upGrad

# CLUSTERING ASSIGNMENT

BY :Advaith Radhakrishnan

# PROBLEM STATEMENT

Identify the top countries that are in dire need of aid and healthcare. Your job is to categorise the countries using some socio-economic and health factors that determine the overall development of the country. Then you need to suggest the countries which the CEO needs to focus on the most

# STEPS INVOLVED/ ANALYSIS APPROACH

1. We start off with importing the dataset and do some changes to it, so that we can analyse the data set in a proper manner.

2. One important step here is to convert child_mort, Income and gdpp to its original values

3. Then we do Exploratory Data Analysis, in order to gather some inferences from the visualization.

4. From the EDA, we proceed to outlier analysis, where we decide to cap the outliers for some of the variables, that is, for child_mort, low values are capped and for the rest of the variables, the higher values are capped.

5. Then we check for the Hopkins statistics, which in our case is ranging between 85-90%. ( Hopkins score > 80% is considered to be good) and then scale the dataframe.

6. Now we will proceed with k-mean clustering, and we need to find out the number of clusters or the value of k.

7. This is done by the elbow curve and silhouette score, and we consider the number of clusters to be 3.

8. Then we do the clustering and cluster profiling, with plotting the various variables.

9. From the cluster profiling, we find out the countries that are in need of aid.

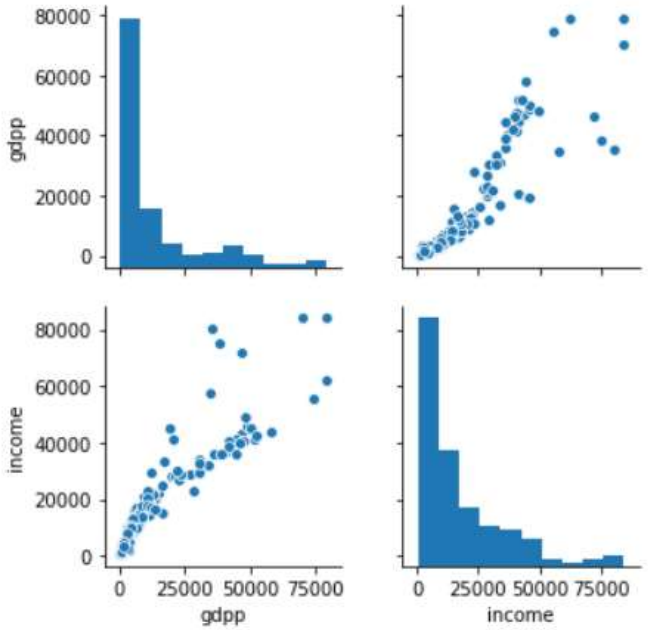
10. We also perform hierarchical clustering, where in we decide the number of clusters using single and complete linkages.

