Assignment 1 - Kmeans Clustering

Liam Duncan - 301476562

March 19, 2024

general info

For silhouette coefficient I followed what is on the slides not what is commonly found online. I used the distance to the centroid of the objects cluster and the centroid of the objects second nearest cluster.

1 Random init

I ran the random algorithm 5 times to get a more accurate representation of the silhouette coefficient. Due to the randomness of the algorthim, results varied drastically. By running it 5 times and taking the averages, the variance was reduced. From this we can see that when k = 3 the best clustering was achieved. This is not surprising when you look at just the locations of the points when they are plotted.

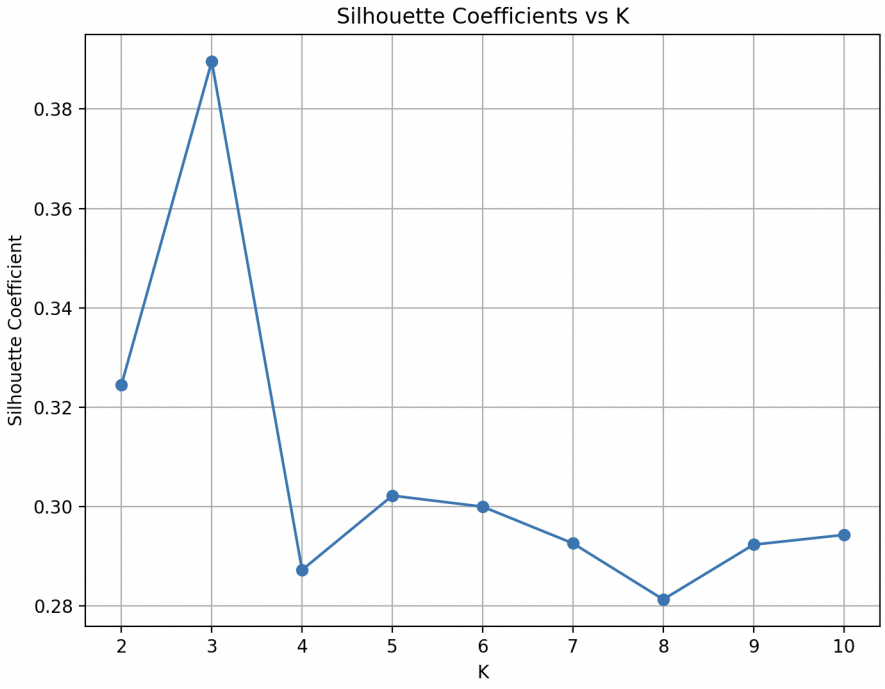

Figure 1: Enter Caption

2 Kmeans++ init

Similar to what I did with the random algorithm, I ran the kmeans++ algorithm 5 times to get a better representation of the silhouette coefficient. From this we can see that in Kmans++, the most accurate was actually when k = 5. This is surprising because based off the look of the data this does not seem to be the likely answer. It looks like it should be 2 or 3. There could be a variety of factors as to why this was the result. There could be an error somewhere in the updating centroids or init centroids that is causing this to be the best choice for K. Not only is the best choice of K surprising, but the best silhouette coefficient of kmeans++ is less than the best of randomness. I think this could be due to the variance of each algorithm. If both these algorithms were run even more times, i.e. 100+ then I think Kmeans++ would eventually on average outperform the random initialization. Overall I think it is tough to directly compare the 2 algorithms when a major part of both algorithms is random. The silhouette coefficient in general is not neccasarily high but when seeing the data plotted its shown that the data is not going to cluster well as there are not very distinct cluster that can be created visually. It could also be due to the fact that we are only having 300 iterations. If we kept it going until there were no more objects changing clusters, then the silhouette coefficient may have increased.

