# Clustering & PCA Assignment

## Analysis

By:

RAJAT GUPTA

# Clustering & PCA Assignment

## Abstract

**Project Brief**

HELP is an NGO is committed to fight poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities.

- CEO of the NGO needs to decide how to use this money strategically and effectively.
- Significant issues that come while making this decision are mostly related to choosing the countries that are in the direst need of aid.

**Goal of data analysis:-**

- Categorize 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

**Business objective**

- Cluster the countries by the factors.
- Use dimensionality reduction using PCA to get the visualisations of clusters in a 2-D form

## Data Extraction and Preparation

- Extract Data and Import them into Python notebook as data frames
- Check for duplicates and null values
- Check outlier treatment and values

## Principal Component Analysis

- Drop non PCA related columns and fit transform
- Identify optimal number of PC using scree plot
- Perform PCA again with optimal number of components and obtain PCA dataset
- Perform outlier analysis and discard outliers

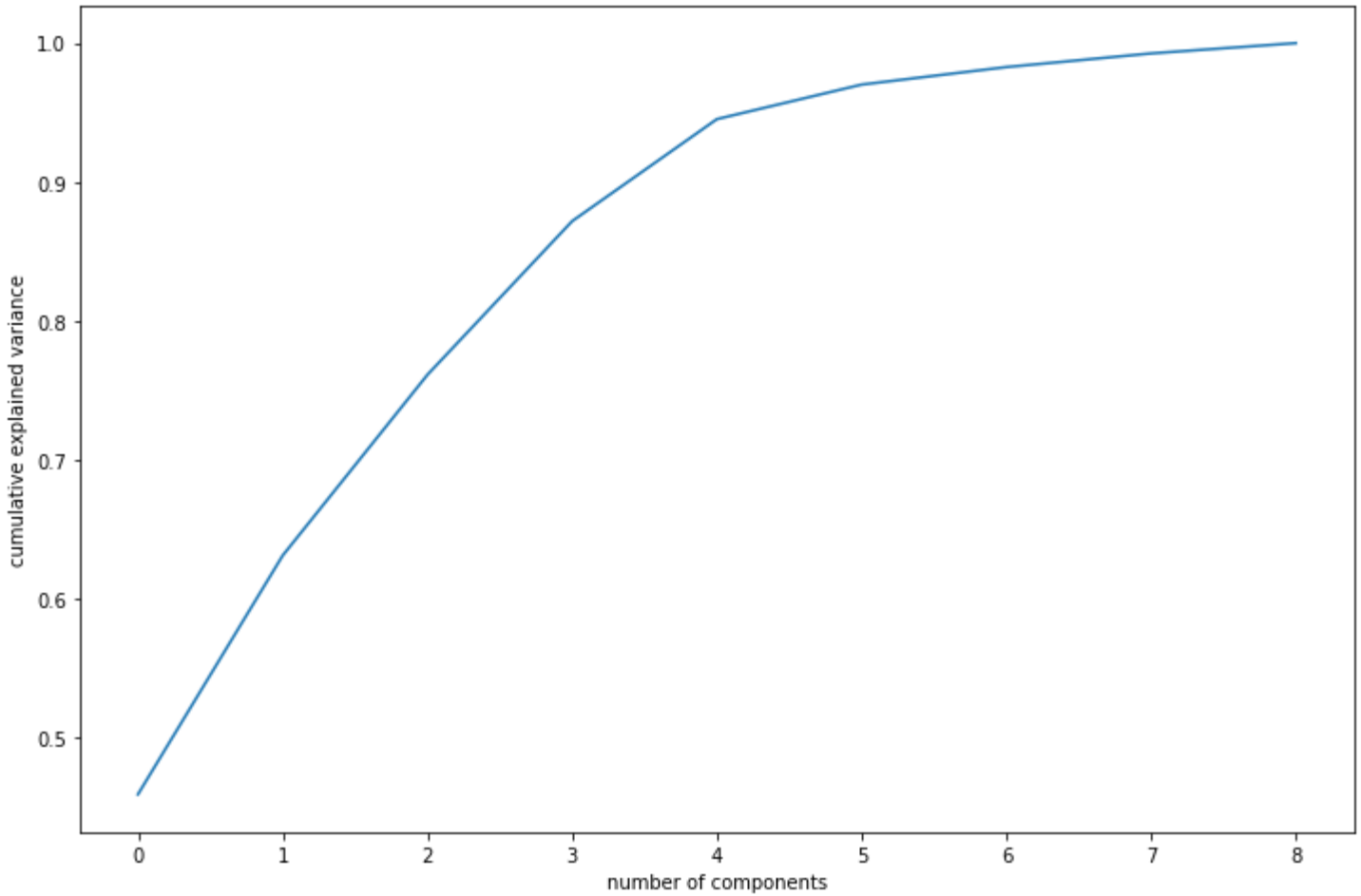
## K Means clustering

- Check if k-means can be performed using hopkins measure
- Perform silhouette and elbow analysis to determine optimal clusters
- Perform clustering with first K to obtain cluster id and join back the clustered data with original dataset
- Find one or more cluster fitting the criteria for funding

