CSL2050:PRML

MINOR PROJECT

Unsupervised Learning on Country Data

By

Narkhede Kartik Sanjay (B21EE041)
Neel Aniruddha Barve (B21EE042)

Indian Institute of Technology, Jodhpur

April, 2023

ABSTRACT

The aim of this report is to categorize the countries using some socio-economic and health factors that determine the overall development of the country, to finally suggest the countries which the CEO needs to focus on the most. 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 has been able to raise around $ 10 million. Now the CEO of the NGO needs to decide how to use this money strategically and effectively. So, the CEO has to make a decision to choose the countries that are in the direst need of aid. In order to achieve this, we use multiple methods for clustering the data into the categories such as ‘Help needed’, ‘Might need help’ and ‘No help needed’.

Data description and preprocessing:

First of all we import the dataset and check if there are any missing values. We see that there are no missing values.

We also check for datatype of each feature whether it is a string or integer. If the datatype is string we convert it to integer using label encoder, however in this dataset there are only integer values.

Using df.describe(), we got an idea of mean values, standard deviation, quartile, minimum and maximum values of each parameter.

We also dropped the column named ‘countries’ as it is not useful for further analysis of data. We will get information about countries using the index number of dataframes.

We use heatmap to obtain correlation between different features.

After looking it , we can see that there is high positive correlation between some features and for some features , there is negative correlation

Positive Correlation shows high dependency between two features and vice versa

Findings from heatmap:

1. High GDP => Low child mortality rate, High total fertility=> high child mortality
2. High child mortality => Low life expectancy
3. Low Life Expectation => High Total Fertility

We use sns.Pairplot() to analyze the relationship between two quantitative variables measured for the same individuals.

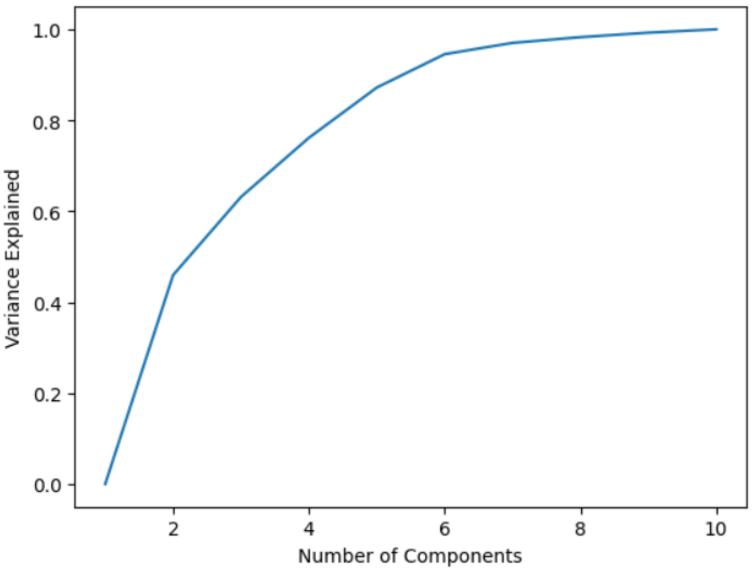
Then we use distplot to visualize the density distribution of each feature .

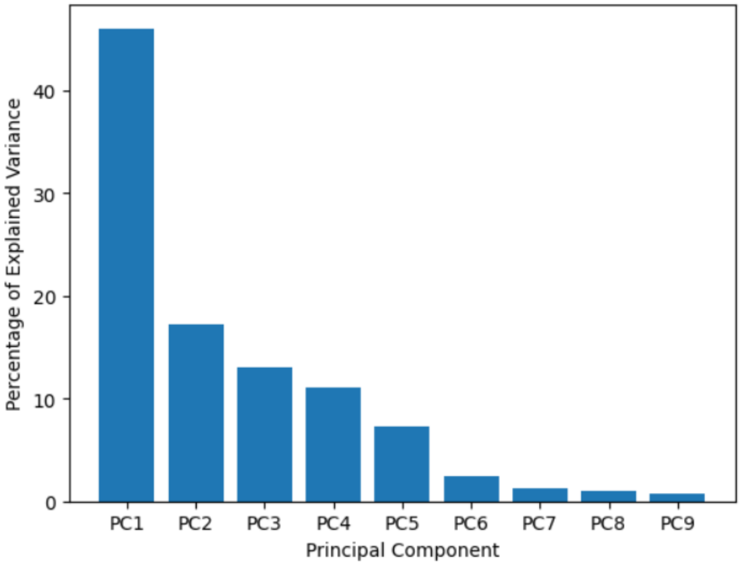
Data preprocessing and visualization:

1)PCA-

PCA (Principal Component Analysis) is a widely used dimensionality reduction technique in machine learning. It is used to reduce the number of features (or dimensions) of a dataset while retaining most of the original information in the data.
We use PCA for the following reasons:
1) Prevents the dominance of one feature
2)Standardization makes the model more robust to outliers and small changes in the input data. This is because the scale of the data is the same for all features, and outliers are scaled down to be similar in magnitude to the rest of the data.
3) helps improve the performance

Note: We are selecting the features with highest variability.



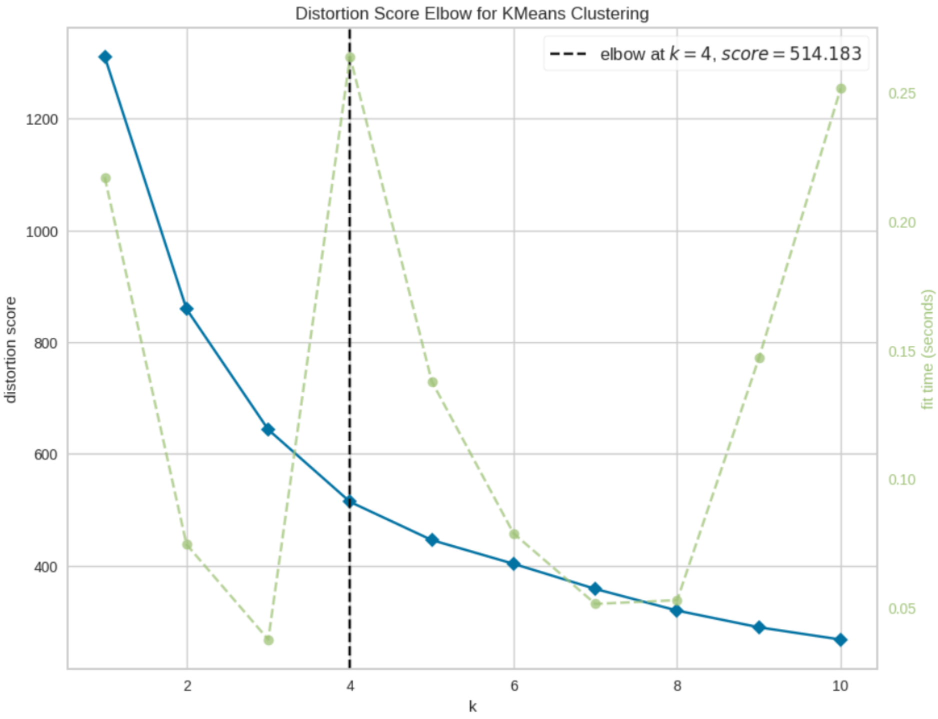


After doing PCA with a standardized dataset, we can see that there are 4 principal components that can explain about 90% of the distribution of the original data. So we dropped the remaining components like PCA5, PCA6, PCA7, PCA8, PCA9 i.ie we selected only the first four principal components.

Finding best cluster number on PCA data:

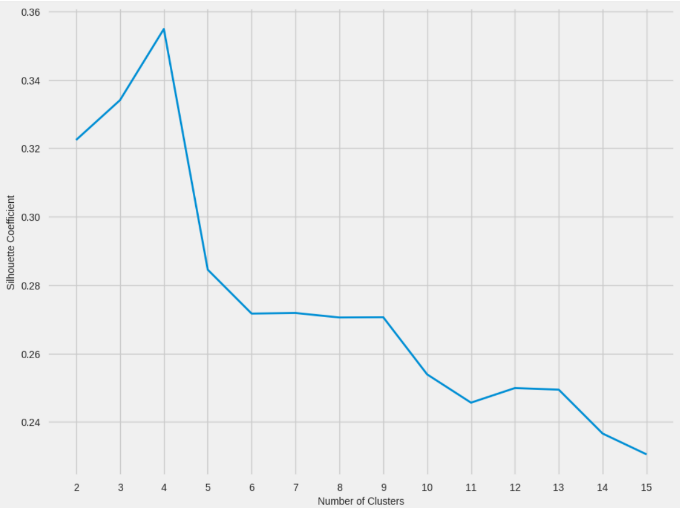
- **<u>K hyperparameter</u>**: It defines the number of clusters or groups the data is to be divided into. For the selection of values of k, we use 2 statistical tests :

  (i) **Elbow Method** : It is a method that plots the sum of squared error for a range of values of k. If this plot looks like an arm, then k is the value that resembles an elbow. From this elbow value, the sum of squared values starts decreasing in a linear fashion and thus is considered as an optimal value.



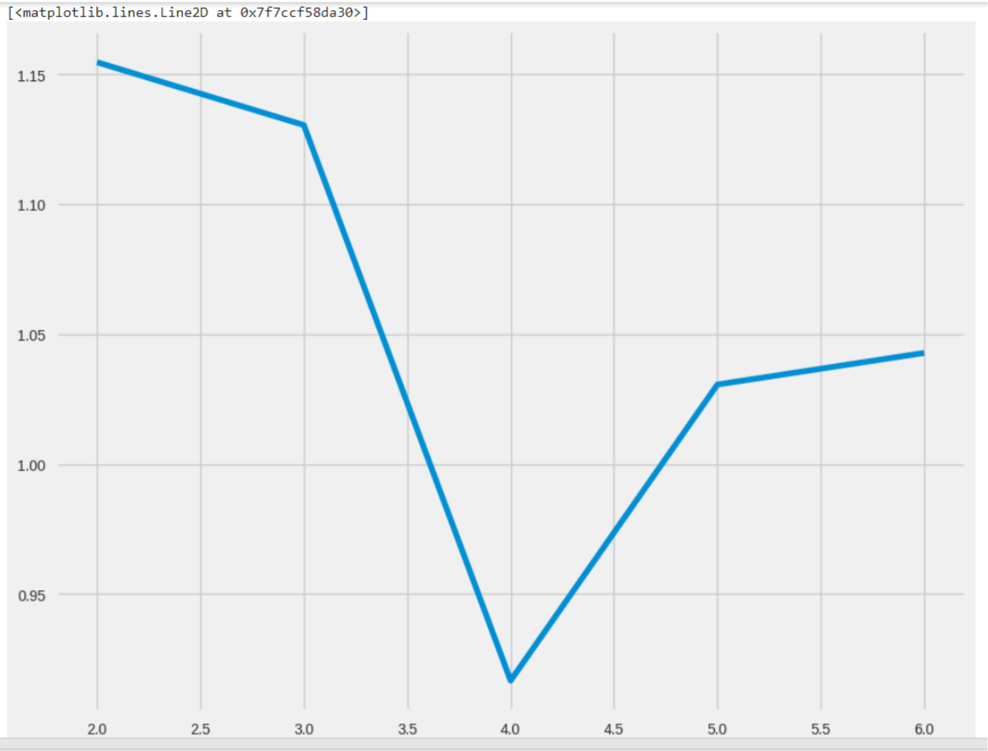
Finding: We are getting elbow at cluster_num=4

  (ii) **Silhouette Score Method** : It is a method that evaluates the quality of clusters in terms of how well data points are clustered with other data points that are similar to each other. This score is calculated using the distance formula and the k value with highest score is selected for modeling.



Finding: We are getting the highest silhouette score at 4 clusters.

(iii) **Davies Bouldin Score** : The minimum score is zero, with lower values indicating better clustering.



```
[<matplotlib.lines.Line2D at 0x7f7ccf58da30>]
```

Finding: for no. of clusters=4, we get minimum Davies Bouldin score.

Conclusion for PCA data: From the Elbow method , Silhouette coefficient and Davies Bouldin Score, we can conclude that the optimal number of clusters for the data is 4.

Finding best cluster number on Scaled data:

We got the similar results as the PCA data, so we are using the number of clusters = 4 for further training models.

We are using the following 7 clustering models for learning on the country data.

CLUSTERING METHODS:

METHOD 1: K-MEANS

METHOD 2: Hierarchical clustering

METHOD 3: DBSCAN

METHOD 4: Spectral Clustering

METHOD 5: Mean Shift Clustering

METHOD 6: Gaussian Mixture Model clustering

METHOD 7: Using Auto-encoder

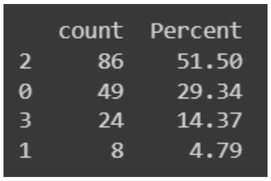
About each method and Reasons for using them:

1. K- Means Clustering

K-means is an unsupervised learning algorithm, which means that it does not require labeled data to learn. Instead, it identifies patterns and structures in the data on its own.
K-means is scalable, which means it can handle large datasets efficiently.
It is generally faster than other clustering algorithms, such as hierarchical clustering or DBSCAN because it is an iterative algorithm that converges quickly to a solution.

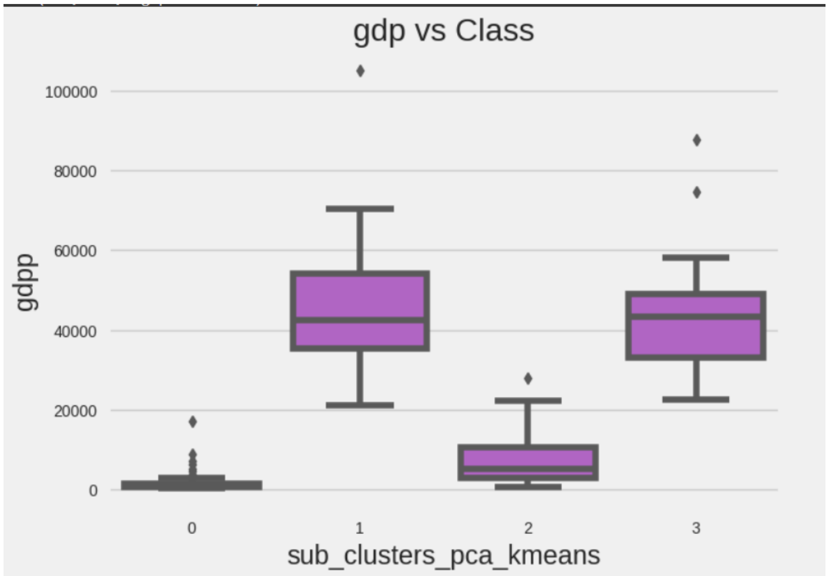
I. K Means on PCA data:

We applied K-means clustering on PCA data with n_clusters = 4. We got the following results:

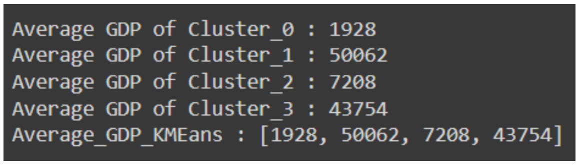


```
   count  Percent
2     86    51.50
0     49    29.34
3     24    14.37
1      8     4.79
```

We draw a boxplot of income, child mortality rate & GDP w.r.t labeled clusters as we do not know which cluster corresponds to the group of countries that are in need. We know that low income, high child mortality & low gdp is a sign of an economically backward nation.

