



Clustering Assignment: HELP-NGO

-By Prajna Deepthi Mantha



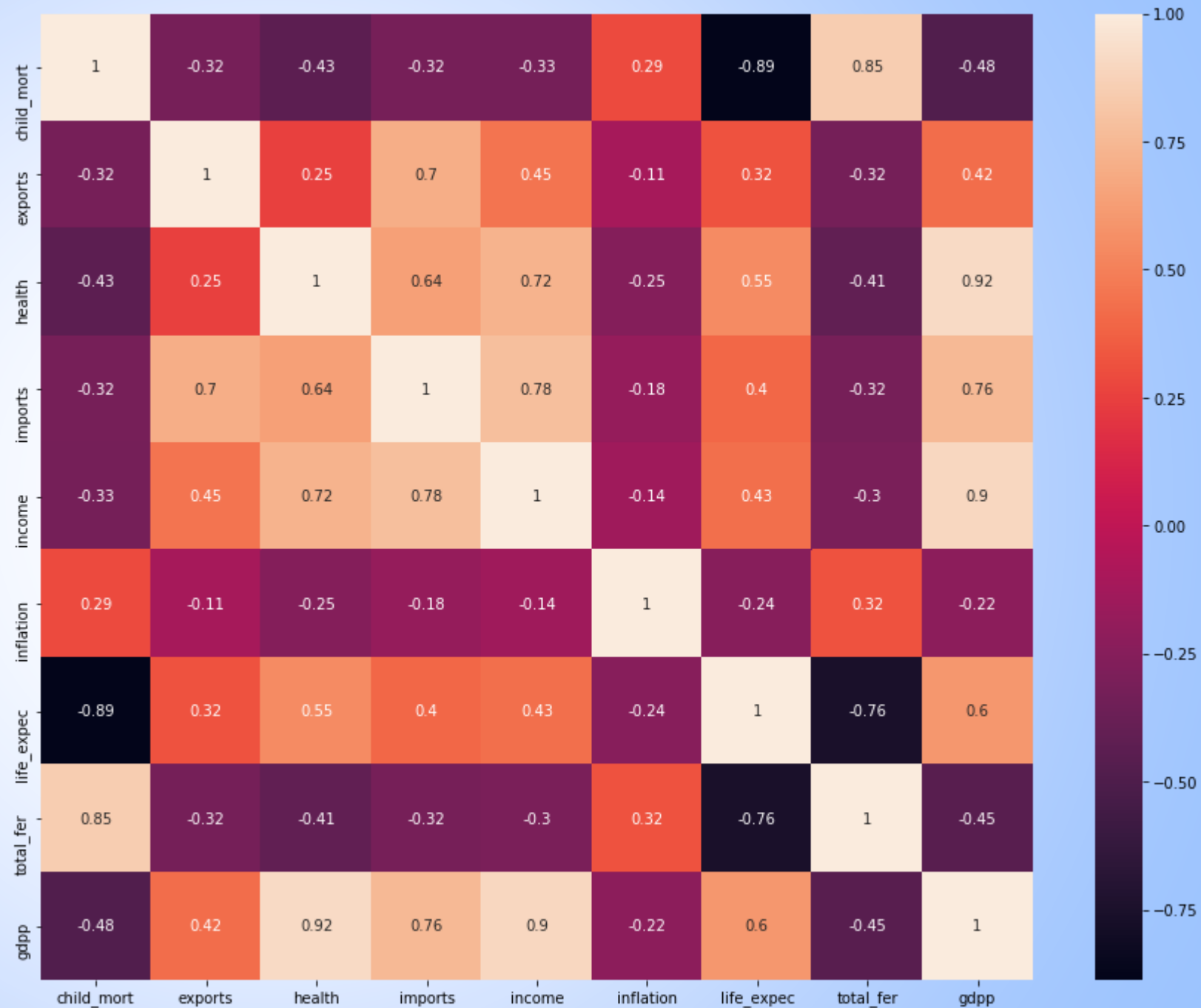
Problem Statement


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.

Data Visualization

Plotting a heatmap to find correlation among attributes.





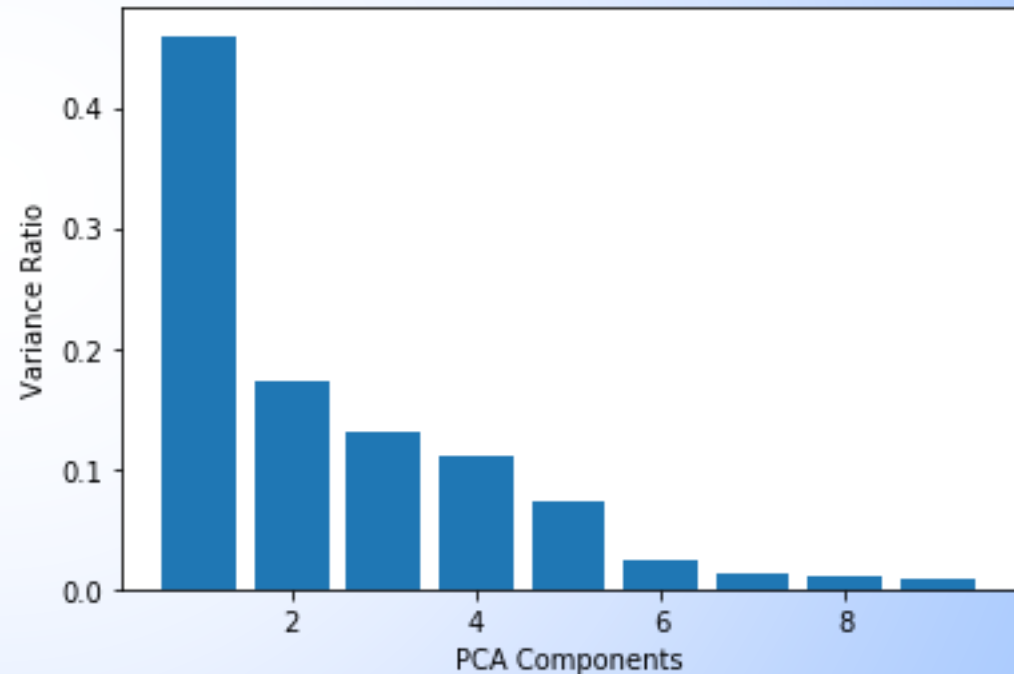
From the above heatmap, we can deduce that there are some attribute that have high correlation among them. We have the option to drop those attributes. But this way, we may lose some information that tends to be very important for our analysis.

