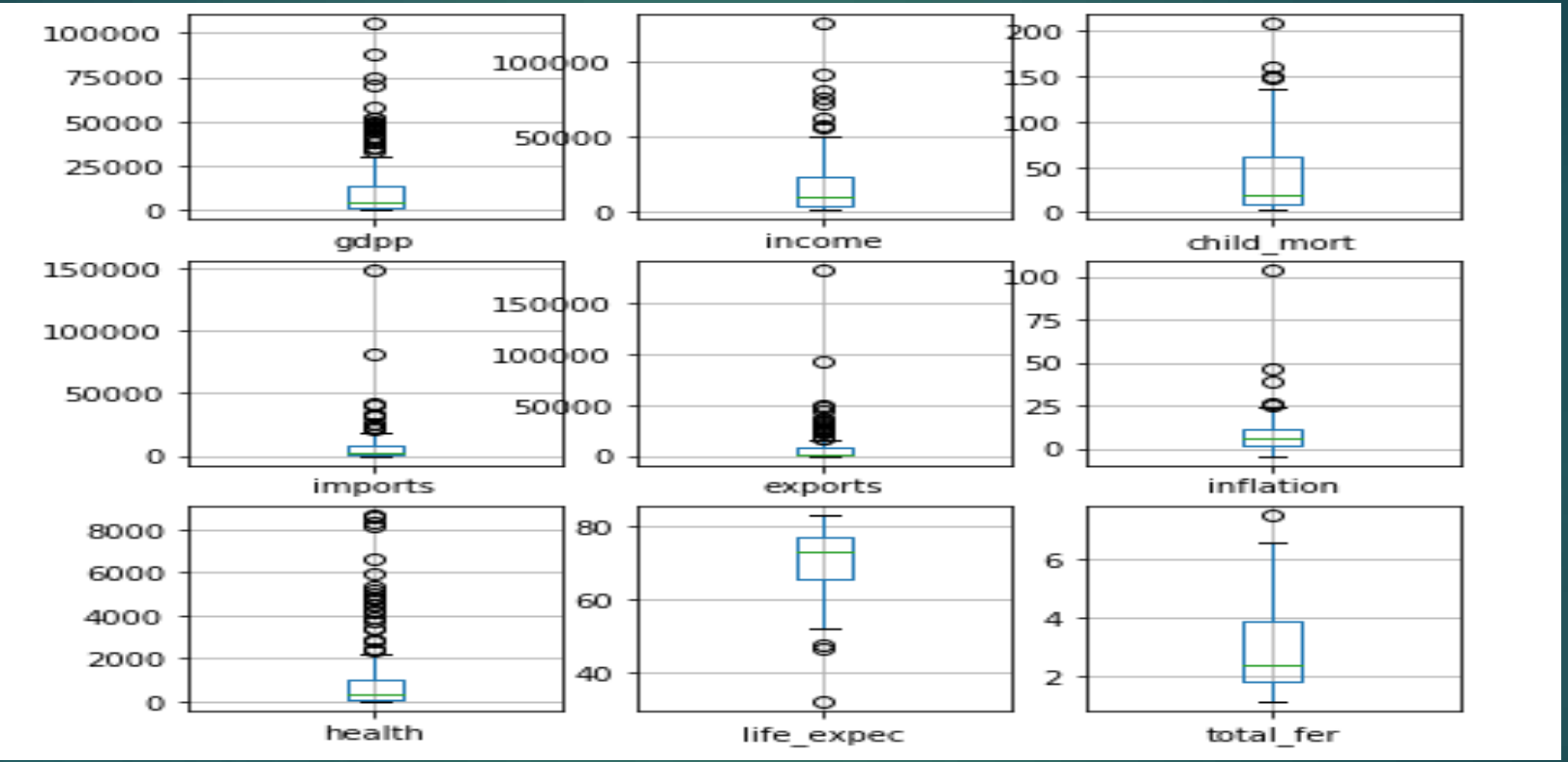
Clustering Assignment

BY RAMYA SREE NARWA

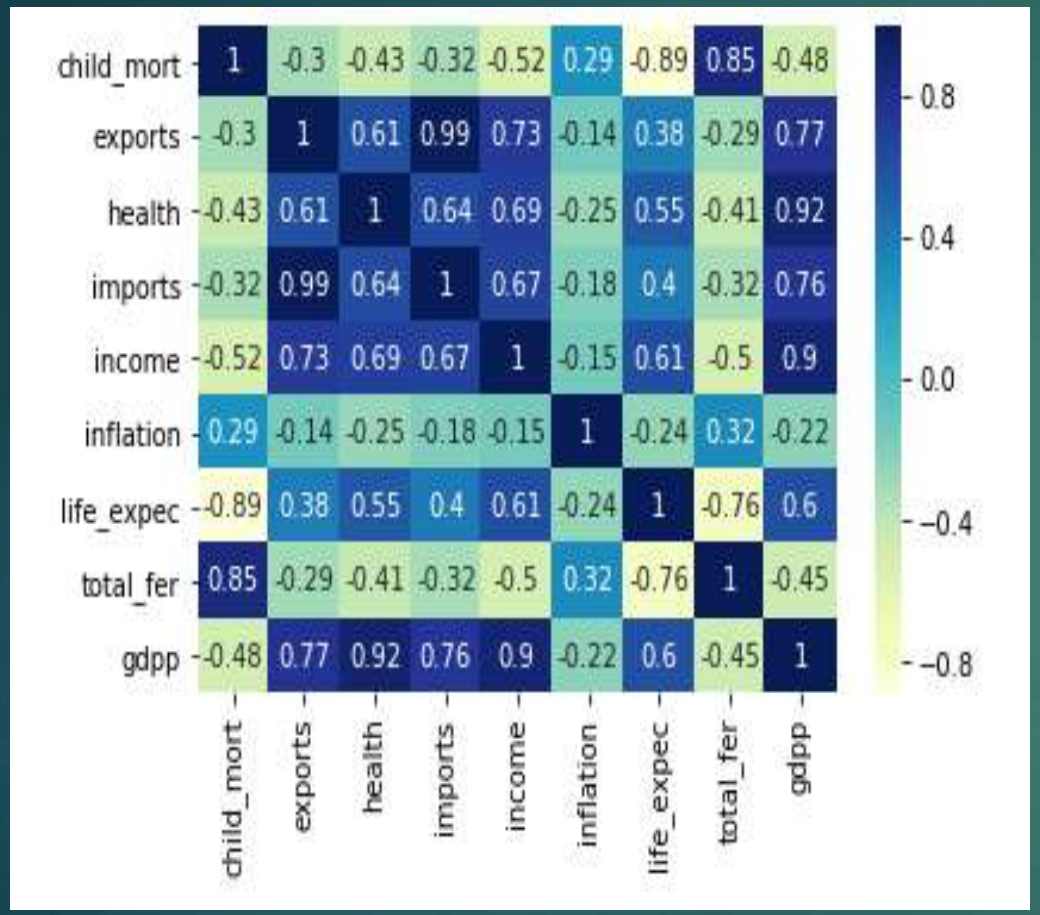
Objective

- 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.
- As an analyst we need to segment the countries and provide with the countries that are in direst need of aid based on the analysis