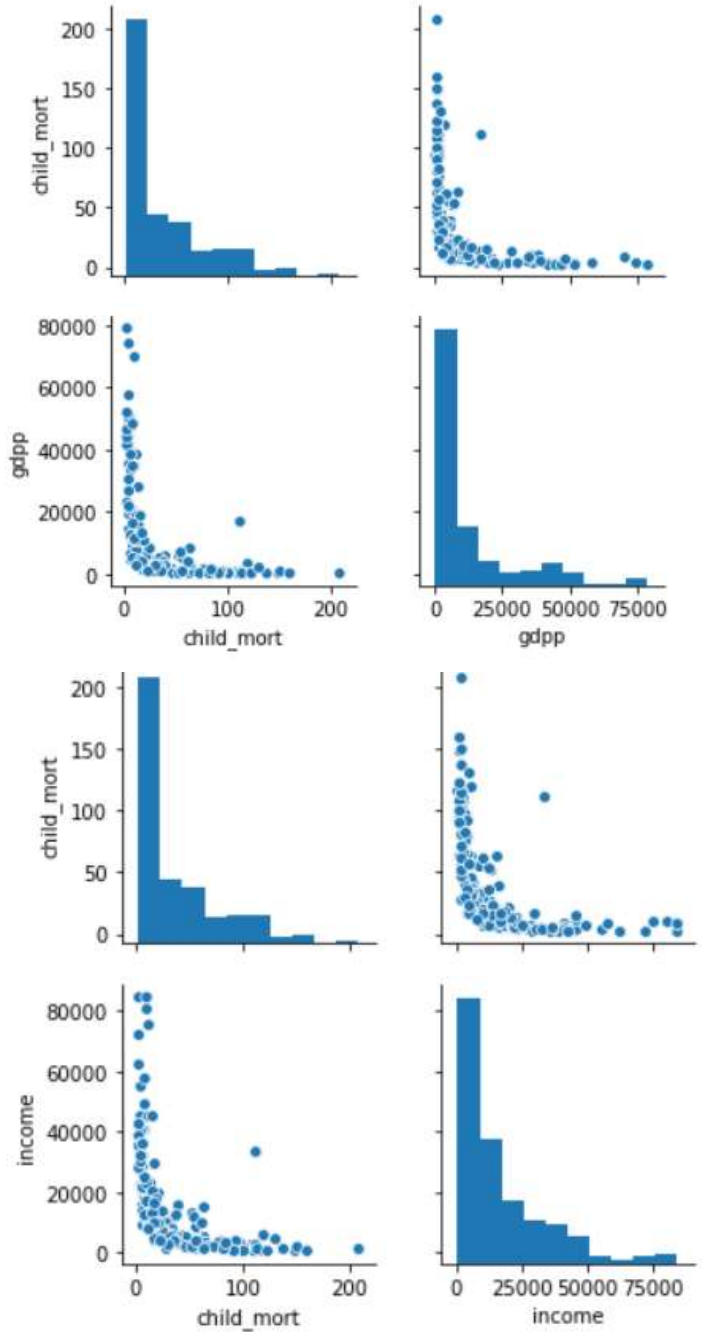
11. Finally we proceed with the clustering and find out the countries that are in need of aid, after plotting the cluster profiling.

# EXPLORATORY DATA ANALYSIS

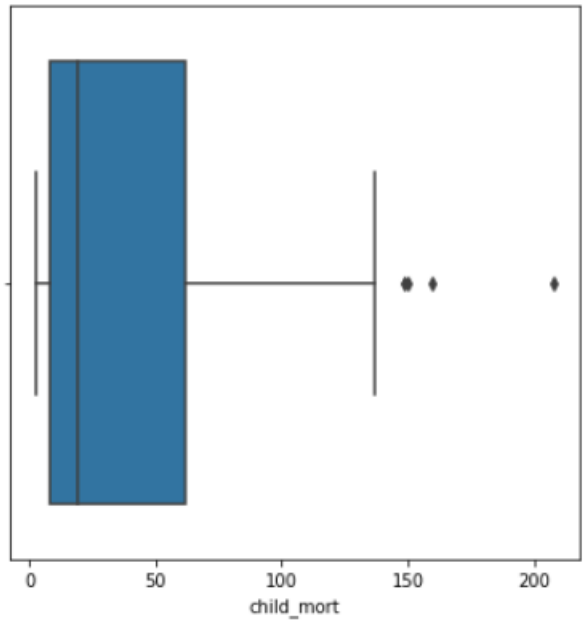
Here we plot child_mort vs gdpp, in order to see how the child_mort Variable varies with gdpp.
Similarly, we do the same pairplot with child_mort vs income and Gdpp vs income.

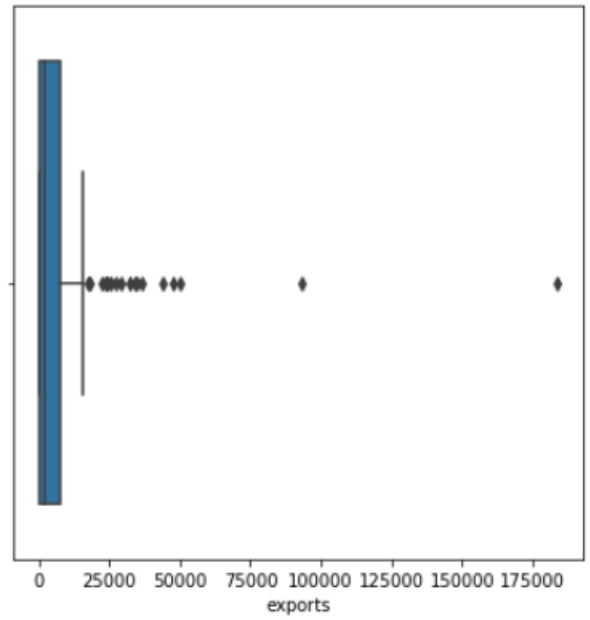
From the plots, we can see the child mortality is high for countries with low gdp and low Income.