So, lets do PCA (Principal Component Analysis) to handle this issue of multicollinearity in the data. Doing this, the multicollinearity can be properly handled without losing or compromising any important information.

Bar graph of Variance ratio for each PCA component

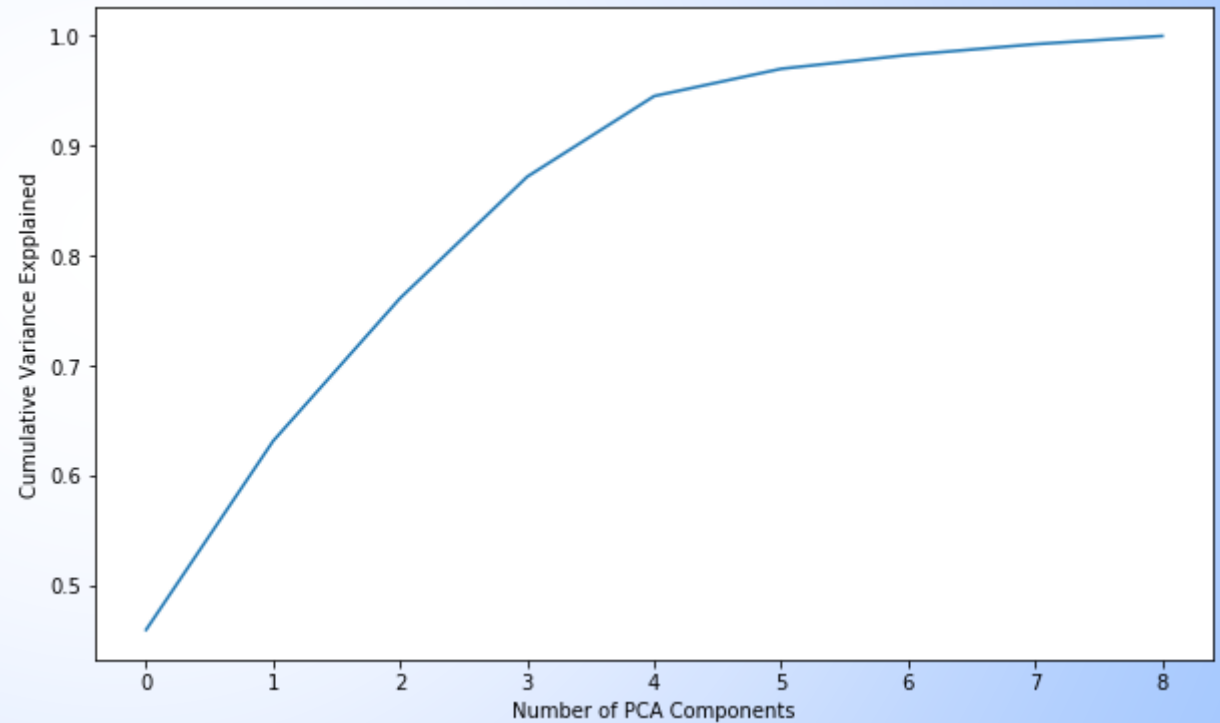
Inferences:

- First component itself explains more than 50% variance
- Second component, variance explained is almost 20%



Plotting Scree plot to visualize the cumulative variance against the number of components.

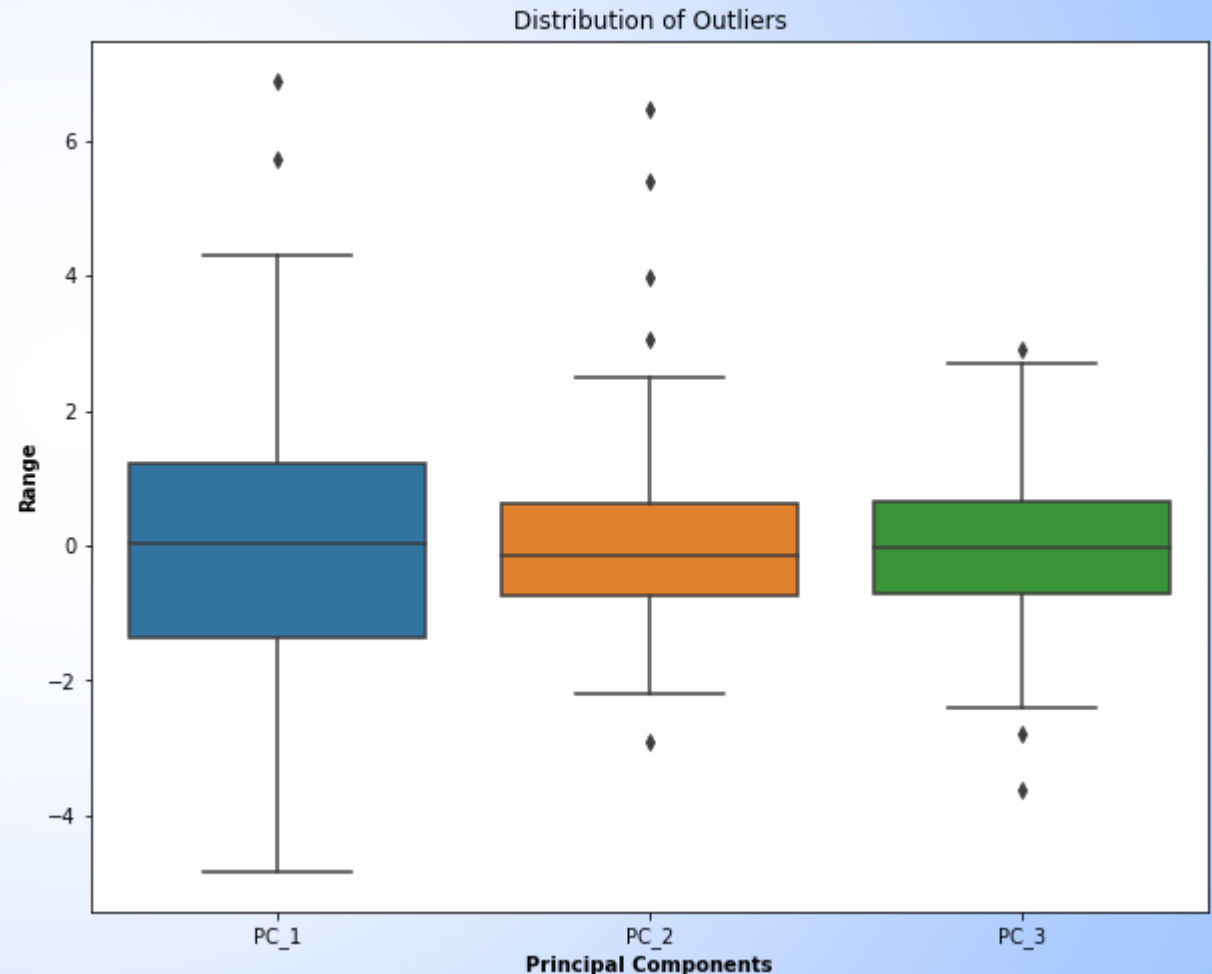
It is clear from scree plot that first 3 principal components are explaining more than 90% variance. So we will use only these three components for clustering the data.



