HELP International

A Humanitarian NGO

By
PAVAN S

REQUIREMENT

HELP has raised $10 million through various funding programs. They plan to use this fund strategically as well as effectively to help the countries that are in dire need of help.

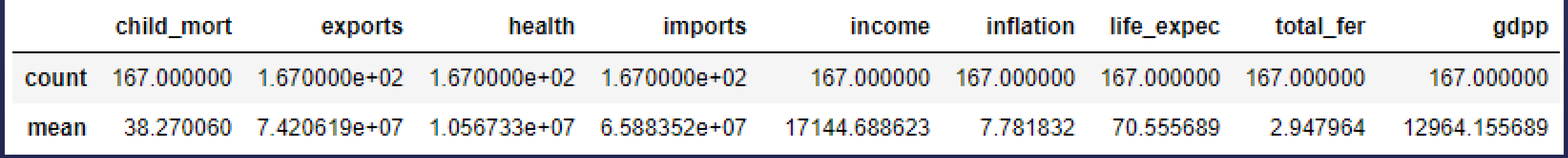
Please help the NGO in choosing the top 5 countries that they need to focus the most taking into consideration the various socio-economic and health factors of the respective countries from the provided data.

STEPS FOLLOWED TO OBTAIN THE OUTCEME

- Data Preparation
- Data Visualization
- Outlier Analysis
- Hopkins Check and Scaling
- Clustering
    - Kmeans Clustering
    - Hierarchical Clustering
- Conclusion

Data Preparation

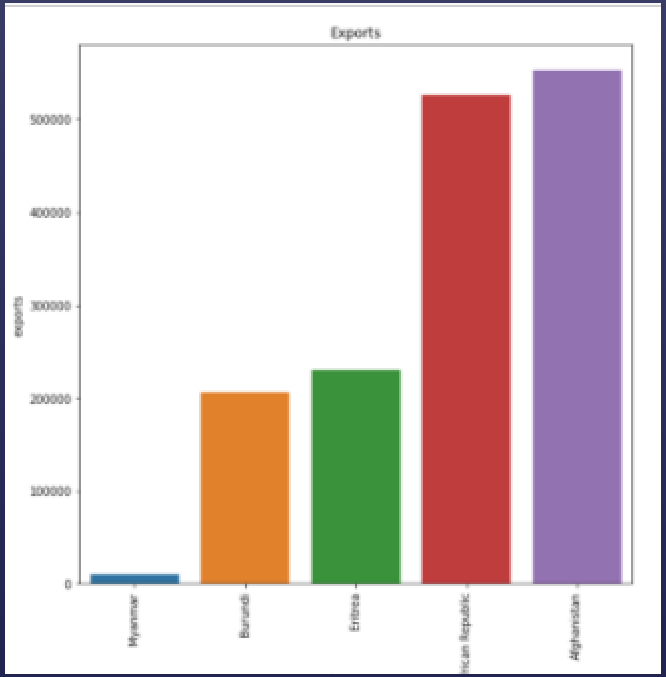
- Converting the factors exports , imports and health in to actual values as they are given as the percentage of the GDPP
- Missing Values check , there were no missing values in the data
- Numerical characteristics of the data , as we see below the mean of the factors are not in a particular scale , for which scaling will be performed in order to help the analysis of data

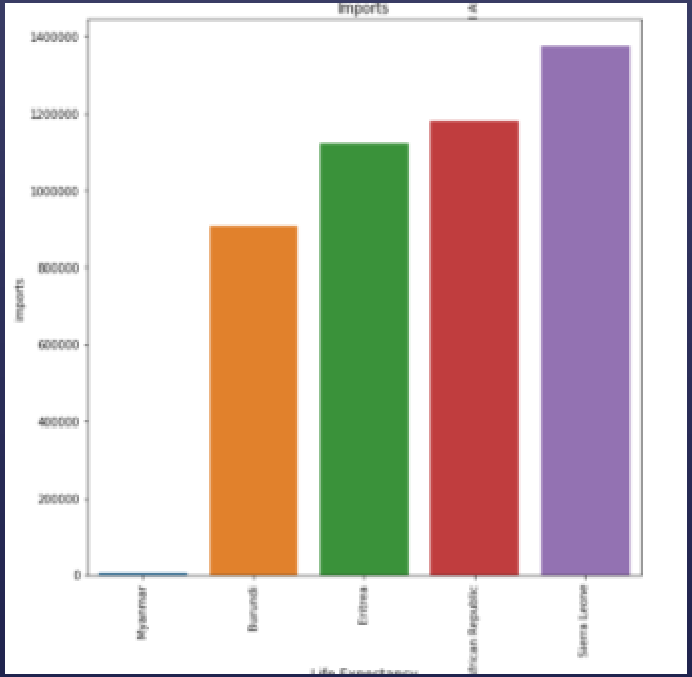


	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
count	167.000000	1.670000e+02	1.670000e+02	1.670000e+02	167.000000	167.000000	167.000000	167.000000	167.000000
mean	38.270060	7.420619e+07	1.056733e+07	6.588352e+07	17144.688623	7.781832	70.555689	2.947964	12964.155689

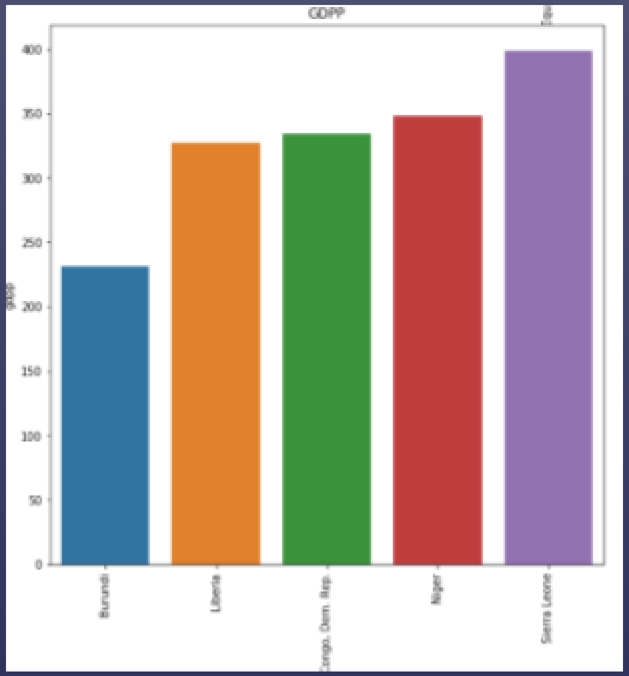
DATA VISULIZATION

- We have compared top 5 countries of each categories based on which that factor alone will the country need help

