CLUSTERING ASSIGNMENT

By

Arun dutt .R

Problem statement

- To Identify 5 top countries that are directly in need of aid.
- For identifying the countries we need to categories the countries using child-mortality ,Income and Gross domestic product per capita (gdpp) . Using these factors to determine the overall development of the country.
- Then we are suggesting these countries to CEO for which more focus is required.

ANALYSIS APPROACH

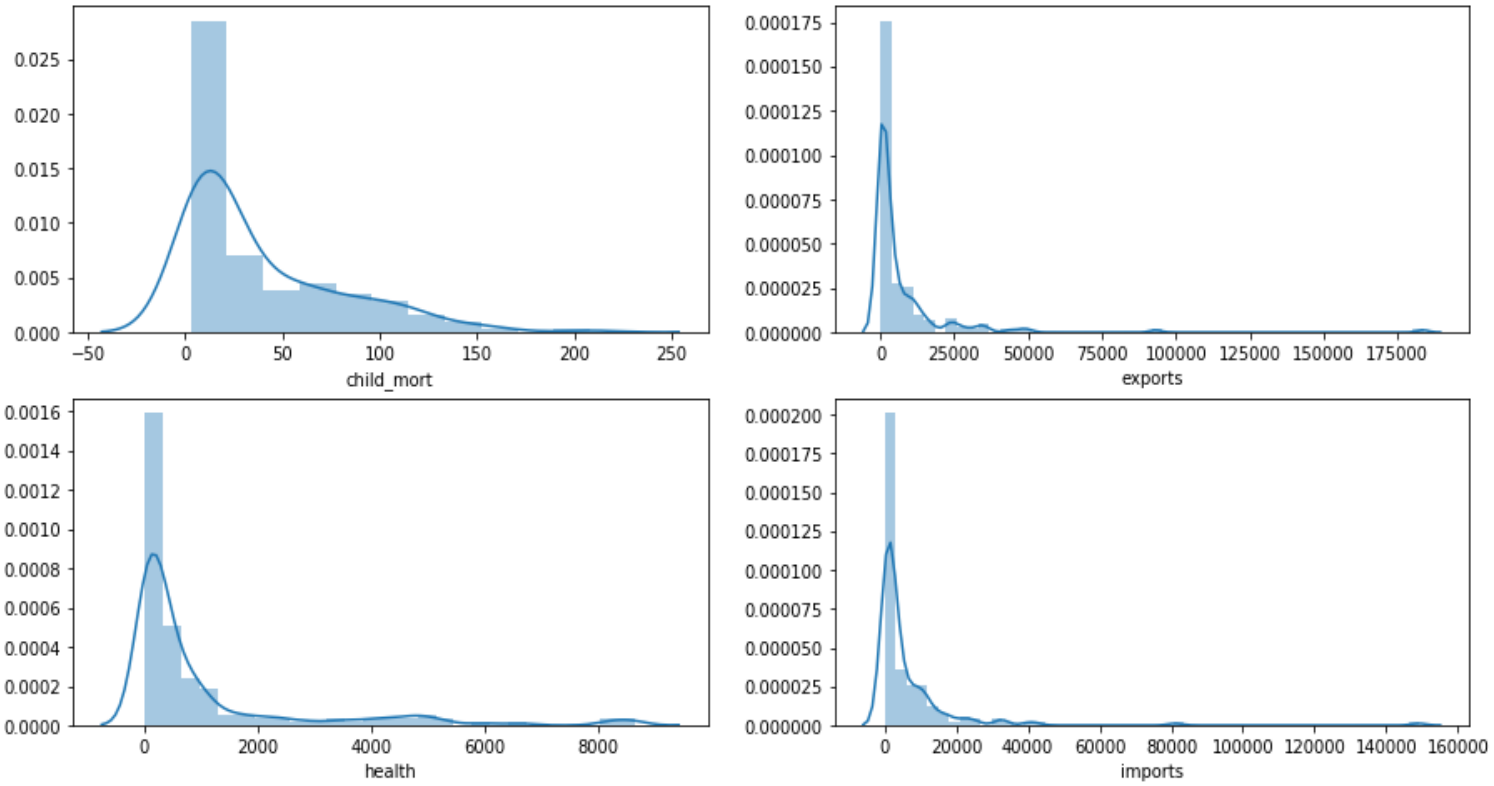
To achieve our problem statement i.e. to find the top 5 counters which are in need of aid , we did the following steps ;

