CLUSTERING ASSIGNMENT

BY : MAYANK TUSHAR

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.

VISUALISING 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:

The clustering process is very sensitive to the presence of outliers in the data. In this case, specifically the outliers need not to be dropped at all. As all the countries are to be considered for evaluation. The extremities should be considered, if any, and may form a characteristic of cluster. Thus, soft capping was implemented.

SCALING DATA:

Standardizing all the continuous variables.

ANALYSIS APPROACH

HOPKINS TEST:

To check if data has tendency to form clusters.

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 analysing dendrogram. Then forming the cluster on scaled data and adding the cluster label to original data for better interpretation. Visualisation of clusters was also done.

DECISION MAKING:

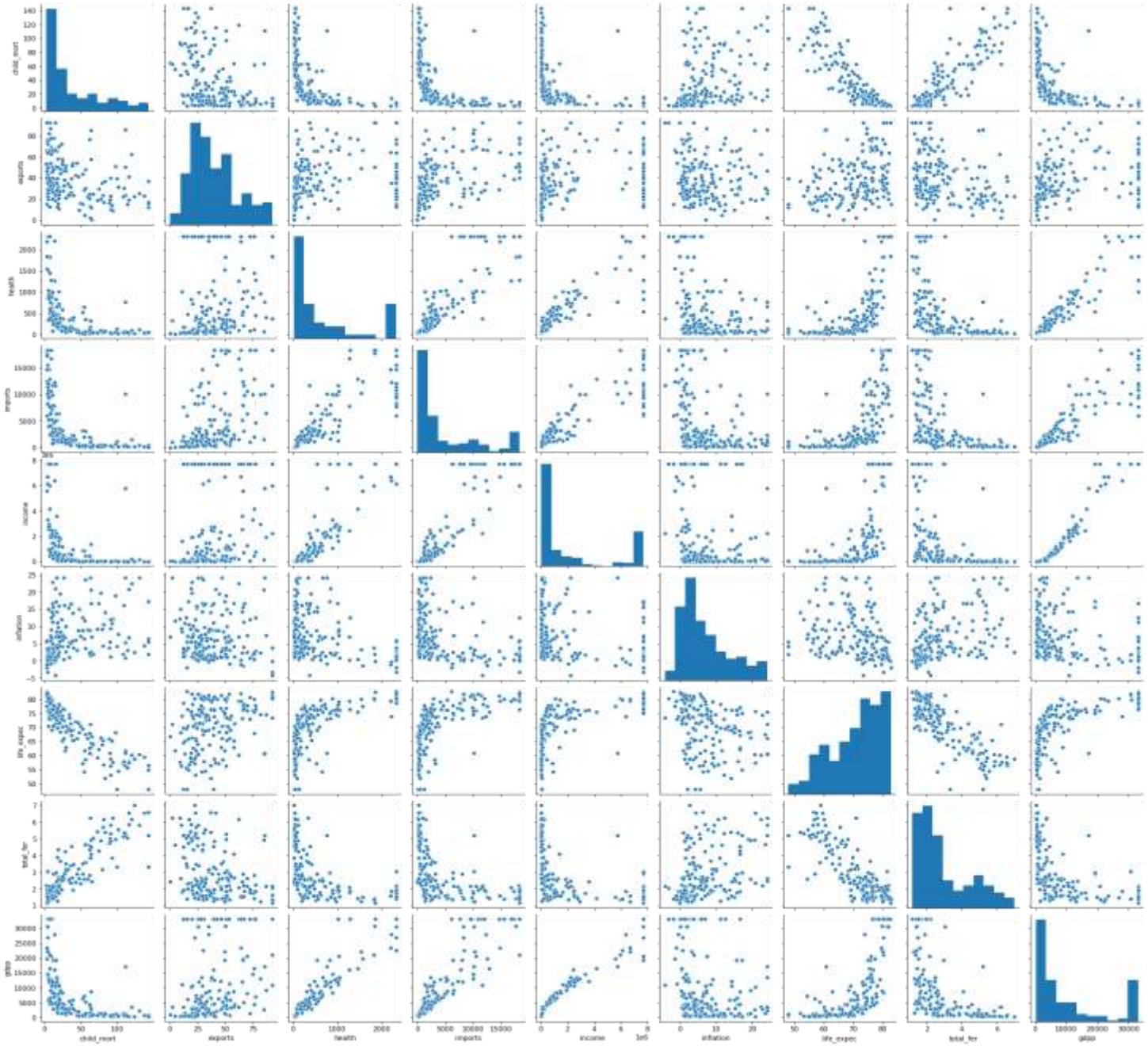
Successfully got same top 10 countries by analyzing both models i.e., Kmeans clustering and Hierarchical clustering which are in dire need of Aid.

VISUALISING DATA

CORRELATION OF THE VARIABLES

Points to be concluded from the graph on the right :

From the pair plot, we can see that there are some variables having very high correlation with respect to each other.



CORRELATION OF THE VARIABLES

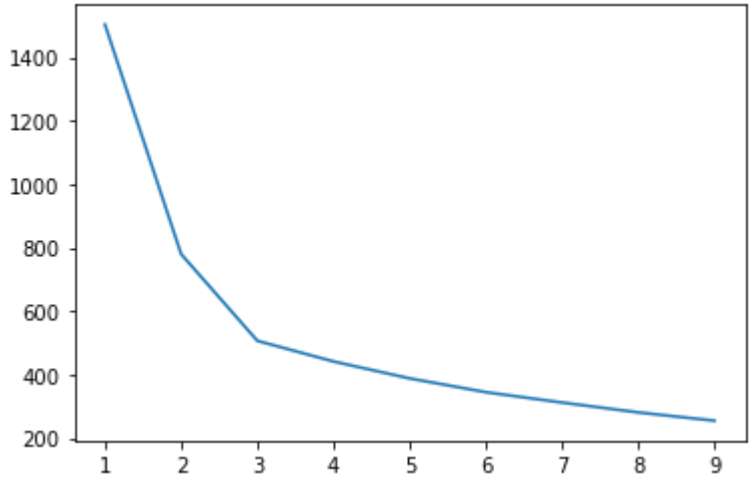
Points to be concluded from the graph on the right :

From the heatmap, we can see that there are some variables having very high correlation with respect to positive and negative.

