CLUSTERING ASSIGNMENT

ARPIT VIJAY
arpitvj1993@gmail.com

OBJECTIVE

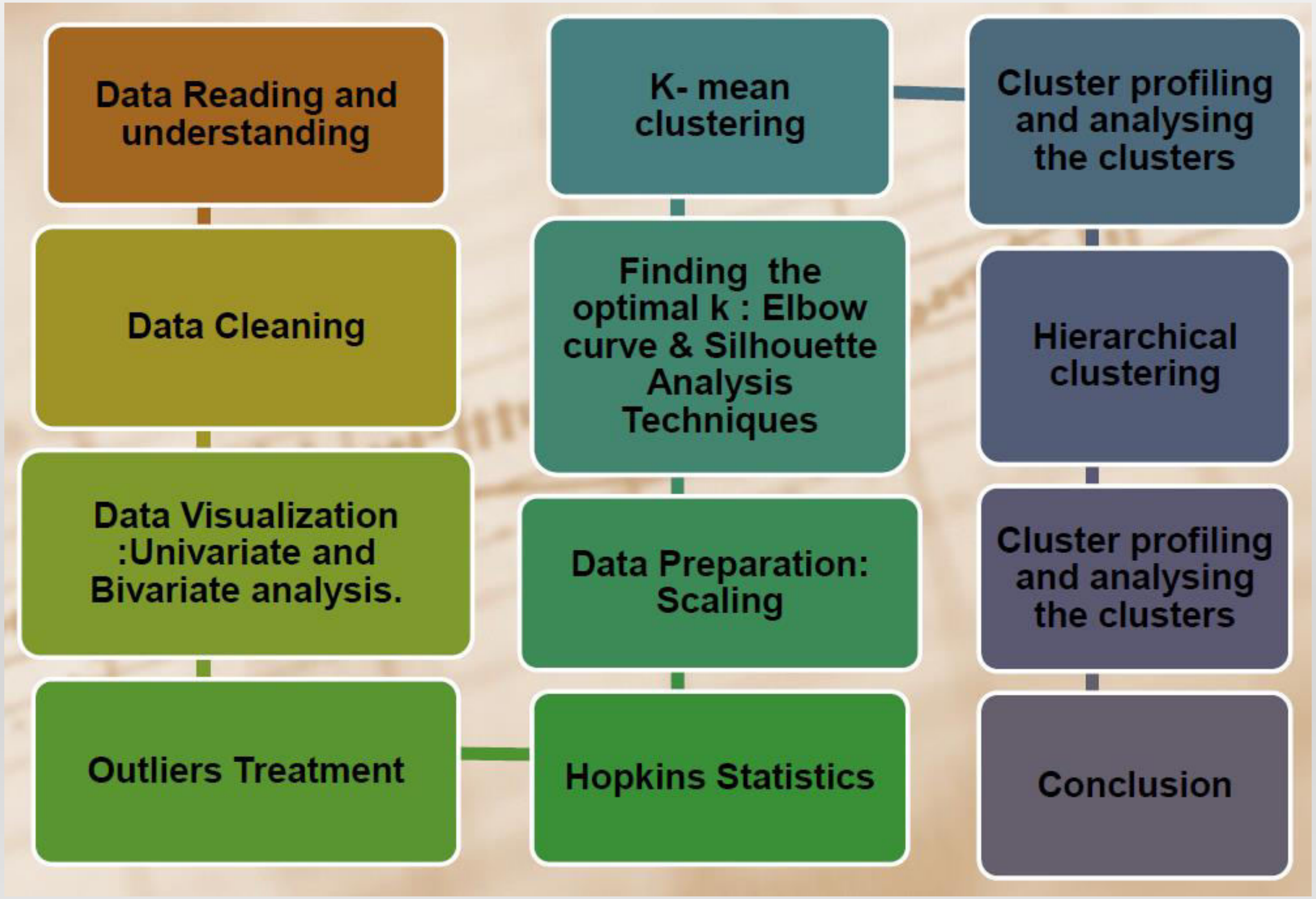
- To categorize the countries using some socio-economic and health factors that determine the overall development of the country.
- Then need to suggest the countries which the CEO needs to focus on the most
- Help International may choose to utilize this Knowledge while making this decision are mostly related to choosing the countries that are in the direct need of aid.

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 programes, 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.

APPROACH APPLIED



Data Reading and understanding

Data Cleaning

Data Visualization :Univariate and Bivariate analysis.