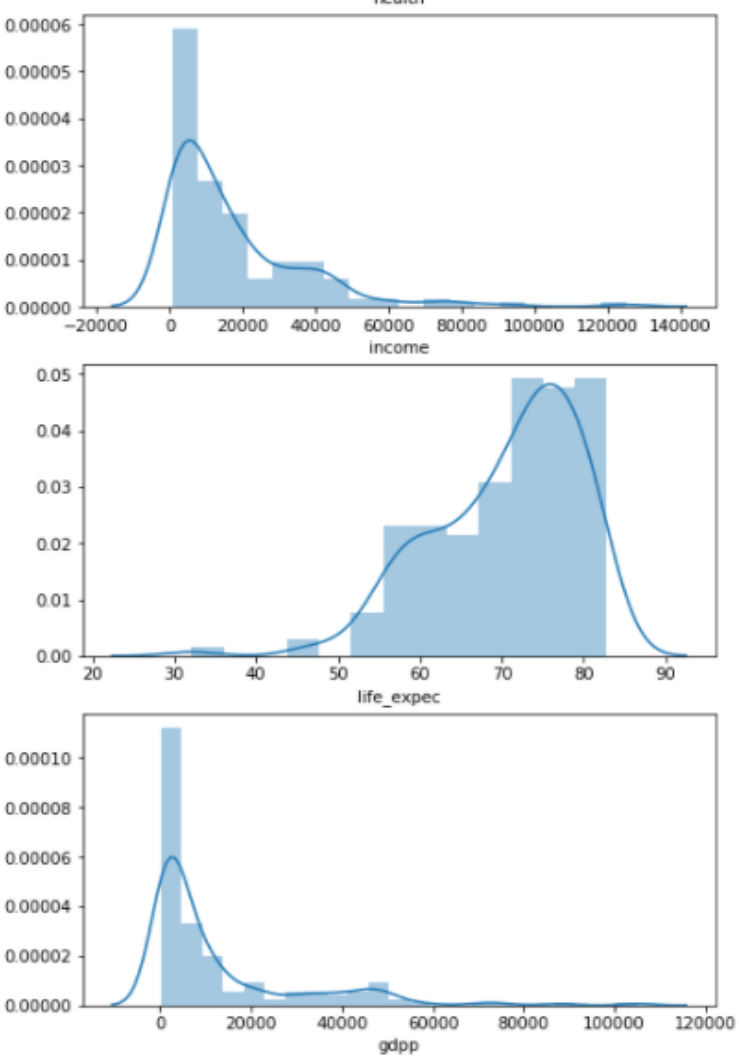
- Data analysis (EDA)
- Outlier treatment
- Calculating the Hopkins statistic & Scaling the data
- Clustering the data (K-mean and Hierarchical)
- Visualise the clusters
- Clustering profiling using “gdpp, child_mort and income”
- Identifying the top 5 countries

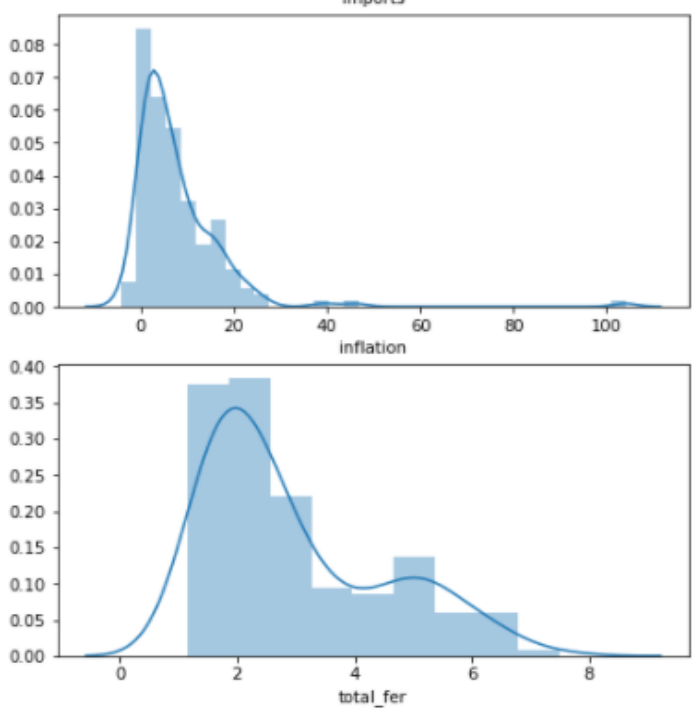
Data analysis (EDA)

- Checking for null values (i.e. no null values found)

- Converting columns (i.e'exports' ,'health', 'imports') which are in percentage of GDP per capita into actual values.



