Clustering Assignment

Presented by

- Pradhan Nayak

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 programs, 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.

And this is where you come in as a data analyst. Your job is to categorize the countries using some socio-economic and health factors that determine the overall development of the country. Then you need to suggest the countries which the CEO needs to focus on the most.

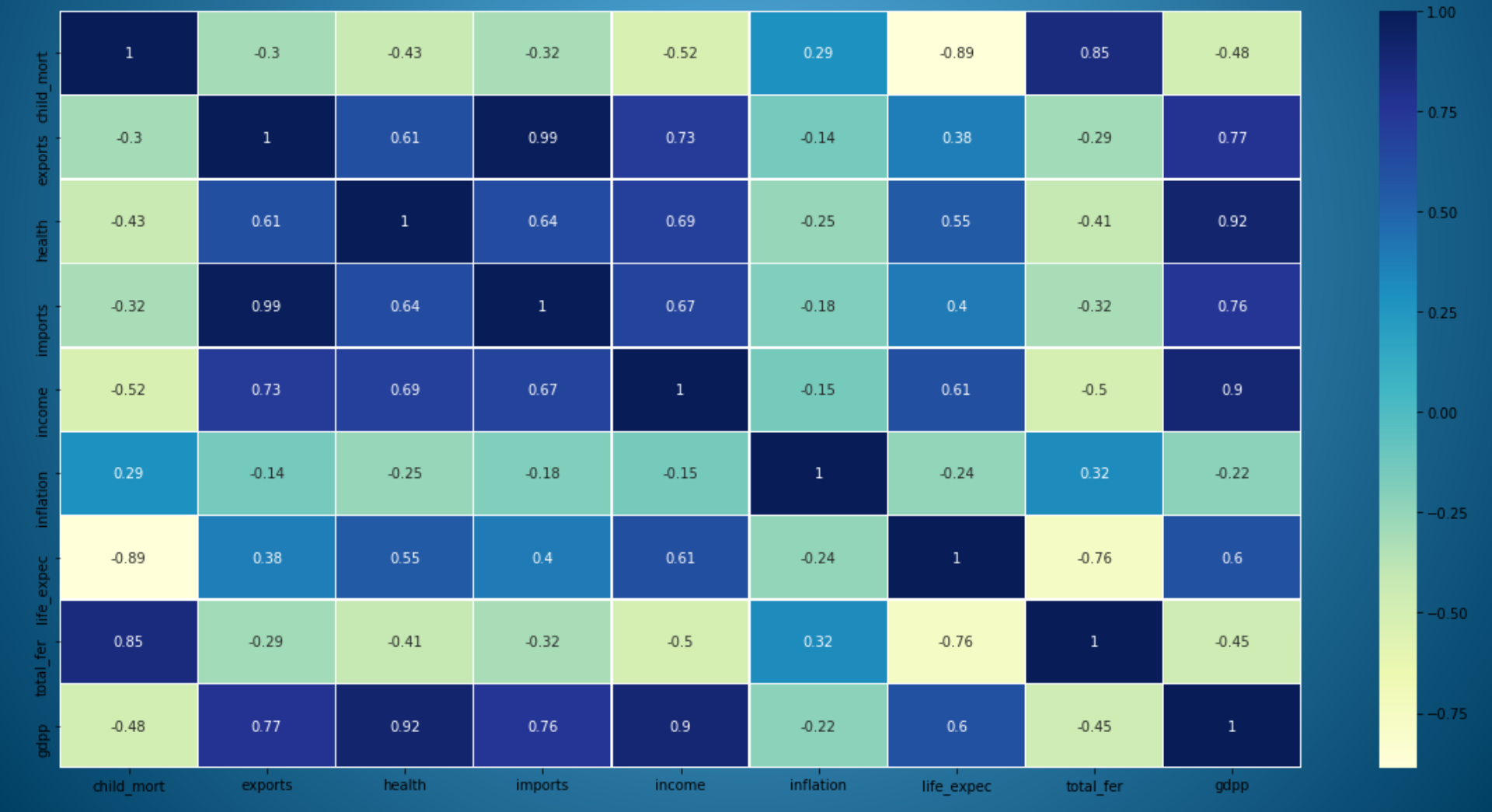
Steps to solve the assignment

1.Data Quality Check: convert imports, exports and health are %age of GDPP to absolute value.
2.EDA: Univariate and Bivariate Analysis: perform distplot and correlation matrix to check the distribution and multicollinearity of the data.
3.Outlier: perform soft capping of gdpp column as it wont affect much for the business model.
4.Hopkin's Test: How different your data is from the randomly scattered data.
5. Scaling: scaling the data with standardscaler preprocessor.
6.Find the best value of k: Silhouette Score and SSD Elbow
7.Final KMeans Analysis
8.Visualization of the clusters using scatterplot
9.Cluster profiling: gdpp, child_mort and income
10.Hierarchial Clustering: Single linkage, Complete linkage, Visualization
11. Analyze the under developed countries and identify the top 10 countries that need financial aid from the NGO.

Visualization of correlation matrix of countries dataset

Points to be concluded from the graph.

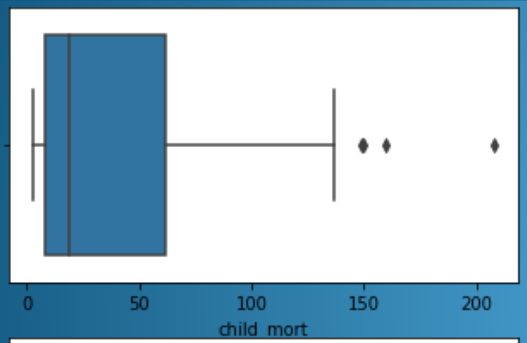
1. exports and imports have a high positive correlations.
2. Health and gdpp has high positive correlations.
3. income and gdpp has high positive correlations.
4. life expectancy and child mort has higher negative correlations.
5. life expectancy and total_fer has also negative correlations.
6. Dataset showing multicollinearity between them.

