Clustering and PCA Assignment

SUBMITTED BY:

JITHIN CHOWDARY KUNKA

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.

After the recent project that included a lot of awareness drives and funding programs, 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 NGO wants to know:

The categories of countries using some socio-economic and health factors that determine overall development of the country. The countries which the CEO needs to focus on the most.

Based on various market surveys, the NGO has gathered a large dataset containing the socio-economic factors of the countries.

Objectives

Our main task is to cluster the countries by socio economic factors mentioned above and present a solution.

We are supposed to use dimensionality reduction using PCA to get the visualizations of the clusters in a 2-D form.

Data Pre Processing

The data was gone through a thorough pre-processing.

The provided data doesn’t contain any missing values or duplicate countries.

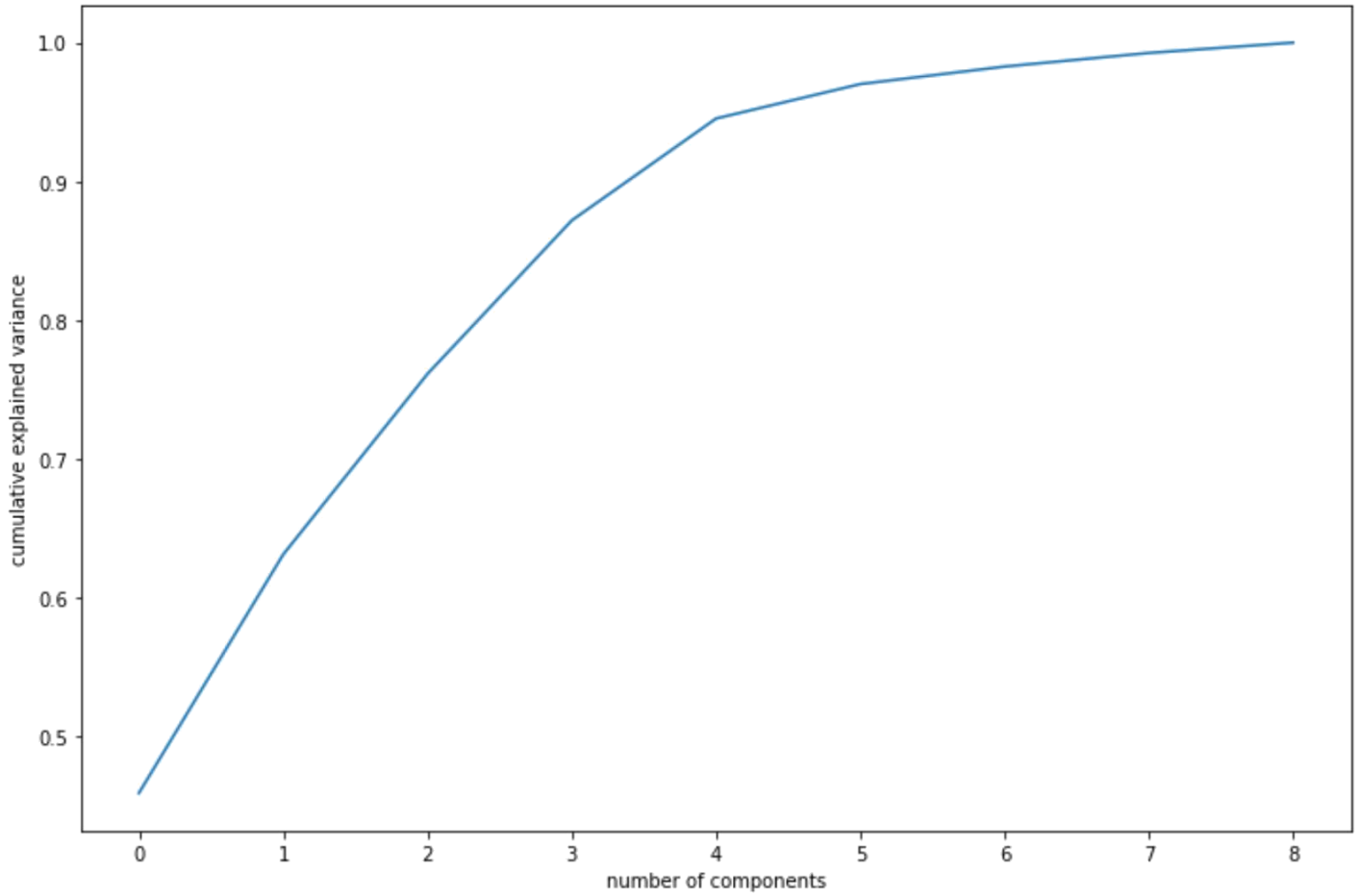
The provided data have certain outliers and they are treated during PCA.

