Clustering Assignment

DONE BY – DHRUV SHARMA

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.

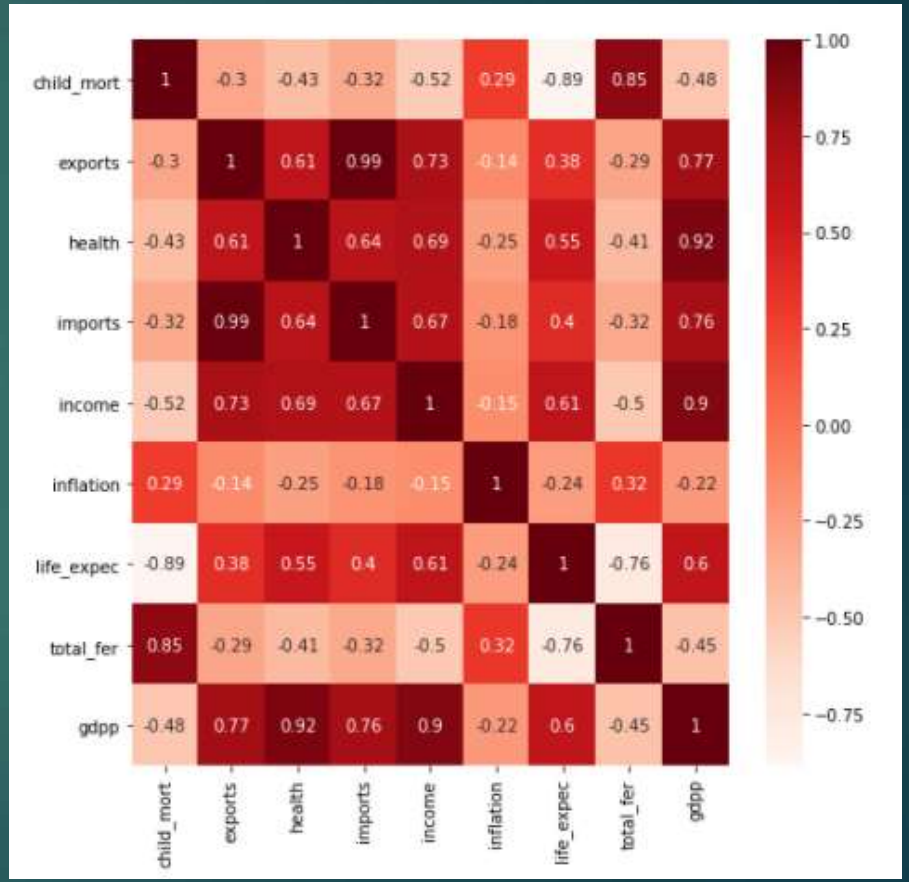
Objective

- The requisite 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.

Exploratory Data analysis and Data visualization

From the heatmap, it can be observed that:

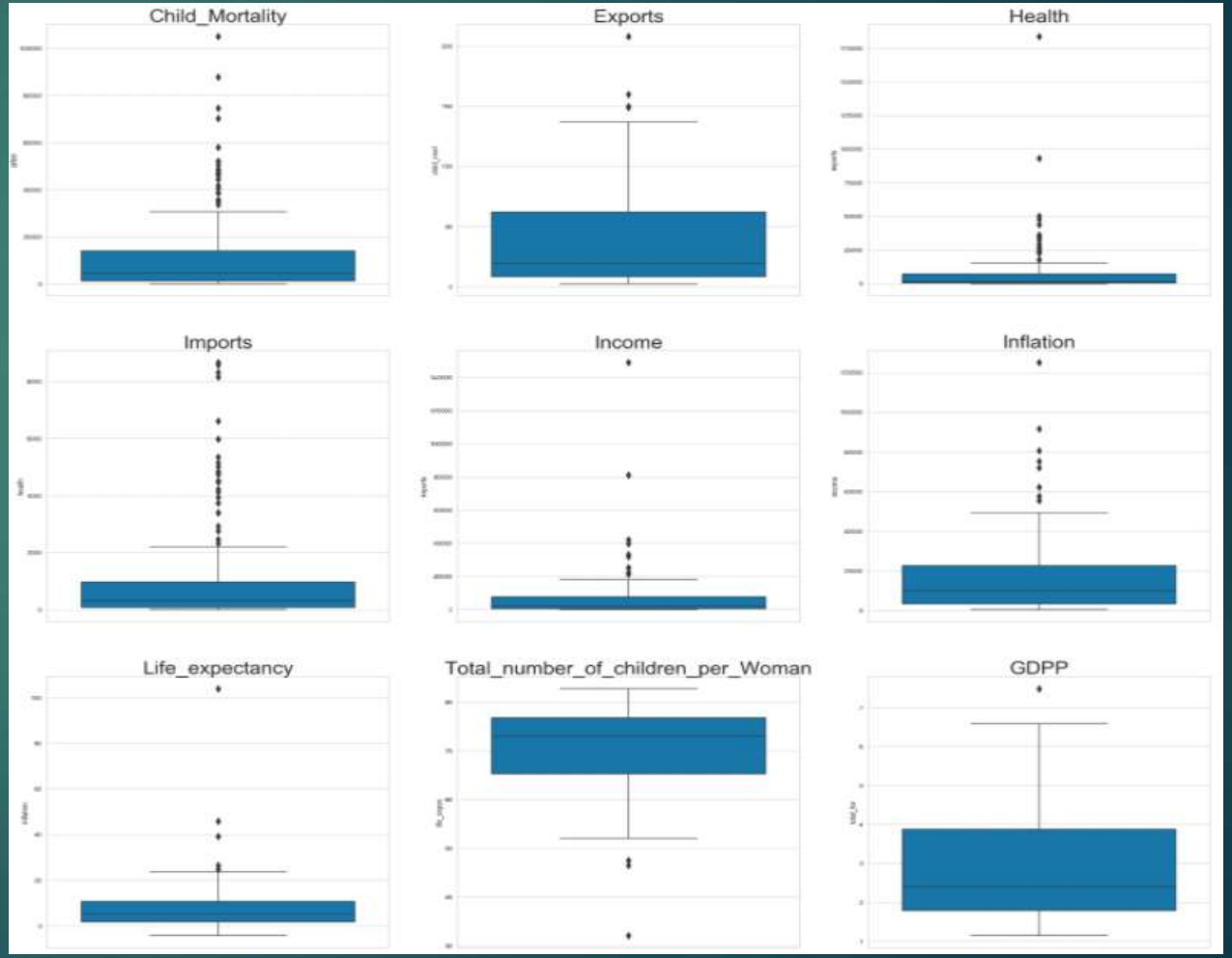
1. Some of the variables have high correlation between them eg- imports,exports, health and gdpp
2. Some features have negative correlation eg- child mortality and life expectancy, gdpp and health etc
3. Thus the dataset consists of multicollinearity.



Visualization of Outliers

All the features consist of outliers in them.

For the purpose of the analysis, we will soft cap the outliers i.e. Removing the datapoints which are above 0.99 quantile of the data range.

