Clustering Assignment

HELP NGO FUNDING DECISION

- BY DEBOPRIYA GHOSH

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. 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.

Objective

- Objective is to segment the countries using some socio-economic and health factors that determine the overall development of the country.

- Then we need to suggest the countries which the CEO needs to focus on the most for considering for NGO Aid.

Analysis Approach

- We will start off with the necessary data inspection and EDA tasks suitable for this dataset - data cleaning, univariate analysis, bivariate analysis etc.

- We will perform the **Outlier Analysis** on the dataset. However, based on the business needs we will either keep the outliers, if it suits the business or remove them. Based on the outlier presence and data volume, we can perform capping or ignore them I required.

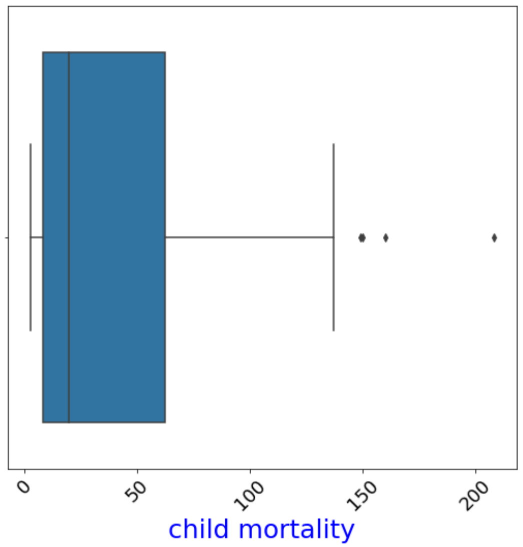
- After that we will try both K-means and Hierarchical clustering (both single and complete linkage) on this dataset to create the clusters. Both the methods may not produce identical results and we might have to choose one of them for the final list of countries.

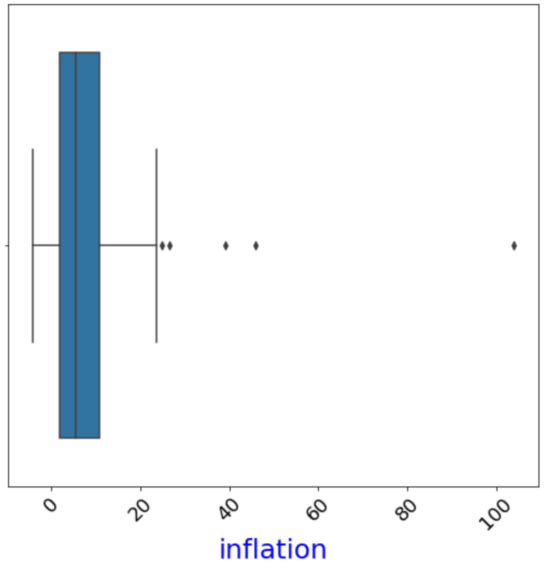
- Analyze the clusters and identify the ones which are in dire need of aid. We will analyze the clusters by comparing how **GDPP**, **child mortality** and **income** are varying for each cluster of countries and differentiate the clusters of developed countries from the clusters of under-developed countries.

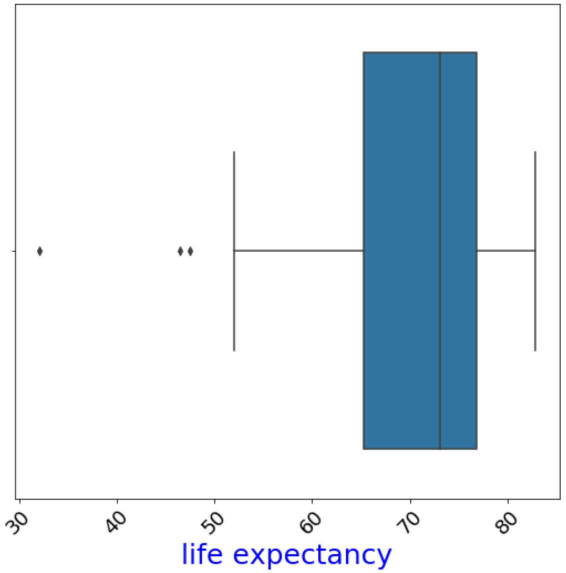
- We need to perform visualizations on the clusters that have been formed. We will plot a scatter plot of all the countries and differentiating the clusters. We will plot bar graph and box plot for cluster profiling.

- Finally, based on our analysis on all socio-economic features we need to report back at least 5 countries which are in direst need of aid from the analysis work that you perform.

Univariate Analysis : Outlier Analysis







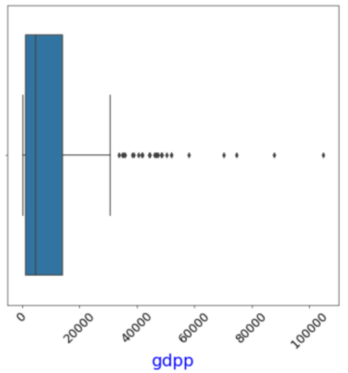
- The ***child mortality*** rate, also under-five mortality rate, refers to the probability of dying between birth and exactly five years of age. From above box plot of child mortality column, we can see that few outliers exist above upper whisker (95% value). That means for some countries the child mortality is extreme high and needs to be consider for aids. Hence, we will not perform any outlier treatment for this column.

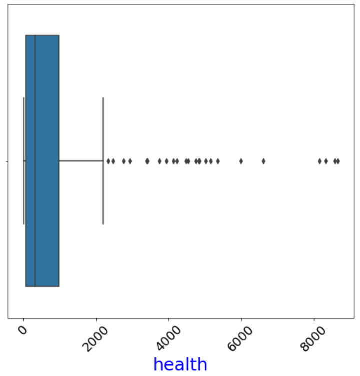
- The ***inflation*** column represents measurement of the annual growth rate of the GDP deflator. More precisely, it signifies a general increase in prices and fall in the purchasing value of money. From above box plot of inflation column, we can see that few outliers exist above upper whisker (95% value). That means for some countries inflation is extreme high and needs to be consider for aids. Hence, we will not perform any outlier treatment for this column.

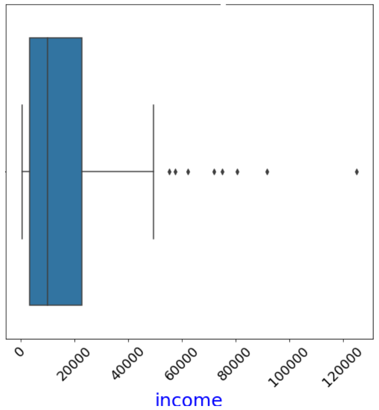
- The ***life expectancy*** column signifies the average number of years a new born child would live. From above box plot of life expectancy column, we can see that few outliers exist below lower whisker (5% value). That means for some countries life expectancy is extreme low and needs to be consider for aids. Hence, we will not perform any outlier treatment for this column.

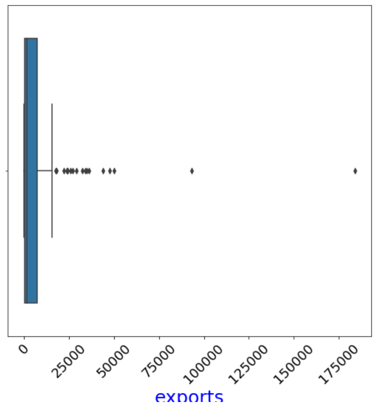
Univariate Analysis : Outlier Analysis

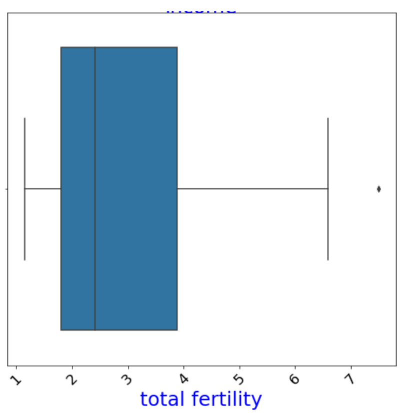
Outlier Treatment: For rest of the columns like imports, health, exports, income etc., we can see that outliers exists above the upper whisker (95% value). That means those countries do not need any aid. Hence, we will perform outlier treatment on those columns.

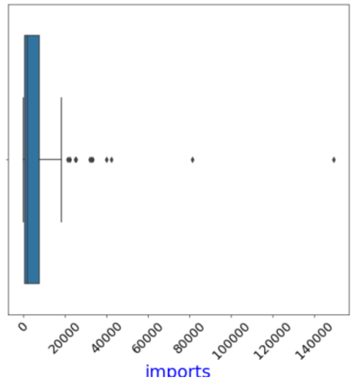




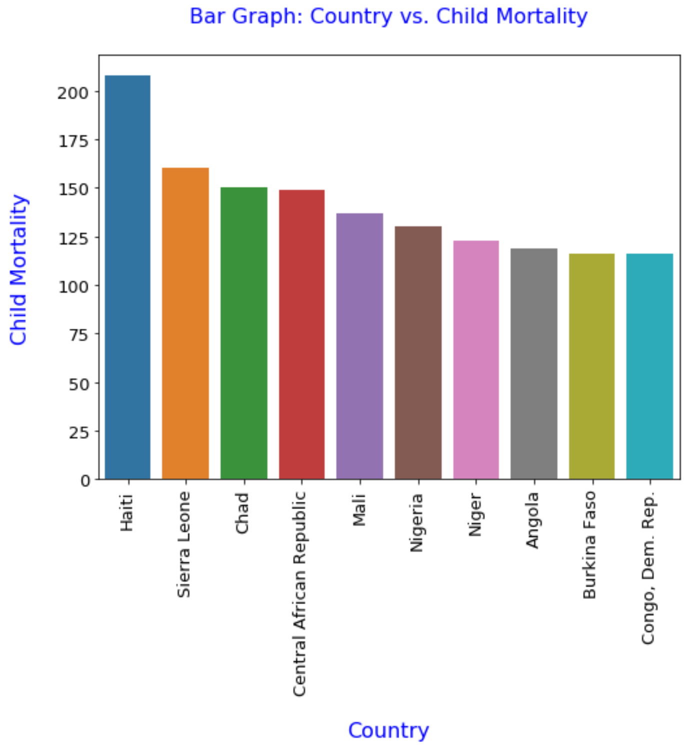






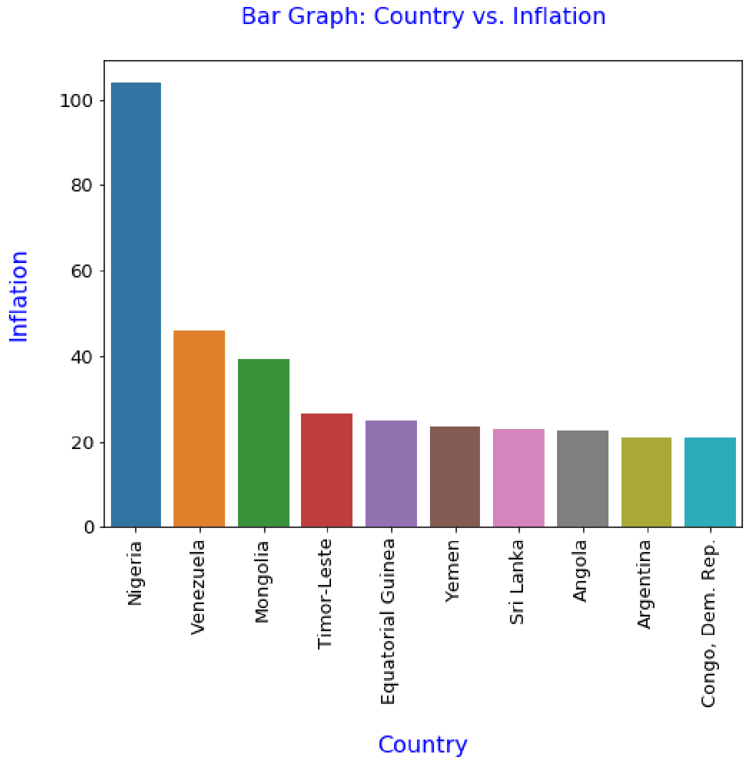


Bivariate Analysis



- From the bar plot(left) between ***Country vs. Child Mortality***, we can see that country Haiti, situated in continent North America has the highest child mortality rate followed by Sierra Leone, Chad, Central African Republic, Mali.

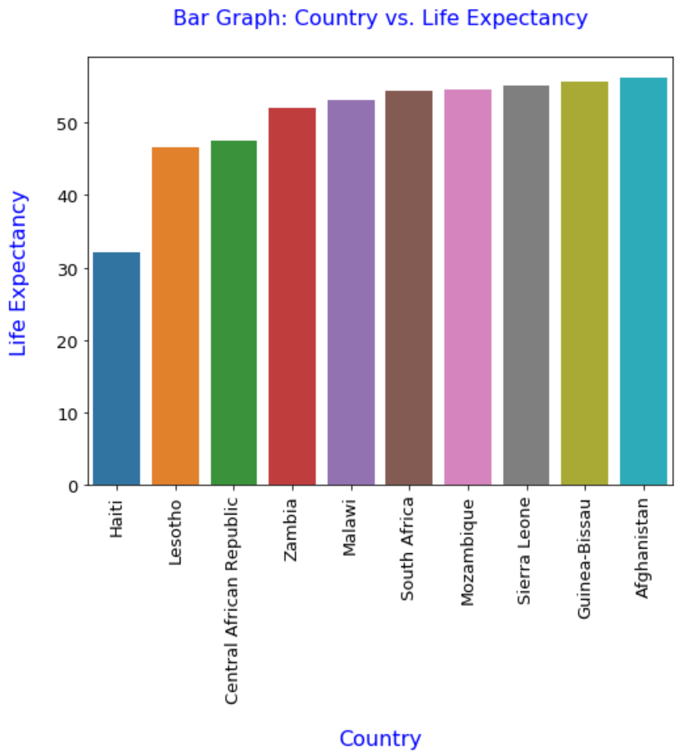
- Hence, based on child mortality rate, we can say that these top 5 countries can be considered for HELP International NGO aid.



- From the bar plot(right) between ***Country vs. Inflation***, we can see that country Nigeria, situated in continent Africa has the highest inflation followed by Venezuela, Mongolia, Timor-Leste, Equatorial Guinea.

- Hence, based on inflation, we can say that these top 5 countries can be considered for HELP International NGO aid.

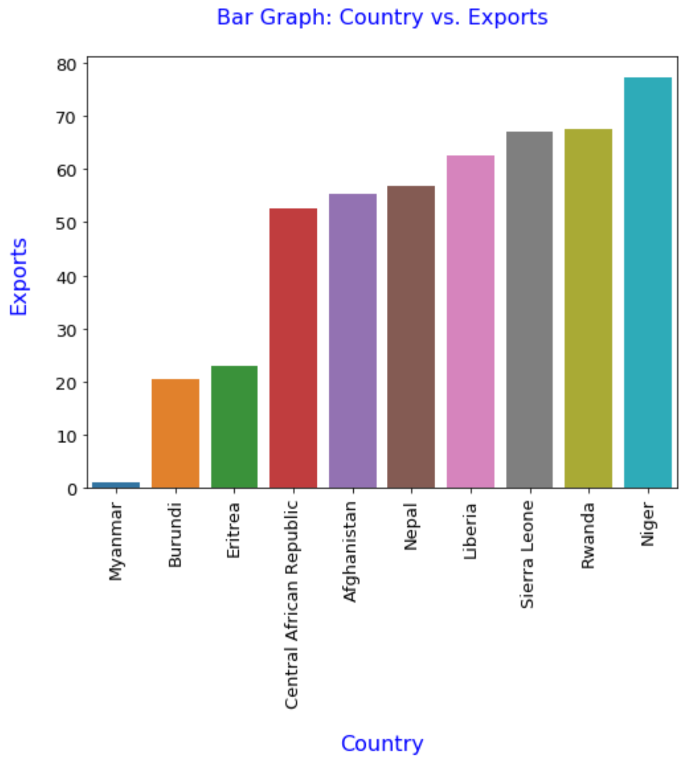
Bivariate Analysis



- From the bar plot(left) between ***Country vs. Life Expectancy***, we can see that country Haiti, situated in continent North America has the lowest life expectancy followed by Lesotho, Central African Republic, Zambia, Malawi.

