Clustering Assignment Part 1

Jay Thakur

May 2020

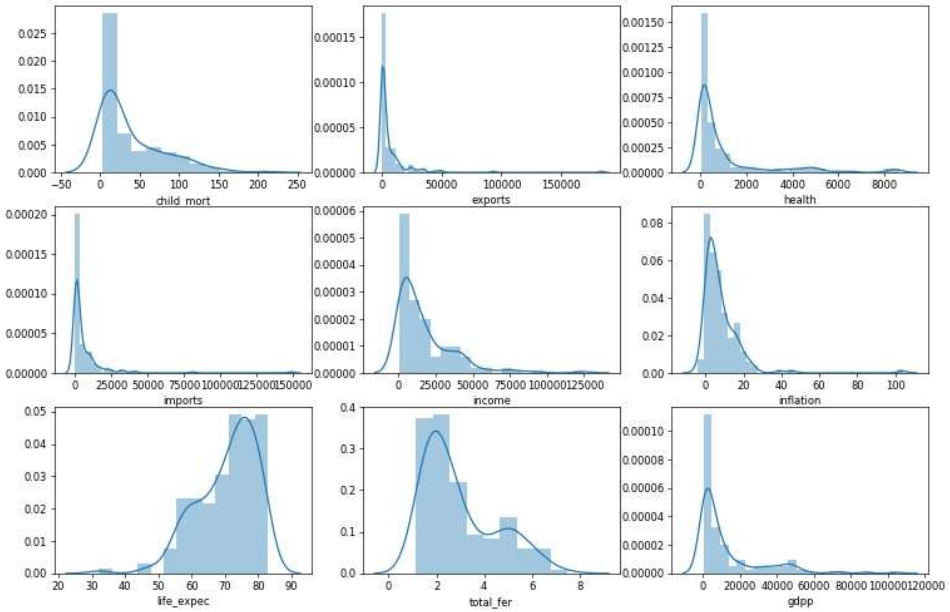
Problem Statement

We need to categories the countries using socio-economic and health factors that determine the overall development of the country. Then we need to suggest the countries which the CEO (HELP International which is an International NGO) needs to focus on the most.

Analysis Approach

- Firstly, the Data Extraction/Reading step was performed over the dataset by reading the data from the .csv file into Dataframe i.e. **country_df.**
- After reading the data into dataframe, we performed Preprocessing part by understanding its shape (no of rows and columns), checking data types and outliers. On the basis of above steps and visualization using Box Plot, we treated some of outliers as data size was low, we didn't cap all the outliers.
- After cleaning data, we ran visualization using Distribution Plot and Box Plot to generate some insights from various variables.
- Now, we are ready to perform Modelling but before that we ran Hopkins Statistics to see how well the data can be clustered, then we rescaled the data for further inspection and finding 'k' (cluster) value.
- We performed Elbow Curve and Silhouette Score to find optimal 'k' value for model building.
- Now, we can build our model using KMean Clustering Algorithm and generate our first result of 'Top 5 Countries in need of urgent Aid'.
- Similarly, we performed Hierarchical Clustering Algorithm for model building, before that we used Linkages (i.e. Single Linkage, Complete Linkage and Average Linkage) to find optimal value for 'k' and among these three Complete Linkage gave suitable value.
- Lastly, we performed model building using Hierarchical Clustering Algorithm and generated our second result of 'Top 5 Countries in need of urgent Aid'.

