HELP International

Clustering Assignment

Ankur sugandhi

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 programmes, 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.
- And this is where you come in as a data analyst. Your job is to categorise the countries using some socio-economic and health factors that determine the overall development of the country. Then you need to suggest the countries which the CEO needs to focus on the most. The datasets containing those socio-economic factors and the corresponding data dictionary are provided below.

Objectives

- Start off with the necessary data inspection and EDA tasks suitable for this dataset - data cleaning, univariate analysis, bivariate analysis etc.
- **Outlier Analysis**: You must perform the Outlier Analysis on the dataset. However, you do have the flexibility of not removing the outliers if it suits the business needs or a lot of countries are getting removed. Hence, all you need to do is find the outliers in the dataset, and then choose whether to keep them or remove them depending on the results you get.
- Try both K-means and Hierarchical clustering(both single and complete linkage) on this dataset to create the clusters. [Note that both the methods may not produce identical results and you might have to choose one of them for the final list of countries.]
- Analyse the clusters and identify the ones which are in dire need of aid. You can analyse the clusters by comparing how these three variables - [**gdpp**, **child_mort** and **income**] vary for each cluster of countries to recognise and differentiate the clusters of developed countries from the clusters of under-developed countries.
- Also, you need to perform visualisations on the clusters that have been formed. You can do this by choosing any two of the three variables mentioned above on the X-Y axes and plotting a scatter plot of all the countries and differentiating the clusters. Make sure you create visualisations for all the three pairs. You can also choose other types of plots like boxplots, etc.
- Both K-means and Hierarchical may give different results because of previous analysis (whether you chose to keep or remove the outliers, how many clusters you chose, etc.) Hence, there might be some subjectivity in the final number of countries that you think should be reported back to the CEO since they depend upon the preceding analysis as well. Here, make sure that you report back at least 5 countries which are in direst need of aid from the analysis work that you perform.

