Clustering & PCA
Country Data

SOURYA PRAKASH PARIDA

Introduction

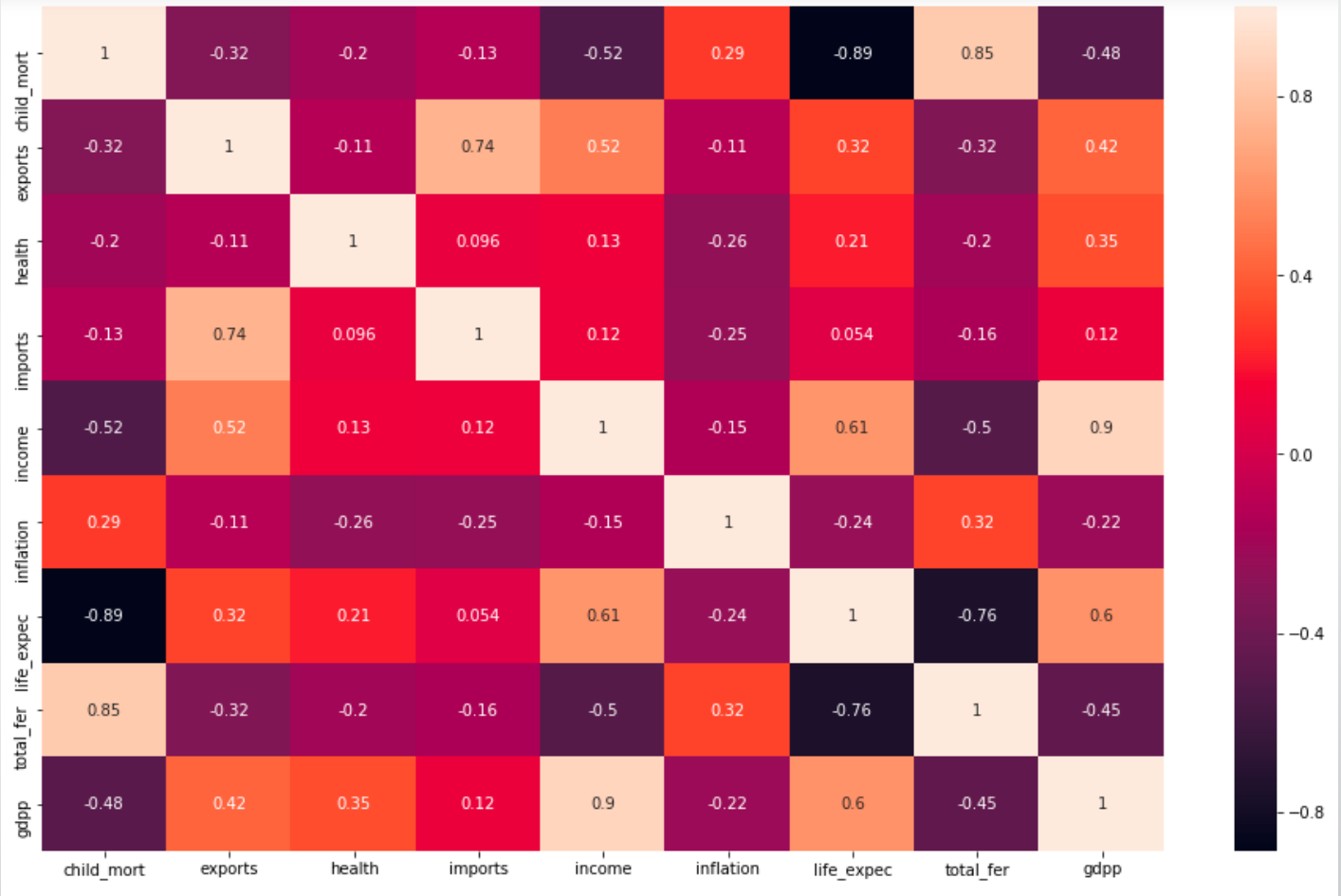
Problem Statement: To categorize the countries using some socio-economic and health factors that determine the overall development of the country, and to suggest the countries that are in the direst need of aid so that the NGO can decide how to use this money strategically and effectively.

Approach: We have used Principal Component Analysis and different clustering methods to group the countries based on the country data available for this analysis.

Assumptions/Exceptions: The data provided to us has outliers which have been treated by removing the countries with outliers (after performing PCA). This has resulted in removal of ~10% of the countries for the analysis. (*Note to reviewer*: This is just one of the many available approaches selected for this assignment to treat the outliers).

