Assignment

OBJECTIVE

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.

Problem Statement

As a data analyst. We need to categorize 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.

By – Gantavya Banga

Data Loading and processing

Loading the data:

[293]:
```
# Loading the Data
data = pd.read_csv('Country-data.csv')
data.head()
```

t[293]:

	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

This dataset has a total of 167 rows and 10 columns. The rows are the individual countries

Data Cleaning

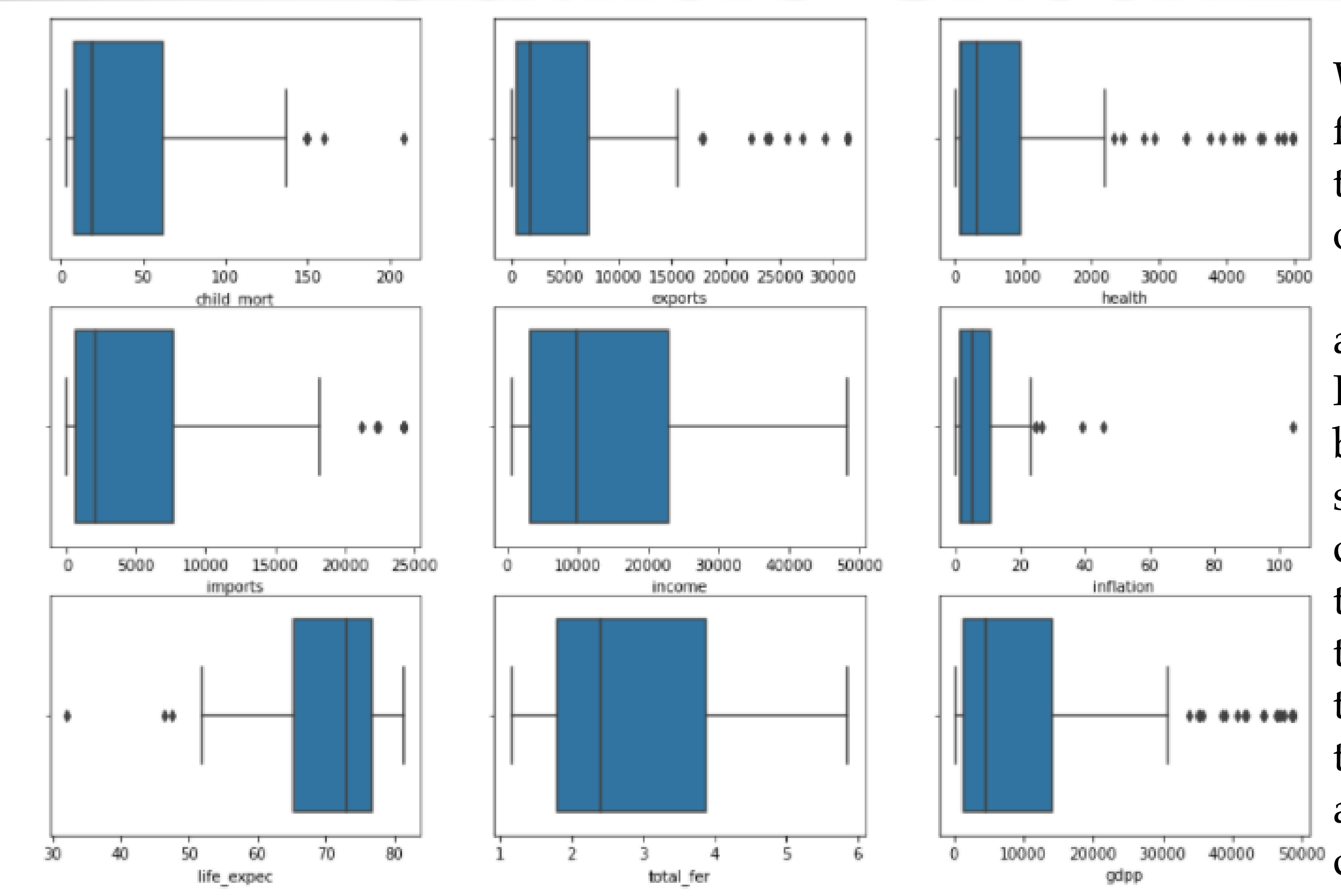
- The Dataset has no Missing Values
- The Dataset has no Duplicates
- Overall it looks like a clean data to be processed

