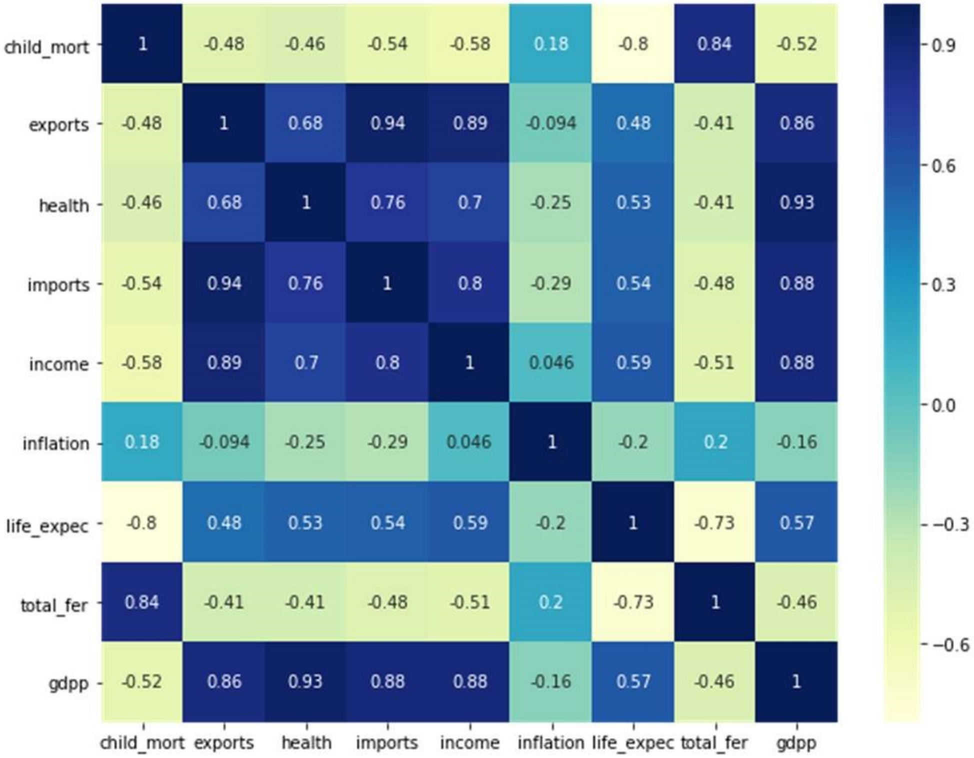
Clustering and PCA Assignment

By – Agam Kushwaha

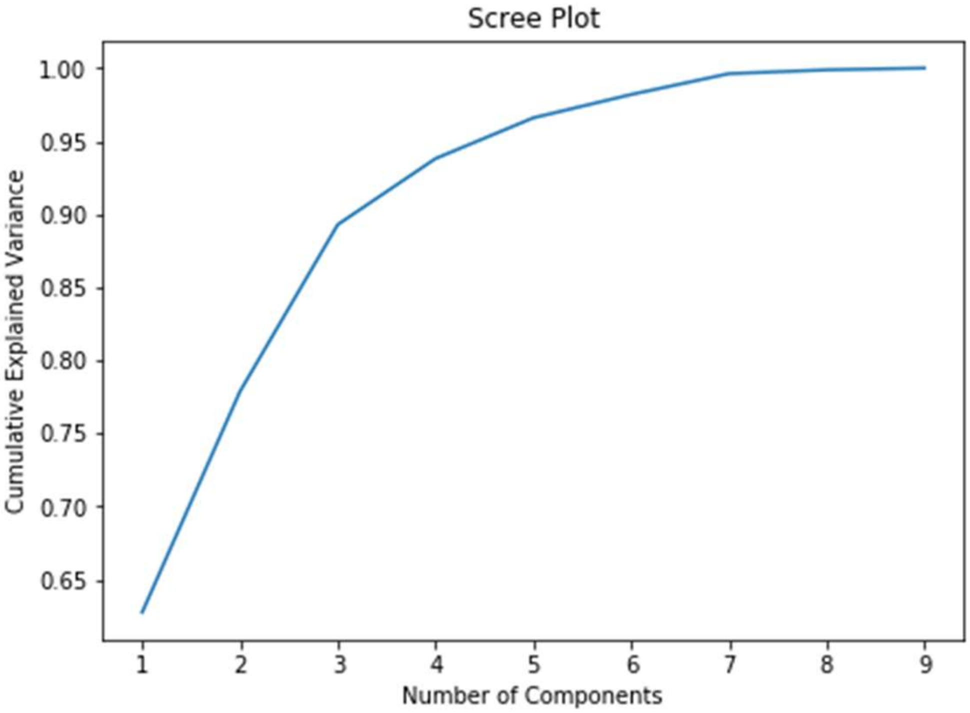
Correlation Matrix



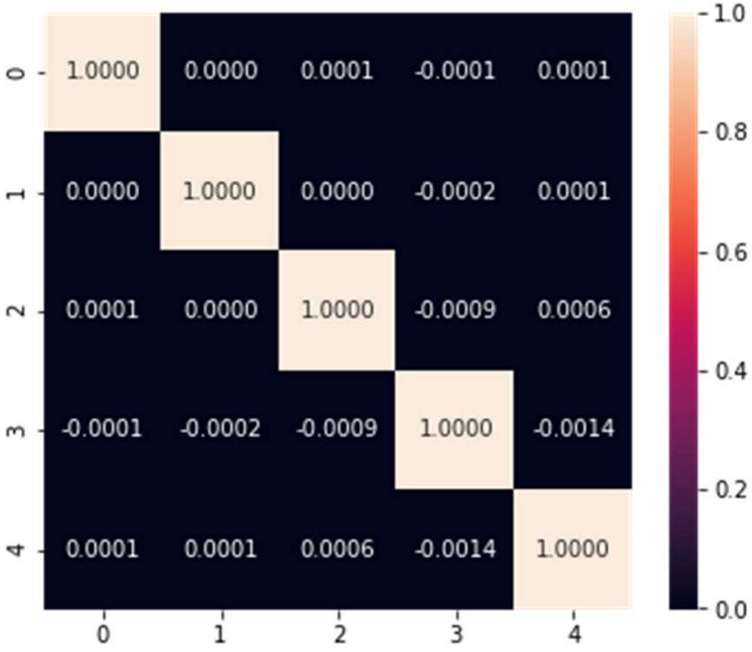
	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
child_mort	1	-0.48	-0.46	-0.54	-0.58	0.18	-0.8	0.84	-0.52
exports	-0.48	1	0.68	0.94	0.89	-0.094	0.48	-0.41	0.86
health	-0.46	0.68	1	0.76	0.7	-0.25	0.53	-0.41	0.93
imports	-0.54	0.94	0.76	1	0.8	-0.29	0.54	-0.48	0.88
income	-0.58	0.89	0.7	0.8	1	0.046	0.59	-0.51	0.88
inflation	0.18	-0.094	-0.25	-0.29	0.046	1	-0.2	0.2	-0.16
life_expec	-0.8	0.48	0.53	0.54	0.59	-0.2	1	-0.73	0.57
total_fer	0.84	-0.41	-0.41	-0.48	-0.51	0.2	-0.73	1	-0.46
gdpp	-0.52	0.86	0.93	0.88	0.88	-0.16	0.57	-0.46	1

