Help Assignment

Application of unsupervised learning & PCA concepts

Contents

I. Problem Statement & Business Goal
II. Method Selection and Advantages
III. Data Visualization , post Data Cleaning
IV. Application of PCA
V. Model building & Final analysis
VI. Conclusion and recommendation

I. Problem Statement & Business Goal

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 funding programs, 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.

Business Goal:

To find out the countries that are in the direst need of aid by using the clustering process by considering some socio-economic and health factors that determine the overall development of the country

II. Method Selection

Challenges in this process :

- Large number of attributes
- Elimination of any attribute(s) will impact the final outcome
- Time and Cost due to large volumes

The approach that we use to overcome the above challenges is to Implement clustering process to group the countries into different clusters based on few parameters and apply the principle component analysis that is based on the concept of dimensionality reduction.

Advantages of PCA approach:

- The large number of dimensions in the data can be reduced to key principal components without losing the original information
- Reduction of processing time and there by reduce "Cost"
- The key attributes as mentioned in the problem like child mortality, income and gross domestic per capita are all considered while grouping the countries into the clusters.

III. Data Visualization, post Data Cleaning

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	Afghanistan	90.2000	10.0000	7.5800	44.9000	1610	9.4400	56.2000	5.8200	553
1	Albania	16.6000	28.0000	6.5500	48.6000	9930	4.4900	76.3000	1.6500	4090
2	Algeria	27.3000	38.4000	4.1700	31.4000	12900	16.1000	76.5000	2.8900	4460
3	Angola	119.0000	62.3000	2.8500	42.9000	5900	22.4000	60.1000	6.1600	3530
4	Antigua and Barbuda	10.3000	45.5000	6.0300	58.9000	19100	1.4400	76.8000	2.1300	12200

The above table contains different attributes that have to be analysed for grouping the countries

III. Data Visualization ,post Data Cleaning

- We can see that child mortality and life expectancy are highly negatively correlated with a correlation of -0.89
- Imports and exports are highly correlated with positive correlation of 0.74
- Similarly total fertility and child mortality are highly correlated with 0.85
- life expectancy and total fertility are highly negatively corelated with -0.76
- Similarly income and GDP are highly positively correlated with 0.90 as the correlation rate .

Thus we need to consider the above points /variables for clustering the countries into similar clusters

IV. Application of PCA

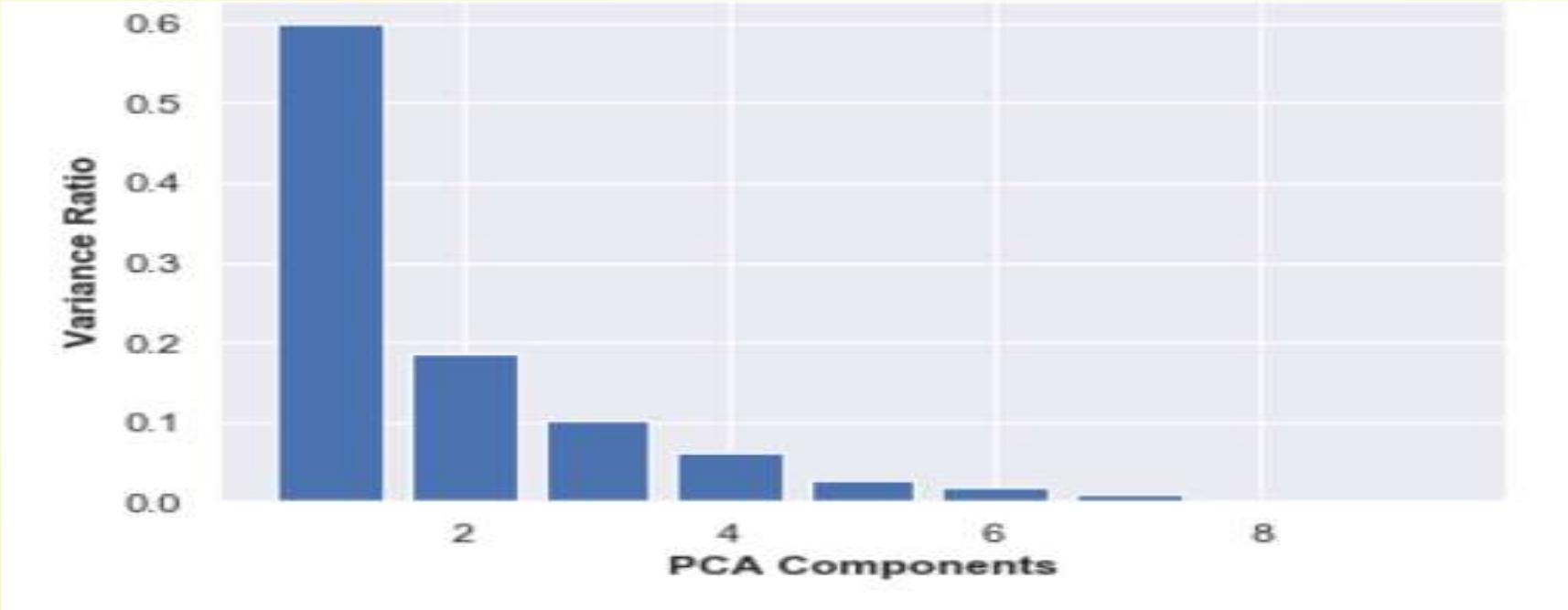
What is Principal component analysis (PCA):

