VSB TECHNICAL UNIVERSITY OF OSTRAVA | FACULTY OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE | DEPARTMENT OF COMPUTER SCIENCE

Fundamentals of Machine Learning

Clustering

Jan Platos

November 15, 2023

Clustering

Clustering

- Clustering techniques apply when there is no class to be predicted.
- The instances are to be divided into natural groups.
- These clusters reflect some mechanism from which instances are drawn.
- A mechanism that causes some instances to bear a stronger resemblance to each other than they do to the remaining instances.
- Clustering naturally requires different techniques than classification.

Clustering

- Clustering produces groups of objects.
- The groups are exclusive or overlapping.
- Assignment to the group may be probabilistic (fuzzy).
- Groups may be hierarchical.
- Most clusters are based on similarity among the the objects represented by distance or similarity function.

Clustering - Distances and Similarity

- How far apart are the objects?
- How close together are the objects?

Clustering - Distances and Similarity

- Distance usually refers to a category of functions that measures difference in a Cartesian space.
- The smaller distances the more closer the objects are.
- To maintain basic properties of the coordinate system, we expect the function to be a Metric:
  1. $d(x, y) \geq 0$ (non-negativity)
  2. $d(x, y) = 0 \Leftrightarrow x = y$ (identity)
  3. $d(x, y) = d(y, x)$ (symmetry)
  4. $d(x, y) \leq d(x, z) + d(z, y)$ (triangle inequality)

Clustering - Distances and Similarity

- Manhattan distance (Taxi Driver/City Block)

$$d(X, Y) = \sum_{i=1}^{N} |x_i - y_i|$$

- Euclidean distance

$$d(X, Y) = \sqrt{\sum_{i=1}^{N} (x_i - y_i)^2}$$

Clustering - Distances and Similarity

- Similarity measures the closeness between objects.
- The higher values the more closer the objects are.
- Similarity does not require Cartesian coordinates.
- Object-related similarity measures may be defined.

Clustering - Distances and Similarity

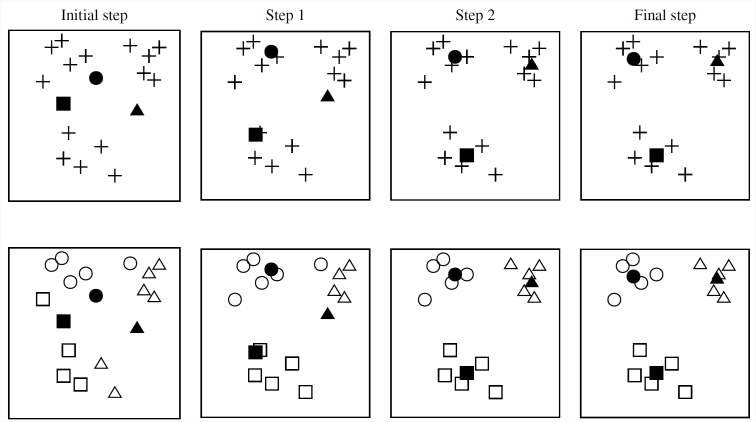
- Cosine Similarity measures the angle between objects in a euclidean space.

$$cos(X, Y) = \frac{X \cdot Y}{\|X\|\|Y\|}$$

Clustering - k-means algorithm

- A basic algorithm that is based on distance using Euclidean distance.
- *k* represents the number of clusters that will be generated.
- The algorithm is iterative and starts with random or pseudo random points.
- Each iteration consists of several defined steps.
- First, all instances are assigned to the closest centroids.
- Then the centroids are moved to the current means.

Clustering - k-means algorithm



Clustering - k-means algorithm

- k-means algorithm is very fast in computation of each iterations.
- k-means requires only few iterations.
- In general, k-means is resistance to the initialization.
- Sometimes, in a really rare cases, a wrong initialization leads to non good clustering.
- The clustering may be repeated several times, but usually it leads to the same result.