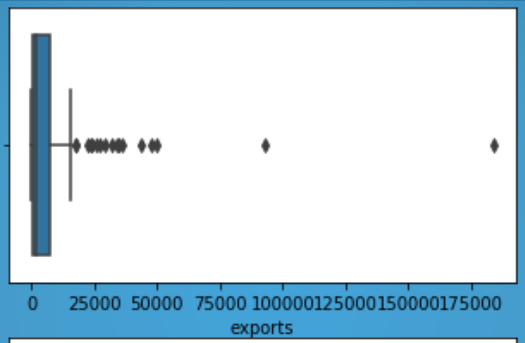


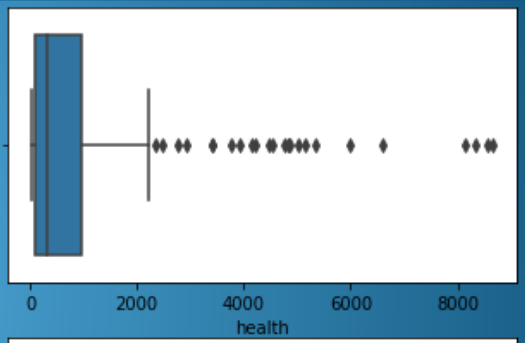
Outlier Treatment

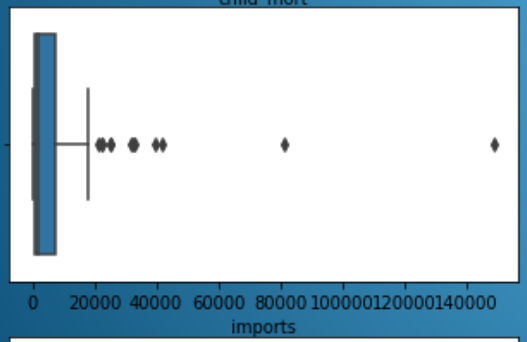
Points to be concluded from the graph.

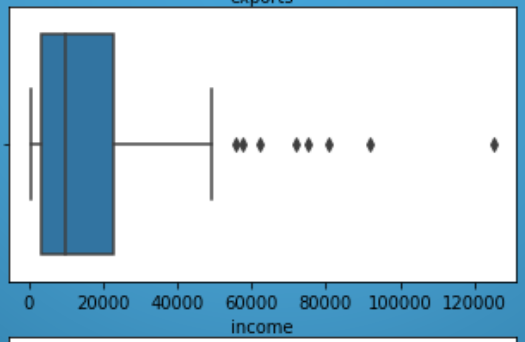
1. There are outliers present in all the columns above.
2. There is no point in removing outliers for column such as child mort, health, inflation, life expectancy as this will impact our business model which could mean we will end up not giving the financial aid to those countries which are dire need.
3. we can also proceed with the analysis without removing the outliers but it can impact our analysis with change in centroids (kmeans value).
4. but we will proceed with our analysis by removing outliers from gdpp so that caping out the higher income countries will not impact our analysis and the business model.

