Clustering of Countries
Assignment Submission Report

- An assignment to cluster a group of countries based on socio-economic factors.

Business Objective

- ***To categorize the list of countries using socio-economic and health factors that determine their overall development.***
- ***This will help the CEO of HELP International(an NGO), to extend financial aid towards countries that are in the direst need of it.***

The objective is thus classified into the following sub-goals:

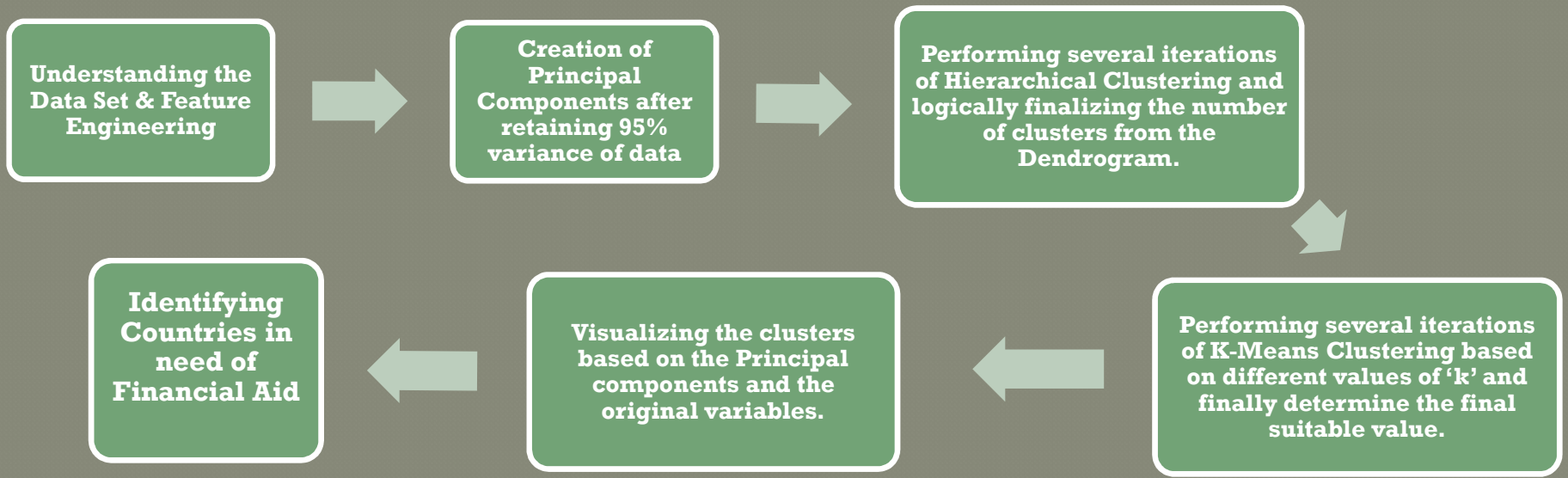
Create the optimal number of PCs so that it results in reduction of dimensionality and multi-collinearity among variables.

Determine the optimal number of clusters required to properly segment the group of countries based on the PCs.

Cluster the countries into pre-determined no of clusters and represent based on the PCs and the original variables.

Problem Solving Methodology

- The approach for this project has been to divide the entire case study into various checkpoints to meet each of the sub-goals using PCA and Clustering techniques. The checkpoints are represented in a sequential flow as below:



Data Preparation & Feature Engineering

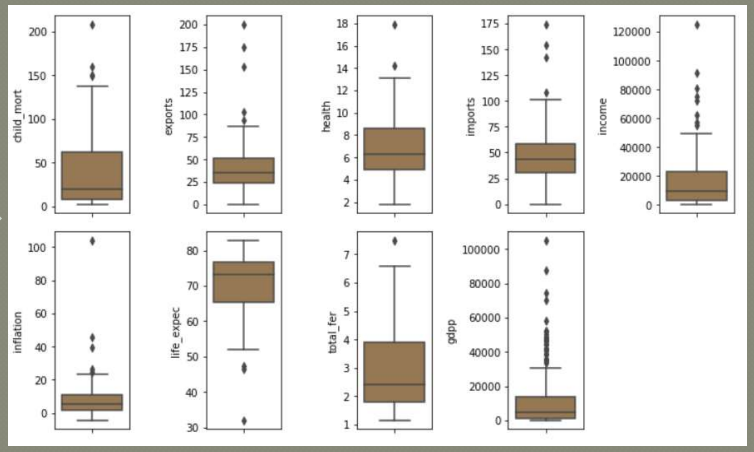
The following data preparation processes were applied to make the data dependable so that it can provide significant business value by improving Decision Making Process:

Feature Scaling

- Scaling all the variables (continuous in this case) to ensure they are on the same scale so that PCA doesn't give more importance to some variables over another.
- Used SKLearn's Standardization package to bring all of the data into a standard normal distribution with mean at zero and standard deviation at one.