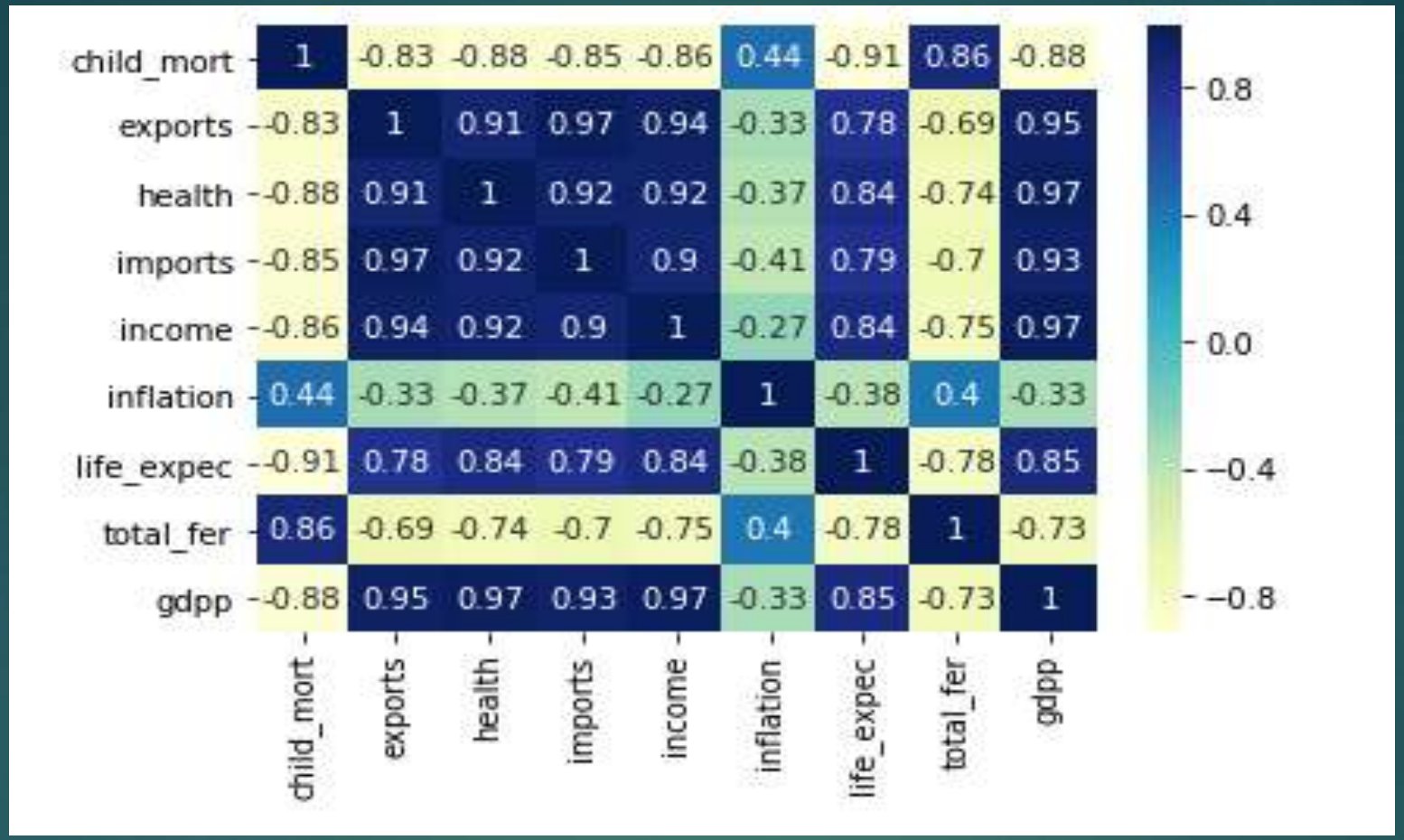
Outlier Analysis



GDPP, Income, Imports, Exports, Inflation and health has outliers but didn't remove/cap them so as to not modify or delete the data.

