Clustering and PCA Assignment

Name: Srinivasa Praneeth

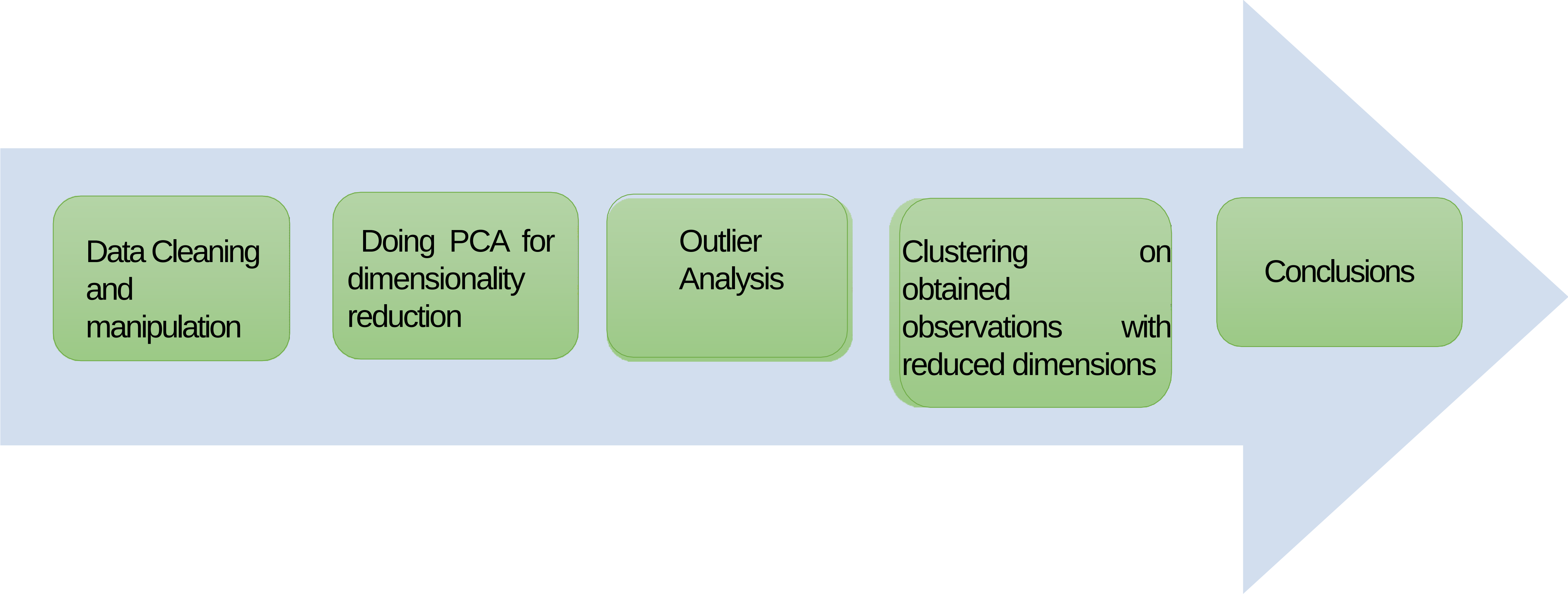
Problem Statement of HELP NGO

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.

They have been able to raise around $ 10 million. 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.

So ,now our job is to categorise the countries using some socio-economic and health factors that determine the overall development of the country so that we can distribute the financial aid correctly to respective countries which are in need of it the most

Analytical Approach to find solution for problem



Data Cleaning and manipulation

Doing PCA for dimensionality reduction

Outlier Analysis

Clustering on obtained observations with reduced dimensions

Conclusions

Data Cleaning and feature scaling:

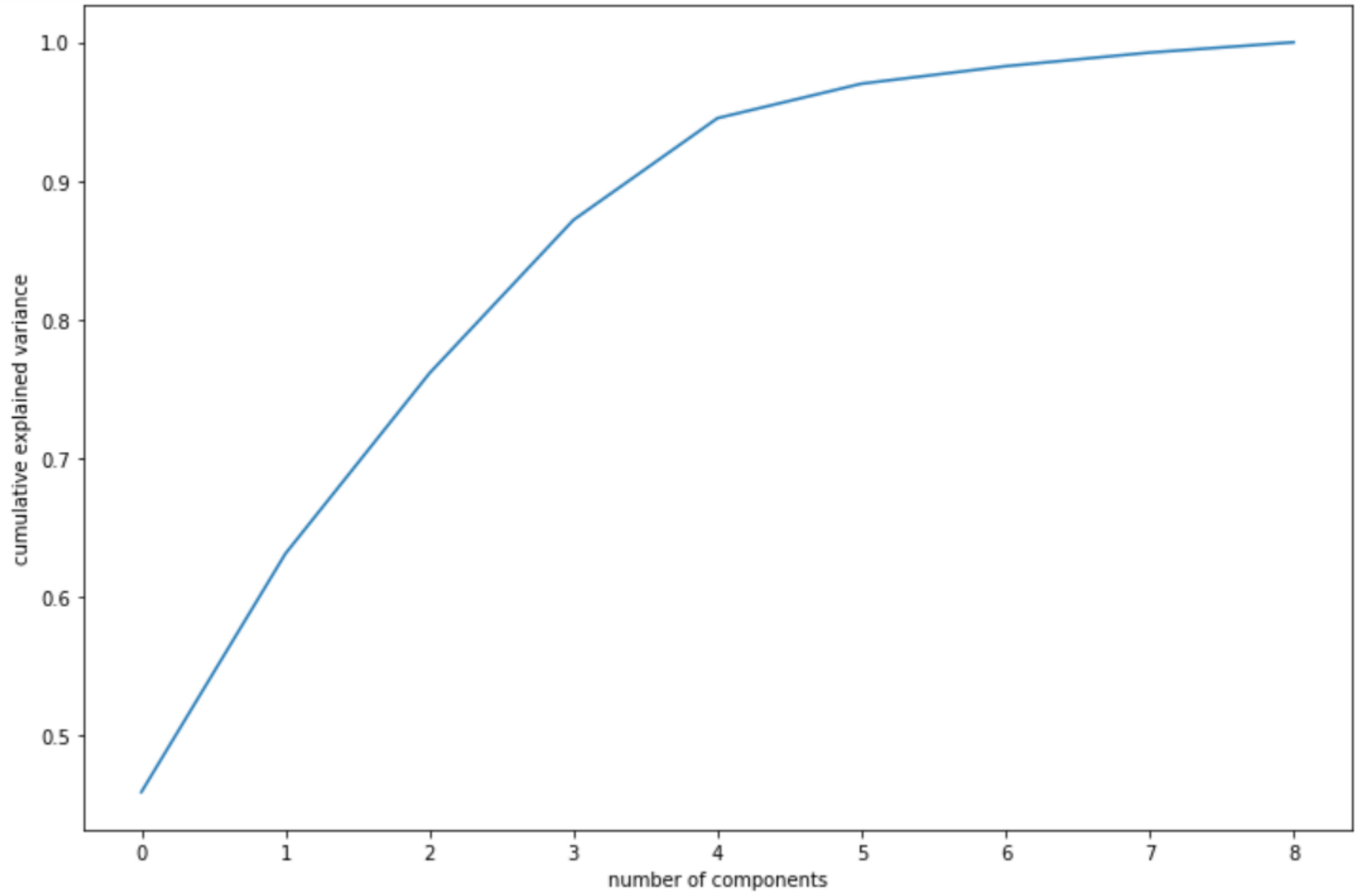
- As major part of analysis is data cleaning ,Analysis begins with importing the data from countries data file into python using the pandas library by read_csv command.

