CLUSTERING OF COUNTRIES

By: Aditya Ranjan Behera

Abstract:-

- **Objective:-**

  We HELP International humanitarian NGO, our main objective is to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities.

- **Problem Statement:-**

  After the recent 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.

Methodology of Analysis of Problem:-

- Data Collection:
  - Import the dataset
  - Sanity check and data cleaning
- Removing Outliers:
  - Removing of outliers as per problem understanding
- Data Visualization:
  - Visualizing dataset to for representation of relation between features of dataset for better understanding of data.
- Standardization Of Data:
  - Standardize dataset and convert the whole dataset into a single scale by using standardscaler for further analysis.

Methodology of Analysis of Problem:-

- Performing PCA:
  - Perform PCA(Principal Component Analysis) on the dataset for dimensionality reduction by using principal components. Calculate the optimal number of principal components by using screeplot and consider five number of principal components.
- Checking Cluster Tendency:
  - By using Hopekine's method check the clusterability of the PCA dataset.
- K-Means Clustering:
  - Calculate the optimal number of clusters by using both sillhoutte average score and Elbow curve method
  - Perform clustering on PCA dataset and form four number of clusters
  - Visualize principal component1 vs principal component2 for each cluster
  - Visualize 'GDPP' vs 'child death' for each cluster
  - Visualize 'Income' vs 'child death' for each cluster
  - Visualize 'GDPP' vs 'Income for each cluster
  - Analyze the name of five countries which are need aid from us

