CLUSTERING ASSIGNMENT.....

Presentation of Major Analysis

By

SUBHAJIT BANERJEE

Reg. mail id- banerjee21subhajit@gmail.com

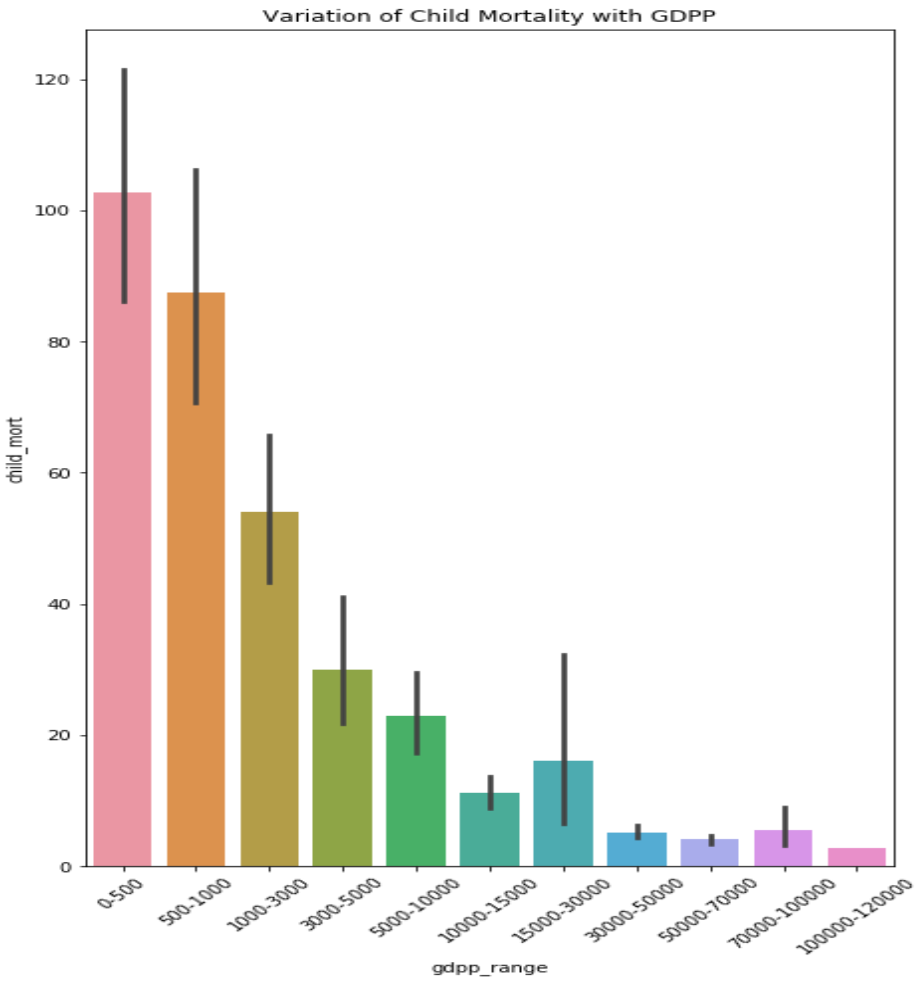
Clustering Assignment for Countries....

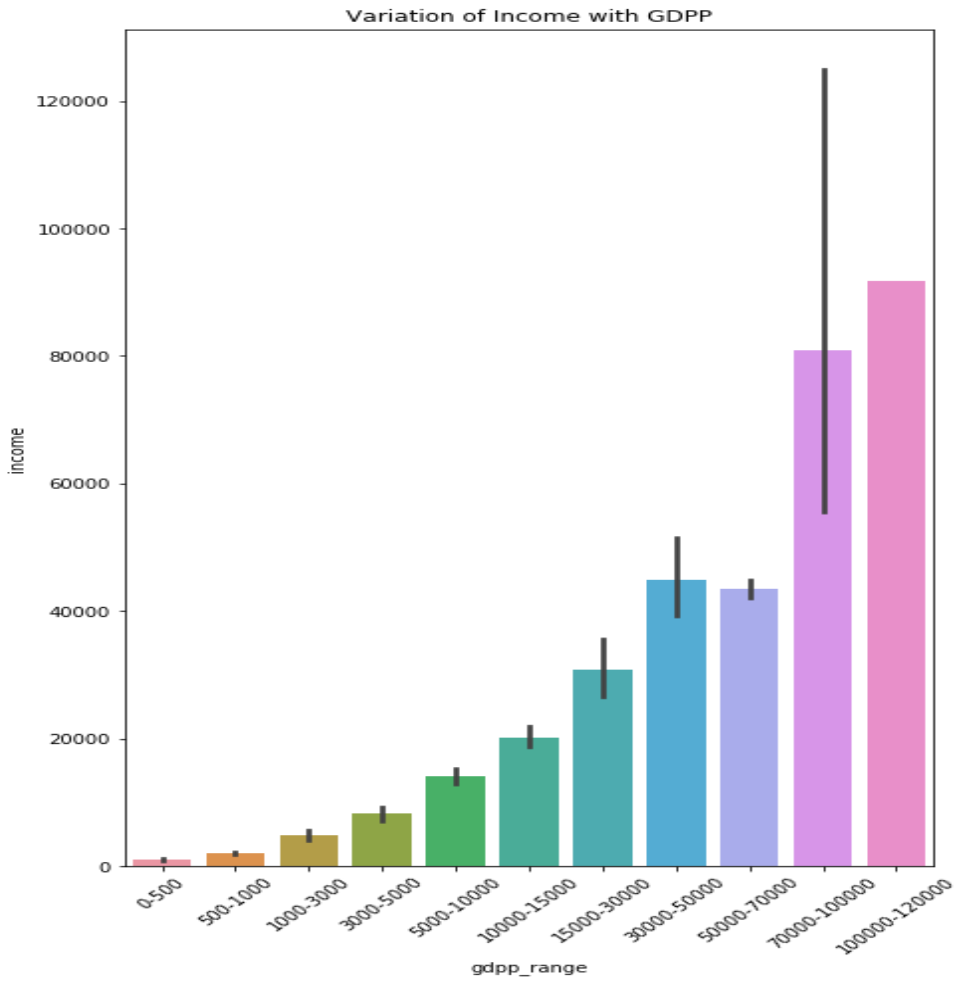
In this assignment I am going to analyse the variables which are have a lead role for countries overall developments, assigning the countries to separate clusters based on these variable factors and visualize and finalise those countries which could be the prime target for the NGO for which they are planning to invest the amount of.

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 you come in as a data analyst. Your job is to categorise the countries using some socio-economic and health factors that determine the overall development of the country. Then you 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.