# OUTLIER ANALYSIS



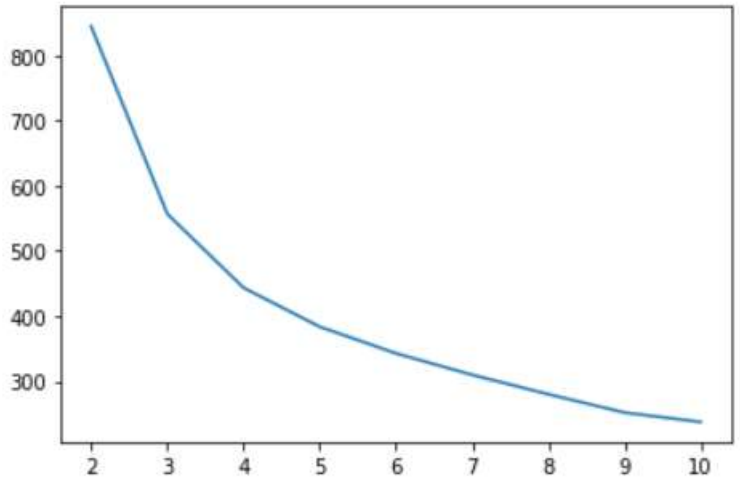

Here, we plot boxplots to fins out if there are outliers present,and we find out all the variables have outliers (example as shown in the above plots, child_mort and exports have outliers.)

Now we decide to cap these outliers present in all the variables. For child_mort variable, we cap all the lower values of the variables.

And for the rest of the variables, we cap the higher values.

This is done because these countries that have high income and gdp and low child mortality wont require any aid.

# K-MEANS CLUSTERING




Now in order to finf out the optimal number of clusters to be considered, we use 2 methods:

1. Elbow Curve

From the above plot, the elbow is formed at 3 and 4. We will select 3 out of both, as the rule being to select the smaller value.

Hence according to the elbow curve, number of clusters to be considered is 3.

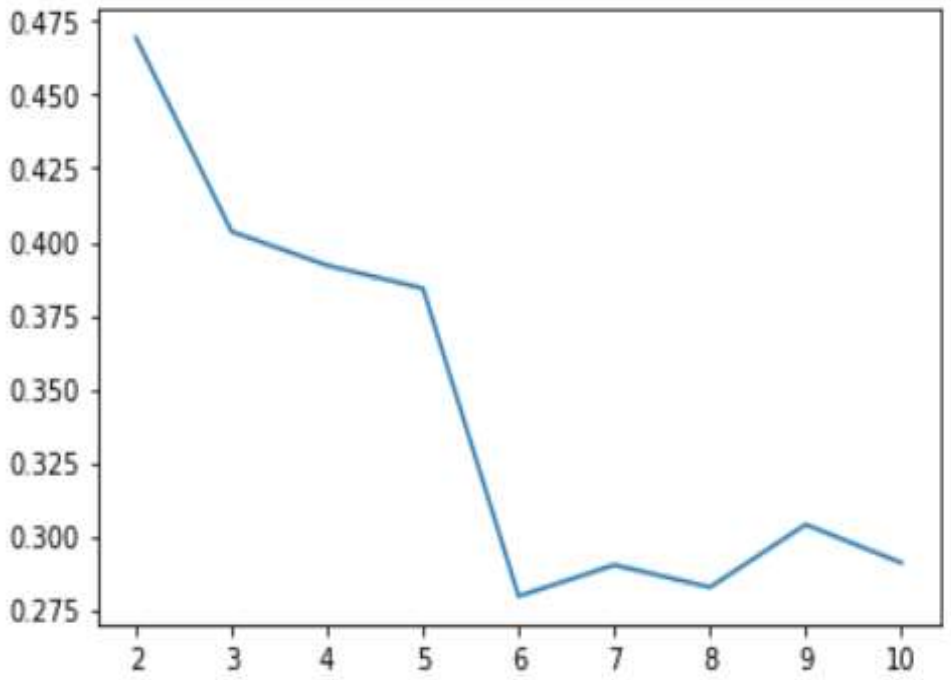
2. Silhouette score

Here, as seen in the plot, the rule is to select the cluster with the highest value, which means 2 in this case.

