HELP International Countries Categorization (Clustering)

PCA and Clustering Assignment Analysis Presentation

Prepared by: Chandrashekhar Purandare

24th Feb 2020

Executive Context

▶ HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities.

▶ HELP International have been able to raise around $ 10 million of fund with recent funding program.

▶ HELP International wants to allocate these funds to countries that are in the direst need of aid based on their socio-economic and health factors.

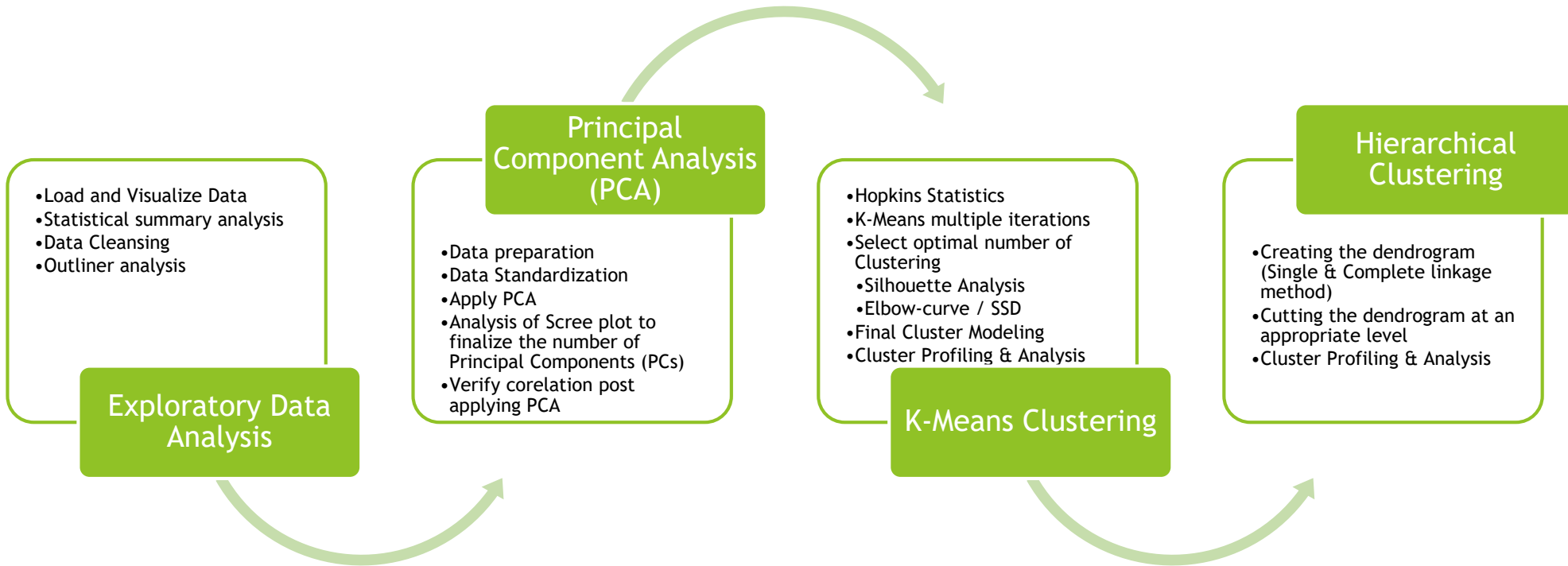▶ This analysis summary intended to drive "*data-driven*" decision making for CEO of the HELP International.

Business Objective

- Categorize the countries using some socio-economic and health factors that determine the overall development of the country.

- Recommend the list of countries which the CEO needs to focus on the most for providing the financial aid

Input Dataset Analysis

- Input dataset contains socio-economic data such as import & export rates, net income /per person, inflation rate which contribute to GDPP of the country.

- Input dataset contains various health factors such as child mortality, life expectancy, total fertility and total health spending for each country.

