PCA and Clustering Assignment

OBJECTIVE

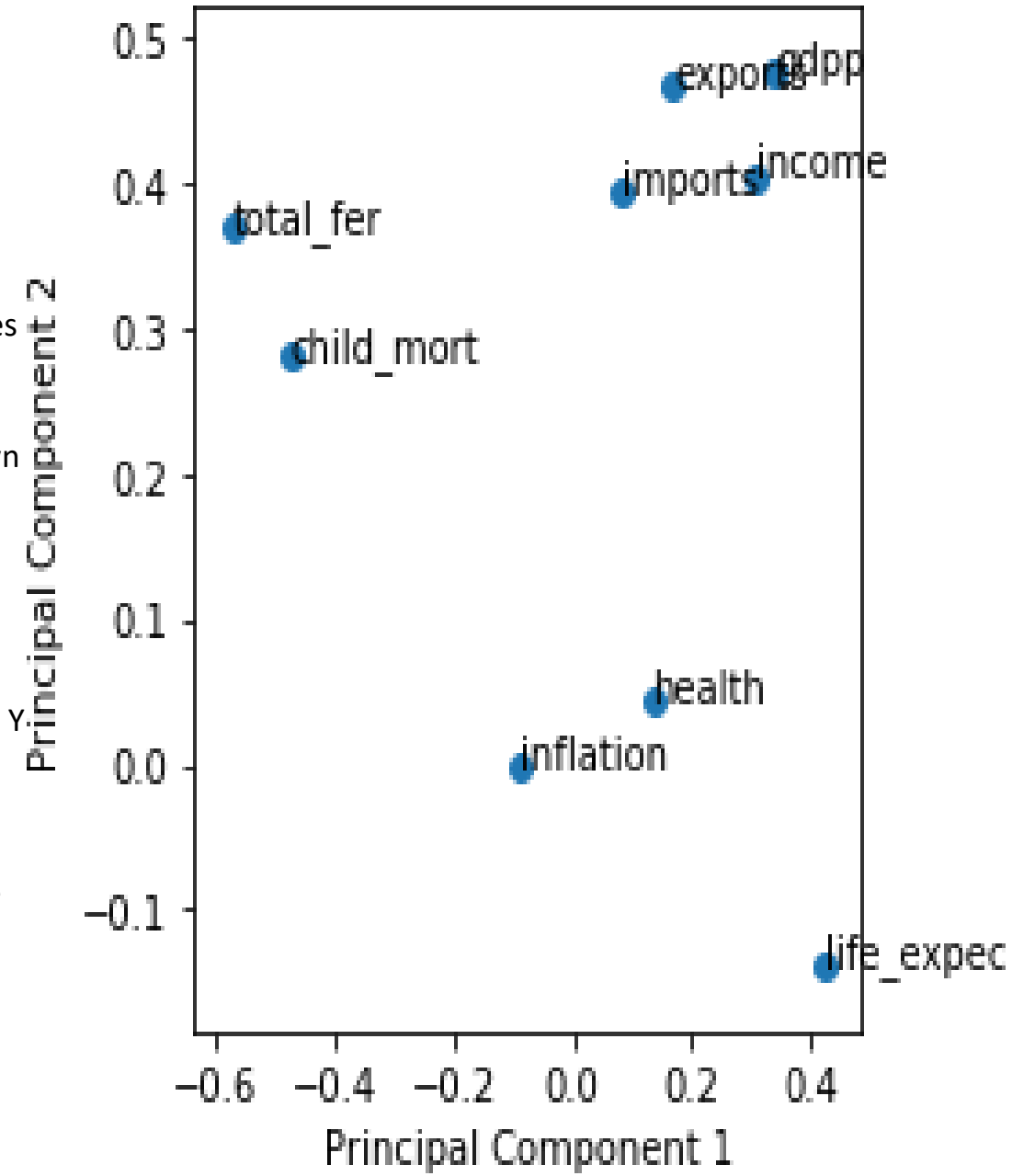
- Help International is an International Humanitarian NGO committed to fight poverty and the people of backward countries with basic amenities and relief during the time of disasters and natural calamities.
- After the recent funding programs, they have been able to raise around $10 million.
- The objective of the assignment is to categorize the countries using some socio-economic and health factors that determine the overall development of the country and suggest the countries which the CEO needs to focus on are the most in dire need of aid.

