Clustering Assignment

Ankit Chand

Data Quality checks and handling missing values:

Data sets were provided

- Country Data

we have considered Country data set and performed necessary actions as explained below.

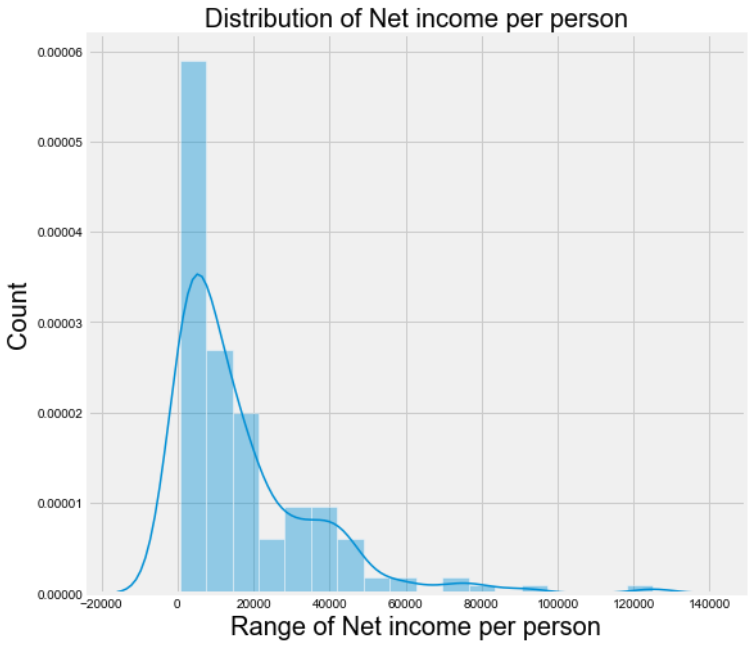
- There are no missing values

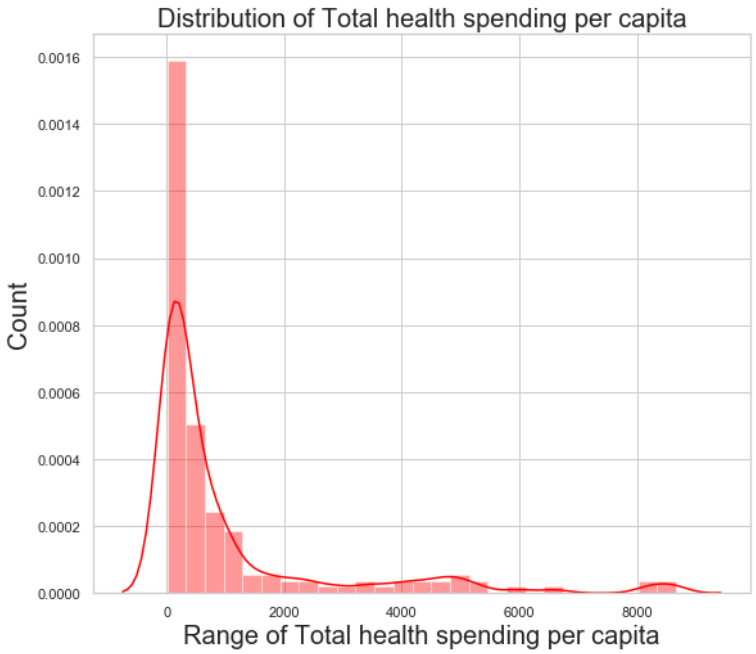
Data Preparation:

While understanding the data, we get to know that exports, import and health columns are in percentage. So, we have prepared our data by transforming into their actual values

Data Visualization:

Income vs health



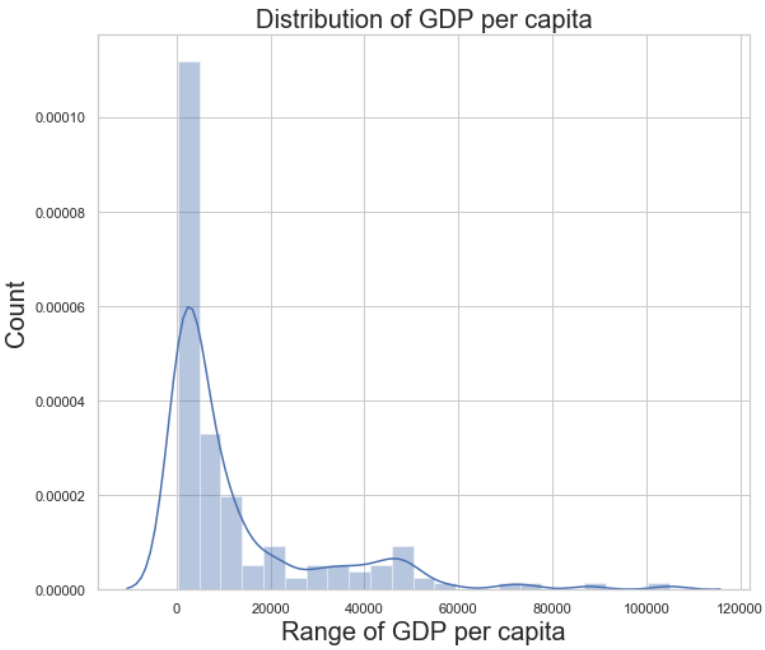


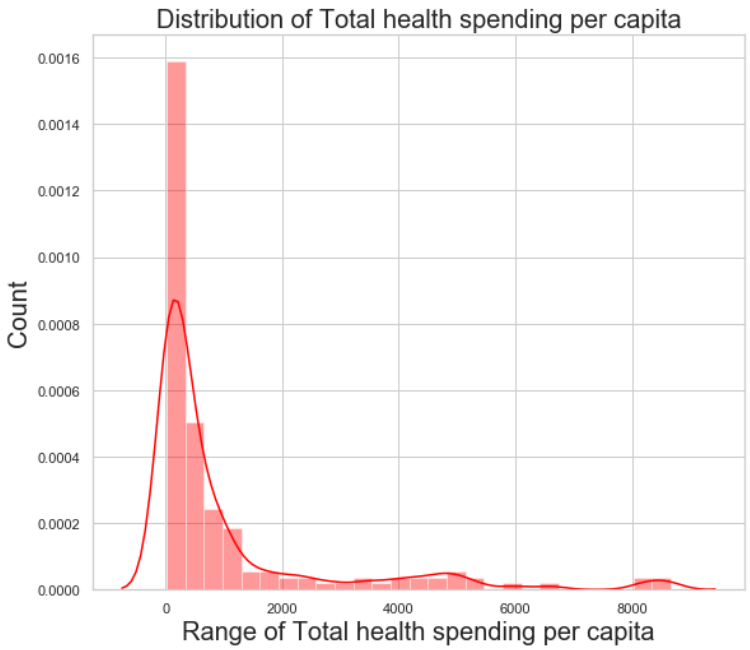
we can infer one thing that There are very few people who earn more than 1000 US Dollars. Most of the people have earnings of around 100-500 US Dollars. Also, we can say that the most of them were in least Income is around 20 US Dollars.

The total health spending per capita is very less due to less income but fewer were able to spend on their health.

