Help International Clustering of Countries

SUBMISSION

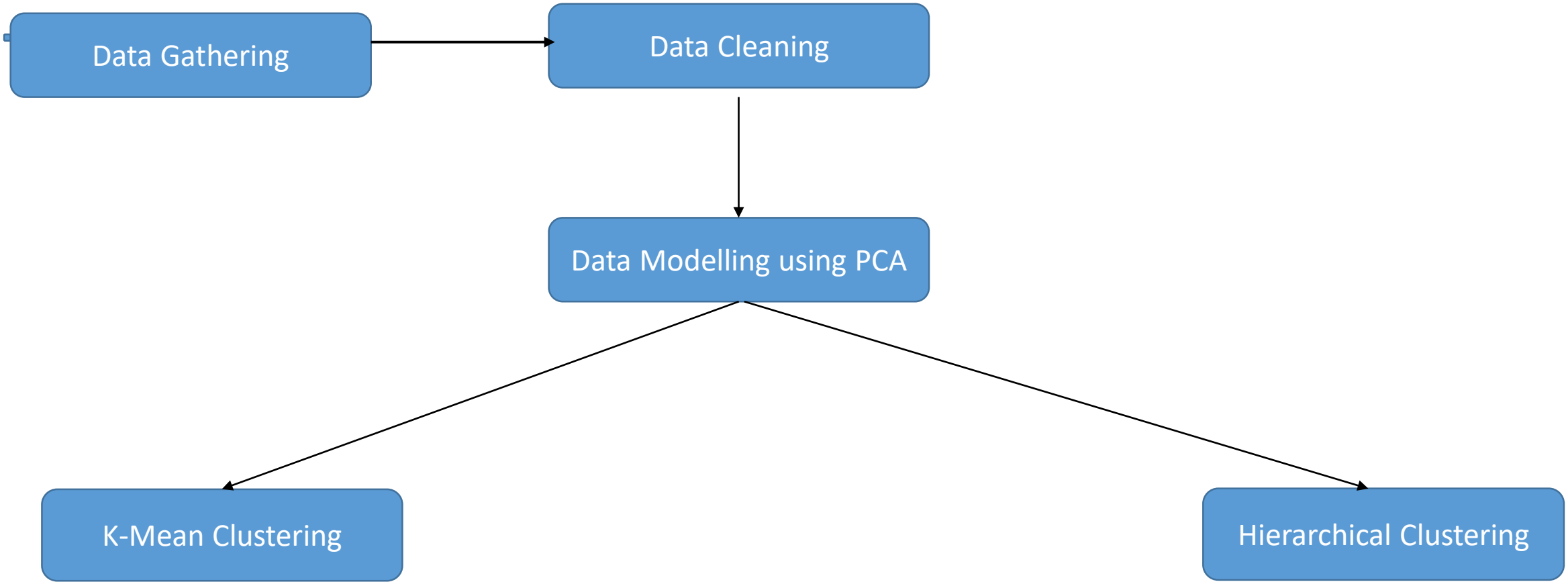
Akarshi Rastogi

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. 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 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.

Problem solving methodology



Analysis

To provide the best and profitable outcome, analysis is divided in further steps:

- **Data Cleaning:**

This is the first step we performed, clean the data and bring it in a state which is suitable for analysis.

Check for the NULL, duplicate values if exists then remove it.

Remove the non required column, if exists.

Check for the outliers in the data using boxplot. Since in this case, data is too small in number so we have not remove the outliers

Data Modelling

Data Modelling has been achieved using PCA(Principal Component Analysis)

1. **Principal Component Analysis**:-

Principal component analysis (PCA) is a statistical procedure that uses orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components in such a way that every principal component tries capturing the maximum of remaining variance. The positions of columns play no role in any calculation.