PCA analysis is one of the powerful method that uses the concept of dimensionality reduction technique . This PCA concept is used when the data contains more number of variables /when we want to reduce the large data to smaller size , without losing the original valuable information of the data frame . By this dimensionality reduction technique the model can be improved with better performance, visualizing complex data in smaller chunks , and helps to analyze /frame the clusters much better for ease of interpretation

WHY PCA for this Analysis :

This data frame contains many variables and large data of countries and to do a critical task of clustering the countries and proper utilisation of the funds , its important to understand what factors can be helpful for us to cluster the countries together and identify such clusters as our target ones to utilise the funds properly.

IV. Application of PCA



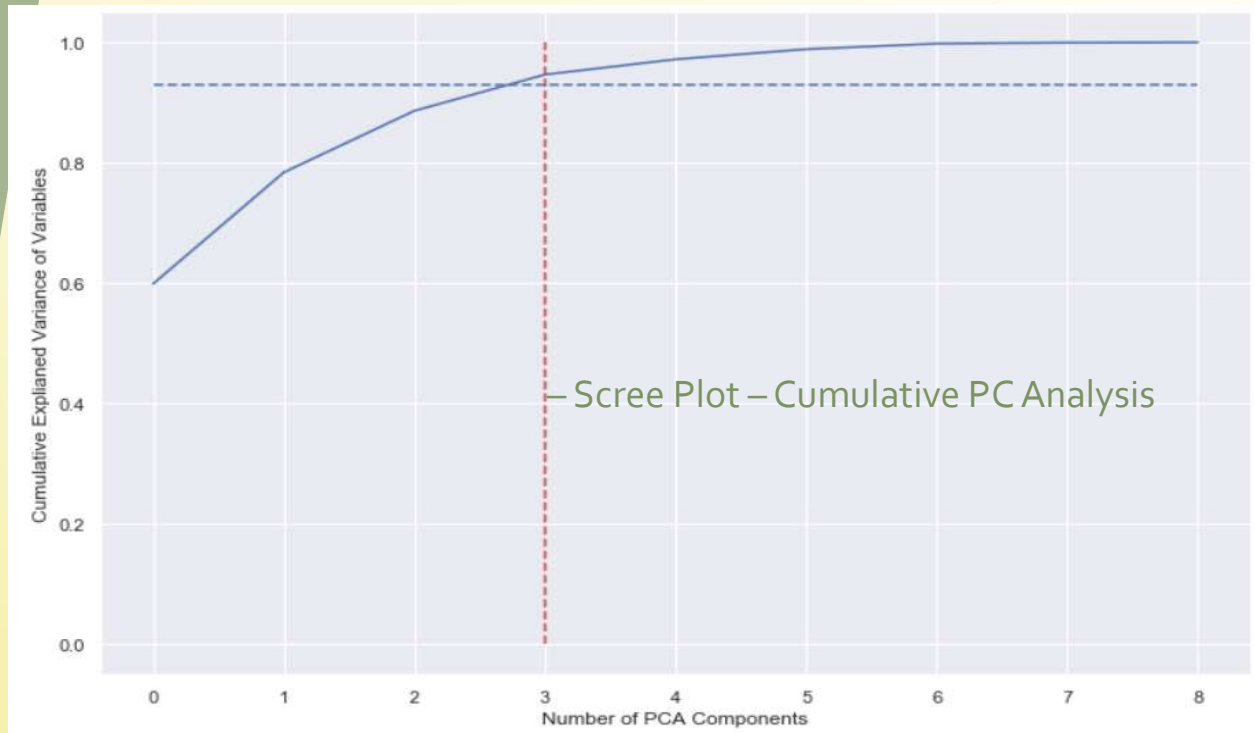
The above graph represents the variance of the data with respect to principal components or attributes.

