Journey of the analysis

Approach adapted to find out the top-10 countries who are in dire need of aid.

Title and Content Layout with List

1. Problem Statement.
2. Roadmap of analysis.
    i. Treatment of outliers.
    ii. Exploratory Data Analysis.
    iii. Data preparation for clustering.
    iv. Methods Used
        a. K-Mean
        b. Hierarchical Clustering
            1. Single method
            2. Complete method
3. Results

Problem Statement :

To cluster the countries using some socio-economic and health factors that determine the overall development of the country., and find out top-most countries who are in dire need of aid, so that the CEO of the NGO can decide how to use the received financial package strategically and effectively.

Roadmap of analysis

Understand Data, and check for any abnormality in data.

Inspection of missing values & outlier, and treat them accordingly.

Exploratory Data Analysis, using pair plot/ and scatter plot as there were only continuous data.

Preparing Data:

1. Check the data, if it is compatible for clustering or not, using HOPKINS score

Hopkins score were calculated 10 times by random selection of data.

Conclusion: Each time, it is more than 80 hence data seems to be good for clustering.

Scaling the data, because of Income and GDP columns as values are really huge as compared to the values of rest of the columns.

Used Standard Scaler to scale

Once data is scaled, check for optimum number of clusters using technique of Silhouette Score and Elbow curve

Applied Hierarchical clustering (single method and complete method)and it turns out that it shows only 2 clusters, and obtained top 10 countries who are in dire need.

Finally applied K-Means on the scaled data for K=3, and obtained top 10 countries who are in dire need.

Final result:

There is no difference in the end result of the countries which we want to select. Hence both method is correct, still if I would prefer, it would be K-Mean method.

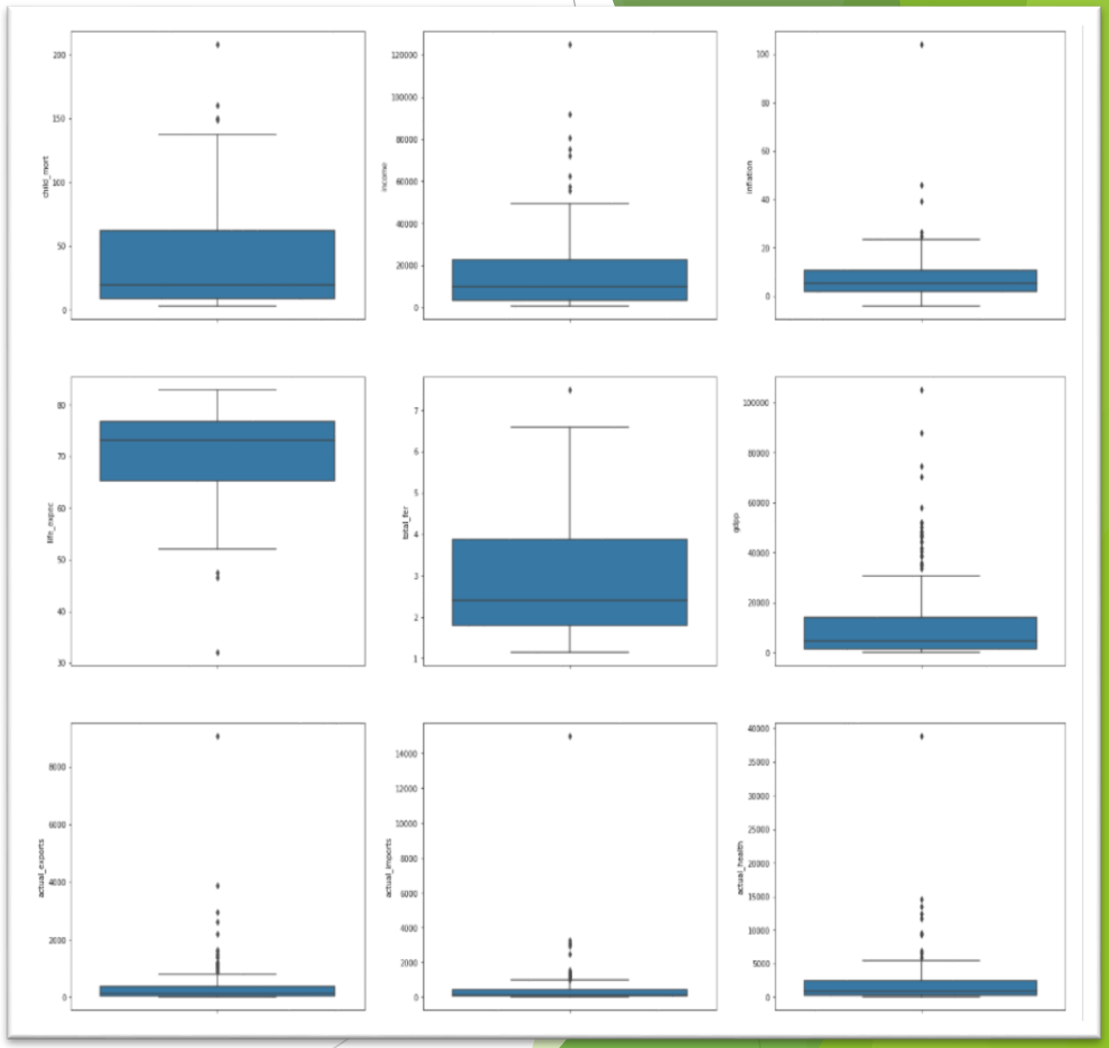
Treatment of outliers.