## Hierarchical clustering

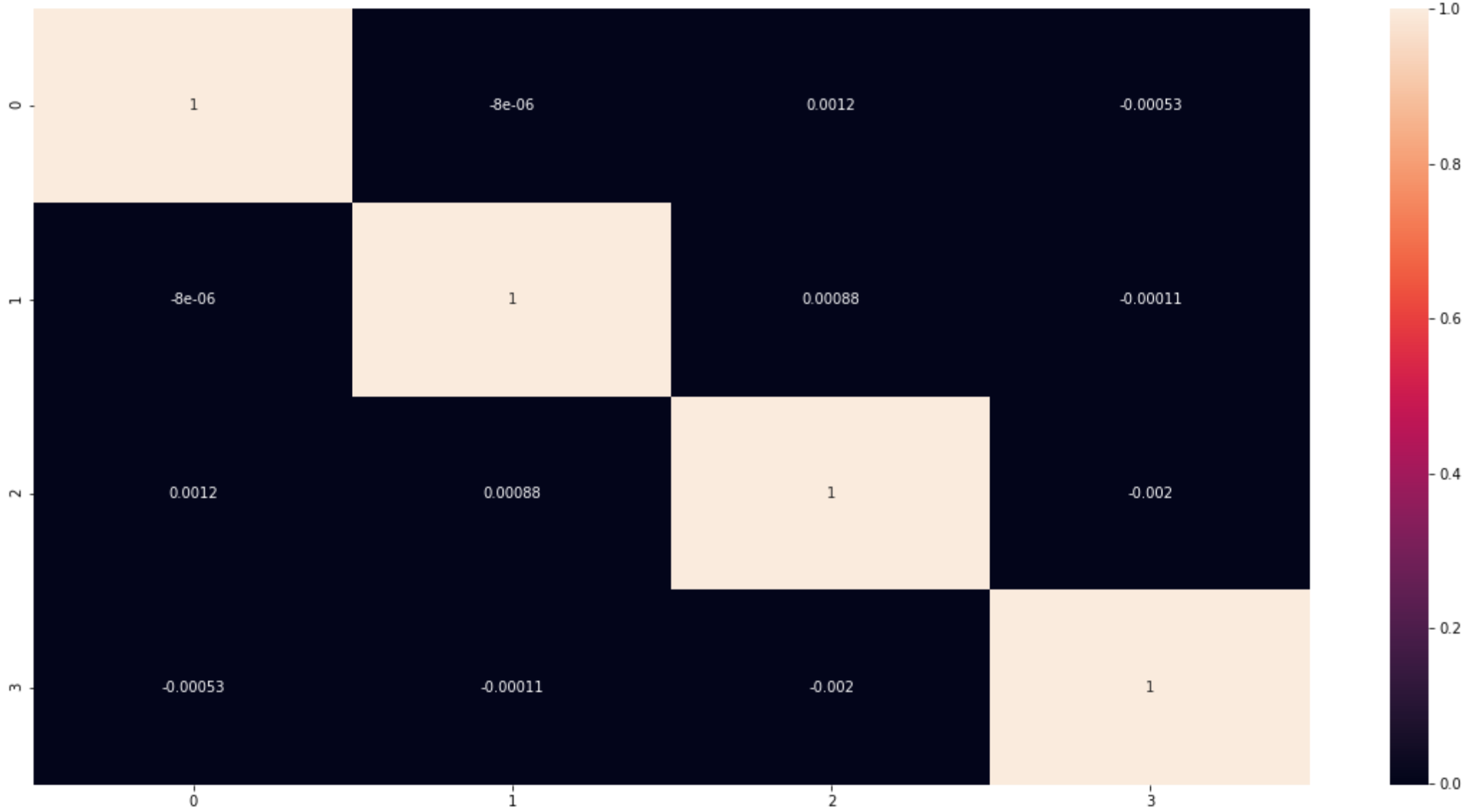
- Perform hierarchical clustering with single and complete linkage on PC dataset and obtain cluster id
- Join back the clustered data with original dataset and perform mean analysis for all columns per cluster
- Perform manual/visual analysis on the outliers countries that were dropped

## Scree plot for choosing components from Principal Component Analysis




Taking no. of components as 4 which are enough to describe 90-95% of the variance

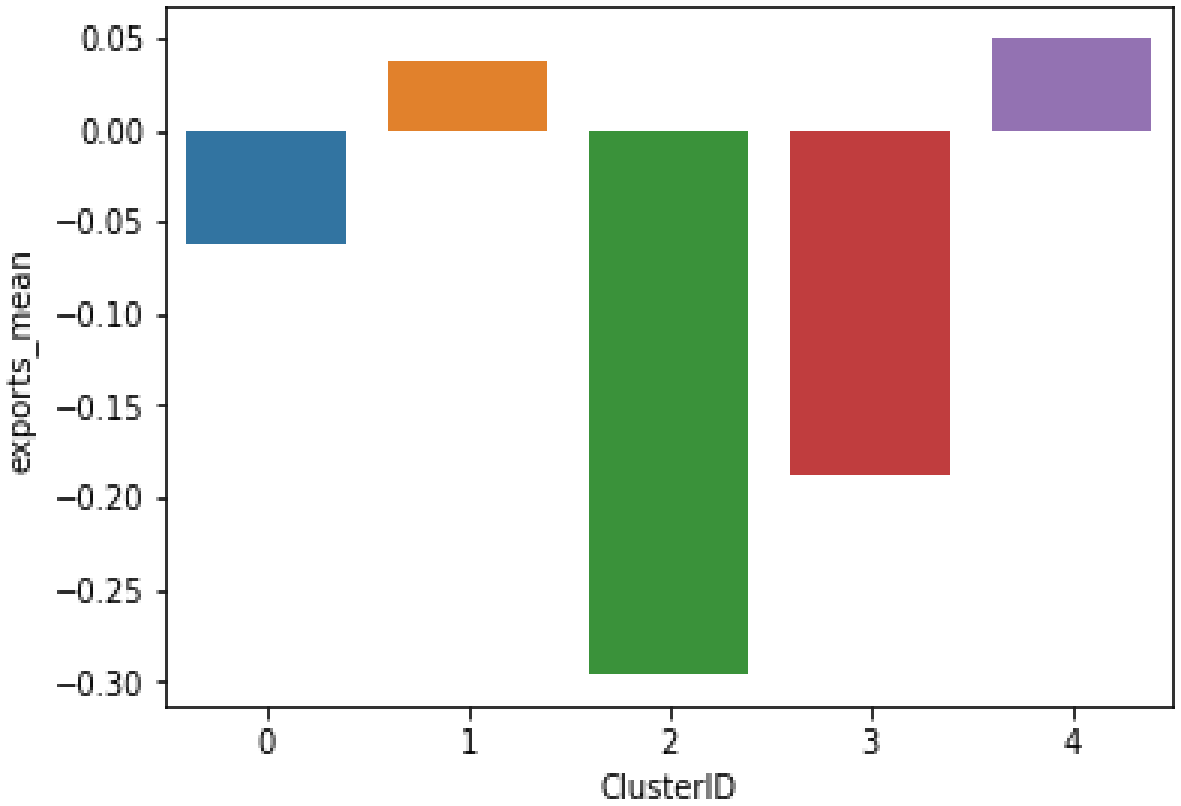
## Correlation matrix for 4 principal components



