CLUSTERING AND PCA ASSIGNMENT

Sushma Subburayan

CLUSTERING AND PCA ASSIGNMENT

Problem Statement: Categorize the countries using some socio-economic and health factors that determine the overall development of the country and suggest at least five countries which are in direst need of aid.

DATA :

Country-data data set containing the socio economic factors of 167 countries

GOALS OF THE ANALYSIS:

To identify at least top five countries which are in the direst need of aid

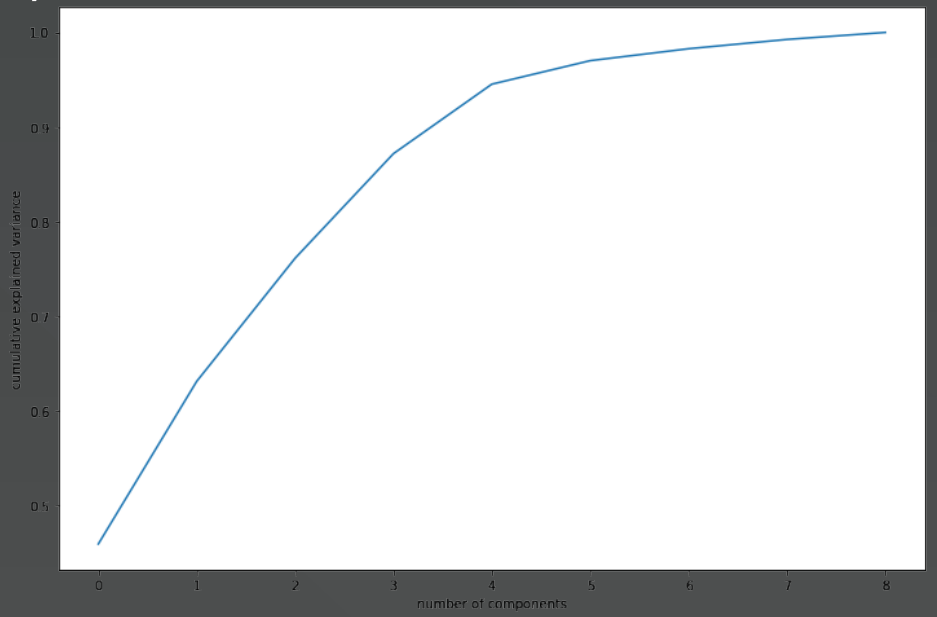
- The following are the socio economic factors given for the countries
  - Name of the country
  - Child Mortality - Death of children under 5 years of age per 1000 live births exports
  - Exports of goods and services. Given as %age of the Total GDP
  - Health - Total health spending as %age of Total GDP
  - imports - Imports of goods and services. Given as %age of the Total GDP
  - Income - Net income per person
  - Inflation - The measurement of the annual growth rate of the Total GDP
  - Life Expectancy - The average number of years a new born child would live if the current mortality patterns are to remain the same
  - Total Fertility - The number of children that would be born to each woman if the current age-fertility rates remain the same.
  - GDPP - The GDP per capita. Calculated as the Total GDP divided by the total population.

- 1 .Read and understood the given data set.
- 2. Cleaned the data
  - a. Checked for null values
  - b. Checked for duplicate records
- 3. Visualized the data
  - a. Created pair plots for all the numeric variables in the data set
  - b. Created box plots for all the categorical variables in the data set
- 4. Data Preparation
  - a. Tried removing outliers(before PCA and after PCA),but as the outlier elimination is leading to loss of many countries and giving incorrect results ,went ahead without eliminating the outliers.
  - b. Scaled all the numeric variables using standard scaler.

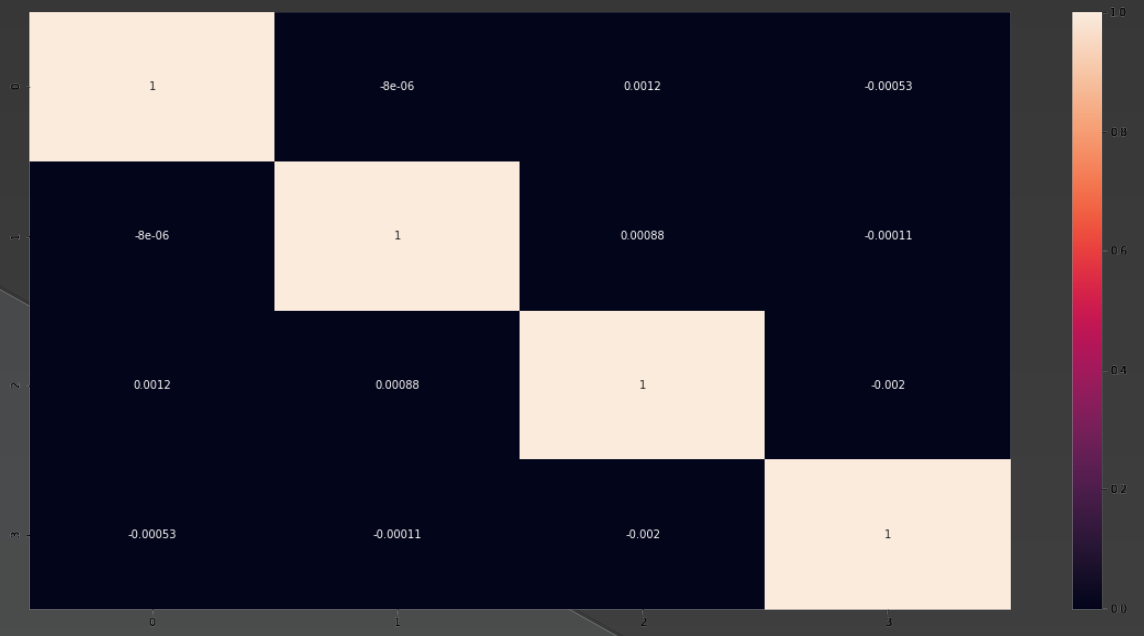
- 5. Hopkins Statistics

  As the Hopkins statistics value was good, went ahead with PCA

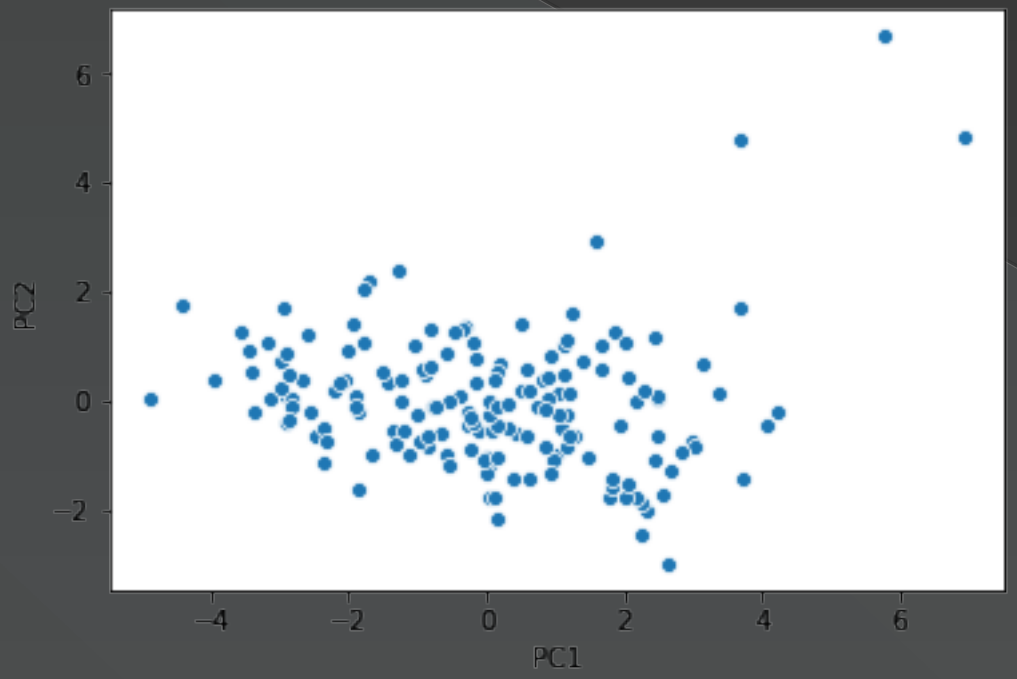
6. PCA
   - a. Performed PCA on the scaled data
   - b.Max varaiance was explained by the first 4 PCS(based on pca.explained_variance_ratio_ and the scree plot)



