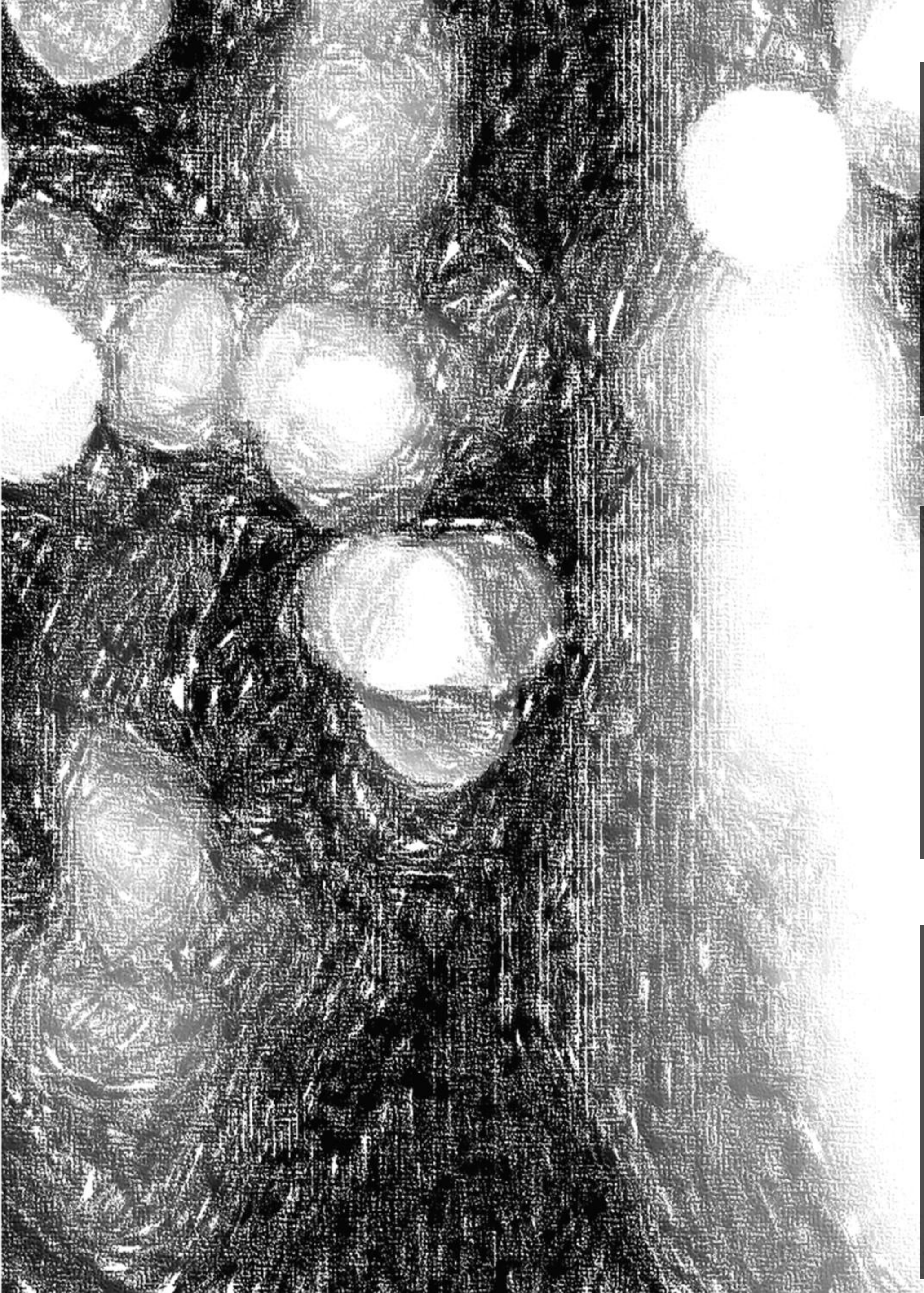
CLUSTERING ASSIGNMENT

Ideated by

Dibyajit Dhara



Objective : To categorize the countries using the socio-economic and health factors that determine the overall development of the country.

Team : Dibyajit Dhara

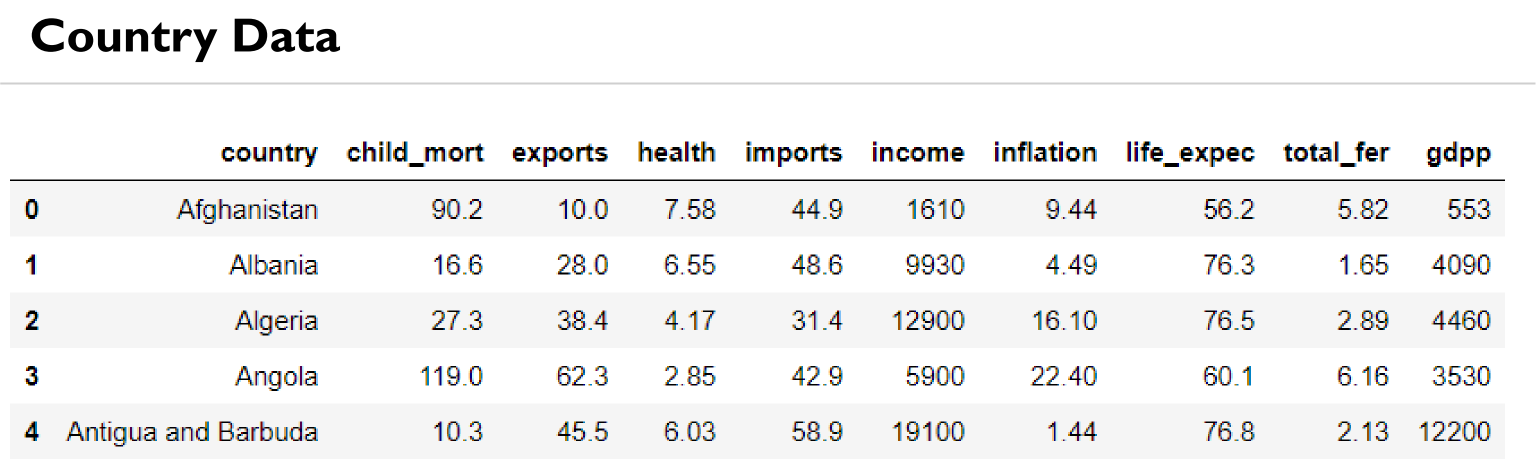
Keys to Decide :

Gdpp : The GDP per capita.

