CLUSTERING OF COUNTRIES

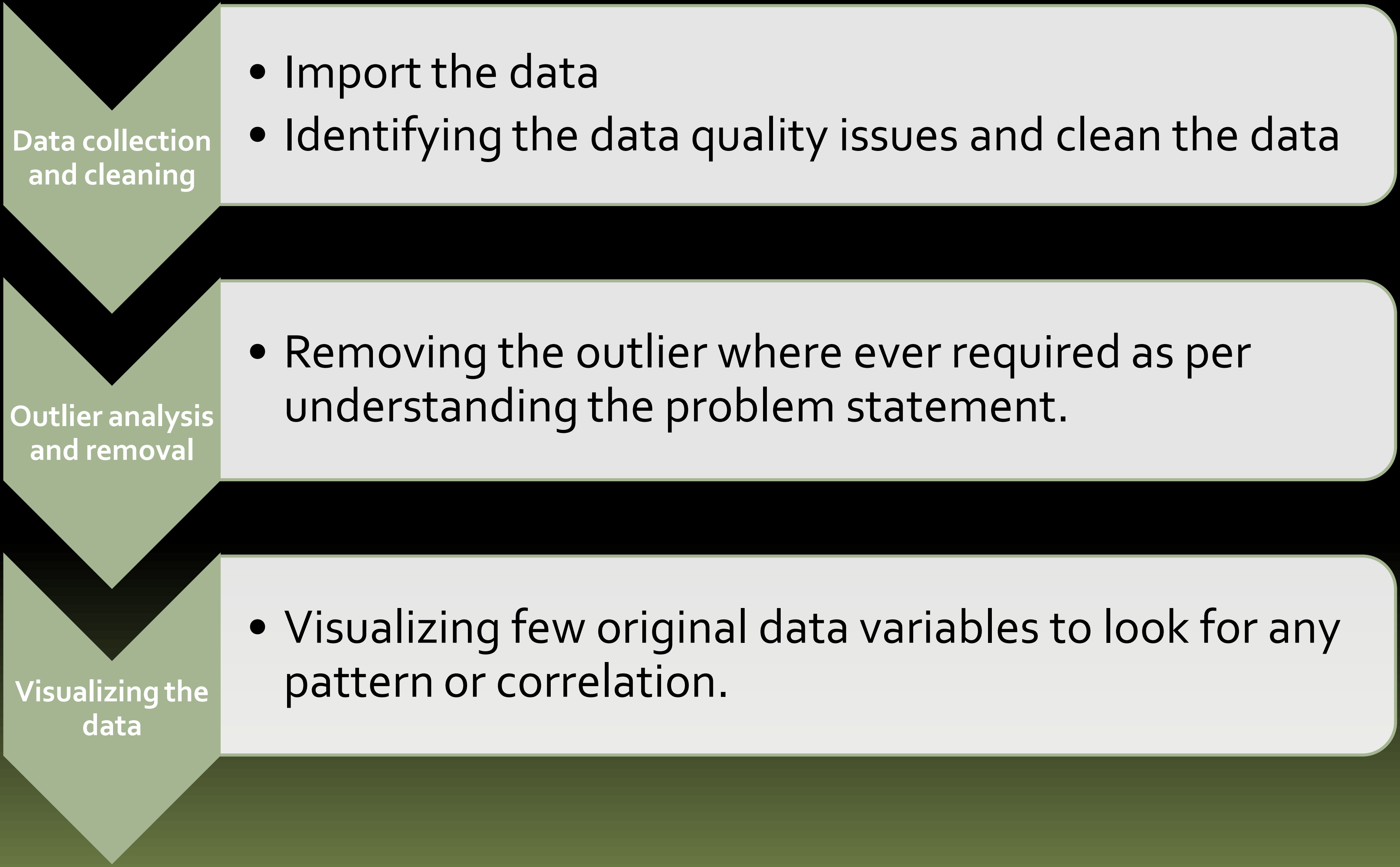
An assignment to cluster a group of countries based on socio-economic factors.

By:Monika Kumari

Aim

- This case study aims is to categorize the countries using some socio-economic and health factors that determine overall development of the country.
- Here we have performed Clustering technique to select the countries which are in direst need of aid by considering socio–economic factor in to consideration.
- This analysis make it easy and help an international humanitarian NGO to provide the top 5 backward countries to provide the basic amenities and relief during the time of disasters and natural calamities. HELP International is an international humanitarian NGO that runs a lot of operational projects from time to time along with advocacy drives to raise awareness as well as for funding purposes.

