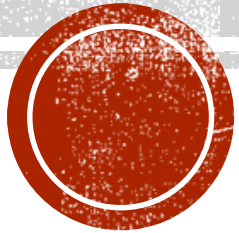
HELP- NGO

CLUSTERING ASSIGNMENT

By – Riya Jain



PROBLEM STATEMENT

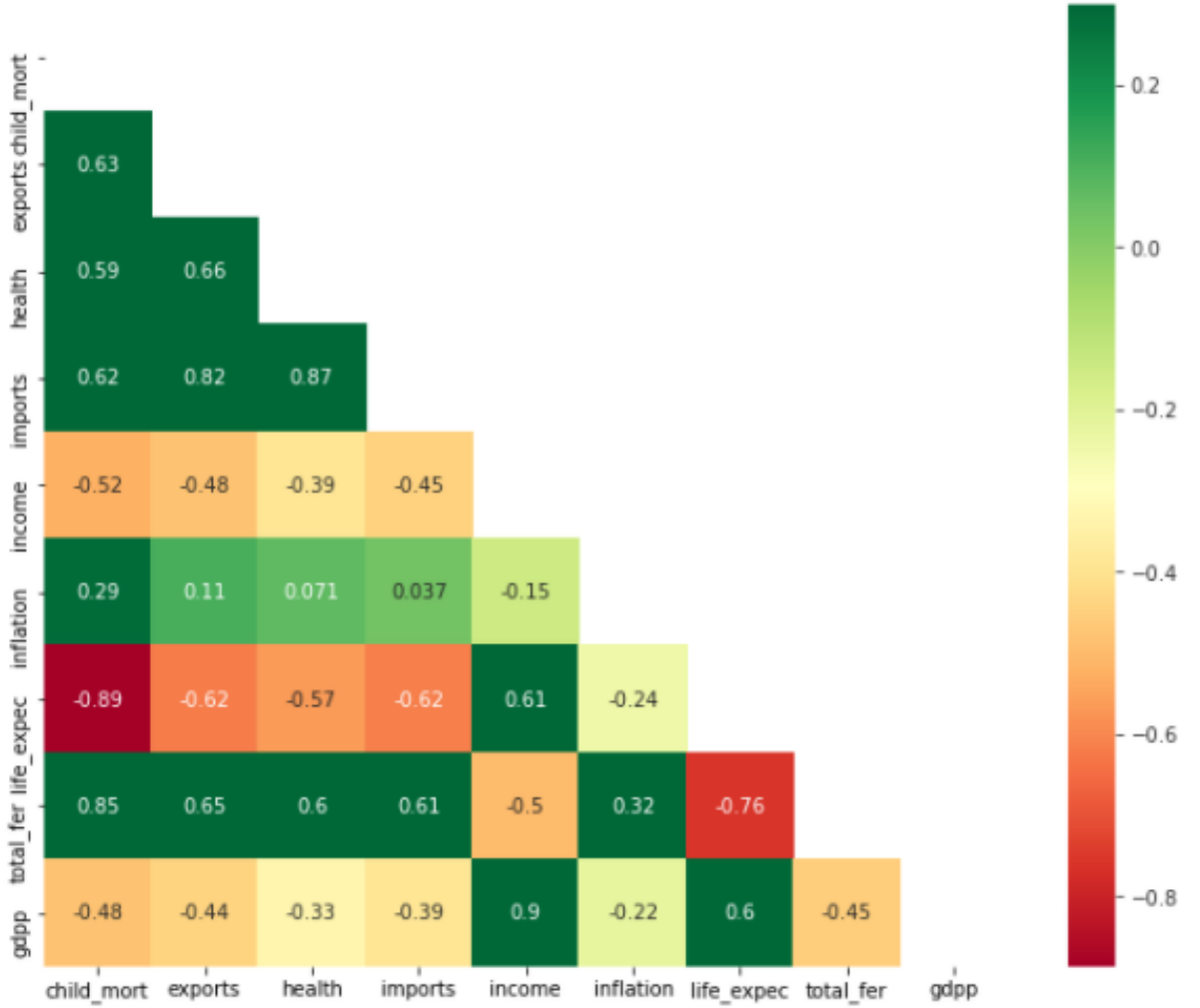
To analyse the facts and figures in dataset of Countries, HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities. Now decision of how to use this money strategically and effectively providing countries who need AID.

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.