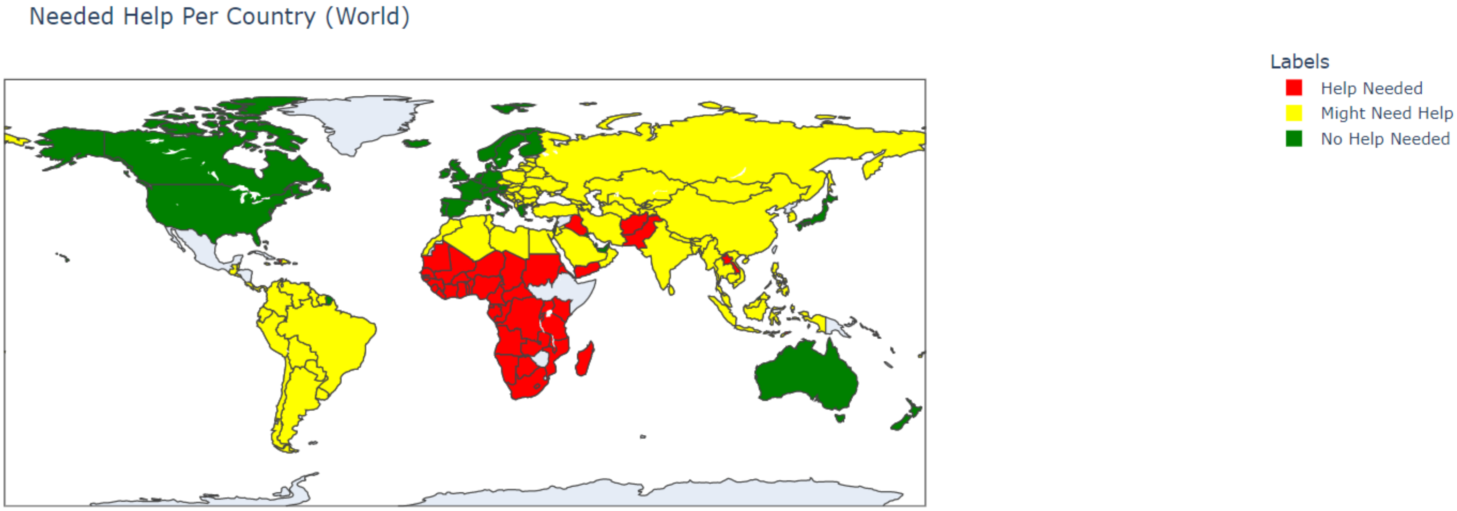


We calculated the average GDP of cluster:

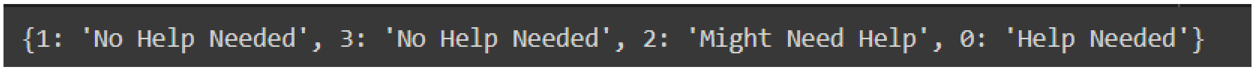


```
Average GDP of Cluster_0 : 1928
Average GDP of Cluster_1 : 50062
Average GDP of Cluster_2 : 7208
Average GDP of Cluster_3 : 43754
Average_GDP_KMEans : [1928, 50062, 7208, 43754]
```

Using this, the cluster assignment function(written from scratch) assigns labels to the clusters.

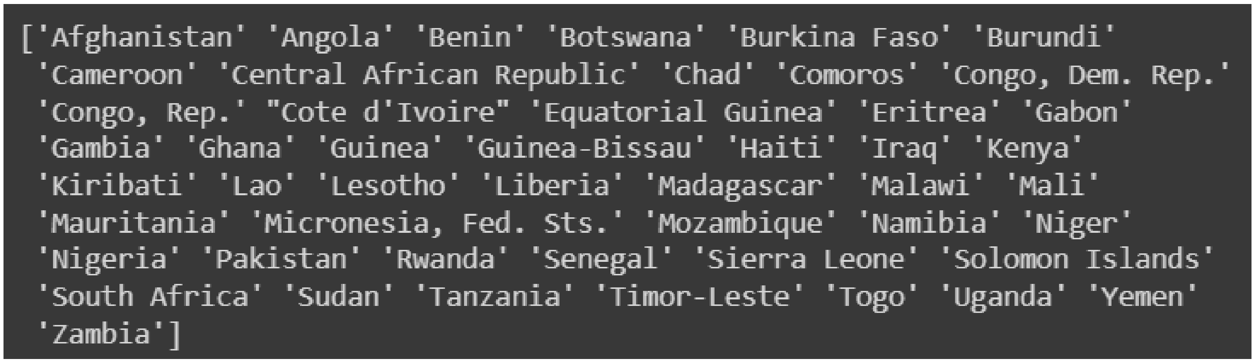


Cluster Assignment:



```
{1: 'No Help Needed', 3: 'No Help Needed', 2: 'Might Need Help', 0: 'Help Needed'}
```

We got the following countries as those who Need help:

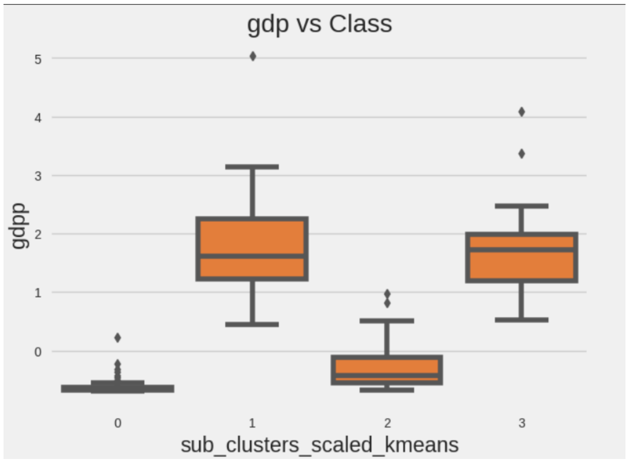


```
['Afghanistan' 'Angola' 'Benin' 'Botswana' 'Burkina Faso' 'Burundi'
 'Cameroon' 'Central African Republic' 'Chad' 'Comoros' 'Congo, Dem. Rep.'
 'Congo, Rep.' "Cote d'Ivoire" 'Equatorial Guinea' 'Eritrea' 'Gabon'
 'Gambia' 'Ghana' 'Guinea' 'Guinea-Bissau' 'Haiti' 'Iraq' 'Kenya'
 'Kiribati' 'Lao' 'Lesotho' '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']
```

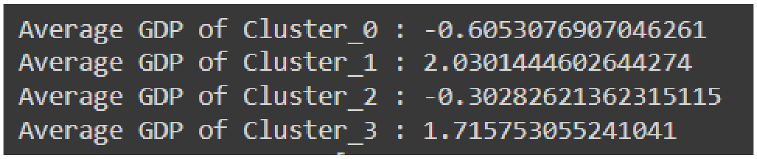
For further analysis, we are also maintaining a dictionary of frequency of needy countries after running it on all models.

II. K-Means on scaled data:

Following are the boxplots:

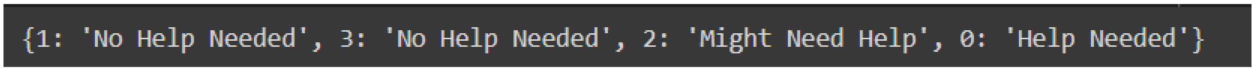


Using boxplots and average GDP of each cluster, we assigned labels to each cluster.

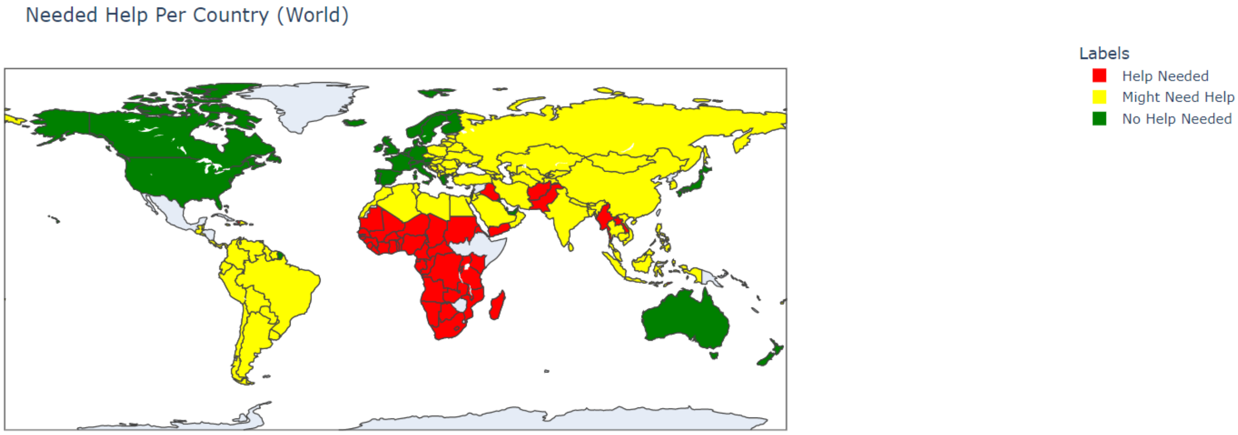


```
Average GDP of Cluster_0 : -0.6053076907046261
Average GDP of Cluster_1 : 2.0301444602644274
Average GDP of Cluster_2 : -0.30282621362315115
Average GDP of Cluster_3 : 1.715753055241041
```

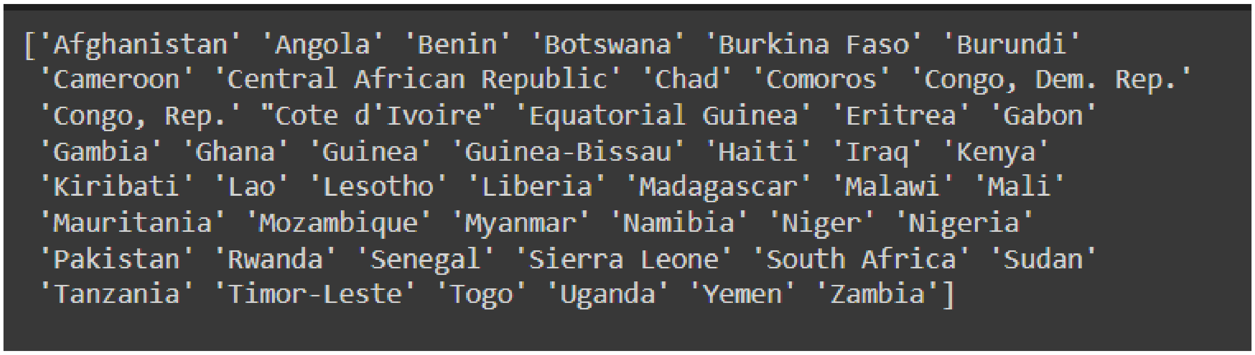
Cluster assignment:



```
{1: 'No Help Needed', 3: 'No Help Needed', 2: 'Might Need Help', 0: 'Help Needed'}
```



Countries that need help:

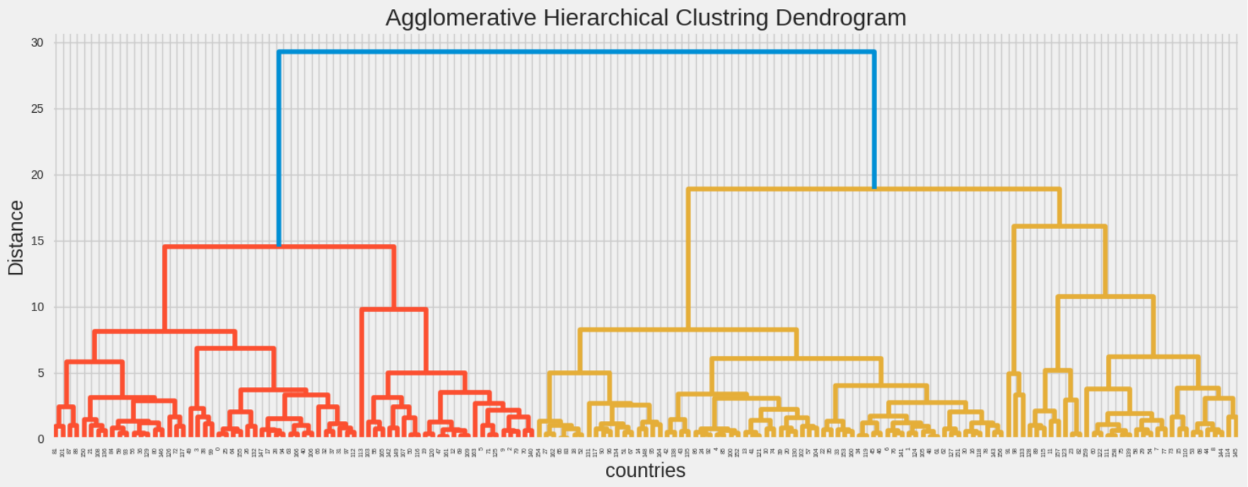


```
['Afghanistan' 'Angola' 'Benin' 'Botswana' 'Burkina Faso' 'Burundi'
 'Cameroon' 'Central African Republic' 'Chad' 'Comoros' 'Congo, Dem. Rep.'
 'Congo, Rep.' "Cote d'Ivoire" 'Equatorial Guinea' 'Eritrea' 'Gabon'
 'Gambia' 'Ghana' 'Guinea' 'Guinea-Bissau' 'Haiti' 'Iraq' 'Kenya'
 'Kiribati' 'Lao' 'Lesotho'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

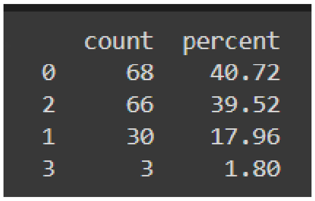
2. <u>Hierarchical clustering</u>:

Hierarchical clustering groups data over a variety of scales by creating a cluster tree or dendrogram. The tree is not a single set of clusters, but rather a multilevel hierarchy, where clusters at one level are joined as clusters at the next level. This allows us to decide the level or scale of clustering that is most appropriate for our application.

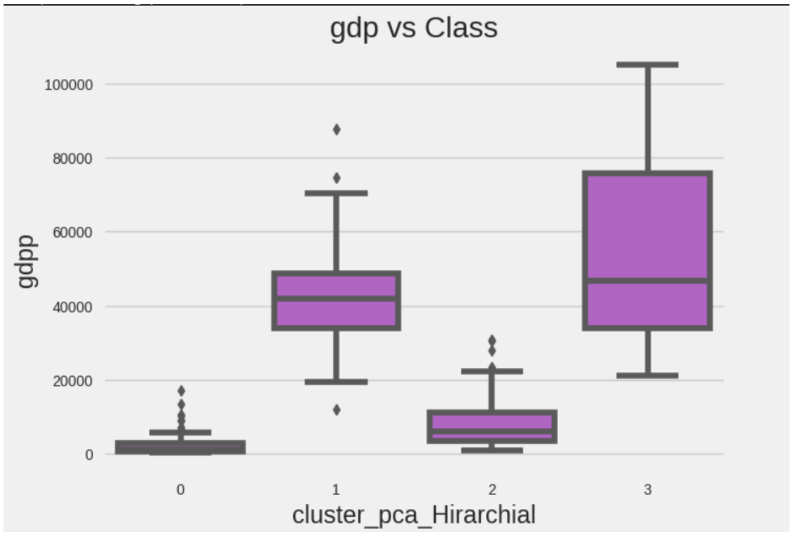
I. In PCA



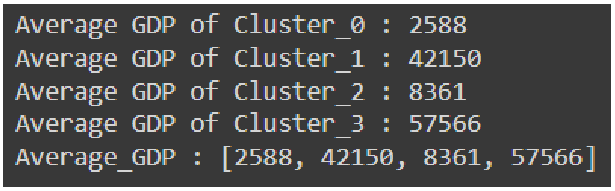
We got the above results on applying hierarchical clustering:



```
   count  percent
0     68    40.72
2     66    39.52
1     30    17.96
3      3     1.80
```

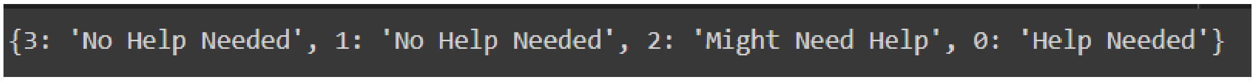


Using boxplots and average GDP of each cluster, we assigned labels to each cluster.

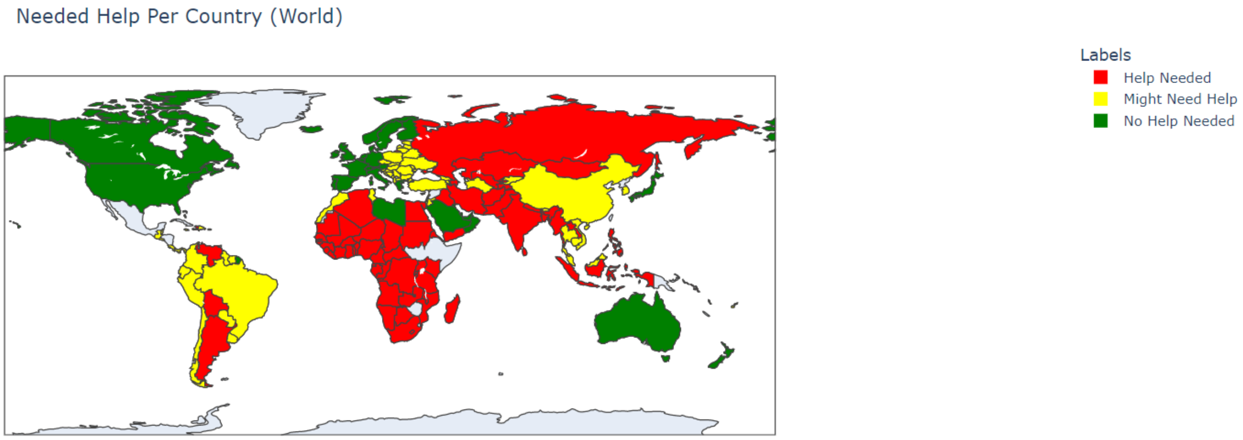


```
Average GDP of Cluster_0 : 2588
Average GDP of Cluster_1 : 42150
Average GDP of Cluster_2 : 8361
Average GDP of Cluster_3 : 57566
Average_GDP : [2588, 42150, 8361, 57566]
```

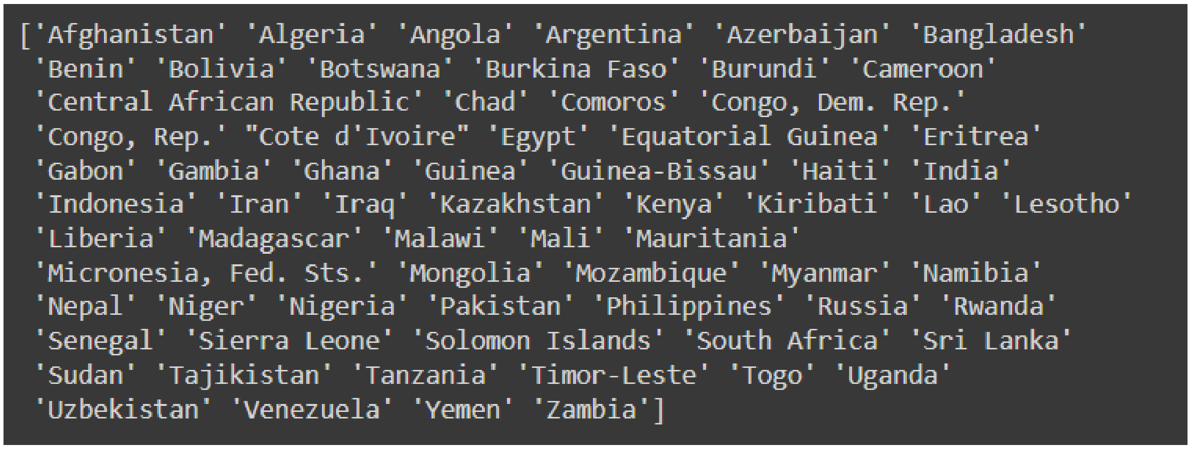
Cluster assignment:



```
{3: 'No Help Needed', 1: 'No Help Needed', 2: 'Might Need Help', 0: 'Help Needed'}
```

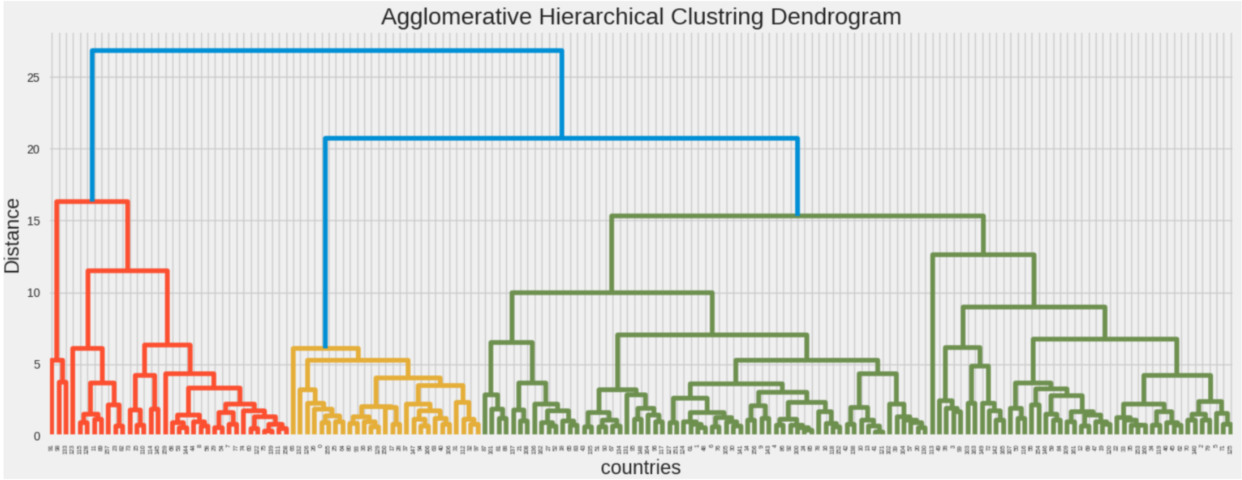


Countries that need help:

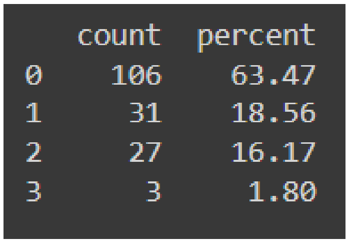


```
['Afghanistan' 'Algeria' 'Angola' 'Argentina' 'Azerbaijan' 'Bangladesh'
 'Benin' 'Bolivia' 'Botswana' 'Burkina Faso' 'Burundi' 'Cameroon'
 'Central African Republic' 'Chad' 'Comoros' 'Congo, Dem. Rep.'
 'Congo, Rep.' "Cote d'Ivoire" 'Egypt' 'Equatorial Guinea' 'Eritrea'
 'Gabon' 'Gambia' 'Ghana' 'Guinea' 'Guinea-Bissau' 'Haiti' 'India'
 'Indonesia' 'Iran' 'Iraq' 'Kazakhstan' 'Kenya' 'Kiribati' 'Lao' 'Lesotho'
 'Liberia' 'Madagascar' 'Malawi' 'Mali' 'Mauritania'
 'Micronesia, Fed. Sts.' 'Mongolia' 'Mozambique' 'Myanmar' 'Namibia'
 'Nepal' 'Niger' 'Nigeria' 'Pakistan' 'Philippines' 'Russia' 'Rwanda'
 'Senegal' 'Sierra Leone' 'Solomon Islands' 'South Africa' 'Sri Lanka'
 'Sudan' 'Tajikistan' 'Tanzania' 'Timor-Leste' 'Togo' 'Uganda'
 'Uzbekistan' 'Venezuela' 'Yemen' 'Zambia']
```

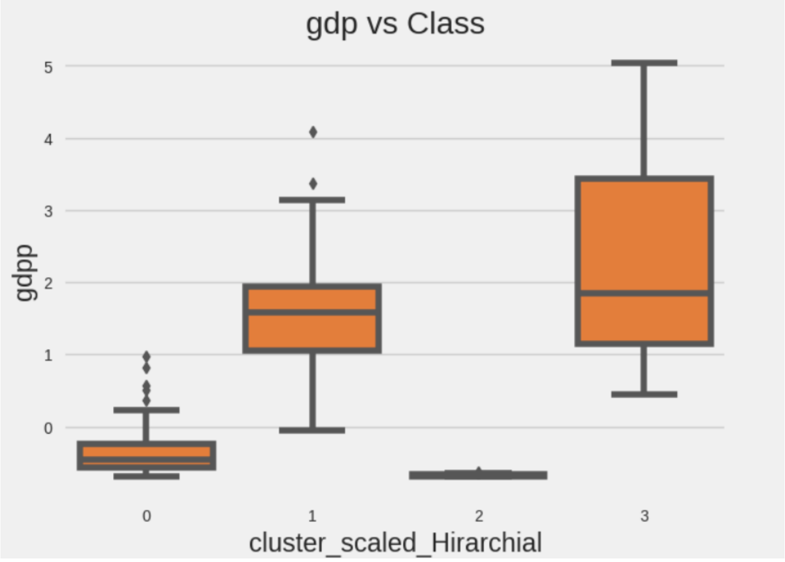
II. On Scaled data



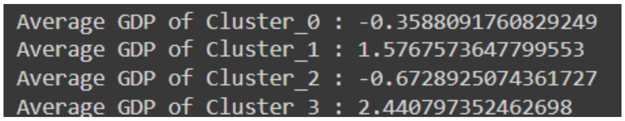
We got the above results on applying hierarchical clustering:



```
   count  percent
0    106    63.47
1     31    18.56
2     27    16.17
3      3     1.80
```

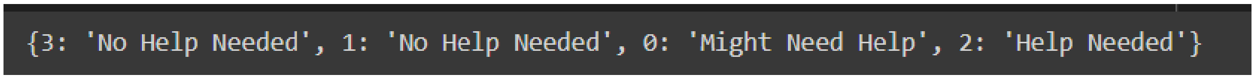


Using boxplots and average GDP of each cluster, we assigned labels to each cluster.



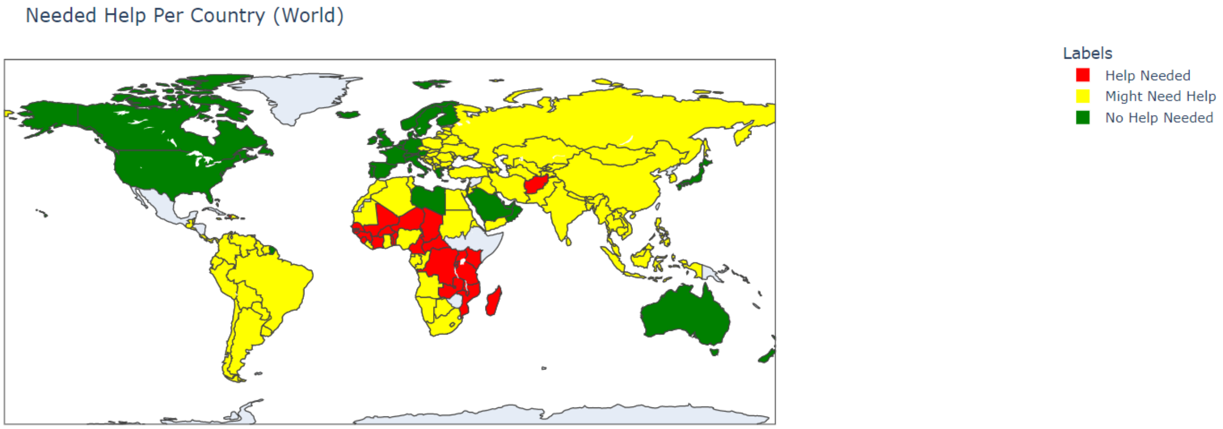
```
Average GDP of Cluster_0 : -0.3588091760829249
Average GDP of Cluster_1 : 1.5767573647799553
Average GDP of Cluster_2 : -0.6728925074361727
Average GDP of Cluster_3 : 2.440797352462698
```

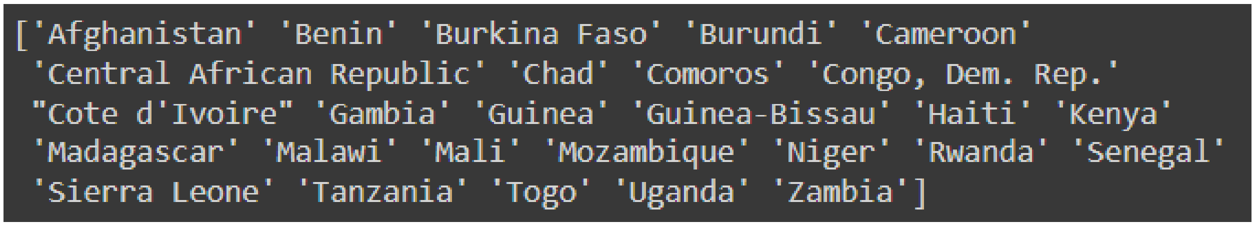
Cluster assignment:



```
{3: 'No Help Needed', 1: 'No Help Needed', 0: 'Might Need Help', 2: 'Help Needed'}
```

Countries that need help:





```
['Afghanistan' 'Benin' 'Burkina Faso' 'Burundi' 'Cameroon'
 'Central African Republic' 'Chad' 'Comoros' 'Congo, Dem. Rep.'
 "Cote d'Ivoire" 'Gambia' 'Guinea' 'Guinea-Bissau' 'Haiti' 'Kenya'
 'Madagascar' 'Malawi' 'Mali' 'Mozambique' 'Niger' 'Rwanda' 'Senegal'
 'Sierra Leone' 'Tanzania' 'Togo' 'Uganda' 'Zambia']
```

3. DBSCAN

It is a clustering algorithm that illustrates an approach based on local density estimation. This approach allows the algorithm to identify clusters of arbitrary shapes.

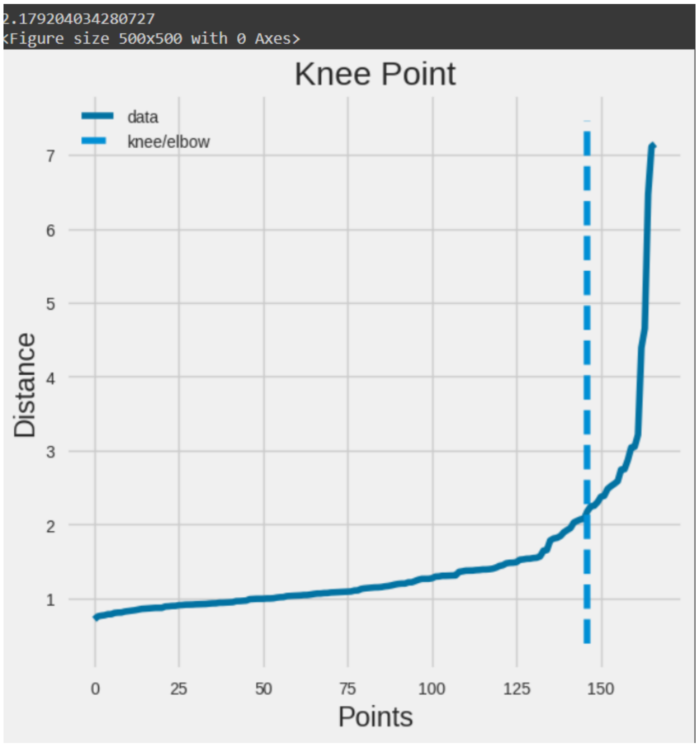
This algorithm defines clusters as continuous regions of high density. It works in following way:

I. For each instance, the algorithm counts how many instances are located within a small distance ε (epsilon) from it. This region is called the instance’s εneighborhood.
II. If an instance has at least min_samples instances in its ε-neighborhood (including itself), then it is considered a core instance. In other words, core instances are those that are located in dense regions.
III. All instances in the neighborhood of a core instance belong to the same cluster.This may include other core instances, therefore a long sequence of neighboring core instances forms a single cluster.
IV. Any instance that is not a core instance and does not have one in its neighborhood is considered an anomaly.

This algorithm works well if all the clusters are dense enough, and they are well separated by low-density regions.

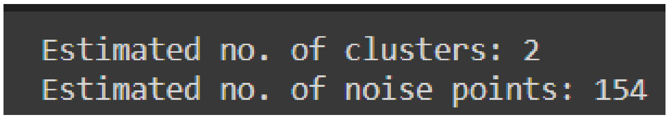
A. Using PCA

Finding optimum value of epsilon:



```
2.179204034280727
<Figure size 500x500 with 0 Axes>
```

We got following results from DBSCAN:

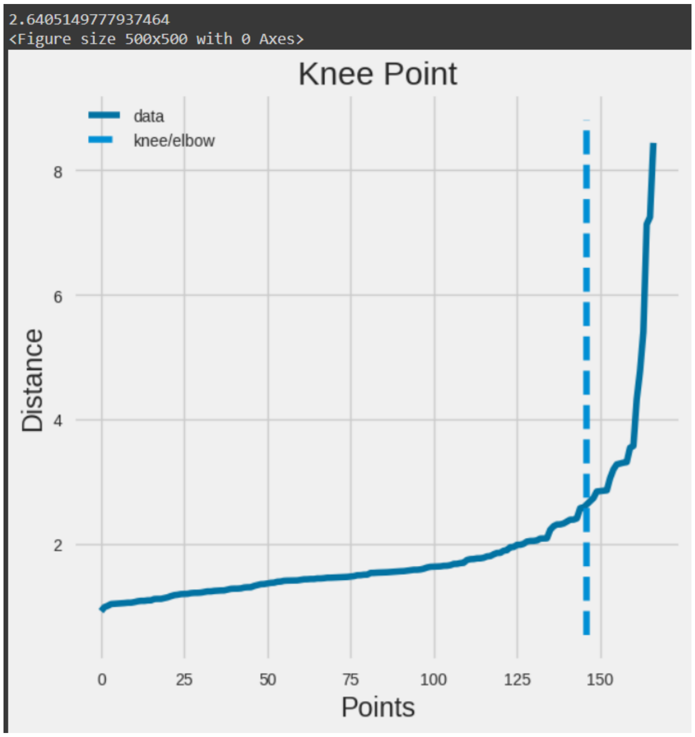


```
Estimated no. of clusters: 2
Estimated no. of noise points: 154
```

Conclusion:

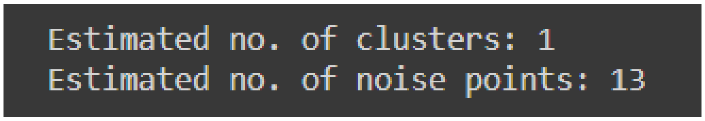
We got only 2 clusters with 154 outliers, so for further analysis, we are not considering this model.

B. Using scaled dataset



```
2.6405149777937464
<Figure size 500x500 with 0 Axes>
```

We got following results:



```
Estimated no. of clusters: 1
Estimated no. of noise points: 13
```

Conclusion:

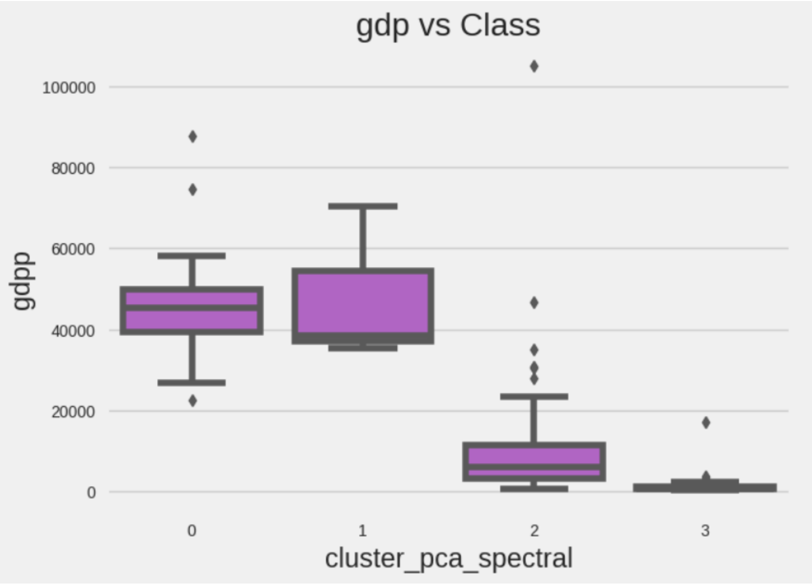
We got only 1 cluster with 13 outliers, so for further analysis, we are not considering this model.

4. Spectral Clustering

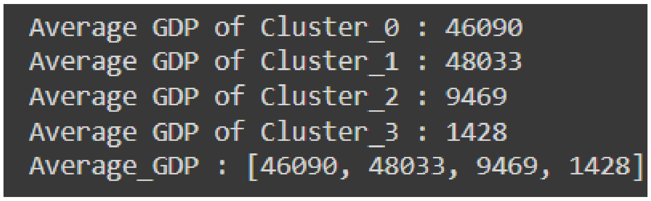
I. Spectral clustering algorithm takes a similarity matrix between the instances and creates a low-dimensional embedding from it (reduces its dimensionality), then it uses another clustering algorithm in this low-dimensional space.

II. Spectral clustering can capture complex cluster structures, however it does not scale well to a large number of instances, and it does not behave well when the clusters have very different sizes.

A. On PCA

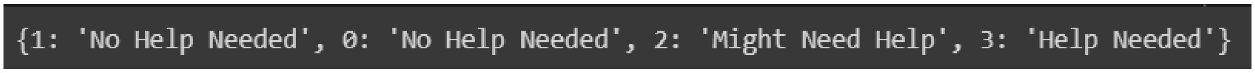


Using boxplots and average GDP of each cluster, we assigned labels to each cluster.

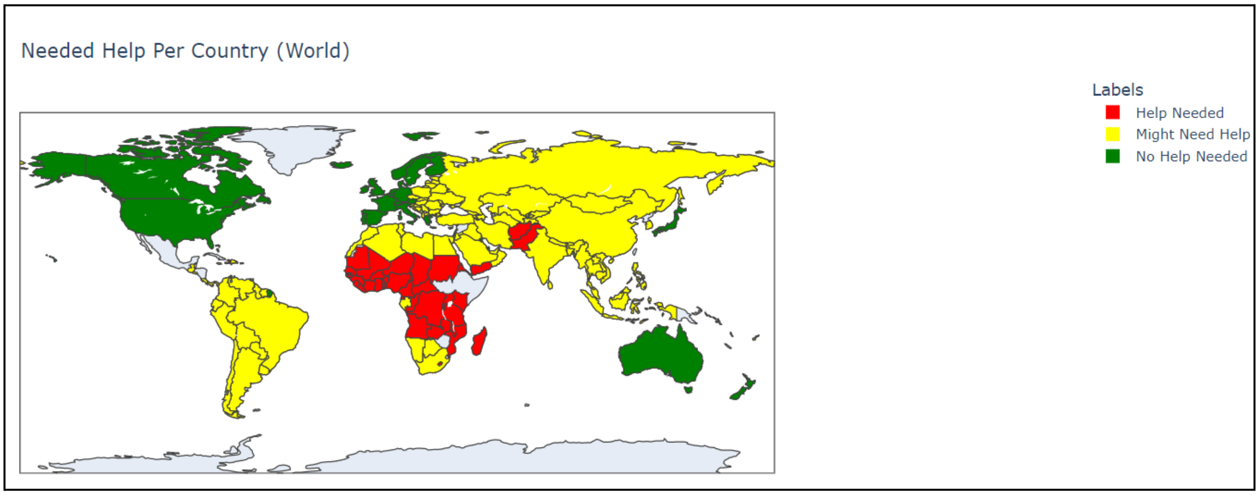


```
Average GDP of Cluster_0 : 46090
Average GDP of Cluster_1 : 48033
Average GDP of Cluster_2 : 9469
Average GDP of Cluster_3 : 1428
Average_GDP : [46090, 48033, 9469, 1428]
```

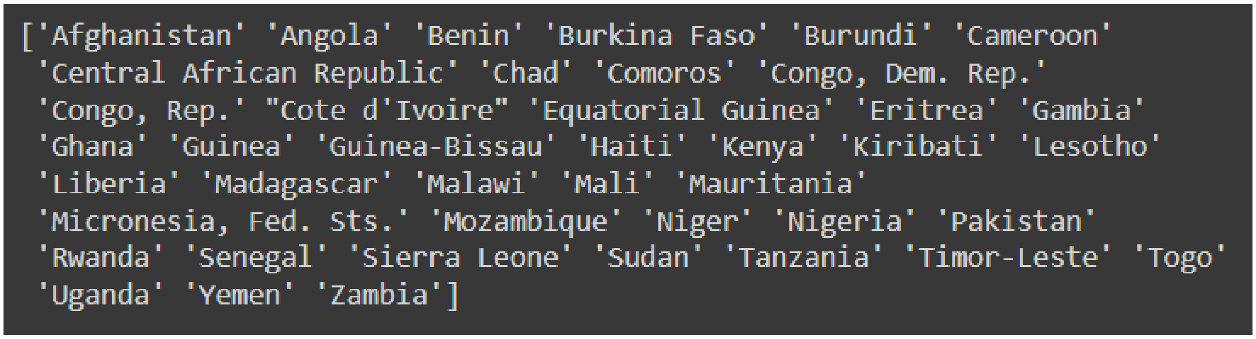
Cluster assignment is as follows:



```
{1: 'No Help Needed', 0: 'No Help Needed', 2: 'Might Need Help', 3: 'Help Needed'}
```

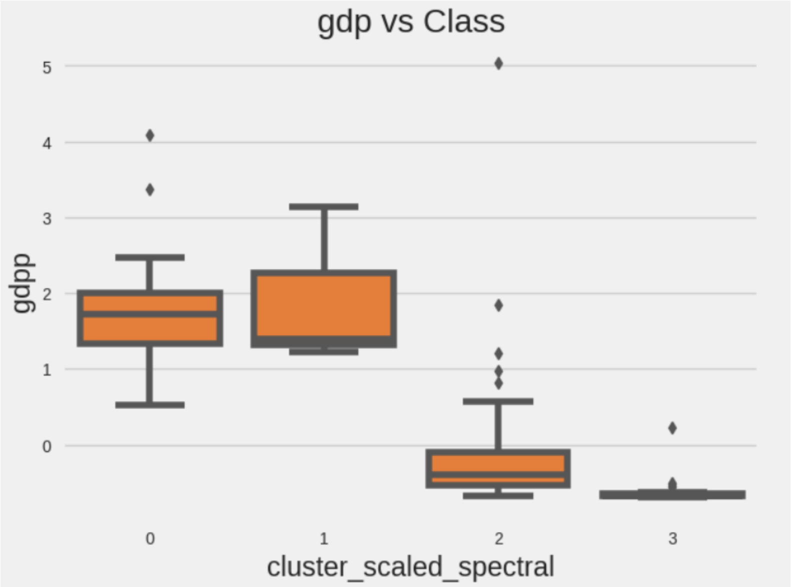


Following countries need help:

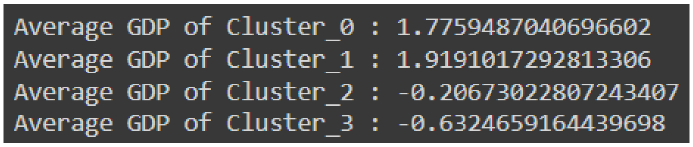


```
['Afghanistan' 'Angola' 'Benin' 'Burkina Faso' 'Burundi' 'Cameroon'
 'Central African Republic' 'Chad' 'Comoros' 'Congo, Dem. Rep.'
 'Congo, Rep.' "Cote d'Ivoire" 'Equatorial Guinea' 'Eritrea' 'Gambia'
 'Ghana' 'Guinea' 'Guinea-Bissau' 'Haiti' 'Kenya' 'Kiribati' 'Lesotho'
 'Liberia' 'Madagascar'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

B. On scaled data

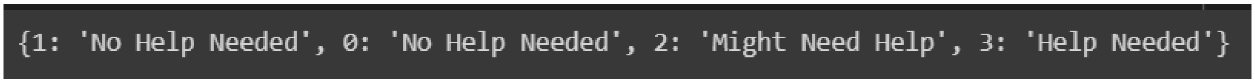


Using boxplots and average GDP of each cluster, we assigned labels to each cluster.

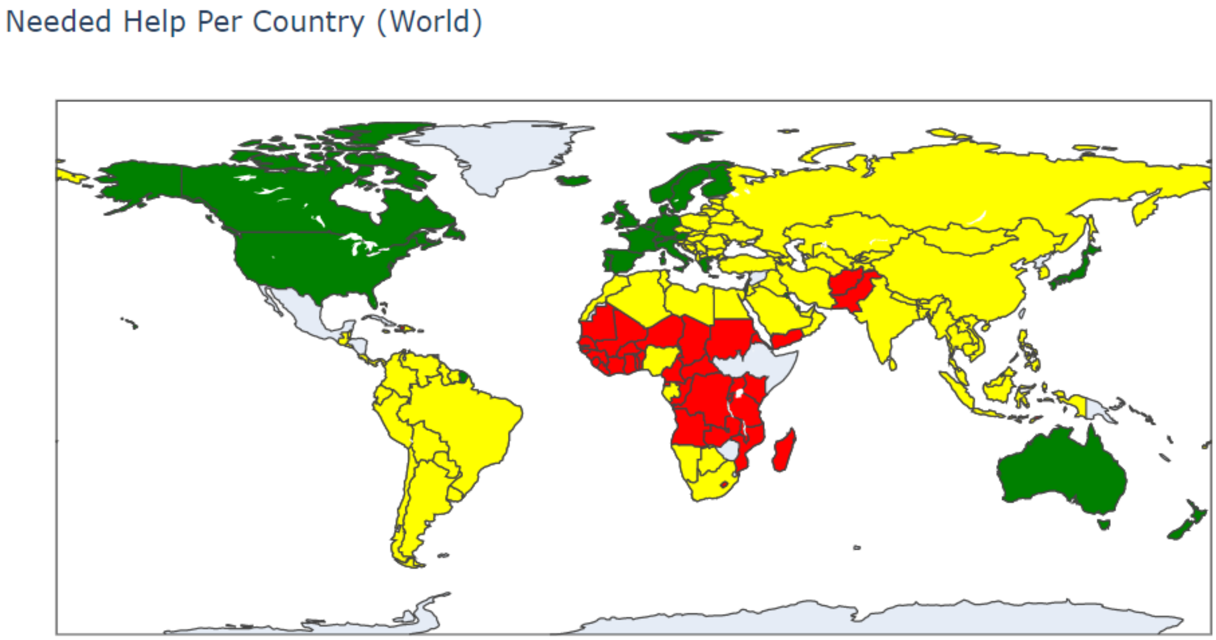


```
Average GDP of Cluster_0 : 1.7759487040696602
Average GDP of Cluster_1 : 1.9191017292813306
Average GDP of Cluster_2 : -0.20673022807243407
Average GDP of Cluster_3 : -0.6324659164439698
```

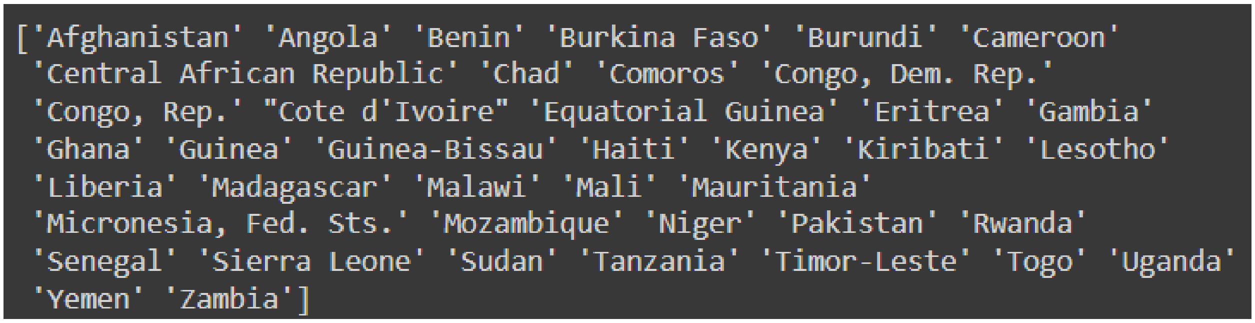
Cluster assignment as follows:



```
{1: 'No Help Needed', 0: 'No Help Needed', 2: 'Might Need Help', 3: 'Help Needed'}
```



Following countries are in need of help:



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

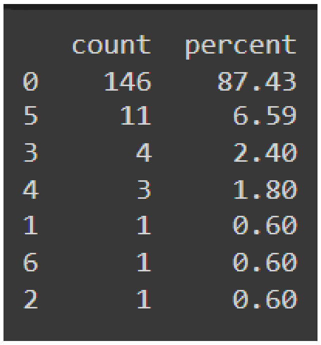
5. Mean Shift Clustering

I. This algorithm starts by placing a circle centered on each instance, then for each circle it computes the mean of all the instances located within it, and it shifts the circle so that it is centered on the mean. Next, it iterates this mean-shift step until all the circles stop moving (i.e., until each of them is centered on the mean of the instances it contains).

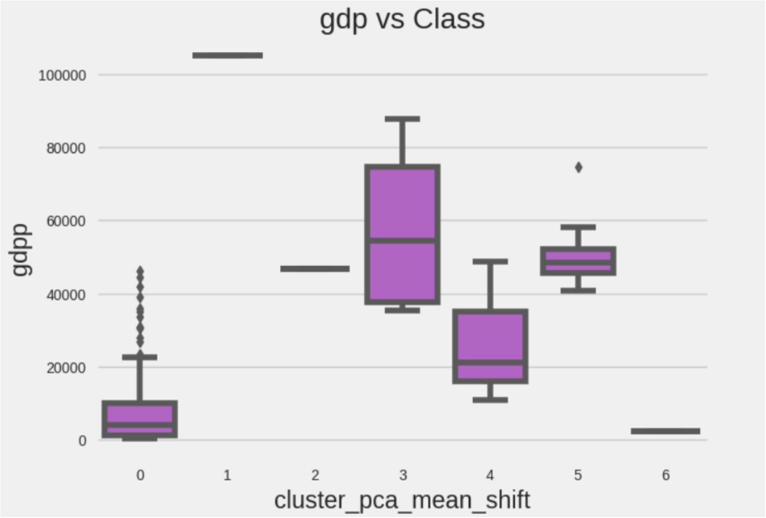
II. This algorithm shifts the circles in the direction of higher density, until each of them has found a local density maximum. Finally, all the instances whose circles have settled in the same place (or close enough) are assigned to the same cluster.

III. It has a single hyperparameter (the radius of the circles, called the bandwidth) and relies on local density estimation; however when they have internal density variations, it tends to break clusters into pieces. Its computational complexity is O(m2), thus not suited for large datasets.

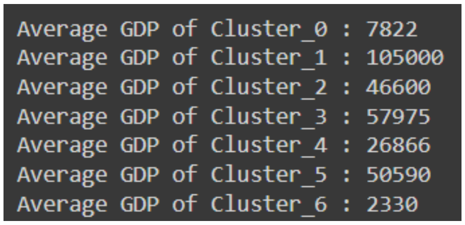
A. On PCA



```
   count  percent
0    146    87.43
5     11     6.59
3      4     2.40
4      3     1.80
1      1     0.60
6      1     0.60
2      1     0.60
```

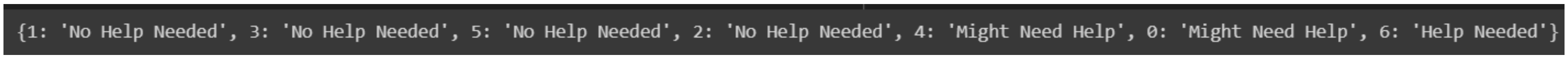


Using boxplots and average GDP of each cluster, we assigned labels to each cluster.

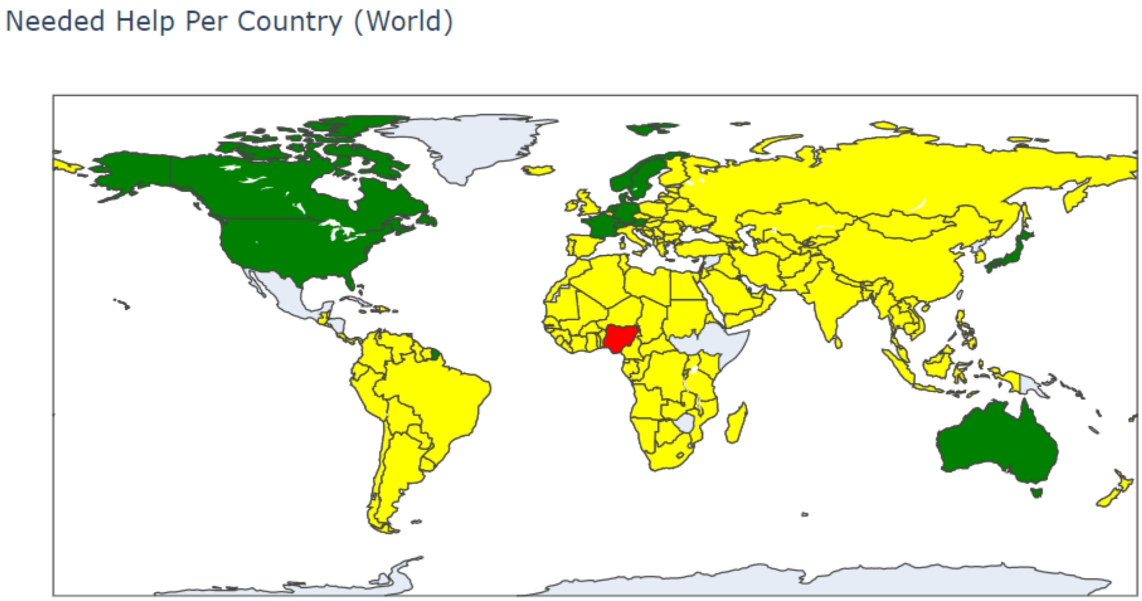


```
Average GDP of Cluster_0 : 7822
Average GDP of Cluster_1 : 105000
Average GDP of Cluster_2 : 46600
Average GDP of Cluster_3 : 57975
Average GDP of Cluster_4 : 26866
Average GDP of Cluster_5 : 50590
Average GDP of Cluster_6 : 2330
```

Clustering assigning as follows:



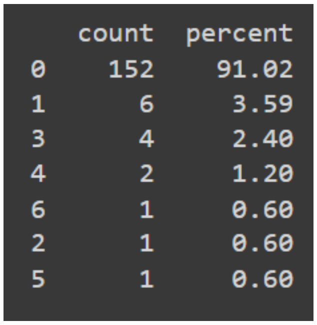
```
{1: 'No Help Needed', 3: 'No Help Needed', 5: 'No Help Needed', 2: 'No Help Needed', 4: 'Might Need Help', 0: 'Might Need Help', 6: 'Help Needed'}
```



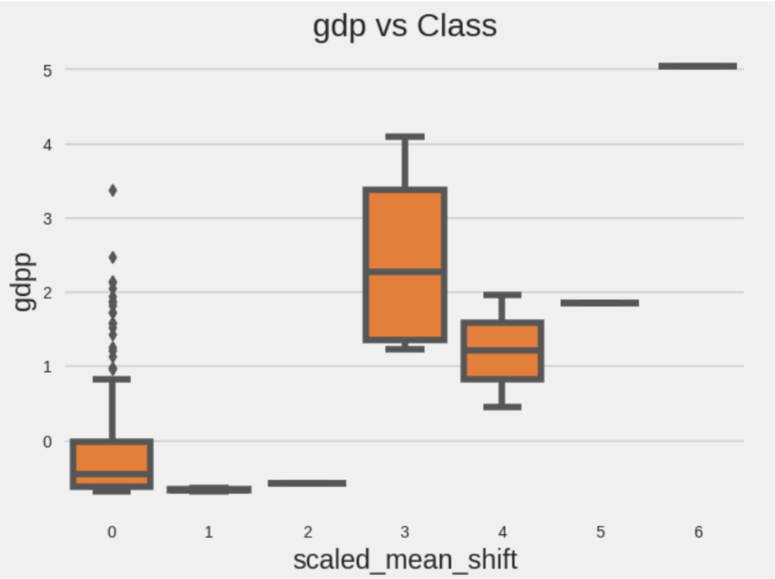
Countries that need help: (only one here)

```
['Nigeria']
```

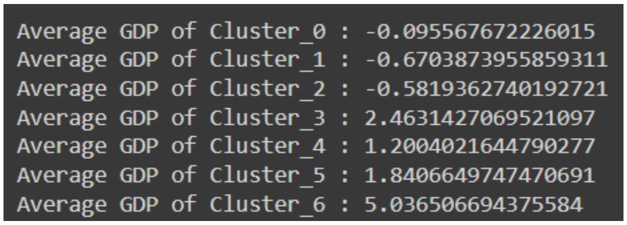
B. On scaled data



```
   count  percent
0    152    91.02
1      6     3.59
3      4     2.40
4      2     1.20
6      1     0.60
2      1     0.60
5      1     0.60
```

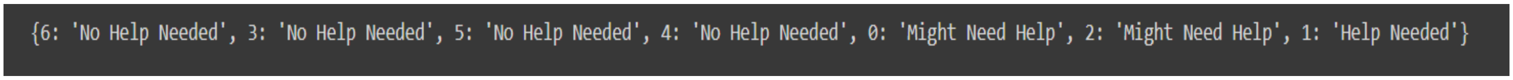


Using boxplots and average GDP of each cluster, we assigned labels to each cluster.