Exploratory Data Analysis

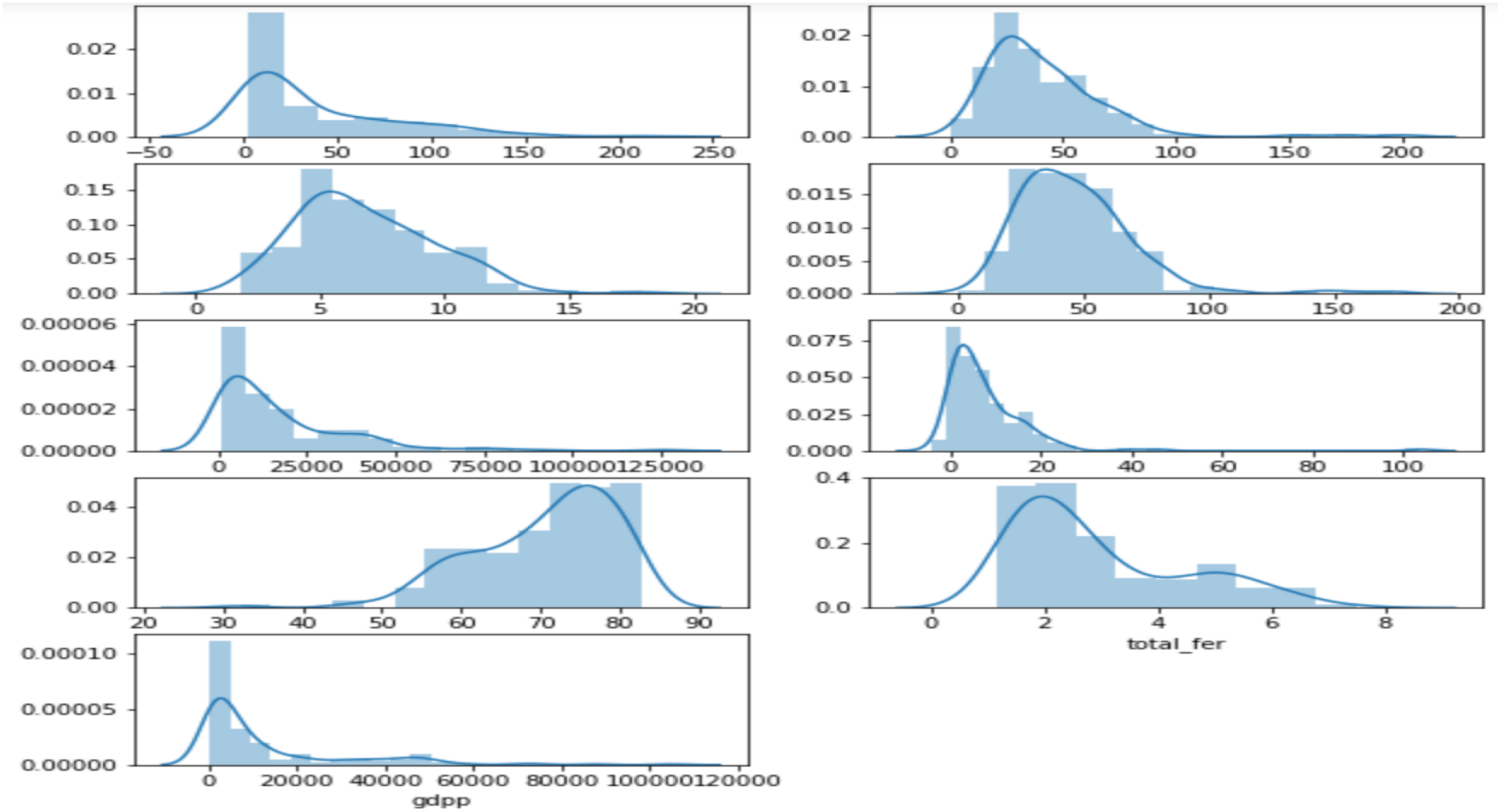
Importing Data

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	Afghanistan	90.2	10.0	7.58	44.9	1610	9.44	56.2	5.82	553
1	Albania	16.6	28.0	6.55	48.6	9930	4.49	76.3	1.65	4090
2	Algeria	27.3	38.4	4.17	31.4	12900	16.10	76.5	2.89	4460
3	Angola	119.0	62.3	2.85	42.9	5900	22.40	60.1	6.16	3530
4	Antigua and Barbuda	10.3	45.5	6.03	58.9	19100	1.44	76.8	2.13	12200

