1. You wish to cluster a dataset consisting of 1,000,000 points on a computer with 16GB of memory. Would you use k-means clustering or hierarchical agglomerative clustering? Explain your answer.
   For full marks, either of the following is acceptable:

   You would use k-means, because it has space complexity $O(N)$ whereas HAC has space complexity $O(N^2)$. (Using Θ rather than O is acceptable.)

   You would use k-means because the HAC dissimilarity matrix would not fit into memory. For 1,000,000 elements, it would have $(10^6)^2 = 10^12$ entries, which would be far too large for 16GB of memory.

2. Consider a dataset X consisting of the following points: 1, 2, 3.5, 8, 10 and 14
   Calculate by hand two iterations of the k-means clustering algorithm for k=2, on the above dataset X, with the two centroids initialised to 1 and 2. What are the clusters and centroids after the first and second iterations?

   Centroids for step 1 are:

   - Centroid 1: 1
   - Centroid 2: 7.5

   Centroids for step 2 are:

   - Centroid 1: $2\frac{1}{6}$
   - Centroid 2: $10\frac{2}{3}$

   FYI, here are the workings:

   Iteration 1

   Reassignment:

   - Cluster 1: 1
   - Cluster 2: 2 3.5 8 10 14

   Recentering:

   - Centroid 1: 1
   - Centroid 2: 7.5

   Iteration 2

   Reassignment:

   - Cluster 1: 1 2 3.5
   - Cluster 2: 8 10 14

   Recentering:

   - Centroid 1: $2\frac{1}{6}$
   - Centroid 2: $10\frac{2}{3}$