- Checked in the data frame whether it has any missing values in any rows and columns and found there are no any missing values

- Checked for any outliers using visualizations

- Using feature standardization, have scaled all the numerical columns except the country feature which is non-numeric column

- After that, started doing the PCA on the numerical data for dimensionality reduction so that we dont loose any information

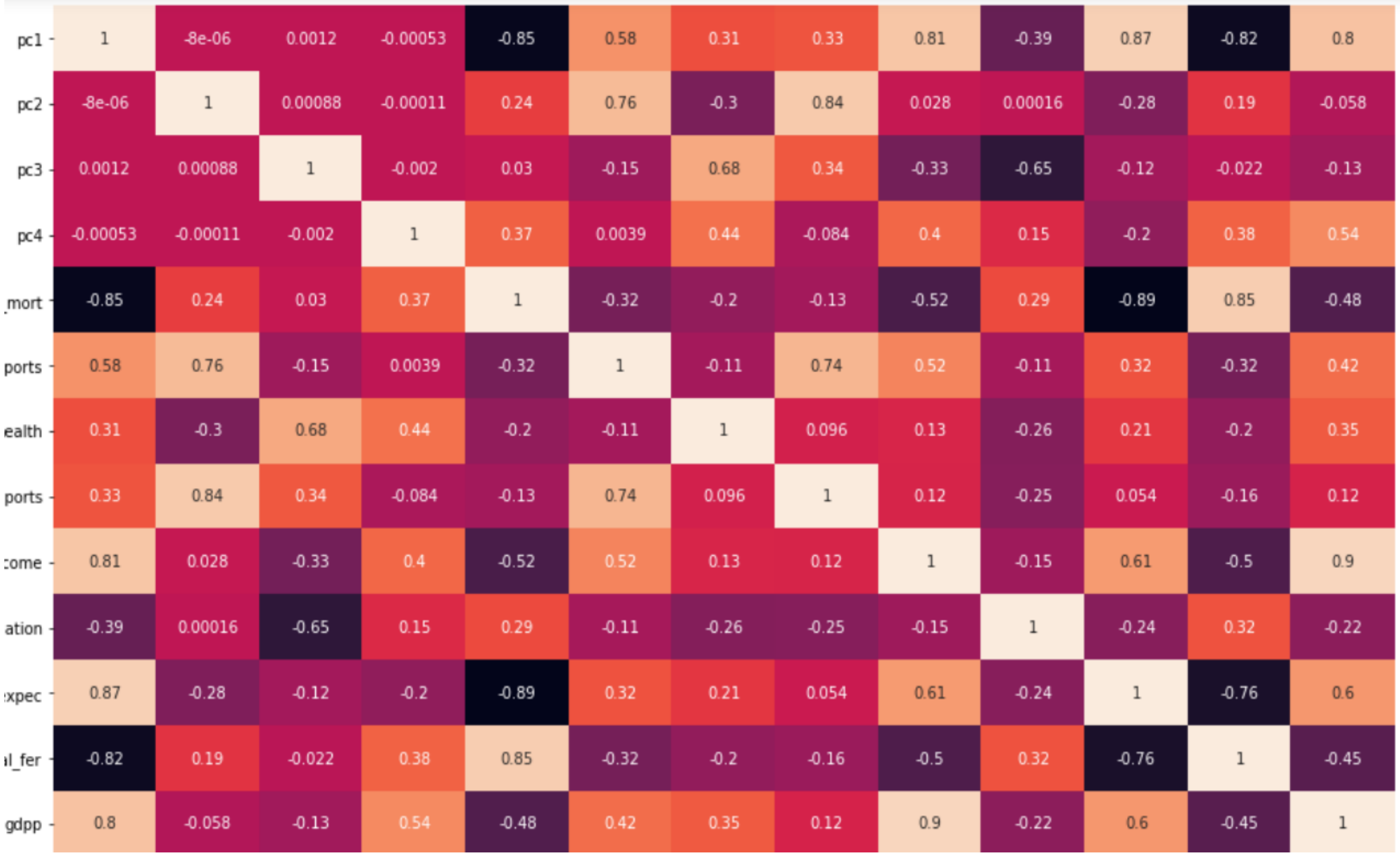


- Then From above scree plot we can assume to get 85-87% variance we can use 4 PC's, i am trying with 87% variance as it is
- mentioned in many sources including discussion forum that better to go for variance between 85-90%

- So, continued with 87% variance is obtained using incremental PCA where we have taken 4 Principal Components with minimum correlation where effectively removed multicollinearity

Now PCA is done and added back the original data to PCA dataset also, so now we checked for outliers and discarded any outliers

After checking the correlation matrix, observed the PCA results as below



pc1	1	-8e-06	0.0012	-0.00053	-0.85	0.58	0.31	0.33	0.81	-0.39	0.87	-0.82	0.8
pc2	-8e-06	1	0.00088	-0.00011	0.24	0.76	-0.3	0.84	0.028	0.00016	-0.28	0.19	-0.058
pc3	0.0012	0.00088	1	-0.002	0.03	-0.15	0.68	0.34	-0.33	-0.65	-0.12	-0.022	-0.13
pc4	-0.00053	-0.00011	-0.002	1	0.37	0.0039	0.44	-0.084	0.4	0.15	-0.2	0.38	0.54
mort	-0.85	0.24	0.03	0.37	1	-0.32	-0.2	-0.13	-0.52	0.29	-0.89	0.85	-0.48
ports	0.58	0.76	-0.15	0.0039	-0.32	1	-0.11	0.74	0.52	-0.11	0.32	-0.32	0.42
ealth	0.31	-0.3	0.68	0.44	-0.2	-0.11	1	0.096	0.13	-0.26	0.21	-0.2	0.35
ports	0.33	0.84	0.34	-0.084	-0.13	0.74	0.096	1	0.12	-0.25	0.054	-0.16	0.12
come	0.81	0.028	-0.33	0.4	-0.52	0.52	0.13	0.12	1	-0.15	0.61	-0.5	0.9
ation	-0.39	0.00016	-0.65	0.15	0.29	-0.11	-0.26	-0.25	-0.15	1	-0.24	0.32	-0.22
xpec	0.87	-0.28	-0.12	-0.2	-0.89	0.32	0.21	0.054	0.61	-0.24	1	-0.76	0.6
l_fer	-0.82	0.19	-0.022	0.38	0.85	-0.32	-0.2	-0.16	-0.5	0.32	-0.76	1	-0.45
gdpp	0.8	-0.058	-0.13	0.54	-0.48	0.42	0.35	0.12	0.9	-0.22	0.6	-0.45	1