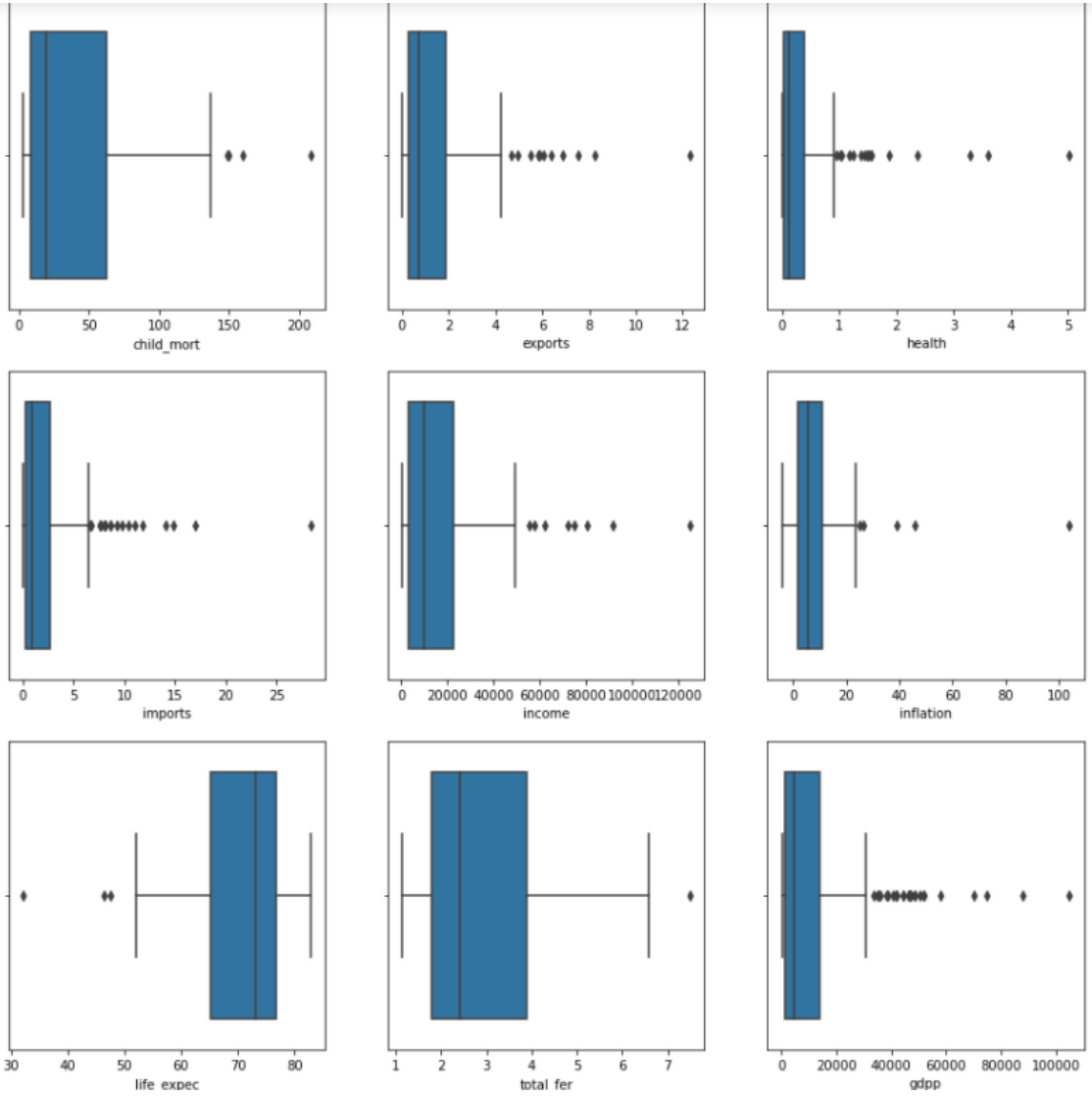
CORRELATION MATRIX FOR NUMERICAL COLUMNS:

- exports and imports are following a linear trend.
- income and life expectancy increases as GDP increase. Which is expected.
- life expectancy decreases with increase fertility and child mortality.

