Clustering and PCA Assignment

Analysis

Presented By:

- Saurabh Dongare

Problem statement

• **HELP International** is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries .

• After the recent funding programs, the NGO have raised 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 Objectives:

• Our Objective is to categorize the countries using some socio-economic and health factors that determine the overall development of the country.

• Then suggest the countries which the CEO needs to focus on the most providing with basic amenities and relief during the time of disasters and natural calamities.

Analysis Approach

Data Understanding (Understanding Problem statement, Columns provided and data dictionary)

Data Cleaning/Standardizing (Eliminating multi co-linearity on highly correlated data)

Using PCA Module(imported) to perform PCA on the data set. Then merging data with original dataset, following with outlier handling.

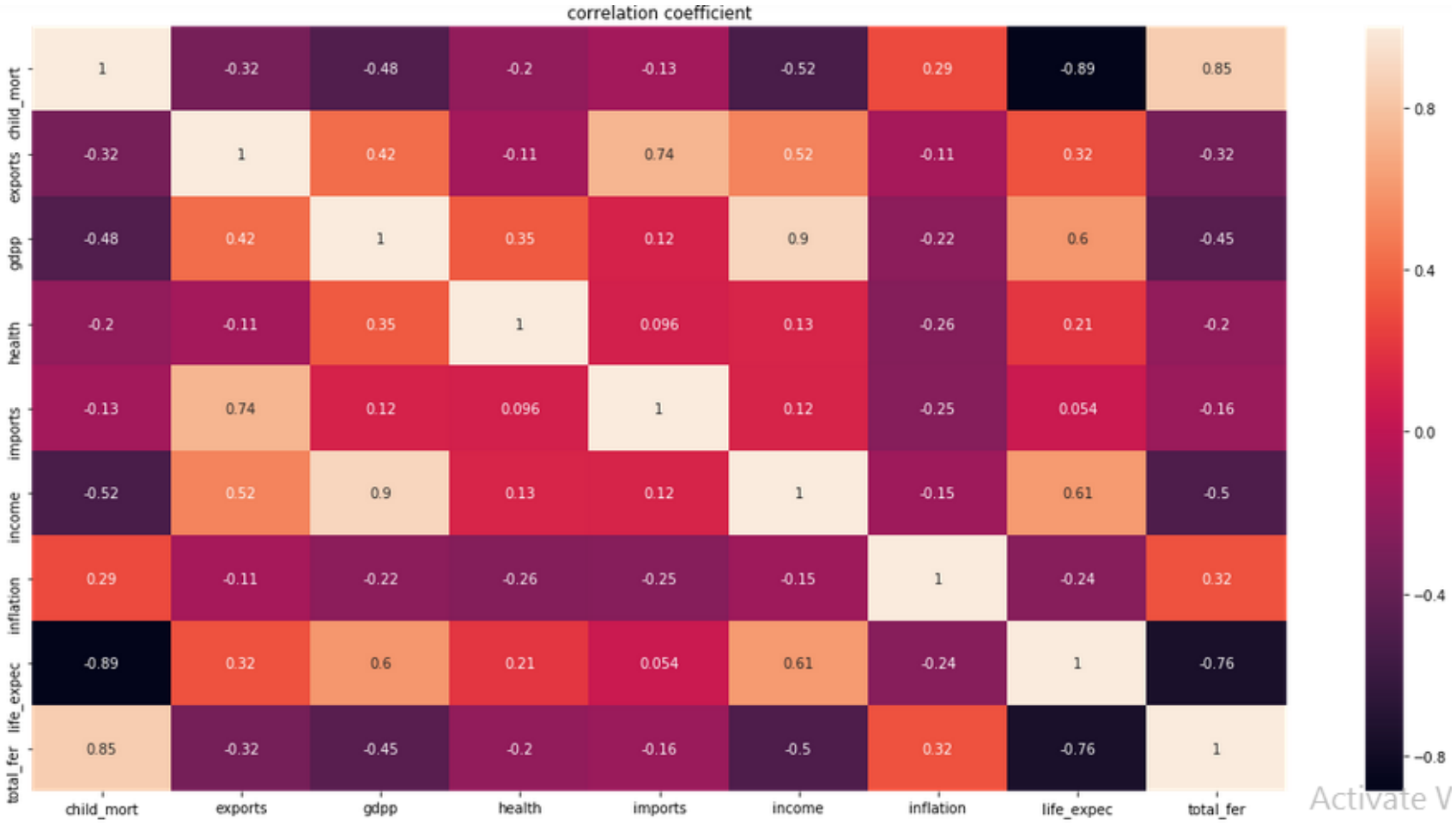
Setting the clusters for the dataset. Followed by performing Hopkins and analyzing the best cluster to use with good tendency

