Clustering Assignment

By Darshan Patel

Abstract

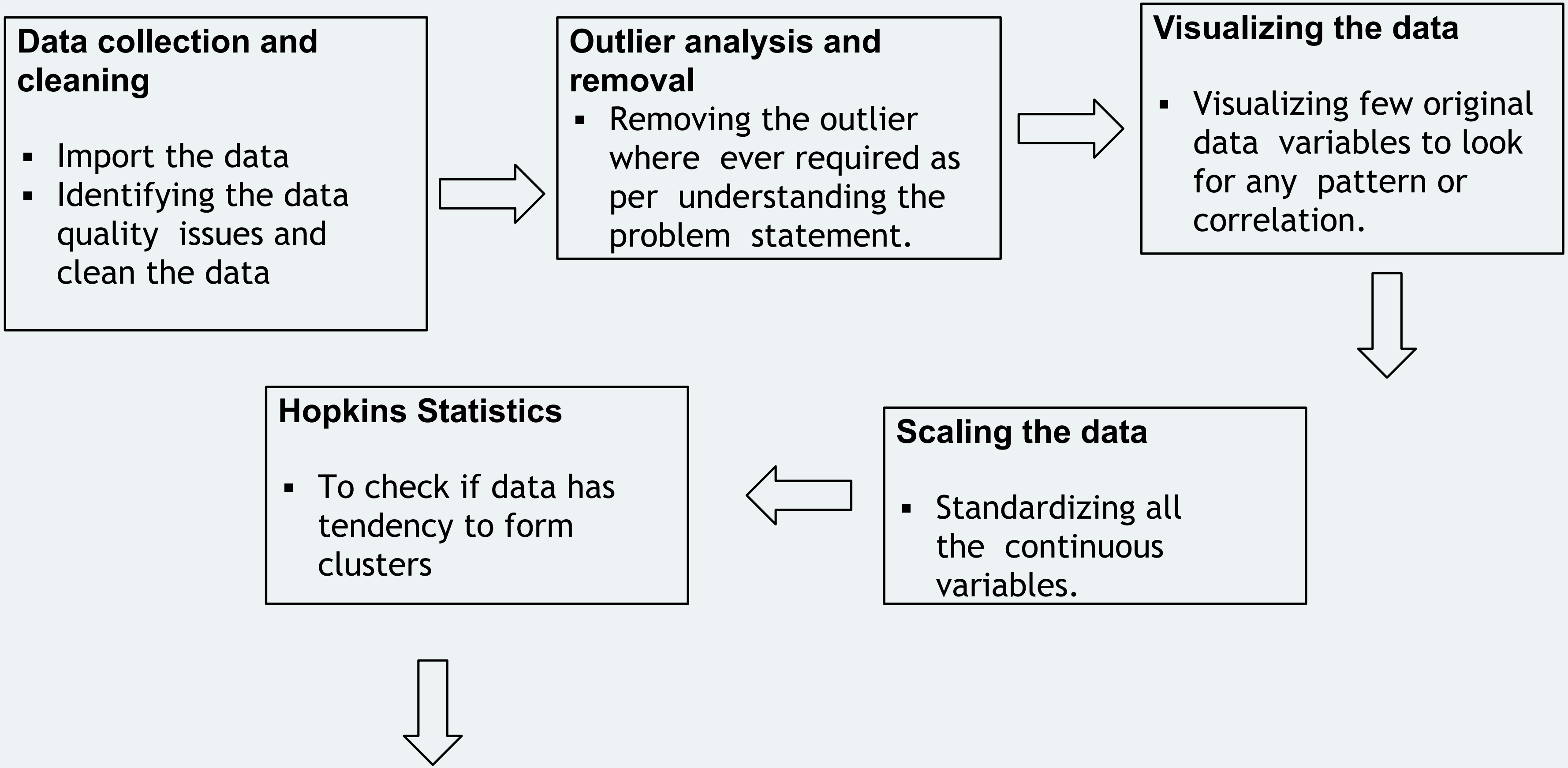
Objective:

We, HELP International humanitarian NGO, committed to fight poverty and provide the people of backward countries with basic amenities and relief during the time of disasters and natural calamities. We run a lot of operational projects from time to time, along with advocacy, drives to raise awareness as well as for funding purposes.

Problem statement:

During the recent funding programmes, we have been able to raise around $ 10 million. As an analyst, we have to come up with the countries list that are in the direst need of aid.

Analysis methodology



Data collection and cleaning

- Import the data
- Identifying the data quality issues and clean the data

Outlier analysis and removal

- Removing the outlier where ever required as per understanding the problem statement.

Visualizing the data

- Visualizing few original data variables to look for any pattern or correlation.

