Clustering & PCA Assignment

SUBMISSION

Submitted by:

Aditya Varma

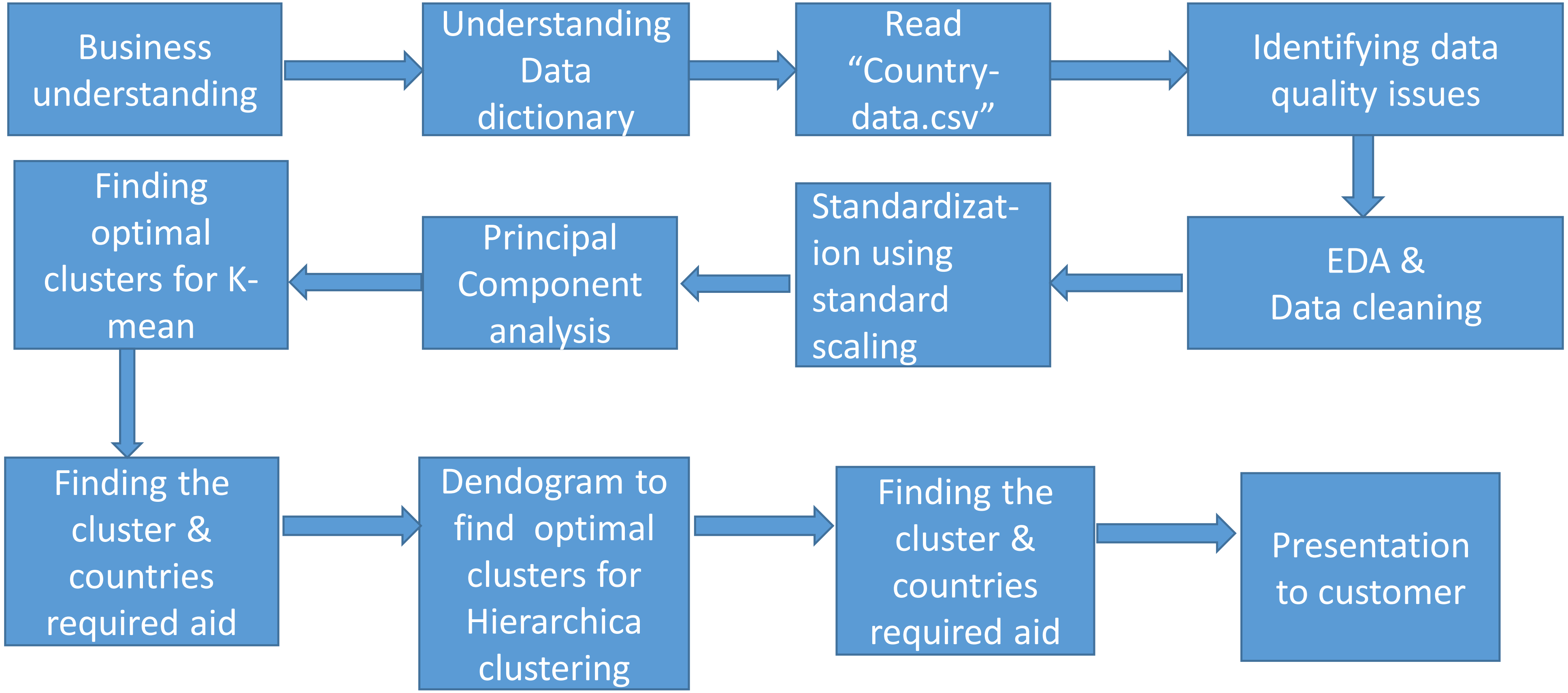
Abstract

Introduction: 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. After the recent funding programs, they have been able to raise around $ 10 million. This analysis is performed to choose the countries that are in the direst need of aid.

Method: The analysis was conducted on data which contains those socio-economic factors such as child mortality, exports, health spending, imports, income, inflation, gdpp, life expectancy, total fertility. Principal component analysis was done on the above data and done dimensionality reduction. K-means clustering and Agglomerate hierarchical clustering was applied on PCA data to find out the countries that are in the direst need of aid.

Results: The findings of this analysis indicates that there are 47 countries which are in need of aid when analysis is done using k-means clustering and 49 countries when analysis is done using Hierarchical clustering.

Problem solving methodology



Business understanding

Understanding Data dictionary

Read “Country-data.csv”

Identifying data quality issues

EDA & Data cleaning

Standardization using standard scaling

Principal Component analysis

Finding optimal clusters for K-mean

Finding the cluster & countries required aid

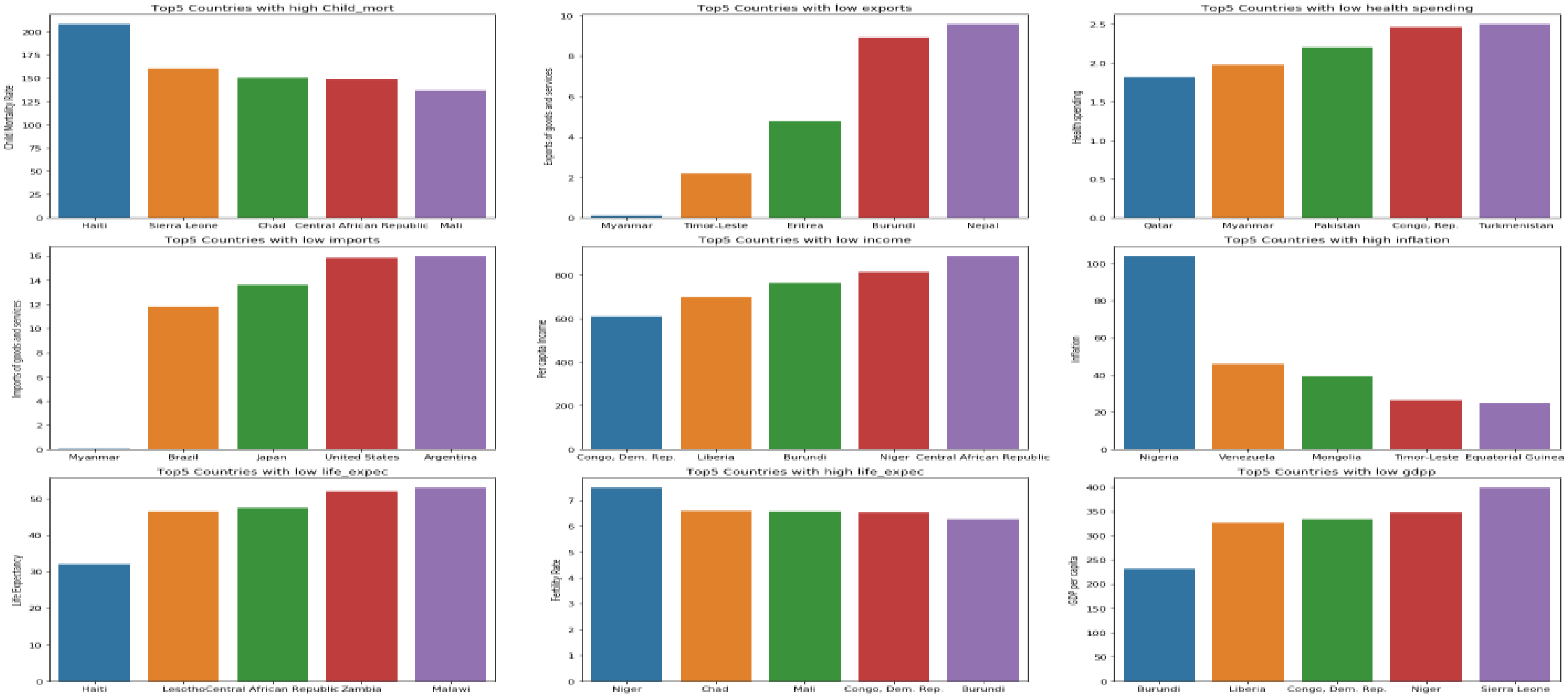
Dendogram to find optimal clusters for Hierarchica clustering