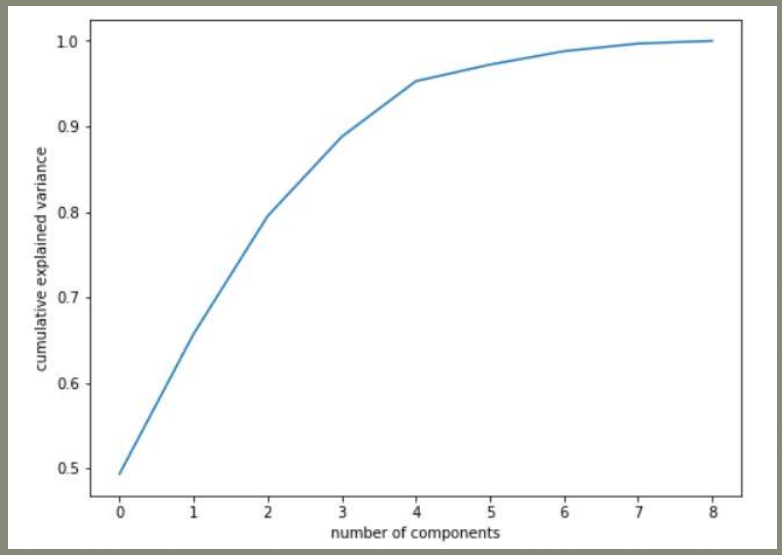
Outlier Treatment

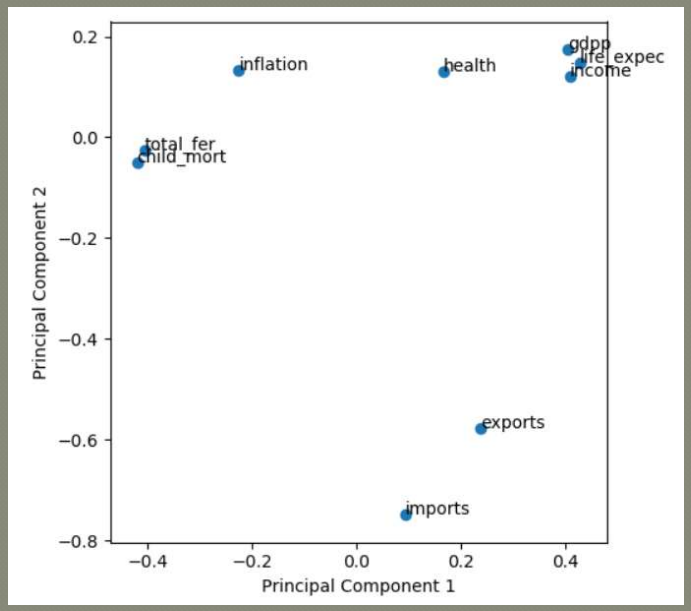
- There were outlier values in many of the variables. However dropping outliers would eliminate some countries from the dataset which might be in true need of aid.
- For the variables where outliers are present, we CAPPED the values above and below a certain pre-decided value.
- This pre-decided value was selected after analyzing the percentile distribution of the variable values.



Creation of Principal Components

- **We create the Scree Plot by plotting the cumulative variance of data against the number of Principal components.**
- **We choose 5 PCs which amounts to 95% cumulative variance in data and significant reduction in dimensionality.**





- **Visualizing the original features on the first 2 Principal Components.**

Hopkins Statistics & Cluster Tendency

- **We perform Hopkins Statistics Test to ensure that the given data has some meaningful clusters is not random.**

- **Hopkins test examines whether data points differ significantly from uniformly distributed data in the multidimensional space and whether it makes sense to create clustering.**

- **Values of Hopkins test range from 0 to 1.**
  - **If the value is between {0.01, …,0.3}, the data is regularly spaced.**
  - **If the value is around 0.5, it is random.**
  - **If the value is between {0.7, …, 0.99}, it has a high tendency to cluster.**

The Hopkins Test value for our dataset is:

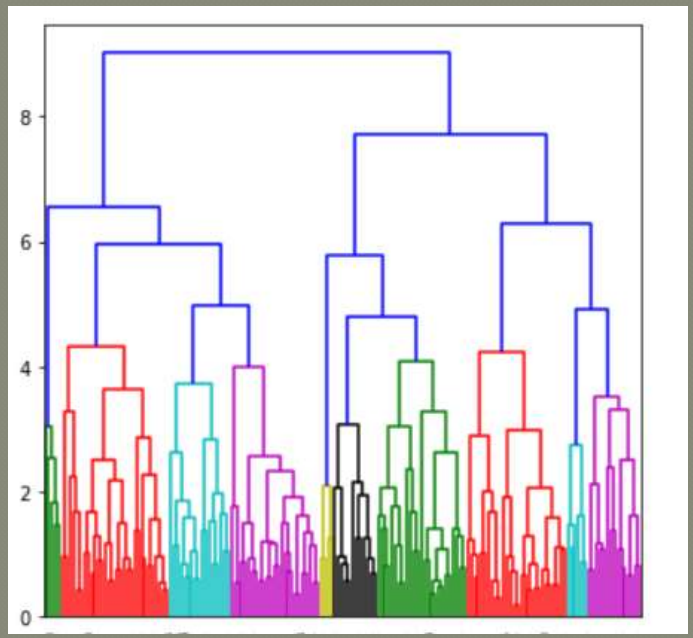
~ .71

Hierarchical Clustering

Hierarchical Clustering is performed in the beginning because it does not have the restriction of deciding the value of K(no of clusters) beforehand. Once we arrive at a satisfactory value, we will perform K-Means Clustering.

Model with 10 Clusters

- **The clustering process uses complete linkage to ensure the cluster are stable and close knit.**
- ***We can see from the 10 clusters created on the right that frequency distribution in the cluster is highly unequeal. The frequency is as high as 30 in one cluster and as low as 4 in another with lot of variation in the middle.***



Dendrogram to display the clusters

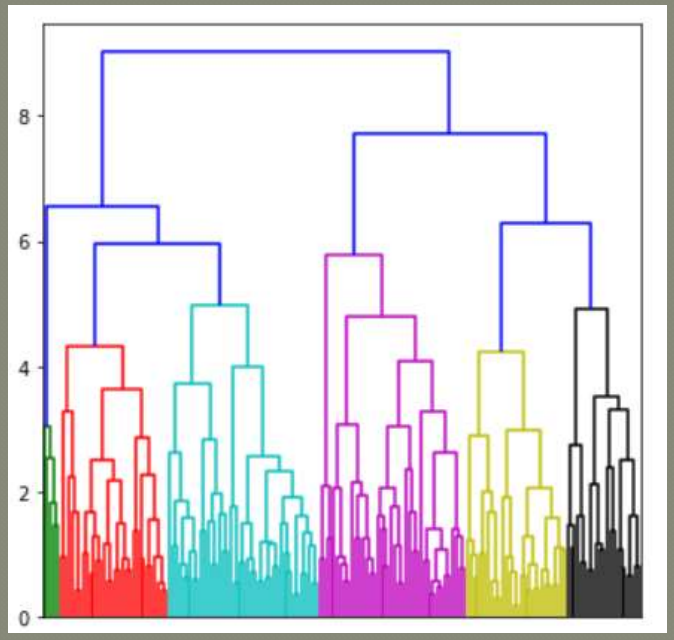
ClusterID	Country
0	30
1	25
2	17
3	5
4	25
5	28
6	15
7	12
8	6
9	4

Number of Countries per Cluster

Hierarchical Clustering contd…

Model with 6 Clusters

- **The clustering process uses complete linkage to ensure the cluster are stable and close knit.**
- ***Initially after creating 10 clusters, we try to see if we can further create clusters that will have more or less a comparative no of countries in each clusters. And also we can try create a lower no of clusters so that visually its easier to determine the countries that are in genuine need of aid.***



Dendrogram to display the clusters

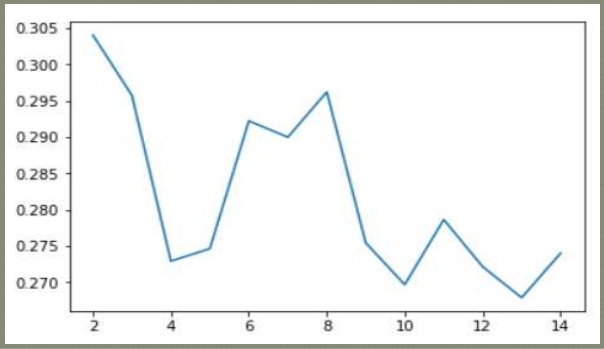
ClusterID	Country
0	30
1	42
2	5
3	41
4	28
5	21

