Assignment
Clustering and PCA

-Harshavardhini.R

-10.06.2019

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 project that included a lot of awareness drives and funding programmes, 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.
- The requirement is to categorise the countries using some socio-economic and health factors that determine the overall development of the country. And then suggest the countries which the CEO needs to focus on the most.

Data understanding and Preparation

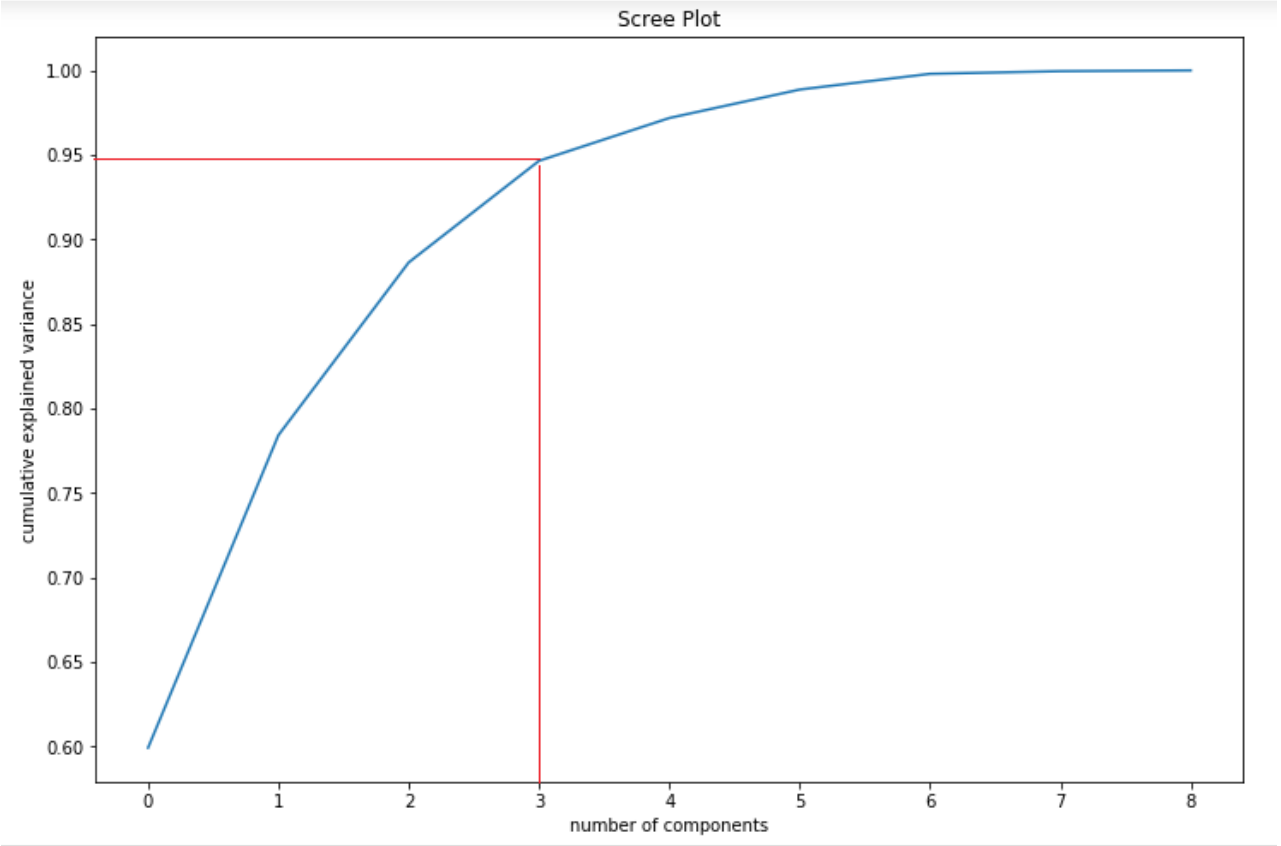
- The data contains country wise parameters – Income per person, GDP per capita, Health Spendings, Child Mortality Rate, Life Expectency, Inflation, Total Fertility, imports and exports.

- Some of the parameters – Health, Imports, Exports and Inflation are expressed as % of he Total GDPP. These parameters need to be converted to the absolute actual values per capita.