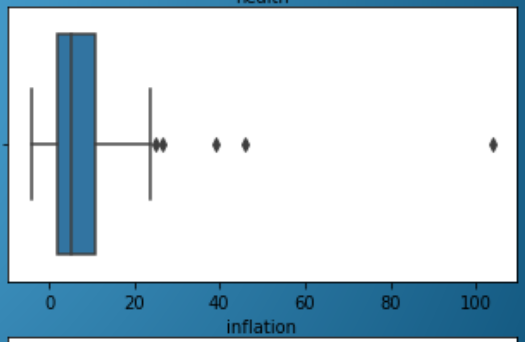


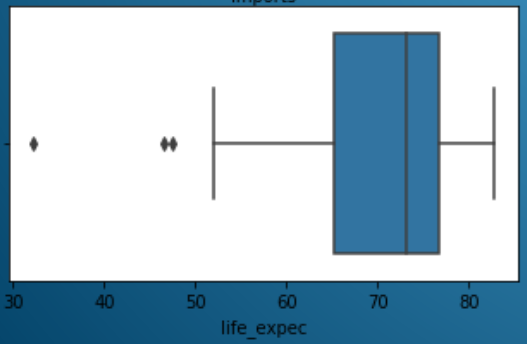


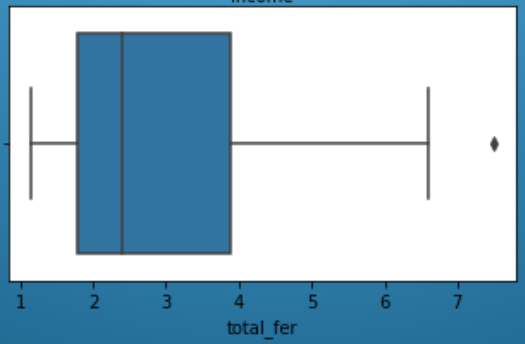


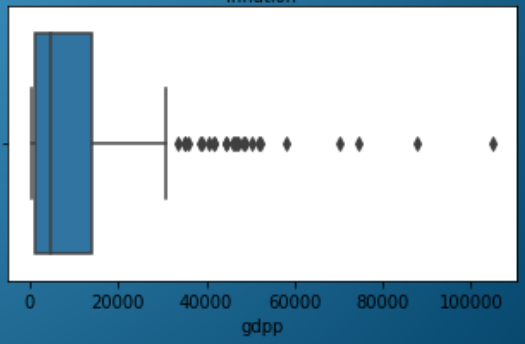








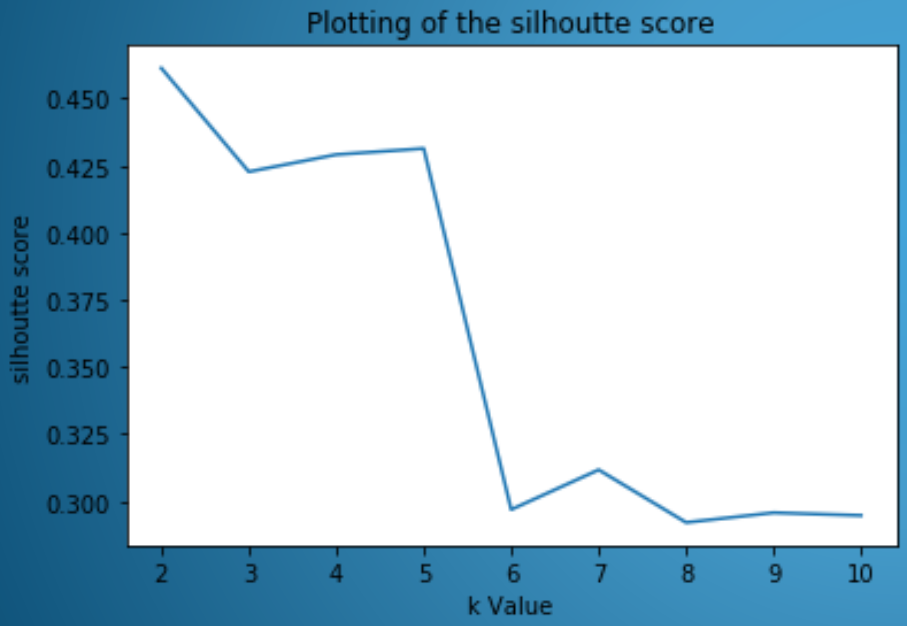


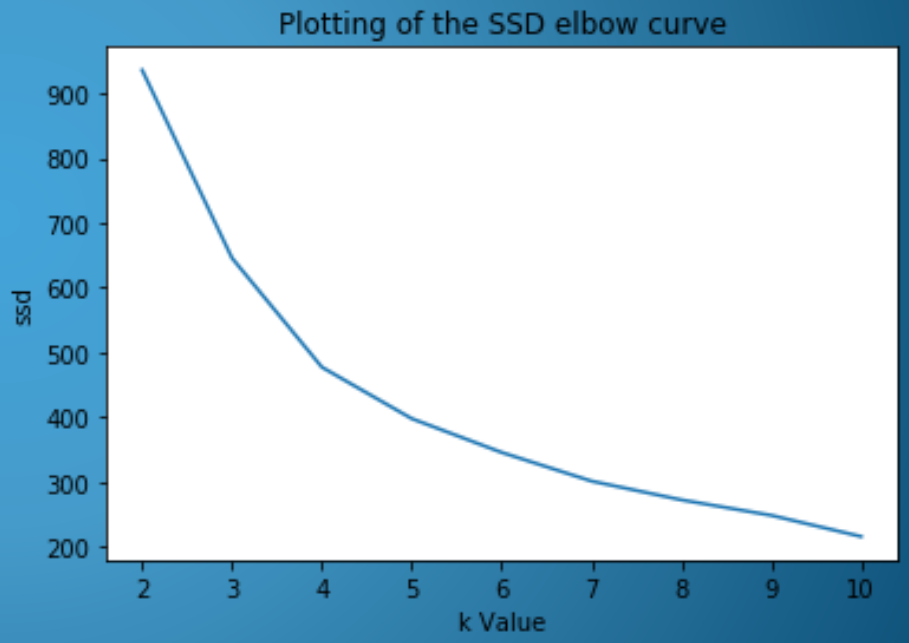


Silhouette score and SSD elbow curve

Points to be concluded from the graph.

1. From the above analysis (silhouette score and SSD Elbow Curve), we could see that 3,4 or 5 clusters are optimal number of clusters to be used.
2. We will try 3 different iterations in K-Means clustering using 3,4 and 5 Clusters and analyze the results.





Hopkins statistics

Points to be concluded from the graph.

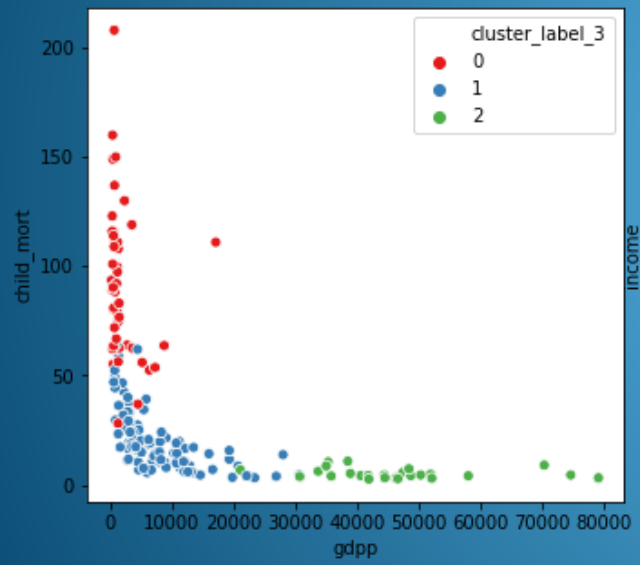
1. To check if the given data will have some meaningful clusters.
2. To check if the given data is not random.
3. To check also know as cluster Tendency.
4. we can understand that our data is 92.78% different from randomly scattered data.
5. Less than 50 - Useless
6. 50-65 - Can Consider
7. 65-84 - Good
8. 85 and Above - Green signa

