IDENTIFICATION OF COUNTRIES FOR AID

Clustering Assignment

By S.Anil

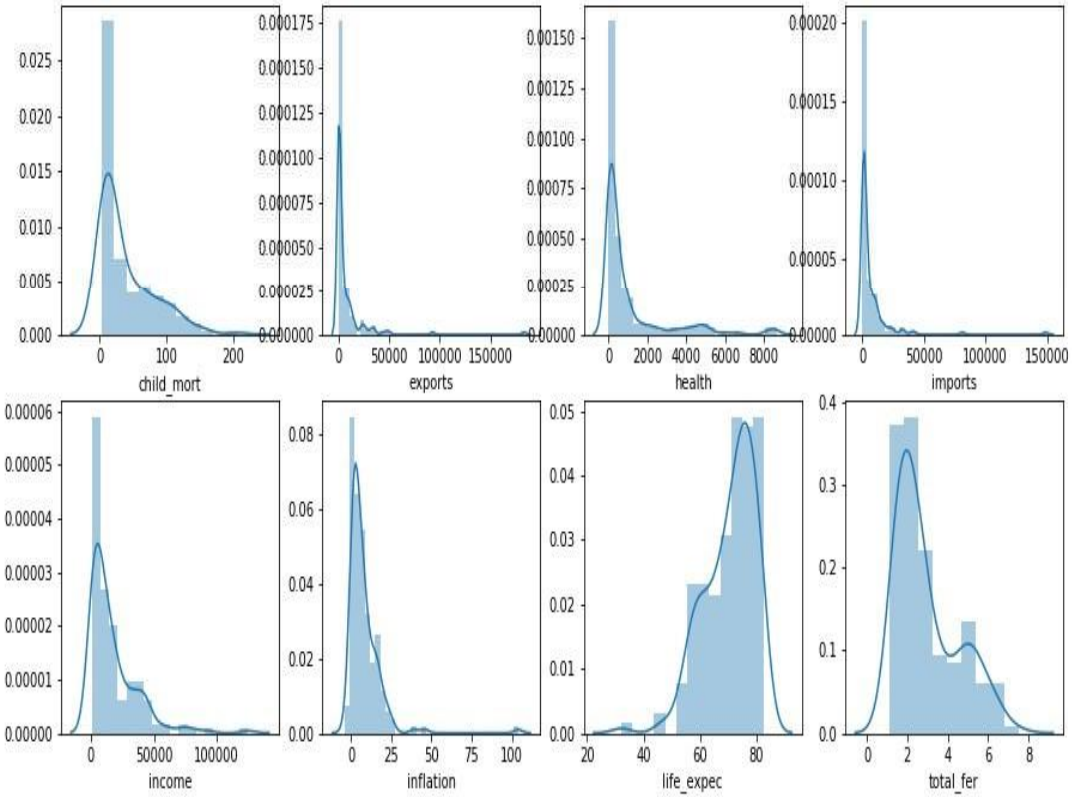
Description of Columns of Data:

- Country - Name of the country
- child_mort - Death of children under 5 years of age per 1000 live births
- exports - Exports of goods and services per capita. Given as %age of the GDP per capita
- health - Total health spending per capita. Given as %age of GDP per capita
- imports - Imports of goods and services per capita. Given as %age of the GDP per capita
- Income - Net income per person
- Inflation - The measurement of the annual growth rate of the Total GDP
- life_expec - The average number of years a new born child would live if the current mortality patterns are to remain the same
- total_fer - The number of children that would be born to each woman if the current age-fertility rates remain the same.
- gdpp - The GDP per capita. Calculated as the Total GDP divided by the total population.

