CLUSTERING AND PCA

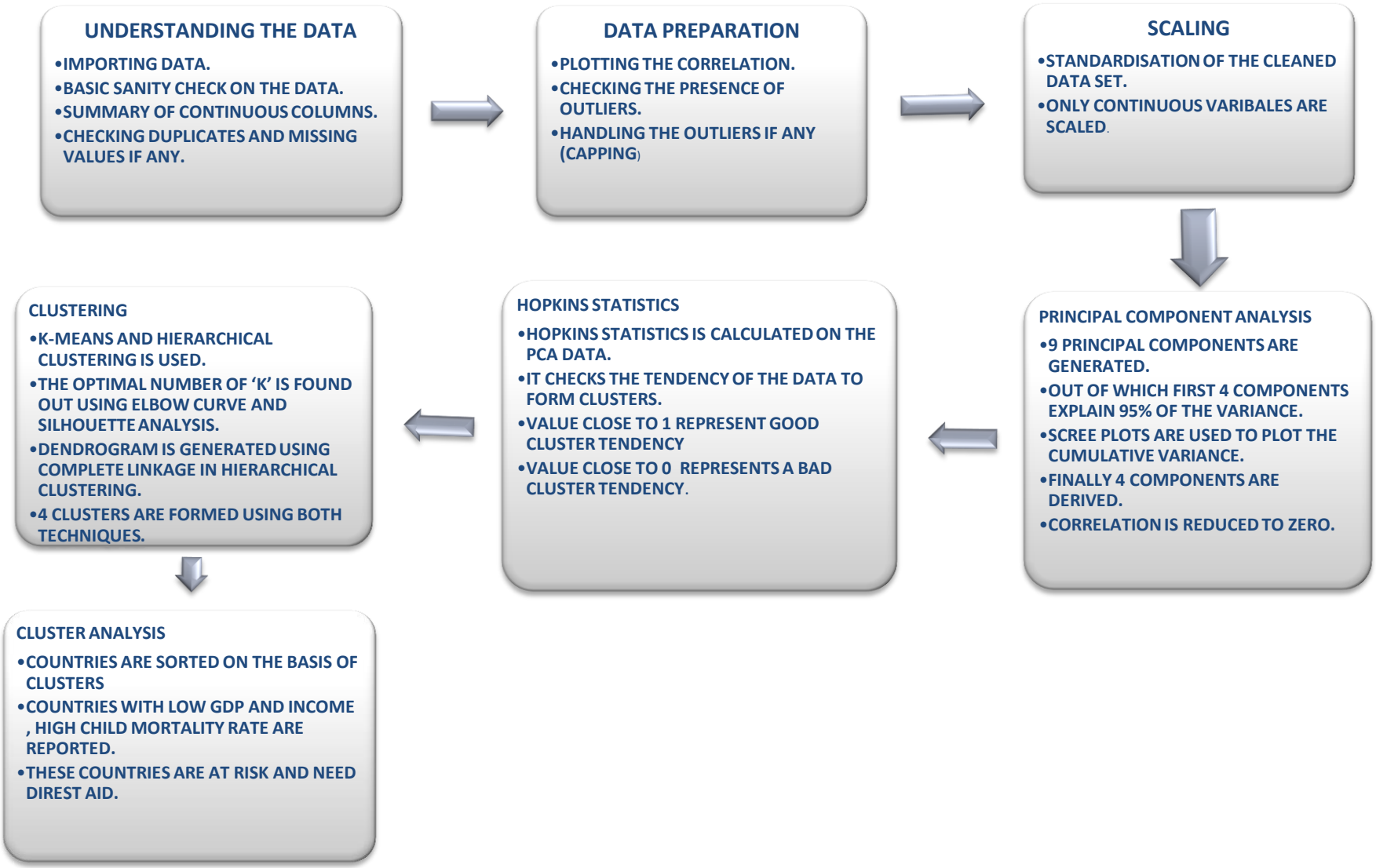
CLUSTERING OF COUNTRIES

- VIDULA AROLKAR

PROBLEM STATEMENT AND OBJECTIVE

- 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.
- We need to categorize the countries using some socio-economic and health factors that determine the overall development of the country.
- Suggest the countries which needs to be focused on the most.