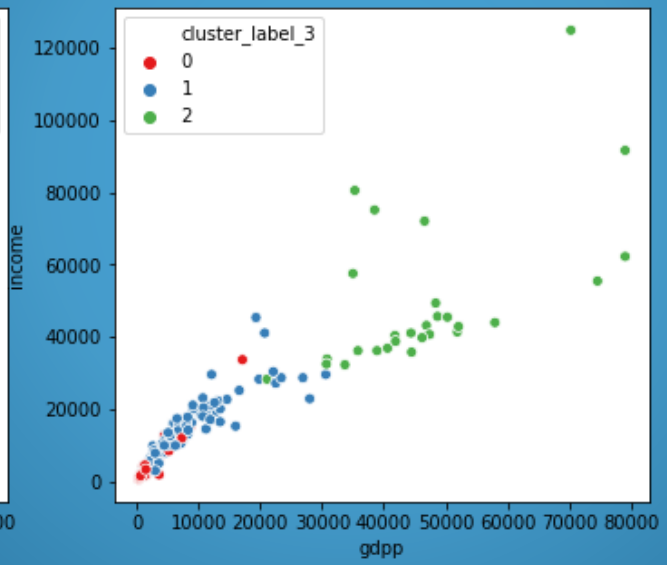
```
from sklearn.neighbors import NearestNeighbors
from random import sample
from numpy.random import uniform
import numpy as np
from math import isnan

def hopkins(X):
    d = X.shape[1]
    #d = len(vars) # columns
    n = len(X) # rows
    m = int(0.1 * n)
    nbrs = NearestNeighbors(n_neighbors=1).fit(X.values)

    rand_X = sample(range(0, n, 1), m)

    ujd = []
    wjd = []
    for j in range(0, m):
        u_dist, _ = nbrs.kneighbors(uniform(np.amin(X,axis=0),np.amax(X,axis=0),d).reshape(1, -1), 2, return_distance=True)
        ujd.append(u_dist[0][1])
        w_dist, _ = nbrs.kneighbors(X.iloc[rand_X[j]].values.reshape(1, -1), 2, return_distance=True)
        wjd.append(w_dist[0][1])

    H = sum(ujd) / (sum(ujd) + sum(wjd))
    if isnan(H):
        print(ujd, wjd)
        H = 0

    return H
```

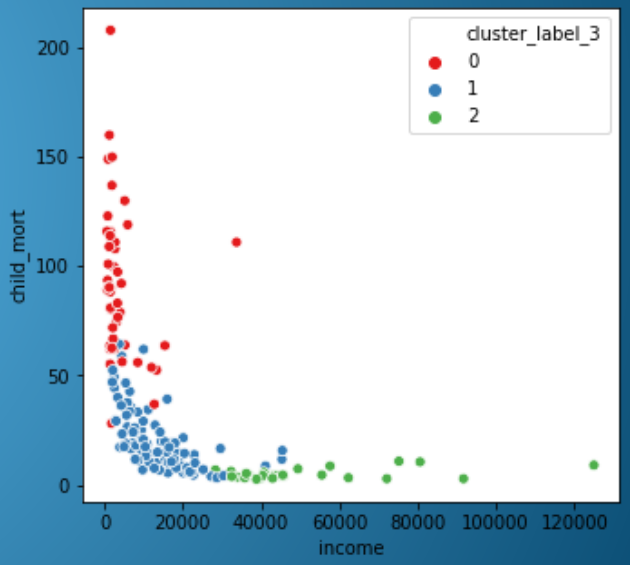
Visualization of kmeans clustering using scatter plot

Points to be concluded from the graph.

1. child_mort seems to be high for lower value of gdpp
2. gdpp increases as income increases.
3. child_mort is high for low income.



