Assignment 4 – Alan Kessler, PREDICT 410 Section 57

The data for this assignment is a series of self-rated personality attributes from 240 individuals. An important step in clustering is to standardize the data so that each variable is given equal weight. The tutorial first uses k-means clustering to group variables. This could be used to reduce multicollinearity or describe personality as a series of factors. For example, for five clusters, “relaxed”, “laidback”, and “easygoing” are all part of the same cluster so one of them could be used as a representative variable in a model using this data. The groups of variables are fairly intuitive and correspond to the big five factors of personality well. The tutorial then discusses the silhouette score, a measure of how well the distance between clusters is maximized and the distance between points within a cluster is minimized. This score is applied to clustering the individual subjects which could be used in targeted marketing where different personalities respond to marketing in unique ways. The largest silhouette score indicative of the strongest clustering specification corresponds to two clusters. The two clusters can be interpreted multiple ways. First, the variable means can be compared across clusters. For example, the first cluster has a much lower rating for “outgoing” than the second cluster. The biggest differences are in “outgoing”, “shy”, and “withdrawn” which supports the extraversion/introversion model of personality. Second, the clusters can be interpreted visually. A scatter plot can show the cluster assignments as different colors across two of the variables used. Additionally, principal components could be used to reduce the dimensionality when visualizing the clusters. The plot of cluster centers indicates how different the clusters are from each other. The results show a clear separation in this case. The final example in the tutorial shows a two-cluster solution for a dataset with four clearly defined clusters. A visual inspection of this data indicates that two-clusters is insufficient.