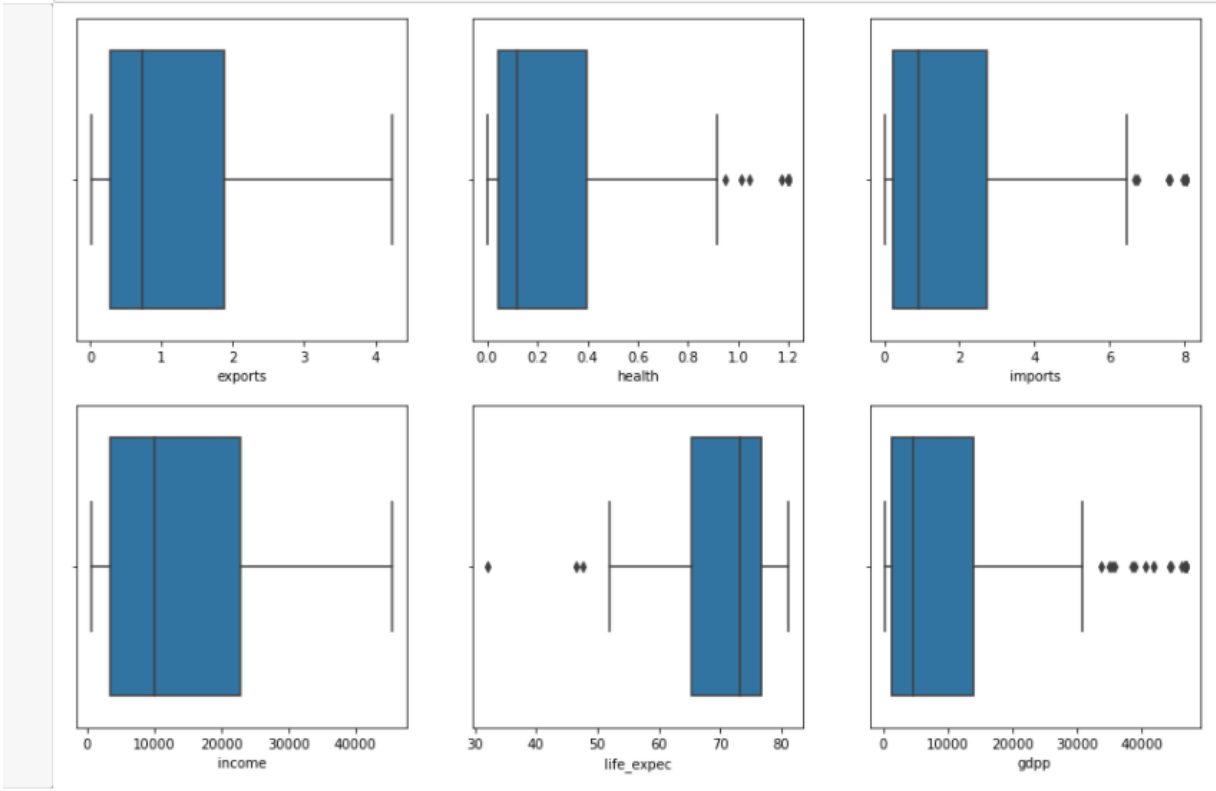


OUTLIER TREATMENT

Before Treatment



After Treatment



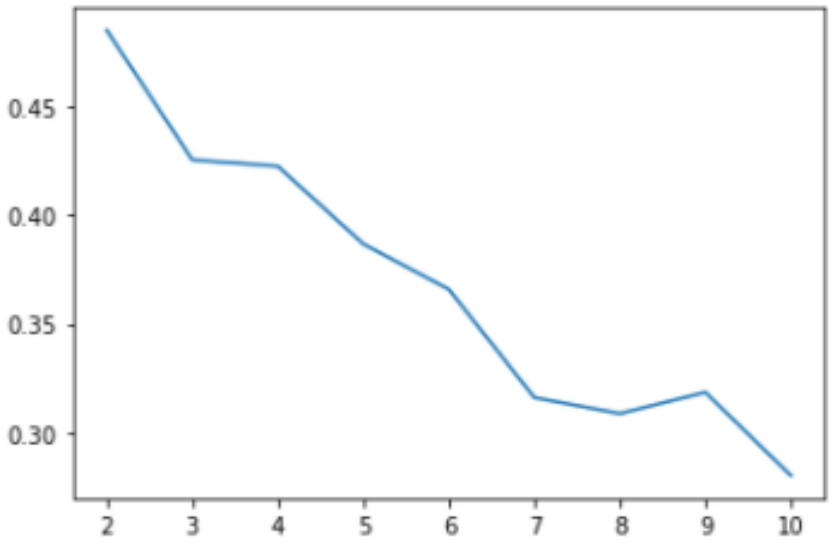
- Capping Upper range outliers for Exports, health, imports, income, life_expect, gdp to 0.93 percentile. They are not removed fully as this would hamper the data.
- No need of Lower range outlier capping.
- Note: capping should be done with care as results may tempered.

HOPKIN'S TEST : TO CHECK CLUSTER TENDENCY

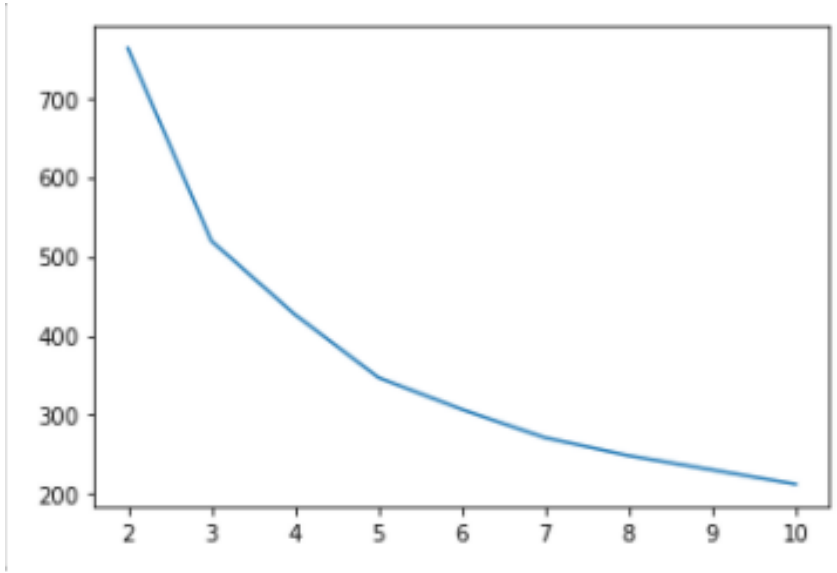
For 10 iteration of HOPKINS TEST value is greater than .86 Thus, data is good for clustering.

After scaling , plotting ssd and silhouette score for Kmeans Clustering to find best value of k