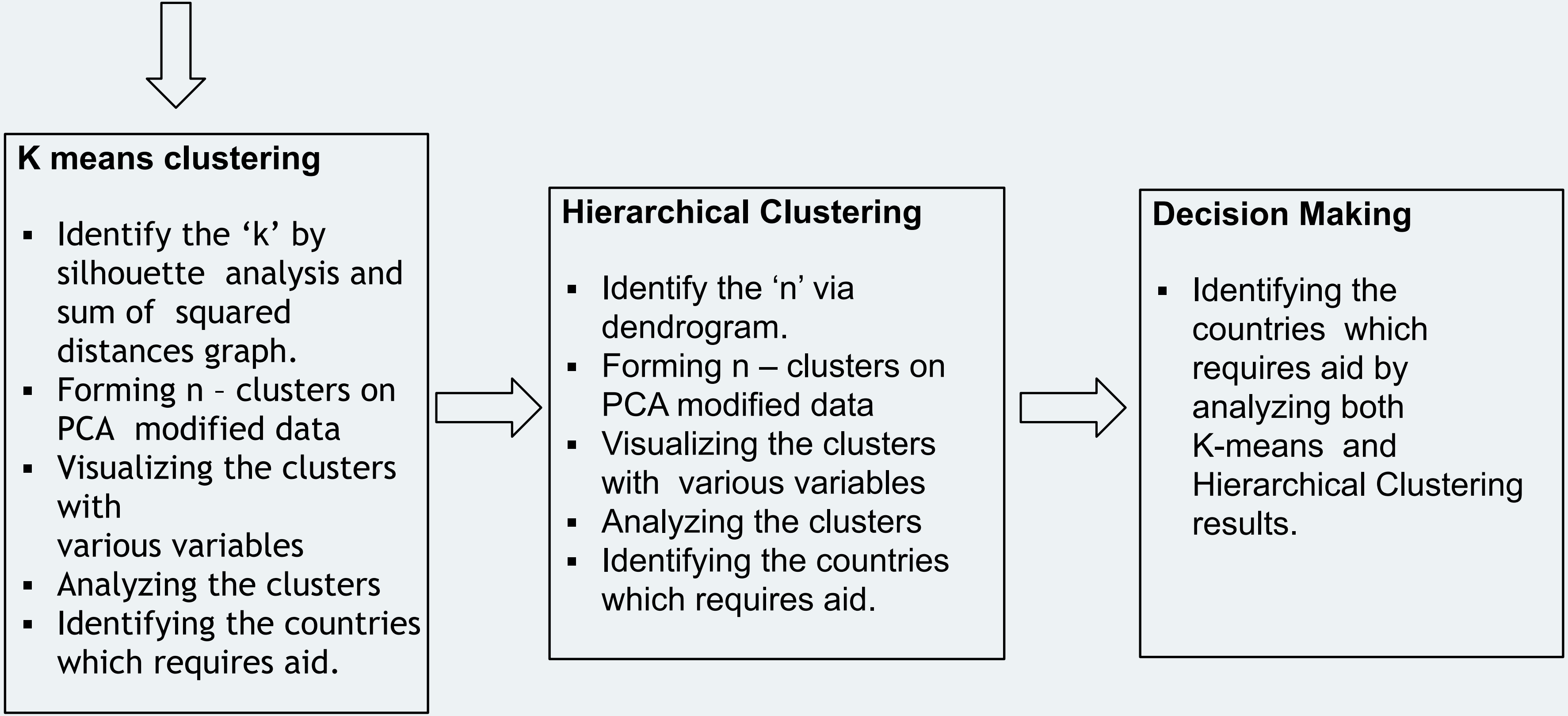
Scaling the data

- Standardizing all the continuous variables.

Hopkins Statistics

- To check if data has tendency to form clusters

Analysis methodology Cont…



K means clustering

- Identify the ‘k’ by silhouette analysis and sum of squared distances graph.
- Forming n – clusters on PCA modified data
- Visualizing the clusters with
  various variables
- Analyzing the clusters
- Identifying the countries which requires aid.

Hierarchical Clustering

- Identify the ‘n’ via dendrogram.
- Forming n – clusters on PCA modified data
- Visualizing the clusters with various variables
- Analyzing the clusters
- Identifying the countries which requires aid.

