CLUSTERING ASSIGNMENT

Analysis methodology

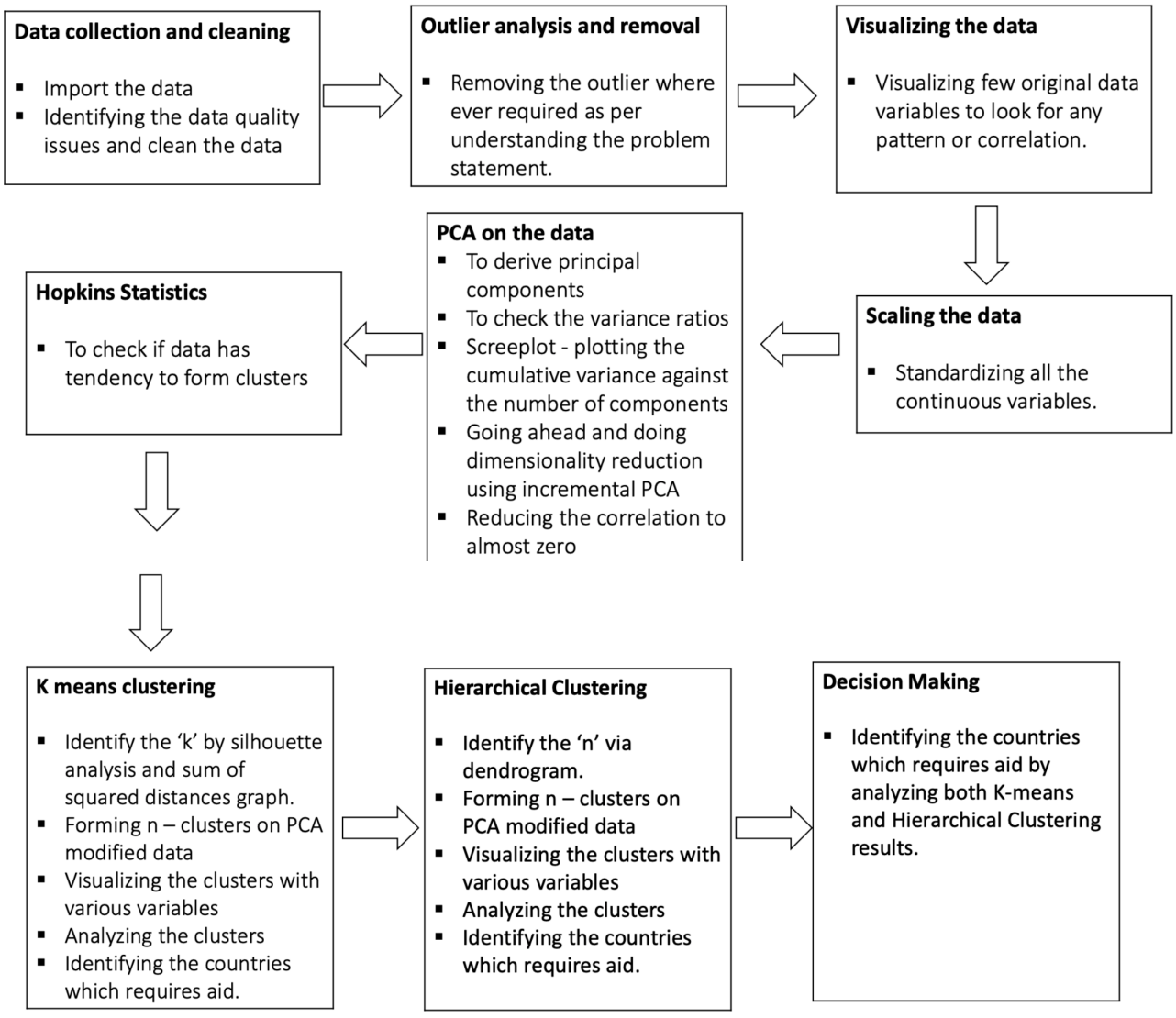
Abstract

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 direst need of aid.