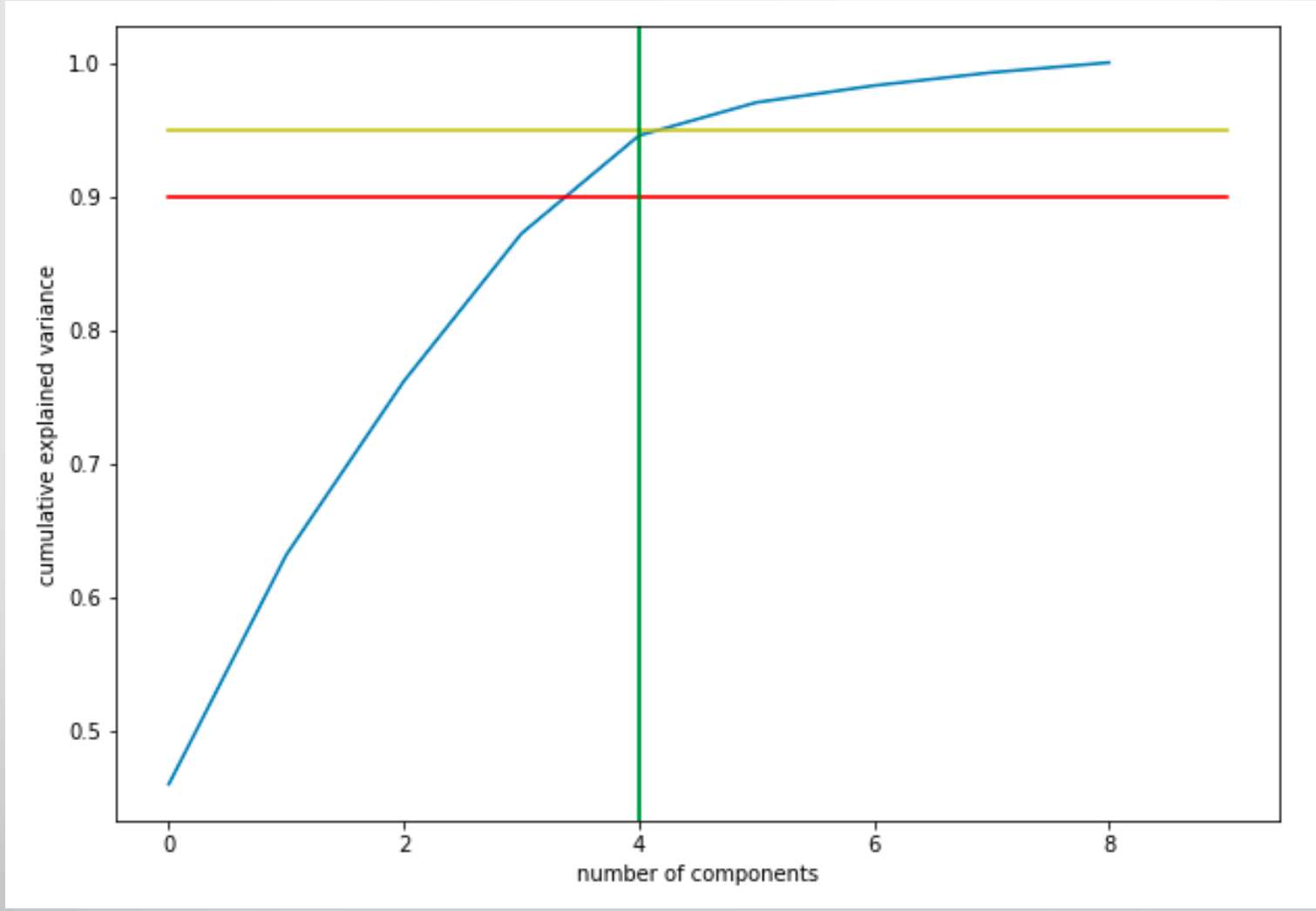
Correlation Matrix - Heatmap

- Income and GDPP are highly correlated: +0.90
- Child Mortality and Total Fertility are highly correlated: +0.85
- Child Mortality and Life Expectancy are highly correlated: -0.89
- Life Expectancy and Total Fertility are highly correlated: -0.76



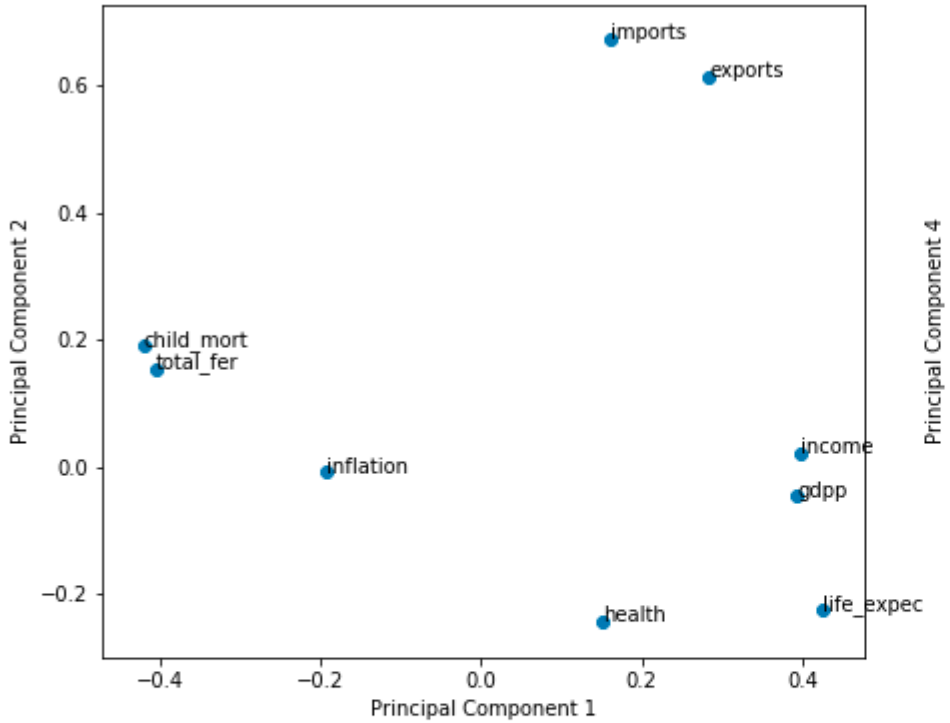
High and linear correlation indicates that PCA can be applied on this dataset

