Clustering & PCA Assignment

HELP International

Submitted By:

Anusha Hitendra Kulkarni

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 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.

Objectives & Goals of Analysis

BUSINESS Objectives:

- The objective is to categorise the countries using some socio-economic and health factors that determine the overall development of the country.
- Suggest the countries which the CEO needs to focus on the most for funding purposes.

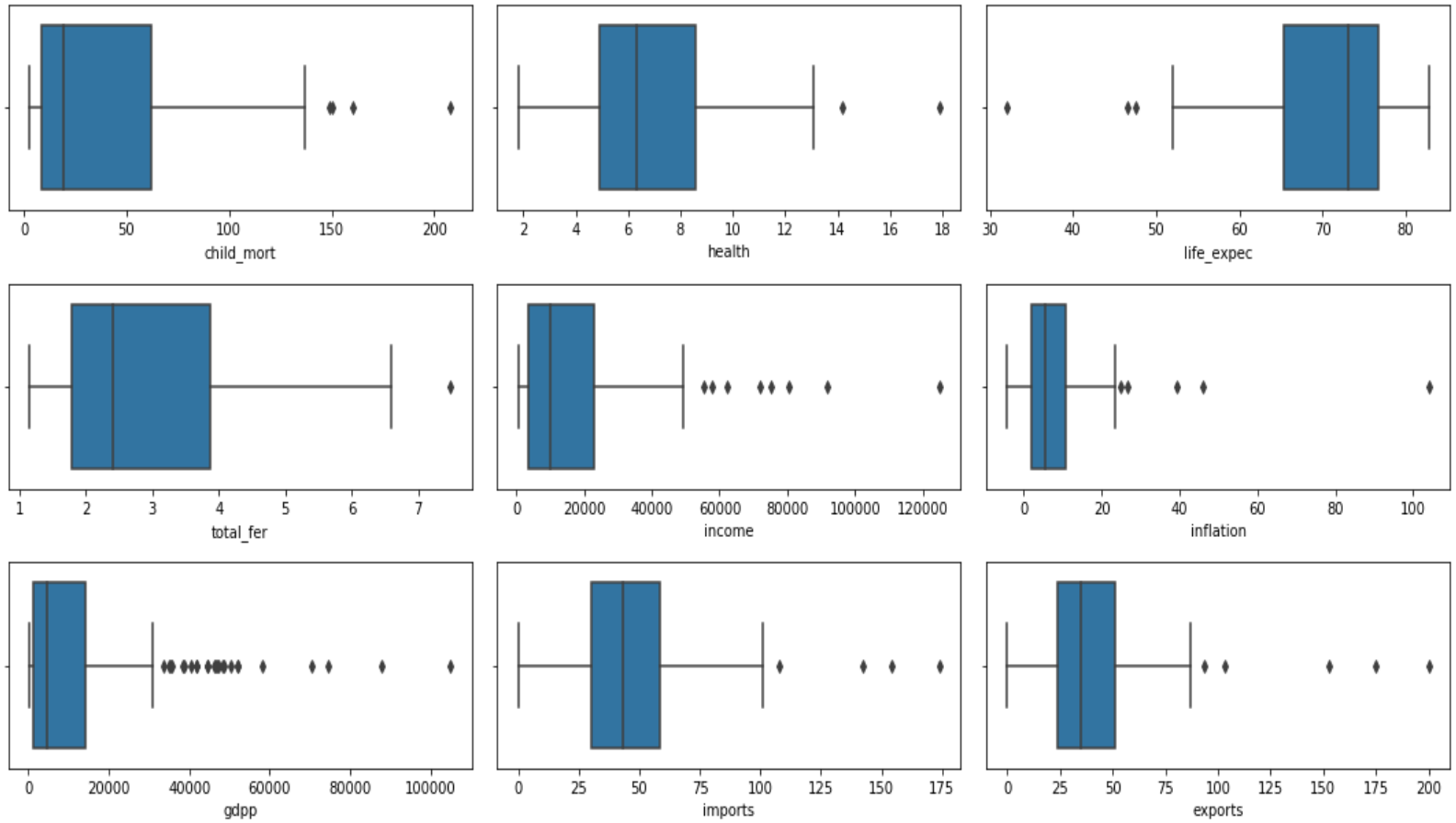
Goals of Analysis:

- Perform PCA on the dataset and obtain the new dataset with the Principal Components.
- Perform K-means and Hierarchical clustering on this dataset and create clusters.
- The final list of countries depends on the number of components that you choose and the number of clusters that you finally form.

