CLUSTERING ASSIGNMENT

BY: ANKIT GUPTA

ASSIGNMENT DETAILS AND PROBLEM

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

OBJECTIVES OF THE CASE STUDY

There might be some subjectivity in the final number of countries that you think should be reported back to the CEO since they depend upon the preceding analysis as well. Here, make sure that you report back at least 5 countries which are in direst need of aid from the analysis work that you perform.

FILES USED AND APPROACH

We have been provided with one dataset for this assignment.

1. Country_data

APPROACH USED

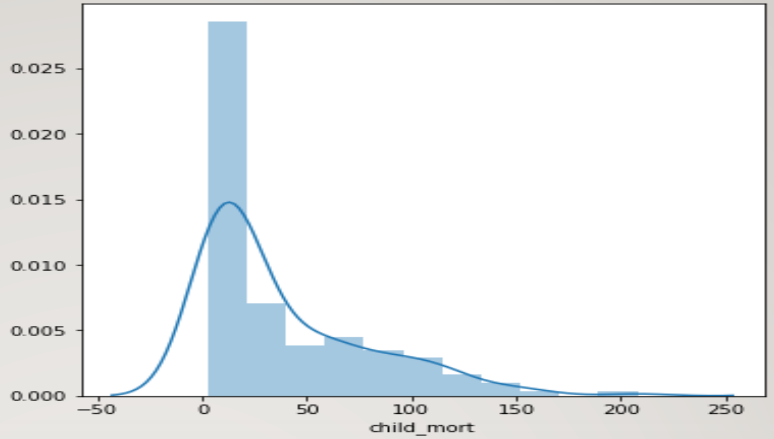
1. Data Understanding.
2. Data preparation for clustering.
3. Outlier treatment
4. Hopkins check
5. b. Clustering
6. K-MEANS
7. Run K-Means and choose K using both Elbow and Silhouette score
8. Run K-Means with the chosen K
9. Visualize the clusters
10. Clustering profiling using "gdpp, child_mort and income"
11. Hierarchical Clustering
12. Use both Single and Complete linkage
13. Choose one method based on the results
14. Visualise the clusters
15. Clustering profiling using "gdpp, child_mort and income"

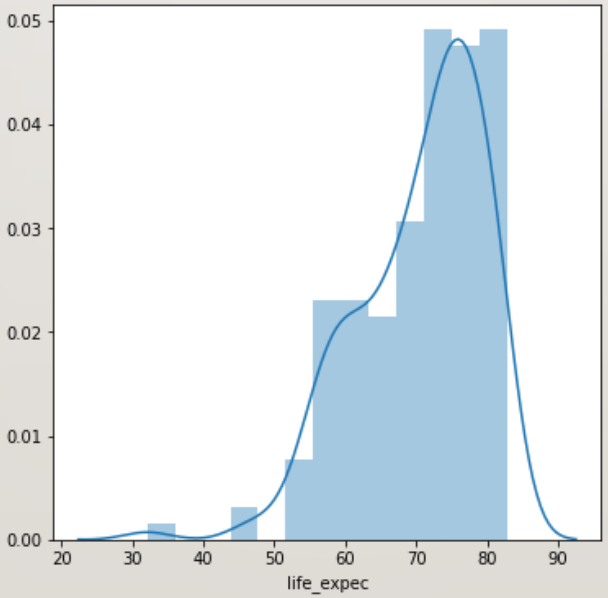
Univariate Analysis:

