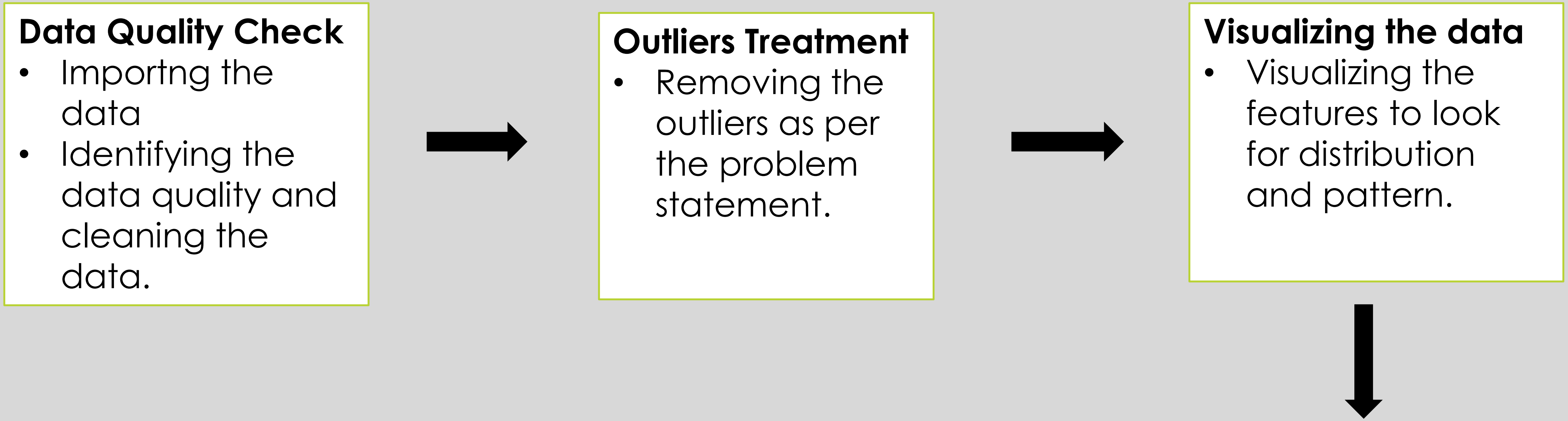
CLUSTERING ASSIGNMENT

Countries Clustering

Problem Statement:

During the recent funding program, HELP NGO have been able to raise around $ 10 Million. Now as a analyst, we have to decide how to use this money strategically and effectively and have to come up with list of countries that are in direst need of aid.

Analysis Approach:



Data Quality Check

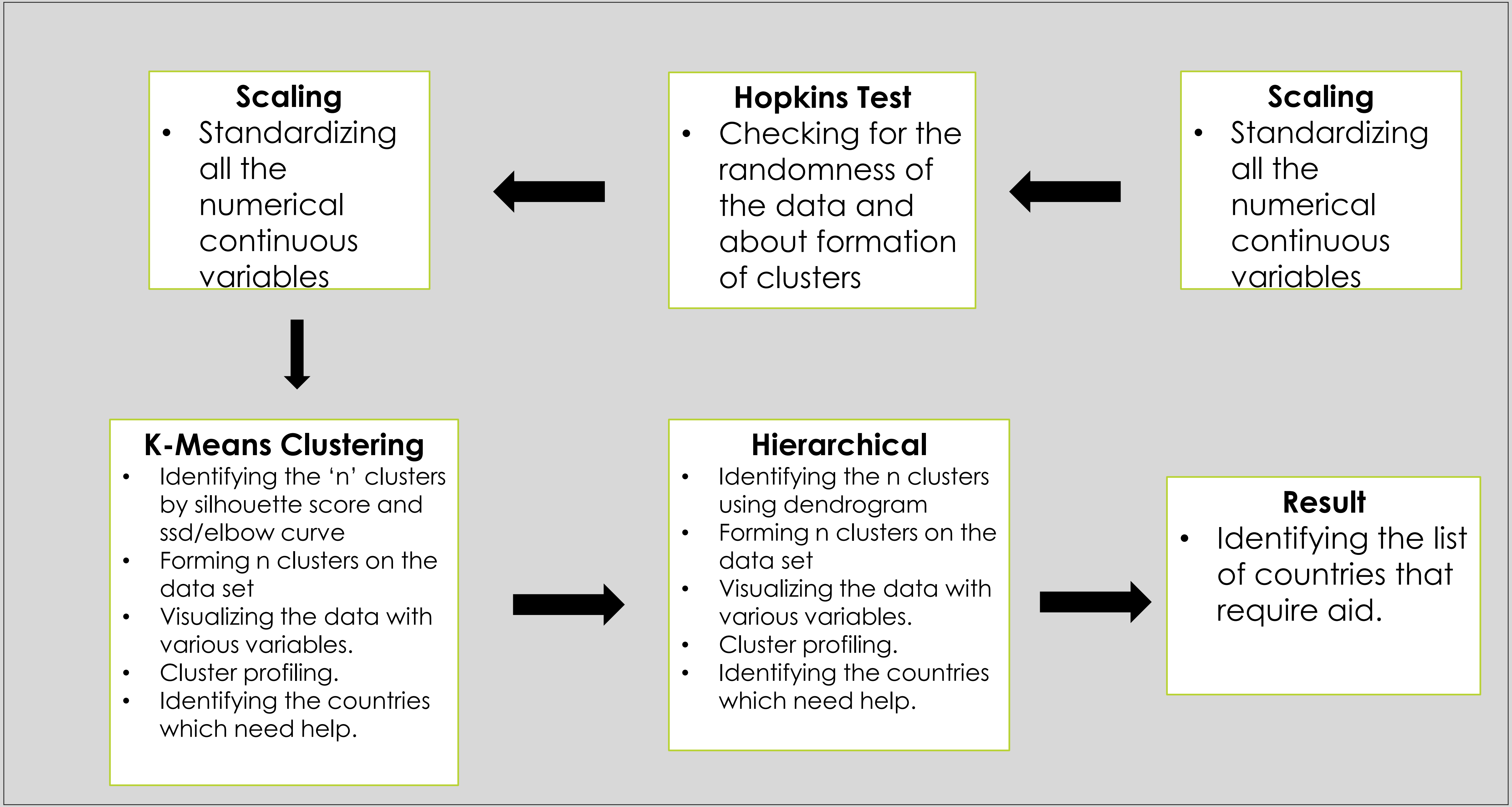
- Importng the data
- Identifying the data quality and cleaning the data.