- From the Country vs. Child Mortality bar plot, we already saw that country Haiti has the highest child mortality rate. So, Haiti can strongly be considered for NGO aid.

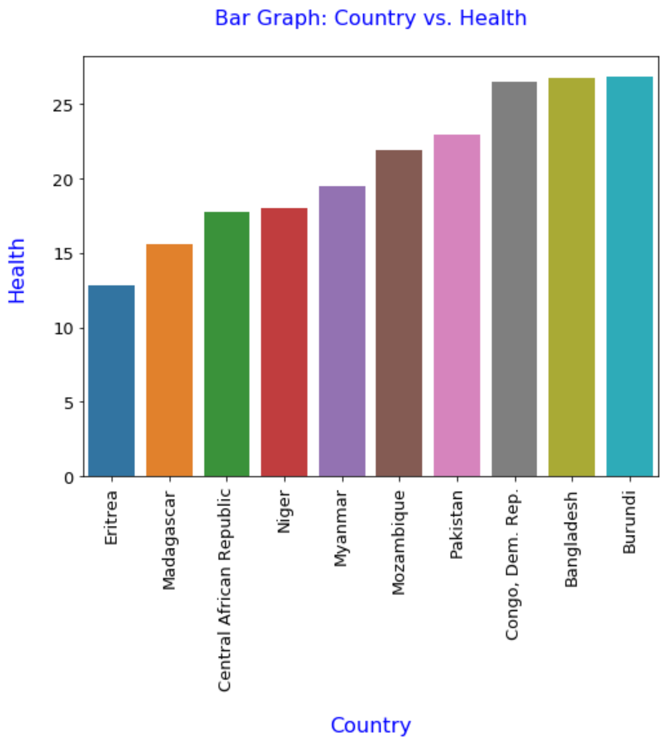
- Hence, based on Life expectancy, we can say that these least 5 countries can be considered for HELP International NGO aid.



- From the bar plot(right) between ***Country vs. Exports***, we can see that country Myanmar, situated in continent Asia has the lowest exports of goods and services followed by Burundi, Eritrea, Central African Republic, Afghanistan.

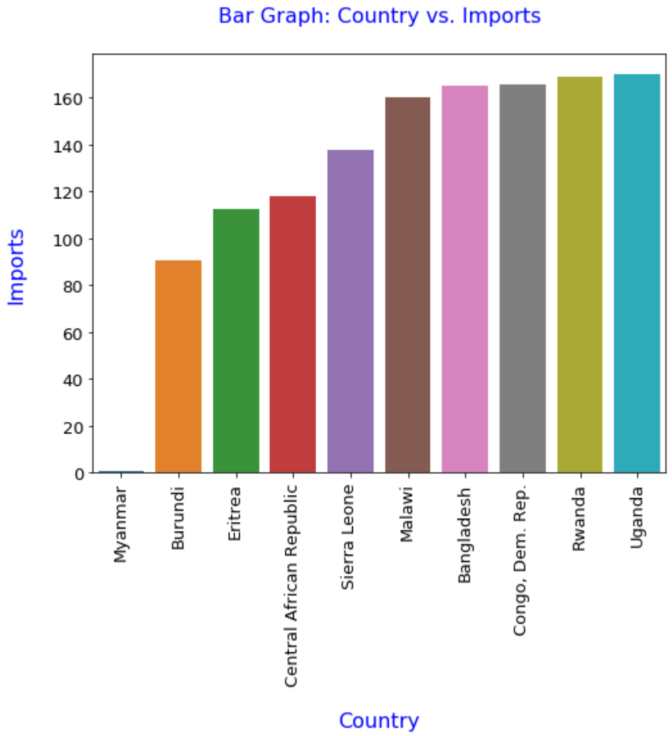
- Hence, based on exports, we can say that these least 5 countries can be considered for HELP International NGO aid.

Bivariate Analysis

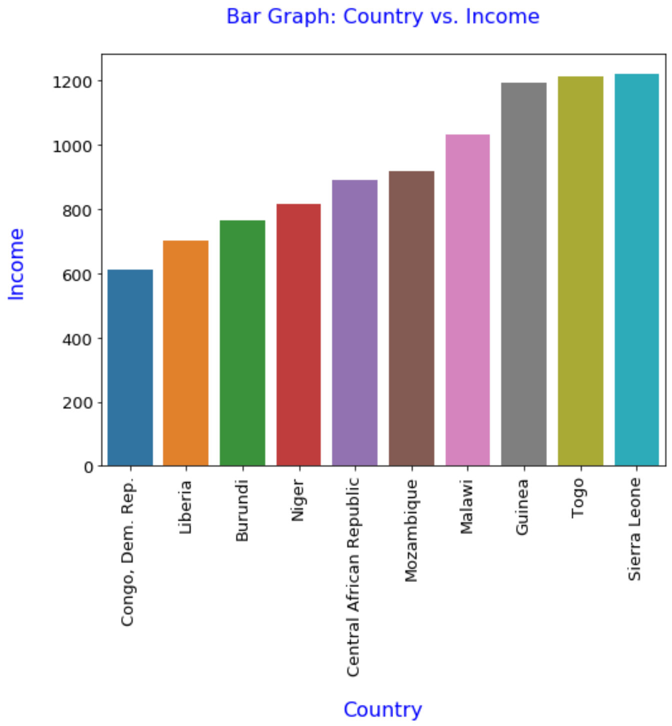


- From the bar plot(left) between ***Country vs. Health***, we can see that country Eritrea, situated in continent Africa spend lowest amount on health followed by Madagascar, Central African Republic, Niger, Myanmar.
- From the Country vs. Exports bar plot, we already saw that country Myanmar has the lowest exports of goods and services. So, Myanmar can strongly be considered for NGO aid.
- Hence, based on Health spend, we can say that these least 5 countries can be considered for HELP International NGO aid.

- From the bar plot(right) between ***Country vs. Imports***, we can see that country Myanmar, situated in continent Asia has lowest imports of goods and services followed by Burundi, Eritrea, Central African Republic, Sierra Leone.
- From the Country vs. Exports and Country vs. Health bar plots, we already saw that country Myanmar has the lowest exports of goods and services as well as it spends low on health. So, Myanmar can strongly be considered for NGO aid.
- Hence, based on Imports spend, we can say that these least 5 countries can be considered for HELP International NGO aid.

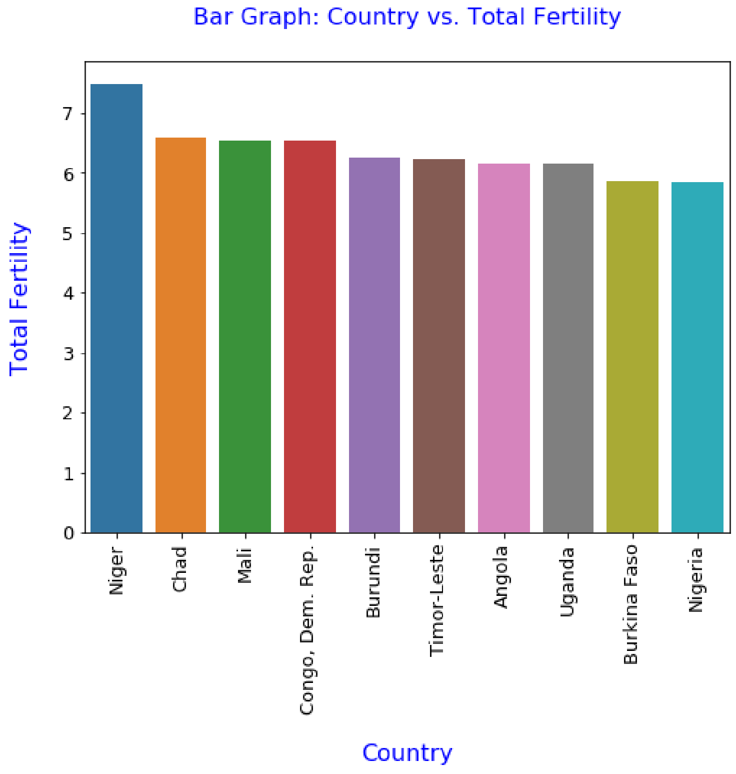


Bivariate Analysis



- From the bar plot(left) between ***Country vs. Income***, we can see that country Congo, Dem. Rep., situated in continent Africa has lowest net income per person followed by Liberia, Burundi, Niger, Central African Republic.

- Hence, based on income per person, we can say that these least 5 countries can be considered for HELP International NGO aid.

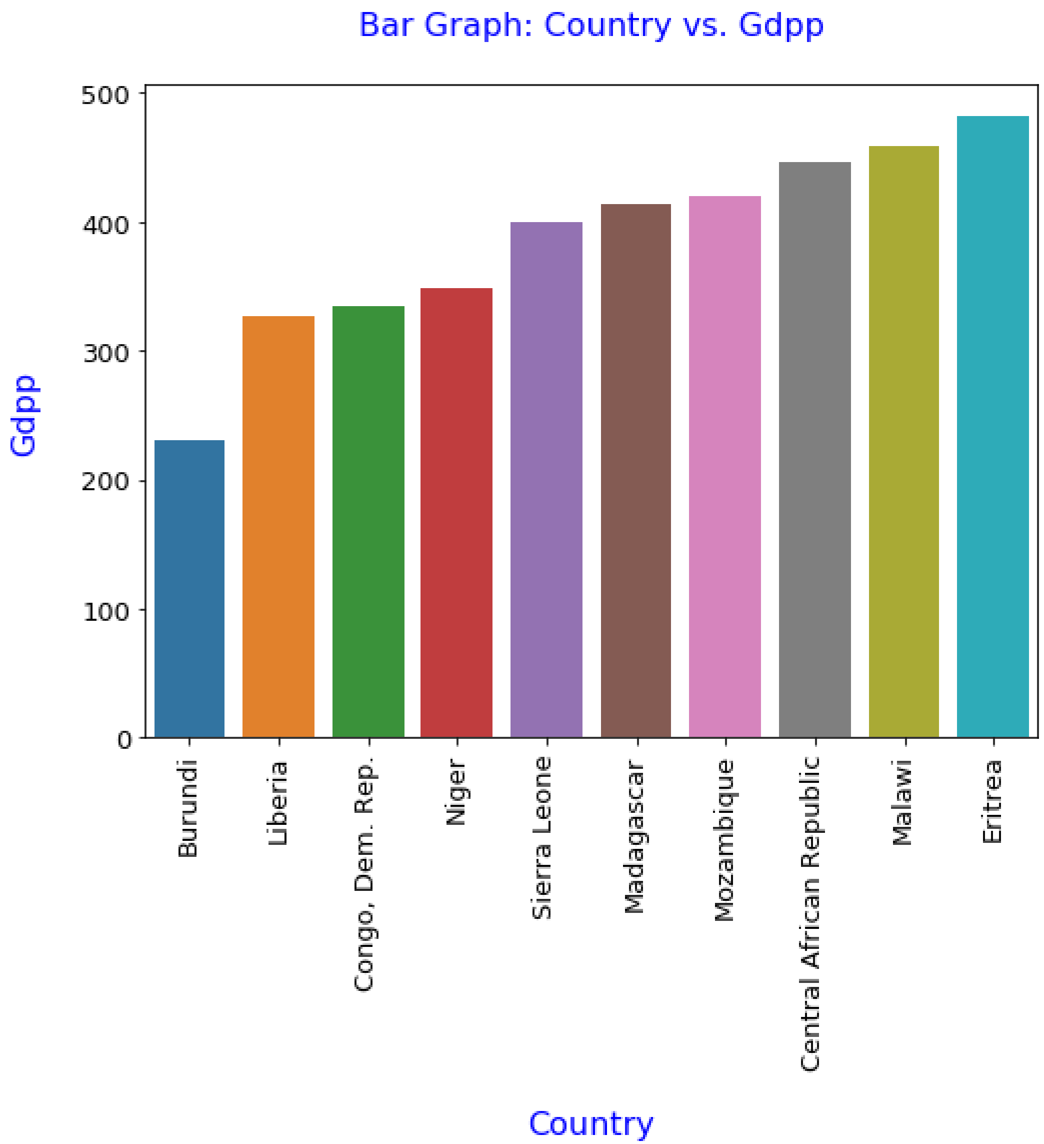


- From the bar plot(right) between ***Country vs. Total Fertility***, we can see that country Niger, situated in continent Africa has highest fertility rate followed by Chad, Mali, Congo, Dem. Rep., Burundi.

- From the Country vs. Income and Country vs. Health bar plots, we already saw that country Niger has the low income per person as well as it spends low on health. So, Niger can strongly be considered for NGO aid.

- Hence, based on total fertility rate, we can say that these top 5 countries can be considered for HELP International NGO aid.

Bivariate Analysis



- From the bar plot between ***Country vs. Gdpp***, we can see that country Burundi, situated in continent Africa has lowest gdpp per capita followed by Liberia, Congo, Dem. Rep., Niger, Sierra Leone.

- From the Country vs. Total Income , Country vs. Imports, Country vs. Exports, Country vs. Total Fertility bar plots, we already saw that country Burundi has the low income per person and it has low imports and exports of goods and services, as well as, having high fertility rate. So, Burundi can strongly be considered for NGO aid.

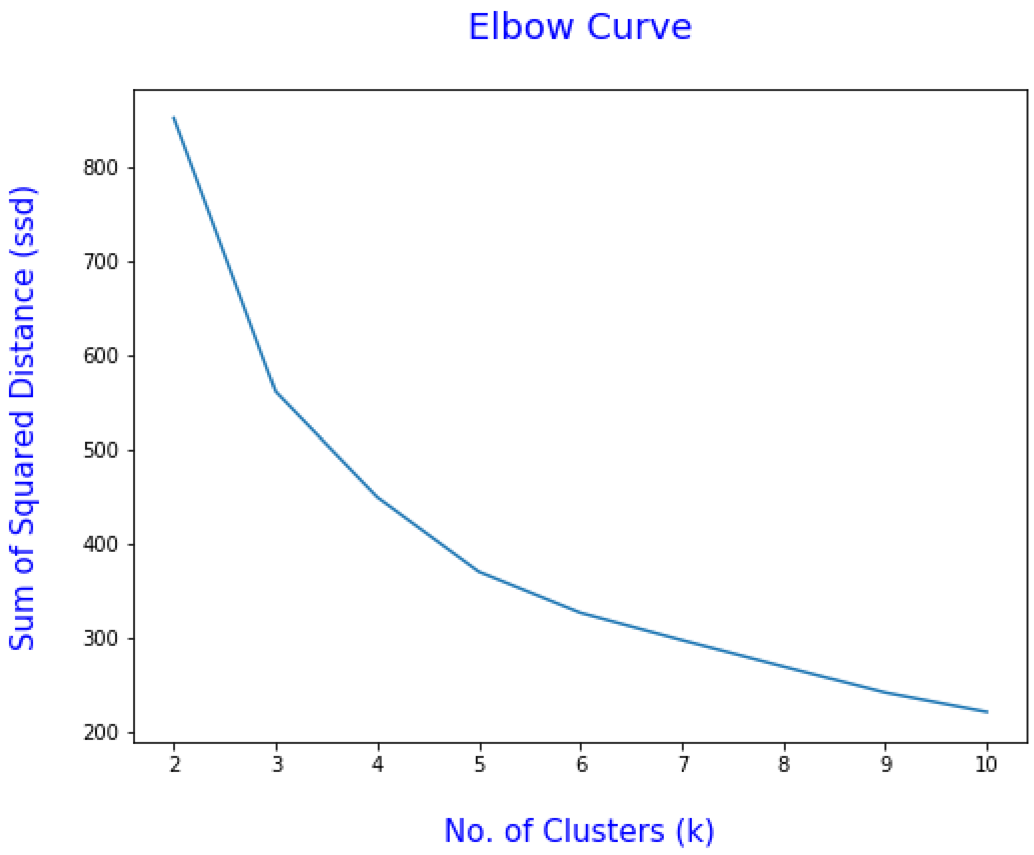
- Hence, based on gdpp per capita, we can say that these least 5 countries can be considered for HELP International NGO aid.

