Clustering Assignment

(Help International)

-Nistha Kumar

Table of Content

Problem Statement

Business Definition

→ HELP International, an International NGO aims for providing aid to the people of backward countries. They need to decide to use the funds effectively by helping the countries in direct needs.

→ Factors determining the financial state of a country:

- High Income, High GDPP, Low Child Mortality: Developed
- Medium Income, Medium GDPP, Medium Child Mortality: Developing
- Low Income, Low GDPP and High Child Mortality: Underdeveloped

Data Project definition

→ Create clusters of countries based on all the factors using 2 approach:

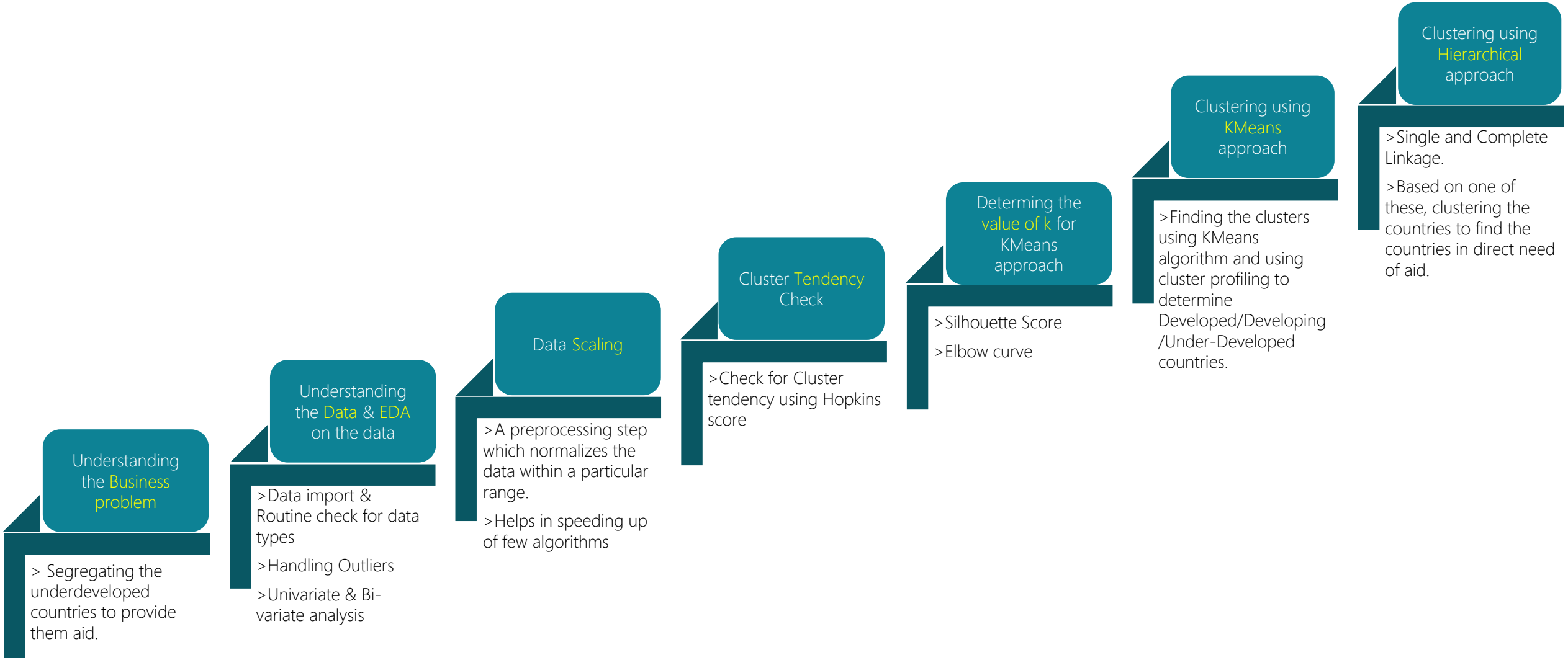
-K-Means

-Hierarchical Clustering

→ Using both the methods , create clusters of countries using all the parameters and then do cluster profiling using parameters: 'gdpp', 'income', and 'child_mort'.

