Clustering of Countries

RAHUL ANAND

Abstract

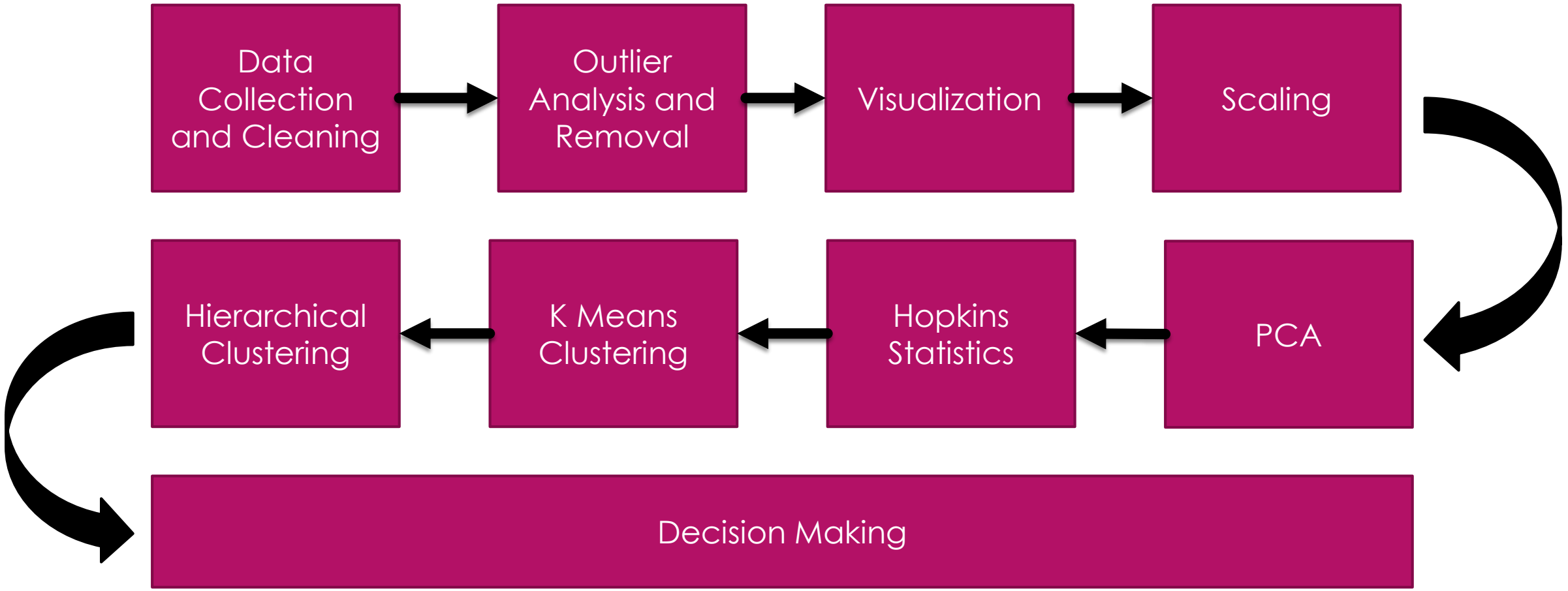
- Objective:

We, HELP International humanitarian NGO, committed to fight poverty and provide the people of backward countries with basic amenities and relief during the time of disasters and natural calamities. We run a lot of operational projects from time to time, along with advocacy, drives to raise awareness as well as for funding purposes.

