Clustering Assignment

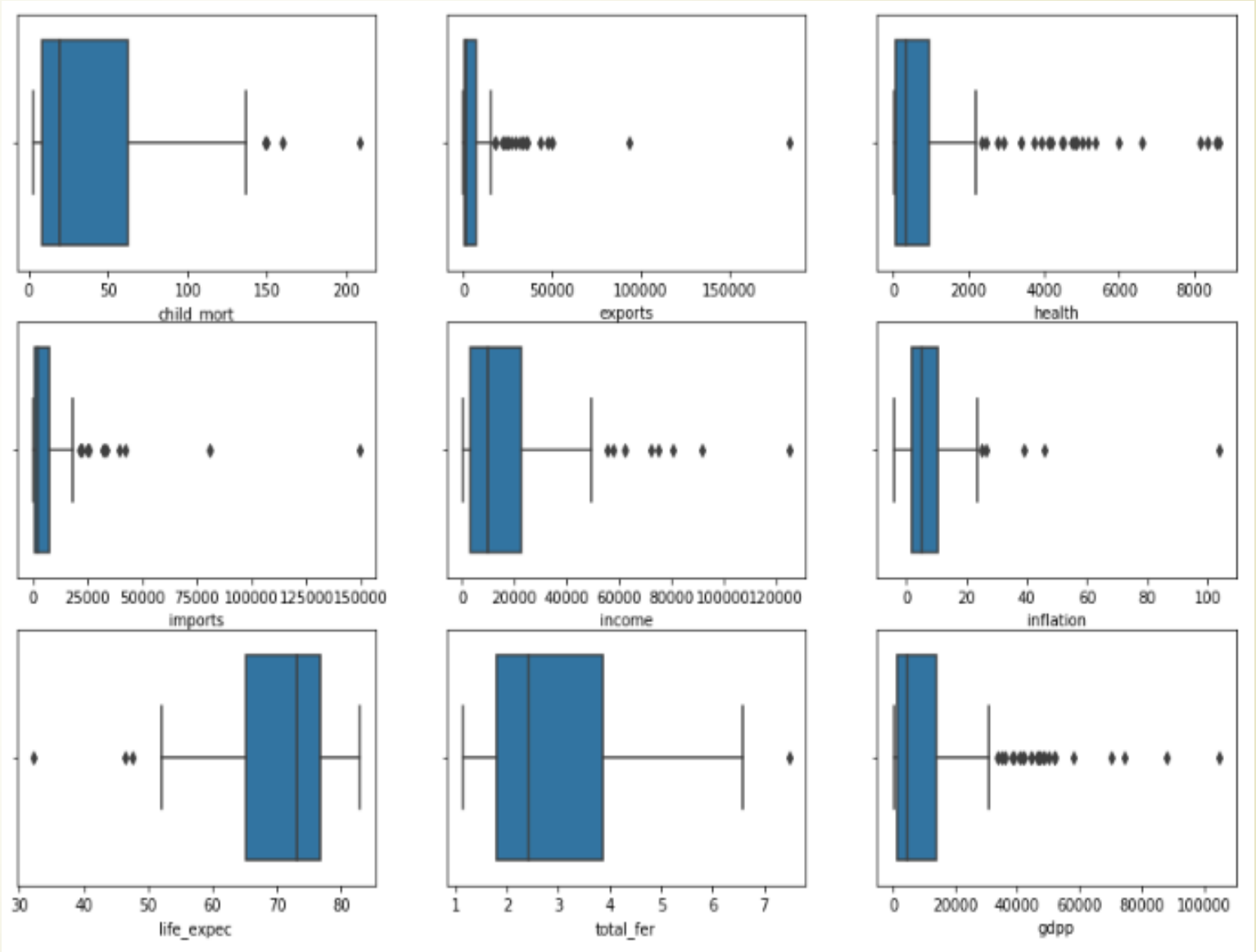
By:- Pawan

Problem Statement

Categorize the countries using socio-economic and health factors that determine the overall development of the country and suggest the countries which needs to be focused on the most.

