Clustering & PCA Assignment

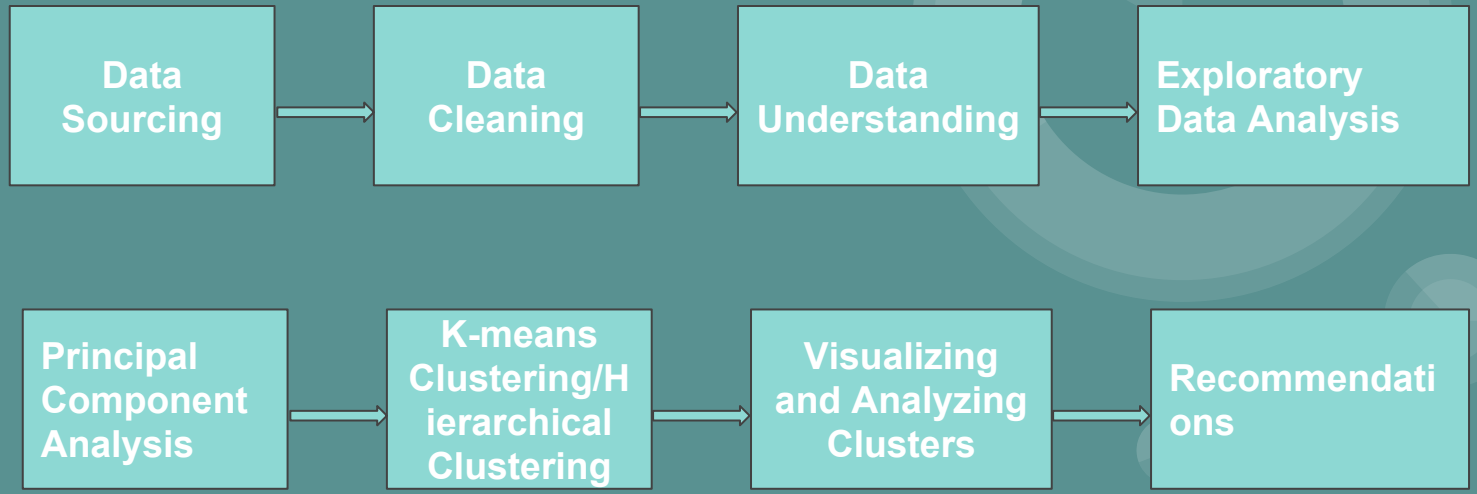
HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities. It runs a lot of operational projects from time to time along with advocacy 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.

Analysis Objectives

To categorise the countries using some socio-economic and health factors that determine the overall development of the country. Also to suggest the countries which the CEO needs to focus on the most. The datasets containing those socio-economic factors and the corresponding data dictionary are provided.

Methodology used for Analysis



Data Sourcing/ Data Cleaning/ Data Understanding

- There are 10 columns and 167 variables; dtypes: float64(7), int64(2), object(1)
- There seems to be a few outliers but they will be removed after applying PCA
- No null values are present in the given dataset
- No missing values are present in the given dataset
- The data has to be standardised/ normalised in order to apply PCA on the dataset
- The dataset does not require any extensive EDA.
- The variables are a mix of economic and social factors
- Economic factors can help to cluster the countries into different clusters and social factors can help to disburse funds for different sectors.

First look at the dataset

	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

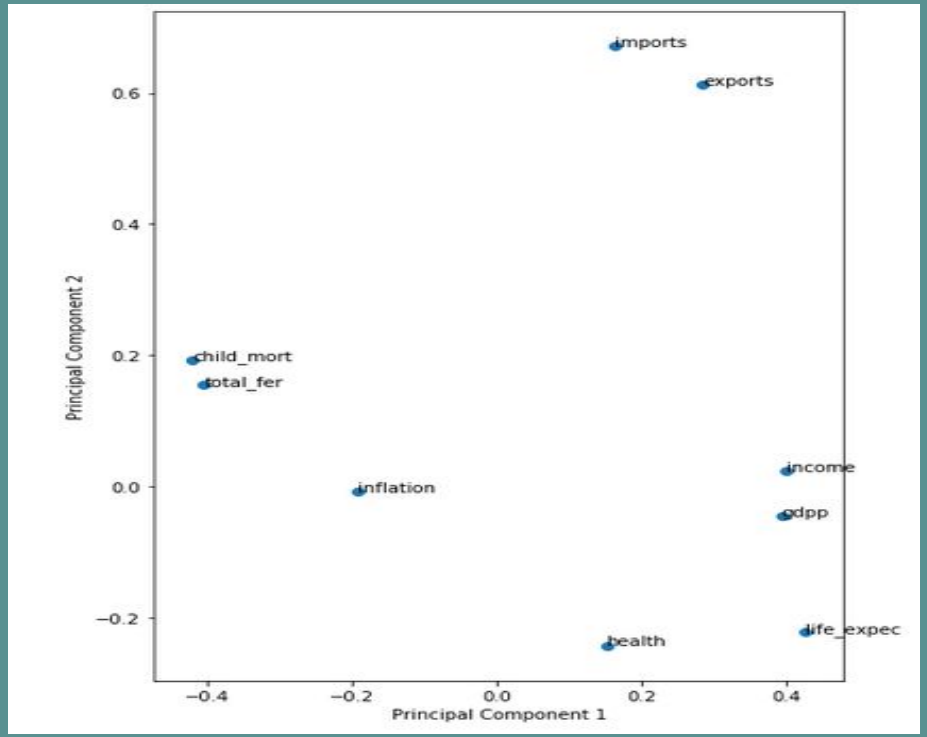
Variables-
Country Name, Child Mortality Rate, Health, Exports, Imports, Income, Inflation, Life Expectancy, Total Fertility Rate, Gross Domestic Product.