From this, we can infer how income correlated with health means if income is more than they can able to spend more on their health if less then unable to spend on their health

GDP vs health

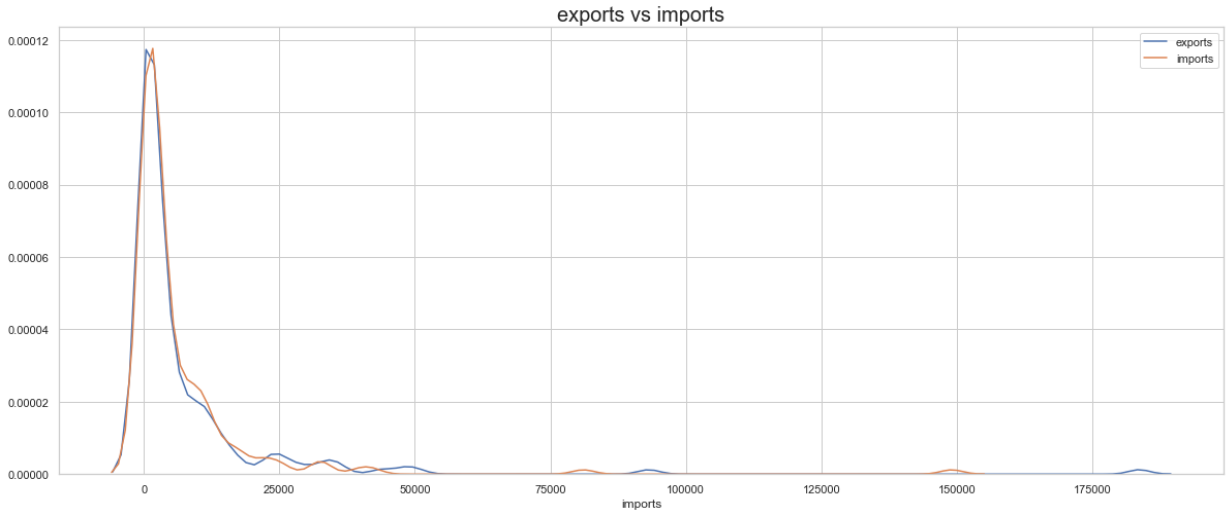




- More GDP per capita lies in range 10 to 15000 and less lies between 18000 to 50000More Total health spending per capita lies in range 10 to 1500 and less lies between 1800 to 5000

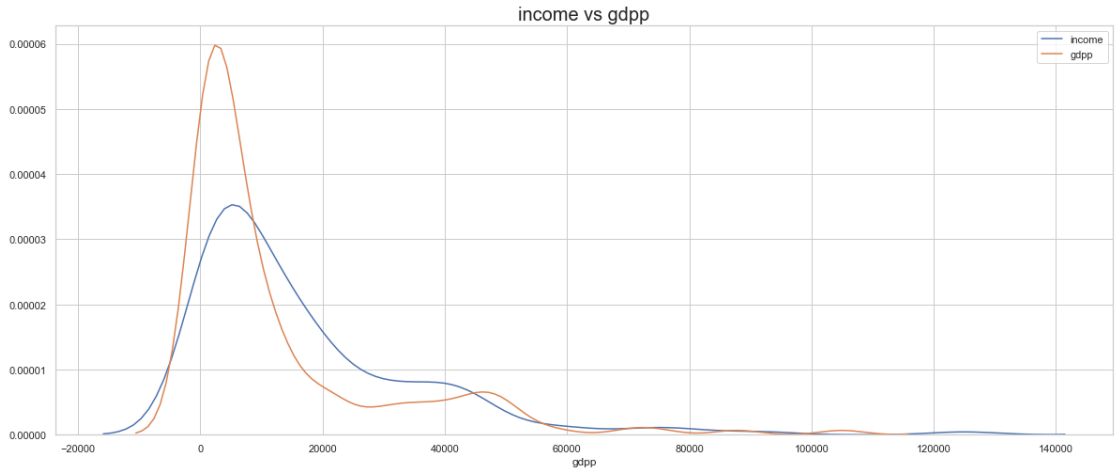
we can infer that if GDP per capita is more than total health spending per capita is also more. So, we can say that total health spending per capita is dependent on GDP per capita

exports vs imports



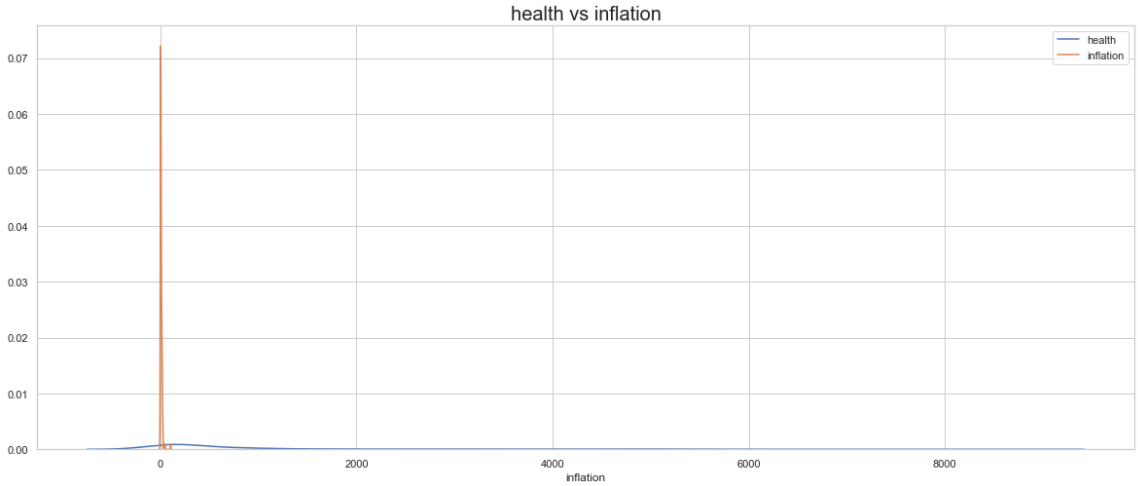
It clearly infers that amount of exports and imports were done approximately the same

income vs GDP



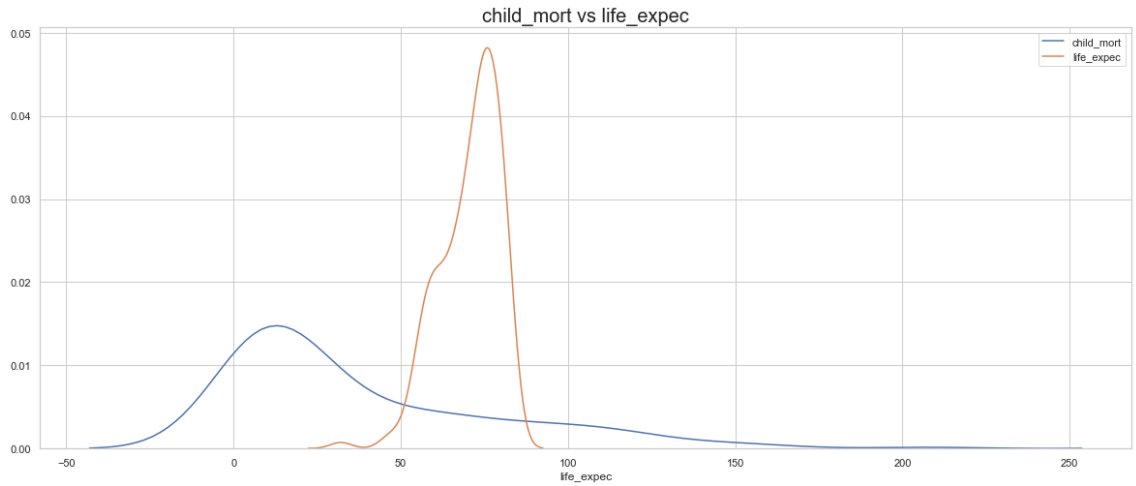
It infers that countries which are moderate earner have high GDP

