CLUSTERING ASSIGNMENT

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.

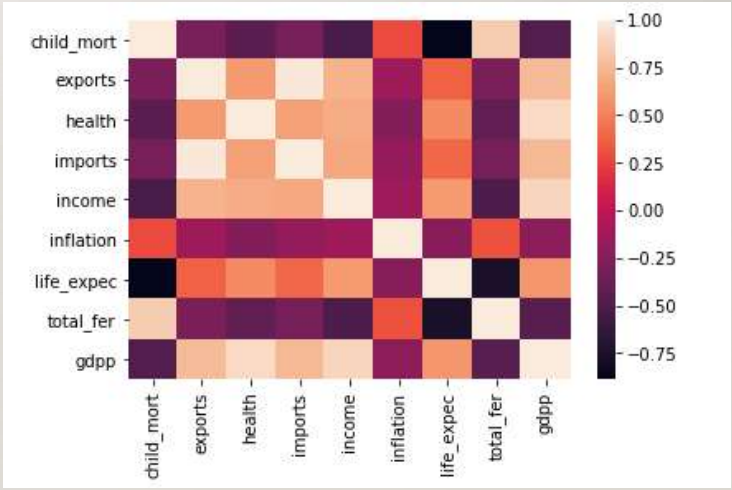
READING THE DATA AND CLEANING

1. We checked the data and remove "Names" column of countries because we needed numerical data for analysis initially
2. We didn't find null values so we didn't need to clean it

PREPROCESS DATA

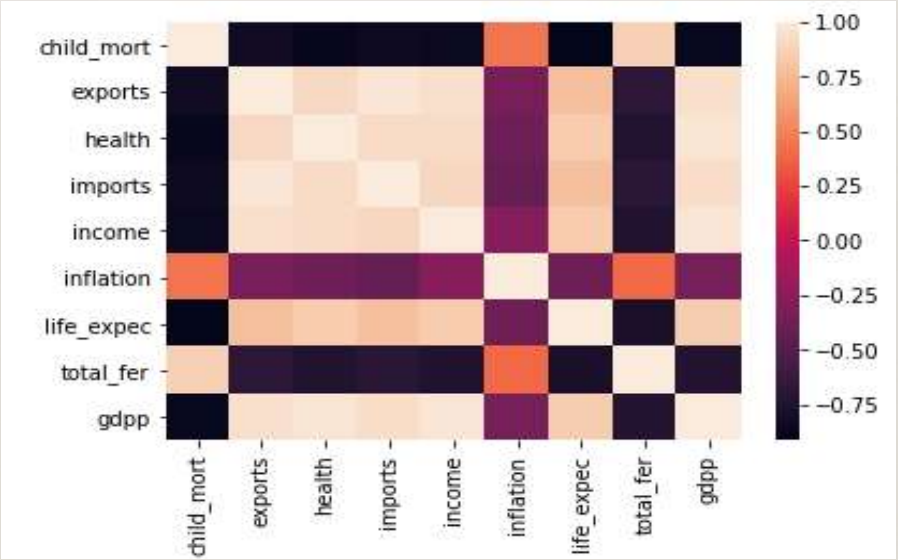
- 1.We Transformed the data using power transformer for better understanding of correlation and multicollinearity of the data
- Before transforming the data,we had skewed data and