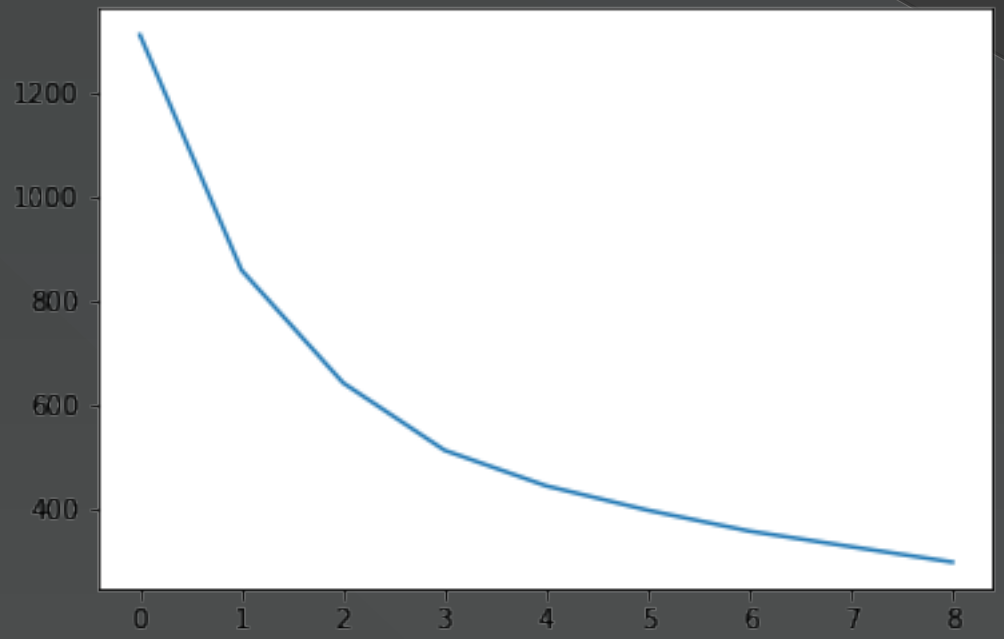
- c.Performed incremental PCA for efficiency with 4 components - four components were selected based on the scree plot
- d.Created correlation matrix and heatmap for the principal components

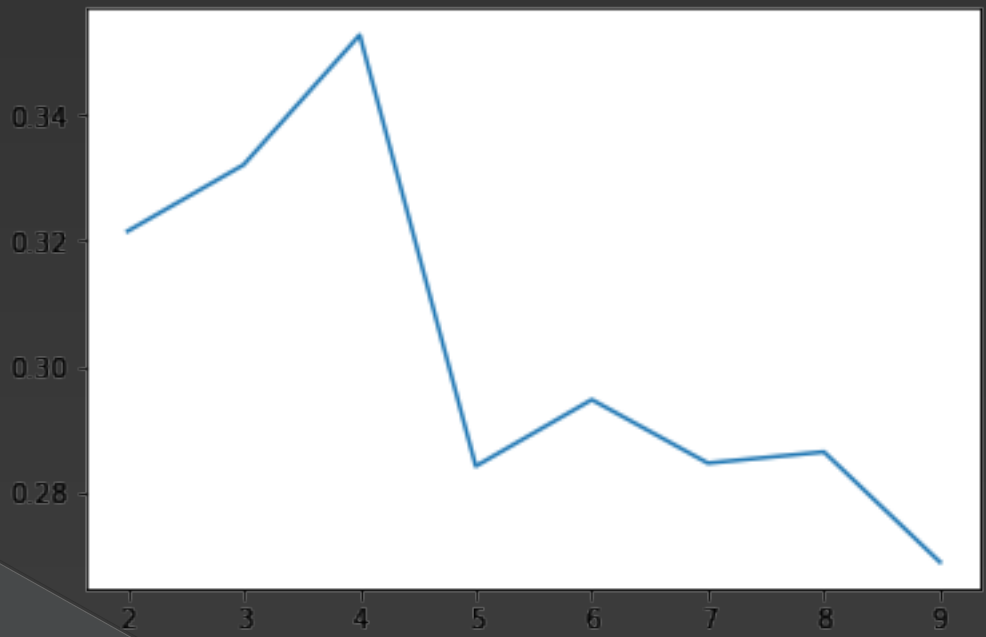


e.Made a scatter plot to visualize the PCs

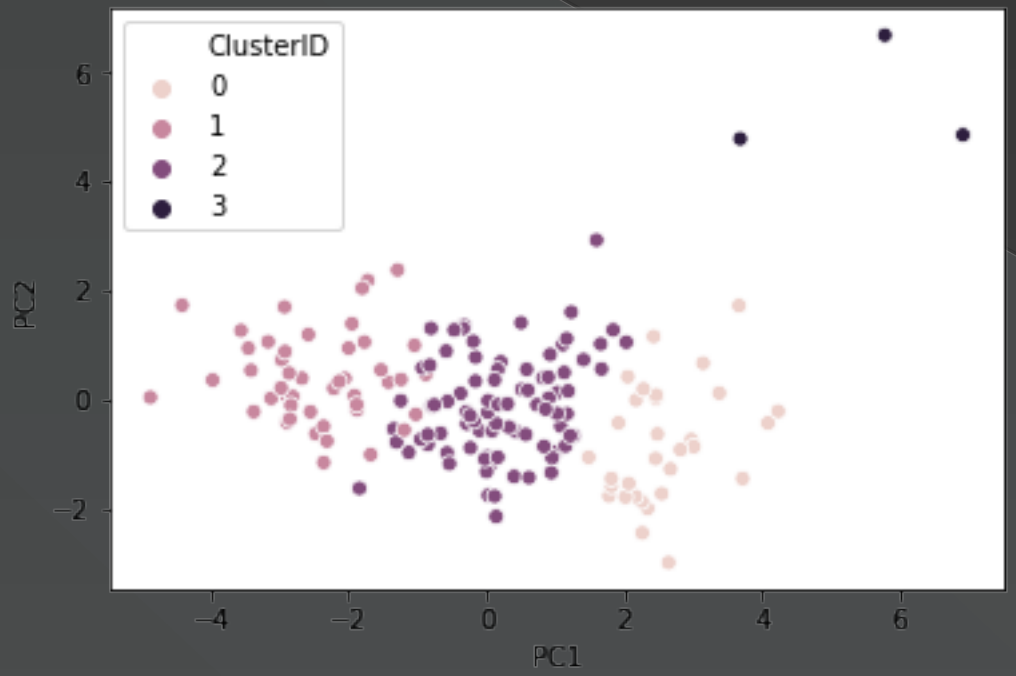


- 7. Hopkins Statistics
  - As the Hopkins statistics for PCs was good, went ahead with K-Means Clustering
- 6. K-Means Clustering
  - a.Did silhouette score analysis and created elbow curve to identify the number of clusters