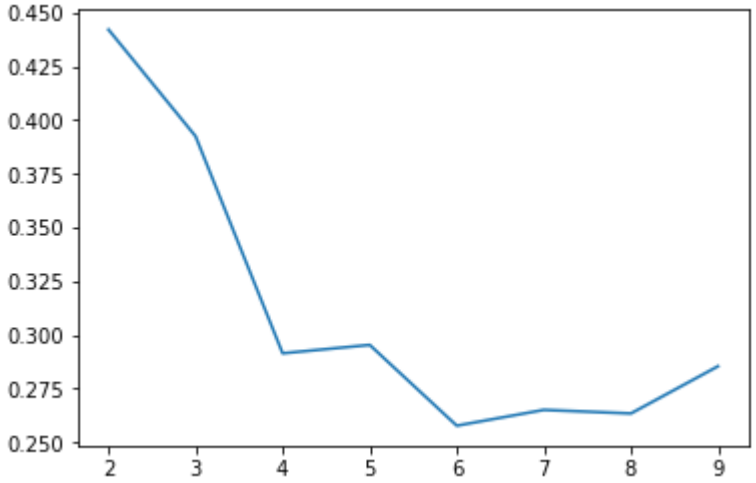
KMEANS CLUSTERING

Points to be concluded from the graph on the right :

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.



Elbow Curve



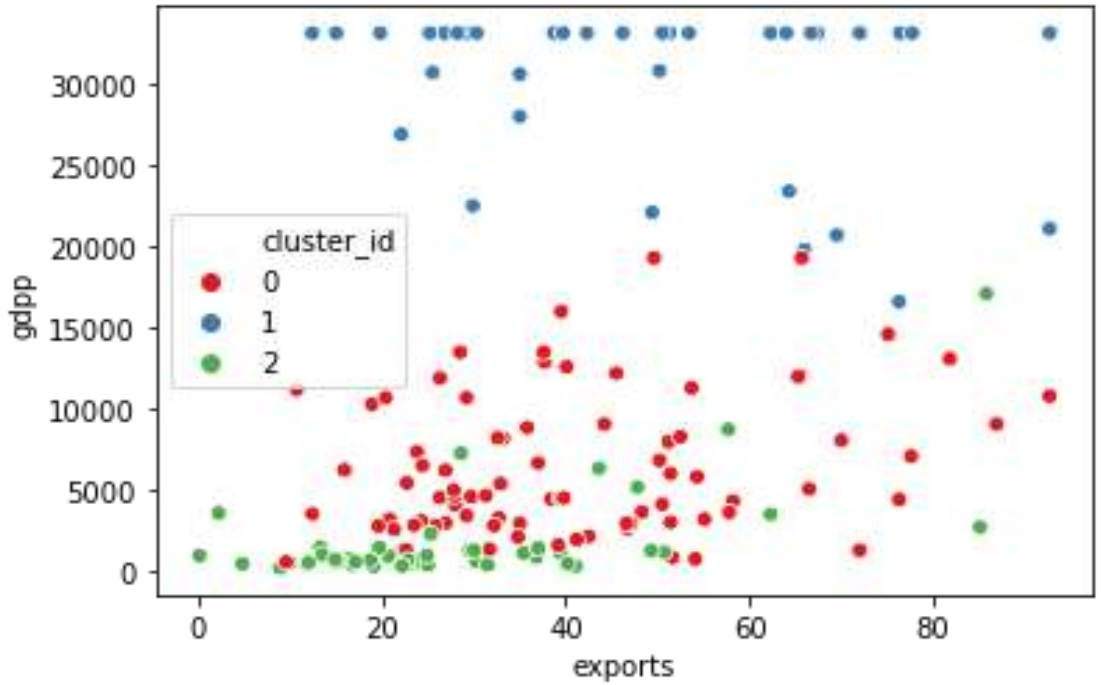
Silhouette Analysis

KMEANS CLUSTER VISUALIZATION

EXPORTS VS GDPP

Points to be concluded from the graph on the right :

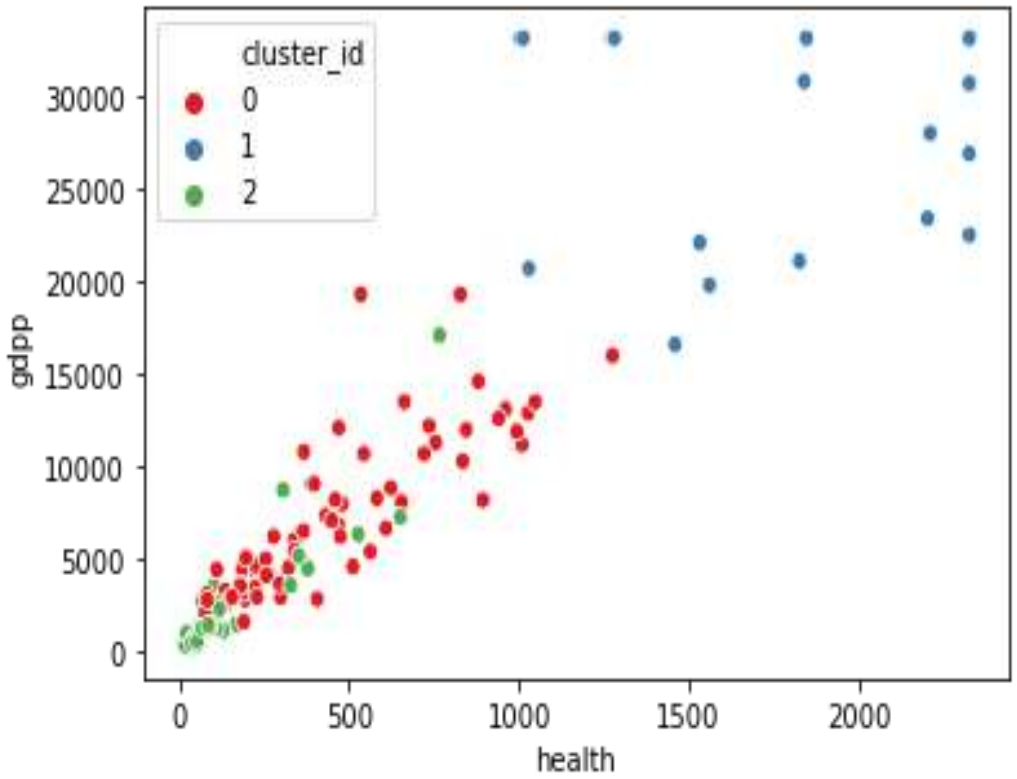
The scatter plot between exports vs gdpp not showing perfect correlation between them.



HEALTH VS GDPP

Points to be concluded from the graph on the right :

The Scatter plot between the Health and Gdpp at some point linear correlations for cluster_id 0 and 2.