Approach towards the Analysis :

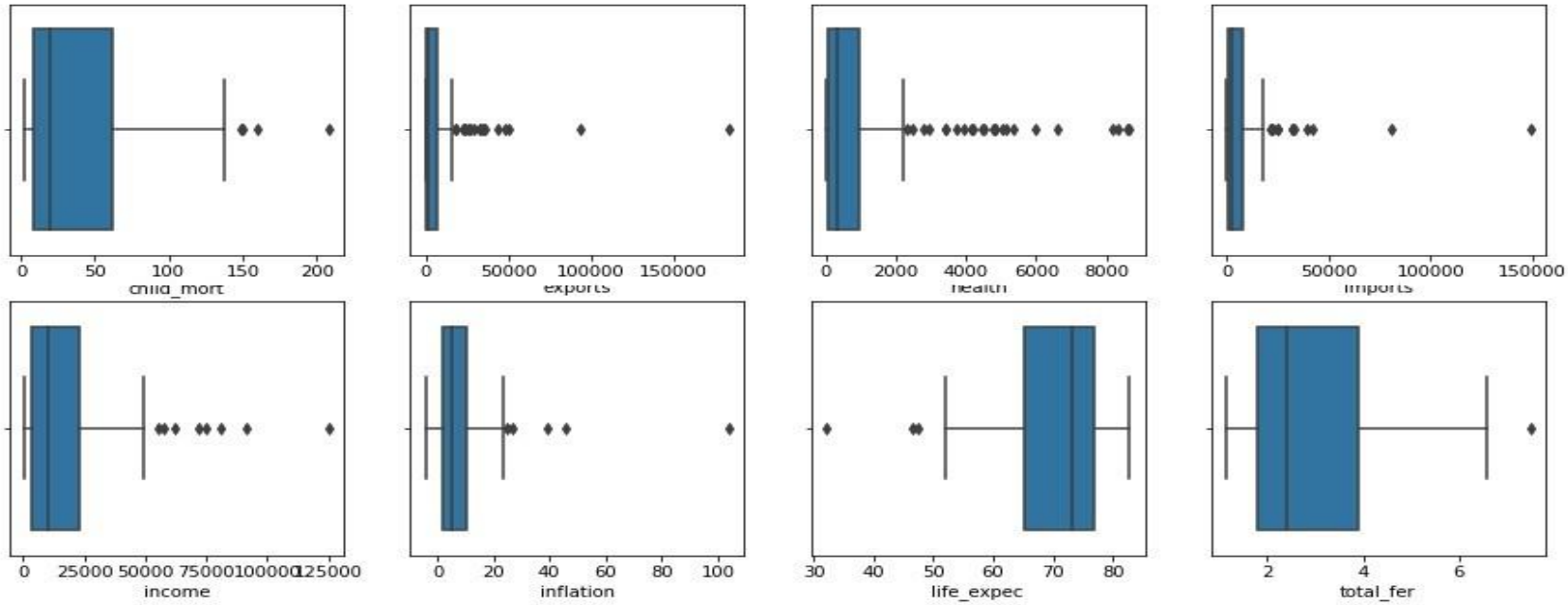
1. EDA
2. Outlier Treatment
3. Scaling
4. Checking the tendency of the data: Hopkins Test
5. Checking the best value for K: SSD, Silhouette method
6. Perform K-Means with the final value of k
7. Visualize the clusters using scatter plot
8. Perform Cluster Profiling: GDPP, CHILD_MORT, INCOME
9. Hierarchical Clustering
10. Single Linkage & Complete Linkage: Dendogram
11. Use the suitable method and perform the final cut
12. Visualize the cluster
13. Perform Cluster Profiling: GDPP, CHILD_MORT, INCOME

Distribution of Data:



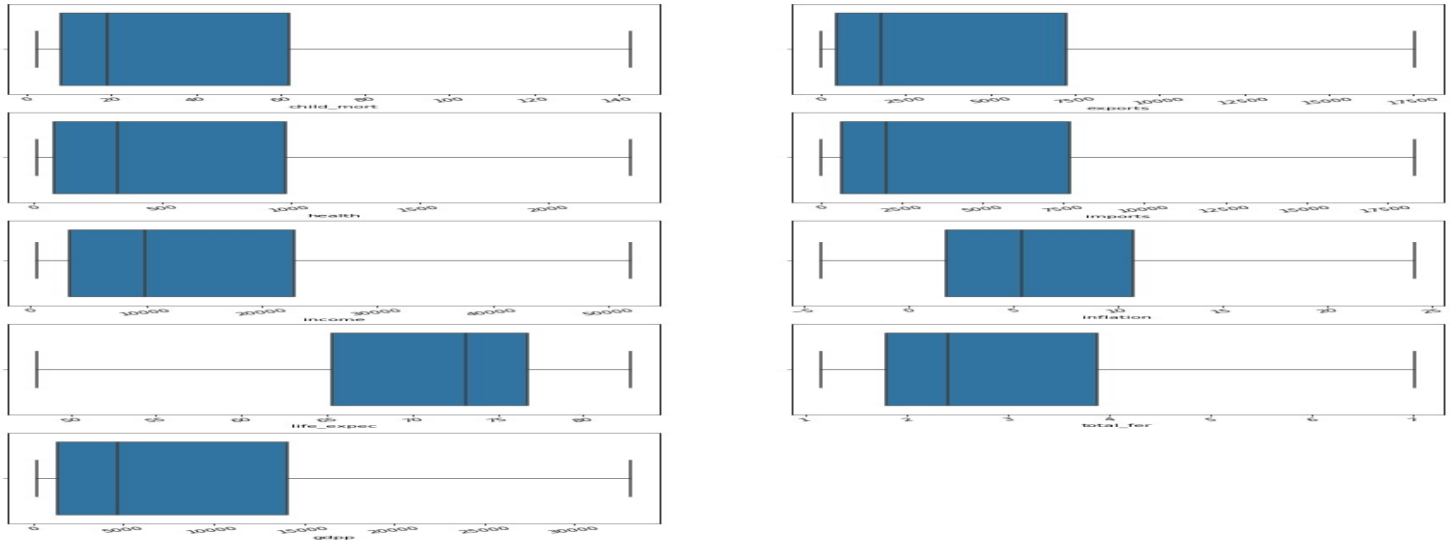
- Distribution of the Countries amongst the various columns is significantly skewed.
- In order to identify the countries the clusters shall be formed on the basis of the characteristics of the distribution and their pattern. (Low, mid and high values)

Outliers :



- **Since all the countries are to be considered, thus we cannot eliminate any of the rows. We will be capping them.**

Data Visualization after Capping :



Now since the data is well with in the ranges. We can proceed with the analysis.

SCALING REQUIRED & HOPKINS TEST

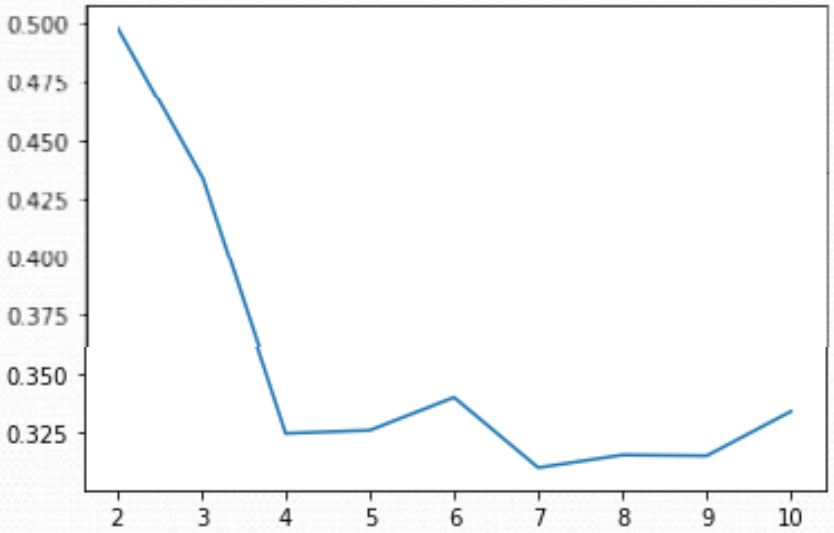
- Standardized Scaling : Duly performed as for clustering and all the attributes of the countries data be on the same scale.

- Hopkins Test :

  As the hopkins value is between $\{0.7, ..., 0.99\}$, data has a high tendency to cluster.