health vs inflation



It infers that those countries which has an annual growth rate of total GDP is more had spent very less on their health

child_mort vs life_expect

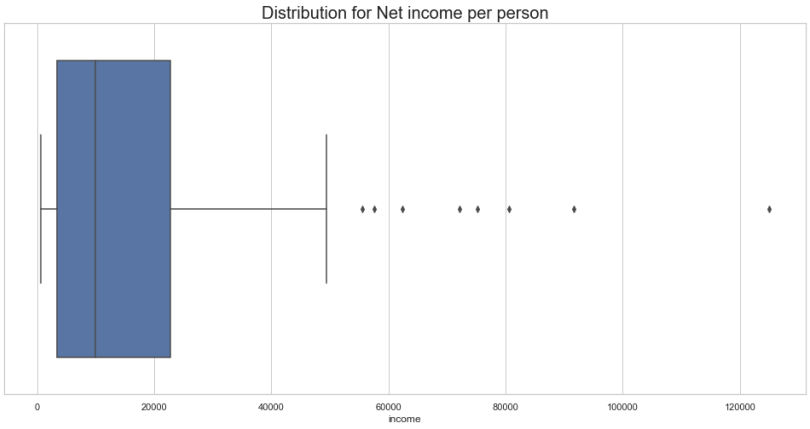


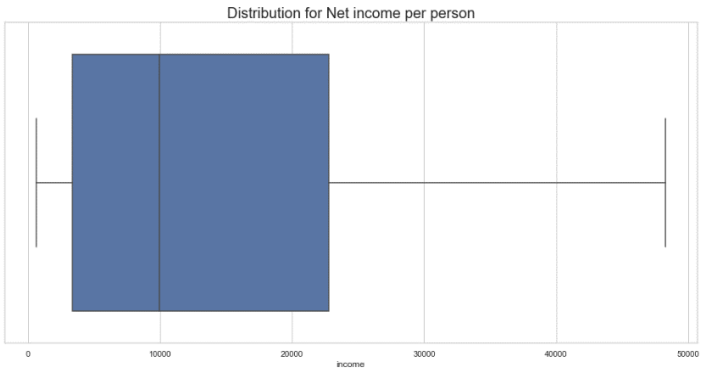
It infers that those countries which has child_mort more has very less life_expec and those countries which have child_mort less has very high life_expec

Outlier Treatment

Analyzing and Interpreting *outlier through* Skewness Values *by plotting boxplot*

income Column



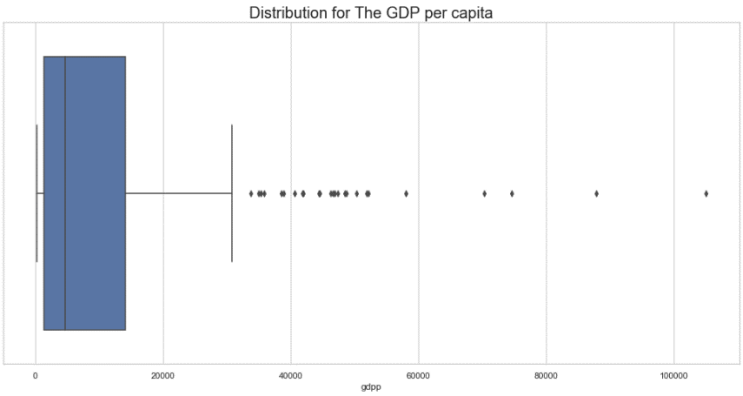


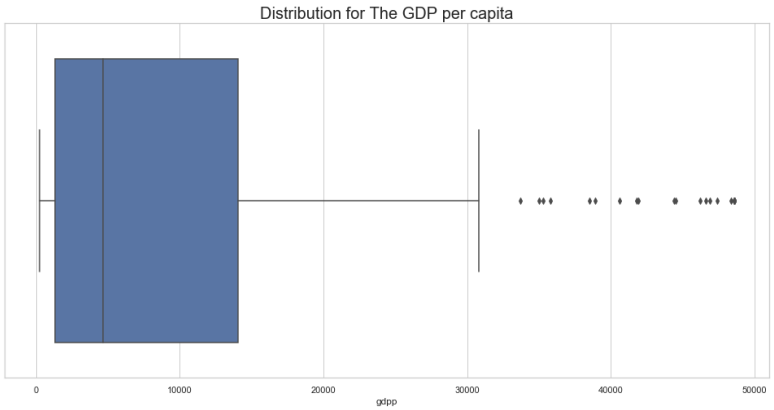
After Looking at the Box plot, it become clear that there are Outliers, and these outliers are also important to be treated for a better clustering model

- we do capping because if we drop them then we may lose countries that are dire need of AID

So, skewness gets overcome.(like in income column earlier skewness was 2.211386 and now it came down to 0)
For now, we can leave those values as it is. And good to go

GDP Column

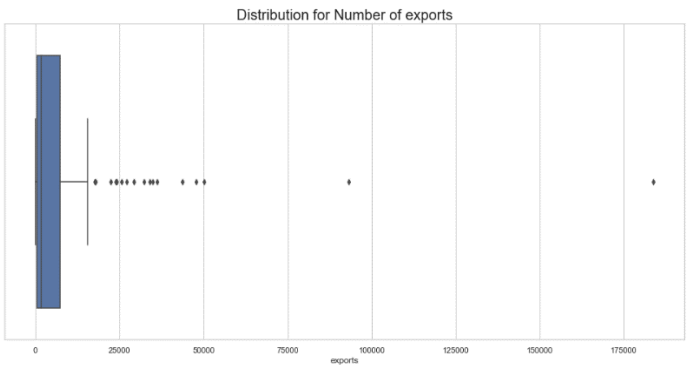


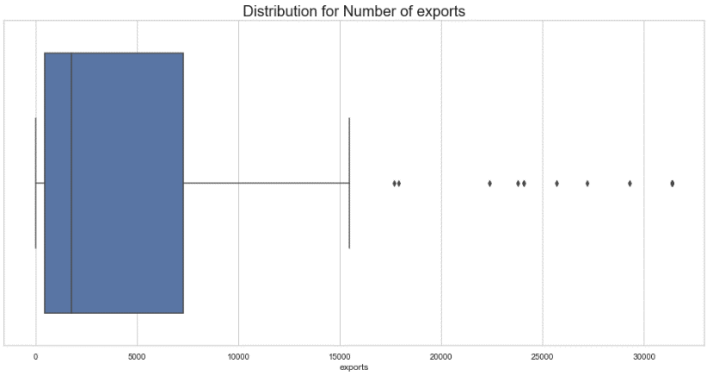


Again, we do capping, any value that is `beyond` the `95th percentile` we will `make` as 95th percentile
Good to go. If I take 90th percentile instead of 95th percentile than many data points can get affected.

