Clustering & PCA Assignment

Prateek Awadhut Gharat

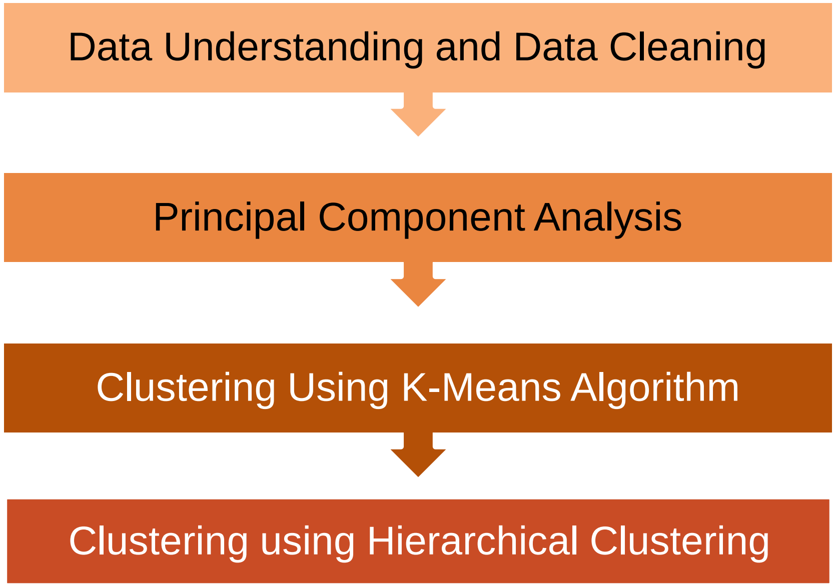
Abstract

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. After the recent funding programs, they have been able to raise around $ 10 million.

The significant issues that come while making this decision are mostly related to choosing the countries that are in the direst need of aid.

To categorize 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.

Methodology



Data Understanding and Data Cleaning

Principal Component Analysis

Clustering Using K-Means Algorithm

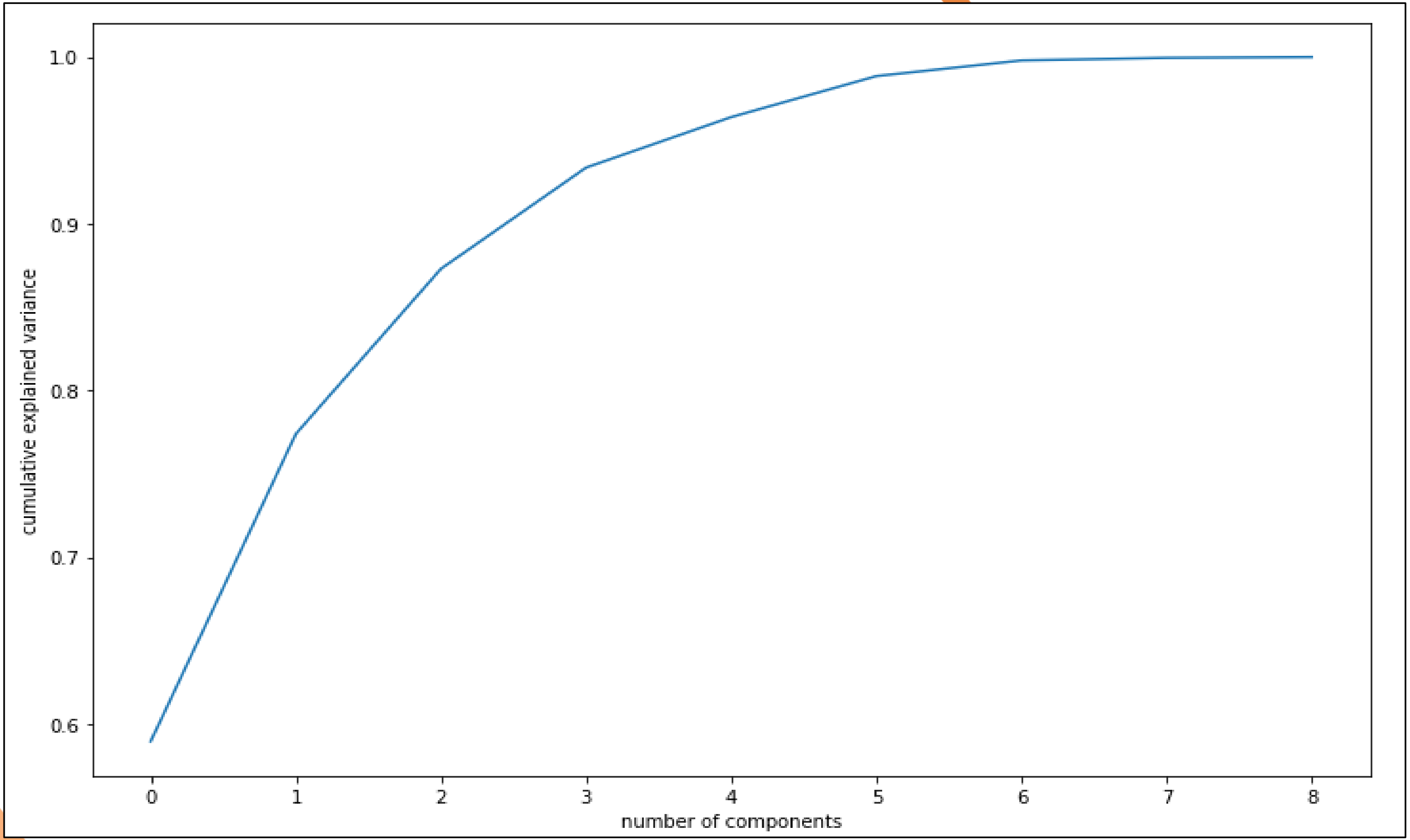
Clustering using Hierarchical Clustering

