HELP INTERNATIONAL

Clustering Assignment

SUBMITTED BY:

RATIK KHANNA

The problem statement

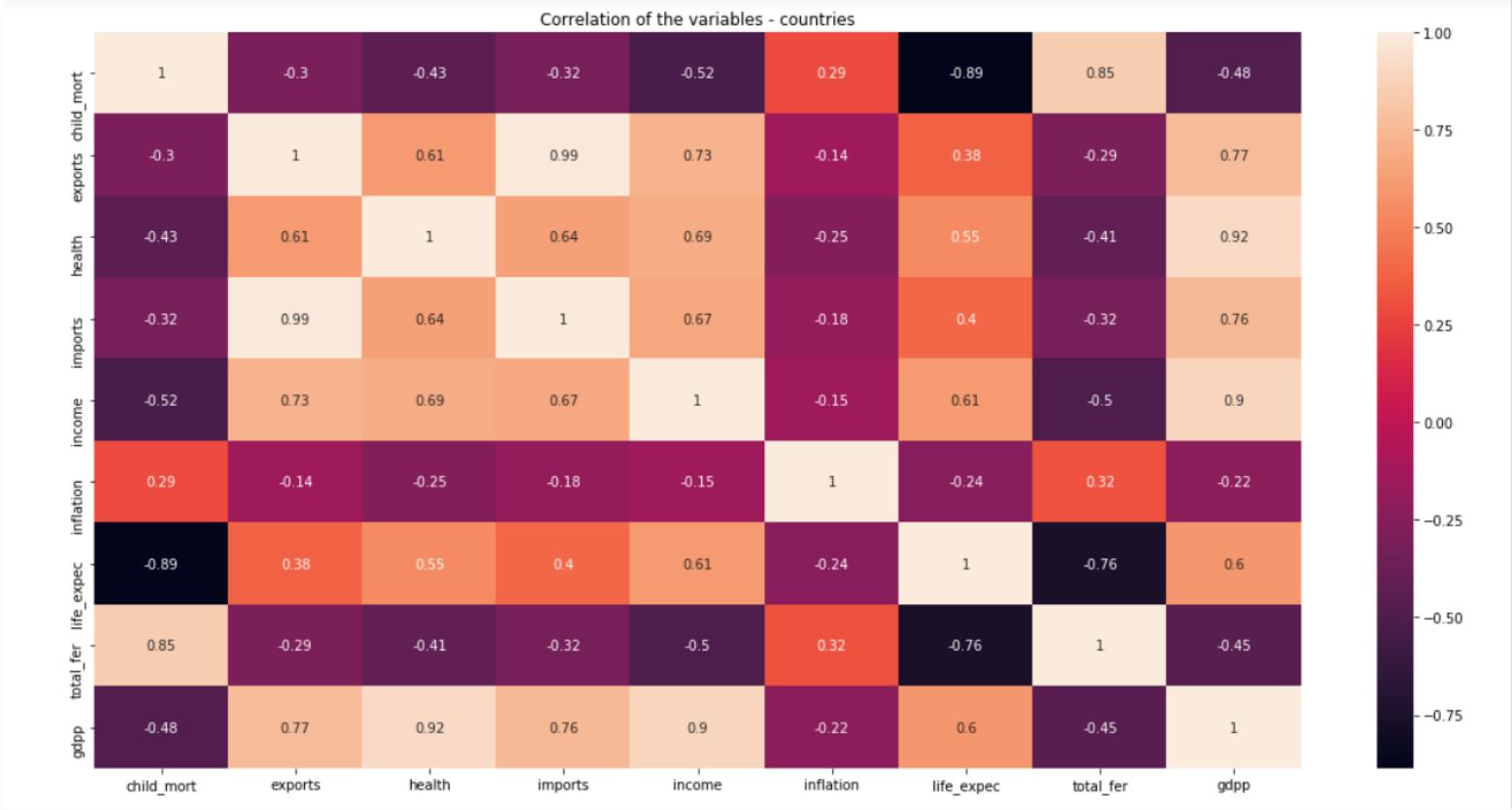
HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries. Now the CEO of the NGO needs to decide how to use $ 10 million strategically and effectively. The significant issues that come while making this decision are mostly related to choosing the countries that are in the direst need of aid. So here I come as a Data Analyst and my job is to categorise the countries using some socioeconomic and health factors that determine the overall development of the country.

The analysis approach

1. Understand the Data – Here I checked the basic information of data like null values, data type, columns, etc.
2. Preparing the Dataset- Here I converted health, import and export columns into a uniform form, so that analysis can be performed easily.
3. Performing EDA- I performed EDA of data and visualized the dataset in order to have better understanding of data set.
4. Treating Outliers- During EDA, found that various columns have outliers, so I treated the outliers.
5. Rescaling Dataset- To standardise the values, rescaling was performed on dataset.
6. Hopkins Check- Hopkins check is performed 10 times on dataset and the value was above 80% in all the cases.
7. K-means Clustering- Elbow curve analysis and Silhouette score was performed to determine the value of ‘k’. The value of k was obtained as 2.
8. 8. Hierarchical Clustering- Hierarchical clustering was performed, both Single Linkage and Complete Linkage was performed, and here also the value of k was obtained as 2.
9. 9. Final Conclusion- Both the clustering suggests same top 5 countries that require aid. Thus, we draw conclusion that Burundi, Liberia, Congo, Niger and Sierra Leone are top 5 countries that require AID from HELP International.