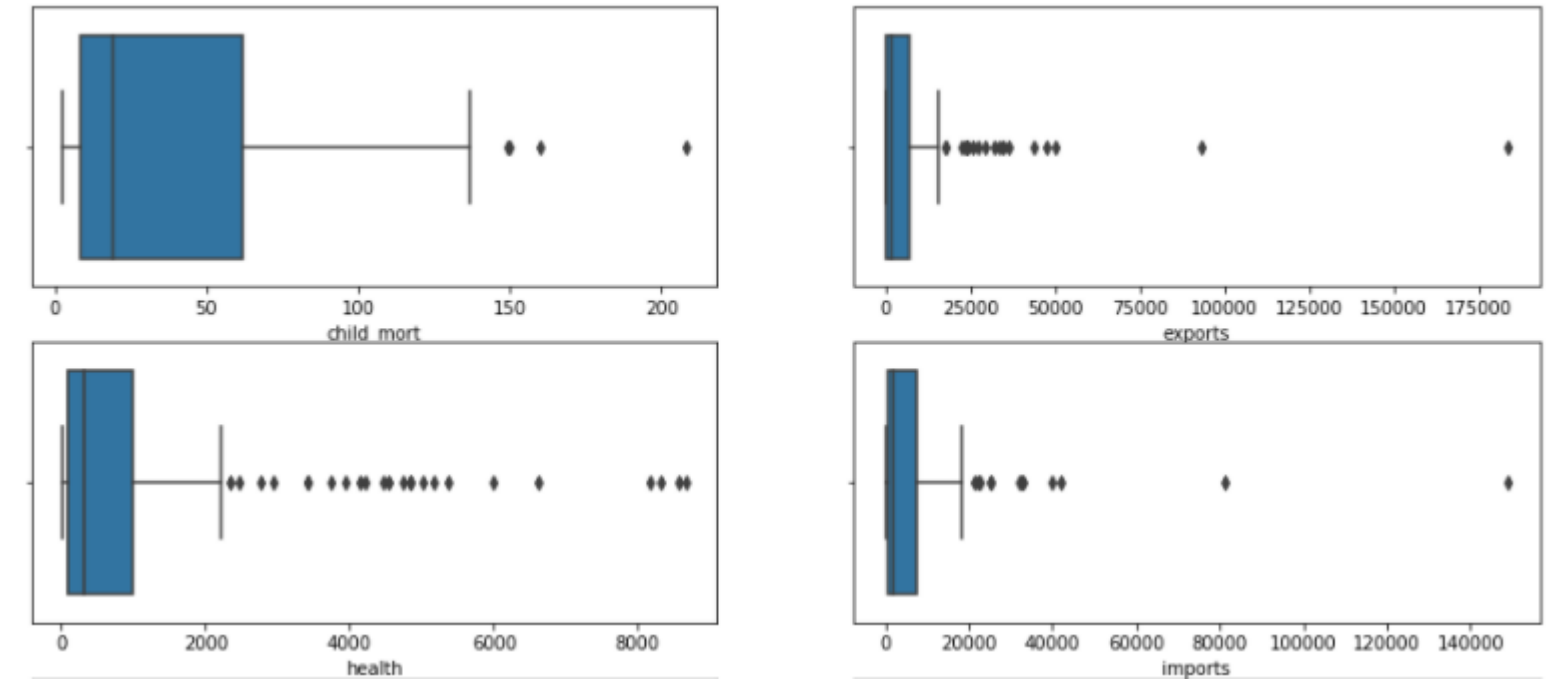


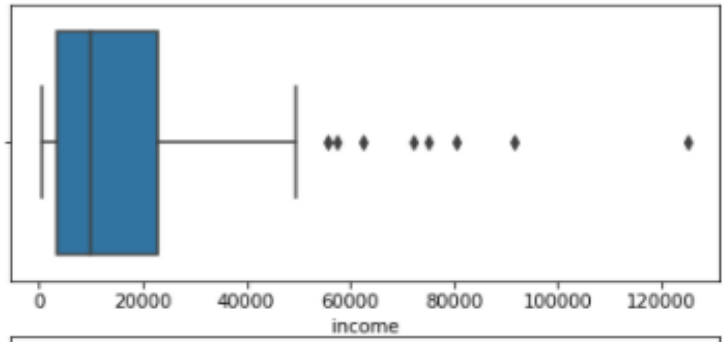


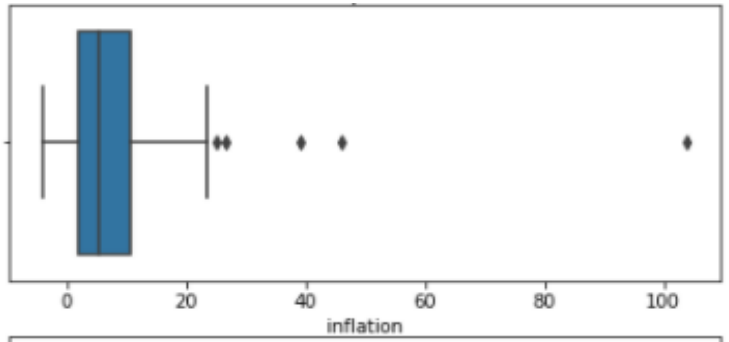
After analyzing this above plots we come to conclusion that we will use Child mortality ,Income and Gdpp for profiling the clustering . As it contains relevant data pattern to achieving our top 5 countries which desperately required aid

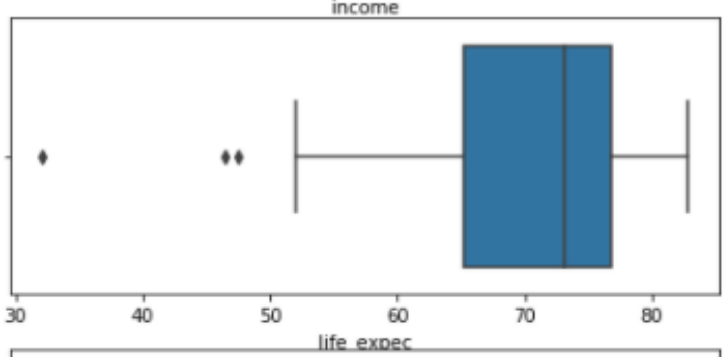
Outlier treatment

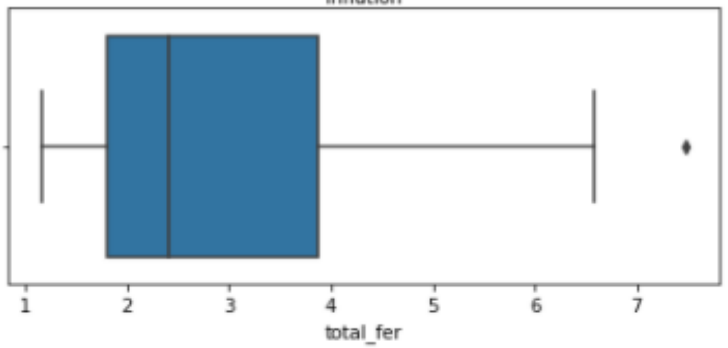
- Plotting boxplot graph to find outliers and Capping them.

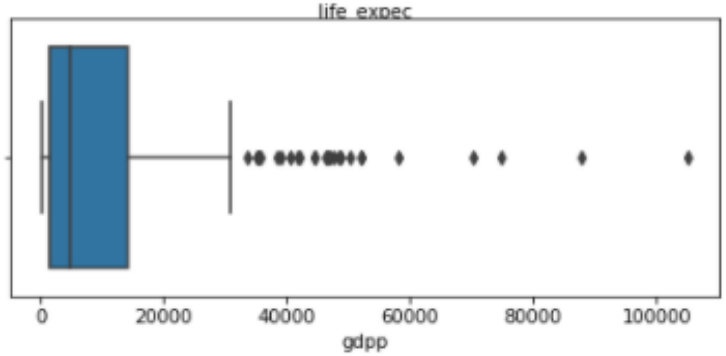












As we can see in box plots that there are outliers in more or else in every column of data. We are not touching the lower end of outlier of every columns, also we are not touching the ‘child_mort’ (lower & upper end)as we need those country with high child mortality and countries which are less economically developed .