Principal Component Analyis

	PC1	PC2	PC3	PC4	Feature
0	-0.419519	0.192884	-0.029544	0.370653	child_mort
1	0.283897	0.613163	0.144761	0.003091	exports
2	0.150838	-0.243087	-0.596632	0.461897	health
3	0.161482	0.671821	-0.299927	-0.071907	imports
4	0.398441	0.022536	0.301548	0.392159	income
5	-0.193173	-0.008404	0.642520	0.150442	inflation
6	0.425839	-0.222707	0.113919	-0.203797	life_expec
7	-0.403729	0.155233	0.019549	0.378304	total_fer
8	0.392645	-0.046022	0.122977	0.531995	gdpp

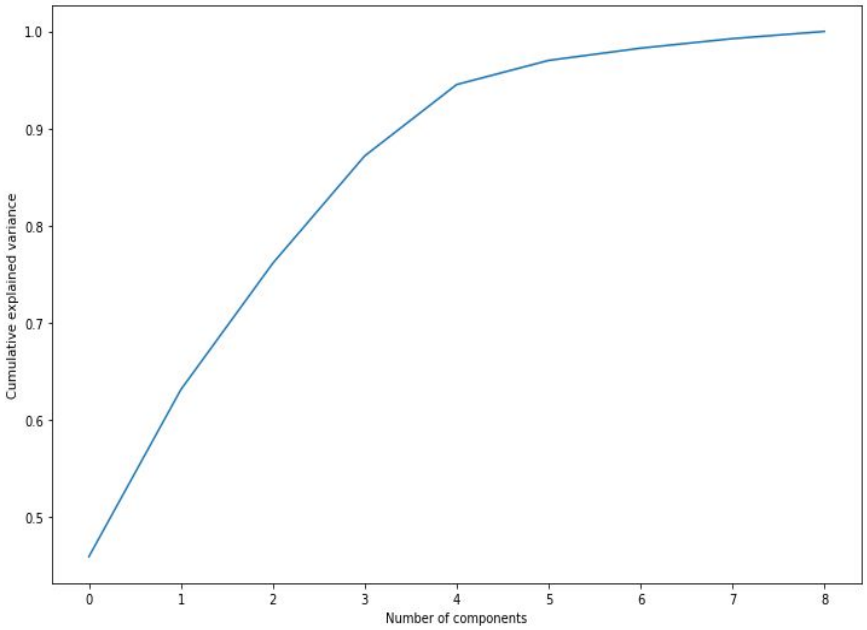
- Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.
- Taking four principal components for each feature variable.

Principal Component Analyis



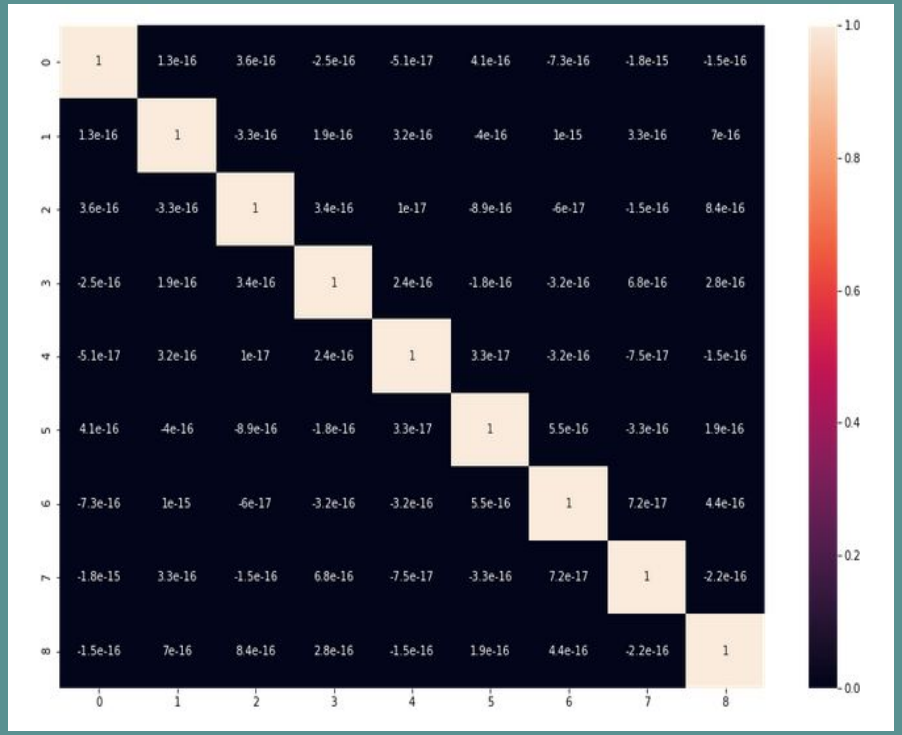
- We can see that the first component is in the direction of life_expec, income, gdpp; Also these three components have the highest loadings.
- The second component is in the direction of imports and exports

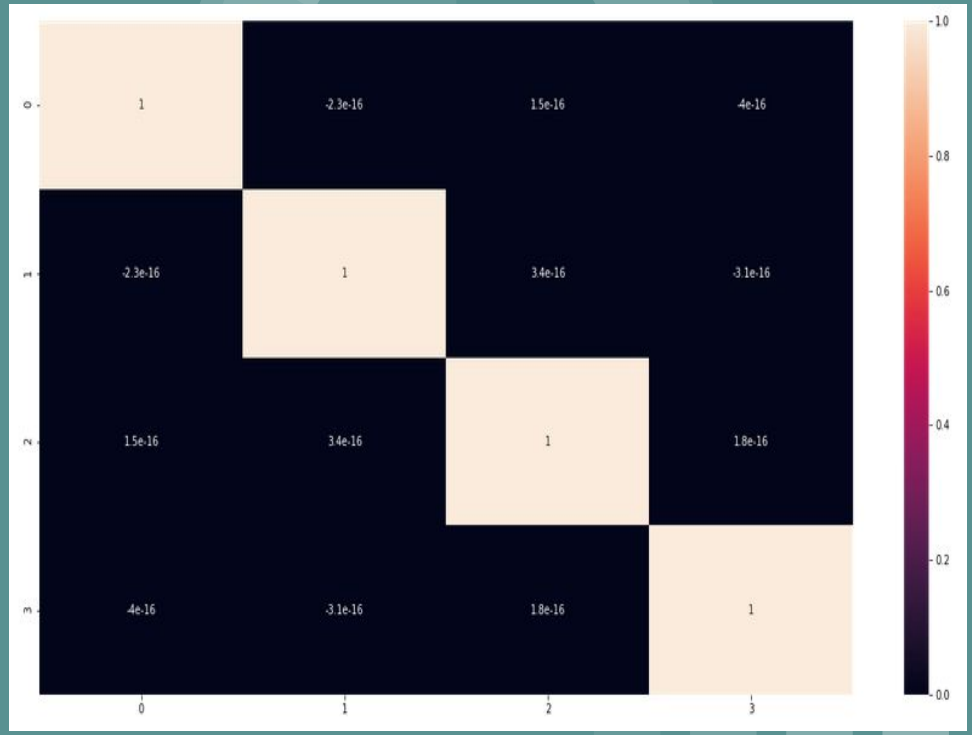
Principal Component Analysis
Scree Plot



- A Scree Plot is a simple line segment plot that shows the fraction of total variance in the data as explained or represented by each PC.
- Therefore, four principal components are enough to explain variance(more than 80%) in out data.

