CLUSTERING ASSIGNMENT

PRASASTI CHOUDHURY

DS C17 FEB 2020, GROUP 2

PROBLEM STATEMENT

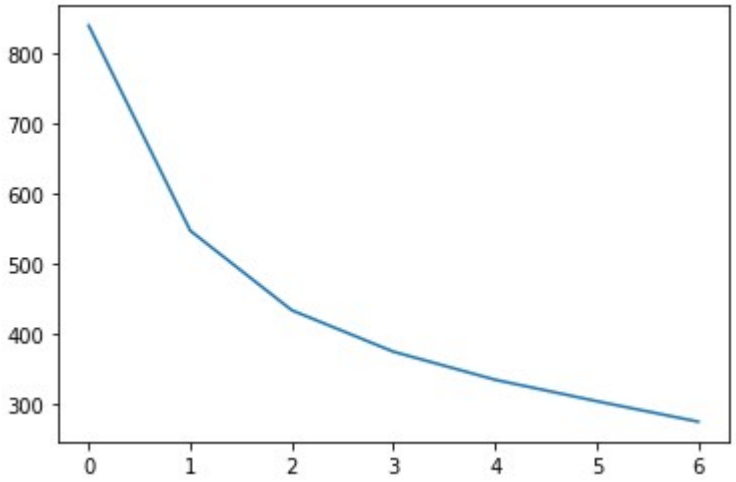
HELP NGO has managed to raise a fund of around $10 million. The CEO of the NGO has decided to help the countries who are in direst needs with this money. As a Data Analyst, the countries needs to be categorised based on its socio-economic and health conditions. The countries with the lowest gdpp, lowest income and the highest child mortality are in direst needs of the fund. Those countries needs to be identified so that the NGO can provide the aid.

ANALYSIS APPROACH

- Started with the data understanding by doing the EDA
- Data preparation by capping of the outliers, scaling of all the continuous variables and performing Hopkins check
- Performing Kmeans clustering:
  - Selecting the optimal K , by analysing from the Silhouette score and the elbow curve
  - Running Kmeans with the chosen K
  - Scatterplots were plotted to check on the cluster that has lowest gdpp, lowest income and highest child mortality rate. Plotting bar graph to confirm the observation
  - The first 10 countries with the above mentioned socio-economic and health conditions were identified based on the cluster that was identified from the graphs
- Performing Hierarchical clustering:
  - Plotting dendrograms for both single linkage and complete linkage- indentifying the number of clusters
  - Scatterplots were plotted again just like it was done in kmeans clustering to get the cluster that has lowest gdpp, lowest income and highest child mortality rate
  - The first 10 countries with the above mentioned socio-economic and health conditions were identified based on the cluster that was identified from the graphs
- Comparing the two countries list from the two clustering methods, its decided which of them are in direst need of the aid

KMEANS CLUSTERING METHOD

- From the Silhouette score and the Elbow curve, it was clearly visible that the optimal k for the number of clusters should be 3 . The countries were clustered in all these 3 clusters and plots were plotted.
- The first 10 countries with the lowest gdpp, lowest income and highest child mortality were identified to be prioritised for the aid.