- Provided data has been cleaned and outliers have been treated.
- Re-scaled the data with Standardized method.
- As per the the heatmap, we see that some variables like total fertility - child mortality , income - gdpp and imports - exports have high correlation.
- It is a good practice to use PCA for removing the multicollinearity.

Principal Component Analysis



Scree Plot shows that **95%** of the variance is being explained by **5 components.**

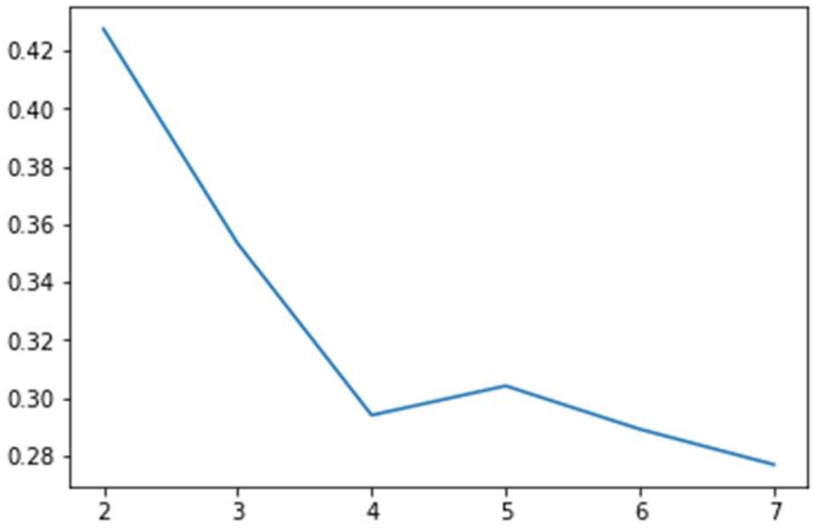


	0	1	2	3	4
0	1.0000	0.0000	0.0001	-0.0001	0.0001
1	0.0000	1.0000	0.0000	-0.0002	0.0001
2	0.0001	0.0000	1.0000	-0.0009	0.0006
3	-0.0001	-0.0002	-0.0009	1.0000	-0.0014
4	0.0001	0.0001	0.0006	-0.0014	1.0000

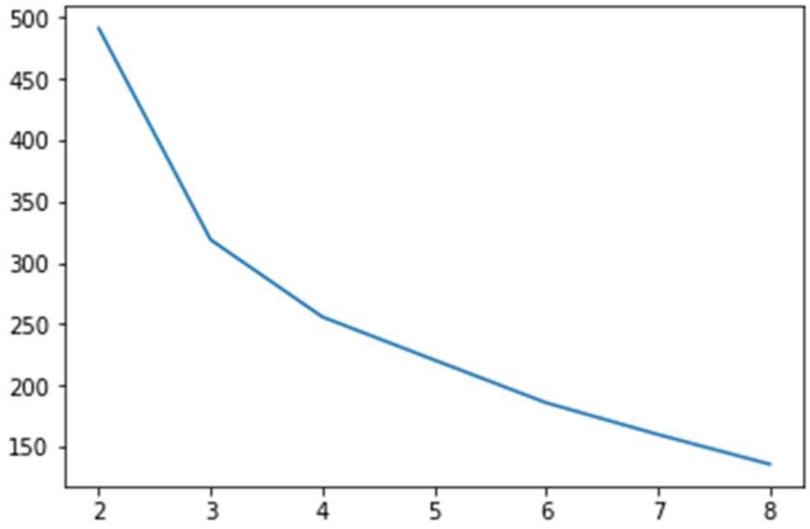
After applying Incremental PCA with 5 components, we see that the correlation in the data has almost reduced to zero.

