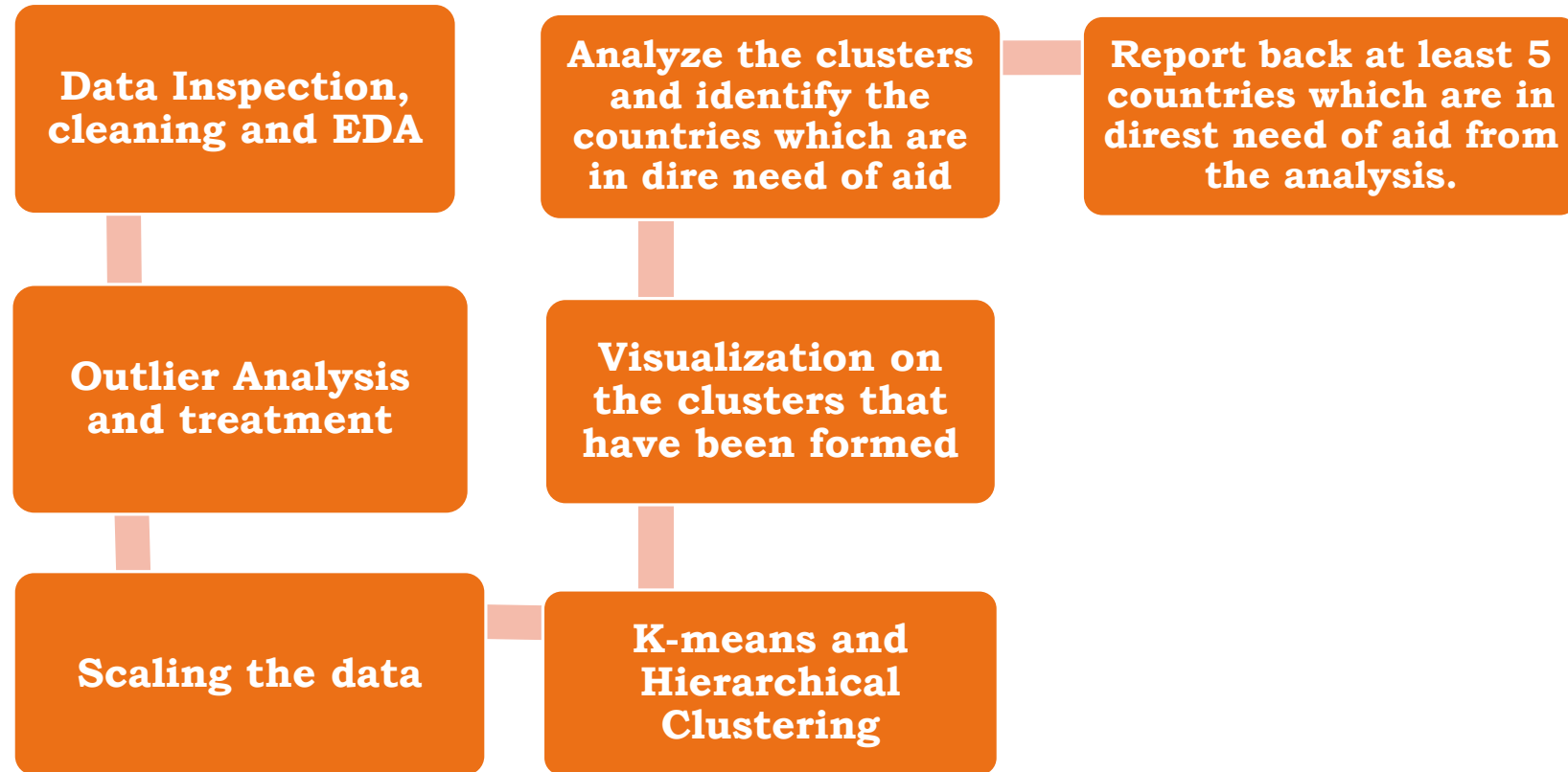# Clustering Assignment

SHRUTI SRIVASTAVA

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 programs, 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.

Analysis approach

Data Inspection, cleaning and EDA

Outlier Analysis and treatment

Scaling the data

K-means and Hierarchical Clustering

Visualization on the clusters that have been formed

Analyze the clusters and identify the countries which are in dire need of aid

Report back at least 5 countries which are in direst need of aid from the analysis.

Data Inspection and cleaning

- There are 167 countries in the dataset with 9 socio-economic and health features related to each of them.

- There are no null values in the dataset.

- No duplicates were found in the data.

- There are outliers present in the data which have been handled

- High correlation exists between some of the variables in the data set.

- The data has been standardized using scaling.

Exploratory Data Analysis

To choose the countries that are in the direst need of aid we need to check some socio-economic and health factors that determine the overall development of the country.

