Clustering Assignment

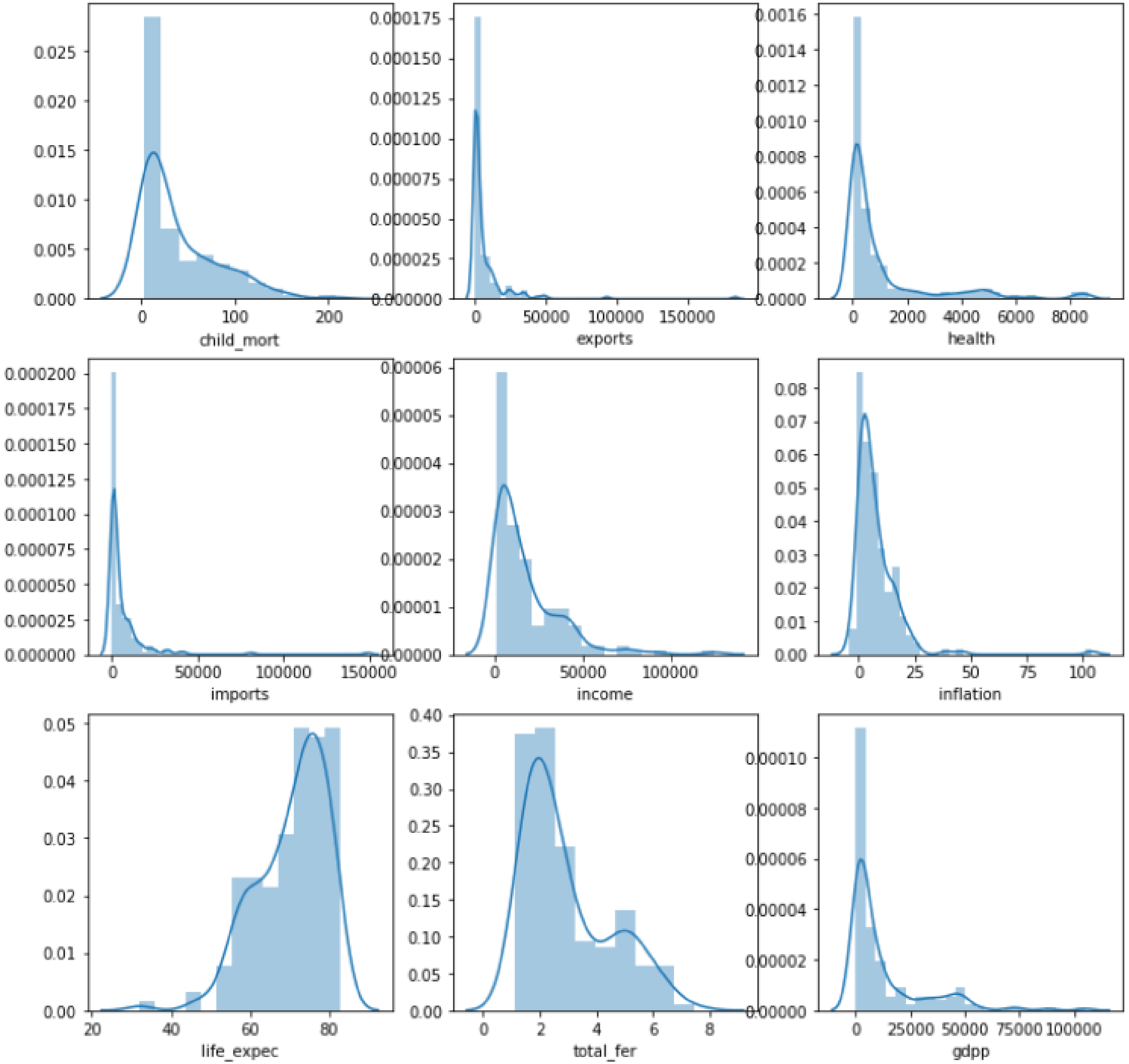
Problem Statement :

- To analyse the facts and figures in dataset of Countries, 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. Now decision of how to use this money strategically and effectively providing countries who need AID.

Analysis Approach :

- Following EDA steps starting from inspecting the dataframe and doing data outlier treatment, Non-Graphical Analysis to Graphical Analysis continuing with Scaling, Checking the tendency of the data: Hopkins Test, finding the best value for K by SSD & silhouette method, Performing KMeans with the final value of k then Visualizing the clusters using scatter plot profiling: GDPP, CHILD_MORT, INCOME after than using Hierarchical clustering (single & complete) finding the required cluster and comparing list of countries from both clusters.

