Clustering Assignment
(Countries Dataset)

BY- SUMIT RAGHUVANSHI

Objective

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.

Problem statement:

During the recent funding programs, NGO 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 Approach:

• Data Collection And Data Cleaning: Importing the data then cleaning it , checking if there are any null values. We found out that some columns were in %of gdpp form so corrected them using the correct formulas.

• Visualizing data: We detected outliers by visualizing the data , outliers were treated according to our problem statement. We also found out that some Variables are highly corelated to each other.

• Outliers Detection and treatment: There were outliers in almost every columns. Some outliers like in gdpp column for example, have outliers on high end of spectrum which we can remove safely because high gdpp countries wont need urgent aid. For this analysis we capped the gdpp column at .95 quantile at upper end and .1 quantile at lower end. We did not capped other variable because that would effect our analysis for finding poorest countries with high child mortality.
• Scaling data: Standardizing all the continuous variables.

• Kmeans Clustering: Identifying the “k” through silhouette analysis and elbow curve. Then forming the cluster on scaled data the adding the cluster id on original data for better interpretation of data. And visualizing the clusters.

• Hierarchical Clustering Identifying optimal number for k by analyzing dendrogram. Then forming the cluster on scaled data and adding the cluster label to original data for better interpretation. Visualization of clusters was also done.

• Decision Making: Successfully identified the top 10 countries by analyzing both model which are in dire need of Aid

Data visualizations
Correlation of the variables: Countries

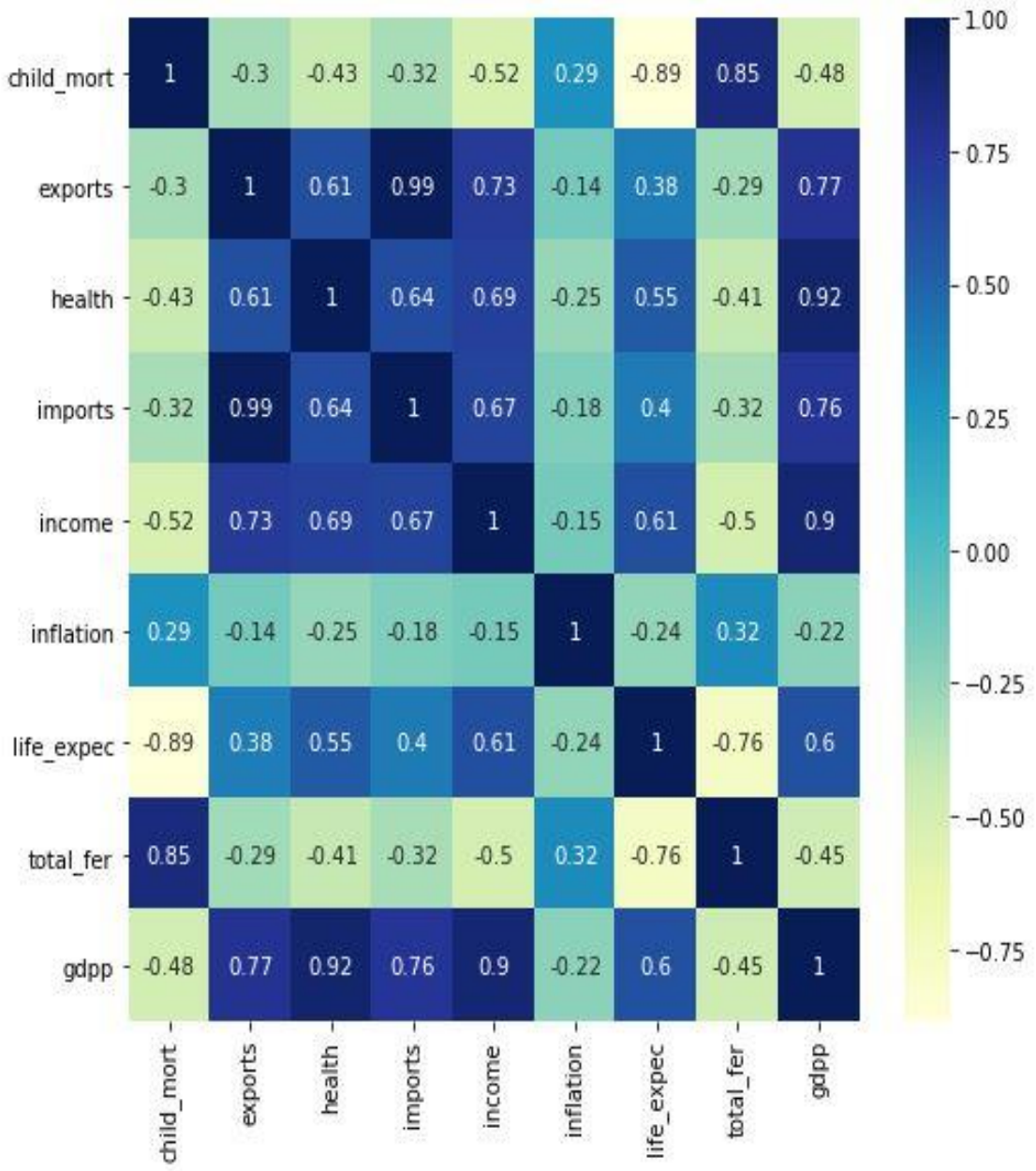
From the heatmap on the right we can conclude below points –