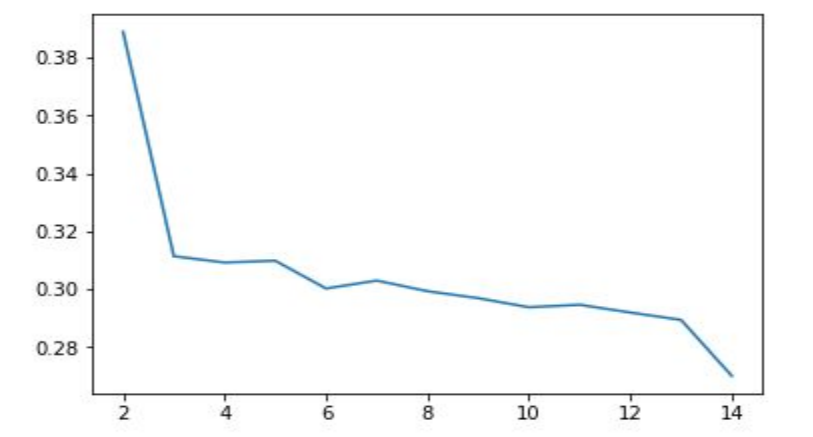
Principal Component Analyis
Correlation- Heat Map





PCA succeeded in removing correlations in data.

K- Means Clustering
Silhouette Analysis and Hopkins Statistics



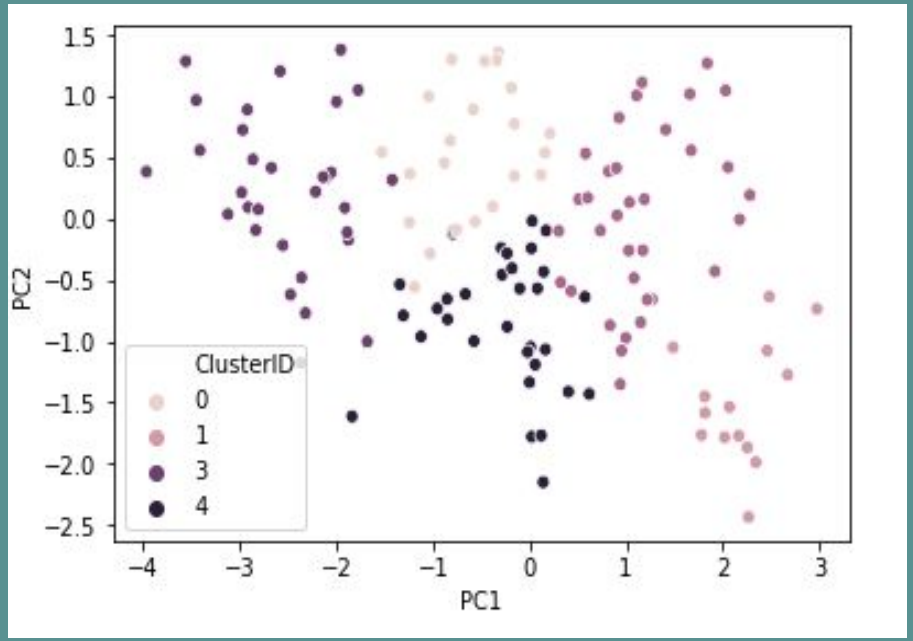
Hopkins Statistics:

The Hopkins statistic, is a statistic which gives a value which indicates the cluster tendency, in other words: how well the data can be clustered.If the value is between {0.7, ..., 0.99}, it has a high tendency to cluster.
For our Dataset; Hopkin statistic value=0.762

Silhouette Analysis is the mean intra-cluster distance to all the points in its own cluster.

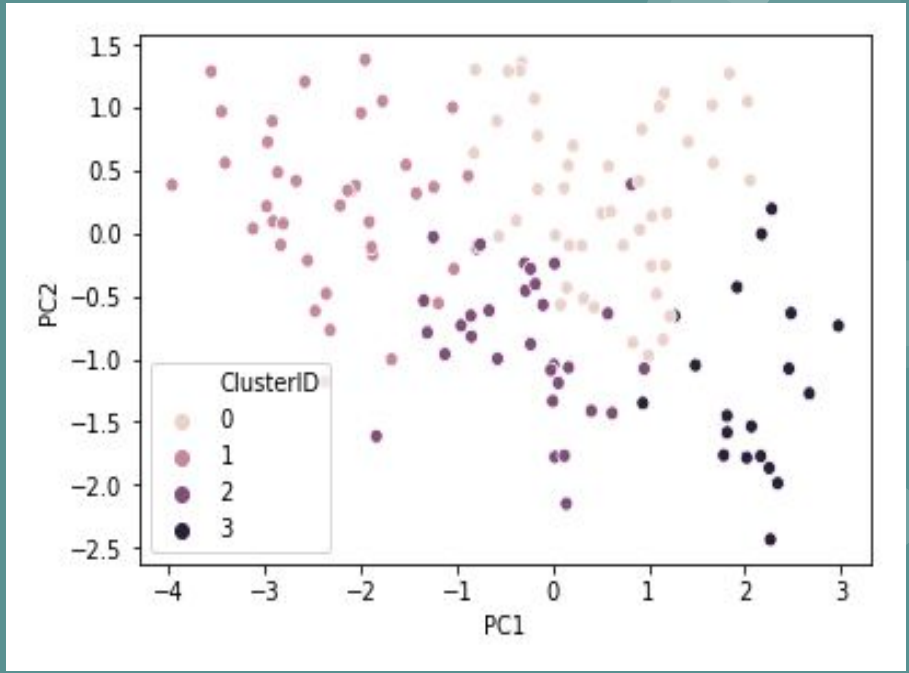
- The value of the silhouette score range lies between -1 to 1.
- A score closer to 1 indicates that the data point is very similar to other data points in the cluster,
- A score closer to -1 indicates that the data point is not similar to the data points in its cluster.

