HELP NGO CASE STUDY SUBMISSION

Name : Ranip Hore

Cohort : 30th Sep , 2018

INTRODUCTION

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 project that included a lot of awareness drives and funding programmes, 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.
- And this is where you come in as a data analyst. Your job is to categorise the countries using some socio-economic and health factors that determine the overall development of the country. Then you need to suggest the countries which the CEO needs to focus on the most.

Objective :

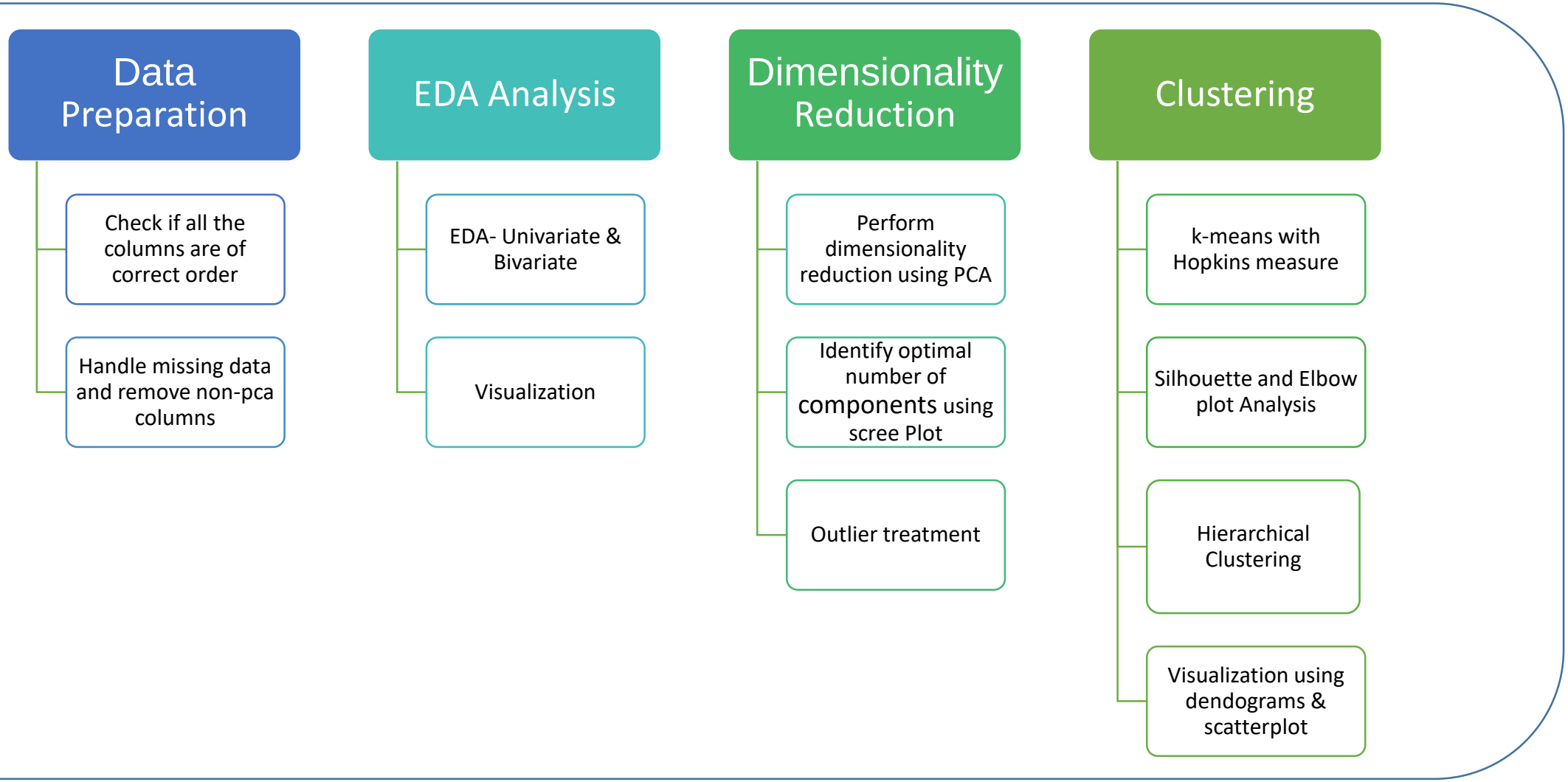
- The main objective is to form clusters of the given countries and analyse based on socio-economic factors to present the recommendations to the CEO so that decision can be taken for which countries that needs immediate attention.

DATA UNDERSTANDING

The dataset comprises of 167 rows and 10 columns. Below are the information for each columns :

- **country** : Name of the country
- **child_mort** : Death of children under 5 years of age per 1000 live births
- **exports** : Exports of goods and services. Given as %age of the Total GDP
- **health** : Total health spending as %age of Total GDP
- **imports** : Imports of goods and services. Given as %age of the Total GDP
- **income** : Net income per person
- **inflation** : The measurement of the annual growth rate of the Total GDP
- **life_expec** : The average number of years a new born child would live if the current mortality patterns are to remain the same
- **total_fer** : The number of children that would be born to each woman if the current age-fertility rates remain the same.
- **gdpp** : The GDP per capita. Calculated as the Total GDP divided by the total population.

PROBLEM SOLVING METHODOLOGY



DATA CLEANSING AND MANIPULATION

- The dataset comprised of 167 rows and 10 columns.
- We checked for NaN columns and found none of the rows having NaN values
- Few columns like Health, exports and imports were given as percentage of Total GDP. Since we don't have neither total gdp nor total population provided in the dataset and instead have gdp per capita(gdpp), we converted the percentage values to absolute values per capita by multiplying the column value with gdpp and dividing it by 100.

Snapshot of dataset before absolute conversion :

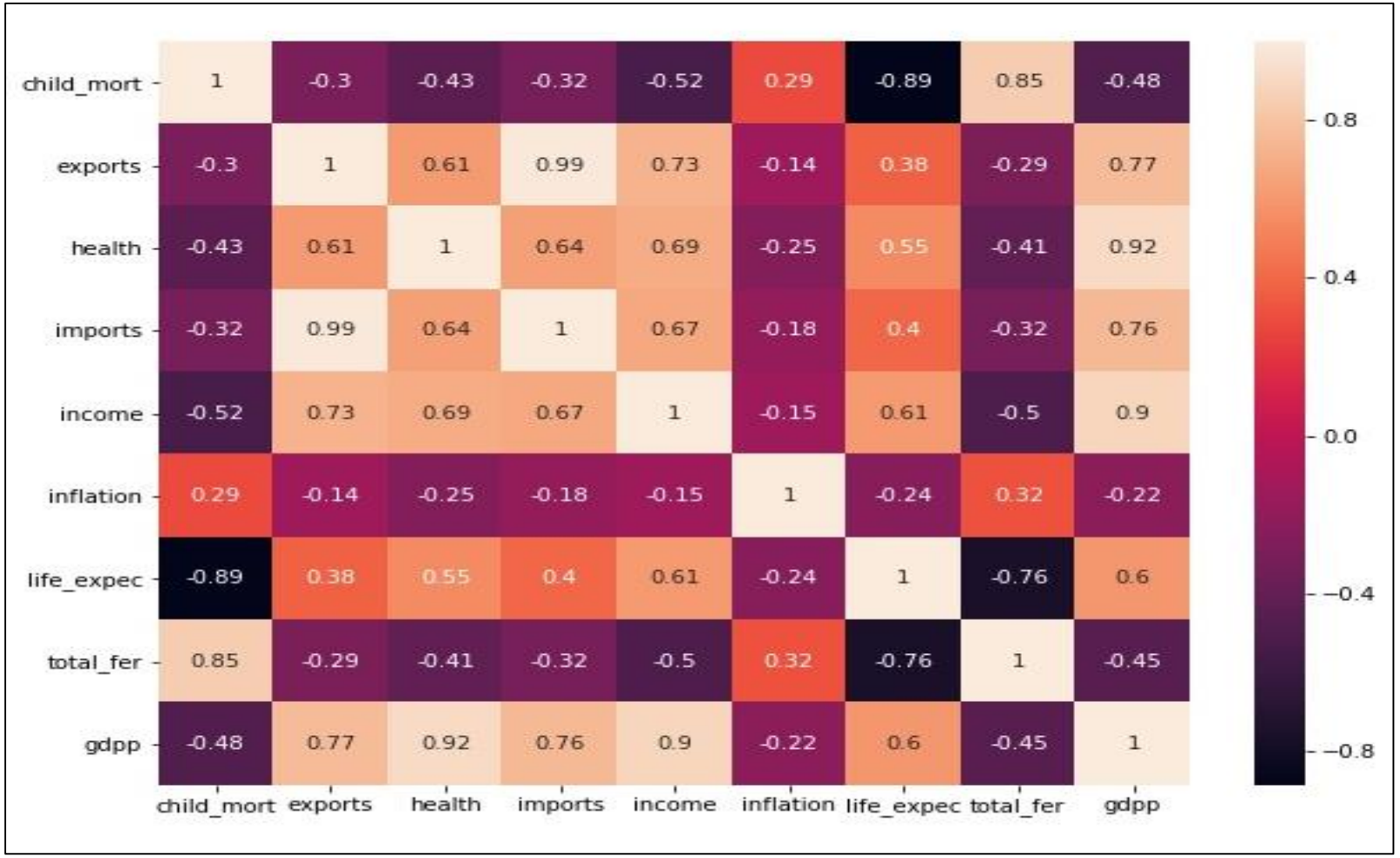
	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	Afghanistan	90.2	10.0	7.58	44.9	1610	9.44	56.2	5.82	553
1	Albania	16.6	28.0	6.55	48.6	9930	4.49	76.3	1.65	4090
2	Algeria	27.3	38.4	4.17	31.4	12900	16.10	76.5	2.89	4460
3	Angola	119.0	62.3	2.85	42.9	5900	22.40	60.1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

