R Notebook

PCA

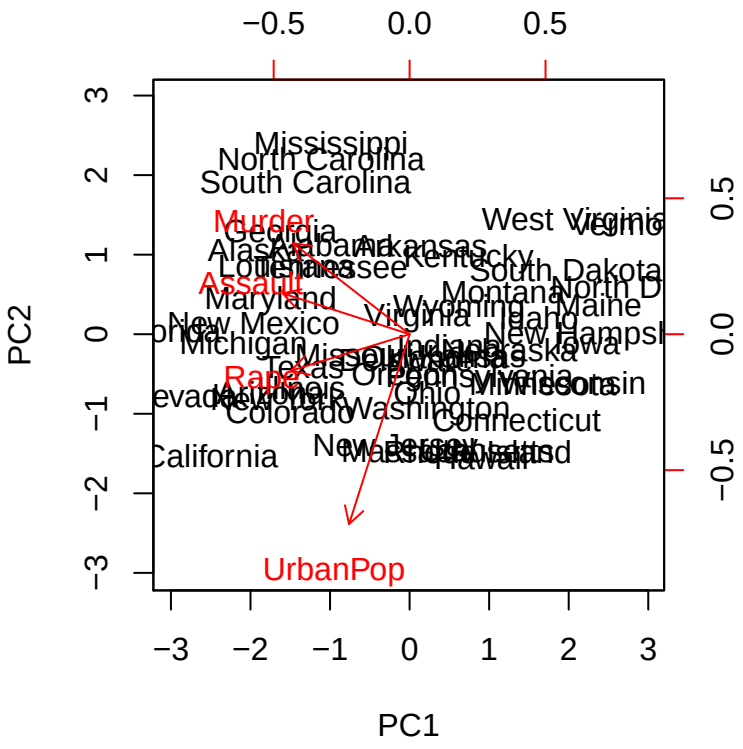
```
USArrests %>% glimpse
```

```
## Observations: 50
## Variables: 4
## $ Murder   <dbl> 13.2, 10.0, 8.1, 8.8, 9.0, 7.9, 3.3, 5.9, 15.4, 17.4,...
## $ Assault  <int> 236, 263, 294, 190, 276, 204, 110, 238, 335, 211, 46,...
## $ UrbanPop <int> 58, 48, 80, 50, 91, 78, 77, 72, 80, 60, 83, 54, 83, 6...
## $ Rape     <dbl> 21.2, 44.5, 31.0, 19.5, 40.6, 38.7, 11.1, 15.8, 31.9,...
```

```
pr<-prcomp(USArrests, scale=TRUE)

pr %>% summary
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4
## Standard deviation     1.5749 0.9949 0.59713 0.41645
## Proportion of Variance 0.6201 0.2474 0.08914 0.04336
## Cumulative Proportion  0.6201 0.8675 0.95664 1.00000
```

```
biplot(pr, scale=0)
```



Variance explained

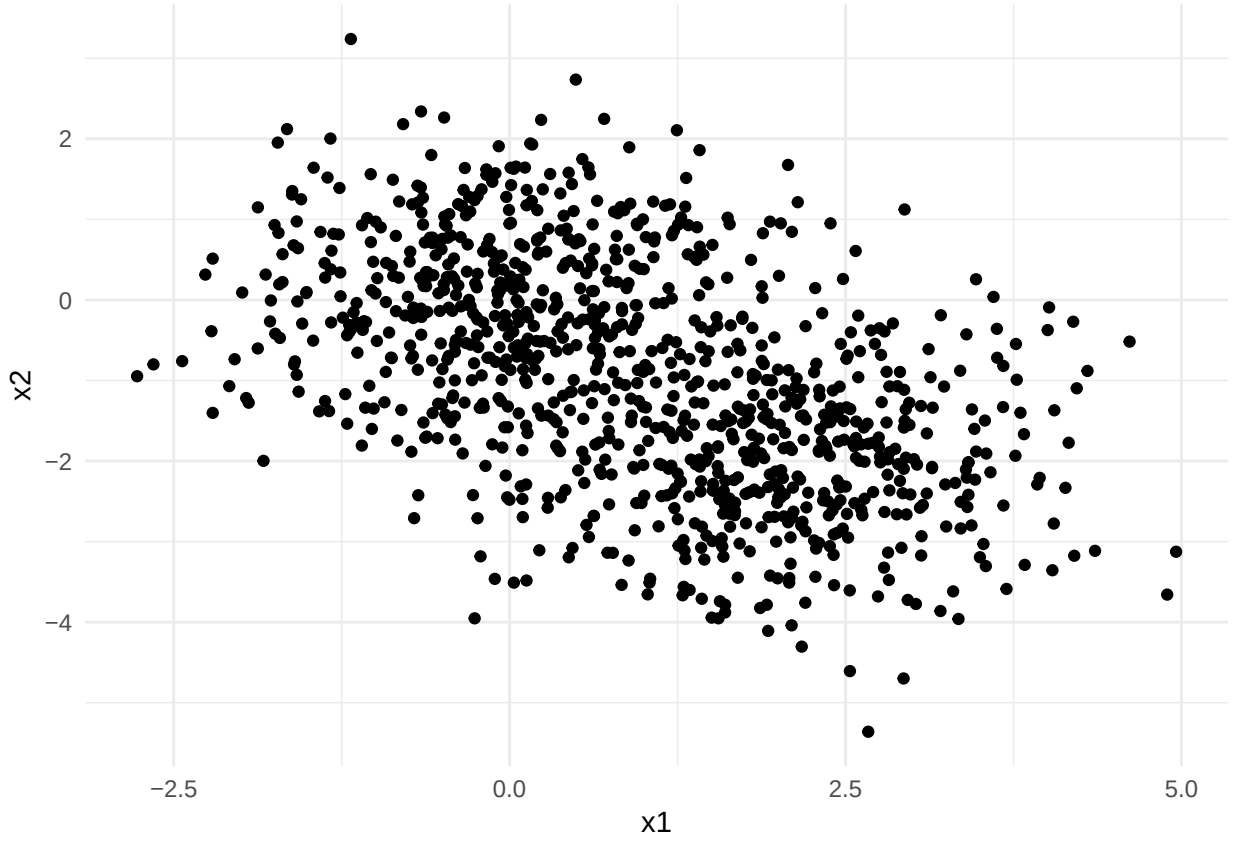
```
pr$sdev
```

```
## [1] 1.5748783 0.9948694 0.5971291 0.4164494
```