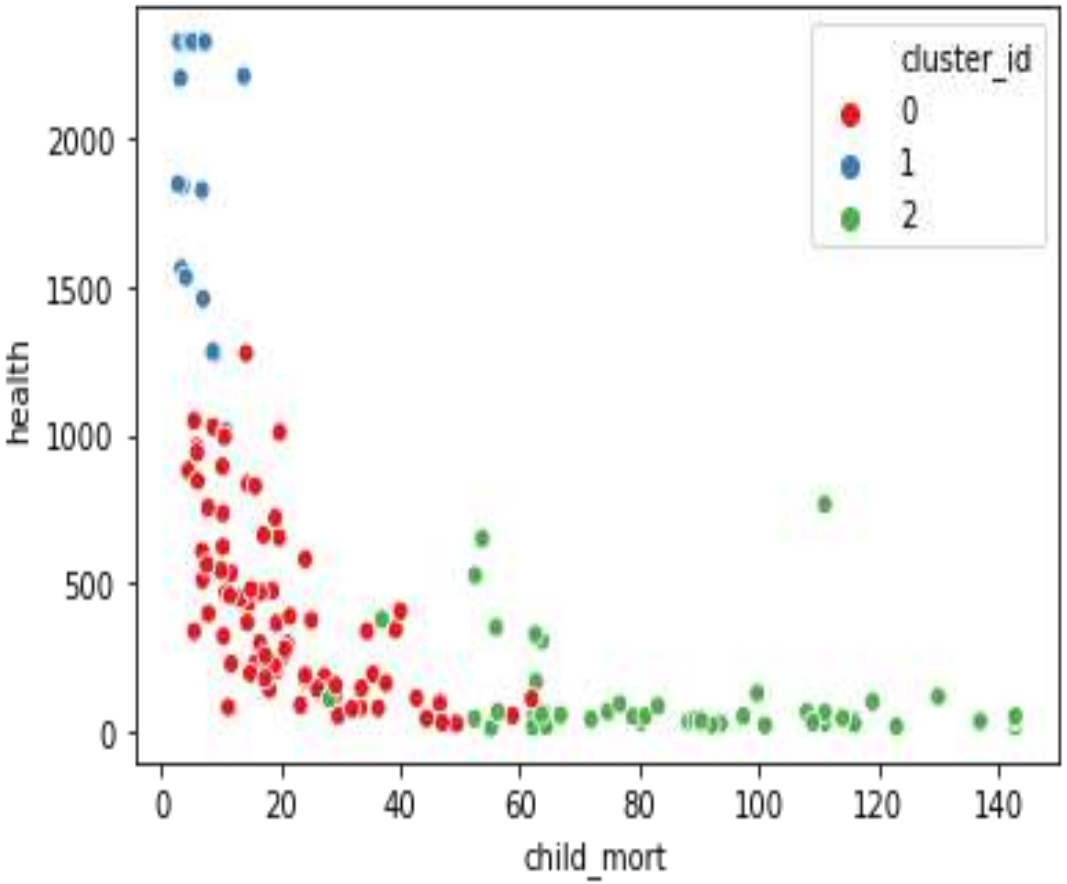


CHILD_MORT VS HEALTH

Points to be concluded from the graph on the right :

For Cluster 2 ,health expenditure is very low and child mortality is very high.

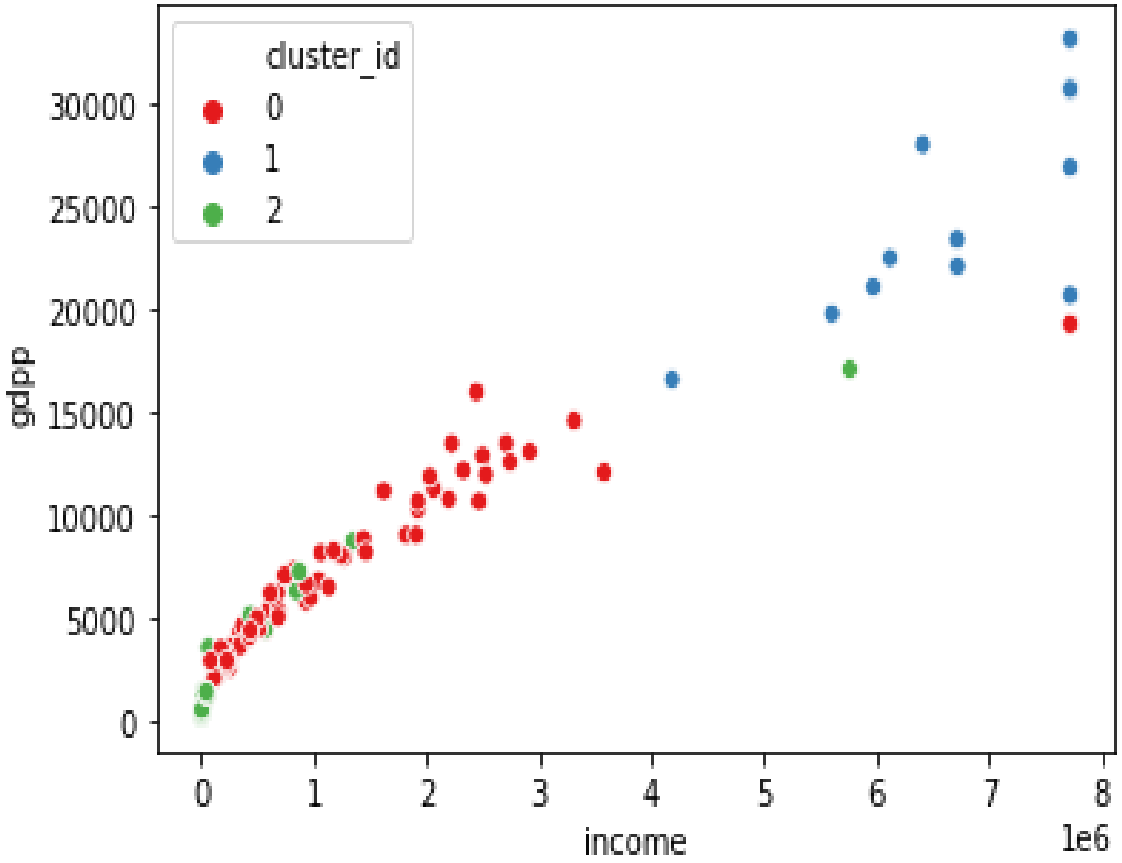
For Cluster 1, health expenditure is high and child mortality is very low.



INCOME VS GDPP

Points to be concluded from the graph on the right :

The scatter plot between the income and gdpp show linear correlation between them for cluster 0 and cluster 2.

