COUNTRY CLUSTERING

CLUSTERING & PCA ASSIGNMENT - PART I

ANALYSIS APPROACH

- **Problem Statement**:- Identification of counties which need economic aid based on the Socio-economic and health factors.

- **Selected approach:-** PCA for dimension reduction followed by clustering for grouping countries

- **The analysis was performed with below steps:**

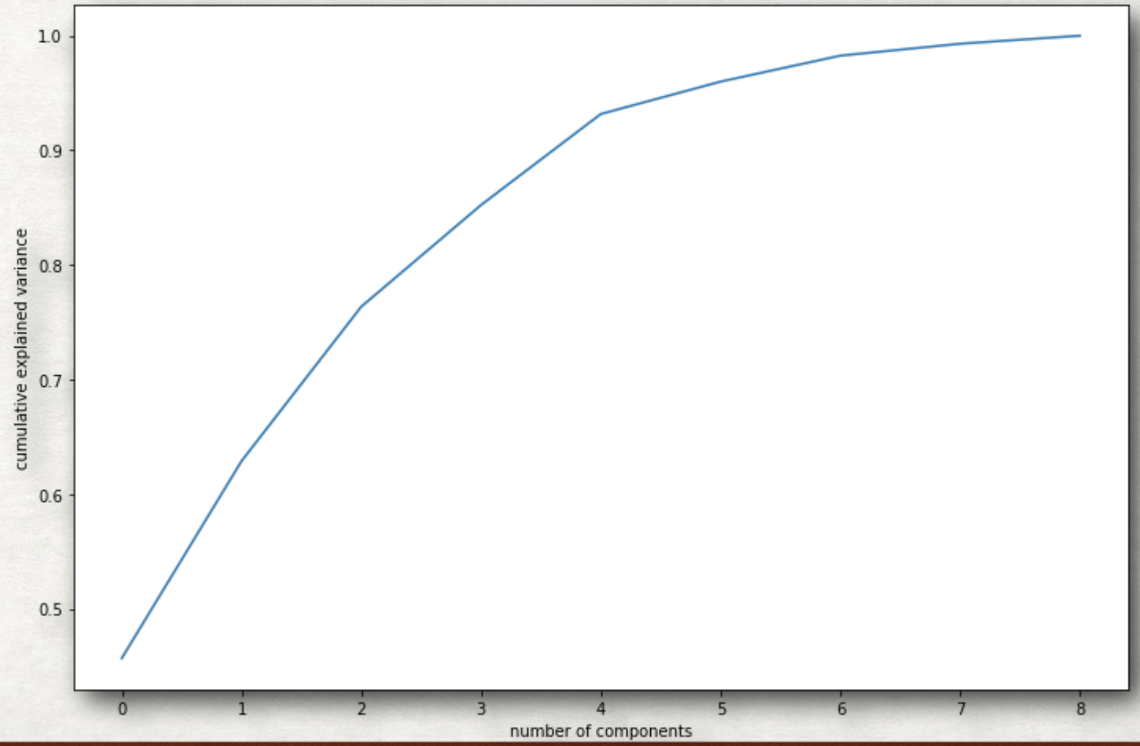
  - EDA - Data cleaning/Missing value analysis-Treatment

  - Outlier analysis and treatment - Pre and Post PCA - decision was taken not to perform outlier removal as this will remove important data

  - Data Standardisation - As clustering results vary on data scale we need to apply data scaling/standardisation

  - PCA - Principle component analysis was performed to select 4 principle components.

  - The 4 components were selected as around 94% of information is explained the the 4 components