So, skewness get overcome.(like in gdpp earlier 2.198079 and now 1.437938)
For now we can leave those values as it is. And good to go

Exports Column

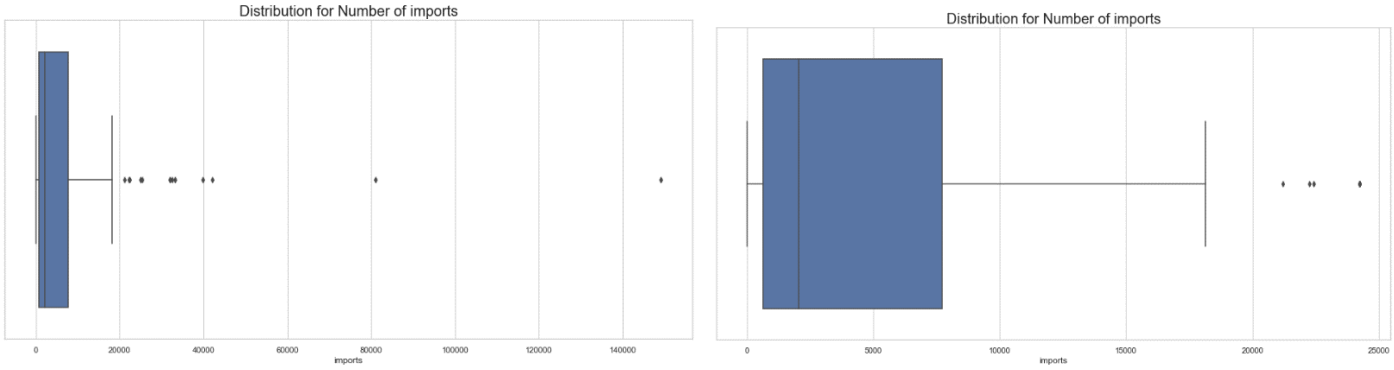




Yes, there are Outliers, and these outliers are also important to be treated for a better clustering model

After capping, still data points were lies in range 22000 to 32000. For good clustering we can leave for now

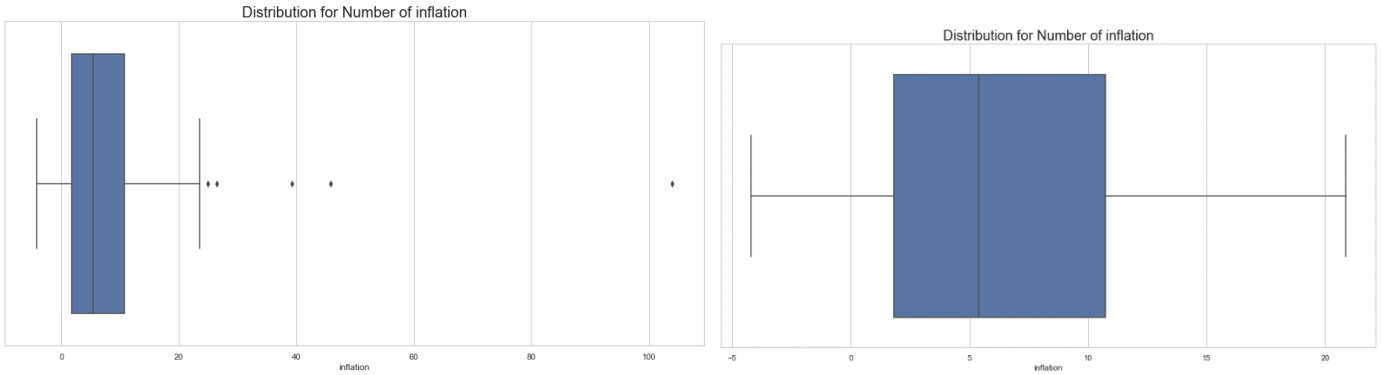
Imports column



After capping, somehow skewness gets overcome.(like in imports earlier 6.558903 and now 1.682672)

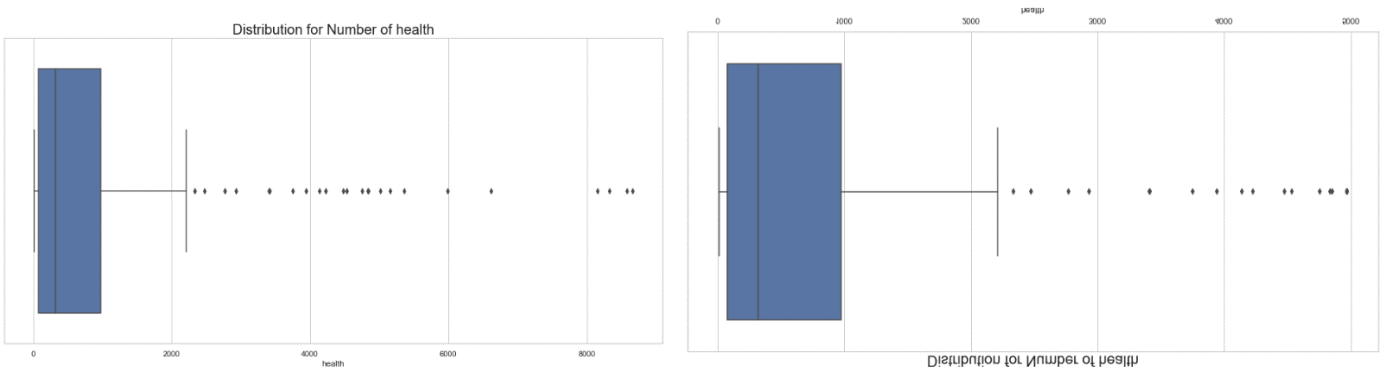
For now, we can leave those values as it is.

Inflation column



After capping, somehow skewness gets overcome.(like in inflation earlier 5.107640 and now 0.802852)

Health column



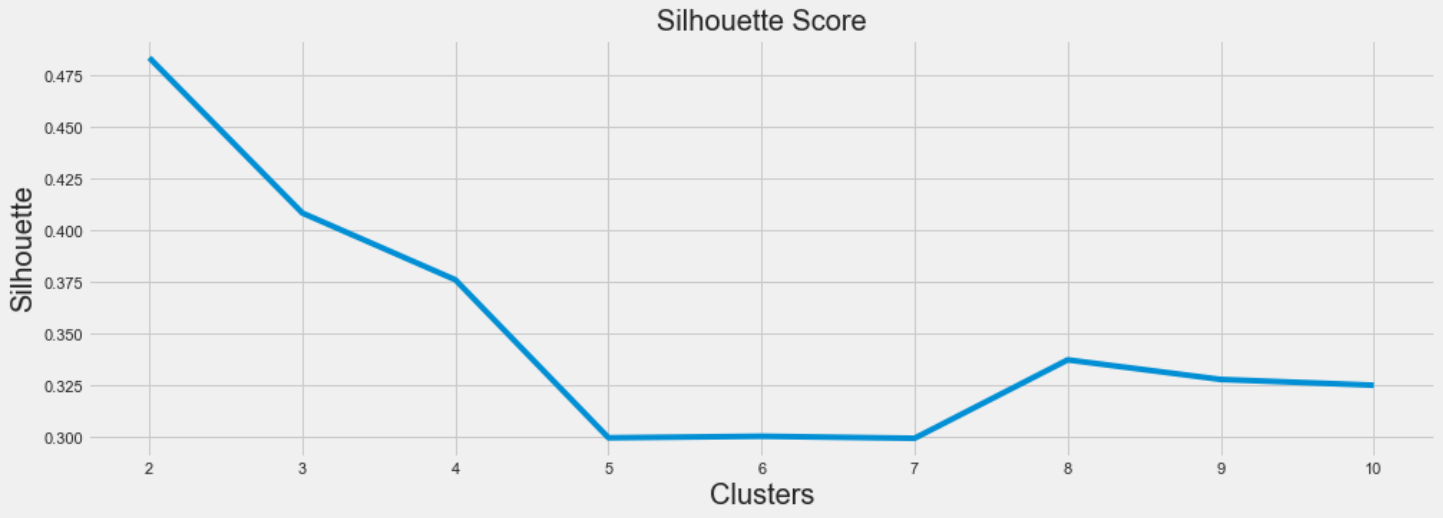
After capping, somehow skewness gets overcome.(like in health earlier 2.503283 and it came to 1.886402)

