Clustering Assignment On HELP International NGO

Presented by

Pratima Kumari

Agenda

- Understanding of Problem Statement
- Target or Objective
- Reading & understanding the data
- Data Analysis
- Data Preparation- Handling Outliers
- Verification of Data, If fit to do Clustering- Hopkins test
- K-Cluster Algorithm for clustering Countries and Cluster Profiling
- Data Analysis
- Hierarchal Clustering
- Inferences

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 programs, they have been able to raise around $ 10 million. They wants to decide how to use this money strategically and effectively to the Countries that are in the direst need of aid.

Objective:

- Find the countries which needs to be focus on the most.
- Using K-Clustering and Hierarchal Clustering Algorithms , Needs to find out the major 5 countries which are in need and can spend money efficiently during Natural Calamities
- The countries are required based on factors with GDPP , Income and Child Mortality. Also can look for health spending, total fertility and life expectancy

Reading And Understanding Data

Here we can see , we have different factors to measure the Countries possibilities who could bein direst need: These are follows:

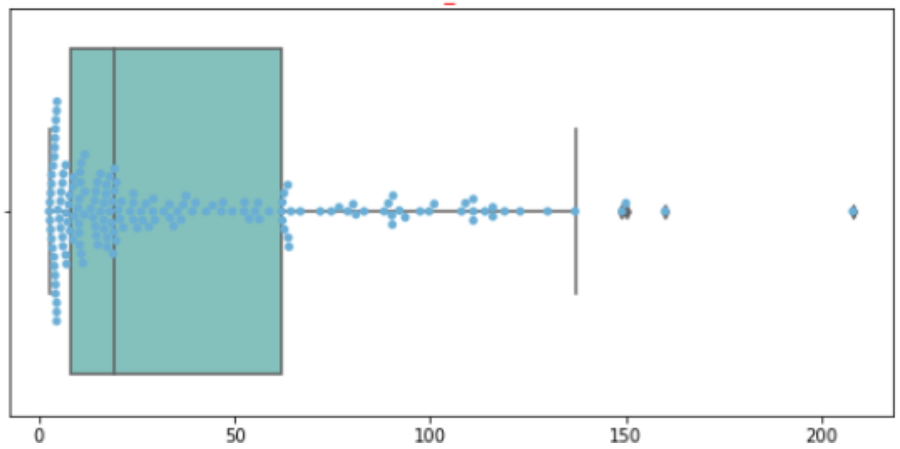
- Country
- Child Mortality
- Exports
- Imports
- Health
- Income
- Inflation
- Life expectancy
- Total fertility
- Gdpp

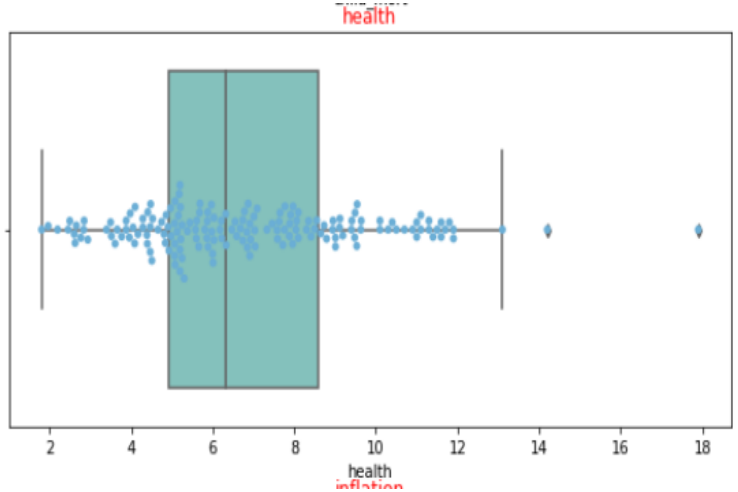
	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	Afghanistan	90.2	55.30	41.9174	248.297	1610	9.44	56.2	5.82	553
1	Albania	16.6	1145.20	267.8950	1987.740	9930	4.49	76.3	1.65	4090
2	Algeria	27.3	1712.64	185.9820	1400.440	12900	16.10	76.5	2.89	4460
3	Angola	119.0	2199.19	100.6050	1514.370	5900	22.40	60.1	6.16	3530
4	Antigua and Barbuda	10.3	5551.00	735.6600	7185.800	19100	1.44	76.8	2.13	12200

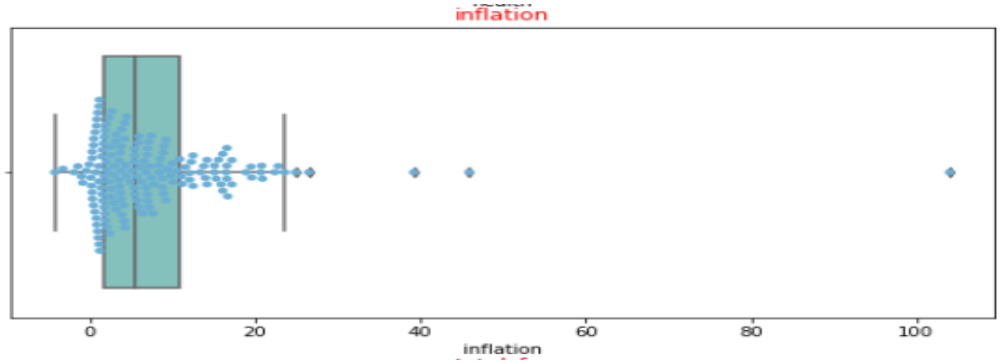
Overall Description of Data

	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	38.270060	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

- It can be seen the highest mortality rate is 208.
- It can be seen the total fertility maximum is 7.49
- In some countries life expectancy is 32
- These indicates that some countries needs to be taken into consideration for disaster or natural calamities.





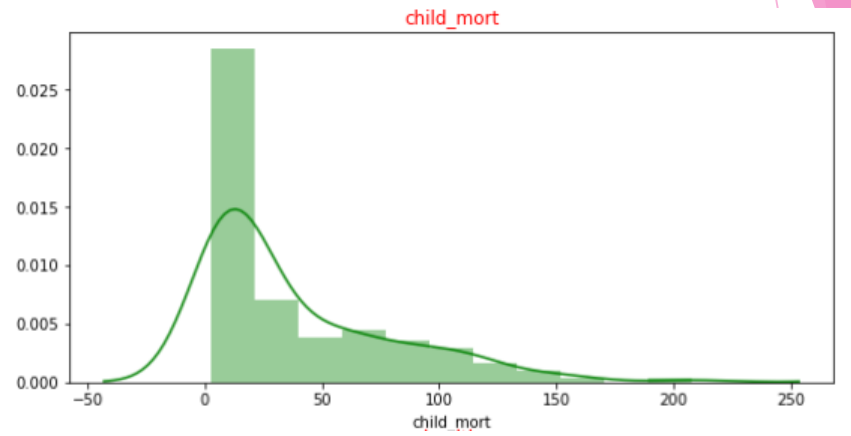


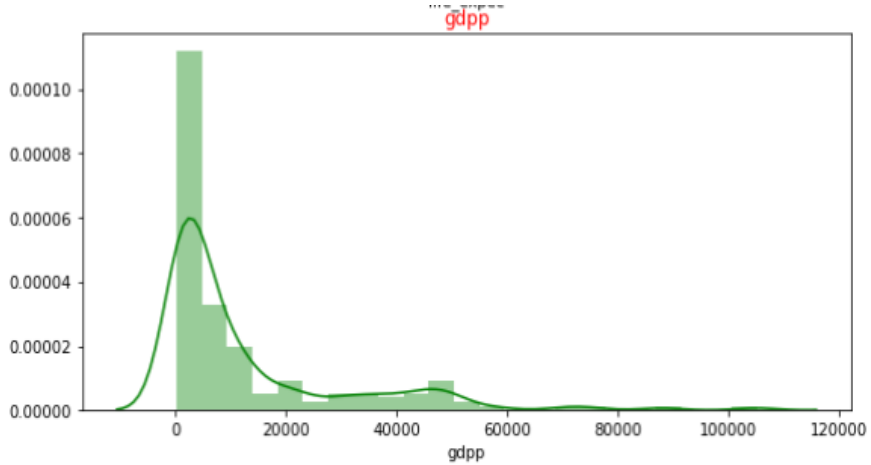
Exploratory Data Analysis

- - Here we can see, we have outliers for almost all the features
- - We can see outliers in child mort means there are few countries whose child mortality rate is much higher than others and hence it is really need aid in natural calamities
- - In health some countries have outlier. That could mean they have good health conditions for each person and life expectancy relatively should be more.
- - In inflation it can be seen few countries are below 0 means the GDP comparative to inflations are not increased.

EDA of child mortality and Gdpp

- It can be seen in next slide that child mortality are in the range of 0-50 , but can also be observed that some countries have child mortality more than 100.
- It can be observed the gdpp mainly lie in three different ranges between 0 and 20k, 40k-60k and more than 70k.

