Clustering & PCA Assignment

NGO Help International

Submitted by–

Aayushi Aggarwal

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. IT RUNS A LOT OF OPERATIONAL PROJECTS FROM TIME TO TIME ALONG WITH ADVOCACY DRIVES TO RAISE AWARENESS AS WELL AS FOR FUNDING PURPOSES. AFTER THE RECENT FUNDING PROGRAMMES, 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.

Data Understanding

- Total available corpus --$10 million
- Total countries – 167
- The data contains country wise parameter as –

  1. Net income per person

  2. The measurement of the annual growth rate of the Total GDP

  3. The average number of years a new born child would live if the current mortality patterns are to remain the same

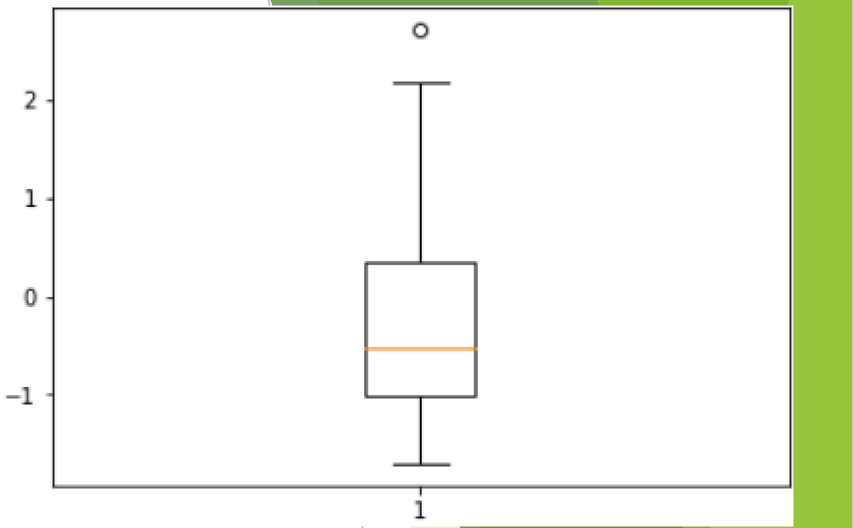
  4. The number of children that would be born to each woman if the current age-fertility rates remain the same.

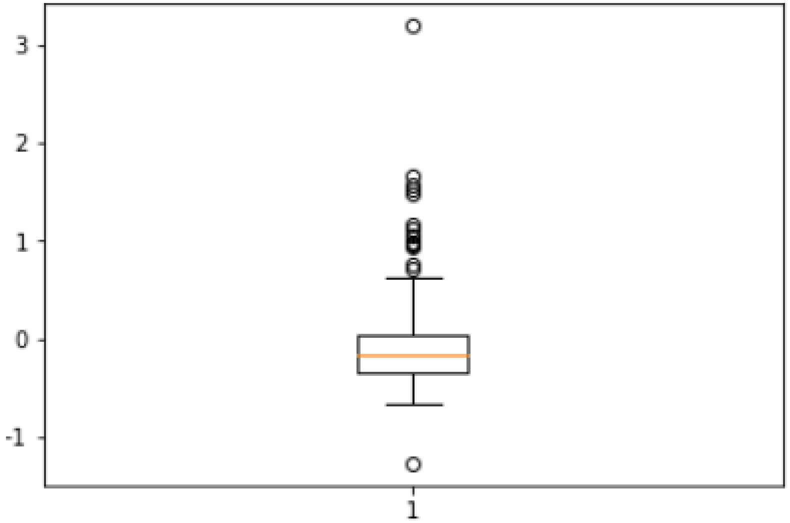
  5. The GDP per capita. Calculated as the Total GDP divided by the total population.

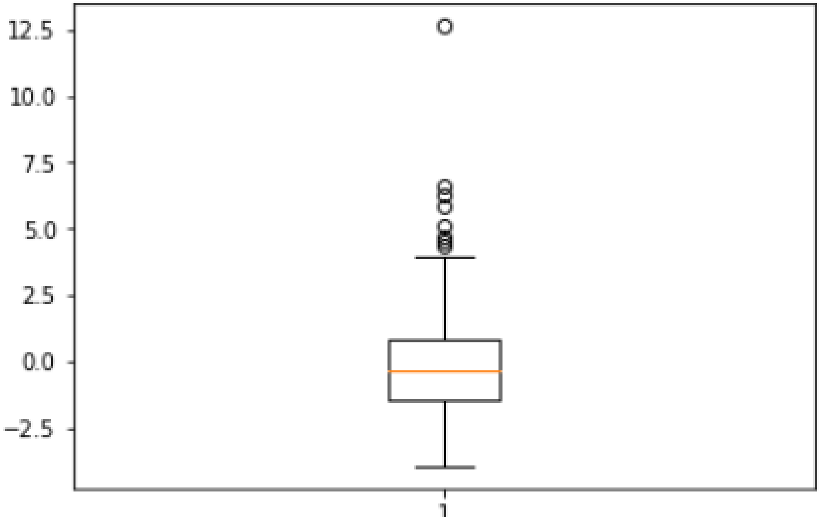
- 6. Death of children under 5 years of age per 1000 live births

