HELP International NGO
Assignment on Clustering

By Tirumala Vedavyas A

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.

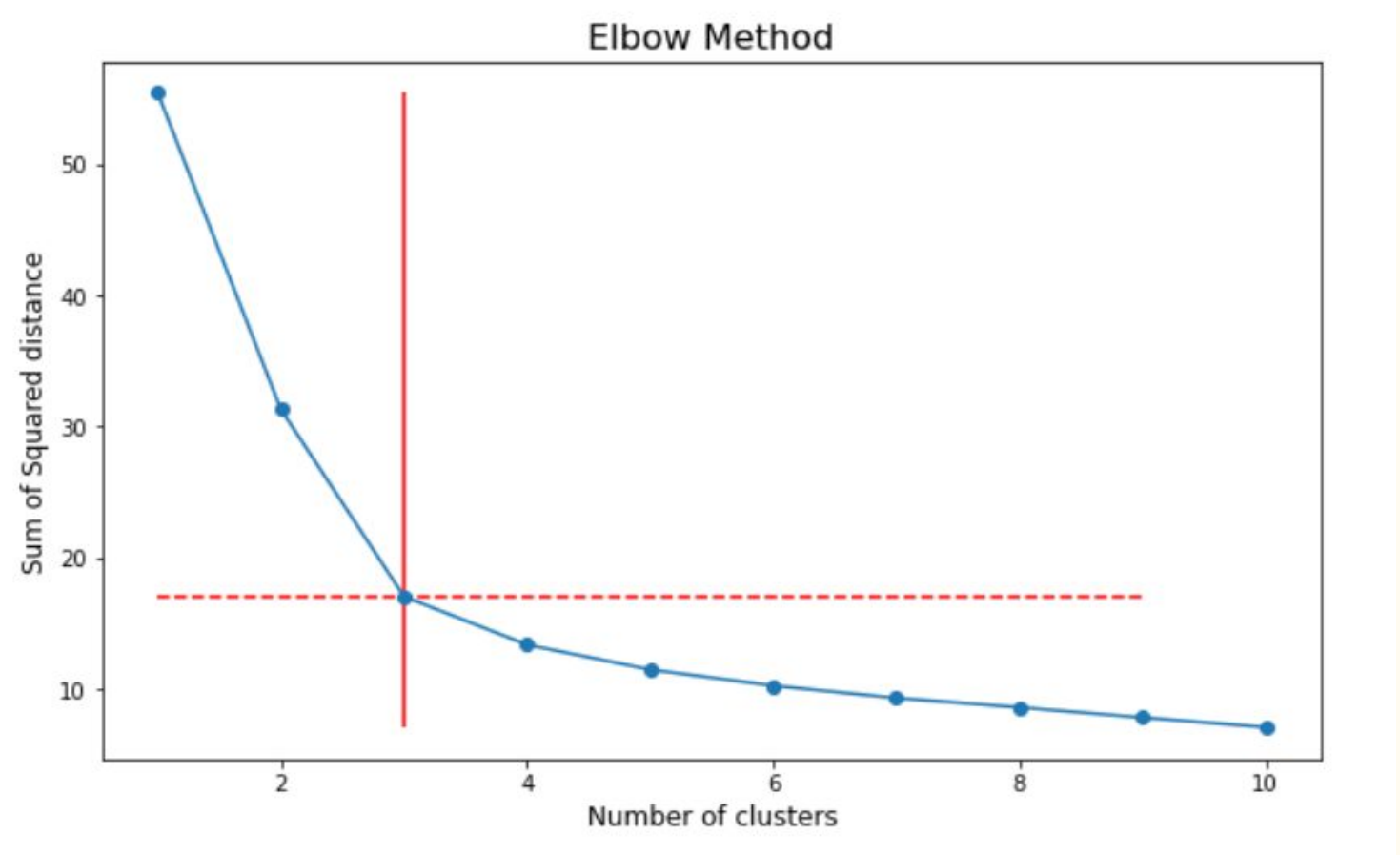
After the recent funding programmes, they have been able to raise around $ 10 million. Now the CEO of the NGO needs to decide how to use this money strategically and effectively. The significant issues that come while making this decision are mostly related to choosing the countries that are in the direst need of aid.

And this is where we come in as a data analyst. Our job is to categorise the countries using some socio-economic and health factors that determine the overall development of the country. Then we need to suggest the countries which the CEO needs to focus on the most. The datasets containing those socio-economic factors and the corresponding data dictionary are provided below

