PCA and Clustering

ASSIGNMENT SUBMISSION

Name:

Vipul Shrivastava

Background

Problem:

- HELP International is an international humanitarian NGO works for fighting poverty and provides basic amenities and relief to the people of backward countries during the time of disasters and natural calamities.
- NGO raised $ 10 million from various awareness drive and funding program to invest in countries strategically and effectively.
- The problem faced by CEO of HELP international is how to choose the countries that are in the direst need of aid.

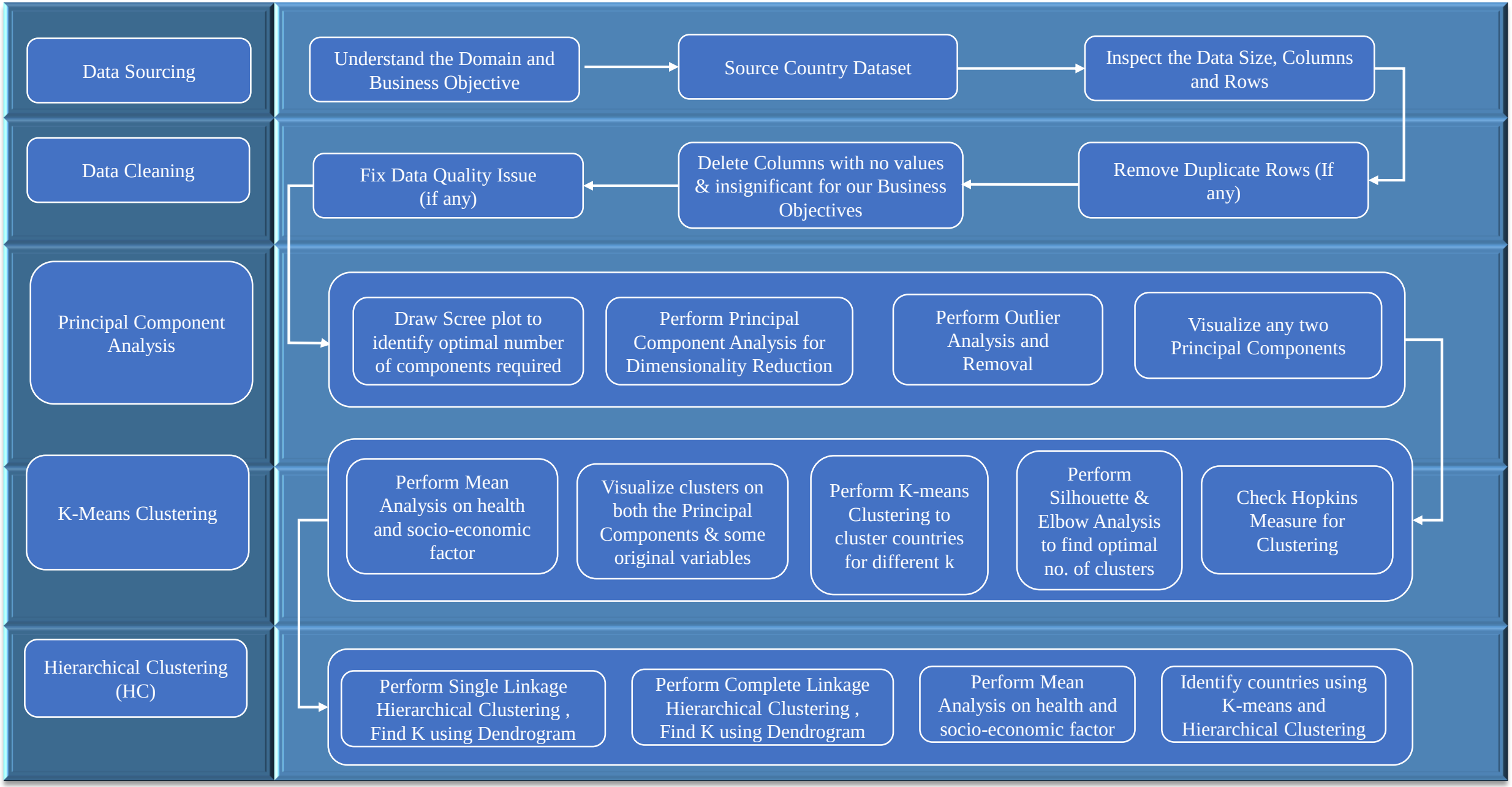
Business Objective:

To identify the countries using some socio-economic and health factors that determine the overall development of the country.

Goals of Analysis:

Categorize those countries into same cluster which has some sort of similarity in their status.

Analysis Approach



Data Understanding

There are 9 socio-economic and health factors (original variables) that determine the overall development of the country.

Column	Description
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.

