PRML MINOR PROJECT

STUDENT-1 NAME	RAUNAK GARG
STUDENT-1 ROLL NO.	B21EE056
STUDENT-2 NAME	SAMARTH SUDHIRKUMAR BHALERAO
STUDENT-2 ROLL NO.	B21EE060

Problem Statement : "HELP International" has been able to raise around $ 10 million. Now the CEO of the NGO needs to decide how to use this money strategically and effectively. So, the CEO has to make a decision to choose the countries that are in the direst need of aid. Hence, your Job as a Data scientist is 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.

Data Provided : We are provided with "***Country Data***". It is a dataset having all the required data of a particular country like its life expectancy, health, mortality rate, gdp, etc. & these all factors are indirectly relatable to each other on the basis of which the decision to invest in the health sector of a particular country can be made.

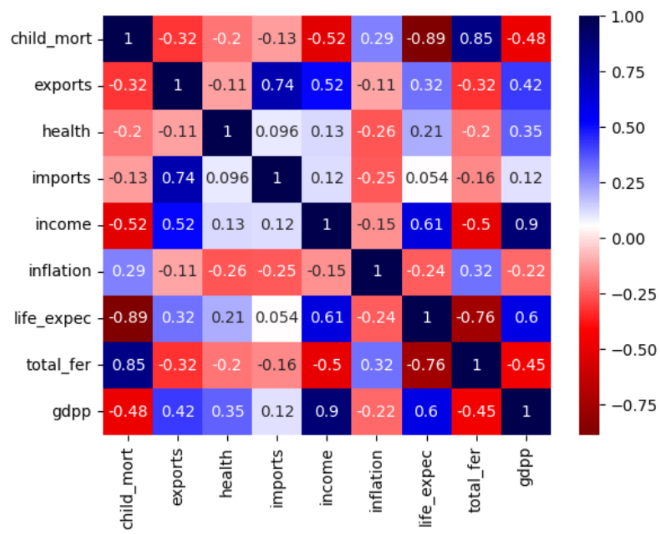
TASK UNDERSTANDING : In this problem statement, as we need to make the decision that which country really need the investment in the health care sector depending on the economic and life-living factors , to we need to make different groups of the countries, i.e. clusters so that the particular group can be identified that needs the investment help.

PipeLine-Flow:
Loading & Preprocessing the data—> Visualize the data so that the understandability of data becomes easy —> Reduce the data dimensions if required —> Apply various methods of clustering and get the best clusters

of the countries —> Identify the cluster type and make the final decision to invest on the the group of countries.

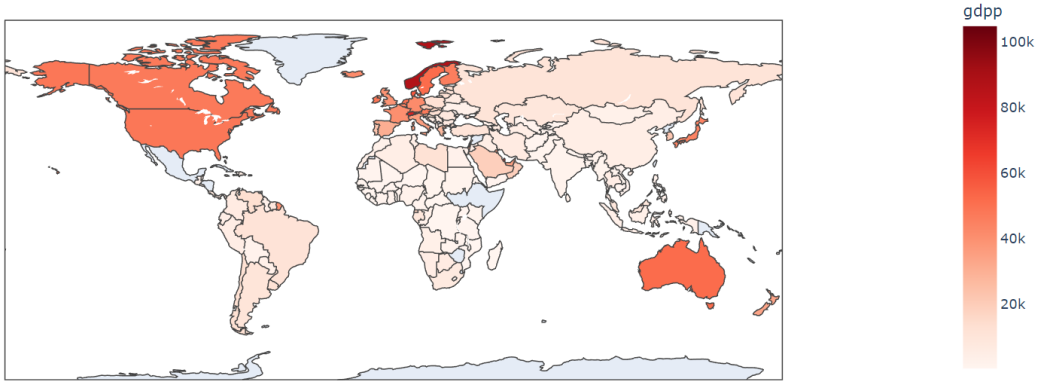
- **Fetching and preprocessing the data -** We have loaded the dataset, obtained the dataframe then we checked for NULL values and found that the dataset has no NULL values,

- **Visualizing the correlation of features in the data and the overall data to proceed further -** We used *heatmaps* to obtain the correlation between the features and we observed that all the features are indirectly dependent on each other and are co-related. On having a read further ahead about the given factors, we can know that all the features will contribute in decision making; so we decided to proceed with all the features and setting the country as our target variable. The heatmap of the correlation.

