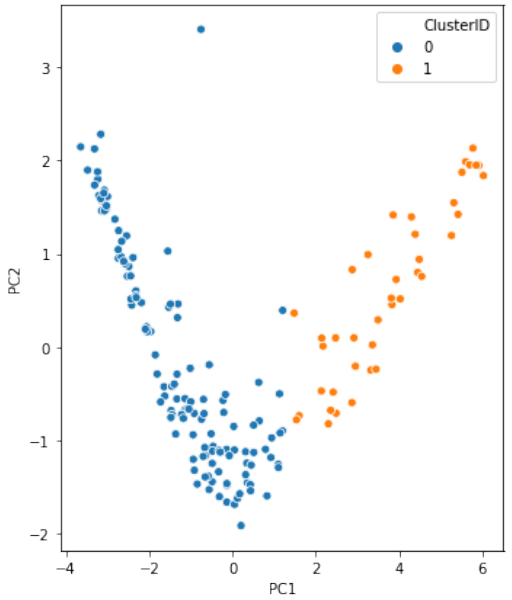


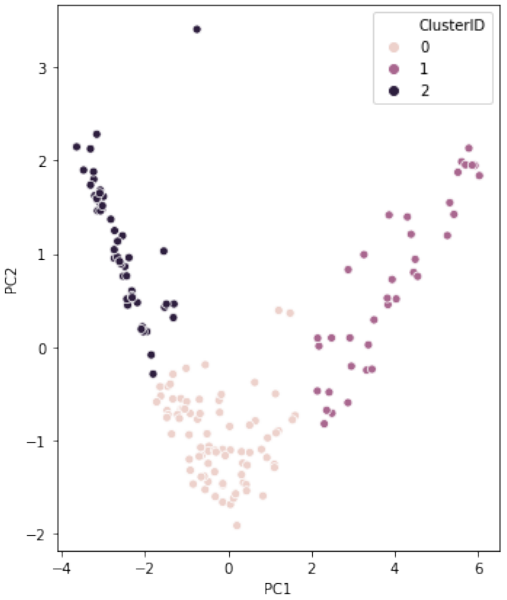


K-Means Clustering

- Silhouette score vs number of clusters is plotted to identify the number of clusters
- Elbow curve is also plotted to complement the observations from the silhouette score
- 2 and 3 clusters were considered for the K-Means and the clusters formed with 3 made more sense to the given problem