- From the above Bivariate analysis, we can conclude that, African countries can be considered mostly for NGO aid.

K-Mean Analysis :
Choosing Value of K using *Elbow Curve*

- Calculate the Within-Cluster-Sum of Squared Errors (ssd) for different values of k and choose the k for which ssd becomes first starts to diminish. So, the point where this distortion declines the most is the elbow point.

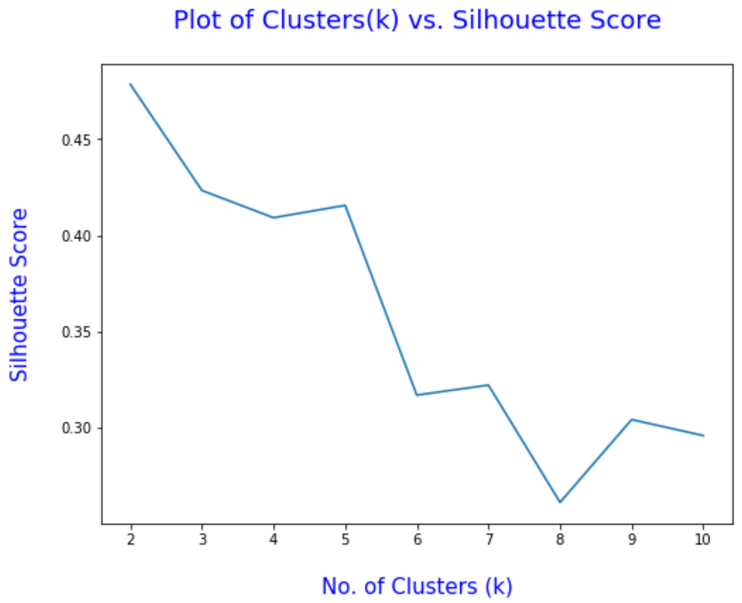
- From the above curve the elbow point will be at k = 3, because at that point, distortion declines the most.

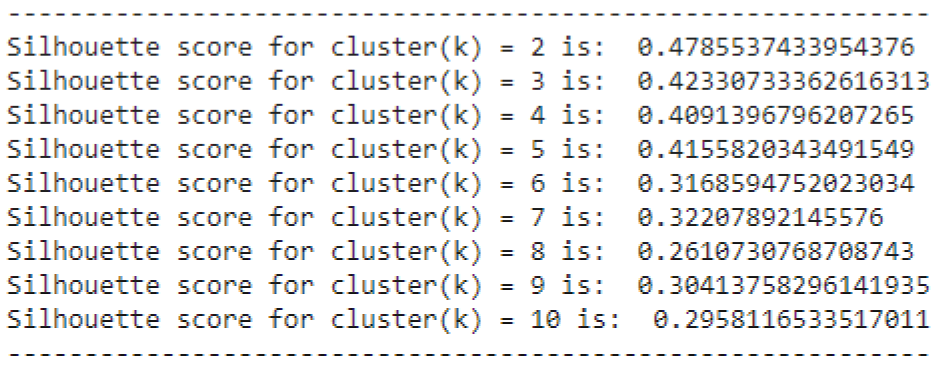


K-Mean Analysis :
Choosing Value of K using *Silhouette score*

- We know that, Silhouette score closer to 1 indicates that the data point is very similar to other data points in the cluster. In our plot, the Silhouette score value is closer to 1, for k = 2. But for business, k = 2 is not a good number of clusters to work on.

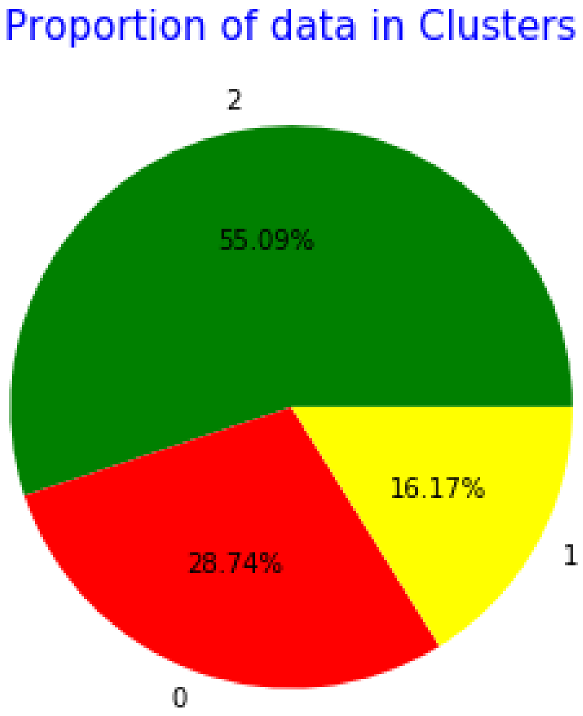
- Hence, from Silhouette score curve, we will consider k = 3 as the optimum cluster number.

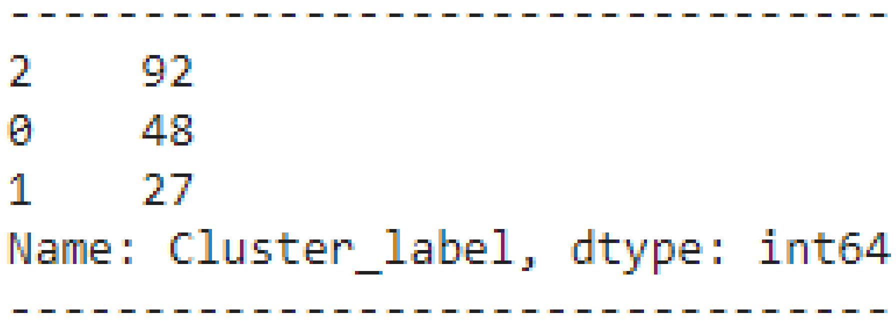




```
------------------------------------------------------------
Silhouette score for cluster(k) = 2 is:  0.4785537433954376
Silhouette score for cluster(k) = 3 is:  0.4233073362616313
Silhouette score for cluster(k) = 4 is:  0.409139679620726 5
Silhouette score for cluster(k) = 5 is:  0.4155820343491549
Silhouette score for cluster(k) = 6 is:  0.3168594752023034
Silhouette score for cluster(k) = 7 is:  0.32207892145576
Silhouette score for cluster(k) = 8 is:  0.2610730768708743
Silhouette score for cluster(k) = 9 is:  0.30413758296141935
Silhouette score for cluster(k) = 10 is:  0.2958116533517011
------------------------------------------------------------
```

K-Mean Analysis :
Data Distribution after *K-Mean Clustering*



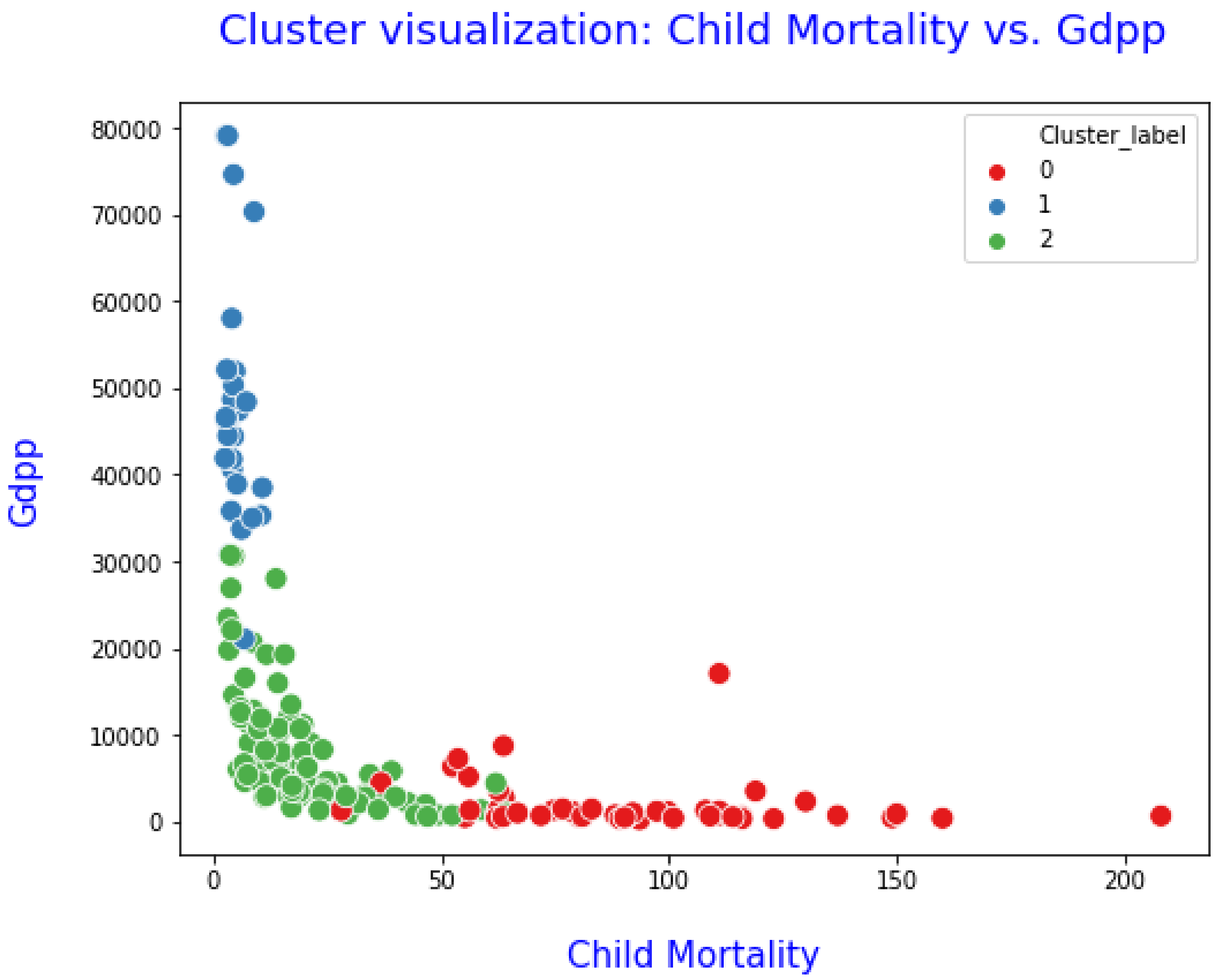


```
-----------------------------------------
2    92
0    48
1    27
Name: Cluster_label, dtype: int64
-----------------------------------------
```

- Data distribution seems goods among three clusters. Hence, we can proceed with visualization and Cluster profiling

