HELP INTERNATIONAL

CLUSTERING OF COUNTRIES

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 programs, 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.

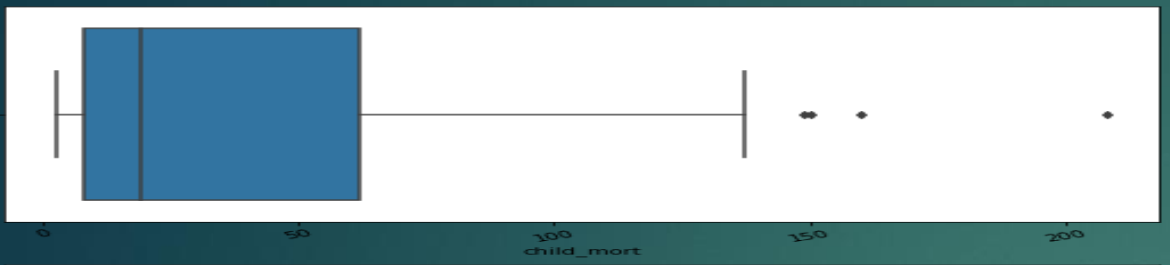
Analysis Approach

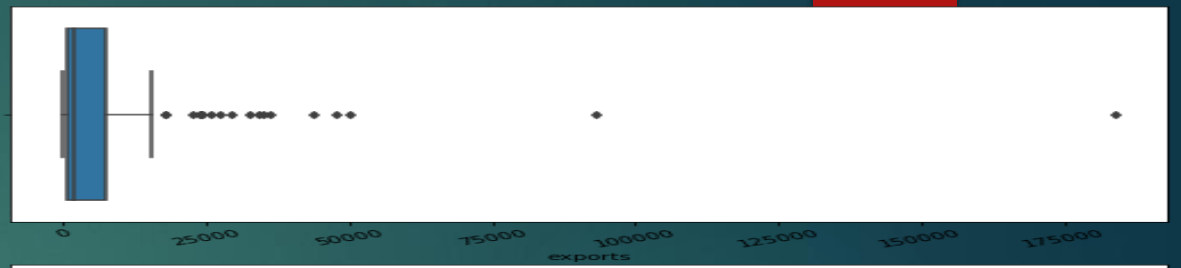
- To categorize the countries based on the socio-economic and health factors that determine the overall development of the country
- To suggest the countries which the CEO needs to focus the most
- The approach we used here to cluster the countries are:
  - Data Inspection and EDA
  - Outlier analysis
  - Hierarchical and K means clustering
  - Analysis of these clustering techniques

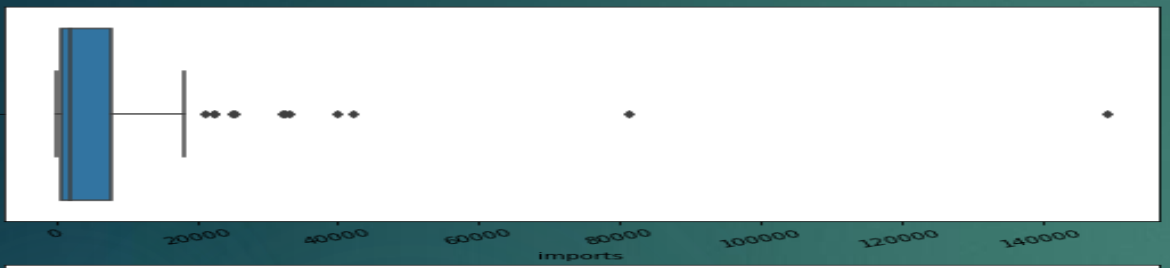
Data Processing

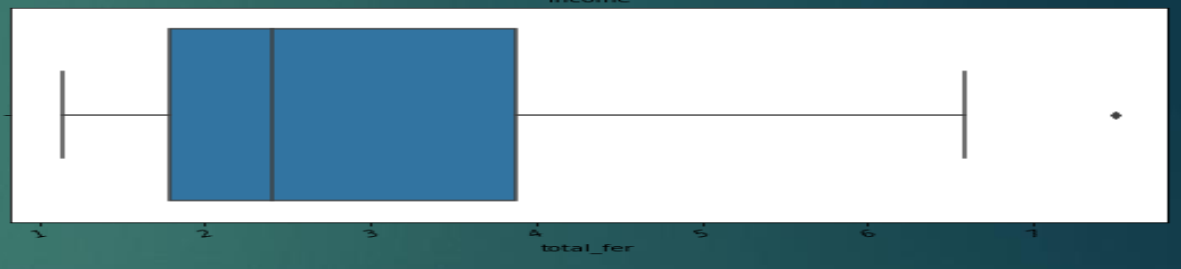
- It was found that there were no null values in the data.
- There were also no duplicate countries in the data
- There were few outliers which was capped later as dropping them would not help us in our analysis
- Converted the columns (Health, imports, exports) from percentage to absolute values
- Later used StandardScaler() to scale the data for modeling

