MIT 880 (k-nn mode seeking clustering) **Assignment:1**

1: Question 1

Consider the data *q_clus.xlsx*. Do a k-nearest neighbour, k-NN, **density** based clustering on the variable X, using mode seeking. Use the following algorithm in answering the question:

- Estimate the density using the k-NN density estimator below.

  $$\hat{f}_{knn}(x_i) = \frac{1}{||x_i - x_k||^2}$$

  Thus the density at point x_i, $\hat{f}_{knn}(x_i)$, is the reciprocal of the squared distance to the $k-th$ nearest neighbour x_k.
- For each of the observations x_i
  - Define a pointer to the observation within the k-nearest neighbours of x_i with the highest k-NN-density.
  - Repeat the process by following pointers from the initial pointer until a pointer that points to itself is found. This will be taken as a local mode of $\hat{f}_{knn}$.
- Assign each point, x_i, that converged to the same mode to the same cluster.

HINT: The attached R code contains two functions, 1) a $k - nn$ density estimation function and 2) a mode seeking function

1. Give the k-NN density estimate for the observed data. Overlay the graph of the density estimate on a histogram of the data.
2. Use the algorithm above to determine the relevant mode(s), comparing different values of k. How many clusters, **c**, does your cluster solution suggest. Motivate your answer.
3. Graphically illustrate the effect of different choices of k on the cluster solution.