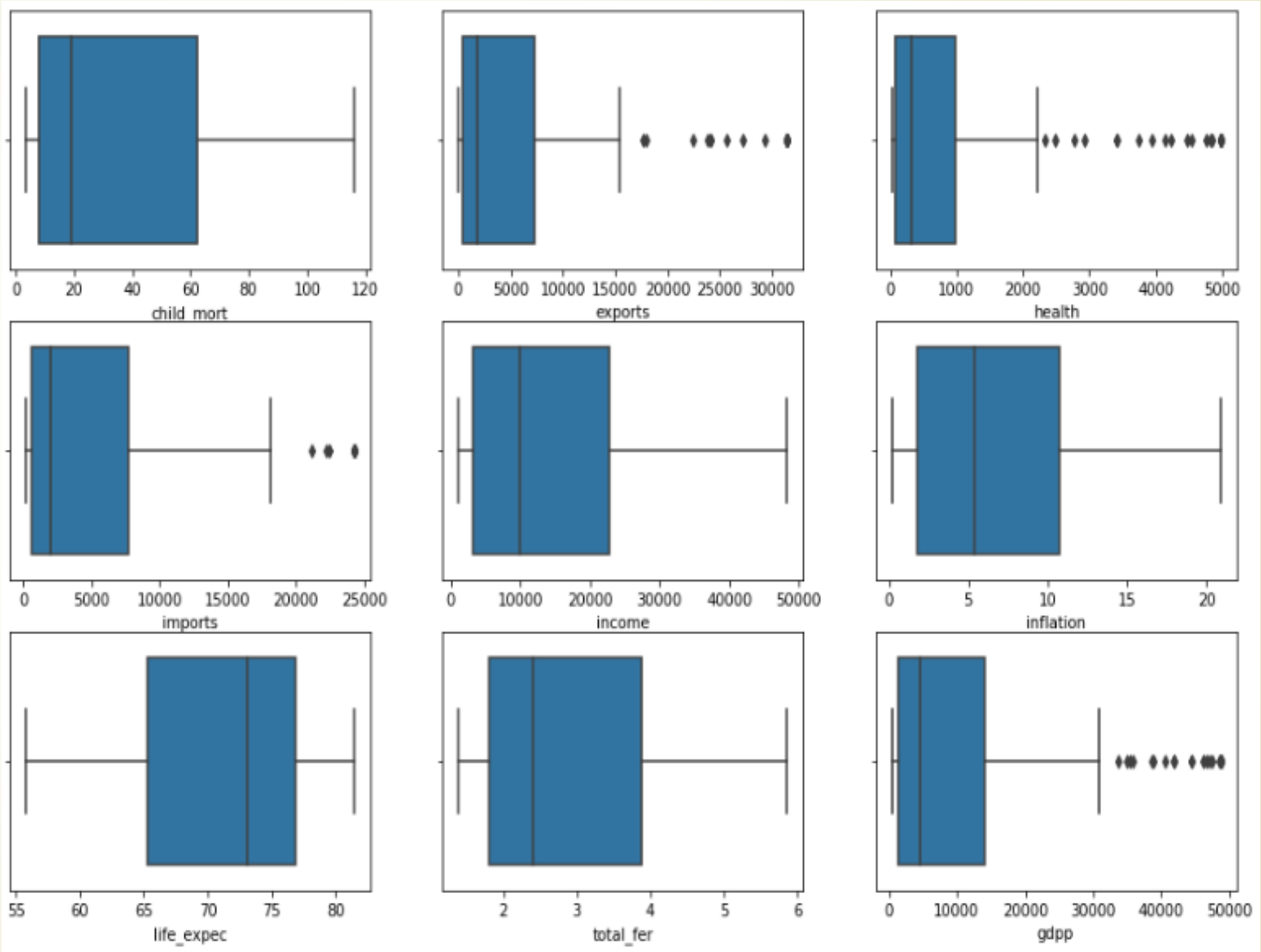
Outlier Analysis

We can clearly see that there are outliers in the data which can influence the model. So we treat the outliers by capping technique as the data is not large.



Outlier Treatment

- As mentioned, we used the capping technique for outlier treatment.
- We used the data <= 0.05 and data >= 0.95 percentile into one group
- As you can observe, most of the outliers have been treated and we can go ahead with this data.
- We can't remove the rest outliers as we might loose important data.