Outliers Treatment

- Removing the outliers as per the problem statement.

Visualizing the data

- Visualizing the features to look for distribution and pattern.



Scaling

- Standardizing all the numerical continuous variables

Hopkins Test

- Checking for the randomness of the data and about formation of clusters

Scaling

- Standardizing all the numerical continuous variables

K-Means Clustering

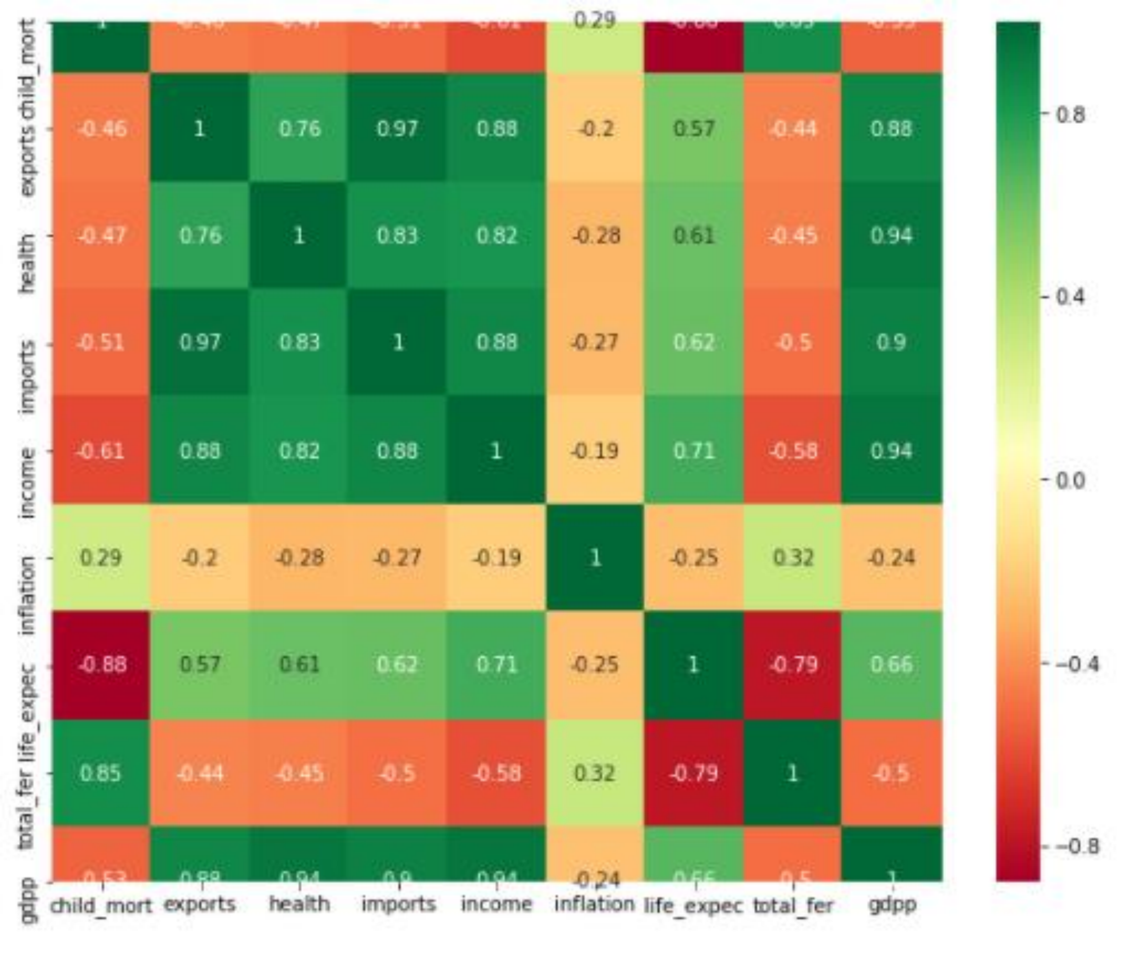
- Identifying the 'n' clusters by silhouette score and ssd/elbow curve
- Forming n clusters on the data set
- Visualizing the data with various variables.
- Cluster profiling.
- Identifying the countries which need help.

Hierarchical

- Identifying the n clusters using dendrogram
- Forming n clusters on the data set
- Visualizing the data with various variables.
- Cluster profiling.
- Identifying the countries which need help.

