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.

So, the problem in hand is to categorize the countries using some socio-economic and health factors that determine the overall development of the country and to suggest the countries which the CEO needs to focus on the most.

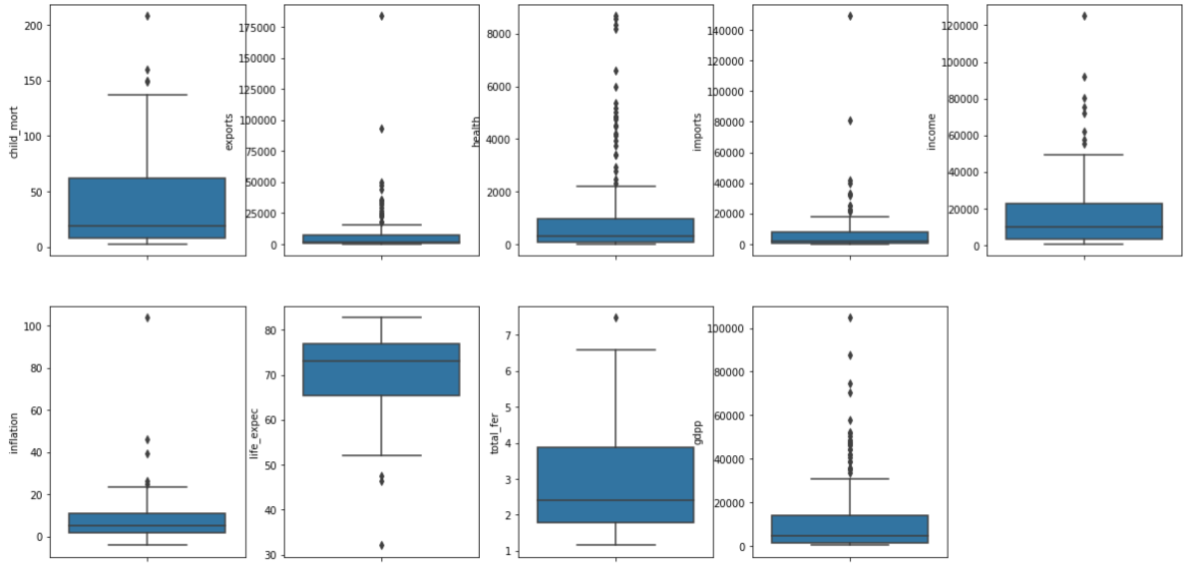
Step by Step Approach

Step1: Reading and understanding the data

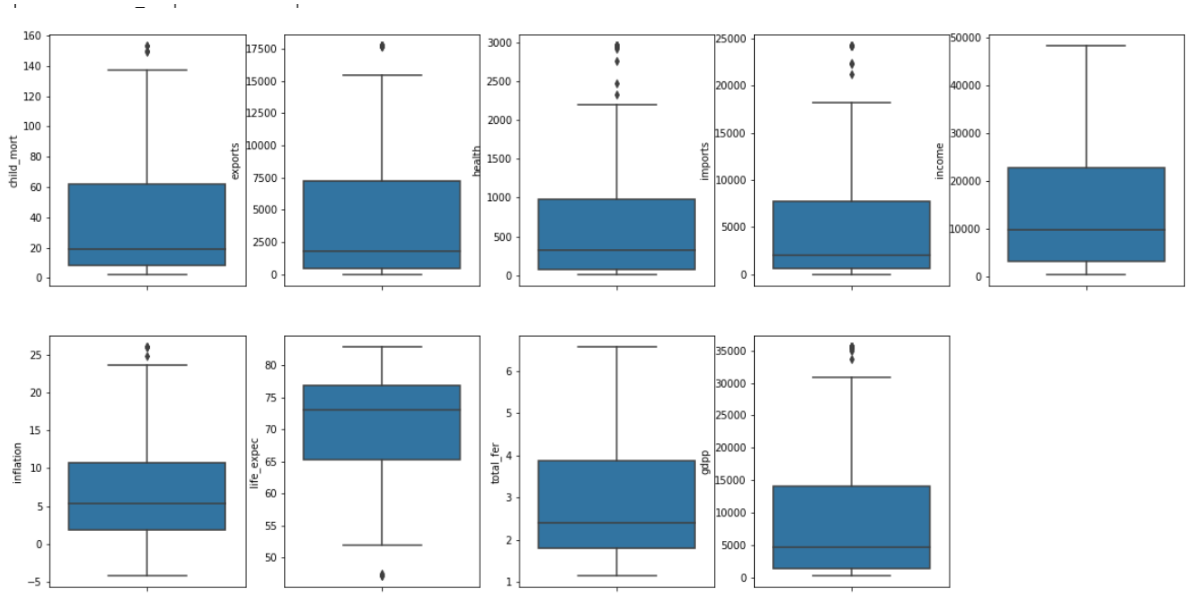
The first step is to understand the dataset and its columns by referring to its data dictionary. Then we need to read it into the notebook and inspect it. In the data dictionary, we can see that the columns “exports”, “health”, “imports” are expressed as percentages of “gdpp”. So, before going ahead, these 3 columns have been converted into normal values. Then, dataset is inspected for any missing values. Since there are no missing values in the data set, the dataset is inspected for any outliers.