- Silhouette Score



- SSD Elbow Curve



Looking at Above Graphs, It would be best if we check for both k=3 and k=4

VISUALIZING CLUSTERS FOR BOTH K=3 & K=4

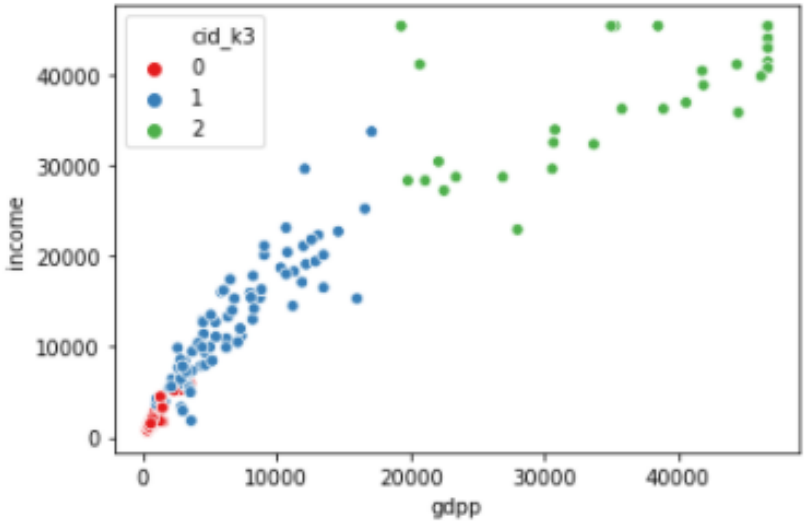
- **Visualize the cluster using scatter plots**
- Plot for:

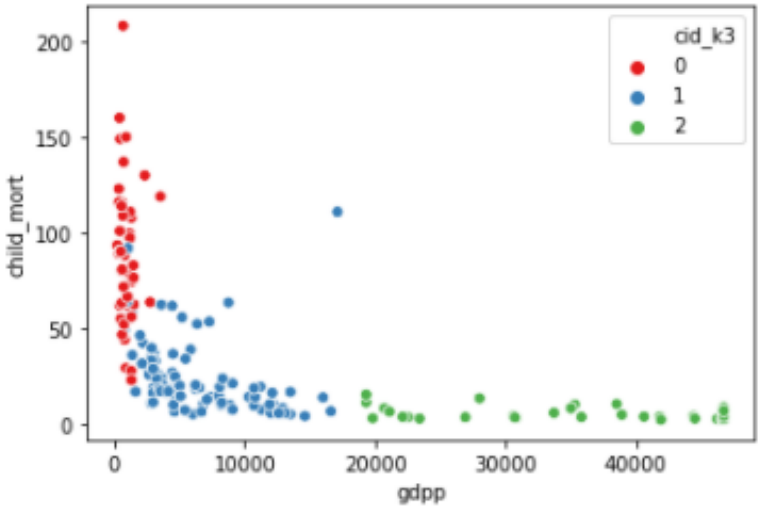
- GDPP

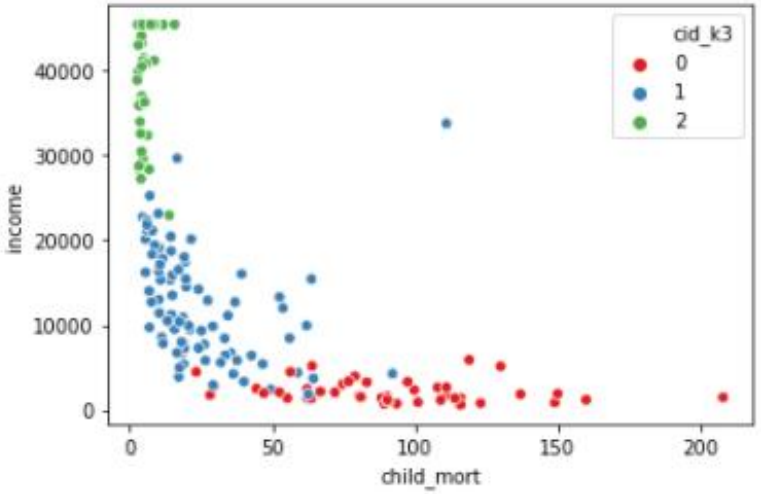
- Income

- Child_Mort

- Choosing these because there distplots look more promising for cluttering





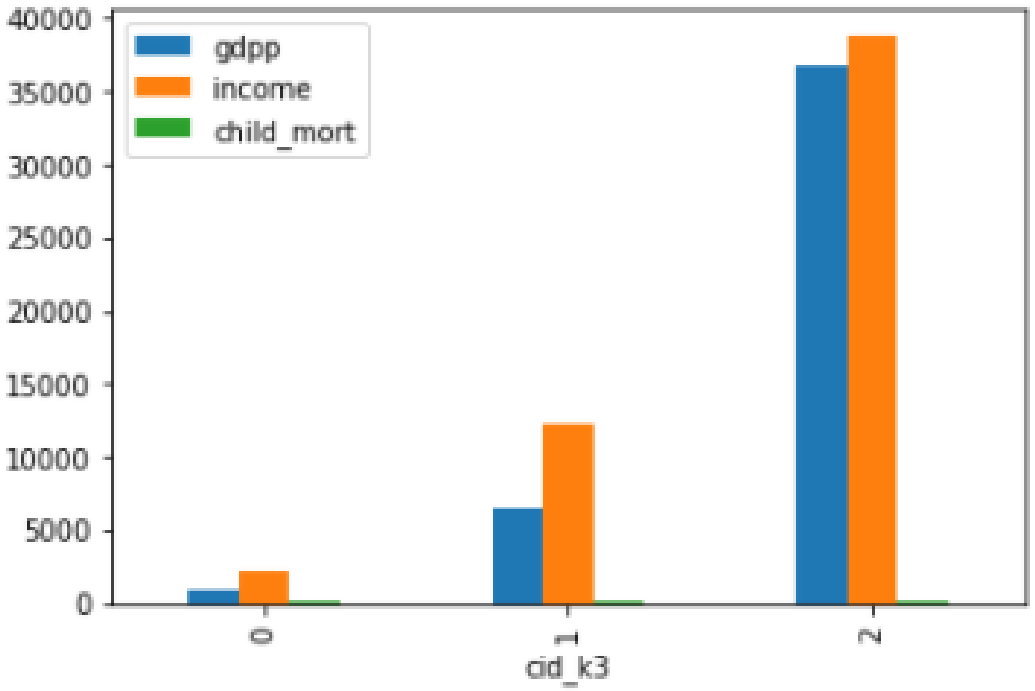


For both k=3 and k=4, getting same distribution of clusters.

