# Country Clustering Assignment

Using K-means and Hierarchical Clustering approach

Krunal Tanna

# Summary

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. It runs a lot of operational projects from time to time along with advocacy drives to raise awareness as well as for funding purposes.

After the recent funding programs, they have been able to raise around $ 10 million. Now the CEO of the NGO needs to decide how to use this money strategically and effectively. The significant issues that come while making this decision are mostly related to choosing the countries that are in the direst need of aid.

I must categorize the countries using some socio-economic and health factors that determine the overall development of the country. Then I need to suggest the countries which the CEO needs to focus on the most.

# Objective

My main task is to cluster the countries by using K-means and Hierarchical Clustering approach by the factors mentioned and then present the solution by providing the list of countries who need the financial aid.

# Approach

My approach to this problem is:

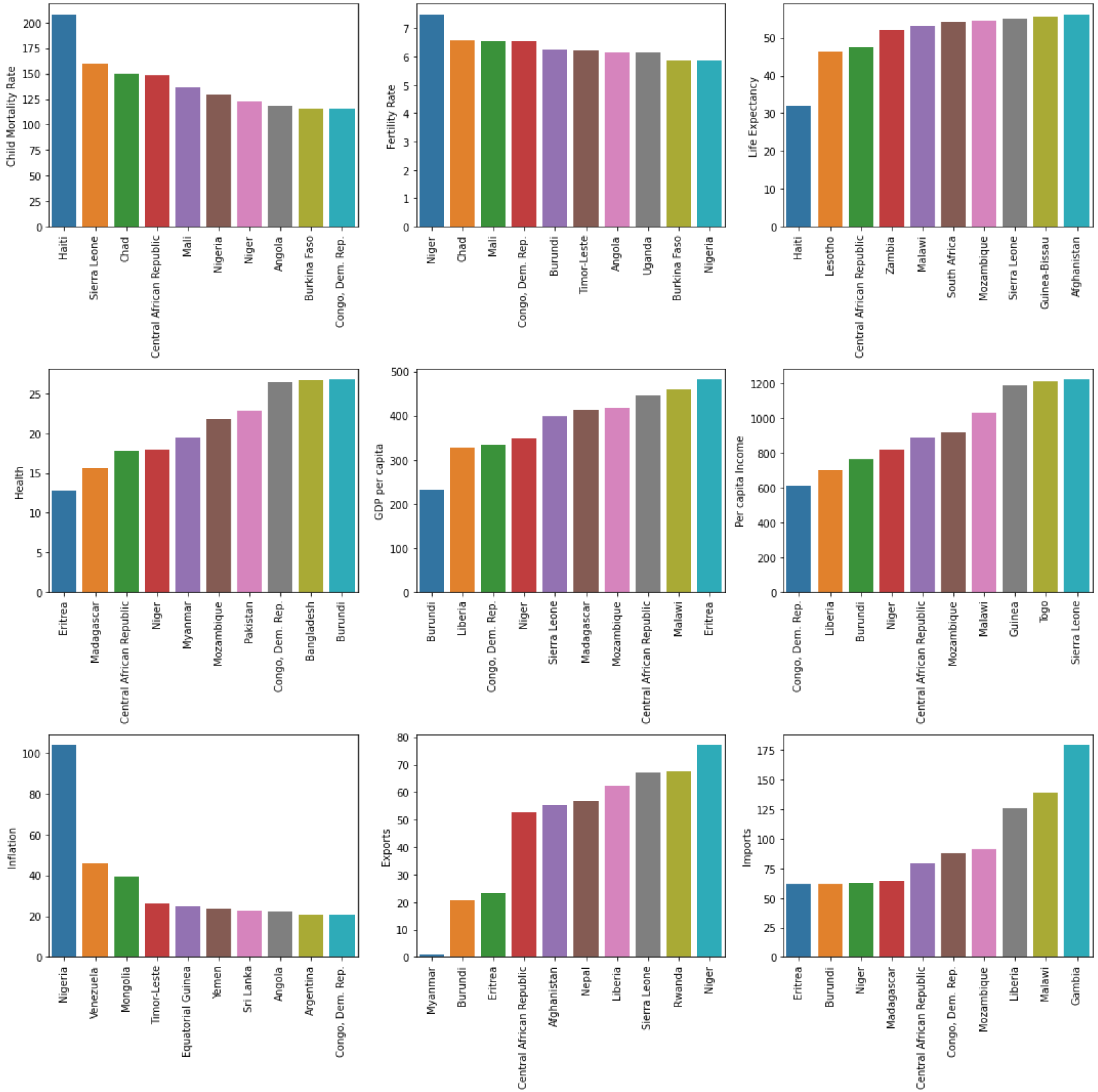
1. Data Reading & Inspection
2. Exploratory Data Analysis
3. Outlier Analysis
4. Hopkins Statistics Test
5. Rescaling the Features
6. Finding the optimal number of clusters – Silhouette and Elbow Curve
7. Model Building – K-means & Hierarchical
8. Conclusion