- First component is explained by nearly 60%
- Second component variance is explained by 20%
- Third ones by nearly 10%.

This concludes that the Combination of four or five Principal components will explain nearly 95 to 96% variance in the data .

IV. Application of PCA

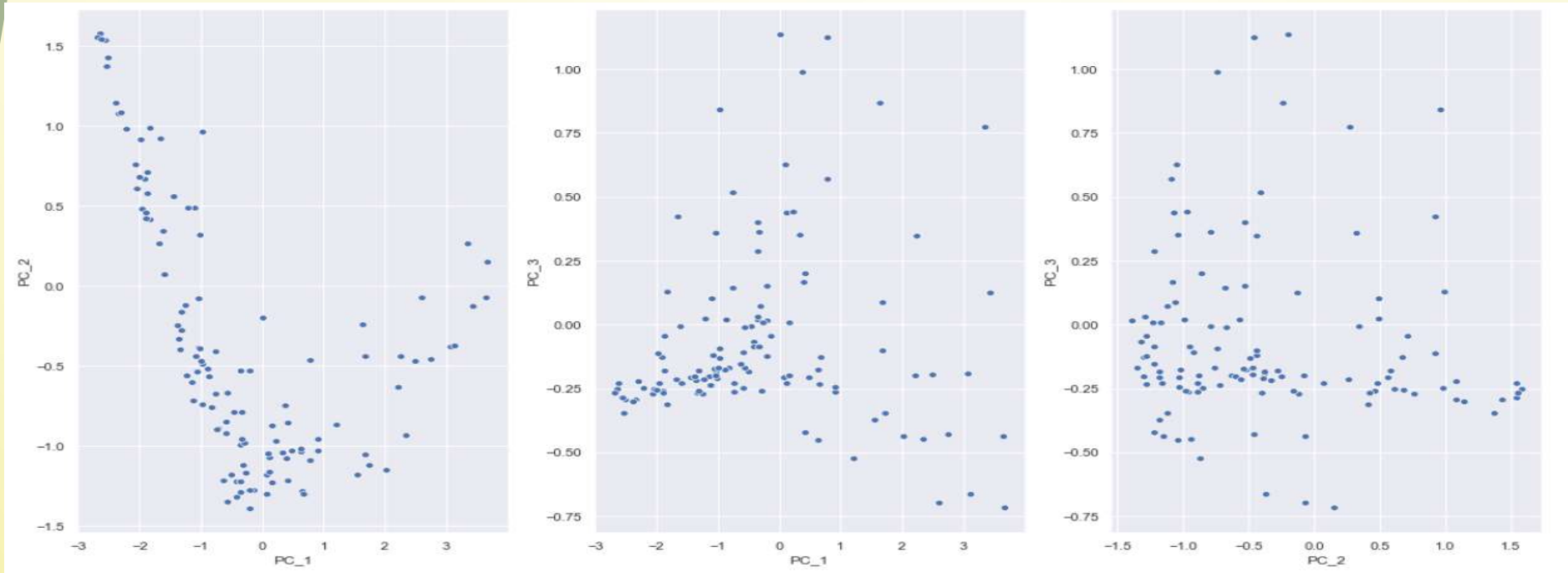
– Scree Plot – Cumulative PC Analysis



	Attribute	PC_1	PC_2	PC_3
0	child_mort	-0.3044	0.5065	-0.0311
1	exports	0.3452	0.3570	-0.2132
2	health	0.3514	0.1113	-0.2294
3	imports	0.3425	0.3216	-0.3207
4	income	0.3898	0.1202	0.2512
5	inflation	0.1982	0.2244	0.8562
6	life_expec	0.3347	-0.4096	0.0199
7	total_fer	-0.2886	0.4883	0.0241
8	gdpp	0.4011	0.1706	-0.0318

From the above scree plot, more than 90% of the variance in the data is explained by the first three principal Components. Thus we shall consider these Principal components for our further analysis and clustering of countries This is explained by cumulative calculation of the PC values.