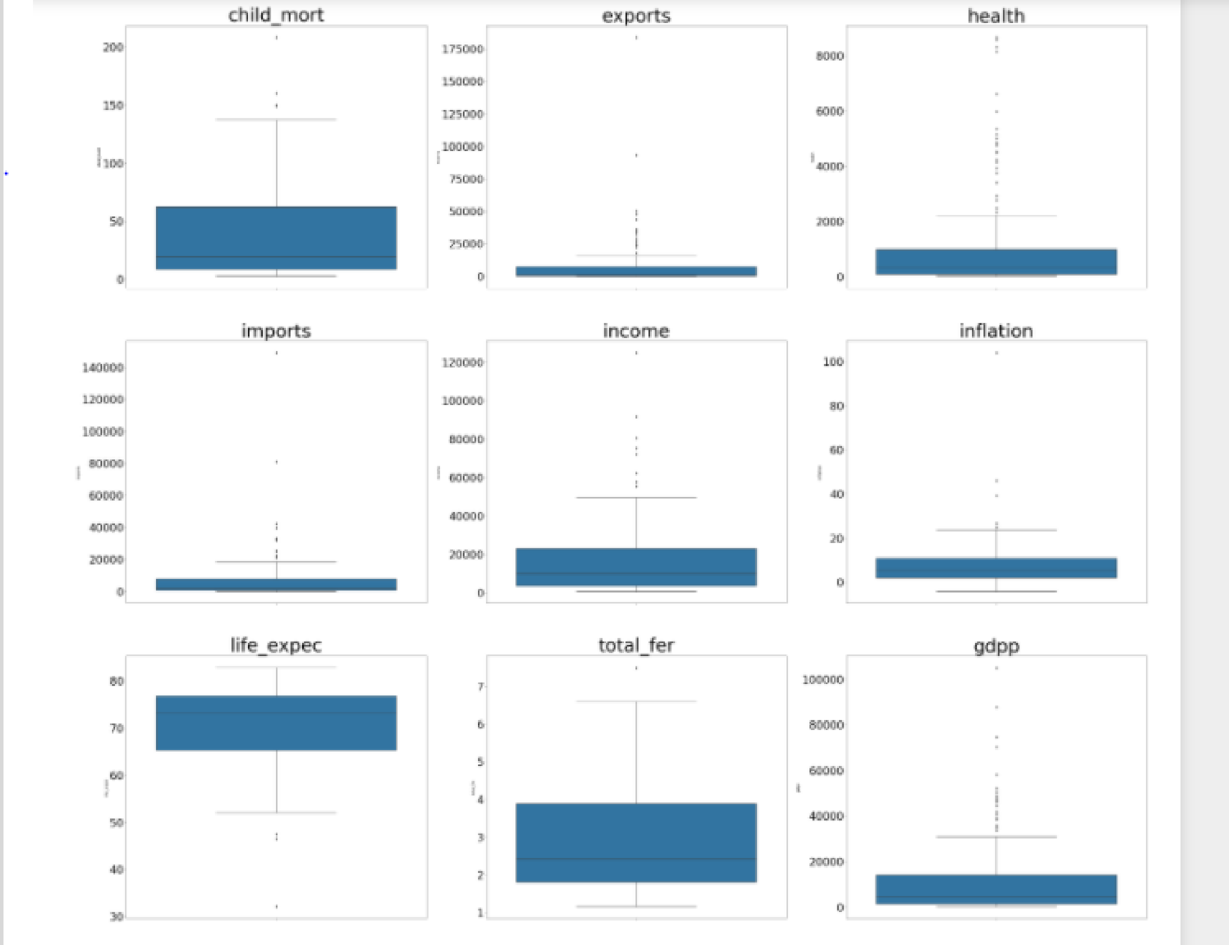
Result

- Identifying the list of countries that require aid.



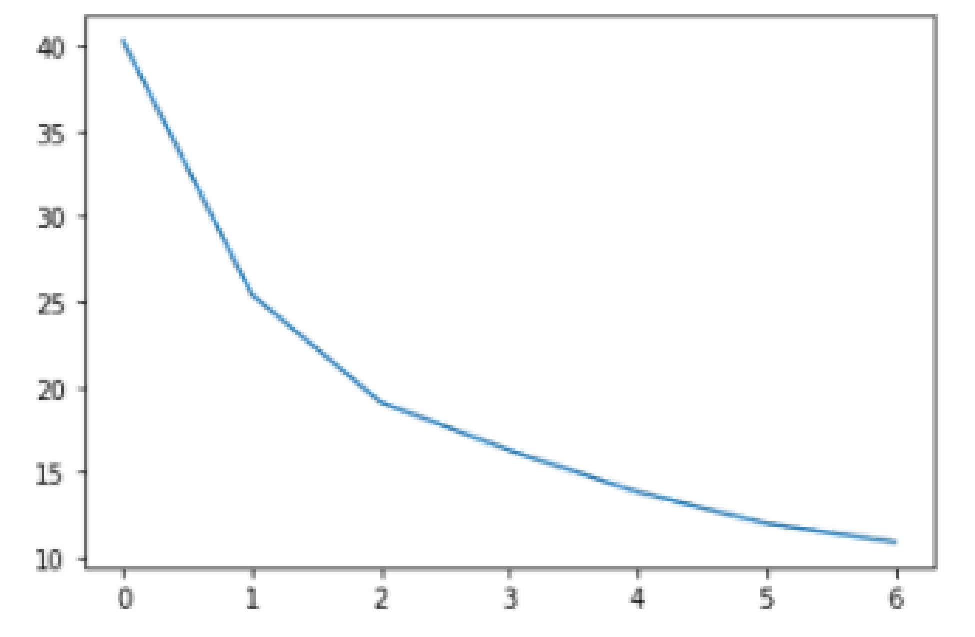
- After data quality check we have removed outliers from few columns which do not require any aid.
- All the values are standardized to get the better performance of the data.
- Looking at the heatmap we can infer that imports, health, exports, gdpp columns have high correlation.

