Clustering and PCA Case Study

Abstract

Problem Statement:

- HELP International is an international humanitarian NGO which 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.
- The NGO is able to raise around 10 Million.
- The CEO of the NGO needs to decide how to use this money strategically and effectively.

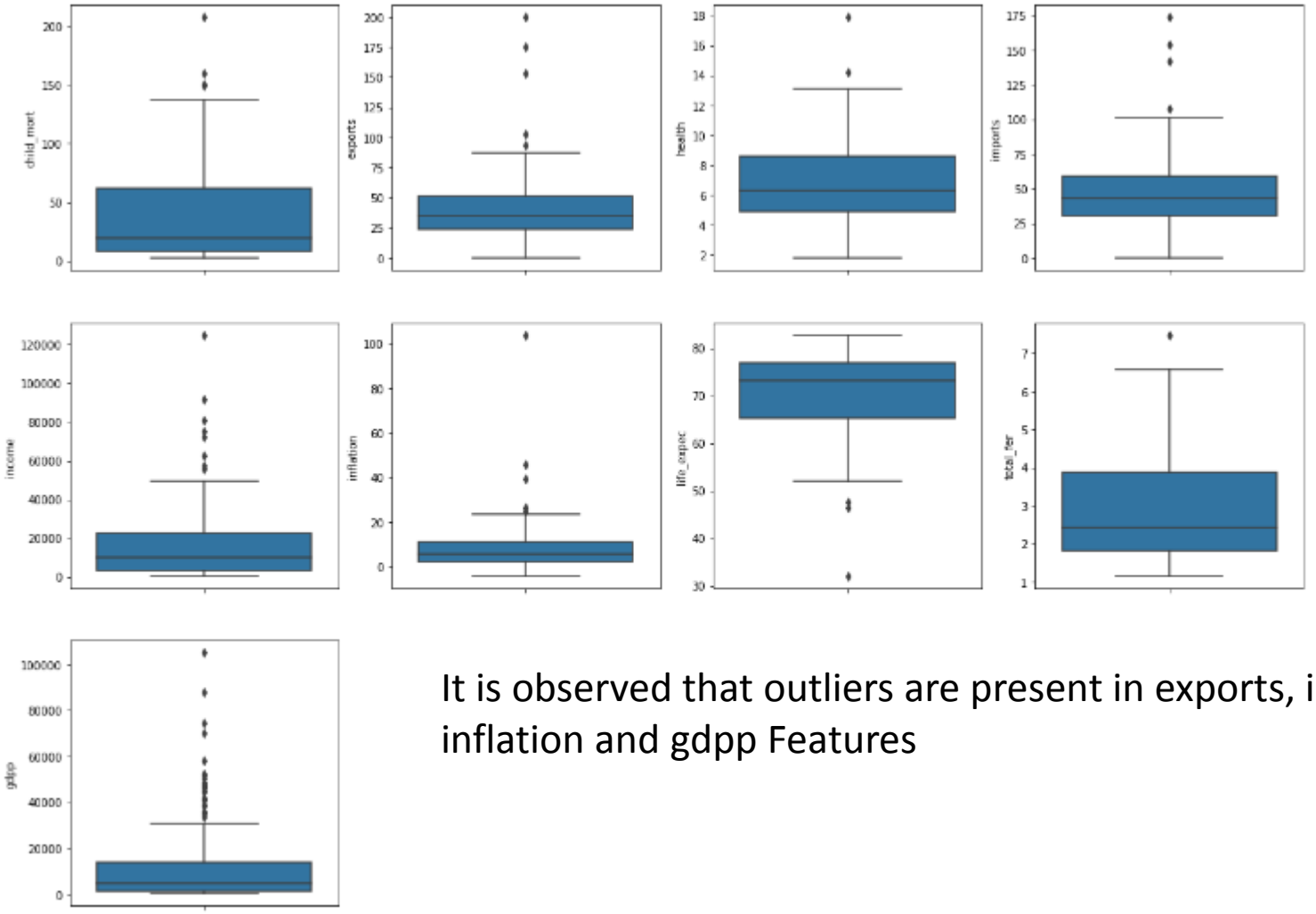
Goal:

- To choose the countries that are in the direst need of aid.