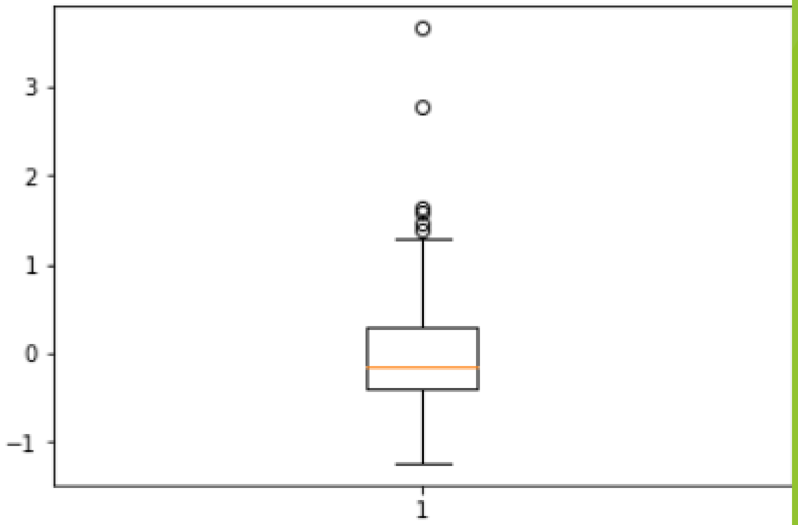
Data Preparation

- Outlier Analysis



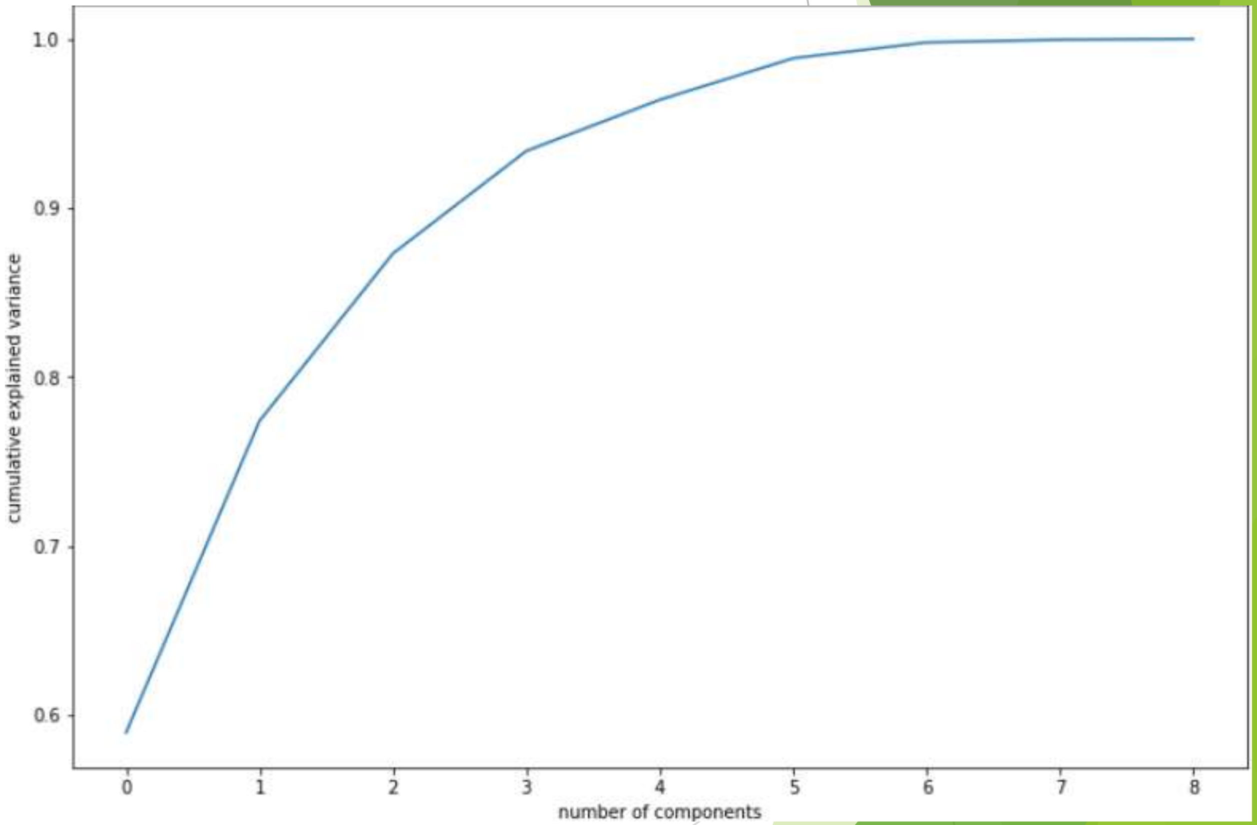






Analysis –PCA(Principal Component Analysis) and Incremental PCA

- On plotting the correlation matrix, it was found that the parameters are highly correlated. So for removing the correlation and for dimensionality reduction, Principal Component Analysis (PCA) needs to be done.
- First, On analysing the Scree Plot, it is found that 4 Principal Components (PC) are enough to explain almost 95% of the variance. So PCA is performed using no. of PCs as 4.

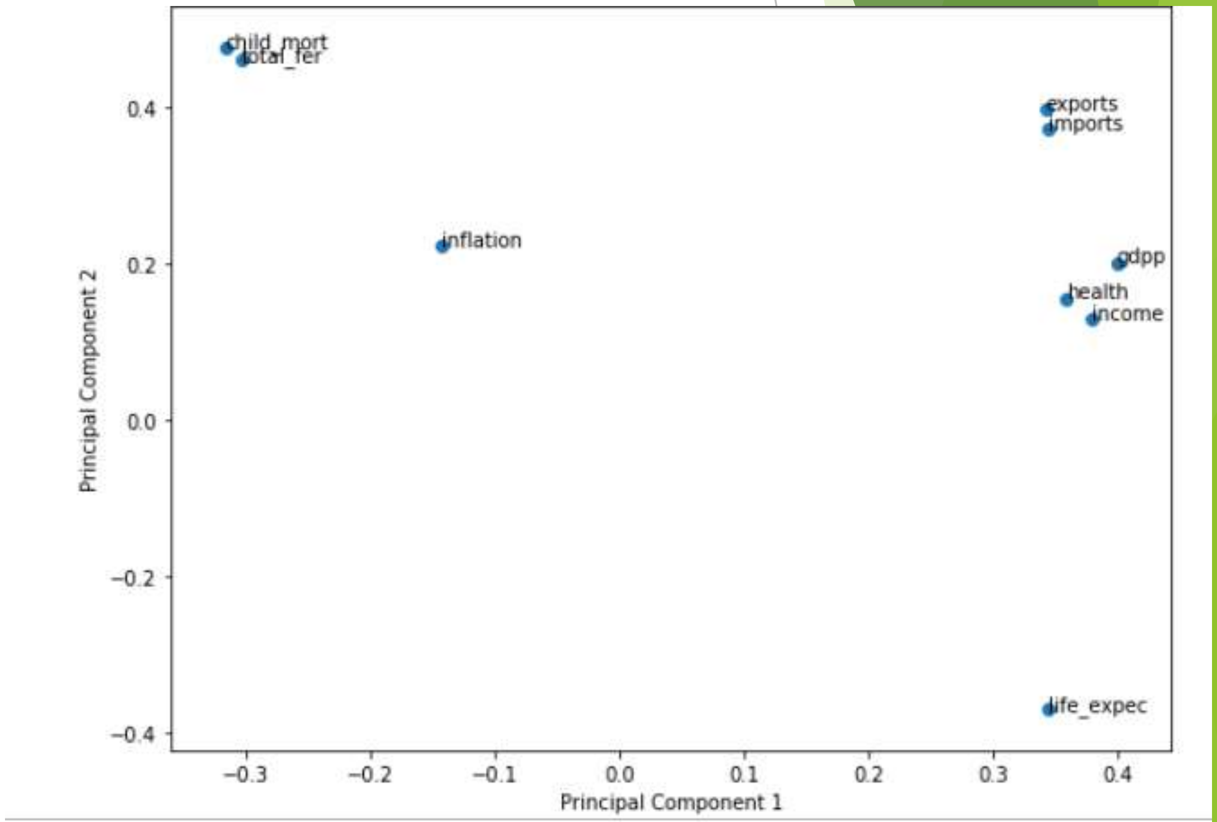


Analysis –PCA(Principal Component Analysis) and Incremental PCA

- The data set is now transformed into PCs. We see that the first component, PC1 is in the direction of 'income', 'gdpp‘, ‘export’, ‘imports’, ‘life expectancy’ and health'. The second component PC2 is in the direction of 'child_mort', 'total_fertility’. So the data points with low PC1 and high PC2 are to be selected for the aid.

