Clustering Assignment
Kaustav Bhattacharjee

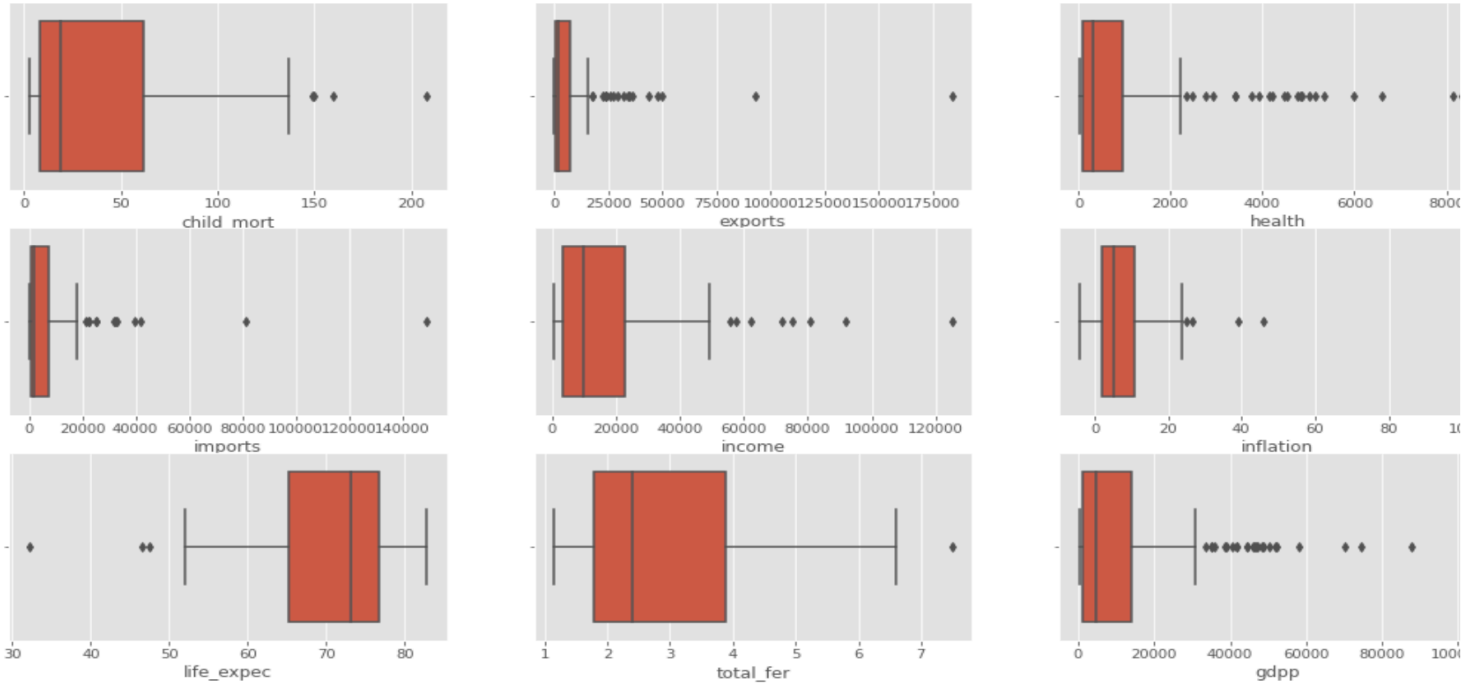
Problem Statement :

Categorise the countries using the given socio-economic and health factors that determine the overall development of the country. Then suggest the countries which needs most attention.

Analysis Approach :

