Clustering of Countries

By - Mukesh Chaurasia

Abstract

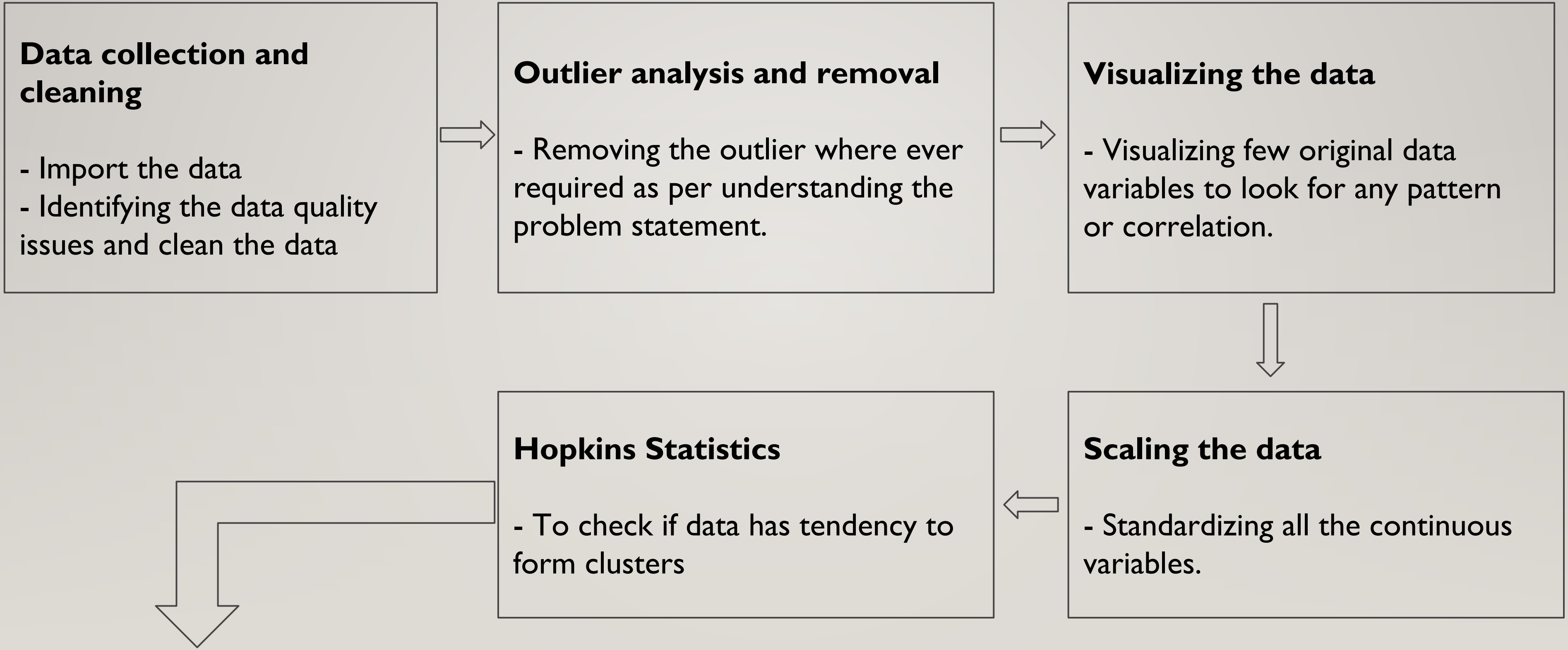
Objective:

We, HELP International humanitarian NGO, committed to fight poverty and provide the people of backward countries with basic amenities and relief during the time of disasters and natural calamities. We run a lot of operational projects from time to time, along with advocacy, drives to raise awareness as well as for funding purposes.

Problem statement:

During the recent funding programmes, we have been able to raise around $ 10 million. As an analyst, we have to come up with the countries list that are in the dire need of aid.

Analysis methodology



Data collection and cleaning

- Import the data
- Identifying the data quality issues and clean the data

Outlier analysis and removal

- Removing the outlier where ever required as per understanding the problem statement.

Visualizing the data

- Visualizing few original data variables to look for any pattern or correlation.

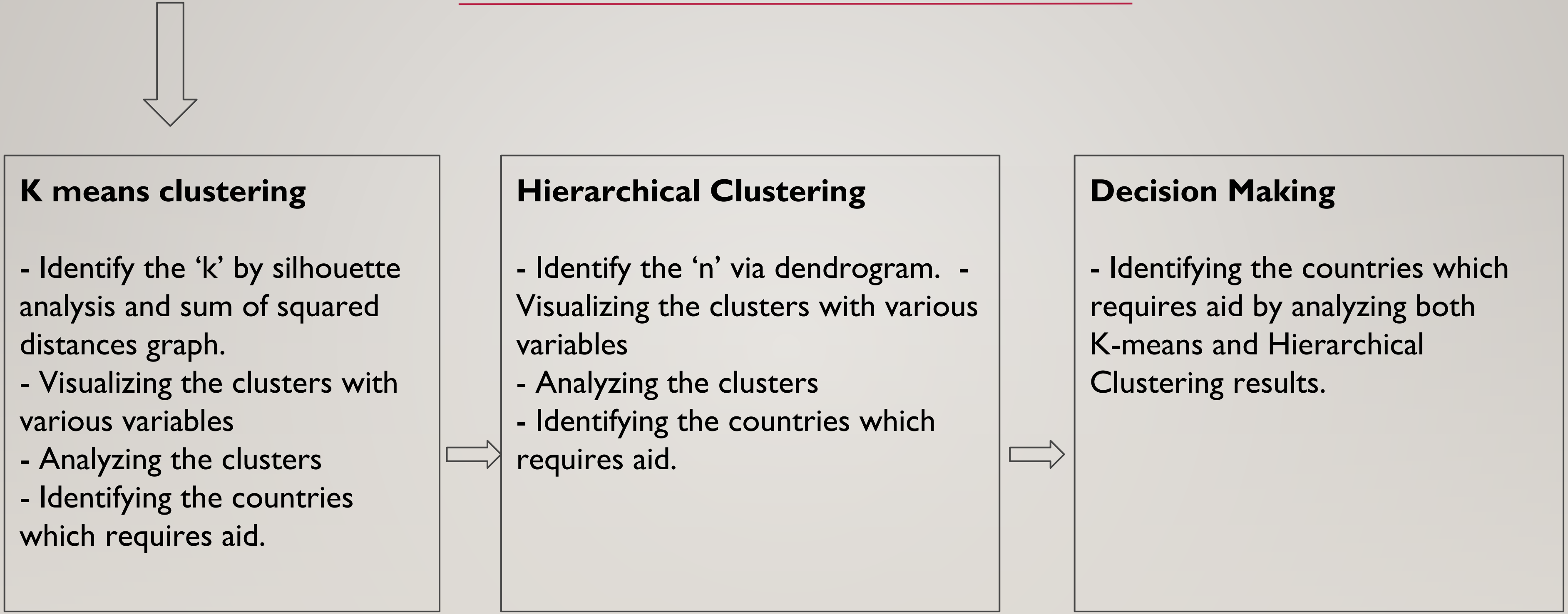
Scaling the data

- Standardizing all the continuous variables.