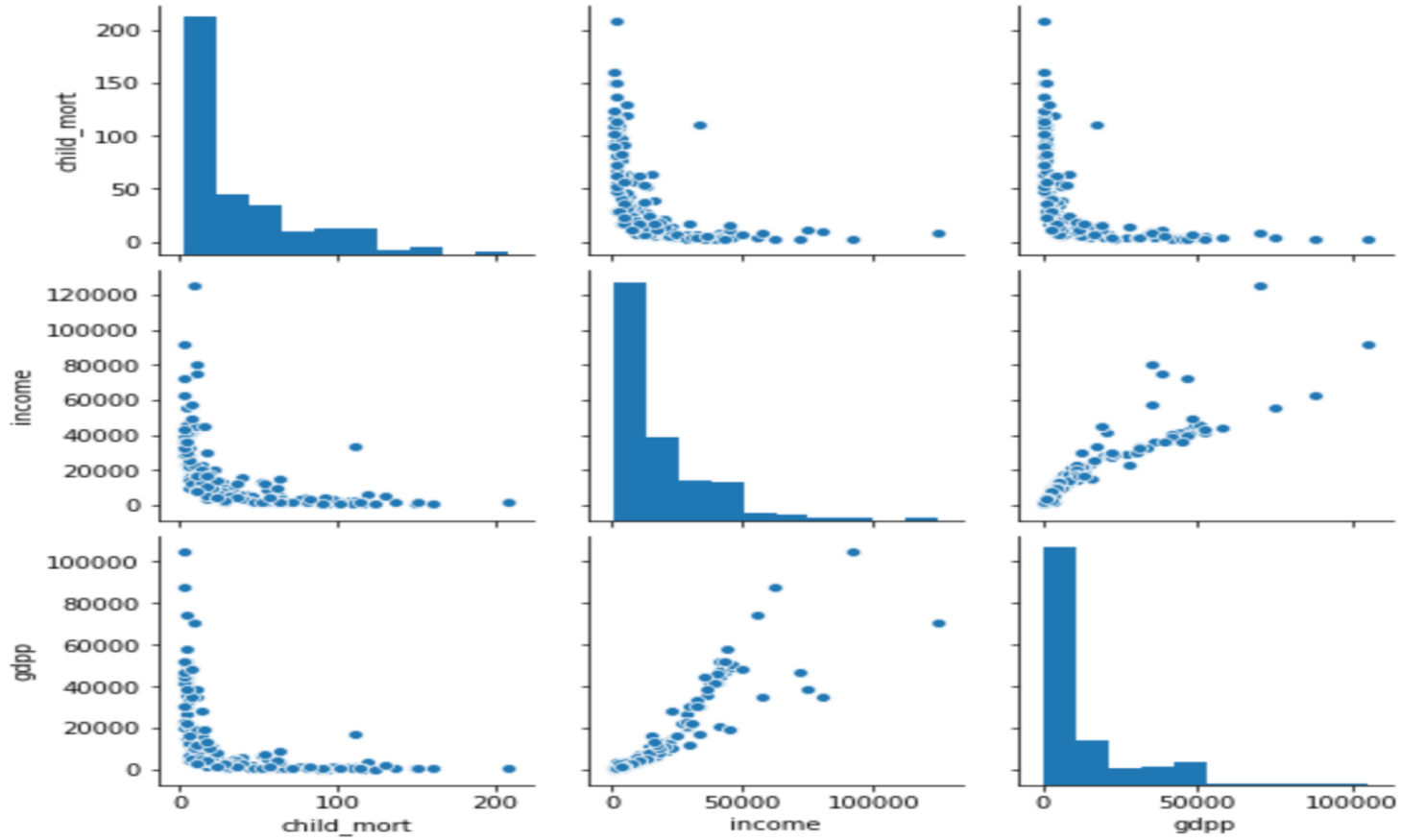
Data Summary

	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	38.270060	41.108976	6.815689	46.890215	17144.688623	7.781832	70.555689	2.947964	12964.155689
std	40.328931	27.412010	2.746837	24.209589	19278.067698	10.570704	8.893172	1.513848	18328.704809
min	2.600000	0.109000	1.810000	0.065900	609.000000	-4.210000	32.100000	1.150000	231.000000
25%	8.250000	23.800000	4.920000	30.200000	3355.000000	1.810000	65.300000	1.795000	1330.000000
50%	19.300000	35.000000	6.320000	43.300000	9960.000000	5.390000	73.100000	2.410000	4660.000000
75%	62.100000	51.350000	8.600000	58.750000	22800.000000	10.750000	76.800000	3.880000	14050.000000
max	208.000000	200.000000	17.900000	174.000000	125000.000000	104.000000	82.800000	7.490000	105000.000000