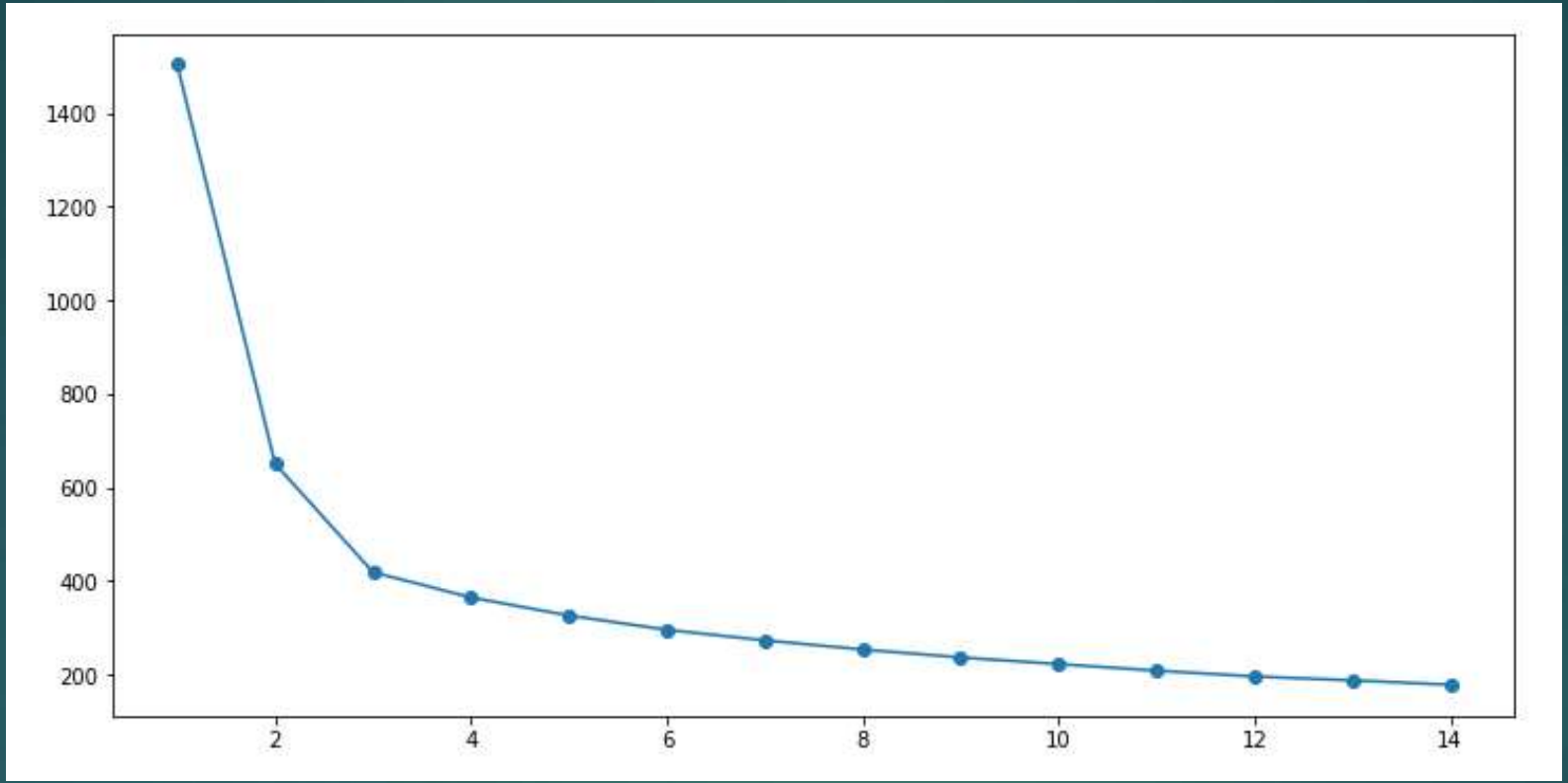


- After cleaning the data, when the scatter plots and heat maps are checked for the data the data seems to be skewed and the correlations are high for most of the variables.
- Therefore applied power transformation for the data to scale and transform the data to remove skew.

