



Clustering Assignment

(Help International)

-Nistha Kumar

Table of Content

1 Problem statement

2 Overall Approach

3 Data Preparation

4 Kmeans Clustering

5 Hierarchical Clustering

6 Conclusion

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

Understanding the Business problem

- > Segregating the underdeveloped countries to provide them aid.

Understanding the Data & EDA on the data

- > Data import & Routine check for data types
- > Handling Outliers
- > Univariate & Bi-variate analysis

Data Scaling

- > A preprocessing step which normalizes the data within a particular range.
- > Helps in speeding up of few algorithms

Cluster Tendency Check

- > Check for Cluster tendency using Hopkins score

Determining the value of k for KMeans approach

- > Silhouette Score
- > Elbow curve

Clustering using KMeans approach

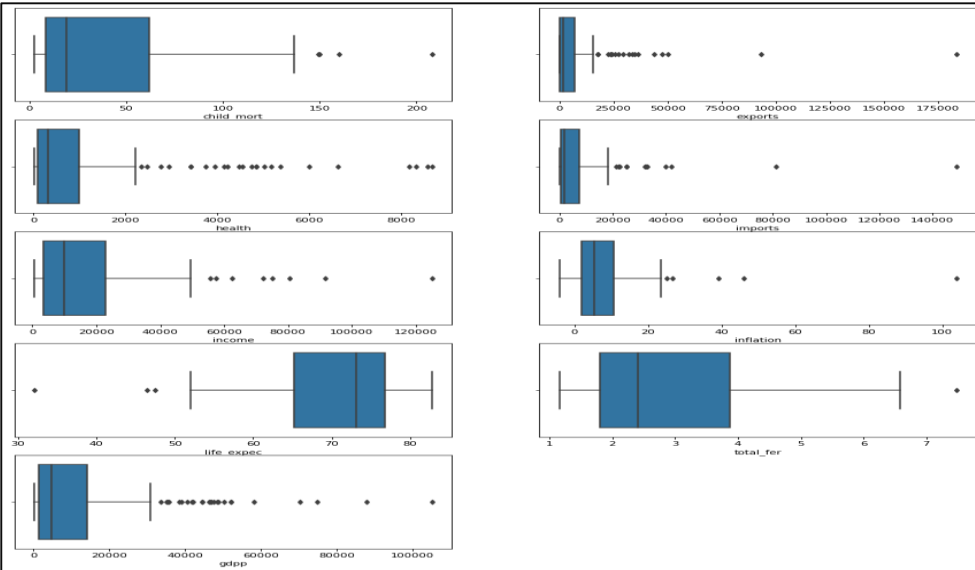
- > Finding the clusters using KMeans algorithm and using cluster profiling to determine Developed/Developing /Under-Developed countries.

Clustering using Hierarchical approach

- > Single and Complete Linkage.
- > Based on one of these, clustering the countries to find the countries in direct need of aid.

