Clustering Assignment

TRIPURA RAJAVARAPU

Problem Statement

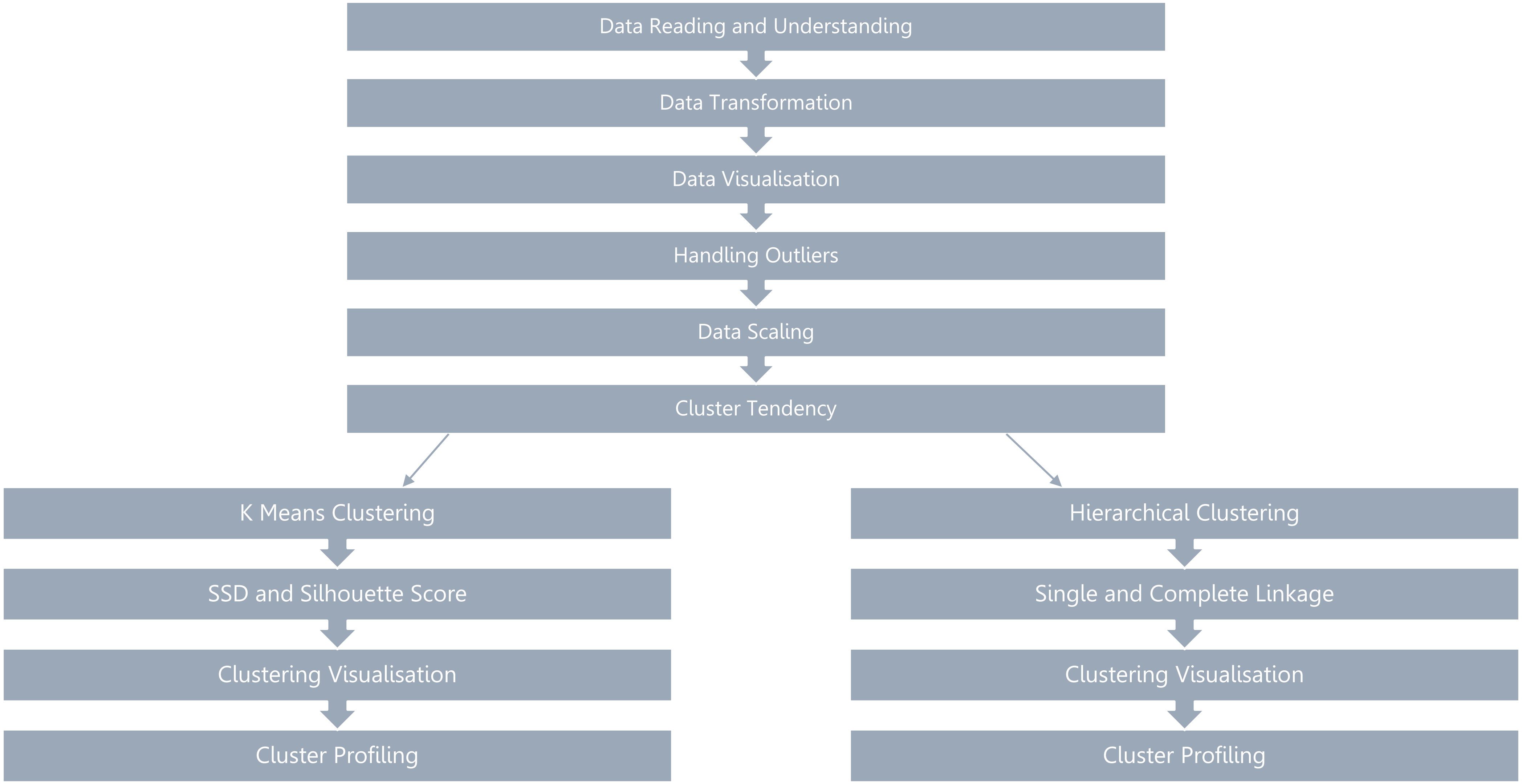
Business Understanding:

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. After the recent funding programmes, they have been able to raise around $ 10 million. Now 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.

Agenda:

