NGO Clustering Presentation

Name- Chetan Jaiprakash Gaiddhane
E-mail ID – g.chetan619@gmail.com

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.

Data Understanding-

The Dataset contains different countries and some of the socio-economic and health factors are given that determine the overall development of the country. Using which we need to suggest the countries which the CEO needs to focus on the most Using Clustering. The datasets containing the following socio-economic factors -

1. country
2. child_mort
3. exports
4. health
5. imports
6. Income
7. Inflation
8. life_expec
9. total_fer
10. gdpp

Analysis Approach-

1. Importing the Required Libraries.
2. Loading the Dataset.
3. Analyzing the Dataset.
4. Checking for Null Values in the dataset.
5. Converting the export, import, health variables into actual values ,as they are given as %age of GDP per capita.
6. Heatmap representation of correlation matrix of Dataset .
7. Performing Univariate Analysis.
8. Performing Bivariate Analysis.
9. Visualizing and Hanlding the Outliers in the Dataset.
10. Calculating the Hopkins statistic.
11. Performing Data Scaling.
12. Plotting the Silhouette score and Elbow curve-ssd.
13. Choosing the value of K for K-means Clustering.
14. Cluster Profiling Visualising the Cluster with GDP, INCOME AND CHID_MORT Only as they the Major contributing Features.
15. Sorting theTop 5 Countries which are in Direst need of Aid.
16. Performing Hierarchical Clustering –Single Linkage.
17. Performing Hierarchical Clustering –Complete Linkage.
18. Choosing the value of K for Hierarchical Clustering.
19. Cluster Profiling Visualising the Cluster with GDP, INCOME AND CHID_MORT Only as they the Major contributing Features.
20. Sorting theTop 5 Countries which are in Direst need of Aid.

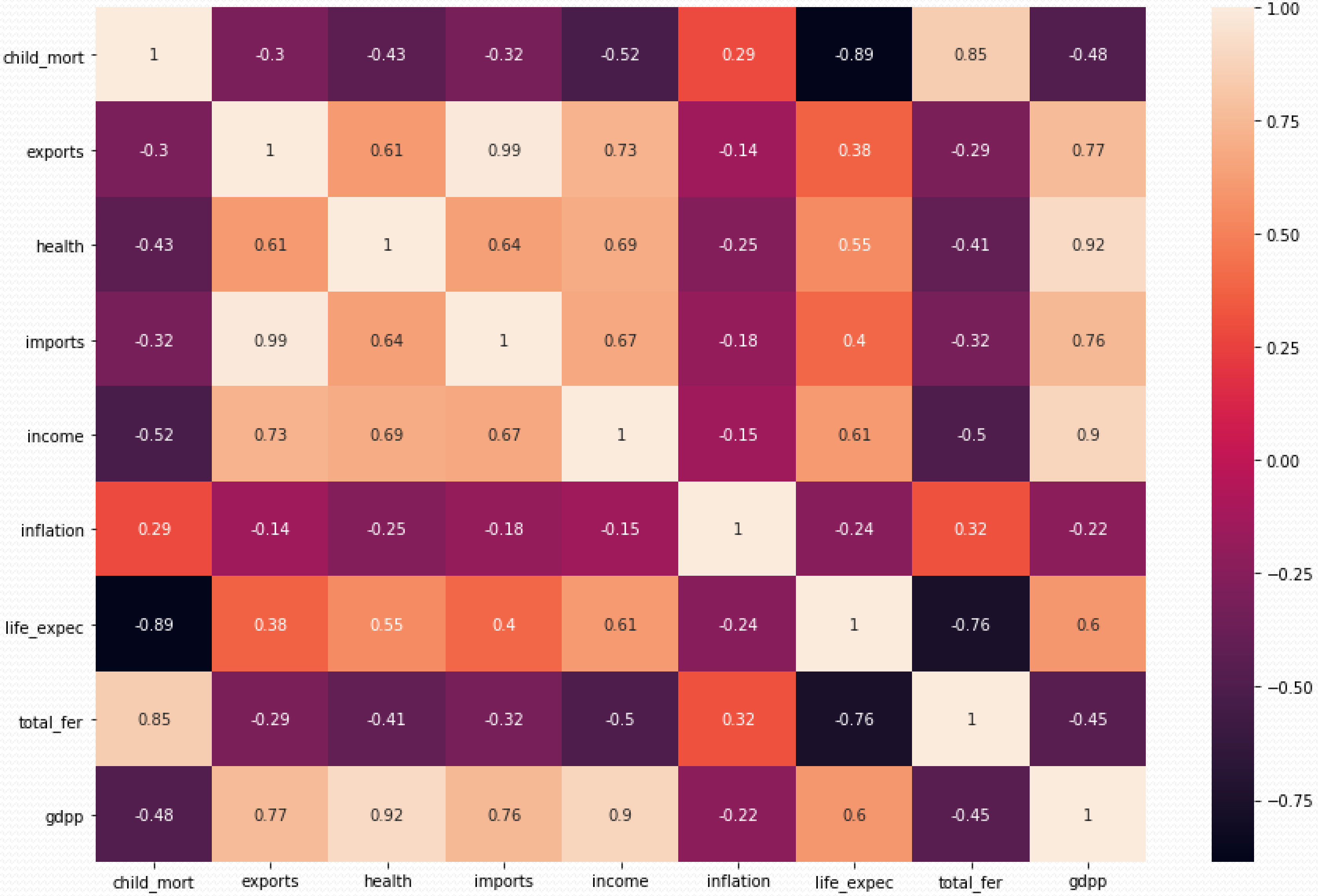
Data Conversion-

	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

-Converting the export, import, health variables into actual values ,as they are given as %age of GDP per capita.

	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

Heatmap representation of correlation of Dataset-

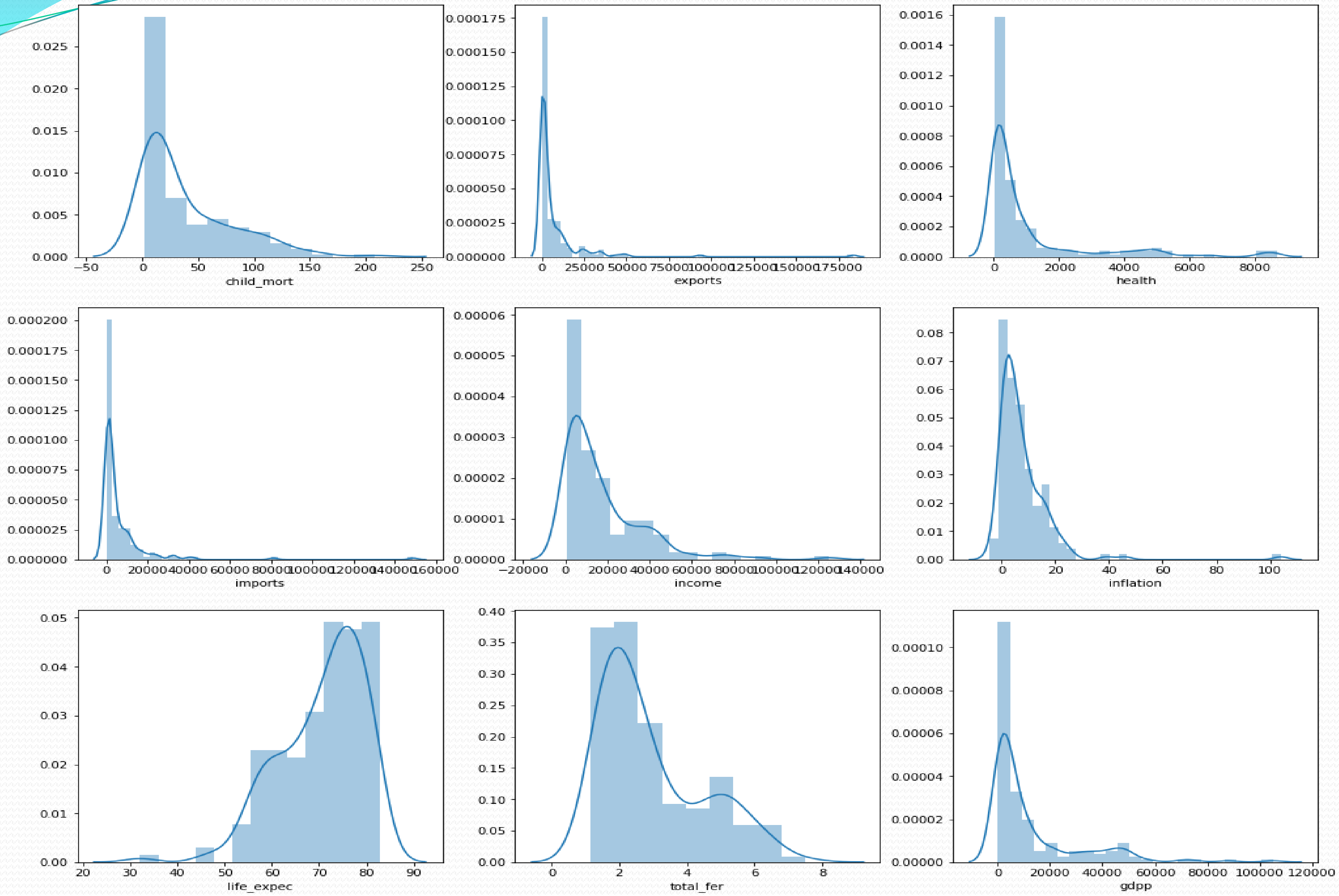


	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
child_mort	1	-0.3	-0.43	-0.32	-0.52	0.29	-0.89	0.85	-0.48
exports	-0.3	1	0.61	0.99	0.73	-0.14	0.38	-0.29	0.77
health	-0.43	0.61	1	0.64	0.69	-0.25	0.55	-0.41	0.92
imports	-0.32	0.99	0.64	1	0.67	-0.18	0.4	-0.32	0.76
income	-0.52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

Insights From Heatmap representation-

1. exports is highly correlated with imports.
2. health, exports, income,imports are highly correlated with gdpp.
3. total_fer is highly positively correlated with child_mort and negatively correlated with life_expec
4. child_mort is having high negative correlation with life_expec.