Approach For Analysis

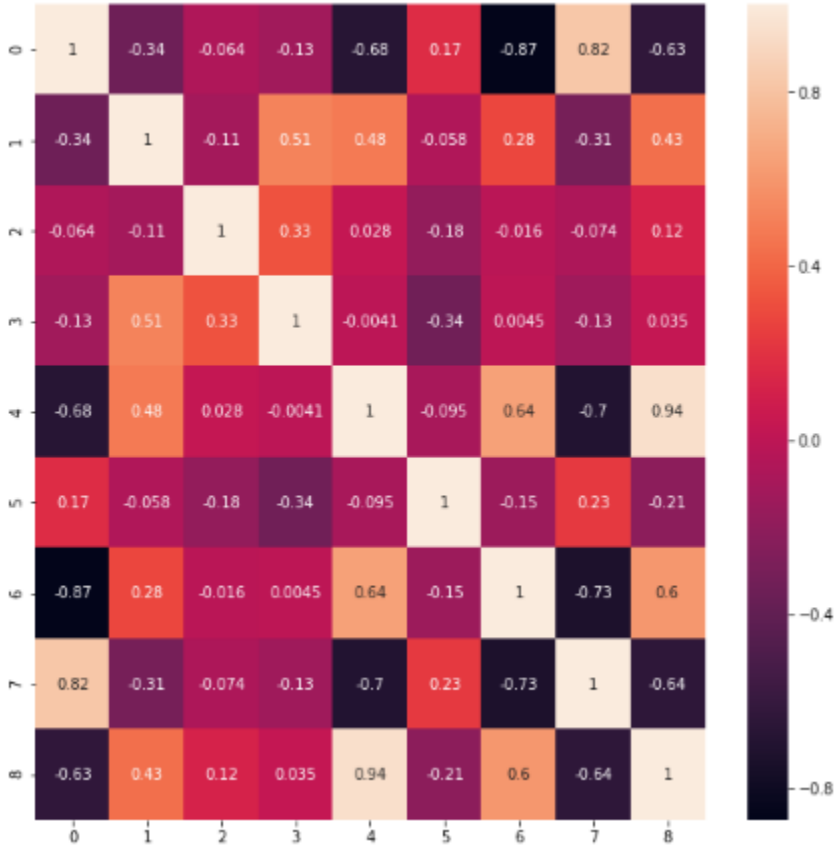
- Data Sourcing
- Performing EDA Analysis
- Remove the outliers
- Scale the Features
- Performed PCA
- Implemented K-Means Clustering
- Implemented Hierarchical clustering

Univariate Analysis



It is observed that outliers are present in exports, income, inflation and gdpp Features

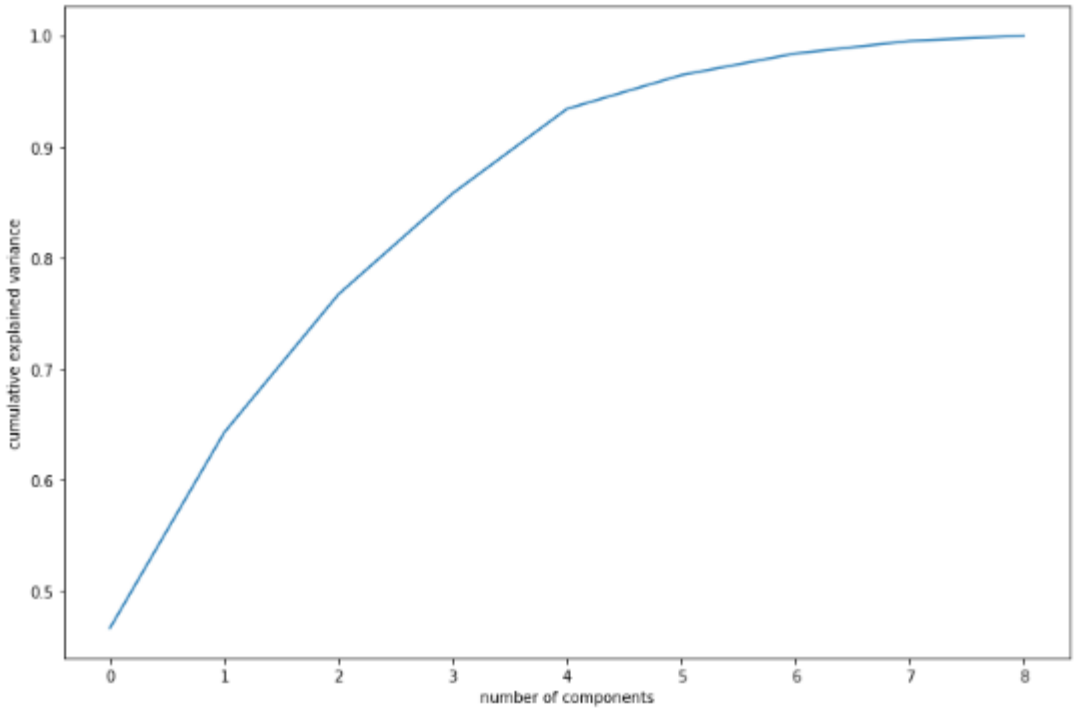
Correlation Matrix



There is correlation between most of the Features.