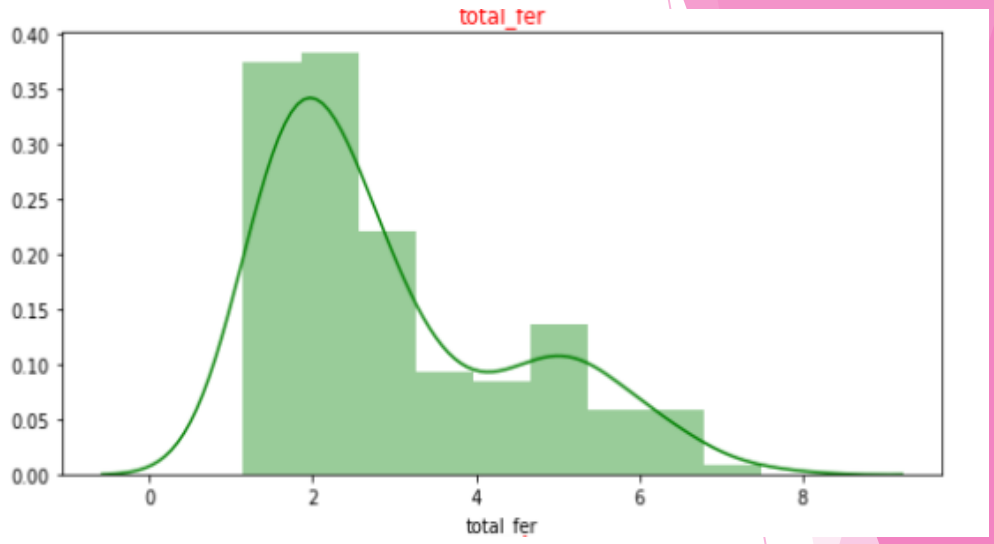


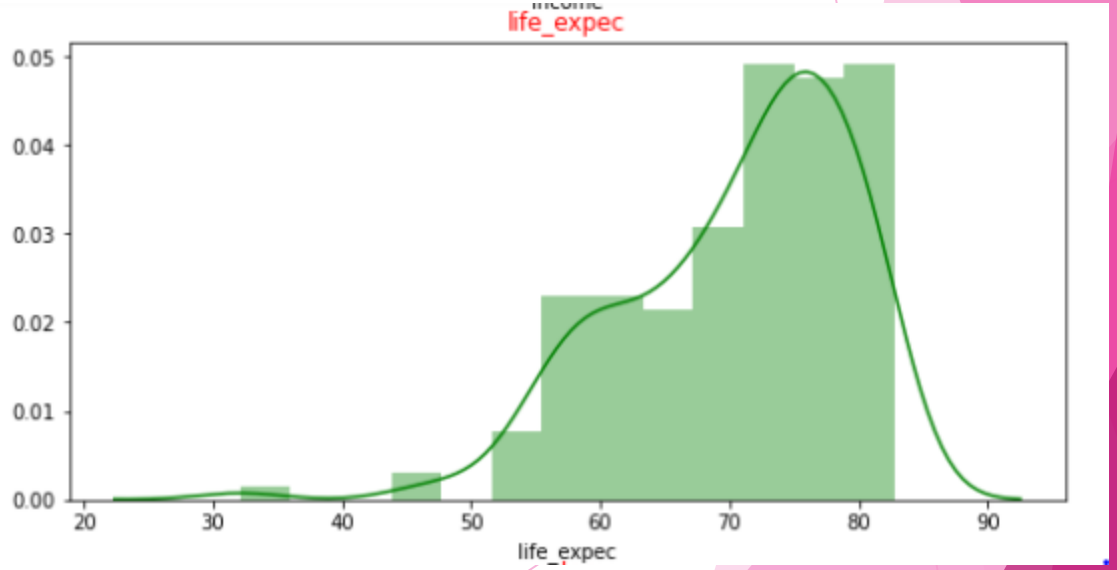


EDA for Total fertility and life expectancy

- In Total fertility , It is observed that most of countries are in the average of 2.5 which is good indication of being developed but some countries have total fertility as more than 6 . So if countries belong to this needs attention in terms of health and other financial factors.

- Also in life expectancy , with low expectancy , it is required more focus. But on an average many countries have more life expectancy.

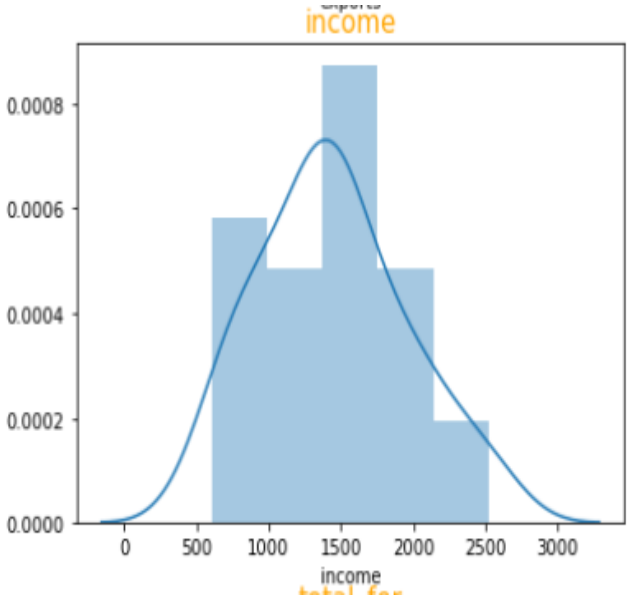


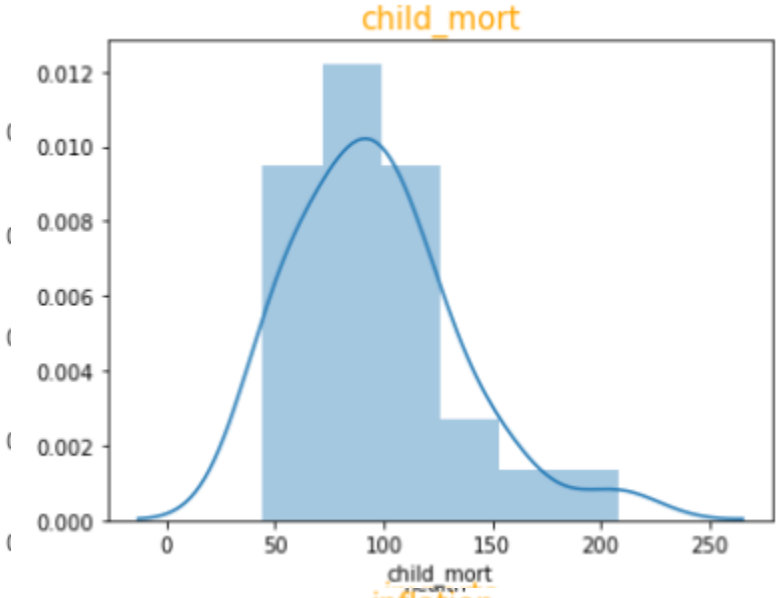


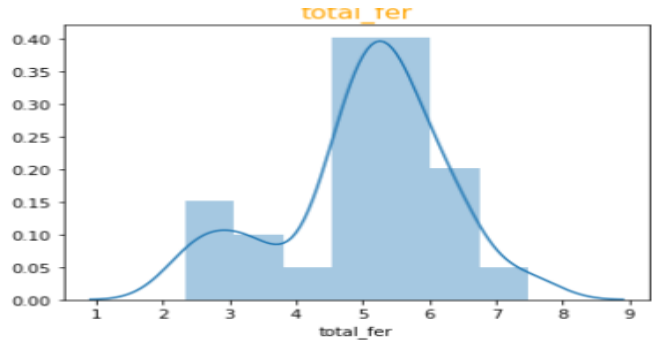
Countries having GDPP< 800

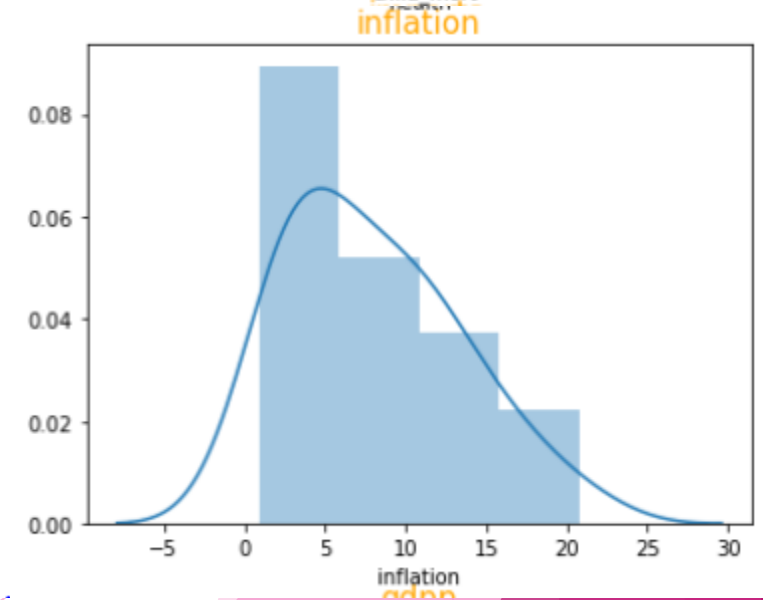
It can be observed that Countries with Low GDPP

- High Child Mortality
- Low Income
- Total Fertility is also very high
- Inflation is very less and import and exports are quiet less









Countries with Child Mortality more than 90

- The Top 5 countries having Highest Child Mortality Rate are:

  - Haiti - Sierra Leon - Chad - Central African Republic - Mali

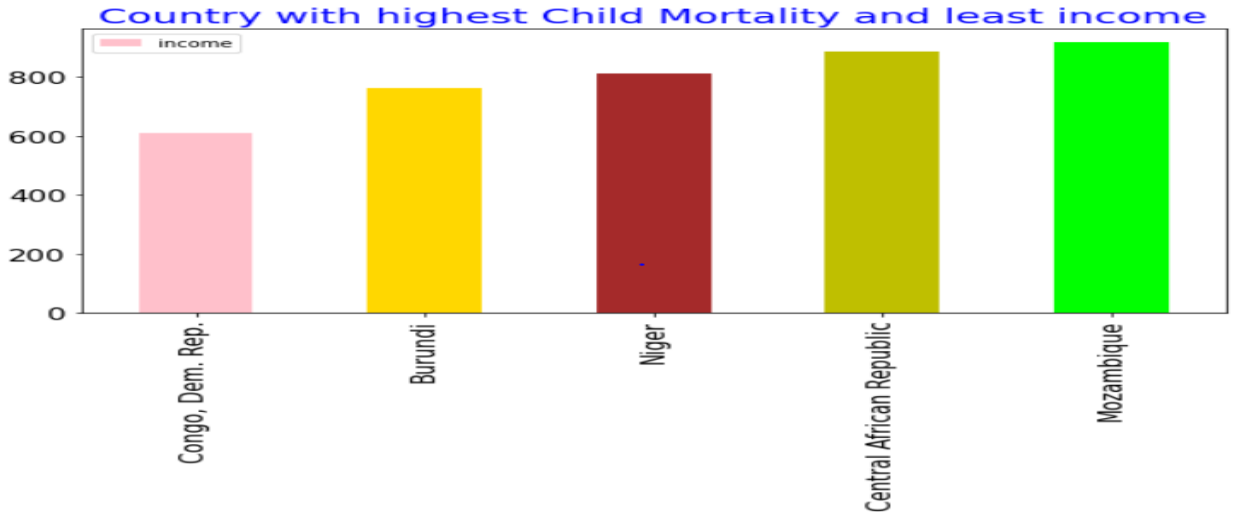
The Countries with High Child Mortality and Also having High Fertility are :