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

Approach for Analysis



Data Reading and Understanding

Data Transformation

Data Visualisation

Handling Outliers

Data Scaling

Cluster Tendency

K Means Clustering

SSD and Silhouette Score

Clustering Visualisation

Cluster Profiling

Hierarchical Clustering

Single and Complete Linkage

Clustering Visualisation

Cluster Profiling

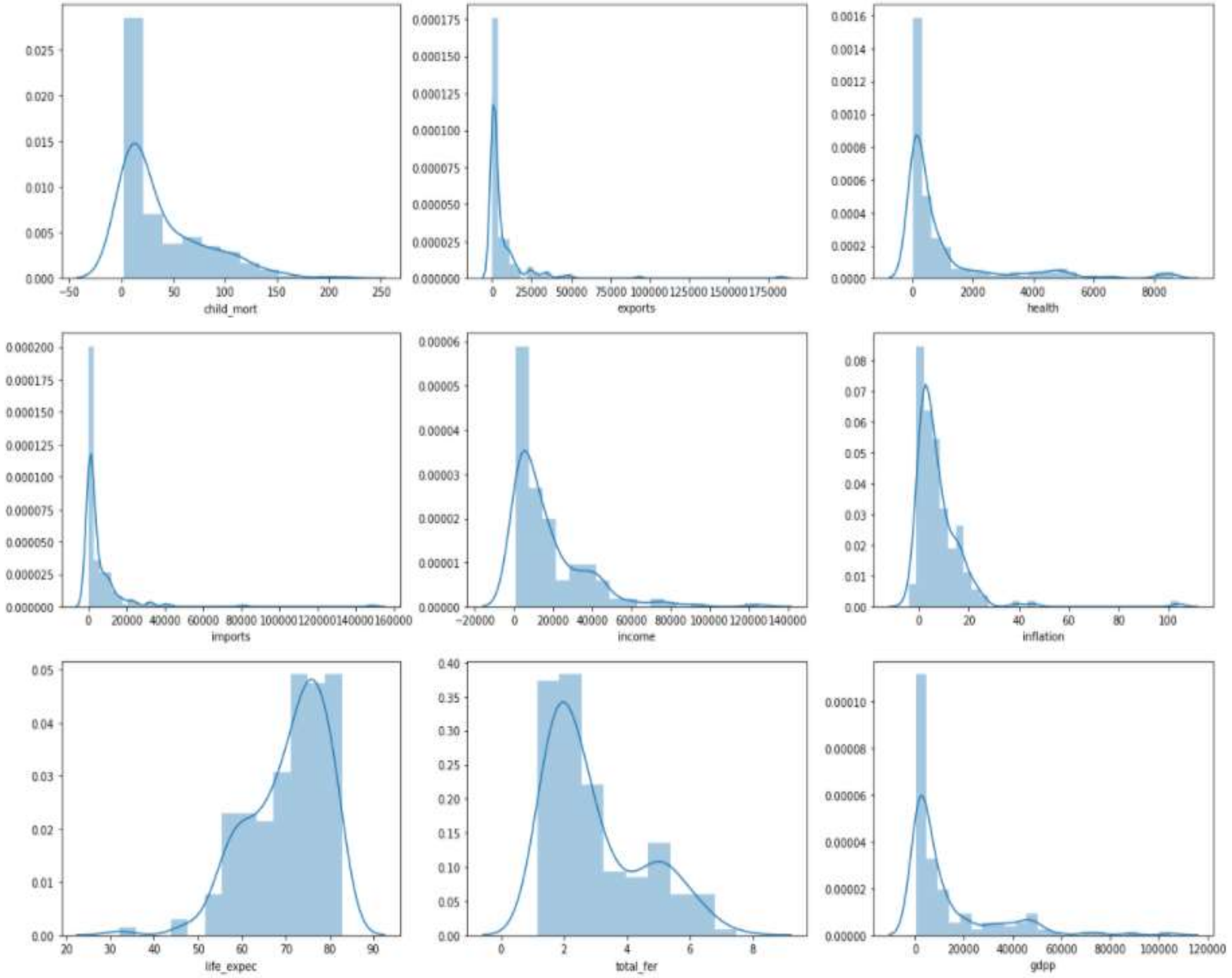
Analysis on Application Data

Univariate Analysis numerical variables

Analysis

- ✓ Inflation, health, child_mort, imports ,exports, inflation are normally distributed .
- ✓ High values of child mort rate are concentrated approximately between 0 and 25
- ✓ Values of gdpp are concentrated between 0 and 20000 and also between 40000 and 60000
- ✓ Total fertility column values are concentrated more between 1 and 3.8 and also between above 4 and 6
- ✓ Most of the countries income is ranging between 0-20000 and also between 30000-50000