- Number of missing/ null values in dataset – None (0)

- Number of duplicate of duplicates in dataset – None (0)

- Columns with large number of outliers (which are managed through clustering)

  - Exports, health, imports

- Columns with small number of outliers (which are addressed)

  - GDPP, life expectancy total fertility

- Columns with large number of outliers (which are not treated for business reasons)

  - Income, inflation

Analysis Methodology

Exploratory Data Analysis
- Load and Visualize Data
- Statistical summary analysis
- Data Cleansing
- Outliner analysis

Principal Component Analysis (PCA)
- Data preparation
- Data Standardization
- Apply PCA
- Analysis of Scree plot to finalize the number of Principal Components (PCs)
- Verify corelation post applying PCA

K-Means Clustering
- Hopkins Statistics
- K-Means multiple iterations
- Select optimal number of Clustering
  - Silhouette Analysis
  - Elbow-curve / SSD
- Final Cluster Modeling
- Cluster Profiling & Analysis

Hierarchical Clustering
- Creating the dendrogram (Single & Complete linkage method)
- Cutting the dendrogram at an appropriate level
- Cluster Profiling & Analysis

EDA Summary

	count	mean	std	min	25%	50%	75%	max
child_mort	167.00000	38.27006	40.32893	2.60000	8.25000	19.30000	62.10000	208.00000
exports	167.00000	41.10898	27.41201	0.10900	23.80000	35.00000	51.35000	200.00000
health	167.00000	6.81569	2.74684	1.81000	4.92000	6.32000	8.60000	17.90000
imports	167.00000	46.89021	24.20959	0.06590	30.20000	43.30000	58.75000	174.00000
income	167.00000	17144.68862	19278.06770	609.00000	3355.00000	9960.00000	22800.00000	125000.00000
inflation	167.00000	7.78183	10.57070	-4.21000	1.81000	5.39000	10.75000	104.00000
life_expec	167.00000	70.55569	8.89317	32.10000	65.30000	73.10000	76.80000	82.80000
total_fer	167.00000	2.94796	1.51385	1.15000	1.79500	2.41000	3.88000	7.49000
gdpp	167.00000	12964.15569	18328.70481	231.00000	1330.00000	4660.00000	14050.00000	105000.00000