However, considering 2 to be the optimal number of clusters would mean, that we are basically dividing the data to 2 halves and this is not right

Hence, we will go with 3 clusters.

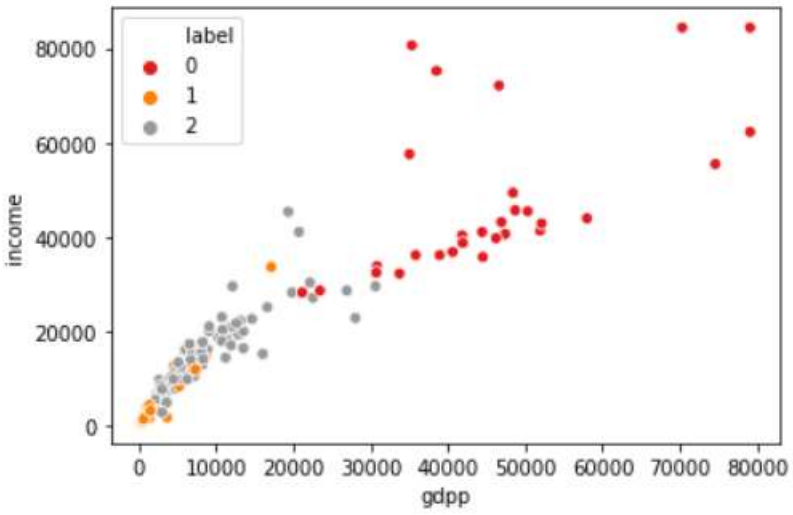
Therefore according to both the methods, optimal number of clusters is 3.

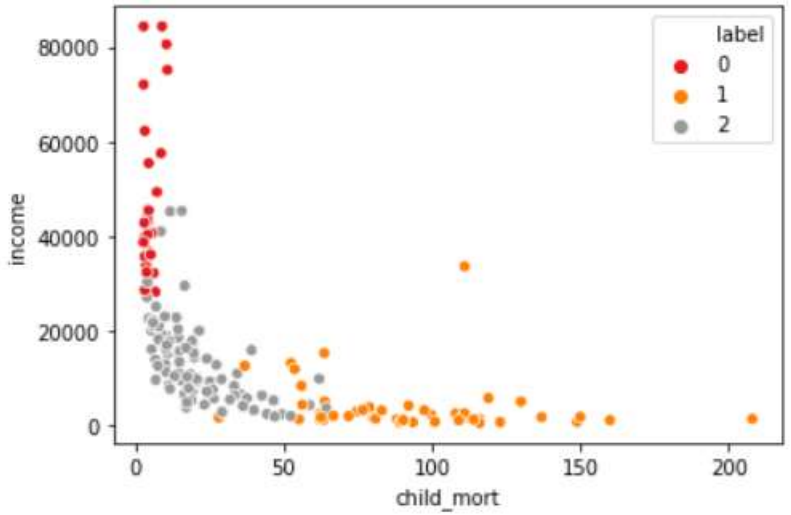


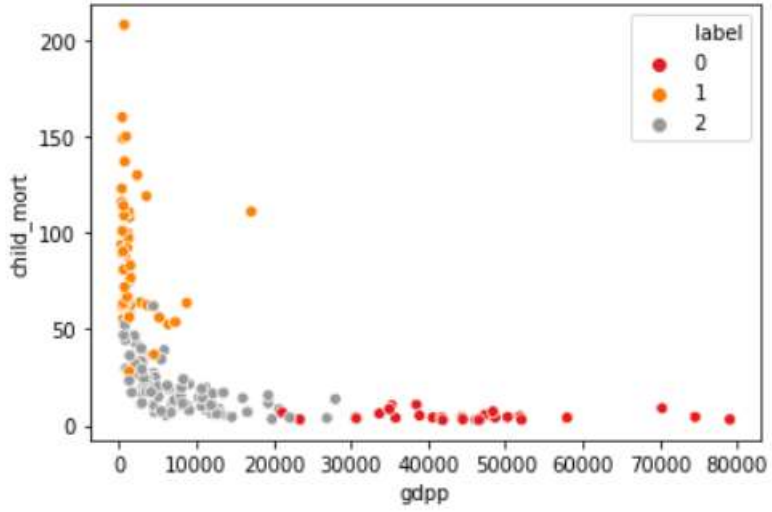
# PLOTTING THE CLUSTERS

We plot the different clusters, considering the variables child_mort, gdpp and income.