Accordingly applying Silhouette Analysis (for nearest cluster) and sum of squares of distance between them

At the end applying Hierarchical clustering on dataset representing through Dendogram.

Heat map showing the correlation among features

- High correlation can be seen in the darker zones and using the PCA we have handled it.

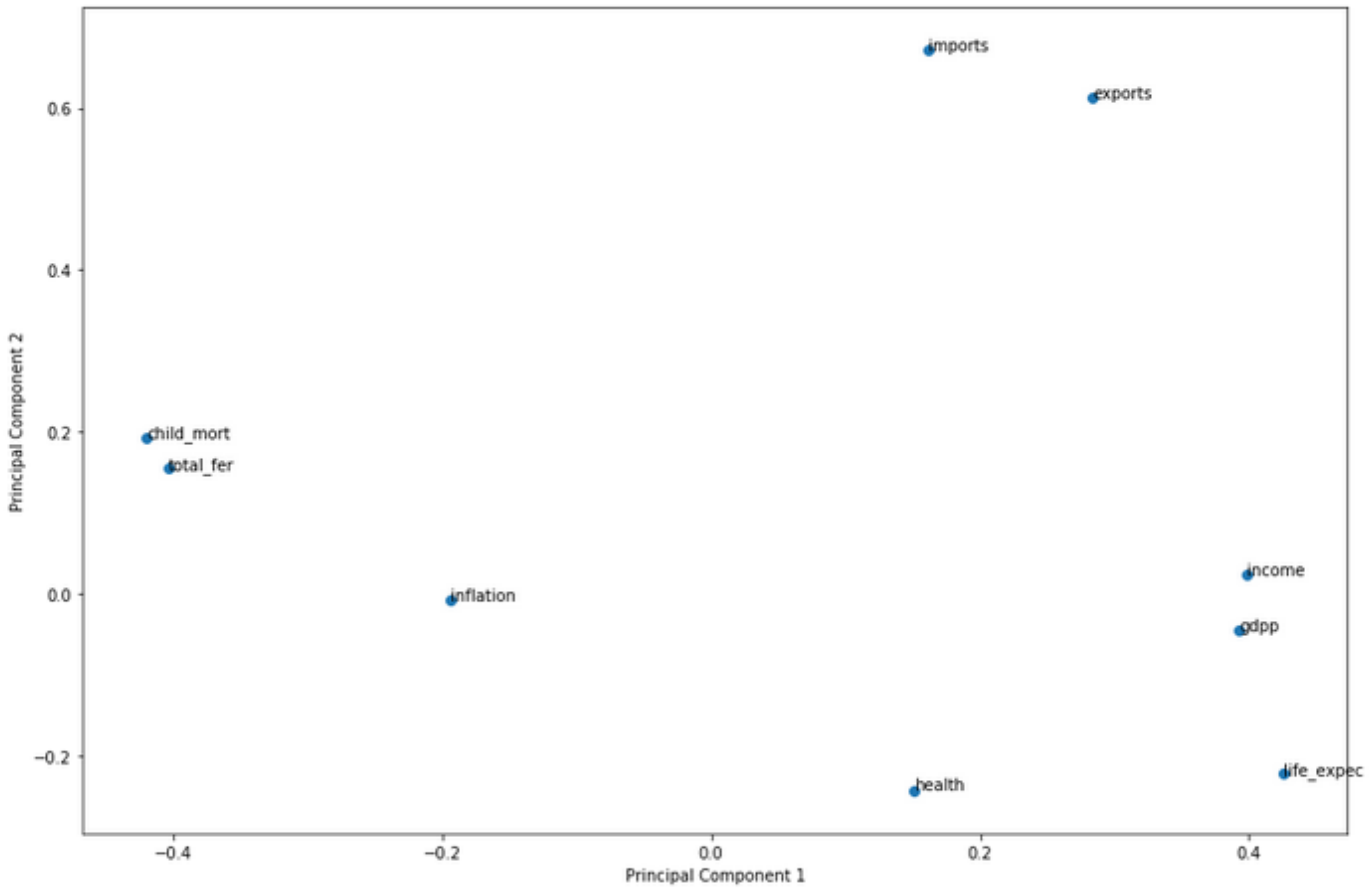


Principle Component Analysis - Overview

- During the analysis the high dimensionality of data was reduced.
- As the name suggest we have chosen four principle components .
