Clustering Assignment

Soumya Prakash Parida

HELP INTERNATIONAL

Objective

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 programs, 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.

Objective

- To categorize the countries using some socio-economic and health factors that determine the overall development of the country.
- To suggest the countries which the CEO needs to focus on the most.

Analysis Process Steps

EDA

Load & Inspect Data
Format & Convert data types
Univariate/Bivariate Analysis.

1

2

Data Preparation

Outlier Observation
Outlier Treatments (if any)
Hopkins Measure

K-Means Clustering

Prepare and scale data
Choose optimum K
Run K-Means
Profile Clusters

3

4

Hierarchical Clustering

With Single Linkage
With Complete Linkage
Choose K, Run clustering
Profile Clusters

Choose Countries

Based on any Clustering output & analysis.
Choose the best 5 countries to receive aid.

5

EDA

From the outset, GDPP and Child Mortality rate look like the major 2 features that can be tagged as independent. But the matrix shows a highly correlated data. Hence, we are going to run the clustering on all the features available.

GDPP

The GDPP seems to be a dominant feature with 4 other features correlated with it.

Income

Although income is correlated with GDPP, it is slightly independent from GDPP calculation.

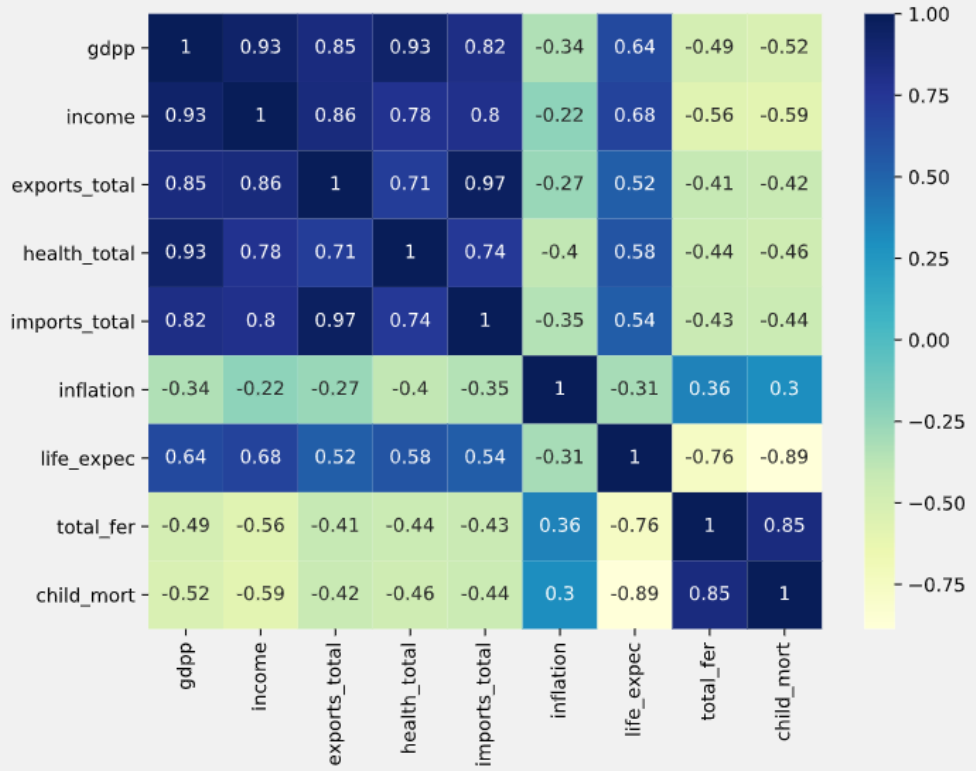
Child Mortality

Child Mortality is correlated with Total Fertility and is negatively correlated with life expectancy.

Inflation

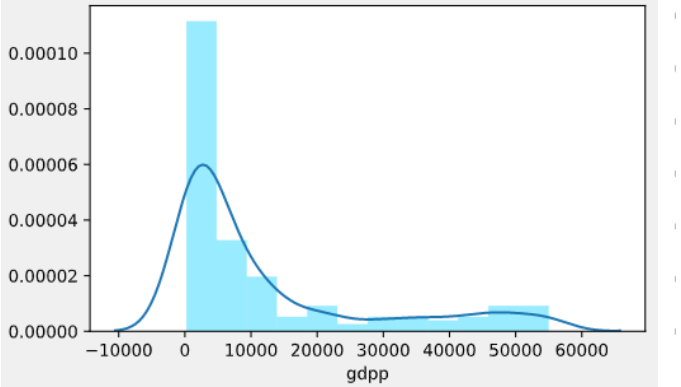
Inflation is another feature that is somewhat correlated or inversely correlated with all other features