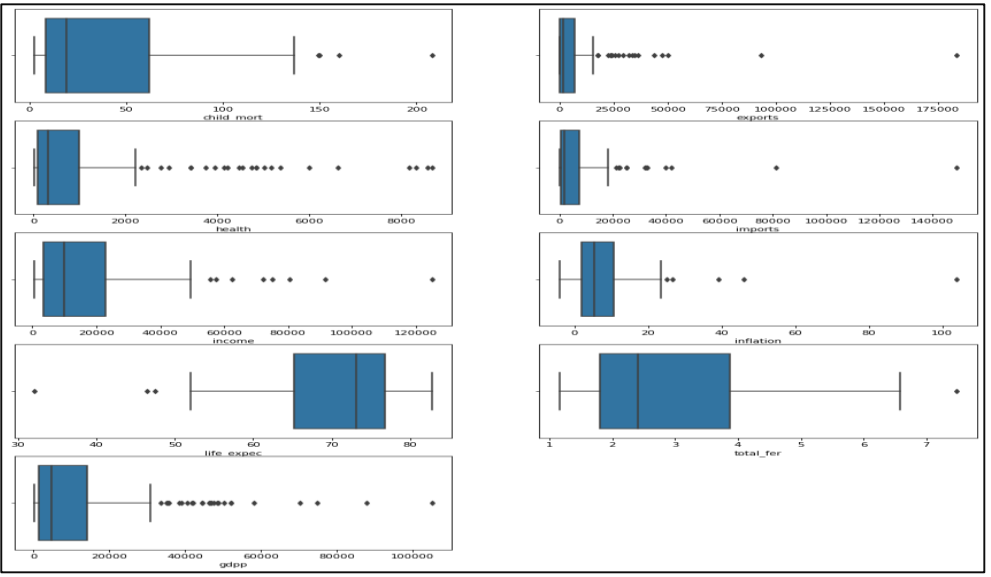
→ Find top 5 countries in need of aid based on any 1 clustering method used