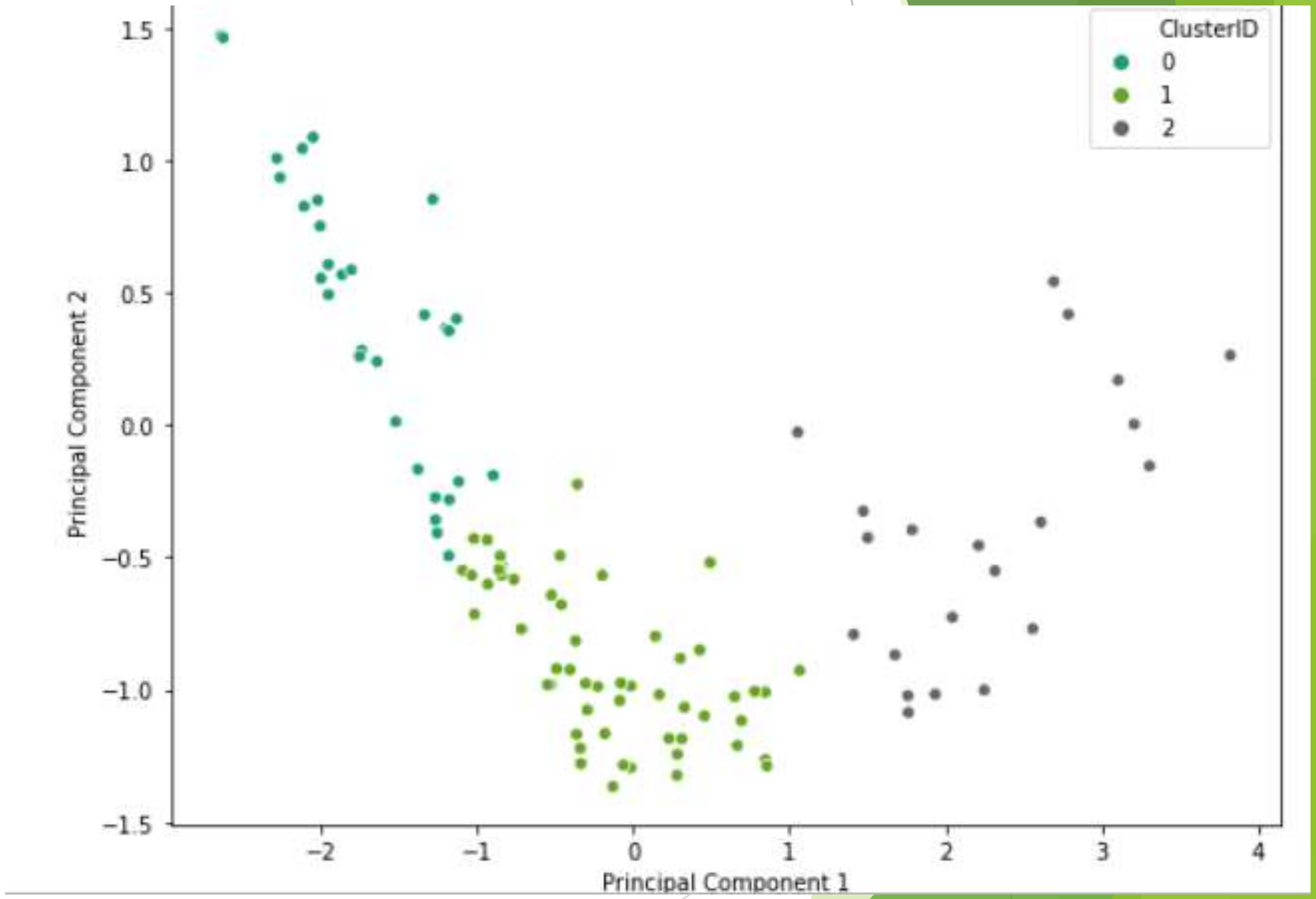
- Now the data needs to be clustered for finding the countries with low PC1 and high PC2.

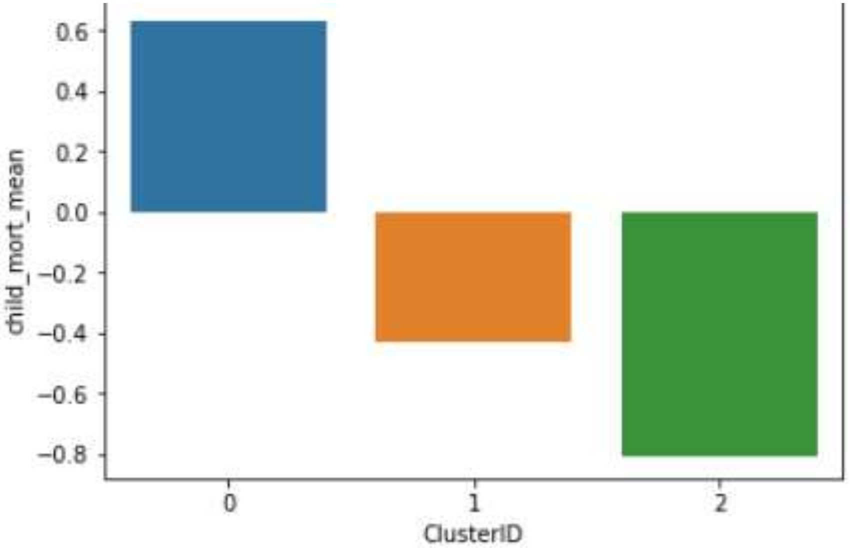
- Both K-Means and Hierarchichal Clustering is done.

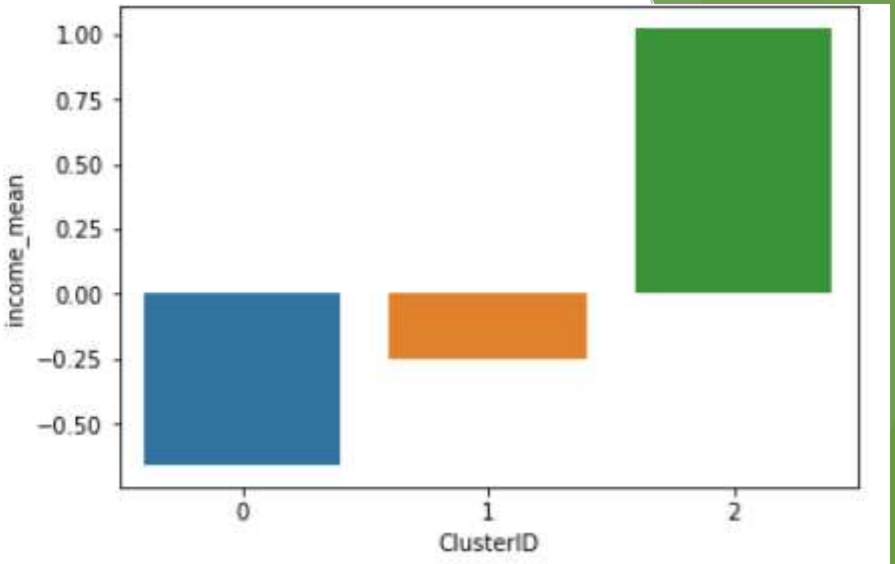


Analysis: Clustering (K-means)