K- Means Clustering
k=5



- We can see from below data that four clusters are enough to represent our data
- Reiterating with four clusters

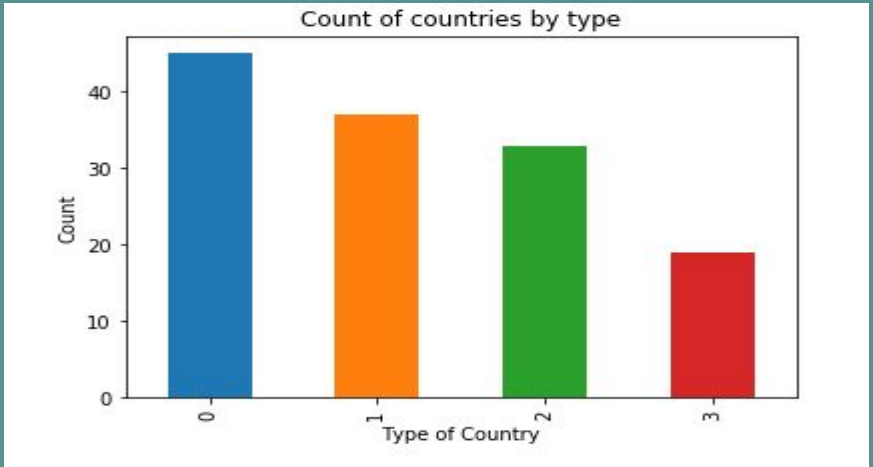
K- Means Clustering
k=4



Scatter Plot between PC1 and PC2 for 4 clusters

K- Means Clustering
k=4

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	PC1	PC2	PC3	PC4	ClusterID
0	Afghanistan	90.2	10.0	7.58	44.9	1610	9.44	56.2	5.82	553	-2.913025	0.095621	-0.718118	1.005255	1
1	Albania	16.6	28.0	6.55	48.6	9930	4.49	76.3	1.65	4090	0.429911	-0.588156	-0.333486	-1.161059	0
2	Algeria	27.3	38.4	4.17	31.4	12900	16.10	76.5	2.89	4460	-0.285225	-0.455174	1.221505	-0.868115	2
3	Angola	119.0	62.3	2.85	42.9	5900	22.40	60.1	6.16	3530	1.033576	0.136659	-0.225721	-0.847063	0
4	Antigua and Barbuda	10.3	45.5	6.03	58.9	19100	1.44	76.8	2.13	12200	0.022407	-1.779187	0.869997	-0.036967	2



- Countries are divided into 4 clusters;
- Highest number of countries belong to Cluster 0

K- Means Clustering
k=4

Cluster 0

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	PC1	PC2	PC3	PC4	ClusterID
1	Albania	16.6	28.0	6.55	48.6	9930	4.490	76.3	1.65	4090	0.429911	-0.588156	-0.333486	-1.161059	0
3	Angola	119.0	62.3	2.85	42.9	5900	22.400	60.1	6.16	3530	1.033576	0.136659	-0.225721	-0.847063	0
10	Bahamas	13.8	35.0	7.89	43.7	22900	-0.393	73.8	1.86	28000	1.670996	0.561162	0.991258	-0.207080	0
12	Bangladesh	49.4	16.0	3.52	21.8	2440	7.140	70.4	2.33	758	1.081374	-0.481970	-0.664355	-0.522505	0
13	Barbados	14.2	39.5	7.97	48.7	15300	0.321	76.7	1.78	16000	0.580025	0.535327	0.486228	-1.035275	0

Cluster 1

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	PC1	PC2	PC3	PC4	ClusterID
0	Afghanistan	90.2	10.0	7.58	44.9	1610	9.44	56.2	5.82	553	-2.913025	0.095621	-0.718118	1.005255	1
15	Belgium	4.5	76.4	10.70	74.7	41100	1.88	80.0	1.86	44400	-2.672314	0.418172	-0.257368	0.278672	1
19	Bolivia	46.6	41.2	4.84	34.3	5410	8.78	71.6	3.20	1980	-0.882088	0.457368	-0.584633	0.406161	1
22	Brazil	19.8	10.7	9.01	11.8	14500	8.41	74.2	1.80	11200	-3.122053	0.038775	-0.455751	1.080918	1
24	Bulgaria	10.8	50.2	6.87	53.0	15300	1.11	73.9	1.57	6840	-2.807909	0.078649	-0.342961	0.543557	1

K- Means Clustering
k=4

Cluster 2

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	PC1	PC2	PC3	PC4	ClusterID
2	Algeria	27.3	38.4	4.17	31.4	12900	16.100	76.5	2.89	4460	-0.285225	-0.455174	1.221505	-0.868115	2
4	Antigua and Barbuda	10.3	45.5	6.03	58.9	19100	1.440	76.8	2.13	12200	0.022407	-1.779187	0.869997	-0.036967	2
5	Argentina	14.5	18.9	8.10	16.0	18700	20.900	75.8	2.37	10300	-0.101584	-0.568252	0.242092	-1.466266	2
8	Austria	4.3	51.3	11.00	47.8	43200	0.873	80.5	1.44	46900	-0.181487	-0.402866	0.867459	-0.438773	2
11	Bahrain	8.6	69.5	4.97	50.9	41100	7.440	76.0	2.16	20700	-1.123851	-0.961397	0.526615	-1.197201	2

Cluster 3

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	PC1	PC2	PC3	PC4	ClusterID
6	Armenia	18.1	20.8	4.40	45.3	6700	7.77	73.3	1.69	3220	2.342165	-1.988459	0.190344	1.105038	3
7	Australia	4.8	19.8	8.73	20.9	41400	1.16	82.0	1.93	51900	2.973764	-0.734689	-0.519766	1.205442	3
9	Azerbaijan	39.2	54.3	5.88	20.7	16000	13.80	69.1	1.92	5840	1.268744	-0.656588	-0.488098	0.055634	3
32	Chad	150.0	36.8	4.53	43.5	1930	6.39	56.5	6.59	897	0.937827	-1.350472	-0.821130	-0.259855	3
35	Colombia	18.6	15.9	7.59	17.8	10900	3.86	76.4	2.01	6250	2.174455	-0.004510	0.257320	-0.311857	3