Number of Countries per Cluster

K-Means Clustering

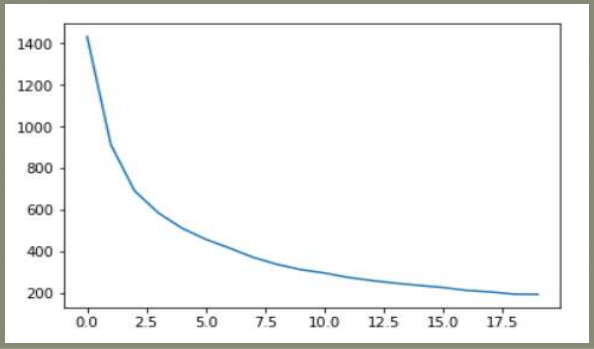
Silhouette Analysis

- From the Silhouette Graph, we see that the optimal no of clusters is 8 followed by 6.
- Thus we build 2 models with both these values separately.
- The one with 6 clusters appears to be more close-knit and stable. Also this is value that we got from our hierarchical model.
- Thus we decide to use 6 clusters finally.



Sum of Squared Distances

- From the Sum of Squared Distances, we see that the elbow in the curve occurs between 6 and 8 which gives an estimate of the optimal number K in K Means.
- This is in accordance with our decision to use 6 clusters for our model.

