CLUSTERING AND PCA ASSIGNMENT

By:

Abhijeet Kumar

PROBLEM STATEMENT

1. Main task is to make clustering models and final cluster of the countries by the features/factors mentioned below and then by help of visualization present solution and recommendations.

2. categories the countries using socio-economic and health factors to decide development of the countries and to suggest the countries to focus most.

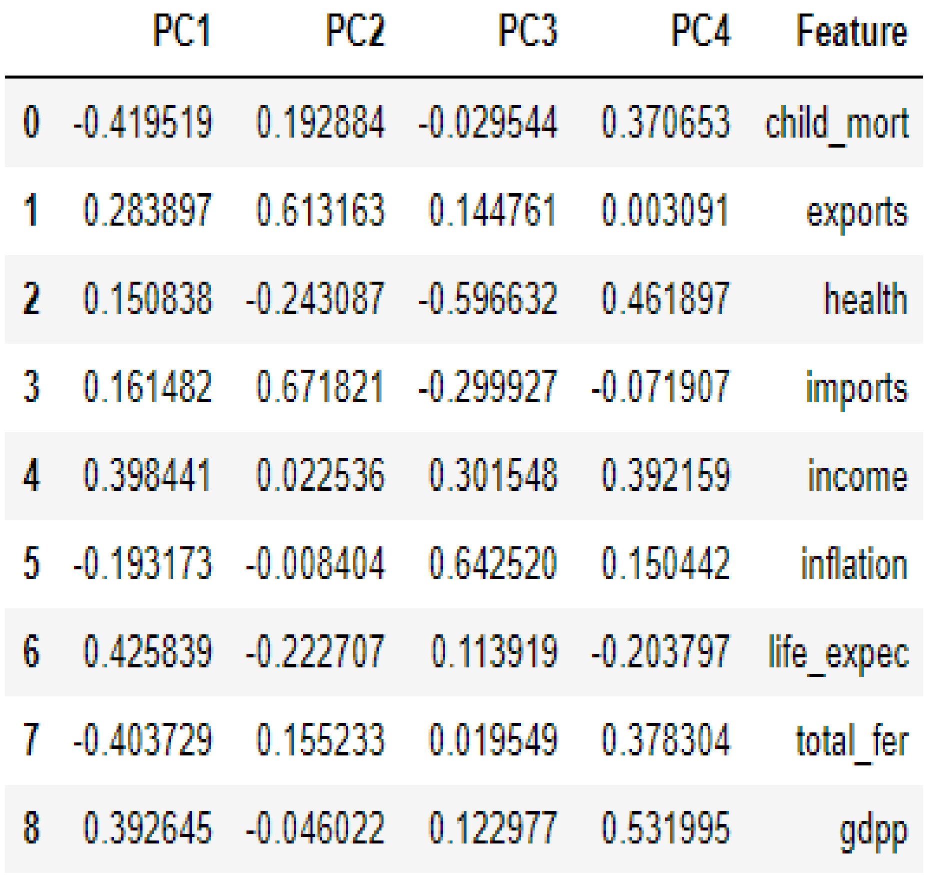
ANALYSIS APPROACH

1. Checking the dataset and doing basic EDA checks.
2. Checking correlation based on highly correlated.
3. Importing PCA-Module and performing PCA on dataset.
4. Merging PCA data with original dataset and outliers treatment.
5. Performing Hoppkins test to analysis dataset is best to use with respect to cluster tendency.
6. Applying Silhouette Analysis and sum of squares of distance between them.
7. Lastly applying K-means and hierarchal clustering on dataset and representing through Dendogram and Barplot

Principal Component Analysis

Objectives of principal component analysis

1. Discover the dimensionality of the data set.
2. Identify new meaningful underlying variables.
3. 4 principal components we have chosen with the help of screen plot.



