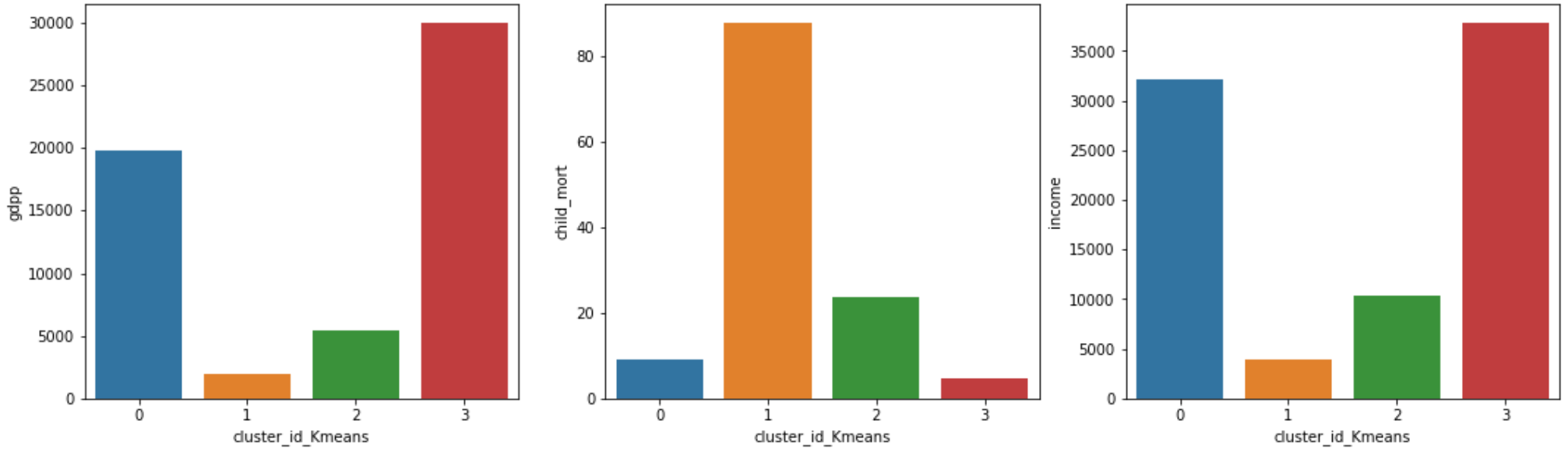
April 2020

Clustering & PCA Assignment

CLUSTERING AND PCA ASSIGNMENT
ISHAR MOHAPATRA

K-MEANS ALGORITHM.

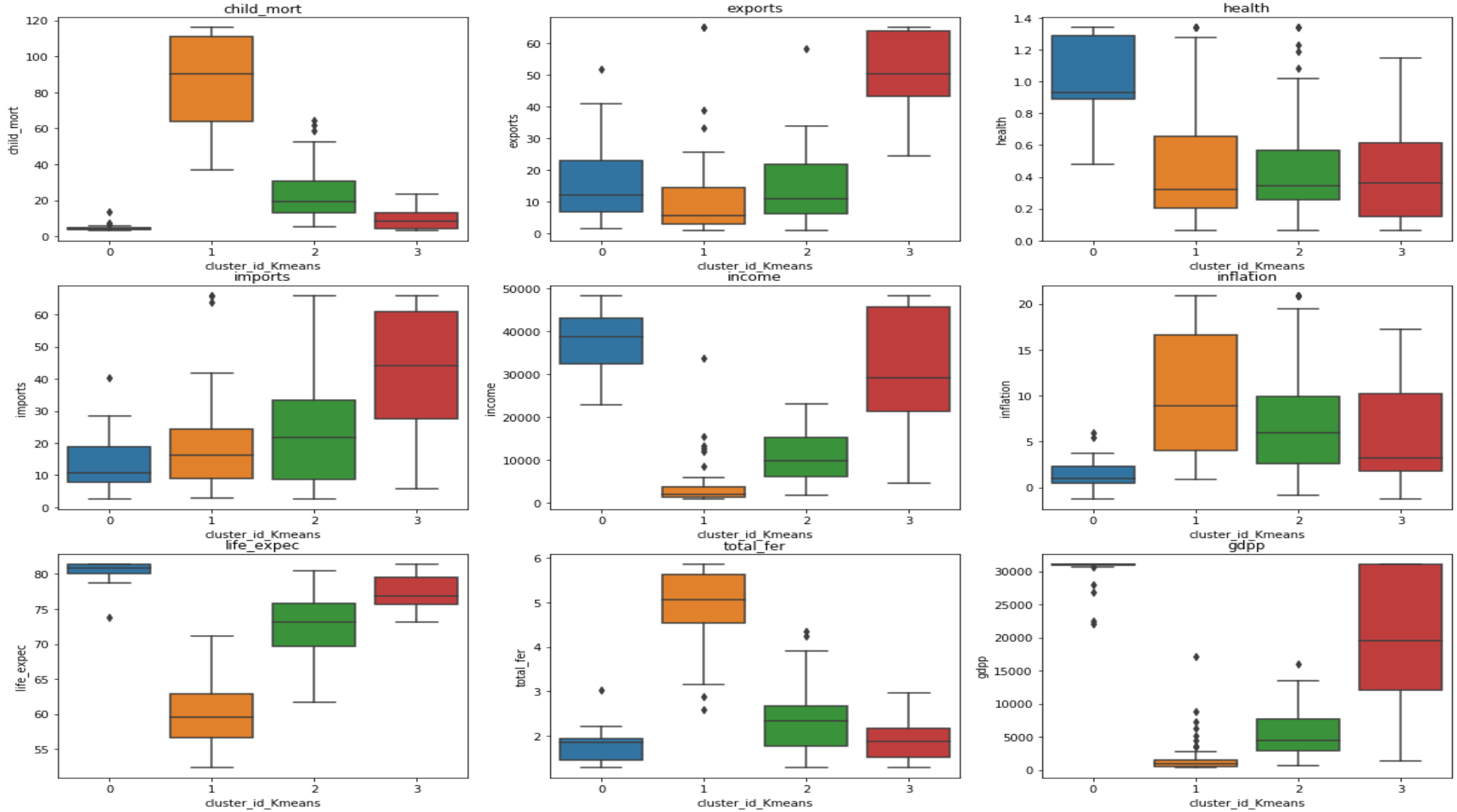
Here is the Bar plot for comparison between clusters based on GDP per capita, child mortality rate & net income person