child_mort : Death of children under 5 years of age per 1000 live births

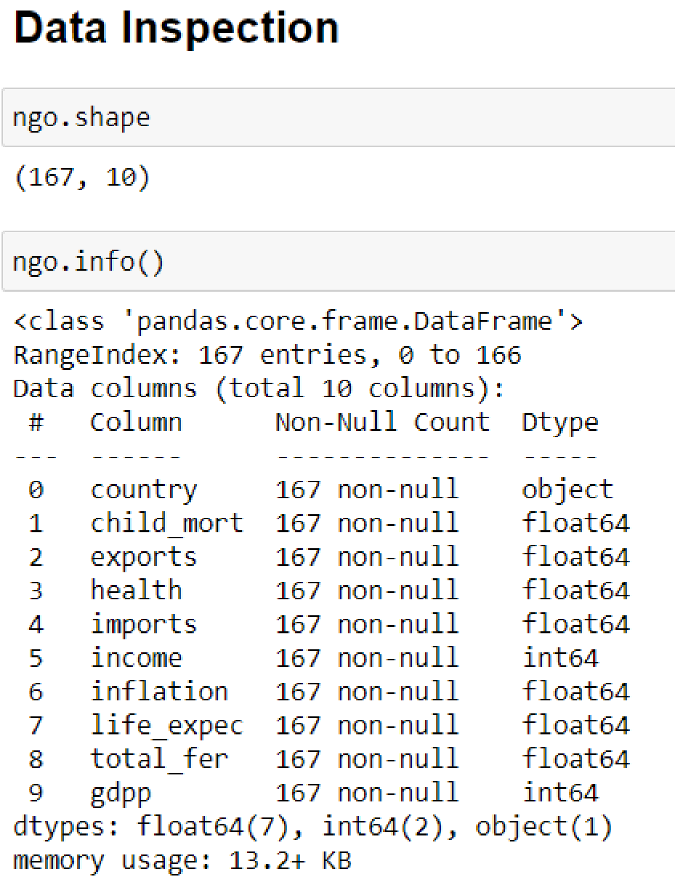
income : Net income per person

INFORMATION ABOUT DATA



Country 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



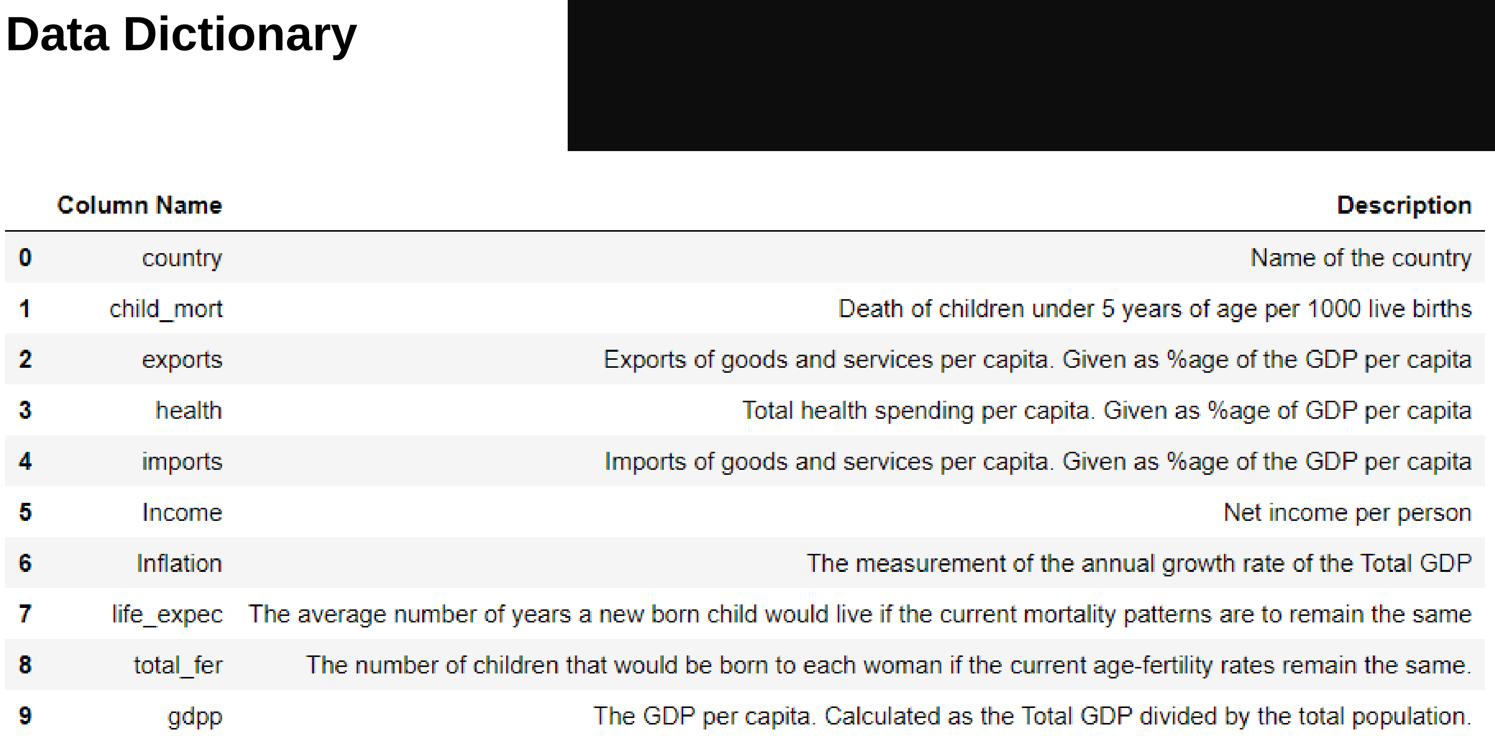
Data Inspection

```
ngo.shape
```

```
(167, 10)
```

```
ngo.info()
```

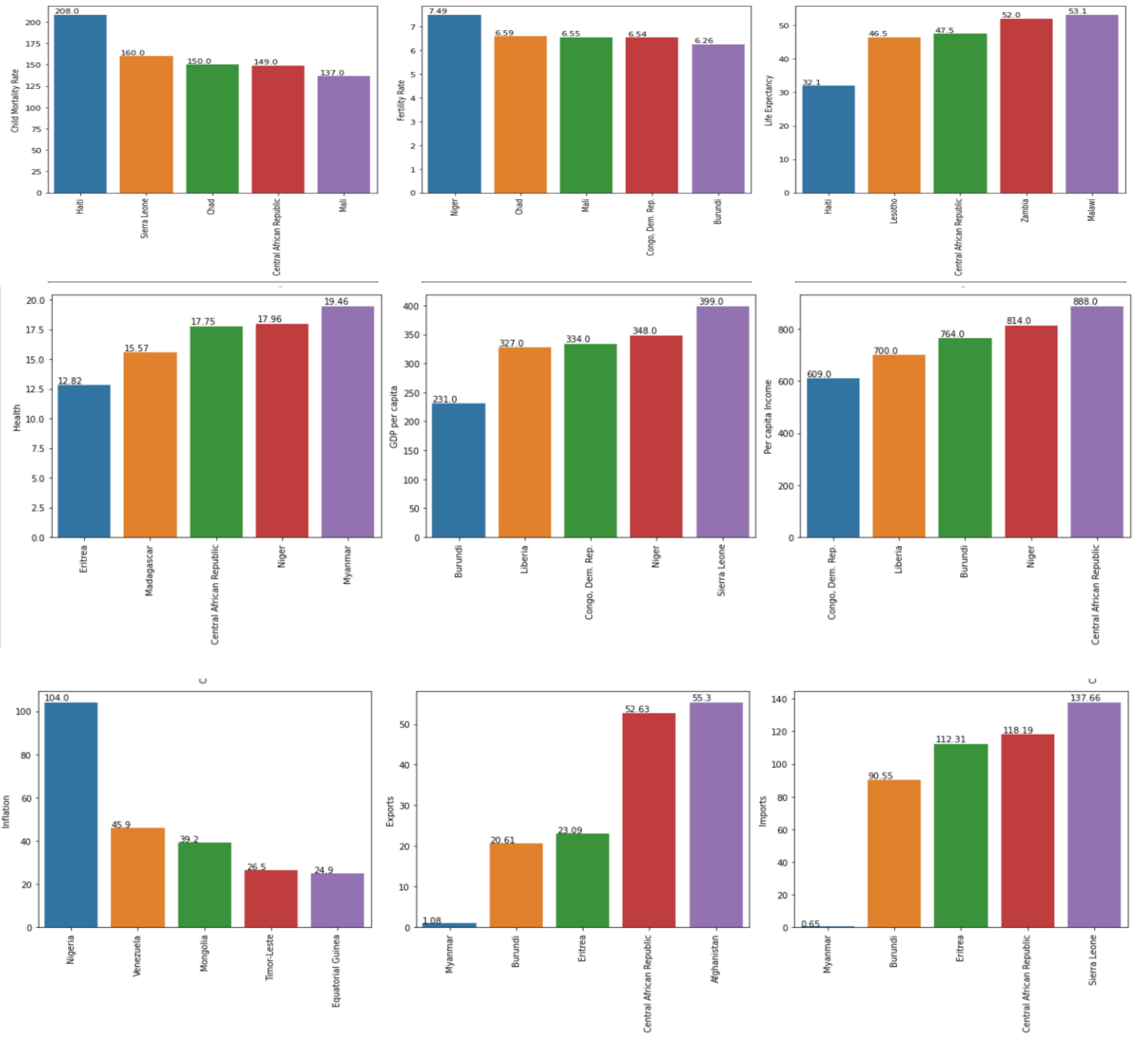
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 167 entries, 0 to 166
Data columns (total 10 columns):
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   country     167 non-null    object 
 1   child_mort  167 non-null    float64
 2   exports     167 non-null    float64
 3   health      167 non-null    float64
 4   imports     167 non-null    float64
 5   income      167 non-null    int64  
 6   inflation   167 non-null    float64
 7   life_expec  167 non-null    float64
 8   total_fer   167 non-null    float64
 9   gdpp        167 non-null    int64  
dtypes: float64(7), int64(2), object(1)
memory usage: 13.2+ KB
```



Data Dictionary

	Column Name	Description
0	country	Name of the country
1	child_mort	Death of children under 5 years of age per 1000 live births
2	exports	Exports of goods and services per capita. Given as %age of the GDP per capita
3	health	Total health spending per capita. Given as %age of GDP per capita
4	imports	Imports of goods and services per capita. Given as %age of the GDP per capita
5	Income	Net income per person
6	Inflation	The measurement of the annual growth rate of the Total GDP
7	life_expec	The average number of years a new born child would live if the current mortality patterns are to remain the same
8	total_fer	The number of children that would be born to each woman if the current age-fertility rates remain the same.