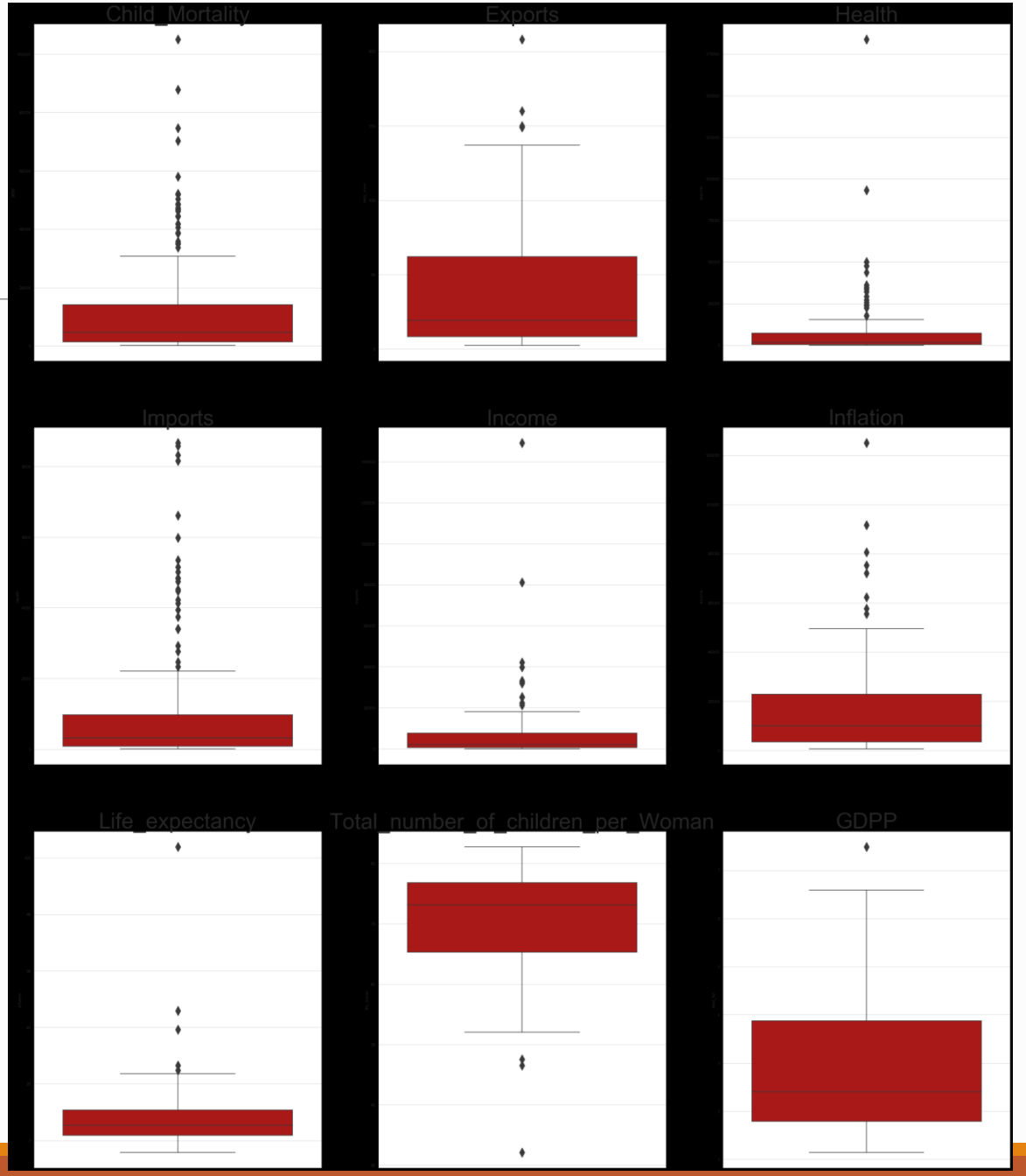
Correlation Heatmap of all Columns

From the above heatmap, we can see that there are some variables having very high correlation with repect to postive and negative.



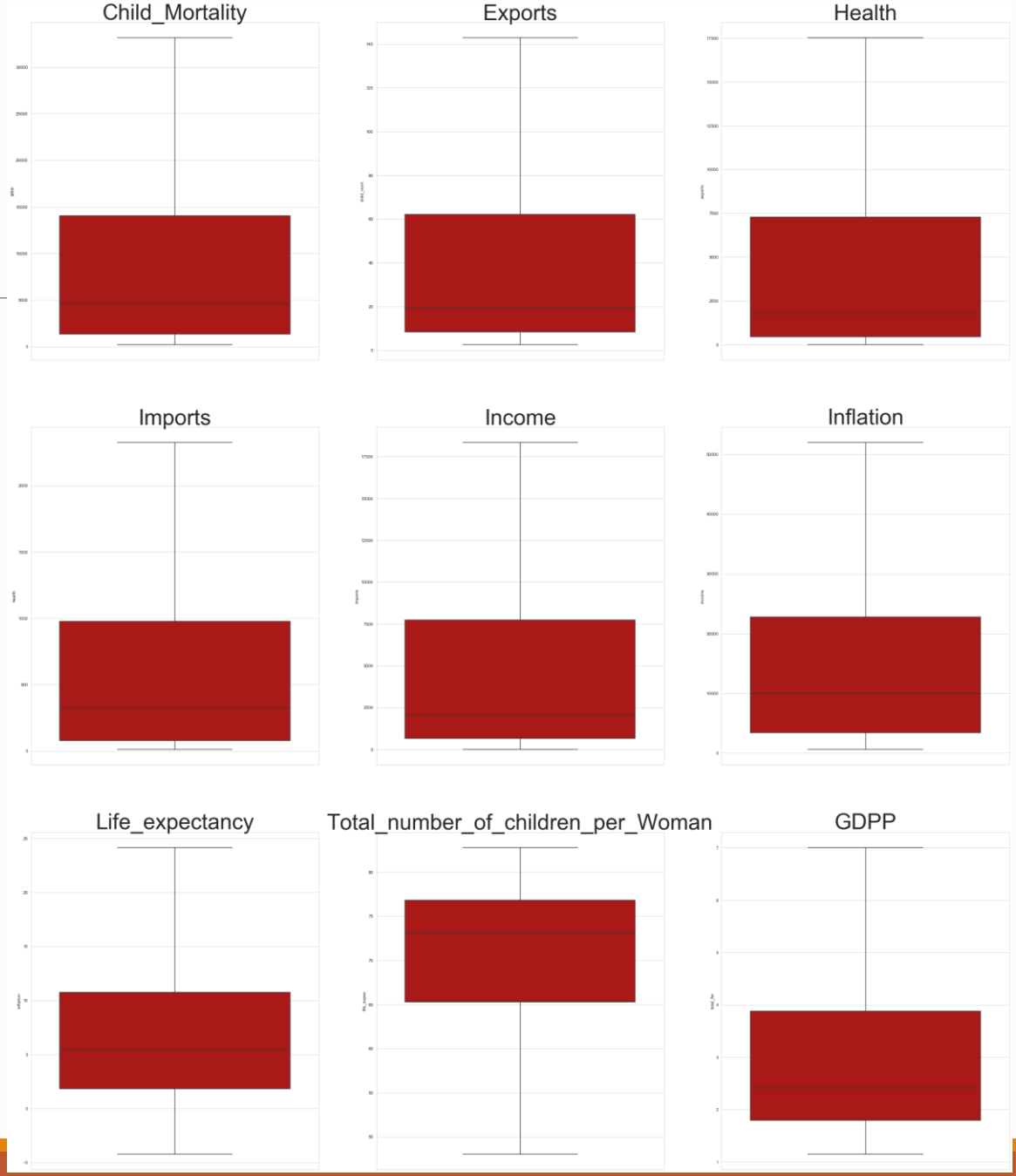
Visualising the outliers with boxplot

From the boxplot, we can conclude that all the variables/components are having outliers. So, will be treating those outliers.



Outlier Treatment

All the outliers were treated, in order to obtain a good value of ‘k’.