Outlier Analysis:

We certainly can find few outliers in the data, from the above boxplot. Since we are working to find the countries that are in dire need of aid, we are not ready to lose any kind of data and would like to analyze as it is.

So, considering the business need, we are not removing any outliers here.

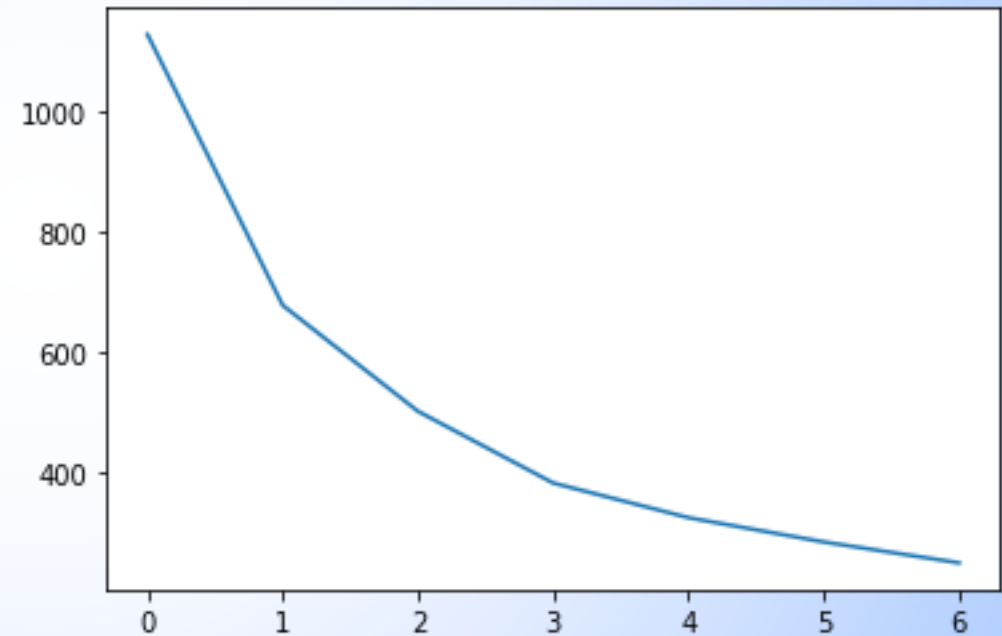




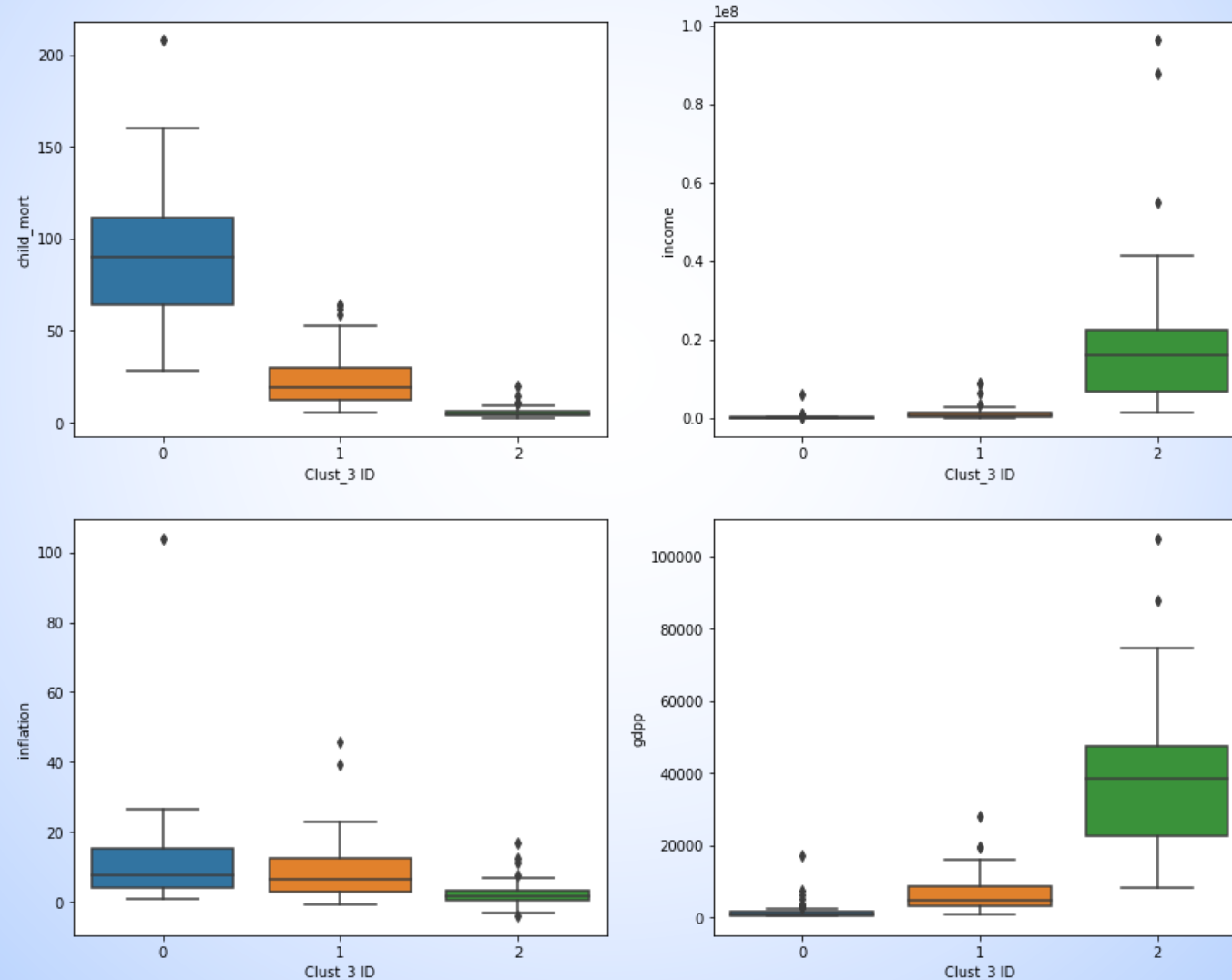
Model Building using K-Means Clustering

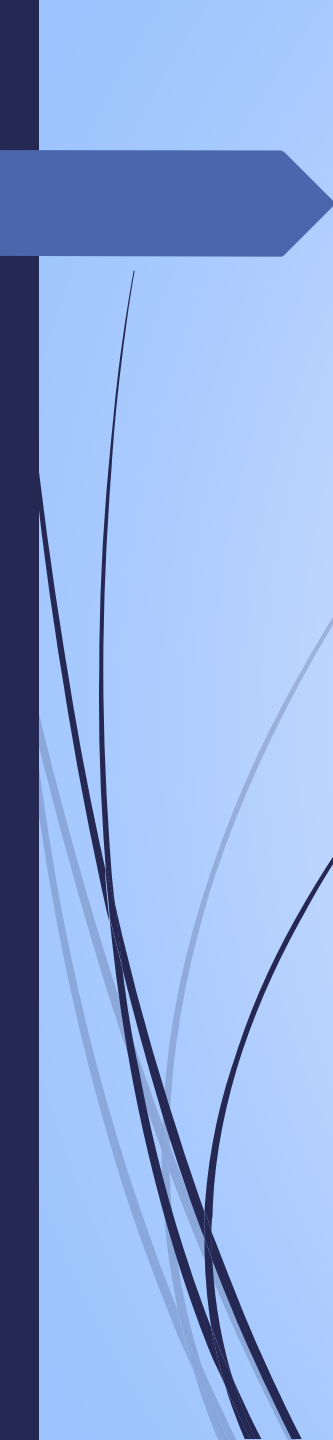
Elbow Curve

From the elbow curve, we can consider 3 or 4 as the initial value of K i.e., initial number of clusters.



Plotting boxplot on attributes "child_mort",
"income", "inflation" and "gdpp" to visualize the data
spread