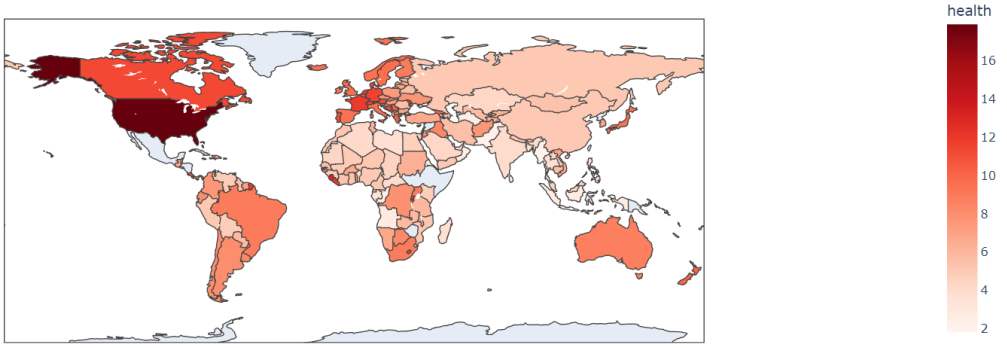


We also visualized the individual data columns using 2D *World Maps* plotting as a function of intensities and statistical graphs.
One example of gdp of country visualized using the world map.

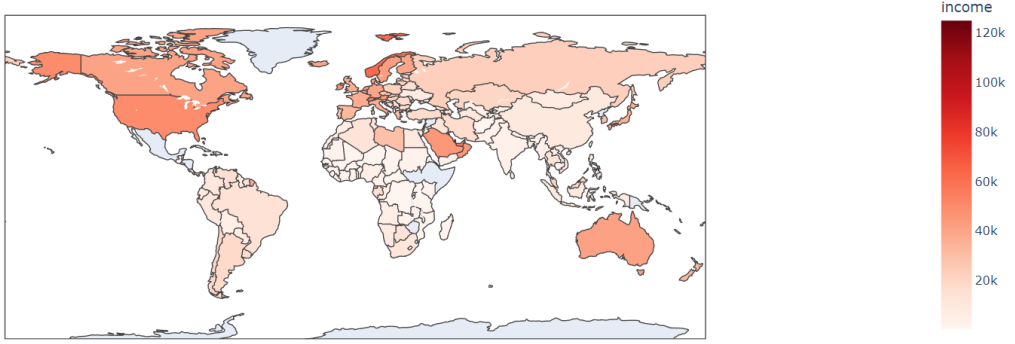
gdpp per country (World)