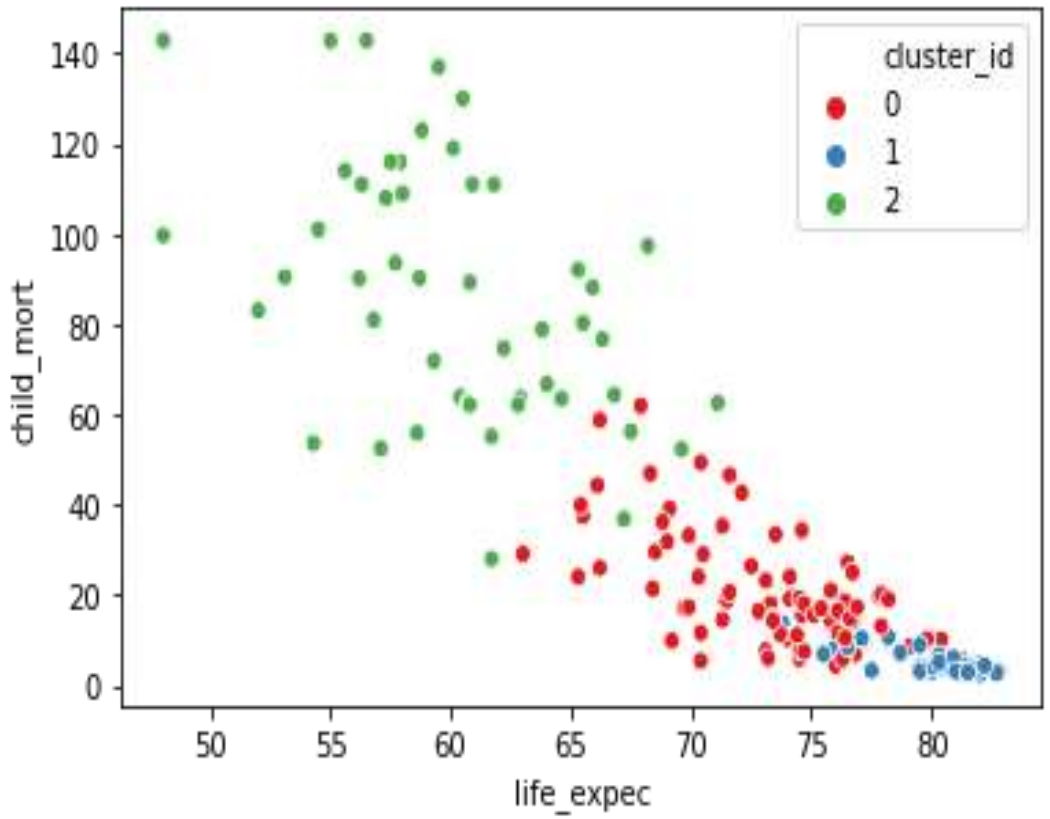


LIFE_EXPEC VS CHILD_MORT

Points to be concluded from the graph on the right :

For Cluster 2 , life expectancy is very low and child mortality is very high.

For Cluster 1, life expectancy is high and child mortality is low.

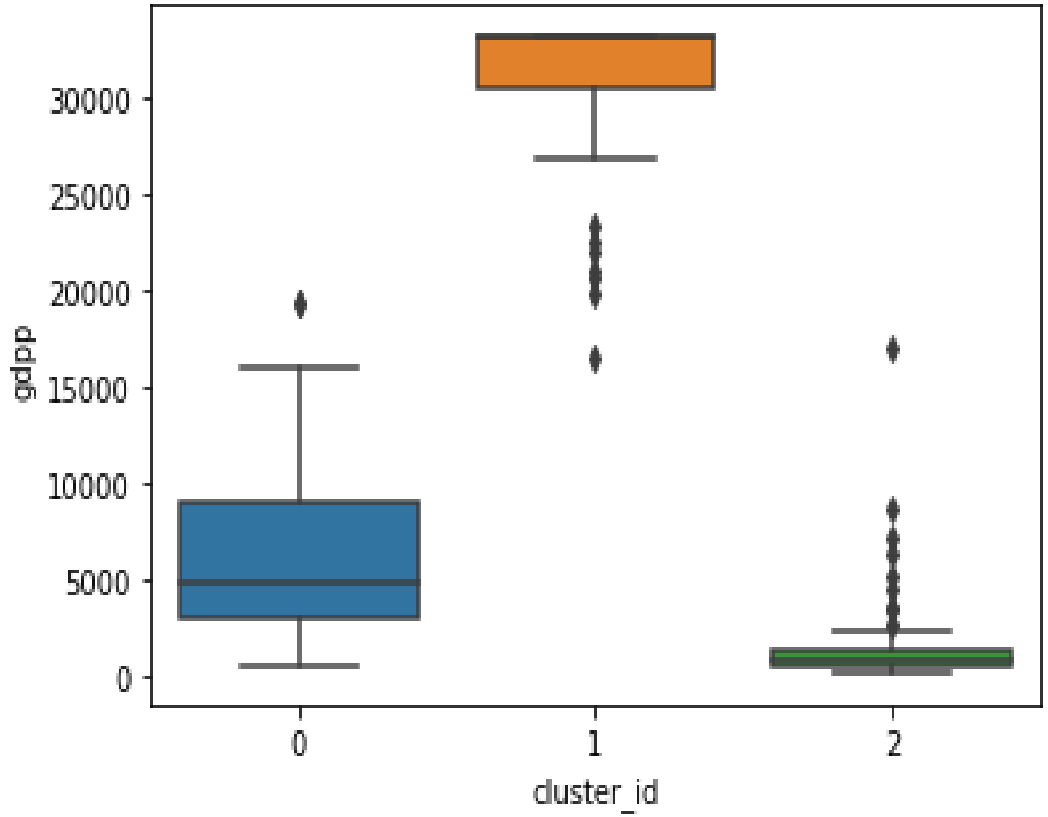


KMEANS CLUSTER PROFILING

CLUSTER_ID VS GDPP

Points to be concluded from the graph on the right :