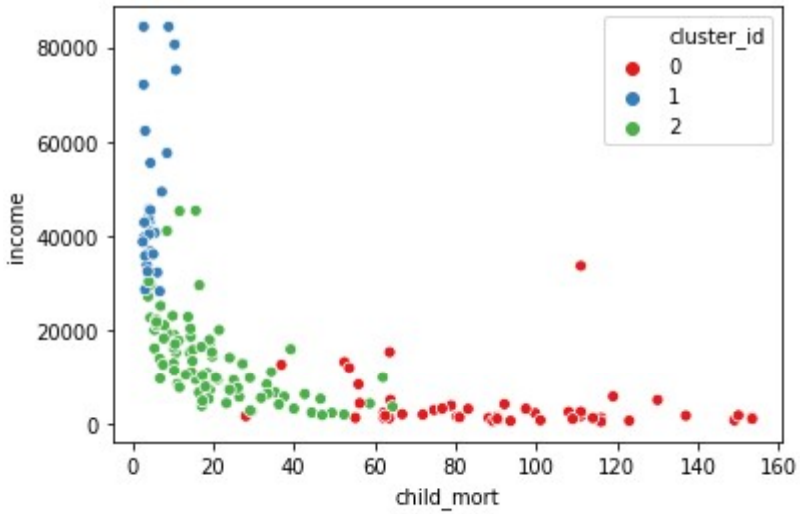


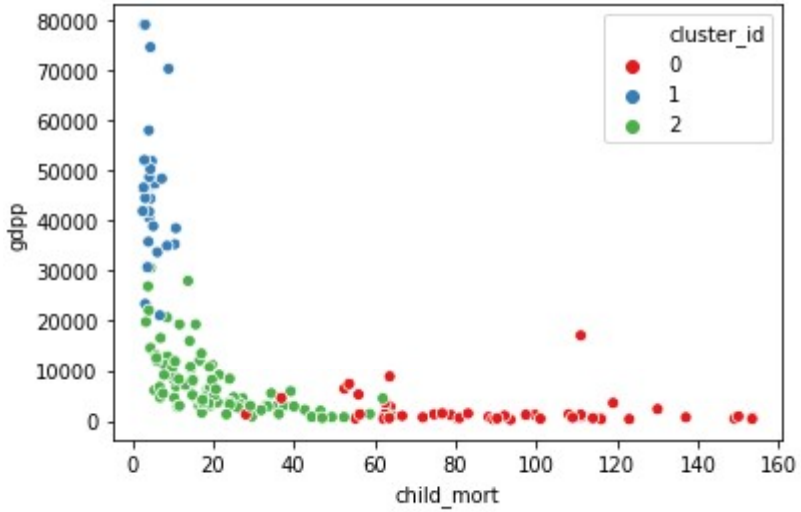
Elbow Curve

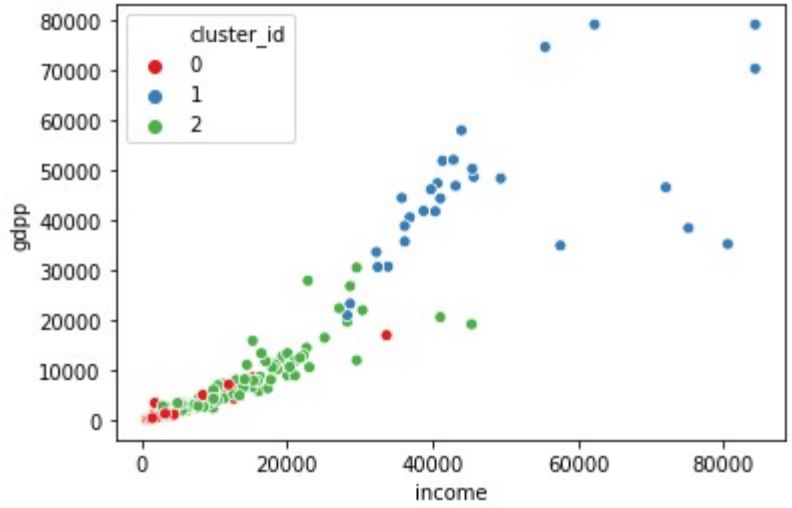
```
For n_clusters=2, the silhouette score is 0.46943108916984133
For n_clusters=3, the silhouette score is 0.4069661349025314
For n_clusters=4, the silhouette score is 0.3952110775478241
For n_clusters=5, the silhouette score is 0.38571514086454484
For n_clusters=6, the silhouette score is 0.30094905601572
For n_clusters=7, the silhouette score is 0.31081934752255896
For n_clusters=8, the silhouette score is 0.3233283616337525
```