- Inspecting dataframe
- Null check
- Data preparation
- EDA
- Outlier treatment
- Scaling
- Hopkins test
- K means clustering
- Hierarchical clustering
- Comparing results



Univariate Analysis

Inference:

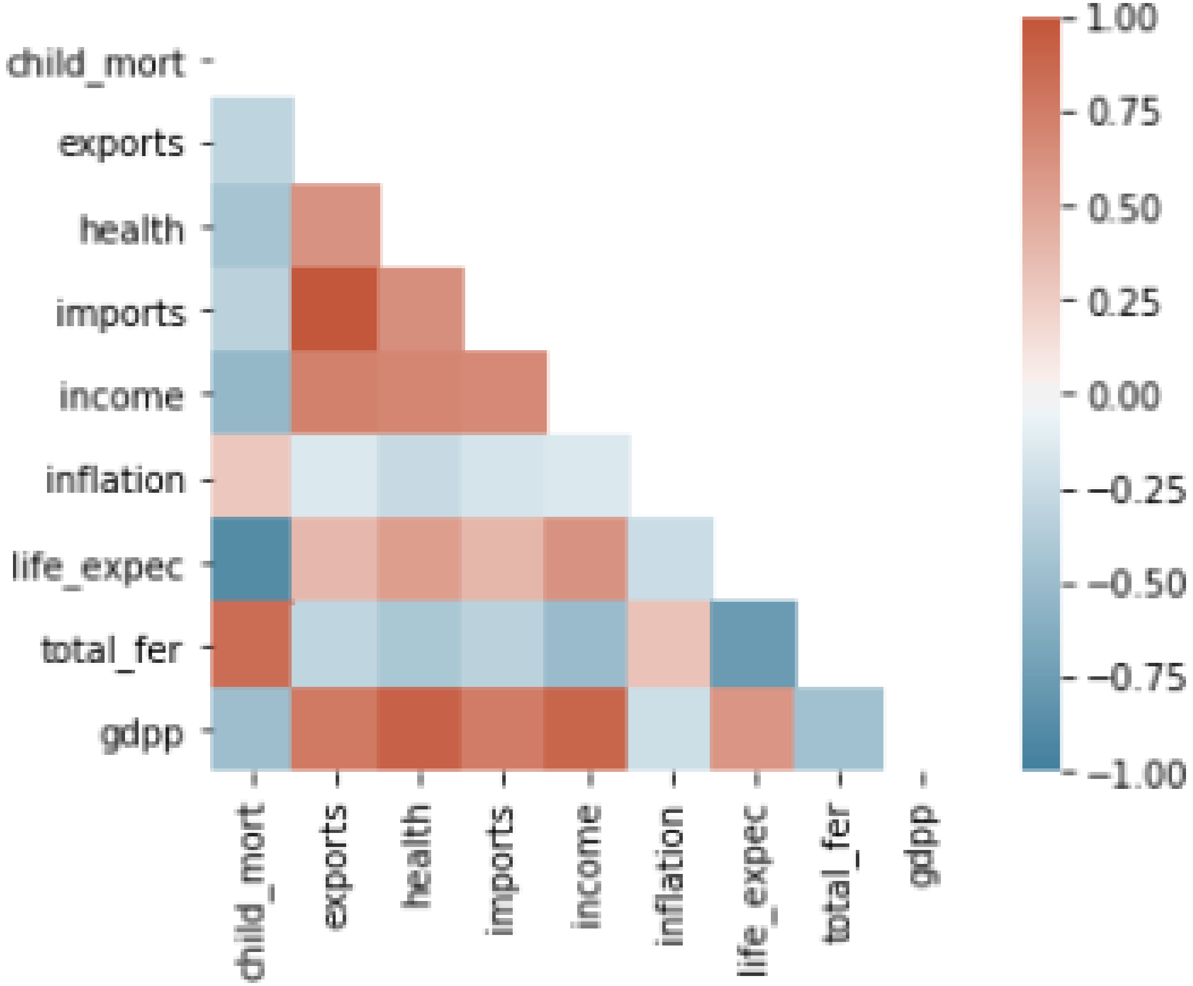
Child mortality, income, gdpp, exports, imports, health inflation have large outlier values.

Correlation of each column with other:

- Imports Exports increases simultaneously

- As GDP grows health and income increases

- Fertility and child moratlity gets infected directly by life_expectancy



Outlier Treatment :