We analyzed the lowest 10 countries for each factor.
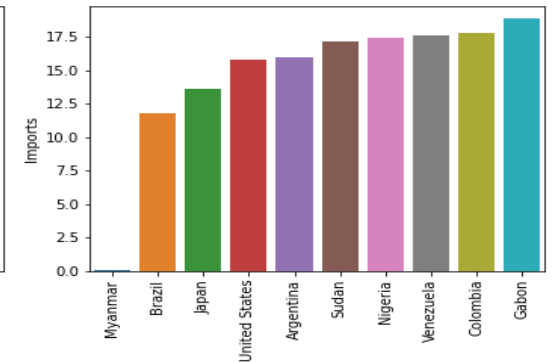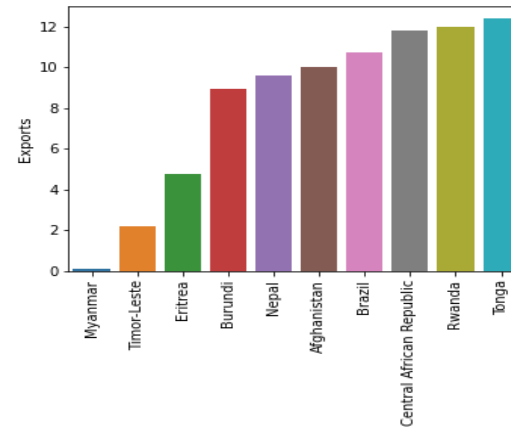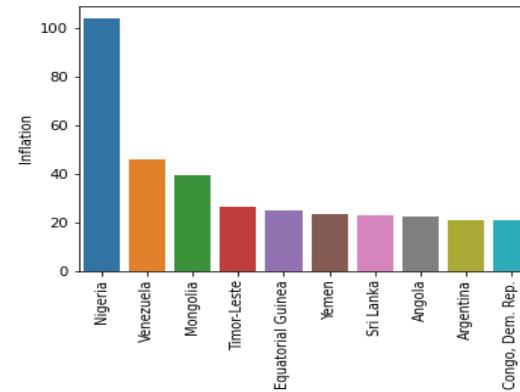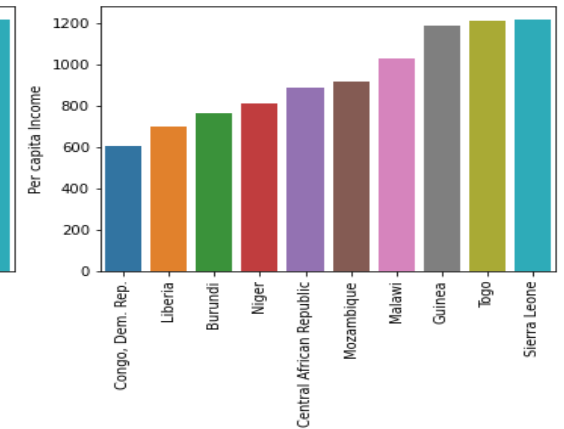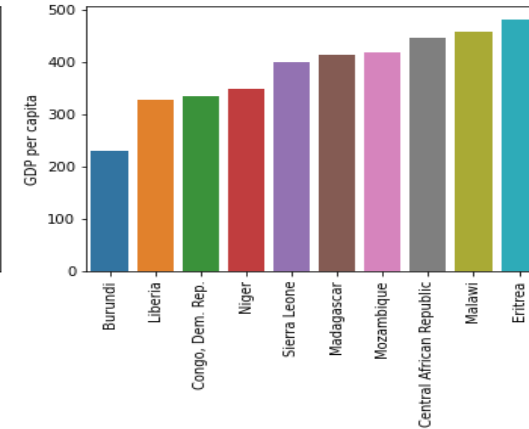