As we can see, there are 3 clusters formed.



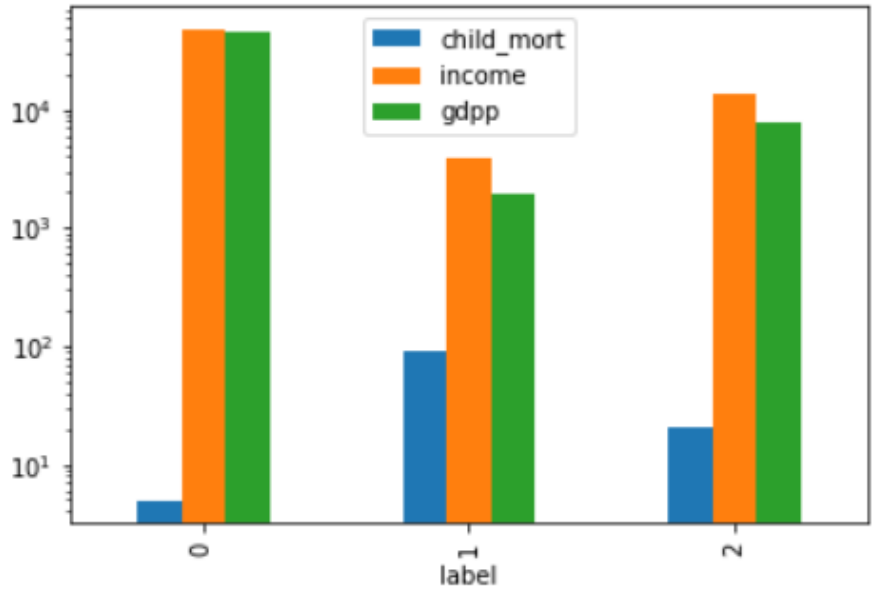



# CLUSTER PROFILING

Here we are looking at the cluster where the income and gdp is low and child mortality is high, which is label=1.

Filter the data for that cluster and we can find out the countries in need for aid.



Filtering the data for the cluster and finding out the top 5 countries in that cluster.

The countries are as follows:

| | country | child_mort | exports | health | imports | income | inflation | life_expec | total_fer | gdpp | label |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 26 | Burundi | 93.6 | 20.6052 | 26.7960 | 90.552 | 764.0 | 12.30 | 57.7 | 6.2600 | 231.0 | 1 |
| 88 | Liberia | 89.3 | 62.4570 | 38.5860 | 302.802 | 700.0 | 5.47 | 60.8 | 5.0200 | 327.0 | 1 |
| 37 | Congo, Dem. Rep. | 116.0 | 137.2740 | 26.4194 | 165.664 | 609.0 | 20.80 | 57.5 | 6.5400 | 334.0 | 1 |
| 112 | Niger | 123.0 | 77.2560 | 17.9568 | 170.868 | 814.0 | 2.55 | 58.8 | 6.5636 | 348.0 | 1 |
| 132 | Sierra Leone | 160.0 | 67.0320 | 52.2690 | 137.655 | 1220.0 | 17.20 | 55.0 | 5.2000 | 399.0 | 1 |