SOLUTION METHODOLOGY



DATA UNDERSTANDING AND CORRELATION

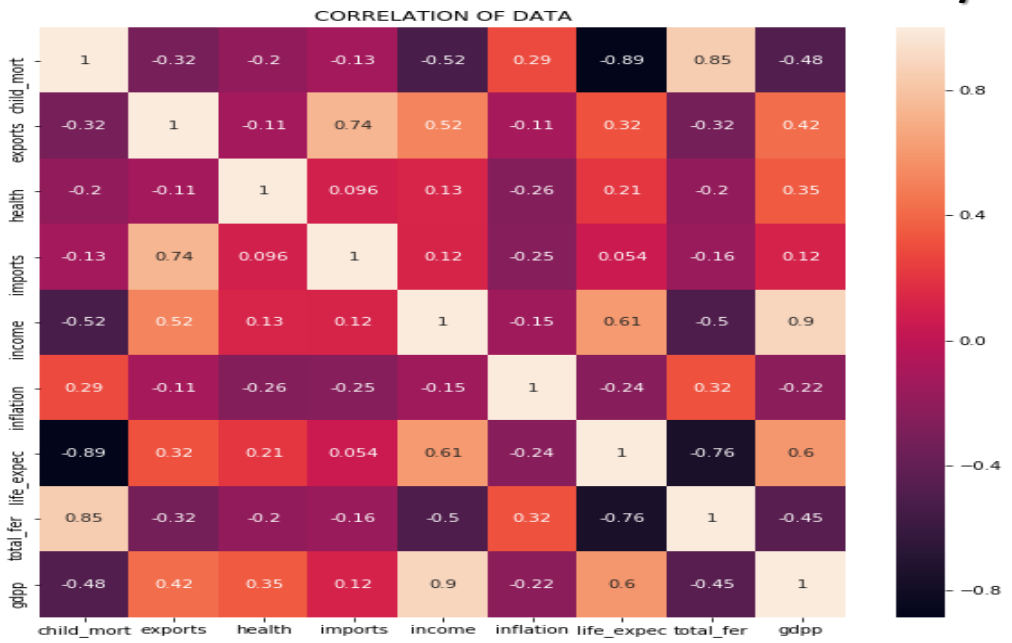
	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	Afghanistan	90.2	10.0	7.58	44.9	1610	9.44	56.2	5.82	553
1	Albania	16.6	28.0	6.55	48.6	9930	4.49	76.3	1.65	4090
2	Algeria	27.3	38.4	4.17	31.4	12900	16.10	76.5	2.89	4460
3	Angola	119.0	62.3	2.85	42.9	5900	22.40	60.1	6.16	3530
4	Antigua and Barbuda	10.3	45.5	6.03	58.9	19100	1.44	76.8	2.13	12200

Following are the columns:

- **Child Mortality**
- **Exports**
- **Health**
- **Imports**
- **Income**
- **Inflation**
- **Life Expectancy**
- **Total Fertility**
- **GDP of the country**

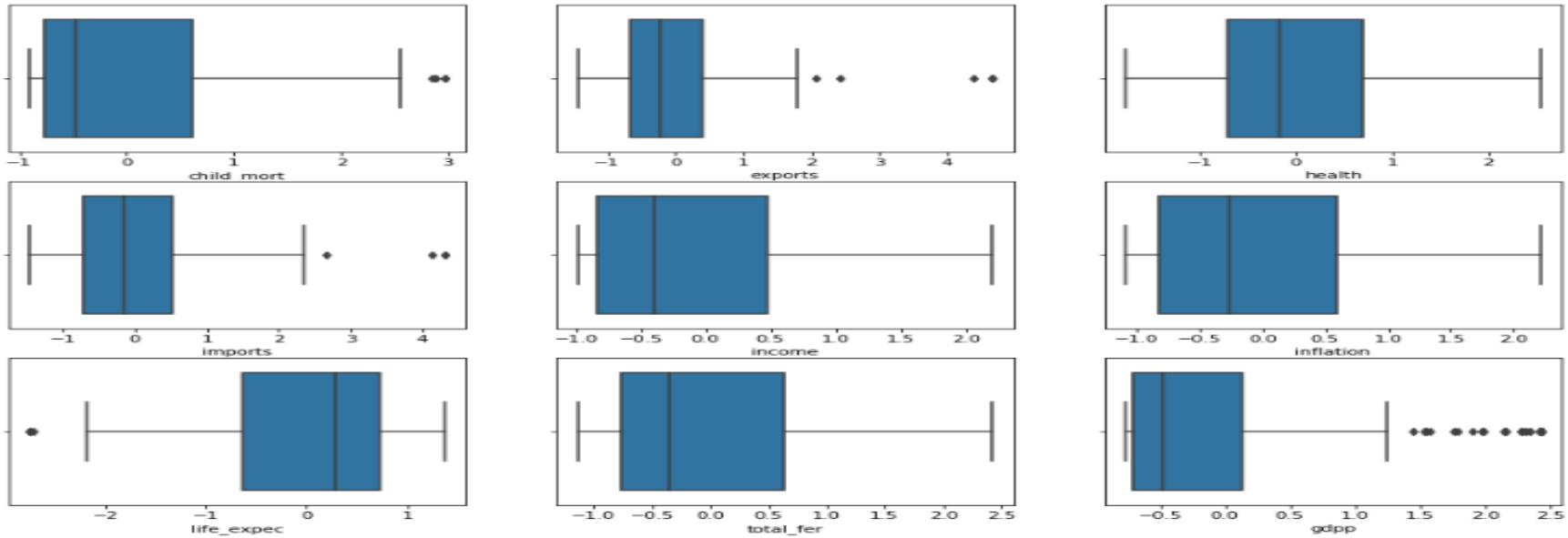
HIGHLY CORRELATED VARIABLES:

- ***Child mort and Total fertility***
- ***Exports and Imports***
- ***GDP and Life expectancy***



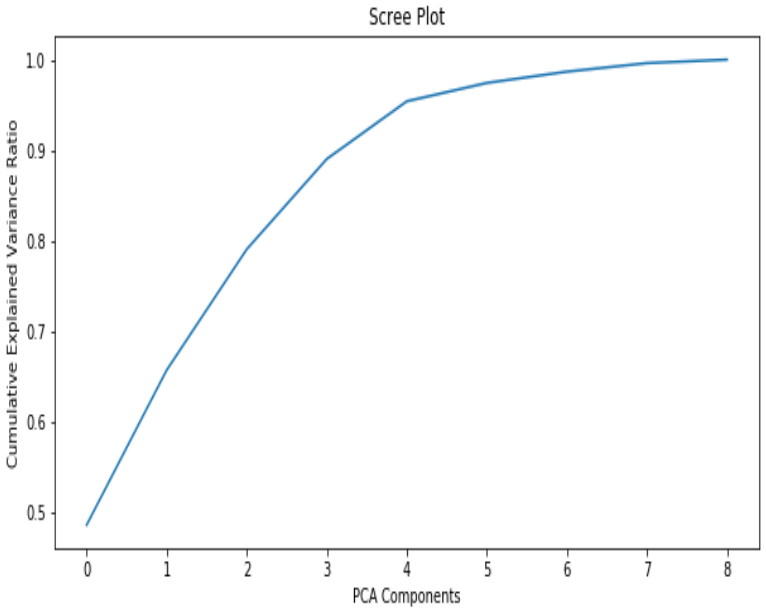
DATA PREPARATION AND VISUALISATION

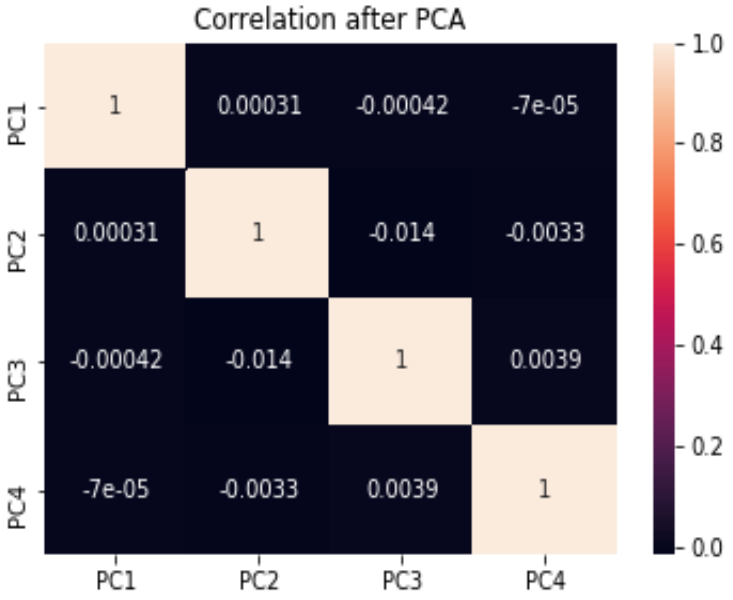
- Missing values were checked and there seems to be no missing values.
- Duplicate data and spelling mistakes are checked.
- Outliers are present more in GDP , income, inflation columns.
- After analysing the data, removing the outliers is not a good option. Hence Capping is preferred.
- There is no large data hence, Soft Capping is preferable. Range for Capping is between 0.01 and 0.99. Only for GDP, Income and Inflation column, the range used is 0.05 and 0.95.



PRINCIPAL COMPONENT ANALYSIS

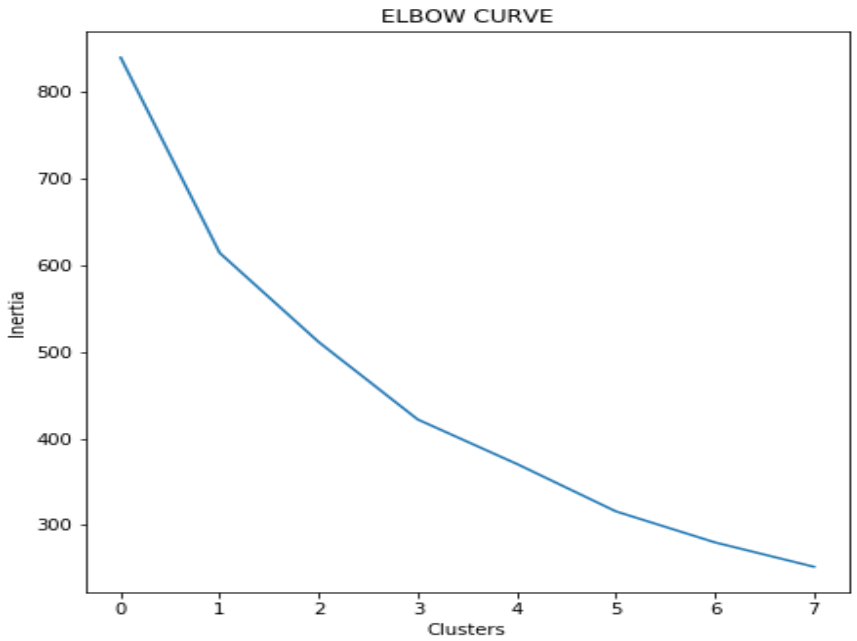
- **SCALING: Before performing PCA on the data set, it need to be scaled. Hence the data is scaled using Standardization. Only the continuous variables are scaled.**
- **PCA: Scree plot is used to determine the Principal Components that are useful.**
- **Scree Plot explained a Variance of almost 95% with first 4 components.**
- **Here, Dimensionality Reduction is achieved by using Incremental PCA with 4 components.**
- **The Correlation obtained after performing PCA is almost 0.**