Hopkins Statistics: By checking my Hopkins score just to understand whether this data is good for clustering or not and it looks like we have a good cluster that can be formed using this data which has a high tendency to cluster.

Scaling: Scaling is performed mostly during model building processes to bring everything to the same scale. Standardized scaling, on the other hand, brings all the data points in a normal distribution with mean zero and standard deviation one. It can improve the efficiency of clustering algorithms. So, it becomes an essential step before clustering as Euclidean distance is very sensitive to the changes in the differences.

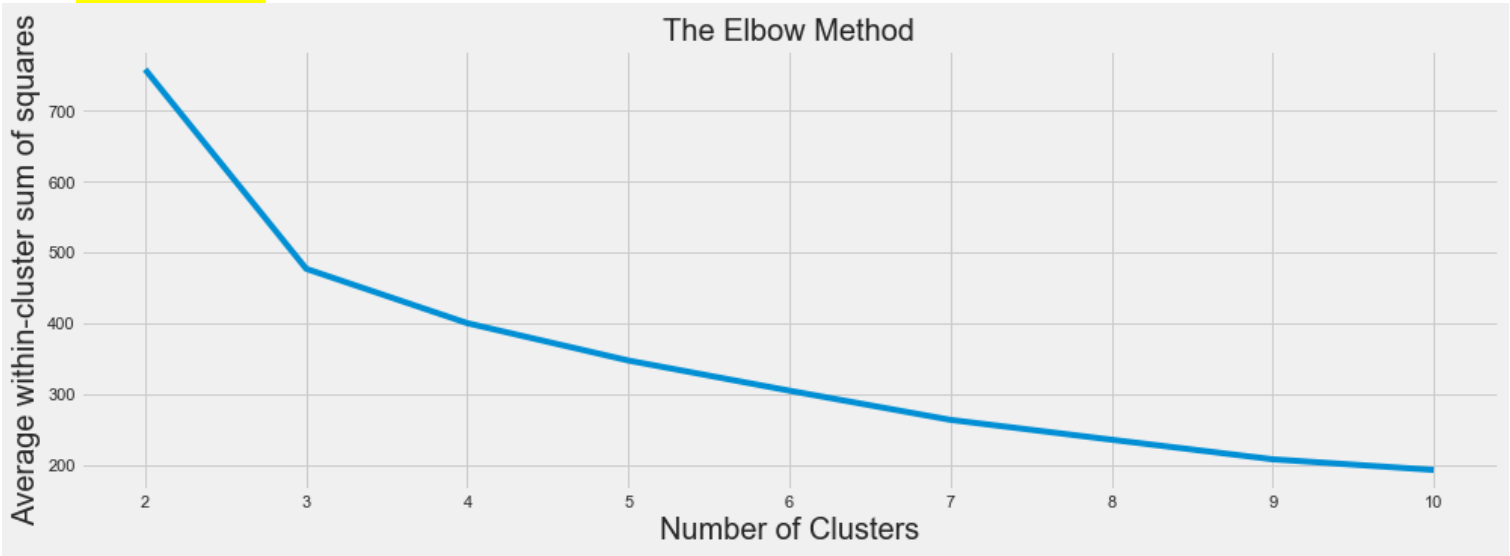
Let's find out Number of Optimal Clusters (value k):

Silhouette Score:



Silhouette Score, we will use that particular value k for which the Silhouette Score is maximum. So, we can see maximum at cluster K=2 but going with k=2 is not a good idea because k=2 basically means that we are just dividing data into two half's and that's why 2 is not always taken as number of clusters. So, we will go with next highest value that is 3.

Elbow Curve:



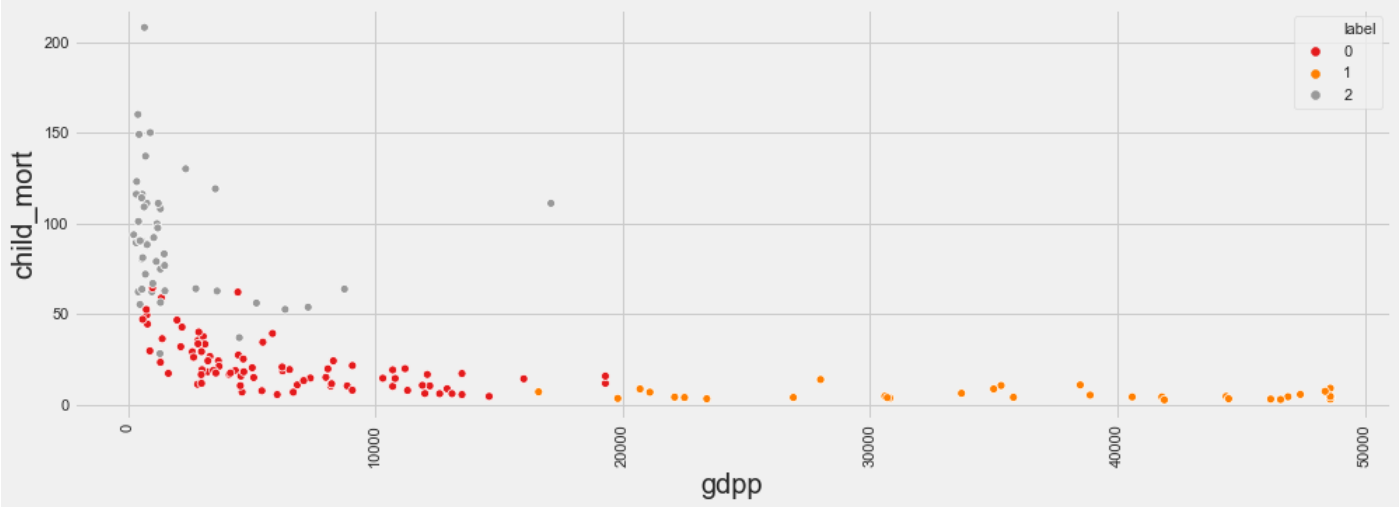
According to my business requirement we need to identify those country which are in urgent need of AID (Country having high child_mort, low income and low GDP). So, I can say that I will go with 3

Clustering Analysis:

K-means Algorithm
Visualizing the Cluster

Analyze the clusters by comparing how these three variables - [GDP, child_mort and income] vary for each cluster of countries to recognize and differentiate the clusters of developed countries from the clusters of under-developed countries.