Correlation Matrix

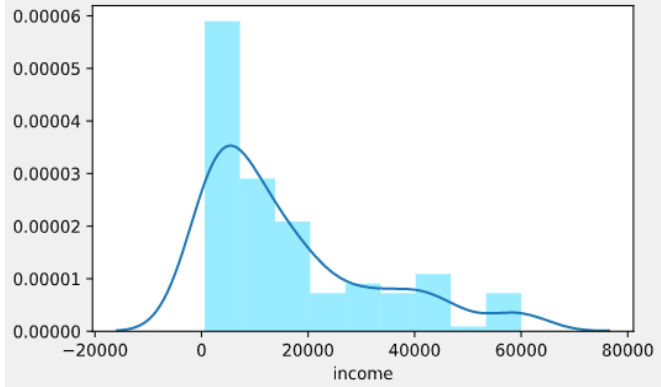


Data Distribution

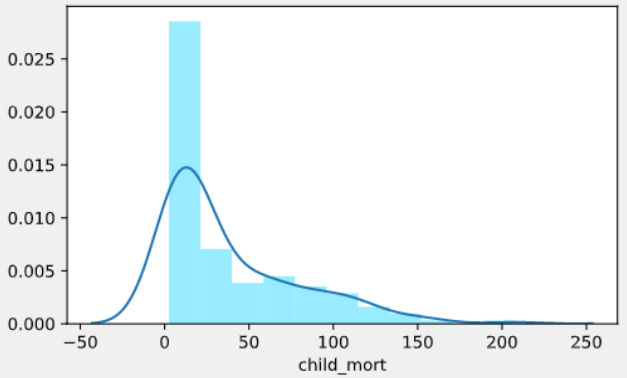
- ✓ The feature distributions are highly skewed to the left, except life expectancy which is right skewed.
- ✓ Since most of the features related to GDPP are skewed towards zero, which may indicate that the dataset had more countries from the underdeveloped nations than developing or developed countries. [*We can achieve a normal distribution for the feature if we obtain the log values instead. But that may make the algorithms unstable and give a skewed outcome. So we will not do that here.*]
- ✓ Interestingly the life expectancy is skewed to the right, which means that most of the countries have a higher life expectancy regardless of the development status.



GDPP is skewed to the left, this may imply that number of under-developed countries is more than developed/developing countries



Income is highly correlated with GDPP and follow the same distribution. However there are a few outliers at the high end of income.



Child mortality rate is also skewed to the left. However this means that that most countries have a lower child mortality rate.

Outliers Treatment

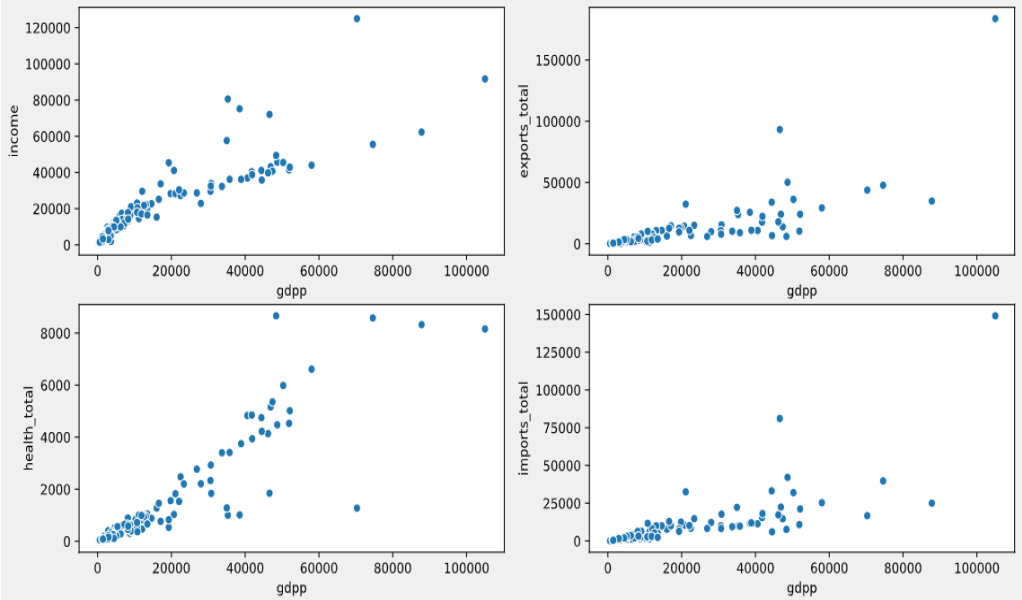
Capping

- ✓ The graph on left shows that these values are highly correlated, and hence all these values should be equivalently spread out towards the end of spectrum.
- ✓ We took a decision on capping these values with values inferred from graph to pack the clusters a bit closer. Let's do this only for the GDPP related features and for higher values only. The reason behind this is, we know that we need to determine countries requiring aid, a general notion would be that countries with higher GDPP parameters may not be figuring in the list anyway.
- ✓ We also cap the 'Inflation' value similarly, and assume anything beyond 20% as 20, since its already a high number.