Univariate Analysis on Numerical Variables



Correlation between Numerical Variables

Analysis

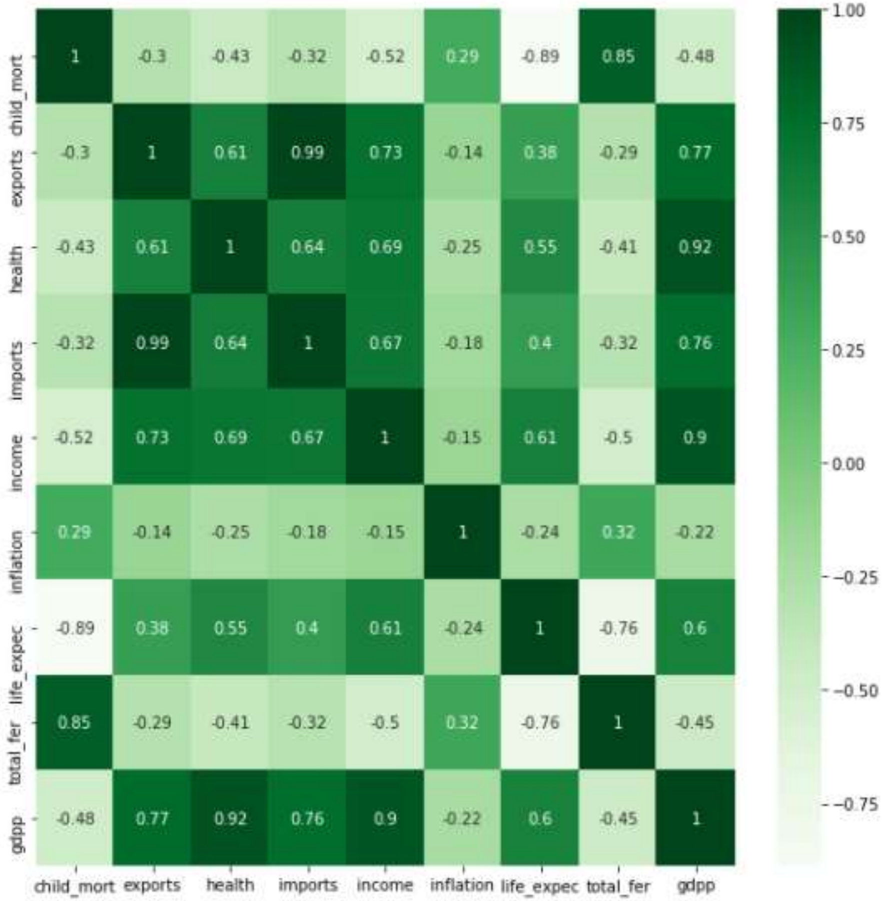
There is more positive correlation between

- ✓ imports and exports is the highest which is 0.99
- ✓ health and gdpp it is 0.92
- ✓ income and gdpp it is 0.9
- ✓ child_mort and total_fer it is 0.85

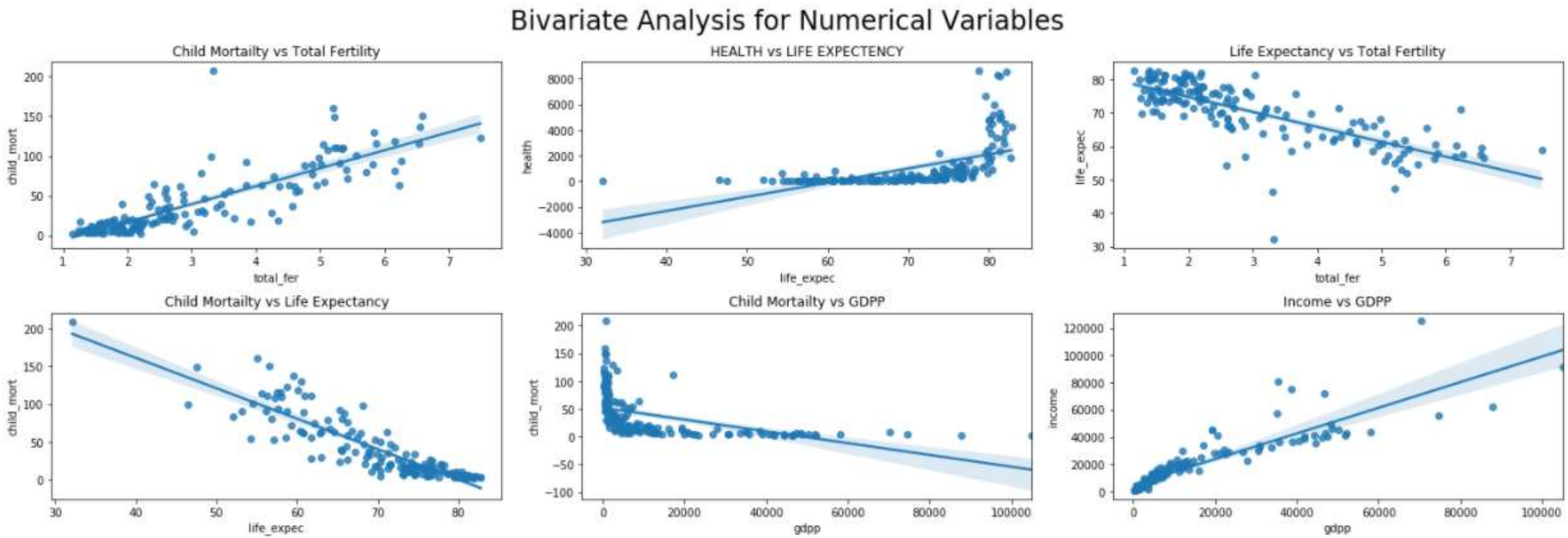
There is more negative correlation between

- ✓ child_mort and life_expec is the lowest which is -0.89
- ✓ life_expec and total_fer it is -0.76

Checking for multi-collinearity



	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



Analysis:

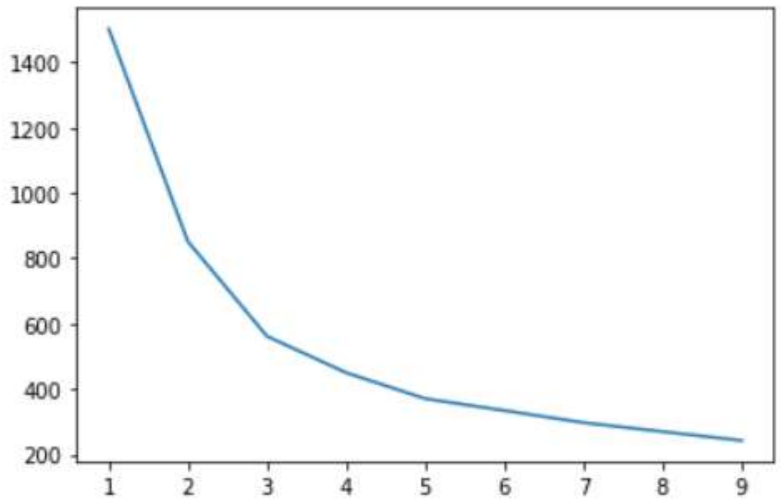
- From the above plot, it is evident that there is high negative correlation between life expectancy and total fertility and child mortality and life expectancy
- There is more positive correlation between total fertility and child mortality and also between gdpp and income

K Means Clustering

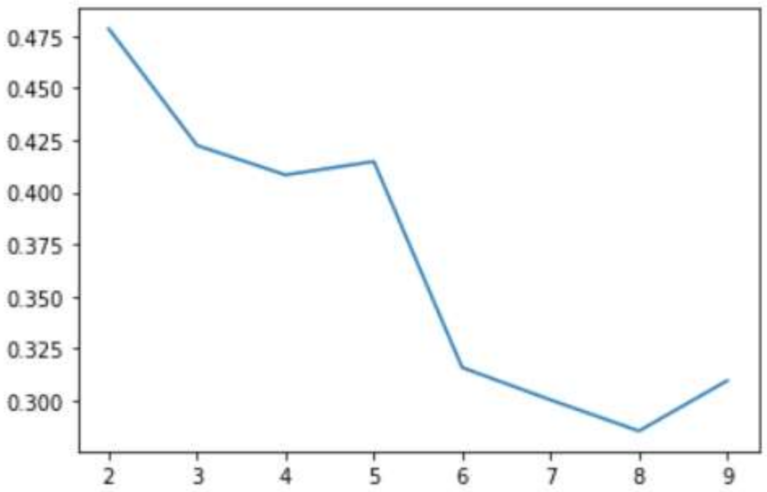
SSD and Silhouette Score

Based on both the graphs the optimal number of clusters is taken as 3 which means k=3

SSD(Elbow Curve)