- **Snapshot of dataset after absolute conversion :**

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	Afghanistan	90.2	55.30	41.9174	248.297	1610	9.44	56.2	5.82	553
1	Albania	16.6	1145.20	267.8950	1987.740	9930	4.49	76.3	1.65	4090
2	Algeria	27.3	1712.64	185.9820	1400.440	12900	16.10	76.5	2.89	4460
3	Angola	119.0	2199.19	100.6050	1514.370	5900	22.40	60.1	6.16	3530
4	Antigua and Barbuda	10.3	5551.00	735.6600	7185.800	19100	1.44	76.8	2.13	12200

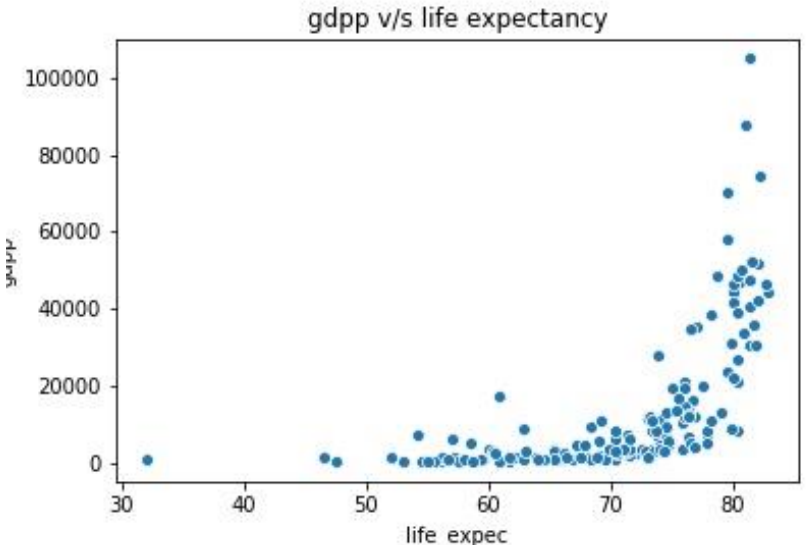
EXPLORATORY DATA ANALYSIS

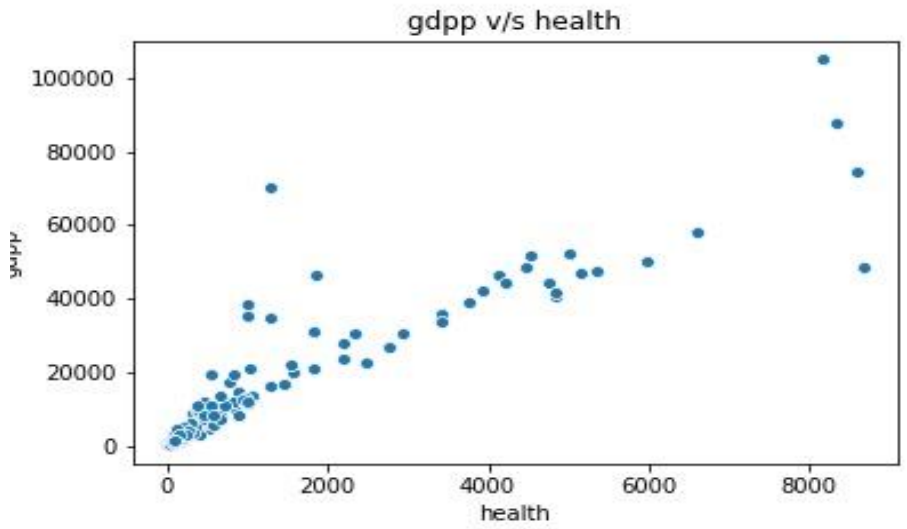
CORRELATION AND MULTICOLLINEARITY CHECK:

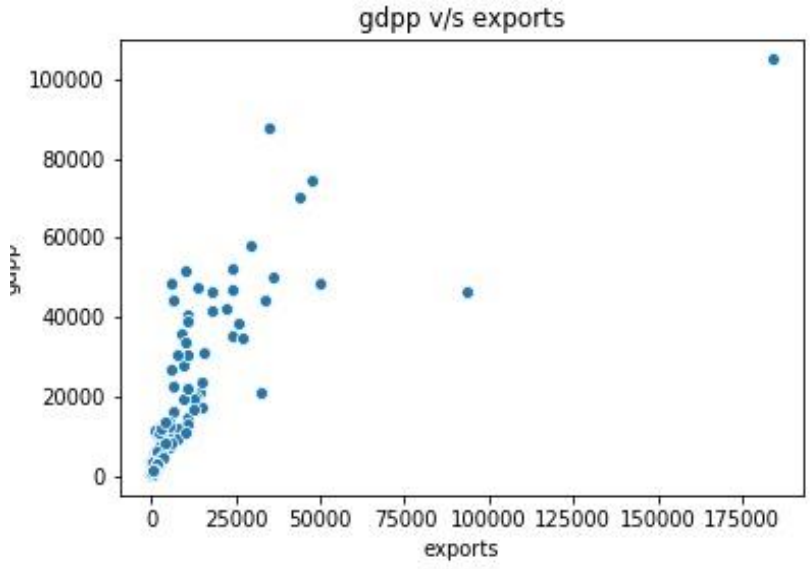


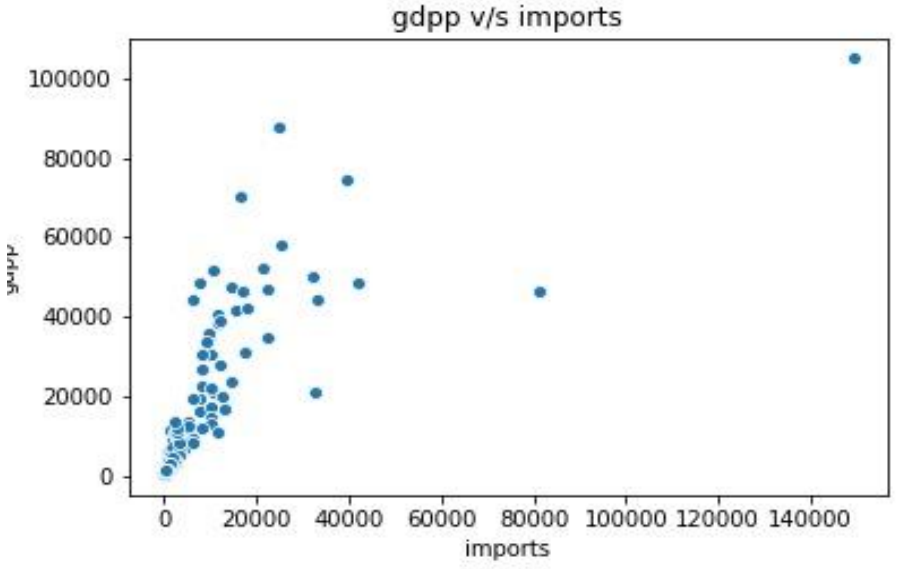
Observation : From the heatmap it can be observed that there is high positive correlation between gdpp and exports, imports, income and life expectancy. Also there is string correlation between income and imports, exports and imports etc.