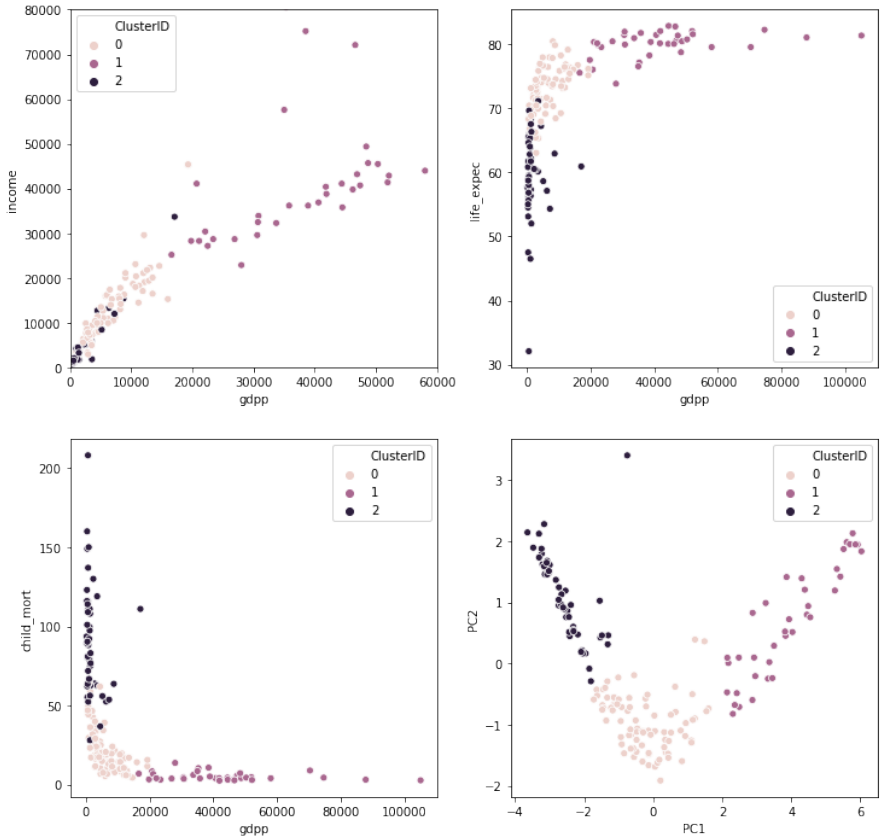






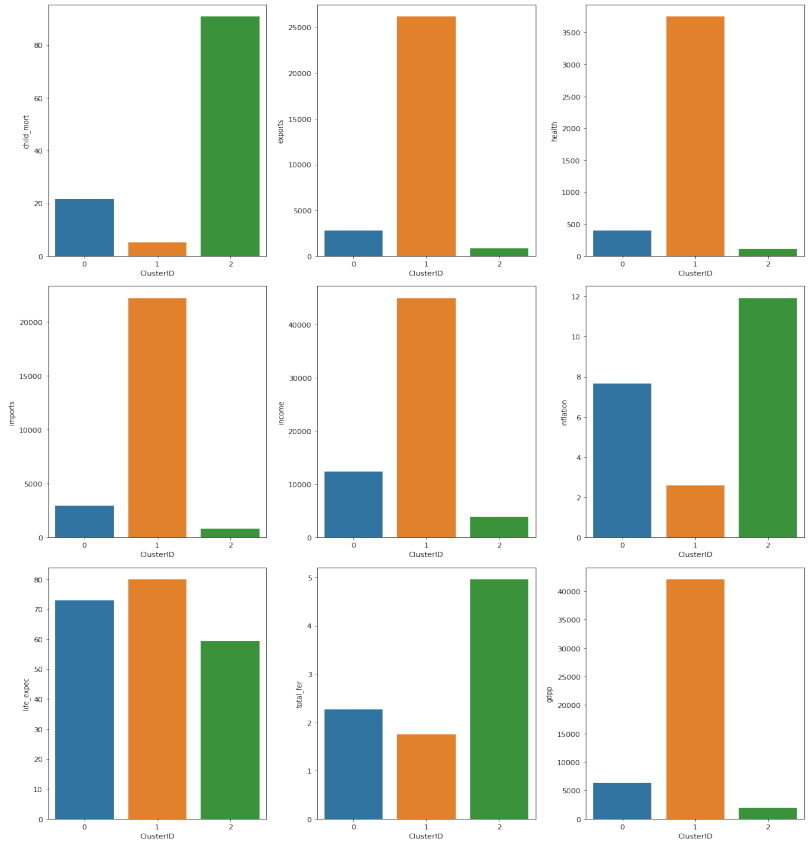
Results of K-Means

- The figure contains scatter plots of gdpp vs life_expec, gdpp vs income,gdpp ,gdpp vs child_mort and distribution in PC1 vs PC2 plane
- The distinction between the 3 clusters formed is visible from the figure
- The cluster2 is the one with low gdpp and high child_mort as well as low gdpp and low life_expec
- Cluster2 is the cluster of interest



