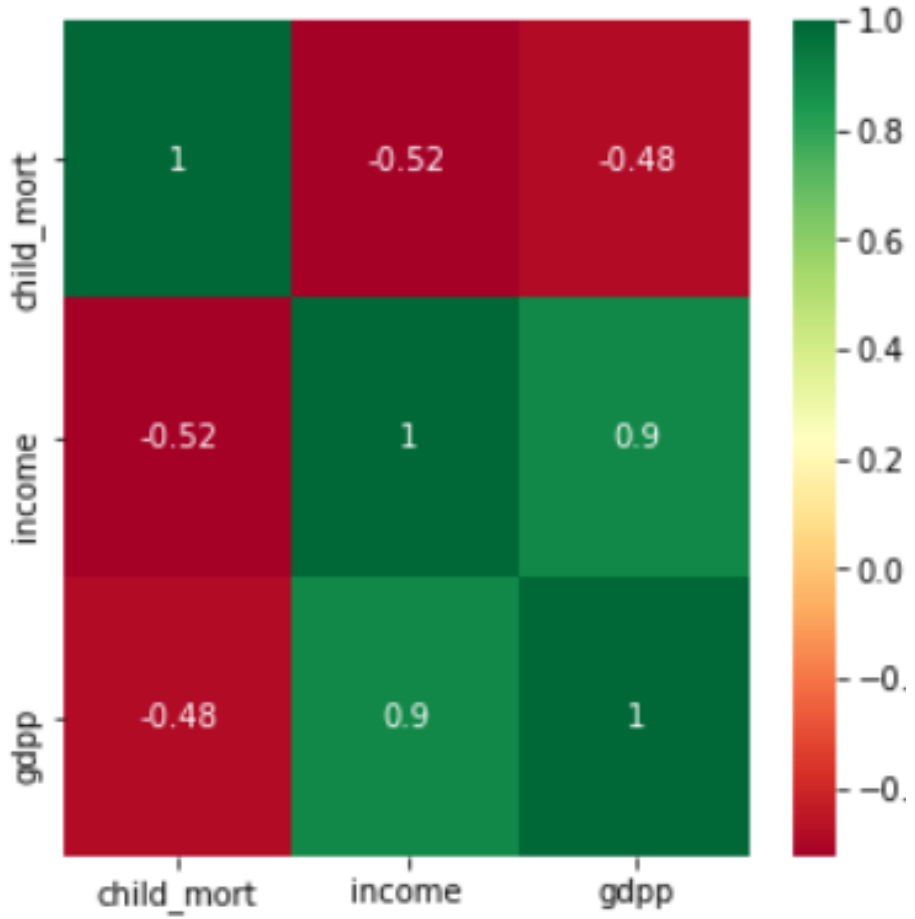
Clustering Assignment presentation to CEO

PROBLEM STATEMENT:

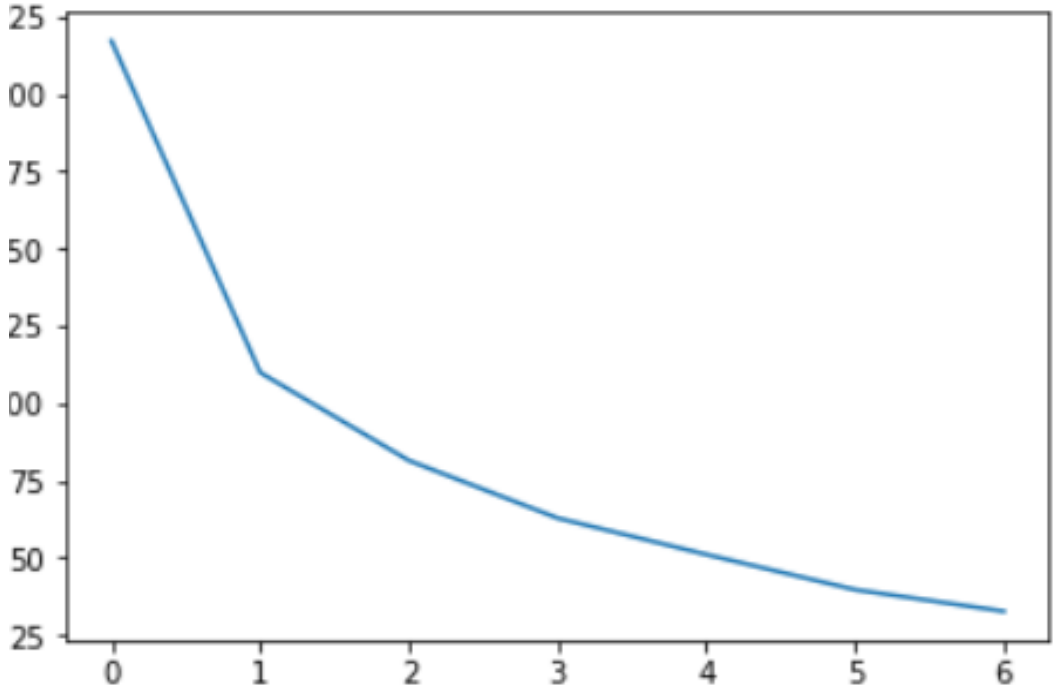
- HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities. It runs a lot of operational projects from time to time along with advocacy drives to raise awareness as well as for funding purposes.
- After the recent funding programmes, they have been able to raise around $ 10 million. Now the CEO of the NGO needs to decide how to use this money strategically and effectively. The significant issues that come while making this decision are mostly related to choosing the countries that are in the direst need of aid.
- And this is where you come in as a data analyst. Your job is to categorise the countries using some socio-economic and health factors that determine the overall development of the country. Then you need to suggest the countries which the CEO needs to focus on the most. The datasets containing those socio-economic factors and the corresponding data dictionary are provided below.

Initial Insights:

- After reading the data, it is observed that there is no missing values and every variable is having correct data type assigned to it. And also checked if there are any duplicates and found out no duplicate rows in the data set.
- After visualizing the target variables using a heatmap, we found out that the child mortality rate is clearly in negative correlation with both income and gross domestic product of a country.

