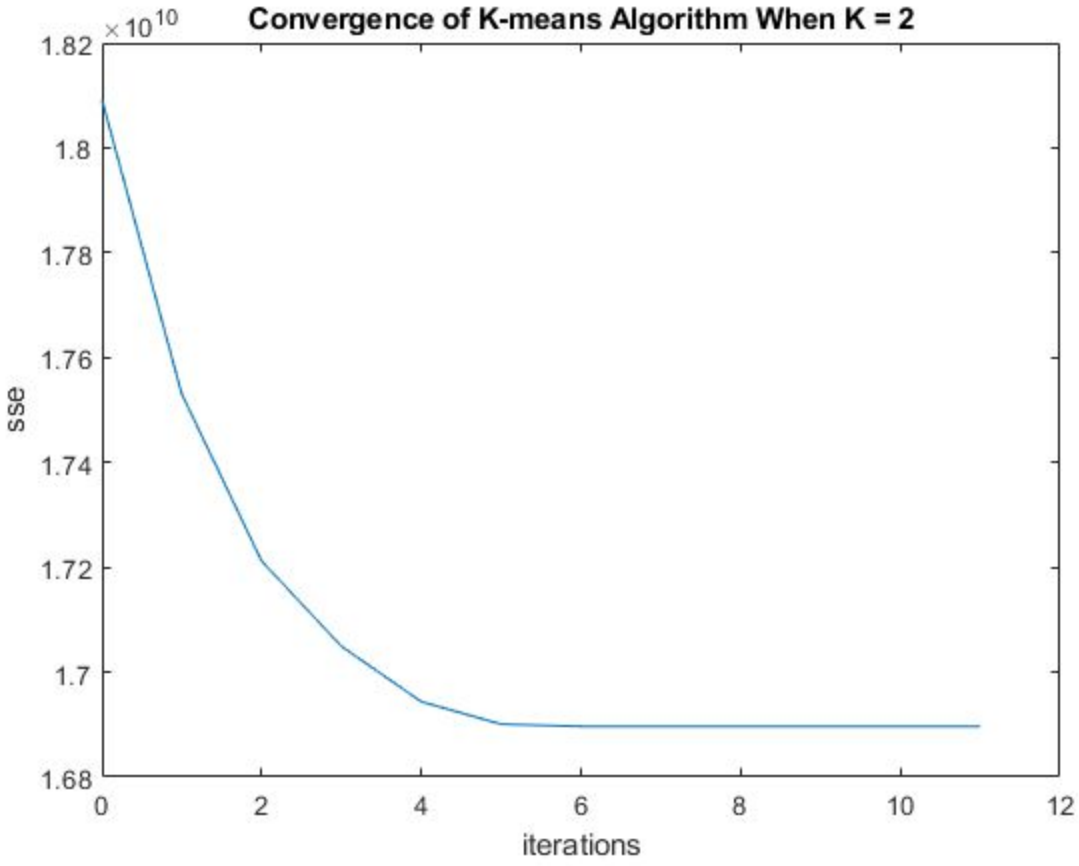
Parker Bruni and Brian Blythe
CS434
5/27/2018