Step2: EDA

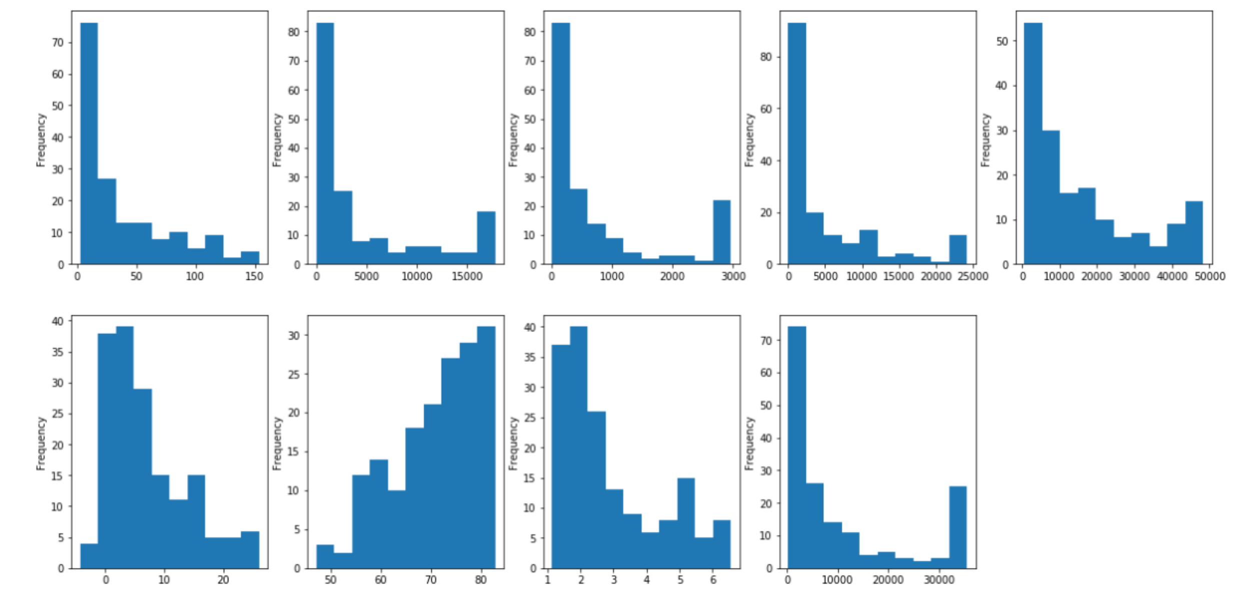
For outlier analysis, using describe function, the maximum, minimum and other values have been inspected for all the columns. Then a box plot for each column have been plotted. The following image shows the box plots for all the columns.