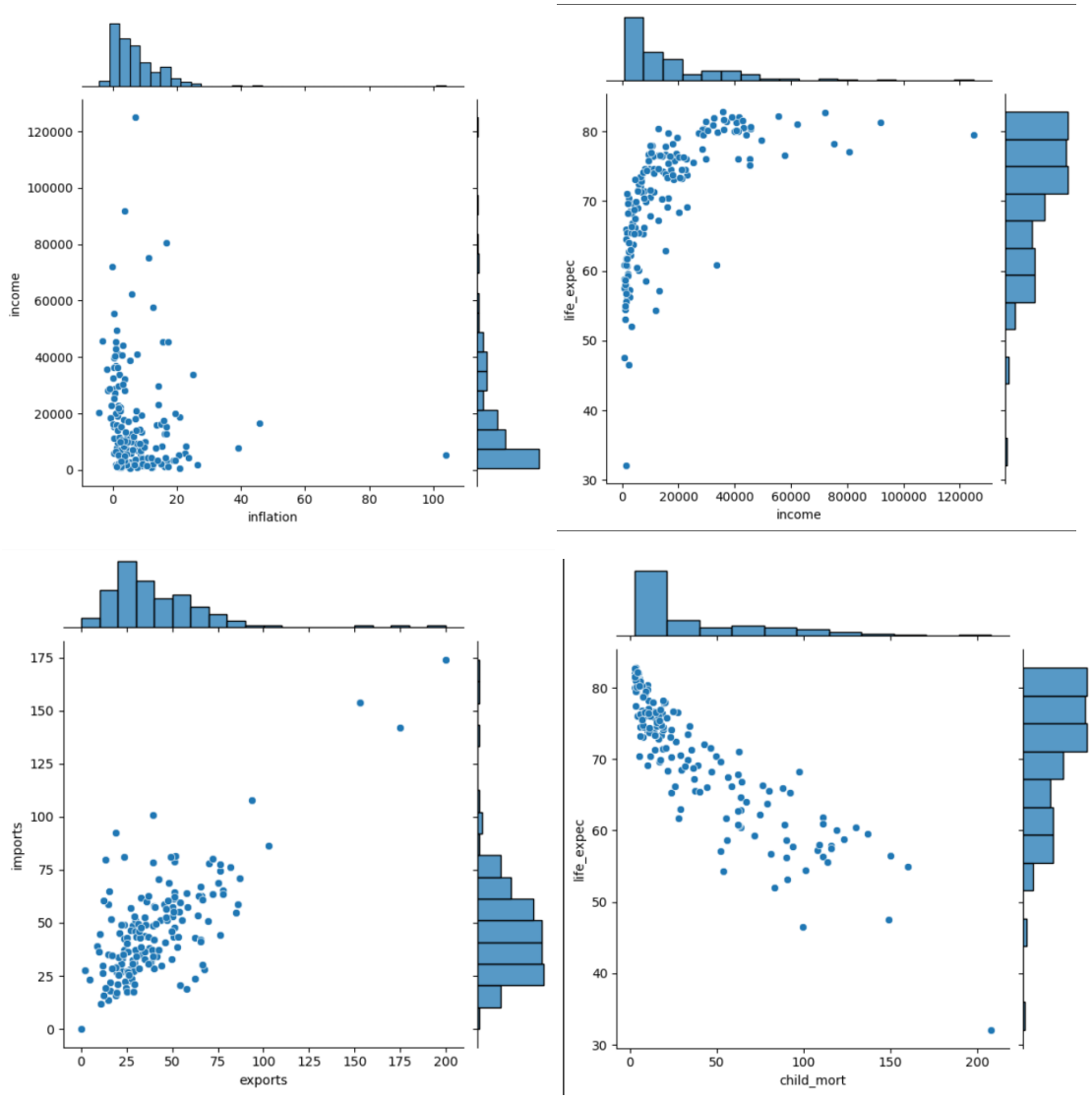
health per country (World)



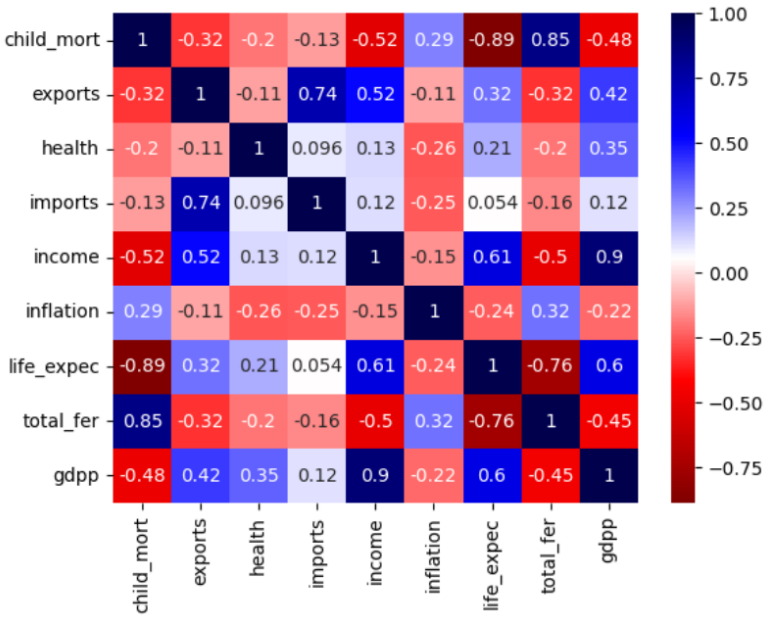
income per country (World)