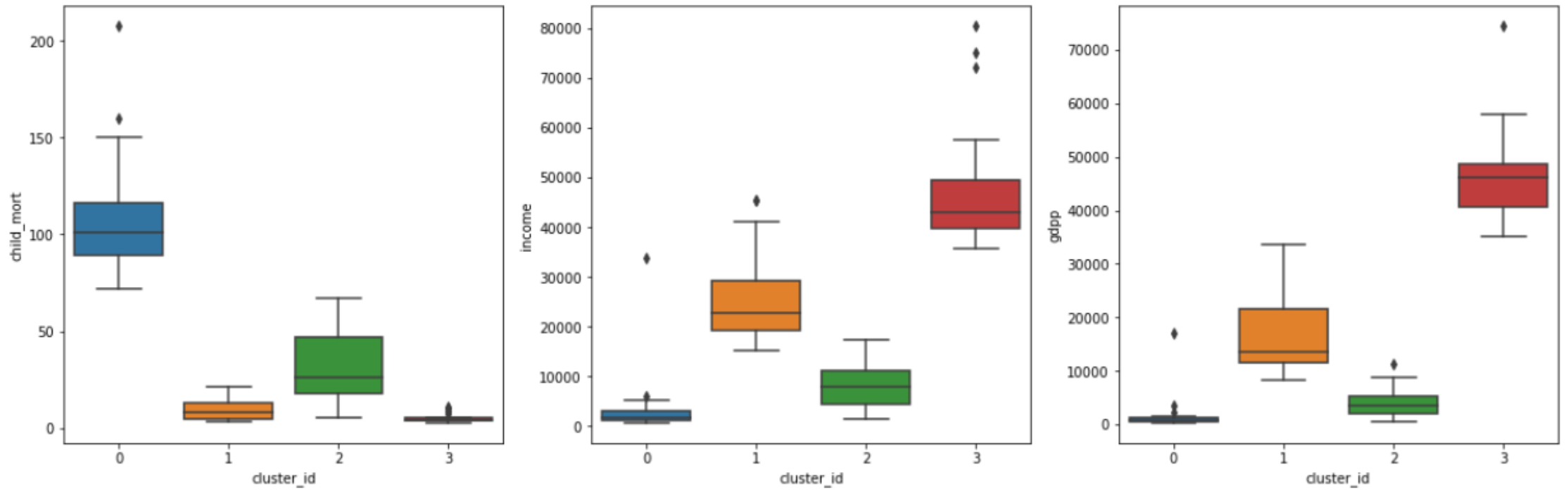


- Observed there are outliers in the target variables and caped upper limits of Income , GDPP. And capped the lower limit of Mortality rates. The restrictions wont effect the decision of which country to concentrate on.
- After Treating Outliers, straight away gone for modelling using K_Means Clustering Algorithm. Did the scaling using standard scalar and fit transform on the data in order to bring all the values around 0.
- Written a code to find the optimal value of clusters and obtained elbow curve from which we will consider the value of clusters.



- Not only using Elbow curve, even with the help of silhouette score, we finally decided the number of cluster = 4 would be optimum.
- After the labelling is done, we observed the trends of mortality rate, income and GDPP using Boxplots.