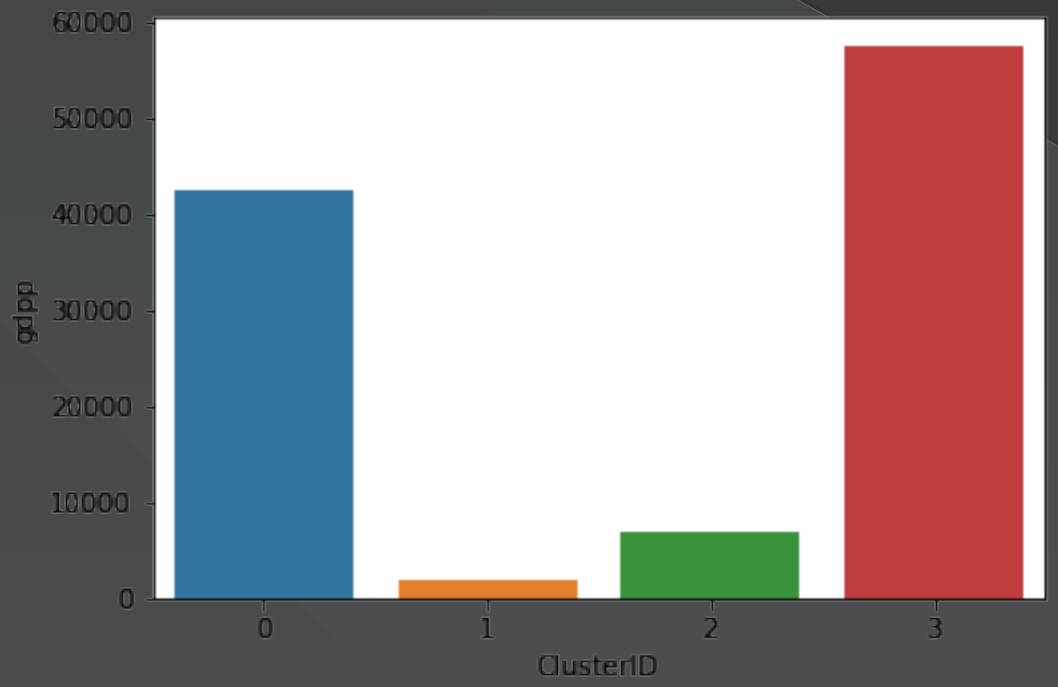


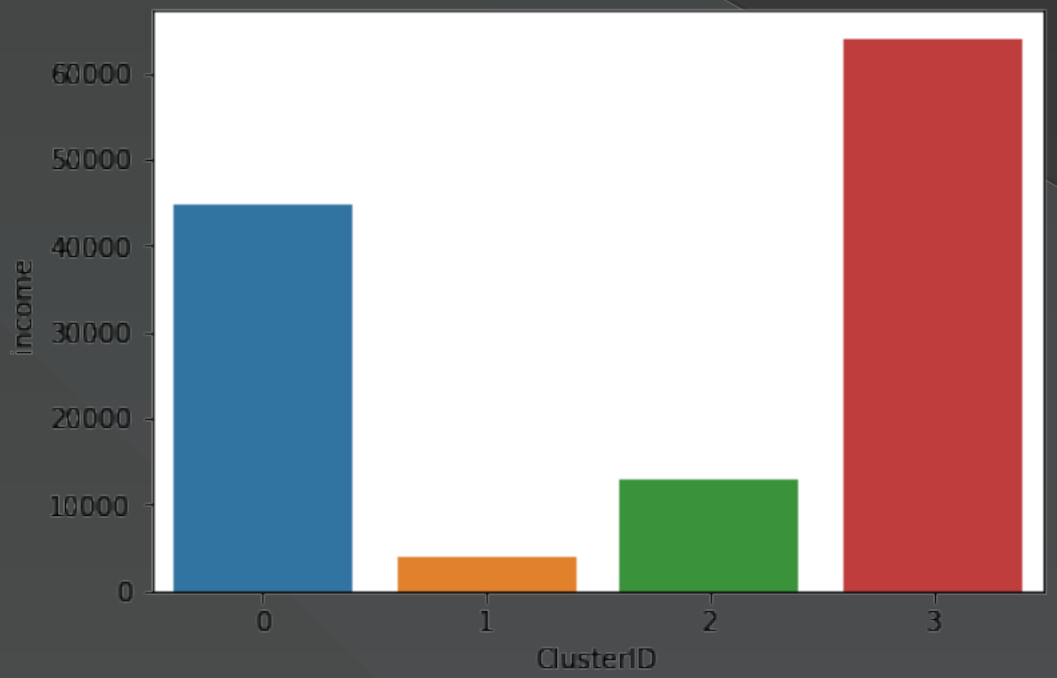


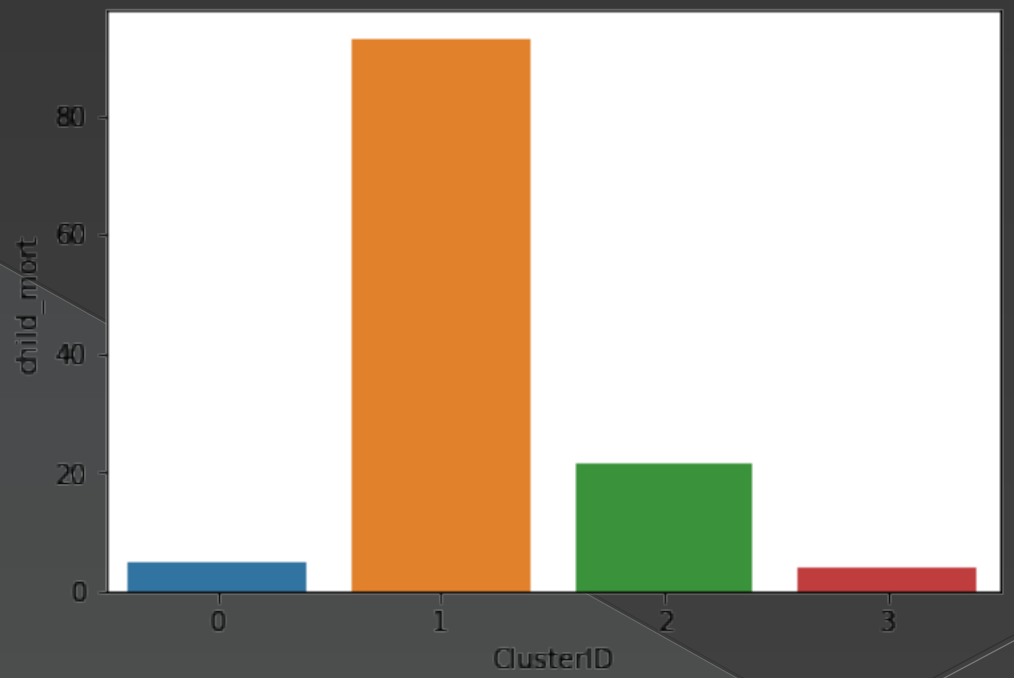
- b.Performed K-Means clustering with four clusters
- c.Analyzed the clusters formed

