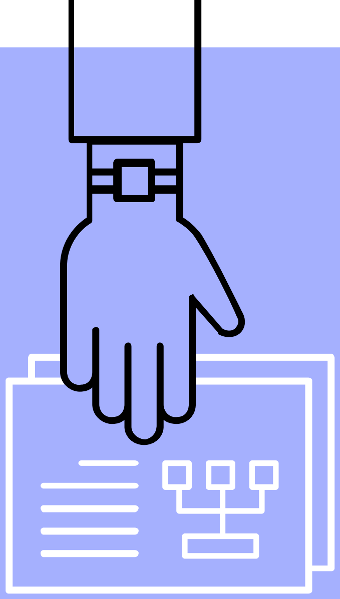
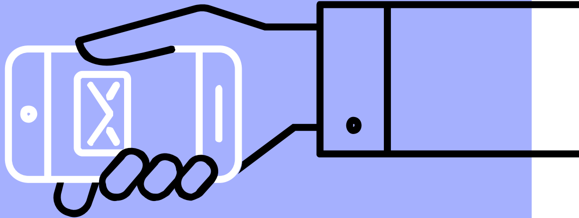
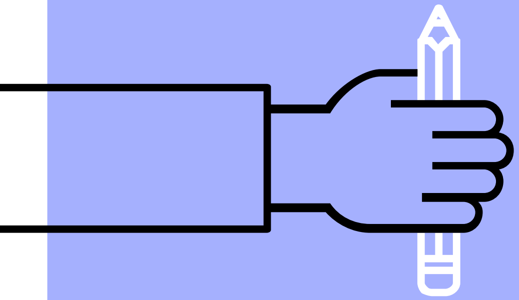
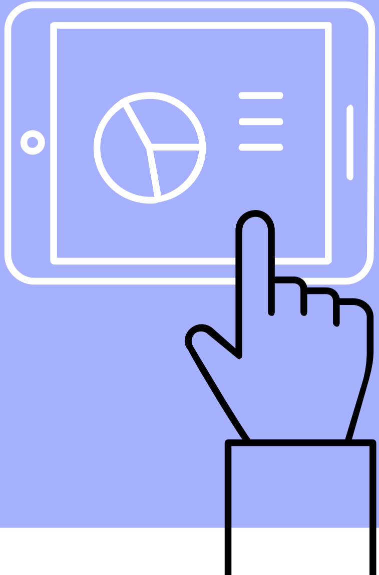




Clustering - The countries needing aid the most

By Ankita Divya





PROBLEM STATEMENT

- This project aims to identify countries which are in need of aid from a HELP NGO the most.

- Using some socio-economic and health factors that determine the overall development of the country we need to categorize such counties using the clustering algorithm.This may be useful for the CEO of the NGO to decide how to use the money strategically and effectively.

- This will ensure that the countries which need the money most will be helped.

Understanding the Data

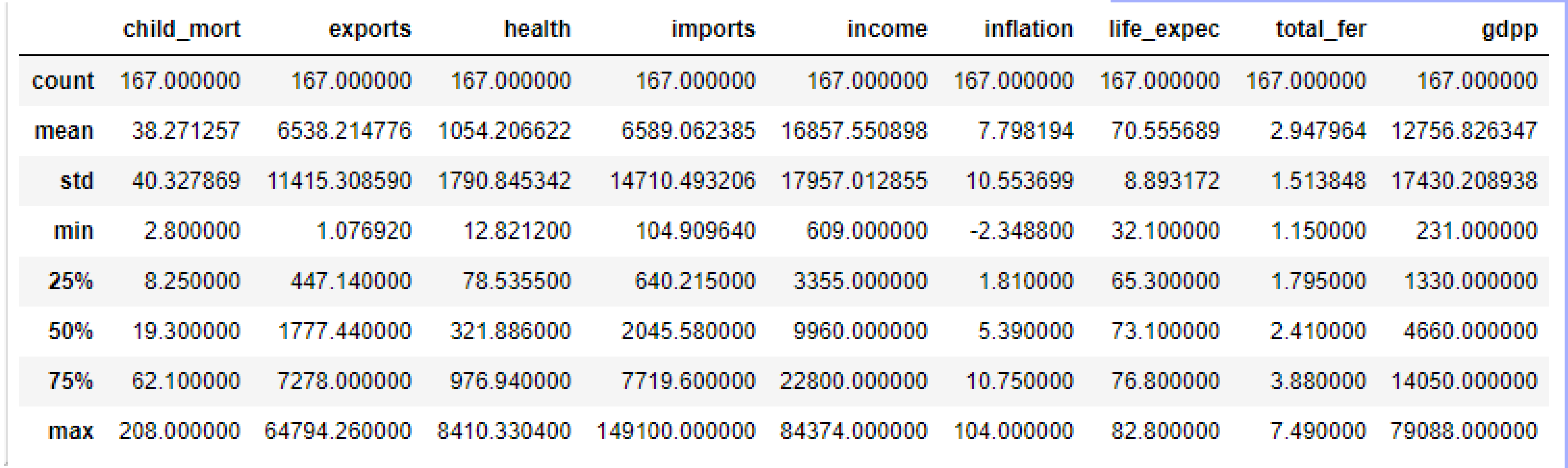
- Understanding the details and feature present in the data.

- As the exports, health and imports are given as %age of the GDP per capita, we converted these to actual values

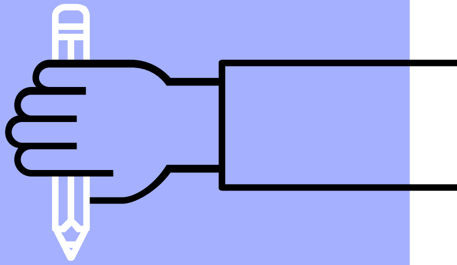
- This will ensure that the countries which need the money most will be helped

Understanding the Data

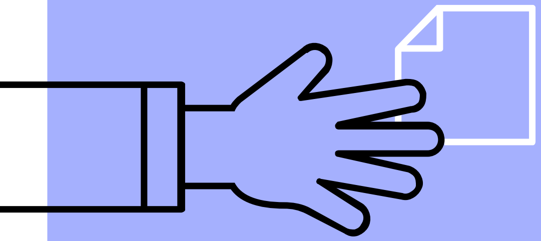
- Understanding the details and different features present in the data.
- As the exports, health and imports are given as %age of the GDP per capita, we converted these to actual values.
- Below is the summary of the data.