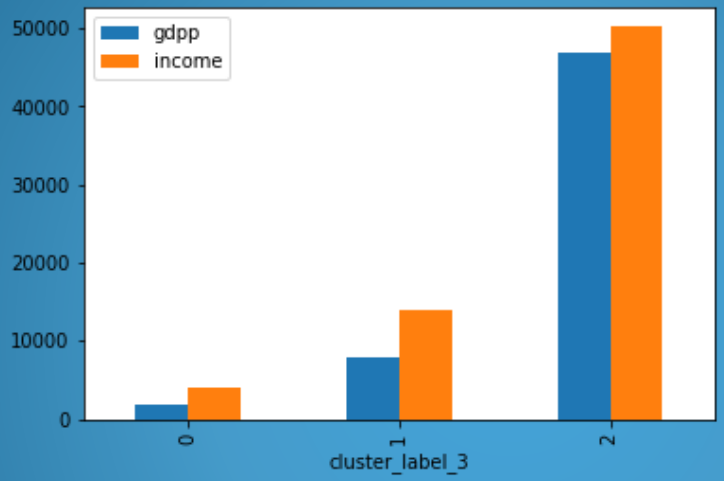


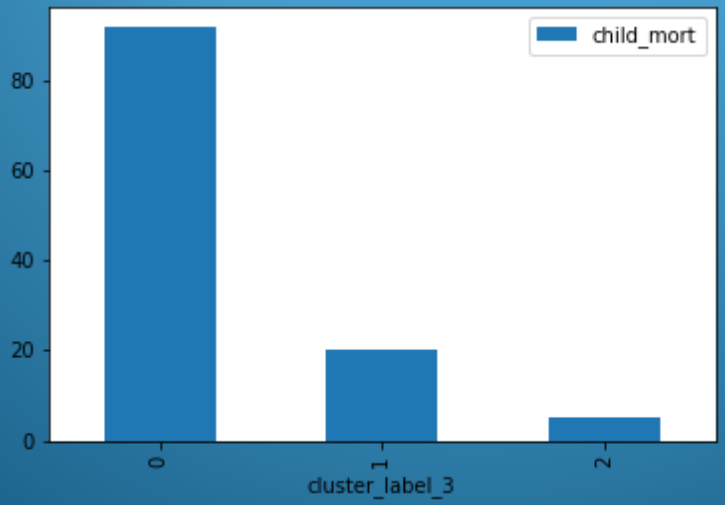


Kmeans Cluster profiling: gdpp, income,child_mort

Points to be concluded from the graph.

1. cluster 0: seems to high child_mort compared to other clusters.
2. cluster 0 : gdpp and income also is lower compared to other clusters.
3. Hence cluster 0 is recommended for the financial aid from the HELP international NGO.

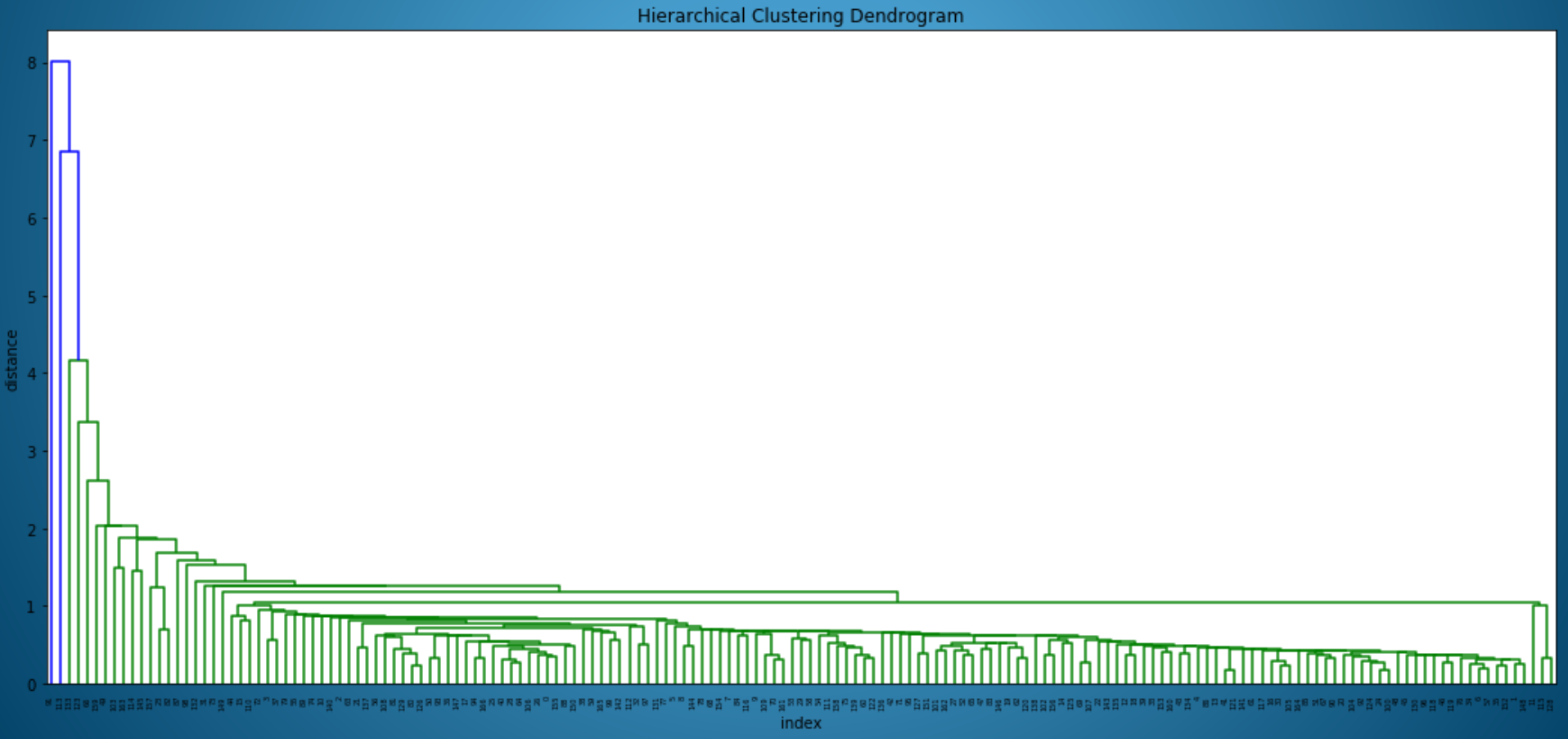




Hierarchical Clustering: Single linkage

Points to be concluded from the graph.

