Assignment on K Means and Hierarchical Clustering

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. After the recent funding programmes, 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.

BUSINESS OBJECTIVE:

To categorise the countries that are in the need of aid based on some social, economic and health factors that determine the overall development of the country.

Data Cleaning : Missing Values and Outliers

- As a first step, the missing values were identified and it was found that there were no NAN values in the dataset
- In order to have a true picture about the country's development , we have converted the variables such as exports, imports and health which were given as percentage of gdpp into absolute value
- Further, the box plot was used to identify the outliers since outliers have a serious impact on the performance of the algorithm and it could prevent optimal clustering.

The box plot analysis indicates the presence of outliers in almost all factors:

1) For variables such as Inflation and Child Mortality we have not treated the high range outliers since by treating high range outliers we may loose some important information about those countries which might have high inflation and high child mortality rate.

2) For all other columns we have treated only the upper range outliers since all these columns indicates the good health of the economy and by treating the lower end outliers, we may end up loosing some information about the countries that are in the need of aid.

Thus we have capped the outliers for reaching optimal clustering.

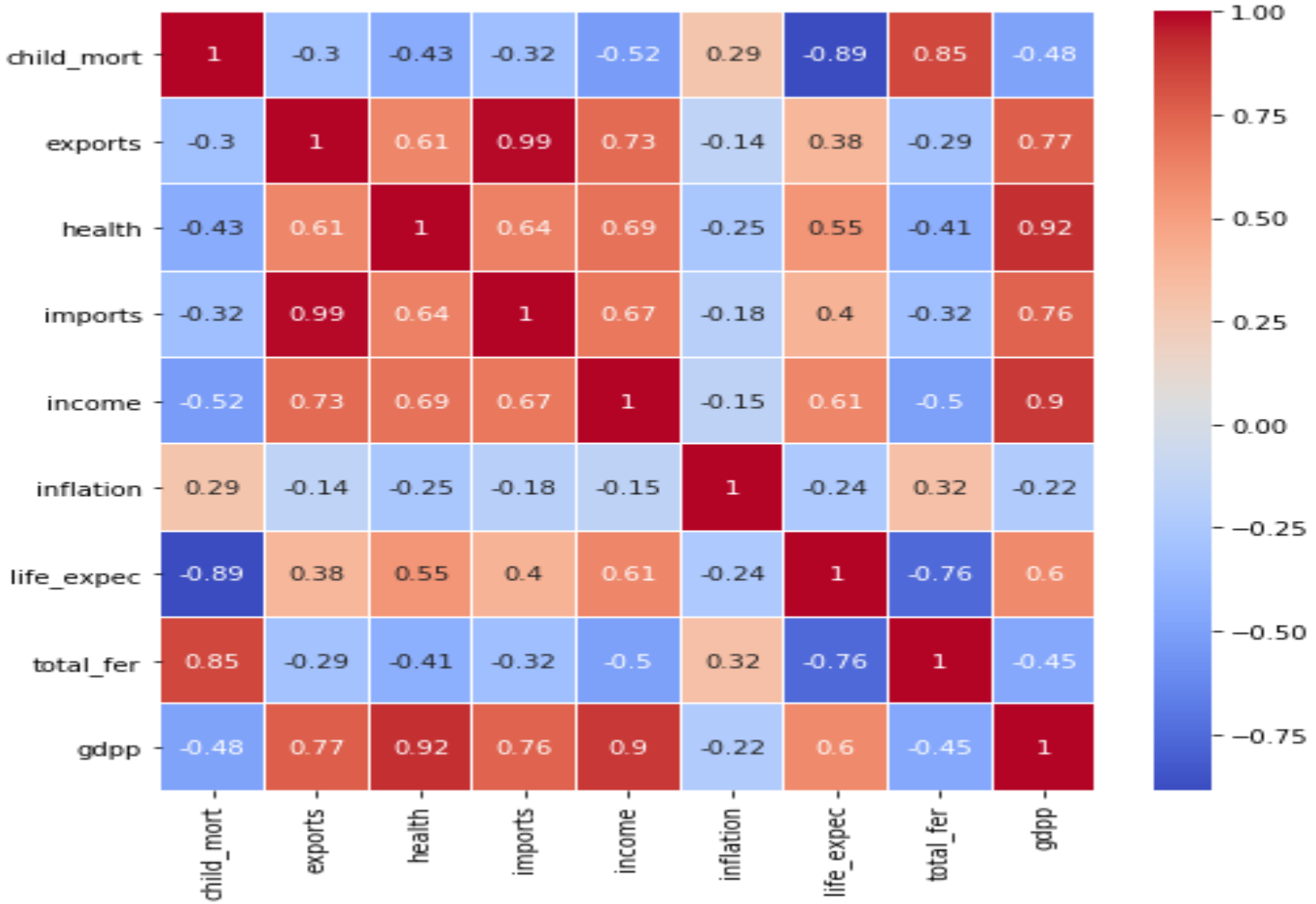
UNIVARIATE AND BIVARIATE ANALYSIS

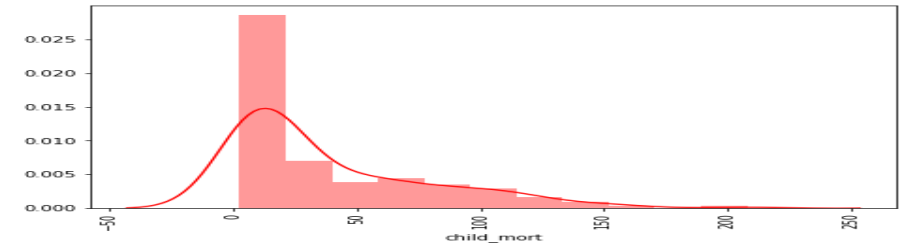
The heat map indicates :

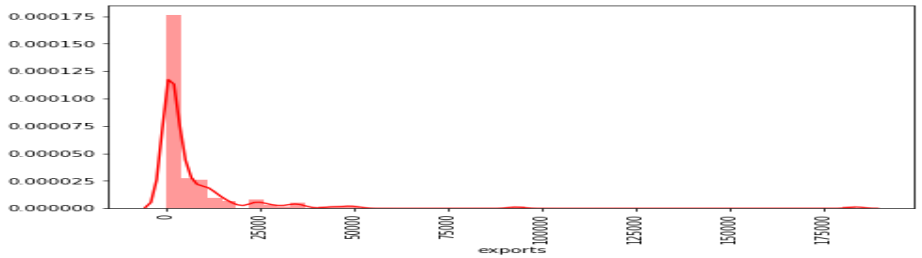
1.High positive correlation between income and GDP per capita.

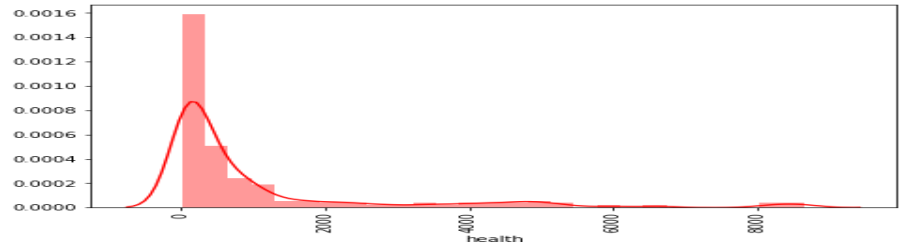
2.High positive correlation between health expenditure and GDP per capita.

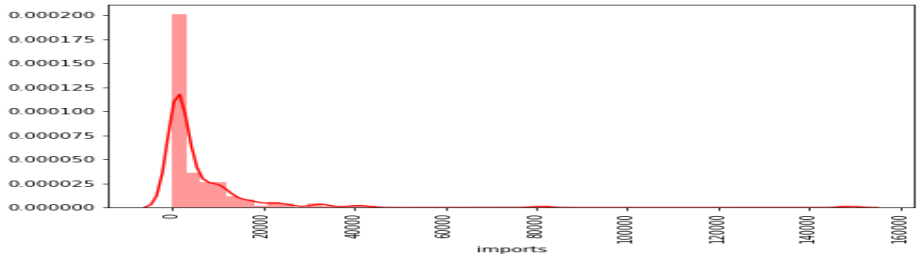
3.High negative correlation between life expectancy and child mortality