Continues on next slide…

ANALYSIS APPROACH

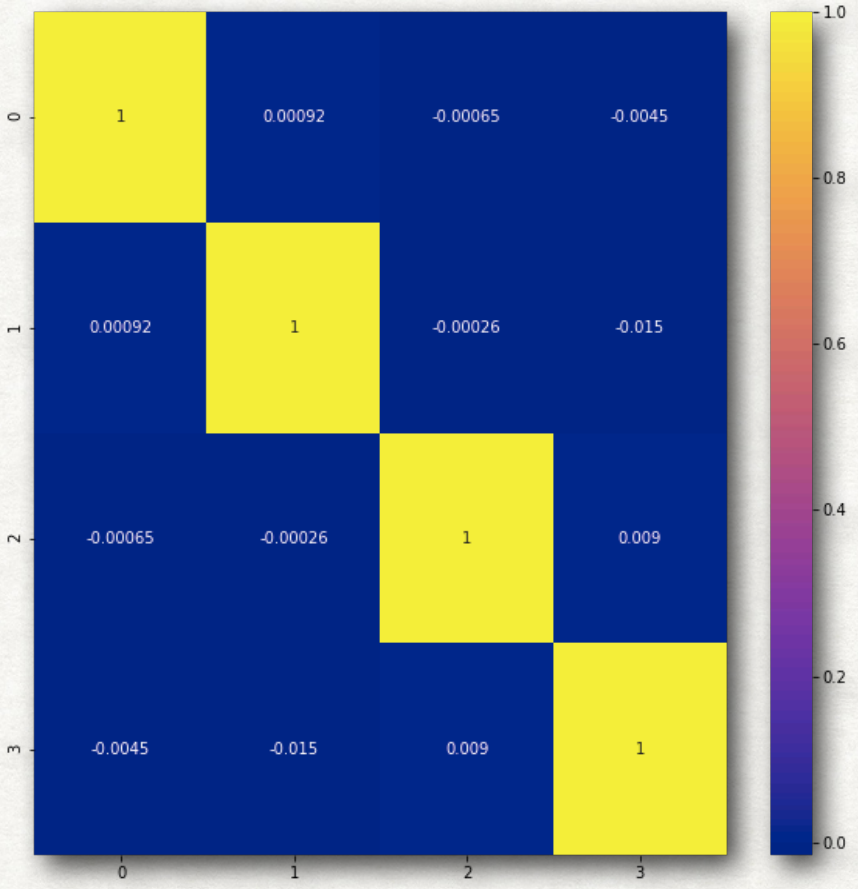
- The correlation between the principle components created is almost zero

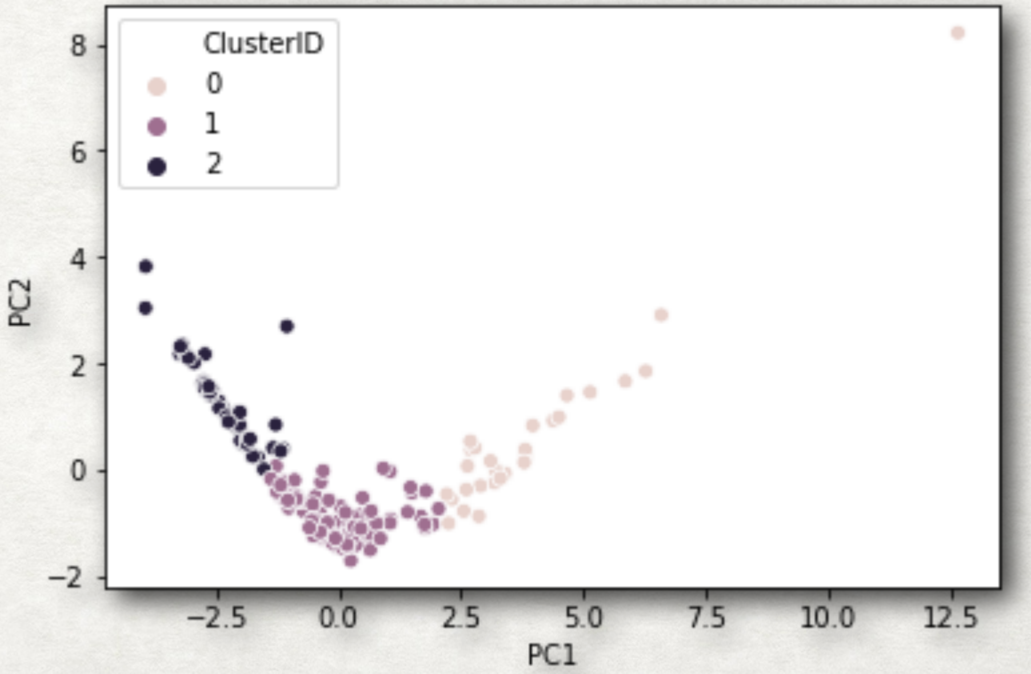
- After PCA is performed, the resultant dataset is used to check for the clustering ability by using Hopkins score which is > 0.90 so the dataset is excellent tendency of clustering

- After the PCA, we applied K-means and Hierarchical clustering algorithm on the dataset.