Looking at the above box-plots it is clear that there are outliers present which can drastically affect our clustering as clustering is distance based method. In order to treat outliers will need to treat upper and lower outliers separately for each column focusing our problem statement.

Hence for each column will perform treatment as follow:

1.child_mort : Will not treat upper outliers as we need to identify countries whom immediate aid can be provided and high child_mort plays imp role in deciding.
2.income : Can fix upper outliers.
3.inflation : There seems to not much of inflation pointers as 99%ile is 41.48 and 100%ile is 104
4.life_expec : No upper outliers but there are few lower outliers which is important for analysis hence will not treat it.
5.total_fer : There seems to not much of difference in 99%ile (6.56) and 100%ile (7.49) hence no need to treat outliers
6.gdpp : Looking at the box-plot it looks like there are several outliers, which in real world we know as developed nations. We need to treat these outliers otherwise it will drastically affect the clustering of lower end gdpp
7.actual_exports : Need to treat upper end outliers
8.actual_imports : Need to treat upper end outliers
9.actual_health : Need to treat upper end outliers

Treatment: Capped the outliers (which need to be treated) in soft rang, i.e. 99%ile



Exploratory Data Analysis

- Since all columns are continuous variable, performed EDA accordingly i.e. distribution graph, not much insights.
- Performed bi-variate analysis on loop while fixing income, child_mortality and GDP.

Insights:

- Child mortality shows strong inverse relationship with income of country.
- Life Expectancy follows direct relationship with income of country & inverse relationship with child_mort
- Age -fertility which is (total_fer) decreases where income is high and increases with increase in child_mort or vice versa.
- Export, import, GDP and expenditure on health show direct relationship with income of country and inverse relationship with child_mort
-

Data preparation for clustering (i)

- Performed Hopkins test for 10 times, to check compatibility of data to form clusters.

```
for i in range(0,10):
    print((i+1),") Hopkins score", round(hopkins(new_data.drop('country', axis=1)
```

executed in 406ms, finished 15:42:30 2020-08-30

```
1 ) Hopkins score 0.93
2 ) Hopkins score 0.94
3 ) Hopkins score 0.93
4 ) Hopkins score 0.94
5 ) Hopkins score 0.89
6 ) Hopkins score 0.9
7 ) Hopkins score 0.86
8 ) Hopkins score 0.94
9 ) Hopkins score 0.81
10 ) Hopkins score 0.87
```

- Looking at the scores, it is concluded that data is distributed widely that form clusters

Data preparation for clustering (ii)

- Once it is established that data can be clustered, need to scale data because of the difference in the values of GDPP/income w.r.t to values of child_mortality column.

Out[516]:

