PCA AND CLUSTERING CASE STUDY

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.

GOAL-

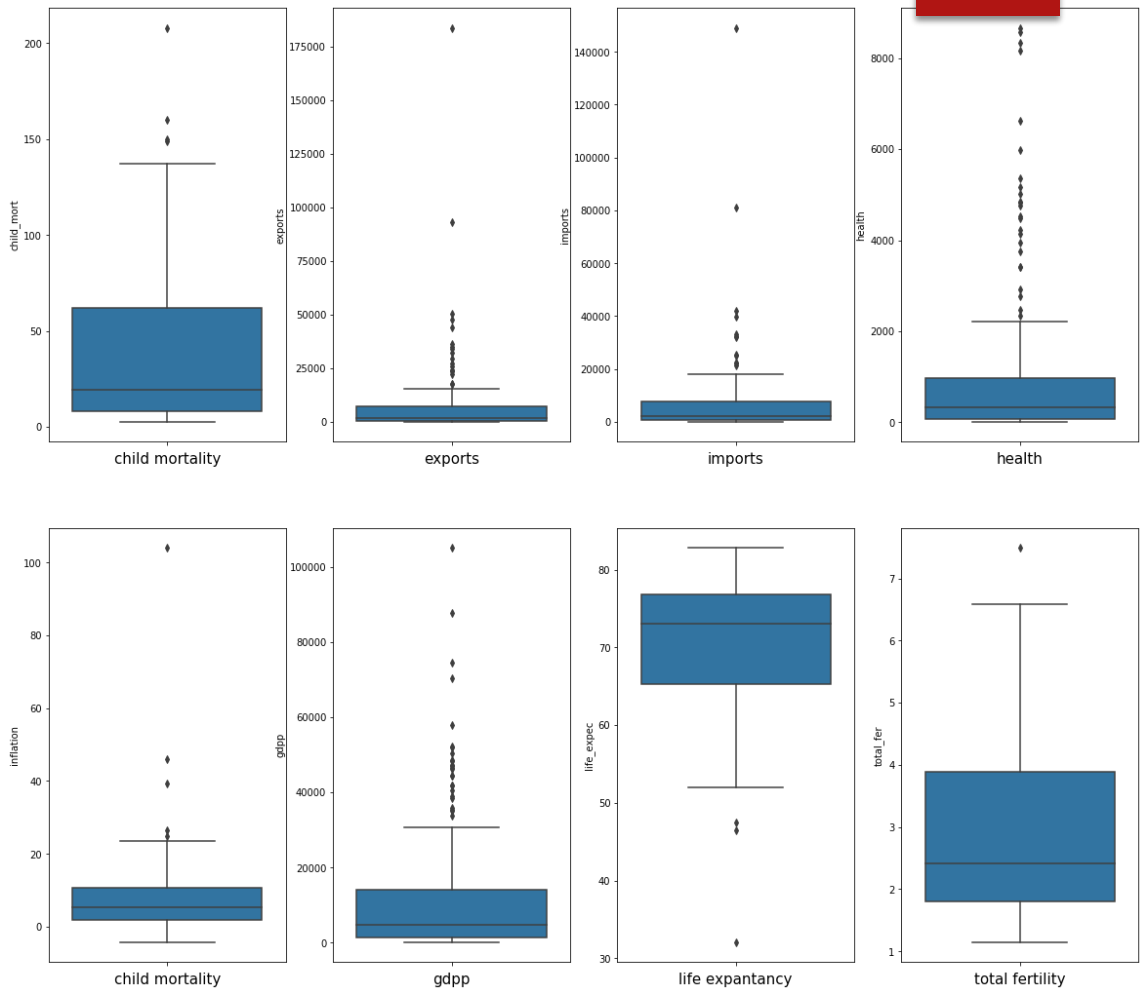
TO FIND OUT COUNTRIES WHICH ARE IN NEED OF HELP.

Approach for analysis.

1)Data Sourcing.
2)Performing EDA analysis.
3)Removed the outliers.
4)Scaled the features.
5)Performed PCA.
6)Implemented K-Means clustering.
7)implemnented hierarchical clustering.