- Data understanding and cleaning includes the variables understanding and treatment of data as per the requirement.
- Converting the data in desired format and application of normalisation and standardisation.
- Identification of Principal Component which explain around 90% of the data.
- Application of PCA and outlier treatment.
- Identification of k values and application of K-Means clustering.
- Identification of clusters by dendrogram

Principal Component Analysis

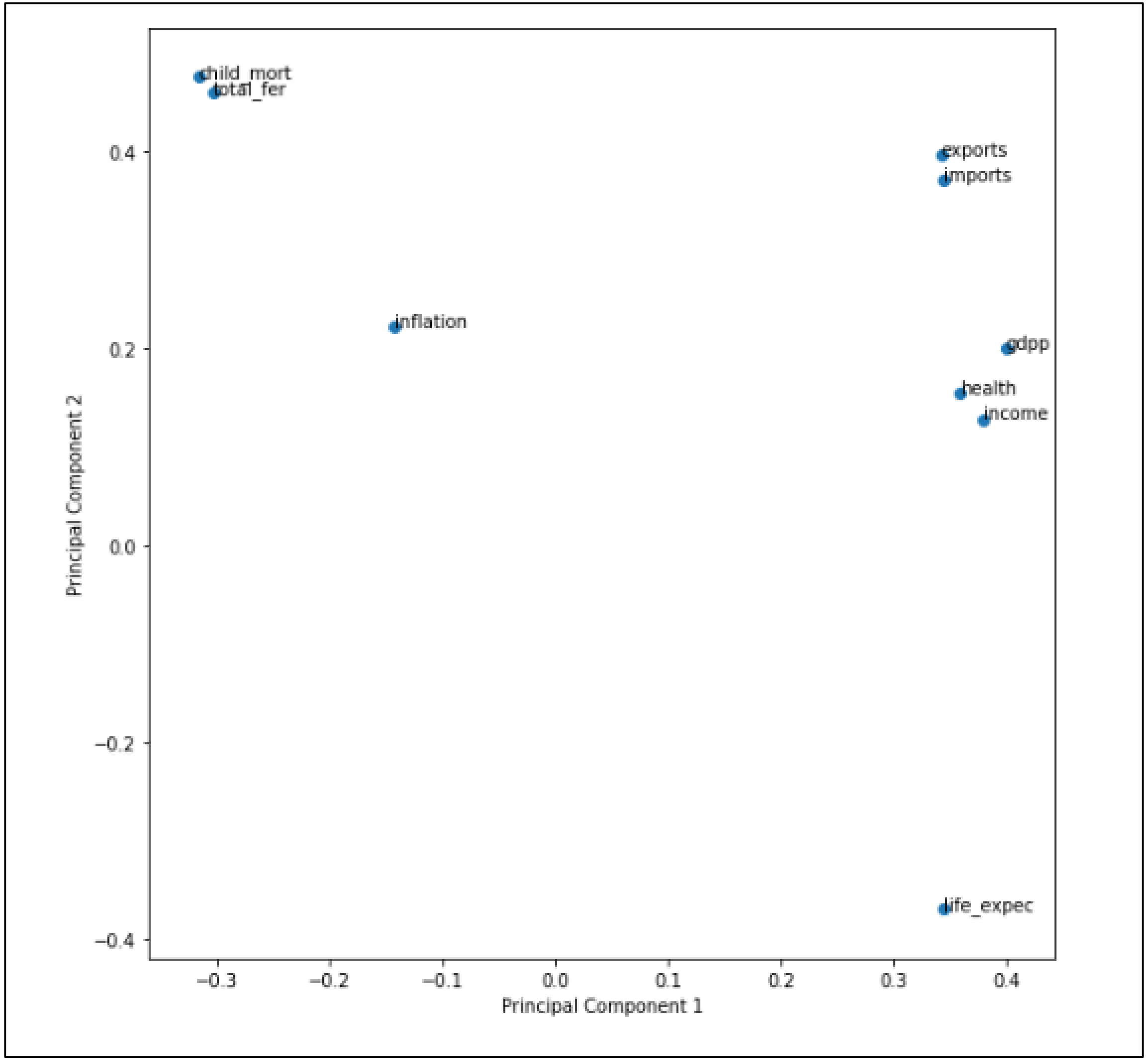
- Principal component analysis is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.
- Standardizing the dataset for PCA implementation.
- Verification of the variance ratio of Principal components.
- Plotting the cumulative variance against the number of components.
- Identification of number of PC which covers around 90% variance in data.
- Built the PCA model using the 3 components (which represents maximum variance).
- Treatment of outliers.