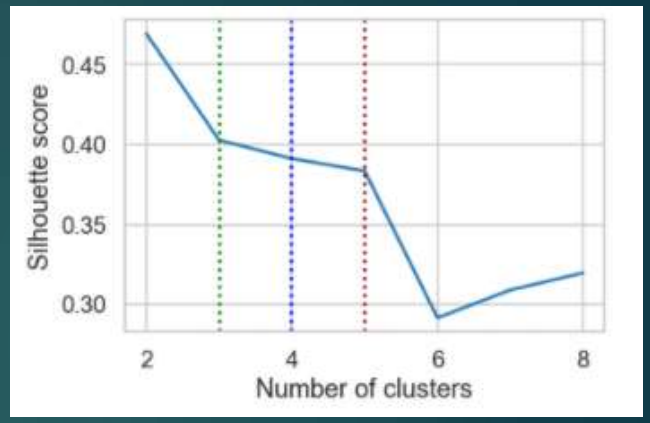


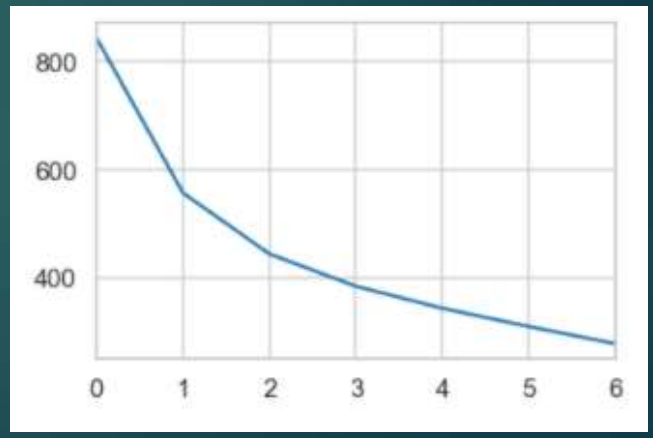
Data Preprocessing

- It was observed that the data does not contain any null values
- There are no duplicates present in the dataset which ensures the data is clean
- The features which consisted of outliers were treated by using soft capping method
- The data was scaled using standard scaler before feeding into the Machine Learning model