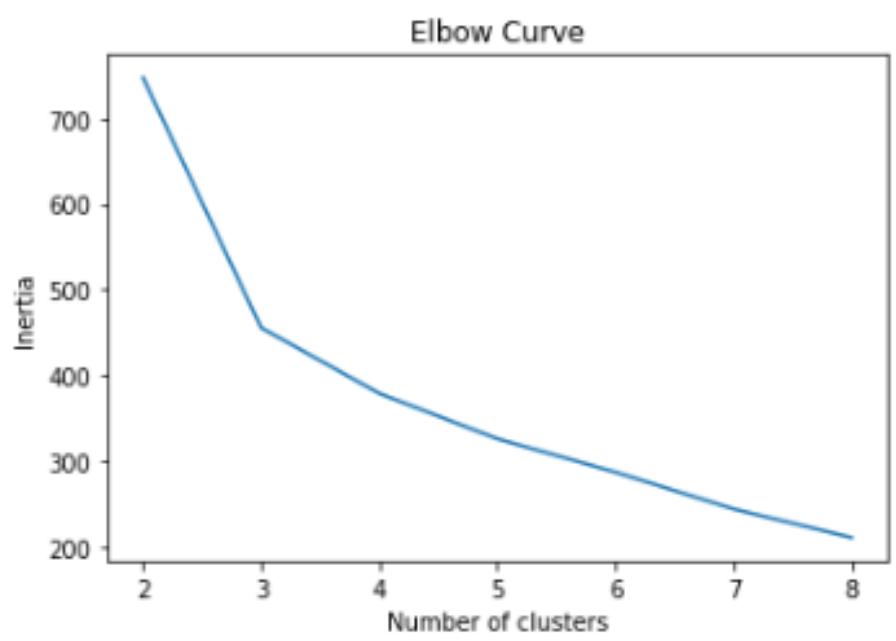


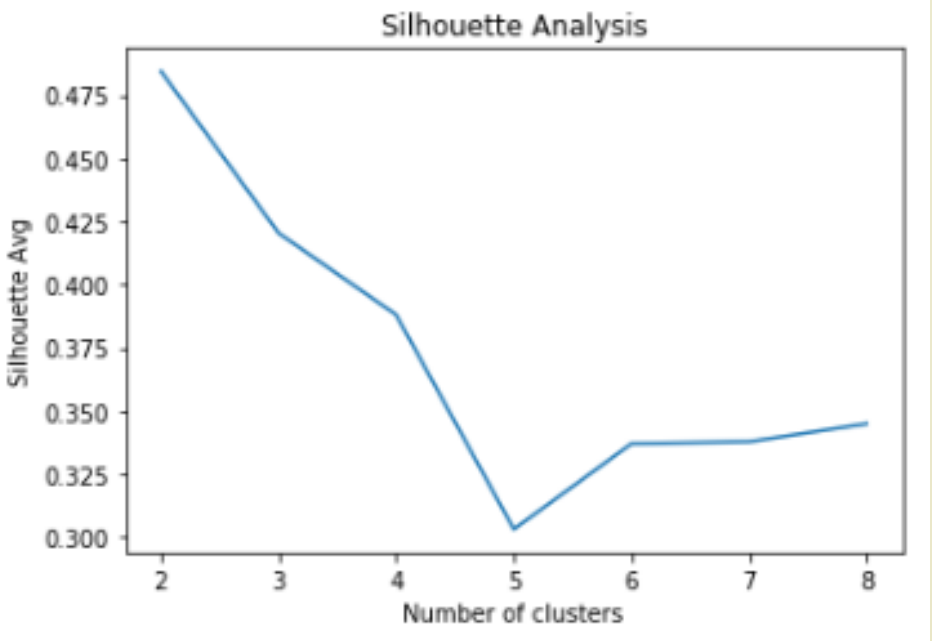
Data Analysis

- We can observe that we have high correlation between few columns such as total fertility with child mortality and life expectancy and so on.
- But we go ahead with the same data.

K-Means Clustering

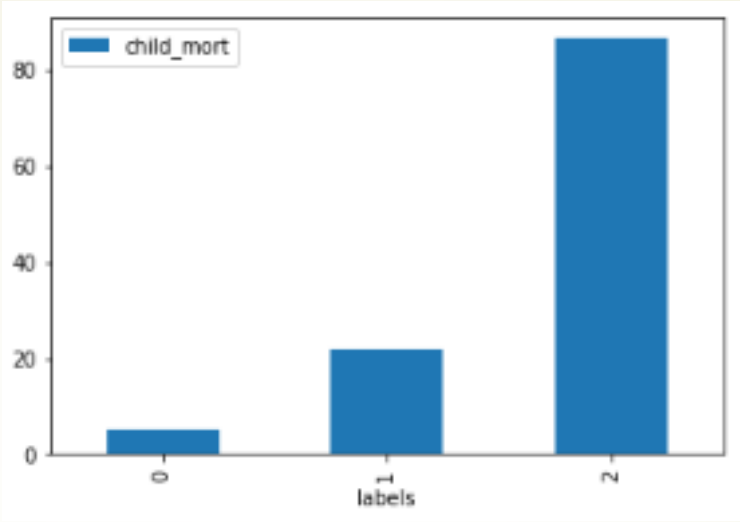
- In order to decide the number of clusters to consider, we performed elbow curve and silhouette analysis.
- The results from both the analysis suggested that 3 was the optimal number of clusters.