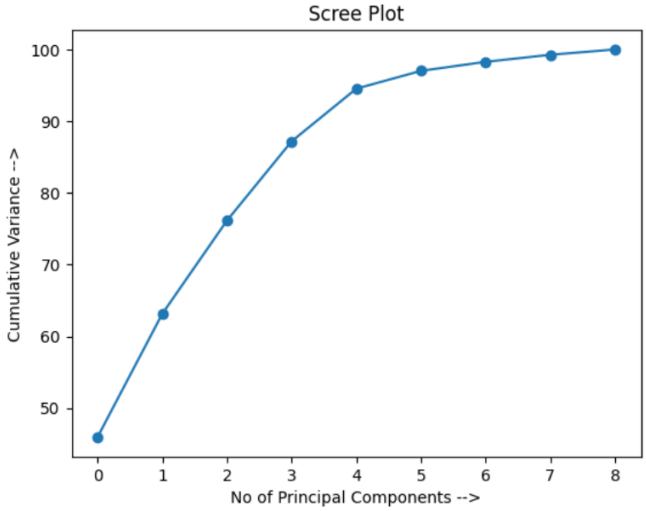
We have also made some joint plots so that we can study the indirect-relation between the given two features as all the features are indirectly related and are affecting the overall factors of clustering.



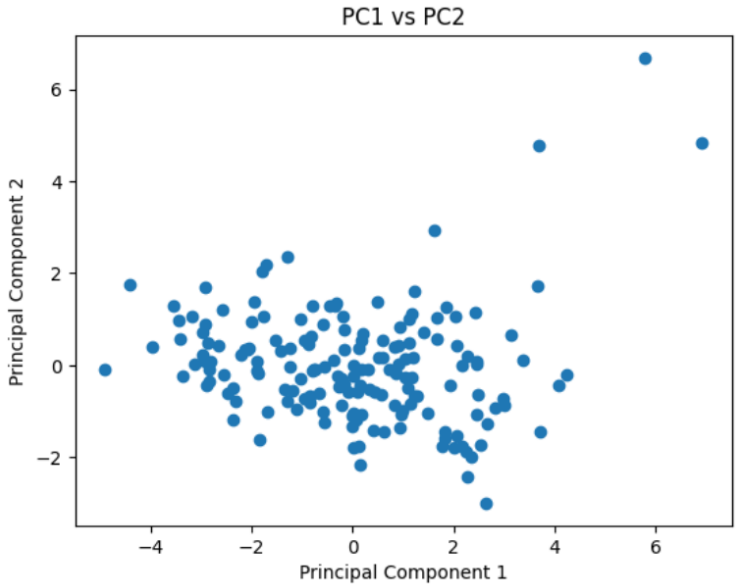
- **Data scaling -** We then scaled the data as we observed that the data values of some features were too big as compared to some other features which will result in biased dimension reduction or further reduction. For scaling purposes we are using the *Standard Scaler* provided by sklearn. So using scaler, we obtained the scaled dataframe. The heatmap of correlation:



- **Dimension Reduction** - We then applied the *dimension reduction technique* to make the data more easy to analyze. We here used the *PCA* , i.e. principal component analysis method to do so. It gave us a total of 9 principal components but on the basis of the "*Scree plot*" we decided to go with the first 5 important principal components.
  To proceed further, we created the final *PCA* dataframe with the first 5 principal components (PC1, PC2 ,etc.) to make its use in further clustering processes.



Also we visualized the PC1 vs PC2 as these are principal components as these contribute with the maximum variances.



- **Clustering -** As we have initially identified that the solution to the problem statement is to obtain a cluster. Thus we are using different *clustering methods* which are unsupervised as we have no initial label information available to obtain the clusters.

  The clustering methods used are mentioned below :
  - KMeans Clustering
  - Spectral Clustering
  - BIRCH Clustering
  - Agglomerative Clustering
  - DBSCAN Clustering

As we have two data frames available to apply clustering thus to compare the models which suits the best, we have obtained the corresponding *"Silhouette scores"* for all the clustering methods applied on both the dataframes, i.e scaled dataframe (df_scaled) & data frame obtained by

PCA (pca_df). Here we are noting down the corresponding Silhouette Score.

The ***silhouette score*** is calculated by comparing the average distance between an object and all other objects in its own cluster to the average distance between the object and all objects in the nearest neighboring cluster. The silhouette score can be used to compare different clustering algorithms and to select the best number of clusters for a given dataset.Thus we are using the silhouette score as one of the tools for evaluation and comparison purposes .