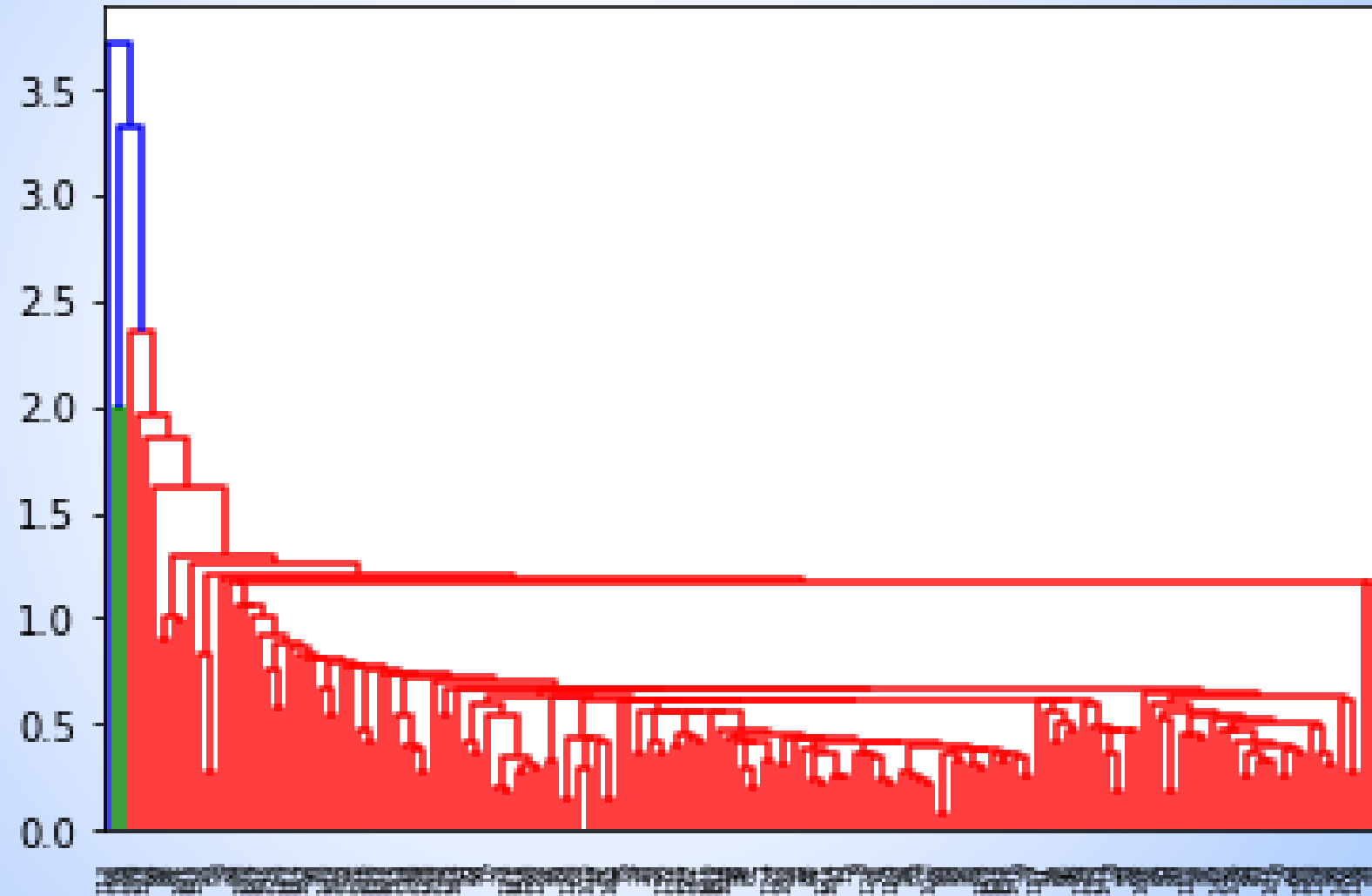
- 
1. For Cluster 0: high child mortality rate, low income, comparatively higher inflation than the countries in other clusters and very low gdpp. These clearly prove that these countries are under-developed and need help.
 2. For Cluster 1: all the four attributes- child mortality, income, inflation and gdpp are moderate. So these countries are not much of a concern to provide help.
 3. For Cluster 2: the countries falling under this cluster have very low child mortality rate, good income, lesser inflation and very good gdpp. These clearly are developed countries. They do not require help from NGO.

Hence we are going to consider those countries that are falling under Cluster 0 to extend help from NGO.