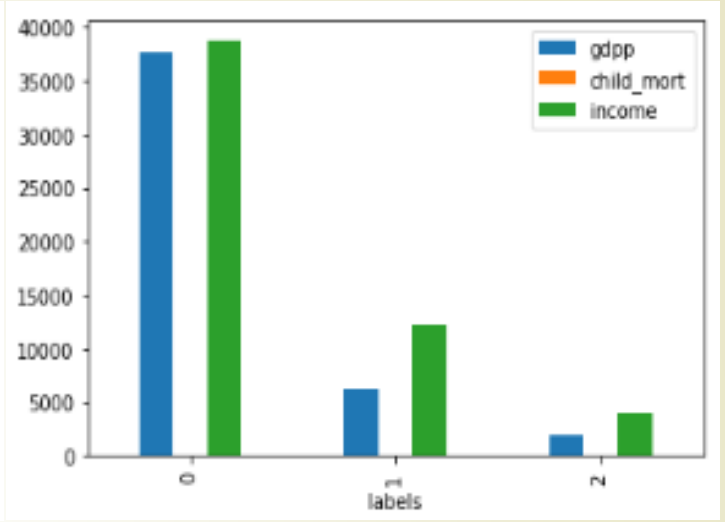


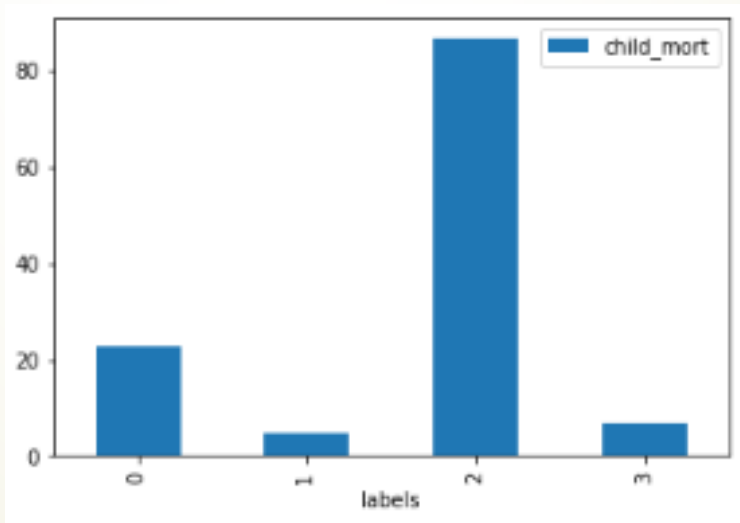


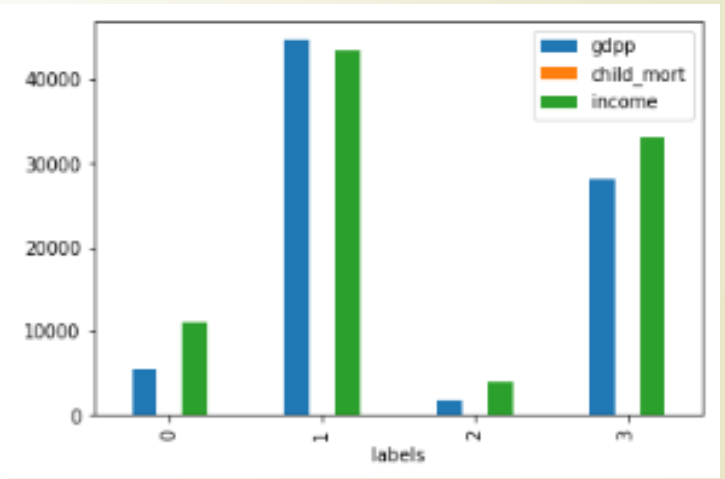
Comparison of different number of clusters for K-Means

- We ran the two iterations for K-Means Clustering providing the values as 3 and 4.
- From the graphical analysis we observe that the clustering with 3 clusters is more appropriate and makes more sense when compared to the 4 clusters.
- Hence we decide to go ahead with 3 clusters.