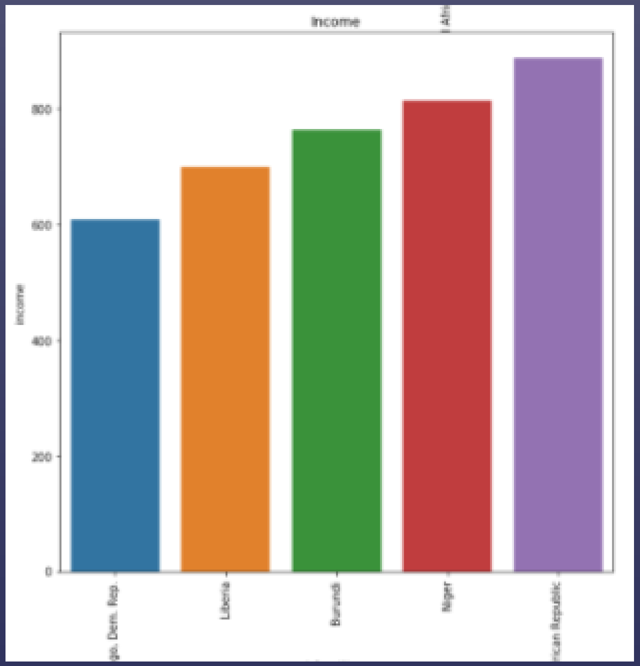




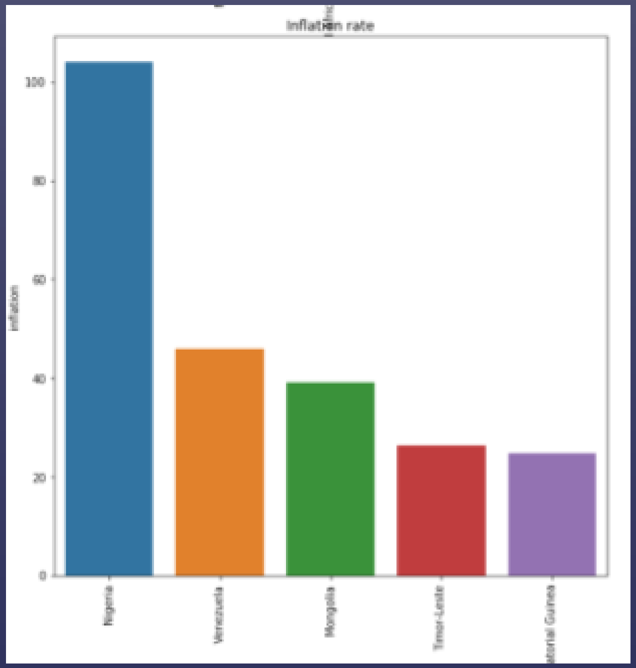
Myanmar has the least export and import services among all the countries.



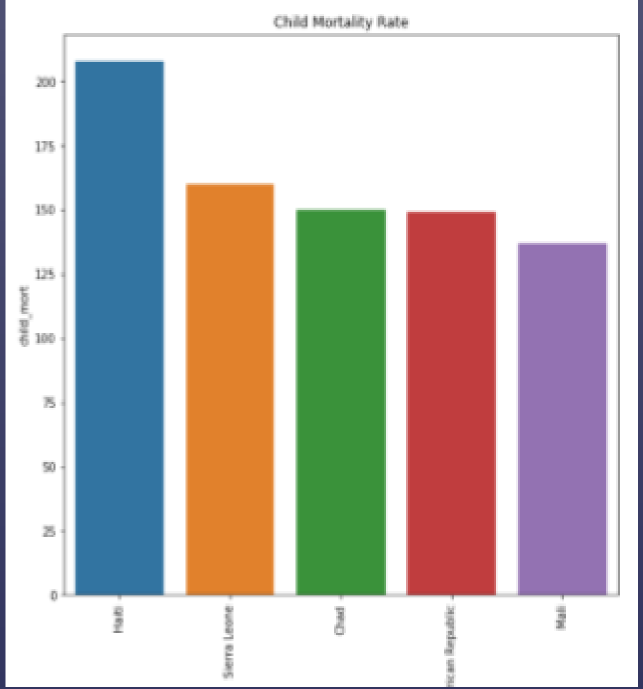
The country with the least GDP per capita is Burundi