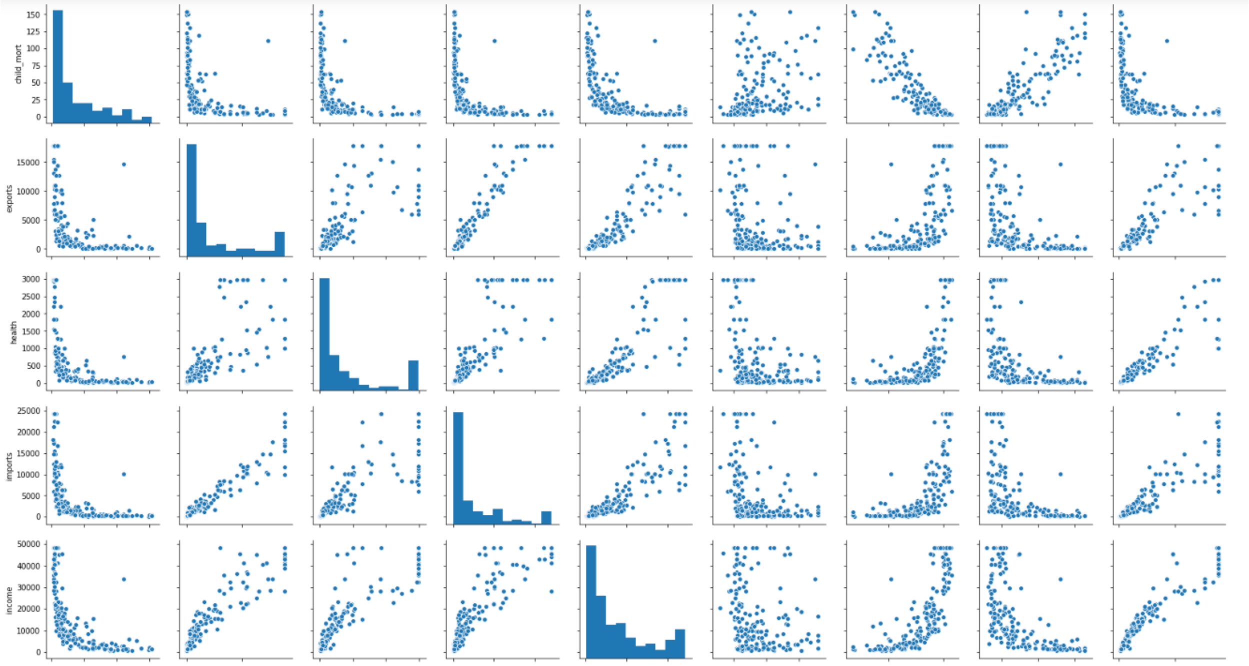
By looking at the above image, all the columns have outliers. All of them have positive outliers except "life_expec" column. "exports", "health", "imports", "income", "gdpp" have significant number of outliers. Deleting the outliers by simply removing them from dataset gets rid of too many countries. Especially for "life_expec", which have negative outliers. Our target is to find the countries that are in direst need, getting rid of the countries that have very less life expectancy gets rid of our target countries. So, the best way to deal with these outliers is to cap them. Instead of setting a hard-set threshold for all the columns, a threshold value for each column has been set up by inspecting the quantile values for each column. Hence all the countries have been preserved while handling the outliers. The below image shows the box plots for all the columns after capping the outliers.