Hopkins Statistics

- To check if data has tendency to form clusters

Analysis methodology cont…



K means clustering

- Identify the ‘k’ by silhouette analysis and sum of squared distances graph.
- Visualizing the clusters with various variables
- Analyzing the clusters
- Identifying the countries which requires aid.

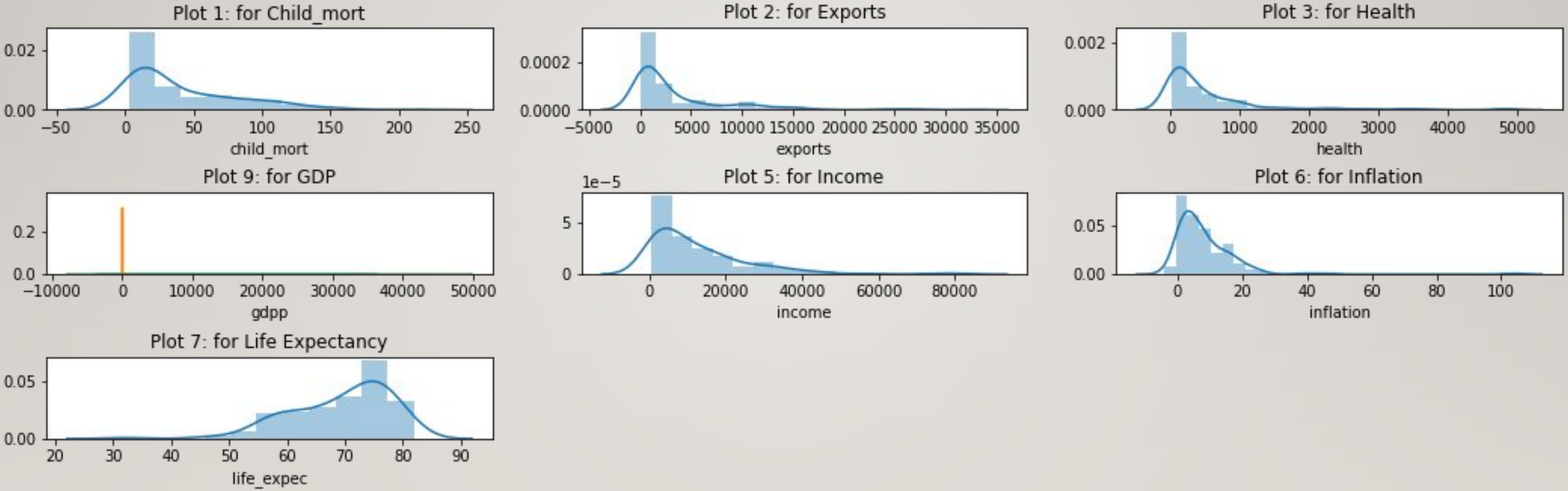
Hierarchical Clustering

- Identify the ‘n’ via dendrogram. - Visualizing the clusters with various variables
- Analyzing the clusters
- Identifying the countries which requires aid.

Decision Making

- Identifying the countries which requires aid by analyzing both K-means and Hierarchical Clustering results.

EDA (Univariate analysis Categorical variables)



EDA (Univariate analysis Categorical variables) Contd...

Plot 1 : Child Mortality

- There are many countries which has very less child mortality that's why it has peak at below 50.

Plot 2 : Exports

- Poor countries has less exports so there is peak at start below 5000.

Plot 3 : Health

- Poor countries do not have high budget for health so there is peak at start at below 1000.

Plot 4 : GDP

- Poor countries do not have high GDP so there is almost flat curve for GDP

Plot 5 : Income

- Per capita income is less for poor countries so there is peek at start close to zero.