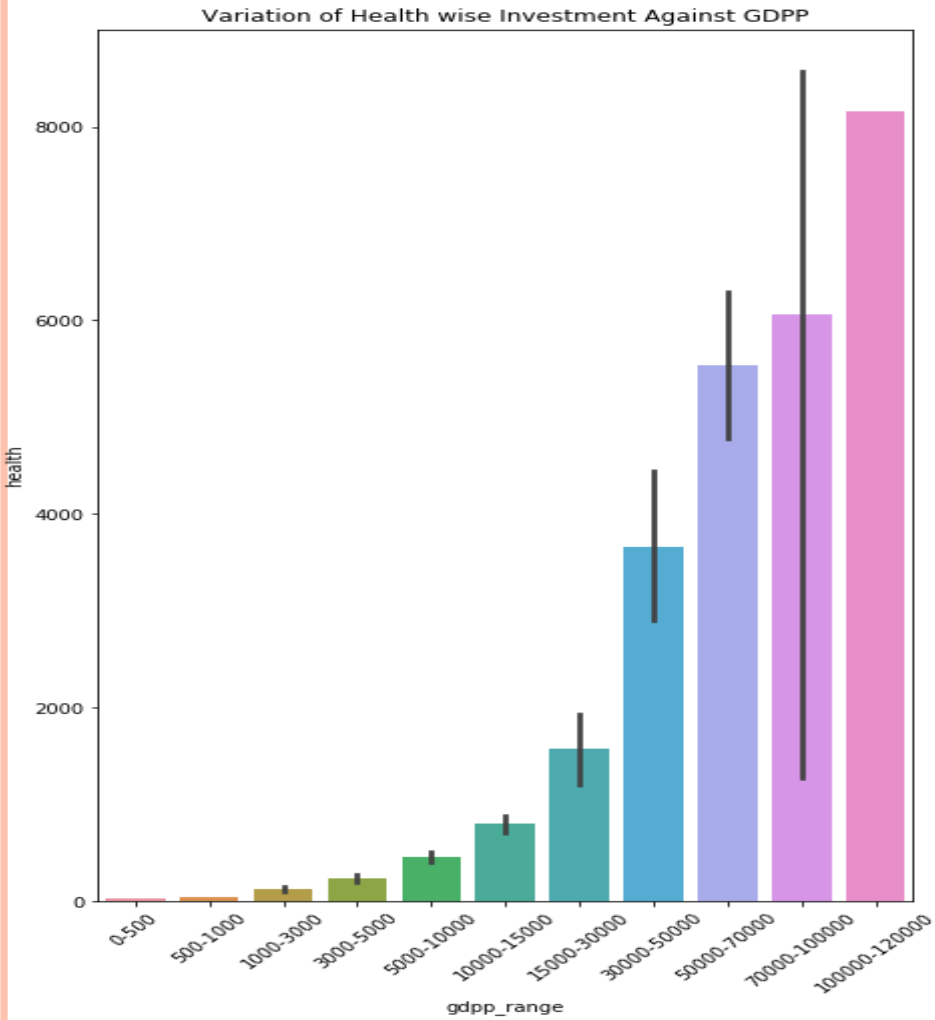
EDA Analysis
Comparison of Child Mortality with GDP and Income

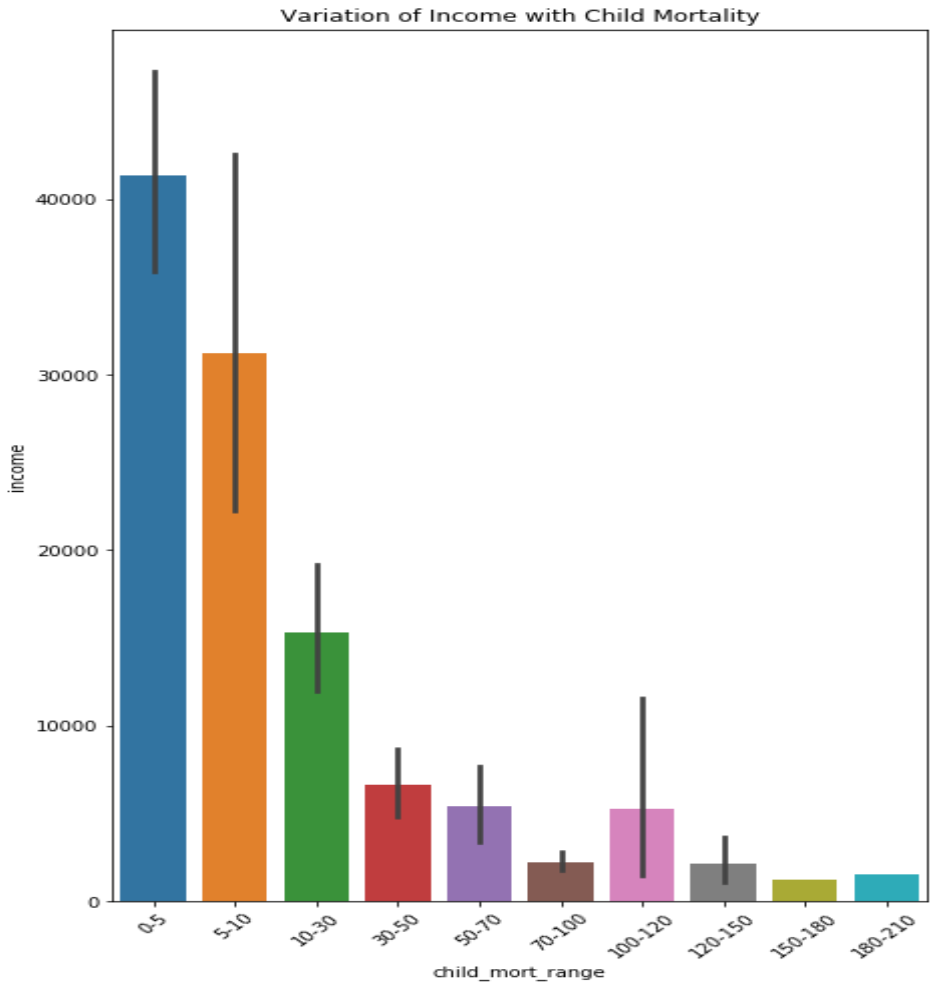




From the above bar plot it is clear that the countries with high GDP and high Income are socio-economically more Strong than the countries with low GDP and low Income. And from the 2nd bar plot it is also clear that, the countries for which GDP is low their Income status is also low and it gradually increases with GDP.

