Clustering Assignment

Submitted by Kratika Sharma

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 program, 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.

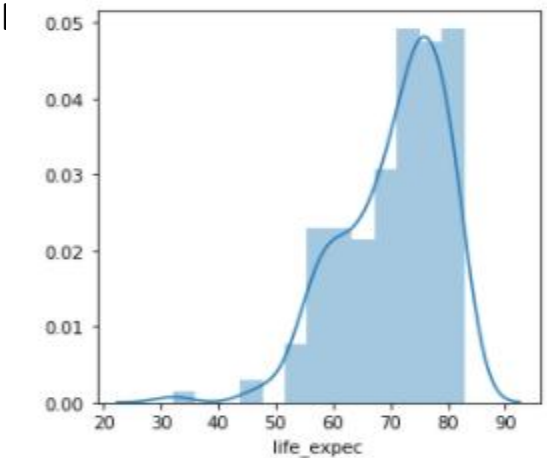
We need to categorize the countries using some socio-economics and health factors that determined the overall development of the country.

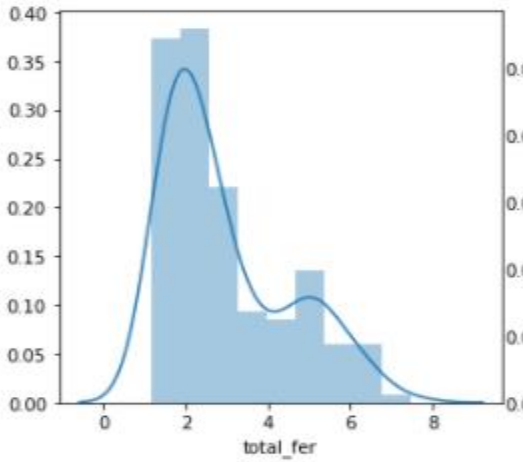
Then we need to suggest the countries which CEO need to focus the most.

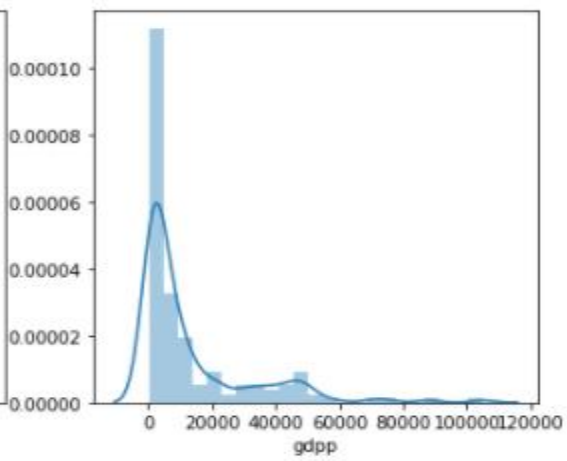
EDA

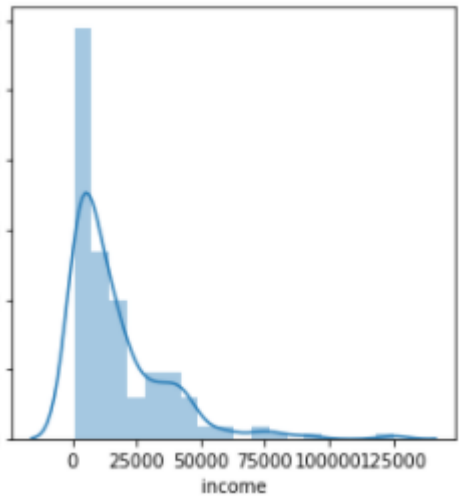
We create distplot for all the fields and observe that

- Child mort, Import, Export, Health are normally distributed. Hence not able to see any pattern
- Here we are able to see pattern for cluster in income, total_fer, life_expec, gdpp