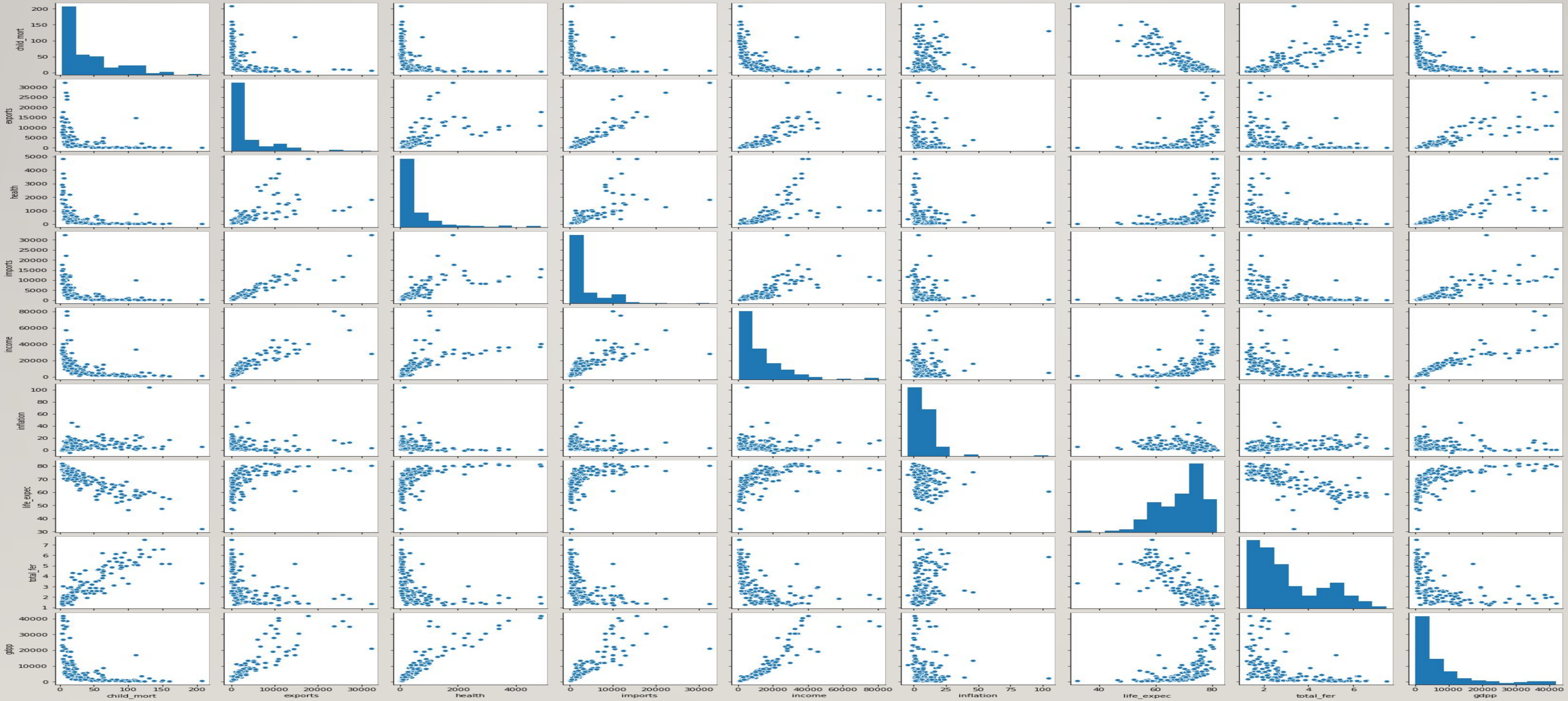
Plot 6 : Inflation

- Inflation rate are less for countries so peak is at start near to zero only.

Plot 7 : Life Expectancy

- Life expectancy is high except few countries so peak is at around 75.

EDA (Bivariate analysis Continuous Continuous)



EDA (Bivariate analysis Continuous Continuous) Contd...

1. GDP vs Child_mort
- it's clearly indicates there is negative correlation between GDP and child mortality. As GDP increases Child mortality decreases

2. GDP vs exports
- GDP and exports has positive correlation. GDP increases exports also increases.

3. GDP vs health
- GDP and health has positive correlation. GDP increases health also increases.

4. GDP vs imports
- GDP and imports has positive correlation. GDP increases imports also increases.

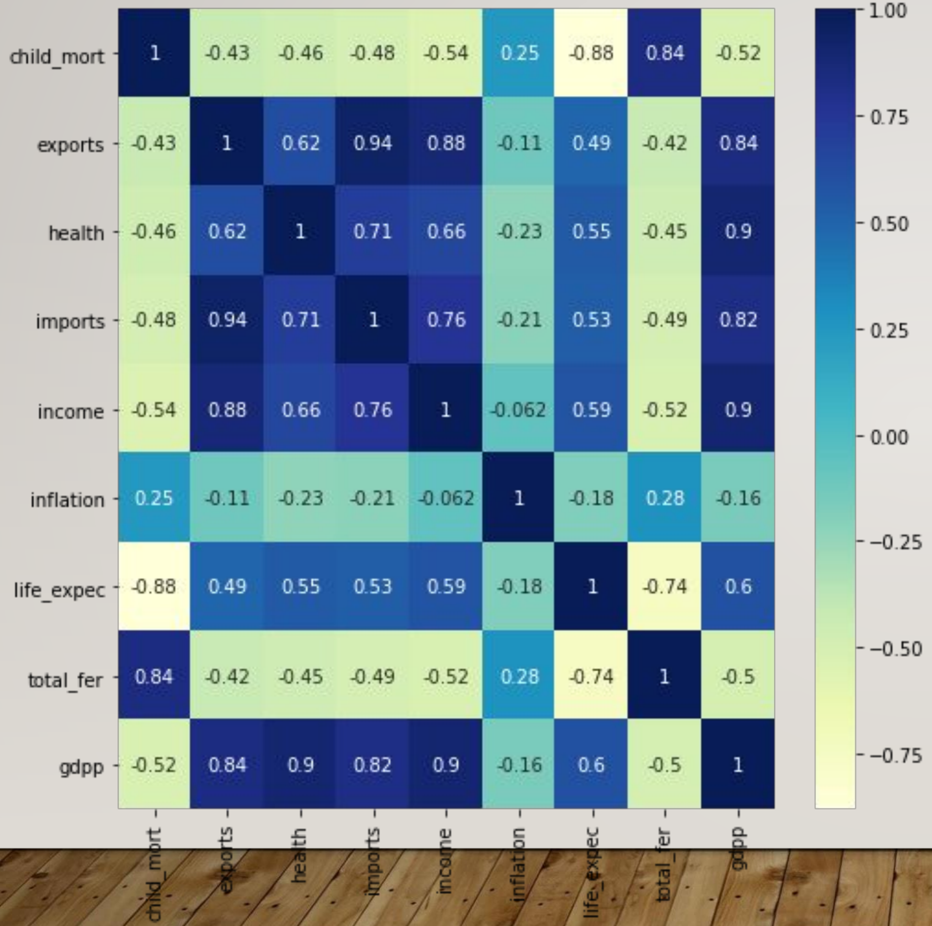
5. GDP vs income
- GDP and income has positive correlation. GDP increases income also increases.

6. GDP vs inflation
- countries having low GDP has higher inflation rate.

7. GDP vs life_expec
- countries having high GDP has higher life expectancy.

8. GDP vs total_fert
- countries with low GDP has higher fertility rate.

Correlation in the data:



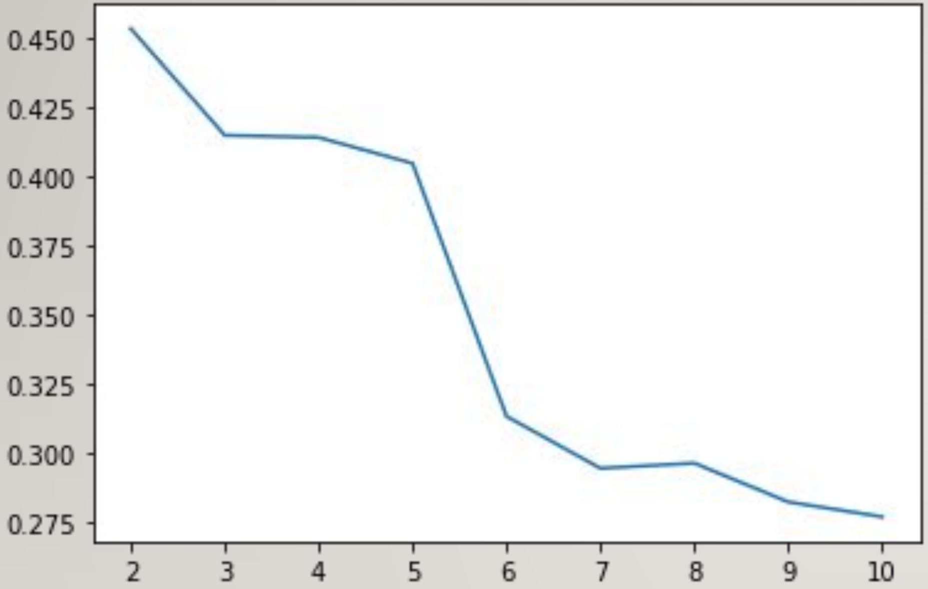
	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
child_mort	1	-0.43	-0.46	-0.48	-0.54	0.25	-0.88	0.84	-0.52
exports	-0.43	1	0.62	0.94	0.88	-0.11	0.49	-0.42	0.84
health	-0.46	0.62	1	0.71	0.66	-0.23	0.55	-0.45	0.9
imports	-0.48	0.94	0.71	1	0.76	-0.21	0.53	-0.49	0.82
income	-0.54	0.88	0.66	0.76	1	-0.062	0.59	-0.52	0.9
inflation	0.25	-0.11	-0.23	-0.21	-0.062	1	-0.18	0.28	-0.16
life_expec	-0.88	0.49	0.55	0.53	0.59	-0.18	1	-0.74	0.6
total_fer	0.84	-0.42	-0.45	-0.49	-0.52	0.28	-0.74	1	-0.5
gdpp	-0.52	0.84	0.9	0.82	0.9	-0.16	0.6	-0.5	1