Outliers Treatment

Hopkins Statistics

Data Preparation: Scaling

Finding the optimal k : Elbow curve & Silhouette Analysis Techniques

K- mean clustering

Cluster profiling and analysing the clusters

Hierarchical clustering

Cluster profiling and analysing the clusters

Conclusion

DATA READING

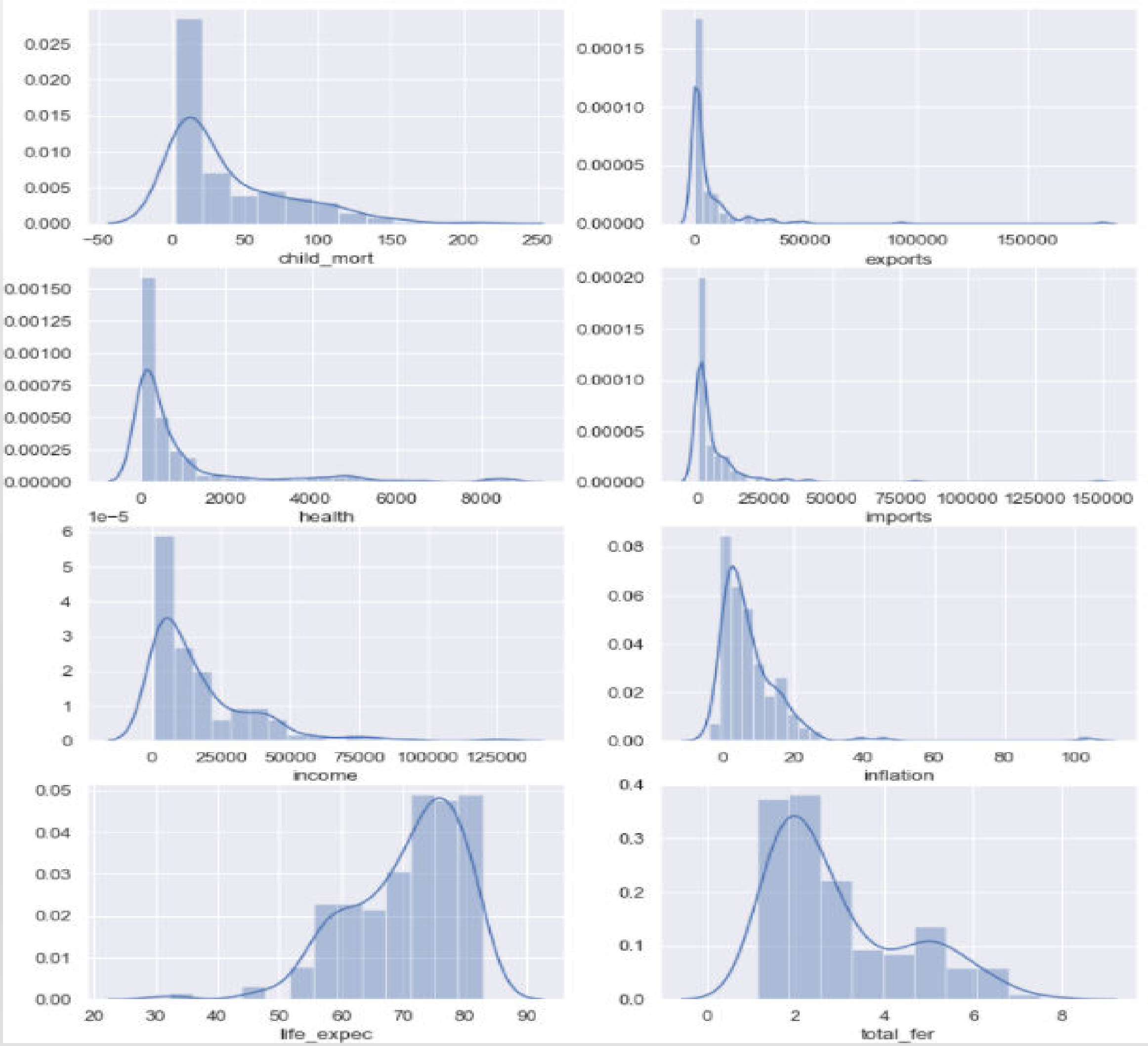
- Imported the data and viewed the dataset.
- Understood the type of data ,read the shape of the dataset.
- Checked the summary of the data.

PREPARING THE DATA & DATA CLEANING

- The dataset was clean enough.
- Checked the null values if any but found to be zero.Checked if any duplicate rows presents and concluded the dataset was free from the duplicate rows too.
- Checked few columns like export ,import and health was in the percentage form changed to the actual value to give us the proper analysis of the data later.

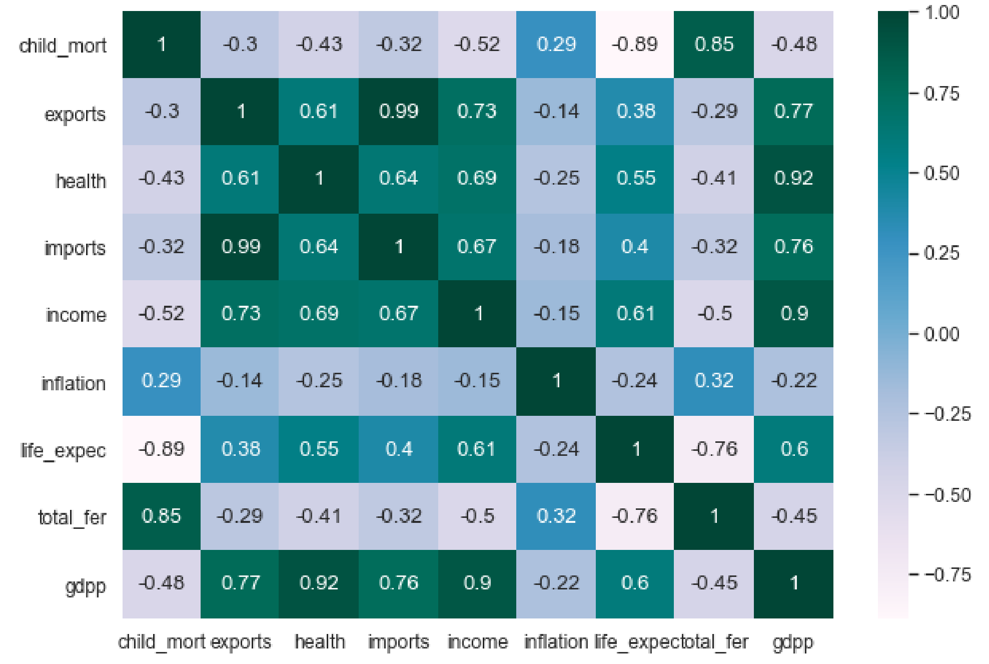
DATA VISUALIZATION(UNIVARIATE ANALYSIS)