Here we can clearly see that countries in cluster 1 requires the aid most, as it has lowest average GDP of 1927 , lowest average net income per person which is about 3952 and highest average child mortality rate which is about 88.

	cluster_id_Kmeans	child_mort	income	gdpp
0	0	4.713043	37837.826087	29970.869565
1	1	87.685106	3952.270638	1927.005957
2	2	23.578873	10408.169014	5422.845070
3	3	9.238462	32070.384615	19729.230769

Here is the Box plot for comparison between K-Means clusters based on the all dimensions.



ABOVE K-MEANS CLUSTERING - INFERENCES

Cluster 1 needs more attention

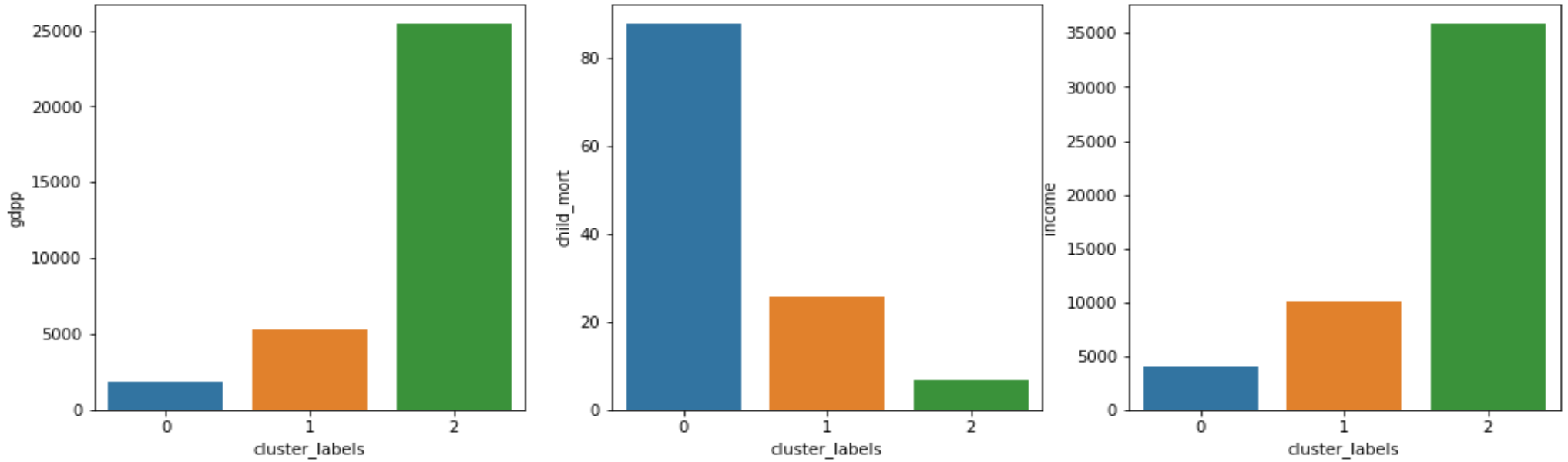
- child_mortality (Death of children under 5 years of age per 1000 live births) is the highest for Cluster 3 out of the 4.
- exports is the lowest for Cluster 3 which contributes to the GDPP. Hence GDP per capita is also lowest for Cluster 3.
- Income of the majority of the people is very less for Cluster 3 and Inflation is more.
- life_expec or the average number of years a new born child would live is very low for Cluster 3.

