Clustering Assignment

(Segmentation of countries for humanitarian ADD by an NGO using K-Means & Hierarchical Clustering)

By: Roushan Kumar

Background:

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.

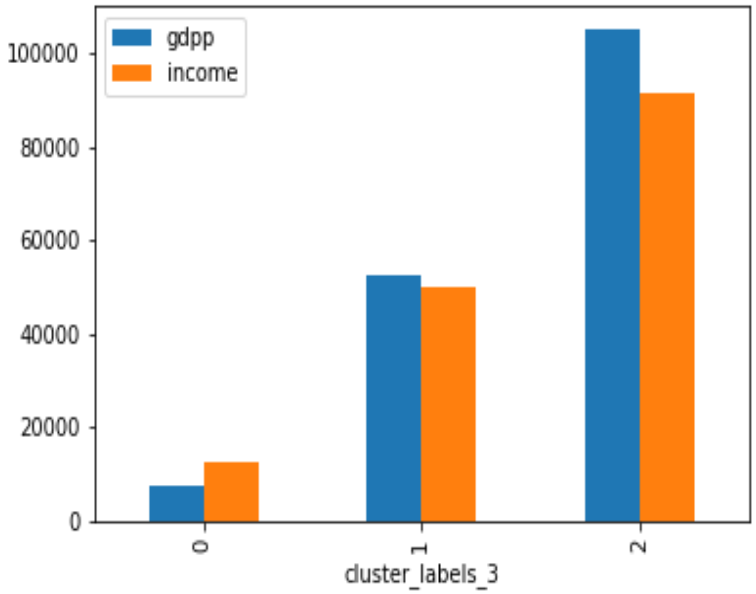
After the recent project that included a lot of awareness drives and funding programmes, they have been able to raise around $ 10 million. Now the CEO of the NGO needs to decide how to use this money strategically and effectively. The significant issues that come while making this decision are mostly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

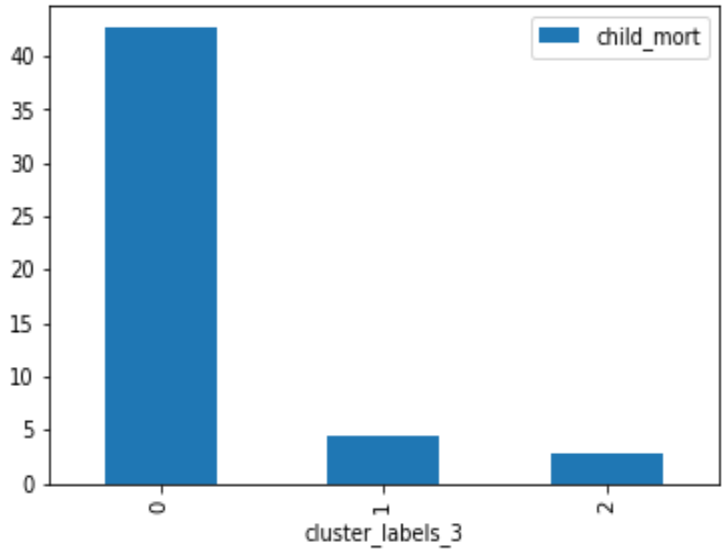
Technical Approach

- Check for missing value, and treatment
- Check for outlier and treatment
- Perform the basic EDA to find the variability and distribution of the data, so as to identify if we need t scaling the data
- Data Scaling if necessary
- Use Hopkins Method to check if the dataset is good enough for a cluster analysis
- Using Hierarchical clustering to identify the optimal cluster value.
- Use Silhouette and Elbow method to validate the optimal cluster values.
- Use K-Means Cluster method to build the final cluster model.
- Analyze the cluster that is representing the countries that will solve the Business Problem.
- Present the final report

Cluster Summary

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 0 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.





Final list of top 5 Undeveloped Country from Cluster

We conclude on the top 5 list of countries from the final cluster (Under Developed Countries) based on the median values of gdpp, income and child_mort.

We filtered countries with.