Data is further standardised for Principal Component Analysis.

PCA

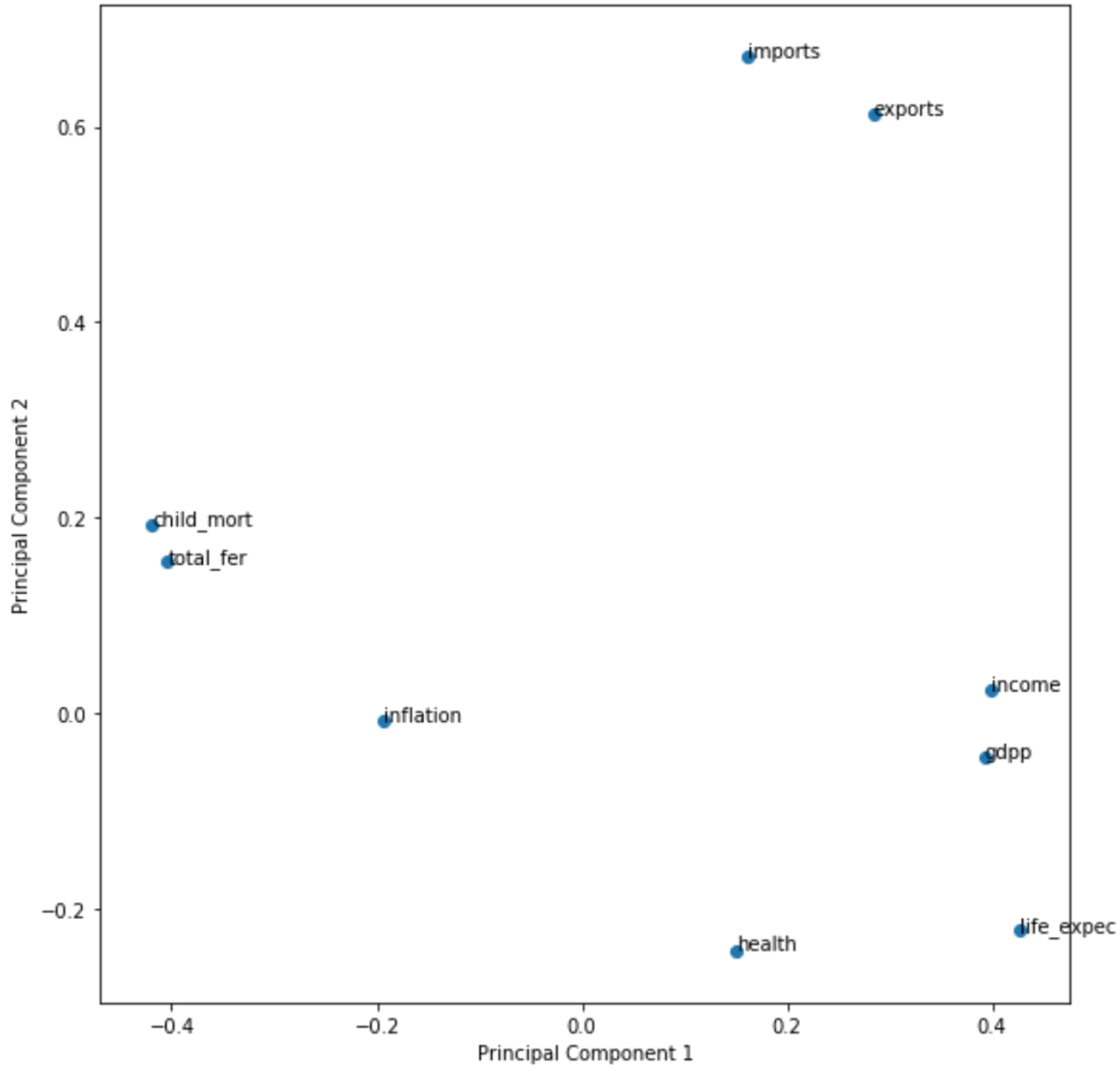
Beside Graph 4 components are enough to describe 95% of the variance in the dataset

We'll choose 4 components for our clustering



PCA

Variable Relation with 2 Principal Components



Clustering

We use both K means and Hierarchical Clustering on our 4 pca components.