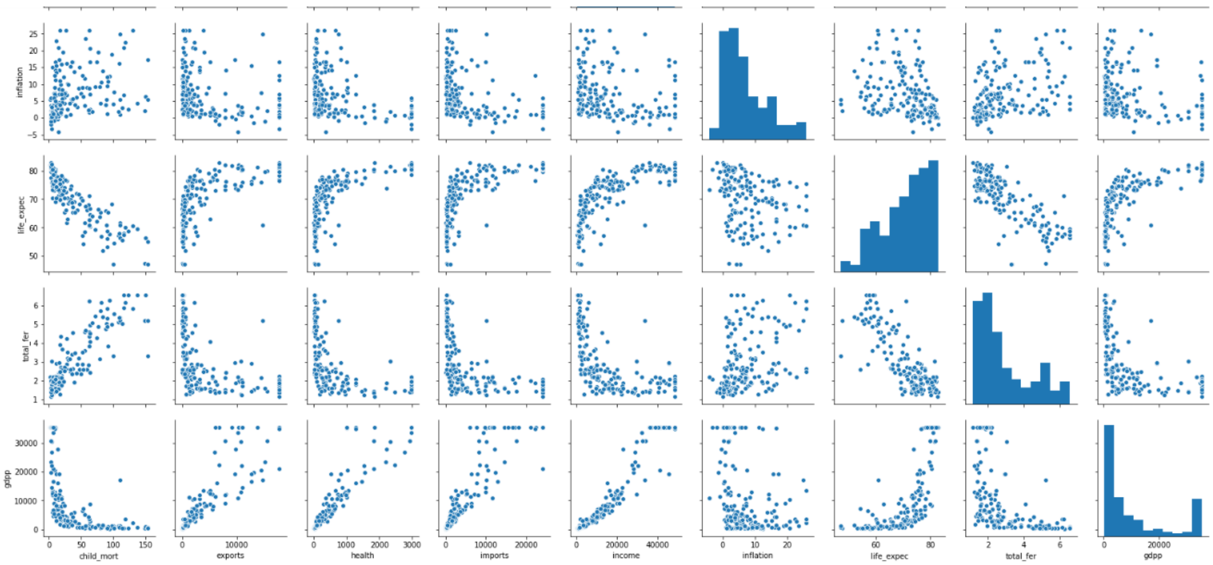


Before jumping into further analysis, Hopkins statistic has been calculated for the dataset and the value came out as 87.72 which is as expected. So, in the next steps of EDA, univariate and bivariate analysis has been performed on all the columns to understand how the data of each column behaves individually and relatively. For univariate analysis, all the columns have been plotted using histogram and for bivariate analysis, a pair plot has been used which shows the scatter plots for all combinations of columns.



Univariate Analysis

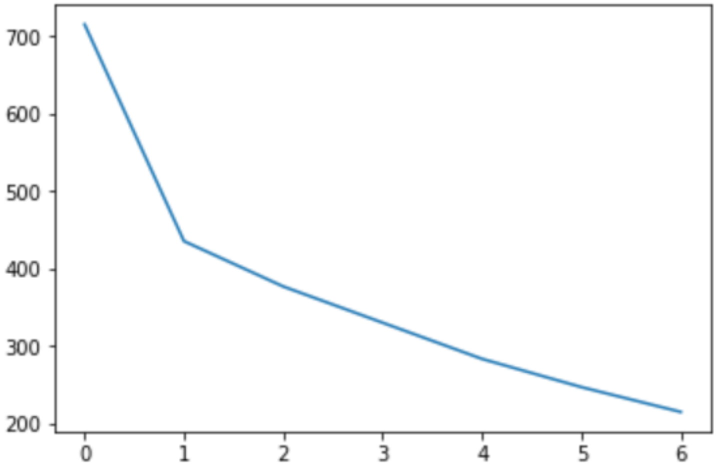




Bivariate Analysis

Step3: Clustering

The data is properly processed and is made ready to be clustered. Firstly, kmeans algorithm is used to cluster the data. Now, to find the number of clusters, elbow curve is plotted, and it shows the elbow bend at 3 clusters. Beyond 3, the inertial difference is not significant.

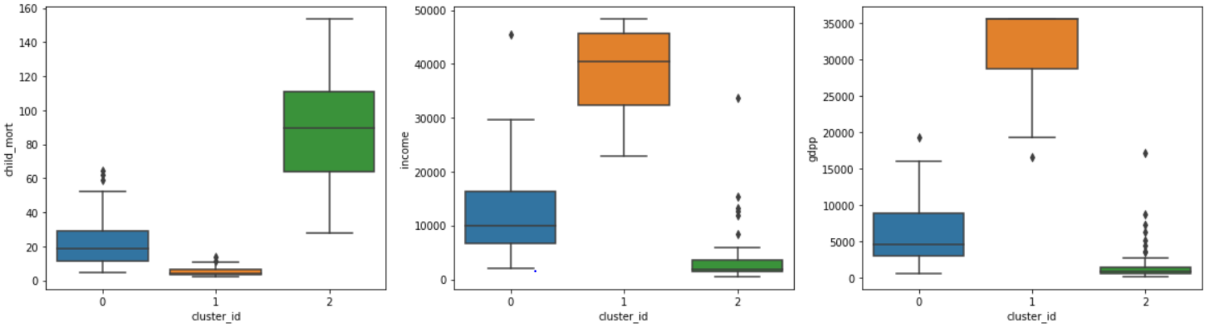


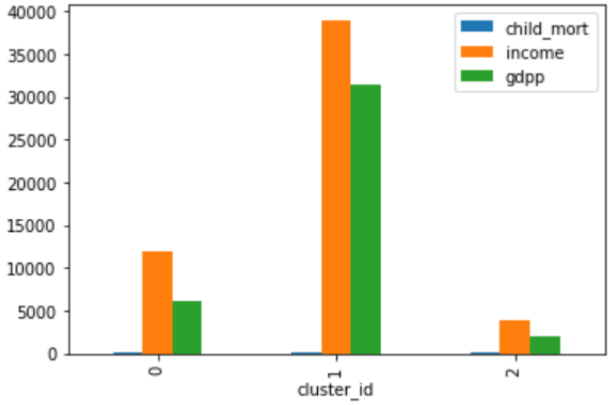
Elbow Curve

To verify this, silhouette score is also used to identify the ideal number of clusters. Silhouette score is calculated for a range of clusters from 2 to 8 and the score is maximum at 2. But the difference of score between 2 and 3 clusters is not huge. So, the data is clustered into 3 clusters.