Data Visualization



90% of information is being explained by first 3 components.

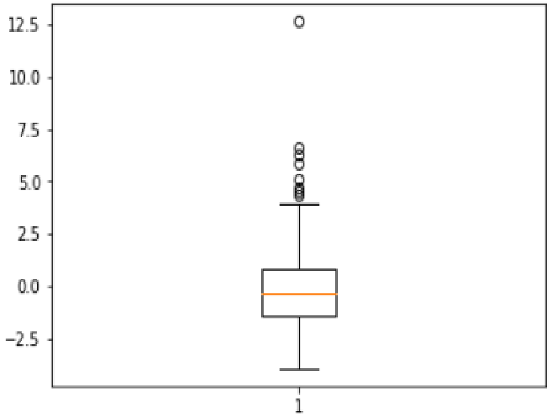
Data Visualization



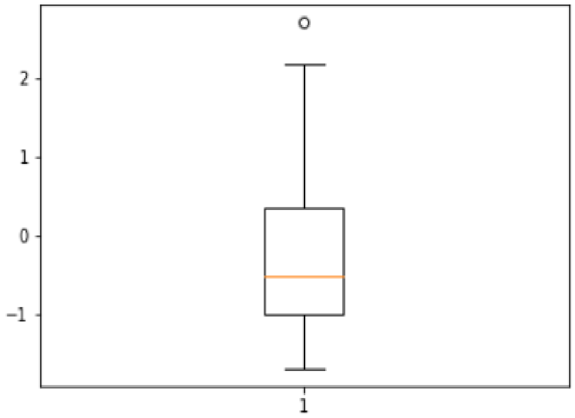
Principal components explaining the variance