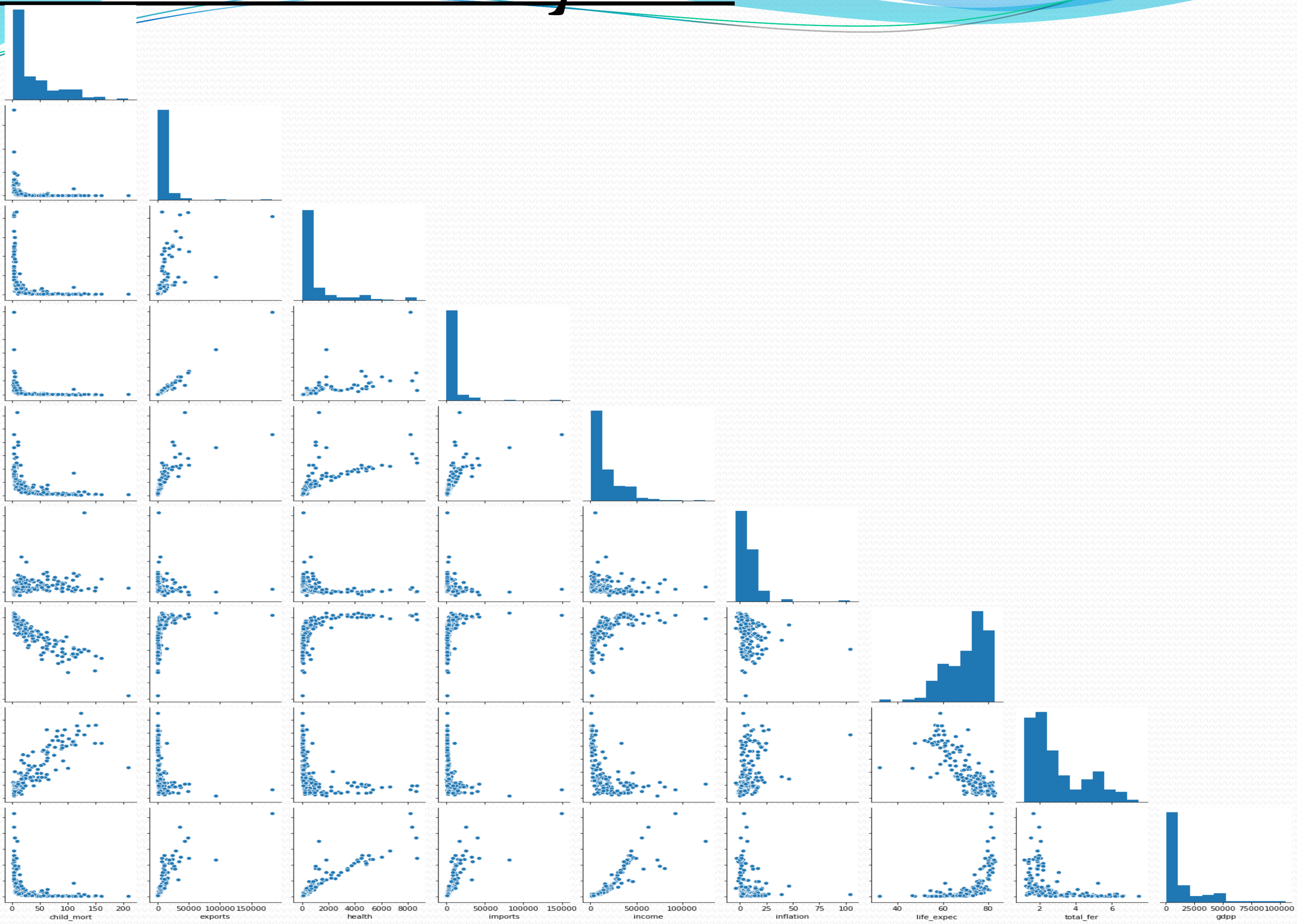
Univariate Analysis -



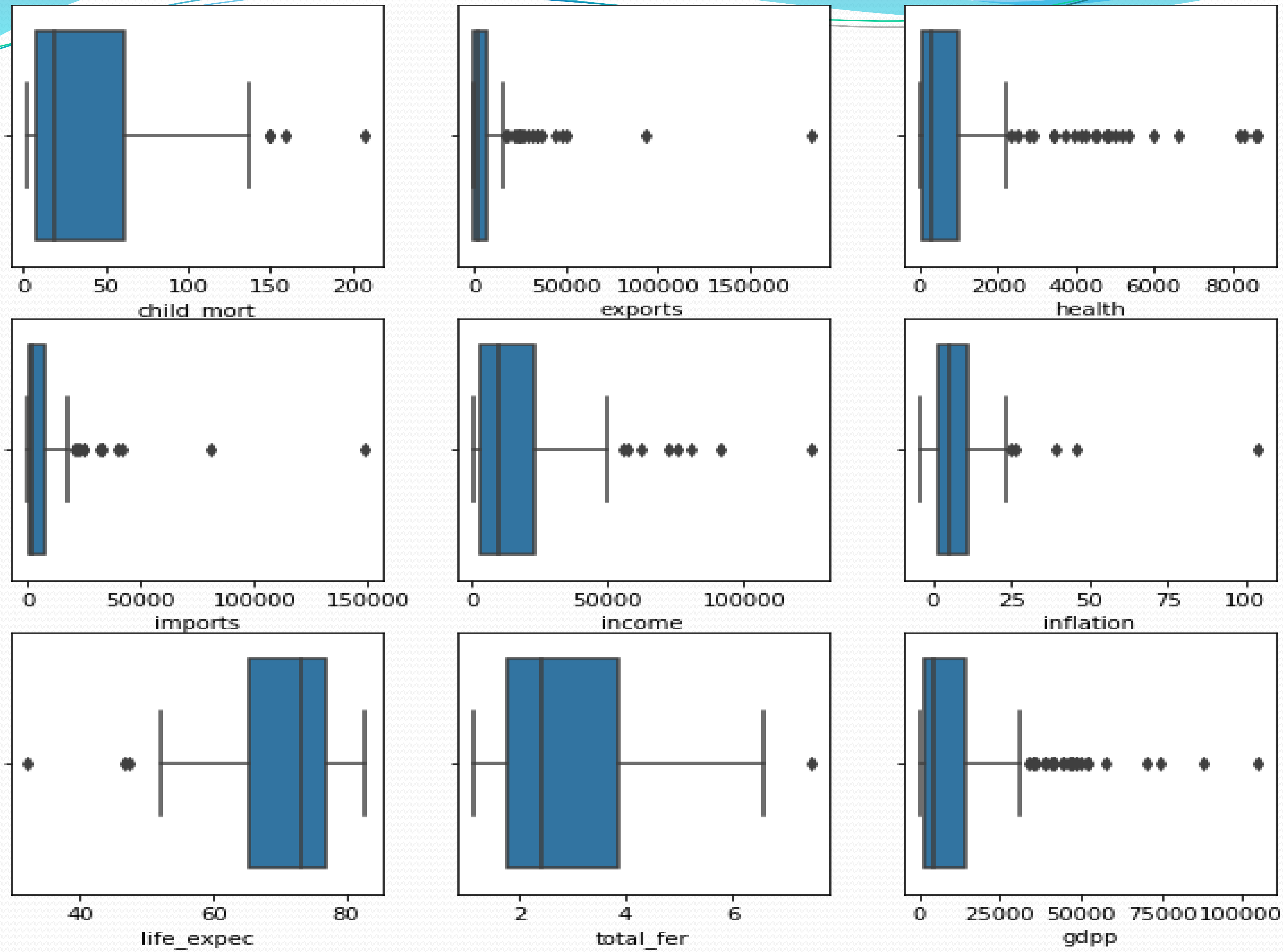
Insights from Above Univariate Analysis-

1. We can see that all the distributions except Life expectancy are formed on the Left side.
2. Export, Import and income has a very wide range of values .
3. In some countries the child mortality rate is very high.
4. some of the countries have very high gdpp and many countries have low or very low gdpp value.
5. Most of the countries have more than 50 years os life expectancy.