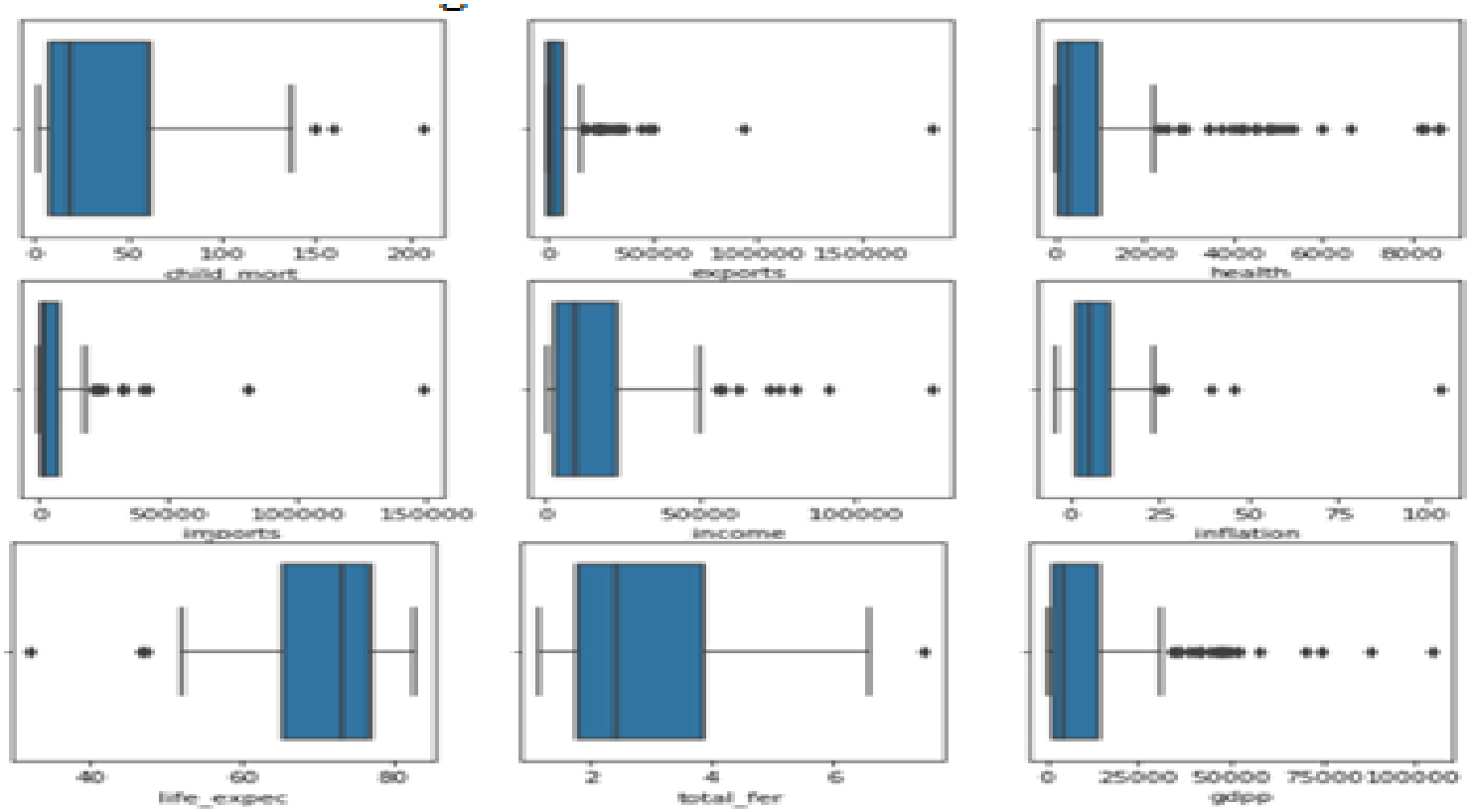






Detecting Outliers and Treatment

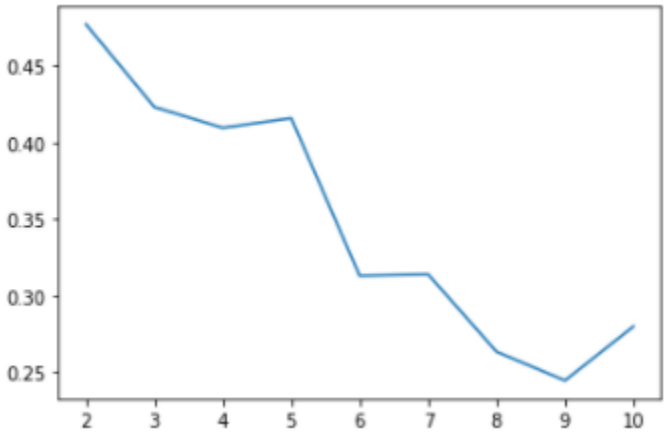
- Child Mort, Inflation: We should not deal with upper range outliers but we can deal with lower range outliers.
- life_expec : We should not deal with lower range outlier
- Outliers Treatment has been done for export, Import, health, total_fer, gdpp column
- We did outlier treatment using **soft scaling method**