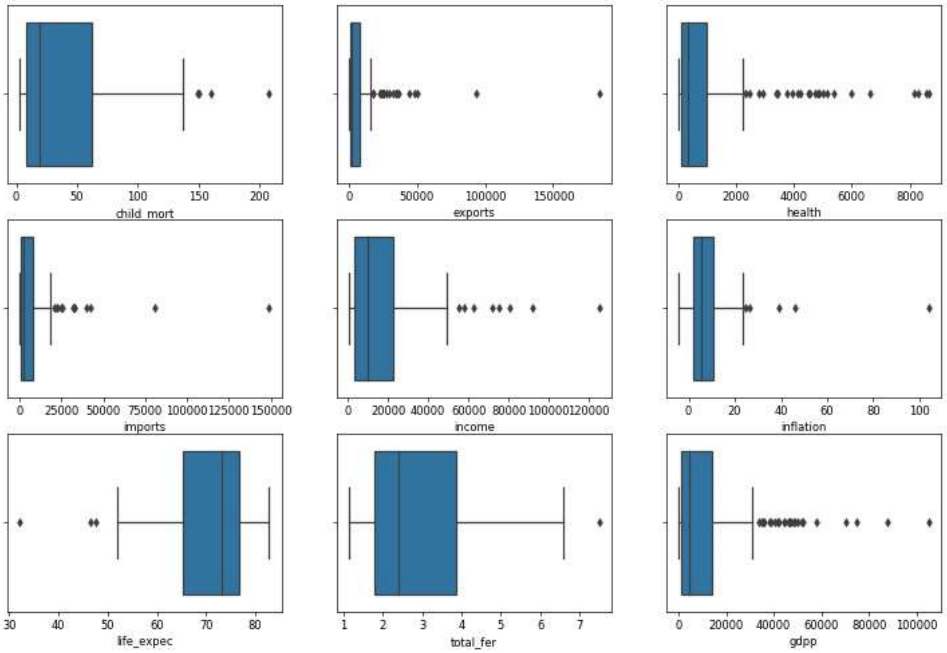
1. General Visualization



Observations:

- 1. `Child Mortality` plot shows that the `child mortality rate per 1000 live births of children(>5) highly ranges from 10 to 150 which is approx 1% to 15% per 1000 live births of children(>5)` and this seems bad.
- 2. Most of the countries seems to `spend around 0 to 2000` in their currency on `health per capita` and `very less countries seems to spend more than 4000 per capita` in their currency.
- 3. For most of the countries the `imports` and `exports` figure almost `seems to equal per capita`.
- 4. High number of people seems to have `income of 10000 to 53000` in most countries.
- 5. In most of the countries the `inflation rate seems to be low which is the result of low demand for goods and services`.
- 6. The `life expectency of new born child` with current mortality rate `ranges between age 50 to age 82` which seems to be moderate but not good.
- 7. The `number of children` to be `born to a woman according to the plot is 2-3` in most countries, than `4-5` for some and than `5-7` for left over countries `which could lead to faster growth in population`.
- 8. For most countries `GDPP` which is the GDP/Total Population of Country approximately `ranges between 5000 to 20000 and so on`.

2. Outlier Visualization



Observations:

- 1. Each and every variables seems to have low or higher or both range of outliers.
- 2. Export, Health and GDPP are the variables with highest high range outliers.
- 3. life_expec is the only low range outliers but very less.
- 4. Some other variables with high range outliers are child_mort, imports, income, inflation and total_fer.

Observations:

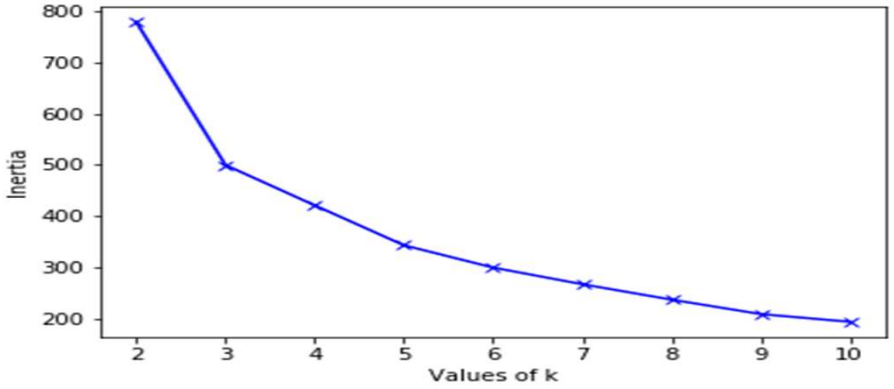
1. We are not going to apply capping to higher range outliers in `child_mort` and `inflation` because those outliers will help us to understand about the countries graph related to these variables and accordingly to distribute required aid money.

2. Lower `gdpp` could also help us in finding countries in need of financial aid. So we are going to remove only higher range outlier under this variable.