Data Visualization

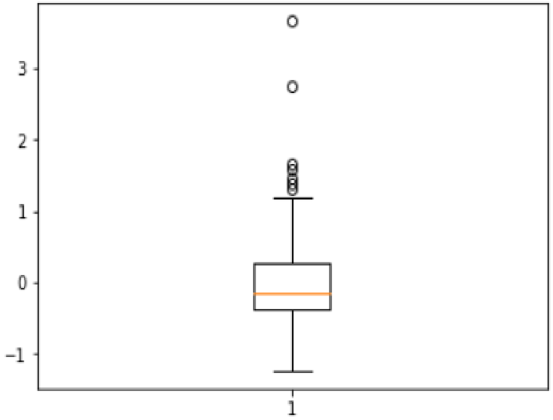
Outlier Treatment



Principal Component 1

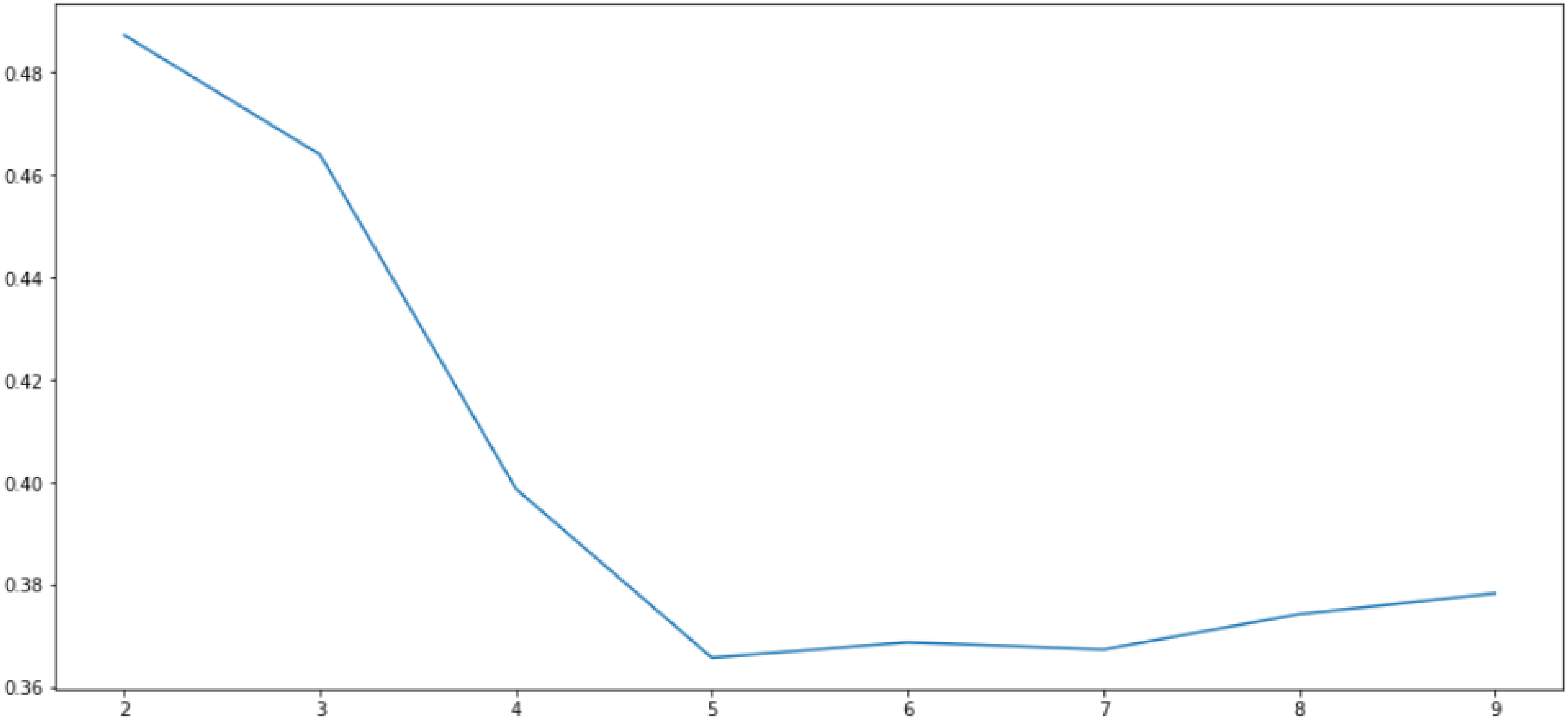


Principal Component 2