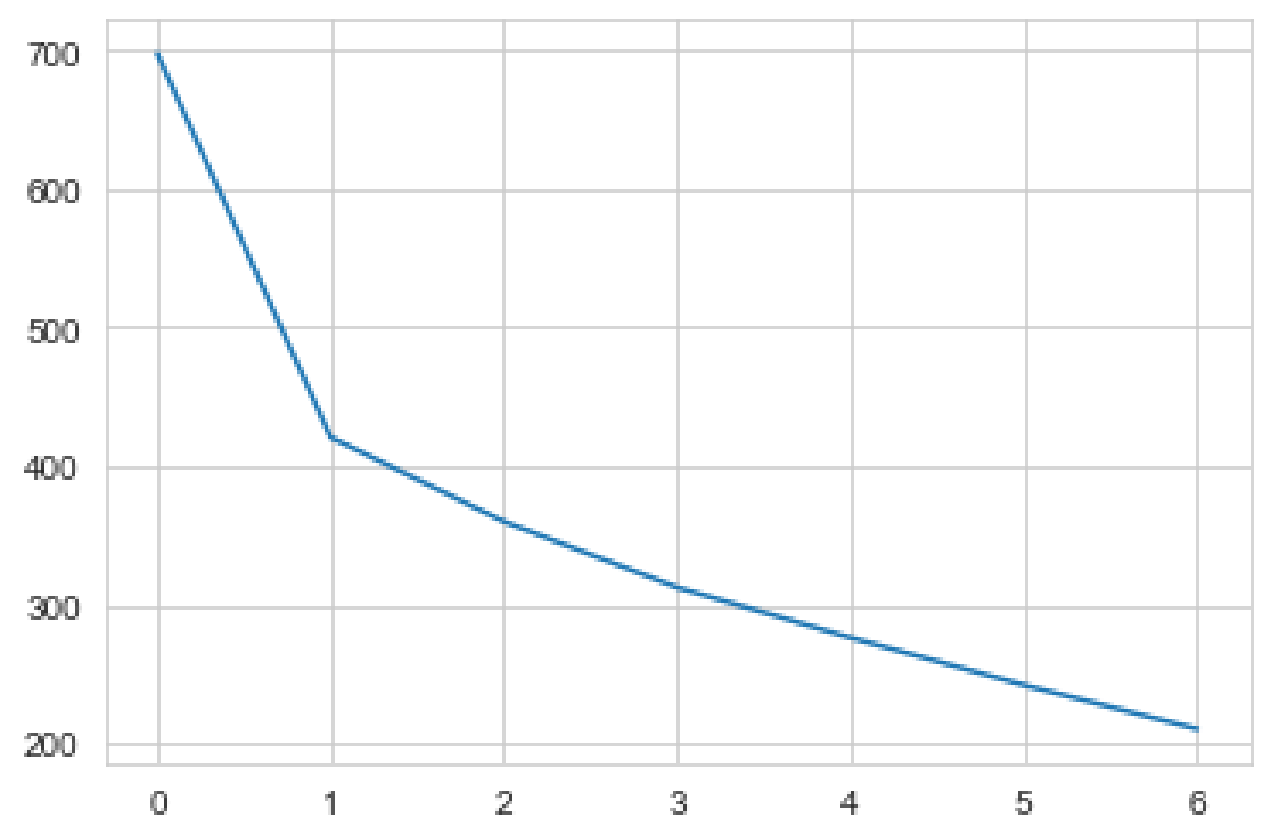


K-means Clustering

K-MEANS IS METHOD OF CLUSTER ANALYSIS USING A PRE-SPECIFIED NO. OF CLUSTERS. IT REQUIRESADVANCE KNOWLEDGE OF ‘K’.

Elbow Curve analysis

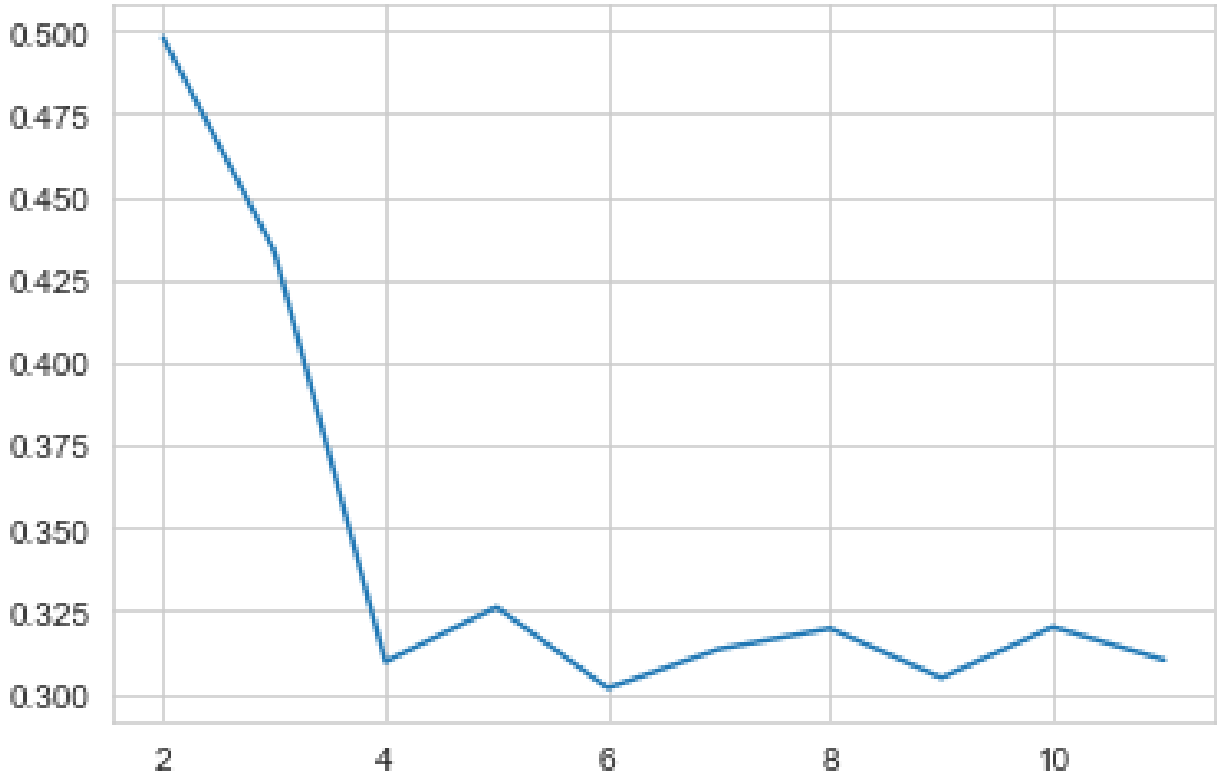
The elbow curve is dropping significantly till 1, i.e. 2 clusters.