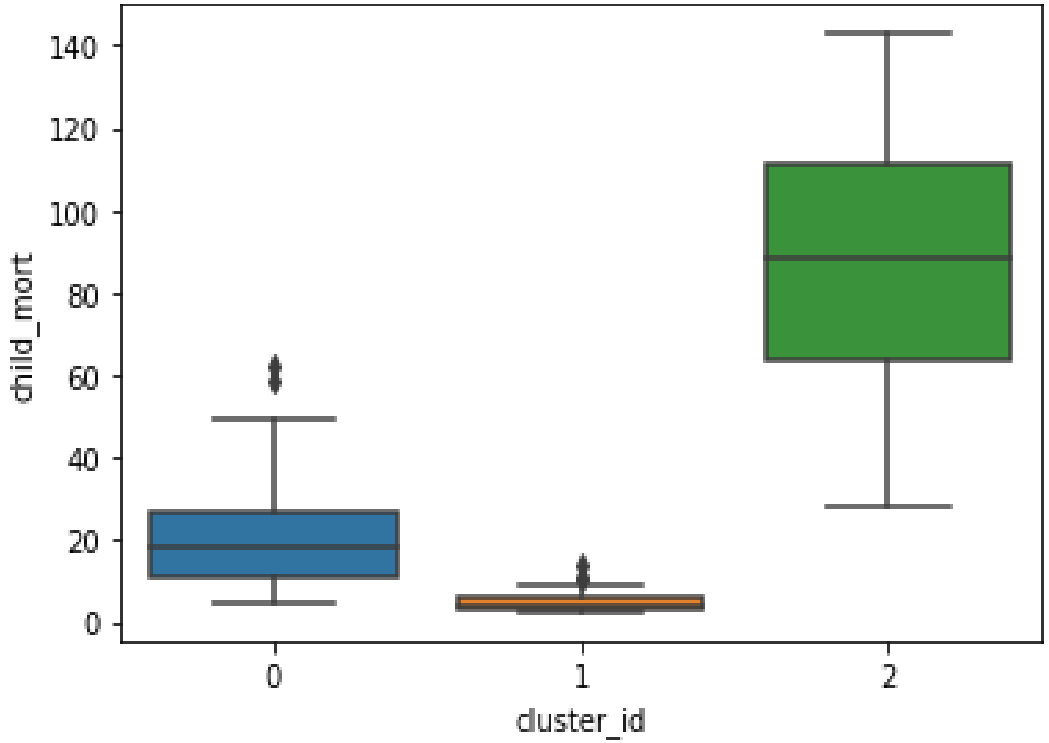
For cluster id 2 gdpp is very low in compare to other cluster id.



CLUSTER_ID VS CHILD_MORT

Points to be concluded from the graph on the right :

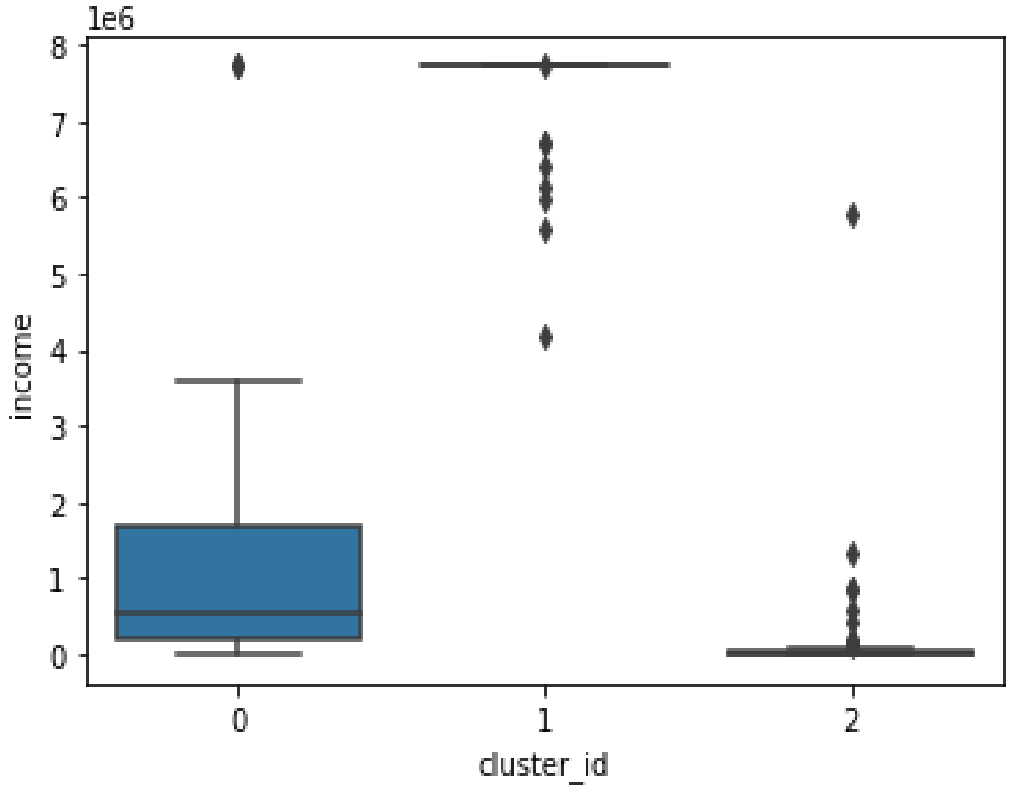
For cluster id 2 child mortality is very high in compare to other cluster id.



CLUSTER_ID VS INCOME

Points to be concluded from the graph on the right :

For cluster id 2 income is very low in compare to other cluster id.



KMEANS CLUSTERING

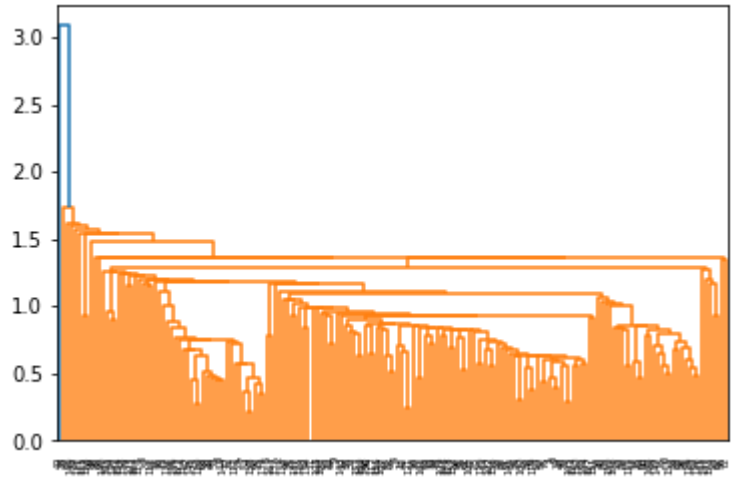
- Top 10 Countries obtained from K-Means Clustering in dire need of Aid, sorted by child mortality as decreasing order and income and Gdpp as increasing order are as Follows :