3. Other variables like `exports` `imports` `health` and `income` are having much higher quartile range which could impact the result and also they are of no use due to much higher range than required or normal values. Hence, we will only remove the higher range outliers in these variables.

K-means Clustering Algorithm

K-means Elbow Curve and Silhouette Score Results for k value




```
For k=2, the Silhoette Score is 0.5
For k=3, the Silhoette Score is 0.43
For k=4, the Silhoette Score is 0.43
For k=5, the Silhoette Score is 0.4
For k=6, the Silhoette Score is 0.32
For k=7, the Silhoette Score is 0.29
For k=8, the Silhoette Score is 0.27
For k=9, the Silhoette Score is 0.31
For k=10, the Silhoette Score is 0.3
```

```
2    82
0    48
1    37
Name: labels, dtype: int64
```

Observations:

- We can see that the elbow point hover over k as 3 and after the plot does not make any drastic bend. `Hence, according to the Elbow Curve output (line plot) we will choose k=3`.
- But to finalize value of k we will check Silhouette Score.

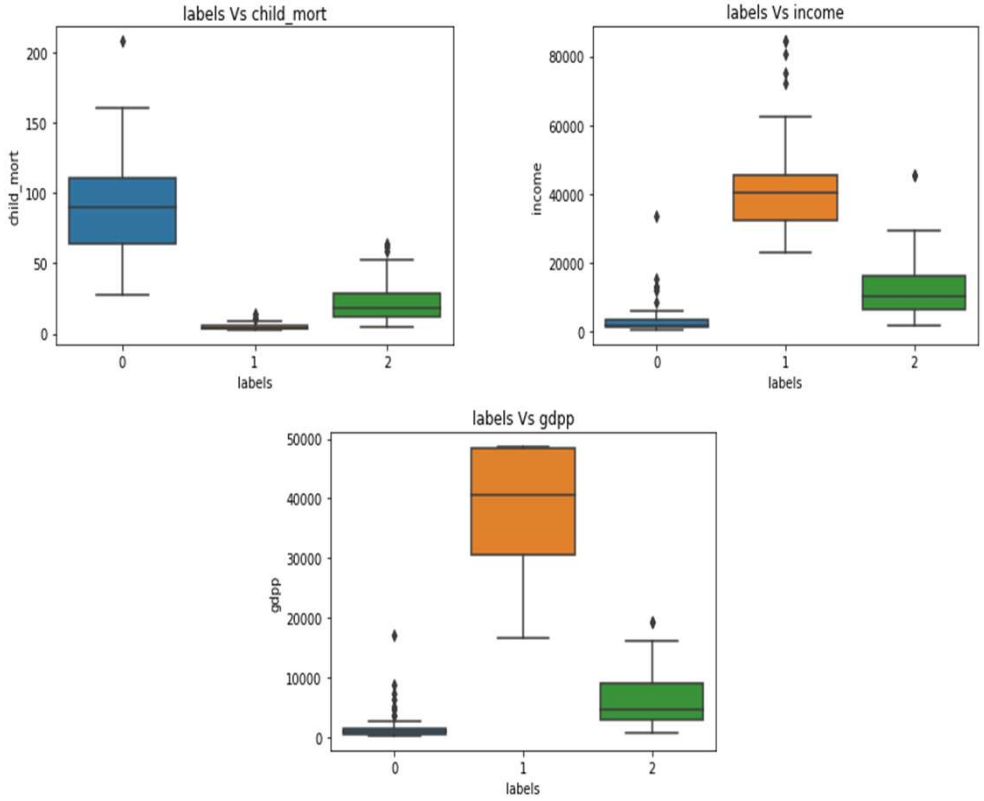
Observations:

- After analyzing Silhouette Score we can say that though `k=3 has second highest Silhouette Score` will go with it, due to the moderate significance.

Observations:

- We can see in 3rd picture the no of values assigned to which labels after building K-means model.