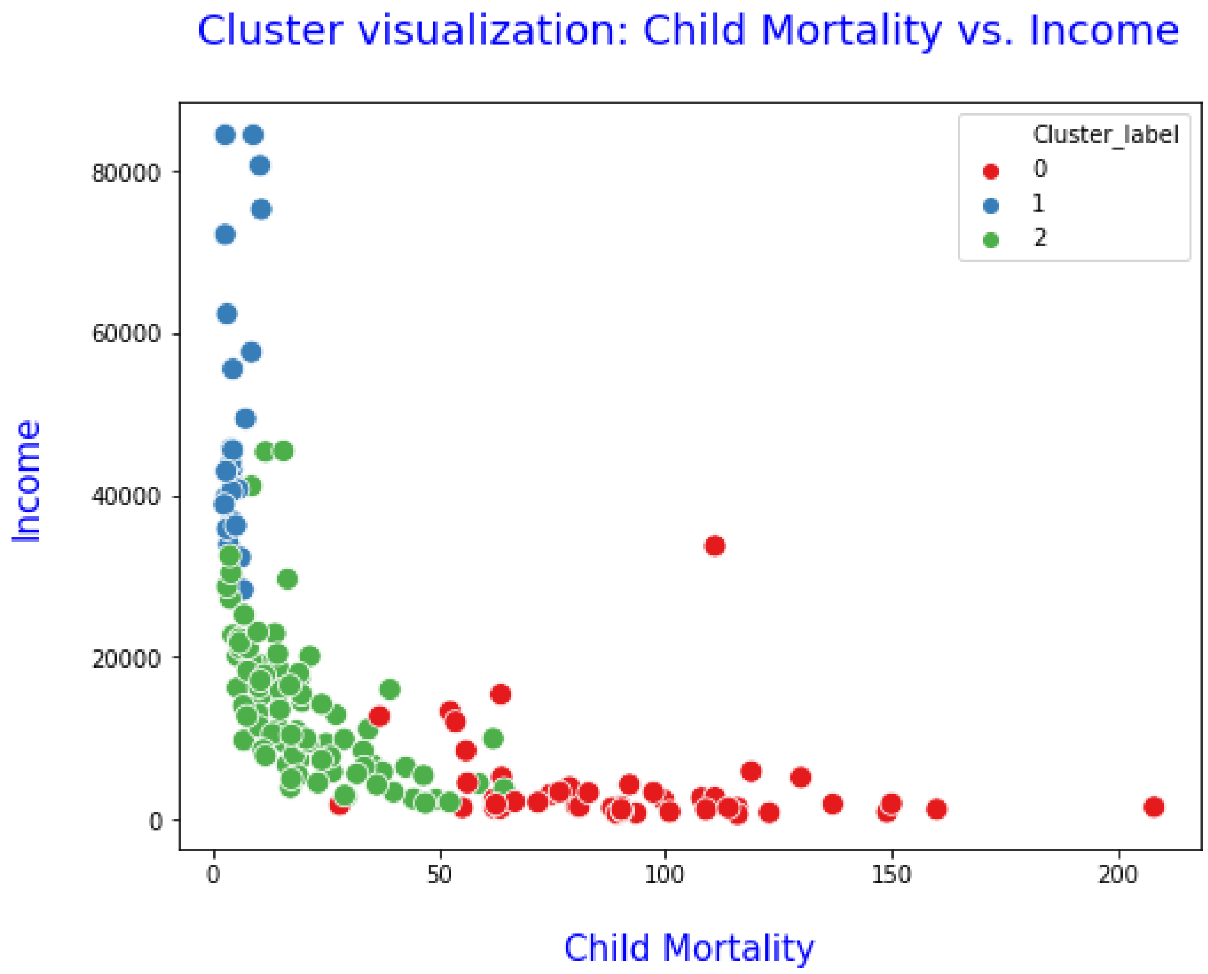
K-Mean Analysis : Cluster Visualization

Child Mortality vs. Gdpp



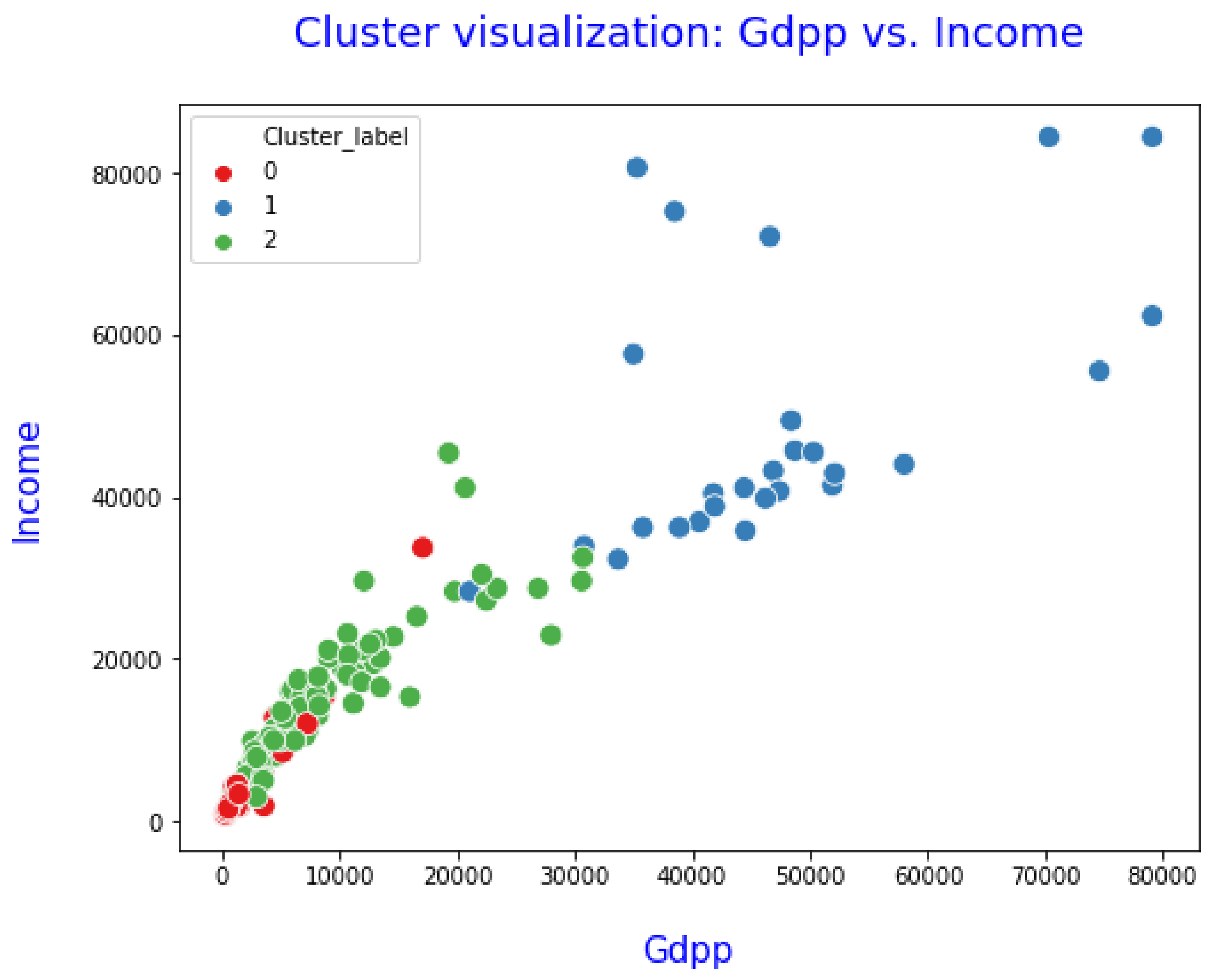
K-Mean Analysis : Cluster Visualization

Child Mortality vs. Income



K-Mean Analysis : Cluster Visualization

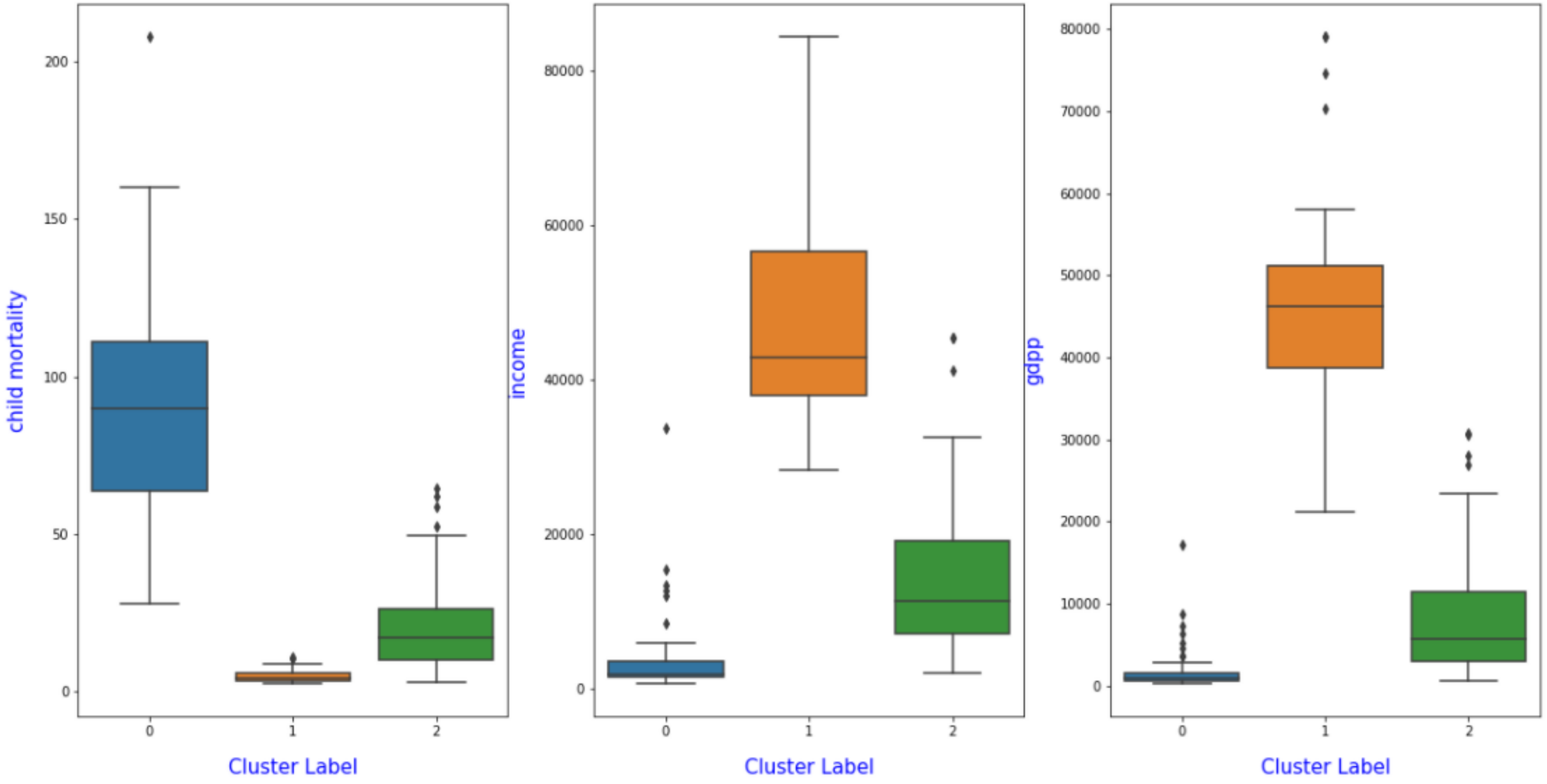
Gdpp vs. Income



K-Mean Analysis : Cluster Profiling

Box Plot for displaying data segmentation among cluster 0, 1, 2 based on below features:

- ***Child Mortality***
- ***Income***
- ***Gdpp***

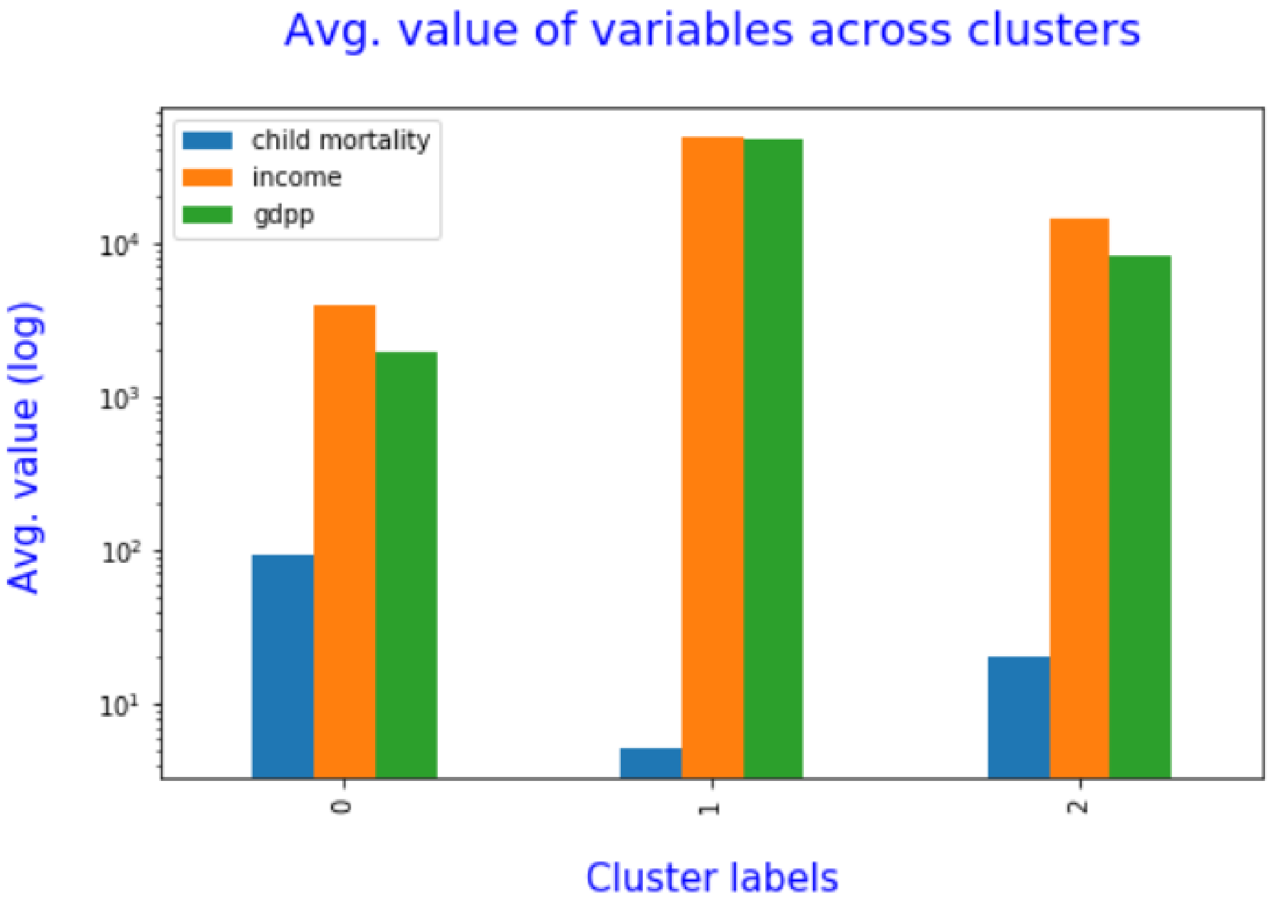


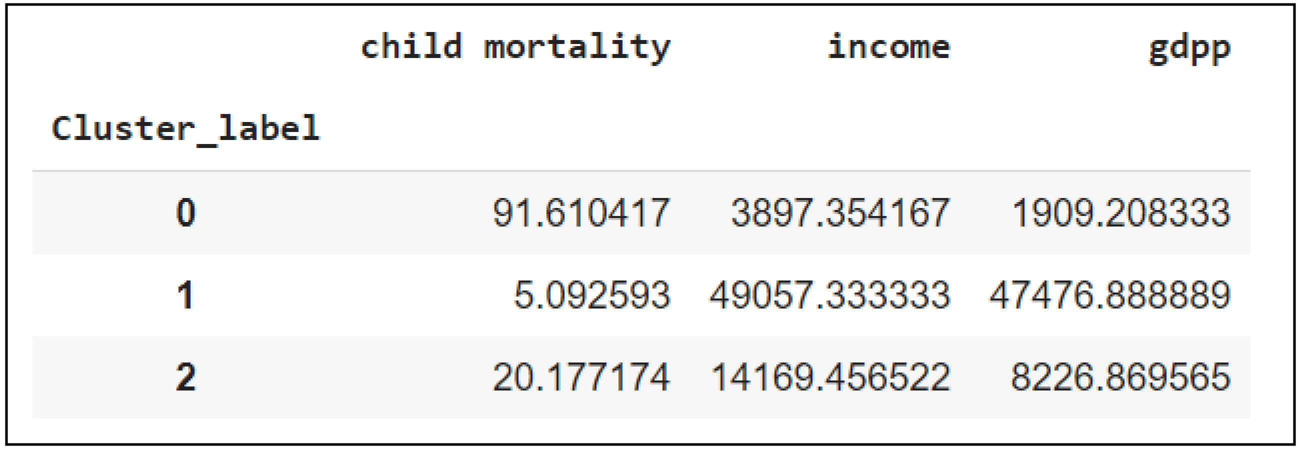
K-Mean Analysis : Cluster Profiling

- From ***Bar plot***, we can see that **Cluster 0** is having Low GDPP, Low Income and High Child Mortality rate. Hence, this segment can be clustered as **Under developed countries and need financial aids from NGO**.

- We can also see that **Cluster 2** is having better GDPP, Income and Child Mortality rate than cluster 0, but worst than cluster 1. Hence, we can clustered this segment of countries as **Developing Countries**.

- As **Cluster 1** is having High GDPP, High Income and Low Child Mortality rate. Hence, this segment can be clustered as **Developed countries**.



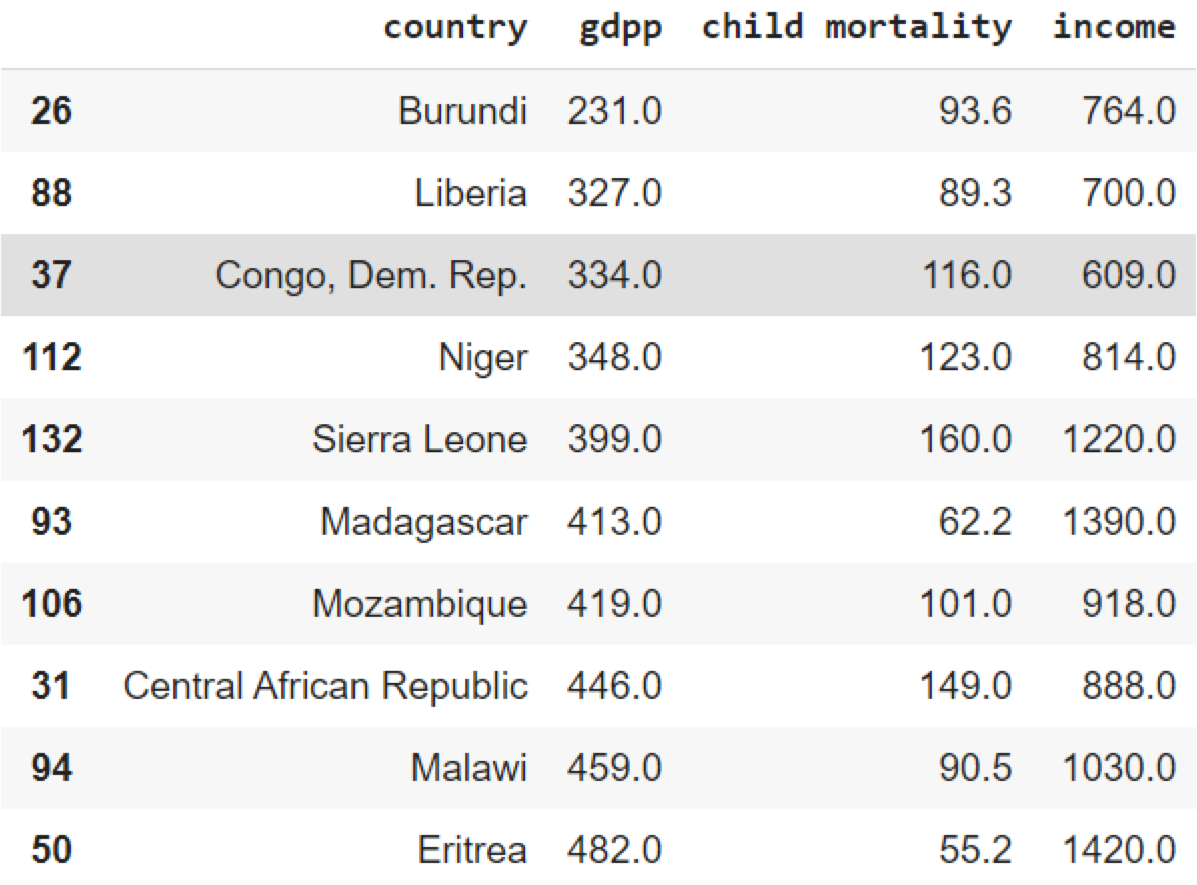


	child mortality	income	gdpp
Cluster_label			
0	91.610417	3897.354167	1909.208333
1	5.092593	49057.333333	47476.888889
2	20.177174	14169.456522	8226.869565