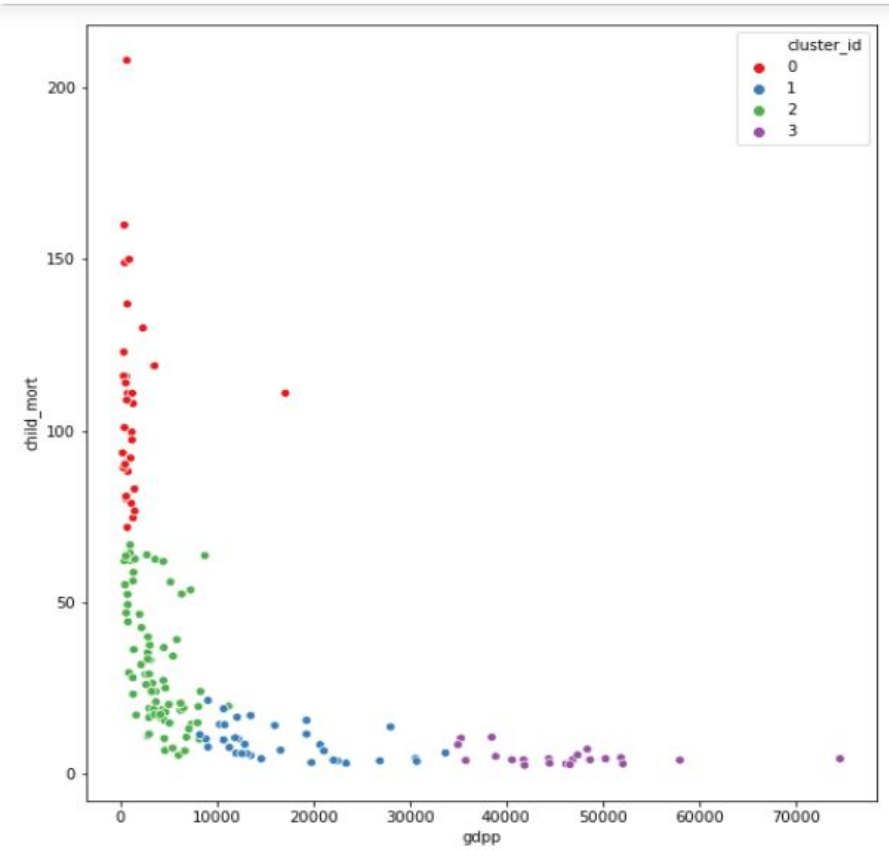
```
plt.show()
```

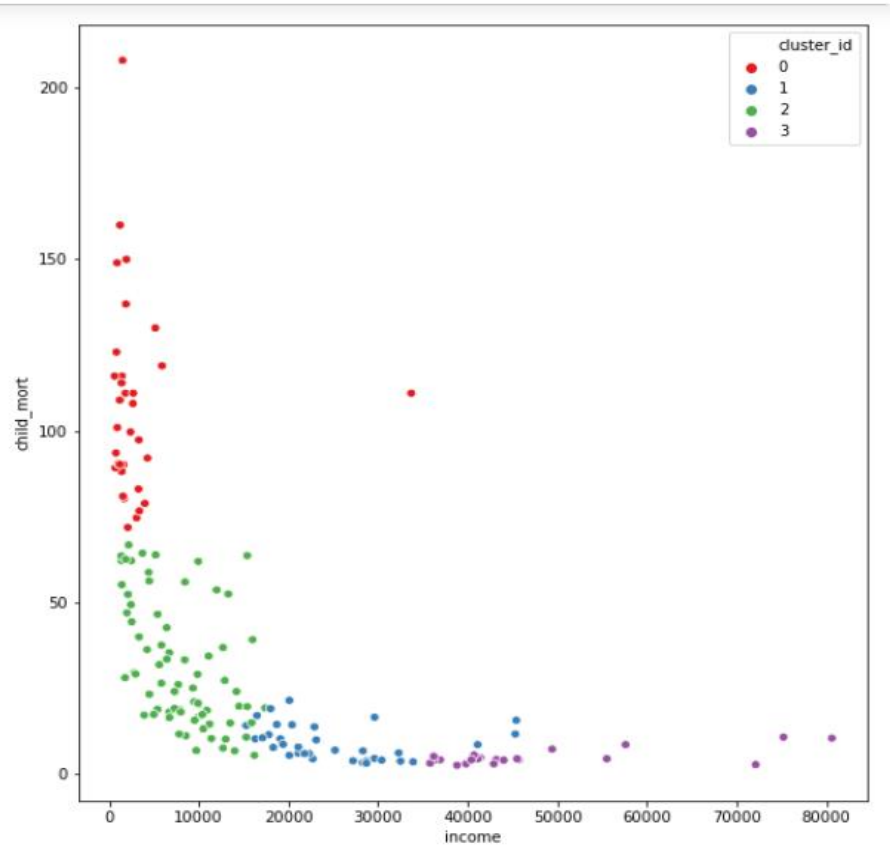


Inferences from BoxPlots:

- From the boxplots, we can infer that the countries in cluster 0 are the most vulnerable and to be concentrated first. We can also observe that the countries in the cluster 0 are having mean child mortality rate of above 100 and their income and GDPs are too low. (K_MEANS Clustering).

