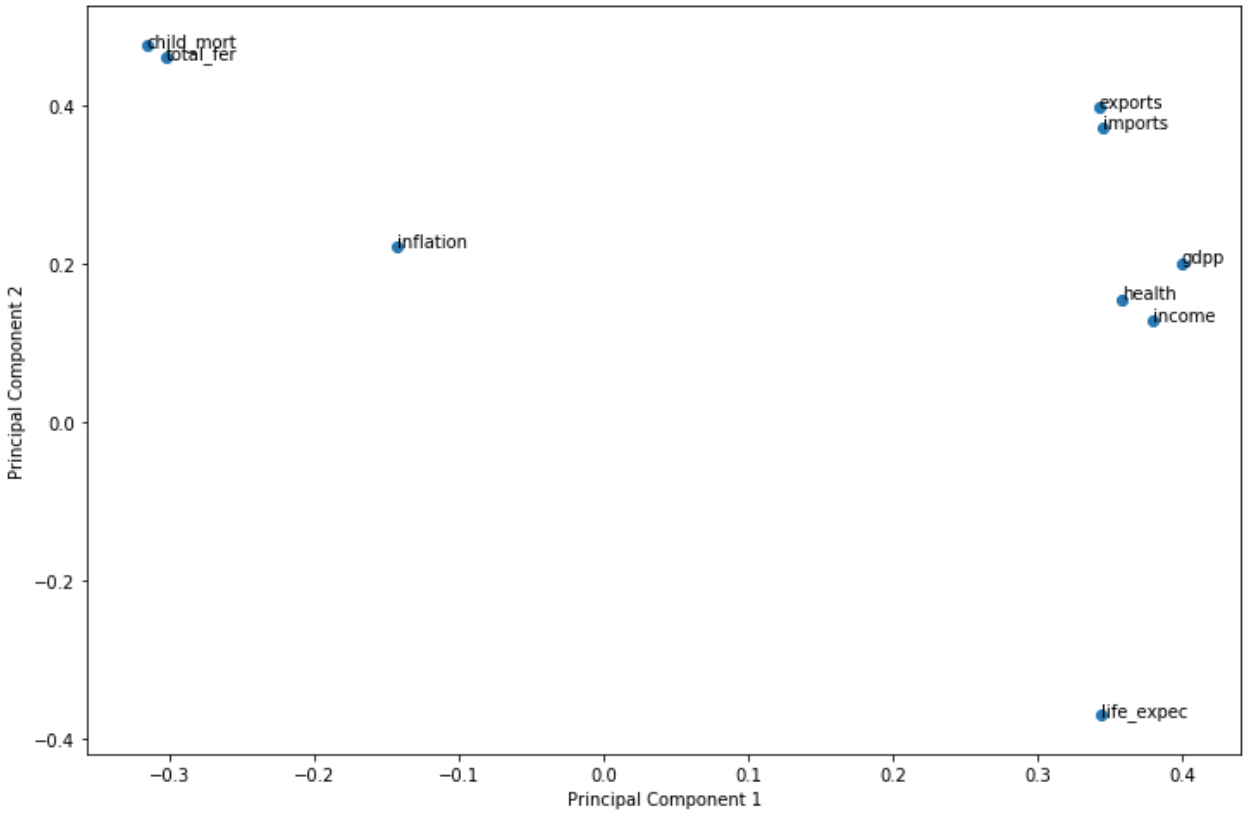
Clustering & PCA

HELP International

19 August 2019



Analysis of Socio-Economic Factors

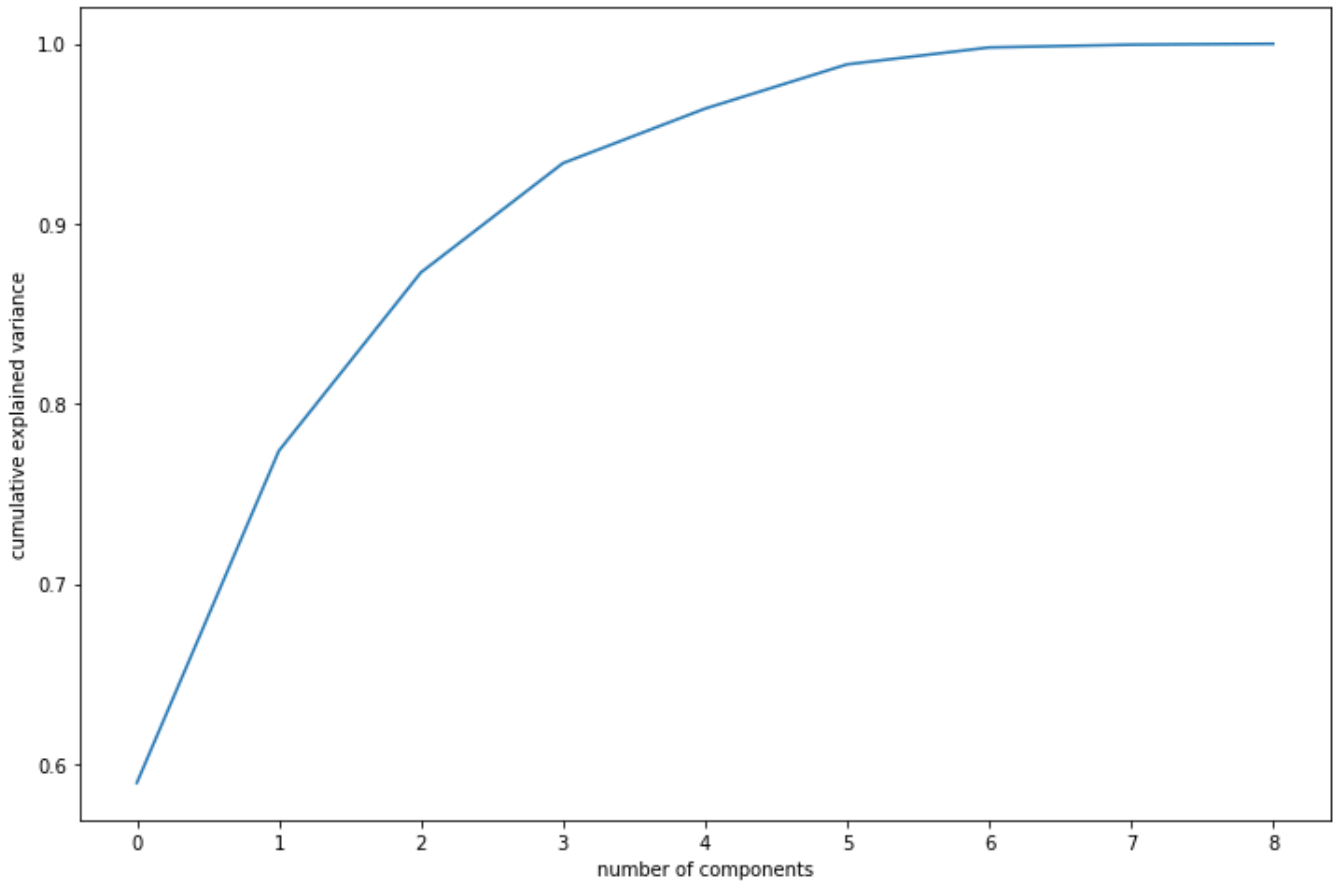
Introduction

We have to analyse the given dataset containing countries along with some of their socio-economic factors to find out the list of countries who are in direst need of aid to fight poverty.

Approach

