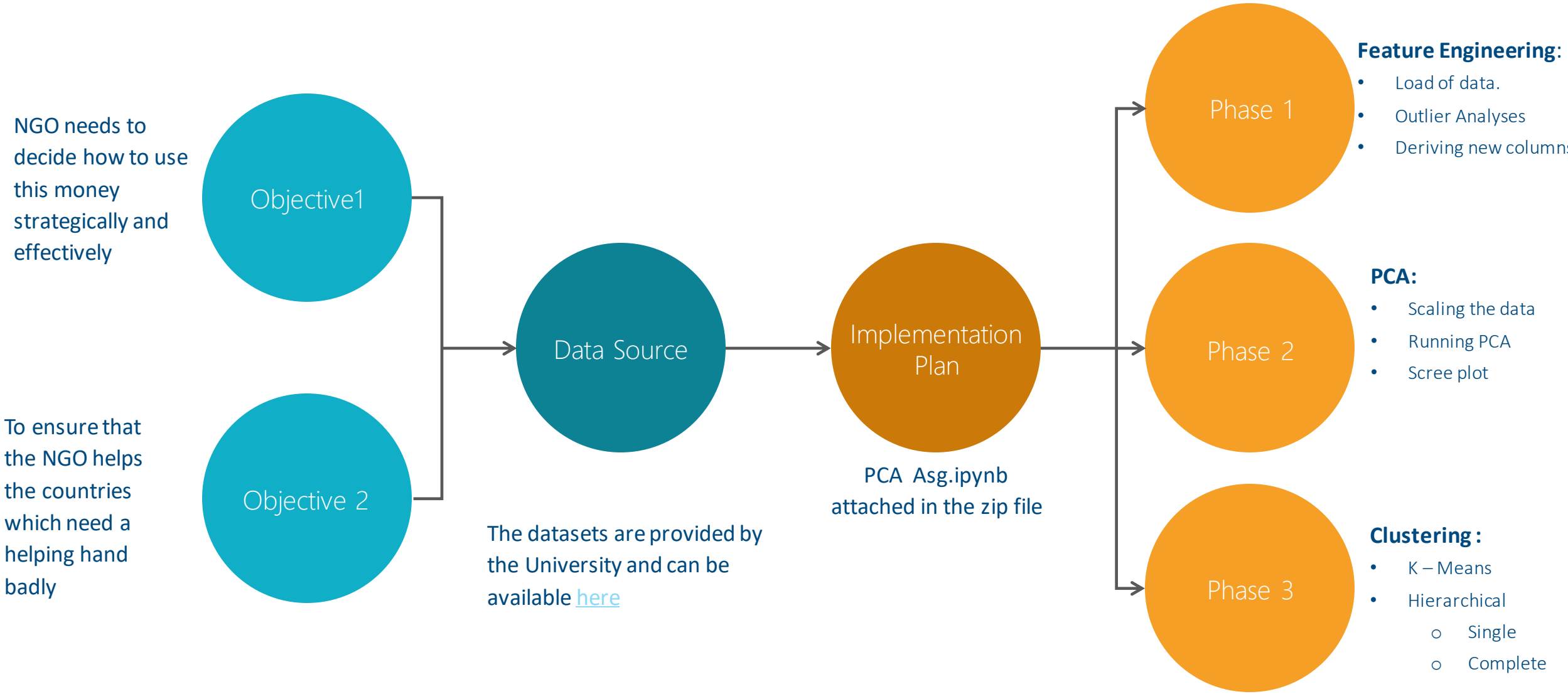
Help International Countries - Clustering

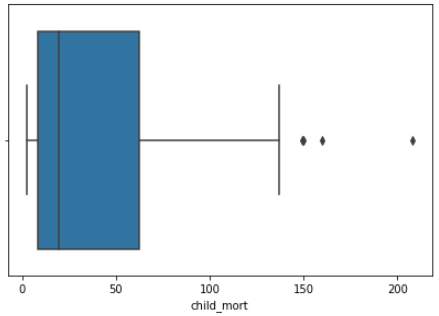
V V SATYA SAI KIRAN KUSUMANCHI

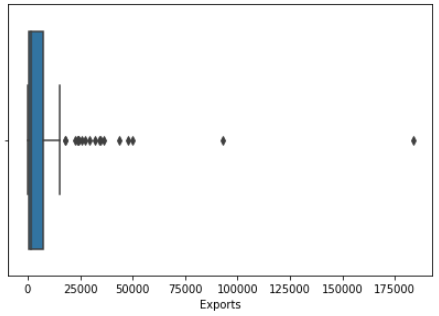
Help International Countries - Clustering