METHOD

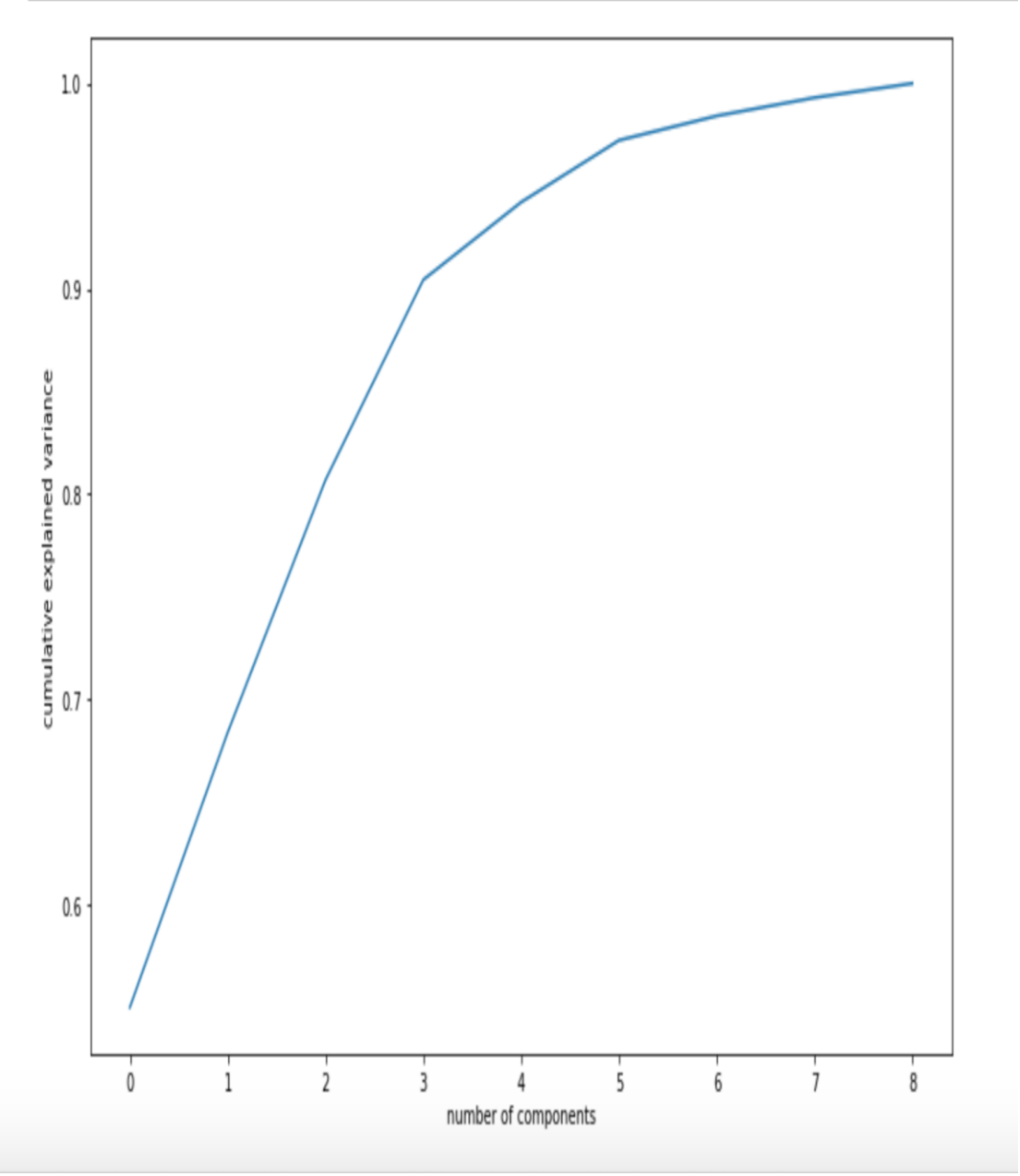
1) Perform Principal Component Analysis (PCA) on the dataset and obtain a new dataset with the principal components.
2) Perform clustering (K-means and Hierarchical) on the new dataset and create clusters.
3) Analyze the clusters and identify the ones which are in dire need of aid. This is done by the method of comparing the variability of the original variables deferring for each cluster and isolating the clusters of developed countries from the clusters of under-developed countries.
4) Additionally, visualizations need to be performed on the clusters by choosing the first two principal components and plotting a scatter plot of all the countries and separating the clusters.
5) The same visualization need to be done on various kinds of plots using any two of the original variables such as GDPP, child mortality and so on.
6) The final list of countries depending on the number of the components and the number of clusters formed must be reported back to the CEO.