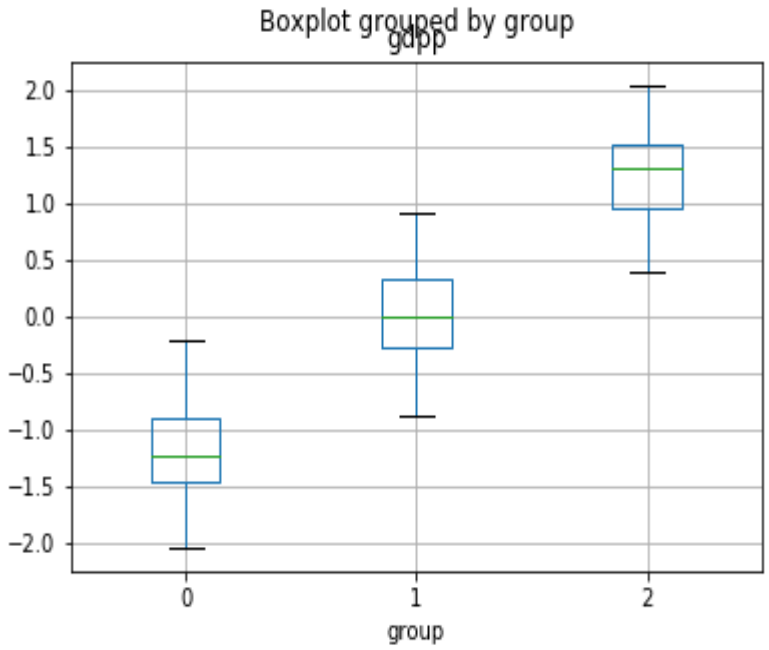


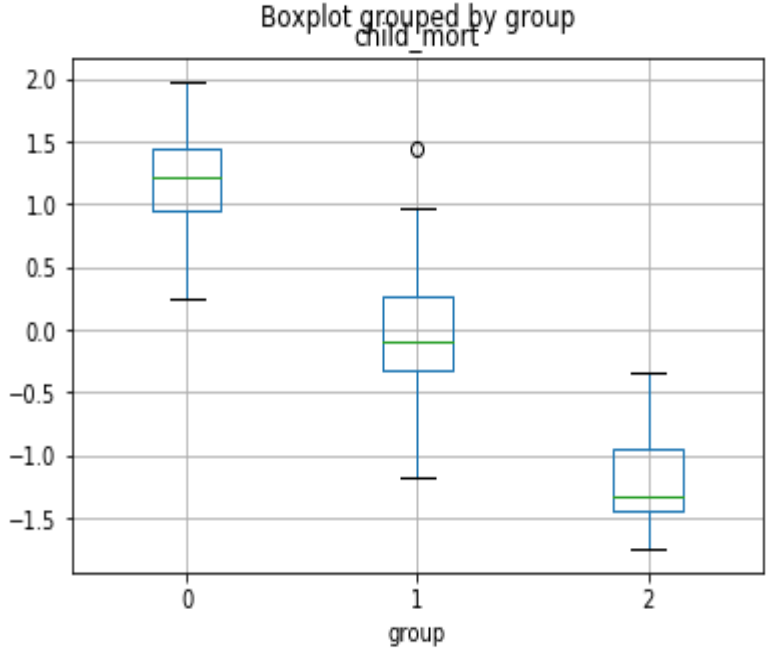
After the data is transformed using power transformer, the data is scaled, no skew and the correlations are significant

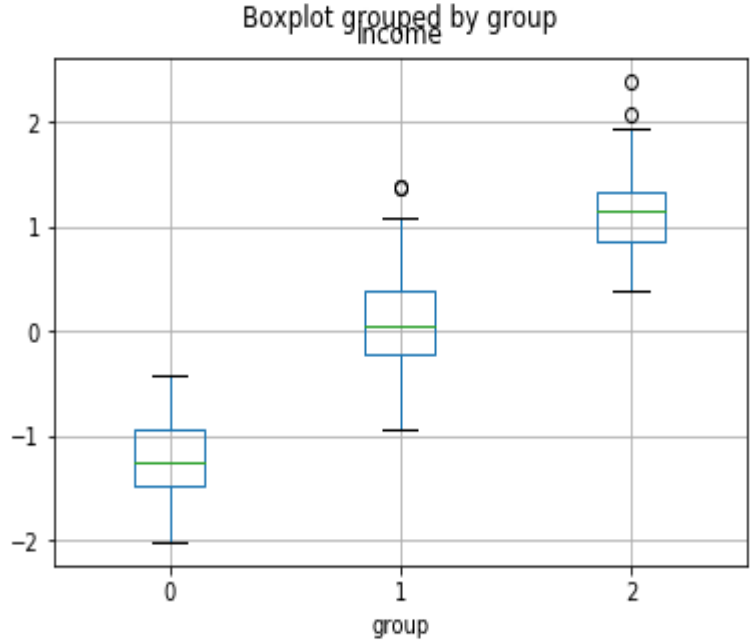
K-Means Clustering



- With the elbow curve above using the sum of squared distances we can see that elbow is in range of 3-4. Therefore considering k=3 for further analysis using K-Means clustering.
- Using Silhouette Analysis also the optimal value of K seems to be 2-3