Bivariate Analysis -



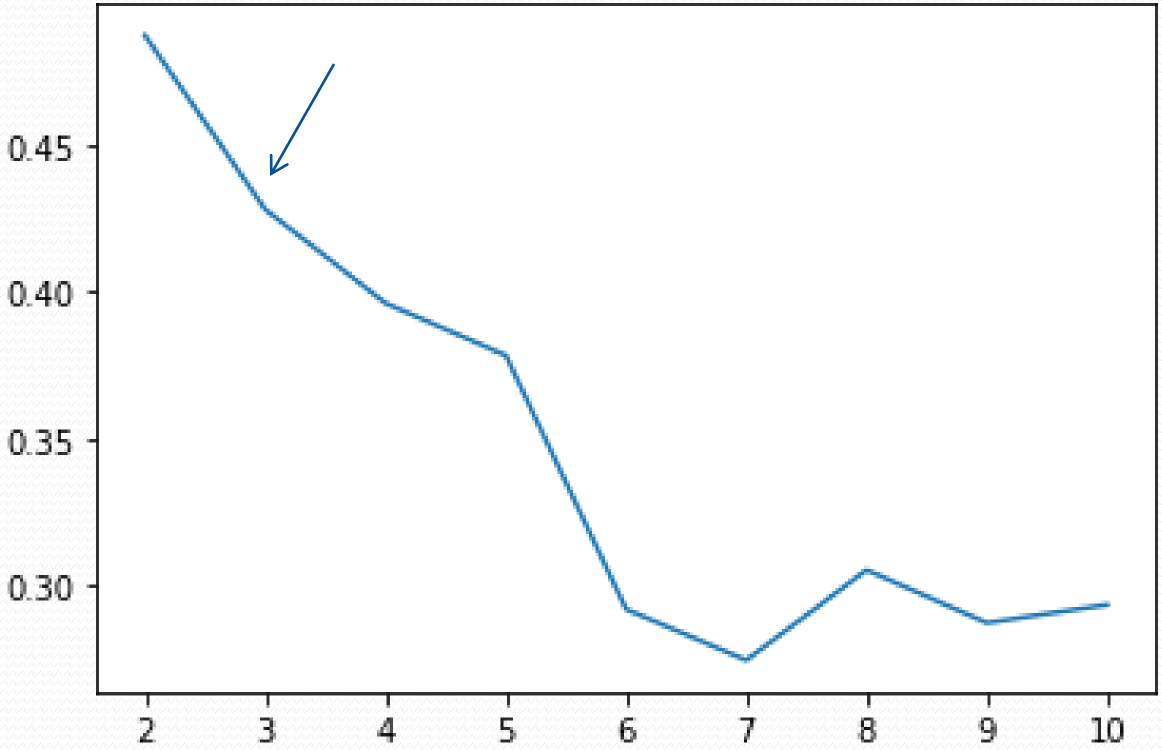
Visualising the Outliers in the Dataset-



Outliers Analysis-

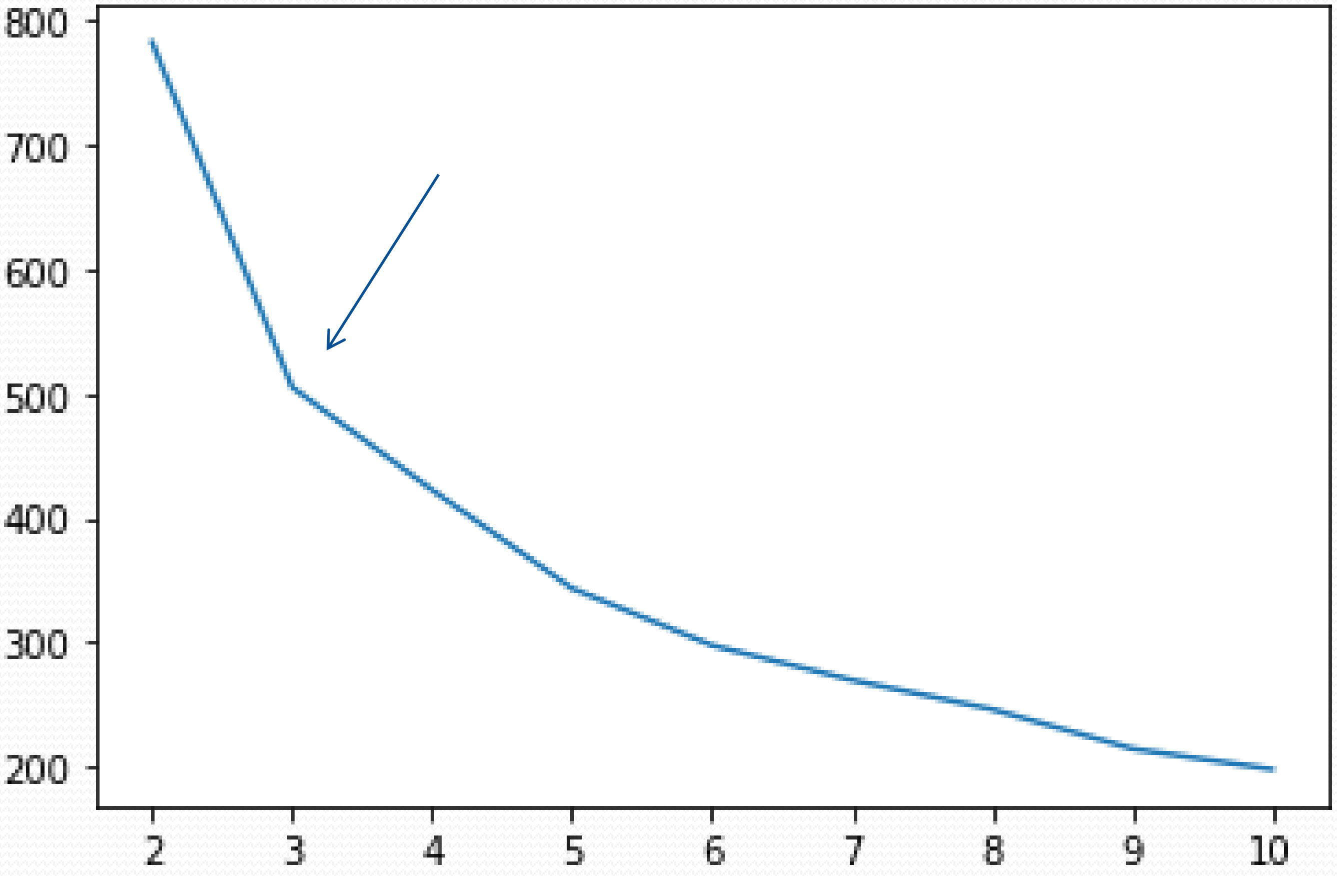
1. Outliers are present in almost all the numeric variables.
2. export,import,income and gdpp varible have some serious outliers which might affect the cluster formation.
3. Outliers in child_mort variable are expected as we need to target this variable for analysis ,High child mortality rate means they need the help.
4. It is not a good idea to remove the outliers as we will loose the data of most the countries ,but if we don’t do something Cluster formation can get affected .
5. So, it is good idea to cap the Higher values of import, export, income and gdpp variables only as the higher values of these variables are of the Developed countries which don’t really need help ,so we can cap the values for better Cluster formation
6. we will cap Only the Upper values for import ,export ,income and gdpp by 0.95 Quantile (Soft Margin Cap).
7. The upper values of these variables will not affect our analysis as they will cap only the developed countries don’t need the Aid .