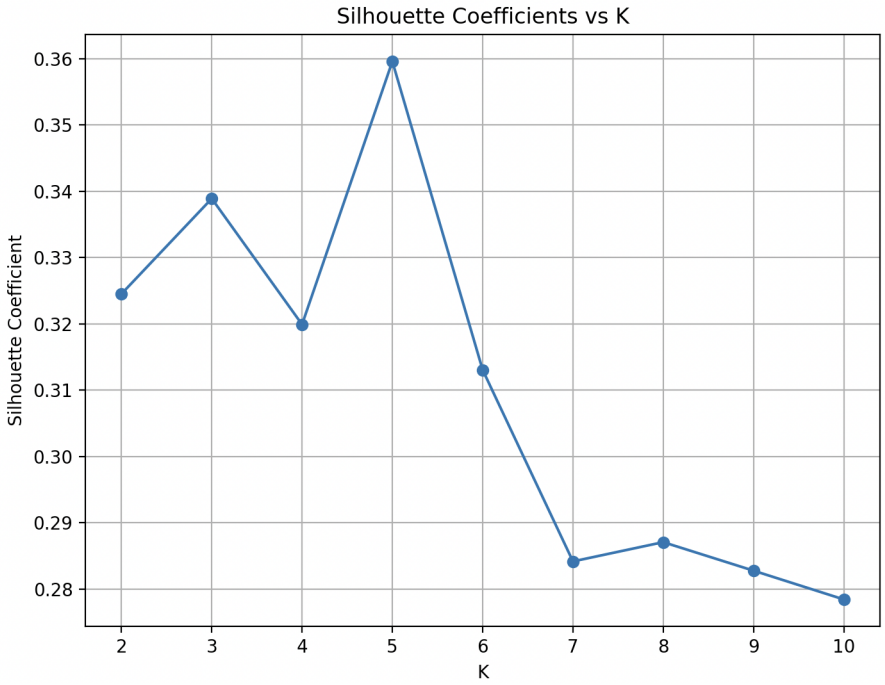

Figure 2: Enter Caption

3 Plots

Based off the plots, it is not surprising that when k = 3 in random, the silhouette was greater than when k = 5 in kmeans++. There is less overlap in the random plot meaning that the clustering is better.

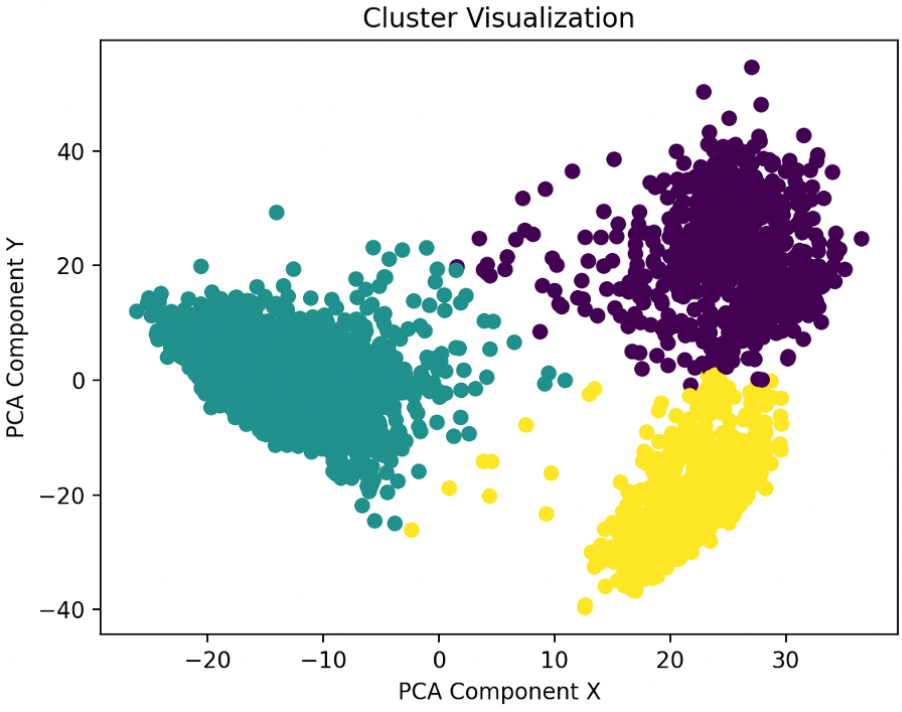

Figure 3: Random where k = 3

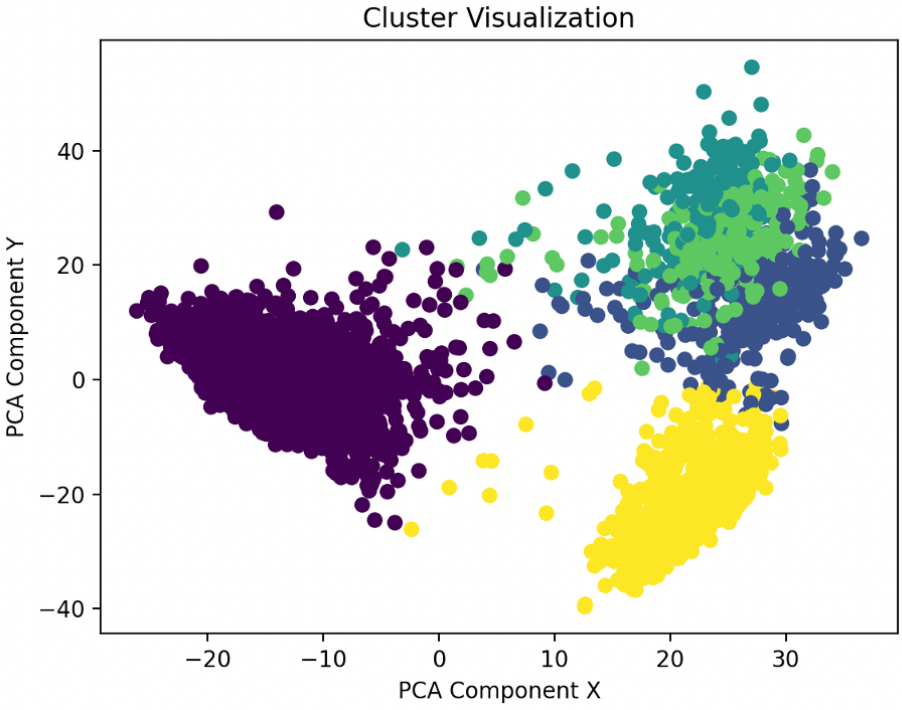

Figure 4: Kmeans++ where k = 5