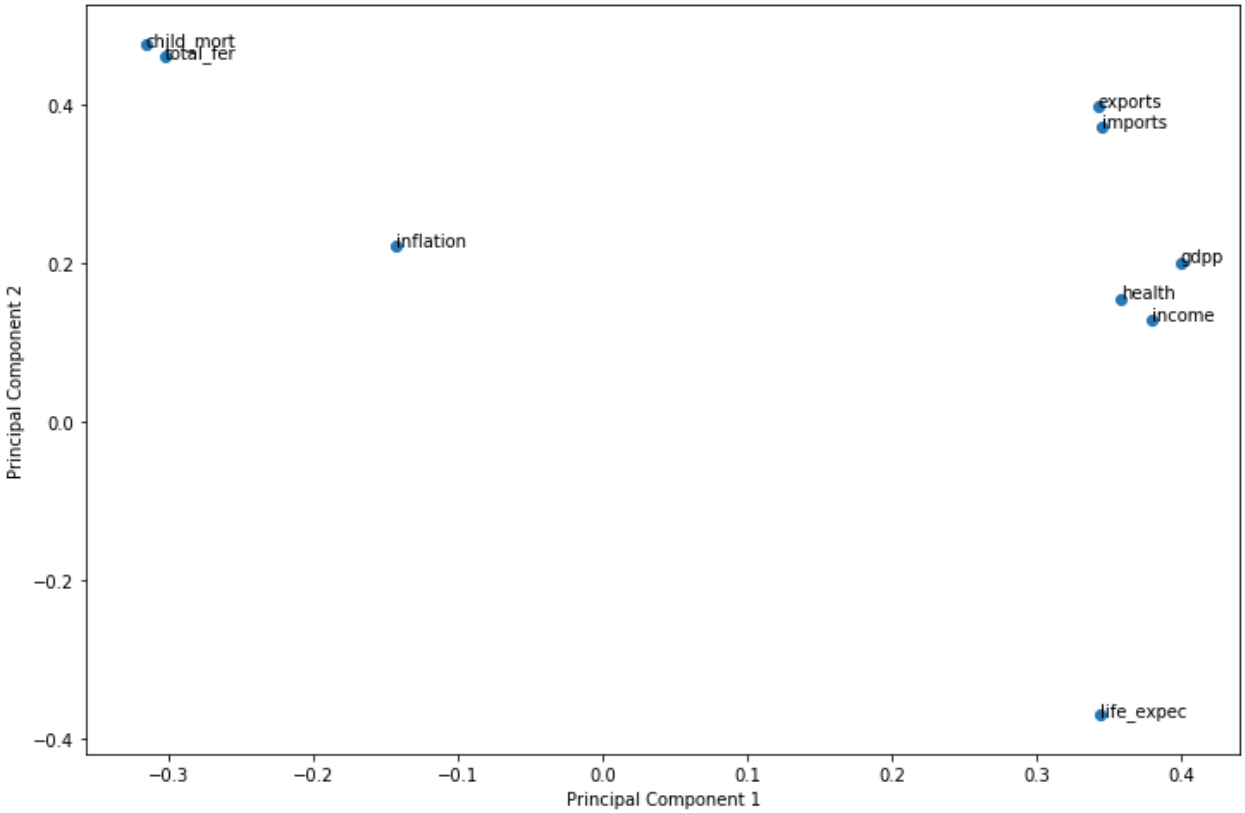
We went ahead and loaded the necessary libraries first and then the data into our jyputer notebook. As always we tried to gather some basic understanding of the data before moving into the PCA and Clustering activity.

Post gathering sufficient feel of the data, we started with the PCA and plotted the scree plot to decide the number of PCs we'll be using in our analysis.