Steps of Analysis

- Data Preparation and Data Cleaning
- Exploratory Data Analysis : Univariate Analysis
- Outliers Analysis
- Standardization(Scaling) of Data
- Principal Component Analysis (PCA) On Data
- Performing K-Means Clustering
- Performing Hierarchical Clustering
- Analysis Of Clusters And Getting the Names of Countries

Outliers Analysis



- There are a number of outliers in the data.
- Since, the K-Means algorithm tries to allocate each of the data point to one of the clusters, outliers have serious impact on the performance of the algorithm and prevent optimal clustering.
- Keeping in mind we need to identify backward countries based on socio economic and health factors.
- We will cap the outliers to values accordingly for analysis.

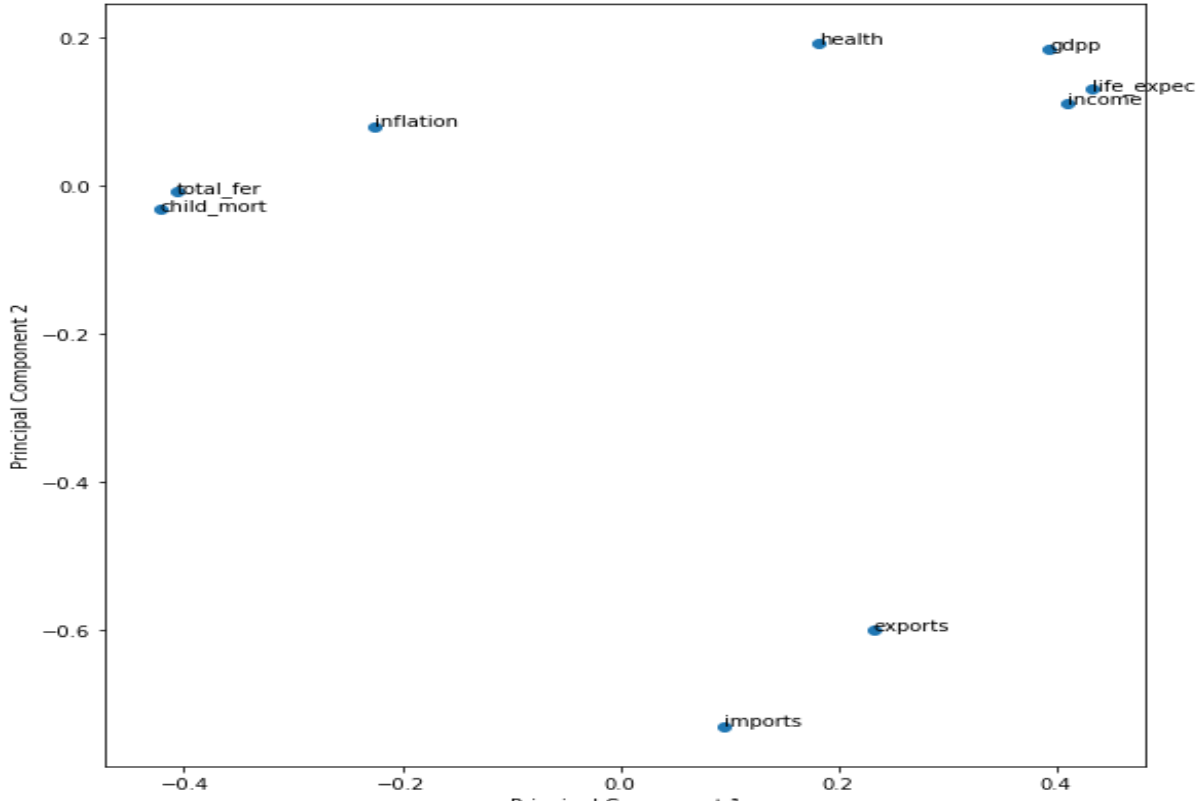
Standardization Of Data

Standardisation of data, that is, converting the m in to z-scores with mean 0 and standard deviation 1, is important for 2 reasons in K-Means algorithm:

- Since we need to compute the Euclidean distance between the data points ,it is important to ensure that the attributes with a larger range of values do not out-weight the attributes with smaller range.
- Thus, scaling down of all attributes to the same normal scale helps in this process.
- The different attributes will have the measures in different units. Thus, standardisation helps in making the attributes unit-free and uniform.