IV. Application of PCA – Scatter Plots



- It is inferred that the variation between PC1 & PC2 components are scattered all over the axis (-3&3)and most of the points in between -1,1
- With respect to PC1&PC3, we do not find that much variance explained between the datapoints but only few points are centered -0.25&1
- With respect to PC2&PC3,most of the points are spread between -1.5&1.5 and also the points are scattered all over.

IV. Application of PCA – Hopkins statistics test

Hopkins statistics test :
The Hopkins statistic test is a way of measuring the cluster tendency of a data set. It will help us to identify if the data is uniformly distributed or not . A value close to 1 tends to indicate the data is highly clustered , random data will tend to result in values around 0.5,and uniformly distributed data will tend to result in values closer to 0

Note: we got a score of 0.81 which is a good score of Hopkins test to proceed with clustering

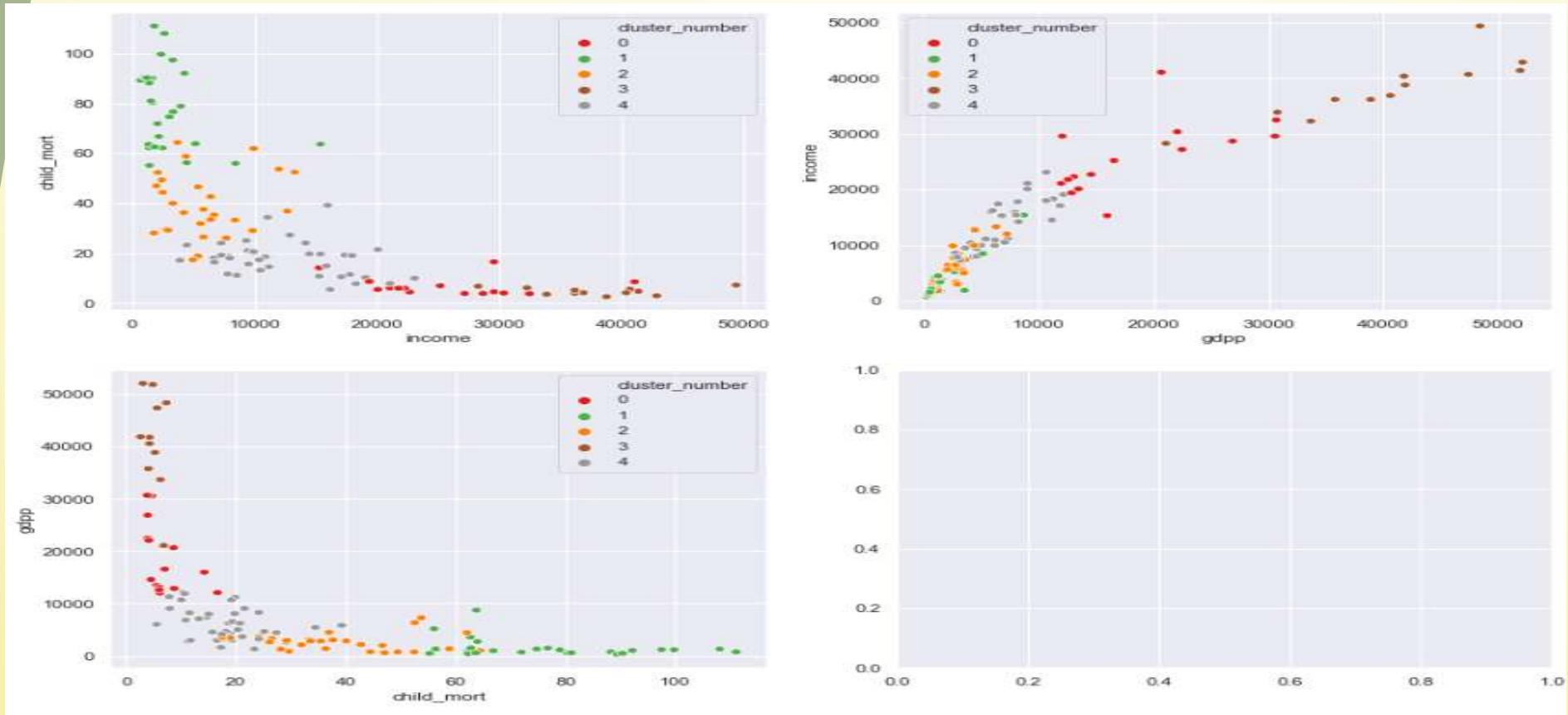
V. Model building & Final analysis

	country	child_mort	exports	imports	health	income	inflation	life_expec	total_fer	gdpp	cluster_number
0	Afghanistan	90.2000	55.3000	248.2970	41.9174	1610	52.2032	56.2000	5.8200	553	1
1	Algeria	27.3000	1712.6400	1400.4400	185.9820	12900	718.0600	76.5000	2.8900	4460	4
2	Antigua and Barbuda	10.3000	5551.0000	7185.8000	735.6600	19100	175.6800	76.8000	2.1300	12200	4
3	Armenia	18.1000	669.7600	1458.6600	141.6800	6700	250.1940	73.3000	1.6900	3220	4
4	Australia	4.8000	10276.2000	10847.1000	4530.8700	41400	602.0400	82.0000	1.9300	51900	3