•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

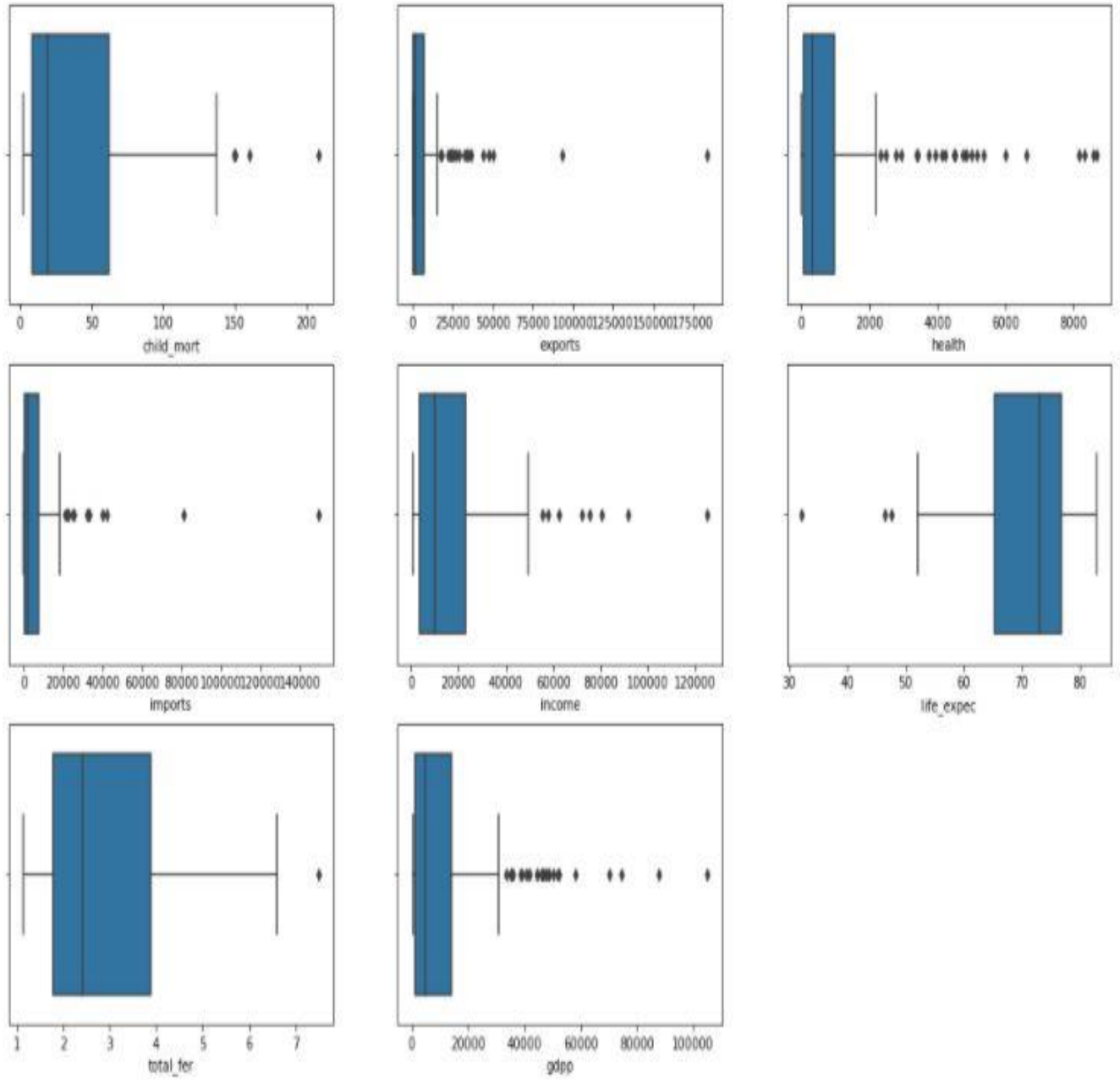
•life_expentency and total_fertility are highly correlated with correlation of -0.76