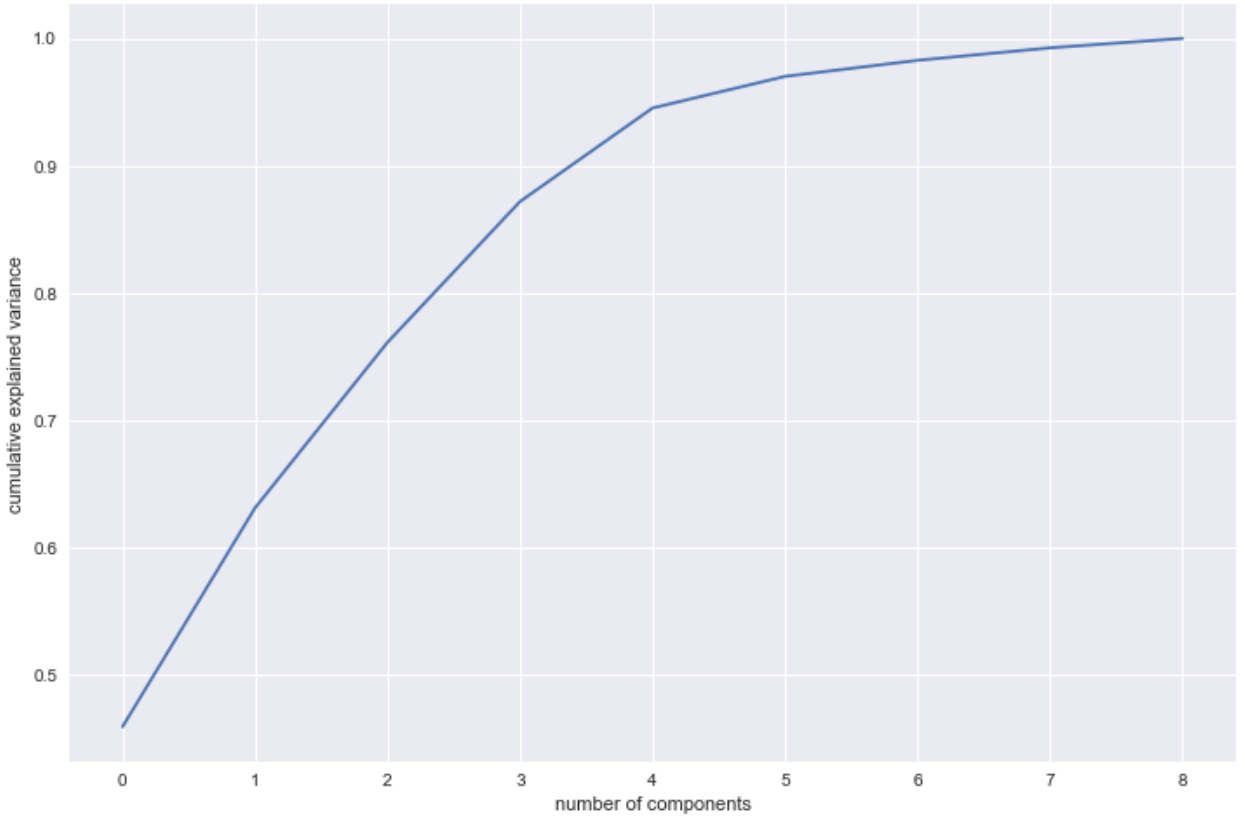
Principal Component Analysis

Principal Component Analysis (PCA)

- Technique used for identification of a smaller number of uncorrelated variables known as **principal components** from a larger set of data.
- It is used to emphasize variation and capture strong patterns in a data set
- It is a Dimensionality Reduction method.

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.
- The PCs are ordered, and by definition are therefore assigned a number label, by decreasing order of contribution to total variance.
- Used to identify optimal number of Principal Components.



