# PCA AND CLUSTERING ASSIGNMENT

## HELP - international humanitarian NGO

Yash Gupta

# OBJECTIVES

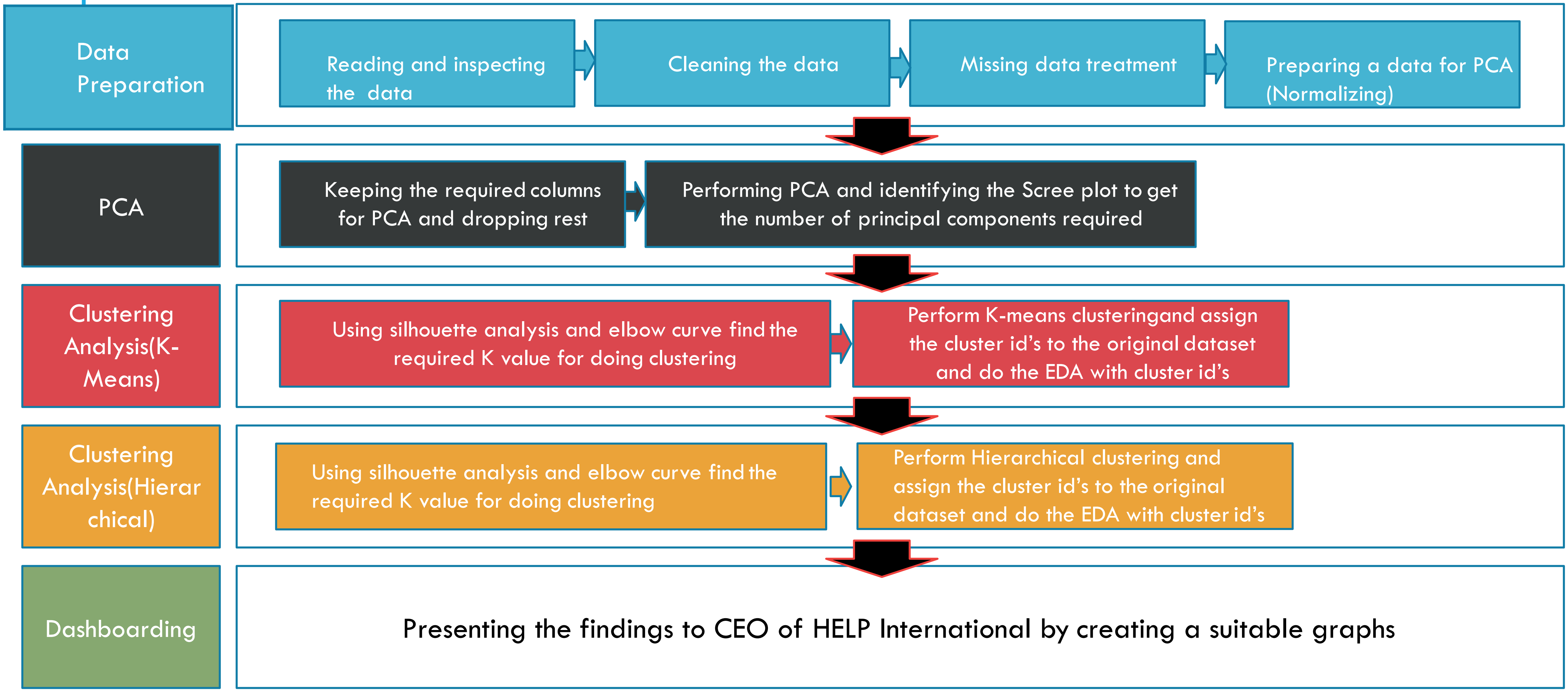
## Problem Statement

The objective of this analysis is to understand the socio-economic factors of the given countries and provide the necessary aid to the countries who are in the direst need.

This was achieved by dividing our analysis into following sub tasks,

- PCA Analysis: Finding out the optimal number of principal components required for explaining the most variance in our given data.
- Clustering analysis: Clustering the countries based on the principal components obtained from the PCA and identifying the required cluster for the helping.
- To suggest the countries which the CEO needs to focus on the most to use this money strategically and effectively.