- Respective feature column is presented along with the components.
- Features dimensionality reduction would be done using this technique handling outliers and removing multi co linearity.

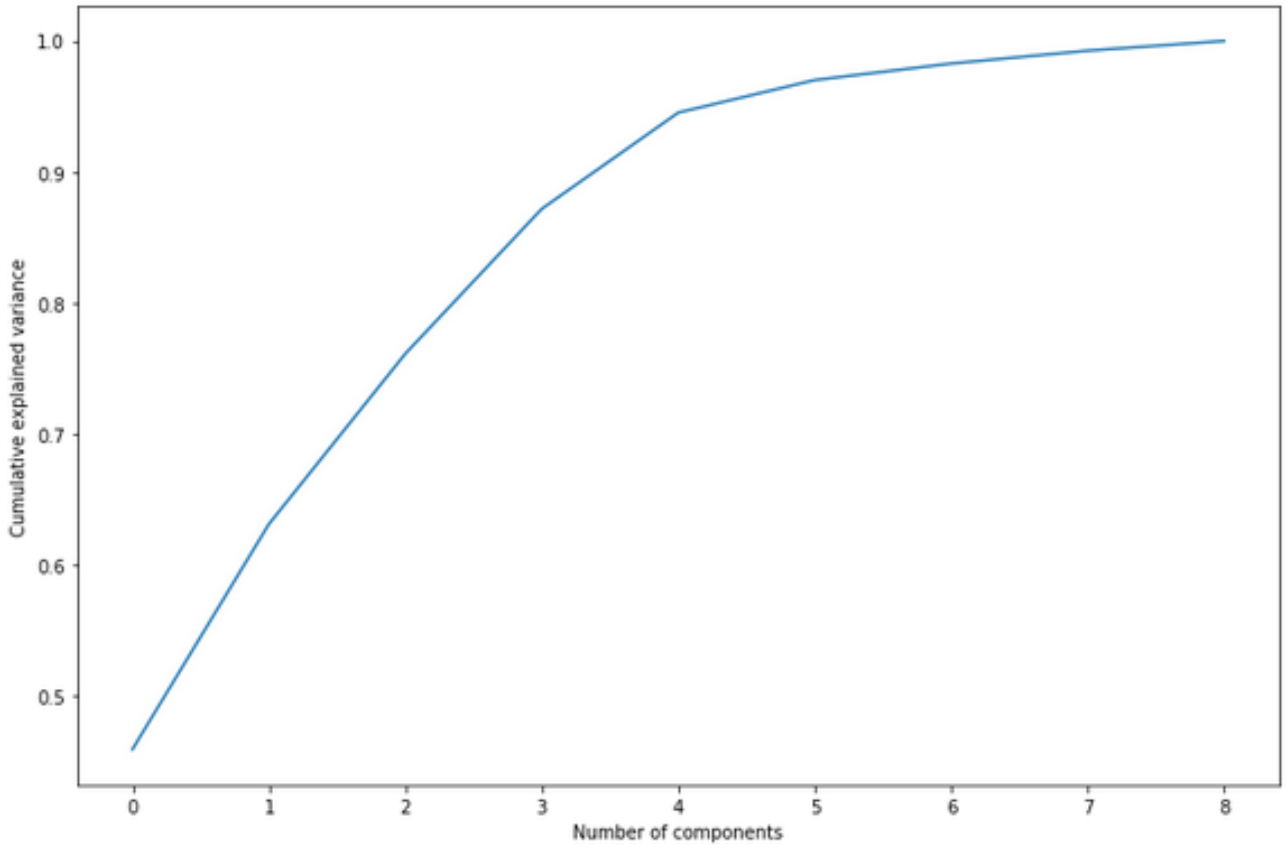
	PC1	PC2	PC3	PC4	Feature
0	-0.419519	0.192884	-0.029544	0.370653	child_mort
1	0.283897	0.613163	0.144761	0.003091	exports
2	0.392645	-0.046022	0.122977	0.531995	gdpp
3	0.150838	-0.243087	-0.596632	0.461897	health
4	0.161482	0.671821	-0.299927	-0.071907	imports