PCA

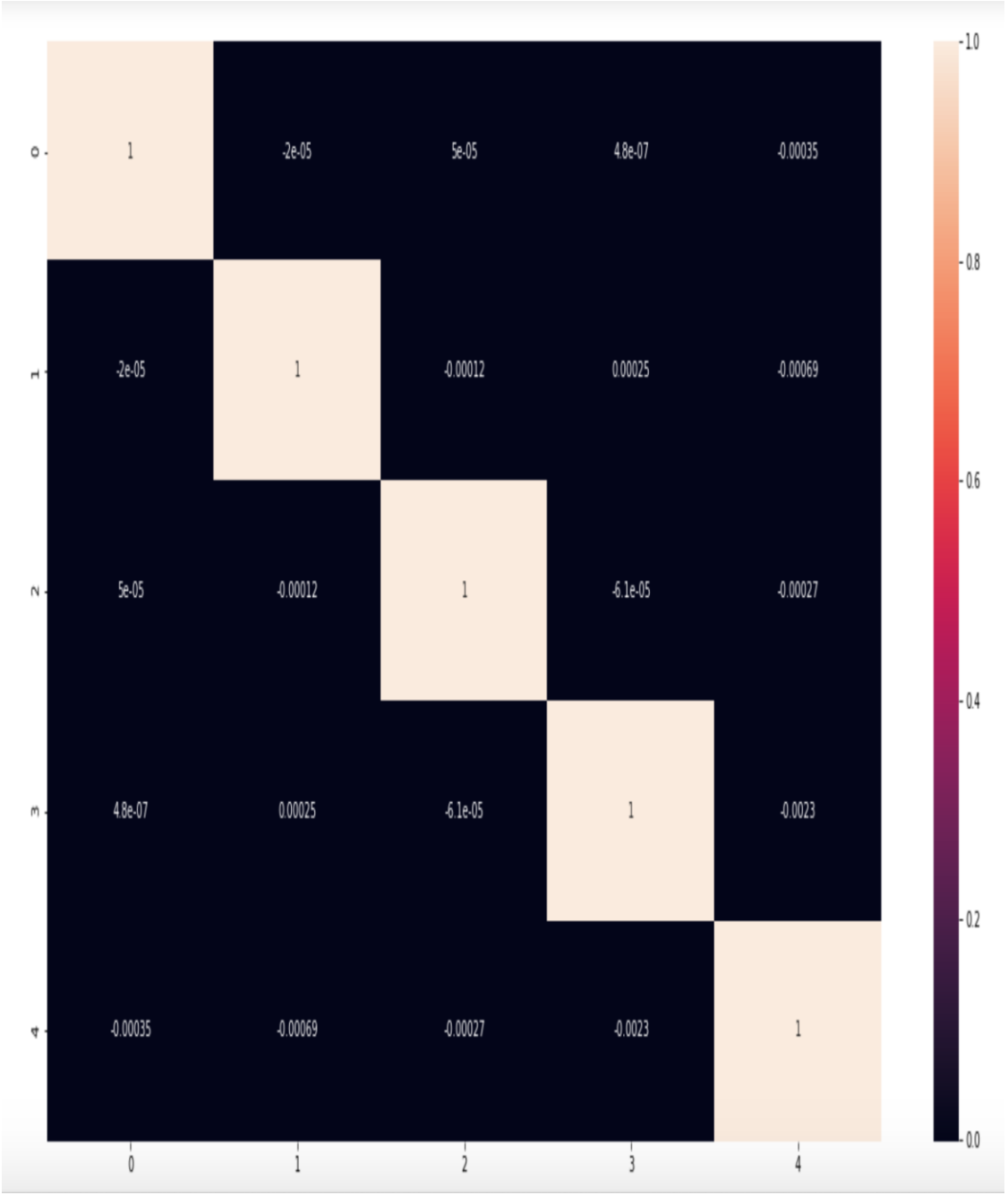
- Principal component analysis is the process to draw out variables as components from a large set of variables of a dataset.
- Generally, it is employed on a symmetric correlation matrix also known as covariance matrix.
- Randomized singular value decomposition solver is applied on the dataset.
- Original features are plotted on the first 2 principal components as axes (X - Y axes) as shown in the figure.
- Figure illustrates that the exports and gdpp are the most influential factors followed by imports and income (hence 2-4 PC's).