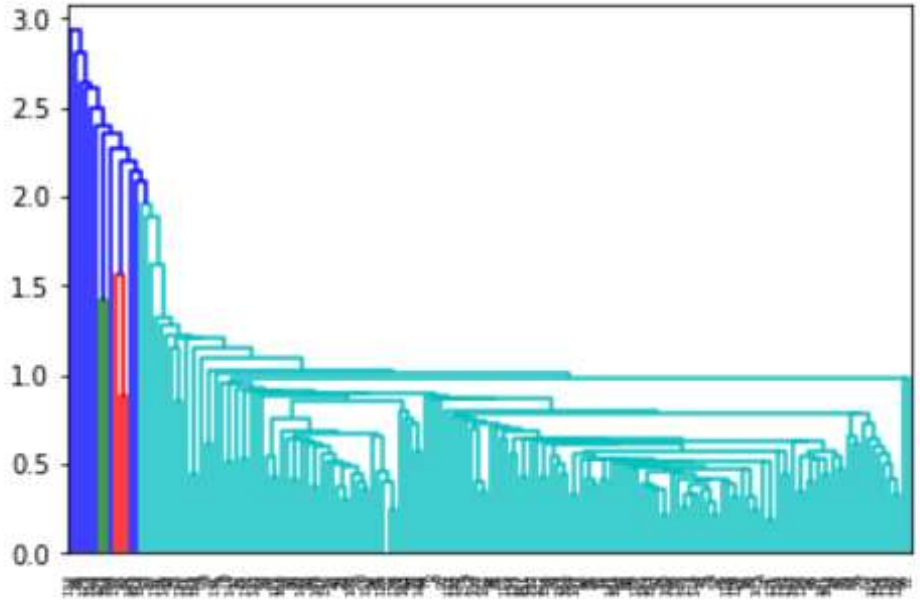
# HIERARCHICAL CLUSTERING

Firstly, we will look at the plots of both single linkage and complete linkage, in order to find out the number of clusters to be chosen.

1.Single Linkage

As we can see, the plot is not at all clear.

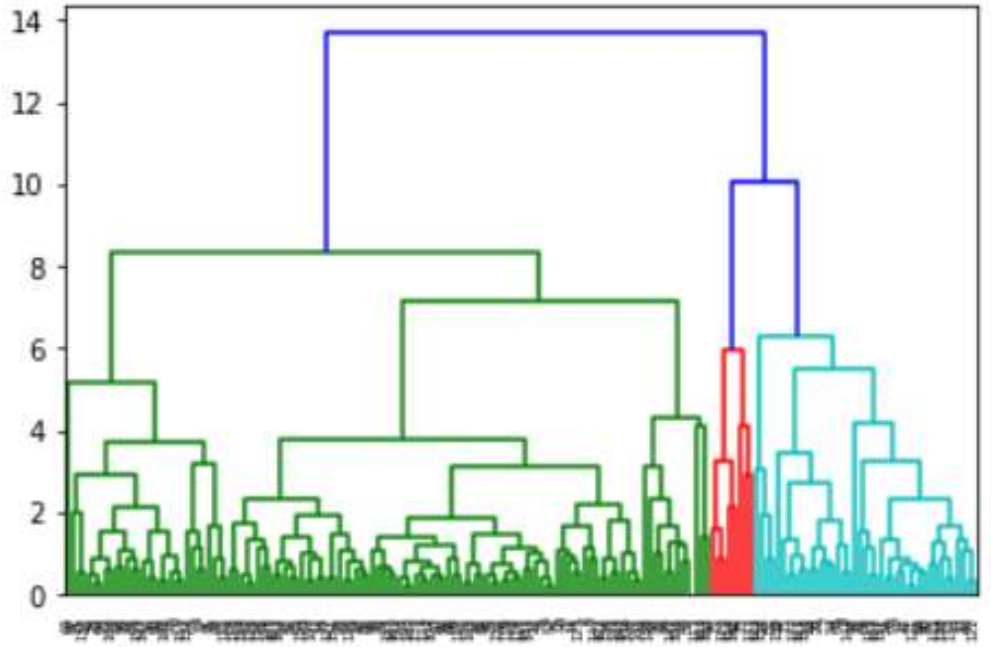
Hence we proceed with complete linkage



## 2.Complete Linkage

Since we get a clearer picture from complete linkage, we can draw a horizontal line at 10 and consider the number of clusters to be 3.