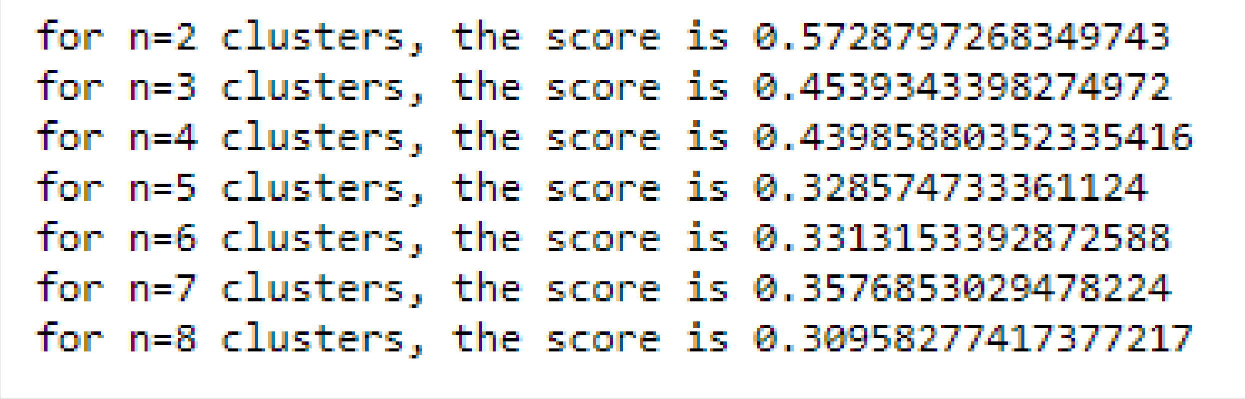
Outliers



- From the figure we can find that Health, GDPP, Income columns are having outliers

K-Means Clustering



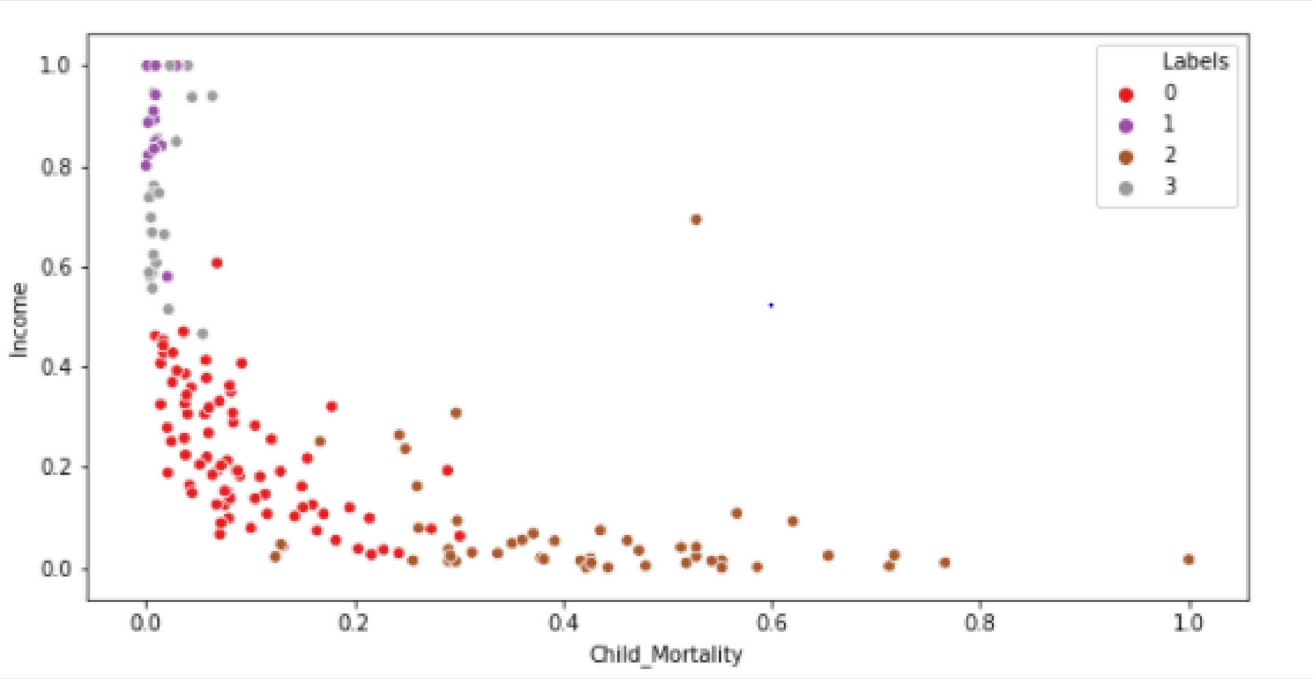


```
for n=2 clusters, the score is 0.5728797268349743
for n=3 clusters, the score is 0.453934398274972
for n=4 clusters, the score is 0.43985880352335416
for n=5 clusters, the score is 0.328574733361124
for n=6 clusters, the score is 0.331315339287258
for n=7 clusters, the score is 0.3576853029478224
for n=8 clusters, the score is 0.30958277417377217
```