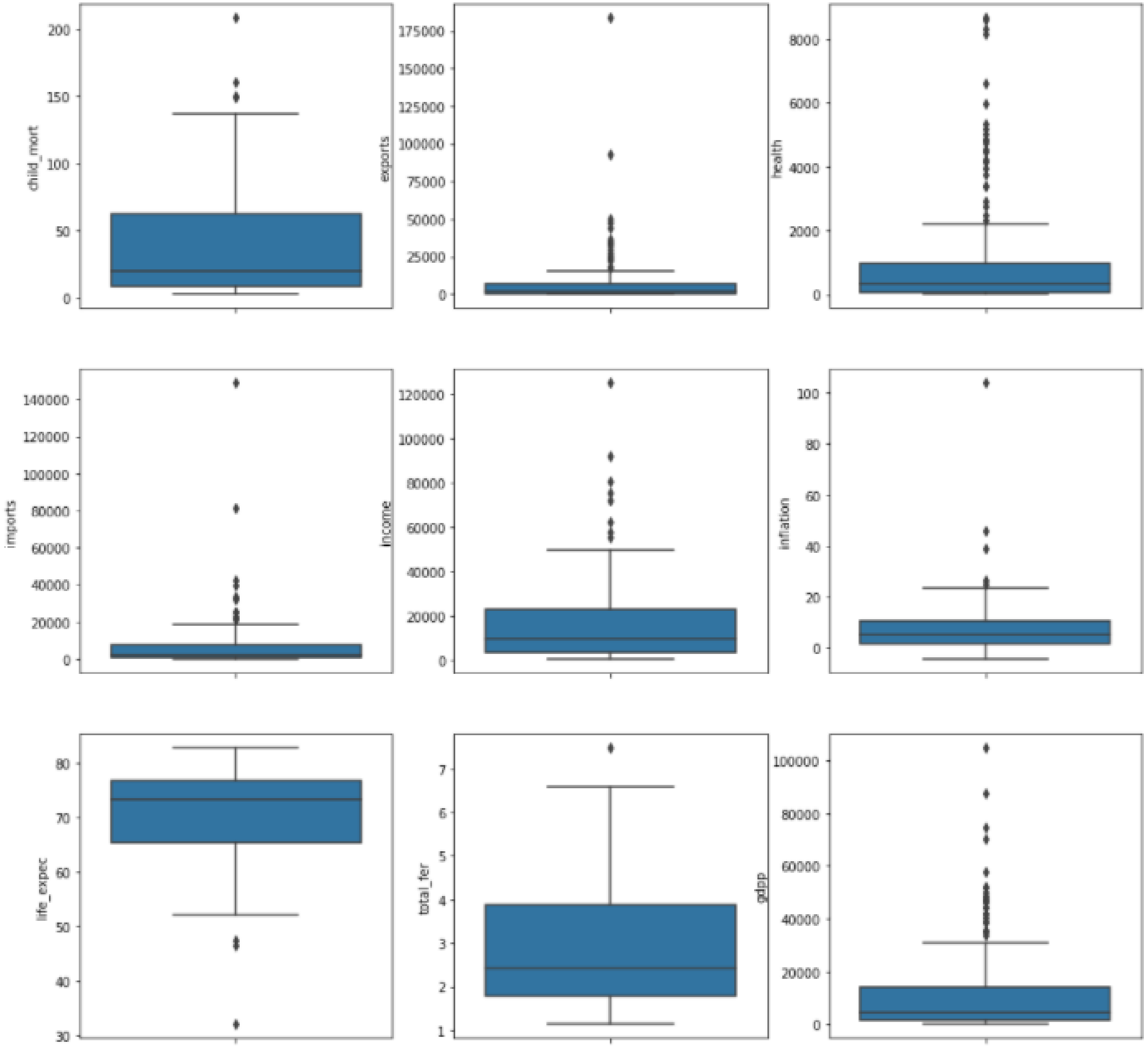
Capping values as per lower outliers:

'imports’ to ‘exports’ quantile to (0.05)

Capping values as per upper outliers:

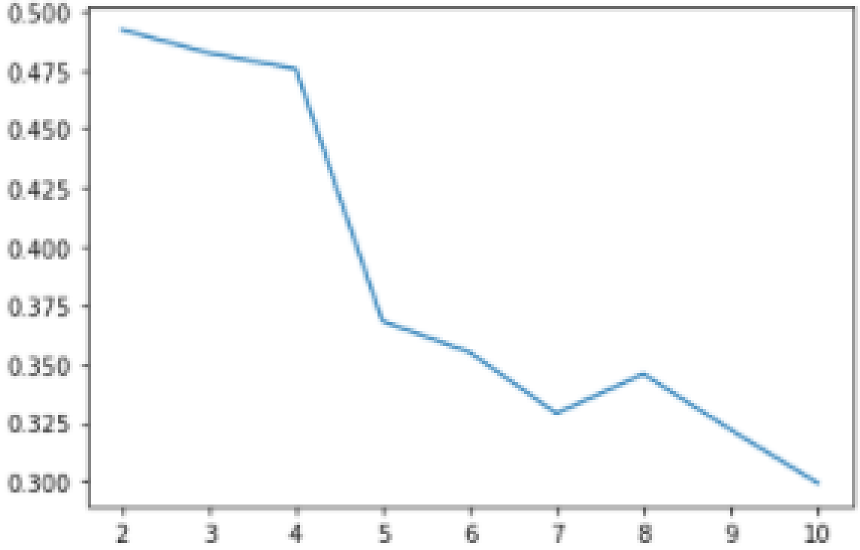
'inflation’, 'gdpp’ to quantile (0.99)
'exports’, 'health’, 'imports’, 'income’ to quantile(0.95)

Note: capping should be done with care as results may tempered.

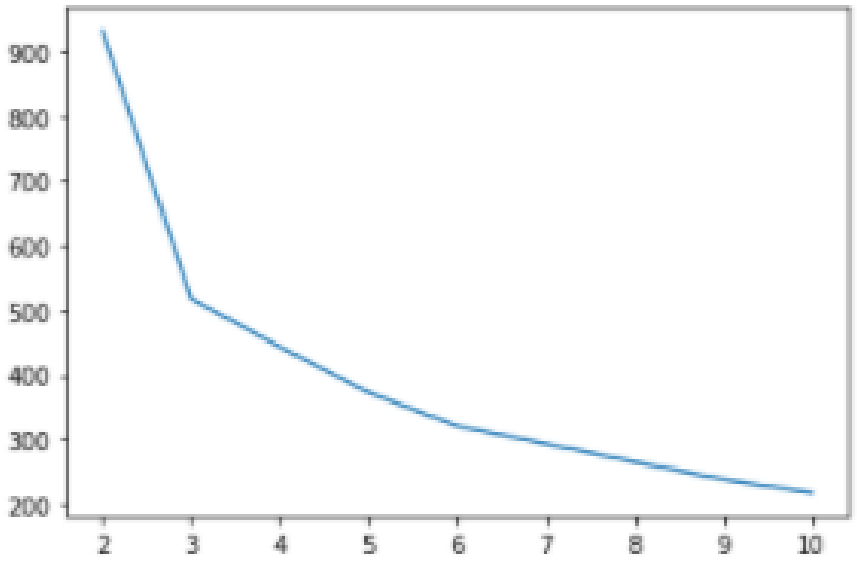


For 10 iteration of HOPKINS TEST value is greater than .85
Thus, data is good for clustering.

