Methods that are used to discover unknown relationships in data are called *unsupervised methods*.

There are 2 classes of unsupervised methods:

1. *Cluster analysis: It* finds groups in your data with similar characteristics.
2. *Association rule mining: It* finds elements or properties in the data that tend to occur together.
3. The goal of the cluster analysis is to group the observations in your data into *clusters* such that every datum in a cluster is more similar to other datums in the same cluster than it is to datums in other clusters.

There are 2 approaches:

1. Hierarchical clustering- it finds nested groups of clusters
2. K-means- It is a quick and popular way of finding clusters in quantitative data.

Clustering can be done based on similarity and dissimilarity (distance). Few of the distance measures are:

 Euclidean distance

 Hamming distance

 Manhattan (city block) distance

 Cosine similarity

* The Euclidean distance between two vectors x and y is defined as edist(x, y) <- sqrt((x[1]-y[1])^2 + (x[2]-y[2])^2 + ...). But it can only makes sense when all the data is real-valued (quantitative). If the data is categorical (in particular, binary), then other distances can be used.
* HAMMING DISTANCE, It is useful when all the variables are categorical, where it counts the number of mismatches: hdist(x, y) <- sum((x[1] != y[1]) + (x[2] != y[2]) + ...).Distance can be defined as 0 if two points are in the same category, and 1 otherwise.
* MANHATTAN (CITY BLOCK) DISTANCE, It measures distance in the number of horizontal and vertical units

it takes to get from one (real-valued) point to the other (no diagonal moves): mdist(x, y) <- sum(abs(x[1]-y[1]) + abs(x[2]-y[2]) + ...) This is also known as *L1 distance* (and squared Euclidean distance is *L2 distance*).

* COSINE SIMILARITY, It is a common similarity metric in text analysis. It measures the smallest angle between two vectors (the angle theta between two vectors is assumed to be between 0 and 90 degrees).