Silhouette Analysis

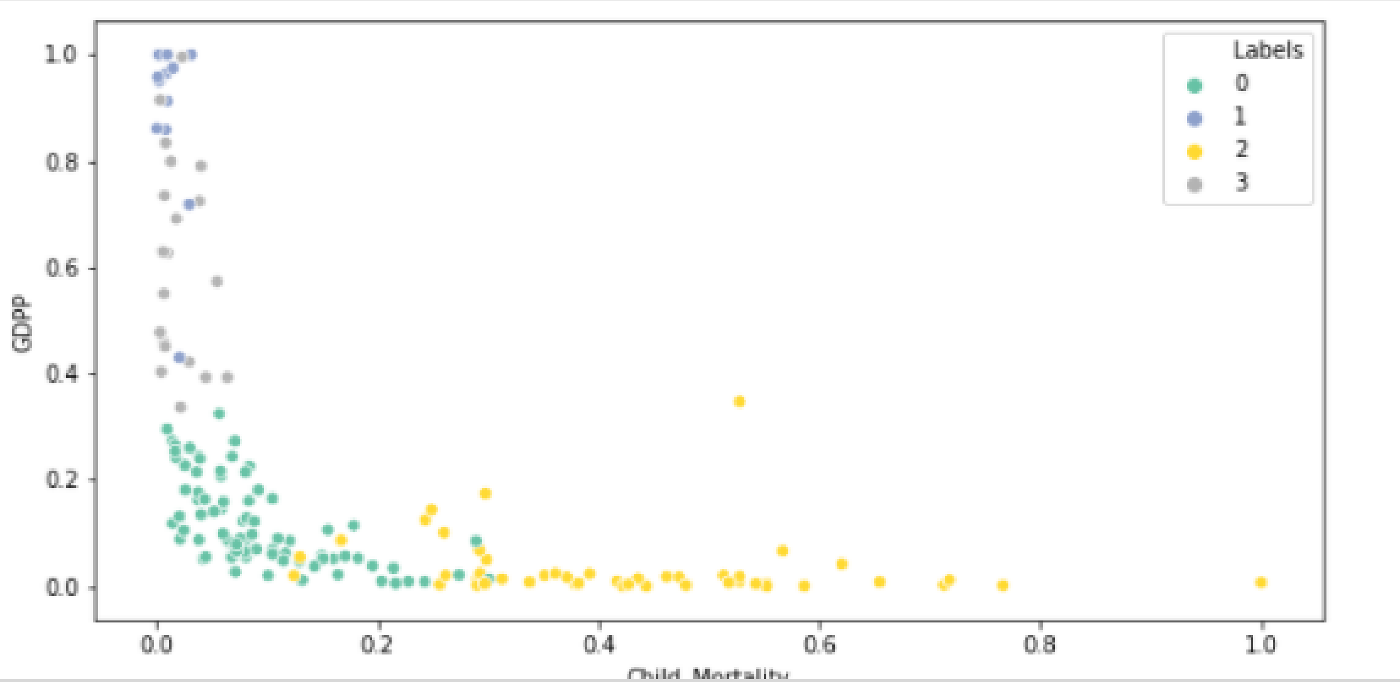
- By seeing the Silhouette analysis we can take n=4 as it is having optimal score

K-Means Clustering



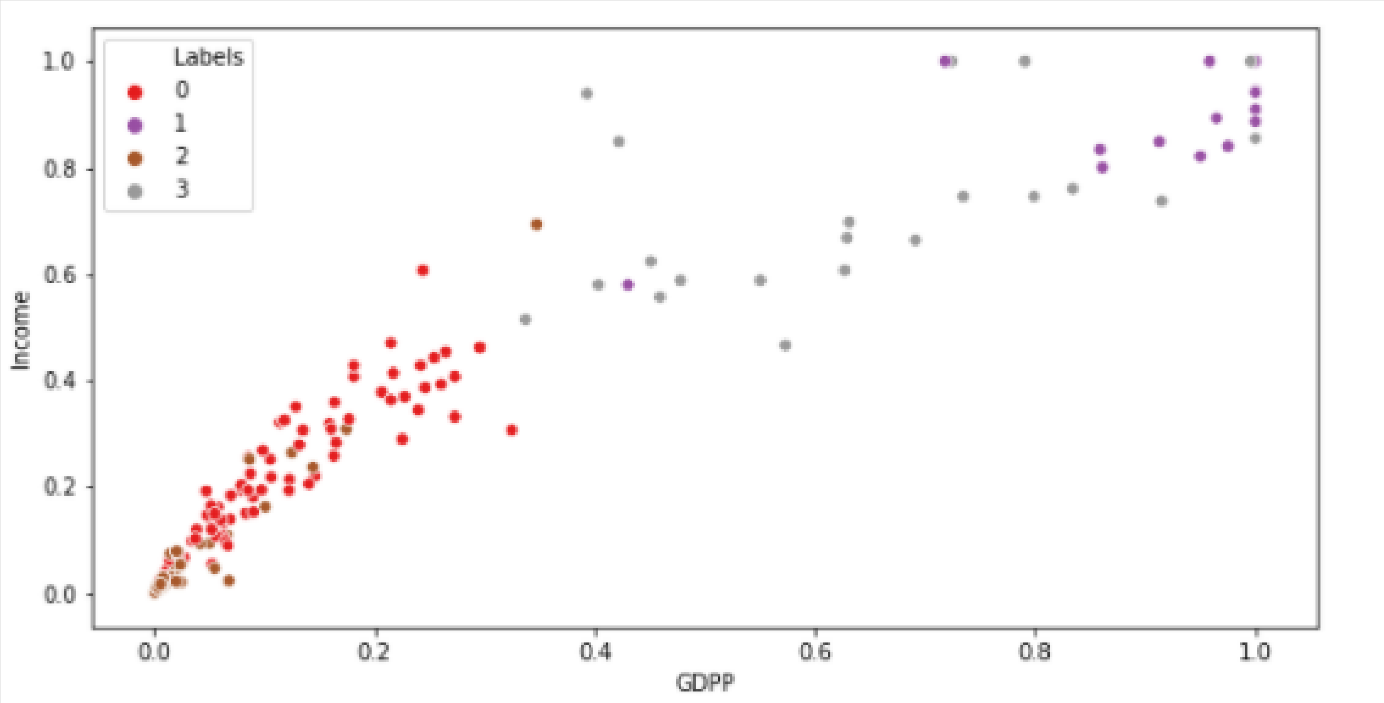
- Scatter plot between Income and Child Moratality, we can see the clusters formed.