Implementation Assignment 4

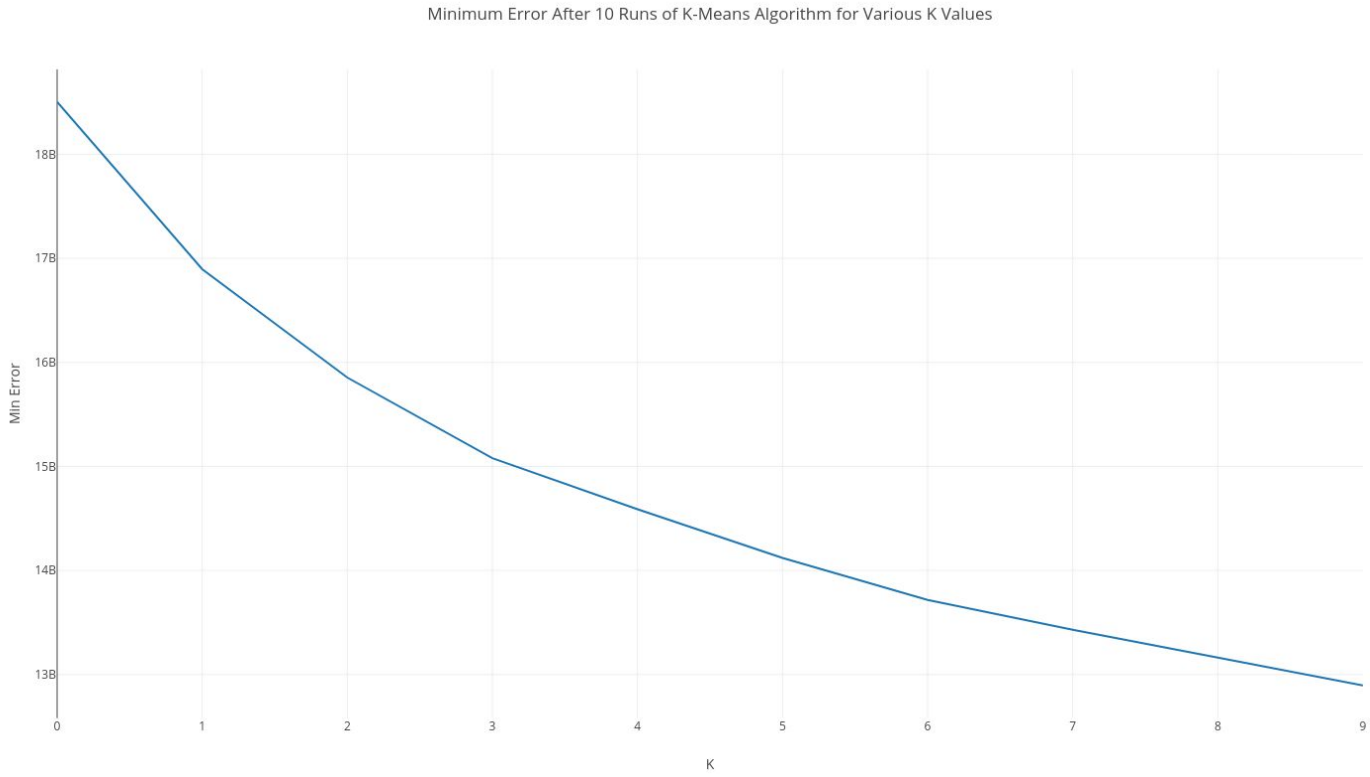
1. Non-hierarchical clustering K-Means algorithm

1.



This is a typical run of convergence for when K = 2. It begins to converge after about 5 iterations, at which point the changes to the sse slow down substantially.

2.



(In case graph text is too small)
Title: Minimum Error After 10 Runs of K-Means Algorithm for Various K Values
Y-axis: Min Error
X-axis: K

We think a proper K value for this dataset is K = 3. We pick this value for K because it appears to be the location where the error for the K-means algorithm as a function of K appears to have the largest second derivative. This means that for K values less than and equal to 3, K-means likely groups data points into three somewhat discrete clusters, but for K values greater than 3, K-means starts to subdivide clusters. This subdivision of clusters reduces the overall sum of squared error but does so at a slower rate for increased K values since the dominant clusters have already been achieved by K = 3.

2. Principal Component Analysis

1.
Top 10 eigen values in decreasing order:

352868.69
267895.87
227632.70
174703.49
130486.76
115542.50
99726.44
90576.06
85326.54
71547.97

For part 3.2 and 3.3
Due to technical difficulties, we were unable to get matlab to install properly on either the osu server or to our local machine. Because of this, we were unable to generate the graphs to view the data visually. However, the mean image and eigenvectors are calculated within the program when executed.