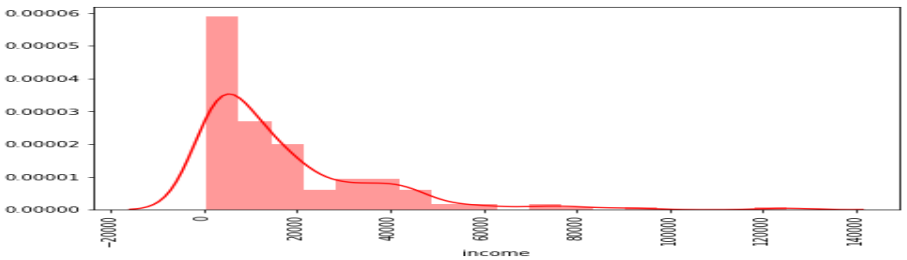


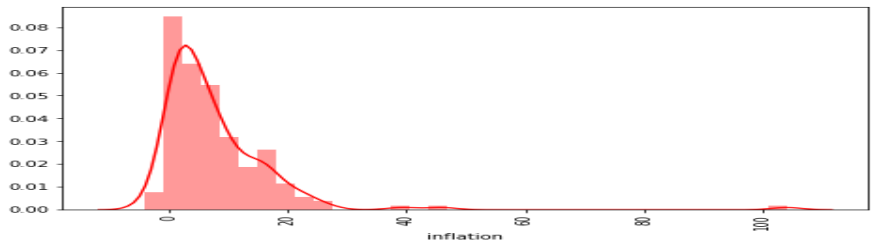


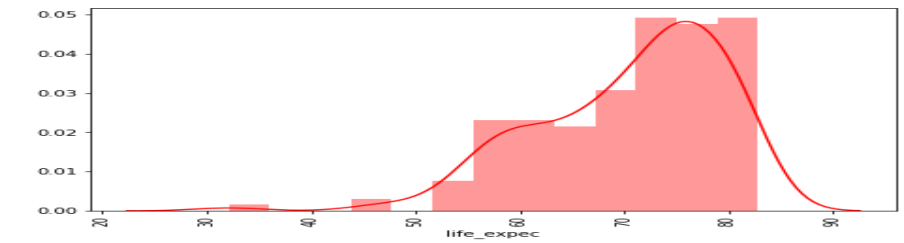


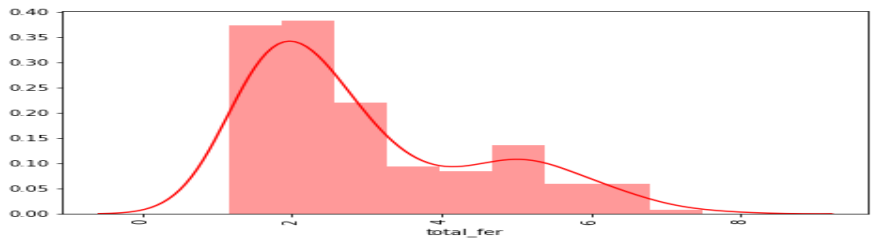


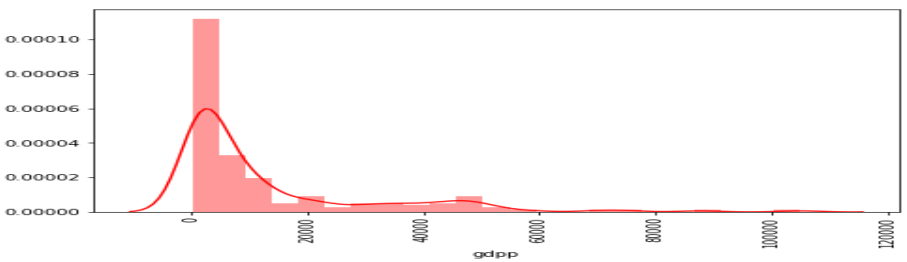












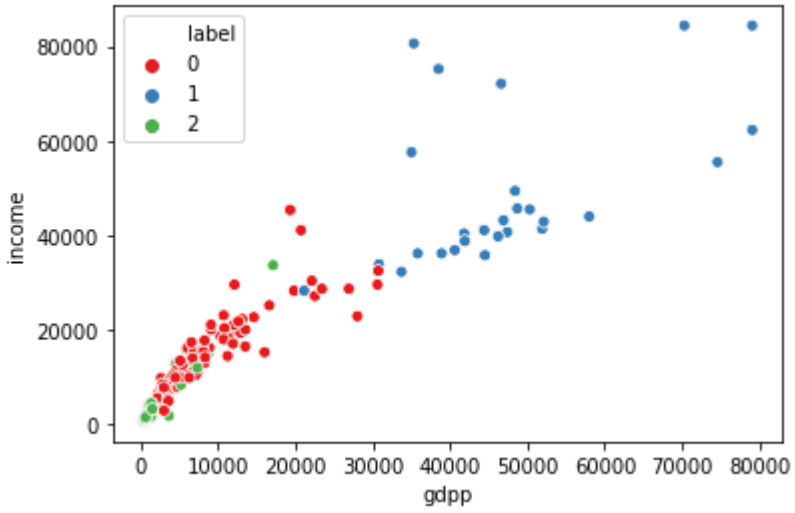
GDP per capita and income followed same pattern . The are lot of countries those gdp per capita and income falls between the range of 0-20000. Similarly Child mortality follow the same distribution.

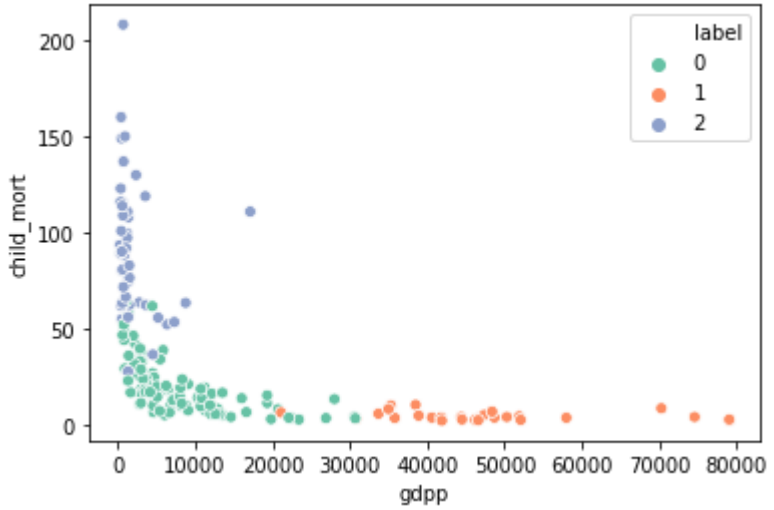
Preparing the data for modeling