Calculating the Hopkins statistic & Scaling the data

The Hopkins statistic, is a statistic which gives a value which indicates the cluster tendency and how well the data can be clustered.

- As we got the value between {0.85 -0.93}, it has a high tendency to cluster.

We used Standardized scaling as the variables are scaled in such a way that their mean is zero and standard deviation is one. So scaling all the values to the same normal scale helps us to form better and good cluster.

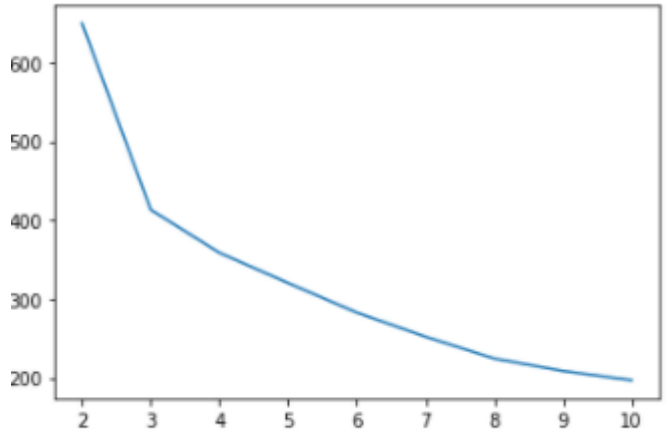
Clustering the data using K-mean

- Finding the value of 'K' by Silhouette Score and Elbow score. Also comparing both and concluding the final value of 'K'

Silhouette Score

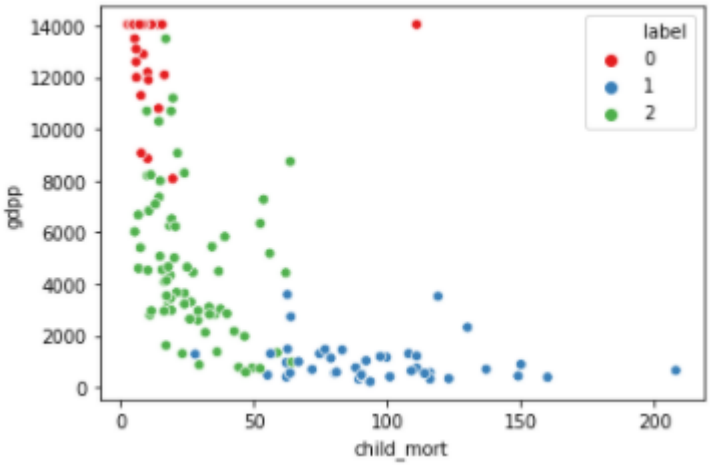
- For n_clusters=2, the silhouette score is 0.4816682906760779
- For n_clusters=3, the silhouette score is 0.4355194970640l694
- For n_clusters=4, the silhouette score is 0.3655245133176908
- For n_clusters=5, the silhouette score is 0.3526696784815590 6
- For n_clusters=6, the silhouette score is 0.337396934446371
- For n_clusters=7, the silhouette score is 0.318591942148701 9
- For n_clusters=8, the silhouette score is 0.298279392494728 9
- For n_clusters=9, the silhouette score is 0.2816593925610514 4
- For n_clusters=10, the silhouette score is 0.2699471612288636 4