Clustering

K Means Clustering

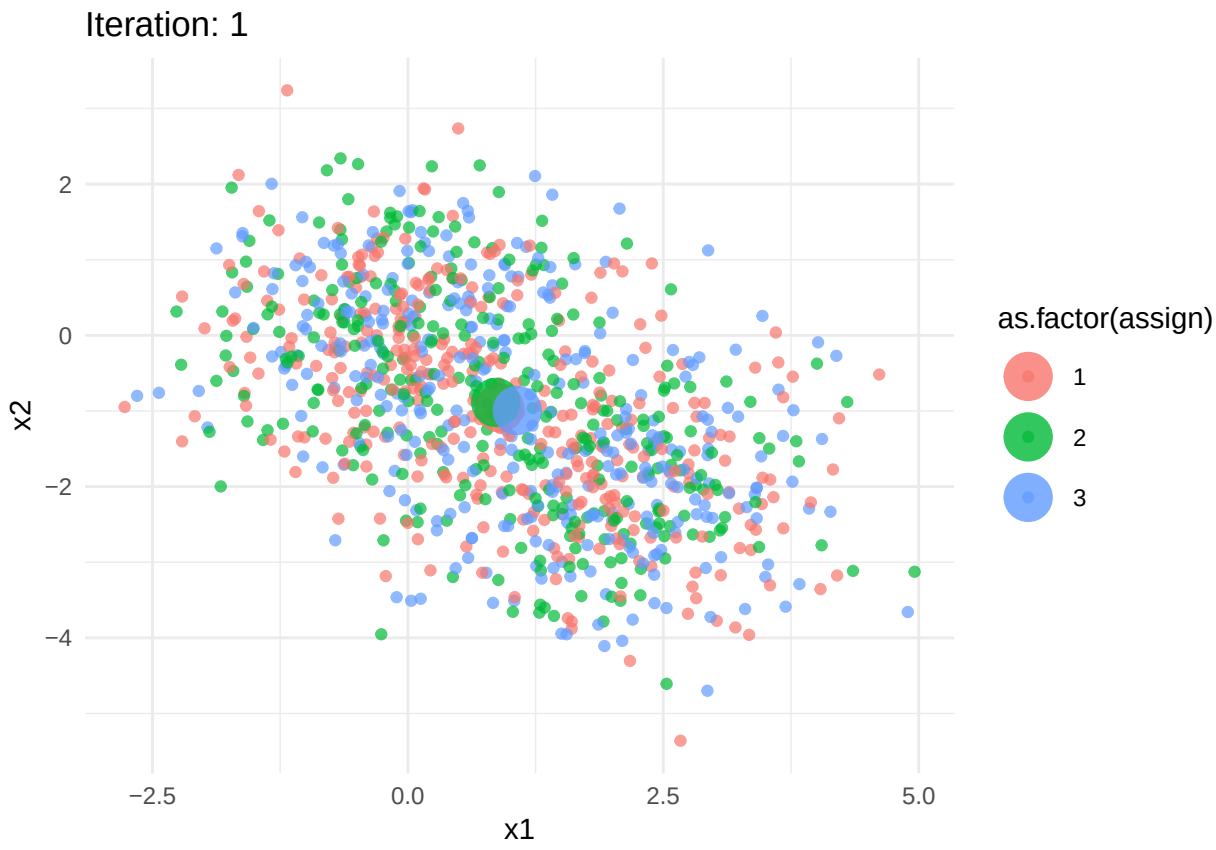
```
d <- data.frame(
  x1 = rnorm(n = 1000),
  x2 = rnorm(1000)
)

cluster <- rnorm(nrow(d)) > 0


d <- d %>%
  mutate(x1 = ifelse(cluster, x1 + 2, x1),
         x2 = ifelse(cluster, x2 -2, x2))

d %>%
  ggplot(aes(x = x1, y = x2)) +
  geom_point()
```



