Solution for clustering the countries

HELP International

RECOMMENDATION

Objectives

1. The task at hand is to cluster the countries by the factors above and then present the solution and recommendations.
2. Categorizing the countries using socio-economic and health factors that determine the overall development of the country.
3. The datasets provided contains socio-economic factors and the corresponding data dictionary.

Methodology implemented

1. K- means clustering and hierarchical clustering is done on the given data of 197 countries, in order to get the final set of countries that need aid.
2. As a first step, an exploratory data analysis is done on the dataset to figure out the trends in a sorted order. To select the final cluster of countries, we can use ‘gdpp’, ‘income’ and ‘child_mort’ as 3 main factors that are distinguishable.
3. Outliers in the data generally seen are the ones that belong to (high income, high gdpp and low child_mort) cluster. They’re joined to the data using a method called outlier capping.
4. The optimal number of cluster is chosen to be 3, using two statistical methods known as elbow curve method, and silhouette analysis.
5. Finally clusters are formed using K- means and hierarchical clustering to obtain the results as shown

Hopkins statistic

1. The Hopkins statistic is a way of measuring the cluster tendency of a data set.
2. A value close to 1 tends to indicate the data is highly clustered, random data will tend to result in values around 0.5, and uniformly distributed data will tend to result in values close to 0.
3. **Our cluster** model when checked for Hopkins test value gave a significant 0.85 – 0.90 value, which conveys us that the cluster tendency is high

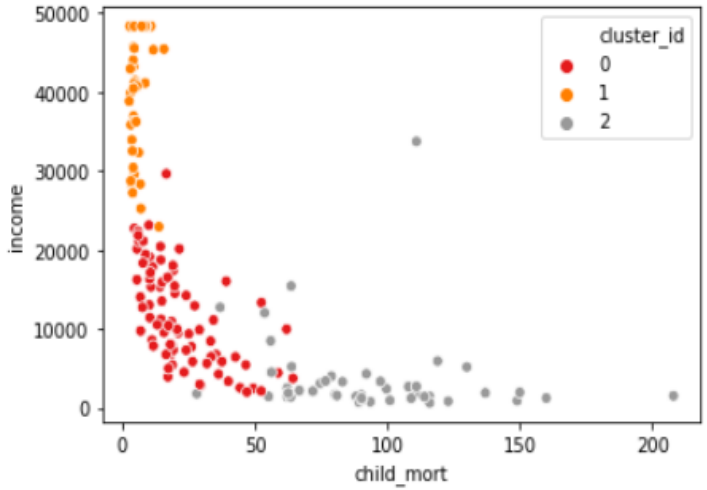
Finding the optimal number of clusters

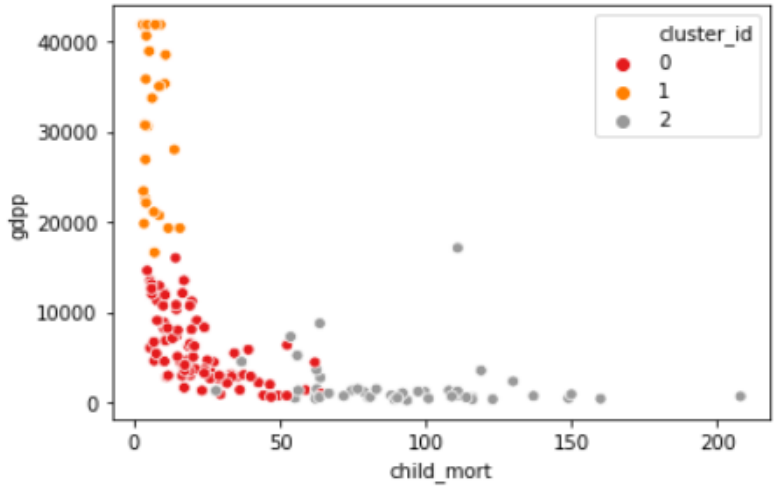
1. Finding the optimal number of clusters is done in two ways here - One is SSD or elbow curve method. The other is silhouette analysis. Elbow curve is minimizing intra cluster distance to maximizing inter - cluster distance, while silhouette analysis gives a value that measures the effectiveness of a given cluster

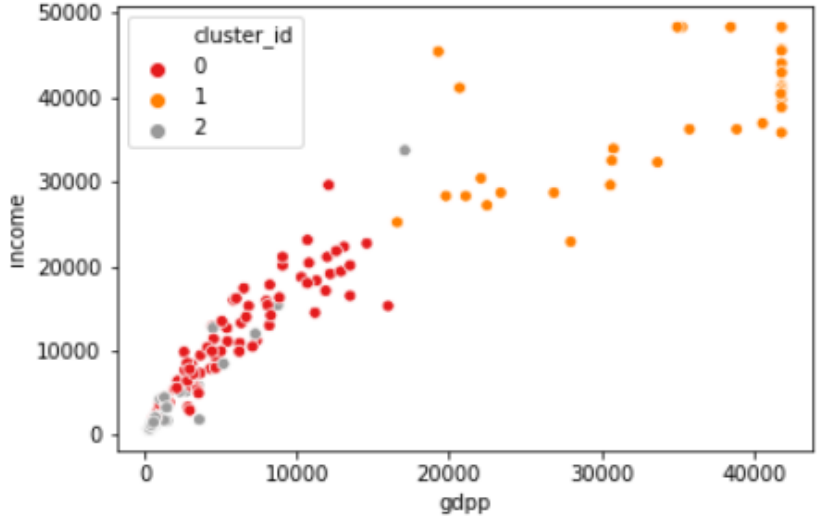
2. The elbow curve forms an 'elbow' at number of clusters = 3). The picked value here is where the elbow forms. The elbow is also formed at number of clusters = 5) but there is a slight decreasing trend at 5, but higher decreasing trend at clusters = 3.

3. Silhouette analysis tells the effectiveness. Higher the value, higher is the formation of effective clusters. We see that the silhouette score is higher when the number of clusters = 2. Second in the line is 3 and 4, which have almost similar scores. But choosing 2 clusters is generally not preferred by business problems, and hence we choose the number of clusters as 3.

Results







The clusters as depicted by legends are 0,1 and 2. In this graph, the countries which are needed to be considered for aid belongs to cluster 2.

Results

	country	child_mort	exports	health	imports	income	inflation	life_expec	total_fer	gdpp	cluster_id	hcluster_labels
37	Congo, Dem. Rep.	116.0	137.2740	26.4194	165.664	609.0	20.80	57.5	5.861	334	2	0
88	Liberia	89.3	62.4570	38.5860	302.802	700.0	5.47	60.8	5.020	327	2	0
26	Burundi	93.6	20.6052	26.7960	90.552	764.0	12.30	57.7	5.861	231	2	0
112	Niger	123.0	77.2560	17.9568	170.868	814.0	2.55	58.8	5.861	348	2	0
31	Central African Republic	149.0	52.6280	17.7508	118.190	888.0	2.01	47.5	5.210	446	2	0

1. Above are the 5 countries in need of aid, with high child_mort, low income and low gdpp.
2. The cluster_id stands for the cluster designated by K-means clustering, and the hcluster_labels represent the clusters designated by hierarchical clustering