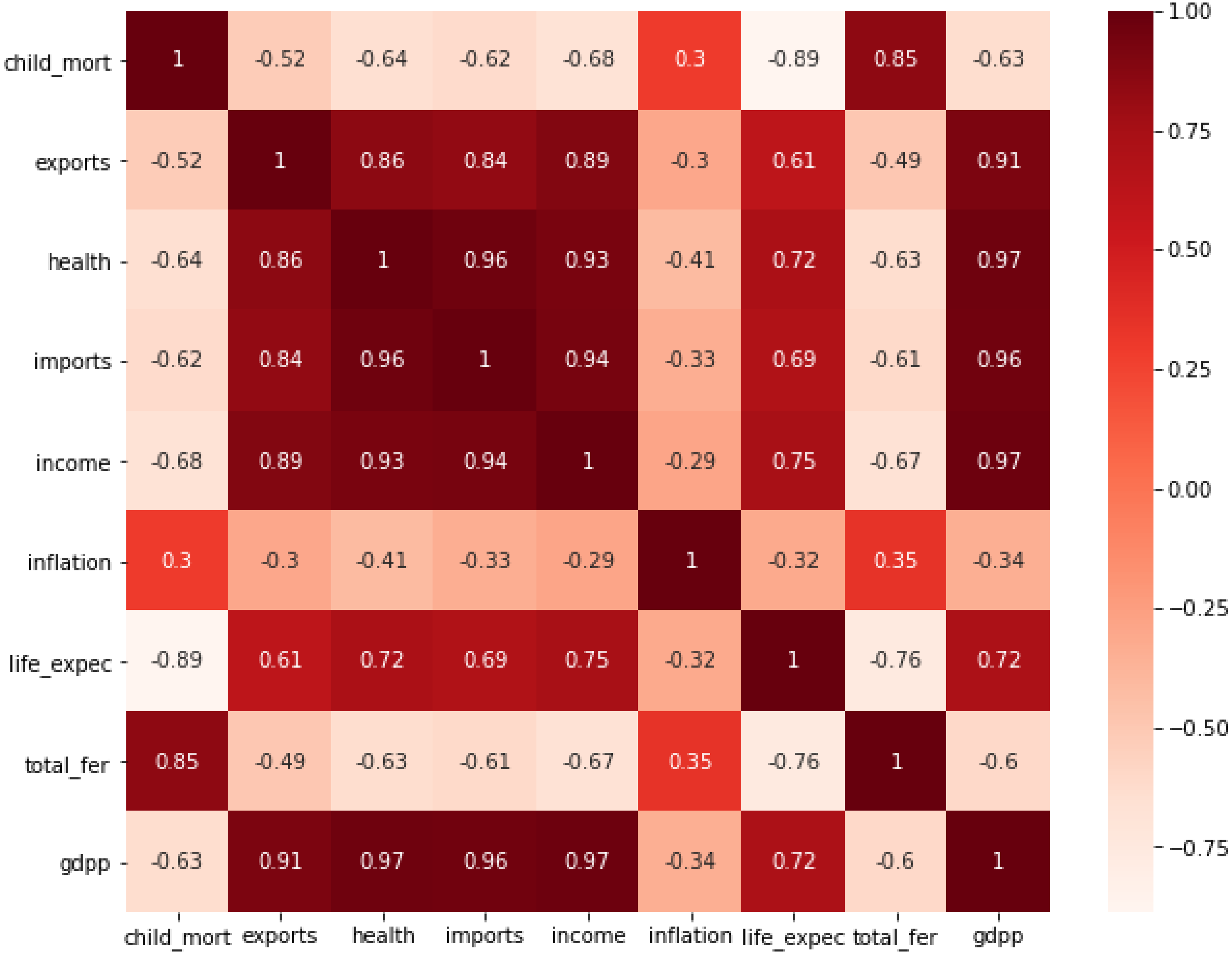
# Data Reading and Inspection

- We have read the data frame by importing pandas library and using .read_csv() function
- We performed data check by checking the shape, info of the data
- We also checked the description of the data i,.e mean, min, max, median.
- We performed null check if in case there are any null/missing values in the data set but luckily there were none.
- From data dictionary, we can find that exports, health and imports values are mentioned in percentage w.r.t gdpp. So we convert ed them in actual values.
- Now our data is ready to be perform EDA.
- Shape of the data is: (167,10)

# Exploratory Data Analysis

- Here we have plotted the bar chart of countries with various parameters.
- From the 1st and 3rd Fig. we can clearly observe that Haiti has the highest Child Mortality Rate and even the Life expectancy is very low.
- From Fig.4 it is observed that Health facilities in Eritrea, Burundi, Niger and Madagascar is extremely poor.
- Fig 5. and Fig. 6 showcases that countries like Burundi, Liberia and Congo Dem Rep has the lowest GDP and income