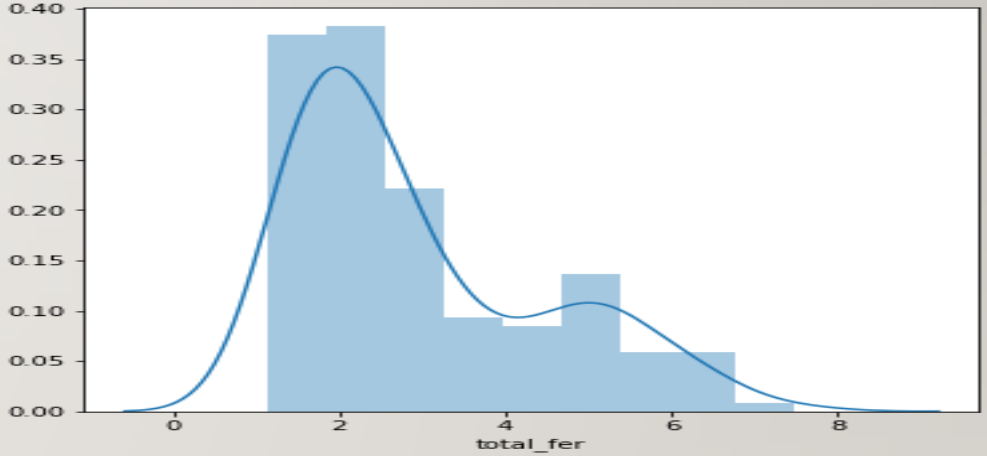
child mortality is more on poor countries