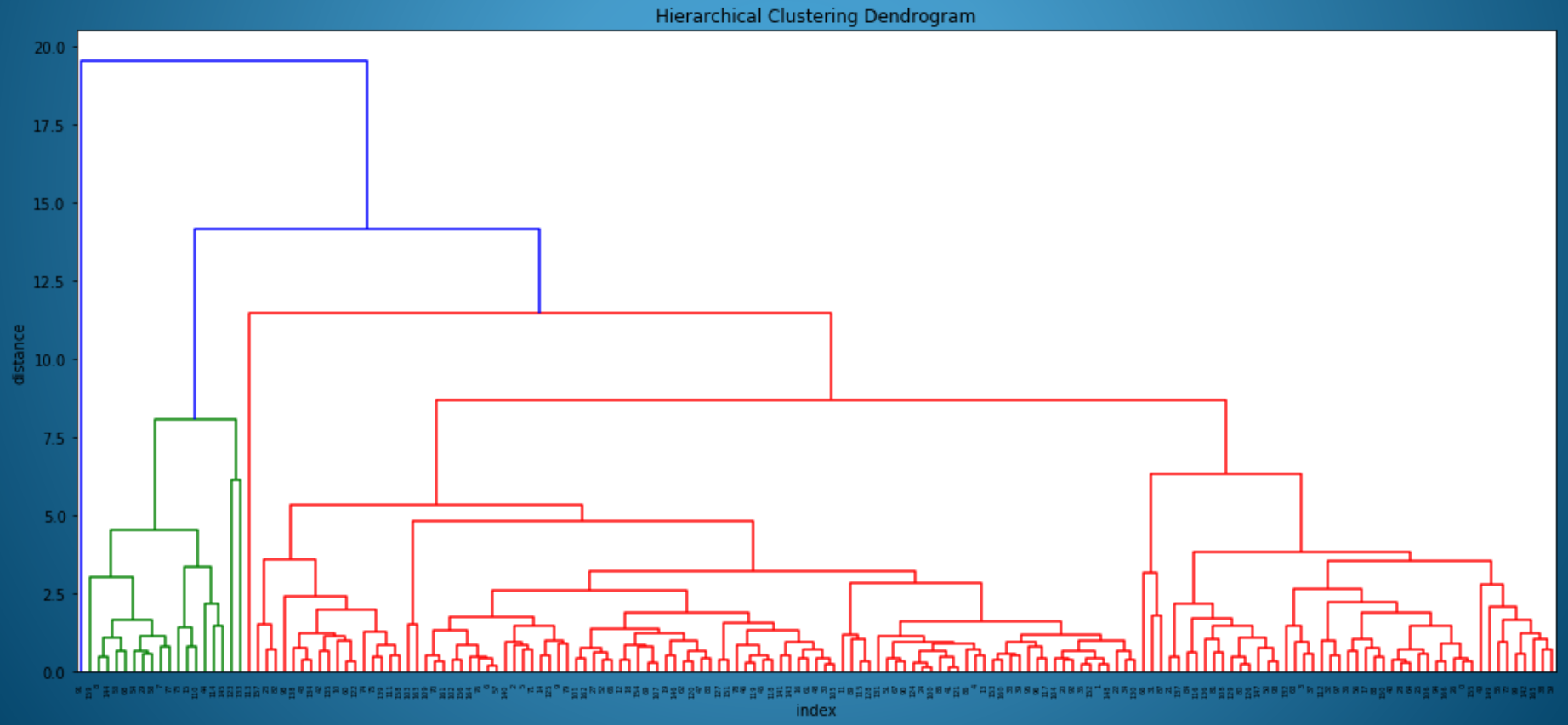
1. Here, the distance between 2 clusters is defined as the shortest distance between points in the two clusters
2. a single linkage-type will produce dendrograms which are not structured properly.
3. we can cut the tree in a threshold value, we will use complete linkage dendrogram for hierarchical clustering.



Hierarchical Clustering: Complete linkage

Points to be concluded from the graph.

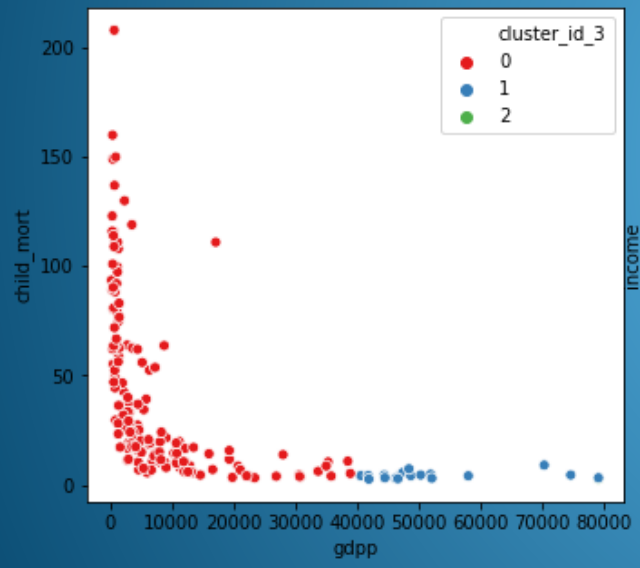
1. This graph shows proper way to decide number of clusters needs to be used by cutting at threshold value.
2. We will cut at 3 branches which will give us 3 clusters