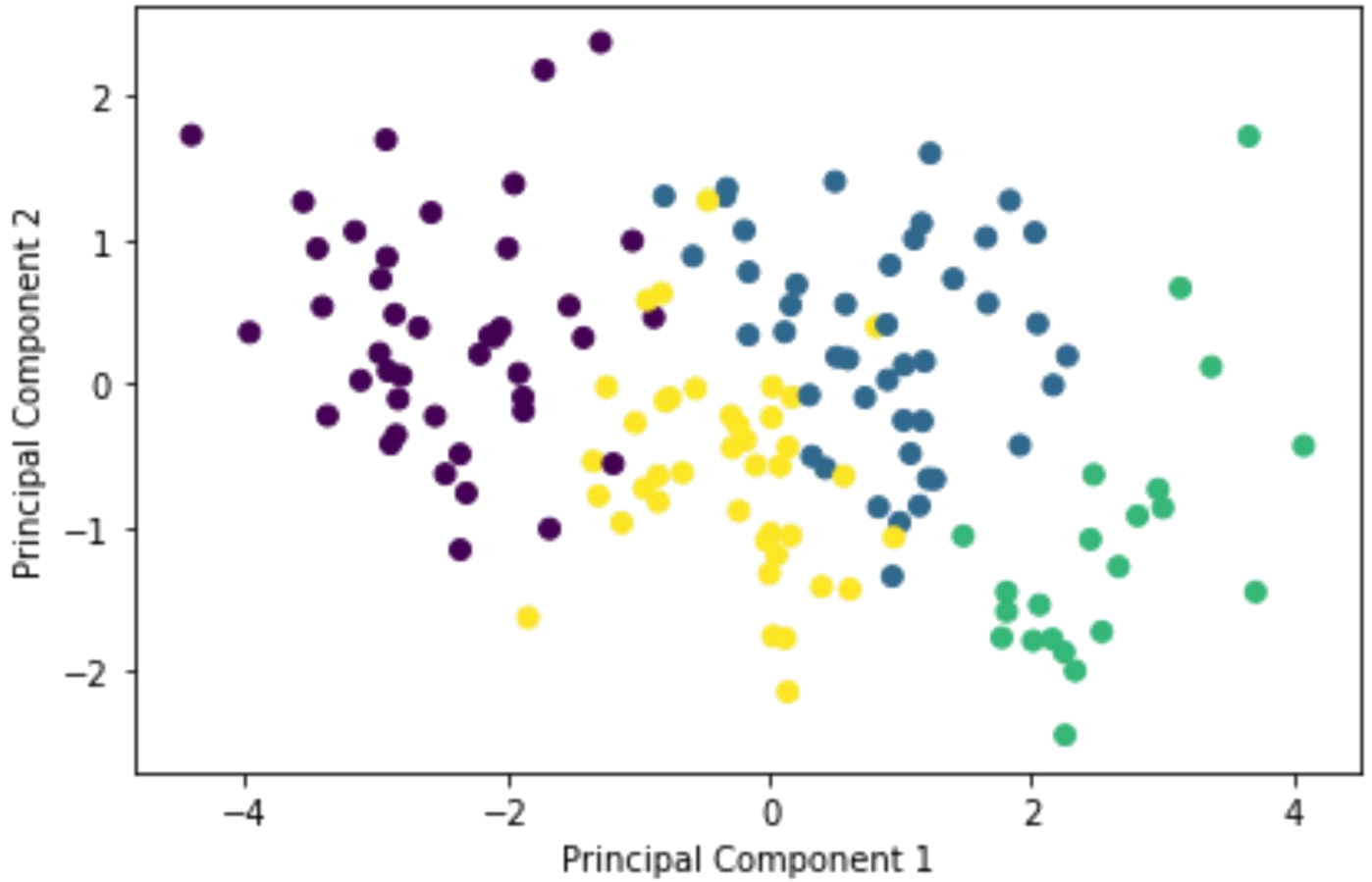
For K means Clustering we take K = 4, 5 Nearest Neighbours

We ran Hopkins Statistic which provide the value, if it is between {0.7, ..., 0.99}, it has a high tendency to cluster.