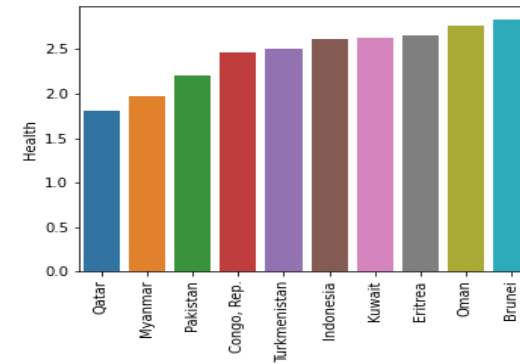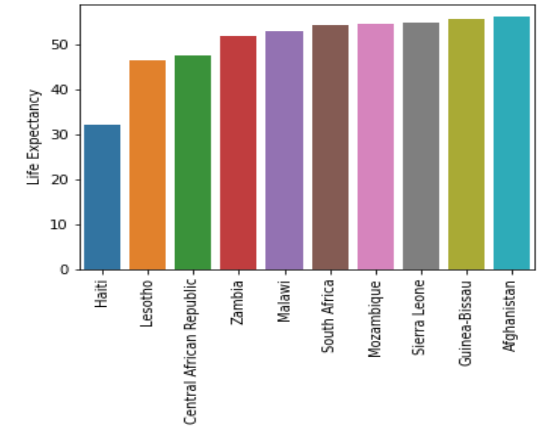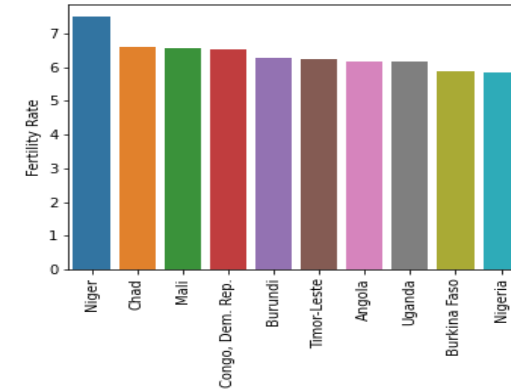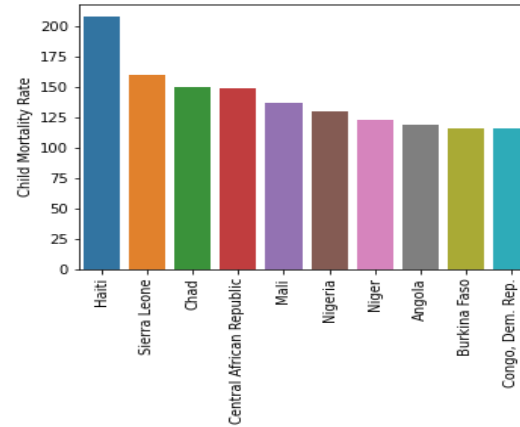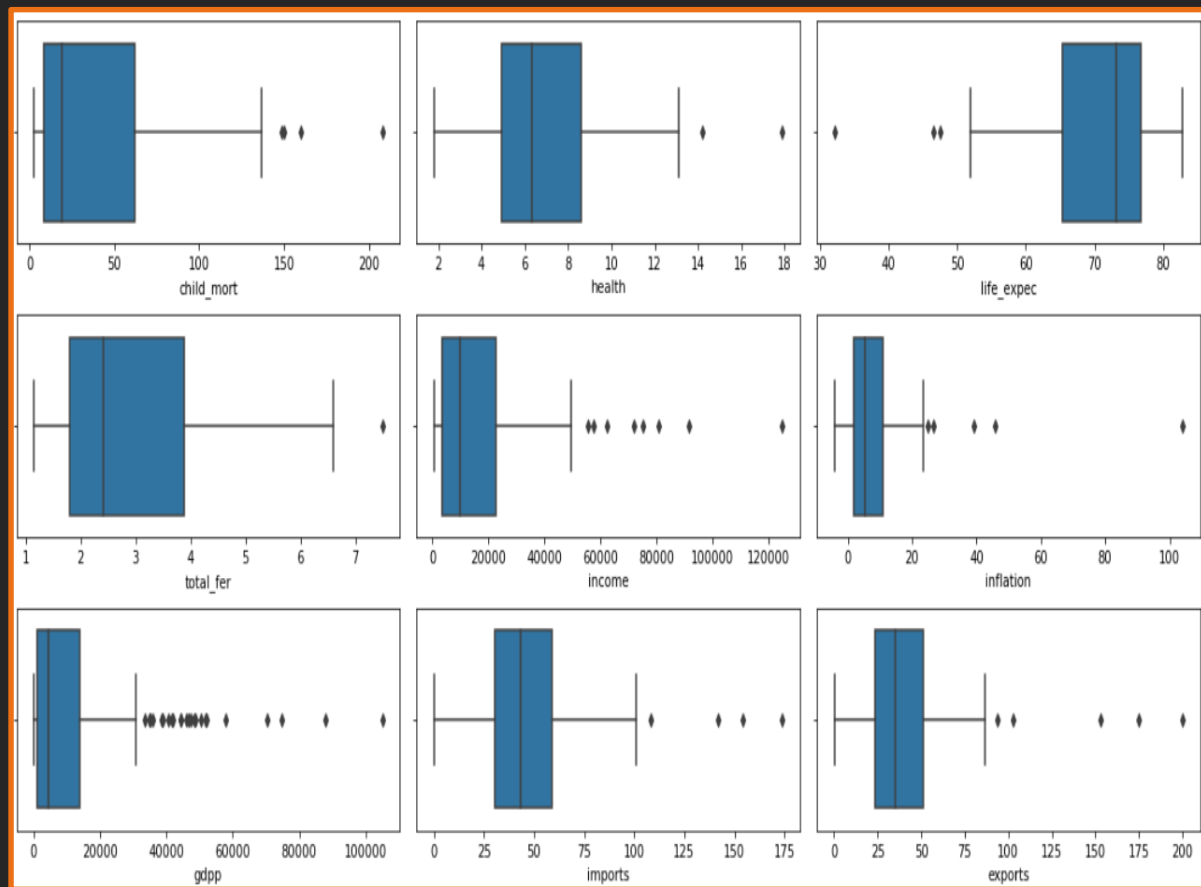
Correlation in the Data