Analysis

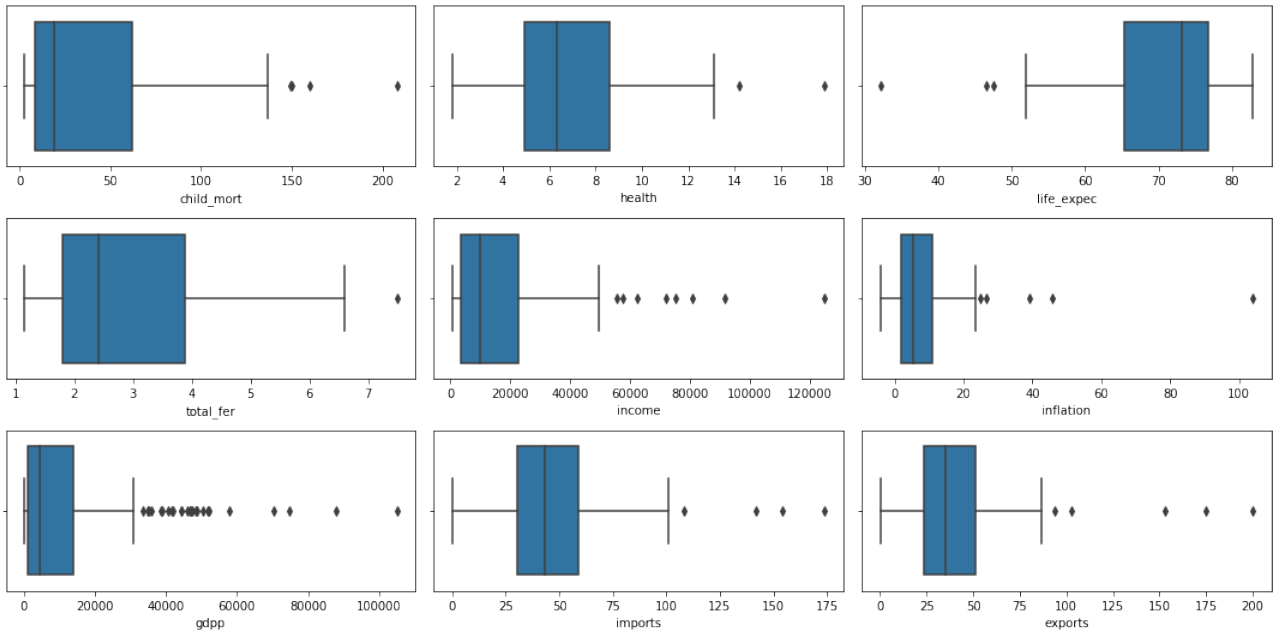
1. Data collection and cleaning

	index	Data_Type	Unique	Nulls	Null_Percent
0	country	object	167	0	0.0
1	child_mort	float64	139	0	0.0
2	exports	float64	147	0	0.0
3	health	float64	147	0	0.0
4	imports	float64	151	0	0.0
5	income	int64	156	0	0.0
6	inflation	float64	156	0	0.0
7	life_expec	float64	127	0	0.0
8	total_fer	float64	138	0	0.0
9	gdpp	int64	157	0	0.0

```
1 country_info.info()
```

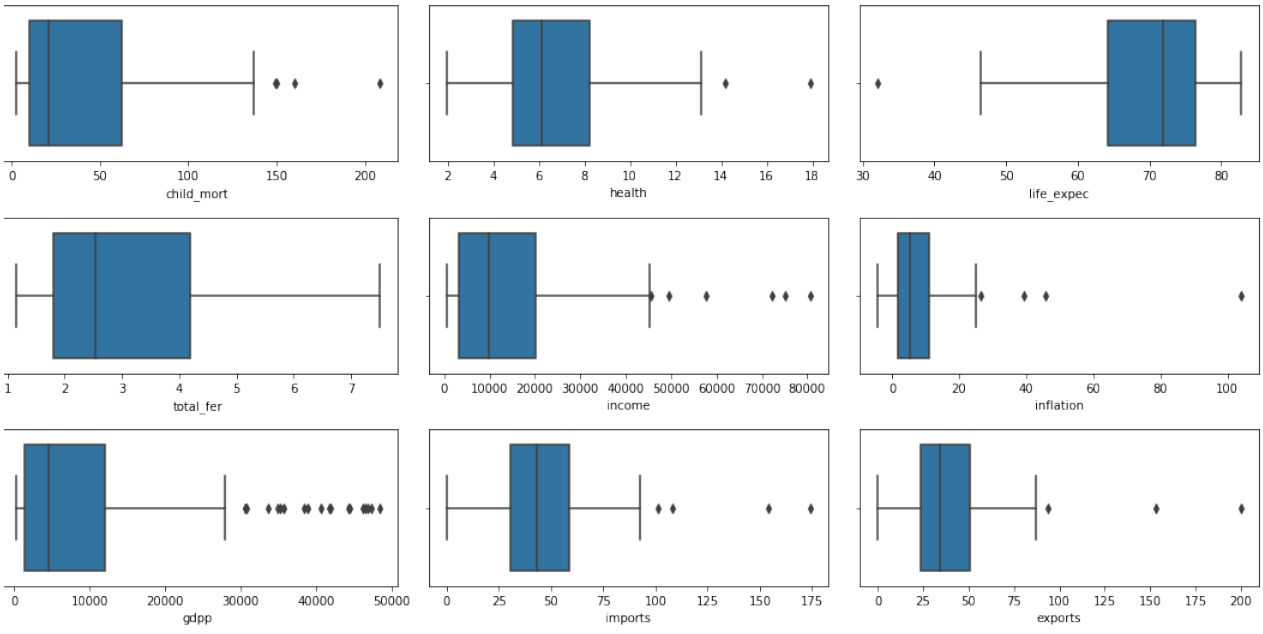
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 167 entries, 0 to 166
Data columns (total 10 columns):
country       167 non-null object
child_mort    167 non-null float64
exports       167 non-null float64
health        167 non-null float64
imports       167 non-null float64
income        167 non-null int64
inflation     167 non-null float64
life_expec    167 non-null float64
total_fer     167 non-null float64
gdpp          167 non-null int64
dtypes: float64(7), int64(2), object(1)
memory usage: 13.1+ KB
```

2.Outlier analysis and removal



We see that gdpp, income and inflation have high outliers.

However let's not remove outliers from inflation as this might lead to loss in country details which are not doing well socio-economically(i.e. countries in direct need of aid).