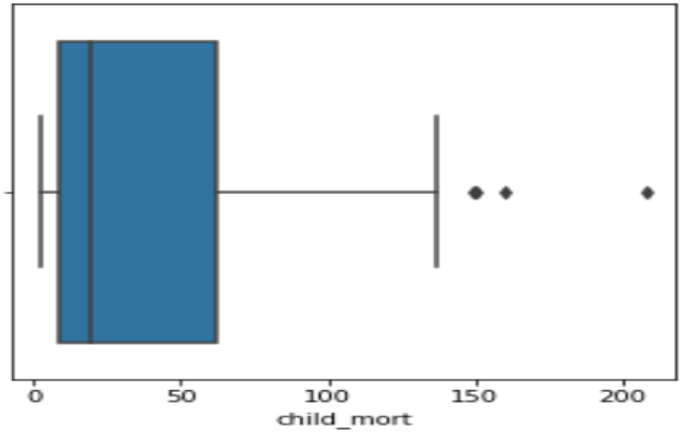
Graphical representation

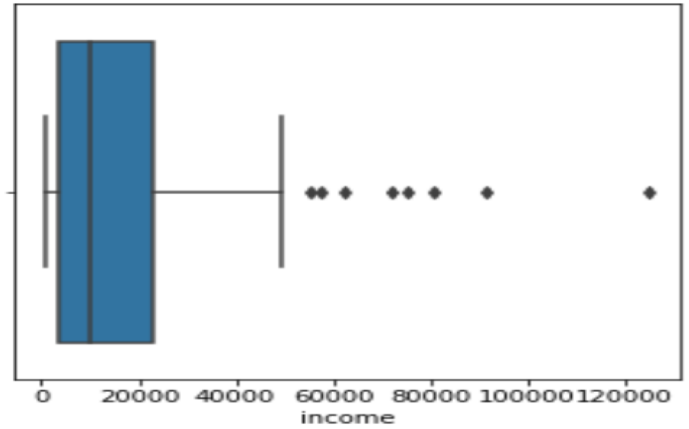


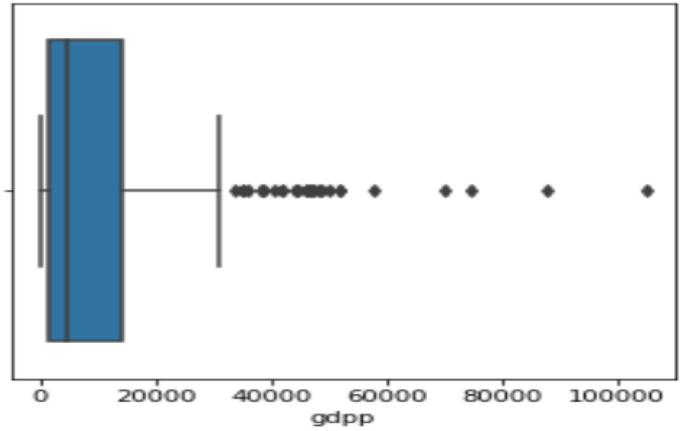
Child_mort, income, gdpp correlation



Outliers


0
50
100
150
200
child_mort


0
20000
40000
60000
80000
100000
120000
income


0
20000
40000
60000
80000
100000
gdpp

Standard scaling

	child_mort	income	gdpp
0	1.291532	-0.851668	-0.702259
1	-0.538949	-0.386946	-0.498726
2	-0.272833	-0.221053	-0.477434
3	2.007808	-0.612045	-0.530950
4	-0.695634	0.125254	-0.032042