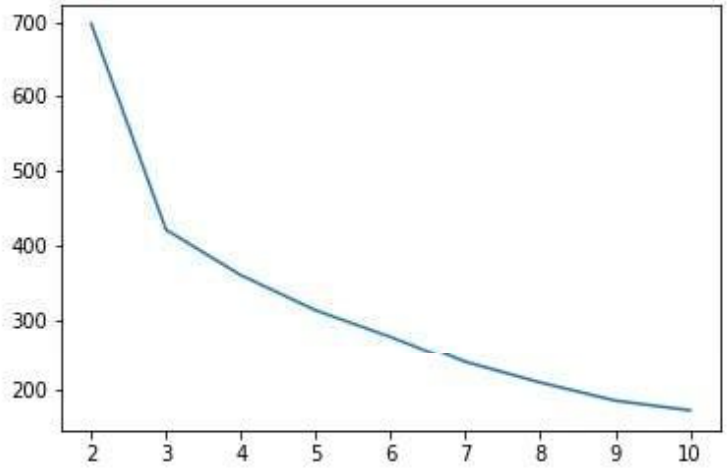
K- Means Clustering :

(X- Axis showing the number of Clusters)



- **Visualization of Silhouette Score :**

Optimum value considered by SSD and silhouette method is 3.

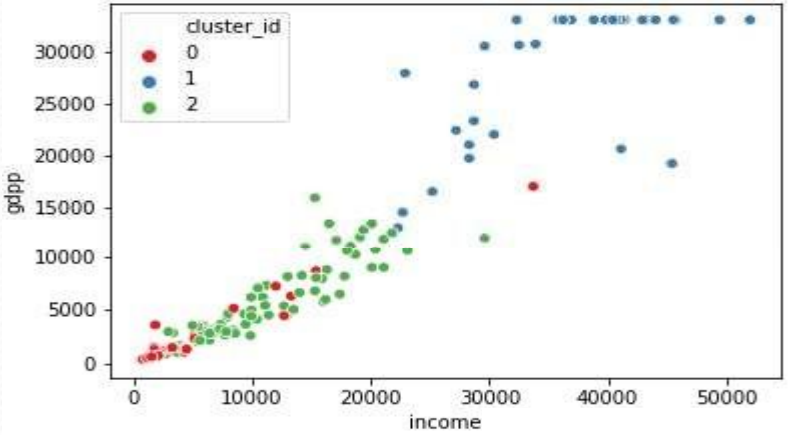


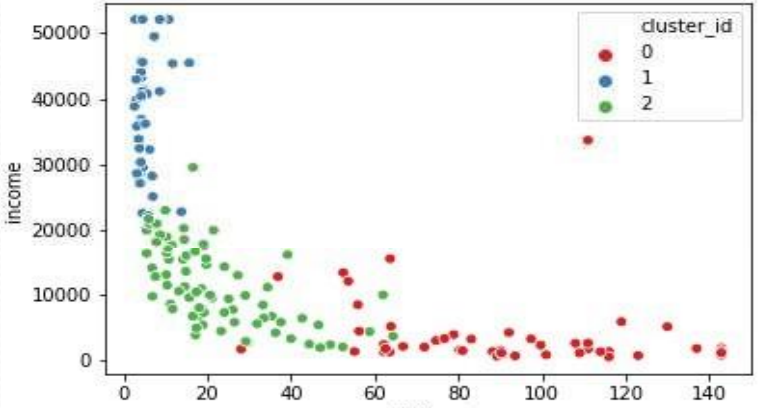
- **Performance Evaluation: Sum of Squared Differences**