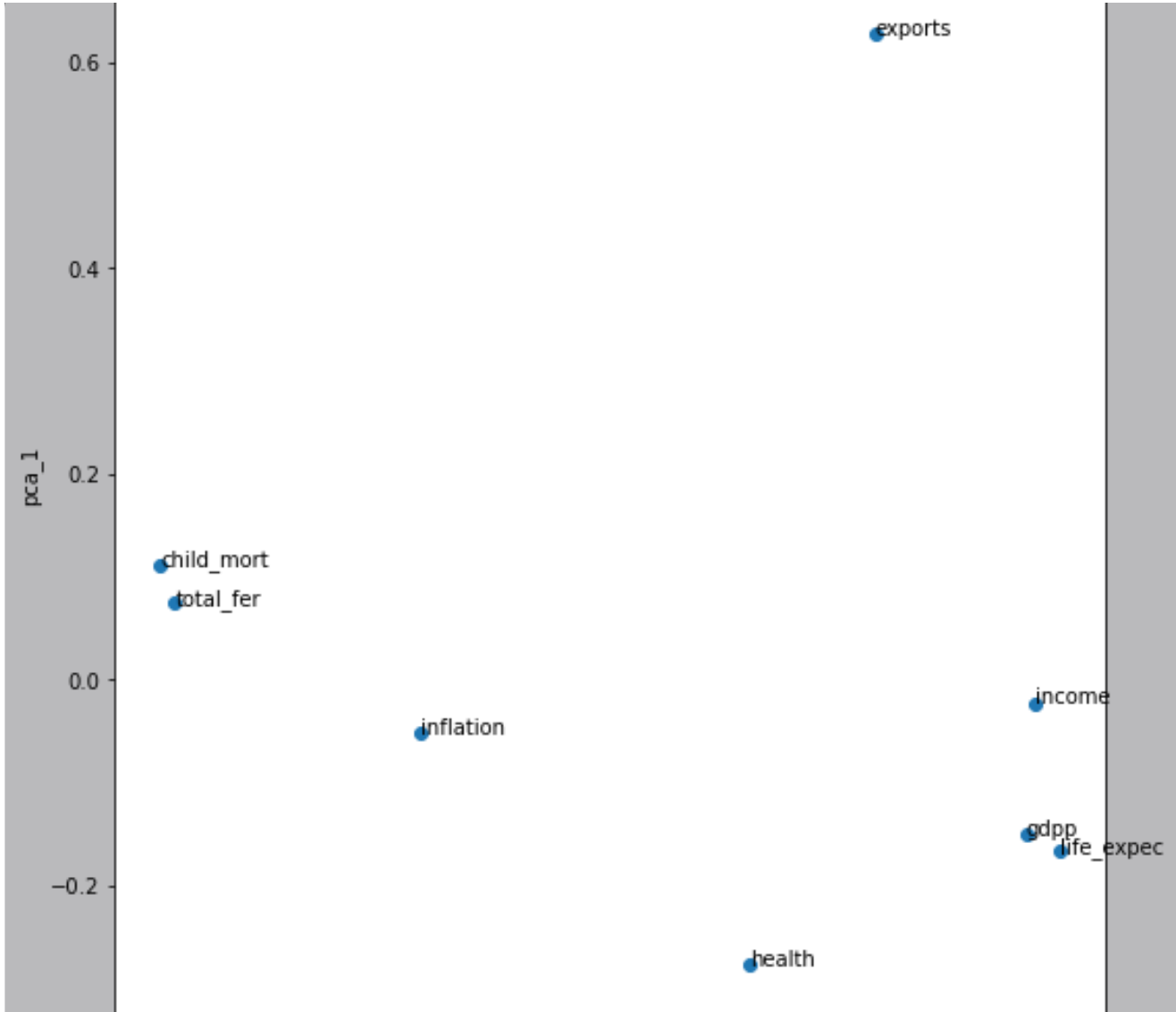
Methodology of Analysis of Problem:-

- Hierarchical Clustering:
  - Calculate the optimal number of clusters by using both single linkage and complete linkage.
  - Perform clustering on PCA dataset and form four number of clusters
  - Visualize principal component1 vs principal component2 for each cluster
  - Visualize ‘GDPP’ vs ‘child death’ for each cluster
  - Visualize ‘Income’ vs ‘child death’ for each cluster
  - Visualize ‘GDPP’ vs ‘Income for each cluster
  - Analyze the name of five countries which are need aid from us.

- Analysis of Results:
  - By analyzing results of K-Means and Hierarchical clustering obtain the name of five countries as per our target.

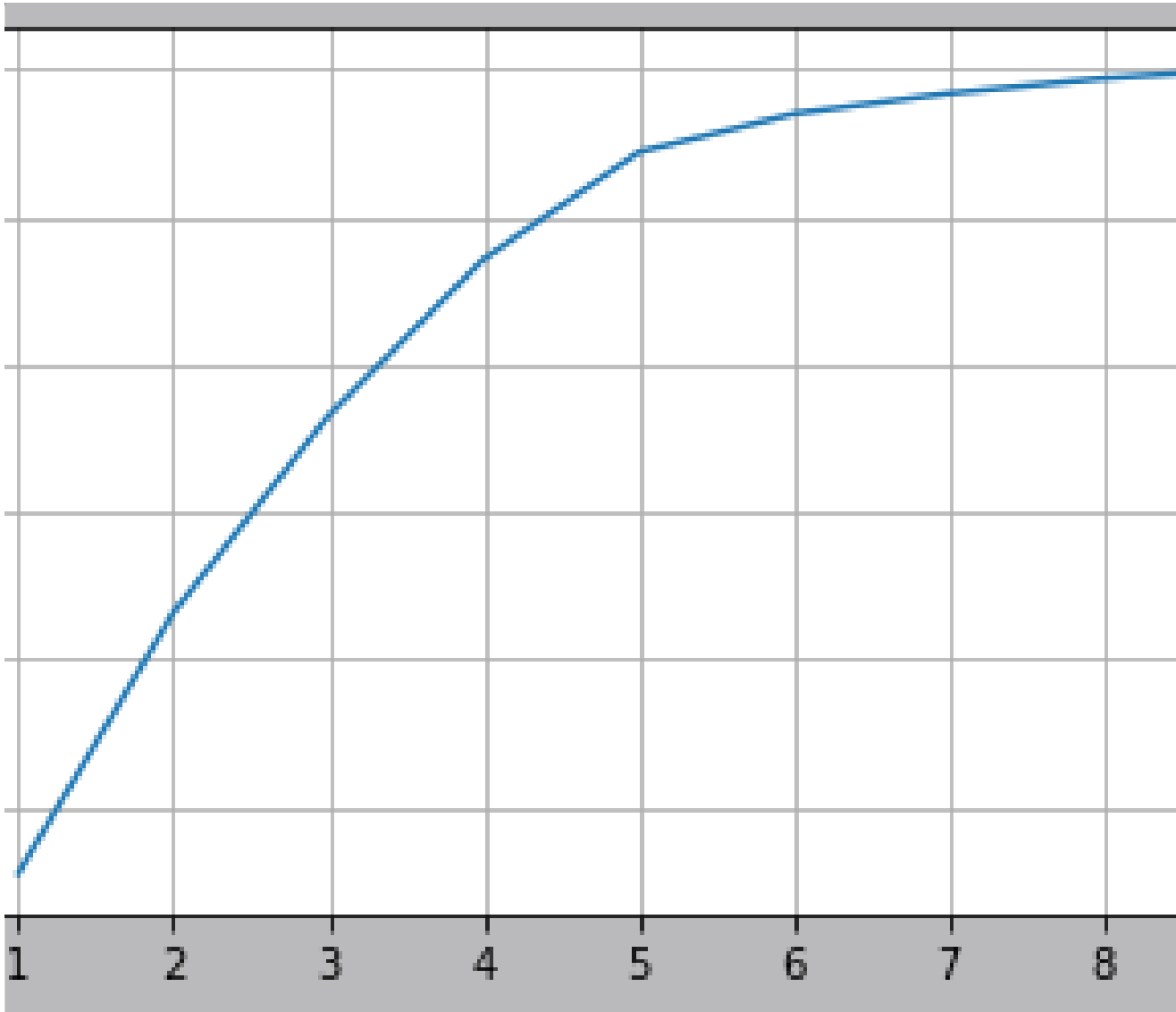
Principal Component Analysis:-

Visualize the distribution of principal component-1 and principal component-2 for each features of the dataset.