We notice that there is high correlation between quite a few variables like total_fer and child_mort, income and gdpp etc.
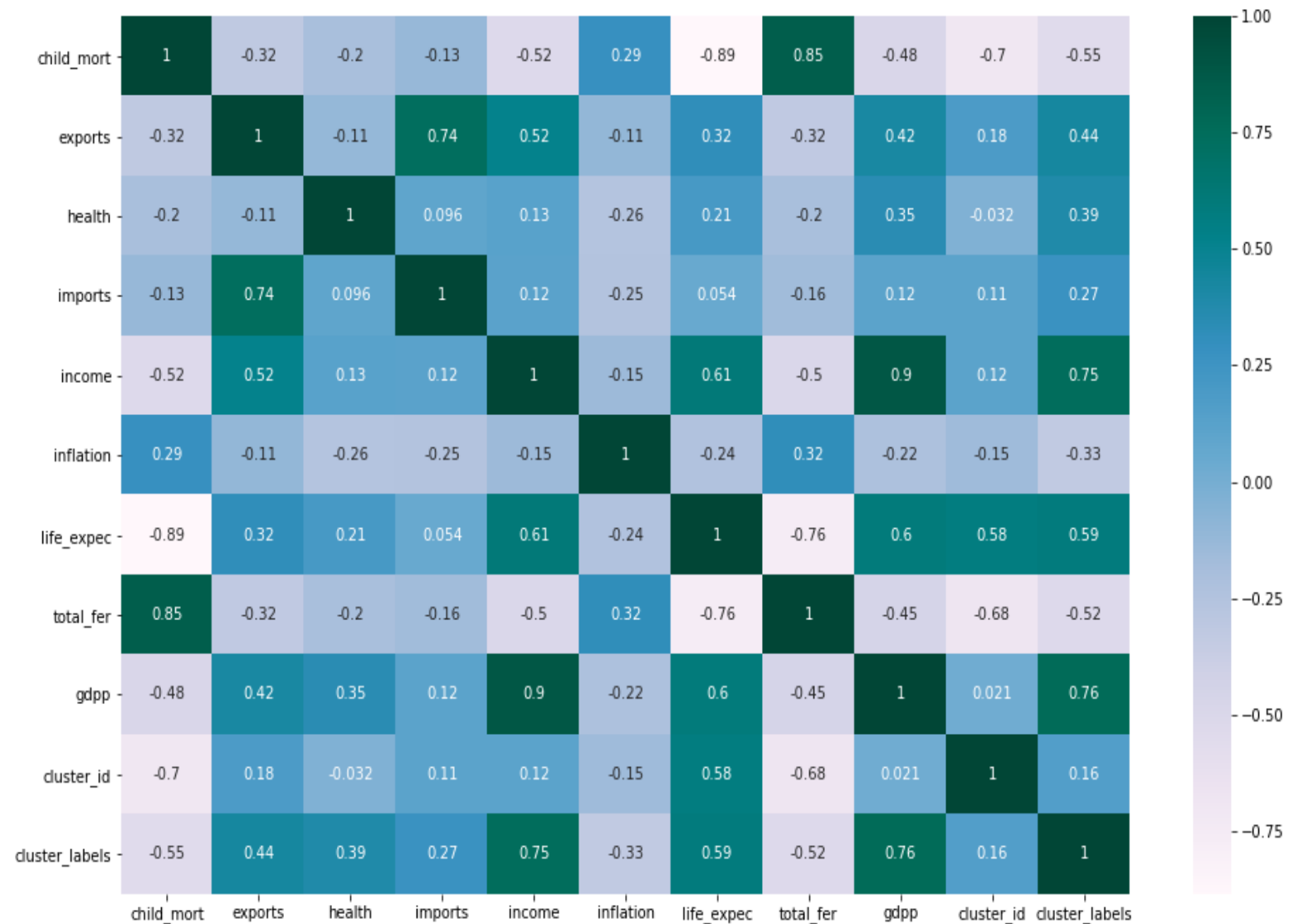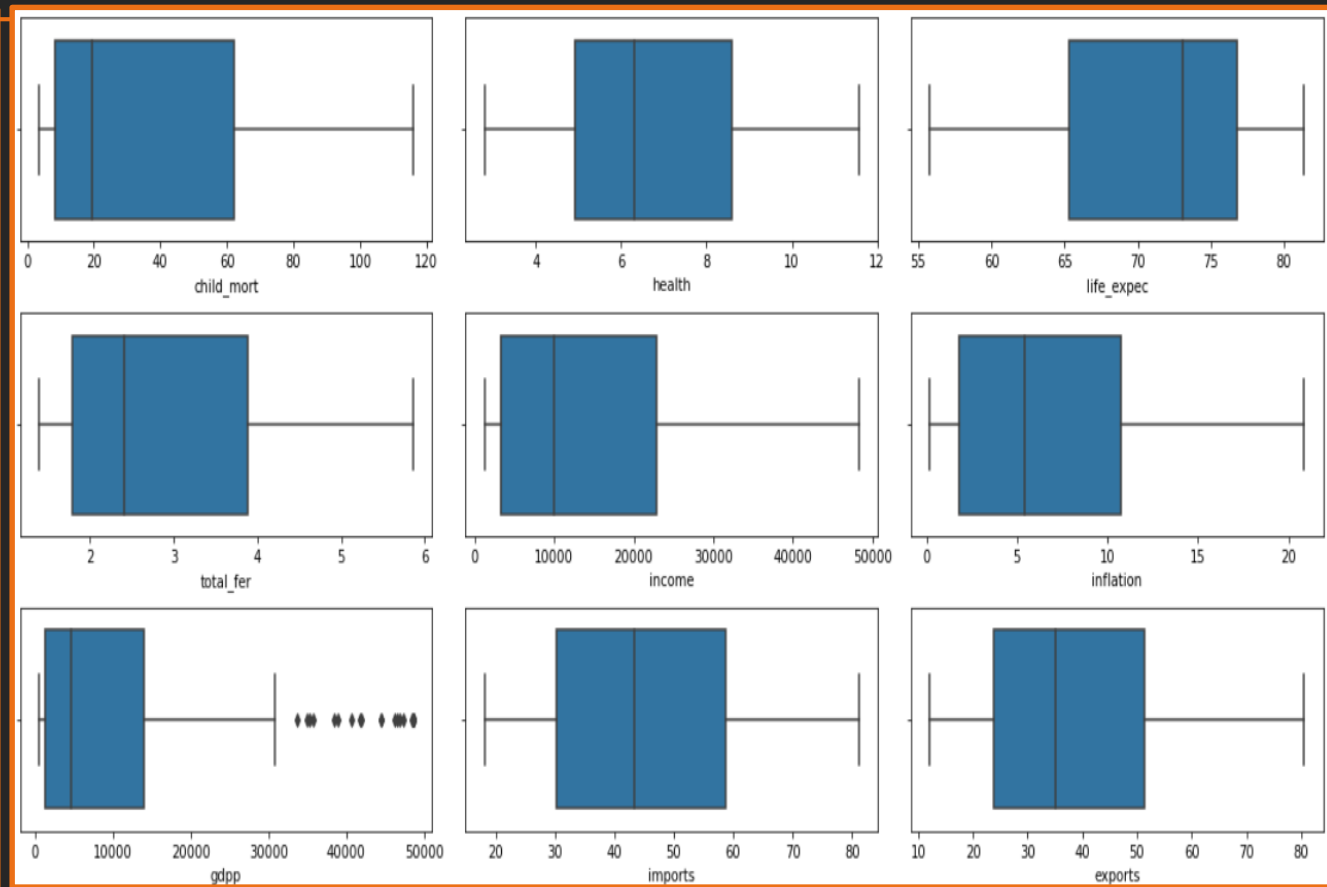
Outlier Analysis