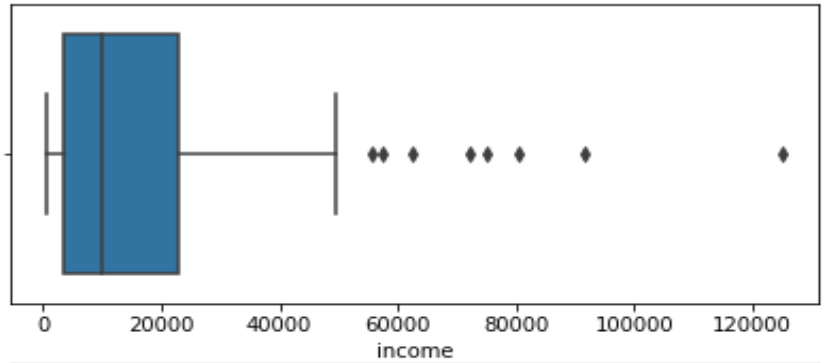
Comparison Of Health with gdp and Income with Child Mortality

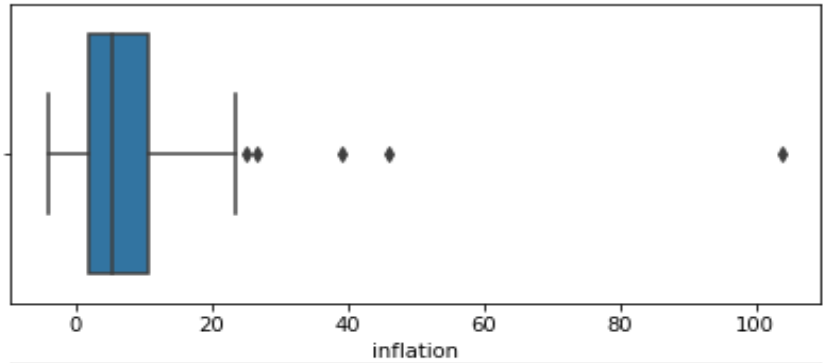


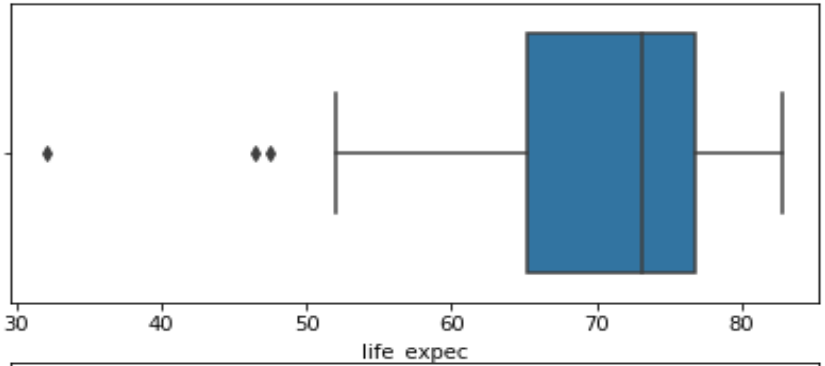


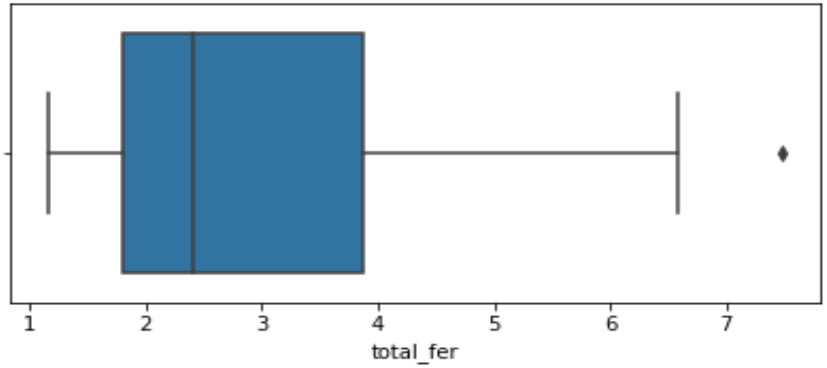
1. **Country with high GDP, Health investment is also high and vice-versa.**
2. **In the other side, Countries having a high Income, are reflecting a very low Child Mortality rate and vice-versa.**

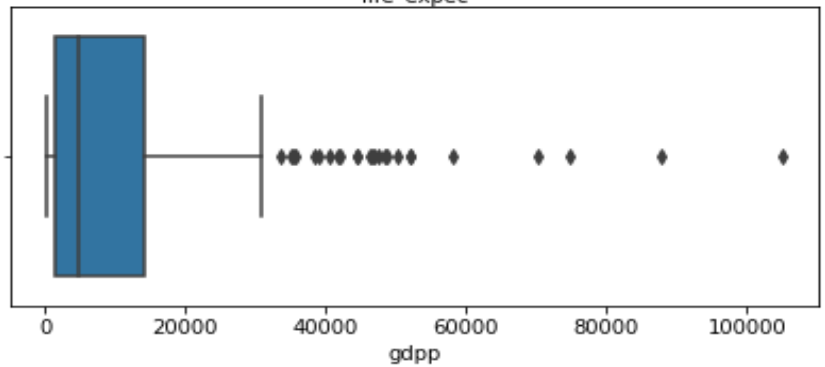
Univeriate Analysis for Numeric Variable (Box Plot)