Plot of the Silhouette statistics -



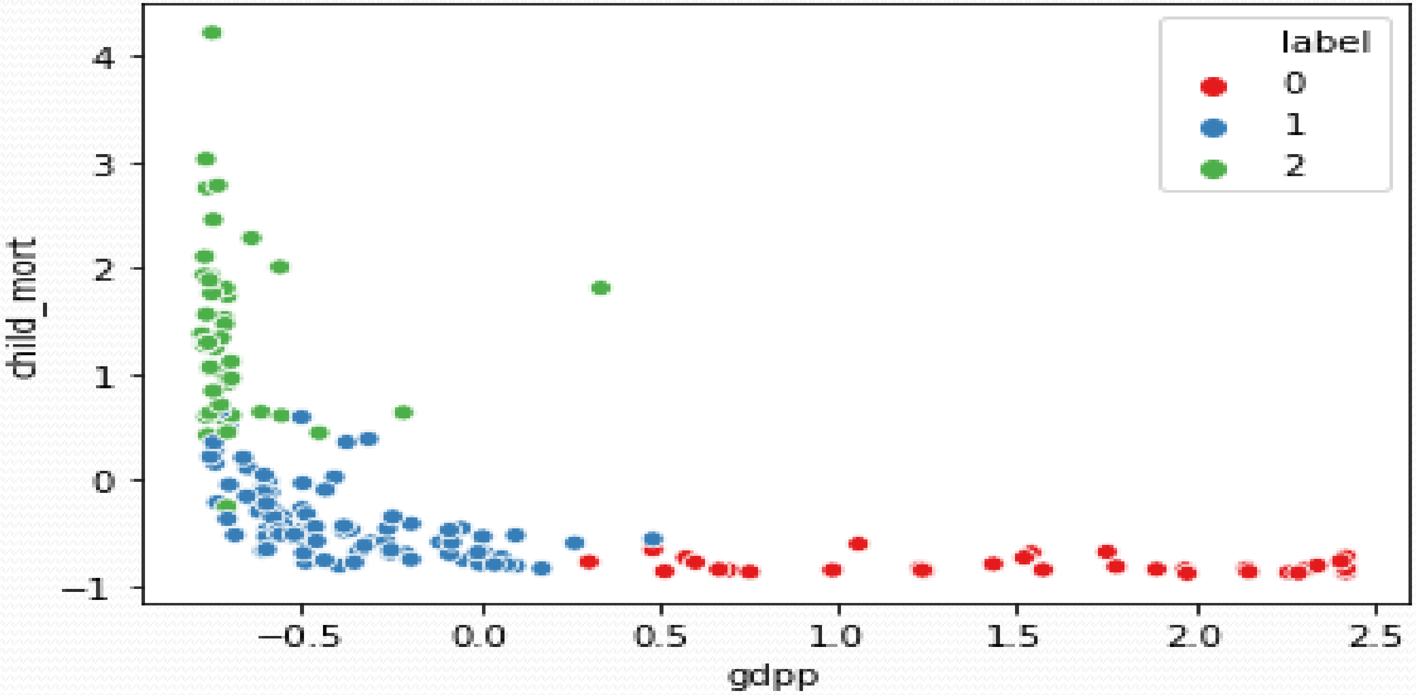
As we can see that there is a significant drop in the silhouette_score when moving from 3 to 4 clusters ,it seem good to select K=3.

Plot of the Elbow curve-ssd statistics-

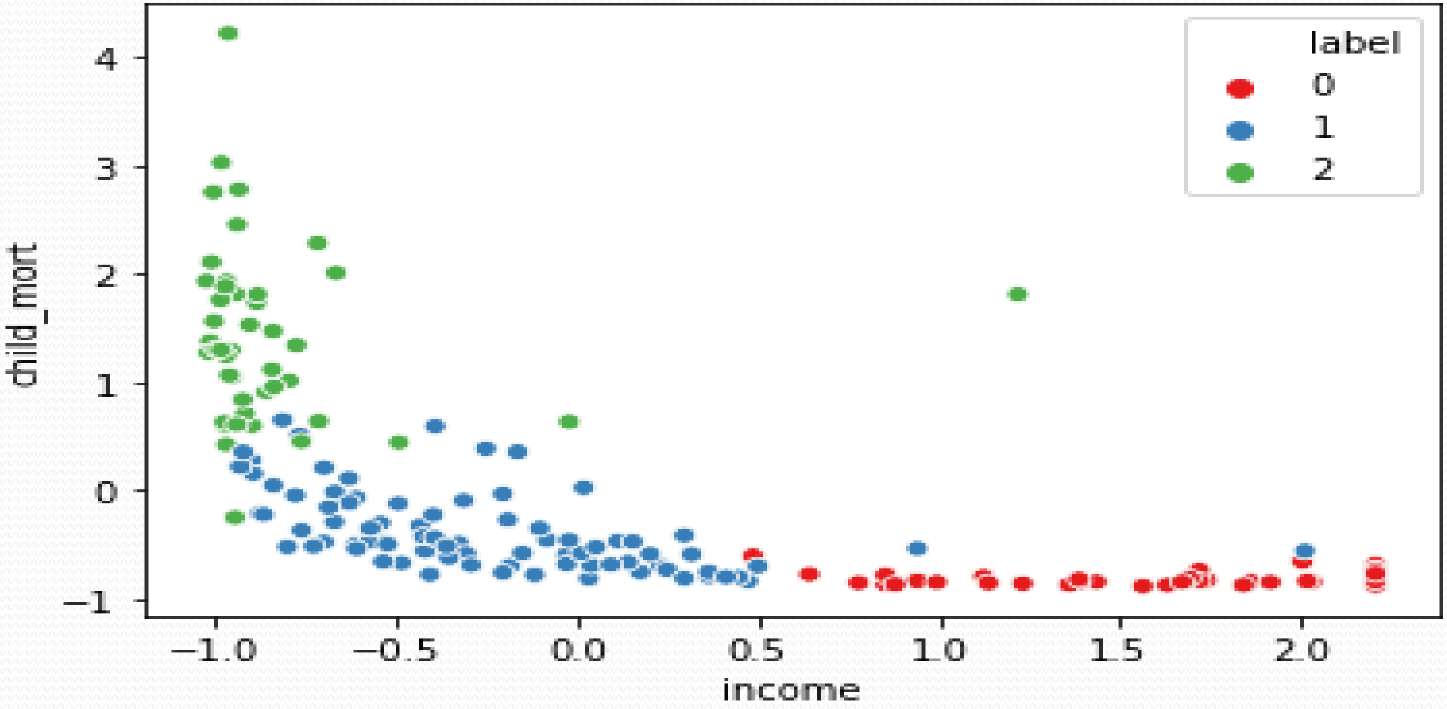


Also in Elbow-Curve Statistic we can see that there is a significant drop in the score when moving from 3 to 4 clusters ,it seem good to select K=3.

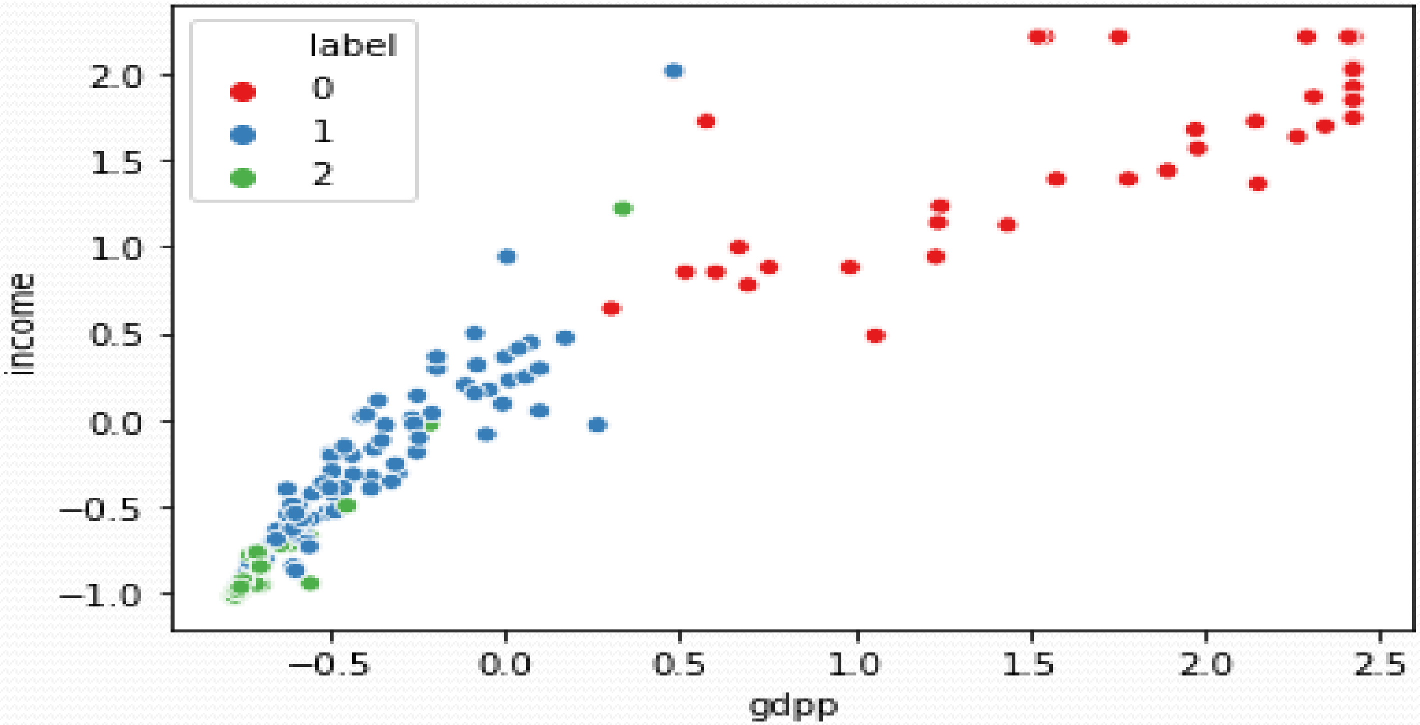
Plotting the cluster between 'gdpp' and 'child_mort' features-



