Problem 2

> *Prove that given a set of n data points in one dimension (i.e. n real numbers) and a number k, the optimal k-means centers can be computed in time $O(kn^2)$.*
> *Hint: Use dynamic programming. Recall that the goal is to minimize the sum of squared distances of the points to their nearest center. Create a dynamic programming table of size $n \times k$. Store in the $(i, j)^{th}$ entry of the table, the optimal value for the problem on the first i points using j centers. Prove that the $(i, j)^{th}$ entry of the table can be computed in linear time from the lexicographically smaller entries.*

The idea is to look at the general intermediate step in the optimization process. Say we have already classified p points into d clusters optimally where $d = j - 1$. The cost function for this scenario is the minimum cost of the optimal classification of p points and that of attributing the remaining $i - p$ points into the remaining cluster, which is just the mean Euclidean distance from each of the remaining points to the remaining cluster center. We can recursively apply this step to the p points classified into d classes to form a dynamic table. Each recursion is $O(pd)$ deep, which in the case of the $n \times k$ problem is $O(nk)$ and there are n such recursions, so overall time complexity is $O(kn^2)$.