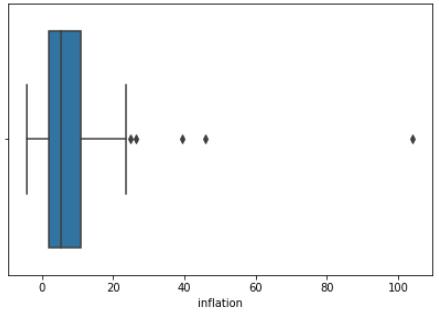


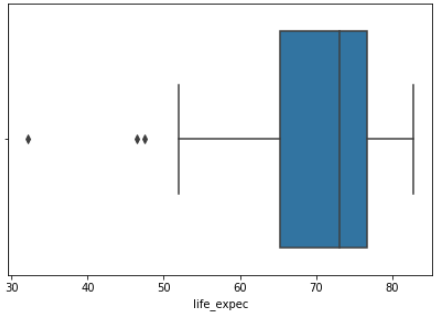
Phase 1: Feature Engineering

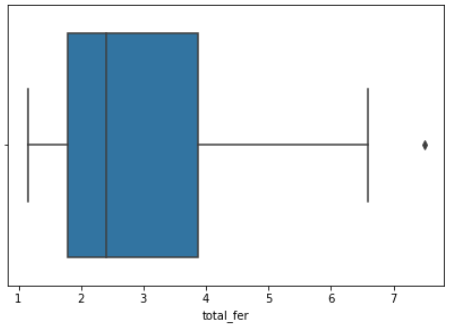
- **Null value checks:** Checked for the NULL values. No Nulls present
- **Outlier Analyses:** Outliers are present in the GDPP, but they are on the higher end of the spectrum i.e., they would be clustered into a single cluster probably by the algorithms
- **Derived Columns: '**exports', 'health', 'imports' are calculated as percentage of the GDP. GDP is not present in the dataset. so we can consider them as the % of GDPP and proceed to calculate the absolute values
  - The effect of the deriving the absolute metrics shall be minimal anyway because we will scale the data to get every feature in relatively same basis

