Help International

Identifying countries in dire need of financial aid.

Anand Yati

Problem Statement

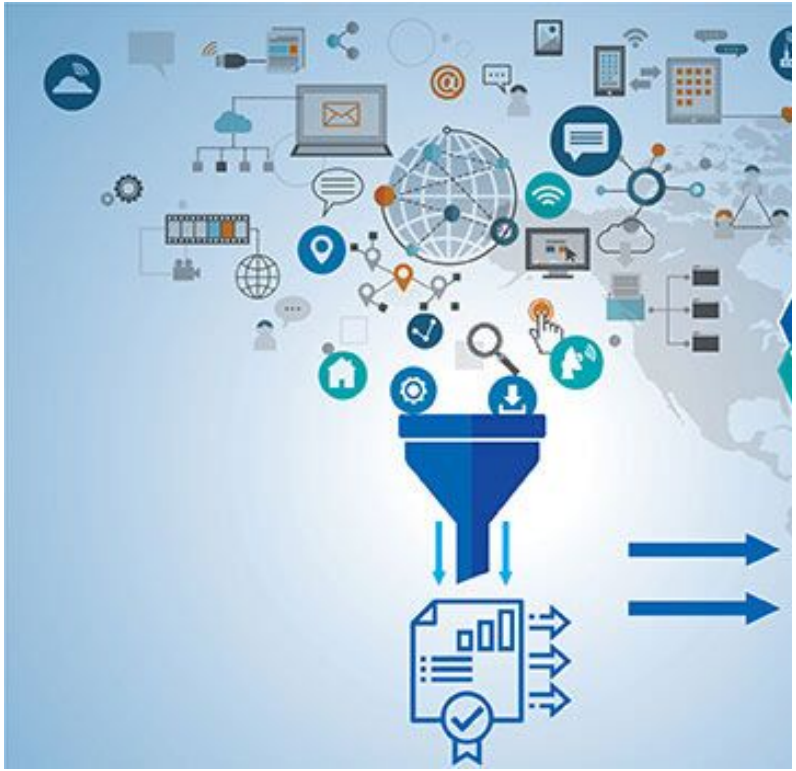
Quantitatively identify countries in dire need of financial aid based on an array of socio-economic indicators. These indicators should be used to create an overall development profile of the country which is then used to segregate countries into a cluster of nations needing assistance.

The output of this exercise would be a prioritised list of countries which can be directly consumed by decision makers of the NGO (CEO, executive audience) to effectively and strategically distribute financial aid.

Solution approach [Steps involved]

1. Procure the data, understand different columns, data types, value ranges etc.
2. Perform exploratory data analysis
3. Clean data:
    a. Transforming columns which were % of GDP to absolute values.
    b. Dropping unnecessary columns (% of GDP columns after obtaining corresponding absolute values)
    c. Missing value treatment (not required in this case)
    d. Data type conversions/encoding (not required in this case)
4. Prepare data for modeling
    a. Outlier treatment (performed)
    b. Scaling (Standard & Min-Max)
5. Clustering:
    a. K means: Choosing K through elbow curve method & Silhouette score method. Clustering and country profiling
    b. Hierarchical: Choosing single vs complete linkage method, cutting tree into clusters and country profiling.
6. Choosing the cluster in most dire need of aid through analysing country profiles based on: income, gdpp, child mortality.
7. Prioritising countries inside the cluster which is in dire need of aid. Prioritisation done based on variables in step 6.
8. Results summarisation and presentation.

Data set



Context: Data on socio-economic indicators of countries.

Variables provided: Country, child mortality per 1000 births, exports/imports/health-spends as a % of GDP, income/person, life expectancy, fertility rate, inflation, GDP per person

Raw dataset:

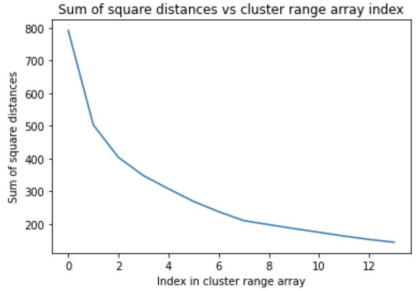
- Columns: 10
- Rows: 167

Data treatments required across columns:

- Some variables (like health spend) are given as % of GDP (another column in dataset) and needs to be converted to absolute values.
- Outlier imputation to inter-quartile range
- Scaling of columns (standard scalar & min-max scalars used)