Finding the cluster & countries required aid

Presentation to customer

Exploratory Data Analysis

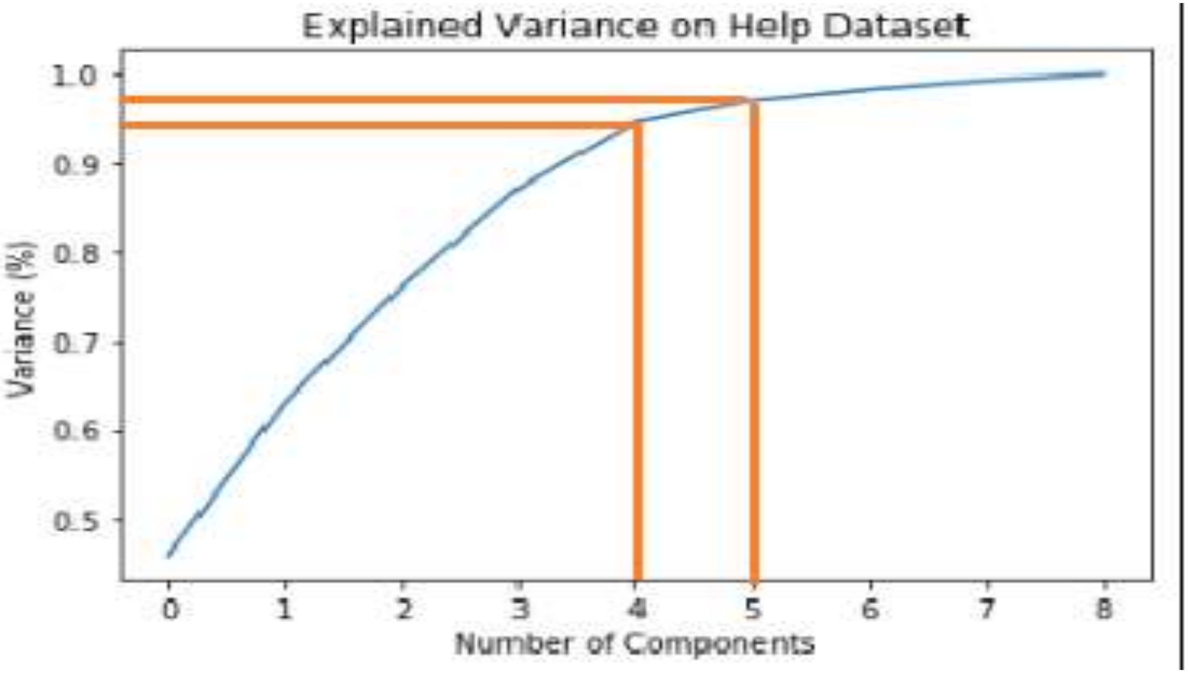
This analysis is done to identify the 5 underdeveloped countries which are in need of the aid at most based on the following factors low life_expectency, low gdpp, low income, low imports & exports, high inflation,high child_mortality, high total_fer



Result: From the above visualization it is clear that the countries which are in need of the financial aid are mostly from the African continent.

Principal Component Analysis

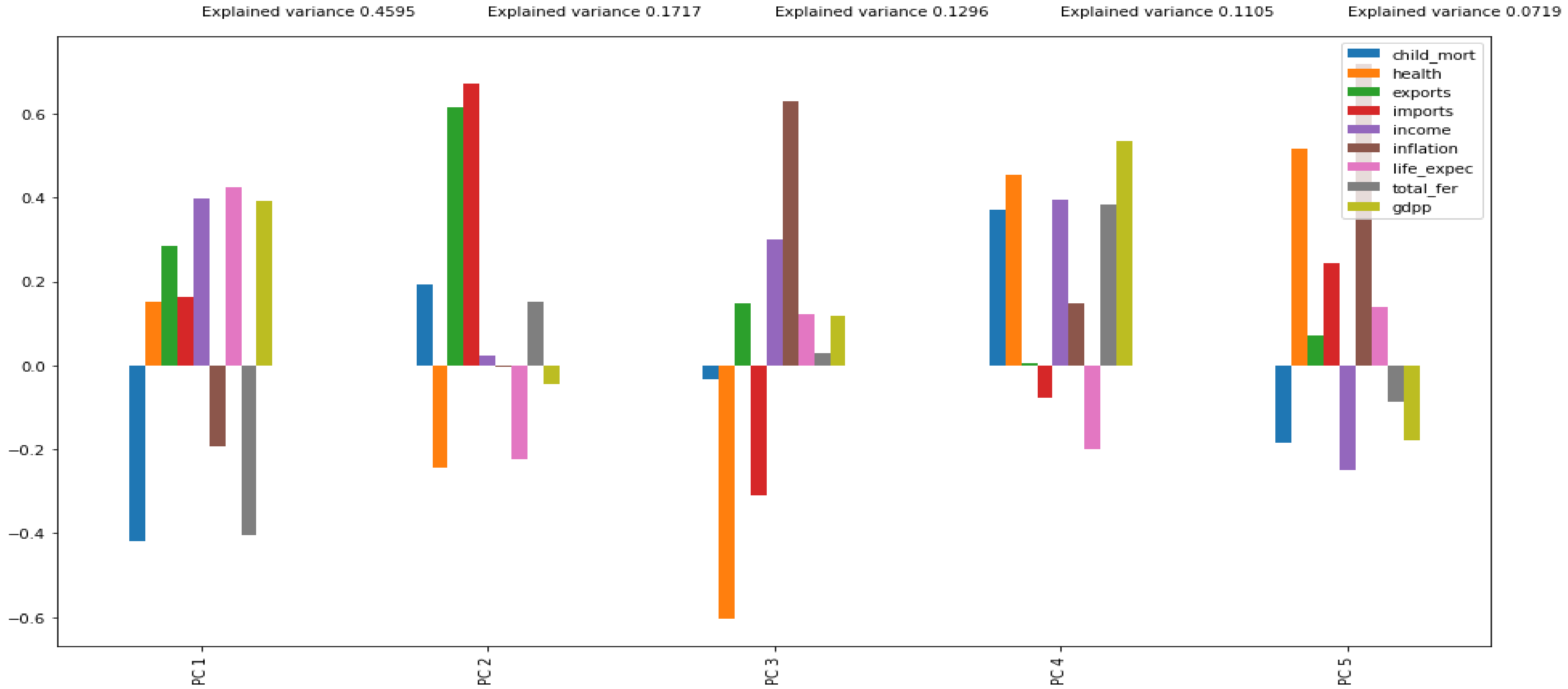
This analysis is done reduce the dimensions from the given data. To perform PCA, first we need to find the number of components which explains the maximum number of variation. This can be observed from scree plot below.