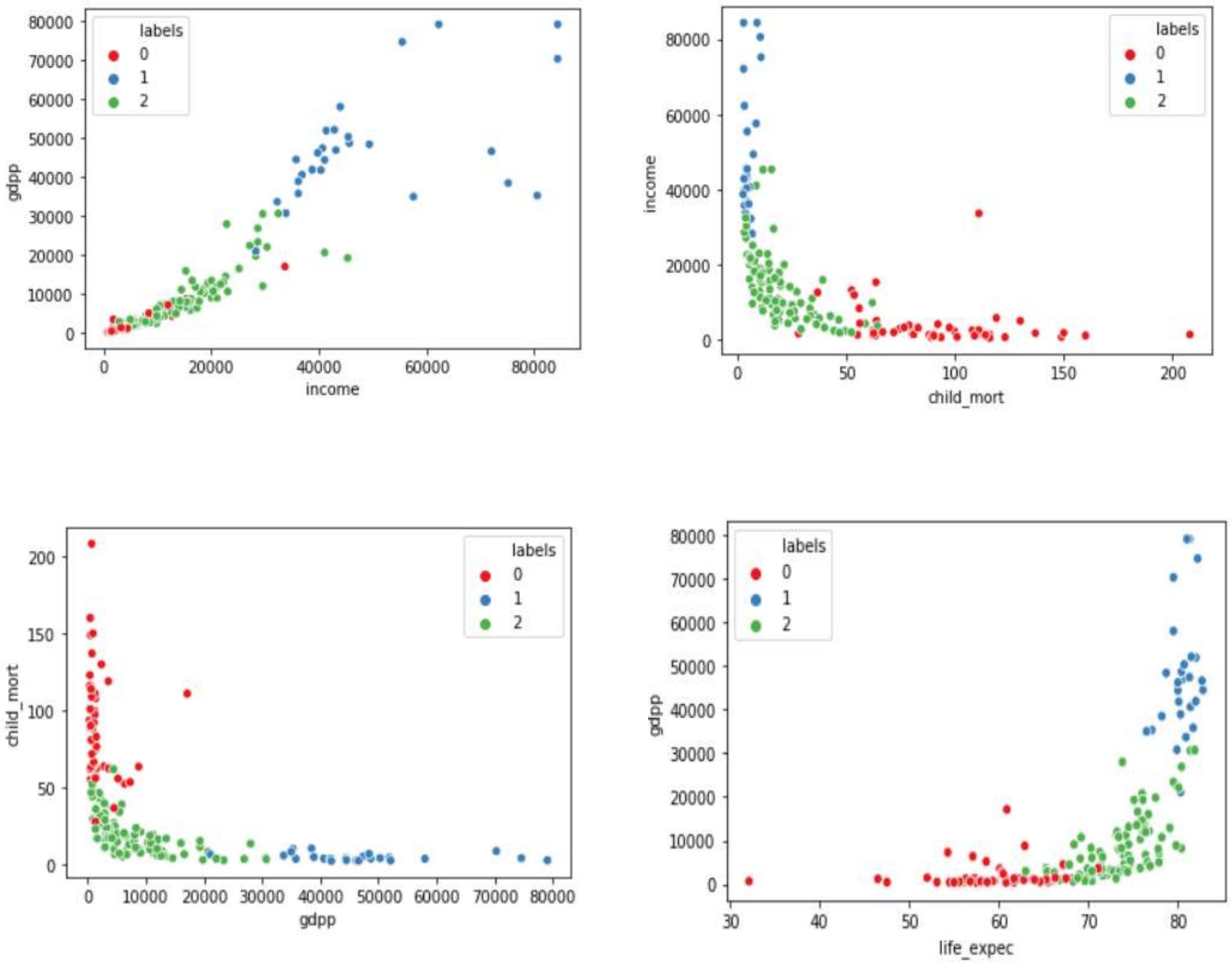
Silhouette Score



Clustering Visualization

Analysis from the graphs

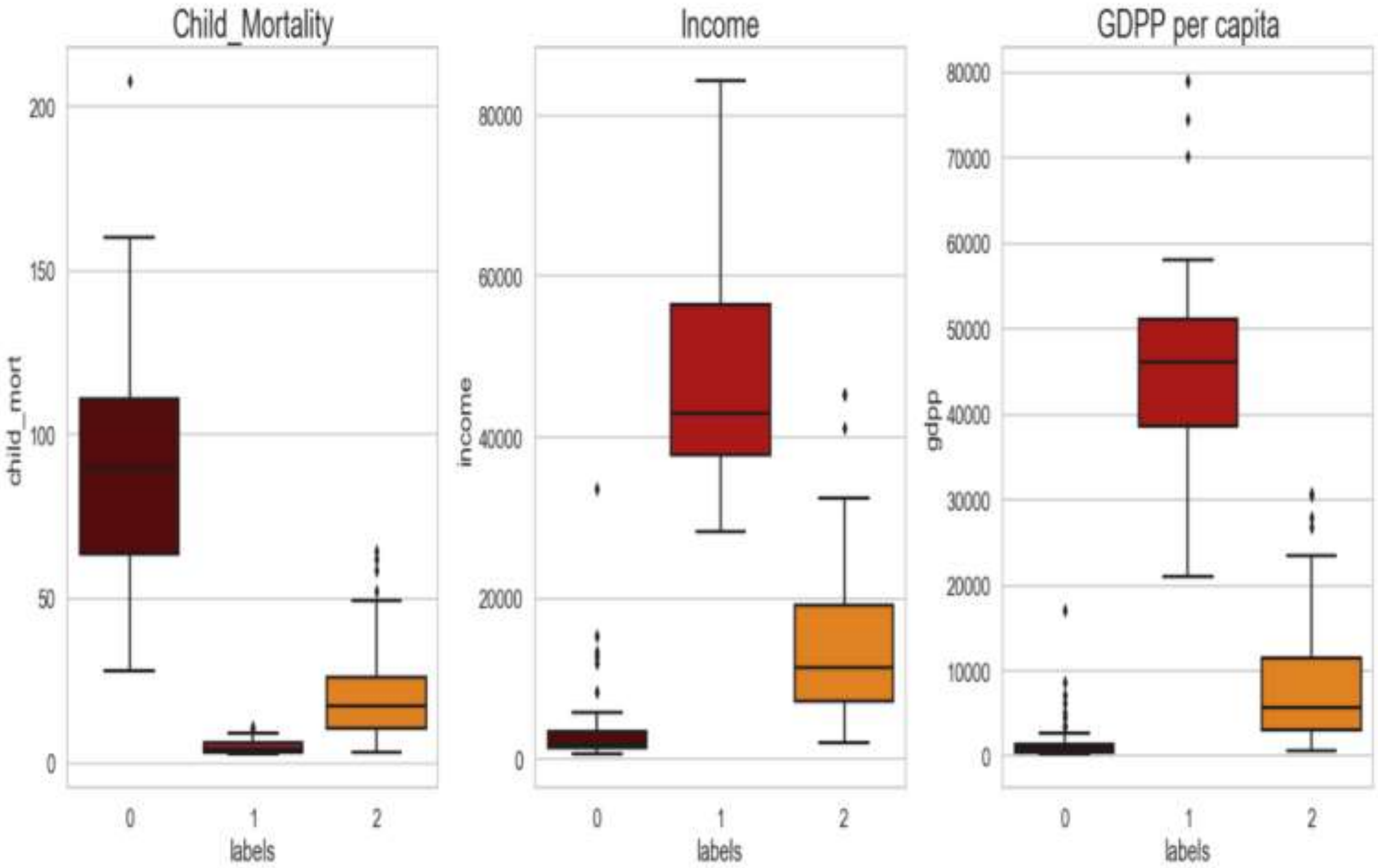
- ✓ We can see that the highest income and highest GDPP belongs to cluster 1.
- ✓ It can be seen that cluster 0 is concentrated with low income and high child mortality rate
- ✓ It is evident that cluster label 0 is having high child mortality rate and low GDPP
- ✓ It can be seen from the graph that the more life expectancy and high GDPP belongs to cluster '1'