Decision Making

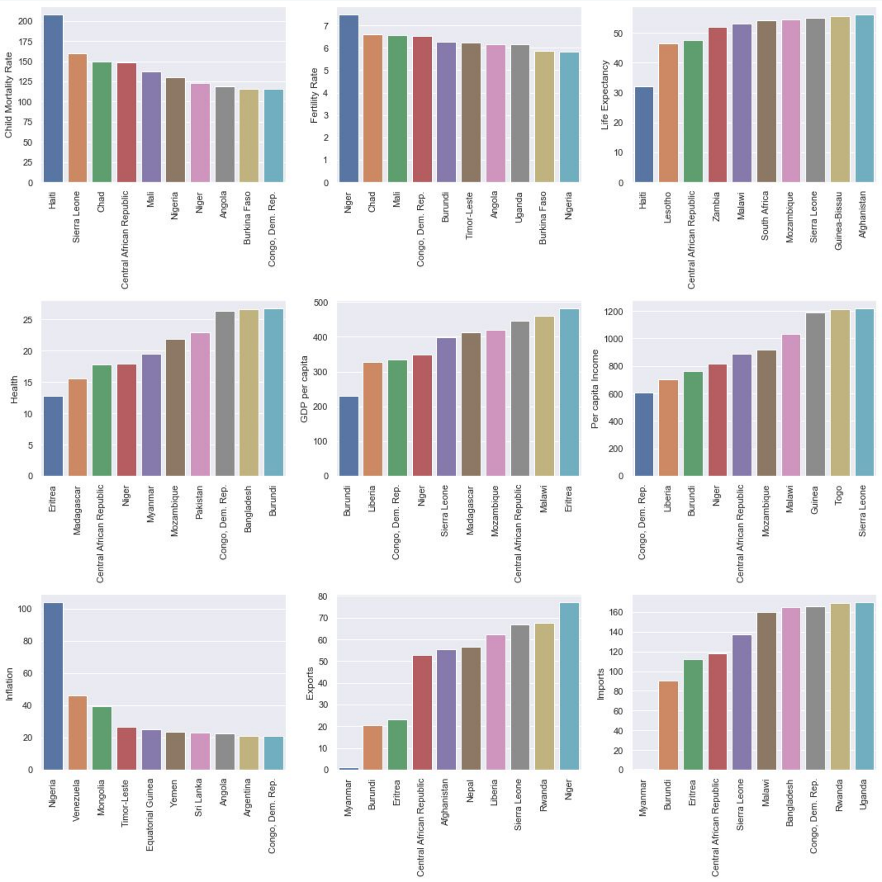
- Identifying the countries which requires aid by analyzing both K-means and Hierarchical Clustering results.

Univariate Analysis

- After data cleaning , we remove outlier from gdpp column because the country with high gdpp would not require any aid as there are already doing good.

- The first barplot based on **Child Mortality** which shows **Haiti** has highest Child Mortality whereas Haiti has **lowest life expectancy rate** in the world.

- From above bar plot we could see the common countries in the profile of **gdpp**, **child_mort** and **income** are: **Congo**, **Dem. Rep**., **Burundi**, **Nige**r,**Central African Republic**