Result: From the above visualization selecting 5 components we can preserve around 95% of the total variance of the data. With this information in our hands, we can implement the PCA for 5 best components we will also verify this by variance explanation plot in the following slide

Principal Component Analysis

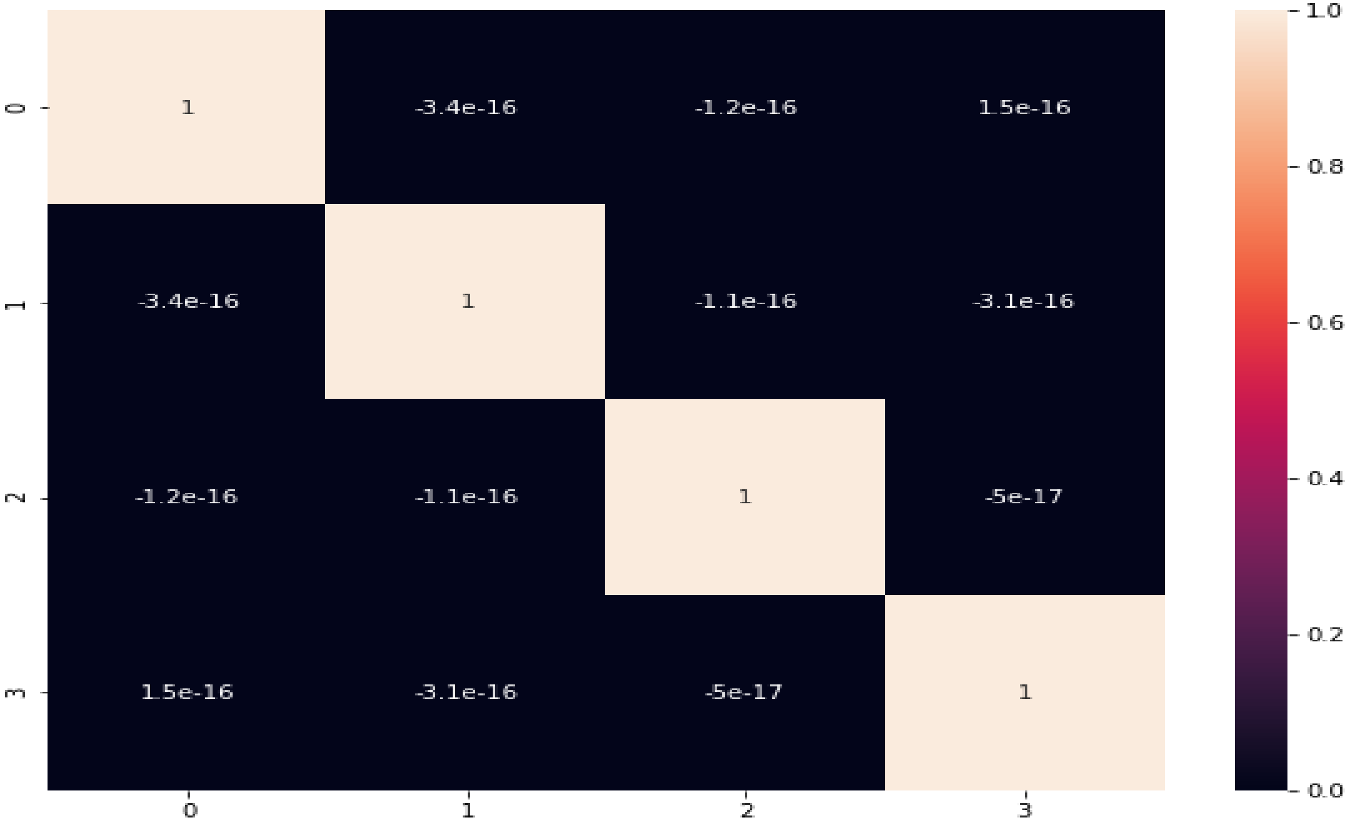
We can also plot explained variance for each principal component to find out percentage of variance explained by each dimension.



Result: First 4 principal components explain the variance of 0.87(0.45+0.17+0.13+0.11) so we can say it's good to use first 4 principal components.

Principal Component Analysis-Covariance matrix

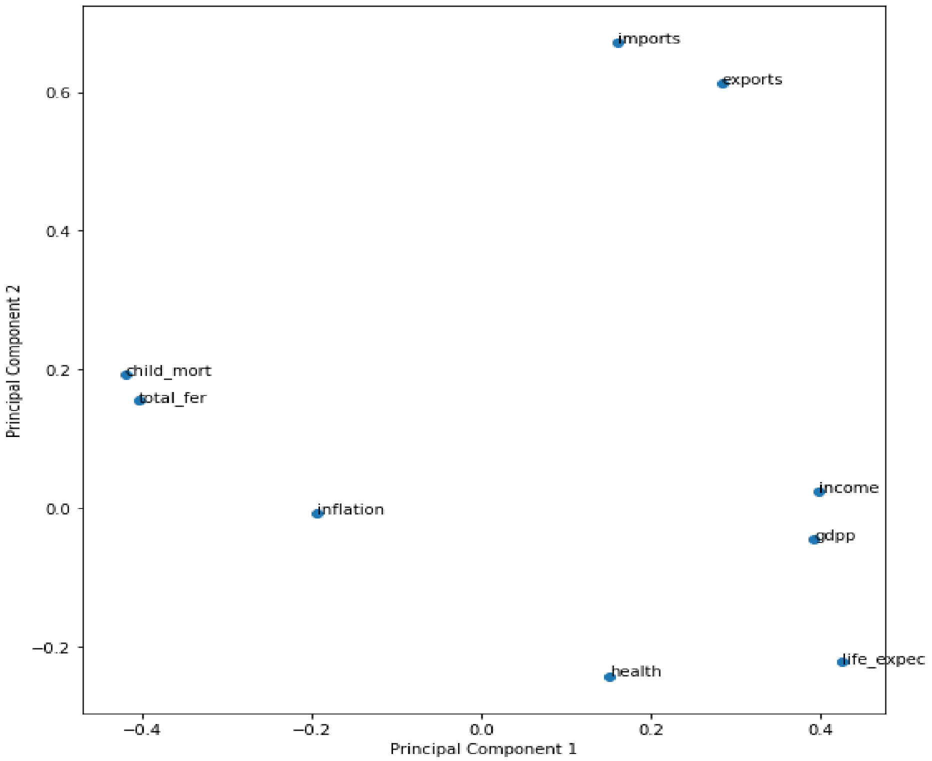
After finding the optimal number of components for PCA and creating a dataframe with principal components we can plot covariance matrix as below to find the correlation among each principal component.



Result: From above correlation plot it can be observed that correlations are very close to 0. Hence there is no correlation between any two components which tells that PCA technique has been applied perfectly to the data

Principal Component Analysis-Scatter plot

Below is the scatter plot plotted for PC1 & PC2 which explains which component is in the which variable direction