Overall Approach



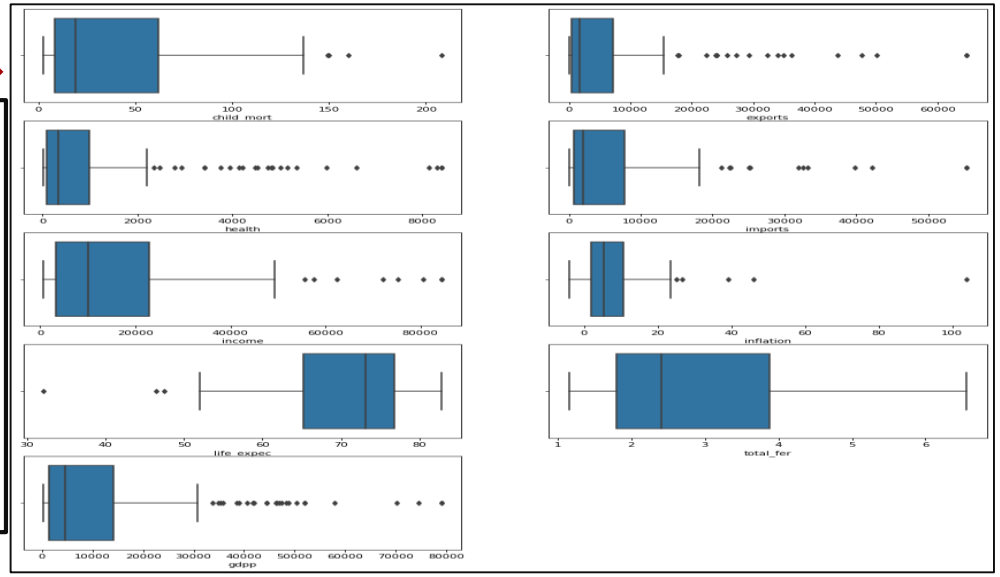
Data Preparation

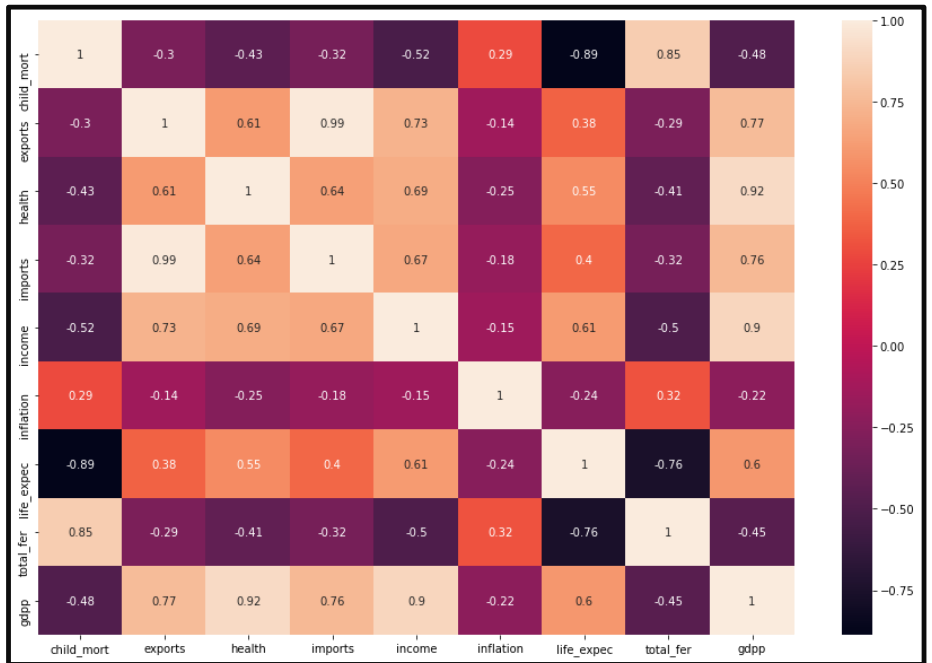
Outlier Treatment



After Outlier Treatment

- Used 'capping' method to treat the outliers(soft handling 99 percentile)
- Higher values of child_mort, Inflation and lower values for all other columns should not be handled as they are strong indicators for countries in need of aid.



