Clustering Assignment

By

Vigneshwari Chandramohan

Business Requirement

- HELP International – an international humanitarian NGO helps the people of backward countries at the time of disaster and natural calamities.
- They have collected around $10 million from their recent funding programme. They want to use this money effectively.
- The CEO of the company has to be provided with the **top 5 countries** that are in dire need of help.

Problem Statement

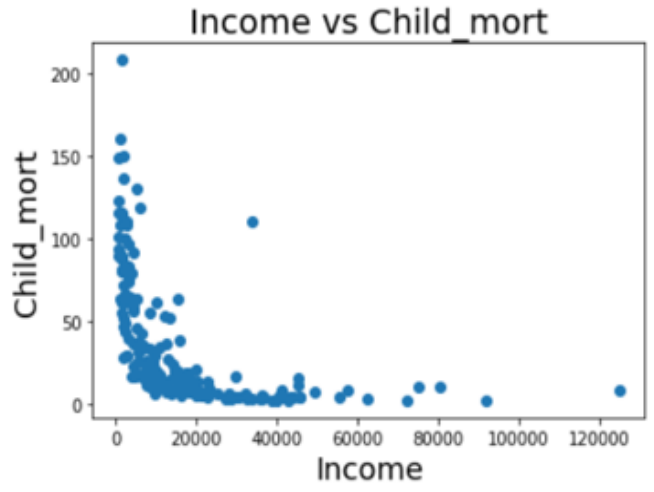
- As a Data Analyst, we have to come up with the names of the 5 countries that are in need of help from the NGO company.
- Based on the criteria - countries that have **low GDPP, low income and high child mortality rate**, the fund has to be distributed to fulfill their requirements.
- Using K-Means and Hierarchical clustering (single and complete linkage), we can cluster the countries based on the selected criteria and accomplish this task in an easier way.

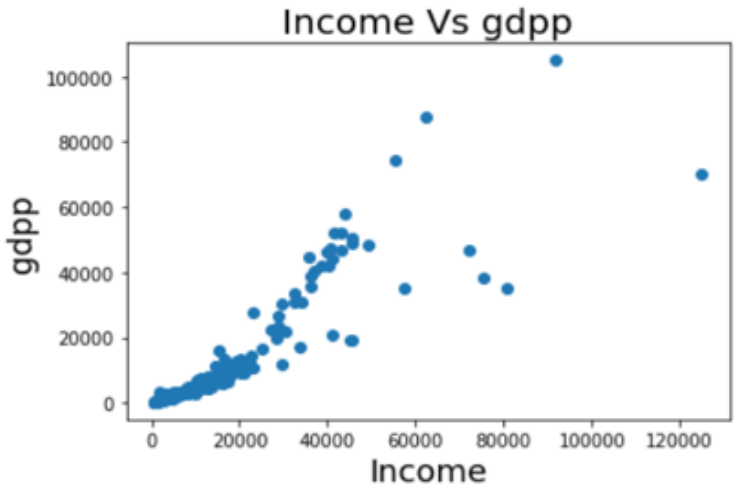
Analysis and Approach:

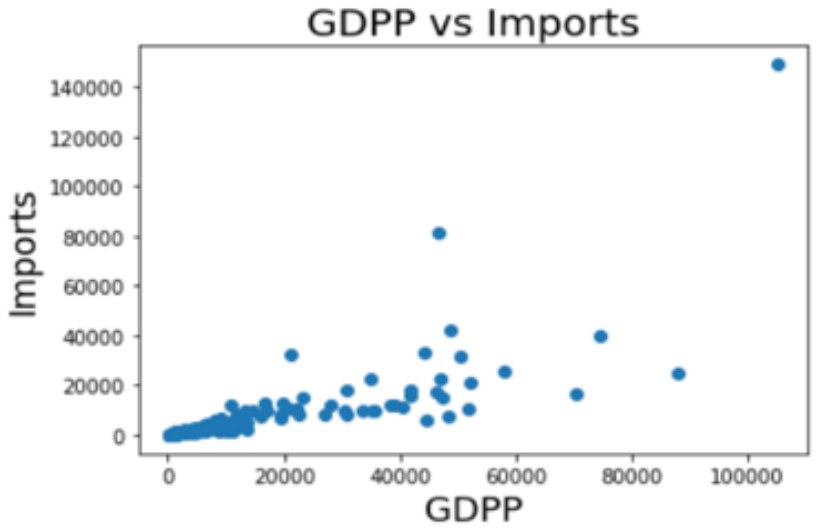
- Using bivariate analysis – we can find the relationship between **gddp, income and child mortality**. On plotting these in a scatter plot, the following things have been observed.
    1. Child_mort increases when the income decreases
    2. Gdpp is low when the income is low and vice versa
    3. Child mortality is high when the Gdpp is low
    4. Exports and Imports are high for the countries whose GDDP is high