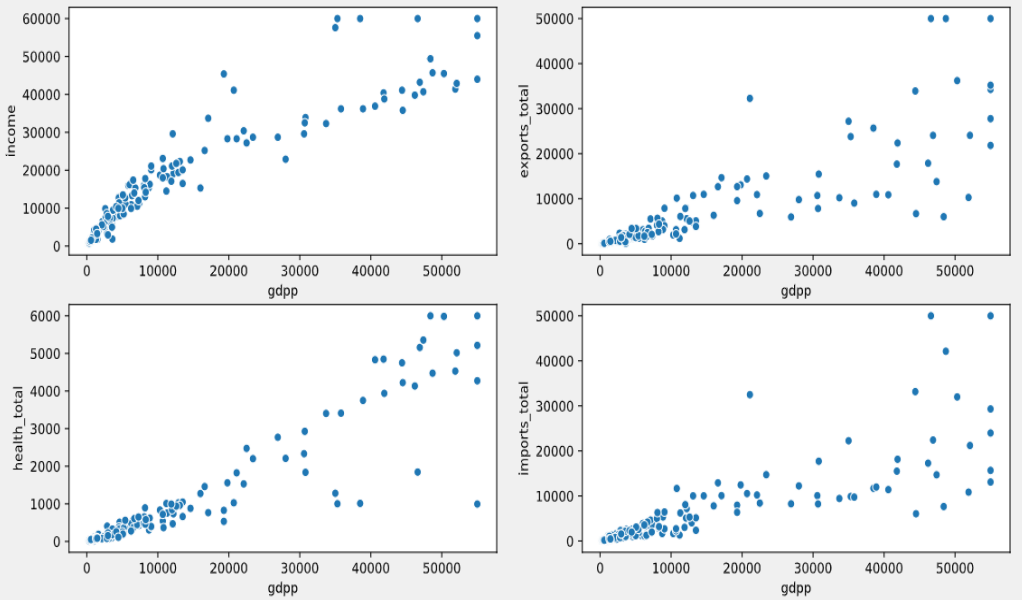
Hopkins Measure : 0.9174157159756519

The Hopkins measure obtained after the outlier's treatment indicates that the dataset is prime for clustering.

Outliers

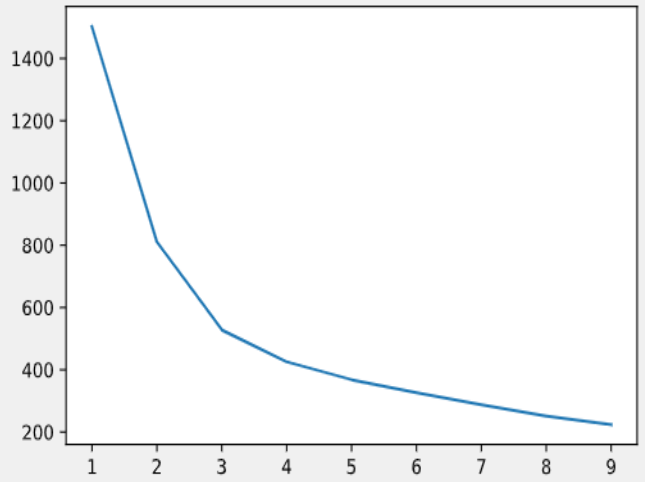


Outliers Treatment



K-Means Clustering

Elbow Method K=3

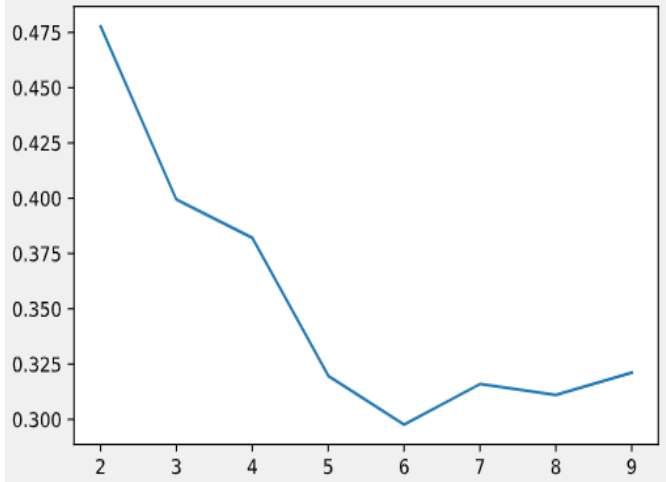


[K = 3]

Based on both the elbow curve and silhouette score analysis, k = 3, seems to be the optimum choice of clusters for k-means.

It should also be noted that, using Silhouette score, even k = 2, option looks viable too. But dividing into 2 clusters does not provide much insight.