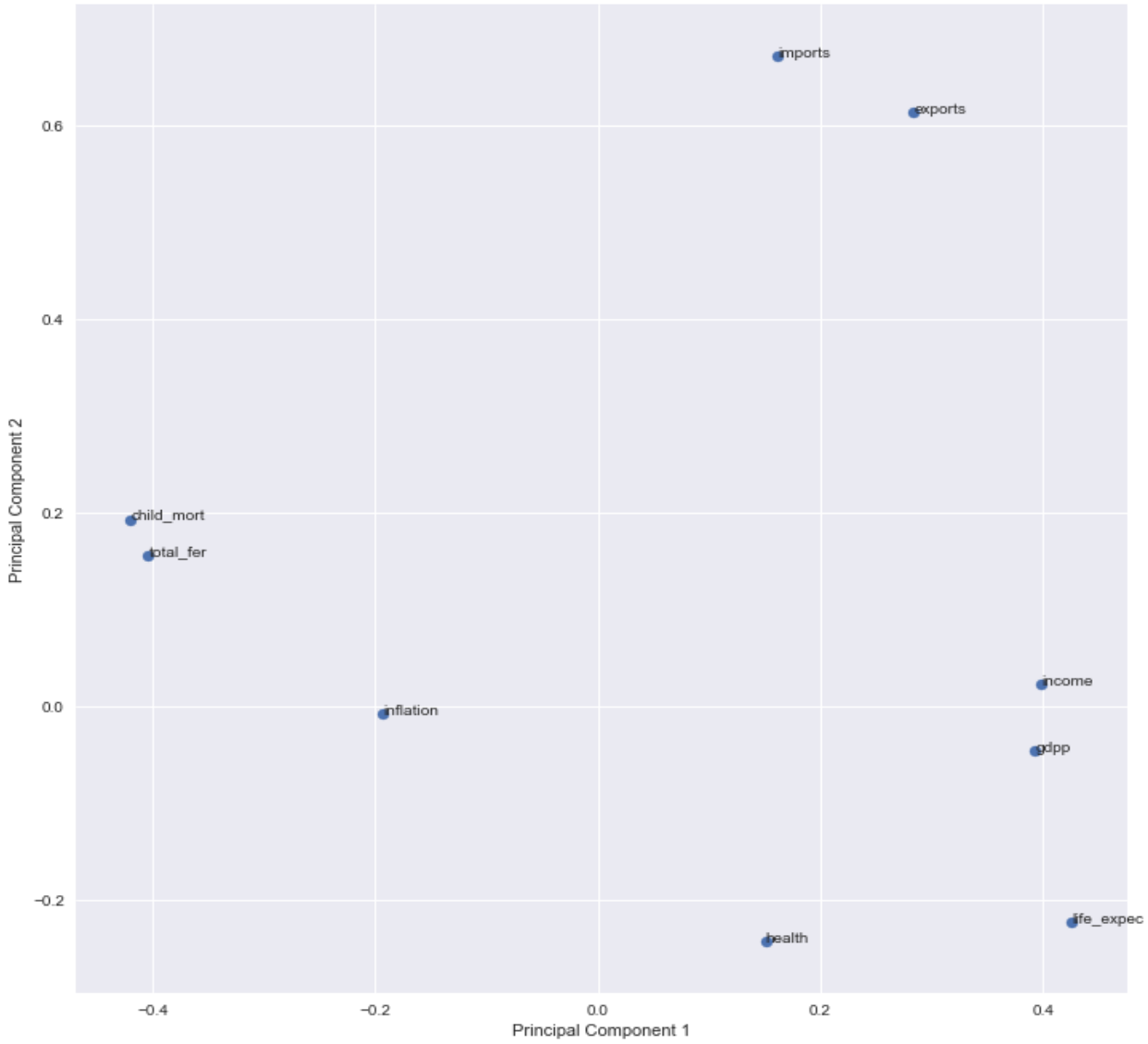
Scree Plot

From the Scree Plot we can see 5 Principal Component can explain more than 95% of variance.

So chosen no. of Principal Component = 5

PCA with No. of Principal Component = 5

- Using Number of Principal Component = 5, Principal Component Analysis is performed.
- We obtained five Principal Component namely PC1, PC2, PC3, PC4 and PC5.
- Visualization is performed for two Principal Component PC1 and PC2 on original variables using scatter plot.
- Scatter plot shows the distribution of different original variables in PC1 and PC2.
- We can also analyse from plot about the original variables clustering tendency.



Scatter Plot for PC1 vs PC2 on original variables

Scatter plot between PC1 and PC2 for different countries

- Image shows the scatter plot between Principal Component 1 and Principal Component 2 for all countries
- We can observe how countries are scattered.
- From the graph we can also identify which countries could belong to which cluster.

K-Means Clustering

K-Means Clustering

- K-means clustering is a simple unsupervised learning algorithm used to solve clustering problems.
- It is used to emphasize variation and capture strong patterns in a data set
- It follows a simple procedure of classifying a given data set into a number of clusters, defined by the letter "k," which is fixed beforehand.

Hopkins Statistics:

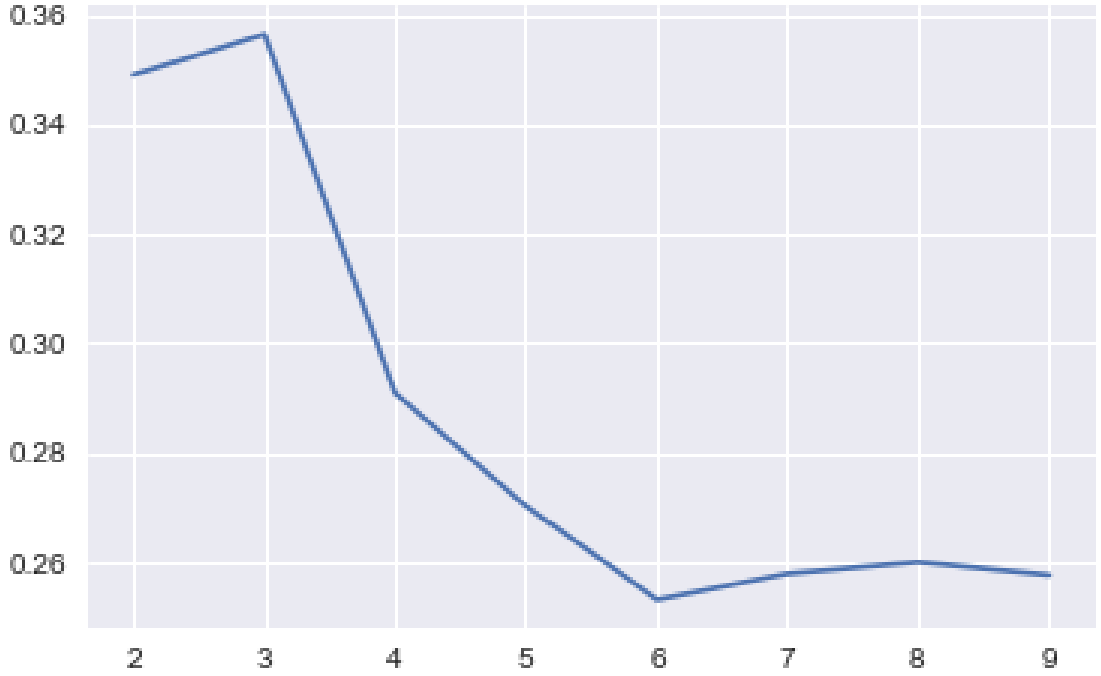
- The Hopkins statistic, is a statistic which gives a value which indicates the cluster tendency, in other words: how well the data can be clustered.
  - If the value is between {0.01, ...,0.3}, the data is regularly spaced.
  - If the value is around 0.5, it is random.
  - If the value is between {0.7, ..., 0.99}, it has a high tendency to cluster.

- When we performed this we got Hopkins measure more than **0.7,** that means the data we have has high tendency to cluster.

Determining optimal number of clusters(k)

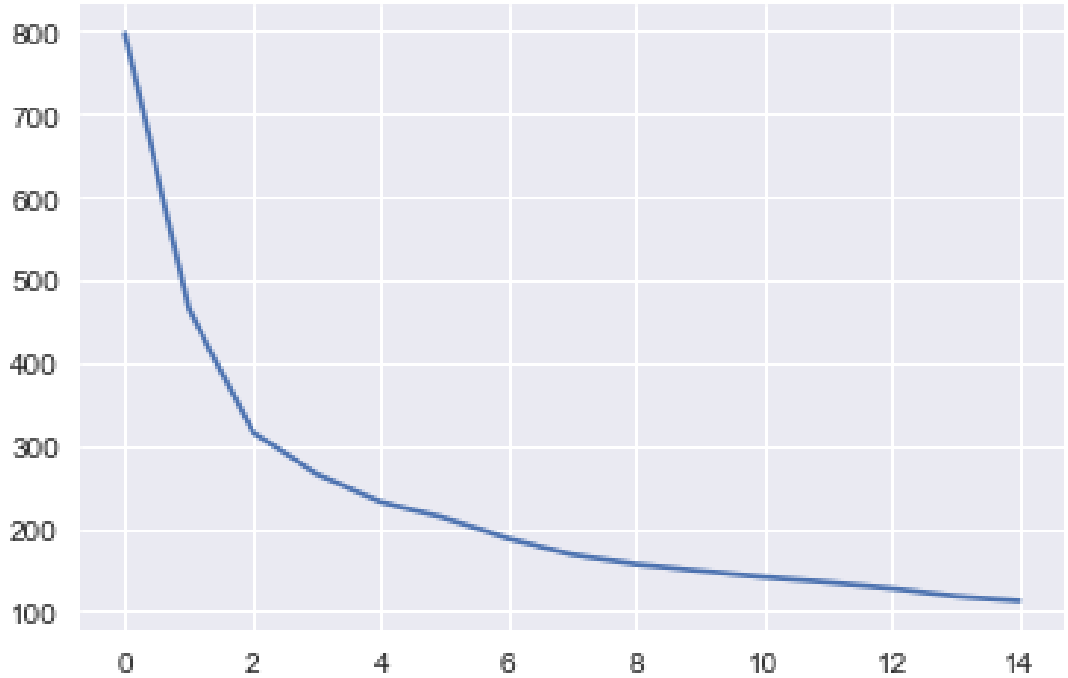
Silhouette Analysis

- 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.