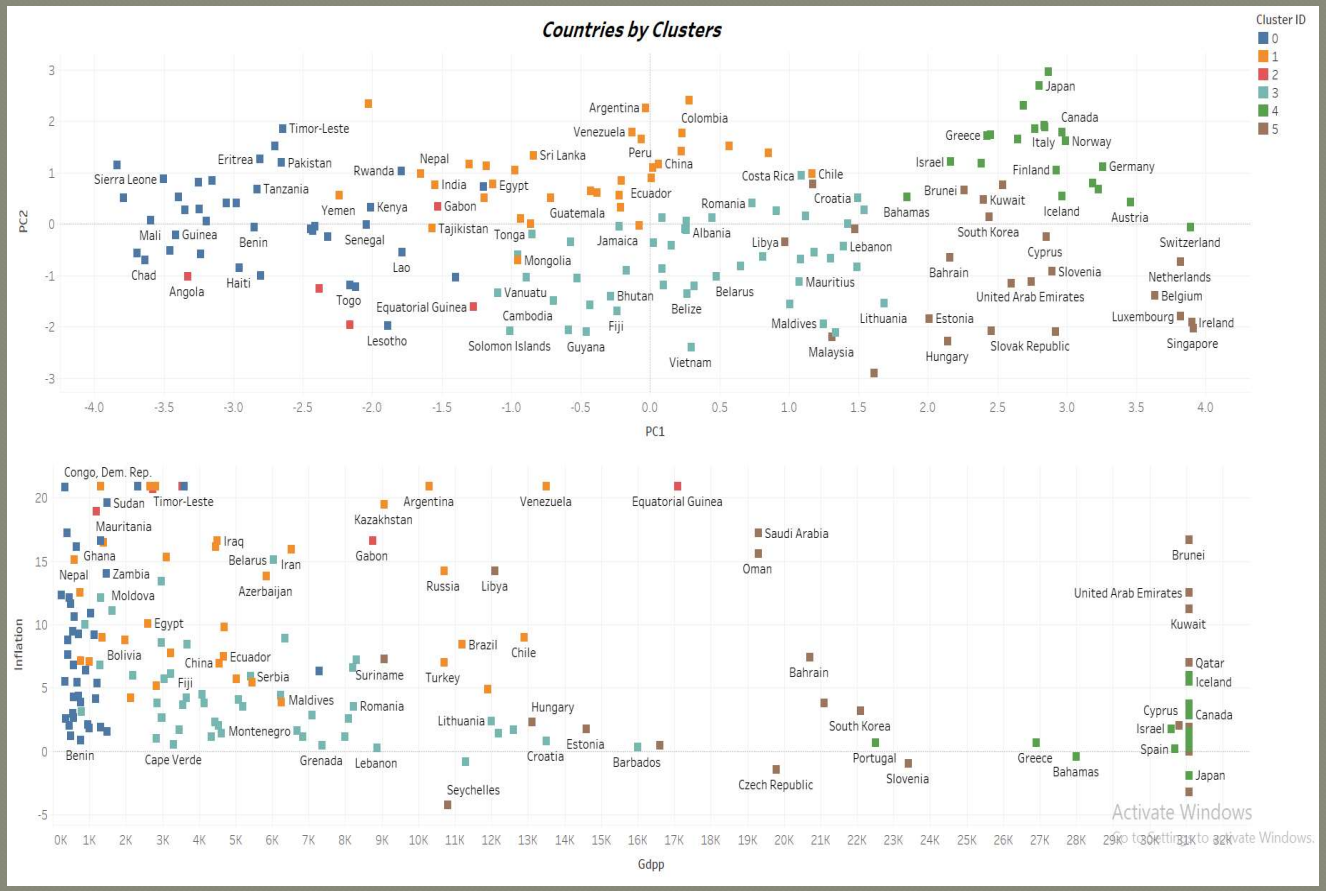


Using the value of 'K' & Centroids from Hierarchical model

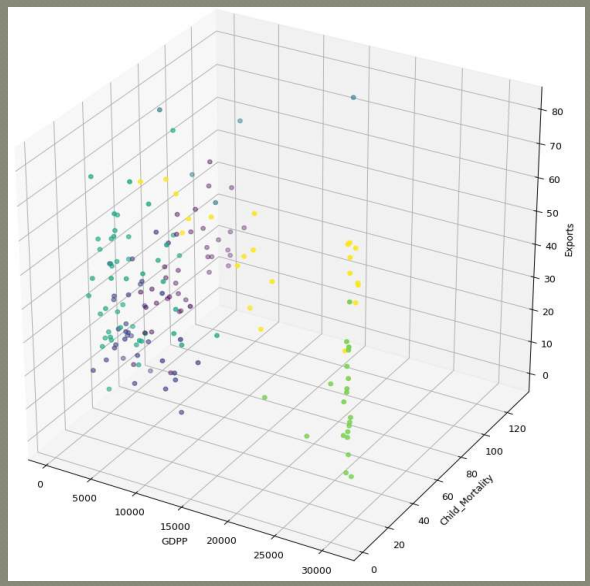
- *Choosing the initial value of K centroids can affect the K-Means algorithm and its final results.*
- *Here, we manually input the cluster centers as received from hierarchical clustering process into the K-Means process so that the K-Means algorithm considers these centers as the initial centroid locations.*
- *Instead of using random centroids, using properly tested centroid locations will improve the efficiency of the clustering process.*