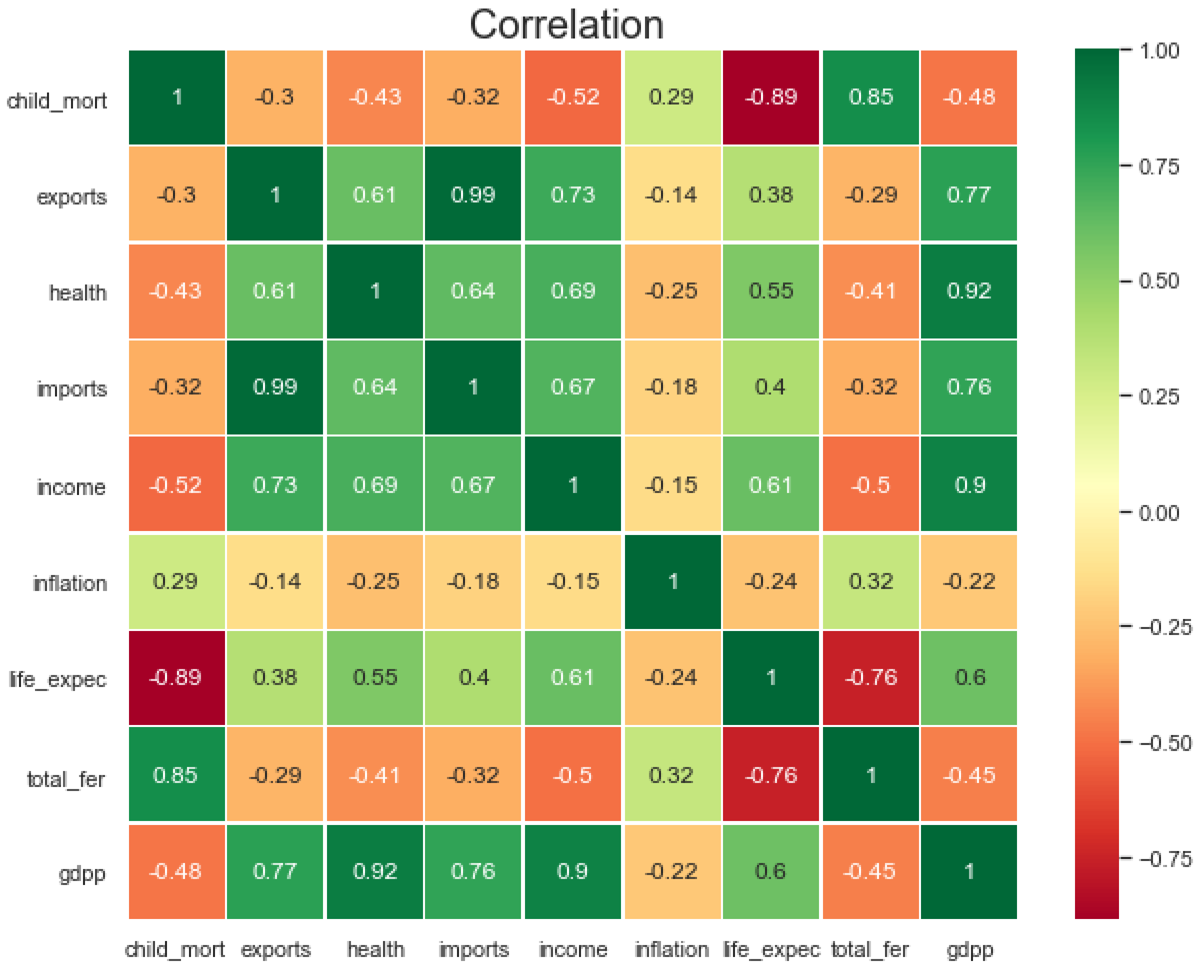


	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
count	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000
mean	38.271257	6538.214776	1054.206622	6589.062385	16857.550898	7.798194	70.555689	2.947964	12756.826347
std	40.327869	11415.308590	1790.845342	14710.493206	17957.012855	10.553699	8.893172	1.513848	17430.208938
min	2.800000	1.076920	12.821200	104.909640	609.000000	-2.348800	32.100000	1.150000	231.000000
25%	8.250000	447.140000	78.535500	640.215000	3355.000000	1.810000	65.300000	1.795000	1330.000000
50%	19.300000	1777.440000	321.886000	2045.580000	9960.000000	5.390000	73.100000	2.410000	4660.000000
75%	62.100000	7278.000000	976.940000	7719.600000	22800.000000	10.750000	76.800000	3.880000	14050.000000
max	208.000000	64794.260000	8410.330400	149100.000000	84374.000000	104.000000	82.800000	7.490000	79088.000000



EXPLORING CORRELATIONS

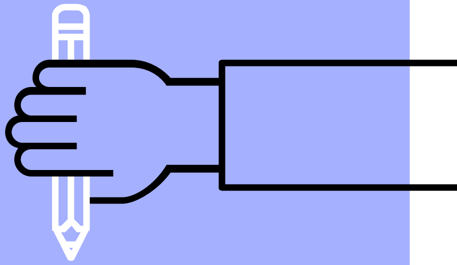




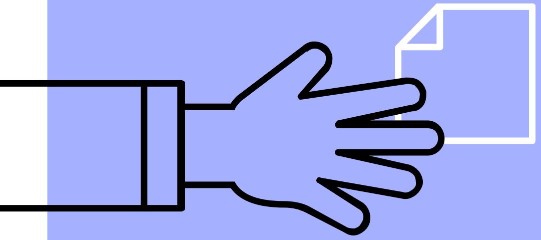
	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
child_mort	1	-0.3	-0.43	-0.32	-0.52	0.29	-0.89	0.85	-0.48
exports	-0.3	1	0.61	0.99	0.73	-0.14	0.38	-0.29	0.77
health	-0.43	0.61	1	0.64	0.69	-0.25	0.55	-0.41	0.92
imports	-0.32	0.99	0.64	1	0.67	-0.18	0.4	-0.32	0.76
income	-0.52	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

Heatmap of the dataframe

- Import and Export are highly positively correlated features with 0.99 while life expectancy and child mortality are highly negatively correlated features.