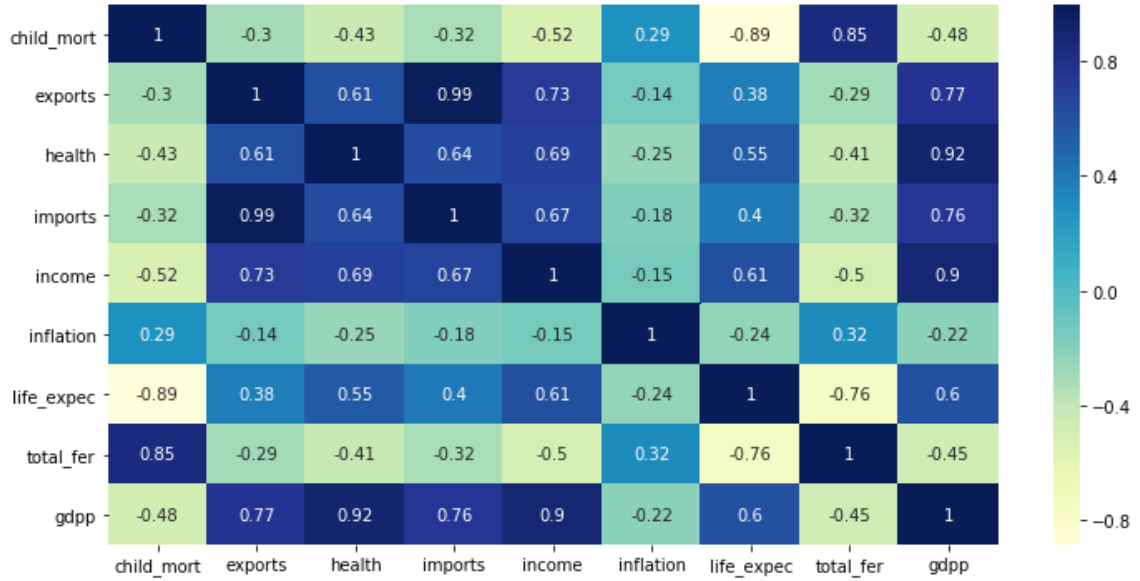
UNIVARIATE ANALYSIS- OUTLIERS ANALYSIS.

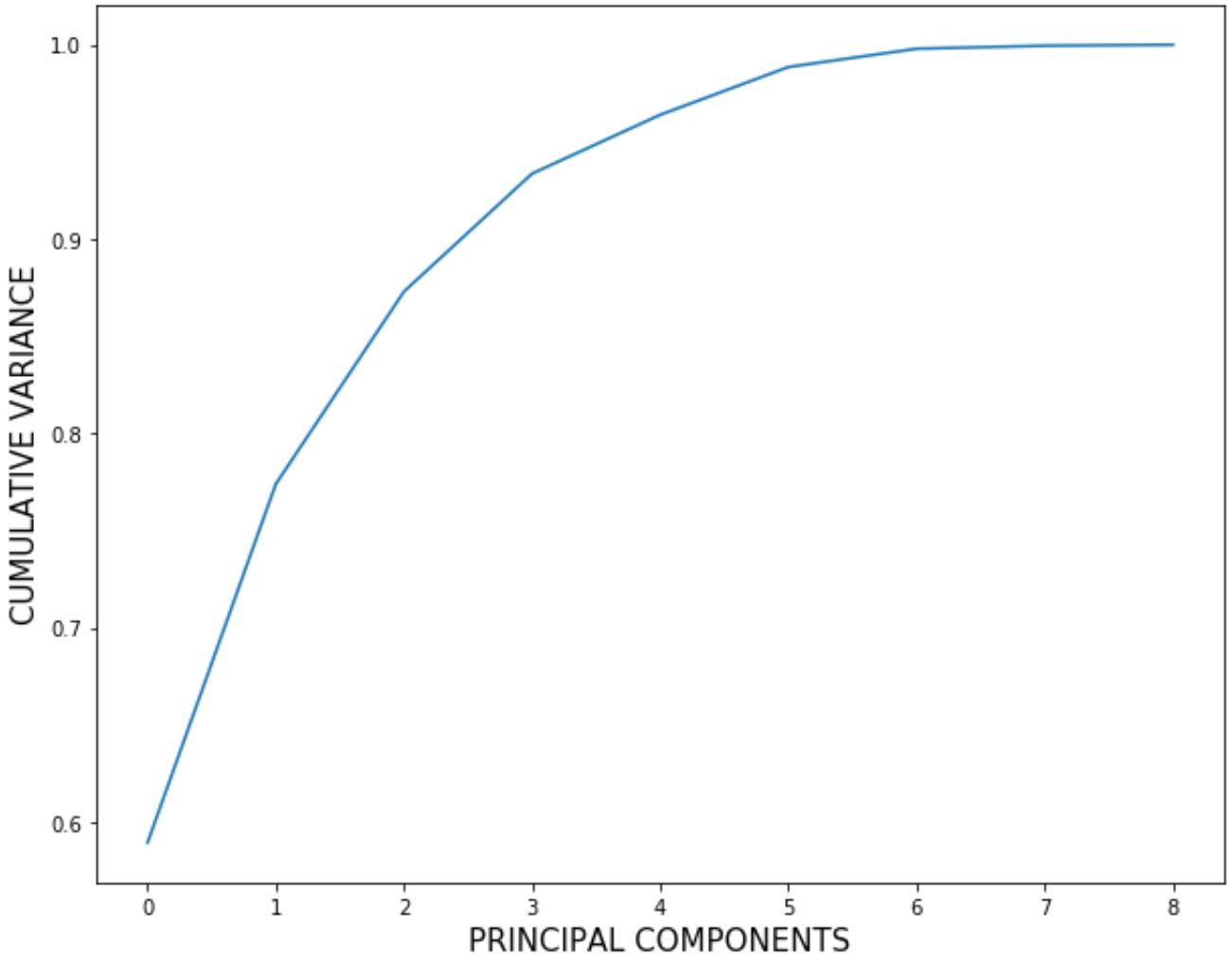
- There were some outliers in the features which were dealt in the analysis.