K-Means Clustering



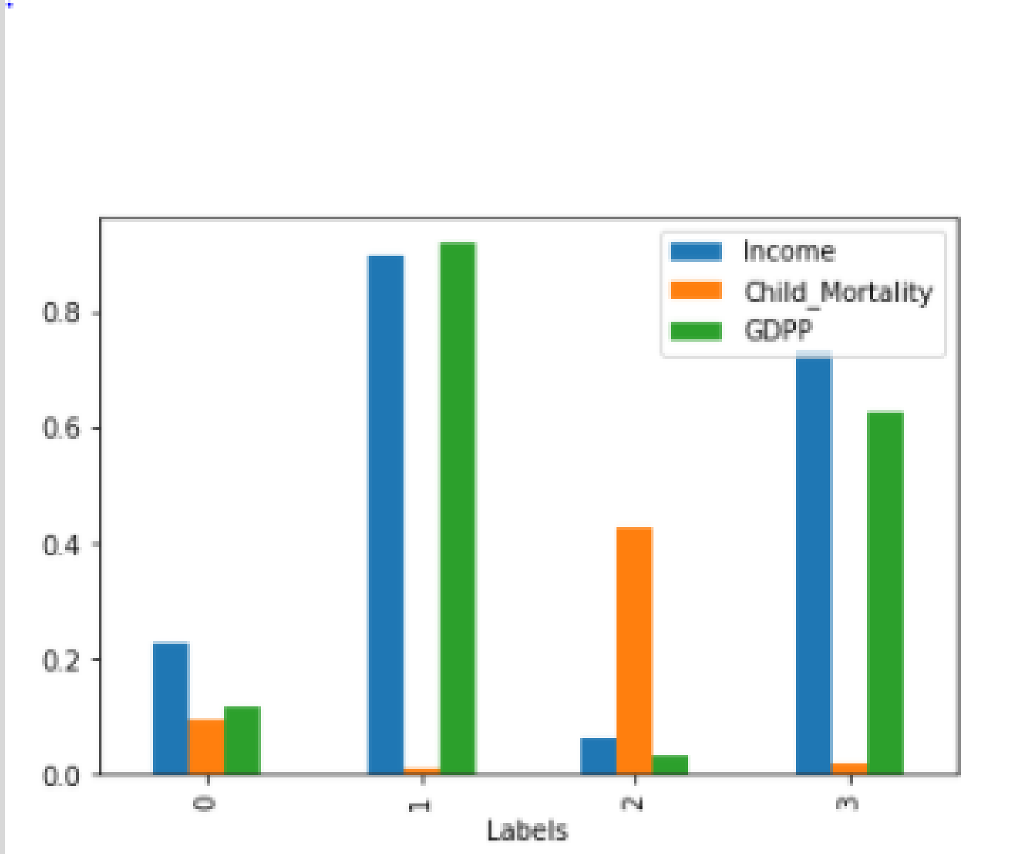
- Scatter plot between GDPP and Child Moratality, we can see the clusters formed.

K-Means Clustering



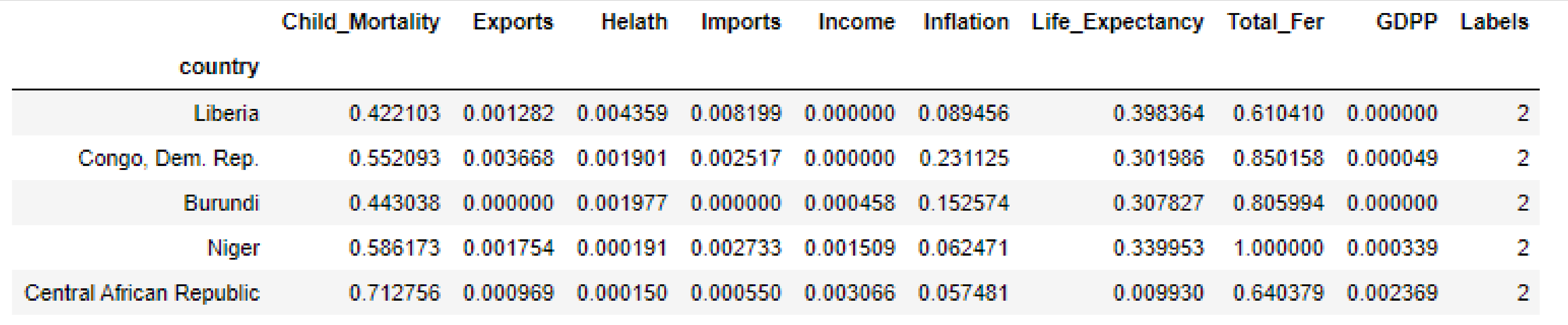
- Scatter plot between GDPP and Income, we can see the clusters formed.