	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

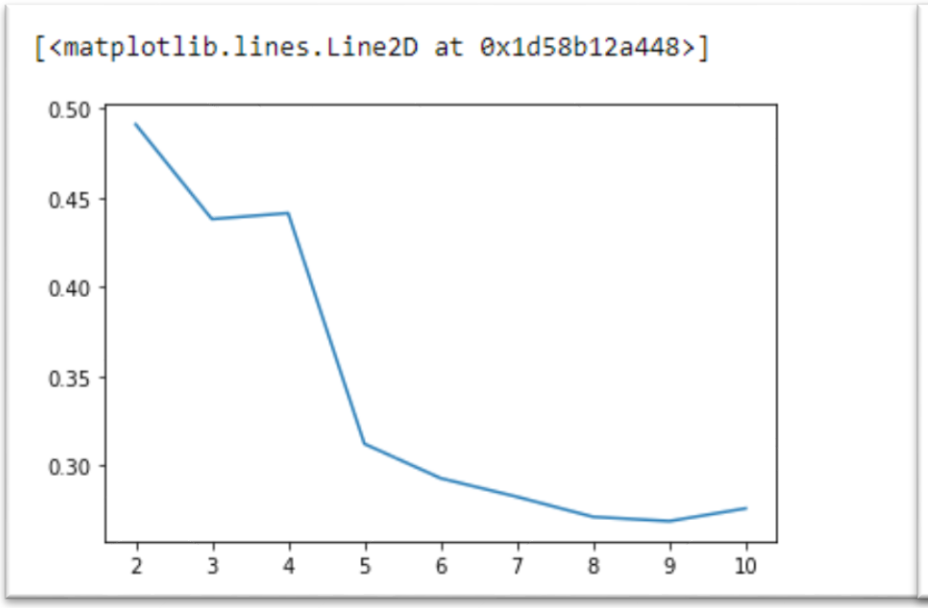
Methods Used

- K-Mean Clustering method:

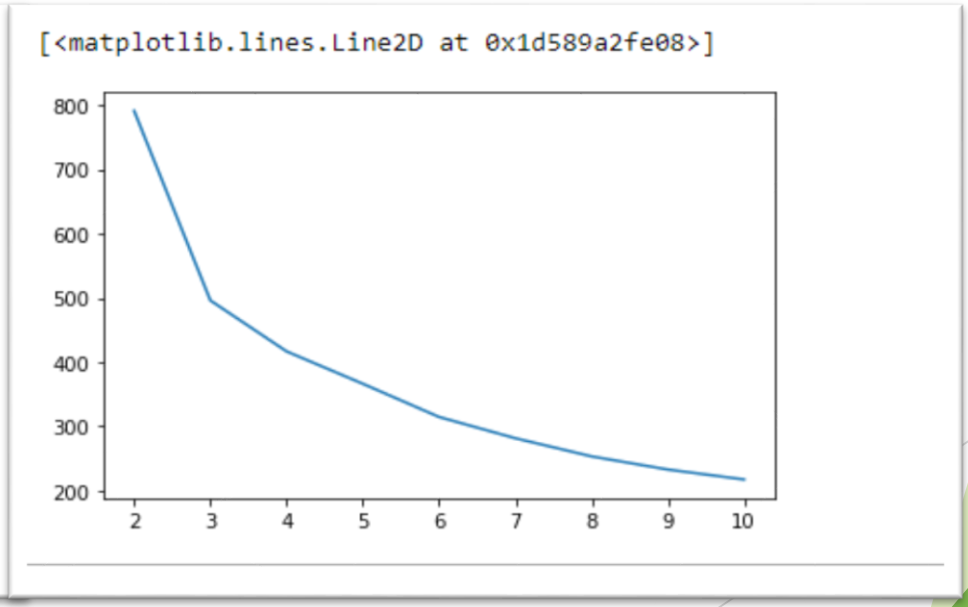
After scaling data, opted to use K-Mean scaling technique to cluster the data. But before applying this method need to find the optimum value of K(clusters).

For that used Silhouette score and Elbow curve analysis to decide optimum number of clusters.