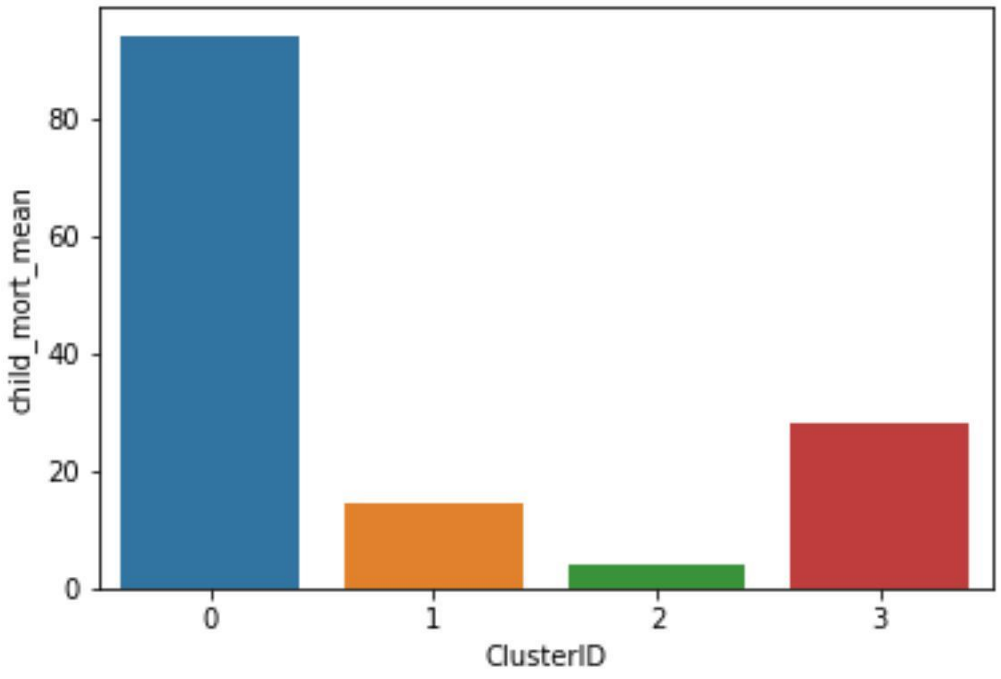
We got a Hopkins Statistic value = 0.75

Clustering

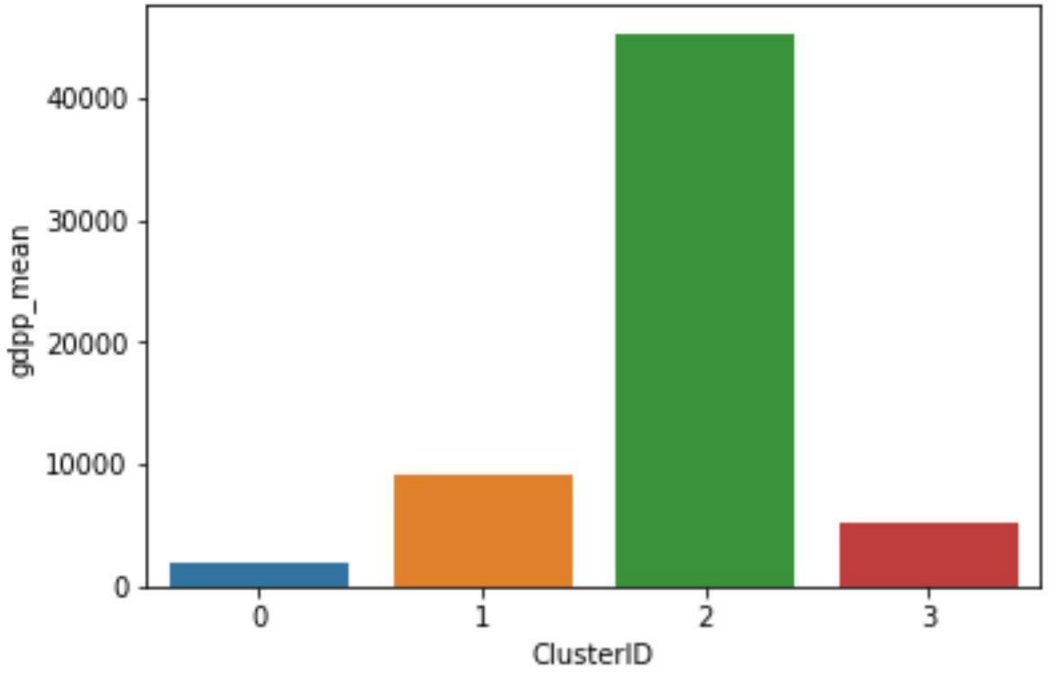
Clusters formed with 2 Principal Components