Clustering results: Number of clusters

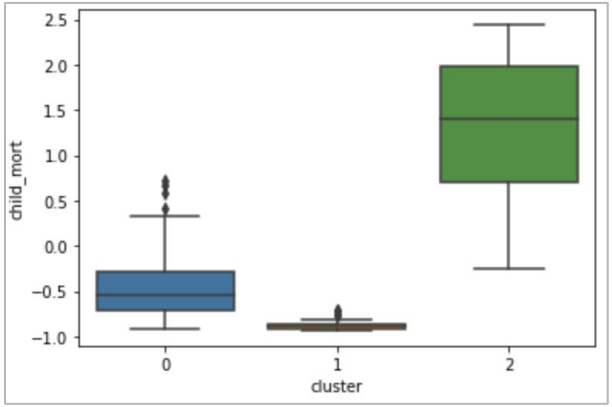
```
ssd = []
cluster_range = [2,3,4,5,6,7,8,9,10,11,12,13,14,15]
for num_clusters in cluster_range:
  kmeans = KMeans(n_clusters=num_clusters, verbose=0,max_iter=50)
  kmeans.fit(country_data_scaled[numeric_columns])
  ssd.append(kmeans.inertia_)

plt.plot(ssd)
plt.xlabel("Index in cluster range array")
plt.ylabel("Sum of square distances")
plt.title("Sum of square distances vs cluster range array index")
plt.show()
```



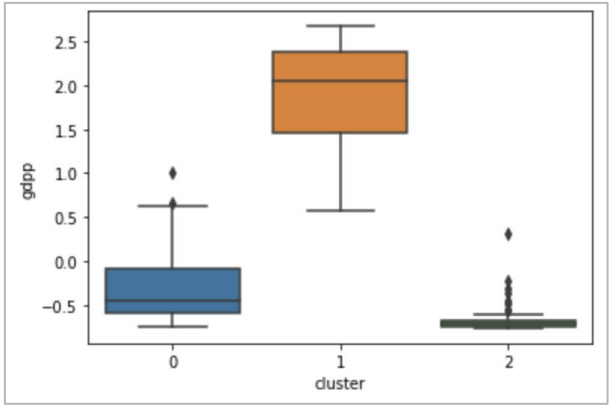

```
For 2 clusters, Silhouette score is: 0.4834991457405332
For 3 clusters, Silhouette score is: 0.41278062890648703
For 4 clusters, Silhouette score is: 0.39594880315682435
For 5 clusters, Silhouette score is: 0.37940355788531294
For 6 clusters, Silhouette score is: 0.29479892228279037
For 7 clusters, Silhouette score is: 0.3208016400988546
For 8 clusters, Silhouette score is: 0.33822235751873186
For 9 clusters, Silhouette score is: 0.31806476709521836
For 10 clusters, Silhouette score is: 0.30391255110072674
For 11 clusters, Silhouette score is: 0.284487682519972
For 12 clusters, Silhouette score is: 0.27643540740556516
For 13 clusters, Silhouette score is: 0.2775904774857871
For 14 clusters, Silhouette score is: 0.26905498259746435
For 15 clusters, Silhouette score is: 0.2718071481688423
```

Elbow curve method points towards 3 being a good cluster count to choose, a high Silhouette score for cluster size 3 confirms it as a choice for number of clusters.

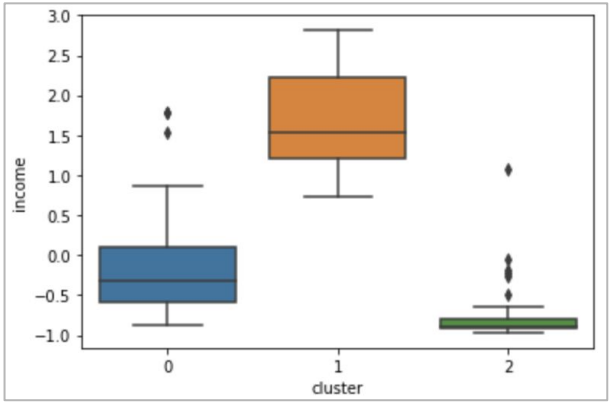
Cluster comparisons: K Means [Univariate]