Clustering - k-means algorithm

- The function that is optimized is defined as follows:

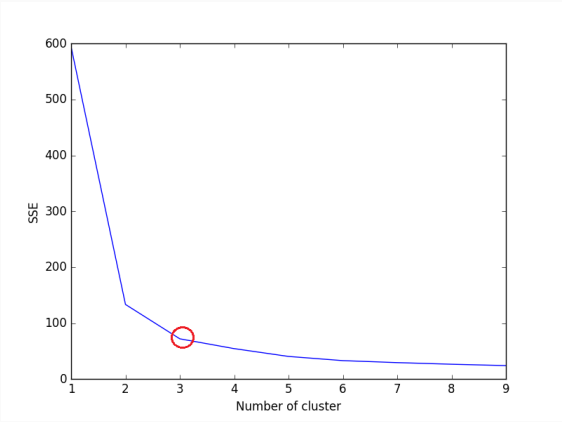
$$\sum_{i=0}^{n} (X_i - C)^2$$

- where X_i is a data point/object/instance and C is the centroid closest to the point x_i.
- It is called a sum of squared distances or sum of squared errors (SSE).
- Due to SSE the algorithm searches for spherical clusters.

Clustering - k-means algorithm

- The number of clusters is selected manually.
- The optimal number of clusters may be evaluated using the SSE function for different number of clusters.
- Consider the increasing number of clusters on a dataset and how the SSE will develop.
- Based on the Elbow method, the optimal k may be selected.

Clustering - k-means algorithm



Clustering - k-means algorithm

- k-means is the basic algorithm or the Representation based clustering.
- Clustering that uses Manhattan distance is called k-median.
- A variant that does not uses distance but similarity is called k-medoid.
- Distance computation may be accelerated using space-indexing data structures (kD-Tree, ball-tree, etc.).
- All the points are assigned to the clusters.

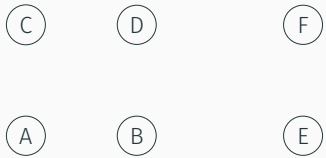
Clustering - Hierarchical Clustering

- Creates a hierarchical structure above the objects from the dataset.
- The different levels of clustering granularity provide different application-specific insights to the data.
- Hierarchical organization of the data allows even better flat cluster

Clustering - Hierarchical Clustering - Algorithm types

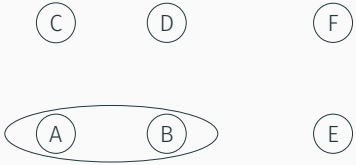
- Bottom-up (agglomerative) methods
  - Individual data objects are agglomerated into higher level clusters.
  - Objective function is used for computing similarity.
- Top-down (divisive) methods
  - Partitioning of the data objects into tree-like structure.
  - A flat clustering algorithm may be used for the partitioning in a given step.
  - A trade-off in balance of the tree between number of clusters and the number of objects in each cluster/leaf.

Clustering - Hierarchical Clustering - Agglomerative Methods

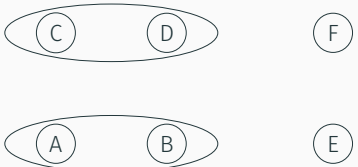


A B C D E F

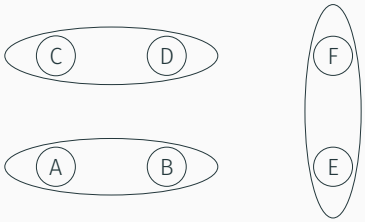
Clustering - Hierarchical Clustering - Agglomerative Methods



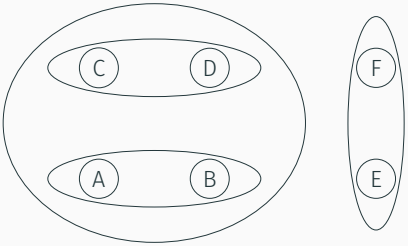
Clustering - Hierarchical Clustering - Agglomerative Methods

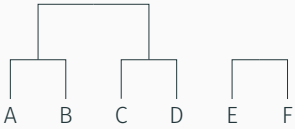


Clustering - Hierarchical Clustering - Agglomerative Methods

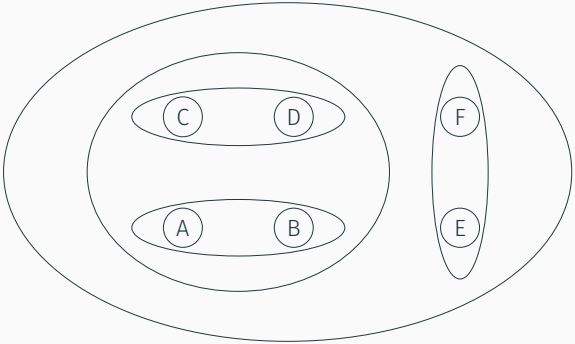


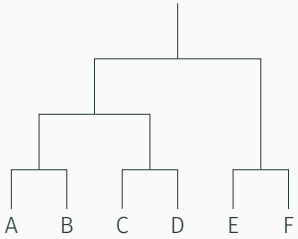
Clustering - Hierarchical Clustering - Agglomerative Methods