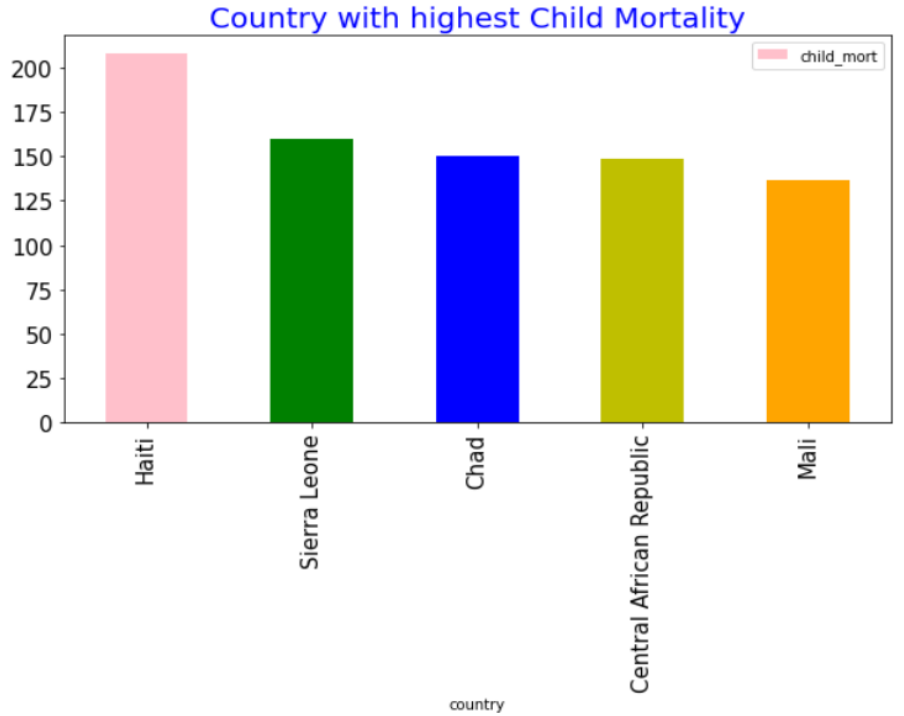
- Nige – Chad – Mali - Congo.Dem.Republic - Burundi

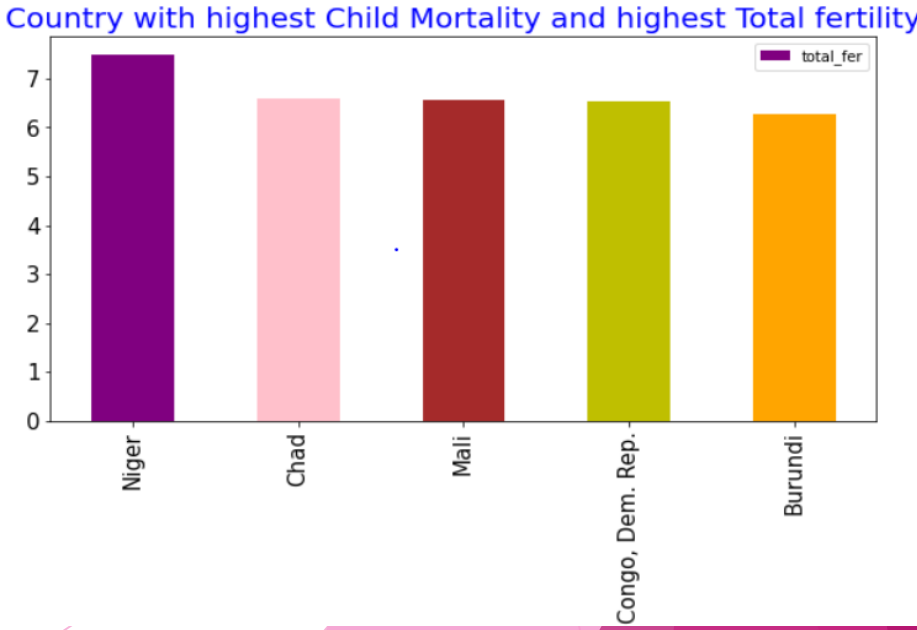
These countries required high focus as there is more need of aid during natural calamities

The Countries with high Child Mortality and very low income are

- Congo.Demo.Rep – Burundi – Niger - Central African Republic - Mozambique





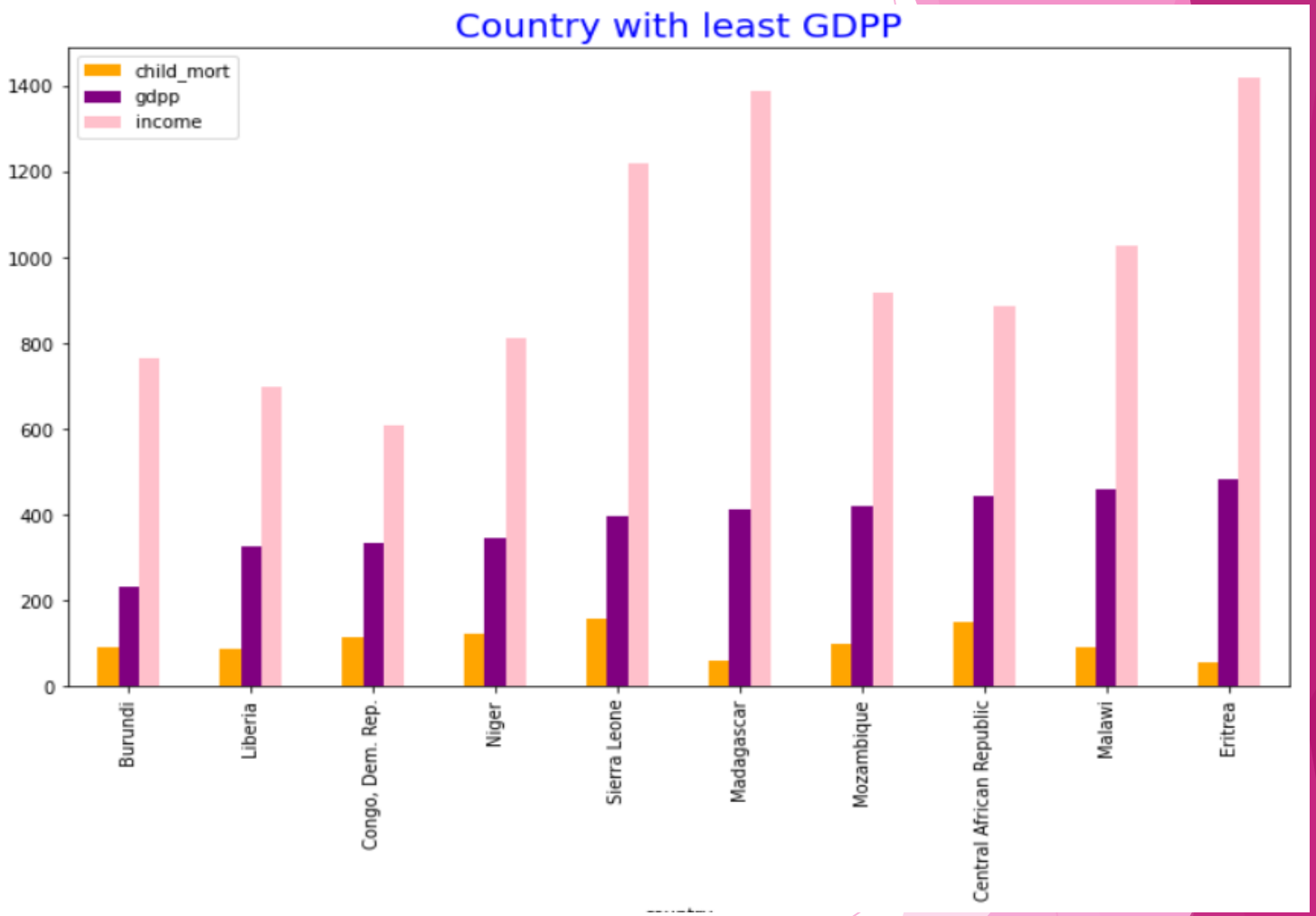


Child Mortality, GDPP, Income among countries with least GDPP

- The Below plot contains the countries which has very low GDPP and among them rate of child mortality and income

- It can be observed that Sierra Leone and Central African Republic has highest rate of Child Mortality.

- It can also be observed with the least income is for Congo. Dem.Republic and Liberria which has also quiet high Child Mortality



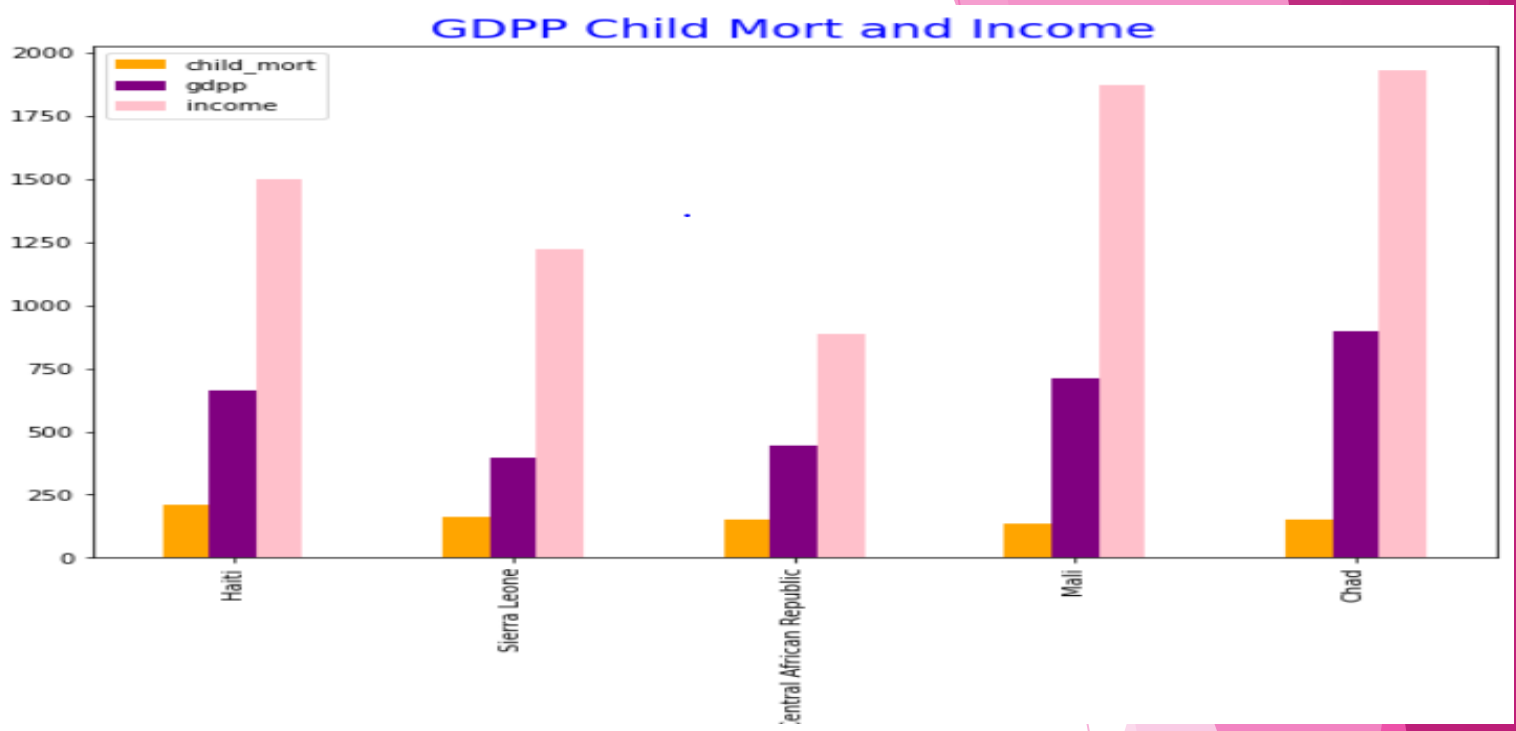
Countries with Less Net Exports[Export-Import] and Child Mortality more than 130