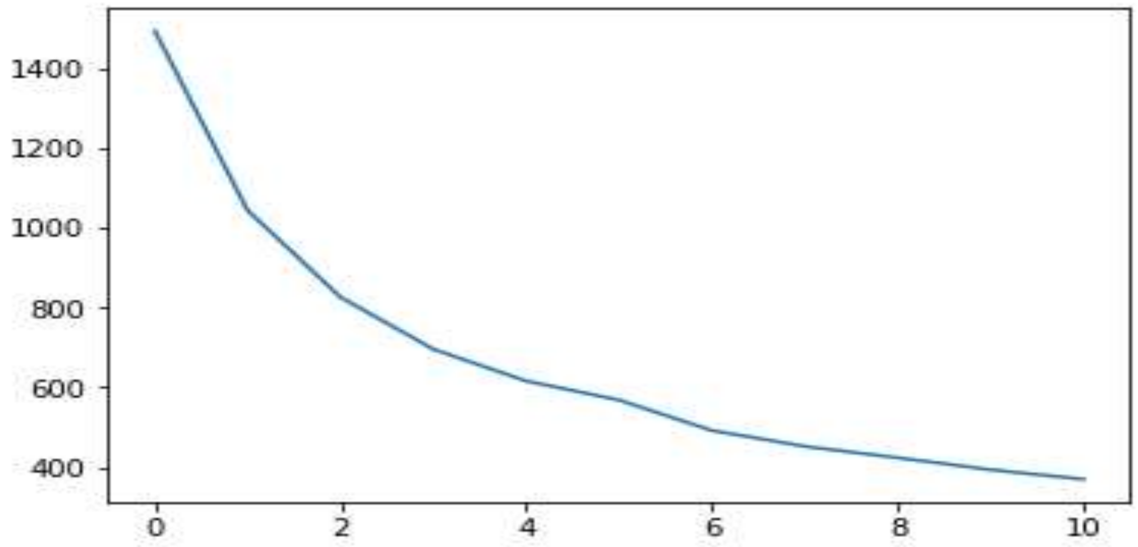


Result: From above plot it can be observed below

- First component is in the direction where the 'health', 'life_exp', 'gdpp', 'income' variables are heavy
- Both components are in the direction where the 'import' & 'export' variables are heavy
- Second component in the direction where the 'child_mort' & 'total_fert' variables are heavy

Clustering analysis (k-means clustering)

To perform k-means clustering we need to find optimal number of clusters before applying k-means. So we can use elbow curve to find number of clusters.



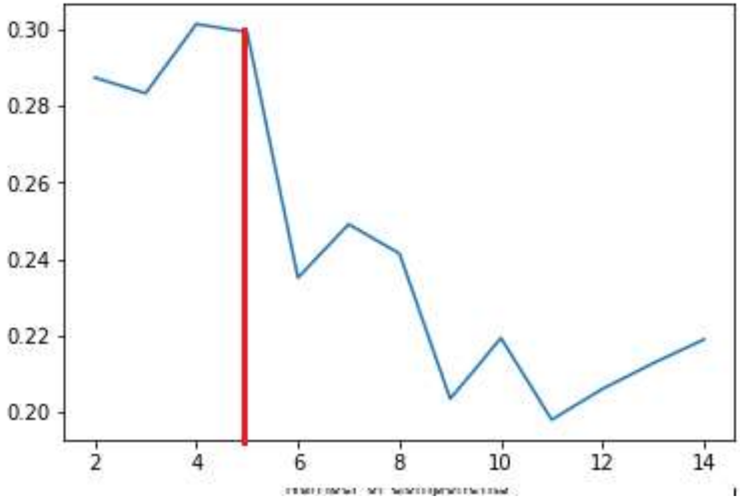
Result: Elbow curve above does not have a clear elbow. Instead, we see a fairly smooth curve, and it's unclear what is the best value of k to choose. In cases like this, we might try a different method for determining the optimal k, such as computing silhouette scores which is explained in following slide.

Clustering analysis (k-means clustering)

Silhouette Score is also used to find the number of clusters. It is defined by the formula

$$\text{silhouette score} = \frac{p-q}{max(p,q)}$$

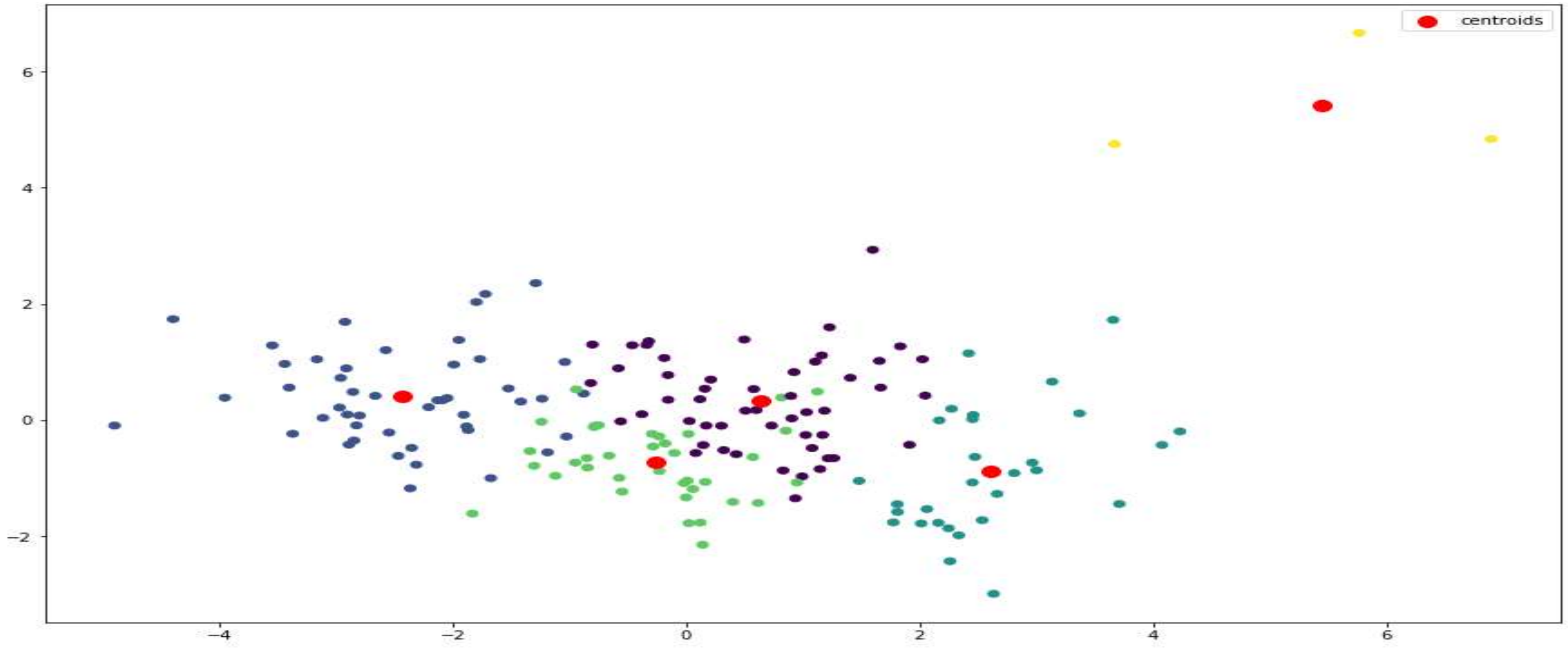
- The value of the silhouette score range lies between -1 to 1.
- A score closer to 1 indicates that the data point is very similar to other data points in the cluster,
- A score closer to -1 indicates that the data point is not similar to the data points in its cluster.