It is understood from the business standpoint and from the question , that Child mortality , income and GDP are the key factors that decide the development of the country .

Lets proceed with analysing all these 3 components for building meaningful clusters

V. Model building & Final analysis



V. Model building & Final analysis

Inferences that can be drawn from the scatter plots :

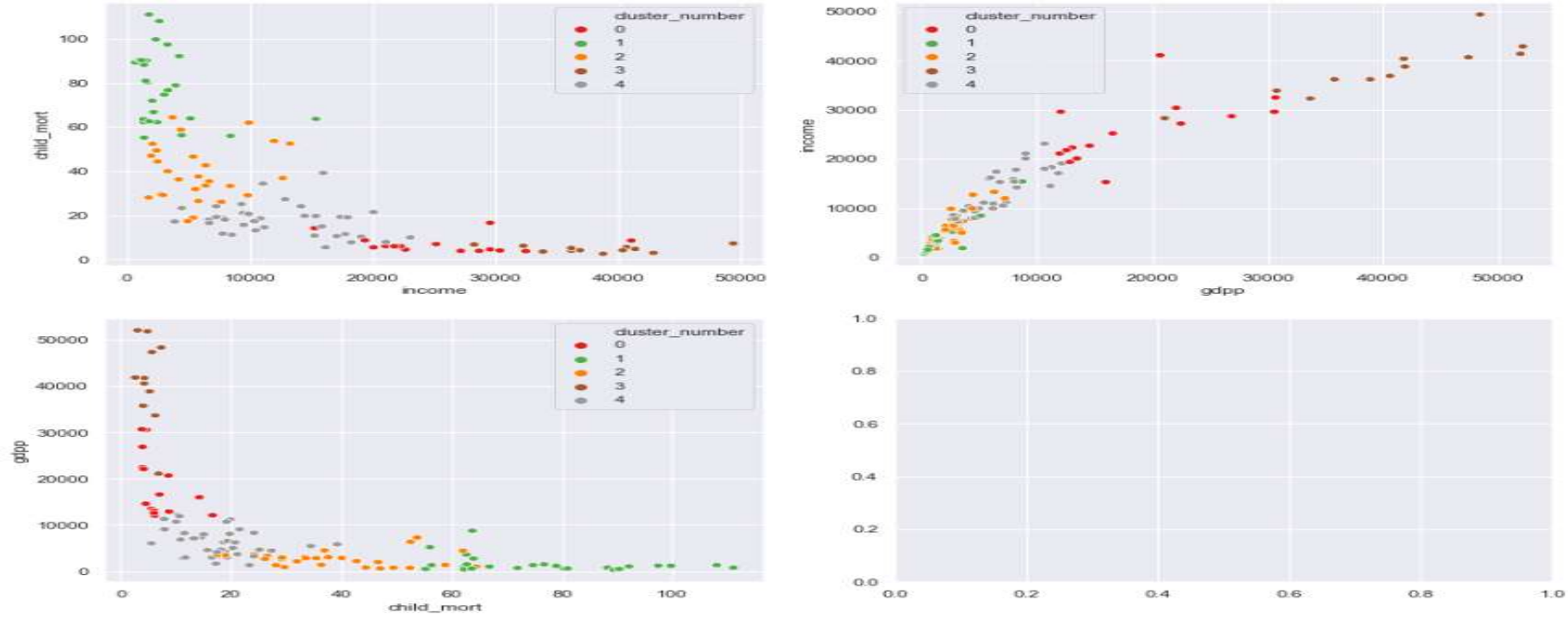
a. There seems to have a close relationship in forming the cluster of the countries with child_ mortality & income .
b. Countries with lower income levels have more child mortality and these form similar clusters . Thus aid is needed for such countries
c. Also we can group the countries based on income and GDP of the nations .
d. As it is known that the more income , the more purchasing capacity and more production which in turn leads to better GDP.
e. We can see that many countries have less than 10,000 level of GDP with very low income levels close to 10,000 & these can be clubbed as a cluster .
f. The next element to consider is GDP & child mortality . It is observed that countries account for such

cluster has child mortality rate between 0-20 level points with low GDP and next category comes with those with child mortality between 20 to 50 level points and GDP still less than 10,000 mark.