- Outlier are to be removed. But on analyzing, it was found that the outliers consist of almost 25% of the data. And if we remove the outliers, some of the countries might miss out on the aid. So they are kept as it is.

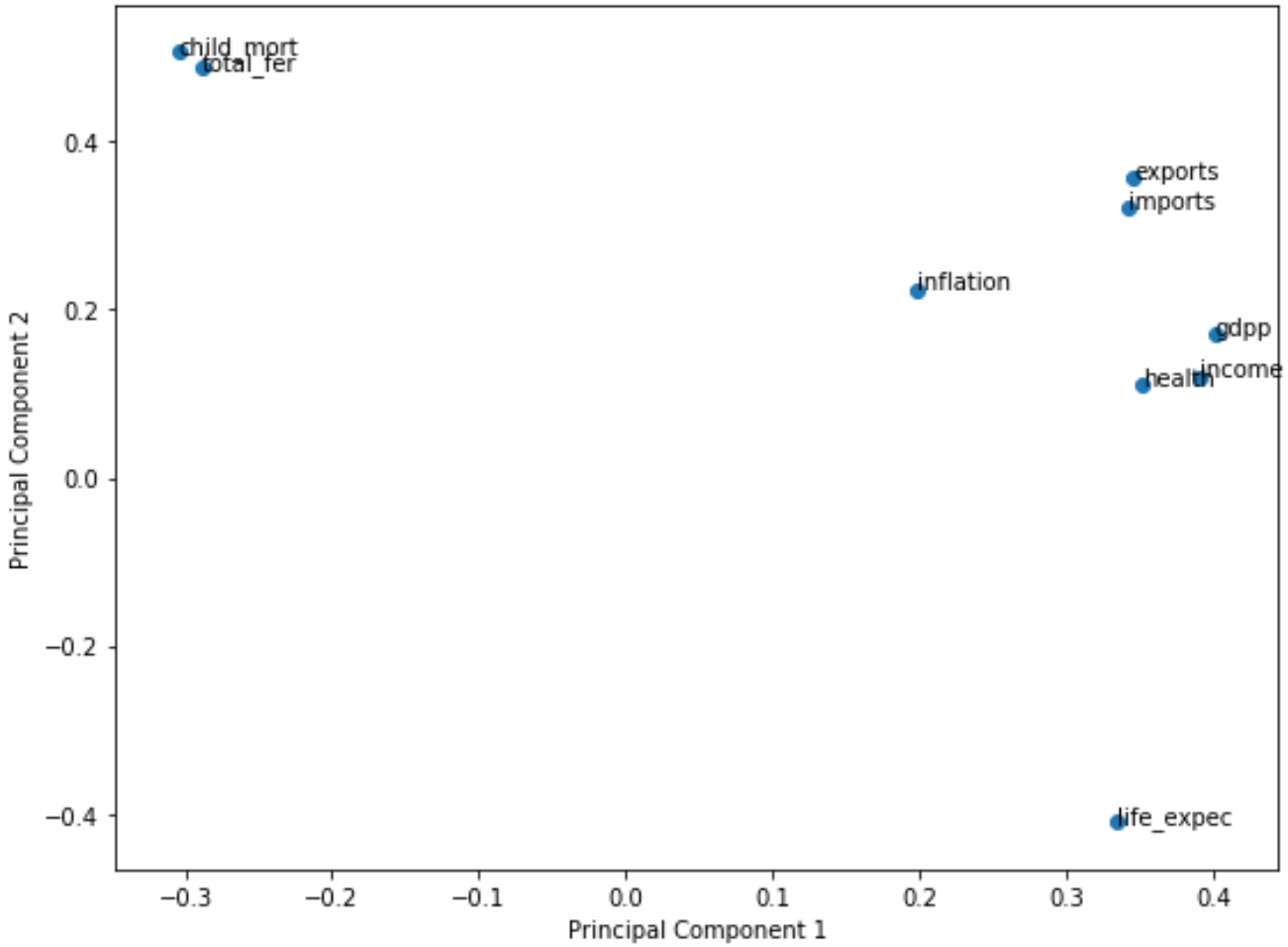
Analysis (PCA)

- On plotting the correlation matrix, it was found that the parameters are highly correlated. So for removing the correlation and for dimensionality reduction, Principal Component Analysis (PCA) needs to be done.
- First, On analysing the Scree Plot, it is found that 4 Principal Components (PC) are enough to explain almost 95% of the variance. So PCA is performed using no. of PCs as 4.