# ANALYSIS APPROACH



Data Preparation

Reading and inspecting the data

Cleaning the data

Missing data treatment

Preparing a data for PCA (Normalizing)

PCA

Keeping the required columns for PCA and dropping rest

Performing PCA and identifying the Scree plot to get the number of principal components required

Clustering Analysis(K-Means)

Using silhouette analysis and elbow curve find the required K value for doing clustering

Perform K-means clusteringand assign the cluster id's to the original dataset and do the EDA with cluster id's

Clustering Analysis(Hierarchical)

Using silhouette analysis and elbow curve find the required K value for doing clustering

Perform Hierarchical clustering and assign the cluster id's to the original dataset and do the EDA with cluster id's

Dashboarding

Presenting the findings to CEO of HELP International by creating a suitable graphs

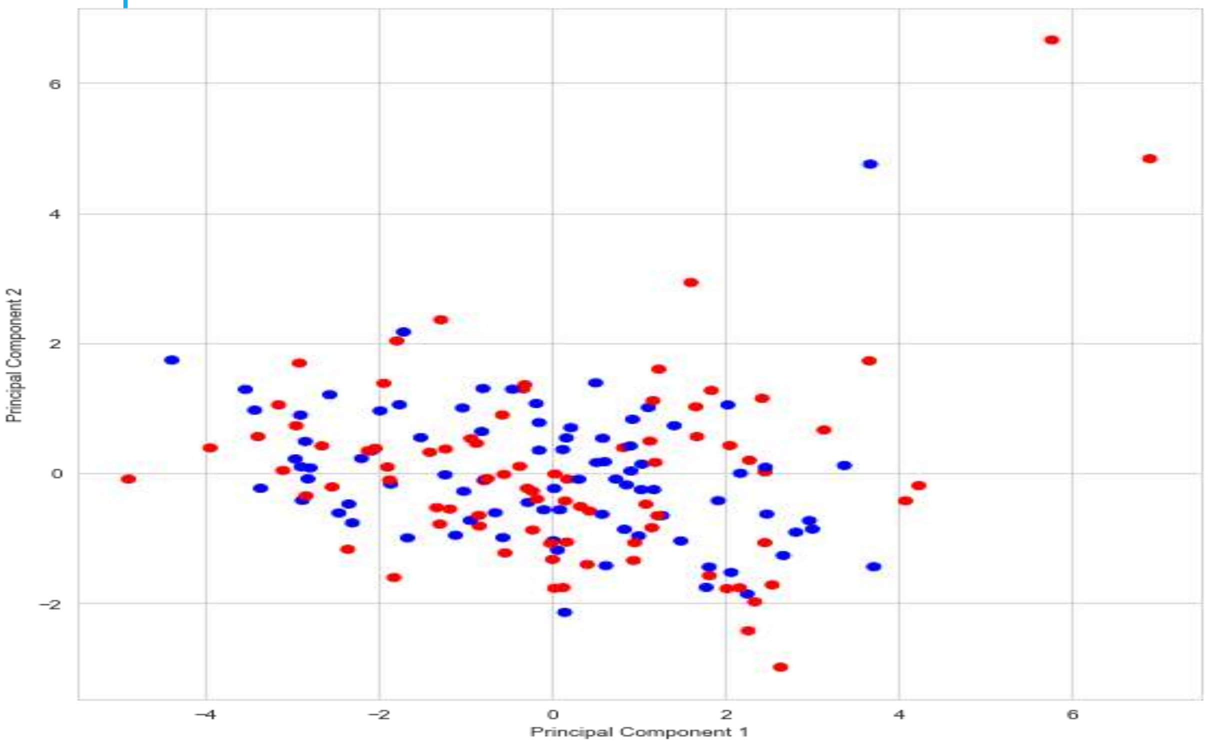
# DATA CLEANING

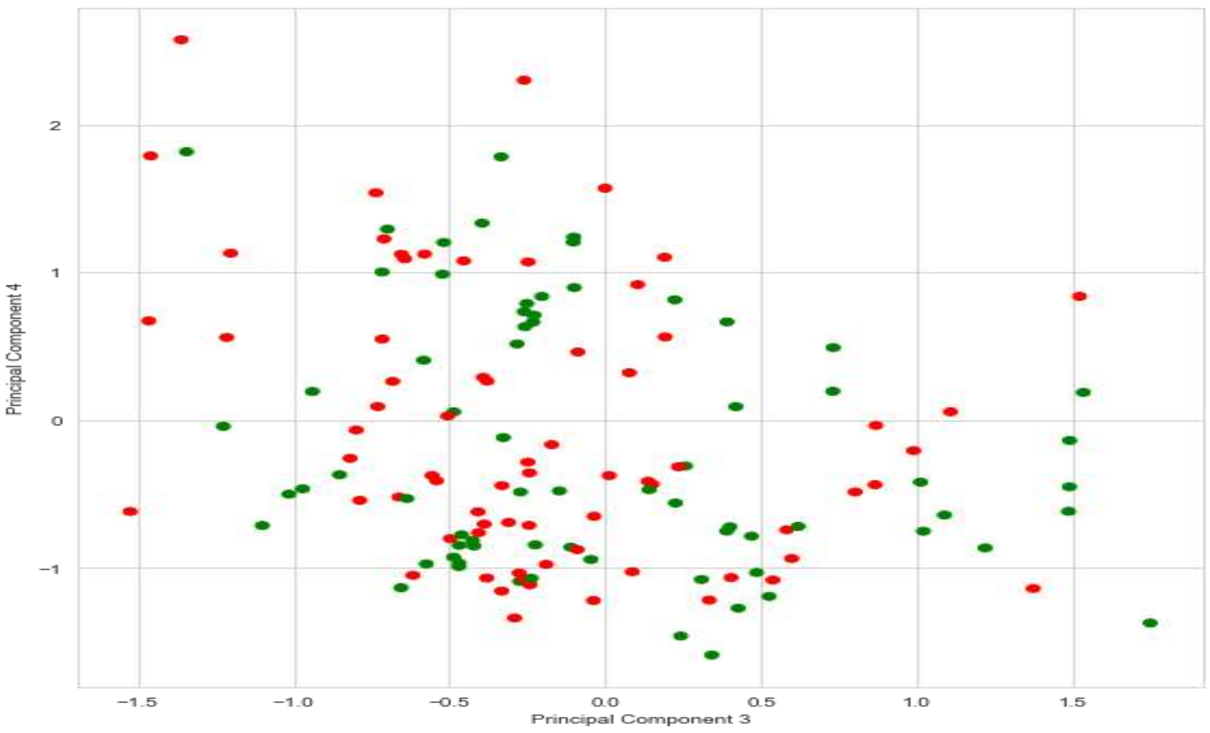
I did preliminary analysis on data to understand and check its feasibility to be used for  the said purpose..

**All the graphs are made from hierarchical clustering

** **Results for K mean clustering with k=5 (clusters) is same as result obtained from Hierarchical clustering which are discussed in details in subsequent slides.**