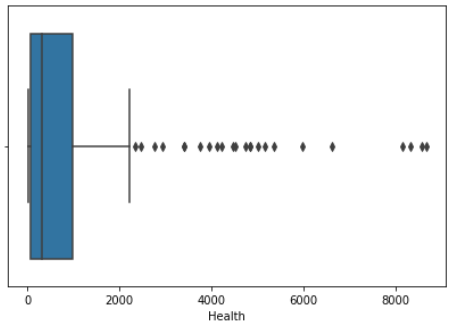


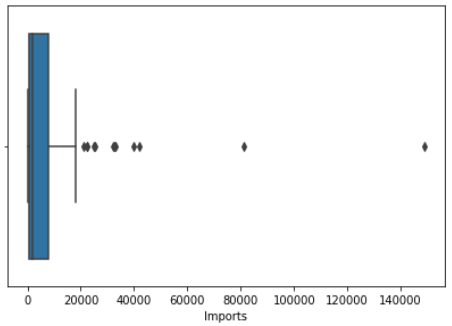


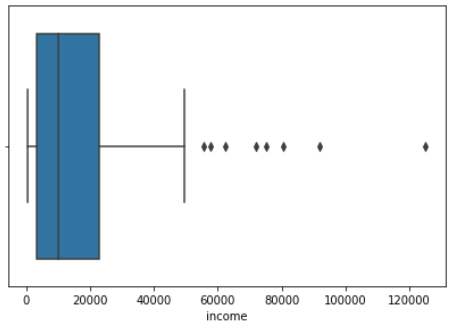






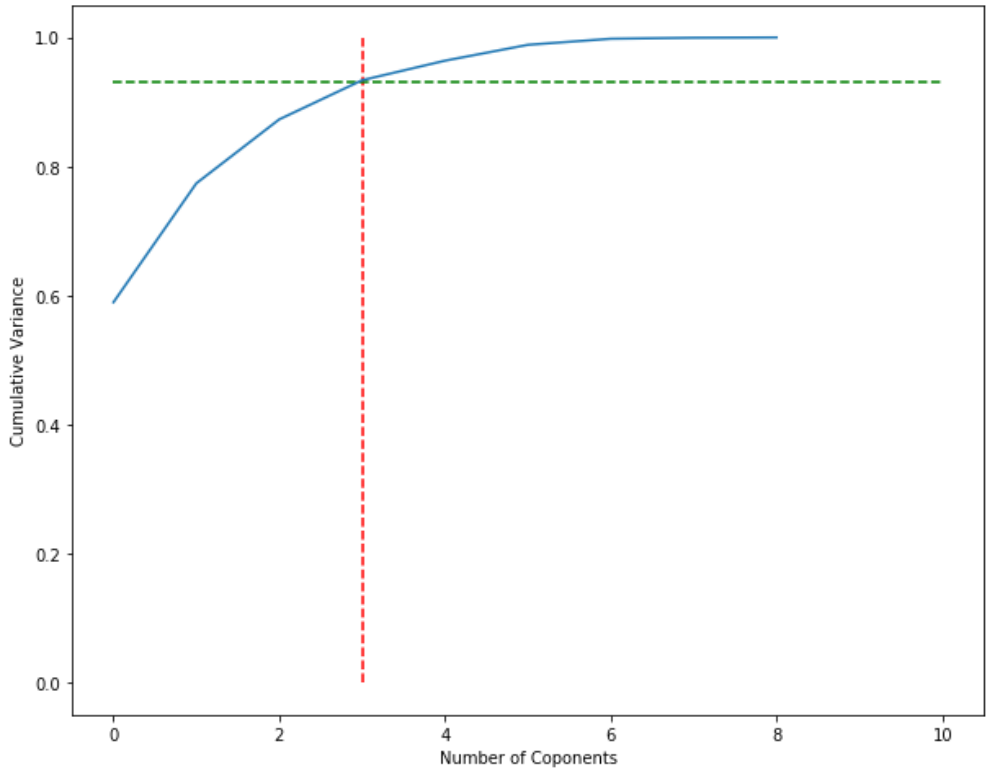






Phase 2: PCA

Scree plot : Number of Components vs. Cumulative variance

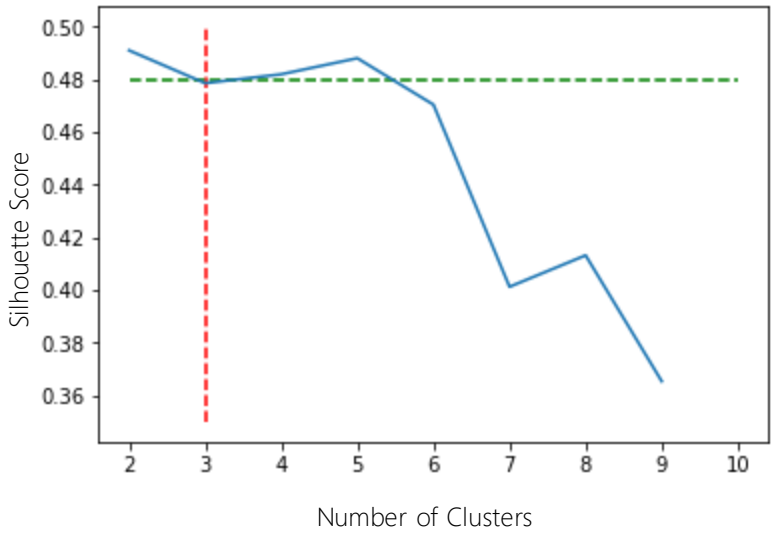


Observations:

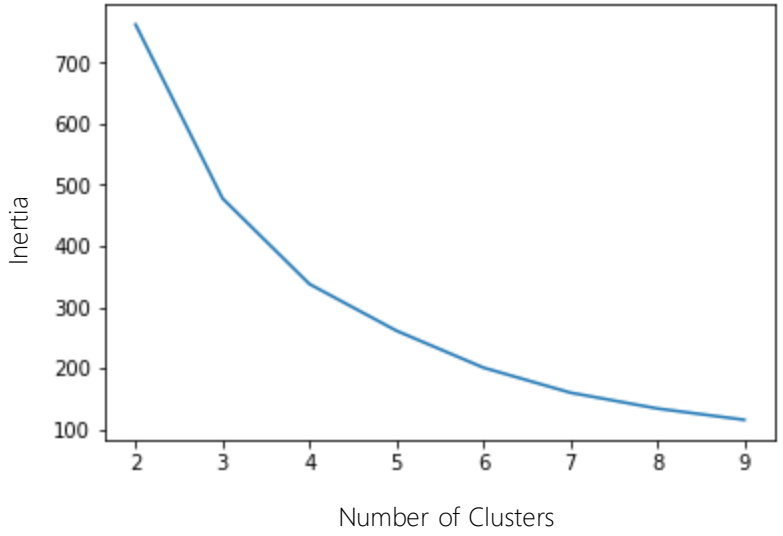
- From the original data set the variance is being explained by the following features:
  - income, gdpp are the main columns which has high variance in the data
- Scaling of the data: Standard Scaling technique.
- PCA: using svd_solver (randomized Algorithm), random_state = 42 for consistency of results
- Scree Plot: Fig on left has # of components on X-axis and Cumulative Variance on Y-Axis
  - the lines indicate that the at 3 principle components could explain total of 93 % variance in the data.

Phase 3: Clustering (K-Means)

Silhouette Analysis



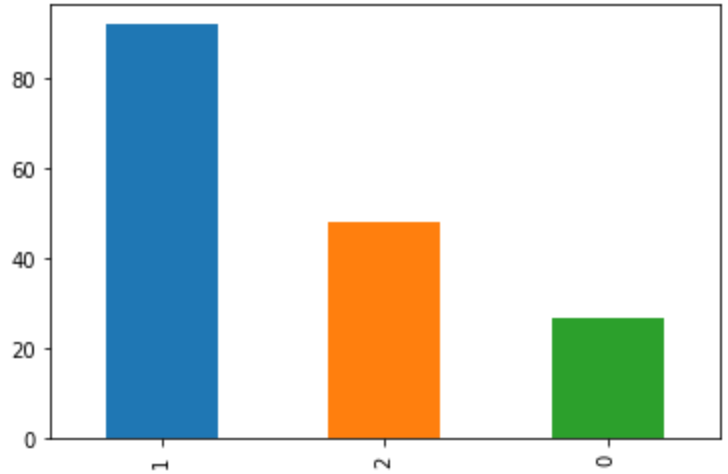
Elbow Plot



Observations

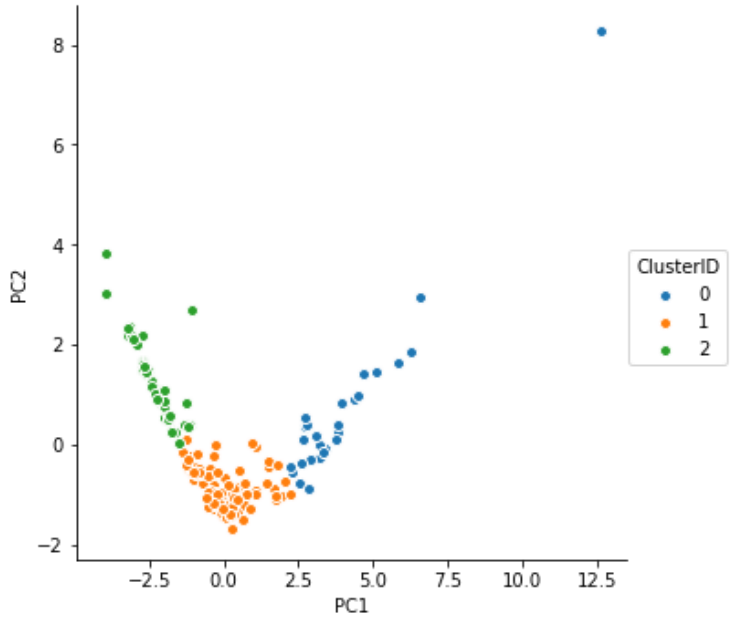
- From the Silhoute Analysis we can see that the at the 3 number of cluster we can see that the silhouette score is 0.48.
- From Elbow Plot the We can see that the Slope of the curve suddenly fall at 3 number of clusters
- So we can see that the optimal number of cluster to consider is 3 by using K-Means clustering algorithm