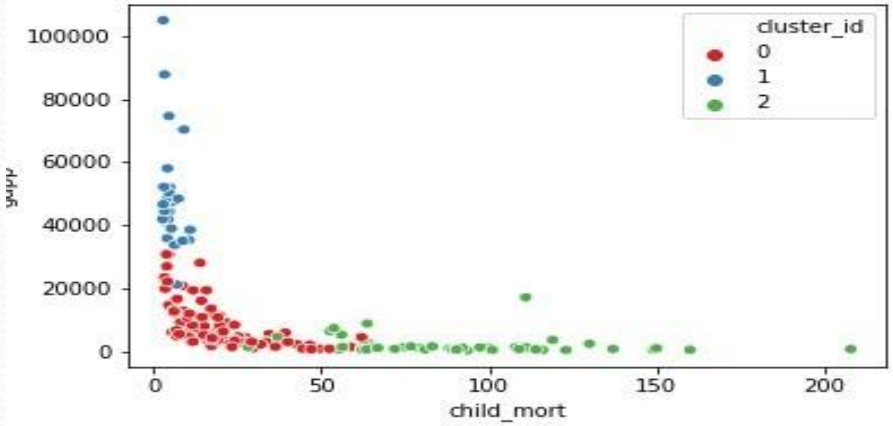
Logically as well the consideration is 3 as countries considered based on their health and econometrics in three categories.

Plotting the Clusters:

Scatter Plots on the basis of Child Mortality, Income and GDPP



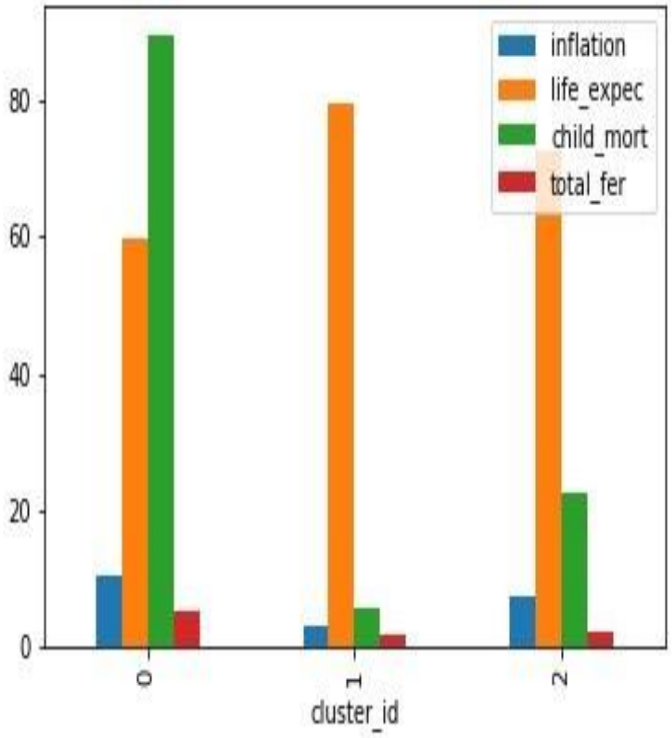


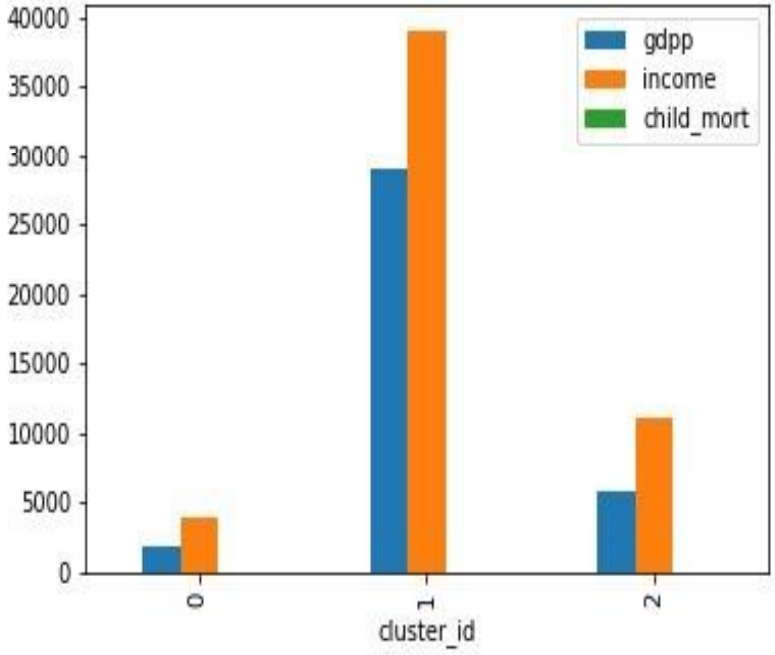


Based on the Cluster Numbers the segmentation on the basis of K-Means methodology Profiles, Following are the derivatives.