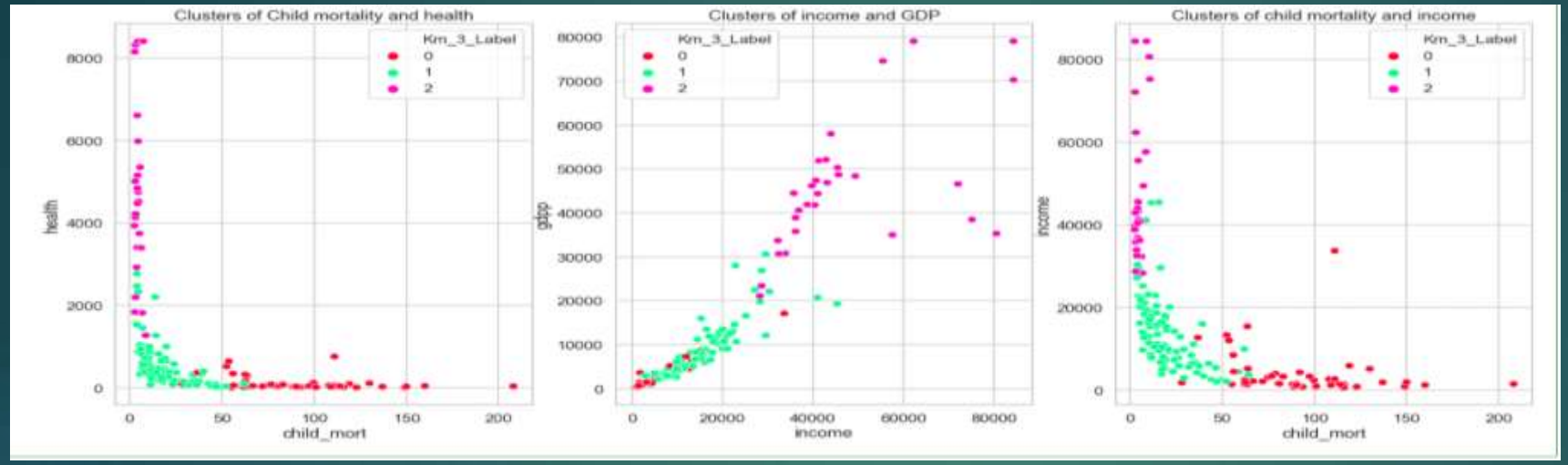
Model Building using K-Means Clustering

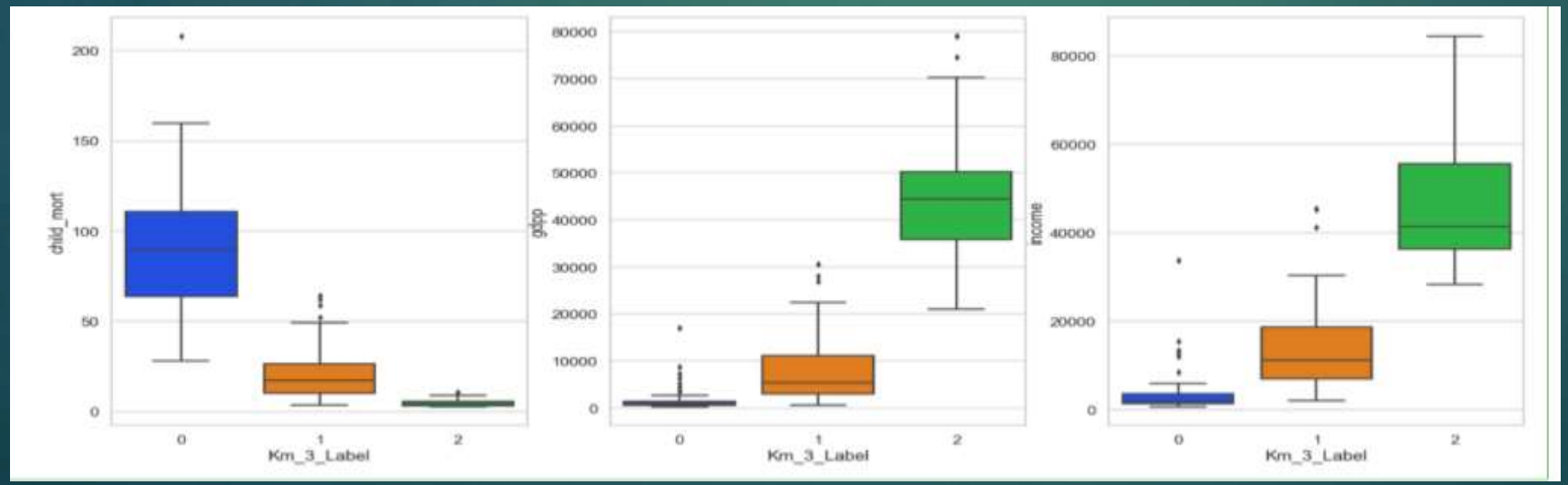
- We need to find the optimal number of k for k-means clustering
- Silhouette analysis was performed on the dataset
- The elbow curve/ SSD method was also performed for finding out the optimal number of clusters to be used
- Thus after looking at the graph, K-means clustering was performed by taking values as 3,4 and 5
- The cluster value of 3 was found to be appropriate for segmenting the data into 3 clusters





