Clustering of countries

Grouping of countries based on socio-economic and health factors

Problem statement:

- The aim of the analysis is to help an international humanitarian NGO,HELP,to strategically plan and use the funds raised only for the under-developed nations.This requires clustering of countries on socio-economic and health factors and come up with the final list of poverty struck countries that are in dire need of aid.

Analysis Approach:

- Exploratory Data analysis (EDA).
- Principal component Analysis(PCA).
- Identifying cluster tendency using Hopkins statistic.
- Identifying optimal number of clusters(k) based on
    - Elbow curve method.(k=3)
    - Silhouette score.(k=5)
    - Business needs.(k=3)
- Clustering and choosing the better output from
    - K-means.
    - Hierarchical.
- Identifying final list of countries.

Exploratory Data Analysis:

- Redundant features – No redundant features have been identified.
- Missing values – No missing values are found.
- Outlier Analysis:

  There is a consistent increase of values till 99th percentile across all the columns but a huge difference between the 99th percentile value and the maximum value in all the columns expect life_expec(Life expectancy),tot_fer(total fertility rate).
- Outlier Treatment:

  Per the business case,since we are looking for under-developed countries,only the outliers(maximum values) in the columns income and gdpp have been dropped and rest all are considered for analysis.
- Identifying Correlation:

  Highly correlated features are Income and GDPP,Imports and exports,Total Fertility rate and Child Mortality.

  Since PCA handles collinear variables,no column is manually dropped.

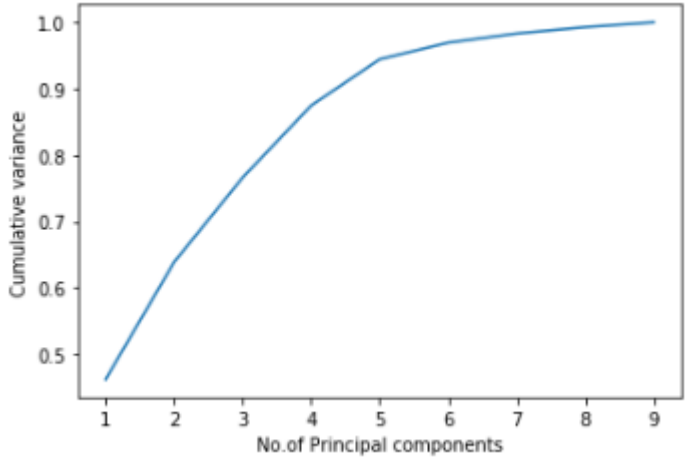
Principal component analysis(PCA):

- Feature scaling:

As the important features like income,gdpp,inflation etc all are on different scales,data is standardized for PCA.

- Scree plot:

As around 90% of variance in data is explained by 5 principal components,5 principal components have been considered for Data modeling.



Hopkins statistic:

- Hopkins statistic is a measure of randomness in the data and tendency of dataset to form clusters.
- Value
  - close to 1 implies data is highly clustered.
  - close to 0.5 implies data is random.
  - close to 0 implies data is uniform.

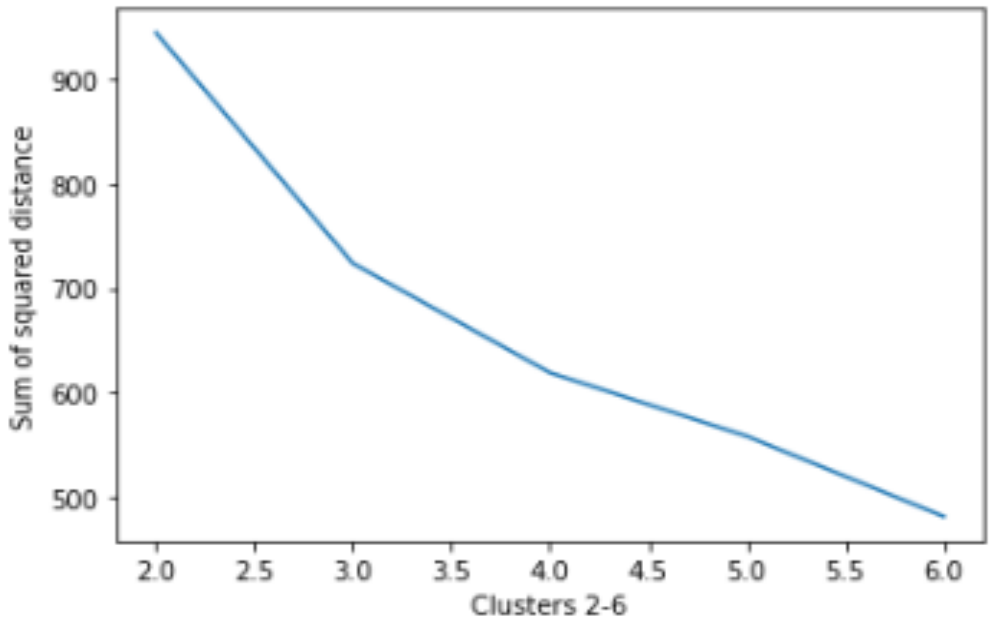
- Hopkin statistic for the given dataset: Highly clustered data.