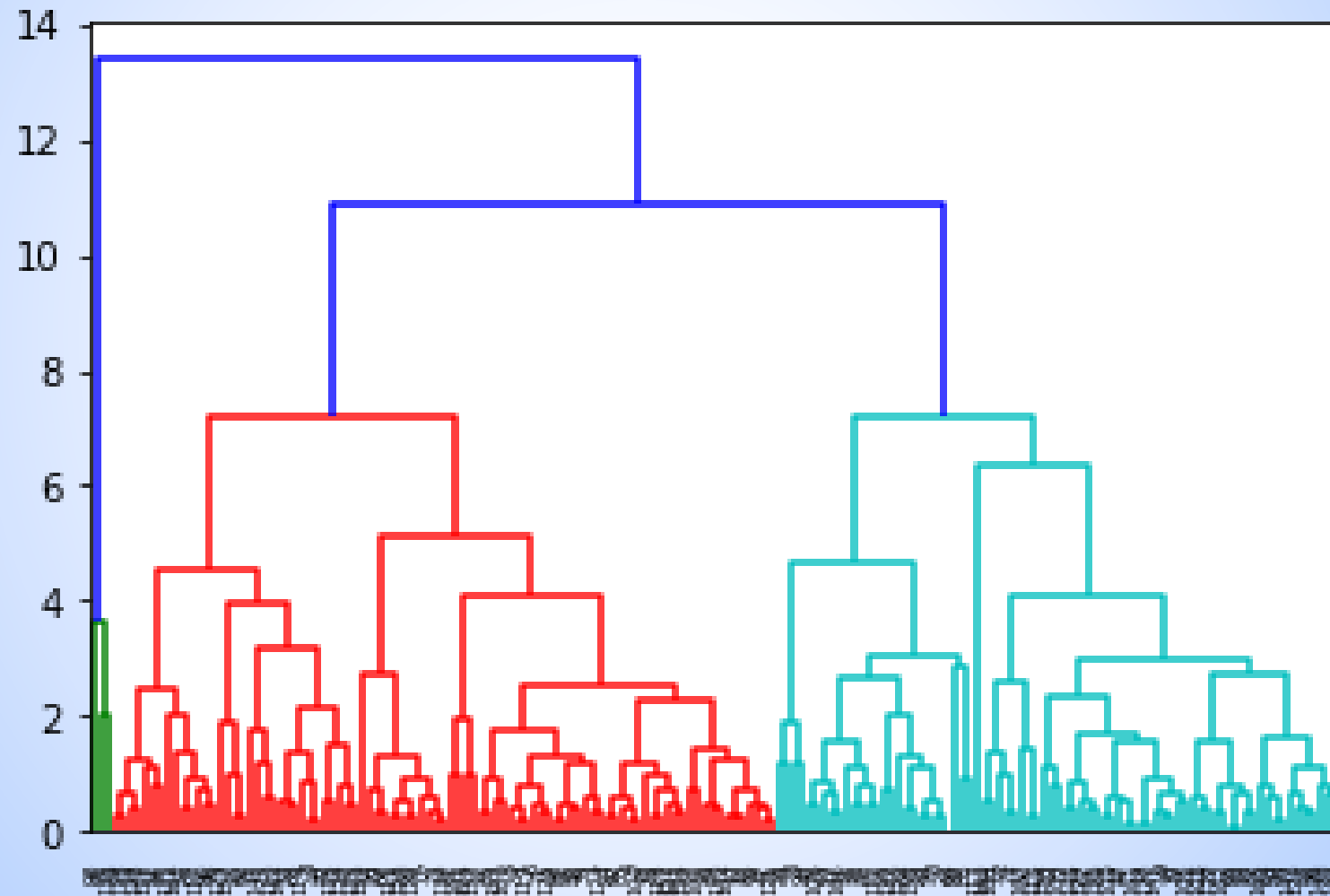


Model Building using Hierarchical Clustering

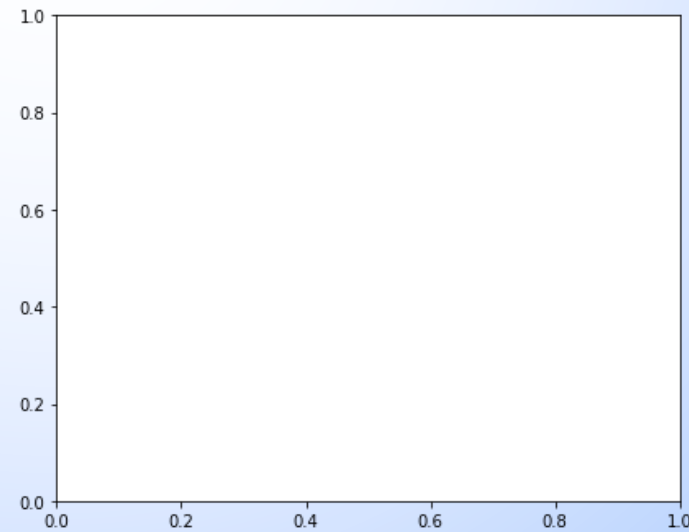