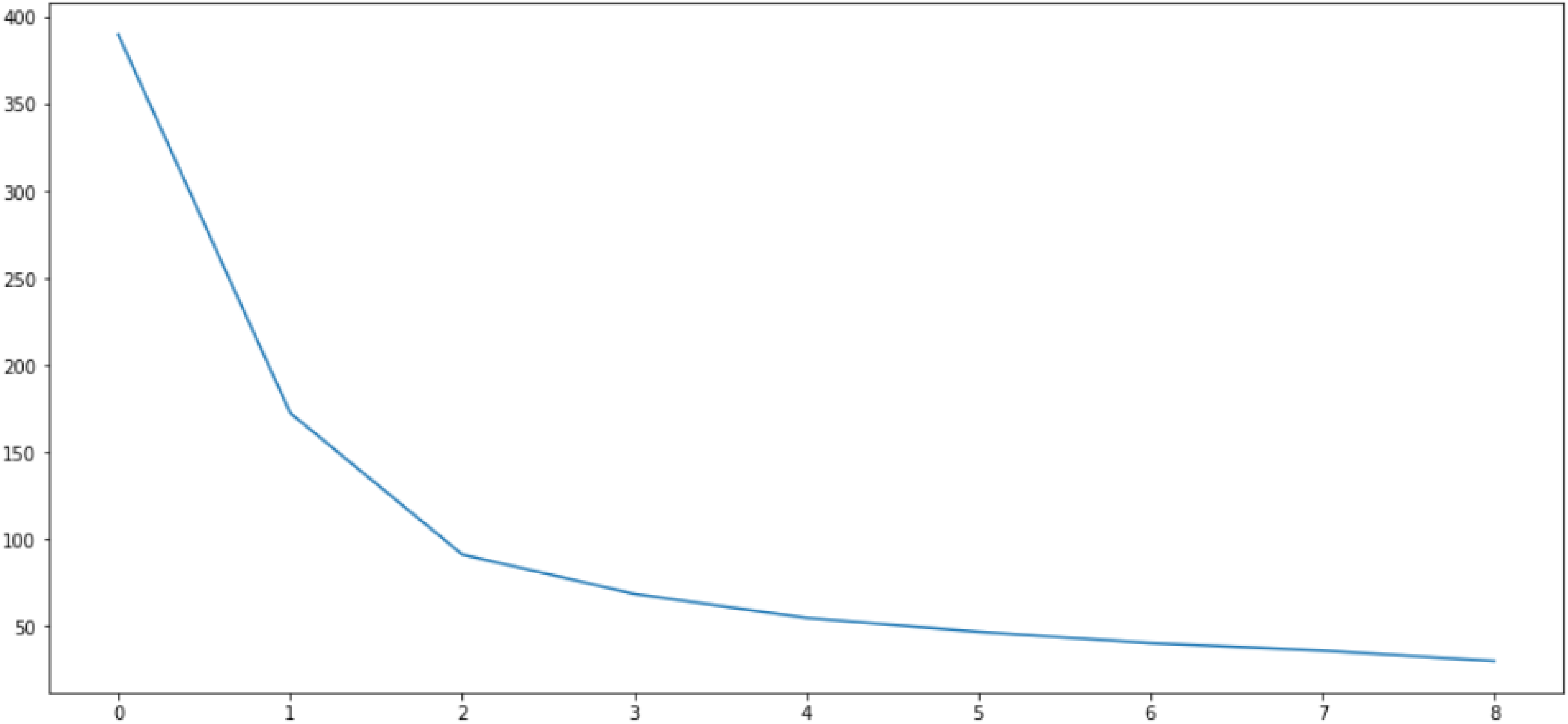
Principal Component 3

Data Visualization



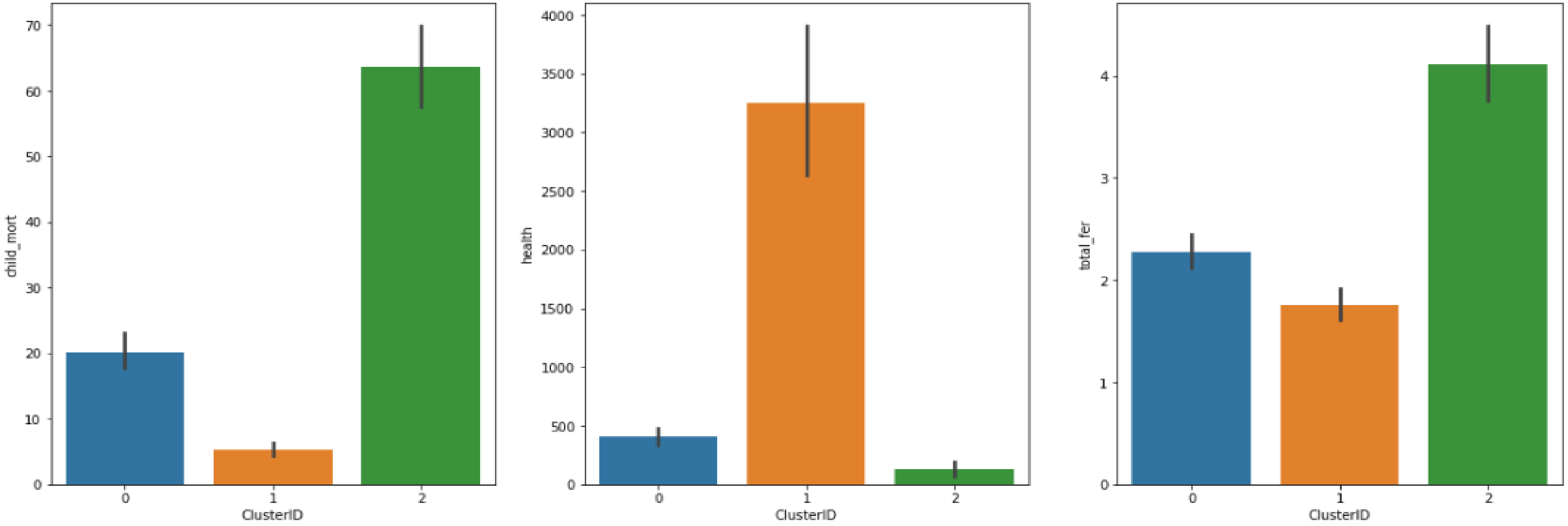
Silhouette score analysis

Data Visualization



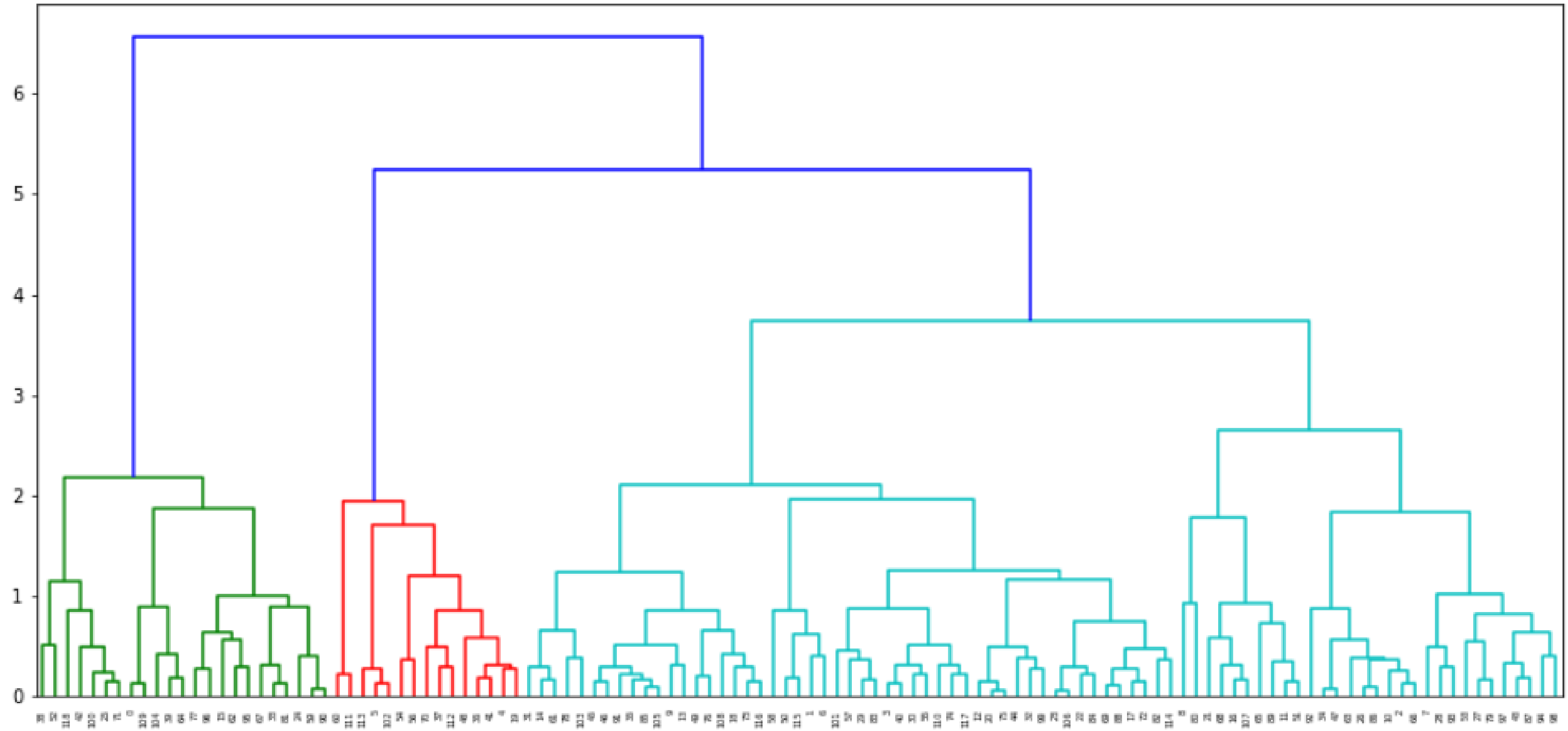