Silhouette Score



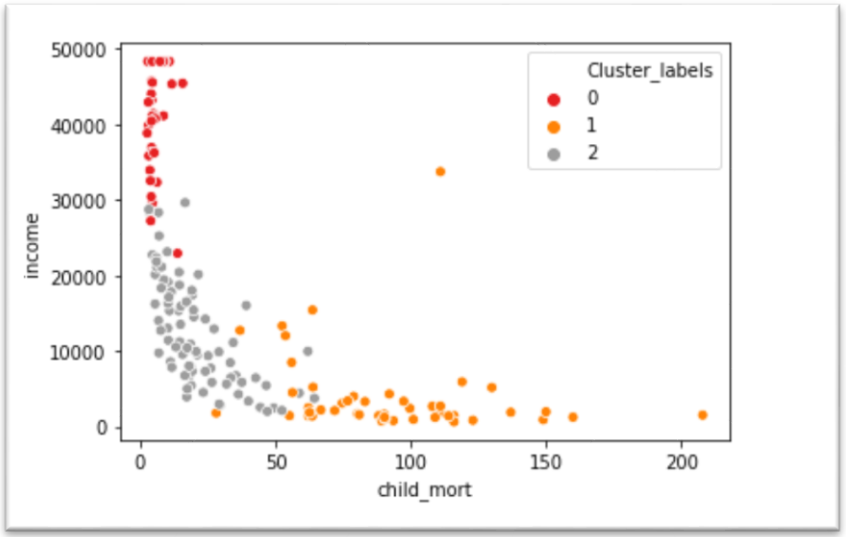
Elbow Curve

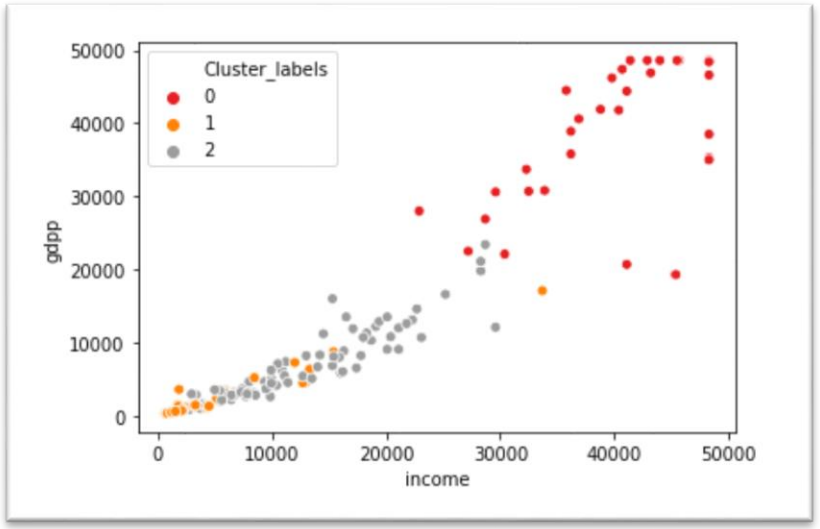


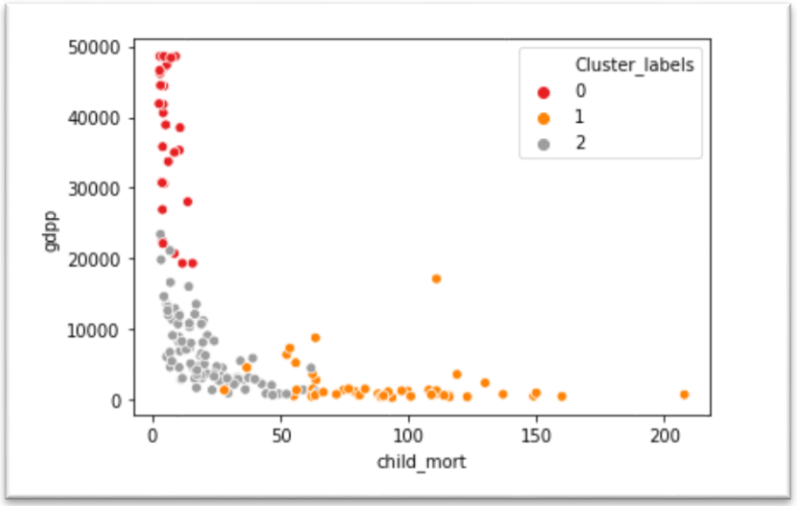
Looking at both, decided to opt for K=3.

K-Mean

- Using K-Mean clustering modelling, got desired results as after plotting all the three features against each other, clusters can be seen distinct.