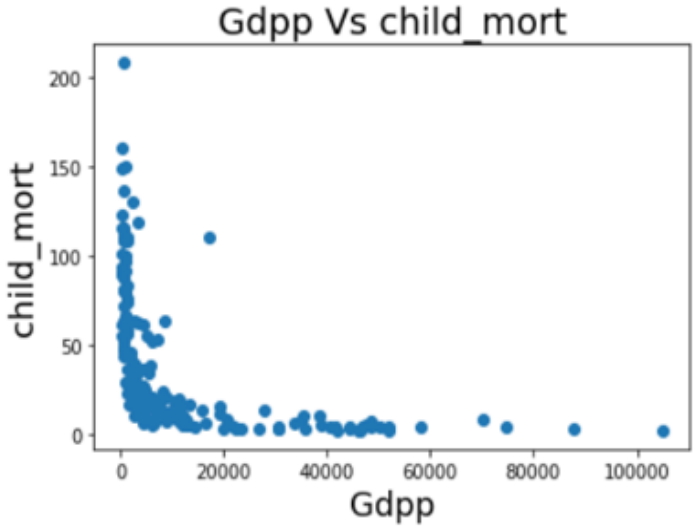
- Using boxplots, we could identify the outliers and remove them if they are not required.
    1. We need to retain the **higher end values** for child_mort and inflation.
    2. For other features - GDPP, exports, imports, income and health, we need to retain the **lower end values**.

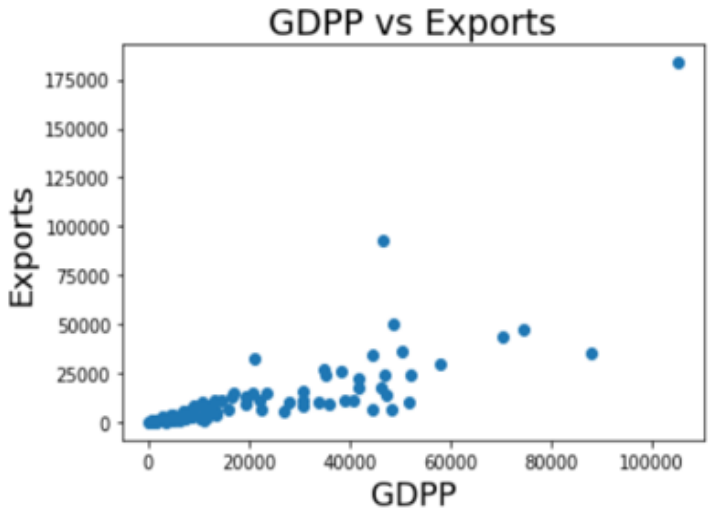
Visualisation from Bivariate Analysis











Approach

- To identify the countries, we can perform **K-Means clustering** and **Hierarchical clustering** (single and complete linkage).
- Before clustering, we need to do scaling and check the **Hopkins statistics**. Hopkins statistics is used to measure the cluster tendency.
- If the Hopkins statistics is greater than 0.8, then we could assume that the given dataset has good cluster tendency.
- After this, we can perform K-Means and Hierarchical clustering to identify the top 5 countries that are in need of help.