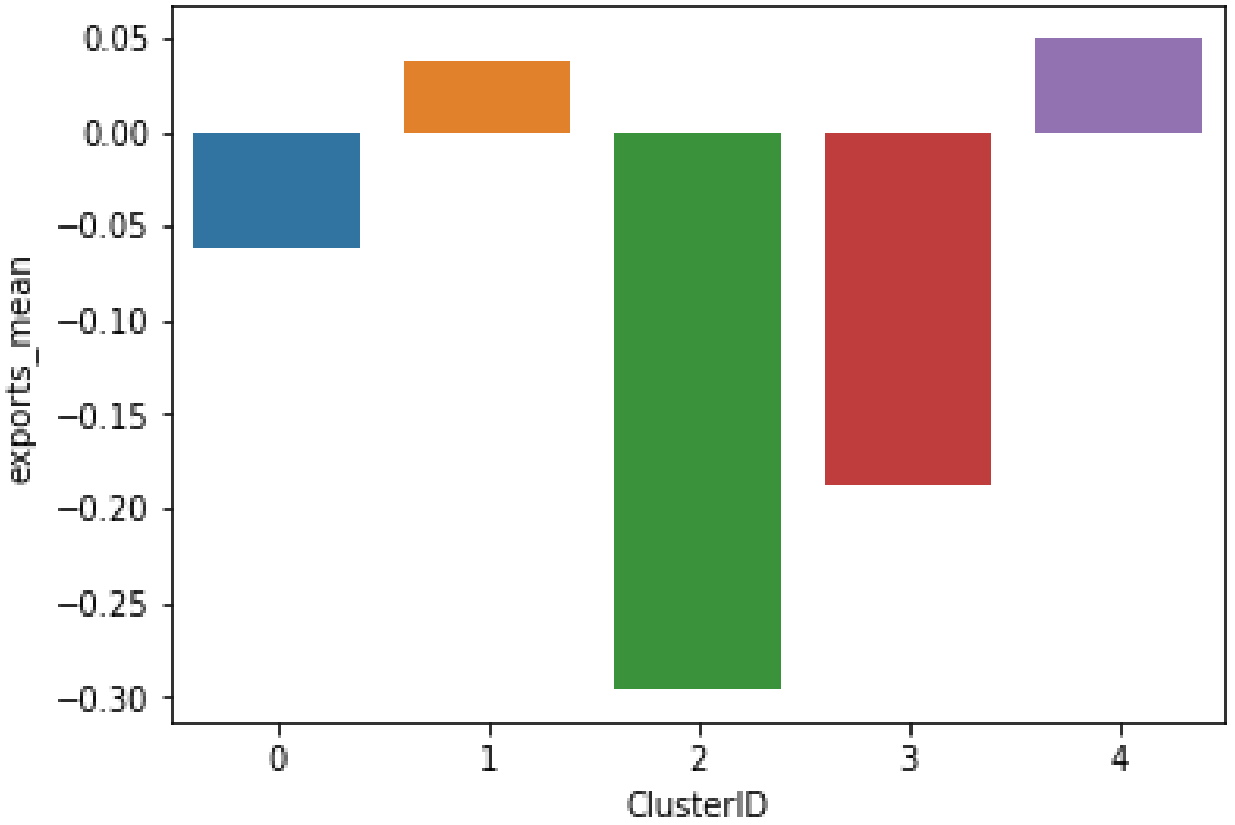

All values are less than 0.05 which mean are variables are good for analysis

## Results from both K means and hierarchical clustering

### Economic factor: exports




K means Clustering