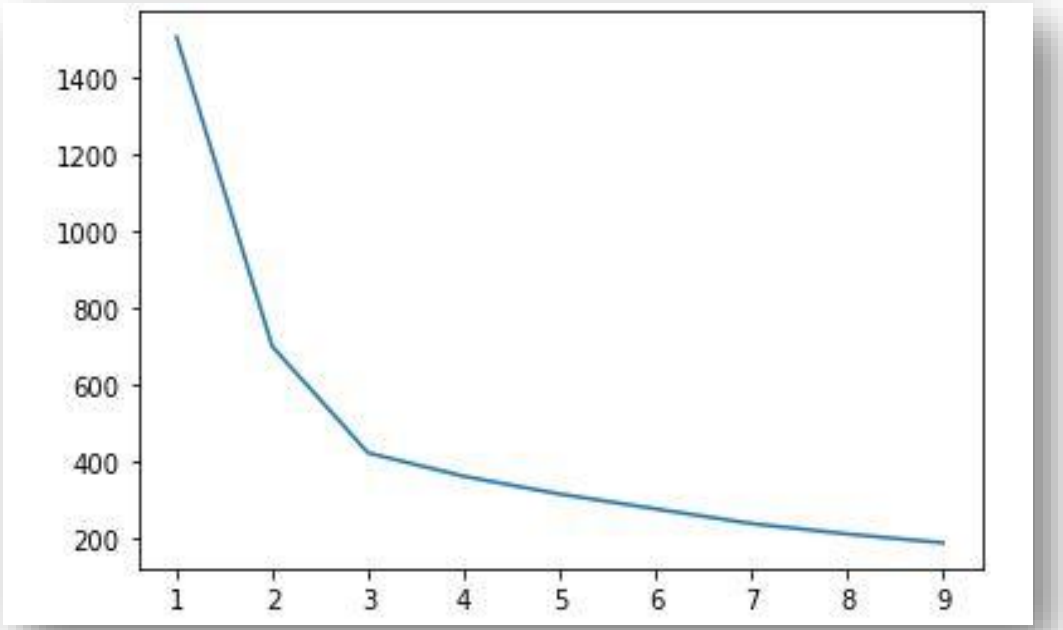
Visualization of Outliers

From the boxplots attached on the left, points to be concluded –

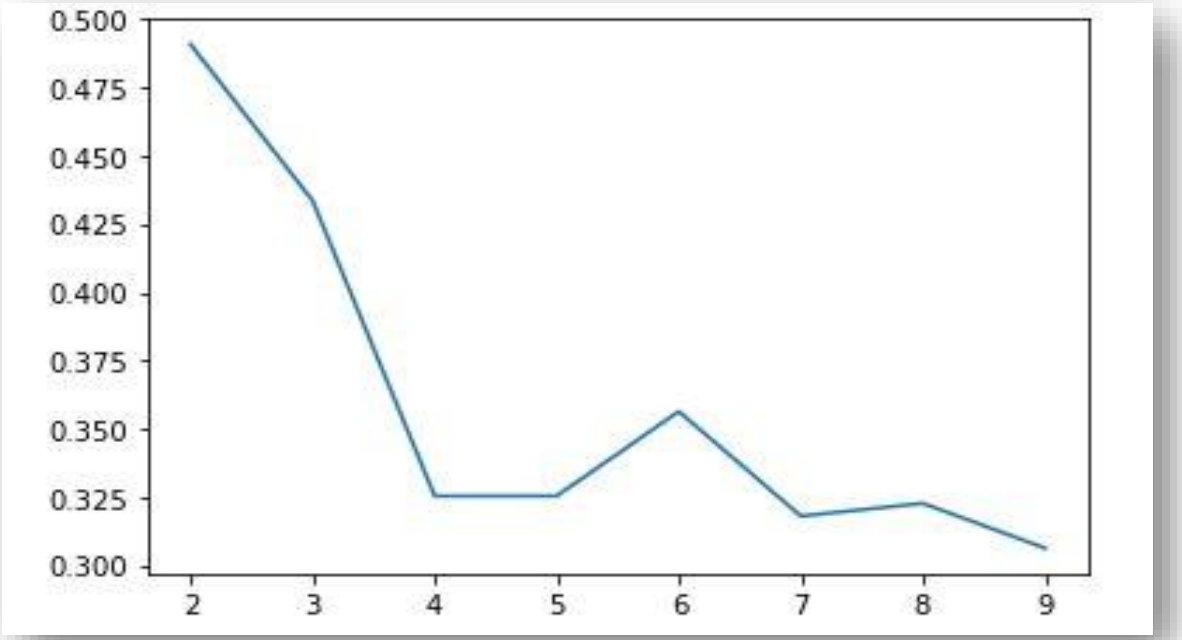
- As we can see that all the boxplots created for the variables are having decent amount of outliers.
- All boxplots except one ‘life_expec’ is having outliers on the bottom of the boxplots which means there are some countries were no. of children is very less compared to the other countries.
- The Imports boxplot is having very thin size of quartiles compared to others.