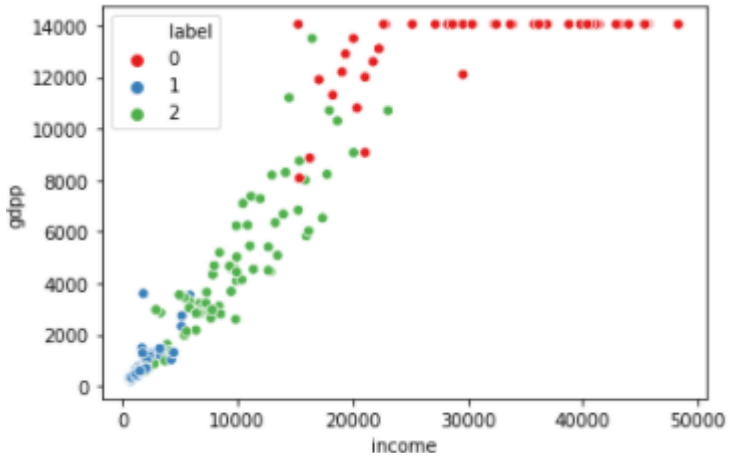
ELBOW CURVE

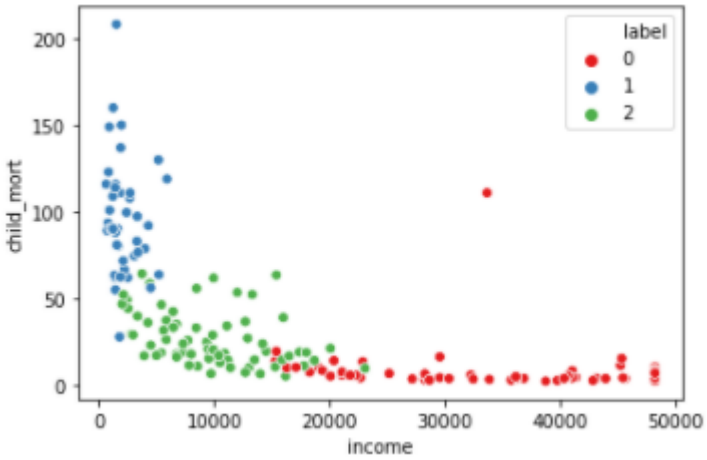


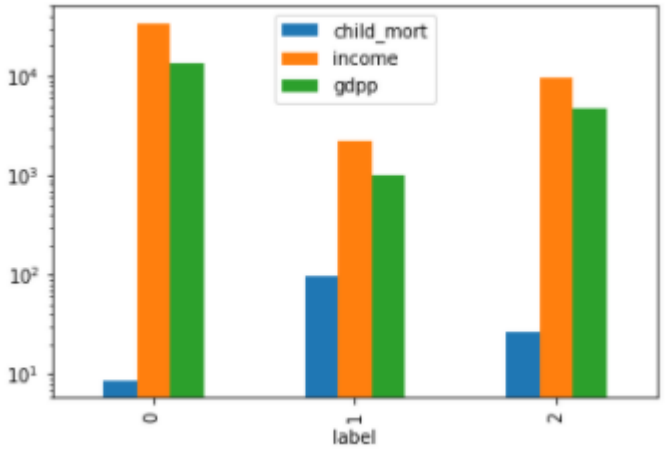
As we see in elbow curve after 3 there is no significant drop and when we saw percentage of scores in Silhouette Score method it suggested us to take 2 but it not recommended in industries to take 2 clusters as it only divide in to part of data . so we are taking 3 clusters as the Silhouette Score is near to 2 clusters .By comparing both elbow and Silhouette Score we came to conclusion that it is suitable to take 3 cluster(K=3).

VISUALISE THE CLUSTERS(K-MEAN)









- As we see in the graphs there is good distribution cluster of data .
- In the bar graph we have taken child_mort , income and Gdpp to know the cluster of countries which required the aid .
- Cluster 1 is containg the countries which in need of aid .

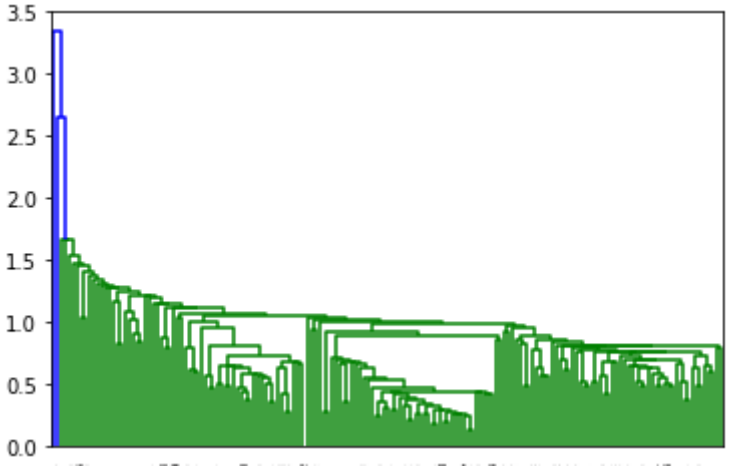
Cluster profiling and top 5 countries(K-mean)