K-Means Clustering

1. Based on the Elbow-curve/SSD and Silhouette score, choose the optimal value for K (number of clusters). After running the K-Means clustering algorithm, each datapoint will be assigned to a cluster.
2. Now, each country in the dataset has been assigned to any one cluster.
3. Based on the cluster properties, we can identify the countries looking for need help.
4. To identify the cluster that has low gdpp, low income and high child mortality we can do profiling.
5. Countries under **cluster 0** are in need of immediate help.

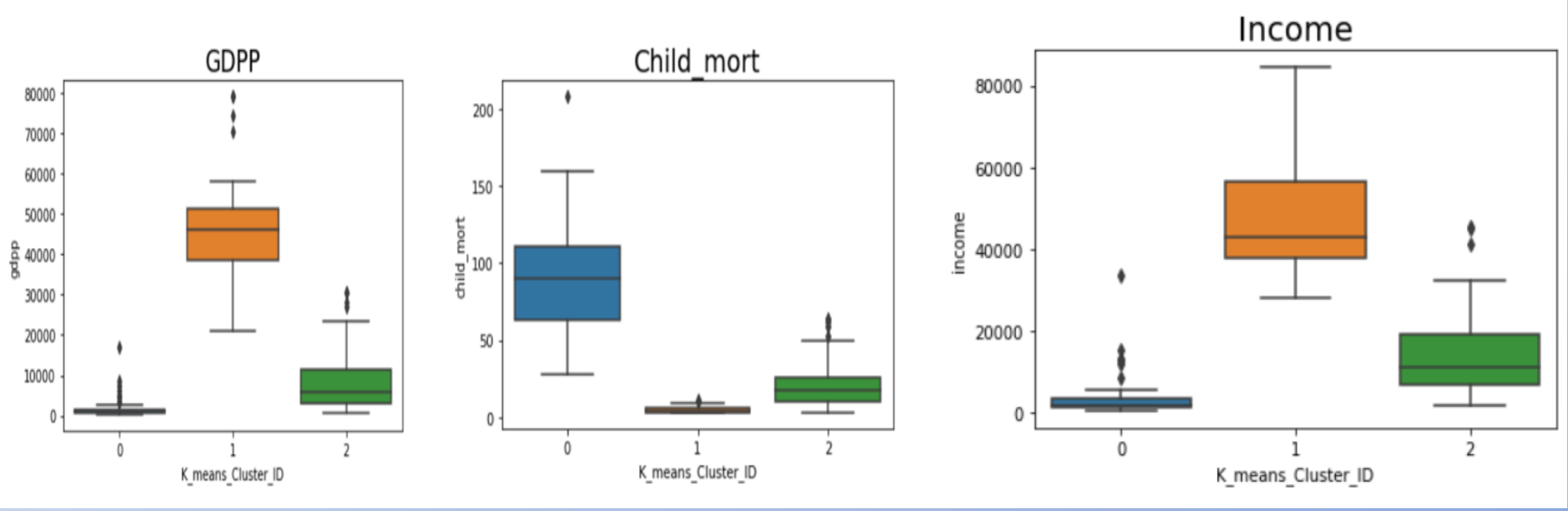
```
country_df[['gdpp', 'income', 'child_mort']].groupby(country_df.K_means_Cluster_ID).mean()
```

K_means_Cluster_ID	gdpp	income	child_mort
0	1909.208333	3897.354167	91.610417
1	47476.888889	49057.333333	5.122222
2	8226.869565	14169.456522	20.177174

On doing the profiling, we came to know that the countries in the cluster 0 has **low gdpp, low income and high child_mort**.