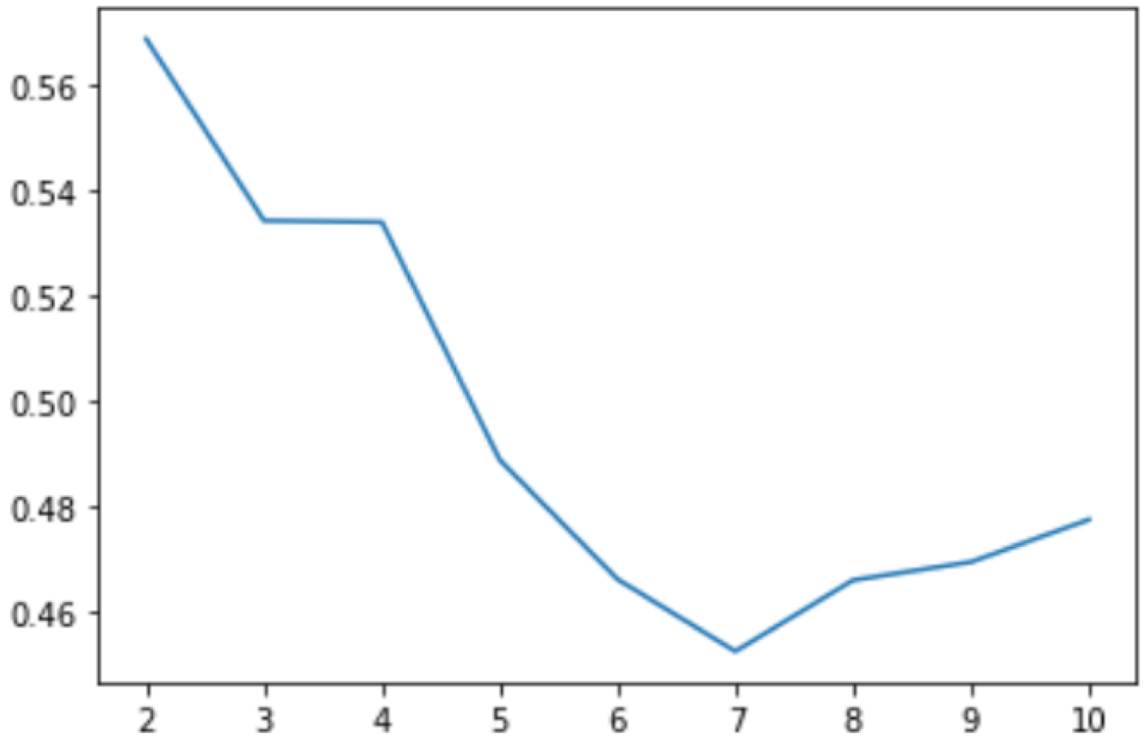
Hopkins

- Range is between 87% to 95%

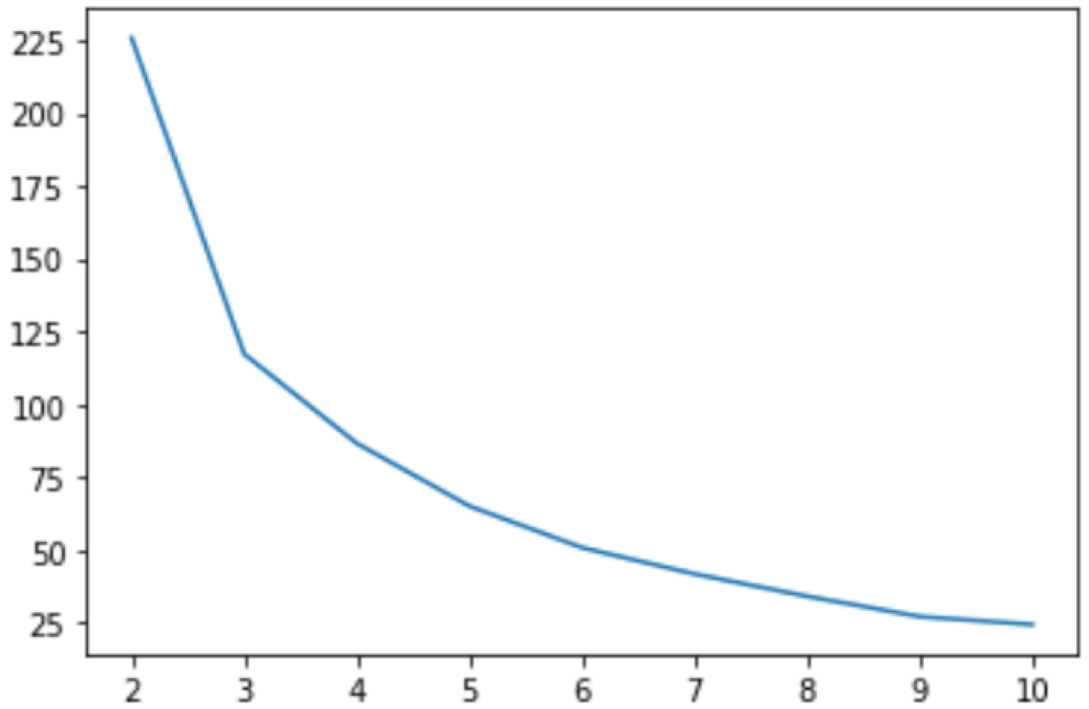
```
In [160]: hopkins(df1.drop('country', axis = 1))

Out[160]: 0.945549028837805
```