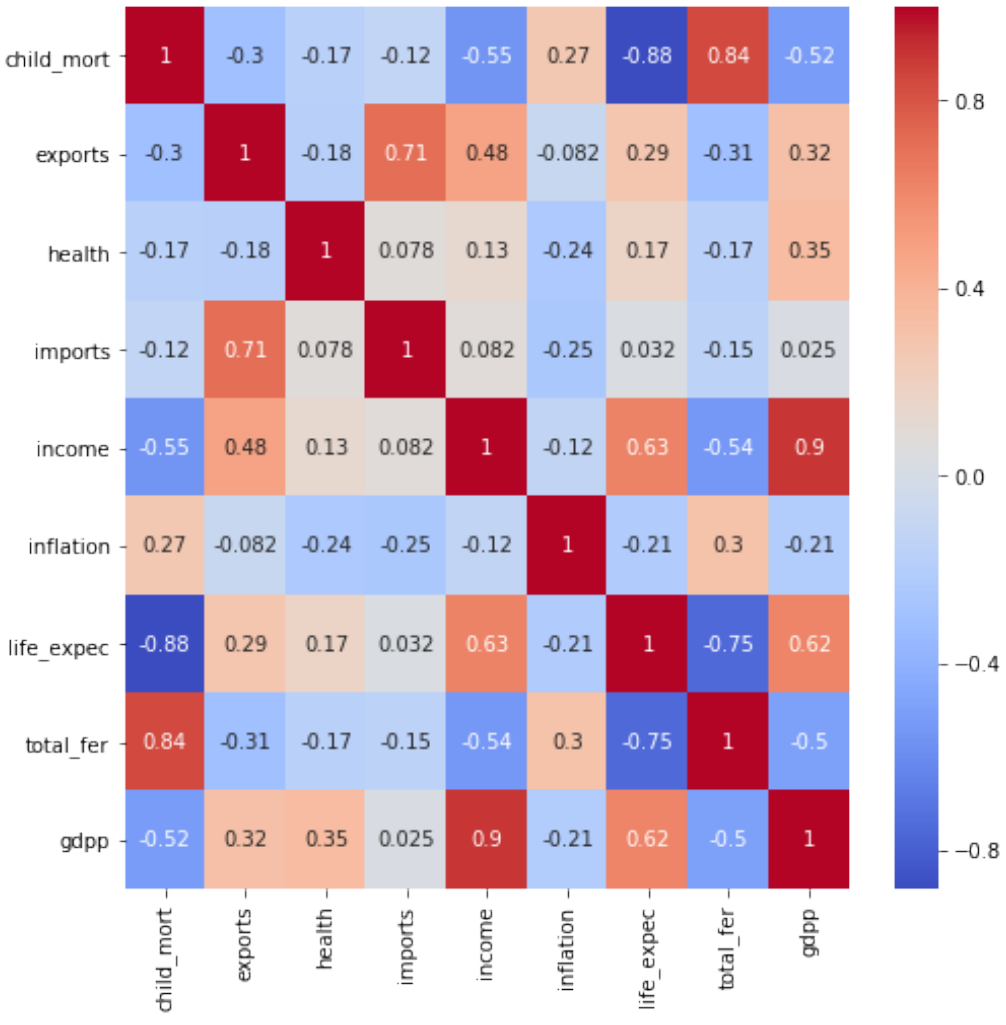
3. Scaling and PCA on data:

	PC1	PC2	Feature
0	-0.421623	-0.032585	child_mort
1	0.232156	-0.599838	exports
2	0.181381	0.191686	health
3	0.094524	-0.729981	imports
4	0.410887	0.110986	income

4.Correlation in the data

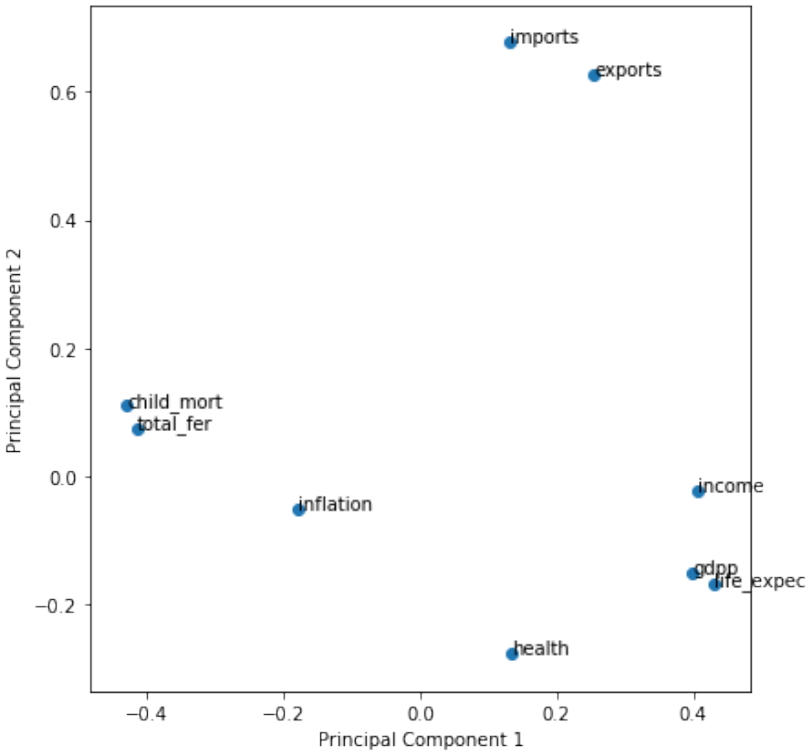
After data cleaning we removed outlier from gdpp column because the country with high gdpp would not require any aid as there are already doing well.