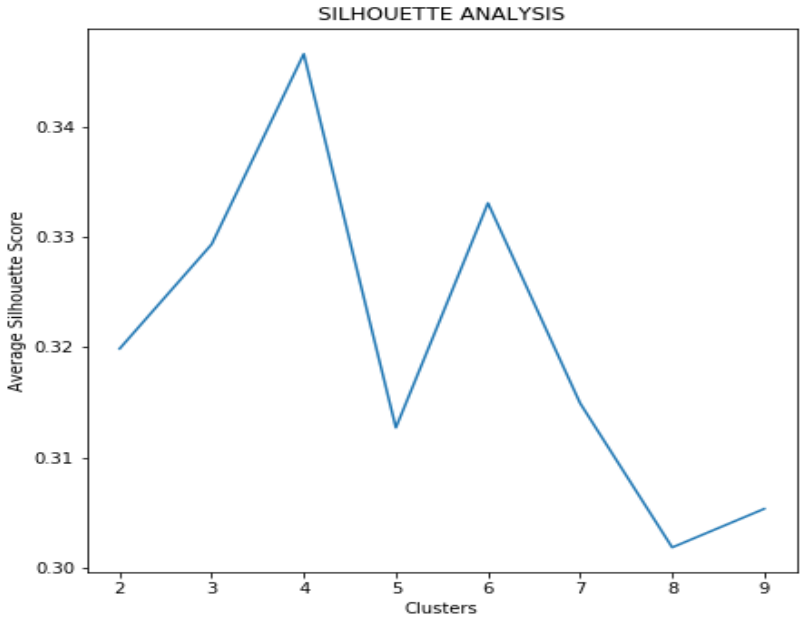




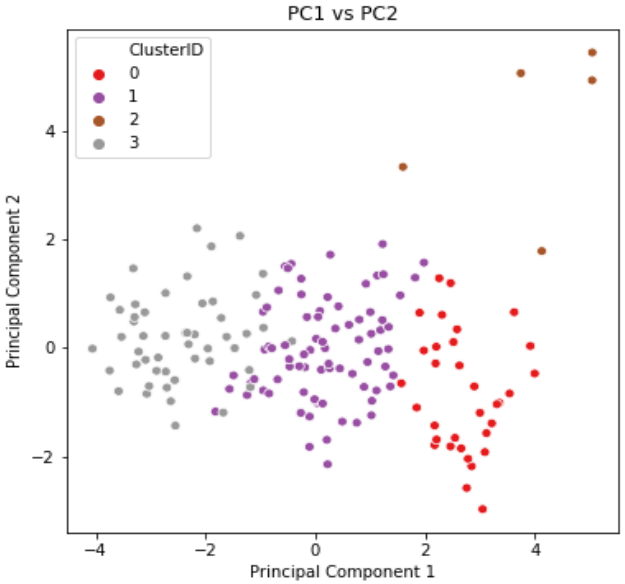
K-MEANS CLUSTERING

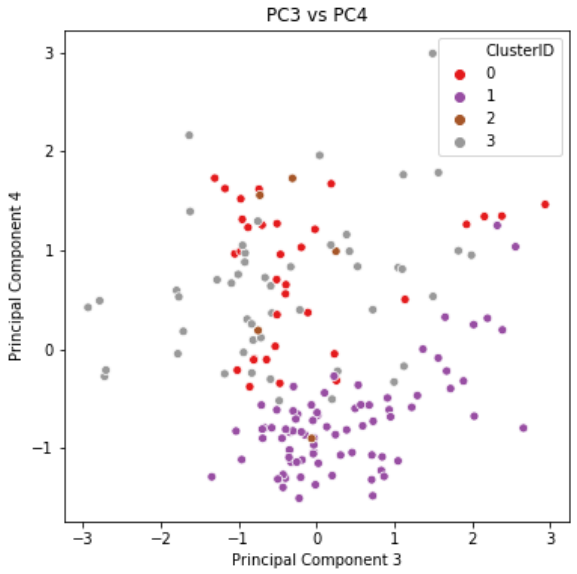
- **HOPKINS STATISTICS: The Hopkins statistics score obtained is close to 1, due to which the data can be considered good for clustering.**
- **K-Means Clustering : It is performed using cluster = 4 for our data.**
- **This optimal number is considered using Elbow Curve (Sum of Squared Distances) and Silhouette Analysis.**
- **The Elbow curve shows that the inertia value is gradually decreasing from cluster 2 to 4.**