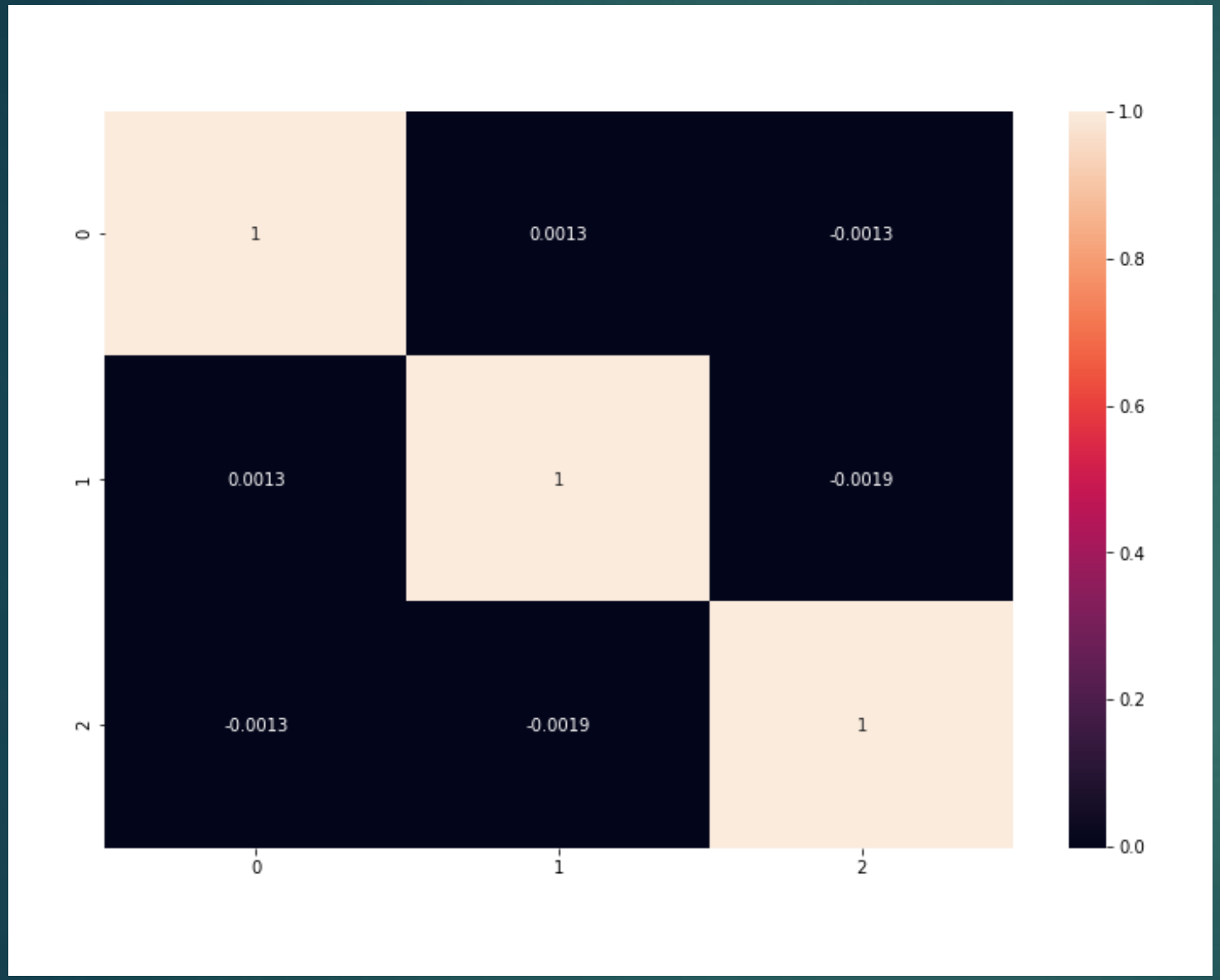
Correlation analysis

- The feature were correlated to each other ,hence by doing PCA ,we can reduce the dimensionality and make our analysis smooth.





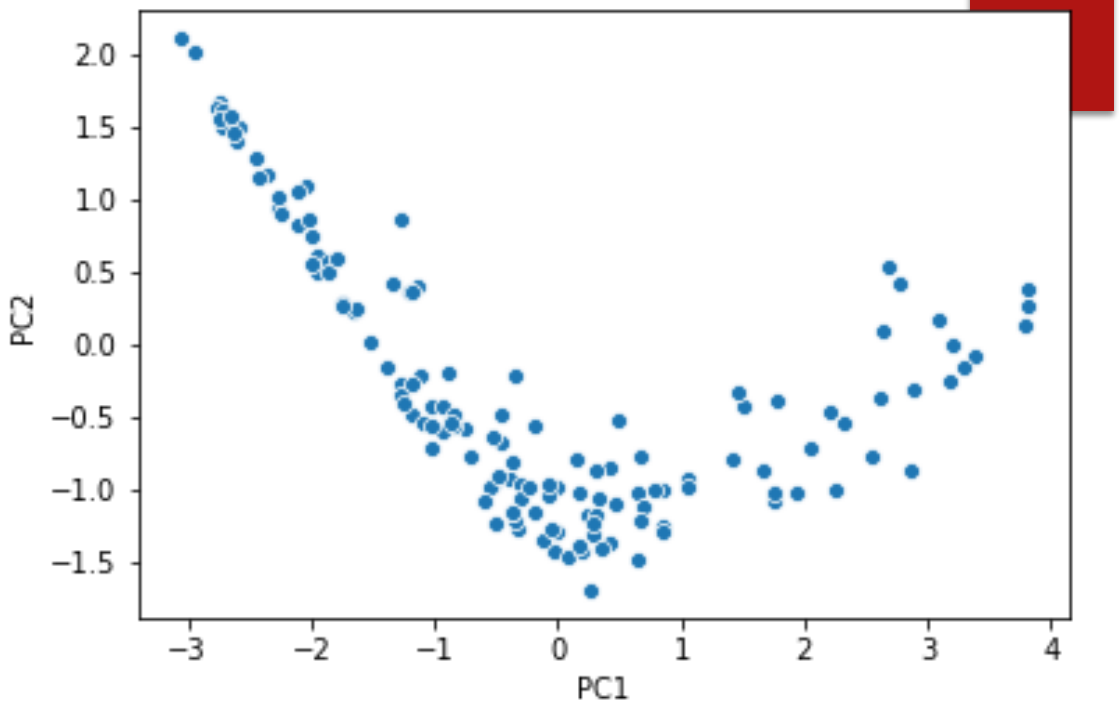
The scree plot—
It can be seen from scree plot that about 88% of variance can be explained by 3 Principal components. Hence 3 PCs were chosen.