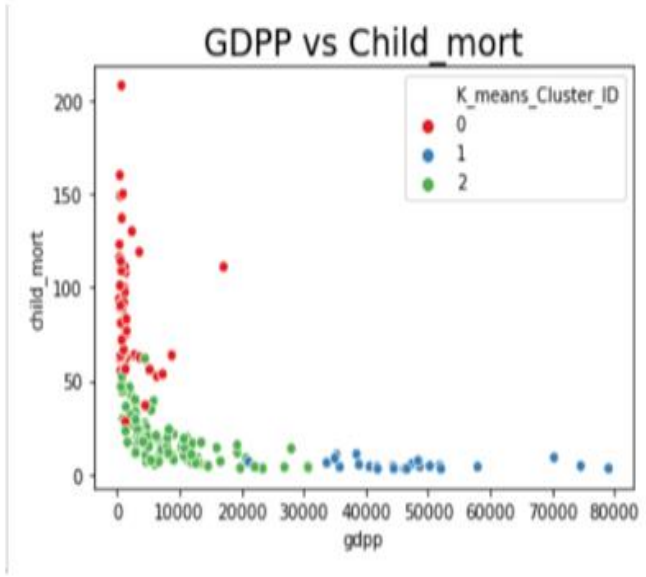
Boxplot –Visualisation (cluster identification)

From the below boxplot visualisation, we can clearly say that, **Cluster 0 has low GDP, high child mortality and low income**.

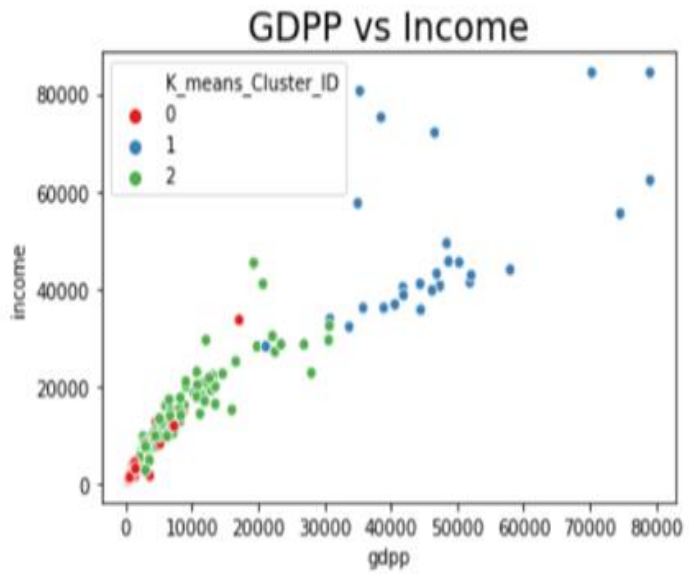


Scatterplot – Visualisation (cluster identification)

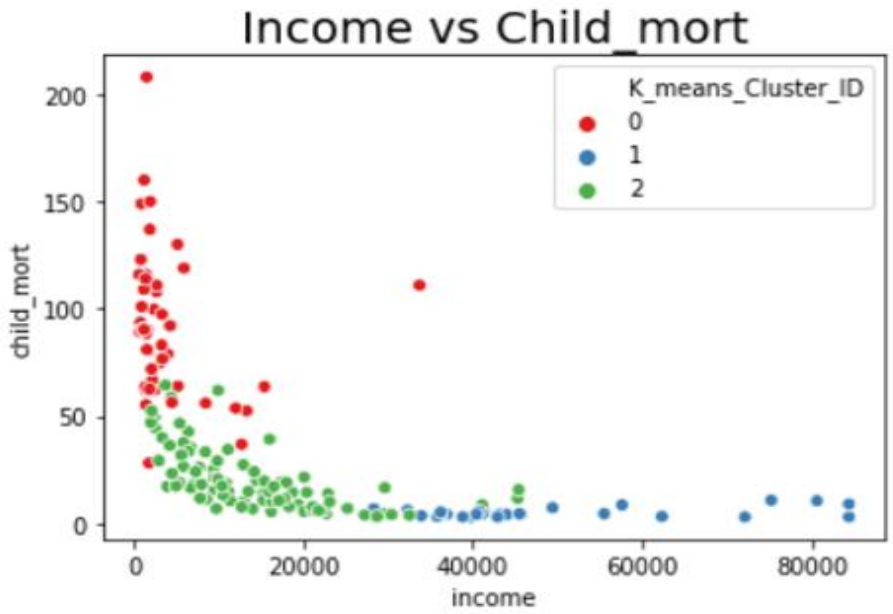
From the below scatterplot visualisation, we can confidently say that, Cluster 0 (**red dots**) has high child mortality with low GDP and low income.