Clustering : K-Means

Silhouette Analysis

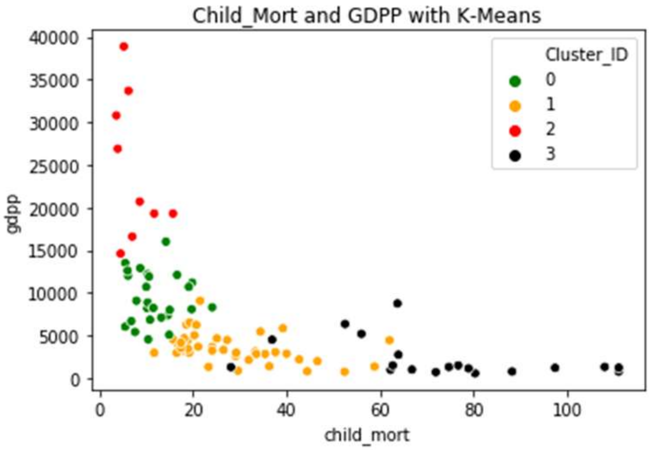


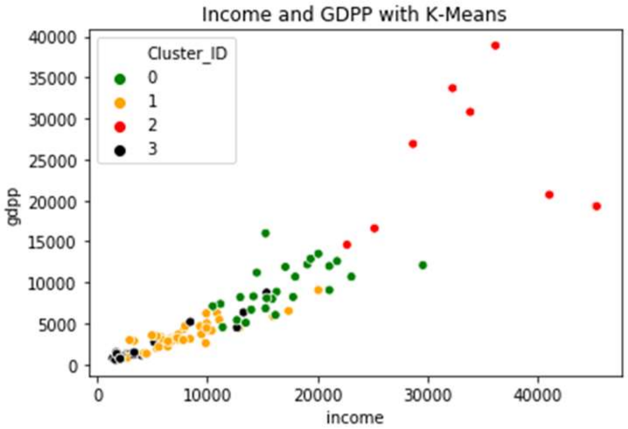
Elbow Curve

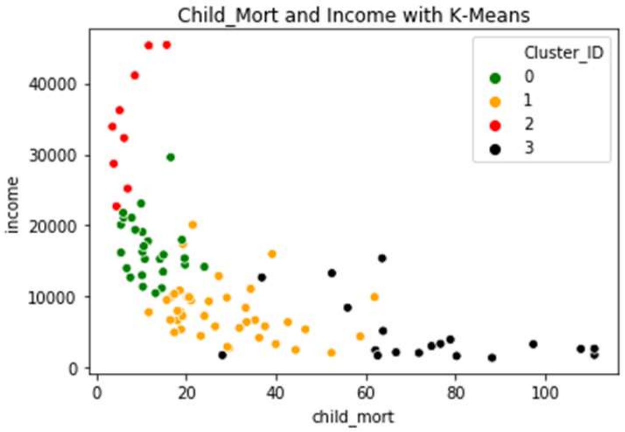


From the above Elbow Curve, we can see that elbow formation is pretty good at cluster no. 4. After this point, any further increase in the cluster, does not change the SSD abruptly. Also, as per the Silhouette Score, it suits 4 clusters would be good. So, let's take k = 4.