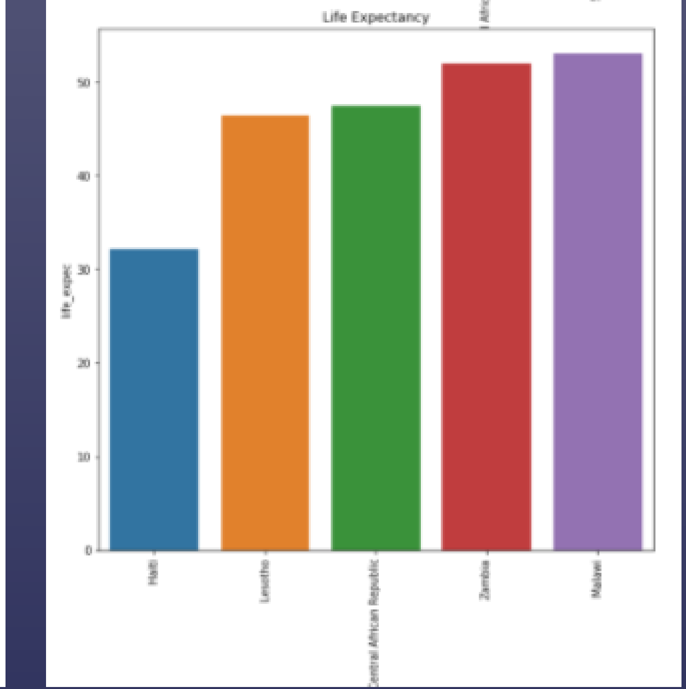


The country with the least per capita income is Congo Democratic Republic

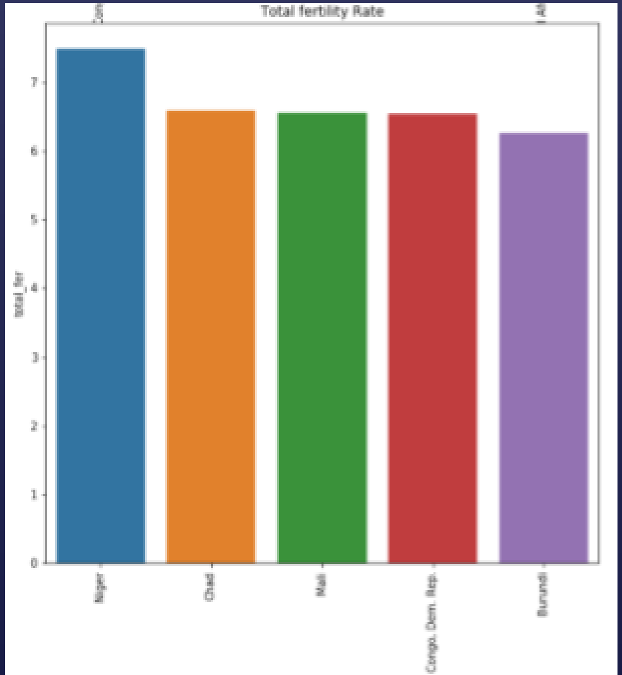


Nigeria has the highest inflation rate among all



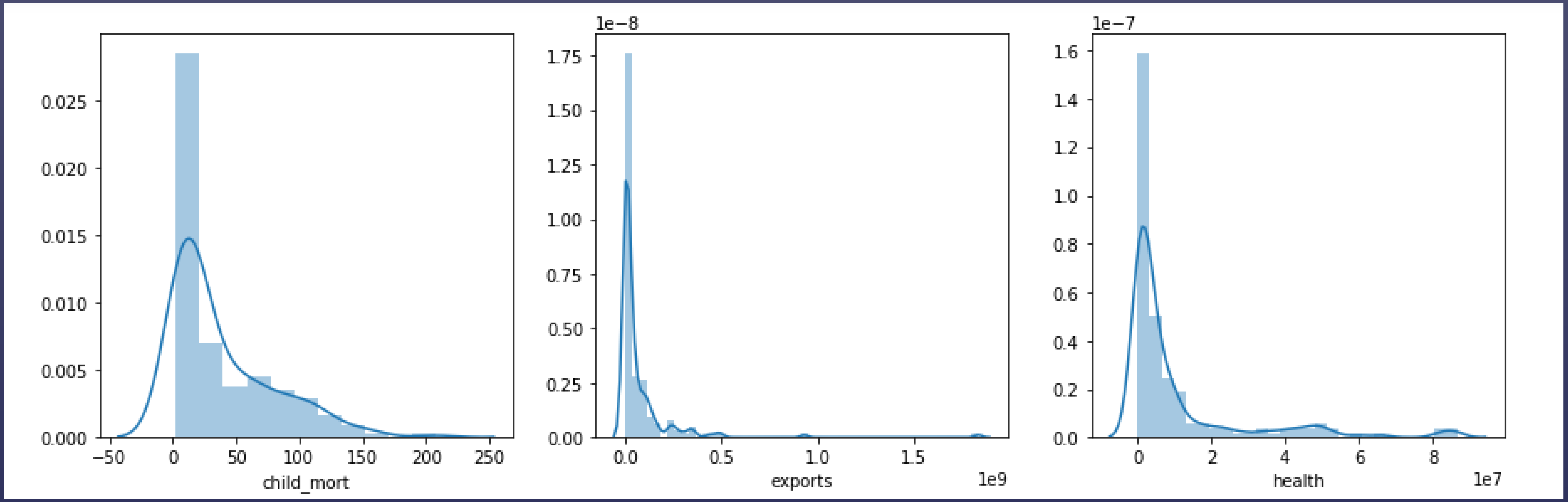


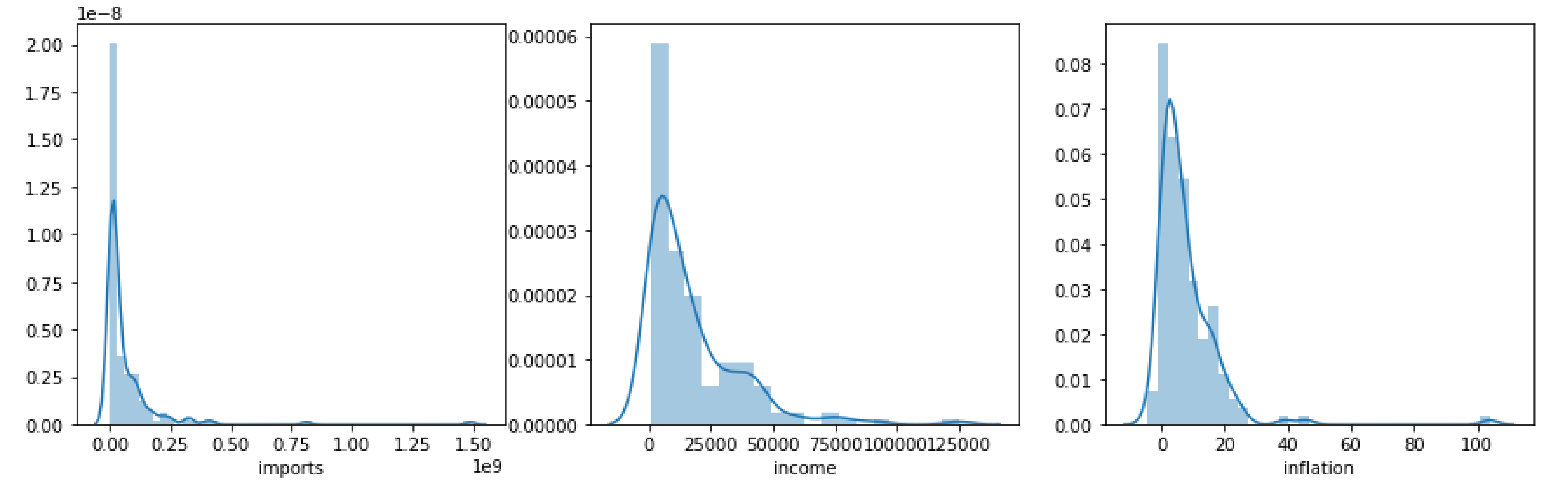
Haiti has highest child mortality rate and the least life expectancy among all.

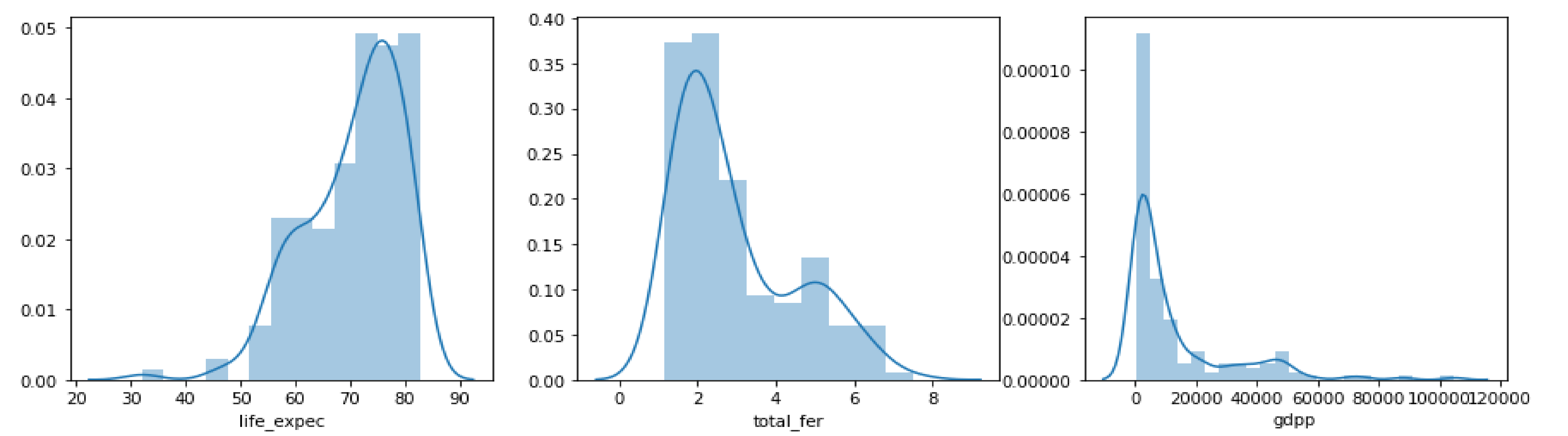


Niger has the highest total fertility rate.