CLUSTER PROFILING FOR K=3 & K=4

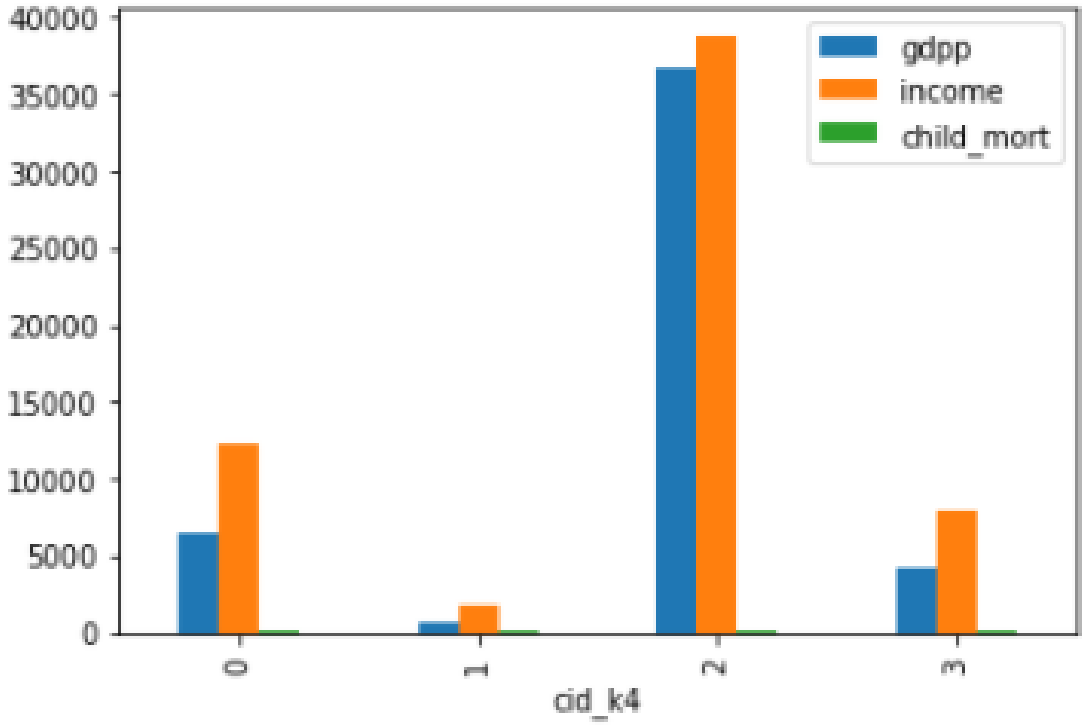
Barplot for different Clusters achieved after K means K==3:

- From this, following inferences can be drawn:
- Cluster 0 : lowest GDPP, Lowest income, Highest child_mort (can be seen from average as well). Countries in this cluster are in dire need of aid and less developed.
- Cluster 1 : low GDP, income and high child_mort Thus, they may need treatment.
- cluster 2 : highest income, GDP and lowest child_mort indicating these countries are developed and does not need aid.



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

- From this, following inferences can be drawn:
- Cluster 0 : low GDPP, Low income, High child_mort. But doesn't seem a good cluster, since it may need treatment.
- Cluster 1 : lowest GDP, income and highest child_mort. Seems like a good cluster to go with, they are in dire need of the aid.
- cluster 2 : highest income, GDP and lowest child_mort indicating these countries are developed and does not need aid.
- cluster 3 : low GDPP, Low income, High child_mort. But doesn't seem a good cluster, since it may need treatment.



TOP 10 COUNTRIES DATASET FOR CLUSTER HAVING LOW 'GDPP', HIGH 'CHILD_MORT', AND LOW 'INCOME' AFTER KMEANS BOTH K==3&4:

K=3

country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cid_k3
Burundi	93.6	3.861472	1.199254	8.008388	764	12.30	57.7	6.26	231	0
Liberia	89.3	4.228753	1.199254	8.008388	700	5.47	60.8	5.02	327	0
Congo, Dem. Rep.	116.0	4.228753	1.199254	8.008388	609	20.80	57.5	6.54	334	0
Niger	123.0	4.228753	1.199254	8.008388	814	2.55	58.8	7.49	348	0
Sierra Leone	160.0	4.210526	1.199254	8.008388	1220	17.20	55.0	5.20	399	0
Madagascar	62.2	4.228753	0.912833	8.008388	1390	8.79	60.8	4.60	413	0
Mozambique	101.0	4.228753	1.199254	8.008388	918	7.64	54.5	5.56	419	0
Central African Republic	149.0	2.645740	0.892377	5.941704	888	2.01	47.5	5.21	446	0
Malawi	90.5	4.228753	1.199254	7.603486	1030	12.10	53.1	5.31	459	0
Eritrea	55.2	0.993776	0.551867	4.834025	1420	11.60	61.7	4.61	482	0