Visualization in K-Means Clustering



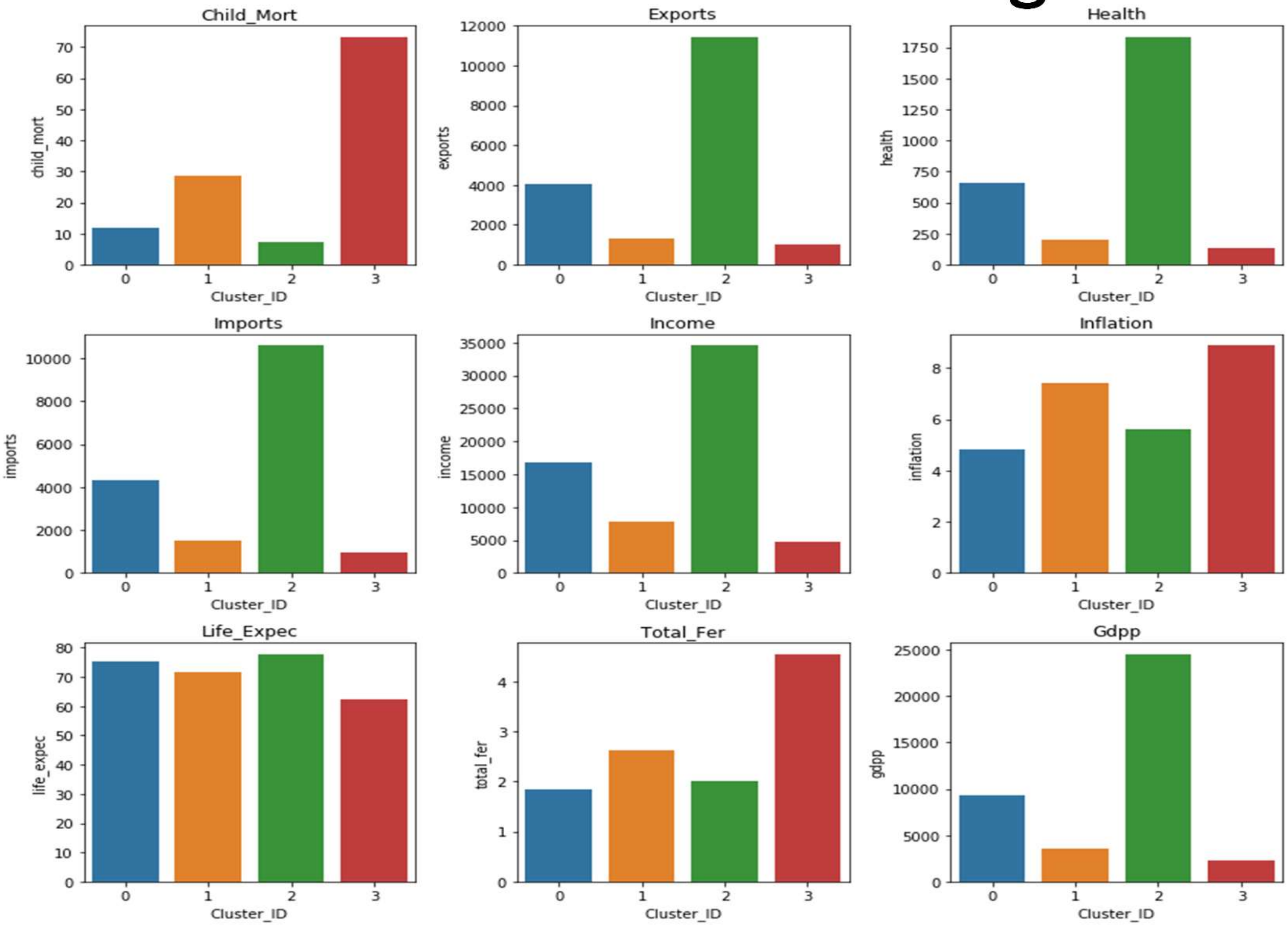




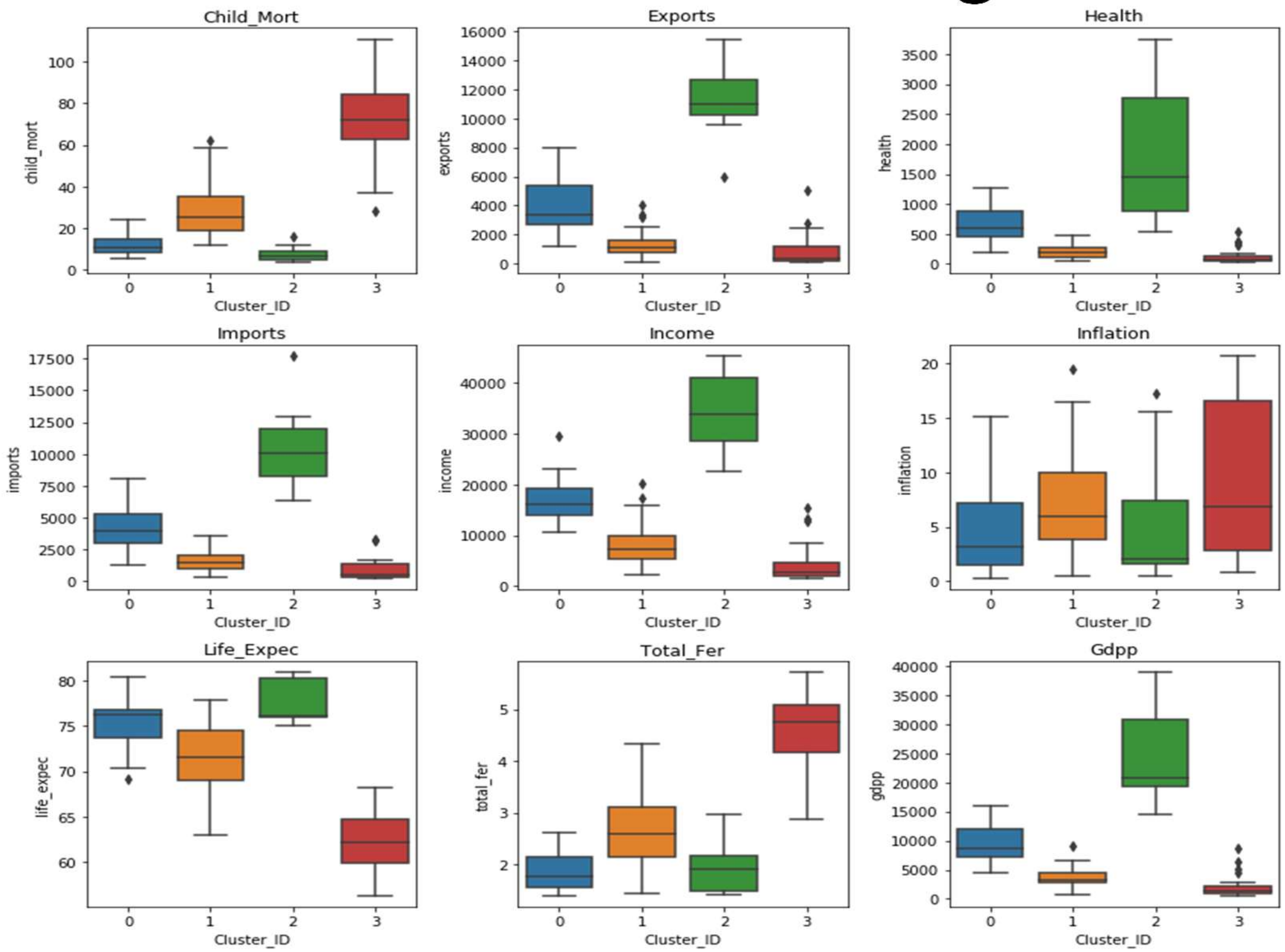
We can see a proper cluster formation after implementing K-Means.

- First graph is a scatter plot between GDPP and Child Mortality. There is high child mortality and low GDPP in Cluster 3.
- Second graph is a scatter plot between GDPP and Income. There is Low income and low GDPP in Cluster 3.
- Third graph is a scatter plot between Income and Child Mortality. There is high child mortality and low income in Cluster 3.