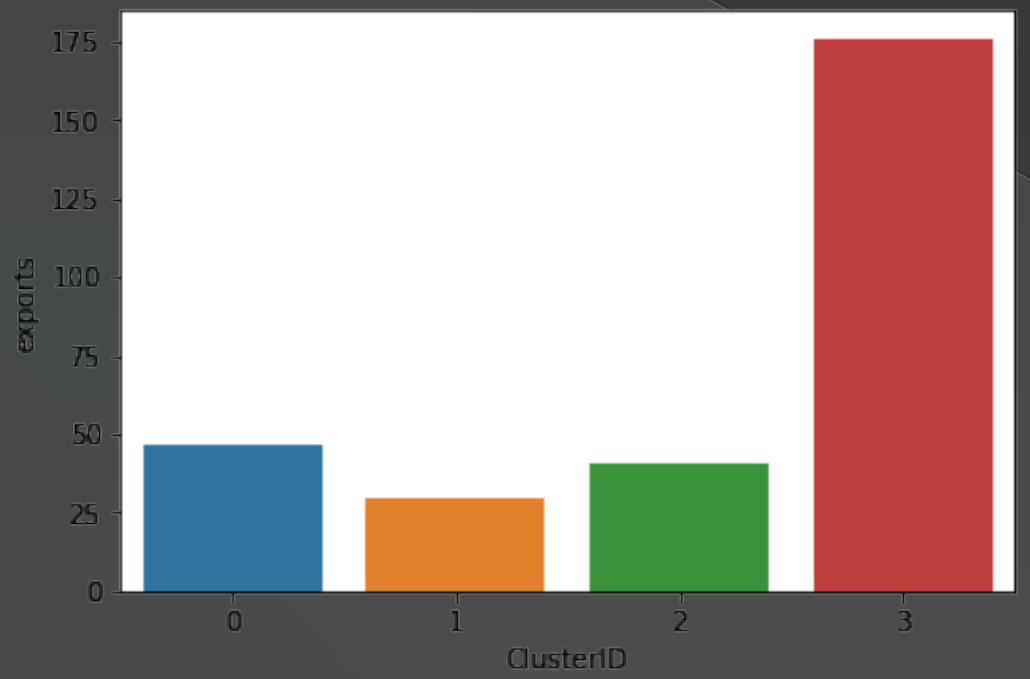


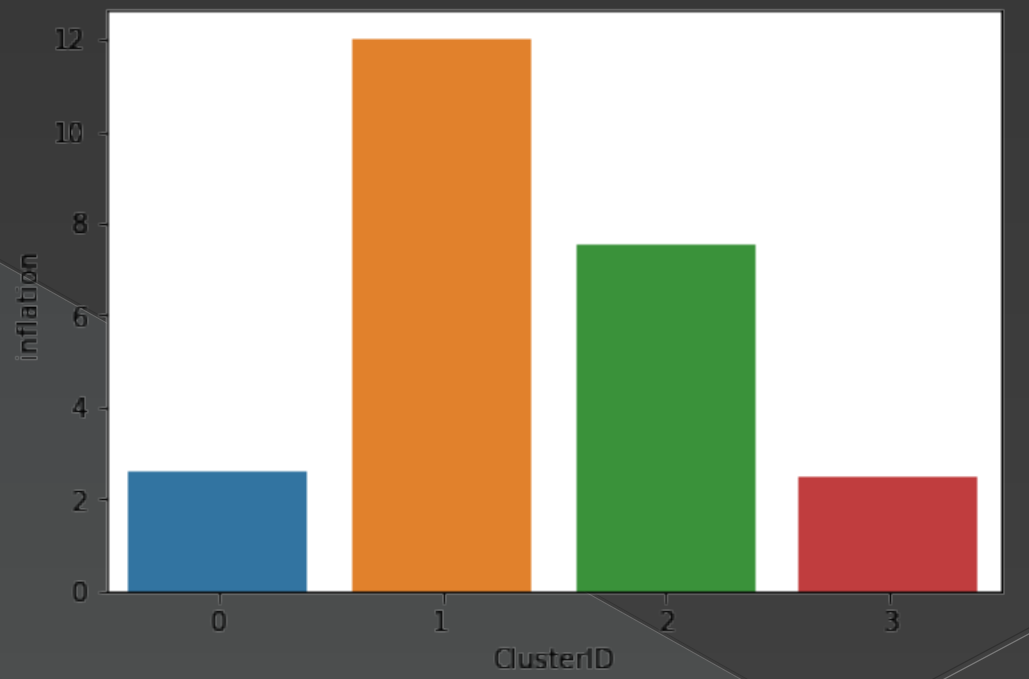
- d.Plotted bar graphs between the clusters and the features and identified the clusters with countries in the direst need of aid(Clusters 1and 2)