Clustering

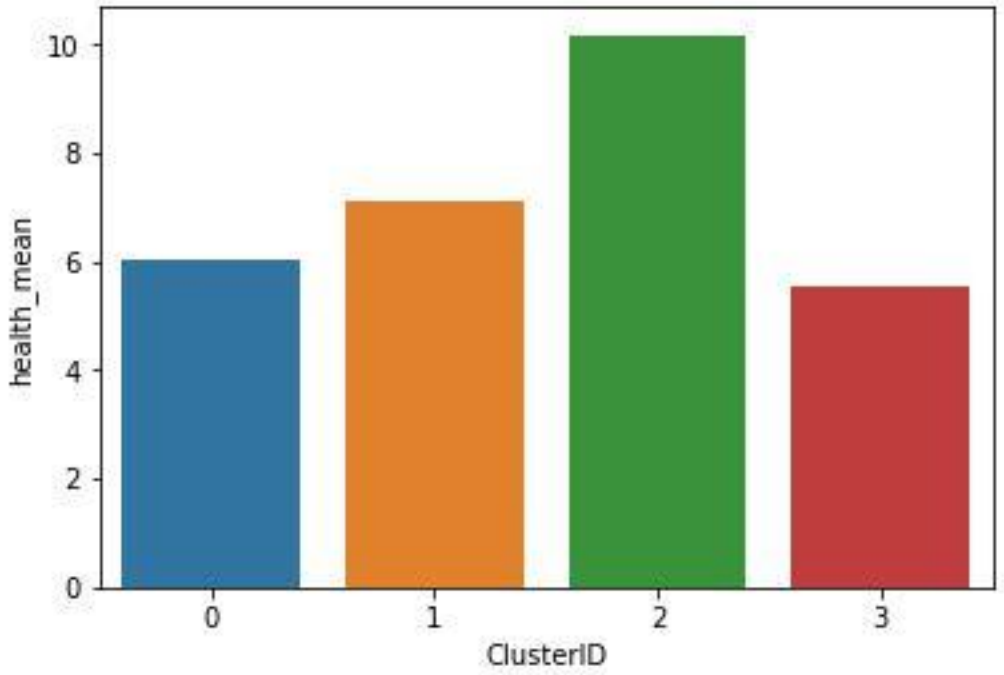


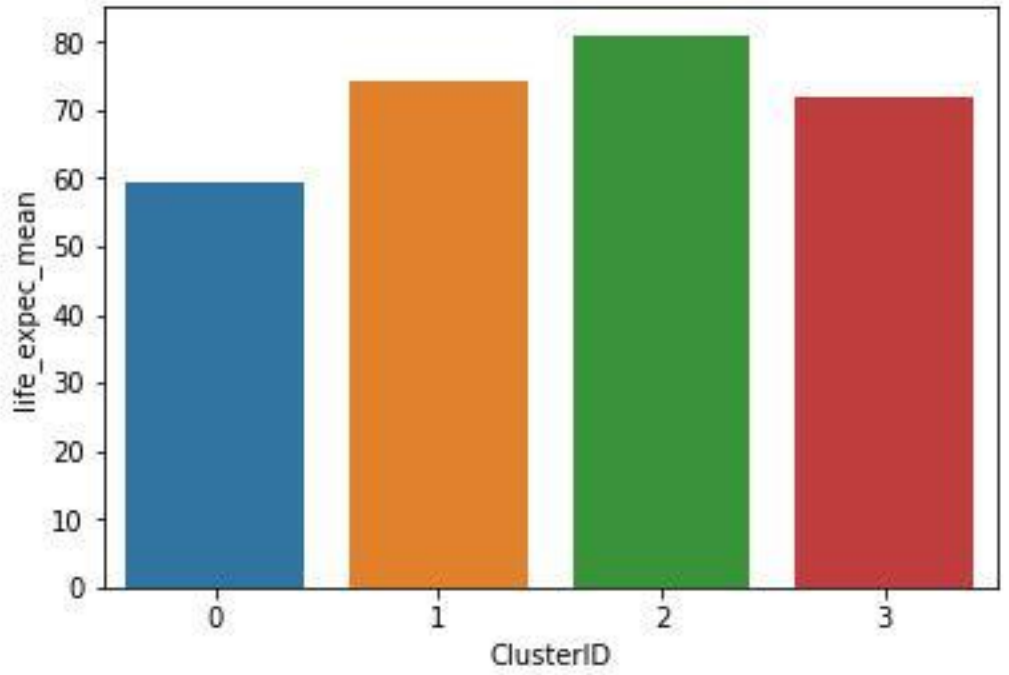
Child Mortality



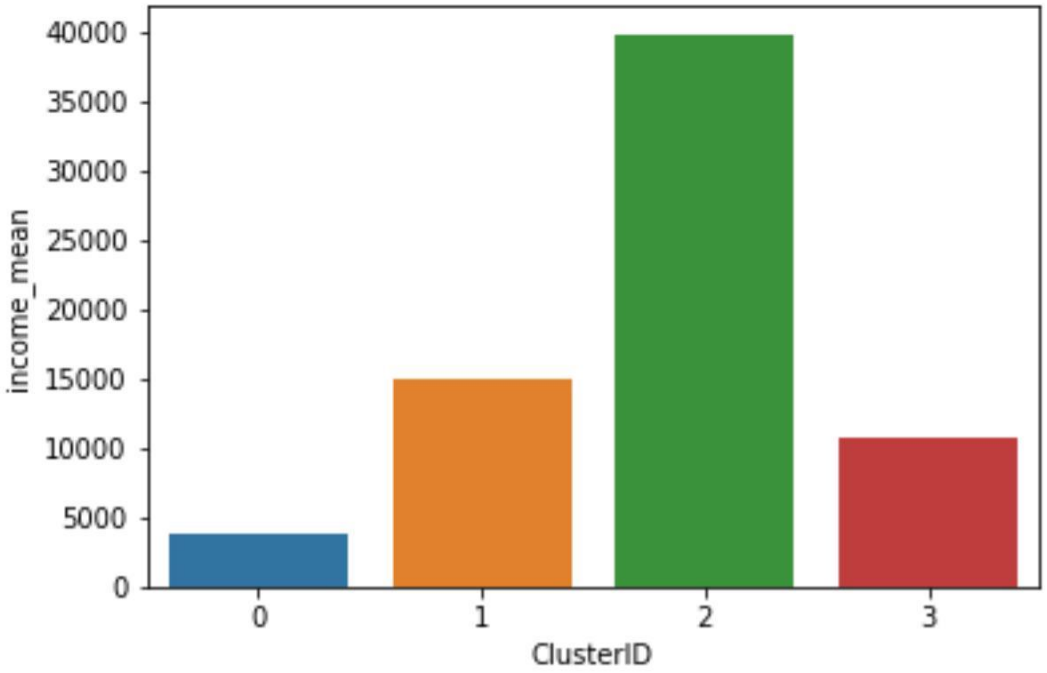
GDP

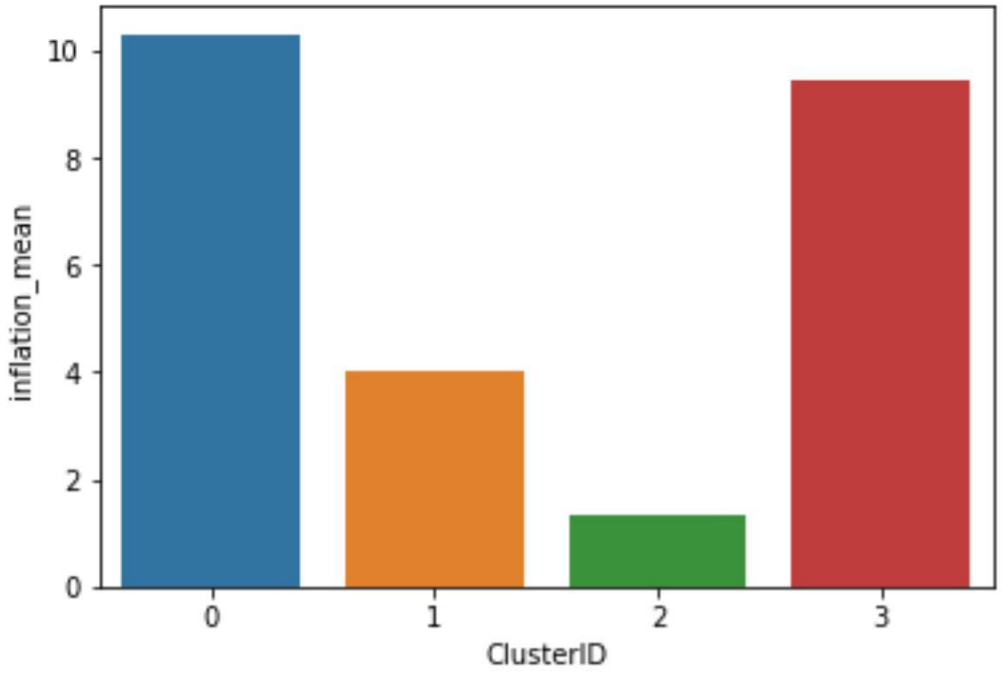
Clustering


health_mean
0
2
4
6
8
10
0
1
2
3
ClusterID

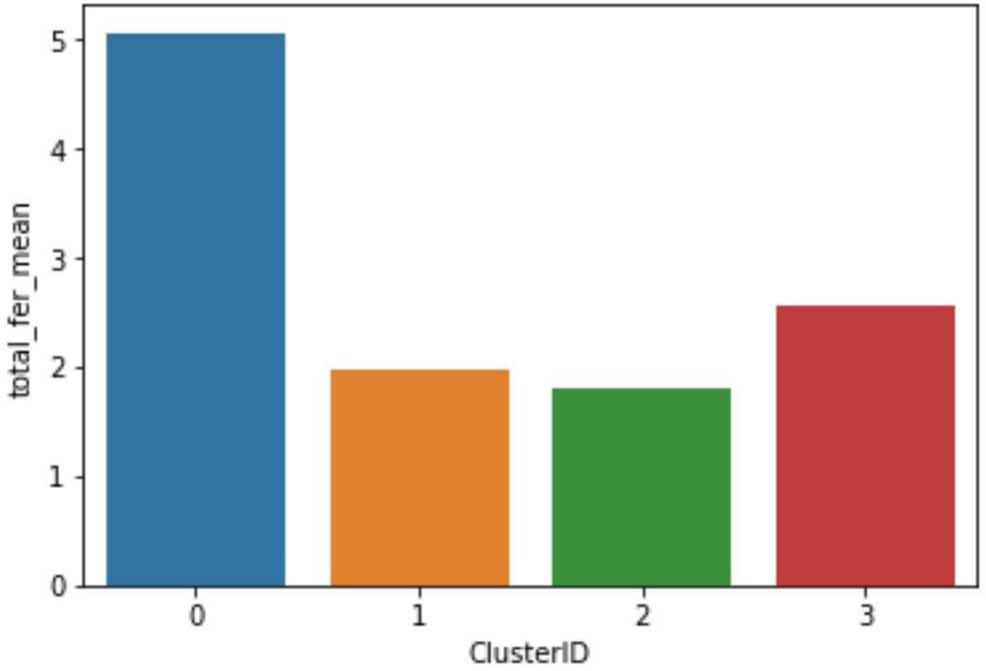

life_expec_mean
0
10
20
30
40
50
60
70
80
0
1
2
3