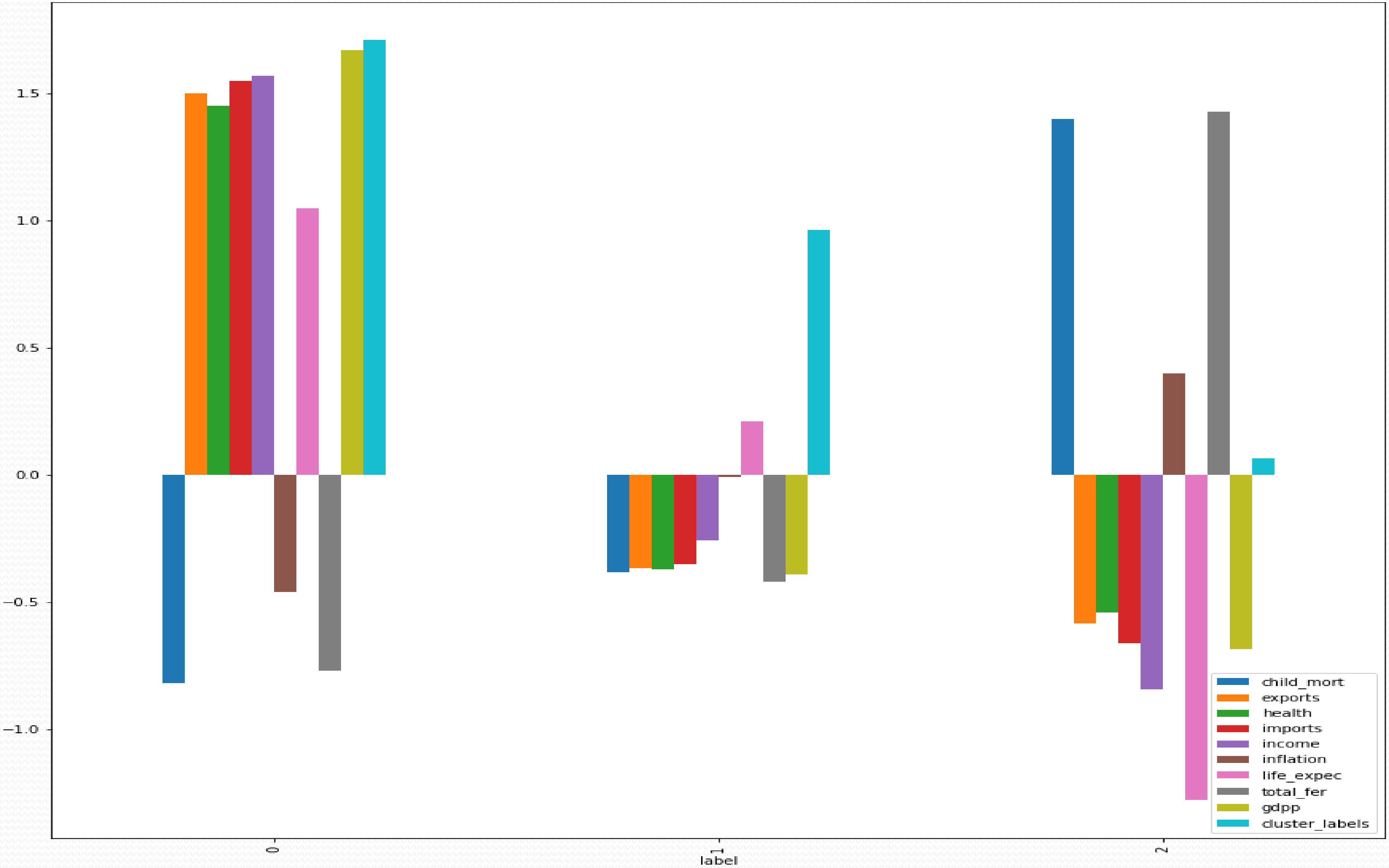
Plotting the cluster between ‘income' and 'child_mort' features-



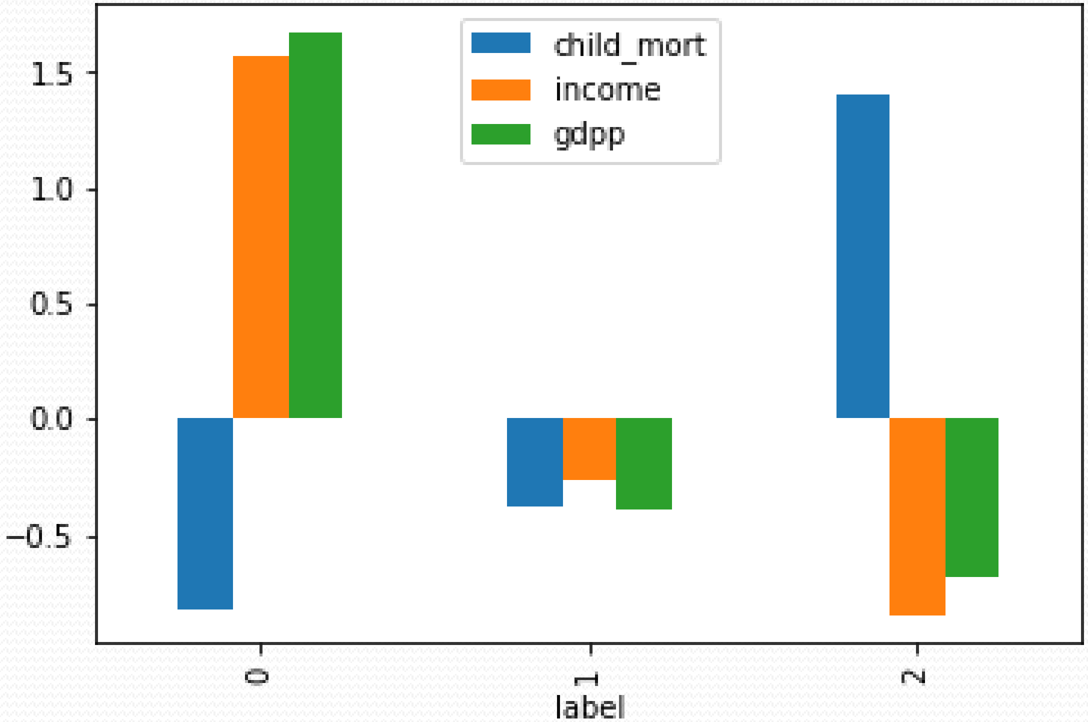
Plotting the cluster between 'gdpp' and ‘income' features-



Distribution of different varibles inside the Clusters -



Visualizing the Cluster with GDPP, INCOME AND CHID_MORT Only as they the Major contributing Features -



1. As we can see in the above figure when we filtered out the data of child_mort, income and gdpp for different clusters .
2. we can see that cluster 2 is having the lowest income and gdpp distribution ,Also we can see that the child_mort is very high in this cluster. This is the cluster we were looking for which is in Direst need of help