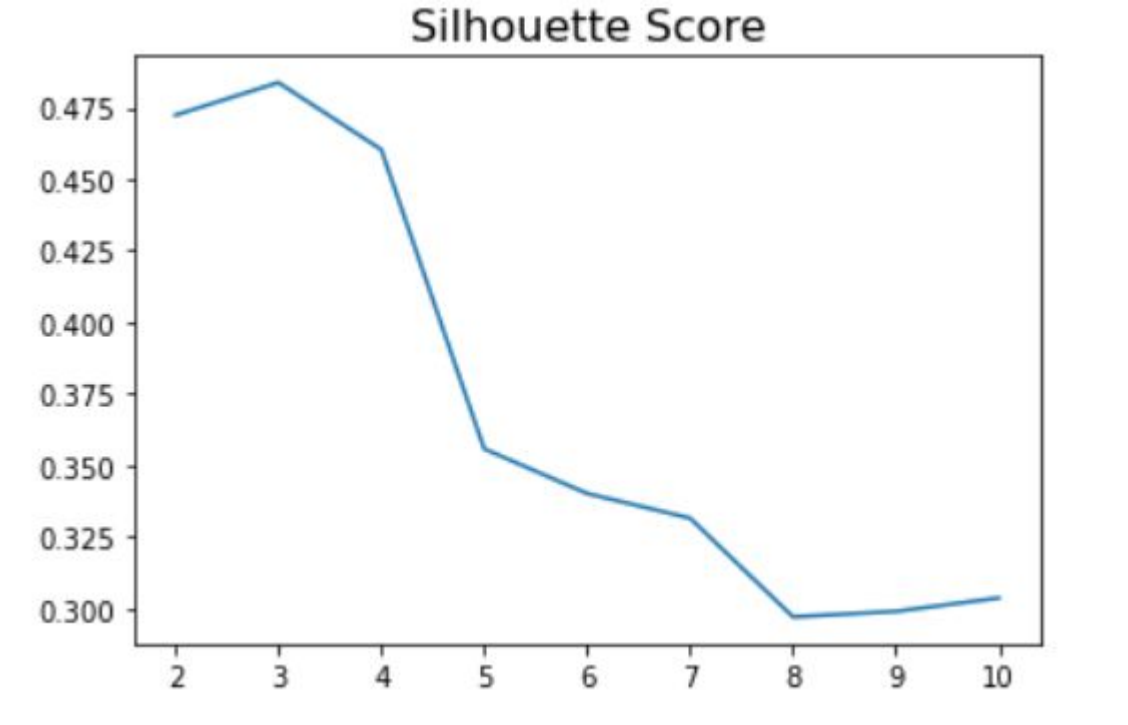
TECHNICAL APPROACH

- Using K-Means Clustering method to build clustering model.
- Use Silhouette and Elbow method to validate the optimal cluster values.
- Using Hierarchical clustering to identify the optimal cluster value.
- Use both single and complete linkage.
- Final model selection and labeling.
- Select model based on cluster results.
- Top 5 countries selection for financial aid based on socio-economic and health factors.

K-Means Clustering

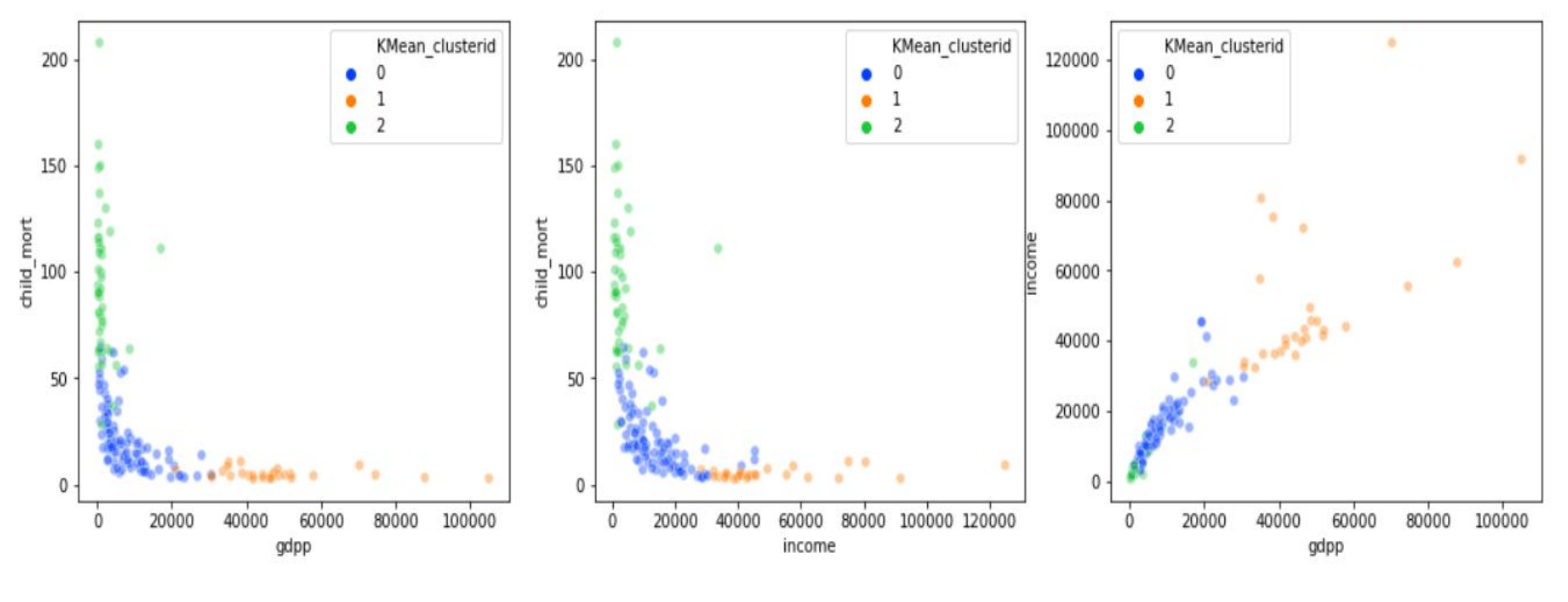


Insight: There is less reduction in sum of squared distance after cluster 3 in elbow method.

