K means++

The parameter r in KMeans++ controls how many times the algorithm randomly initializes centroids, which is important because KMeans++ is sensitive to where the centroids start. If the initial placement is poor, the algorithm can get stuck in the local minima. Running the algorithm multiple times with different initializations increases the chances of finding a better solution. When r is small, the algorithm doesn't have enough chances to explore good initializations, so the results can vary a lot, and the clustering might not be great. Moderate values of r give the algorithm enough opportunities to find better centroids while keeping the computational cost reasonable.If r is too large the improvement in clustering accuracy becomes minimal, but the computational cost increases significantly, making it inefficient. For the iris dataset, using an r value between 5 is a good choice, here we can see the QE does not vary at all after R=2 implying it is easy to find the global minima extremely quickly.

	R	Quantization Error
0	1	78.940841
1	2	78.945066
2	3	78.940841
3	4	78.940841
4	5	78.940841
5	10	78.940841
6	15	78.940841
7	20	78.940841
8	25	78.940841
9	30	78.940841
10	35	78.940841
11	45	78.940841
12	50	78.940841

Fuzzy Kmeans –

The parameters r (number of random iterations) and p (fuzziness parameter) play crucial roles in the fuzzy c-means algorithm. The parameter r determines how many times the algorithm runs with different initializations of cluster centers, helping to avoid local minima and improve clustering accuracy. The influence of these parameters is reflected in the quantization errors. Increasing r generally reduces both the fuzzy quantization error (J) and the hard quantization error by improving the quality of the initial cluster centers. The choice of p affects the balance between hard and soft clustering, with a well-chosen value minimizing both errors. For datasets with well-separated clusters, a lower p is preferable, while overlapping clusters benefit from a higher p. For clustering 50% of the Iris dataset, values with r = 15 and p = 2 provides a good balance between accuracy and computational efficiency.

	r	p	fuzzy_error	hard_error
0	5	1.5	74.436534	79.015015
1	10	1.5	74.436534	79.015021
2	15	1.5	74.436534	79.015020
3	20	1.5	74.436534	79.015011
4	25	1.5	74.436534	79.015010
5	30	1.5	74.436534	79.015010
6	35	1.5	74.436534	79.015014
7	45	1.5	74.436534	79.015009
8	50	1.5	74.436534	79.015010
9	5	2.0	60.546814	79.433514
10	10	2.0	60.546814	79.431905
11	15	2.0	60.546814	79.433314
12	20	2.0	60.546814	79.433324
13	25	2.0	60.546813	79.432637
14	30	2.0	60.546814	79.431786
15	35	2.0	60.546814	79.433225
16	45	2.0	60.546814	79.433309
17	50	2.0	60.546813	79.432780
18	5	2.5	43.564082	80.139361
19	10	2.5	43.564082	80.138803
20	15	2.5	43.564082	80.139368
21	20	2.5	43.564082	80.139192
22	25	2.5	43.564082	80.138556
23	30	2.5	43.564082	80.139365
24	35	2.5	43.564082	80.137855
25	45	2.5	43.564082	80.139351
26	50	2.5	43.564082	80.139359
27	5	3.0	29.081108	80.969584
28	10	3.0	29.081107	80.970270
29	15	3.0	29.081107	80.970449
30	20	3.0	29.081107	80.970947
31	25	3.0	29.081108	80.970104
32	30	3.0	29.081108	80.971365
33	35	3.0	29.081108	80.968817
34	45	3.0	29.081107	80.971307
35	50	3.0	29.081108	80.971527

Spectral Clustering -

The choice of **σ** in spectral clustering significantly impacts the similarity graph and clustering performance. A small **σ** creates a sparse graph where many points are disconnected, failing to capture the global structure of the data, leading to poor clustering. Conversely, a large **σ** results in an overly dense graph where most points are connected, losing the local structure and also degrading performance. An optimal **σ** strikes a balance between local and global structure, typically yielding the best clustering results. Here we suggest a sigma of 2

Sigma	Accuracy
0.5	0.566667
1	0.553333
1.5	0.560000
2	0.646667
2.5	0.646667
5	0.633333
10	0.553333