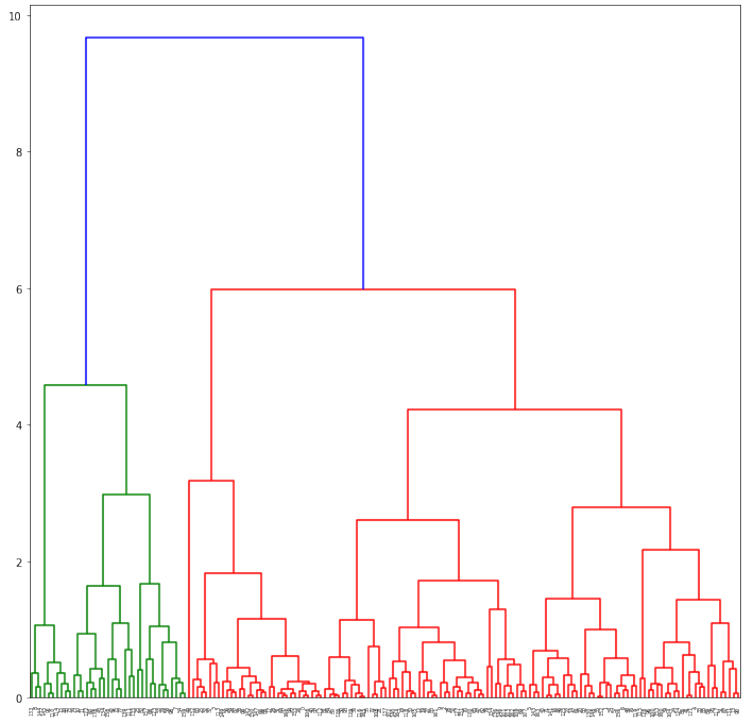
Results of K-Means

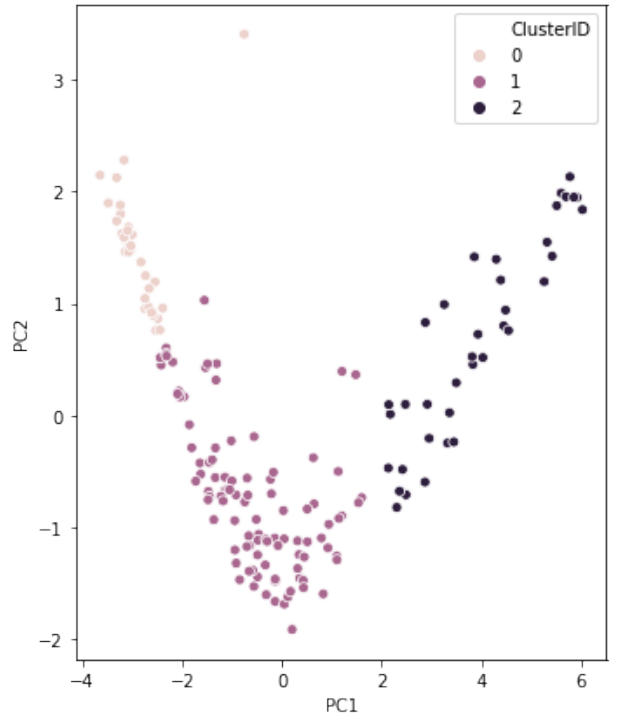
- The figure shows the variation of different features across the clusters
- It can been seen that cluster2 is the cluster of our interest as it has the lowest gdpp , high child mortality, low life expectancy , low income and low amount spend on health
- This supports the observation from the previous plots
- There are 49 countries will fall into the cluster2

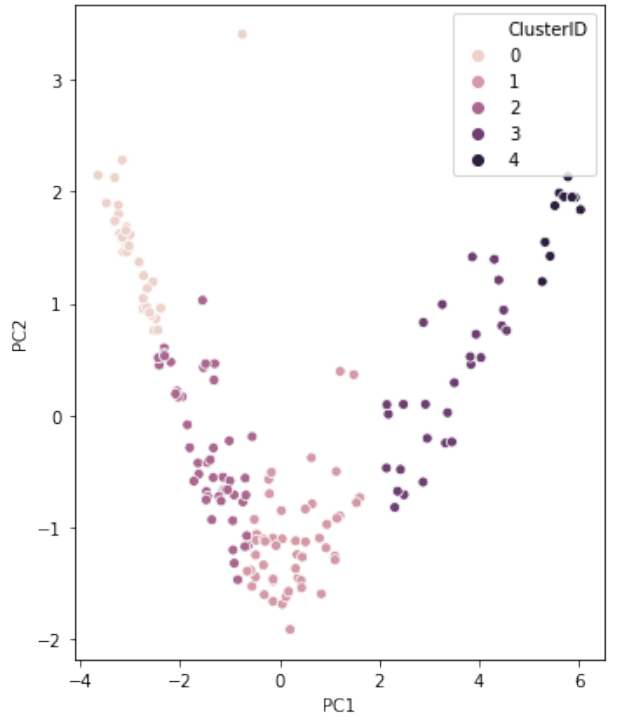


Hierarchical Clustering

- Hierarchical clustering is performed and the results for 3 and 5 clusters are compared
- The results showed that the cluster of interest remaining unchanged in both cases, therefore 3 clusters are considered for the analysis