Analysis methodology



Data collection and cleaning

- Import the data
- Identifying the data quality issues and clean the data

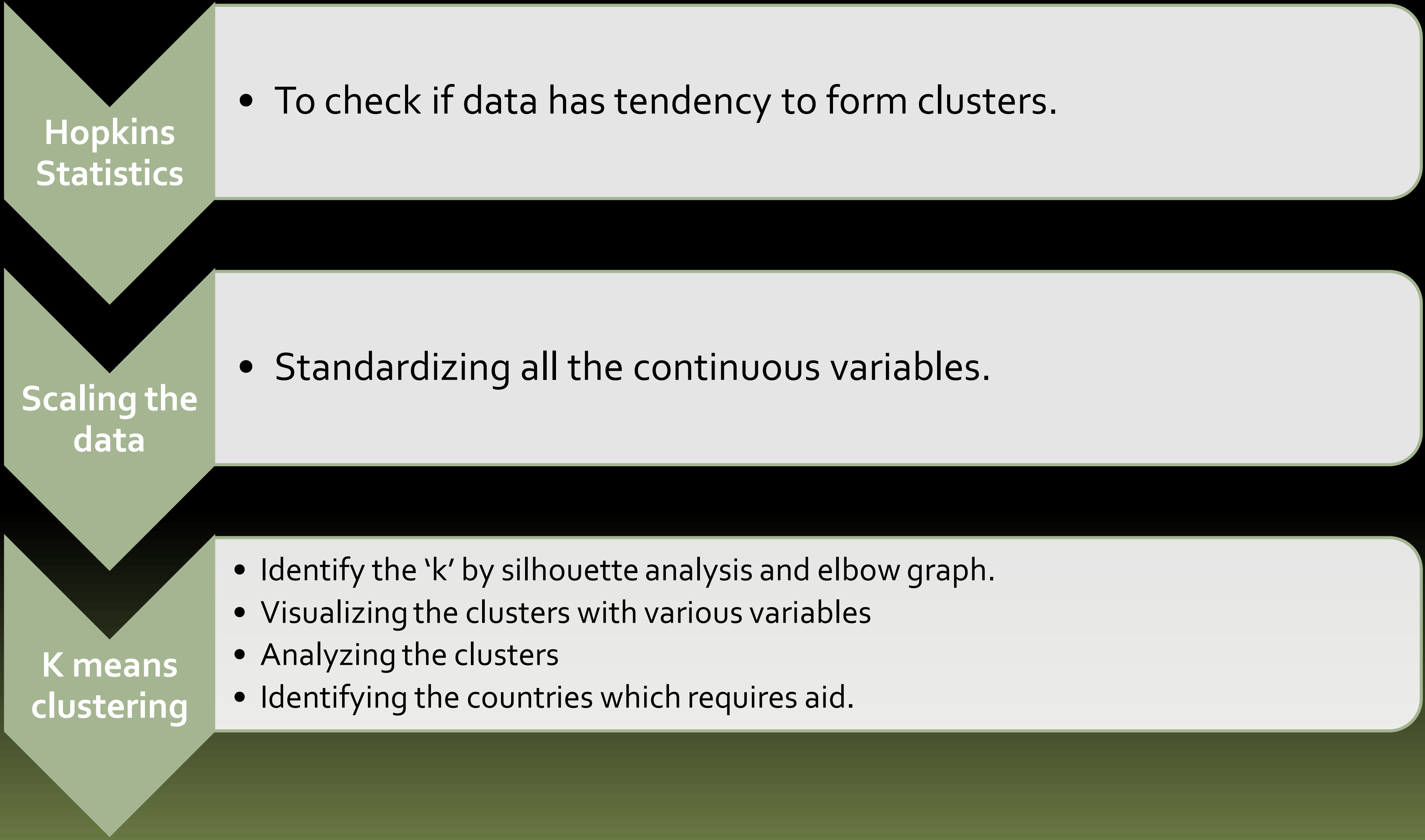
Outlier analysis and removal

- Removing the outlier where ever required as per understanding the problem statement.

Visualizing the data

- Visualizing few original data variables to look for any pattern or correlation.

Analysis methodology Contd………



Hopkins Statistics

- To check if data has tendency to form clusters.

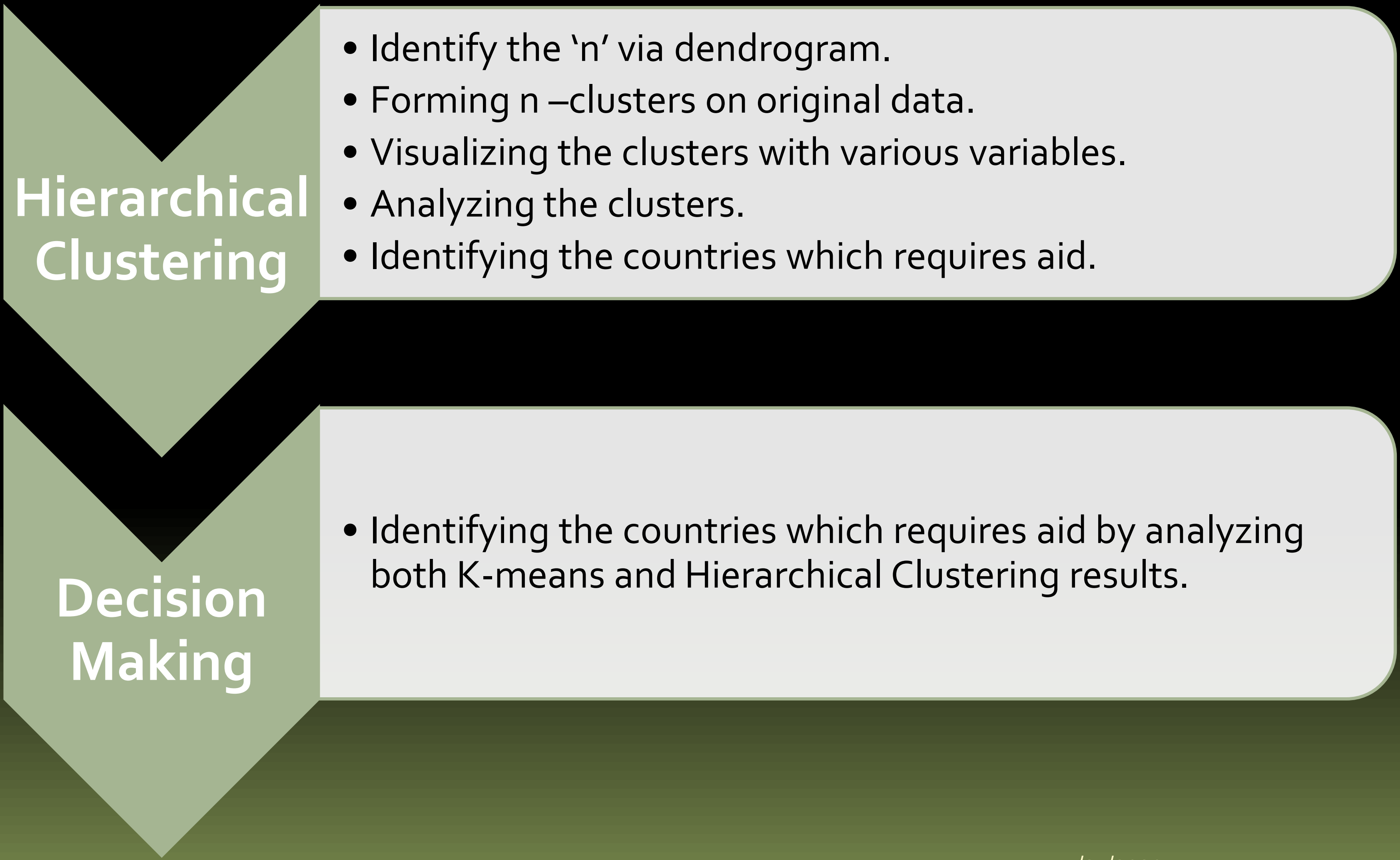
Scaling the data

- Standardizing all the continuous variables.