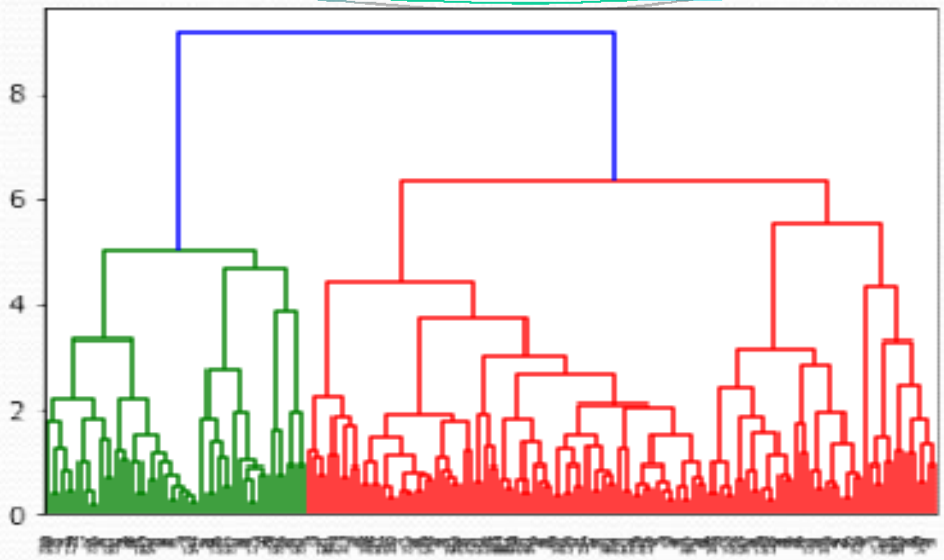
- **Cluster – 0 : UN-DEVELOPED COUNTRIES**
  - Child Mortality – Extremely High
  - Income – Extremely Low
  - GDPP – Extremely Low
- **Cluster – 1 : DEVELOPING COUNTRIES**
  - Child Mortality – Not so High
  - Income – Not so Low
  - GDPP – Not so Low
- **Cluster – 2 : DEVELOPING COUNTRIES**
  - Child Mortality – Low
  - Income – High
  - GDPP – High

Visual Representation of the clusters and their profiles :


inflation
life_expec
child_mort
total_fer
80
60
40
20
0
0
1
2
cluster_id

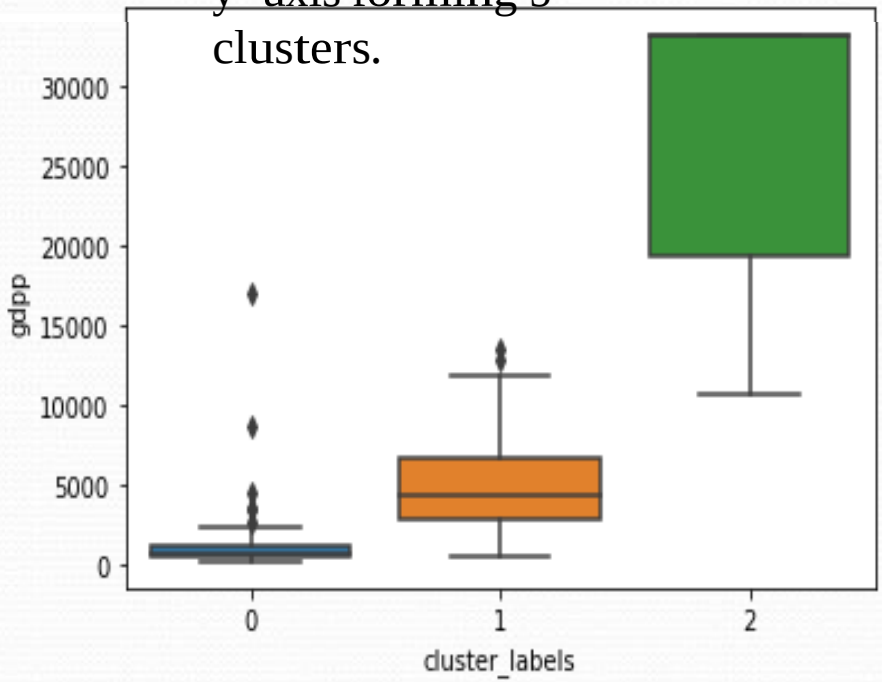

gdpp
income
child_mort
40000
35000
30000
25000
20000
15000
10000
5000
0
0
1
2
cluster_id