  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 Fas

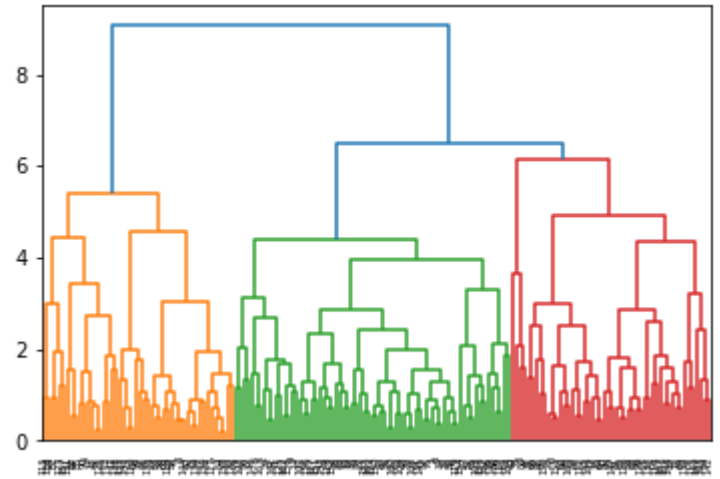
HIERARCHICAL CLUSTERING

Points to be concluded from the graph on the right :

- Single Linkage : Clusters are not forming correctly in single linkage so we will not use this modelling.
- Multiple Linkage : Single linkage is not interpretable so we will go with complete linkage model.
- We will go with the complete linkage hierarchical clustering because single linkage is more complex and cluster formation was not good in single linkage.



Single Linkage



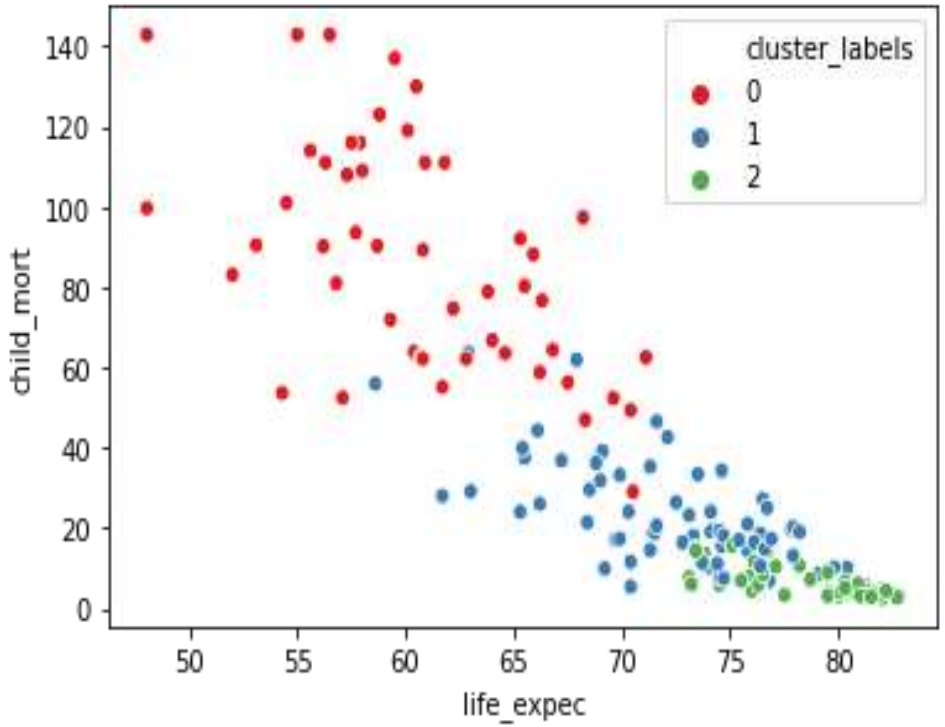
Multiple Linkage

HEIRARICHAL CLUSTERS VISUALIZATION

LIFE_EXPEC VS CHILD_MORT

Points to be concluded from the graph on the right :

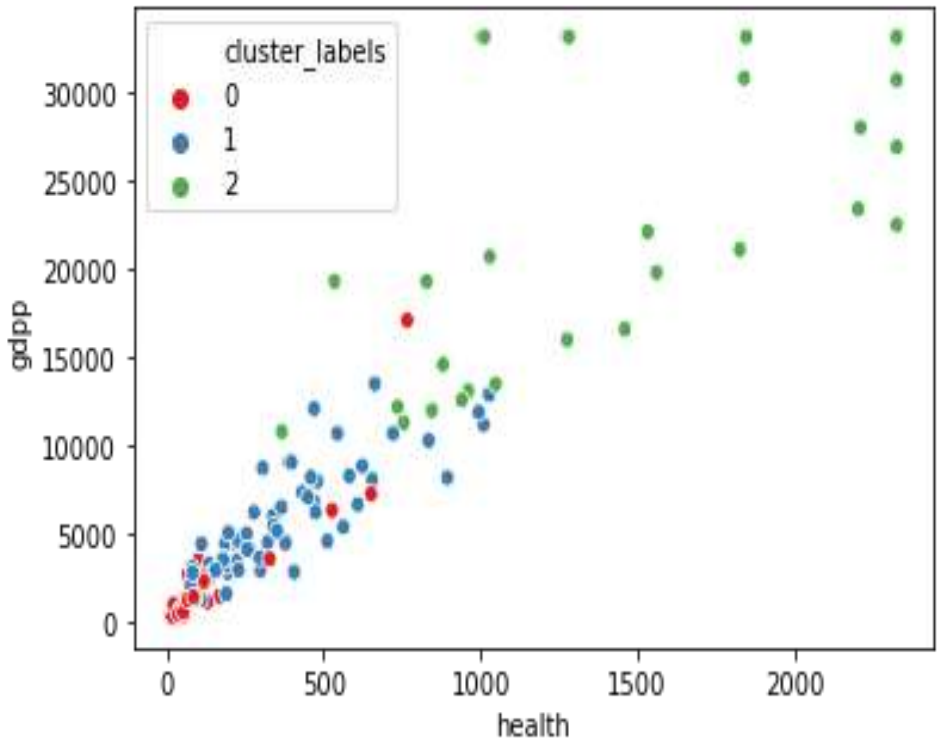
- For Cluster labels 0, life expectancy is very low and child mortality is very high.
- For Cluster labels 1, life expectancy is high and child mortality is low.