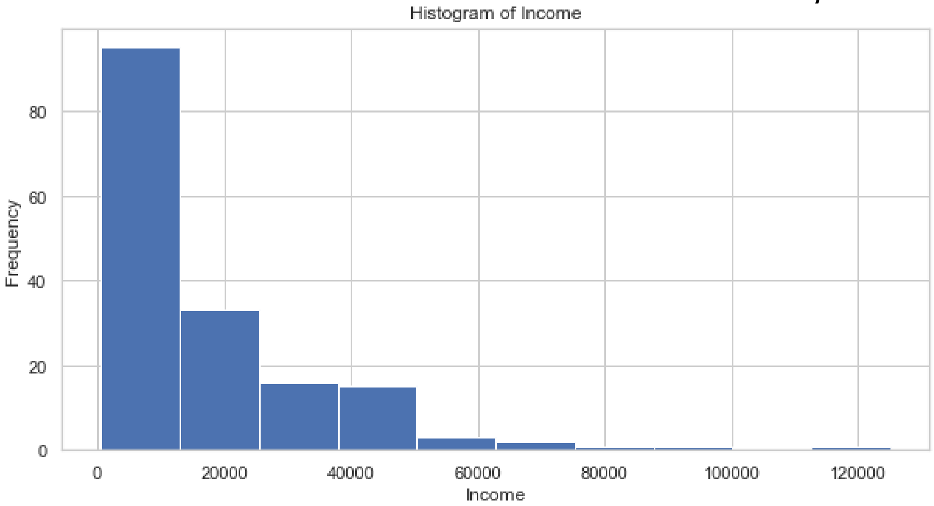
UNIVARIATE ANALYSIS OF NUMERICAL VARIABLES

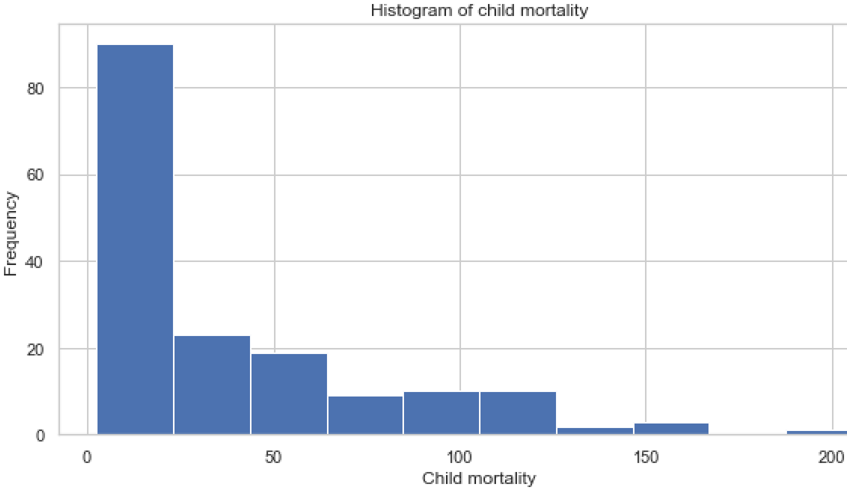


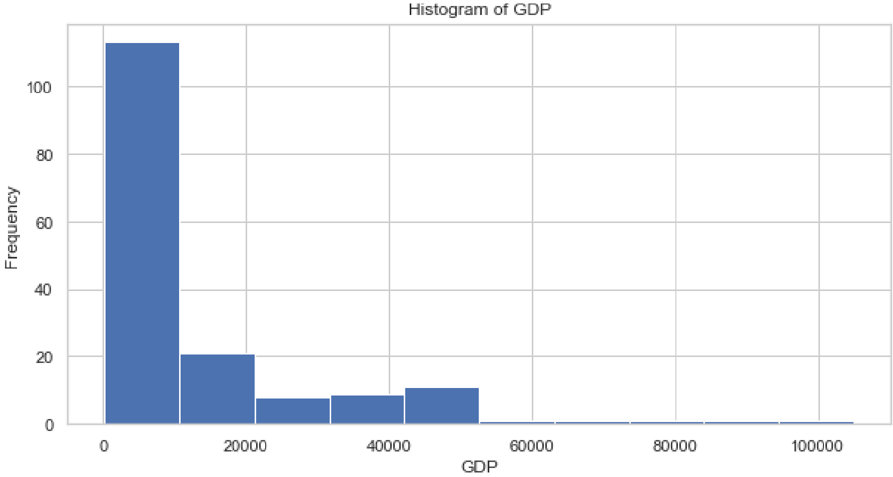
Variables considered:

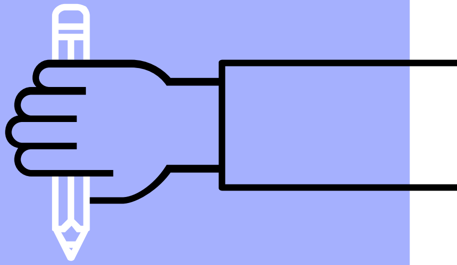
- Child Mortality
- Income
- GDP

DISTRIBUTION OF INCOME , CHILD MORTALITY AND GDP





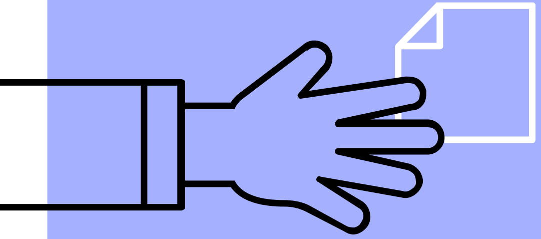


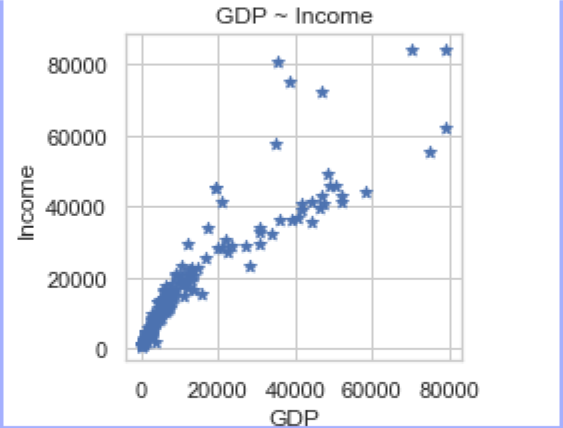


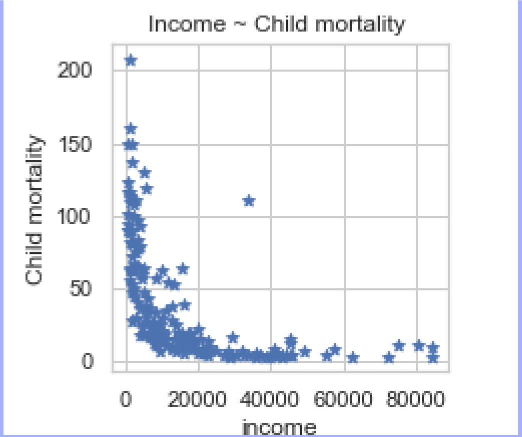
BIVARIATE ANALYSIS

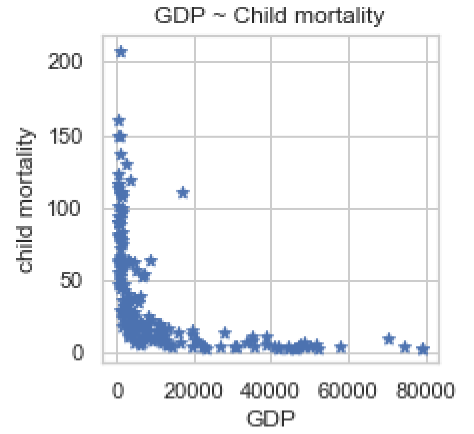
Between:

- GDP and Income
- Income and child mortality
- GDP and Child mortality







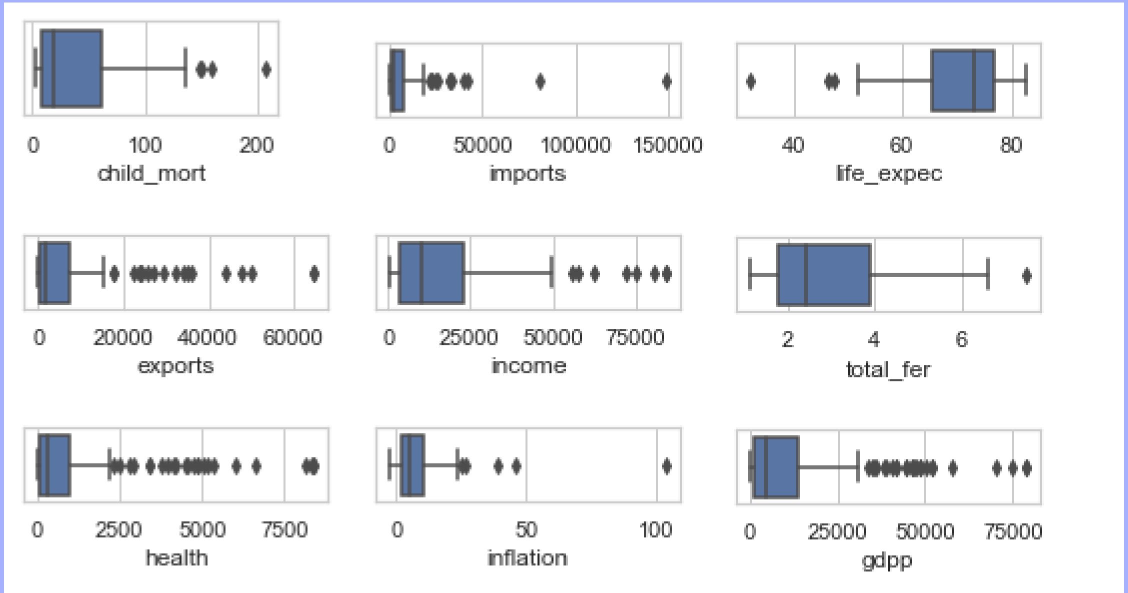


- ➔ Income and GDP are positively related, we can observe from the graph that as GDP increases Income also increases

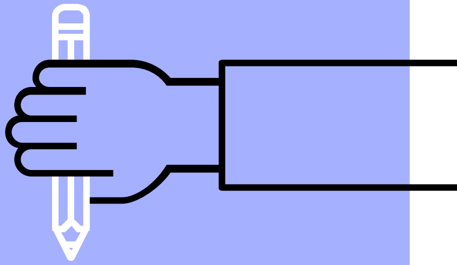
- ➔ We could observe from the above plot between Income ~ child mortality, for low income the child mortality is high and as income increase child mortality is also very less

- ➔ We could observe from the above plot of GDP and child mortality that for low GDP the child mortality is high and as GDP increase child mortality is also very less

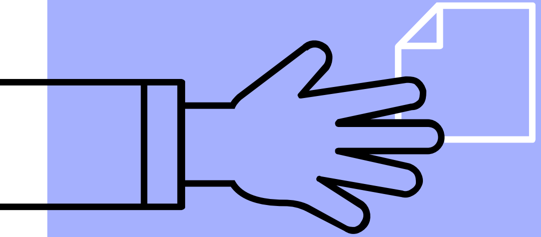
Checking outliers for different numeric columns



After capping lower values of child mortality, inflation and import. Also capping higher values of export, income, health and GDP. We could see all the continuous variables have outliers.



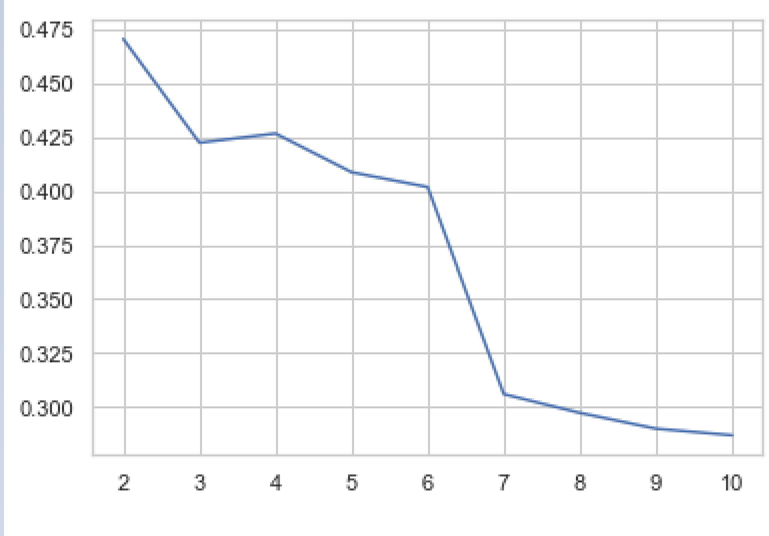
K-means Clustering



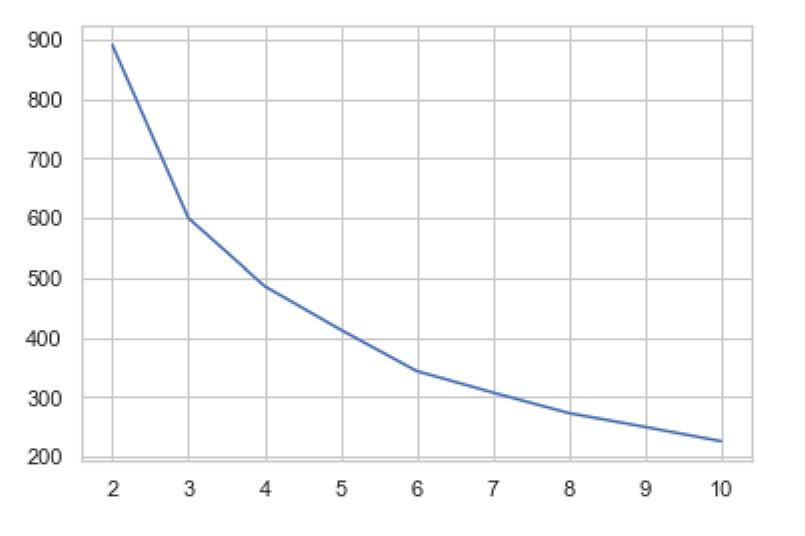
Variables considered for cluster profiling:

- Child Mortality
- Income
- GDP

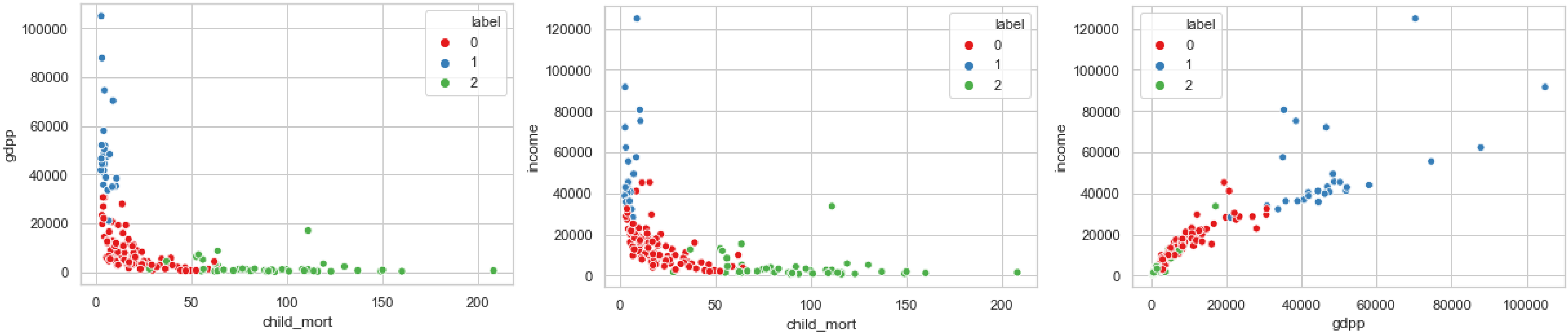
Silhouette Score



Elbow-Curve



- We concluded that the Optimum number of cluster here could be 3. As we can see a change of slope from steep to shallow (an elbow) at 3, we can determine that the optimal number of clusters will be 3 here.
- Also, using the average silhouette method which computes the average silhouette of observations for different values of k. Even though the average silhouette is maximum at k = 2, but for k = 3 also the score is quite high. So the optimal number of clusters is k = 3.