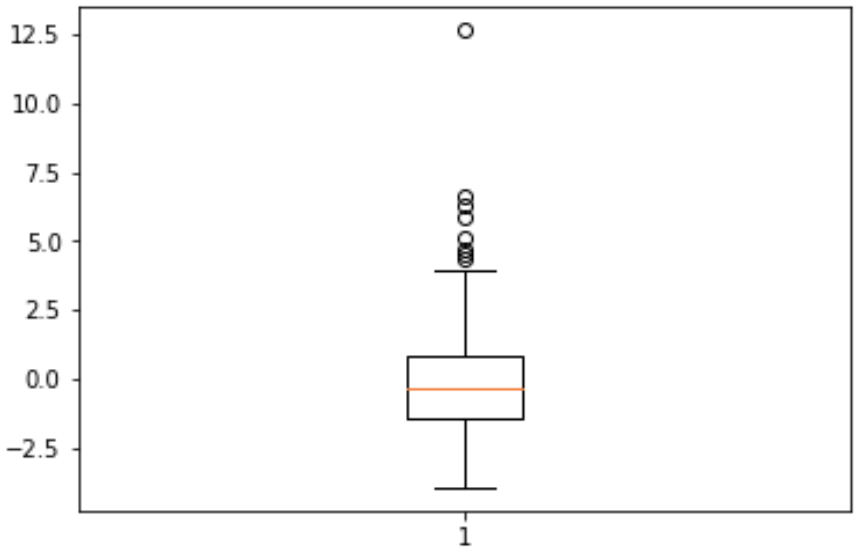
As checked two PCs were able to capture somewhere around 88% of variance. Hence we decide to use 2 PCs for our analysis.

Post this we analysed the socio-economic factors by plotting them with help of PCs.

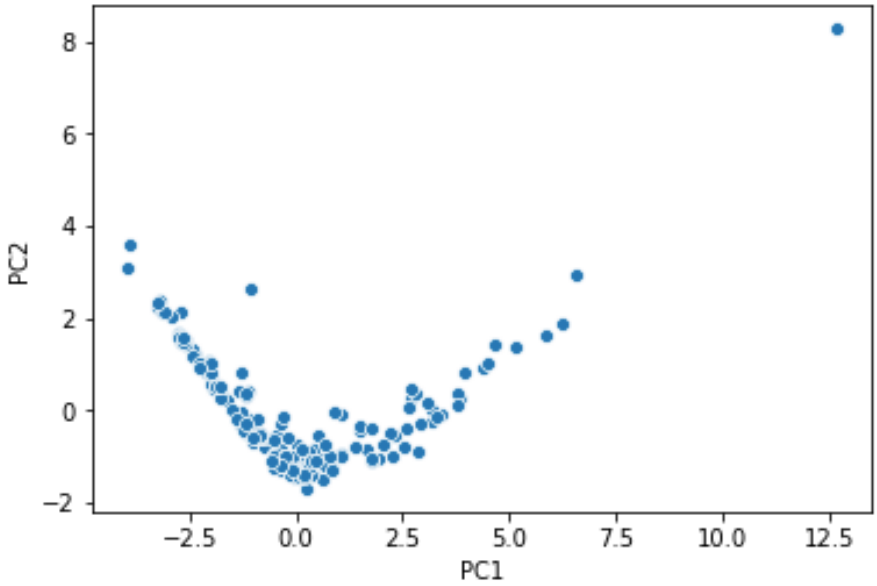


As we can see exports, imports gdpp, health and income are some of the positive factors whereas child_mort is one of the negative factors.

Then we checked for outliers in our PCs. We decided to keep the outliers, since they were in groups.

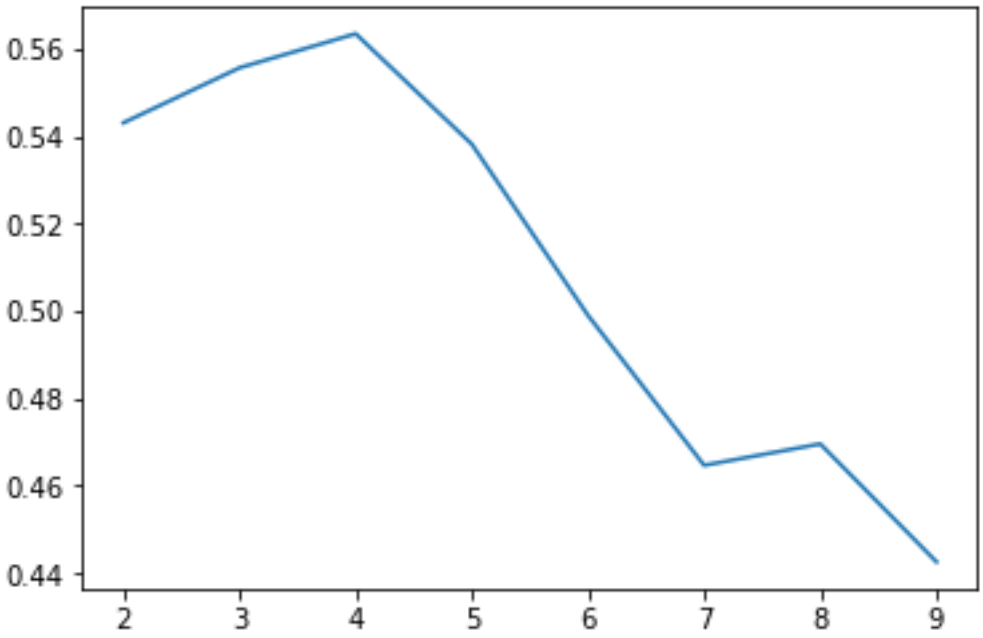


Before moving on to clustering we decided to visualise the PCs.



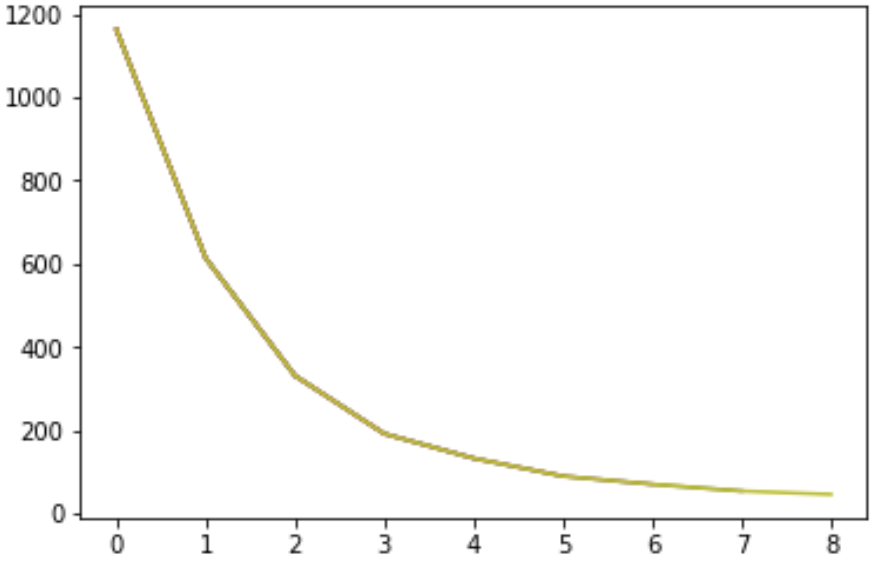
Then we check Hopkins statistic which was 0.97, definitely greater than 0.5 and it meant that clustering made sense.