# RESULT of PCA

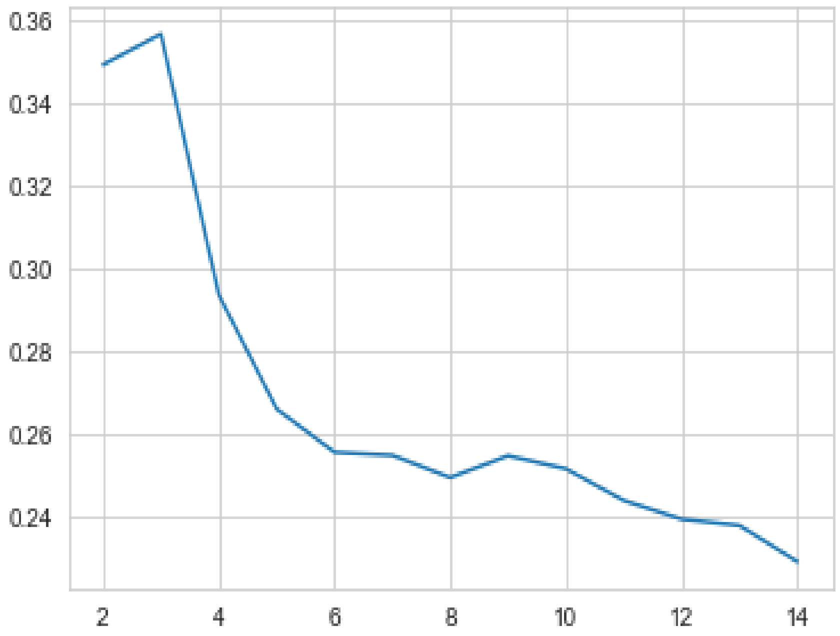


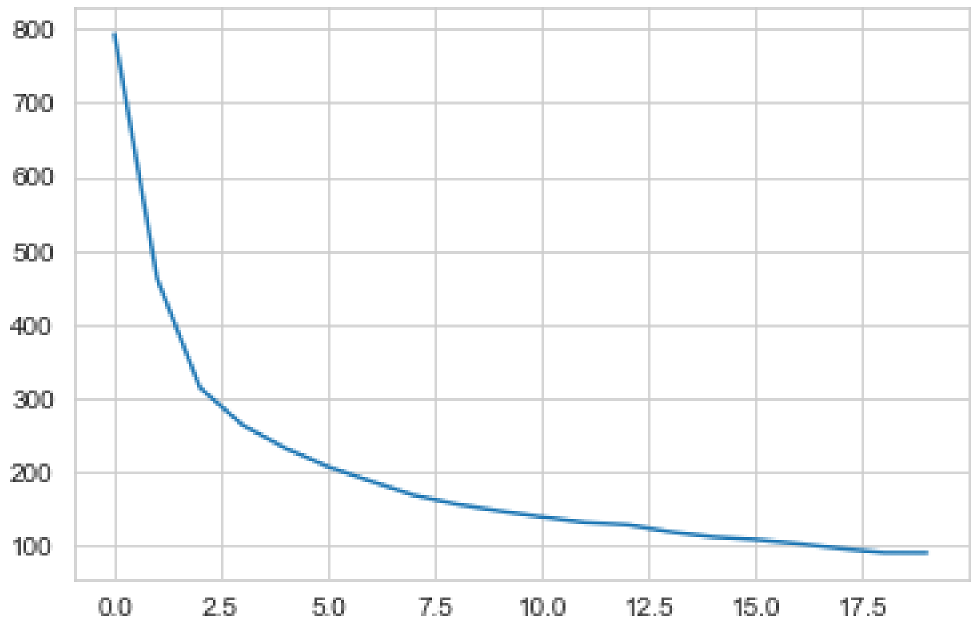


- PCA is done for Dimensionality Reduction .
- Around 90% variance is explained by 5 principal components.
- Max correlation is 0.007 , min correlation is -0.002 so correlations are indeed very close to 0

# RESULT of K mean Clustering for k = 5

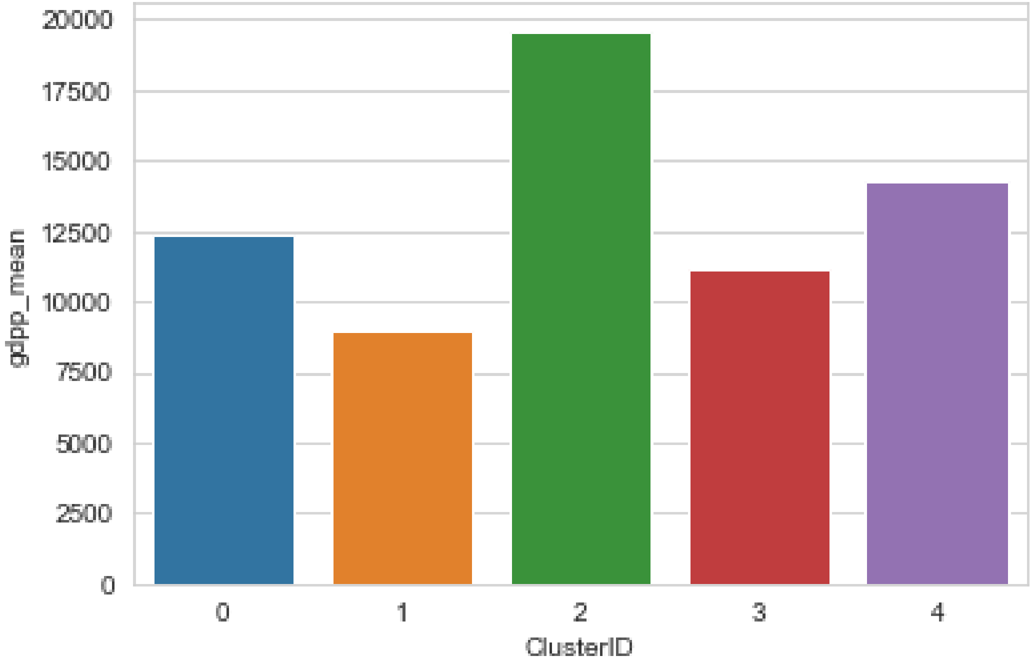
Hopkins Value is 0.76 so dataset has high tendency to cluster.