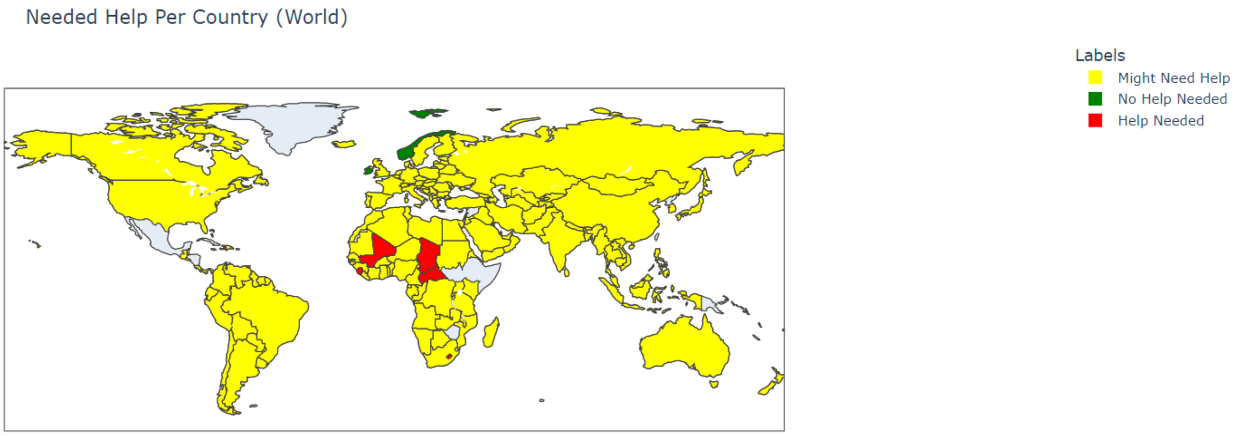


```
Average GDP of Cluster_0 : -0.095567672226015
Average GDP of Cluster_1 : -0.6703873955859311
Average GDP of Cluster_2 : -0.5819362740192721
Average GDP of Cluster_3 : 2.4631427069521097
Average GDP of Cluster_4 : 1.2004021644790277
Average GDP of Cluster_5 : 1.8406649747470691
Average GDP of Cluster_6 : 5.036506694375584
```

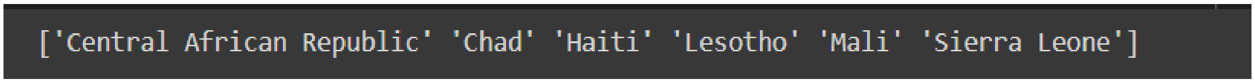
Cluster assigning:



```
{6: 'No Help Needed', 3: 'No Help Needed', 5: 'No Help Needed', 4: 'No Help Needed', 0: 'Might Need Help', 2: 'Might Need Help', 1: 'Help Needed'}
```



Countries that require help:

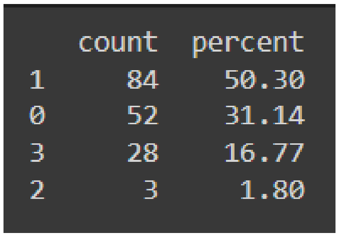


```
['Central African Republic' 'Chad' 'Haiti' 'Lesotho' 'Mali' 'Sierra Leone']
```

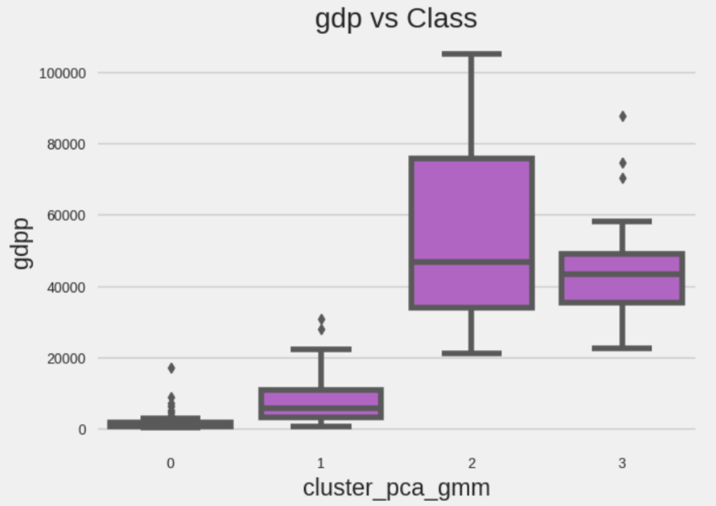
6. Gaussian Mixture Model clustering

It is a probabilistic model that assumes that the instances were generated from a mixture of several Gaussian distributions whose parameters are unknown. All the instances generated from a single Gaussian distribution form a cluster that usually looks like an ellipsoid. Each cluster can have a different ellipsoidal shape, size, density and orientation.

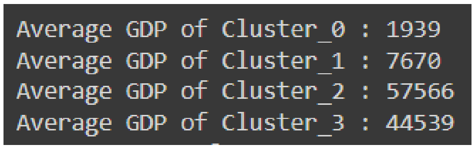
A. On PCA



```
   count  percent
1     84    50.30
0     52    31.14
3     28    16.77
2      3     1.80
```

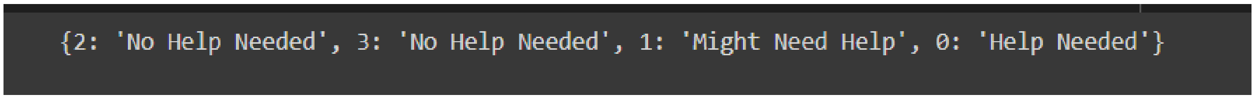


Using boxplots and average GDP of each cluster, we assigned labels to each cluster.

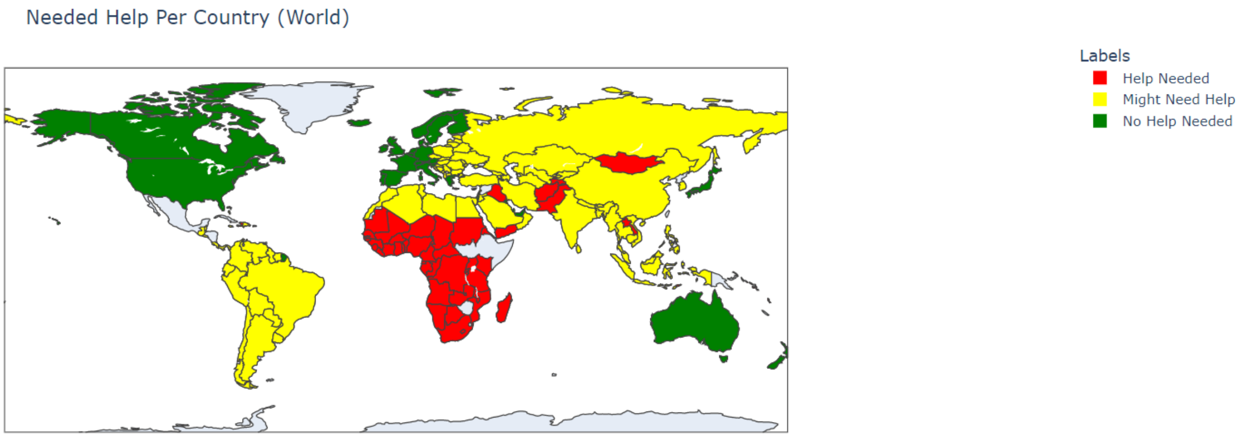


```
Average GDP of Cluster_0 : 1939
Average GDP of Cluster_1 : 7670
Average GDP of Cluster_2 : 57566
Average GDP of Cluster_3 : 44539
```

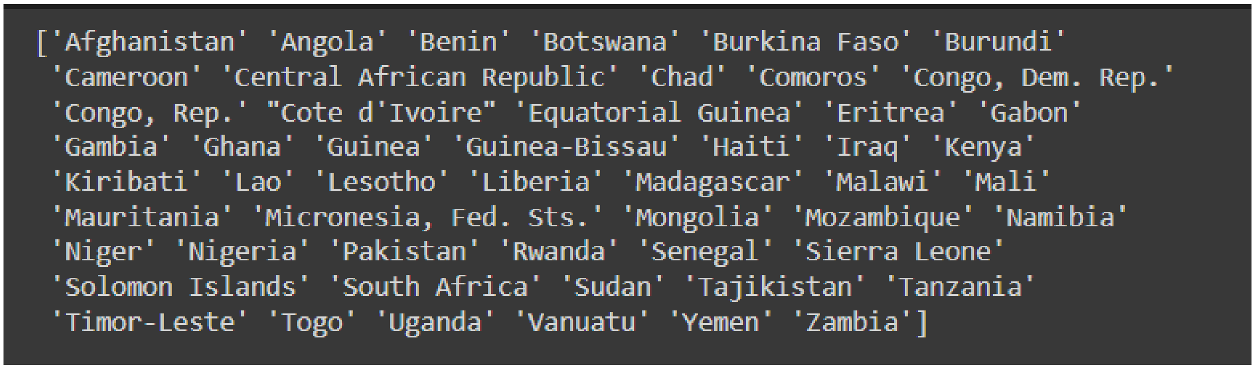
Clustering assignment:



```
{2: 'No Help Needed', 3: 'No Help Needed', 1: 'Might Need Help', 0: 'Help Needed'}
```

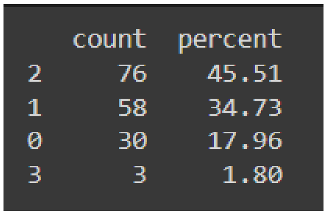


Countries that need help:

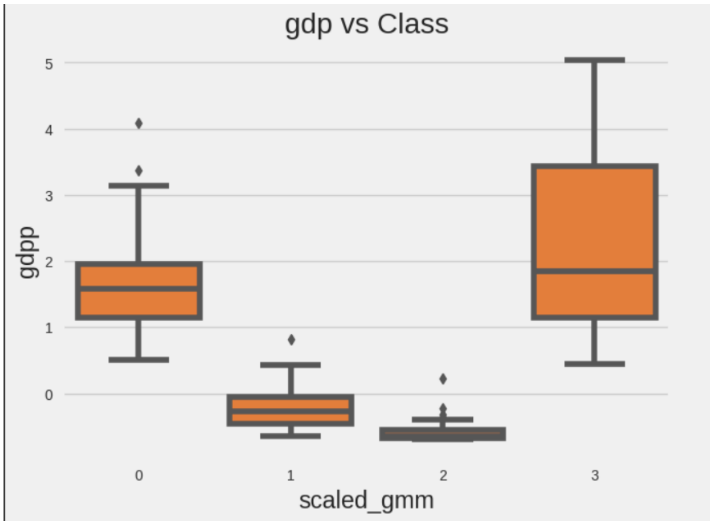


```
['Afghanistan' 'Angola' 'Benin' 'Botswana' 'Burkina Faso' 'Burundi'
 'Cameroon' 'Central African Republic' 'Chad' 'Comoros' 'Congo, Dem. Rep.'
 'Congo, Rep.' "Cote d'Ivoire" 'Equatorial Guinea' 'Eritrea' 'Gabon'
 'Gambia' 'Ghana' 'Guinea' 'Guinea-Bissau' 'Haiti' 'Iraq' 'Kenya'
 'Kiribati' 'Lao' 'Lesotho'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

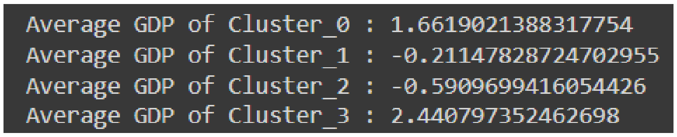
B. On scaled data