After scaling , plotting ssd and silhouette_score for Kmeans Clustering



silhouette_score



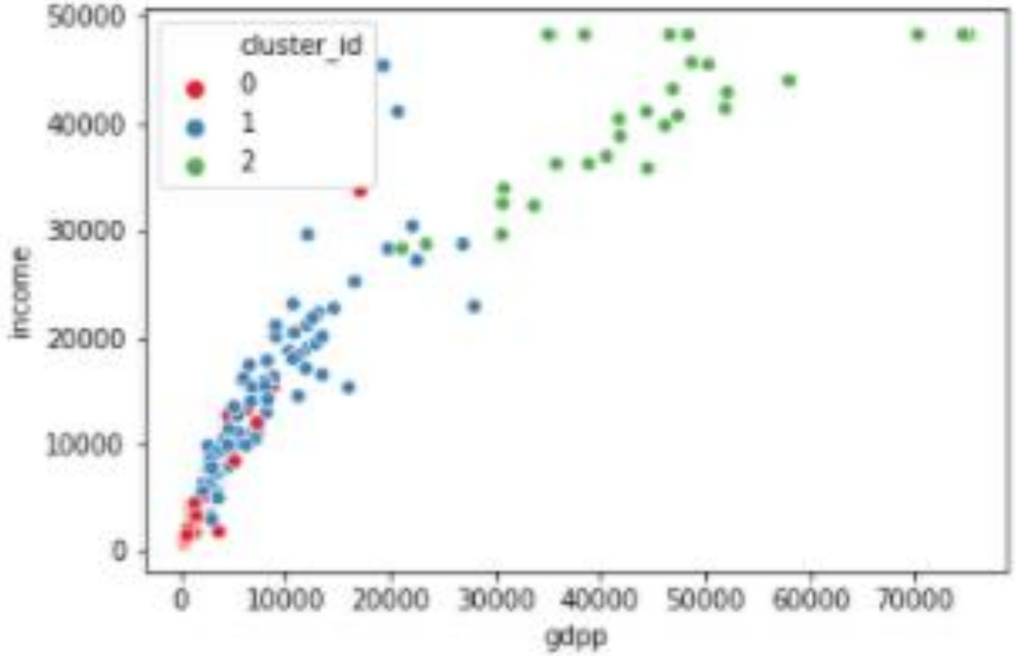
SSD : elbow curve

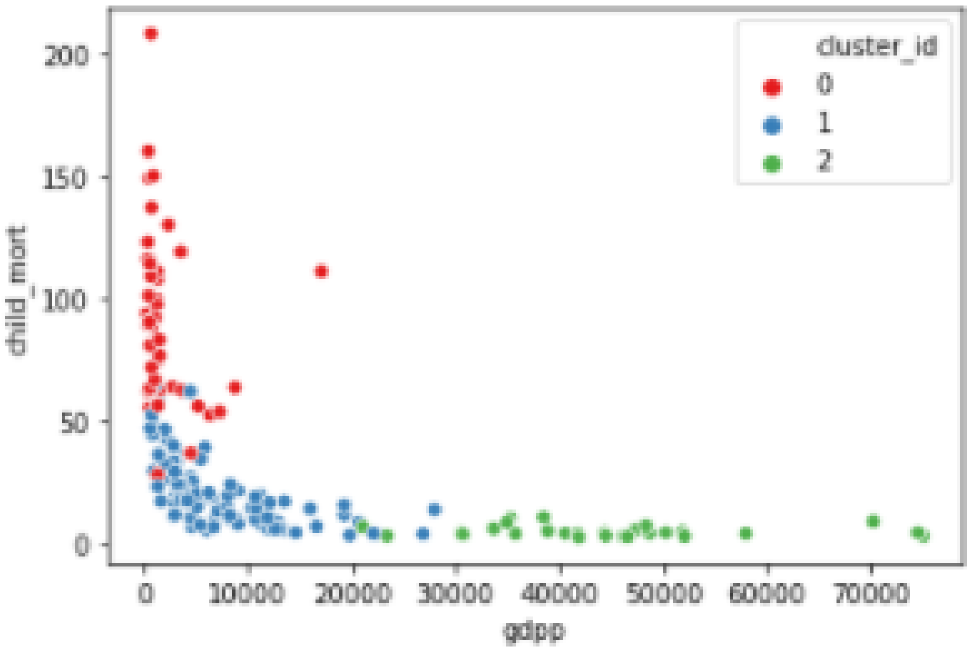
Thus, plotting for k==3 and k==4

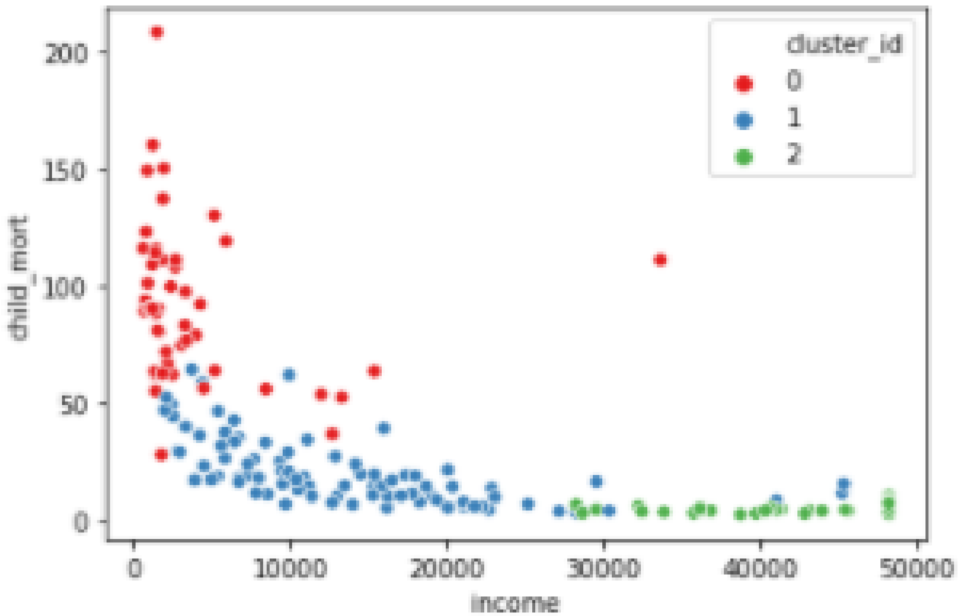