Using PCA, number of Features can be reduced

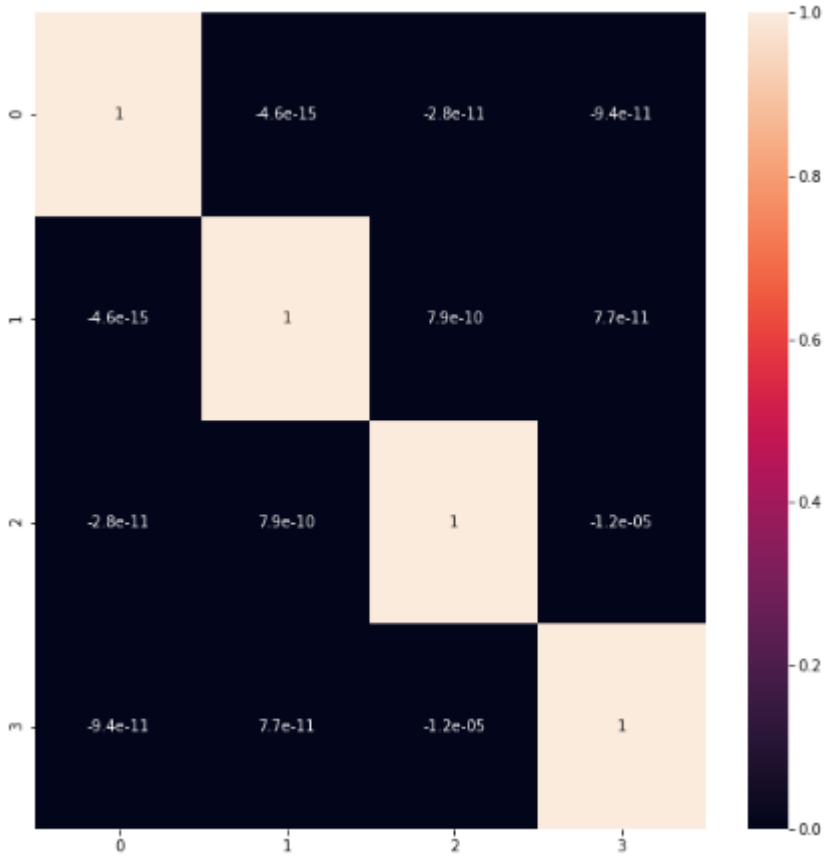
Scree Plot



It is observed that, 5 components are enough to describe 93% of the variance in the dataset