Univariate Analysis

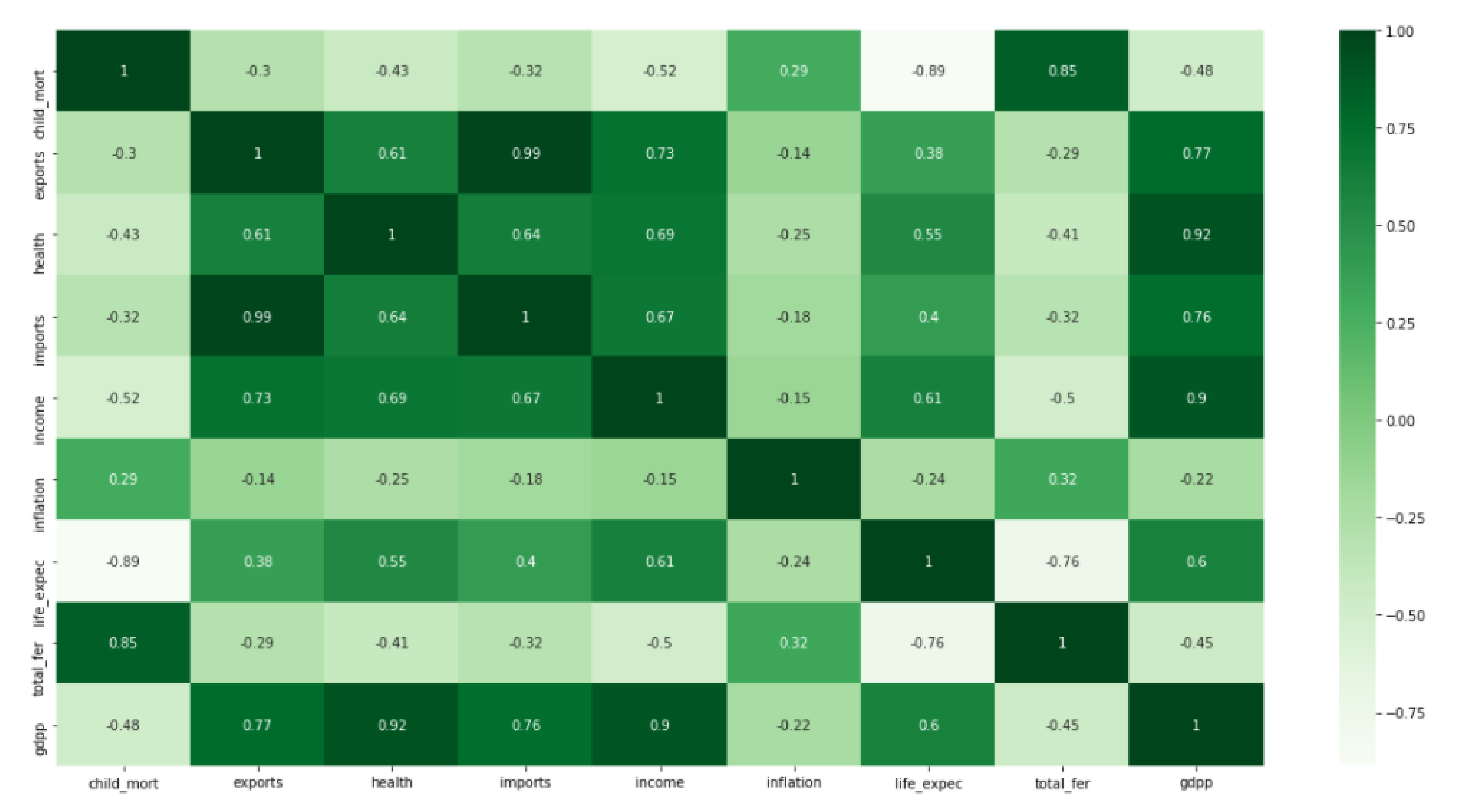






- All the factors are left skewed (where the mean of the data is less than the median) except for the life expectancy
- Except for total_fertility and GDP per capita all the factors have unimodal distribution where the latter have bimodal distribution (representing two groups of data)