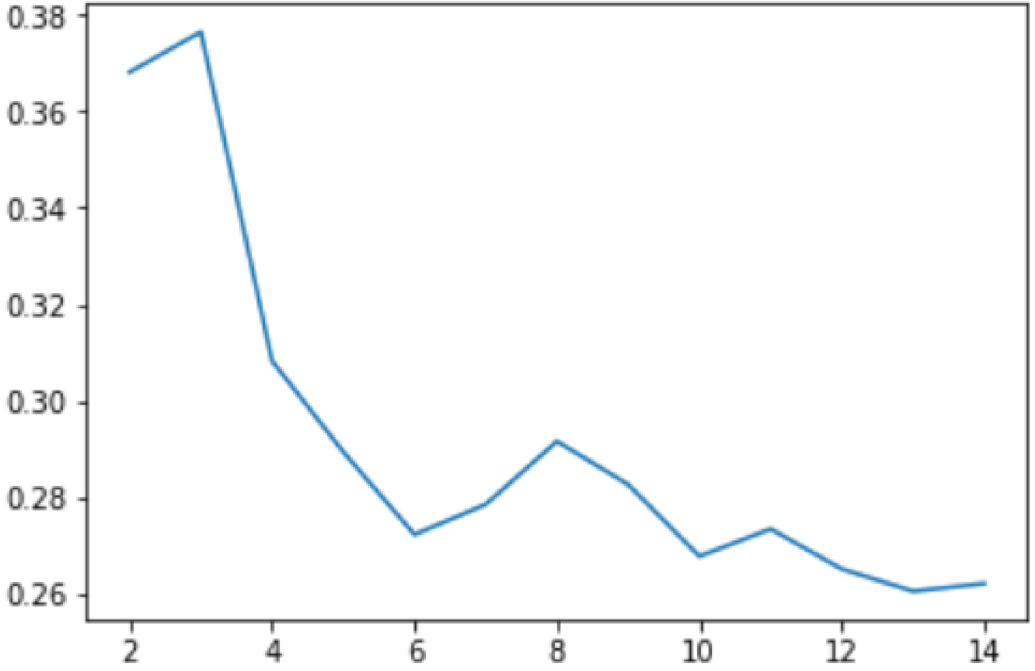
we can say that PC1 is positively correlated with life_expectancy, income and next gdpp also there. PC1 is negatively correlated with total_fer and child_mortality so PC1 is well explaining those features.

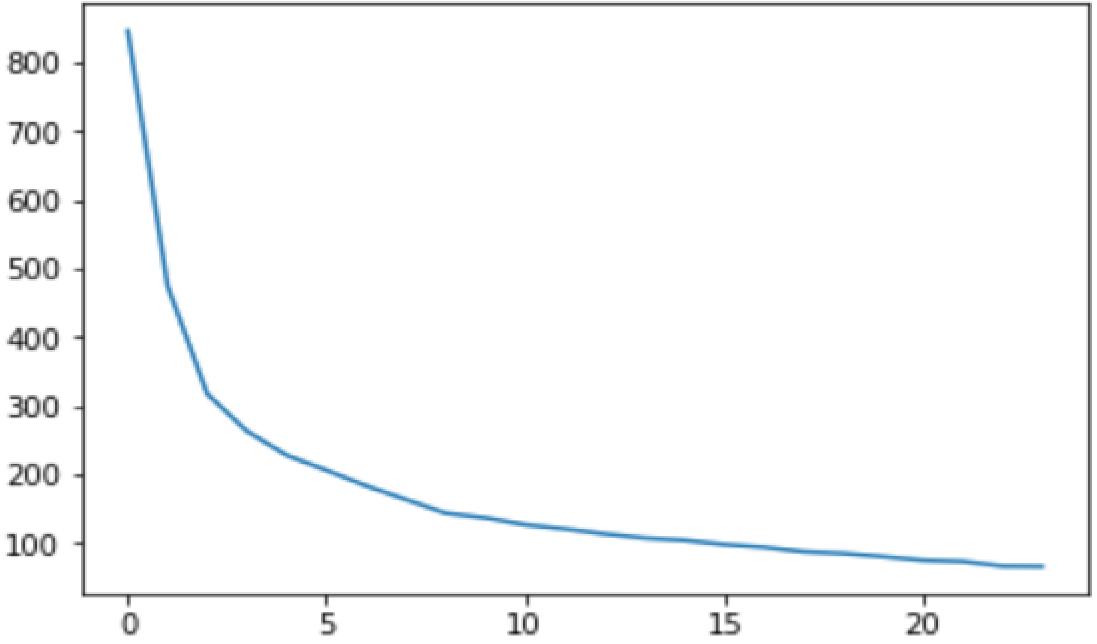
PC2 is highly positively correlated with exports and imports

PC3 is highly positively correlated with health and negatively correlated with inflation, so PC3 well explaining those

Moving to form clusters by k-means and hierarchial clusteringon the obtained PCA with original data