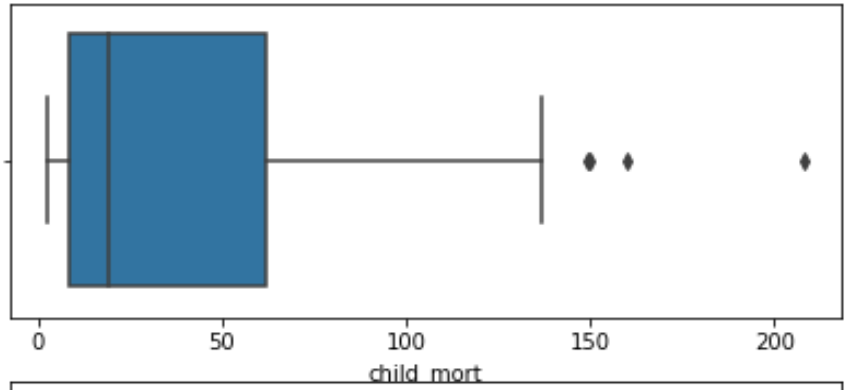


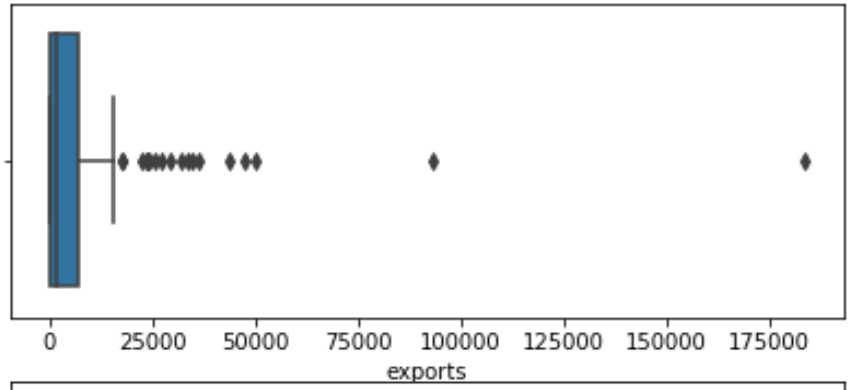


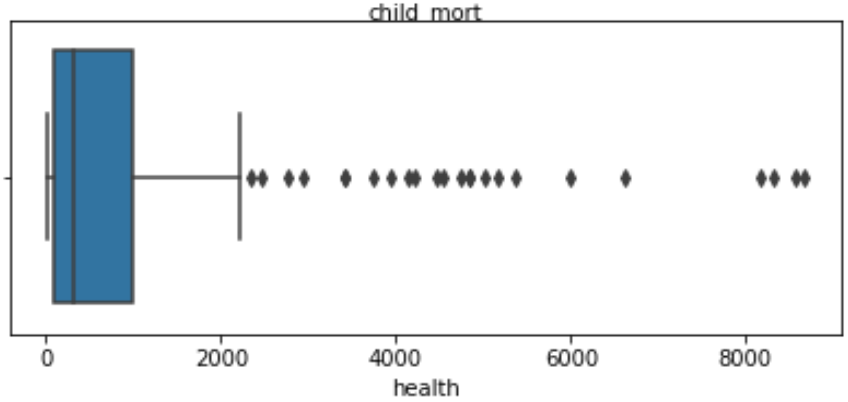


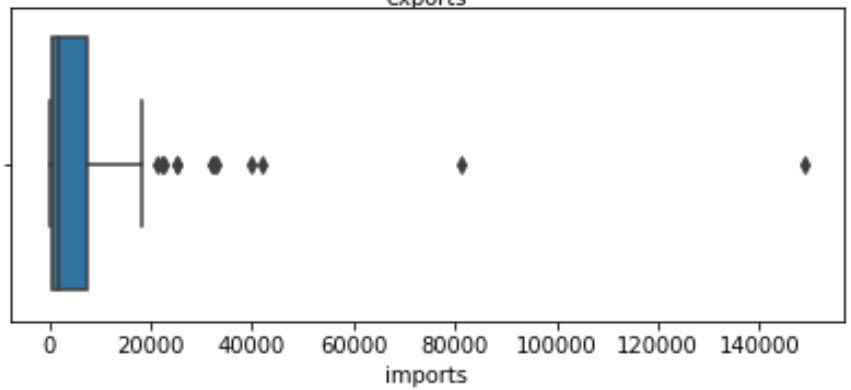
Contd/…..

Univeriate Analysis for Numeric Variable (Box Plot)







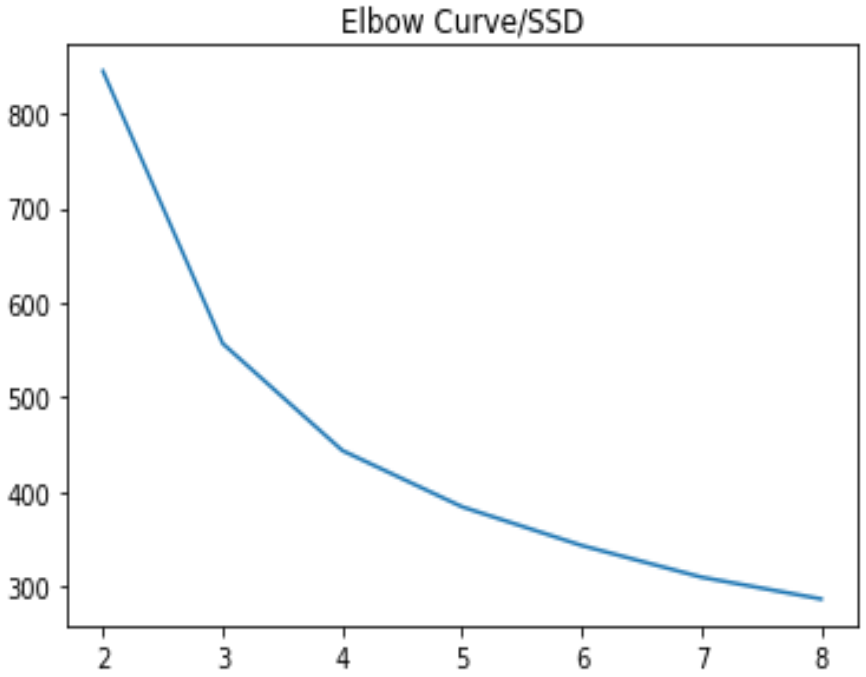


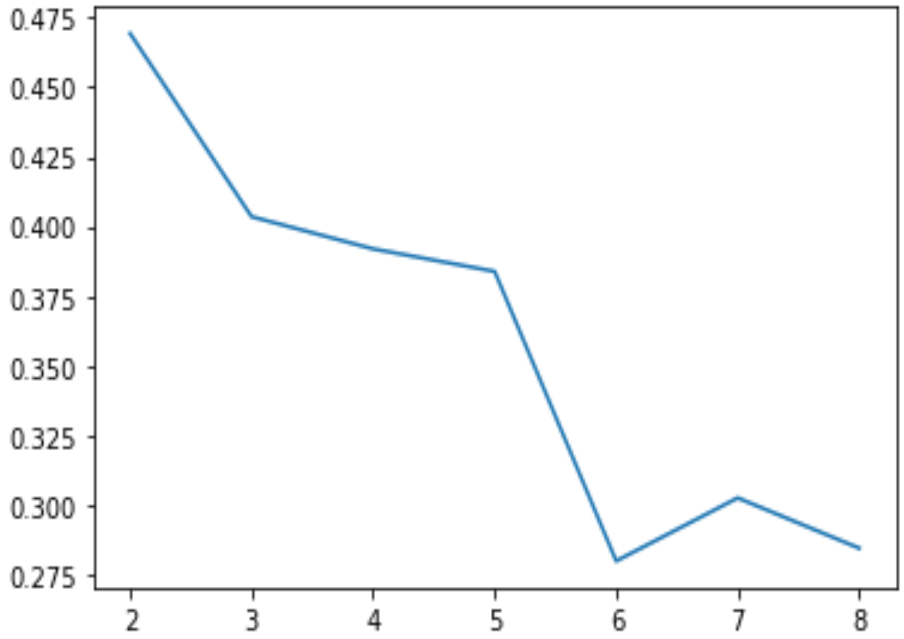
From the above boxplots it is clear that most of the variable have outliers at its upper range except one i.e. "life_expec". For Child Mortality it has very few outliers and it clearly indicate those countries, which we can intuitively say, which are in direst need of AID. Median for Child mort are lies in the 1st quartile range.

As almost all the column have outliers so before proceed for Modelling we have to teat the same. As our Data is small so we are not removing any outliers but make them **cap and treat only the upper limit to its 99 percentile**. As this range is a **Soft Range** so it will not alter our dataset so much.

But here we should keep in Mid that we will not do anything with our "child mort" column, because these outlined countries are our main target countries for which the NGO can prepare their investments.