K-Means : Visualization Through Bar Plot



K-Means : Visualization Through Box Plot



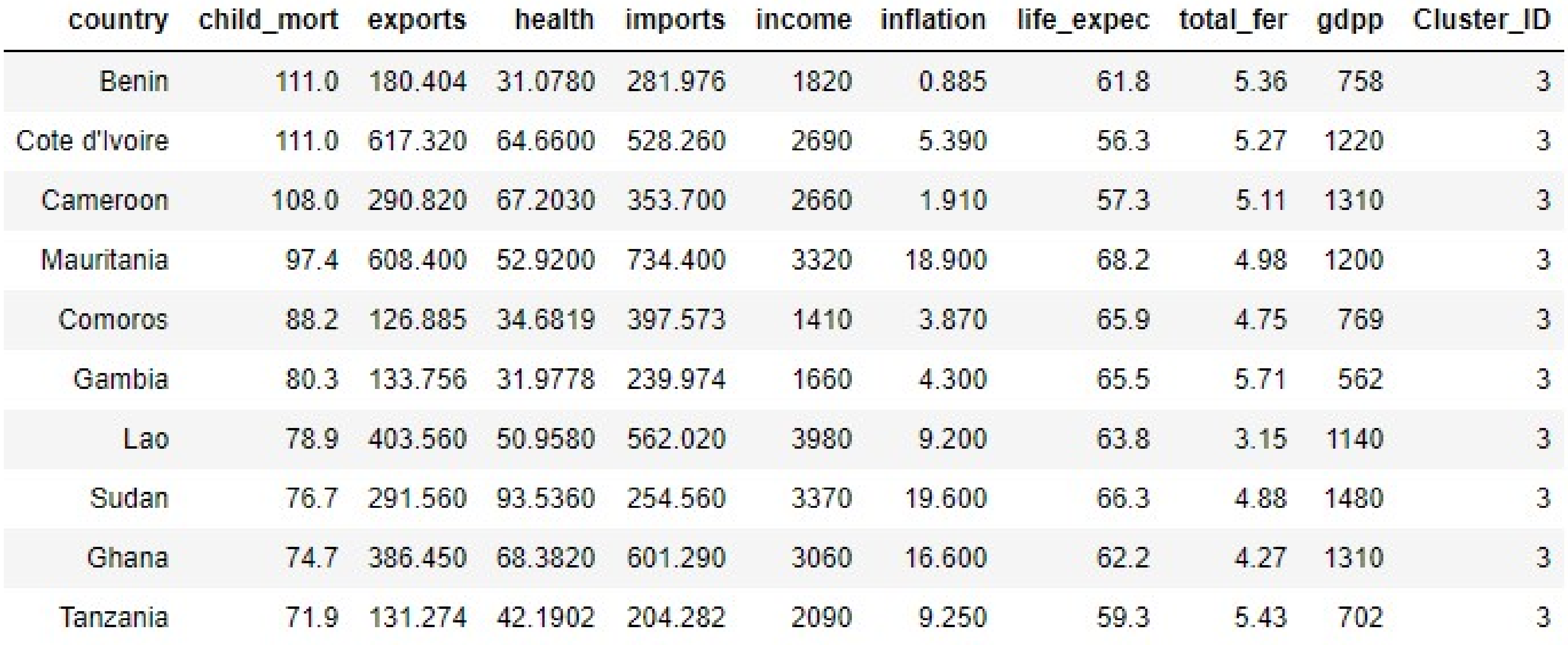
Analysis : K-Means Clustering

After implementing K-Means Clustering, if we look at previously mentioned Bar plots and Boxplots, we can observe that Cluster 3 needs more attention. Below points explain that Cluster 3 has:

- Highest Child Mortality
- Lowest Income
- Highest Inflation
- Lowest Life Expectancy
- Highest Total Fertility
- Lowest GDPP
- Lowest Health Spending

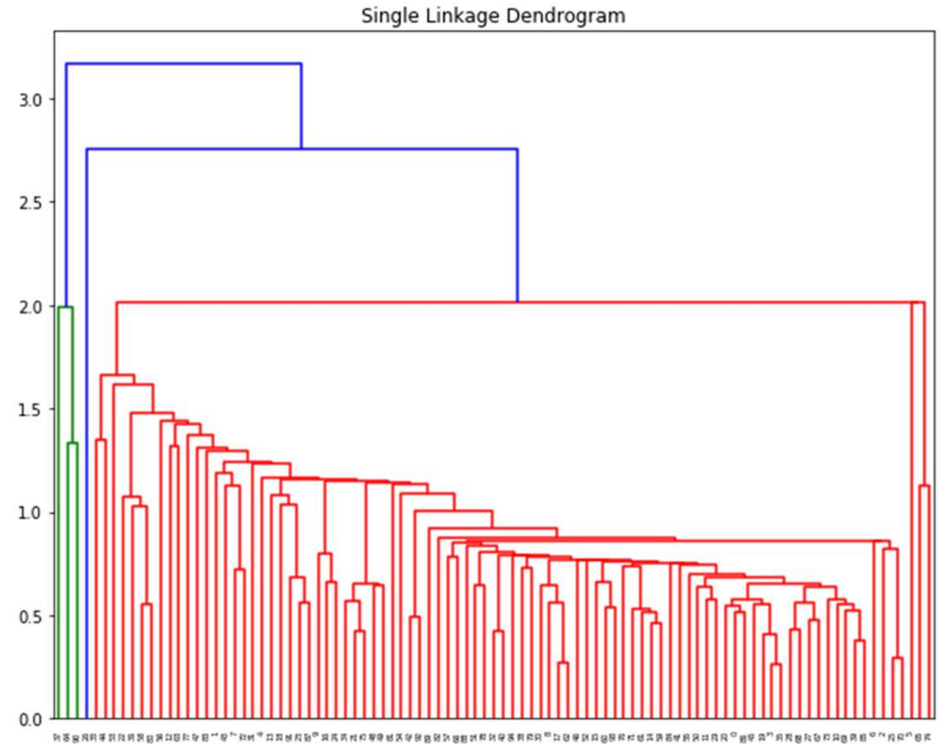
K-Means : Countries in Cluster 3

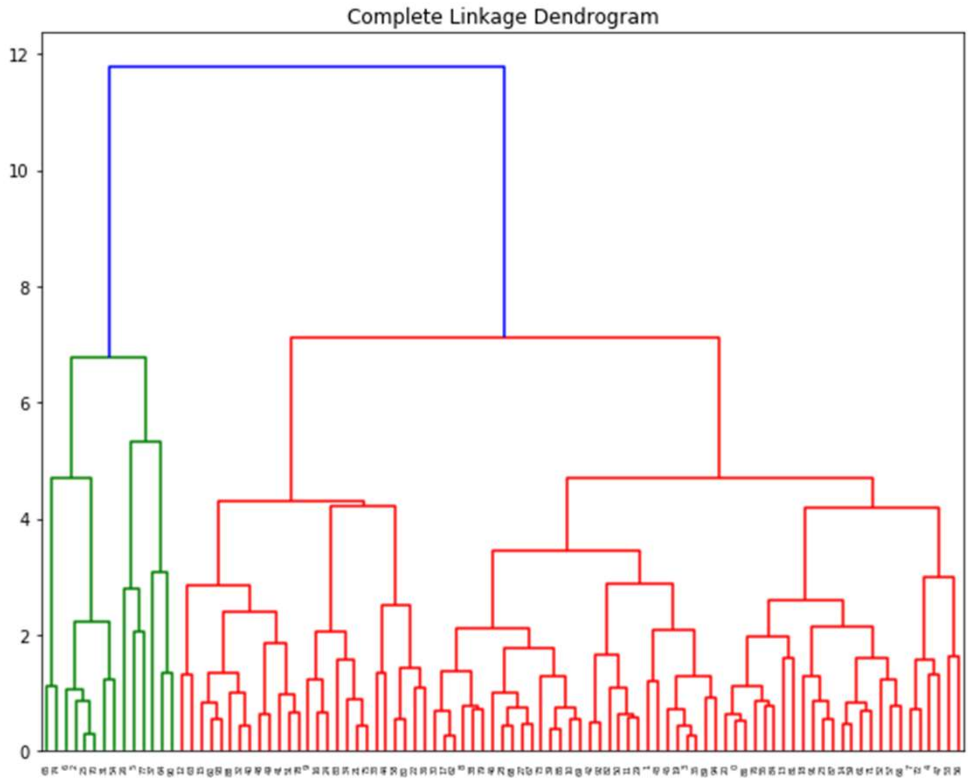
10 countries under Cluster 3 are mentioned below:



country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	Cluster_ID
Benin	111.0	180.404	31.0780	281.976	1820	0.885	61.8	5.36	758	3
Cote d'Ivoire	111.0	617.320	64.6600	528.260	2690	5.390	56.3	5.27	1220	3
Cameroon	108.0	290.820	67.2030	353.700	2660	1.910	57.3	5.11	1310	3
Mauritania	97.4	608.400	52.9200	734.400	3320	18.900	68.2	4.98	1200	3
Comoros	88.2	126.885	34.6819	397.573	1410	3.870	65.9	4.75	769	3
Gambia	80.3	133.756	31.9778	239.974	1660	4.300	65.5	5.71	562	3
Lao	78.9	403.560	50.9580	562.020	3980	9.200	63.8	3.15	1140	3
Sudan	76.7	291.560	93.5360	254.560	3370	19.600	66.3	4.88	1480	3
Ghana	74.7	386.450	68.3820	601.290	3060	16.600	62.2	4.27	1310	3
Tanzania	71.9	131.274	42.1902	204.282	2090	9.250	59.3	5.43	702	3

Clustering : Hierarchical

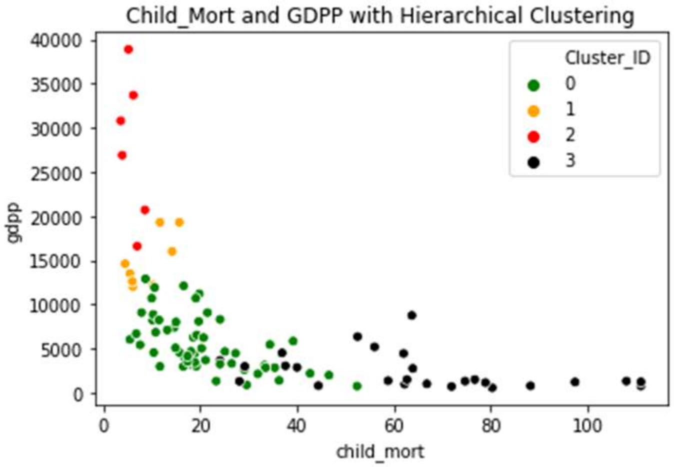


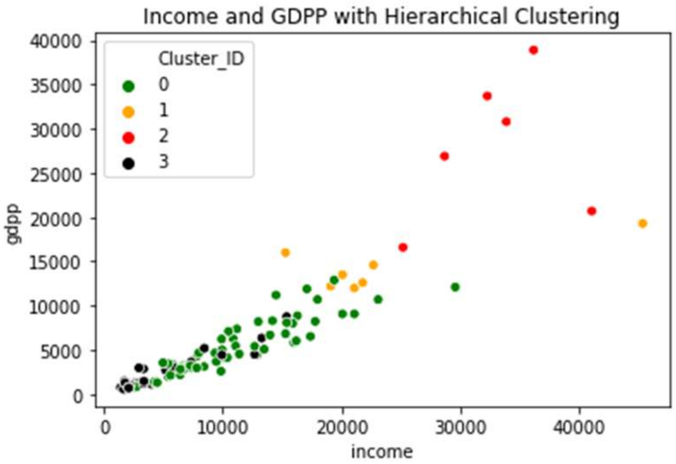


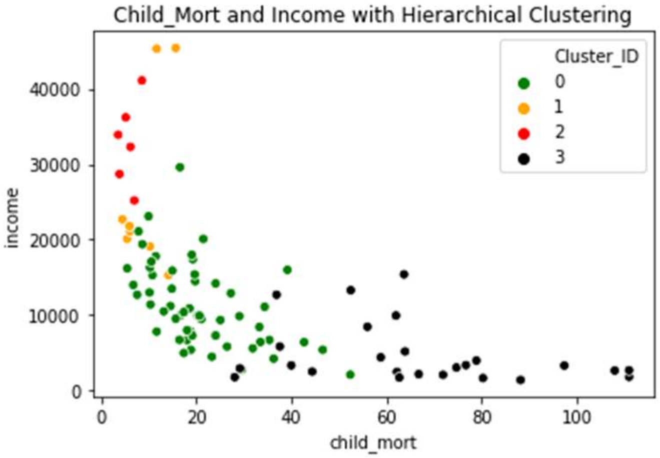
There is no proper visualization in Single Linkage Dendrogram.

- Visualization is proper in Complete Linkage Dendrogram.
- Let’s cut the tree at the height to get 4 clusters.