Scatterplot for K means :

(same plots for k==3 and k==4)

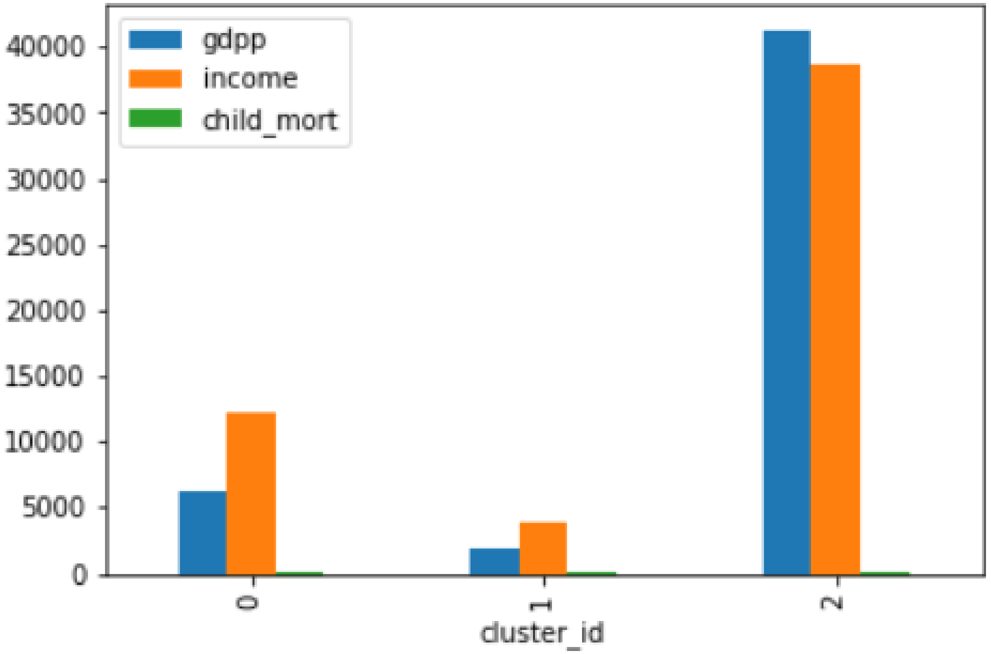
- GDPP
- Income
- Child_Mort







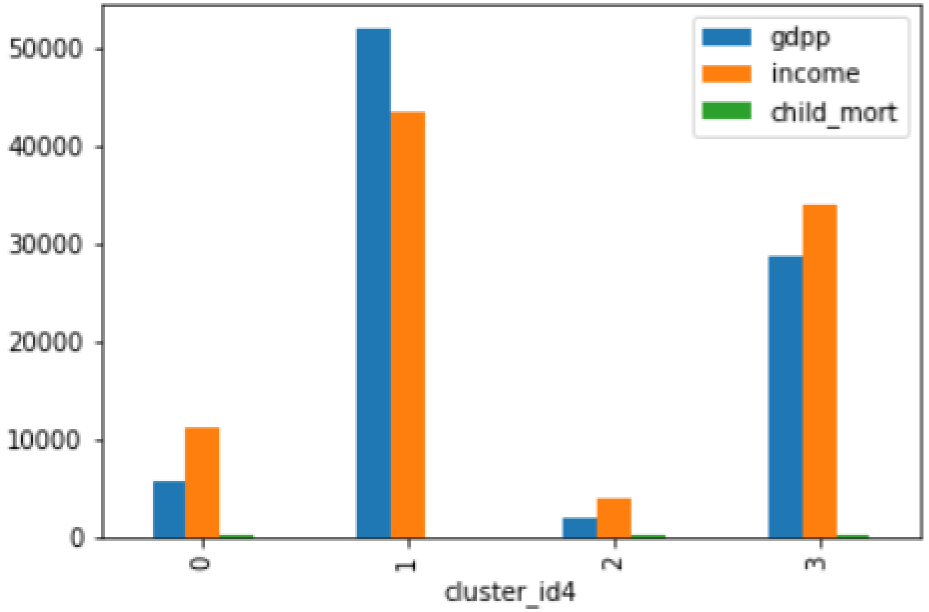
Barplot for different Clusters achieved after K means K==3:



Using the above chart about clusters, drawing the following inferences:

- Cluster 0 has low income, child_mort and GDP than others indicating these countries have stunted growth and need aid.
- Cluster 1 has good income, child_mort and GDP. Thus, they don't need much treatment.
- Cluster 2 has highest income, child_mort and GDP indicating these countries are developed and does not need aid.

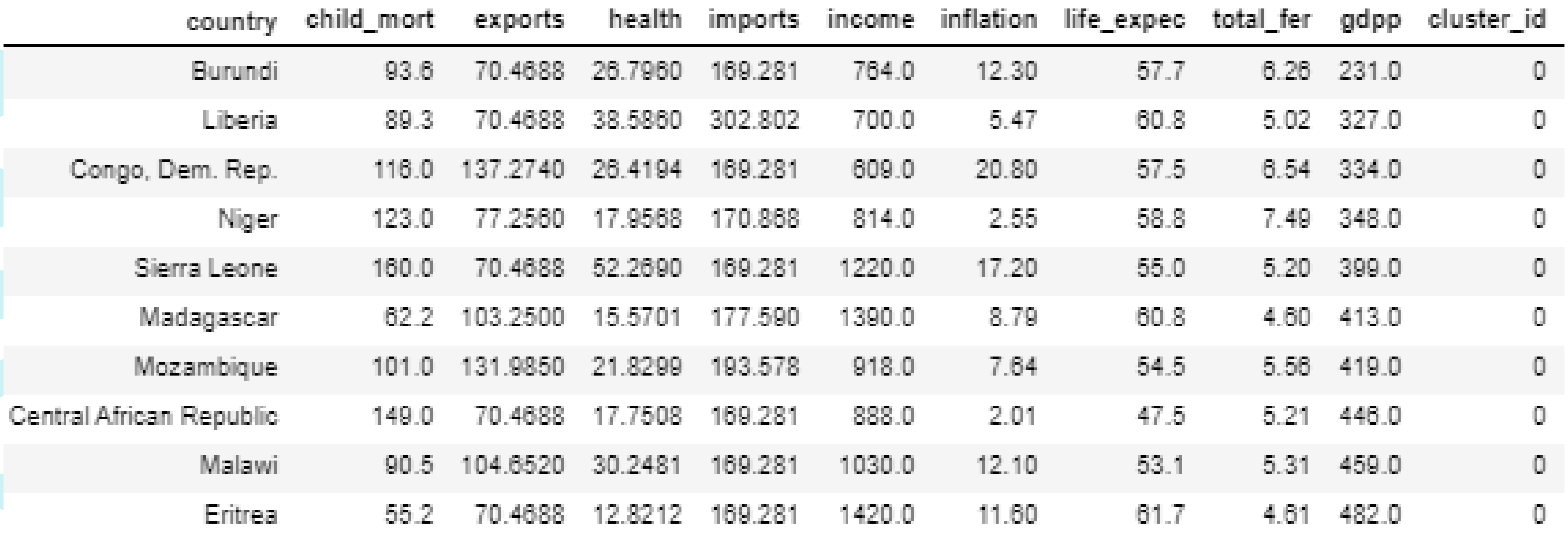
Barplot for different Clusters achieved after K means K==4:



Using the above chart about clusters, drawing the following inferences:

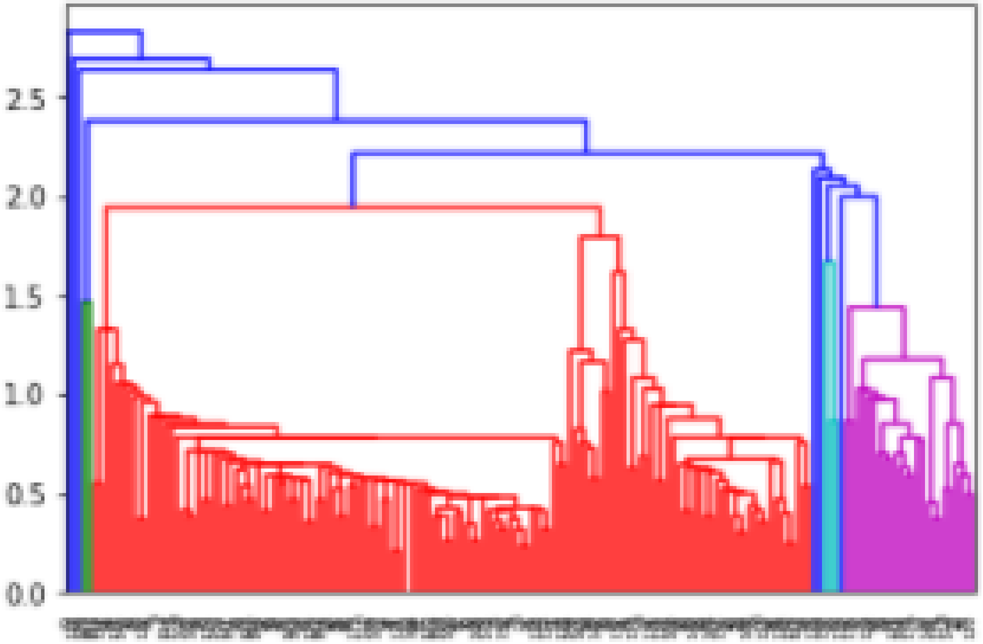
- Cluster 0 has low income , GDPP and high child_mort than others indicating these countries have stunted growth and may need aid.
- Cluster 1 has highest GDPP, income and lowest child_mort indicating these countries are developed and does not need aid.
- Cluster 2 has lowest GDPP, income and highest child_mort. Thus, they don't need much treatment.
- Cluster 3 has good income, GDP and low child_mort. Thus, they don't need much treatment.

