(0) At http://www.stat.cmu.edu/~larry/all-of-statistics/=data/faithful.dat find some two dimensional data. Implement a k-means algorithm for two clusters using just the first (!) column only. Evaluate your algorithm visually.

(1) Apply your method only to the second column. Is it better or worse?

(2) Generalize your algorithm to vector valued data and an arbitrary number of clusters. Apply it to the full data set with both columns

(3) Suppose for the i-th data point the first column has value c1(i) and the second c2(i) . Is there a new c(i) = a * c1(i) + b * c2(i) with suitable well picked a and b such that clustering based on c(i) is better than the one done on c1(i) or c2(i) alone.