Visualization in Hierarchical Clustering


Child_Mort and GDPP with Hierarchical Clustering

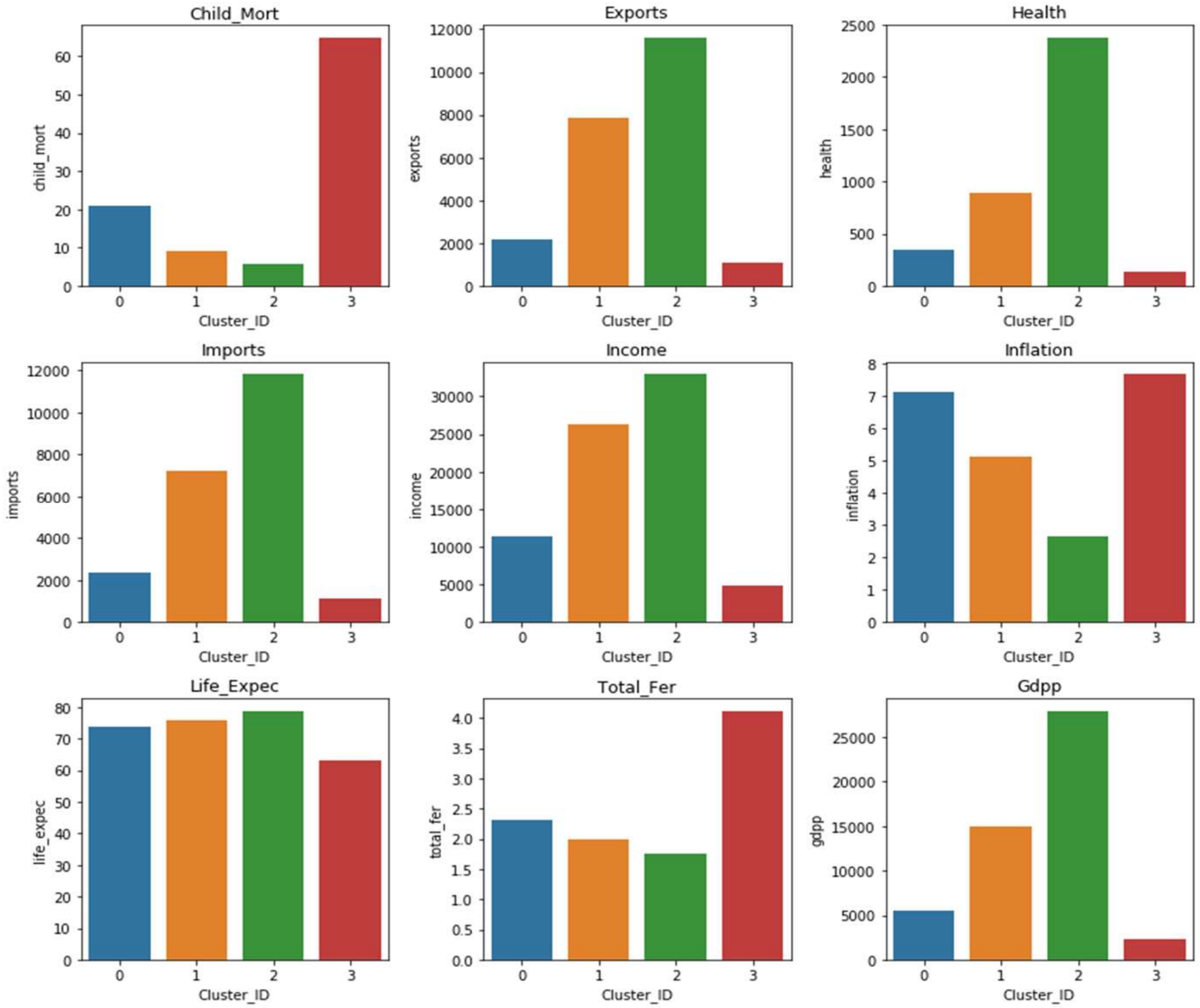

Income and GDPP with Hierarchical Clustering


Child_Mort and Income with Hierarchical Clustering

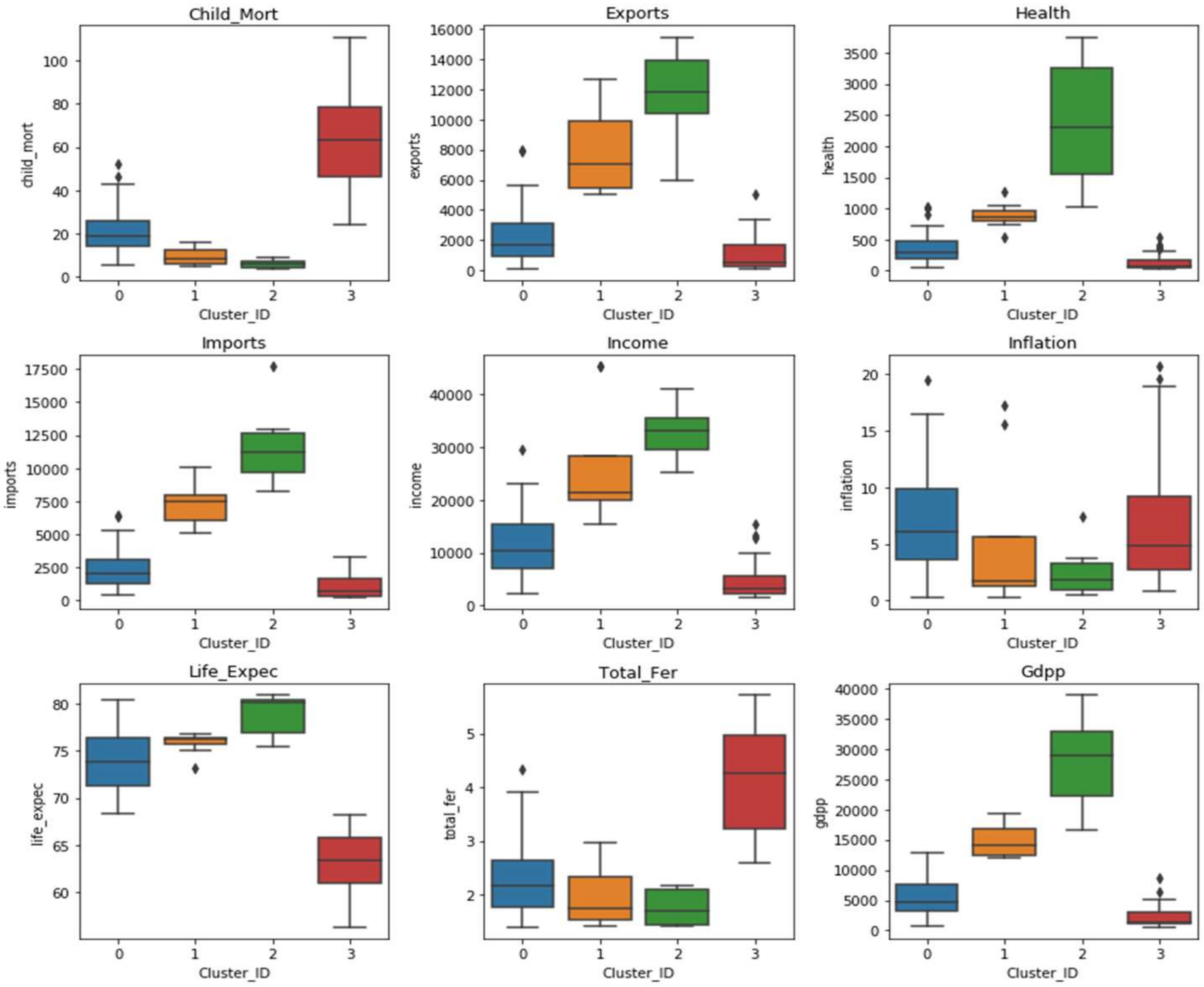
We can see a proper cluster of data points after implementing Hierarchical Clustering.