We see high correlation between

- gdpp and income
- total_fer and child_mort
- imports and exports

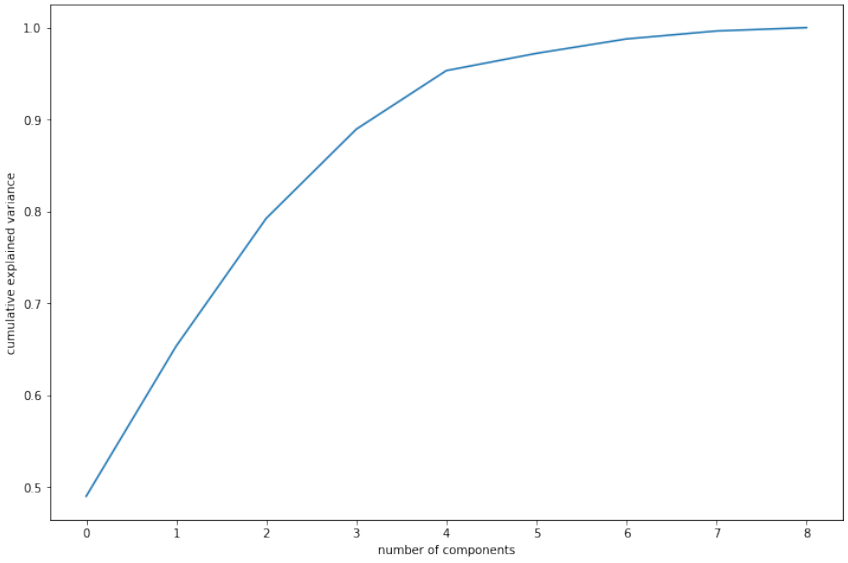


5.Principal Component Analysis (PCA)

PCA on data



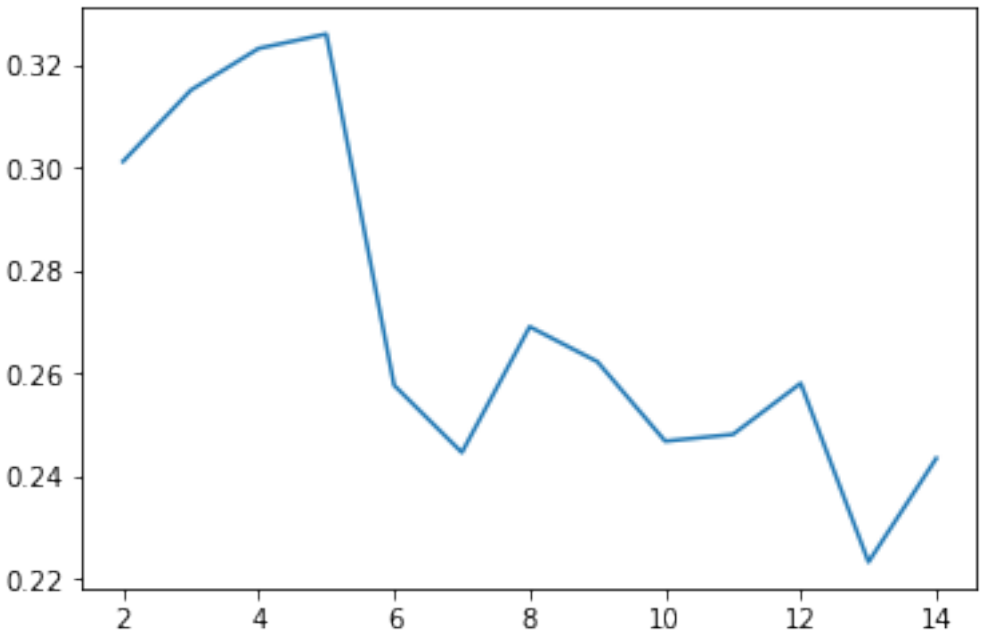
Visualising the features along PC1 and PC2



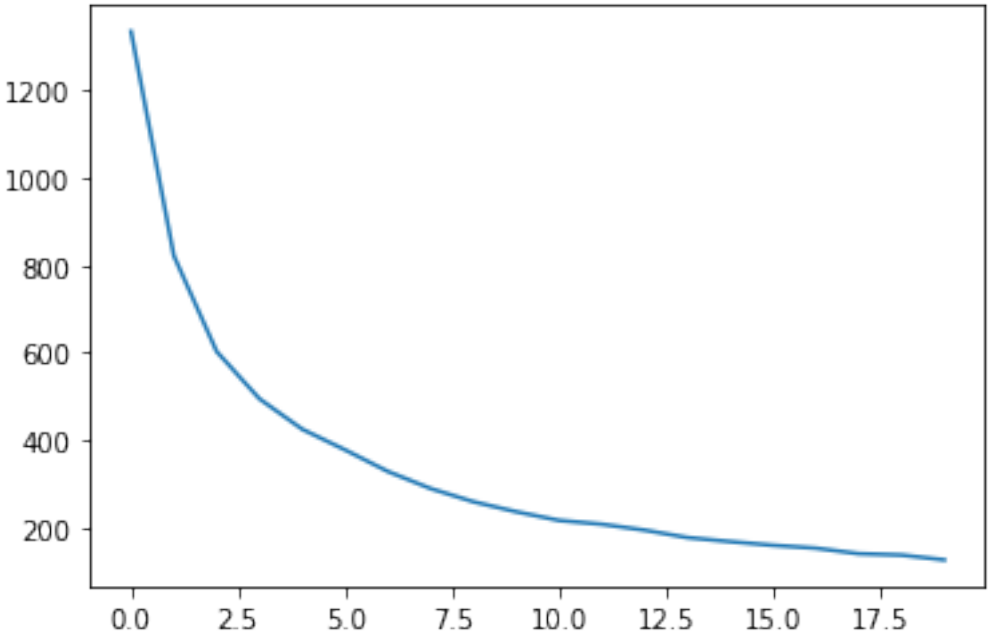