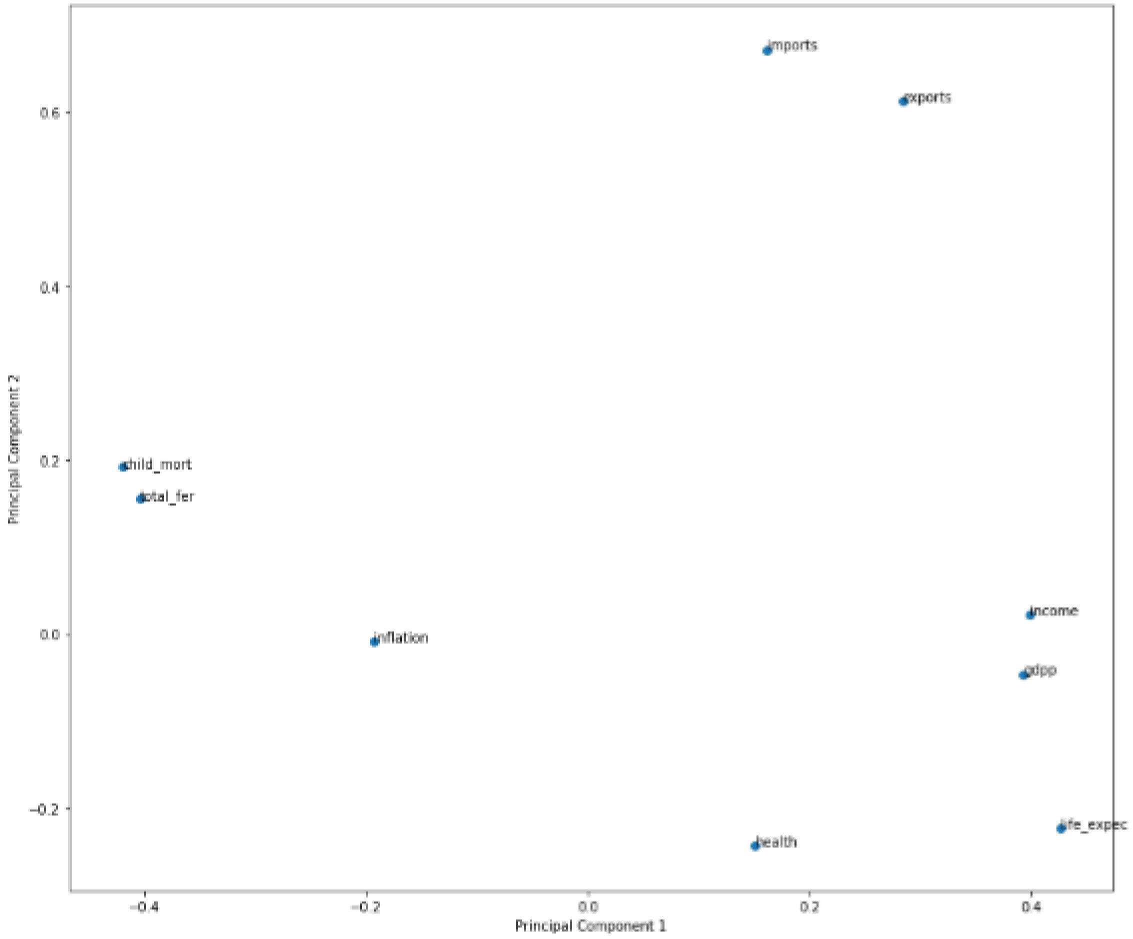
	PC1	PC2	PC3	PC4	Feature
0	-0.419519	0.192884	-0.029544	0.370653	child_mort
1	0.283897	0.613163	0.144761	0.003091	exports
2	0.150838	-0.243087	-0.596632	0.461897	health
3	0.161482	0.671821	-0.299927	-0.071907	imports
4	0.398441	0.022536	0.301548	0.392159	income
5	-0.193173	-0.008404	0.642520	0.150442	inflation
6	0.425839	-0.222707	0.113919	-0.203797	life_expec
7	-0.403729	0.155233	0.019549	0.378304	total_fer
8	0.392645	-0.046022	0.122977	0.531995	gdpp

PCA1 vs PCA 2 (Projection on 2D)

This scatter plot is generated from the principal components (PC1,PC2).