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 167 entries, 0 to 166
Data columns (total 10 columns):
country        167 non-null object
child_mort     167 non-null float64
exports        167 non-null float64
health         167 non-null float64
imports        167 non-null float64
income         167 non-null int64
inflation      167 non-null float64
life_expec     167 non-null float64
total_fer      167 non-null float64
gdpp           167 non-null int64
dtypes: float64(7), int64(2), object(1)
memory usage: 13.2+ KB
```
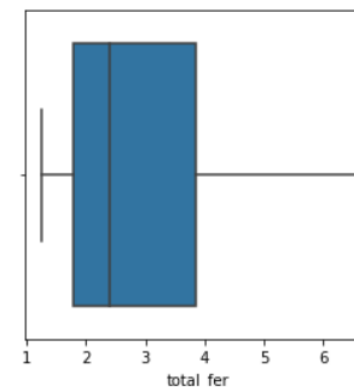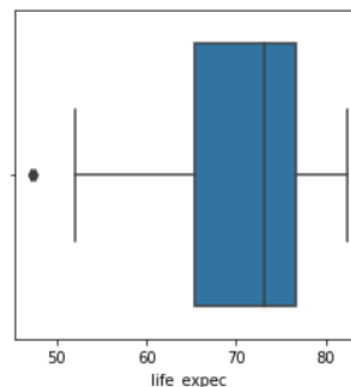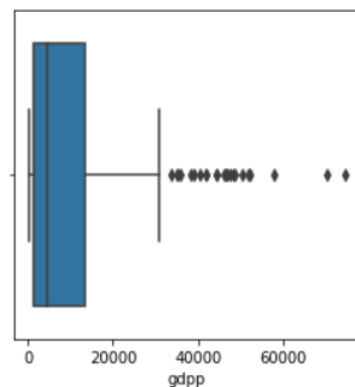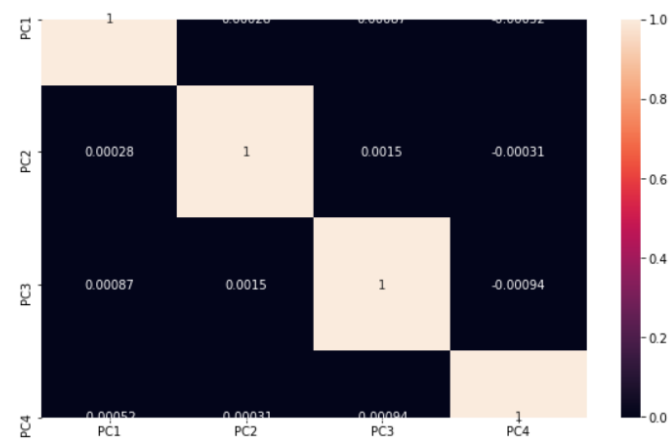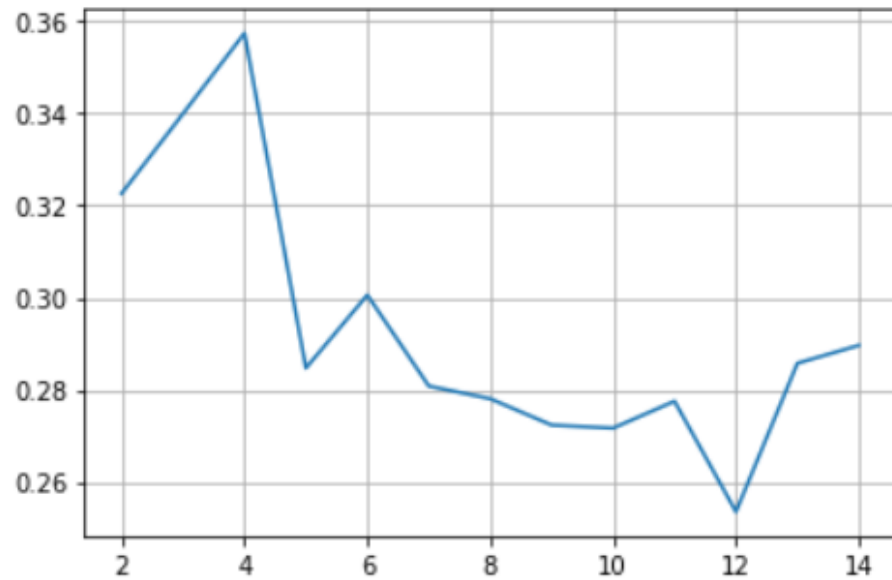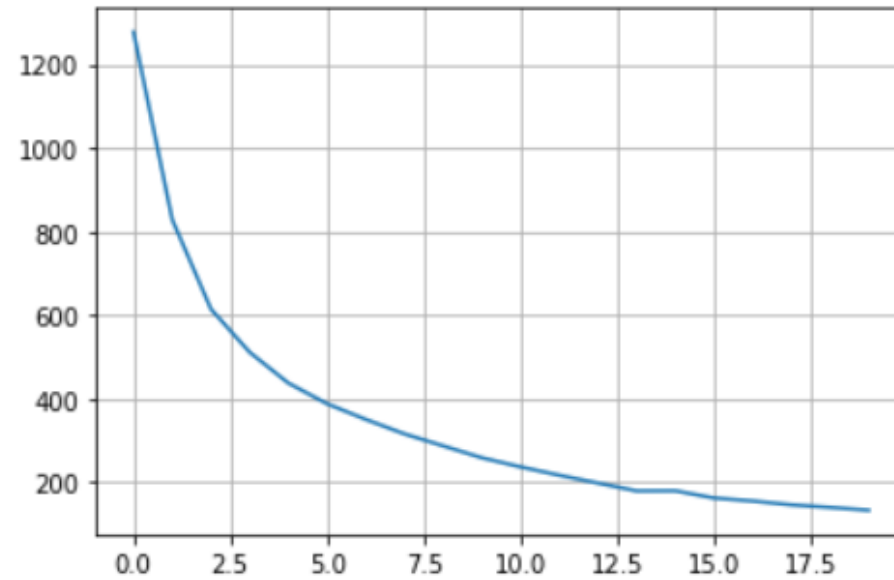
PCA Summary
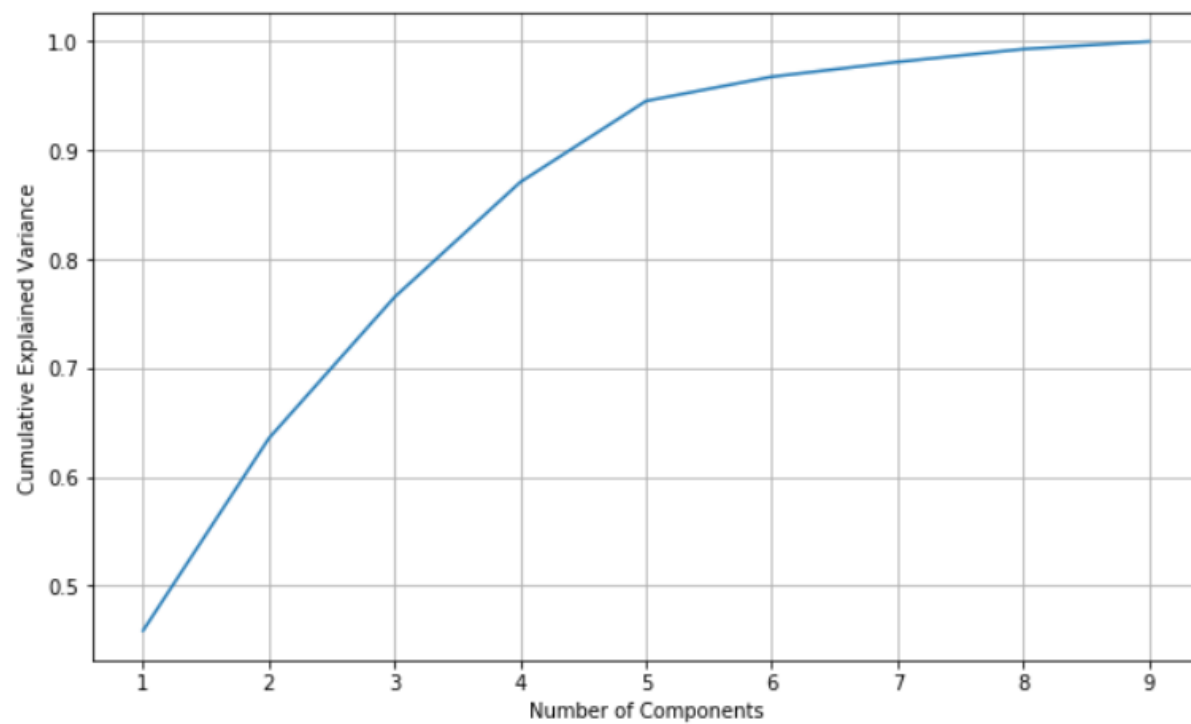
