How to select k

One issue with the k-means clustering is how to pick the value of k. Often we are interested in exploring different values, there might be different clustering structures at different scales, for example. Sometimes, however, we want to be able to say that the data shows evidence for some specific number of clusters. Sadly, this rarely works out, or at least, we rarely get a completely satisfactory answer.

We will start out describing how we think it might work. The idea is to define a measure of the quality of the clusters. If

$$\mathcal{C} = \{C_1, C_2, \ldots, C_k\} \tag{1}$$

is a set of clusters with corresponding centroids $\mathbf{c}_i$ then one such quantity is

$$J(\mathcal{C}) = \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in C_i} [d(\mathbf{x}_j, \mathbf{c}_i)]^2 = \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in C_i} |\mathbf{x}_j - \mathbf{c}_i|^2 \tag{2}$$

This is sometimes called the *degree of dissimilarity* since it measures how much the points differ from their centroid. Now it should be clear that this number goes down as k increases, the question is what shape that descending curve has. The claim is that there will be a *dogs' leg*, a sort of corner where $J(\mathcal{C})$ levels off suddenly, this means that adding new centroids isn't as useful anymore and so the dogs' leg corresponds to the best value of k.

In Fig. 1 there are some points with very clear clusters, it is an artificial dataset intended to have this structure. In Fig. 2 the value of $J(\mathcal{C})$ for the clustering discovered through k-means has been plotted. You can see that the fall in the value of $J(\mathcal{C})$ does level out at around five, but, even for this most ideal of datasets, it isn't super obvious" So, in summary, use J to assess the best value of k, but don't be surprised if it isn't super conclusive!

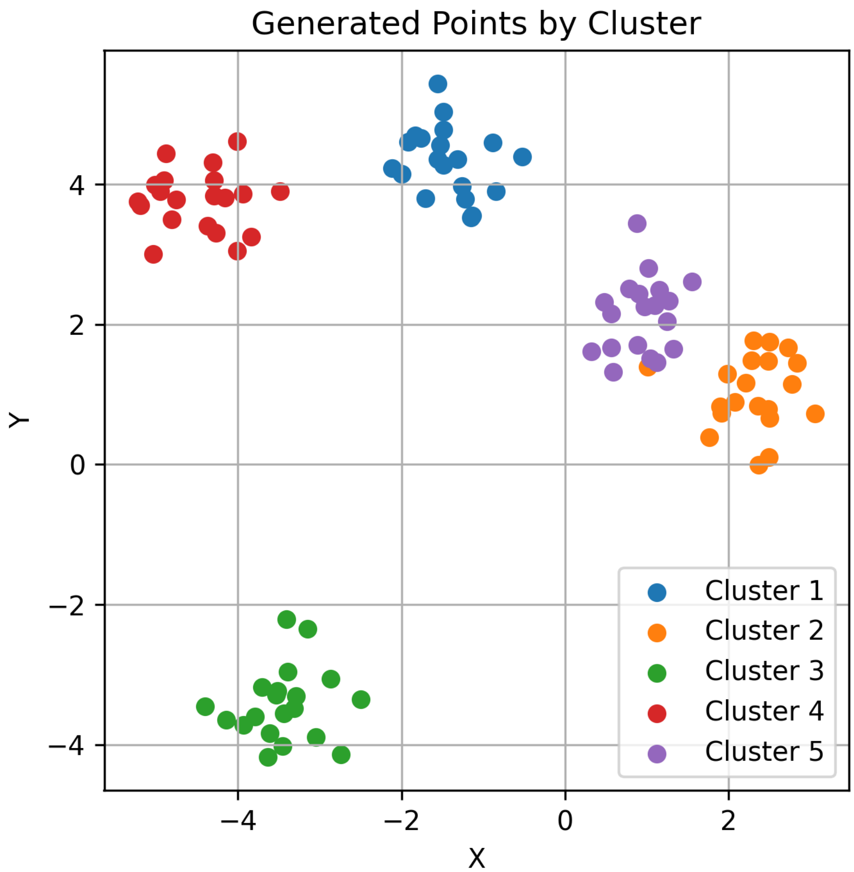

Figure 1: In this simulated data set five centers have been picked at random in a ten by ten box, for each center 20 points have been picked using a normal distribution with standard deviation 0.5.

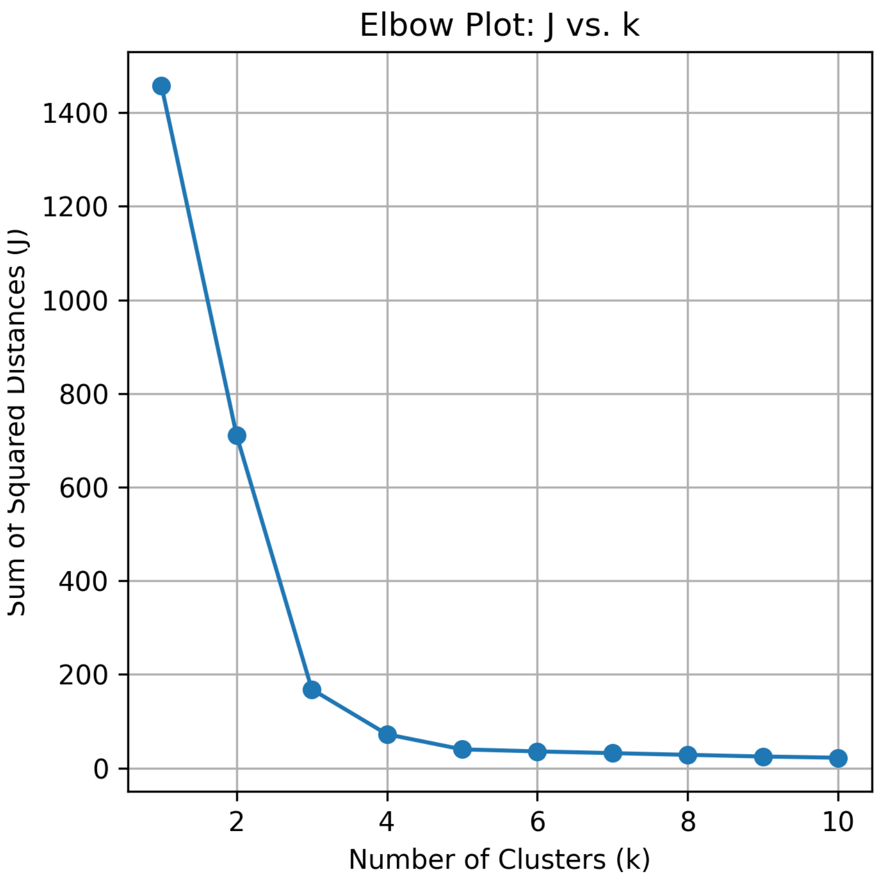

Figure 2: For the simulated data the degree of dissimilarity has been calculated and is plotted against k.