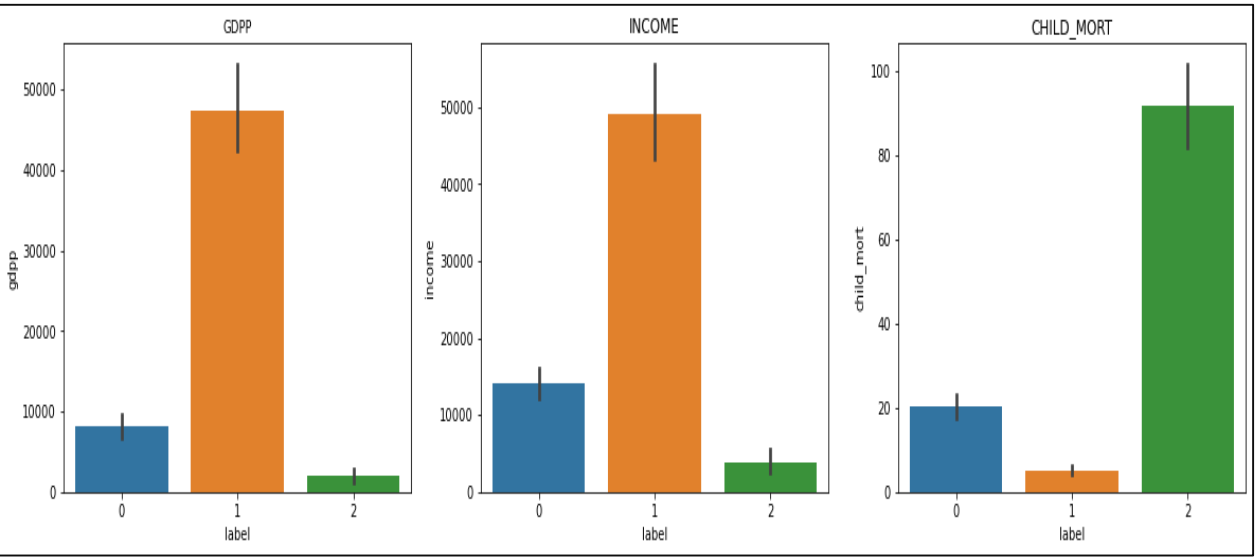


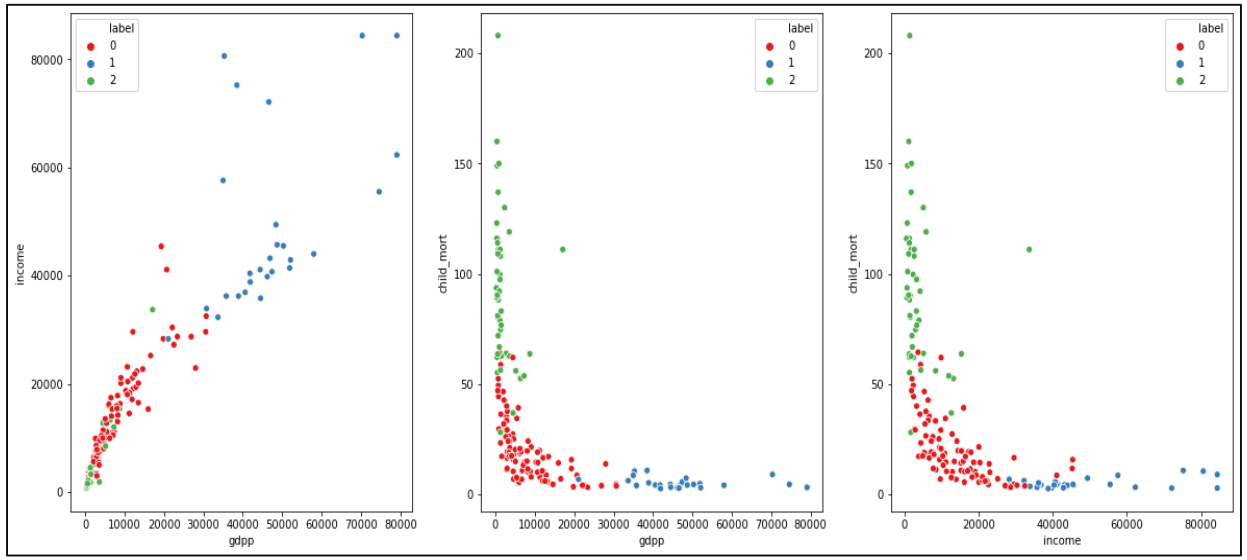
Insights From Correlation Heatmap

- Imports & Exports are highly correlated
- Health, Income, Exports, Imports are having high correlation with GDPP
- Child mortality rate is having high negative correlation with Life Expectancy
- Total Fertility is having high positive correlation with Child Mortality whereas has high negative correlation with Life Expectancy.

K-means Clustering

Based on Silhouette Score and Elbow curve method, value of k chosen is 3





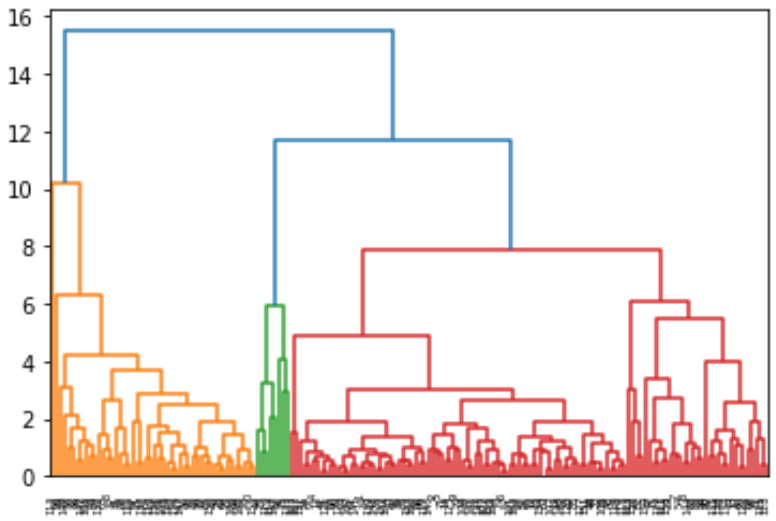
- **All developed countries are having high gdpp, high income and low child_mort (label 1)**
- **Developing countries have medium gdpp, medium income and medium child_mort (label 0)**
- **Under-developed countries have least gdpp, least income and highest child_mort (label 2)**