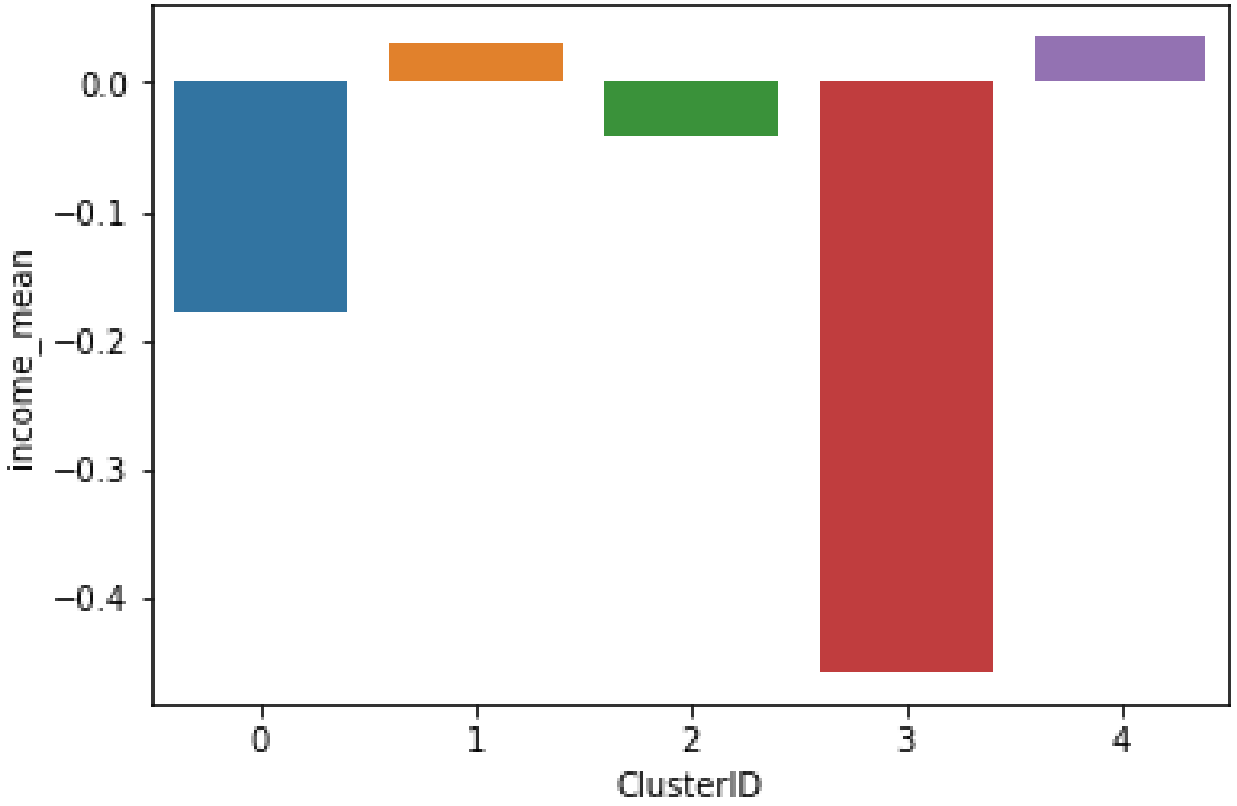



Hierarchical clustering

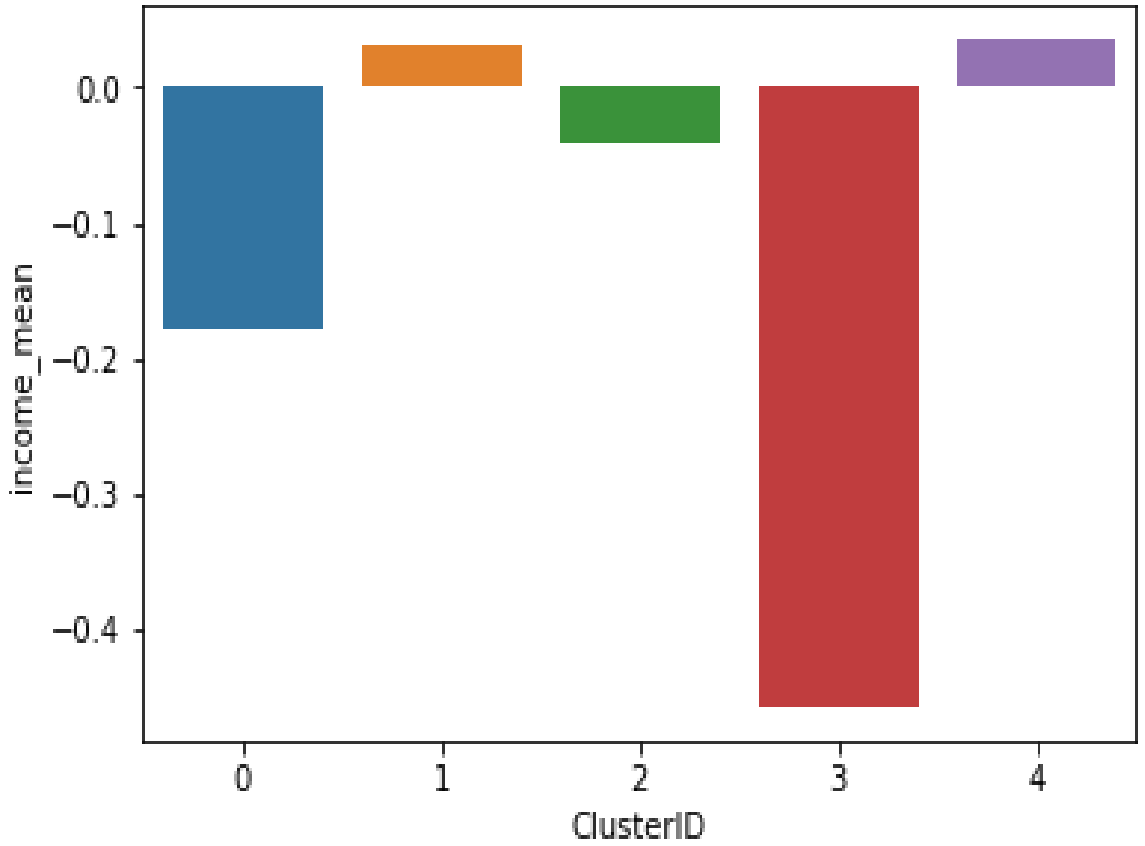
Cluster ID 2 and 3 need most of the help for exports economic factor