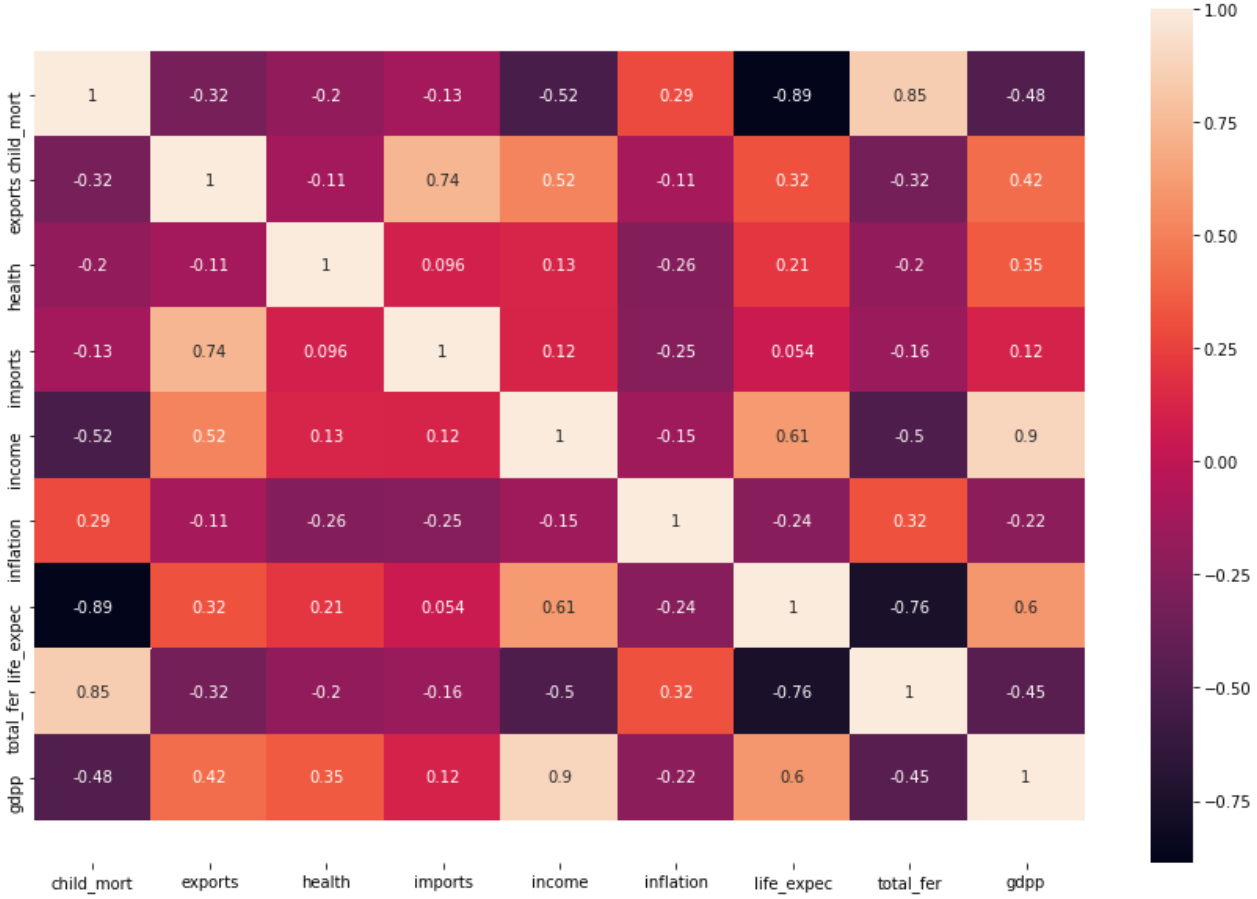
- Problem statement:

During the recent funding programmes, we have been able to raise around $ 10 million. As an analyst, we have to come up with the countries list that are in the direst need of aid.

Analysis Methodology



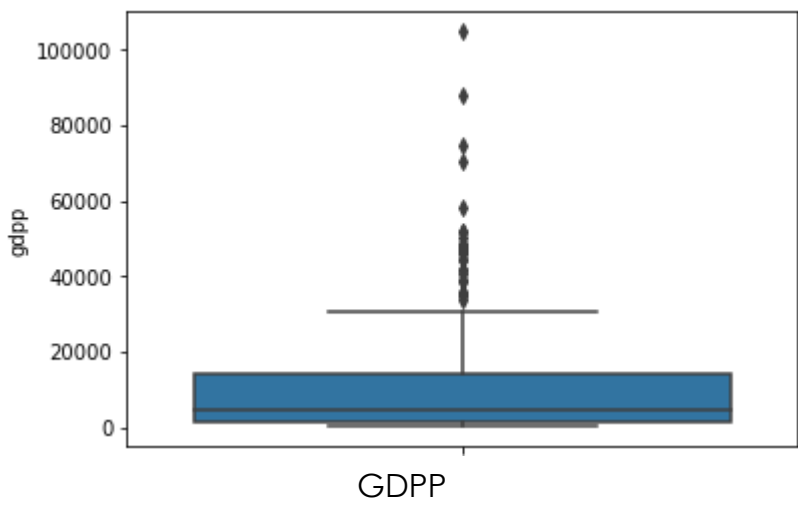
Correlation in the data



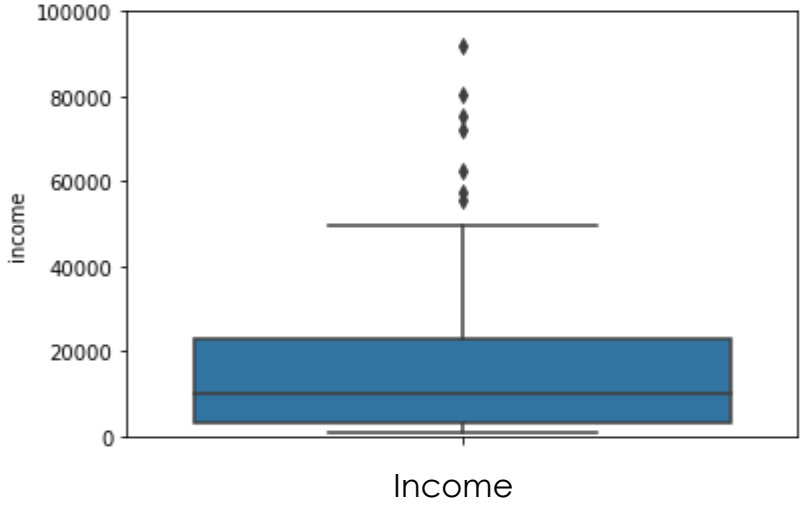
Inference:

- child_mortality and life_expentency are highly correlated with correlation of -0.89
- child_mortality and total_fertility are highly correlated with correlation of 0.85
- imports and exports are highly correlated with correlation of 0.74
- life_expentency and total_fertility are highly correlated with correlation of -0.7

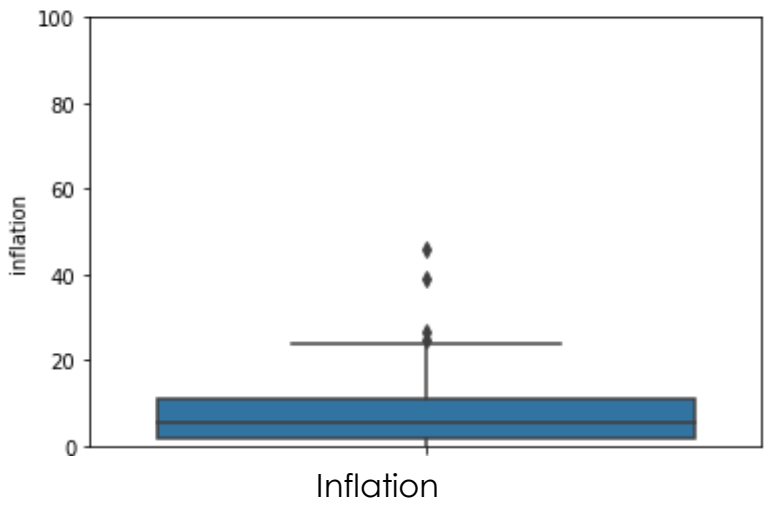
Outliers Treatment



GDPP



Income



Inflation

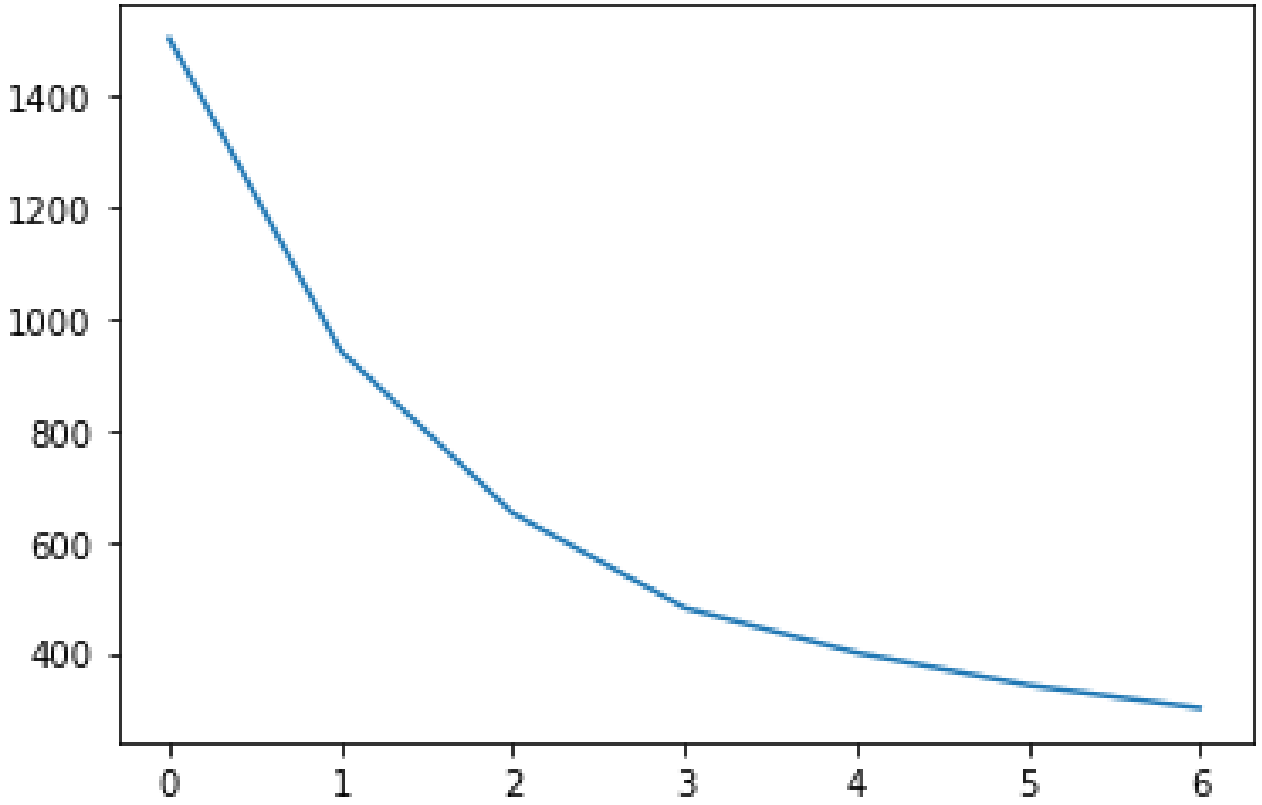
We see that gdpp, income and inflation column has high outliers. However we have not removed outliers as this might lead to loss in country details which are not doing well- socio-economically(countries with direst need of aid).