ANALYSIS OF COMPONENTS ACROSS THE CLUSTERS:

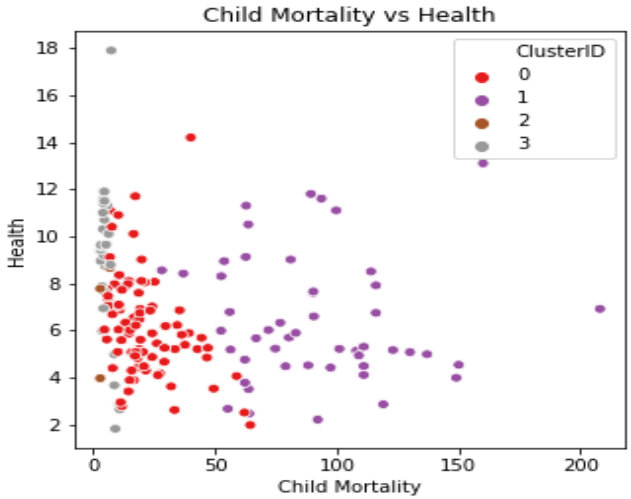


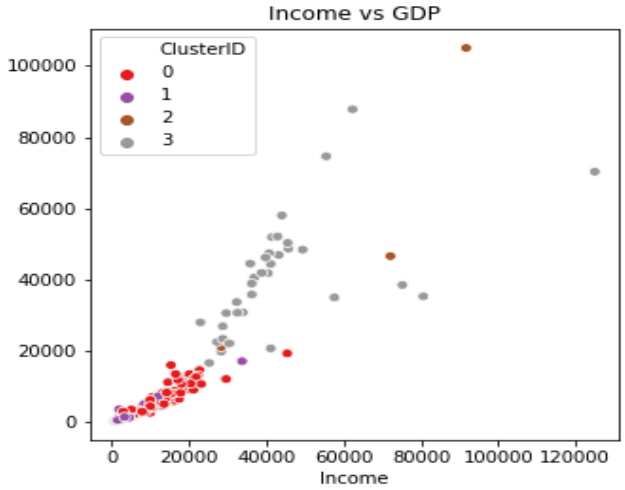


We can see that most of the cluster in PC1 & PC2 are distinct.
PC3 and PC4 also has most of the variance explained.

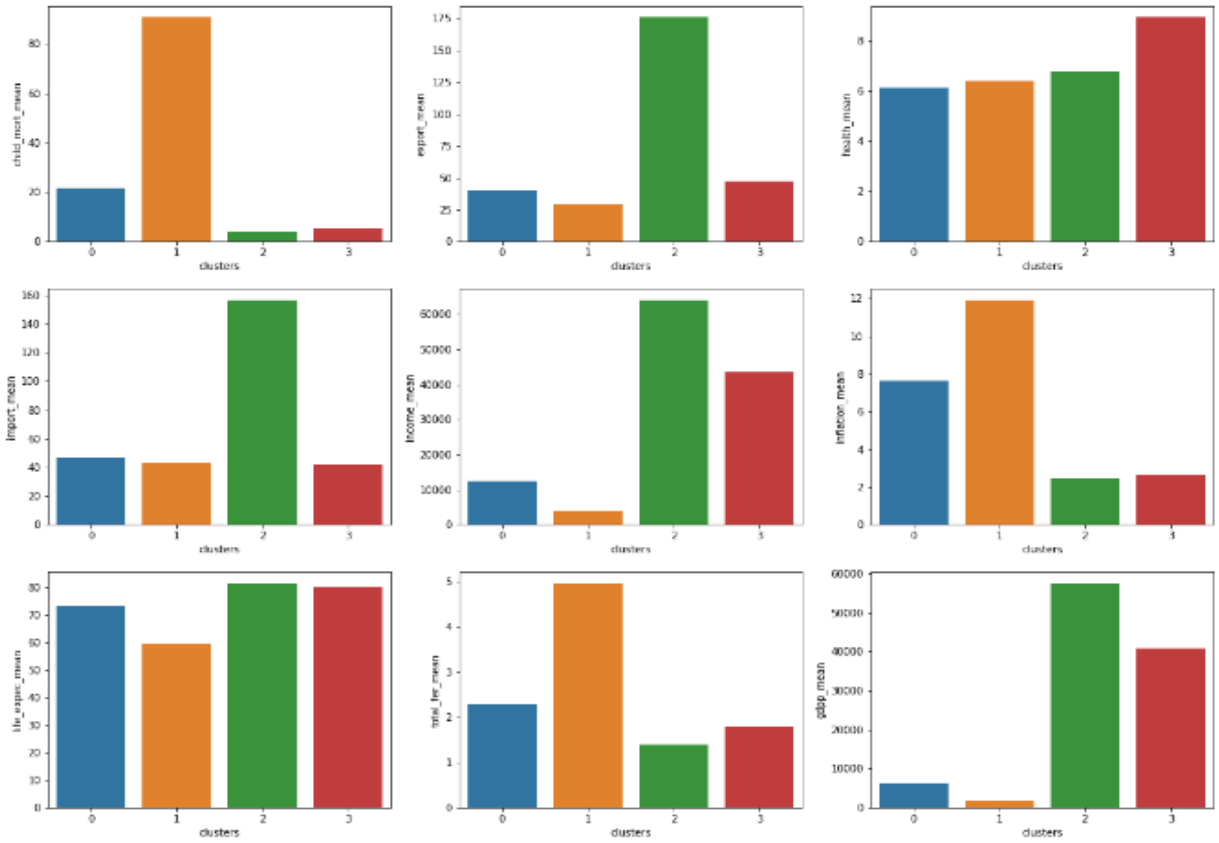
CLUSTER ANALYSIS FOR ACTUAL VARIABLES:

- **The plot shows that the income and GDP for cluster 1 and are very low.**
- **Also the Child mortality for cluster 1 is high.**





FINAL ANALYSIS USING K-MEANS CLUSTERING:



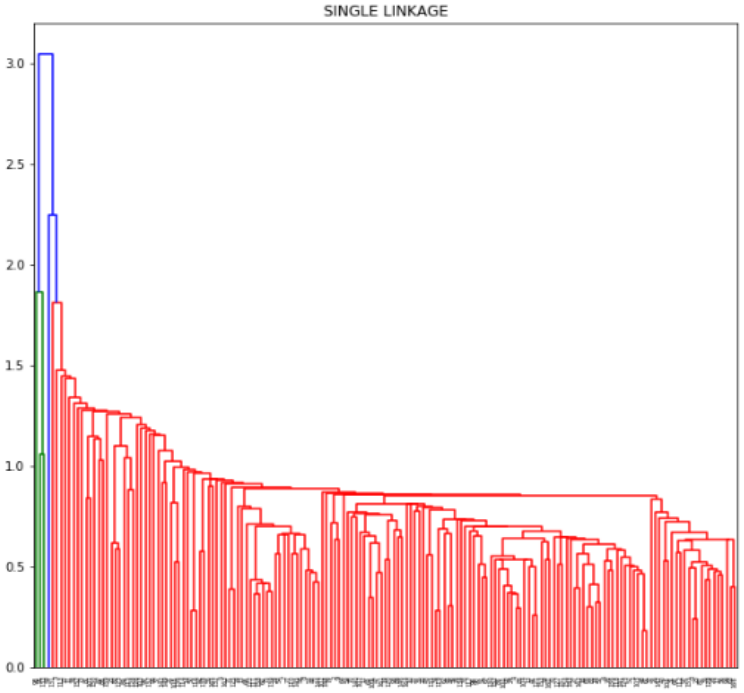
Countries that need Aid:

1. **BURUNDI**
2. **LIBERIA**
3. **CONGO,DEM. REP.**
4. **NIGER**
5. **SIERRA LEONE**

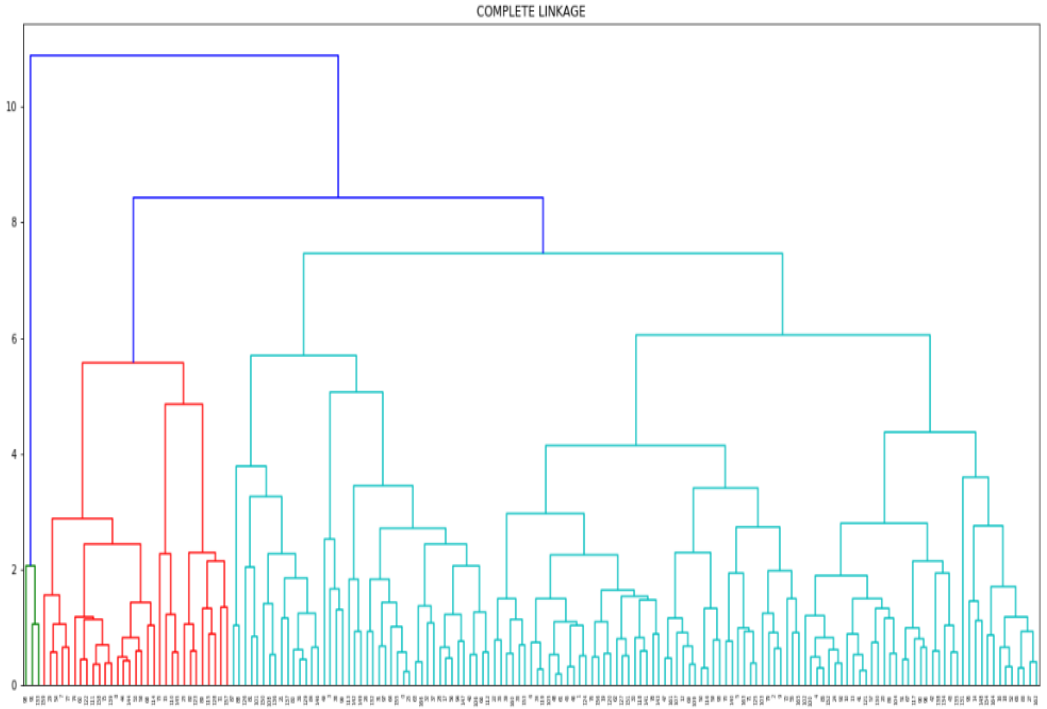
- **From the bar plots , we can conclude that cluster 1 is the cluster of concern.**
- **It has high mortality rate.**
- **Cluster 1 has low Income and GDP.**
- **The total fertility rate is also high for cluster 1.**
- **It also has high life expectancy.**