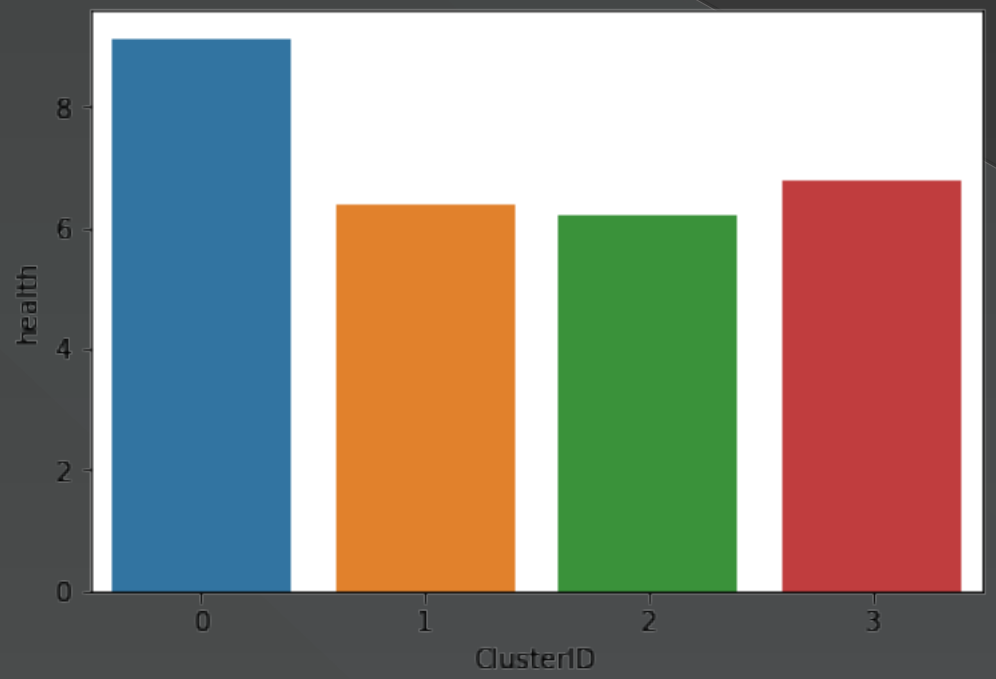


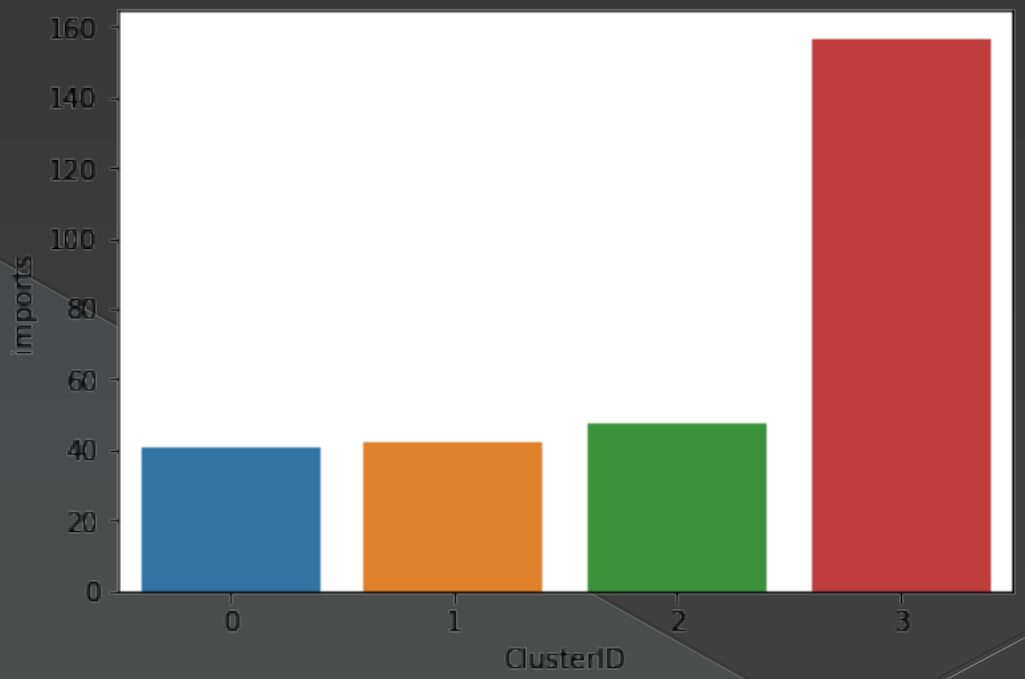

60000
50000
40000
30000
20000
10000
0
income
0
1
2
3
ClusterID

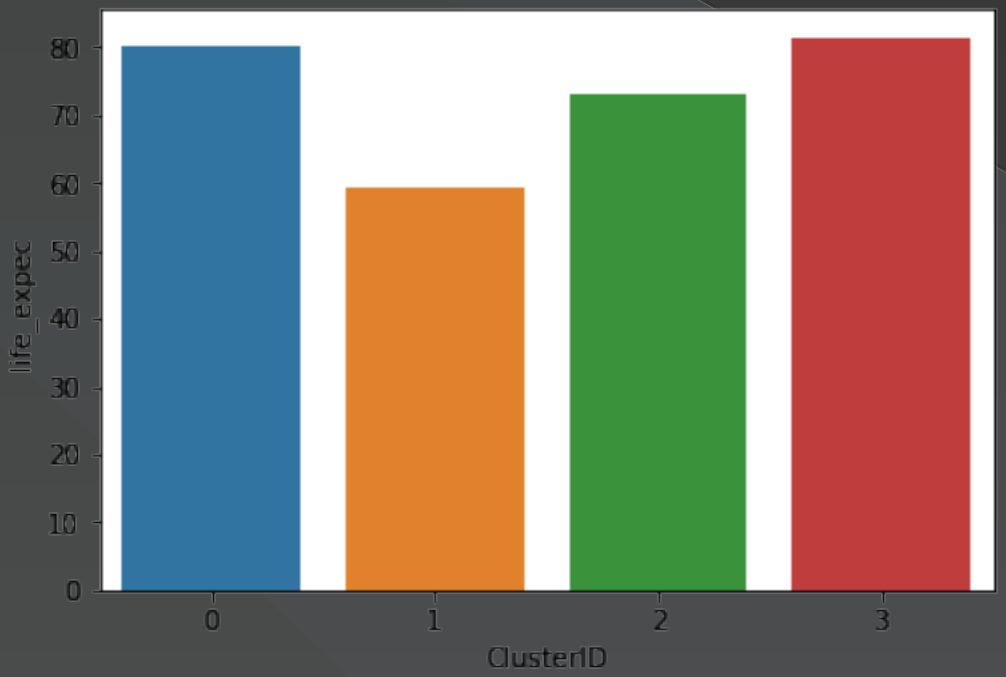

80
60
40
20
0
child_mort
0
1
2
3
ClusterID

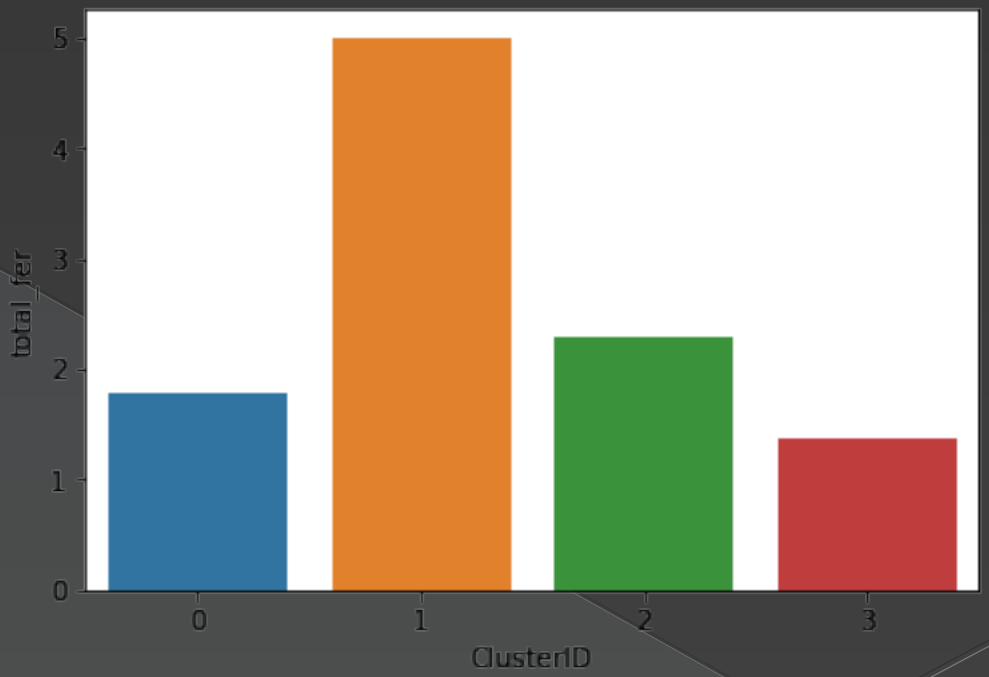

175
150
125
100
75
50
25
0
exports
0
1
2
3
ClusterID


12
10
8
6
4
2
0
inflation
0
1
2
3
ClusterID


health
0
2
4