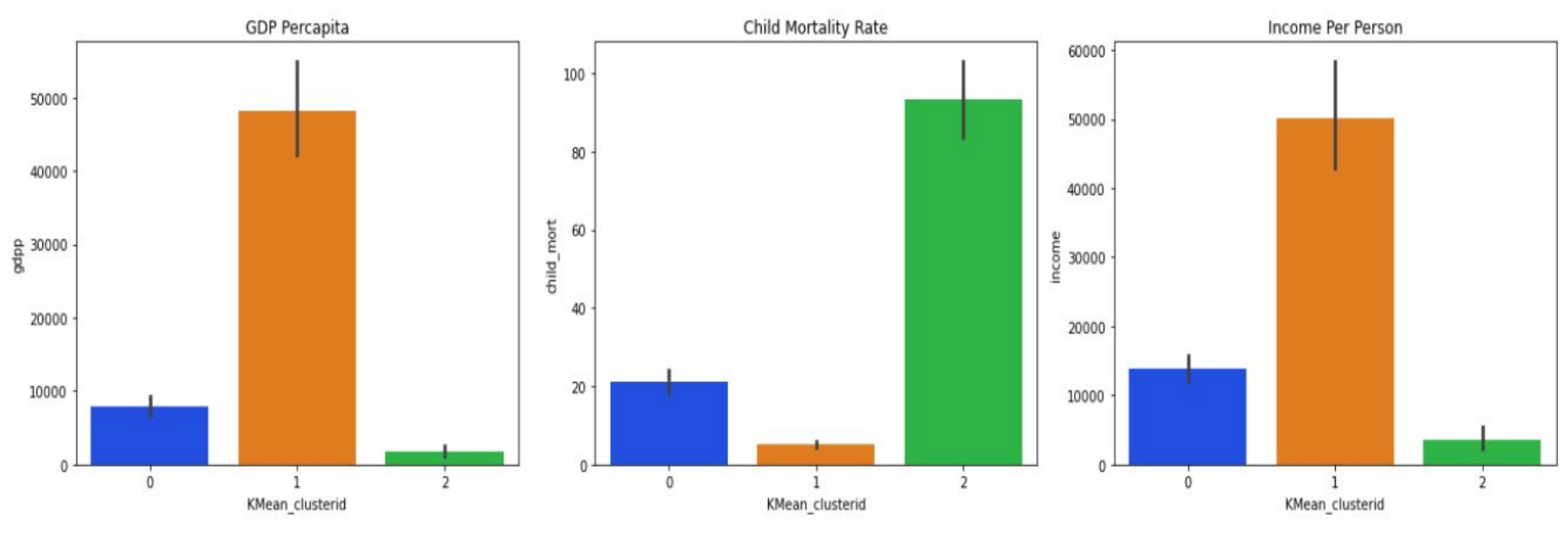


Insight: The maximum silhouette score is at cluster 3.

Scatter plot on various variables to visualize the clusters based on them

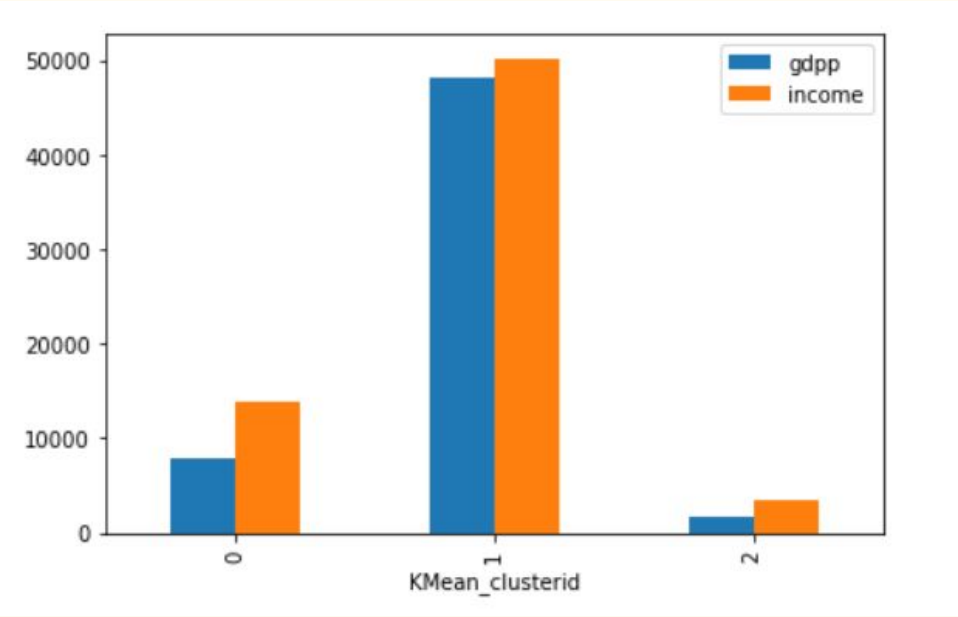


Visualising clusters

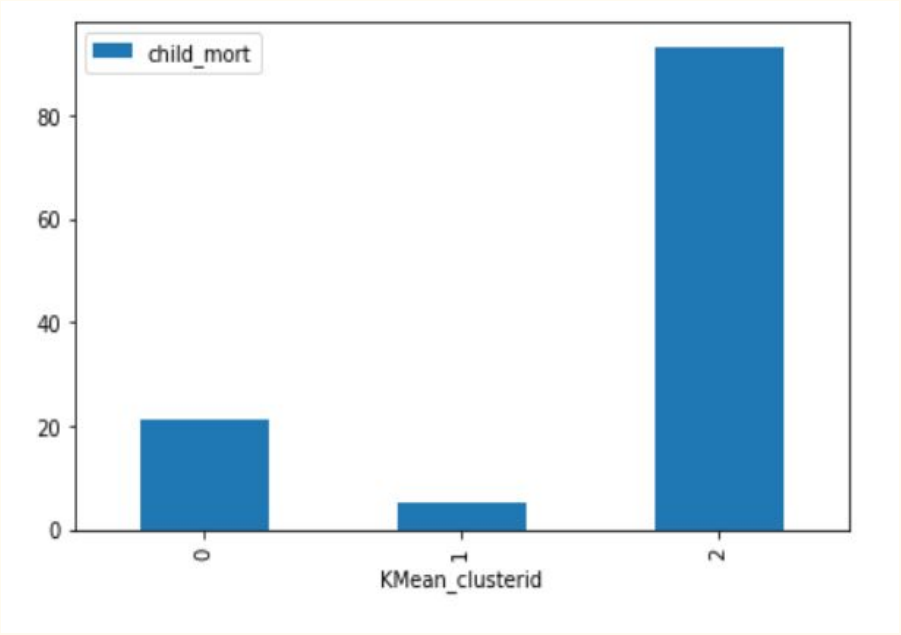


INSIGHT:It's clearly showing that the cluster 2 is having highest Child Mortality and lowest Income & GDPP and comes under undeveloped countries.