HIERARCHICAL CLUSTERING

- Single Linkage and Complete Linkage ,both were performed on the PCA dataset.
- It is observed that Single linkage does not provide a clear view of the data . Hence , Complete Linkage is preferred.



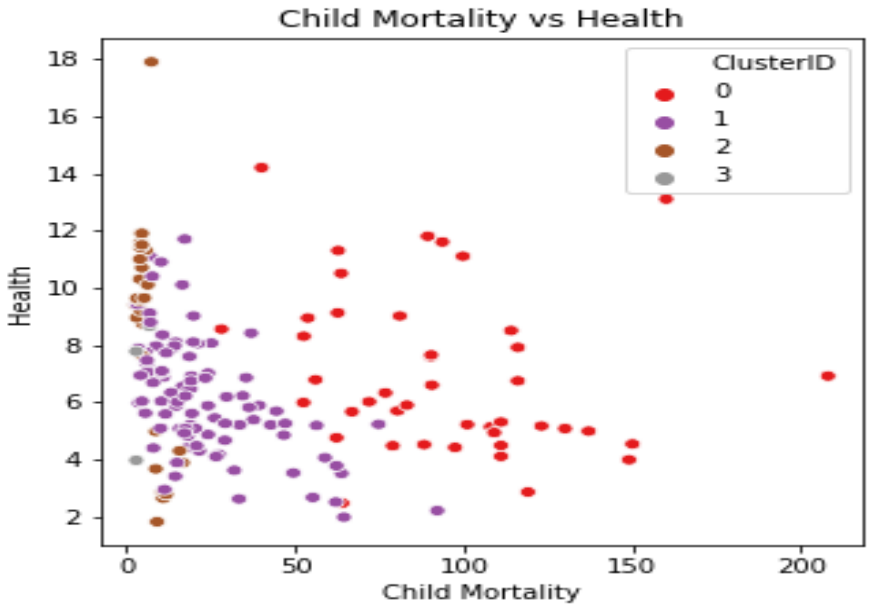
Single Linkage



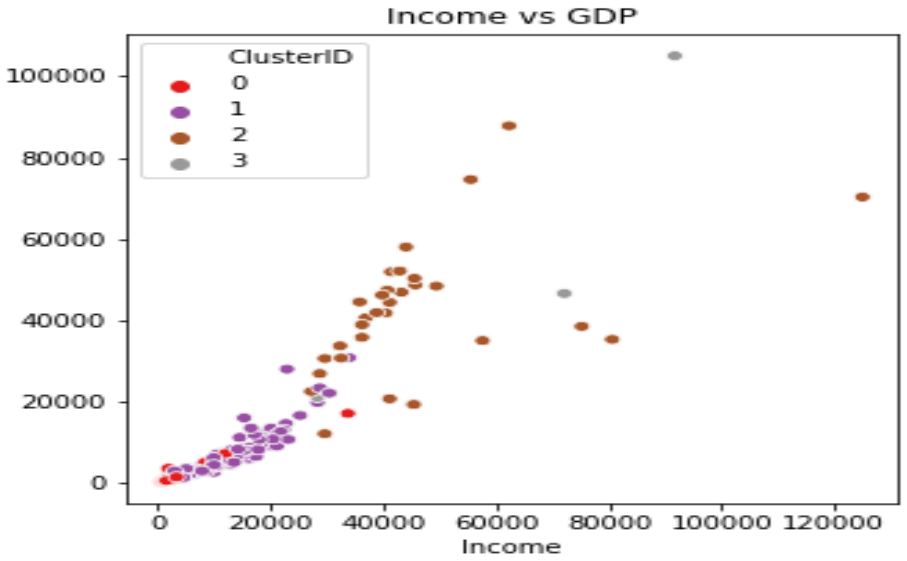
Complete Linkage

- When the Dendrogram is cut vertically at the position 4, we get 4 clusters.
- Hence , the data is grouped into 4 clusters.