Data Cleaning

In [294]:
```
# Converting exports,imports and health spending percentages to normal values.

data['exports'] = data['exports']*data['gdpp']/100

data['imports'] = data['imports']*data['gdpp']/100

data['health'] = data['health']*data['gdpp']/100
```

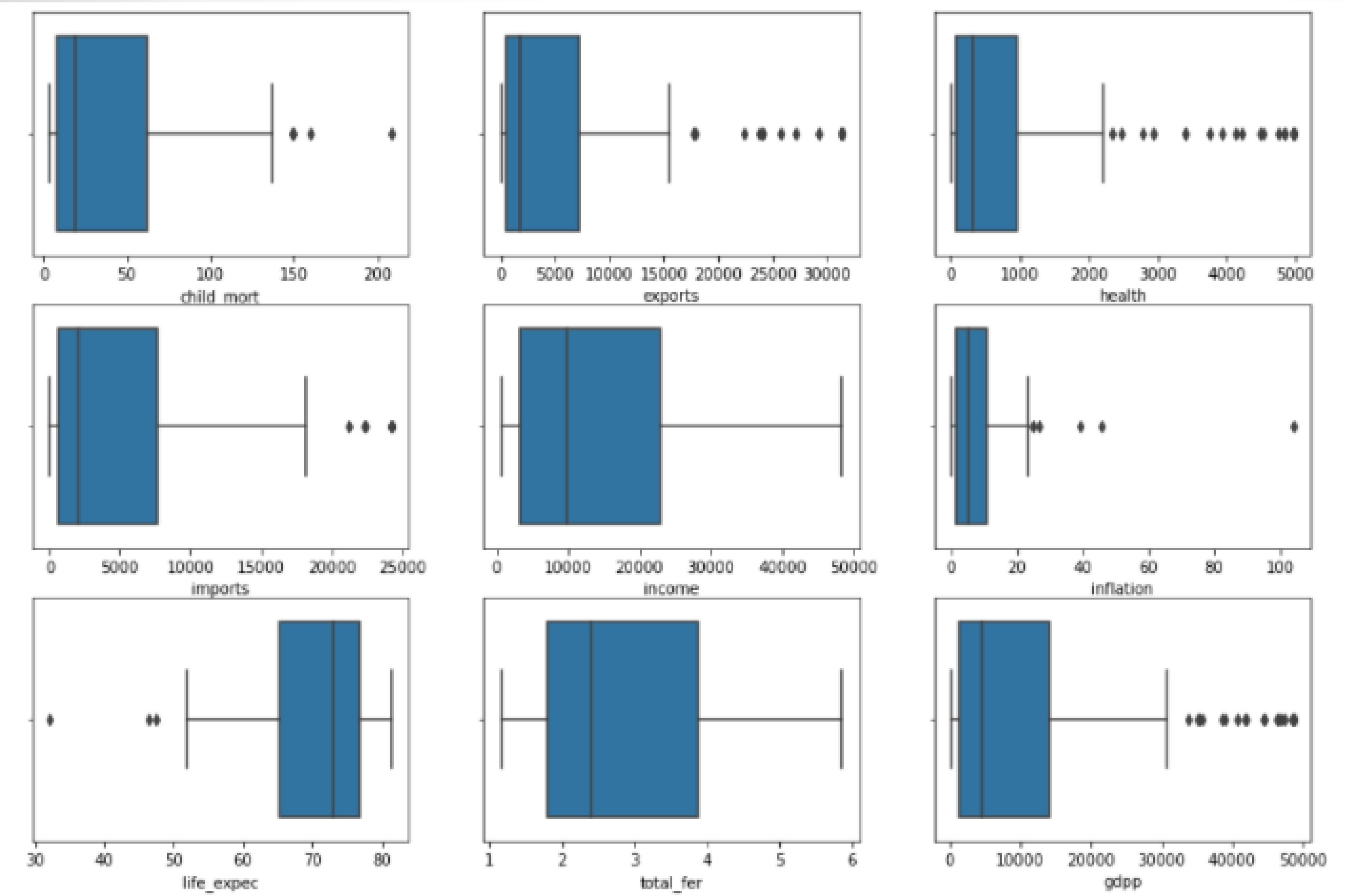
We converted the exports,imports and health spending percentages to absolute Values for better data understanding and processing.