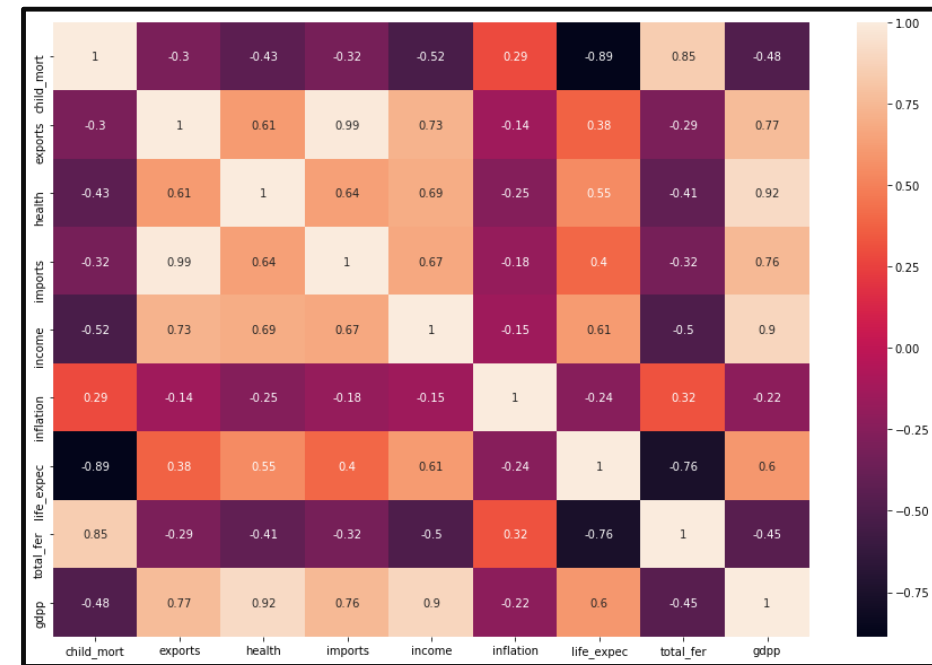
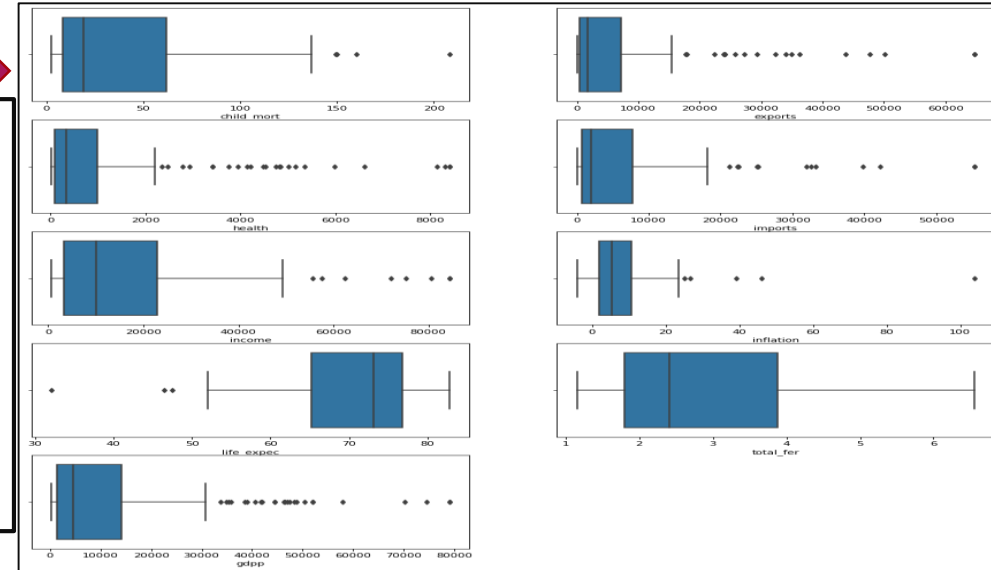
Data Preparation

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.

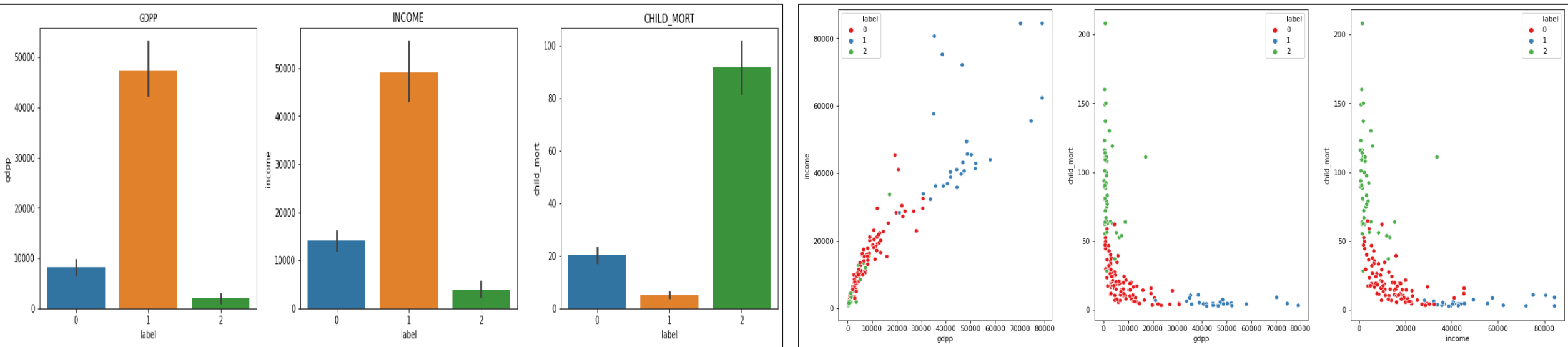


Insights From Correlation Heatmap

- 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 gdp, high income and low child_mort (label 1)
- Developing countries have medium gdp, medium income and medium child_mort (label 0)
- Under-developed countries have least gdp, least income and highest child_mort (label 2)