6
8
0
1
2
3
ClusterID

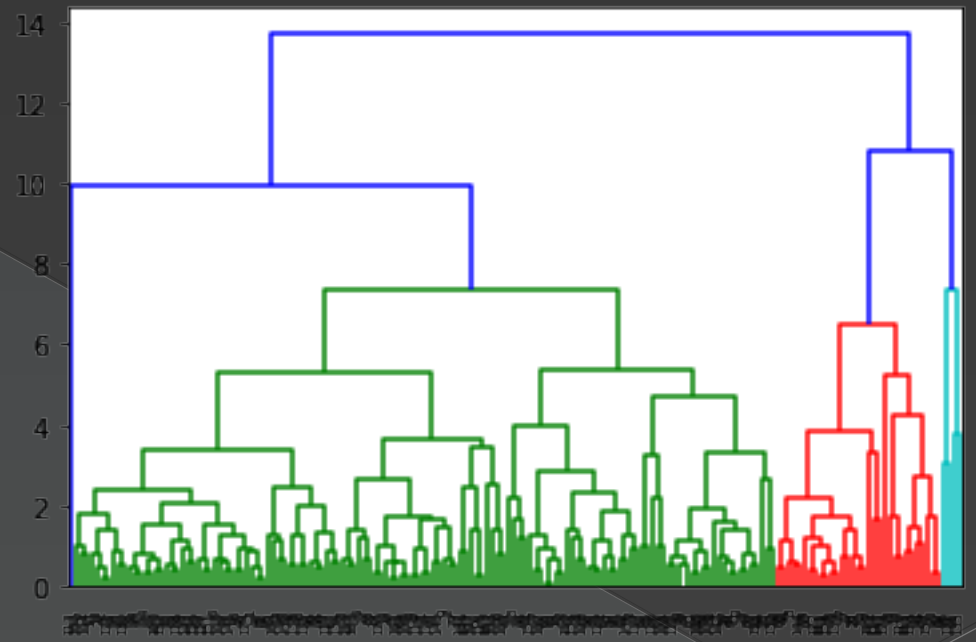

imports
0
20
40
60
80
100
120
140
160
0
1
2
3
ClusterID


life_expec
80
70
60
50
40
30
20
10
0
0
1
2
3
ClusterID

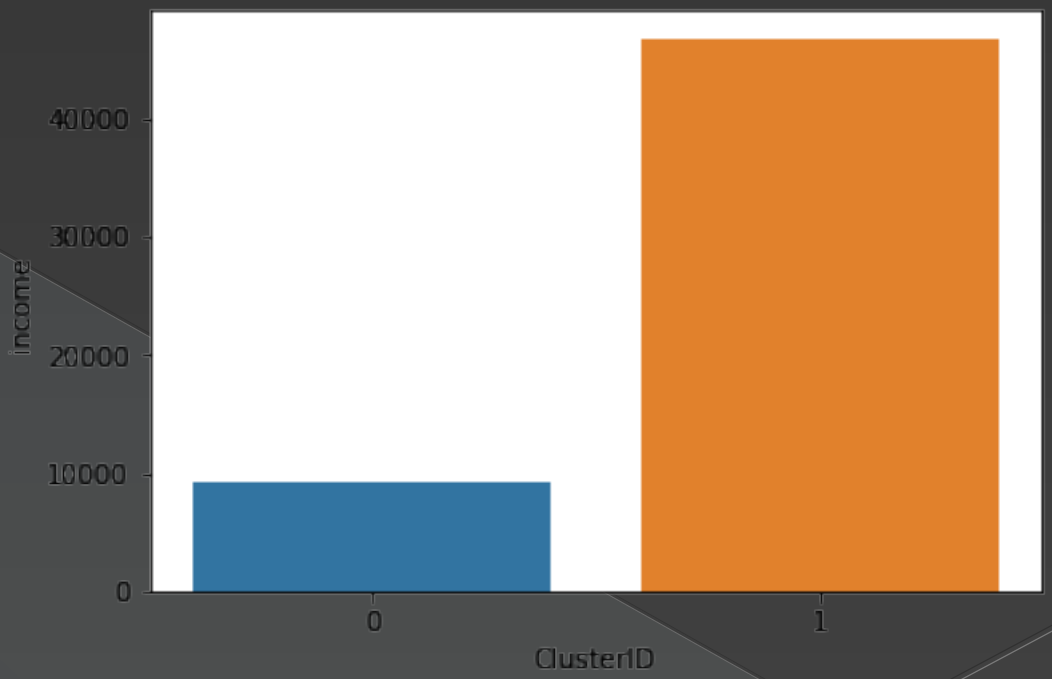

total_fer
5
4
3
2
1
0
0
1
2
3
ClusterID

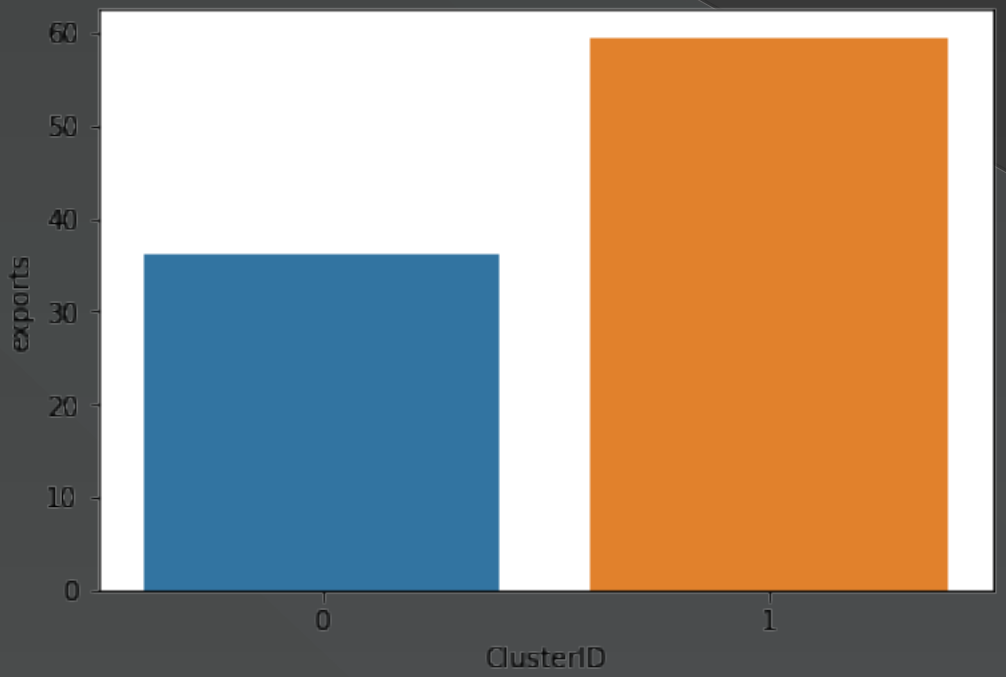
- e.Identified top five countries from the cluster which are in direst need of the aid