Analysis (PCA)

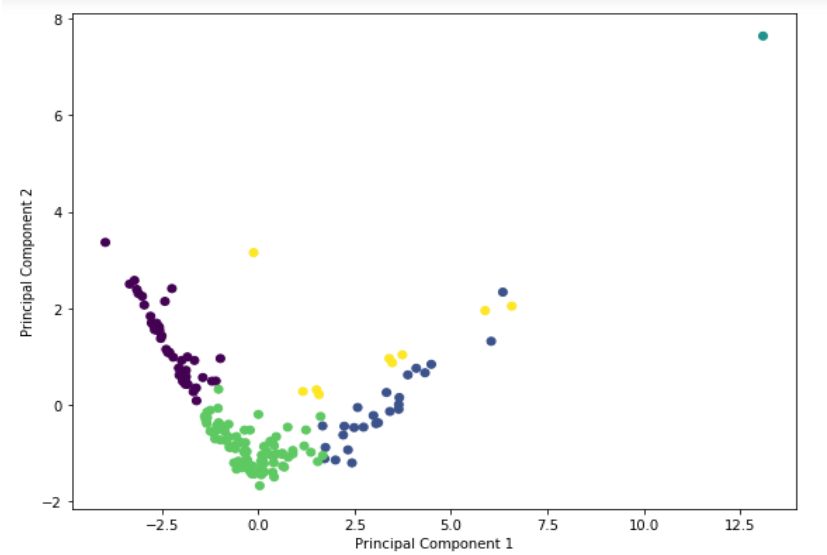
- The data set is now transformed into PCs. We see that the first component, PC1 is in the direction of 'income', 'gdpp‘, ‘export’, ‘imports’, ‘life expectancy’ and 'health'. The second component PC2 is in the direction of 'child_mort', 'total_fertility’. So the data points with low PC1 and high PC2 are to be selected for the aid.
- Now the data needs to be clustered for finding the countries with low PC1 and high PC2.
- Both K-Means and Hierarchichal Clustering is done.