life expentancy is high in between 60 to 80

total fertilation is not normally distributted but maximum is 2





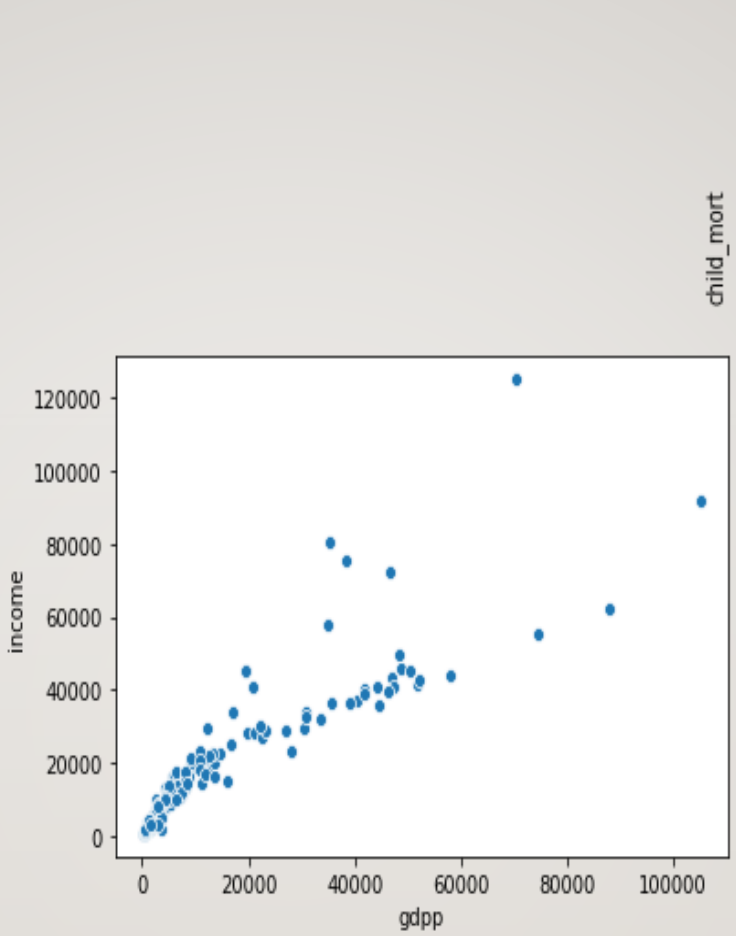


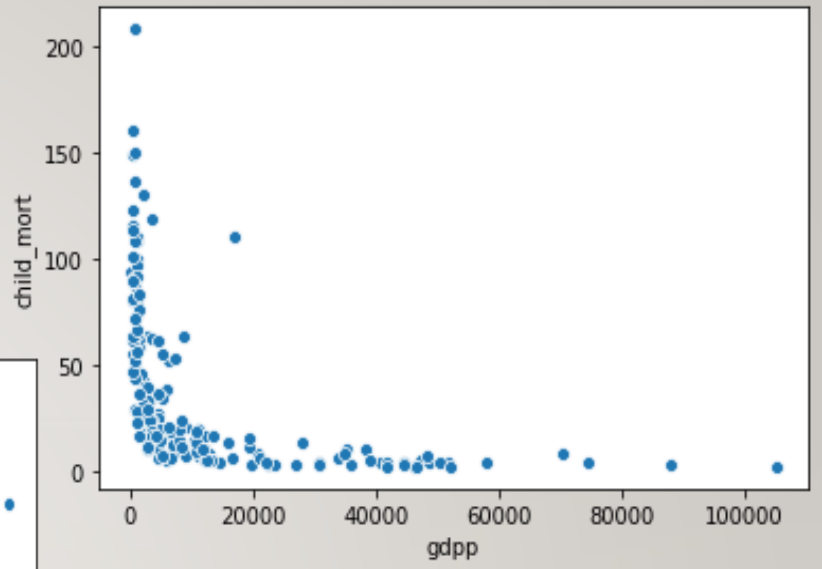
BIVARIATE ANALYSIS:

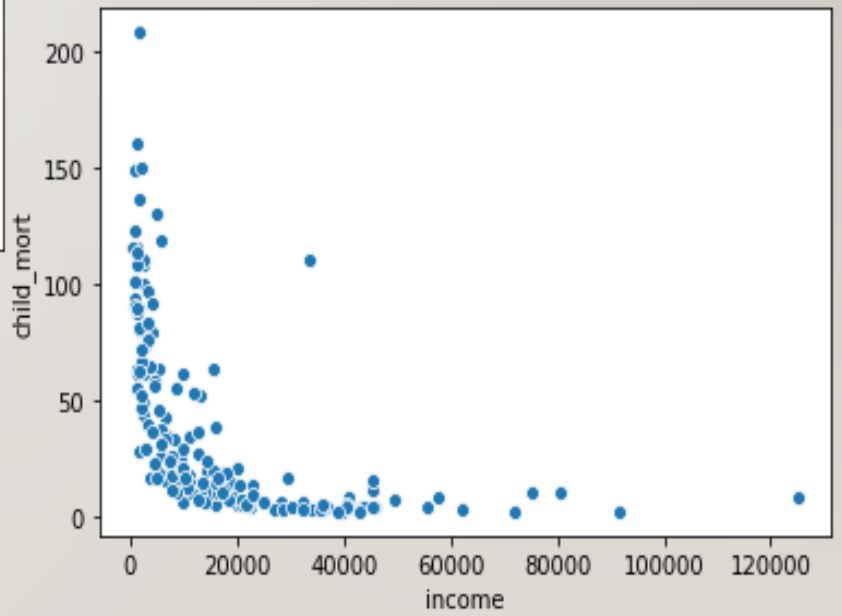
as we seen low income countries have low gdp