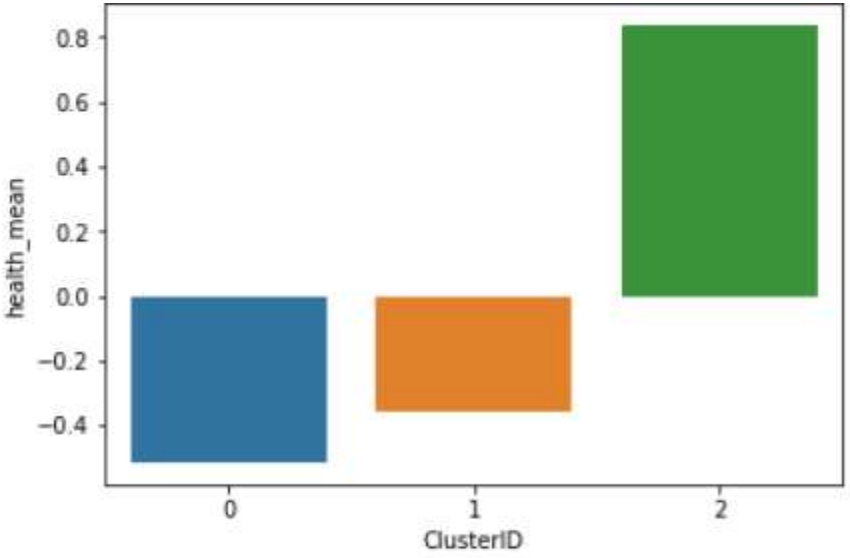
- First Silhouette Score and Sum of Squared
- Distances is calculated for various values of k,
- from 3 to 8.
- On Analysing, it was found that both k=3 are
- suitable..
- Therefore, Clustering is done with both k=3
- After Clustering, for each Cluster Id, the
- mean values of all PCs and some original
- features are calculated and plotted.

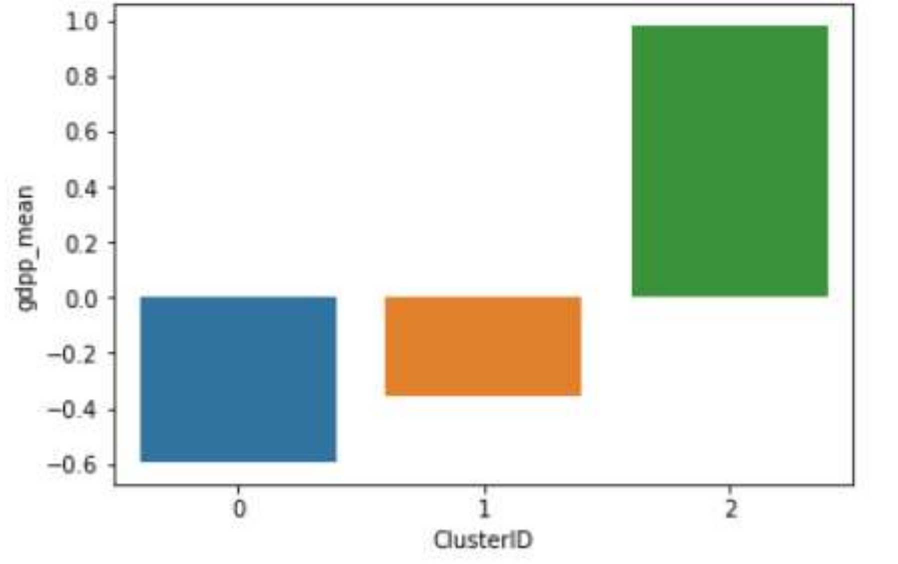






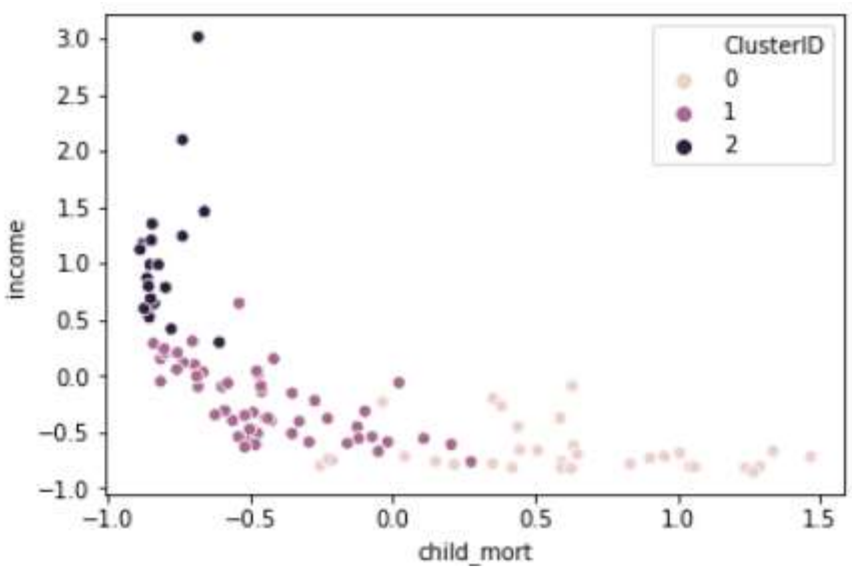
We can observe countries in cluster 0 are in dire need of aid.