- First graph is a scatter plot between GDPP and Child Mortality. There is high child mortality and low GDPP in Cluster 3.
- Second graph is a scatter plot between GDPP and Income. There is Low income and low GDPP in Cluster 3.
- Third graph is a scatter plot between Income and Child Mortality. There is high child mortality and low income in Cluster 3.

Hierarchical : Visualization Through Bar Plot



Hierarchical : Visualization Through Box Plot



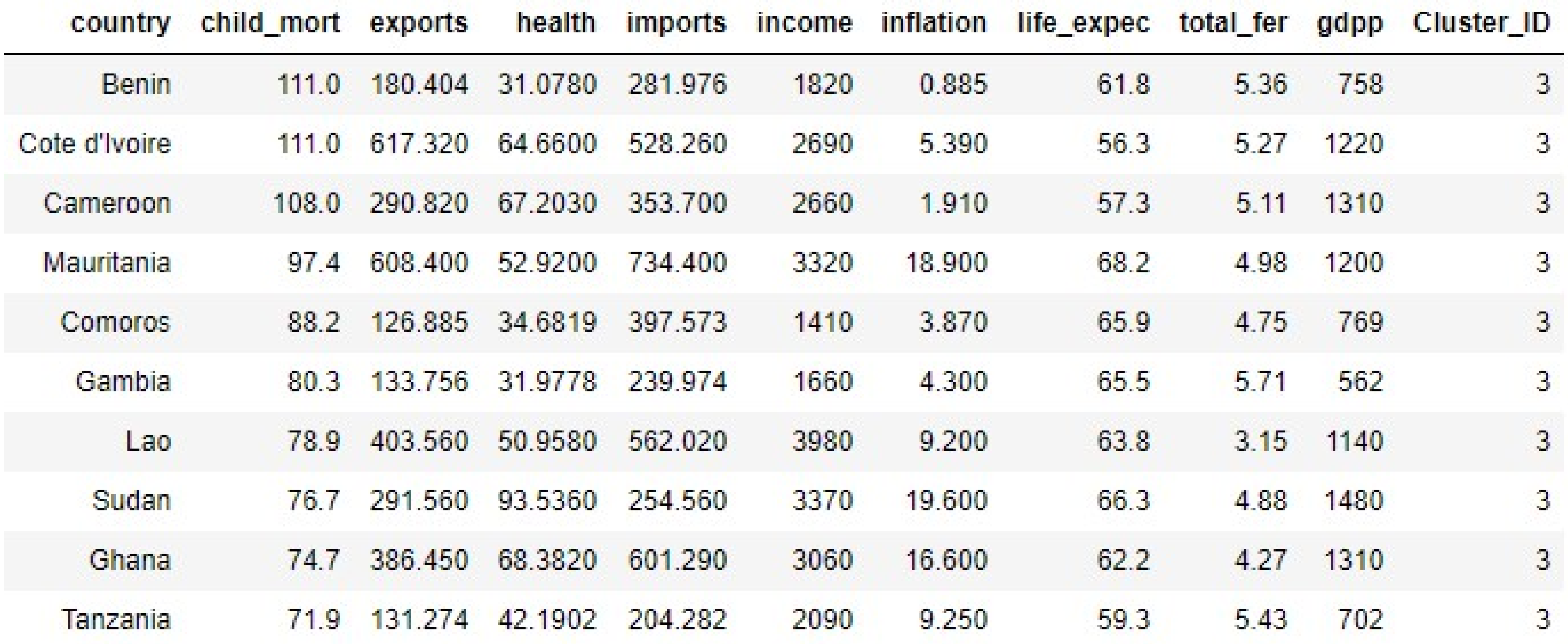
Analysis : Hierarchical Clustering

After implementing Hierarchical Clustering, if we look at previously mentioned Bar plots and Boxplots, we can observe that Cluster 3 needs more attention. Below points explain that Cluster 3 has:

- Highest Child Mortality
- Lowest Income
- Highest Inflation
- Lowest Life Expectancy
- Highest Total Fertility
- Lowest GDPP
- Lowest Health Spending

Hierarchical : Countries in Cluster 3

10 countries under Cluster 3 are mentioned below:



country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	Cluster_ID
Benin	111.0	180.404	31.0780	281.976	1820	0.885	61.8	5.36	758	3
Cote d'Ivoire	111.0	617.320	64.6600	528.260	2690	5.390	56.3	5.27	1220	3
Cameroon	108.0	290.820	67.2030	353.700	2660	1.910	57.3	5.11	1310	3
Mauritania	97.4	608.400	52.9200	734.400	3320	18.900	68.2	4.98	1200	3
Comoros	88.2	126.885	34.6819	397.573	1410	3.870	65.9	4.75	769	3
Gambia	80.3	133.756	31.9778	239.974	1660	4.300	65.5	5.71	562	3
Lao	78.9	403.560	50.9580	562.020	3980	9.200	63.8	3.15	1140	3
Sudan	76.7	291.560	93.5360	254.560	3370	19.600	66.3	4.88	1480	3
Ghana	74.7	386.450	68.3820	601.290	3060	16.600	62.2	4.27	1310	3
Tanzania	71.9	131.274	42.1902	204.282	2090	9.250	59.3	5.43	702	3

Conclusion

As per the analysis using both K-Means and Hierarchical Clustering, we got same list of countries. So, below are the countries that are in direst need of aid by considering socio – economic factor into consideration:

1. Benin

2. Cote d'Ivoire

3. Cameroon

4. Mauritania

5. Comoros

6. Gambia

7. Lao

8. Sudan

9. Ghana

10. Tanzania