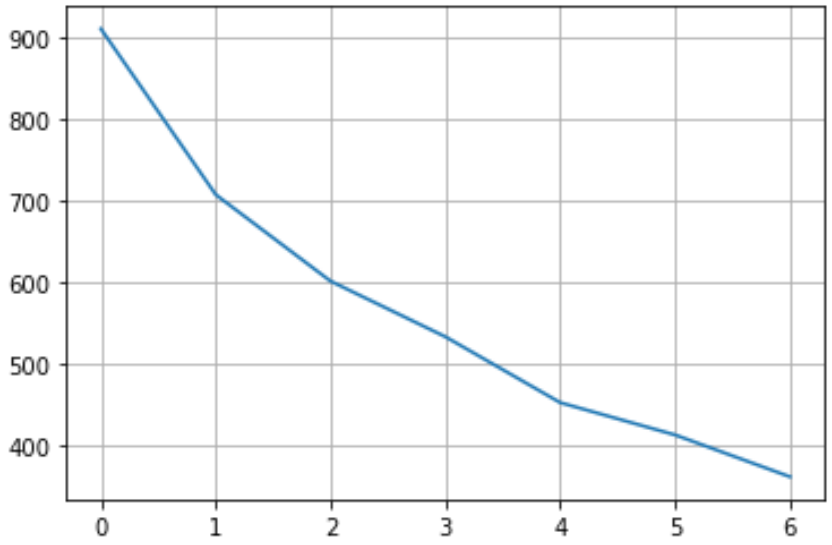
Principal Component Analysis:-

By visualizing this screeplot we can conclude that to use 5 number of principal components. Because 95% of variance are explained by 5 principal components.