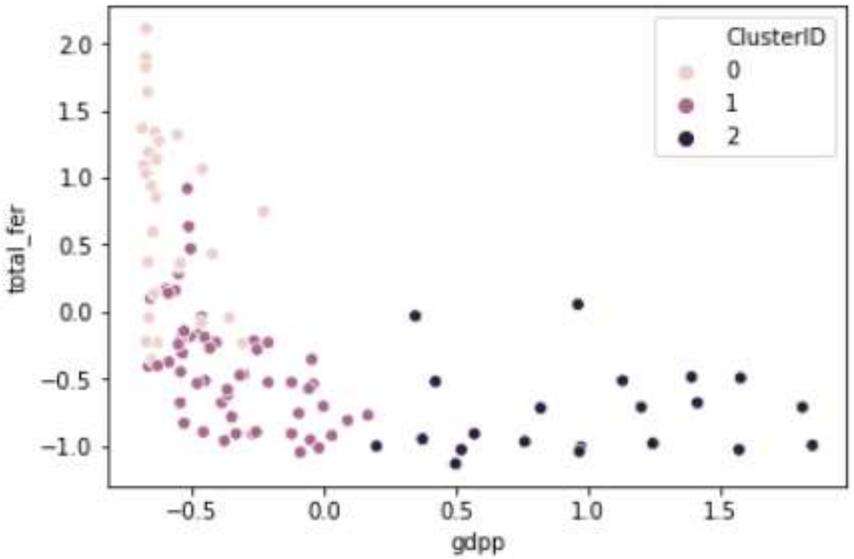


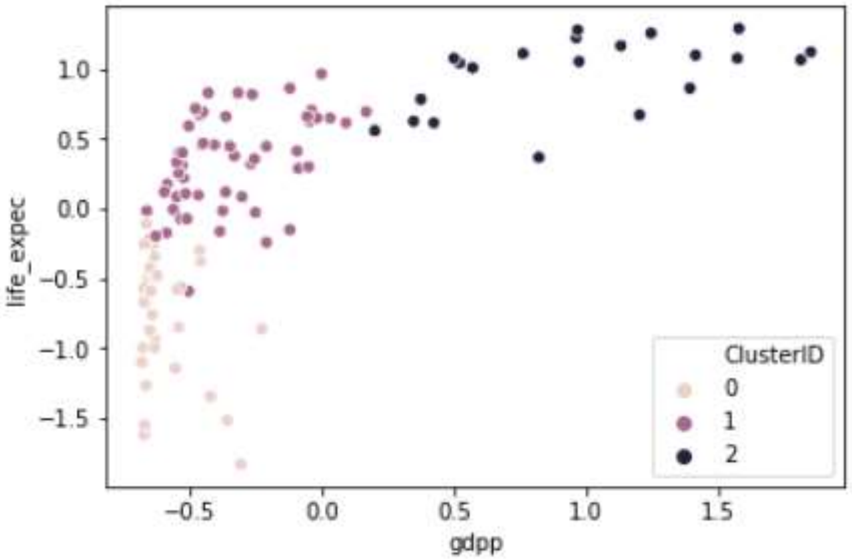


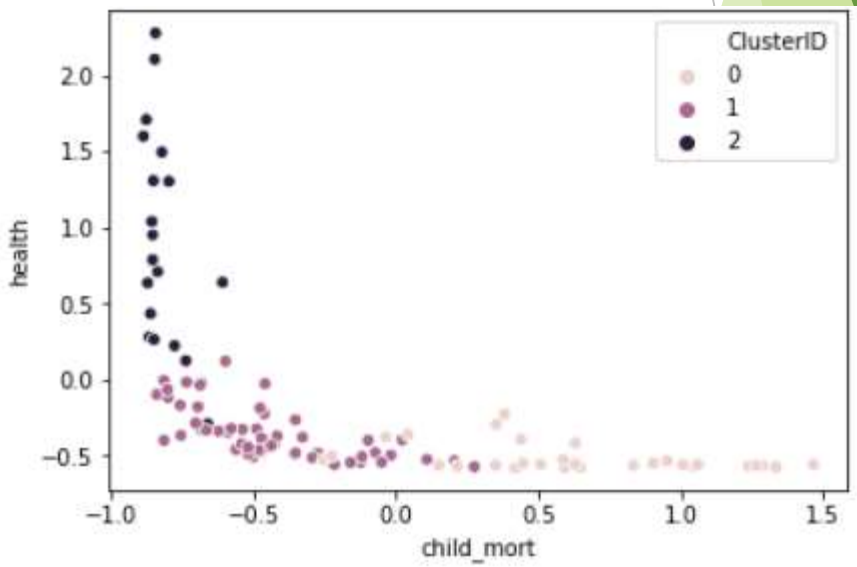
Observations

- We can observe countries in cluster 0 are in dire need of aid.
- It is having high child mortality and high total_fer
- It is having low income, low health, low life_expec and low gdpp
- These are some of the factors which majorly help in distinguishing developed countries from under developed countries.
- Let’s see some scatter plots to make our above observations concrete


ClusterID
0
1
2
income
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
child_mort
−1.0
−0.5
0.0
0.5
1.0
1.5

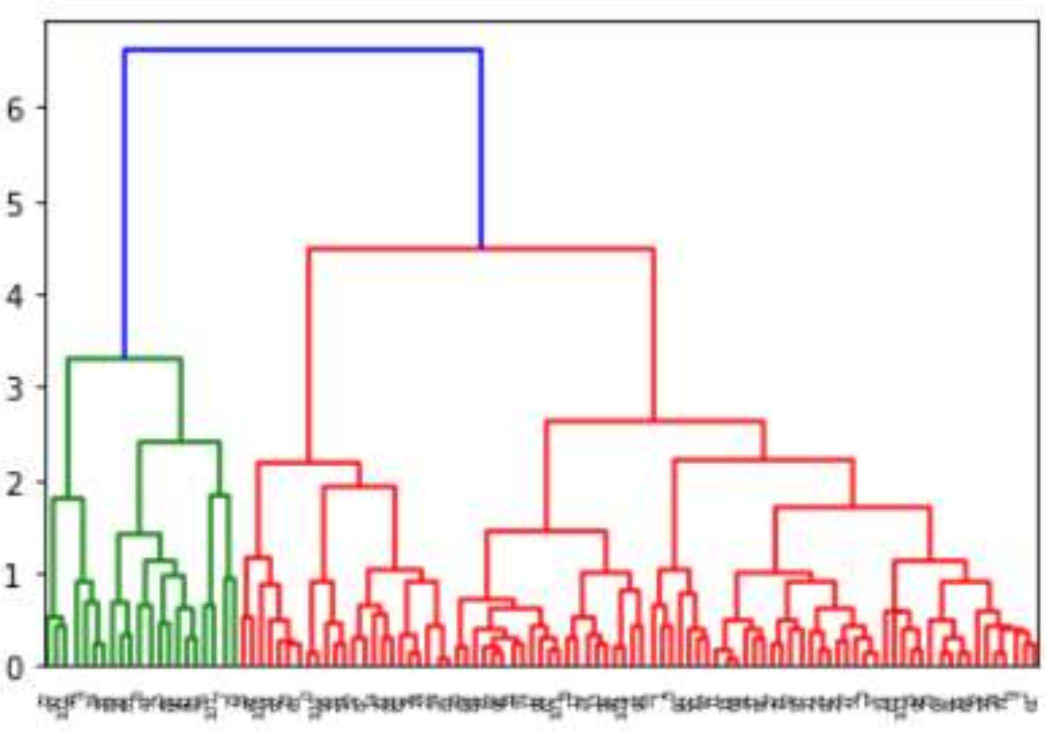

ClusterID
0
1
2
total_fer
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
gdpp
−0.5
0.0
0.5
1.0
1.5