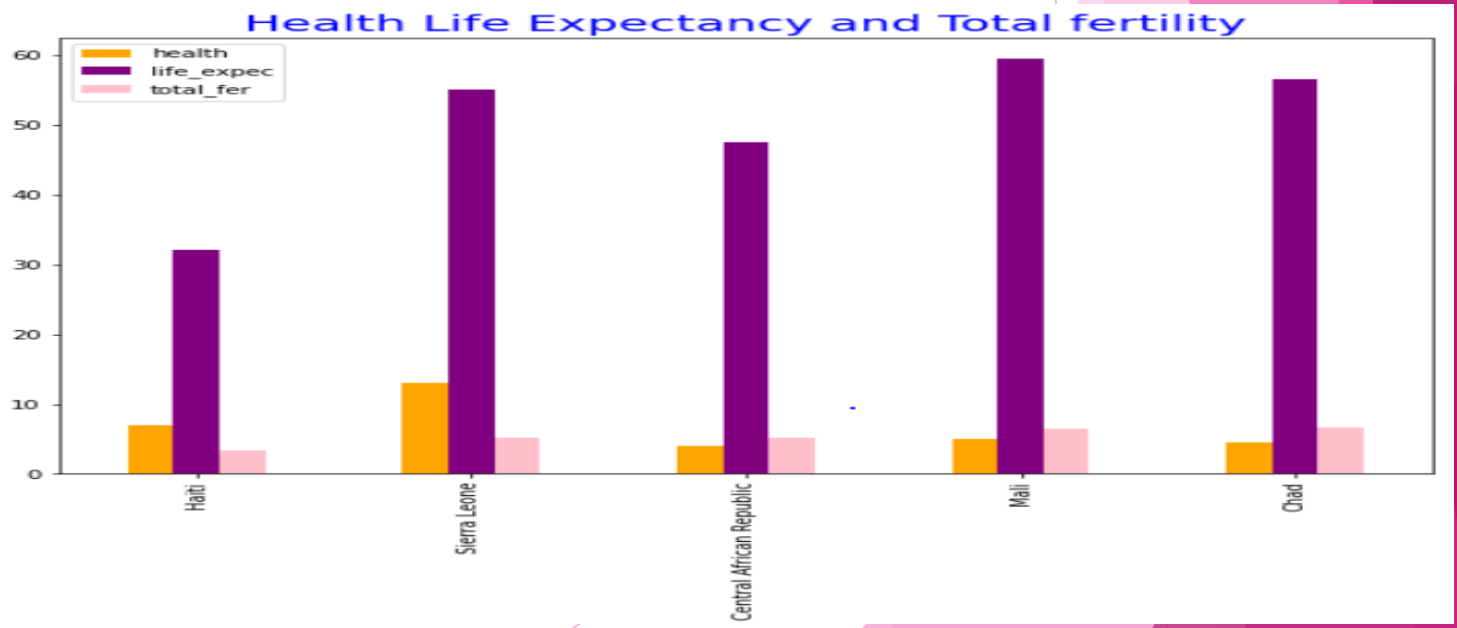
- The above plot is top countries who is having least net export(export - import) and child mortality more than 130.

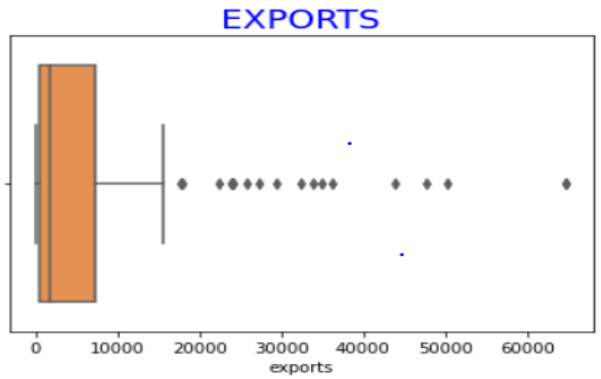
- It can be observed the countries with less net exports also have less GDPP compare to other developed countries as well as less income.

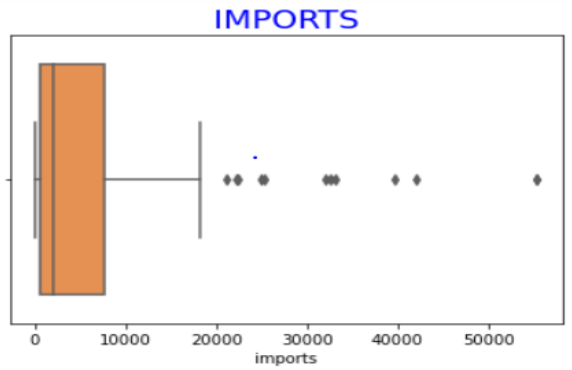
- Hence these countries required are under developed and required more attention during Natural Calamities.

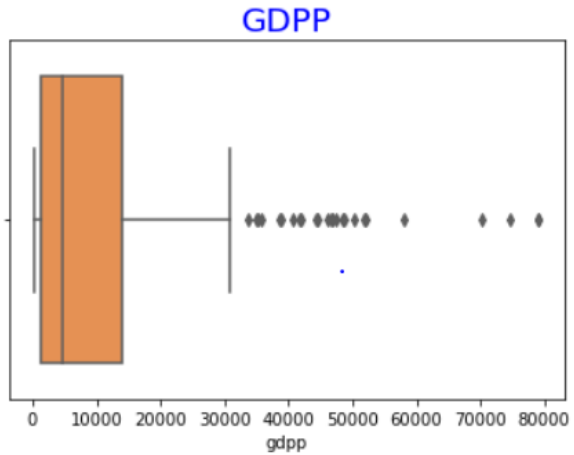
- Also can be observed that Haiti has very less life expectancy, health spending is very low. Total fertility is more as for Chad and Mali.

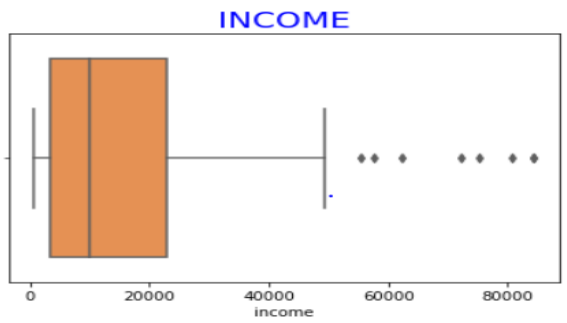


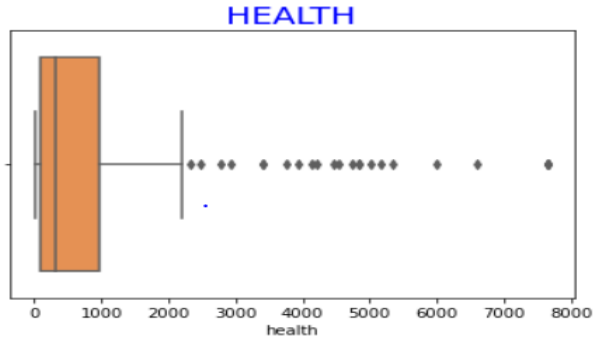












Handling Outliers

- The exports is capped from the highest values to 0.99 percentile.
- - The Imports is capped from the highest values to 0.99 percentile.
- The Income is capped from the highest values to 0.99 percentile
- - The health is capped from the highest values to 0.98 percentile
- - The GDPP is capped from the highest values to 0.99 percentile

HOPKINS TEST

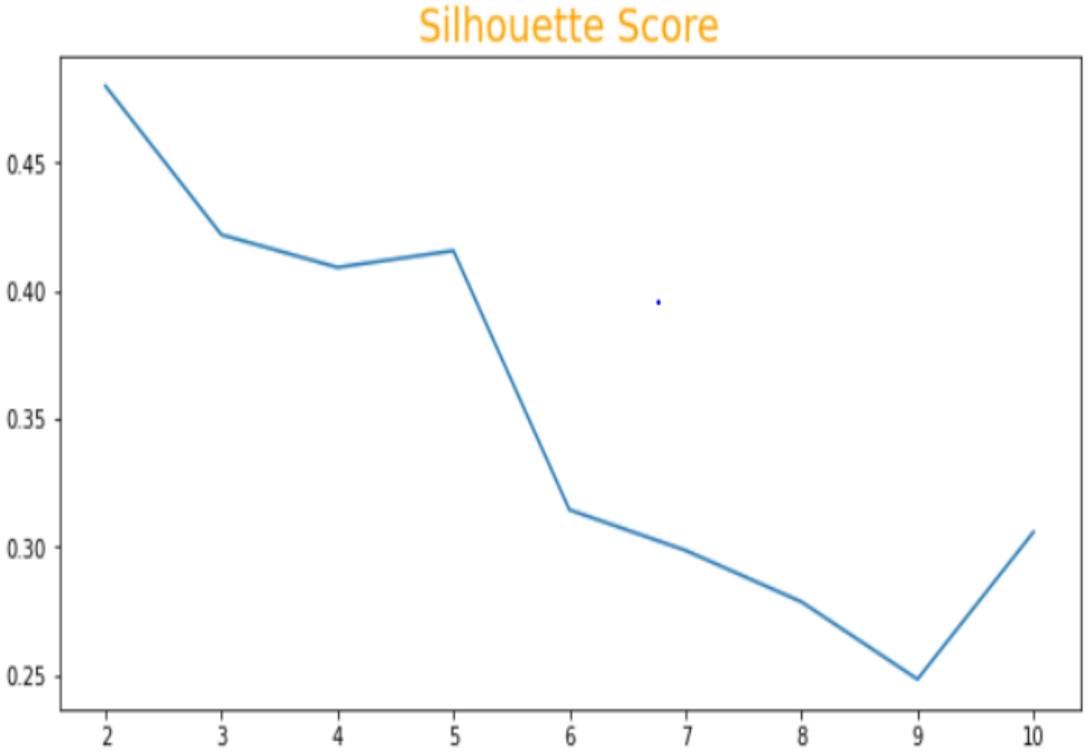
- The Data is fit to do Clustering and can be used for Hierarchal Clustering Also.
- The Value Obtained is :