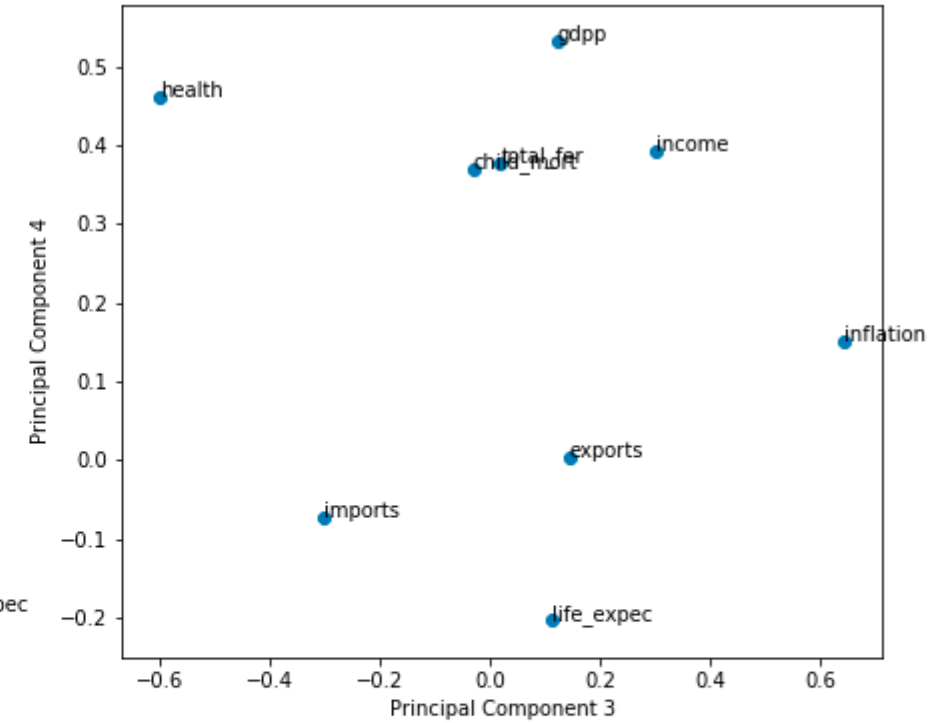
Principal Component Analysis



Observation: When no. of PCs = 4, the cumulative explained variance is ~95%, indicating that the first 4 PCs are sufficient for clustering.