9	gdpp	The GDP per capita. Calculated as the Total GDP divided by the total population.

LET’S VISUALIZE



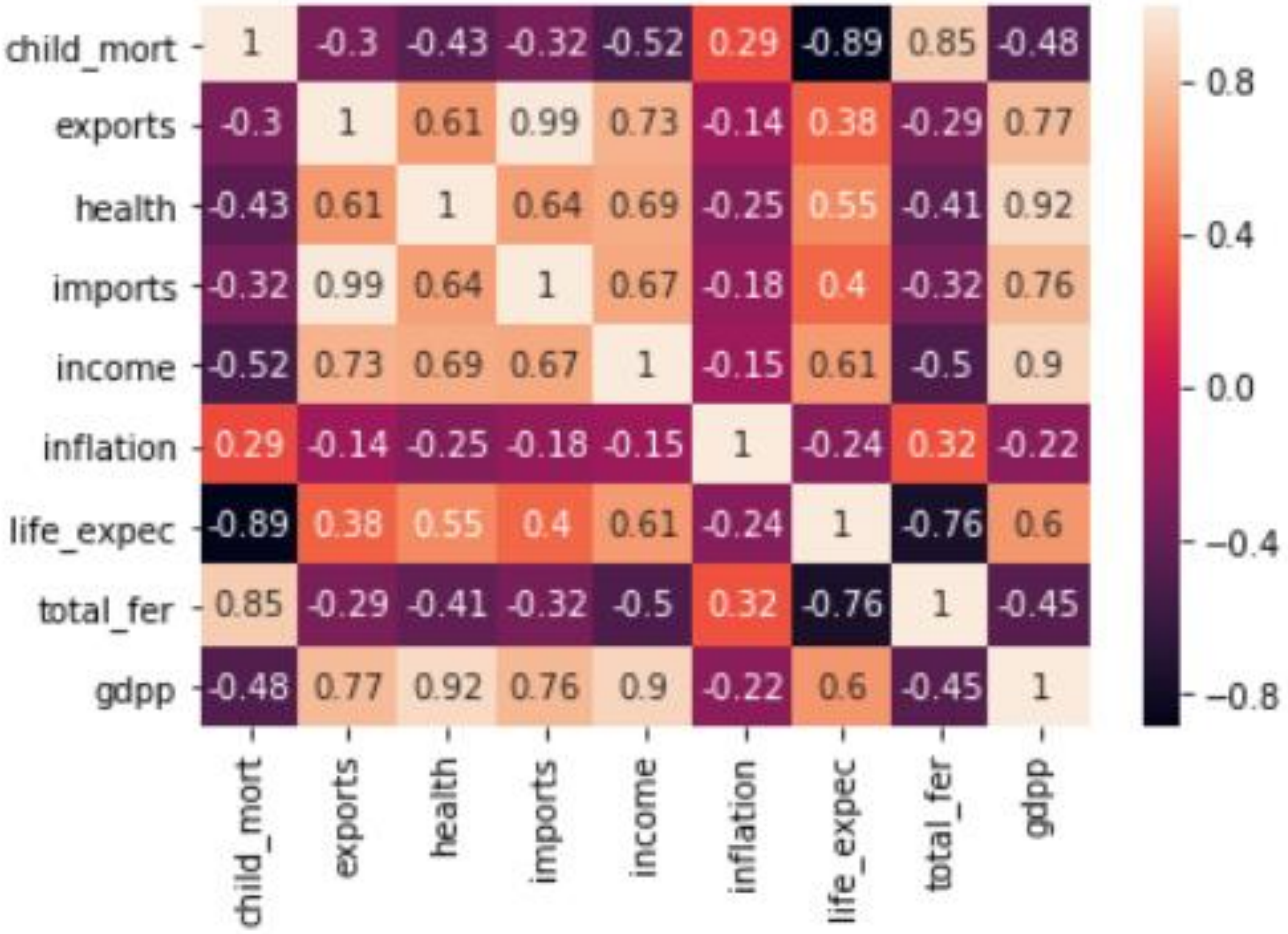
Insights:

- 1. Mainly African countries has high child mortality rate and very low income ,gdpp is also low
- 2. 'Haaiti' have maximum child mortality and minimum life expectancy
- 3. There is very low import and export in 'Mayanmar'
- 4. There is some countries which has a negative inflation rate (i.e seychelles,japan,ireland,check republic etc)
- 5. 'Qatar' has maximum income and high export rate
- 6. 'Luxemberg' is maximum in import,export and gdpp and also have a high income rate

LET’S UNDERSTAND DATA

Insights:

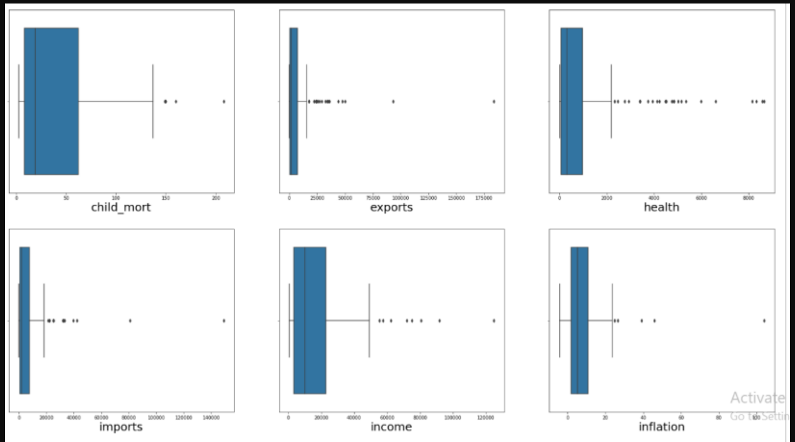
There are some features which are very positively co related like (child mortality and total fertility,import and export,income and export , gdp and export , gdp and health , gdp and income , gdp and import etc)
There are some features which are very negatively co related like (life expectency and child mortality , life expectency and total fertility)



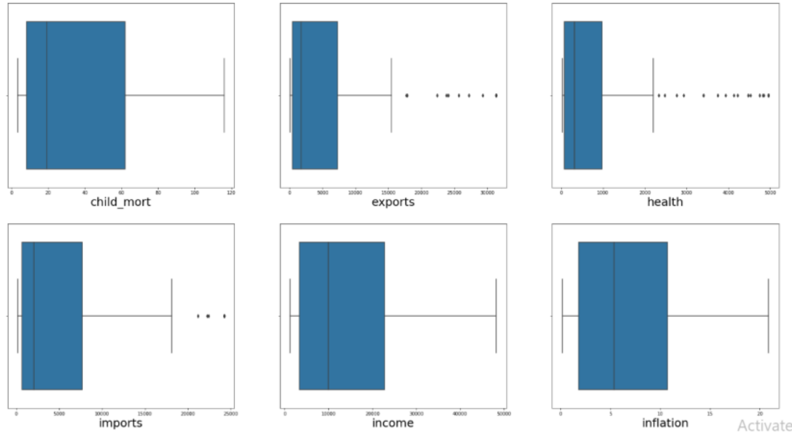
	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
child_mort	1	-0.3	-0.43	-0.32	-0.52	0.29	-0.89	0.85	-0.48
exports	-0.3	1	0.61	0.99	0.73	-0.14	0.38	-0.29	0.77
health	-0.43	0.61	1	0.64	0.69	-0.25	0.55	-0.41	0.92
imports	-0.32	0.99	0.64	1	0.67	-0.18	0.4	-0.32	0.76
income	-0.52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

CAPPING SCALING AND HOPKIN'S TEST

Before Capping



After Capping



1. Percentile Capping for outliers handling

- ✓ We use Winsorization technique at 5th and 95th percentile which implies values that are less than the value at 1st percentile are replaced by the value at 1st percentile, and values that are greater than the value at 99th percentile are replaced by the value at 99th percentile.

2. Hopkins Statistics Test

The Hopkins statistic (introduced by Brian Hopkins and John Gordon Skellam) is a way of measuring the cluster tendency of a data set.It acts as a statistical hypothesis test where the null hypothesis is that the data is generated by a Poisson point process and are thus uniformly randomly distributed. A value close to 1 tends to indicate the data is highly clustered, random data will tend to result in values around 0.5, and uniformly distributed data will tend to result in values close to 0.