47 Countries are clustered here:

['Afghanistan', 'Angola', 'Benin', 'Botswana', 'Burkina Faso', 'Burundi', 'Cameroon', 'Central African Republic', 'Chad', 'Comoros', 'Congo, Dem. Rep.', 'Congo, Rep.', "Cote d'Ivoire", 'Equatorial Guinea', 'Eritrea', 'Gabon', 'Gambia', 'Ghana', 'Guinea', 'Guinea-Bissau', 'Haiti', 'Iraq', 'Kenya', 'Kiribati', 'Lao', 'Lesotho', 'Liberia', 'Madagascar', 'Malawi', 'Mali', 'Mauritania', 'Mozambique', 'Namibia', 'Niger', 'Nigeria', 'Pakistan', 'Rwanda', 'Senegal', 'Sierra Leone', 'Solomon Islands', 'South Africa', 'Sudan', 'Tanzania', 'Timor-Leste', 'Togo', 'Uganda', 'Yemen', 'Zambia']

HIERARCHICAL CLUSTERING ALGORITHM.

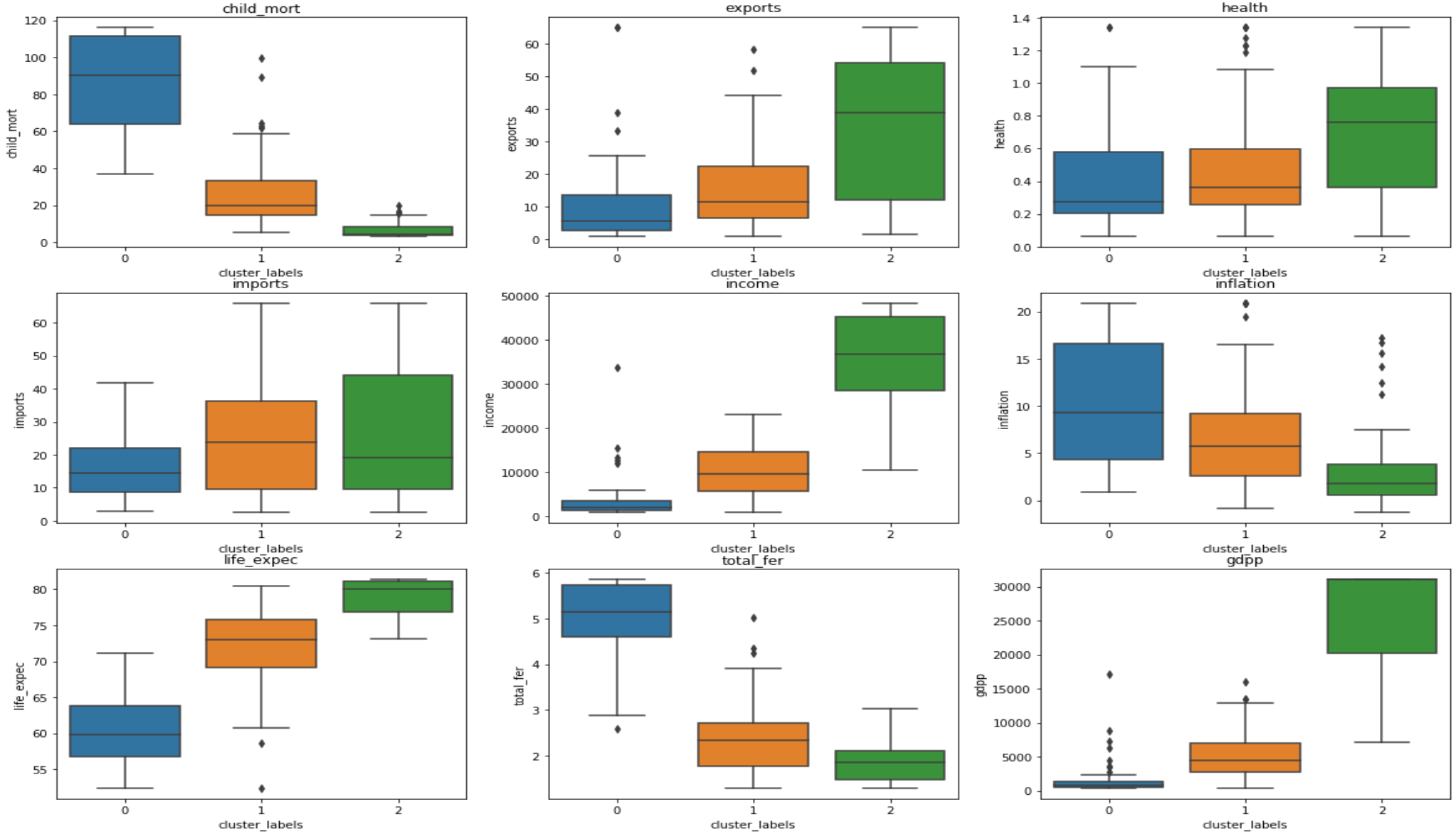
Here is the Bar plot for comparison between clusters based on GDP per capita, child mortality rate & net income person



Here we can clearly see that countries in cluster 0 requires the aid most as it has lowest average GDP of 1888 , lowest average net income per person which is about 3964 and highest average child mortality rate which is about 88.

	cluster_labels	child_mort	income	gdpp
0	0	87.861364	3964.978182	1888.476364
1	1	25.889474	10108.785263	5248.530526
2	2	6.604255	35874.680851	25444.680851

Here is the Box plot for comparison between Hierarchical clusters based on the all dimensions.