  0.939097

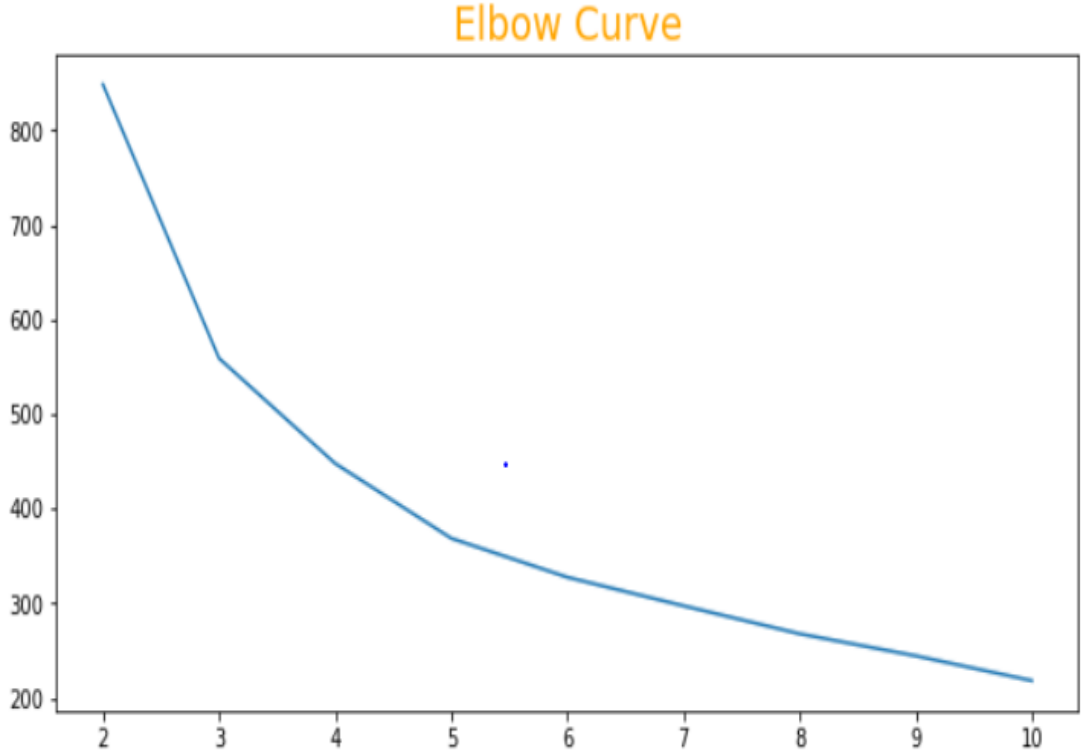
Clustering of Countries: Choosing K Value

- It can be observed that Silhouette Score is maximum for value 3 which indicates that we K can be choosen as 3 for Clustering
-It can be observed that The steep slope or depth is at point 3 , which indicated , K-Clustering can be done using K as 3.
Below plots can be observed and K cluster can be chosen as 3.

Silhouette Score

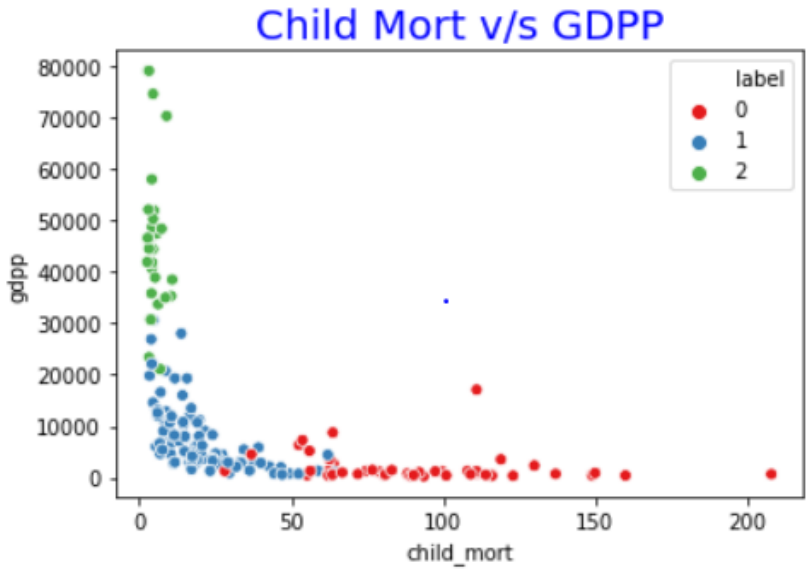


Elbow Curve



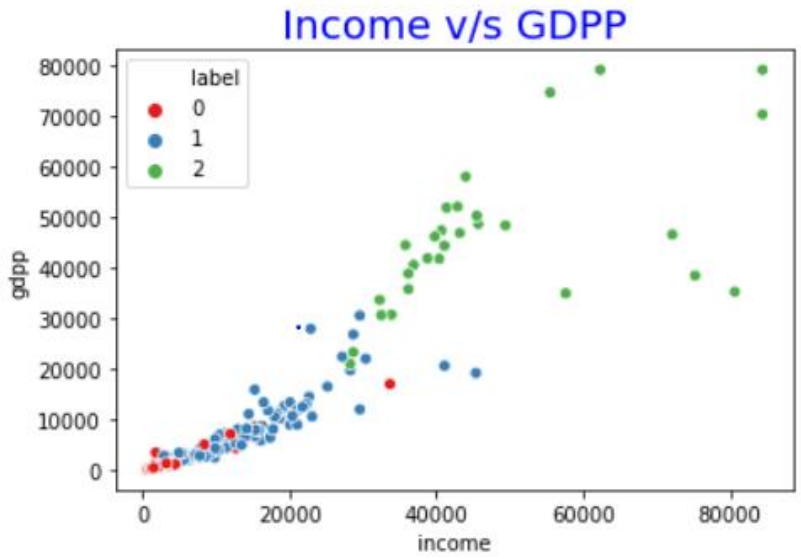
K-Clustering Algorithm

child mortality versus GDPP



- - It can be seen that as Child mortality increases , the gdpp for red colored cluster is getting decreased and it is very low.
- - The cluster with green color is having very low child mortality rate , having high GDP.
- - The cluster with blue is having moderate where GDP is moderately high and child mortality is however low.

Income versus GDPP

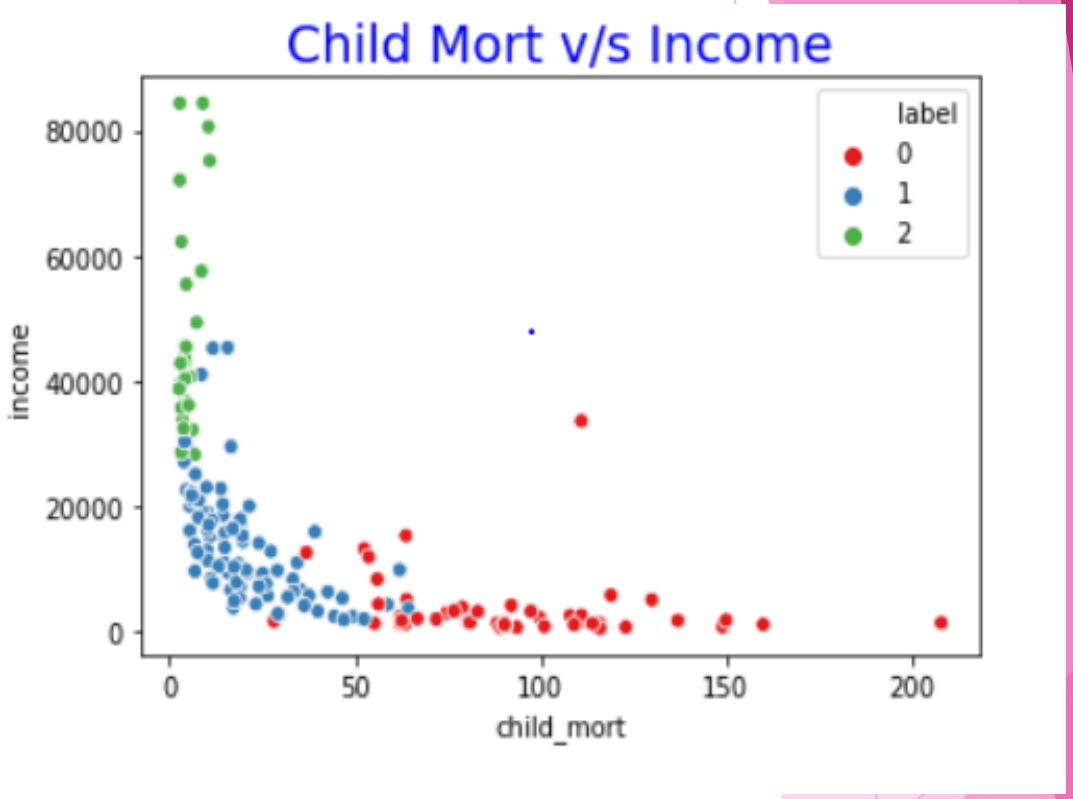


- - In the Above scatter plot, it can be observed that Cluster green where income is high , the GDPP is also high.
- - In Blue color cluster , the income is moderately high and GDPP is accordingly the same
- - whereas red cluster, the income is very low and same with GDPP