Correlations were not that clear

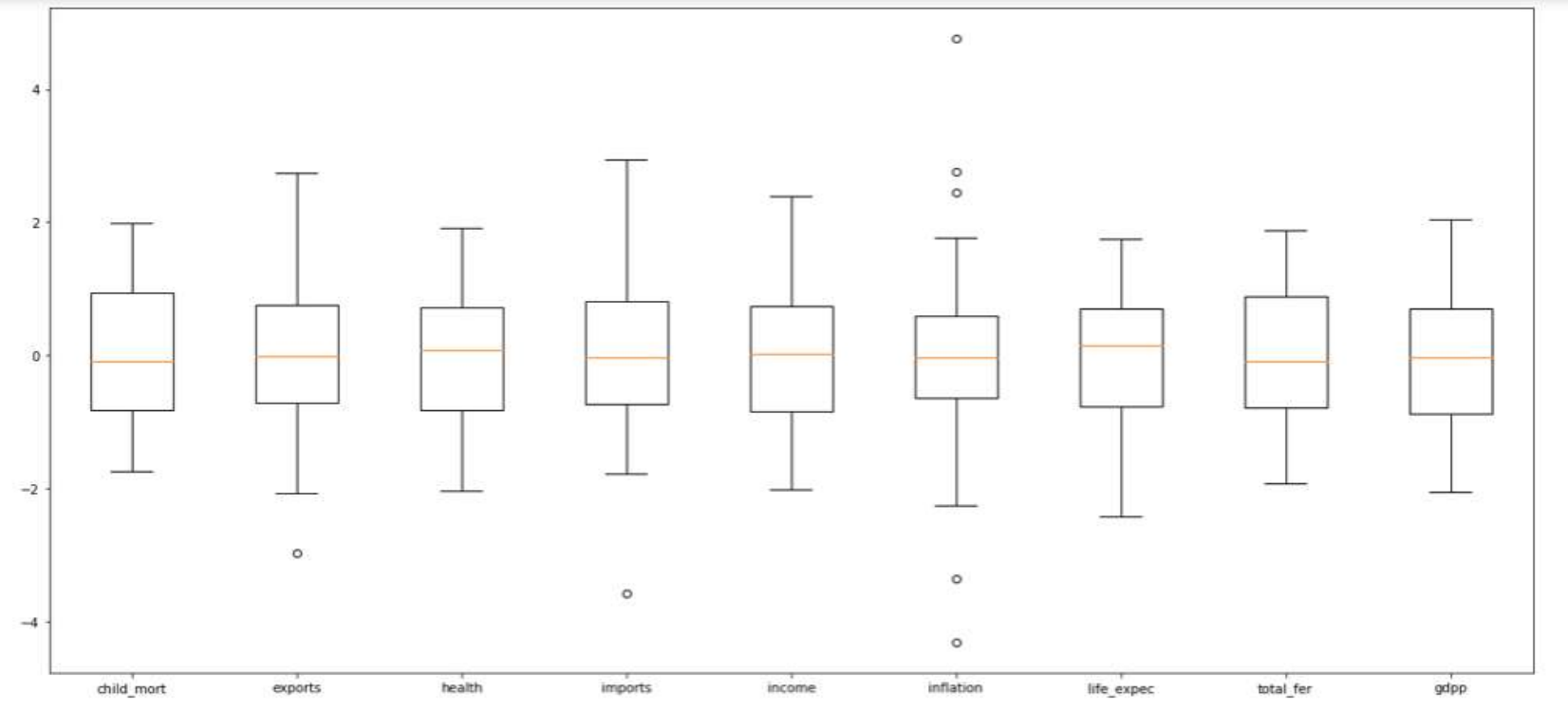


- After transforming the data we had below picture of correlation. We can see that correlation between features is much more clearer.

We validated it by checking skewness which moved closer to zero



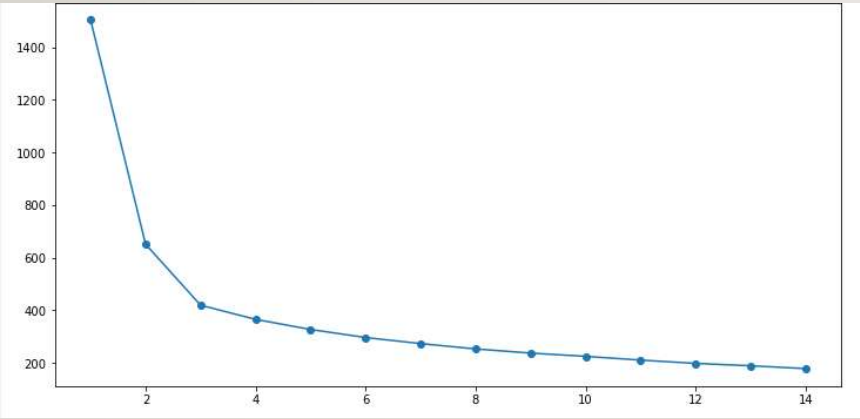
OUTLIERS AND OUTLIER TREATMENT



- There are few outliers in inflation,most of the features dont have outliers so we are not going to do any outlier treatment