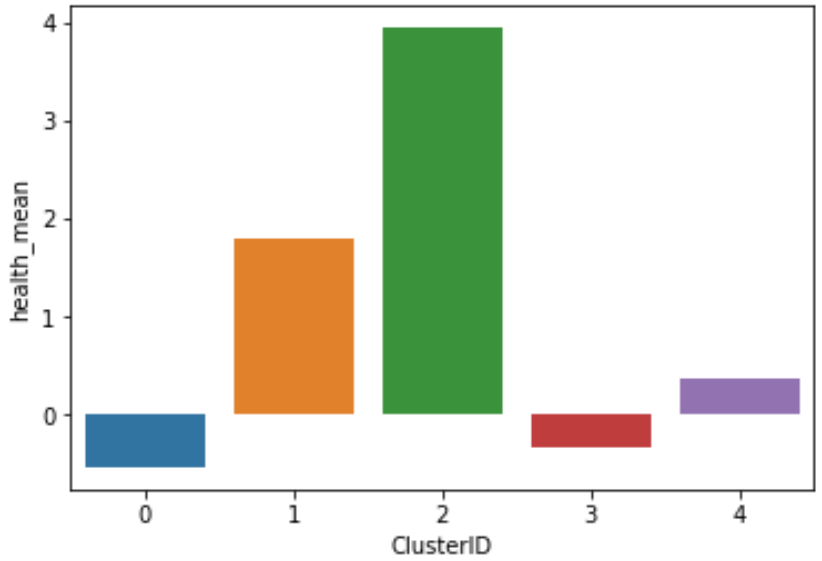


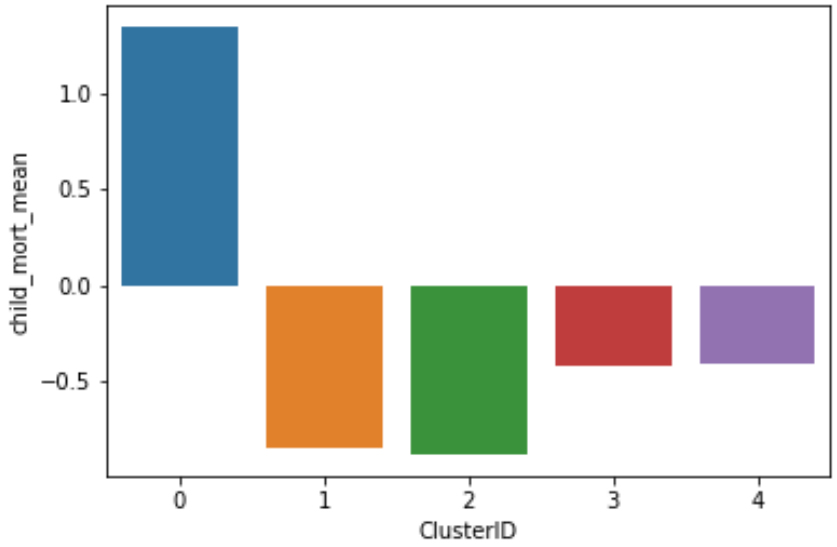
Analysis: Clustering (K-means)

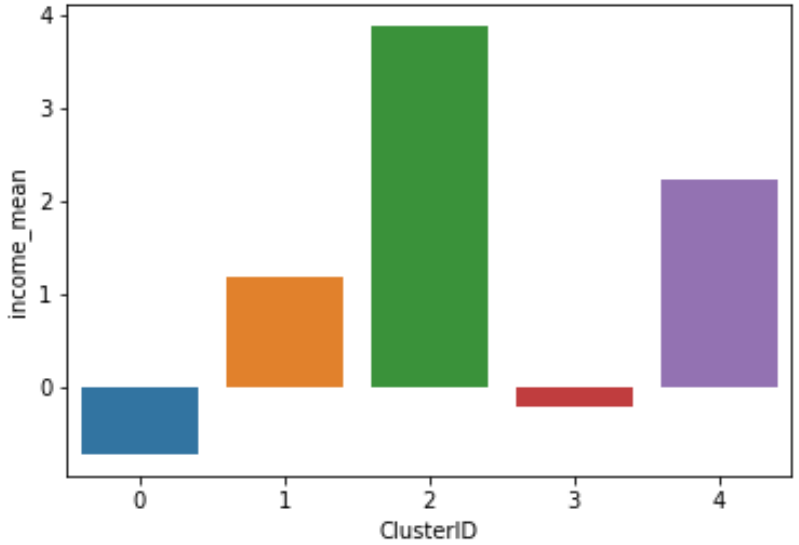
- First Silhouette Score and Sum of Squared Distances is calculated for various values of k, from 2 to 9.
- On Analysing, it was found that both k=4 or k=5 are suitable..
- Therefore, Clustering is done with both k=4 and k=5.
- The clusters chosen for both k-values have more or less the same result. But result of k=5 are more appropriate.
- After Clustering, for each Cluster Id, the mean values of all PCs and some original features are calculated and plotted.

K-Means Clustering (k=5)