Silhouette Score K=2 or 3

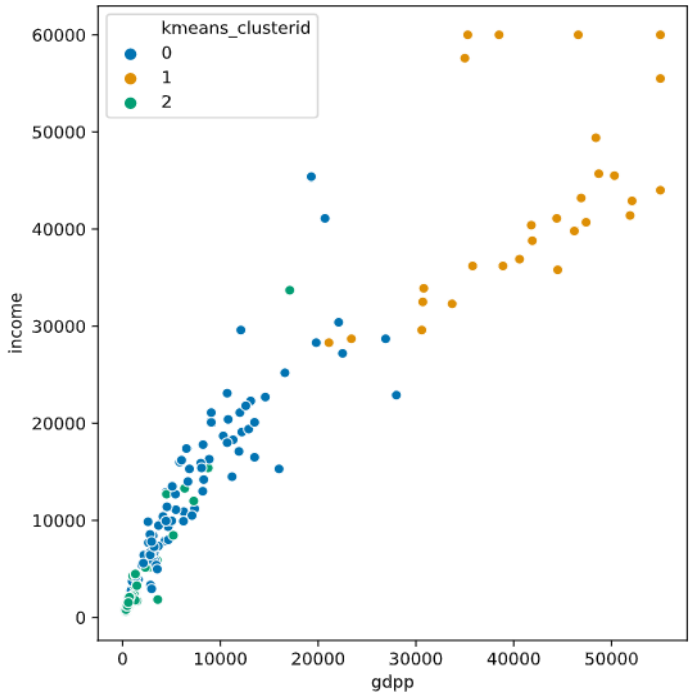


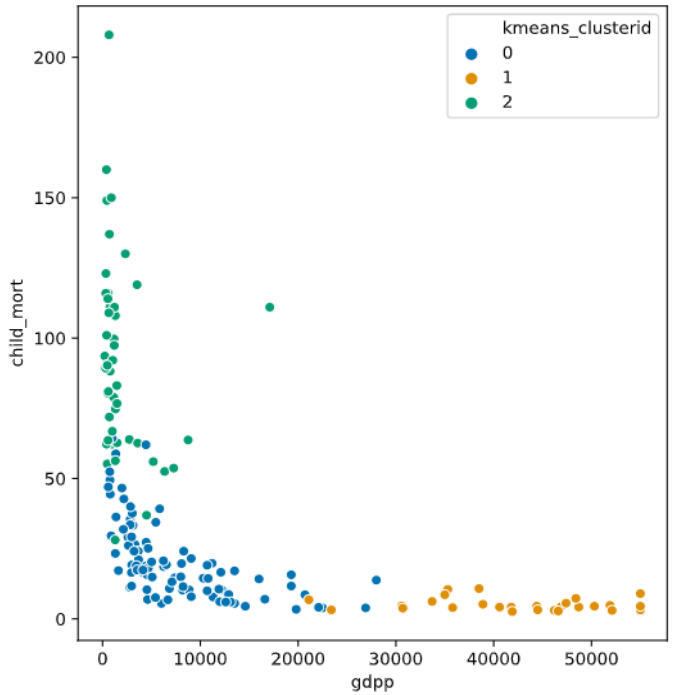
K-Means Clustering

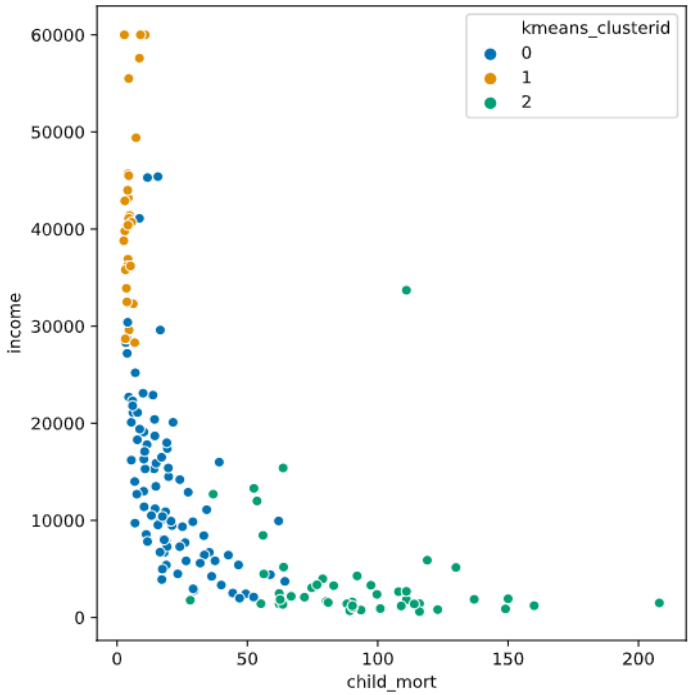
Clustering Output:

The below plots clearly define the 3 categories obtained through k-means clustering.

- ✓ Cluster 0: Has average GDPP/income with mid level child mortality rate.
- ✓ Cluster 1: Group of developed nations with very high GDPP /income with very low child mortality.
- ✓ Cluster 2: This cluster has high child mortality rate with low GDPP/income. There are quite a few outliers here though regarding GDPP, which tells us that the child mortality rate is a significant factor for this cluster.