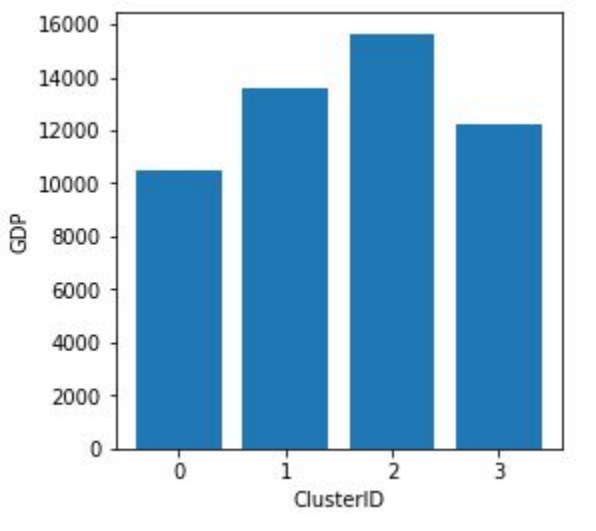
K- Means Clustering
k=4

Mean of Country Data by Cluster for further analysis

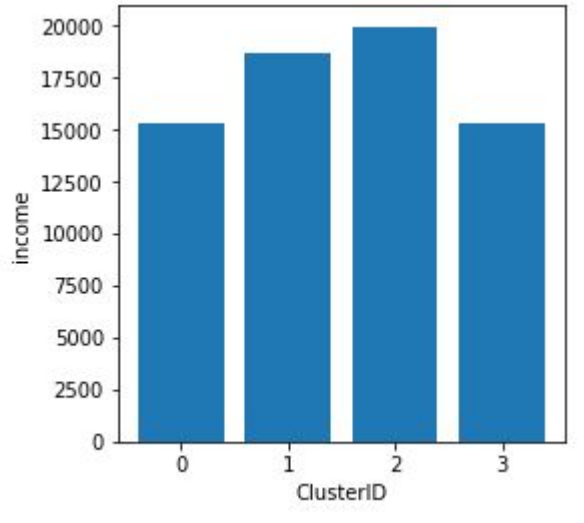
ClusterID	gdpp	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	ClusterID
0	10515.622222	36.313333	43.180200	6.453556	51.348131	15283.777778	4.657733	70.777778	2.810667	0
1	13581.675676	39.454054	42.562432	6.646216	43.875676	18668.918919	7.540730	70.372973	3.024324	1
2	15651.787879	31.563636	43.290909	7.139697	50.581818	19967.848485	8.042030	72.260606	2.738788	2
3	12250.368421	65.147368	33.667368	6.786316	41.726316	15315.473684	12.894632	66.473684	3.558421	3

K- Means Clustering(k=4)
Analysis of the clusters

Mean GDP by Cluster



Mean Income by Cluster

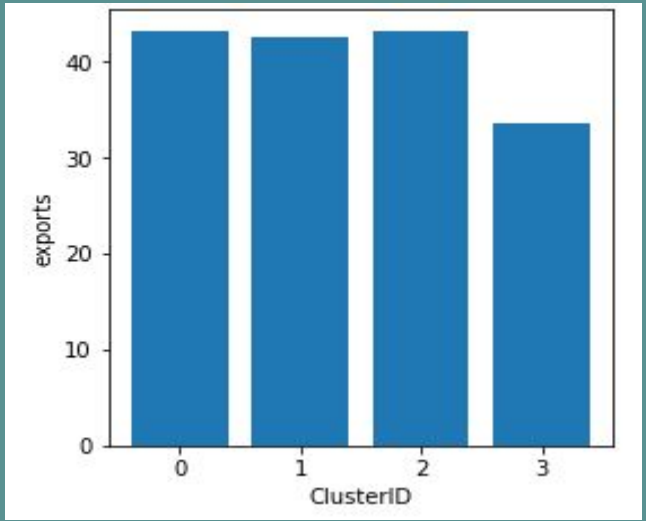


- When allocating money for creating job, skilling people and employment so that the GDP of the companies increases--countries in Cluster 0 and 3 should be focussed.

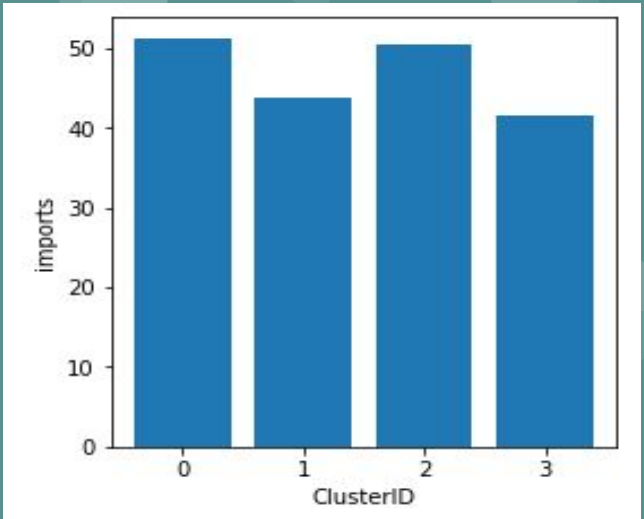
- Mean Income of the countries in Cluster 0 and 3 is less

K- Means Clustering(k=4)
Analysis of the clusters

Mean exports by Cluster



Mean imports by Cluster

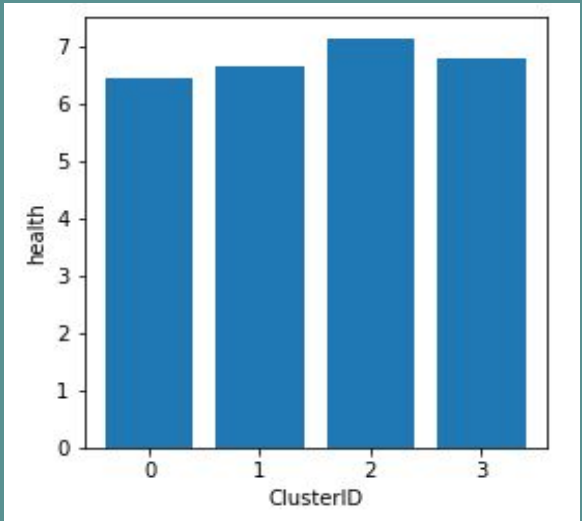


- The exports of countries in Cluster 1, Cluster 2 and Cluster 0 are comparable; Funds should given to help countries in Cluster 3 to increase there exports;