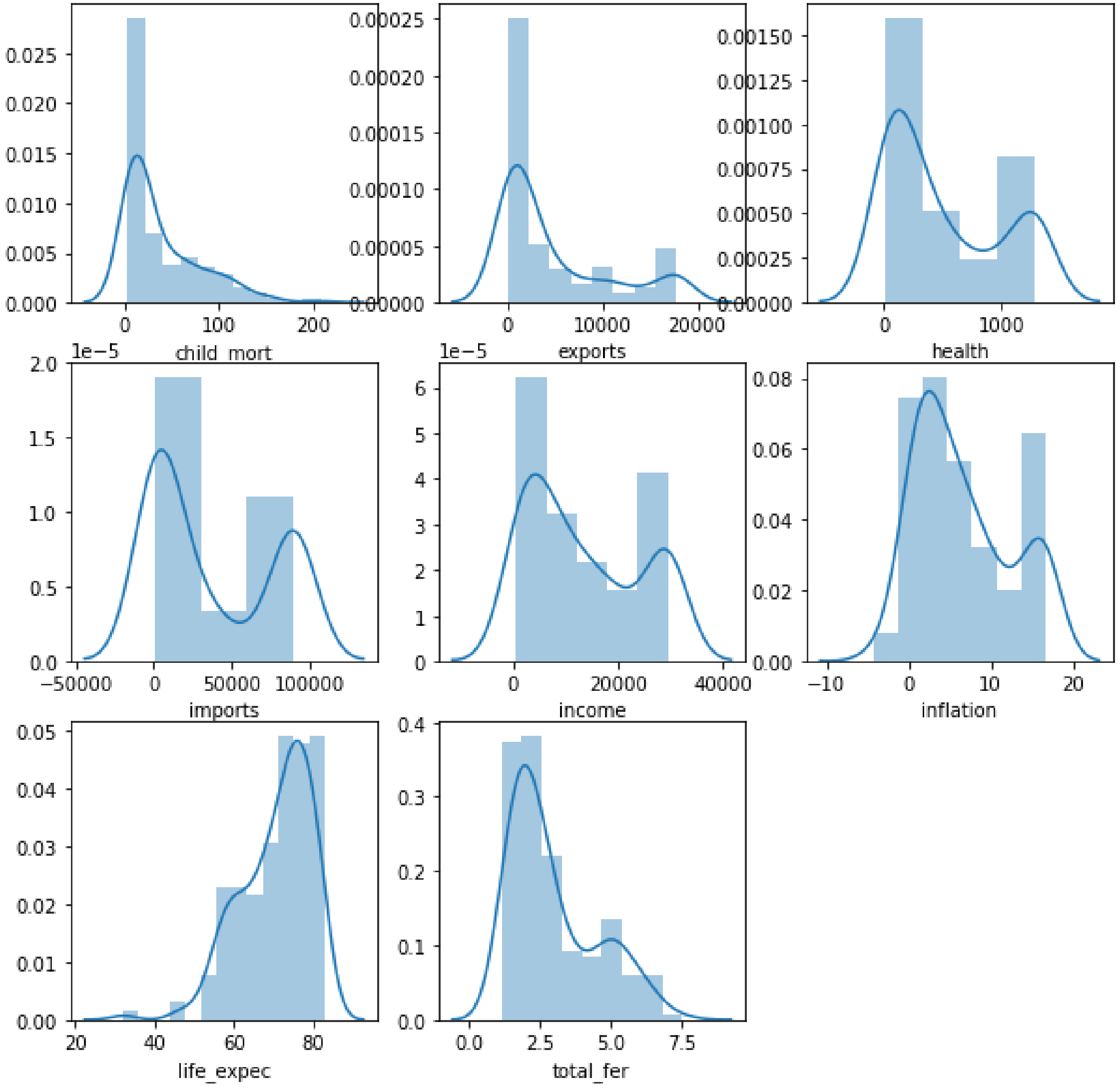




|  | child_mort | exports | health | imports | income | inflation | life_expec | total_fer | gdpp |
|---|---|---|---|---|---|---|---|---|---|
| child_mort | 1 | -0.52 | -0.64 | -0.62 | -0.68 | 0.3 | -0.89 | 0.85 | -0.63 |
| exports | -0.52 | 1 | 0.86 | 0.84 | 0.89 | -0.3 | 0.61 | -0.49 | 0.91 |
| health | -0.64 | 0.86 | 1 | 0.96 | 0.93 | -0.41 | 0.72 | -0.63 | 0.97 |
| imports | -0.62 | 0.84 | 0.96 | 1 | 0.94 | -0.33 | 0.69 | -0.61 | 0.96 |
| income | -0.68 | 0.89 | 0.93 | 0.94 | 1 | -0.29 | 0.75 | -0.67 | 0.97 |
| inflation | 0.3 | -0.3 | -0.41 | -0.33 | -0.29 | 1 | -0.32 | 0.35 | -0.34 |
| life_expec | -0.89 | 0.61 | 0.72 | 0.69 | 0.75 | -0.32 | 1 | -0.76 | 0.72 |
| total_fer | 0.85 | -0.49 | -0.63 | -0.61 | -0.67 | 0.35 | -0.76 | 1 | -0.6 |
| gdpp | -0.63 | 0.91 | 0.97 | 0.96 | 0.97 | -0.34 | 0.72 | -0.6 | 1 |