Top 10 countries dataset for cluster having low ‘gdpp’, high 'child_mort’, and low 'income’ after Kmeans both K==3&4: (same set of countries)

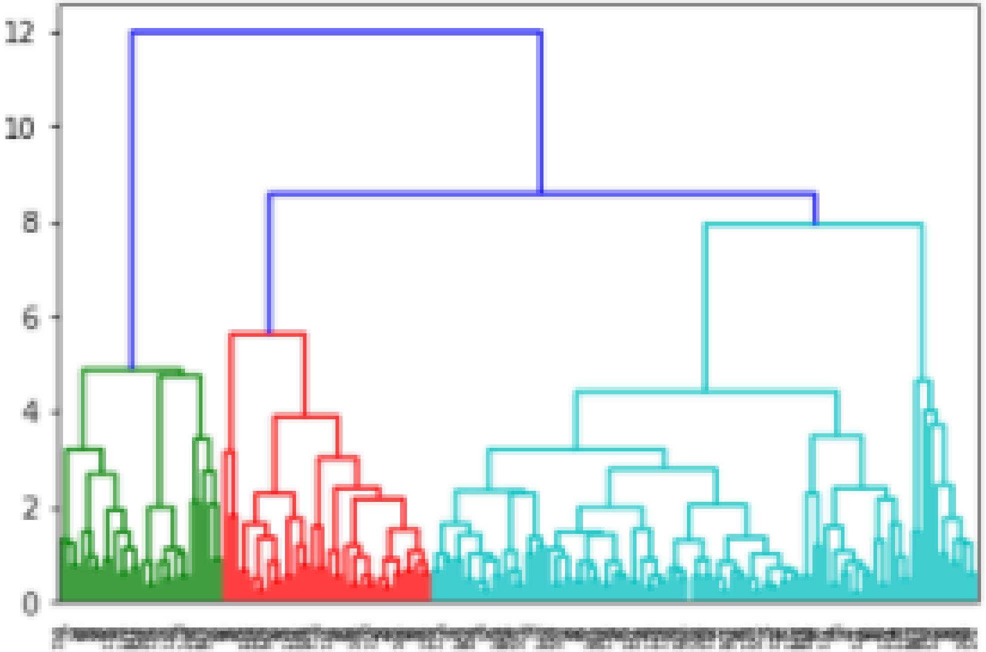


country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_id
Burundi	93.6	70.4688	26.7960	169.281	764.0	12.30	57.7	6.26	231.0	0
Liberia	89.3	70.4688	38.5860	302.802	700.0	5.47	60.8	5.02	327.0	0
Congo, Dem. Rep.	116.0	137.2740	26.4194	169.281	609.0	20.80	57.5	6.54	334.0	0
Niger	123.0	77.2560	17.9568	170.868	814.0	2.55	58.8	7.49	348.0	0
Sierra Leone	160.0	70.4688	52.2690	169.281	1220.0	17.20	55.0	5.20	399.0	0
Madagascar	62.2	103.2500	15.5701	177.590	1390.0	8.79	60.8	4.60	413.0	0
Mozambique	101.0	131.9850	21.8299	193.578	918.0	7.64	54.5	5.56	419.0	0
Central African Republic	149.0	70.4688	17.7508	169.281	888.0	2.01	47.5	5.21	446.0	0
Malawi	90.5	104.6520	30.2481	169.281	1030.0	12.10	53.1	5.31	459.0	0
Eritrea	55.2	70.4688	12.8212	169.281	1420.0	11.60	61.7	4.61	482.0	0

Hierarchical Clustering :