K-Means Clustering



Elbow Curve



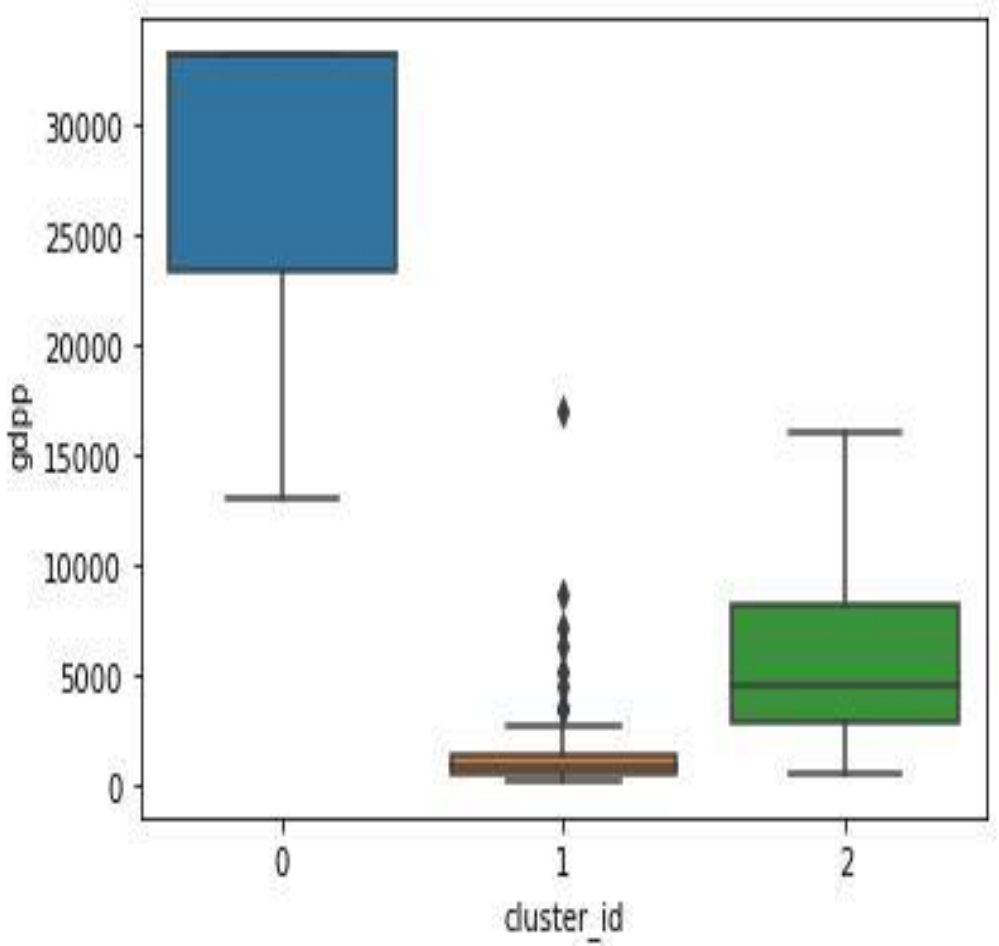
Silhouette Curve

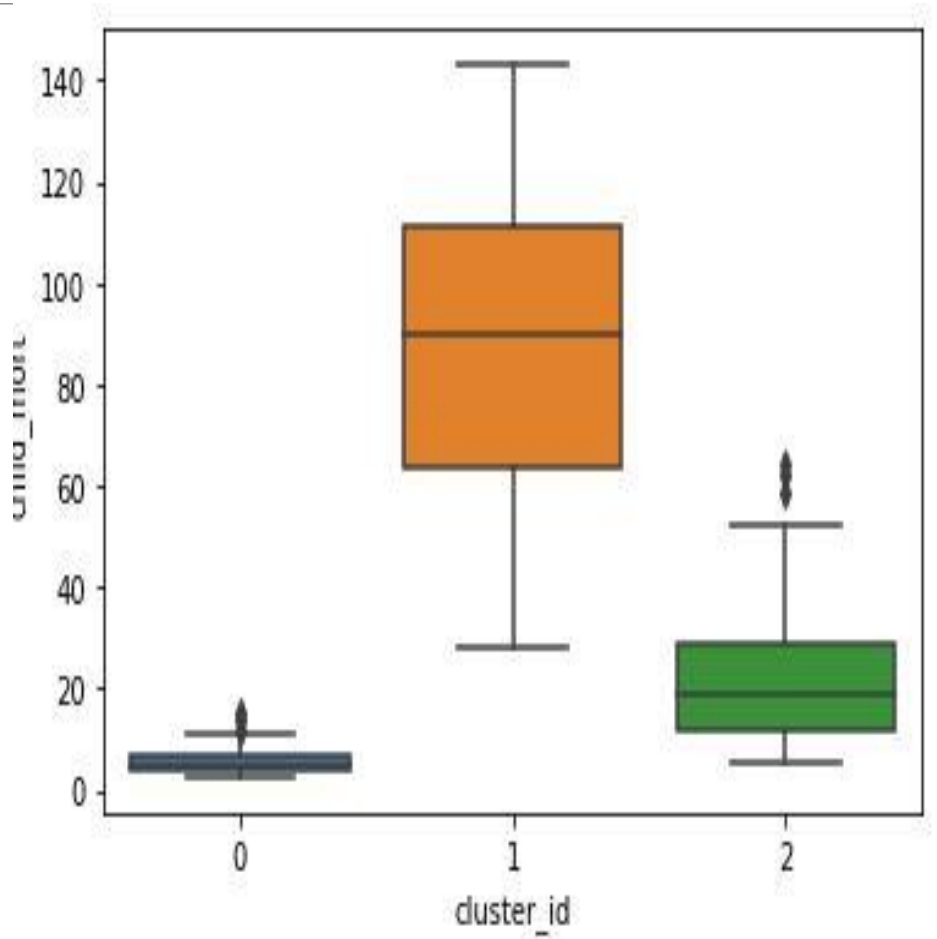
We can see in silhouette analysis that highest peak is at 2, but 2 is never a good number for clustering, on the other hand elbow curve has elbow at 3. So we will go with k=3.