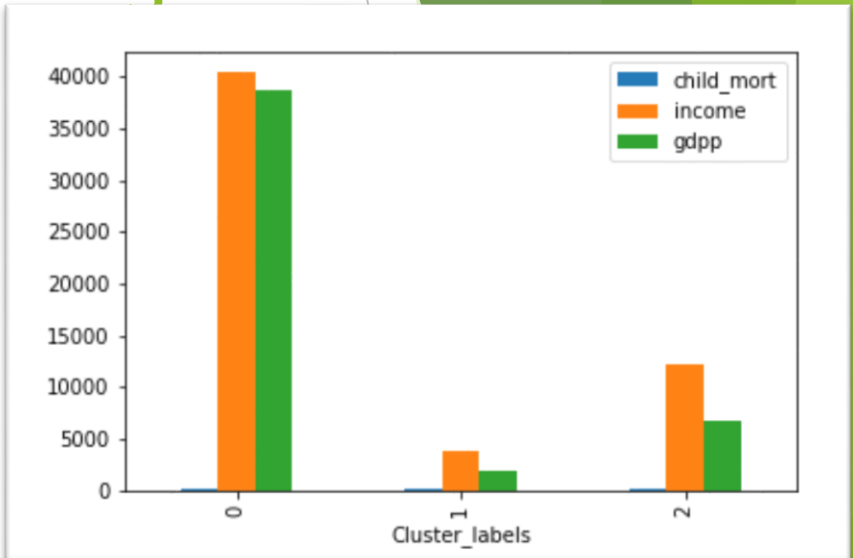


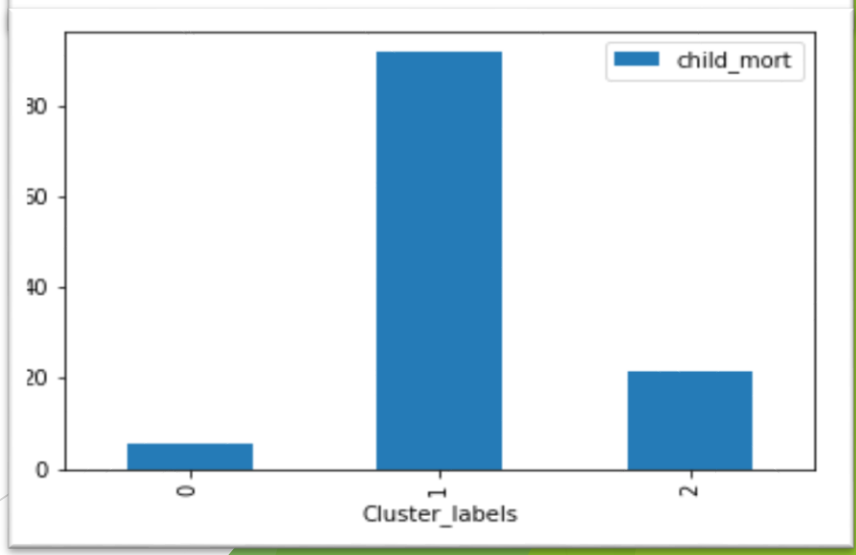
Top 10 countries in dire need of aid, based on features child_mort, income and GDP using K-Mean

- First, did the cluster profiling based on features, child_mort, income and GDP.
- It seems to be a clear difference between 3 clusters based on income and gdpp, however for child_mortality it seems to be not clear because of scaling of income and gdpp.

Insights:

- 1. Cluster 0: High Income, high gdpp but low child_mortality.
- 2. Cluster 1: Low Income, low gdpp and high child_mortality.
- 3. Cluster 2: medium income, medium gdp and medium child_mortality





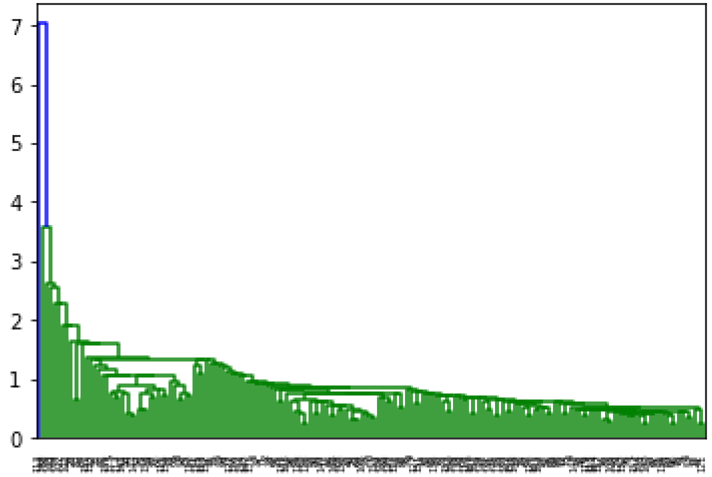
Result using K-Means: Top 10 countries who are in dire need of aid are :

- 1. Haiti
- 2. Sierra Leone
- 3. Chad
- 4. Central African Republic
- 5. Mali
- 6. Nigeria
- 7. Niger
- 8. Angola
- 9. Congo, Dem. Rep.
- 10.Burkina Faso

	country	child_mort	income	inflation	life_expec	total_fer	gdpp
66	Haiti	208.0	1500.0	5.45	32.1	3.33	662
132	Sierra Leone	160.0	1220.0	17.20	55.0	5.20	399
32	Chad	150.0	1930.0	6.39	56.5	6.59	897
31	Central African Republic	149.0	888.0	2.01	47.5	5.21	446
97	Mali	137.0	1870.0	4.37	59.5	6.55	708
113	Nigeria	130.0	5150.0	104.00	60.5	5.84	2330
112	Niger	123.0	814.0	2.55	58.8	7.49	348
3	Angola	119.0	5900.0	22.40	60.1	6.16	3530
37	Congo, Dem. Rep.	116.0	609.0	20.80	57.5	6.54	334
25	Burkina Faso	116.0	1430.0	6.81	57.9	5.87	575