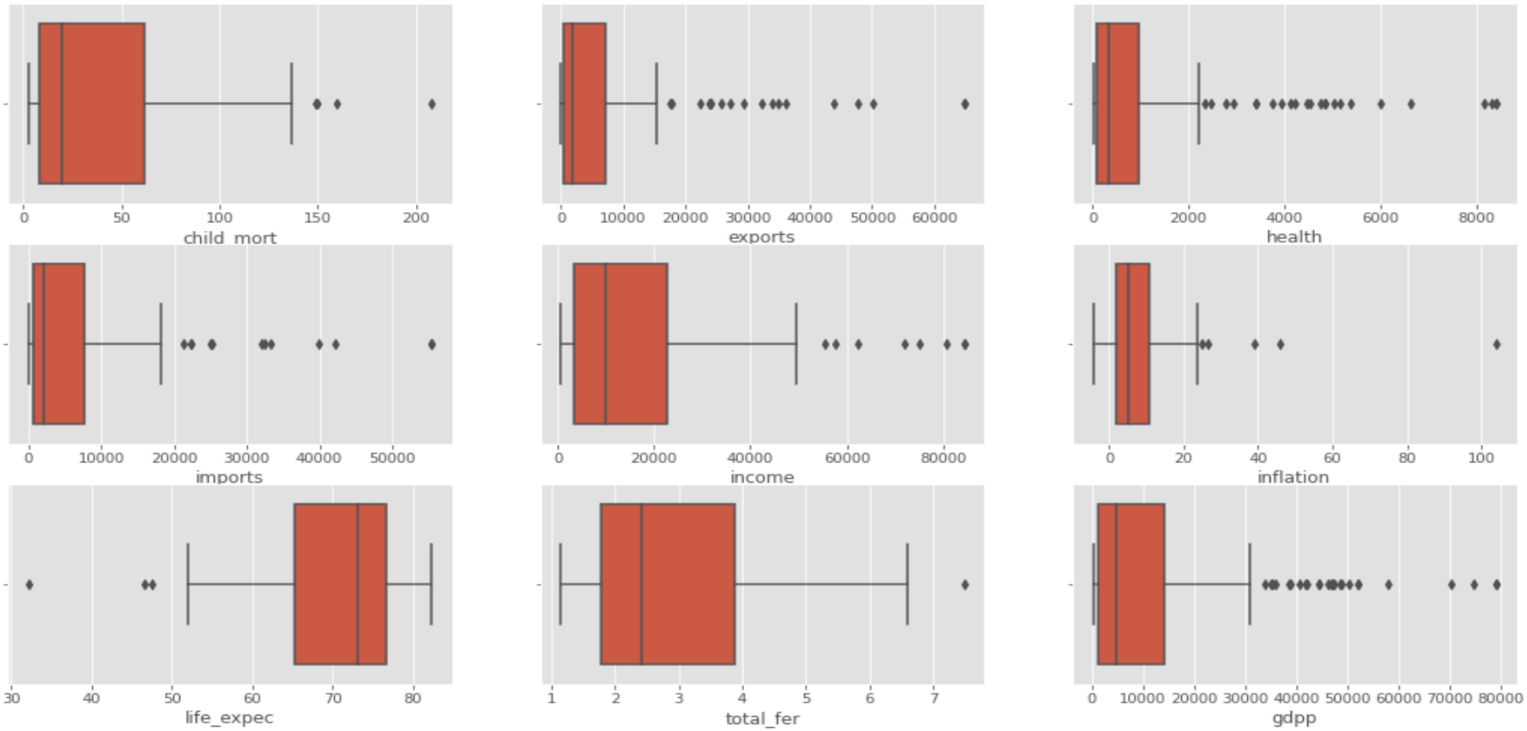
- After going through the data set it is obvious that we have to segregate similar countries and then get the list of countries with a dire need of Aid.
- After applying **Hopkins Statistics,** We can further confirm that the randomness of the data is suitable for Clustering, hence, I am going to use KMeans and Hierarchical Clustering on this dataset to get further details.
- Before going into the clustering we have to pre process the data so that we get more interpretable results in terms of busyness.
- Pre processing of the data is explained in the next slide.

Data Pre Processing :



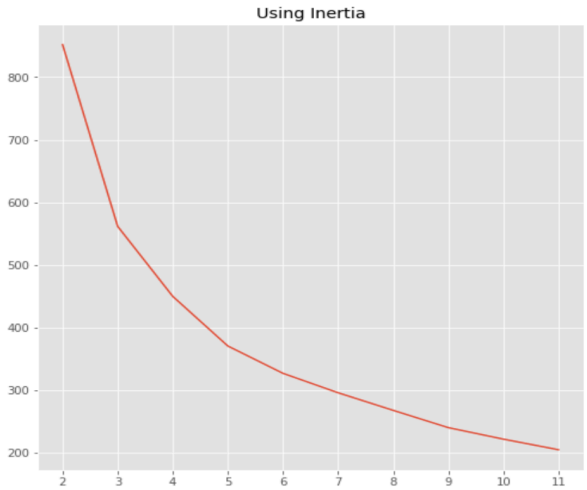
- **From the above box and whisker plot, we can see that there are outliers which might impact our analysis, hence, the outliers needs to be handled before going forward with the analysis.**
- **As our objective is to find out the countries with urgent need of aid, we can cap the variables in such a way that countries with a higher 'exports' value, 'health' conditions, 'imports' value, 'income', and 'gdpp' are capped in such a way that the high values from these variables doesn't effect our analysis. And for the variables, 'child_mort' and 'inflation', there are no outliers in the lower range which could actually have unwanted effect on our analysis, so, we do not need to perform any outliers treatment fore these two variable.**