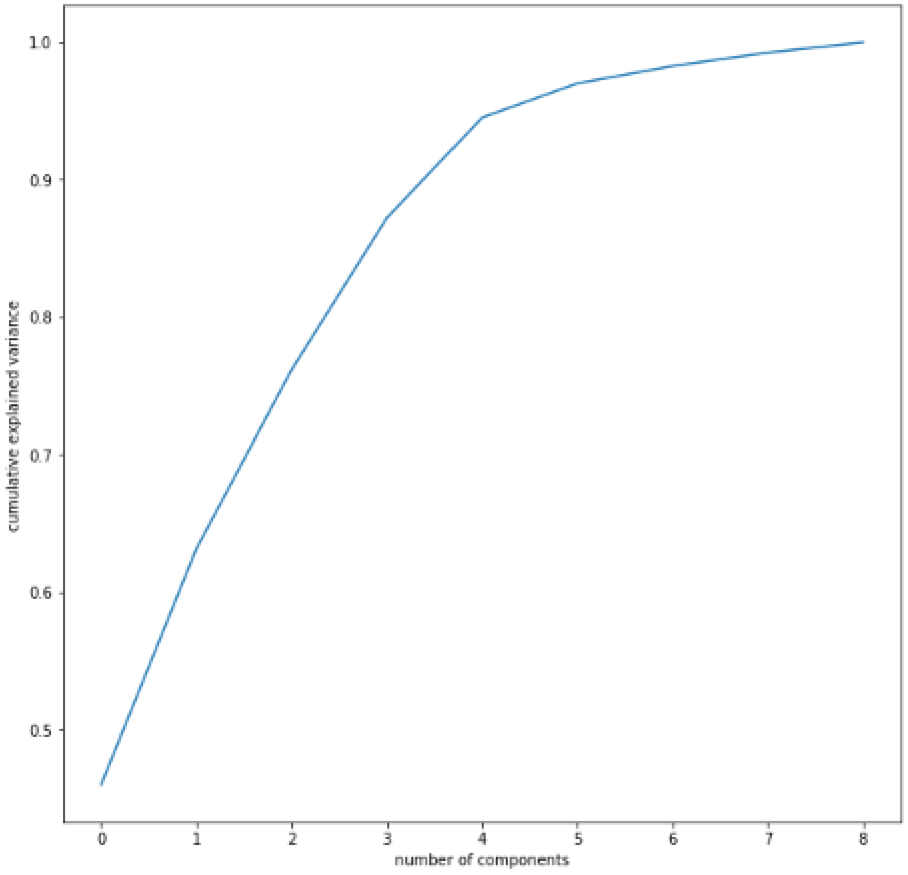
We see that the outliers are seen in the right side of graph.

It also shows how the feature (income,gdpp,life_expc, etc.) are highly depending on the country distributed.