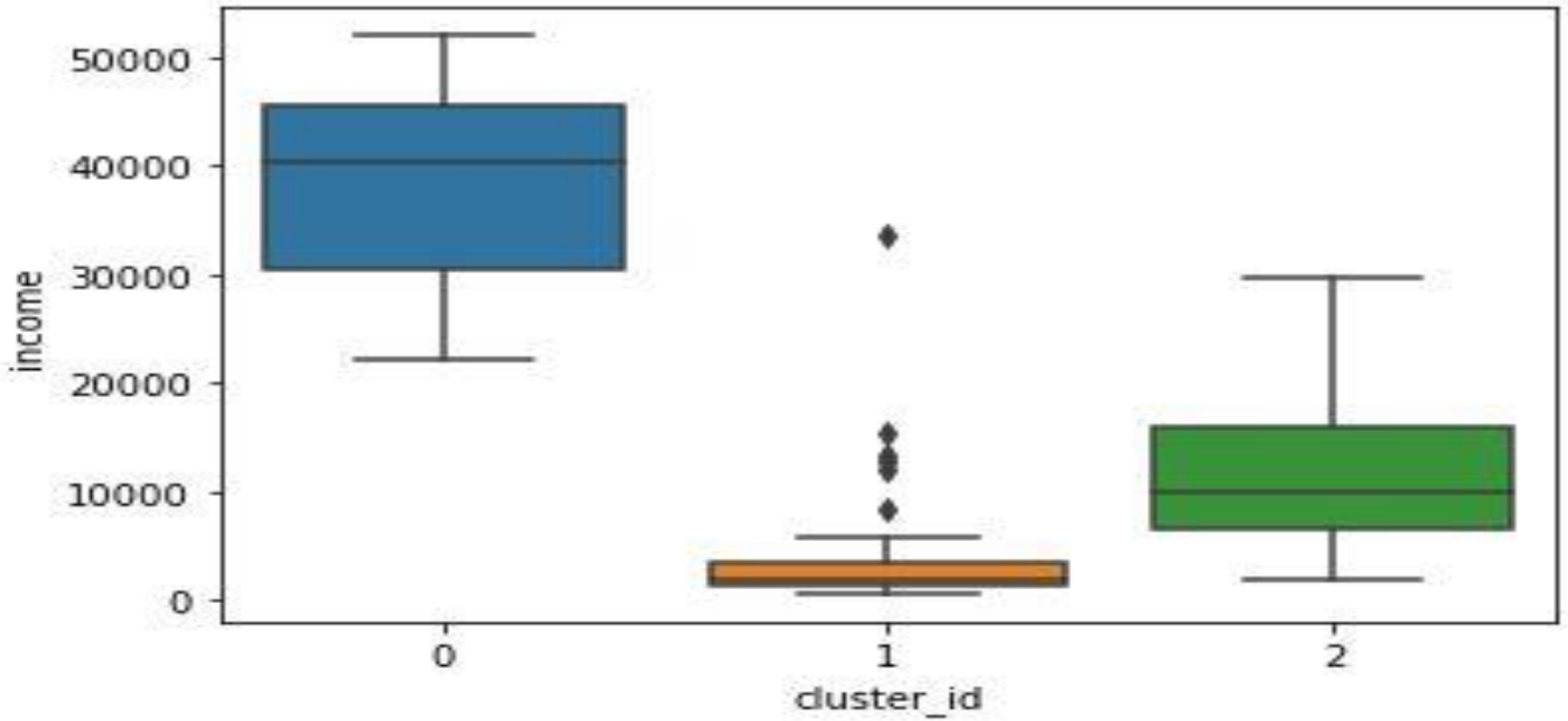
KMeans Clustering Analysis

- We are trying to find the optimum value of the k-value based on the business requirements.
- So, to achieve this we used silhouette analysis to find the score of range of cluster values.
- The silhouette scores after the implementations are below
    - For no. of cluster=2, silhouette score is 0.498013162580517 7
    - For no. of cluster=3, silhouette score is 0.433601771392598 6
    - For no. of cluster=4, silhouette score is 0.324381028570179 5
    - For no. of cluster=5, silhouette score is 0.32467504953979 91
    - For no. of cluster=6, silhouette score is 0.29921334336486793
    - For no. of cluster=7, silhouette score is 0.31316743132103253
    - For no. of cluster=8, silhouette score is 0.31195076705799 57
- We found that cluster =2 is having highest score hence k value should be 2 but, if we choose k value as ‘2’ , it will not suit business needs.
- Hence, we use k value as ‘3’ since it is giving precise information also fulfills business needs.

Visualization of the original variables with clusters







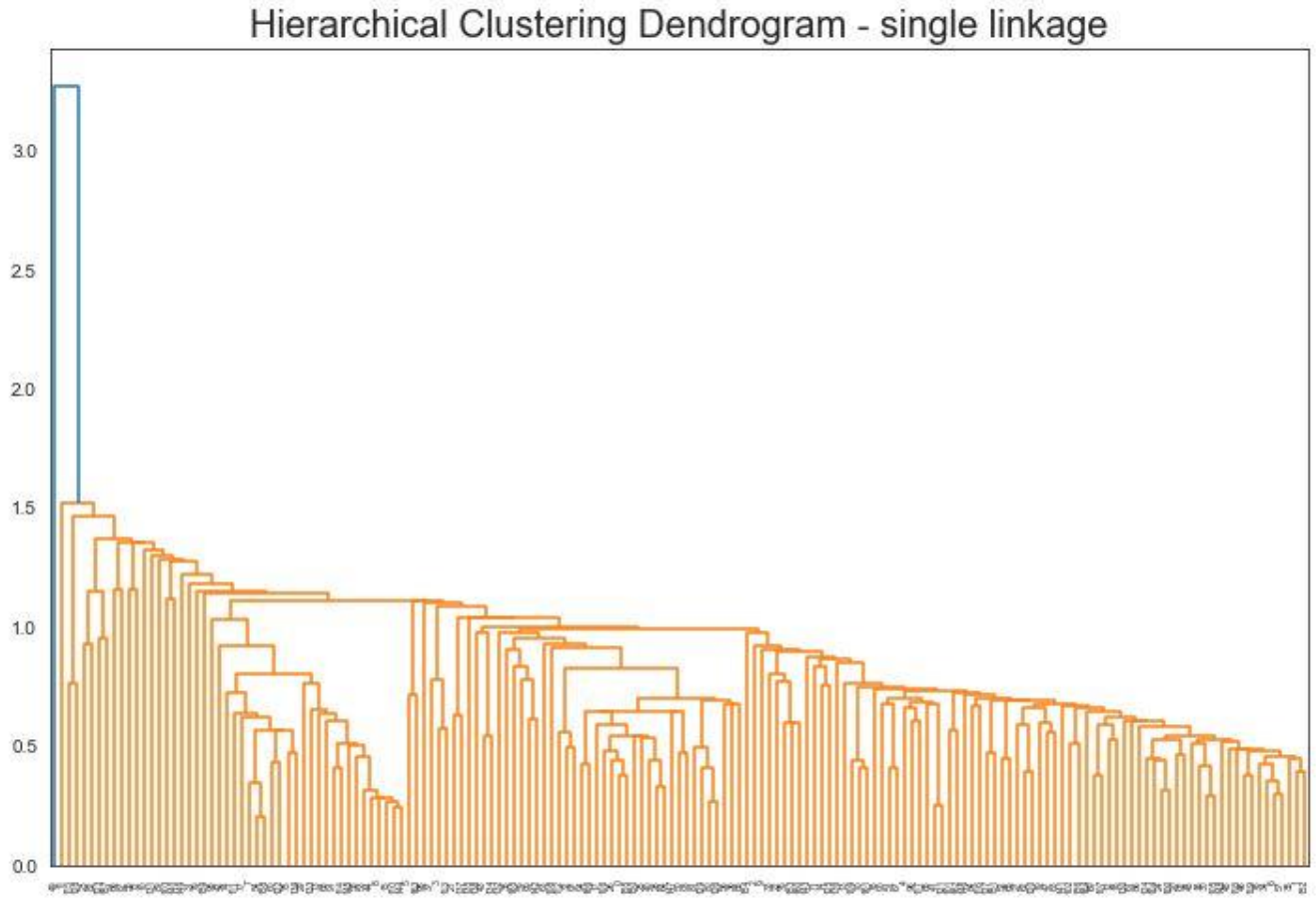
For cluster id 1 both gdpp and income are very low and child mortality is very high. So this cluster is our main focus for providing aid to countries.