Visualization Using Scatter Plot

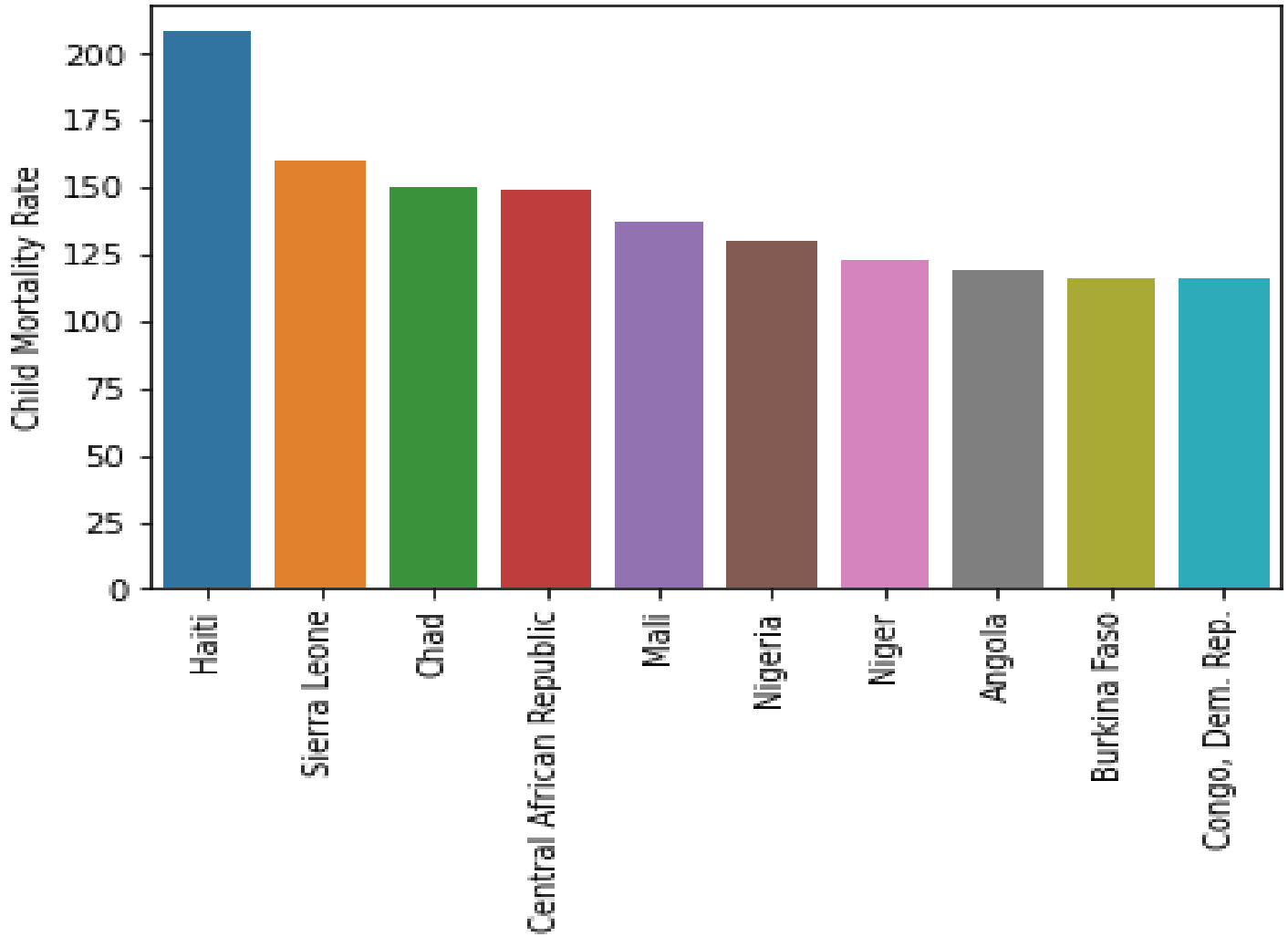




Inferences from Scatter plot

- The Red dots represent the countries with high mortality rate and low income group. So they are the countries to be first concentrated on.

Final plots:


Child Mortality Rate
200
175
150
125
100
75
50
25
0
Haiti
Sierra Leone
Chad
Central African Republic
Mali
Nigeria
Niger
Angola
Burkina Faso
Congo, Dem. Rep.

Inferences from Bar chats

Conclusion:

1. We can conclude that the Clusters with cluster_id **0** using K_Means algorithm and cluster_labels **1** using hierarchical clustering are the most vulnerable countries and need help with immediate effect.

- 2. The Countries **HAITI, Sierra Leone, Chad, Central African Republic**, **Mali** are the top five countries which needs immediate attention from **HELP international**.