3. Rescaling the Features

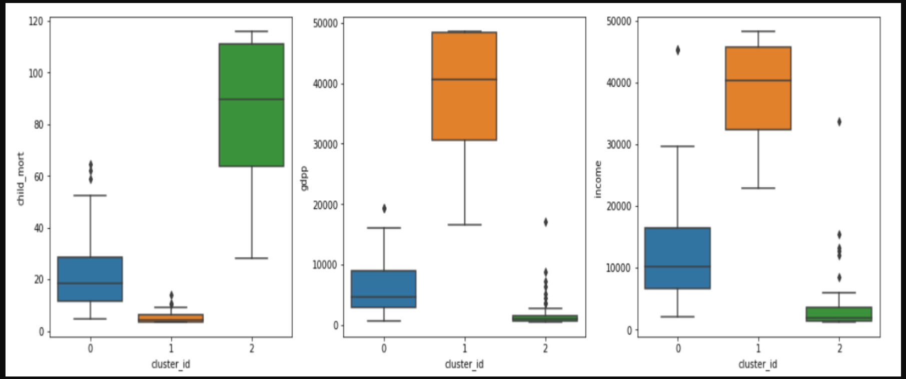
- ✓ We will use Standardisation Scaling which will convert our data where data's mean is 0 & sigma is 1

MODEL BUILDING

K- means Clustering

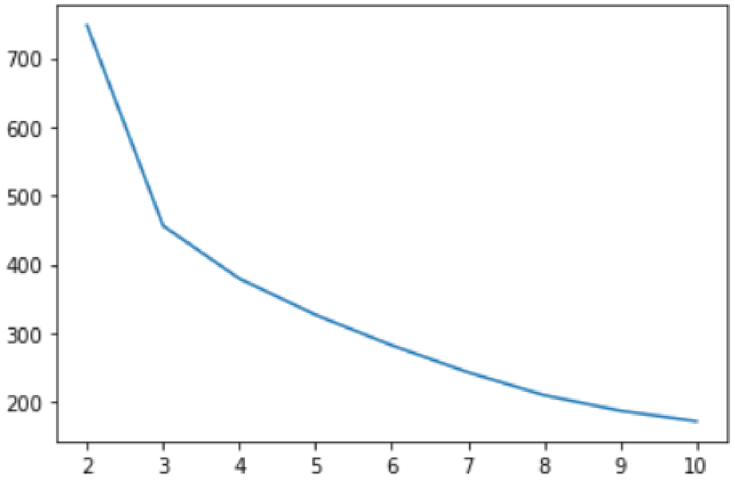
+ **Finding the Optimal Number of Clusters :** We opt for n= 3 clusters using Elbow curve, Silhouette Analysis & Business understanding

+ **Cluster Profiling for Clusters Formed**

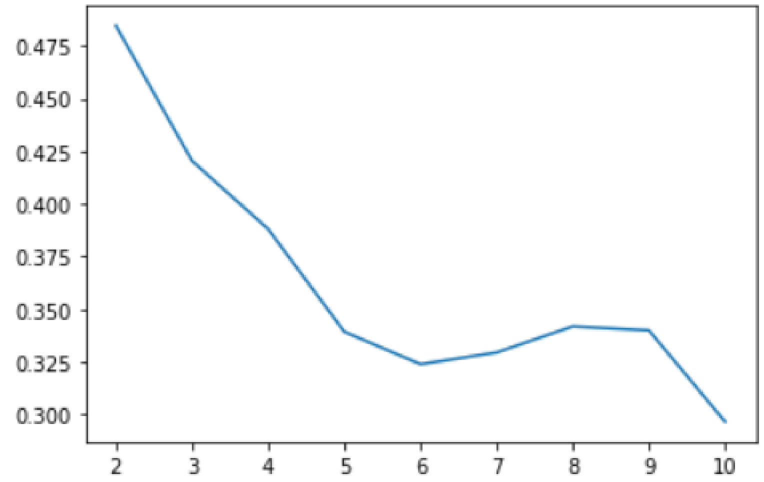


As per problem, Cluster 2 suits our requirement.

1. Elbow Curve



2. Silhouette Analysis



Final Result

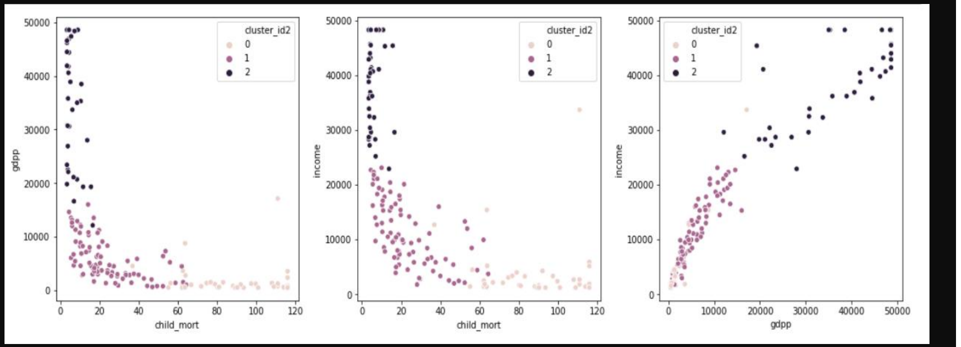
1. Sierra Leone
2. Niger
3. Mali
4. Central African Republic
5. Chad

MODEL BUILDING

Hierarchical Clustering

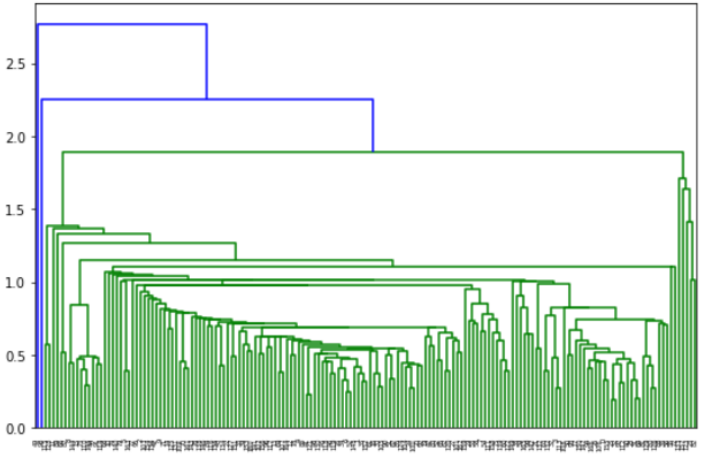
+ Finding the Optimal Number of Clusters : We opt for n= 3 clusters using Business understanding & visualization of Single linkage & Complete Linkage

+ Cluster Profiling for Clusters Formed

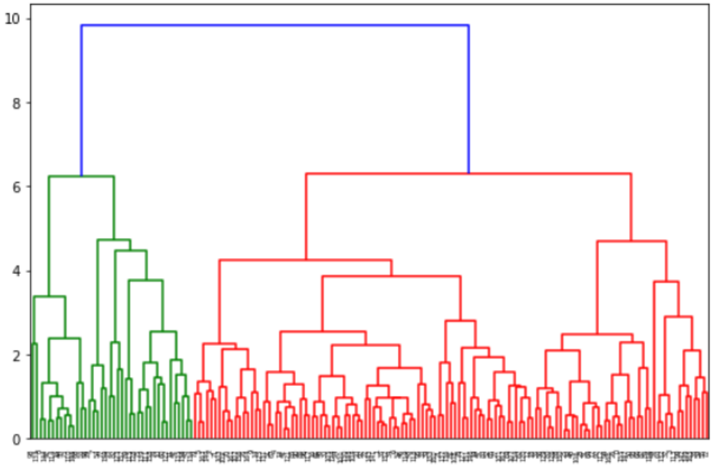


As per problem, Cluster 0 suits our requirement.