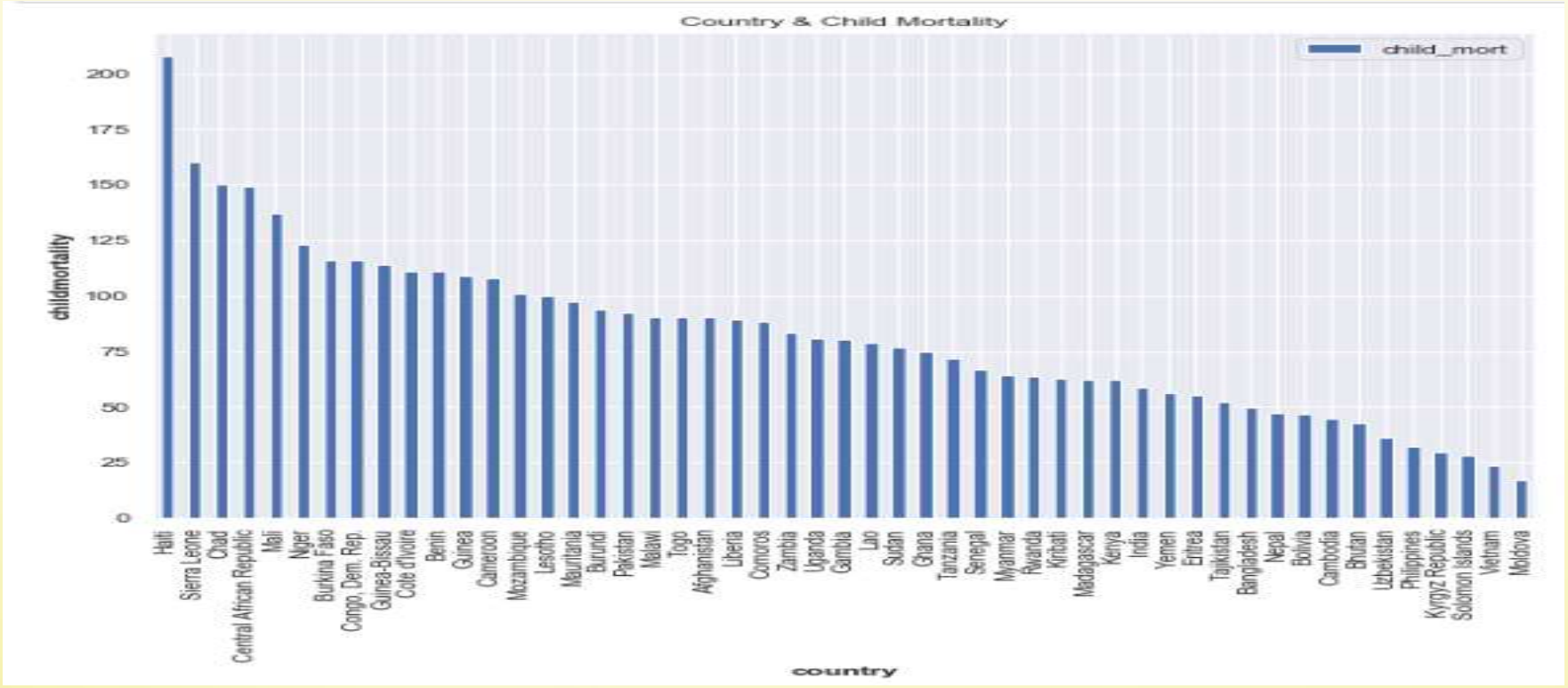
we need to extend financial aid to such countries & they form one another cluster.