ClusterID

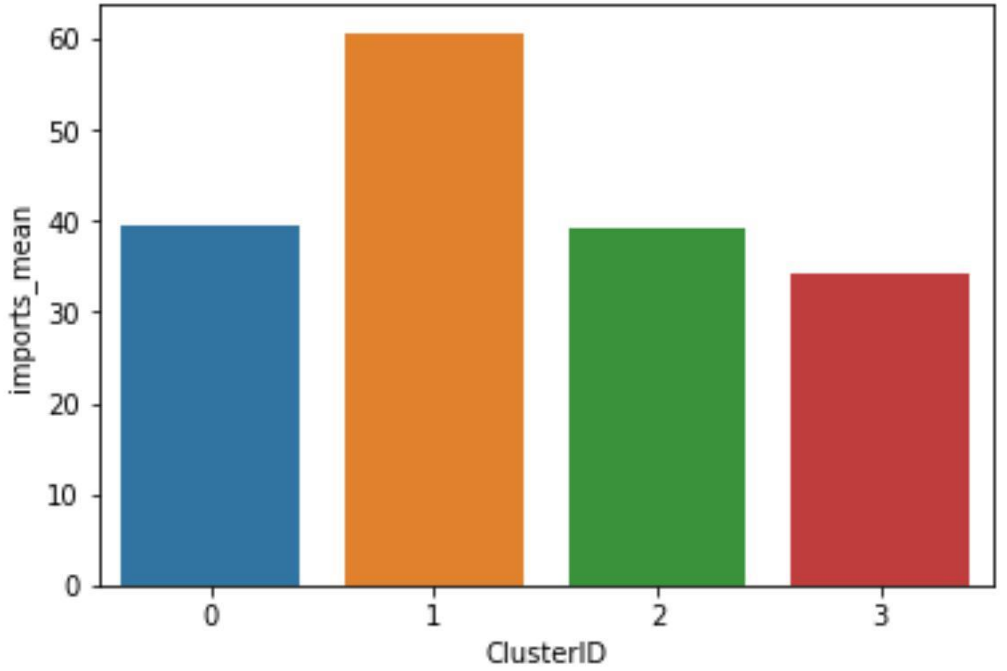
Clustering


income_mean
40000
35000
30000
25000
20000
15000
10000
5000
0
0
1
2
3
ClusterID


inflation_mean
10
8
6
4
2
0
0
1
2
3
ClusterID

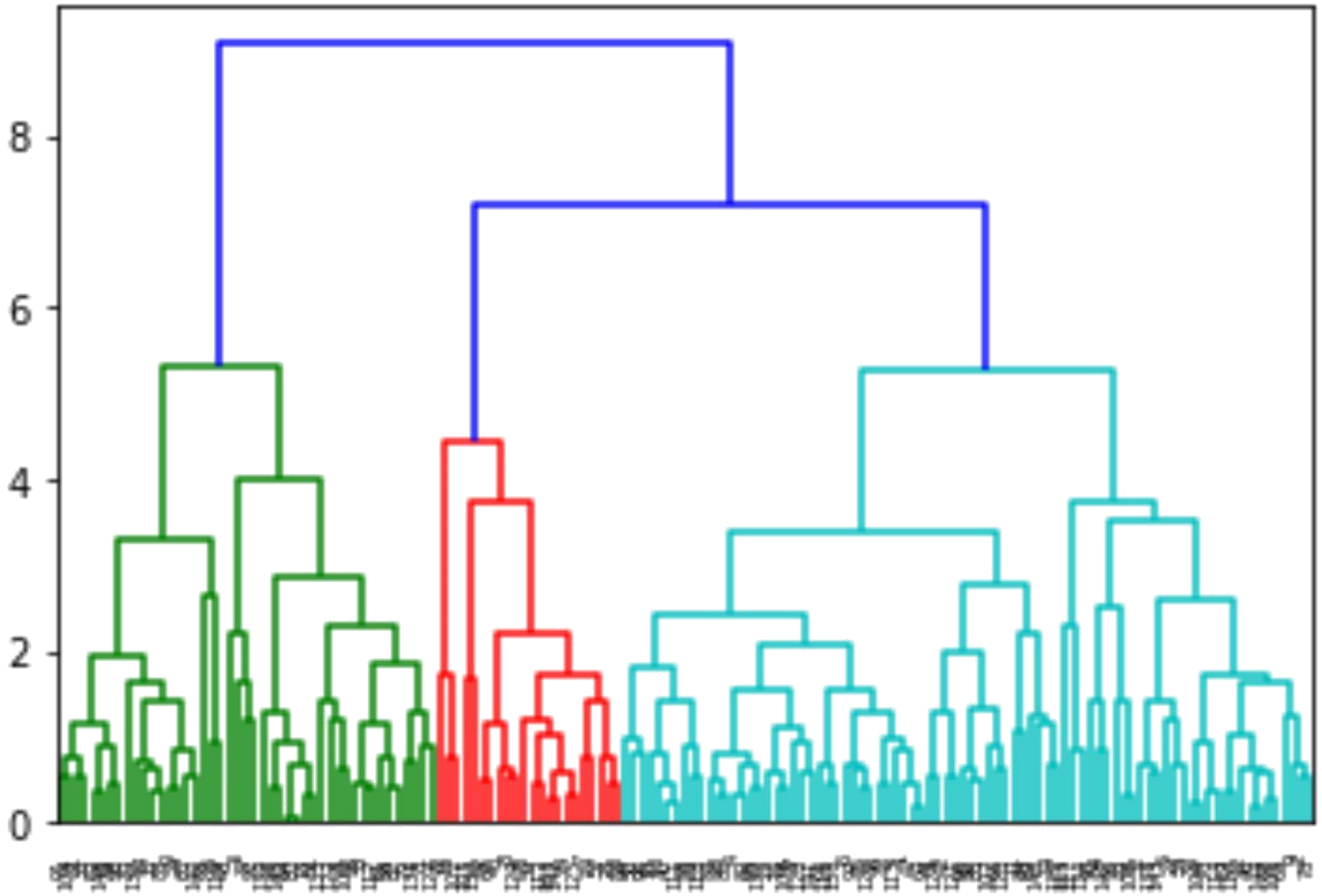
Clustering


total_fer_mean
0
1
2
3
4
5
ClusterID
0
1
2
3


imports_mean
0
10
20
30
40
50
60
ClusterID
0
1
2
3

Hierarchical Clustering

Beside Hierarchical Clustering shows no: of clusters it can be divided to



Conclusions

- We followed the below mentioned steps:
  - With all the socio economic data provided we carried out data cleaning, data preparation to confirm the data is ready for further analysis
  - In PCA we found out 4 components are enough with 95% Variance
  - Hopkins Statistic with 0.75 shows that formation of clusters is highly likely
  - Finally mean of different factors were plotted for all the clusters
  - Clusters with low income, low GDPP, high child mortality will be considered for maximum fund investment.

Countries belonging to such categories can be seen in the end result of the code.

THANK YOU