ABOVE HIERARCHICAL CLUSTERING - INFERENCES

Cluster 0 needs more attention which is similar to Cluster 1 of K-Means

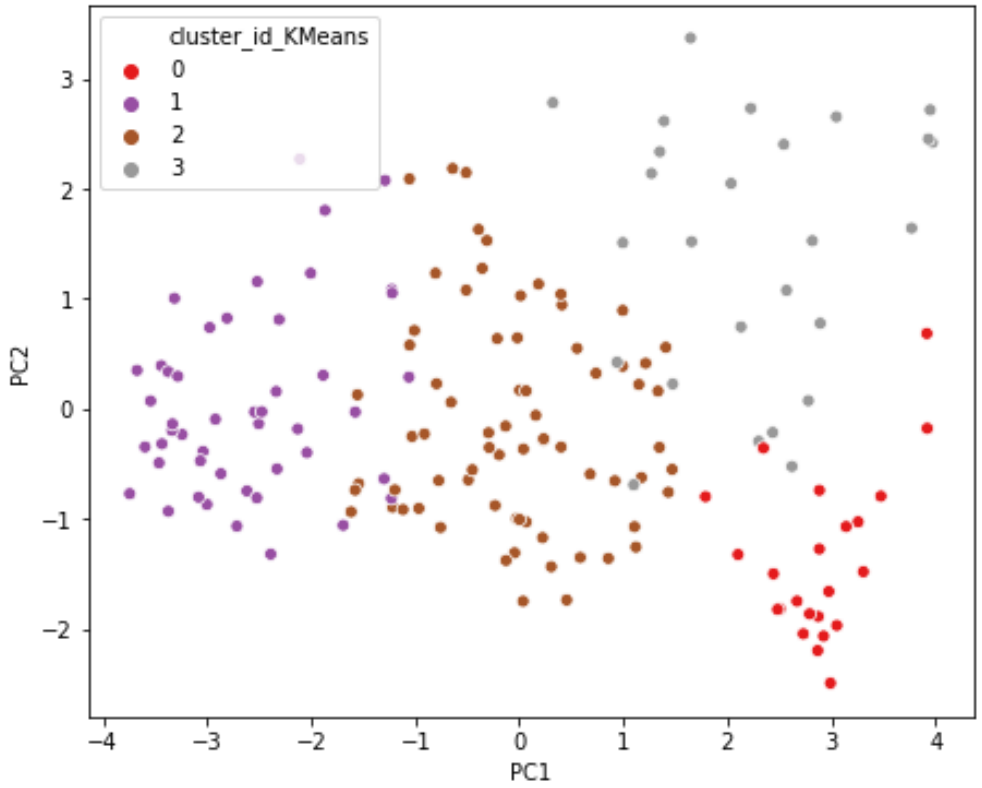
- child_mortality (Death of children under 5 years of age per 1000 live births) is the highest for Cluster 0.
- exports & imports is the lowest for Cluster 0 which contributes to the GDPP. Hence GDP per capita is also lowest for Category 0.
- Income of the people is very less for Cluster 0 and Inflation is more.
- life_expec or the average number of years a new born child would live is also very low for Cluster 0.
- expenses on health care is also less for Cluster 0 countries.

