Clustering & PCA Assignment

Problem statement:

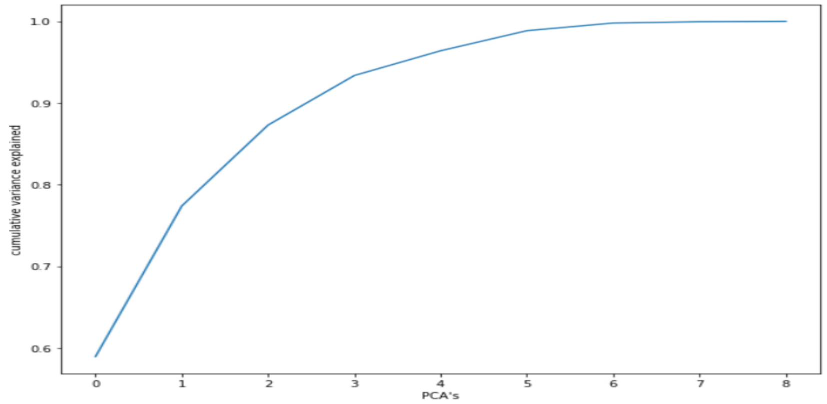
International humanitarian NGO that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities.

After the recent 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.

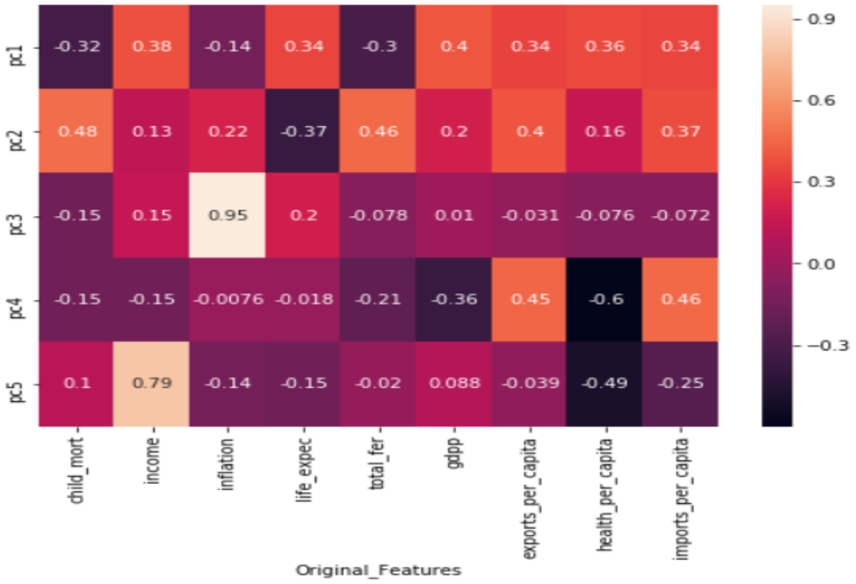
So we need to provide names of top 5 countries which are in the direst need of aid.

Analysis approach:

1. Reading the data
2. Data cleaning if any quality issues like spelling mistakes in countries, etc are there in data but couldn't find any quality issues
3. Converting the values which are in percentage to their actual values like exports, health, imports
4. Data preparation like scaling to perform PCA latter, choosing standard scaler since PCA assumes mean to be centralised i.e. 0
5. Performing PCA
6. Choosing number of PCA which explains most of the variance using siere plot in this case we chose 5 principal components out of 8 which explained about 96% of variance

