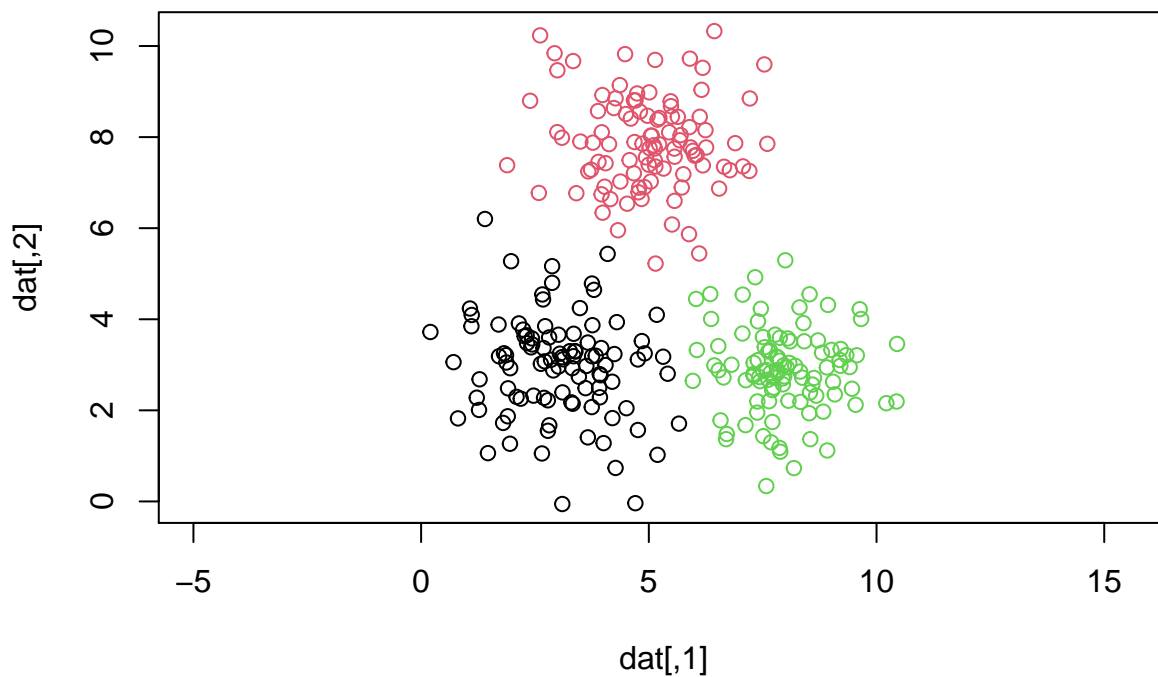
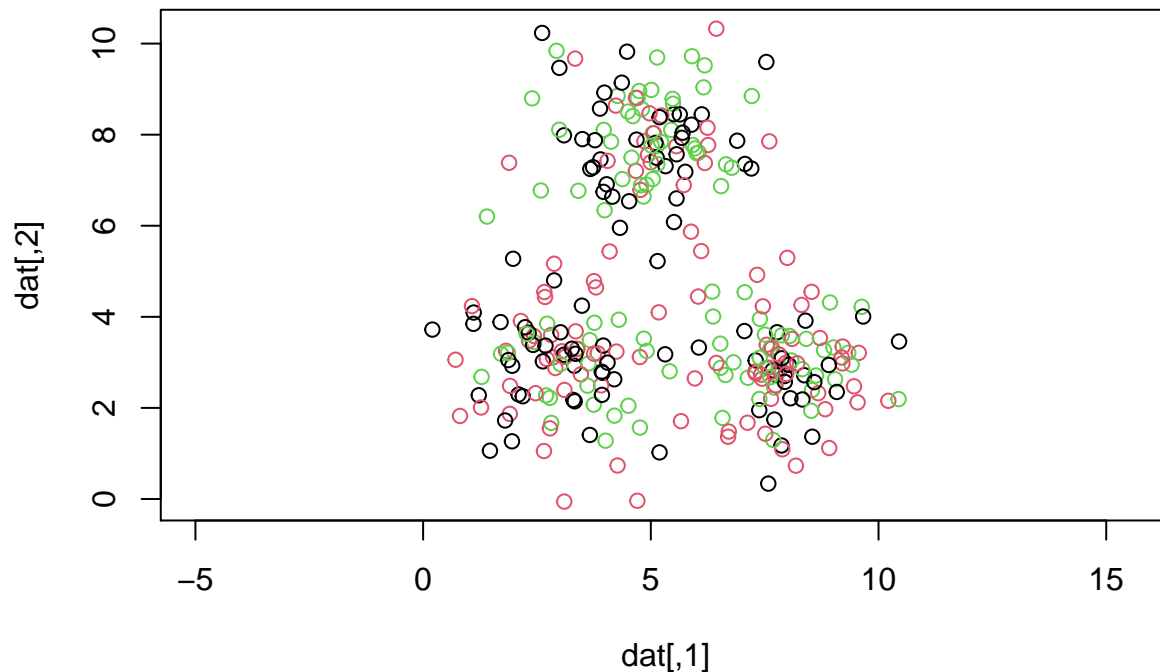


## K-means Clustering

```
#Data
set.seed(2020)
RNGkind(sample.kind = "Rejection")
library(mvtnorm)
cv <- matrix(c(1, 0, 0, 1), ncol = 2)
j <- rmvnorm(100, mean = c(3, 3), sigma = cv)
k <- rmvnorm(100, mean = c(5, 8), sigma = cv)
l <- rmvnorm(100, mean = c(8, 3), sigma = cv)
dat <- rbind(j, k, l)
true_groups <- as.factor(c(rep("j", 100), rep("k", 100), rep("l", 100)))