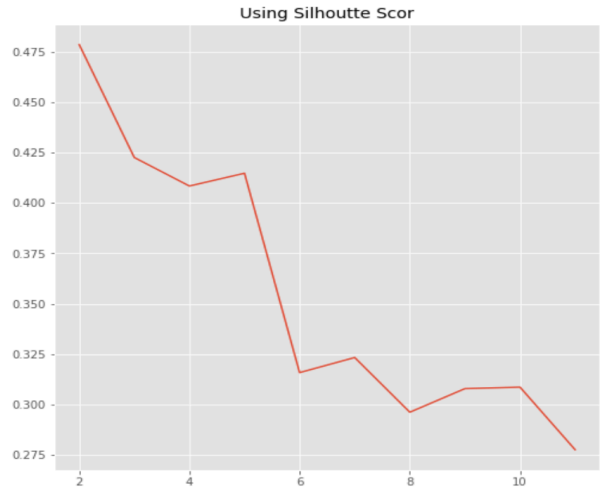
Data Distribution After Preprocessing :



distributions for each of the numeric columns using a box plot after capping outliers at 99 percentile for the variables mentioned in the previous slide.

Determining Number of Clusters Using Inertia and Silhouette Score and a Brief Description of The Clustering Methodology:



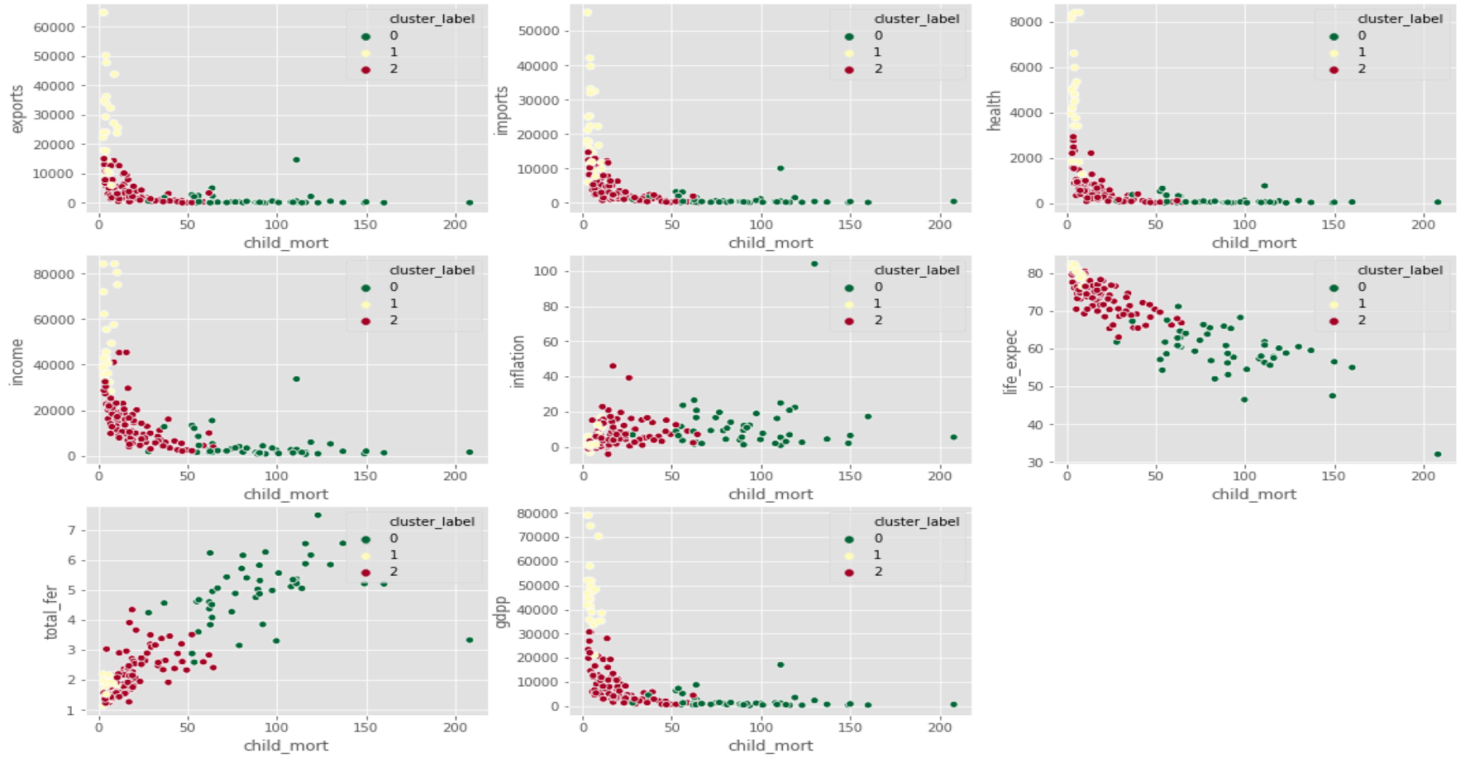


- **From the above charts, we can see that going with *K = 3* clusters would be a good option for us for this dataset.**