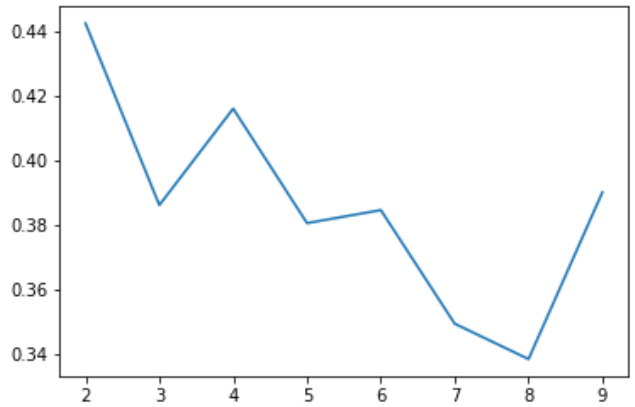


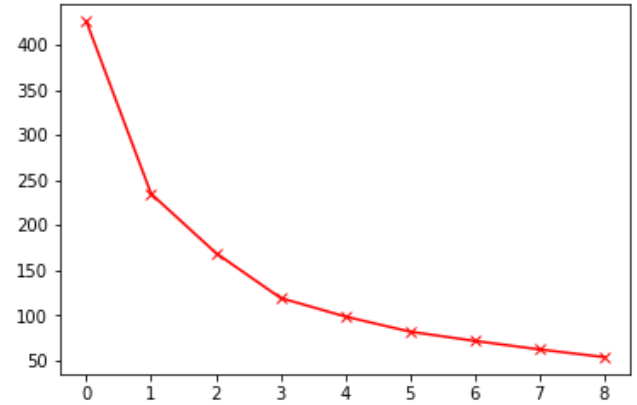
There is no correlation between any 2 components and it can be verified that PCs are orthogonal to each other.

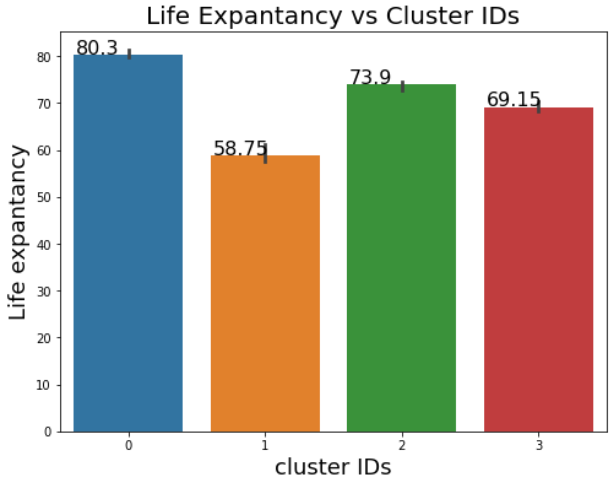
For clustering –
the Hopkins ratio was .81 which is very good to do the clusters

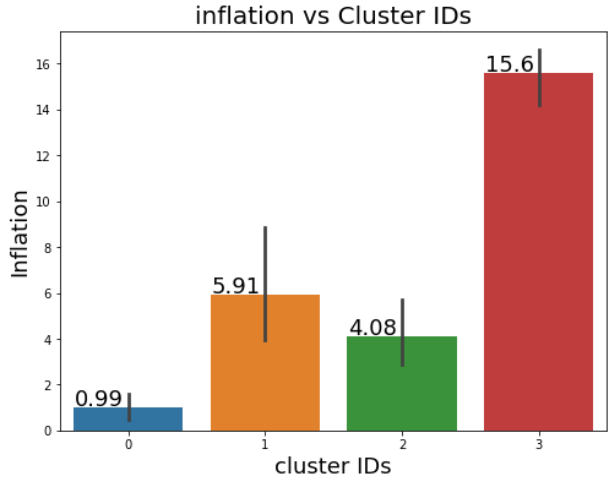
- The silhouette graph that the data is very consistent.
- From the elbow curve ,we can see that there was a sharp bent at k=4
- So we can take k=4.