- Now we will do cluster 1 profiling by taking 'child_mort(high) ,income(low) and Gdpp(low).
- The top 5 countries according to K-mean clustering are

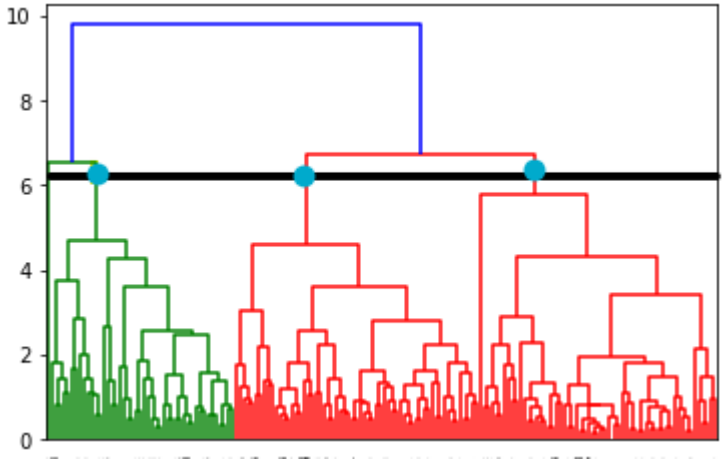
country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	label
Haiti	208.0	101.286	45.7442	428.314	1500.0	5.45	32.1	3.3300	662	1
Sierra Leone	160.0	67.032	52.2690	137.655	1220.0	17.20	55.0	5.2000	399	1
Chad	150.0	330.096	40.6341	390.195	1930.0	6.39	56.5	6.5636	897	1
Central African Republic	149.0	52.628	17.7508	118.190	888.0	2.01	47.5	5.2100	446	1
Mali	137.0	161.424	35.2584	248.508	1870.0	4.37	59.5	6.5500	708	1

Clustering the data (Hierarchical)