Silhouette Score Analysis

```
For n_clusters=2, the silhouette score is 0.4980131625805177
For n_clusters=3, the silhouette score is 0.43360177139259876
For n_clusters=4, the silhouette score is 0.3095930455357734
For n_clusters=5, the silhouette score is 0.32611022128924727
For n_clusters=6, the silhouette score is 0.3016613783800461
For n_clusters=7, the silhouette score is 0.3134484542778152
For n_clusters=8, the silhouette score is 0.3196820701950583
For n_clusters=9, the silhouette score is 0.304763208260997
For n_clusters=10, the silhouette score is 0.3201376728297778
For n_clusters=11, the silhouette score is 0.31018208489087756
```

The silhoutte score is high for 2 clusters.



Cluster obtained from K-means clustering.

```
0    123
1     44
Name: cluster_id, dtype: int64
```

The elbow curve suggests 2 clusters will be optimum and silhouette score is high for 2 clusters. Thus, we will be taking the value of k = 2.

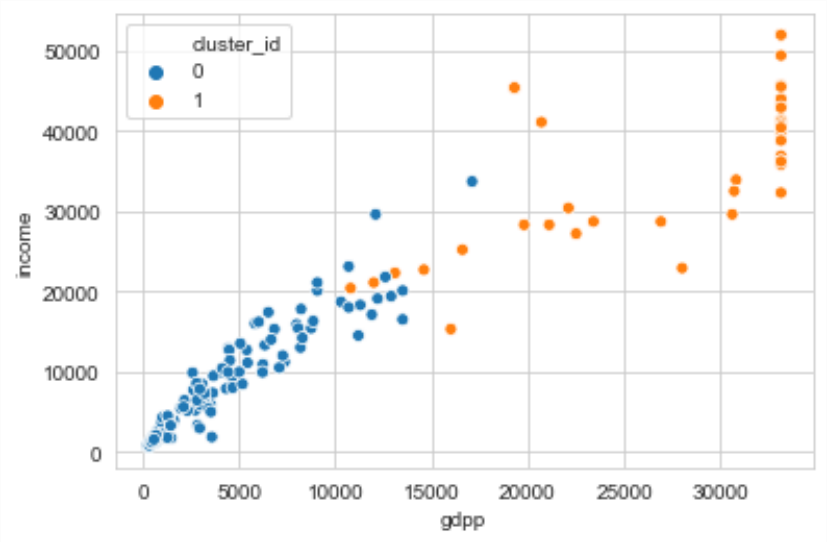
Column Means

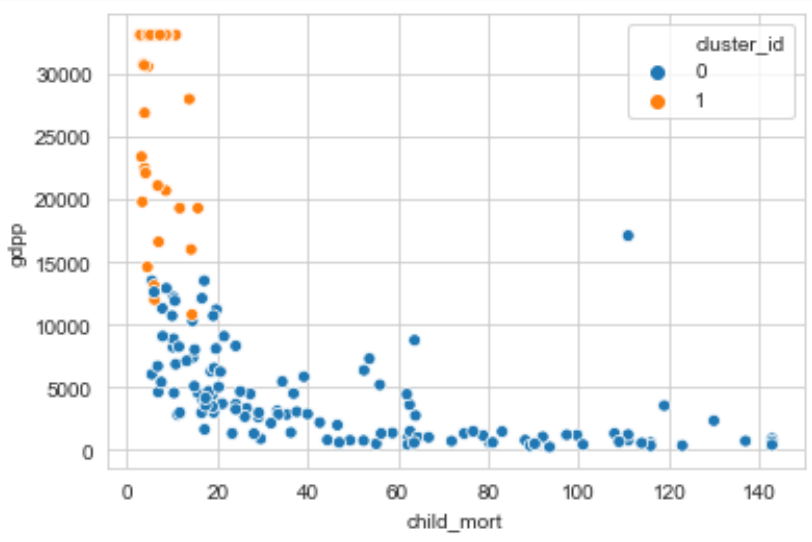
```
# mean of colums - gdpp, income, child_mort
ctr_df.groupby('cluster_id')[['gdpp', 'income', 'child_mort']].mean()
```

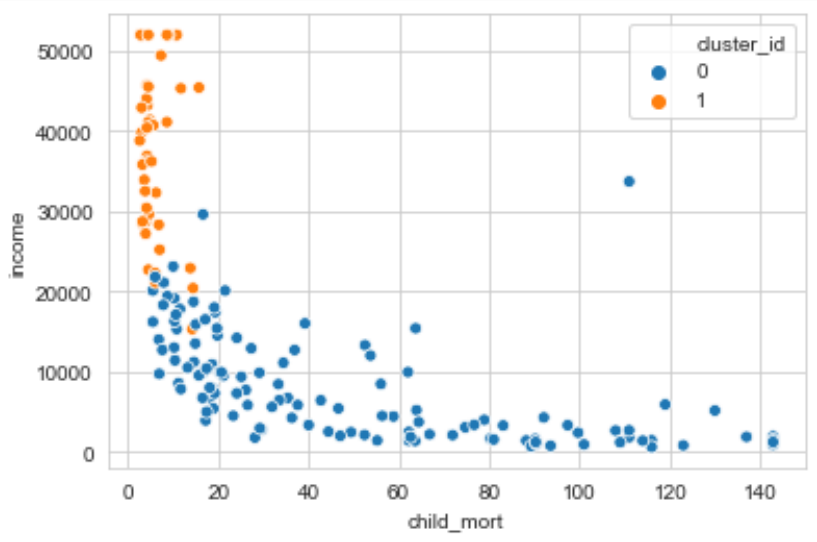
cluster_id	gdpp	income	child_mort
0	4082.227642	8159.048780	49.017886
1	27876.136364	37621.363636	6.054545

The mean of gdpp, income and child_mort is low in case of cluster_labels = 0 and hight in case of cluster_labels = 1