- All the variables are not Normally distributed
- Most of the variables are Positively skewd like child_mort, exports, health, income, total_fer, gdpp.
- The variables are Negatively skewd is life_expec



DATA VISUALIZATION(BIVARIATE ANALYSIS)

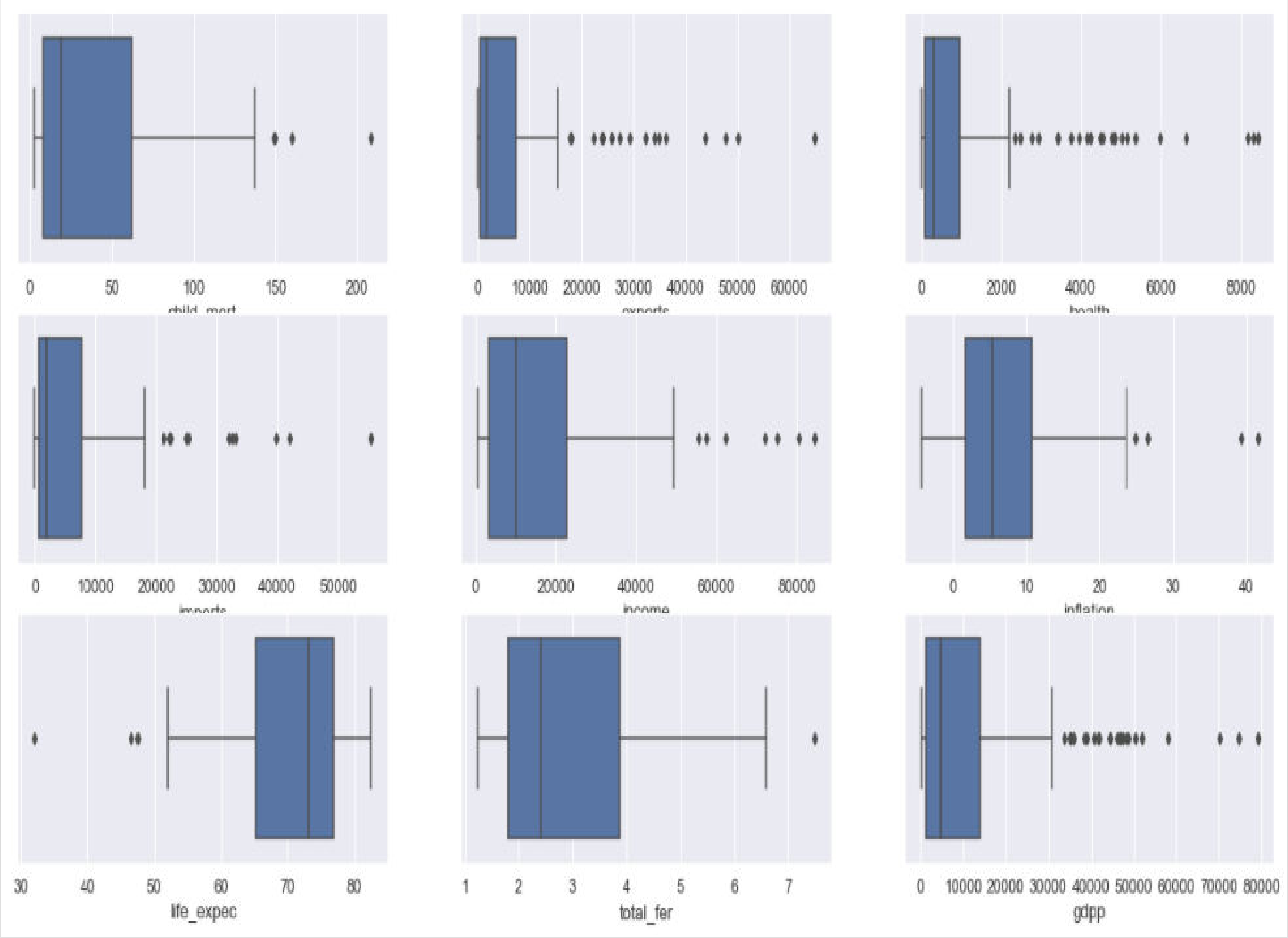
- This is the Bivariate Analysis of all the Numeric Variables which shows the correlation of all the parameters i.e which parameters are highly correlated
- Export vs Income- 0.73
- Health vs Income – 0.69
- The child_mort and the life_expec shows high negative correlation with each other.
- Countries with good infrastructure ,health facility and high income, GDPP and less Child Mortality doesn't require aid.



	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

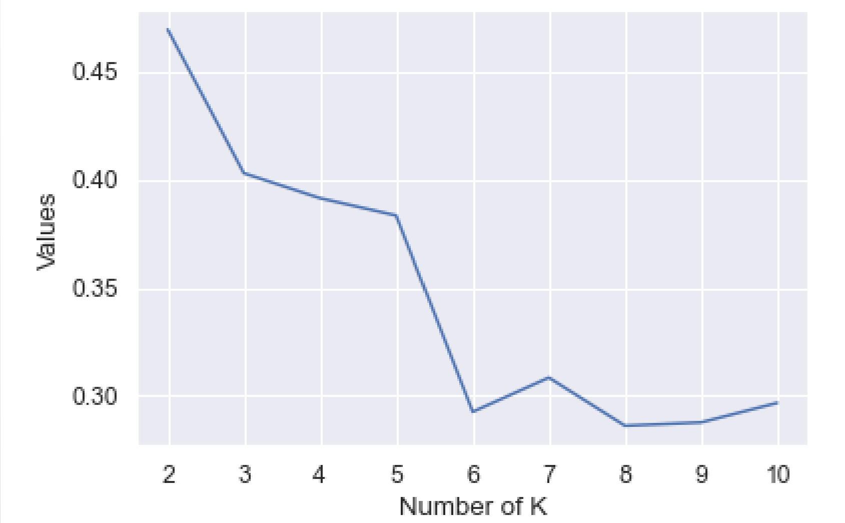
OUTLIER TREATMENT

- There are a lot of outliers in our features, however due to limited data we cannot drop the rows, thus now we need to perform capping to all our features.