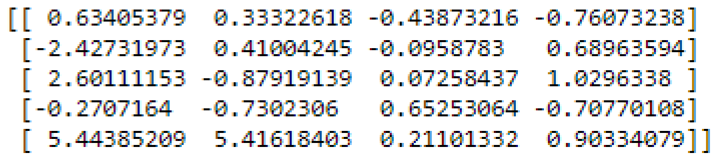
Result: From *above plot we can consider 5 as optimal number for k.*

Clustering analysis (k-means clustering)

When we fit the k-means with 5 clusters on the PCA data and visualize we will get the below clusters with centroids as below



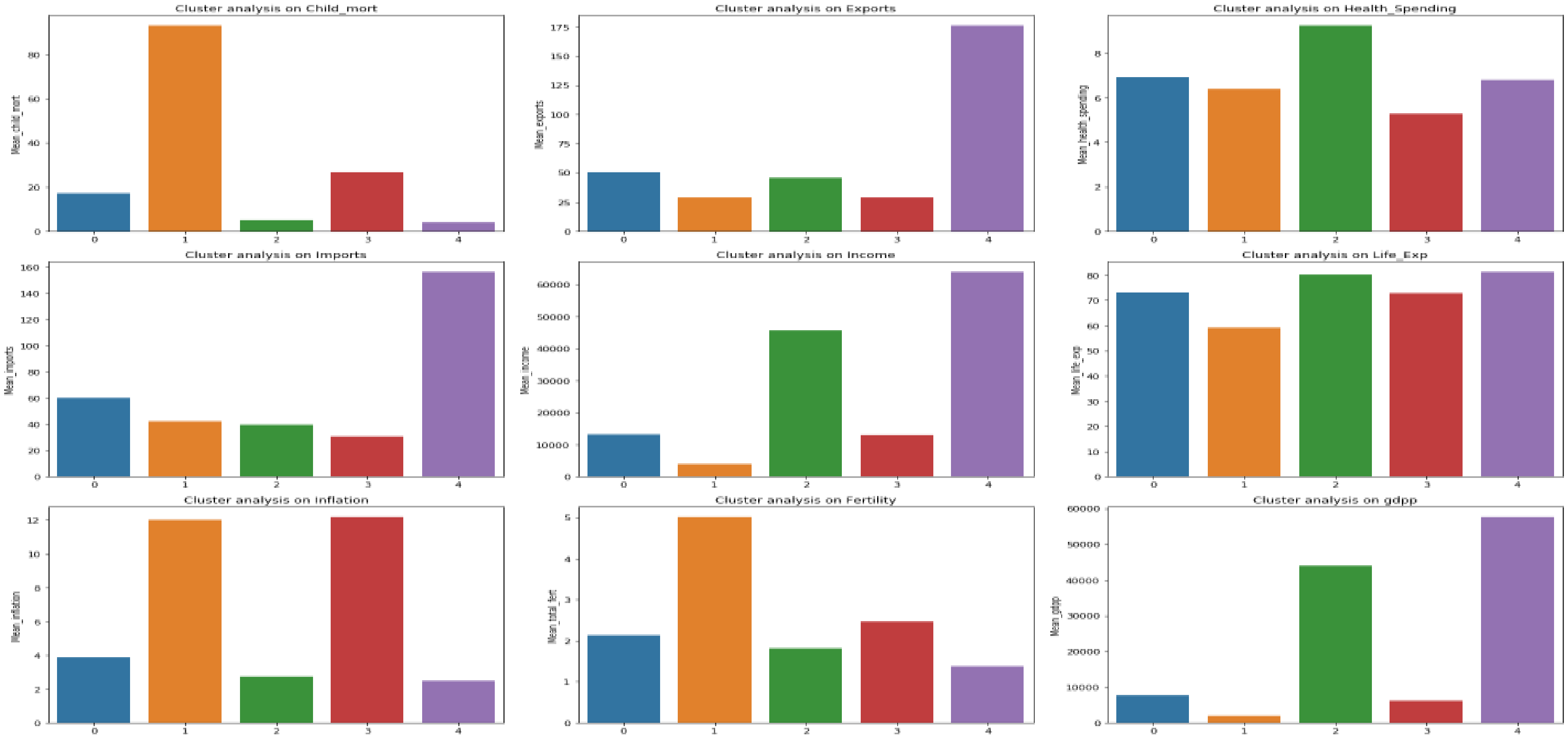
Result: We get the cluster centers as below



```
[[ 0.63405379  0.33322618 -0.43873216 -0.76073238]
 [-2.42731973  0.41004245 -0.0958783   0.68963594]
 [ 2.60111153 -0.87919139  0.07258437  1.0296338 ]
 [-0.2707164  -0.7302306   0.65253064 -0.70770108]
 [ 5.44385209  5.41618403  0.21101332  0.90334079]]
```

Conclusions- k-means clustering

When we plot the bar chart as below from the clusters obtained from k-means we will get the plot as below

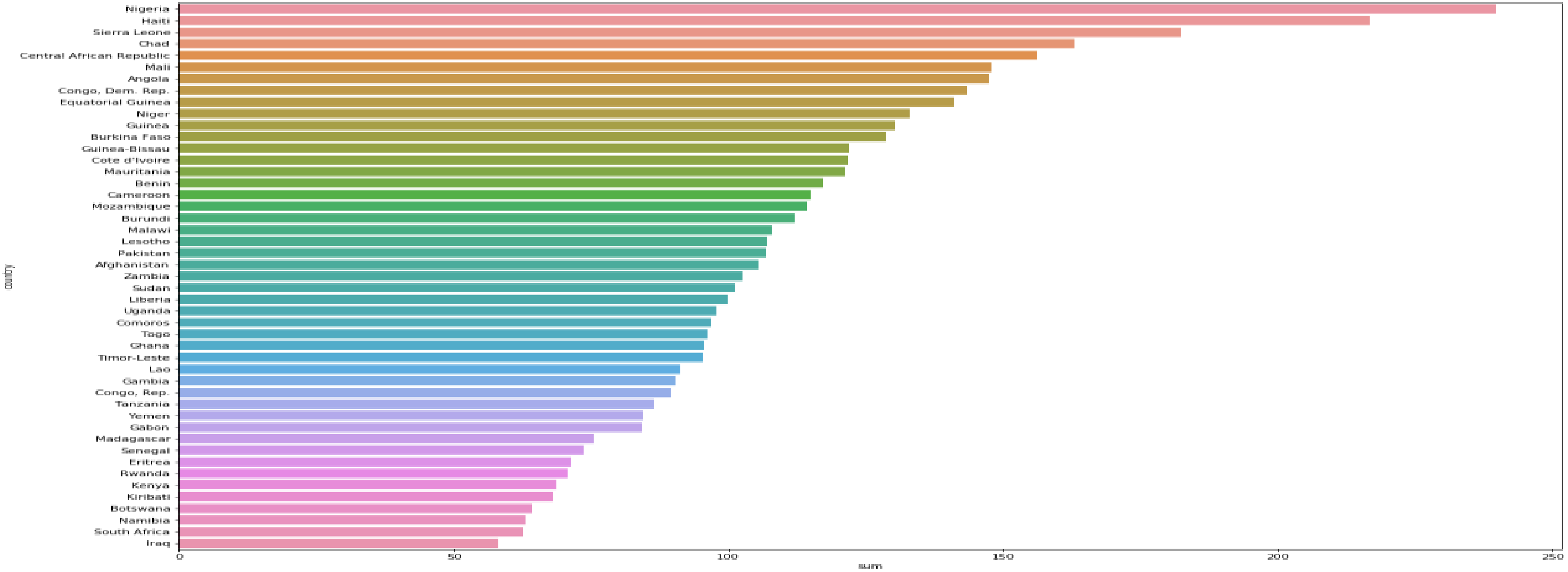


Result: From above plots it is evident that Cluster with ClusterID 1, is the cluster which consists of under developed countries and cluster with ClusterID 4, is the cluster which consists of developed countries