Hierarchical Clustering

Single Linkage

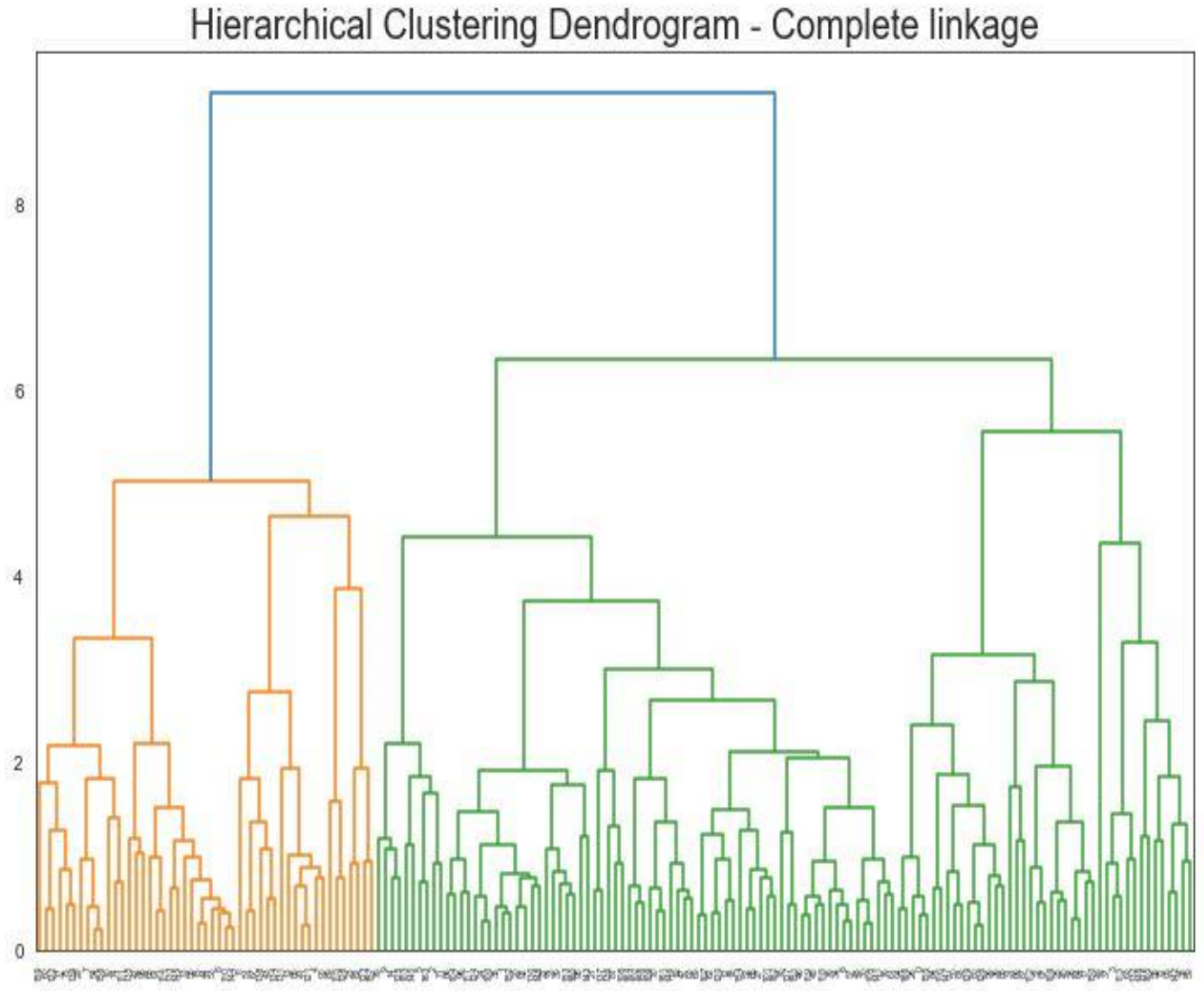
This is another method to find our low development countries.

As we can see from the graph of linkage dendrogram, it is not quite visible and doesn't not suits properly with the dataset because we can cut the tree in a threshold value, we will use complete linkage dendrogram for hierarchical clustering.