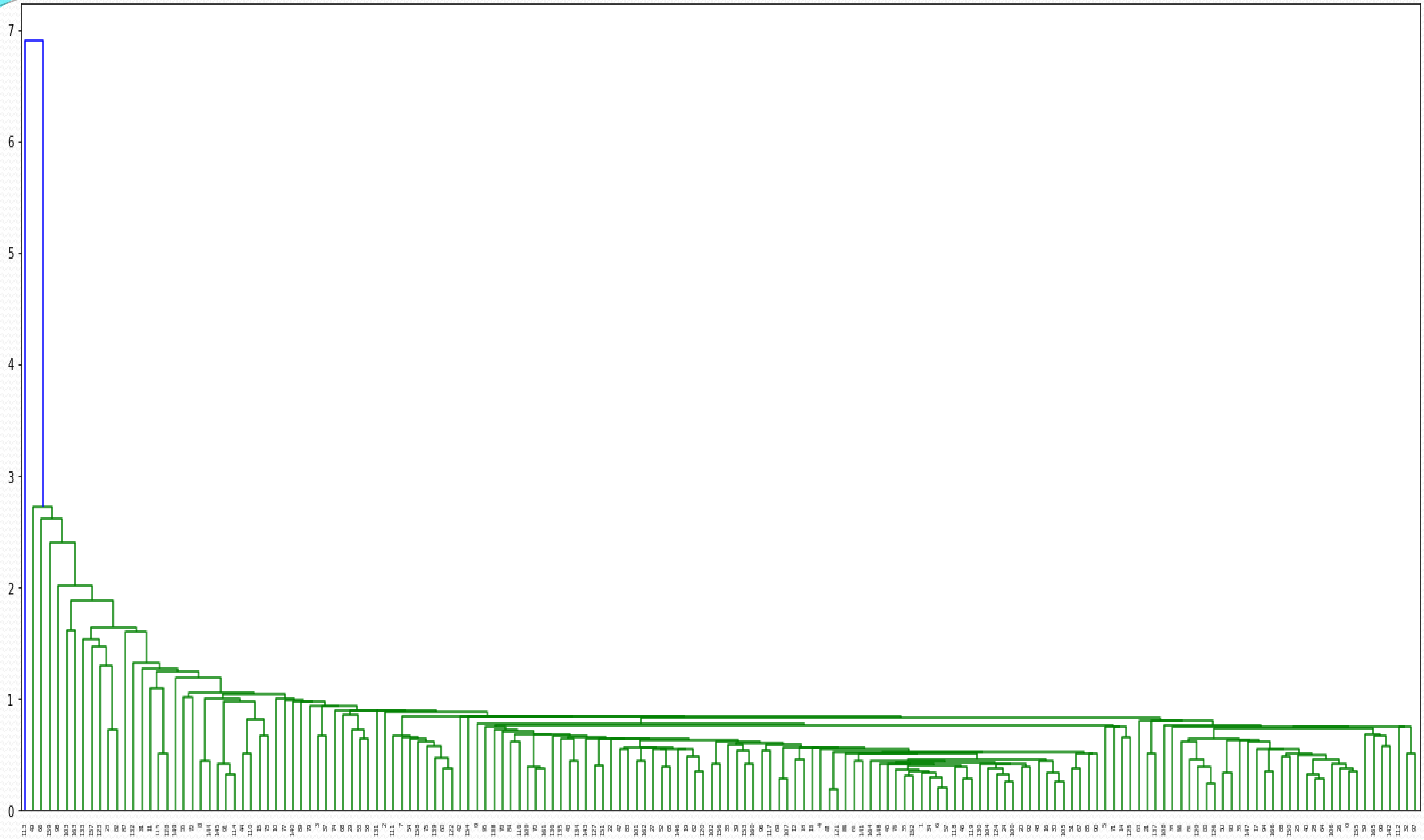
Top 5 Countries which are in Direst need of Aid as Found in K-Means Clustering :-

After Performing the K-means Clustering on the given dataset following are Top 5 Countries which are in Direst need of Aid, these Countries have lowest income and gdpp ,Also the Child Mortality Rate is highest . These Countries are as follows –

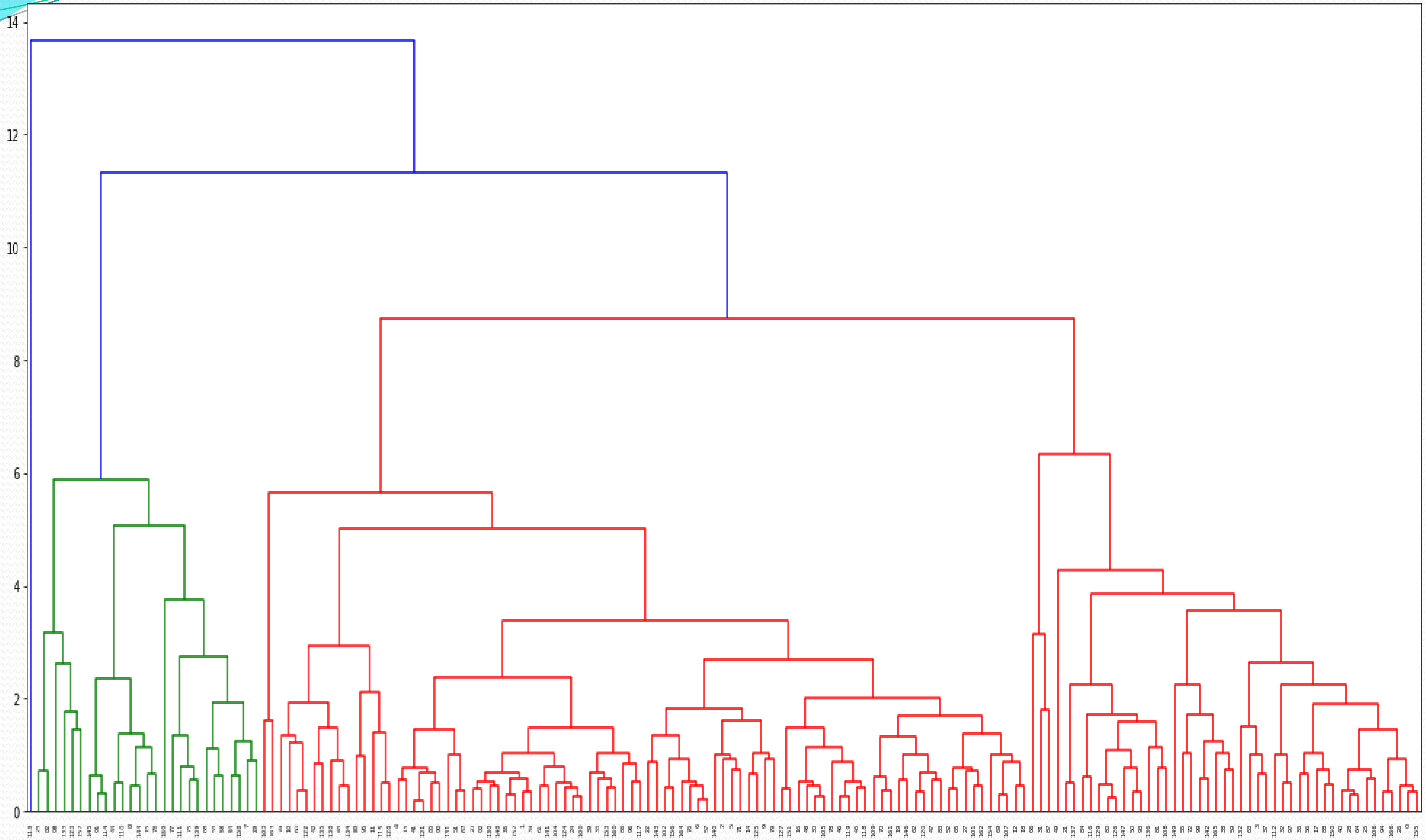
1. Congo, Dem. Rep.
2. Liberia
3. Burundi
4. Niger
5. Central African Republic