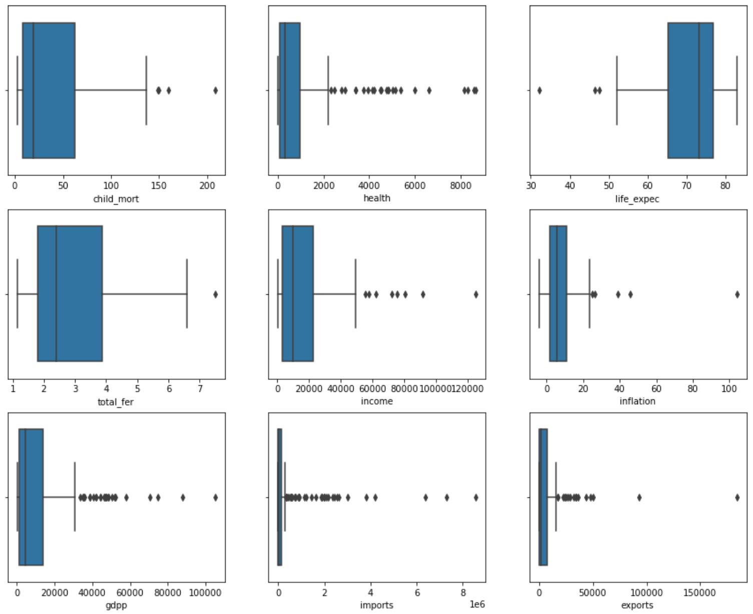
# Exploratory Data Analysis

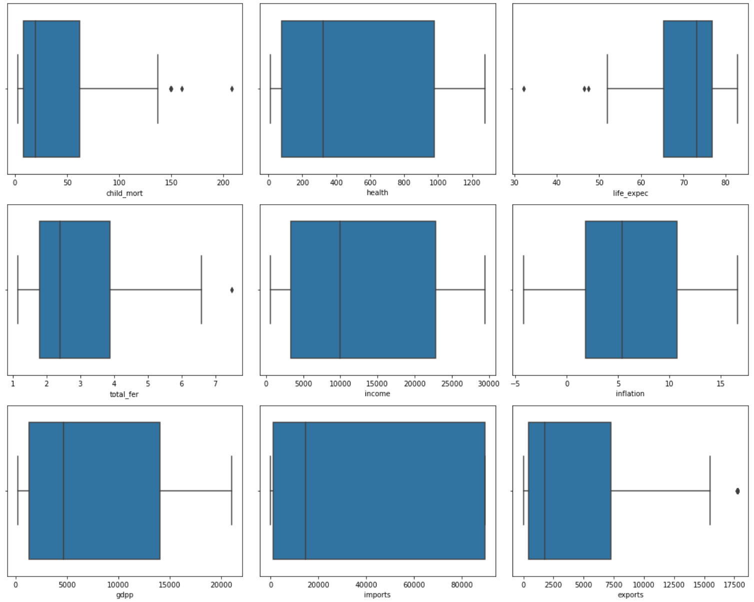
- Here we have plotted Heatmap to find correlation between different variables
- child_mortality and life_expentency are highly correlated with correlation of -0.89
- child_mortality and total_fertility are highly correlated with correlation of 0.85
- imports and exports are highly correlated with correlation of 0.74
- life_expentency and total_fertility are highly correlated with correlation of -0.76



# Exploratory Data Analysis

- We have plotted a histogram here to get an overall idea of the distribution of the dataset.
- From the distribution, It looks like 2/3 clusters will be optimal.





# Outlier Analysis

- To check if the data has outliers, we have plotted the boxplot.
- As we can see from the first chart that variables like exports/imports, health, income, inflation and gdpp have high number of outliers.
- So, we will be capping the upper range of outliers by handling them using capping techniques where in we will limit the extreme variables up to a particular range.
- We also have few outliers in Child Mortality, but we wont cap them since it is our target variable, and we need the countries which have high number of child mortality.
- After capping, we have plotted the boxplot again which can be seen in second figure where we do not have outliers left.

# Hopkins Statistics Test

- The Hopkins statistic (introduced by Brian Hopkins and John Gordon Skellam) is a way of measuring the cluster tendency of a data set. It acts as a statistical hypothesis test where the null hypothesis is that the data is generated by a Poisson point process and are thus uniformly randomly distributed. A value close to 1 tends to indicate the data is highly clustered, random data will tend to result in values around 0.5, and uniformly distributed data will tend to result in values close to 0.

- 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.

- After applying Hopkins test, score was above 0.85 which means it has a high tendency of cluster formation.

# Rescaling the feature

- There are two common ways of rescaling:
- Min-Max scaling
- Standardization (mean-0, sigma-1)

Here, have used Standardization Scaling.

# Finding the optimal number of clusters

- There are two common ways of finding the optimal number of clusters:

- Silhouette Analysis:

  Silhouette score=(p−q)/max(p,q)

  p is the mean distance to the points in the nearest cluster that the data point is not a part of

  q is the mean intra-cluster distance to all the points in its own cluster.

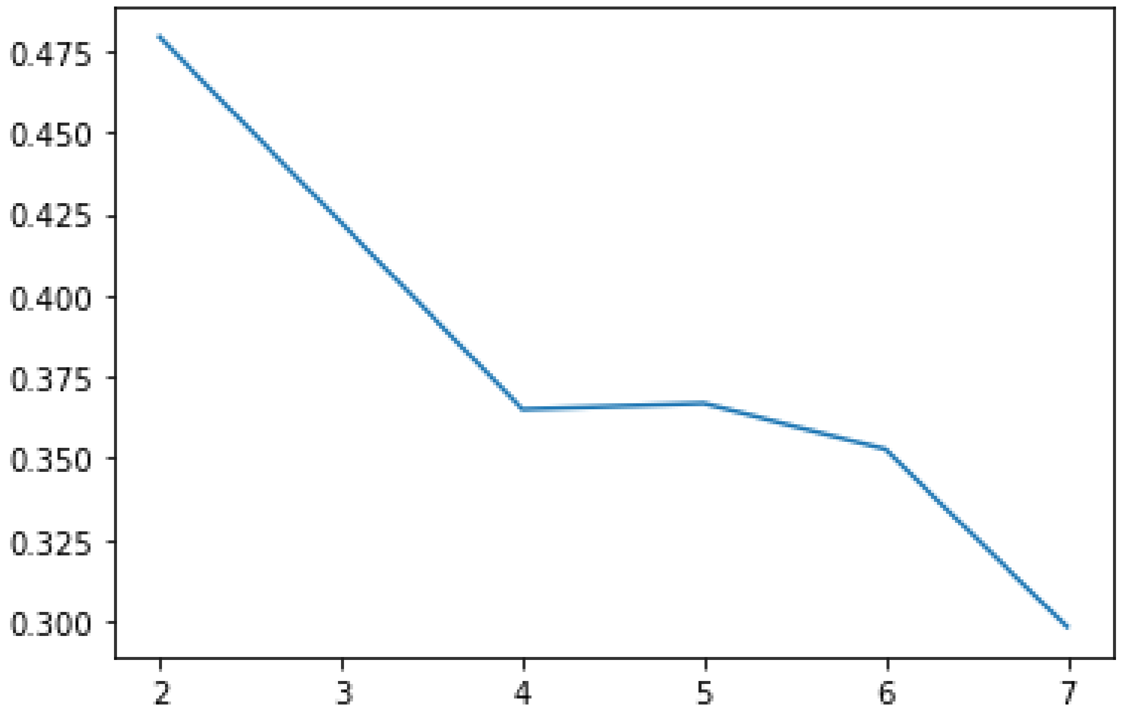
  The value of the silhouette score range lies between -1 to 1.

