PCA and Clustering Assignment

Understanding the data

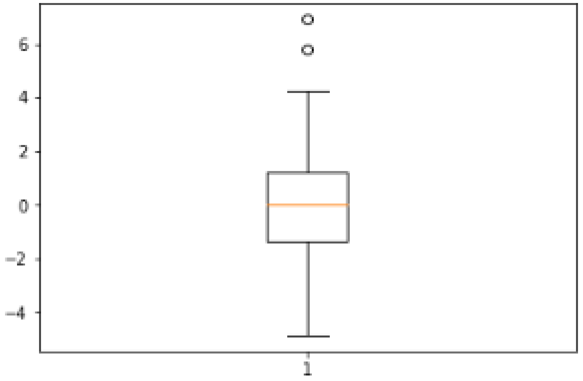
- We go through the data dictionary and try to understand the data
- we have to categorize the countries using some socio-economic and health factors that determine the overall development of the country.
- Then you need to suggest the countries which the CEO needs to focus on the most.
- The datasets containing those socio-economic factors and the corresponding data dictionary are provided below.

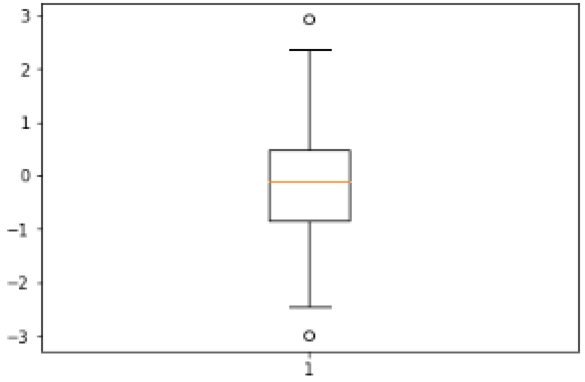
Clean the data

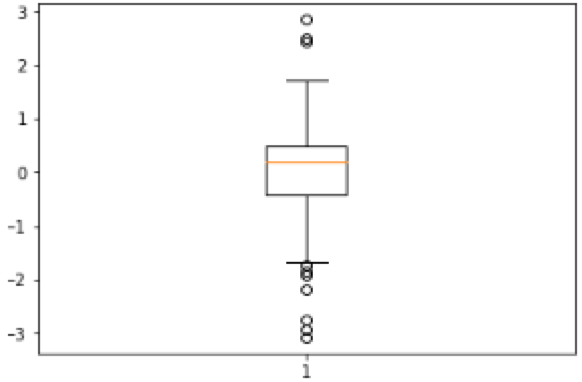
- We convert datatypes from int to float to keep it same as other columns
- We check for missing values and treat them if necessary
- We drop the unwanted columns

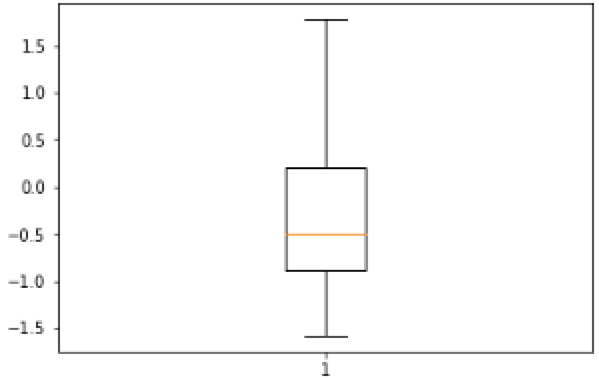
Prepare the data for modelling

- We perform standardization
- We also scale the variables
- Drop the unnecessary columns
- We do the fit_transform on our data
- Outlier treatment