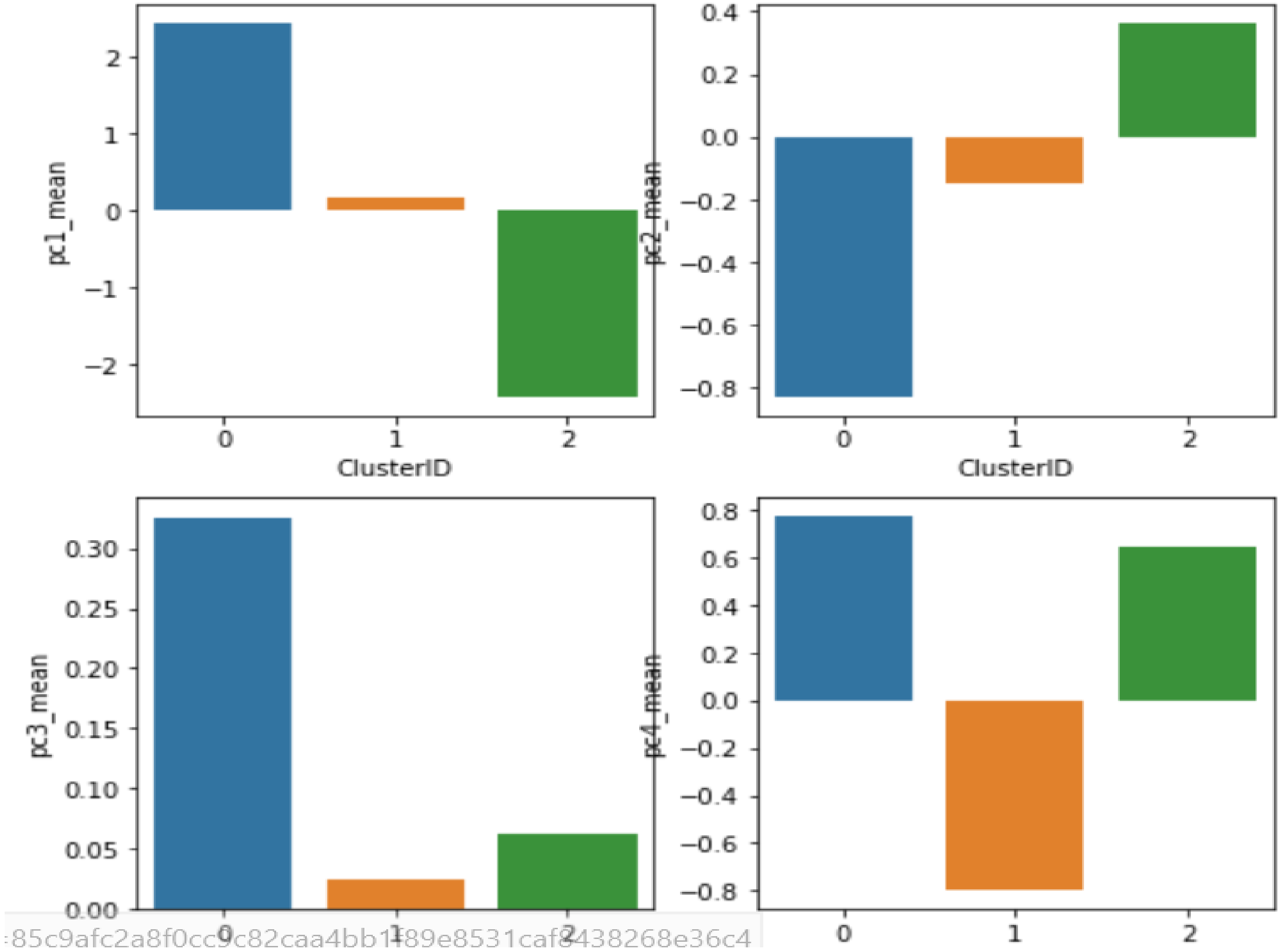
Now After this, we need to do clustering , but we need to check whether it can be done using hopkins statistics
As we got the value of 0.75, we have proceeded to do clustering.
After this, have done the Silhouette Analysis and drawn the elbow curve which suggests the optimal value of K.





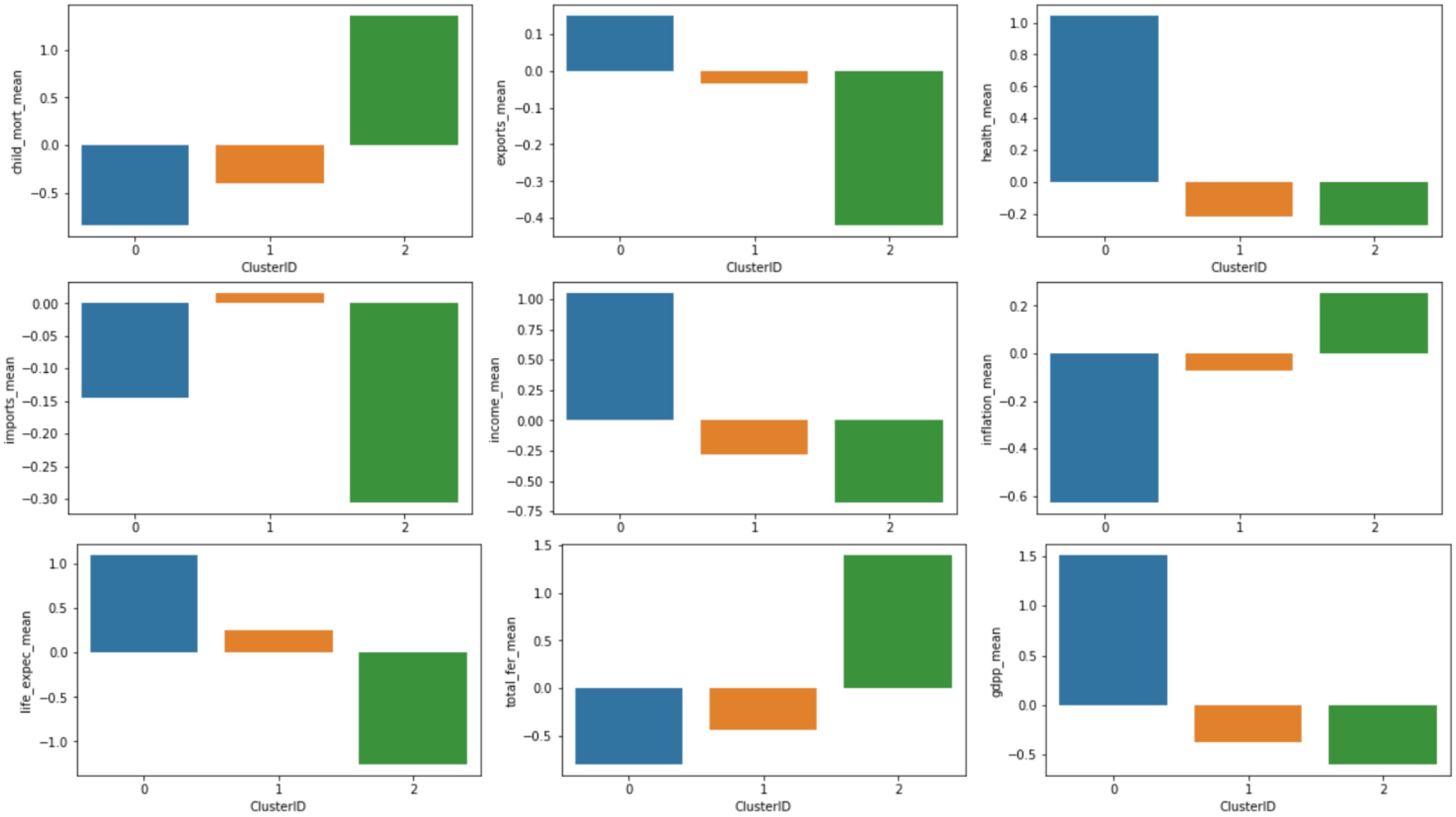
Both the methods are saying most optimal value is 3 , and checked clustering with k=3 and also checked with 4 and 5 as Per elbow curve.