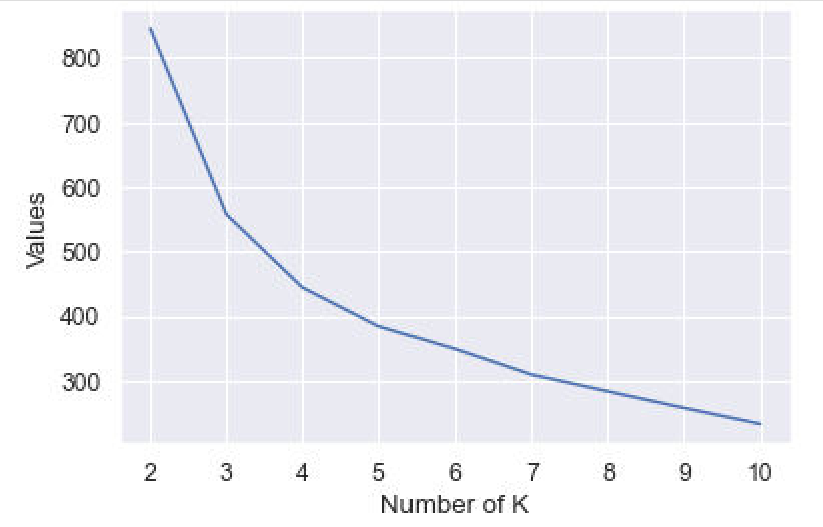


K-MEAN CLUSTERING

- Silhouette Analysis



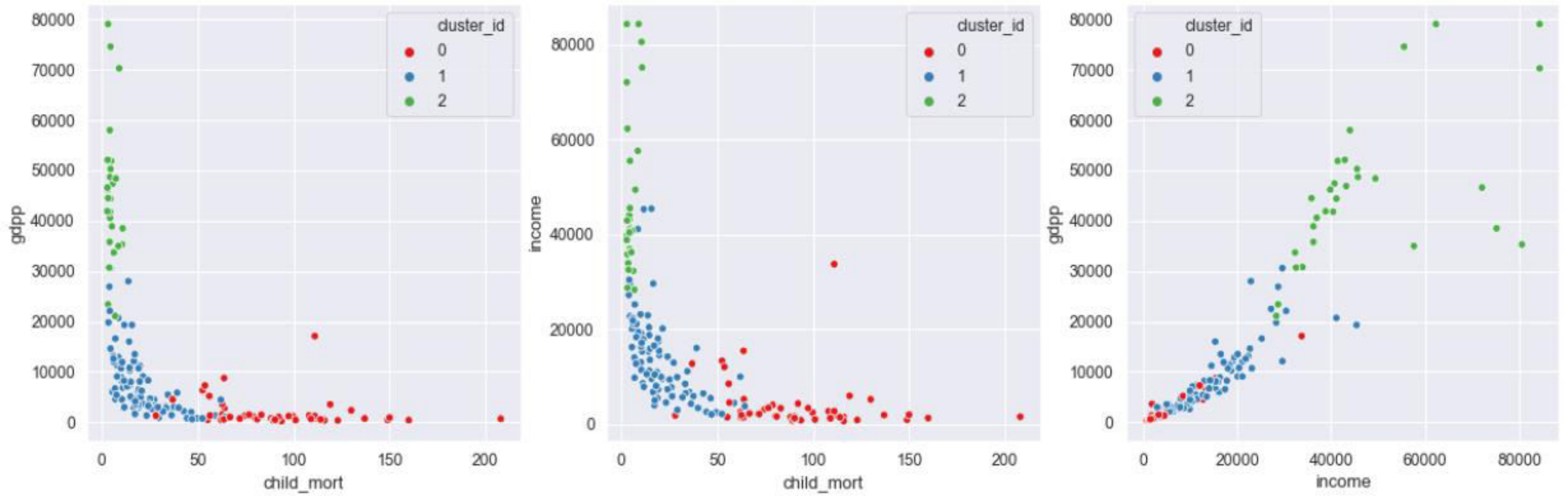
- Elbow Curve(Sum of Squared distances)



Scores for different k values can be seen. It says k=2 is the best choice.