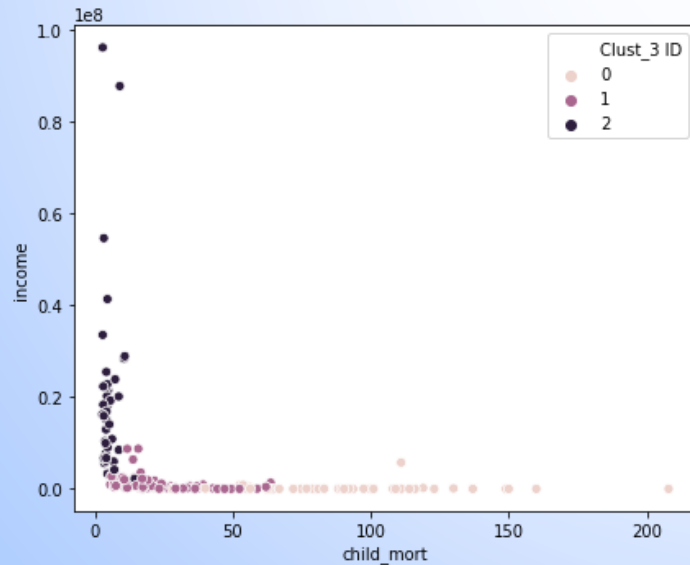
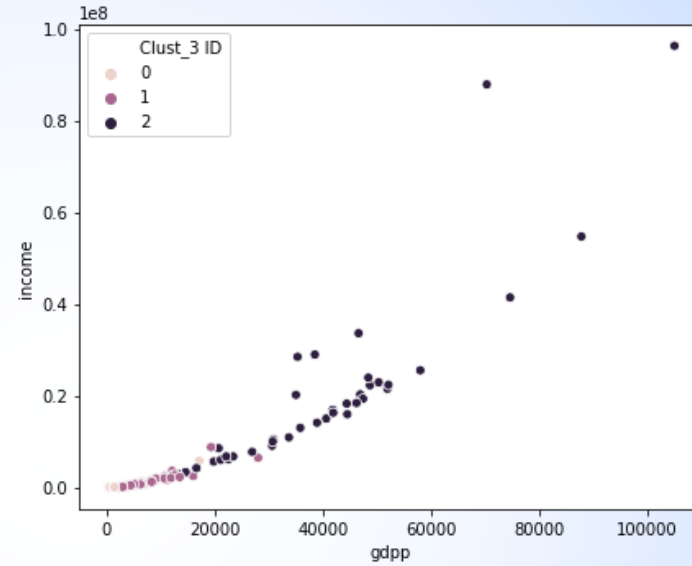
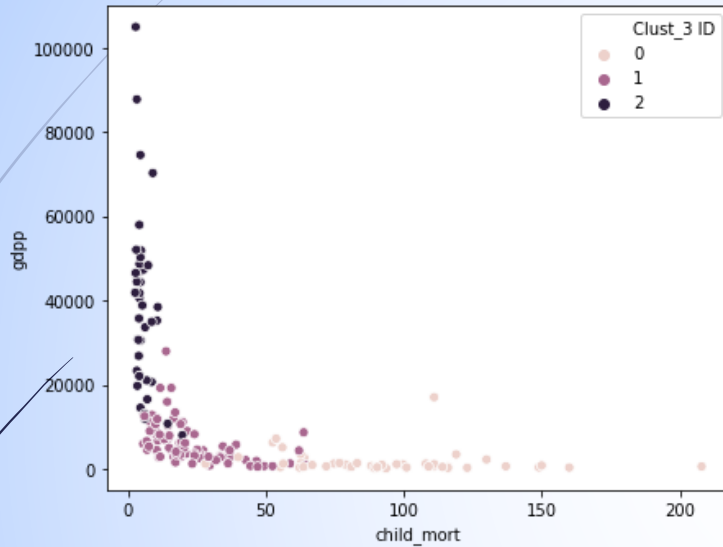
Single Linkage