## Results from both K means and hierarchical clustering

### Economic factor: Income




K means Clustering

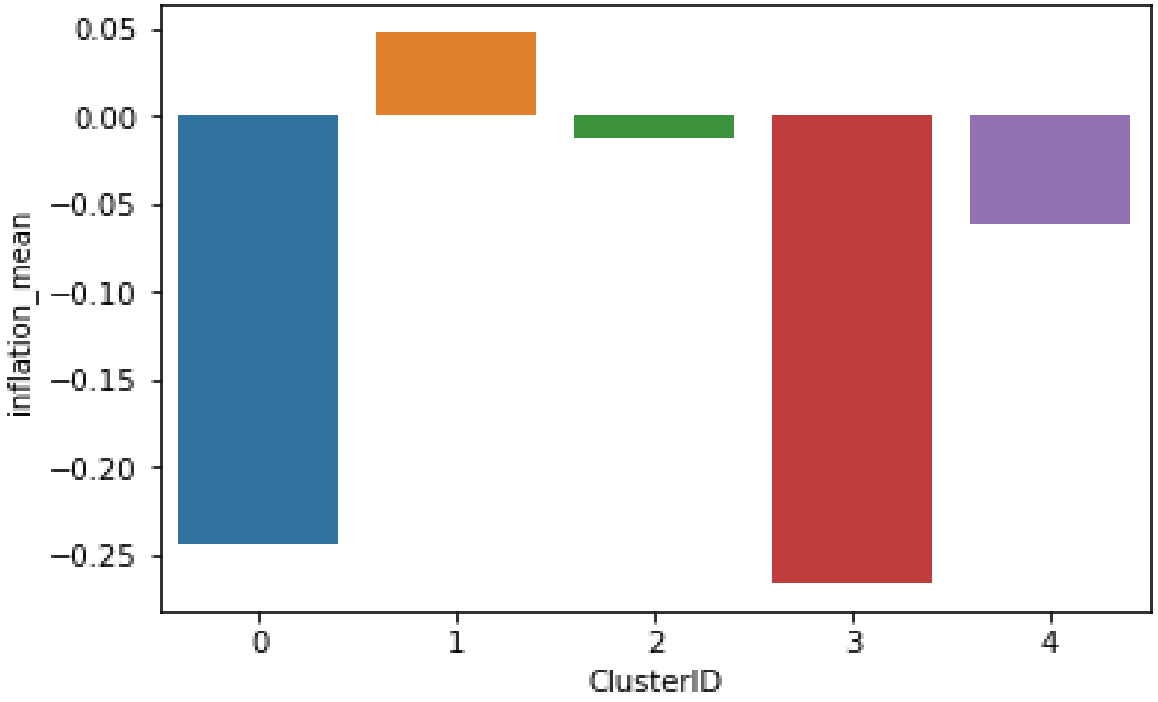



Hierarchical clustering

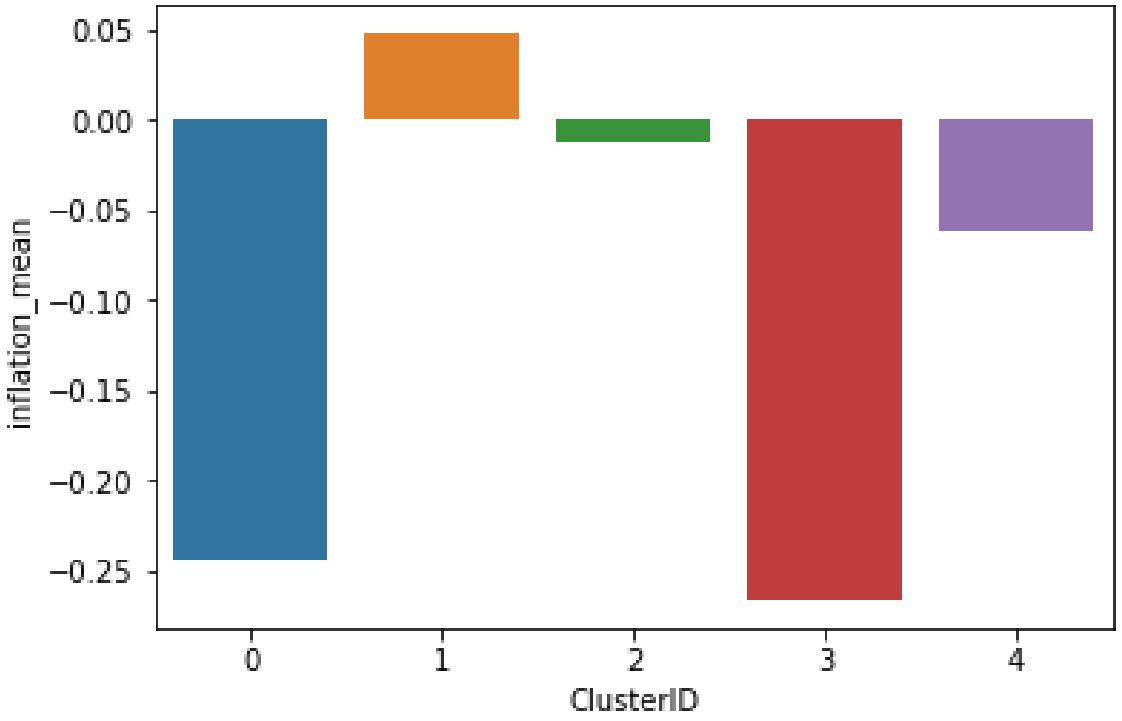
Cluster ID 3 need most of the help for Income economic factor