Clustering - Hierarchical Clustering - Agglomerative Methods





Clustering - Hierarchical Clustering - Agglomerative Methods

- Iterative approach starting with individual data object.
- Two clusters are merged in each iteration.
- Each merging step reduces the number of clusters by one.
- A carefully selected measure for computation of the distance between individual objects need to be defined.
- A proper strategy for measuring the distance between clusters need to be defined also.
- A distance matrix should be stored in a memory, the computational complexity increases when not.

Clustering - Hierarchical Clustering - Agglomerative Methods

Algorithm 1: AgglomerativeMerge(Dataset: D)

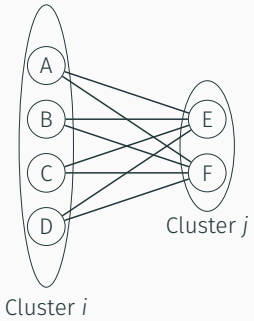
```
1  begin
2      Initialize n × n distance matrix M using D;
3      repeat
4          Pick the closest pair of clusters i and j using M;
5          Merge clusters i and j;
6          Delete rows/columns i and j from M and create a new row and
             column for newly merged cluster;
7          Update the entries of the new row and column of M;
8      until termination criterion;
9      return current merged cluster set
10 end
```

Group Similarity Computation

- Distance between two groups of objects need to be computed.
- The distance is a function of the distances between all pairs of objects from different clusters.

$$D(C_i, C_j) = func_{\forall x \in C_i, \forall y \in C_j}\left(d(x, y)\right)$$

- Each criterion has different advantages and disadvantages.



Clustering - Hierarchical Clustering - Agglomerative Methods

- Best (single) linkage.
  - The distance is equal to the minimum distance between all pairs of objects.
  - Its corresponds to the closes pair of objects between the two groups.
  - Very efficient approach in discovering clusters of arbitrary shape.
  - Very sensitive to noise that connects different clusters.

$$D(C_i, C_j) = \min_{\forall x \in C_i, \forall y \in C_j} \{d(x, y)\}$$

- Worst (complete) linkage:
  - The distance is equal to the maximum distance between all pairs of objects.
  - Its corresponds to the farthest pair of objects between the two groups.
  - This criterion attempts to minimize the maximum diameter of a cluster.

$$D(C_i, C_j) = \max_{\forall x \in C_i, \forall y \in C_j} \{d(x, y)\}$$

Clustering - Clustering Evaluation

- The clustering groups similar objects into groups.
- The clusters created by each algorithm may differ.
- Most important part is the comparison the properties of the objects between groups.
- The ideal way is to evaluate the aggregated features of the groups similarly to the exploration analysis.
- The groups should differ in at least single feature.

Clustering - Clustering Evaluation

- Internal Validation Criteria
  - Sum of Square Distances to Centroids
  - Intra-cluster to Inter-cluster distance ratio.
  - Silhouette coefficient
  - Probabilistic measure
- External Validation Criteria
  - Purity
  - Gini index
  - Entropy

Clustering - Clustering Evaluation

Internal Validation Criteria

- Useful when no external criteria is available.
- The major problem if internal criteria is that they are biased toward one or another algorithms.
- The criteria is usually borrowed from the objective function used by certain algorithms.
- The main usage of these criteria is for comparison of the algorithm from the same class or different run of the same algorithm.

Clustering - Clustering Evaluation

Sum of Square Distances to Centroids

- Useful when centroids are determined – mainly distance-based algorithms.
- The sum of squared distances of each point to corresponding centroid is used as a quality measure.
- The smaller value indicate better clustering quality.

$$SSQ = \sum_{X \in D} dist(X, C)^2$$

- Where C is the closest centroid to X.

Clustering - Clustering Evaluation

Intra-cluster to Inter-cluster distance ratio

- Based on sets of random pairs of objects.
- The P is a set of pairs that belong to the same cluster.
- The Q is a set of pairs that does not belong to the same cluster.
- The average distances are defined as follows:

$$Intra = \frac{1}{|P|} \sum_{(X_i, X_j) \in P} dist\,(X_i, X_j) \qquad Inter = \frac{1}{|Q|} \sum_{(X_i, X_j) \in Q} dist\,(X_i, X_j)$$

- The ratio $Intra/Inter$ is a quality measure. Smaller values means higher quality.