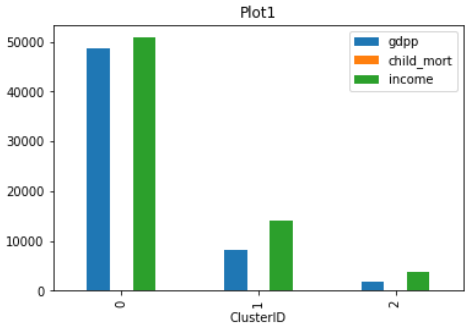
of Countries Segmented thru Clusters

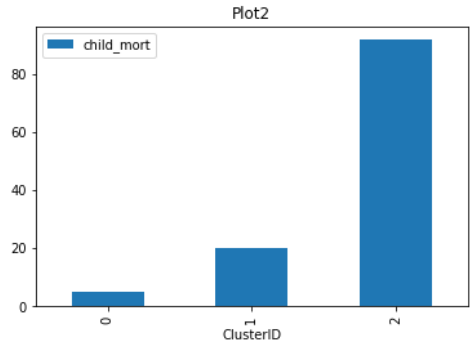


Phase 3: Clustering (K-Means)

K Means Cluster – Profiling & Analysis





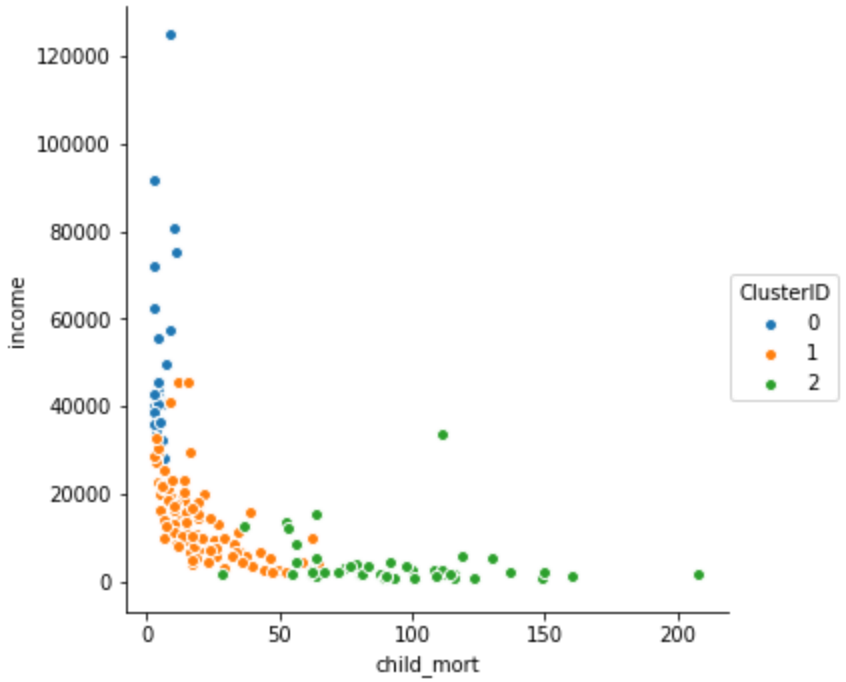


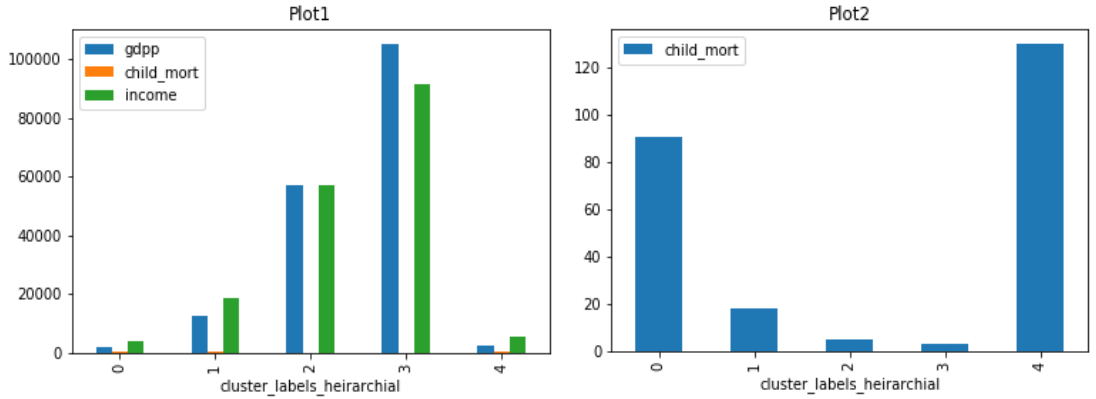
Observations

- From the beside plots : Plot 1 : As the Child mortality is very less compared to the income and the gdpp we were not able to see the representation of the child mortality, so plotted the average values of the child mortality in the plot2 with respect to ClusterID
- This clearly explain how the clusters are divided by the K-Means Clustering:
  - Cluster 0 : These are the Rich/ High Income(/gdpp) countries with almost less than 10 children dying within the 5 years of birth. This indicates the reach of health and food facilities in the cluster 0 countries
  - Cluster 1 : These re the middle/ Avg Income(/gdpp) countries with on average 20+ children dies within the 5 years of birth. The health and food facilities could be below than that of the Rich countries
  - Cluster 2 : These are the poor countries with very low Income and very low gdpp relative to the countries in cluster 0 & 1. With the child mortality more than 80/90. these are the countries which could be having the least reach of health, sanitation and food facilities