• Please find below the scatter plot showing the relation between PC1 and PC2. Outliers are clearly visible in the dataset provided. Also we could see that the features child_mort, exports and so forth depending on the country are distributed.



Scree Plot

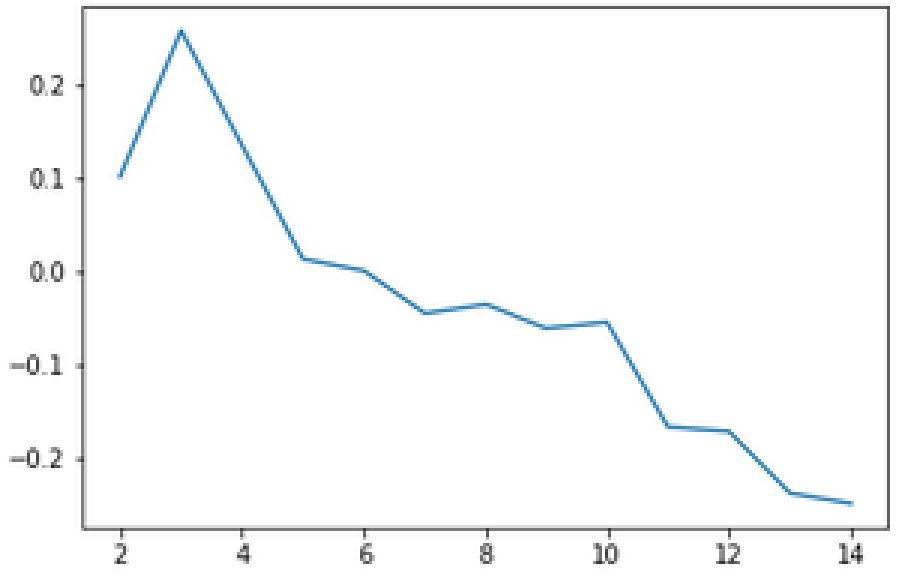
Below shown is the Scree Plot showing the variance(ideally should be between 80-90%) and on the basis of this the decision of component selection could be done.