K-means Outcome

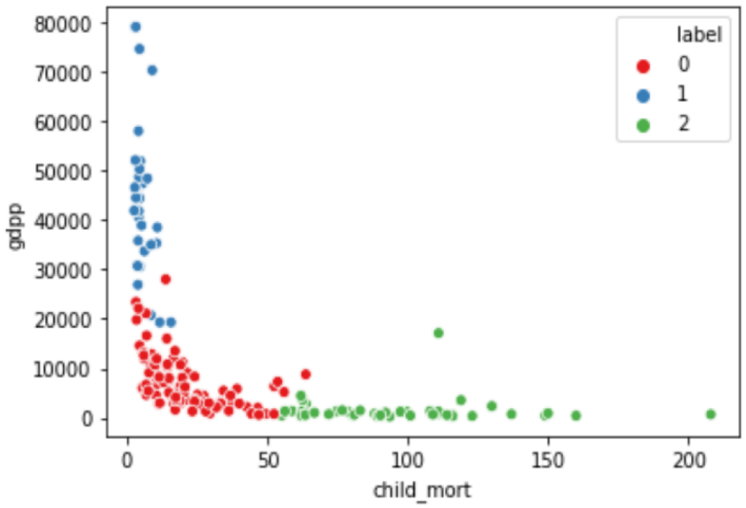
Silhouette score

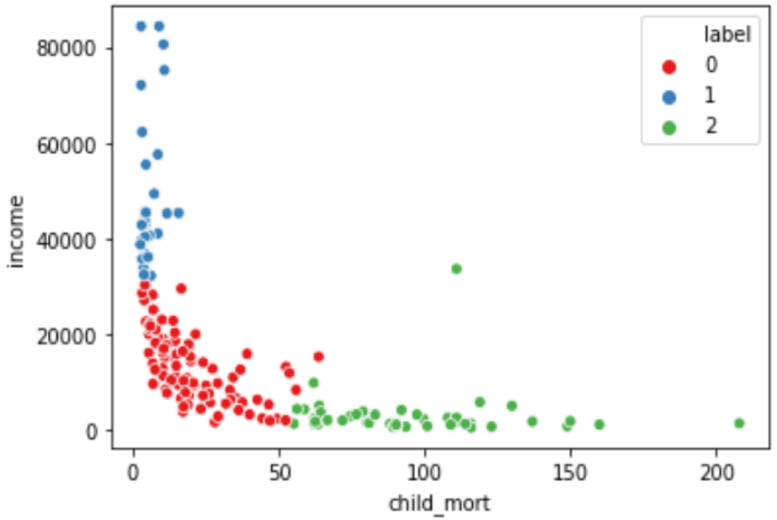


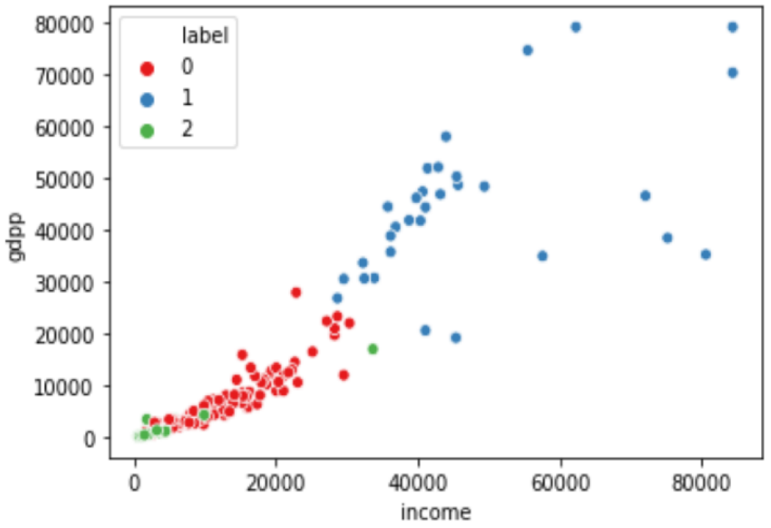
Elbow curve



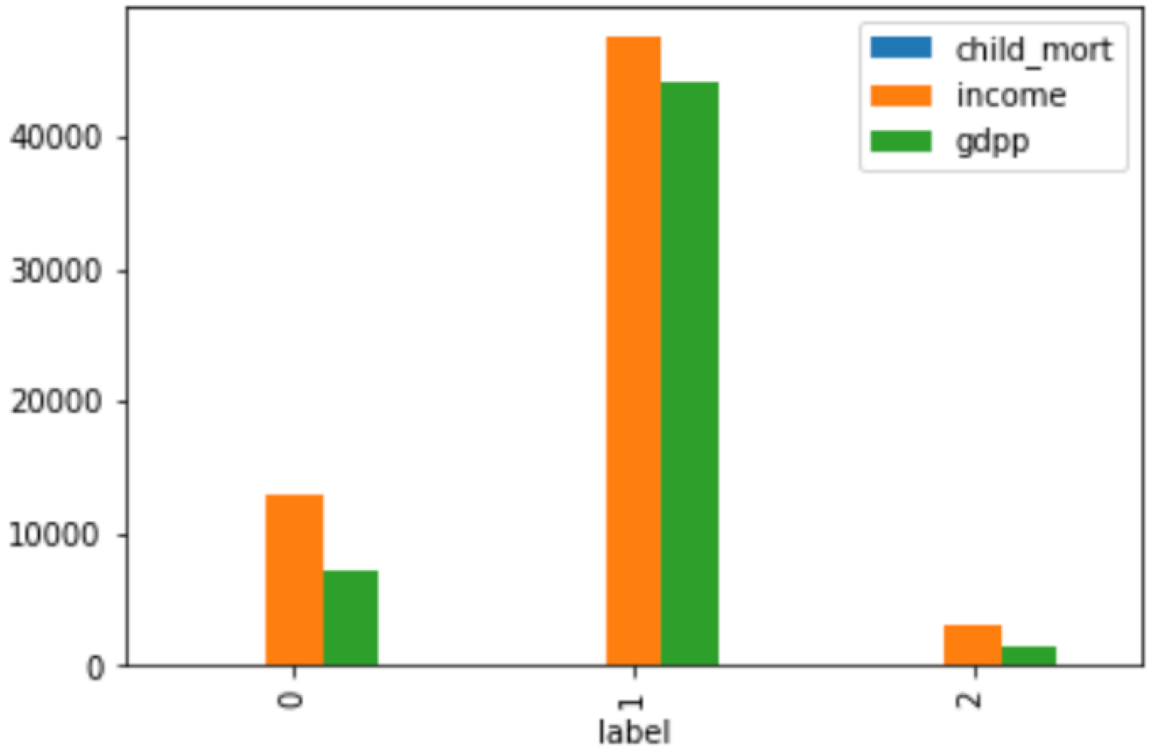
For k=3 (Clusters based on features







Cluster profiling

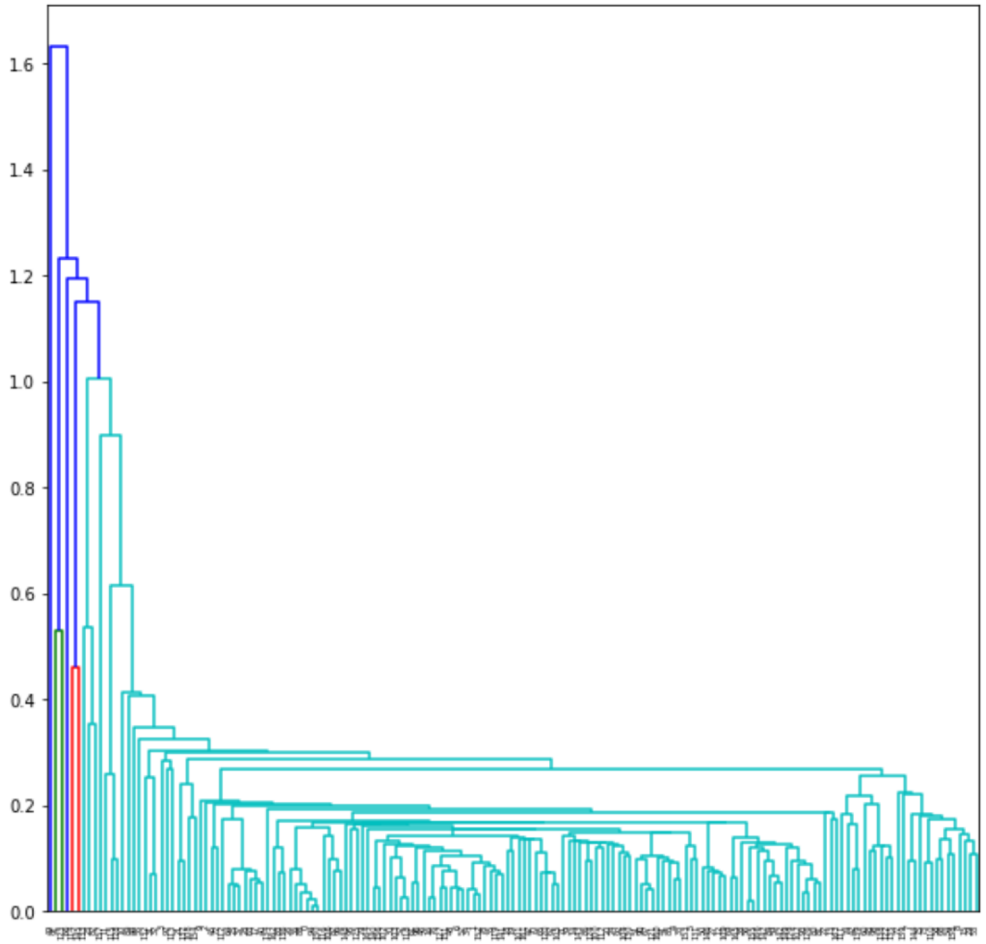