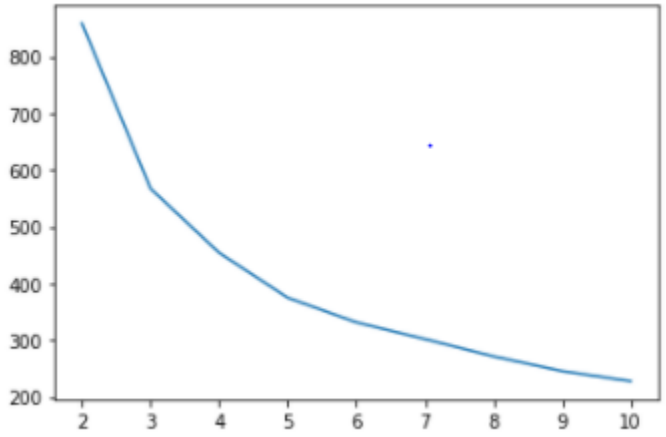
K Means Clustering

- According to Silhouette Analysis 2,3 and 5 is higher then all other cluster. As per industry standards 2 cluster are not considered.
- Looking at this elbow curve we can consider number of cluster as 3

Silhouette Analysis

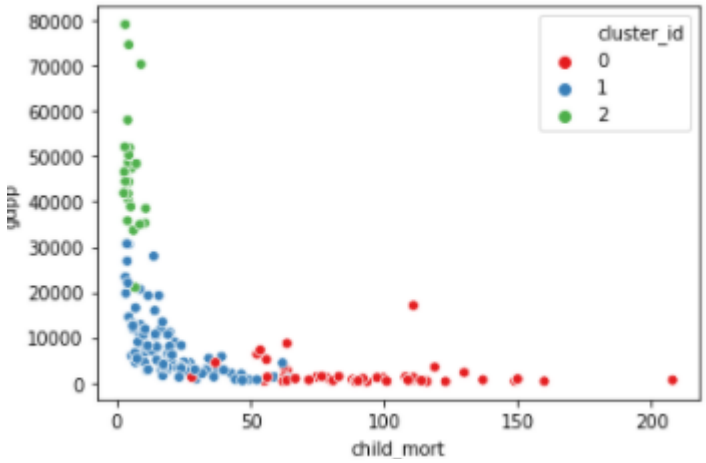


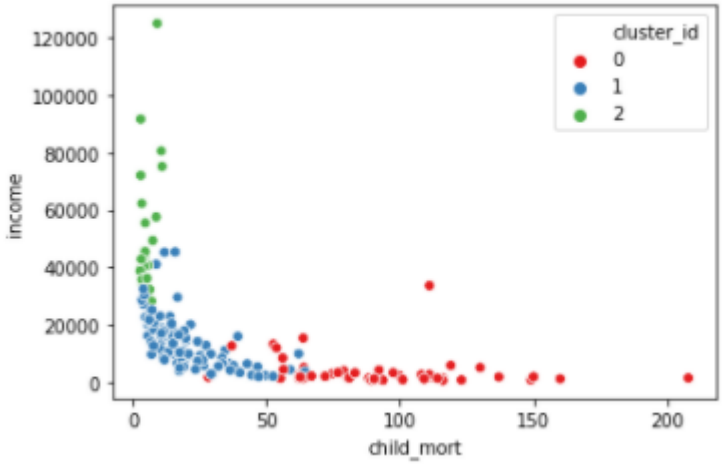
Elbow Curve

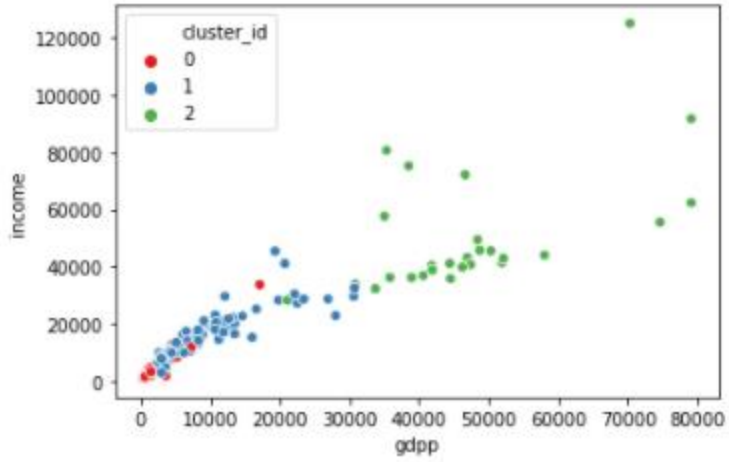


Plot

Here we able to identify proper three cluster for child mort and gdpp, gdpp and income, gdpp and child mort and income