CLUSTER ANALYSIS ON THE ACTUAL VARIABLES:

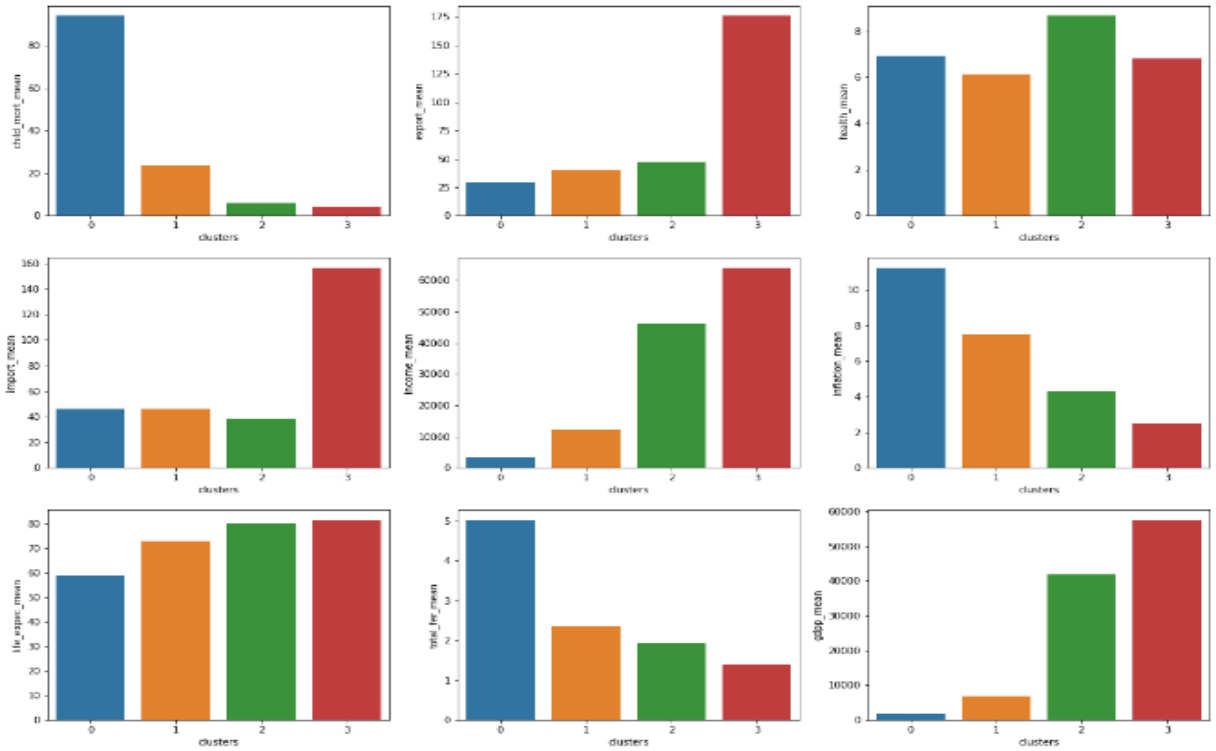


- From the scatterplot , we can see that Child Mortality Rate is high in Cluster 0.
- Health remains stable in all the clusters.



- It is observed that Cluster 0 and 1 has low income .
- GDP rate is also low for Cluster 0.

FINAL ANALYSIS USING HIERARCHICAL CLUSTERING:



Countries that need Aid:

1. **BURUNDI**
2. **LIBERIA**
3. **CONGO,DEM. REP.**
4. **NIGER**
5. **SIERRA LEONE**

- **From the bar plots , we can conclude that cluster 0 is the cluster of concern.**
- **The mortality rate is high in Cluster 0.**
- **Cluster 0 has low Income and GDP.**
- **Average Exports and Imports are low in cluster 0.**
- **It also has high total fertility.**
- **Inflation is high in Cluster 0.**

CONCLUSION

- **From both the Clustering techniques, same Countries are derived.**

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
26	Burundi	93.6	8.92	11.60	39.2	764	12.30	57.7	6.26	231
88	Liberia	89.3	19.10	11.80	92.6	700	5.47	60.8	5.02	327
37	Congo, Dem. Rep.	116.0	41.10	7.91	49.6	609	20.80	57.5	6.54	334
112	Niger	123.0	22.20	5.16	49.1	814	2.55	58.8	7.49	348
132	Sierra Leone	160.0	16.80	13.10	34.5	1220	17.20	55.0	5.20	399

- **Hence these are the top 5 countries that need direst aid.**
- **After considering the 3 socio-economic factors , Child Mortality rate, Income and GDP, these countries has poor performance.**
- **Hence essential help should be provided to such countries to maintain the fertility rate, which thus will stable the Life expectancy.**
- **Also Income in these countries is low, so one can provide any additional health insurance to persons in these countries.**
- **The import and exports are also low , which is resulting into high inflation and thus GDP of those countries is falling.**
- **Providing funds for insurances to such countries might increase the health factor and thereby reduce the child mortality rate.**