Dimension distribution across Principal Components





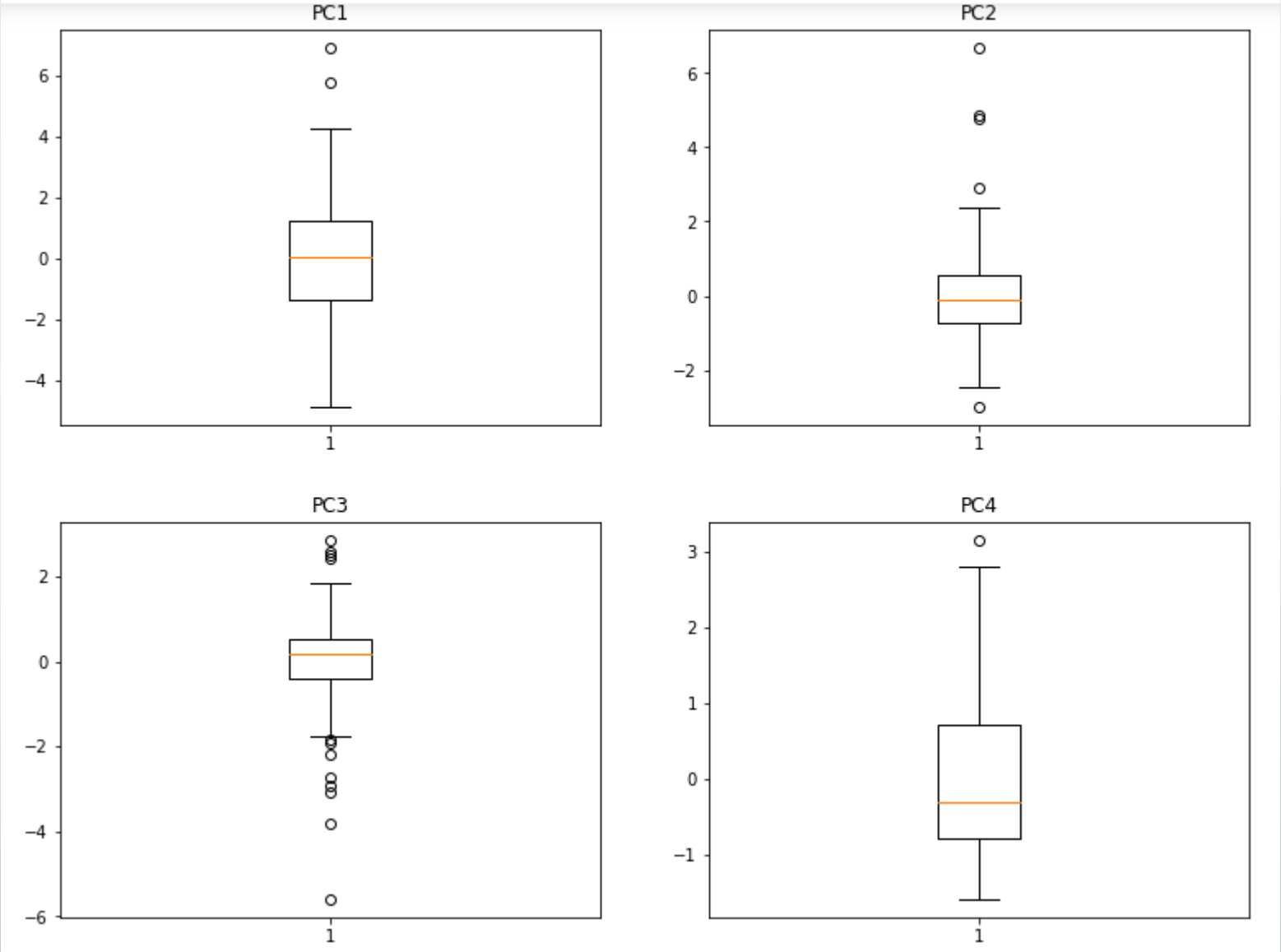
Observations:

- Income, GDPP and Life Expectancy are loaded into PC1
- Income, GDPP, Health and Child Mortality are loaded into PC4
- Imports and Exports are loaded into PC2
- Inflation is loaded into PC3