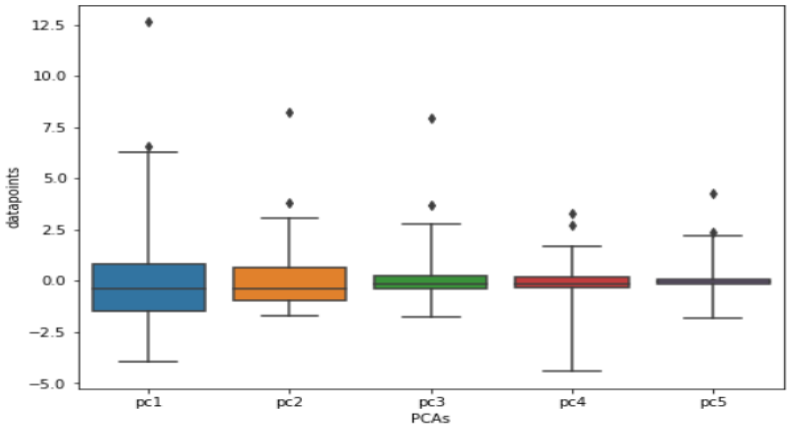
   by observing above plot if we choose 5 as number of components we get around 96% variance which is pretty good and after that it does increase much
7. Observing how principal components are located in original axis(original features)



Observations from heatmap

- theme1(pc1) explains all features except inflation
- theme2(pc2) explains child_mort,life_expect,total_fer,exports_percapita,imports_percapita
- theme3(pc3) expalins inflation
- theme4(pc4) explains gdpp,exports_percapita,imports_percapita,health_percapita
- theme5(pc5) explains income and health

8. Now representing all data points in our new axis i.e. PCA's (performing basis transformation)
9. Checking are there any outliers by ploting box plot for our new axis

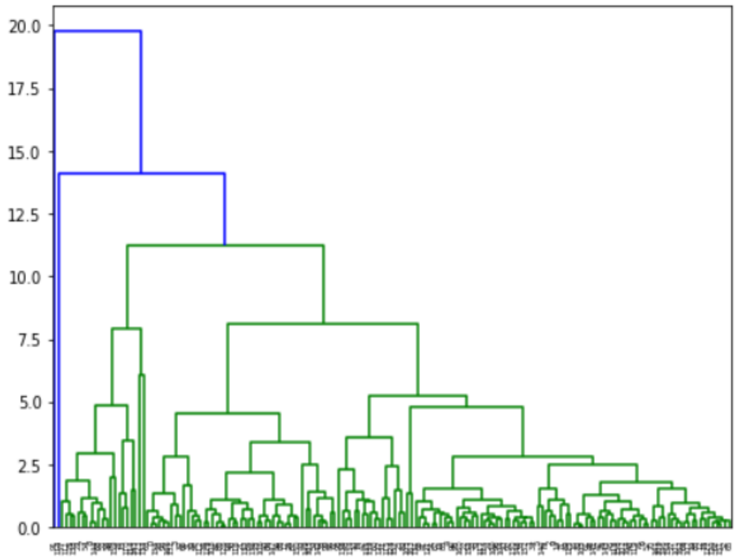


There are some outliers as we can see in our dataset but we are not sure so keeping them for further analysis.

10. Checking whether Clustering is possible for data points of new axis i.e. pca's using Hopkins statistics, we got 0.87 which is greater than 0.5 therefore there is chance of forming clusters in our dataset
11. 1st applying hierarchical clustering because this type of clustering doesn't prerequire number of clusters so we can decide number of clusters after observing dendogram and decide how many clusters we can get based on
    a. Population inside each cluster, more the population better so that it is easy for business to take decision
    b. Clusters must be far away from each other

    After choosing number of clusters in hierarchical clustering we will get an idea of number of clusters for K-means clustering which prerequire number of clusters

    dendogram for hierarchical clustering :