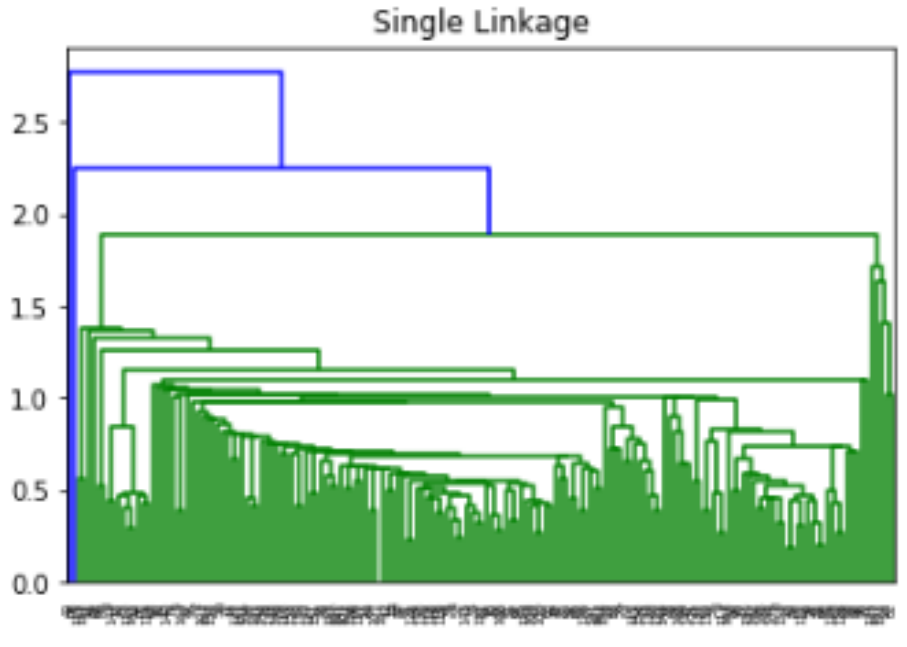


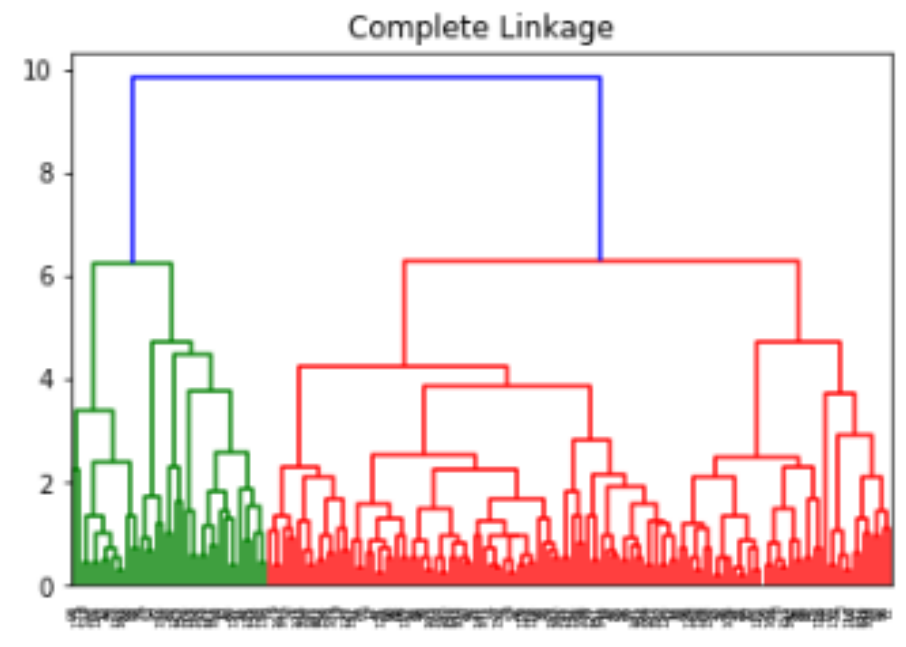




Hierarchical Clustering

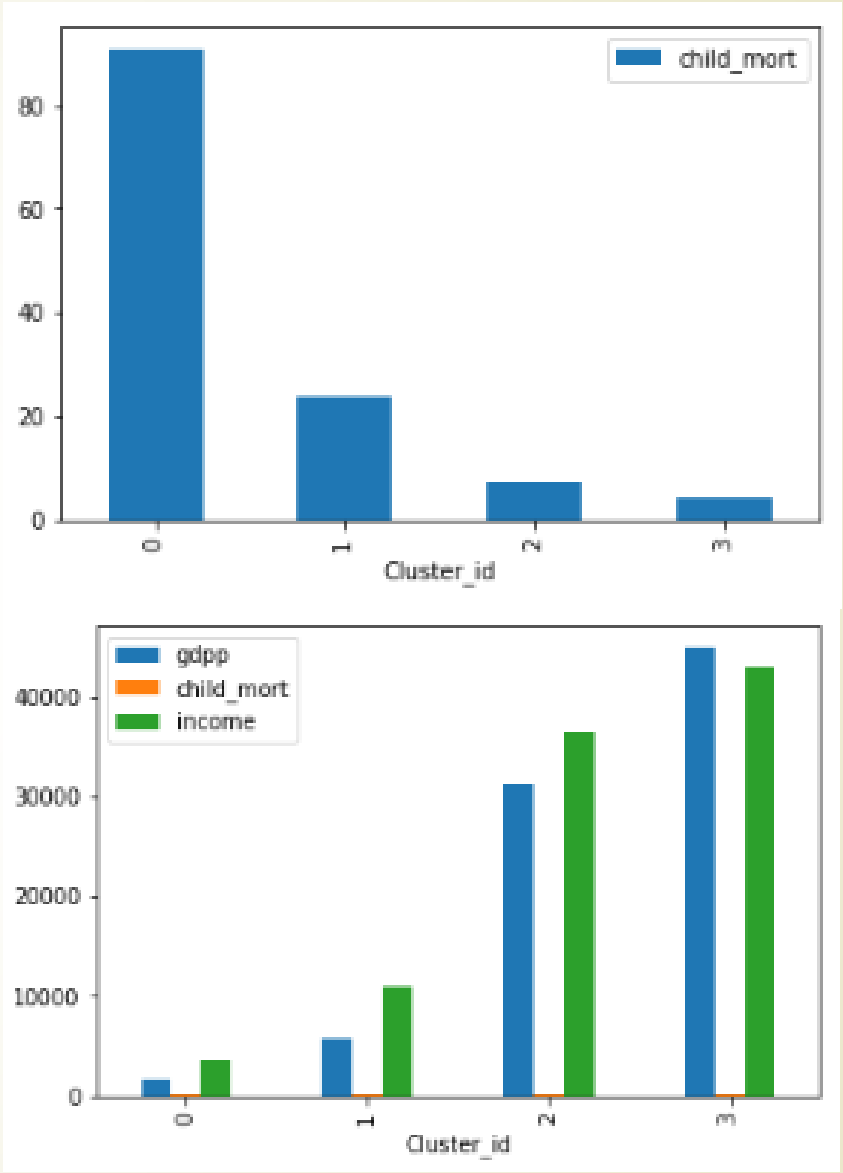
- For hierarchical clustering, we used both single linkage and complete linkage.
- Complete linkage showed a better dendrogram and suggested 4 clusters.