Hierarchical Clustering – Single method

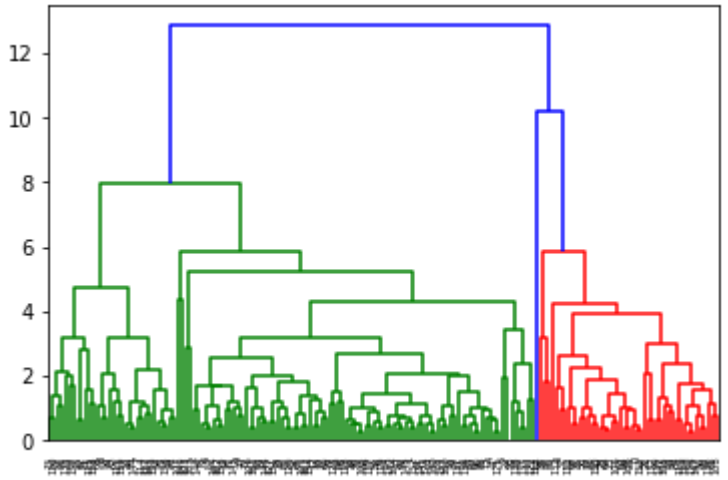
- Performed Hierarchical clustering, with single linkage but result obtained are not satisfactory.
- Taken, K=3 shows that, it did not distinguishes cluster distinctly.




```
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0])
```

Hierarchical Clustering – complete method

- Performed Hierarchical clustering, with complete method but result obtained are not satisfactory.
- Taken, K=3 shows that, it did not distinguishes cluster distinctly.

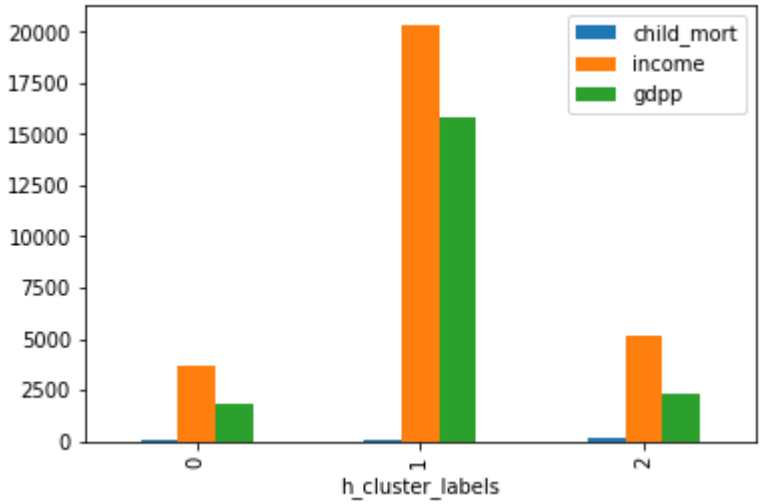


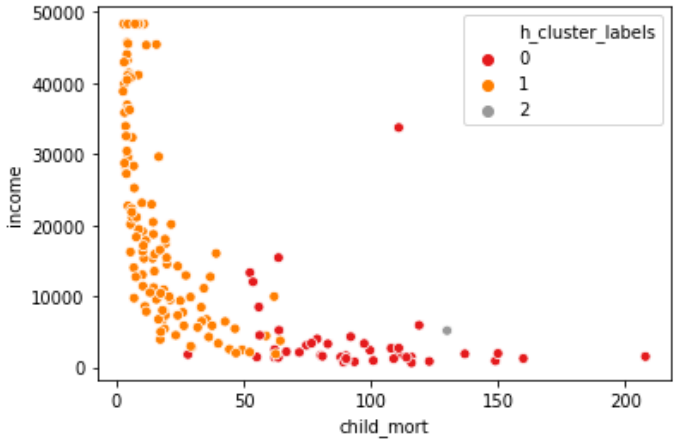

```
array([0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0,
       1, 1, 1, 0, 0, 1, 0, 1, 1, 0, 0, 1, 1, 1, 0, 0, 0, 1, 0, 1, 1, 1,
       1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 0, 0, 1, 1, 0, 1, 1, 1, 0, 0, 1,
       0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 0, 1, 1, 0,
       0, 1, 1, 1, 1, 0, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1,
       1, 1, 0, 2, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1,
       0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1, 1,
       1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0])
```