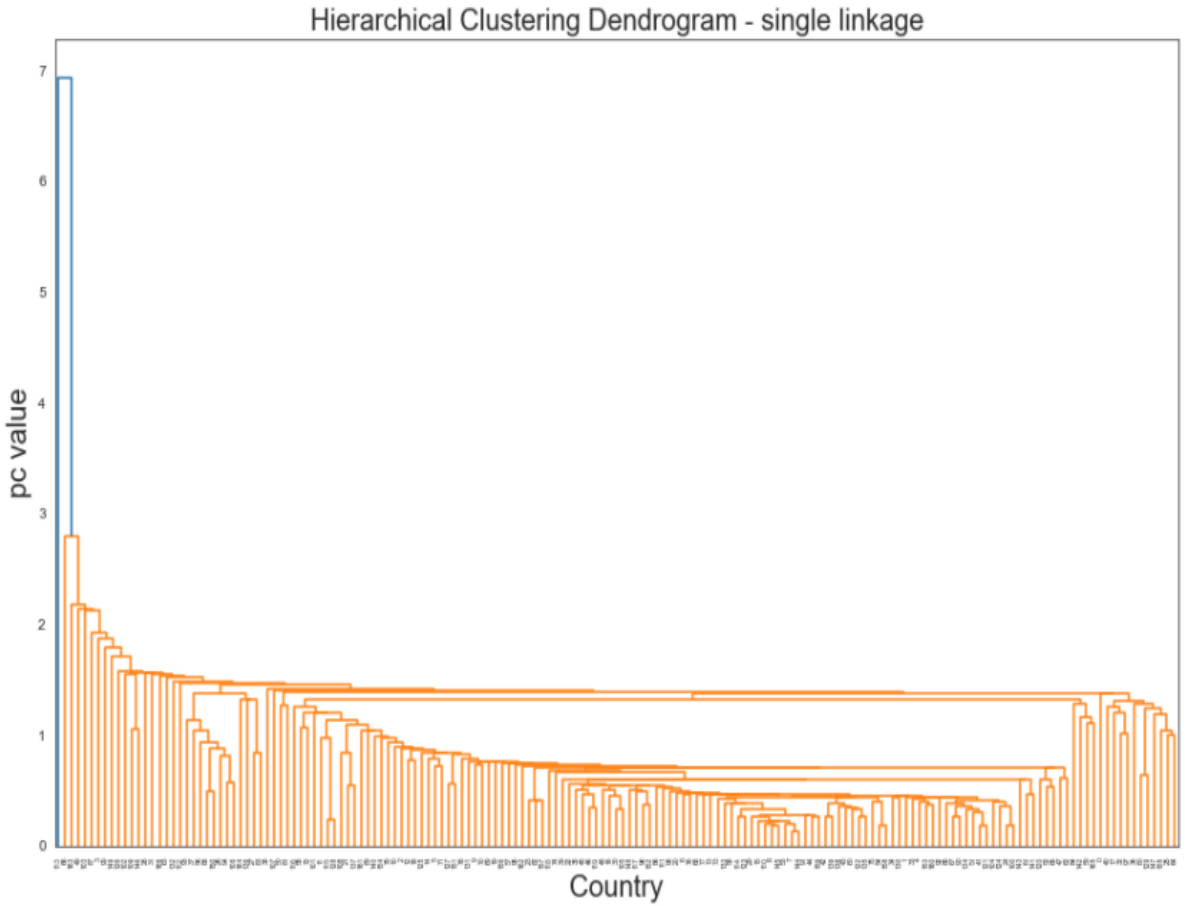
K=4

country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cid_k3	cid_k4
Burundi	93.6	3.861472	1.199254	8.008388	764	12.30	57.7	6.26	231	0	1
Liberia	89.3	4.228753	1.199254	8.008388	700	5.47	60.8	5.02	327	0	1
Congo, Dem. Rep.	116.0	4.228753	1.199254	8.008388	609	20.80	57.5	6.54	334	0	1
Niger	123.0	4.228753	1.199254	8.008388	814	2.55	58.8	7.49	348	0	1
Sierra Leone	160.0	4.210526	1.199254	8.008388	1220	17.20	55.0	5.20	399	0	1
Madagascar	62.2	4.228753	0.912833	8.008388	1390	8.79	60.8	4.60	413	0	1
Mozambique	101.0	4.228753	1.199254	8.008388	918	7.64	54.5	5.56	419	0	1
Central African Republic	149.0	2.645740	0.892377	5.941704	888	2.01	47.5	5.21	446	0	1
Malawi	90.5	4.228753	1.199254	7.603486	1030	12.10	53.1	5.31	459	0	1
Togo	90.3	4.228753	1.199254	8.008388	1210	1.18	58.7	4.87	488	0	1