K means clustering

- Identify the 'k' by silhouette analysis and elbow graph.
- Visualizing the clusters with various variables
- Analyzing the clusters
- Identifying the countries which requires aid.

Analysis methodology Contd….



Hierarchical Clustering

- Identify the ‘n’ via dendrogram.
- Forming n –clusters on original data.
- Visualizing the clusters with various variables.
- Analyzing the clusters.
- Identifying the countries which requires aid.

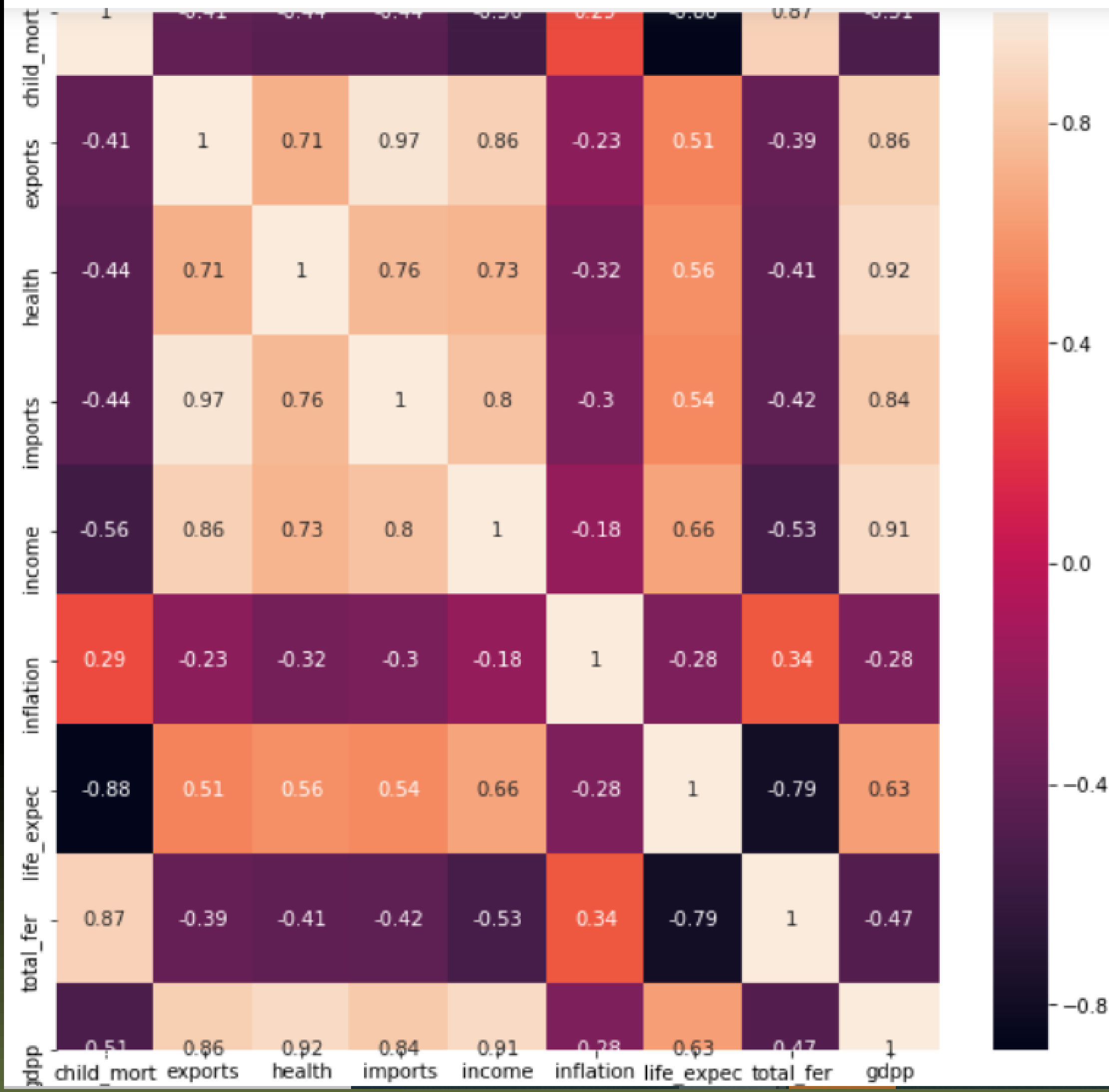
Decision Making

- Identifying the countries which requires aid by analyzing both K-means and Hierarchical Clustering results.

Correlation in the data

- After data cleaning , we removed outlier by using the capping technique because the country with high gdpp would not require any aid as there are already doing good.
- We did standardized scaling to standardize all parameters on cleaned, outlier removed data.

- Looking at the heatmap, we see that few variables like (total fertility, child mortality) , (income , gdpp) and (imports and exports) have high correlation.