Randomily assign each point to K clusters

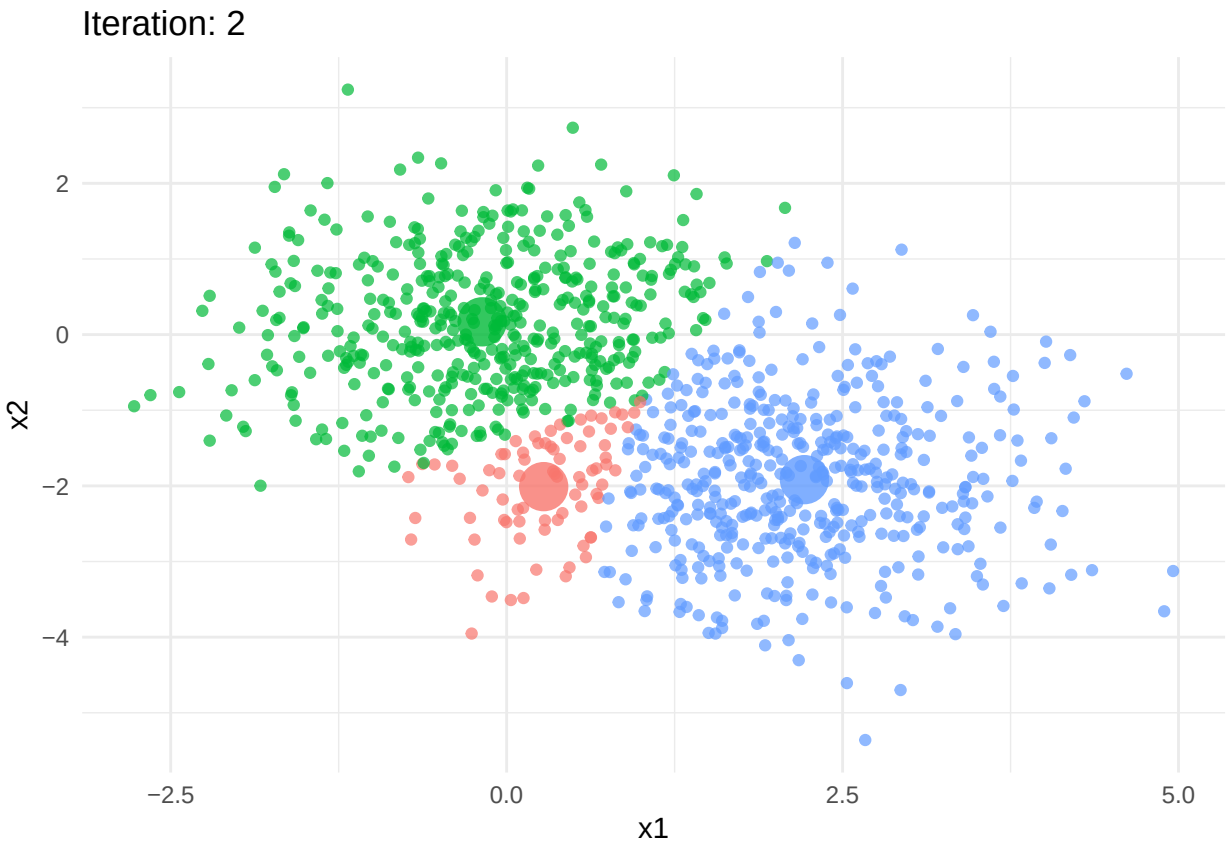
```
k <- 3
k_seq <- 1:k

d$assign <- sample(k_seq, size = nrow(d), replace = T)
```

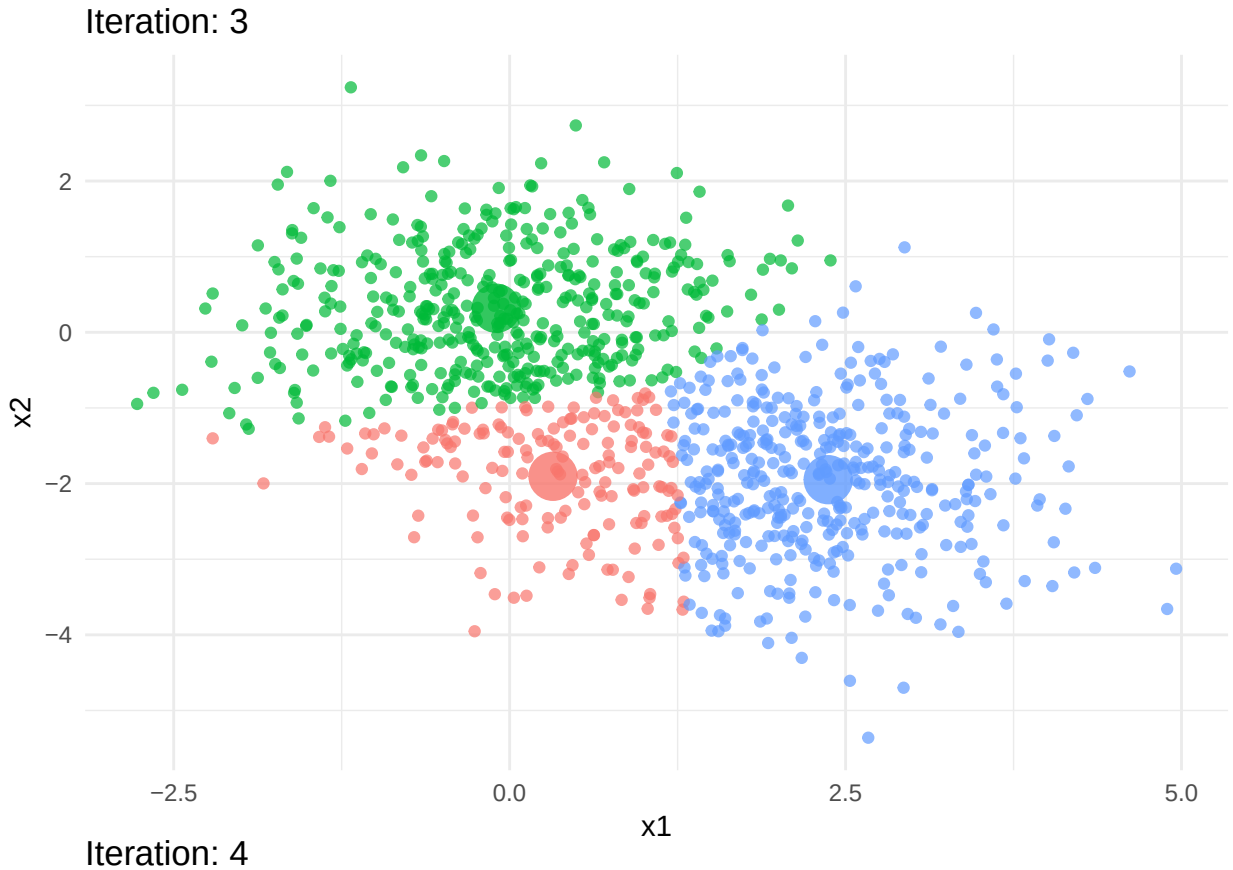
```r
centroids <- d %>%
  group_by(assign) %>%
  summarise_at(vars(x1, x2), funs(cent = mean))

iteration <- 1
d %>%
  ggplot(aes(x = x1, y = x2, color = as.factor(assign))) +
  geom_point( alpha = 0.7) +
  geom_point(aes(x = x1_cent, y = x2_cent), data = centroids, size = 8, alpha = 0.8) +
  labs(title = paste0("Iteration: ", iteration))
```