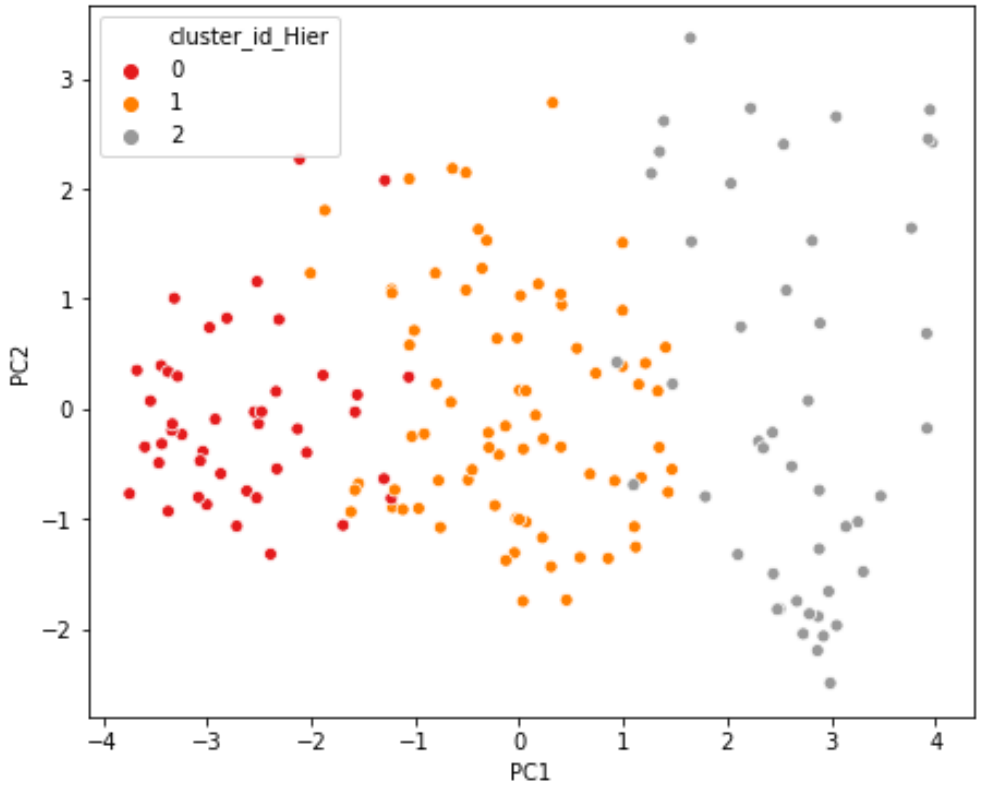
44 Countries are clustered here:

['Afghanistan', 'Angola', 'Benin', 'Botswana', 'Burkina Faso', 'Burundi', 'Cameroon', 'Central African Republic', 'Chad', 'Comoros', 'Congo, Dem. Rep.', 'Congo, Rep.', "Cote d'Ivoire", 'Equatorial Guinea', 'Eritrea', 'Gabon', 'Gambia', 'Ghana', 'Guinea', 'Guinea-Bissau', 'Haiti', 'Iraq', 'Kenya', 'Lao', 'Madagascar', 'Malawi', 'Mali', 'Mauritania', 'Mozambique', 'Niger', 'Nigeria', 'Pakistan', 'Rwanda', 'Senegal', 'Sierra Leone', 'South Africa', 'Sudan', 'Tajikistan', 'Tanzania', 'Timor-Leste', 'Togo', 'Uganda', 'Yemen', 'Zambia']

COMPARING BOTH CLUSTERS

Here we have compared K-Means cluster vs Hierarchical Cluster based on Principal Component 1 vs Principal Component 2.