K=3 with K-Means



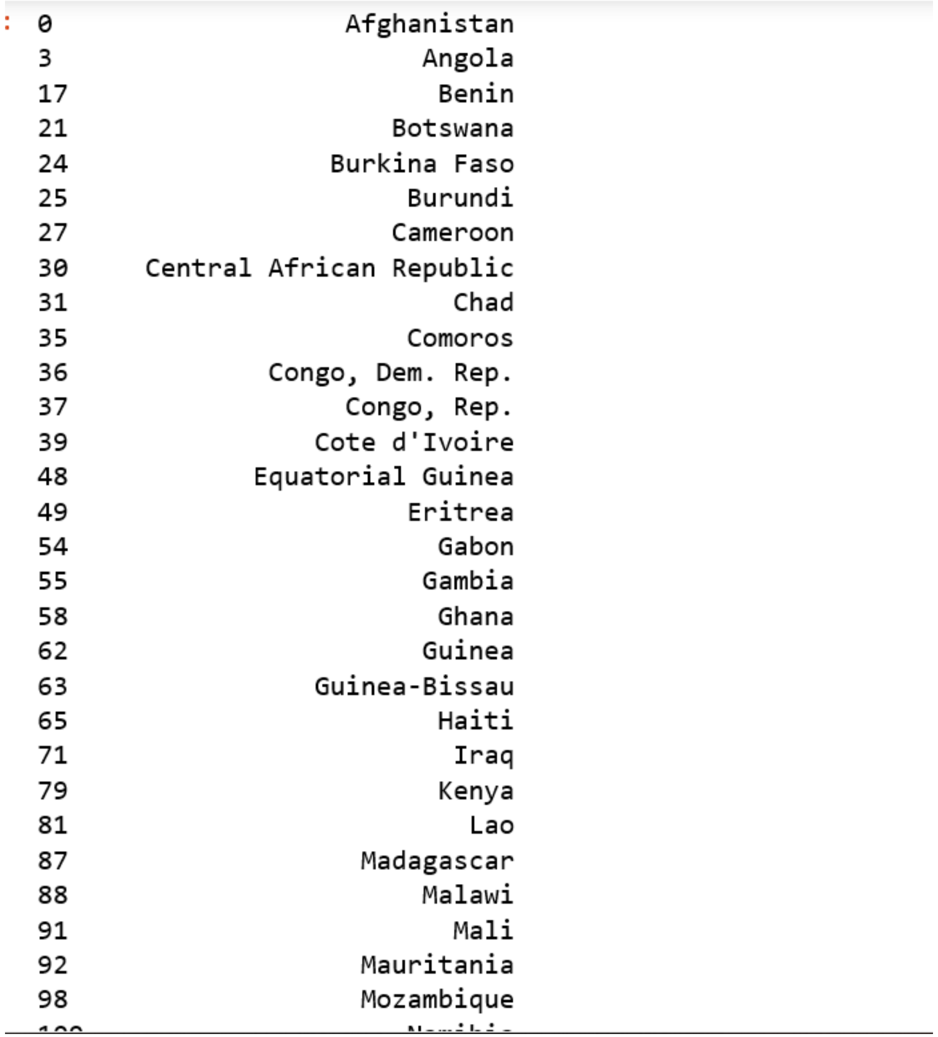
PC_1 mean is low for cluster 2 and from Correlation we got PC1 is positively correlated with life_expectancy, income and next gdpp also there. PC1 is negatively correlated with total_fer and child_mortality so PC1 is well explaining those features.

We will see the features behaviour in next slide



After analyzing the visualizations of PC's and original variables we can conclude that PC1's mean is very low for clusterid 2 which implies that Clusterid 2 has low income,low gdpp, high child_mort and high fert_rate.PC3's mean is little high for clusterid 2 compared to clusterid1 which implies that cluster2 is having high inflation and low health and low imports and exports of clusterid2.So clusterid 2 here needs more aid from the foundation.

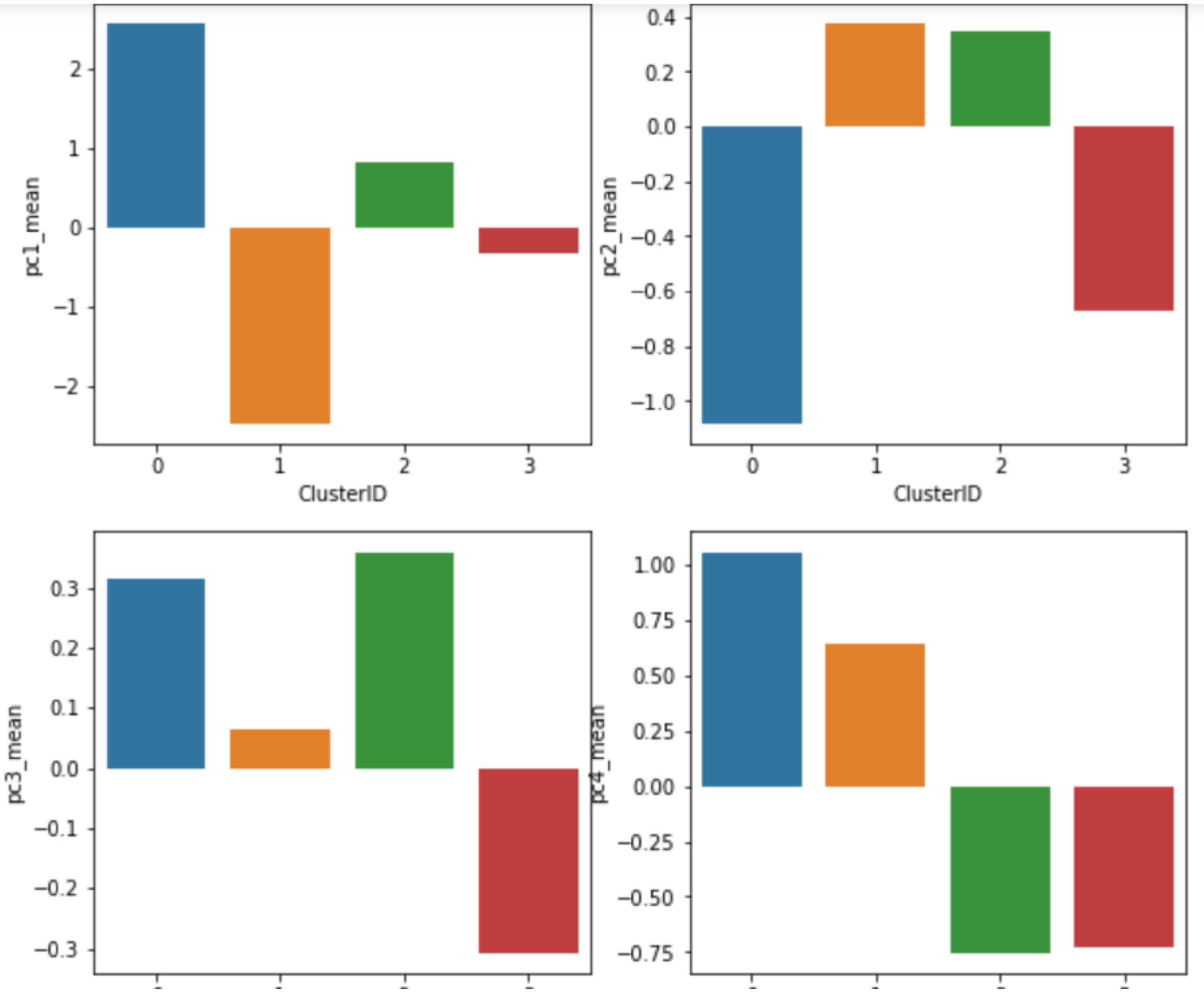
List of few countries which are in aid for finacnical aid which we have selected cluster2