Outliers Check



We can observe from the boxplots that there are outliers present in almost all the columns. However, We will be carefully selecting the columns and using the soft-Capping technique to cap the outliers using the lower range and upper range capping method.

Outliers Treatment



We made use of upper range method and the lower range to carefully cap the outliers depending on each column.

We went for lower range capping with inflation and child_motality columns, rest all the columns we went with upper range capping

We can now observe that most of the outliers are dealt with. There are still some outliers which can be seen in the boxplot , however we will be ignoring them as its important for the data analysis

HOPKINS CHECK

HOPSKINS CHECK

```
## Check the HOPKINS
from sklearn.neighbors import NearestNeighbors
from random import sample
from numpy.random import uniform
import numpy as np
from math import isnan

def hopkins(X):
    d = X.shape[1]
    #d = len(vars) # columns
    n = len(X) # rows
    m = int(0.1 * n)
    nbrs = NearestNeighbors(n_neighbors=1).fit(X.values)

    rand_X = sample(range(0, n, 1), m)

    ujd = []
    wjd = []
    for j in range(0, m):
        u_dist, _ = nbrs.kneighbors(uniform(np.amin(X,axis=0),np.amax(X,axis=0),d).reshape(1, -1), 2, return_distance=True)
        ujd.append(u_dist[0][1])
        w_dist, _ = nbrs.kneighbors(X.iloc[rand_X[j]].values.reshape(1, -1), 2, return_distance=True)
        wjd.append(w_dist[0][1])

    H = sum(ujd) / (sum(ujd) + sum(wjd))
    if isnan(H):
        print(ujd, wjd)
        H = 0

    return H
```

```
hopkins(data.drop(['country'],axis=1))
```

```
0.9284122191851386
```

We know that anything over 0.80 is a good number for clustering and in this case we have got 0.92 which is a good signal for us. And we can hereby conclude that this is a good dataset for clustering

SCALING THE DATA

SCALING

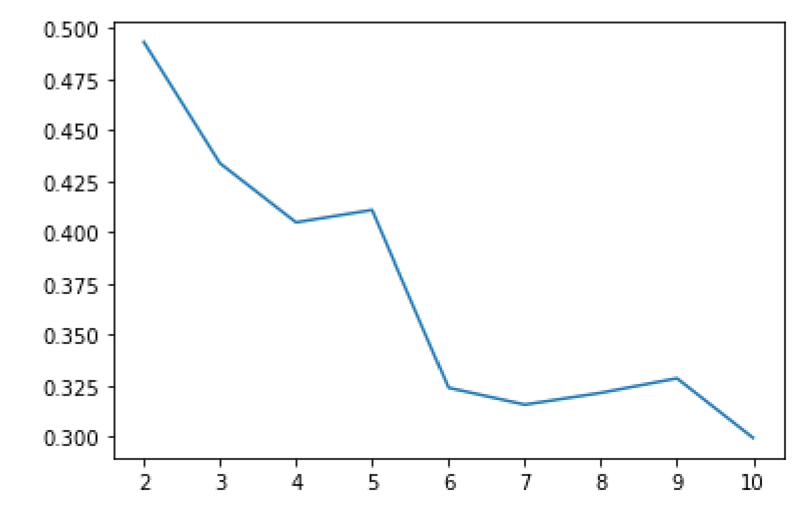
```
### Scaling

from sklearn.preprocessing import StandardScaler
ss = StandardScaler()
data2 = ss.fit_transform(data.drop(['country'],axis=1))
```

```
data2 = pd.DataFrame(data2)
data2.columns = data1.columns
data2.head()
```