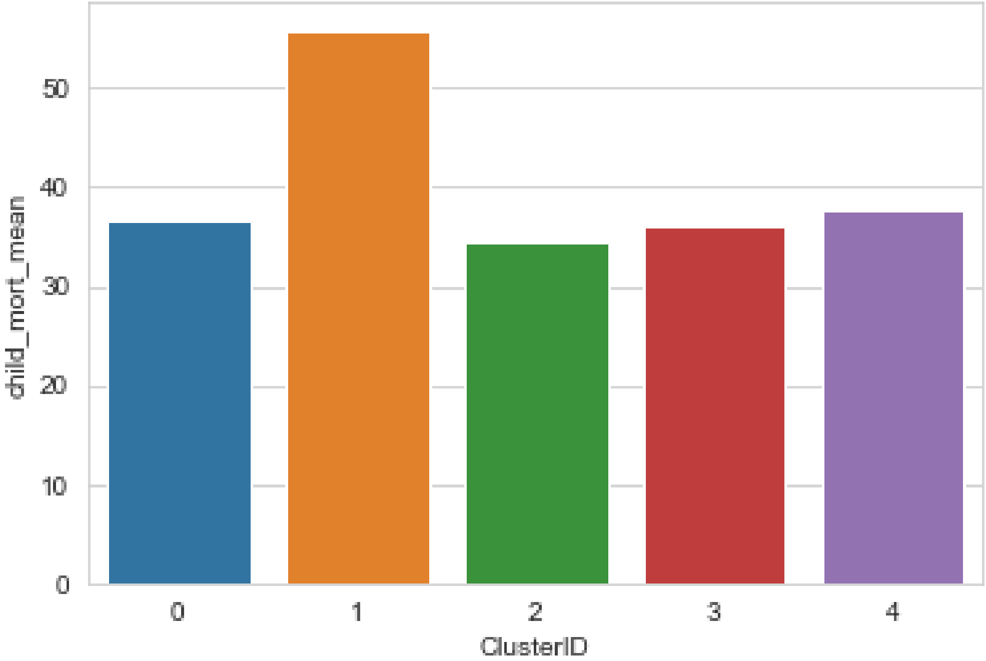


- Silhouette Analysis – form above graph  k =  5 seems to be good value for forming clusters.