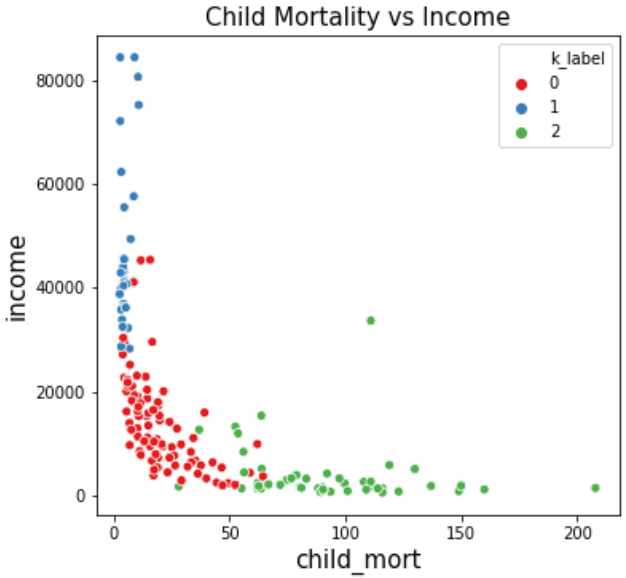
Applying K_Means Algorithm
Checking Elbow Curve and Silhouette Score for optimum Cluster

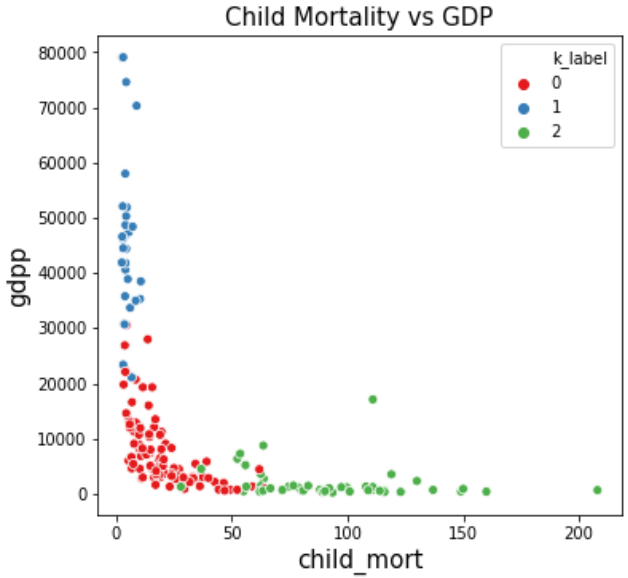


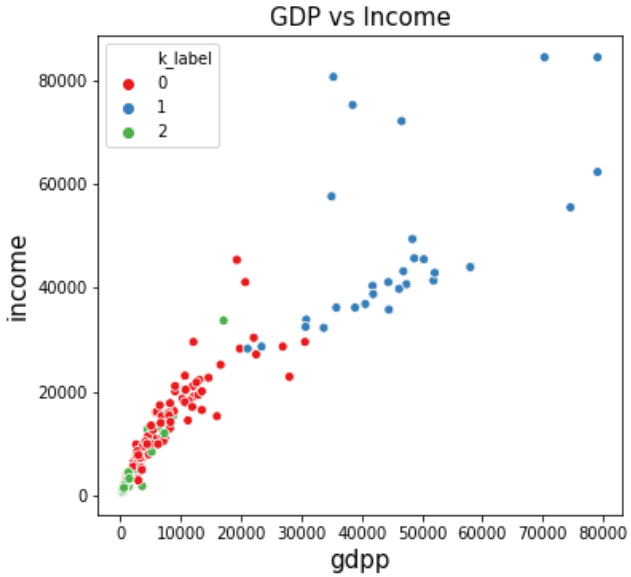


So after checking the **Elbow Curve** & **Silhouette Score**, both the curve are indicating the 1st bending at label 3. So we can choose our **final K as 3**. Because We have seen that in the **Sum of Squared Distance** Curve the 1st knee like bend has come at the level of 3. and in the silhouette score, the score of k=3 is also considerable for clustering. Though in the silhouette score for 2 is quite impressive from score of 3, but intuitively we can say from the point of our dataset that choosing the cluster number at 2 might not give us a optimum result that we can expect for. So for better clustering and better outcome of our analysis, we will choose our final number of cluster as 3.

Plotting the Cluster(K Means)