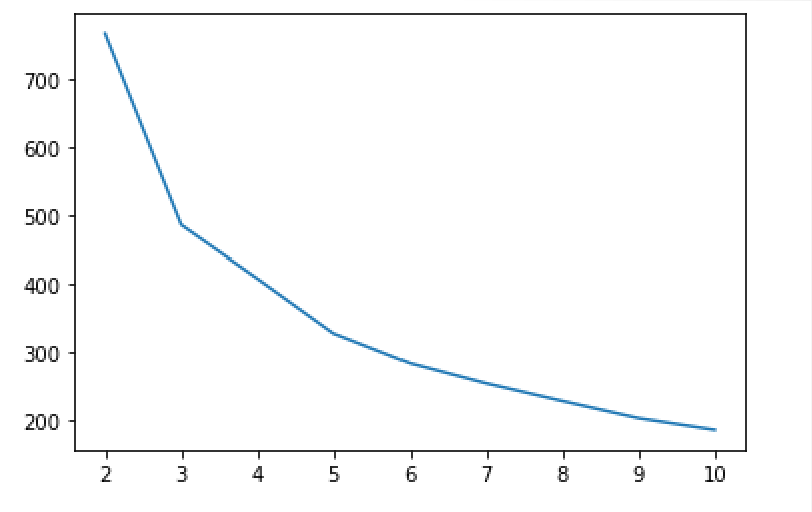
	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	1.291607	-0.669581	-0.629435	-0.732729	-0.958349	0.150169	-1.623180	2.016421	-0.757362
1	-0.539812	-0.542172	-0.473489	-0.472182	-0.394006	-0.322868	0.654823	-0.880535	-0.523321
2	-0.273560	-0.475838	-0.530017	-0.560152	-0.192552	0.786618	0.677490	-0.019090	-0.498838
3	2.008250	-0.418960	-0.588935	-0.543087	-0.667360	1.388664	-1.181180	2.044904	-0.560376
4	-0.696578	-0.027134	-0.150685	0.306422	0.227992	-0.614335	0.711490	-0.547072	0.013312

Scaling is one of the crucial step in the cluster analysis and in this data, we made use of standard scaler tool from Sklearn library to standardize the whole data