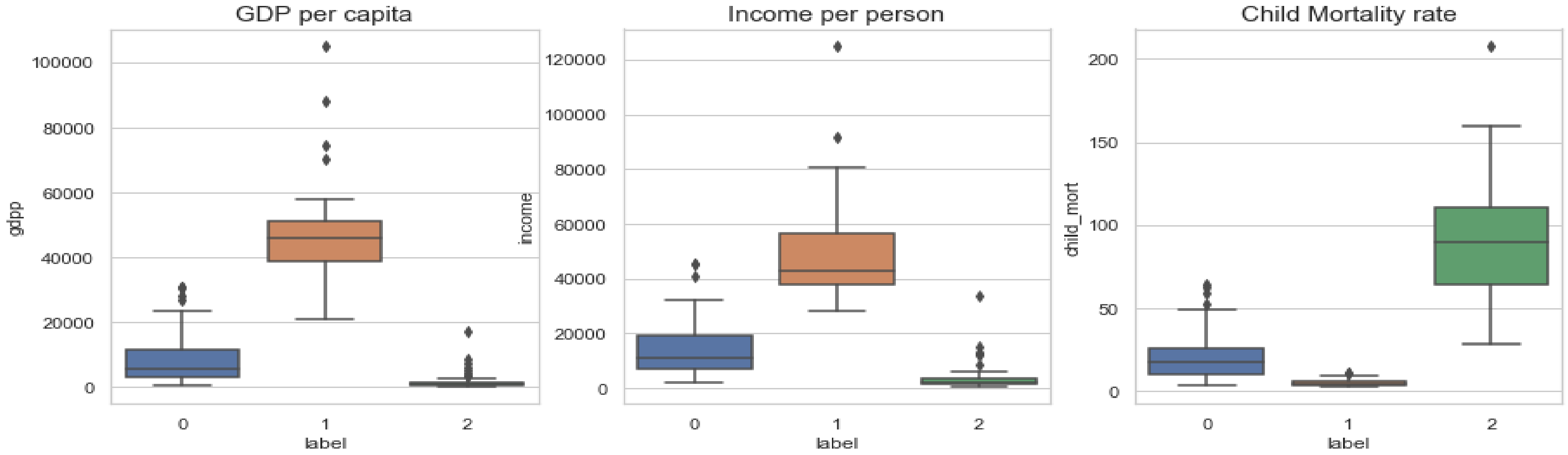


Scatter Plots of the clusters formed

- The three clusters formed for Low child mortality and high GDP, average child mortality and average GDP , high child mortality and low GDP.

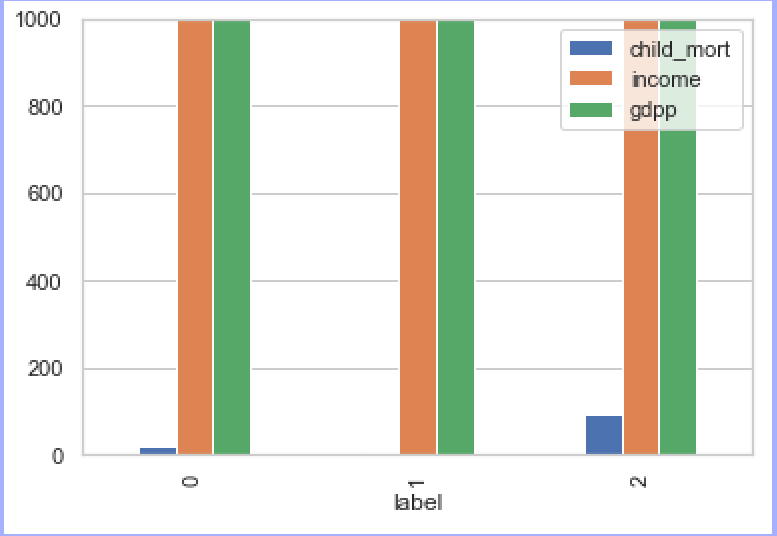
- The three clusters formed for Low child mortality and high income, average child mortality and average income, high child mortality and low income.

- The three clusters formed for high income and high GDP, average income and average GDP ,low income and low GDP.



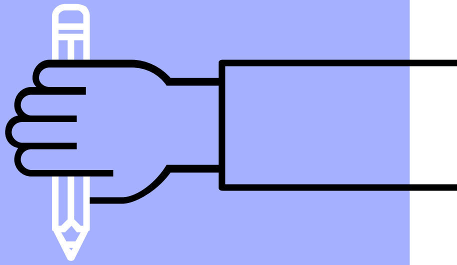
Box Plots of the clusters formed

- Cluster 0: Average GDP, Average Income and Average child mortality
- Cluster 1: High GDP, High Income and Low child mortality
- Cluster 2: Low GDP, Low Income and High child mortality

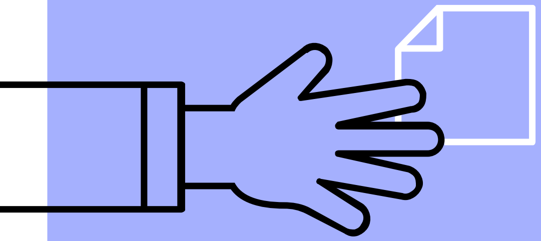


Cluster Profiling

- Cluster 0: Average GDP, Average Income and Average child mortality
- Cluster 1: High GDP, High Income and Low child mortality
- Cluster 2: Low GDP, Low Income and High child mortality
- The cluster which will need aid the most will be the cluster 2