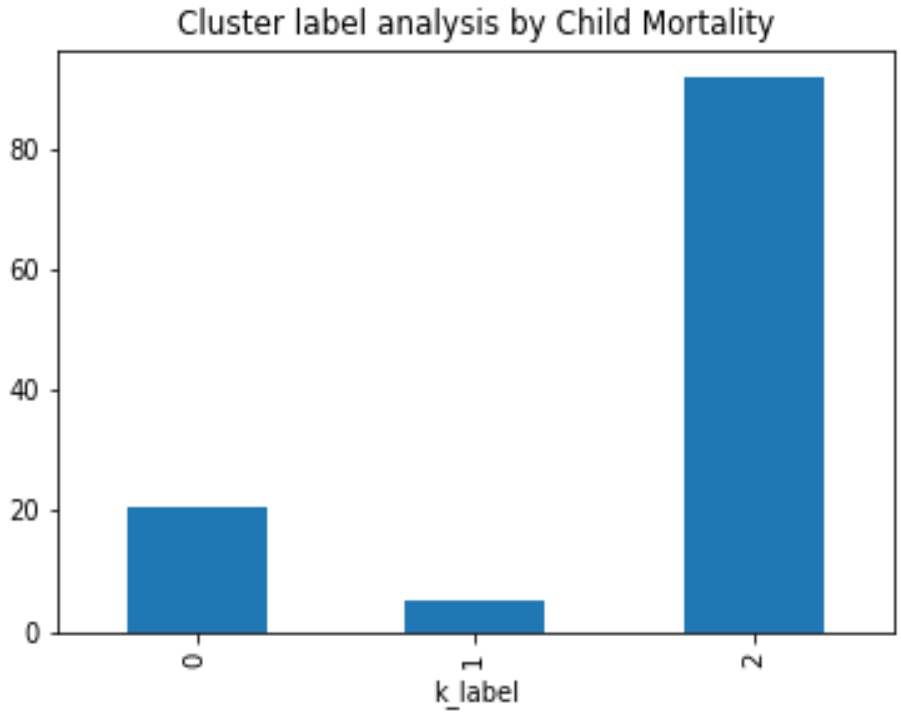


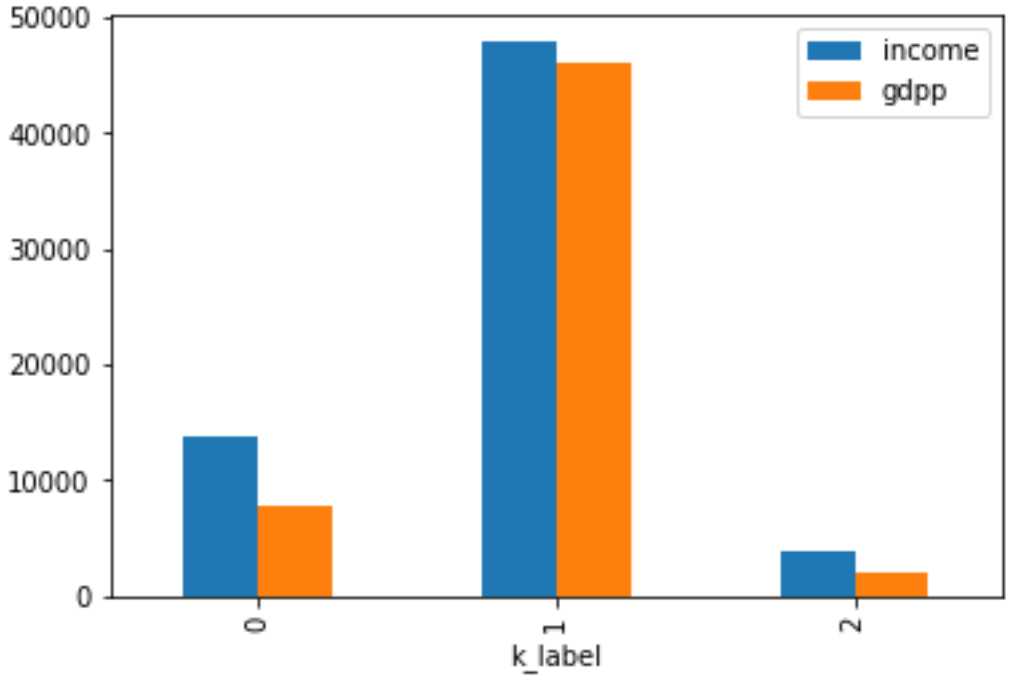




So after performing the K_Means algorithm and assigning the data point to its appropriate cluster number, and now by comparing the data points by the 3 major variables, we have seen that all the countries which are literally Socio-Economically very much poor and for which the child Mortality rate are much more high are **assigned to cluster number 2.** So we can intuitively say that these are the countries which are really in the direst need of AID and all the major investment should be planed for development by taking these countries into consideration. **We can be far more clear by moving ahead our analysis further.**

Cluster Profiling(K Means Algo)
Comparing the Cluster by Child Mortality, Income and GDP



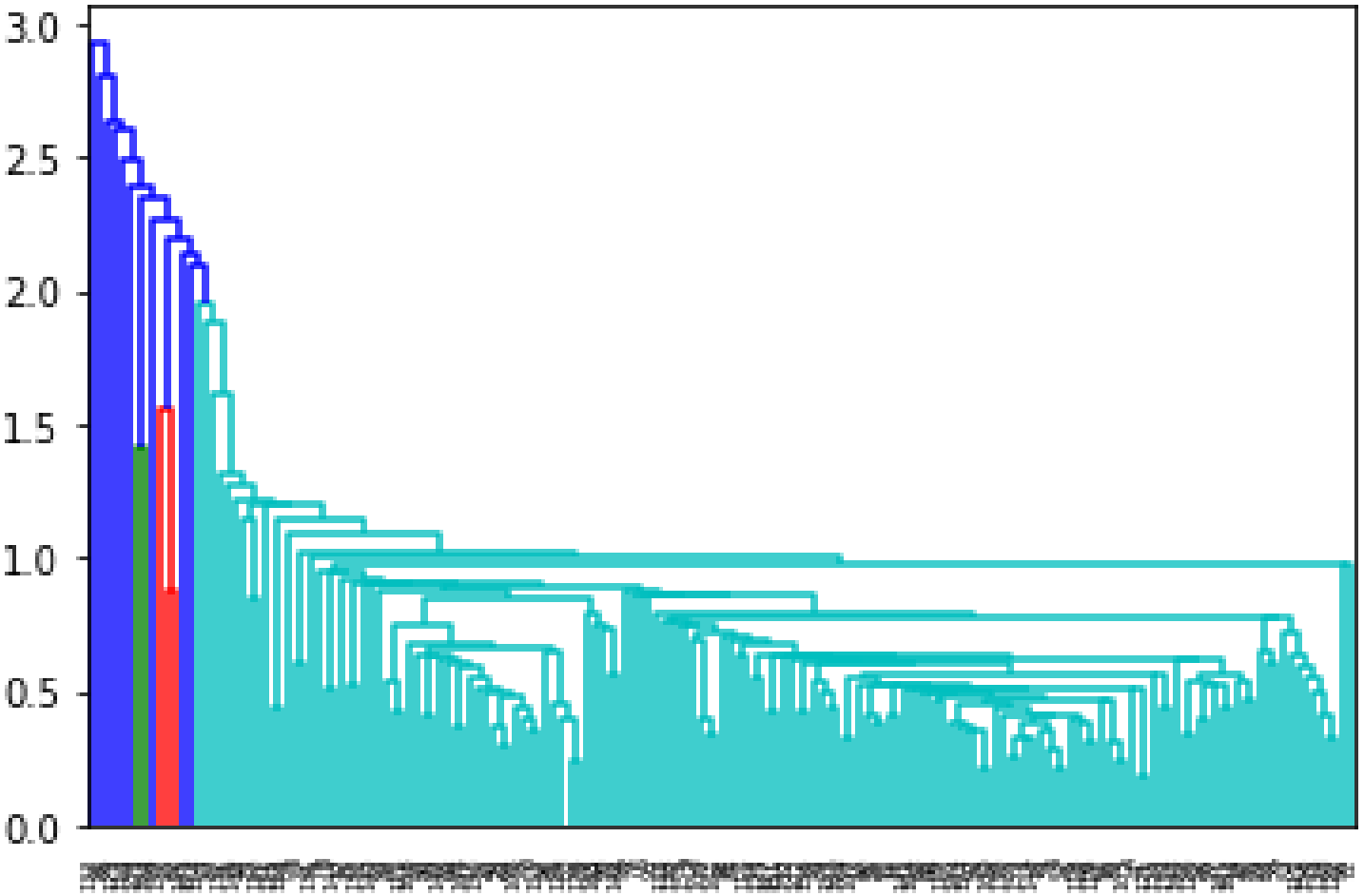


From the bar plot it clear that majorly effected countries by Child Mortality rate are all assign in Number 2. So these are our prime target countries.

From the above plot we can say that label 2 is our prime target countries. So now we have filter the dataset by label 2 and rest the all cases as of now for time being and lets check the Hierarchical clustering and lets see what kind of result it returns.