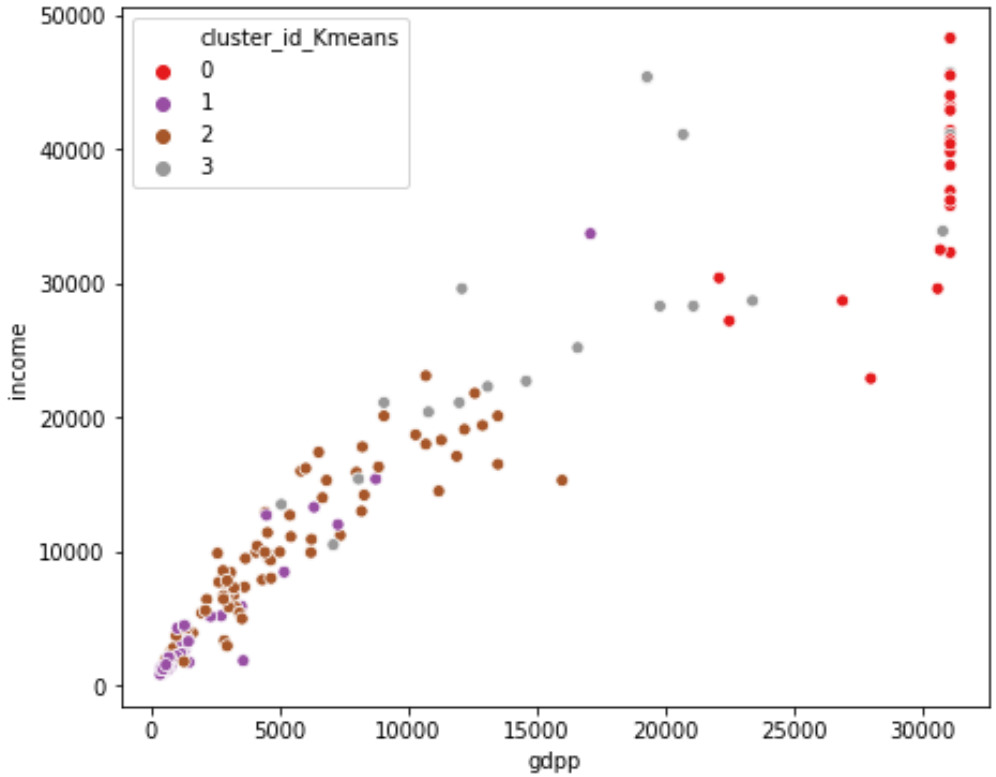


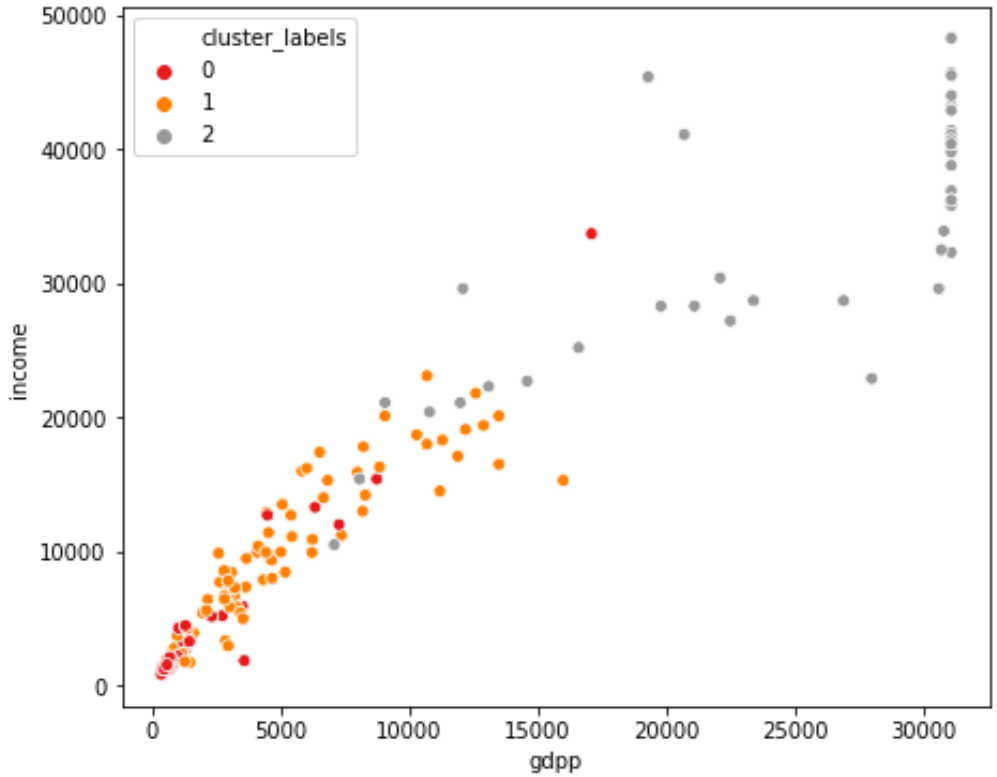


Here PC1 and PC2 explains about 65 % of the data.

COMPARING BOTH CLUSTERS

Here we have compared K-Means cluster vs Hierarchical Cluster based on GDP per Capita vs net person income.