Visualising the clusters

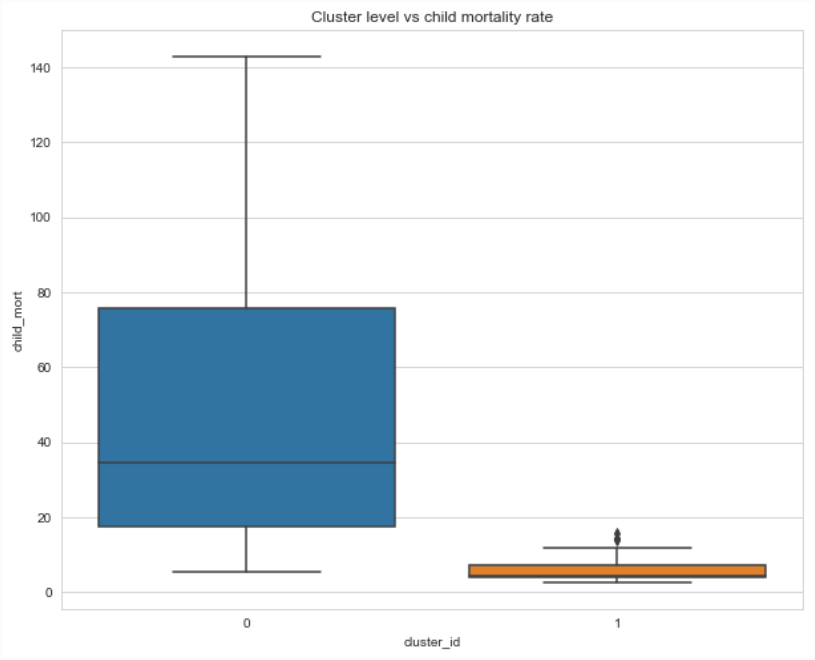


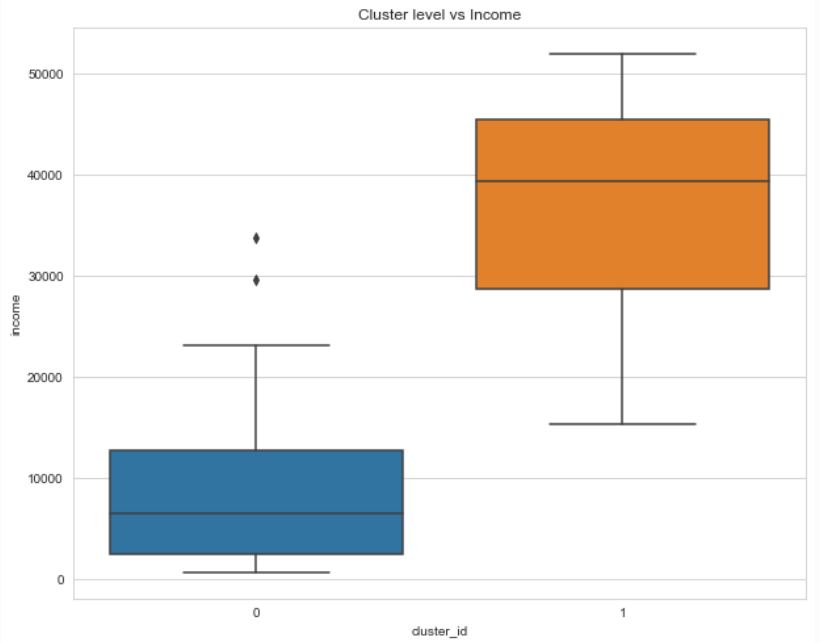


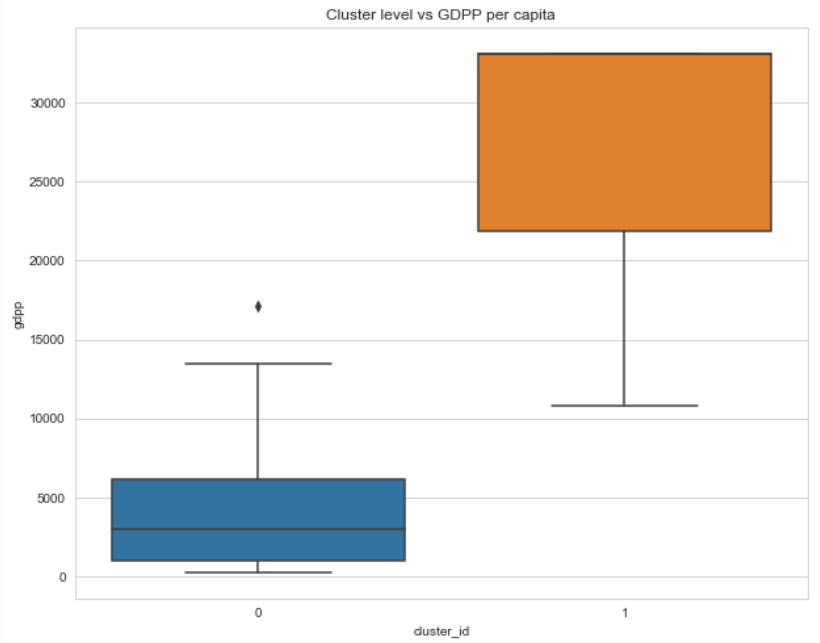


The above Visualization shows us that the gdpp, income and child_mort of cluster having cluster_id = 0 our very bad, while social and econmoic indicators of cluster having cluster_id = 1 are good.

Visualising Distribution of Cluster Labels







Cluster level vs child mortality rate

Cluster level vs income

Cluster level vs GDPP per capita

The above Visualization shows us that the gdpp, income and child_mort of cluster having cluster_id = 0 our very bad, while social and econmoic indicators of cluster having cluster_id = 1 are good.

Countries Obtained After K-means Clustering

country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_id	cluster_labels
Burundi	93.600	20.6052	26.7960	90.552	764.0	12.30	57.7	6.2600	231.0	0	0
Liberia	89.300	62.4570	38.5860	302.802	700.0	5.47	60.8	5.0200	327.0	0	0
Congo, Dem. Rep.	116.000	137.2740	26.4194	165.664	609.0	20.80	57.5	6.5400	334.0	0	0
Niger	123.000	77.2560	17.9568	170.868	814.0	2.55	58.8	7.0075	348.0	0	0
Sierra Leone	142.875	67.0320	52.2690	137.655	1220.0	17.20	55.0	5.2000	399.0	0	0

The K-means clustering shows us that the Countries that are in direst need of aid is 123. However, as per our analysis the top 5 countries that require aid are Burundi, Liberia, Congo, Niger and Sierra Leone. These 5 countries have very lowest social and economic indicators among the countries with low social and economic indicators.