Single linkage clustering

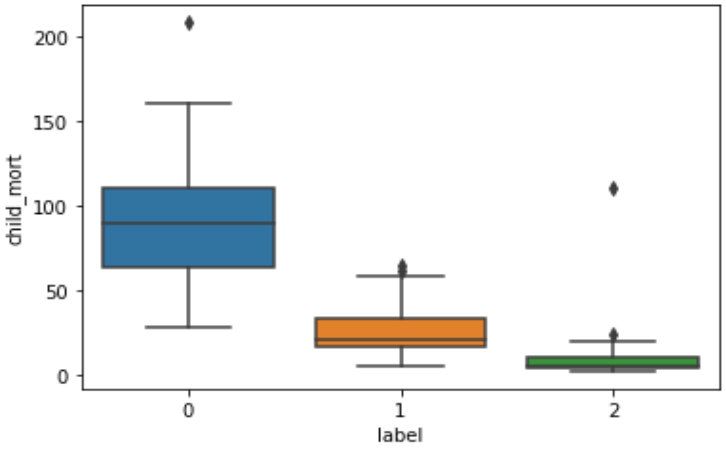


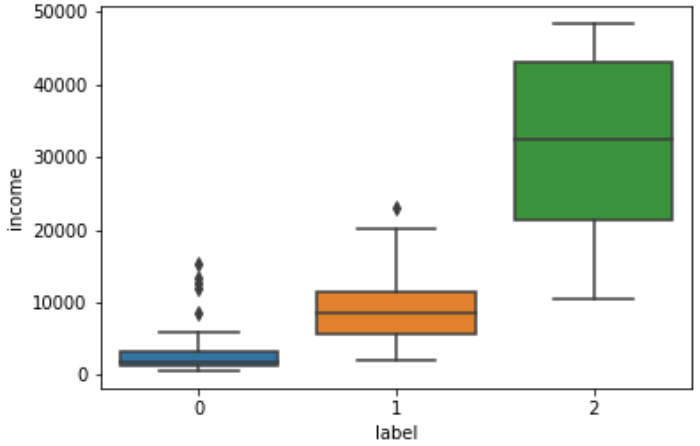
Complete linkage clustering

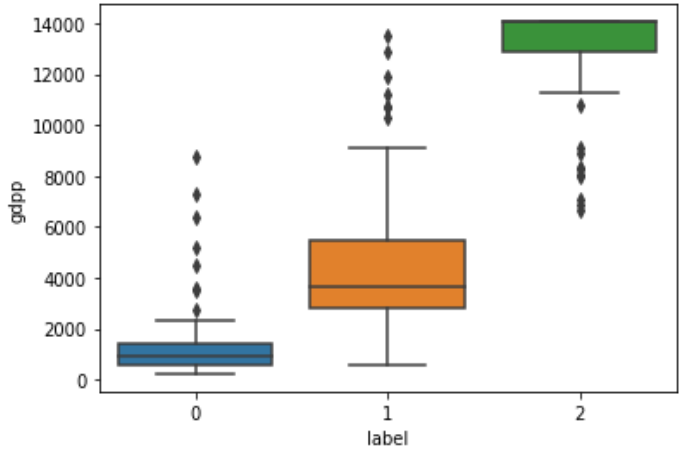


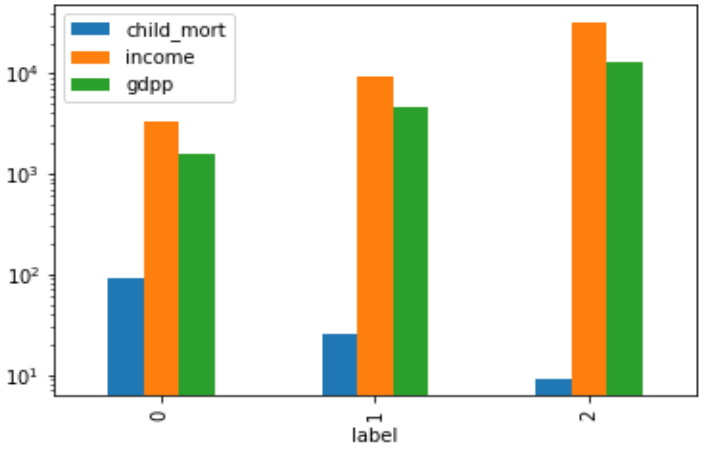
- We have used complete linkage hierarchical clustering no get the suitable no of cluster as compared to single linkage clustering.
- The clusters is obtained by cutting the dendrogram by divisive clustering method at an appropriate level.
- Thus if we cut the dendrogram at dissimilarity measure of 6, so we obtain 3 clusters.

VISUALISE THE CLUSTERS(HIERARCHICAL)









- As we see in the graphs there is good distribution cluster of data .
- In the bar graph we have taken child_mort , income and Gdpp to know the cluster of countries which required the aid .
- Cluster '0' is contain the countries which in need of aid .

Cluster profiling and top 5 countries(Hierarchical)

- Now we will do cluster 0 profiling by taking 'child_mort(high) ,income(low) and Gdpp(low).

- The top 5 countries according to Hierarchical clustering are

country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	label
Haiti	208.0	101.286	45.7442	428.314	1500.0	5.45	32.1	3.3300	662	0
Sierra Leone	160.0	67.032	52.2690	137.655	1220.0	17.20	55.0	5.2000	399	0
Chad	150.0	330.096	40.6341	390.195	1930.0	6.39	56.5	6.5636	897	0
Central African Republic	149.0	52.628	17.7508	118.190	888.0	2.01	47.5	5.2100	446	0
Mali	137.0	161.424	35.2584	248.508	1870.0	4.37	59.5	6.5500	708	0

Conclusion

- The top 5 countries according to K-mean clustering are 'Haiti', 'Sierra Leone', 'Chad', 'Central African Republic', 'Mali'.
- The top 5 countries according to Hierarchical clustering are 'Haiti', 'Sierra Leone', 'Chad', 'Central African Republic', 'Mali'.
- By looking at K-mean and Hierarchical clustering data ,we are getting same countries in both the cases ,so we recommend to the CEO that we have to focus more on these 5 countries for providing them aid.

Thank You