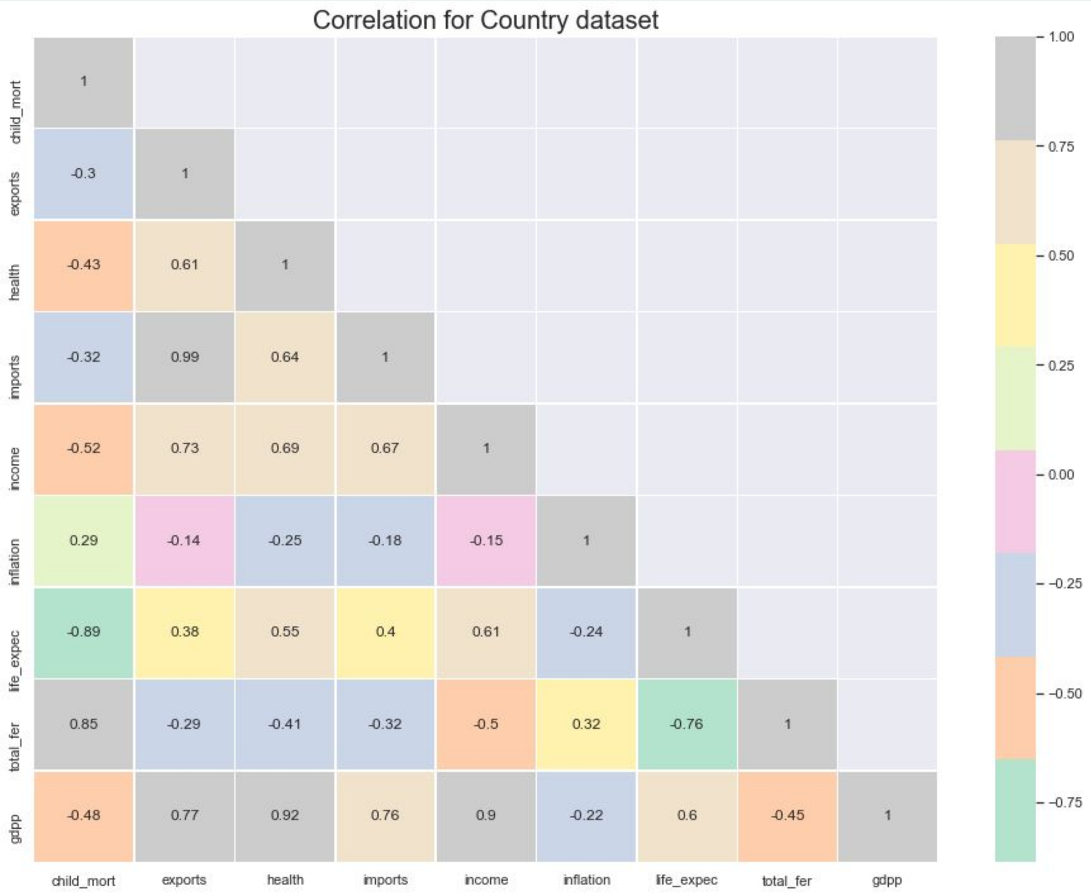
Bivariate Analysis and Outlier Analysis

- Looking at the heatmap, we see that few variables like (total fertility, child mortality) , (income , gdpp) and (imports and exports) have high correlation.

- **child_mort, inflation:** High Child mortality and higher inflation are matter of concern so we will not doing treatment for these features.

- **gdpp** : It has outlier at higher level. We will compute outlier by performing Interquartile Range (0.99 percentile)

- **life_expec** : It has outliers below the lower hinge, But again it is our matter of concern so we will not impute these values.

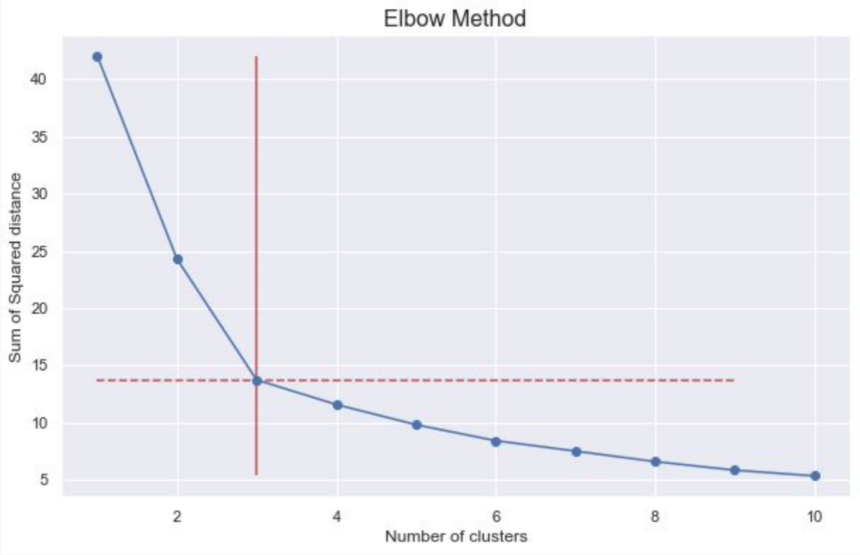


Correlation for Country dataset

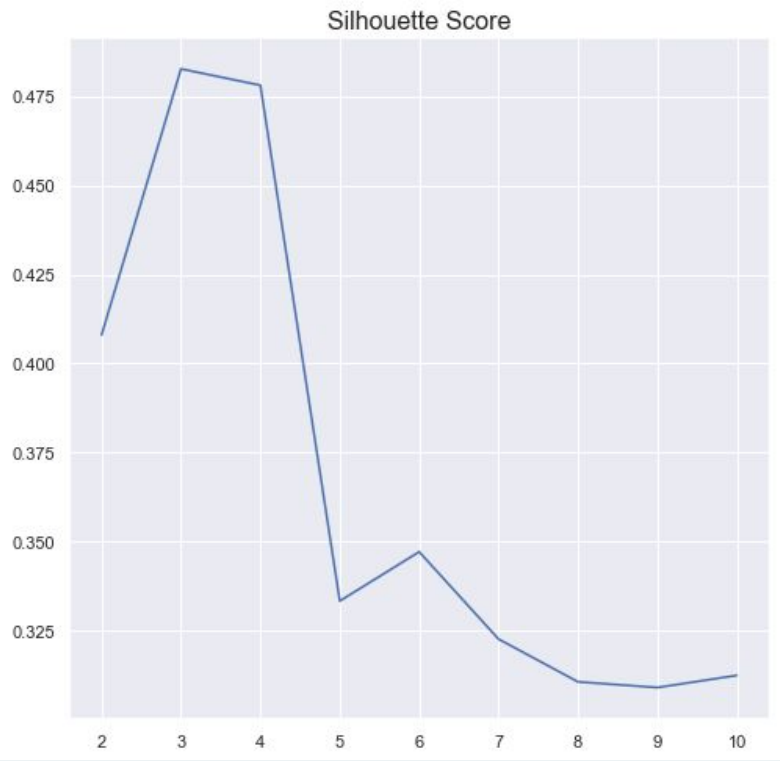
	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp
child_mort	1								
exports	-0.3	1							
health	-0.43	0.61	1						
imports	-0.32	0.99	0.64	1					
income	-0.52	0.73	0.69	0.67	1				
inflation	0.29	-0.14	-0.25	-0.18	-0.15	1			
life_expec	-0.89	0.38	0.55	0.4	0.61	-0.24	1		
total_fer	0.85	-0.29	-0.41	-0.32	-0.5	0.32	-0.76	1	
gdpp	-0.48	0.77	0.92	0.76	0.9	-0.22	0.6	-0.45	1