EXPLORATORY DATA ANALYSIS[Contd.]





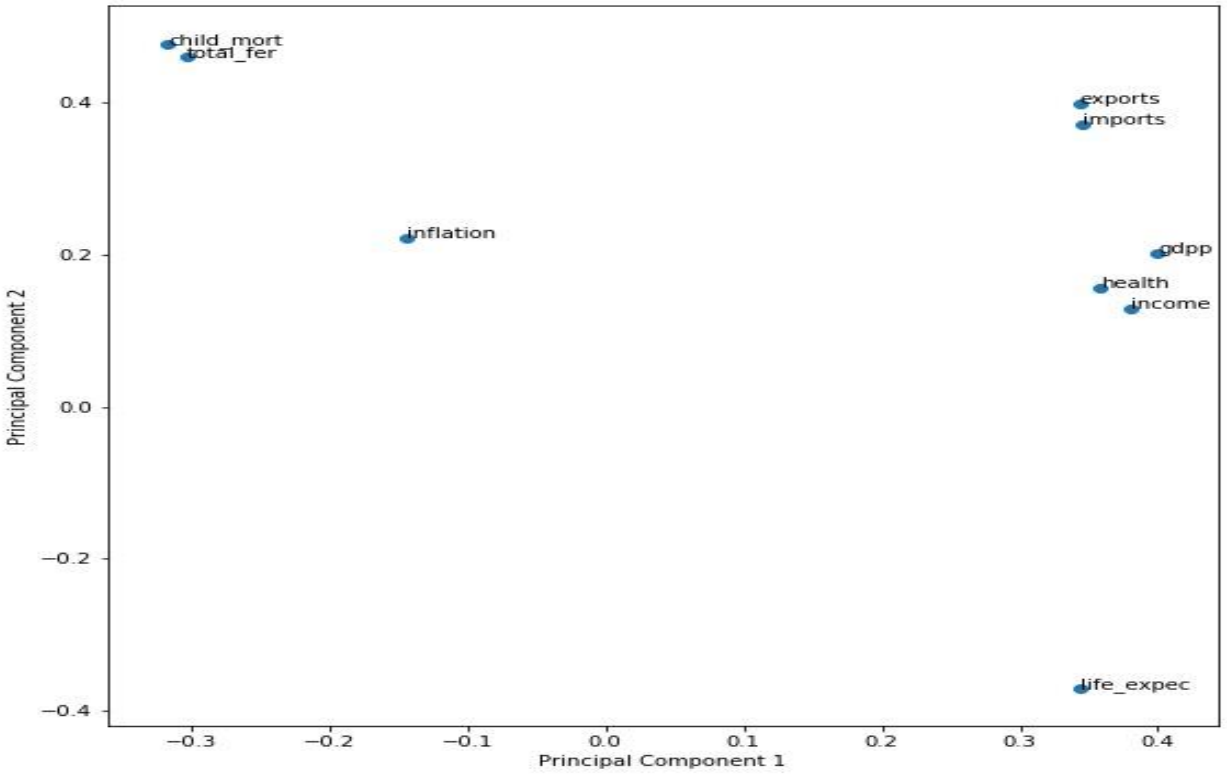


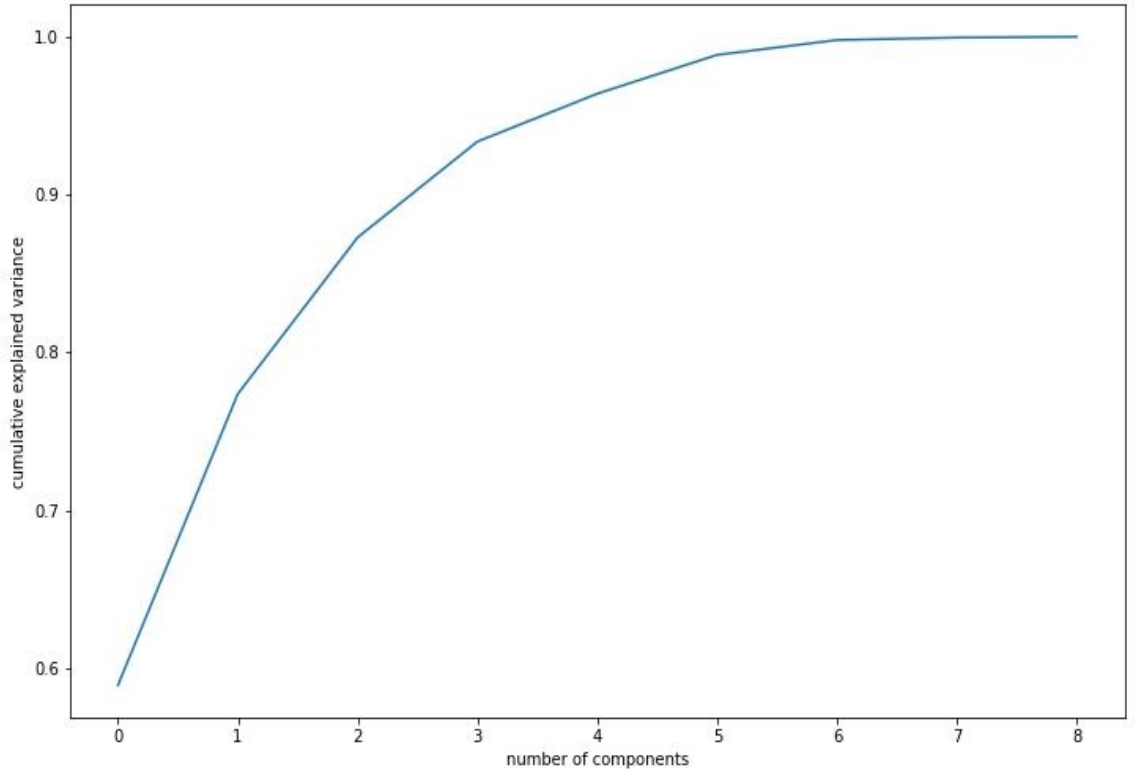


Observation : As deduced from the heatmap, we could see that the gdpp has strong correlation with exports and imports. Life expectancy is also better for countries with high gdp.

Dimensionality Reduction Using PCA

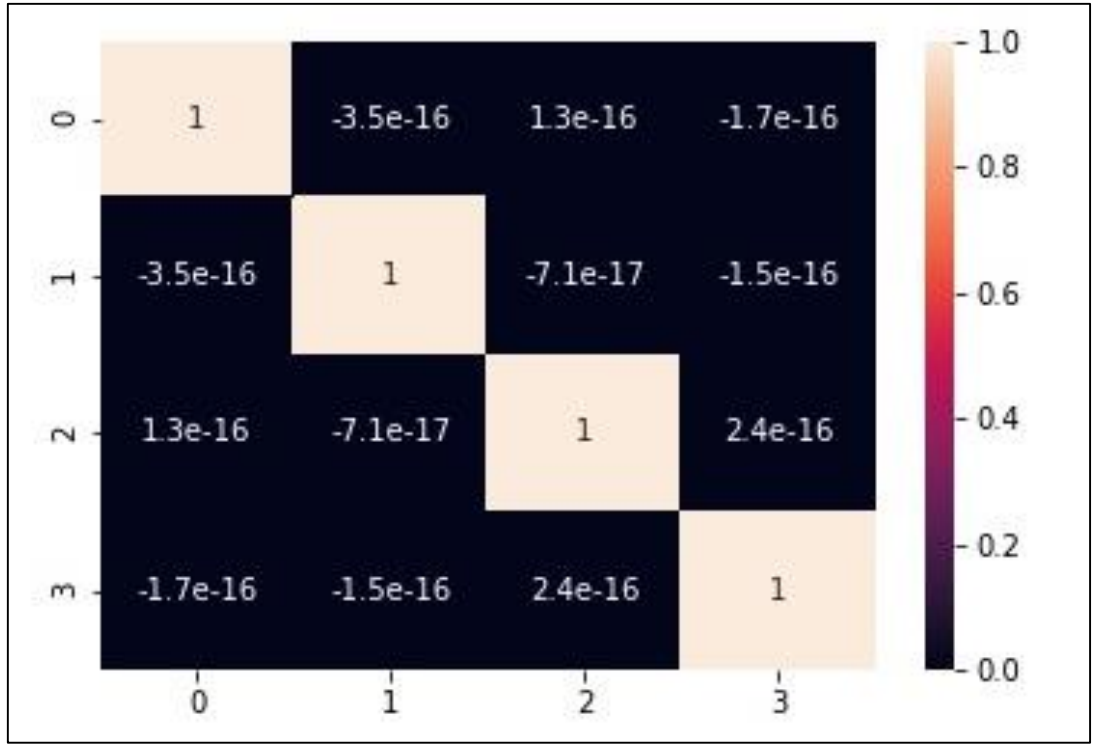
- PCA(Principal Component Analysis) has been performed on the dataset to reduce the number of features.
- Before applying PCA, the categorical column **country** has been removed from the dataset.
- The data is scaled using sklearn StandardScaler() to normalise it.
- We have then performed **fit_transform** and identified optimal number of components from the scree plot.