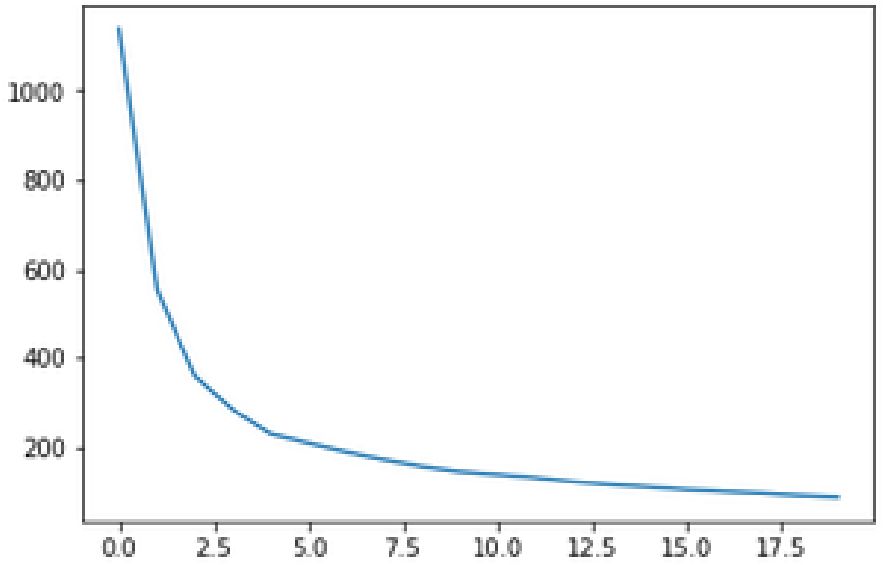
Approach towards Hopkins Analysis

- It indicates how the cluster tendency is and how are they formed.
- 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
- So if the hopkins value is between 0.7 and 0.9 then its we know we could invest in them.