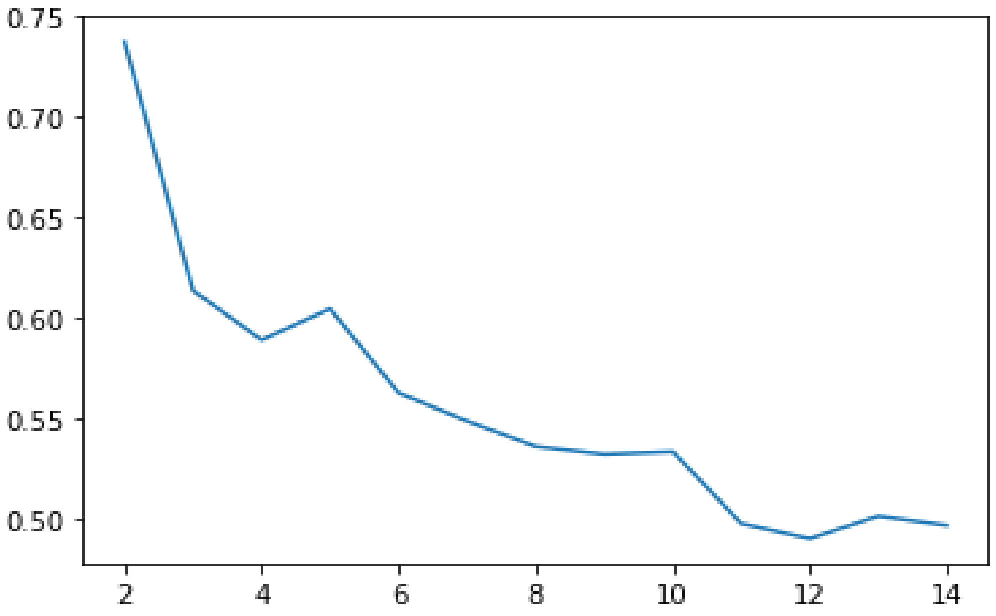


Scree Plot

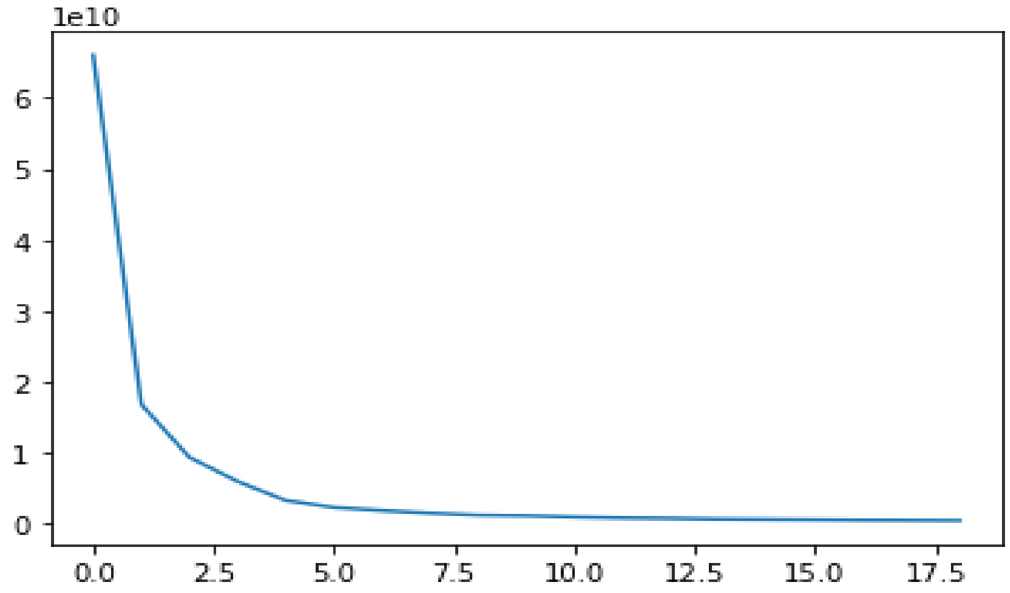
Graph shown here is the scree plot of the variance which is 95% of the information that contain by Four PCs and on the basis of this the four components we can proceed further.