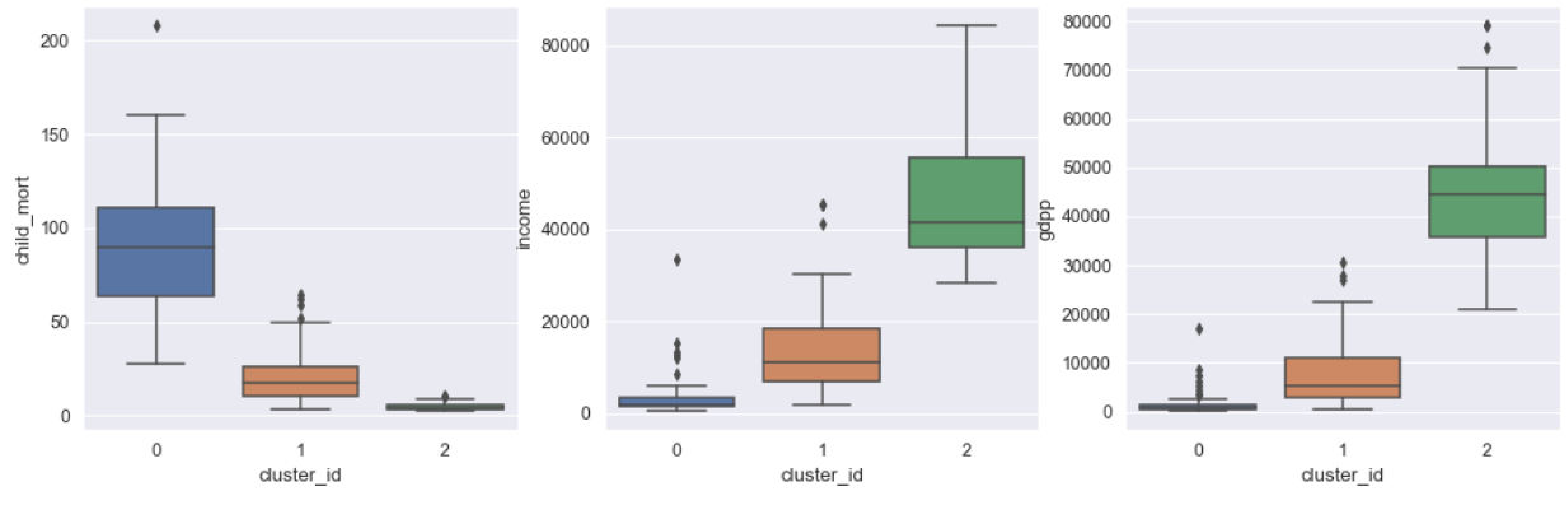
Choosing k=3,because we are doing clustering and we do not wish to divide the graph into 2

K-MEAN CLUSTERING



- The countries in red shows some behaviors like high child mortality , Low gdpp and low income .
- So the countries belonging to such clusters can be considered for the need of the aid.

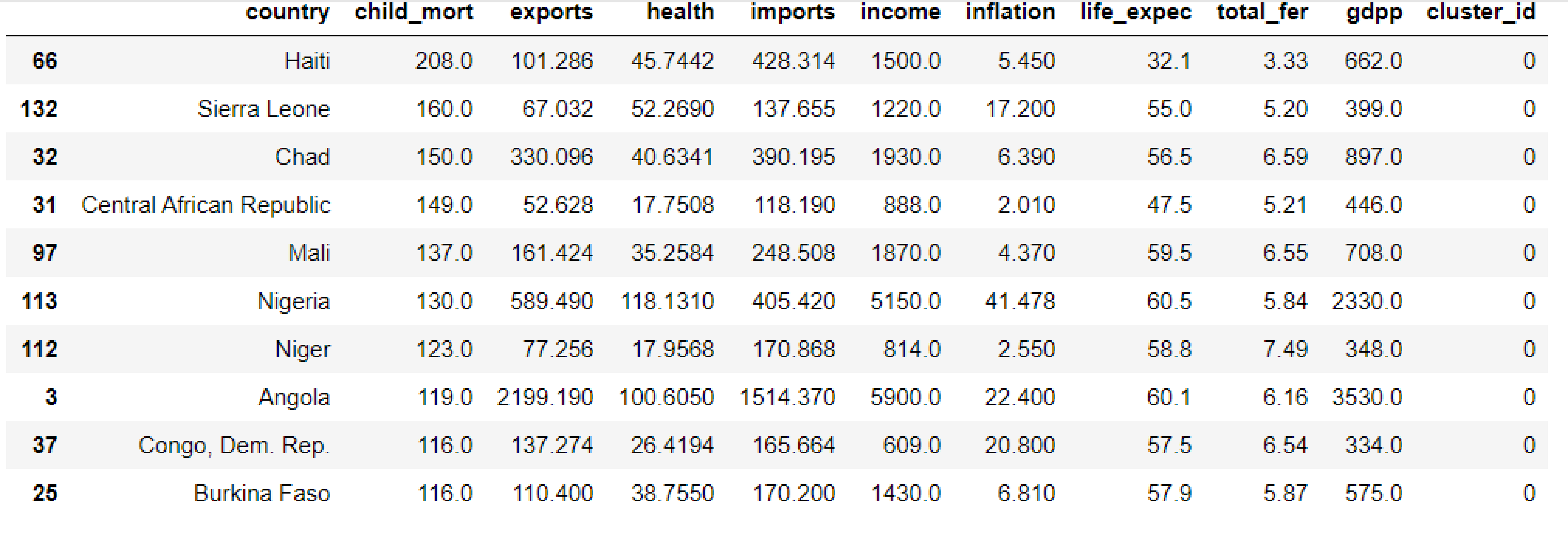
K-MEAN CLUSTERING



- Cluster 0 is the most affected cluster as depicted from the boxplots above. As it is the cluster with the highest child_mort, and low income and low gdpp