Income, gdpp vs cluster_labels



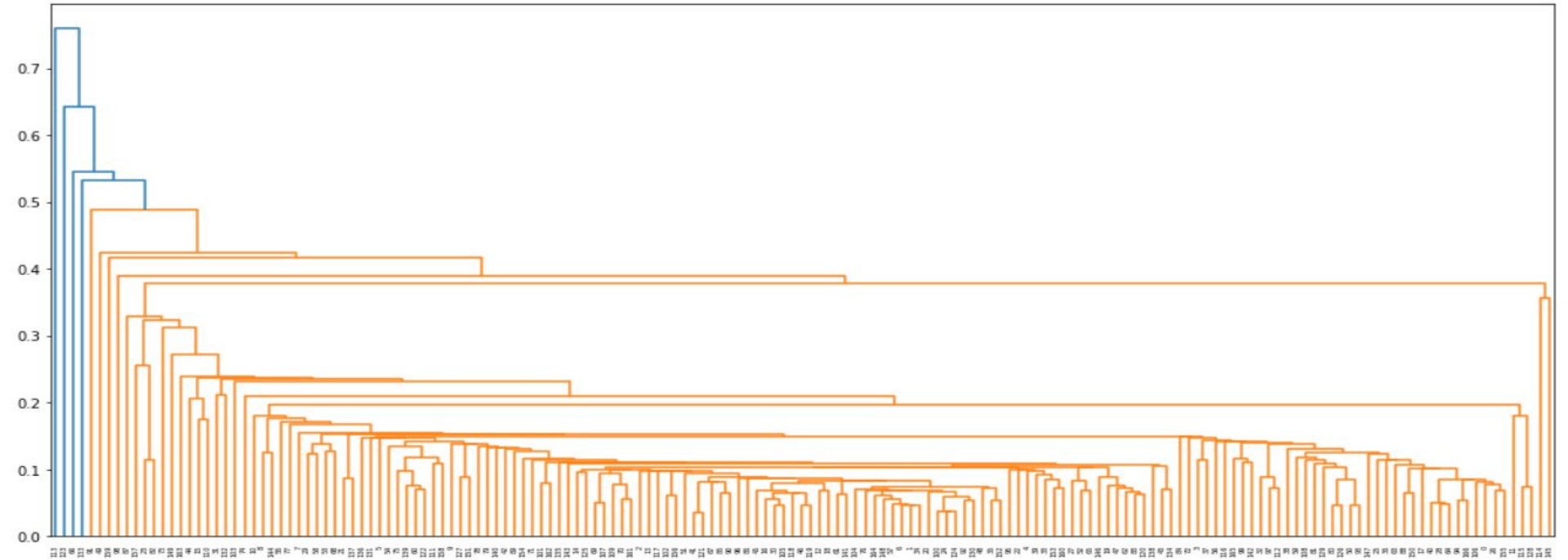
Child_mort vs cluster_labels



- Cluster 2 has the Highest Average Child Mortality rate of ~92 when compared to other clusters, and Lowest average GDPP & Income of ~ 1909 & 3897 respectively.
- All these figures clearly makes this cluster the best candidate for the financial aid from NGO. We could also see that Cluster 2 comprises of ~29% of overall data, and has ~48 observations in comparison to 167 total observations.

Hierarchical Clustering

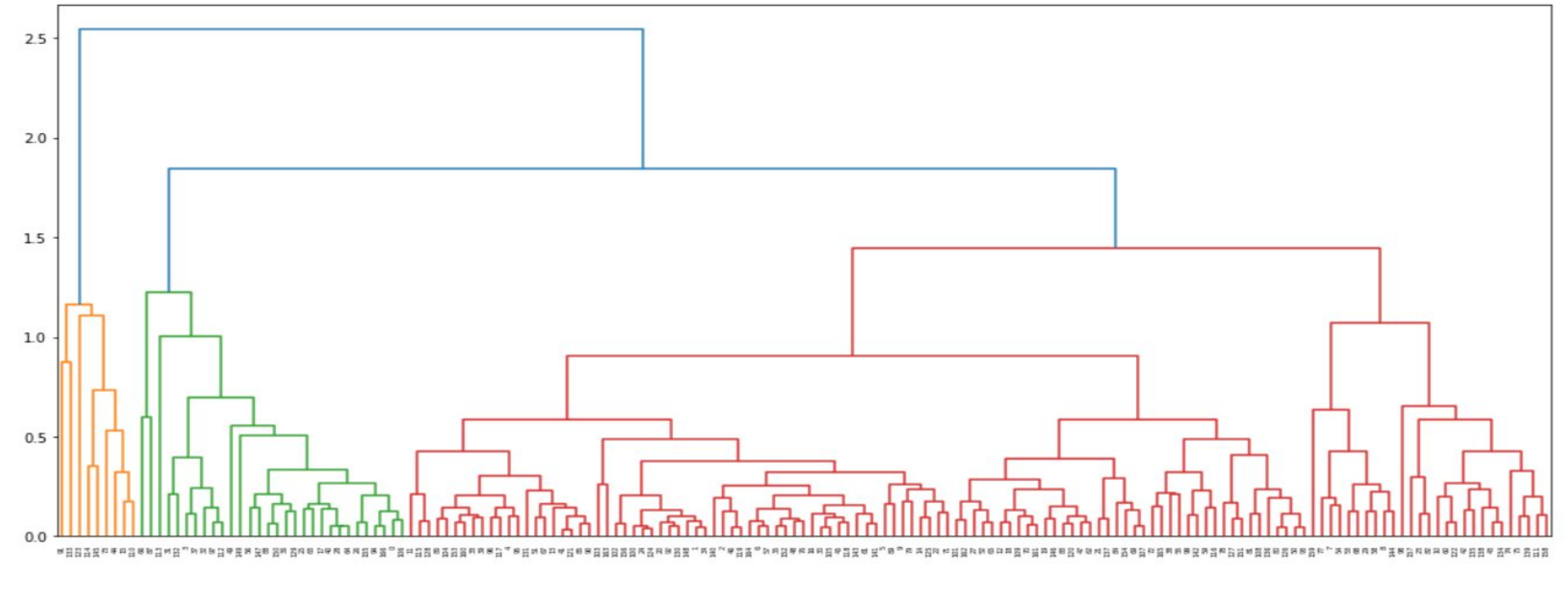
Single Linkage



Insights:

- The clusters of the single linkage are not truly satisfying. The single linkage method appears to be placing each outlier in its own cluster.
- As you can clearly see, single linkage doesn't produce a good enough result for us to analyse the clusters. Hence, we need to go ahead and utilise the complete linkage method and then analyse the clusters once again.

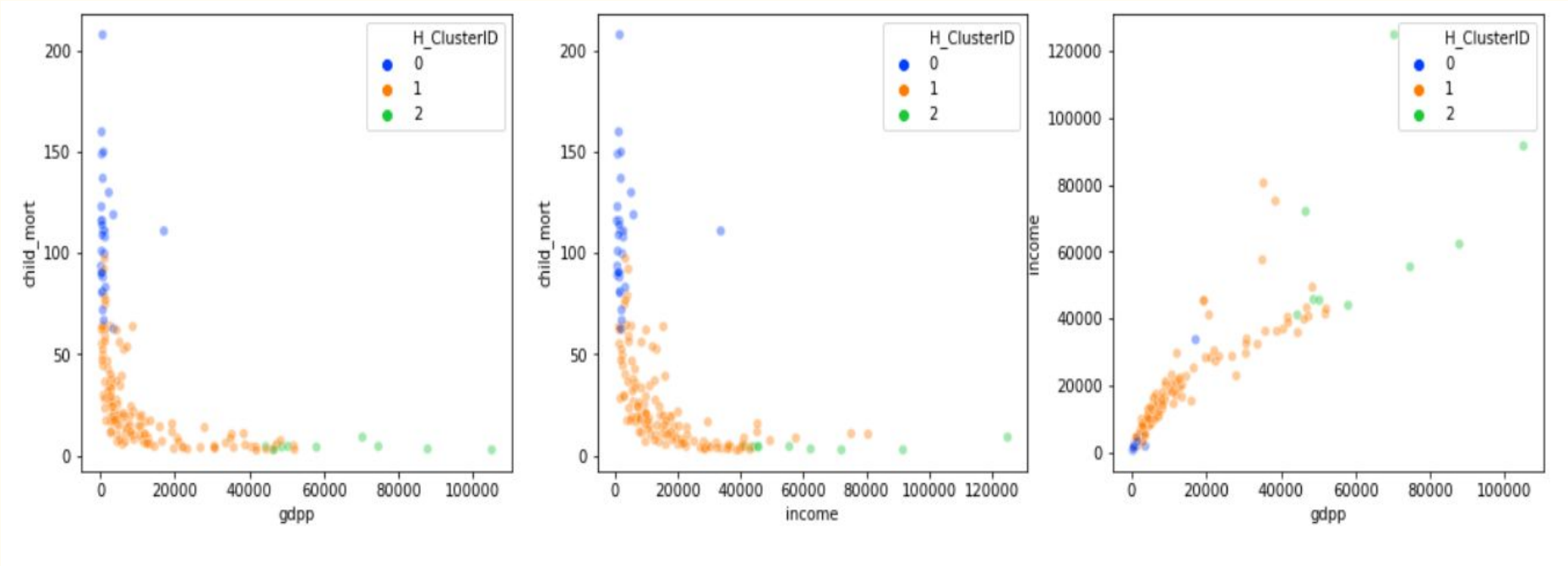
Complete Linkage



Insights:

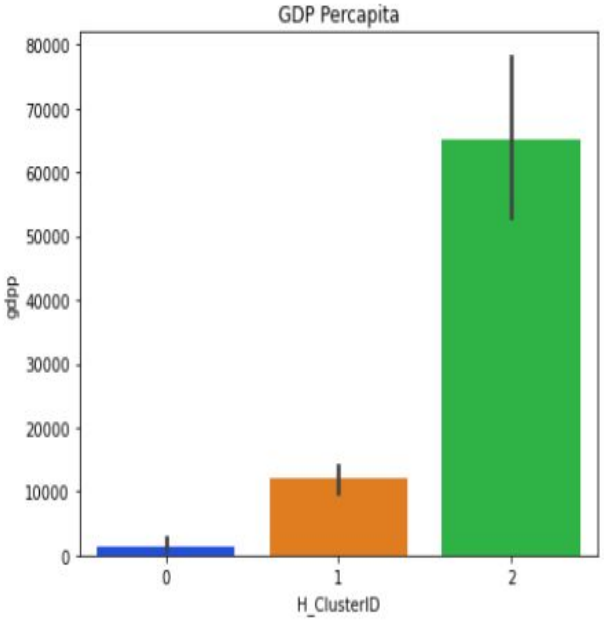
- From the above Dendrograms, it is evident that 'Complete Linkage' give a better cluster formation.
- So we will use Complete linkage output for our further analysis.
- We will build two iterations of clustering with 3 & 4 clusters and analyse the output.