Hopkins Score

```
In [29]: hopkins(country_norm)
Out[29]: 0.9465330438245912
```

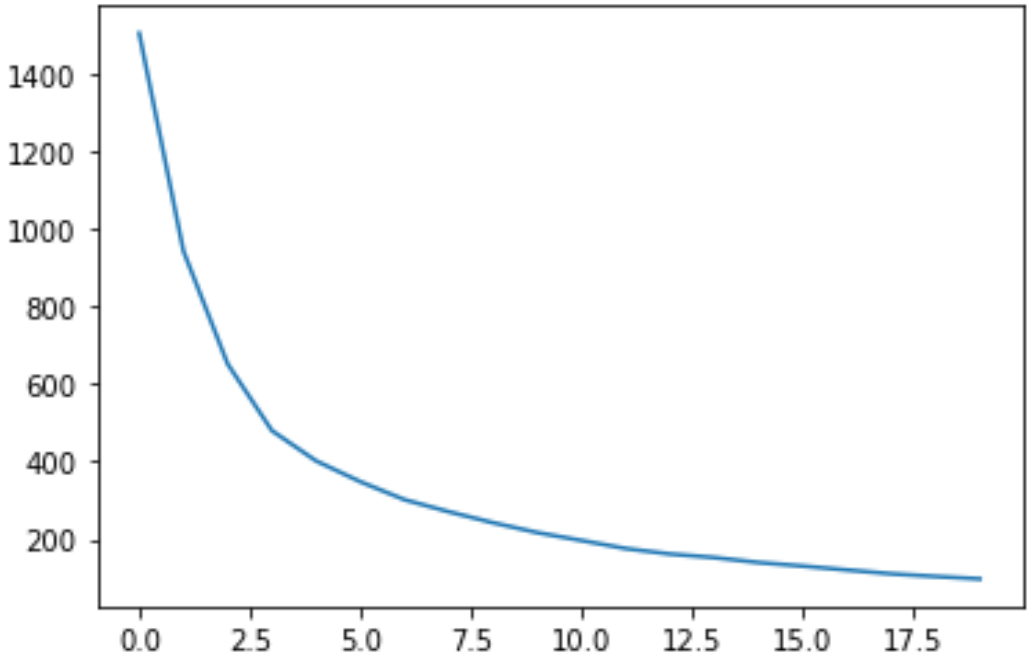
Inference:
0.95 is a good Hopkins score for Clustering.

Elbow Curve for Optimal cluster



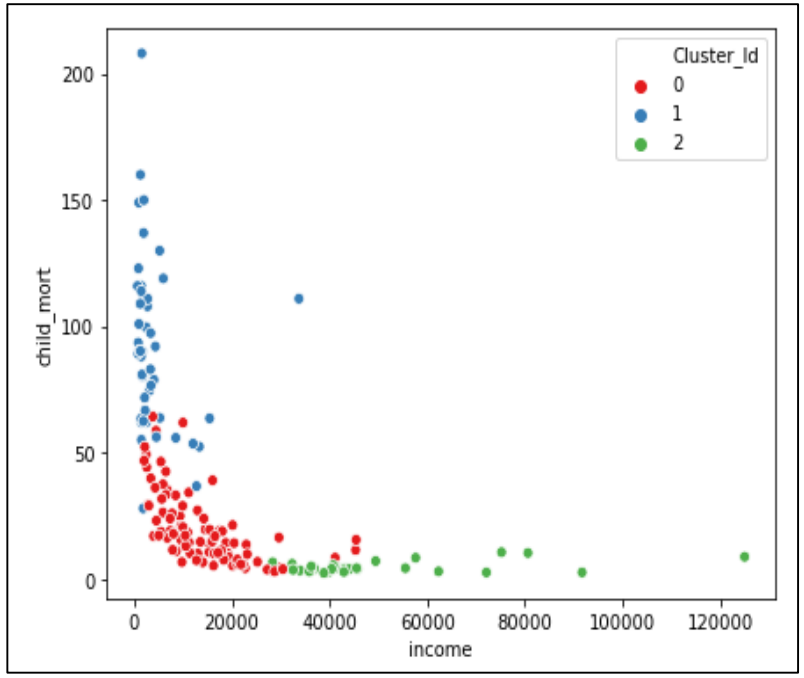
Inference:
Looking at the elbow curve it looks good to proceed with either 3 or 4 clusters.