1. Taking the graph above into consideration, I am going with 3 clusters here, I will first use KMeans clustering, which will use the Euclidian distance between datapoints from the scaled data set, and the KMeans++ initialization methods to calculate the centroids and then reiterate the process until the centroids stop converging any more.

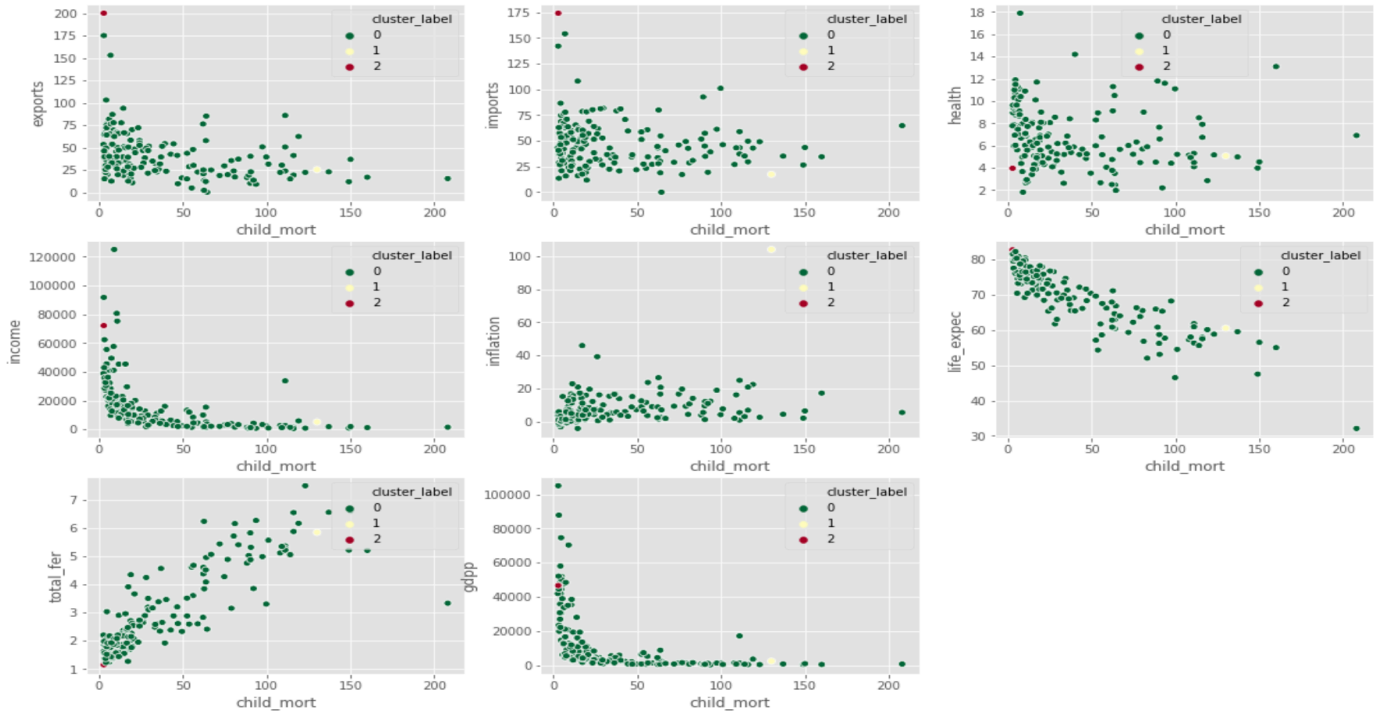
2. After that I will use Hierarchical Clustering's Single Linkage Method and Complete linkage method to create two different Dendrograms, using the Agglomerative method of Hierarchical Clustering. And then, using the cut tree method we will create K=3 Clusters.