ClusterID	Country
0	38
1	33
2	5
3	47
4	21
5	23

2D & 3D Visualization of the Clusters



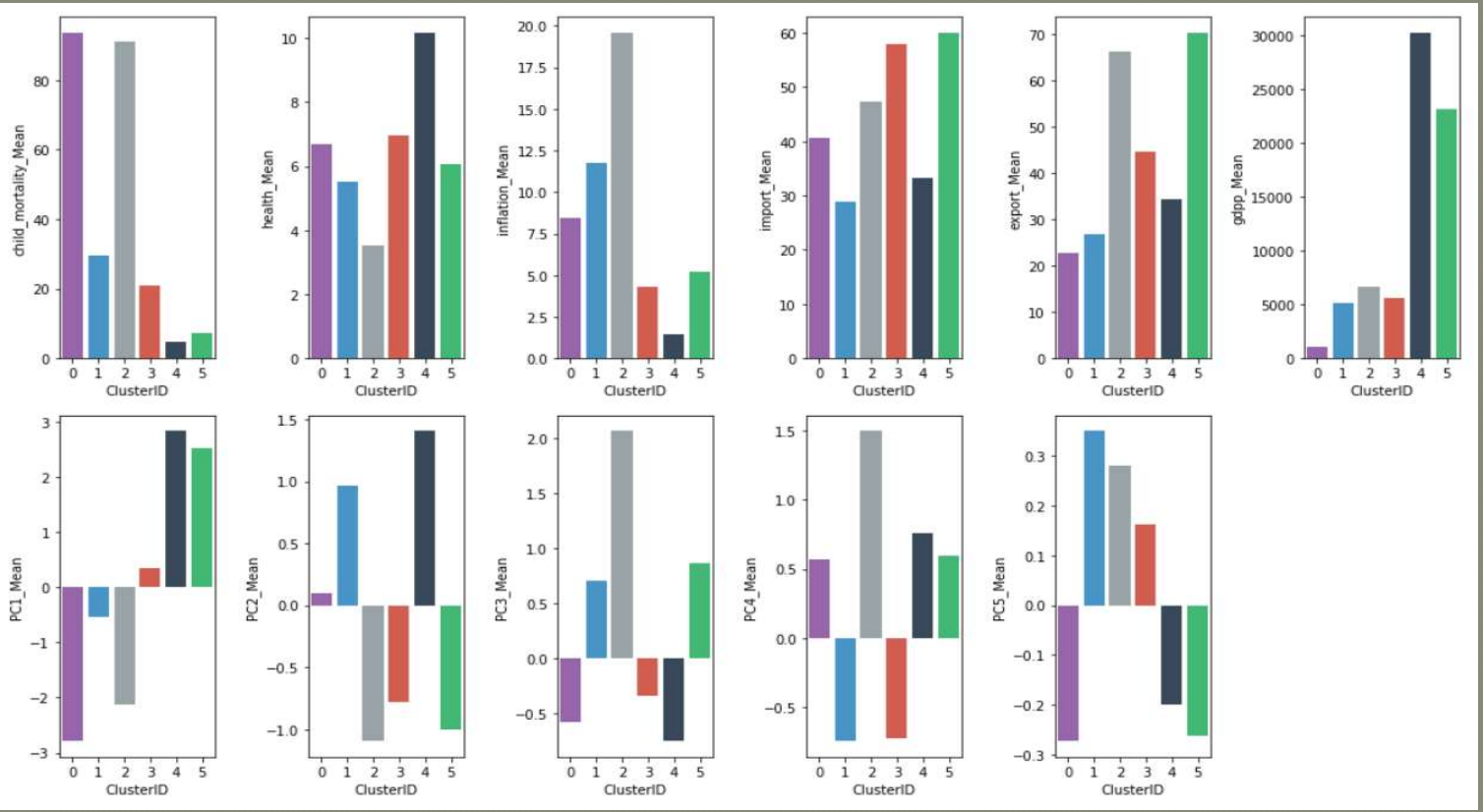
Countries are being visualized on the right based on the first 2 Principal Components and 2 of the given features - GDPP & Inflation.



•3D representation of Countries segmented into clusters based on 3 of the given features - GDPP, Child Mortality & Exports.

Comparing the Clusters

Plotting the mean values for PCs and other socio-economic factors per cluster.



Observations

- Countries belonging to Clusters '0' and '2' have high child mortality rates for children under 5 years of age.
- Cluster '0' has very low gdpp value followed by clusters '1', '2' and '3'.
- Total spending on health is poorest for countries belonging to cluster '2'.
- Cluster 4 has countries with the strongest socio-economic and health status like USA, the European giants -UK, Germany, France and some Asian leaders like Japan, Israel etc.

Identifying Countries in Need of Aid

Mean of Principal Component values and original variables for each of the 6 clusters created.

ClusterID	child_mortality_Mean	health_Mean	inflation_Mean	import_Mean	export_Mean	gdpp_Mean	PC1_Mean	PC2_Mean	PC3_Mean	PC4_Mean	PC5_Mean
0	93.66	6.68	8.42	40.55	22.83	1044.26	-2.79	0.10	-0.58	0.57	-0.27
1	29.47	5.53	11.74	28.96	26.78	5171.70	-0.53	0.96	0.70	-0.75	0.35
2	91.00	3.54	19.59	47.32	66.37	6664.00	-2.14	-1.10	2.07	1.50	0.28
3	20.86	6.94	4.29	58.04	44.77	5644.17	0.35	-0.79	-0.34	-0.73	0.16
4	4.73	10.17	1.47	33.27	34.30	30292.38	2.84	1.41	-0.75	0.76	-0.20
5	7.44	6.05	5.18	60.02	70.36	23155.65	2.52	-1.00	0.86	0.60	-0.26

Identifying countries in need of Financial Aid from the clusters:

Cluster# 0

- **Highest child mortality** rate.
- **Lowest GDPP** value.
- **38** countries in this cluster are mostly **African** countries like -
- **Cameroon, Haiti, Kenya, Madagascar, Nigeria, Rwanda, Senegal, Sudan, South Africa, etc**
- and some occasional **Asian** countries like -
- **Pakistan and Afghanistan.**

Cluster# 2

- **Lowest total spending on Health**
- **2nd highest child mortality** rate
- **3rd lowest GDPP** value
- Countries belonging to this cluster are **all 5 African** countries. The 5 countries that belong to this cluster are:
  - **Angola, Congo Rep.,Equatorial Guinea, Gabon, Mauritania**