0.88

Optimal number of clusters:

- Elbow curve Method:

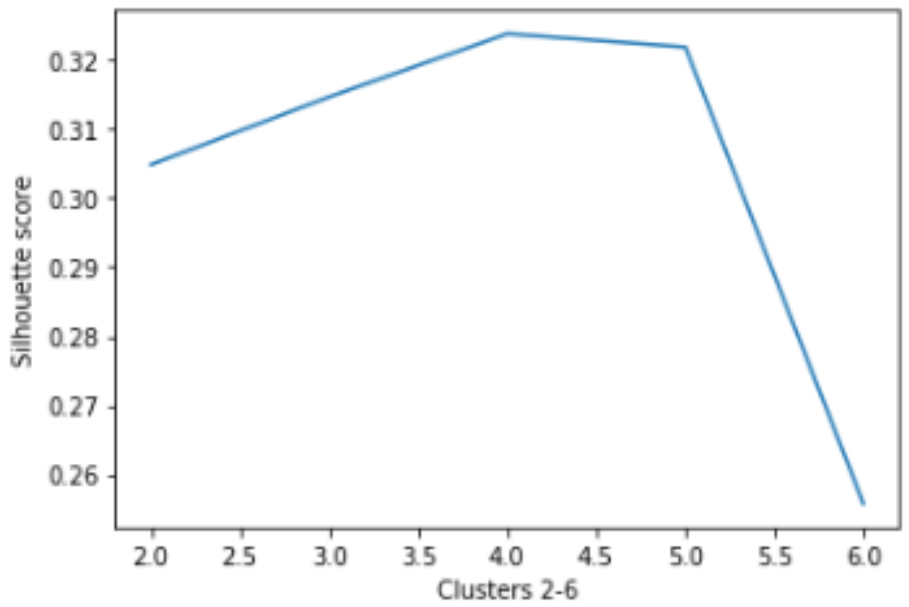
Recommended number of clusters :3



Optimal number of clusters(cont..)

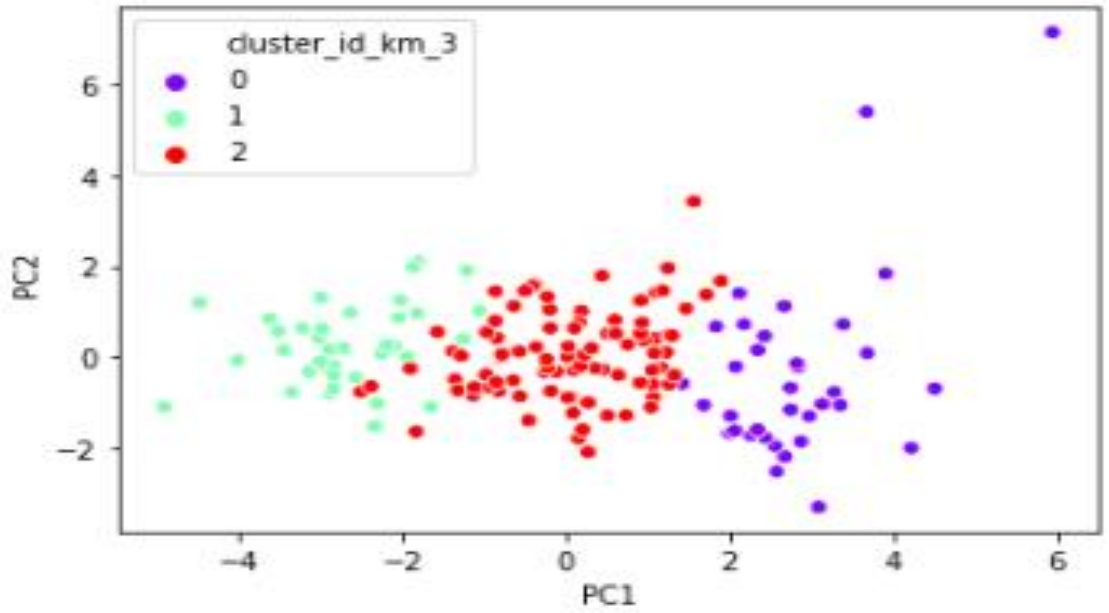
- Silhouette score:

From the global maxima,recommended number of clusters : 5



Clustering using K-means:

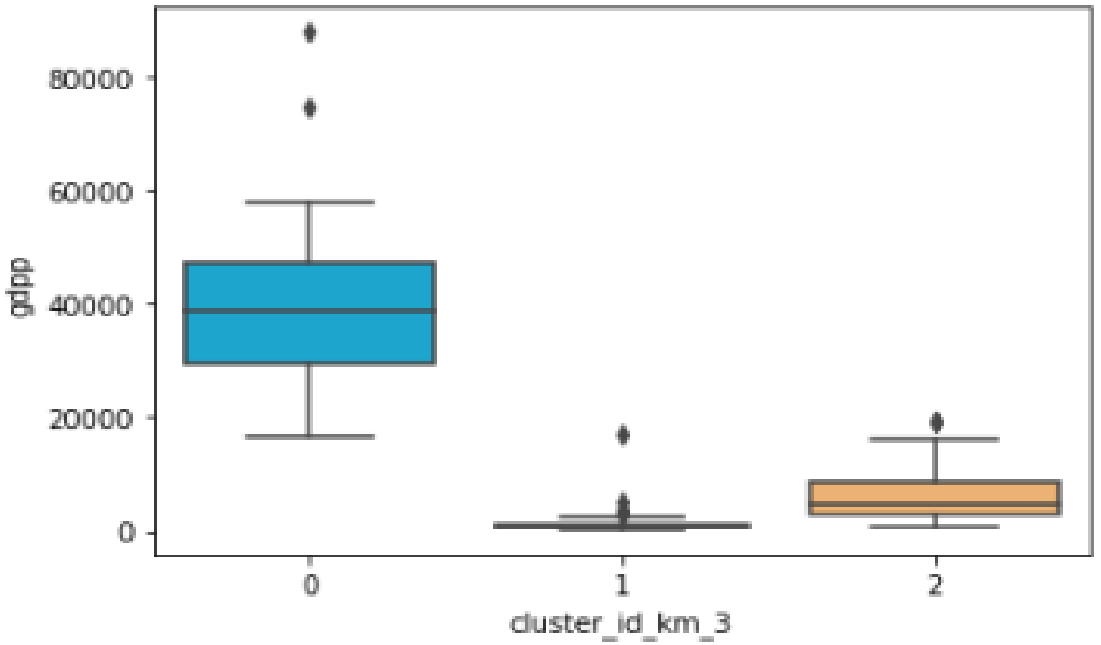
Clustering using k = 3 clusters:



Cluster Analysis:

- Cluster analysis using original variables – GDPP

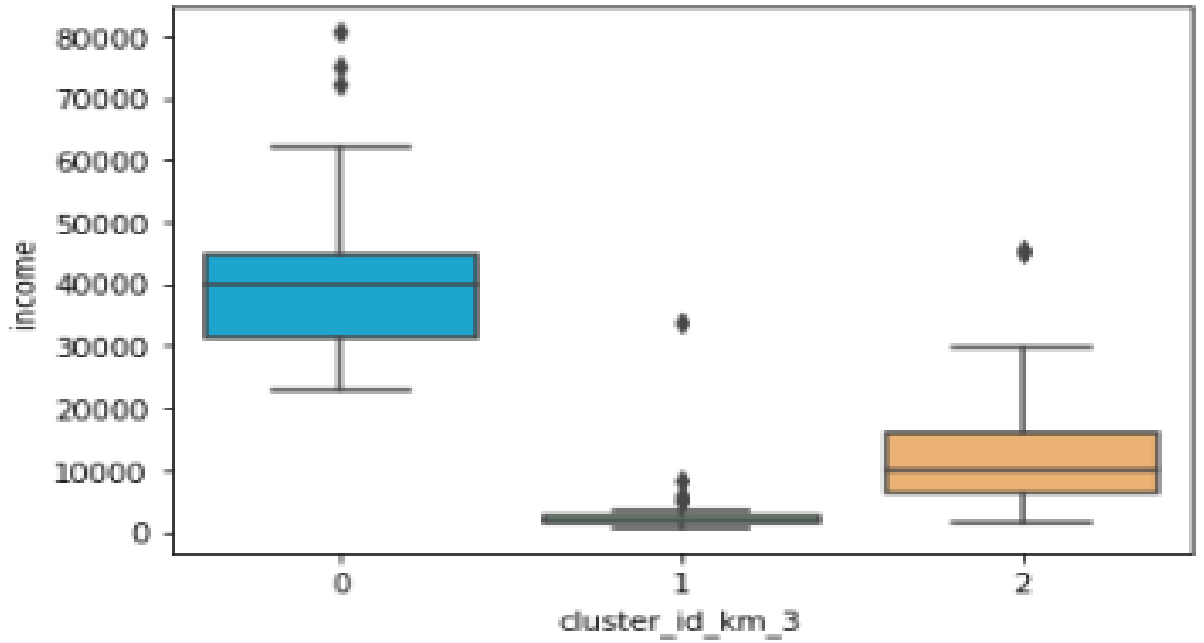
  Observation: Cluster 1 has recorded a low GDPP.



Cluster Analysis(cont..):

- Cluster analysis using original variables – Income

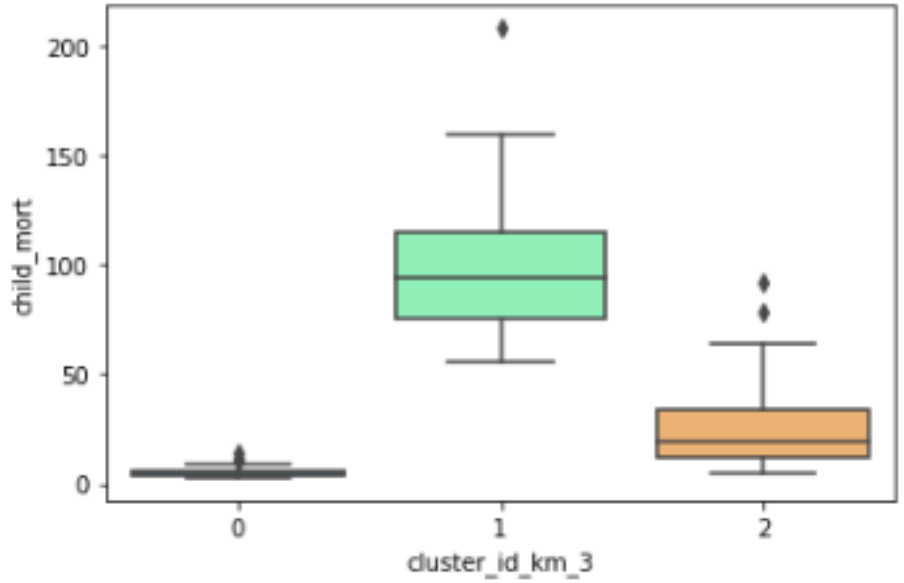
  Observation: Cluster 1 has recorded a low Income.