Cluster 0 - has lower GDPP and higher child_mort



Cluster 0 - has lower GDPP and low income



Cluster 0 - has lower income and higher child_mort

Top 5 countries in Cluster 0 by K-Means Clustering:

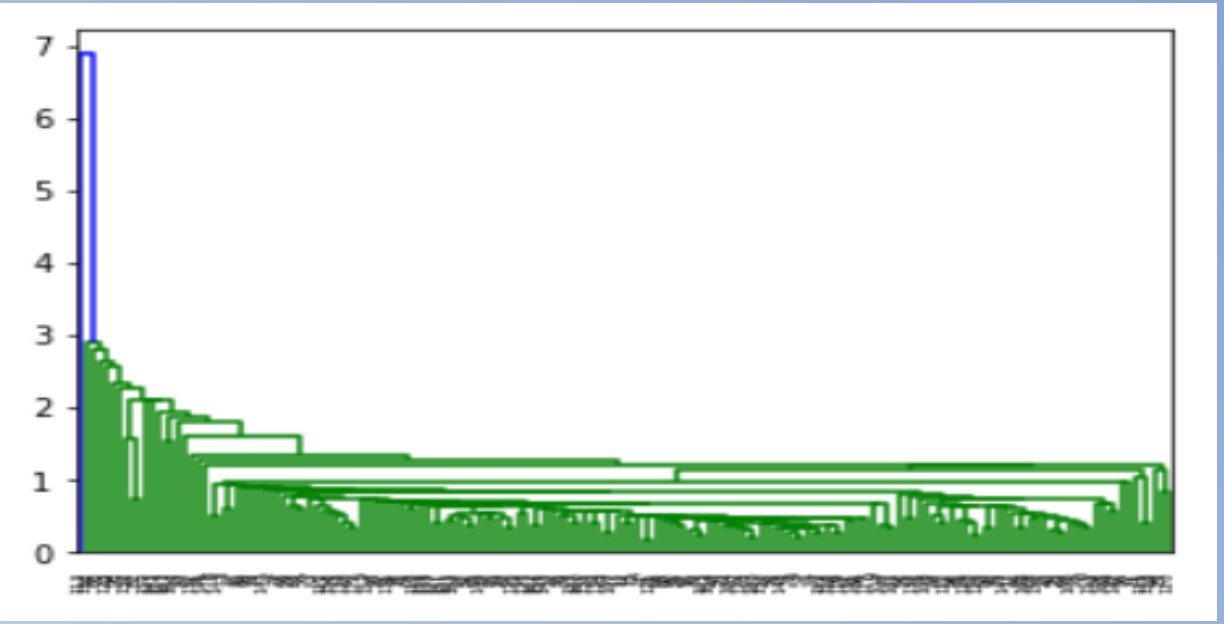
	country	gdpp	income	child_mort	K_means_Cluster_ID
26	Burundi	231.0	764.0	93.6	0
88	Liberia	327.0	700.0	89.3	0
37	Congo, Dem. Rep.	334.0	609.0	116.0	0
112	Niger	348.0	814.0	123.0	0
132	Sierra Leone	399.0	1220.0	160.0	0

Hierarchical Clustering

Using the K-value from Elbow-curve/SSD and Silhouette score, we can perform Hierarchical clustering.

Single Linkage Method:

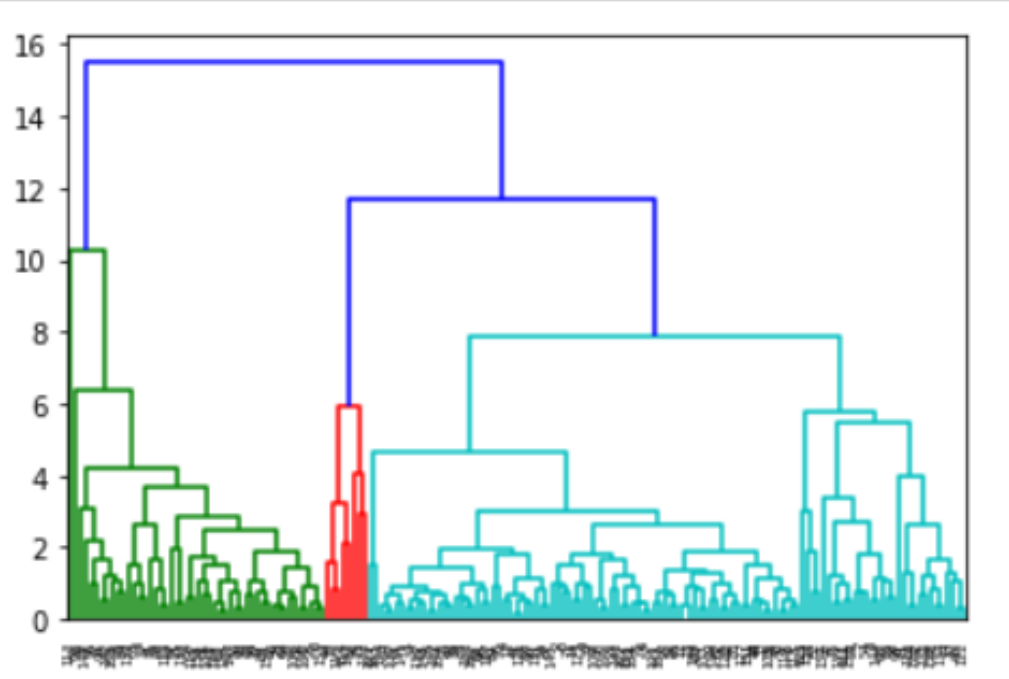
In Single linkage method, distance between 2 clusters is defined as the shortest distance between points in the two clusters. But, we could not clearly visualise the **dendrogram**.



Hierarchical Clustering

Complete linkage method:

In this method, the distance between 2 clusters is defined as the maximum distance between any 2 points in the clusters. So, the **dendrogram** can be visualised clearly.



- Using the K-value from the K-Means algorithm, we can cut the dendrogram (created using complete linkage) and identify the cluster labels. Now, the datapoints has been assigned to clusters.
- Based on the cluster properties, we can identify the countries that need help.
- To identify the cluster that has low gdpp, low income and high child mortality we can do profiling.
- Countries under **cluster 0** are in need of immediate help.

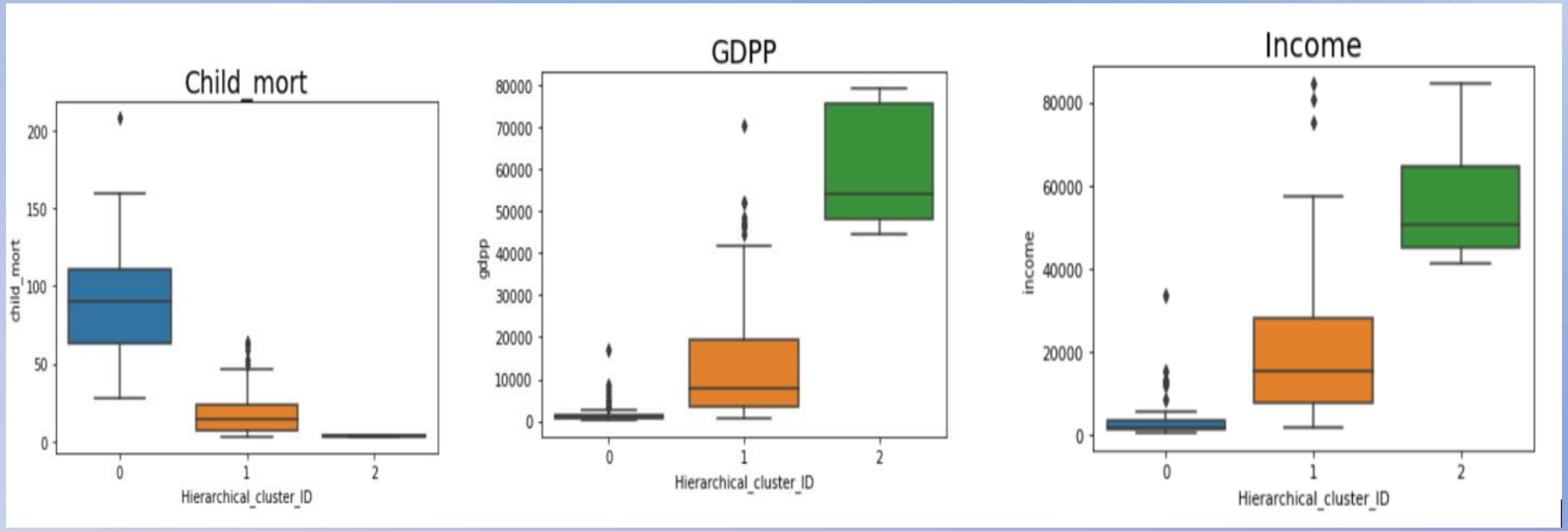
```
country_df[['gdpp', 'income', 'child_mort']].groupby(country_df.Hierarchical_cluster_ID).mean()
```