- Silhouette Analysis

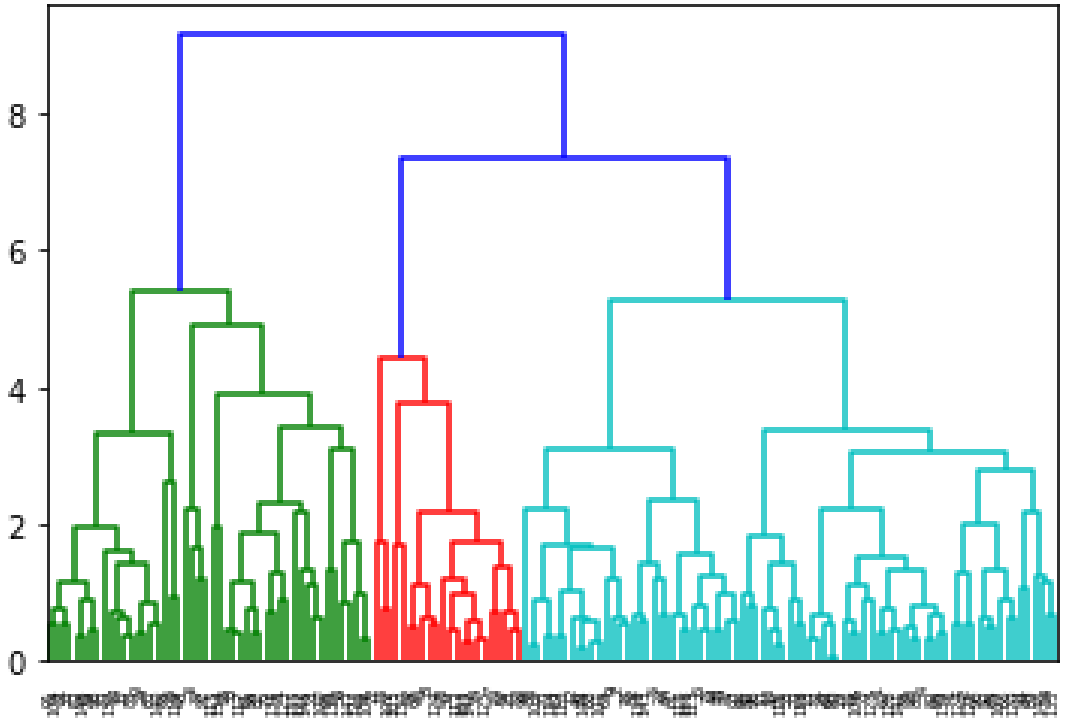


- Sum of squared distance



Hierarchical clustering

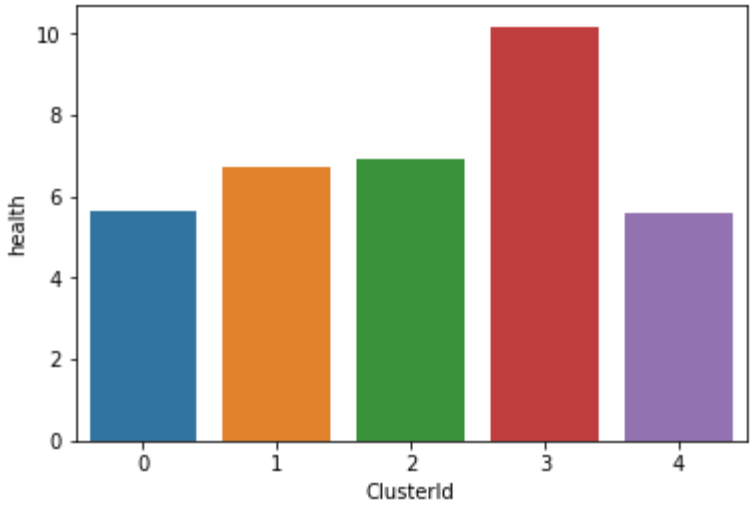
- Now this clustering is another way of representing the clusters in our dataset.
- So the diagram below shows is known as the Dendogram, where we can how the dataset for the components are divided and they are dependent on each other

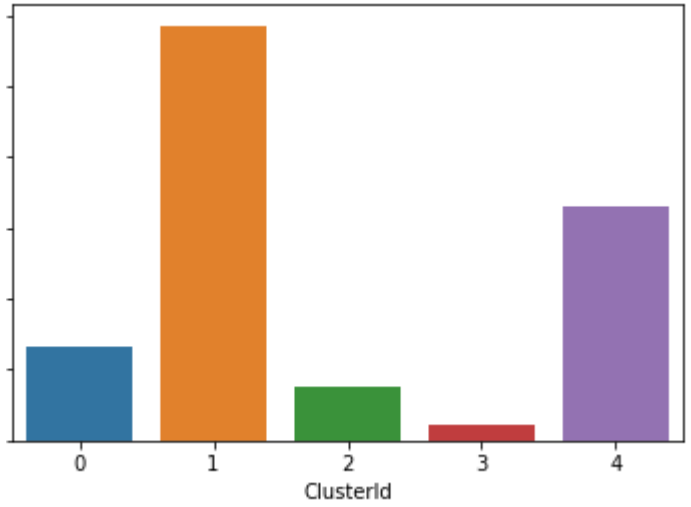


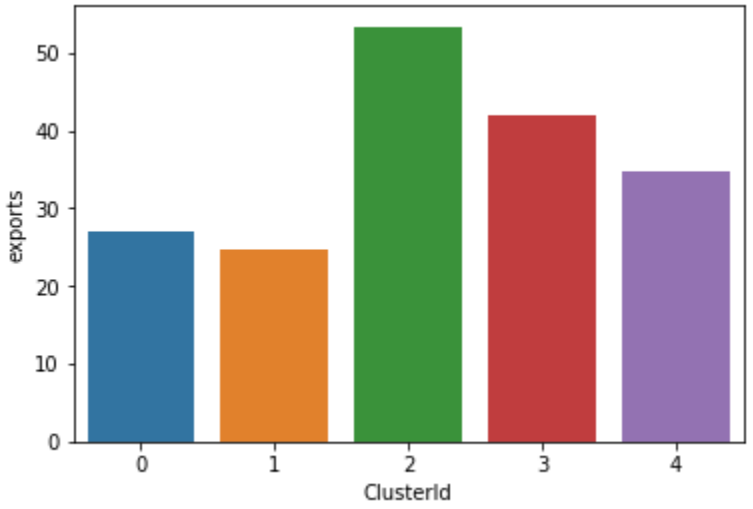
Identifying the Cluster with Poor Country