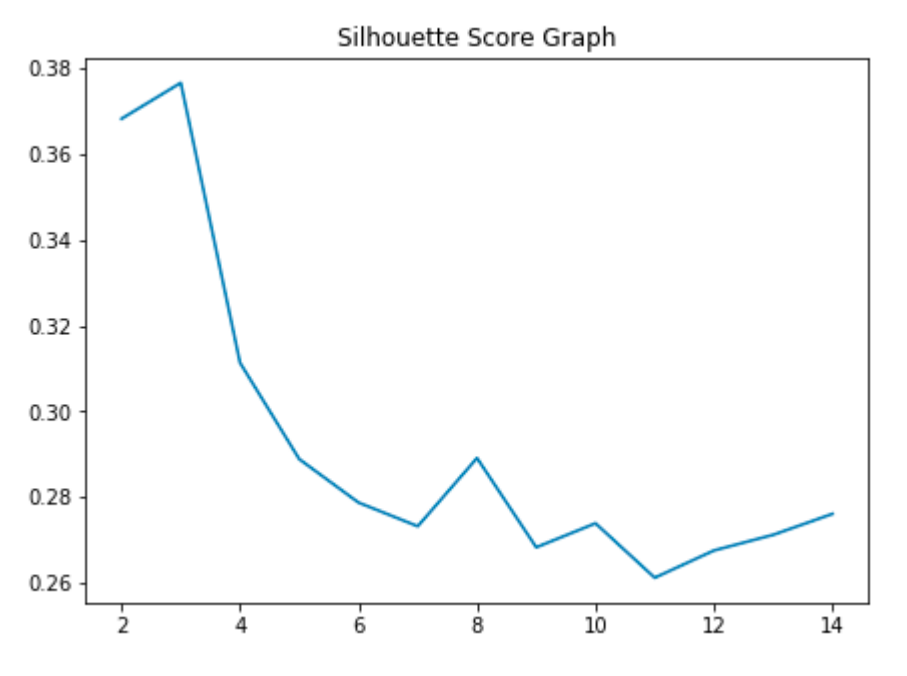
Principal Components - Outliers

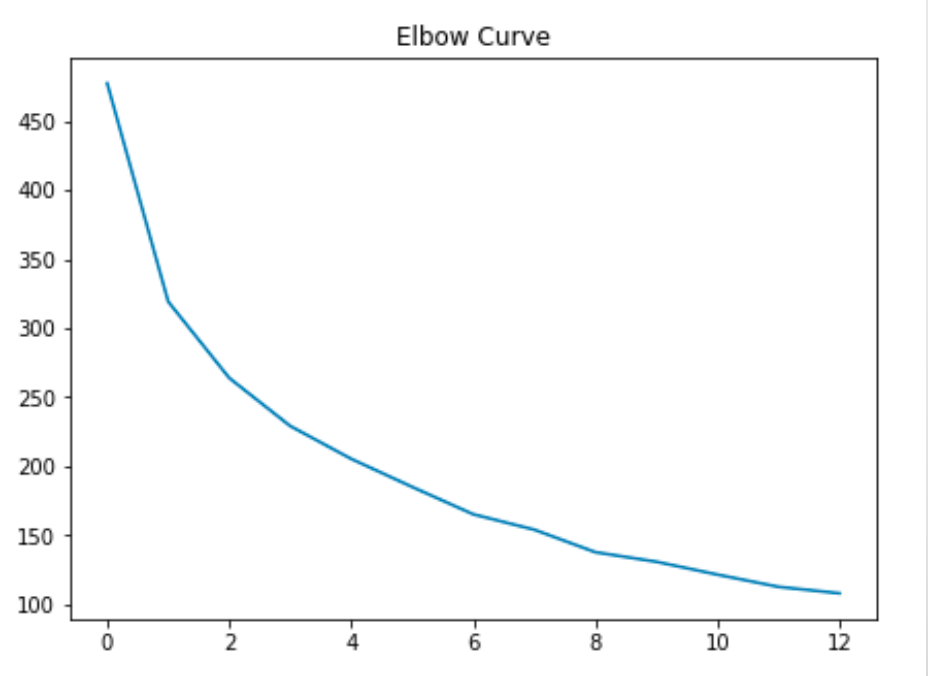
The Principal Components had outliers which were treated using IQR method, resulting in removal of 18 countries' rows and thus reducing the no. of rows from 167 to 149.

The Hopkins score is **0.74** which is a good indication that the dataset has a good tendency for clustering.



K-Means Clustering Method

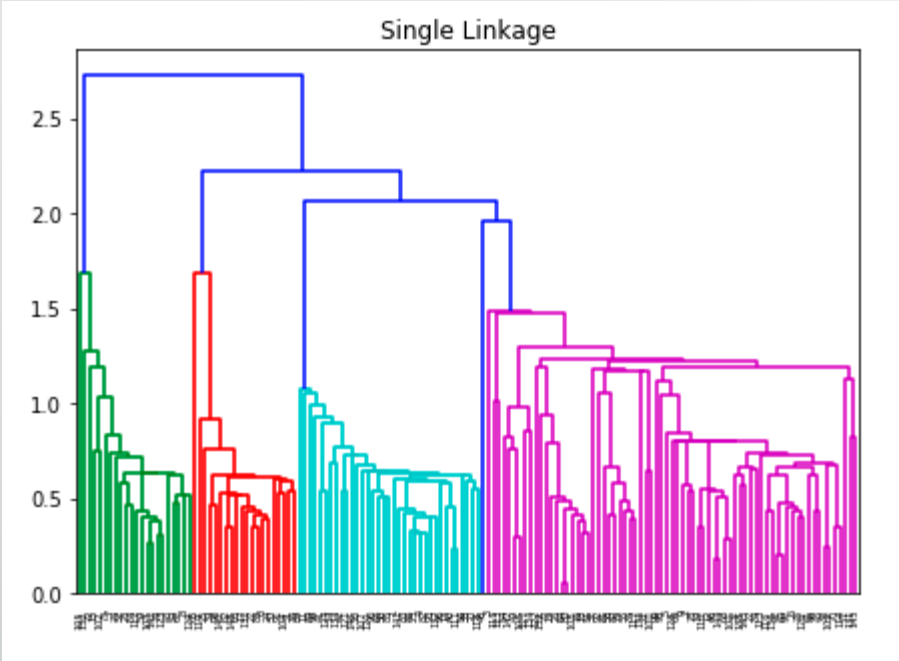


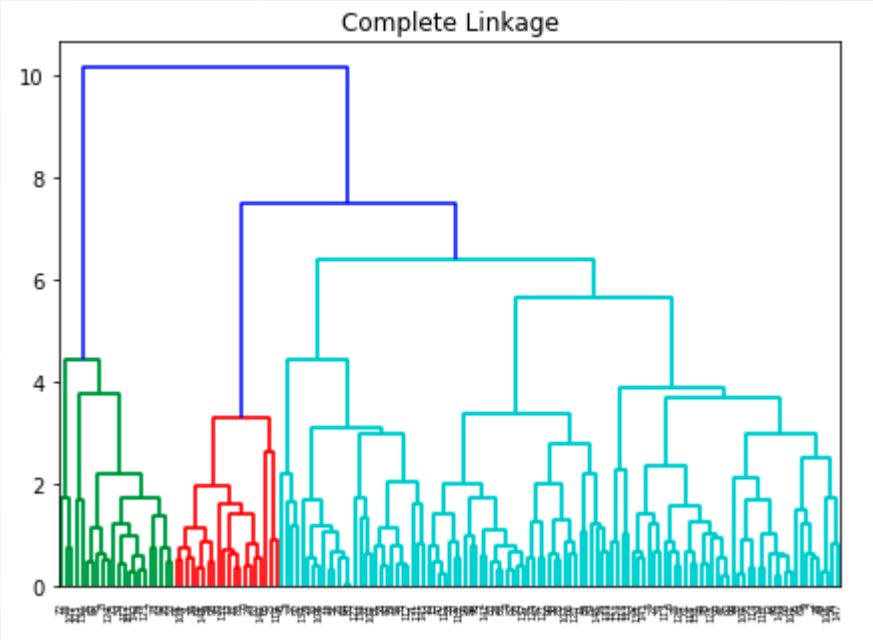


From the above graph, the optimum no. of clusters was chosen to be 4, i.e. k =4.