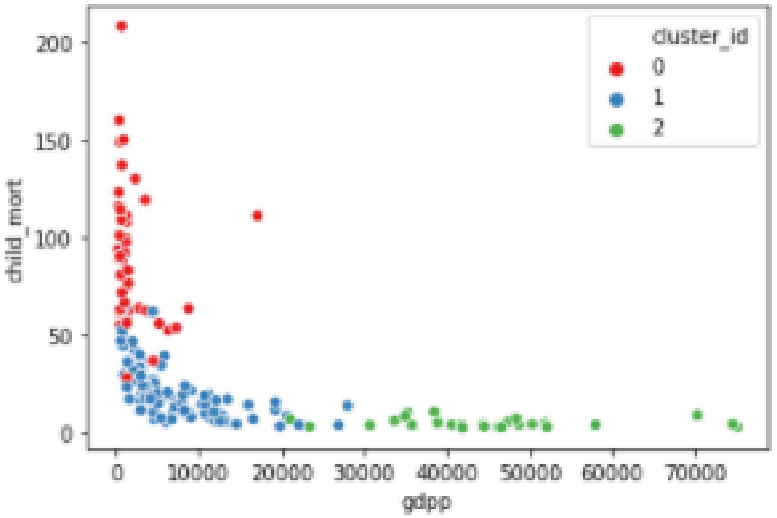
Single Linkage Dendrogram

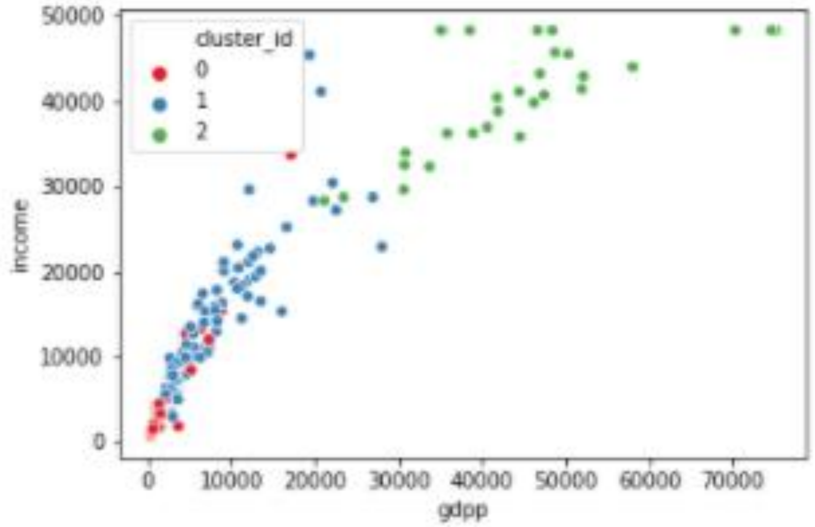


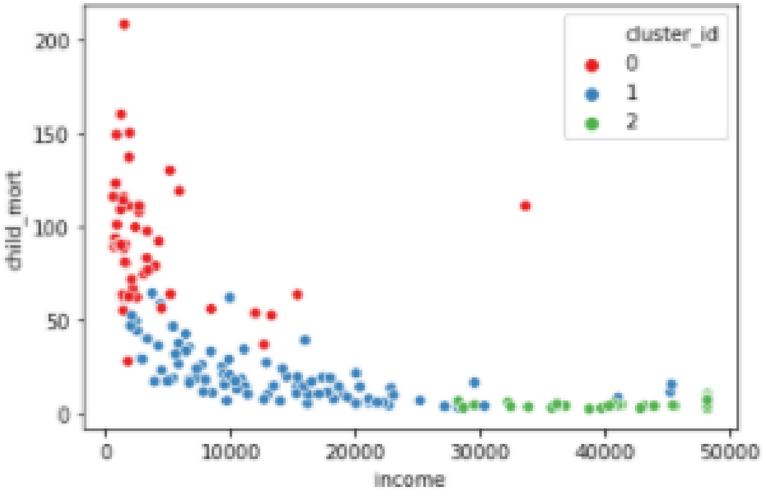
Complete Linkage Dendrogram

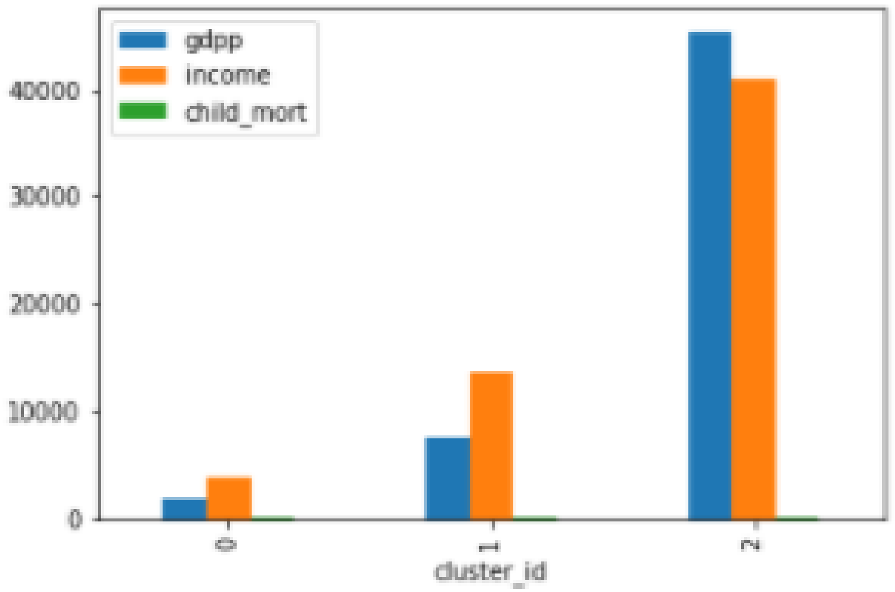
Getting same Scatterplot and Bar plot for Hierarchical clustering-

Clusters = 3 as per complete linkage dendrogram









Thus, lists of countries are same from both the Clustering Technique.

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

Thank You