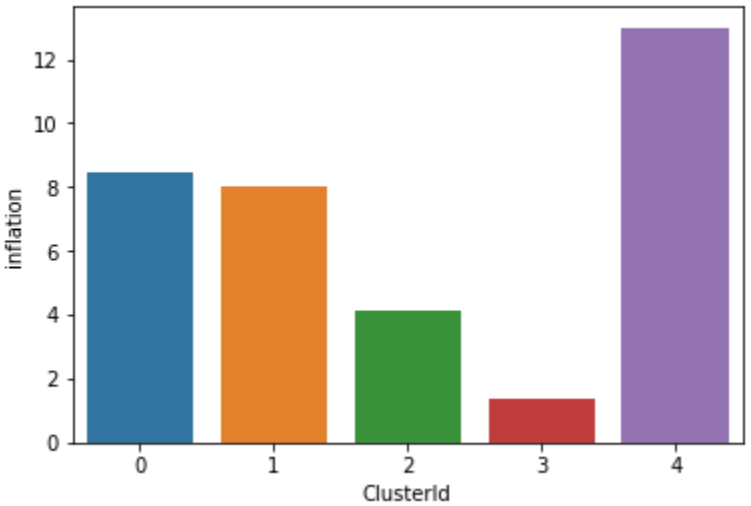
Below clusters show frequency : Similarly there are more graphs shown

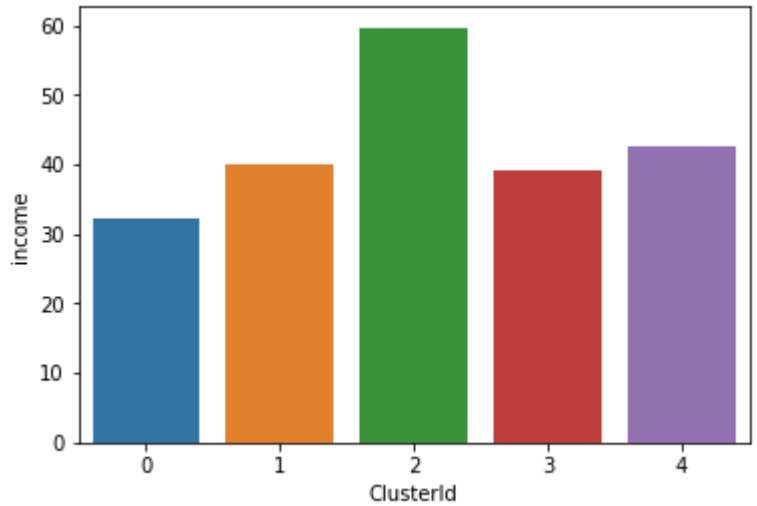
High loan amount, interest rates and amount invested
leads to more default