Complete Linkage

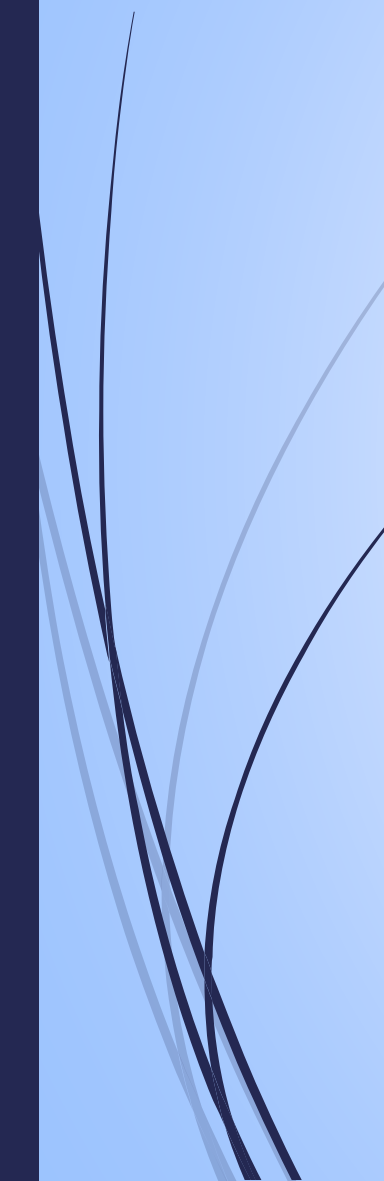


Plotting scatter plot on original attributes to visualize the spread of data

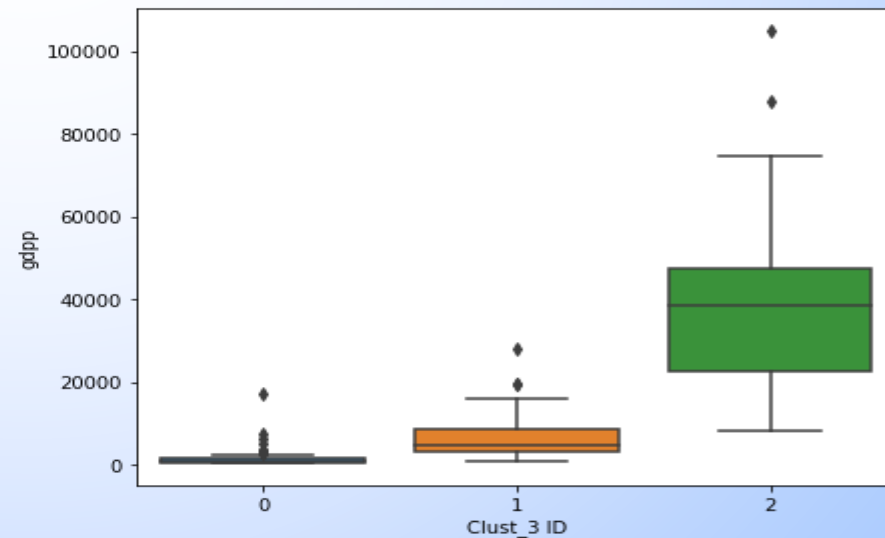
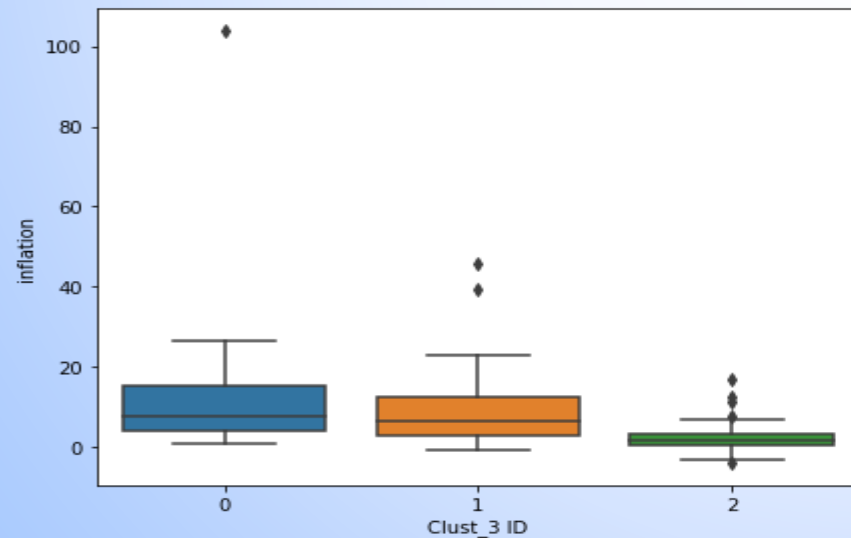
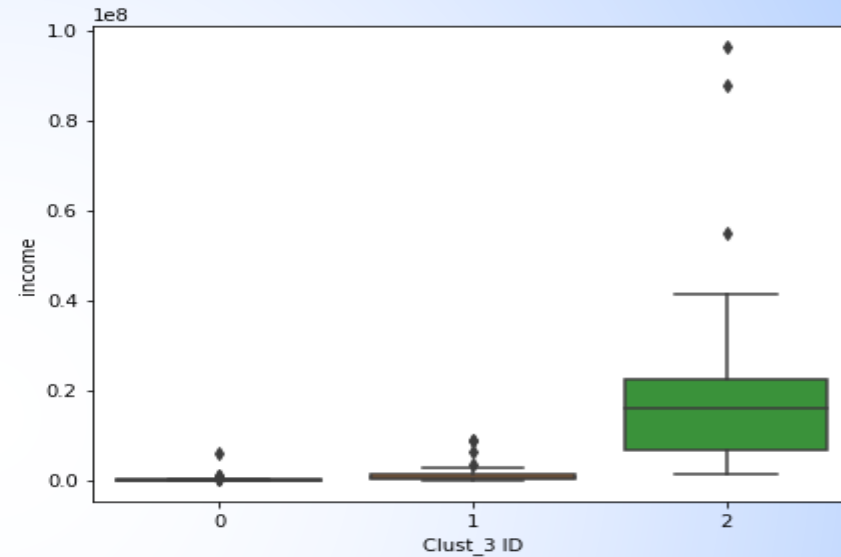
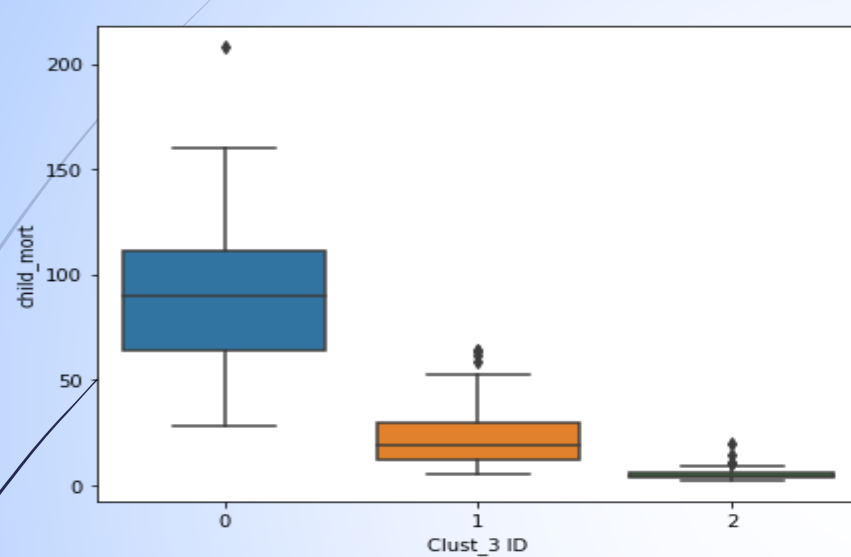




