**Kmeans Algorithm**

Questions:

* How is similarity defined?

Definition:

* K-means is simple: group similar data points together and discover underlying patterns. To achieve this objective, K-means looks for a fixed number (k) of clusters in a dataset.”
* K-means is an unsupervised learning model, which means that it has no target variable and does not require labeled data.
* Define target number K, which refers to the number of centroids.
* Every data point is allocated to each of the clusters through reducing the in-cluster sum of squares.
* In other words, the K-means algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible.
* The ‘means’ in the K-means refers to averaging of the data; that is, finding the centroid.

Clustering Types:

* Hard clustering: in hard clustering, each data object or point either belongs to a cluster completely or not. For example in the Uber dataset, each location belongs to either one borough or the other.
* Soft clustering: in soft clustering, a data point can belong to more than one cluster with some probability or likelihood value. For example, you could identify some locations as the border points belonging to two or more boroughs.

**Algorithm Types**:

Connectivity-based clustering: the main idea behind this clustering is that data points that are closer in the data space are more related (similar) than to data points farther away. The clusters are formed by connecting data points according to their distance. At different distances, different clusters will form and can be represented using a dendrogram, which gives away why they are also commonly called "hierarchical clustering". These methods do not produce a unique partitioning of the dataset, rather a hierarchy from which the user still needs to choose appropriate clusters by choosing the level where they want to cluster. They are also not very robust towards outliers, which might show up as additional clusters or even cause other clusters to merge.

Centroid-based clustering: in this type of clustering, clusters are represented by a central vector or a centroid. This centroid might not necessarily be a member of the dataset. This is an iterative clustering algorithms in which the notion of similarity is derived by how close a data point is to the centroid of the cluster. k-means is a centroid based clustering, and will you see this topic more in detail later on in the tutorial.

Distribution-based clustering: this clustering is very closely related to statistics: distributional modeling. Clustering is based on the notion of how probable is it for a data point to belong to a certain distribution, such as the Gaussian distribution, for example. Data points in a cluster belong to the same distribution. These models have a strong theoritical foundation, however they often suffer from overfitting. Gaussian mixture models, using the expectation-maximization algorithm is a famous distribution based clustering method.

Density-based methods search the data space for areas of varied density of data points. Clusters are defined as areas of higher density within the data space compared to other regions. Data points in the sparse areas are usually considered to be noise and/or border points. The drawback with these methods is that they expect some kind of density guide or parameters to detect cluster borders. DBSCAN and OPTICS are some prominent density based clustering.

**Algorithm** (steps):

1. Generate k nodes with random locations on the Euclidean space.
2. Assign each point to the node who has the least Euclidean squared distance.
3. Calculate new means for each cluster than then become our new centroids.
4. Assign each point to the new centroids.
5. Repeat process until you reach a certain threshold.

**References**:

Kmodes:

<https://analyticsdefined.com/using-k-modes-clustering-categorical-data/>

Kmeans (python)

<https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1>

Kmeans (R)

<https://www.datacamp.com/community/tutorials/k-means-clustering-r>

Data

<https://www.kaggle.com/fivethirtyeight/uber-pickups-in-new-york-city/downloads/uber-pickups-in-new-york-city.zip/2>

Wiki

<https://en.wikipedia.org/wiki/K-means_clustering>