Project 3 Answers
Aban Khan, Sam Feld, Matt Glennon, Jay Salgado

1a. There are several assumptions that the k-means algorithm makes of the input data. Despite using 4D initialization to determine our centroids it is still possible to choose poor initial points which will affect the outcome of k-means. The algorithm also assumes that the clusters can be cleanly isolated into identifiable clusters per Voronoi partitioning, each cluster is separate, that each cluster is similar in density and size, and are spherical.

1b. K-means clustering appears to work for easygaussian1-5, easygaussian7-9, hardgaussian1-9. It appears to somewhat work for diffstddev2, easygaussian6 and stretched2. It appears to not work for bullseye2 and diffdensity2. I determined this based on the WCSS. A higher WCSS means a greater deviation of points from the centroid or overlap between clusters which is present in data files such as diffdensity2.

1c. For the problems where k-means appears to fail the primary assumption that is seemingly violated is the isolation of the data points into identifiable clusters and the data being close together. In particular, bullseye2 is a difficult distribution to work with because there is a clear cluster sitting in the center surrounded by points in a circle. As a result, the WCSS is high because we only identify 2 centroids when the distribution of the data is so far apart. Additionally, diffdensity2, while having two somewhat discernible clusters has some overlap. The green cluster towards the left has less density and points that are further away as compared to the red cluster to the right. Due to the green cluster possessing more outcast data points the WCSS increases due to the distance.

1d. Yes, it is possible for k-means to fail if no assumptions are violated. If our initial centroids are not chosen properly we can converge to a local optima rather than what the clustering should be. It might also be that the dataset we are evaluating isn't meant to be evaluated with k-means due to the distribution of data where there is no clear placement of a centroid.

2a. Iterated use of the k-means algorithm can help with quality of output clusters because it allows the algorithm to run multiple attempts in order to find a smaller WCSS which can increase the quality of our output. It's possible that if we run it only once our centroids were ill-placed and didn't choose the most optimal placement. By repeating this process over and over again we are allowing multiple trials to choose the best possible representation of our data as we can.

2b. In general, it appears that iterations of k-means only improved a few select cases. bullseye2 and diffdensity2 weren't affected, however, easygaussian5 resulted in 0.51 which was the minimum value of the three runs done in question 1. This trend of the smallest value from the initial 3 trials being the min WCSS value of this question continues with easygaussian6-easygaussian9. All of the hardgaussian files remained the same, however.

2c. The iterated approach helps for clustering problems where we have multiple local optima or if the clusters are overlapping with each other.

3b. When running this technique, every single test for every single file returned 10 clusters except for Run #8 k on diffstddev2 and easygaussian3. As was mentioned in office hours, we checked the functions for calculating the min and max values for each dimension and the function which generates random data. We confirmed these were working properly. The data sets I expect for this technique to consistently succeed in are those that follow the assumptions of k-means: the clusters can be cleanly isolated into identifiable clusters per Voronoi partitioning, each cluster is separate, that each cluster is similar in density and size, and are spherical. These assumptions mean each cluster is identifiable and thus the assumed clusters in a dataset would be accurate. Data sets in this project that would represent this are easygaussian1-5, easygaussian7-9 and hardgaussian1-9.

3c. The data sets I expect to vary for a discerned number of clusters are those that don't necessarily violate the assumptions made but make it difficult to discern clearly where data might belong. Ideally, the k-values will vary from run to run. One example are data sets where the data has uneven densities or sizes such as diffdensity2 or datasets where the data slightly overlaps such as easygaussian8.

3d. The datasets I expect to be incorrect are those that return the wrong k-assumption. Data sets that are heavily spread out, non-spherical or difficult to discern. One example are data sets where the data is spread out such as hardgaussian1 or datasets where the data overlaps such as hardgaussian3.