Now we tried silhouette curve to check for optimal 'k'.



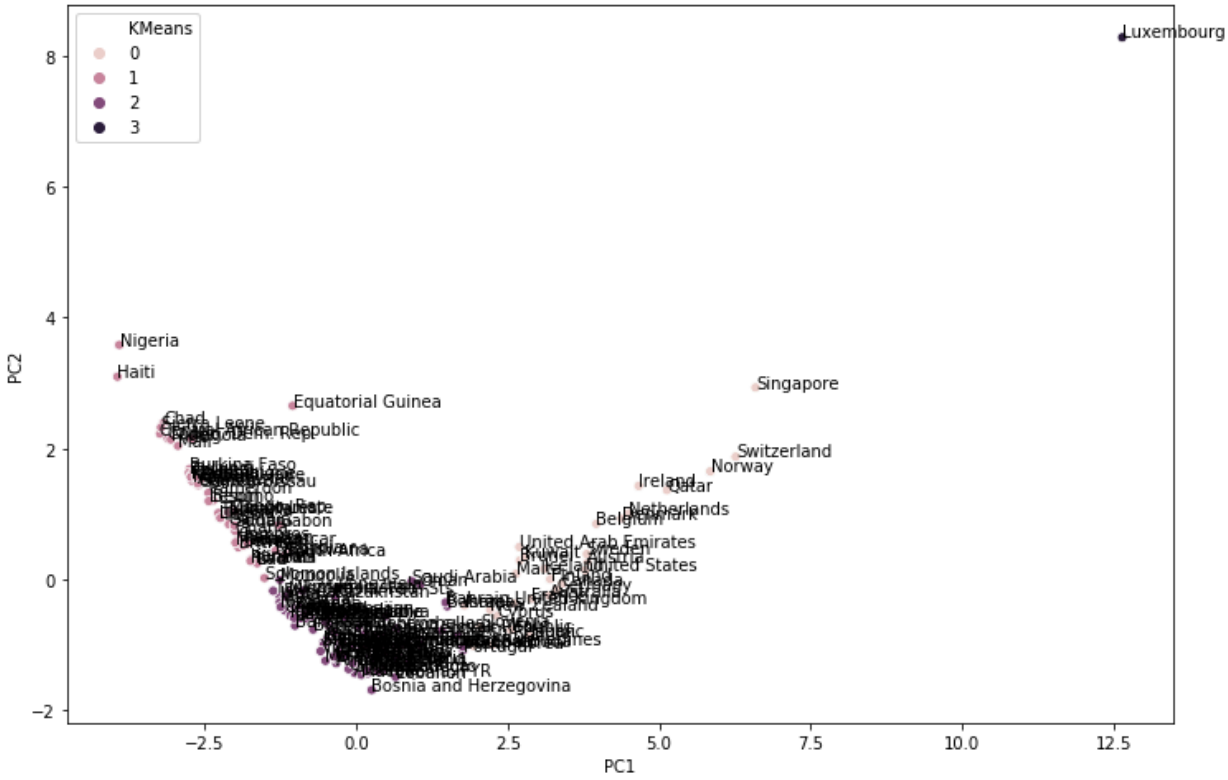
At cluster-4 the score was at highest in silhouette score.

Then we tried elbow curve to verify the ‘k’.

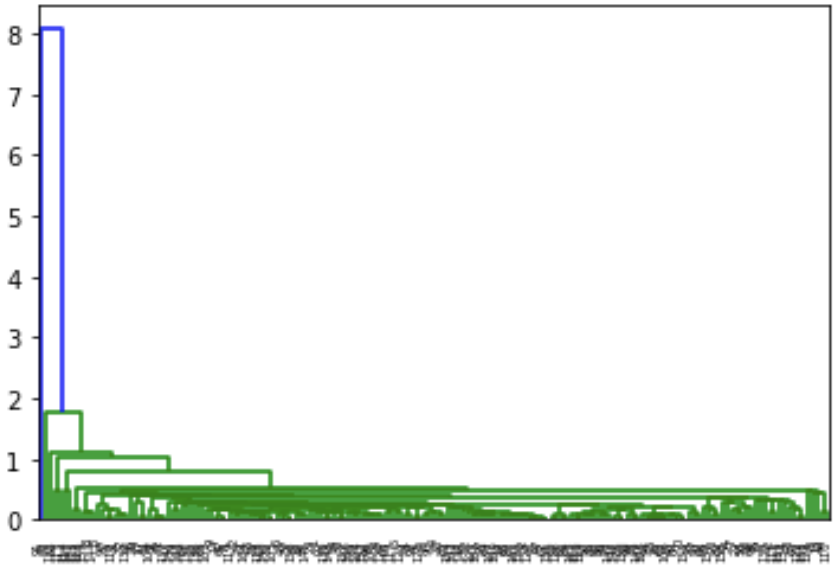


Over here too 4th cluster made sense as the sum of squared errors was at low while keeping the number of clusters at minimum.