K-Means number of cluster Indicators

Elbow curve

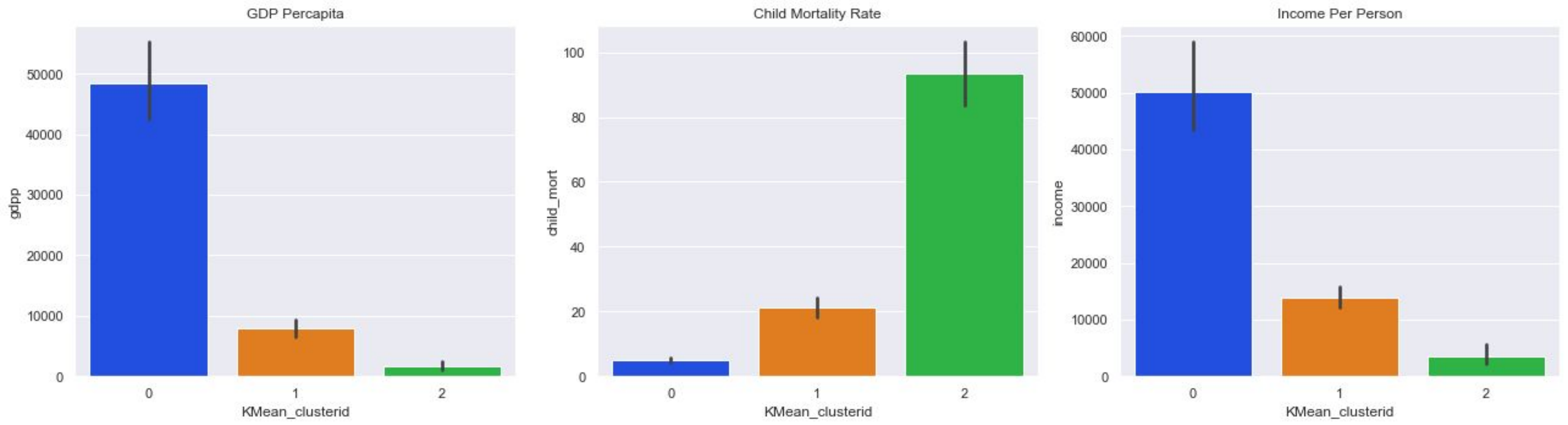


Silhouette Analysis



By looking **silhouette analysis,** we see the highest peak is at **k =3** and in sum of squared distances graph , we see that the **elbow curve** is at **3** , so we are going ahead with **k as 3.**

K-means clustering



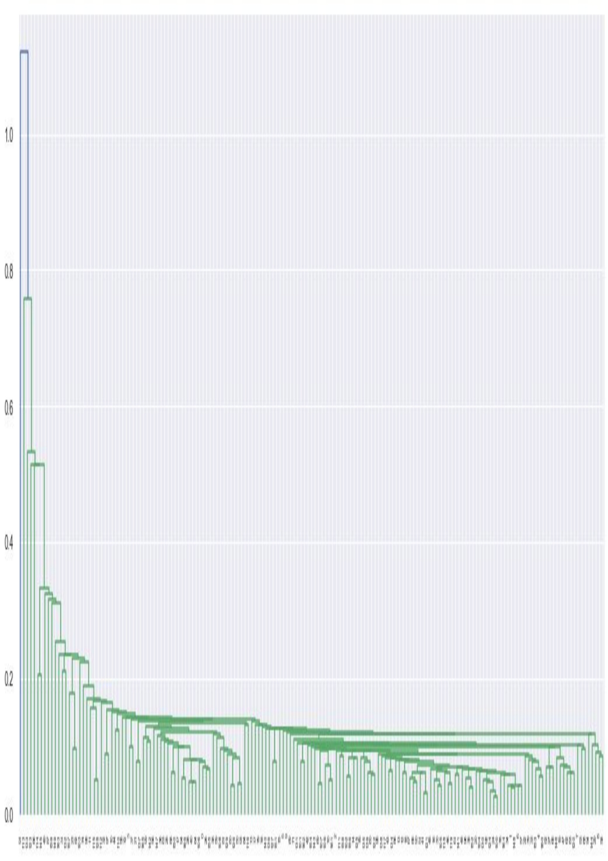
Interpretation of Clusters:
- Cluster 2 has the Highest average Child Mortality rate of ~90 when compared to other clusters, and Lowest average GDPP & Income of ~ 5000 & 9000 respectively.
- All these figures clearly makes this cluster the best candidate for the financial aid from NGO. We could also see that Cluster 2 comprises of ~28% of overall data, and has ~48 observations in comparison to 167 total observations