Hierarchical clustering

The process of Hierarchical Clustering involves either clustering sub-clusters (data points in the first iteration) into larger clusters in a bottom-up manner or dividing a larger cluster into smaller subclusters in a top-down manner. During both the types of hierarchical clustering, the distance between two sub-clusters needs to be computed. The different types of linkages describe the different approaches to measure the distance between two sub-clusters of data points. The different types of linkages are: -

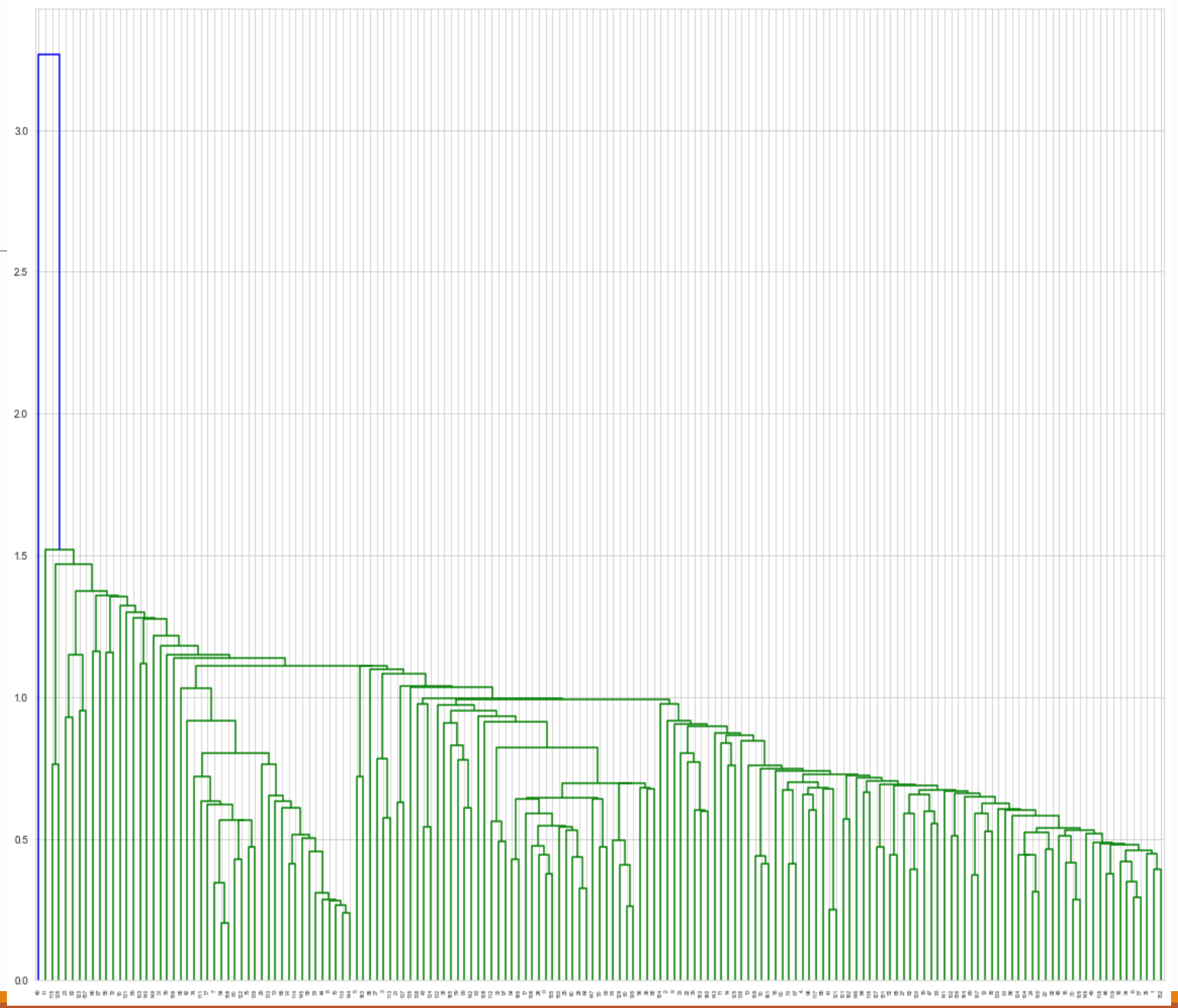
Single Linkage: Here, the distance between 2 clusters is defined as the shortest distance between points in the two clusters.

Complete Linkage: Here, the distance between 2 clusters is defined as the maximum distance between any 2 points in the clusters.

Average Linkage: Here, the distance between 2 clusters is defined as the average distance between every point of one cluster to every other point of the other cluster.

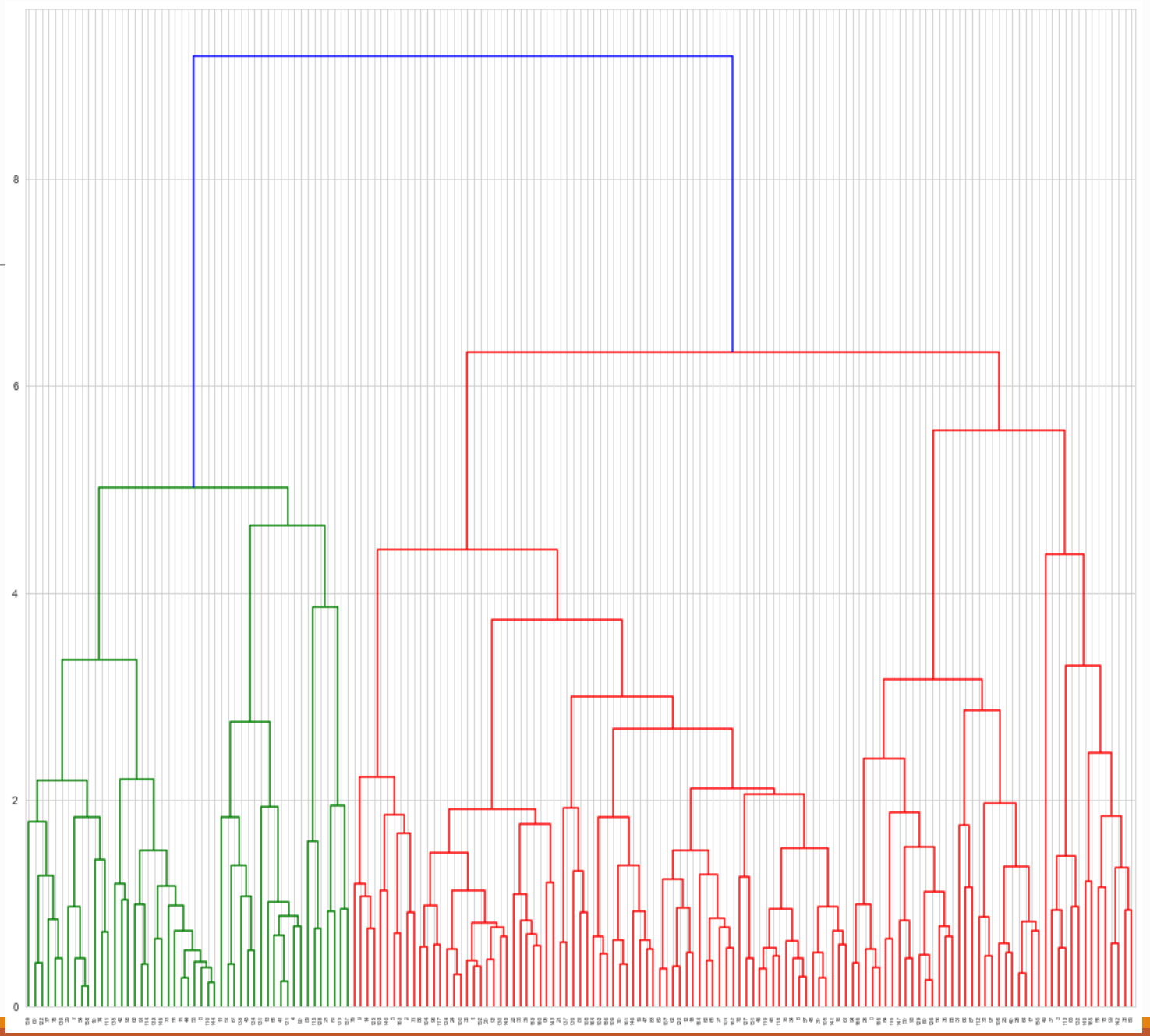
Single Linkage

It's not clearly visible in single linkage dendogram hence we will go for complete linkage which gives us proper results.