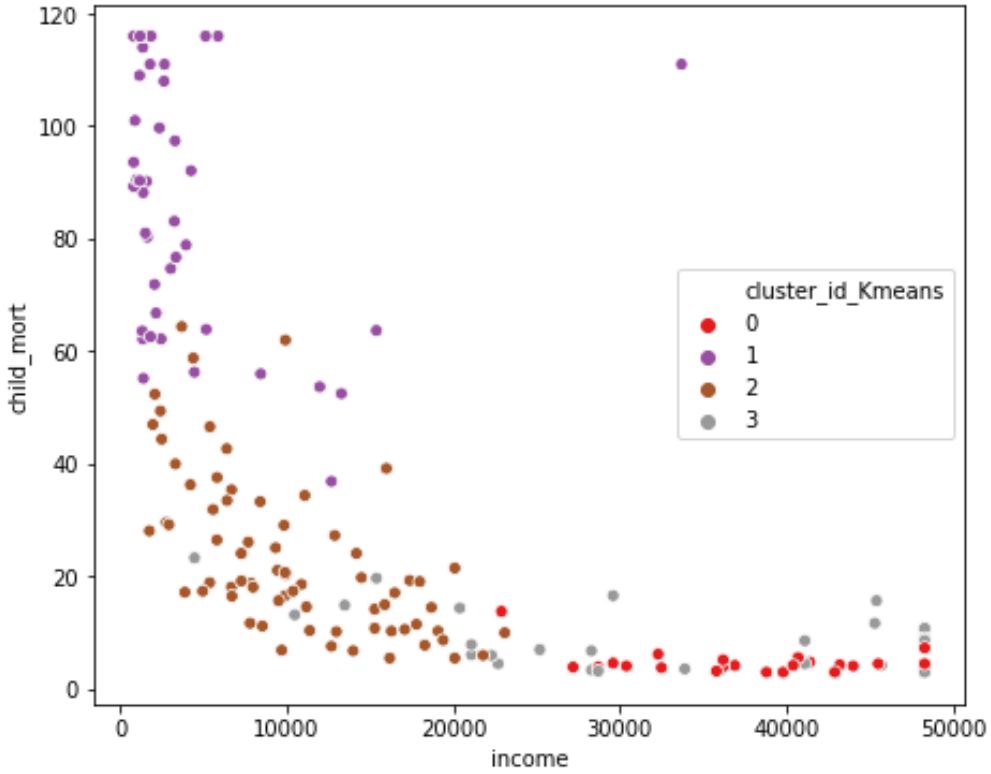


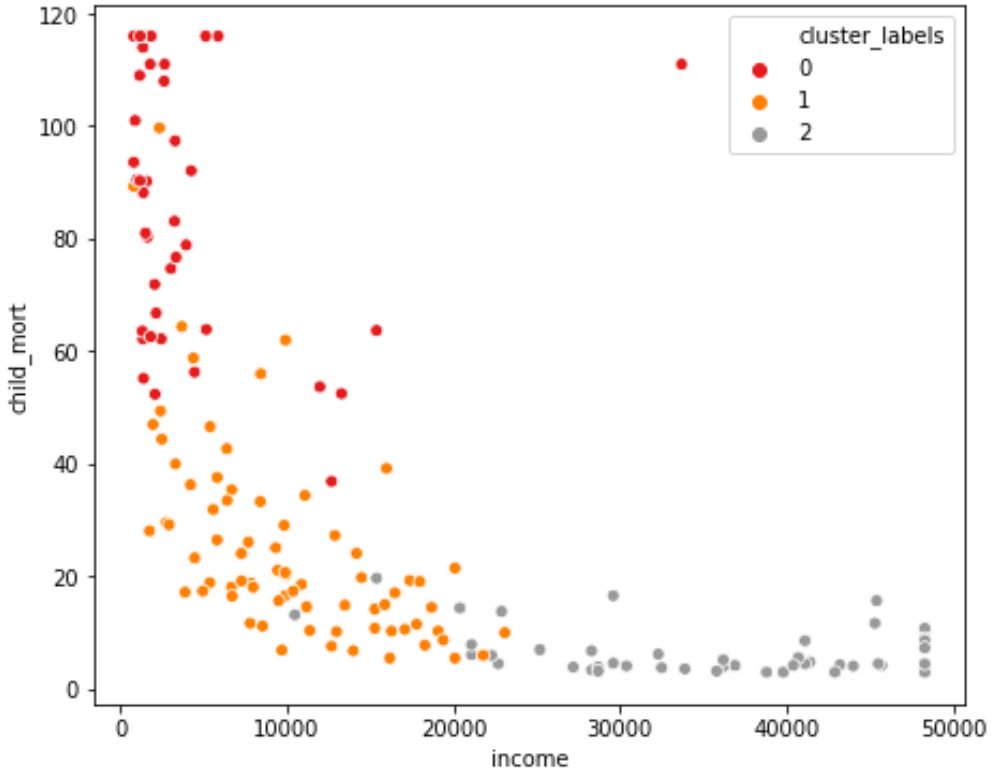


Due to capping of outliers we can several point in top right corner, hierarchical clustering has perfectly clustered them into 1 cluster.

COMPARING BOTH CLUSTERS

Here we have compared K-Means cluster vs Hierarchical Cluster based on net person income vs child mortality rate.