- Imports of a country should not be high as it devalues its currency and also increases debt;

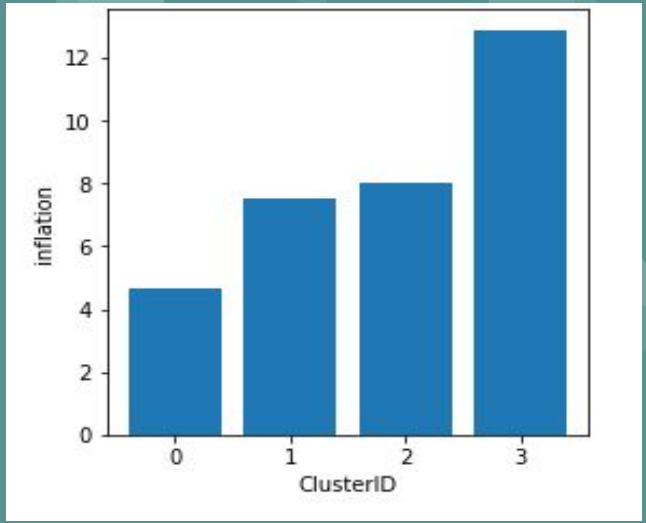
K- Means Clustering(k=4)
Analysis of the clusters

Mean health by Cluster



- Spendings on health is comparable for all the three clusters; Other factors such as life expectancy and child mortality will be a more defining factor for funds disbursal

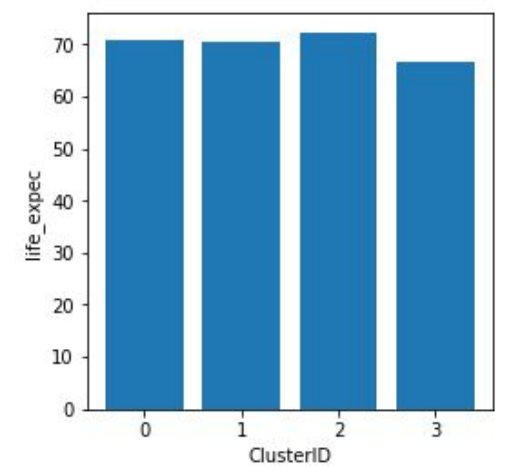
Mean inflation by Cluster

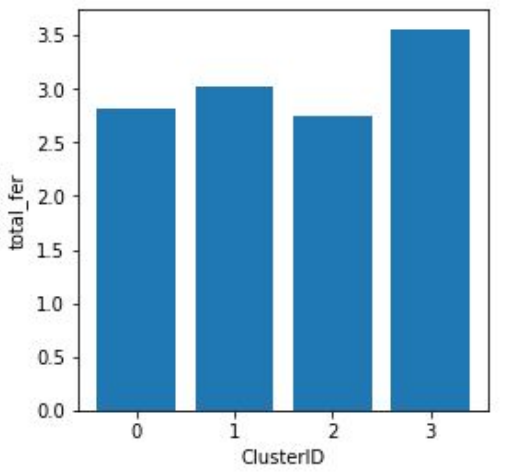


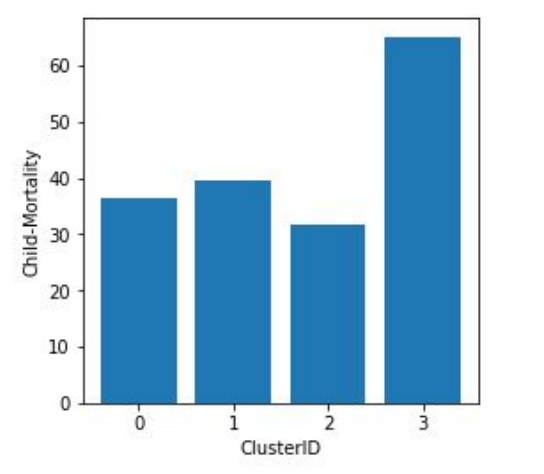
- Countries in Cluster 3 show high inflation; Measures to curb the same should be taken by appropriate fund disbursal.

K- Means Clustering(k=4)
Analysis of the clusters

Mean life expectancy, total fertility rate and child mortality by Cluster







- Life expectancy is comparable in almost all the clusters; Though this is a mean data which requires further scrutinization to understand medical facilities in countries which enhances life expectancy
- Cluster 3 shows high total fertility which is not good for the health of both mother and her children; Higher TFR means
- # greater population which is a burden on our limited resources; Funds should be spent

K- Means Clustering(k=4)
Analysis of the clusters

Mean life expectancy, total fertility rate and child mortality by Cluster

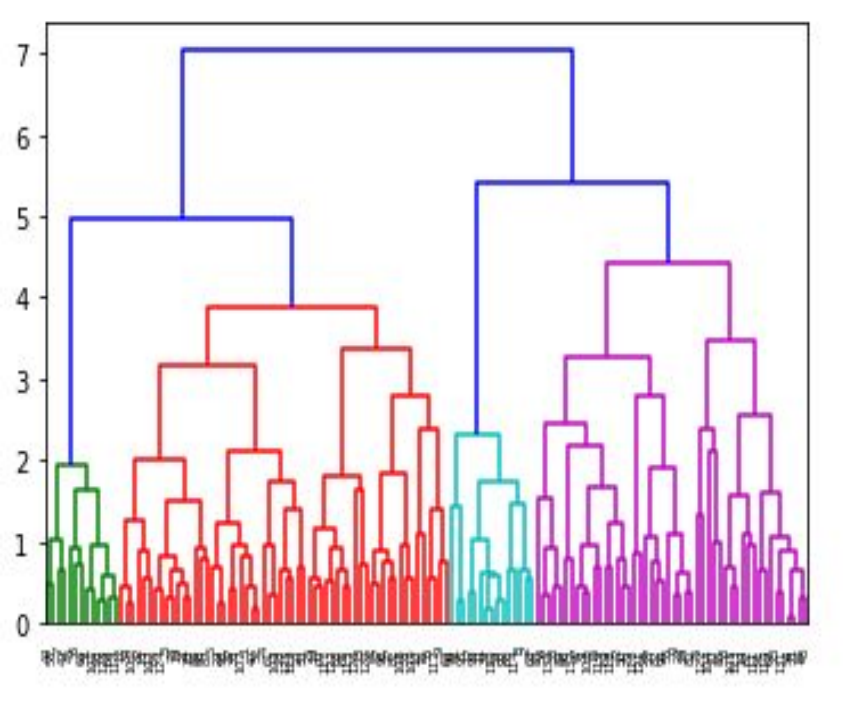
- Life expectancy is comparable in almost all the clusters; Though this is a mean data which requires further scrutinization to understand medical facilities in countries which enhances life expectancy