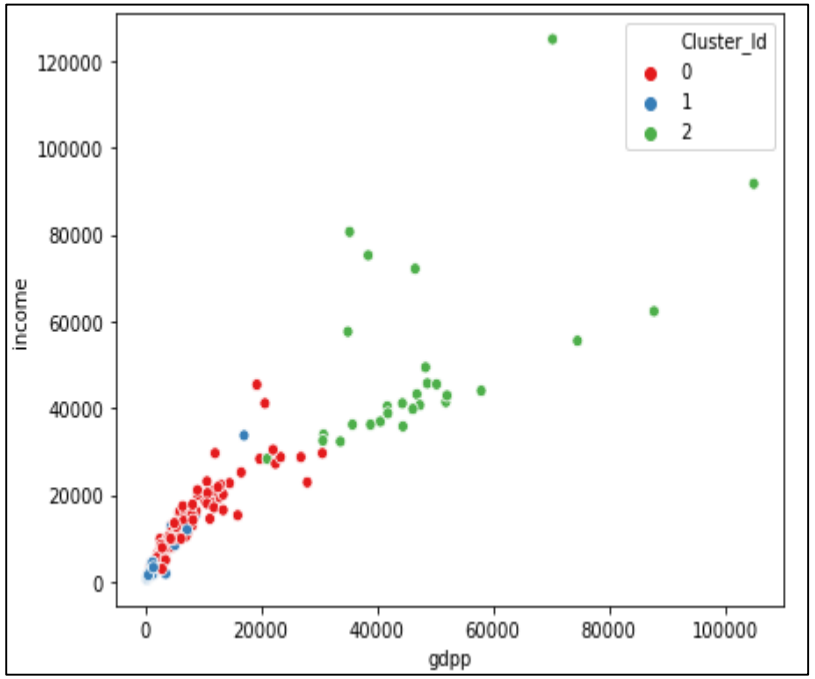
Silhouette Analysis

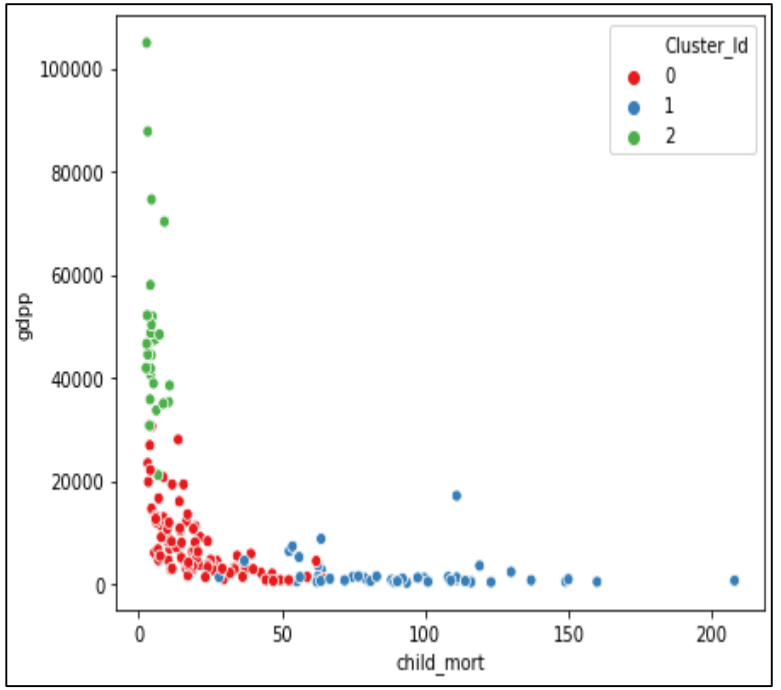


For n_clusters=2, the silhouette score is 0.45863306035476264
For n_clusters=3, the silhouette score is 0.4218615812599681
For n_clusters=4, the silhouette score is 0.42673357397704514
For n_clusters=5, the silhouette score is 0.4324001169216119
For n_clusters=6, the silhouette score is 0.39279369617575527
For n_clusters=7, the silhouette score is 0.3068220382518731
For n_clusters=8, the silhouette score is 0.26474839748627066