Elbow curve method

K-Mean Algorithm



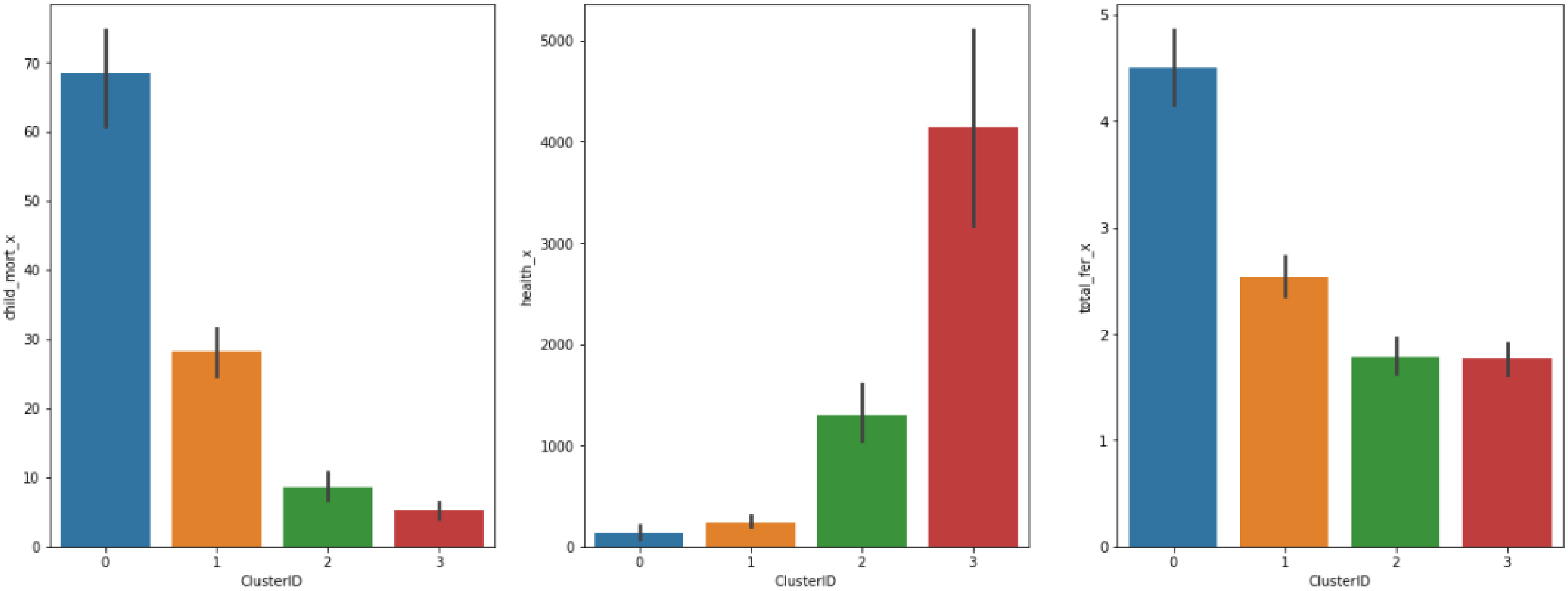
- After identification of K values from silhouette score analysis and elbow curve method, we apply K-Mean Algorithm and identify 3 clusters.
- Cluster 2 has countries that are in the direst need of aid with poor in Child mortality, health spending and total fertility rate.

