CLUSTERING ASSIGNMENT

Satadhriti Chakrabarty
DS C17 FEB 2020, GROUP 2

PROBLEM STATEMENT

Our NGO has raised a fund of around $10 million that will be used to aid economically backward countries in times of natural disasters. In this situation, we must figure out the countries which need financial aid at the earliest on the grounds of high child mortality, low income, low GDP per capita and other socio-economic metrics. A list of countries (5-10) will be highly beneficial for our NGO to prioritize and help those in the direst of states and lie at the bottom of the socio-economic ladder.

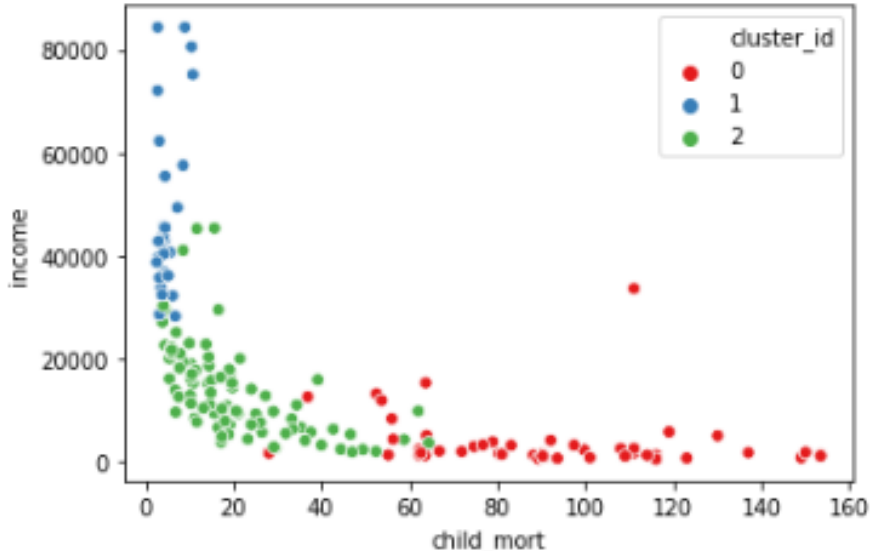
ANALYSIS APPROACH

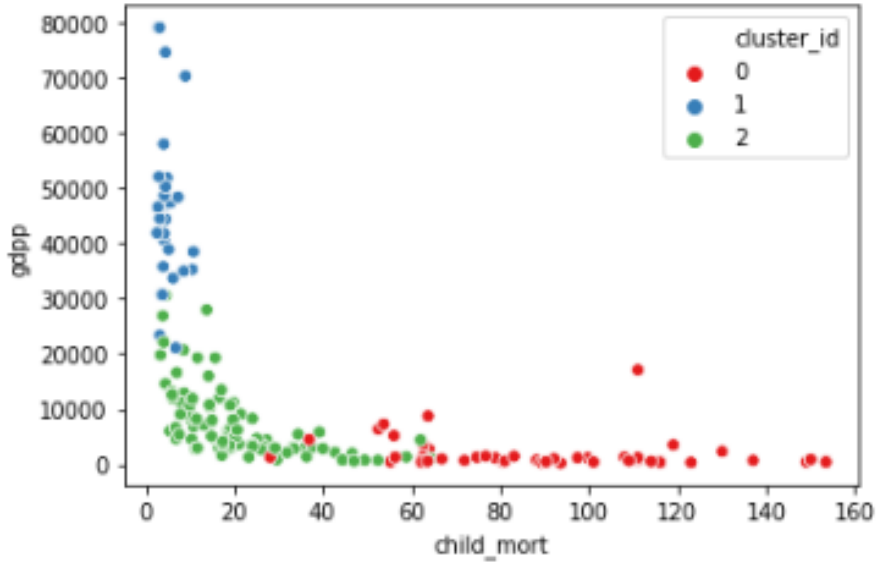
- Capping of outliers for the variables and scaling for all the continuous variables
- Hopkins score was checked to see if the data was suitable for clustering.
- From the Elbow curve and the Silhouette analysis, the optimal number of clusters for K-means clustering was selected
- Three most important variables viz. child mortality, income and GDP per capita were used to cluster the countries effectively for both K-means and Hierarchical clustering
- Scatterplots for each pair of the three above most important variables were plotted to identify the cluster with the highest child mortality, lowest income and lowest GDP
- A bar plot denoting each cluster against the three above variables was plotted to further confirm the results
- Top 10 Countries belonging to the cluster with the worst performance on the three metrics were figured out for both K-means and Hierarchical Clustering
- Results of K-means and Hierarchical clustering were compared and having similar results for both further confirmed the perfect approach used for clustering