## Results from both K means and hierarchical clustering

### Economic factor: Inflation




K means Clustering

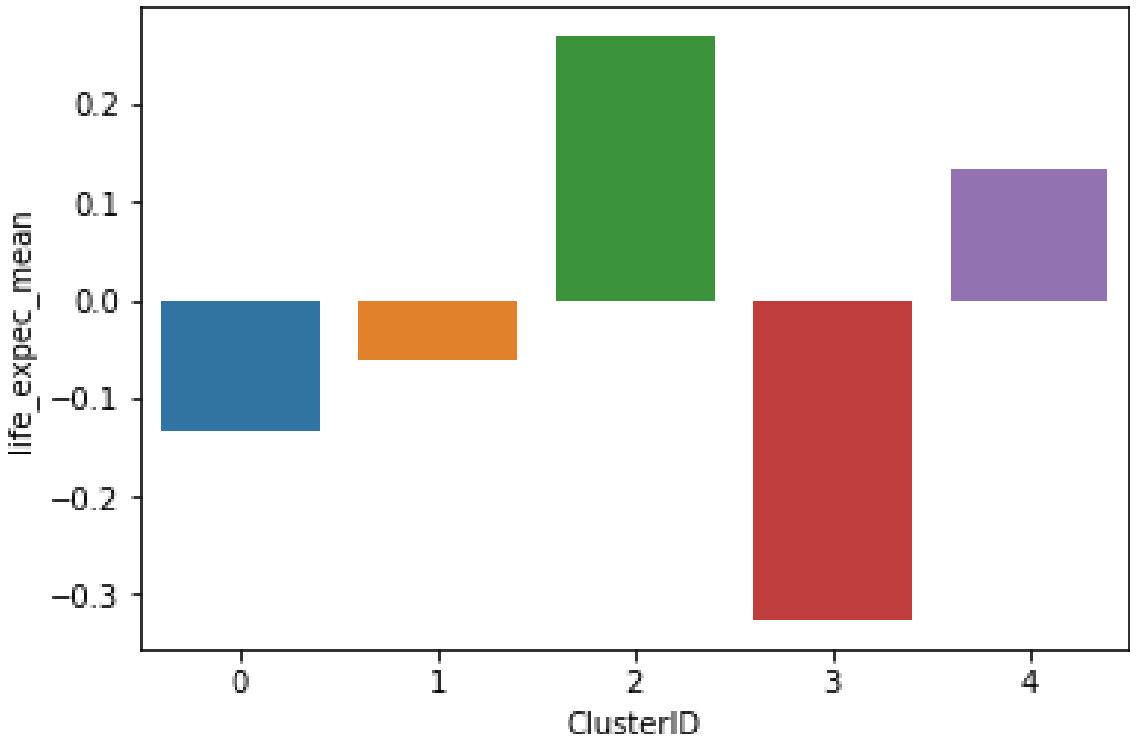



Hierarchical clustering

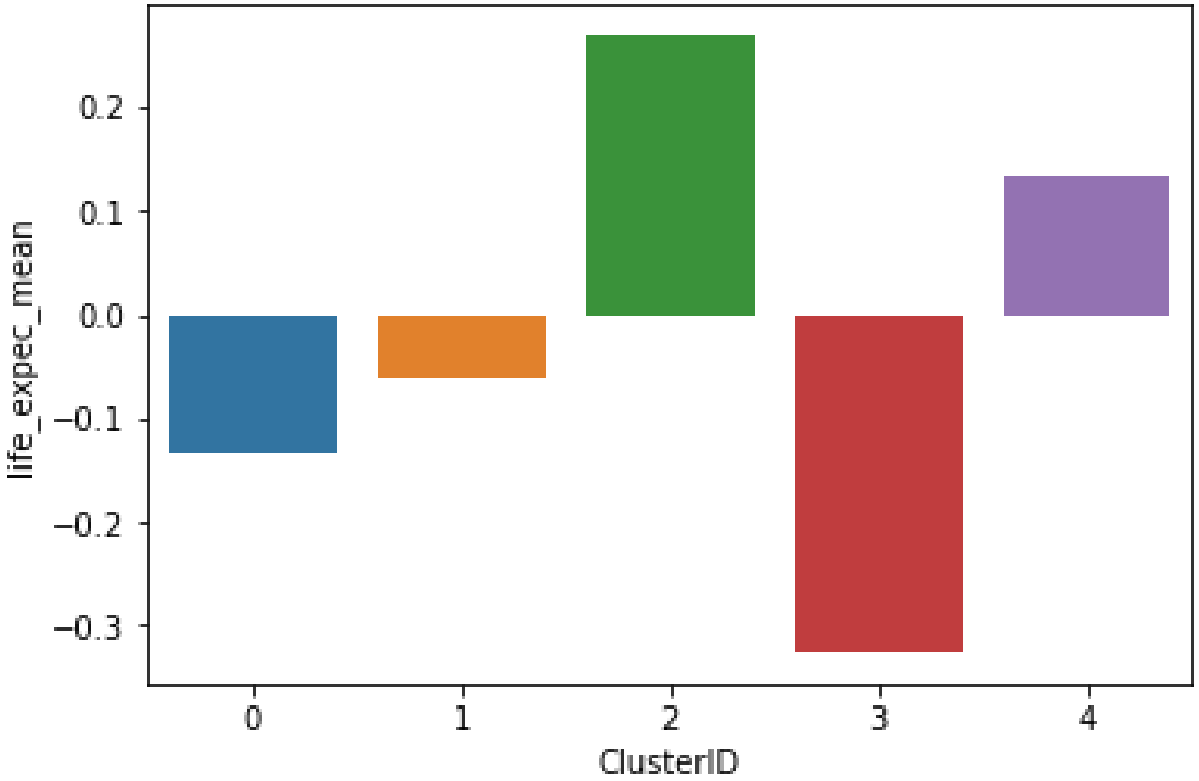
Cluster ID 0 and 3 need most of the help for Inflation economic factor