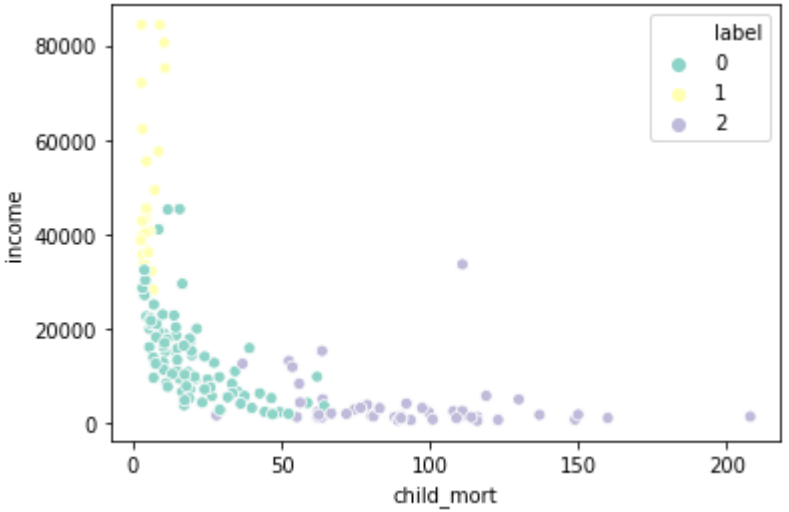
- Standardized Scaling : In order to avoid the out-weighting the attributes with smaller range, standardized scaling was done to scale down all attributes to the same normal scale.
- The Hopkins statistic was performed to check the cluster tendency of a data set.It acts as a statistical hypothesis test where the null hypothesis is that the data is generated by a Poisson point process and are thus uniformly randomly distributed. In our analysis, the Hopkins score was 0.88.