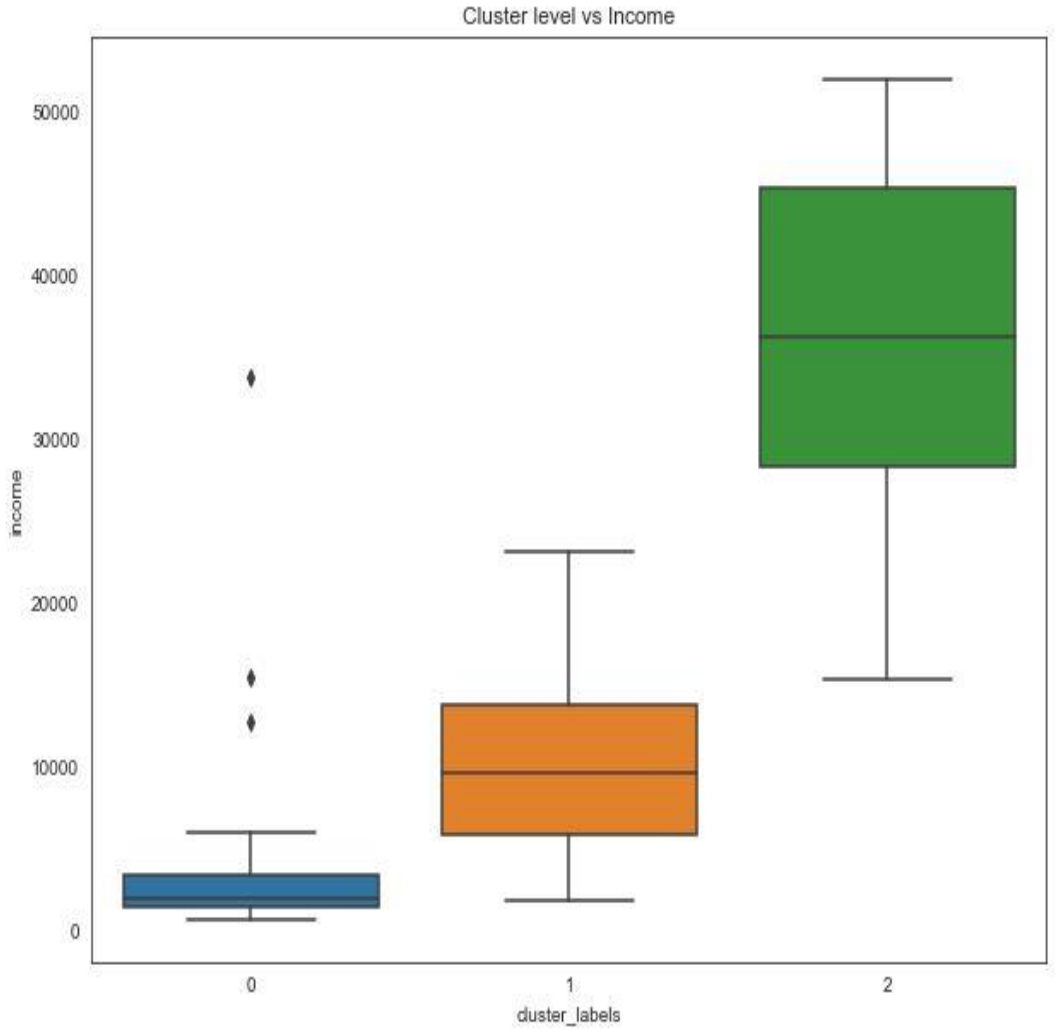


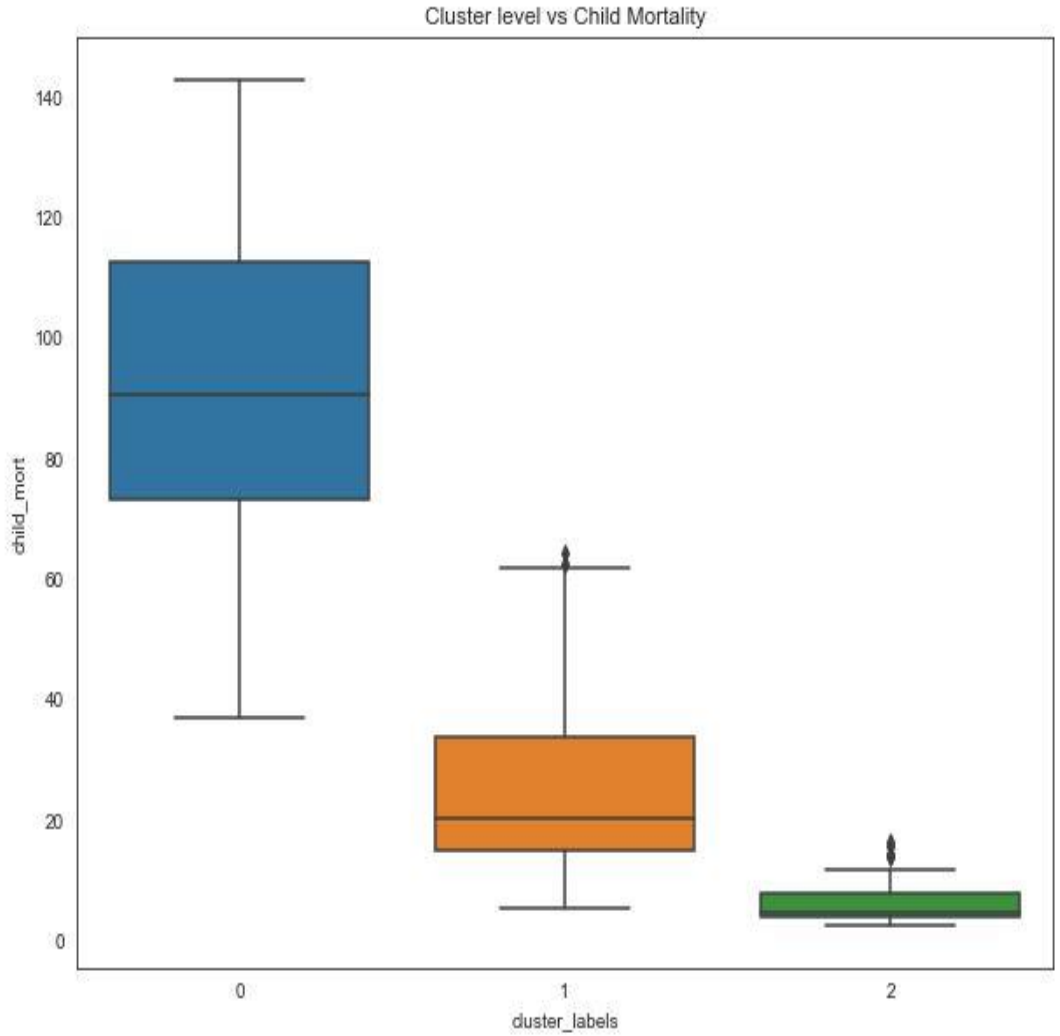
COMPLETE LINKAGE

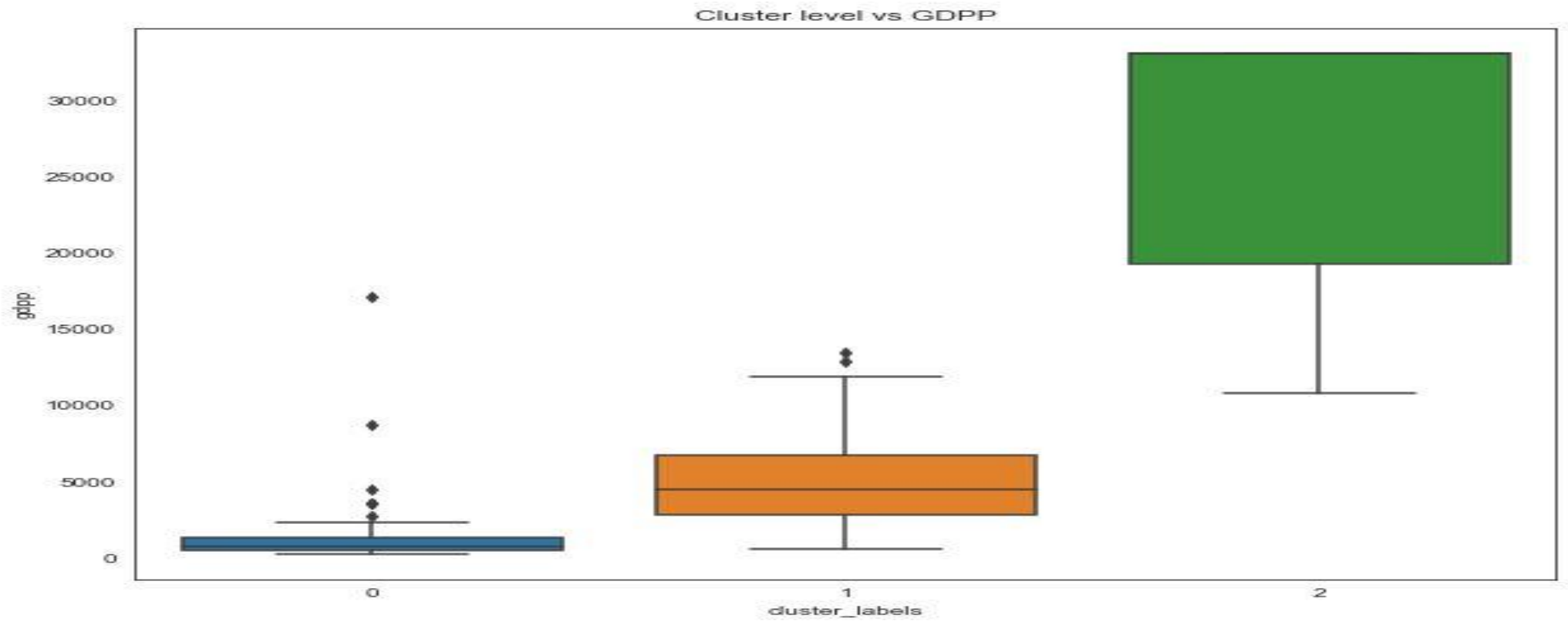
Points noted from the graph on the right –
This graph shows proper way to decide number of clusters needs to be used by cutting at threshold value.
We will cut at 3 branches which will give us 3 clusters



Visualization of the original variables with clusters







For Cluster 0 both income and gdpp is vey low and child mortality is very high. So this cluster will be our focus for countries that are in dire need of aid.

Final list of countries

We got same top 10 countries from both hierarchical and k-means model that are in dire need of Aid.

Following are the countries name requiring aid:

1. Sierra Leone
2. Central African Republic
3. Haiti
4. Chad
5. Mali
6. Nigeria
7. Niger
8. Angola
9. Congo , Dem. Rep.
10. Burkina Faso

These are the categorize of countries using some socio-economic and health factors that determine the overall development of the country.