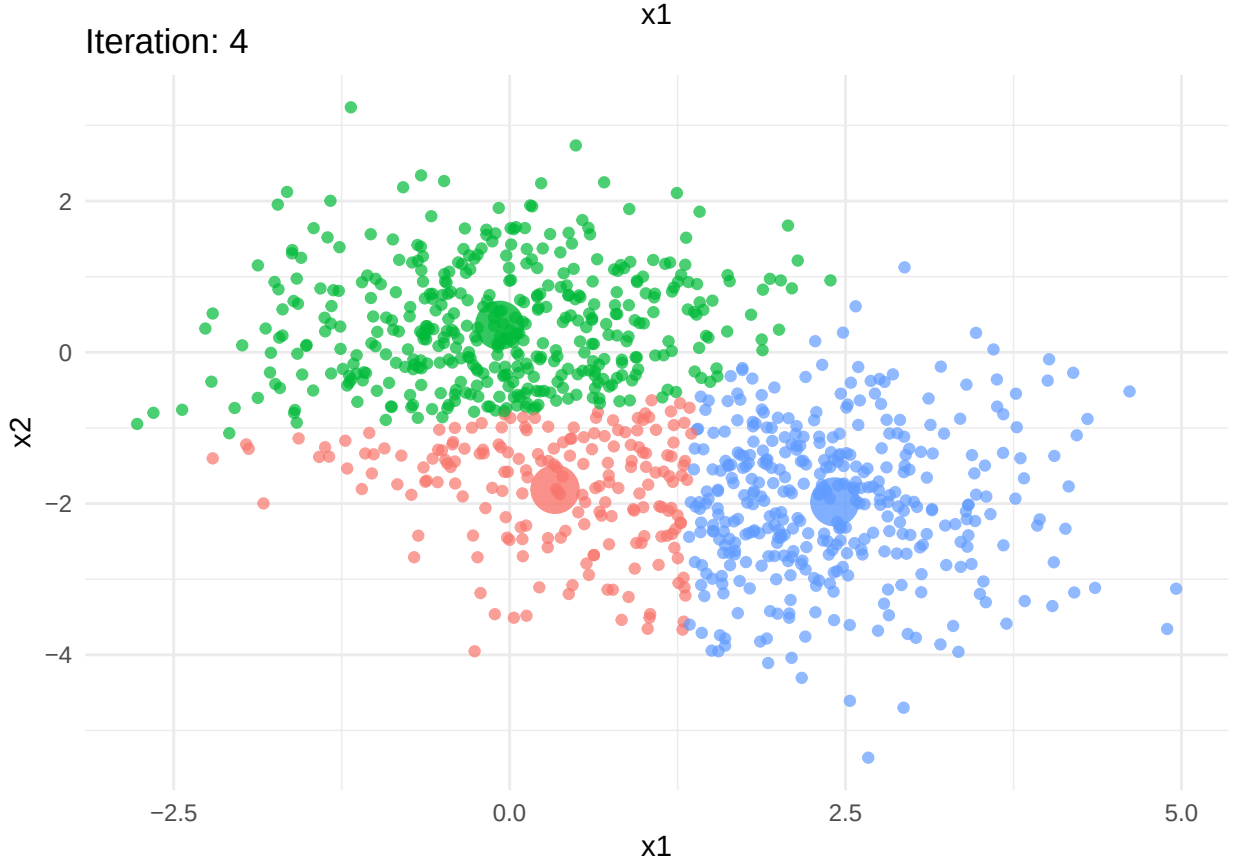


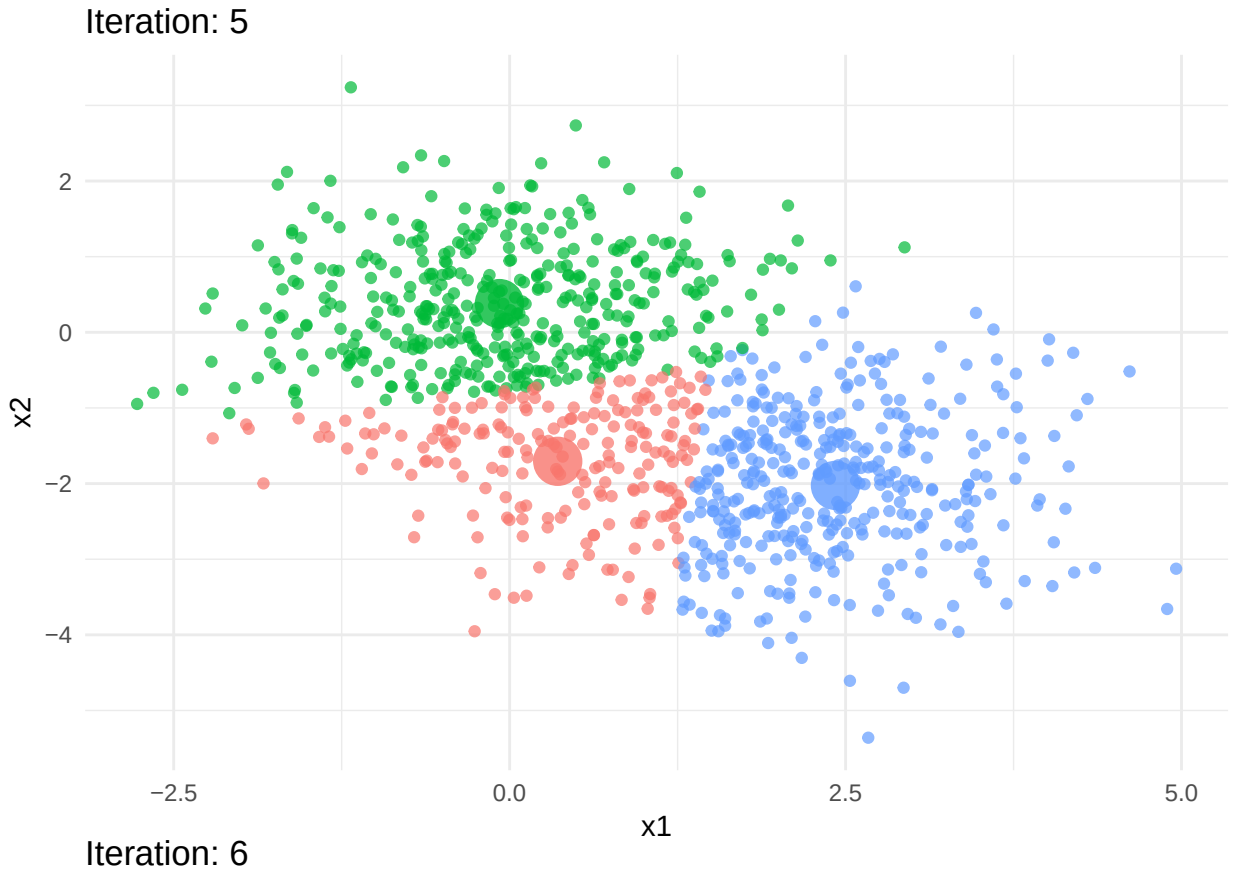

```r
d$assign <- NULL
cent_old <- data.frame()

while(is.character(all_equal(cent_old, centroids)) & iteration <= 100){
  iteration <- iteration + 1
  d <- d %>%
    crossing(centroids) %>%
    mutate(euc_dist = sqrt((x1_cent - x1)^2 + (x2_cent - x2)^2)) %>%
    group_by(x1,x2) %>%
    filter(euc_dist == min(euc_dist)) %>%
    select(x1, x2, assign)

  cent_old <- centroids
  centroids <- d %>%
    group_by(assign) %>%
    summarise_at(vars(x1, x2), funs(cent = mean))
```