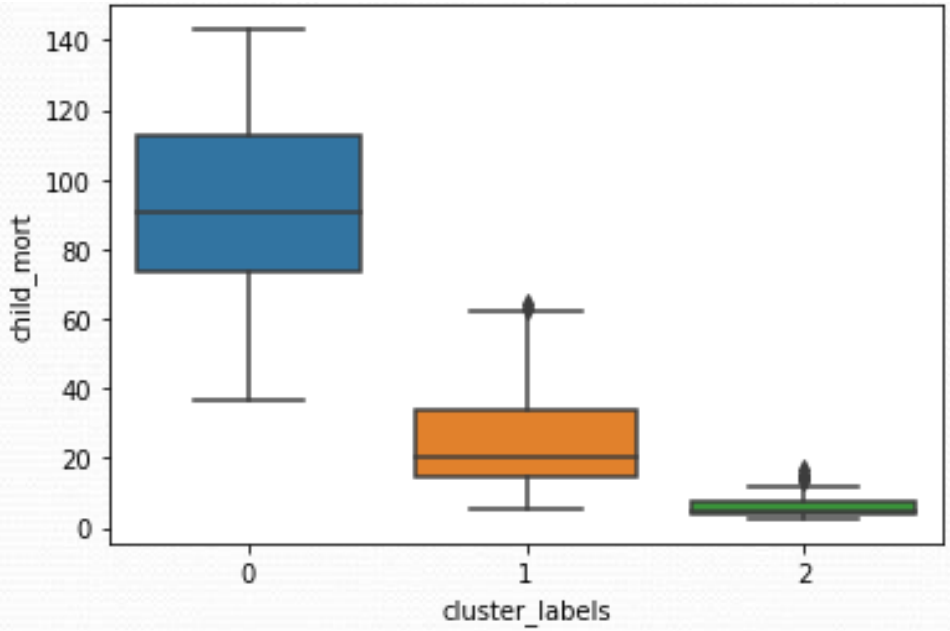
HIERARICHAL CLUSTERING



Complete Linkage –

1. Cutting tree at 6 at the y- axis forming 3 clusters.





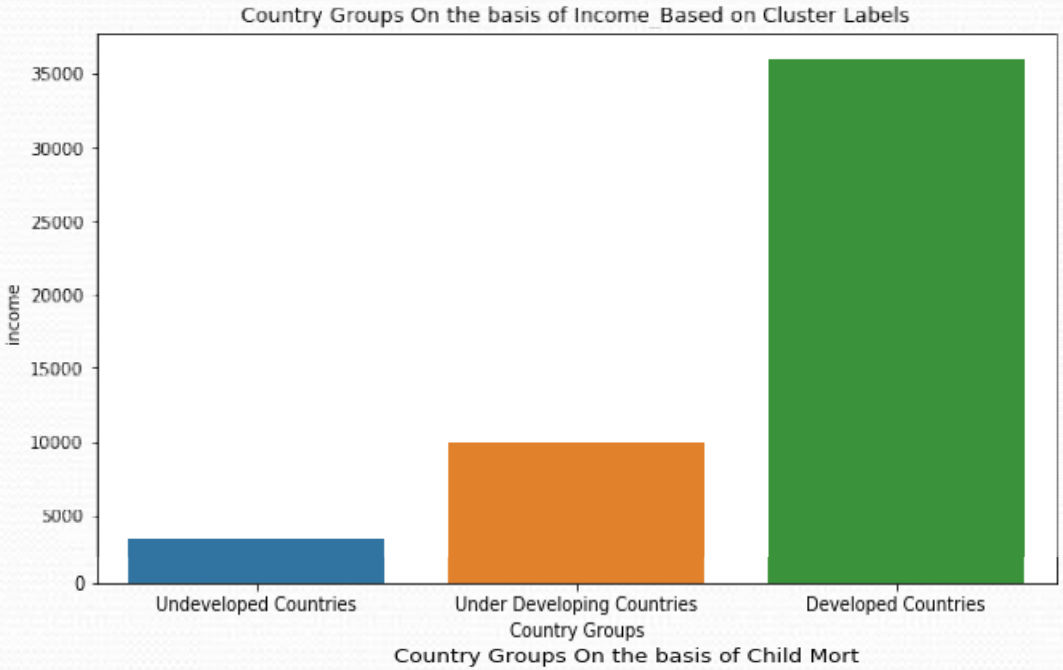
INFERENCE HIERARCHAL CLUSTERING

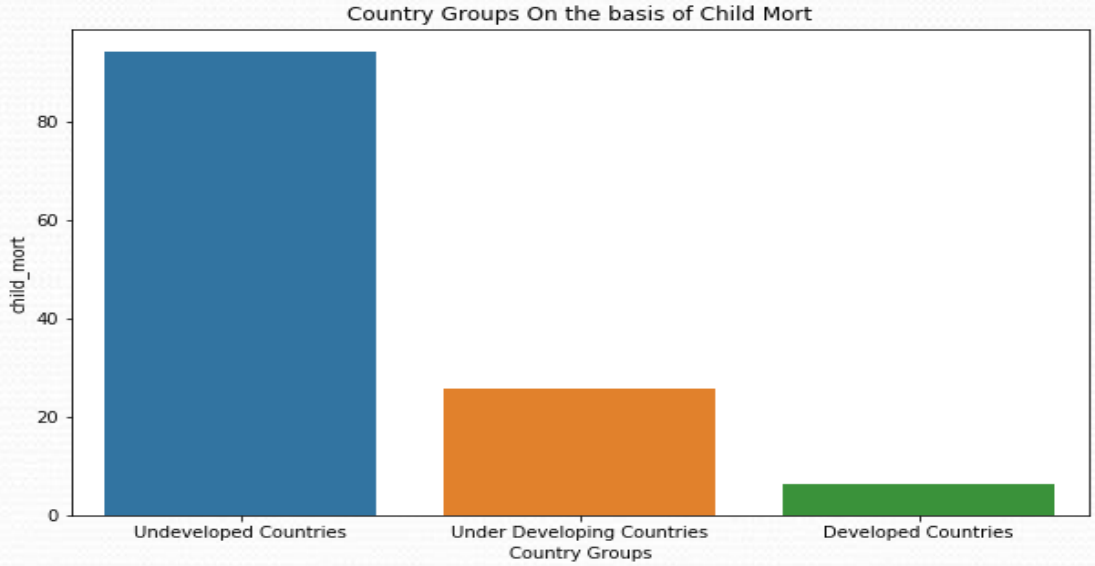
(Based on the mean values of the countries):

Cluster – 0 : Undeveloped Countries (43 Countries)

Cluster – 1 : Under-Developing Countries (49 Countries)

Cluster – 2 : Developed Countries (49 Countries)





CONCLUSION

- On visualizing the clusters on the basis of performing Cluster Profiling: **GDPP, CHILD_MORT, INCOME &** Using both the results, reporting the countries that are in need of the AID: The top 10 Countries which require the aid (Based on the above Cluster Profiling)
- On the basis of Hierarchal Clustering(Reasons for choosing Hierarchal Clustering):

- Eliminating the k Means limitation of predefined consideration of number of clusters.
- Data set is small.
- Slightly more data points considered in Un-developed Countries Segment (Cluster- labels = 0)
-

Final Conclusion :

The top 10 Countries requiring the aid are hereunder :

1. Central African Republic
2. Sierra Leone
3. Haiti
4. Chad
5. Mali
6. Nigeria
7. Niger
8. Angola
9. Congo, Dem. Rep.
10. Burkina Faso