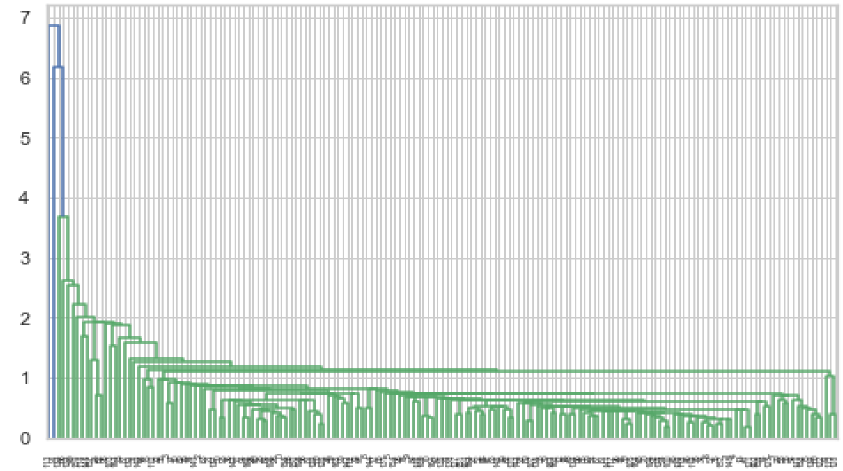


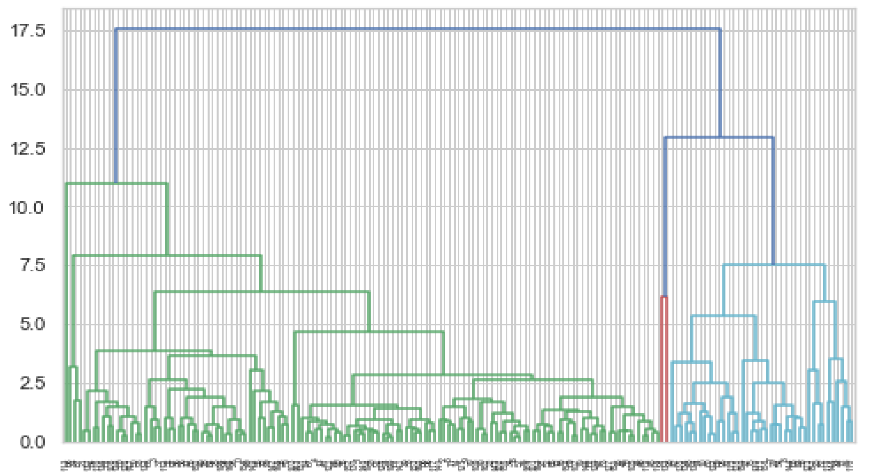
Hierarchical Clustering



Variables considered for cluster profiling:

- Child Mortality
- Income
- GDP

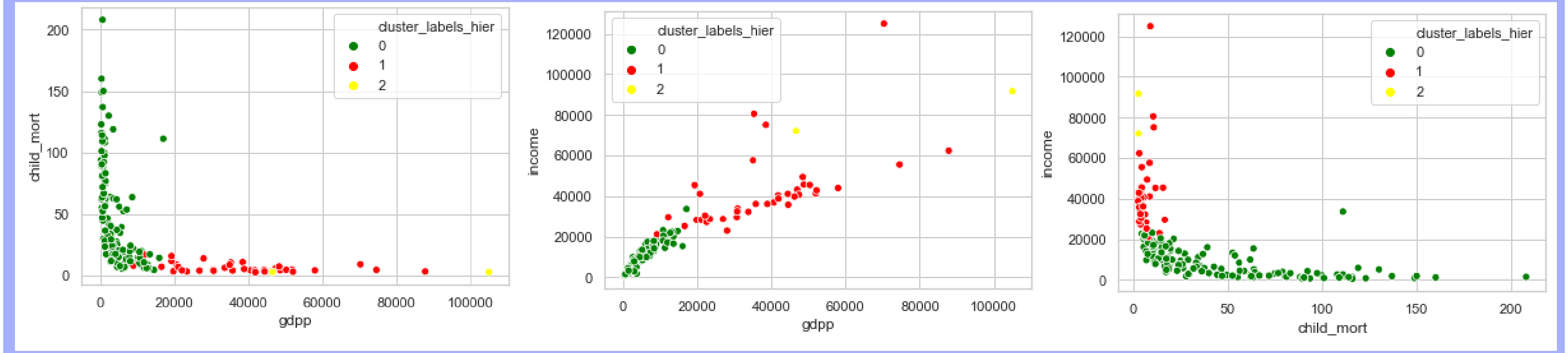




Single Linkage

Complete Linkage

- We can see 3 prominent clusters in complete linkage which are in green, red, and sky-blue.
- Cutting the dendrogram vertically such that n_clusters = 3

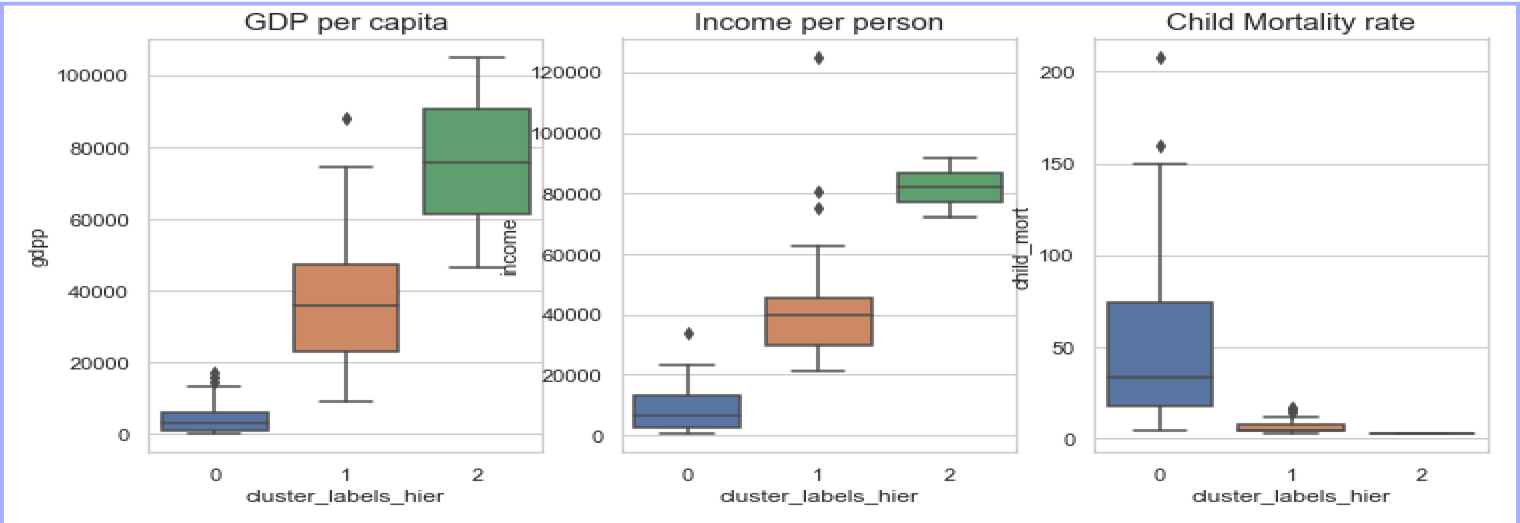


Scatter Plots formed showing three distinct clusters

- The three clusters formed for Low child mortality and high GDP, average child mortality and average GDP , high child mortality and low GDP.