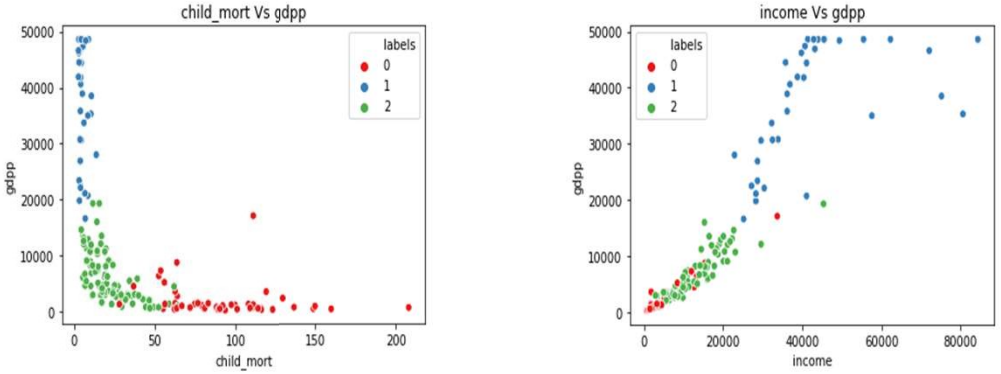
K-means Model Box Plot Visualization

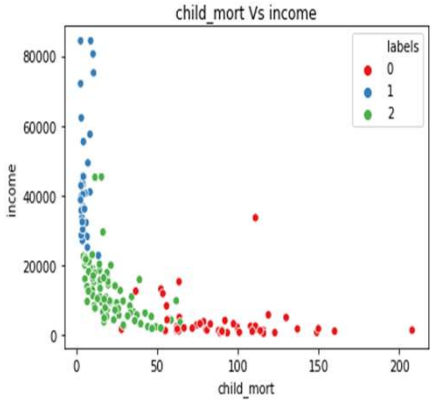


Observations:

- Plot 1 indicates that Countries under label 0 has higher mortality rate with average approx. 98-100 child mortality per 1000 child(age <5) followed by label 2 and label 1 Countries with average child mortality 20-25 and 8-10 over 1000 child (age <5) respectively.
- Plot 2 indicates the Countries under label 0 have lower income per capita i.e. approx. average income 1000 followed by Countries under the label 1 and 2 having average income per captia as approx. 41000 and 10000.
- Plot 3 indicates that the Countries under label 0 have lower gdpp with approx. gdpp 1000-2000 followed by Countries under the label 2 and 1 having average gdpp as approx. 5000 and 40000.

