Clustering & PCA Assignment

SUBMISSION

Submitted by:

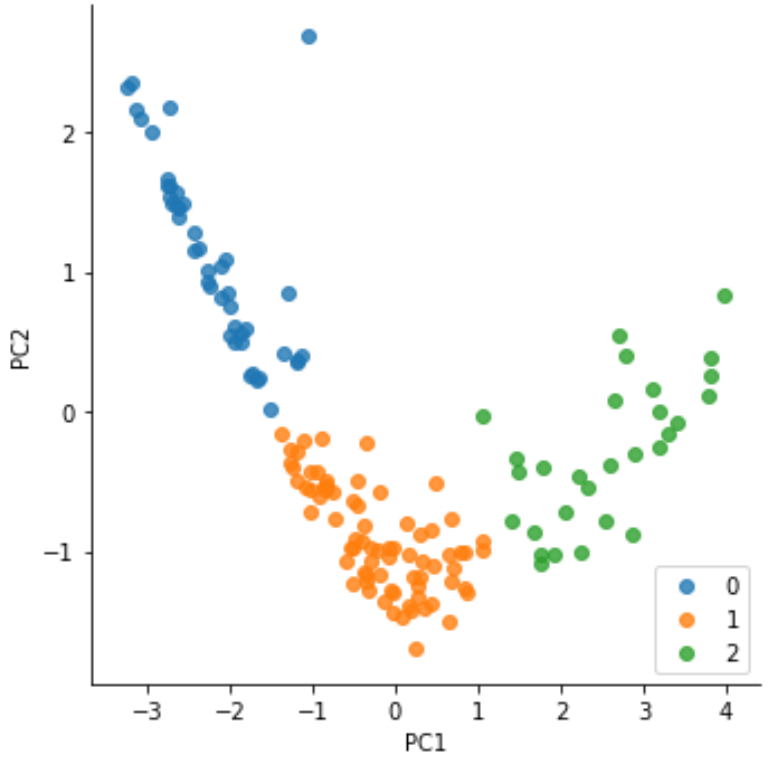
Rajasekhar Battula

Problem statement and Approach

- HELP international have been able to raise around $ 10 million. The significant issue is to chose the countries that are in the direst need of aid.
- Below approach is followed to choose the countries.
  - As part of data cleaning, feature imports, exports and income is converted in terms of GDPP.
  - After initial EDA on the given data, PCA is performed on the data and observed that 3 Principal components explains about 87.3% of total variance in data.
  - On this PCA data, outlier analysis is performed and 18 countries are identified as outliers. Now the total count of countries is 149.
  - Both K-means (with k= 4) and Hierarchical Clustering (3 clusters) are applied on the above data.
  - A total of 3 clusters are finalized looking at their socio- economic and health factors.
  - Now, plotted the outliers and original data in the same plot and mapped the outliers to resp. clusters.

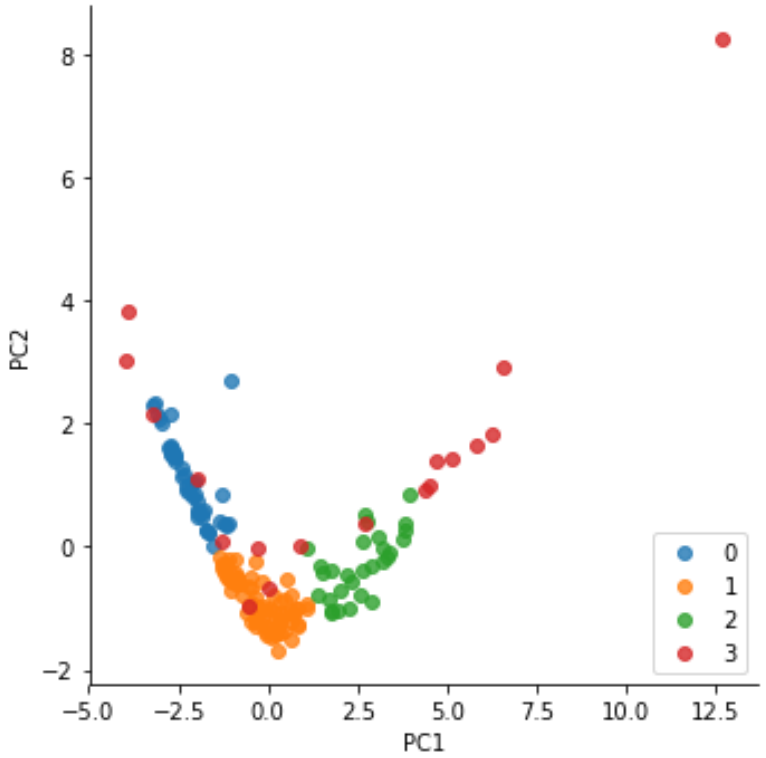
Clusters with and without outlier counties