Child mortality across clusters



GDP per person across clusters

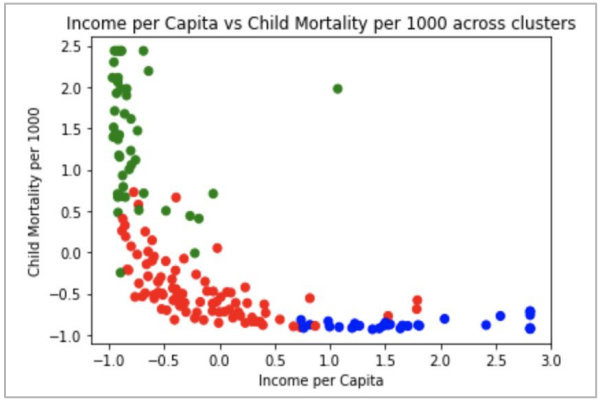


Income per person across clusters

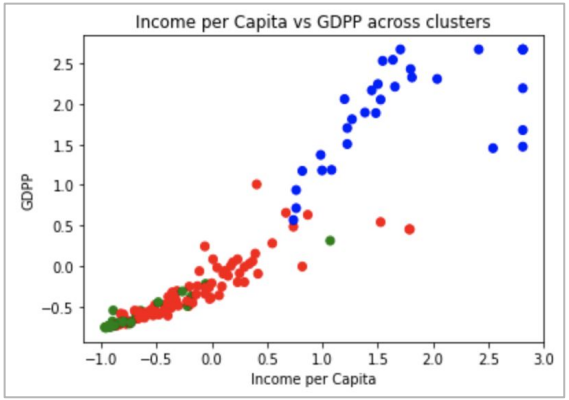
Insight: Cluster 2 has highest child mortality rate along with lowest GDP and Income per person.

Thus Cluster 2 countries are prime candidates for financial aid.

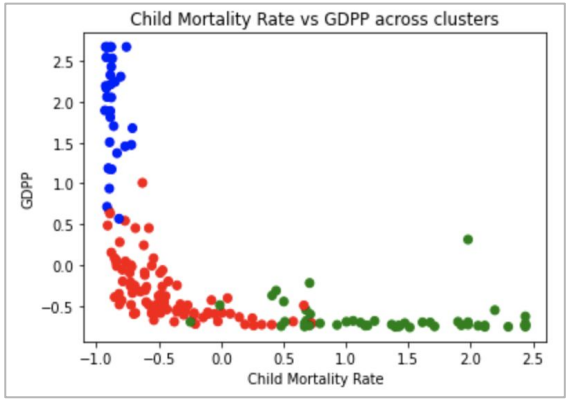
Cluster comparisons: K Means [Bivariate]



Child mortality Vs. Income per person across clusters



Income per Capita vs GDPP across clusters

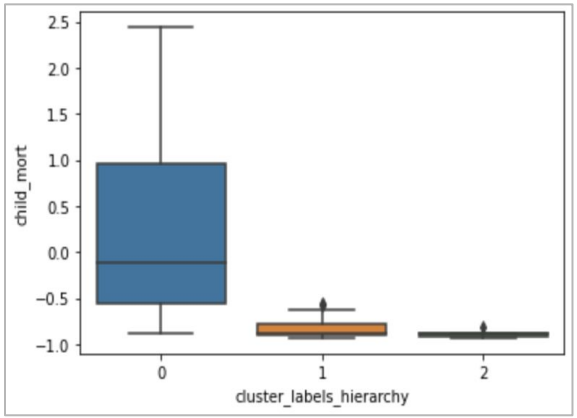


Child Mortality Rate vs GDPP across clusters

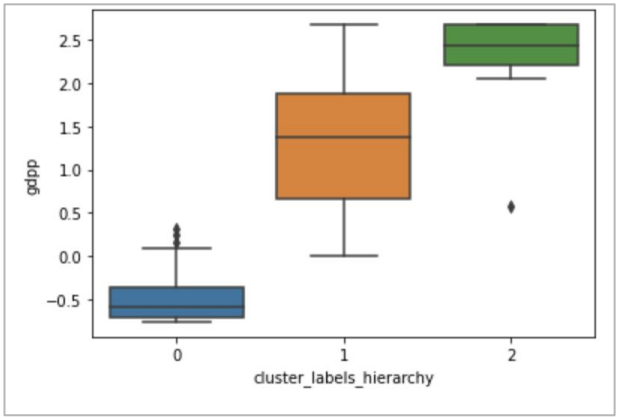
Insight: Cluster 2 [in Green color] has highest child mortality along with lowest income/GDP per capita

Thus Cluster 2 countries are prime candidates for financial aid.