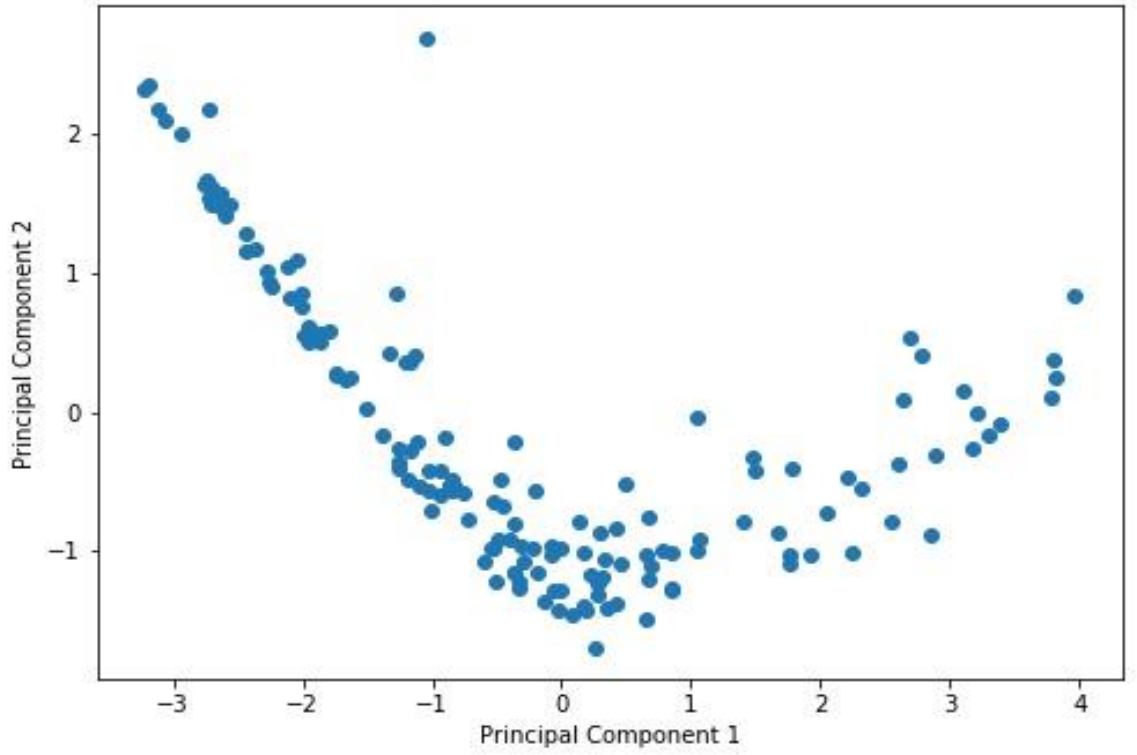




PCA[Contd.]

- From the scree-plot it can be seen that 4 components are explaining about 93% of the variance. So we proceeded with 4 principal components.
- There is no correlation between the four components as evident from the below plot.
- Also we have plotted PC1 and PC2 for 2D visualization.



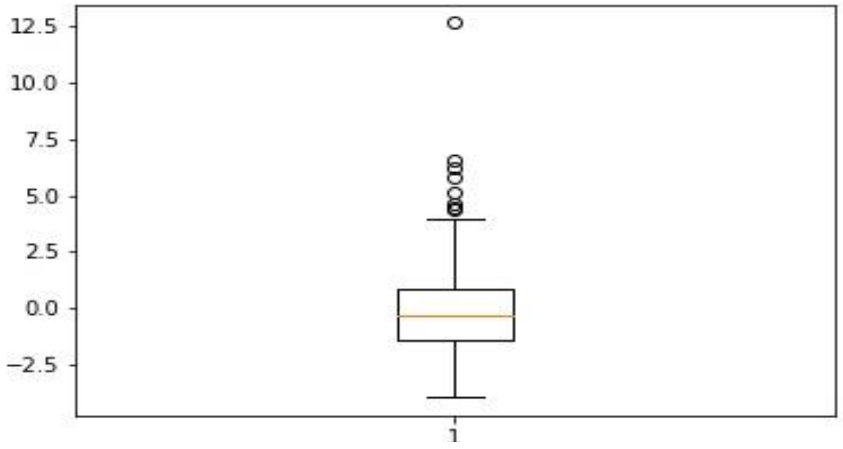


PCA[Contd.]

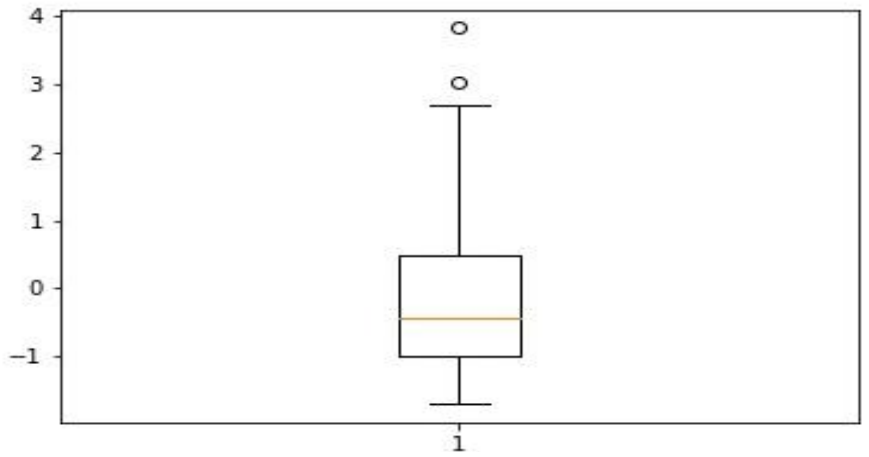
Outlier Removal :

Outliers were present in the original dataset, which needs to be removed now before going for clustering. We had removed the outliers from the four principal components by IQR method and percentile range as 0.25 and 0.75.