Observations from above dendogram

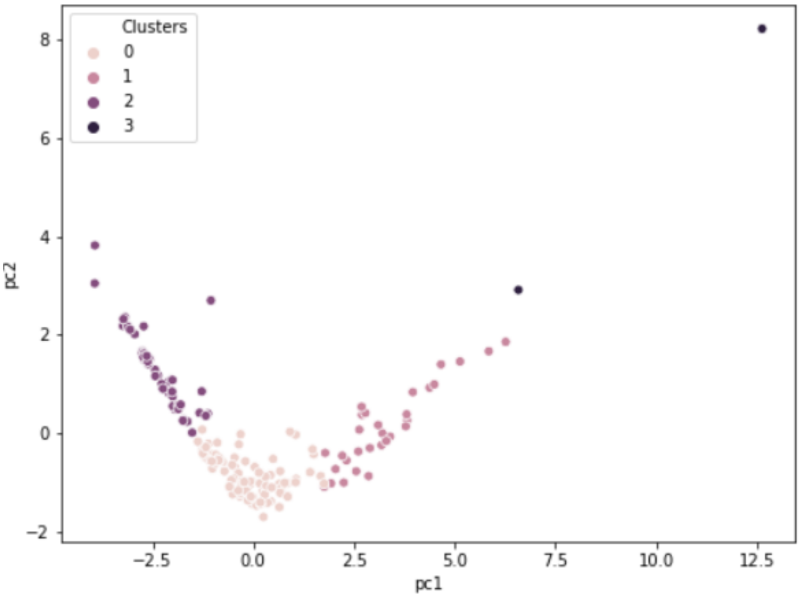
- if choose 4 clusters the distance between clusters is significant and also we will good number of clusters
- and also we need to check for good number of population within each clusters because if only few countries stay wthin cluster then it becomes difficult for CEO to take seperate decision for only those few countries
- if we choose 6 clusters some clusters will be having very less populated so going with 4

12. observing the population of clusters:

```
cluster 0    144
cluster 1     21
cluster 3      1
cluster 2      1
```

    we can see that cluster 2 and 3 contains very less population so we can ignore those and analyse them seperately
13. Performing K-means clustering:

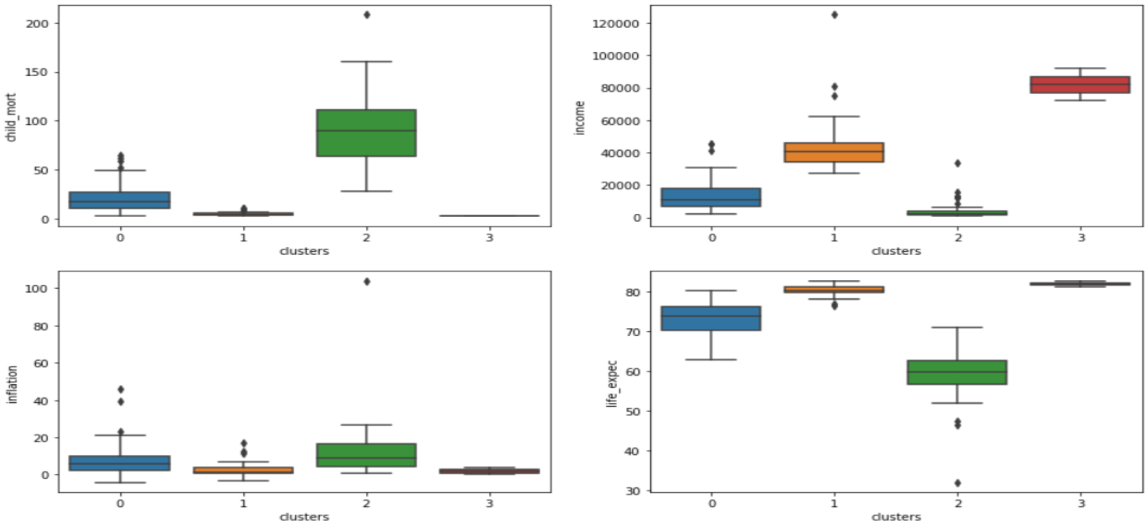
- Finding the number of clusters using elbow curve and silhoute analysis and also we got idea of number of clusters to be chosen from hierarchical clustering we found that 4 clusters are good if chosen in business point of view so choosing the same
- Performing K-means clustering with number of clusters 4 and visualising the clusters on 1st 2 PCA's since they capture maximum variance