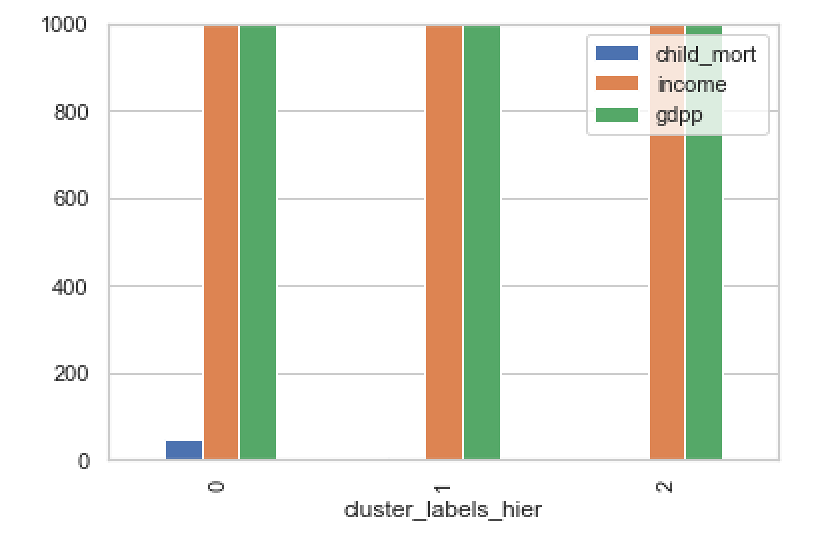
- The three clusters formed for Low child mortality and high income, average child mortality and average income, high child mortality and low income.

- The three clusters formed for high income and high GDP, average income and average GDP ,low income and low GDP.



Box Plots of the clusters formed

- Cluster 0: Low GDP, Low Income and High child mortality
- Cluster 1: Average GDP, Average Income and Average child mortality
- Cluster 2: High GDP, High Income and Low child mortality



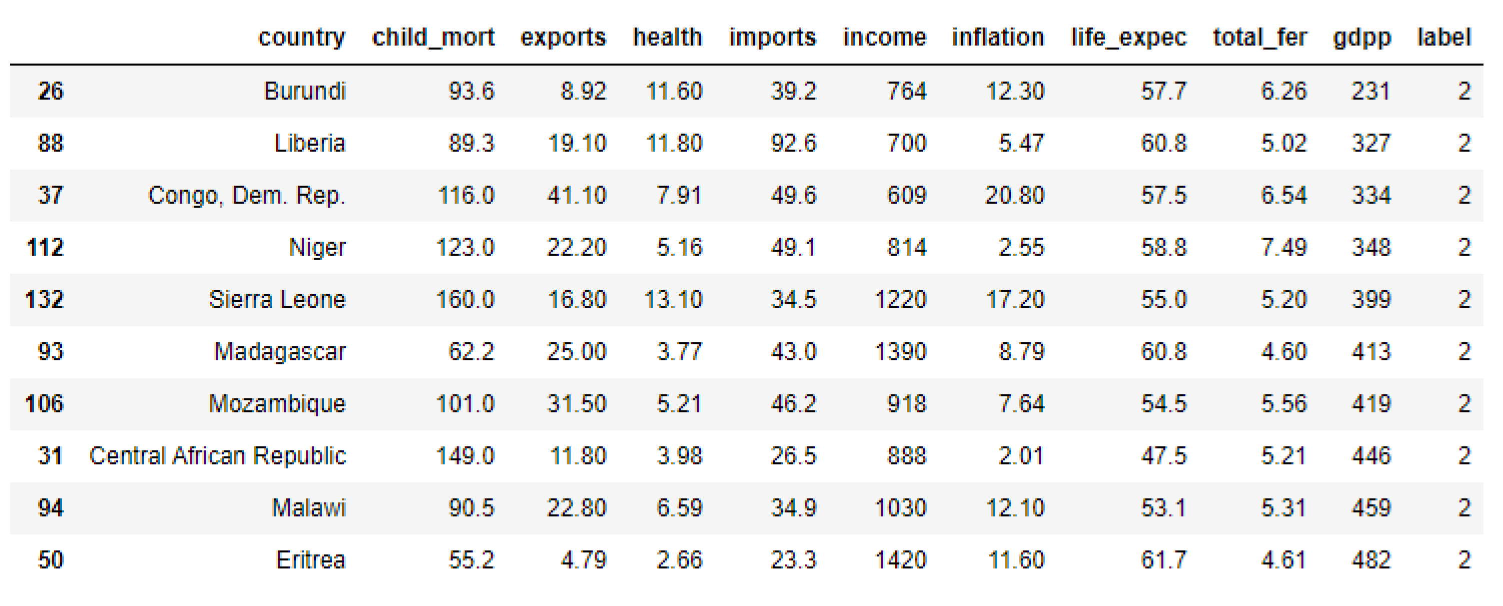
Cluster Profiling

- Cluster 0: High GDP, High Income and Low child mortality

- Cluster 1: Average GDP, Average Income and Average child mortality

- Cluster 2: Low GDP, Low Income and High child mortality

- The cluster which will need aid the most will be the cluster 0



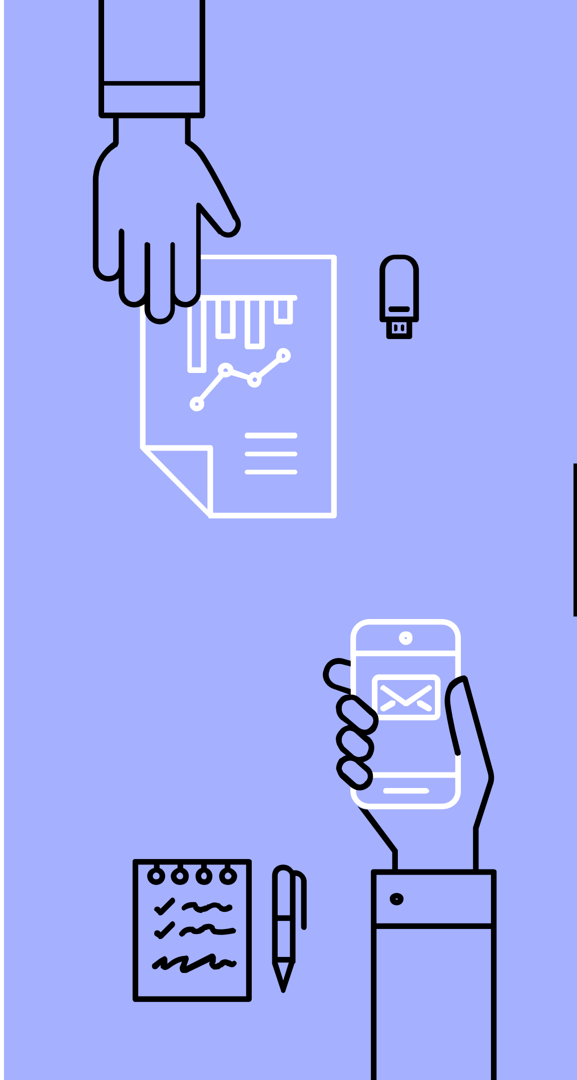
	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	label
26	Burundi	93.6	8.92	11.60	39.2	764	12.30	57.7	6.26	231	2
88	Liberia	89.3	19.10	11.80	92.6	700	5.47	60.8	5.02	327	2
37	Congo, Dem. Rep.	116.0	41.10	7.91	49.6	609	20.80	57.5	6.54	334	2
112	Niger	123.0	22.20	5.16	49.1	814	2.55	58.8	7.49	348	2
132	Sierra Leone	160.0	16.80	13.10	34.5	1220	17.20	55.0	5.20	399	2
93	Madagascar	62.2	25.00	3.77	43.0	1390	8.79	60.8	4.60	413	2
106	Mozambique	101.0	31.50	5.21	46.2	918	7.64	54.5	5.56	419	2
31	Central African Republic	149.0	11.80	3.98	26.5	888	2.01	47.5	5.21	446	2
94	Malawi	90.5	22.80	6.59	34.9	1030	12.10	53.1	5.31	459	2
50	Eritrea	55.2	4.79	2.66	23.3	1420	11.60	61.7	4.61	482	2