Hopkins Statistics

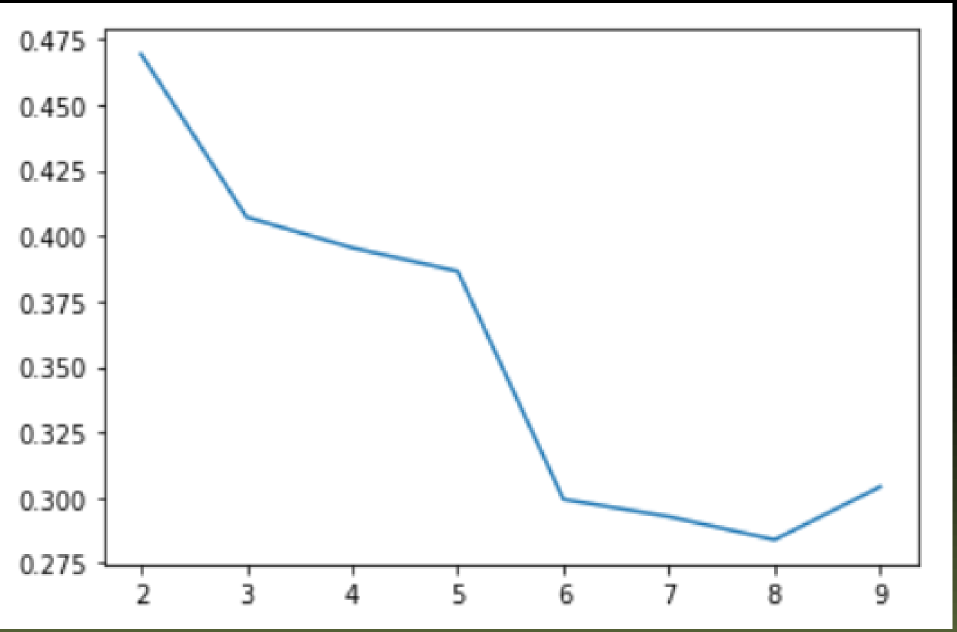
- We perform Hopkins Statistics Test to ensure that the given data has some meaningful clusters is not random.
- Hopkins test examines whether data points differ significantly from uniformly distributed data in multidimensional space and whether it make sense to create clustering.

The Hopkins Test value for our dataset is : ~ 88

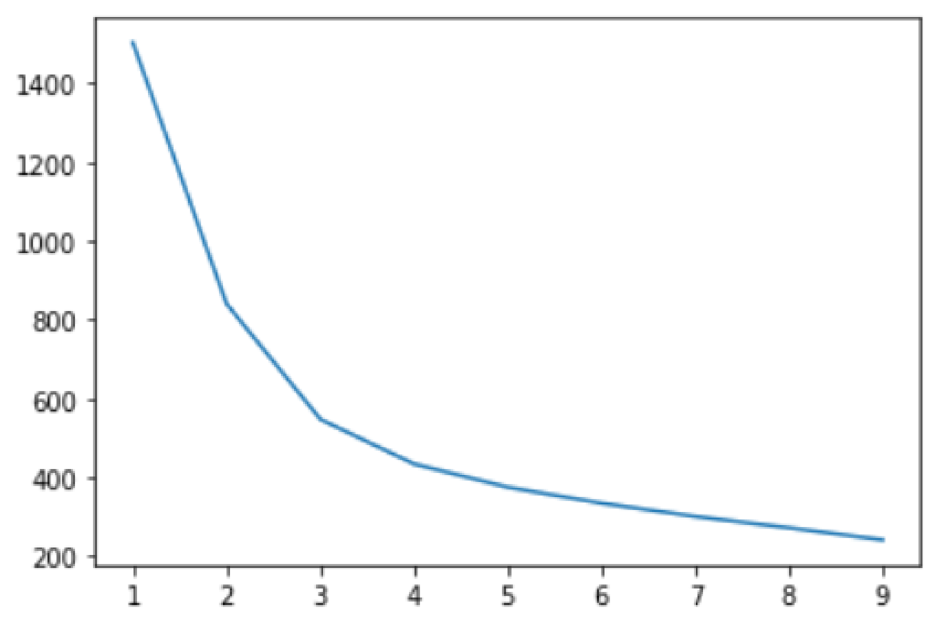
Selecting the optimal cluster number

From the below Silhouette and elbow curve, we see that the optimal no of clusters is 3 followed by 5.   Thus we build 2 models with both these values separately.