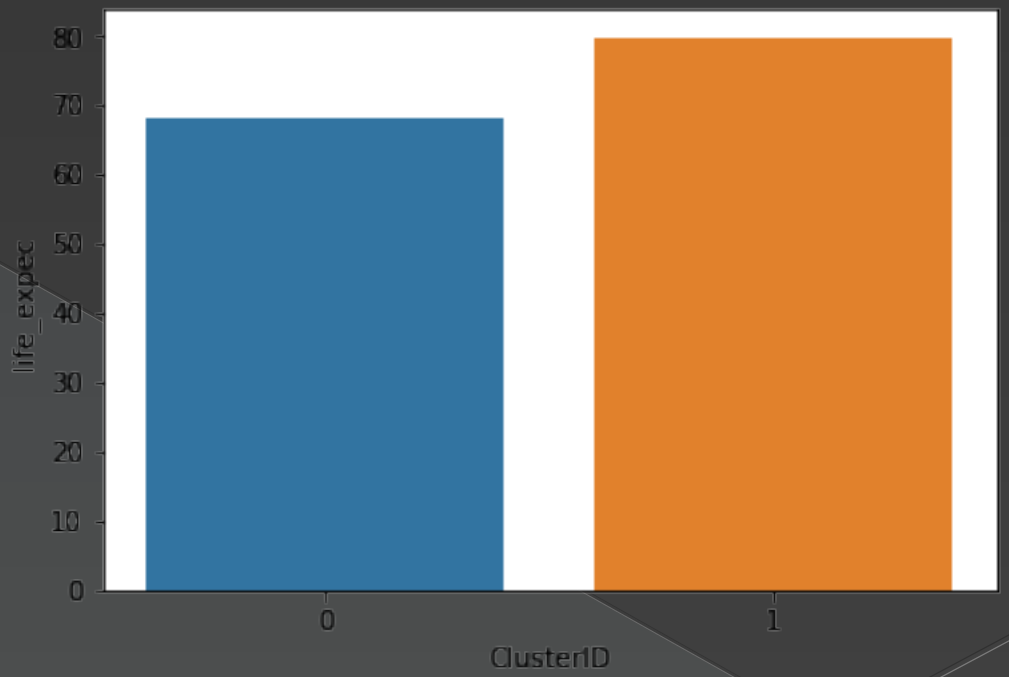
- 7.Hierarchical Clustering
  - a.Performed Hierarchical clustering with two clusters

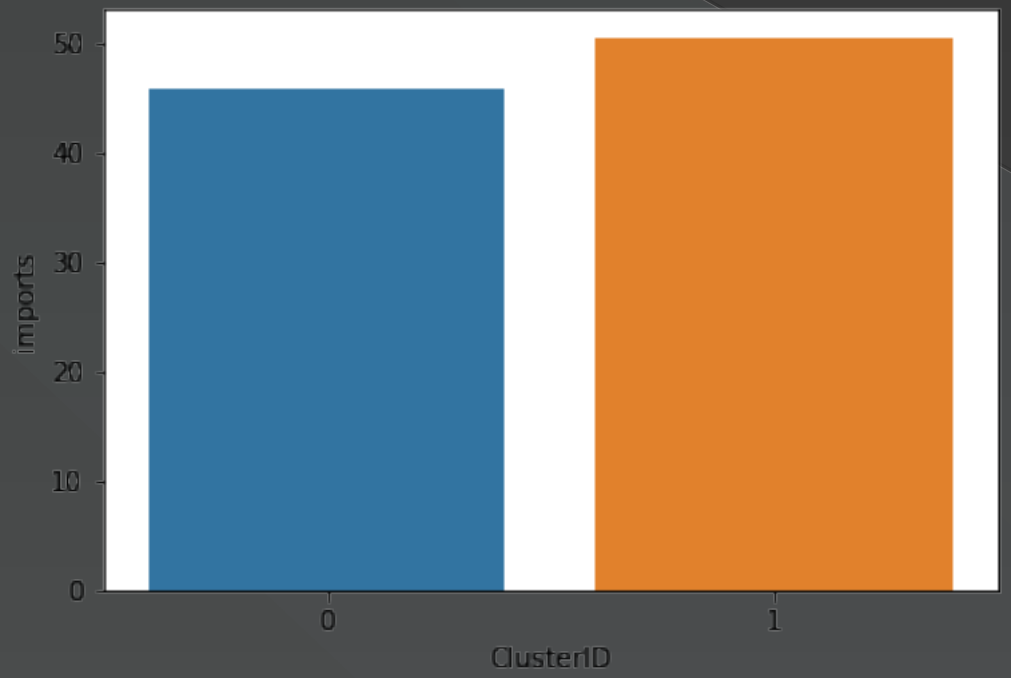


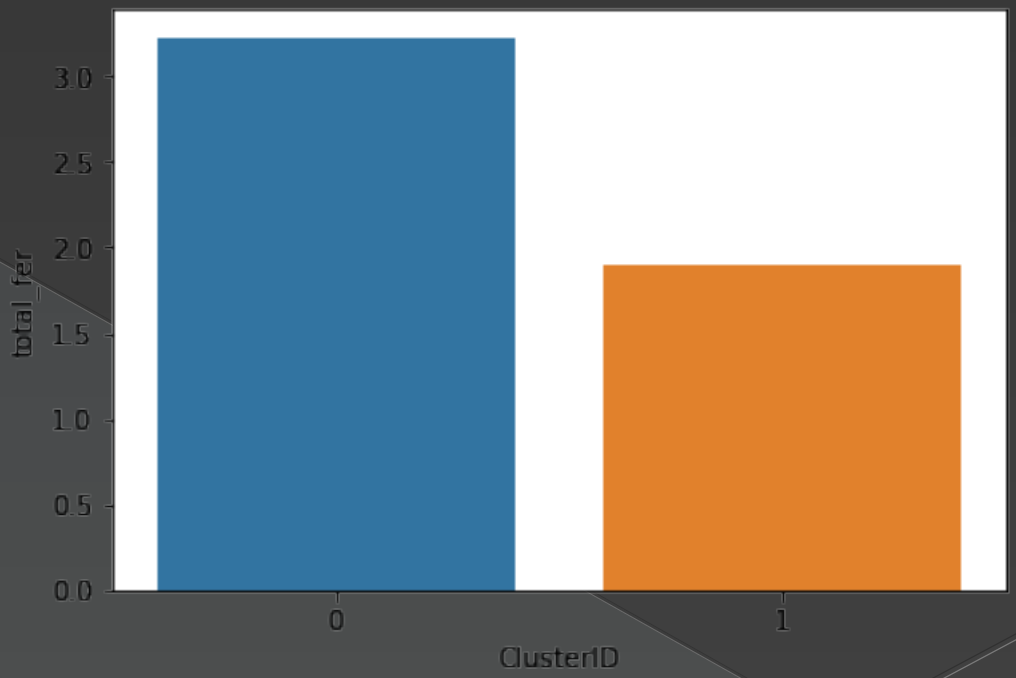
- b.Analyzed the clusters formed
- c.Plotted bar graphs between the clusters and the features and identified the clusters with countries in the direst need of aid(Clusters 0)