K Means Clustering

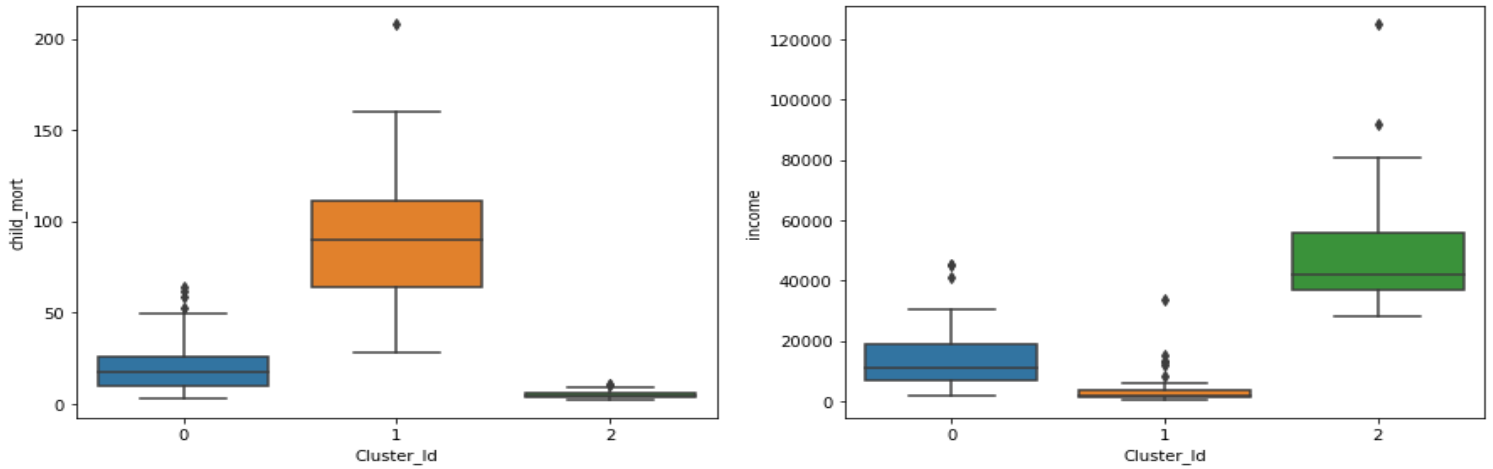


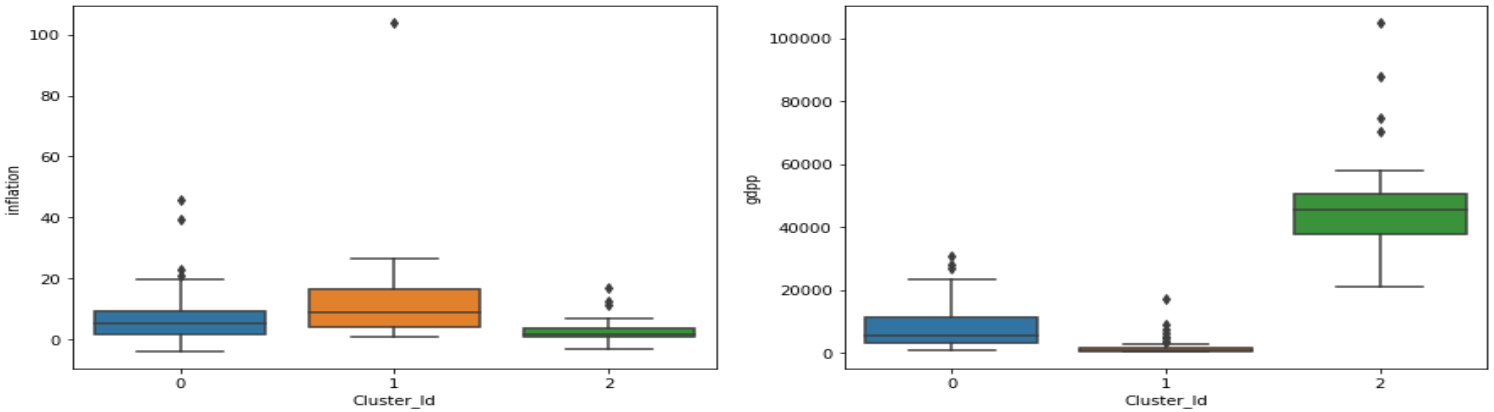




Scatter plot on Original attributes to visualize the spread of the data

K Means Clustering

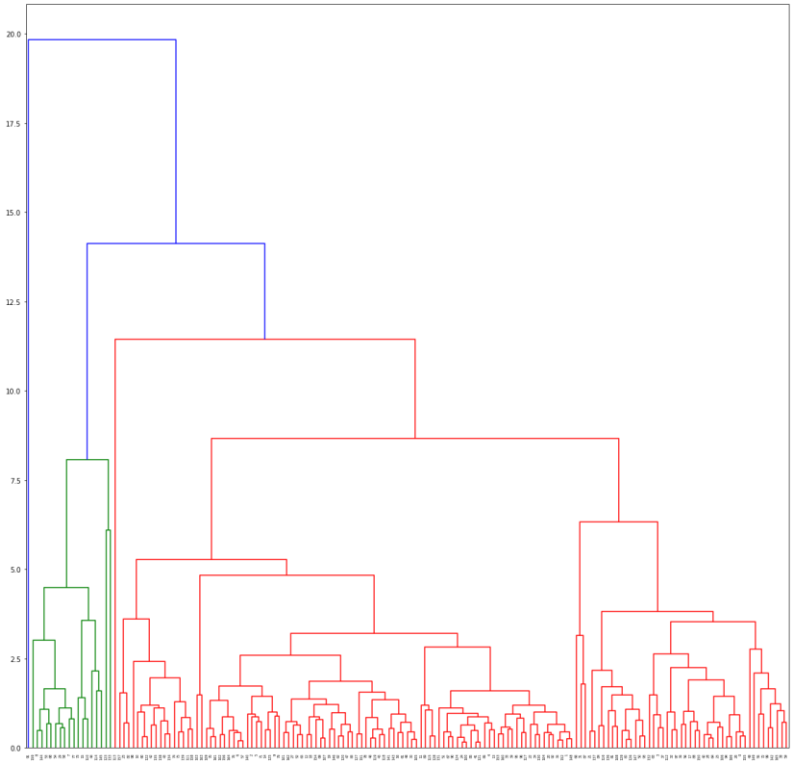




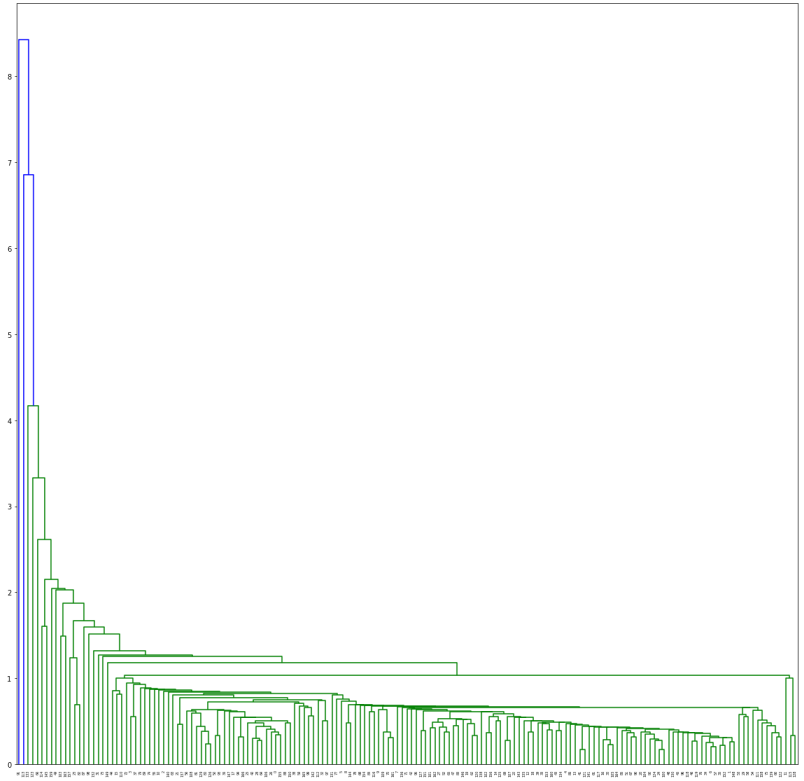
Inference:

- Child Mortality is highest for Cluster 1.This cluster need some aid.
- Income and Gdpp are measures of development. Higher the per capita income and gdpp better is the country's development.
- Income per capita and gdpp seems lowest for countries in cluster 1. Hence, these countries need some help.