Plot of clusters
without outliers

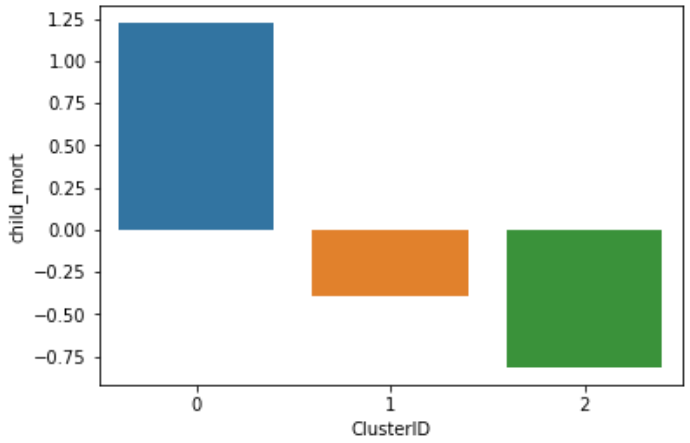


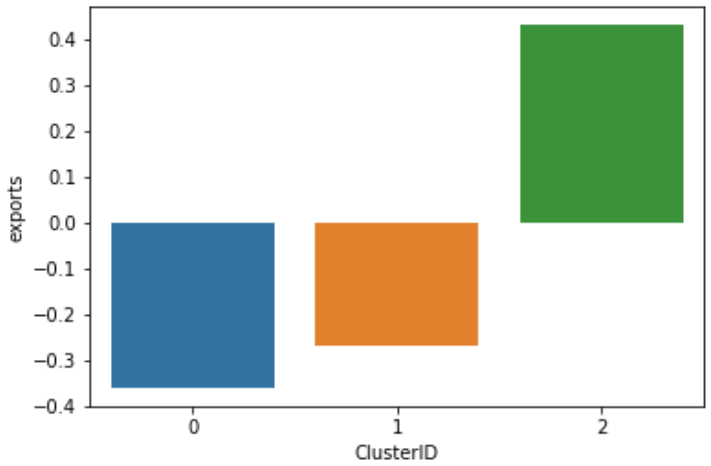
Cluster 0 is where there is need of help.