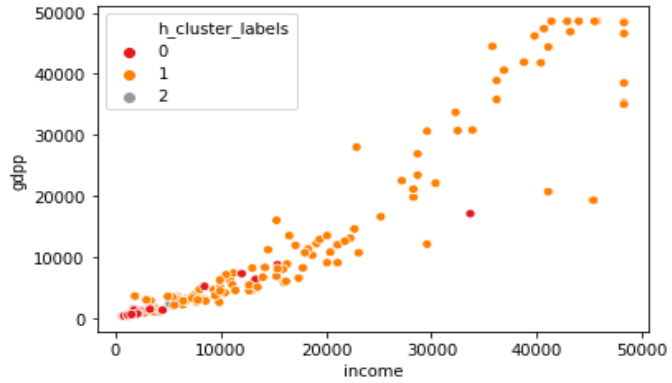
Hierarchical Clustering

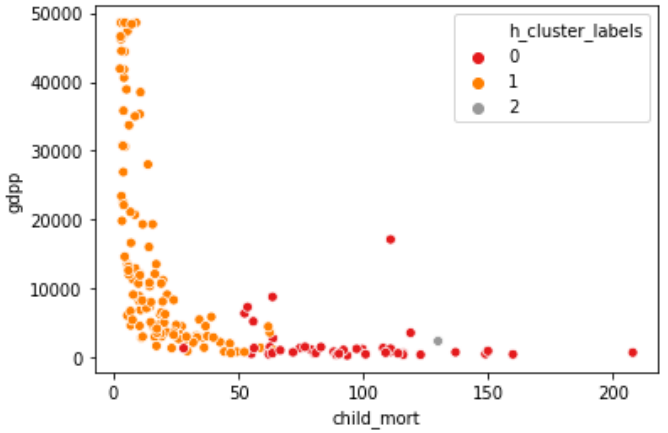
Clearly it is evident that using Hierarchical clustering method, n=3 does not give 3 distinct clusters.

Hence, took n=2







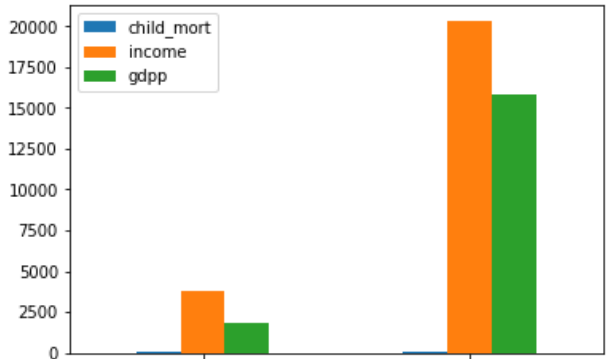


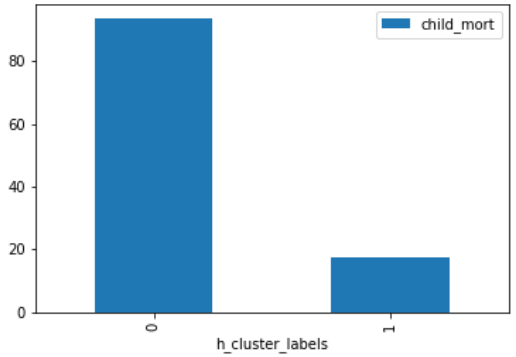
Hierarchical Clustering, using n=2

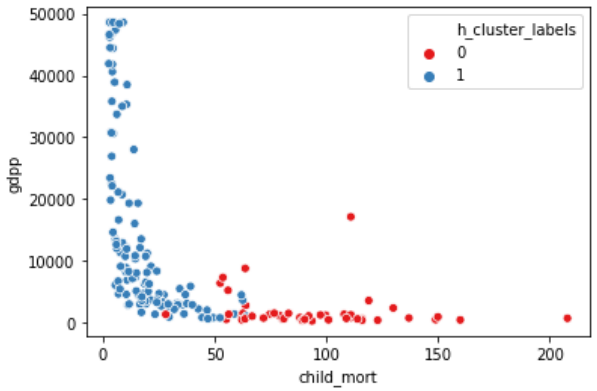
Using n=2, it is clear that data is clustered in two categories.