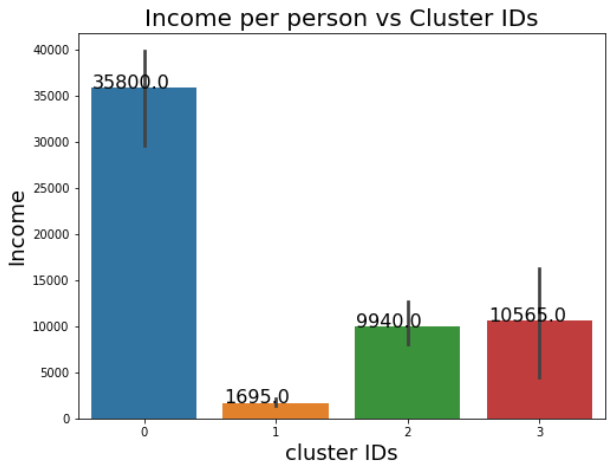




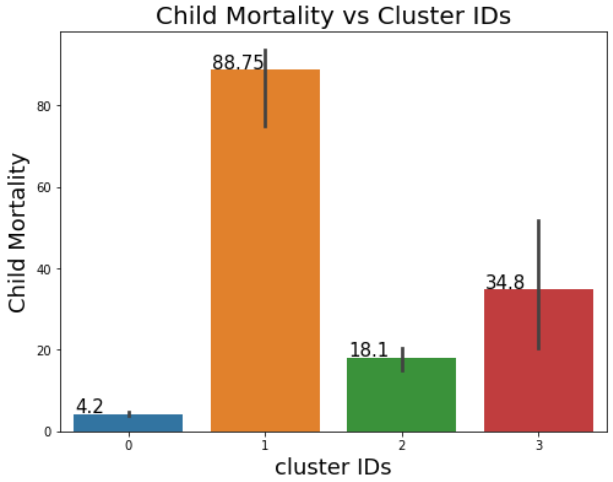


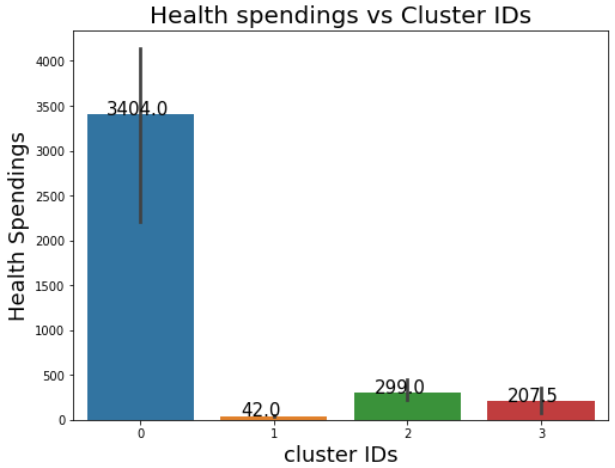


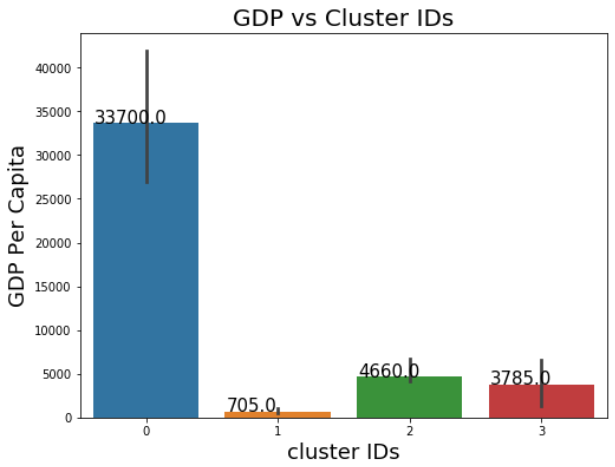




Visualising graphs in K means method



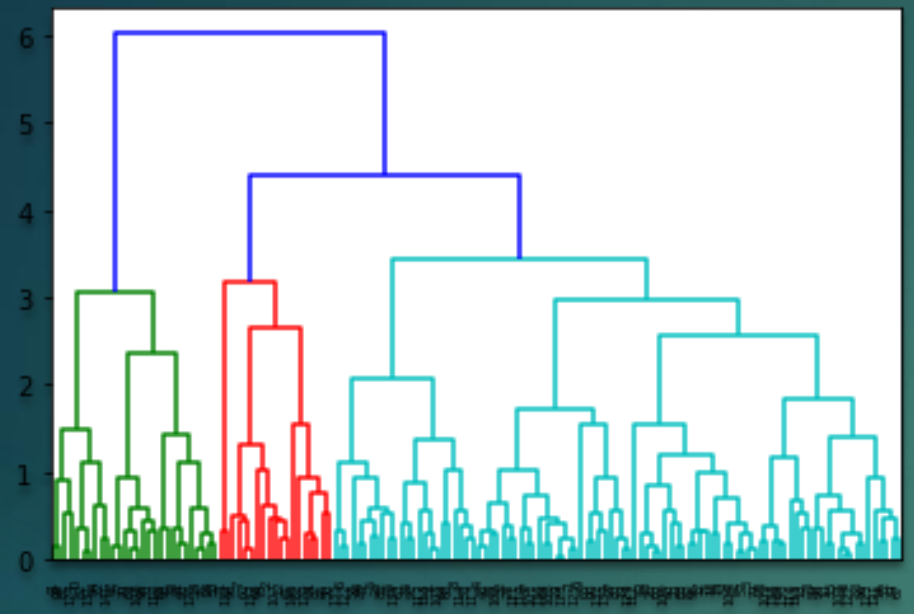


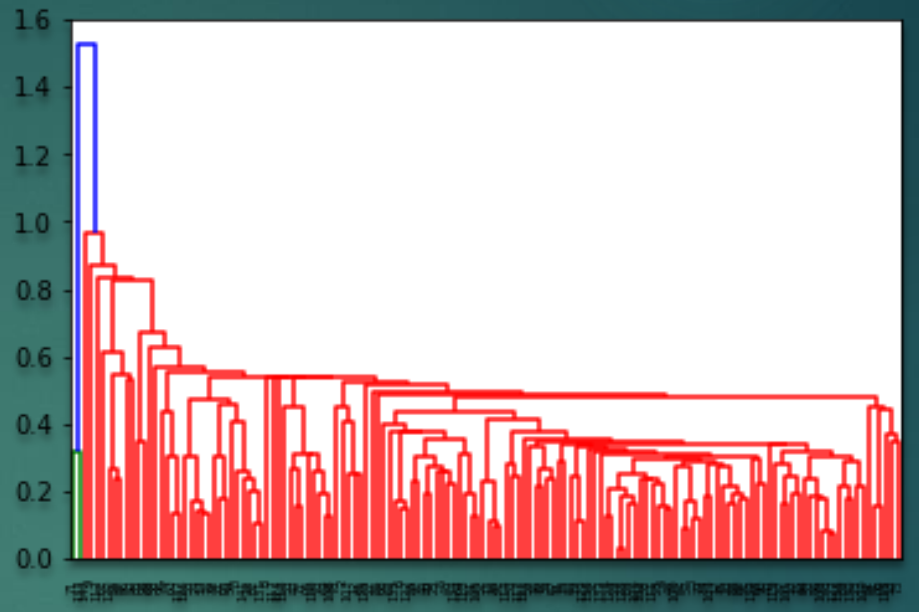


Analysis graphs of K means method of clustering

Analysing the k means method

- From the graphs it is evident that cluster 1 countries lag behind in all the variables necessary for the development of the country such as ghd ,income per person, spending on health while high on variables like child mortality ,inflation.