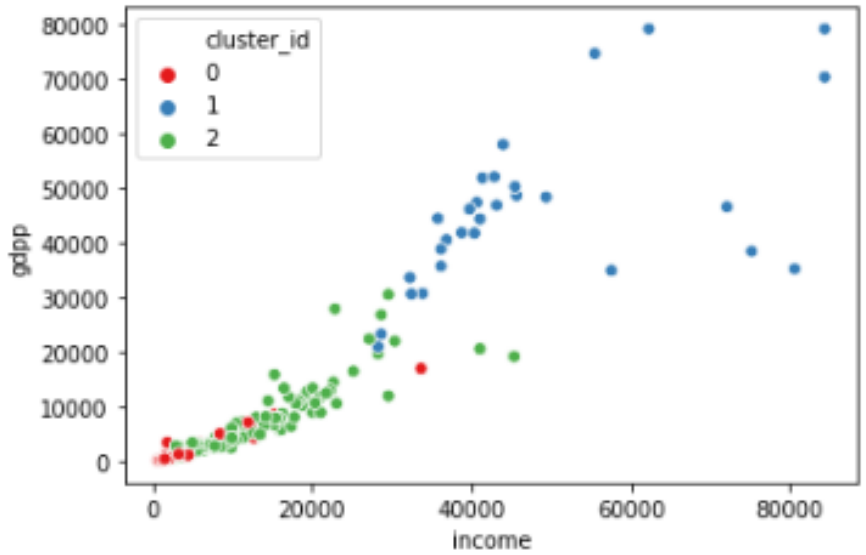
K-MEANS CLUSTERING

- The combination of the Elbow curve and the Silhouette scores assert that the optimal number of clusters must be 3 and from the business point of view, 3 would be the ideal number of clusters to help prioritize our money at hand. The Silhouette score for number of clusters as 4, is marginally smaller than that for the number of clusters as 3 and hence confirm the choice of 3 as the best possible number of clusters.
- The countries belonging to the cluster with the highest child mortality, lowest income and lowest GDP per capita will be the countries which will be the first of the lot to receive the financial aid and out of this cluster, we need to figure out the top 10 countries which have the worst socio-economic metrics.

VISUALIZATIONS OF K-MEANS CLUSTERING

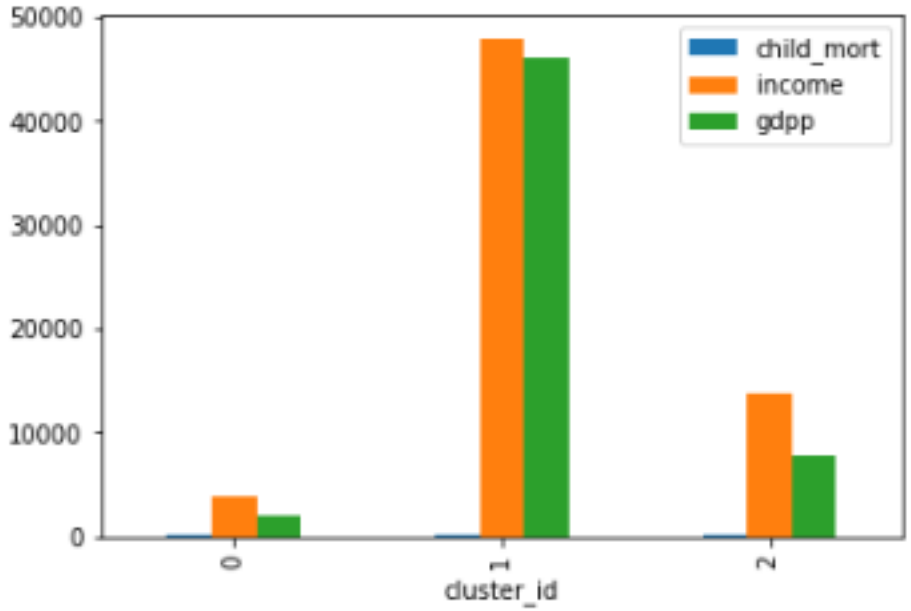






From the figures, it is certain that the countries belonging to Cluster_id 0 are the countries with the highest child mortality, lowest income and lowest GDP while countries belonging to Cluster_id 1 are the countries with the best performance on the metrics