K-means Model Scatter Plot Visualization

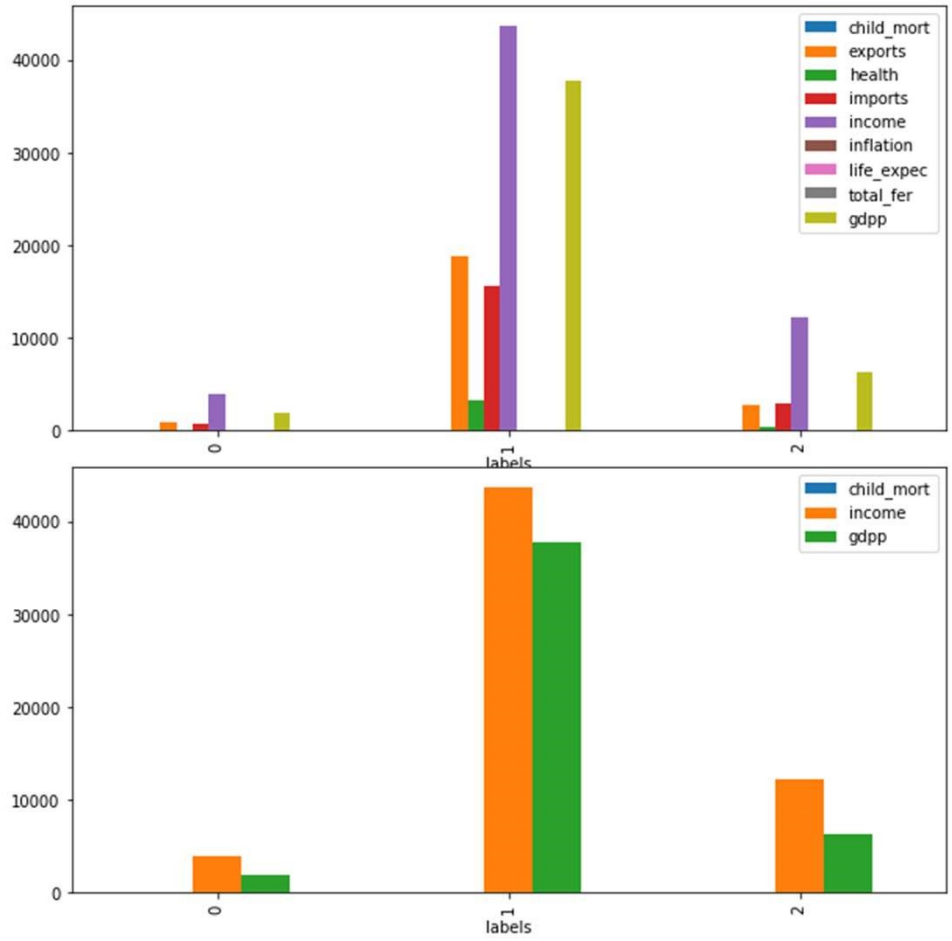




Observations

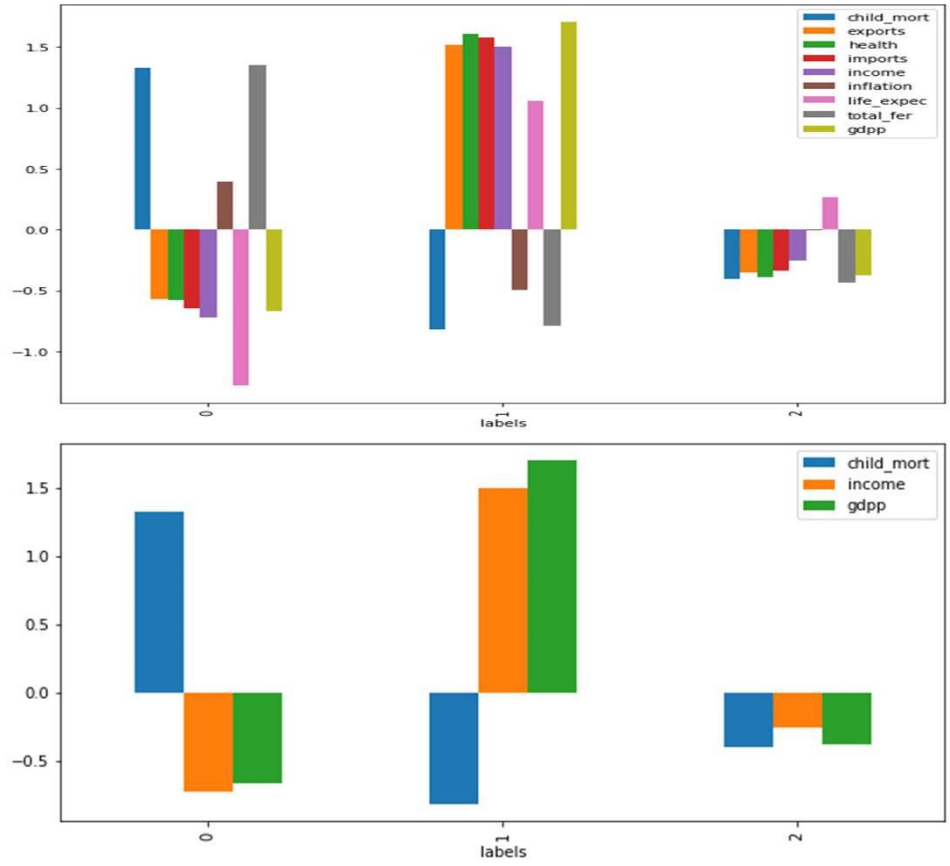
- Even scatter plot suggest same pattern for Countries under label 0 and 2 like
  - Lower GDPP, Higher Child Mortality rate per 1000 child(age <5).
  - Lower GDPP, Lower Income per Capita.
  - Lower Income, Higher Child Mortality rate per 1000 child(age <5).
- And vice versa

K-means Model Profiling(Unscaled Data)



As we can see that due to much higher values of `income` and `gdpp` we are unable to see the bar plot for `child_mort`. Hence we will try to use scaled dataframe of same data and plot the same bar plot.