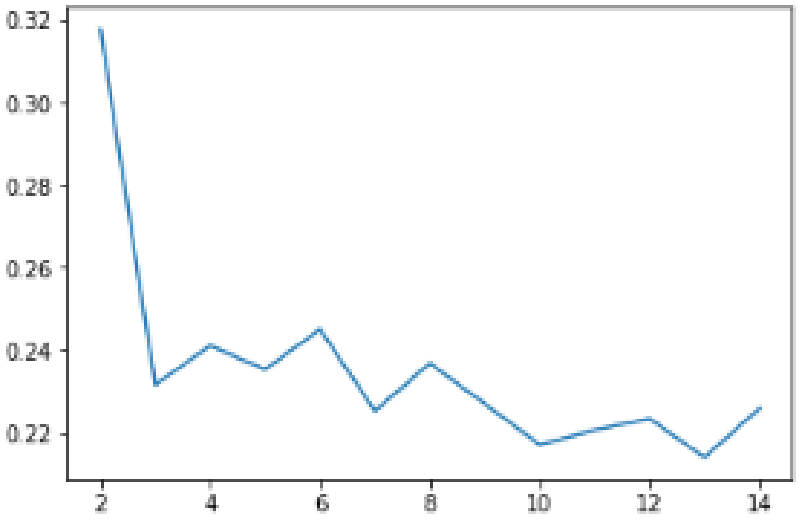
Hence 5 components are choosen

Correlation Matrix for PCA



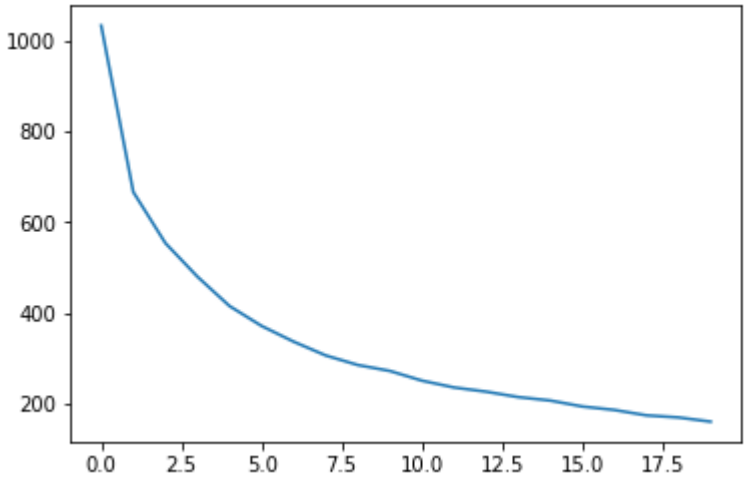
- **max corr:** 7.896799804358591e-10
  min corr: -1.2409147872206981e-05
- Indeed - there is no correlation between any two components.
- Multicollinearity is effectively removed
- Models will be much more stable

K- Means Analysis



Hopkins Value : 71% which indicates that clustering can be performed on the data

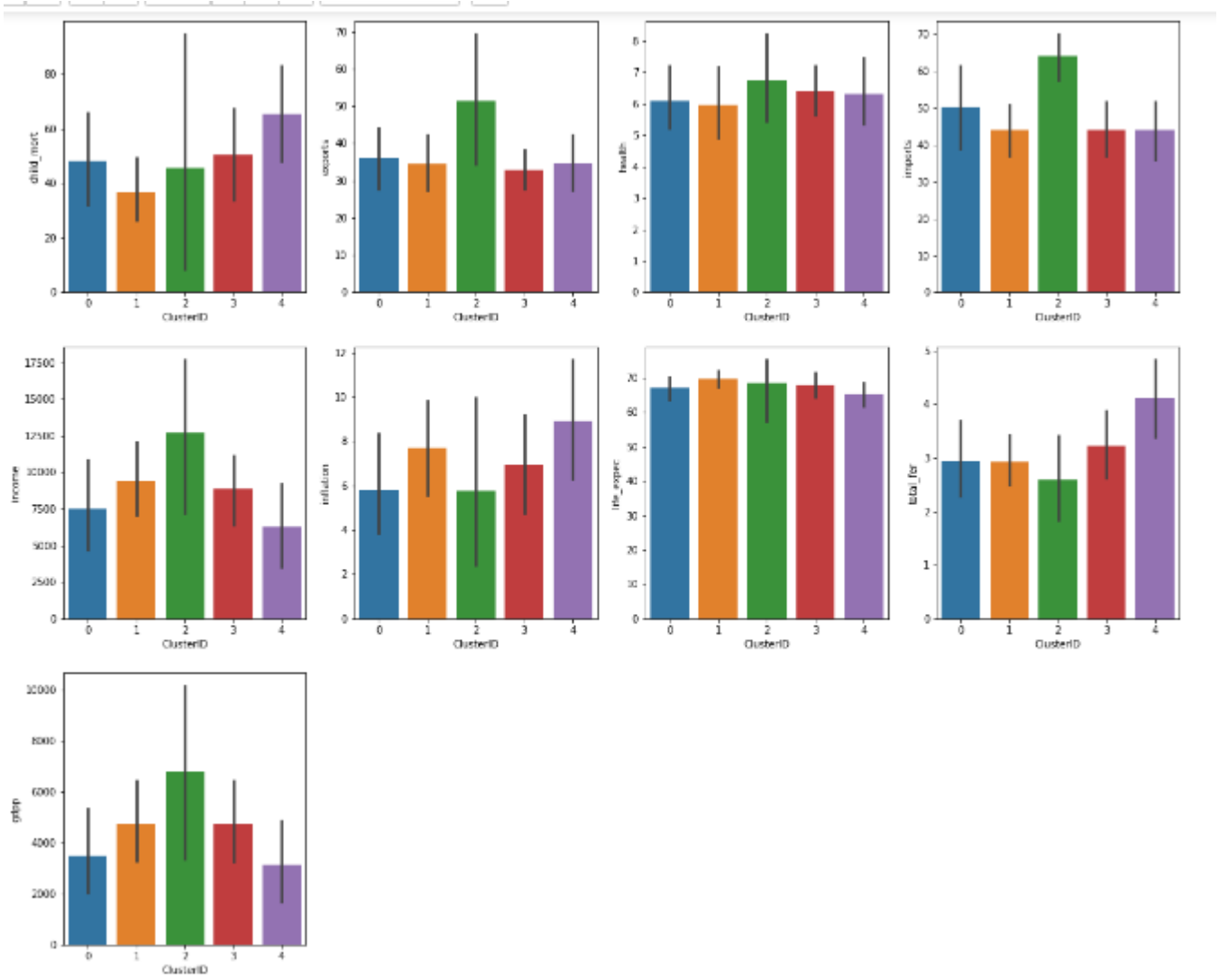
From the Silhouette graph for all the k values, we can say that there is a consistency within the data