Once the kmeans algorithm is fitted to the dataset, the labels have been attached to their respective columns of the original dataset as new column called “cluster_id”

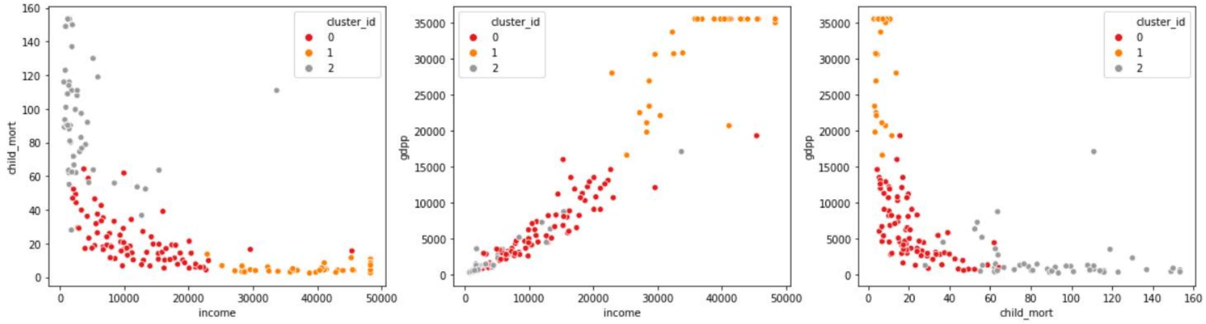
Now using this column, cluster profiling has been performed for “child_mort”,” income”, ”gdpp”. For this, both box plots and bar graphs have been used.



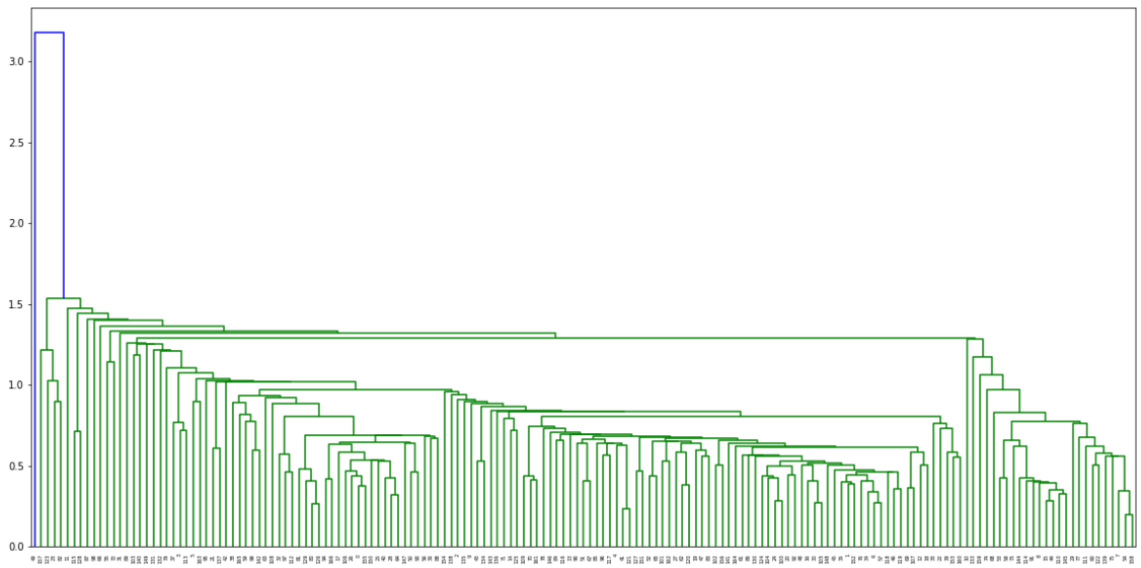


The above bar graph shows the mean of child_mort, income and gdpp for all three clusters. By looking at the above graphs, it is clear that the cluster 2 has the countries that are in direst need.