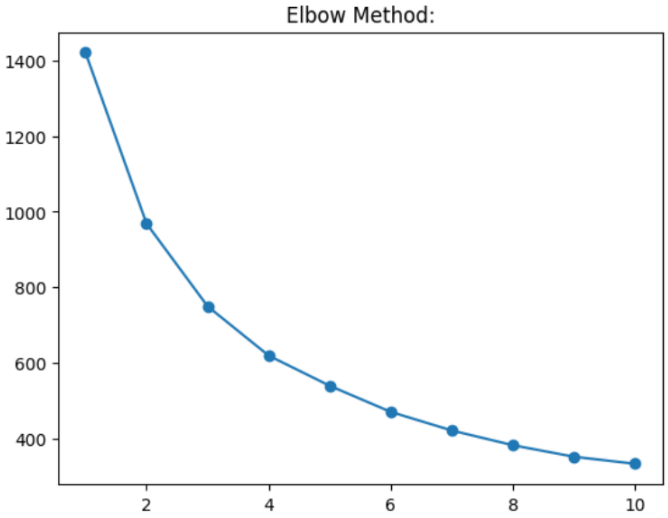
Clustering on Dataset	Silhouette Score
KMeans on PCA data	0.32718347402877235
KMeans on Scaled data	0.2975089381471821
Spectral on PCA data	0.29167897323118375
Spectral on Scaled data	0.2724597604806676
BIRCH on PCA data	0.3063359428343731
BIRCH on Scaled data	0.284327833626243
Agglomerative on PCA data	0.31413405010910866
Agglomerative on Scaled data	0.24811891847692066

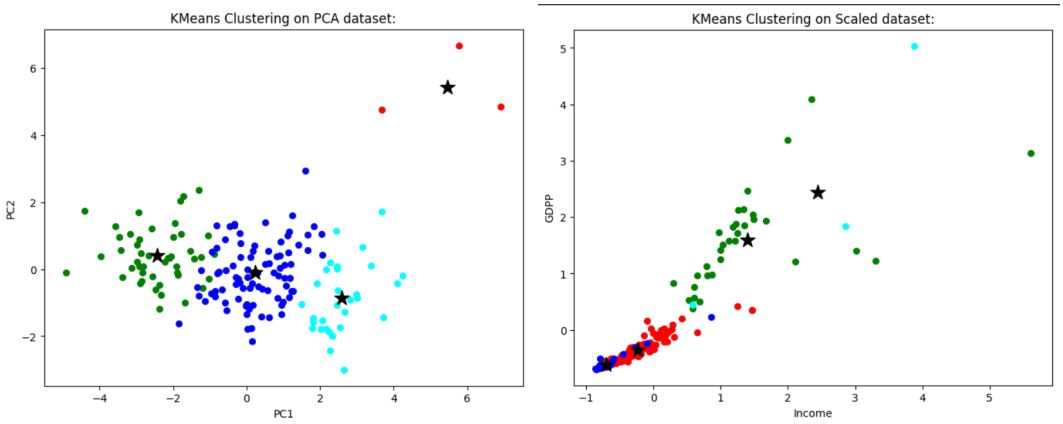
- **Clustering Outcomes**

☐ **K-Means**

We performed the K-means clustering on both the PCA scaled data and only scaled for the number of clusters defined as 4. We have obtained this cluster number as 4 using the elbow method. Also this number is logically valid as the worldwide countries can be majorly classified in 4 categories starting from most developed to countries

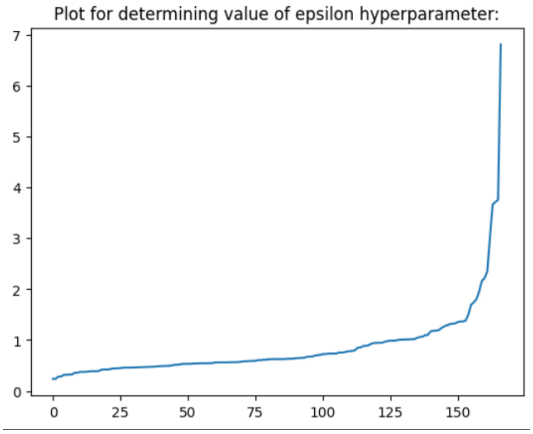
having very much scope of development. Thus we now have both k-mean clusters for PCA and without PCA scaled data.