VI. Conclusion and Recommendation

After applying the concepts of K Means Clustering and Hierarchical Clustering Methods that follow the concept of Euclidean distance of data forming from the respective clusters and using the dendrogram method of calculating the mean distance of the data point from its clusters we finally arrive at those principle components: Child Mortality, income and GDP as the key parameters for arriving at the clusters for the countries. Please refer the below scatter Plots of the clusters (0, 1, 2, and 4 as cluster numbers) for reference

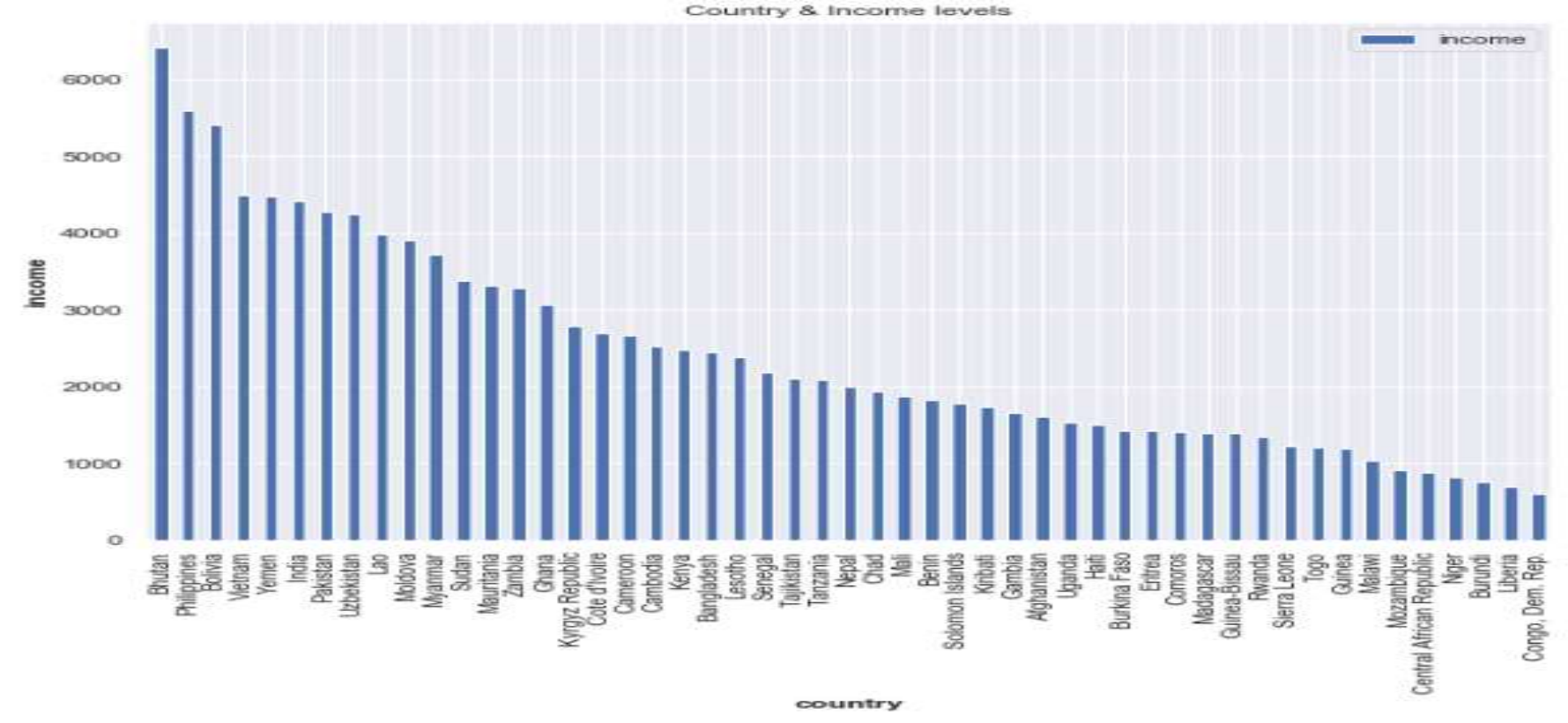


VI. Conclusion and Recommendation



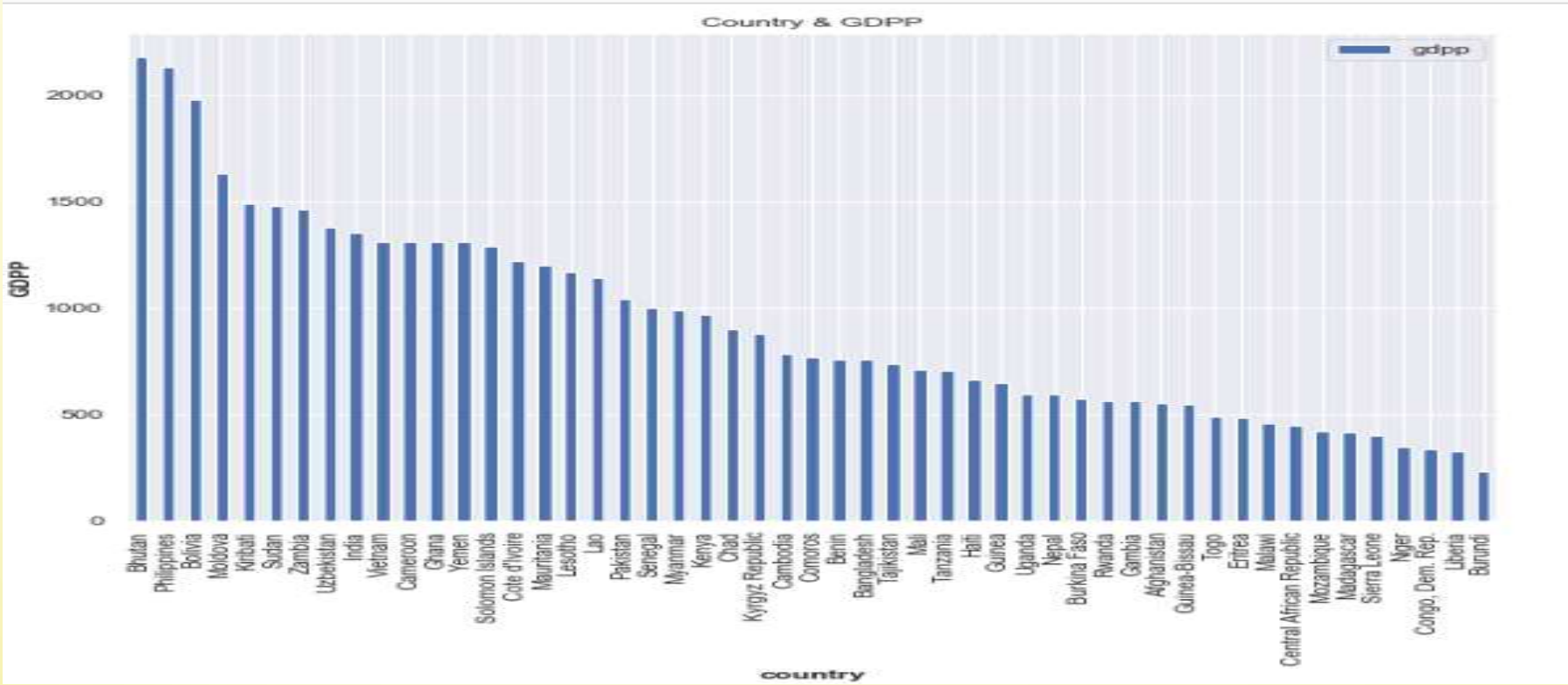
Note: We can see the list of countries that have high child mortality that needed aid

VI. Conclusion and Recommendation



Note: We can see the list of countries that have very lower income levels that needed aid

VI. Conclusion and Recommendation



Note: We can see the list of countries that have very lower GDP that needed aid

VI. Conclusion and Recommendation