Comparison of K-Means and Hierarchical Clustering

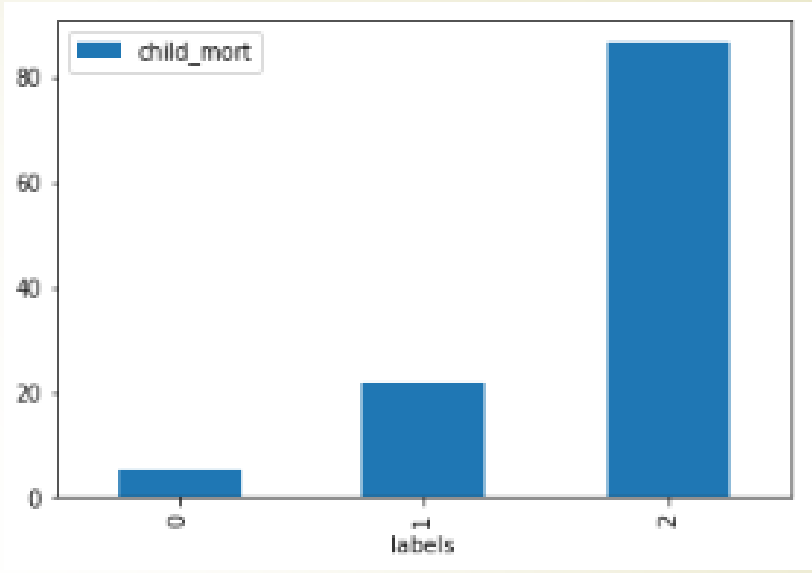
- From Hierarchical Clustering we observe that the number of appropriate clusters is 4.
- After analyzing the clusters we find that it behaves in the same way as K-Means for 4 clusters and hence not that effective.
- Hence we decide to go ahead with K-Means clustering technique with 3 clusters.

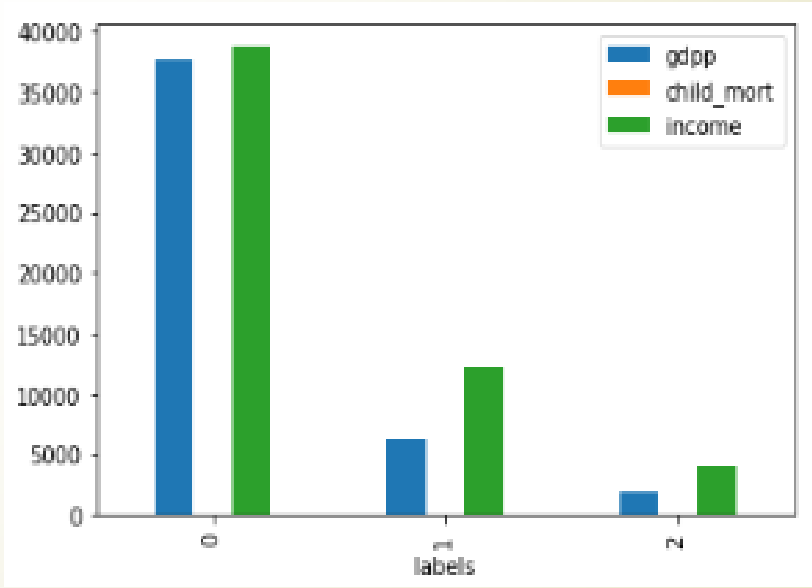


Analyzing the clusters to find the country names.

- Based on the 3 variables – GDPP, Income and Child mortality we need to find the cluster with low GDPP, low Income and high Child mortality rate.