Underdeveloped Countries:
Burundi
Liberia
Congo, Dem. Rep.
Niger
Sierra Leone

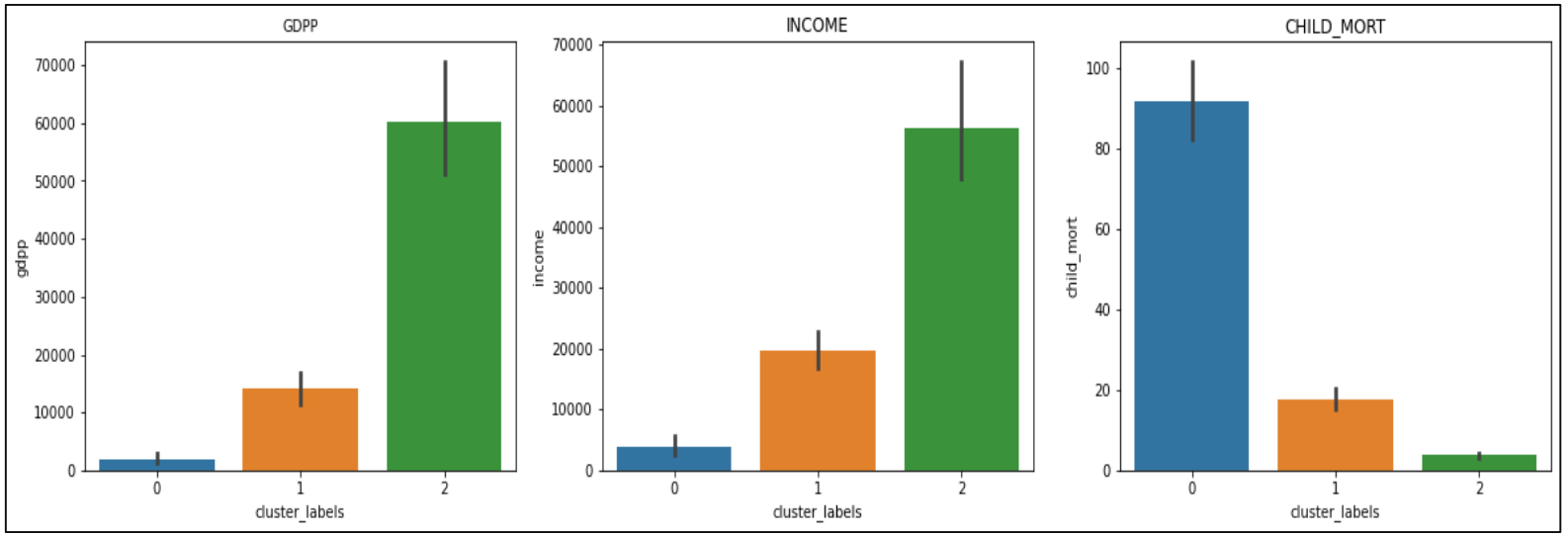
Developing Countries:
Nepal
Tajikistan
Bangladesh
Cambodia
Kyrgyz Republic

Developed Countries:
Malta
Cyprus
New Zealand
United Arab Emirates
Brunei

Hierarchical Clustering



Complete Linkage Dendogram
- Number of clusters selected is 3



- **All developed countries are having high gdpp, income and low child_mort (label 2)**
- **Developing countries have medium gdpp, income and medium child_mort (label 1)**
- **Under-developed countries have least gdpp, income and highest child_mort (label 0)**

Underdeveloped Countries:
Burundi
Liberia
Congo, Dem. Rep.
Niger
Sierra Leone

Developing Countries:
Nepal
Tajikistan
Bangladesh
Cambodia
Kyrgyz Republic

Developed Countries:
Belgium
Singapore
Ireland
Netherlands
Denmark

Conclusion

- After comparing both K-Means & Hierarchical, I'm going ahead with K-Means as the clustering of data is more uniform(balanced data).
- After grouping the data based on socio-economic and health factors, we can determine the overall development of the countries and hence find the countries in direst need of aid.

Countries in need of aid

1. Afghanistan
2. Angola
3. Benin
4. Botswana
5. Burkina Faso
6. Burundi
7. Cameroon
8. Central African Republic
9. Chad
10. Comoros
11. Congo, Dem. Rep.
12. Congo, Rep.
13. Cote d'Ivoire
14. Equatorial Guinea
15. Eritrea
16. Gabon
17. Gambia
18. Ghana
19. Guinea
20. Guinea-Bissau
21. Haiti
22. Iraq
23. Kenya
24. Kiribati
25. Lao
26. Lesotha
27. Liberia
28. Madagascar
29. Malawi
30. Mali
31. Mauritania
32. Mozambique
33. Namibia
34. Niger
35. Nigeria
36. Pakistan
37. Rwanda
38. Senegal
39. Sierra Leone
40. Solomon Islands
41. South Africa
42. Sudan
43. Tanzania
44. Timor-Leste
45. Togo
46. Uganda
47. Yemen
48. Zambia

Top 5 countries in need of aid (most under-developed countries):

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

Thank You