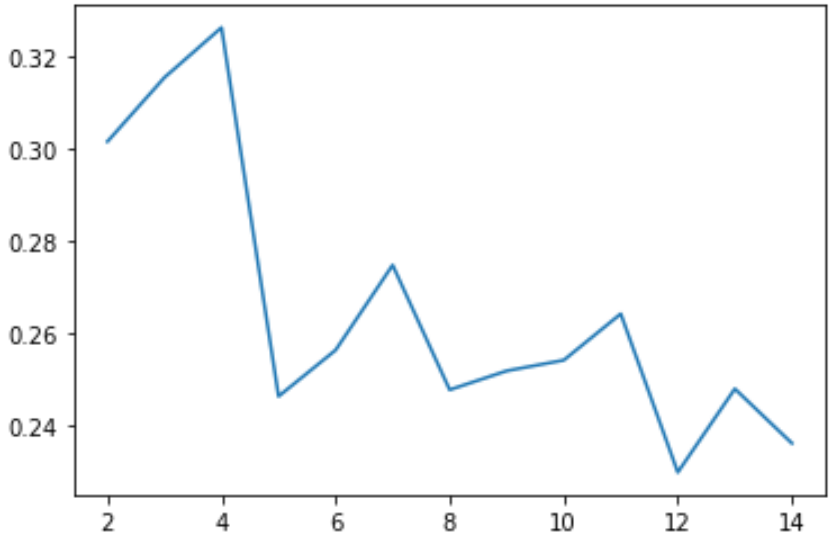


K-Means Clustering:-

ELBOW CURVE METHOD

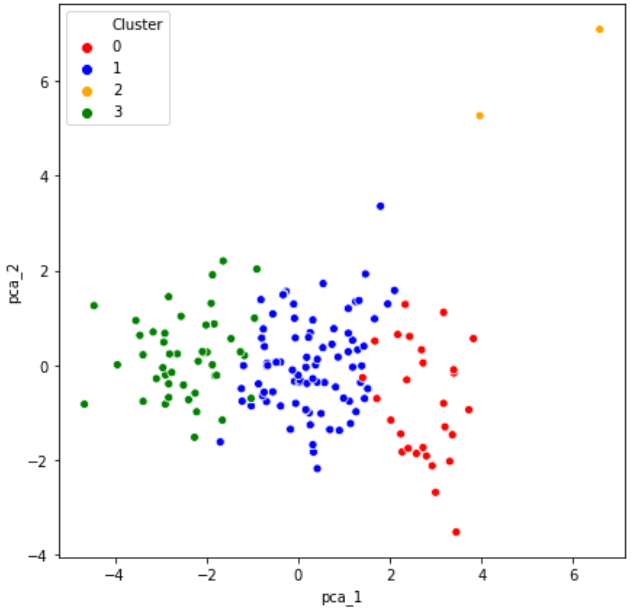


SILHOUETTE SCORE METHOD

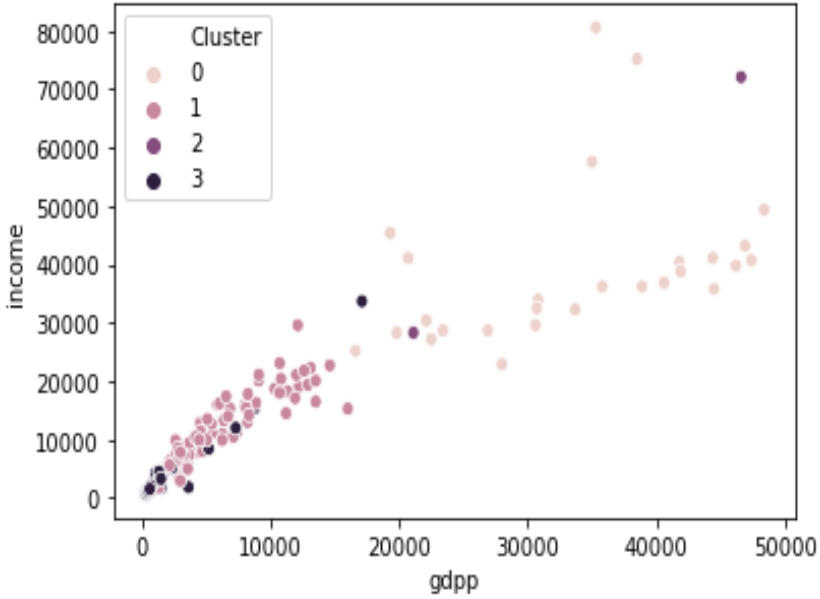


K-Means Clustering:-

- Visualization of different feature for each cluster:-



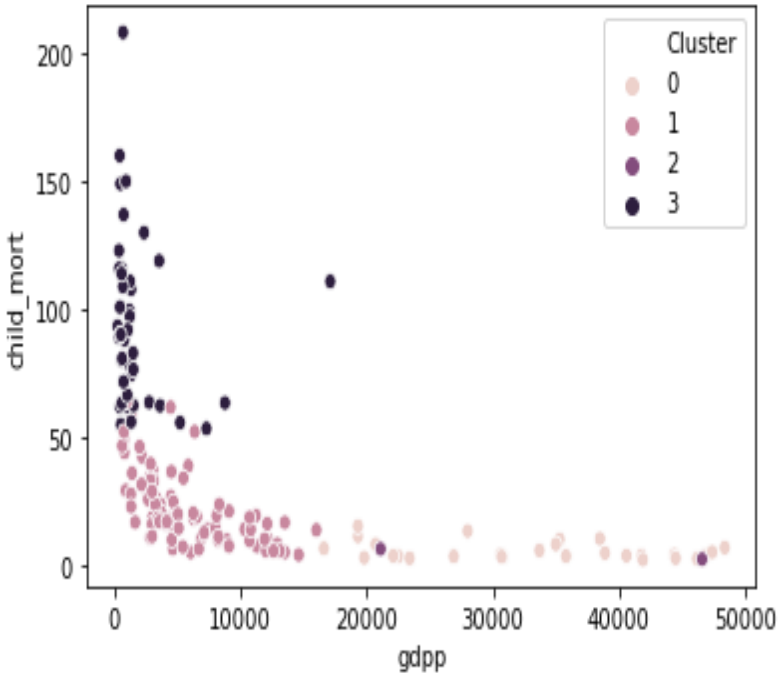
Scatter plot of pca-1 and pca-2 for each cluster



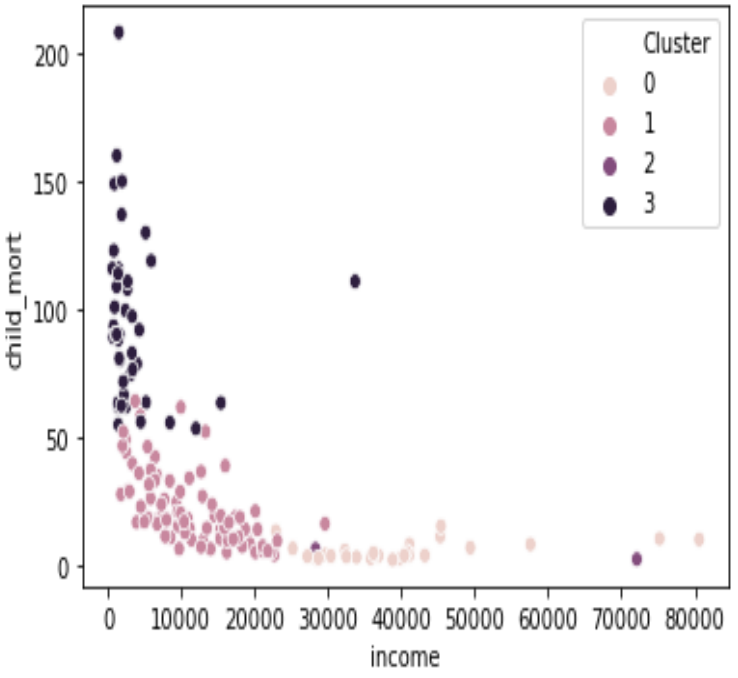
Scatter plot of income and gdpp for each cluster