Top 5 countries with direst need for funding – k-means

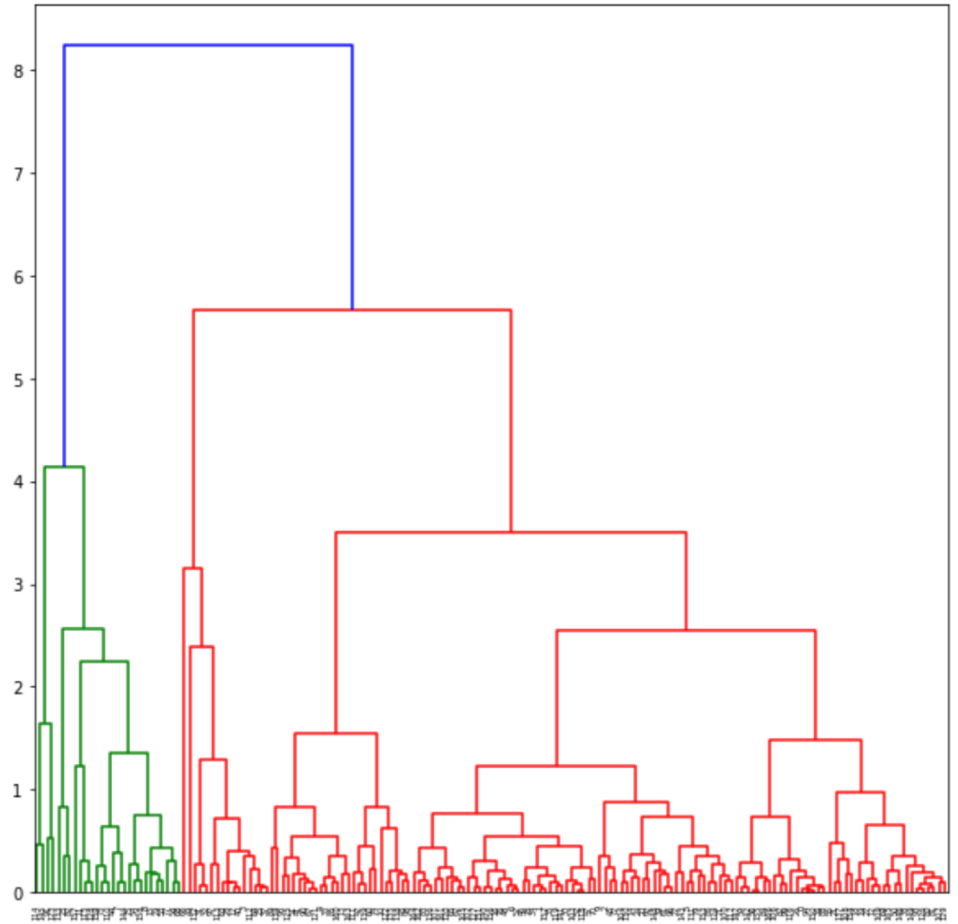
	child_mort	income	gdpp	country	label
66	208.0	1500.0	662.0	Haiti	2
132	160.0	1220.0	399.0	Sierra Leone	2
32	150.0	1930.0	897.0	Chad	2
31	149.0	888.0	446.0	Central African Republic	2
97	137.0	1870.0	708.0	Mali	2