EDA for K-Means Clustering

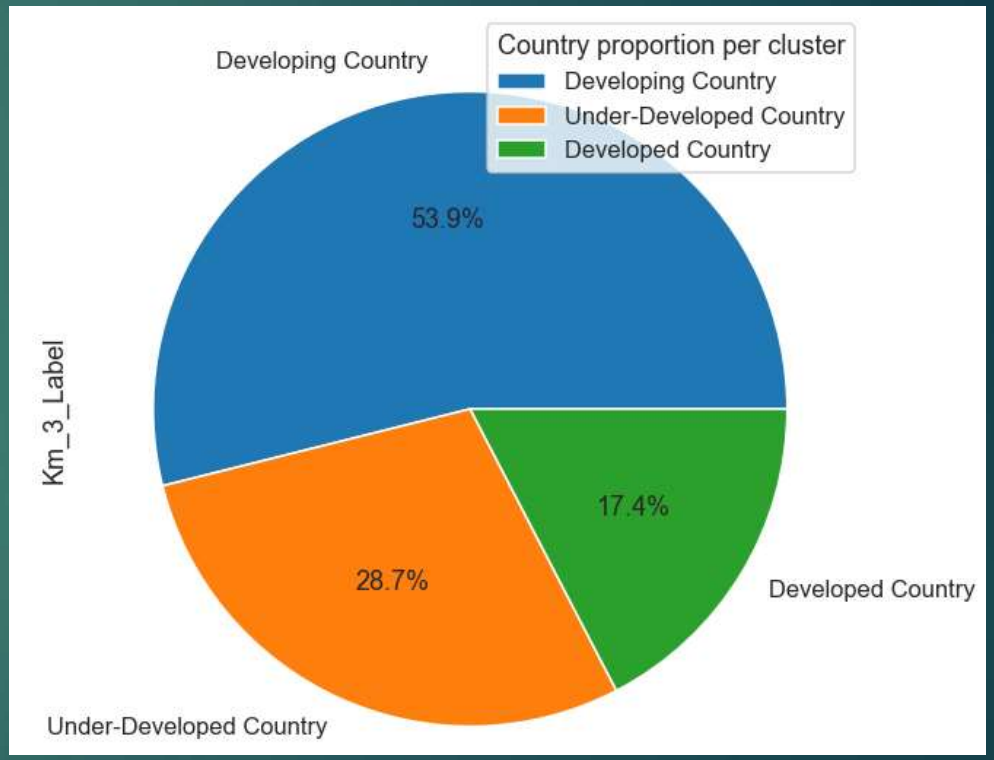




K-Means model Interpretation and Conclusion

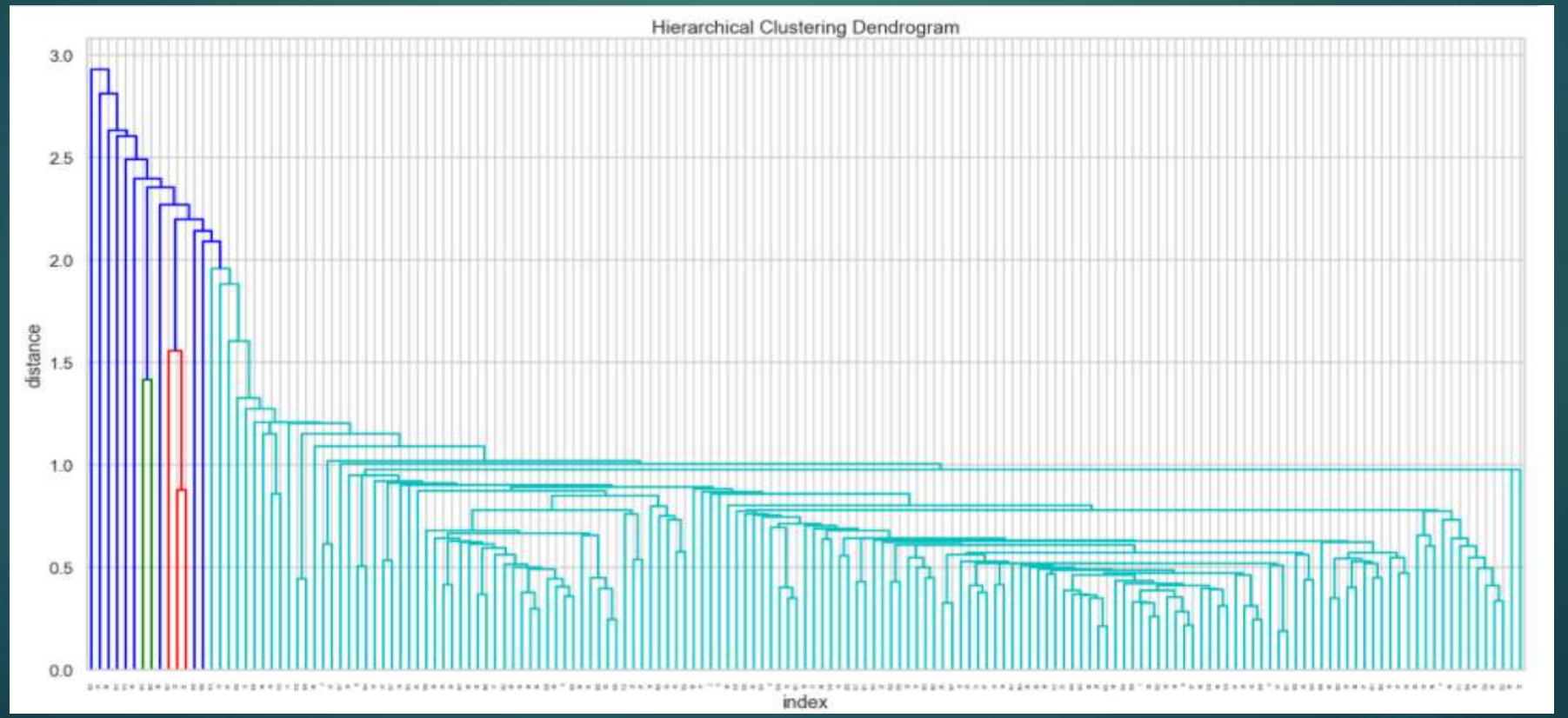
The top 5 countries in need of financial aid according to K-Means clustering are:

1. Burundi
2. Liberia
3. Congo, Dem. Rep
4. Niger
5. Sierra Leone

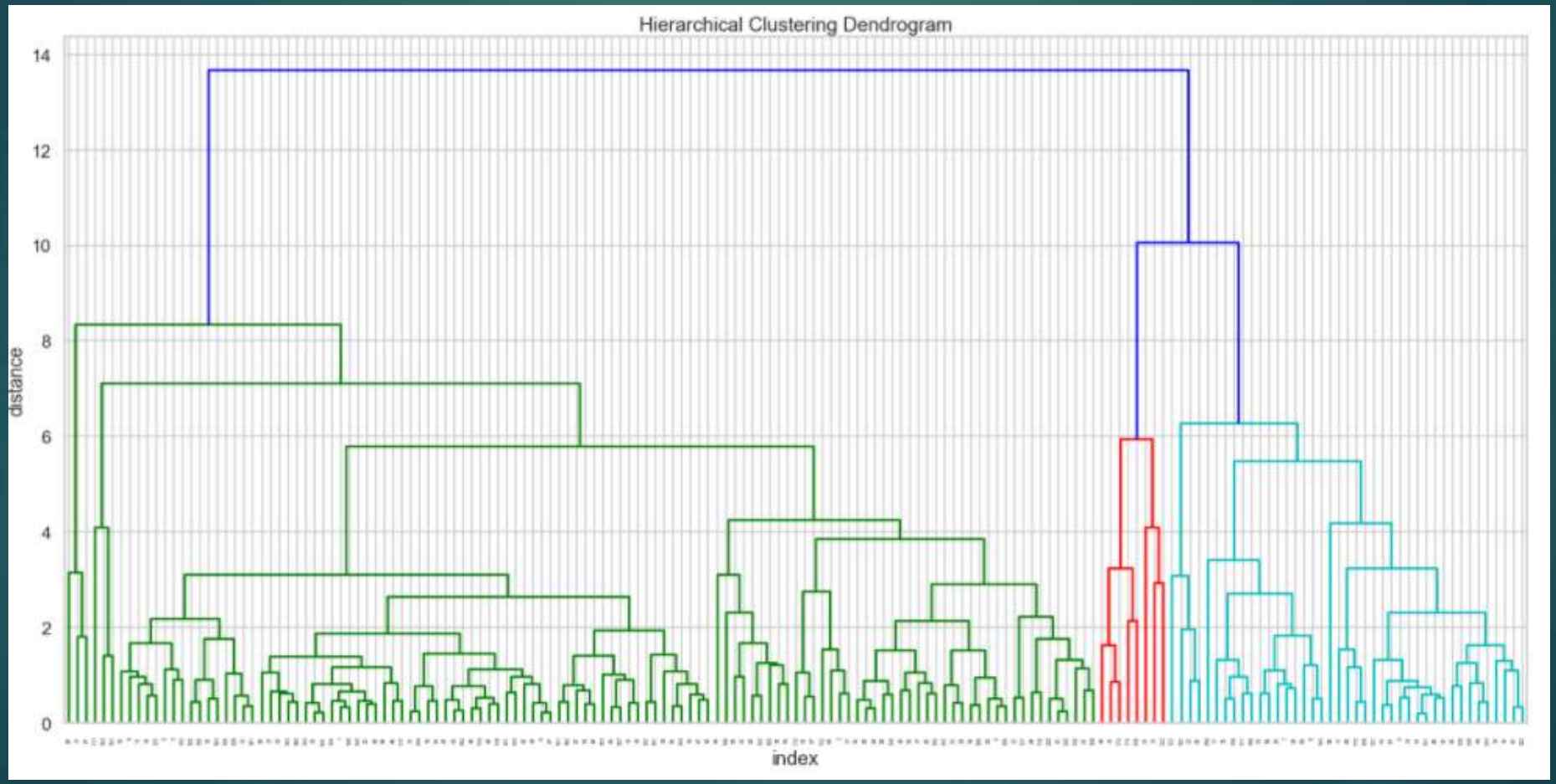


Model Building using Hierarchical Clustering

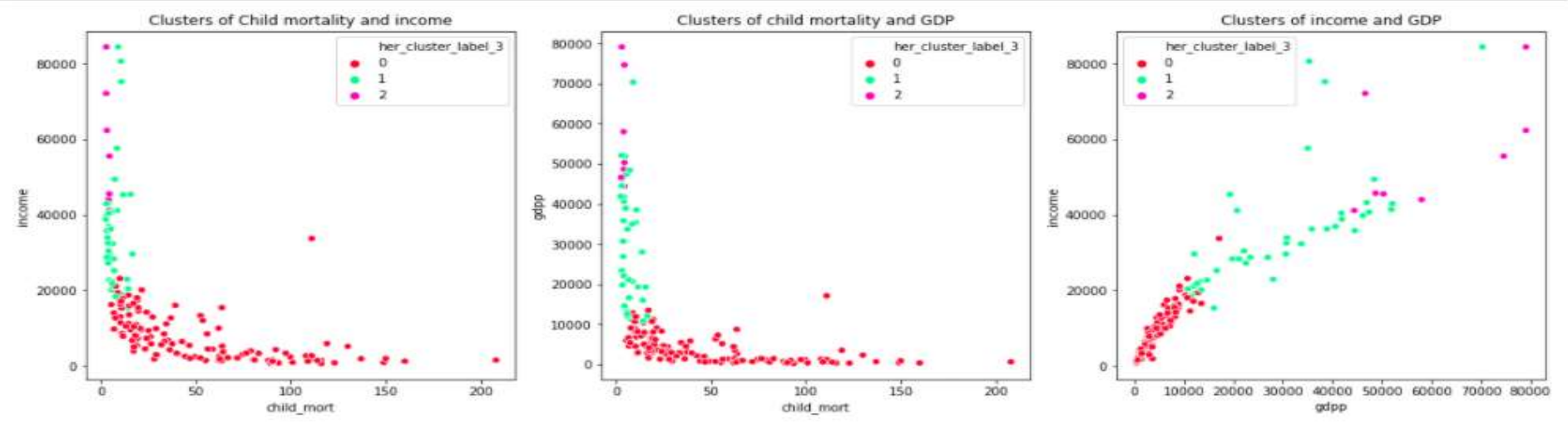
- Hierarchical clustering using Single Linkage Method proved to