K-MEAN CLUSTERING

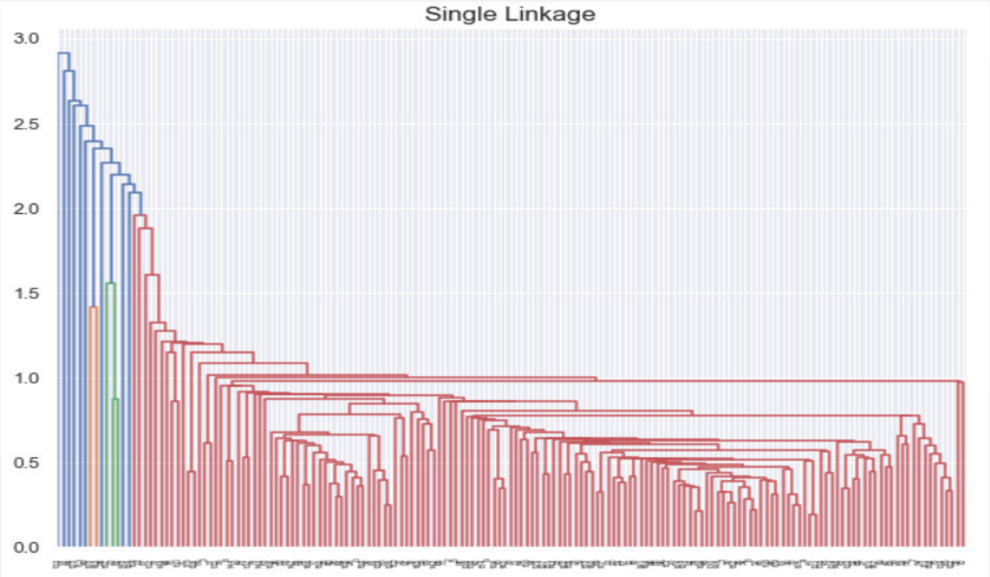
- Countries those are in the urgent need of the aid according to k-mean clustering

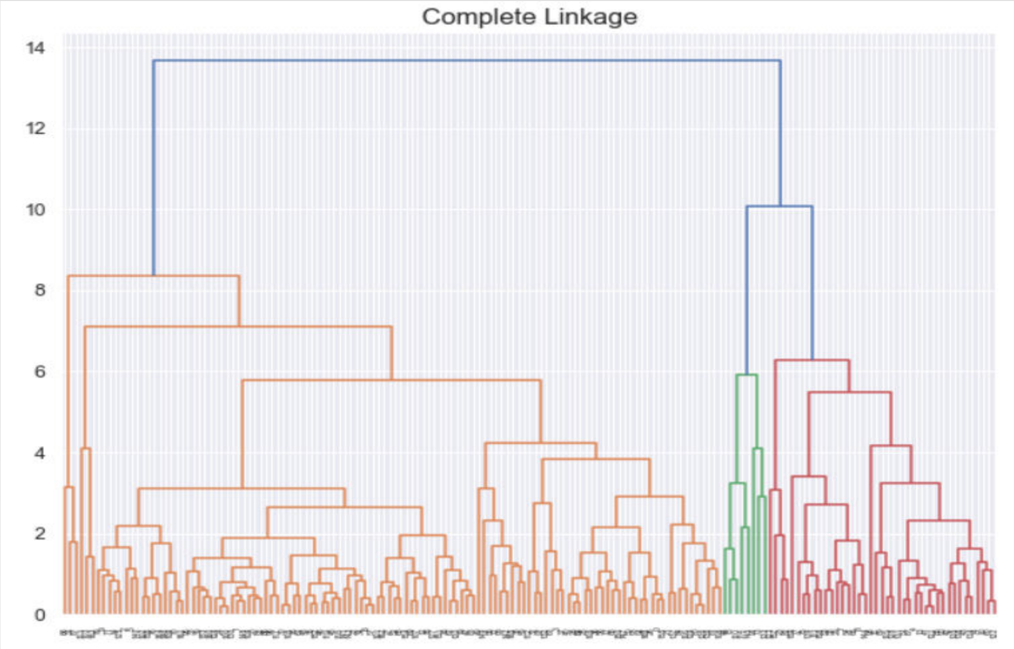


	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_id
66	Haiti	208.0	101.286	45.7442	428.314	1500.0	5.450	32.1	3.33	662.0	0
132	Sierra Leone	160.0	67.032	52.2690	137.655	1220.0	17.200	55.0	5.20	399.0	0
32	Chad	150.0	330.096	40.6341	390.195	1930.0	6.390	56.5	6.59	897.0	0
31	Central African Republic	149.0	52.628	17.7508	118.190	888.0	2.010	47.5	5.21	446.0	0
97	Mali	137.0	161.424	35.2584	248.508	1870.0	4.370	59.5	6.55	708.0	0
113	Nigeria	130.0	589.490	118.1310	405.420	5150.0	41.478	60.5	5.84	2330.0	0
112	Niger	123.0	77.256	17.9568	170.868	814.0	2.550	58.8	7.49	348.0	0
3	Angola	119.0	2199.190	100.6050	1514.370	5900.0	22.400	60.1	6.16	3530.0	0
37	Congo, Dem. Rep.	116.0	137.274	26.4194	165.664	609.0	20.800	57.5	6.54	334.0	0
25	Burkina Faso	116.0	110.400	38.7550	170.200	1430.0	6.810	57.9	5.87	575.0	0