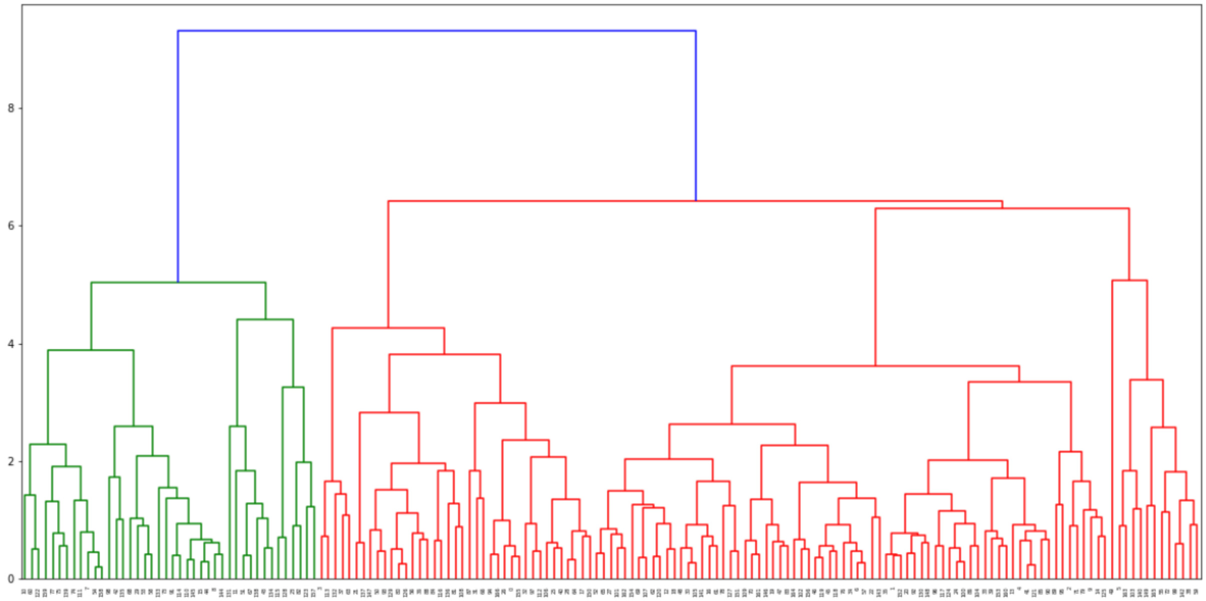
And finally, scatterplots have also been plotted for the target columns which shows the data cluster wise.



Now, the hierarchical clustering is used to perform the clustering operation. Initially, single linkage is used to plot the dendrogram. But, as it is not easy to interpret and cut, complete linkage is used to plot the dendrogram and this clearly shows the tree in a structured form which is easy to interpret and cut.