Hierarchical Clustering

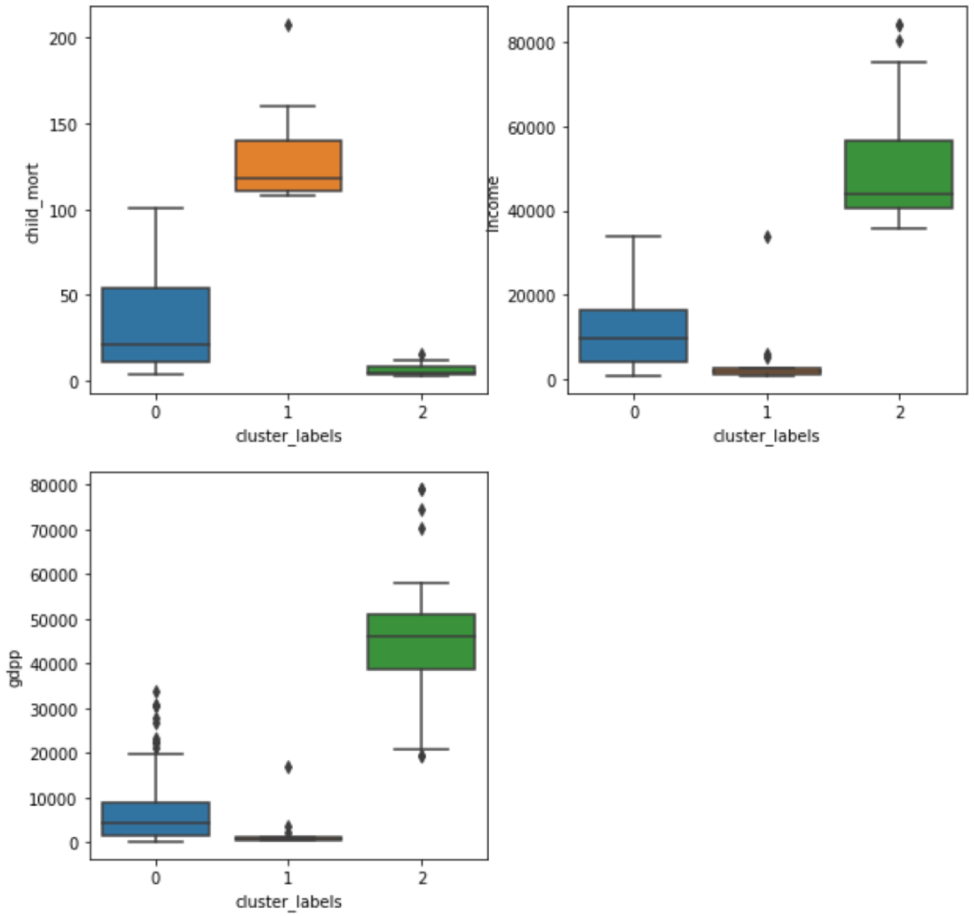
Single linkage



Complete linkage



Boxplot for income, child_mort, gdpp



Top 5 countries with direst need for funding – Hierarchical Clustering

	child_mort	income	gdpp	country	cluster_labels
66	208.0	1500.0	662.0	Haiti	1
132	160.0	1220.0	399.0	Sierra Leone	1
32	150.0	1930.0	897.0	Chad	1
31	149.0	888.0	446.0	Central African Republic	1
97	137.0	1870.0	708.0	Mali	1

Final result

- The output for both K-Means and Hierarchical are the same