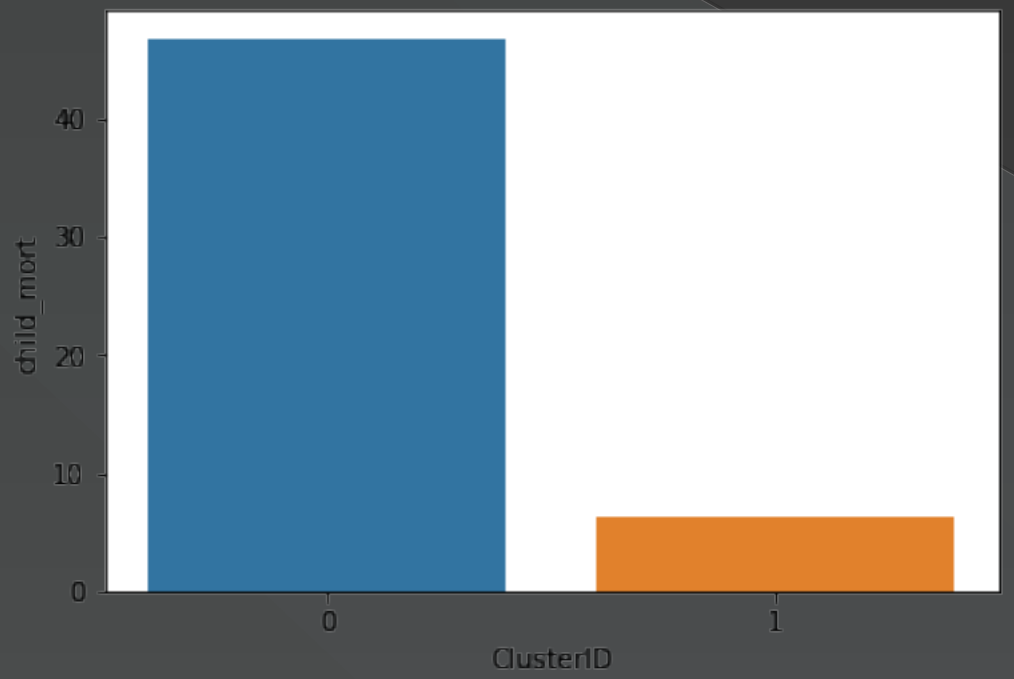


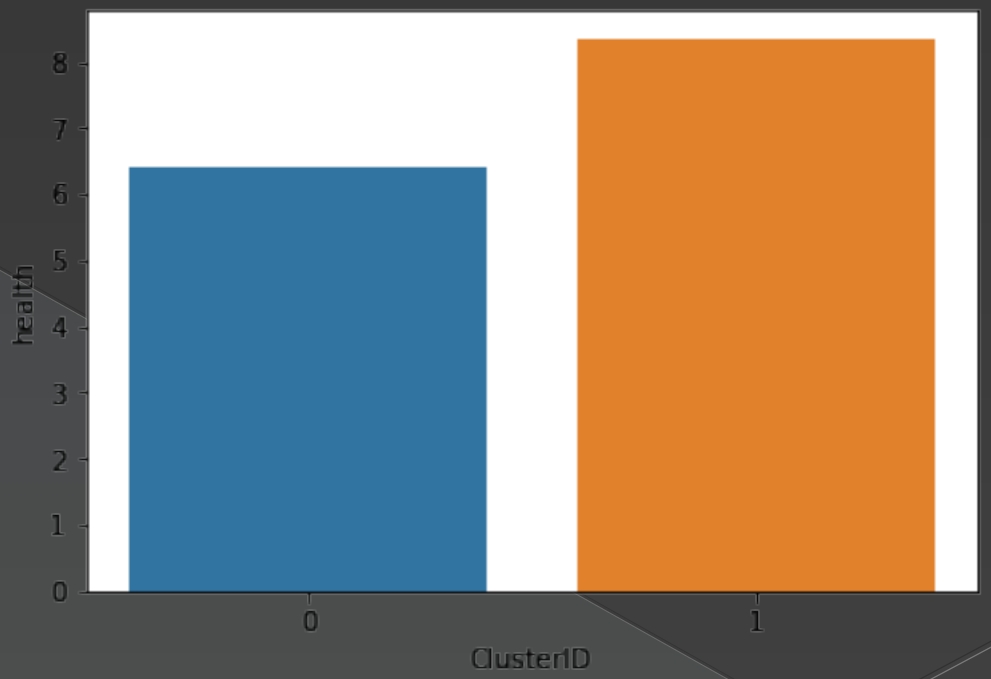

exports
60
50
40
30
20
10
0
0
1
ClusterID

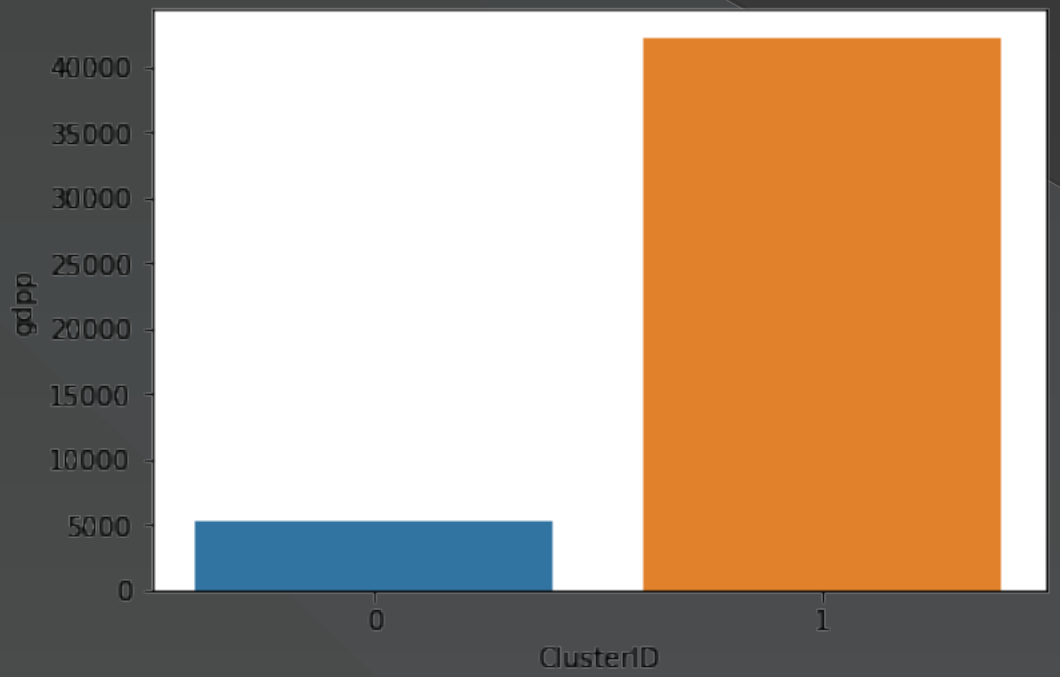

life_expec
80
70
60
50
40
30
20
10
0
0
1
ClusterID


50
40
30
20
10
0
imports
0
1
ClusterID


3.0
2.5
2.0
1.5
1.0
0.5
0.0
total_fer
0
1
ClusterID


child_mort
40
30
20
10
0
0
1
ClusterID


health
8
7
6
5
4
3
2
1
0
0
1
ClusterID


40000
35000
30000
25000
20000
15000
10000
5000
0
gdpp
0
1
ClusterID

- d.Identified top five countries from the cluster which are in direst need of the aid

- 8. The list of top five countries which are in the direst need of aid are:
  - 1.Congo, Dem. Rep.
  - 2.Burundi
  - 3.Liberia
  - 4.Niger
  - 5.Central African Republic

THANK YOU