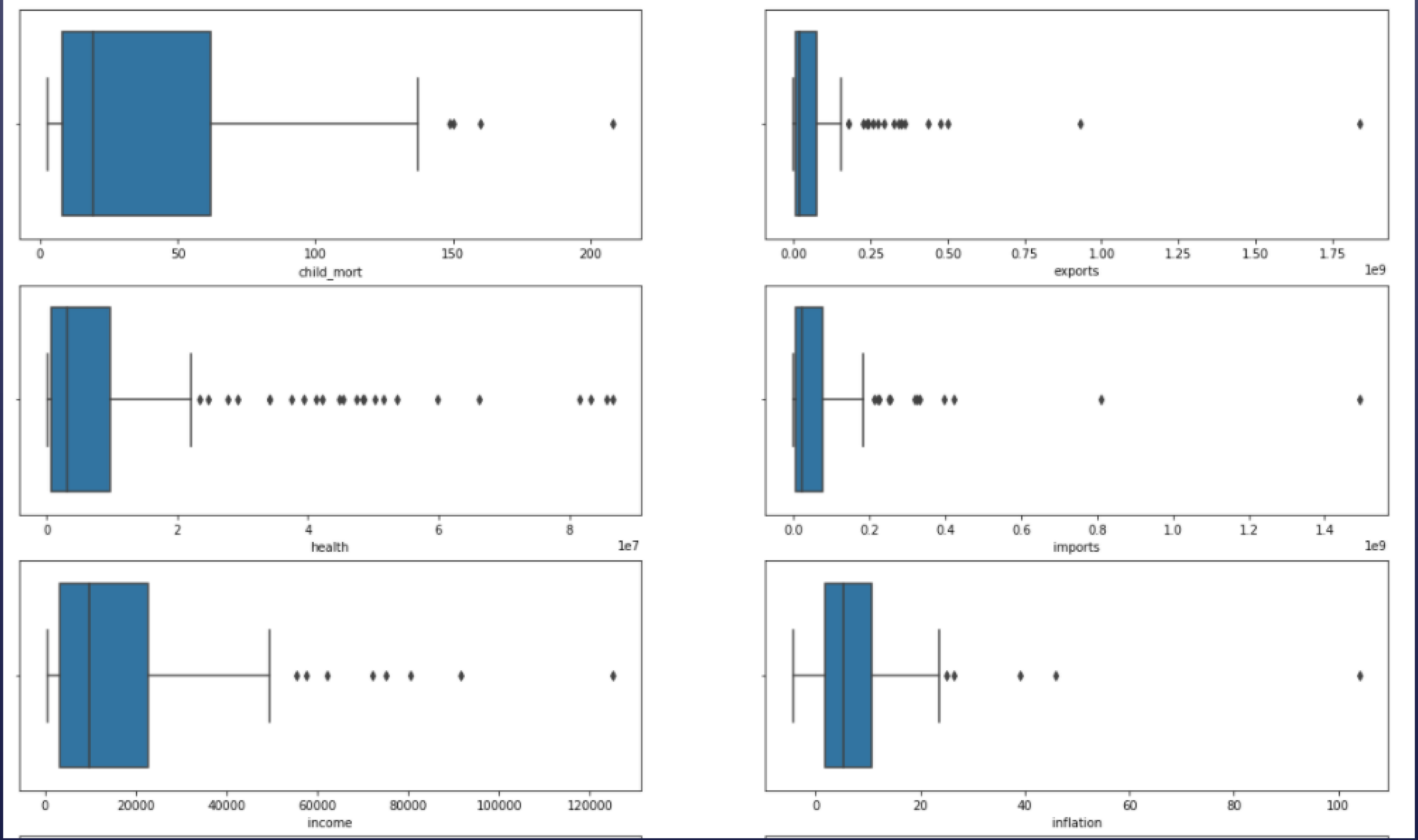
Bivariate Analysis

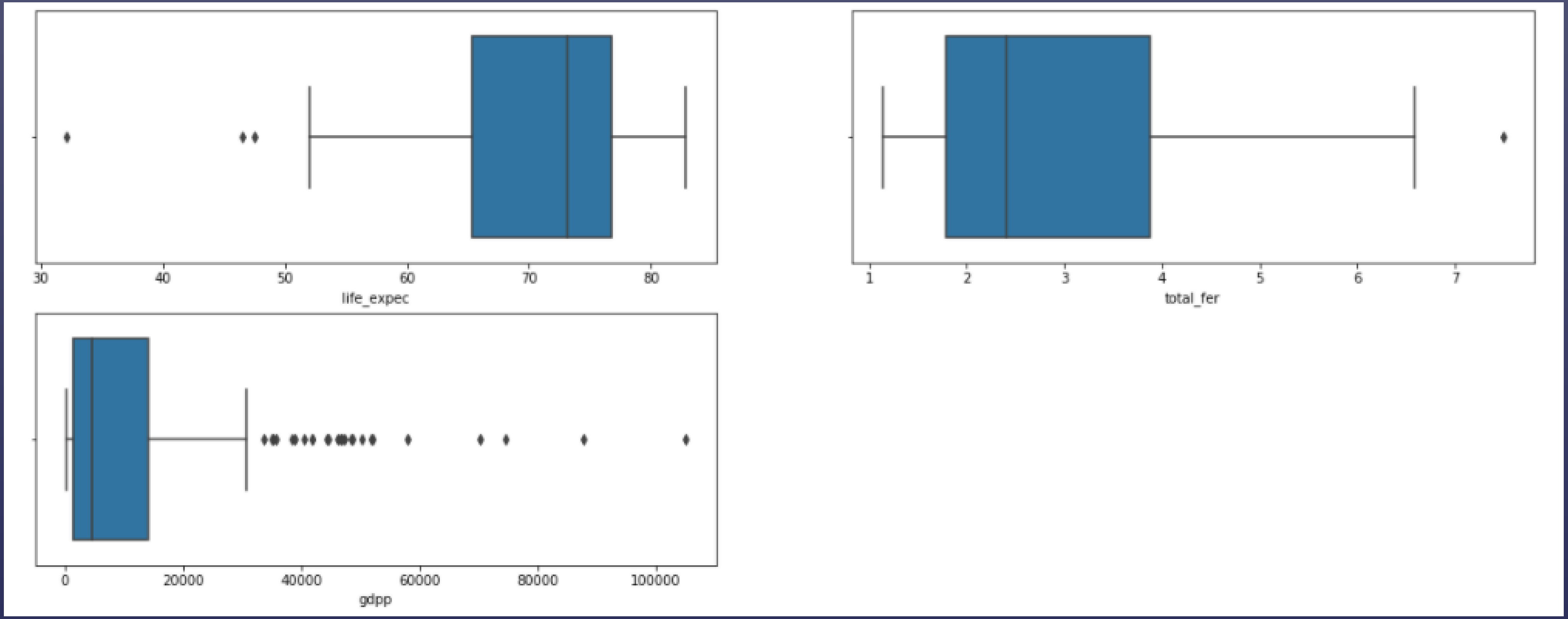


	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
child_mort	1	-0.3	-0.43	-0.32	-0.52	0.29	-0.89	0.85	-0.48
exports	-0.3	1	0.61	0.99	0.73	-0.14	0.38	-0.29	0.77
health	-0.43	0.61	1	0.64	0.69	-0.25	0.55	-0.41	0.92
imports	-0.32	0.99	0.64	1	0.67	-0.18	0.4	-0.32	0.76
income	-0.52	0.73	0.69	0.67	1	-0.15	0.61	-0.5	0.9
inflation	0.29	-0.14	-0.25	-0.18	-0.15	1	-0.24	0.32	-0.22
life_expec	-0.89	0.38	0.55	0.4	0.61	-0.24	1	-0.76	0.6
total_fer	0.85	-0.29	-0.41	-0.32	-0.5	0.32	-0.76	1	-0.45
gdpp	-0.48	0.77	0.92	0.76	0.9	-0.22	0.6	-0.45	1

- imports and exports have a high correlation of 0.99
- child_mort and total_fer have high correlation of 0.85
- child_mort and life_expec has negative correlation of -0.89
- life_expec and child_mort has negative correlation of -0.76
- If the gdpp is high
  - - child mortality is low
  - - Income is high
  - - Inflation is low
  - - Life expectancy is high
  - - Total fertility is low

OUTLIER ANALYSIS





- GDPP and Health have too many outliers
- Higher outlier in child mortality and lower outlier in life expectancy are not capped as these countries may miss out on help
- Export, imports, health, income, inflation, gdpp and total fertility are soft capped (but not removed as they have business requirement

HOPKIN’s CHECK

- We measure the clustering tendency(whether the clusters can be formed or not) of the data set through a parameter called Hopkin’s check.

- If HC < 0.5 Data set has bad tendency for clustering ,nearing 0 no tendency

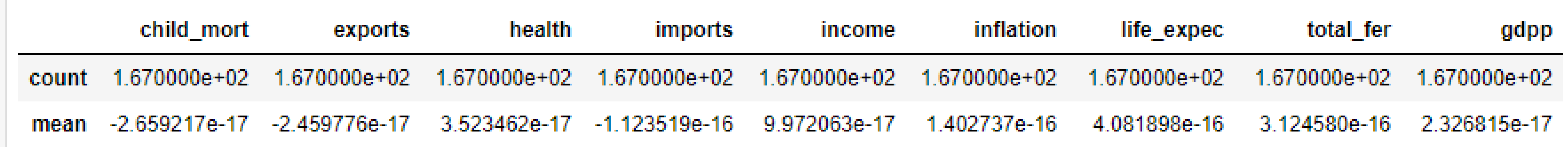
- Else HC > 0.75 Data set has good tendency for clustering , nearing 1 perfects clusters can be formed

For the provided data set the Hopkin’s Value is > 0.9 so it exhibits very good clustering tendency

SCALING

- We perform scaling in order to standardize the independent data features . It is necessary to perform scaling else the algorithm tends to consider greater values , higher and smaller values ,lower regardless of the unit of the value.

- We have used Standard Scaler , where it scales the feature so it has a distribution with 0 mean and 1 is variance