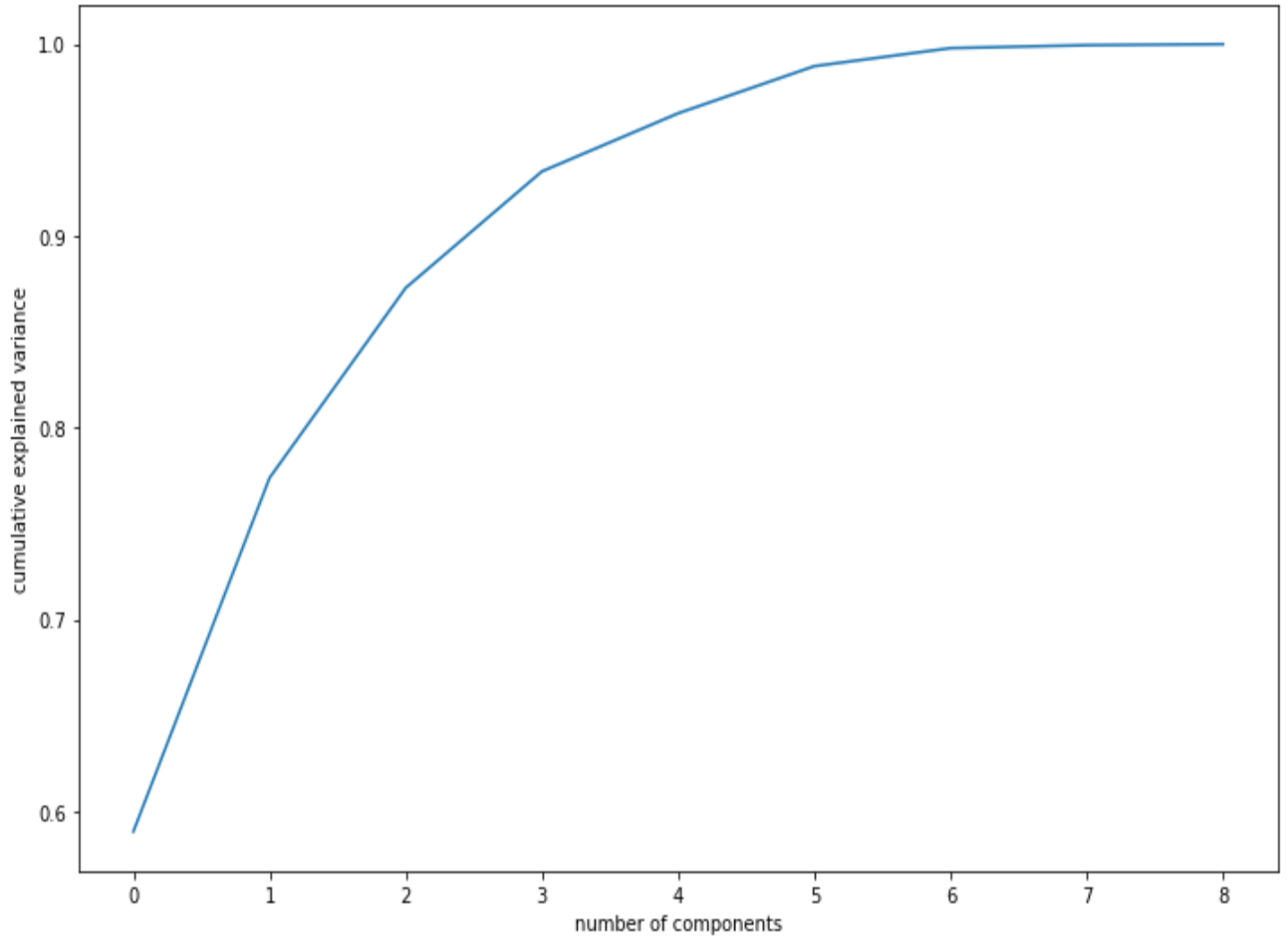
This will give the PCs of the original dataset. Note that the PCs are aligned along the rows. Thus the first PC is denoted by the first array of the m PCA finds the principal components z1,z2....zp such that each principal component is independent (i.e. perpendicular) to each other and captures as much variance as possible. The principal components are linear combinations of the original features.atrix that you obtain by applying this function pca.components_.

PCA

Scree Plot:
Plotting has been done among Number of Components and Cumulative explained Variance

Clearly over 95% of the data is properly explained by the first 4 principal components.

4 principal components an be analysed via the curve points.