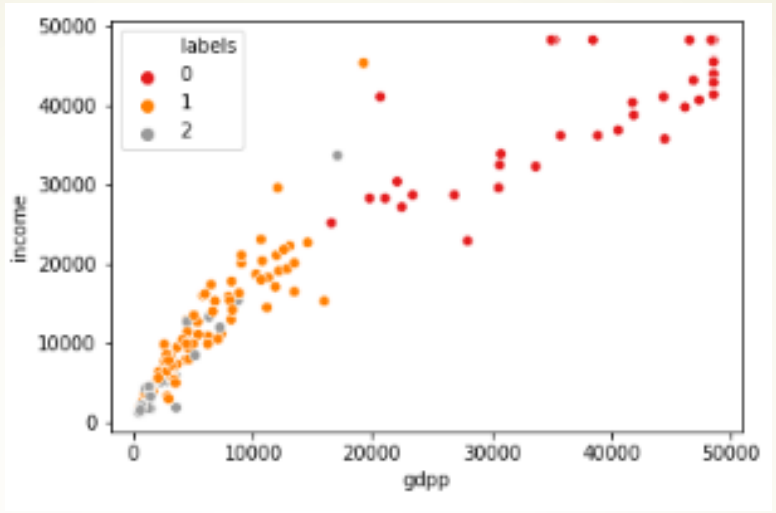
- Hence we find that the cluster 2 meets the requirements and we need to find the top 5 countries in that cluster with low GDPP, low Income and high Child mortality rate.

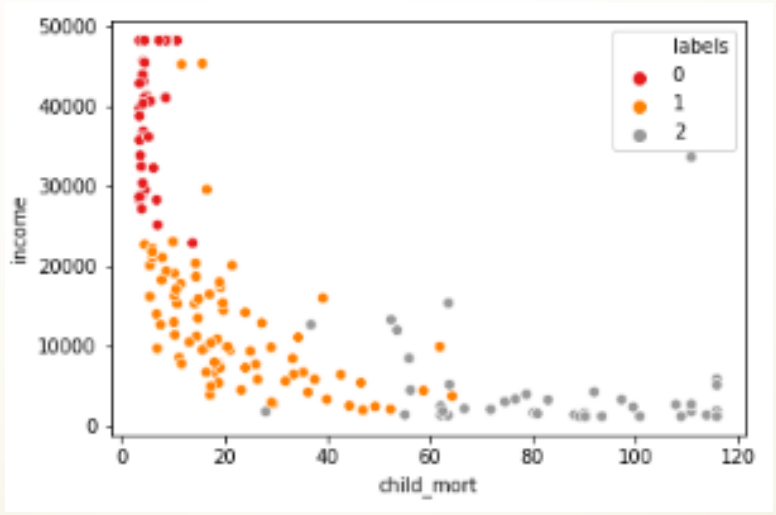


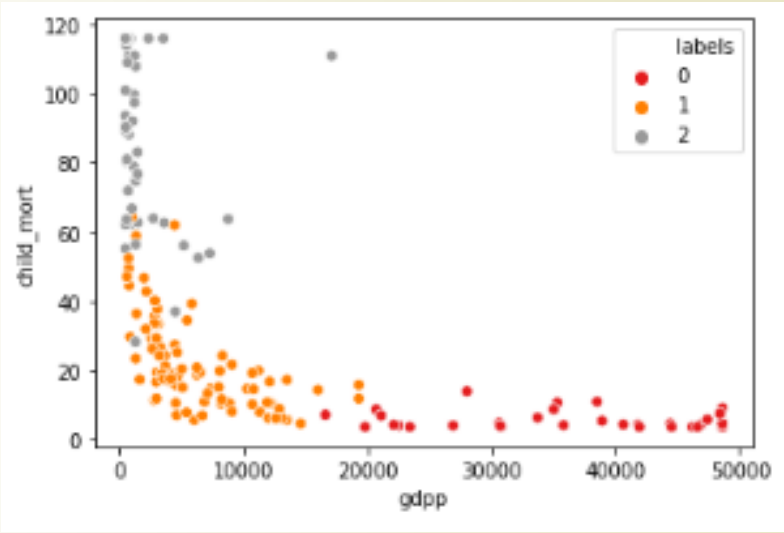


These scatter plots also support the previous analysis

- We find below from these scatter plots:
  - Low GDPP corresponds to low income.
  - Low Income corresponds with high Child mortality rate
  - Low GDPP corresponds to high Child mortality rate.