the countries in which there is low gdp child mortality is maximum

the countries where income is low is high child mortality



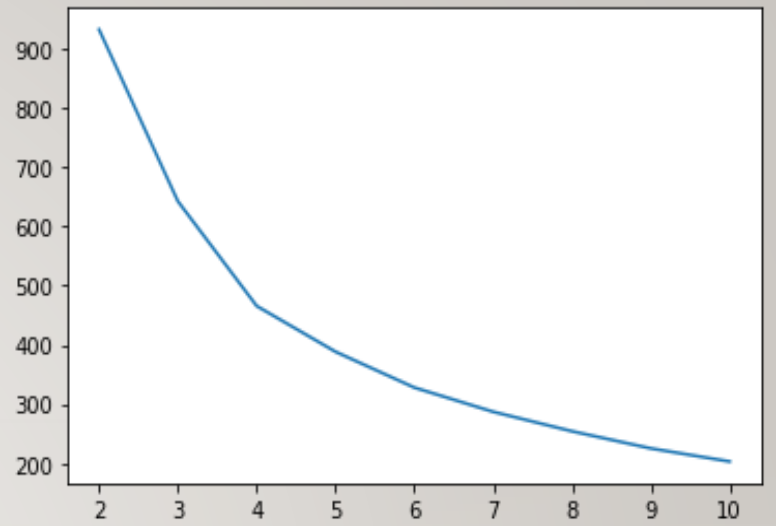




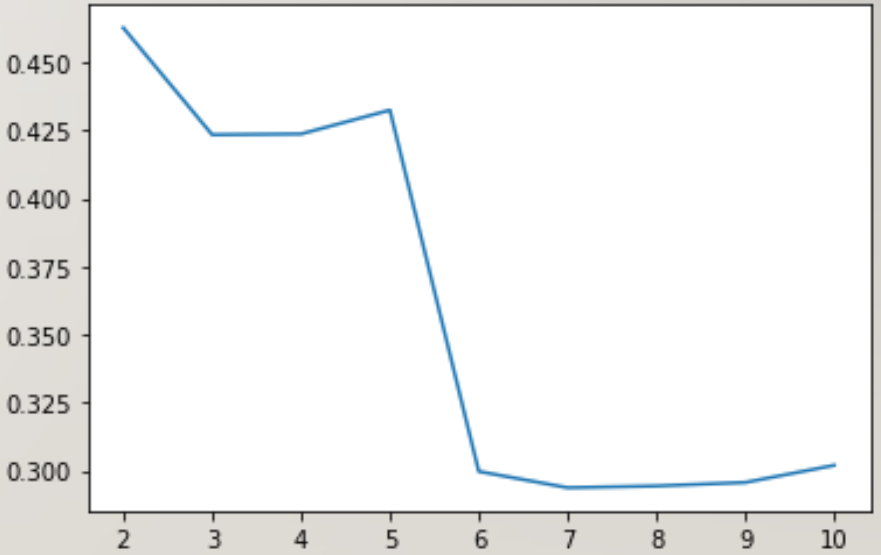
FOR CHOOSING OF K:

As we can see that both elbow curve and silhouette score denotes that k =3 is the bets clustering value for analysis

ELBOW CURVE



SILHOUETTE ANALYSIS



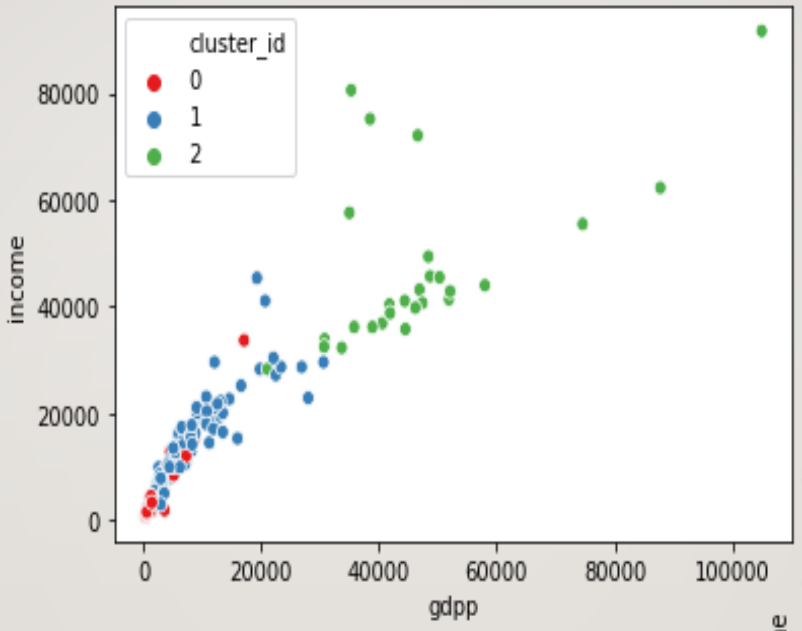
Plotting the cluster with respect to data :