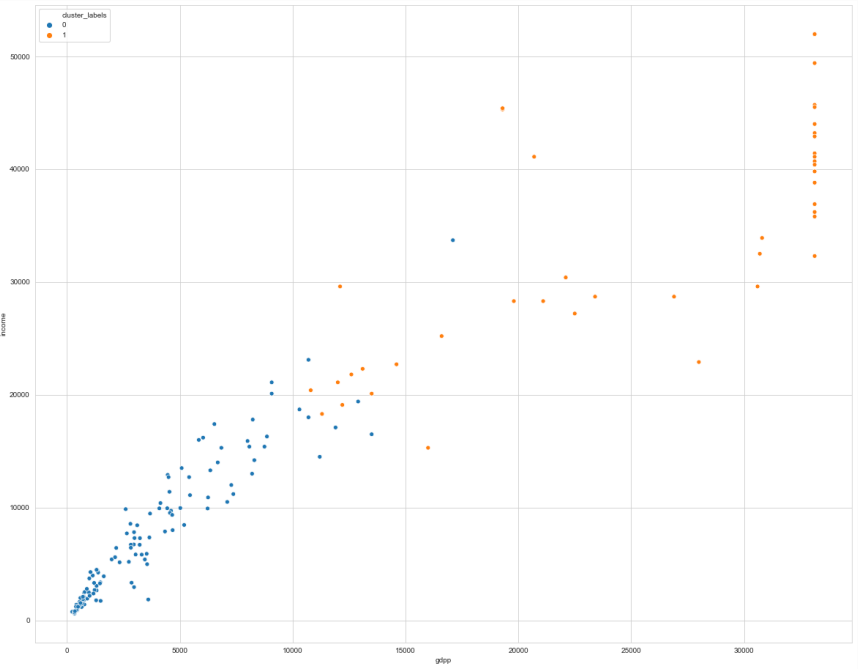


Complete Linkage

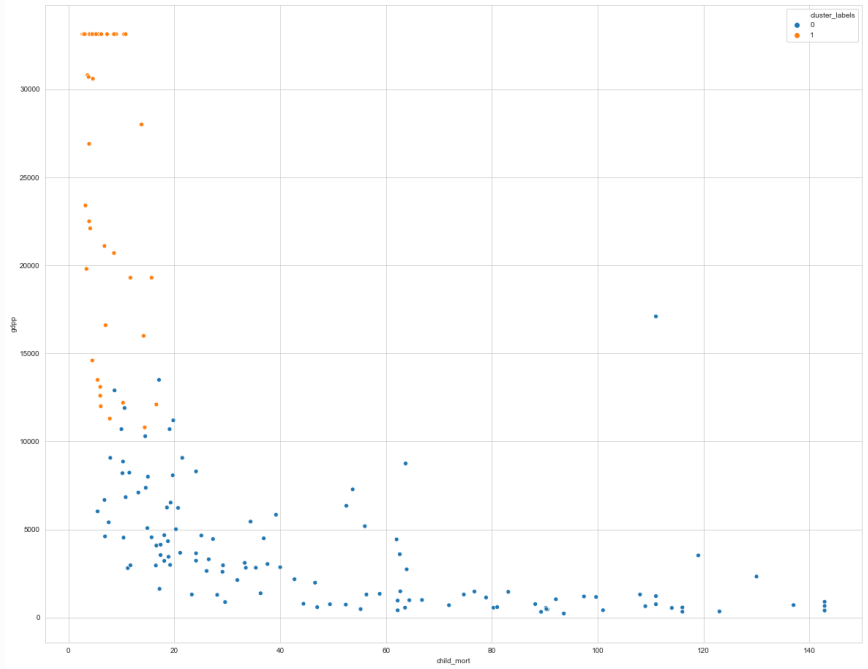
The complete linkage shows us that max distance. So we cut the dendrogram into 2 clusters.



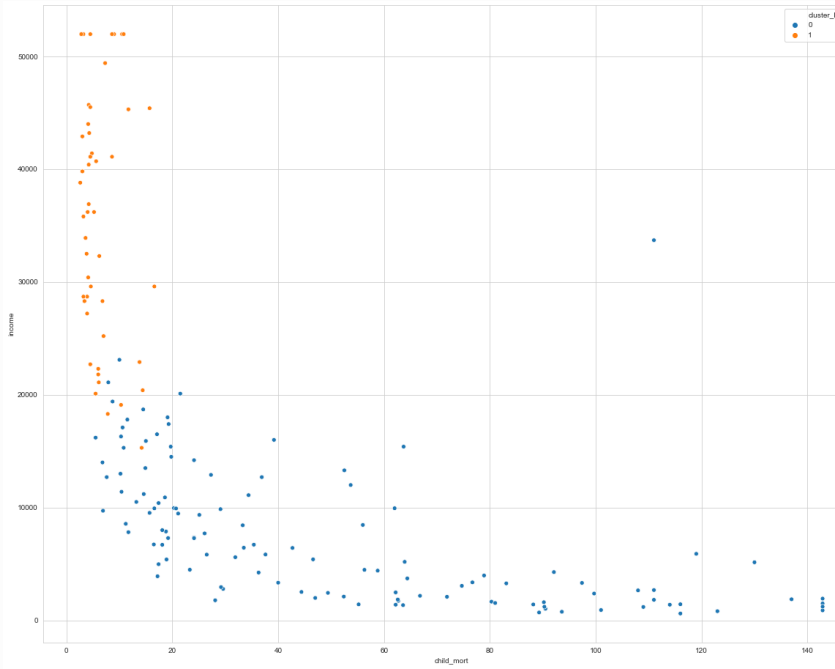
Visualising clusters obtained after hierarchical clustering



GDPP vs income



Child mortality rate vs GDPP



Child mortality rate vs income

The above Visualization shows us that the gdpp, income and child_mort of cluster having cluster_id = 0 our very bad, while social and econmoic indicators of cluster having cluster_id = 1 are good.