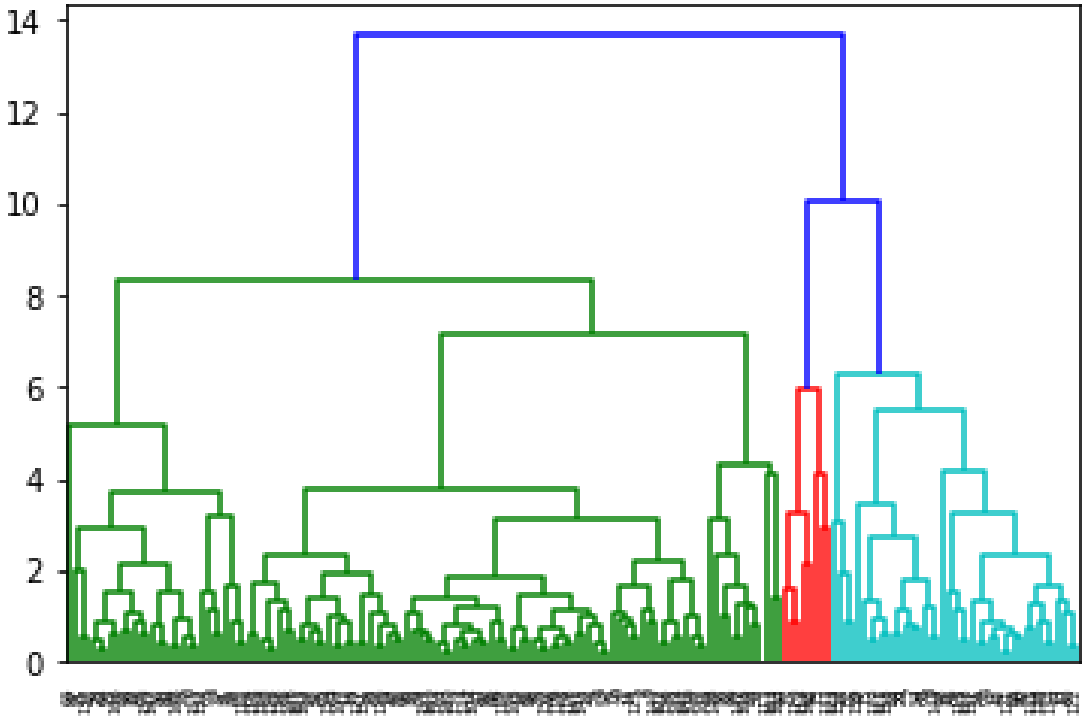
Applying of Hierarchical Clustering Method

Checking Single Linkage Dendrogram



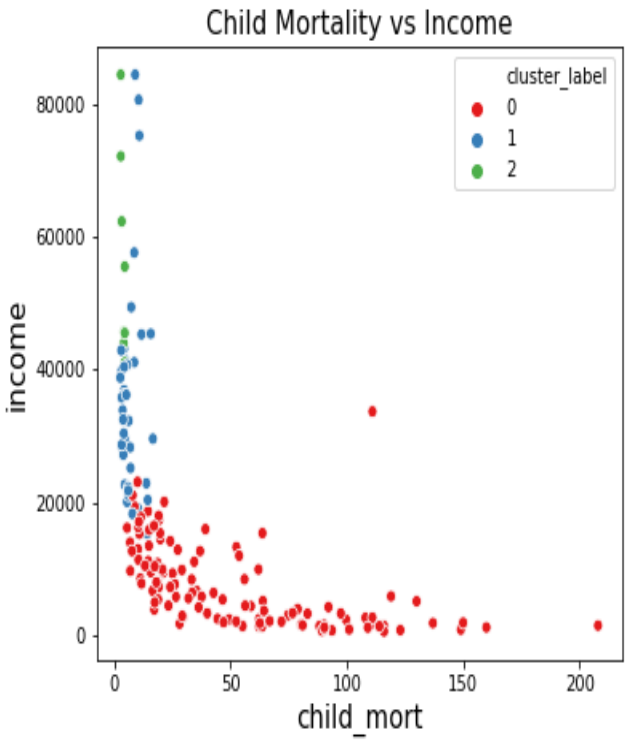
Applying of Hierarchical Clustering Method

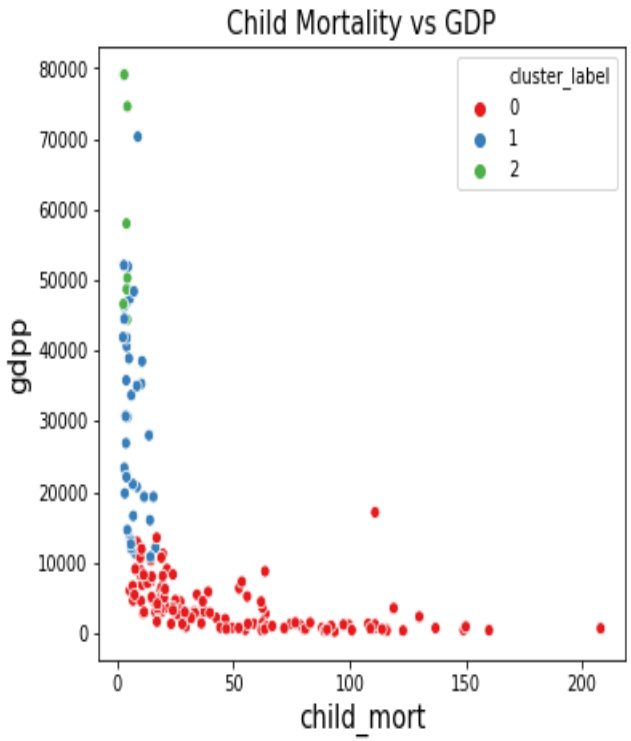
Checking Multiple Linkage Dendrogram

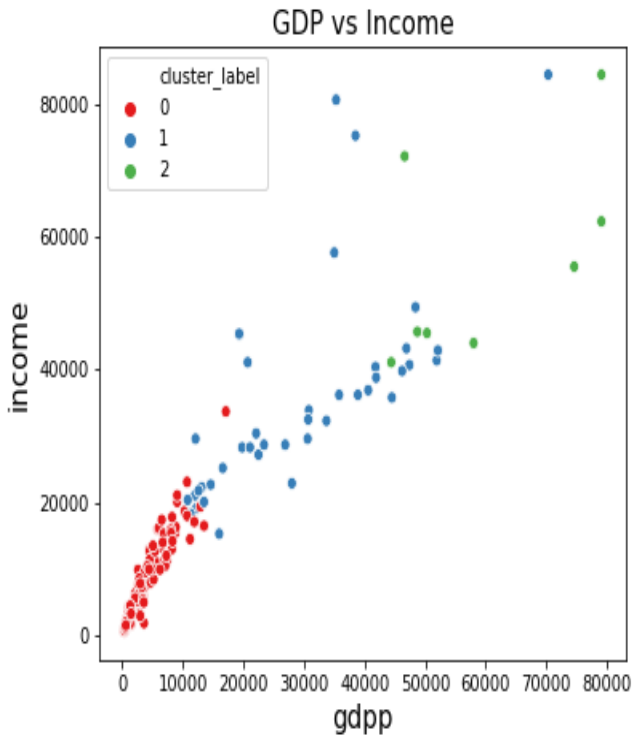


Now this dendrogram is far intuitive enough. we can clearly seen 3 different set of clusters by its colour green, red and sky. Majority of data point are filtered by green followed by sky and Red. So if we cut the tree and draw a horizontal line between 8 and 10 we will get 3 distinct set of cluster data which can give us far more interpretable understanding from our business point of view. **So based on this observation we have apply the cut tree method and set the cluster number 3 and assigning the label to the dataset**.