CONFIRMATION OF THE CLUSTERS IN K-MEANS CLUSTERING

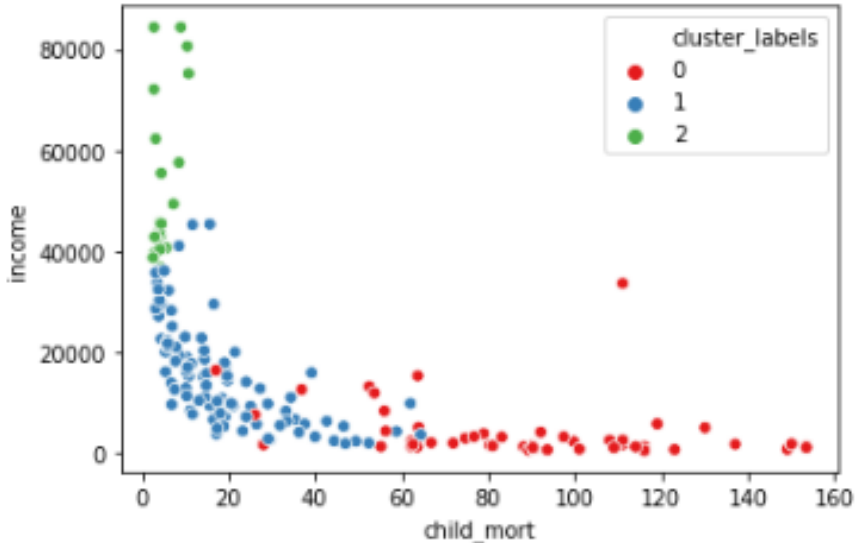


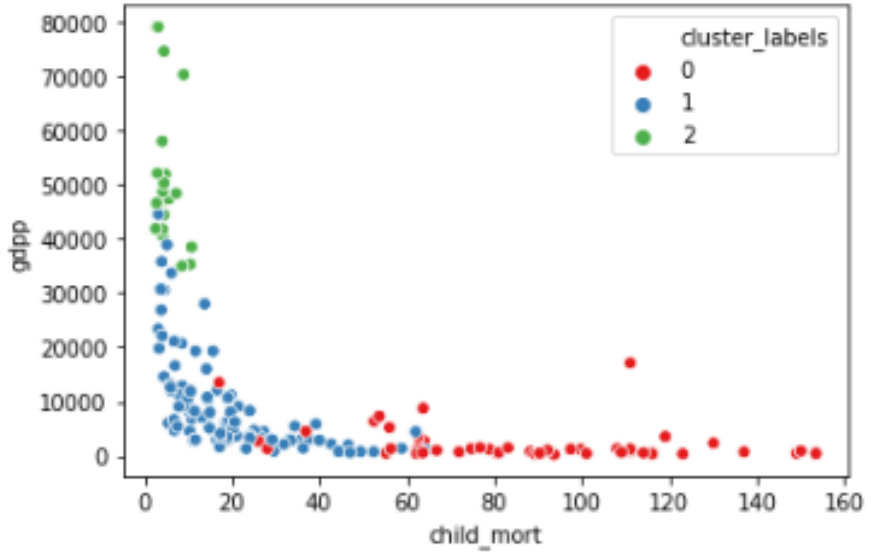
The above plot further confirm that countries belonging to cluster_id 0 are in dire states in terms of child mortality, income and GDP per capita and need aid at the earliest.

