Clustering

find group in your data

no target variable to be modeled

The *k*-means Clustering Algorithm

The *k*-means algorithm is perhaps the most commonly used clustering method. Having been studied for several decades, it serves as the foundation for many more sophisticated clustering techniques. The following table list a couple of reasons why *k*-means is still widely used.

Strengths

1
- Use simple principles that can be used in non-statistical terms

2
- Highly flexible, it can be adapted with simple adjustments

3
- Perform well enough under many real-world use cases

Weaknesses

1
- Not as sophisticated as more modern clustering algorithms

2
- Because it uses an element of random chance, it is not guaranteed to find the optimal set of clusters

3
- Requires a reasonable guess as to how many clusters naturally exist in the data

The *k*-means algorithm assigns each of the n data points to one of the k clusters, where k is a predetermined number. The goal is to minimize the distances within each cluster and maximize the difference between the clusters.

k-means Steps

1. k cluster "*center*" points are created at random.
2. For each observation:
   (a) The distance between each observation and the k center points is calculated.
   (b) The observation is assigned to the cluster of the nearest center point.
3. The center points are moved to the means (i.e., centers) of their respective clusters.
4. Steps 2 and 3 are repeated until no observation changes in cluster membership.