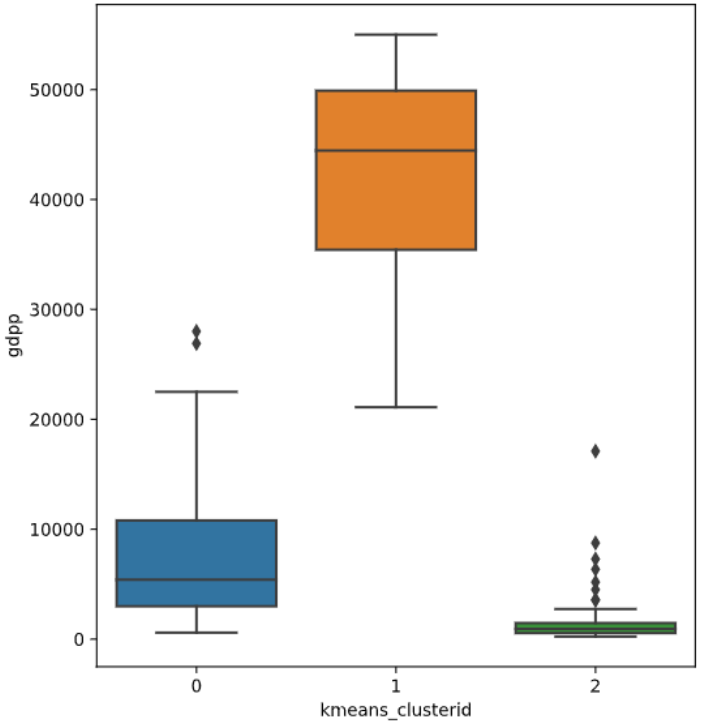


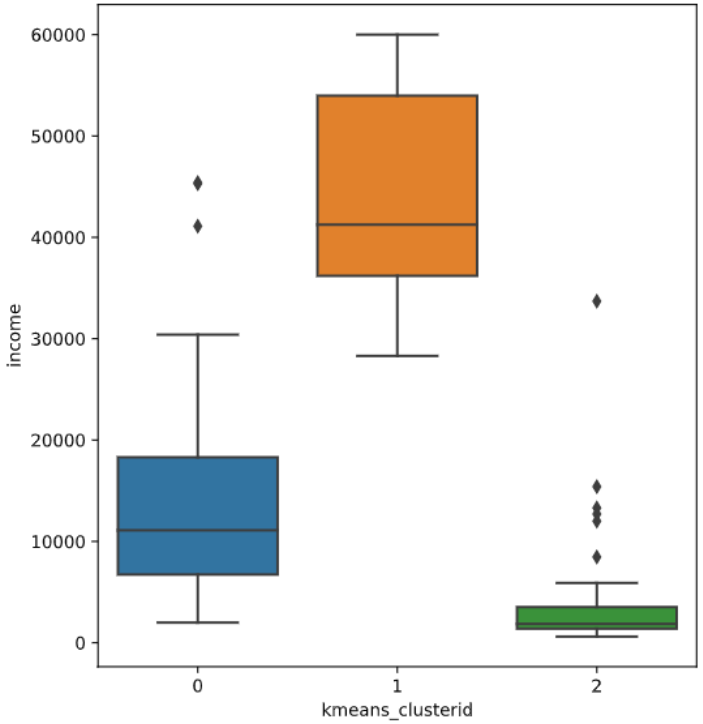


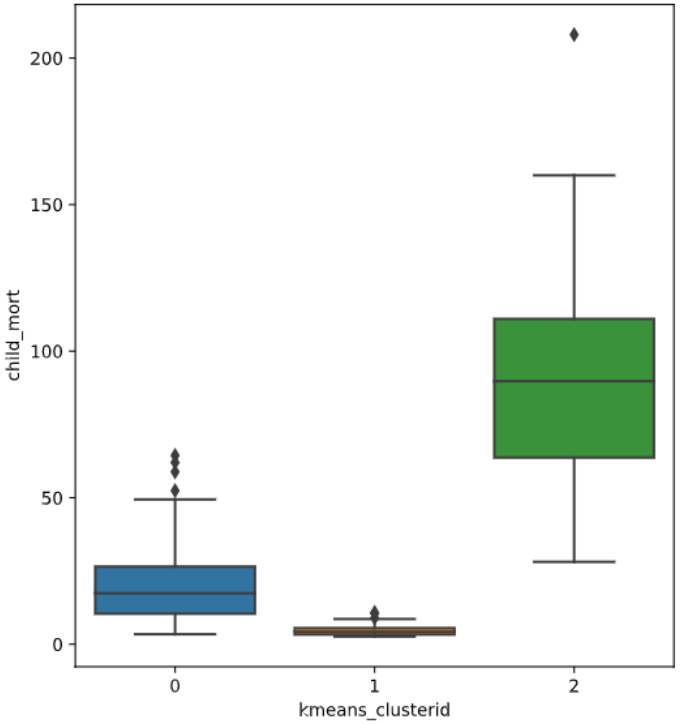
K-Means Clustering

Cluster Profiling:

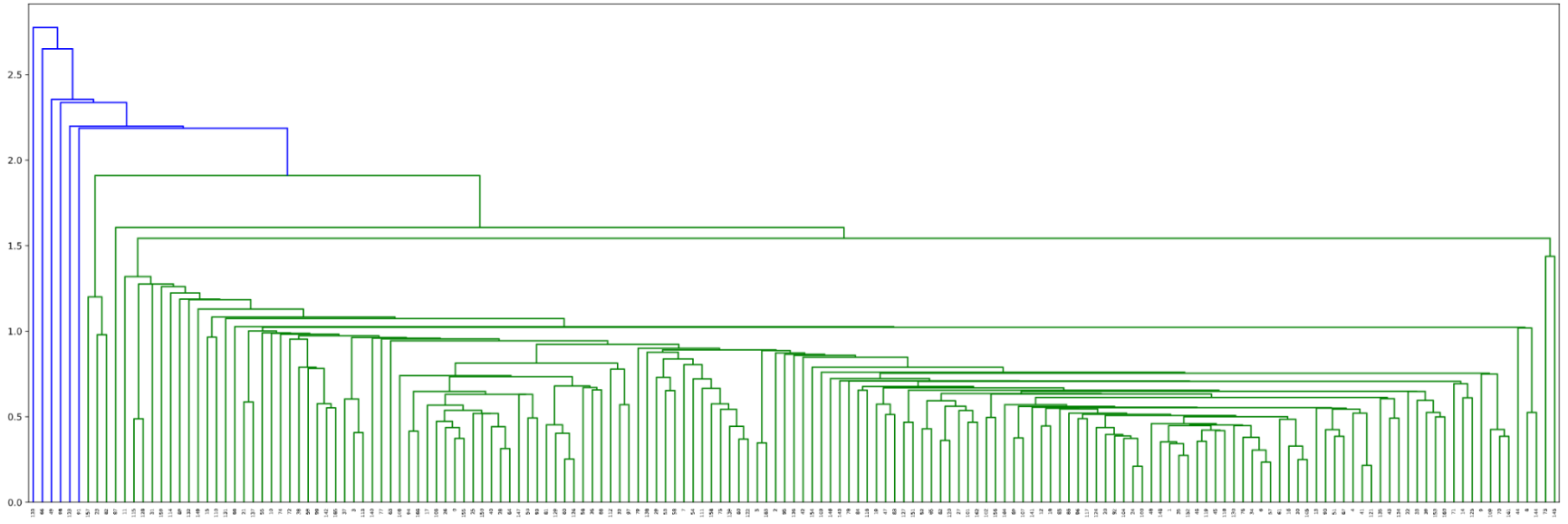
The below box plots throw some good light upon the trends for the 3 clusters identified.
Cluster 0, seems to be moderate in all aspects, i.e. GDPP is average, and Child Mortality rate is mostly below 50.
Cluster 1, which we identified to be developed countries, has significant high GDPP (~45K mean) and Income (~40K mean) and the child mortality rate is close to zero.
Cluster 2, identified as mostly underdeveloped nations, have GDPP and Income at the bottom with mean below 1K, whereas the Child Mortality mean value is closer to 100.