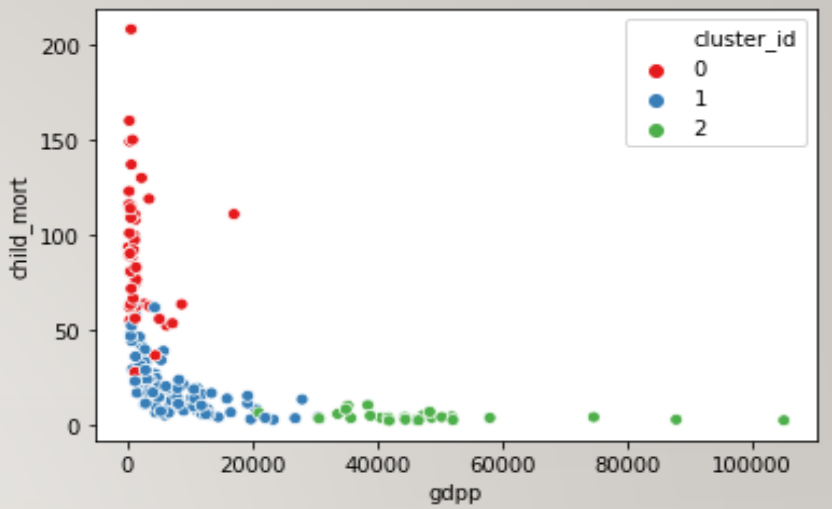
we use gdp, income,child mortality

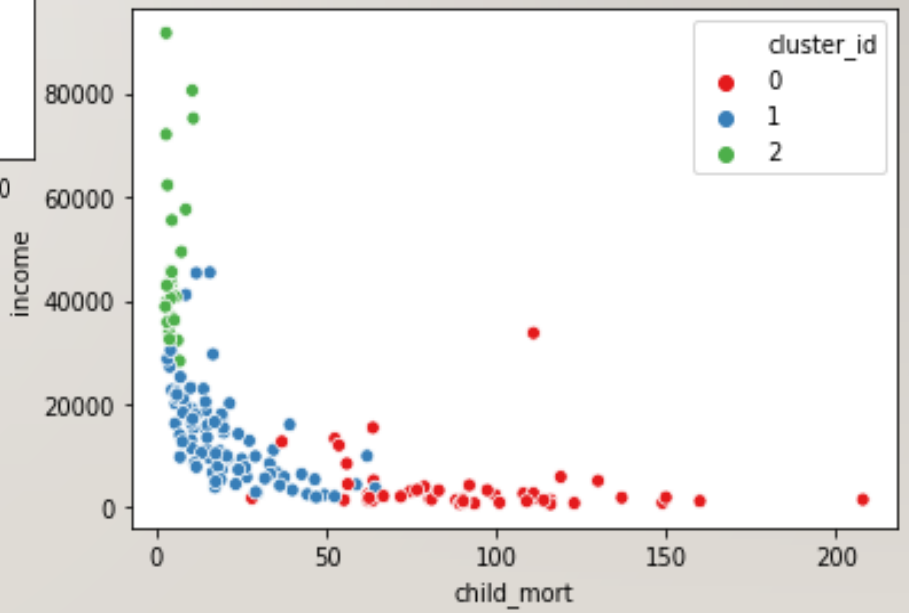
as we seen from the plot cluster id 2 has maximum gdp and income