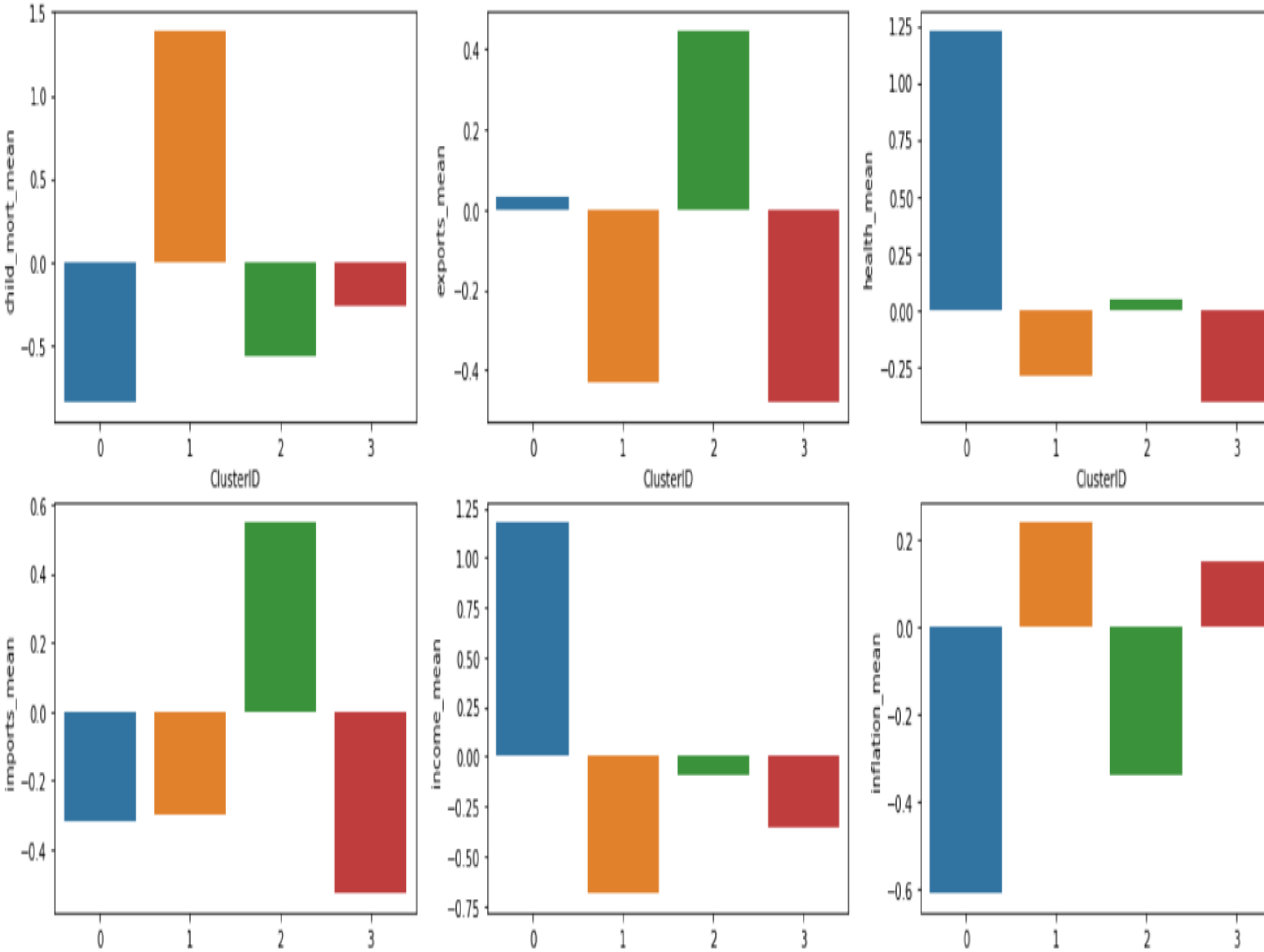
```
0                    Afghanistan
3                         Angola
17                         Benin
21                      Botswana
24                  Burkina Faso
25                       Burundi
27                      Cameroon
30      Central African Republic
31                          Chad
35                       Comoros
36              Congo, Dem. Rep.
37                   Congo, Rep.
39                 Cote d'Ivoire
48             Equatorial Guinea
49                       Eritrea
54                         Gabon
55                        Gambia
58                         Ghana
62                        Guinea
63                 Guinea-Bissau
65                         Haiti
71                          Iraq
79                         Kenya
81                           Lao
87                    Madagascar
88                        Malawi
91                          Mali
92                    Mauritania
98                    Mozambique
100                      Namibia
```

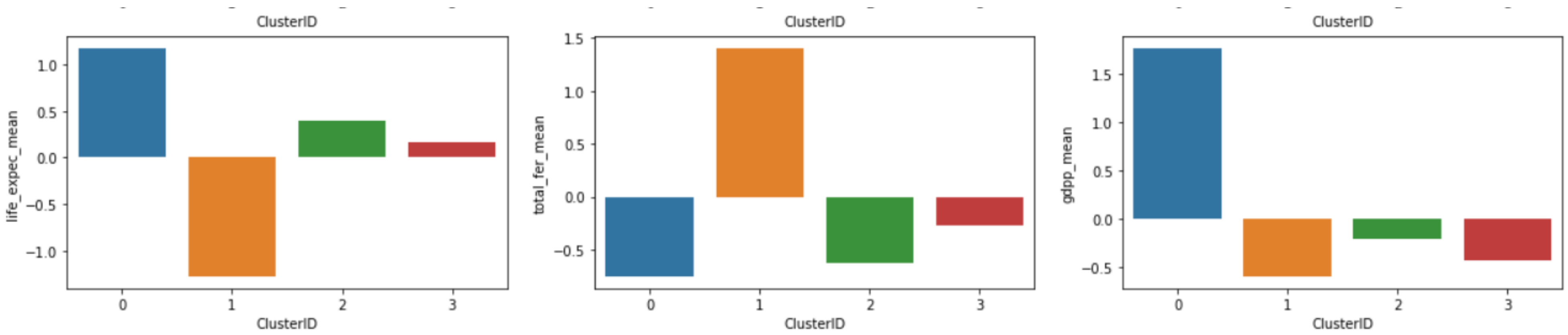
K-means with K=4

Visualizing the PC’s with 4 clusters

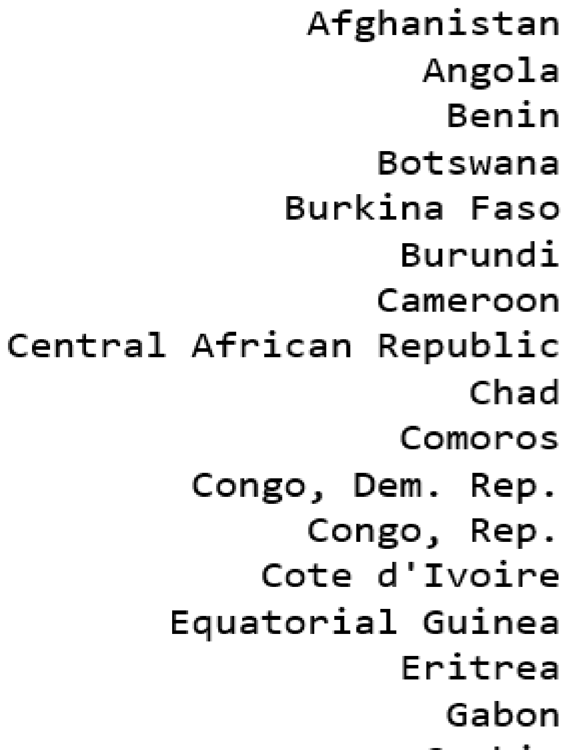


Visualizing the features with mean of features as per the cluster





SO from above visualizations we can observe that PC1's mean is very low for cluster id '1' which implies that Clusterid 1 has low income,low gdpp, high child_mort and high fert_rate and low life_expec and it has high inflation compared to clusterid 3 and clusterid 2, so we can say that cluster1 is looking for more aid