Elbow analysis

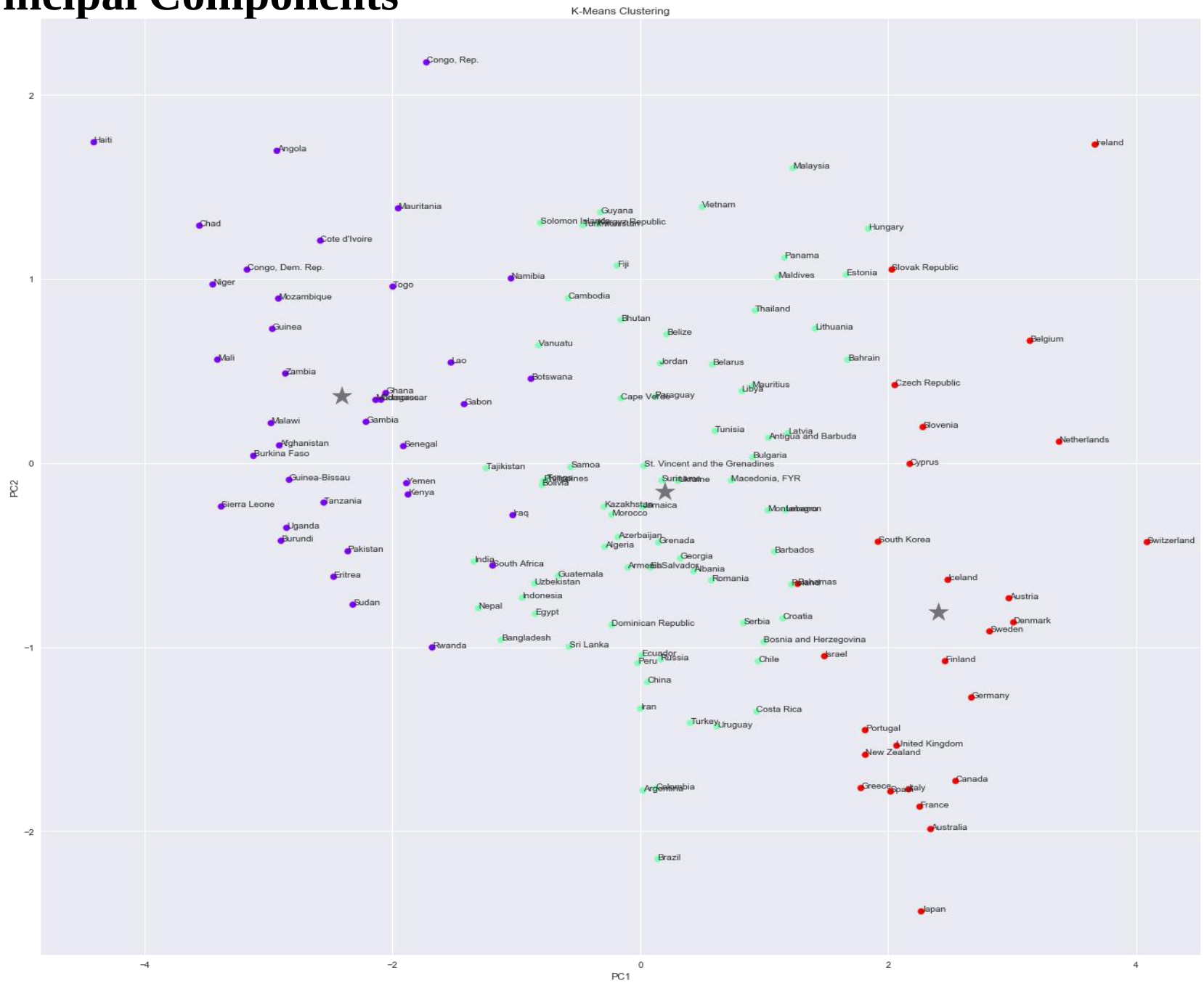
- The graph looks like elbow of human being
- As the number of cluster increases sum of squared distance decreases.



From Silhouette Analysis and Elbow analysis we identify optimal number of cluster(k) = 3