The k-means method with k=4 was used to cluster the countries.

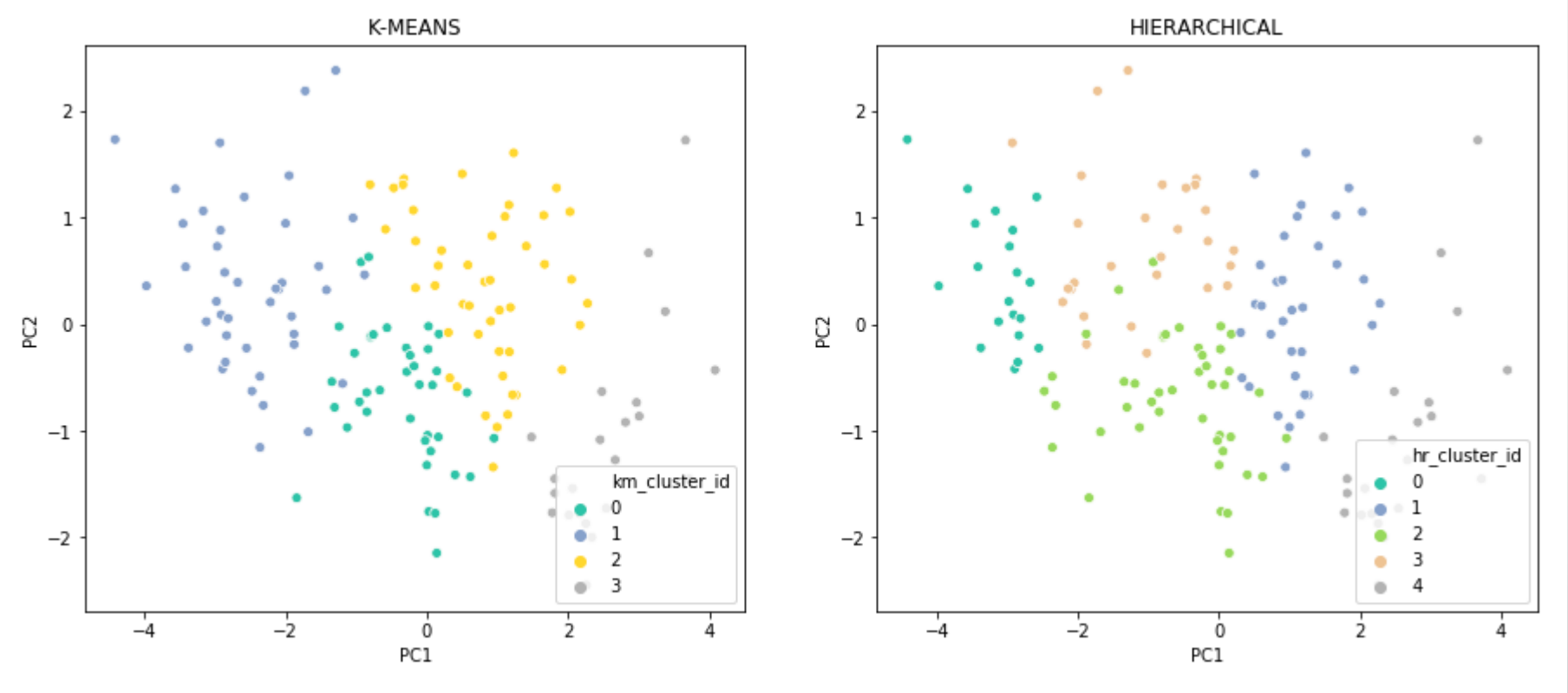
Hierarchical Method



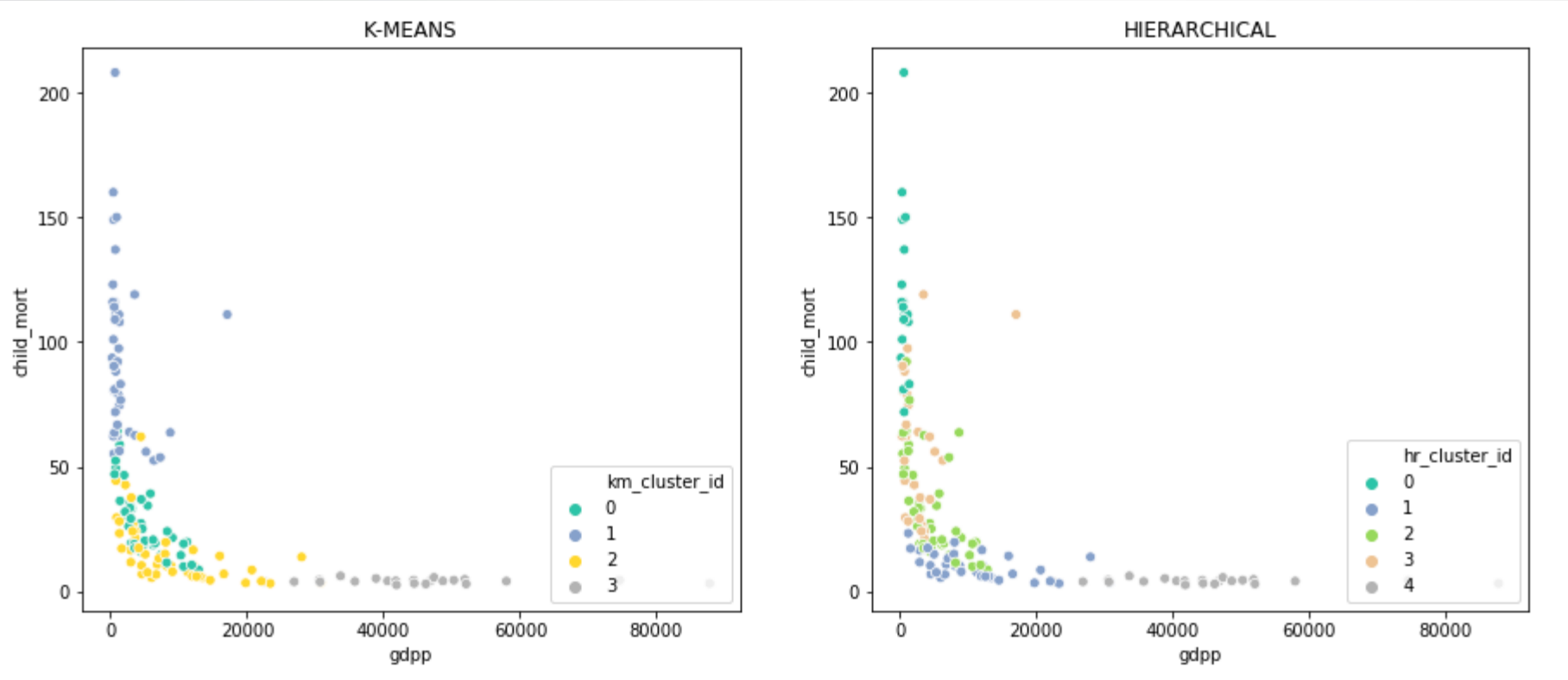


- In the above dendrograms, the complete-linkage hierarchical clustering