Principal Component Analysis(PCA)

Principal Component Analysis, finds uncorrelated components that are linear combinations of the original variables and capture the variance/ information in the data. Another interpretation is that it finds the best lower dimensional approximation of the data.

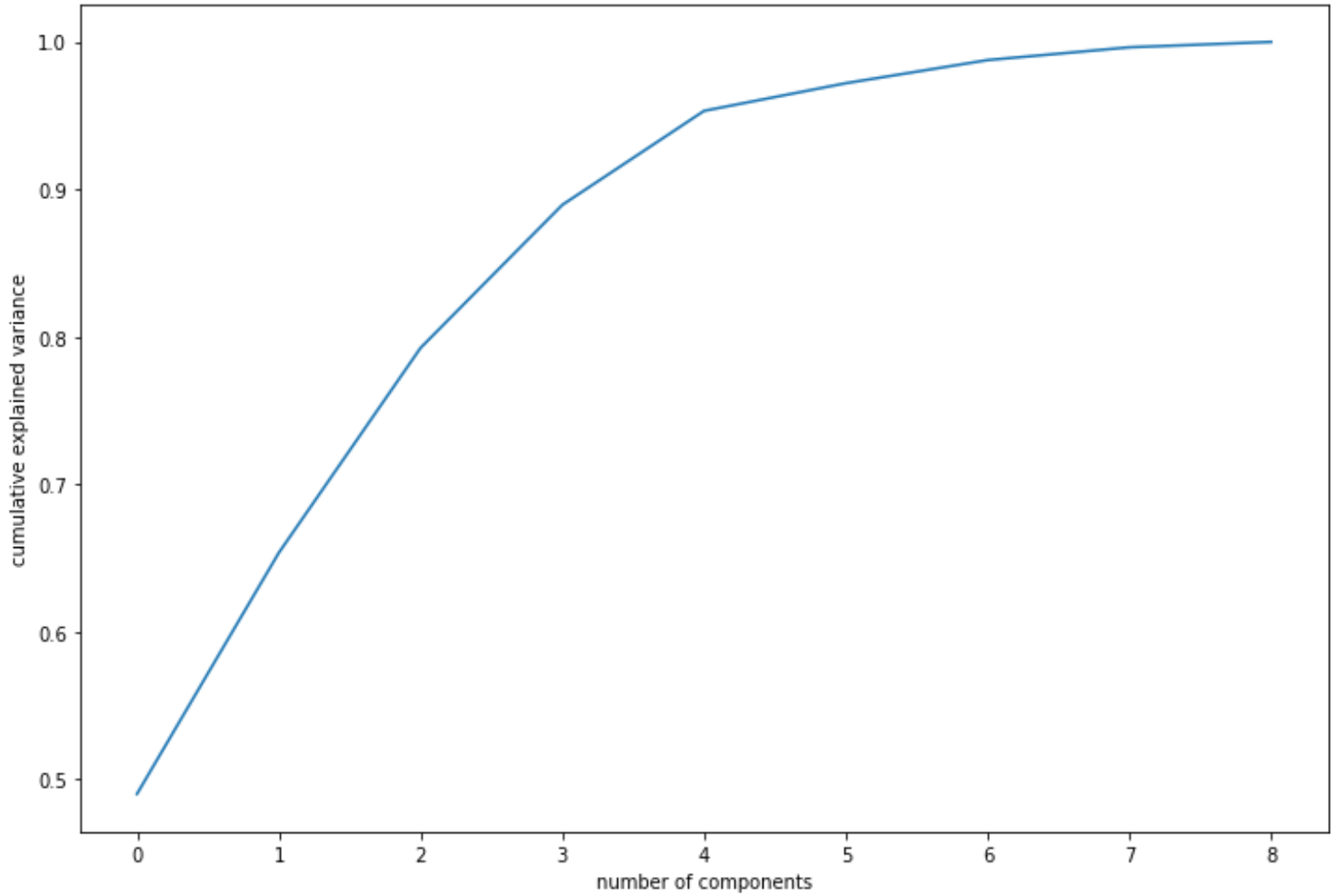


ScreePlot

Looking at the screeplot to assess the number of needed principal components.

