# Question 5

Yufei Yin

## Question 5

- Consider the following two points:

  The kmeans algorithm is guaranteed to converge after a finite number of iterations (that is, if iters is large enough, eventually the two steps of the kmeans iterations will not change the cluster locations or the cluster assignments of the data items).

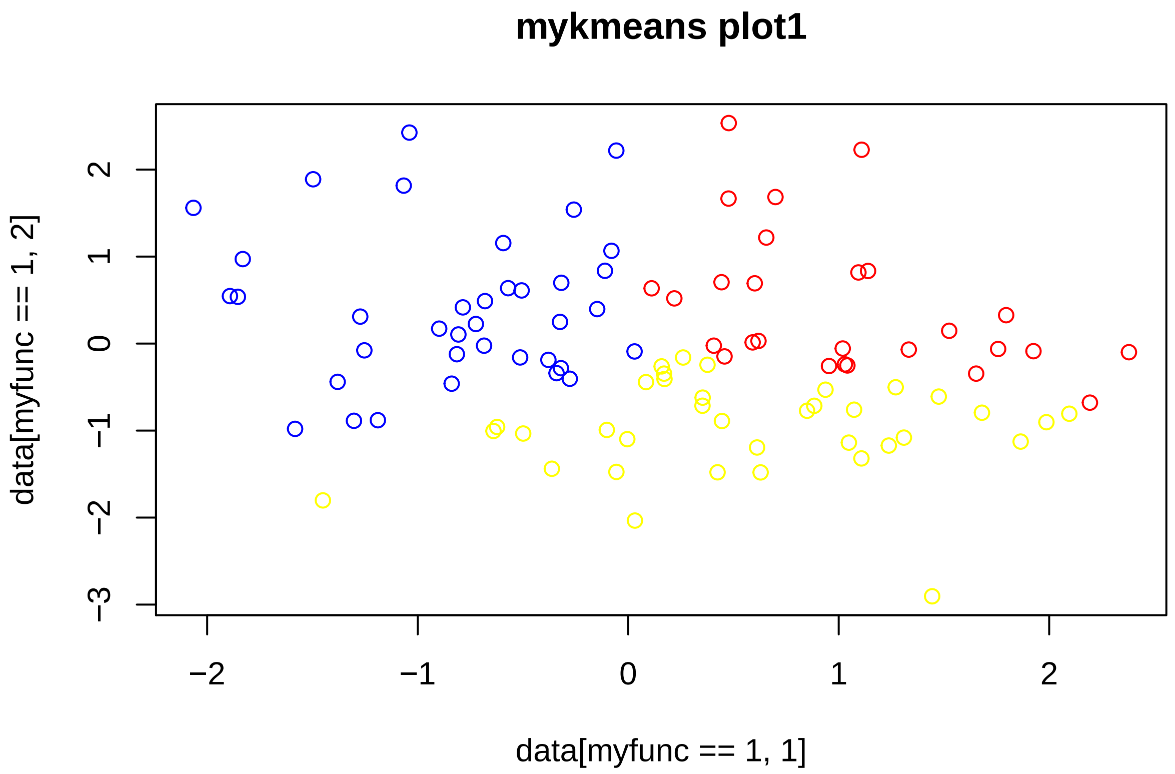
  If the kmeans algorithm is run twice with two different random initializations, then the solutions the two runs converge to may be different.

- Create a (small?) dataset that demonstrates the phenomenon whereby two runs with two different random initializations converge to two different solutions. You may create the dataset any way you want, and fix the random initializations any way you want. Provide the dataset and the initializations and code and plots demonstrating the difference.