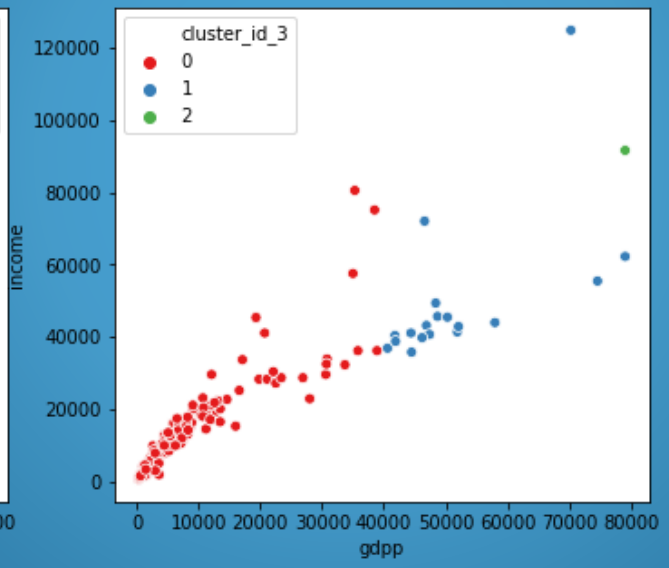


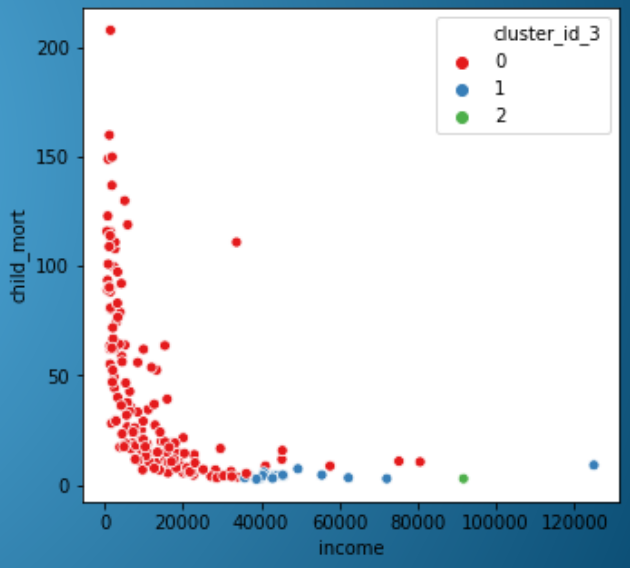
visualization of hierarchical clustering using scatter plot

Points to be concluded from the graph.

1. child_mort is high when gdpp is low.
2. income seems to be low when gdpp is also low.
3. child_mort is high when income is low.



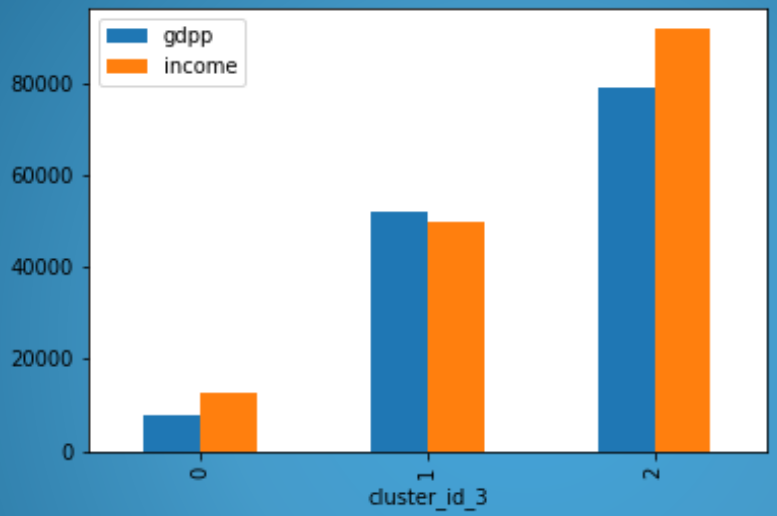


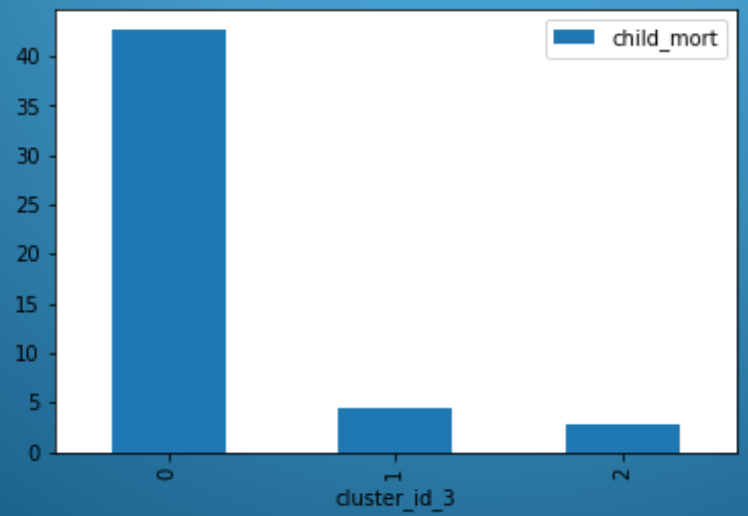


Hierarchical Cluster profiling: gdpp, income,child_mort

Points to be concluded from the graph.

1. child_mort is high for cluster 0
2. income and gdpp is low for cluster 0
3. segmentation of clusters clear identifies clustero: under developed countries, cluster 1: developing countries, cluster 2: developed countries.