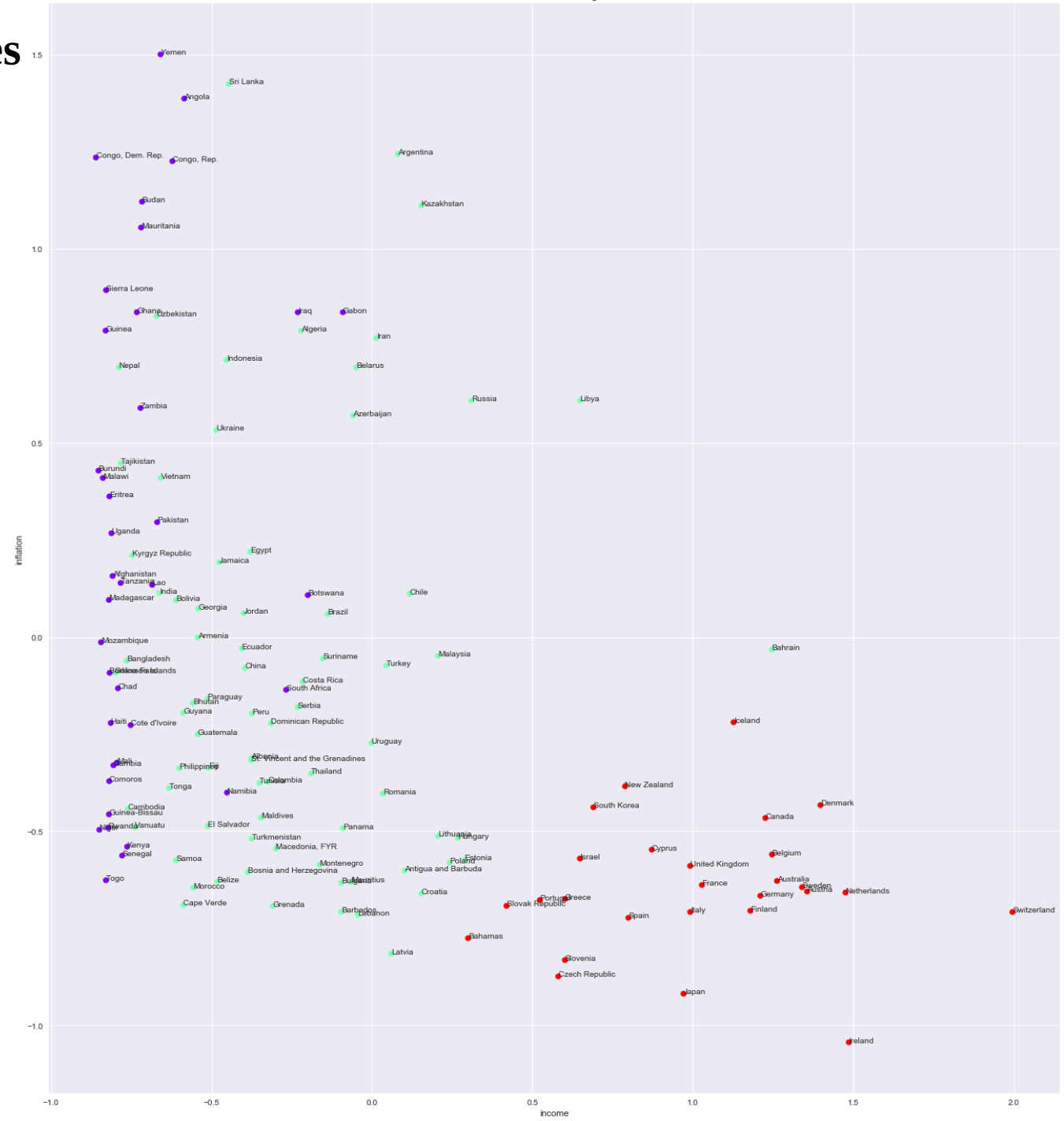
Visualization of clusters on Principal Components

- Using number of cluster = 3, we performed K-means clustering on identified principal components.
- The clusters are visualized on principal component PC1 and PC2
- From the graph we can observe different countries are in different colour and star are the cluster centres.
- The countries having same colours belongs to same clusters and have some pattern in common.



Visualization of clusters on Original Variables (Income and inflation)

- The clusters are visualized on original variables income and inflation
- From the graph we can observe different countries are in different colour and star are the cluster centres. The countries having same colours belongs to same clusters and have some pattern in common.
- We can observe that when net income per person (income) is low, inflation is high and these are collected into one cluster.



Result of Clustering

K-Means Clustering

- When we performed clustering with k=3, we found 38 countries belong to cluster 0(first cluster), 76 countries belong to cluster 1(second cluster) and 27 countries belong to cluster 2 (third cluster)

- Graph shows 38 countries belong to cluster 0.