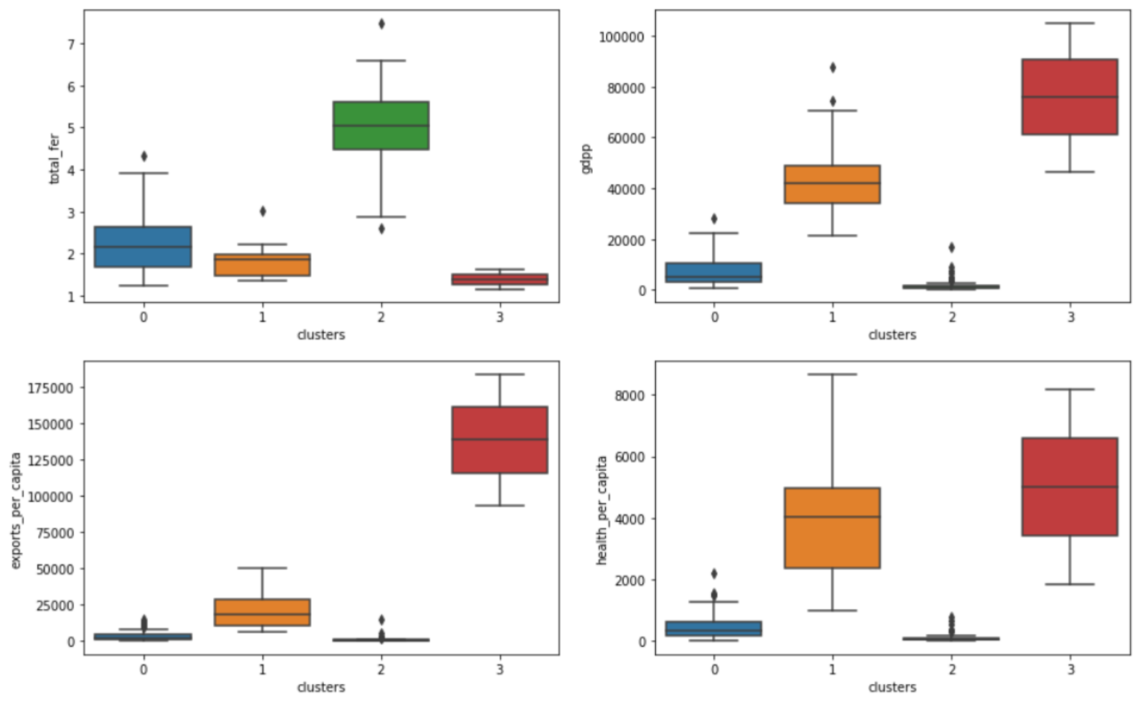


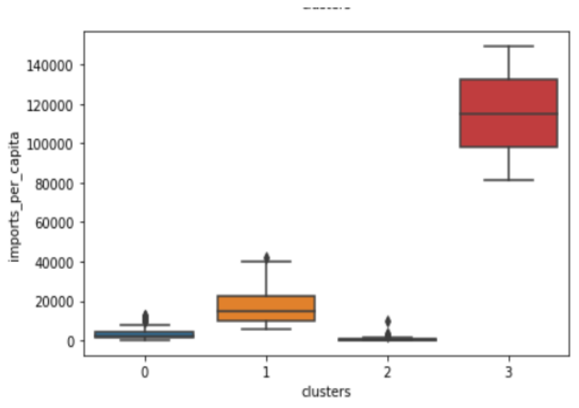
14. we can choose kmeans clustering for proceeding with analysis since it gives clusters with good population except 3 cluster which contains only 2 countries and also clusters seems intuitive from the above plot which seems better than hierarchical clustering
    population of each clusters

```
cluster0    87
cluster2    48
cluster1    30
cluster3     2
```

15. Now analysing clusters on original data set







Observations from the above plots

- cluster 1 and 3 are developed countries because they have high gdpp,high income and low child_mort
- cluster 2 is developing because they have average gdpp,average income and average child_mort
- cluster 0 is under developed because they have low gdpp,low income and high child_mort

16. Deciding clusters whether are developed, under developed, developing based on gdpp, income, child_mort and inflation, Using binning technique
We get these countries which are in direst need of help:

    1. Burund
    2. Congo,Dem. Rep
    3. Guinea,

4. Mauritania
5. Sierra Leone

These countries have high child_mort, low net income per person, low gdpp and high inflation