HEALTH VS GDPP

Points to be concluded from the graph on the right :

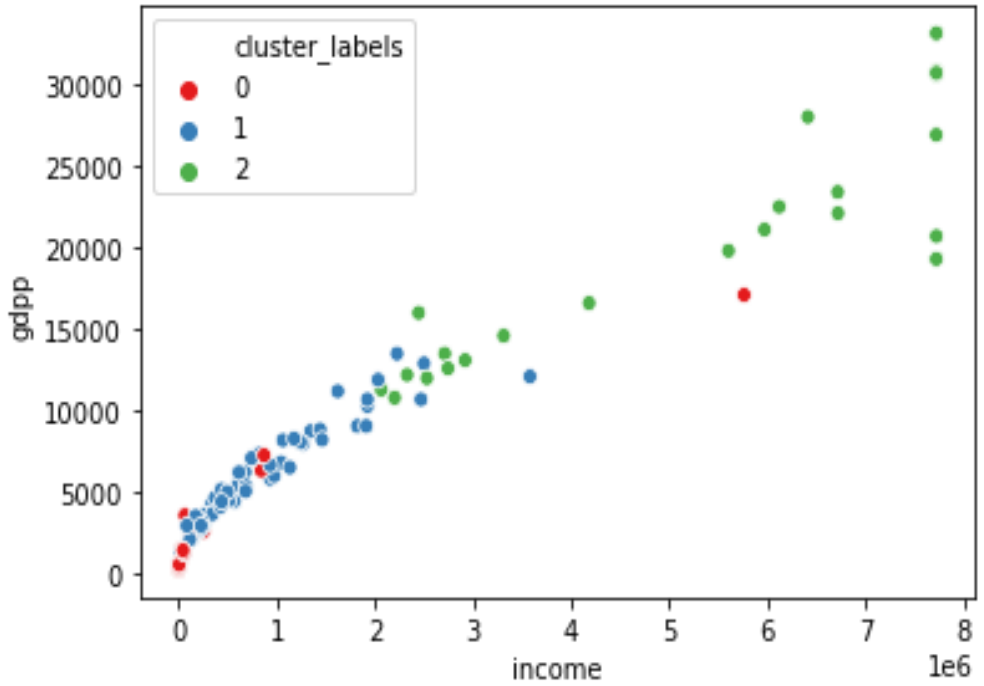
The Scatter plot between the health and gdpp at some point linear correlations for cluster_labels 0 and cluster_labels 1.



INCOME VS GDPP

Points to be concluded from the graph on the right :

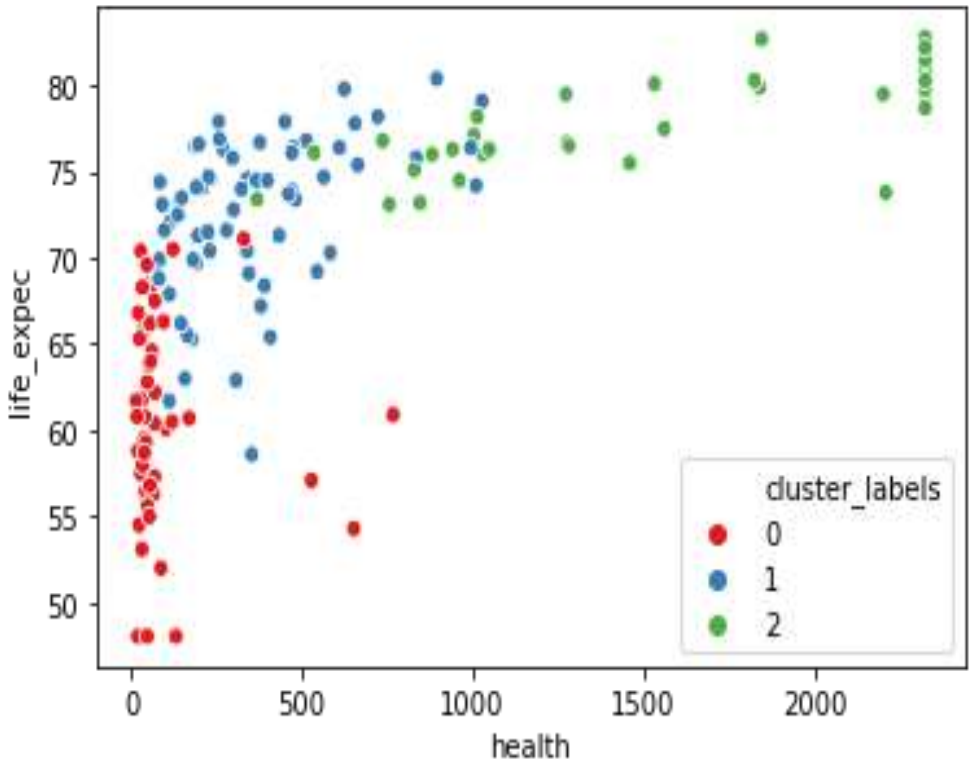
The scatter plot between the income and gdpp show some linear correlation between them for cluster_labels 0 and cluster_labels 1.



HEALTH VS LIFE_EXPEC

Points to be concluded from the graph on the right :

The scatter plot between the health and life_expec show that cluster_labels 0 and cluster_labels 1 is high life expenditure and low total health spent per capita.

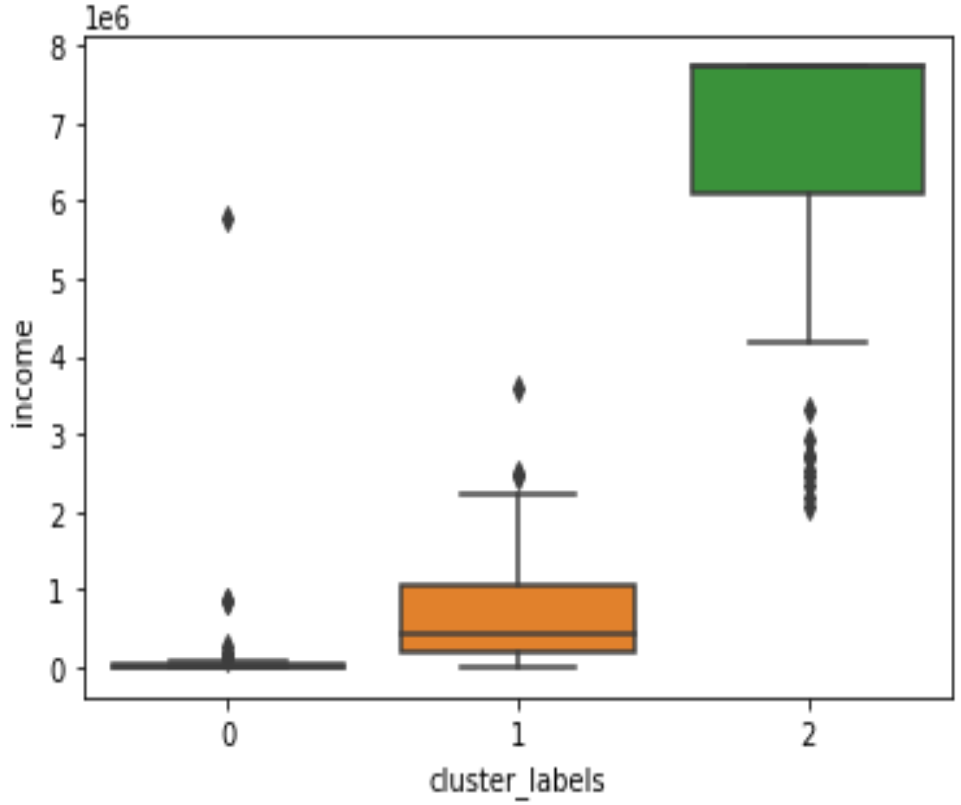


HEIRARICHAL CLUSTERING PROFILING

CLUSTER_LABELS VS INCOME

Points to be concluded from the graph on the right :