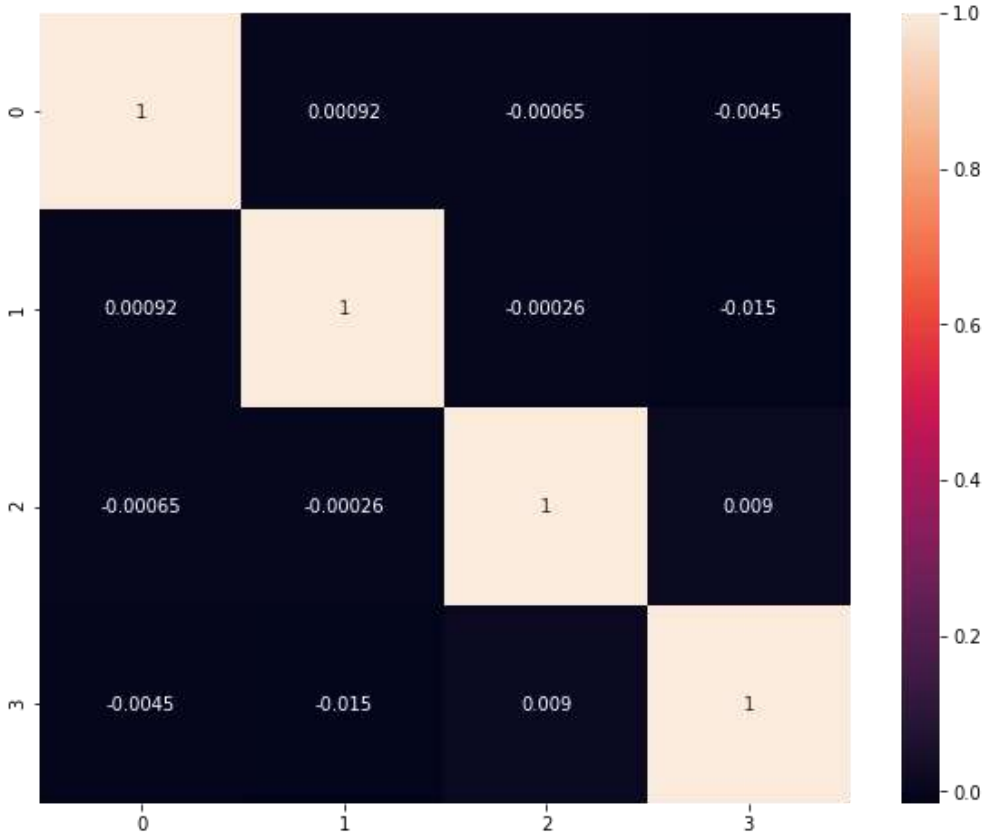


PCA

Up until now, you had learnt how to create a scree-plot and measure the number of components required to explain a certain percentage of variance. Now after you choose the number of components, you need to project the data onto only those number of components.