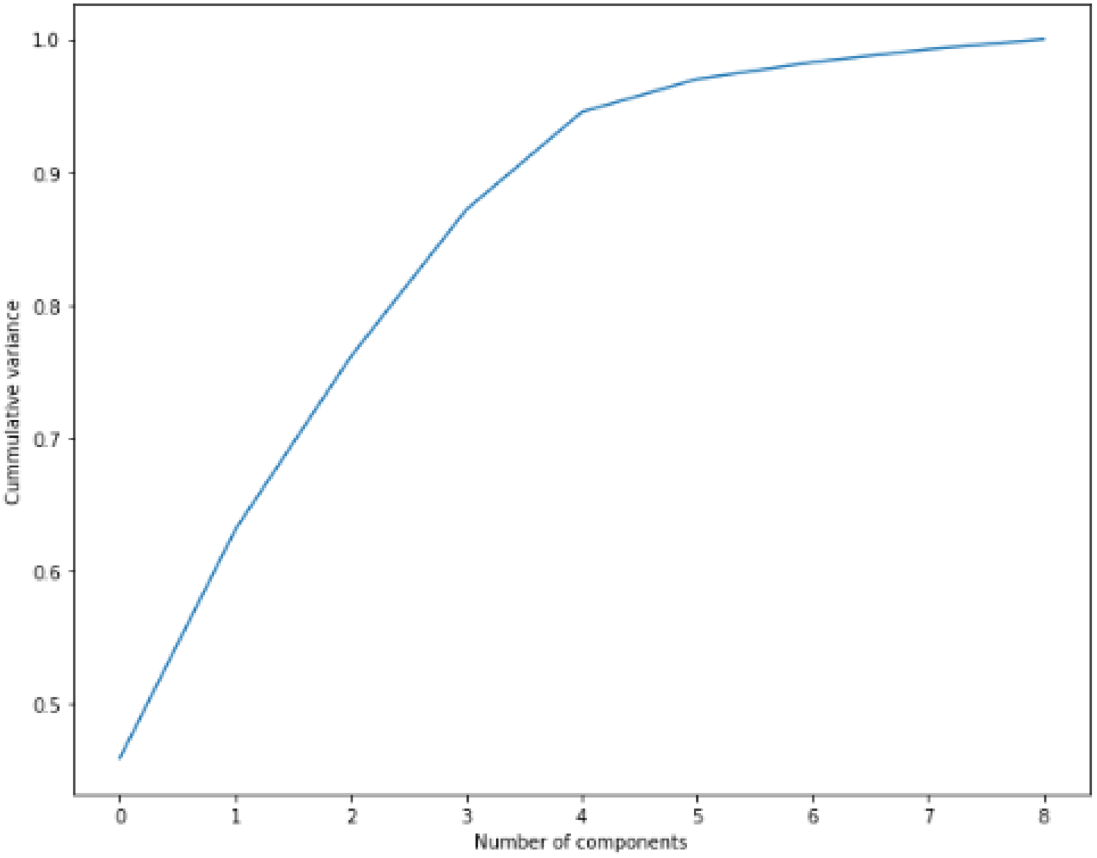






PCA

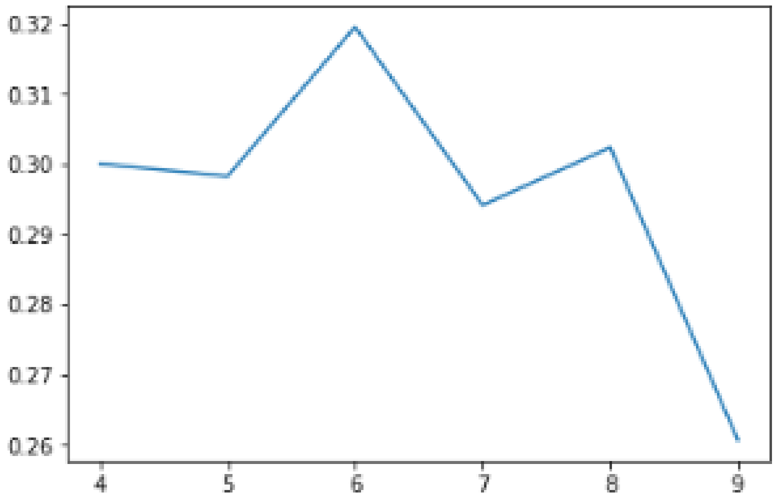
- We instantiate the PCA module
- Fit the data
- We can view the pca.components_