Underdeveloped Countries:

Burundi
Liberia
Congo, Dem. Rep.
Niger
Sierra Leone

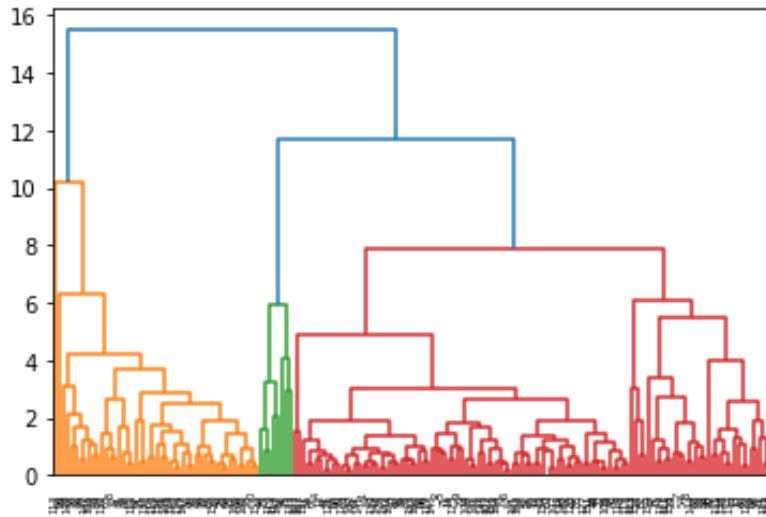
Developing Countries:

Nepal
Tajikistan
Bangladesh
Cambodia
Kyrgyz Republic

Developed Countries:

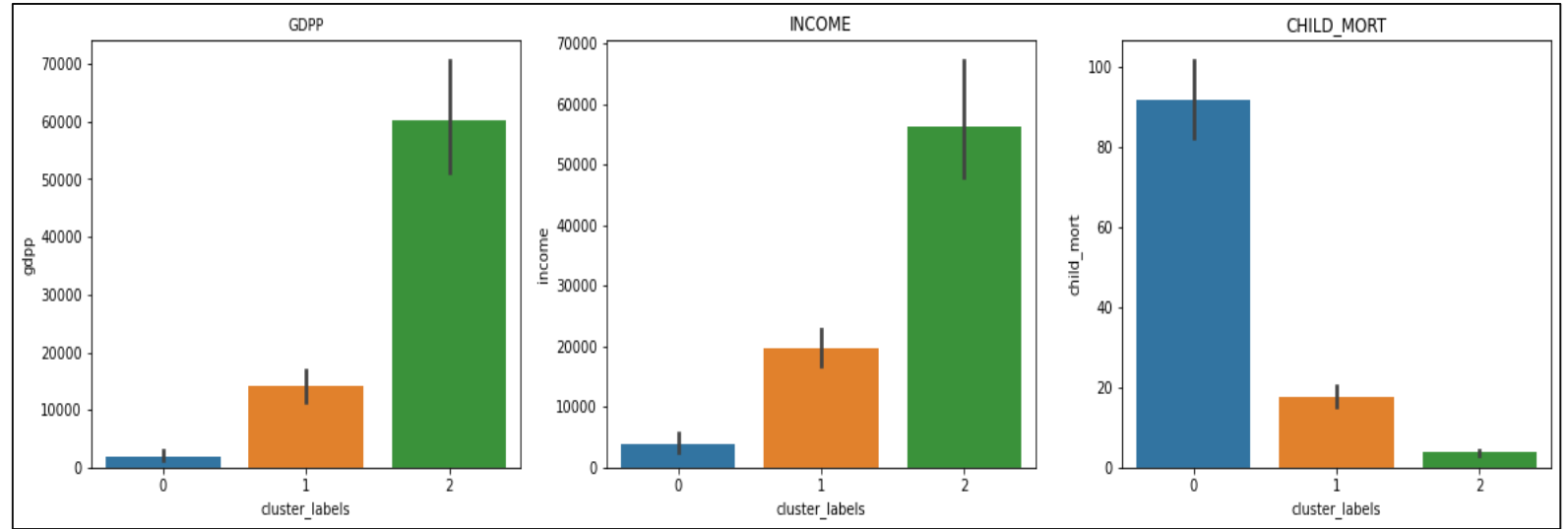
Malta
Cyprus
New Zealand
United Arab Emirates
Brunei

Hierarchical Clustering.



Complete Linkage Dendrogram

- Number of clusters selected is 3



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

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

Countries in need of aid

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