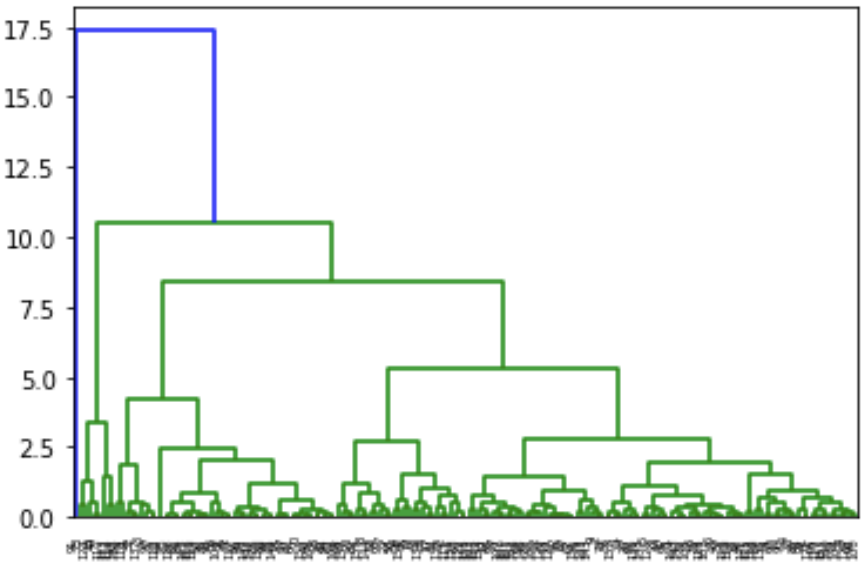
Then we used K-means algorithm on our scaled PCs, below is the representation.



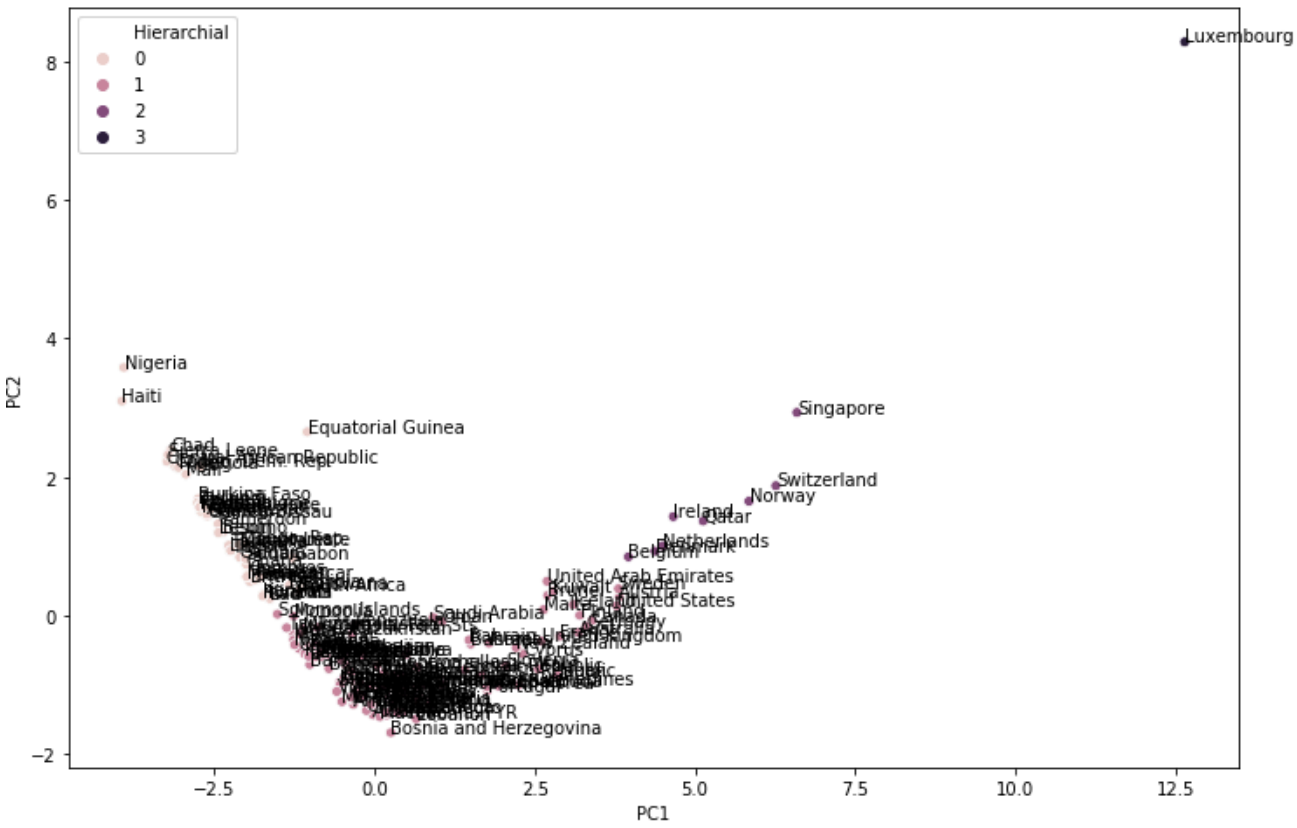
Post K-means we tried Hierarchical clustering too. At first, we choose single linkage to begin with.