Cluster Analysis(cont..):

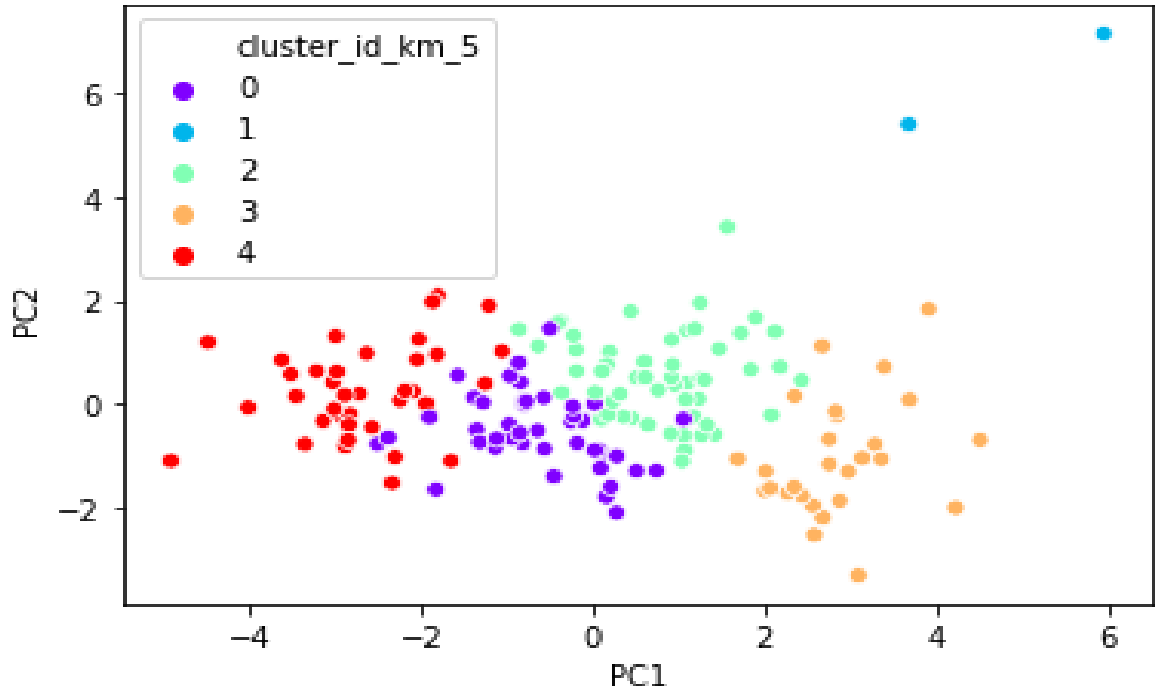
- Cluster analysis using original variables – Child Mortality

  Observation: Cluster 1 has recorded a high Child mortality.



Clustering using K-means:

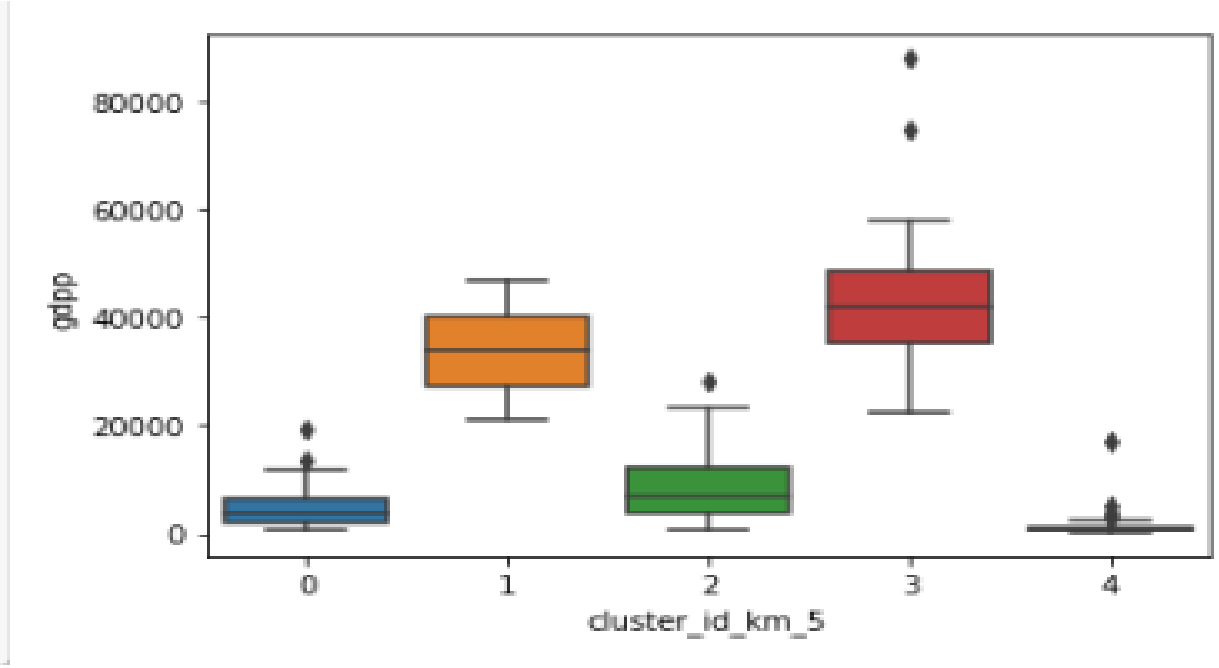
Clustering using k = 5 clusters:



Cluster Analysis:

- Cluster analysis using original variables – GDPP

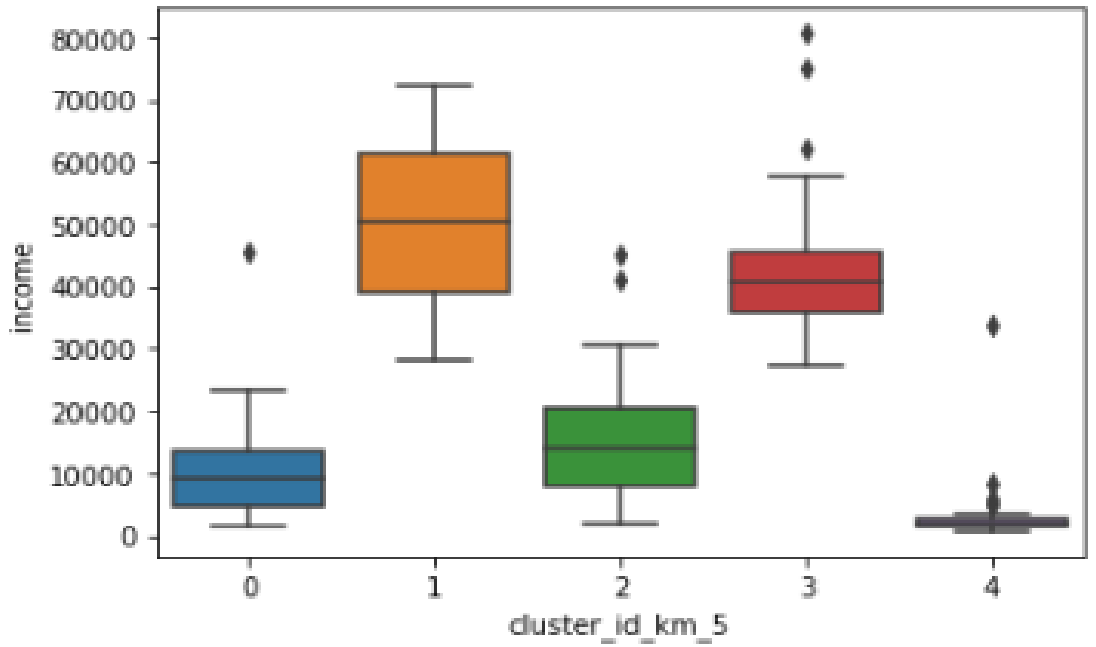
  Observation: Cluster 4 has recorded a low GDPP.



Cluster Analysis(cont..):

- Cluster analysis using original variables – Income

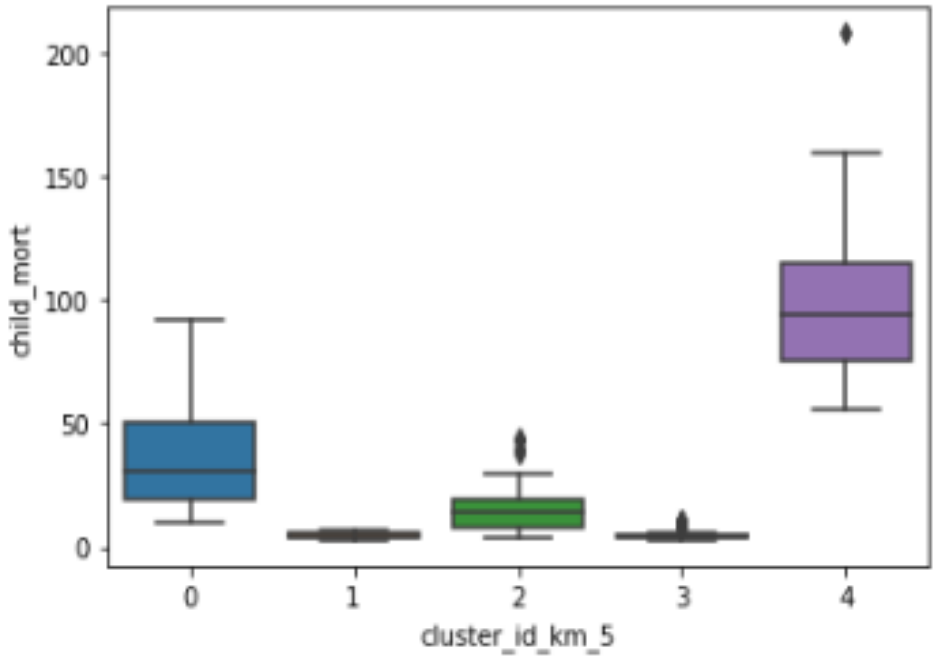
  Observation: Cluster 4 has recorded a low Income.