- The selection of the number of clusters for K-means was done by using Elbow analysis and Avg. Silhouette score analysis. K=3 was selected

- After creating and applying K-means clustering model on the PCA generated dataset, we get 3 clusters which partitions our country dataset





Continues on next slide…

ANALYSIS APPROACH

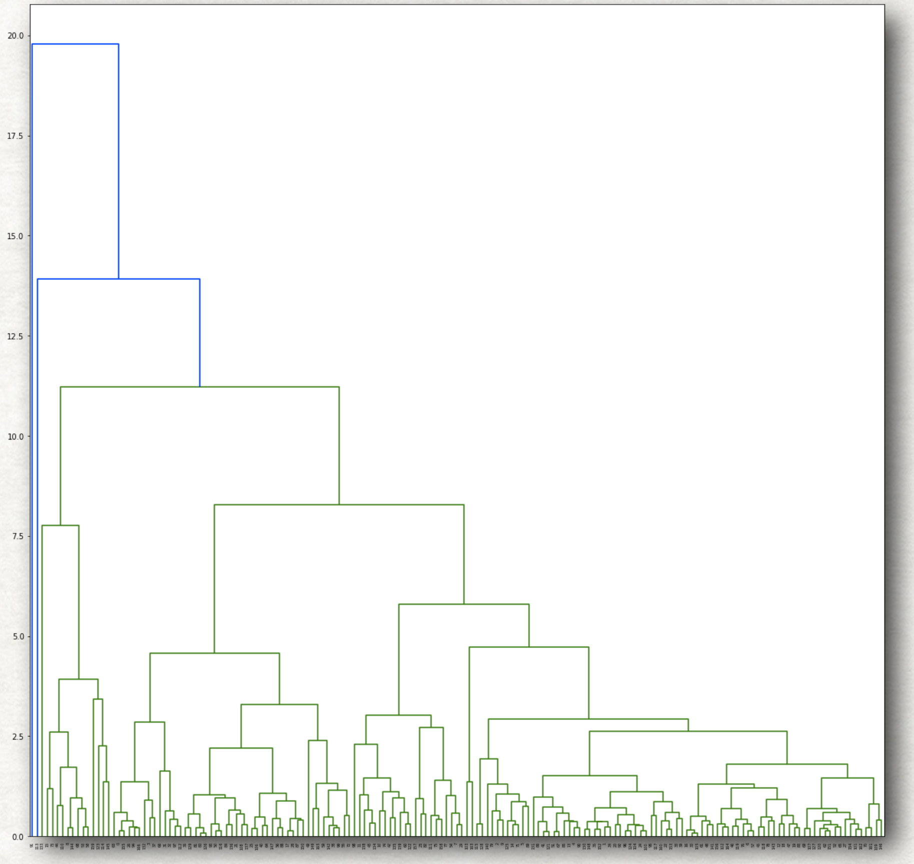
- The Hierarchical clustering model with single and Complete linkage was created to check which model gives us best results

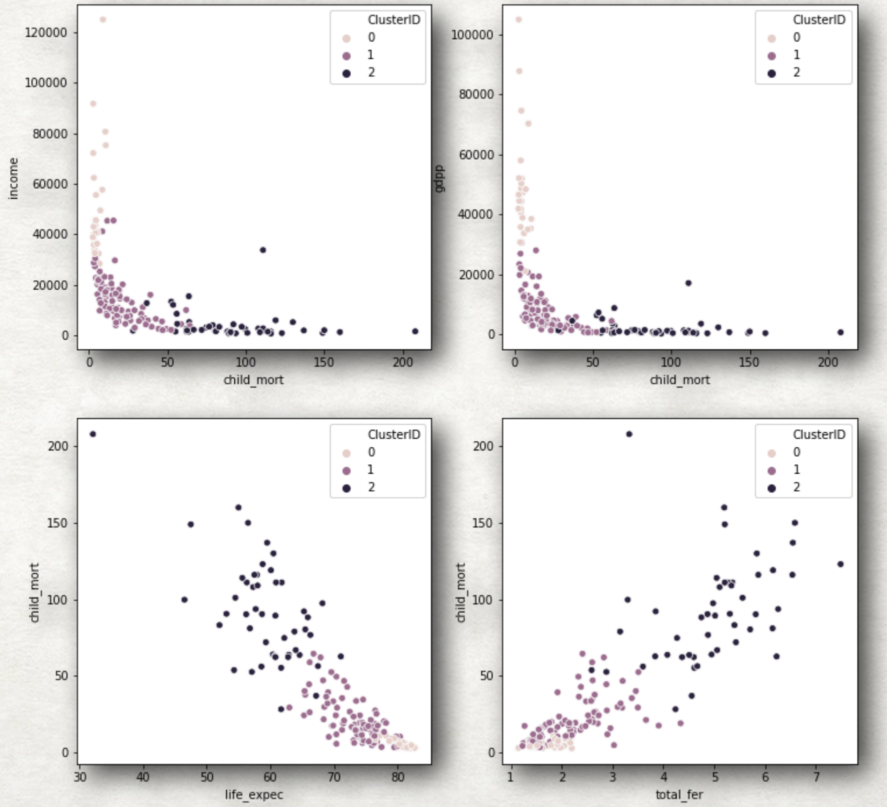
- The Hierarchical clustering resulted in many clusters with 1/2 countries in each clusters, this is result of the outliers in the dataset.