```
   count  percent
2     76    45.51
1     58    34.73
0     30    17.96
3      3     1.80
```

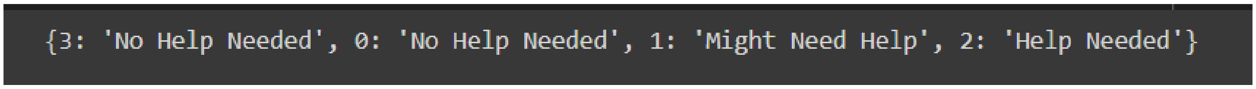


Using boxplots and average GDP of each cluster, we assigned labels to each cluster.

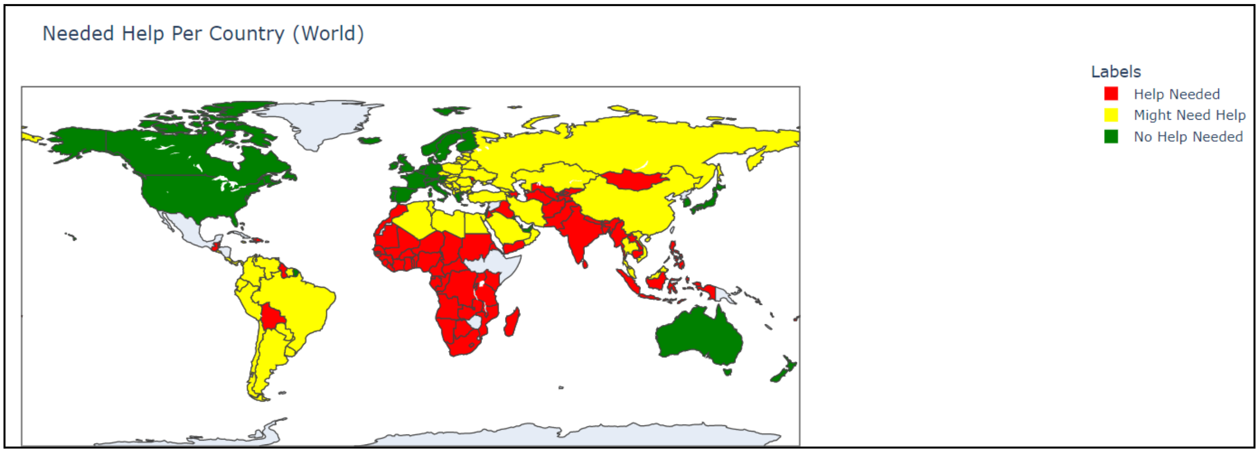


```
Average GDP of Cluster_0 : 1.6619021388317754
Average GDP of Cluster_1 : -0.21147828724702955
Average GDP of Cluster_2 : -0.5909699416054426
Average GDP of Cluster_3 : 2.440797352462698
```

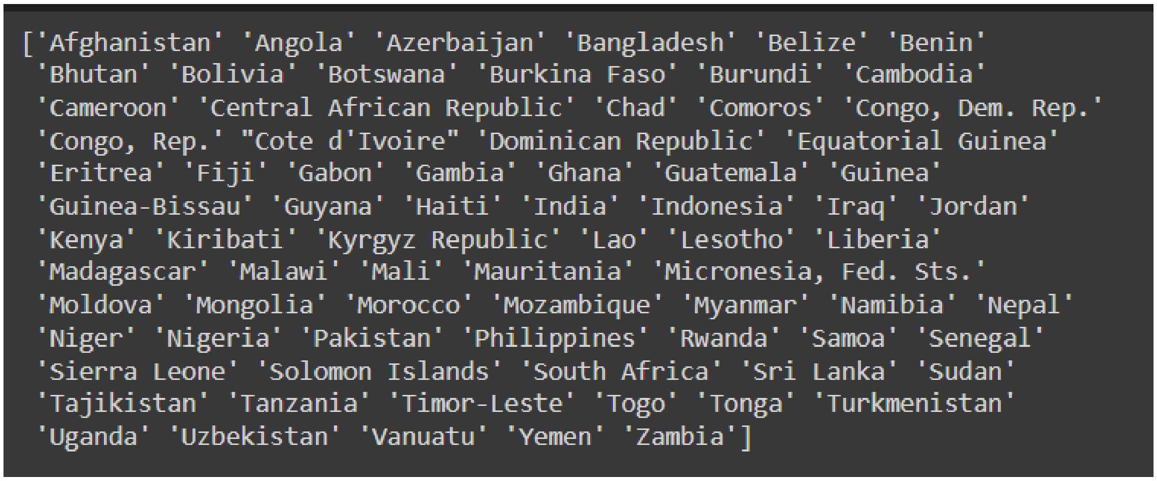
Clustering assignment:



```
{3: 'No Help Needed', 0: 'No Help Needed', 1: 'Might Need Help', 2: 'Help Needed'}
```



Countries that need help:



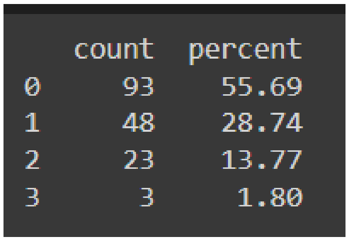
```
['Afghanistan' 'Angola' 'Azerbaijan' 'Bangladesh' 'Belize' 'Benin'
 'Bhutan' 'Bolivia' 'Botswana' 'Burkina Faso' 'Burundi' 'Cambodia'
 'Cameroon' 'Central African Republic' 'Chad' 'Comoros' 'Congo, Dem. Rep.'
 'Congo, Rep.' "Cote d'Ivoire" 'Dominican Republic' 'Equatorial Guinea'
 'Eritrea' 'Fiji' 'Gabon' 'Gambia' 'Ghana' 'Guatemala' 'Guinea'
 'Guinea-Bissau' 'Guyana' 'Haiti' 'India' 'Indonesia' 'Iraq' 'Jordan'
 'Kenya' 'Kiribati' 'Kyrgyz Republic' 'Lao' 'Lesotho' 'Liberia'
 'Madagascar' 'Malawi' 'Mali' 'Mauritania' 'Micronesia, Fed. Sts.'
 'Moldova' 'Mongolia' 'Morocco' 'Mozambique' 'Myanmar' 'Namibia' 'Nepal'
 'Niger' 'Nigeria' 'Pakistan' 'Philippines' 'Rwanda' 'Samoa' 'Senegal'
 'Sierra Leone' 'Solomon Islands' 'South Africa' 'Sri Lanka' 'Sudan'
 'Tajikistan' 'Tanzania' 'Timor-Leste' 'Togo' 'Tonga' 'Turkmenistan'
 'Uganda' 'Uzbekistan' 'Vanuatu' 'Yemen' 'Zambia']
```

7. Auto-encoding using tensorflow

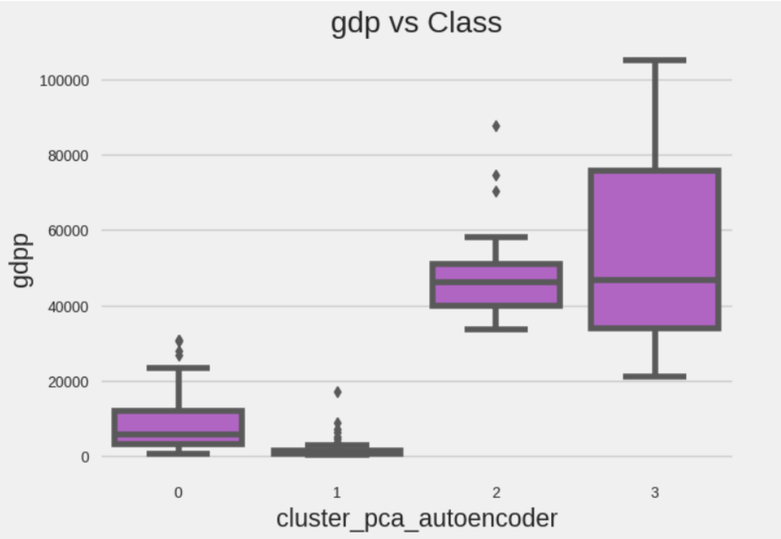
I. Autoencoders are a type of neural network that are used for unsupervised learning. They are commonly used for tasks such as image and speech recognition, data compression, and feature extraction.

II. An autoencoder consists of an encoder and a decoder, which are typically symmetrical. The encoder reduces the input data to a lower-dimensional representation, and the decoder attempts to reconstruct the original data from this representation.

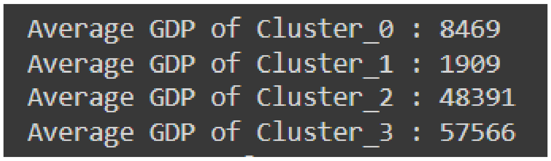
A. On PCA



```
   count  percent
0     93    55.69
1     48    28.74
2     23    13.77
3      3     1.80
```

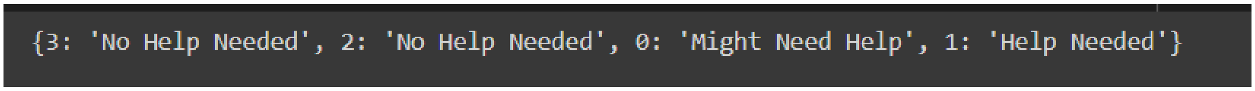


Using boxplots and average GDP of each cluster, we assigned labels to each cluster.

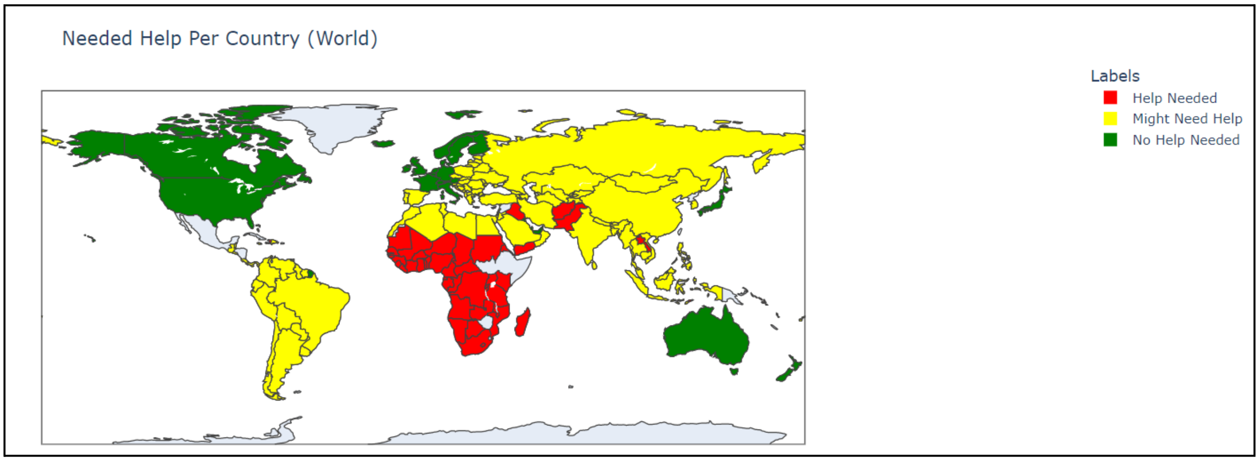


```
Average GDP of Cluster_0 : 8469
Average GDP of Cluster_1 : 1909
Average GDP of Cluster_2 : 48391
Average GDP of Cluster_3 : 57566
```

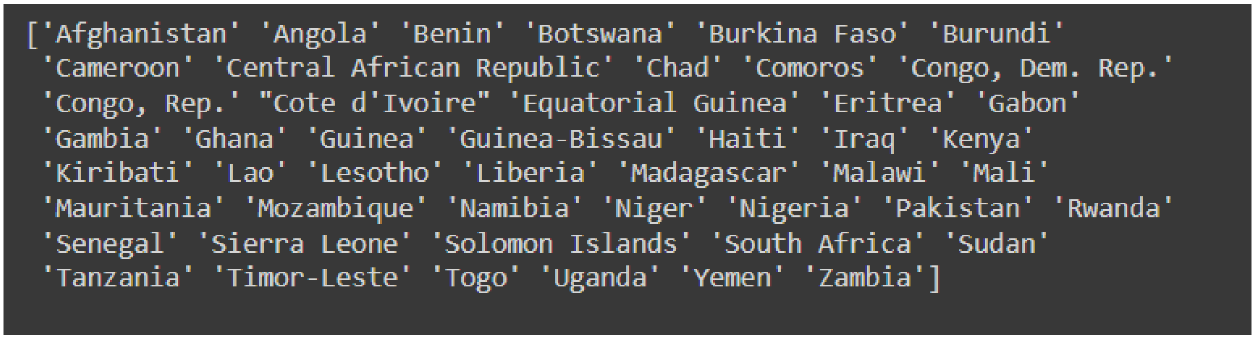
Clustering assigning as follows:



```
{3: 'No Help Needed', 2: 'No Help Needed', 0: 'Might Need Help', 1: 'Help Needed'}
```

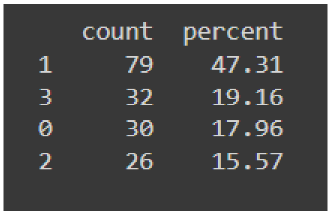


Countries that need help:

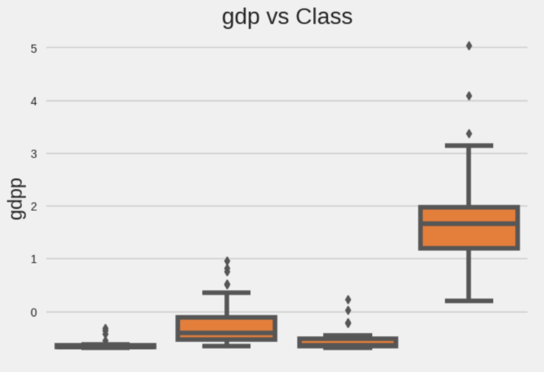


```
['Afghanistan' 'Angola' 'Benin' 'Botswana' 'Burkina Faso' 'Burundi'
 'Cameroon' 'Central African Republic' 'Chad' 'Comoros' 'Congo, Dem. Rep.'
 'Congo, Rep.' "Cote d'Ivoire" 'Equatorial Guinea' 'Eritrea' 'Gabon'
 'Gambia' 'Ghana' 'Guinea' 'Guinea-Bissau' 'Haiti' 'Iraq' 'Kenya'
 'Kiribati' 'Lao' 'Lesotho'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

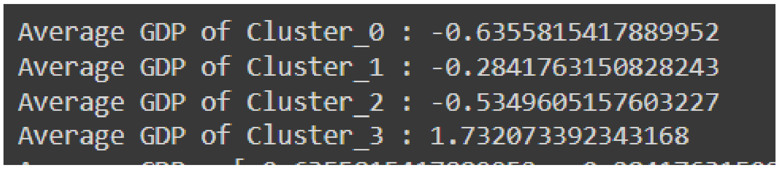
B. On scaled data



```
   count  percent
1     79    47.31
3     32    19.16
0     30    17.96
2     26    15.57
```

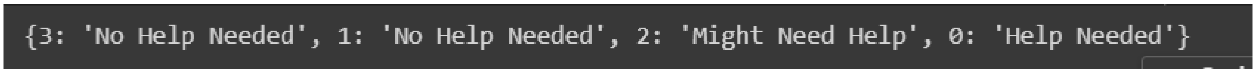


Using boxplots and average GDP of each cluster, we assigned labels to each cluster.