Clustering - Clustering Evaluation

Silhouette Coefficient

- Compares similar distances as the previous one.
- $Davg_i^{in}$ is the average distance of X_i to data points within the cluster.
- $Dmin_i^{out}$ is the minimum of the average distances to all other clusters.

$$S_i = \frac{Dmin_i^{out} - Davg_i^{in}}{\max\left\{Dmin_i^{out}, Davg_i^{in}\right\}}$$

- The overall silhouette coefficient is the average of the data point-specific coefficients.
- The value is in the range $\{-1, 1\}$. Large positive values indicate highly separated clustering, large negative value indicate a ”mixing” between clusters.

Clustering - Clustering Evaluation

External Validation Criteria

- These criteria are available when the ground truth is know.
- In the real datasets, the ground truth is usually not know.
- An approximation may be achieved using available class labels.
- These labels should not correspond to the natural clusters.
- Despite these problems, external evaluation criteria are preferable.
- The number of natural clusters may not reflect the number of classes.

Clustering - Clustering Evaluation

- When the number of determined clusters and the number of classes is equal, a confusion matrix is useful.

Cluster Indices	1	2	3	4
1	97	0	2	1
2	5	191	1	3
3	4	3	87	6
4	0	0	5	195

Cluster Indices	1	2	3	4
1	33	30	17	20
2	51	101	24	24
3	24	23	31	22
4	46	40	44	70

Clustering - Clustering Interpretations

- The quality of the clusters is not as important as the meaning.
- The detected clusters represents the group of objects.
- The groups should contain different type of object to be meaningful.
- Evaluation of the meaning may be done using exploration analysis between clusters.

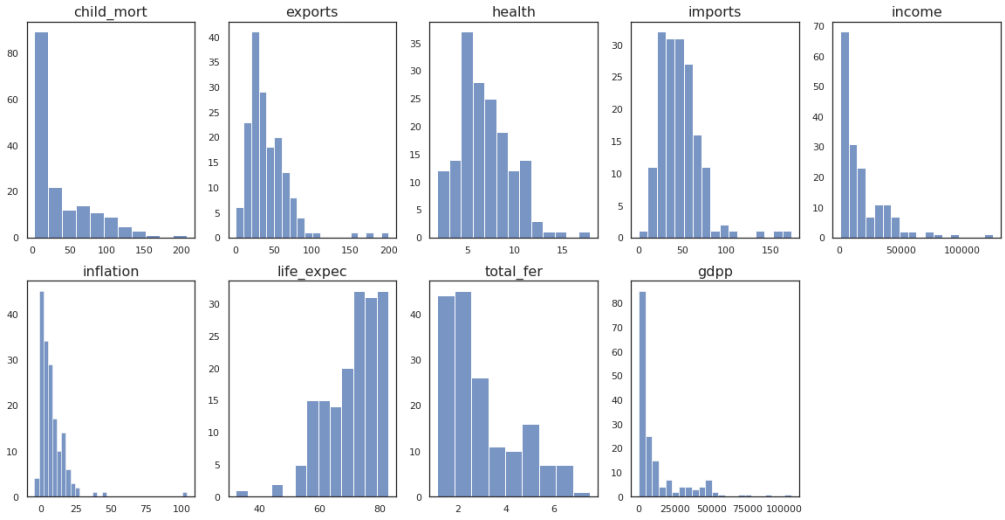
Clustering - Clustering Interpretations - Example

- **Unsupervised Learning on Country Data** (Link)
- **Objective:** To categorize the countries using socio-economic and health factors that determine the overall development of the country.
- **Shape:** 167 rows, 10 columns.
- **Description:** Float numbers with different distribution.