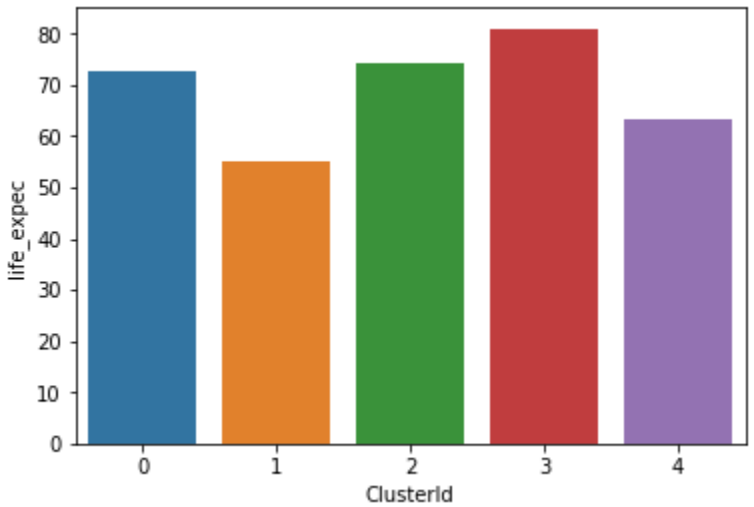


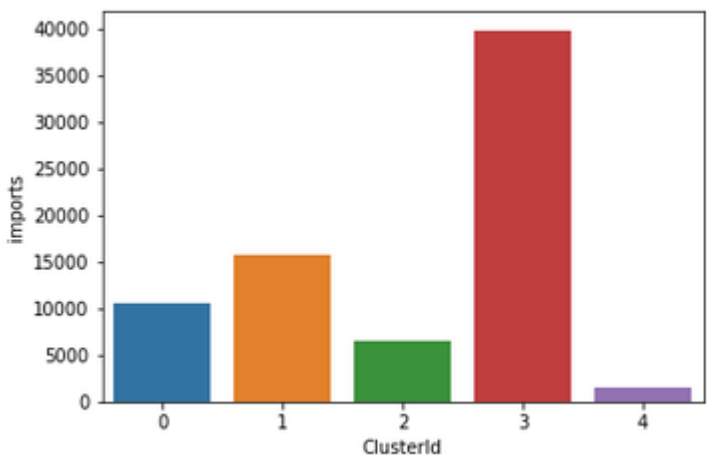


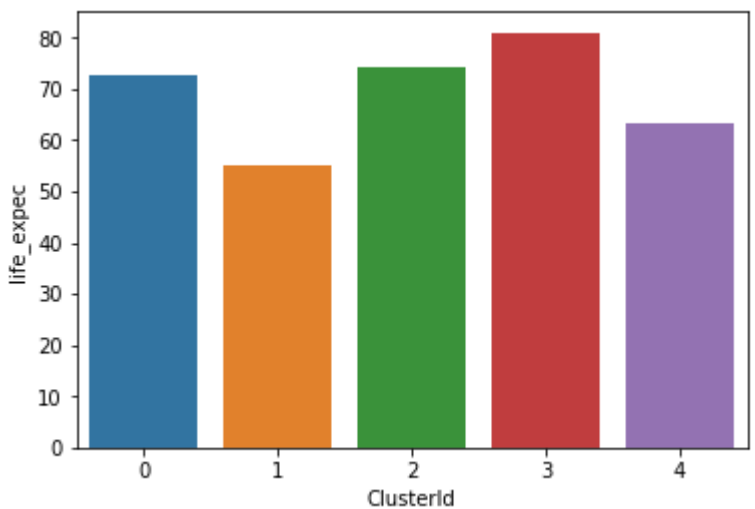


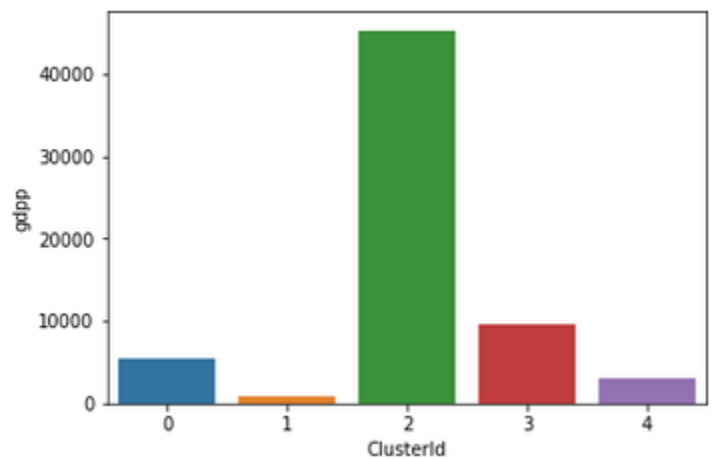

inflation
ClusterId


income
ClusterId


life_expec
ClusterId


imports
ClusterId


life_expec
ClusterId


gdpp
ClusterId

Result Analysis -- Conclusion

Hence we conclude by this analysis that :

1) After clustering cluster 4 has the lowest frequency, meaning this cluster contains the countries in need.
- (Please note this clustering may change dynamically everytime we run the code.)
2) We have seen total of 23 country under this cluster who's socio economic growth is little, which needs help from NGO
3) Below is the countries list which needs the help:

	ClusterId	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
0	4	Afghanistan	90.2	10.00	7.58	44.9	1610	9.440	56.2	5.82	553
17	4	Benin	111.0	23.80	4.10	37.2	1820	0.885	61.8	5.36	758
25	4	Burkina Faso	116.0	19.20	6.74	29.6	1430	6.810	57.9	5.87	575
26	4	Burundi	93.6	8.92	11.60	39.2	764	12.300	57.7	6.26	231
28	4	Cameroon	108.0	22.20	5.13	27.0	2660	1.910	57.3	5.11	1310