Conclusions- k-means clustering

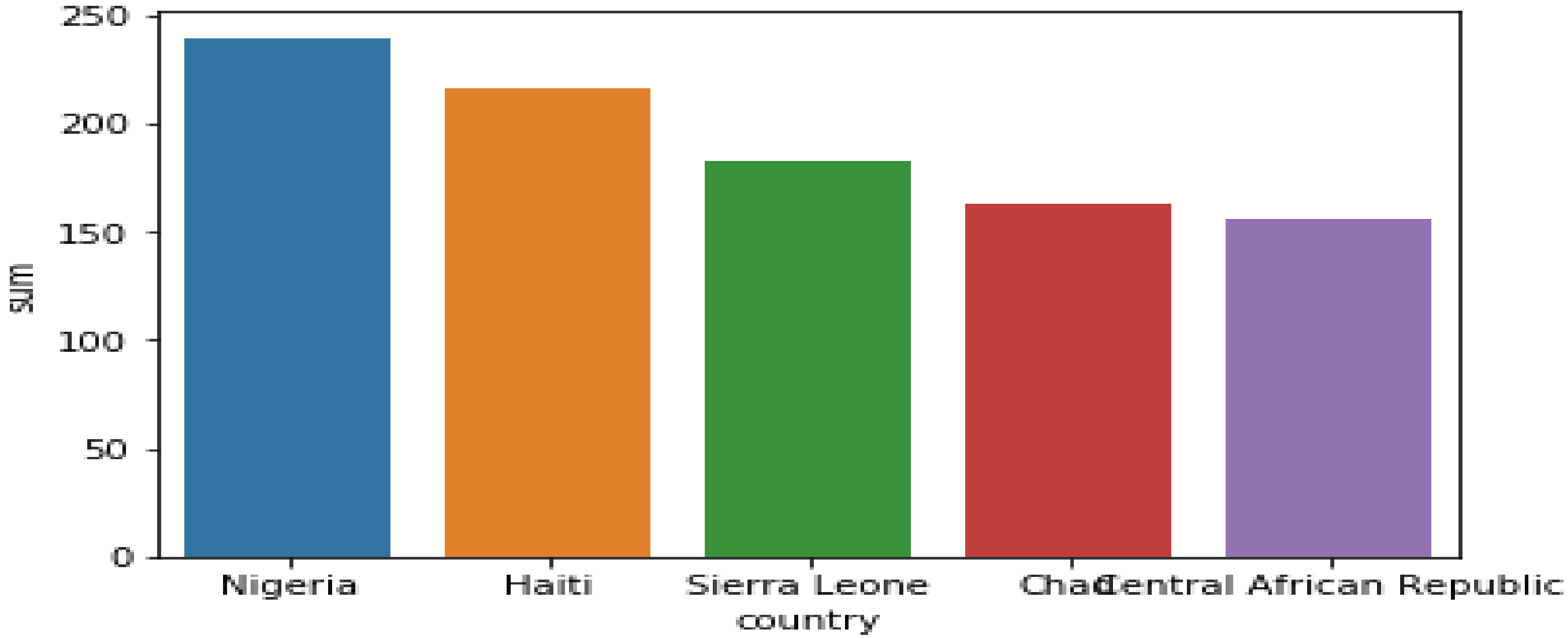
Final list of underdeveloped countries: In cluster 1 we got totally **47 list of under developed countries** and they are as follows

'Afghanistan', 'Angola', 'Benin', 'Botswana', 'Burkina Faso', 'Burundi', 'Cameroon', 'Central African Republic', 'Chad', 'Comoros', 'Congo, Dem. Rep.', 'Congo, Rep.', "Cote d'Ivoire", 'Equatorial Guinea', 'Eritrea', 'Gabon', 'Gambia', 'Ghana', 'Guinea', 'Guinea-Bissau', 'Haiti', 'Iraq', 'Kenya', 'Kiribati', 'Lao', 'Lesotho', 'Liberia', 'Madagascar', 'Malawi', 'Mali', 'Mauritania', 'Micronesia, Fed. Sts.', 'Mozambique', 'Namibia', 'Niger', 'Nigeria', 'Pakistan', 'Rwanda', 'Senegal', 'Sierra Leone', 'Solomon Islands', 'South Africa', 'Sudan', 'Tanzania', 'Timor-Leste', 'Togo', 'Uganda', 'Yemen', 'Zambia‘



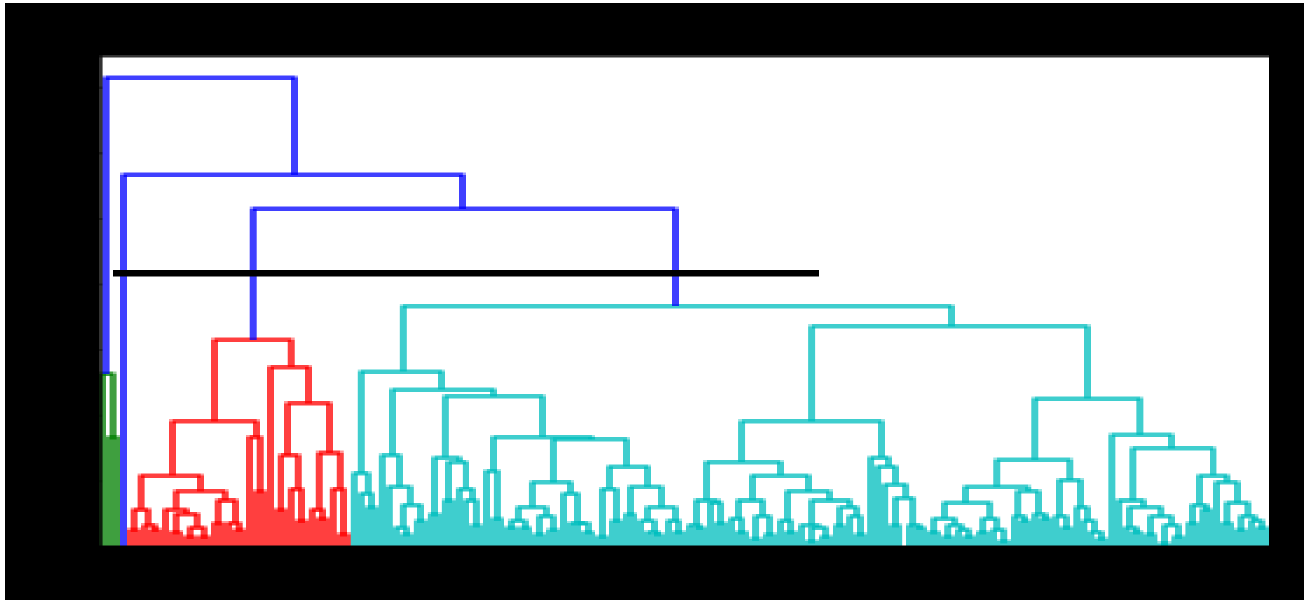
Conclusions- k-means clustering

Top 5 countries which are in dire need of financial aid: Below is the top 5 countries which requires financial aid. To find the Top 5 countries within cluster 1 we will consider 'child_mort', 'inflation', & 'total_fer' as factors to select top 5 countries as if the the countries which have high sum of these 3 columns can be considered as top 5. The top 5 countries which are in dire need of aid is **Nigeria, Haiti, Sierra Leone, Chad, Central African Republic**



Clustering analysis (Hierarchical clustering)