plot(dat, col=true_groups, asp = 1)
```



```
set.seed(2020)
assignments <- factor(sample(c(1, 2, 3), 300, replace = TRUE)) # initial groupings
plot(dat, col = assignments, asp = 1) #initial plots
```



```
#K-means Clustering Algorithm
k <- 3
set.seed(2019)
assignments <- factor(sample(c(1,2,3), 300, replace = TRUE))
distances <- function(point, means){
  outputs <- matrix(NA, nrow=nrow(point), ncol=nrow(means))
  for (i in 1:nrow(means)) {
    centroid <- means[i,]
    outputs[i,] <- apply((dat=rep(1, nrow(point))%o%centroid)^2, 1, sum)
  }
  return(outputs)
}

centroids <- function(point) {
  clusters <- levels(assignments)
  outputs <- matrix(NA, nrow=k, ncol=ncol(point))
  for (i in 1:length(clusters)) {
    cluster <- clusters[i]
    outputs[i,] <- apply(point[assignments == cluster, ], 2, mean)
  }
  return(outputs)
}

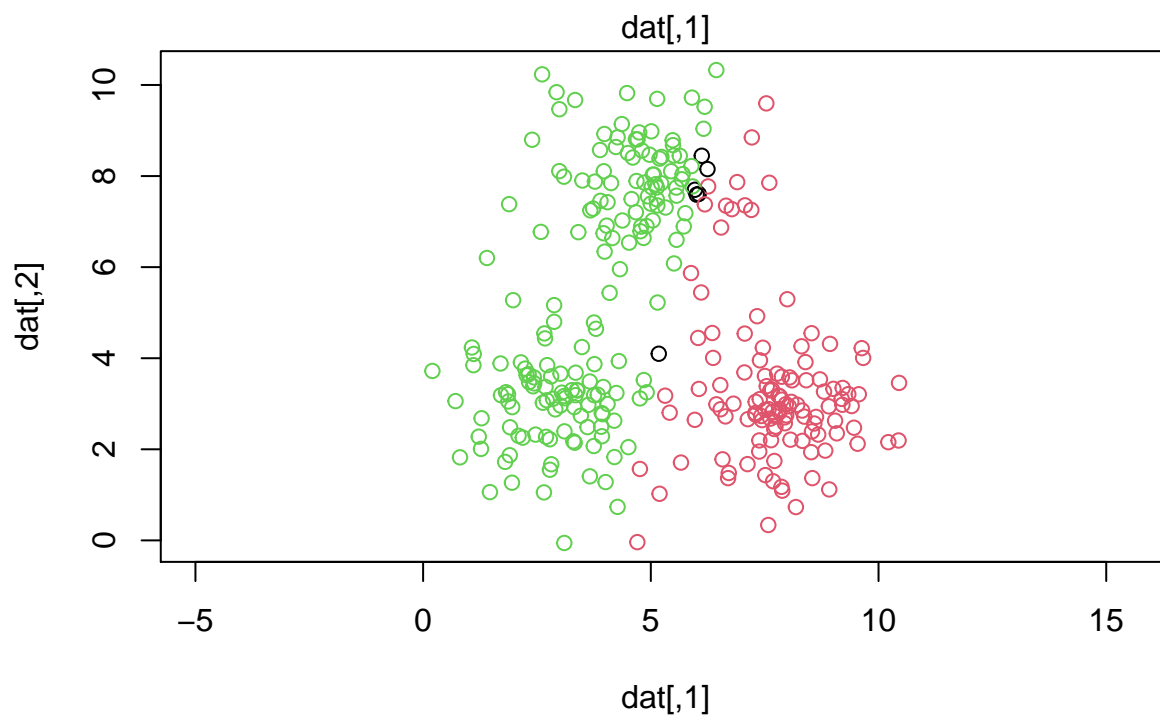
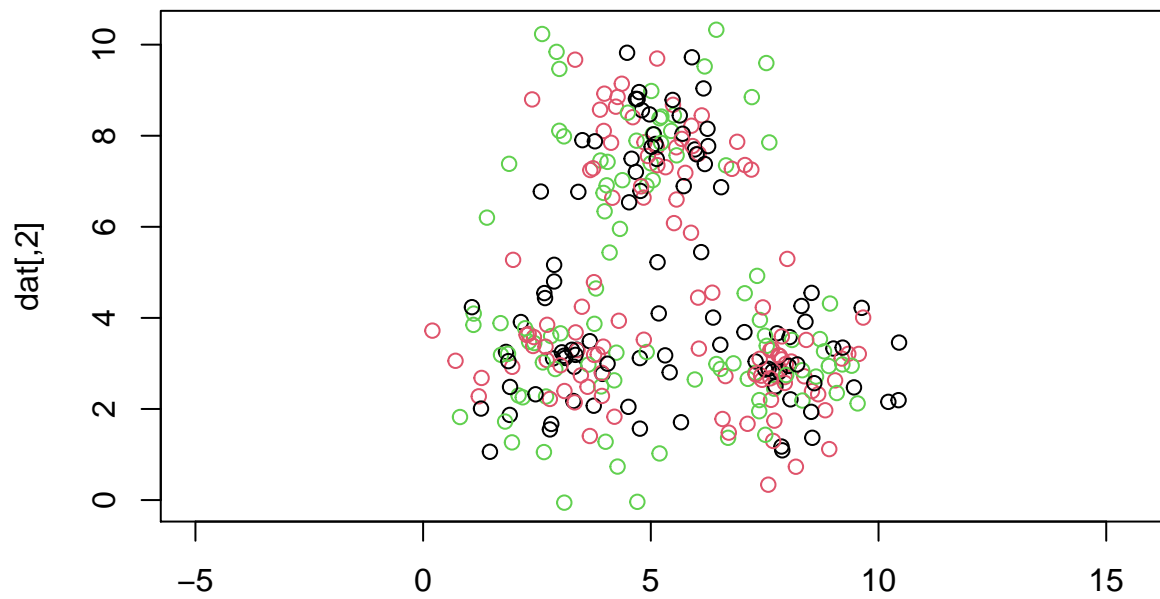
d <- distances(dat, centroids(dat))

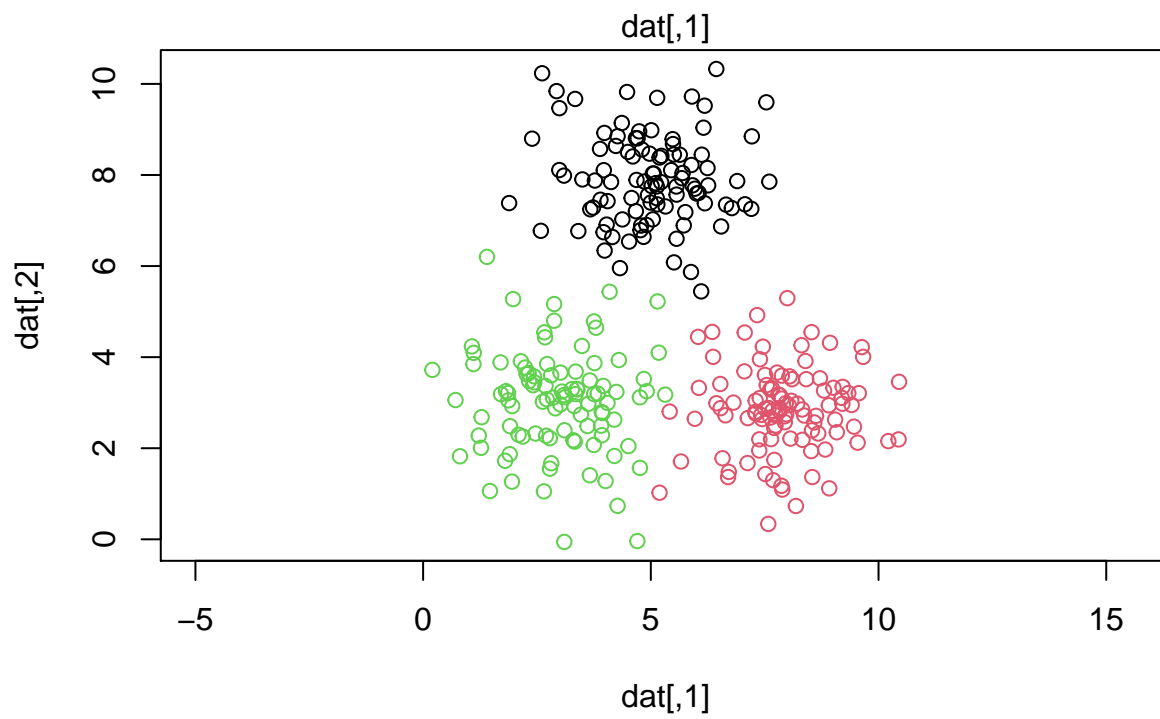
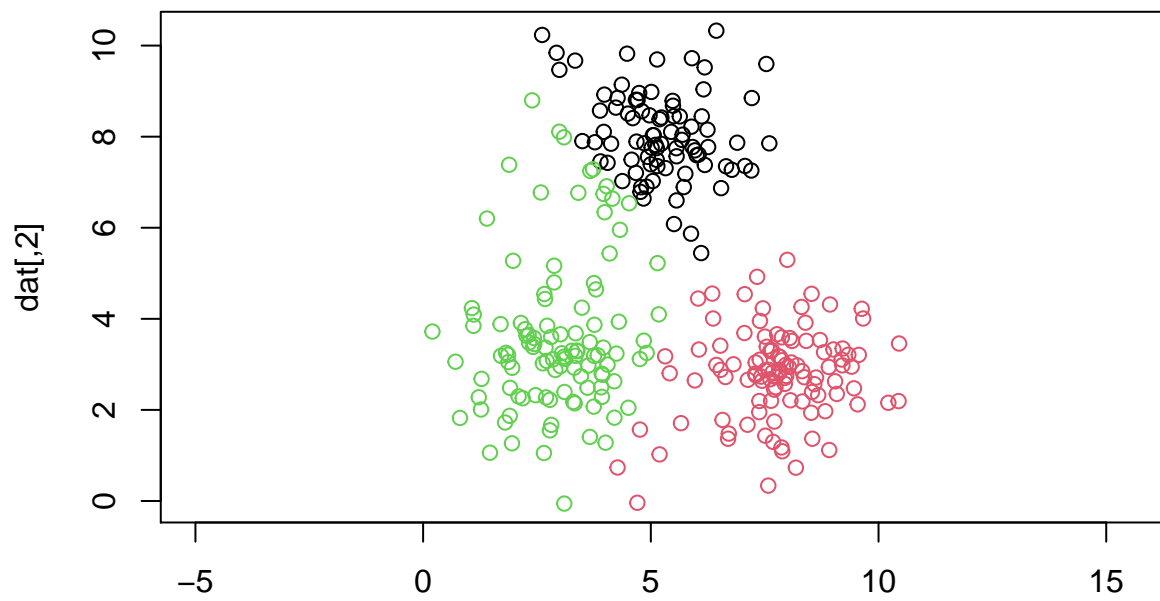
converged = FALSE
i = 0
while(!converged){
  plot(dat, col = assignments, asp = 1)
  d <- distances(dat, centroids(dat))
  prev.assignments <- assignments
  assignments <- factor(apply(d, 1, which.min))
}
```