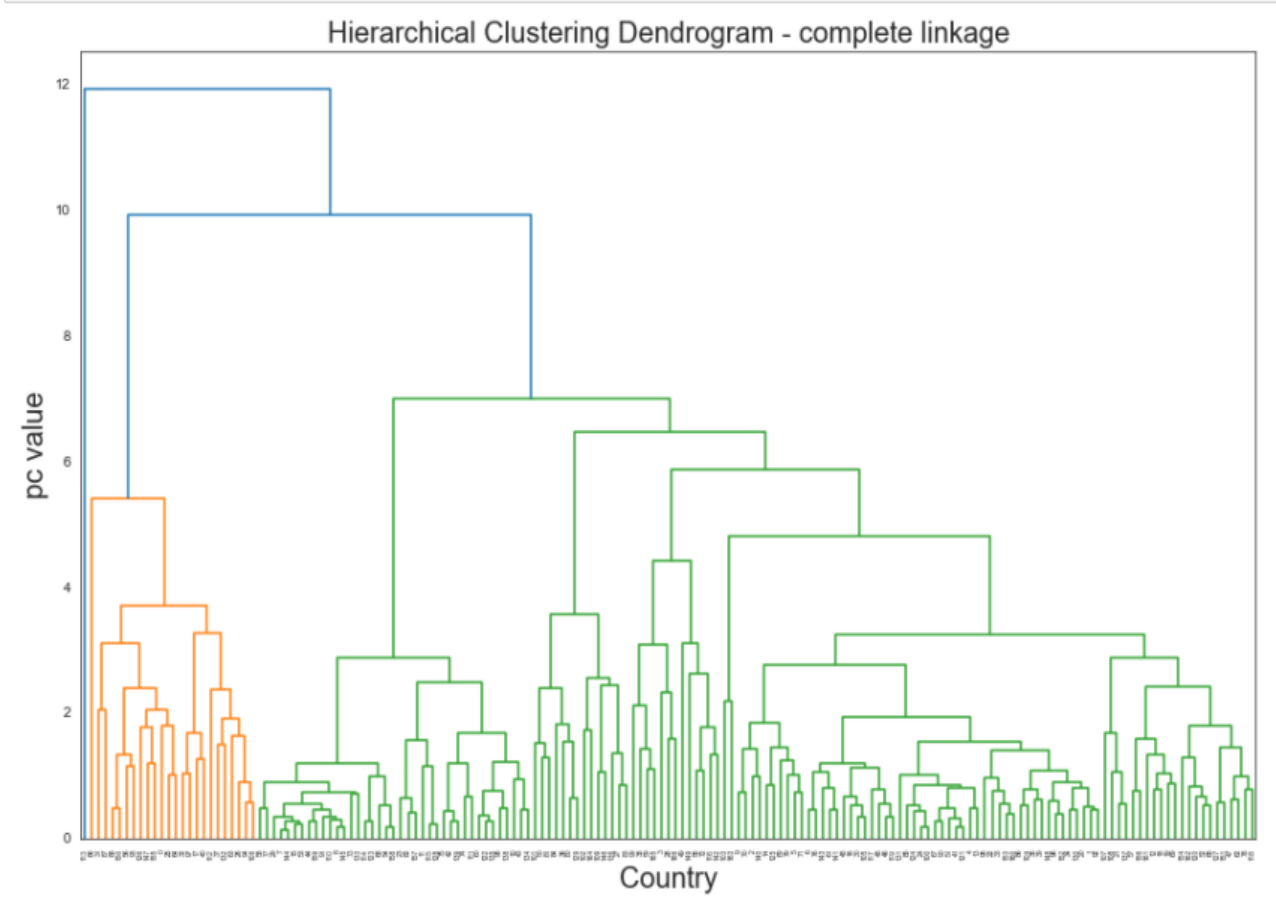
We can see that for both k=3 and k =4, almost all the countries are same in top 10.

HIERARCHICAL CLUSTERING :