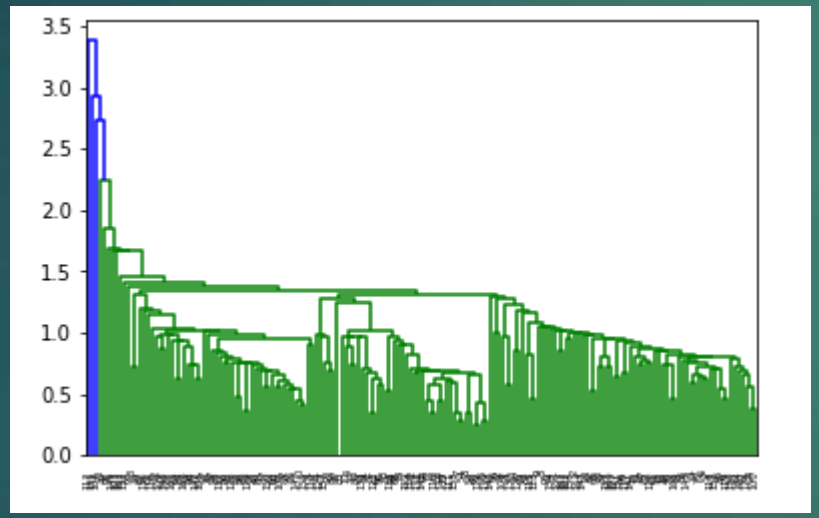
Using the boxplots it can be concluded that the countries with group 0 has low GDPP, high child mortality and low income which makes them the area of concern for aid.

Top 10 countries that would need aid based on the analysis using K-Means clustering are:

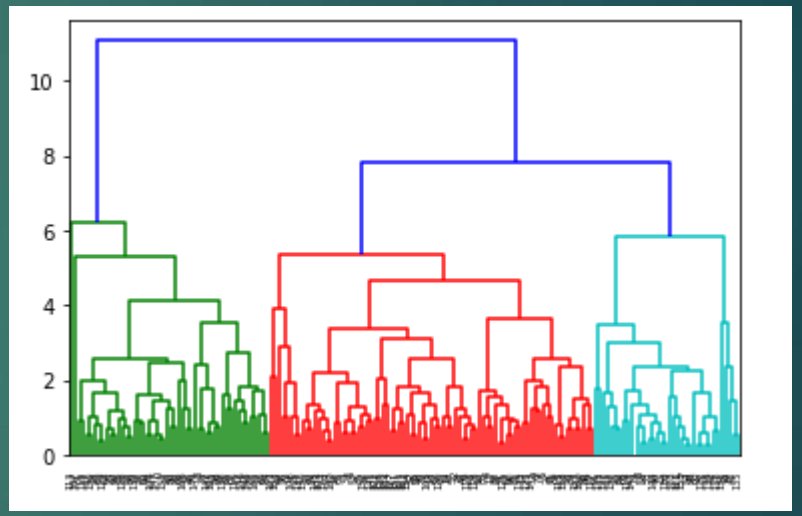
country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	group
Burundi	1.298780	-2.063795	-1.541130	-1.770468	-1.867384	0.738196	-1.422077	1.645014	-2.048077	0
Liberia	1.257999	-1.606617	-1.294897	-1.138832	-1.923514	-0.013415	-1.181773	1.323867	-1.814890	0
Congo, Dem. Rep.	1.484194	-1.259150	-1.550729	-1.460581	-2.011864	1.405163	-1.436428	1.703294	-1.800675	0
Niger	1.534638	-1.514535	-1.813260	-1.444379	-1.826407	-0.484814	-1.340700	1.873128	-1.773118	0
Sierra Leone	1.759989	-1.576152	-1.091915	-1.556923	-1.558710	1.145065	-1.604688	1.378219	-1.681334	0
Madagascar	0.942840	-1.386846	-1.910343	-1.424122	-1.470105	0.390124	-1.181773	1.183867	-1.658188	0
Mozambique	1.364657	-1.276898	-1.680340	-1.378683	-1.747977	0.260760	-1.635948	1.478218	-1.648508	0
Central African Republic	1.699129	-1.679792	-1.821115	-1.635352	-1.769743	-0.594558	-1.998305	1.381149	-1.606595	0
Malawi	1.269579	-1.380850	-1.459061	-1.478143	-1.672009	0.719890	-1.719425	1.409950	-1.587312	0
Eritrea	0.838250	-2.018839	-2.042388	-1.661445	-1.455493	0.673414	-1.105457	1.187437	-1.554497	0