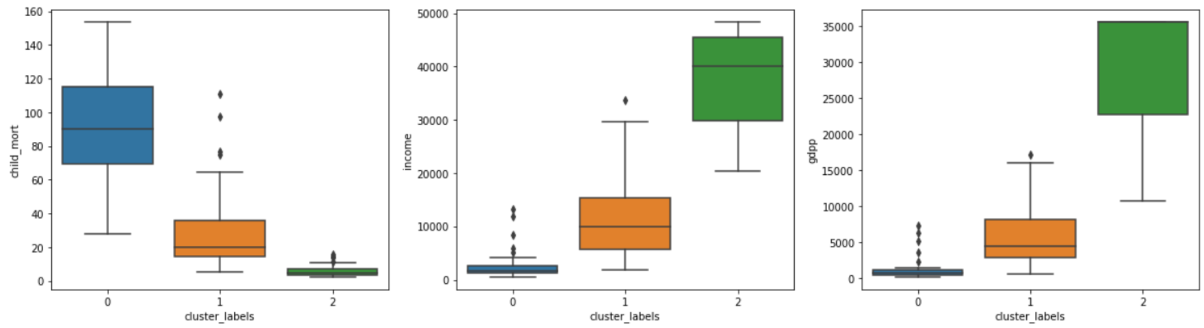


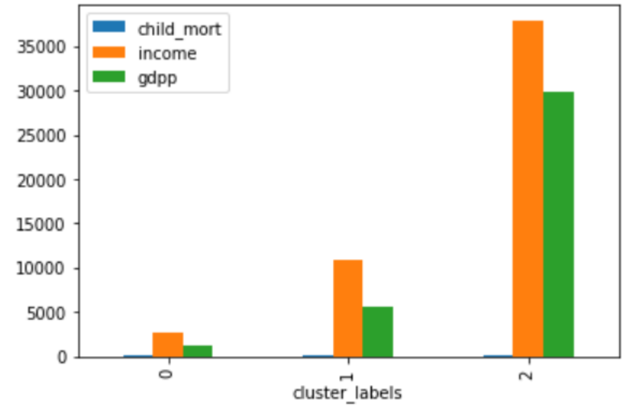
Dendrogram using single linkage



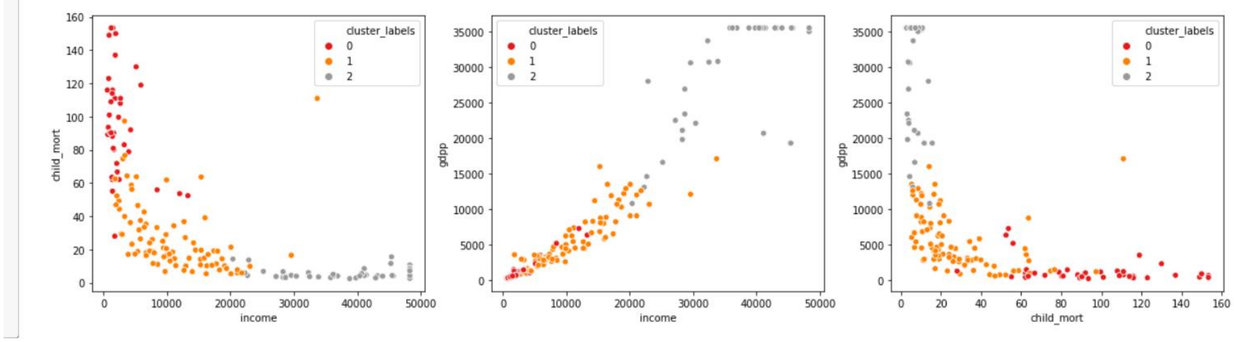
Dendrogram using complete linkage

Although, after looking at the dendrogram, it is easy to cut it into two clusters, I preferred 3 as it made the scatterplots of cluster wise data look better and easy to interpret. After cutting the tree into 3 clusters, I assigned the cluster labels to the original dataset as a new column called "cluster_labels". Then cluster profiling has been performed just as for the kmeans algorithm and bar plot and box plots are shown below.



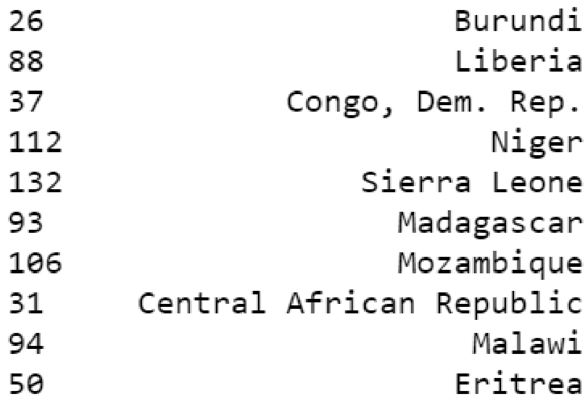


By looking at the above graphs, it is clear that the cluster 0 has countries that are in direst need. And finally, scatterplots have also been plotted for the target columns which shows the data cluster wise.



Step4: Finding the countries that are in direst need

Now in order to find the countries that are in direst need for the funding, the original dataset has been filtered to extract the countries with cluster_id as 2 (from KMeans clustering) and countries with cluster_labels as 0 (from hierarchical clustering) and both of these separate dataframes are individually sorted based on income, gdpp (should be low) and child_mort(should be high). Once these two data frames are sorted, the top 10 countries have been extracted and it is identified that both the algorithms gave same countries as output. And hence the CEO of HELP international should release the funding urgently for these below mentioned countries.



```
26                    Burundi
88                    Liberia
37          Congo, Dem. Rep.
112                     Niger
132              Sierra Leone
93                 Madagascar
106                Mozambique
31   Central African Republic
94                     Malawi
50                    Eritrea
```