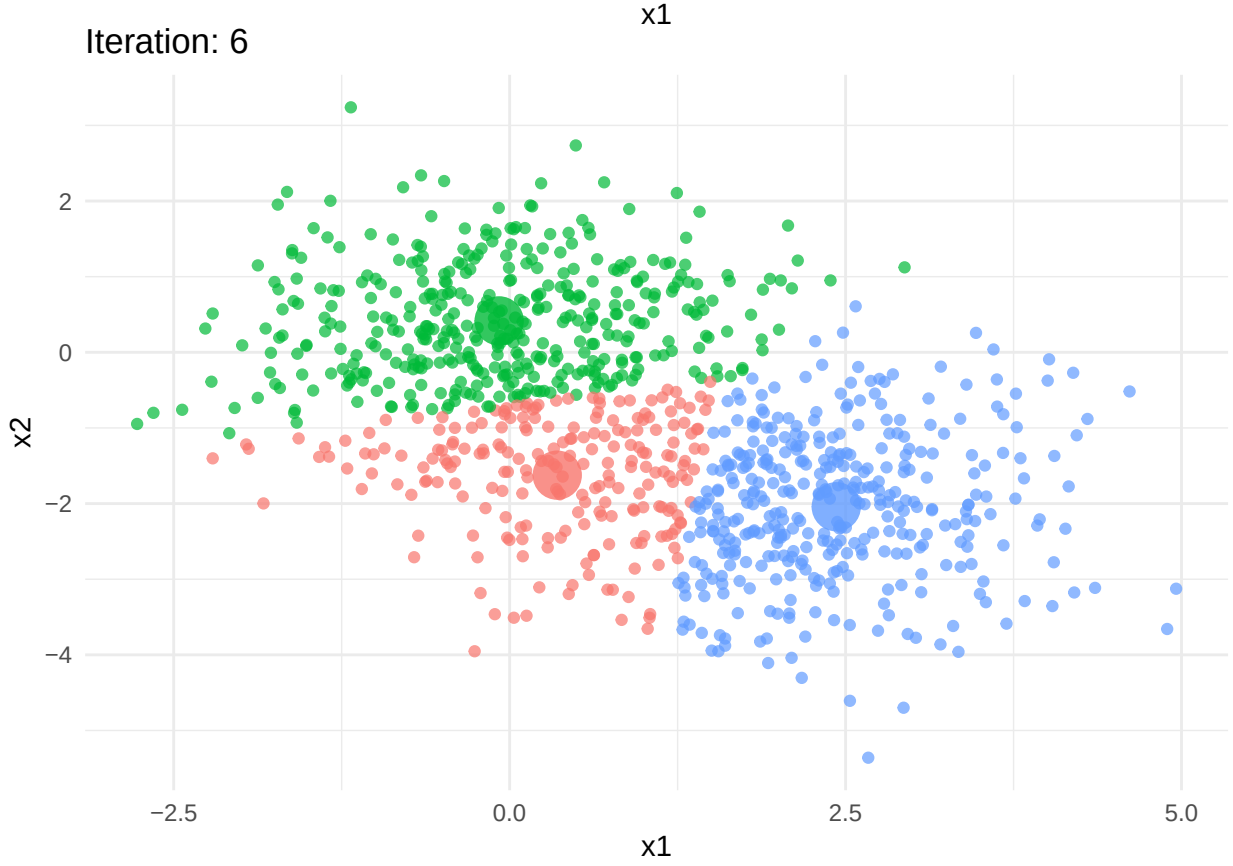
```
  d_plot <- d %>%
    ggplot(aes(x = x1, y = x2, color = as.factor(assign))) +
    geom_point( alpha = 0.7) +
    geom_point(aes(x = x1_cent, y = x2_cent), data = centroids, size = 8, alpha = 0.8)  +
  labs(title = paste0("Iteration: ", iteration)) +
    theme(legend.position = "none")
  print(d_plot)
  d$assign <- NULL
}
```