GDP and child_mort

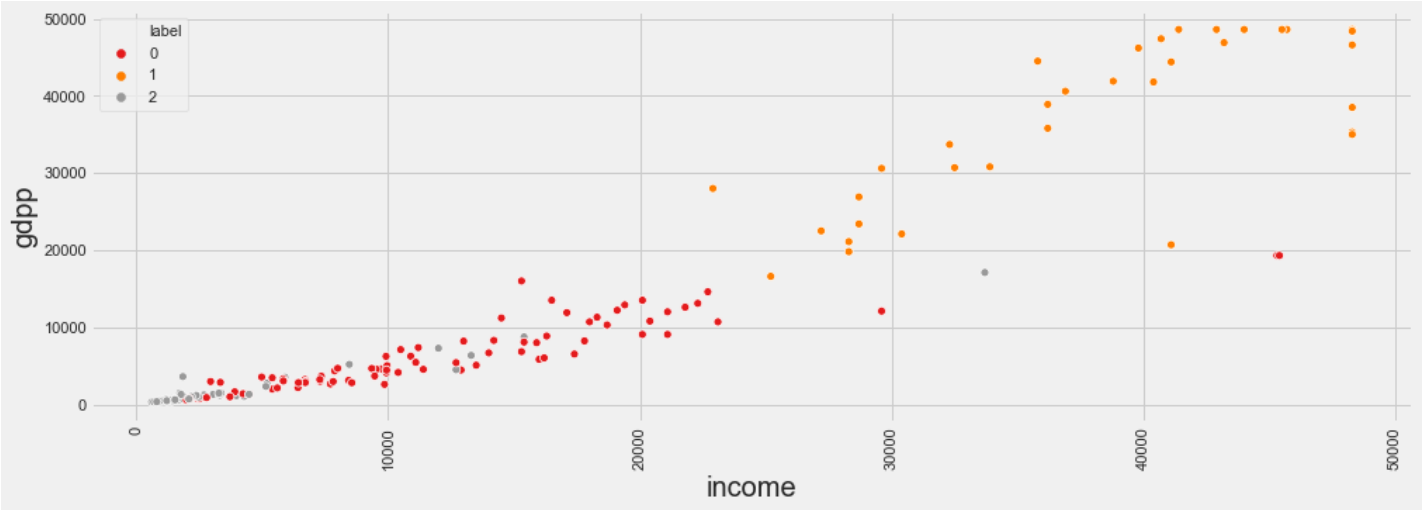


we can infer that

- cluster 2 are those whose child_mort is very high and GDP is very low
- cluster 0 are those whose child_mort and GDP is moderately high
- cluster 1 are those whose child_mort is very less and GDP is very high

According to our business requirement, cluster 2 are those country which are in urgent need of AID(Country having high child_mort, low income and low GDP)

income and GDP

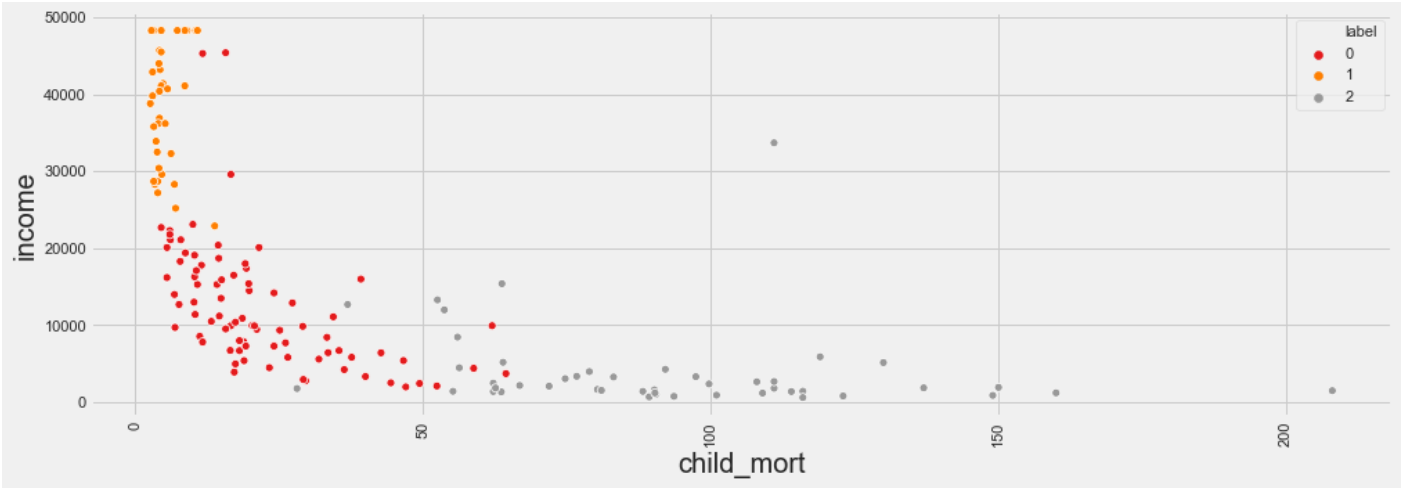


we can infer that

- cluster 2 are those whose income and GDP is very low
- cluster 0 are those whose income and GDP is moderate
- cluster 1 are those whose income and GDP is very high

According to our business requirement, cluster 2 are those country which are in urgent need of AID(Country having high child_mort, low income and low GDP)

child_mort and income

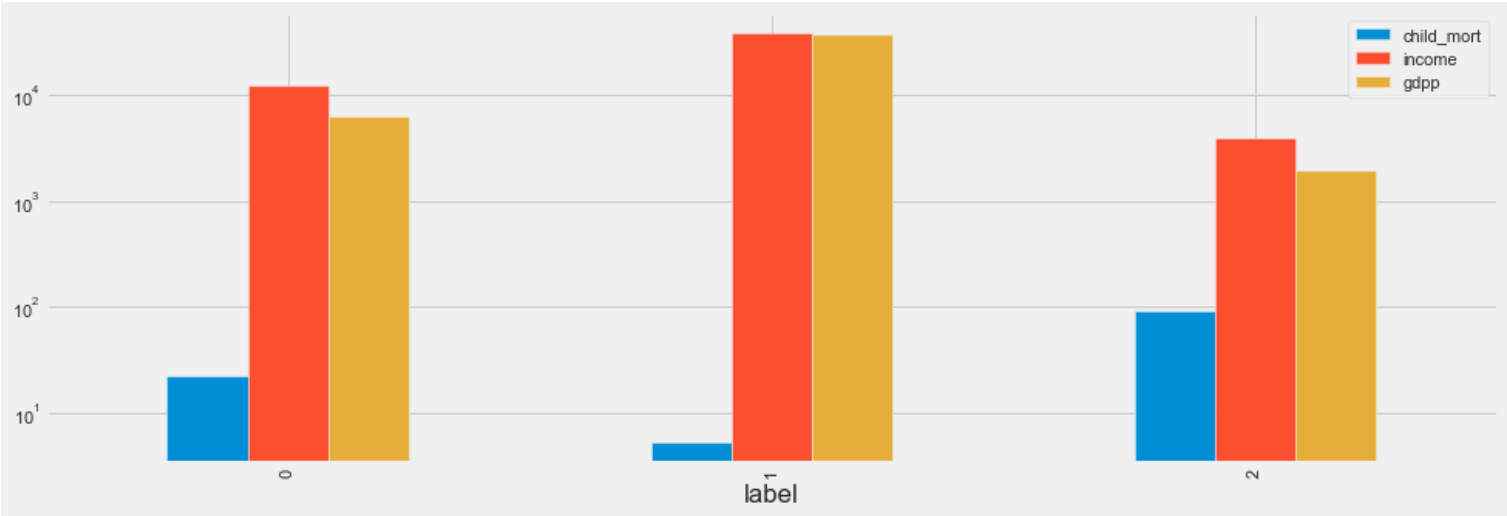


we can infer that

- cluster 2 are those whose income is very low and child_mort is very high
- cluster 0 are those whose income and child_mort is moderate
- cluster 1 are those whose income is very high and child_mort is very low