We notice that there are quite a few outliers in the data and both k-means and Hierarchical clustering algorithms are sensitive to outliers, hence we have handled the outliers by capping them.
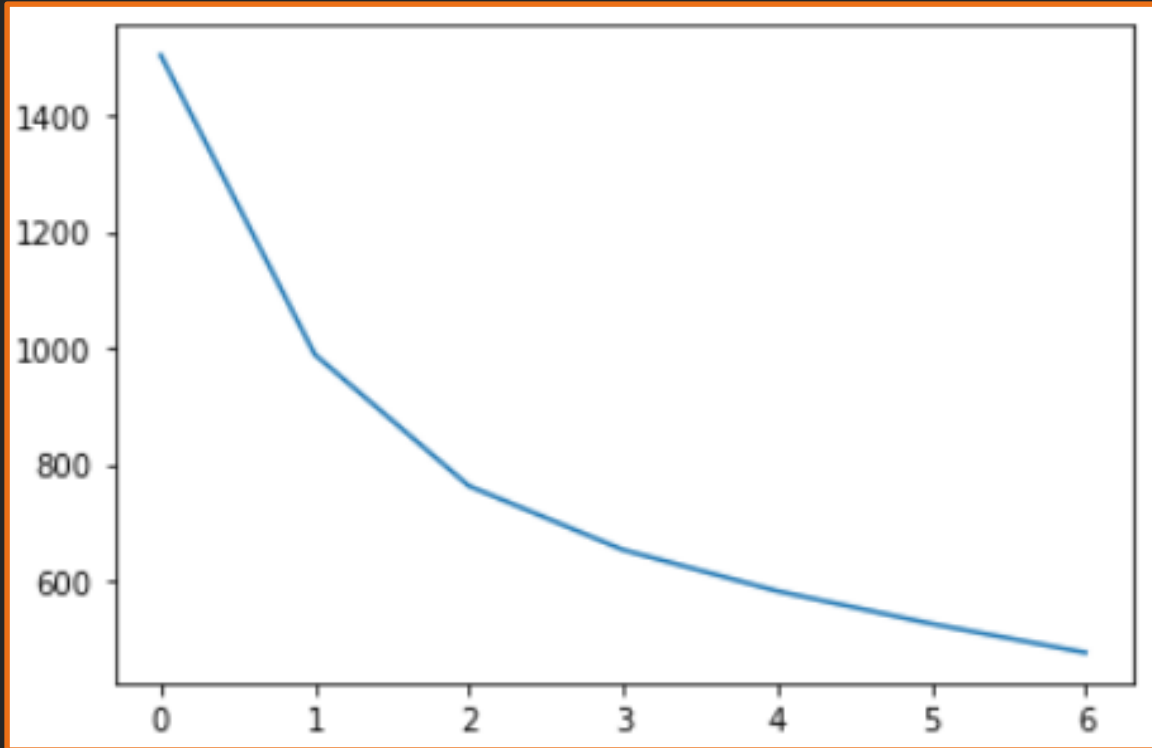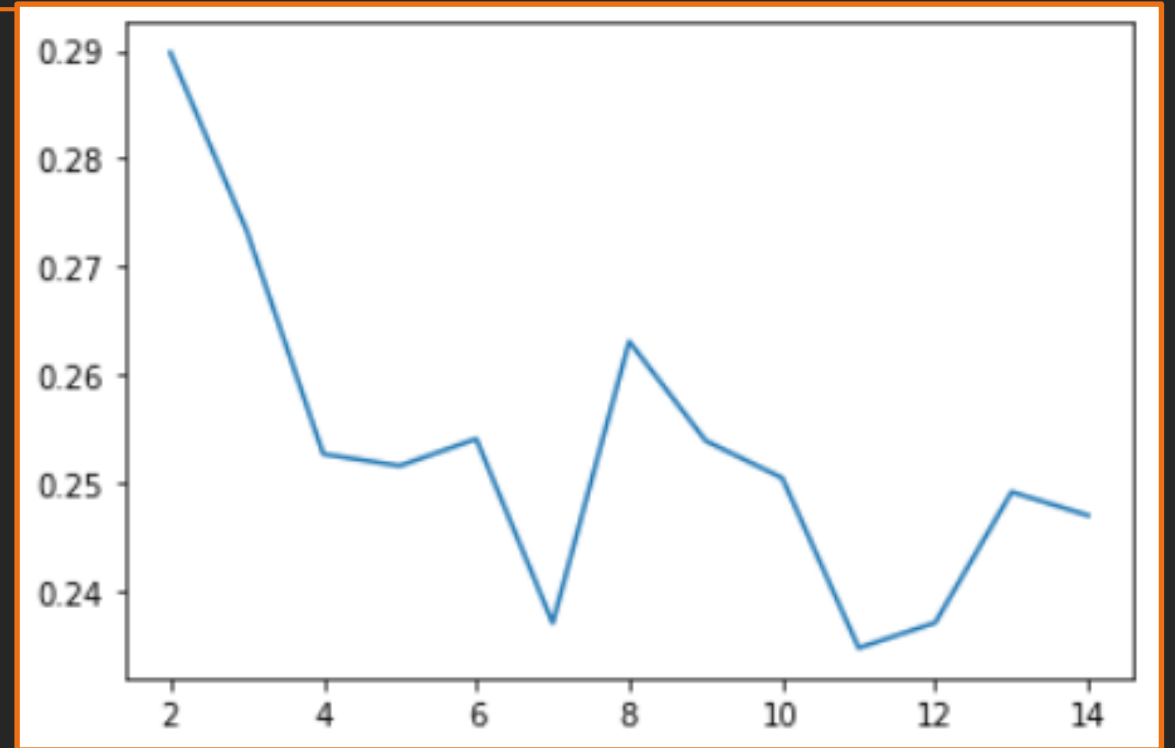
BEFORE   **AFTER**
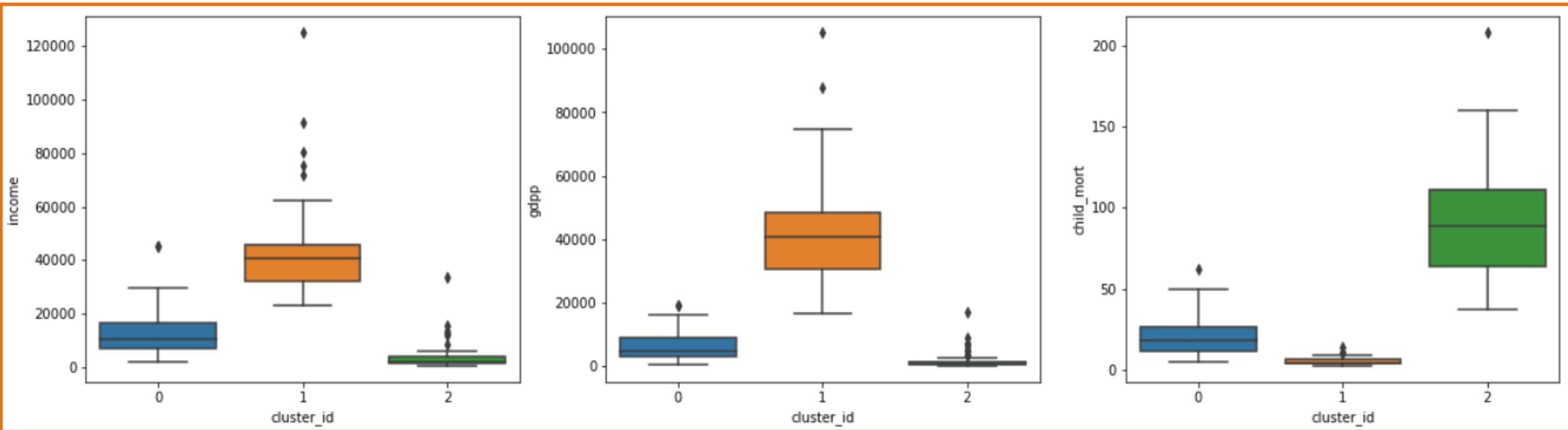
K-Means Clustering

