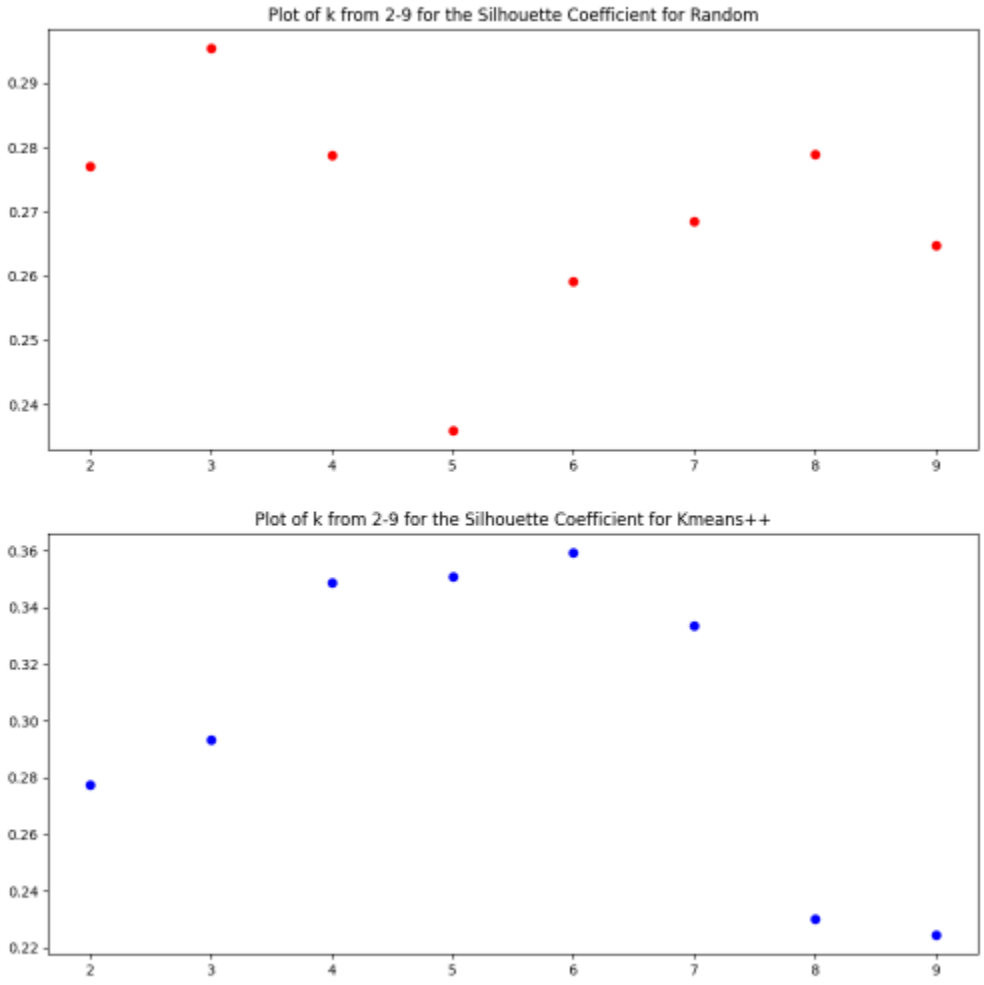
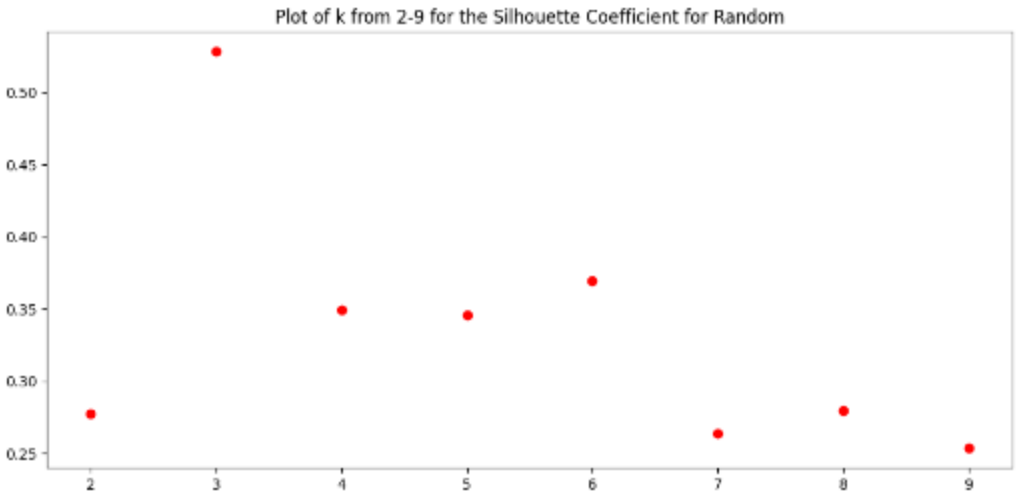
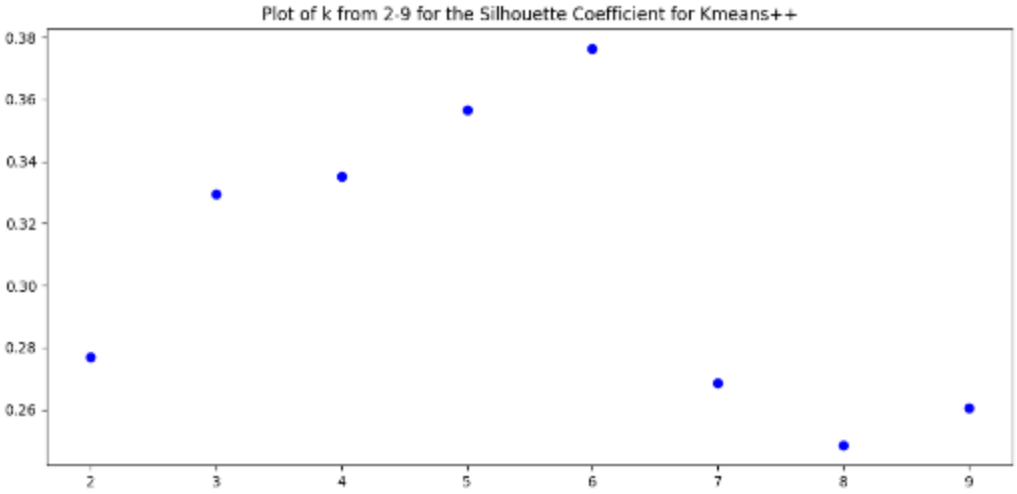
CMPT459 D100 Fall 2022
Data Mining
Instructor: Martin Ester
Assignment 2

Daniel Tham 301358046

Tasks 2 and 3. The best k for random initialization with the Silhouette Coefficient is k = 3. The best k for kmeans++ initialization with the Silhouette Coefficient is k = 6.







Here's another iteration of running the k-means on random initialization and kmeans++. It's entirely possible this could be lucky as many of my other tests had k = 3 being the best k, while some runs had the k = 2 or k =9 having the highest in the random initialization.

As for the kmeans++ initialization, even though while testing, k = 6 was more frequent, the overall silhouette coefficients seem higher than the random initialization. A reason could be the random initialization is picking random points for the initialization of the centroids, while kmeans++ will only randomize the first centroid and determine the rest through the points that are furthest away from the previous point. This will go through all the centroids in the dataset and determine the clusters more efficiently which in turn, produces higher silhouette coefficients.

The coloring of the clusters is not perfect. As we notice there are a couple of outliers between the borders of each cluster.