be inappropriate as the results were not interpretable.

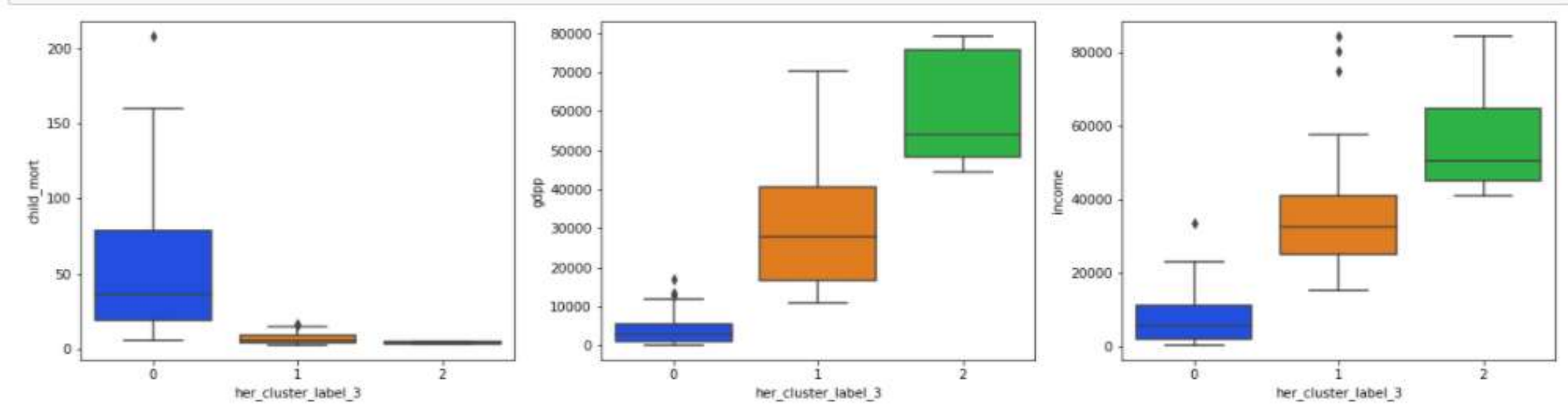


- Hierarchical clustering using complete linkage method was observed to be interpretable and a threshold value of 8 was chosen to be appropriate for clustering