silhouette score analysis

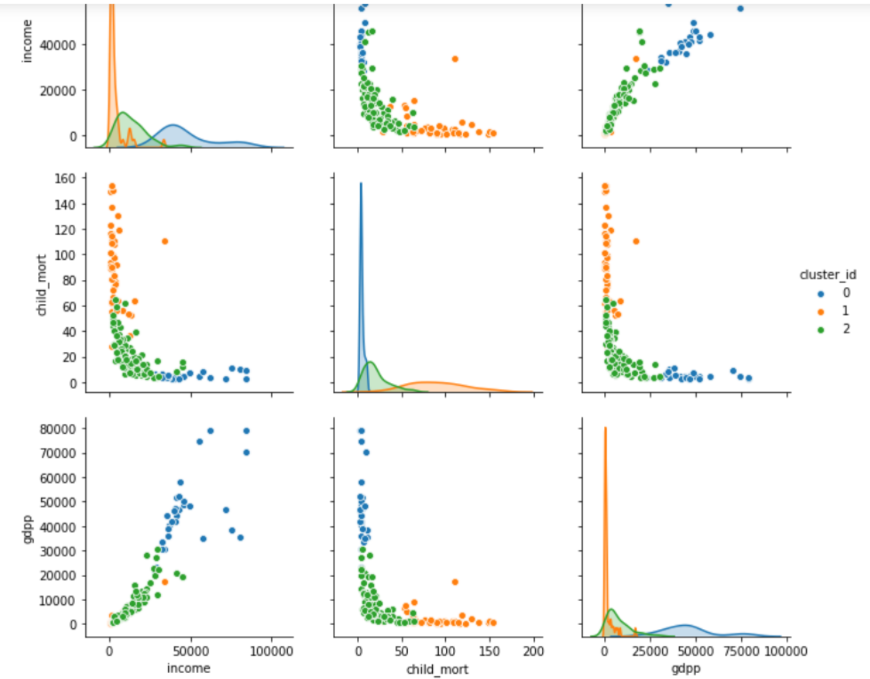


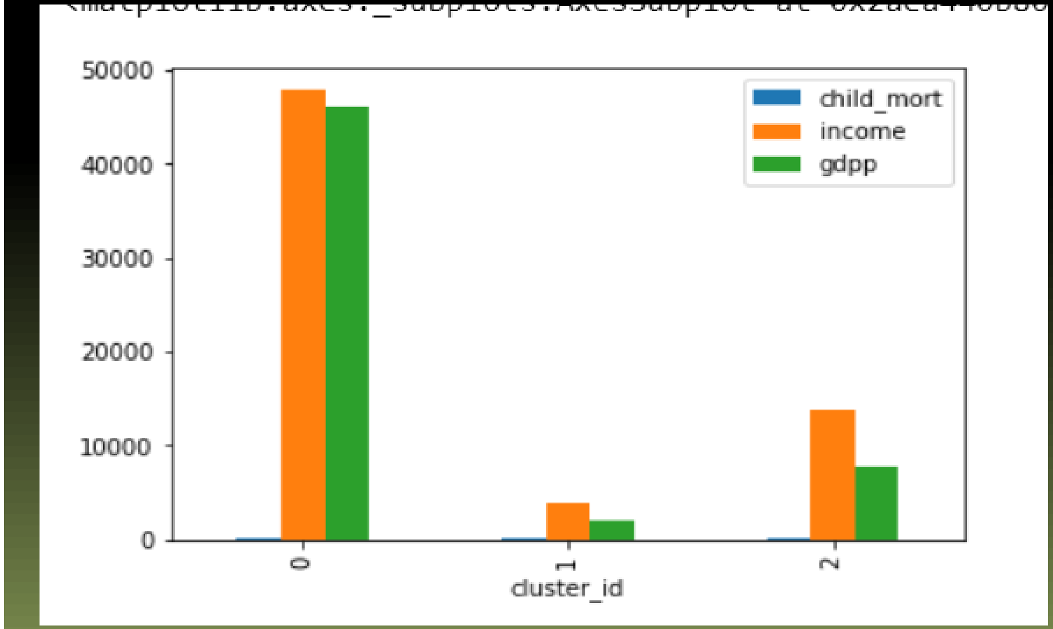
elbow curve method



K Means

By using K-Means Clustering Technique we found that certain cluster like child_mortality, income and gdpp helps us to identify the that the cluster_id 0 - is in a medium state, 2 are in a good condition and the countries under cluster_id 1 are in the direst need of aid.

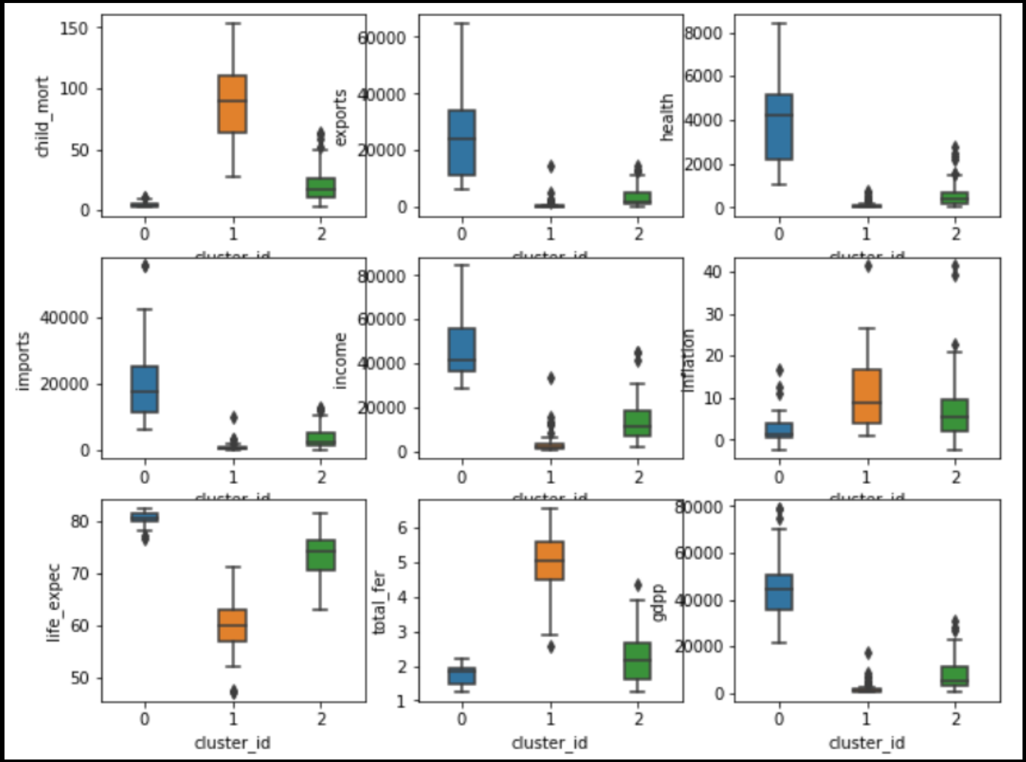


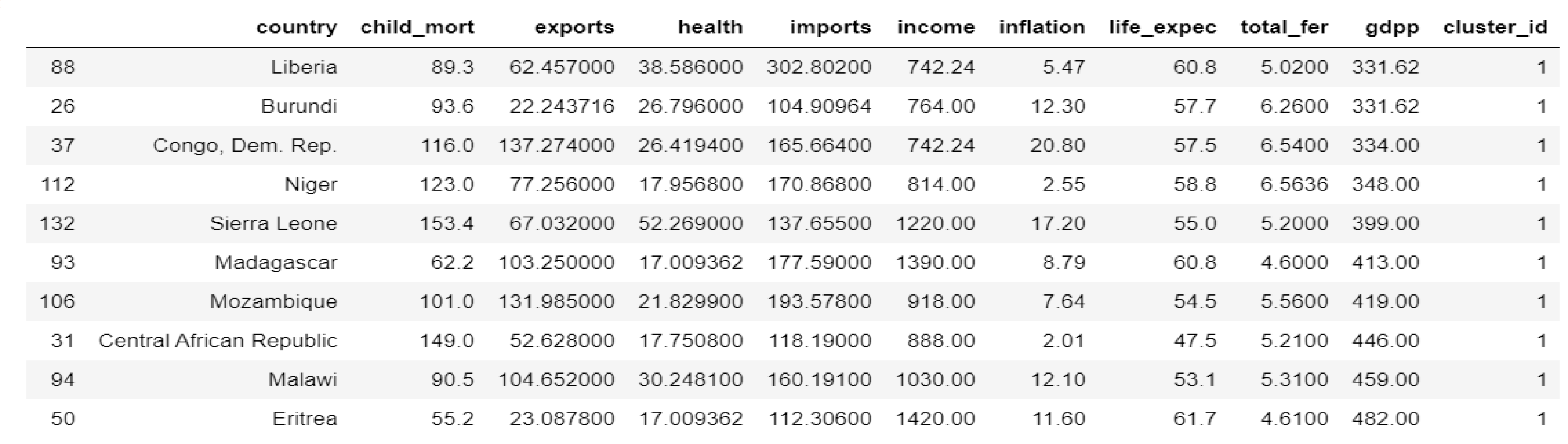


K Means Contd.....