  A score closer to 1 indicates that the data point is very similar to other data points in the cluster,

  A score closer to -1 indicates that the data point is not like the data points in its cluster.

- Elbow curve:

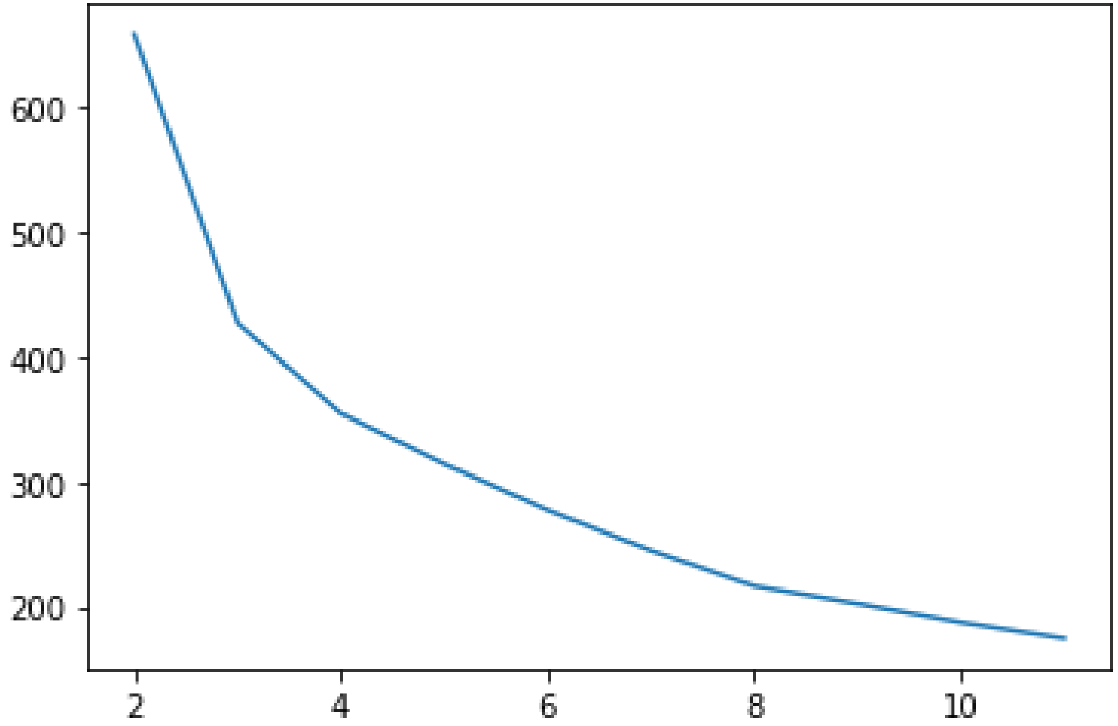
  A fundamental step for any unsupervised algorithm is to determine the optimal number of clusters into which the data may be clustered. The Elbow Method is one of the most popular methods to determine this optimal value of k.



# Silhouette Analysis

- Checked the Silhouette score from range 2-8
- For cluster with n=3, the silhouette score was 0.42 which was found to be an optimal score
- So silhouette analysis suggests, there should be 3 clusters.