child mortality is more in cluster 0 as it has low gdp and high child mortality

child mortality is high in cluster 0 as it have low income

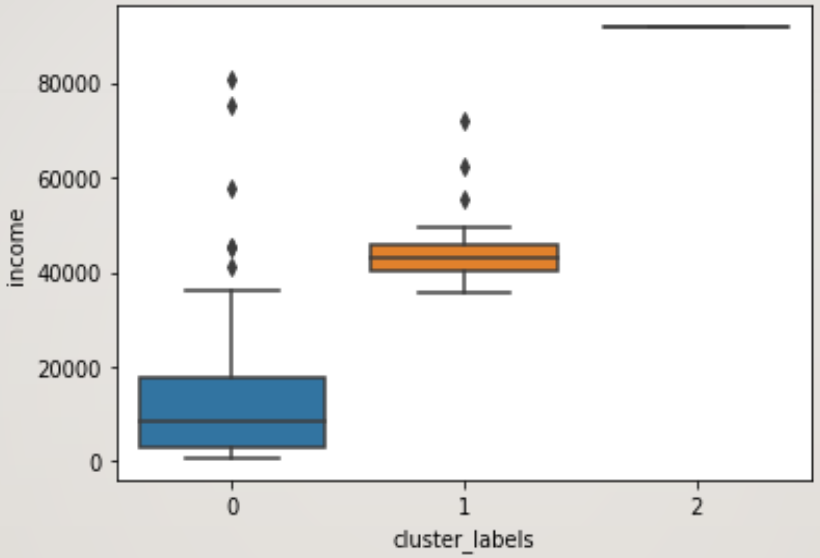


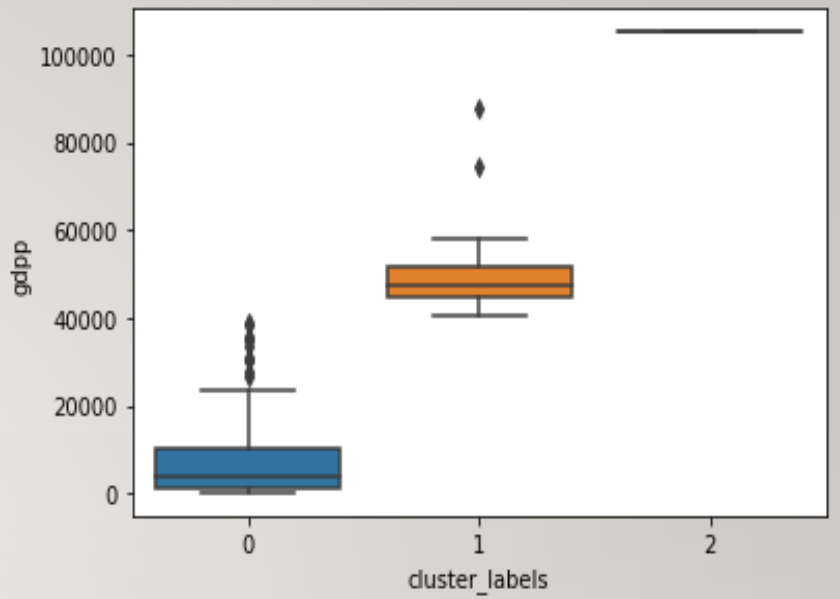


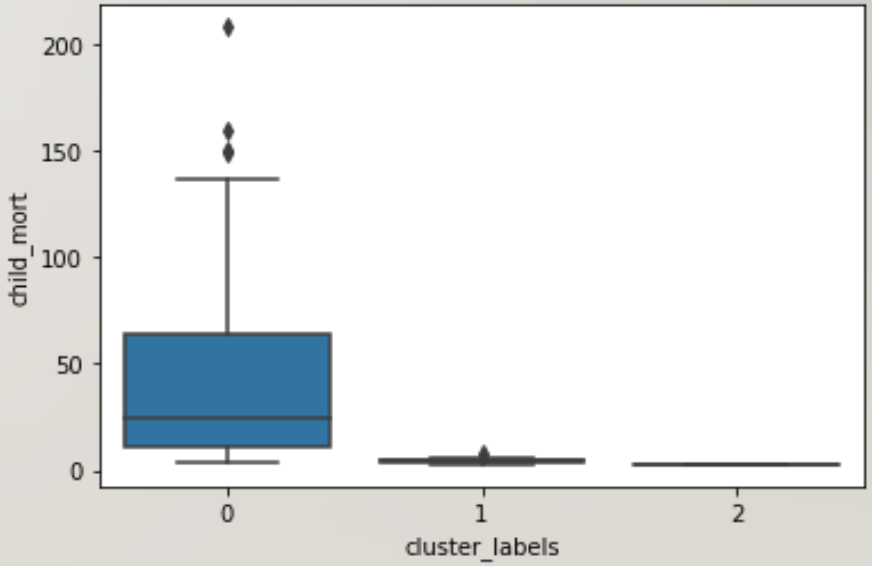


HIERARCHICAL CLUSTERING:

we have to focus on cluster label o countries in which there is low gdp low income and high child mortality as these type of countries need Direct aid from HELP organistion







Conclusion and Recommendation

we have to focus on cluster label o countries in which there is low gdp low income and high child mortality as these type of countries need
Direct aid from HELP organistion

Top 5 countries which need direct aid on the basis of conclusion are as follows:
1.HAITI
2.SIERRA LEONE
3.CHAD
4.CENTRAL AFRICAN REPUBLIC
5.MALI

THANK YOU