Hierarchical Clustering

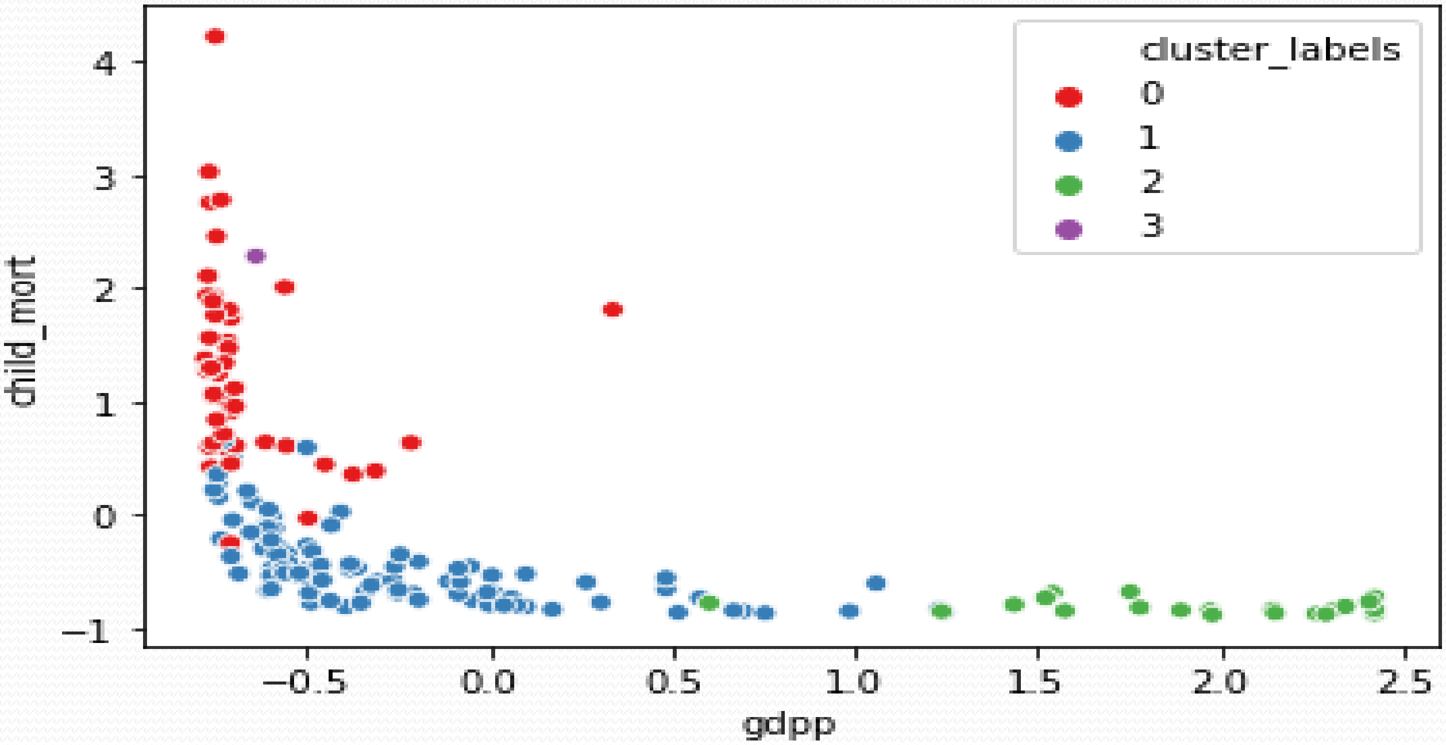
Single Linkage Dendrogram-



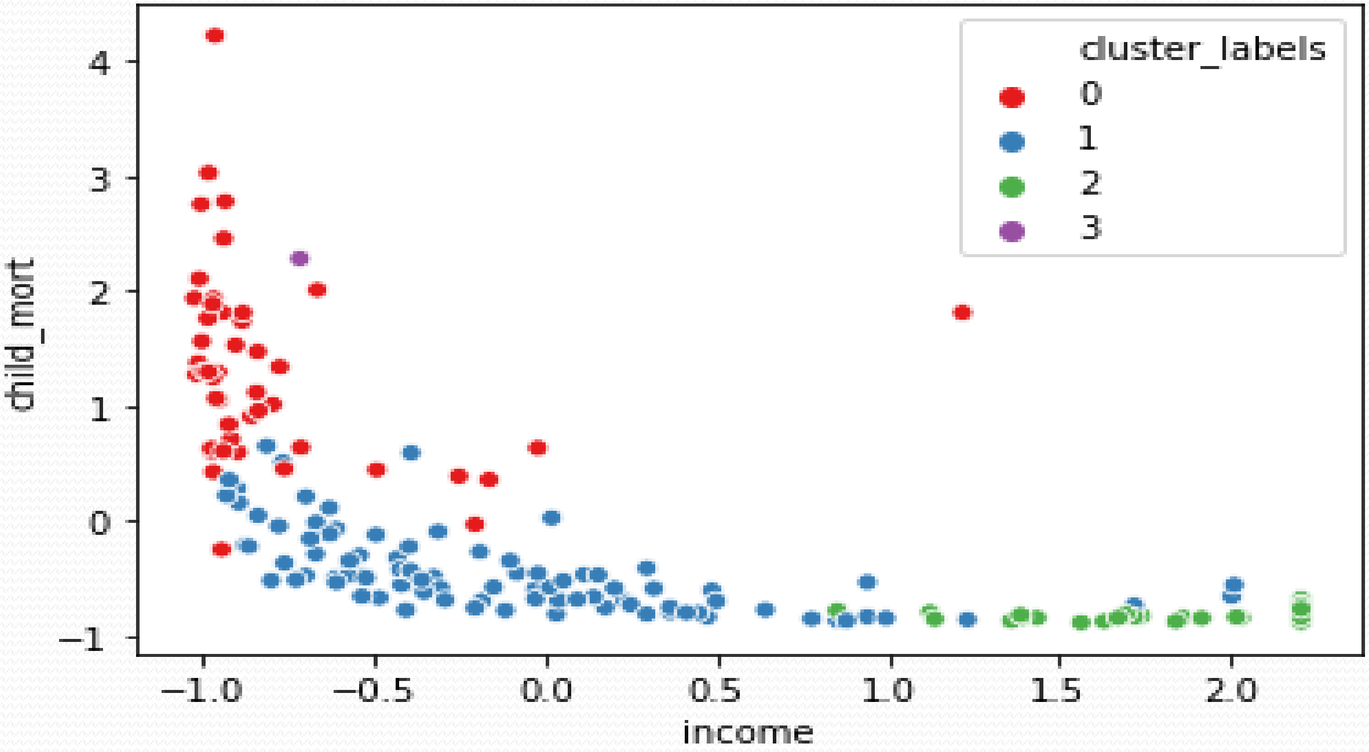
Complete Linkage Dendrogram-



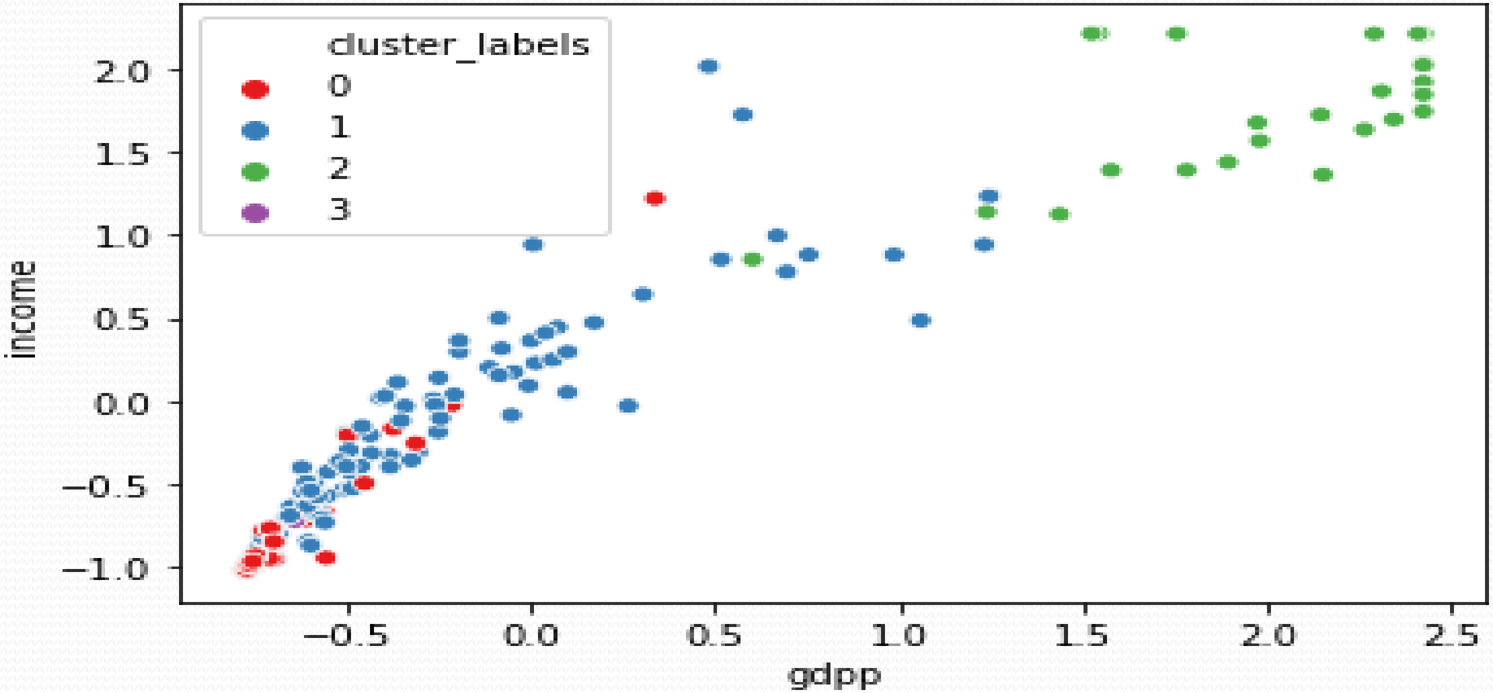
Plotting the cluster between 'gdpp' and 'child_mort' features-