We see that feature like gdpp, life expectancy and income are along the direction of PC1 and other features like total fertility and child mortality are along PC2 direction

6. K-Means Clustering

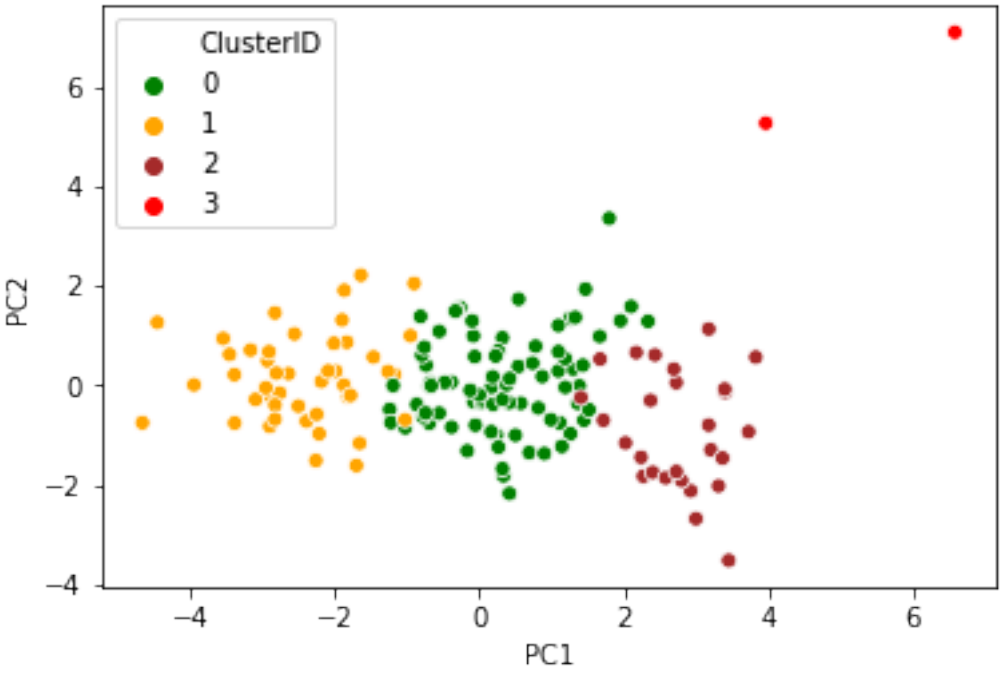
Silhouette analysis:



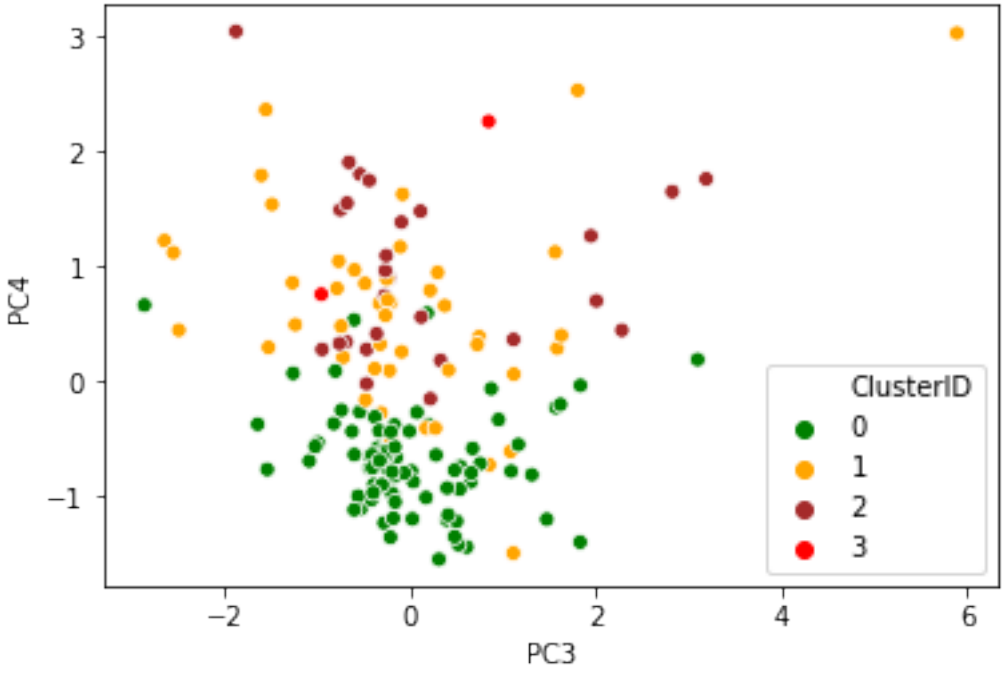