Conclusion : From the above analysis we can see the list of the countries that have high child mortality , low income and low GDP and that need aid . Hence by using the concept of PCA , which in turn uses the concept of dimensionality reduction , we can reduce the large number of dimensions into smaller parts and not losing any information of the original data , able to arrive at the list of countries that need aid to improve their current status .

Index No	Country	Index No	Country	Index No	Country	Index No	Country
0	Afghanistan	14	Eritrea	28	Malawi	42	Sudan
1	Bangladesh	15	Gambia	29	Mali	43	Tajikistan
2	Benin	16	Ghana	30	Mauritania	44	Tanzania
3	Bhutan	17	Guinea	31	Moldova	45	Togo
4	Bolivia	18	Guinea-Bissau	32	Mozambique	46	Uganda
5	Burkina Faso	19	Haiti	33	Myanmar	47	Uzbekistan
6	Burundi	20	India	34	Nepal	48	Vietnam
7	Cambodia	21	Kenya	35	Niger	49	Yemen
8	Cameroon	22	Kiribati	36	Pakistan	50	Zambia
9	Central African Republic	23	Kyrgyz Republic	37	Philippines		
10	Chad	24	Lao	38	Rwanda		
11	Comoros	25	Lesotho	39	Senegal		
12	Congo, Dem. Rep.	26	Liberia	40	Sierra Leone		
13	Cote d'Ivoire	27	Madagascar	41	Solomon Islands		