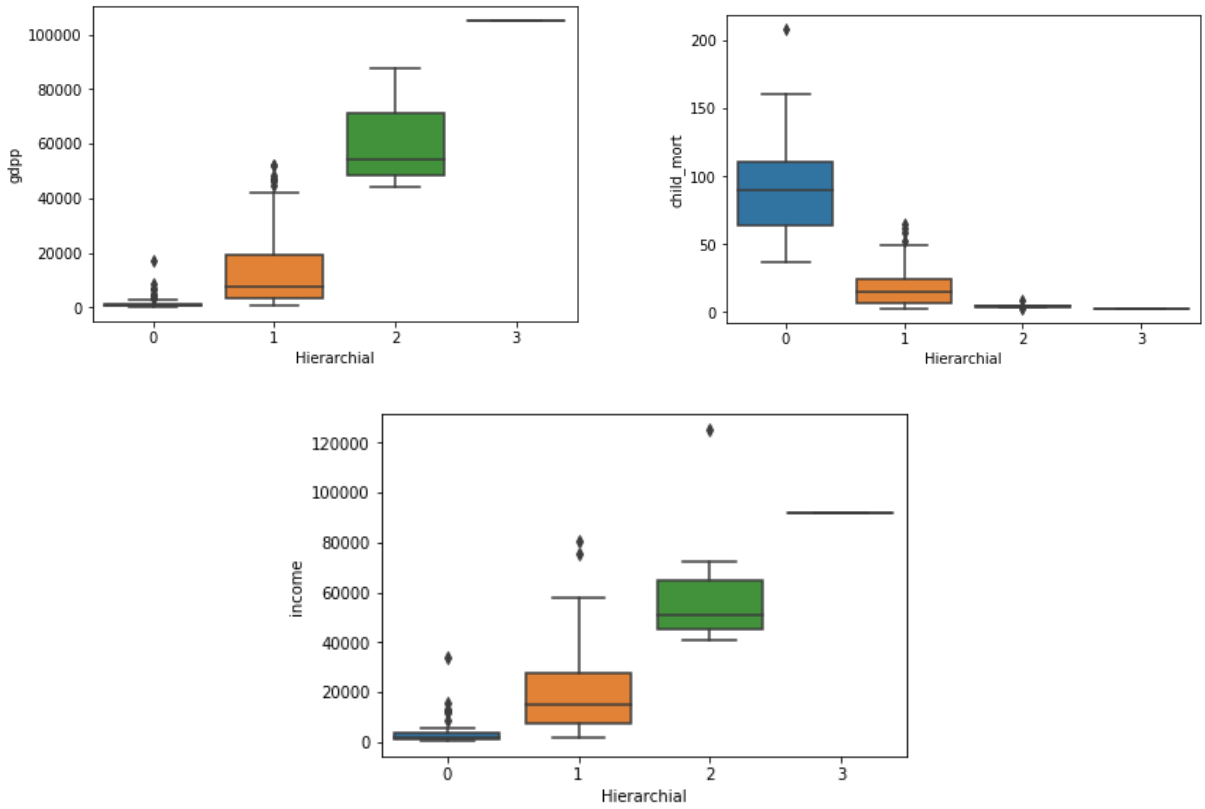
As evident from the dendrogram we then tried complete linkage as single linkage was not clear at all to decide on any number of clusters.



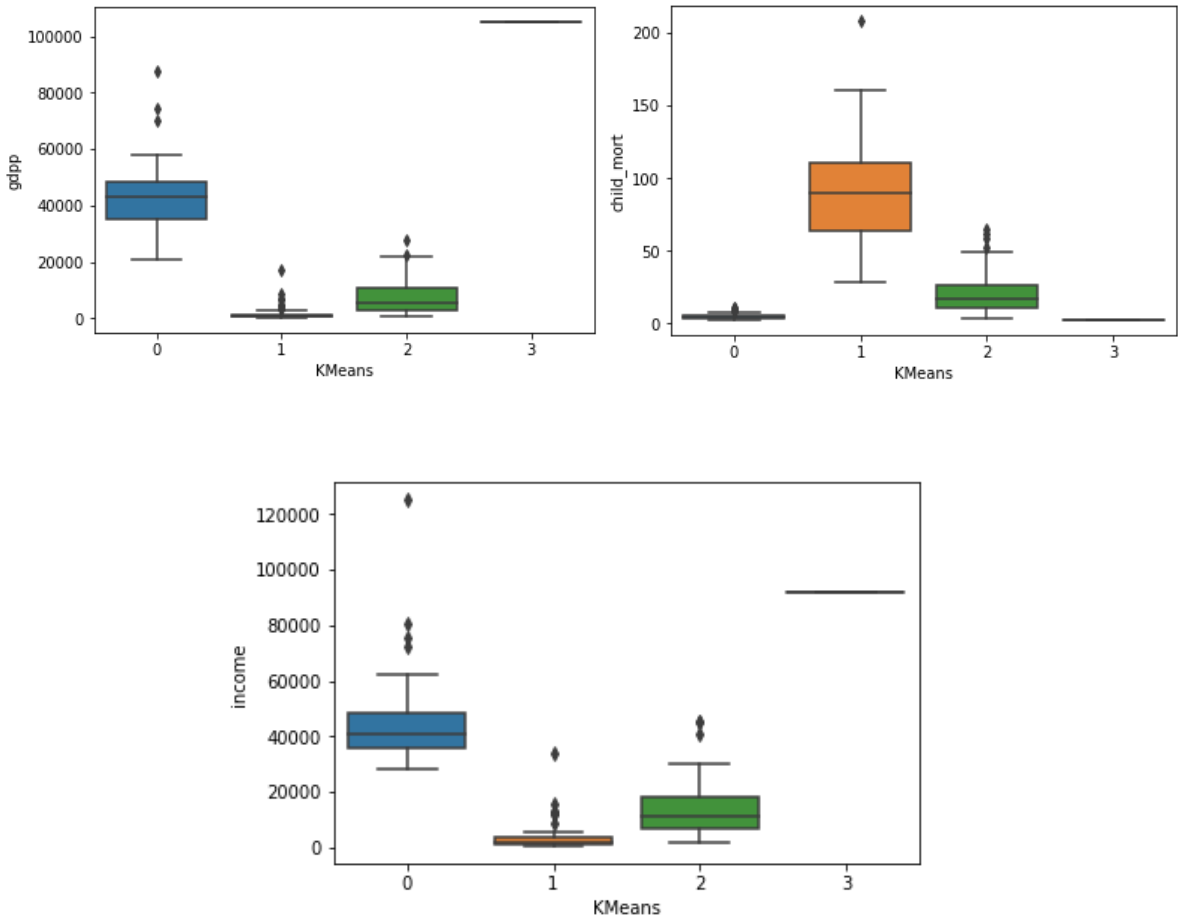
As we had thought, complete linkage was a clear winner here. We were able to visualise the required number of clusters and start off with our analysis. However when we plotted a scatter plot using Hierarchical clusters the results were the same as of K-means.