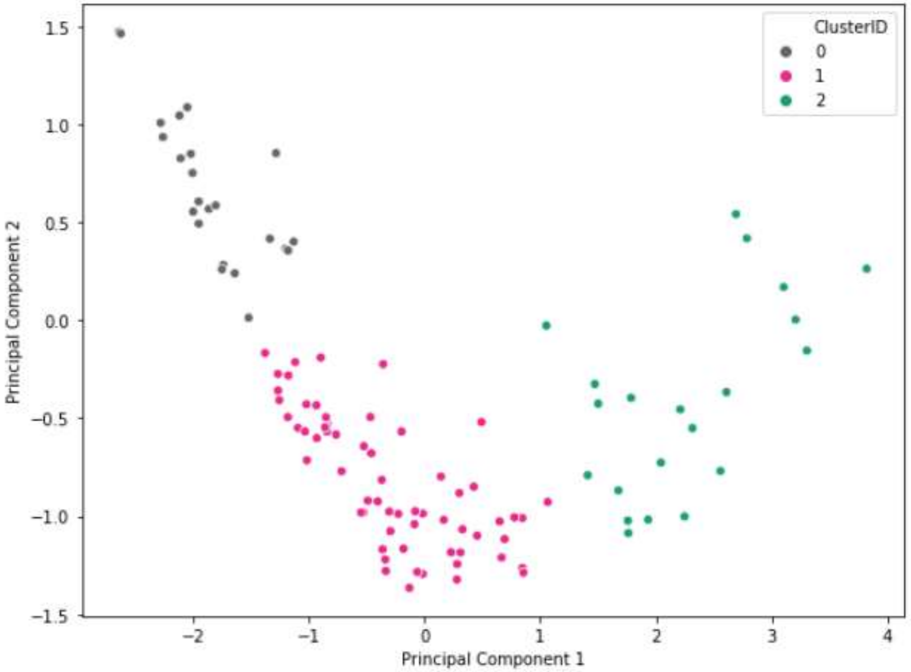

ClusterID
0
1
2
life_expec
1.0
0.5
0.0
−0.5
−1.0
−1.5
gdpp
−0.5
0.0
0.5
1.0
1.5

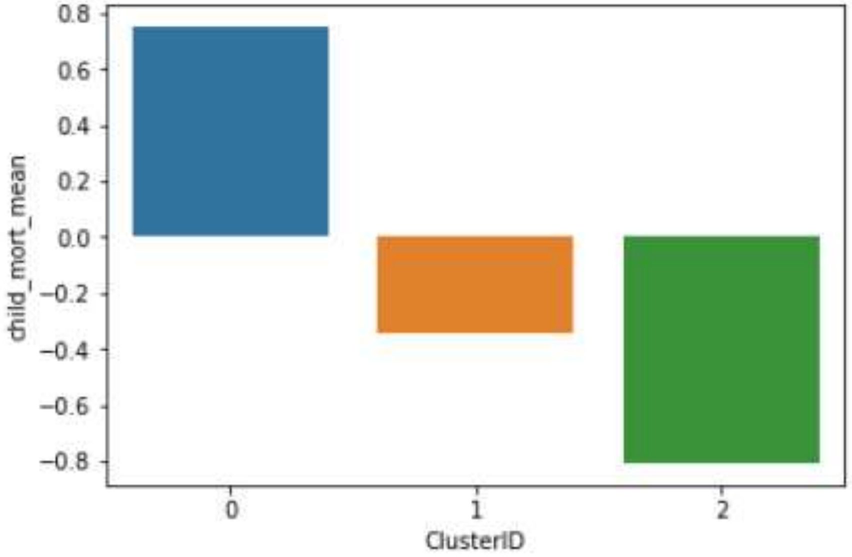

ClusterID
0
1
2
health
2.0
1.5
1.0
0.5
0.0
−0.5
child_mort
−1.0
−0.5
0.0
0.5
1.0
1.5

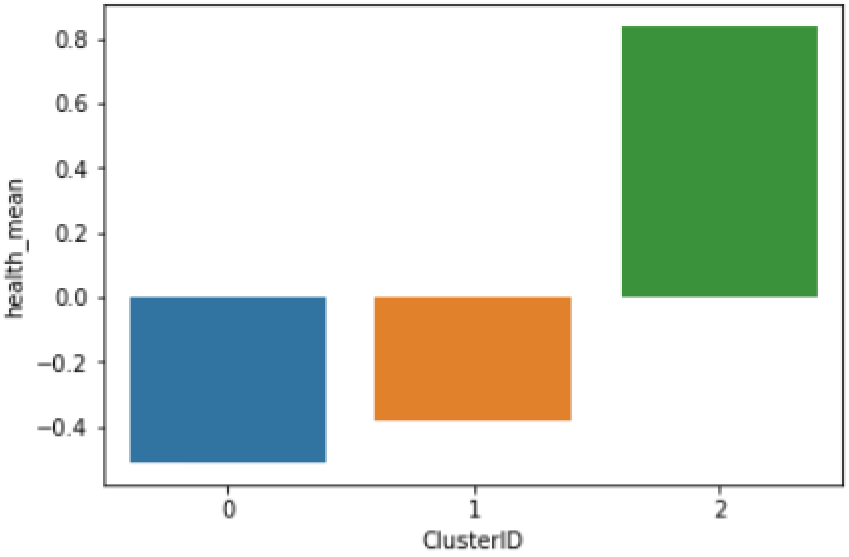
Analysis: Clustering (Hierarchichal)

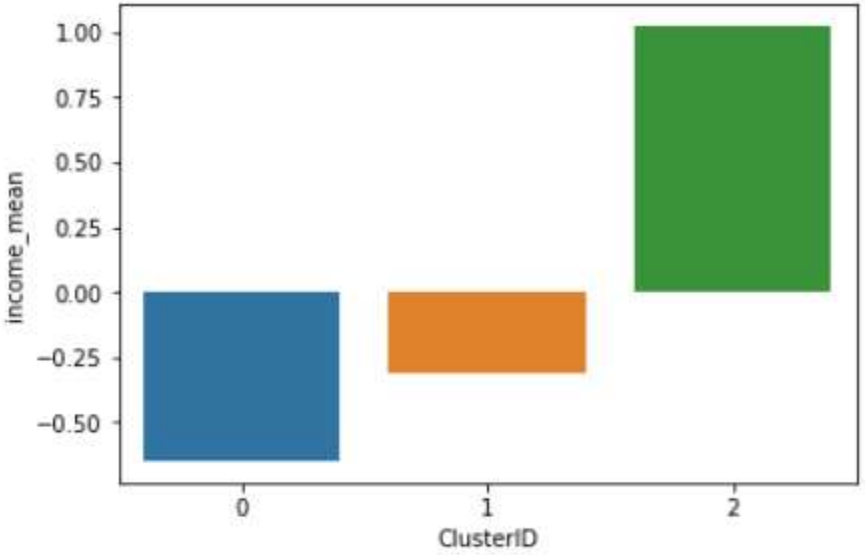
- First Dendrogram is created and it is cut at an appropriate place to obtain the no. of clusters (3)
- Now clustering is done with no of clusters as 3.
- For each Cluster Id, the mean values of all PCs and some original features are calculated and plotted.