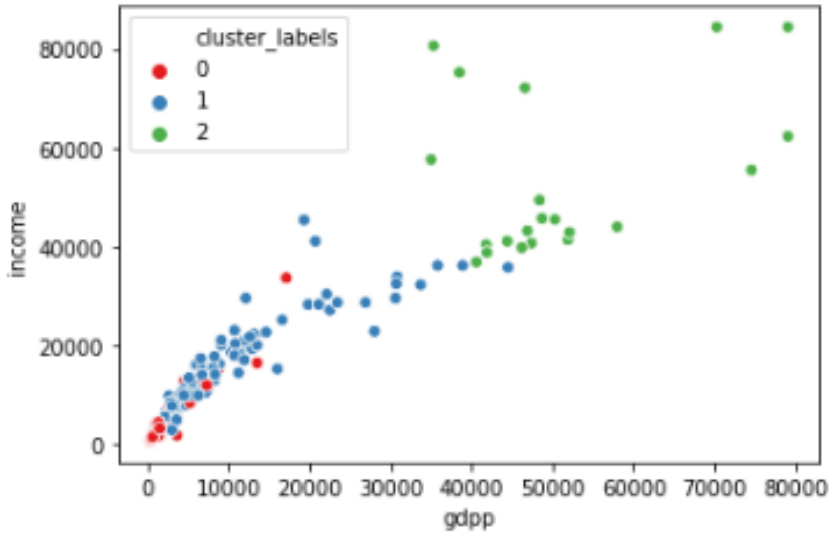
HIERARCHICAL CLUSTERING

- Single Linkage is not used as the interpretation would have been very difficult.
- Complete Linkage, on the other hand, provide a clearer view and given our understanding of the K-means clustering, we choose the number of clusters as 3 from the dendogram and compare the results with K-means clustering

VISUALIZATIONS OF HIERARCHICAL CLUSTERING

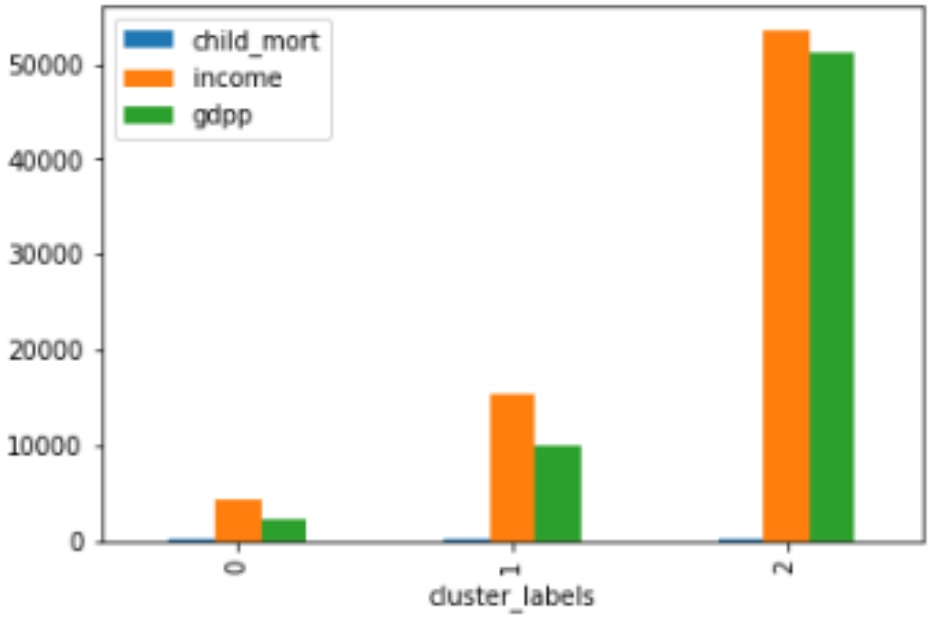






From the figures, it is certain that the countries belonging to Cluster_id 0 are the countries with the highest child mortality, lowest income and lowest GDP while countries belonging to Cluster_id 1 are the countries with the best performance on the metrics

CONFIRMATION OF THE CLUSTERS IN HIERARCHICAL CLUSTERING



The above plot further confirm that countries belonging to cluster_id 0 are in dire states in terms of child mortality, income and GDP per capita and need aid at the earliest.

FINAL RESULT

Combining the two results, we get a list of 10 countries that appear to be uniform. The countries with the worst performing metrics i.e. highest child mortality, lowest income and lowest GDP are:

i) Sierra Leone

ii) Haiti

iii) Chad

iv) Central African Republic

v) Mali

vi) Nigeria

vii) Niger

viii) Angola

ix) Congo Dem. Rep.

x) Burkina Faso

*The above countries should be prioritized above others in case of a future natural calamity or disaster. It is interesting to see that 9 out of these 10 countries are in **Africa** with the only exception being Haiti.*