- After data cleaning , we removed outlier from gdpp column because the country with high gdpp would not require any aid as there are already doing good.

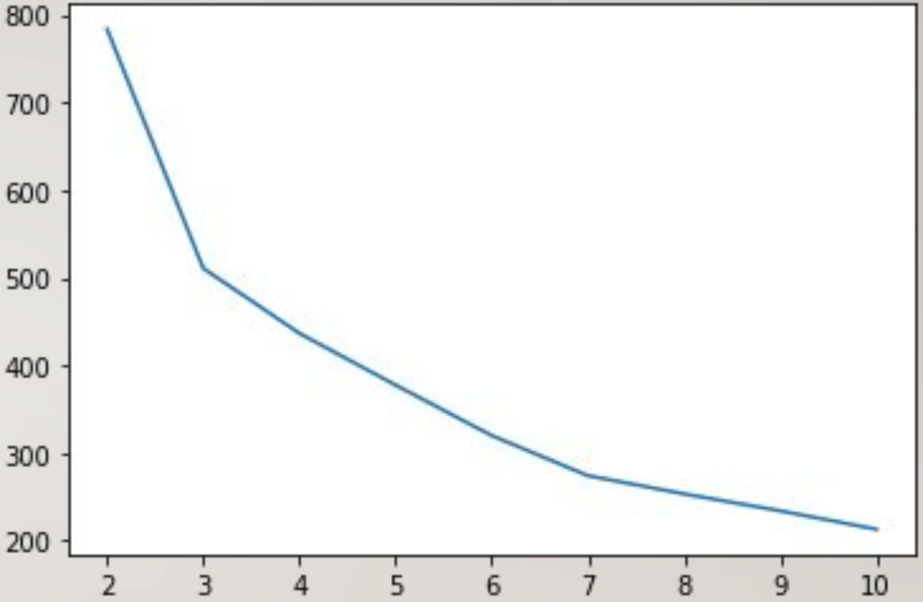
- We did standardized scaling to standardize all parameters on cleaned, outlier removed data.

- Looking at the heatmap, we see that few variables like (total fertility, child mortality) , (income , gdpp) and (imports and exports) have high correlation.

K-means clustering



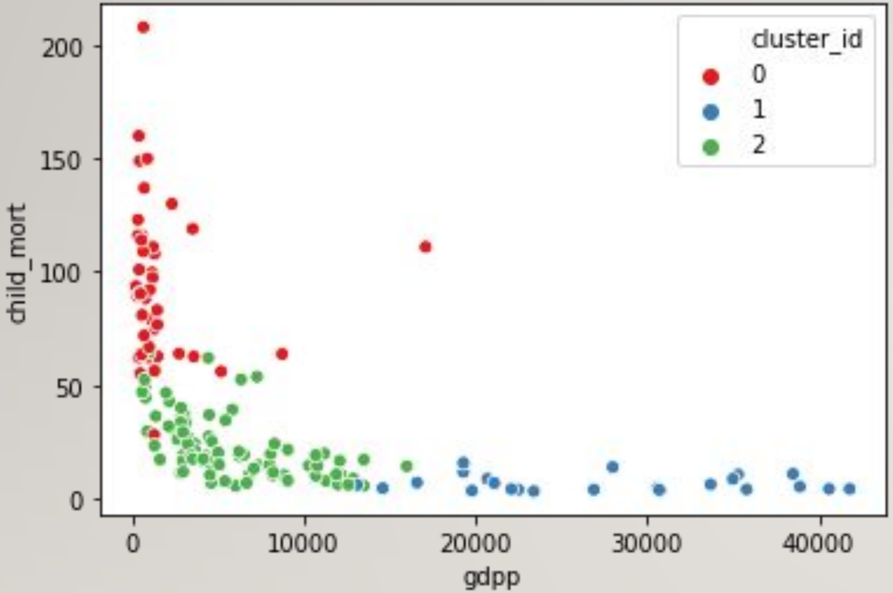
Silhouette Analysis



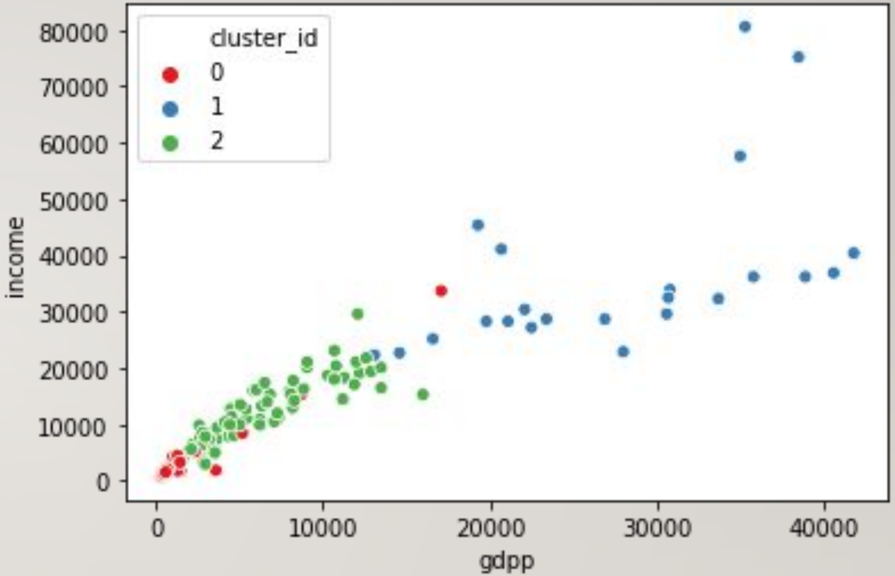
Sum of Squared Distances

By looking silhouette analysis, we see the highest peak is at k =3 and in sum of squared distances graph , we see that the elbow is also at 3 , so we are going ahead with k as 3.