- The basic technique to compute the principal components is to plot the original features on the first 2 principal components as axes.
- This helps in choosing the “optimal” number of principal components ,known as a **“Scree plot”**
- The y-axis represents the cumulative variance captured and the x-axis represents the number of principal components.
- We can infer from the scree plot shown in the figure that 5 components are adequate to describe around 97% of the variance in the dataset.
- Hence 5 components are selected for modelling.

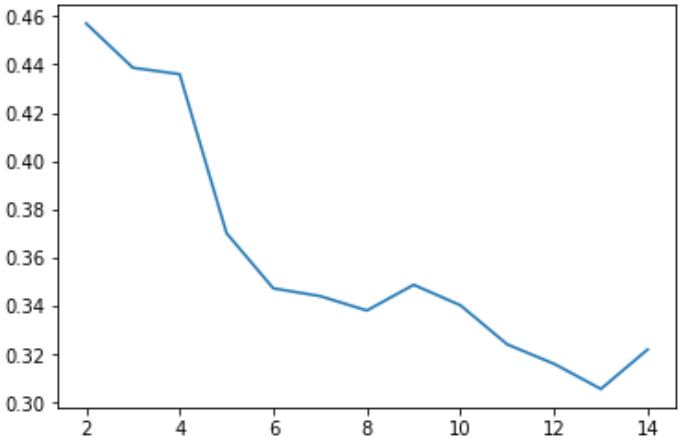


- The next step is getting the data onto the principal components by basis transformation.
- In order to achieve this, a correlation matrix of the principal components was created.
- We can infer from the heatmap that there is little but no correlation.
- Lastly, the column country from the original dataset was merged with the new dataset on which PCA was applied.



K-Means Clustering

- K-means clustering begins with computing Hopkins statistic.
- Hopkins statistic, is a statistic which gives a values which indicates the cluster tendency or in other words, how well the data can be clustered.



- In order to determine the number of clusters, silhouette score is calculated using the mean intra-cluster distance (p) and the mean nearest-cluster distance (q) for each sample.
- Silhouette score = $\frac{p-q}{\max(p,q)}$
- 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 for k=2

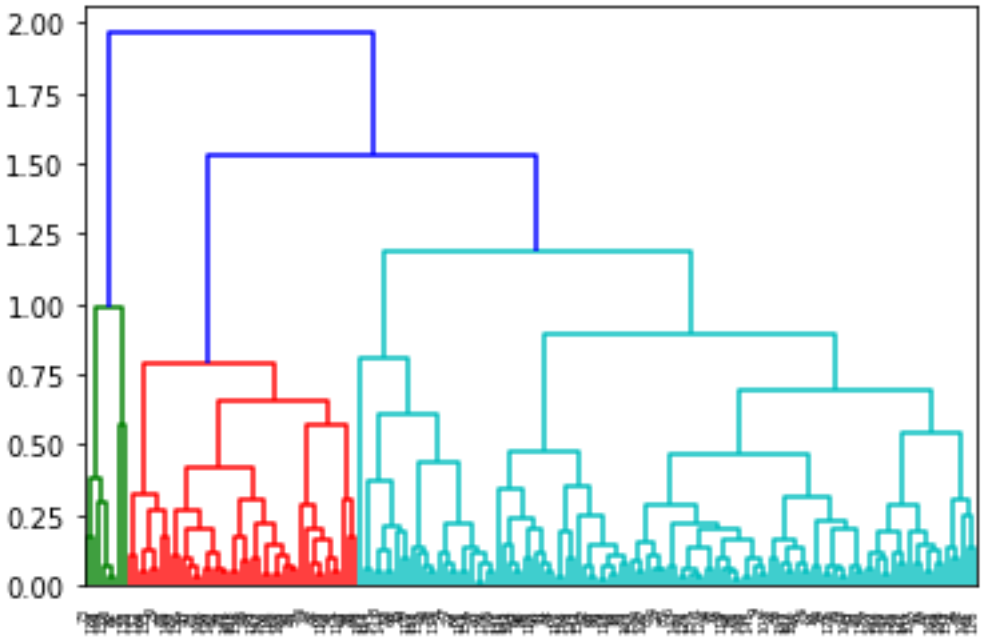
- The clusters are now analyzed to identify the countries which are in dire need of aid.
- The final list of countries depends on the number of components chosen and the number of clusters formed.
- The output is seen in the image.

```
countries.sort_values(by=['gdpp','income']).head(10)
```