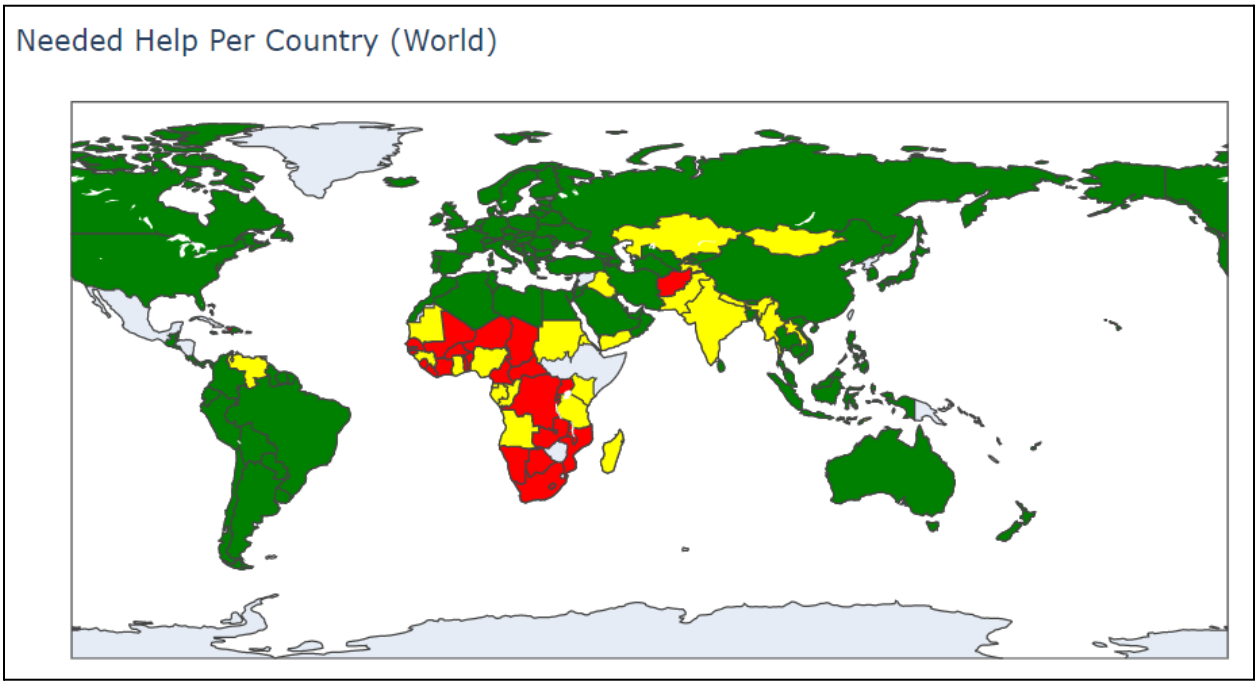


```
Average GDP of Cluster_0 : -0.6355815417889952
Average GDP of Cluster_1 : -0.2841763150828243
Average GDP of Cluster_2 : -0.5349605157603227
Average GDP of Cluster_3 : 1.732073392343168
```

Clustering assigning as follows:

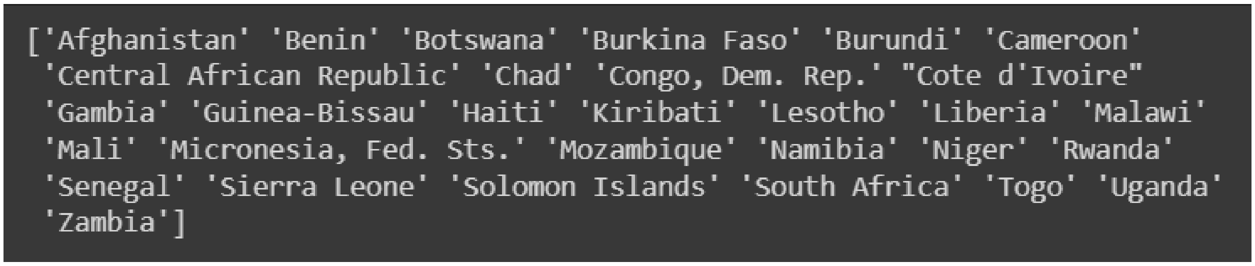


```
{3: 'No Help Needed', 1: 'No Help Needed', 2: 'Might Need Help', 0: 'Help Needed'}
```



Needed Help Per Country (World)

Following countries are in need:

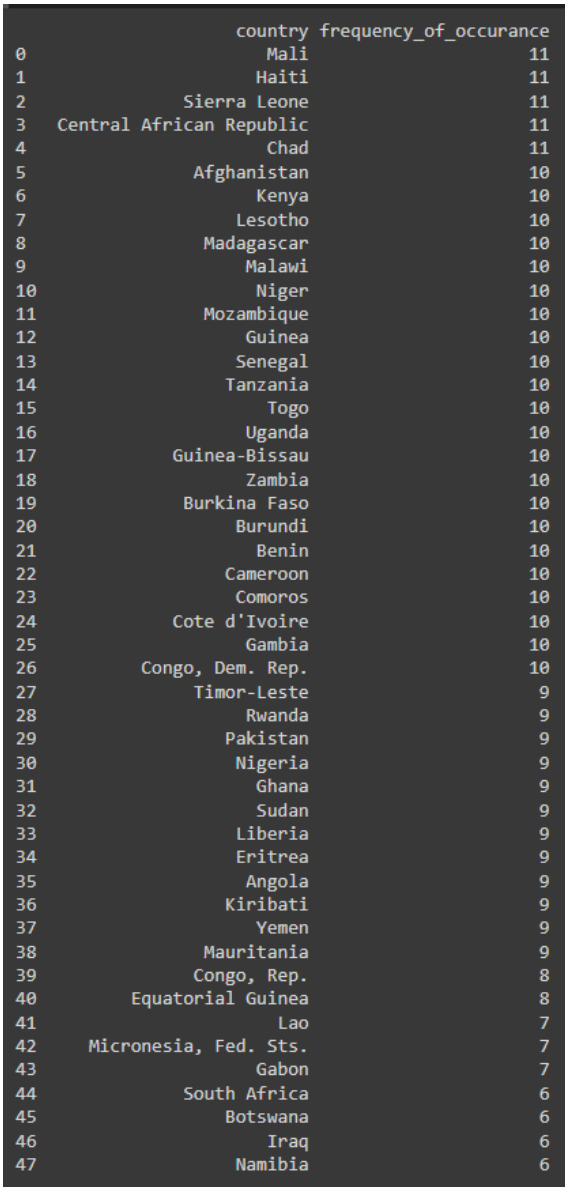


```
['Afghanistan' 'Benin' 'Botswana' 'Burkina Faso' 'Burundi' 'Cameroon'
 'Central African Republic' 'Chad' 'Congo, Dem. Rep.' "Cote d'Ivoire"
 'Gambia' 'Guinea-Bissau' 'Haiti' 'Kiribati' 'Lesotho' 'Liberia' 'Malawi'
 'Mali' 'Micronesia, Fed. Sts.' 'Mozambique' 'Namibia' 'Niger' 'Rwanda'
 'Senegal' 'Sierra Leone' 'Solomon Islands' 'South Africa' 'Togo' 'Uganda'
 'Zambia']
```

CONCLUSION

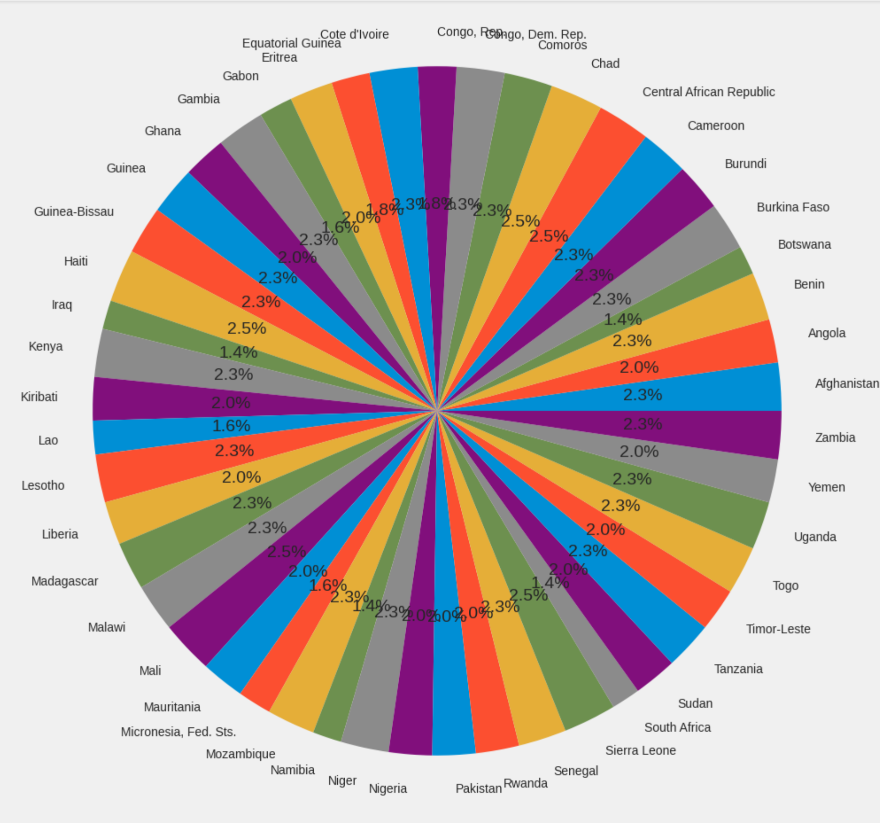
- We discarded the DBSCAN method, because it is not showing good results.
- Rest of the models are giving appropriate results that support each other.
- As it is difficult to discard other clustering methods ,We will use all other methods and maintain the frequency of occurrence of each country in the ‘Needy country’ cluster using Dictionary.
- To remove anomalies in the model results, we are using a method similar to voting.
- We will use a threshold to remove the countries due to the anomalies in the model and place the countries in a sequence so that we know how to distribute our money.

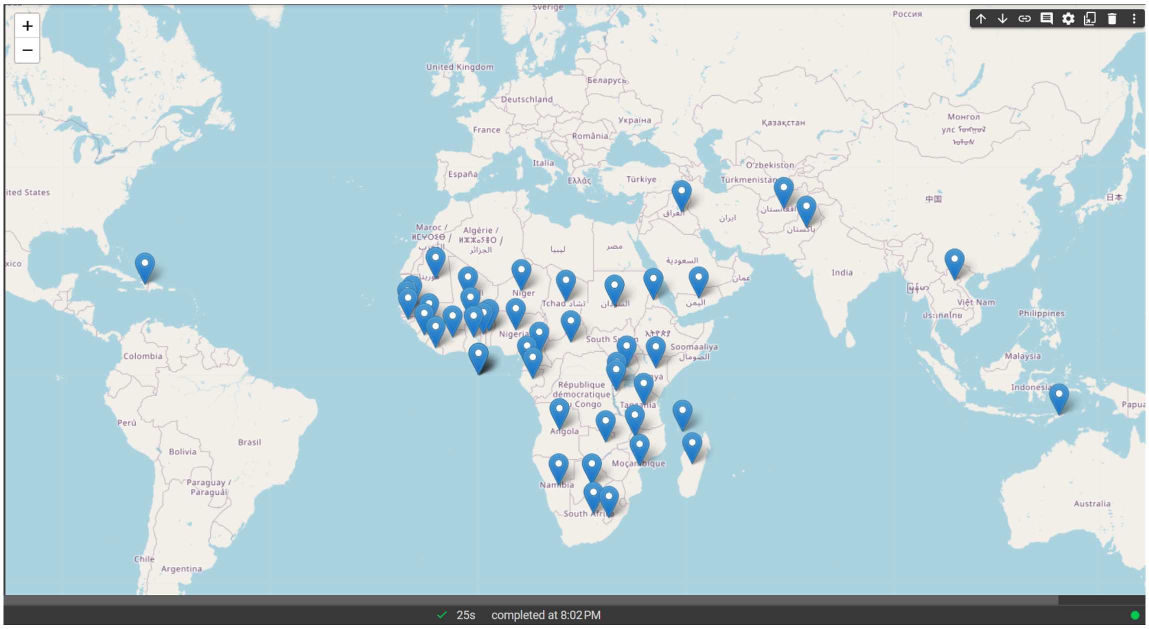
According to their decreasing order of frequency, the CEOs of NGO can invest money in the corresponding countries.



```
                     country  frequency_of_occurance
0                       Mali                      11
1                      Haiti                      11
2               Sierra Leone                      11
3   Central African Republic                      11
4                       Chad                      11
5                Afghanistan                      10
6                      Kenya                      10
7                    Lesotho                      10
8                 Madagascar                      10
9                     Malawi                      10
10                     Niger                      10
11                Mozambique                      10
12                    Guinea                      10
13                   Senegal                      10
14                  Tanzania                      10
15                      Togo                      10
16                    Uganda                      10
17             Guinea-Bissau                      10
18                    Zambia                      10
19              Burkina Faso                      10
20                   Burundi                      10
21                     Benin                      10
22                  Cameroon                      10
23                   Comoros                      10
24             Cote d'Ivoire                      10
25                    Gambia                      10
26          Congo, Dem. Rep.                      10
27               Timor-Leste                       9
28                    Rwanda                       9
29                  Pakistan                       9
30                   Nigeria                       9
31                     Ghana                       9
32                     Sudan                       9
33                   Liberia                       9
34                   Eritrea                       9
35                    Angola                       9
36                  Kiribati                       9
37                     Yemen                       9
38                Mauritania                       9
39              Congo, Rep.                        8
40         Equatorial Guinea                       8
41                       Lao                       7
42     Micronesia, Fed. Sts.                       7
43                     Gabon                       7
44              South Africa                       6
45                  Botswana                       6
46                      Iraq                       6
47                   Namibia                       6
```

Visualization for which is given in the below pie chart:





✓ 25s completed at 8:02 PM

Above plot is of the geographical location of the needy countries on the world map. We see that most of the countries are from the African continent and are thus at a higher priority.