Sum of squared distances:

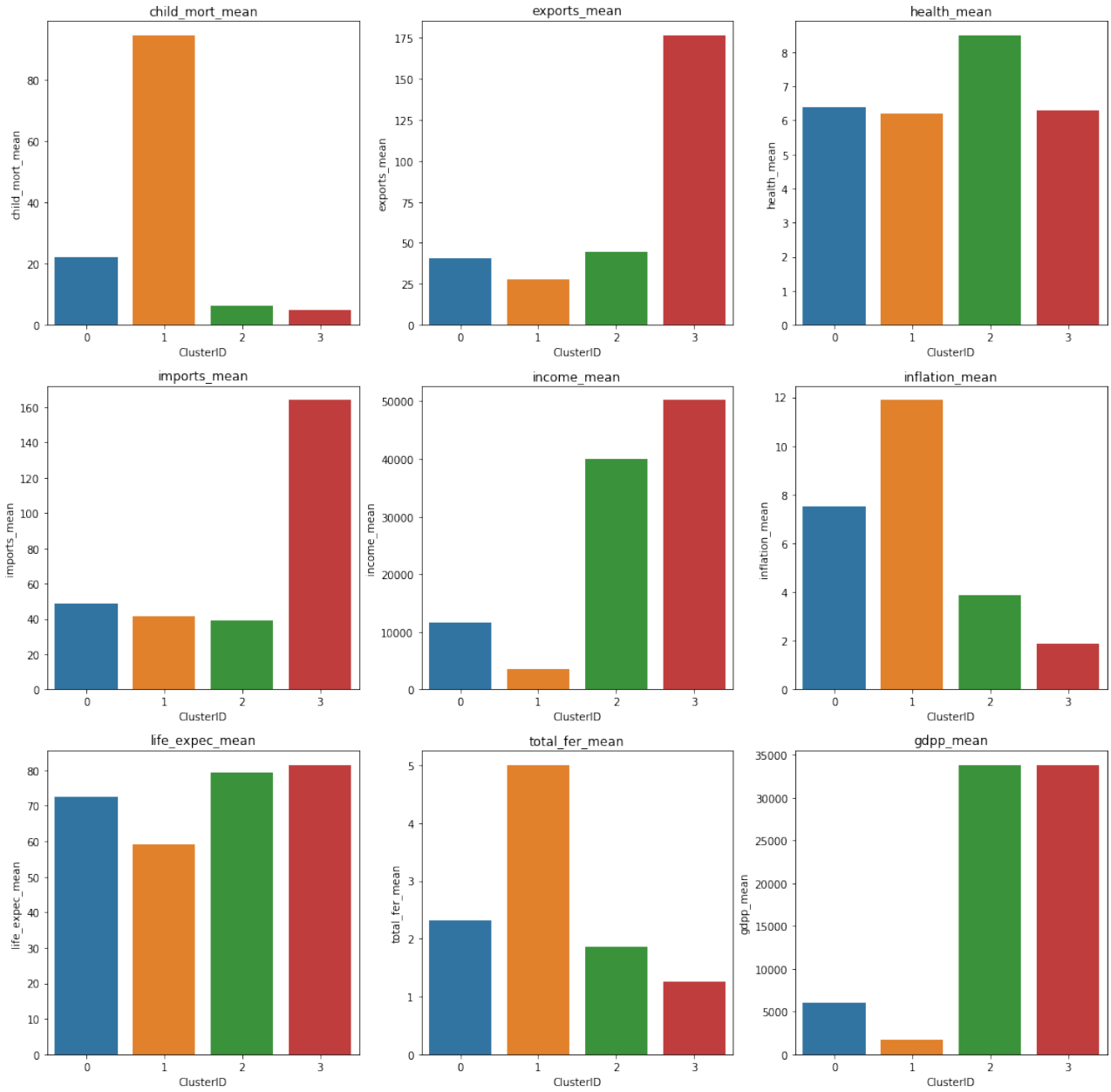


scatter plot for PC1, PC2 and clusterId



scatter plot for PC3, PC4 and clusterId





Countries requiring direct aid according to K- Means Clustering:

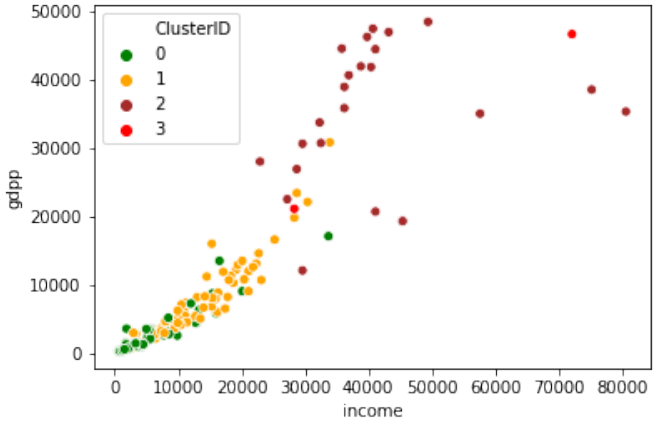
- Burundi
- Liberia
- Congo, Dem. Rep.