Finding Out the Cluster With Countries which are in Need of Aid Using Kmeans Clustering:



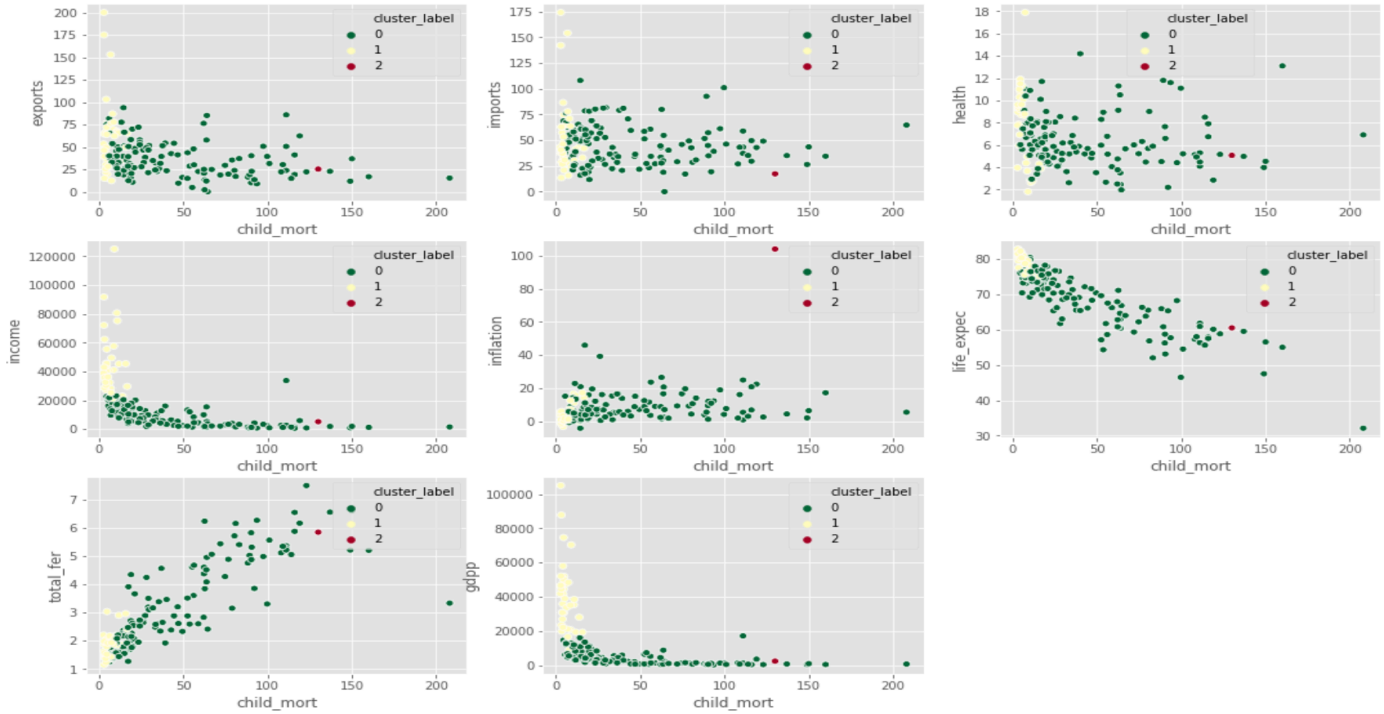
- **From the above scatter plots we can clearly see that cluster *K=0* has the similar Countries which are in more need of aid than that of Countries in the other two clusters.**

Finding Out the Cluster With Countries which are in Need of Aid Using Hierarchical Clustering Using Single Linkage Method:



- **From the above scatter plots we can clearly see that cluster *K=0* has the similar Countries which are in more need of aid than that of Countries in the other two clusters.**

Finding Out the Cluster With Countries which are in Need of Aid Using Hierarchical Clustering Using Complete Linkage Method:



- **From the above scatter plots we can clearly see that cluster *K=0* has the similar Countries which are in more need of aid than that of Countries in the other two clusters.**

Conclusions:

1. **The organization should be focusing on the below mentioned top 10 Countries, as, from all three method used in the analysis, these are the top 10 countries that has been derived.**

2. **<u>List of Countries (In Alphabetical Order):</u>**
    1. **Angola**
    2. **Burkina Faso**
    3. **Central African Republic**
    4. **Chad**
    5. **Congo-Dem.Rep.**
    6. **Guinea-Bissau**
    7. **Haiti**
    8. **Mali**
    9. **Niger**
    10. **Sierra Leone.**