K-Mean Analysis :
Country Identified for NGO AID

- **Top 10 Country** identified using K-Means clustering based on following features:

  - ***Child Mortality***
  - ***Income***
  - ***Gdpp***



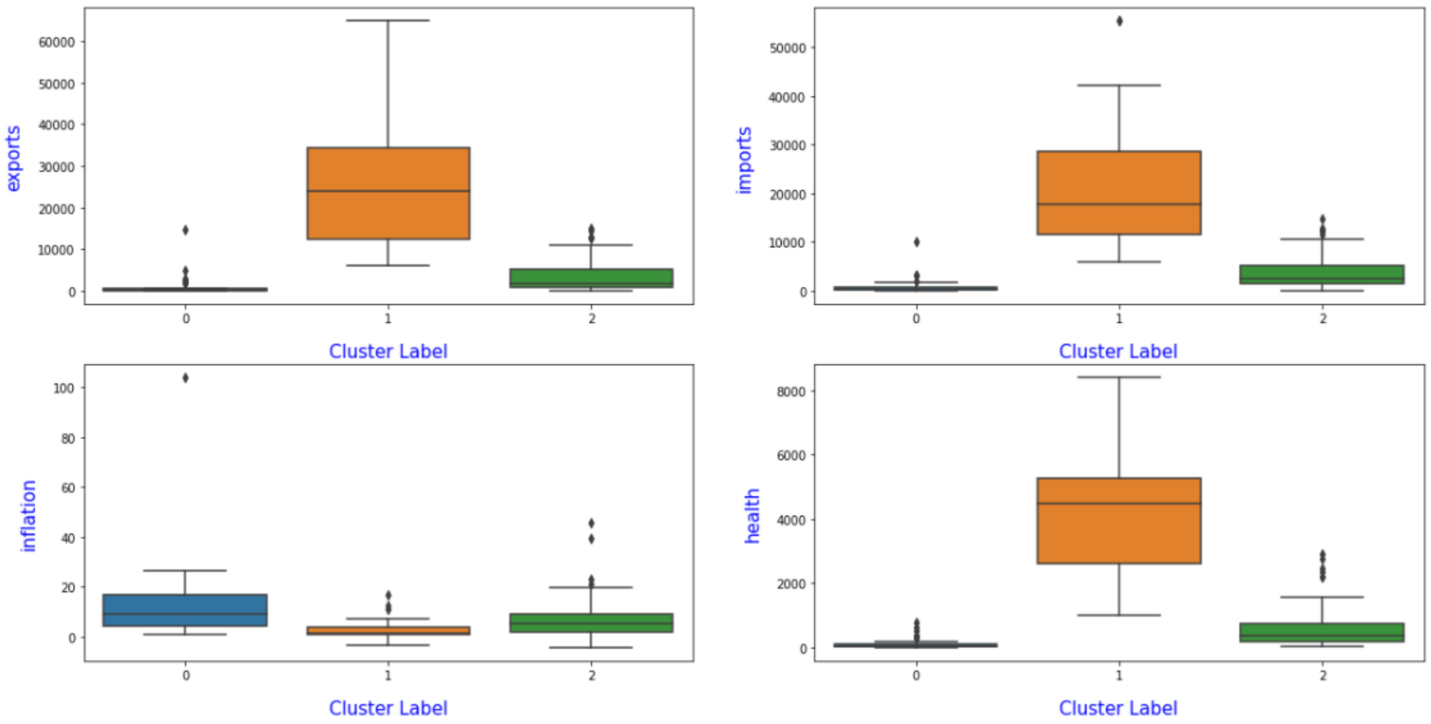
	country	gdpp	child mortality	income
26	Burundi	231.0	93.6	764.0
88	Liberia	327.0	89.3	700.0
37	Congo, Dem. Rep.	334.0	116.0	609.0
112	Niger	348.0	123.0	814.0
132	Sierra Leone	399.0	160.0	1220.0
93	Madagascar	413.0	62.2	1390.0
106	Mozambique	419.0	101.0	918.0
31	Central African Republic	446.0	149.0	888.0
94	Malawi	459.0	90.5	1030.0
50	Eritrea	482.0	55.2	1420.0

K-Mean Analysis : Cluster Profiling

(Other Socio-economic factors)

Box Plot for displaying data segmentation among cluster 0, 1, 2 based on **other socio-economic factors**:

- ***Health***
- ***Imports***
- ***Exports***
- ***Inflation***



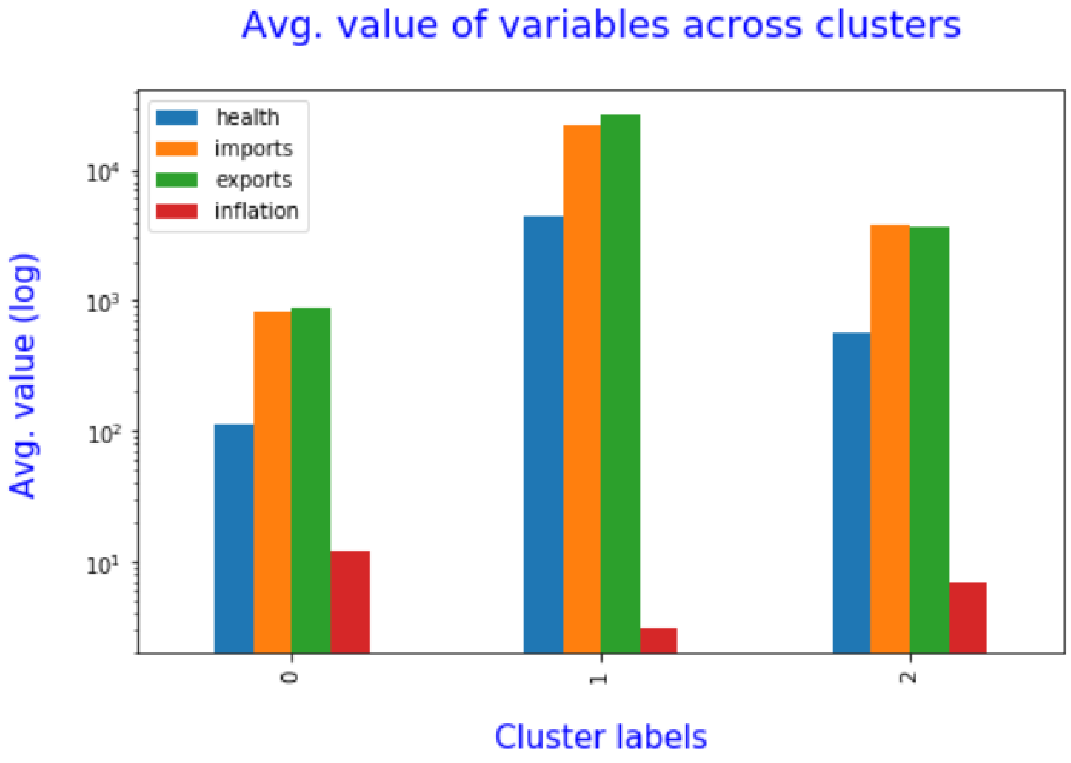
K-Mean Analysis : Cluster Profiling

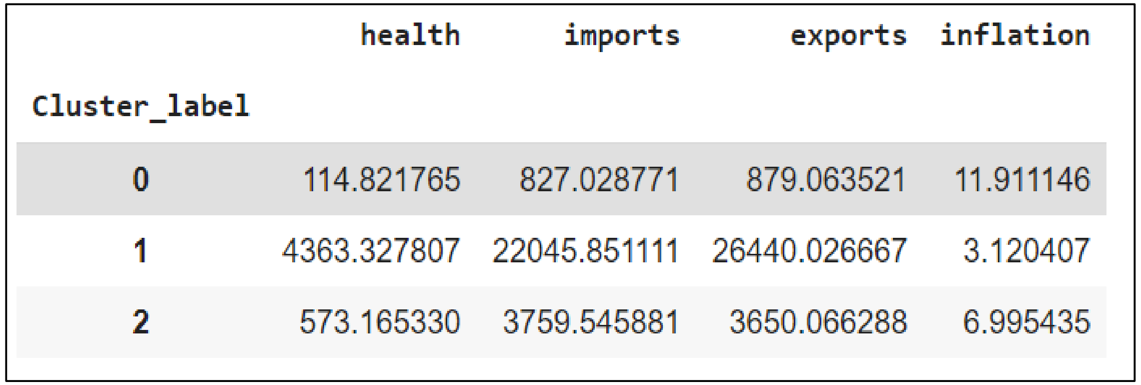
(Other Socio-economic factors)

- From above plot, we can see that **Cluster 0** is having Low Health spent, Low Imports of goods and services, Low Exports of goods and services and High Inflation. Hence, this segment can be clustered as **Under developed countries and need financial aids from NGO.**

- We can also see that **Cluster 2** is having better Health spent, Imports of goods and services, Exports of goods and services and Inflation than cluster 0, but worst than cluster 1. Hence, we can clustered this segment of countries as **Developing Countries**.

- As **Cluster 1** is having High Health spent, High Imports and Exports of goods and services and Low Inflation. Hence, this segment can be clustered as **Developed countries.**



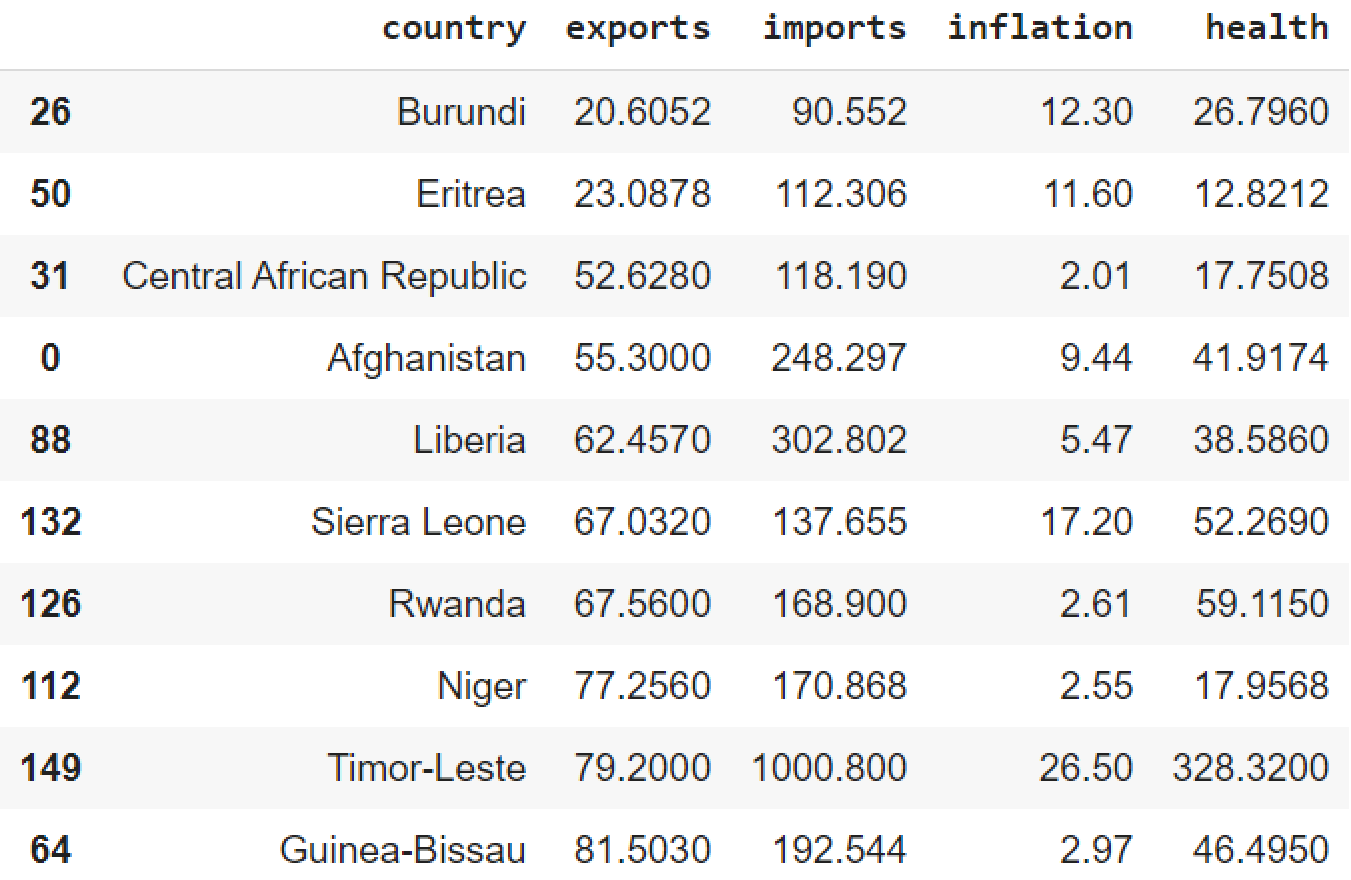


	health	imports	exports	inflation
Cluster_label				
0	114.821765	827.028771	879.063521	11.911146
1	4363.327807	22045.851111	26440.026667	3.120407
2	573.165330	3759.545881	3650.066288	6.995435

K-Mean Analysis :
Country Identified for NGO AID

(Other Socio-economic factors)