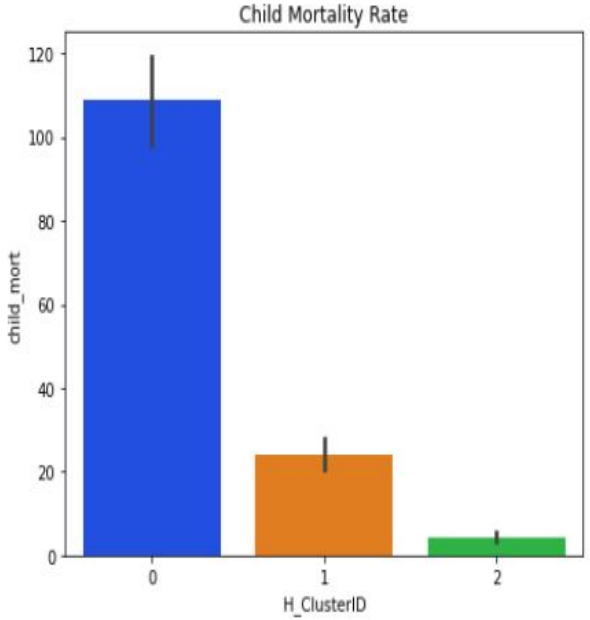
Scatter plot on various variables to visualize the clusters based on them

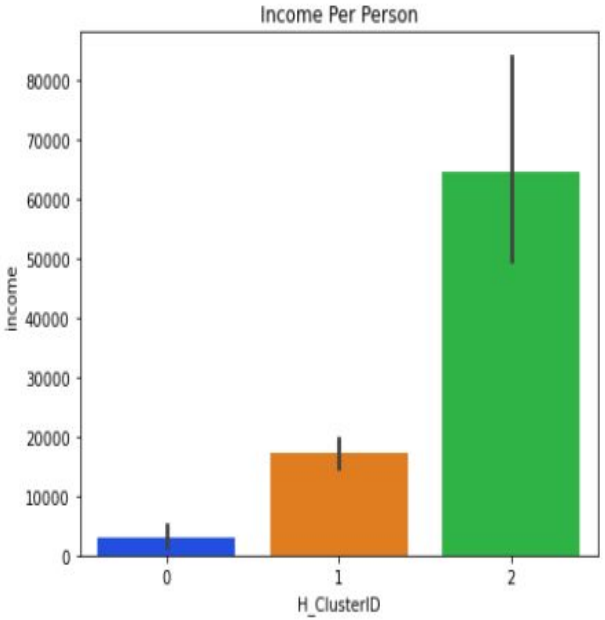


From above graphs, we can say that clusters are clearly visible.

Visualising clusters

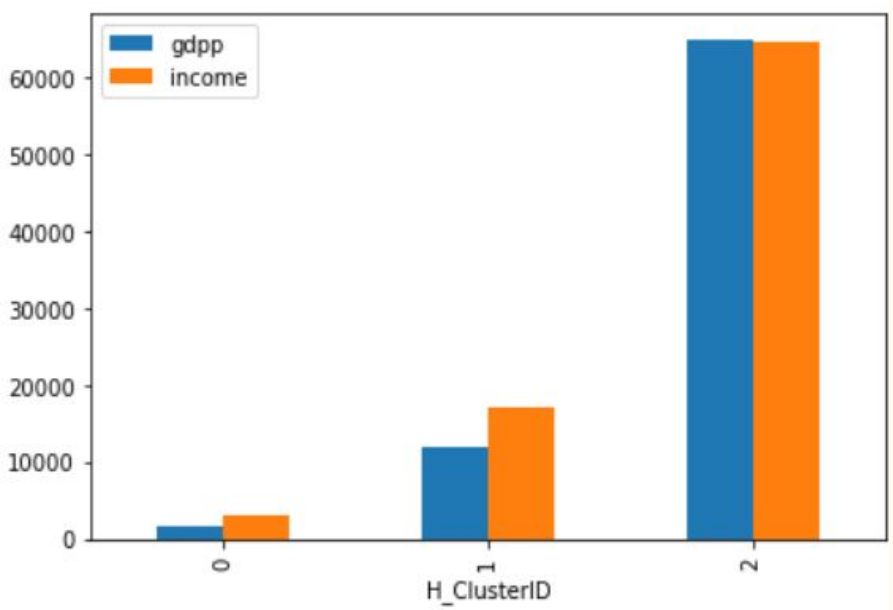




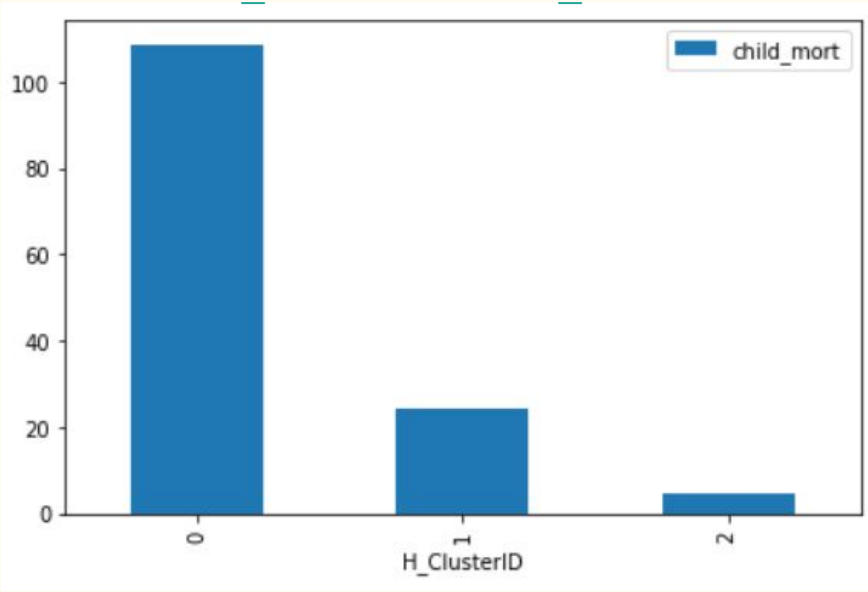


INSIGHT: It's clearly showing that the cluster 0 having highest Child Mortality and lowest Income & GDPP and its comes under undeveloped countries.