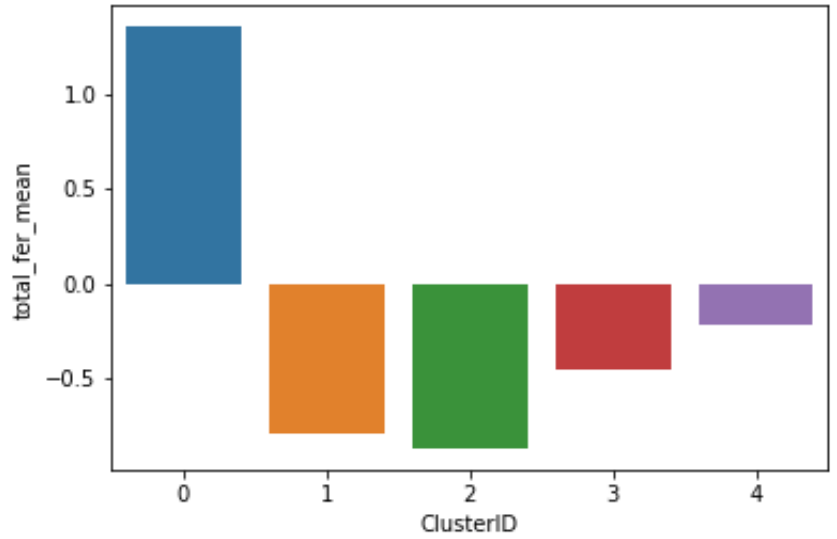


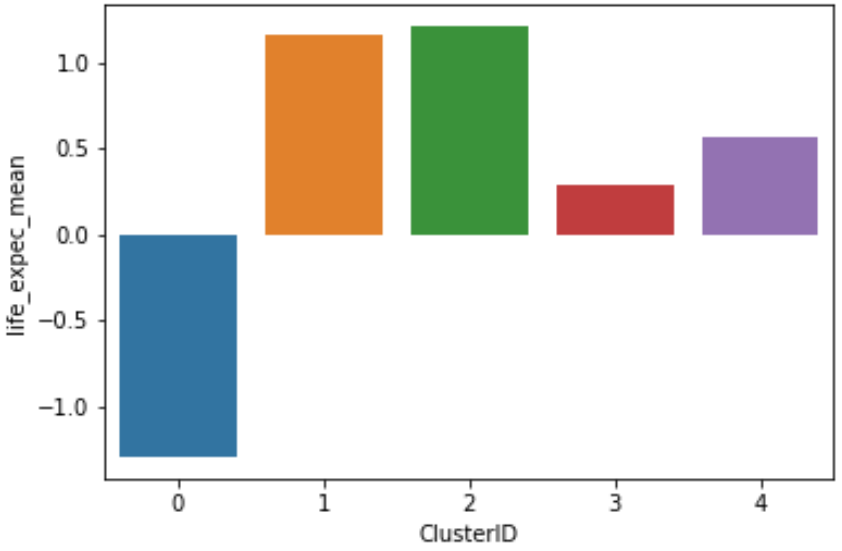




K-Means Clustering (k=5)

Cluster 0 has high child mortality rate, high fertility, low income, low health and low life expectancy. Therefore cluster 0 is the appropriate choice for the aid.