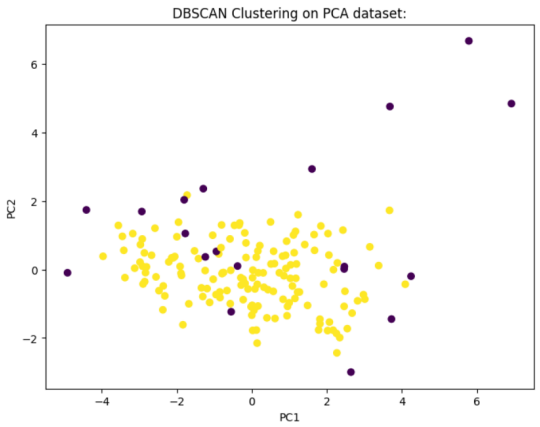




☐ DBSCAN

DBSCAN method is applied to both the dataframes and the labels are obtained but we will not proceed with this clustering method as this method is mostly used to identify the *outliers*, and in our case, the outlier will correspond to both extremely developed and extremely underdeveloped countries and both will be labeled same, so it is not suitable for solving our problem statement.





☐ **Spectral Clustering**

We applied the Spectral Clustering method on both the PCA and scaled dataset to obtain 4 clusters (You can choose any but to standardize the process, we are using number of clusters as 4).We obtained the clusters and corresponding silhouette score and in all multiple runs, we found that k-means clustering is performing Better than spectral clustering method on the basis of silhouette score thus till now k-mean clustering is considered for further consideration,