As per our K mean clusters-
Cluster - 1 are a of concern due
to:

- Low gdpp
- Low income
- High childmortality



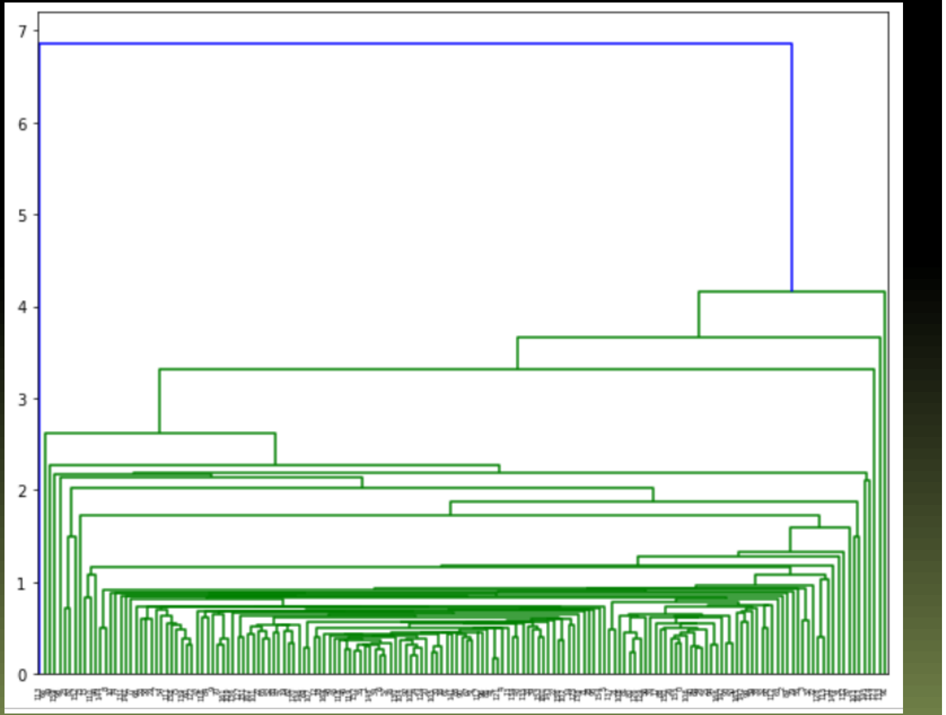


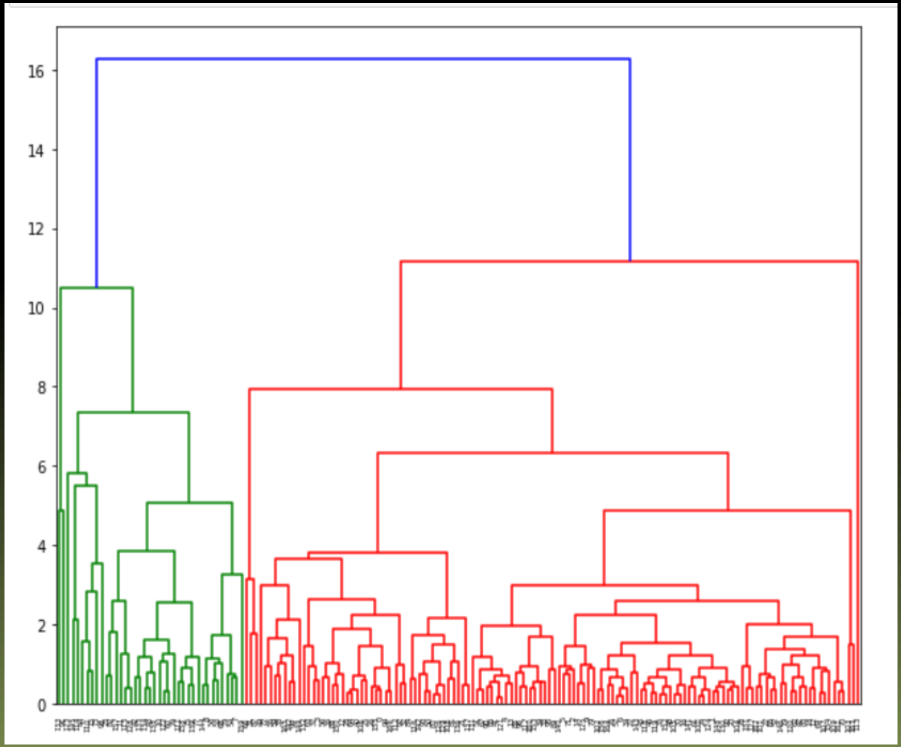
	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_id
88	Liberia	89.3	62.457000	38.586000	302.80200	742.24	5.47	60.8	5.0200	331.62	1
26	Burundi	93.6	22.243716	26.796000	104.90964	764.00	12.30	57.7	6.2600	331.62	1
37	Congo, Dem. Rep.	116.0	137.274000	26.419400	165.66400	742.24	20.80	57.5	6.5400	334.00	1
112	Niger	123.0	77.256000	17.956800	170.86800	814.00	2.55	58.8	6.5636	348.00	1
132	Sierra Leone	153.4	67.032000	52.269000	137.65500	1220.00	17.20	55.0	5.2000	399.00	1
93	Madagascar	62.2	103.250000	17.009362	177.59000	1390.00	8.79	60.8	4.6000	413.00	1
106	Mozambique	101.0	131.985000	21.829900	193.57800	918.00	7.64	54.5	5.5600	419.00	1
31	Central African Republic	149.0	52.628000	17.750800	118.19000	888.00	2.01	47.5	5.2100	446.00	1
94	Malawi	90.5	104.652000	30.248100	160.19100	1030.00	12.10	53.1	5.3100	459.00	1
50	Eritrea	55.2	23.087800	17.009362	112.30600	1420.00	11.60	61.7	4.6100	482.00	1