Based on the Elbow curve and silhouette analysis we can conclude that K=3 will be the appropriate number of clusters to go with.
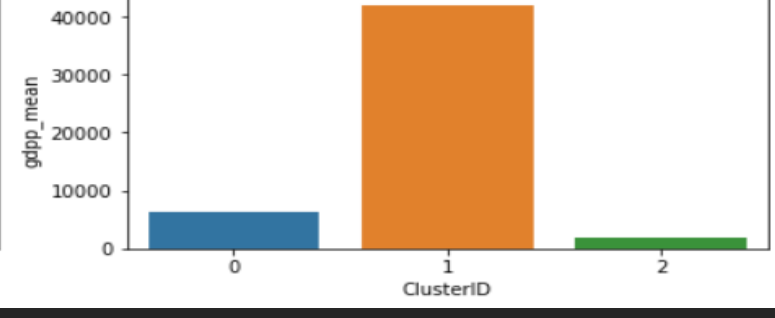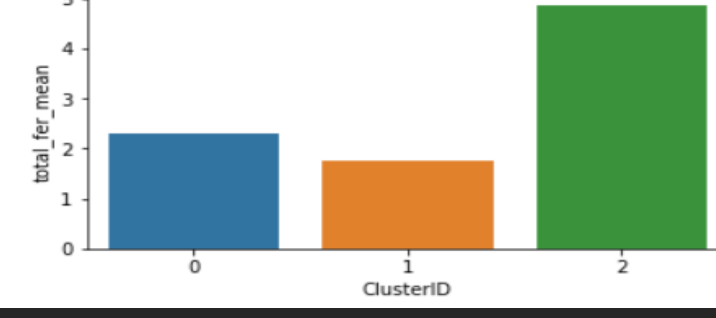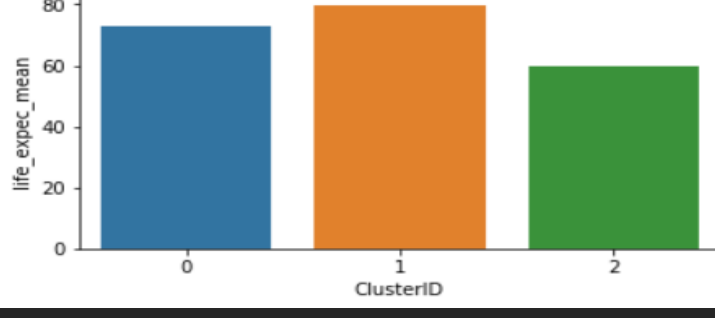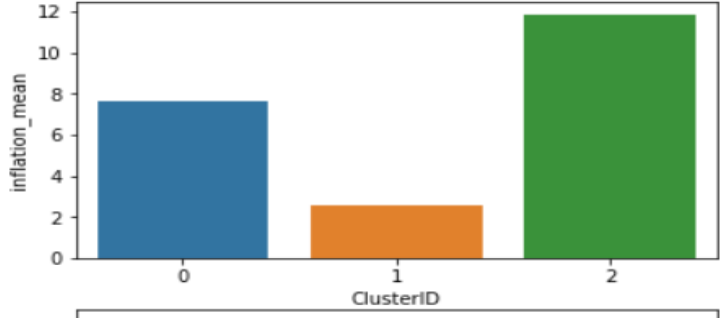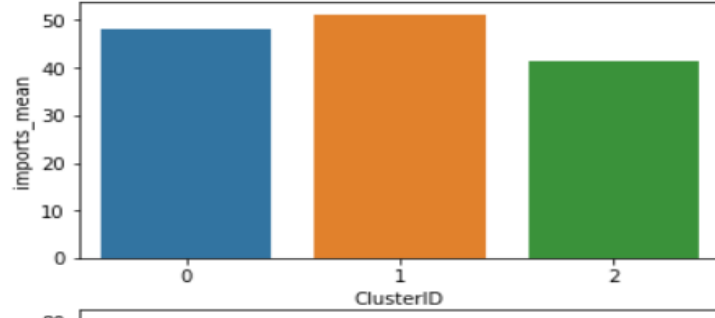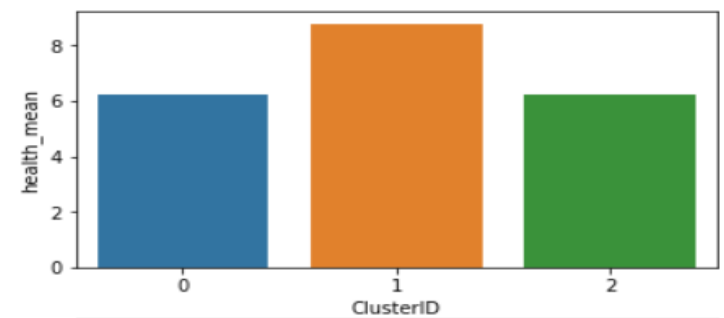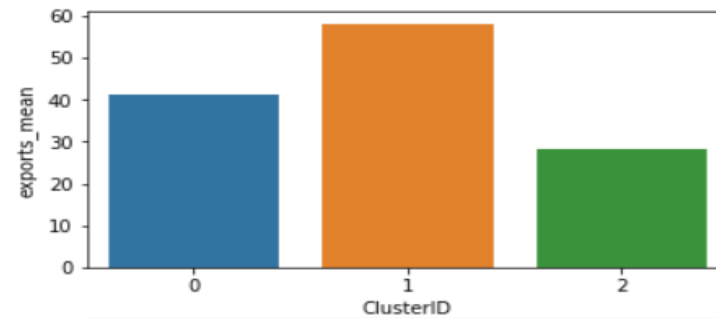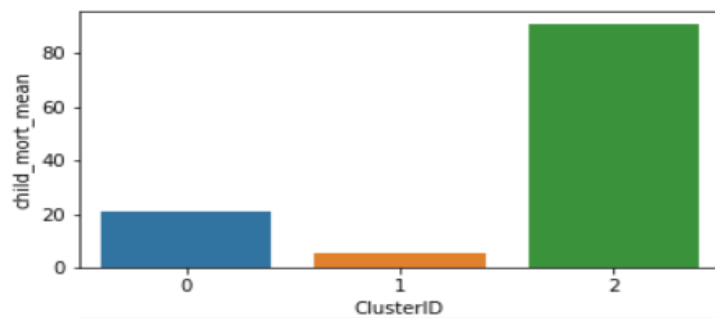
Sum of Squared Distances (SSD)   **Silhouette Analysis**