K-means clustering

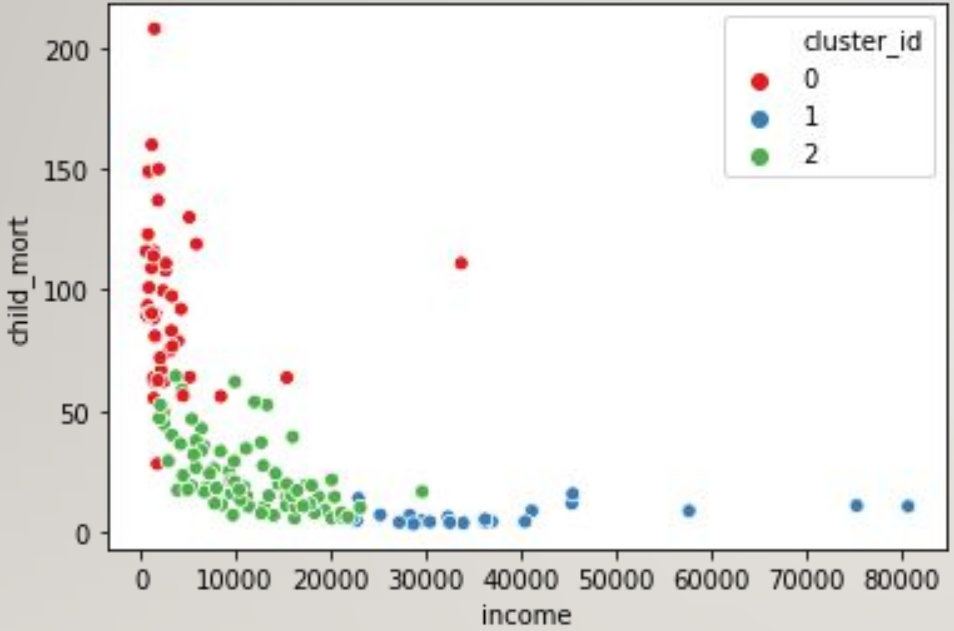


Scatter plot of gdpp and child_mort for various clusters. We see that for cluster 0 , gdpp is low and child mortality is very high.



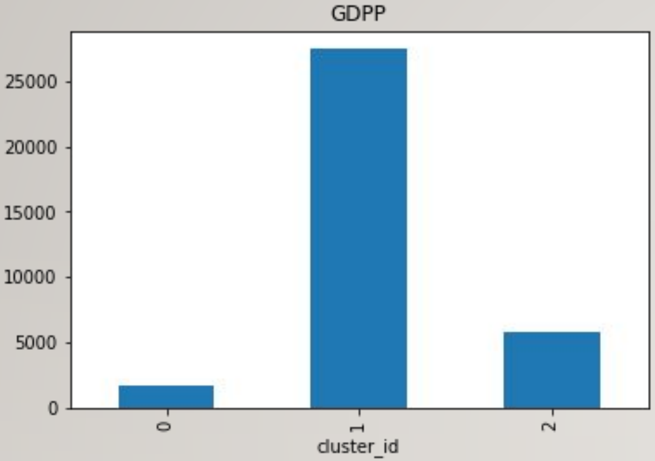
Scatter plot of gdpp and income for various clusters. We see that for cluster 0 , both gdpp and net income per person are very low.

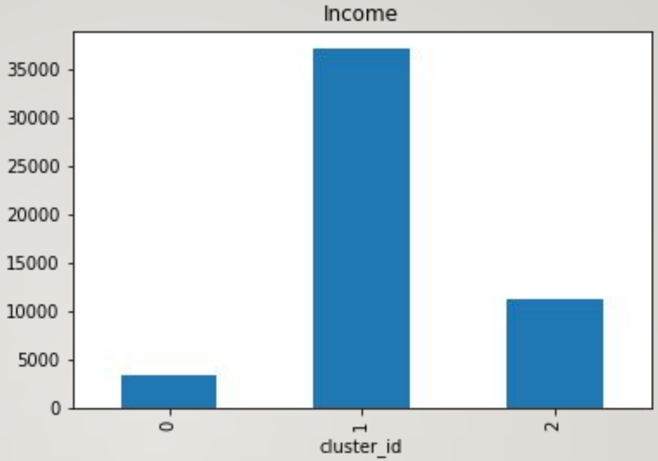
K-means clustering

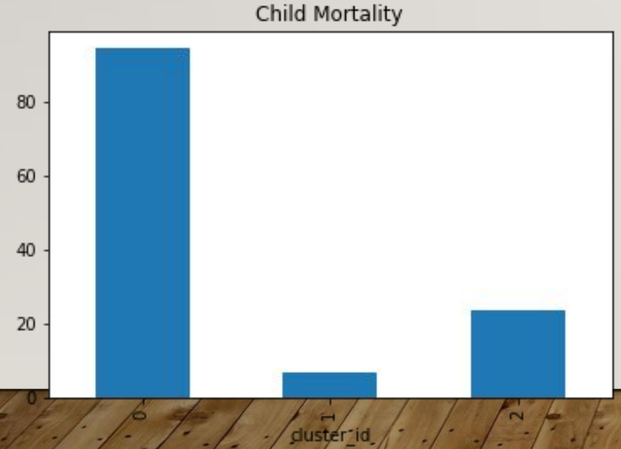


Scatter plot of income and child_mort for various clusters. We see that for cluster 0 , income is low and child mortality is very high.