- The countries which CEO should be focussing based on low GDP, low income and high child mortality are mentioned above

Conclusion

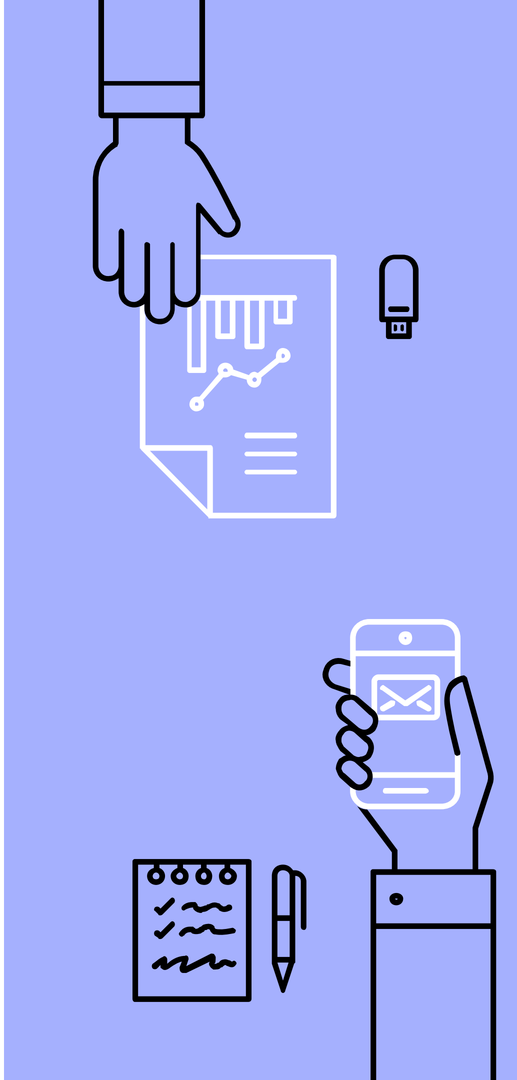
- Using the K-means and hierarchical clustering, the countries which need the aid are most decided on the factors like high child mortality, income and GDP. Both clustering gave the similar result.

- After comparing both the K-means and hierarchical clustering algorithms, based on the clusters formed and clarity in the plots we can conclude that K-means is having relative balanced no of countries in all clusters. Hence we can consider K-means as final approach.



Conclusion

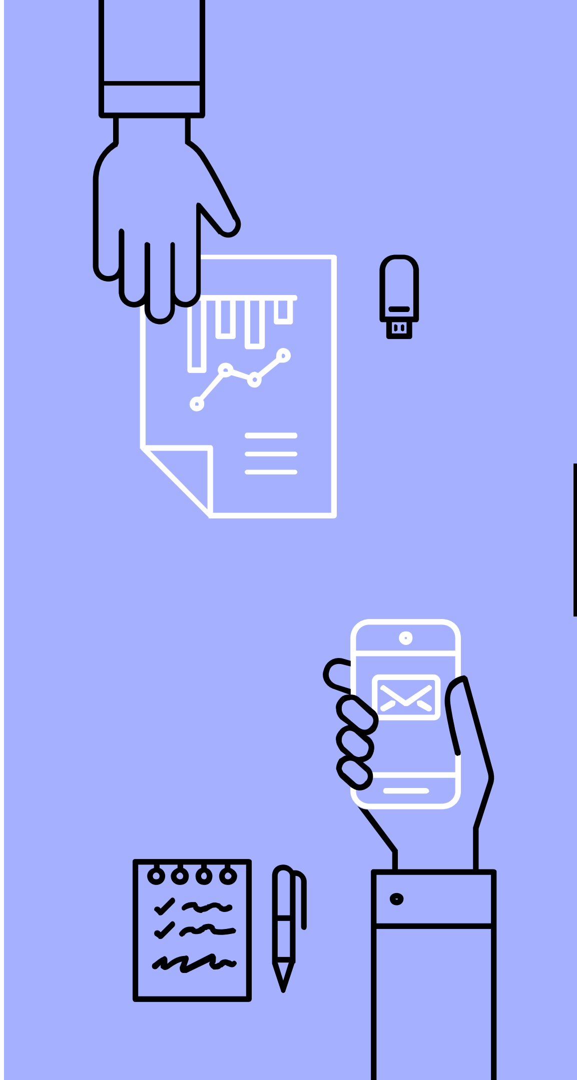
- Companies available in K-mean clustering in cluster 2 and in hierarchical clustering to cluster 0 are the countries which needs aid as it has lowest GDP, lowest income and highest child mortality. As in both the methods, the countries clustered for underdeveloped countries was almost same. i.e. deciding no. of clusters as 3 was profitable.

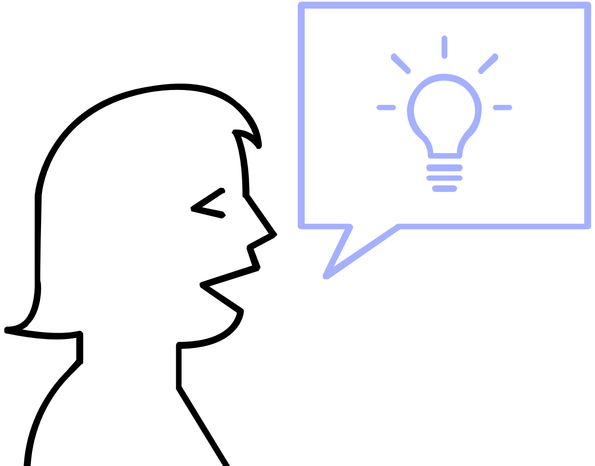


Recommendation

The countries which CEO should be focussing based on low GDP, low income and high child mortality on are as follows:-

- Burundi
- Liberia
- Congo, Dem. Rep.
- Niger
- Sierra Leone
- Madagascar





CONCLUSION...

- Using the K-means and hierarchical clustering, the countries which need the aid are most decided on the factors like high child mortality, income and GDP. Both clustering gave the similar result.

- After comparing both the K-means and hierarchical clustering algorithms, based on the clusters formed and clarity in the plots we can conclude that K-means is having relative balanced no of countries in all clusters. Hence we can consider K-means as final approach.

- Companies available in K-mean clustering in cluster 2 and in hierarchical clustering to cluster 0 are the countries which needs aid as it has lowest GDP, lowest income and highest child mortality. As in both the methods, the countries clustered for underdeveloped countries was almost same. i.e. deciding no. of clusters as 3 was profitable.

THANKS!