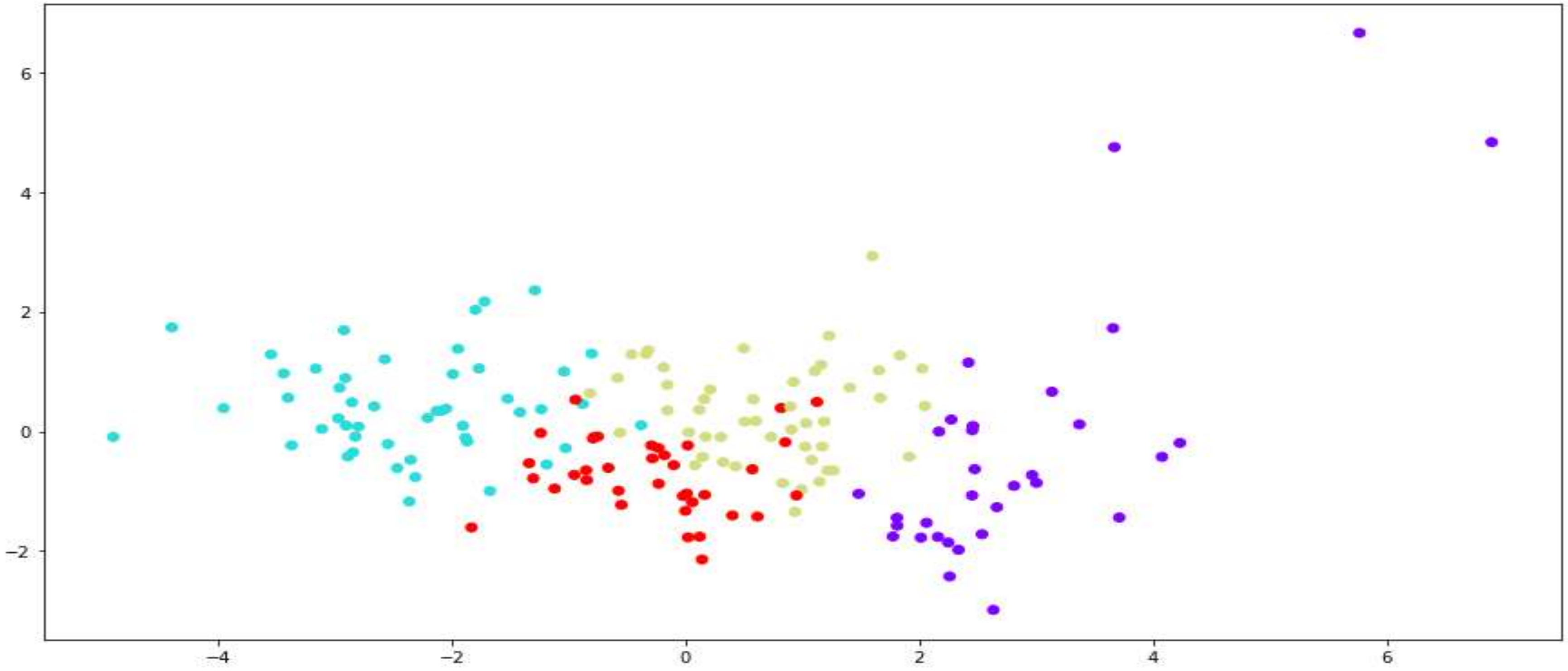
To perform Hierarchical clustering we need to find optimal number of clusters. So we can use dendogram.



- **Result**: From above dendogram it can be observed that algorithm has clustered data into 3 clusters by default(3 different colors). But to avoid under fitting but to get the better accuracy we will find out the number of clusters by checking the largest vertical distance without any horizontal line passing through it . So we draw a new horizontal black line that passes through the blue line. Since it crosses the blue line at four points, therefore the number of clusters will be 4.

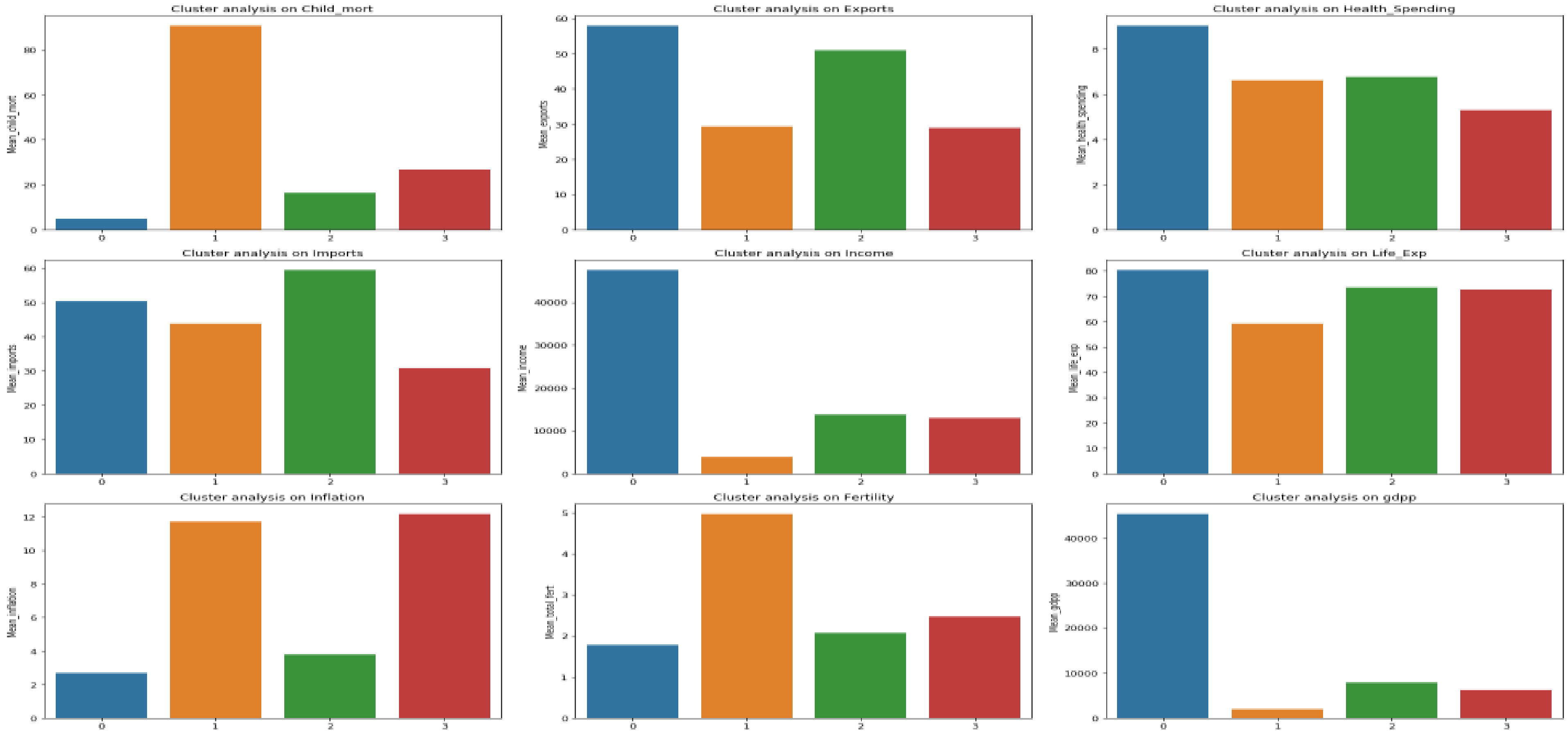
Clustering analysis (Hierarchical clustering)

Below is the scatter plot for clusters formed as part of Hierarchical clustering



Conclusions- Hierarchical clustering

When we plot the bar chart as below from the clusters obtained from hierarchical clustering we will get the plot as below