K-means clustering





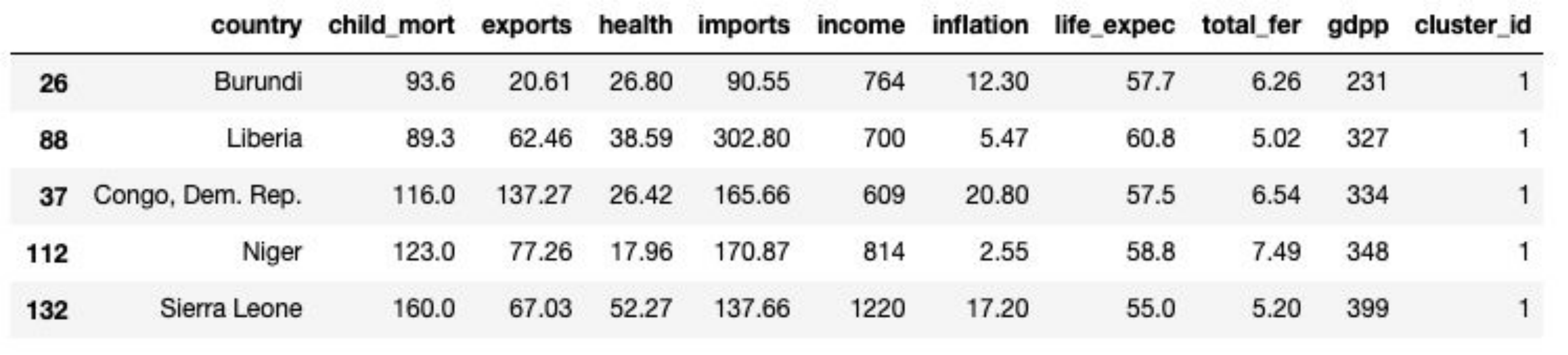


As per K- means cluster algorithm Cluster 0 is in need of help due to

- Low gdpp
- Low income
- High child mortality

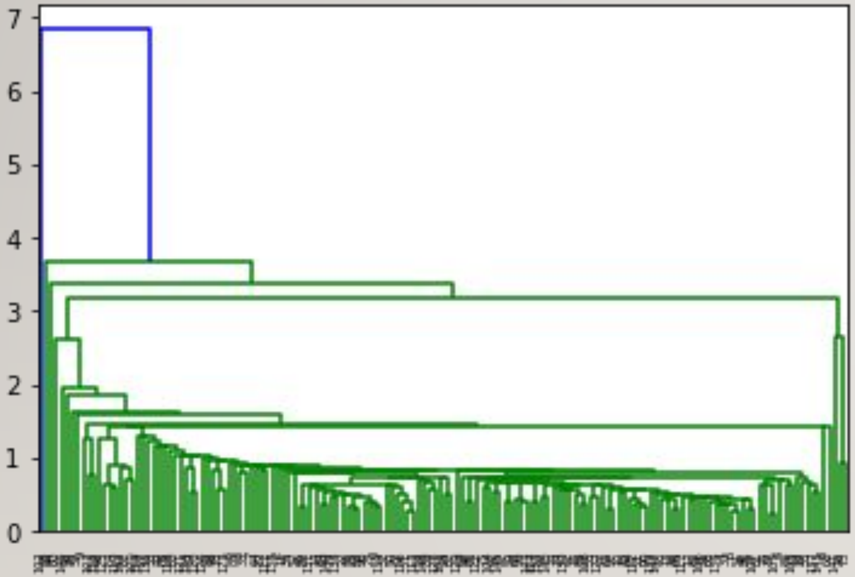
K-means clustering

5 countries under cluster 0 which need help are:

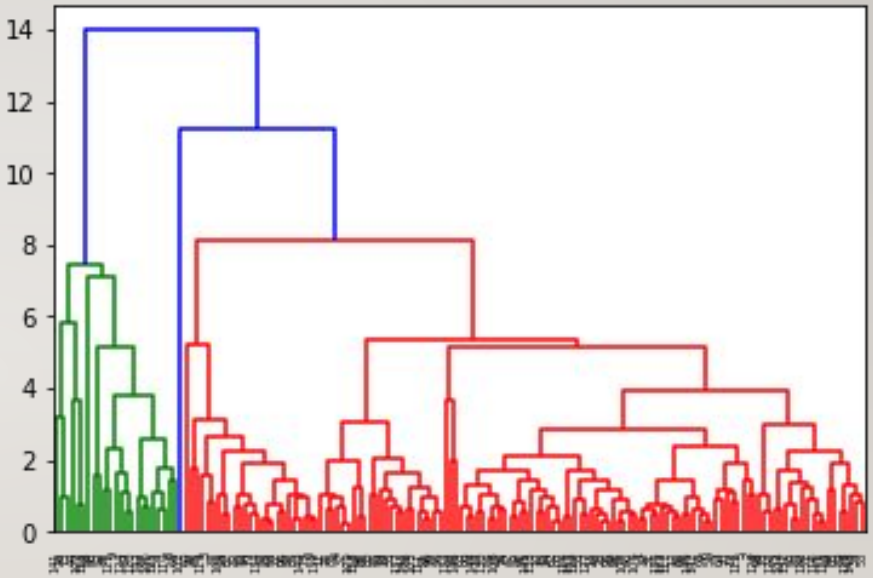


	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_id
26	Burundi	93.6	20.61	26.80	90.55	764	12.30	57.7	6.26	231	1
88	Liberia	89.3	62.46	38.59	302.80	700	5.47	60.8	5.02	327	1
37	Congo, Dem. Rep.	116.0	137.27	26.42	165.66	609	20.80	57.5	6.54	334	1
112	Niger	123.0	77.26	17.96	170.87	814	2.55	58.8	7.49	348	1
132	Sierra Leone	160.0	67.03	52.27	137.66	1220	17.20	55.0	5.20	399	1