1. Lowest gdpp
2. Lowest income and
3. Highest child_mort

:

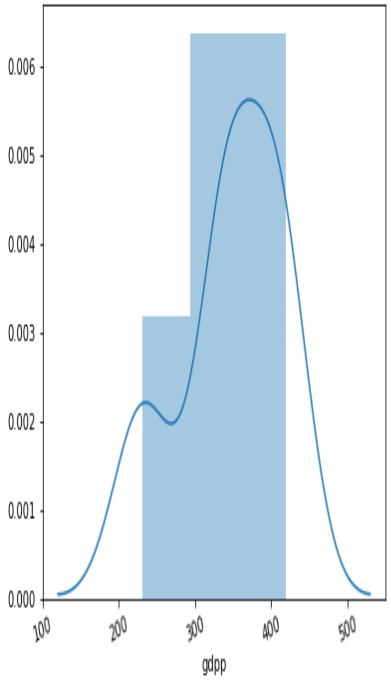
	country	gdpp	income	child_mort
26	Burundi	231	764	93.6
37	Congo, Dem. Rep.	334	609	116.0
112	Niger	348	814	123.0
132	Sierra Leone	399	1220	160.0
106	Mozambique	419	918	101.0

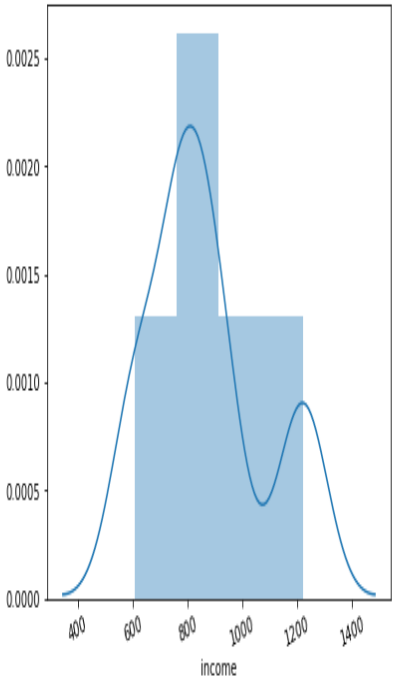
Statistics

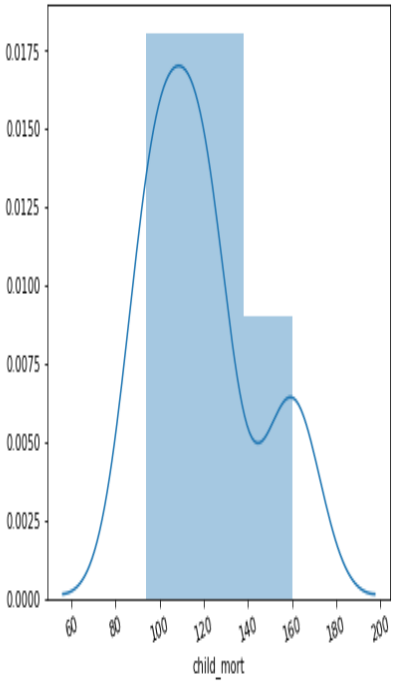
Min GDPP =231
Min INCOME =609
Min CHILD_MORT =90

Max GDPP = 553
Max INCOME = 1610
Max CHILD_MORT = 160

Median GDPP =432.5
Median INCOME =974
Median CHILD_MORT =107







	gdpp	income	child_mort
count	5.000000	5.000000	5.00000
mean	346.200000	865.000000	118.72000
std	73.332803	227.580755	25.85908
min	231.000000	609.000000	93.60000
25%	334.000000	764.000000	101.00000
50%	348.000000	814.000000	116.00000
75%	399.000000	918.000000	123.00000
max	419.000000	1220.000000	160.00000

Descriptive statistics of top 5 recommended countries

Conclusion

- We 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, we have below the top countries under our 'Under Developed 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.

Top 5 countries recommended for financial Aid

	country	gdpp	income	child_mort
26	Burundi	231	764	93.6
37	Congo, Dem. Rep.	334	609	116.0
112	Niger	348	814	123.0
132	Sierra Leone	399	1220	160.0
106	Mozambique	419	918	101.0

Thank You