K-MEANS ANALYSIS

- The Elbow curve and Silhouette analysis indicates that the we could initiate the clustering with k as 3 .
- K means has classified the different countries into 3 clusters based on socio-economic features. The analysis illustrates that most of the countries fall under cluster 1 (0)
- *From the business understanding we have learnt that Child_Mortality, Income, Gdpp are some important factors which decides the development of any country. Therefore, we are proceeding with the clustering analysis with these three variables.*

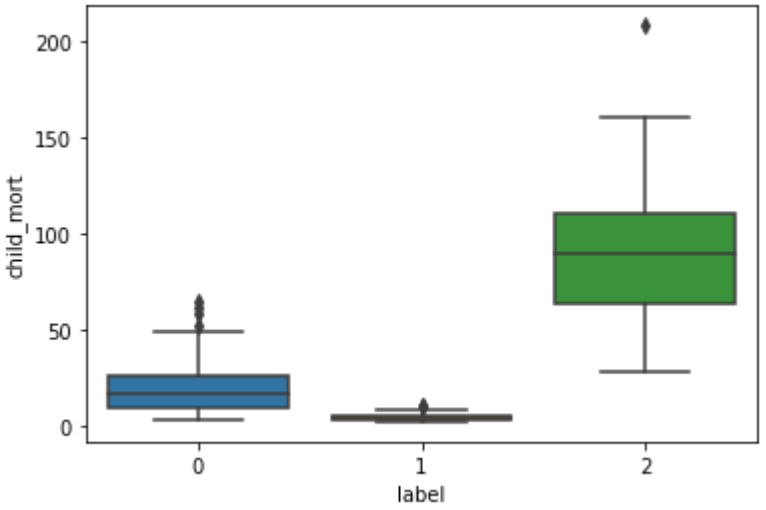




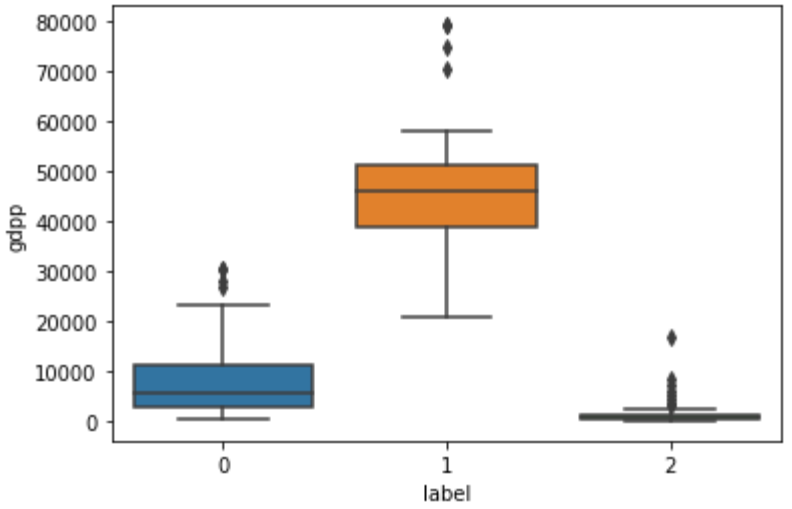


K-MEANS ANALYSIS

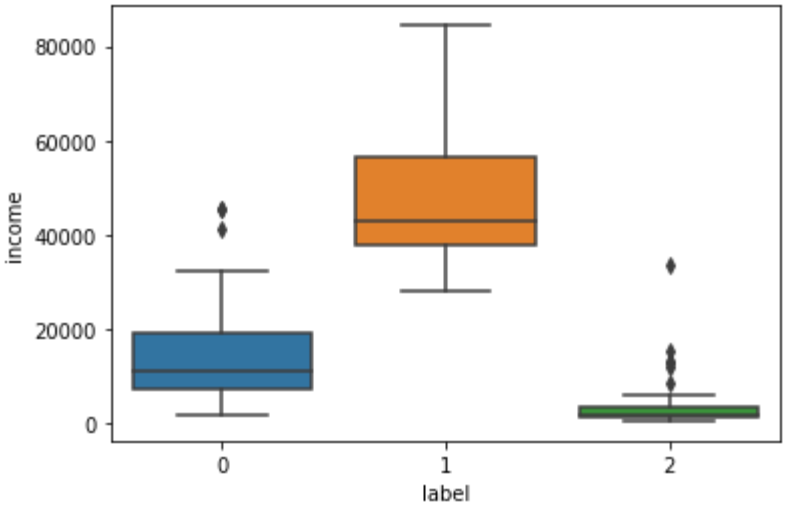
The Box plot and Bar analysis clearly indicates that the child mortality rate is higher in cluster 3 (labeled as 2). Income and Gdpp are measures of development. Higher the per capita income and gdpp better is the country's development. Income per capita and gdpp seems lowest for countries in cluster 3.It reflects that countries in cluster 3 need an urgent aid.