Hierarchical Clustering



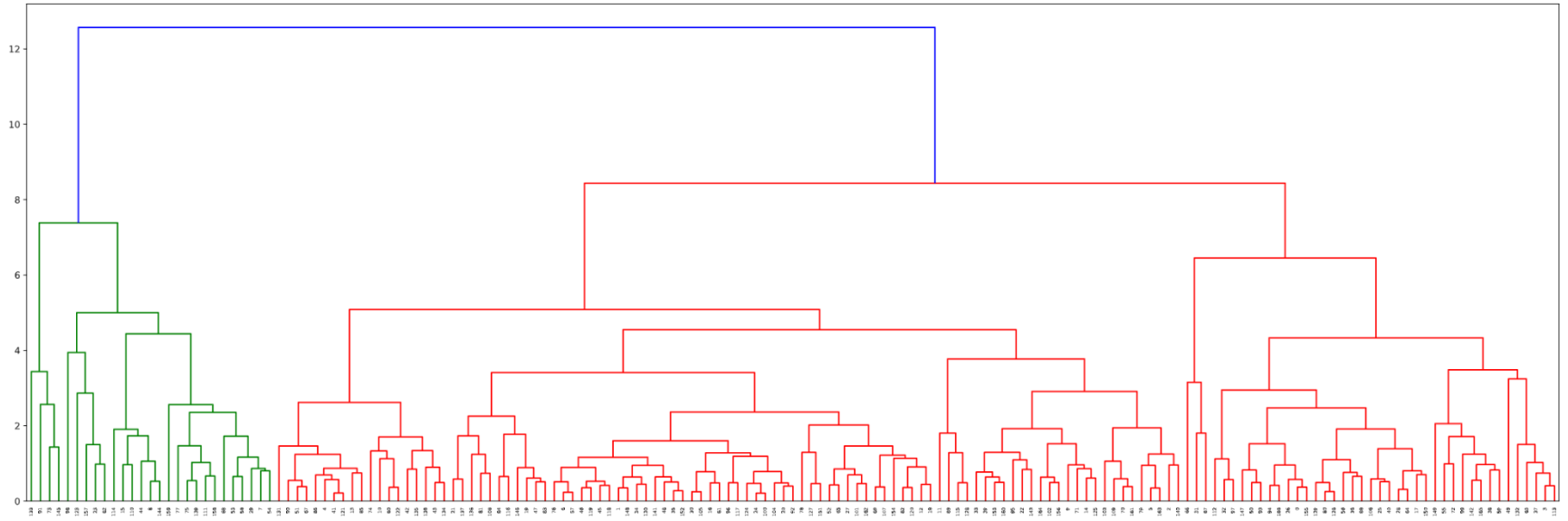
Single Linkage:

The single linkage method does not clearly imply any clustering. Although there are smaller groups at one side of the dendrogram.

Complete Linkage:

The dendrogram with complete linkage looks good for hierarchical clustering, as it has clearly segregated clusters. With 8 features, it gives us 3 groups and with 6 features it gives 5 clusters. This finding is similar as that from k-means preprocessing.