Then we plotted few box plots of three socio-economic factors namely - child_mort, income and gdpp using Hierarchical clustering.



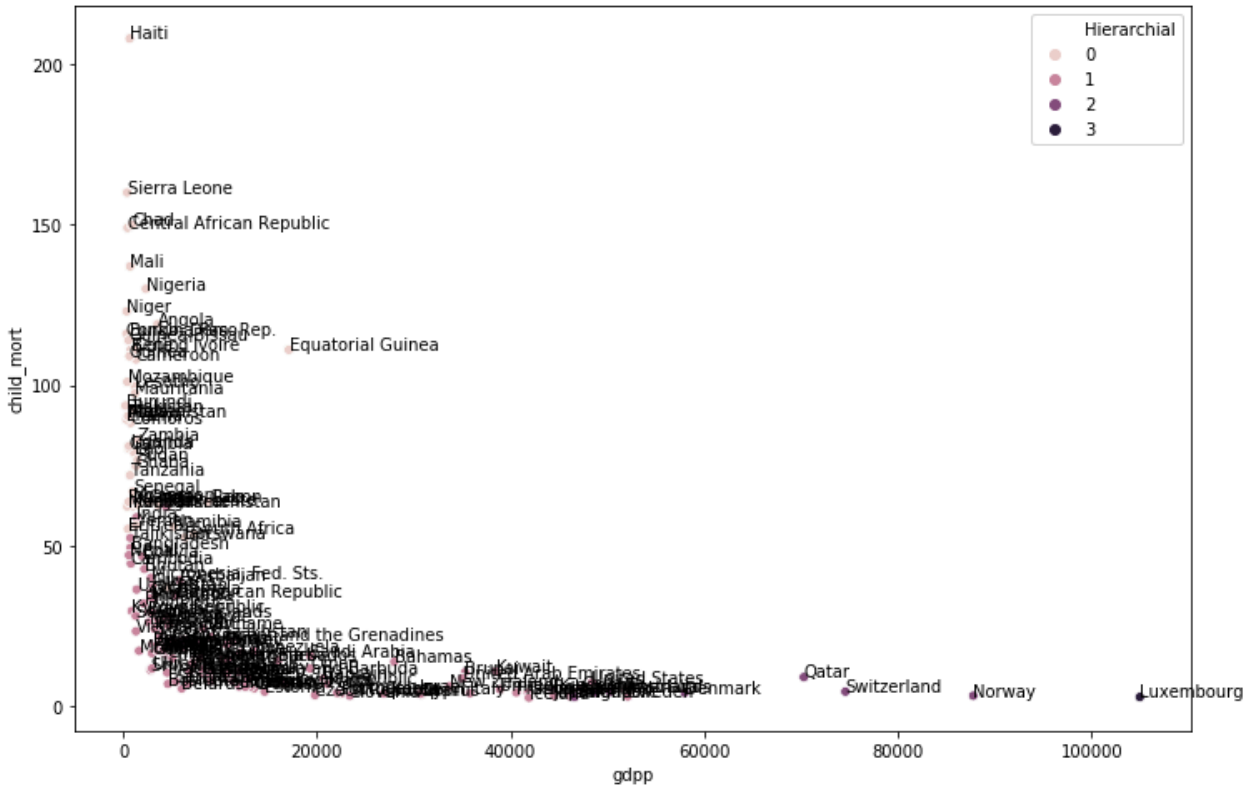
We did the same analysis of factors using K-means clustering.