HIERARCHICAL CLUSTERING

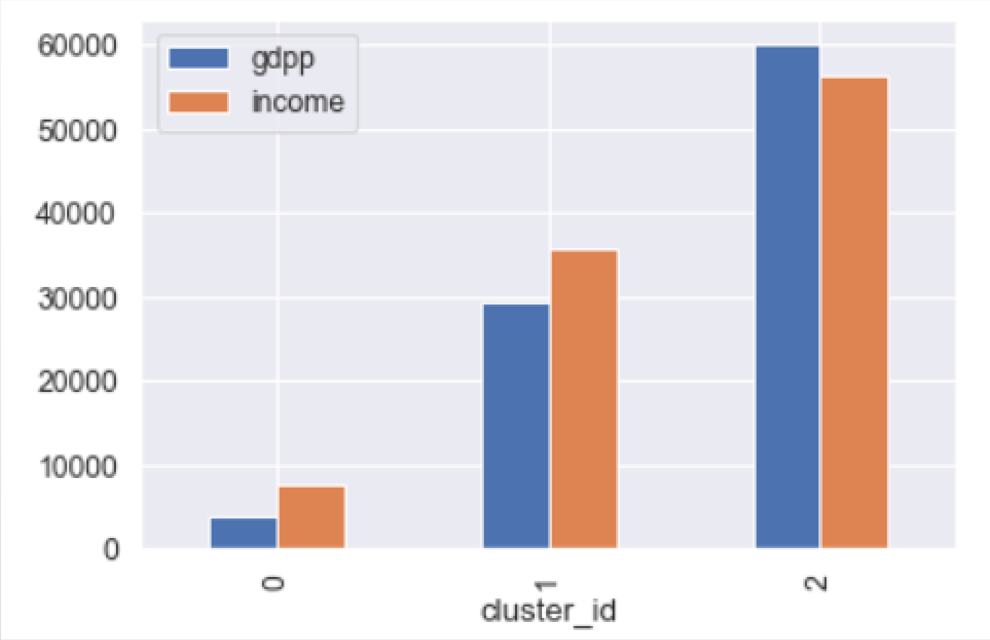
- Single linkage Hierarchical clustering was not very clear so we will look into the complete linkage
- Complete linkage Hierarchical clustering gives us a option to refine the countries into 3 clusters and analyze

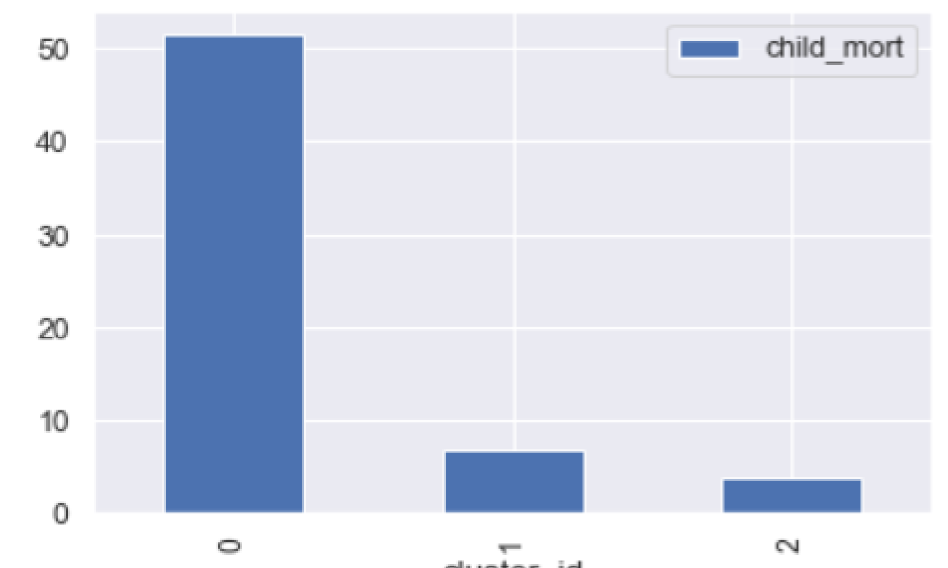




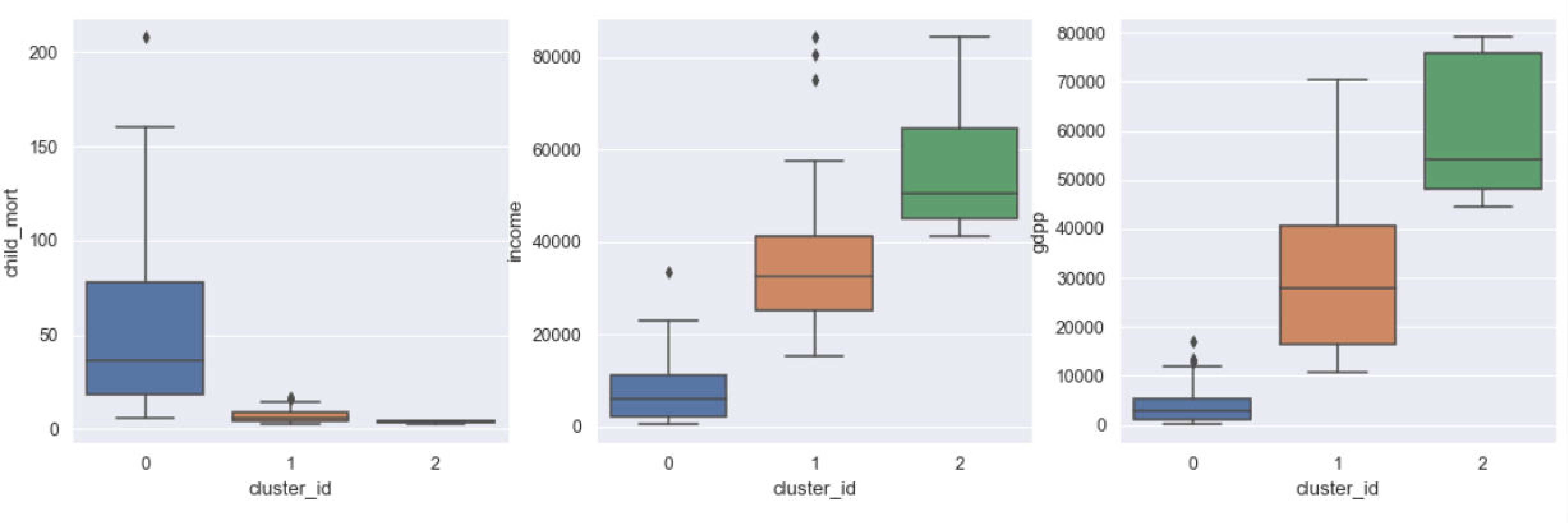
HIERARCHICAL CLUSTERING

- Cluster 0 is the target cluster as it has high child_mort and the low gdpp and income