Clustering using Hierarchical Method

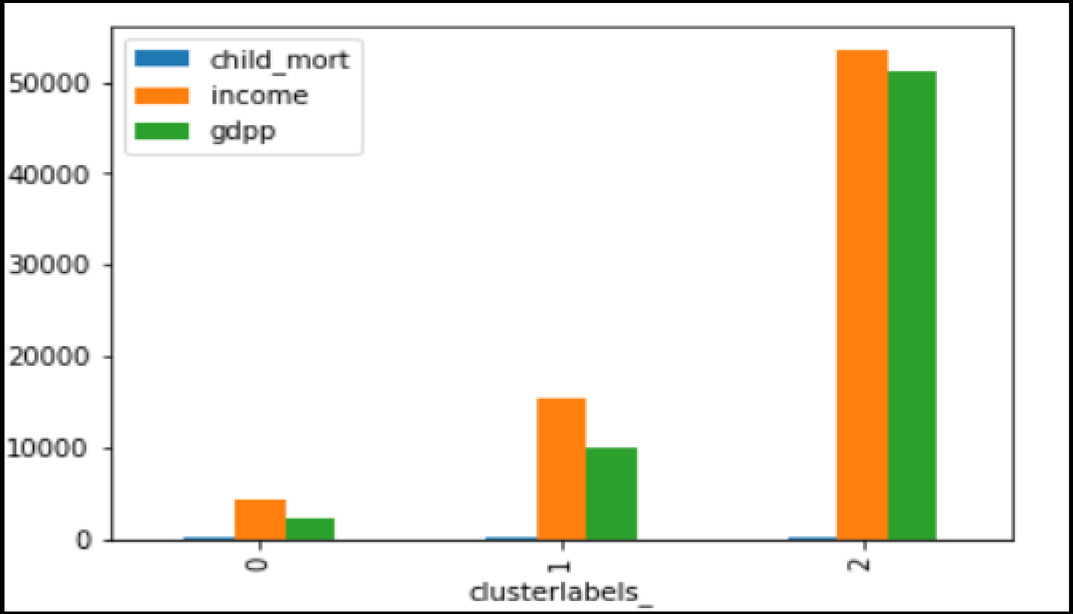
- The clustering process uses are hierarchical clustering single method and complete linkage to ensure the cluster are stable and close knit.
- We are going for **Complete method hierarchical** clustering as below single method clustering is not clear. By looking at this dendogram taking n-clusters as 3.

Single method hierarchical clustering



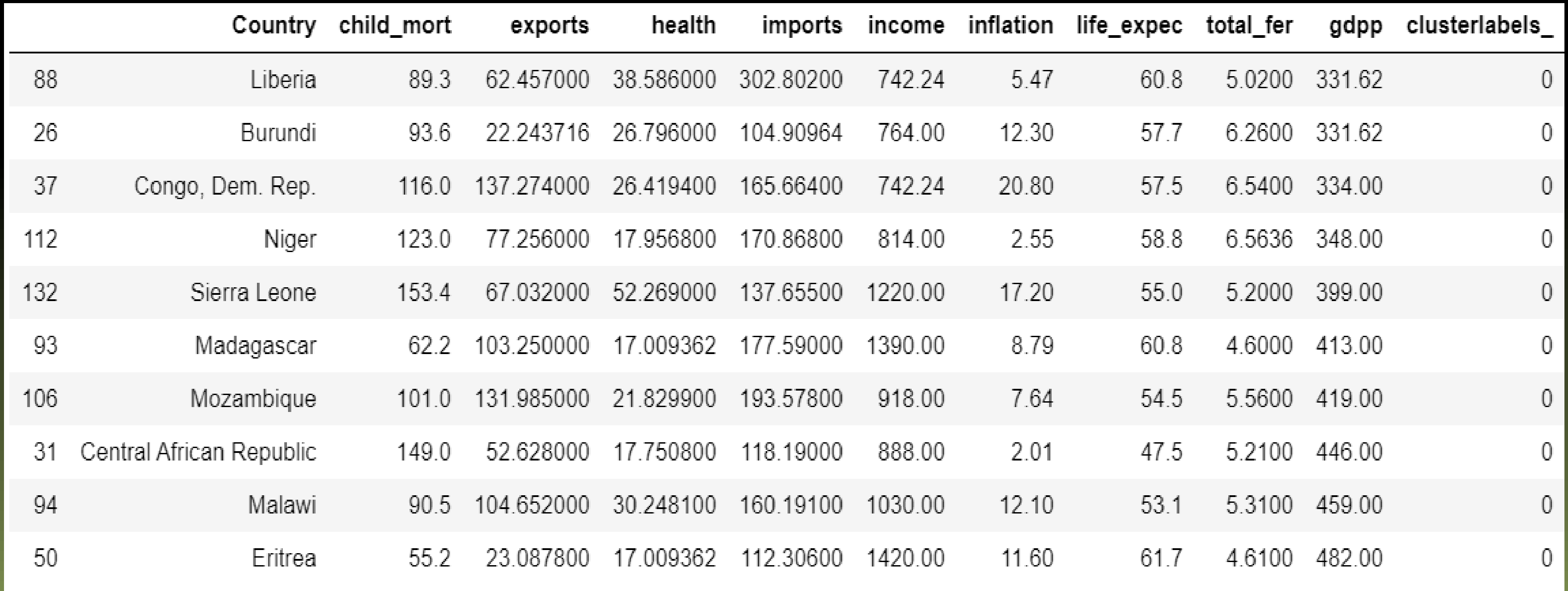


Hierarchical Clustering



As per our Hierarchical clusters- Cluster - O are a of concern due to:

- Low gdpp
- Low income
- High childmortality



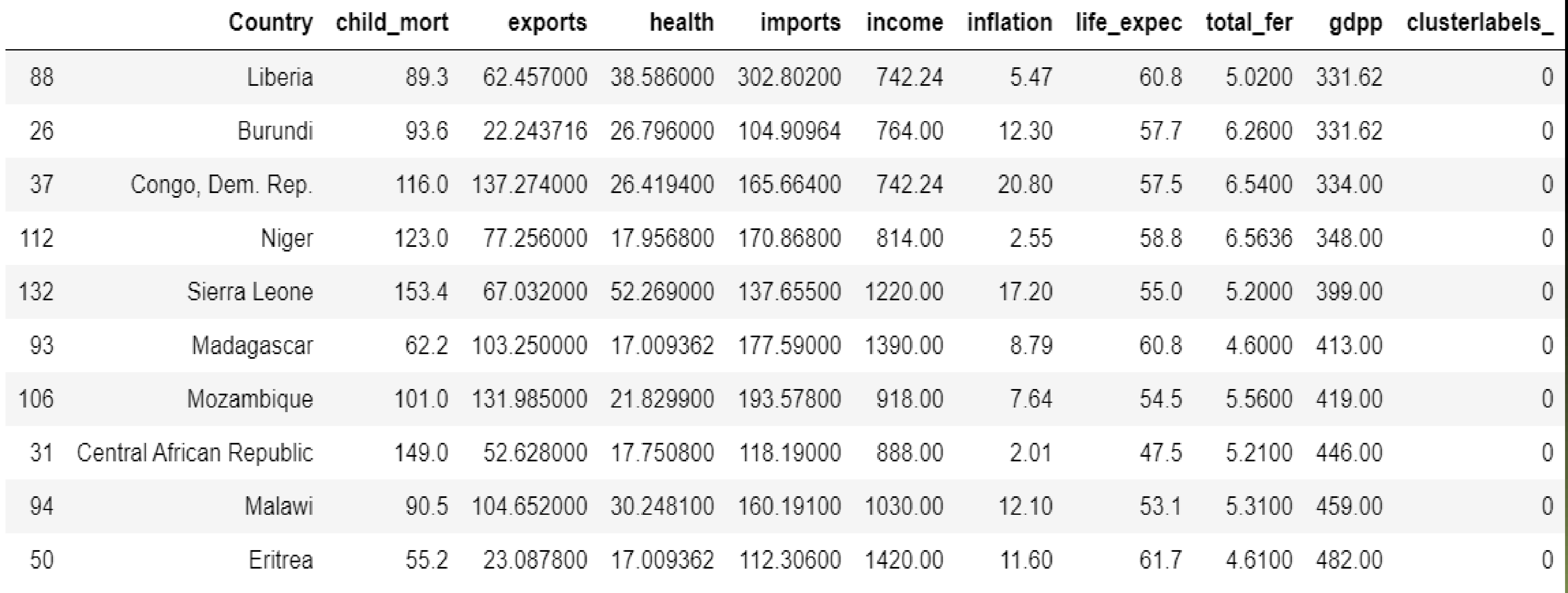
	Country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	clusterlabels_
88	Liberia	89.3	62.457000	38.586000	302.80200	742.24	5.47	60.8	5.0200	331.62	0
26	Burundi	93.6	22.243716	26.796000	104.90964	764.00	12.30	57.7	6.2600	331.62	0
37	Congo, Dem. Rep.	116.0	137.274000	26.419400	165.66400	742.24	20.80	57.5	6.5400	334.00	0
112	Niger	123.0	77.256000	17.956800	170.86800	814.00	2.55	58.8	6.5636	348.00	0
132	Sierra Leone	153.4	67.032000	52.269000	137.65500	1220.00	17.20	55.0	5.2000	399.00	0
93	Madagascar	62.2	103.250000	17.009362	177.59000	1390.00	8.79	60.8	4.6000	413.00	0
106	Mozambique	101.0	131.985000	21.829900	193.57800	918.00	7.64	54.5	5.5600	419.00	0
31	Central African Republic	149.0	52.628000	17.750800	118.19000	888.00	2.01	47.5	5.2100	446.00	0
94	Malawi	90.5	104.652000	30.248100	160.19100	1030.00	12.10	53.1	5.3100	459.00	0
50	Eritrea	55.2	23.087800	17.009362	112.30600	1420.00	11.60	61.7	4.6100	482.00	0

Summary

As by both k-means and Hierarchical clustering method-we have got same countries which requires aid. The following are the countries which are in direst need of aid by considering socio–economic factor in to consideration:

The order of precedence given to the features is ggdp, child_mort and then the income and the top 5 countries are

❑ Liberia
❑ Burundi
❑ Congo, Dem. Rep.
❑ Niger
❑ Sierra Leone



	Country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	clusterlabels_
88	Liberia	89.3	62.457000	38.586000	302.80200	742.24	5.47	60.8	5.0200	331.62	0
26	Burundi	93.6	22.243716	26.796000	104.90964	764.00	12.30	57.7	6.2600	331.62	0
37	Congo, Dem. Rep.	116.0	137.274000	26.419400	165.66400	742.24	20.80	57.5	6.5400	334.00	0
112	Niger	123.0	77.256000	17.956800	170.86800	814.00	2.55	58.8	6.5636	348.00	0
132	Sierra Leone	153.4	67.032000	52.269000	137.65500	1220.00	17.20	55.0	5.2000	399.00	0
93	Madagascar	62.2	103.250000	17.009362	177.59000	1390.00	8.79	60.8	4.6000	413.00	0
106	Mozambique	101.0	131.985000	21.829900	193.57800	918.00	7.64	54.5	5.5600	419.00	0
31	Central African Republic	149.0	52.628000	17.750800	118.19000	888.00	2.01	47.5	5.2100	446.00	0
94	Malawi	90.5	104.652000	30.248100	160.19100	1030.00	12.10	53.1	5.3100	459.00	0
50	Eritrea	55.2	23.087800	17.009362	112.30600	1420.00	11.60	61.7	4.6100	482.00	0

Thank You!