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	ClusterID
26	Burundi	0.443038	0.044079	0.608452	0.224994	0.001246	0.152574	0.504931	0.805994	0.000000	1
88	Liberia	0.422103	0.095007	0.620883	0.532007	0.000732	0.089456	0.566075	0.610410	0.000916	1
37	Congo, Dem. Rep.	0.552093	0.205067	0.379117	0.284787	0.000000	0.231125	0.500986	0.850158	0.000983	1
112	Niger	0.586173	0.110515	0.208204	0.281912	0.001648	0.062471	0.526627	1.000000	0.001117	1
132	Sierra Leone	0.786310	0.083501	0.701678	0.197972	0.004912	0.197856	0.451677	0.838801	0.001604	1
93	Madagascar	0.290166	0.124523	0.121815	0.246841	0.006279	0.120137	0.566075	0.544164	0.001737	1
106	Mozambique	0.479065	0.157041	0.211311	0.265239	0.002484	0.109509	0.441815	0.695584	0.001794	1
31	Central African Republic	0.712756	0.058487	0.134866	0.151978	0.002243	0.057481	0.303748	0.640379	0.002052	1
94	Malawi	0.427945	0.113517	0.297079	0.200272	0.003384	0.150725	0.414201	0.656151	0.002176	1
50	Eritrea	0.256086	0.023418	0.052828	0.133580	0.006520	0.146105	0.583826	0.545741	0.002396	1

HIERARCHICAL CLUSTERING

- The output of the hierarchical clustering algorithm is an inverted tree-shaped structure, called the dendrogram.
- This algorithm starts with all the data points assigned to a cluster of their own. Then two nearest clusters are merged into the same cluster. In the end, this algorithm terminates when there is only a single cluster left.
- There are types known as single,complete linkages.



HIERARCHICAL CLUSTERING

In [125]:
```
set1=countries.sort_values(by=['gdpp','income']).head(10)
set1
```

Out[125]:

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	ClusterID
88	Liberia	89.3	19.10	11.80	92.6	700	5.47	60.8	5.02	327	1
37	Congo, Dem. Rep.	116.0	41.10	7.91	49.6	609	20.80	57.5	6.54	334	1
112	Niger	123.0	22.20	5.16	49.1	814	2.55	58.8	7.49	348	1
106	Mozambique	101.0	31.50	5.21	46.2	918	7.64	54.5	5.56	419	1
31	Central African Republic	149.0	11.80	3.98	26.5	888	2.01	47.5	5.21	446	1
64	Guinea-Bissau	114.0	14.90	8.50	35.2	1390	2.97	55.6	5.05	547	1
56	Gambia	80.3	23.80	5.69	42.7	1660	4.30	65.5	5.71	562	1
126	Rwanda	63.6	12.00	10.50	30.0	1350	2.61	64.6	4.51	563	1
109	Nepal	47.0	9.58	5.25	36.4	1990	15.10	68.3	2.61	592	1
116	Pakistan	92.1	13.50	2.20	19.4	4280	10.90	65.3	3.85	1040	1

In [127]:
```
set2=countries.sort_values(by=['child_mort'],ascending=False).head(10)
set2
```

Out[127]:

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	ClusterID
31	Central African Republic	149.0	11.8	3.98	26.5	888	2.01	47.5	5.21	446	1
112	Niger	123.0	22.2	5.16	49.1	814	2.55	58.8	7.49	348	1
37	Congo, Dem. Rep.	116.0	41.1	7.91	49.6	609	20.80	57.5	6.54	334	1
64	Guinea-Bissau	114.0	14.9	8.50	35.2	1390	2.97	55.6	5.05	547	1
106	Mozambique	101.0	31.5	5.21	46.2	918	7.64	54.5	5.56	419	1
87	Lesotho	99.7	39.4	11.10	101.0	2380	4.15	46.5	3.30	1170	1
116	Pakistan	92.1	13.5	2.20	19.4	4280	10.90	65.3	3.85	1040	1
88	Liberia	89.3	19.1	11.80	92.6	700	5.47	60.8	5.02	327	1
56	Gambia	80.3	23.8	5.69	42.7	1660	4.30	65.5	5.71	562	1
55	Gabon	63.7	57.7	3.50	18.9	15400	16.60	62.9	4.08	8750	1

Summary

- Countries output from clustering can be concentrated on.
- The socio-economic factors like gpp,income,child_mort are chosen.
- The higher child_mort and low income and gpp countries selected.
- Reasons on why the child mortality is higher should be found.