- We check the explained variance ratio of the components
- We plot the number of components vs cumulative variance

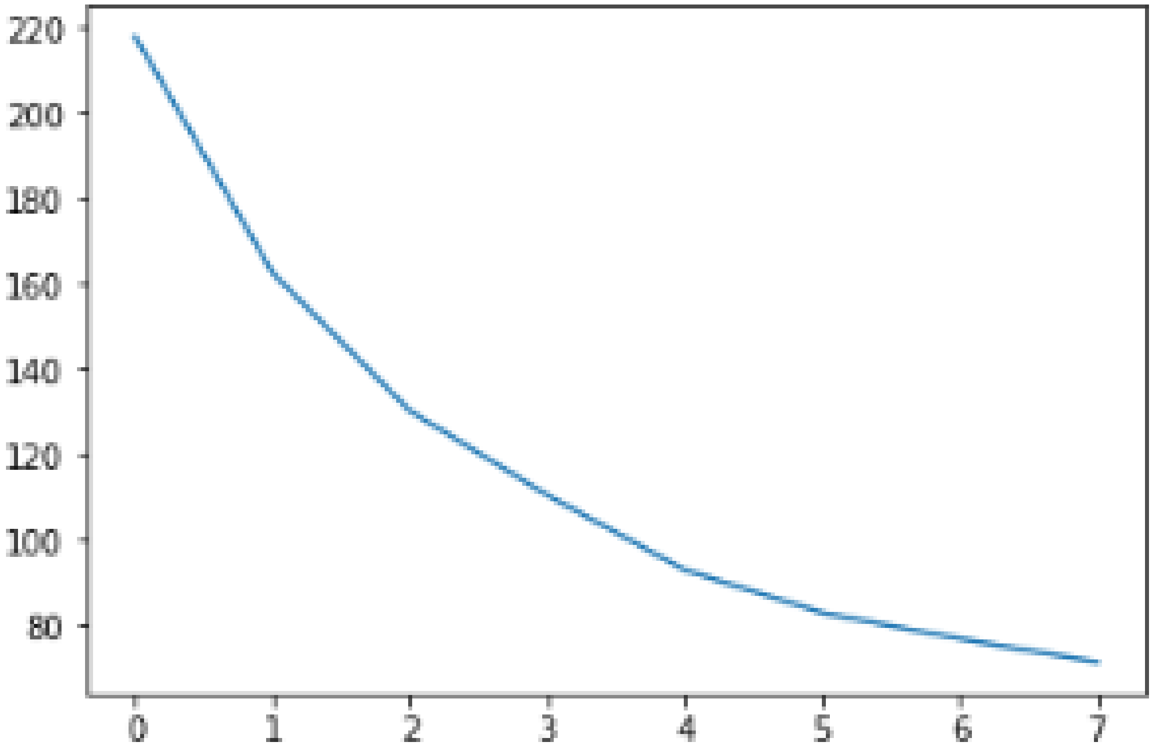
Checking the cluster tendency on PCA dataset

- Hopkins measure
- We see the randomness of the data by this measure
- Hopkins compares out dataset with a set of random variables and gives us a
- Hopkins measure
- Higher the Hopkins measure better the data set
- Silhouette score
- We find the silhouette score as well and select the max peak



