**Clustering**

**Overview**

* Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters).
* Primarily used for exploratory data analysis and business applications like customer segmentation, product segmentation, market segmentation.

In this class we will explore a cluster technique called k-means & DBSCAN and understand how they works.

* Introduction to k-means() clustering
* Scaling of data before cluster analysis
* Elbow Method and Dendogram to find out optimal number of clusters

Can a clustering algorithm group them together by how nearer they are to each other?

Using K-means clustering technique

* k-mean calculates the distance between the points and the center using euclidean distance and then allocates the points to different clusters.

This dataset contains sample of beer brands and analyzes to understand the type of beer the is manufactured and what kind of market is targets. This gives an insight into different segments of market and presence of different beers in different segments.

**Attribute Description:**

* name - the beer brand
* calories - calories per ounce
* sodium
* alcohol - alcohol percentage present
* cost - in dollars

**Clustering evaluation (Finding optimal number of clusters)**

This explains how to find optimal number of clusters in a given dataset by using various techniques. Different techniques discussed here are

* Dendogram
* Elbow method
* Silhoutte score Analysis We will first load the data into dataframe and scale the features and create clusters. And then various metrics are calculated to validate the number of cluster creations and what will be the optimal number of clusters.

The [Silhouette Coefficient](http://scikit-learn.org/stable/modules/clustering.html#silhouette-coefficient) is a common metric for evaluating clustering "performance" in situations when the "true" cluster assignments are not known.

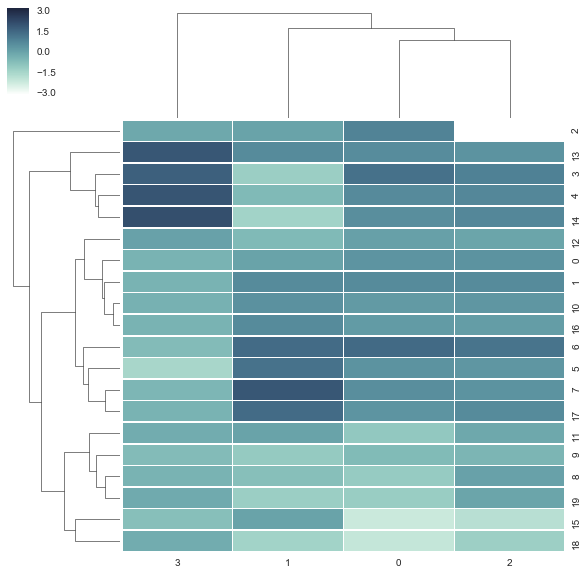
A Silhouette Coefficient is calculated for **each observation**:

*SC*=*b*−*amax*(*a*,*b*)SC=b−amax(a,b)

* a = mean distance to all other points in **its cluster**
* b = mean distance to all other points in **the next nearest cluster**

It ranges from -1 (worst) to 1 (best). A **global score** is calculated by taking the mean score for all observations.

* The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation).
* The silhouette ranges from -1 to 1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.
* If most objects have a high value, then the clustering configuration is appropriate. If many points have a low or negative value, then the clustering configuration may have too many or too few clusters.
* The silhouette can be calculated with any distance metric, such as the Euclidean distance or the Manhattan distance.

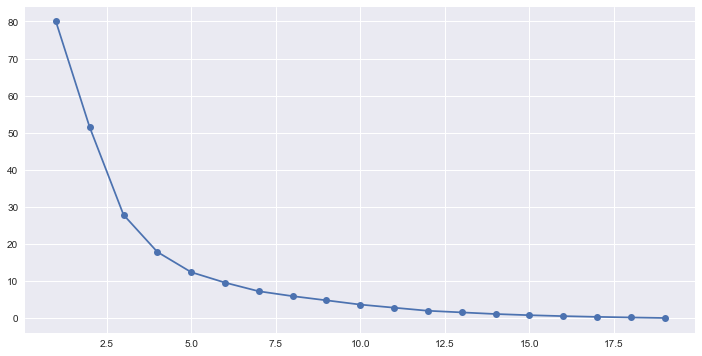
**Using Dendogram**

* Dendogram shows the distance between any two observations in a dataset. The vertical axis determines the distance. The longer the axis, the larger the distance.
* The clustermap feature in seaborn provides the dendogram. It also displays the distance between observations as well as the features. But we are mostly interested in observations.

**Note:**

* From the diagram it can observed that the number of clusters or groups of observations present are between 3 to 6. Based on business context the number of clusters can be created.
* The dendogram shows there are 5 disctinct clusters. Or we can consolidate them into 4 clusters. Too many segments may be harder to interpret, so we will stick to 4 segments here.

**Elbow Analysis**

* The Elbow method is a method of interpretation and validation of consistency within cluster analysis designed to help finding the appropriate number of clusters in a dataset.
* Explained Variance: This method looks at the percentage of variance explained as a function of the number of clusters: One should choose a number of clusters so that adding another cluster doesn't give much better modeling of the data.
* if one plots the percentage of variance explained by the clusters against the number of clusters the first clusters will add much information (explain a lot of variance), but at some point the marginal gain in explained variance will drop, giving an angle in the graph. The number of clusters is chosen at this point, hence the "elbow criterion"
* Given a set of observations (x1, x2, …, xn), where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k (≤ n) sets S = {S1, S2, …, Sk} so as to minimize the within-cluster sum of squares (WCSS) (sum of distance functions of each point in the cluster to the K center). -In other words, its objective is to find: argminS∑i=1k∑x∈Si∥x−μi∥2

# Note:

* The elbow diagram shows that the gain in explained variance reduces significantly from 3 to 4 to 5. So, optimal number of clusters could either 4 or 5.
* The actual number of clusters chosen can be finally based on business context and convenience of dealing with number of segments or clusters.