provides a much better picture of clustering.
- Hence, we can divide the countries at level between 4-6 (i.e. into 5 clusters) for our analysis

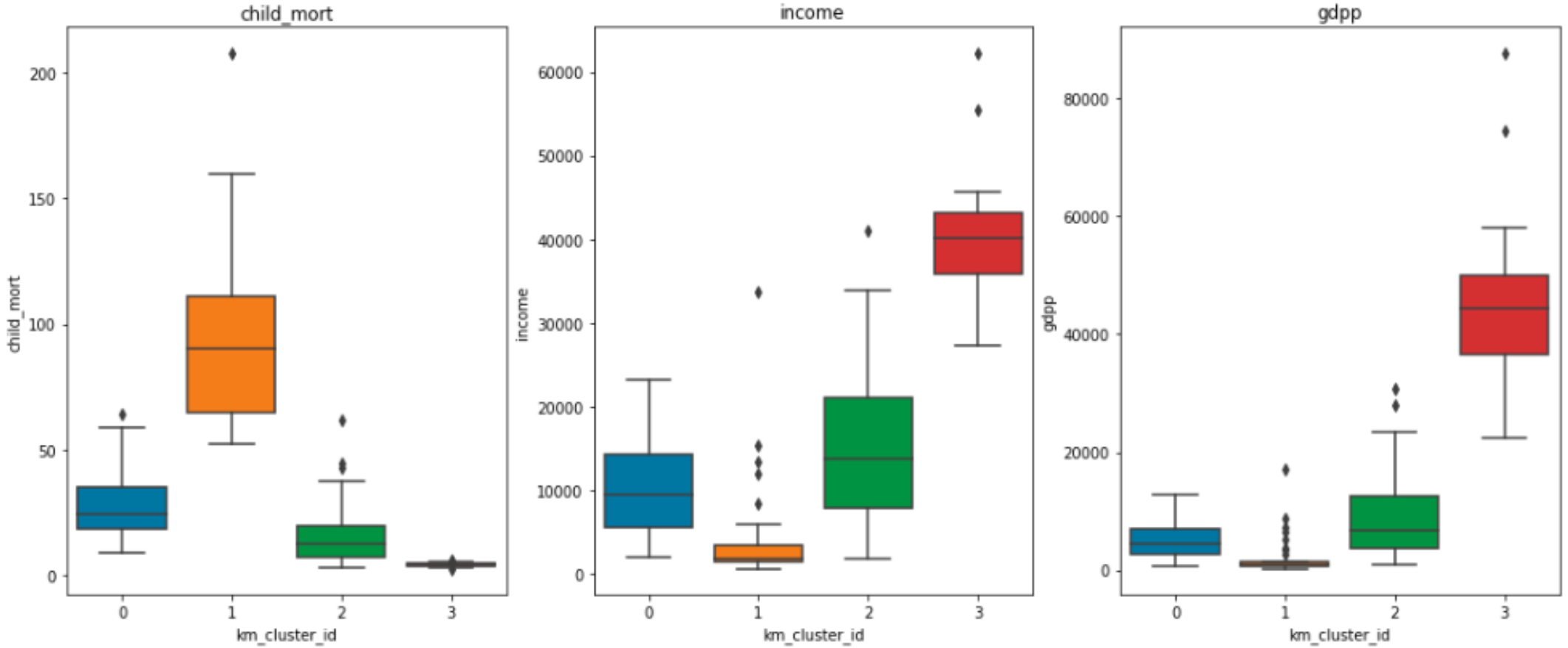
Comparison between k-means and Hierarchical Clusters