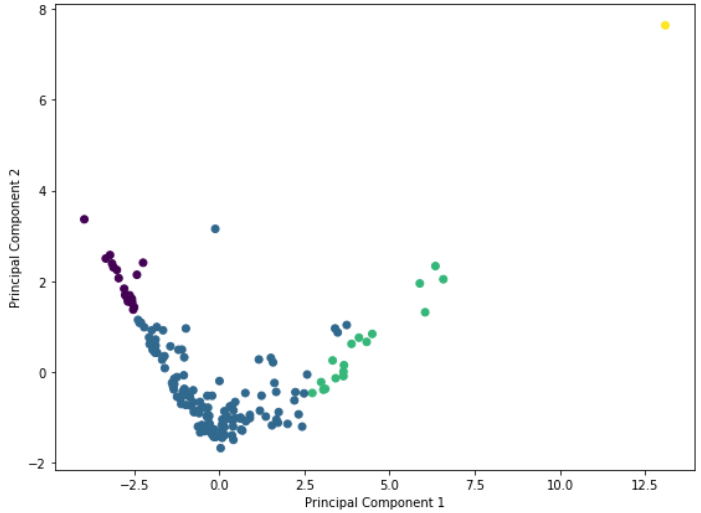


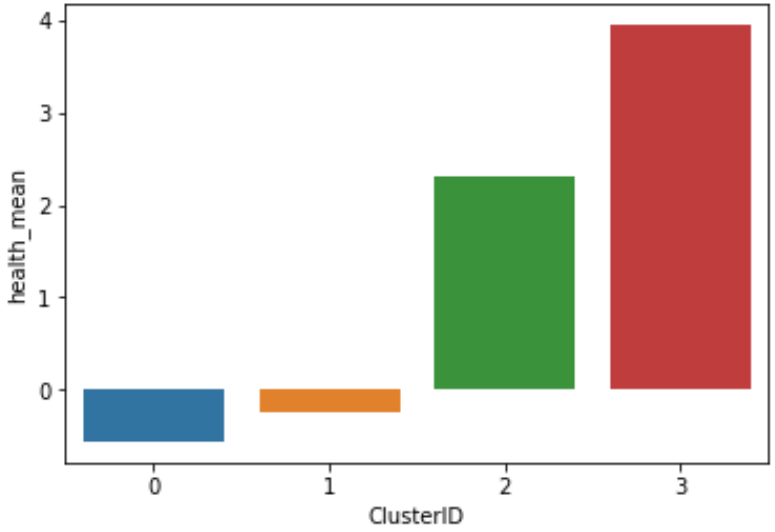


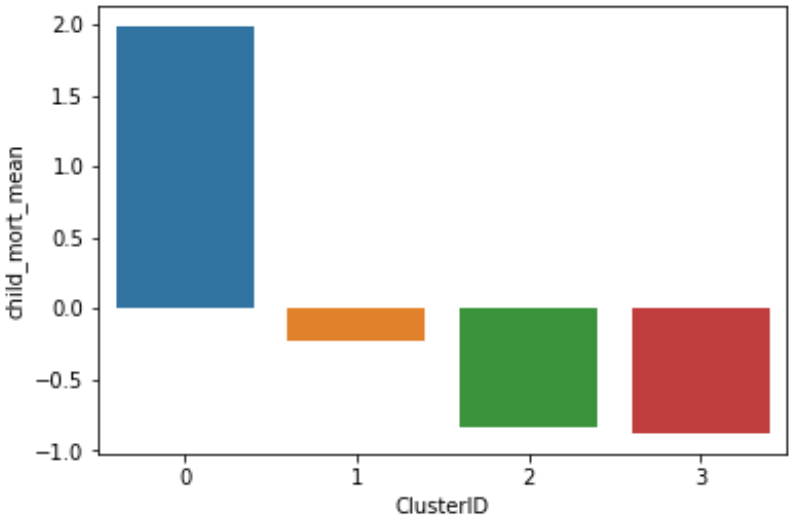
Analysis: Clustering (Hierarchichal)

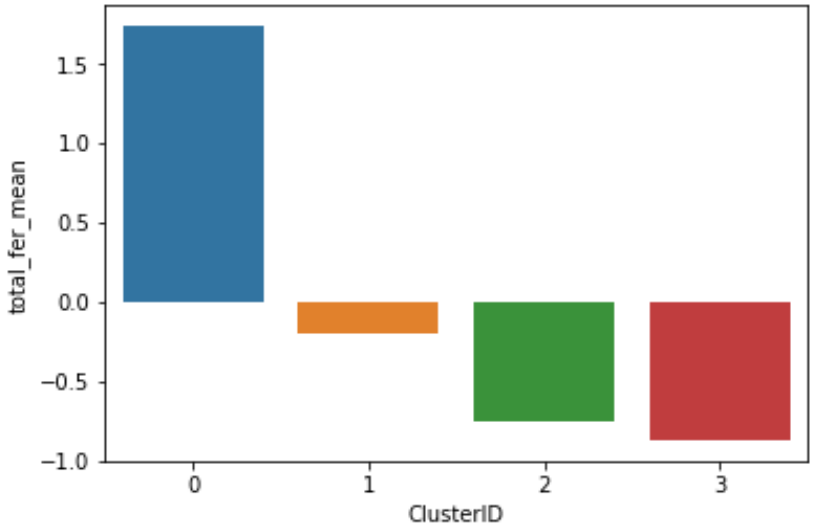
- First Dendrogram is created and it is cut at an appropriate place to obtain the no. of clusters (4)
- Now clustering is done with no of clusters as 4.
- For each Cluster Id, the mean values of all PCs and some original features are calculated and plotted.
- The plots are :

