Clustering Notes

* Centeroid = data point, mid point of multidimensial space, middle mid-point
* Randomly pick points off the graph (doesn’t have to be random
* Group them closer to the chosen point and then do it again and again and again
* There are no labels, there is no preconceived notions going in
* Closest using Euclidean distance
* k-means works in all dimensions (starting at 2)
* recalc each group and finding the point in the middle of the group you made in their first centroid
* “gravity clusters” center of gravity – find the centroid
* every point has to be in a cluster, even if it’s an outlier

how to choose K

* random can create outliers but common
* you can cross validate with random variables on the same K and getting the same clusters means its working well

How do you determine Similarity

* basically an average of all points
* you can make that average no matter how many dimensios because it does it by index
* parametric = figure it out.
* Cohesion = how close they are   
  and separation = how far away are the clusters from each other
  + Much like k-neighbors there is a blend between the two is the ideal
  + Silhouette coefficient (I love this name)
  + Low cohesion = compact, higher = more dispersed
  + High separation = centeroids are far apart
  + Ideal is low cohesion, high separation (clear clusters, make it easy to see)
  + If clusters overlap you get a negative (want the SC closer to 1)
  + But it is unsupervised, so there is no “right” answer unless you decide
    - Needs human interpretation
* Kmeans only works with numbers
  + Euclidean distance only works with numbers, can’t measure the distance of names

Scaled

* Scaling it is telling you how far away it is from the average from its column
* StandardScaler from sklearn – will fit transform all the columns and go forward
  + Name it a different variable so they are separate
  + Fit to the scale and interpret with the original (easier to see as a human interpreter)
  + Better clusters, put the data back into the original data frame so it is easier to see
  + Sklearn = metrics to show SC
* DBscan is non parametric
  + Doesn’t assume that it is a circle, can move it around