According to our business requirement, cluster 2 are those country which are in urgent need of AID(Country having high child_mort, low income and low GDP)

Cluster Profiling: Trying to understand better that what one cluster talking about



we can see by bar plot that

- cluster 2 are those types of countries that are having high child_mort and moderate income and GDP
- cluster 1 are those types of countries that are having low child_mort, high income and GDP
- cluster 0 are those types of countries that are having moderate child_mort and high income and GDP

Therefore, cluster 2 are those country which are in urgent need of AID(Country having high child_mort, low income and low GDP)

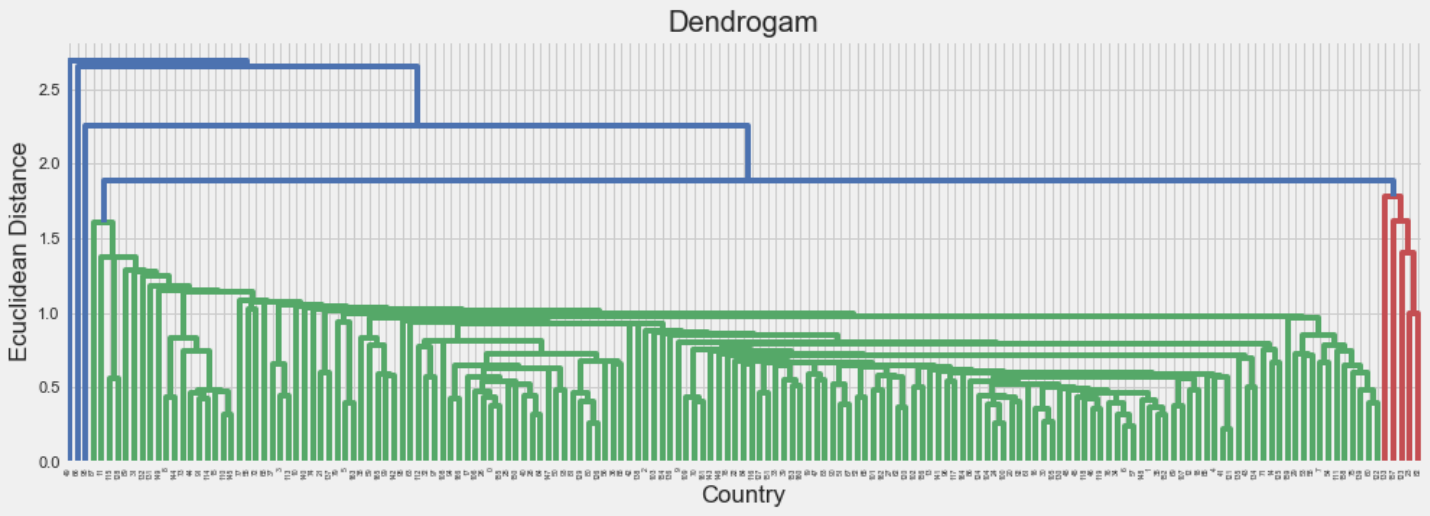
My top 5 countries are:

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	label
66	Haiti	208.0	101.29	45.74	428.31	1500.0	5.45	32.1	3.33	662	2
132	Sierra Leone	160.0	67.03	52.27	137.66	1220.0	17.20	55.0	5.20	399	2
32	Chad	150.0	330.10	40.63	390.20	1930.0	6.39	56.5	6.59	897	2
31	Central African Republic	149.0	52.63	17.75	118.19	888.0	2.01	47.5	5.21	446	2
97	Mali	137.0	161.42	35.26	248.51	1870.0	4.37	59.5	6.55	708	2

Hierarchical Clustering

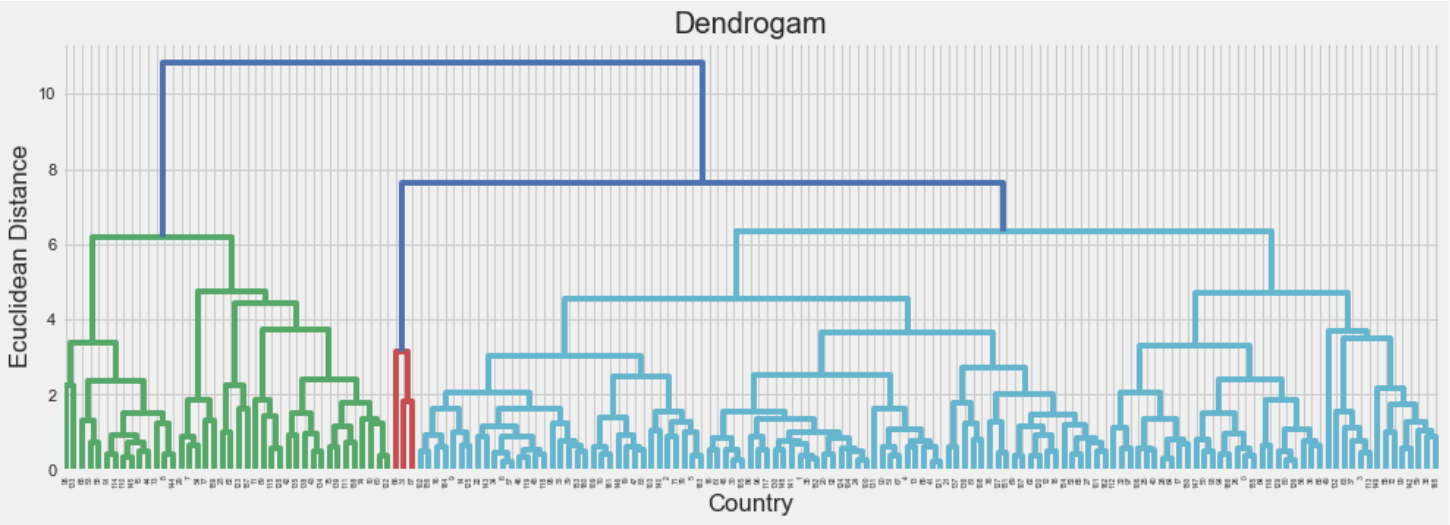
Using Dendrograms to find the number of Optimal Clusters