Cluster Profiling

The countries with low income and GDPP and high child mortality are the countries which need aid

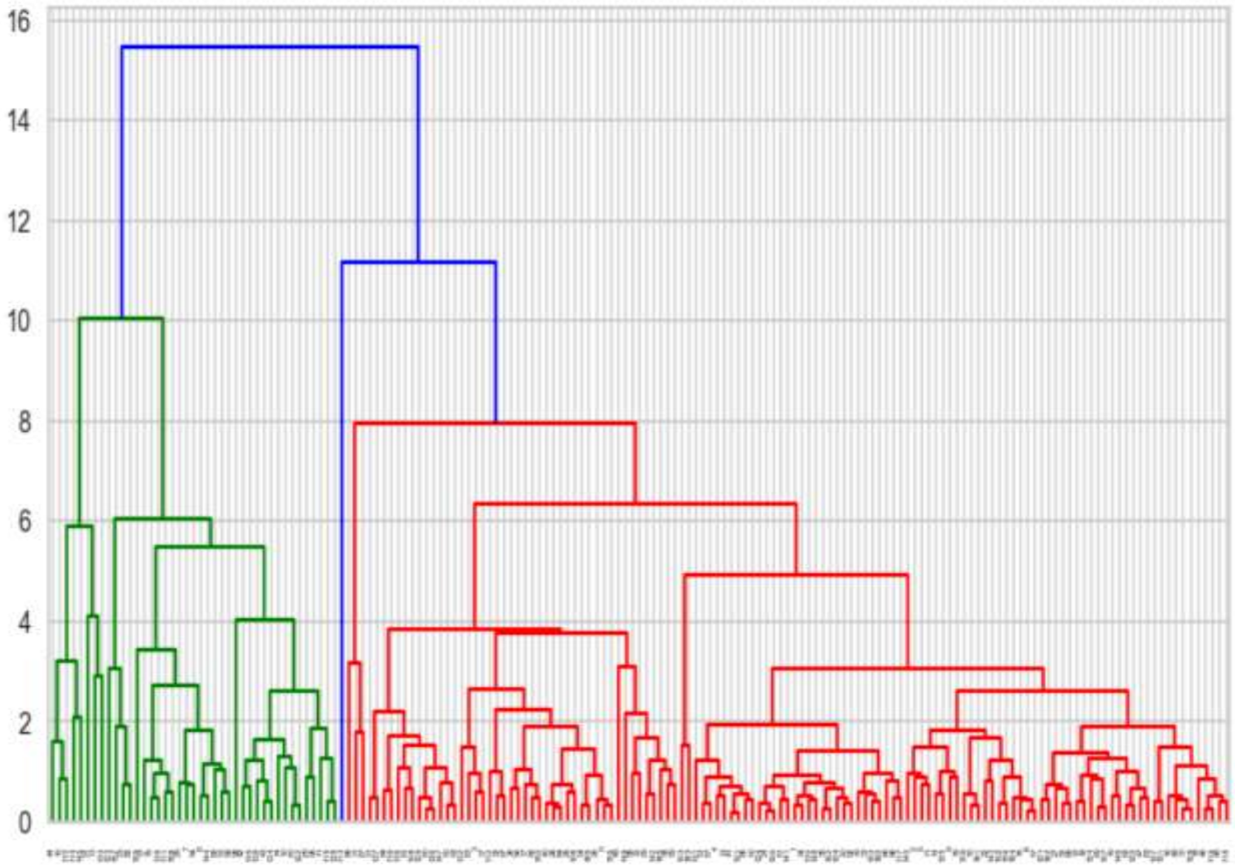
It is evident from the graph that the cluster 0 meet the requirement