Phase 3: Clustering (Hierarchical)

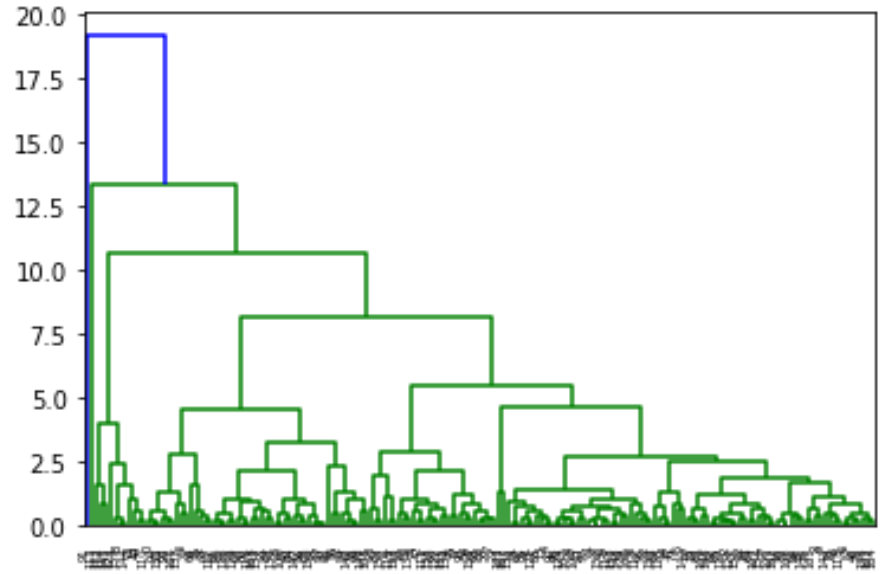
Hierarchical Cluster – Profiling & Analysis





Observations

- We have clustered the data into 5 clusters, here are the observations from the clusters
- The Clusters 0 and 4 are having the least income and gdpp, relatively high child mortality
- The Clusters 2 and 3 are having the better income and gdpp and relatively low child mortality
- The cluster 1 looks like mostly the middle income, gdpp and considerable child mortality countries
- NGO can focus on the cluster 0 and 4.



Inferences

In both Clustering techniques we have decided which clusters should NGOP focus on. Considering the socio-economic , Health Spend and other factors we have prepared the list and below is the observation:

If we observe one thing here all the top 10 countries, we have decided that NGO should focus on based on the K-Means Clustering are also present in the Hierarchical Clustering cluster also(all in cluster 0). Here are the list of the countries for reference again:

- Burundi
- Liberia
- Congo, Dem. Rep.
- Niger
- Sierra Leone
- Madagascar
- Mozambique
- Central African Republic
- Malawi
- Eritrea

Thank You