Cluster Profiling:



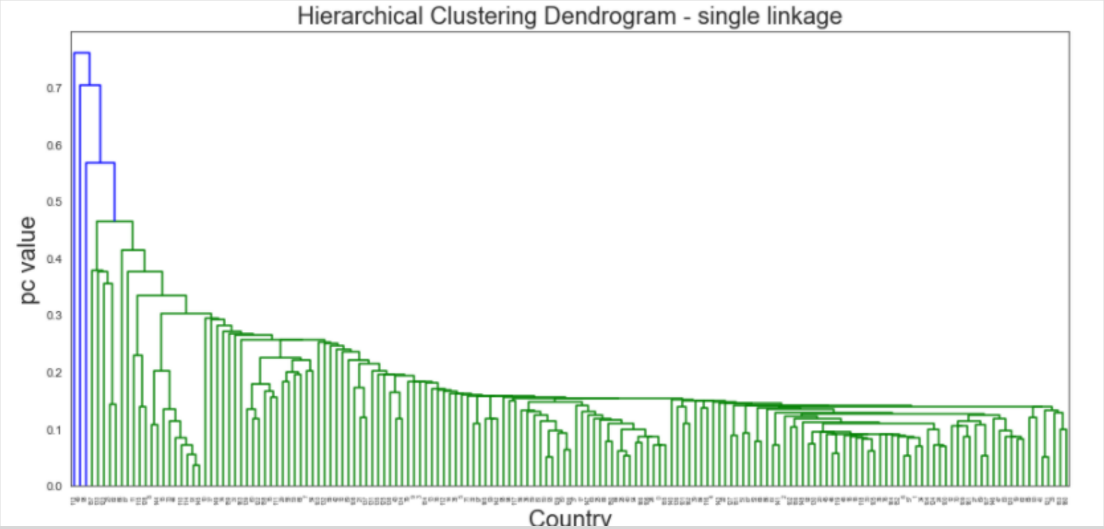
- From the figure we can see that the cluster n=2 is having the lowest income and gdpp and highest child mortality.