Hierarchical clustering

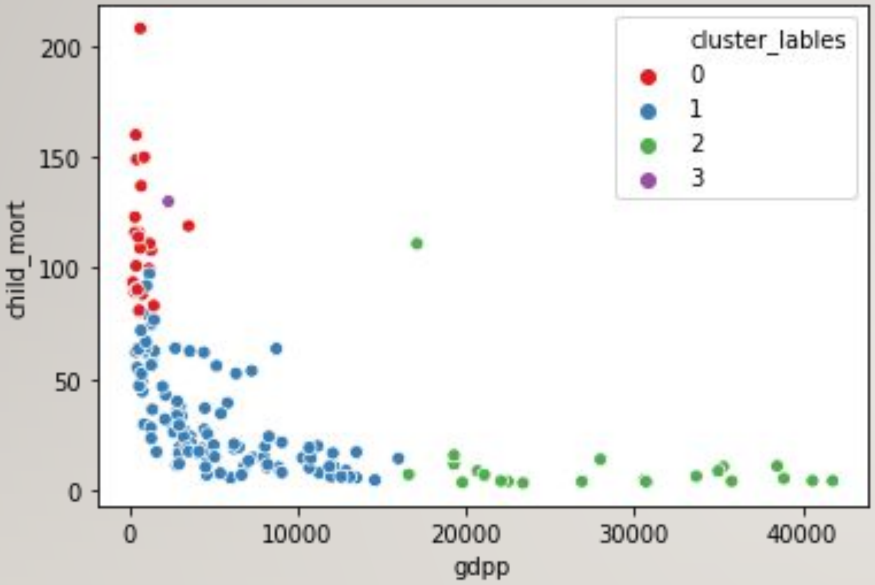


Single method hierarchical clustering

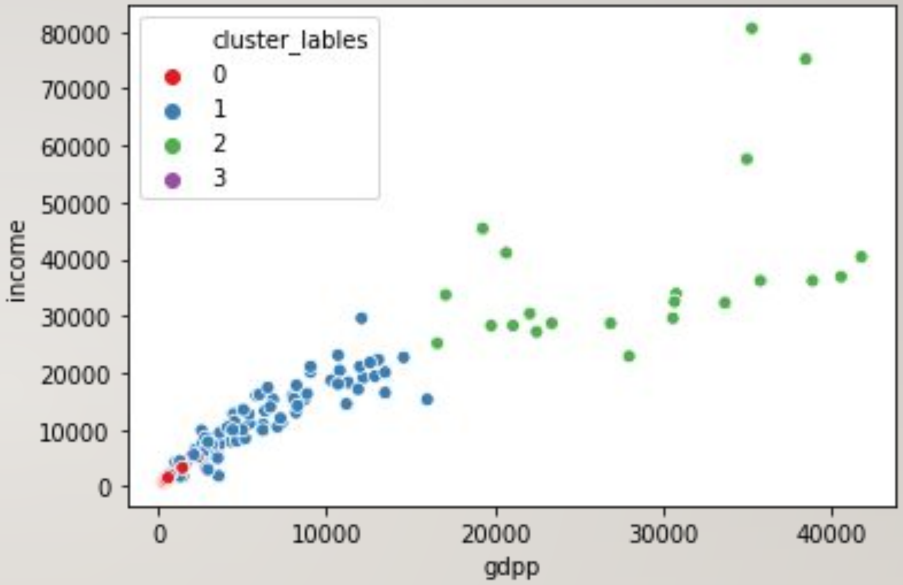


We are going for **Complete method hierarchical clustering** as single method clustering is not clear. By looking at this dendogram taking n-clusters as 4.

Hierarchical clustering

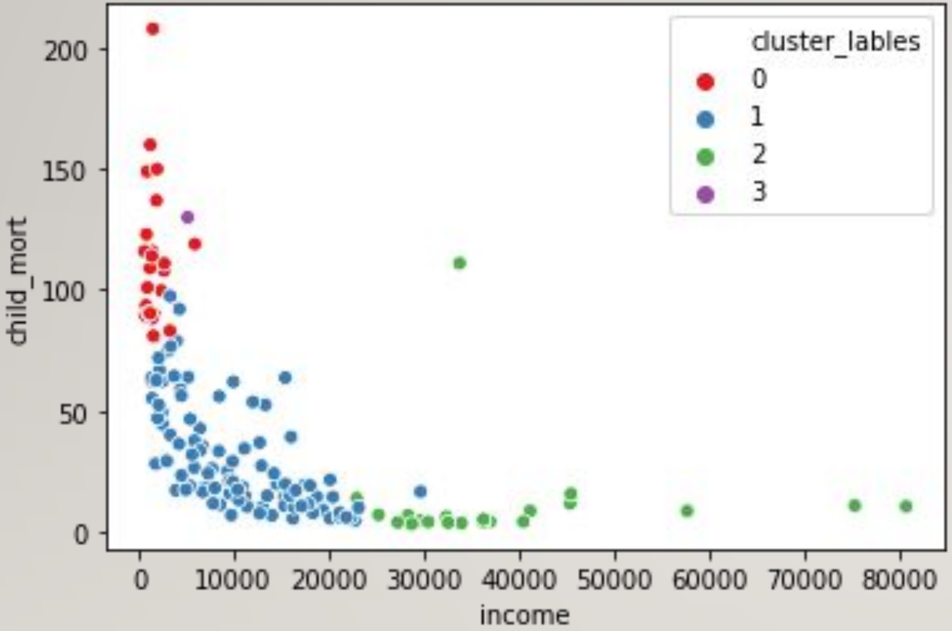


Scatter plot of gdpp and child_mort for various clusters. We see that for cluster 0 , gdpp is low and child mortality is very high.



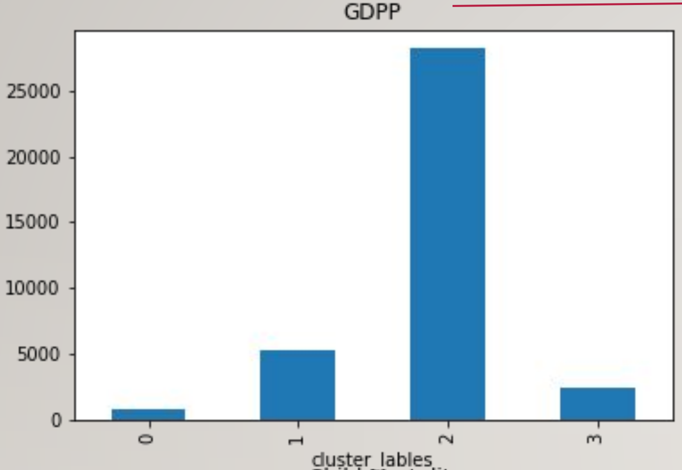
Scatter plot of gdpp and income for various clusters. We see that for cluster 0 , both gdpp and net income per person are very low.

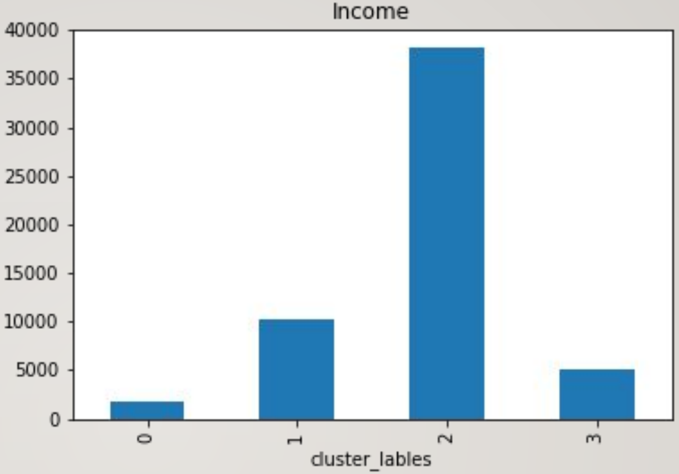
Hierarchical clustering

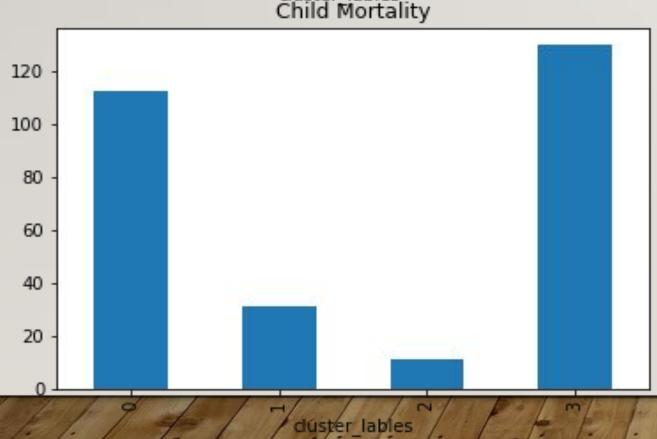


Scatter plot of income and child_mort for various clusters. We see that for cluster 0 , income is low and child mortality is very high.