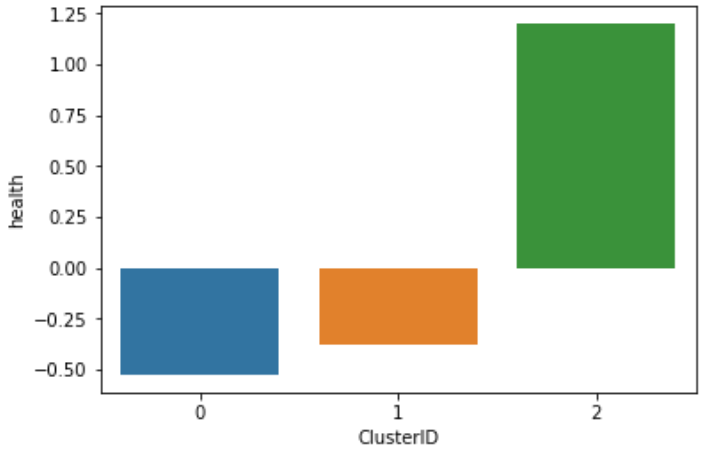
Plot of clusters
with outliers

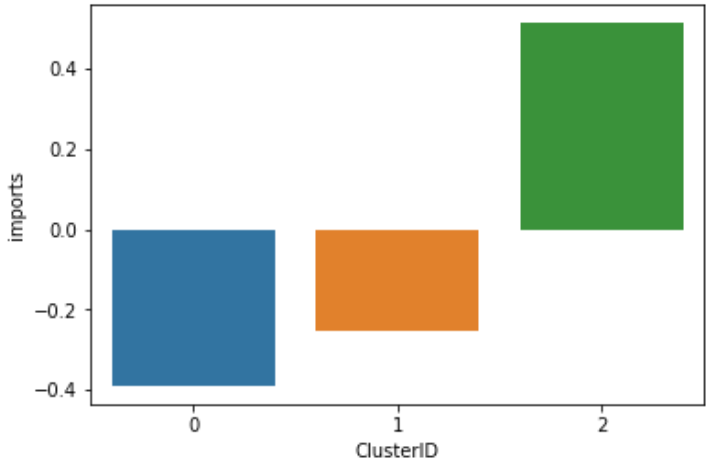


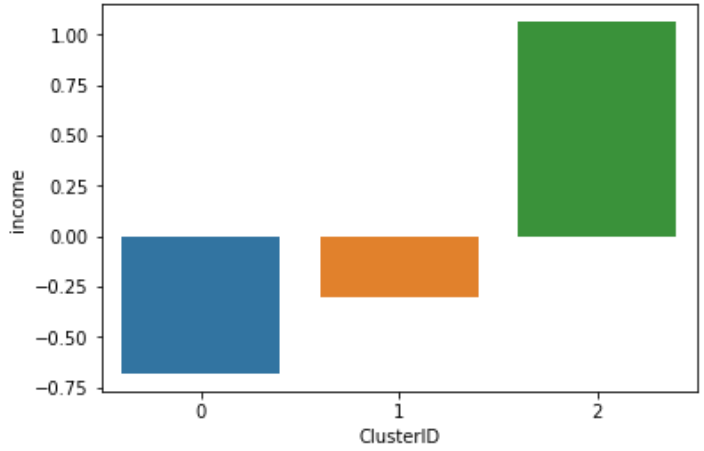
Mean analysis of clusters

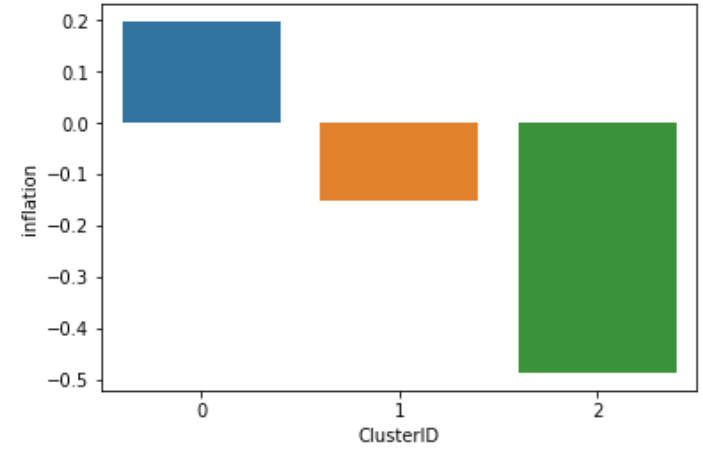




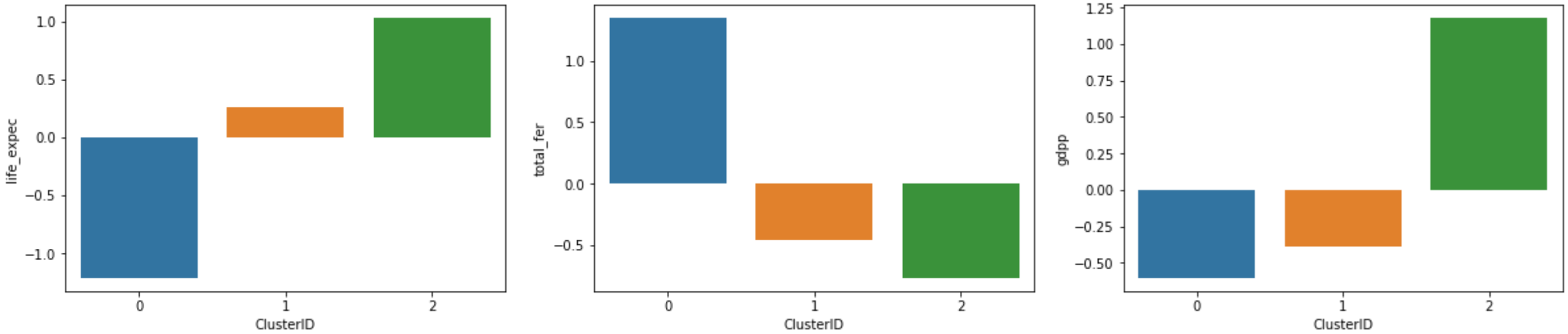






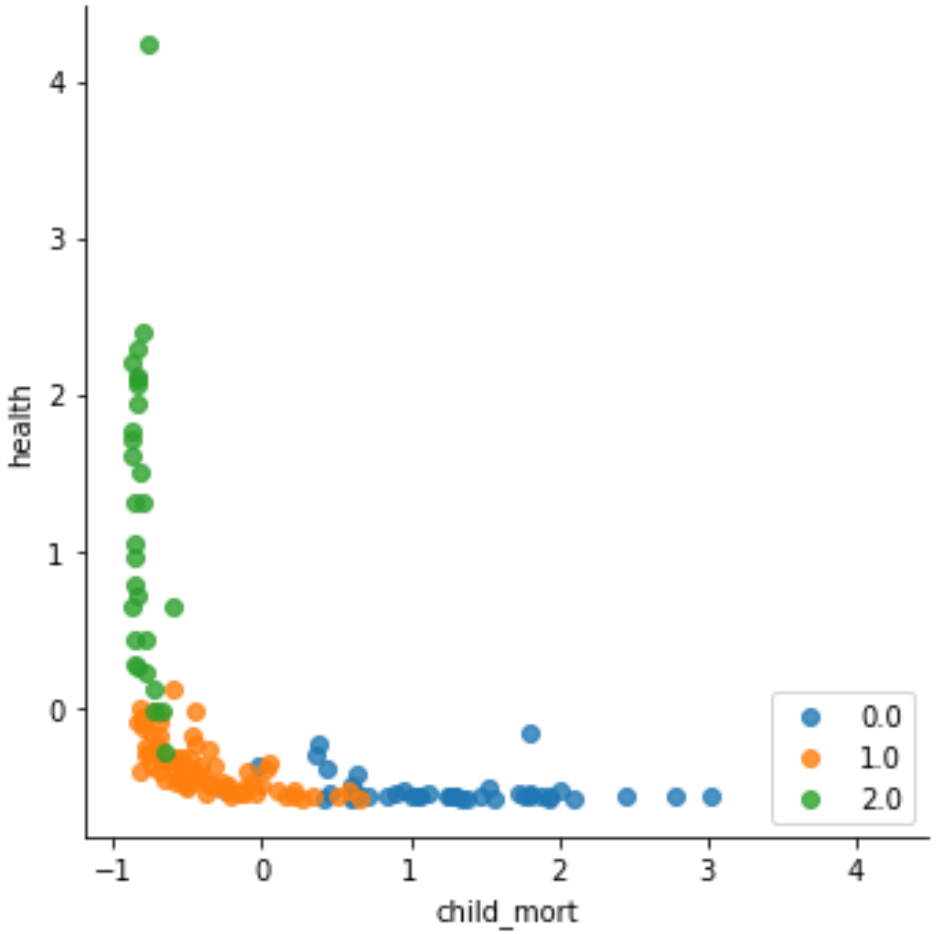


Mean analysis of clusters (Contd.)



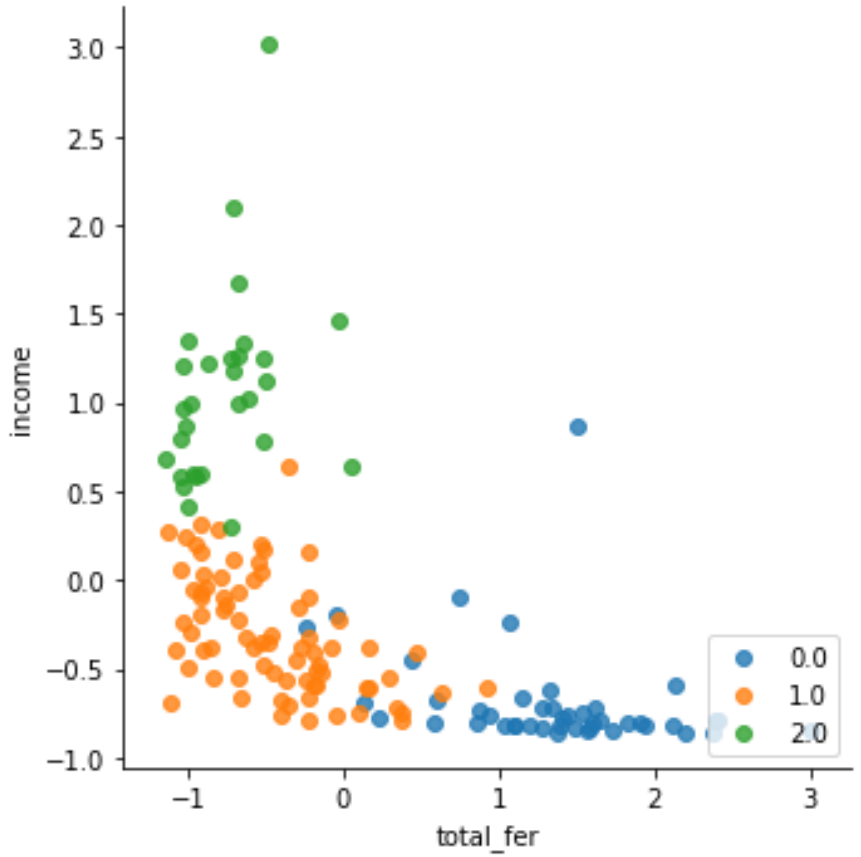
All the barplots show that cluster 0 is having low income per person, high child mortality, low gdpp, high inflation, high total fertility, low life expectancy and low in imports, exports and health. This is the cluster which requires attention from NGO

Plot for Child Mortality vs health of final clusters



It is clearly visible that cluster 0 has high child mortality and less health conditions which are in need of help.

Plot for Total fertility vs income per person of final clusters



It is clearly visible that cluster 0 is having lower income per person and has high total fertility which is difficult to run family.

Conclusions

Output of this project says there are 48 nations which are in need of help, 38 nations are good in having all amenities and not required help and the rest of 81 nations are in the middle cluster.
These are the 48 countries which belong to “cluster 0”.

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