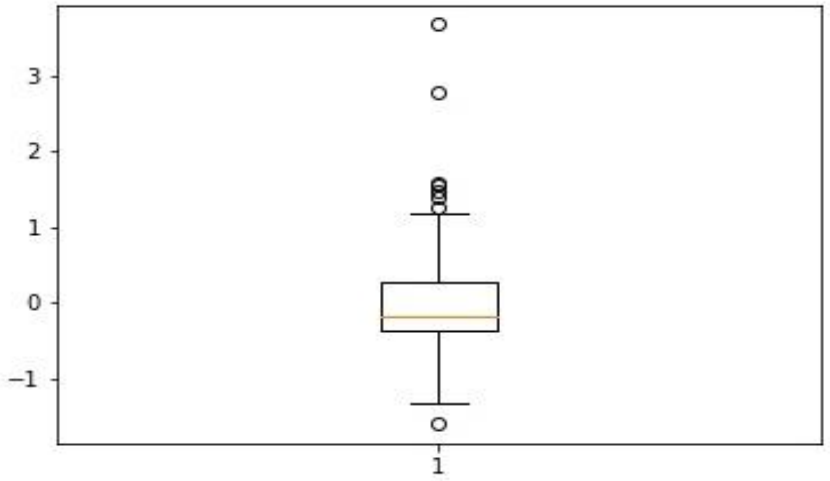
PC1



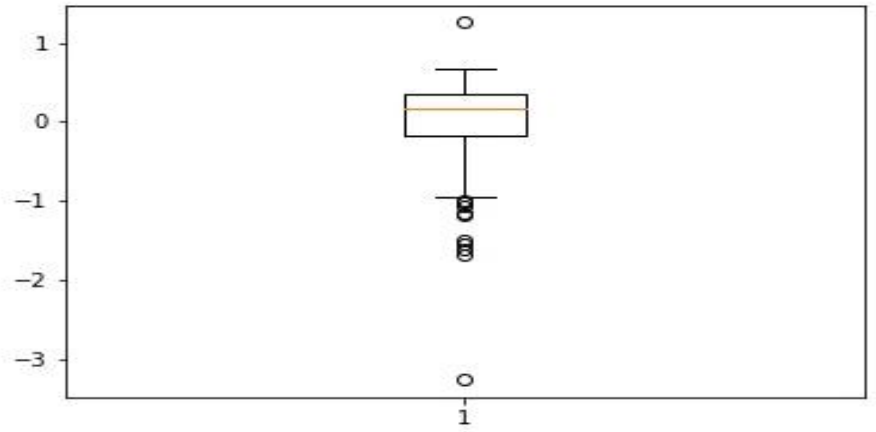
PC2



PC3



PC4

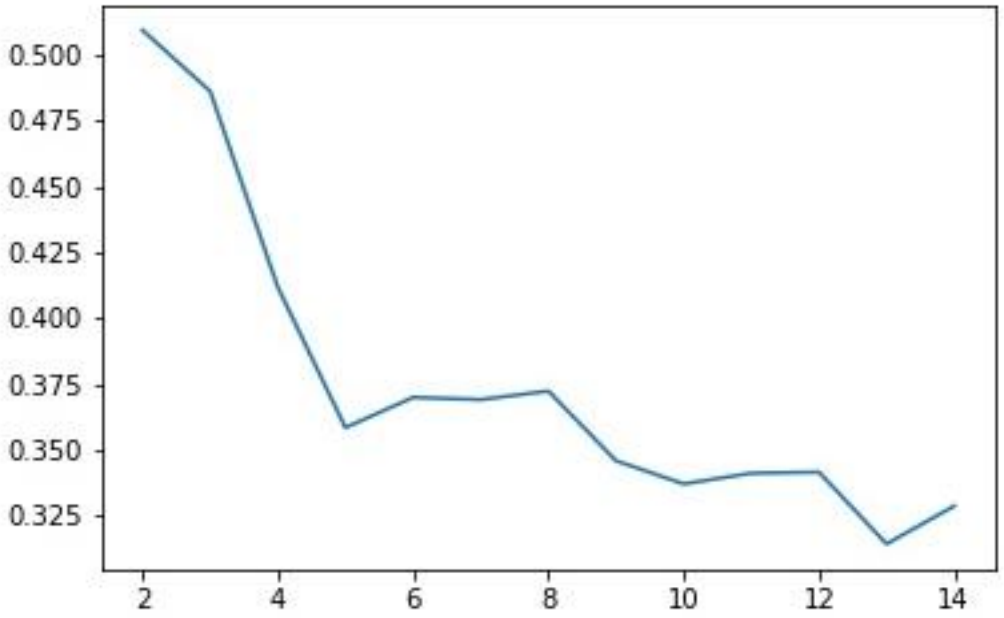


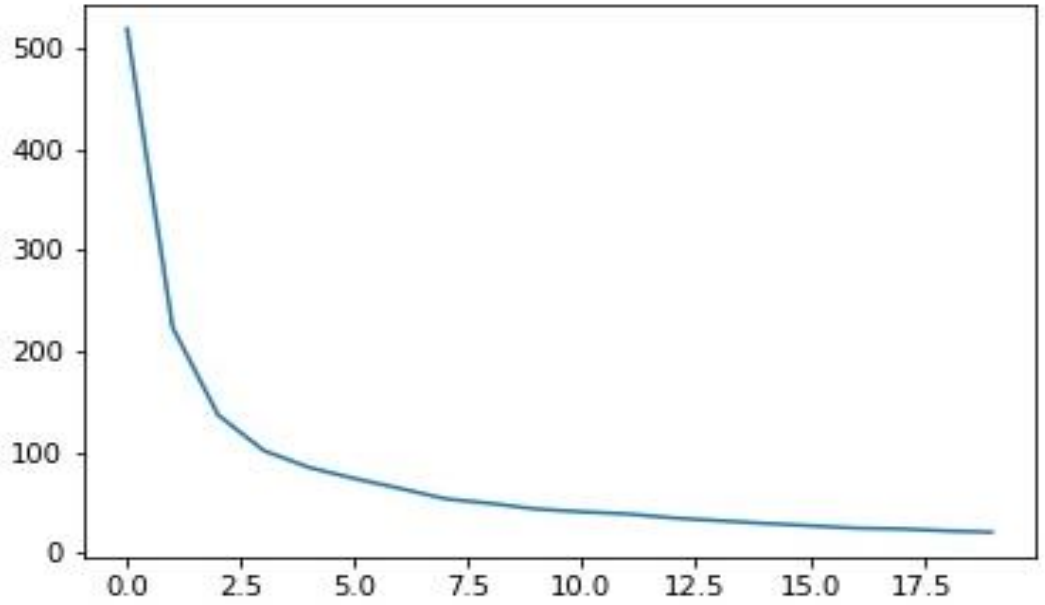
Box plot of Principal Components before outlier removal

Clustering

K-Means Clustering:

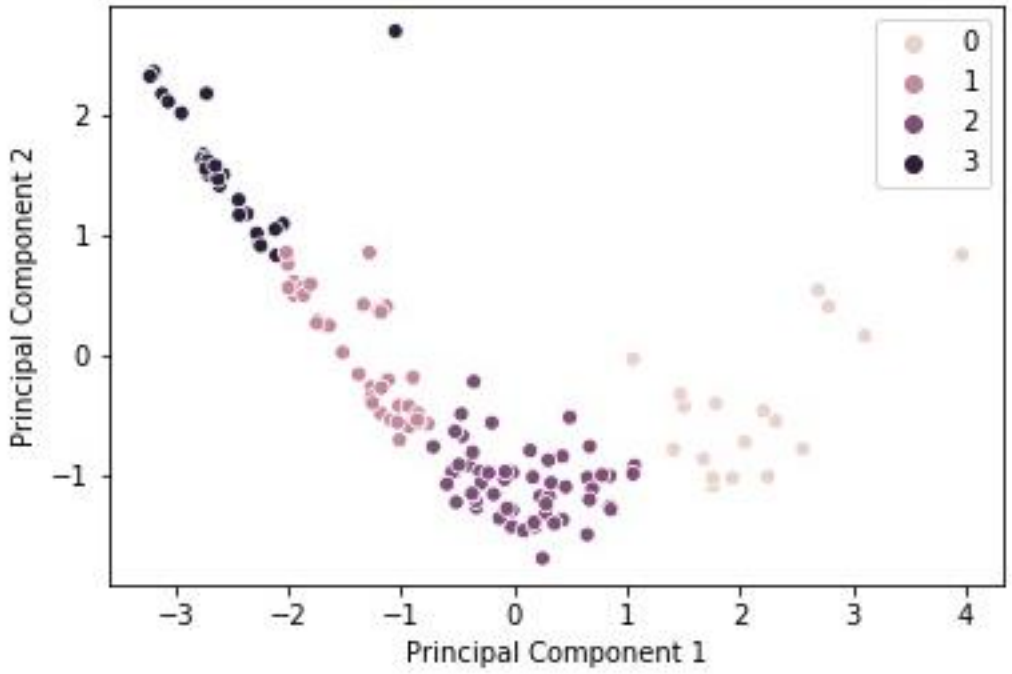
- After removing outliers from the PCA dataset, we needed to create clusters for analysis.
- From Hopkins statistics we inspected if k-means can be performed on the dataset. Since the score came as 0.84 , it is ideal for clustering.
- Then we did **Silhouette** and **Sum of squared distance** analysis to find the optimal number of clusters. From the below plots we chose number of **clusters as k=4** and proceed with our analysis.
- Later we again did the same analysis is done with number of clusters as 5.