- Single Linkage Dendrogram

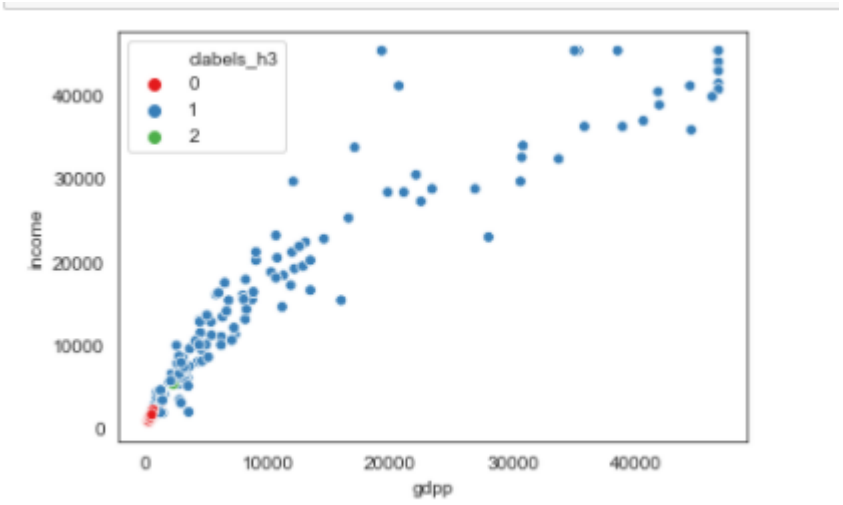


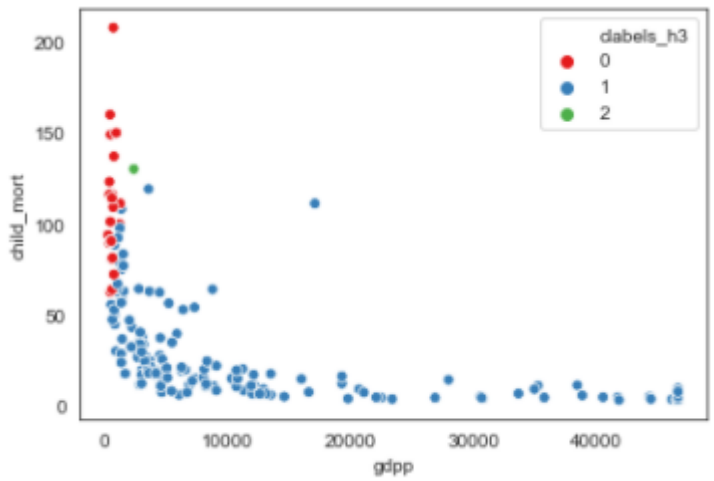
- Complete Linkage Dendrogram

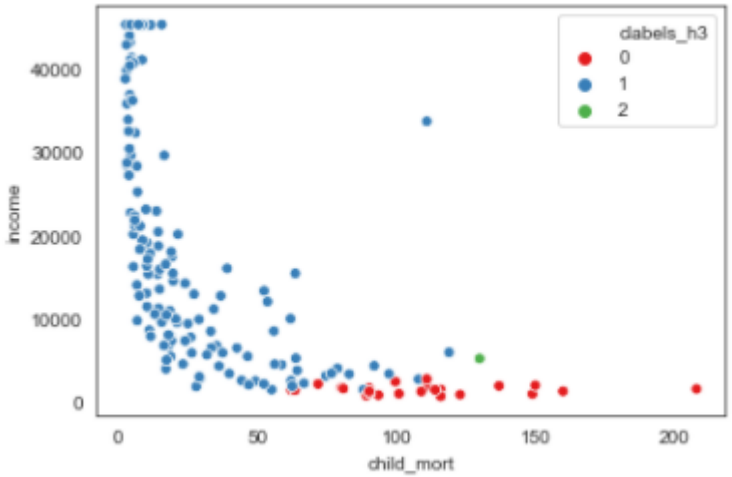


Going with 3 clusters

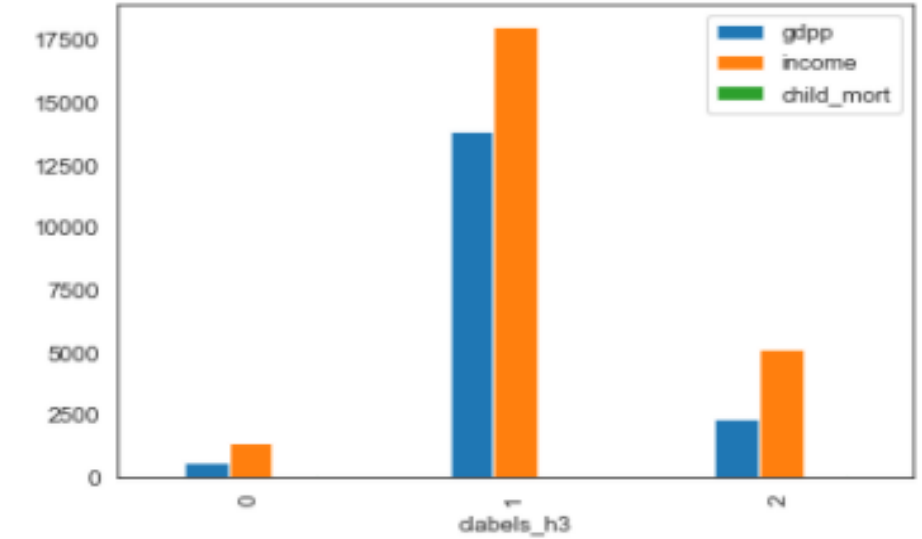
GRAPHS FROM HIERARCHICAL CLUSTERING







- Not the same graphs as we got from k=3 and k=4. Cluster 2 is not that significant



It is clear from here that cluster 0 is the cluster we are looking for as it has low GDP and income and high child_mort.

TOP 10 COUNTRIES WHO NEED AID FROM HIERARCHAL CLUSTERING

country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cid_k3	cid_k4	clabels_h3
Burundi	93.6	3.861472	1.199254	8.008388	764	12.30	57.7	6.26	231	0	1	0
Liberia	89.3	4.228753	1.199254	8.008388	700	5.47	60.8	5.02	327	0	1	0
Congo, Dem. Rep.	116.0	4.228753	1.199254	8.008388	609	20.80	57.5	6.54	334	0	1	0
Niger	123.0	4.228753	1.199254	8.008388	814	2.55	58.8	7.49	348	0	1	0
Sierra Leone	160.0	4.210526	1.199254	8.008388	1220	17.20	55.0	5.20	399	0	1	0
Madagascar	62.2	4.228753	0.912833	8.008388	1390	8.79	60.8	4.60	413	0	1	0
Mozambique	101.0	4.228753	1.199254	8.008388	918	7.64	54.5	5.56	419	0	1	0
Central African Republic	149.0	2.645740	0.892377	5.941704	888	2.01	47.5	5.21	446	0	1	0
Malawi	90.5	4.228753	1.199254	7.603486	1030	12.10	53.1	5.31	459	0	1	0
Togo	90.3	4.228753	1.199254	8.008388	1210	1.18	58.7	4.87	488	0	1	0

THUS, LISTS OF COUNTRIES ARE ALMOST SAME FROM BOTH THE CLUSTERING TECHNIQUE. AFTER COMBINING THE COUNTRIES OF BOTH:

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

THANK YOU