From the Elbow curve, we can say that from k =5, there is a steady decrease in the value.

Hence we can consider K =5

K-Means Clustering Analysis



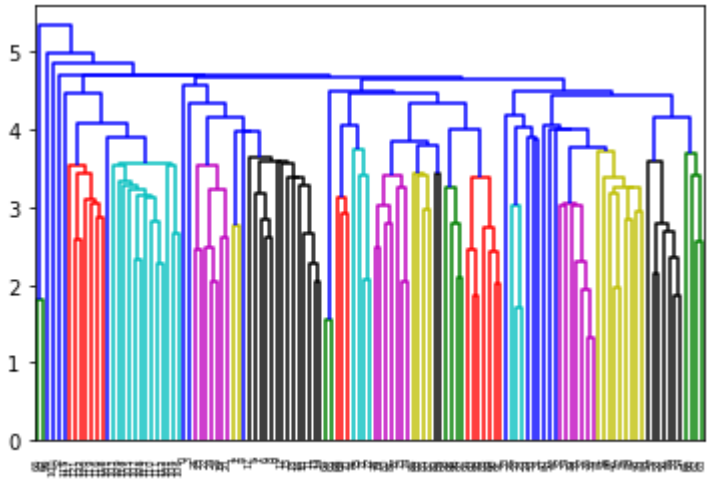
K-Means Observations

- There is no significant variation between the clusters for features exports, health, life_expec and imports
- child_mort rate is high in cluster 2 and low in cluster 4
- Income is high in cluster 0 and low in cluster 2
- Inflation is high for cluster 2&4 and low for cluster 0 & 1
- total_fer is high for cluster 2
- gdpp is high for cluster 0 and low for cluster 1 & 2

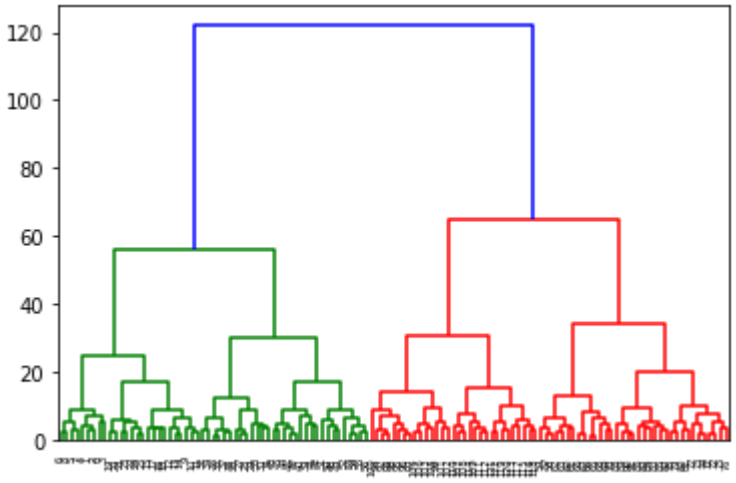
Thus we can say that countries within Cluster 2 need more aid than other countries

