Unsupervised Learning Lab Sheet

1.

(a) The dataset in the table below contains 10 examples, each described by 4 numeric features. These 10 examples have been randomly assigned to two clusters, *C1* and *C2*, in order to initialise the *k*-Means algorithm. The assignments are as follows:

Item	f1	f2	f3	f4
x1	5.1	3.8	1.6	0.2
x2	4.6	3.2	1.4	0.2
x3	5.3	3.7	1.5	0.2
x4	5	3.3	1.4	0.2
x5	7	3.2	4.7	1.4
x6	6.4	3.2	4.5	1.5
x7	6.9	3.1	4.9	1.5
x8	5.5	2.3	4	1.3
x9	6.5	2.8	4.6	1.5
x10	5.7	2.8	4.5	1.3

C1 = { x1, x3, x7, x8 } C2 = { x2, x4, x5, x6, x9, x10

Based on the table above and the cluster assignments, calculate the centroid vector for each cluster.

(b) Based on the centroids calculated in part (a), which clusters will the examples *x1* and *x10* next be assigned to? Calculate distances using the Euclidean distance measure.

2. In the notebook `11 Clustering` *k*-Means clustering is applied with Euclidean distance to the *Penguins unlabelled* dataset with the `n_init` parameter set to 1.
   Report the *intra-cluster distance* for clusterings with different numbers of clusters: *k=2*, *k=3* and *k=4*.

   Repeat the above process again, but change the random seed parameter for k-Means. Are the intra-cluster scores identical?

   Repeat again with n_init set to 50. Does this make a difference?