Cluster 3 : High Child mortality rate

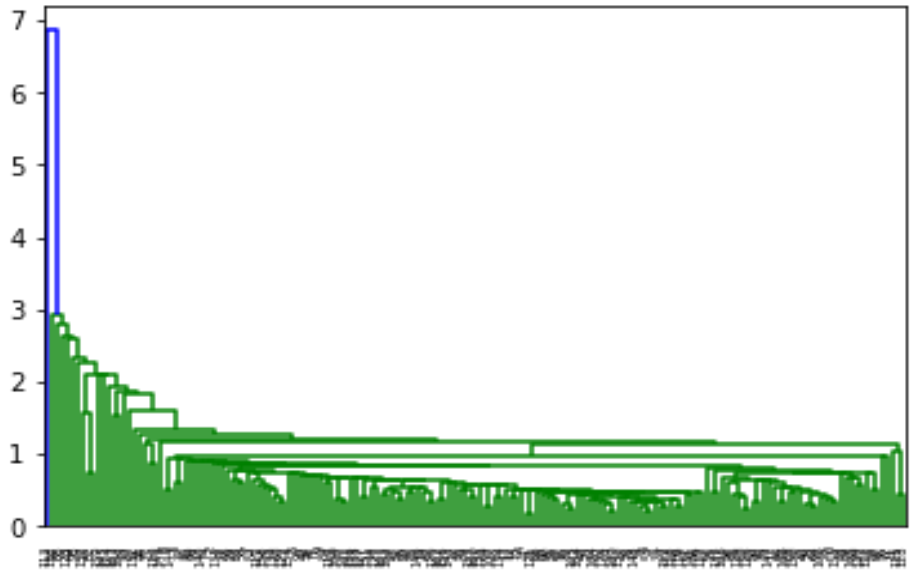


Cluster 3:Low GDPP

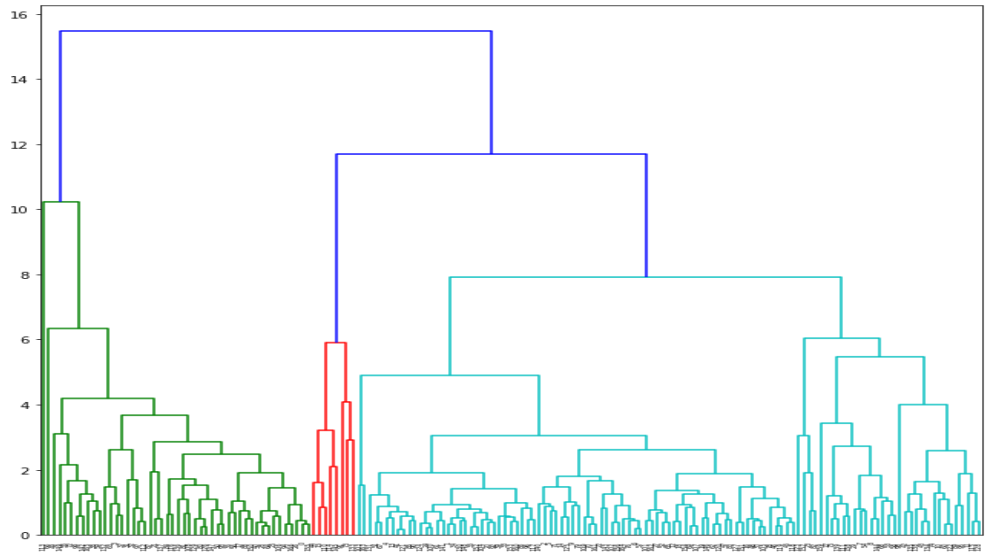


Cluster 3: Low Income

Hiereachical Clustering



Single Dendrogram

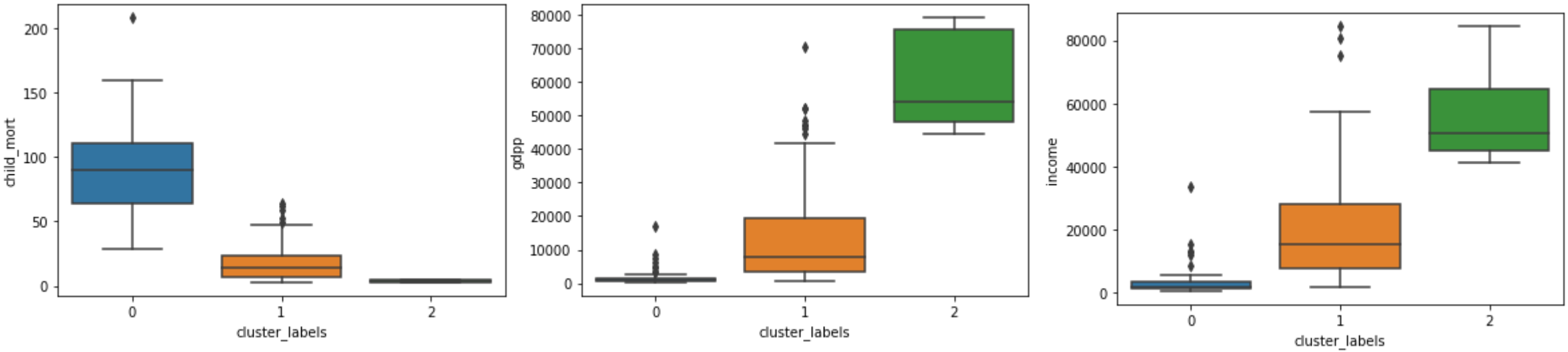


Complete Dendrogram

We have cut the tree at height of approx. 11 to get 3 clusters and consider the three important variables i.e GDP per capita , Income and Child mortality rate for clustering analysis.

BIVARIATE ANALYSIS AFTER MERGING

Based on this graph (below) we can infer that countries in cluster 1 (labeled as 0) in need of aid since they have low GDP and Income and high child mortality rate. Therefore, the CEO should focus on countries in cluster 1.



Final Analysis

Thus we have performed both K-means and Hierarchical clustering and found that the clusters formed in both the cases are same . However, we will proceed with the clusters formed by hierarchical clustering since we don't need to pre determine the value of k in this case and further hierarchal clustering is suitable for most of the distributions.

Thus based on the factors i.e GDP per capita, child mortality, and income which plays a vital role in deciding the development status of the country , we derived at the 3 clusters.

Based on those clusters we have identified the following countries which are in dire need of aid.

1. Haiti
2. Sierra leone
3. Chad
4. Central African Republic
5. Mali

THANK YOU