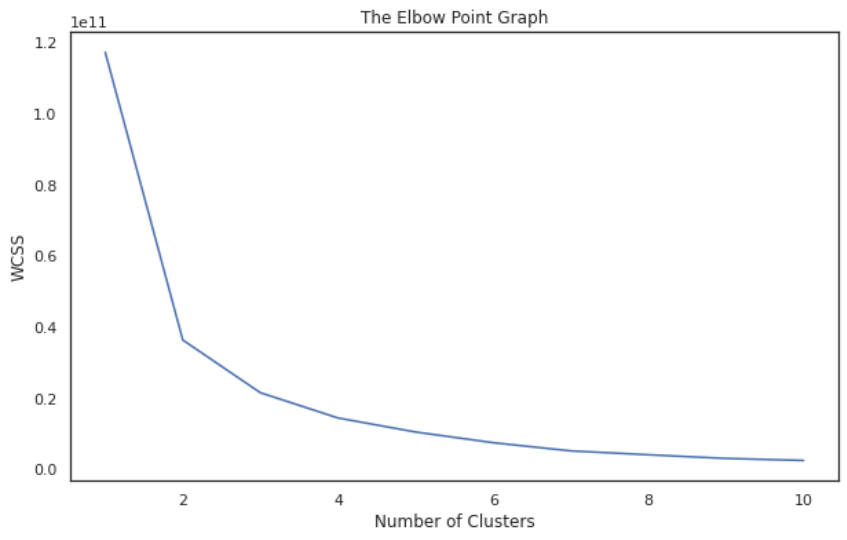
Clustering - Clustering Interpretations - Example - Columns

country	Name of the country
child_mort	Death of children under 5 years of age per 1000 live births
exports	Exports of goods and services per capita. Given as percentage of the GDP per capita
health	Total health spending per capita. Given as percentage of GDP per capita
imports	Imports of goods and services per capita. Given as percentage of the GDP per capita
Income	Net income per person
Inflation	The measurement of the annual growth rate of the Total GDP
life_expec	The average number of years a new born child would live if the current mortality patterns are to remain the same
total_fer	The number of children that would be born to each woman if the current age-fertility rates remain the same.
gdpp	The GDP per capita. Calculated as the Total GDP divided by the total population.

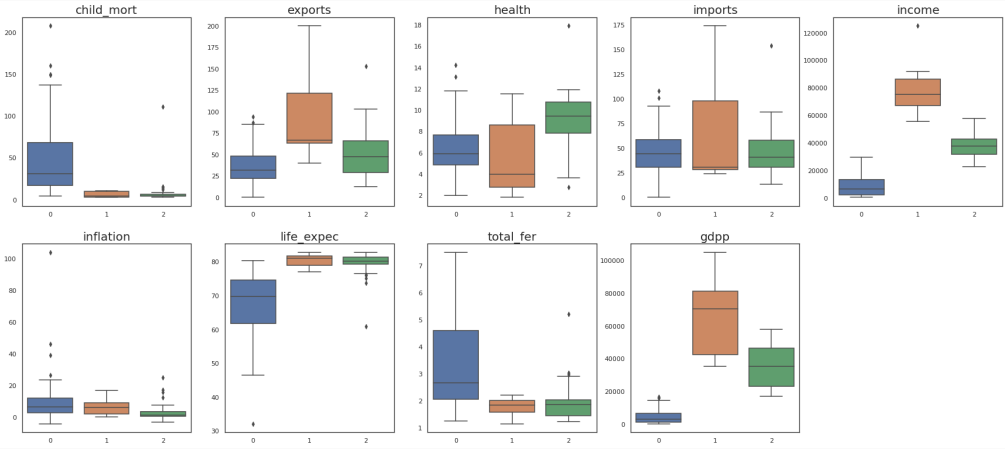
Clustering - Clustering Interpretations - Example



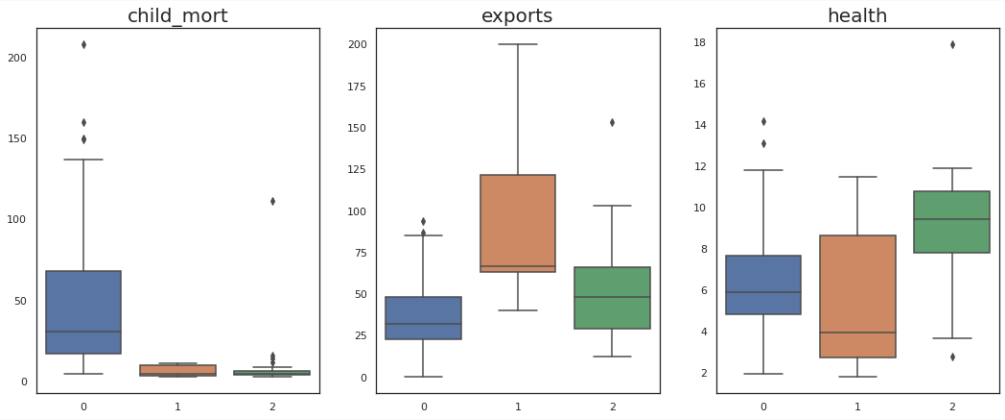
Clustering - Clustering Interpretations - Example