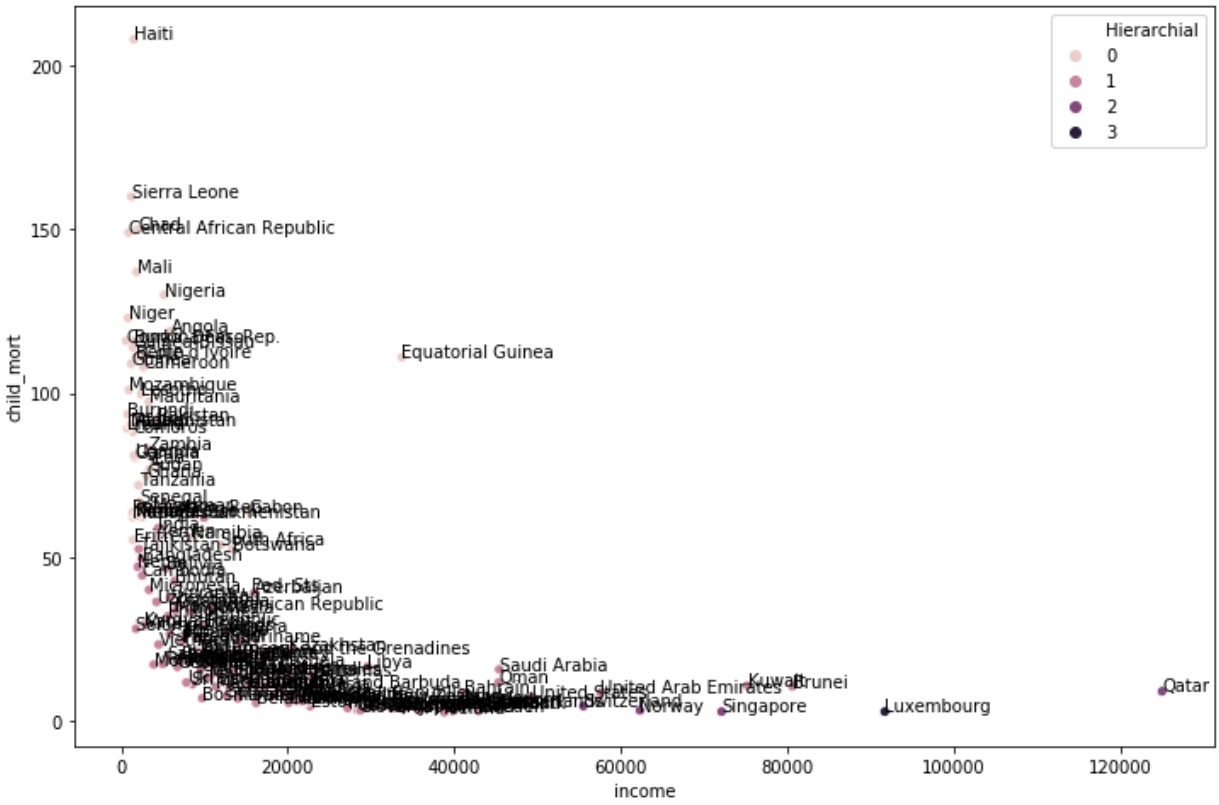


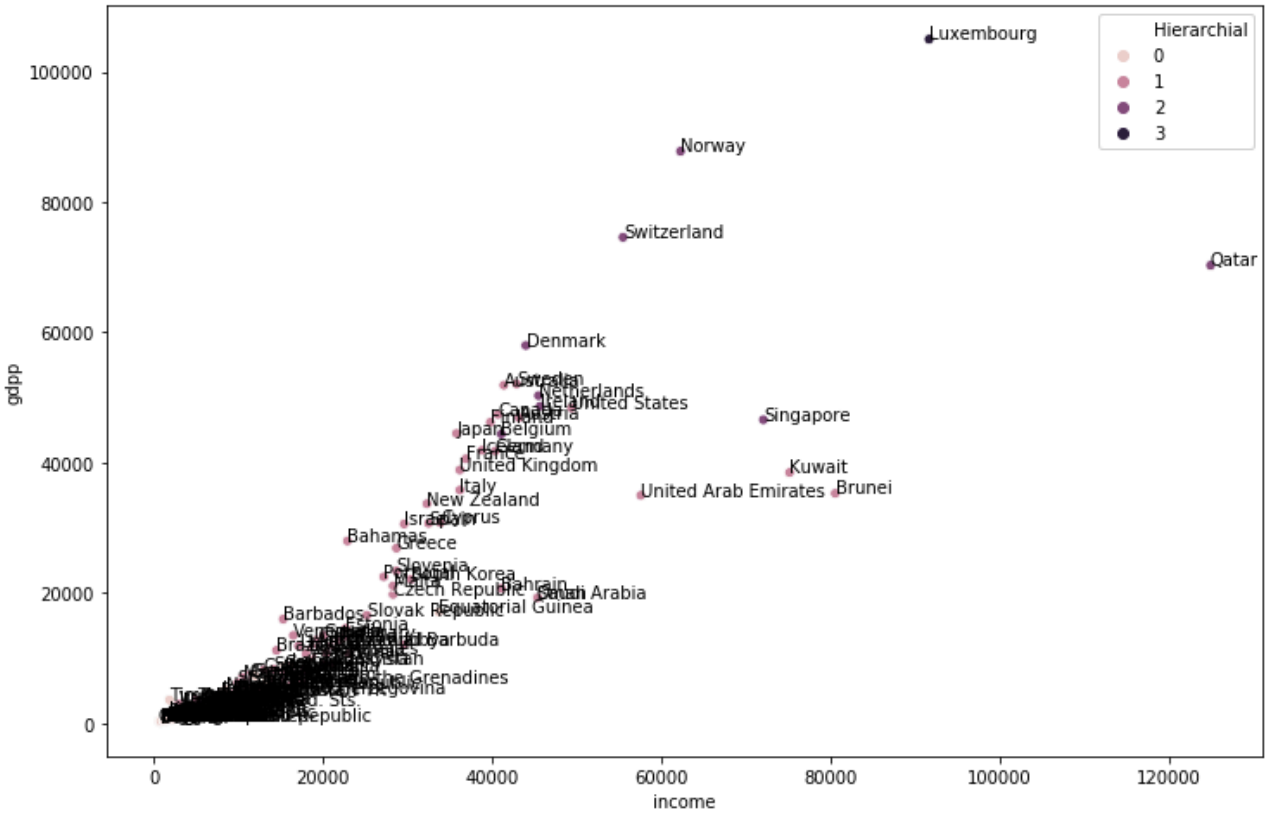
However the results were a little bit difficult to explain as the clusters in K-means did not follow an incremental approach as opposed to Hierarchical clustering.

Then we plotted few representations of countries on how they are fairing among our previously selected socio-economic factors and the results were same in both the clustering methods. However please find the final list of countries who are in direst need of aid mentioned below.

Afghanistan, Bangladesh, Cambodia, Comoros, Eritrea, Gambia, Ghana, Kenya, Kiribati, Kyrgyz Republic, Liberia, Madagascar, Malawi, Moldova, Myanmar, Nepal, Rwanda, Senegal, Solomon Islands, Sudan, Tajikistan, Tanzania, Togo, Uganda and Zambia.


Hierarchial
0
1
2
3
Haiti
Sierra Leone
Central African Republic
Mali
Nigeria
Niger
Equatorial Guinea
Qatar
Switzerland
Norway
Luxembourg
child_mort
gdpp
0
50
100
150
200
0
20000
40000
60000
80000
100000


Hierarchial
0
1
2
3
Haiti
Sierra Leone
Central African Republic
Mali
Nigeria
Niger
Equatorial Guinea
Saudi Arabia
Qatar
Luxembourg
Singapore
child_mort
income
0
50
100
150
200
0
20000
40000
60000
80000
100000
120000


Hierarchial
0
1
2
3
Luxembourg
Norway
Switzerland
Qatar
Denmark
Singapore
Kuwait
United Arab Emirates
Brunei
United Kingdom
New Zealand
Bahamas
Greece
Barbados
Slovak Republic
Equatorial Guinea
Grenadines
gdpp
income
0
20000
40000
60000
80000
100000
120000