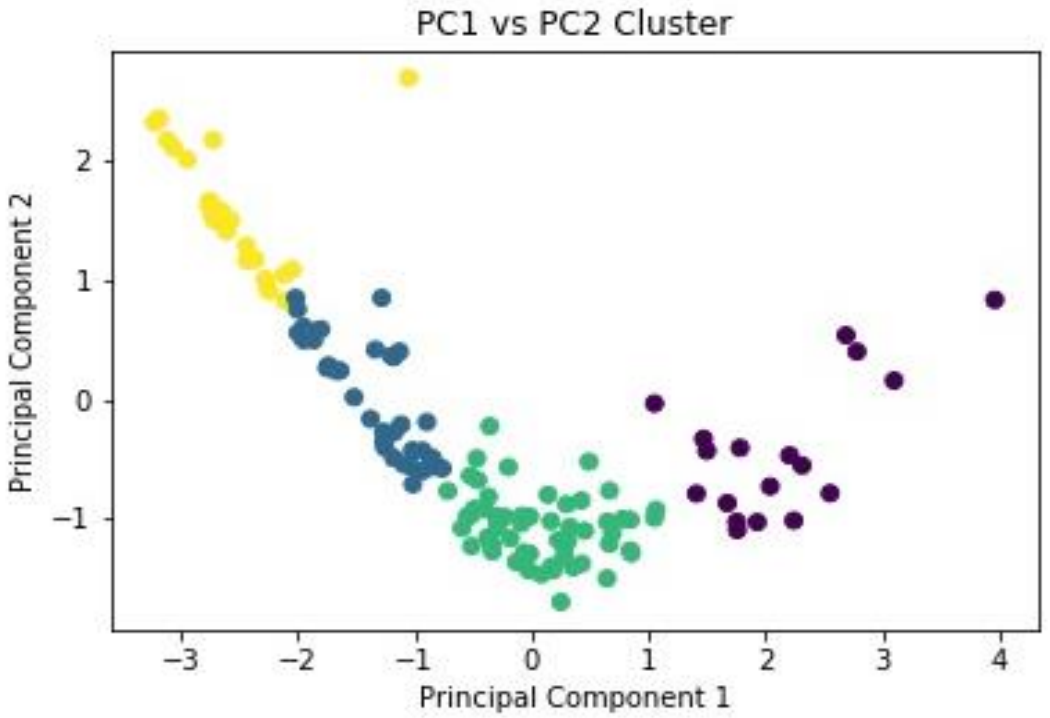




Clustering[Contd.]

Principal Components Cluster Analysis :





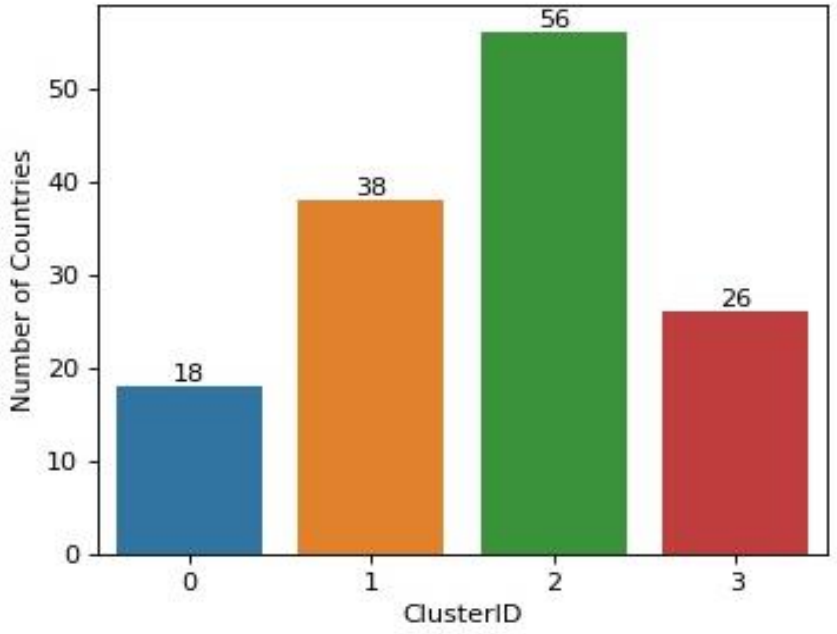
	ClusterID	PC1_mean	PC2_mean	PC3_mean	PC4_mean
0	0	2.123572	-0.422559	0.043955	-0.223509
1	1	-1.326986	-0.012854	-0.151375	0.153868
2	2	0.082870	-1.063954	0.053754	0.343528
3	3	-2.551250	1.562366	-0.485307	-0.266584

The mean of the principal components were computed for each clusters. While for cluster 0 the mean of PC1 is high whereas for cluster 1 the mean of PC4 is highest. For much better understanding we will join the cluster id with the original dataset in next step.

Clustering[Contd.]

- Clusters formed are then merged with the original dataset and mean of each of the variables are calculated for understanding the clusters formed.

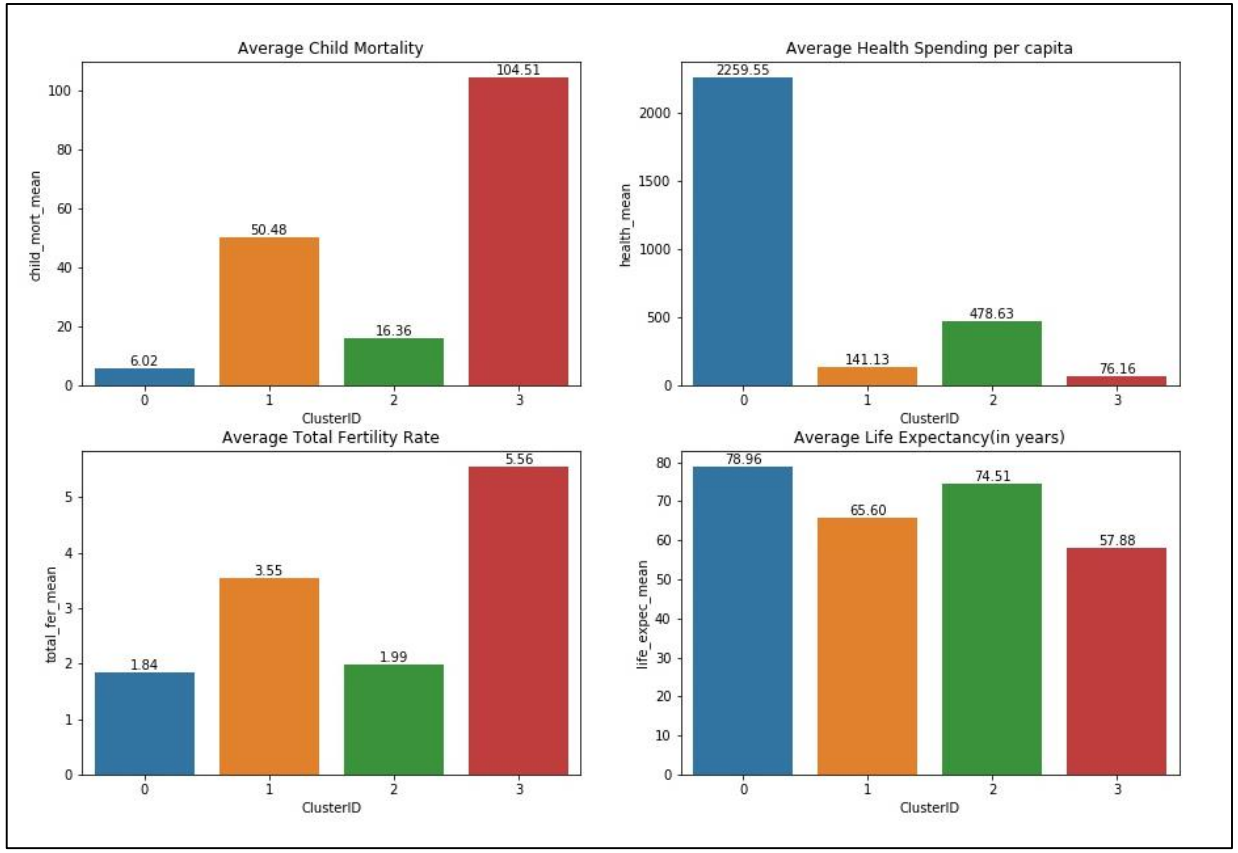
	ClusterID	child_mort_mean	exports_mean	health_mean	imports_mean	income_mean	inflation_mean	life_expec_mean	total_fer_mean	gdpp_mean
0	0	6.016667	14646.661111	2259.547778	13230.394444	36388.888889	3.568333	78.955556	1.840000	28927.777778
1	1	50.481579	880.094219	141.126345	1018.011897	5126.315789	8.396842	65.602632	3.554211	2352.789474
2	2	16.355357	3264.368214	478.629804	3500.688036	13692.678571	5.758018	74.507143	1.993571	7207.678571
3	3	104.507692	923.082046	76.155192	779.279462	3189.038462	10.104038	57.880769	5.557692	1510.769231