Silhouette Analysis



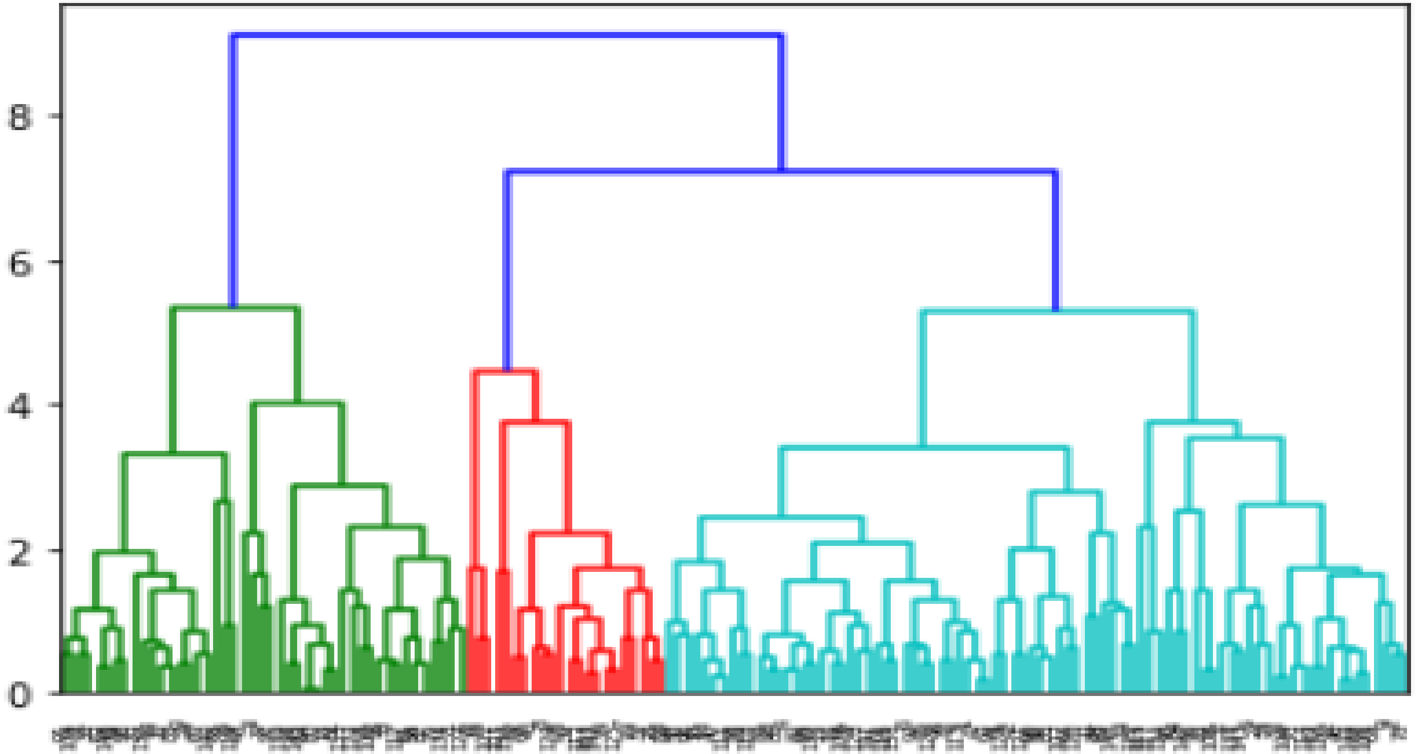
Sum of Squared Distance



Hierarchical Clustering

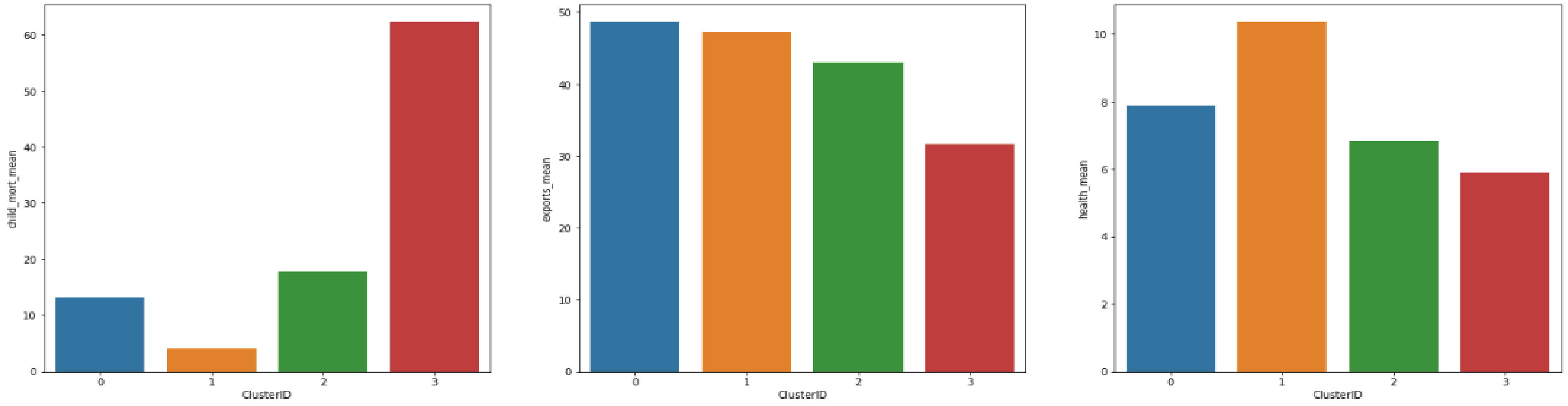
- Hierarchical graph clustering is another way to representing the clusters in dataset.

- Plot shown here is known as the Dendogram , where we can see how the country dataset for the components are divided and they are dependent on each other.