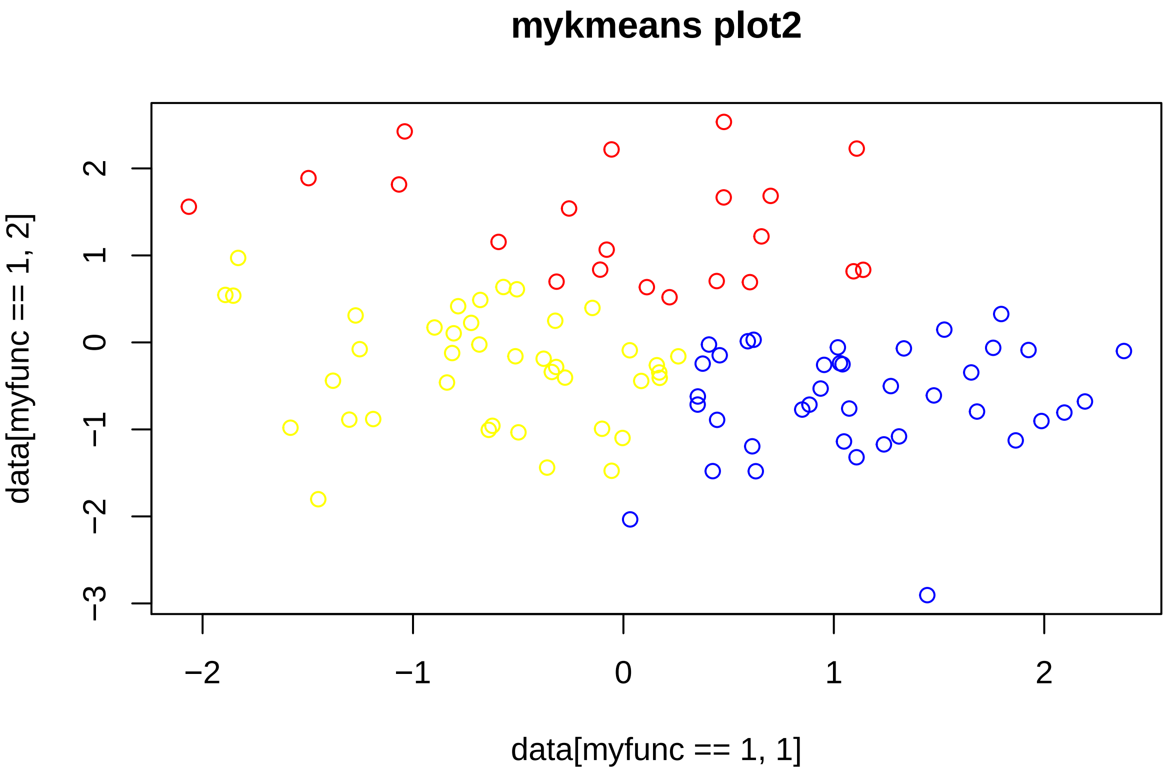
```r
mykmeans = function(x, k, iters){
  N = dim(x)[1]
  D = dim(x)[2]

  centres = matrix(NA, k, D)
  clusters = rep(NA, N) # each entry between 1 and K

  for (i in 1:N){
    clusters[i] = sample.int(k, 1)
  }

  for (iter in 1:iters){
    for (k in 1:k){
      for (d in 1:D){
        centres[k, d] = mean(x[clusters == k, d])
      }
    }
    distanceMatrix <- matrix(NA, nrow=N, ncol=k)
    for(i in 1:k) {
      distanceMatrix[,i] <- sqrt(rowSums(t(t(x)-centres[i,])^2))
    }
    clusters <- apply(distanceMatrix, 1, which.min)
    centres <- apply(x, 2, tapply, clusters, mean)
  }
  return(list(locations=centres, assignment=clusters))
}
```

```r
data = data.frame(x = rnorm(100), y = rnorm(100))
set.seed(240)
res = mykmeans(data, 3, 1000)
```

```
myfunc = res$assignment
plot(data[myfunc == 1, 1], data[myfunc == 1, 2], col = "blue",
     xlim = range(data[,1]), ylim = range(data[,2]),
     main = "mykmeans plot1")
points(data[myfunc == 2, 1], data[myfunc == 2, 2], col = "red")
points(data[myfunc == 3, 1], data[myfunc == 3, 2], col = "yellow")
```



```
set.seed(123)
res = mykmeans(data, 3, 1000)
myfunc = res$assignment
plot(data[myfunc == 1, 1], data[myfunc == 1, 2], col = "blue",
     xlim = range(data[,1]), ylim = range(data[,2]),
     main = "mykmeans plot2")
points(data[myfunc == 2, 1], data[myfunc == 2, 2], col = "red")
points(data[myfunc == 3, 1], data[myfunc == 3, 2], col = "yellow")
```


mykmeans plot2
data[myfunc == 1, 2]
2
1
0
−1
−2
−3
−2
−1
0
1
2
data[myfunc == 1, 1]