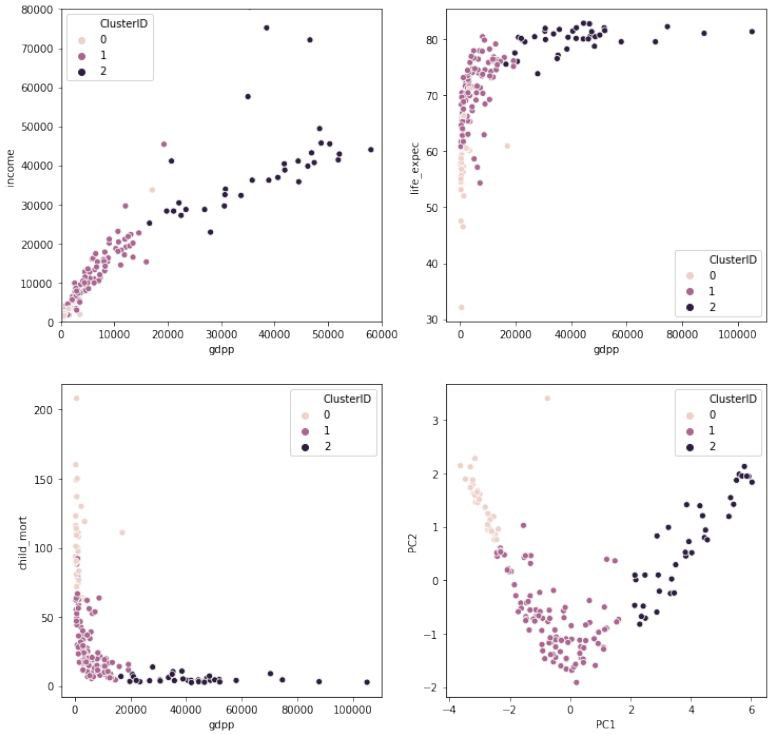


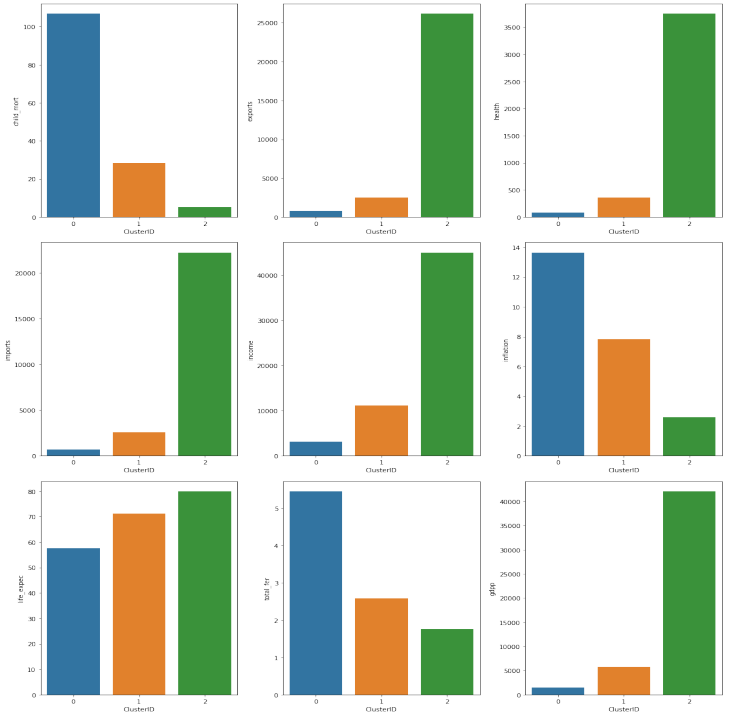




Results of Hierarchical Clustering

- The results observed in Hierarchical are similar to the ones observed in K-Means
- Cluster0 is the interested cluster and it has 32 countries





Conclusion:

- The clusters obtained from both K-Means and Hierarchical are overlapped with 32 countries, which is the entire cluster obtained from Hierarchical
- Since the results on K-Means depend on initial assumptions, we can consider the overlapped 32 countries as the ones that require assistance
- Among these 32 countries, I have considered 'gdpp' as the major factor as it is one of the major economic factor for the development and implementing welfare schemes and shortlisted 10 countries
- **Burundi, Liberia, Congo, Dem. Rep., Niger, Sierra Leone, Mozambique, Central African Republic, Malawi, Togo, Guinea-Bissau is the list of countries that HELP should focus on**