child mortality versus Income

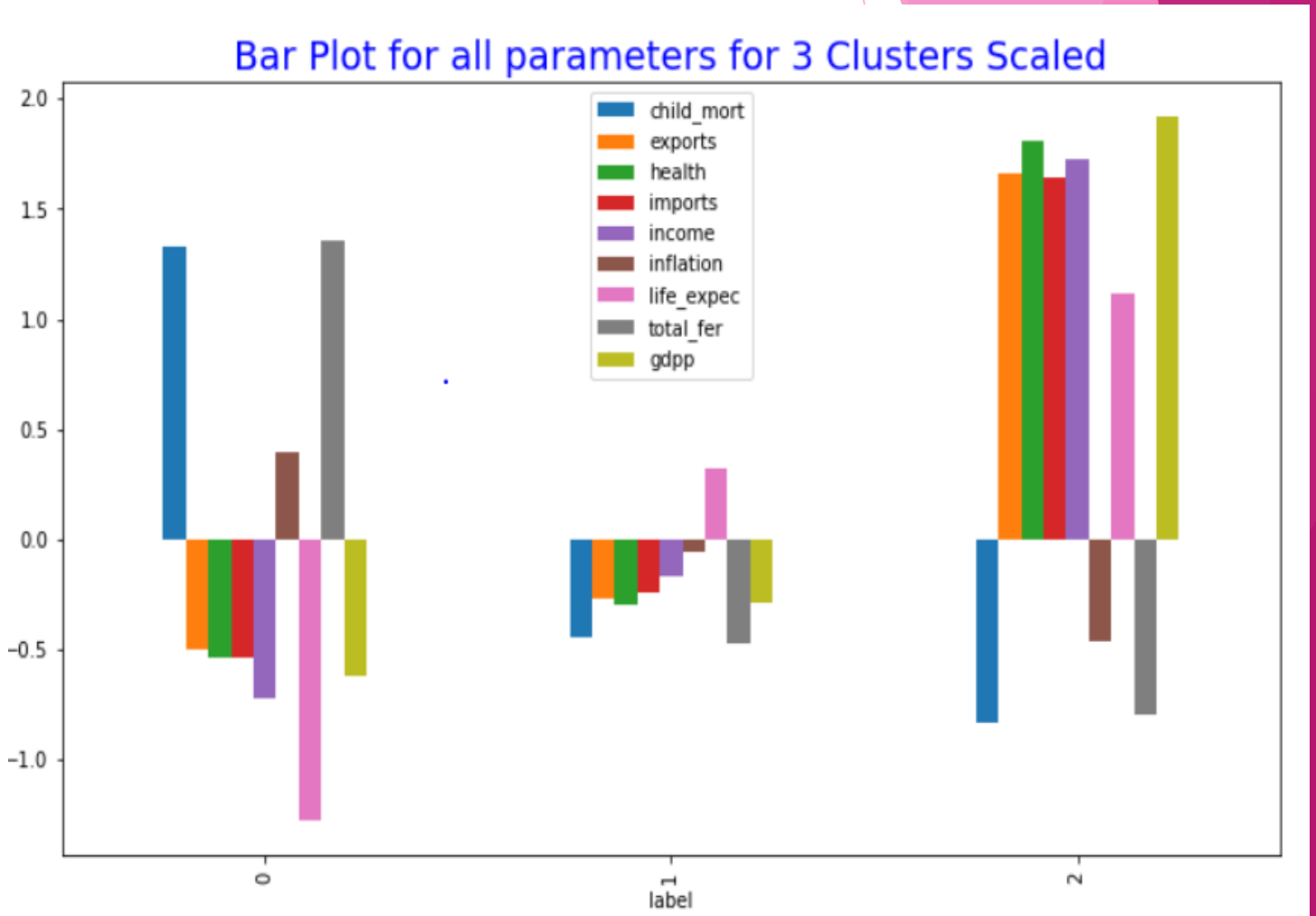
- It can be observed from above plot that , Where income is more that is green cluster, child mortality is not there.

- Whereas in red cluster , as child mortality increases , income per GDPP is then very low.



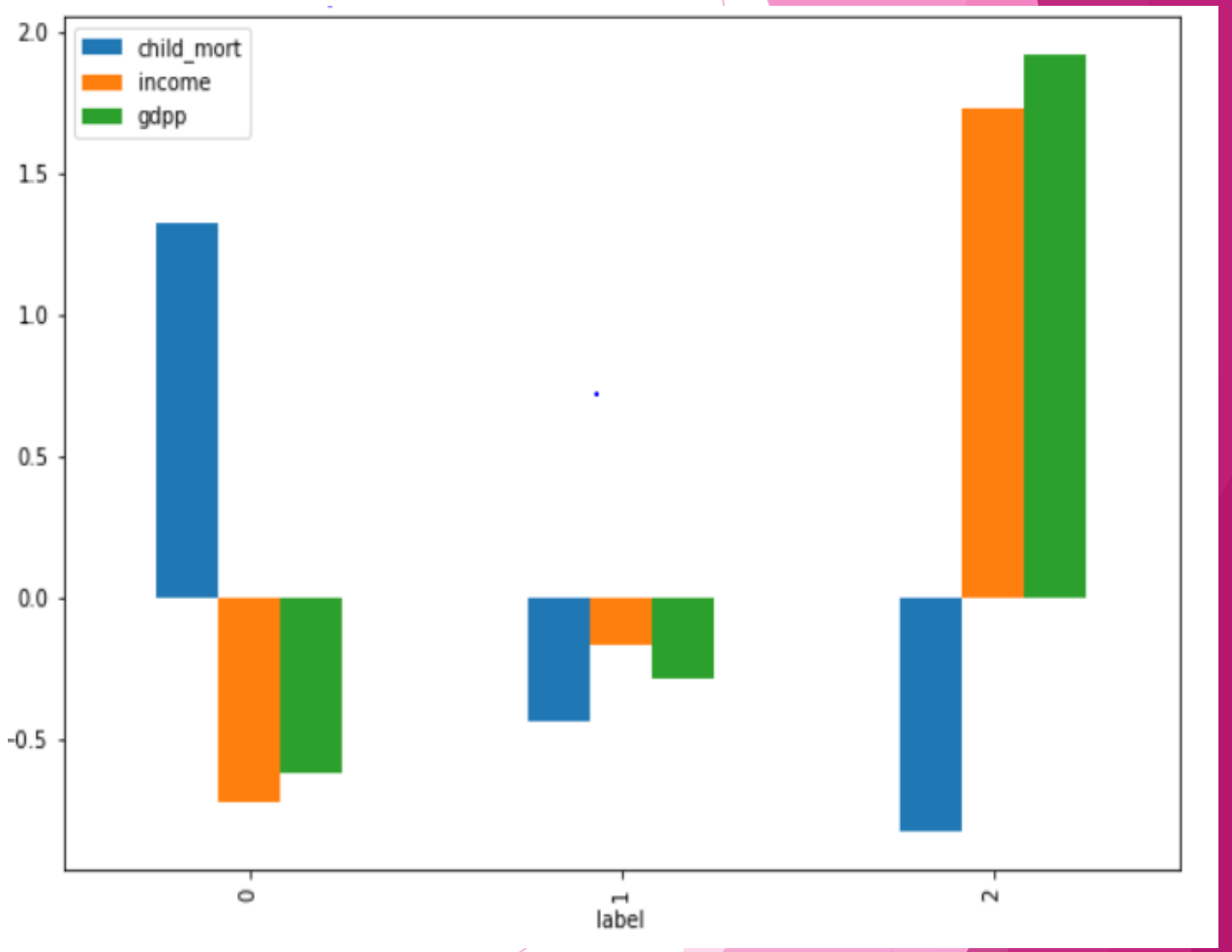
Cluster profiling

- It can be observed that for cluster 0 , The Child mortality , total fertility and inflation is high , whereas other important factors like GDP, income, exports, imports, life expectancy , health is verry low.
- - This means cluster 0 required high attention during calamities. The countries belonging to this group require direst aid.
- It can be observed for cluster 2 , that GDP , health, exports, income, life expectancy , imports, are high and also it can be seen that child mortality , total fertility is very low.
- - This means countries under this cluster are well developed and required less help and aid during calamities.



Analysing Cluster Data

- - The bar plot of GDPP, CHILD_MORT and INCOME.
- - It can be observed that cluster 0 has low income and low Gdpp but child mortality is very high.
- - This indicates that the countries belong to this cluster are under developed and in need of attention more focus during natural calamities.
- - Also it can been seen that luster 1 has low income and low Gdpp too but compare to cluster 0 it is not that low. Also child mortality is very low. This can be considered as developing countries.
- - Cluster 2 as very high income and gdpp and very low child mortality. The countries belongs to the cluster are well developed.

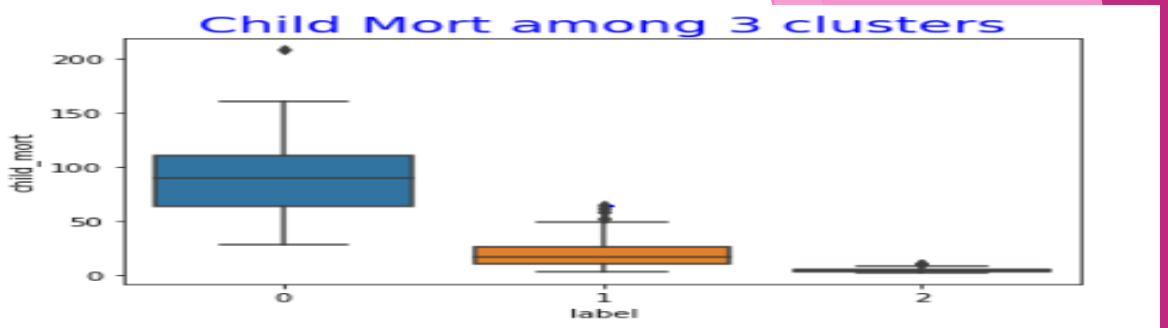


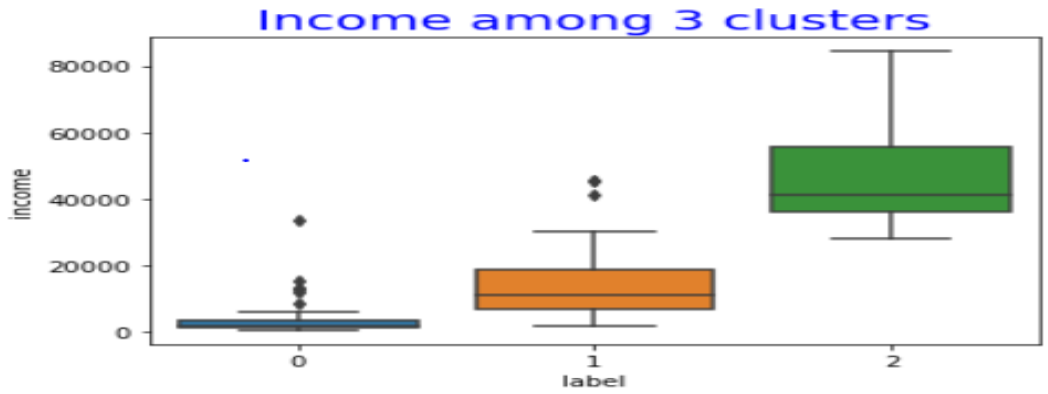
Child Mortality, Income and GDPP variation among Cluster

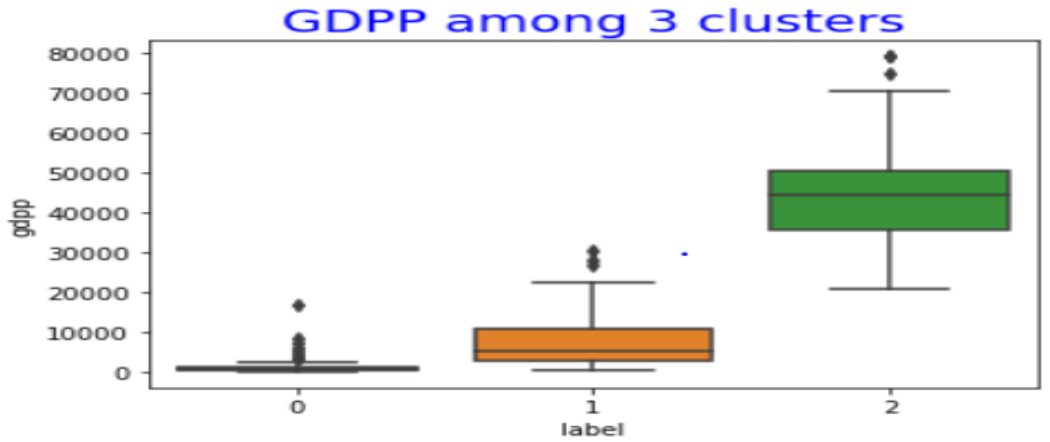
- It can be observed that ,the cluster 0 has very high child mortality rate

- It can be observed that gdpp is very low for cluster 0 compare to other clusters.