Hierarchical clustering





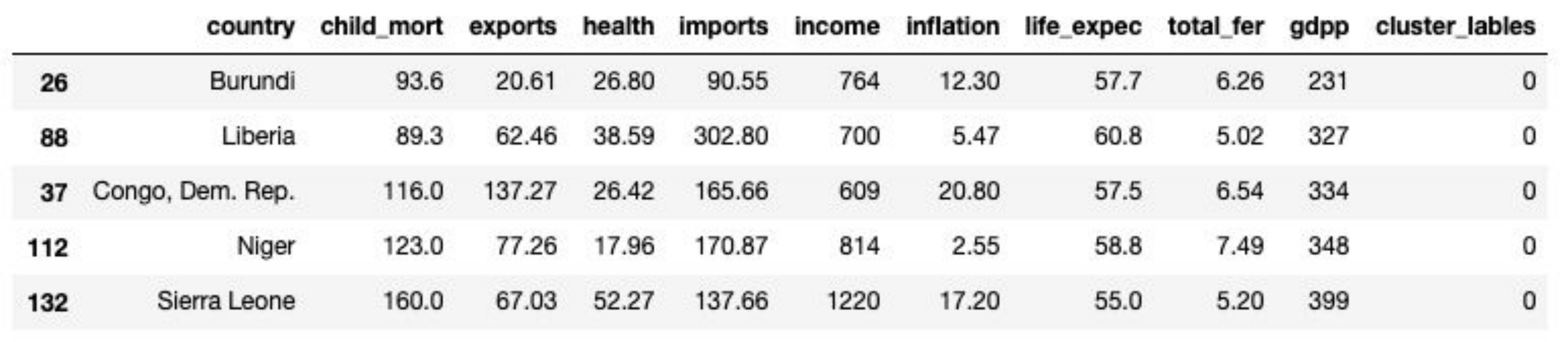


As per Hierarchical clustering algorithm Cluster 0 is in need of help due to

- Low gdpp
- Low income
- High child mortality

Hierarchical clustering

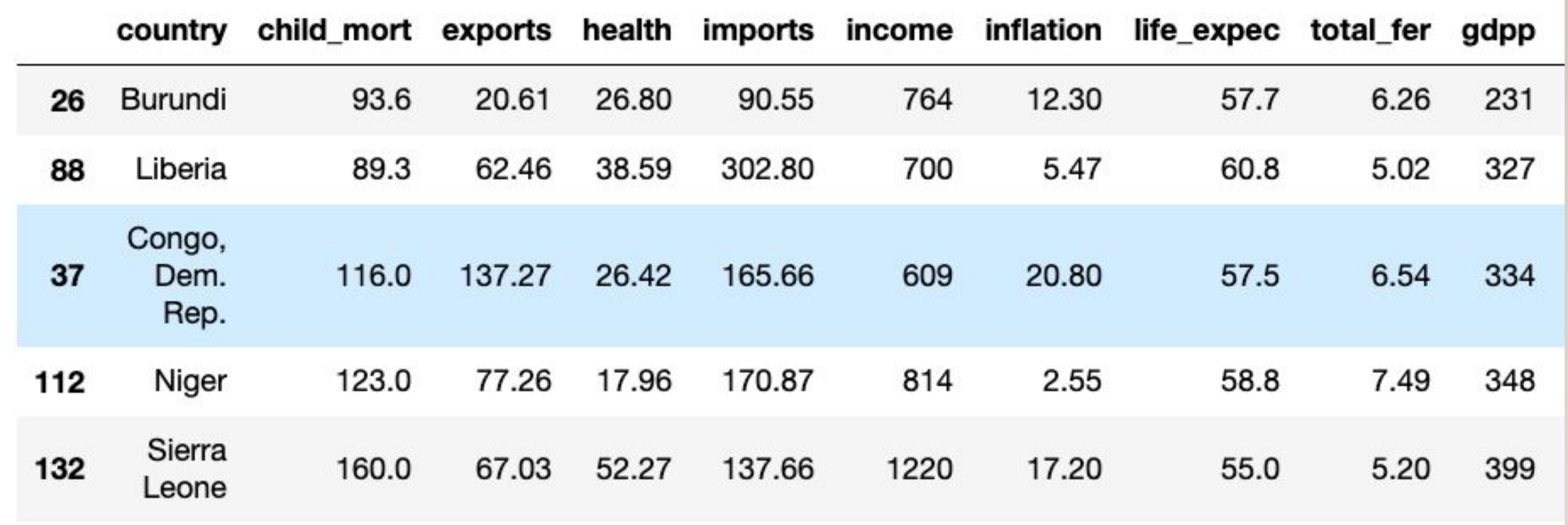
5 countries under cluster 0 which need help are:



	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_lables
26	Burundi	93.6	20.61	26.80	90.55	764	12.30	57.7	6.26	231	0
88	Liberia	89.3	62.46	38.59	302.80	700	5.47	60.8	5.02	327	0
37	Congo, Dem. Rep.	116.0	137.27	26.42	165.66	609	20.80	57.5	6.54	334	0
112	Niger	123.0	77.26	17.96	170.87	814	2.55	58.8	7.49	348	0
132	Sierra Leone	160.0	67.03	52.27	137.66	1220	17.20	55.0	5.20	399	0

Summary

As by both K means and Hierarchical clustering method - we have got same countries which requires aid. The following are the countries which are in dire need of aid by considering socio – economic factor into consideration:



	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
26	Burundi	93.6	20.61	26.80	90.55	764	12.30	57.7	6.26	231
88	Liberia	89.3	62.46	38.59	302.80	700	5.47	60.8	5.02	327
37	Congo, Dem. Rep.	116.0	137.27	26.42	165.66	609	20.80	57.5	6.54	334
112	Niger	123.0	77.26	17.96	170.87	814	2.55	58.8	7.49	348
132	Sierra Leone	160.0	67.03	52.27	137.66	1220	17.20	55.0	5.20	399

Thank You