Hierarchical_cluster_ID	gdpp	income	child_mort
0	1909.208333	3897.354167	91.610417
1	14035.783784	19617.693694	17.690090
2	60097.000000	56321.750000	3.875000

On doing the profiling, we came to know that the countries in the cluster 0 have **low gdpp,low income and high child_mort**.

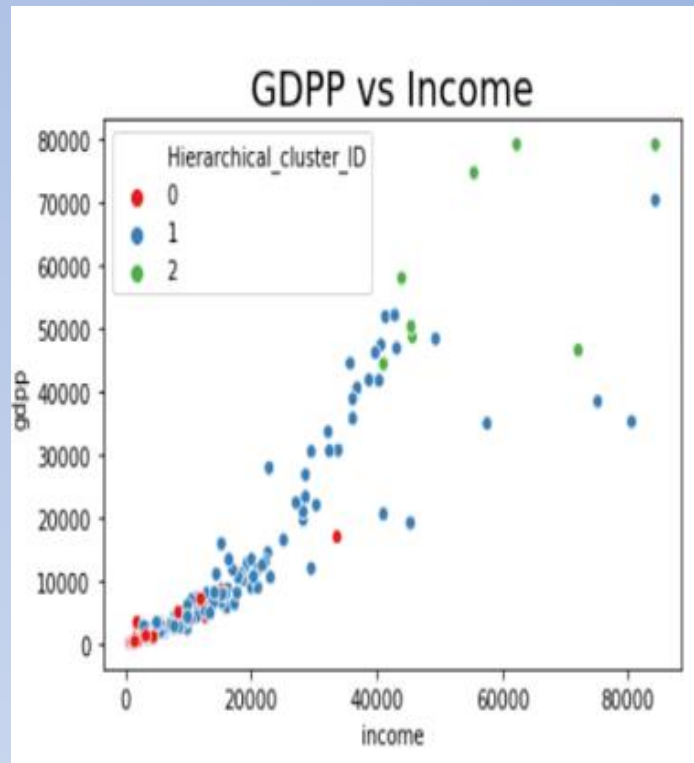
Boxplot –Visualisation (cluster identification)