Outliers in the data



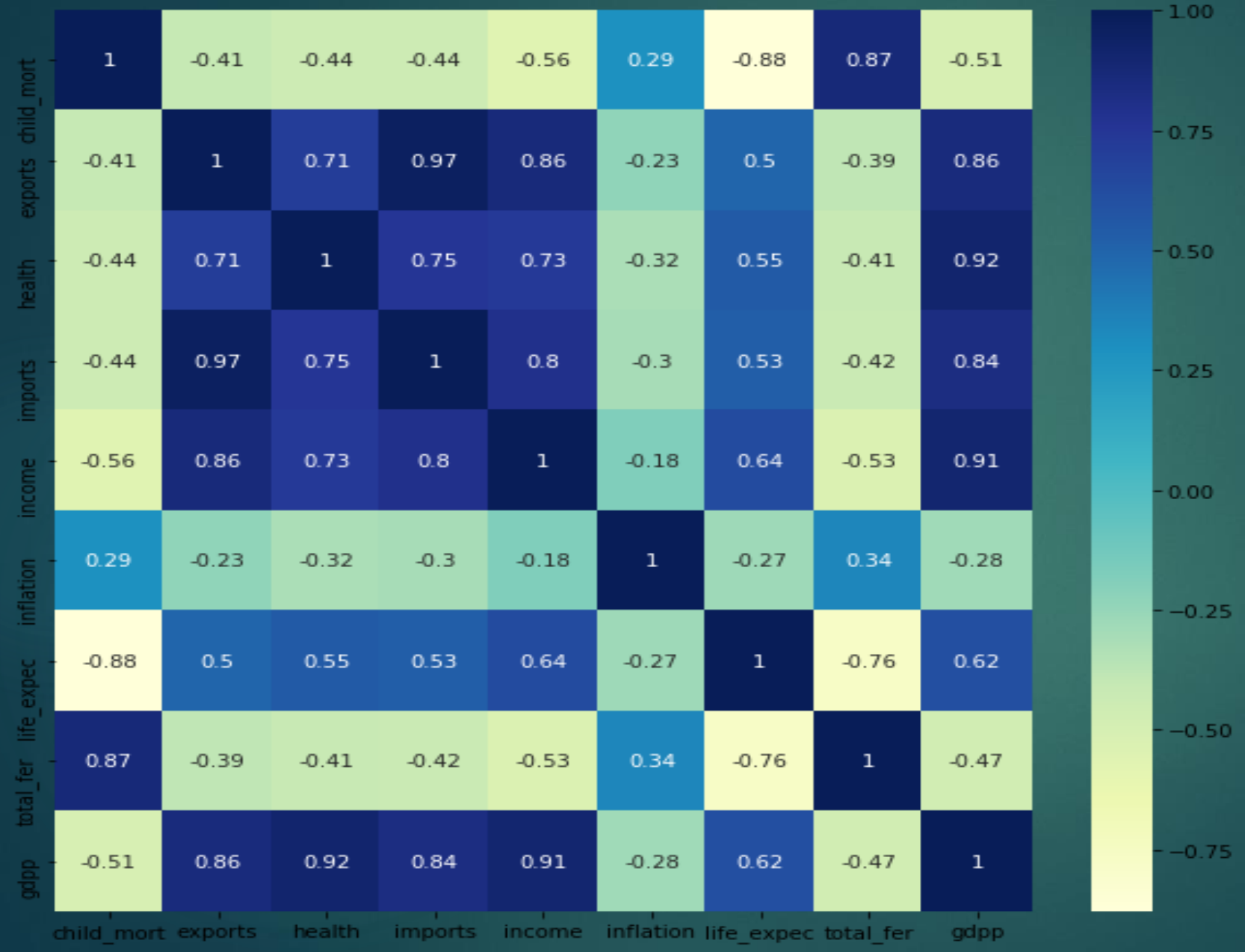






These outliers were treated using capping method as every country is valuable to us and we cannot drop them