# Elbow Curve

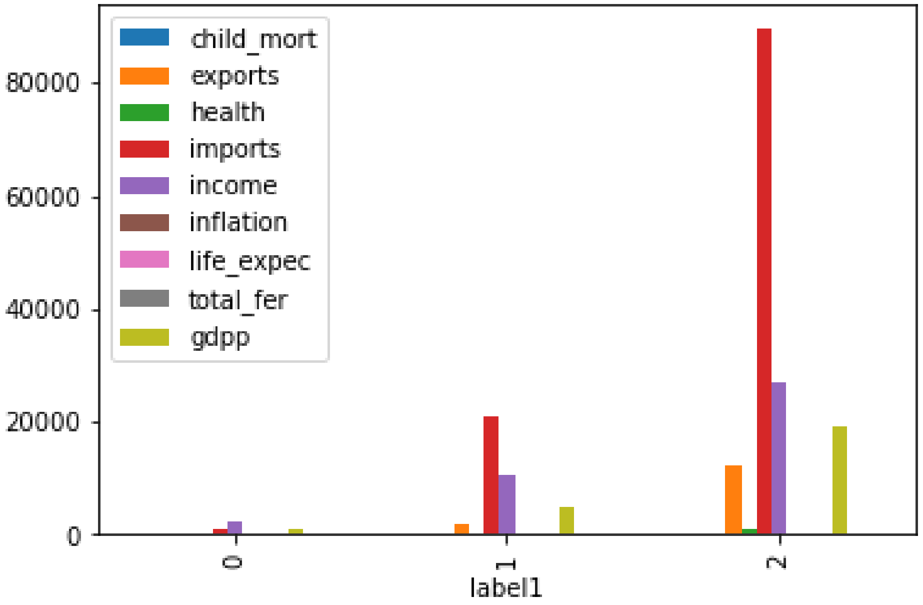


- From the elbow curve it is observed that curve is bend at near n=3
- So elbow curve also suggest the optimal number of clusters to be 3.

- Since both, the analysis suggests 3 clusters, n=3 is used to perform K-means and Hierarchical test.

# Model Building: K-means Clustering

- Applied K-means with cluster =3 and random state = 50 and fitted the object into the scaled dataset.
- After applying value_counts() on the resultant dataset, it is observed: 1 = 75, 2 = 50, 0 = 42. Thus we have sufficient countries in each label.
- Visualized the dataset by plotting bar chart w.r.t labels:



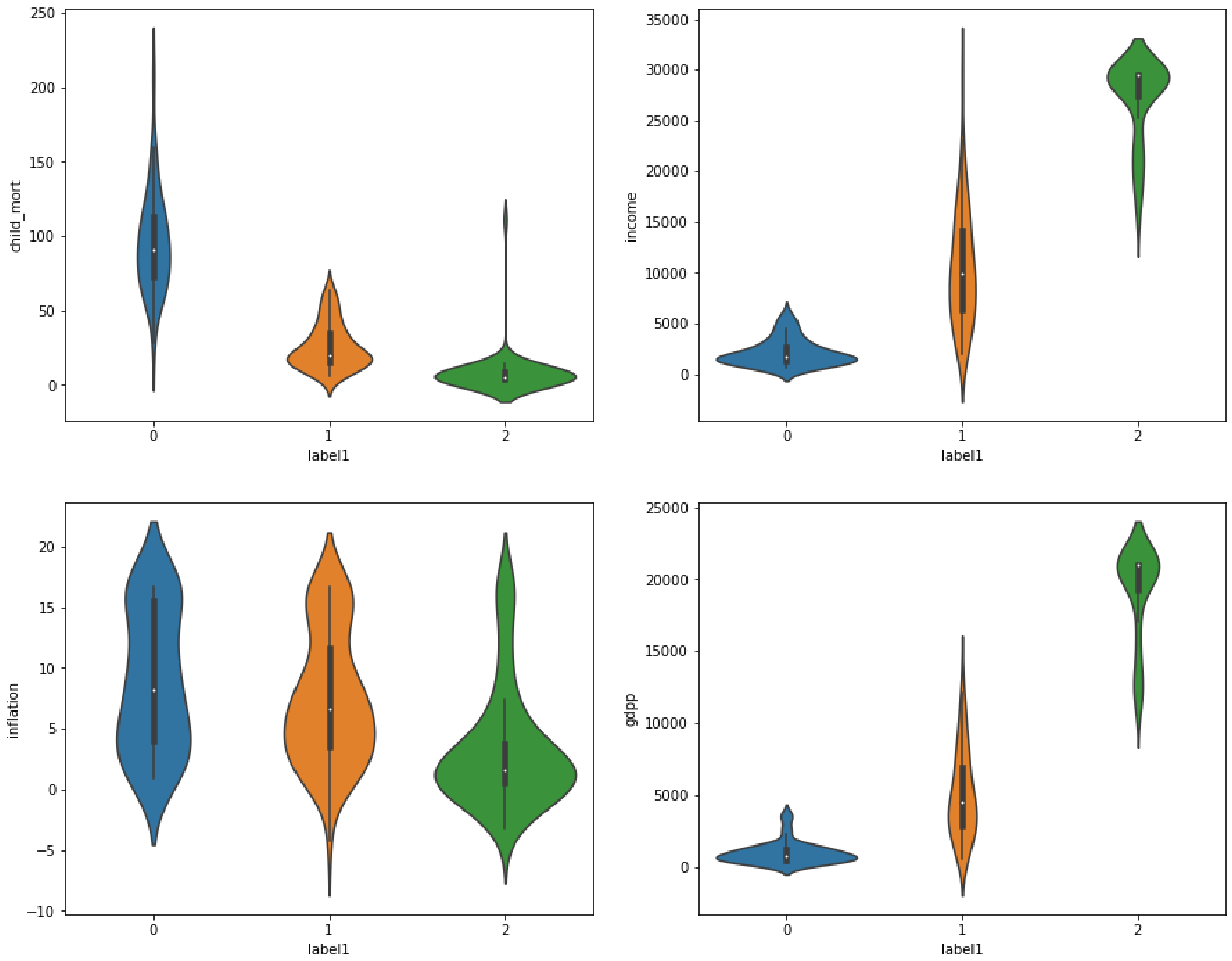
- From the visualization, it is observed that **Label=0** have countries with poor financial and health conditions

# Model Building: K-means Clustering

- As per our requirement, we need to sort all the countries with poor financial, health and socio-economic condition on basis of gdpp, child_mort and income.