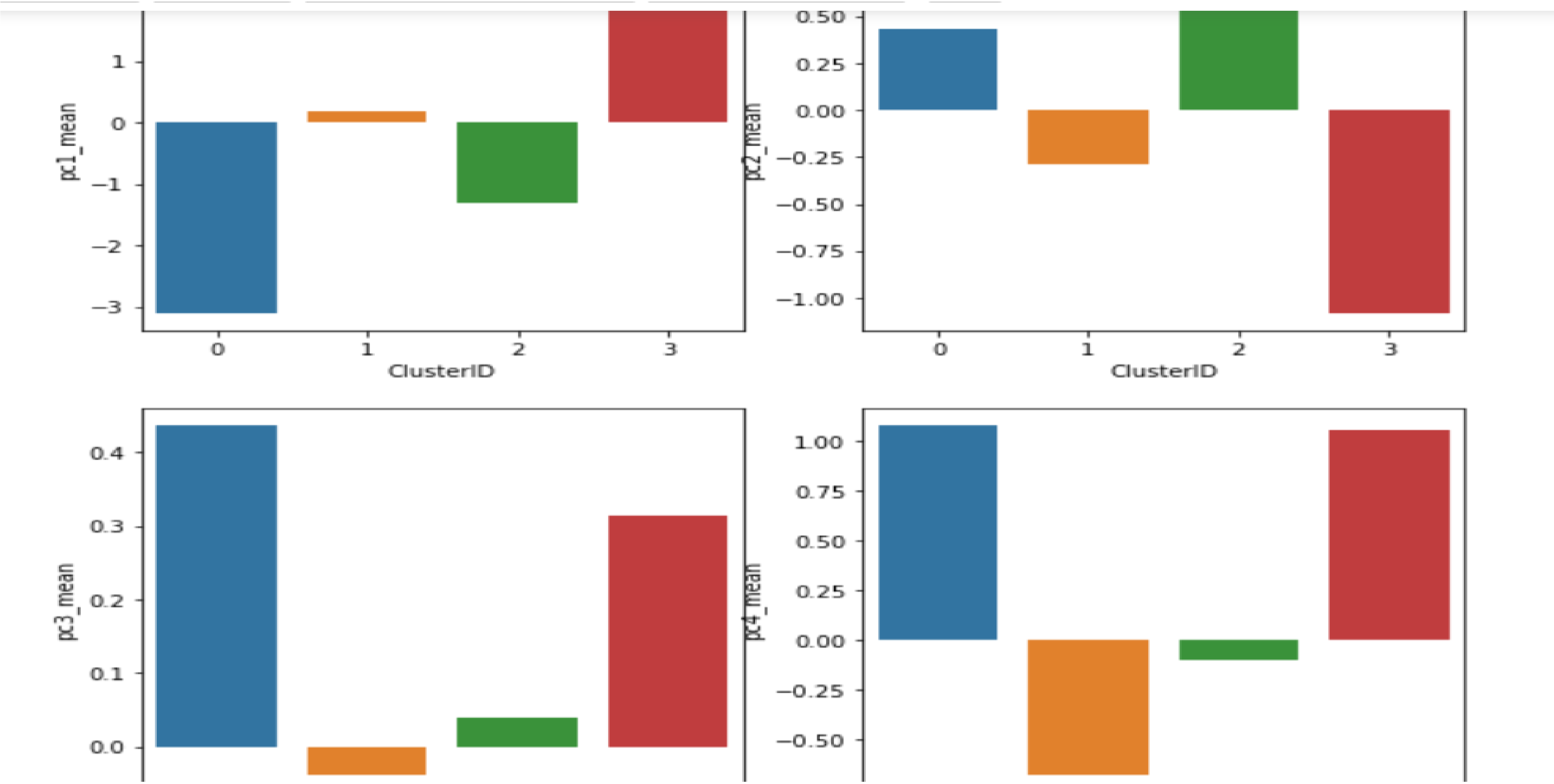


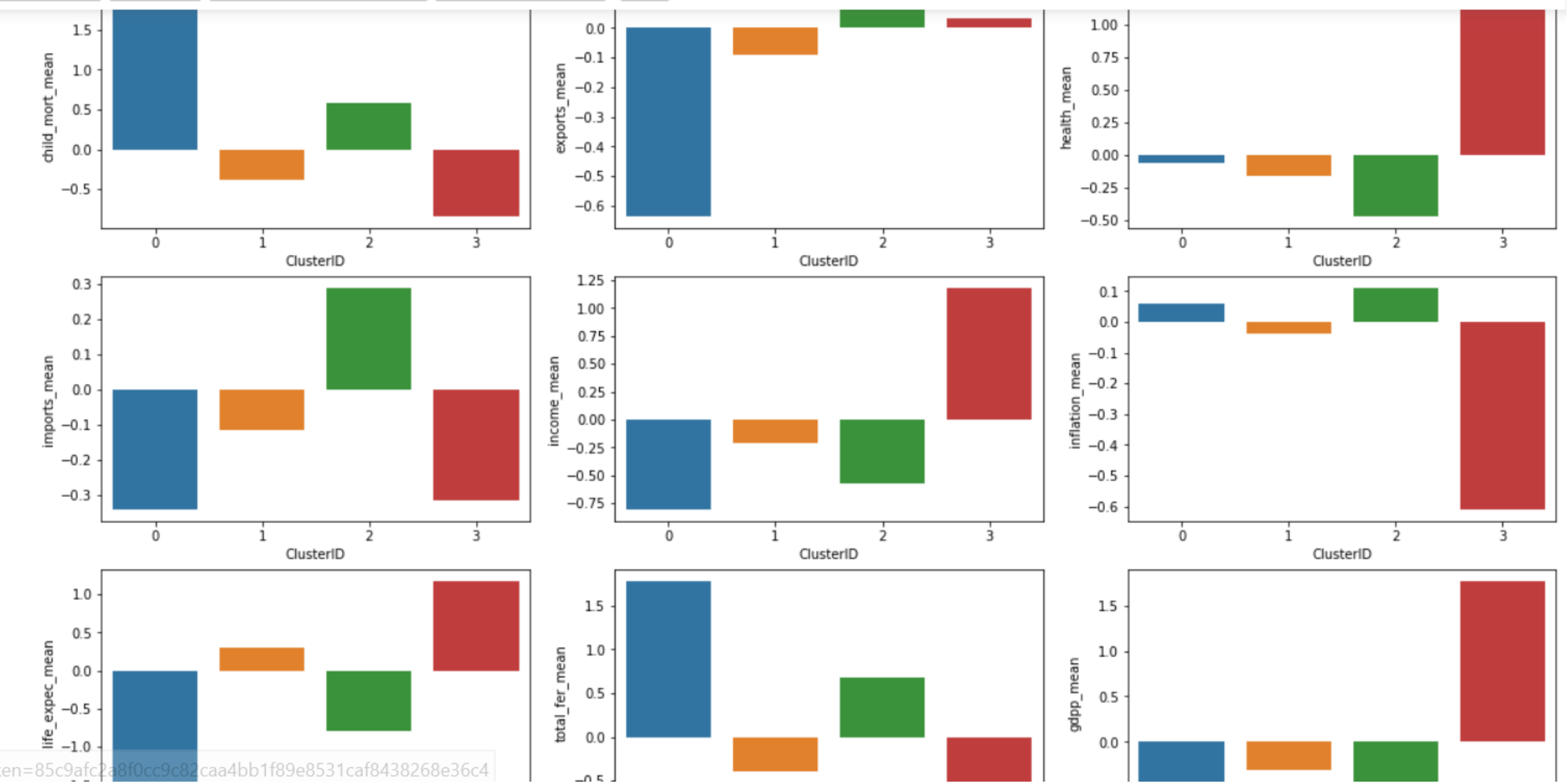
```
             Afghanistan
                  Angola
                   Benin
                Botswana
            Burkina Faso
                 Burundi
                Cameroon
Central African Republic
                    Chad
                 Comoros
        Congo, Dem. Rep.
             Congo, Rep.
           Cote d'Ivoire
       Equatorial Guinea
                 Eritrea
                   Gabon
```

Few top countries which are looking for aid by taking when k=4 clusters selcting The PC1 as main component

Hierarchical Clustering

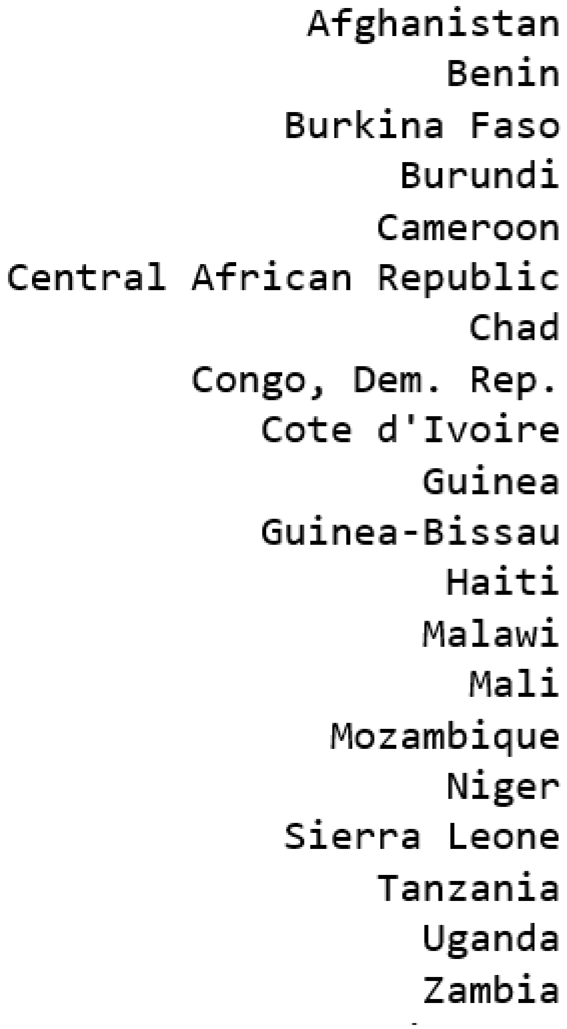
By using complete and single linkage method we have done the hierarchical clustering and obtained the results.





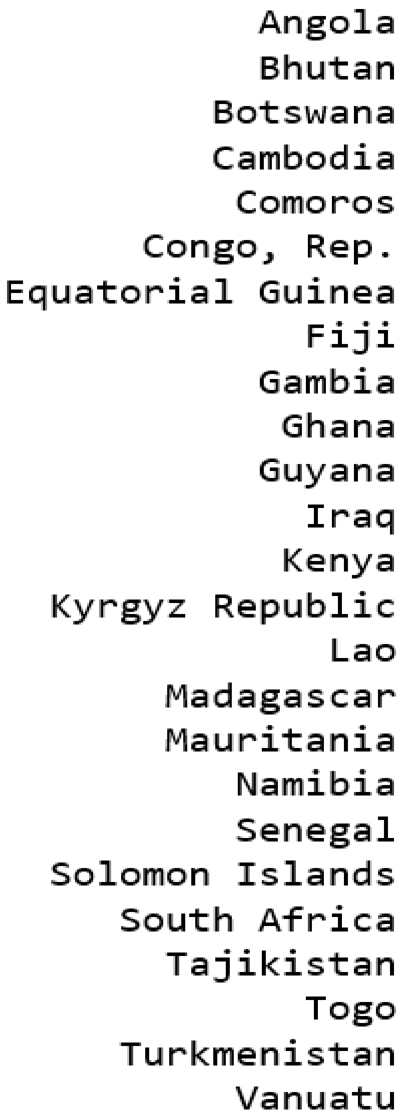
from above graaphs we can check for cluster id 0 which has pc1 mean is v ery low which means low gdpp and income, high child_mort ,high life_expec, high total_fer

Cluster 0 selected countries



```
             Afghanistan
                   Benin
            Burkina Faso
                 Burundi
                Cameroon
Central African Republic
                    Chad
        Congo, Dem. Rep.
           Cote d'Ivoire
                  Guinea
           Guinea-Bissau
                   Haiti
                  Malawi
                    Mali
              Mozambique
                   Niger
            Sierra Leone
                Tanzania
                  Uganda
                  Zambia
```

Cluster 2 selected countries



```
           Angola
           Bhutan
         Botswana
         Cambodia
          Comoros
      Congo, Rep.
Equatorial Guinea
             Fiji
           Gambia
            Ghana
           Guyana
             Iraq
            Kenya
  Kyrgyz Republic
              Lao
       Madagascar
       Mauritania
          Namibia
          Senegal
  Solomon Islands
     South Africa
       Tajikistan
             Togo
     Turkmenistan
          Vanuatu
```

apart from cluster id 0, cluster id2 also looking for financial aid, so from hierarchical clustering, we have got the Above countries list which are same as results we got for k-means. So we can conclude that for any clustering method, We have received similar list of countries which are looking for more financial aid .