Cluster Analysis (Kmeans)

-Above three boxplots indicate that **Cluster-2** has :

- High child mortality
- Low income
- Low gdpp
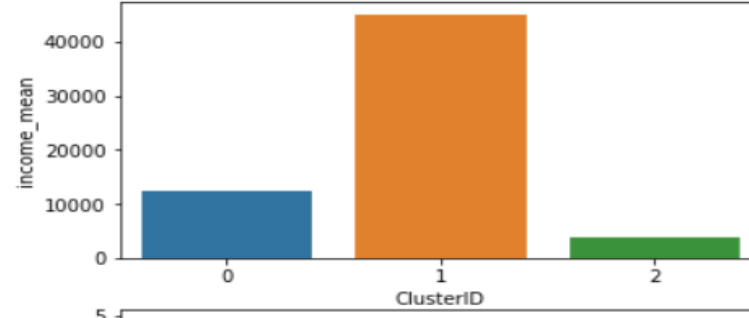
Cluster Analysis (Kmeans)

From the graphs we see **Cluster – 2** has:

- It has highest child mortality
- Lowest income
- Highest Inflation
- Comparatively low life expectancy
- Highest total fertility
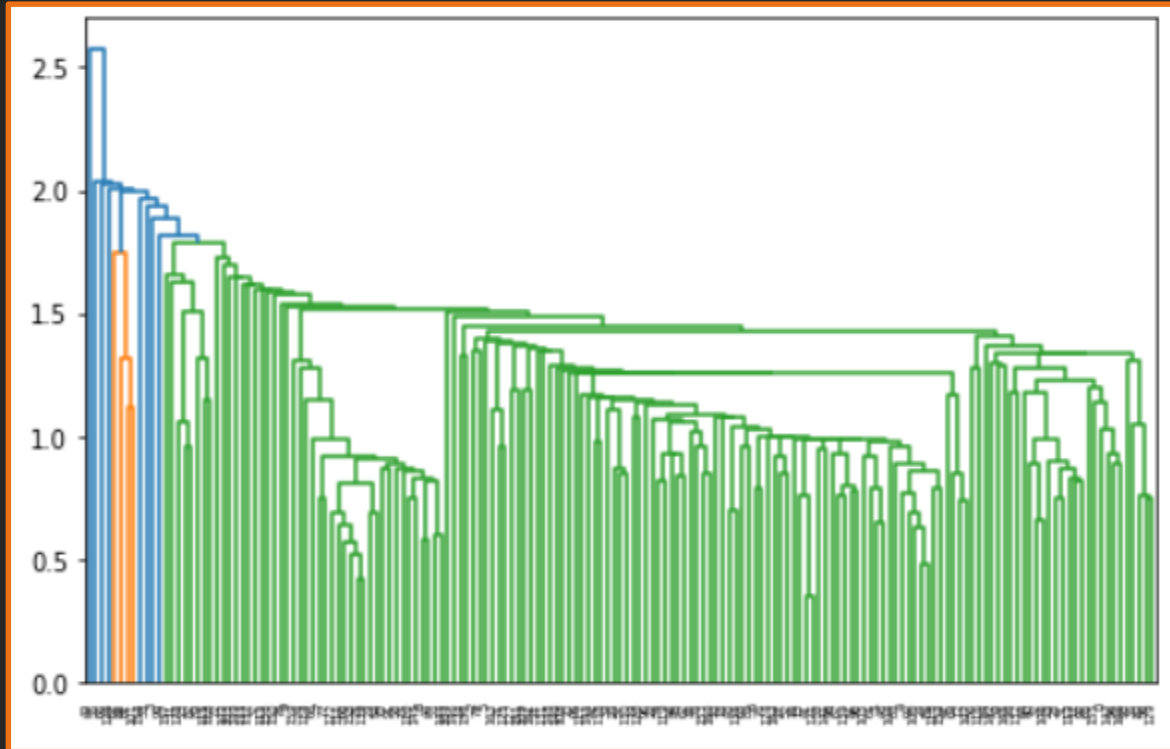- Lowest gdpp.

Countries
(K-Means Clustering)

As per K- means clustering, the country which are direst need of aid are:

- Burundi
- Liberia
- Congo, Dem. Rep
- Niger
- Sierra Leone
- Madagascar
- Mozambique
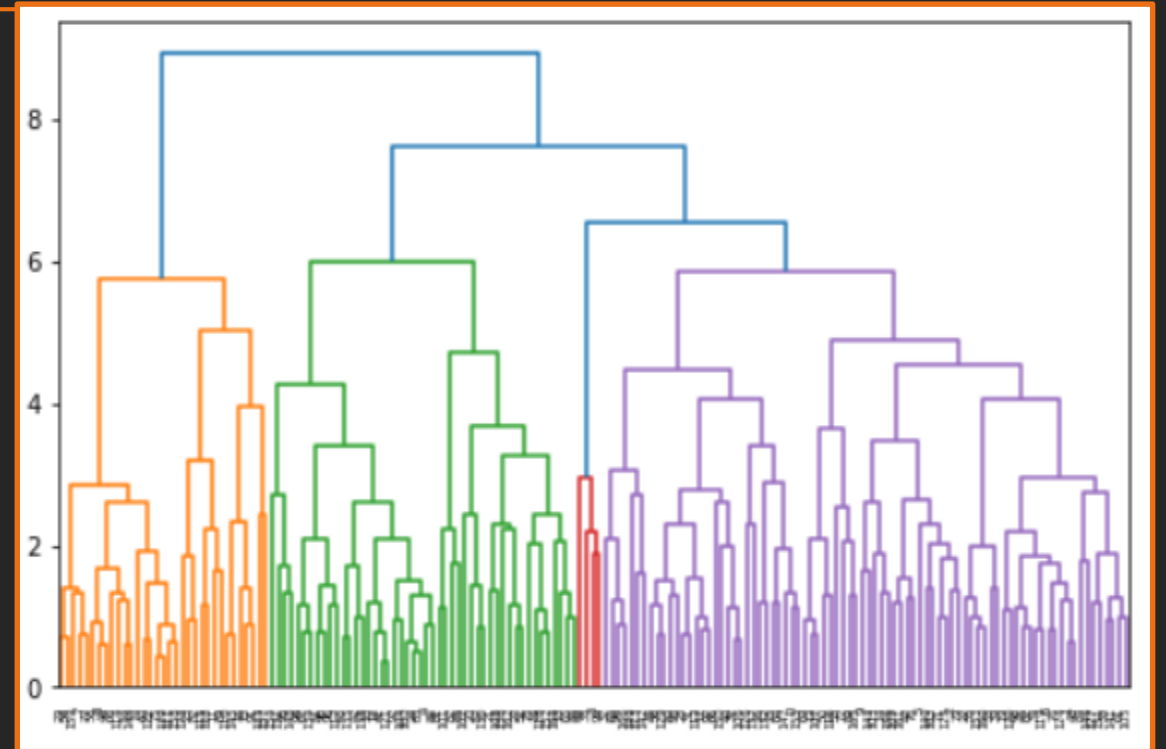- Central African Republic
- Malawi
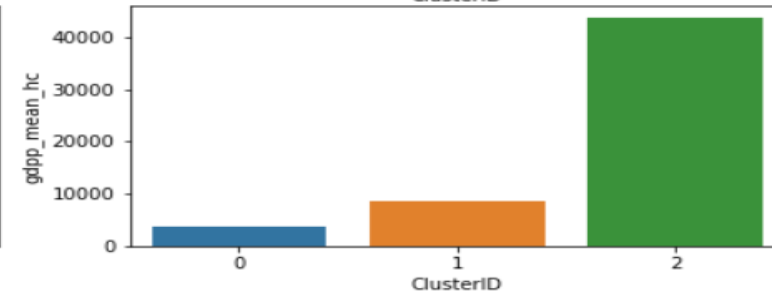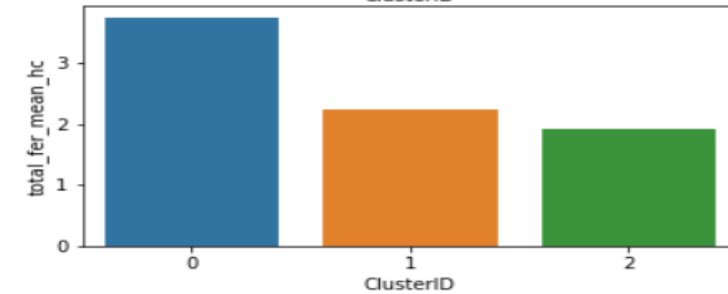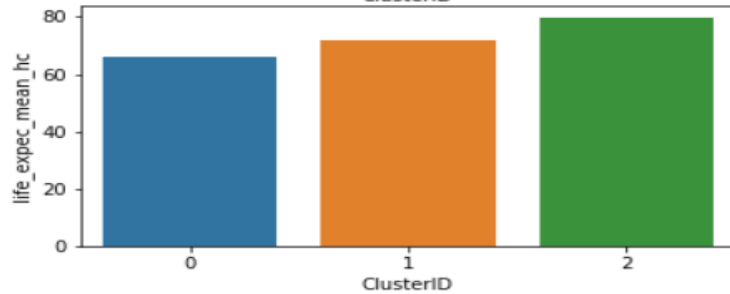- Eritrea