Hierarchical Clustering

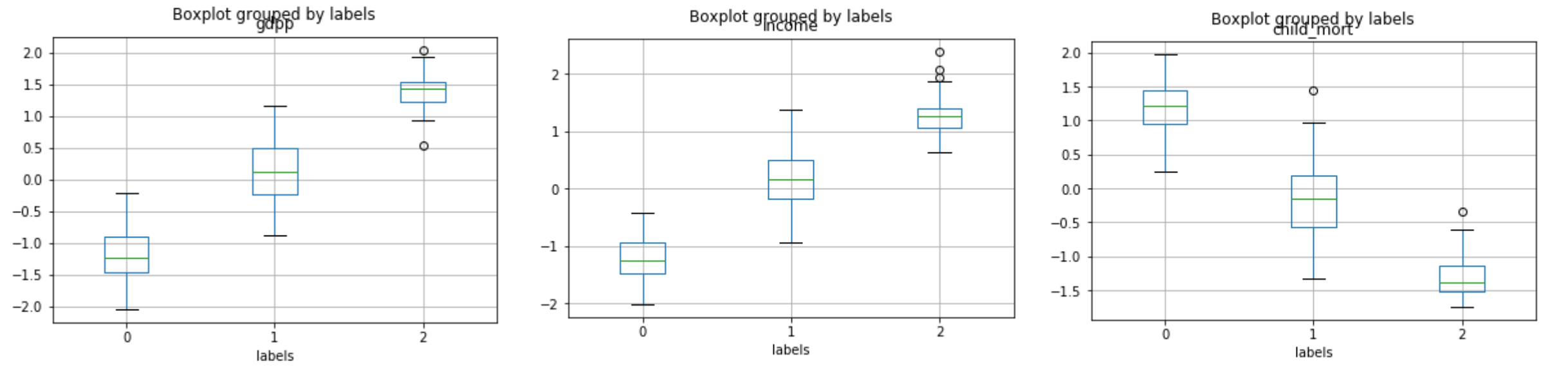
Single linkage



Complete Linkage



Using complete linkage dendogram we can see that three distinct clusters can be formed. Therefore proceeding the hierarchical clustering using k=3.



Based on the box plots the countries of group 0 has low GDPP, high child mortality and low income which makes them the area of concern for aid.

Top 10 countries that would need aid based on the analysis using hierarchical clustering are:

country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	group	labels
Burundi	1.298780	-2.063795	-1.541130	-1.770468	-1.867384	0.738196	-1.422077	1.645014	-2.048077	0	0
Liberia	1.257999	-1.606617	-1.294897	-1.138832	-1.923514	-0.013415	-1.181773	1.323867	-1.814890	0	0
Congo, Dem. Rep.	1.484194	-1.259150	-1.550729	-1.460581	-2.011864	1.405163	-1.436428	1.703294	-1.800675	0	0
Niger	1.534638	-1.514535	-1.813260	-1.444379	-1.826407	-0.484814	-1.340700	1.873128	-1.773118	0	0
Sierra Leone	1.759989	-1.576152	-1.091915	-1.556923	-1.558710	1.145065	-1.604688	1.378219	-1.681334	0	0
Madagascar	0.942840	-1.386846	-1.910343	-1.424122	-1.470105	0.390124	-1.181773	1.183867	-1.658188	0	0
Mozambique	1.364657	-1.276898	-1.680340	-1.378683	-1.747977	0.260760	-1.635948	1.478218	-1.648508	0	0
Central African Republic	1.699129	-1.679792	-1.821115	-1.635352	-1.769743	-0.594558	-1.998305	1.381149	-1.606595	0	0
Malawi	1.269579	-1.380850	-1.459061	-1.478143	-1.672009	0.719890	-1.719425	1.409950	-1.587312	0	0
Eritrea	0.838250	-2.018839	-2.042388	-1.661445	-1.455493	0.673414	-1.105457	1.187437	-1.554497	0	0

Conclusion

Using both the clustering methods, the countries in need of aid are same. Therefore the countries which are in dire need of the aid first by considering socio-economic factors are the following:

- Burundi
- Liberia
- Congo, Dem, Rep
- Niger
- Sierra Leone
- Madagascar
- Mozambique
- Central African Republic
- Malawi
- Eritrea