Creating correlation matrix for the principal components –

we expect little to no correlation

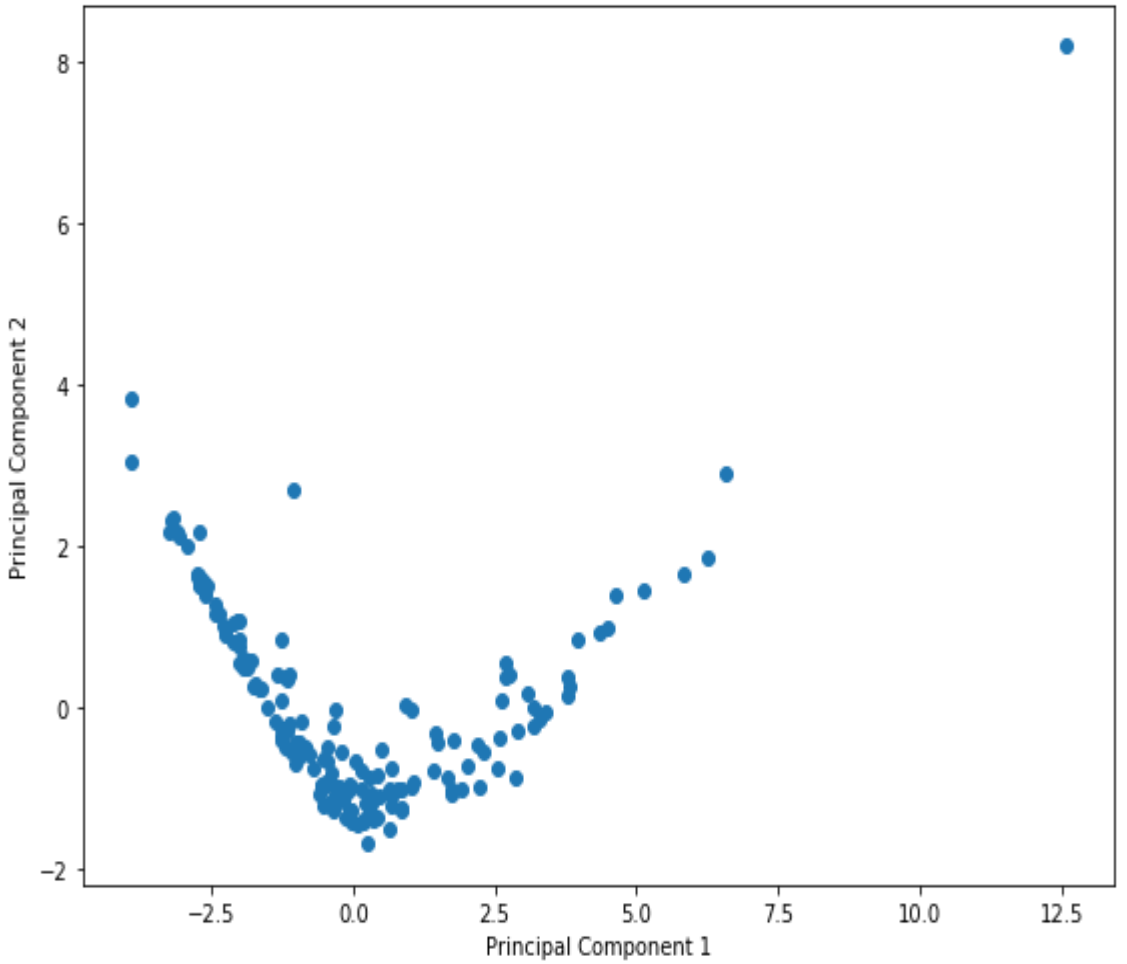


PCA

PCA helped us solve the problem of multicollinearity (and thus model instability), loss of information due to dropping variables, and we don't need to use iterative feature selection procedures.
Here from the image we can identify how data is segregated over a x-y plane.

Now, after PCA we calculated the “HOPKIN’S STATISTICS” to check if clustering is possible on the data or not.
HOPKIN STATISTICS value calculated is **0.908**

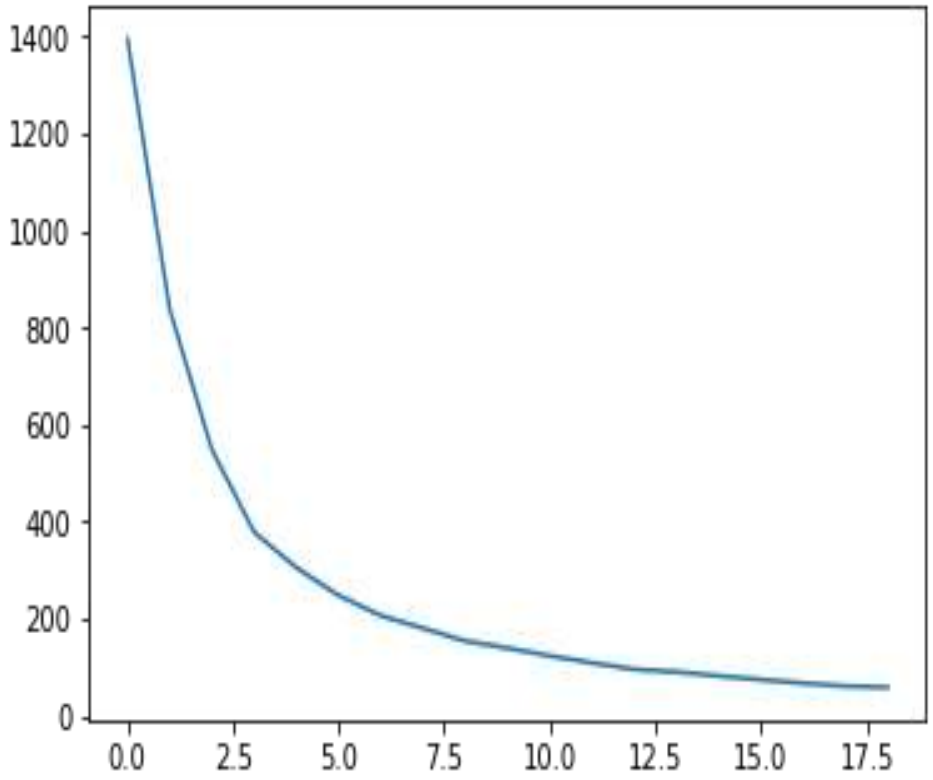
Since the value > 0.5 hence clustering is possible