Plotting the Cluster(Hierarchical)



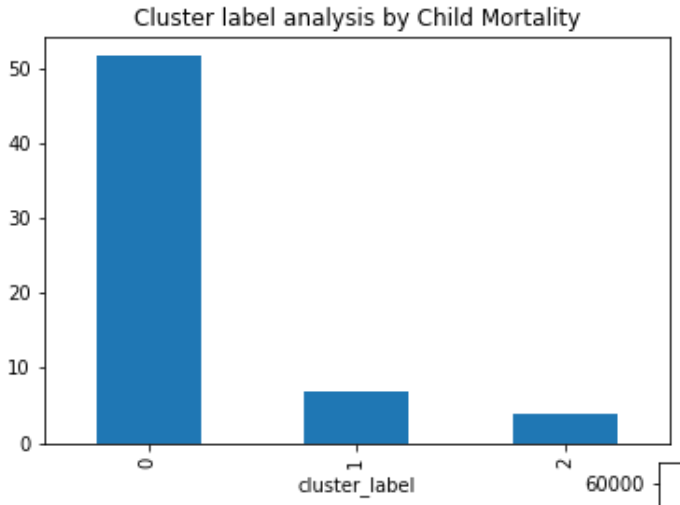


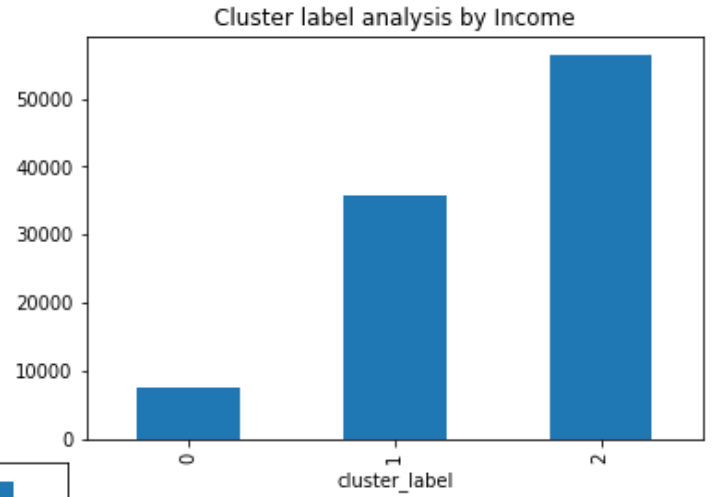


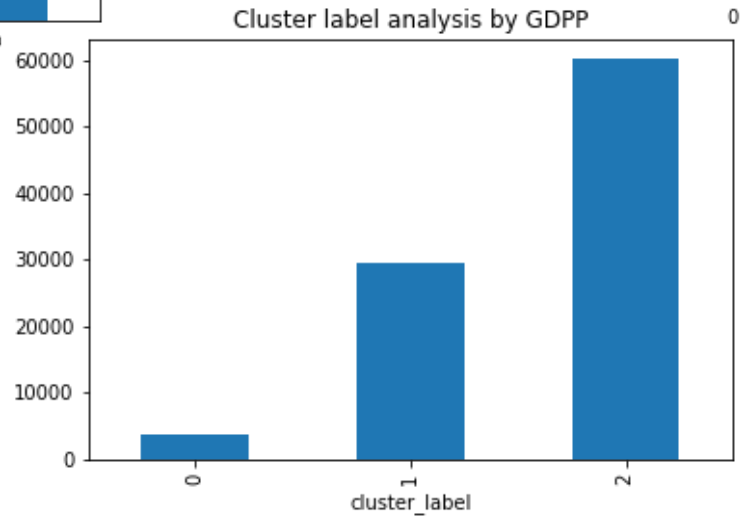
Observation are enclosed

After this visualization, we can understand that **all the countries which have high Child Mortality rate are assigned at 0 label, previously which we have seen at lebel 2**. And for GDP and Income plot, **the low GDP countries with low Income also have the same assignment at lebel 0**.

Cluster Profiling(Hierarchical)
Comparing the Cluster by Child Mortality, Income and GDP







What we have got in the K Means clustering, in Hierarchical clustering we have got a different result, which we can understand from this cluster profiling by 3 major variable Child Mortality, Income and GDPP. Now this times high Child Mortality rate countries are assigned at label 0 which was previously at label 2. And the same think happens for Income and GDPP variable also.

Finalizing the list of Countries

- So we choose to finalise our countries based on K Means algorithms.

- After filtering the data by ***'k_label'*** we have got the top five countries according to Child Mortality Rate, which are really in direst need of AID, are - ***Haiti, Sierra Leone, Chad, Central African Republic, Mali.***

- Apart from that according to Income status and GDPP, we can also consider countries like - ***Burundi, Congo, Dem. Rep.*** This countries are also low in Income level and low GDP.

Conclusion And Recommendation .....

After perform the EDA and the K-Means & Hierarchical clustering we can come to the conclusion that follows...

- **For a countries major development is basically depends on some major factors i.e. Income per capita, GDP, Health infrastructures, Child Mortality etc.**
- **If a country's GDP is low and also per capita Income is low then surely it can say that this particular country is Socio-Economically very week, where as if these two factor for a country is high then this country hold a strong position socio-economically. These are the major two factors that can play a significant role for any country's development.**
- **Apart from this if the Child Mortality rate is very high for a country, then it can surely say that in terms of Health infrastructures this country belongs to a non healthy condition.**
- **So we can recommend ate the NGO that they can take into consideration of 1st 5 countries filtering by *Child Mortality* rate i.e. *Haiti, Sierra Leone, Chad, Central African Republic, Mali*. Because from our basic Understanding we know that if a country's child mortality is high then obviously their health infrastructure is very much poor. So for a NGO this could be their prime responsibility to keep ahead their hand for those countries' health development and plan their amount of investment accordingly.**
- **Apart from that the NGO can also look after for the countries which GDP and Income status is very low like - *Burundi, Congo, Dem. Rep.* They also can look after all those countries which are being cluster at label 0 and plan their investment accordingly. Because all the countries which are been clustered at label 0, they are either Health and Socio-economically undeveloped or poorly developed. But the major focus to be surrounded for all those names which we have mentioned earlier.**