Inferences:

- Clusters formed are almost non-overlapping.
 - It is clear from above scatter plots that as the gdpp increases, the child mortality rate is decreasing.
 - The countries with higher income observe lower child mortalities
 - income is increasing with gdpp and those countries with higher gdpp have higher values of income.
- 

Plotting boxplot on attributes "child_mort", "income", "inflation" and "gdpp"






Inferences:

From the above boxplot we see that the results of Hierarchical clustering are almost similar to those observed with K-Means clustering algorithm.

One major difference we observed between results of hierarchical clustering and K-means clustering is that the "income" and "gdpp" are well clustered/ explained using k-Means clustering algorithm than hierarchical clustering. So let us consider K-Means clustering algorithm in this analysis and do our final analysis to identify the countries that are in the direst need of aid.



Final Analysis



We got countries that are in need of aid in the Cluster 0. Let us now make a list of those countries by considering the “mean” values of attributes “child mortality”, “income”, “GDP per capita” to make our final list of countries that are in need of aid.

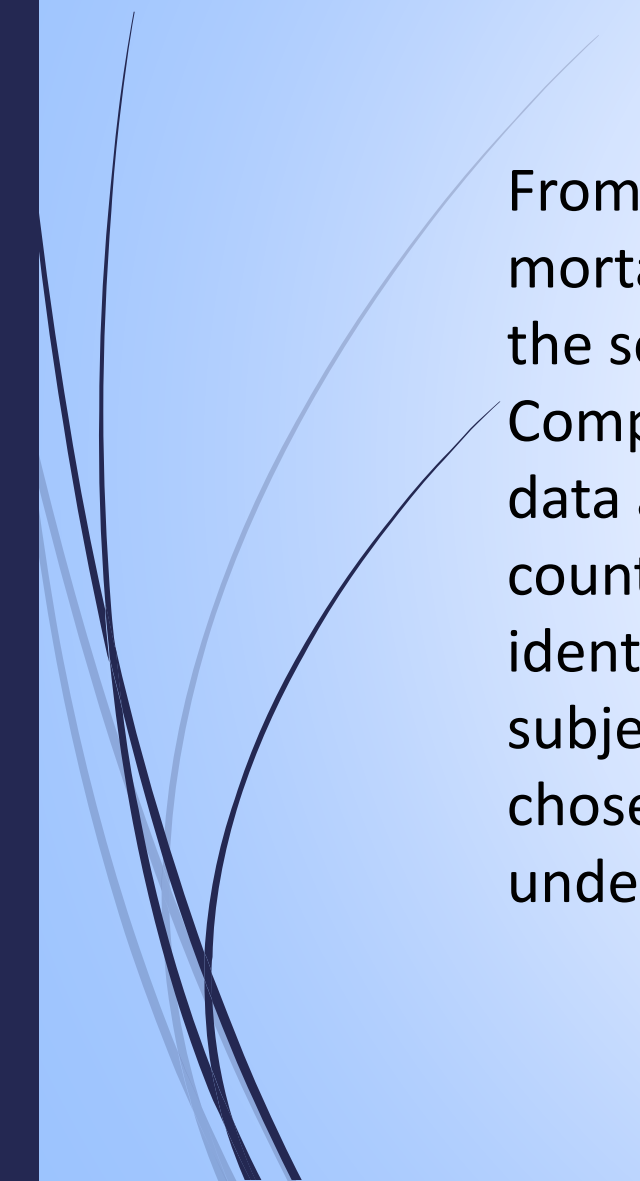


Following is the final list of Countries that are in direst need of aid, based on socio-economic standards.

1. Benin
 2. Burkina Faso
 3. Burundi
 4. Central African Republic
 5. Congo, Dem. Rep.
 6. Guinea
 7. Guinea-Bissau
 8. Haiti
 9. Mali
 10. Mozambique
 11. Niger
 12. Sierra Leone
- 



Conclusion



From the business understanding, we identified few factors like child mortality, GDP Per capita, income, etc. as vital attributes to determine the socio-economic standard of each country. We used Principal Component Analysis (PCA) to handle multicollinearity present in our data and also to reduce the dimensions, followed by clustering the countries based on their development status. We succeeded in identifying 12 countries that are in direct need of aid. This list is subject to changes as it is based on few factors like number of clusters chosen, number of attributes considered based on business understanding, the clustering algorithm used to build the model.