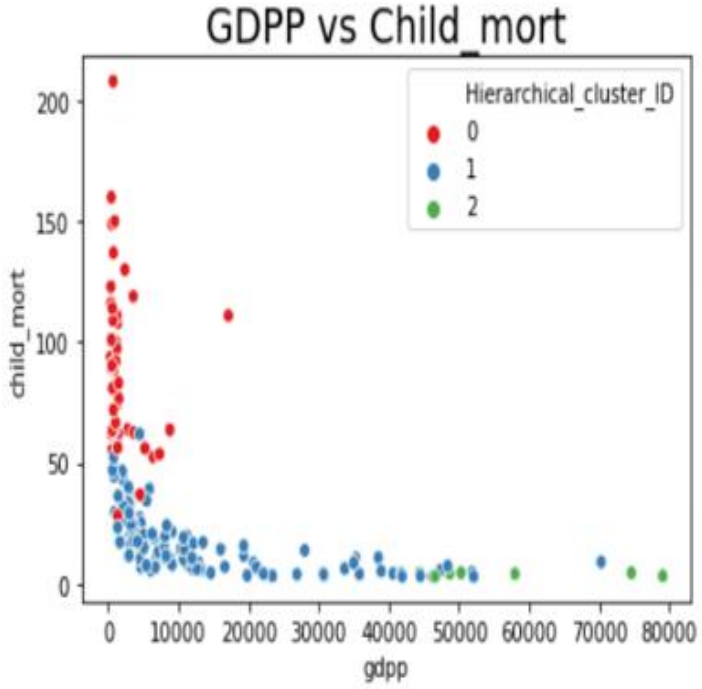
From the below boxplot visualisation, we can clearly say that, **Cluster 0 has low GDP, high child mortality and low income**.

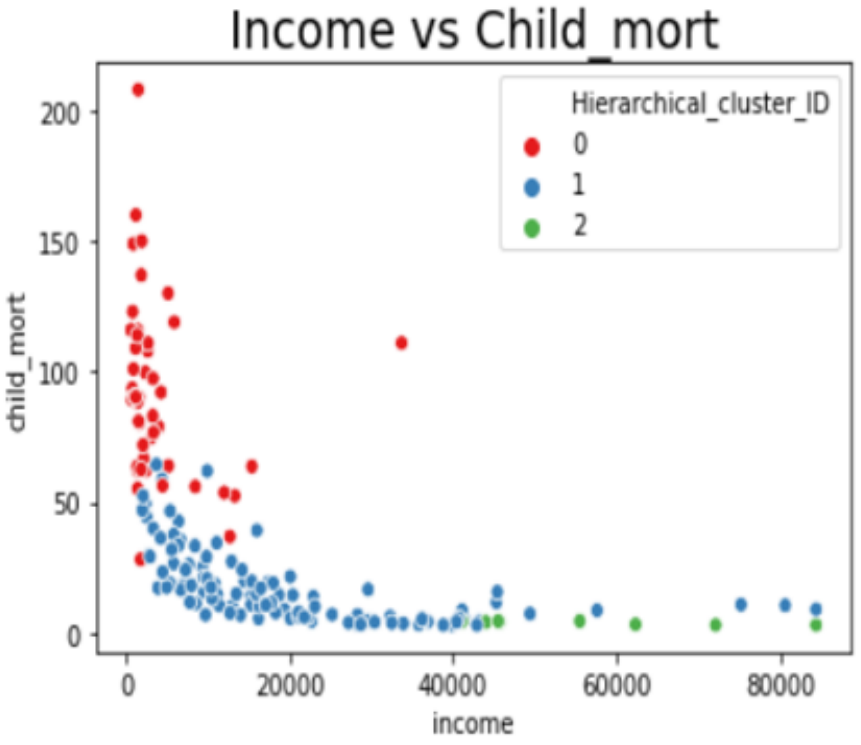


Scatterplot –Visualisation (cluster identification)

From the below scatterplot visualisation, we can confidently say that, Cluster 0 (**red dots**) has high child mortality with low GDP and low income.







Top 5 countries in Cluster 0 by Hierarchical Clustering:

	country	gdpp	income	child_mort	Hierarchical_cluster_ID
26	Burundi	231.0	764.0	93.6	0
88	Liberia	327.0	700.0	89.3	0
37	Congo, Dem. Rep.	334.0	609.0	116.0	0
112	Niger	348.0	814.0	123.0	0
132	Sierra Leone	399.0	1220.0	160.0	0

Recommendations

From both K-Means clustering and Hierarchical clustering, we have obtained the **same top-5 countries**, which have low GDP, high child mortality and low income. HELP International could immediately focus on these 5 countries to aid in their basic amenities.

1. Burundi
2. Liberia
3. Congo, Dem. Rep.
4. Niger
5. Sierra Leone