(a)(b) If you were to choose the best K for this data based on the the plot you generated in part (i), what value of K might you choose? Justify your choice.

Based on the few plots of K versus MSE we generated on different runs through the algorithm, I might choose 8 as the value for K. This is because 8 has a low MSE on all three plots. However, on the second plot, MSE increases from K= 5 to 6 before decreasing at K = 7. The second plot also has MSE increasing slightly at K = 9, then decreasing although on the other two plots they decrease at this value of K. I would choose 8 because it seems to avoid the increase in K=6 and also avoids overfitting effects by choosing a larger value of K.

(b)(a) What differences do you notice between the clusters produced using the min versus the max distance metric? Does this make sense given the deﬁnition of the metrics?

The clusters for max in general cover more data points while in min there are two large clusters and two clusters each containing just one data point. In doing so, max splits a cluster that min recognized as a single cluster into three different clusters. This makes sense as min will only look at minimum distances between clusters, which preclude the two outlier-like points from ever being merged into a cluster since they are so far away. Max does include these two outliers because they lie closer to a large cluster than the large clusters they belong to lie to each other. The clusters in the middle in max reach an equilibrium of how to divide up into 3 different clusters once their maxes are far enough apart. They fit our expectations in that min should produce elongated clusters that create separate clusters for outliers while max produce compact clusters.

(b)(b) What differences do you notice between the clusters produced using the mean versus the centroid distance metric? Does this make sense given the deﬁnition of the metrics?

The mean and centroid clusters are fairly similar, the only difference being the location of the green cluster: slightly on the left arm of the top cluster in mean, and on the right arm of the top cluster in centroid. This makes sense as both of them are expected to be tradeoffs between compactness and elongation. We would expect them to have similar-looking clusters with small differences since the data given does not have huge variation within clusters and looks like it only has a low number of true clusters.