**Results from both K means and hierarchical clustering**

**Social factor: Life expectancy**




K means Clustering



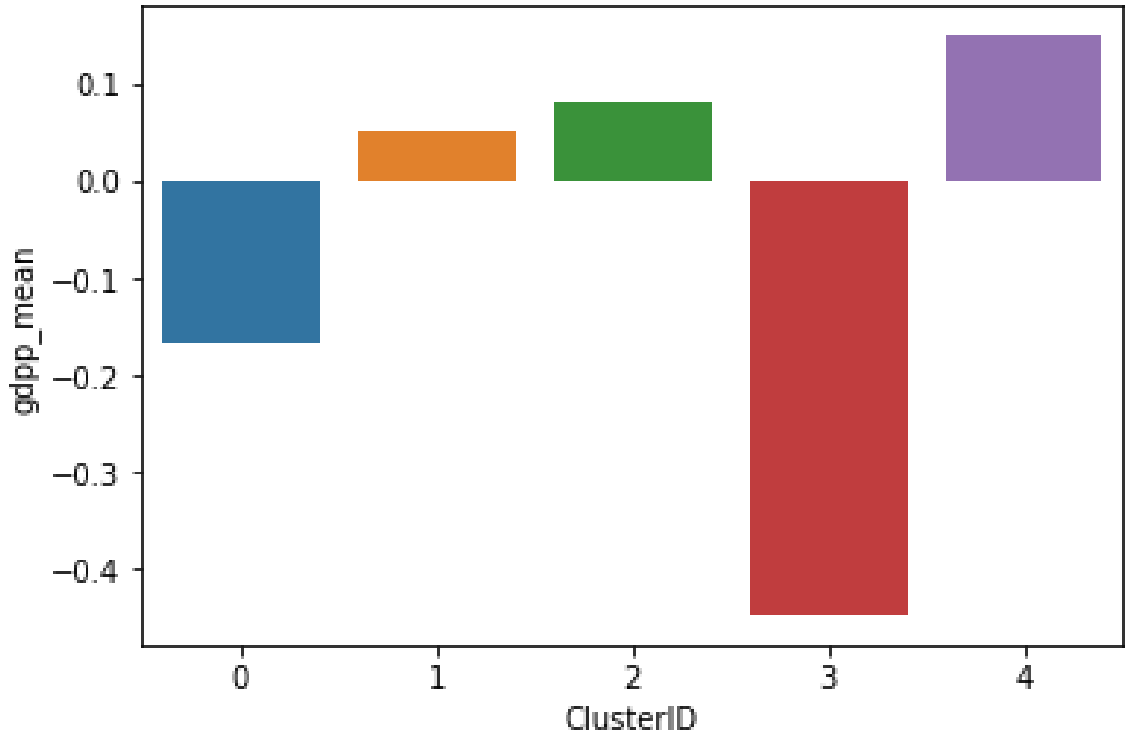

Hierarchical clustering

# Results

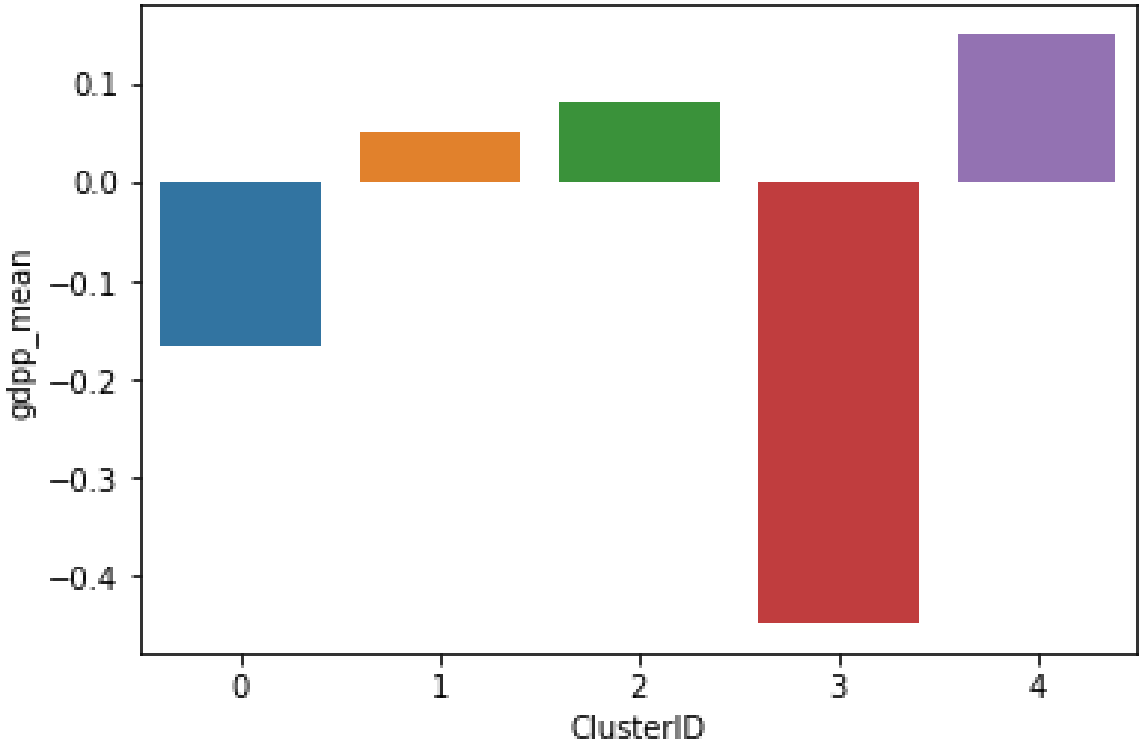
Cluster ID 3 need most of the help for Life expectancy social factor