ClusterID
0
1
2
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
Principal Component 2
−2
−1
0
1
2
3
4
Principal Component 1

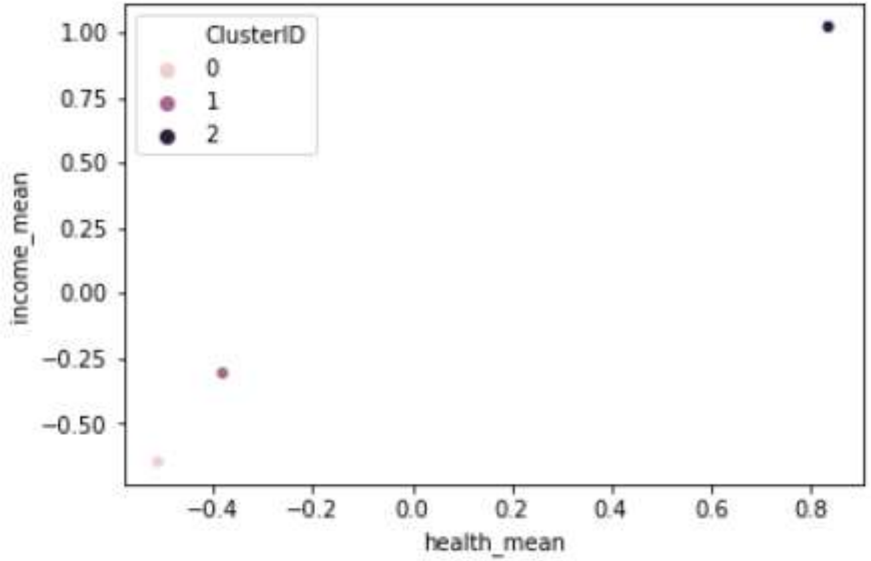

0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
child_mort_mean
0
1
2
ClusterID

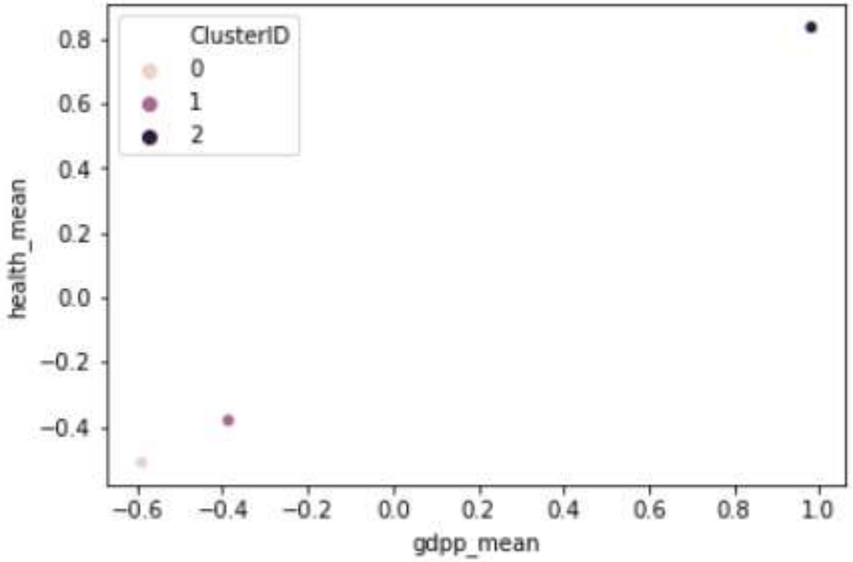

0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
health_mean
0
1
2
ClusterID


1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
income_mean
0
1
2
ClusterID

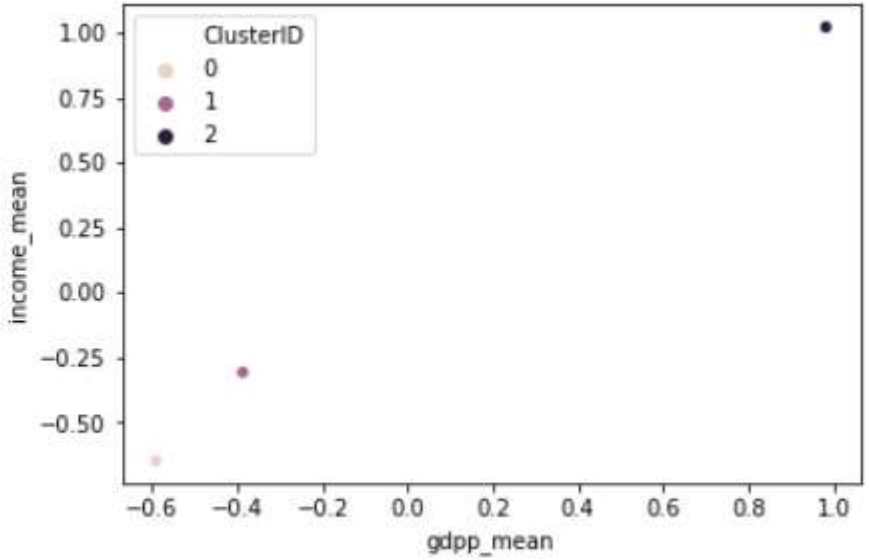
Observations

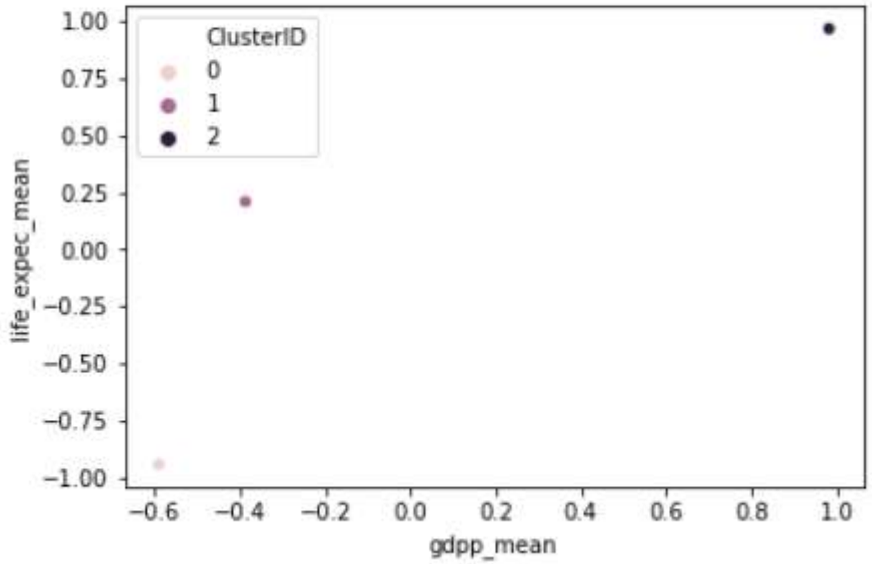
- We can observe countries in cluster 0 are in dire need of aid.
- It is having high child mortality and high total_fer
- It is having low income, low health, low life_expec and low gdpp
- These are some of the factors which majorly help in distinguishing developed countries from under developed countries.
- Let’s see some scatter plots to make our above observations concrete