1. Cluster 0: High Income, high gdpp but low child_mortality.

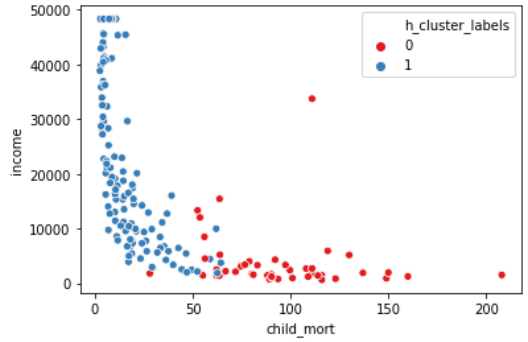
2. Cluster 1: Low Income, low gdpp and high child_mortality.

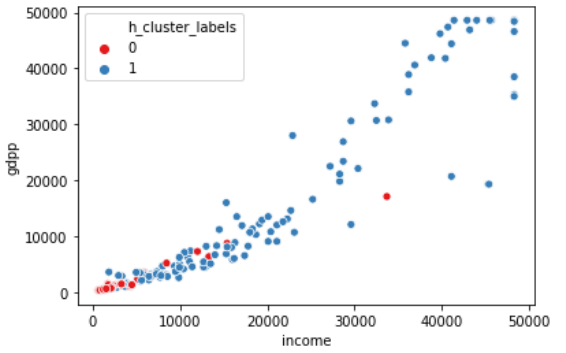







```
<matplotlib.axes._subplots.AxesSubplot at 0x1d58de57a48>
```





Result using Hierarchical clustering Top 10 countries who are in dire need of aid are :

- 1. Haiti
- 2. Sierra Leone
- 3. Chad
- 4. Central African Republic
- 5. Mali
- 6. Nigeria
- 7. Niger
- 8. Angola
- 9. Congo, Dem. Rep.
- 10. Burkina Faso

	country	child_mort	income	inflation	life_expec	total_fer	gdpp	ac
66	Haiti	208.0	1500.0	5.45	32.1	3.33	662	
132	Sierra Leone	160.0	1220.0	17.20	55.0	5.20	399	
32	Chad	150.0	1930.0	6.39	56.5	6.59	897	
31	Central African Republic	149.0	888.0	2.01	47.5	5.21	446	
97	Mali	137.0	1870.0	4.37	59.5	6.55	708	
113	Nigeria	130.0	5150.0	104.00	60.5	5.84	2330	
112	Niger	123.0	814.0	2.55	58.8	7.49	348	
3	Angola	119.0	5900.0	22.40	60.1	6.16	3530	
37	Congo, Dem. Rep.	116.0	609.0	20.80	57.5	6.54	334	
25	Burkina Faso	116.0	1430.0	6.81	57.9	5.87	575	

Conclusion

- There is no difference in the end result of the countries which we want to select. Hence both method is correct, still if I would prefer, it would be K-Mean method.

K-MEAN

	country	child_mort	income	inflation	life_expec	total_fer	gdpp	a
66	Haiti	208.0	1500.0	5.45	32.1	3.33	662	
132	Sierra Leone	160.0	1220.0	17.20	55.0	5.20	399	
32	Chad	150.0	1930.0	6.39	56.5	6.59	897	
31	Central African Republic	149.0	888.0	2.01	47.5	5.21	446	
97	Mali	137.0	1870.0	4.37	59.5	6.55	708	
113	Nigeria	130.0	5150.0	104.00	60.5	5.84	2330	
112	Niger	123.0	814.0	2.55	58.8	7.49	348	
3	Angola	119.0	5900.0	22.40	60.1	6.16	3530	
37	Congo, Dem. Rep.	116.0	609.0	20.80	57.5	6.54	334	
25	Burkina Faso	116.0	1430.0	6.81	57.9	5.87	575	

HIERARCHICAL

	country	child_mort	income	inflation	life_expec	total_fer	gdpp	ac
66	Haiti	208.0	1500.0	5.45	32.1	3.33	662	
132	Sierra Leone	160.0	1220.0	17.20	55.0	5.20	399	
32	Chad	150.0	1930.0	6.39	56.5	6.59	897	
31	Central African Republic	149.0	888.0	2.01	47.5	5.21	446	
97	Mali	137.0	1870.0	4.37	59.5	6.55	708	
113	Nigeria	130.0	5150.0	104.00	60.5	5.84	2330	
112	Niger	123.0	814.0	2.55	58.8	7.49	348	
3	Angola	119.0	5900.0	22.40	60.1	6.16	3530	
37	Congo, Dem. Rep.	116.0	609.0	20.80	57.5	6.54	334	
25	Burkina Faso	116.0	1430.0	6.81	57.9	5.87	575	