# RESULTS (1/3)

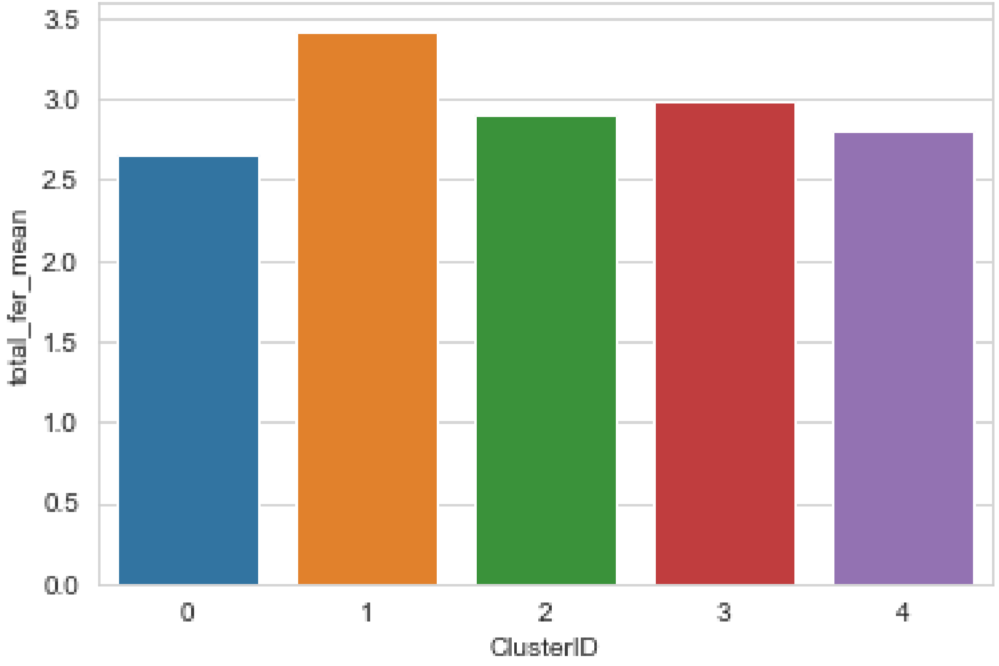


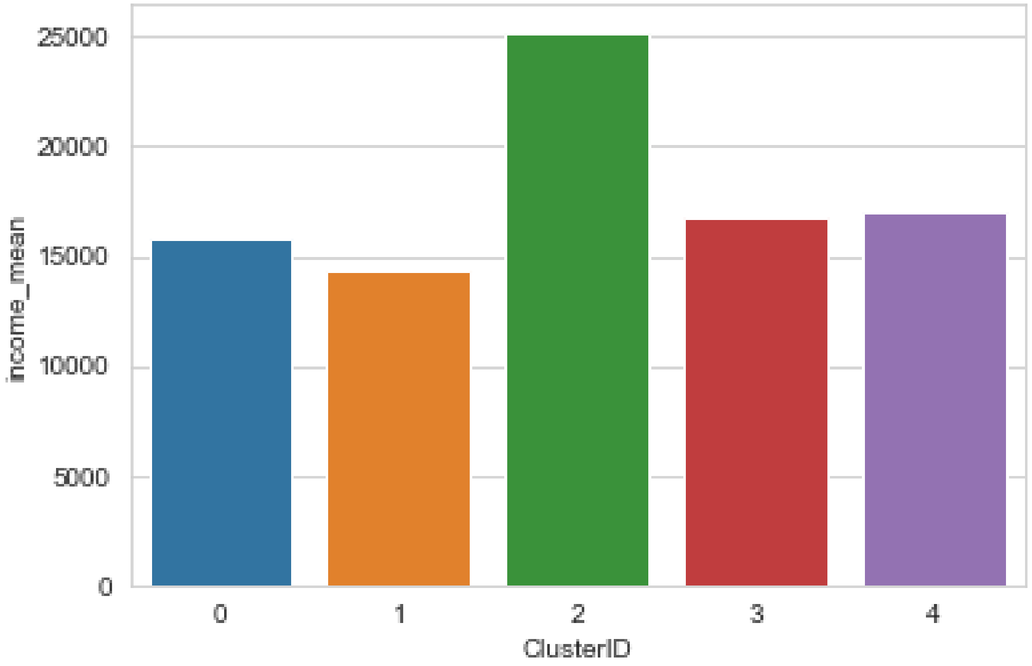
•Countries in the cluster 1 has the lowest GDP compared to other countries



•Countries in the cluster 1 has the highest Death of children under 5 years of age per 1000 live births compared to other countries