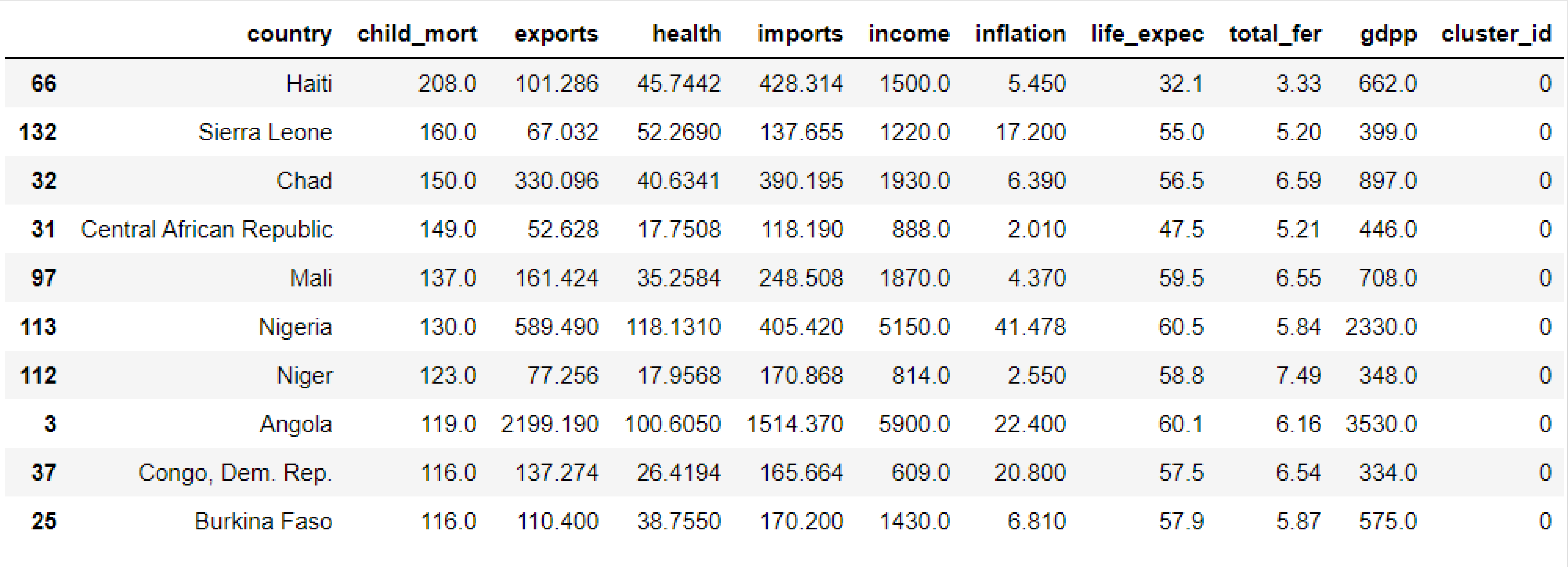
HIERARCHICAL CLUSTERING



- The countries in the cluster 0 shows the behavior of high child mortality, low income and low gdpp .So we can go ahead with the countries in this cluster .

HIERARCHICAL CLUSTERING

- Countries those are in the urgent need of the aid according to Hierarchical clustering



	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_id
66	Haiti	208.0	101.286	45.7442	428.314	1500.0	5.450	32.1	3.33	662.0	0
132	Sierra Leone	160.0	67.032	52.2690	137.655	1220.0	17.200	55.0	5.20	399.0	0
32	Chad	150.0	330.096	40.6341	390.195	1930.0	6.390	56.5	6.59	897.0	0
31	Central African Republic	149.0	52.628	17.7508	118.190	888.0	2.010	47.5	5.21	446.0	0
97	Mali	137.0	161.424	35.2584	248.508	1870.0	4.370	59.5	6.55	708.0	0
113	Nigeria	130.0	589.490	118.1310	405.420	5150.0	41.478	60.5	5.84	2330.0	0
112	Niger	123.0	77.256	17.9568	170.868	814.0	2.550	58.8	7.49	348.0	0
3	Angola	119.0	2199.190	100.6050	1514.370	5900.0	22.400	60.1	6.16	3530.0	0
37	Congo, Dem. Rep.	116.0	137.274	26.4194	165.664	609.0	20.800	57.5	6.54	334.0	0
25	Burkina Faso	116.0	110.400	38.7550	170.200	1430.0	6.810	57.9	5.87	575.0	0

CONCLUSION & RECOMMENDATION

- For a countries major development is basically depends on some major factors i.e. Income per capita, GDP, Health infrastructures, Child Mortality etc.

- If a country's GDP is low and also per capita Income is low then surely it can say that this particular country is Socio-Economically very week, where as if these two factor for a country is high then this country hold a strong position socio-economically. These are the major two factors that can play a significant role for any country's development

- Apart from this if the Child Mortality rate is very high for a country, then it can surely say that in terms of Health infrastructures this country belongs to a non healthy condition.

- So we can recommend the NGO that they can take into consideration of 1st 7 countries filtering by Child Mortality rate i.e. **Haiti, Sierra Leone, Chad, Central African Republic, Mali, Nigeria, Niger**. Because from our basic Understanding we know that if a country's child mortality is high then obviously their health infrastructure is very much poor. So for a NGO this could be their prime responsibility to keep ahead their hand for those countries' health development and plan their amount of investment accordingly.