K-means clustering

5 Countries that are in dire need of aid are :

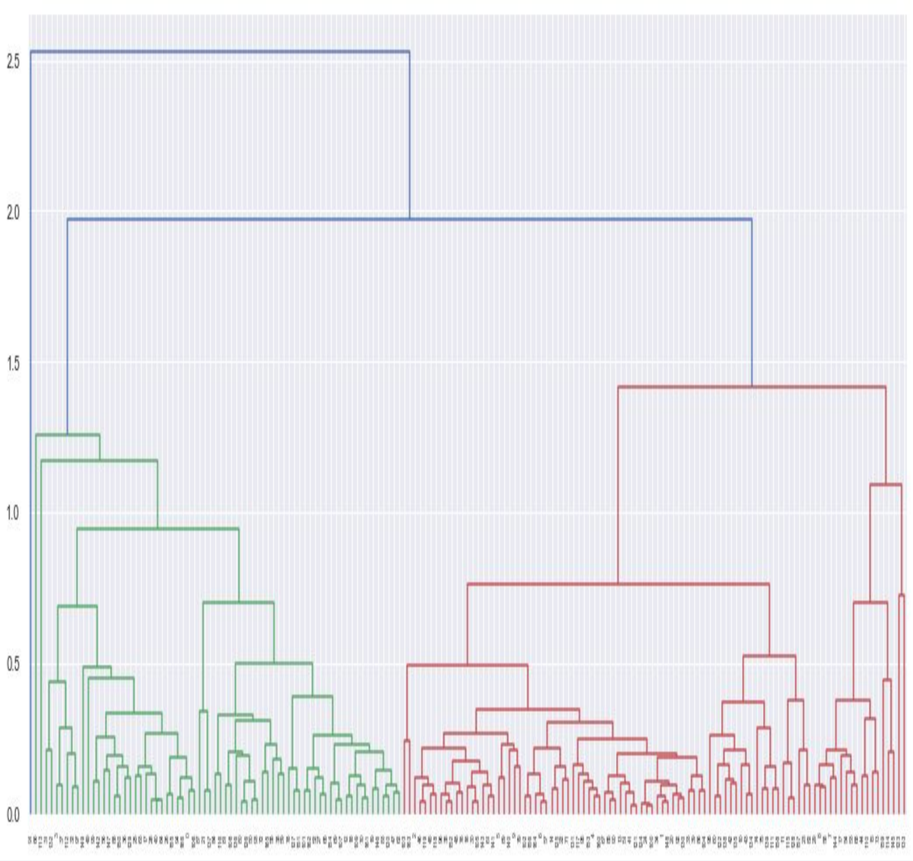
- Haiti
- Sierra Leone
- Chad
- Central African Republic
- Mali