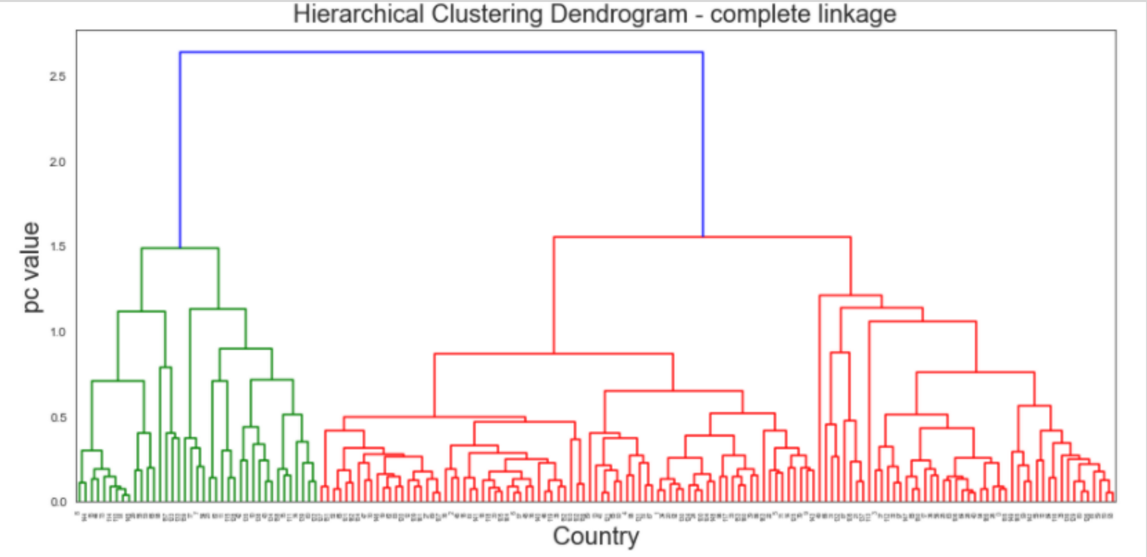
Countries under K-Means



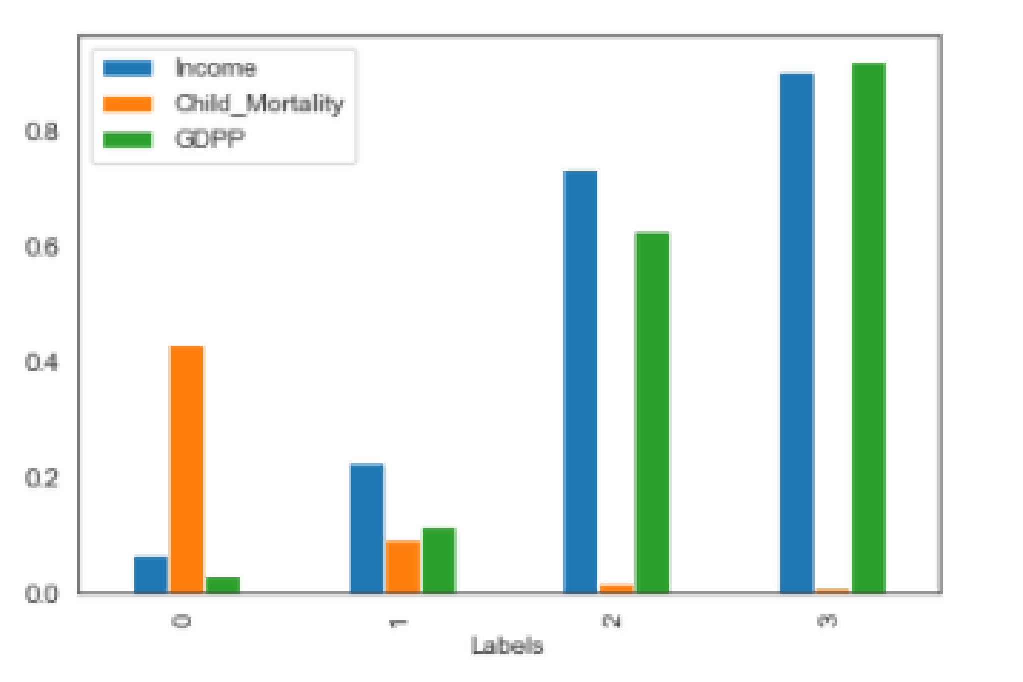
	Child_Mortality	Exports	Helath	Imports	Income	Inflation	Life_Expectancy	Total_Fer	GDPP	Labels
country										
Liberia	0.422103	0.001282	0.004359	0.008199	0.000000	0.089456	0.398364	0.610410	0.000000	2
Congo, Dem. Rep.	0.552093	0.003668	0.001901	0.002517	0.000000	0.231125	0.301986	0.850158	0.000049	2
Burundi	0.443038	0.000000	0.001977	0.000000	0.000458	0.152574	0.307827	0.805994	0.000000	2
Niger	0.586173	0.001754	0.000191	0.002733	0.001509	0.062471	0.339953	1.000000	0.000339	2
Central African Republic	0.712756	0.000969	0.000150	0.000550	0.003066	0.057481	0.009930	0.640379	0.002369	2

Hierarchical Clustering:



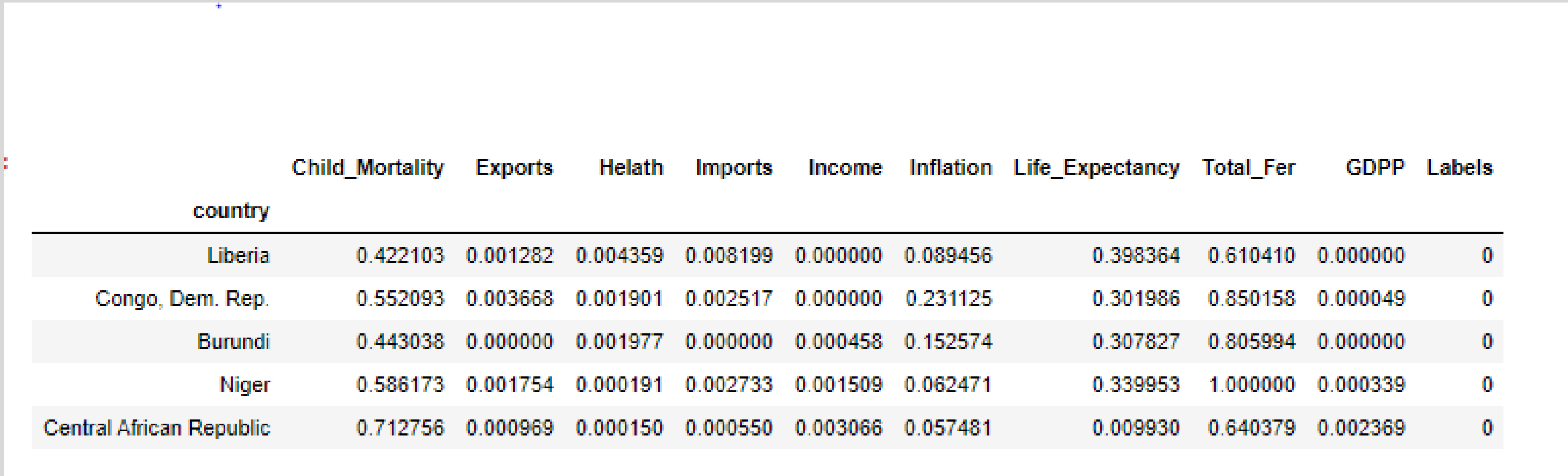


Cluster Profiling



- We can observe the cluster n=0 is having highest child mortality, low income and low gdpp

Countries list



country	Child_Mortality	Exports	Helath	Imports	Income	Inflation	Life_Expectancy	Total_Fer	GDPP	Labels
Liberia	0.422103	0.001282	0.004359	0.008199	0.000000	0.089456	0.398364	0.610410	0.000000	0
Congo, Dem. Rep.	0.552093	0.003668	0.001901	0.002517	0.000000	0.231125	0.301986	0.850158	0.000049	0
Burundi	0.443038	0.000000	0.001977	0.000000	0.000458	0.152574	0.307827	0.805994	0.000000	0
Niger	0.586173	0.001754	0.000191	0.002733	0.001509	0.062471	0.339953	1.000000	0.000339	0
Central African Republic	0.712756	0.000969	0.000150	0.000550	0.003066	0.057481	0.009930	0.640379	0.002369	0

- We can see that both K-Means and hierarchical clustering are giving same results

Summary

- So after the analysis by both K-Means and Hierarchical clustering we found out that both are giving the same countries which are at the bottom list which require financial aid.
- Countries are:

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

Title Lorem Ipsum

LOREM IPSUM DOLOR SIT AMET, CONSECTETUER ADIPISCING ELIT.

NUNC VIVERRA IMPERDIET ENIM. FUSCE EST. VIVAMUS A TELLUS.

PELLENTESQUE HABITANT MORBI TRISTIQUE SENECTUS ET NETUS.