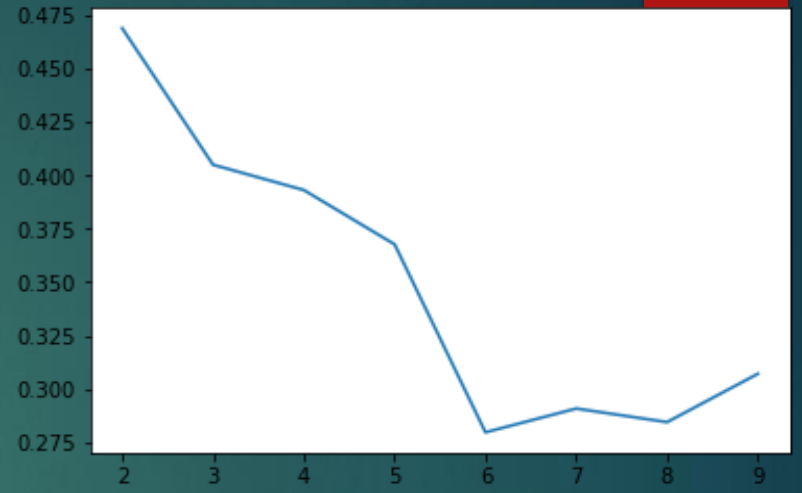
Correlation of the columns



- This is the correlation heat map of the data.
- As you can see child mortality and inflation are related to total fertility.
- Exports are related to Imports
- Health and Income are related to gdpp.
- Life expectancy is related to income

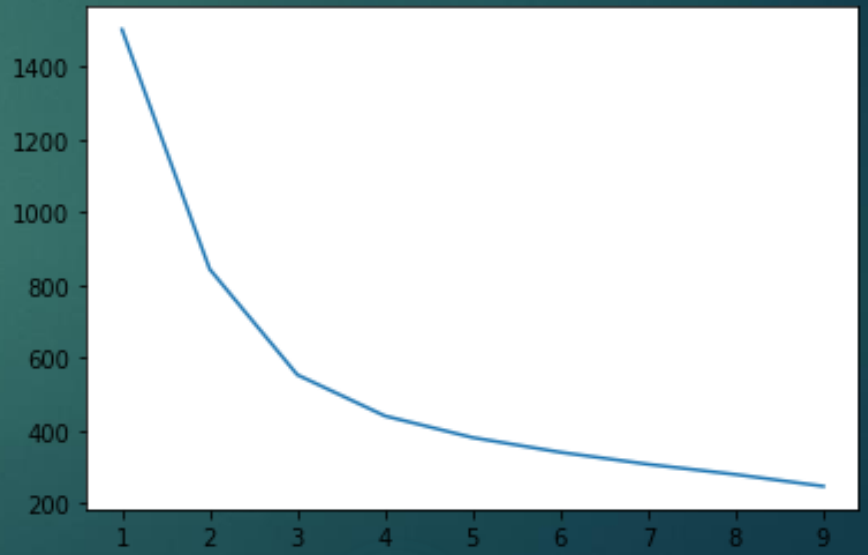
Clustering

- Before clustering we did Hopkins statistics where the measure turned out to be 94, which means there are high chances of clustering.
- Then we scaled the data frame using StandardScaler().
- We did a silhouette analysis and below are the scores:
  - For n_clusters=2, the silhouette score is 0.46882809604417697
  - For n_clusters=3, the silhouette score is 0.4049550565492161
  - For n_clusters=4, the silhouette score is 0.393118398962204
  - For n_clusters=5, the silhouette score is 0.3673844426351726
  - For n_clusters=6, the silhouette score is 0.32287535995201294
  - For n_clusters=7, the silhouette score is 0.3117900673588419
  - For n_clusters=8, the silhouette score is 0.3172384345774441
  - For n_clusters=9, the silhouette score is 0.3104880410350571