K-means Model Profiling(Scaled Data)



Observations

- Looking at both the bar plots we can clearly say that Countries under label 0 need an urgent Aid as their `child_mort` is higher, `income per capita` is lower and `gdpp` is also lower than Countries under label 1 and 2.

Final Suggestion to the CEO of HELP International NGO

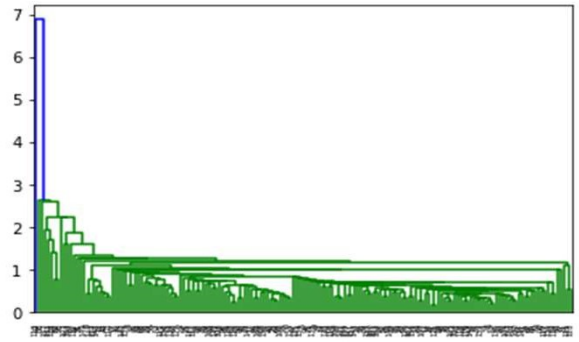
On the basis of the KMeans and Hierarchical model building I would suggest to provide an urgent AID to countries

- Burundi
- Liberia
- Congo, Dem. Rep.
- Niger
- Sierra Leone

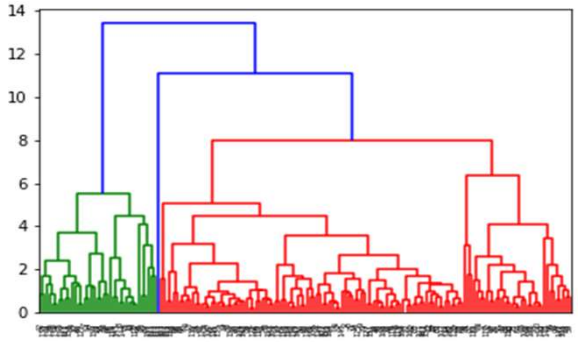
Hierarchical Clustering Algorithm

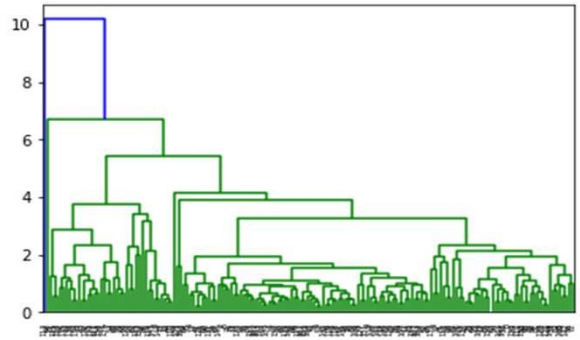
All Three Linkage Results

Single Linkage



Complete Linkage






```
0    165
2      1
1      1
Name: hLabels, dtype: int64
```

Average Linkage

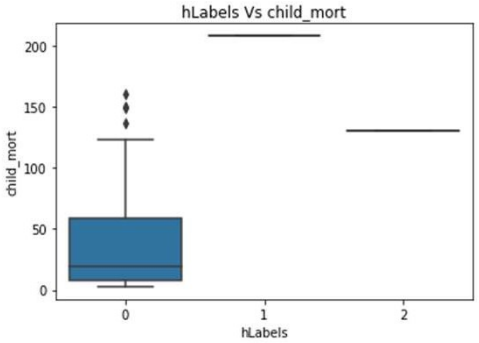
Observations:

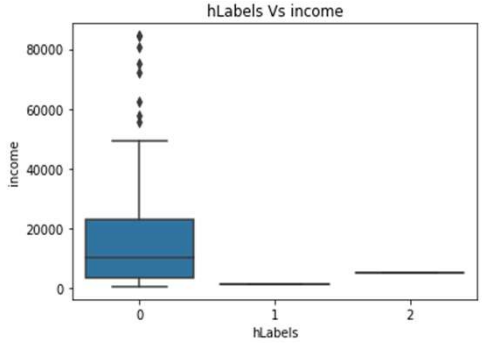
- Plotting of all three linkages we could conclude that `Complete Linkage` gives more clear visualization of the data we are using by answering the question of `What should be the best k value` and after observing complete linkage plotting we could clearly say that value of k should 3 i.e. `number of clusters k = 3`.

