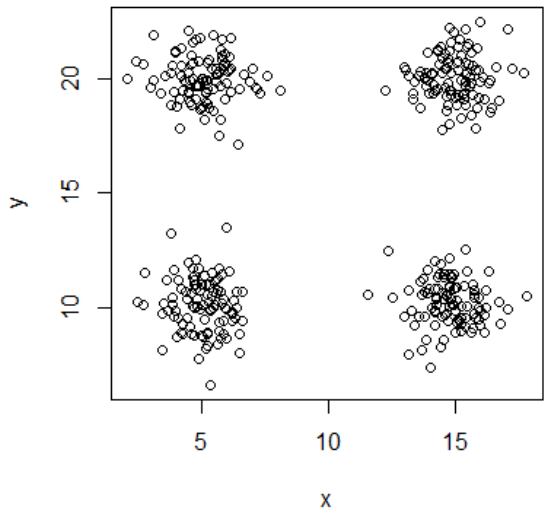
Math 5364 Homework 17

1. Generate a data set similar to the one displayed below, where each of the four clusters has 100 points.



   (a) Perform a K-means clustering with $K = 4$ and 1000 repetitions.

   (b) Plot the points and color them based on which clusters they are in.

   (c) Find the total, total within, and between sums of squares.

   (d) Find the centers of the clusters.

   (e) Suppose we want at least one of our repetitions of K-means to have the property that every cluster contains exactly one initial centroid. How many repetitions would be necessary to ensure that this happens with at least 99% probability?

   (f) Can you find a choice of initial centers that does not result in the optimal clusters? What is the total within sum of squares for that clustering?

2. Perform a K-means clustering with $K = 2$ and 1000 repetitions for the `wdbc` data set. What classification accuracy would be achieved if the clusters were used to predict the diagnosis in this data set?