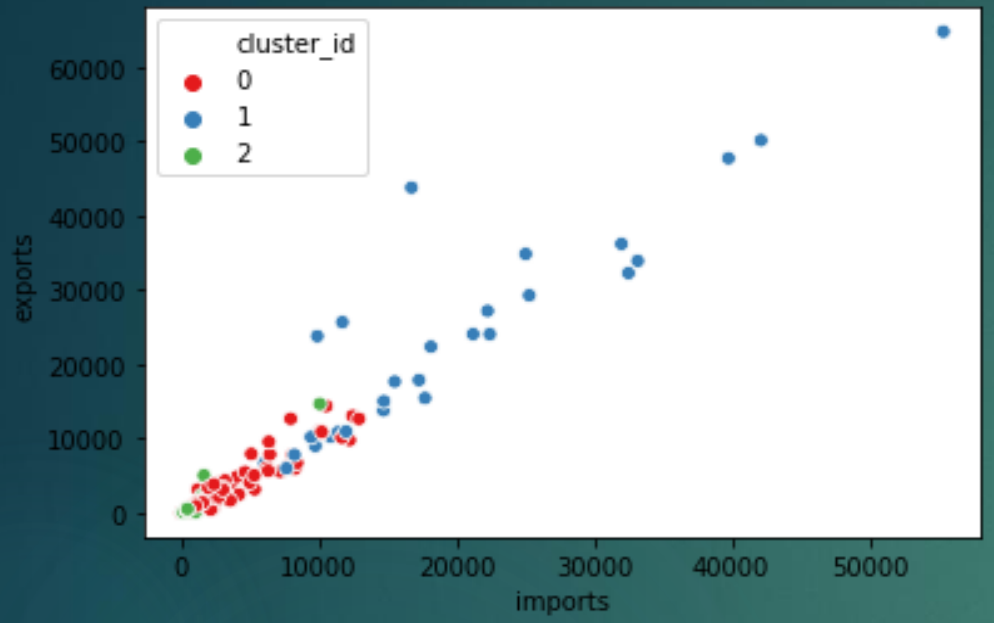
Silhouette analysis

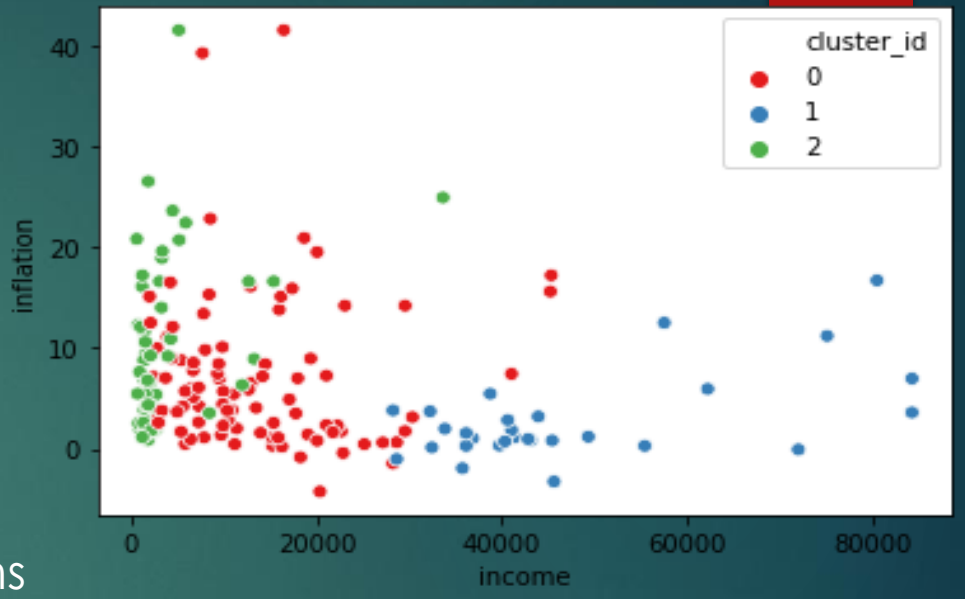
- From the above analysis the perfect k seems to be 3.
- To confirm it further we did an elbow curve/SSD analysis and our optimal cluster seems to be 3 itself.