- The results of K-means clustering are more significant and helped in identification of countries based on the high child mortality, low income and GDPP

- The high child mortality, low income and GDPP countries belong to the underdeveloped country cluster.

- Around 48 countries belong to the underdeveloped country cluster.





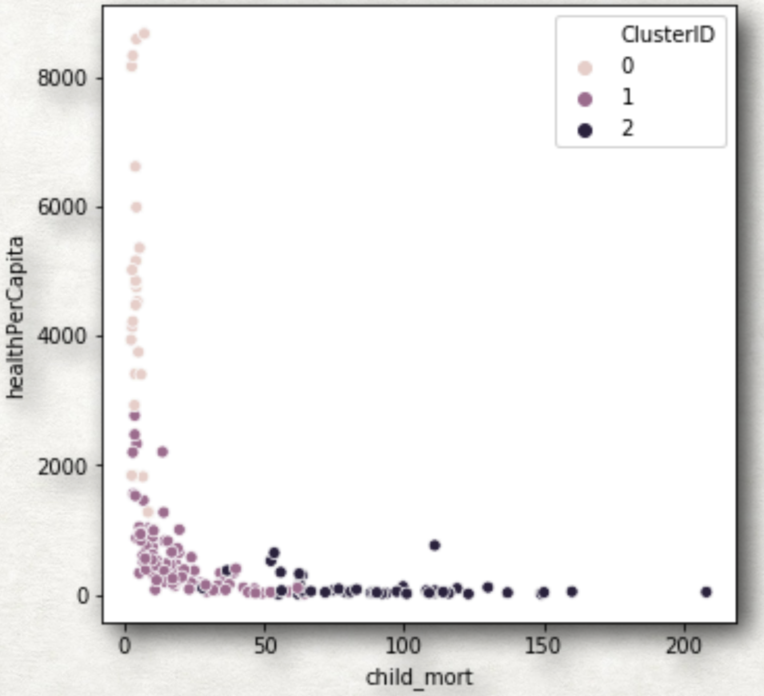
Continues on next slide…

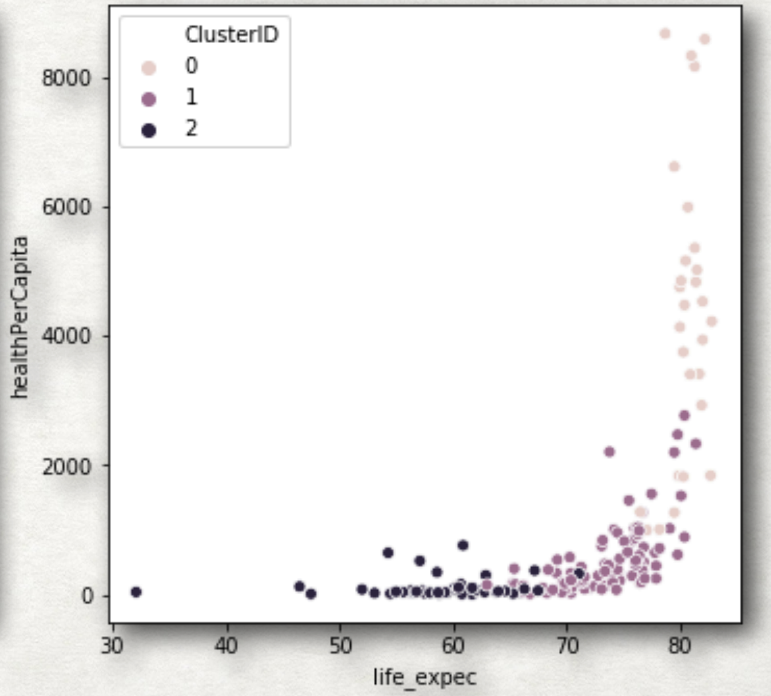
RECOMMENDATION

- Based on the analysis of the dataset with PCA and Clustering we have 17 countries which we have selected with following parameters from the Under-developed countries after binning the data
    - Income less than 3897
    - GDPP less than 1909
    - Child mortality more than 92

Below are top 5 countries which we can focus as they are in need of the aid as they have quite low spending on health and which is affecting their child mortality and life expectancy

- Haiti
- **Sierra Leone**
- **Chad**
- **Central African Republic**
- **Mali**


ClusterID
healthPerCapita
child_mort


ClusterID
healthPerCapita
life_expec