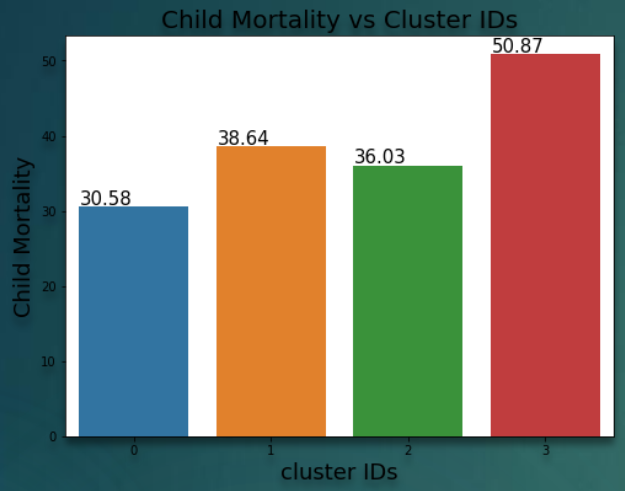


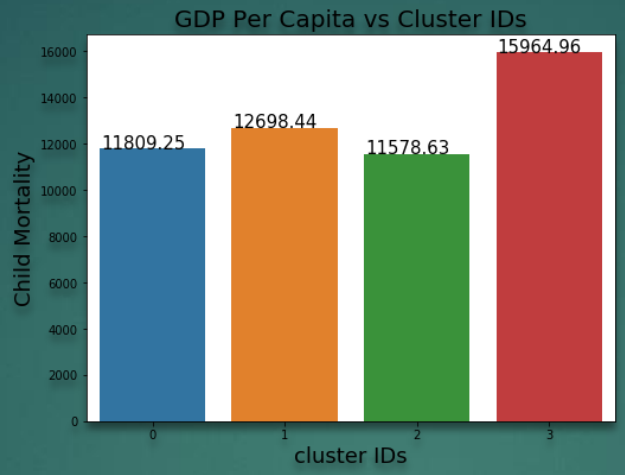


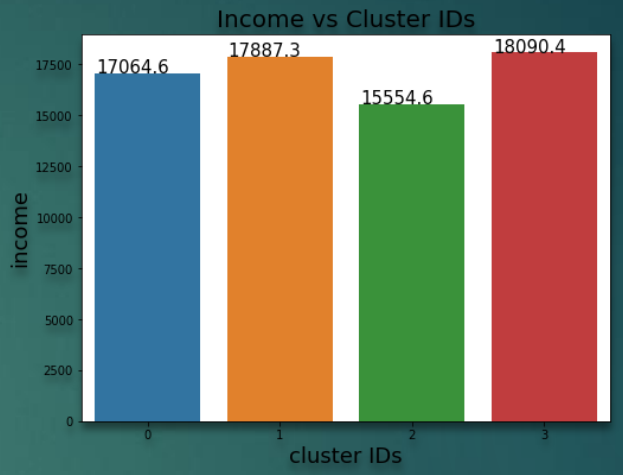
Complete linkage

single linkage

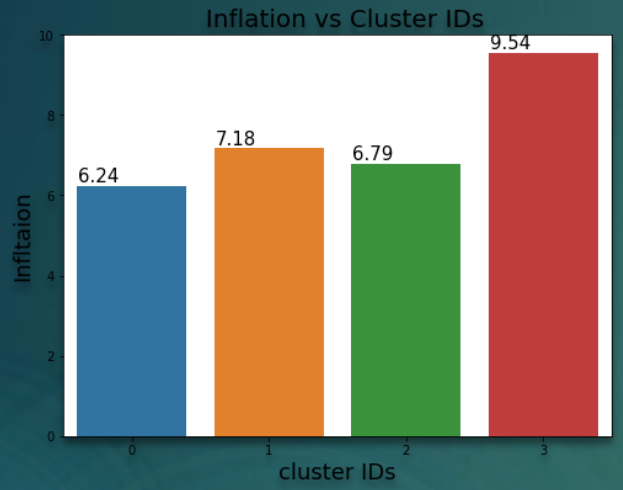
- Single linkage is very complex to understand
- Hence, we considered complete linkage and took k=4

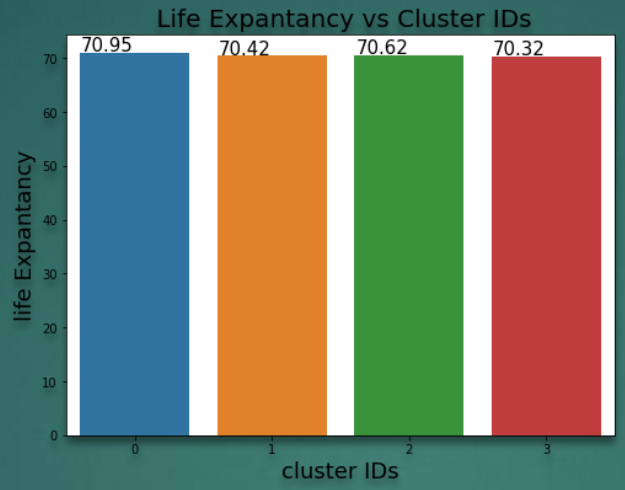


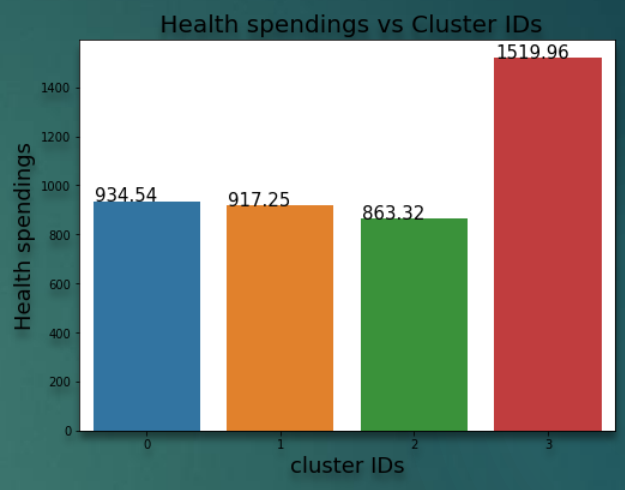




Visualising graphs obtained through hierarchical method







Visualising graphs obtained through hierarchical method

Analysis of clusters formed through hierarchical clustering.

- 1) Life expectancy is lowest in cluster-0.
- 2) Inflation is highest in cluster-2.
- 3) Cluster -0 countries spend the least in health among clusters.
- 4) cluster -0 countries are low in GDP.
- 5) Child mortality is highest in cluster-0 countries.
- 6) Income per person lowest in cluster-0 countries.

- Hence we can say that countries belonging to cluster -0 are in need of help.

Selection of countries which need help.

We can say that all the countries obtained thru hierarchical clustering are there in countries obtained thru K-means so we can say that countries obtained thru hierarchical clustering are the countries which are in dire need of help , as they are common in both analyses, so in total there are 28 countries which need help.

Outliers removal analysis

- .
- Removed the upper range of percentiles , i. e above 95 percentiles and kept all the record lower than it so that in the process of outlier removal, we don't miss out those countries which lie at very low percentiles and indeed are those countries which are in dire need of help.