Single Linkage



consider shortest distance between two points in each cluster

Complete Linkage

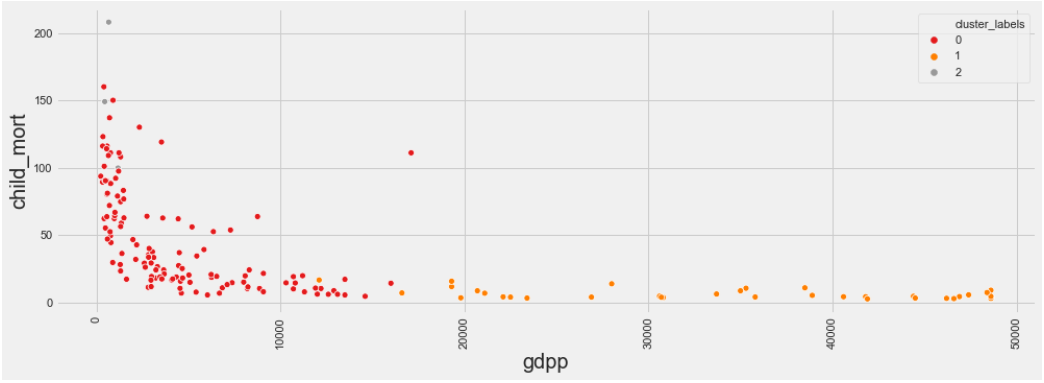


considered the longest distance between 2 points in each cluster

Based on the result we can choose complete linkage for Hierarchical Clustering because it does well in separating clusters if there is noise between clusters and we will cut our dendrogram at 7 which we will get 3 clusters

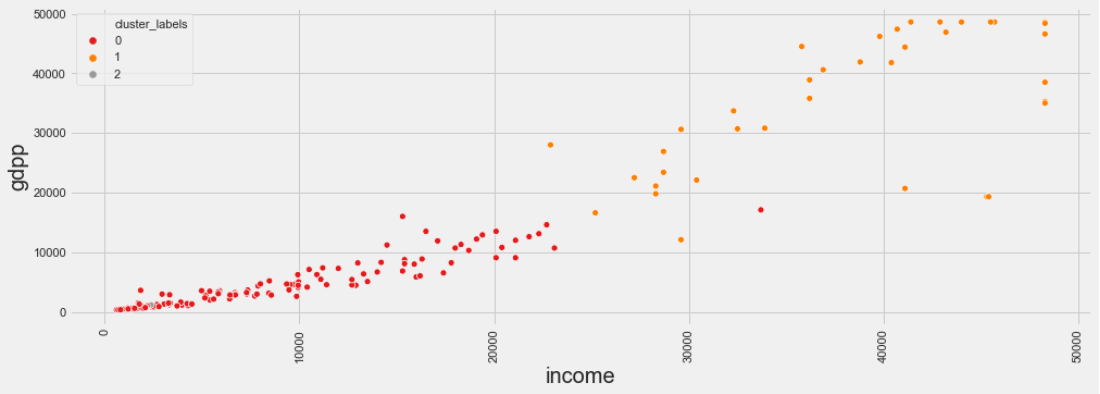
Visualizing the Clusters of Hierarchical Clustering

GDP and child_mort



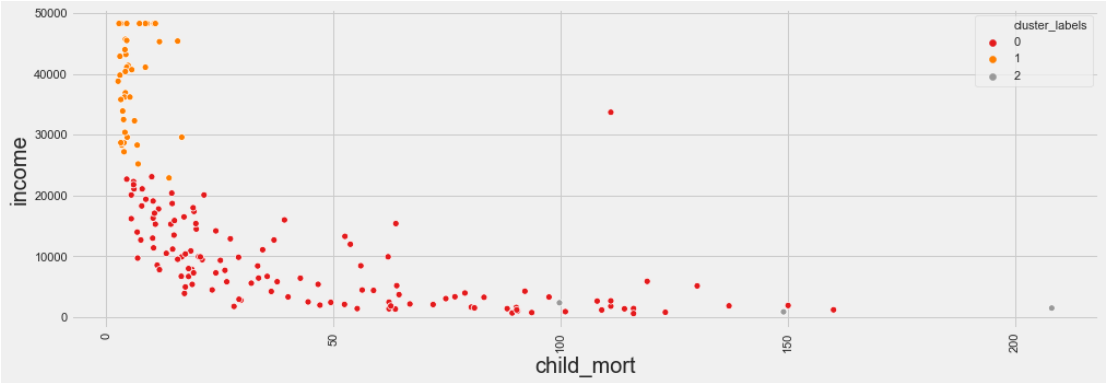
- cluster 2 are those whose child_mort is very high and GDP is very low
- cluster 0 are those whose child_mort and GDP is moderately high
- cluster 1 are those whose child_mort is very less and GDP is very high

income and GDP

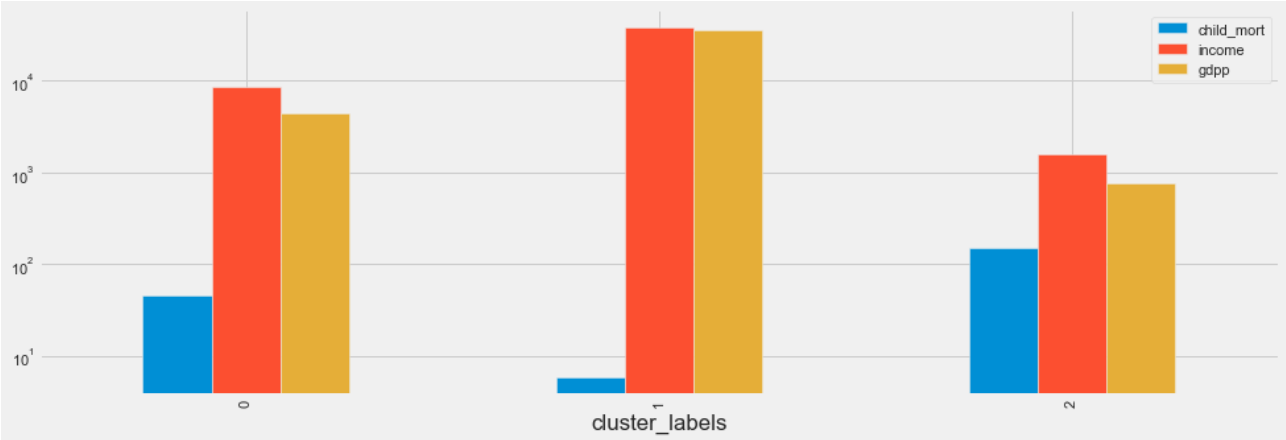


- cluster 2 are those whose income and GDP is very low
- cluster 0 are those whose income and GDP is moderate
- cluster 1 are those whose income and GDP is very high

child_mort and income



- cluster 2 are those whose income is very low and child_mort is very high
- cluster 0 are those whose income and child_mort is moderate
- cluster 1 are those whose income is very high and child_mort is very low



According to our business requirement, cluster 2 are those country which are in urgent need of AID(Country having high child_mort, low income and low GDP)

Conclusion

So, from this analysis we can by doing Hierarchical and K-means Clustering we have got the same countries that are in dire need of AID.

The countries which are in dire need of aid based on some socio-economic and health factors are Haiti, Central African Republic, Lesotho, Sierra Leone, Mali and Chad.