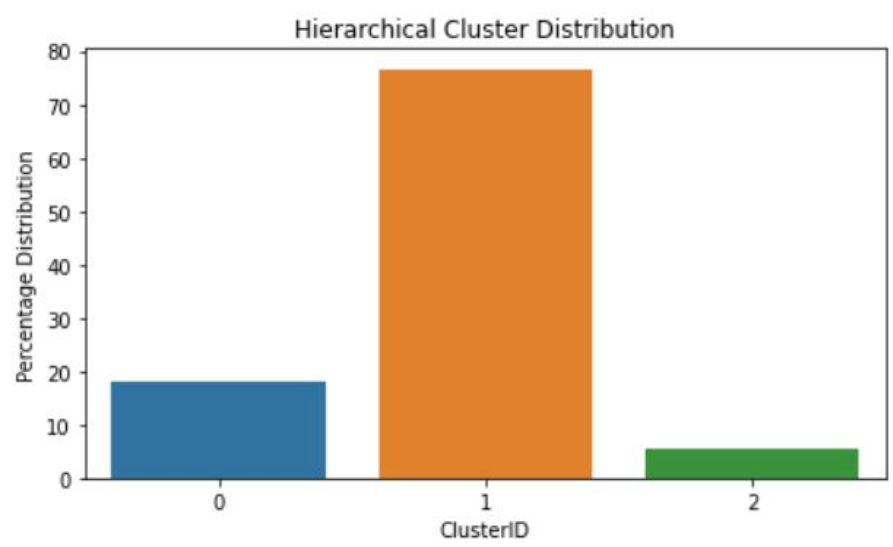
Income, gdpp vs cluster_labels

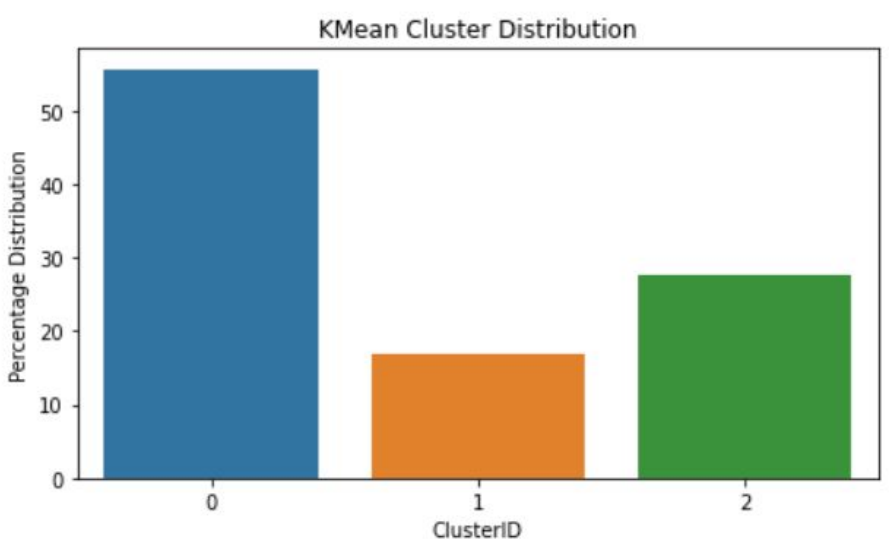


Child_mort vs cluster_labels



- Cluster 0 has the Highest average Child Mortality rate of ~42 when compared to other 3 clusters, and Lowest average GDPP & Income of ~ 7551 & 12641 respectively.
- All these figures clearly makes this cluster the best candidate for the financial aid from NGO.
- We could also see that Cluster 1 comprises of ~89% of overall data, and has ~148 observations in comparison to 167 total observations This seems to be a problem.
- This means that Hierarchical clustering is not giving us a good result as 89% of the data points are segmented into that cluster.
- We also saw that increasing the cluster number is not solving this problem. We will perform K-Means Clustering and check how that turns out to be.