- It can be observed that for cluster 0 , it has very low income compare to other clusters







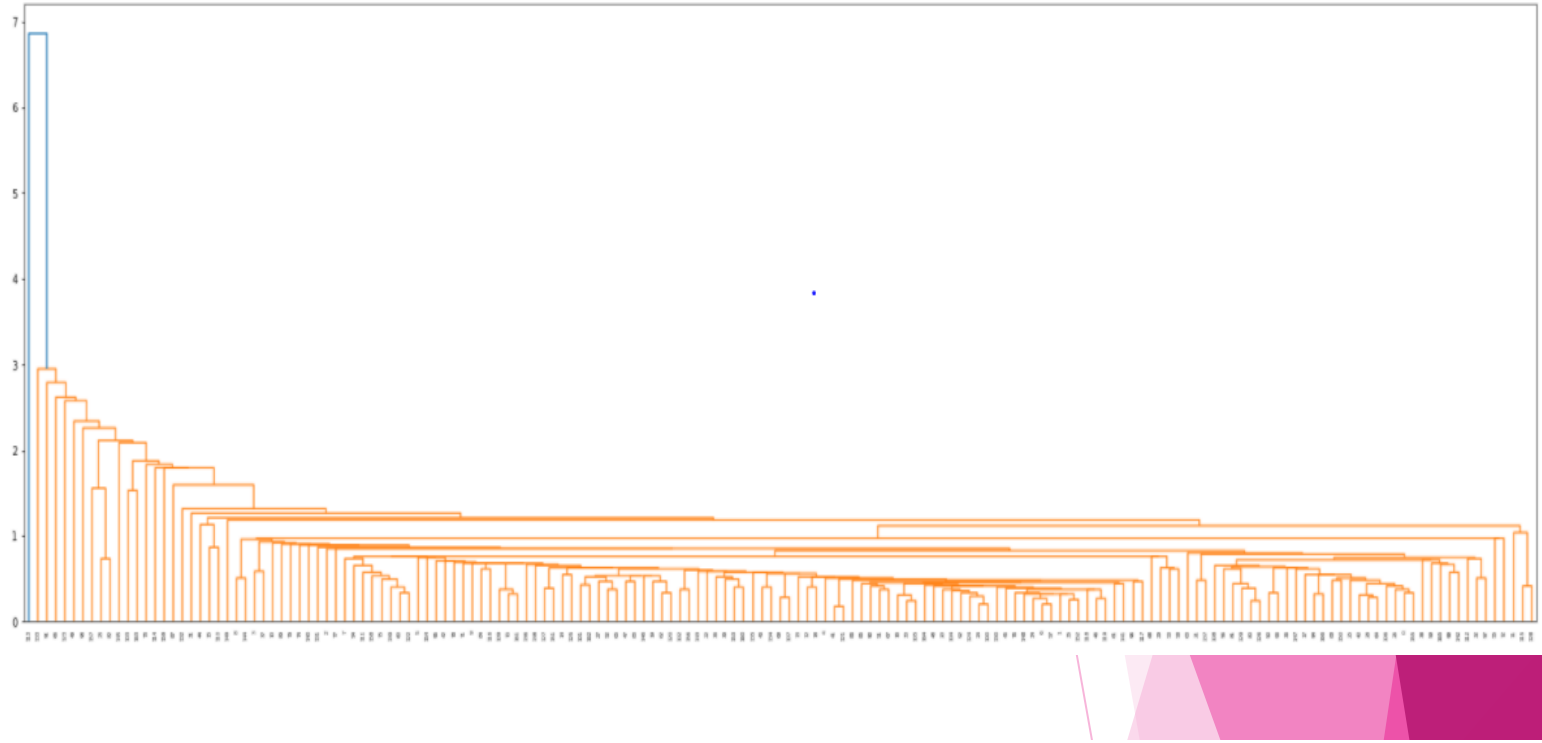
Countries In Direst Need

- From the Eucledian Distance , measure for GDPP , income and Child Mort from Centroid, above list of countries which has near to it.
- The parameters like low inome . low gdpp , high child mortality , life expectancy low, total fertiltiy more and health spending is less.
- - From above with different parameters that can be observed following countries required actions during Natural calamaties.
- - Congo, Dem. Rep.
- - Niger
- - Central Africa Republic
- - Sierra Leone
- - Haiti

	child_mort	income	gdpp	label	health	life_expec	total_fer	country	ED
10	116.0	609.0	334.0	0	26.4194	57.5	6.54	Congo, Dem. Rep.	704.901833
26	89.3	700.0	327.0	0	38.5860	60.8	5.02	Liberia	778.518424
5	93.6	764.0	231.0	0	26.7960	57.7	6.26	Burundi	804.342469
33	123.0	814.0	348.0	0	17.9568	58.8	7.49	Niger	894.493321
7	149.0	888.0	446.0	0	17.7508	47.5	5.21	Central African Republic	1005.540055
31	101.0	918.0	419.0	0	21.8299	54.5	5.56	Mozambique	1014.925063
28	90.5	1030.0	459.0	0	30.2481	53.1	5.31	Malawi	1132.076889
38	160.0	1220.0	399.0	0	52.2690	55.0	5.20	Sierra Leone	1294.234759
44	90.3	1210.0	488.0	0	37.3320	58.7	4.87	Togo	1308.633740
18	109.0	1190.0	648.0	0	31.9464	58.0	5.34	Guinea	1360.194969
27	62.2	1390.0	413.0	0	15.5701	60.8	4.60	Madagascar	1452.206440
36	63.6	1350.0	563.0	0	59.1150	64.6	4.51	Rwanda	1464.925016
19	114.0	1390.0	547.0	0	46.4950	55.6	5.05	Guinea-Bissau	1498.899782
14	55.2	1420.0	482.0	0	12.8212	61.7	4.61	Eritrea	1501.427618
4	116.0	1430.0	575.0	0	38.7550	57.9	5.87	Burkina Faso	1546.435941
9	88.2	1410.0	769.0	0	34.6819	65.9	4.75	Comoros	1609.350912
20	208.0	1500.0	662.0	0	45.7442	32.1	3.33	Haiti	1653.468168
45	81.0	1540.0	595.0	0	53.6095	56.8	6.15	Uganda	1653.767278
0	90.2	1610.0	553.0	0	41.9174	56.2	5.82	Afghanistan	1705.529171
16	80.3	1660.0	562.0	0	31.9778	65.5	5.71	Gambia	1755.216994