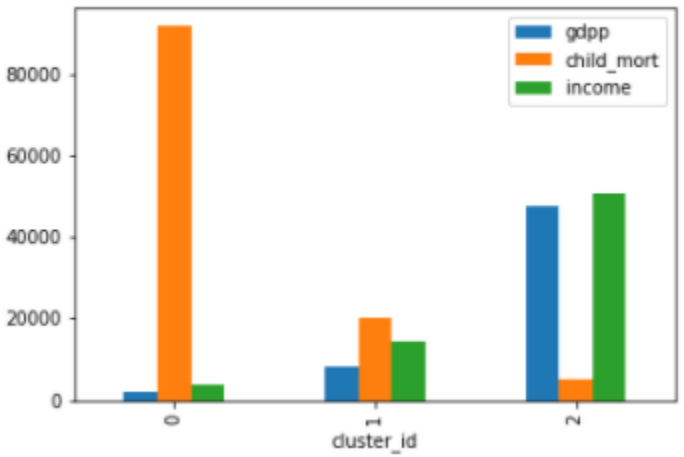


Cluster Profiling and Countries

Cluster profiling has been done based on Low GDPP , Low income and high child mort.

- From mean values we observe that where ever income and gdpp is low child mortality is high
- Looking at bar graph we can clearly say that 0th cluster has low income , low gdp , high child mort hence we can conclude that countries belong to cluster 0 are in the direst need of aid.

cluster_id	gdpp	child_mort	income
0	1911.400833	91.610417	3897.354167
1	8226.869565	20.177174	14169.456522
2	47476.888889	5.092593	50833.333333

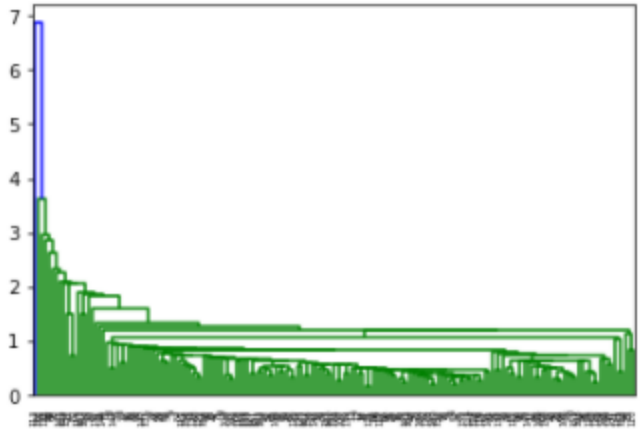


Below are the countries are in the direst need of aid:

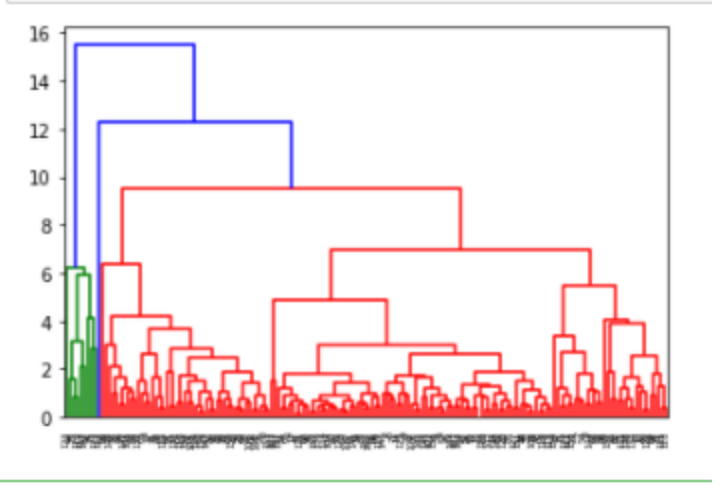
1. Haiti
2. Sierra Leone
3. Chad
4. Central African Republic
5. Mali

