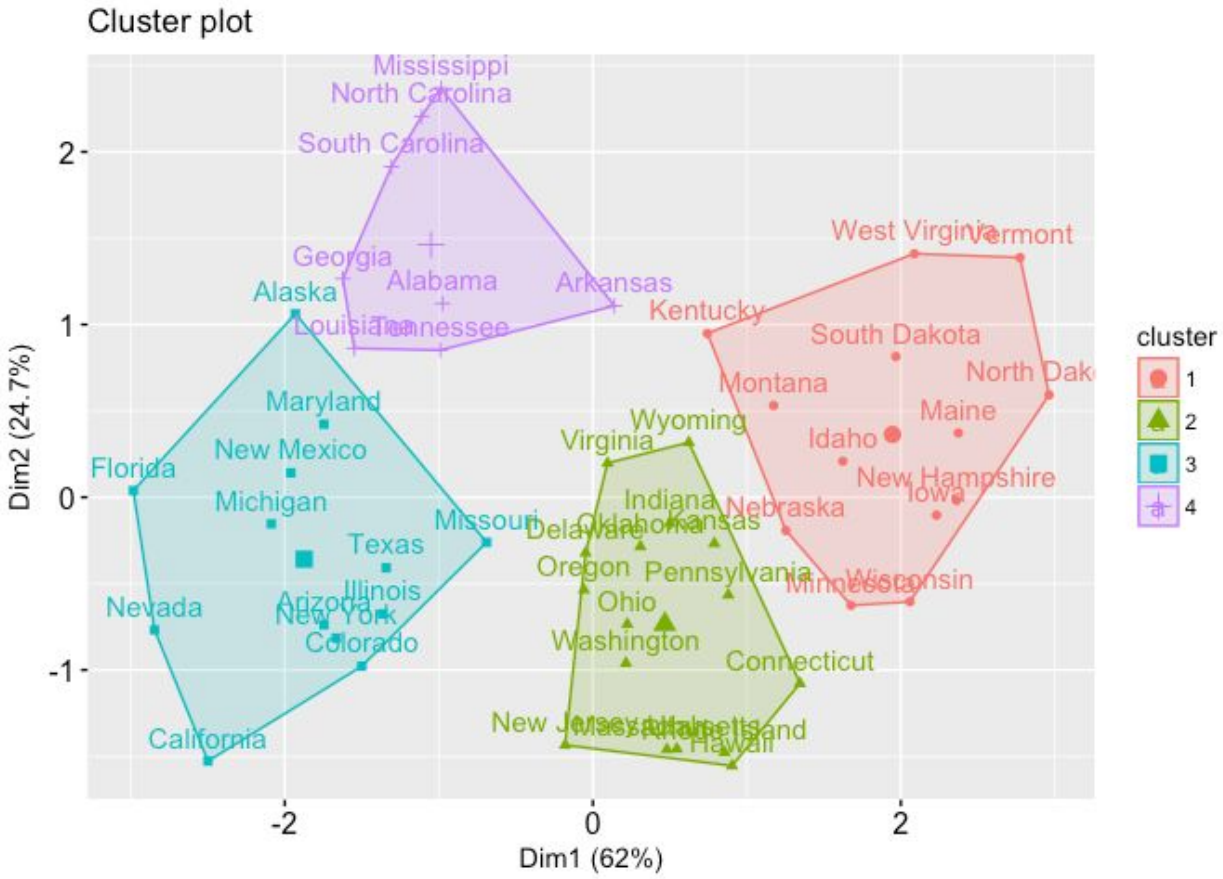
Clustering



What is it?

Grouping particular group of objects based on similar characteristics.When we cluster observations, we want observations in the same group to be similar and observations in different groups to be dissimilar. Because there isn't a response variable, this is an unsupervised method, which implies that it seeks to find relationships between the nn observations without being trained by a response variable.

What does it do?

The classification of observations into groups requires some methods for computing the distance or the (dis)similarity between each pair of observations. The result of this computation is known as a dissimilarity or distance matrix. There are many methods to calculate this distance information; the choice of distance measures is a critical step in clustering. It defines how the similarity of two elements (x, y) is calculated and it will influence the shape of the clusters.

There are several different ways to implement this partitioning, based on distinct models. Distinct algorithms are applied to each model, differentiating its properties and results. These models are distinguished by their organization and type of relationship between them. The most important ones are:

Centralized	each cluster is represented by a single vector mean, and an object value is compared to these mean values
Distributed	the cluster is built using statistical distributions
Connectivity	the connectivity on these models is based on a distance function between elements
Group	algorithms have only group information
Graph	cluster organization and relationship between members is defined by a graph linked structure
Density	members of the cluster are grouped by regions where observations are dense and similar

When is it used?
Examples for using this method can be found here: https://data-flair.training/blogs/clustering-in-r-tutorial/

Personal Experiences with the technique: Trying to find USM institution peers. For example, what institutions of higher education are similar to University of Maryland, Baltimore County?

What kind of data does clustering require:
The clustering algorithm calculated the "mean" of the values, hence continuous numerical data is well suited for this. A lower number of dimensions is preferred. If the data is categorical, then it must be converted to numerical to proceed.
The most common technique used is ***K-means*** clustering (splitting a dataset into a set of k groups.)

R packages and Functions

Tutorials and MOOCs & Reference Materials

- https://uc-r.github.io/kmeans_clustering
  Includes R packages:

```
library(cluster)    # clustering algorithms
library(factoextra)    # clustering algorithms & visualization
library(hclust)
```

- MOOC on Edx - "Cluster Analysis": https://www.edx.org/course/cluster-analysis-utarlingtonx-link-la-cax

- MOOC on Coursera - “Cluster Analysis in Data Mining” https://www.coursera.org/learn/cluster-analysis

References

- “K-Means Cluster Analysis.” *Hierarchical Cluster Analysis · UC Business Analytics R Programming Guide*, uc-r.github.io/kmeans_clustering.
- Image source: https://uc-r.github.io/kmeans_clustering