Scree Plot

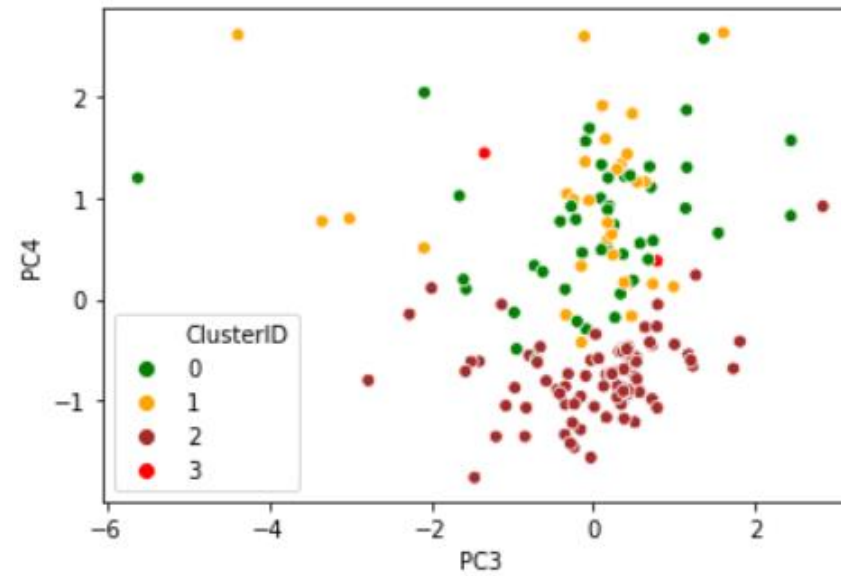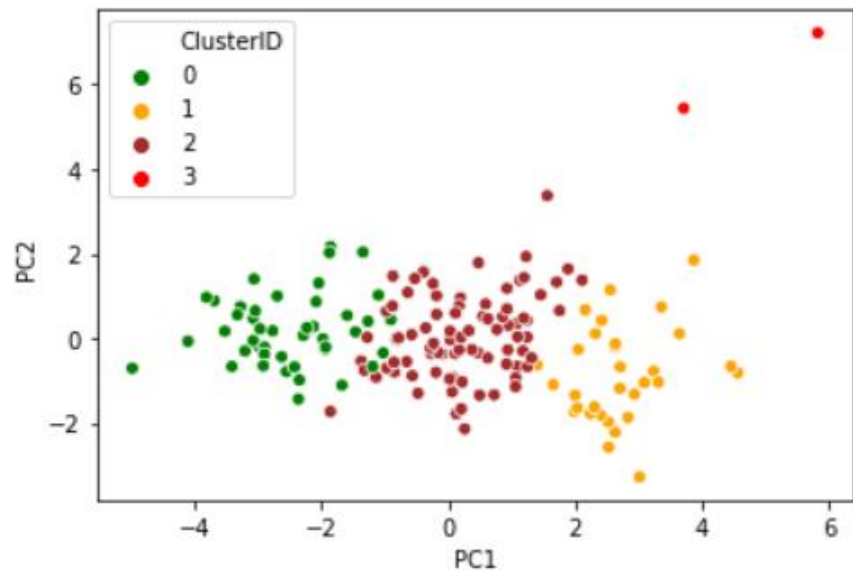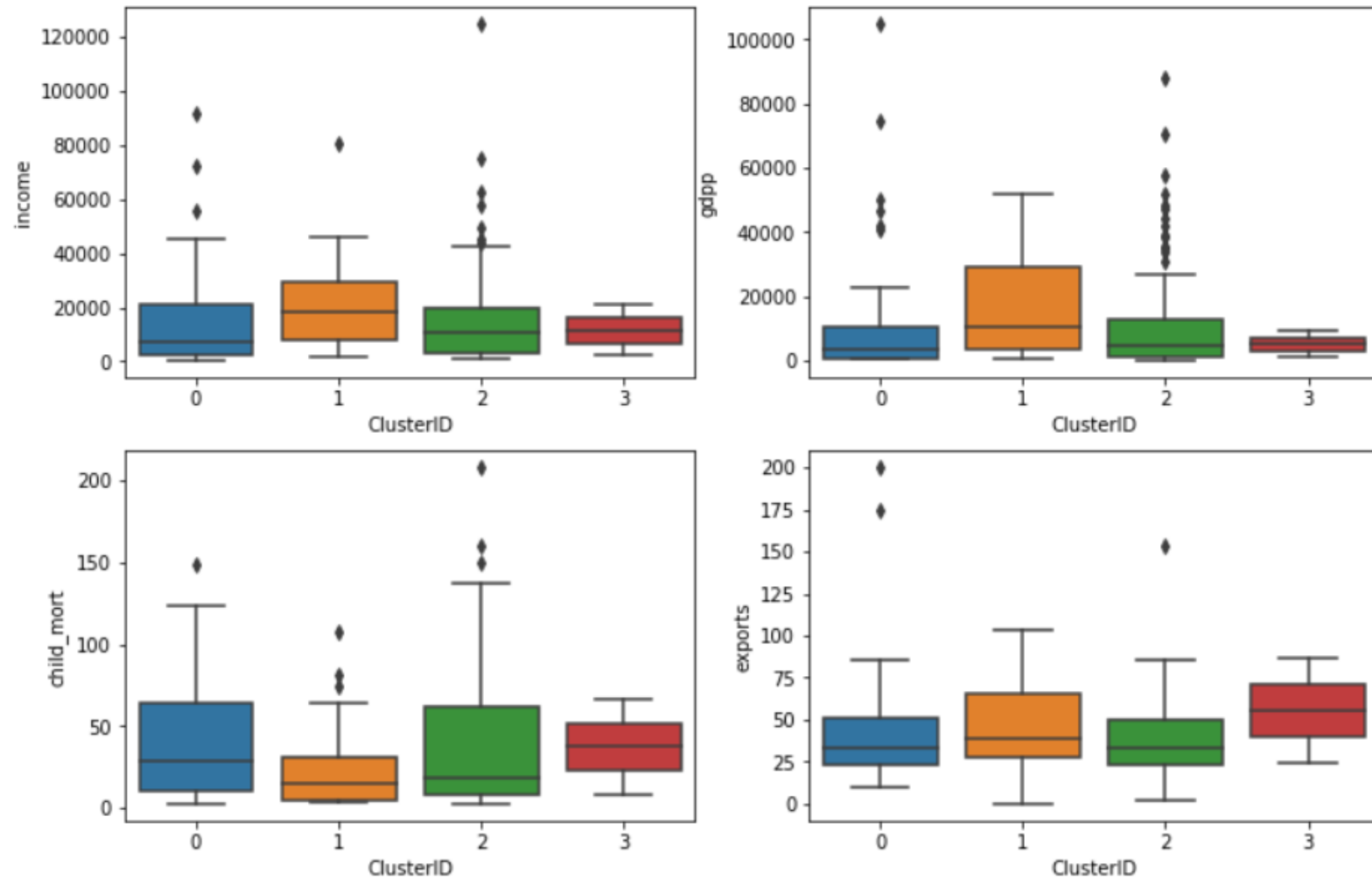Corelation Matrix

K-Means Clustering Summary
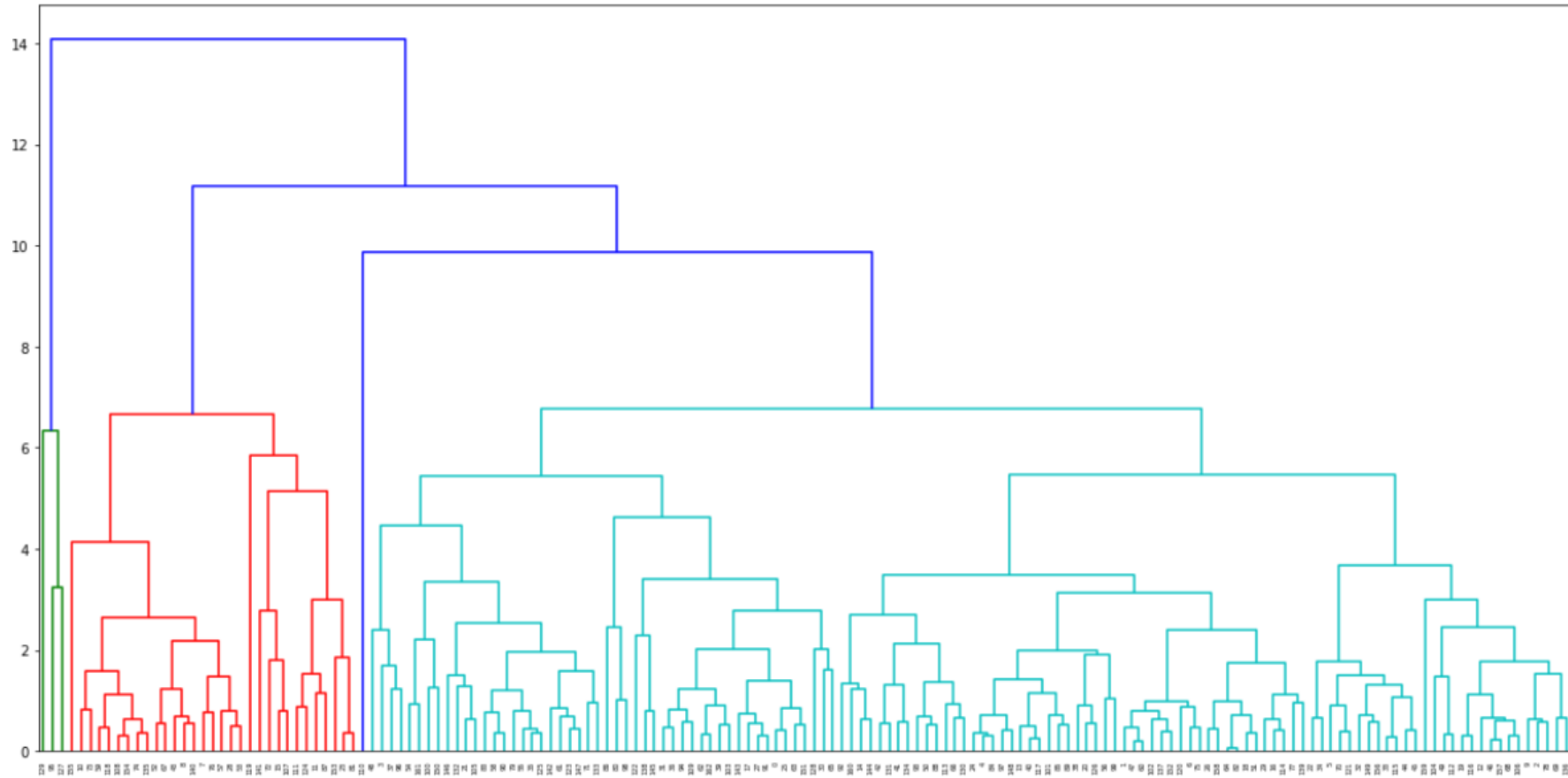

Silhouette Plot

Elbow-curve / SSD

K-Means Cluster Visualization

K-Means Cluster Analysis

Hierarchical Cluster Visualization

Final list of countries

We used two types of clustering models - K-means and Hierarchical Clustering to first identify the cluster which can segregate the countries based on socio-economic and health factors.

We have then considered the top 10 countries which are in direst need of aid based on the socio-economic and health factors that determine the overall development of the country. Following factors taken into considerations to decide the top 10 countries which require aid on priority basis.

- **Child mortality rate is high.**
- **Life expectancy is low**
- **Countries with lower net income per person**
- **Countries with lower GDPP rate.**

As per Hierarchical Clustering, the countries which require aid are:

- Burundi
- Liberia
- Congo, Dem. Rep
- Niger
- Sierra Leone

- Madagascar
- Mozambique
- Central African Republic
- Malawi
- Eritrea

Thank you!