Hierarchical Clustering

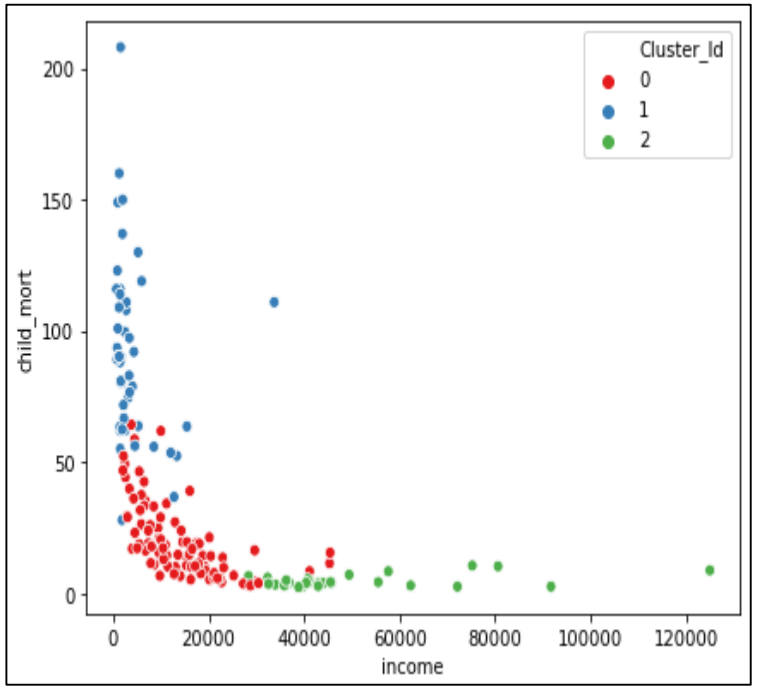


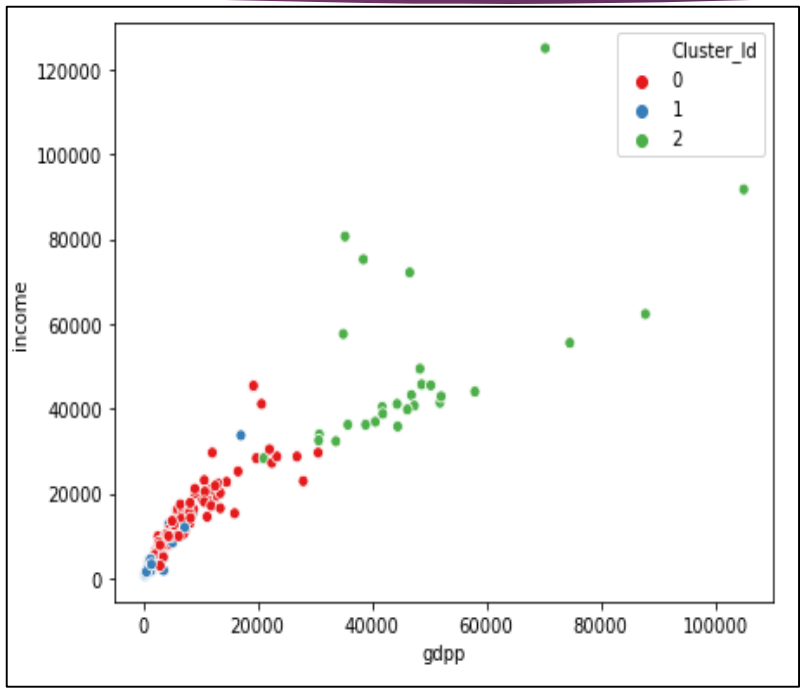
Complete method hierarchical clustering

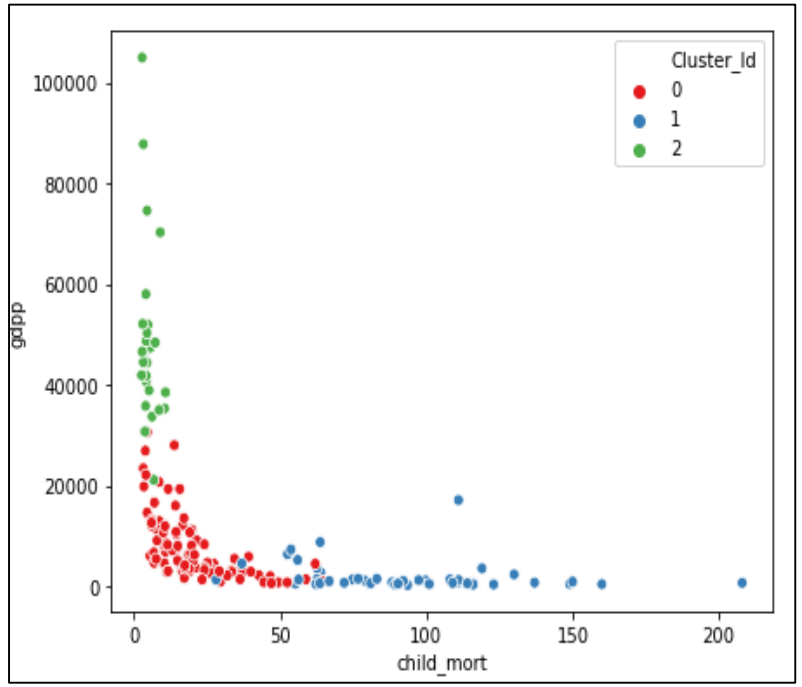


Single method hierarchical clustering

Hierarchical Clustering







Scatter plot on Original attributes to visualize the spread of the data

Summary

- We have analyzed both K-means and Hierarchical clustering and found clusters formed are identical.
- The clusters formed in both the cases are not that great.
- we will proceed with the clusters formed by K-means and based on the information provided by the final clusters we will deduce the final list of countries which are in need of aid.
- Based on clusters we have identified the list of countries which are in dire need of aid.
- The list of countries are subject to change as it is based on the few factors like `Number of Clusters chosen`, `Clustering method used` etc. Which we have used to build the model.

```
0                  Burkina Faso
1                       Burundi
2     Central African Republic
3               Congo, Dem. Rep.
4                        Guinea
5                 Guinea-Bissau
6                         Haiti
7                    Mozambique
8                         Niger
9                  Sierra Leone
Name: country, dtype: object
```

List of countries which are in dire need of aid.

Thank You