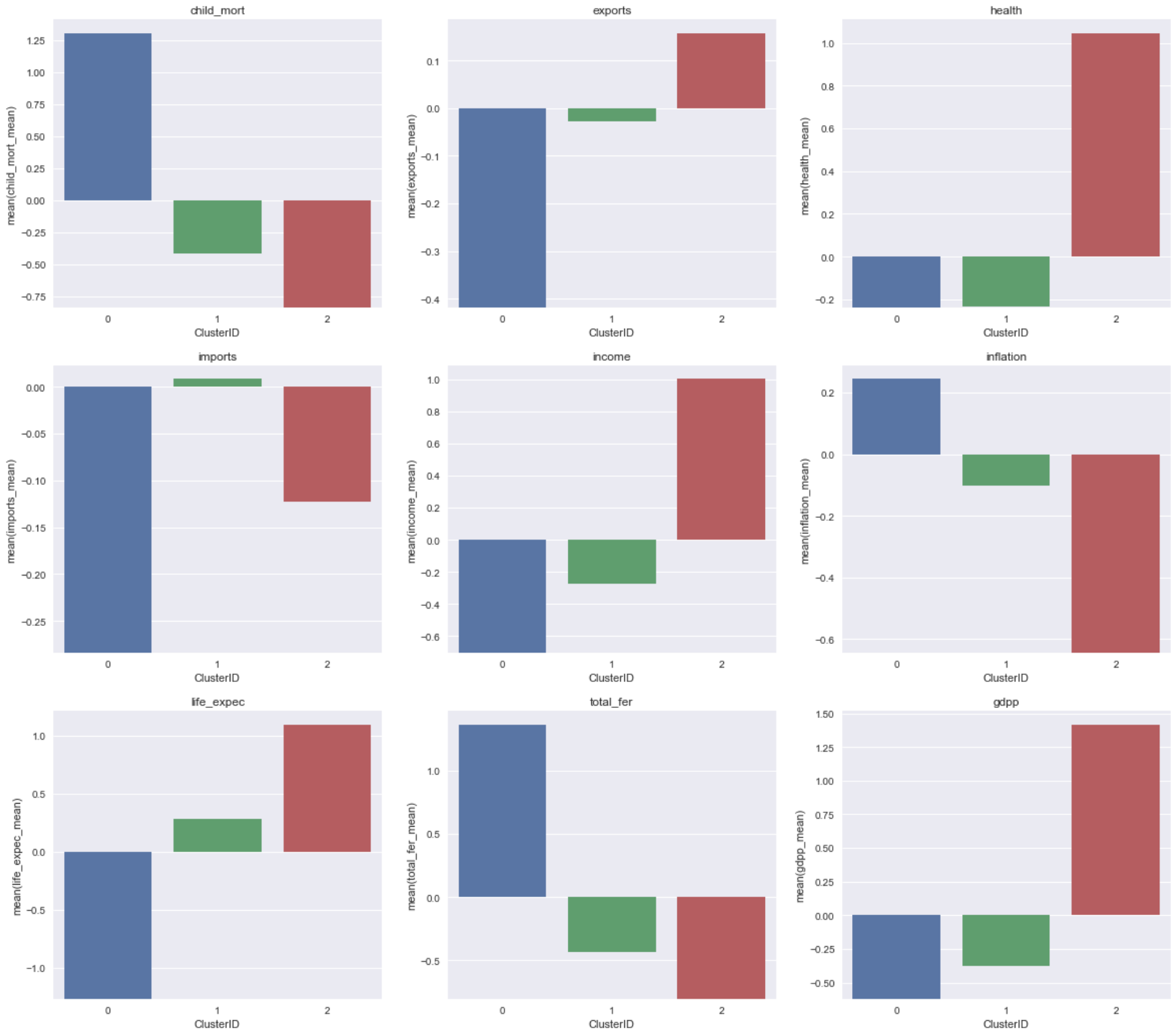
	country
ClusterId	
0	38
1	76
2	27

	ClusterId	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	0	Afghanistan	1.291532	-1.138280	0.279088	-0.082455	-0.808245	0.157336	-1.619092	1.902882	-0.679180
3	0	Angola	2.007808	0.775381	-1.448071	-0.165315	-0.585043	1.387054	-1.179234	2.128151	-0.516268
21	0	Botswana	0.353908	0.091147	0.541996	0.182698	-0.200033	0.107996	-1.517586	-0.045030	-0.361949
25	0	Burkina Faso	1.933196	-0.801651	-0.027638	-0.716337	-0.817611	-0.092213	-1.427359	1.936010	-0.677976
26	0	Burundi	1.376093	-1.177797	1.746991	-0.318607	-0.852261	0.428709	-1.449916	2.194407	-0.696801
32	0	Chad	2.778798	-0.157666	-0.834619	-0.140457	-0.791596	-0.132065	-1.585257	2.413050	-0.660355
36	0	Comoros	1.241791	-0.900444	-0.841922	0.199270	-0.818651	-0.371177	-0.525087	1.193948	-0.667359
37	0	Congo, Dem. Rep.	1.933196	-0.000328	0.399588	0.112267	-0.860326	1.235237	-1.472473	2.379922	-0.691164
38	0	Congo, Rep.	0.637434	1.609635	-1.590479	0.323561	-0.621984	1.225748	-1.145399	1.326459	-0.559500
40	0	Cote d'Ivoire	1.808842	0.347277	-0.553454	-0.148743	-0.752055	-0.226950	-1.607814	1.538477	-0.642679
50	0	Eritrea	0.421059	-1.328914	-1.517449	-0.977347	-0.818131	0.362289	-0.998780	1.101191	-0.683065
55	0	Gabon	0.632460	0.607067	-1.210723	-1.159640	-0.090773	0.836717	-0.863439	0.750036	-0.230613
56	0	Gambia	1.045313	-0.633337	-0.411045	-0.173601	-0.805644	-0.330376	-0.570201	1.830001	-0.678687
59	0	Ghana	0.906037	-0.424773	-0.582666	-0.041025	-0.732804	0.836717	-0.942388	0.875922	-0.637754
63	0	Guinea	1.759101	-0.395501	-0.688559	-0.152886	-0.830097	0.789274	-1.416081	1.584856	-0.673981
64	0	Guinea-Bissau	1.883454	-0.958988	0.615026	-0.484328	-0.819692	-0.456574	-1.686762	1.392715	-0.679508
66	0	Haiti	4.221297	-0.944352	0.034438	0.737863	-0.813969	-0.221257	-4.337186	0.253120	-0.673215
72	0	Iraq	-0.034074	-0.062532	0.582163	-0.529901	-0.231250	0.836717	-0.378468	1.068063	-0.463187
80	0	Kenya	0.595154	-0.746766	-0.754286	-0.550616	-0.762981	-0.540073	-0.874718	0.942177	-0.656524
84	0	Lao	1.010494	-0.208892	-0.856528	0.099838	-0.684938	0.134564	-0.761934	0.133860	-0.647057
93	0	Madagascar	0.595154	-0.589429	-1.112133	-0.161172	-0.819692	0.095661	-1.100286	1.094565	-0.686841
94	0	Malawi	1.298994	-0.669927	-0.082410	-0.496757	-0.838422	0.409732	-1.968722	1.564979	-0.684324
97	0	Mali	2.455480	-0.669927	-0.670302	-0.488471	-0.794718	-0.323734	-1.246905	2.386548	-0.670697
99	0	Mauritania	1.470601	0.350936	-0.878437	0.592857	-0.719277	1.054954	-0.265684	1.346336	-0.643774
106	0	Mozambique	1.560136	-0.351593	-0.586317	-0.028596	-0.844249	-0.013458	-1.810825	1.730618	-0.686513
108	0	Namibia	0.440955	0.244825	-0.013032	0.572142	-0.451851	-0.400591	-1.348410	0.432010	-0.425428
