Rich Gruss

Practical Data Science With R, Chapter 8

1. You want to verify that the clusters you've found are intrinsic to the data and not just an artifact of the clustering algorithm. One way to do this is with bootstrap resampling, a process that assesses the stability of a cluster by re-clustering on a new data set formed by sampling the old dataset with repetition. Another way to evaluate the clustering is to calculate the total within sum of squares (WSS), which measures the total distance of the observations from their centroids. WSS decreases as k grows, but the rate of decrease will flatten after the optimal k. A third method is the Callinki-Harabaz index, which is the ratio of between-cluster distance to within-cluster distance.
2. The kmeans algorithm isn't guaranteed to have a unique stopping point. In fact, it can go on forever, so in some cases you need to specify a maximum number of iterations.
3. Association rules describe the probability that two products are purchased together. The "confidence" that a=>b is proportion of the percentage of transactions that include both a and b to the percentage of transactions that that include a. So if a if purchased in 20% of the transaction, and a and b are purchased together 10% of the time, the confidence that a=>b is .5. R provides the function apriori() to derive association rules from a transactions object. "Lift" is a measure of how real a pattern is. The larger it is, the less likely a pattern occurred by chance.