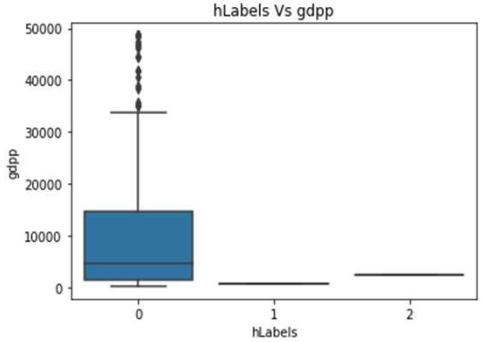
Observations:

- We can see in 4th picture the no of values assigned to which labels after building K-means model.

Hierarchical Model Box Plot Visualization



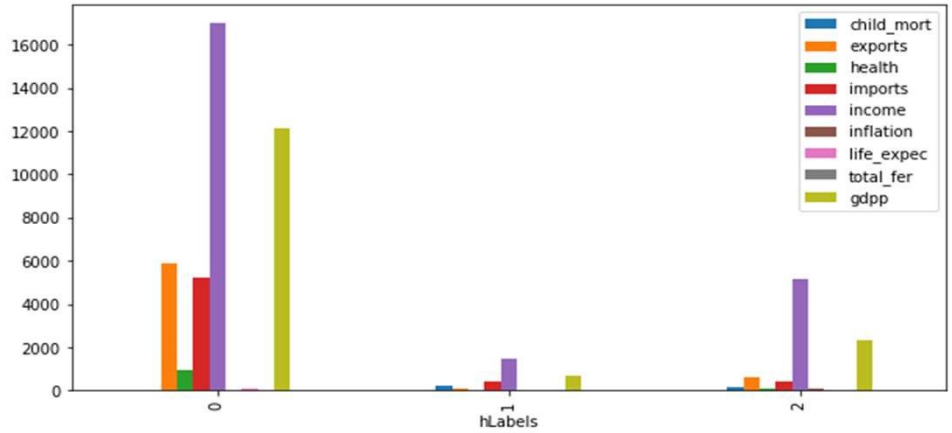


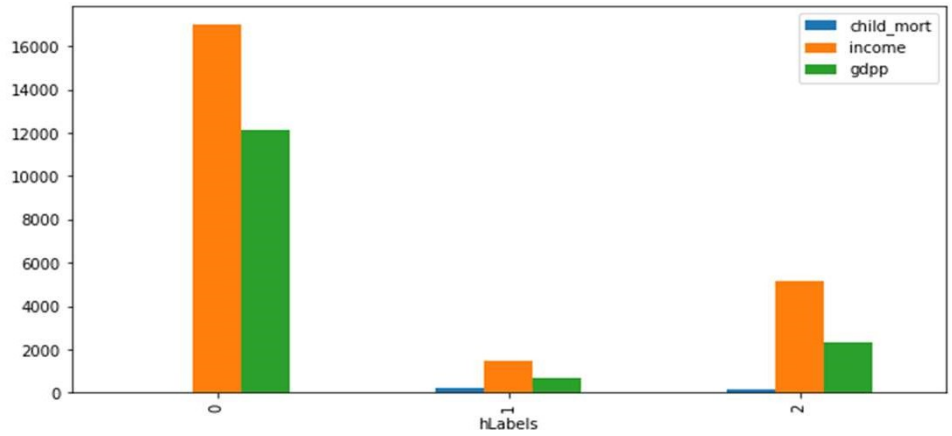


Observations:

- Though `child_mort` seems to be highest for label 1 Country but we won't consider it as there is only one country under that label.
- Similarly, we are also going to consider only label 0 Countries to visualize.
- Similar type of results can be seen in the Scatter plot of the Hierarchical Clustering Model