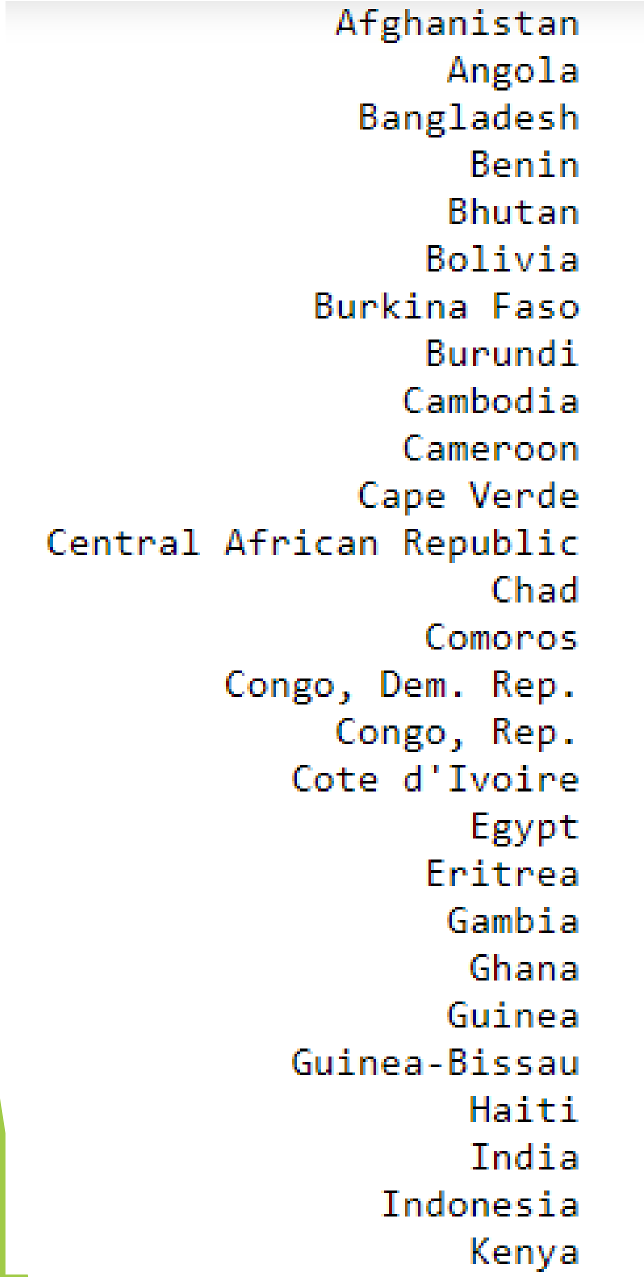


However some countries are missing in the Cluster 0 of the Hierarchichal Clustering. So we choose the Cluster obtained by the K-Means

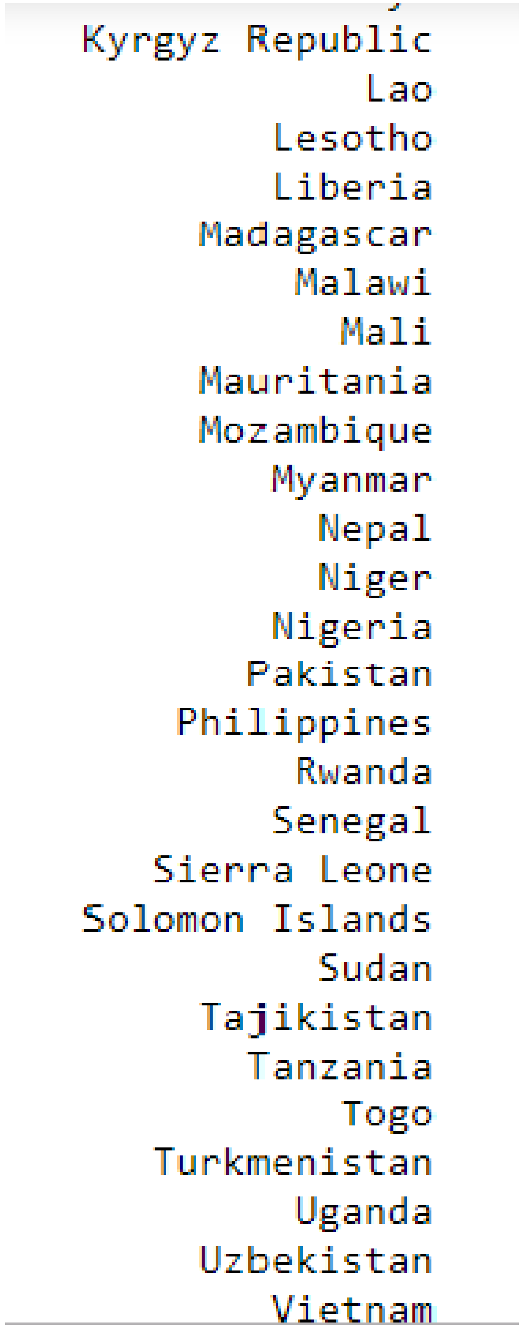




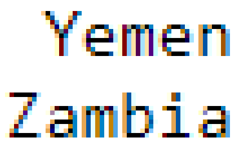
Countries in dire need of aid



```
             Afghanistan
                  Angola
              Bangladesh
                   Benin
                  Bhutan
                 Bolivia
            Burkina Faso
                 Burundi
                Cambodia
                Cameroon
              Cape Verde
Central African Republic
                    Chad
                 Comoros
        Congo, Dem. Rep.
            Congo, Rep.
           Cote d'Ivoire
                   Egypt
                 Eritrea
                  Gambia
                   Ghana
                  Guinea
           Guinea-Bissau
                   Haiti
                   India
               Indonesia
                   Kenya
```



```
Kyrgyz Republic
            Lao
        Lesotho
        Liberia
     Madagascar
         Malawi
           Mali
     Mauritania
     Mozambique
        Myanmar
          Nepal
          Niger
        Nigeria
       Pakistan
    Philippines
         Rwanda
        Senegal
   Sierra Leone
Solomon Islands
          Sudan
     Tajikistan
       Tanzania
           Togo
   Turkmenistan
         Uganda
     Uzbekistan
        Vietnam
```



```
 Yemen
Zambia
```

Conclusion

- 1. The above mentioned countries have low income, low GDPP,low life_expec,low health and high child mortality rate, high total_fer.
- 2. Hence the above mentioned countries are best for maximum fund investment.
- 3. K-means and Hierarchical don't produce identical insights.This depends on the way the principal components and the final number of clusters are chosen.

Recommendations

- 1. More jobs to be created to reduce the unemployment. Time to time hike should be given.
- 2. Infrastructure should be maintained such as educational institutions, hospitals etc.
- 3. Clealiness and hospitals should be there to improve the health conditions in the surroundings.
- 4. GDP includes what is spent on environmental protection, healthcare, and education,This should be kept in mind!
- 5. Implementation of each and eveything is more important after taking the reforms.