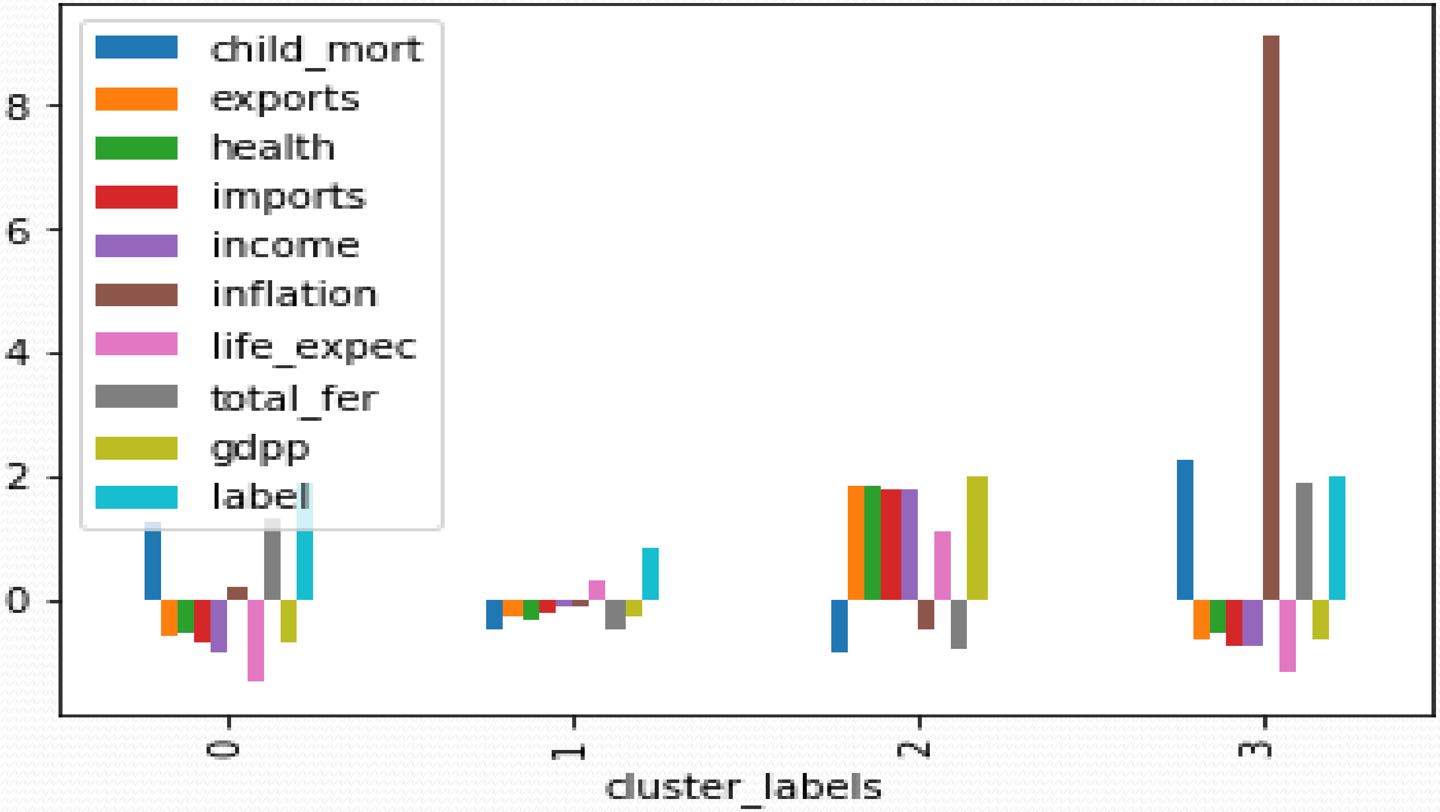
Plotting the cluster between ‘income' and 'child_mort' features-



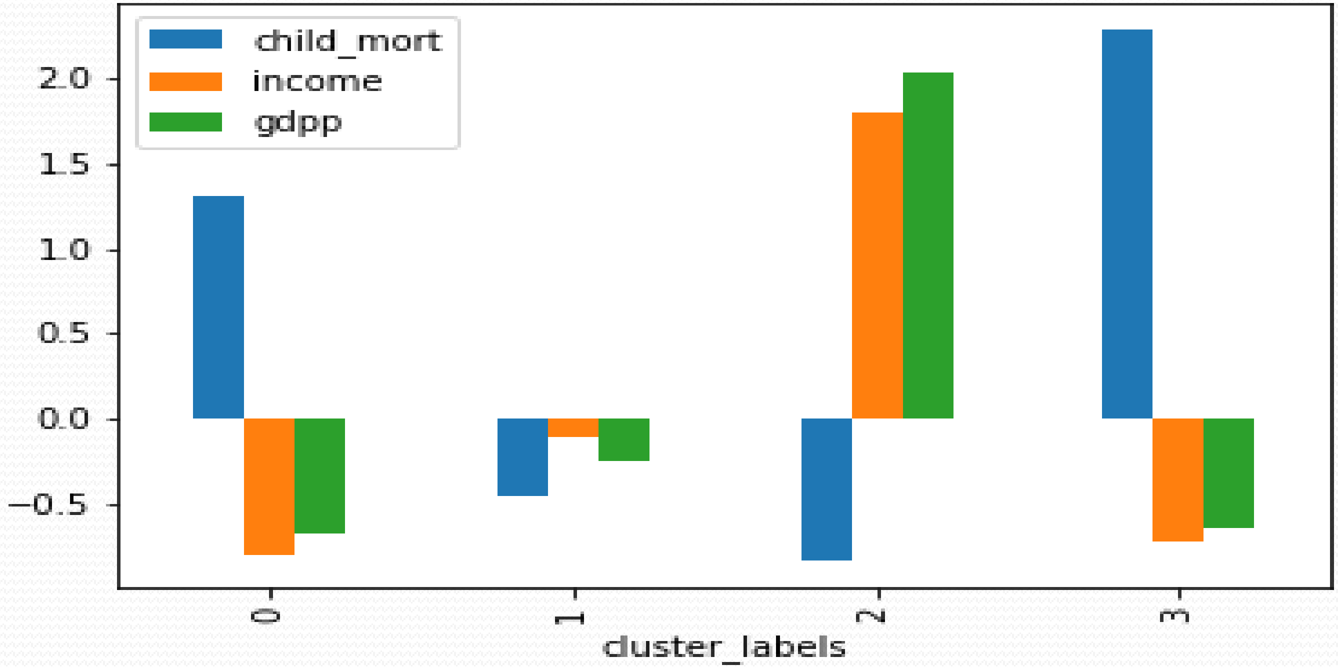
Plotting the cluster between 'gdpp' and ‘income' features-



Distribution of different varibles inside the Clusters -



Visualizing the Cluster with GDPP, INCOME AND CHID_MORT Only as they the Major contributing Features -



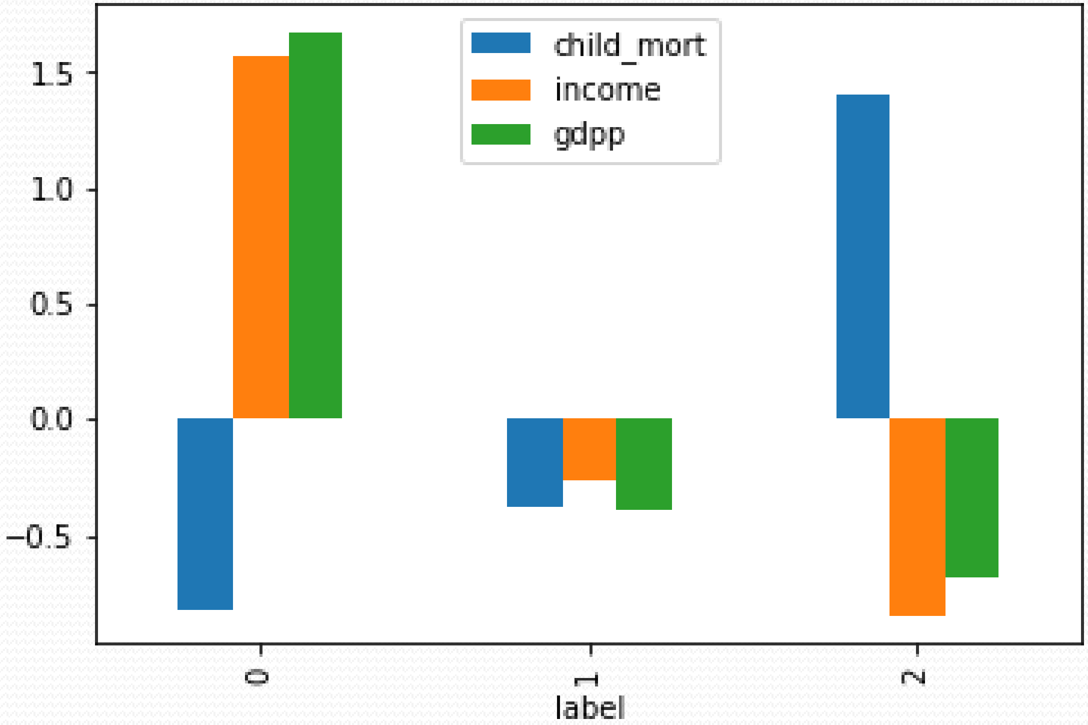
1. In the above figure when we filtered out the data of child_mort, income and gdpp for different clusters .
2. we can see that cluster 0 is having the lowest income and gdpp distribution ,Also we can see that the child_mort is very high in this cluster. This is the cluster we were looking for which is in Direst need of help.
3. Therefore we will select Cluster 0 for the Analysis of top 5 countries.

Top 5 Countries which are in Direst need of Aid as Found in Complete Linkage Hierarchical Clustering :-

After Performing the Complete Linkage Hierarchical Clustering on the given dataset following are Top 5 Countries which are in Direst need of Aid Are same as what we have obtained in the K-means Clustering Model, these Countries have lowest income and gdpp ,Also the Child Mortality Rate is highest . These Countries are as follows -

1. Congo, Dem. Rep.
2. Liberia
3. Burundi
4. Niger
5. Central African Republic

Visualizing the Cluster with GDPP, INCOME AND CHID_MORT Only as they the Major contributing Features -


child_mort
income
gdpp
label