Hierarchical Model Profiling(Unscaled Data)

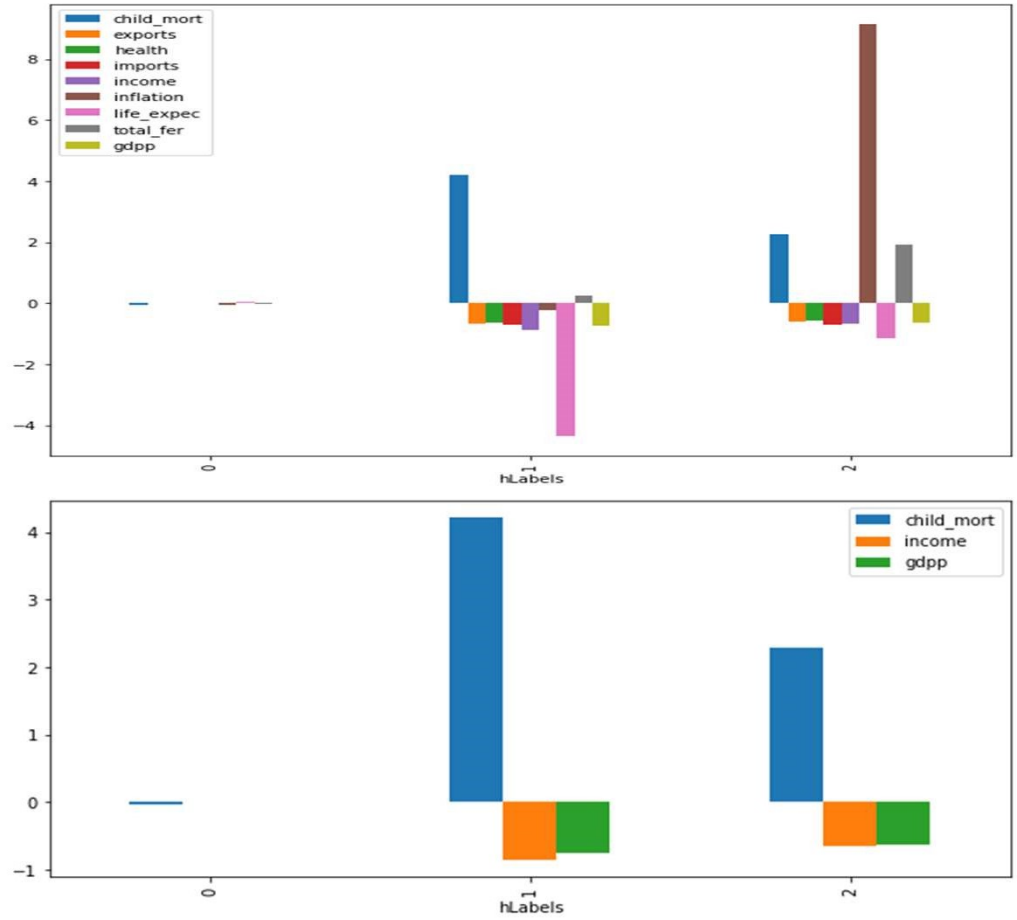




Observations:

- As we can see that due to much higher values of `income` and `gdpp` we are unable to see the bar plot for `child_mort` of label 0. Hence we will try to use scaled dataframe of same data and plot the same bar plot.

Hierarchical Model Profiling(Scaled Data)



Observations:

- In Hierarchical clustering Bar Plotting we can see that the label 1 has higher `child_mort` rate, lower `income` and `gdpp` which suggest that the Countries under Label 1 should be considered for direst Aid on extreme urgent basis.
- But there is a catch in this as there are only one values under label 1 and 2 we cannot choose them on the basis of visualization due to insufficient values labelled under them.
- Hence we will select those countries having higher `child_mort` rate, lower `income` and `gdpp` but in the label 0.

Final Suggestion to the CEO of HELP International NGO

On the basis of the Hierarchical model building I would suggest to provide an urgent AID to countries

- Burundi
- Liberia
- Congo, Dem. Rep.
- Niger
- Sierra Leone