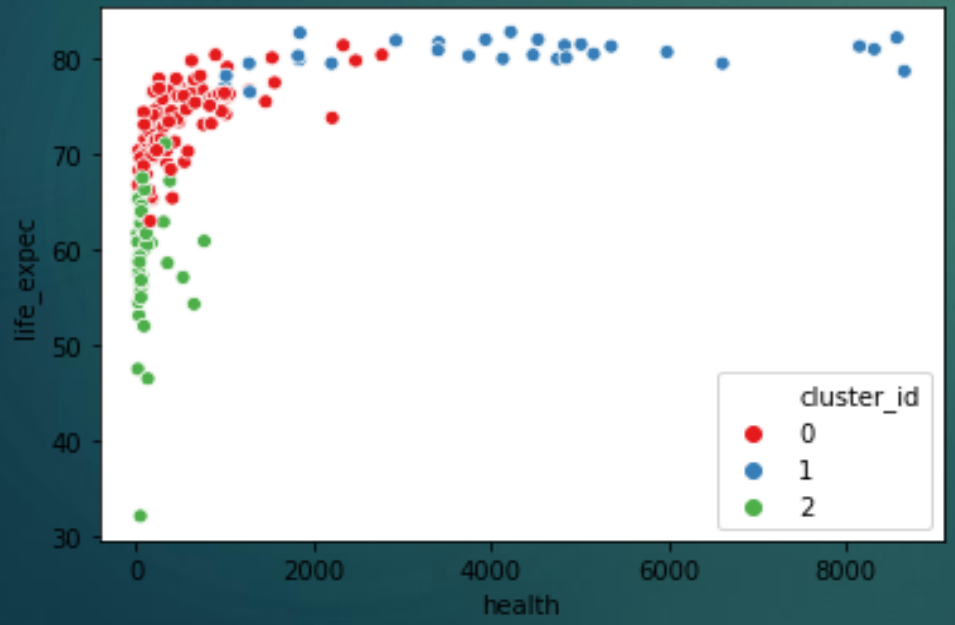
Elbow curve/SSD

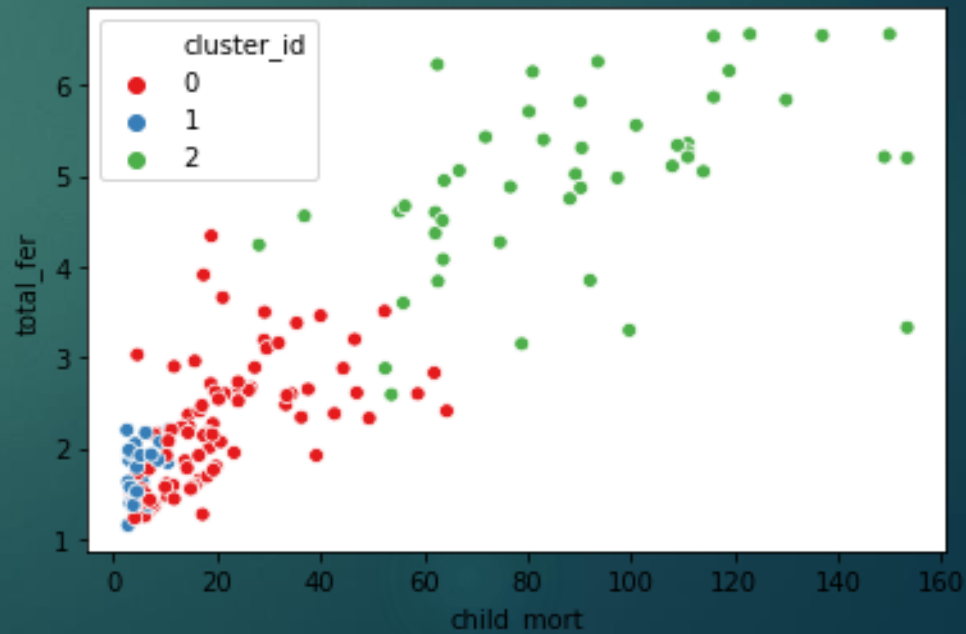
K-means clustering visualizations



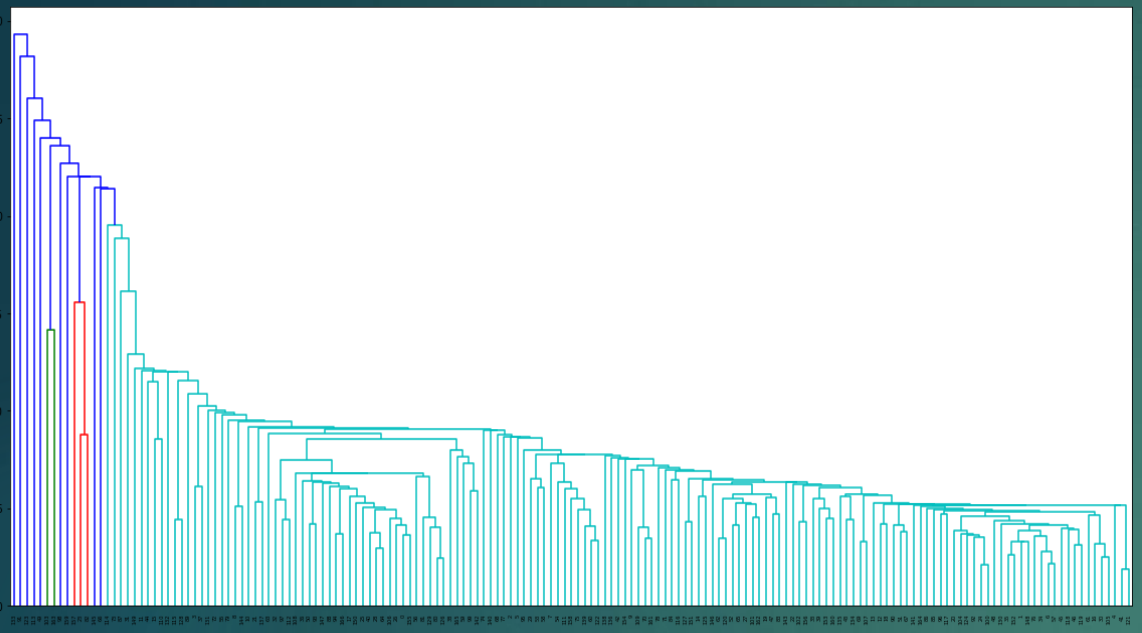


The optimal cluster id seems to be 2 as seen in these plots

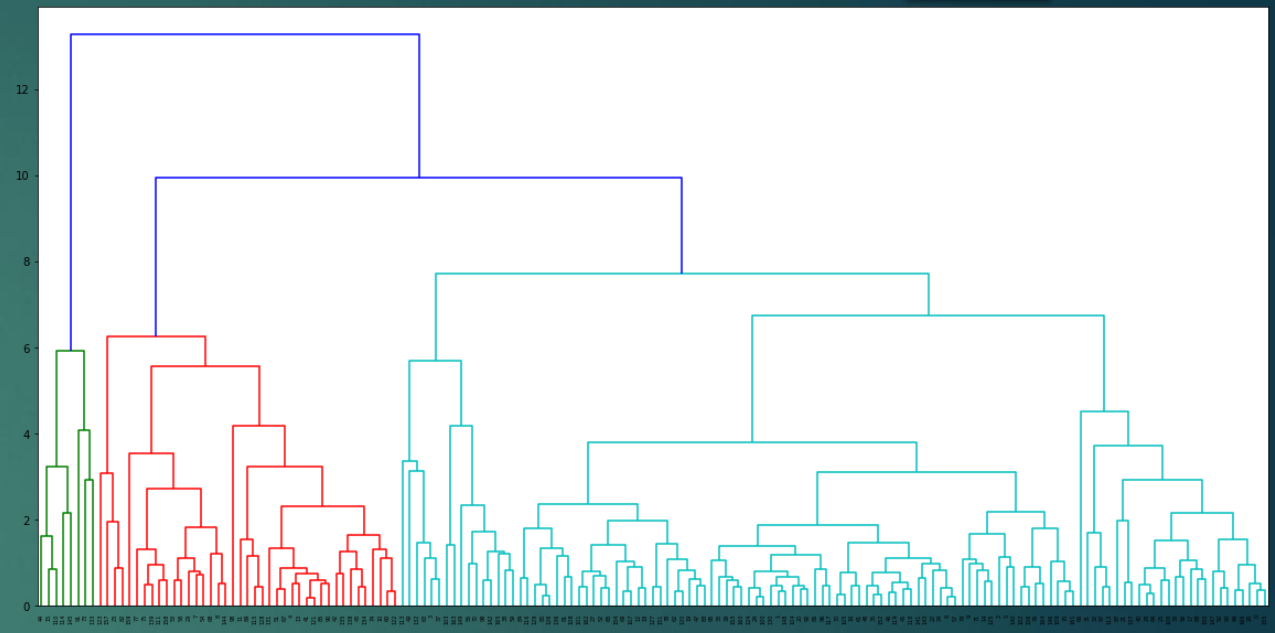




Hierarchical Clustering



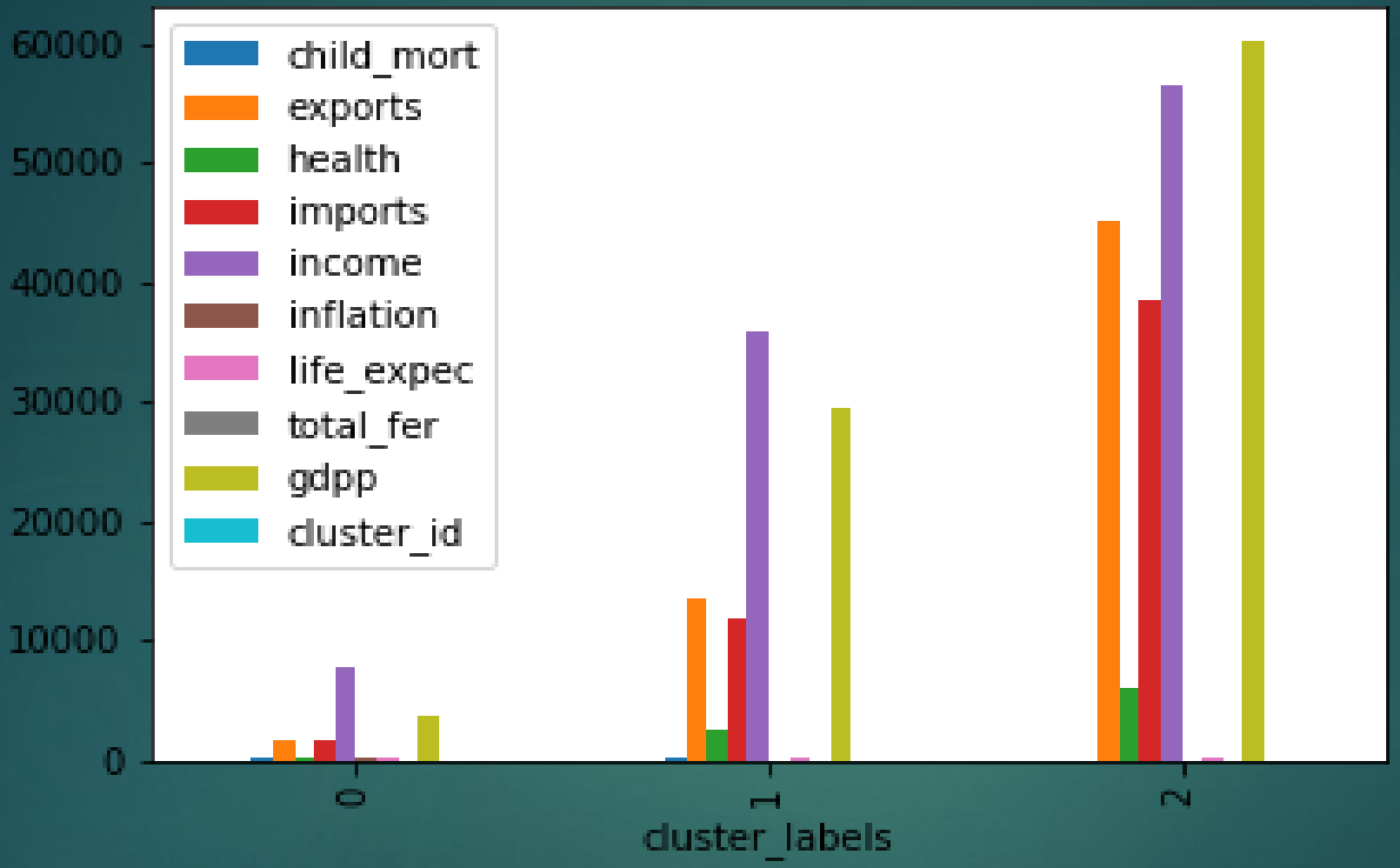
Single Linkage



Complete Linkage

- Complete Linkage seems more convenient and understandable than single linkage, hence taking it into consideration we optimal cluster as 3.
- And finally we choose the top 10 countries

Plot of the cluster which needs help



As you can see the cluster id of 0 are the one that needs aid

Conclusion

- We categorized the countries based on socio-economic and health factors and sorted them based on their gdpp, child mortality and income.
- The following are the countries that we need to focus the most

The Top 10 Countries:

Burundi

Liberia

Congo, Dem. Rep.

Niger

Sierra Leone

Madagascar

Mozambique

Central African Republic

Malawi

Eritrea