- Cluster 3 shows high total fertility which is not good for the health of both mother and her children; Higher TFR means greater population which is a burden on our limited resources; Funds should be spent on educating couples as well as providing them with birth control measures. Focus should be on countries in Cluster 3.

- Child mortality seems to be on the higher side in Cluster 3. Hence funds should be disbursed for countries in cluster 3
for children health, nutrition,maternity care, vaccination and education.

Hierarchical Clustering

k=4



Hierarchical Clustering dendrogram

A dendrogram is a type of tree diagram showing hierarchical clustering — relationships between similar sets of data.

Taking number of clusters as 4;
Also taking a clue from
k-means clustering

Hierarchical Clustering- Clustered Data

Cluster 0

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	ClusterID
0	Afghanistan	90.2	10.0	7.58	44.9	1610	9.44	56.2	5.82	553	0.0
15	Belgium	4.5	76.4	10.70	74.7	41100	1.88	80.0	1.86	44400	0.0
16	Belize	18.8	58.2	5.20	57.5	7880	1.14	71.4	2.71	4340	0.0
19	Bolivia	46.6	41.2	4.84	34.3	5410	8.78	71.6	3.20	1980	0.0
22	Brazil	19.8	10.7	9.01	11.8	14500	8.41	74.2	1.80	11200	0.0

Cluster 1

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	ClusterID
1	Albania	16.6	28.0	6.55	48.6	9930	4.49	76.3	1.65	4090	1.0
3	Angola	119.0	62.3	2.85	42.9	5900	22.40	60.1	6.16	3530	1.0
4	Antigua and Barbuda	10.3	45.5	6.03	58.9	19100	1.44	76.8	2.13	12200	1.0
5	Argentina	14.5	18.9	8.10	16.0	18700	20.90	75.8	2.37	10300	1.0
9	Azerbaijan	39.2	54.3	5.88	20.7	16000	13.80	69.1	1.92	5840	1.0

Hierarchical Clustering- Clustered Data

Cluster 2

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	ClusterID
2	Algeria	27.3	38.4	4.17	31.4	12900	16.100	76.5	2.89	4460	2.0
8	Austria	4.3	51.3	11.00	47.8	43200	0.873	80.5	1.44	46900	2.0
11	Bahrain	8.6	69.5	4.97	50.9	41100	7.440	76.0	2.16	20700	2.0
17	Benin	111.0	23.8	4.10	37.2	1820	0.885	61.8	5.36	758	2.0
39	Costa Rica	10.2	33.2	10.90	35.0	13000	6.570	80.4	1.92	8200	2.0

Cluster 3

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	ClusterID
6	Armenia	18.1	20.8	4.40	45.3	6700	7.77	73.3	1.69	3220	3.0
7	Australia	4.8	19.8	8.73	20.9	41400	1.16	82.0	1.93	51900	3.0
44	Denmark	4.1	50.5	11.40	43.6	44000	3.22	79.5	1.87	58000	3.0
45	Dominican Republic	34.4	22.7	6.22	33.3	11100	5.44	74.6	2.60	5450	3.0
49	Equatorial Guinea	111.0	85.8	4.48	58.9	33700	24.90	60.9	5.21	17100	3.0

Hierarchical Clustering

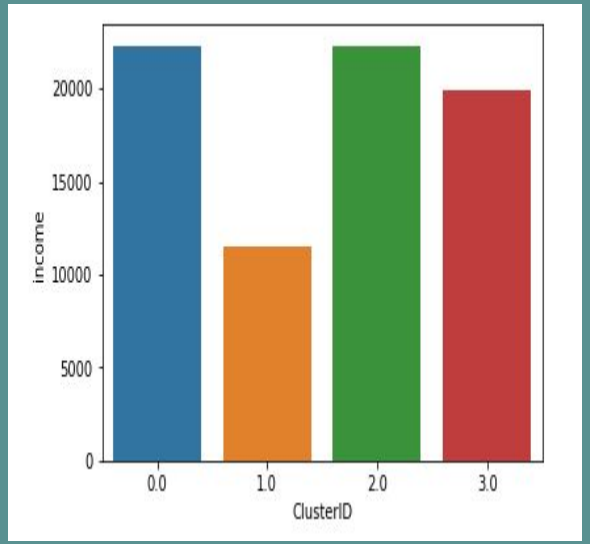
Analysis of Mean Data by Cluster

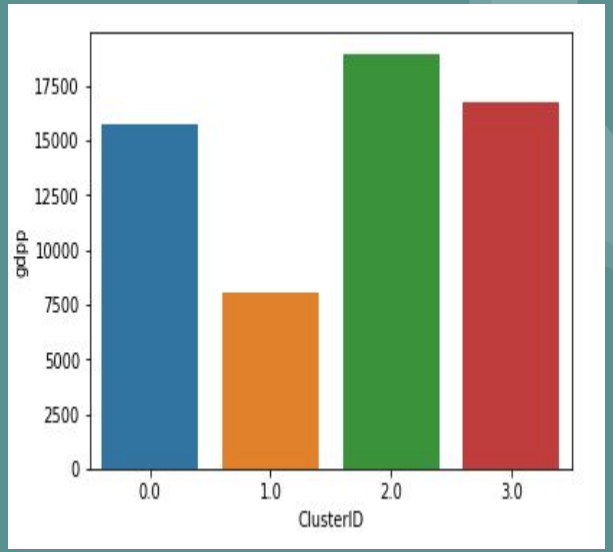
Mean of Country Data by Cluster for further analysis