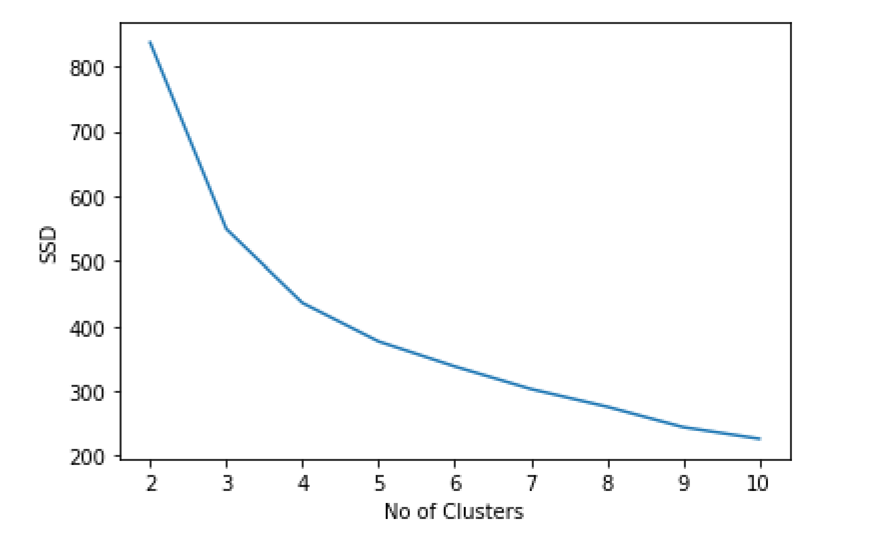


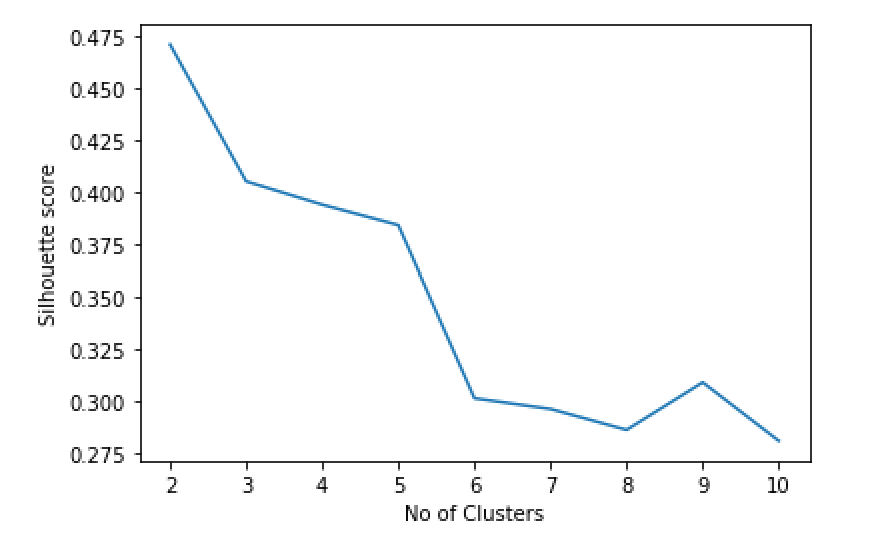
	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
count	1.670000e+02	1.670000e+02	1.670000e+02	1.670000e+02	1.670000e+02	1.670000e+02	1.670000e+02	1.670000e+02	1.670000e+02
mean	-2.659217e-17	-2.459776e-17	3.523462e-17	-1.123519e-16	9.972063e-17	1.402737e-16	4.081898e-16	3.124580e-16	2.326815e-17

CLUSTERING

- **KMeans Clustering (Use the library sklearn)**

- We select a initial K value , for which we have used Silhouette score and Elbow score to select a optimal K value .

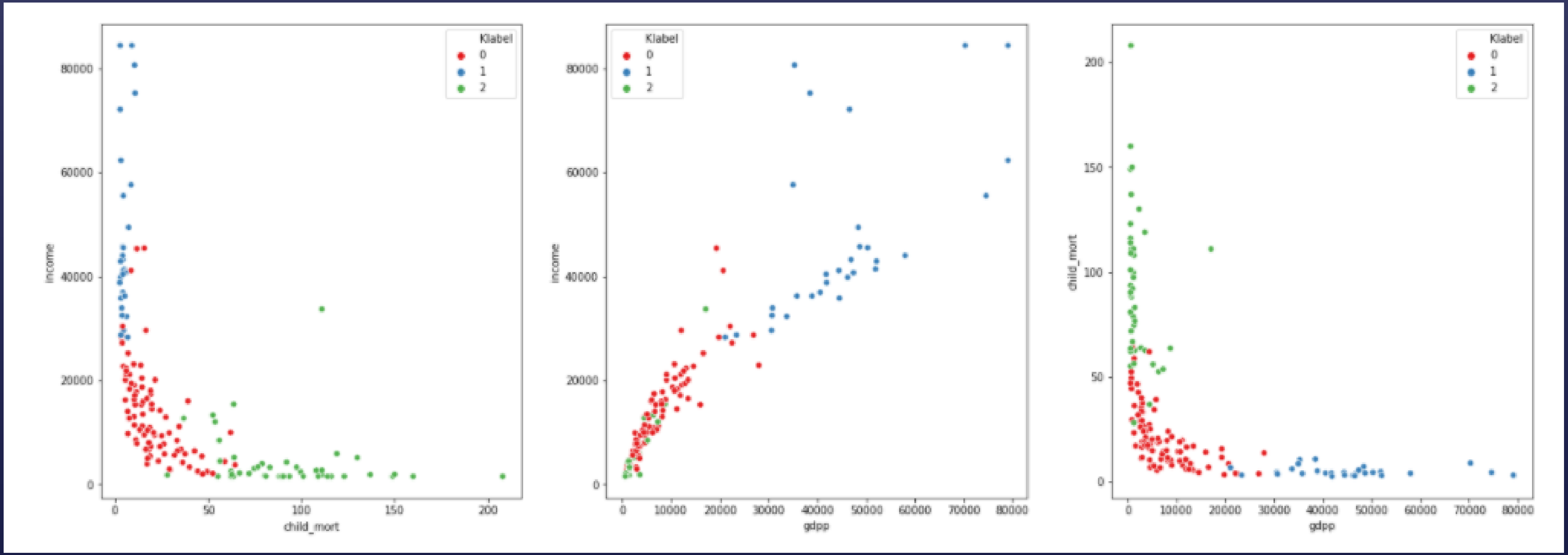


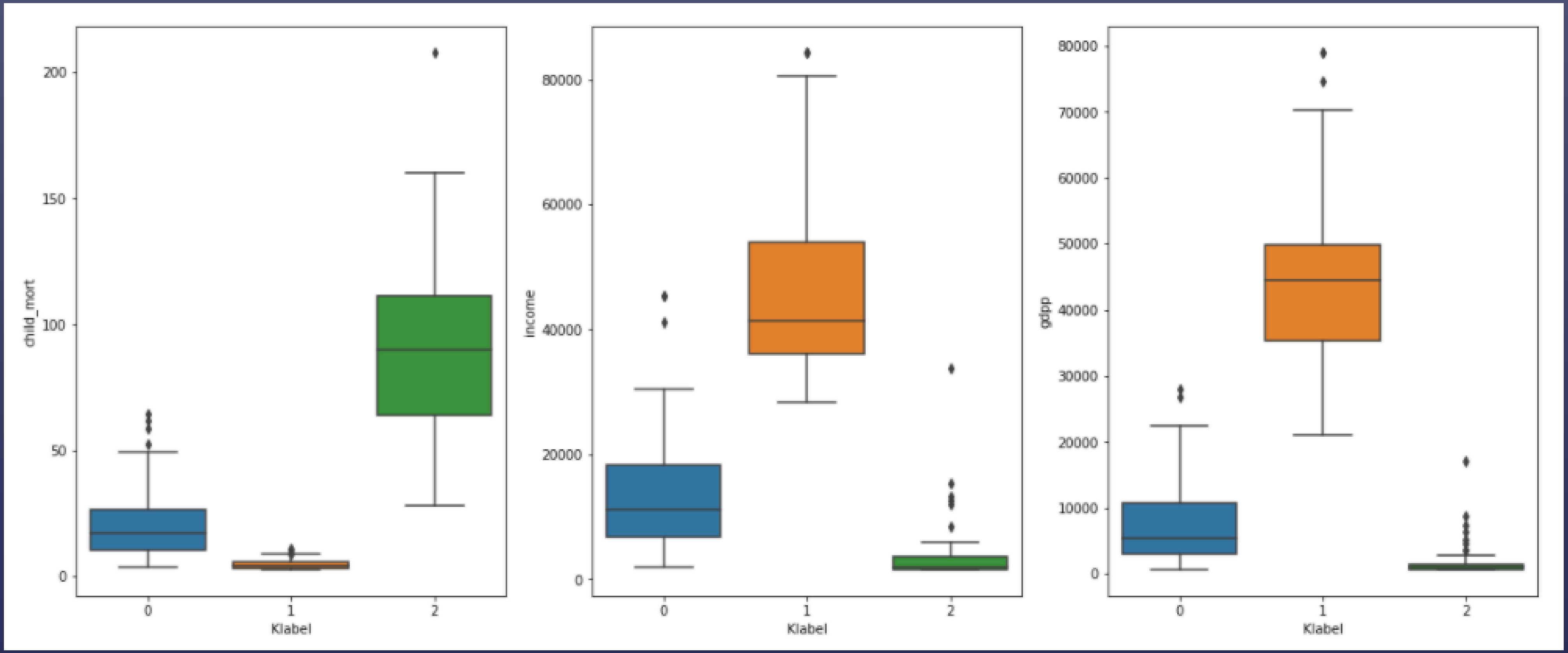


From the above mentioned parameters we select k=3 i.e. mainly data can be categorized into 3 clusters

- It categorizes each item to its nearest mean and after each addition the center of the cluster gets updated
- Process gets repeated to a certain iterations else till we get a stable clusters.