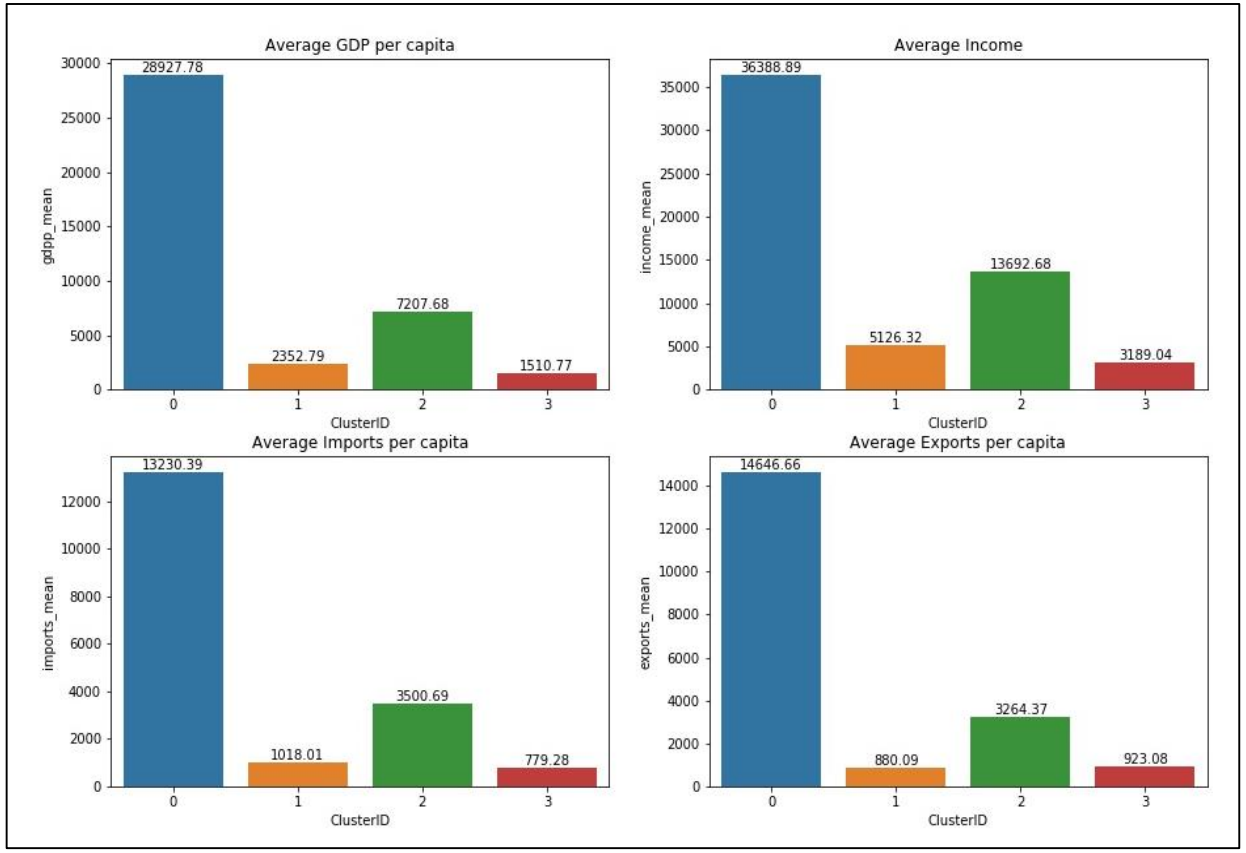


Countries are divided among the four clusters. Cluster 2 have the maximum number of countries while cluster 0 as the least number of countries. From the above data we will derive insights about the clusters formed in the next step.

Clustering[Contd.]

Socio-economic Analysis of the Clusters:





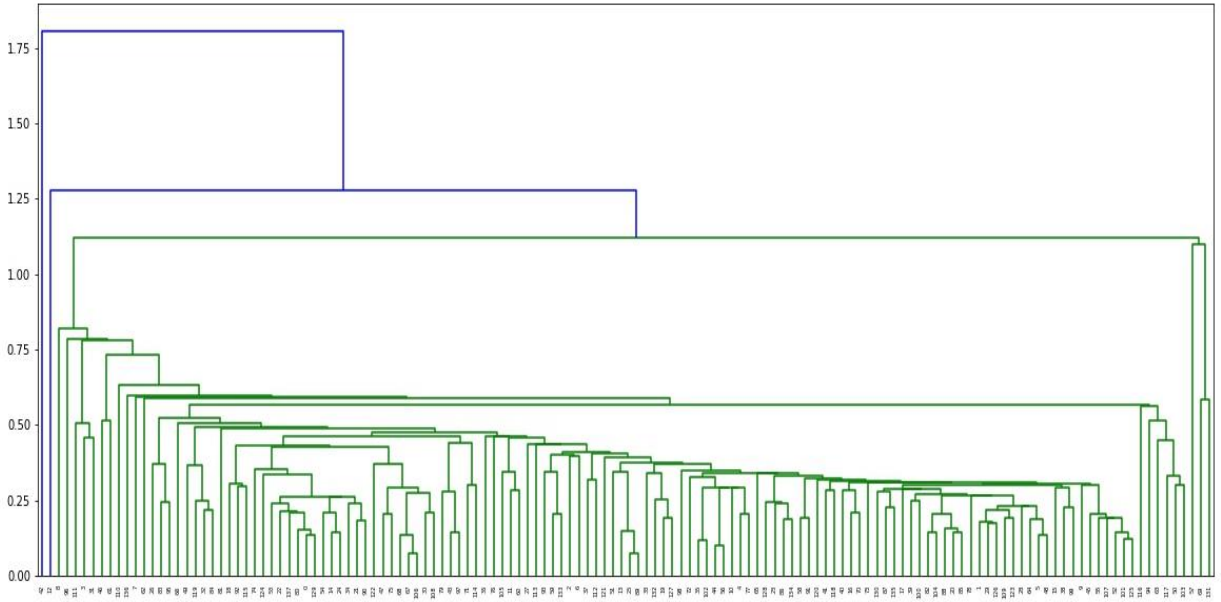
Insight :

- Clusters 1 and 3 are having the highest child mortality rate and total fertility rate, and lowest health spending ,gdp per capita, income, exports and imports. Hence it can be said that these clusters denoted the underdeveloped or developing countries. Cluster 3 is the most affected cluster with the lowest gdpp among all other clusters.
- Cluster 0 represents the countries with high gdpp, income, health spending and least mortality rate. They are the developed countries.
- Cluster 2 falls in between them, with an average gdpp and income.

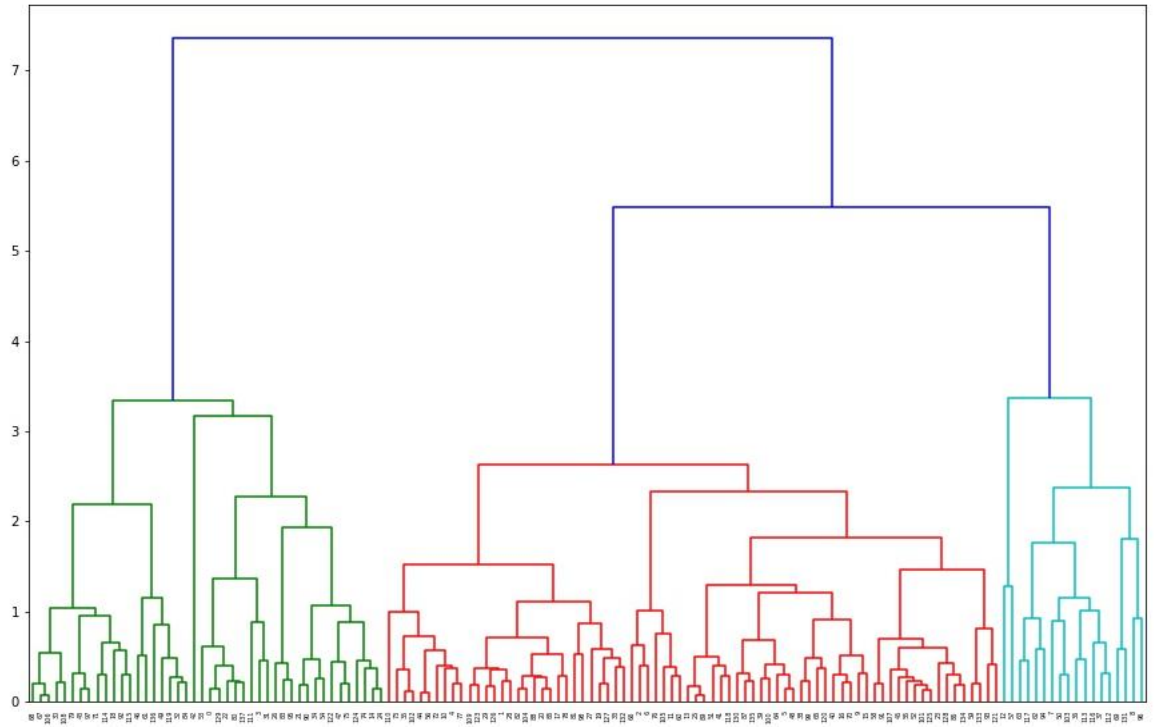
Clustering[Contd.]