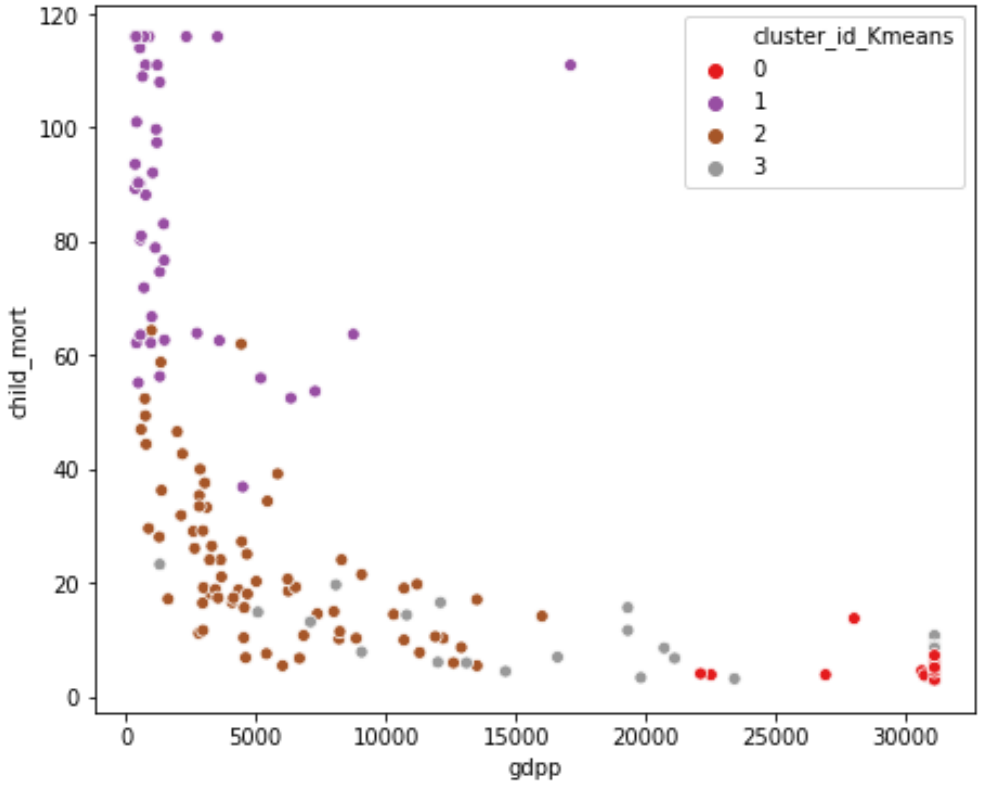


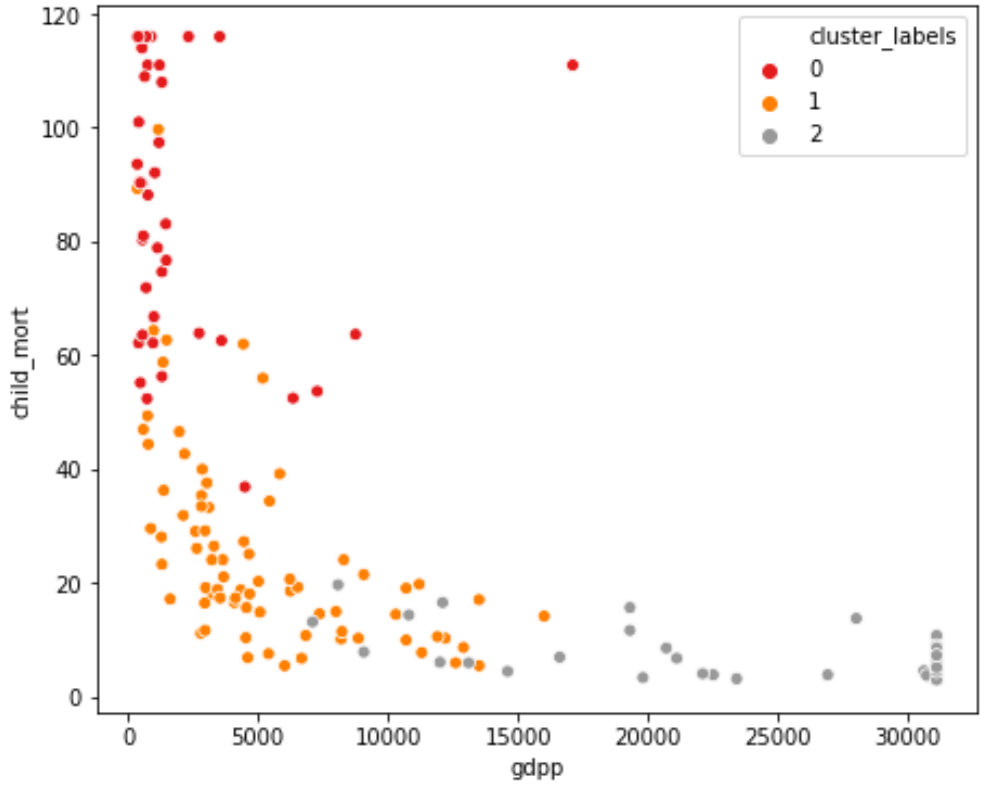


Here we can clearly notice that where net person income is less, child mortality rate is higher and where net person income is higher child mortality rate is less.

COMPARING BOTH CLUSTERS

Here we have compared K-Means cluster vs Hierarchical Cluster based on GDP per Capita vs child mortality.





We can see a similar trend here also, country whose GDP is less has child mortality rate higher and those having high GDP have child mortality is less.

FINAL RECOMMENDATION **for HELP International based on socio-economic and health factors that determine the overall development of the country from both the clustering algorithm are**

- Congo, Dem. Rep.
- Niger
- Burundi
- Sierra Leone
- Liberia
- Madagascar

- *Congo, Dem. Rep & Niger* has a child mortality rate of 116, net income per person of 837, GDP per capita of 364 and Inflation rates of 20.8 & 2.55 respectively.
- Burundi has a child mortality rate of 94, net income per person of 837, GDP per capita of 364 and Inflation rates of 12.
- Sierra Leone has a child mortality rate of 116, net income per person of 1220, GDP per capita of 399 and Inflation rates of 17.
- Liberia and Madagascar has a child mortality rate 89 & 62, net income per person of 837 & 1390 and GDP per capita of 364 & 413 respectively.