**Results from both K means and hierarchical clustering**

**Economic factor: GDPP**




K means Clustering

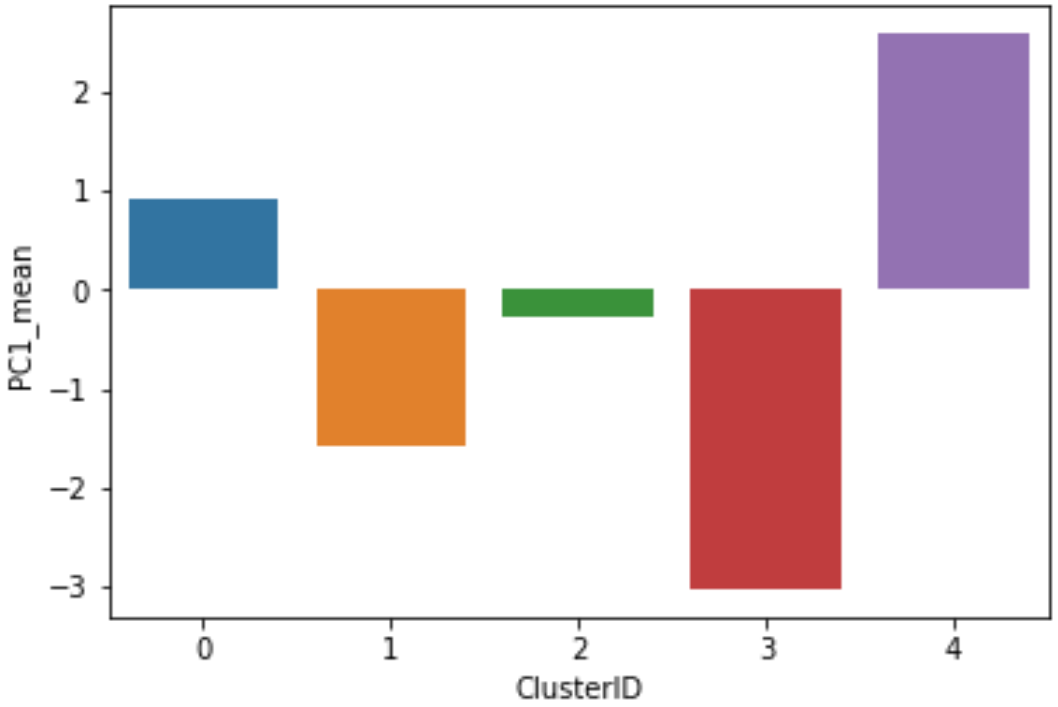



Hierarchical clustering

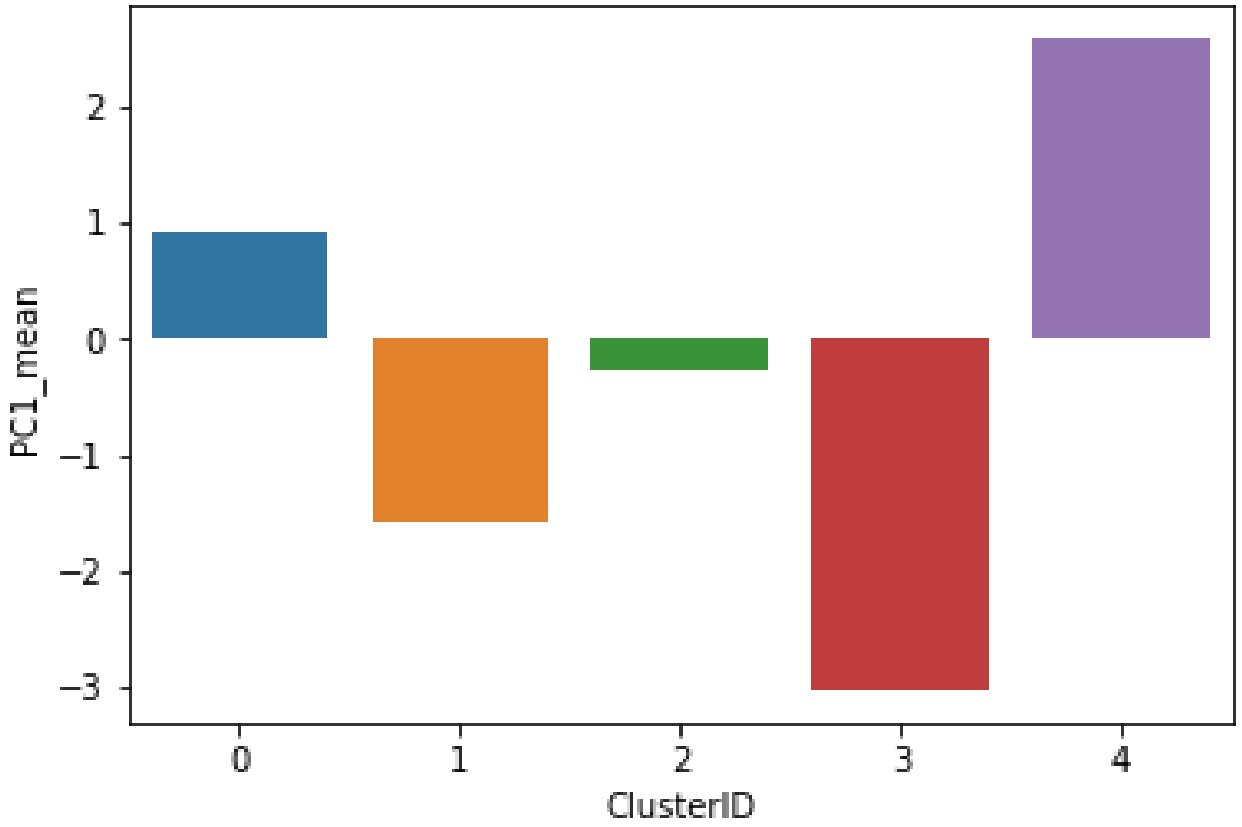
# Results

Cluster ID 3 need most of the help for GDPP economic factor

**Results from both K means and hierarchical clustering**




K means Clustering




Hierarchical clustering

# Results

Cluster ID 3 need most of the help

Some of the Countries which are in need of aid most:

1) Afghanistan
2) Benin
3) Bulgeria
4) Burkina Faso
5) Cambodia
6) Cape Verde
7) Central African Republic
8) Comoros
9) Costa Rice
10) Guatemala
11) Guinea
12) Guyana
13) Montenegro
14) Peru
15) Sierra Leone
16) Slovak Republic
17) Slovenia
18) Thailand

# Results

Disclaimer: Results published here at the time of code run and will change every time code will be run.