Hierarchical Clustering

- Single linkage : Here Single linkage doesn't making any sense

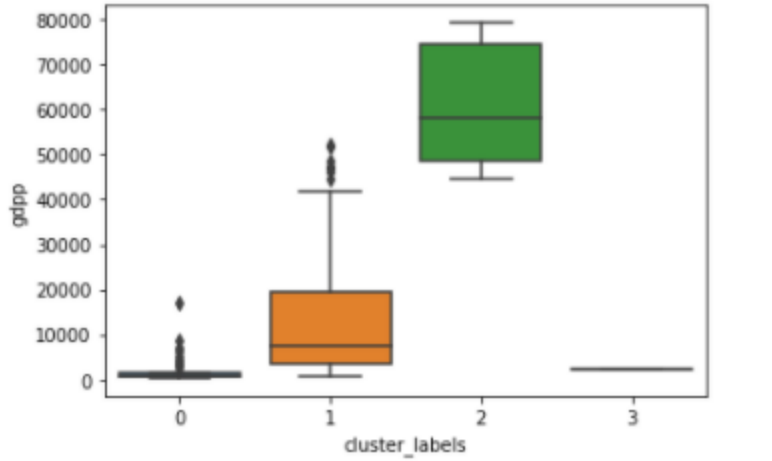


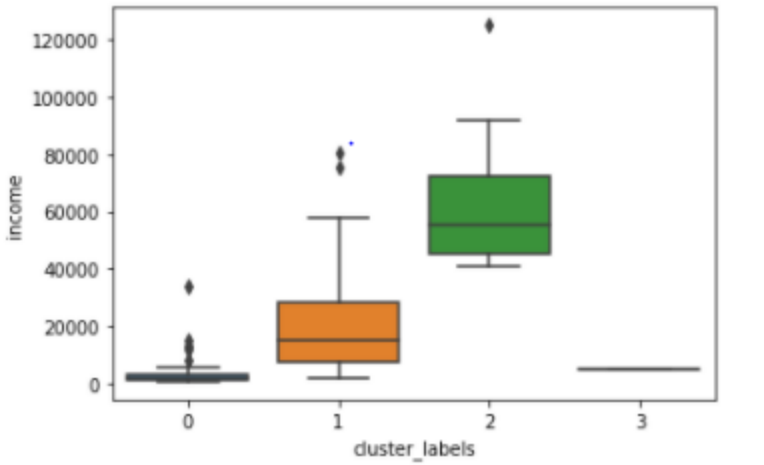
- Complete linkage : Looking at this dendogram we can go with either 3 clusters or 4. I am taking number of cluster as 4 here.

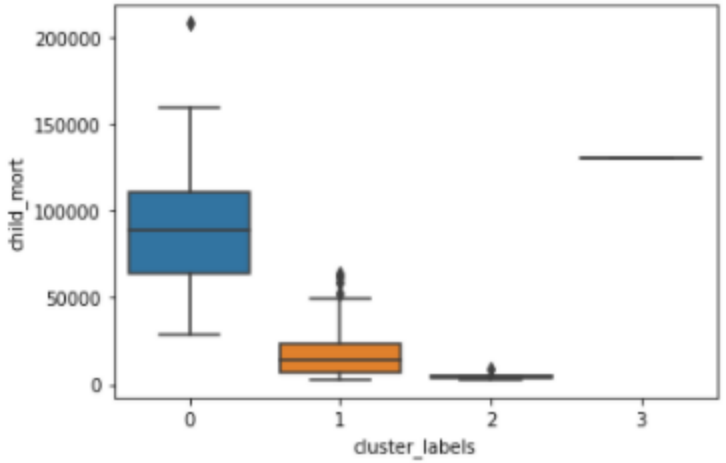


Plots

We can observe here for cluster 0 has high child mort , low income and low gdpp





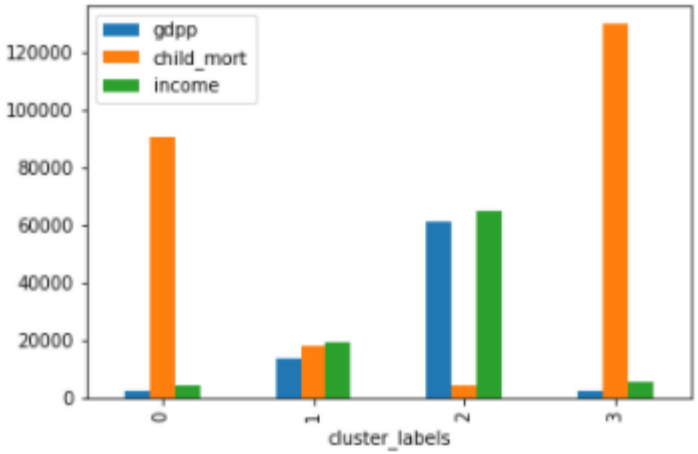


Cluster Profiling and Countries

Cluster profiling has been done based on Low GDPP , Low income and high child mort.

- From mean values we observe that where ever income and gdpp is low child mortality is high
- Looking at bar graph we can clearly say that 0th cluster has low income , low gdp , high child mort hence we can conclude that countries belong to cluster 0 are in the direst need of aid.

cluster_labels	gdpp	child_mort	income
0	1902.494468	90793.617021	3870.702128
1	13524.290909	17765.454545	19029.000000
2	61230.666667	4400.000000	64766.666667
3	2330.000000	130000.000000	5150.000000



Below are the countries are in the direst need of aid:

1. Haiti
2. Sierra Leone
3. Chad
4. Central African Republic
5. Mali

Thanks You