Hierarchical Clustering

Complete Linkage

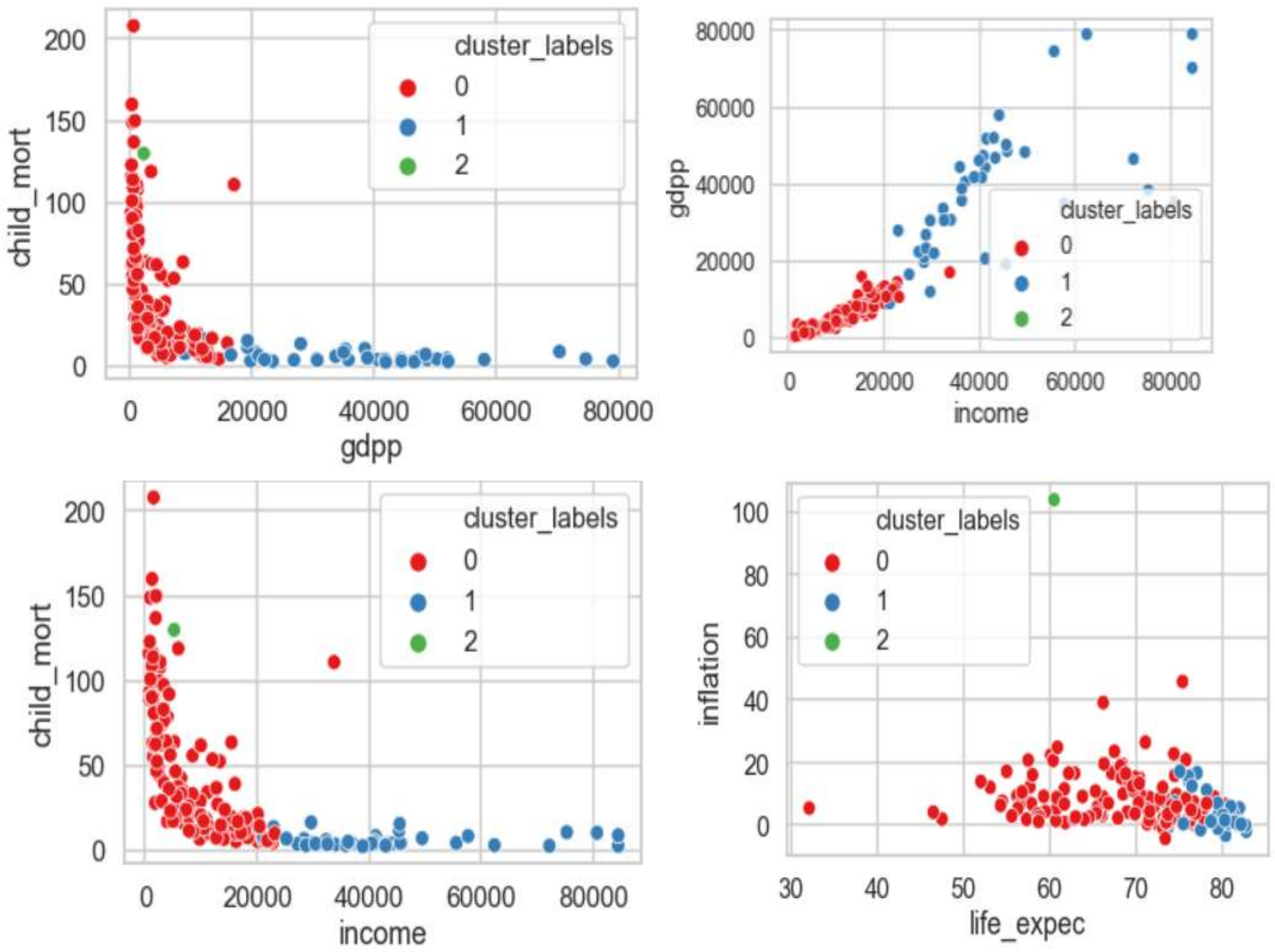
Looking at dendrogram of hierarchical clustering there seem to be 3 clusters