K-Means Clustering:-

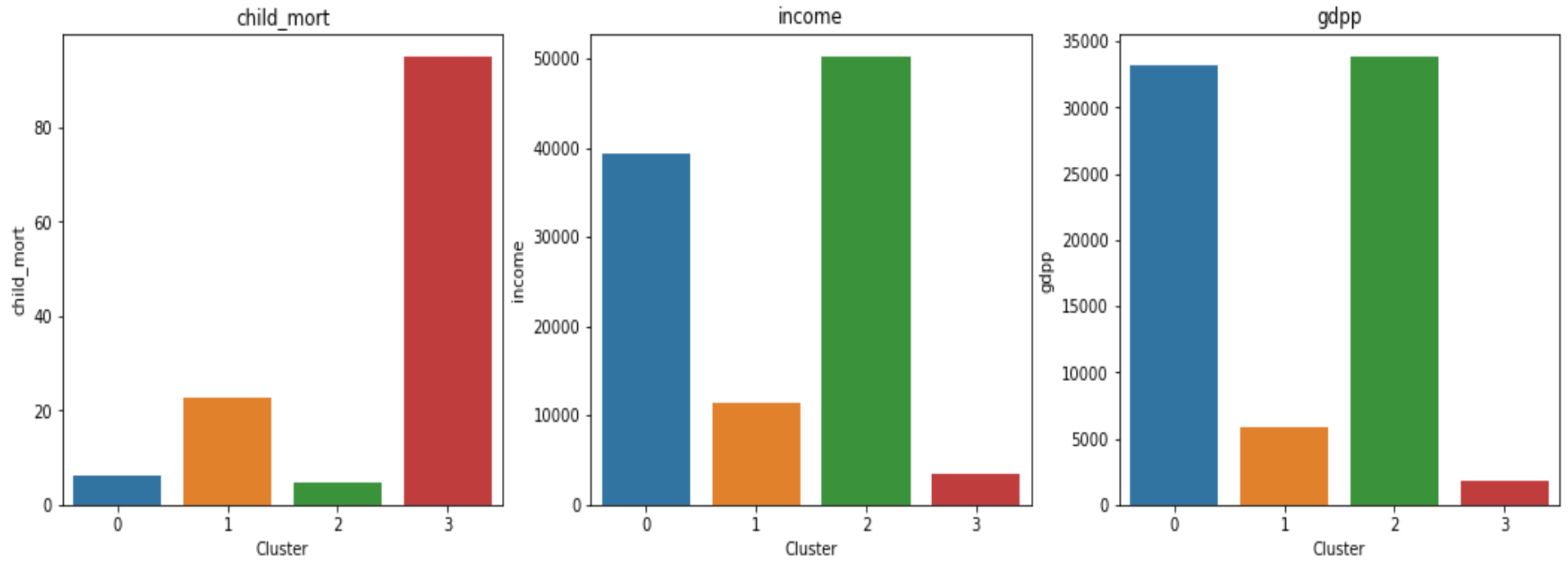
- Scatter plot of gdpp and child_mort for each cluster



- Scatter plot of income and child_mort for each cluster



K-Means Clustering:-



- Visualization of distribution of features like “gdpp”, “income”, “child_mort”
- As per our task we require low “gdpp”, low “income” and high “child_mort” so here we taking cluster-1 to concern

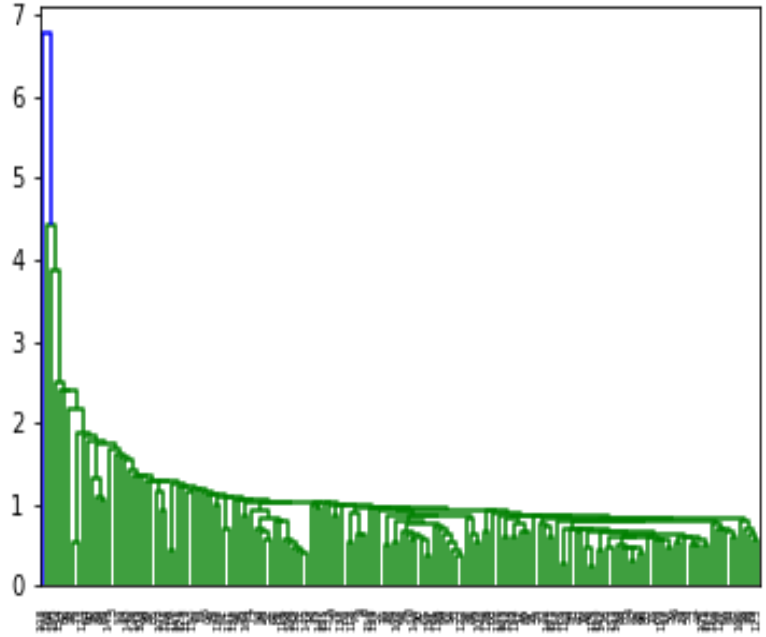
K-Means Clustering:-

- The name of listed five countries are:

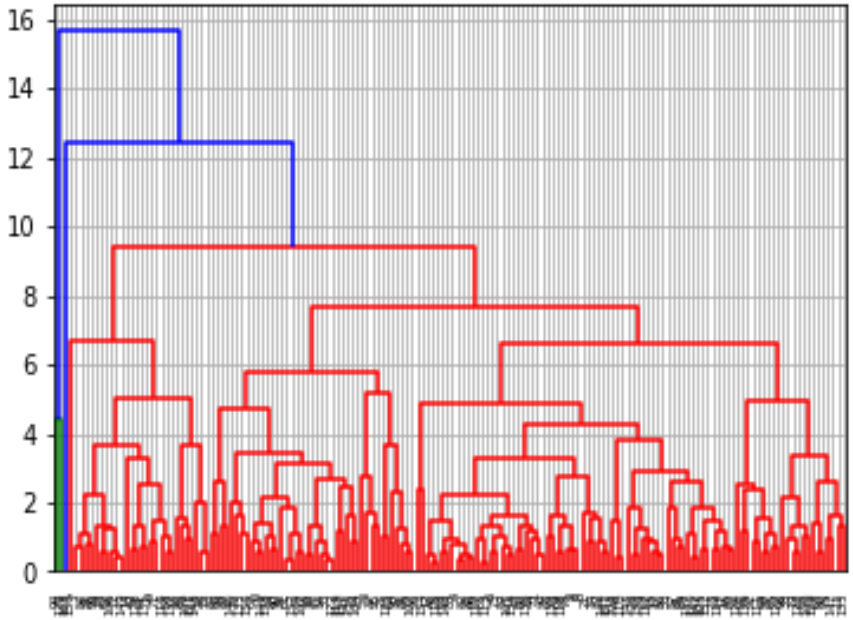
	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	Cluster
25	Burundi	93.6	8.92	11.60	39.2	764	12.30	57.7	6.26	231	3
85	Liberia	89.3	19.10	11.80	92.6	700	5.47	60.8	5.02	327	3
36	Congo, Dem. Rep.	116.0	41.10	7.91	49.6	609	20.80	57.5	6.54	334	3
107	Niger	123.0	22.20	5.16	49.1	814	2.55	58.8	7.49	348	3
125	Sierra Leone	160.0	16.80	13.10	34.5	1220	17.20	55.0	5.20	399	3