Hierarchical Clustering

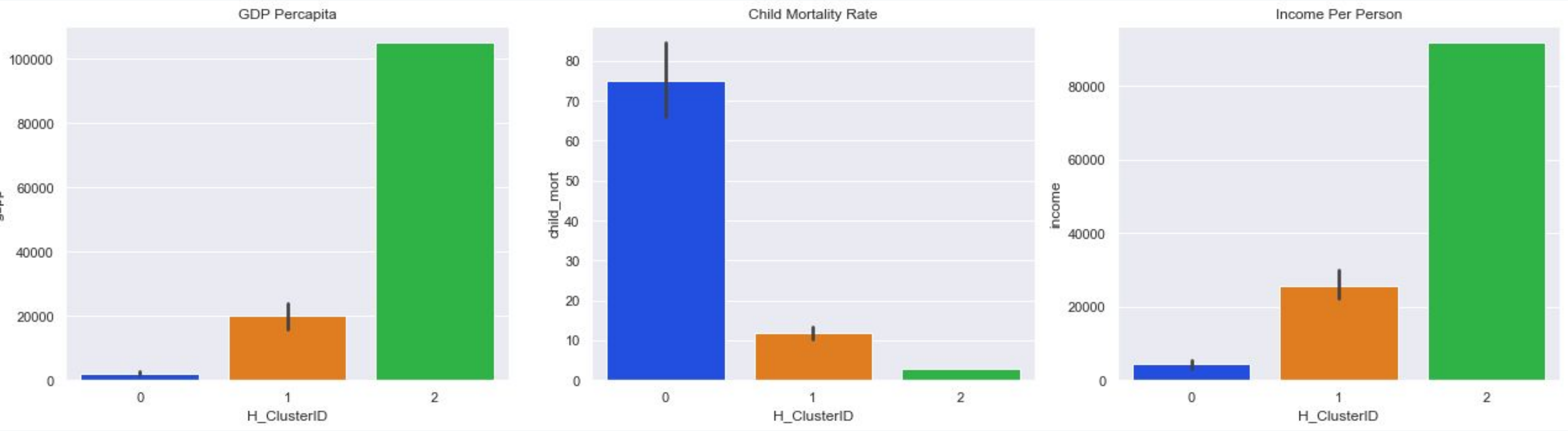


Single method hierarchical clustering

We are going for **Complete method hierarchical clustering** as single method clustering is not clear. By looking at this dendrogram taking n-clusters as 3

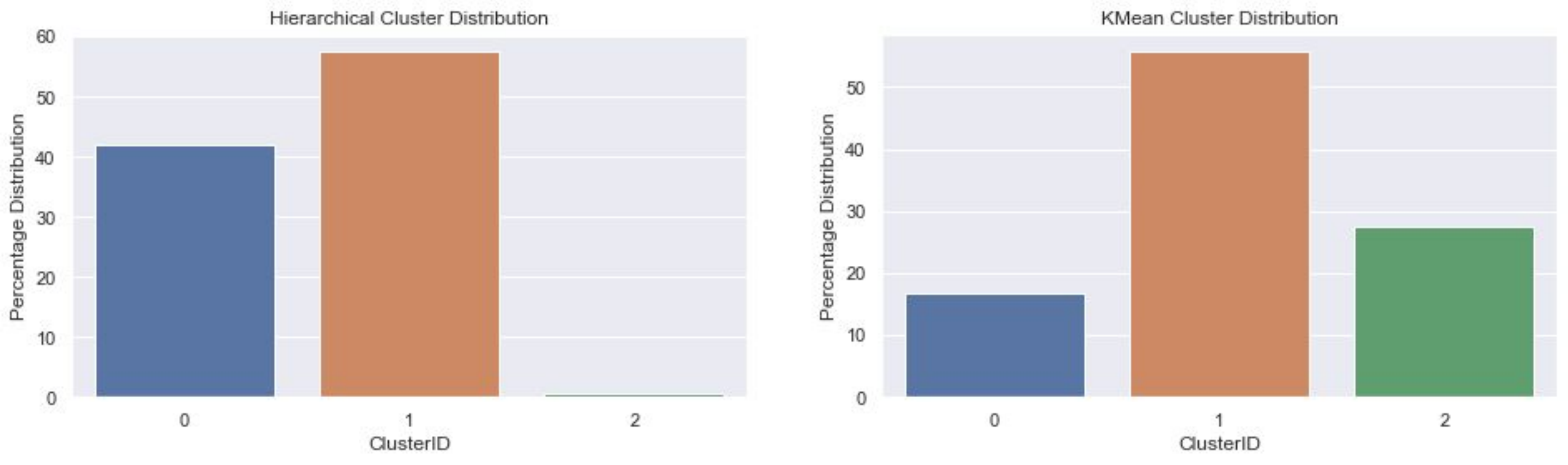


Interpretation of Clusters:



- Cluster 0 has the Highest average Child Mortality rate of ~75 when compared to other 3 clusters, and Lowest average GDPP & Income of ~ 7551 & 12641 respectively.
- All these figures clearly makes this cluster the best candidate for the financial aid from NGO.
- We could also see that Cluster 1 comprises of ~89% of overall data, and has ~148 observations in comparison to 167 total observations This seems to be a problem.
- This means that Hierarchical clustering is not giving us a good result as 89% of the data points are segmented into that cluster.

Clustering Model Selection



- We also saw that increasing the cluster number is not solving this problem. We will perform K-Means Clustering and check how that turns out to be.

- From above analysis we could see KMean is having better distributed cluster. So we will select final model as KMean cluster and doing profiling considering the labels accordingly.

Summary

- Performed CLUSTERING on the socio-economic data provided for various countries to identify countries to recommend for Financial Aid from the NGO.

- Based on our Clustering Analysis, I have identified the top countries under our 'Undeveloped Countries' cluster which are in dire need of the Financial Aid. This output is purely based on the dataset we used and various analytical methodology we performed.

- There are some countries which spend well on health for the people living in that country. For ex: US. Such countries can be skipped. And focus more on Haiti, Sierra Leone, Central African Republic and Mali where the total health spending is too less.