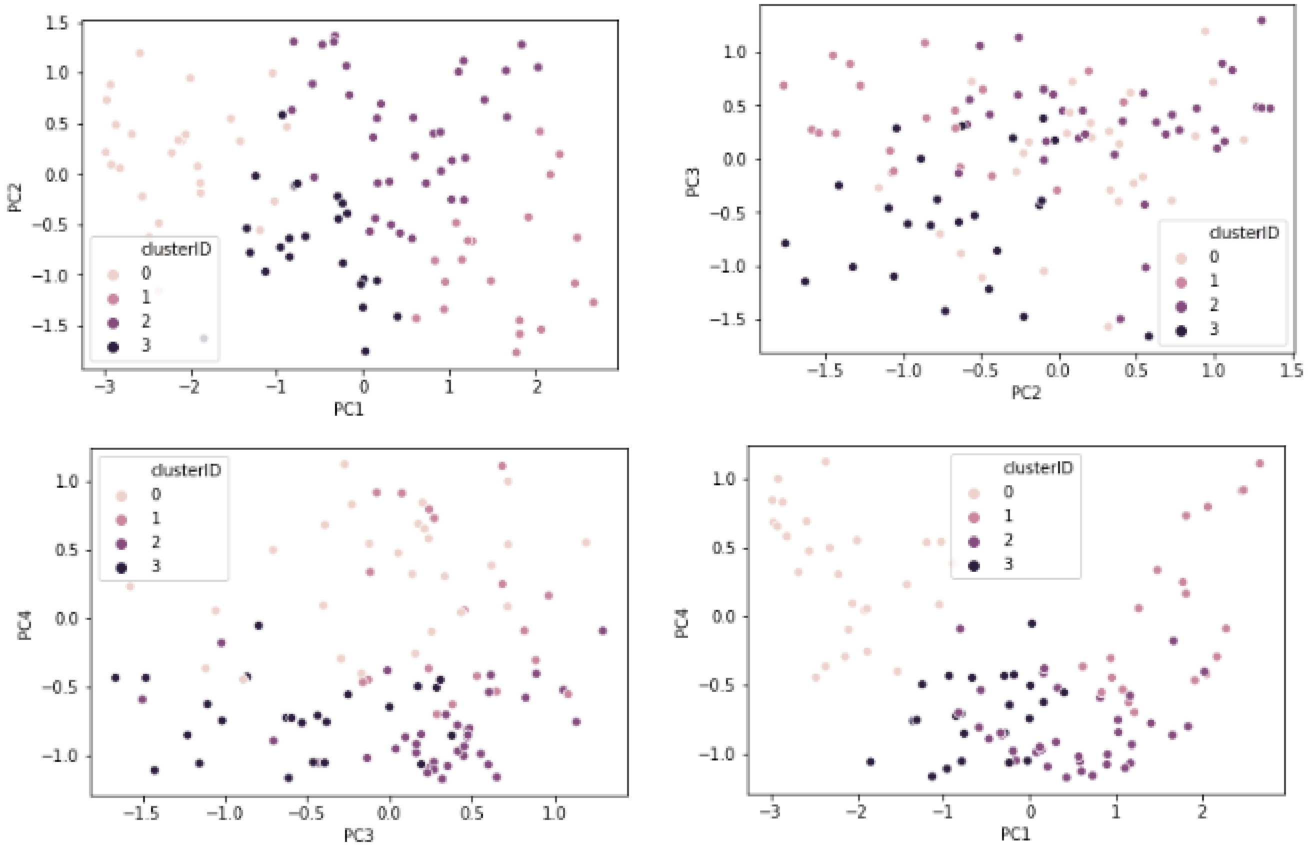
Elbow curve method

- We select the optimal number of clusters looking at the elbow curve element



Cluster modelling and visualization

- We plot the clusters of PCA



Number of countries

- By comparing the countries with high child mortality and low gdpp and income we found 20 countries that are in dire need of help.
- Countries:
- Bahamas,Barbados,Chile,Costa Rica,Croatia,Cyprus,Czech Republic,Finland,Germany,Greece,Iceland,Israel,New Zealand,Poland,Portugal,Serbia,Slovenia,South Korea,United Kingdom,Uruguay