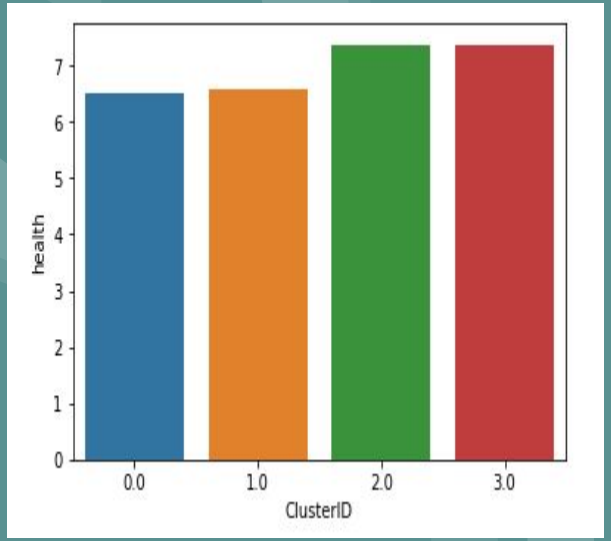
gdpp	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	ClusterID
15719.666667	35.020833	45.410625	6.502292	45.493750	22300.416667	6.668521	71.412500	2.865625	0.0
8062.362069	44.148276	38.996362	6.595345	50.621826	11470.051724	6.790534	69.155172	3.111207	1.0
18976.533333	25.380000	44.940000	7.375333	46.280000	22306.666667	6.523467	74.780000	2.448000	2.0
16783.230769	57.769231	36.200000	7.353846	44.776923	19859.230769	14.399846	67.376923	3.203846	3.0

Hierarchical Clustering(k=4)
Analysis of the clusters

Mean Income, GDP and health by Cluster

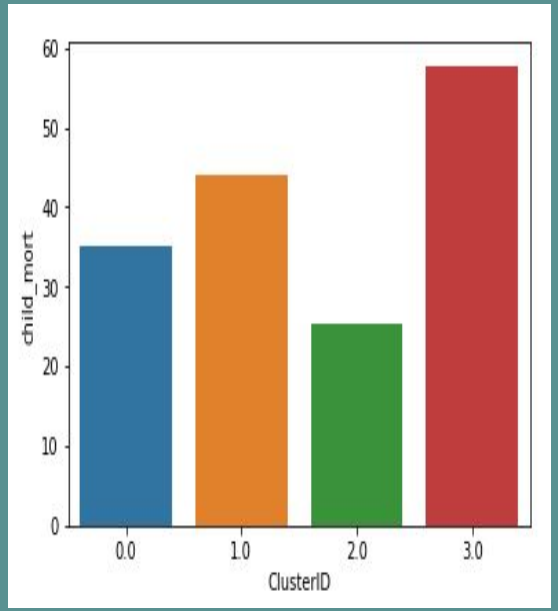


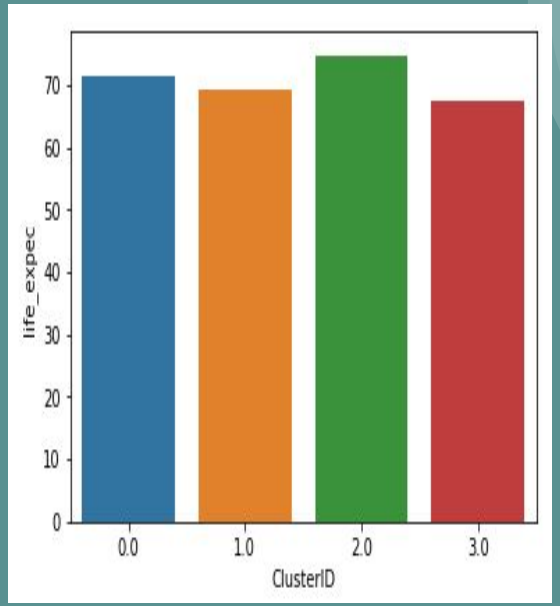


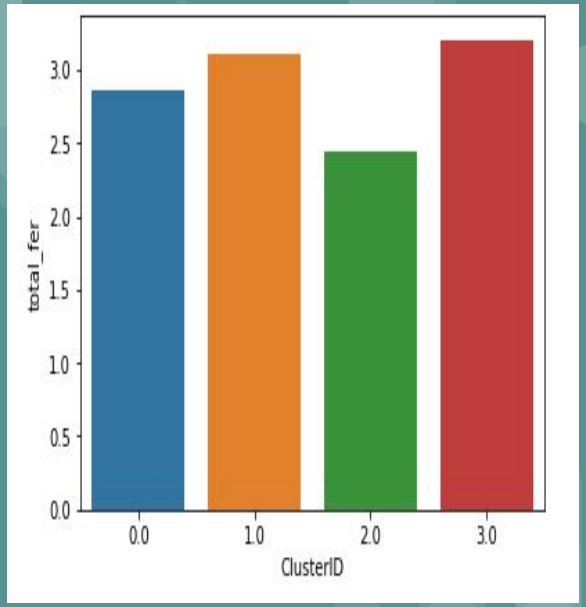


Hierarchical Clustering(k=4)
Analysis of the clusters

Mean Child Mortality, Life Expectancy and Total Fertility Rate by Cluster







Summary

1. K-means Clustering or Hierarchical clustering- With K=4 we get same result for both the methods;
2. Recommendations-
   a. Income and GDP of countries in Cluster in 0 and 3 is comparatively less; Therefore, when allocating money for creating job, skilling people and employment , countries in Cluster 0 and 3 should be focussed.
   b. Inflation is also high in countries in Cluster 3 which is an indicator of bad economy
   c. Child mortality and total fertility rate is also on the higher side in the Cluster 3
   d. Therefore, countries in the Cluster 3 are in dire need of aids

Cluster 3- Armenia, Azerbaijan, Chad, Equatorial Guinea, Haiti, Niger, Nigeria, Nepal etc. are a few countries that belong to least developed countries in the world.

Total no=15 countries

Funding should be provided for education, health, nutrition and disease control. Skilling and employment should also be brought into focus.