The clusters we have got by using the 3 main factors gdpp , income and child mortality



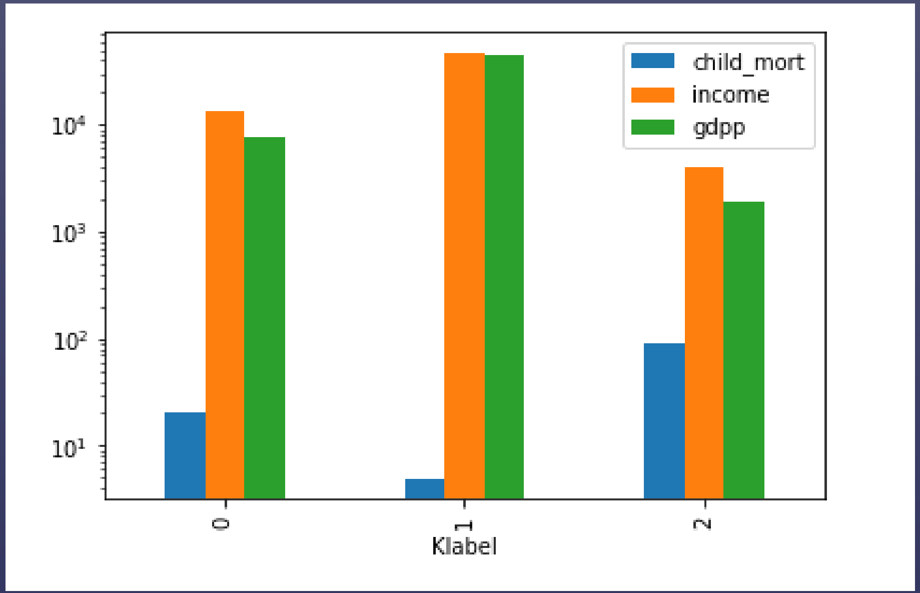


From the above boxplots we confer what does each cluster signifies

Label =0 , low child mortality , income & gdpp - developing countries

Label =1 , low child morality , high income and gdpp - developed countries

Label =2 , high child mortality , low income and gdpp – under developed countries



We select the top 5 countries that belong to cluster label 2 , since they have less income and less GDPP but the child death rate is High

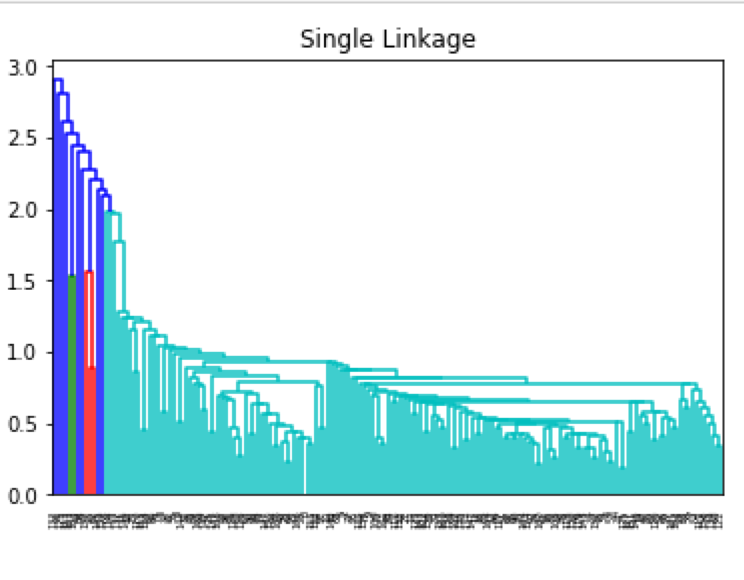
The top 5 countries that NGO needs to help according to Kmeans

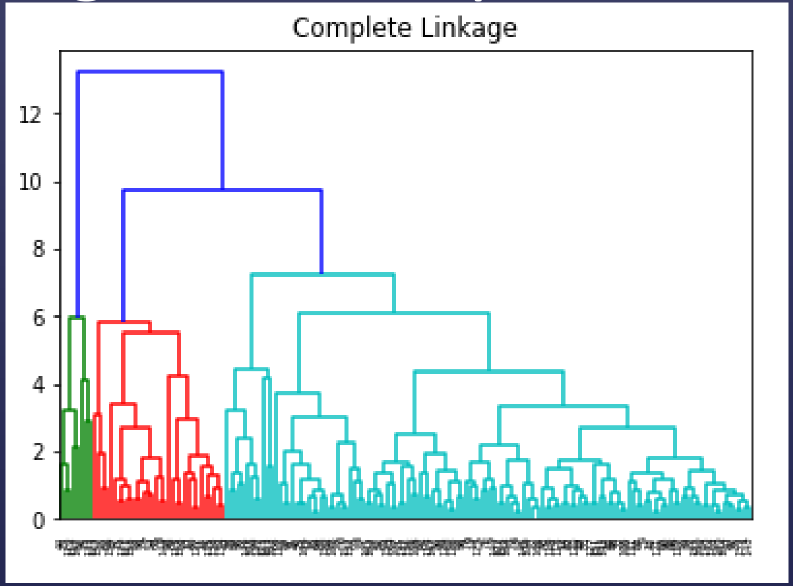
- Sierra Leone
- Central African Republic
- Niger
- Burkina Faso
- Congo, Dem. Rep

