Due on April 15, 2020

1. File `universities.csv` has variables useful to compare or rank major universities. For the schools in that file these variables include

   X_1 = average SAT score of new freshmen,

   X_2 = percentage of new freshmen in top 10% of high school class,

   X_3 = percentage of applicants accepted,

   X_4 = student-faculty ratio

   X_5 = estimated annual expenses and X6 = graduation rate (%)

   a) (10 pts.) Find Principal Components. Contruct a biplot of the universities in PC axes. From this biplot, can you identify (by inspection) clusters of universities?

   b) (20 pts.) Calculate Euclidian distances between pairs of universities. Use them to cluster the universities using complete and single linkage. For each type of linkage construct the dendrogram and cut it for $K = 4$ clusters.

   c) (20 pts.) Use K-means clustering to cluster the universities into $K = 4$ clusters. Use `set.seed(2)`.

2. File `brands.csv` contains data on breakfast cereals produced by three different American manufacturers: General Mills (G), Kellogg (K), and Quaker (Q).

   a) (10 pts.) Find Principal Components and plot the cereals in PC axes, use different color for different manufacturers. Does it appear as if some manufacturers are associated with more nutritional elements (high protein, low fat, high fiber, low sugar, etc.)?

   b) (20 pts.) Calculate Euclidian distances between pairs of cereal brands. Use them to cluster the cereals using complete and single linkage. For each type of linkage construct the dendrogram and cut it for four clusters. Construct a dataframe with the names of the breakfast cereals, and a column showing the assigned cluster to each row.

   c) (20 pts.) Use K-means clustering to cluster the cereals into $K = 4$ clusters. Add a column showing the assigned cluster to the dataframe created in part (b). How different are the clusters?

Report must show the student's last name, first name, and USC ID.