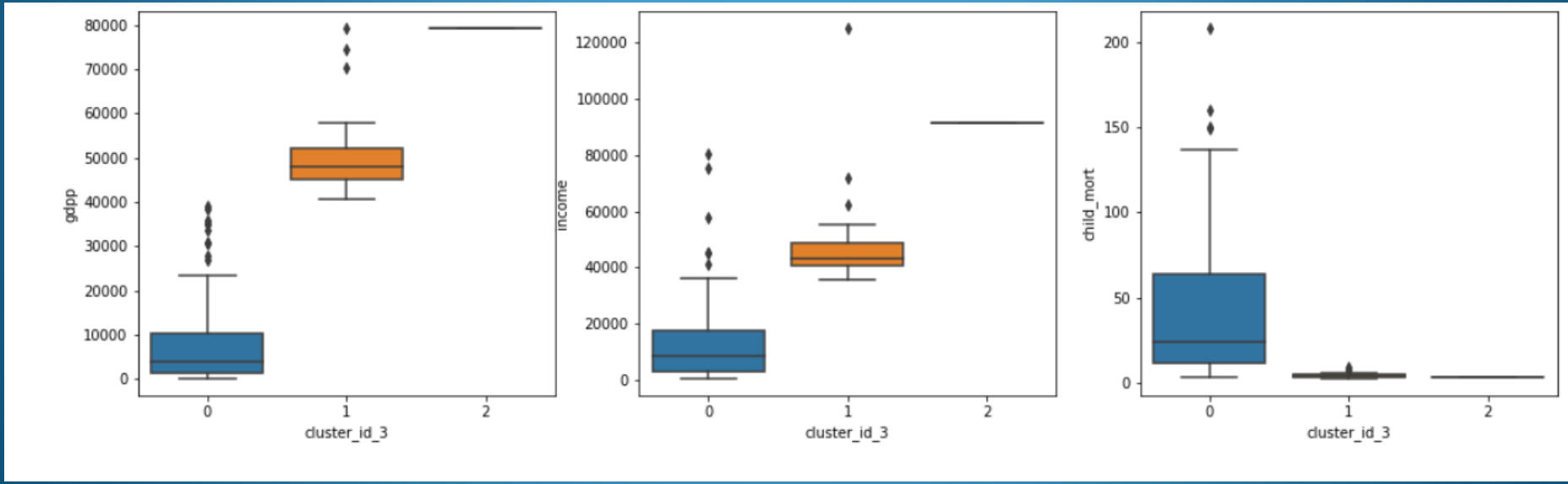




Hierarchical clustering: comparison of gdpp vs income vs child_mort with k=3

Points to be concluded from the graph.

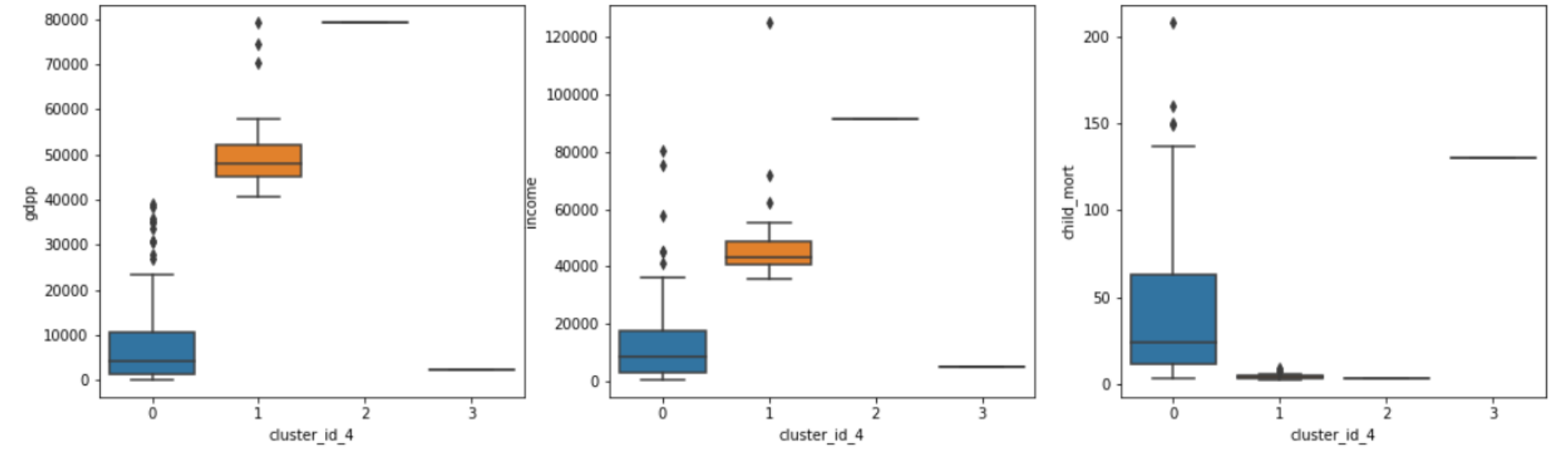
1.outliers identified for gdpp for cluster 0 and 1.
2.outliers identified for income for cluster 0 and 1.
3.outliers identified for child_mort for cluster 0 and 1
4. no data points are grouped in cluster 2.



Hierarchical clustering: comparison of gdpp vs income vs child_mort with k=4

Points to be concluded from the graph.

1. outliers identified for gdpp for cluster 0 and 1.
2. outliers identified for income for cluster 0 and 1. 3.outliers identified for child_mort for cluster 0 and 1
3. no data points are grouped in cluster 2 and 3



Final Conclusion:

Points to be concluded from the graph.

1. performed k means clustering and hierarchical clustering to identify countries that need Financial Aid from HELP international NGO.
2. Based on Hierarchical clustering analysis we concluded that a cluster k=3 gave us better segmentation but due to data imbalance we dropped Hierarchical clustering.
3. Finally we performed k means clustering and found k=3,4,5 as optimal clusters and after final analysis we concluded k=3 gives a better cluster balance and segmentation of countries into different clusters.
4. we identified 10 countries that are in dire need of financial aid from our final set of analysis.

country	child_mort	income	gdpp
Burundi	93.6	764	231.0
Congo, Dem. Rep.	116.0	609	334.0
Niger	123.0	814	348.0
Sierra Leone	160.0	1220	399.0
Mozambique	101.0	918	419.0
Central African Republic	149.0	888	446.0
Malawi	90.5	1030	459.0
Togo	90.3	1210	488.0
Guinea-Bissau	114.0	1390	547.0
Afghanistan	90.2	1610	553.0

Thank YOU