Silhouette Score

VISUALISATION OF KMEANS CLUSTERING

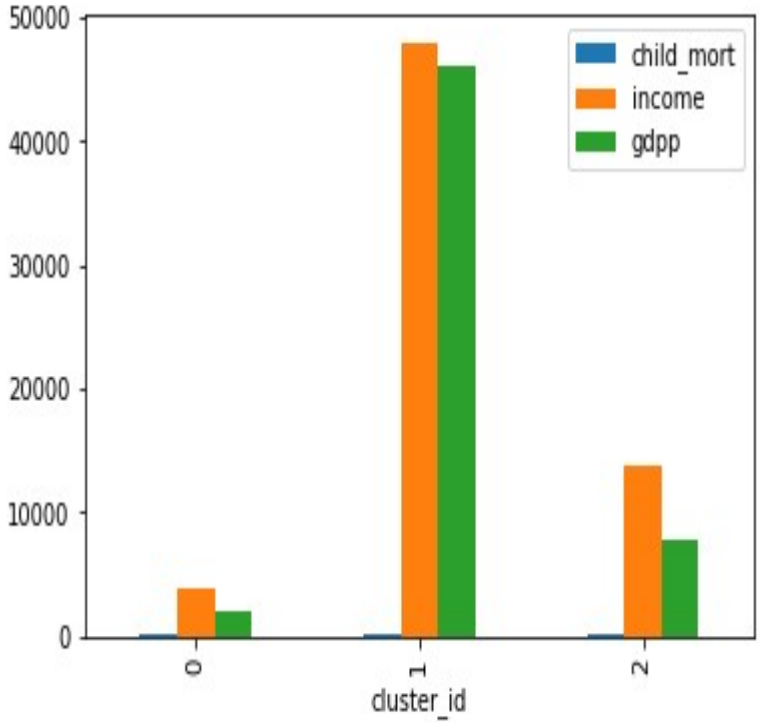






From the above scatter plots it can be clearly seen that Cluster 0 countries have low gdpp, low income and high child mortality. Where as the cluster 1 countries have the highest income, highest gdpp and lowest child mortality. So the aid must be provided to the cluster 0 countries.

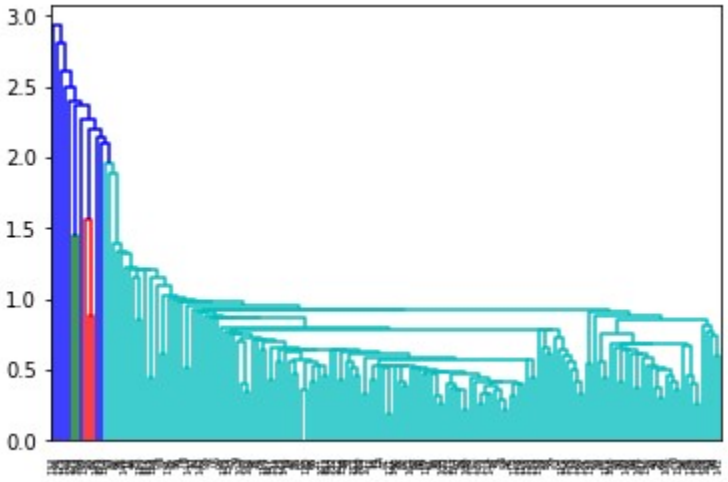
BAR GRAPH VISUALISATION



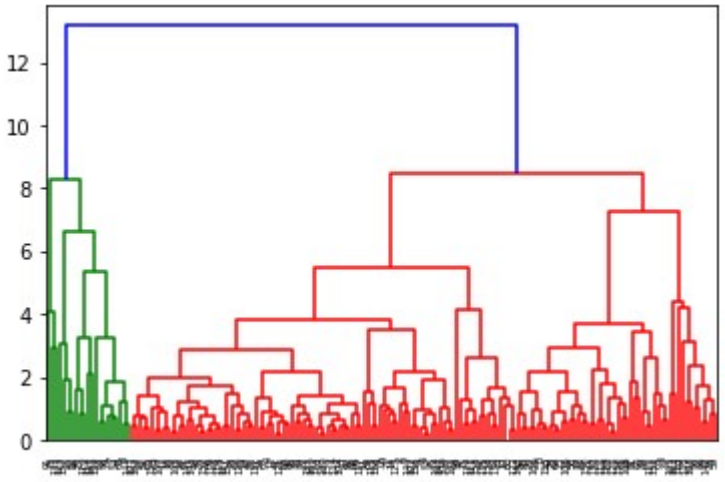
From this bar plot it can be confirmed that the Cluster 0 countries needs the aid.