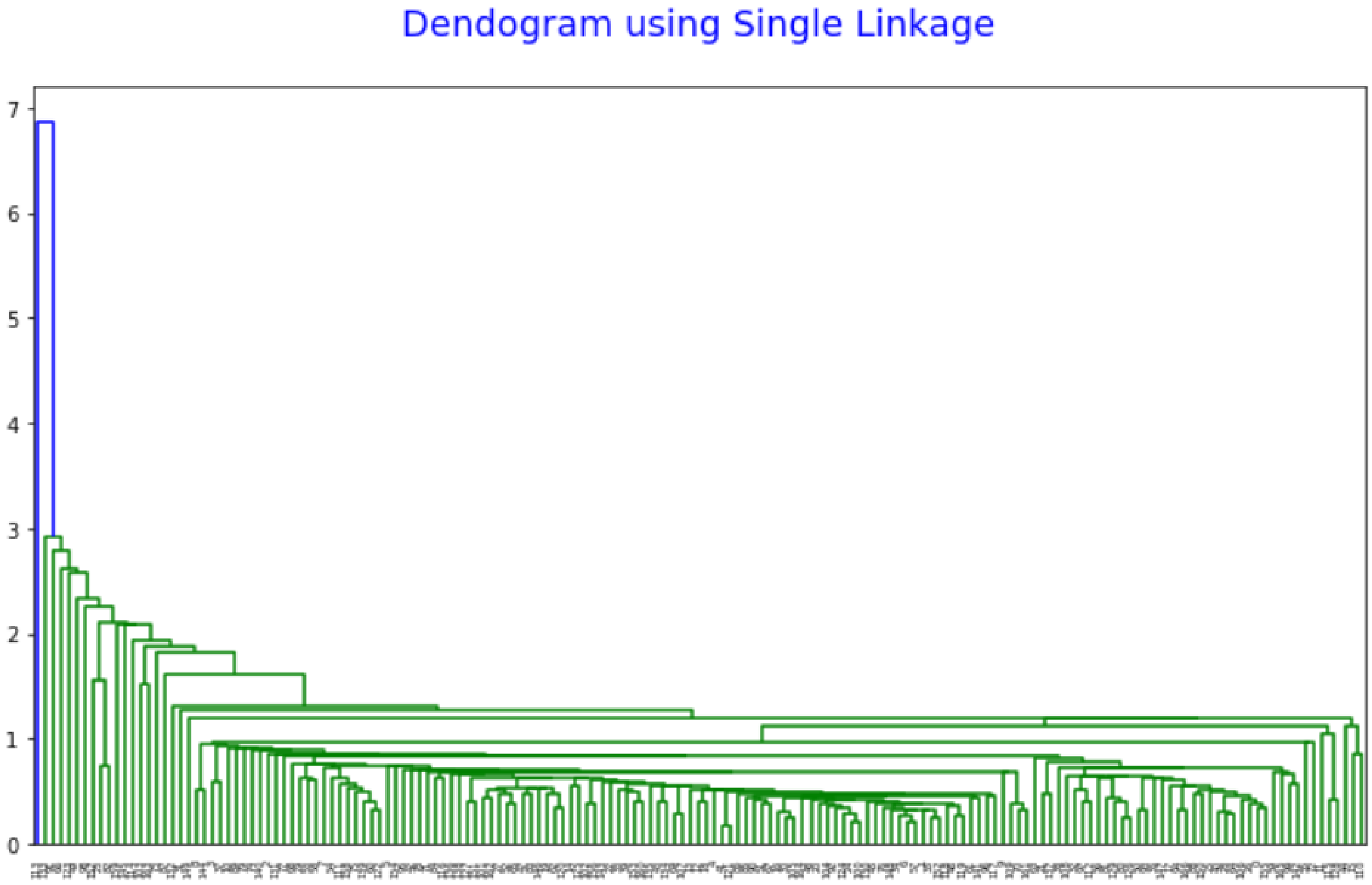
- **Top 10 Country** identified using K-Means clustering based on following **other socio-economic factors**:
  - ***Health***
  - ***Imports***
  - ***Exports***
  - ***Inflation***



	country	exports	imports	inflation	health
26	Burundi	20.6052	90.552	12.30	26.7960
50	Eritrea	23.0878	112.306	11.60	12.8212
31	Central African Republic	52.6280	118.190	2.01	17.7508
0	Afghanistan	55.3000	248.297	9.44	41.9174
88	Liberia	62.4570	302.802	5.47	38.5860
132	Sierra Leone	67.0320	137.655	17.20	52.2690
126	Rwanda	67.5600	168.900	2.61	59.1150
112	Niger	77.2560	170.868	2.55	17.9568
149	Timor-Leste	79.2000	1000.800	26.50	328.3200
64	Guinea-Bissau	81.5030	192.544	2.97	46.4950

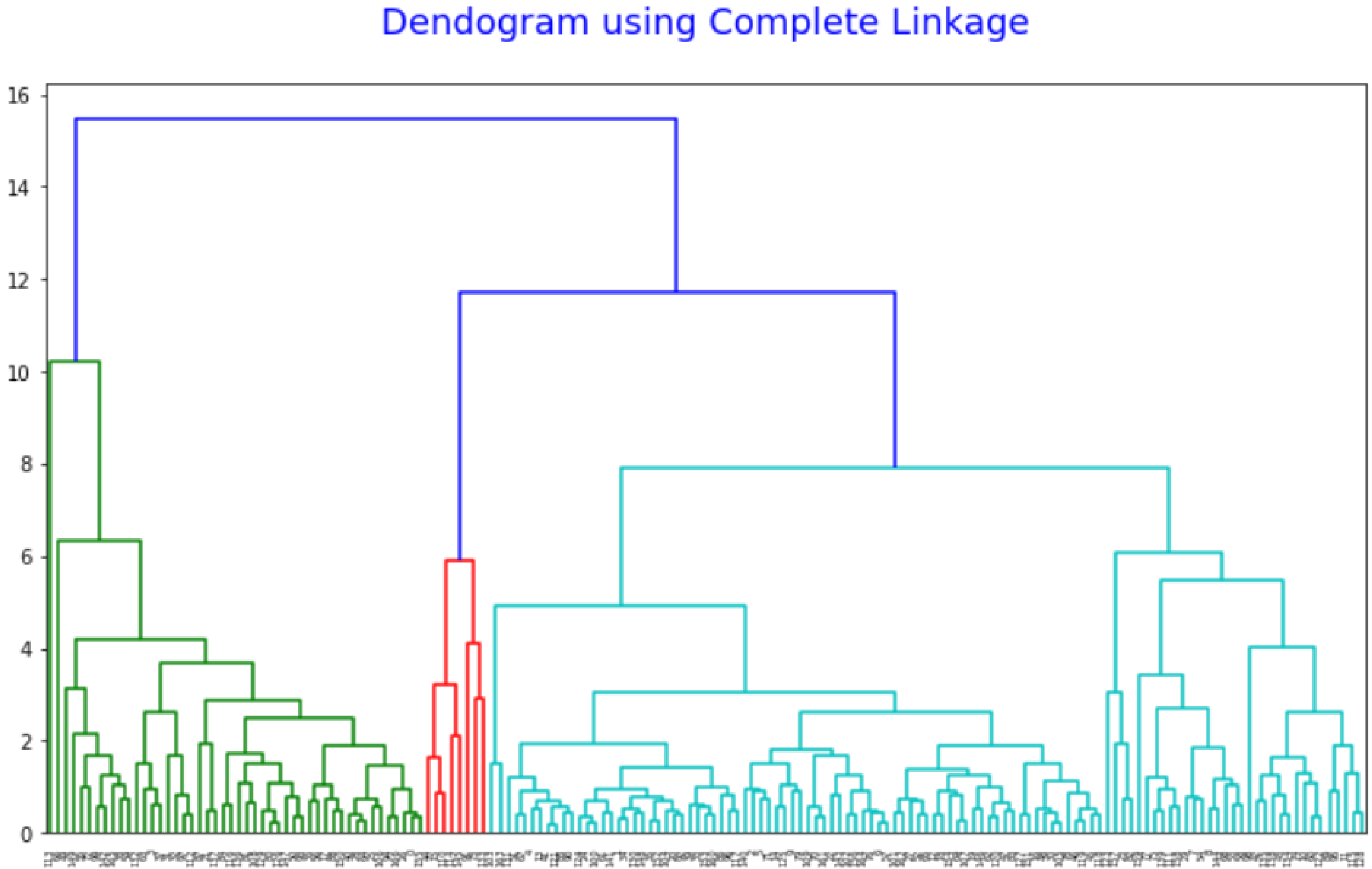
Hierarchical Clustering:

Dendrogram Using
Single Linkage



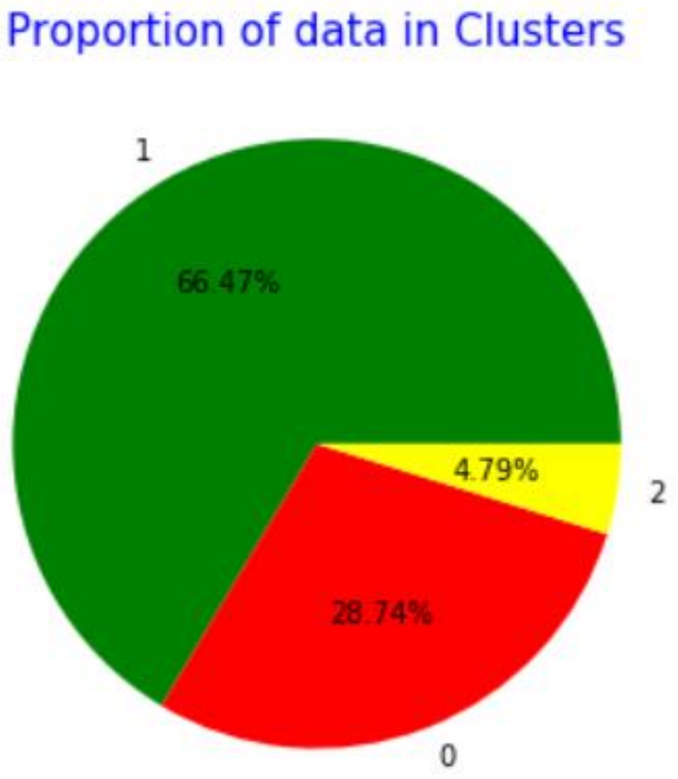
Hierarchical Clustering:

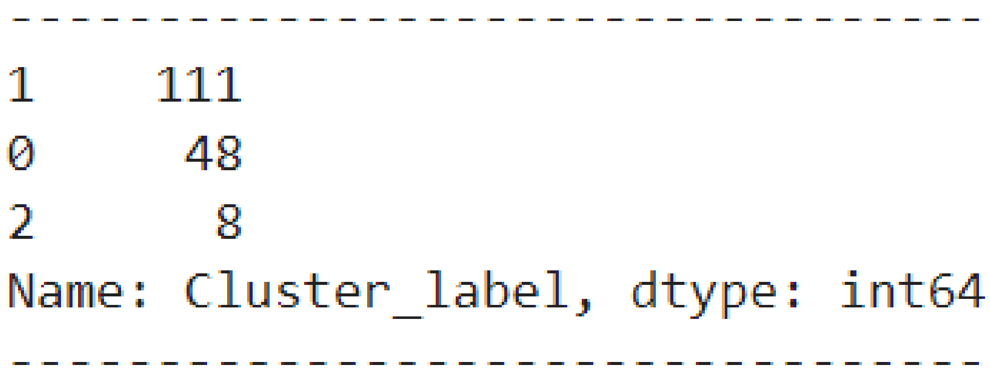
Dendrogram Using
Complete Linkage



- From the above Dendrograms, it is evident that ***Complete Linkage*** gives a better cluster formation. So we will use Complete linkage output for our further analysis.

Hierarchical Clustering:
Data Distribution after *Clustering*


Proportion of data in Clusters

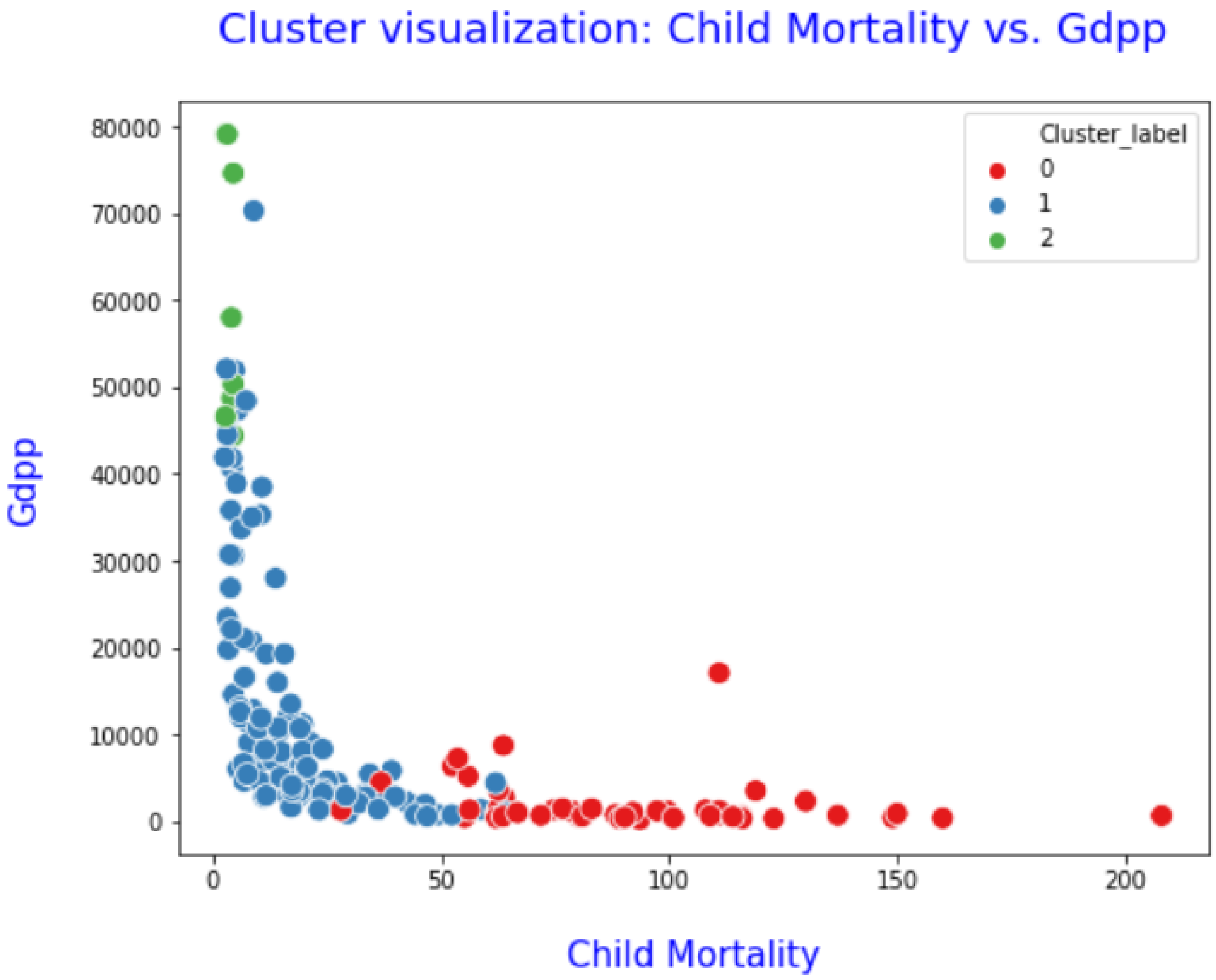


```
-----------------------------------
1    111
0     48
2      8
Name: Cluster_label, dtype: int64
-----------------------------------
```

- Data distribution seems goods among three clusters. Hence, we can proceed with visualization and Cluster profiling

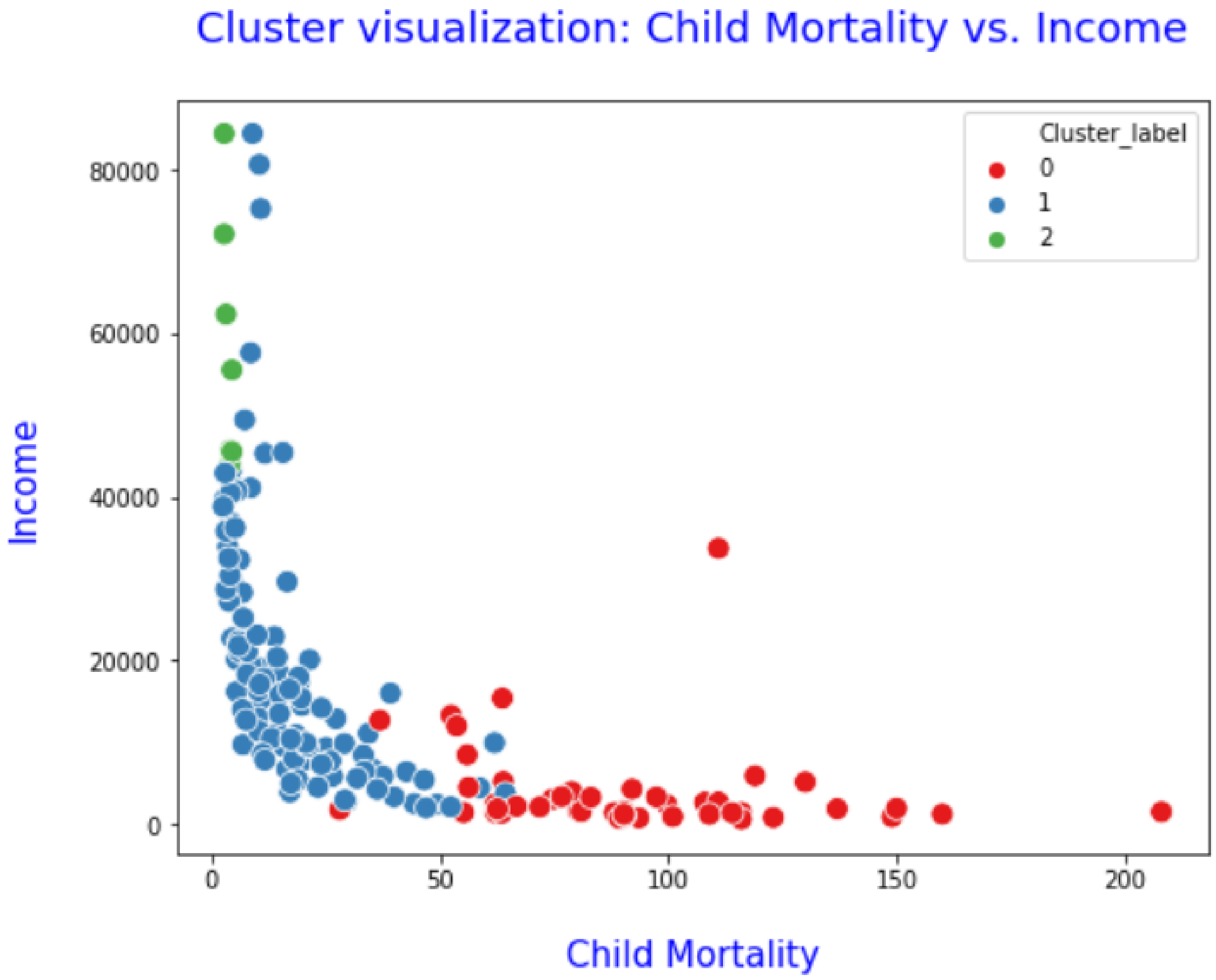
Hierarchical Clustering: Cluster Visualization