We will use cut-tree to identify our clusters at k = 3

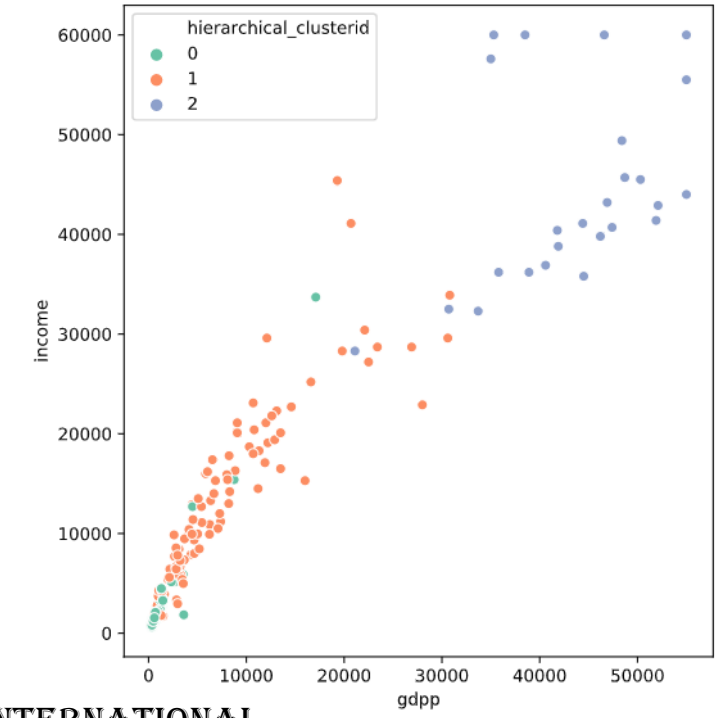


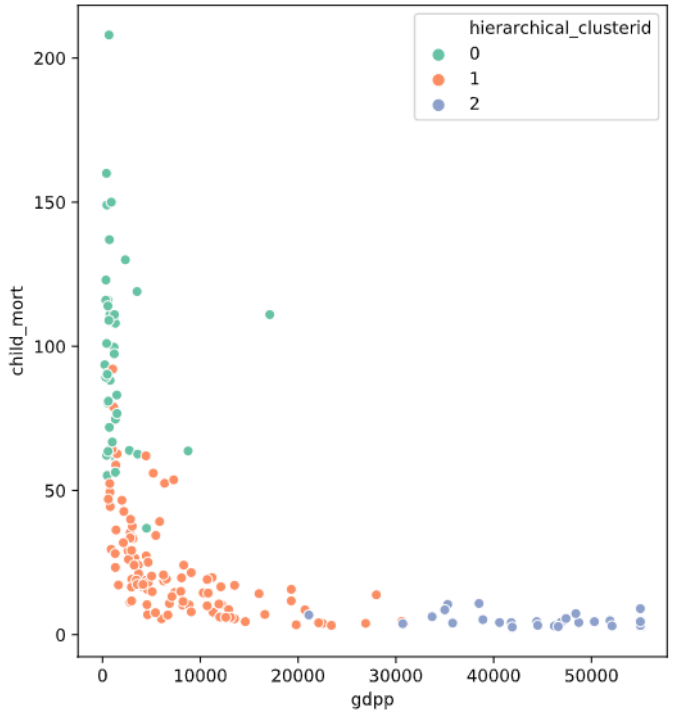
Hierarchical Clustering

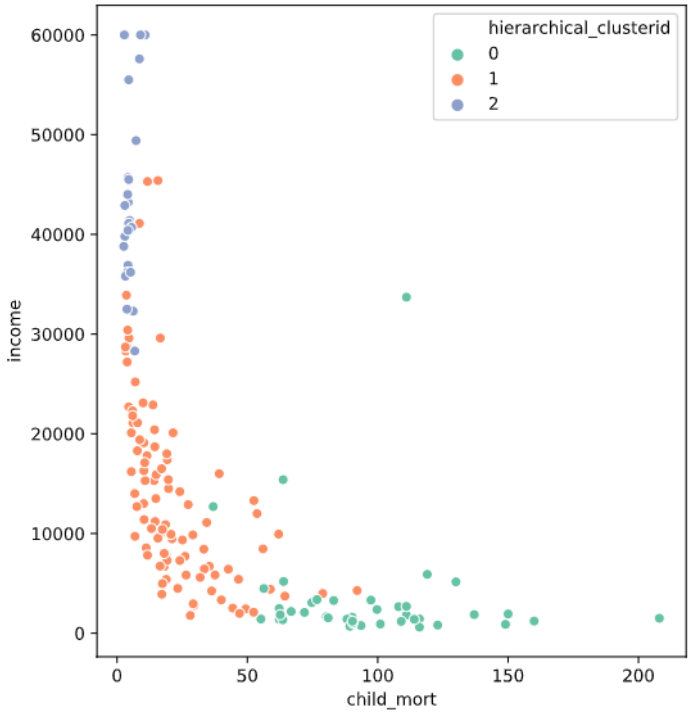
Clustering Output:

The below plots clearly define the 3 categories obtained through hierarchical clustering. This clustering method produced almost similar results as k-means clustering, although there are some shifts between clusters. However the edge cases remain the same which mean the outcome may remain similar.

- ✓ Cluster 0: This cluster has high child mortality rate with low GDPP/Income. There are quite a few outliers here though regarding GDPP, which tells us that the child mortality rate is a significant factor for this cluster.
- ✓ Cluster 1: Has average GDPP /income with mid level child mortality rate.
- ✓ Cluster 2: Group of developed nations with very high GDPP /income with very low child mortality.