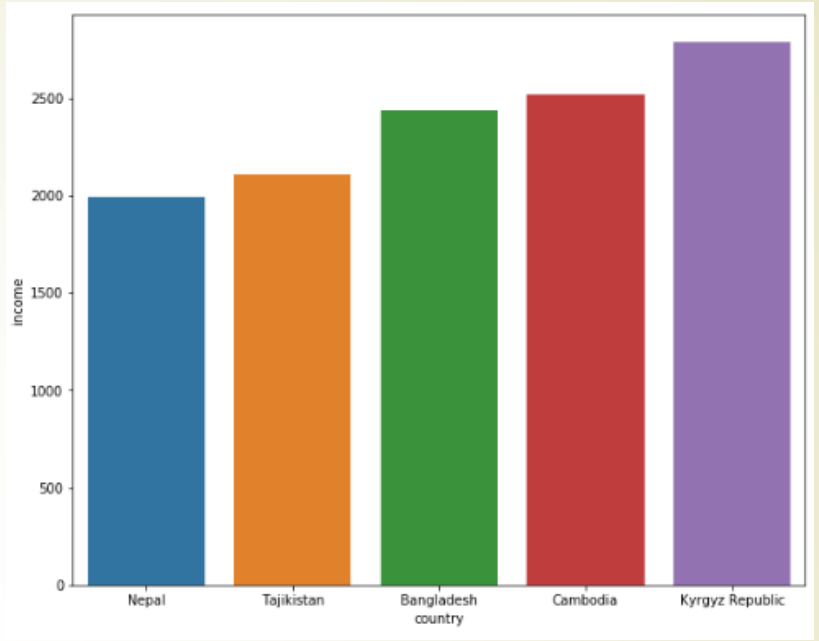


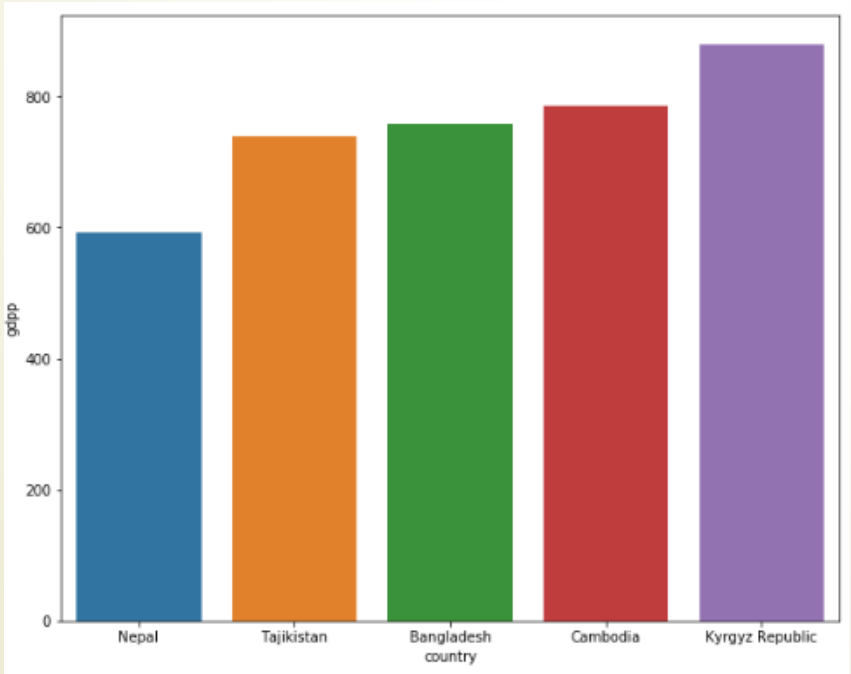
Countries with direst need of aid

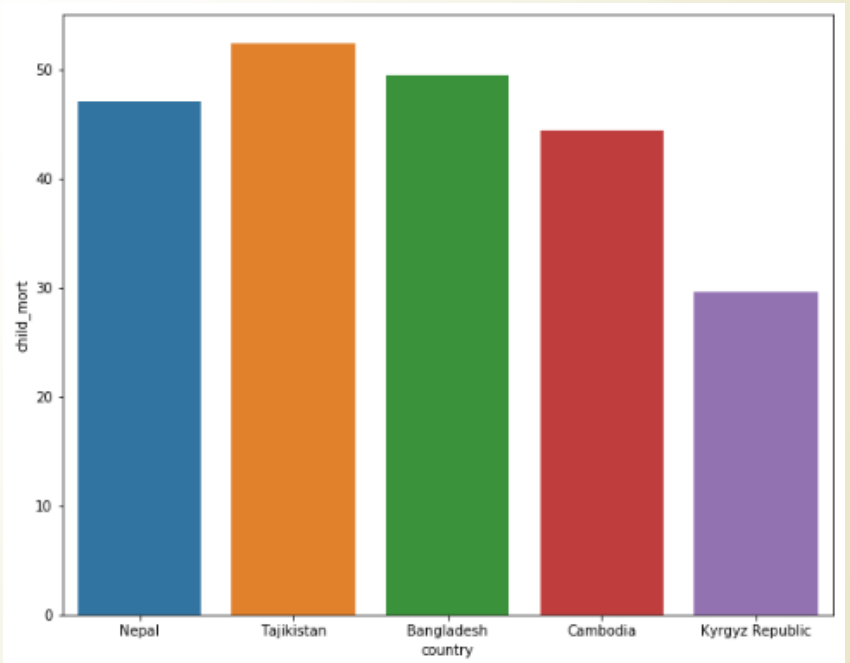
- Based on the analysis we find the below countries which are in direst need of the aid:
  - Nepal
  - Tajikistan
  - Bangladesh
  - Cambodia
  - Kyrgyz Republic

	country	child_mort	income	gdpp
0	Nepal	47.0	1990.0	592.0
1	Tajikistan	52.4	2110.0	738.0
2	Bangladesh	49.4	2440.0	758.0
3	Cambodia	44.4	2520.0	786.0
4	Kyrgyz Republic	29.6	2790.0	880.0

These bar charts also suggest the same







Thank You