Q) How do we determine a good "k" (for kMeans), the number of clusters?

Ans) We can use "NearestNeighbor" function here. Steps are as follows (Using both kMeans & DBSCAN)

Step-1

Run "NearestNeighbor" fn of Sklearn on the whole dataset and obtain nearest neighbor of each element and distance b/w them

Step-2

Sort the output arrow accordance with the distance and plot it

~~Step-3~~

Step-3

From plot we can choose appropriate value of 'epsilon' for Nearest Neighbor fn & ~~Then run NearestNeighbor fn again with~~

Step - 4

Run DBSCAN with that epsilon value and then we will get cluster numbers from there.

Step-5

Put that number to KMeans

Step-6

Verify the goodness of the model by running Kmeans.inertia A good model should be having low value for Kmeans.inertia

Or else For KMeans

We can use Kmeans.inertia directly for different values of K and then choose appropriate K which will be with low value for inertia