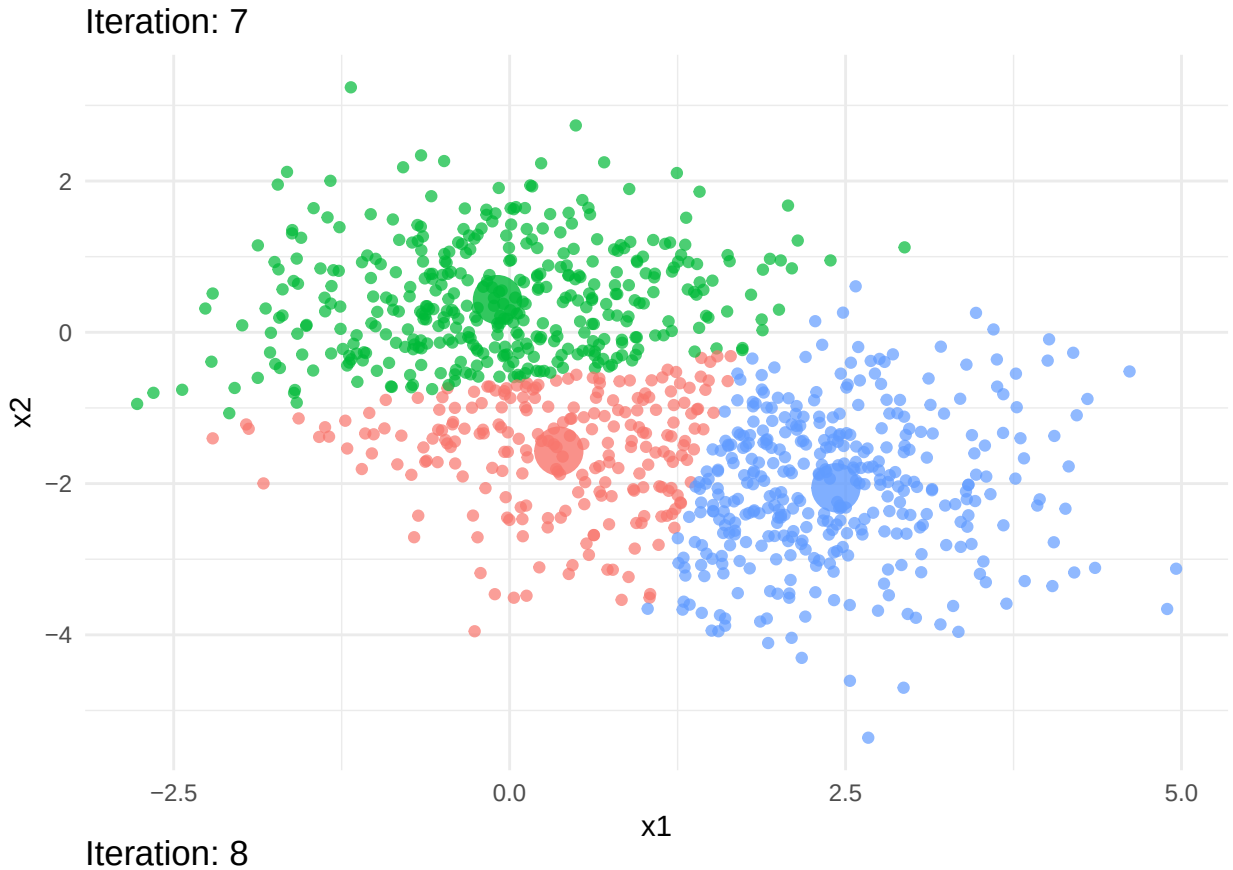


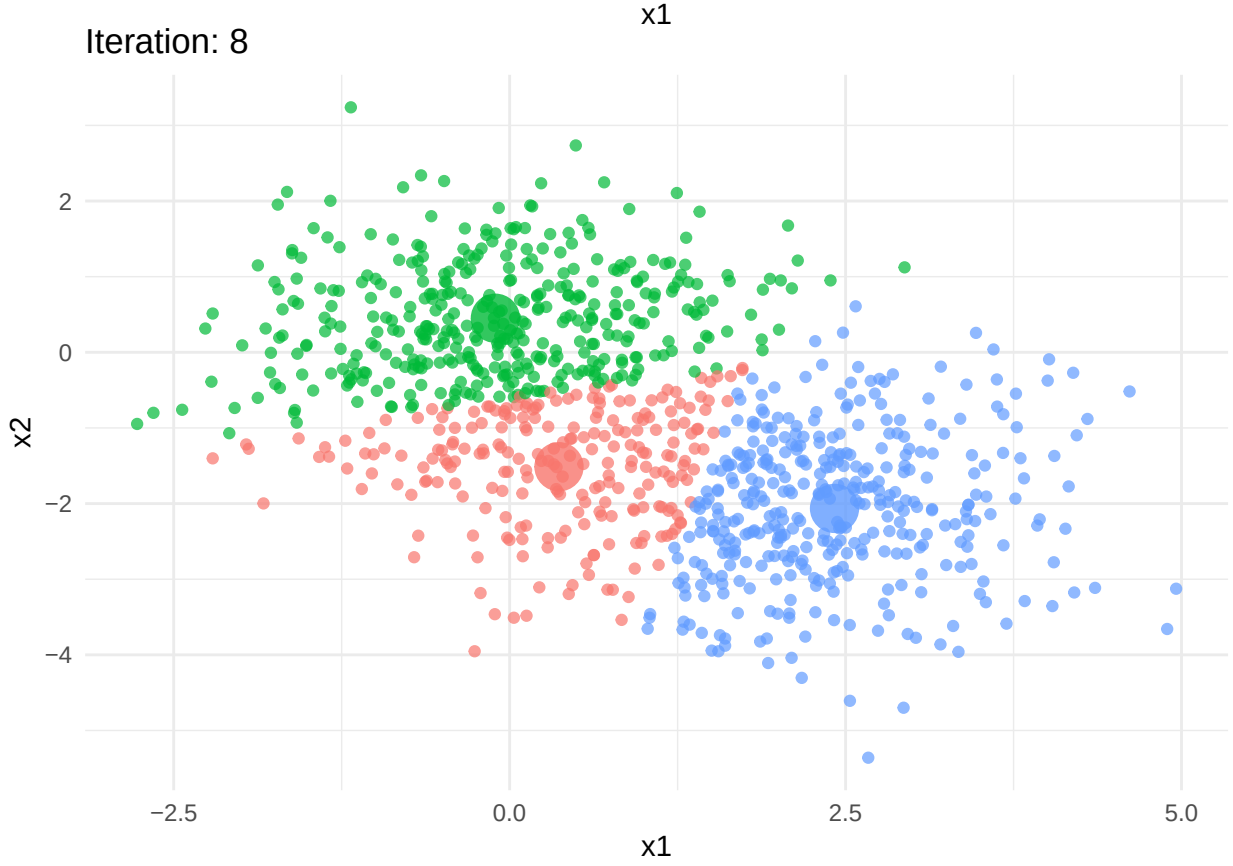

Iteration: 3
2
0