Child Mortality vs. Gdpp



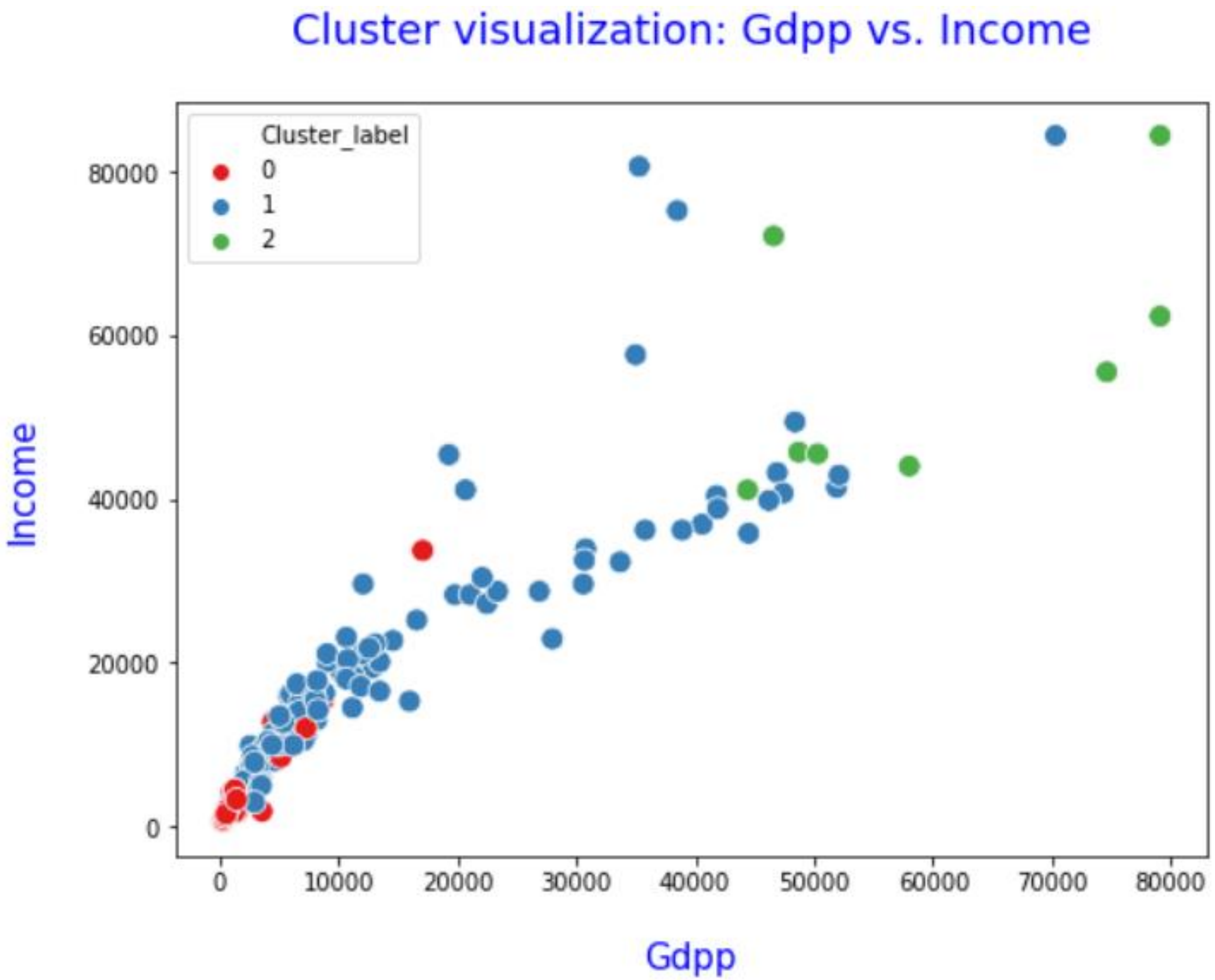
Hierarchical Clustering: Cluster Visualization

Child Mortality vs. Income



Hierarchical Clustering : Cluster Visualization

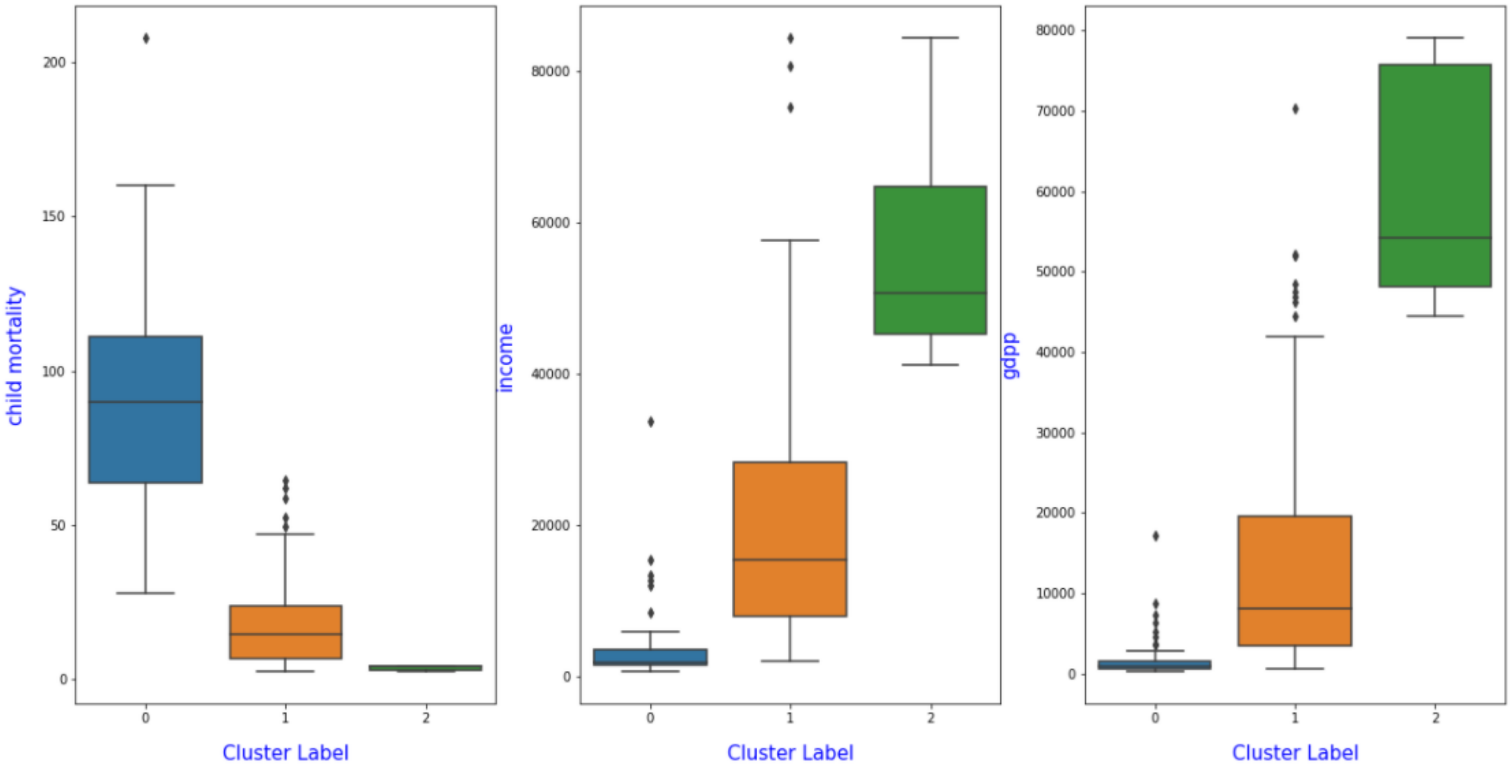
Gdpp vs. Income



Hierarchical Clustering: Cluster Profiling

Box Plot for displaying data segmentation among cluster 0, 1, 2 based on below features:

- ***Child Mortality***
- ***Income***
- ***Gdpp***

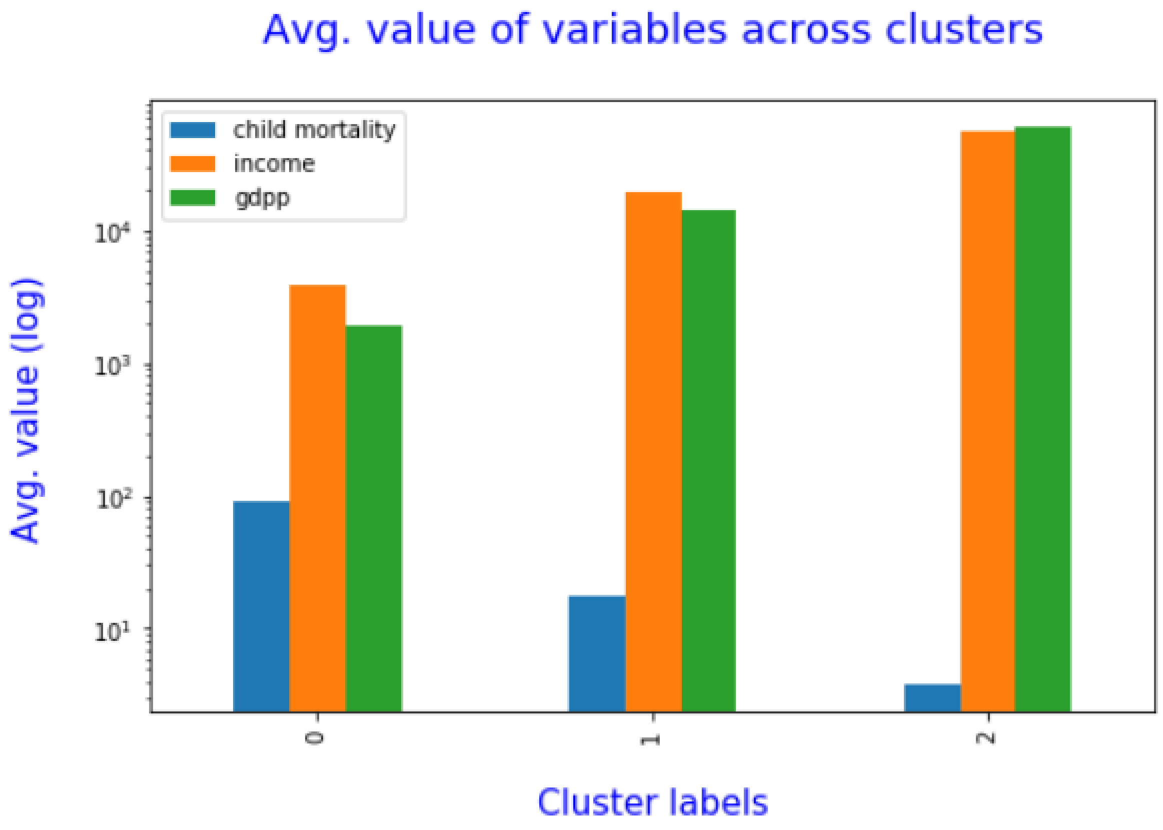


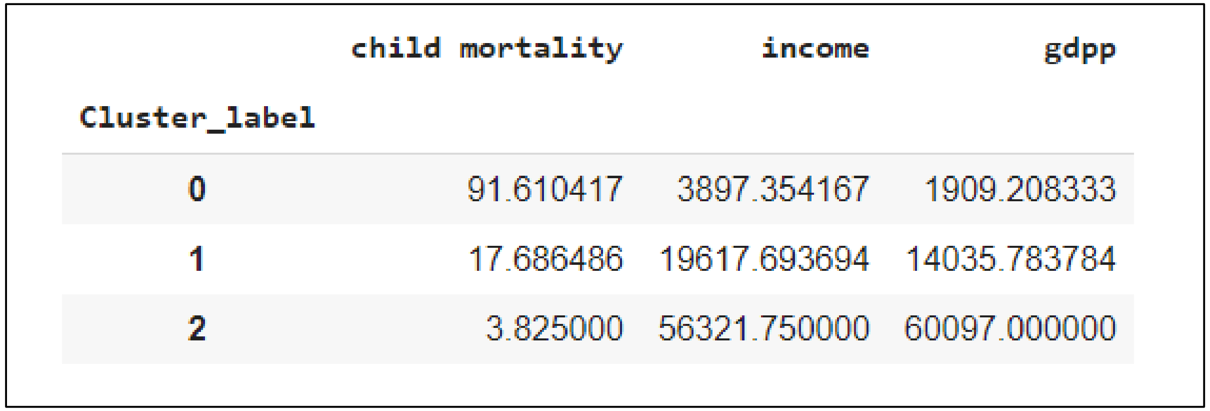
Hierarchical Clustering: Cluster Profiling

- From ***Bar plot***, we can see that **Cluster 0** is having Low GDPP, Low Income and High Child Mortality rate. Hence, this segment can be clustered as **Under developed countries and need financial aids from NGO**.

- We can also see that **Cluster 2** is having better GDPP, Income and Child Mortality rate than cluster 0, but worst than cluster 1. Hence, we can clustered this segment of countries as **Developing Countries**.

- As **Cluster 1** is having High GDPP, High Income and Low Child Mortality rate. Hence, this segment can be clustered as **Developed countries**.