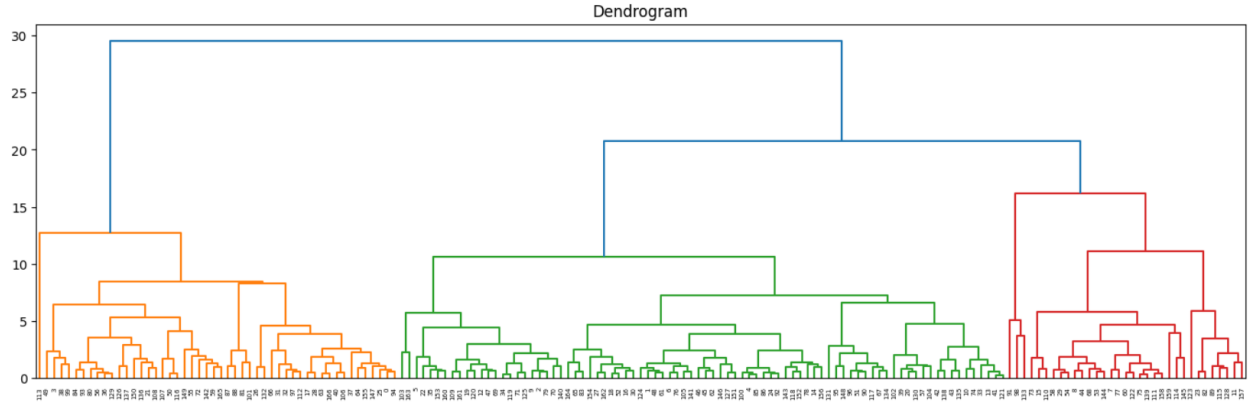
☐ **Birch Clustering**

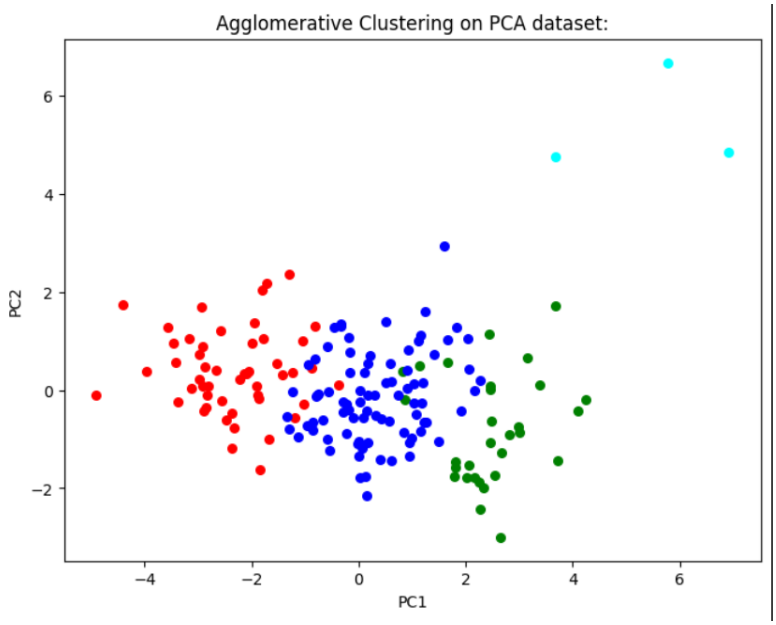
We are now using “Birch Clustering” .The **Birch** builds a tree called the Clustering Feature Tree (CFT) for the given data. For birch clustering, we need to pass the number of clusters, threshold & the branching clusters. The branching factor limits the number of subclusters in a node and the threshold limits the distance between the entering sample and the existing subclusters. We set the number of clusters again as 4 to maintain the standard flow, obtaining the labels & silhouette score.

☐ **Hierarchical clustering**

Hierarchical clustering is a general family of clustering algorithms that build nested clusters by merging or splitting them successively. This hierarchy of clusters is represented as a tree (or dendrogram). From

the hierarchical method of classification, we can observe the distribution of countries in each of the hierarchical levels. We are using agglomerative clustering. Its Bottom-up approach treats each data as a singleton cluster at the outset and then successively agglomerates pairs of clusters until all clusters have been merged into a single cluster that contains all data. We obtained the corresponding final labels but to a restricted number of clustering, which may indirectly result in slightly lower silhouette score than k-means clustering in all multiple runs, we are considering with the k-means till now for final decision making.





▶️ We now have performed all the clustering methods and the required processes to final labels and are considering further with *k-means clustering* on the reasons mentioned above in each method and obtained silhouette score after multiple runs (almost same in all runs).

Analytics📃

Analytics to make final conclusions on the obtained clusters to *label* them with the particular class that which cluster belongs to which cluster-id.

For assigning final labels to cluster id obtained from the k-means clustering, we are considering the multiple factors from the data frame corresponding to each of the countries.
As we have obtained 4 clusters, we are labeling the clusters as follows:

- Most Help Needed → Most backward in terms of health and finance parameters.
- Need Help → This are the countries which don't have the worst conditions but have so much scope of development.
- Might Help Needed → This group has the countries which have nice developing rates and need almost no external help.
- No Help Needed → Most developed countries.

Factors taken in consideration - We took the average values of the factors like, gdp, life-expectancy rate, mortality rate, health etc to judge the clusters.

Countries having high mortality rate , less health, low life expectancy and indirectly less GDP are the countries requiring the help in terms of investment in the health sector.

So we obtained the final classified labeled cluster lists.

Also, we got the sorted list of countries in the labeled class of countries requiring help the most in the order of countries requiring help the most to least on the basis of financial factors like gdp.

Moreover to make the decision specifically, i,e. Amount of money to be distributed in top n (here n is taken 7) in the way such that the money is distributed according to the requirement, we created the function on the basis of GDP (indirectly related to country's financial worth) to distribute the ***10 Million dollars*** and gave the final sheet that how much amount is required to be distributed to which country. Thus the problem is now precisely ***solved*** and we got the ***final precise decision*** about the investment required.

About the cost- distribution function

We have the sorted list of countries in which we have all the countries which require help the most. We also have the corresponding gdp value for each country thus we summed up the gdp for n countries, divided the gdp of each country by the sum of n country's gdp & obtained the respective fractions. Now we subtracted the fractions from unity as the country having least fraction value requires the most money for growth and thus we got the percentage distribution of money among the n countries corresponding to the resulting final fraction value.

Investment Statistics

The following are the investing statistics if the NGO wants to distribute the amount in 7 most needy countries in the terms of investment required in the health care sector. Here again, the code is flexible to obtain the amount of distribution of the given amount among n number of most needy countries.

Burundi – $ 1510859.301227573

Liberia – $ 1446108.18831782

Congo, Dem. Rep. – $ 1441386.7530014836

Niger - $ 1431943.8823688116

Sierra Leone - - $ 1397544.8536355053

Madagascar - $ 1388101.9830028329

Mozambique - $ 1384055.0384459735