- Looks like **4** components are enough to describe 95% of the variance in the dataset

- We'll choose 4 components for our Modelling.



K-means Clustering

The **Hopkins statistic**, is a statistic which gives a value which indicates the cluster tendency, in other words: how well the data can be clustered.

- 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.

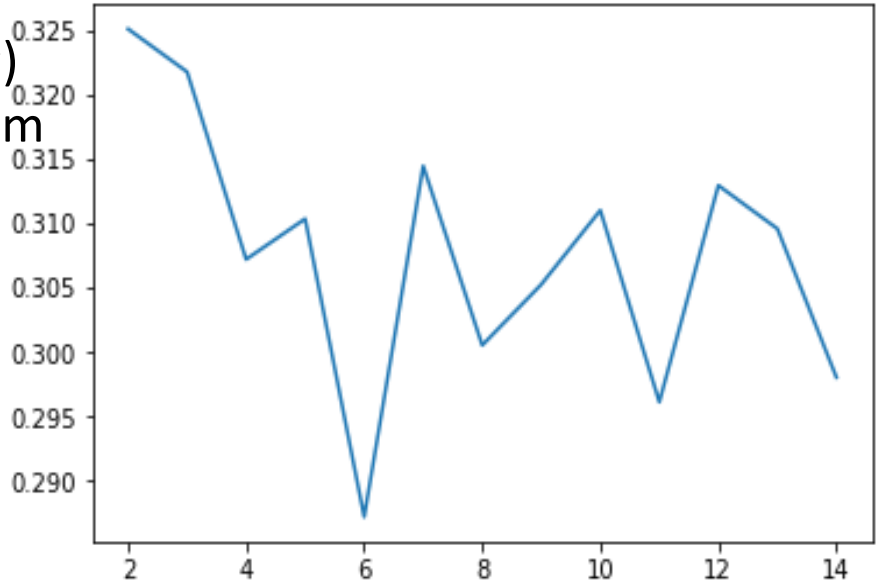
We are getting values of Hopkins Statistic above 0.7 so high tendency to cluster.

Choosing the number of clusters K in advance

1.Silhouette Analysis Method:

Compute clustering algorithm (i.e, k-means clustering) for different values of k. For instance, by varying k from 2 to 15clusters.

- For each k, calculate the average silhouette of observations (avg.sil).
- Plot the curve of avg.sil according to the number of clusters k.
- The location of the maximum is considered as the appropriate number of clusters.



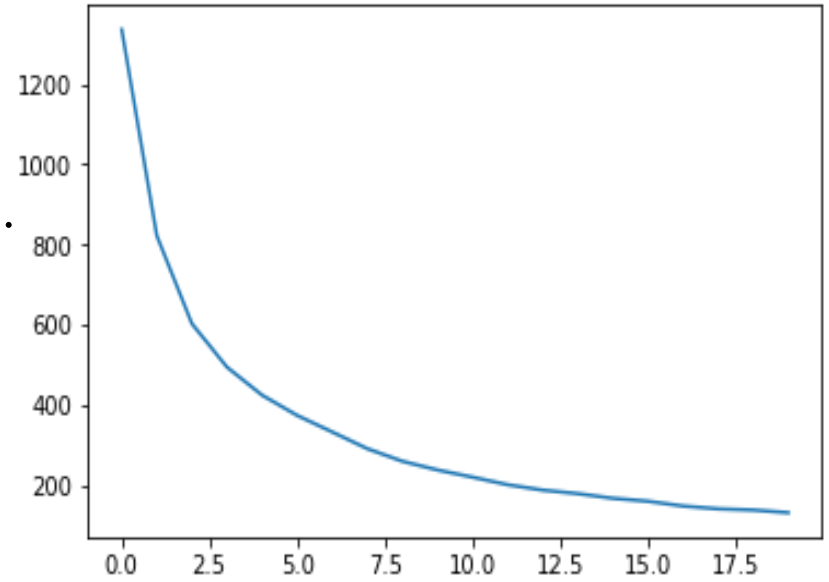
2. Elbow method:-

•Compute clustering algorithm (i.e. k-means clustering) for different values of k.