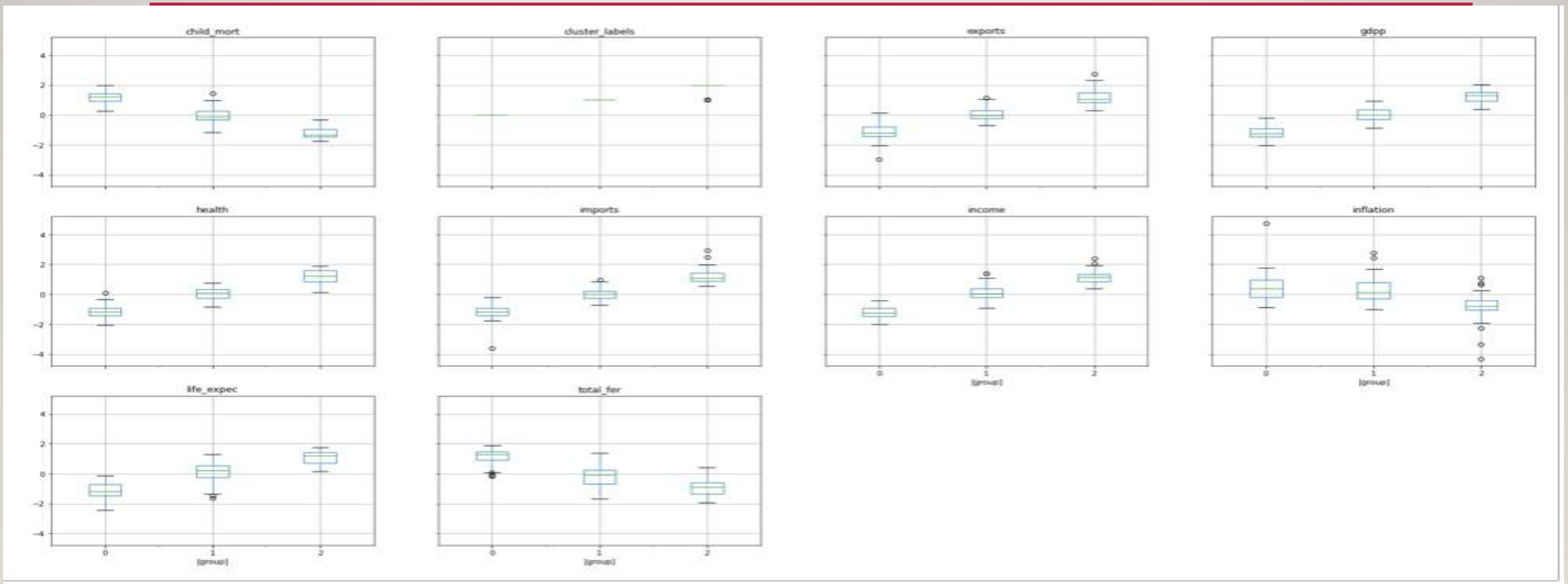
K-MEANS IMPLEMENTATION

- Common challeng with K means is that you need to know how much clusters to expect.Figuring out number of clusters is not obvious in data so we try different numbers and check their sillhoutte coefficient.
- The silhouette coefficient for a data point measures how similar it is to its assigned cluster from -1 (dissimilar) to 1 (similar).The elbow method can be used to determine the number of clusters as well.



- We Can see that there is lesser deviation from 2 to 3 so it looks like 3 clusters

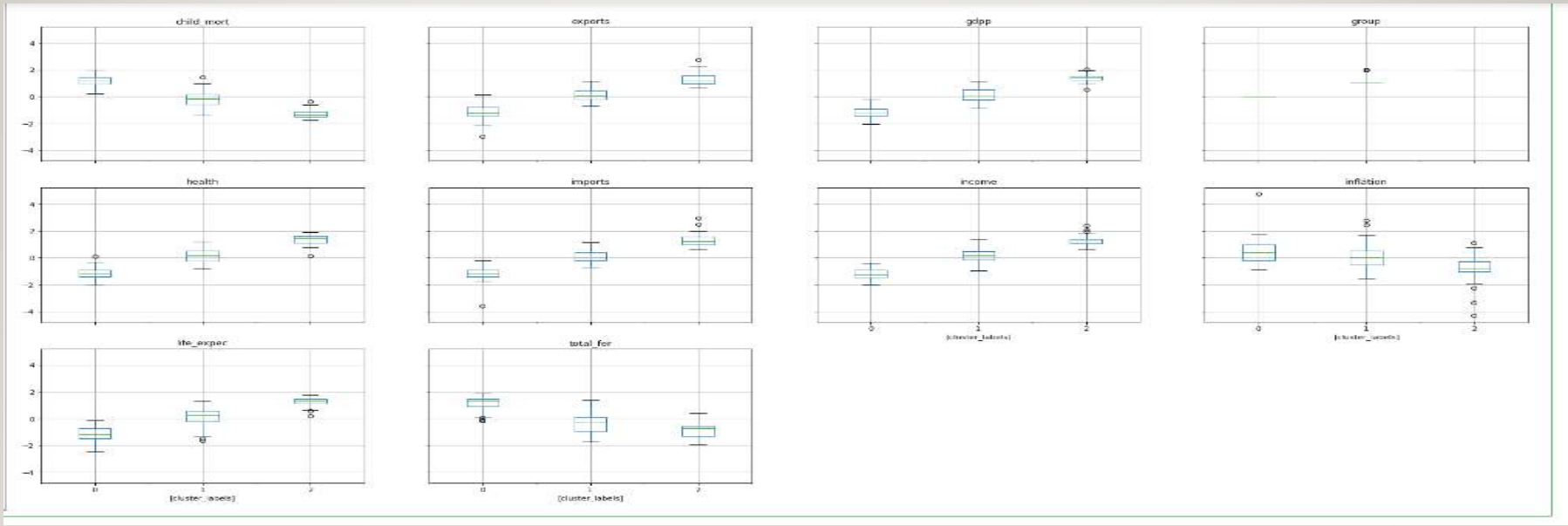
GRAPH AFTER CLUSTERING USING K-MEANS



SOME CONCLUSIONS FROM ABOVE BOX PLOTS

- 1.Group 0 has highest child mortality rate,highest imports,lowest exports,lowest gppp,lowest income, high inflation,low life expetancy and very high fetility rate
- 2.Group 1 scores better on all social indicators than group 0.There are developing countries
- 3.Group 2 scores best on all social indicators and it indicates that they are developed countries

USING CLUSTERING:



- Conlusions of using clustering are same as k means

CONCLUSION:
TOP 5 COUNTRIES WHICH NEED HELP ARE

- Barundi
- Liberia
- Congo, Dem. Rep.
- Niger
- Sierra Leone