−2
−4
x2
−2.5
0.0
2.5
5.0
x1

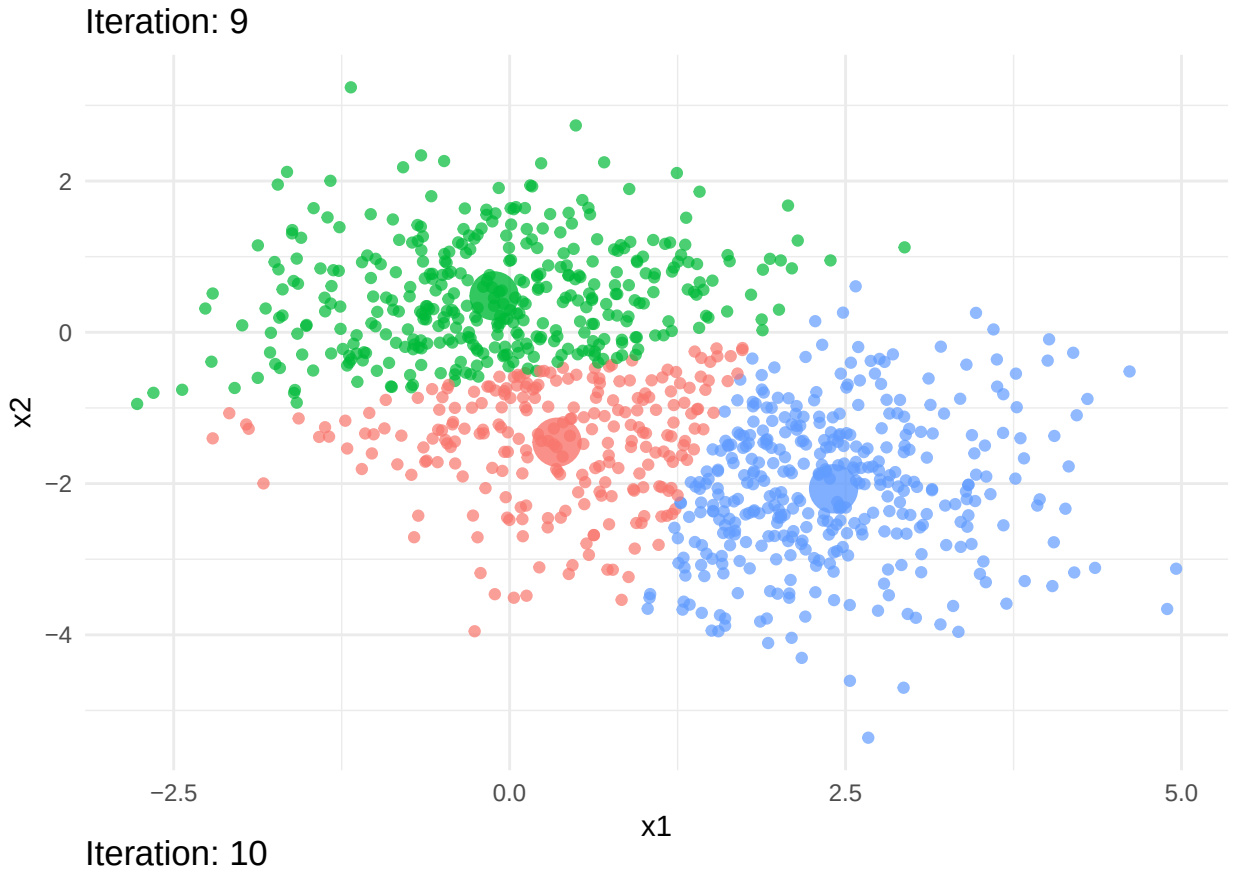

Iteration: 4
2
0
−2
−4
x2
−2.5