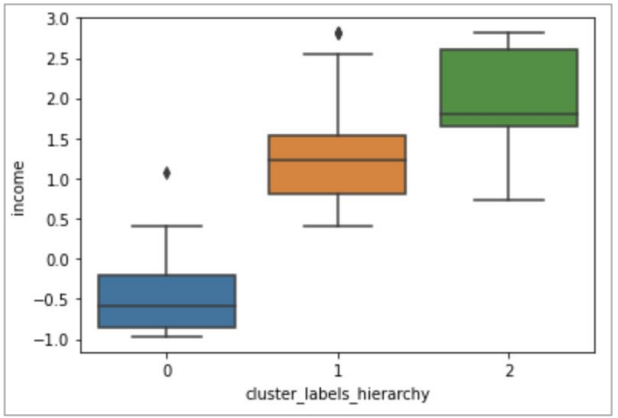
Cluster comparisons: Hierarchical [Univariate]



Child mortality across clusters



GDP per person across clusters

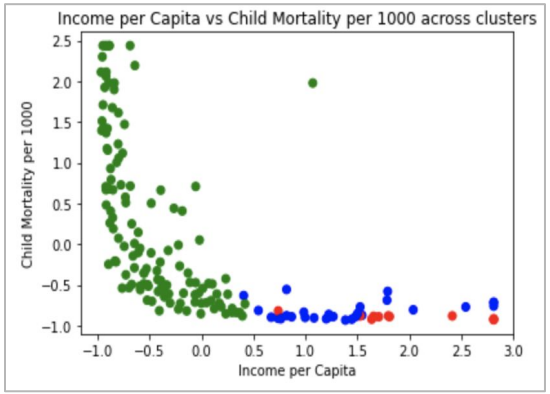


Income per person across clusters

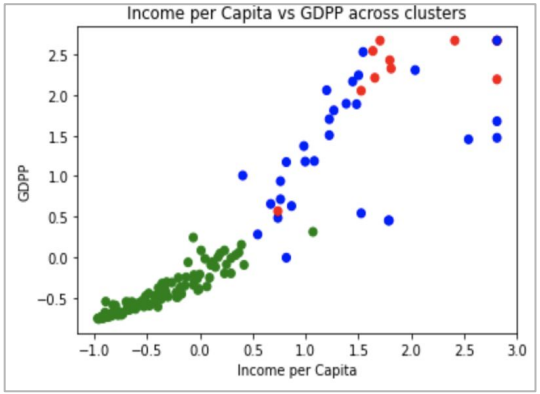
Insight: Cluster 0 has highest child mortality rate along with lowest GDP and Income per person. Thus Cluster 0 countries are prime candidates for financial aid.

Note: Although same countries are clustered together across KMeans and Hierarchical clustering approaches, but there cluster label differs as it changes across iterations randomly.

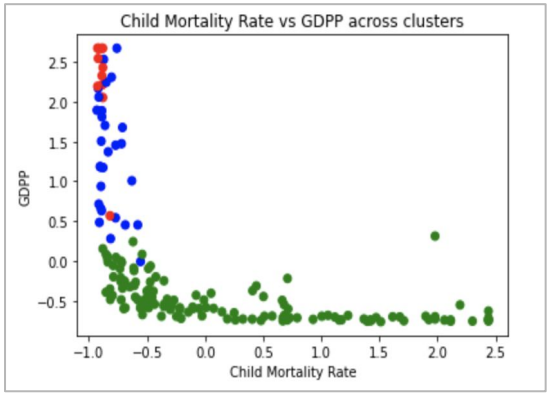
Cluster comparisons: K Means [Bivariate]



Child mortality Vs. Income per person across clusters



Income per Capita vs GDPP across clusters



Child Mortality Rate vs GDPP across clusters

Insight: Cluster in Green color has highest child mortality along with lowest income/GDP per capita

Thus Cluster 2 countries are prime candidates for financial aid.

Filtering the countries from target cluster and prioritising them

Countries from Green cluster (cluster labeled as 2 in K-Means and 0 in Hierarchical clustering) were extracted and a composite metric was calculated for ranking of countries within cluster.

Composite metric = min_max_scaled_values ((child mortality + (1-GDPP) + (1-income))

Notes:

- Min_max scaling was done to compress outlier treated data into a uniform range of 0 to 1, this eliminated effect of a higher range of values inside a column and made all 3 variables (child mortality, GDPP & income) uniform.
- GDPP and Income are subtracted from 1 before adding as they are inversely proportional with child mortality.

Final prioritised list of top 10 countries in dire need of financial aid

Countries	Prioritisation score
Central African Republic	2.98
Sierra Leone	2.97
Haiti	2.947
Niger	2.945
Mali	2.93
Chad	2.92
Congo Dem. Rep	2.9
Burkina Faso	2.86
Guinea-Bissau	2.84
Guinea	2.80

Insights:

The list of top 10 countries in need of aid is same from both "kMeans" and "Hierarchical" clustering methods.