Hierarchical Clustering:

Hierarchical Clustering is performed on the PCA reduced dataset. Both single and complete linkage clustering is performed and cluster_ids are determined. Below are the dendograms obtained from single and complete linkage.



Single Linkage

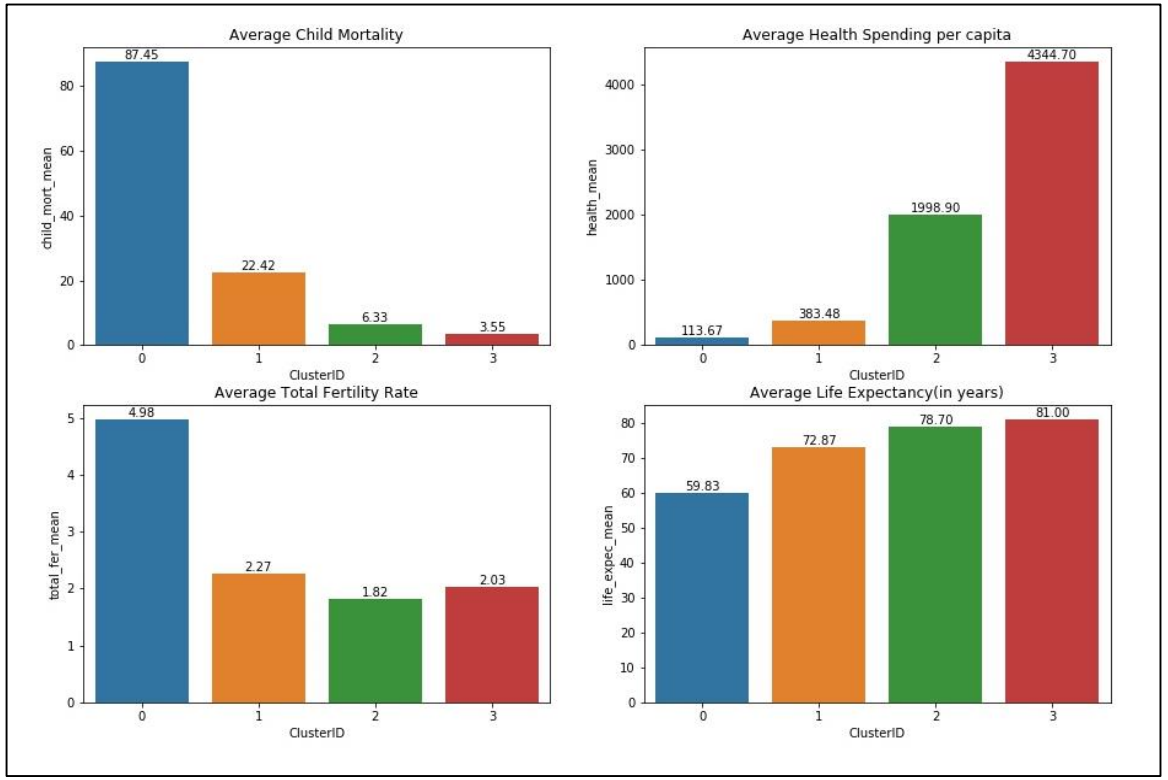


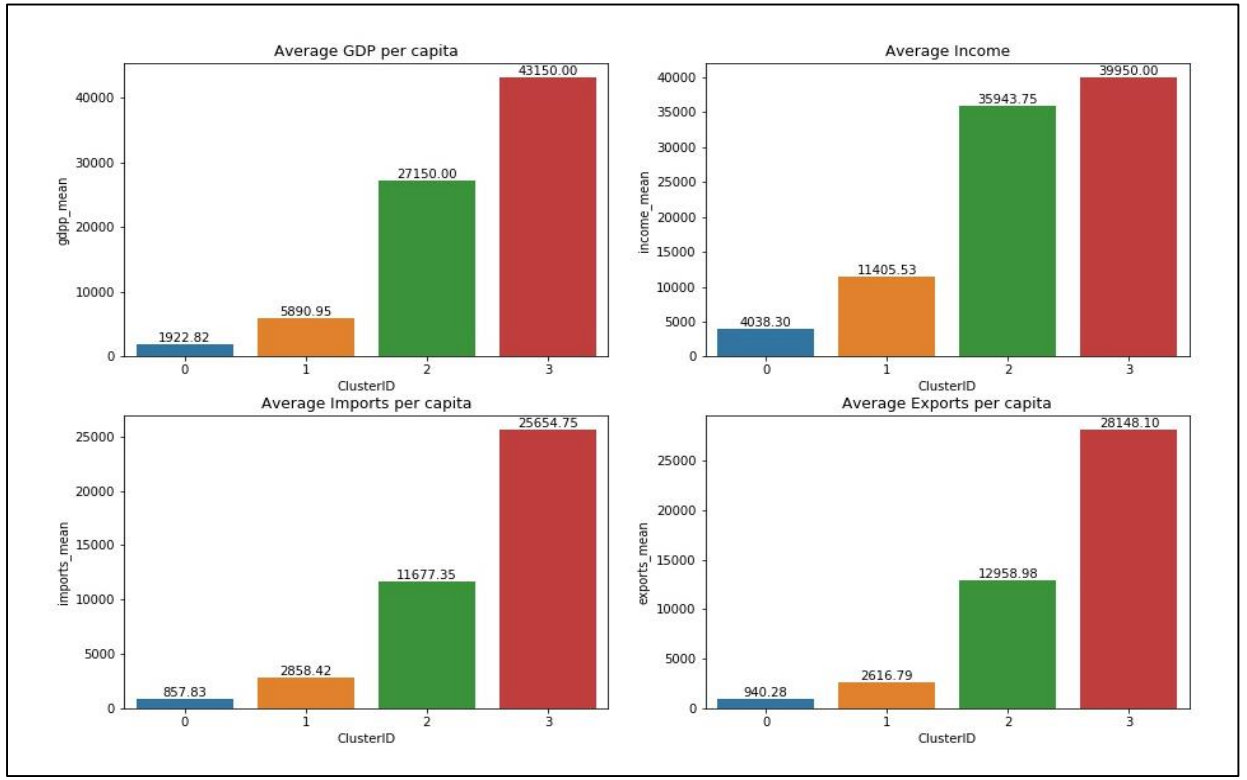
Complete Linkage

We cut the dendogram to give us 4 clusters on which we will do our analysis.

Clustering[Contd.]

Socio-economic Analysis of the Clusters from Hierarchical Method:





Insight :

- Clusters 0 and 1 are having the highest child mortality rate and total fertility rate, and lowest health spending ,gdp per capita, income, exports and imports. Hence it can be said that these clusters denoted the underdeveloped or developing countries. Cluster 0 is the most affected cluster with the lowest gdpp among all other clusters.
- Cluster 3 represents the countries with high gdpp, income, health spending and least mortality rate. They are the developed countries.
- Cluster 2 falls in between them, with an average gdpp and income.

CONCLUSION

From the plots obtained after K-Means and Hierarchical Clustering, we can arrive at the following conclusion :

- Countries which are part of cluster-3 in K-means clustering and cluster-0 in Hierarchical clustering are the countries which are having very high child mortality rate and least gdpp and income. Their child fertility rate is also high which means that the population tends to increase and this combined with the poor gdpp and income, imports/exports reflects that these countries are in dire need of aid and support.
- The countries that falls under this cluster are third world countries like **Afghanistan, Angola, Benin, Burkino, Congo Republic, Uganda, Zambia** etc.
- According to Wikipedia (https://en.wikipedia.org/wiki/Poverty_in_Afghanistan), Afghanistan is one amongst the poorest countries in the world. In Afghanistan, poverty is widespread in rural and urban areas.
- NGO Help should direct their funds for the social betterment of these countries.