Financial Aid required countries on priority basis:

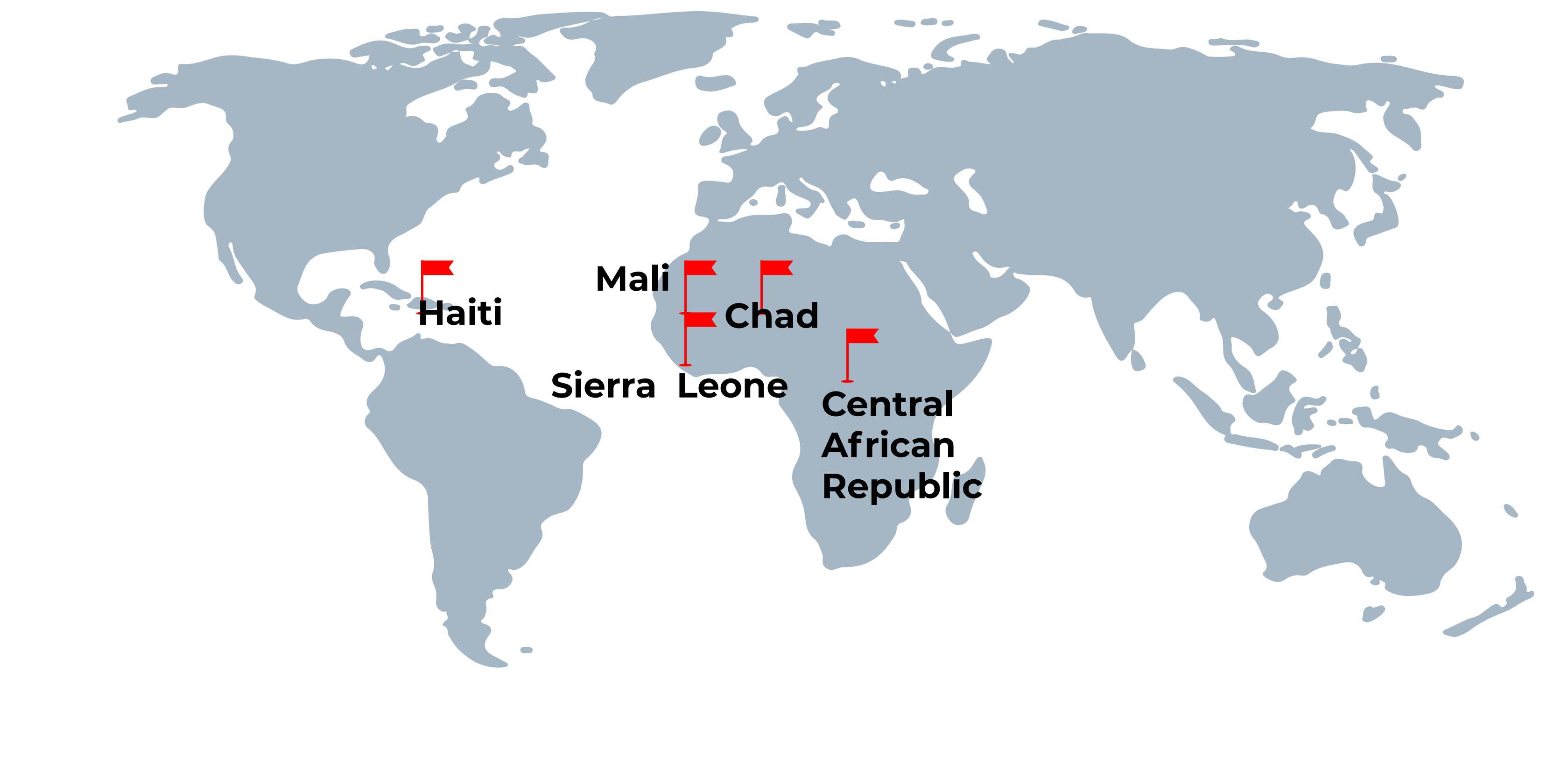
1. **Haiti**
2. **Sierra Leone**
3. **Chad**
4. **the Central African Republic**
5. **Mali**

Haiti As from our Analysis we get to know that Haiti has a low Life expectancy, High Child Mortality, Less fertility these factors implicate that population of Haiti is shrinking and Haiti is in dire need of aid to balance out the population and have access to the healthcare system.

Sierra Leone has the second-highest Child Mortality rate within African nations. With high imports and fewer exports imply that the country is heavily dependent on goods and service coming from other countries. low income and GDP of the country makes Sierra Leone more prone for the need for aid.

C**entral African Republic** with above-average fertility rate we tend to see that it has third highest child mortality rate in the world which implies that even with high fertility rate and low health rate .people of central African republic do not get adequate healthcare facilities. Hence the reason for an immediate need for aid to the Central African Republic.

Top 5 Countries dire need for aid


Haiti
Mali
Chad
Sierra Leone
Central African Republic