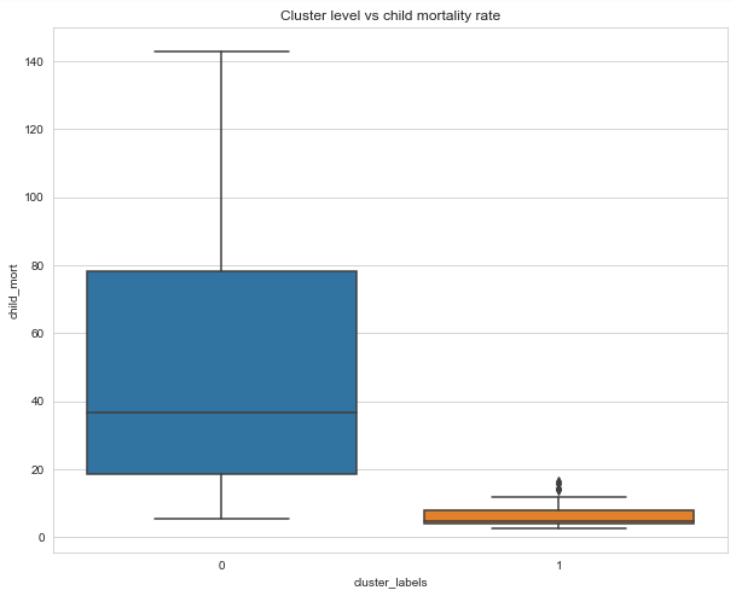
Column Means

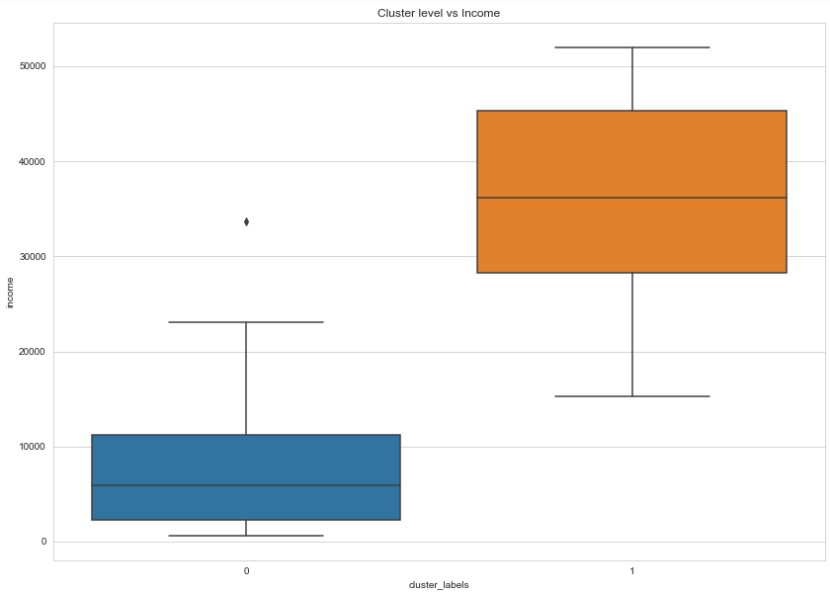
```
# mean of colums - gdpp, income, child_mort
ctr_df.groupby('cluster_labels')[['gdpp', 'income', 'child_mort']].mean()
```

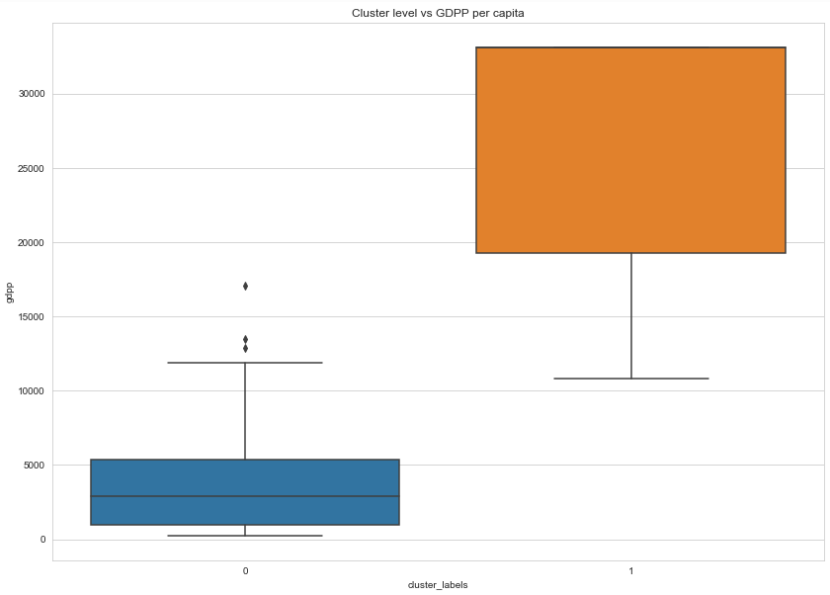
cluster_labels	gdpp	income	child_mort
0	3732.322034	7581.889831	50.703390
1	26290.816327	36004.897959	6.379592

The mean of gdpp, income and child_mort is low in case of cluster_labels = 0 and hight in case of cluster_labels = 1

Visualising Distribution of Cluster Labels







Cluster level vs child mortality rate

Cluster level vs income

Cluster level vs GDPP per capita

All the above visualization shows us that the social and economic indicators of cluster having cluster_id = 0 our very bad, while social and econmoic indicators of cluster having cluster_id = 1 are good.

Countries Obtained After Hierarchical Clustering

country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_id	cluster_labels
Burundi	93.600	20.6052	26.7960	90.552	764.0	12.30	57.7	6.2600	231.0	0	0
Liberia	89.300	62.4570	38.5860	302.802	700.0	5.47	60.8	5.0200	327.0	0	0
Congo, Dem. Rep.	116.000	137.2740	26.4194	165.664	609.0	20.80	57.5	6.5400	334.0	0	0
Niger	123.000	77.2560	17.9568	170.868	814.0	2.55	58.8	7.0075	348.0	0	0
Sierra Leone	142.875	67.0320	52.2690	137.655	1220.0	17.20	55.0	5.2000	399.0	0	0

The hierarchal clustering shows us that the Countries that are in direst need of aid is 118. However, as per our analysis the top 5 countries that require aid are Burundi, Liberia, Congo, Niger and Sierra Leone. These 5 countries have very lowest social and economic indicators among the countries with low social and economic indicators.

Final Conclusion

Both the clustering suggests same top 5 countries that require aid. Thus, we draw conclusion that final list of countries that require AID from HELP International. (top 5 countries):

1. **Burundi**
2. **Liberia**
3. **Congo**
4. **Niger**
5. **Sierra Leone**