K-Mean Clustering

the K-means algorithm is a clustering algorithm that takes N data points and groups them into K clusters.
KMeans() stores the sum of the squared distance of the points to their respective clusters centres as inertia. In other words, inertia represents, how tightly the different clusters are formed

However, in this plot, you can notice a distinct elbow. Beyond the elbow point, the additional (marginal) decrease in inertia with each increase in the cluster number is not very prominent. Thus, the elbow in the curve gives an estimate of the optimal number K in K Means.

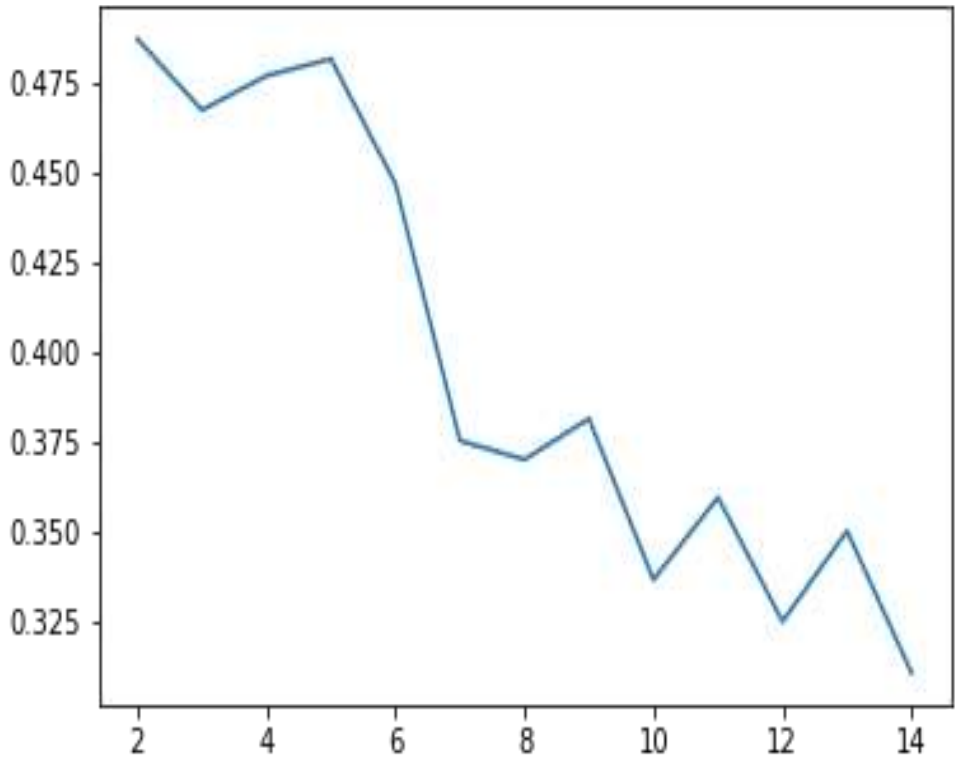


K-Mean Clustering

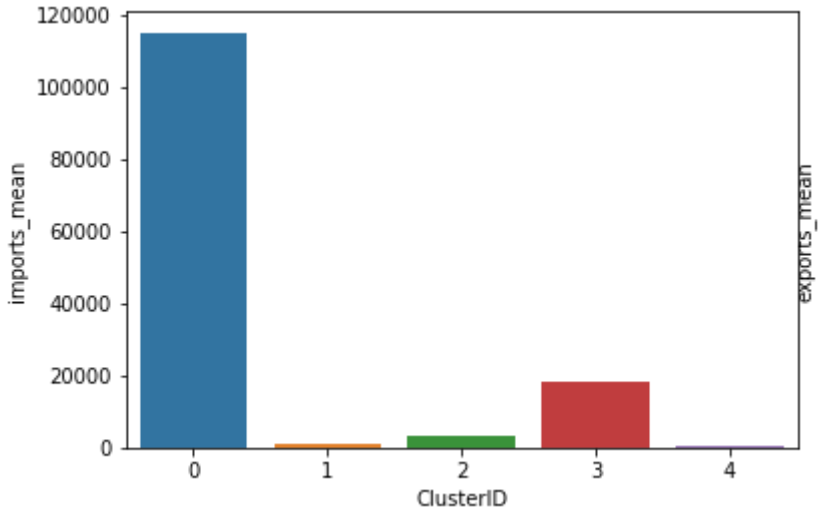
Then we calculated the **Silhouette curve** to find the possibility of number of clusters which can be formed.

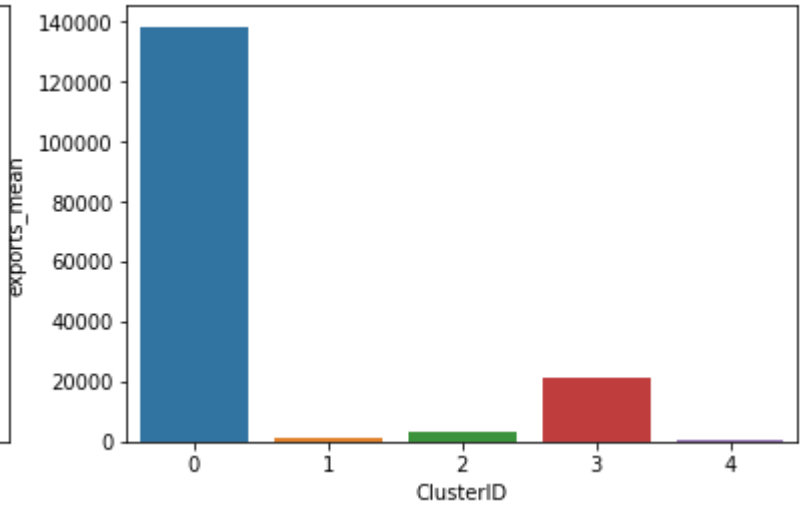
From the Silhouette score Analysis, we find that there are five curves seen in the graph.

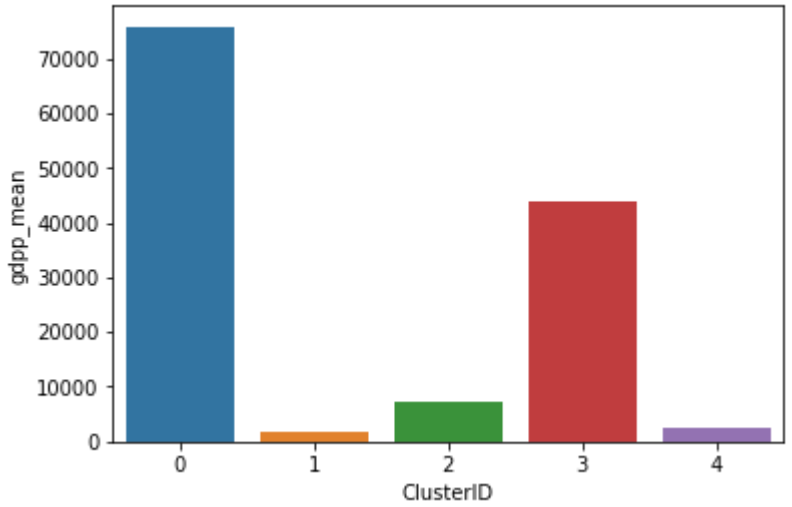
Hence, number of clusters to be taken will be 5