•For each k, calculate the total within-cluster sum of square (wss).

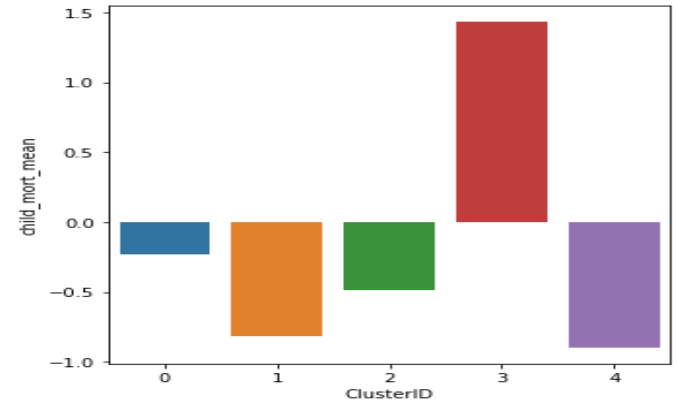
•Plot the curve of wss according to the number of clusters k.

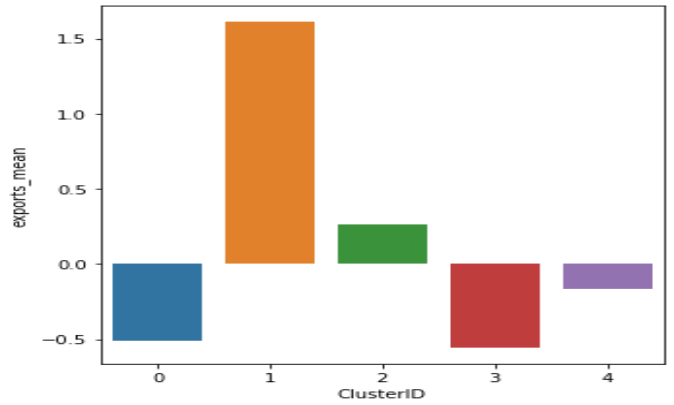
•The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.

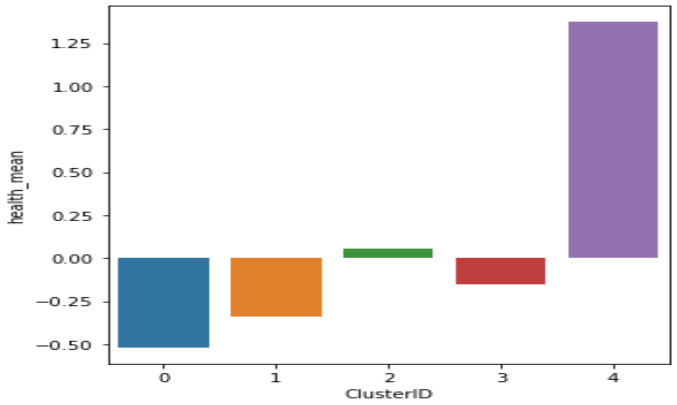


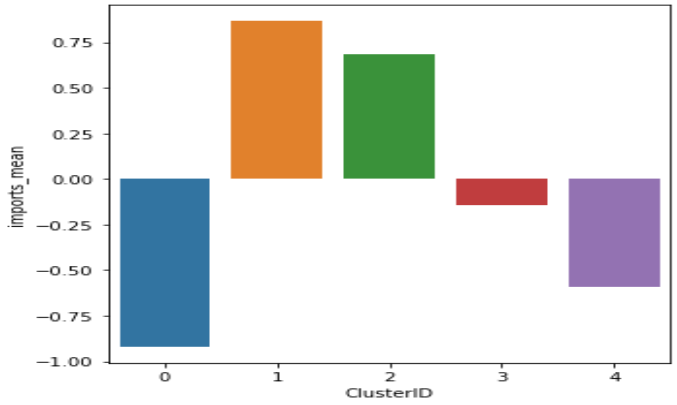
We have chosen number of cluster K as 5 by analysing above two methods.