- Thus after sorting, Violin plot is plotted to check the distribution of the data for child mortality, income and gdpp.

- It is observed that child mortality is high for label 0 and income and gdpp is very low.



# Model Building: Hierarchical Clustering

- Hierarchical clustering involves creating clusters that have a predetermined ordering from top to bottom. For example, all files and folders on the hard disk are organized in a hierarchy. There are two types of hierarchical clustering: Single Linkage and Complete Linkage

- Cut the tree at height of approx. 7 to get 3 clusters and checked if it get any better cluster formation.

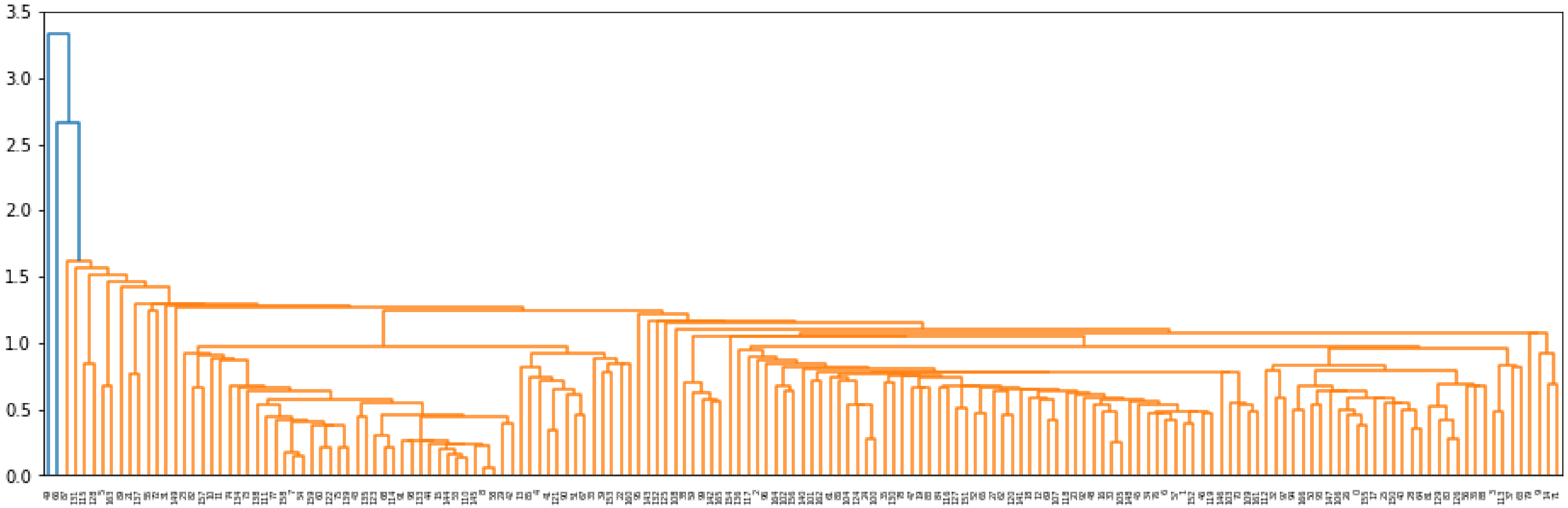
- Assigned Labels to it and merged with the existing dataframe.

- After applying value_counts() on the resultant dataset, it is observed: 1 = 83, 2 = 44, 0 = 40. Thus we have sufficient countries in each label.

- Now I have data frame with both labels.

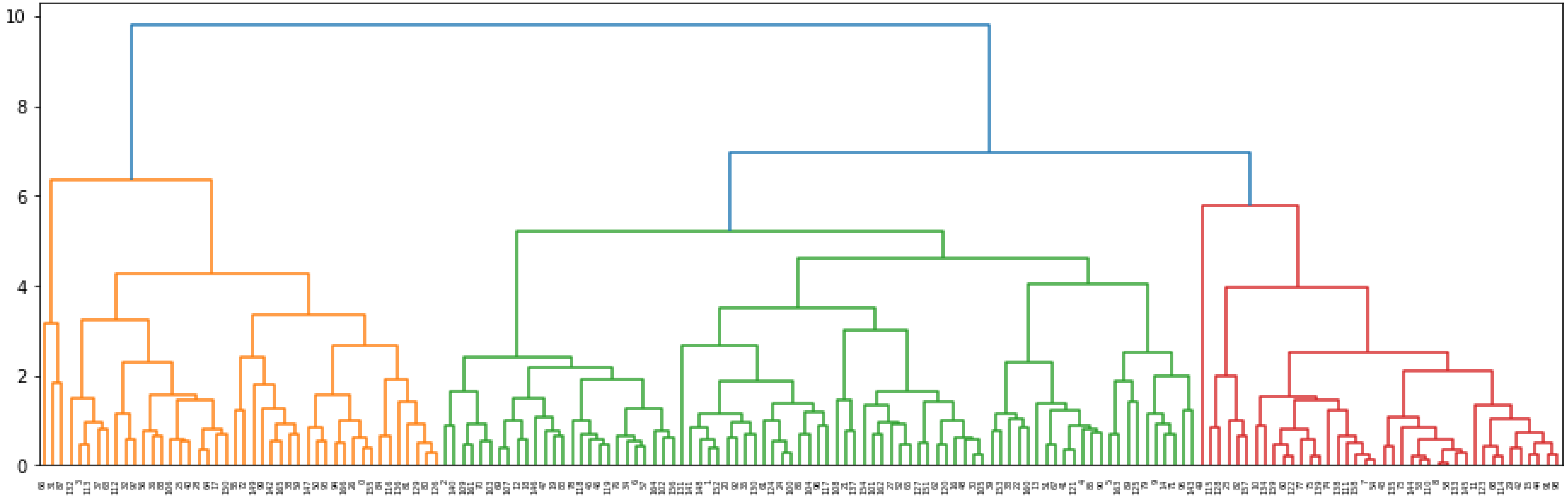
# Hierarchical Clustering: Single Linkage