Hierarchical Clustering

We notice that single linkage visualization does seem to be clear enough. Hence we went with Complete linkage.

Single Linkage

Complete Linkage
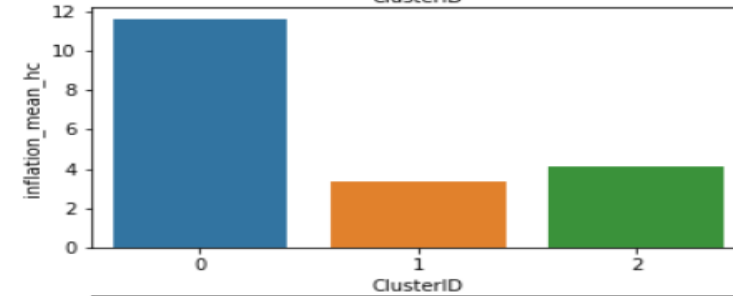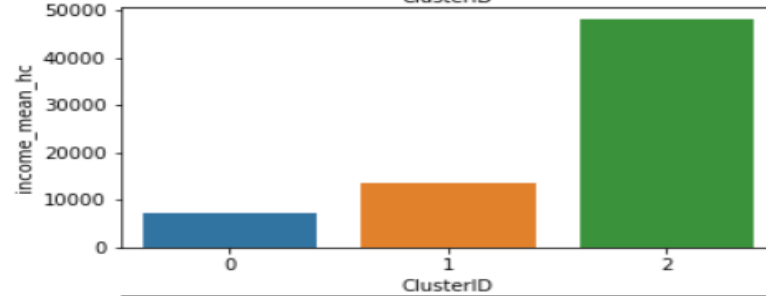
Cluster Analysis (Hierarchical)

From the graphs we see **Cluster – 0** has:
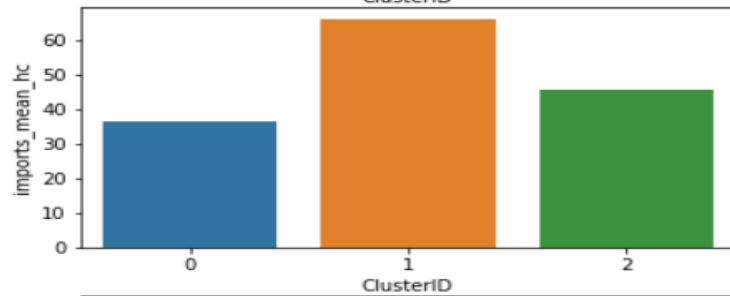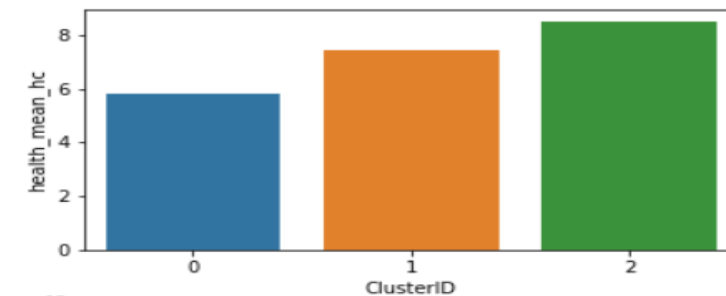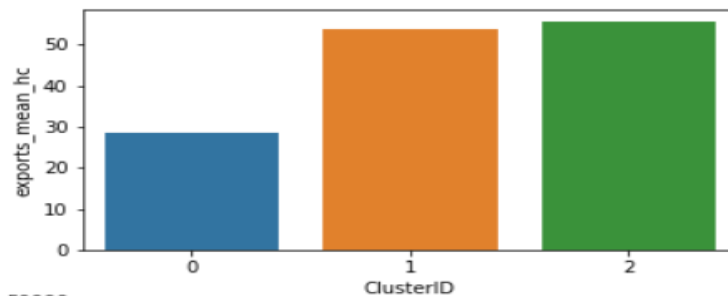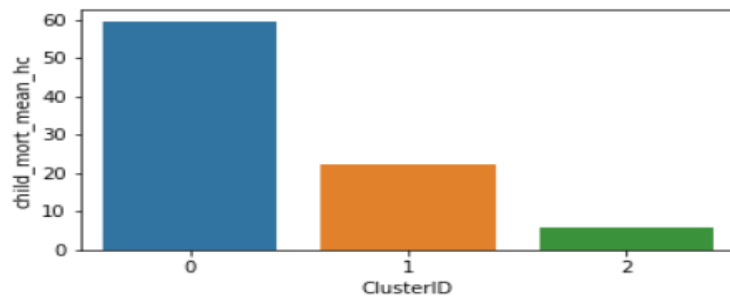
- It has highest child mortality
- Lowest income
- Highest Inflation
- Comparatively low life expectancy
- Highest total fertility
- Lowest gdpp.

Countries (Hierarchical Clustering)

As per hierarchical clustering, the country which are direst need of aid are:

- Burundi
- Congo, Dem. Rep
- Niger
- Sierra Leone
- Madagascar
- Mozambique
- Central African Republic
- Malawi
- Eritrea
- Togo

As by both K means and Hierarchical clustering method - we have got almost same countries which requires aid. The following are the countries which are in direst need of aid by considering socio – economic factor into consideration:

- Burundi
- Congo, Dem. Rep.
- Niger
- Sierra Leone
- Madagascar
- Mozambique
- Central African Republic
- Malawi
- Eritrea

Conclusion