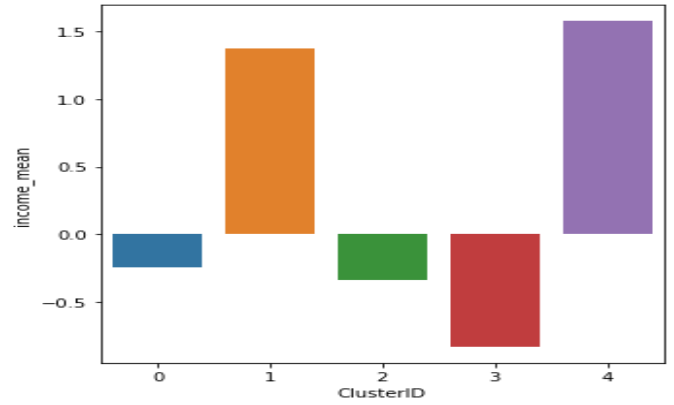
Analysis of Clusters Formed By K-Means Algorithm

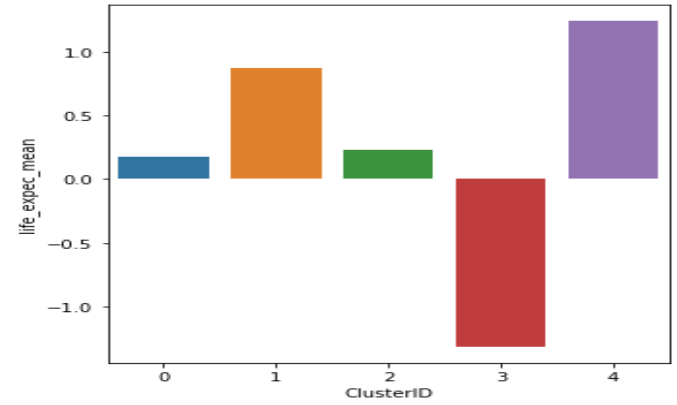


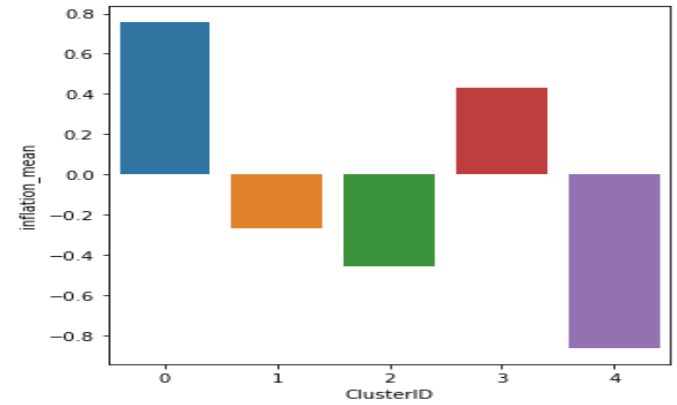


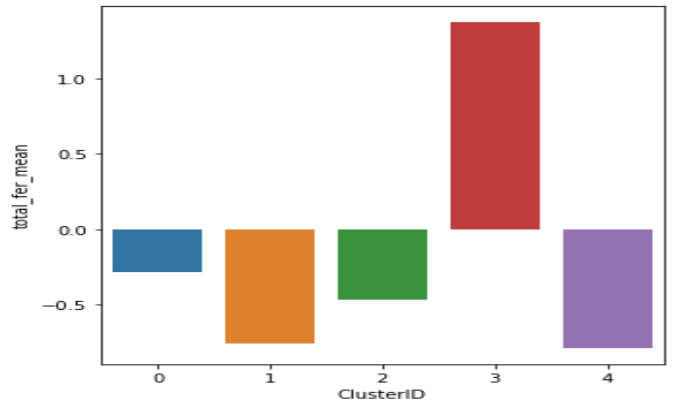


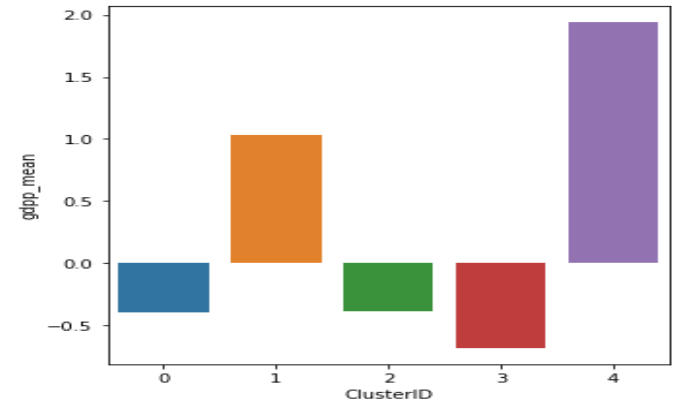






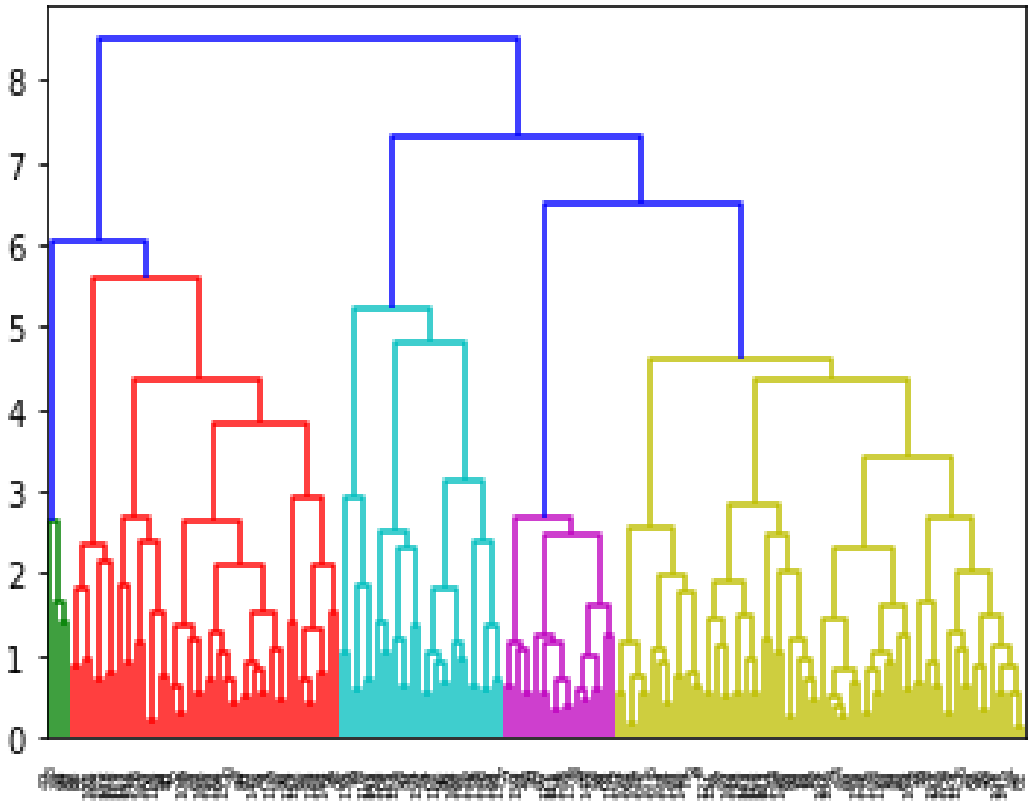




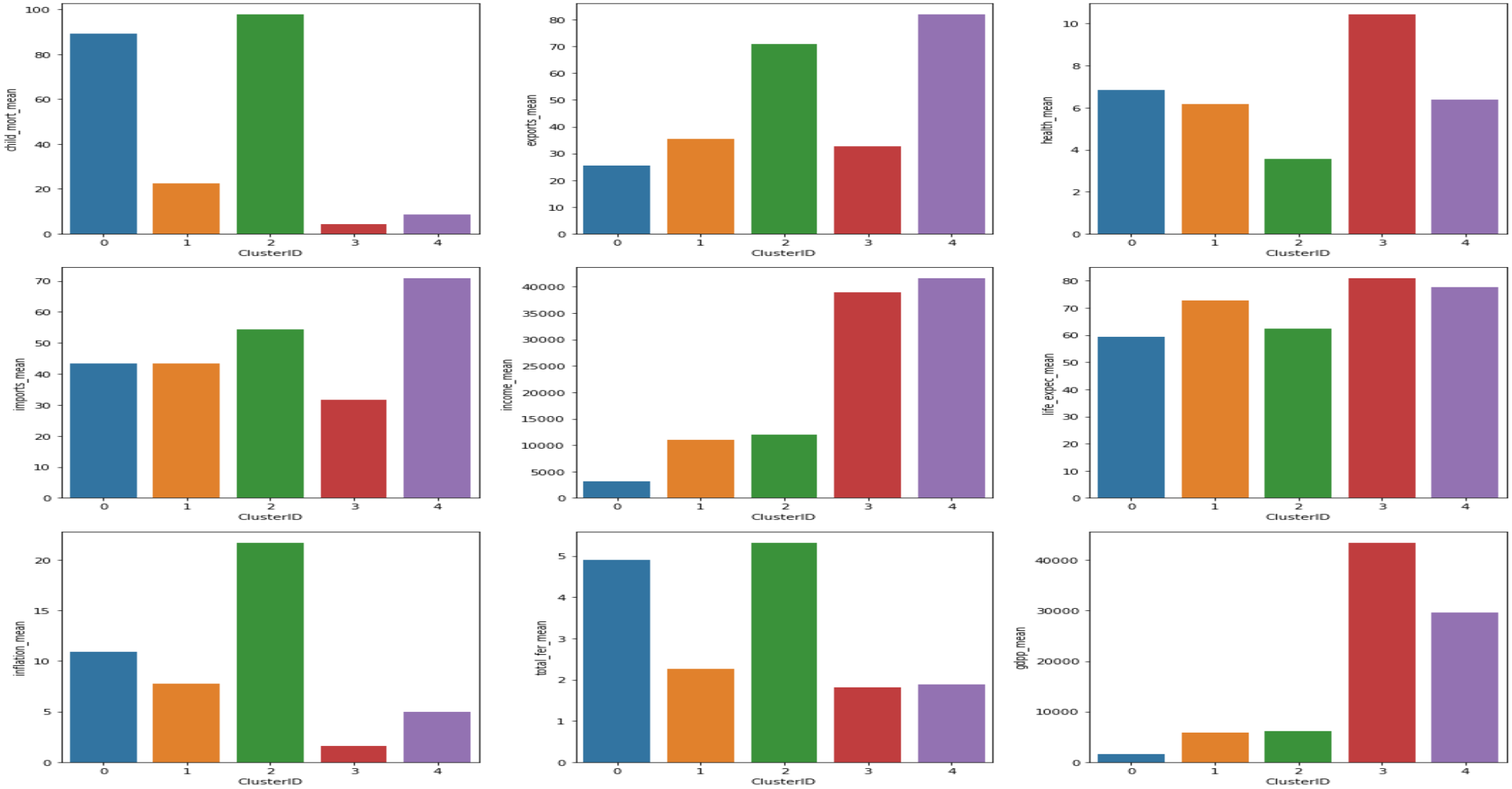


Hierarchical clustering

- The result of the hierarchical clustering algorithm is shown by a dendrogram, which starts with all the data points as separate clusters and indicates at what level of dissimilarity any two clusters were joined