Hierarchical Clustering Algorithm



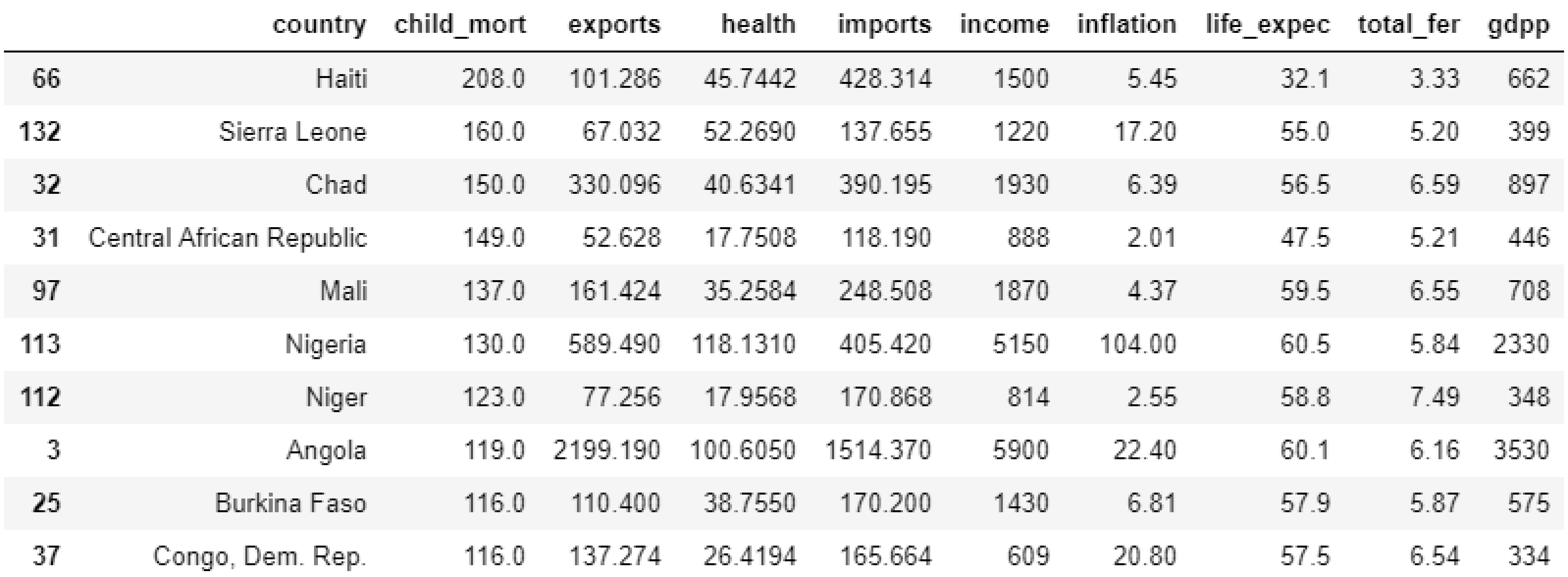
- Hierarchical clustering on using complete method dendrogram and cutting the tree by drawing the parallel line to x axis passing through value 3 on y axis and we obtain 4 clusters

Hierarchical Clustering Algorithm



Cluster 0 has countries that are in the direst need of aid with poor in Child mortality, health spending and total fertility rate.

Observations



	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
66	Haiti	208.0	101.286	45.7442	428.314	1500	5.45	32.1	3.33	662
132	Sierra Leone	160.0	67.032	52.2690	137.655	1220	17.20	55.0	5.20	399
32	Chad	150.0	330.096	40.6341	390.195	1930	6.39	56.5	6.59	897
31	Central African Republic	149.0	52.628	17.7508	118.190	888	2.01	47.5	5.21	446
97	Mali	137.0	161.424	35.2584	248.508	1870	4.37	59.5	6.55	708
113	Nigeria	130.0	589.490	118.1310	405.420	5150	104.00	60.5	5.84	2330
112	Niger	123.0	77.256	17.9568	170.868	814	2.55	58.8	7.49	348
3	Angola	119.0	2199.190	100.6050	1514.370	5900	22.40	60.1	6.16	3530
25	Burkina Faso	116.0	110.400	38.7550	170.200	1430	6.81	57.9	5.87	575
37	Congo, Dem. Rep.	116.0	137.274	26.4194	165.664	609	20.80	57.5	6.54	334

Using either of the clustering methods we are ending up with the same resultant countries that are in the direst need of aid

Summary

- The top 10 countries that are in the direst need of aid are Haiti, Sierra Leone, Chad, Central African Republic, Mali, Nigeria, Niger, Angola, Burkina Faso, Congo, Dem. Rep.
- These Countries are identified based on Socio economical and health factors such as Child Mortality rate, Health Spending, Total Fertility rate.
- Recommendation is to spend funds in these countries by spending on health that will reduces Child Mortality rate.
- Also create awareness programs on birth control which will reduces Total Fertility rate.
- By improving these factors we can improve gdp per capita and that will helps the country to develop.

Thank You