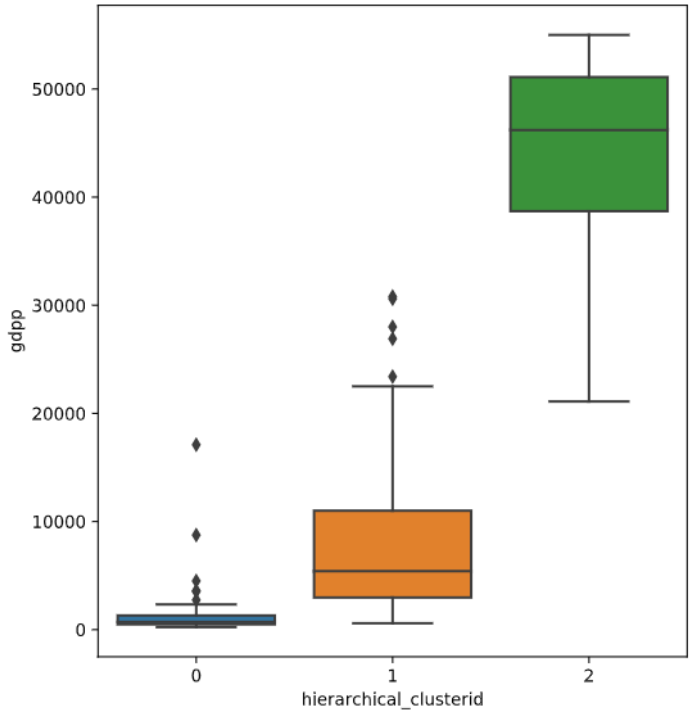


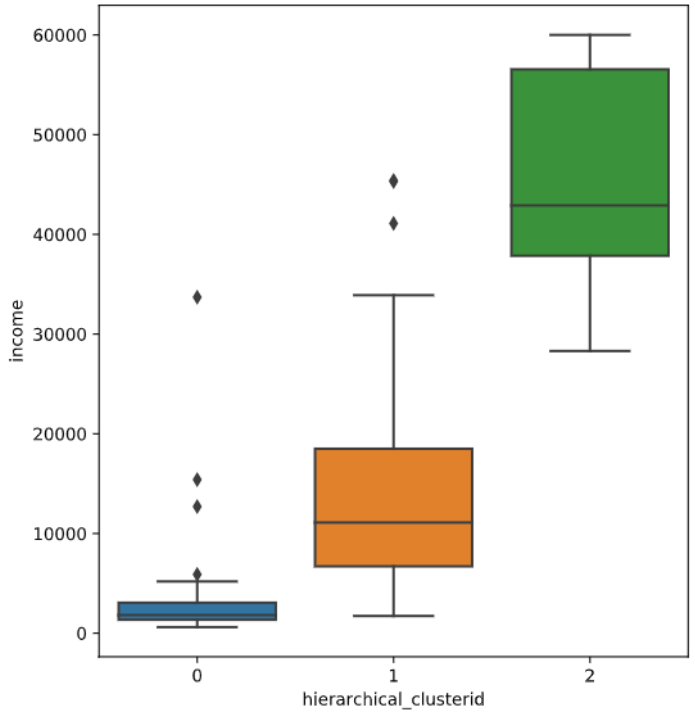
Hierarchical Clustering

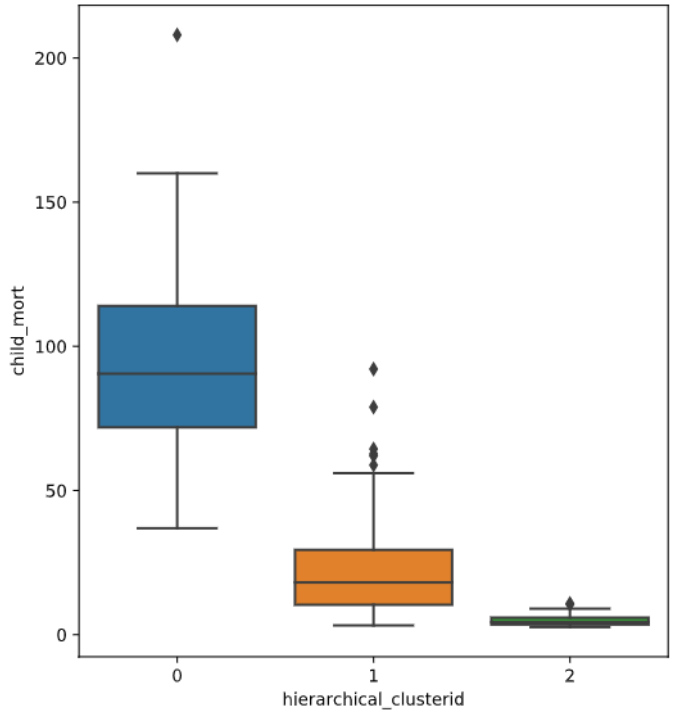
Cluster Profiling:

The below box plots throw some light upon the trends for the 3 clusters identified.

Cluster 0, identified as mostly underdeveloped nations, have GDPP and Income at the bottom with mean below 1K, whereas the Child Mortality mean value is closer to 100.
Cluster 1, seems to be moderate in all aspects, i.e. Income is moderate (~10K Mean), and Child Mortality rate is mostly around 20.
Cluster 2, which we identified to be developed countries, has significant high GDPP (~45K mean) and Income (~40K mean) and the child mortality rate is close to zero.







	Congo, Dem. Rep.	Niger	Central African Republic	Sierra Leone	Haiti
Child Mortality	116	123	149	160	208
Income	669	814	888	1220	1500
GDPP	334	348	446	399	662

Using both K-Means and Hierarchical clustering we arrived at 3 major clusters.

- Low GDPP/High Child Mortality
- Moderate GDPP/Low Child Mortality
- High GDPP/Insignificant Child Mortality

For distribution of aid, hence we can target the cluster with Low GDPP/High Child Mortality. Within this group, to tend to find the country with direst need, we sort them by descending child mortality and ascending GDPP.

The top 5 countries are obtained from the *Hierarchical Clustering* output.


THANK YOU
Soumya Prakash Parida