- Once we obtain the dendrogram, the clusters can be obtained by cutting the dendrogram at an appropriate level. The number of vertical lines intersecting the cutting line represents the number of clusters.
- Thus if we cut the dendrogram at dissimilarity measure of 0.6, we obtain 5 clusters.



Analysis of Clusters Formed By Hierarchical clustering Algorithm



List Of Countries Need To Be Focused More

- Cluster with ClusterID as 0, is the cluster of most backward country.

- Countries on which we require to focus more are:

  Afghanistan', 'Benin', 'Botswana', 'Burkina Faso', 'Burundi', 'Cameroon', 'Central African Republic', 'Chad', 'Comoros', 'Congo, Dem. Rep.', "Cote d'Ivoire", 'Eritrea', 'Gabon', 'Gambia', 'Ghana', 'Guinea', 'Guinea-Bissau', 'Haiti', 'Iraq', 'Kenya', 'Kiribati', 'Lao', 'Lesotho', 'Liberia', 'Madagascar', 'Malawi', 'Mali', 'Micronesia, Fed. Sts.', 'Mozambique', 'Namibia', 'Niger', 'Nigeria', 'Pakistan', 'Rwanda', 'Senegal', 'Sierra Leone', 'Solomon Islands', 'South Africa', 'Sudan', 'Tajikistan', 'Tanzania', 'Timor-Leste', 'Togo', 'Uganda', 'Yemen', 'Zambia"