Intro to Machine Learning
Case Study: Clustering and Unsupervised Learning

by James D. Wilson (University of San Francisco)

1. Using R or Python, create a pipeline for clustering algorithms that we've discussed in class. You do *not* need to hard code each clustering algorithm by hand. You can work in groups of up to 4 on this case study and only one assignment needs to be turned in per group. This task is flexible, but make sure that the pipeline includes the following components:

   (a) **Input**:
      i. an $n \times p$ data matrix or table X whose rows are observations and columns are variables/measurements.
      ii. the number of clusters k
      iii. choice of 1 of at least 3 distance metrics *dist* (e.g., "Euclidean")
      iv. choice of 1 of at least 3 linkage metrics *linkage* (e.g., "single")

   (b) **Output**: the cluster assignments for each observation based on the application of (at least) the following algorithms
      i. Spectral clustering
      ii. Hierarchical clustering (agglomerative and/or divisive)
      iii. k-means
      iv. k-mediods

   (c) **Visualizations**:
      i. a dendrogram for the rows of X
      ii. a heatmap of the confusion matrix whose (i, j)th element is the proportion of times observation i and observation j are clustered together based on the above 5 methods
      iii. a 4×4 grid of biplots for the first 4 pcs of the columns of X with points colored based on the cluster assignment for each point *for each of the 5 clustering methods* applied

2. Apply your algorithm from above to demonstrate its utility on any data set of your choosing. Briefly describe the data set and why clustering is an important task to consider for this data. Discuss the results of your clustering analysis on this data set. Do the clusters provide meaningful insights to your data? How can you tell? Remember that you can always use kaggle.com or resources like the UCI repository for Machine Learning for examples of data.