K-Means Clustering



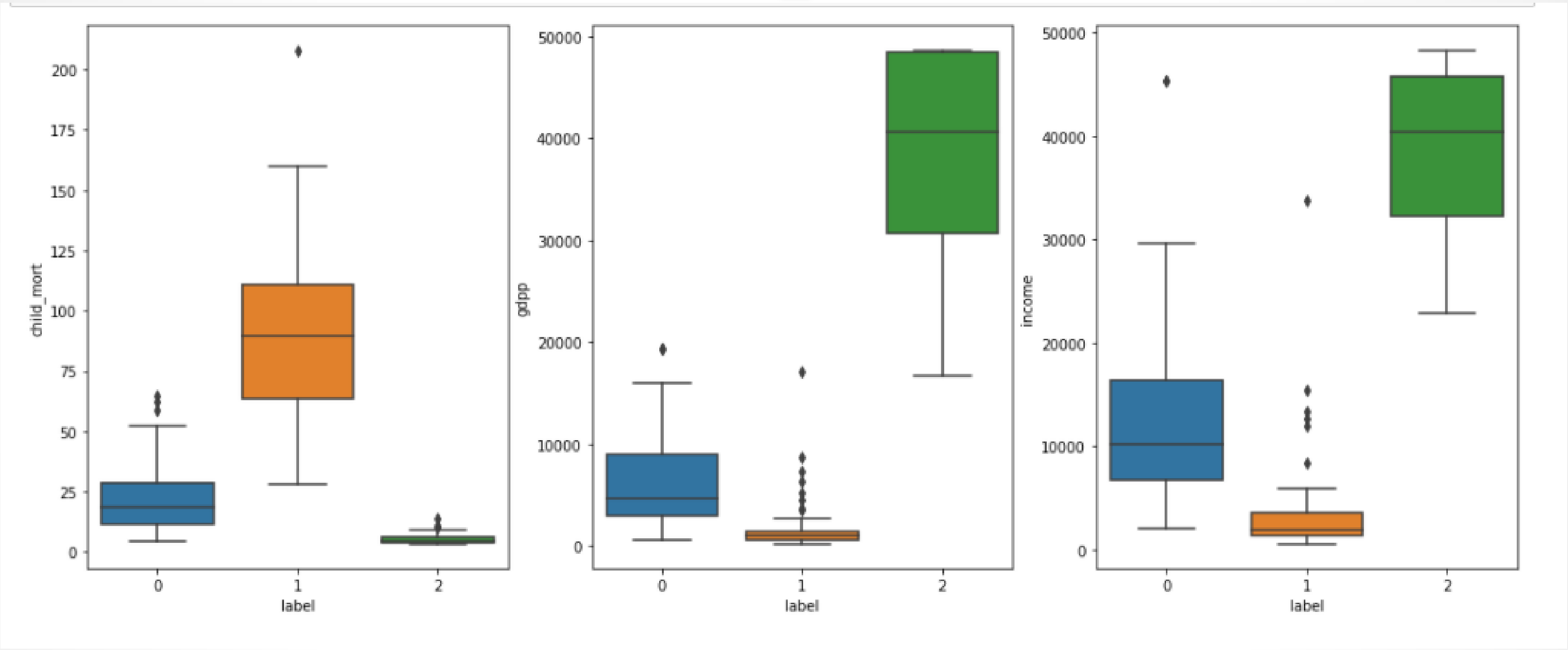
Silhouette Curve



Elbow Curve

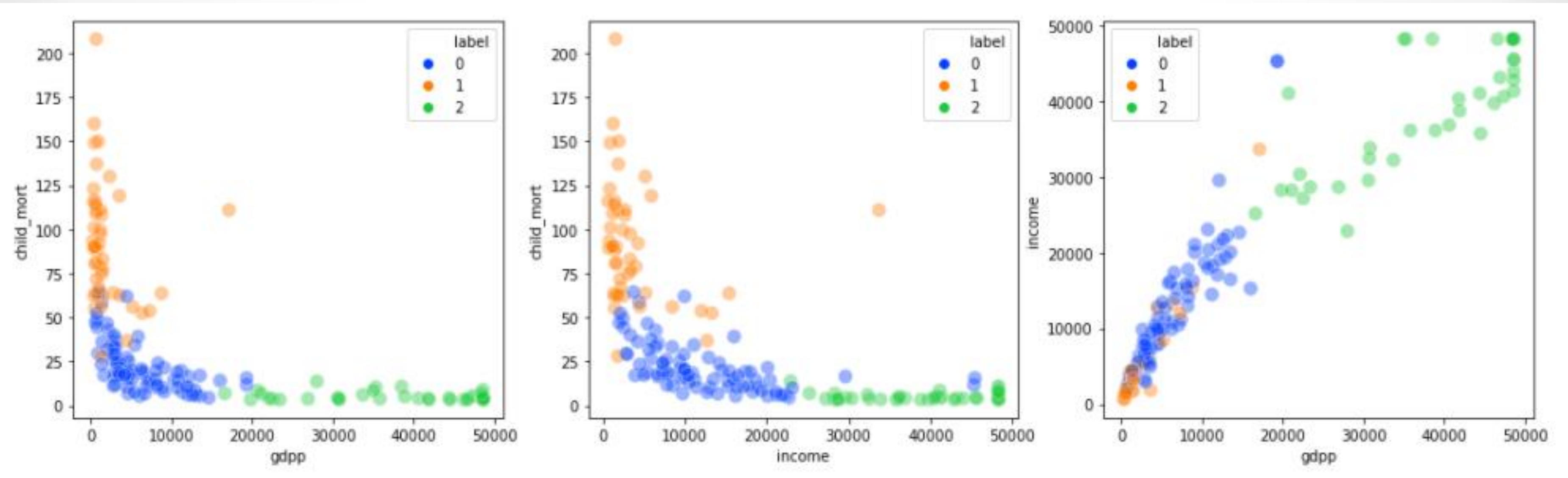
By looking silhouette analysis, we see the highest peak is at k =3 and 5
and in sum of squared distances graph , we see that the elbow is in the range of 3 to 5 ,
so we are going ahead with k as 3.

K-means clustering Boxplot



We can observe from this boxplot that cluster with label 1 has got the countries with the highest child motality rate. Also Label 1 has got the least GDPP and least income. Therefore, label 1 seems to have the countries really struggling at the minute

K-means clustering Scatterplot



From the above scatter plot we can obsrve that countries with low gdpp, low income, hight child motality rate is forming cluster 1 in orange colour and is in far left side of the charts. These will be countries we will be focusing on to decise which country should be given financial aid.