- **Hierarchical Clustering (Use the library sklearn)**

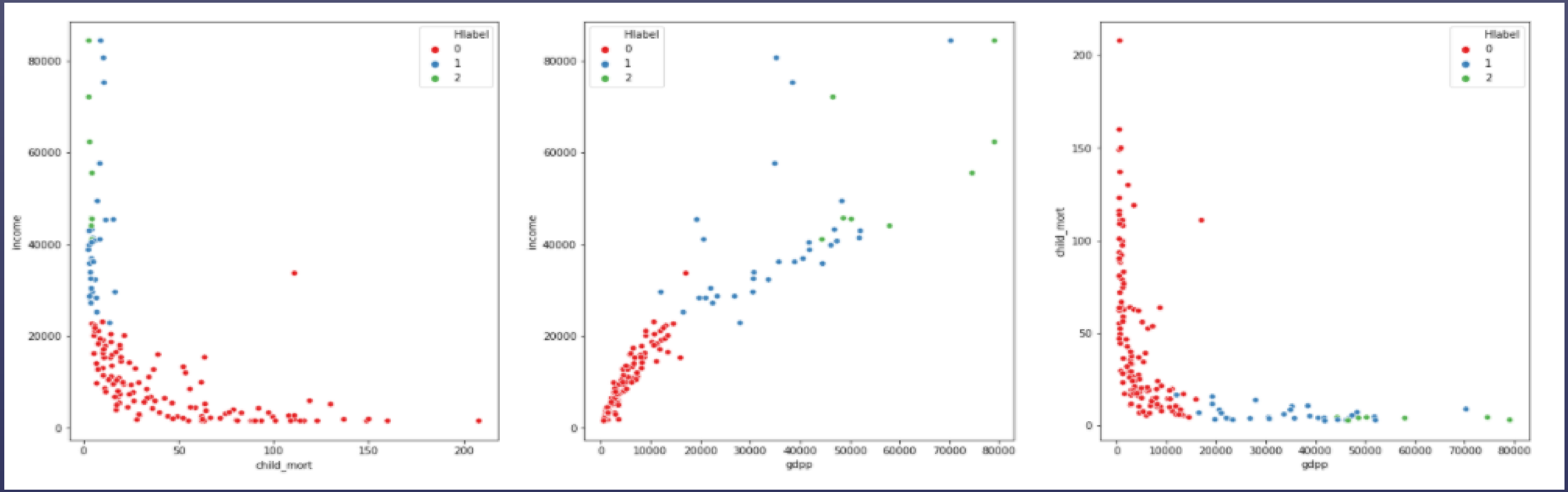
- In this technique we obtain a tree structure /dendrogram for similar data points.

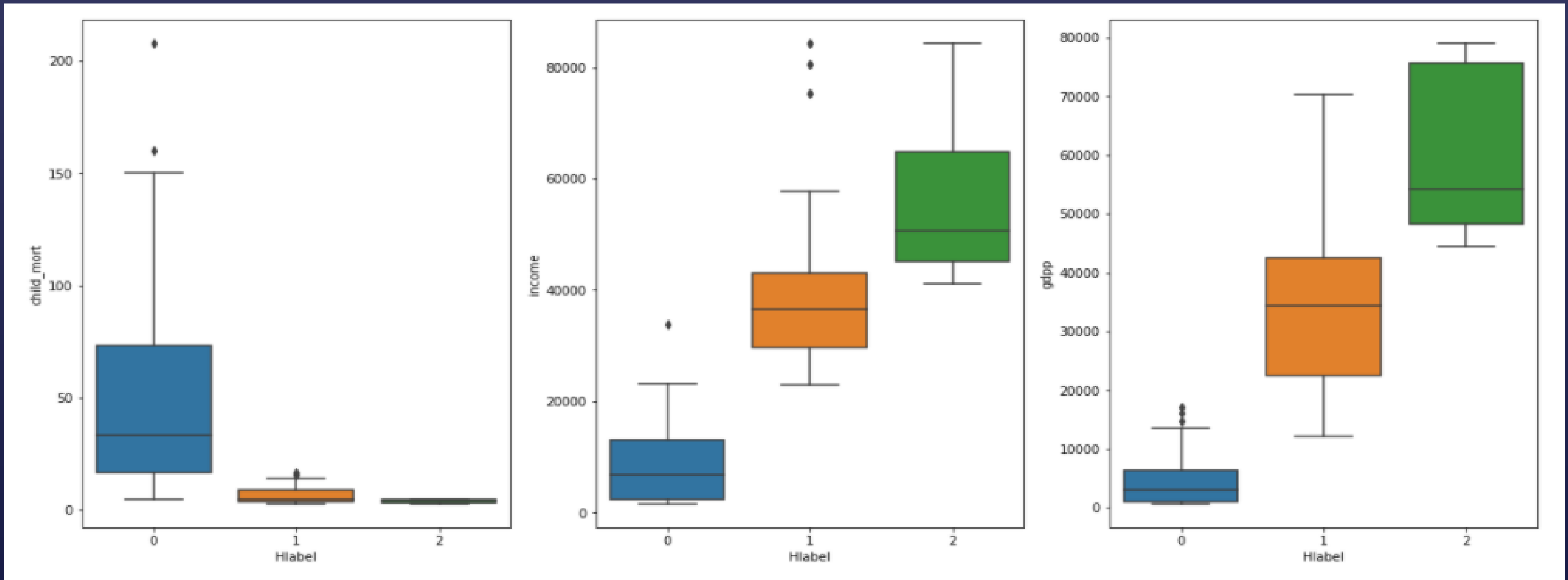
- We have performed single linkage and complete linkage



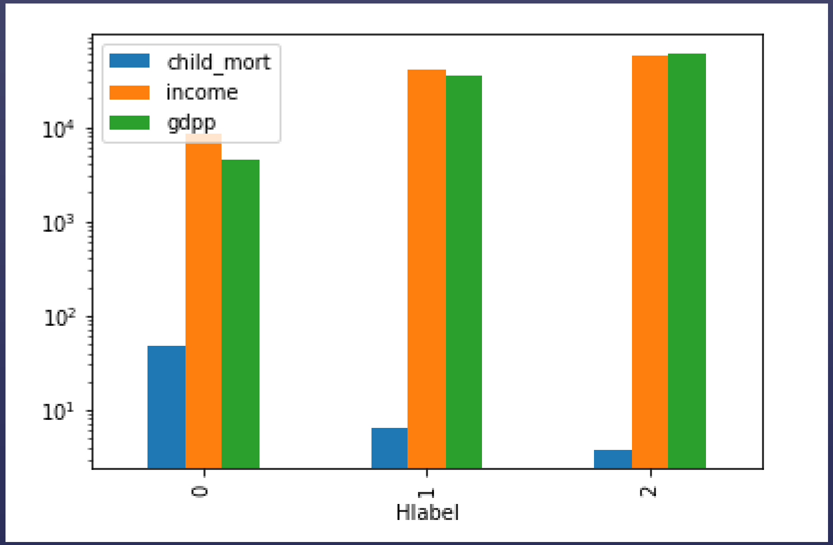


- Complete linkage provides a more structured view than that of single linkage structure


Hlabel
0
1
2
income
child_mort
0
20000
40000
60000
80000
0
50
100
150
200
gdpp
0
10000
20000
30000
40000
50000
60000
70000
80000
child_mort


child_mort
0
50
100
150
200
Hlabel
0
1
2
income
0
20000
40000
60000
80000
gdpp
0
10000
20000
30000
40000
50000
60000
70000
80000

- We are not able to distinguish the cluster categories since the cluster size varies significantly. So we have performed cluster profiling to get more insights to which cluster signifies what.



Label =0 , high child mortality , low income and gdpp – under developed countries

Label = 1 , low child mortality , high income and gdpp – developed countries

Label = 2 ,medium child mortality ,income and gdpp – developing countries

- Even though we have divided into 3 clusters what the cluster signifies is ambiguous even after the visualization.

We select the top 5 countries that belong to cluster label 0 , since they have less income and less GDPP but the child death rate is High

The top 5 countries that NGO needs to help according to Hierarchical clustering

- Sierra Leone
- Central African Republic
- Niger
- Burkina Faso
- Congo, Dem. Rep

Conclusion

- We prefer Kmeans over hierarchical clustering because
  - We get more significant plots in Kmeans
  - We can particularly divide the clusters in to under developed , developing and developed countries.

The 5 countries that on which you need to be focused on helping are

- Sierra Leone
- Central African Republic
- Niger
- Burkina Faso
- Congo, Dem. Rep

THANK YOU