David Smith
STAT 4630
Homework 3

1.

See R file.

2.

See R file.

3.

See R file.

4.

a) The proof is as follows:

$$\frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2 = \frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}x_{ij}^2 - \frac{2}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}x_{ij}x_{i'j} + \frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}x_{i'j} =$$

$$\sum_{i\in C_k}\sum_{j=1}^{p}x_{ij}^2 - 2\sum_{i\in C_k}\sum_{j=1}^{p}x_{ij}\bar{x}_{kj} + \sum_{i\in C_k}\sum_{j=1}^{p}x_{ij}^2 = 2\sum_{i\in C_k}\sum_{j=1}^{p}x_{ij}^2 - 4\sum_{i\in C_k}\sum_{j=1}^{p}x_{ij}\bar{x}_{kj} +$$

$$2\sum_{i\in C_k}\sum_{j=1}^{p}x_{ij}\bar{x}_{kj} = 2\sum_{i\in C_k}\sum_{j=1}^{p}x_{ij}^2 - 4\sum_{i\in C_k}\sum_{j=1}^{p}x_{ij}\bar{x}_{kj} + 2|C_k|\sum_{j=1}^{p}\bar{x}_{kj}^2 = 2\sum_{i\in C_k}\sum_{j=1}^{p}x_{ij}^2 -$$

$$4\sum_{i\in C_k}\sum_{j=1}^{p}x_{ij}\bar{x}_{kj} + 2\sum_{i\in C_k}\sum_{j=1}^{p}\bar{x}_{kj}^2 = 2\sum_{i\in C_k}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{kj})^2$$

b) In steps 1 and 2a, nothing changes. In step 2b, since each observation is reassigned to the cluster whose centroid is closest, each x_{ij} is now closer to $\bar{x}_{kj}$ when we sum over all j features, and so the right-hand side of (10.12) decreases. Thus, the left-hand side also decreases, and so (10.11) decreases at each step when we sum over all K clusters.

5.

See R file.