1. Single linkage



2. Complete Linkage



Final Result

1. Sierra Leone
2. Niger
3. Mali
4. Central African Republic
5. Chad

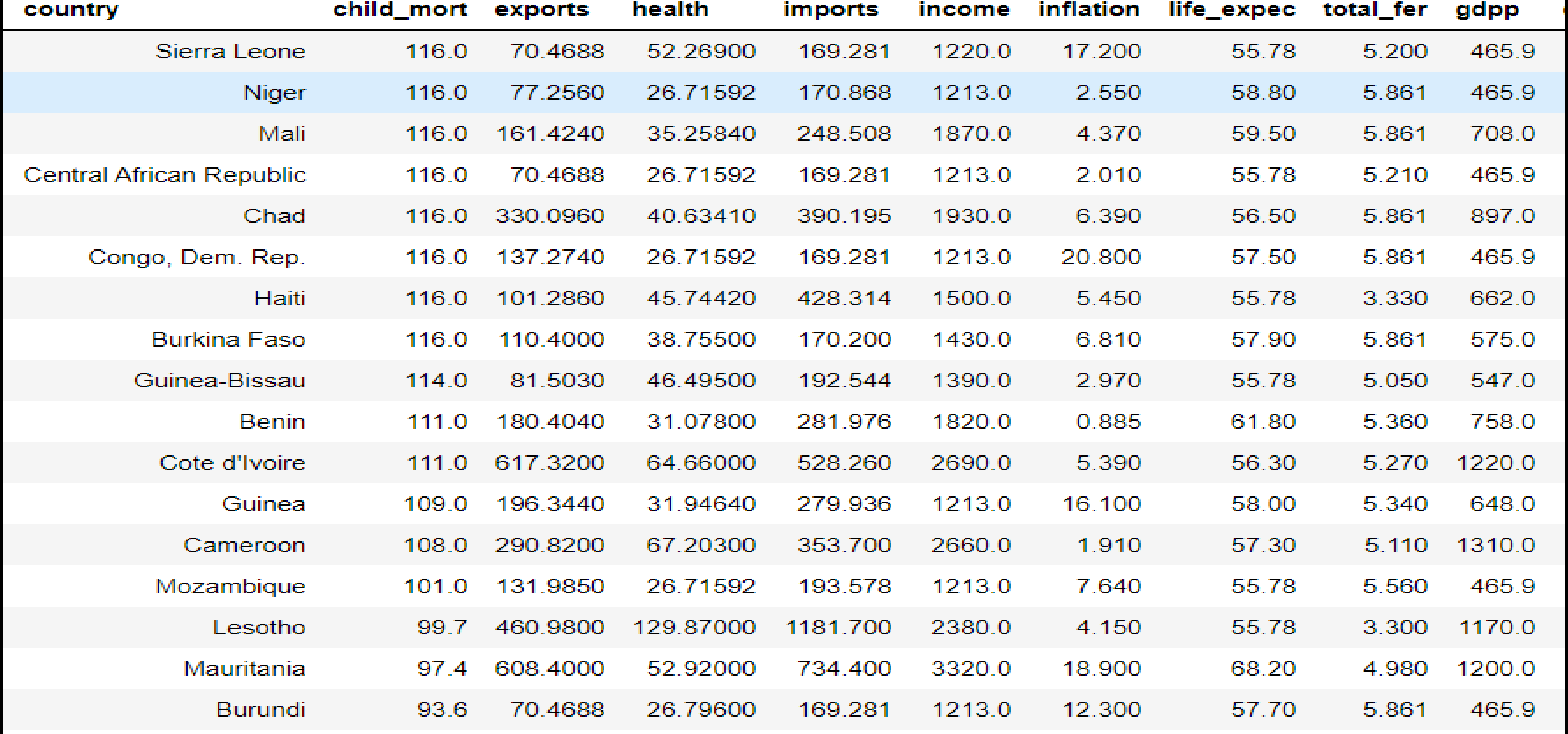
K- MEANS

V/S

HIERARCHICAL CLUSTERING

- ✓ We have analysed both K-means and Hierarchical clustering and found clusters formed are identical.