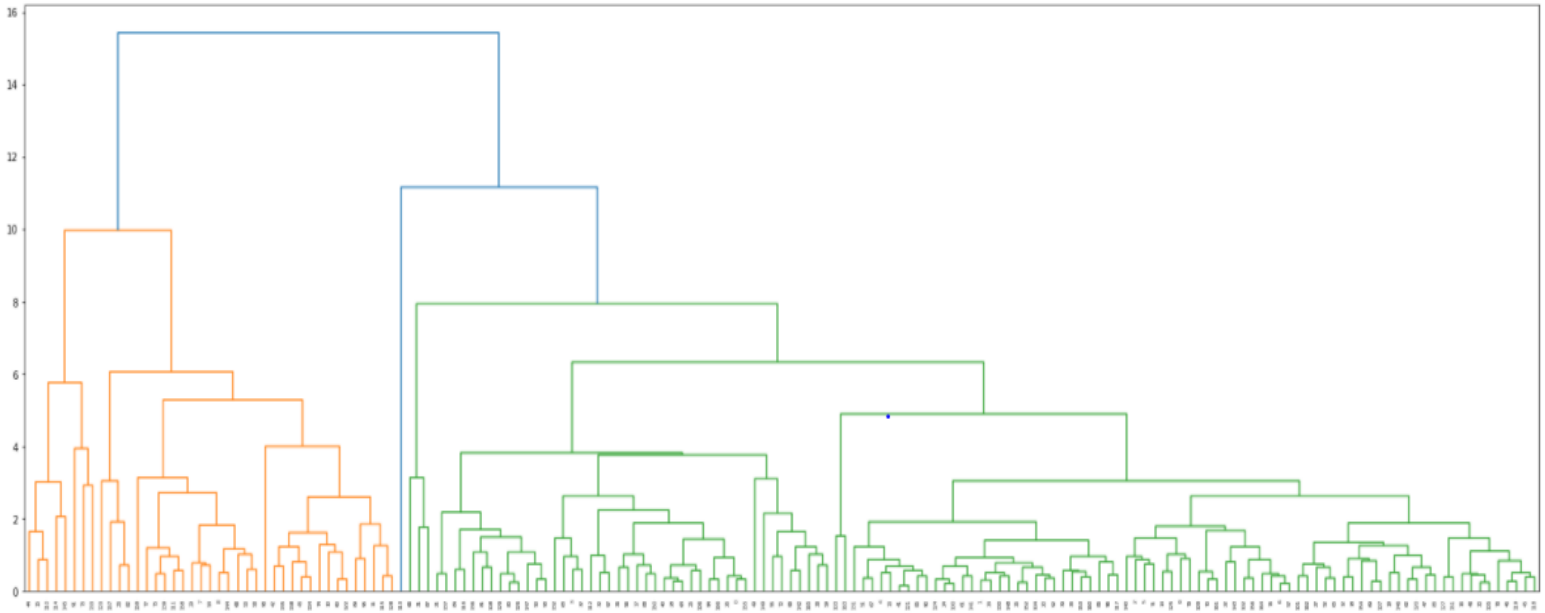
Hierarchal Clustering

Single Linkage

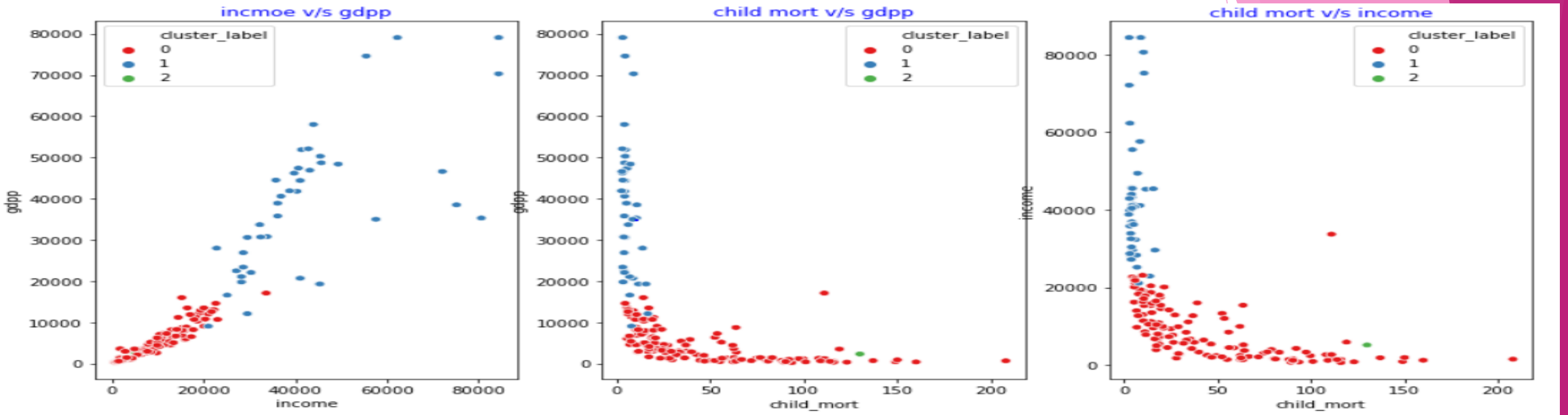


As it can be observed from the above plot , It is very difficult to analyze and cluster the countries into different groups.

Complete Linkage



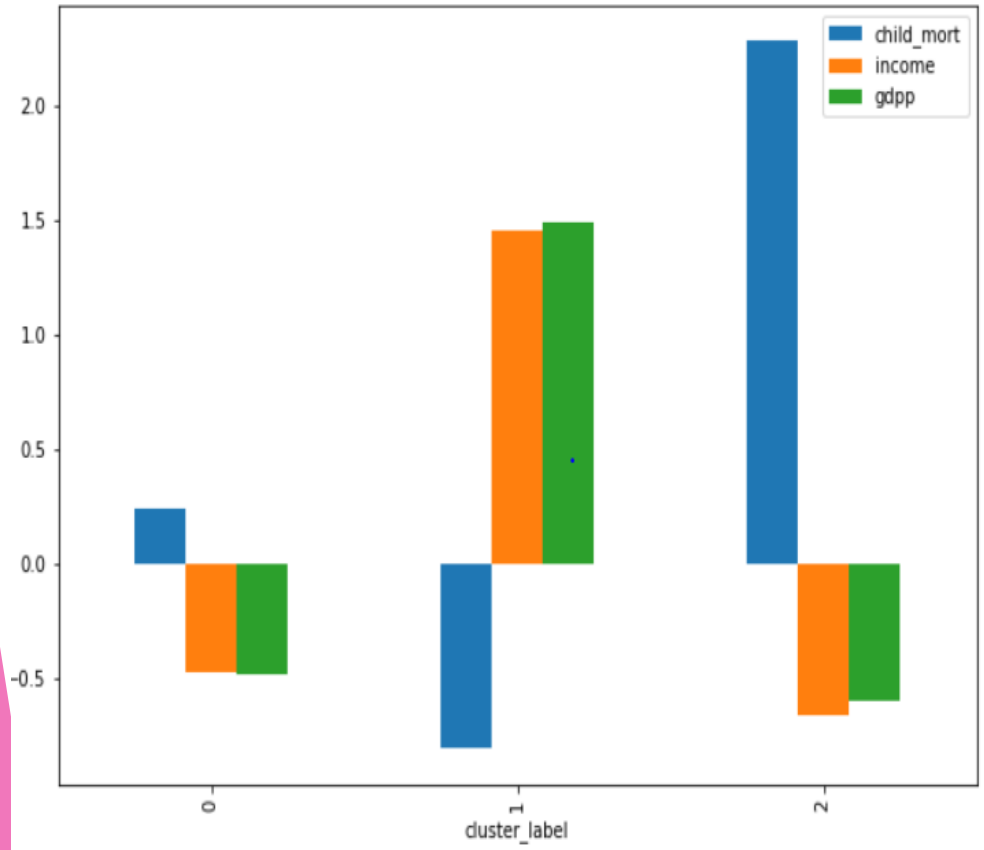
- It can be seen from the plot which as three cluster.
- The countries can be divided into 3 different clusters
- orange , blue and green are highlighted 3 clusters

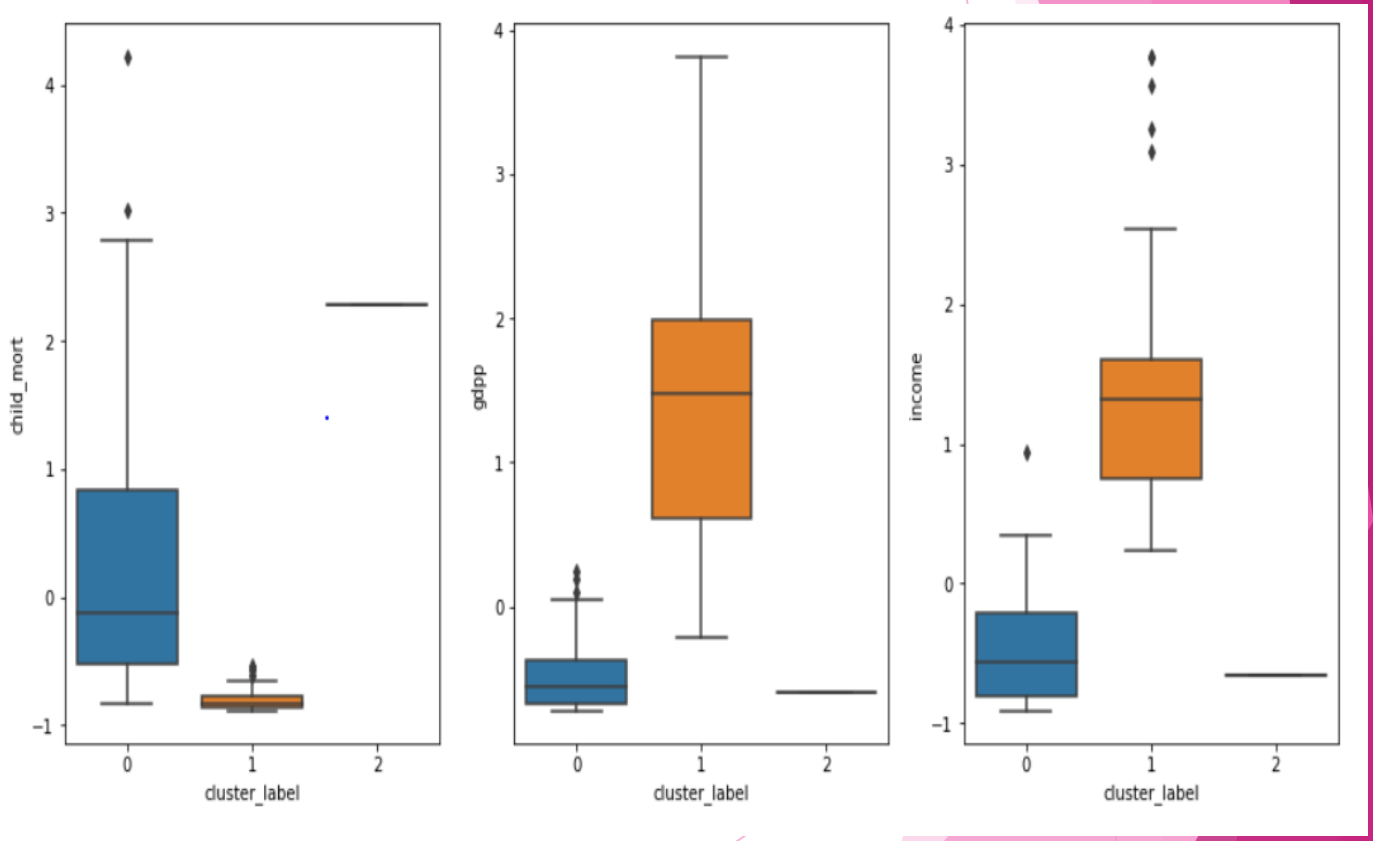


Analysis of Clusters

It can be seen that cluster 0 and cluster 1 are majority occupied and cluster 2 has only one country.

From the plots it can be seen , Comparative to other clusters,The cluster 2 which is have high child mortality and low income and low GDPP.





INFERENCES

- Top Countries which would required more attention during natural Calamites are following below:
- - Central Africa Republic
- - Niger
- - Congo, Dem. Rep.
- - Haiti
- - Sierra Leone

- Central Africa Republic:- The Child mortality is very high and GDPP and Income is very low. Total fertility is also high and health spending is very low.

- Niger :- Niger has very low gdpp and low income and also bery high mortality
, also total fertility rate is also high. Health Spending is very low

Central African republic

Central Africa Republic	
child_mort	149
income	888
gdpp	446
health	17.7508
life_expec	47.5
total_fer	5.21

Niger

Niger	
child_mort	123
income	814
gdpp	348
health	17.957
life_expec	58.8
total_fer	7.49

- Congo, Dem. Rep. : It has very least GDPP and income. Health spending is very low and total fertility rate is also high

- Haiti :- it has very low gdpp and also quiet low income and child mortality is high , life expectancy is also very low. Health spending is very low

Congo,Dem.Rep.

Congo, Dem. Rep.	
child_mort	116
income	609
gdpp	334
health	26.419
life_expec	57.5
total_fer	6.54

Haiti

Haiti	
child_mort	208
income	1500
gdpp	662
health	45.744
life_expec	32.1
total_fer	3.33

Sierra Leone

- Sierra Leone :- in Sierra leone, The child mortality is highest and gdpp is very low.

Sierra Leone	
child_mort	160
income	1220
gdpp	399
health	52.269
life_expec	55
total_fer	5.2

THANK YOU