Clustering Visualization

Analysis from the graphs

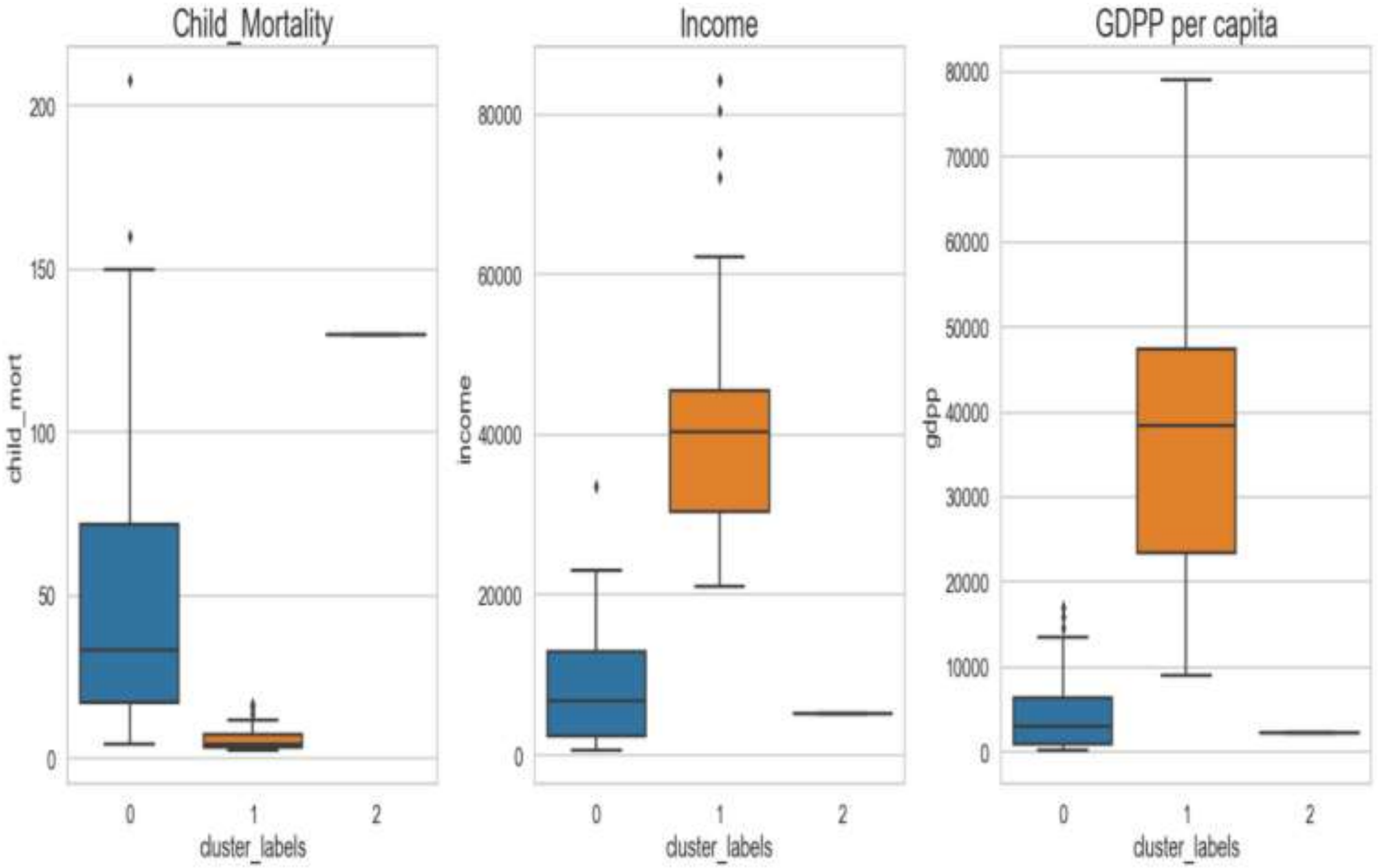
- ✓ We can see that the low GDPP[0-20000] and child mortality ranging between [0-200] are having cluster label as ‘0’
- ✓ It is also evident that there is low income and low GDPP for cluster label 0
- ✓ It can be seen that the low income and high mortality rate are having cluster label as ‘0’
- ✓ We can see that the inflation below 20 and high life expectancy belongs to cluster '1'



Cluster Profiling

The countries with low income and GDPP and high child mortality are the countries which need aid

It is evident from the graph that the cluster 0 meet the requirement



Conclusions

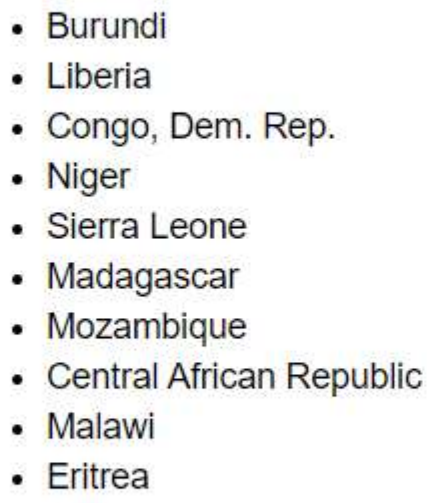
K-means clustering:

- Countries that are direst need of aid-Total 48 countries are in this category

Hierarchical clustering :

- Countries that are direst need of aid-Total 125 countries are in this category

The top 10 countries derived are same by both Hierarchial and KMeans clustering. Below are the list of countries which are in direst need of aid:

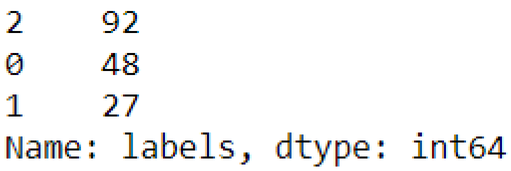


- Burundi
- Liberia
- Congo, Dem. Rep.
- Niger
- Sierra Leone
- Madagascar
- Mozambique
- Central African Republic
- Malawi
- Eritrea

Recommended Clustering procedure

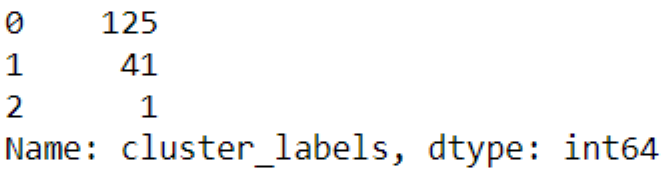
Although, both the clustering methods gave top 10 countries as same, through K-Means clustering better clusters division is done when compared to Hierarchical clustering as shown below. Hence, K-Means clustering gave precise information than Hierarchical clustering.

K Means Clustering



```
2    92
0    48
1    27
Name: labels, dtype: int64
```

Hierarchical Clustering



```
0    125
1     41
2      1
Name: cluster_labels, dtype: int64
```

THANK YOU