Hierarchical Clustering

Single Linkage

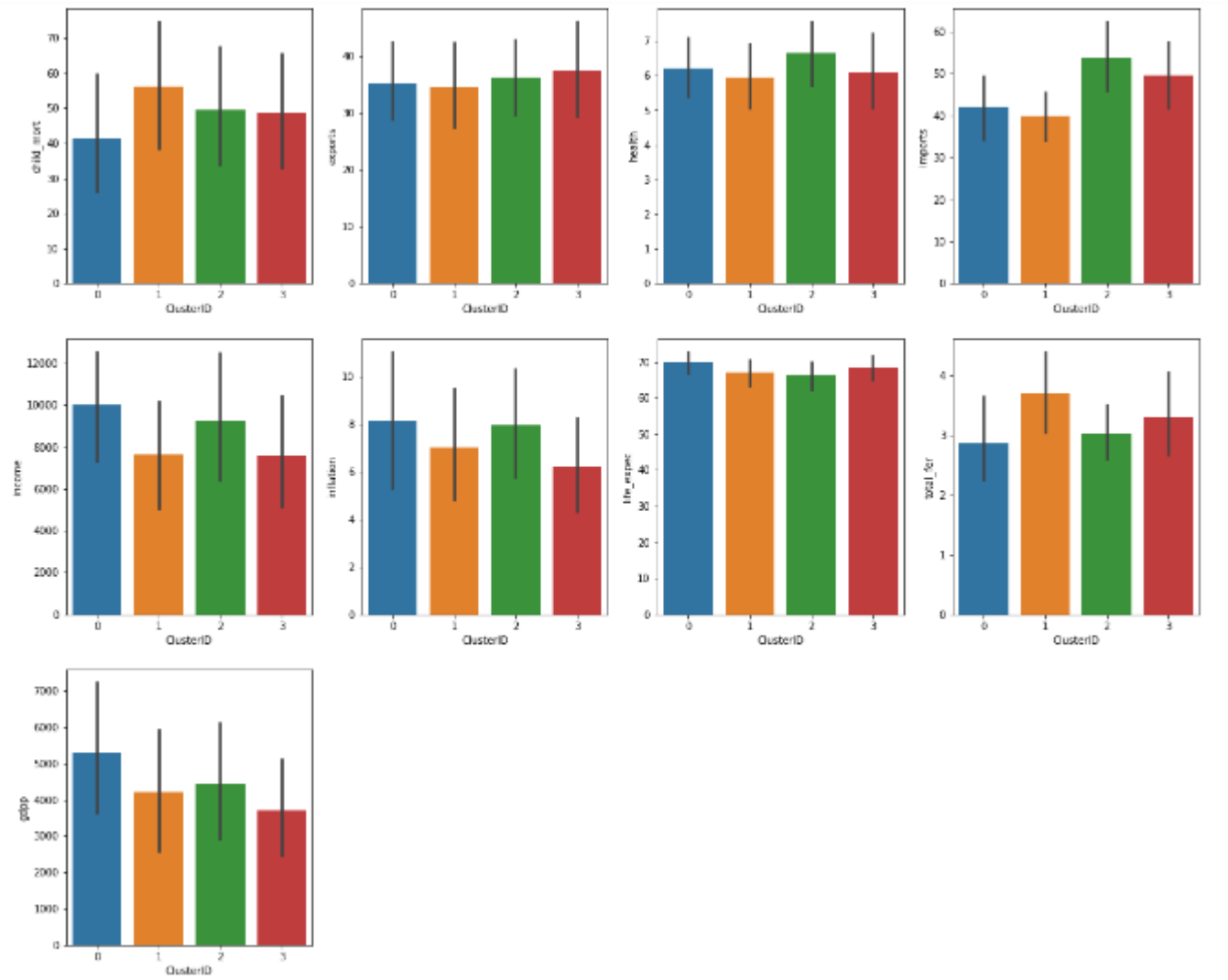


Complete Linkage



Single Linkage graph is not very understandable
Hence, we have considered complete with 4 as the number of clusters

Hierarchical Clustering Analysis



Hierarchical Clustering Observations

- There is no significant variation between the clusters for features exports, health, life_expec
- child_mort rate is high in cluster 1 and low in cluster 2
- imports high in cluster 2
- Income is high in cluster 2 and low in cluster 3
- Inflation is high for cluster 1 and low for cluster 3
- total_fer is high for cluster 1
- gdpp is high for cluster 2 and low for cluster 3

Hence countries which are in cluster 0 and cluster 3 need aid.

Conclusion

Below are the considerations taken into account

- Countries with Child Mortality higher than 50
- Countries with Inflation value as 8 and gdpp value below 3000 should be considered