Cluster Analysis(cont..):

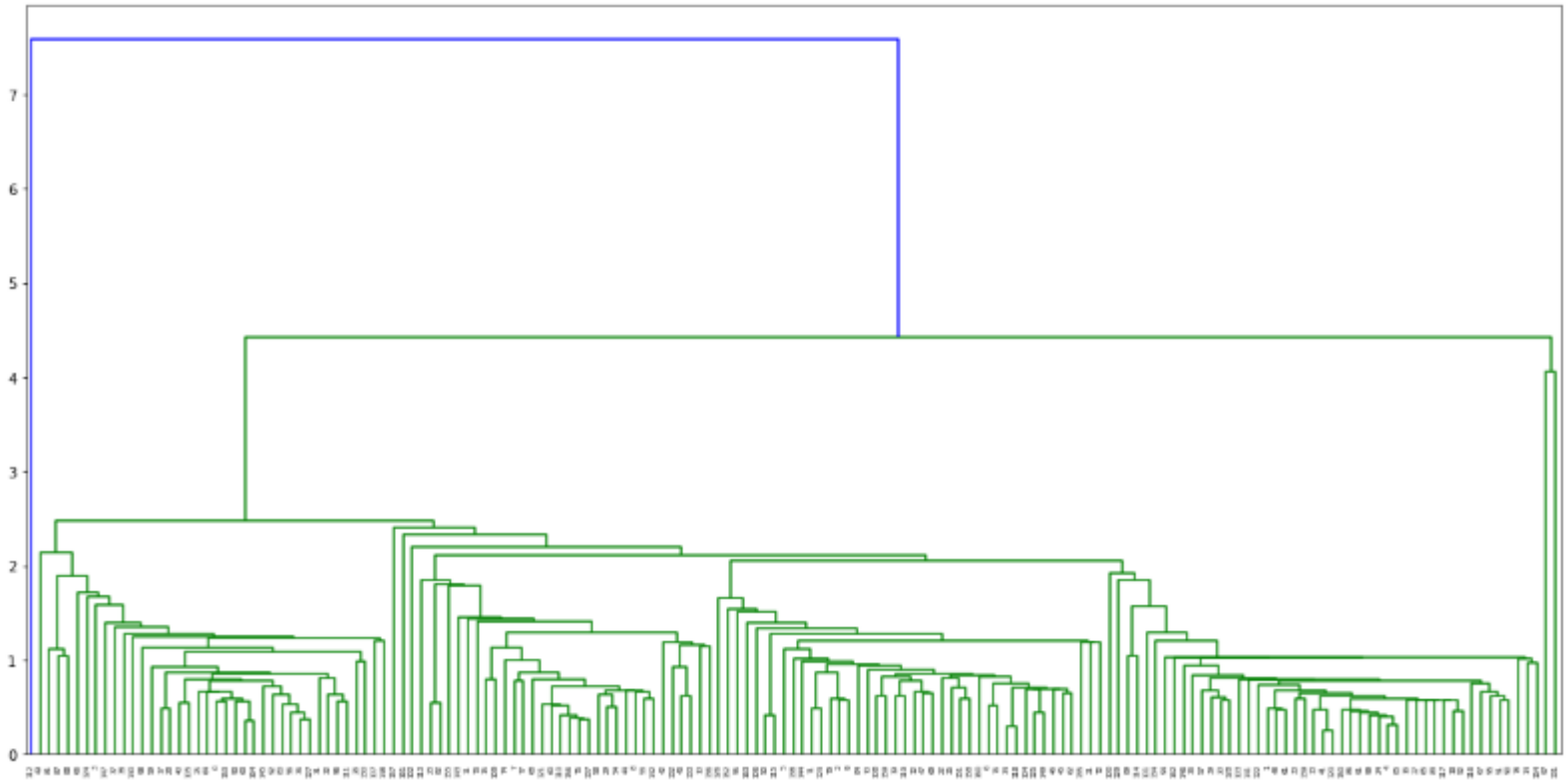
- Cluster analysis using original variables – Child Mortality

  Observation: Cluster 4 has recorded a high Child mortality.