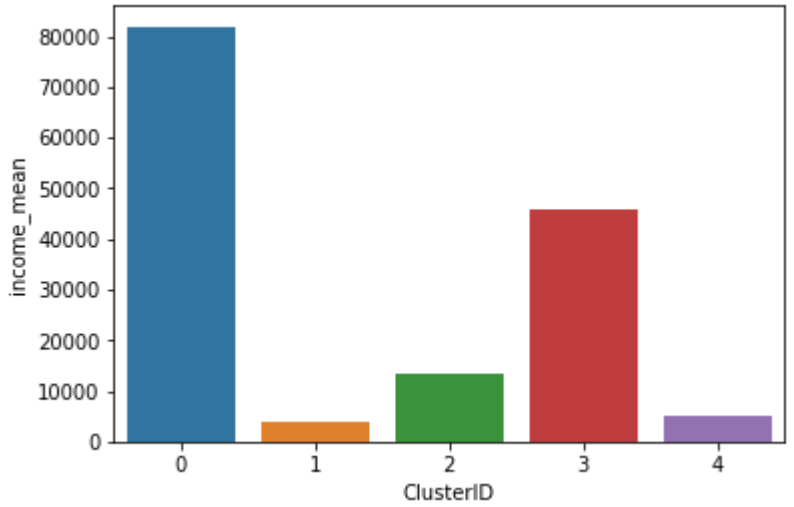


We found that cluster 0 and 3 was the best country from the growth point of view. This country has higher Gdpp, more life expectance, high income.
On the other hand, cluster 1 contains the countries in need of proper aid.









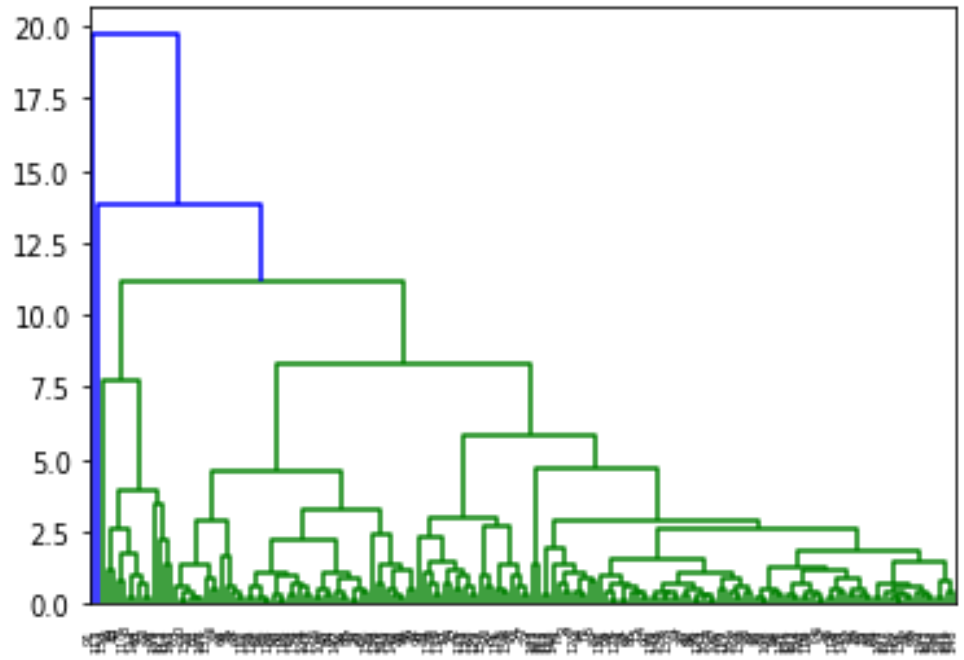
Hierarchical Clustering

One of the major considerations in using the K-means algorithm is deciding the value of K beforehand. The hierarchical clustering algorithm does not have this restriction.

The output of the hierarchical clustering algorithm is quite different from the K-mean algorithm as well. It results in an inverted tree-shaped structure, called the dendrogram.

Here, we can see the complete linkage of the hierarchical

Clustering. Then we cut along the y- axis in such a way that maximum amount of information can be achieved.



As observed from the Hierarchical and K-Mean Clustering graphs, Cluster ID 1 seems the most ill need of aid cluster as per the factors analysed:

low gdpp low net income high child mortality and low life expectancy.

So, the five country which are in direct need of proper aid are : **`uganda`, `togo` , `tazania` , `sierra leone`, `Rwanda`**