Heirarchichal Clustering



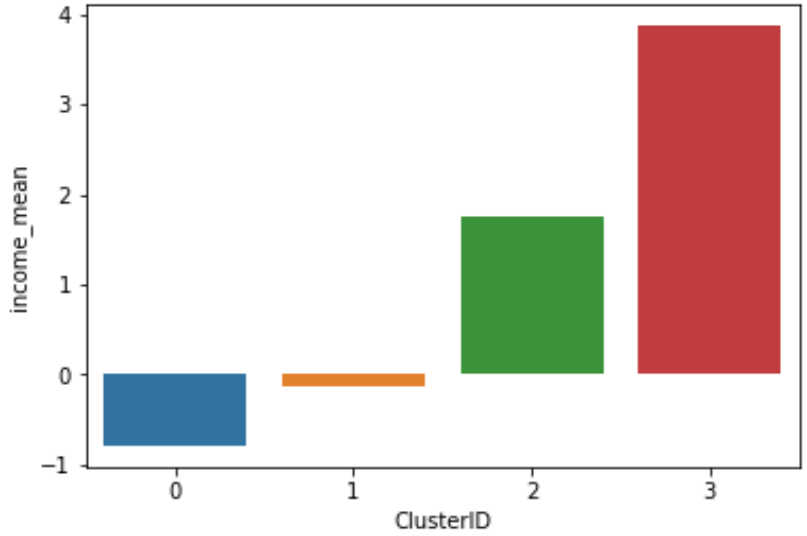


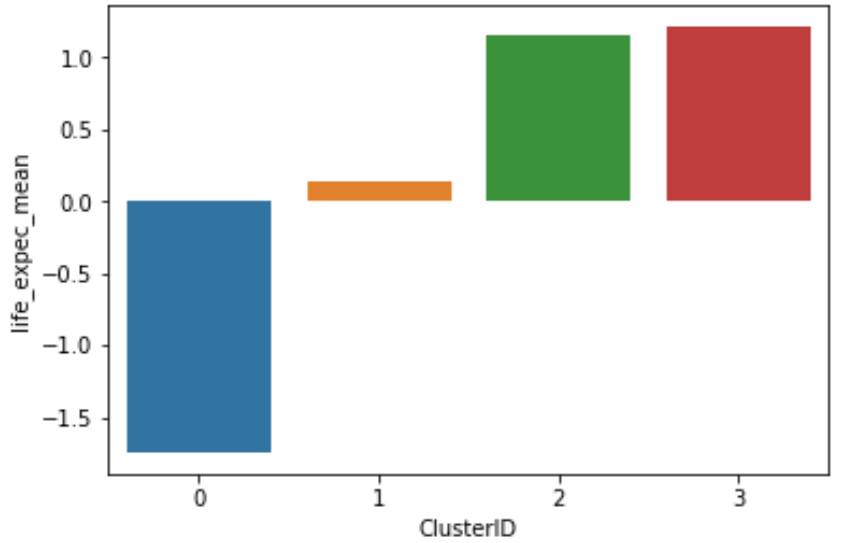




Heirarchichal Clustering

- Cluster 0 has high child mortality rate, high fertility, low income, low health and low life expectancy. Therefore cluster 0 is the appropriate choice for the aid.
- However some countries are missing in the Cluster 0 of the Heirarchichal Clustering. So we choose the Cluster obtained by the K-Means.





List of Countries which are in the direst need of aid

(The green marked countries can be put in the 2nd priority list as they are comparatively better off then the remaining countries)

1. Afghanistan
2. Angola
3. Benin
4. Botswana
5. Burkina Faso
6. Burundi
7. Cameroon
8. Central African Republic
9. Chad
10. Comoros
11. Congo Dem. Rep.
12. Congo Rep.
13. Cote dIvoire
14. Eritrea
15. Gabon
16. Gambia
17. Ghana
18. Guinea
19. Guinea-Bissau
20. Haiti
21. Kenya
22. Kiribati
23. Lao
24. Lesotho
25. Liberia
26. Madagascar
27. Malawi
28. Mali
29. Mauritania
30. Mozambique
31. Namibia
32. Niger
33. Nigeria
34. Pakistan
35. Rwanda
36. Senegal
37. Sierra Leone
38. Solomon Islands
39. South Africa
40. Sudan
41. Tanzania
42. Timor-Leste
43. Togo
44. Uganda
45. Yemen
46. Zambia