Hierarchical Clustering:-

- Single Linkage Method



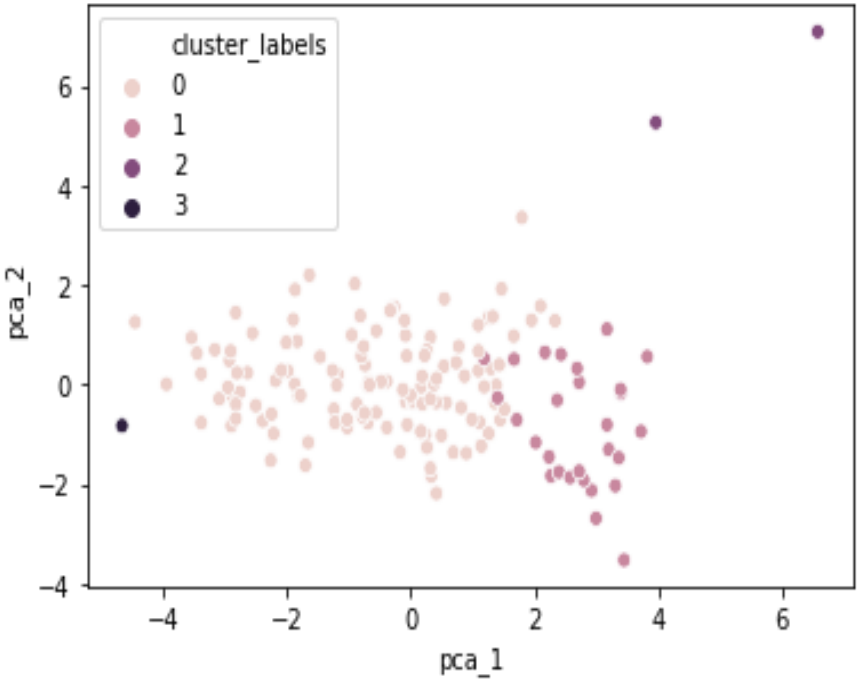
- Complete Linkage Method



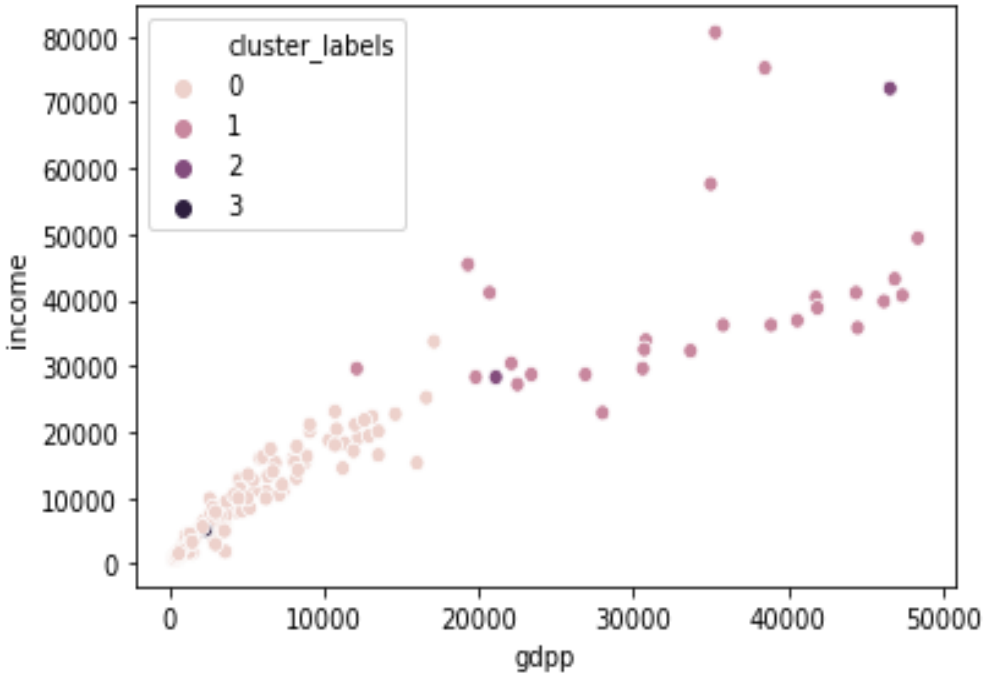
- By considering complete linkage method I have considered 4 number of clusters for clustering dataset.