Comparison between k-means and Hierarchical Clusters

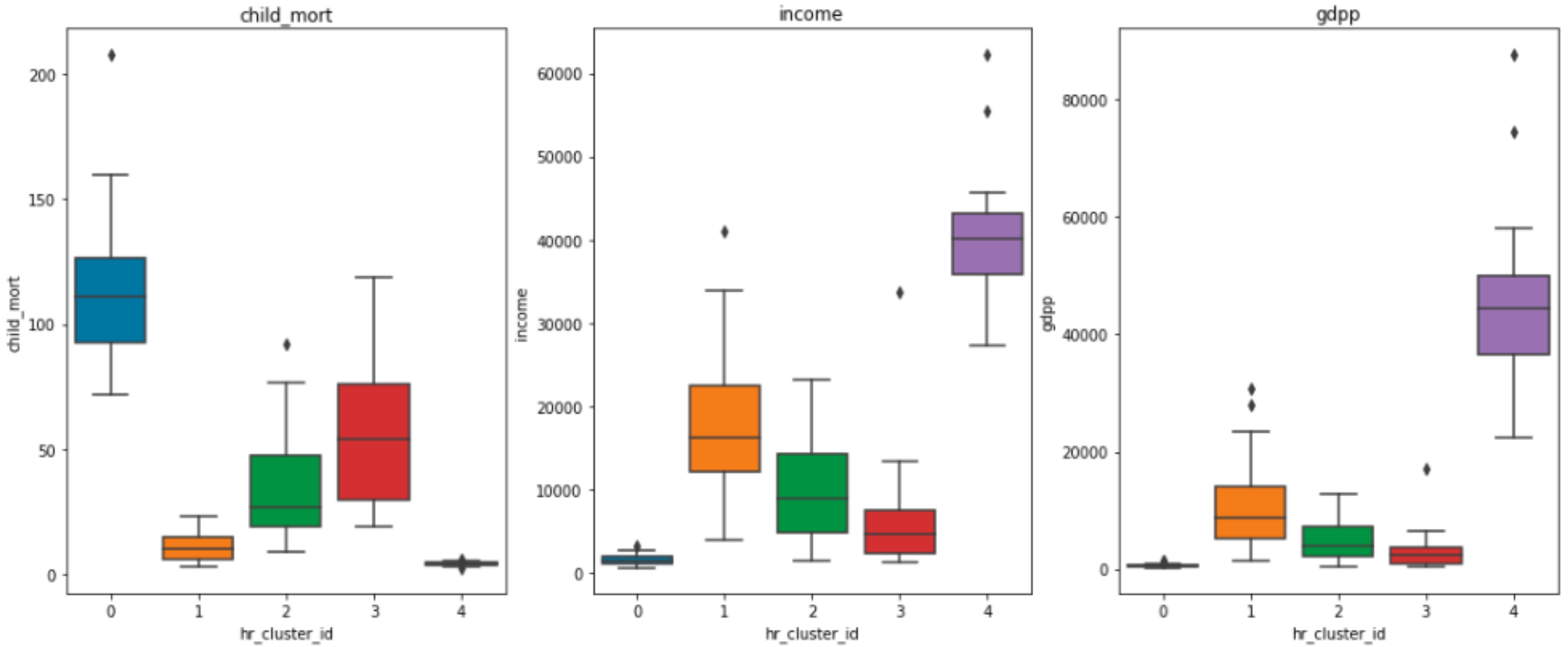


GDDP, Income and Child Mortality: k-means Clusters



Observation (k-means): Cluster 1 has the highest child mortality with low income/gdpp

GDDP, Income and Child Mortality: Hierarchical Clusters



Observation (Hierarchical): Cluster 0 has the highest child mortality with low income/gdpp

Final Results and Observations

In k-means, the clusters have lot of outliers and overlapping values whereas the box-plot for hierarchical clustered countries have less overlaps. Therefore, the hierarchical clustering result is the preferred one.

The following countries fall in the cluster of countries with Lowest GDPP/Income and/or Highest Child Mortality Rate and therefore, in dire need of aid.

- 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