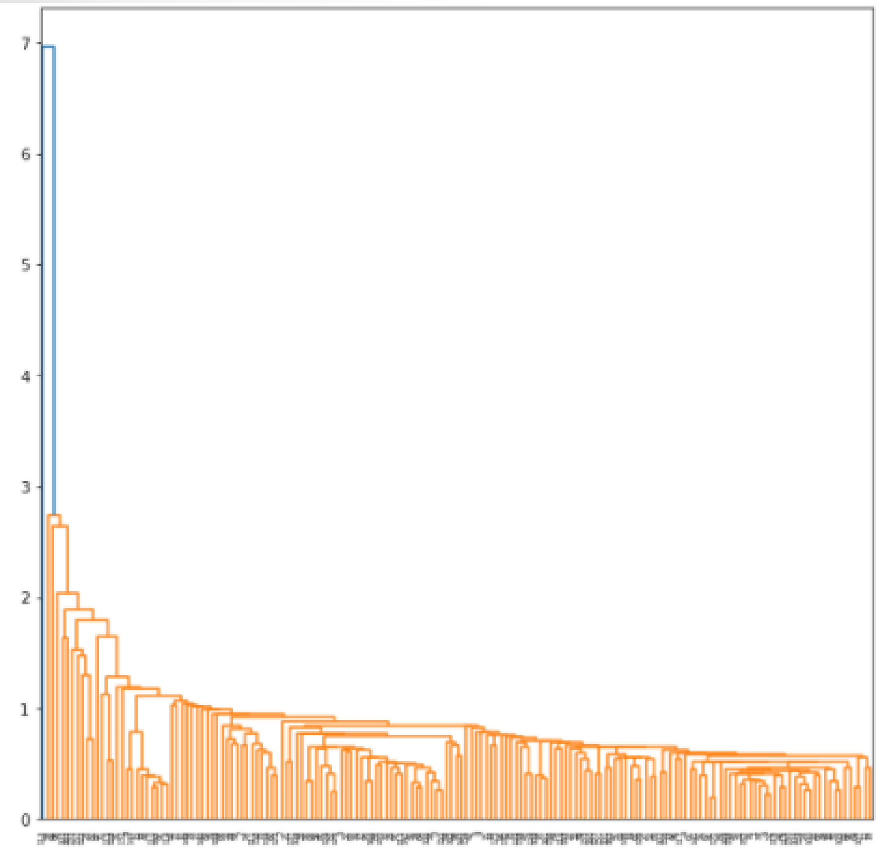
Top 10 countries in need of aid under K-means clustering method and label 1

	country	child_mort	income	gdpp
26	Burundi	93.6	764.0	231
88	Liberia	89.3	700.0	327
37	Congo, Dem. Rep.	116.0	609.0	334
112	Niger	123.0	814.0	348
132	Sierra Leone	160.0	1220.0	399
93	Madagascar	62.2	1390.0	413
106	Mozambique	101.0	918.0	419
31	Central African Republic	149.0	888.0	446
94	Malawi	90.5	1030.0	459
50	Eritrea	55.2	1420.0	482

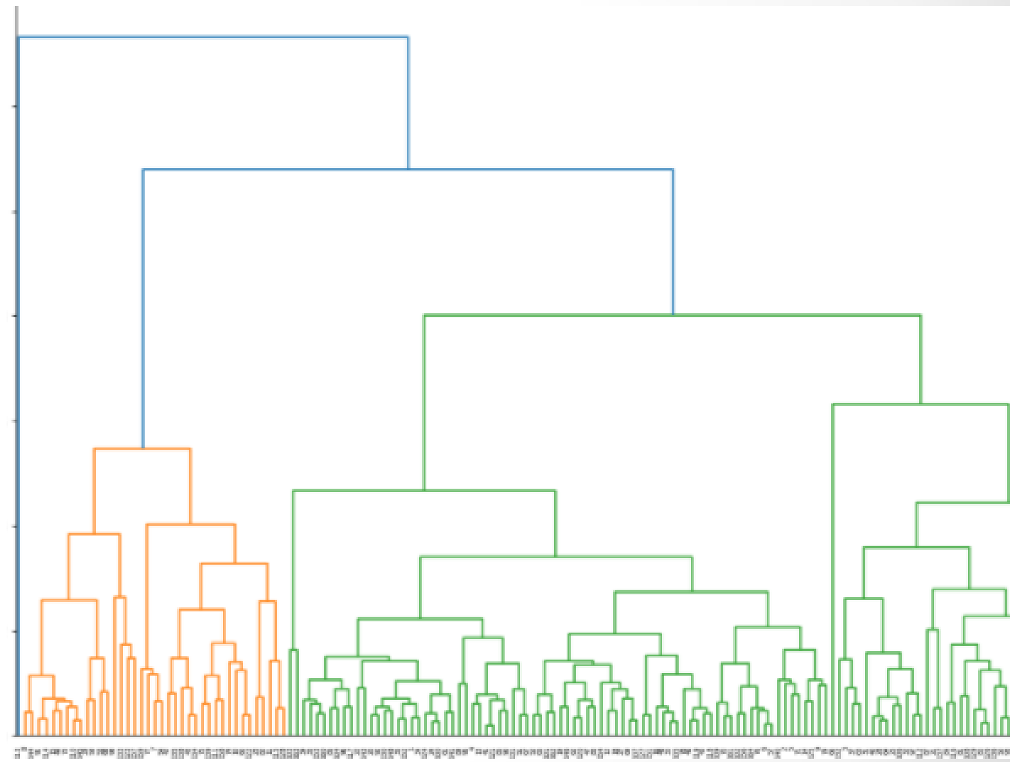
These are the top 10 countries which are currently struggling the most in terms of low gdpp, low income and high child mortality. These countries should be prioritized and given all the financial and medial aid.