112	0	Niger	2.107290	-0.691881	-0.604575	0.091552	-0.849660	-0.496426	-1.325854	3.009349	-0.690398
116	0	Pakistan	1.338787	-1.010215	-1.685418	-1.138924	-0.669330	0.295869	-0.592758	0.597649	-0.652529
126	0	Rwanda	0.629973	-1.065100	1.345326	-0.699765	-0.821773	-0.490732	-0.671706	1.034935	-0.678632
129	0	Senegal	0.709559	-0.593088	-0.422000	-0.273034	-0.778589	-0.562845	-0.739377	1.399340	-0.654718
132	0	Sierra Leone	3.027505	-0.889467	2.294716	-0.513329	-0.828536	0.893648	-1.754433	1.492098	-0.687607
137	0	South Africa	0.383753	-0.457704	0.775692	-0.807483	-0.267670	-0.135860	-1.833382	-0.237171	-0.311056
142	0	Sudan	0.955778	-0.783356	-0.181001	-1.230071	-0.716675	1.121374	-0.479974	1.280080	-0.628451
147	0	Tanzania	0.836399	-0.819946	-0.294197	-0.737052	-0.783272	0.139308	-1.269462	1.644486	-0.671026
150	0	Togo	1.294019	-0.033260	0.304649	0.431279	-0.829057	-0.626419	-1.337132	1.273455	-0.682737
155	0	Uganda	1.062722	-0.878490	0.801253	-0.757767	-0.811887	0.267404	-1.551422	2.121525	-0.676881
165	0	Yemen	0.448417	-0.406478	-0.597272	-0.517472	-0.658924	1.500916	-0.344633	1.140944	-0.637754
166	0	Zambia	1.114951	-0.150348	-0.338015	-0.662477	-0.721358	0.590015	-2.092785	1.624609	-0.629546

Result of Clustering

Mean analysis on original variable

- We performed mean analysis on original variables for each cluster i.e. clusters 0,1 and 2.
- From these graph we can observe that countries belong to cluster 0 has
  - High child_mort
  - Low exports
  - Low spending on health
  - Low income per person
  - High inflation
  - Low life_exp
  - Very lof gdpp
- On above criteria we can say countries belong to cluster 0 need more funding.



Conclusion

- We performed Principal Component Analysis with number of component =5
- We performed k-means clustering for k=3, And found that
  - 38 countries get clustered into Cluster 1 (ClusterId = 0)
  - 76 countries get clustered into Cluster 2 (ClusterId = 1)
  - 27 countries get clustered into Cluster 3 (ClusterId = 2)
- We observed that countries belong to Cluster 1(ClusterId=0) needs more amount of funding for their overall development.
- HELP international need to focus on the countries having
  - High child_mort
  - Low exports
  - Low spending on health
  - Low income per person
  - High inflation
  - Low life_exp
  - Very lof gdpp
- Some of the countries need more funding are:

Afghanistan, Angola, Botswana, Burkina Faso, Burundi, Chad, Comoros, Congo, Cote d'Ivoire, Ghana, Guinea, Guinea-Bissau, Haiti, Iraq, Kenya, Lao, Madagascar, Malawi, Mali, Mauritania, Mozambique, Namibia, Niger, Pakistan, Rwanda, Senegal, Sierra Leone, South Africa, Sudan, Tanzania, Uganda, Yemen, Zambia