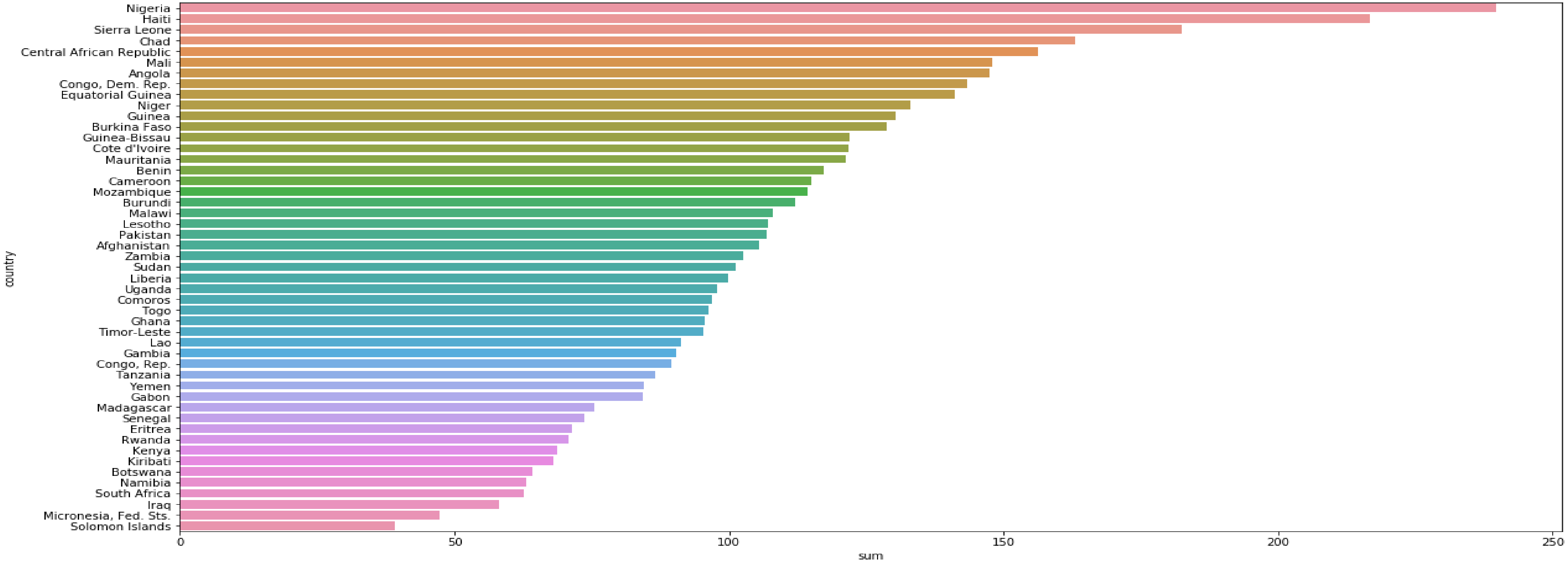


Result: From above plots it is evident that Cluster with ClusterID 1, is the cluster which consists of under developed countries and cluster with ClusterID 0, is the cluster which consists of developed countries

Conclusions- Hierarchical clustering

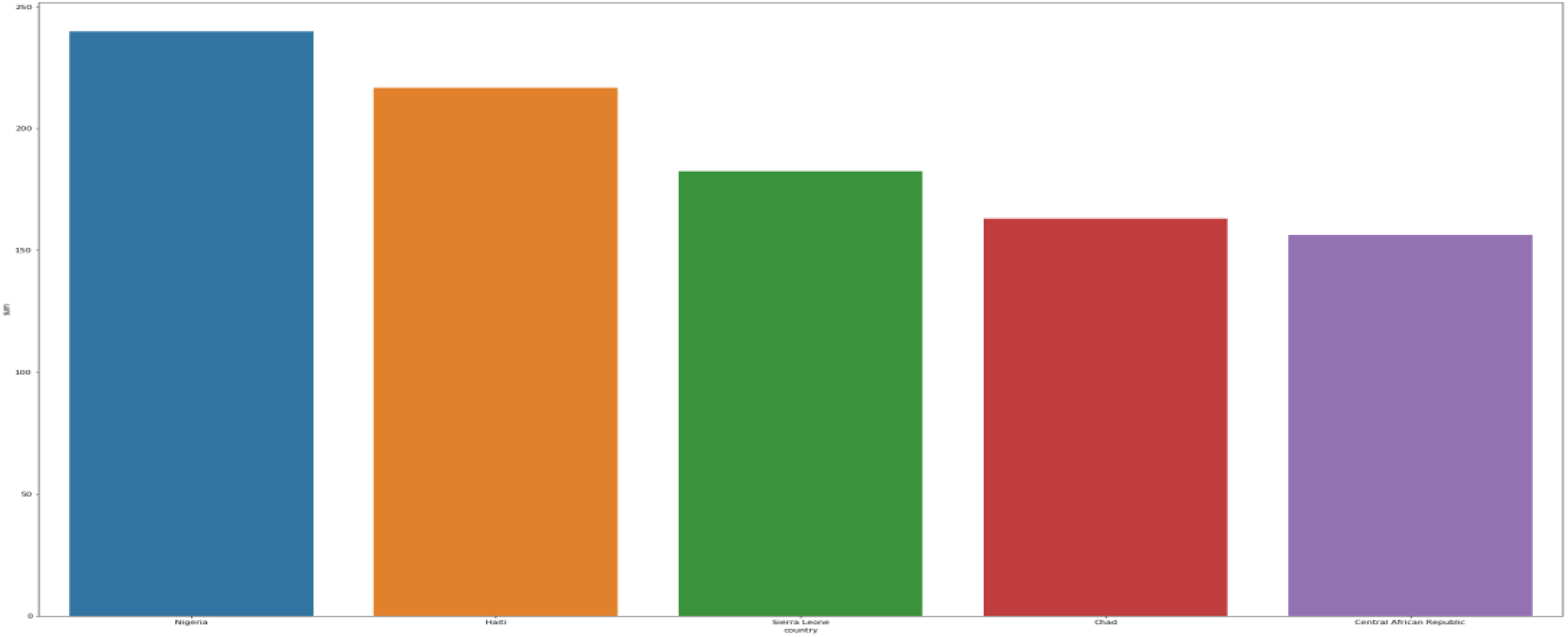
Final list of underdeveloped countries: In cluster 1 we got totally **49 list of under developed countries** and they are as follows

'Afghanistan', 'Angola', 'Benin', 'Botswana', 'Burkina Faso', 'Burundi', 'Cameroon', 'Central African Republic', 'Chad', 'Comoros', 'Congo, Dem. Rep.', 'Congo, Rep.', "Cote d'Ivoire", 'Equatorial Guinea', 'Eritrea', 'Gabon', 'Gambia', 'Ghana', 'Guinea', 'Guinea-Bissau', 'Haiti', 'Iraq', 'Kenya', 'Kiribati', 'Lao', 'Lesotho', 'Liberia', 'Madagascar', 'Malawi', 'Mali', 'Mauritania', 'Micronesia, Fed. Sts.', 'Mozambique', 'Namibia', 'Niger', 'Nigeria', 'Pakistan', 'Rwanda', 'Senegal', 'Sierra Leone', 'Solomon Islands', 'South Africa', 'Sudan', 'Tanzania', 'Timor-Leste', 'Togo', 'Uganda', 'Yemen', 'Zambia‘



Conclusions- Hierarchical clustering

Top 5 countries which are in dire need of financial aid: Below is the top 5 countries which requires financial aid. To find the Top 5 countries within cluster 1 we will consider 'child_mort', 'inflation', & 'total_fer' as factors to select top 5 countries as if the the countries which have high sum of these 3 columns can be considered as top 5. The top 5 countries which are in dire need of aid is **Nigeria, Haiti, Sierra Leone, Chad, Central African Republic**



Thank you