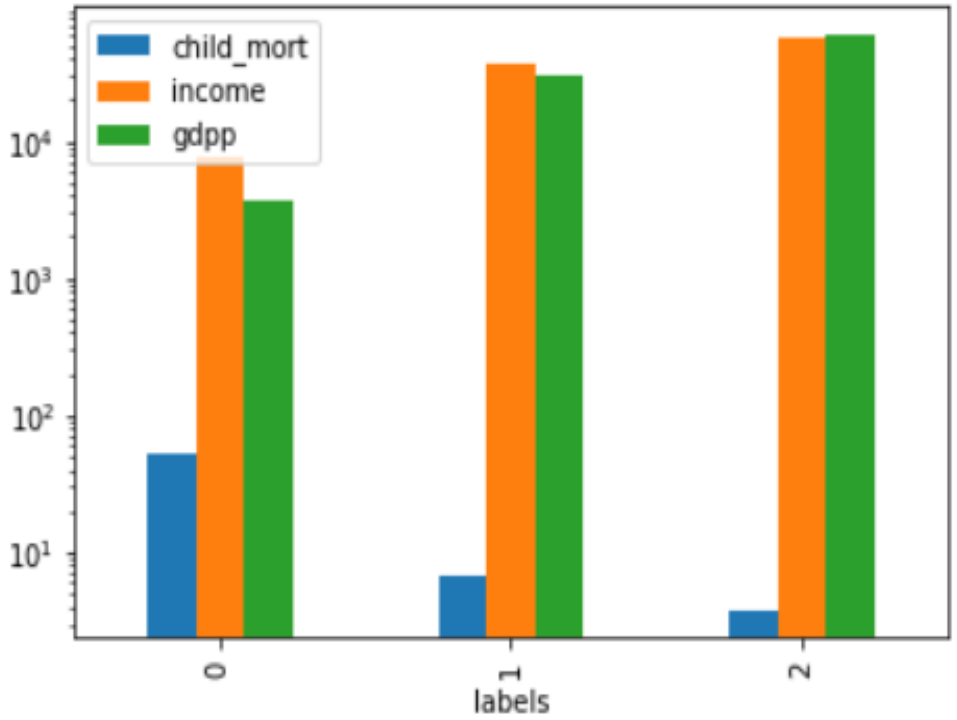
Hence we take the number of clusters to be 3.



# CLUSTER PROFILING

Now from the plot, we can say that the cluster 1 had high child_mort along with low income and GDP.

Hence we consider this cluster to find out the countries in need of aid.



Filtering the data for the cluster and finding out the top 5 countries in that cluster.

The countries are as follows:

| | country | child_mort | exports | health | imports | income | inflation | life_expec | total_fer | gdpp | labels |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 26 | Burundi | 93.6 | 20.6052 | 26.7960 | 90.552 | 764.0 | 12.30 | 57.7 | 6.2600 | 231.0 | 0 |
| 88 | Liberia | 89.3 | 62.4570 | 38.5860 | 302.802 | 700.0 | 5.47 | 60.8 | 5.0200 | 327.0 | 0 |
| 37 | Congo, Dem. Rep. | 116.0 | 137.2740 | 26.4194 | 165.664 | 609.0 | 20.80 | 57.5 | 6.5400 | 334.0 | 0 |
| 112 | Niger | 123.0 | 77.2560 | 17.9568 | 170.868 | 814.0 | 2.55 | 58.8 | 6.5636 | 348.0 | 0 |
| 132 | Sierra Leone | 160.0 | 67.0320 | 52.2690 | 137.655 | 1220.0 | 17.20 | 55.0 | 5.2000 | 399.0 | 0 |

# CONCLUSION

-In all of the above mentioned countries, we can see the following inferences:

1.Child mortality per 100 births is between 90 and 160, in the mentioned countries. In comparison to other countries like singapore with 2.8 per 1000 and US with 5.8 per 1000, shows that the mentioned 5 countries have a high child mortality

2. Similarly comparing the net income per person, it is between 750-1200, which is again very low when compared to other countries, whose income lies between 20000 to 75000

3. Looking at the GDP of the above mentioned countries, = between 200 and 400Comparing with other countries, GDP is between 12000 and 80000, which again shows the mentioned countries have very less GDP

4.We can also look at life expentency, which is between 55 and 60 .Other countries life expentancy ranges anywhere between 75 and 80

# RECOMMENDATION

Therefore, the top 5 countries which the CEO needs to focus on the most and provide the necessary aid and healthcare should be:

1. Burundi
2. Liberia
3. Congo Dem. Rep.
4. Niger
5. Sierra Leone