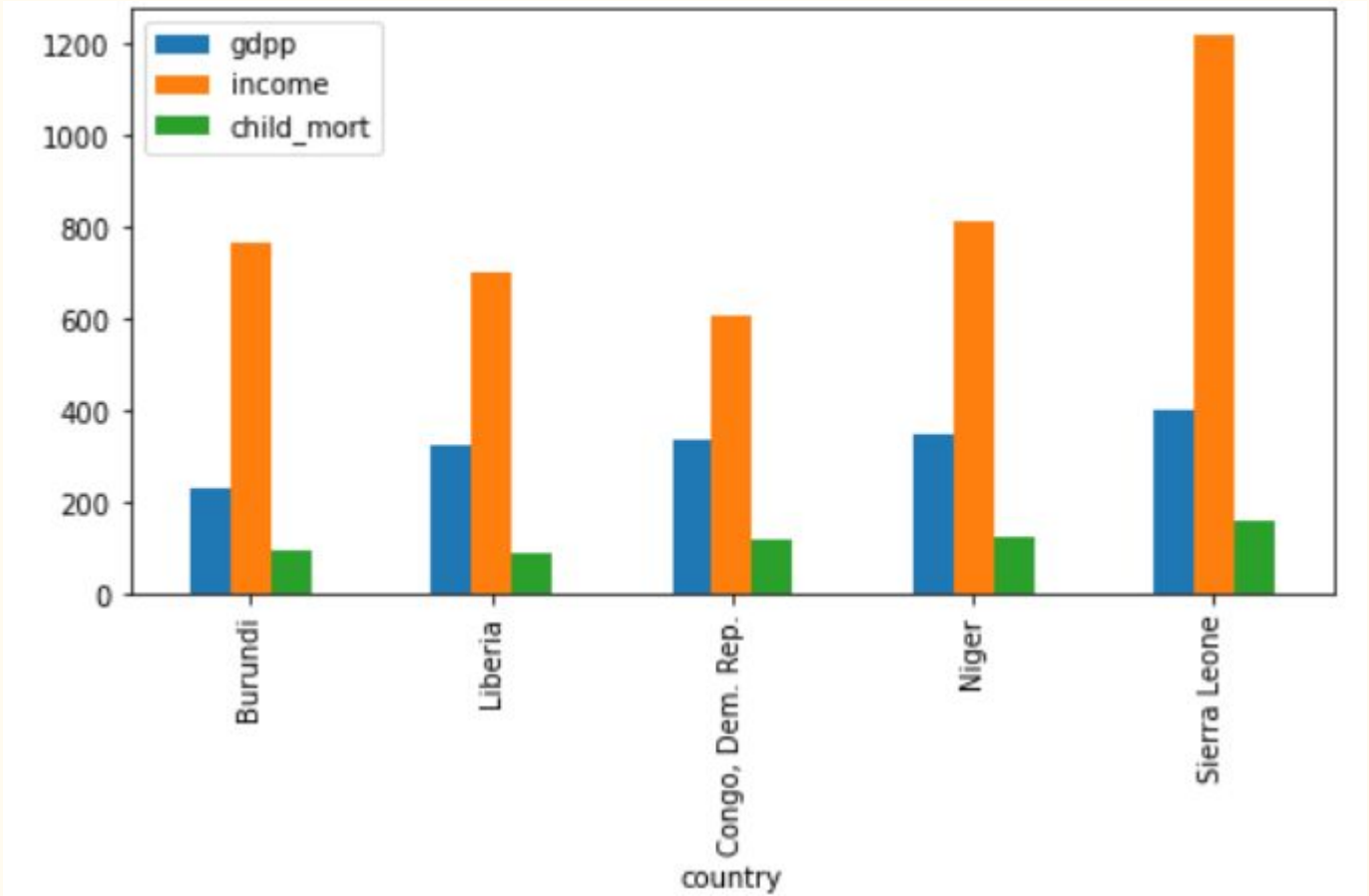




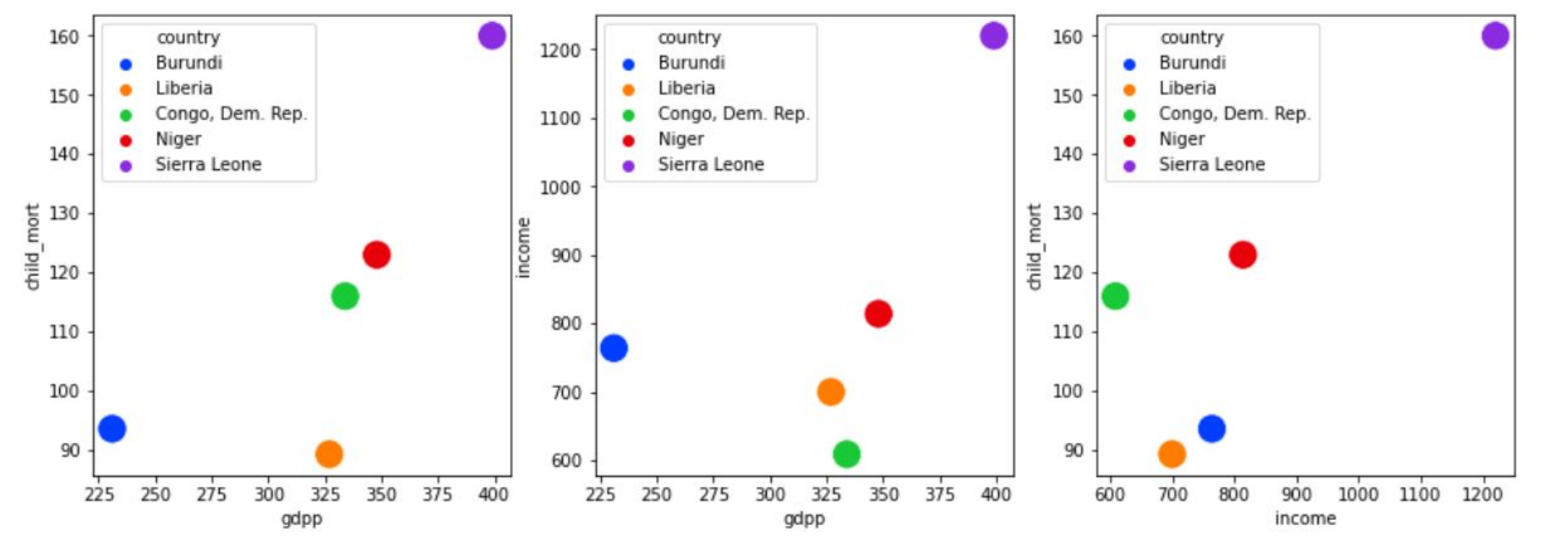
Cluster Summary

- From above analysis we could see KMean is having better distributed cluster. So we will select final model as K Means cluster and doing profiling considering the labels accordingly.
- Kindly note that both the model has resulted the same countries as top 5 undeveloped countries.
- By comparing averages of K-means we can conclude that
  - Cluster 1 belongs to Undeveloped Countries,
  - Cluster 2 belongs to Developed Countries,
  - Cluster 0 belongs to Developing Countries.

Top 5 Countries for financial Aid



Bivariate Analysis of Cluster 'Under_Developed_Countries' (recommended 5)



Conclusion:

Performed CLUSTERING on the socio-economic data provided for various countries to identify countries to recommend for Financial Aid from the NGO.
Based on our Clustering Analysis, I have identified the top countries under our 'Undeveloped Countries' cluster which are in dire need of the Financial Aid. This output is purely based on the dataset we used and various analytical methodology we performed.

The Final list of Countries for Financial Aid on priority basis are:

1. Burundi
2. Liberia
3. Congo, Dem. Rep.
4. Niger
5. Sierra Leone