- Niger
- Sierra Leone
- Madagascar
- Mozambique
- Central African Republic
- Malawi
- Eritrea
- Togo
- Guinea-Bissau
- Afghanistan
- Gambia
- Rwanda
- Burkina Faso

7.Hierarchical Clustering

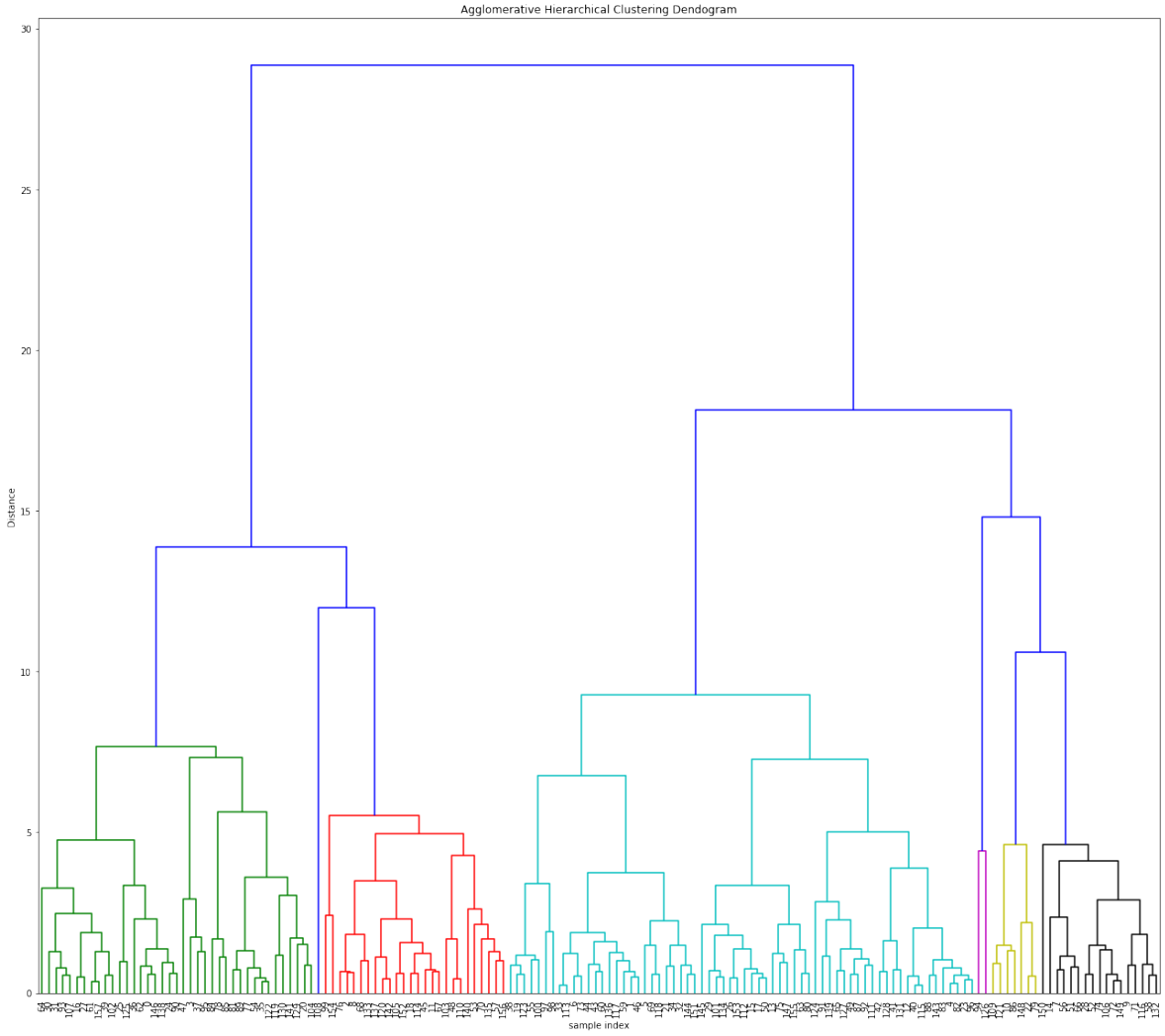
As per our hierarchical clusters:

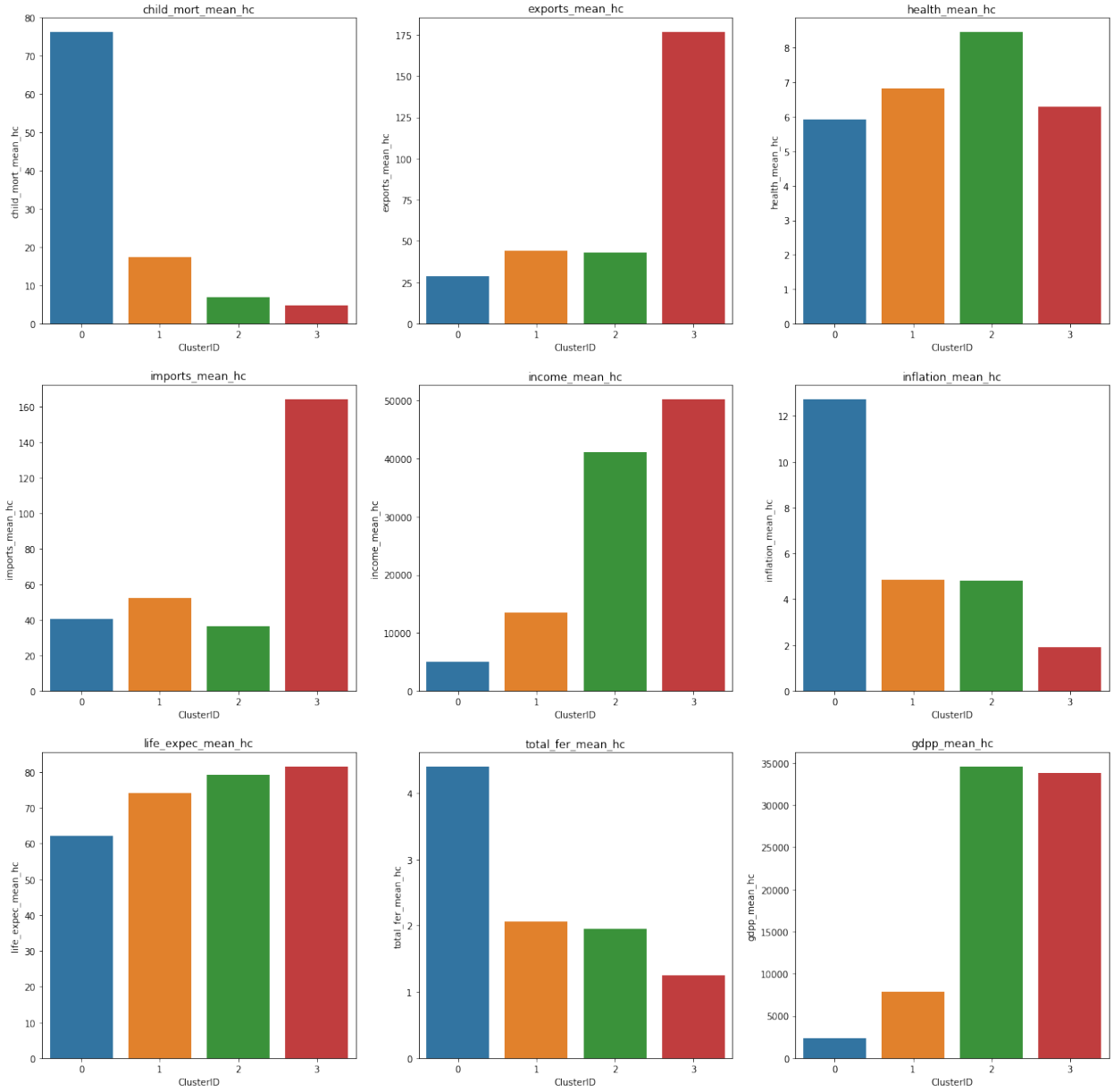
This is area of concern due to:

Low gdpp
Low income
High child mortality
High inflation
High total fertility



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





Considering the business aspect,ignoring cluster 3 as it has just two countries). Upon inspecting the graph we are certain that cluster 0 is our cluster of concern. Because:

- It has highest child mortality
- Lowest income
- Highest Inflation
- Compartively low life expectancy
- Highest total fertility
- Which all in turn leads to lowest gdpp.

Summary:

As both K-means and hierarchical clustering method- we have got some countries which requires aid.

Recommendations

Cluster with ClusterID as 0, is the cluster of most backward country. Countries on which we require to focus more are