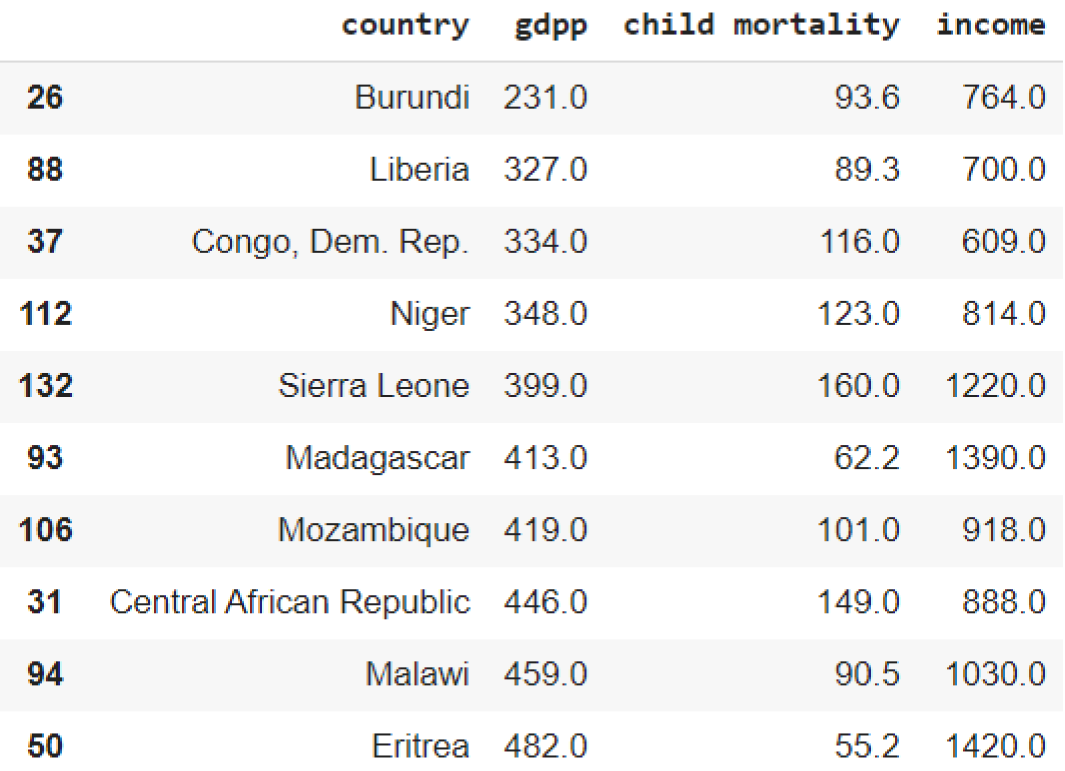




	child mortality	income	gdpp
Cluster_label			
0	91.610417	3897.354167	1909.208333
1	17.686486	19617.693694	14035.783784
2	3.825000	56321.750000	60097.000000

Hierarchical Clustering:
Country Identified for NGO AID

- **Top 10 Country** identified using Hierarchical clustering based on following features:
  - ***Child Mortality***
  - ***Income***
  - ***Gdpp***



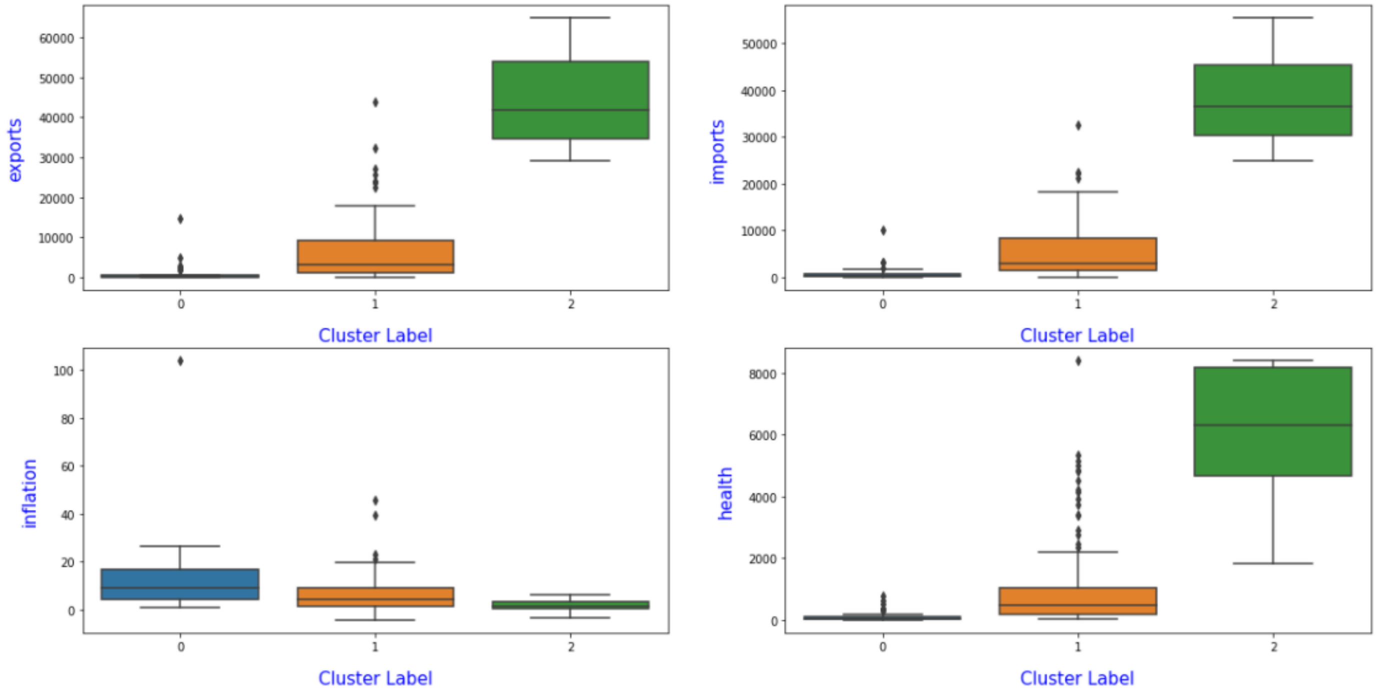
	country	gdpp	child mortality	income
26	Burundi	231.0	93.6	764.0
88	Liberia	327.0	89.3	700.0
37	Congo, Dem. Rep.	334.0	116.0	609.0
112	Niger	348.0	123.0	814.0
132	Sierra Leone	399.0	160.0	1220.0
93	Madagascar	413.0	62.2	1390.0
106	Mozambique	419.0	101.0	918.0
31	Central African Republic	446.0	149.0	888.0
94	Malawi	459.0	90.5	1030.0
50	Eritrea	482.0	55.2	1420.0

Hierarchical Clustering: Cluster Profiling

(Other Socio-economic factors)

Box Plot for displaying data segmentation among cluster 0, 1, 2 based on **other socio-economic factors**:

- ***Health***
- ***Imports***
- ***Exports***
- ***Inflation***



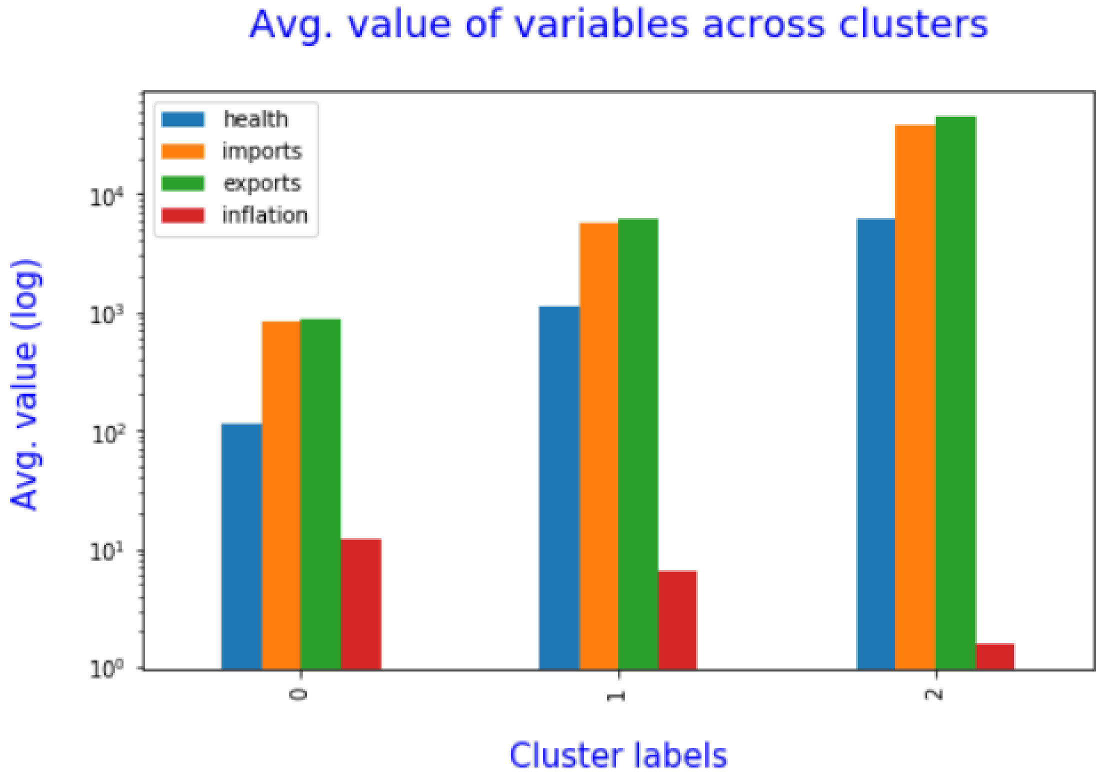
Hierarchical Clustering: Cluster Profiling

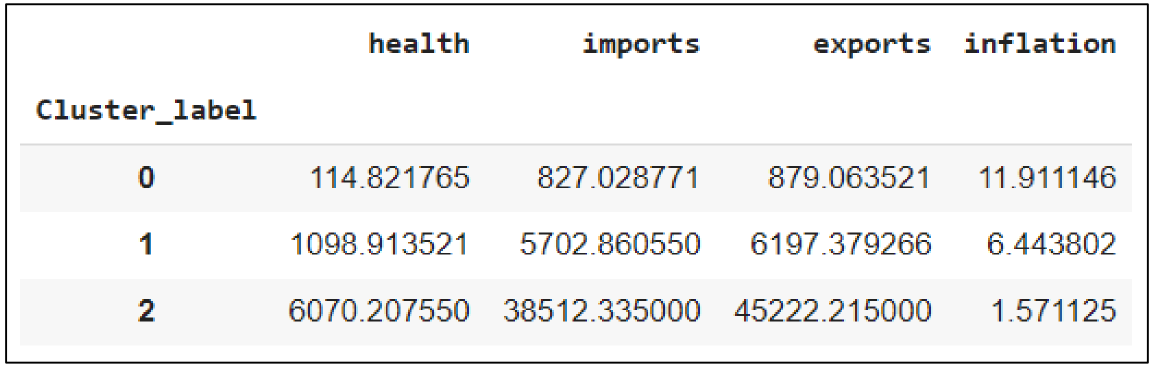
(Other Socio-economic factors)

- From above plot, we can see that **Cluster 0** is having Low Health spent, Low Imports of goods and services, Low Exports of goods and services and High Inflation. Hence, this segment can be clustered as **Under developed countries and need financial aids from NGO.**

- We can also see that **Cluster 2** is having better Health spent, Imports of goods and services, Exports of goods and services and Inflation than cluster 0, but worst than cluster 1. Hence, we can clustered this segment of countries as **Developing Countries**.

- As **Cluster 1** is having High Health spent, High Imports and Exports of goods and services and Low Inflation. Hence, this segment can be clustered as **Developed countries.**



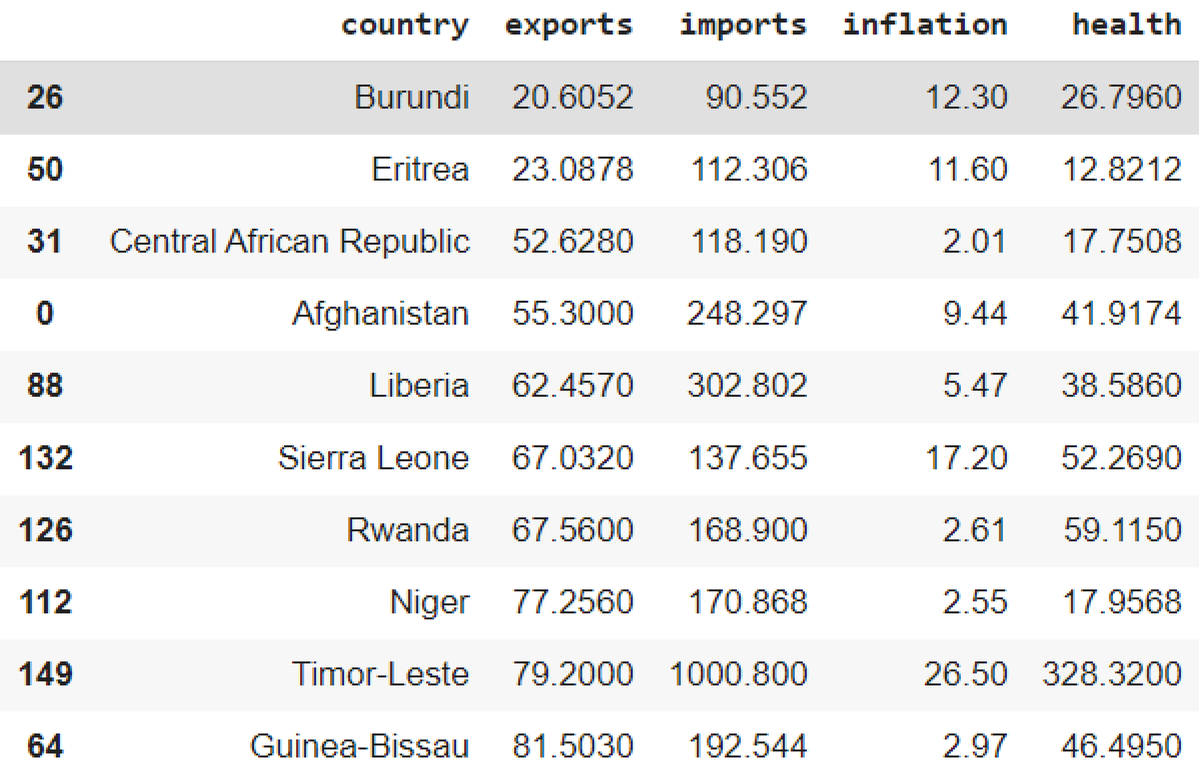


Cluster_label	health	imports	exports	inflation
0	114.821765	827.028771	879.063521	11.911146
1	1098.913521	5702.860550	6197.379266	6.443802
2	6070.207550	38512.335000	45222.215000	1.571125

Hierarchical Clustering: Country Identified for NGO AID

(Other Socio-economic factors)

- **Top 10 Country** identified using Hierarchical clustering based on following **other socio-economic factors**:
  - ***Health***
  - ***Imports***
  - ***Exports***
  - ***Inflation***



	country	exports	imports	inflation	health
26	Burundi	20.6052	90.552	12.30	26.7960
50	Eritrea	23.0878	112.306	11.60	12.8212
31	Central African Republic	52.6280	118.190	2.01	17.7508
0	Afghanistan	55.3000	248.297	9.44	41.9174
88	Liberia	62.4570	302.802	5.47	38.5860
132	Sierra Leone	67.0320	137.655	17.20	52.2690
126	Rwanda	67.5600	168.900	2.61	59.1150
112	Niger	77.2560	170.868	2.55	17.9568
149	Timor-Leste	79.2000	1000.800	26.50	328.3200
64	Guinea-Bissau	81.5030	192.544	2.97	46.4950

Conclusion :

(Countries identifies based on Gdpp, child mortality and income)

Top 10 **Countries identified for NGO AID**, which are Under developed countries, based on ***Gdpp, child mortality and income features*** using both **K-Means and Hierarchical clustering** are as follows:

1. Burundi
2. Liberia
3. Congo, Dem. Rep.
4. Niger
5. Sierra Leone
6. Madagascar
7. Mozambique
8. Central African Republic
9. Malawi
10. Eritrea

Reasons for AID for these countries:

- ***High Child Mortality***
- ***Low GDPP***
- ***Low Income***

Conclusion :

(Countries identifies based on other socio-economic factors)

Top 10 countries identified for NGO AID, which are Under developed countries, ***based on other socio-economic factors like exports, imports, inflation and health etc***. features using both **K-Means and Hierarchical clustering** are as follows:

1. Burundi
2. Eritrea
3. Central African Republic
4. Afghanistan
5. Liberia
6. Sierra Leone
7. Rwanda
8. Niger
9. Timor-Leste
10. Guinea-Bissau

Reasons for AID for these countries:

- ***High Inflation***
- ***Low Exports and Imports of goods and services***
- ***Low spent on health***

Conclusion :

If we considered **all features, including socio-economic factors**, then below are the list of 6 **Under Developed countries** that needs to be considered for NGO AID.

1. Burundi
2. Liberia
3. Niger
4. Central African Republic
5. Sierra Leone
6. Eritrea

END