The high-degree of overlap in countries in dire need of aid showcases robust clustering through both kMeans and Hierarchical methods

Top 25 countries requiring aid from 2 clustering methods

K - Means Clustering

	country	child_mort	income	gdpp	cluster	composite_metrics
0	Central African Republic	1.000000	0.008431	0.012745	2	2.978823
1	Sierra Leone	1.000000	0.018464	0.009959	2	2.971577
2	Haiti	1.000000	0.026926	0.025550	2	2.947524
3	Niger	0.948929	0.006195	0.006936	2	2.935798
4	Mali	1.000000	0.038107	0.028277	2	2.933616
5	Chad	1.000000	0.039920	0.039481	2	2.920599
6	Congo, Dem. Rep.	0.878934	0.000000	0.006106	2	2.872829
7	Burkina Faso	0.878934	0.024810	0.020392	2	2.833732
8	Guinea-Bissau	0.858936	0.023602	0.018733	2	2.816602
9	Guinea	0.808940	0.017558	0.024720	2	2.766662
10	Benin	0.828938	0.036596	0.031241	2	2.761101
11	Nigeria	1.000000	0.137228	0.124429	2	2.738343
12	Mozambique	0.728946	0.009338	0.011145	2	2.708463
13	Cote d'Ivoire	0.828938	0.062887	0.058628	2	2.707423
14	Cameroon	0.798940	0.061981	0.063963	2	2.672996
15	Burundi	0.654951	0.004684	0.000000	2	2.650267
16	Lesotho	0.715947	0.053519	0.055664	2	2.606763
17	Liberia	0.611954	0.002750	0.005691	2	2.603513
18	Malawi	0.623953	0.012722	0.013516	2	2.597715
19	Togo	0.621954	0.018162	0.015235	2	2.588557
20	Afghanistan	0.620954	0.030250	0.019088	2	2.571616
21	Mauritania	0.692948	0.081926	0.057443	2	2.553580
22	Angola	0.908932	0.159892	0.195566	2	2.553474
23	Comoros	0.600955	0.024206	0.031893	2	2.544856
24	Pakistan	0.639952	0.110937	0.047958	2	2.481058

Hierarchical Clustering [Cluster: k-means cluster number, cluster_labels_hierarchy: Hierarchical clustering labels]

	country	child_mort	income	gdpp	cluster	cluster_labels_hierarchy	composite_metrics
0	Central African Republic	1.000000	0.008431	0.012745	4	0	2.978823
1	Sierra Leone	1.000000	0.018464	0.009959	4	0	2.971577
2	Haiti	1.000000	0.026926	0.025550	4	0	2.947524
3	Niger	0.958680	0.006195	0.006936	4	0	2.945549
4	Mali	1.000000	0.038107	0.028277	4	0	2.933616
5	Chad	1.000000	0.039920	0.039481	4	0	2.920599
6	Congo, Dem. Rep.	0.902049	0.000000	0.006106	0	0	2.895943
7	Burkina Faso	0.902049	0.024810	0.020392	4	0	2.856846
8	Guinea-Bissau	0.885869	0.023602	0.018733	4	0	2.843535
9	Guinea	0.845418	0.017558	0.024720	4	0	2.803141
10	Benin	0.861599	0.036596	0.031241	4	0	2.793762
11	Mozambique	0.780697	0.009338	0.011145	4	0	2.760215
12	Cote d'Ivoire	0.861599	0.062887	0.058628	4	0	2.740083
13	Nigeria	1.000000	0.137228	0.124429	0	0	2.738343
14	Burundi	0.720830	0.004684	0.000000	4	0	2.716146
15	Cameroon	0.837328	0.061981	0.063963	4	0	2.711384
16	Liberia	0.686043	0.002750	0.005691	4	0	2.677602
17	Malawi	0.695751	0.012722	0.013516	4	0	2.669512
18	Lesotho	0.770180	0.053519	0.055664	4	0	2.660997
19	Togo	0.694133	0.018162	0.015235	4	0	2.660736
20	Afghanistan	0.693324	0.030250	0.019088	4	0	2.643986
21	Comoros	0.677144	0.024206	0.031893	4	0	2.621045
22	Mauritania	0.751573	0.081926	0.057443	0	0	2.612205
23	Angola	0.926320	0.159892	0.195566	0	0	2.570861
24	Uganda	0.618895	0.028135	0.021578	4	0	2.569182