# RESULTS(2/3)

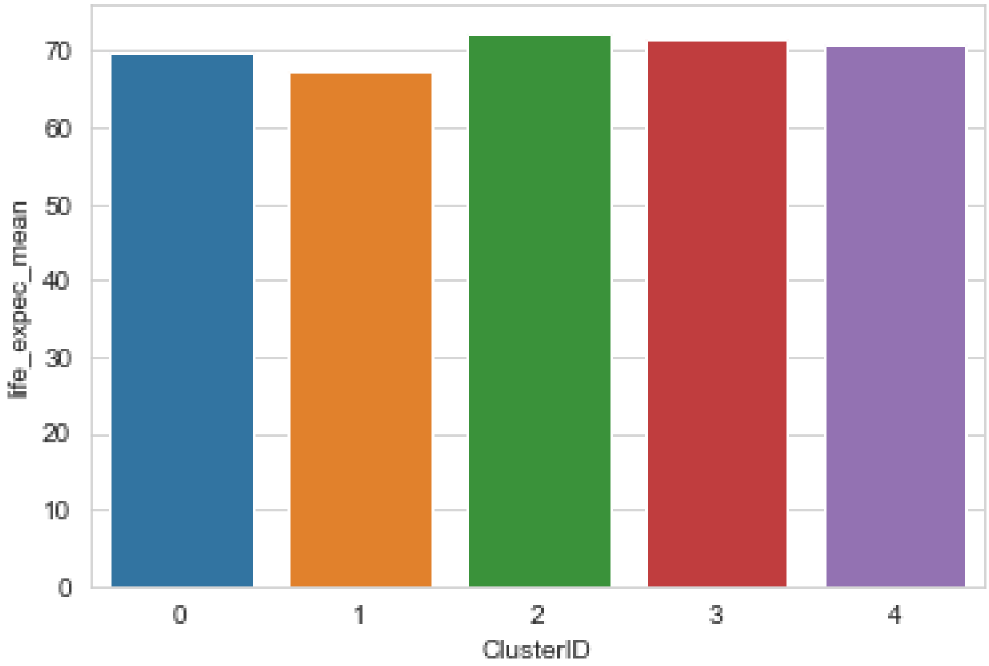




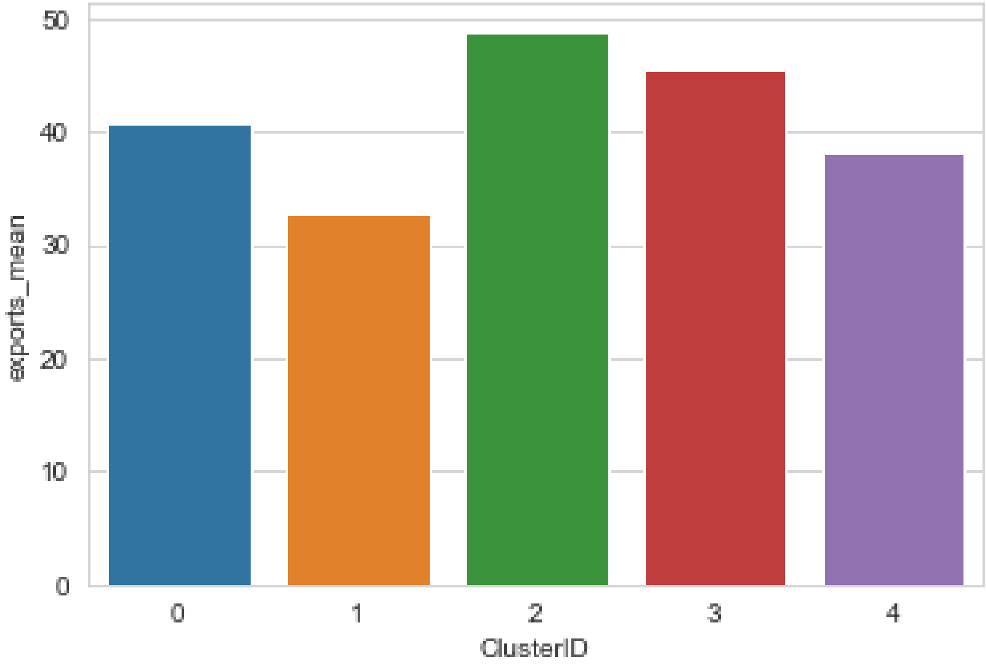
- Countries in the cluster 1 has the highest total fertility rate compared to other countries

- Countries in the cluster 1 has the lowest income compared to other countries

# RESULTS(3/3)



•Countries in the cluster 1 has the lowest life expectancy (age) compared to other countries



•Countries in the cluster 1 has the lowest exports compared to other countries

# CONCLUSIONS

## Recommendation

Following are the recommendations,

- Countries which are in direst need are the countries in the cluster 1.
- After comparing with both the clustering techniques (K mean clustering and Hierarchical Clustering) we have obtained with the below list of countries which can be considered to utilize raised around $ 10 million funds .
- **Afghanistan , Angola , Brunei , Bulgaria , Cameroon , Chile , Comoros , Ghana , Grenada , Lao , Lesotho , Macedonia, FYR , Mali , New Zealand , Samoa, Sierra Leone ,Sri Lanka**

Many of the above countries are from Africa which needs the help.