- ✓ The time complexity of K Means is linear i.e. O(n) while that of hierarchical clustering is quadratic i.e. O(n2)
- ✓ Difference in execution was not experienced as dataset was very small.
- ✓ K Means clustering requires prior knowledge of K i.e. no. of clusters we want to divide our data into. But, we can stop at whatever number of clusters we find appropriate in hierarchical clustering by interpreting the dendrogram
- ✓ So, we will proceed with the clusters formed by K-means and based on the information provided by the final clusters we will deduce the final list of countries which are in need of aid

From K-Means, we found that cluster 2 has the requirement where we need those country whose Child mortality rate is high, Income & GDPP is low.
From cluster 2 my best recommendation based upon low child mortality rate and high income and high gdpp

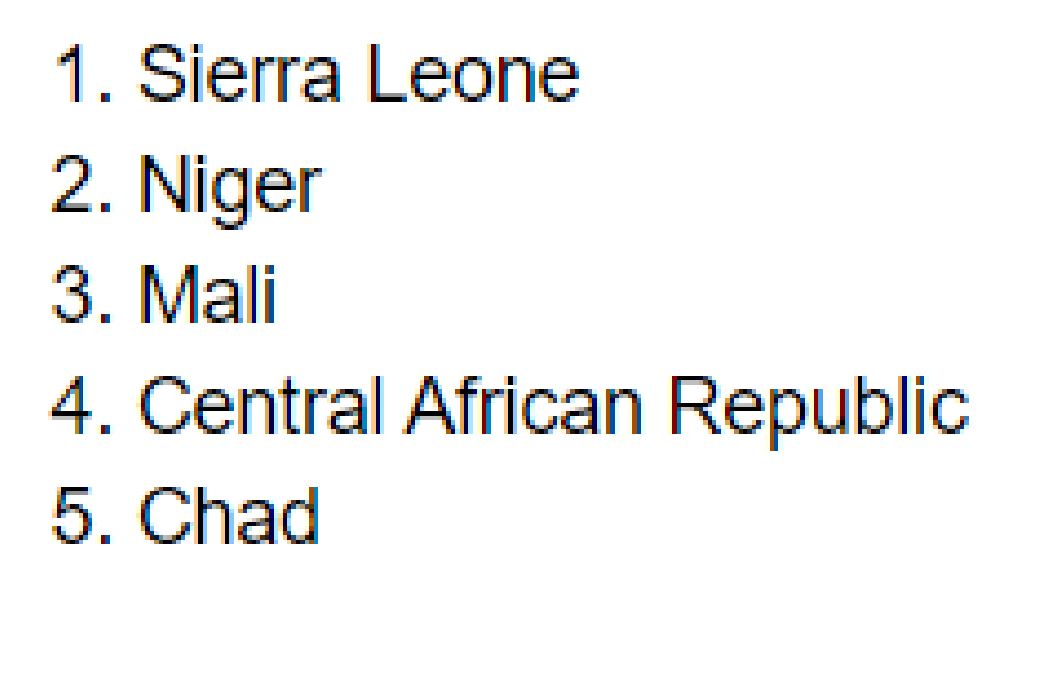


country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
Sierra Leone	116.0	70.4688	52.26900	169.281	1220.0	17.200	55.78	5.200	465.9
Niger	116.0	77.2560	26.71592	170.868	1213.0	2.550	58.80	5.861	465.9
Mali	116.0	161.4240	35.25840	248.508	1870.0	4.370	59.50	5.861	708.0
Central African Republic	116.0	70.4688	26.71592	169.281	1213.0	2.010	55.78	5.210	465.9
Chad	116.0	330.0960	40.63410	390.195	1930.0	6.390	56.50	5.861	897.0
Congo, Dem. Rep.	116.0	137.2740	26.71592	169.281	1213.0	20.800	57.50	5.861	465.9
Haiti	116.0	101.2860	45.74420	428.314	1500.0	5.450	55.78	3.330	662.0
Burkina Faso	116.0	110.4000	38.75500	170.200	1430.0	6.810	57.90	5.861	575.0
Guinea-Bissau	114.0	81.5030	46.49500	192.544	1390.0	2.970	55.78	5.050	547.0
Benin	111.0	180.4040	31.07800	281.976	1820.0	0.885	61.80	5.360	758.0
Cote d'Ivoire	111.0	617.3200	64.66000	528.260	2690.0	5.390	56.30	5.270	1220.0
Guinea	109.0	196.3440	31.94640	279.936	1213.0	16.100	58.00	5.340	648.0
Cameroon	108.0	290.8200	67.20300	353.700	2660.0	1.910	57.30	5.110	1310.0
Mozambique	101.0	131.9850	26.71592	193.578	1213.0	7.640	55.78	5.560	465.9
Lesotho	99.7	460.9800	129.87000	1181.700	2380.0	4.150	55.78	3.300	1170.0
Mauritania	97.4	608.4000	52.92000	734.400	3320.0	18.900	68.20	4.980	1200.0
Burundi	93.6	70.4688	26.79600	169.281	1213.0	12.300	57.70	5.861	465.9

Fund is $10M. We need to focus on few countries whose Child mortality rate is high, Income & GDPP is low.

Both the Methods suggested same countries.

Hence, we recommend following countries for the aid:



1. Sierra Leone
2. Niger
3. Mali
4. Central African Republic
5. Chad

- Thank you

UPGRAD