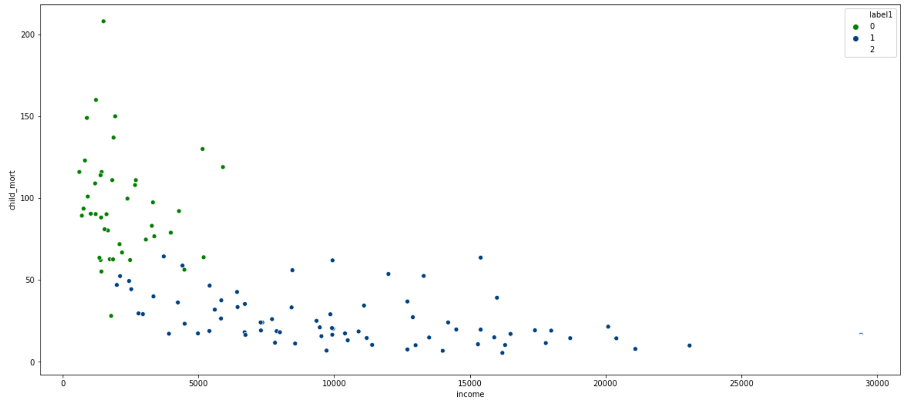
- In single linkage hierarchical clustering, the distance between two clusters is defined as the shortest distance between two points in each cluster. For example, the distance between clusters “r” and “s” to the left is equal to the length of the arrow between their two closest points.
- Performed Single Linkage and plotted its graph.

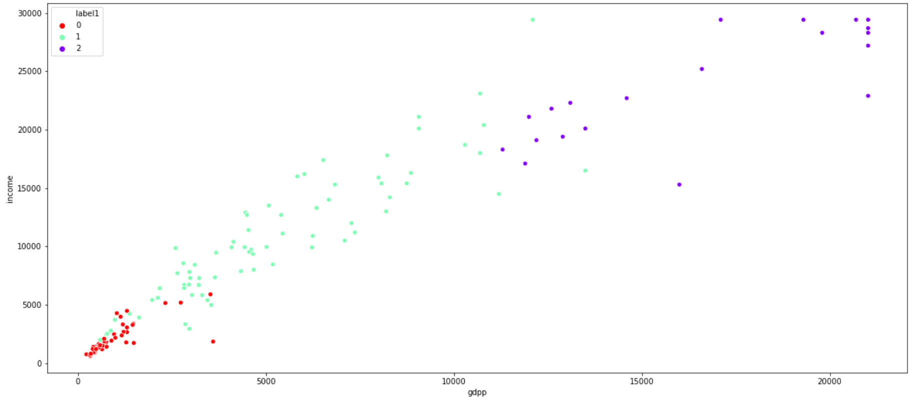


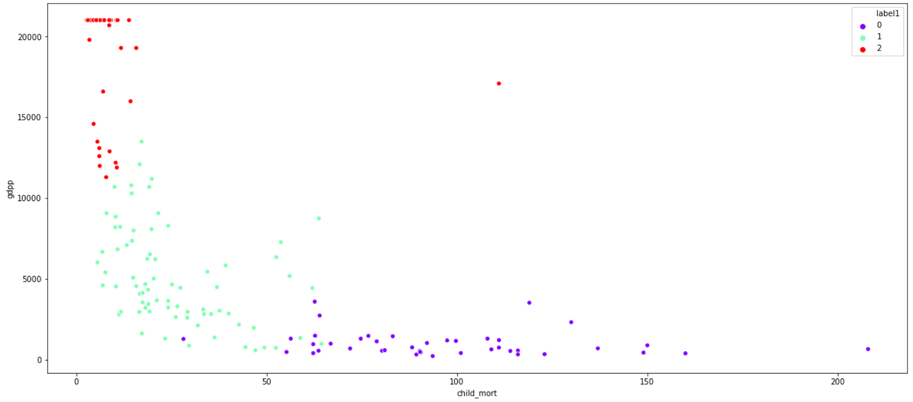
# Hierarchical Clustering: Complete Linkage

- In complete linkage hierarchical clustering, the distance between two clusters is defined as the longest distance between two points in each cluster. For example, the distance between clusters “r” and “s” to the left is equal to the length of the arrow between their two furthest points.
- Performed Complete Linkage and plotted its graph.









# Visualization of Scatter Plot for target variables

- Plotted the Scatter plot for target variables:
- Child mortality
- Income
- Gdpp

- The different clusters are clearly visible for all the labels.

# Conclusion

- After performing K-means and Hierarchical clustering, we have the list of countries with poor socio-economic conditions having the utmost need of Financial Aid.
- Among such countries, the top 5 countries having worse condition on based on target variables are:

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