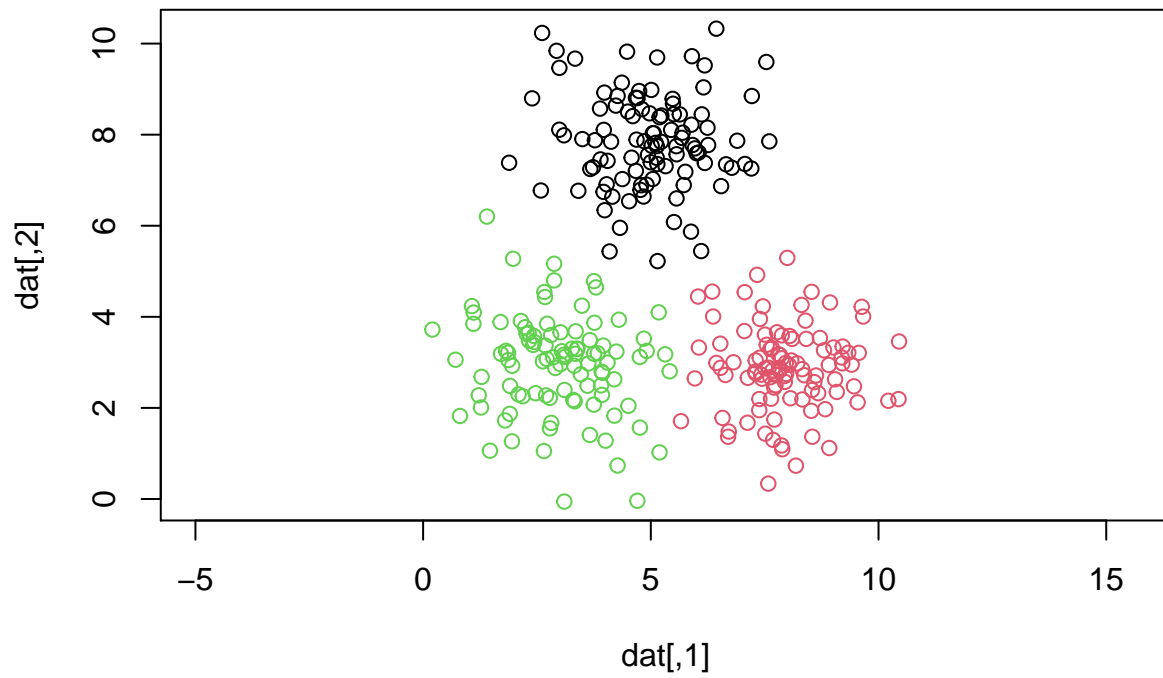
```

i <- i + 1
if (all(prev.assignments == assignments)) {
  converged = TRUE
}
}

```







```
plot(dat, col = assignments, asp = 1)
assignments <- factor(as.character(assignments), levels=c(3,1,2), labels=c("j", "k", "l"))
table(assignments, true_groups)
```

```
##           true_groups
## assignments  j    k    l
##           j  98   0   0
##           k   1 100   0
##           l   1   0 100
```