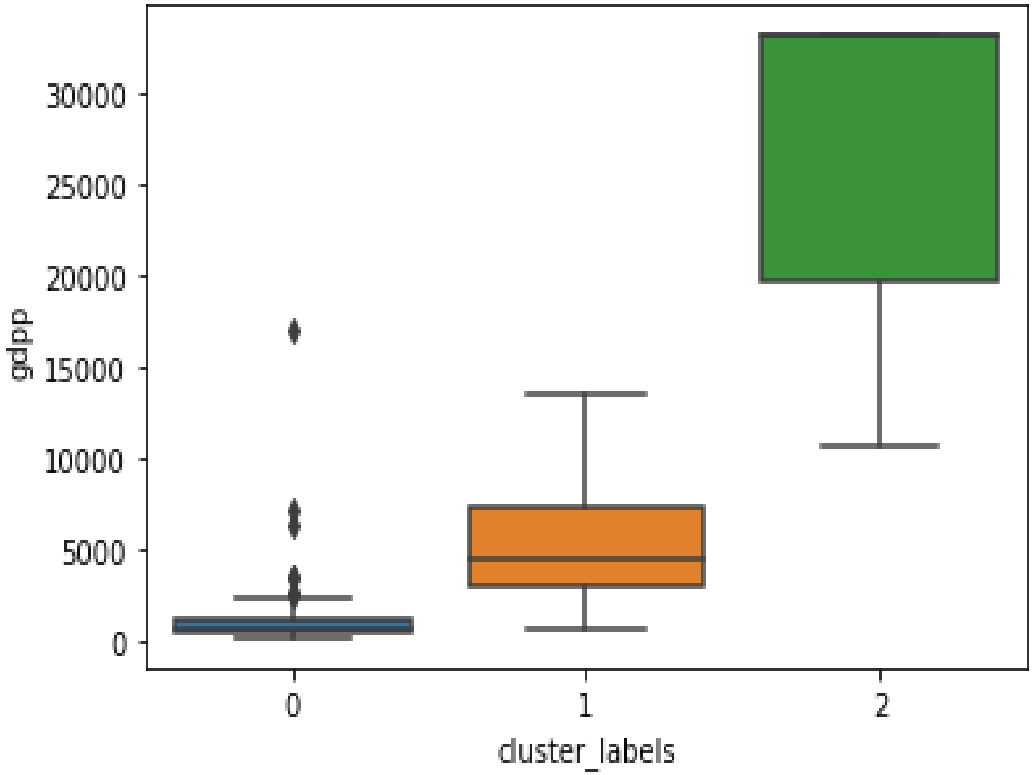
For Cluster_labels 0 income is vey low in compare to the other cluster_labels.



CLUSTER_LABELS VS GDPP

Points to be concluded from the graph on the right :

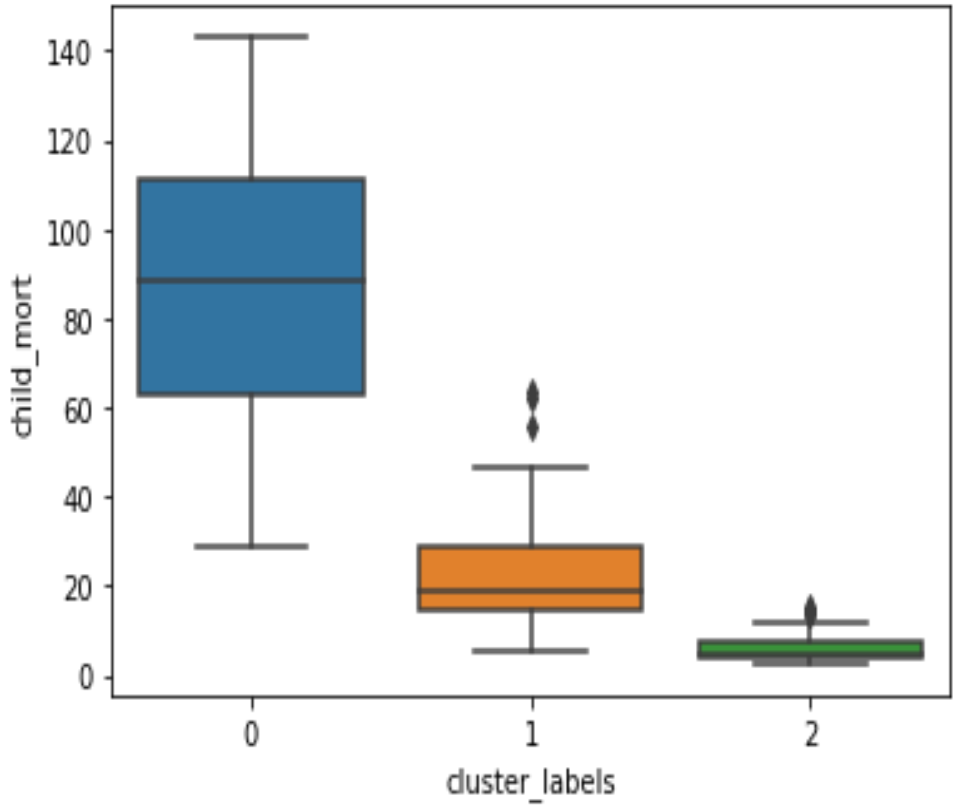
For Cluster_labels 0 gdpp is vey low in compare to the other cluster_labels.



CLUSTER_LABELS VS CHILD_MORT

Points to be concluded from the graph on the right :

For Cluster_labels 0 child_mort is vey high in compare to the other cluster_labels.



HIERARCHICAL CLUSTERING

- Top 10 countries obtained from hierarchical clustering in dire need, sorted by child mortality in decreasing order and income and gdpp in increasing order are as follows :

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 Fas

SUMMARY

Successfully got same top 10 countries by analyzing both models i.e., Kmeans clustering and Hierarchical clustering which 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 Fas

THANK YOU