HIERARCHICAL CLUSTERING

- Both single linkage and complete linkage dendrograms were plotted. But the visualisation in single linkage is not clear, so according to the number of clusters were chosen according to the complete linkage.
- The logical clustering was of the number 4, but in the 4th cluster only 3 countries were coming, which is not proper from the business perspective. So went with 3 as the number of clusters.

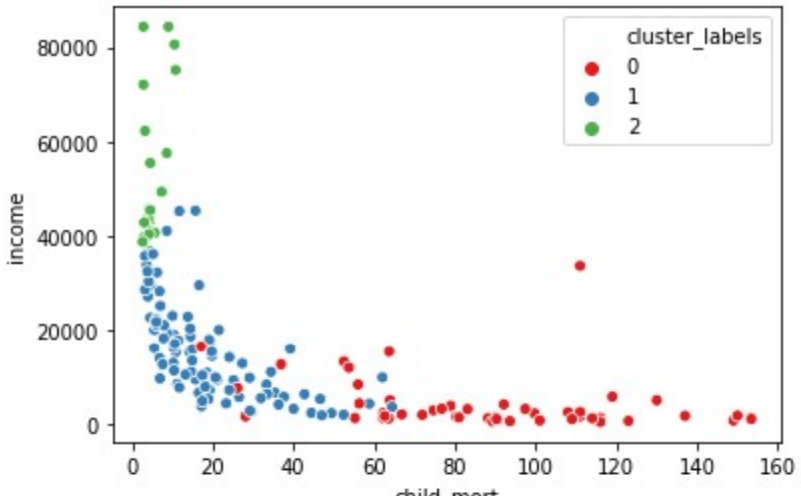


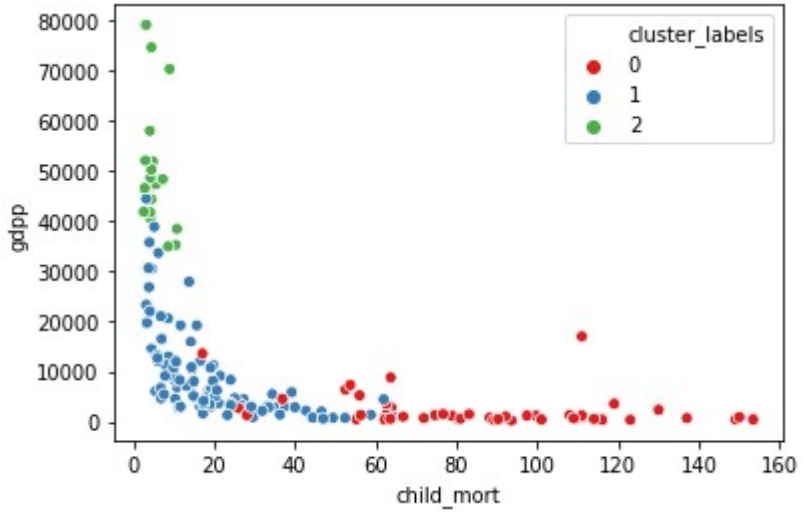
Single Linkage

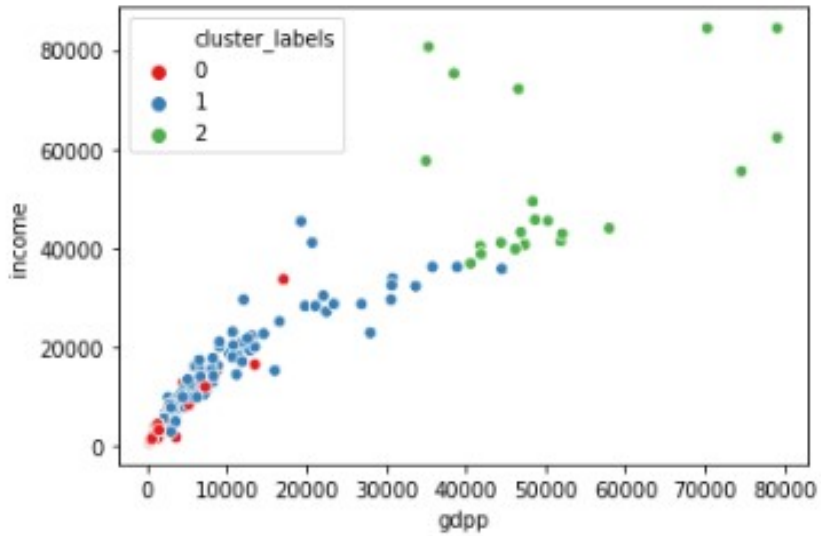


Complete Linkage

VISUALISATION OF HIERARCHICAL CLUSTERING

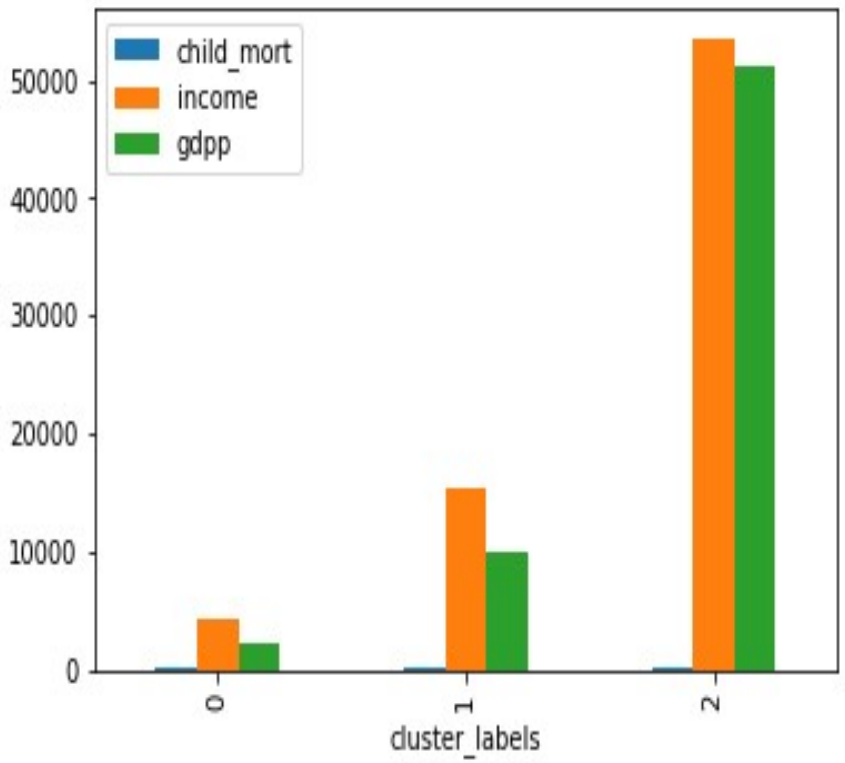






From the above plot it can be seen that:
Cluster0- low income, low gdpp, high mortality
Cluster1- midranged income,gdpp and child mortality
Cluster2- high income, high gdpp and low child mortality

BAR GRAPH VISUALISATION



The mean plot of the clusters shows that the Cluster 0 has the lowest income,lowest gdpp and highest child mortality wheres as the Cluster 2 has the highest gdpp, income and lowest child mortality

FINAL OBSERVATION

- Comparing the two results from both the clustering, the 10 countries which needed the most help financially due to their low socio-economic and health conditions were found out.
  - Sierra Leone
  - Haiti
  - Chad
  - Central African Republic
  - Mali
  - Nigeria
  - Niger
  - Angola
  - Congo Dem. Republic
  - Burkina Faso

- The above countries must be prioritised for the NGO aid.