Hierarchical Clustering:-

- Visualizing scatter plot between pca_1 and pca_2 for each cluster

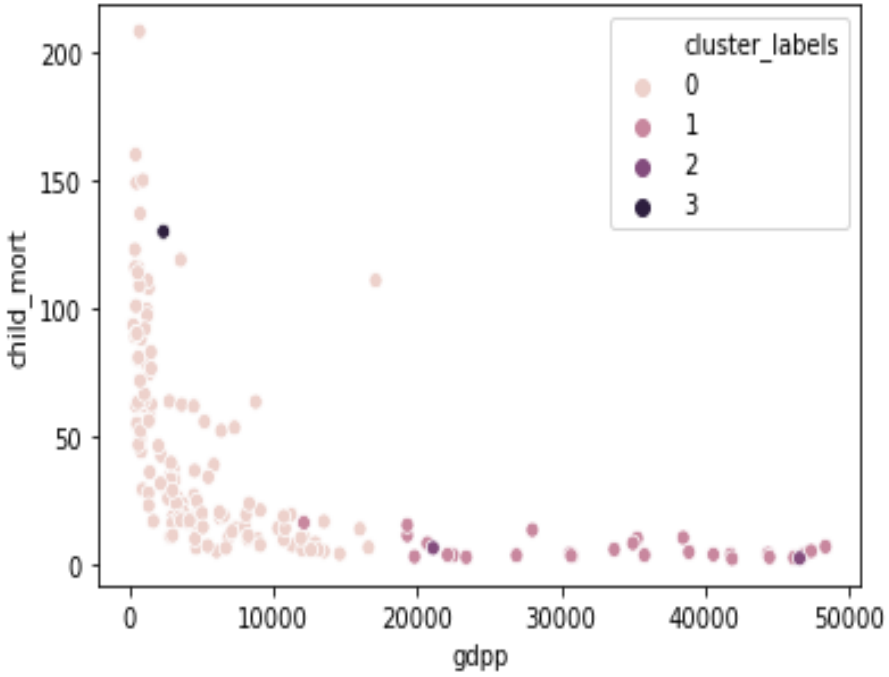


- Visualizing scatter plot between gdpp and income for each cluster

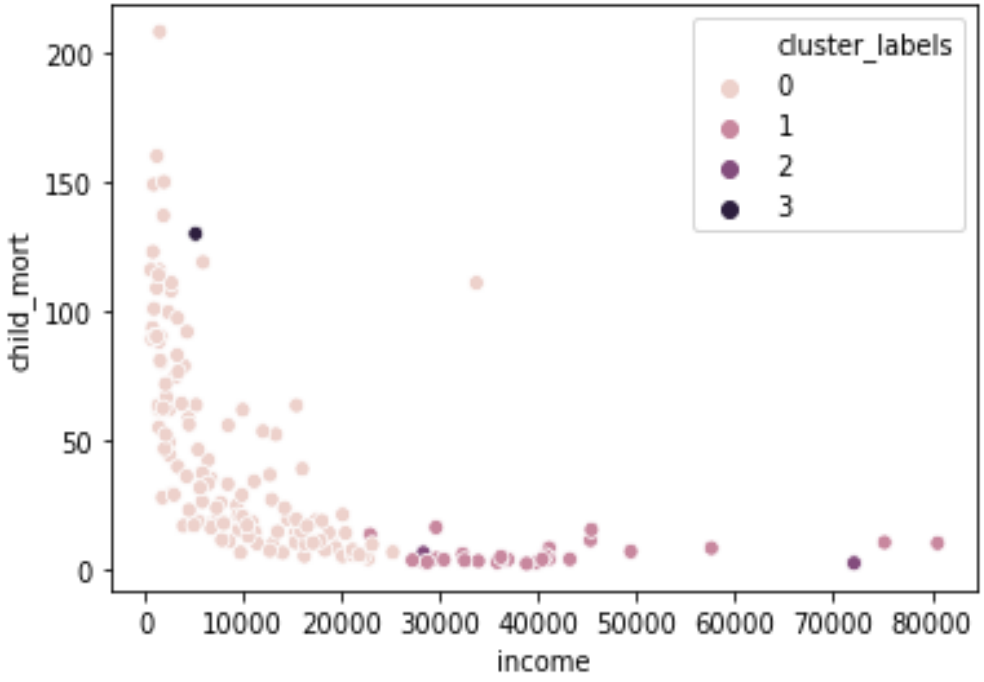


Hierarchical Clustering:-

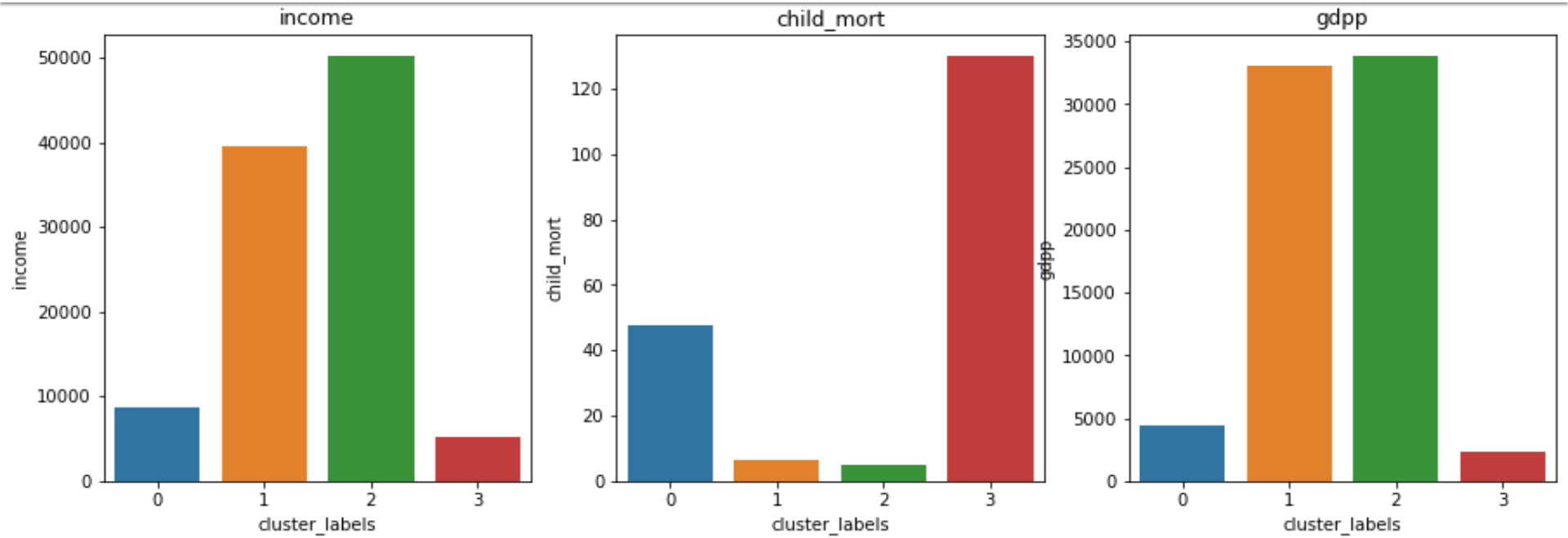
- Visualizing scatter plot between gdpp and child_mortnp



- Visualizing scatter plot between income and child_mort



K-Means Clustering:-



- Visualization of distribution of features like “gdpp”, “income”, “child_mort”
- As per our task we require low “gdpp”, low “income” and high “child_mort” so here we taking cluster-1 to concern

Hierarchical Clustering:-

- The name of listed five countries are:

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_labels
25	Burundi	93.6	8.92	11.60	39.2	764	12.30	57.7	6.26	231	0
85	Liberia	89.3	19.10	11.80	92.6	700	5.47	60.8	5.02	327	0
36	Congo, Dem. Rep.	116.0	41.10	7.91	49.6	609	20.80	57.5	6.54	334	0
107	Niger	123.0	22.20	5.16	49.1	814	2.55	58.8	7.49	348	0
125	Sierra Leone	160.0	16.80	13.10	34.5	1220	17.20	55.0	5.20	399	0

Decision Making:-

- From the analysis of K-Means clustering and Hierarchical clustering we get the same results from both the clusters. So the list of the name of countries which need aid from our side are:

Name of countries are:

Burundi

Liberia

Congo, Dem. Rep

Niger

Sierra leone

THANK YOU