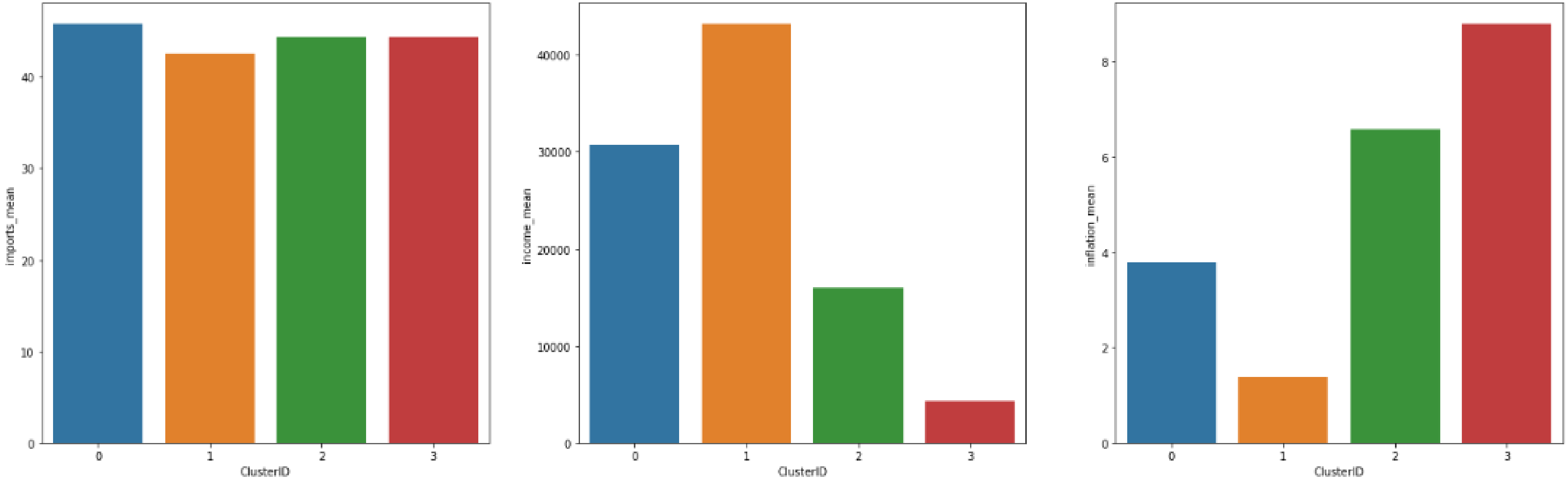


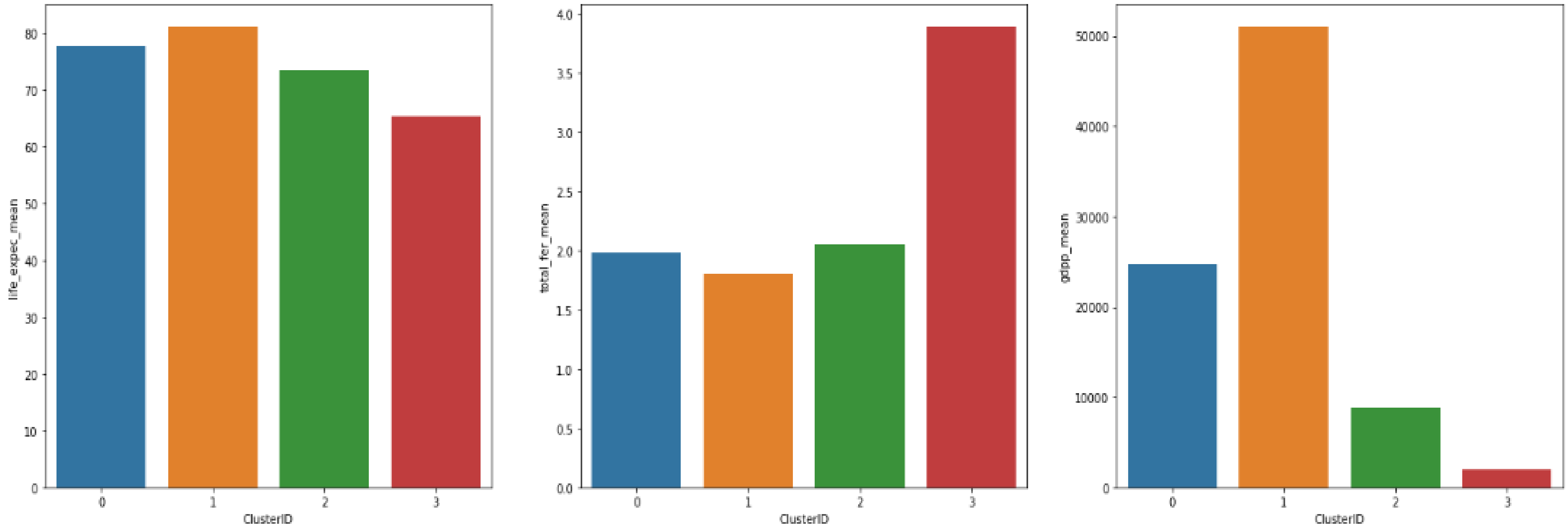
ANALYSIS :

IDENTIFYING THE CLUSTER WITH POOR COUNTRY



The Barplots here shows the Cluster 1 is poorest country which need support in crisis. As you can see the Gdpp is low and Child mortality is high of cluster 1 and life expectance is low and inflation is high of cluster one as their is significant gap between import and export.


imports_mean
0
10
20
30
40
0
1
2
3
ClusterID
income_mean
0
10000
20000
30000
40000
0
1
2
3
ClusterID
inflation_mean
0
2
4
6
8
0
1
2
3
ClusterID


life_expec_mean
0
10
20
30
40
50
60
70
80
0
1
2
3
ClusterID
total_fer_mean
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
0
1
2
3
ClusterID
gdpp_mean
0
10000
20000
30000
40000
50000
0
1
2
3
ClusterID

RECOMMENDATION

We needs to focus on backward countries which need more basic amenities and relief during the time of disasters and natural calamities.

a) We need to work on public awareness and education.
b) Survival plans should provide with basic information on what hazardous events are most likely to occurred in particular countries.
c) Community-wide planning and education system should be encouraged.
d) Disaster education is essential in the training of the government and private sector professionals, emergency management, and emergency service providers who have the major responsibility in disasters and natural calamities.