Hierarchical Clustering

Single Linkage method

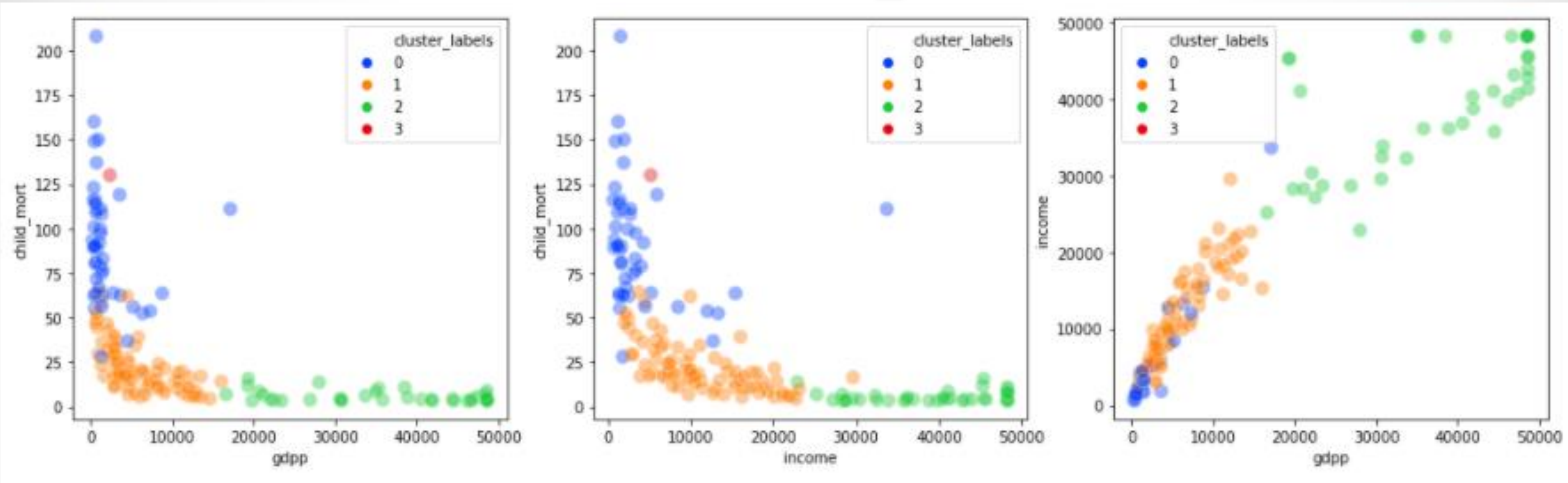


Complete Linkage method



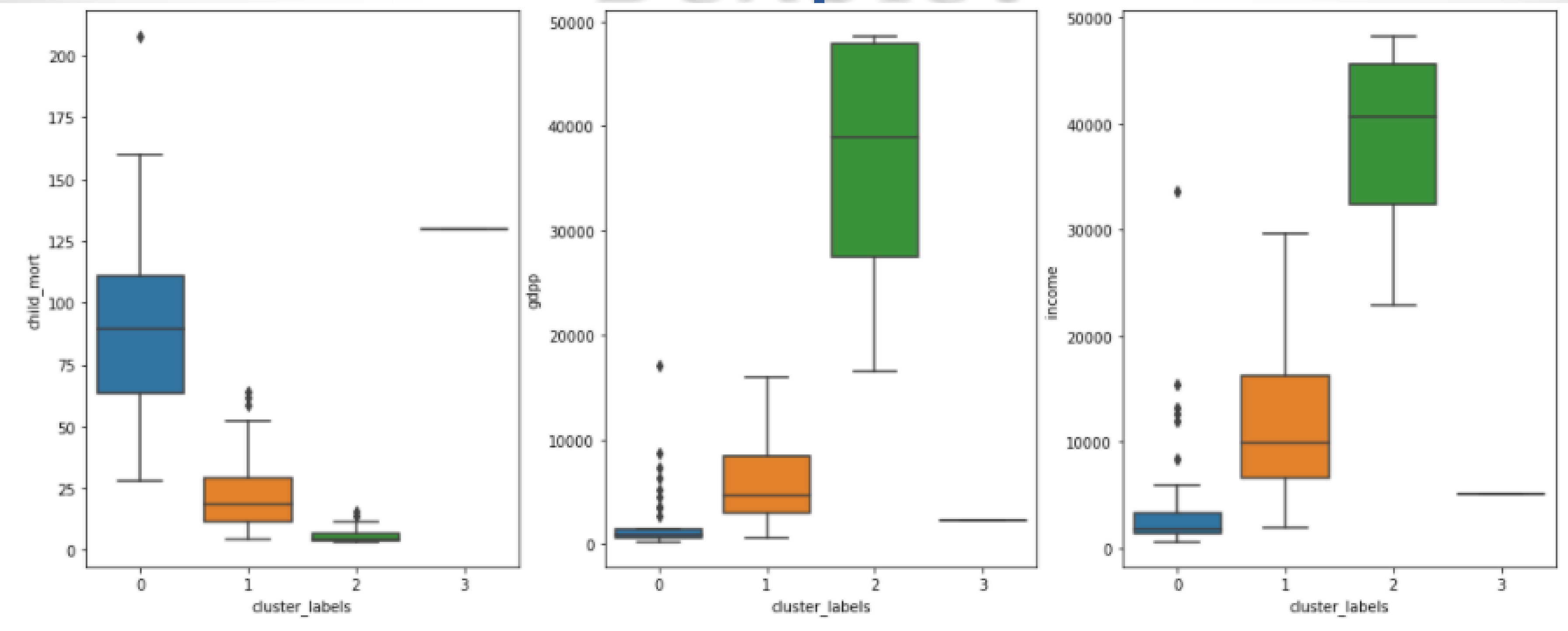
We are going to go with Complete linkage clustering method as single linkage clustering is not clear. By looking at this dendogram taking n-clusters = 4

Hierarchical clustering Scatterplot



From the above scatter plot we can obsrve that countries with low gdpp, low income, hight child motality rate is forming cluster 0 in orange colour and is in far left side of the charts. These will be countries we will be focusing on to decise which country should be given financial aid.

Hierarchical clustering Boxplot



We can observe from this boxplot that cluster with label 0 has got the countries with the highest child motality rate. Also Label 0 has got the least GDPP and least income. Therefore, label 0 seems to have the countries really struggling at the minute

Top 10 countries in need of aid under Hierarchical clustering method and label 0

	country	child_mort	income	gdpp
26	Burundi	93.6	764.0	231
88	Liberia	89.3	700.0	327
37	Congo, Dem. Rep.	116.0	609.0	334
112	Niger	123.0	814.0	348
132	Sierra Leone	160.0	1220.0	399
93	Madagascar	62.2	1390.0	413
106	Mozambique	101.0	918.0	419
31	Central African Republic	149.0	888.0	446
94	Malawi	90.5	1030.0	459
50	Eritrea	55.2	1420.0	482

These are the top 10 countries which are currently struggling the most in terms of low gdpp, low income and high child mortality. These countries should be prioritized and given all the financial and medial aid.

CONCLUSION

We can observe that we got the same sets of countries using k-means and Hierarchical clustering method and sorted the using some socio-economic and health factors that determine the overall development of the country.

	country	child_mort	income	gdpp
26	Burundi	93.6	764.0	231
88	Liberia	89.3	700.0	327
37	Congo, Dem. Rep.	116.0	609.0	334
112	Niger	123.0	814.0	348
132	Sierra Leone	160.0	1220.0	399
93	Madagascar	62.2	1390.0	413
106	Mozambique	101.0	918.0	419
31	Central African Republic	149.0	888.0	446
94	Malawi	90.5	1030.0	459
50	Eritrea	55.2	1420.0	482