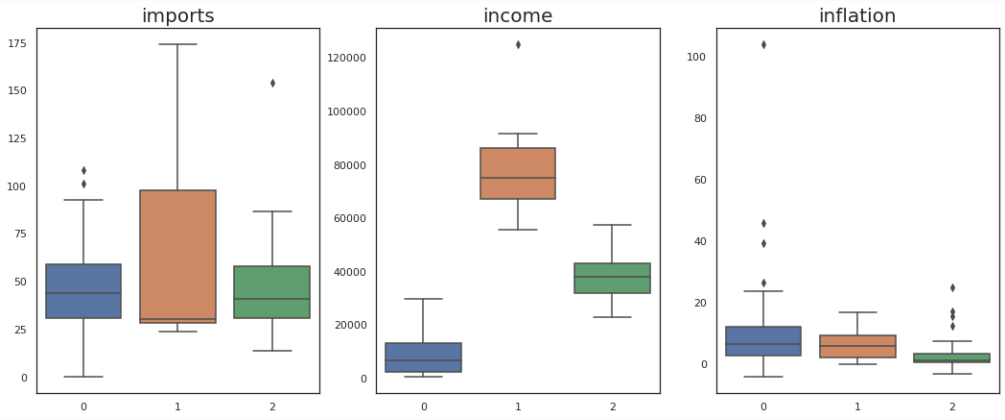
Clustering - Clustering Interpretations - Example



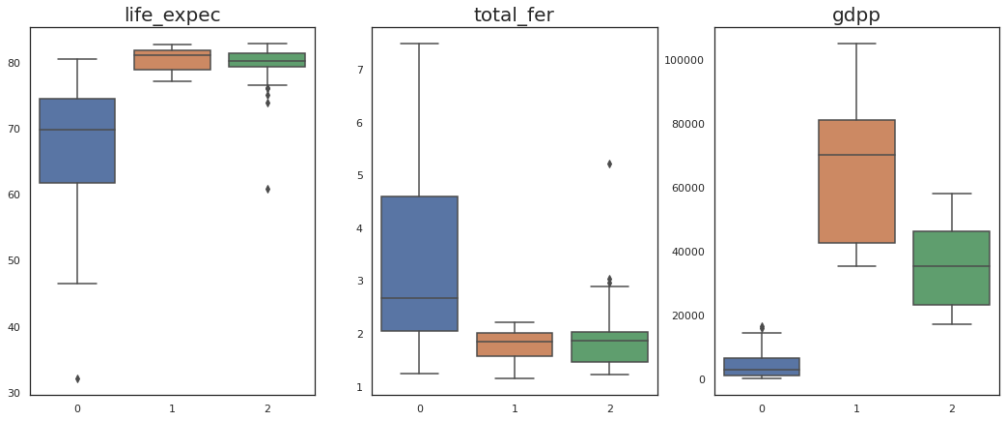
Clustering - Clustering Interpretations - Example



Clustering - Clustering Interpretations - Example



Clustering - Clustering Interpretations - Example



Clustering - Clustering Interpretations - Example

Cluster 0	Afghanistan, Albania, Algeria, Angola, Antigua and Barbuda, Argentina, Armenia, Azerbaijan, Bangladesh, Barbados, Belarus, Belize, Benin, Bhutan, Bolivia, Bosnia and Herzegovina, Botswana, Brazil, Bulgaria, Burkina Faso, Burundi, Cambodia, Cameroon, Cape Verde, Central African Republic, Chad, Chile, China, Colombia, Comoros, Congo, Dem. Rep., Congo, Rep., ...
Cluster 1	Brunei, Kuwait, Luxembourg, Norway, Qatar, Singapore, Switzerland
Cluster 2	Australia, Austria, Bahamas, Bahrain, Belgium, Canada, Cyprus, Czech Republic, Denmark, Equatorial Guinea, Finland, France, Germany, Greece, Iceland, Ireland, Israel, Italy, Japan, Malta, Netherlands, New Zealand, Oman, Portugal, Saudi Arabia, Slovenia, South Korea, Spain, Sweden, United Arab Emirates, United Kingdom, United States

Questions