'Afghanistan', 'Benin', 'Botswana', 'Burkina Faso', 'Burundi', 'Cameroon', 'Central African Republic', 'Chad', 'Comoros', 'Congo, Dem. Rep.', "Cote d'Ivoire", 'Eritrea', 'Gabon', 'Gambia', 'Ghana', 'Guinea', 'Guinea-Bissau', 'Haiti', 'Iraq', 'Kenya', 'Kiribati', 'Lao', 'Lesotho', 'Liberia', 'Madagascar', 'Malawi', 'Mali', 'Micronesia, Fed. Sts.', 'Mozambique', 'Namibia', 'Niger', 'Nigeria', 'Pakistan', 'Rwanda', 'Senegal', 'Sierra Leone', 'Solomon Islands', 'South Africa', 'Sudan', 'Tajikistan', 'Tanzania', 'Timor-Leste', 'Togo', 'Uganda', 'Yemen', ‘Zambia'

Final Recommendation:

We got same countries by both K-means and Heirarchical Clustering; The following are the countries which are in direct need of aid by considering Socio–Economic Factors:

- Burundi
- Liberia
- Congo, Dem. Rep.
- Niger
- Sierra Leone
- Madagascar
- Mozambique
- Central African Republic
- Malawi
- Eritrea
- Togo
- Guinea-Bissau
- Afghanistan
- Gambia
- Rwanda
- Burkina Faso