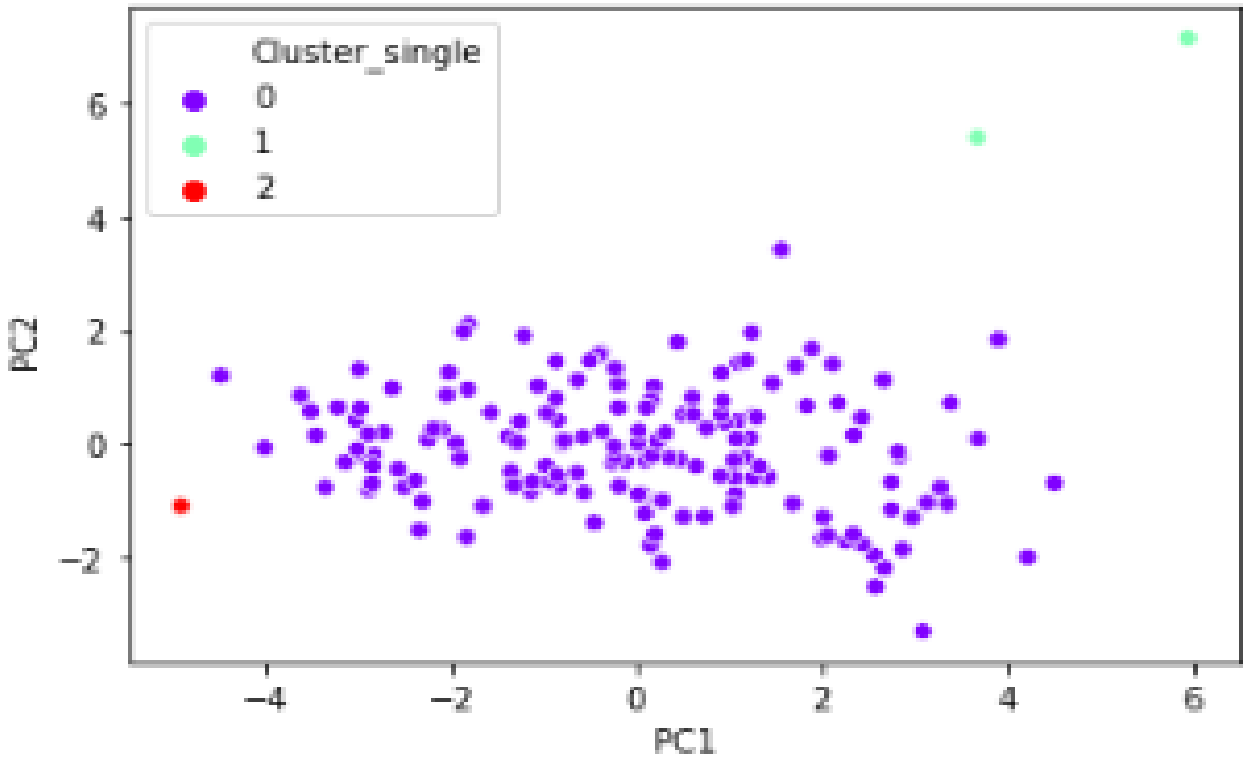


Hierarchical clustering:

- Single linkage Dendrogram:

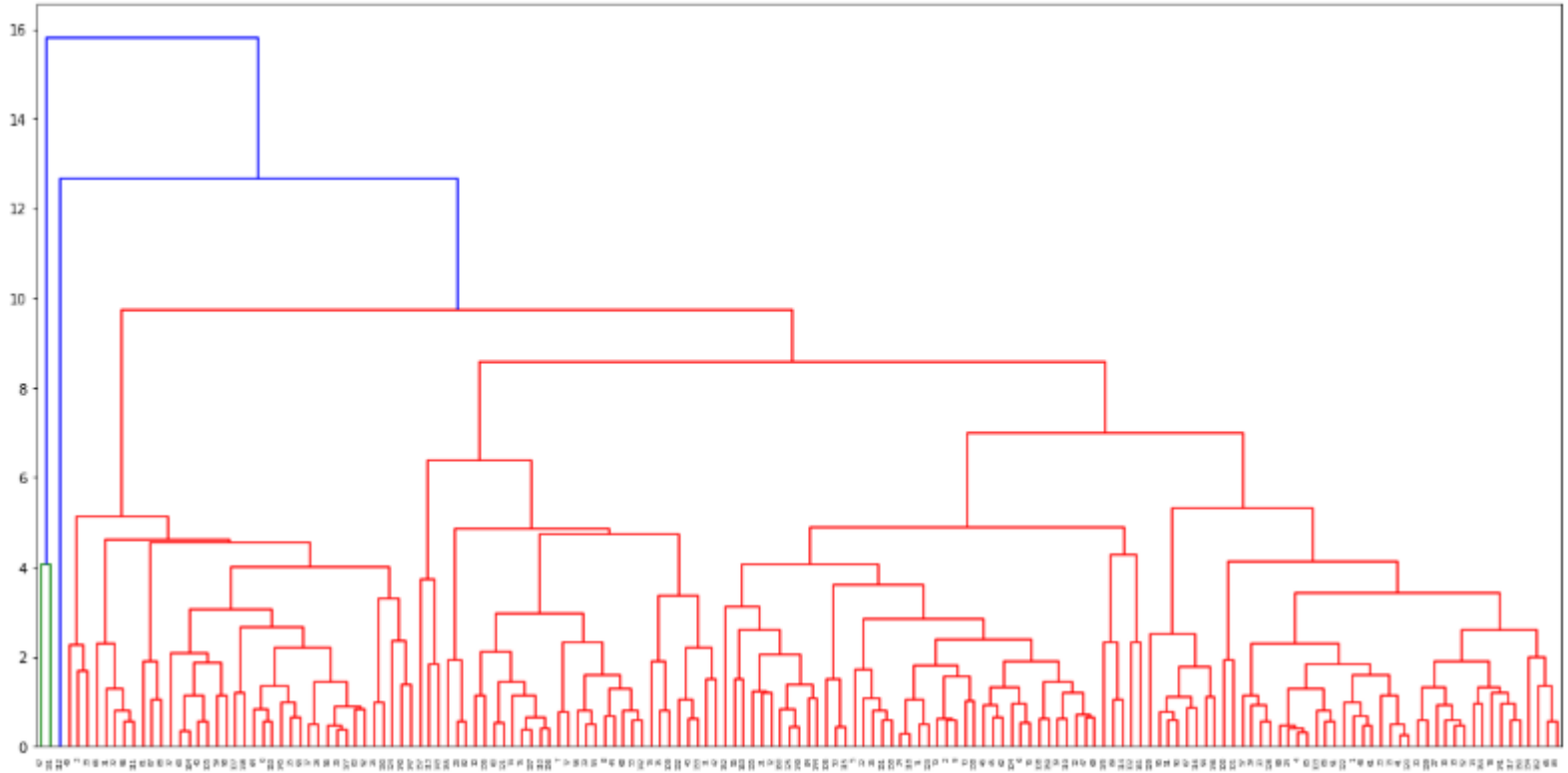


Cluster Analysis (single linkage):