EDA for hierarchical clustering

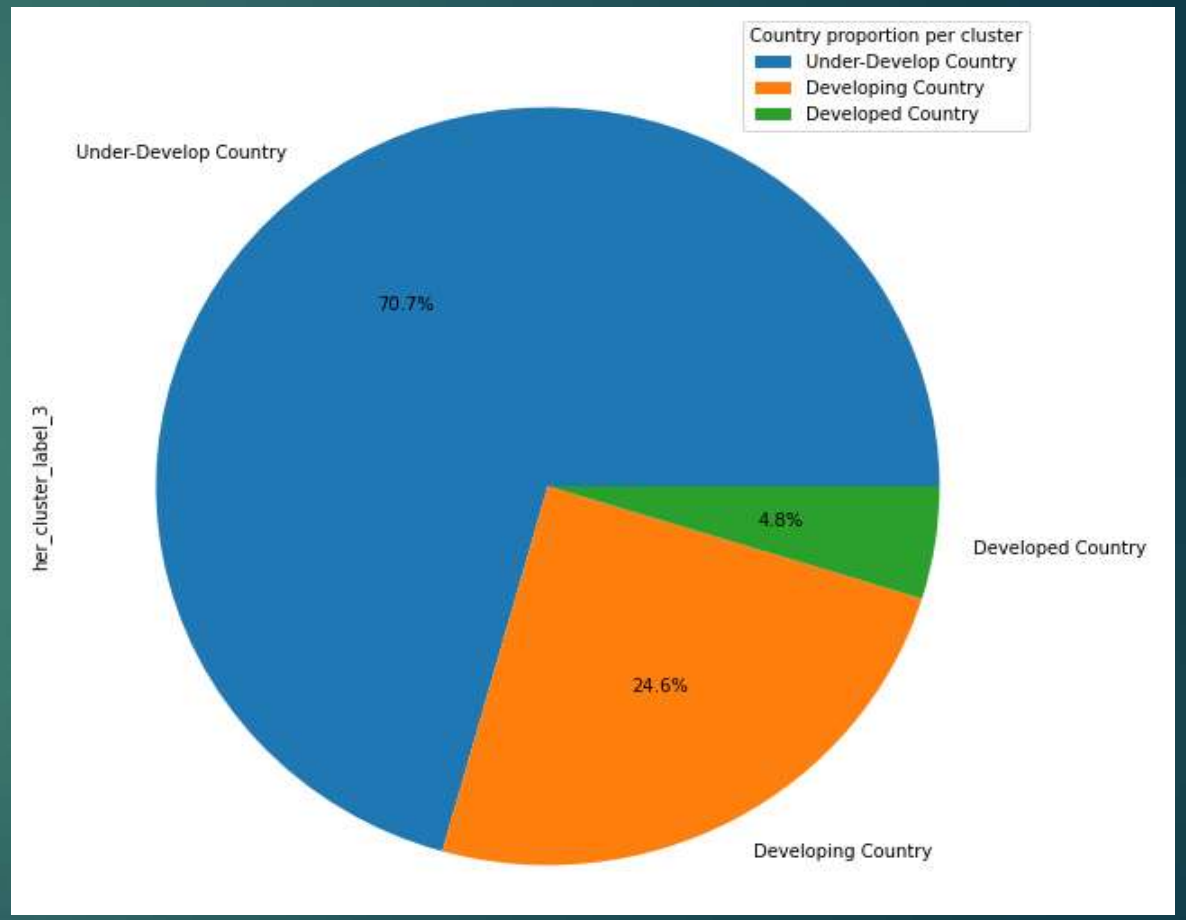




Hierarchical model Interpretation and Conclusion

The top 5 countries in need of financial aid according to Hierarchical clustering are:

1. Burundi
2. Liberia
3. Congo, Dem. Rep
4. Niger
5. Sierra Leone



Conclusion

- After clustering the countries based on K-Means clustering and Hierarchical clustering, it was found that the countries which are in need of financial aid are the same for both the approaches
- These countries need to be focused by the NGO and financial aid should be provided to them
- The 5 countries which are in dire need of financial aid are:
  - 1. Burundi
  - 2. Liberia
  - 3. Congo, Dem. Rep
  - 4. Niger
  - 5. Sierra Leone

Thank you