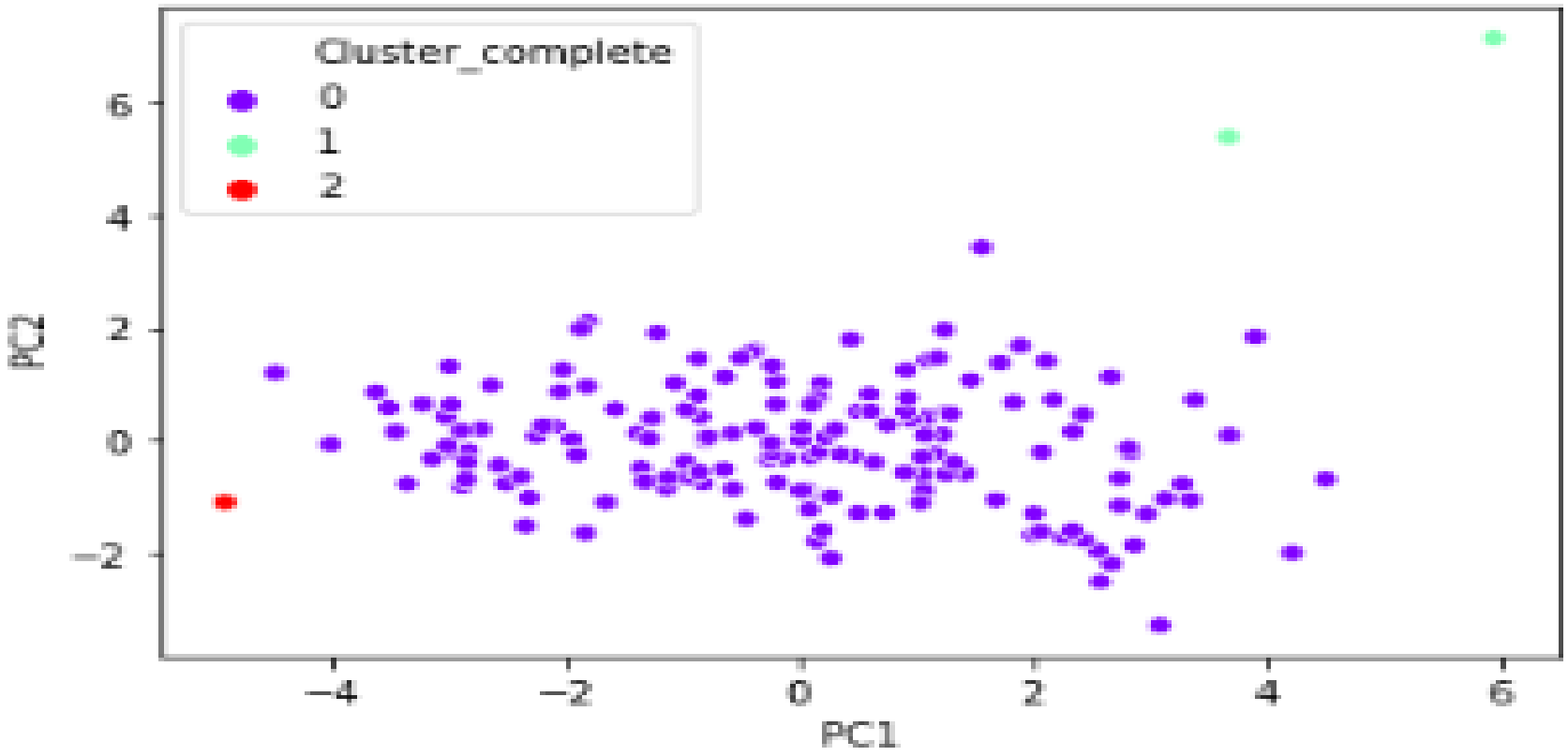


Hierarchical clustering:

- Complete linkage Dendrogram:



Cluster Analysis (complete linkage):



Inferences from the clusters:

- After analyzing the clusters from the scatter plots, data is optimally clustered(more balanced) using K-means.

- Hierarchical clustering(both single and complete linkage) generated imbalanced clusters with around 95% of data points in one cluster and rest 5% in 2 clusters.

- After analyzing data from the K-Means clustering (both with k=3 and k=5 clusters),it has been observed that both the models resulted in same set of 39 under-developed nations that need aid.

Final list of countries:

- From the 39,the top 5 countries that are in dire need of aid (based on income,gdpp and child mortality) are:

1.Burundi
2.Liberia
3.Congo, Dem. Rep.
4.Niger
5.Sierra Leone