0.0
2.5
5.0
x1

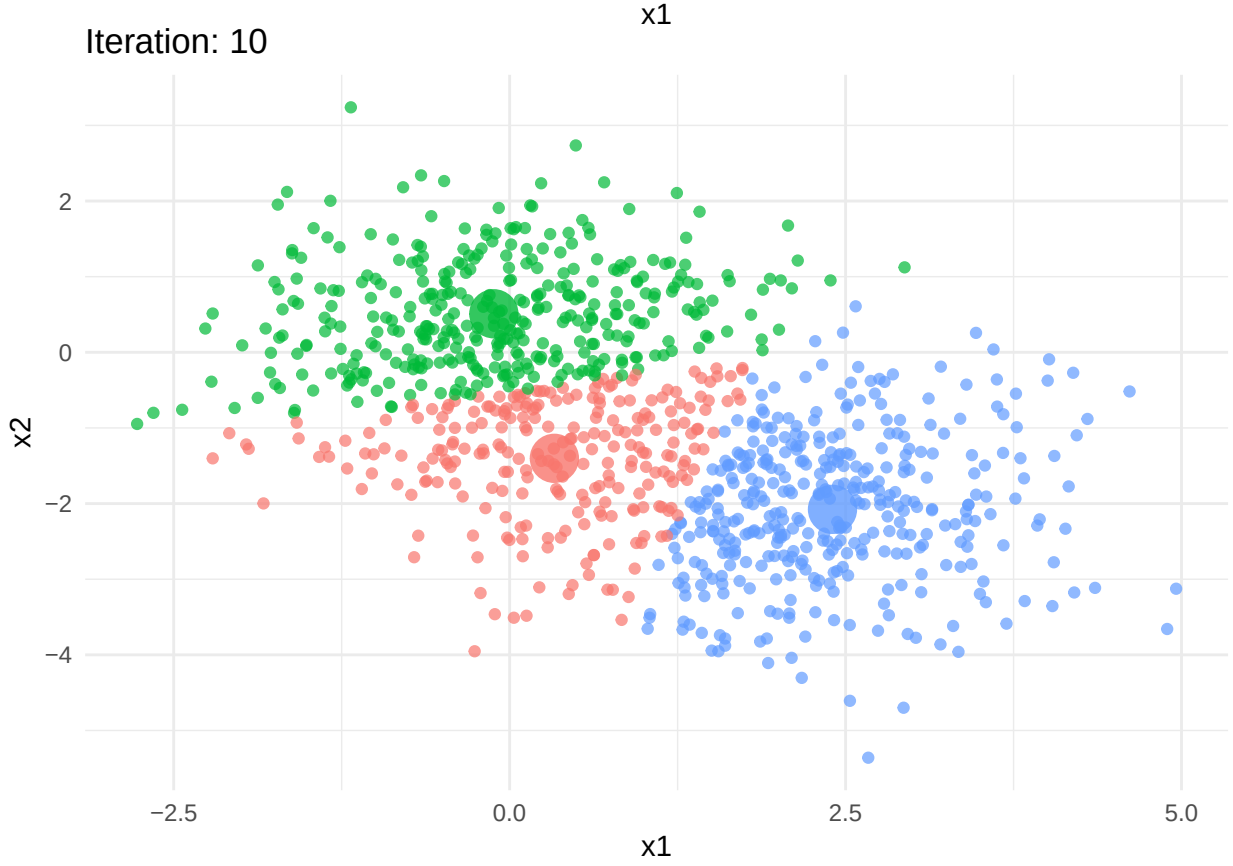

Iteration: 5
2
0
−2
−4
x2
−2.5
0.0
2.5
5.0
x1

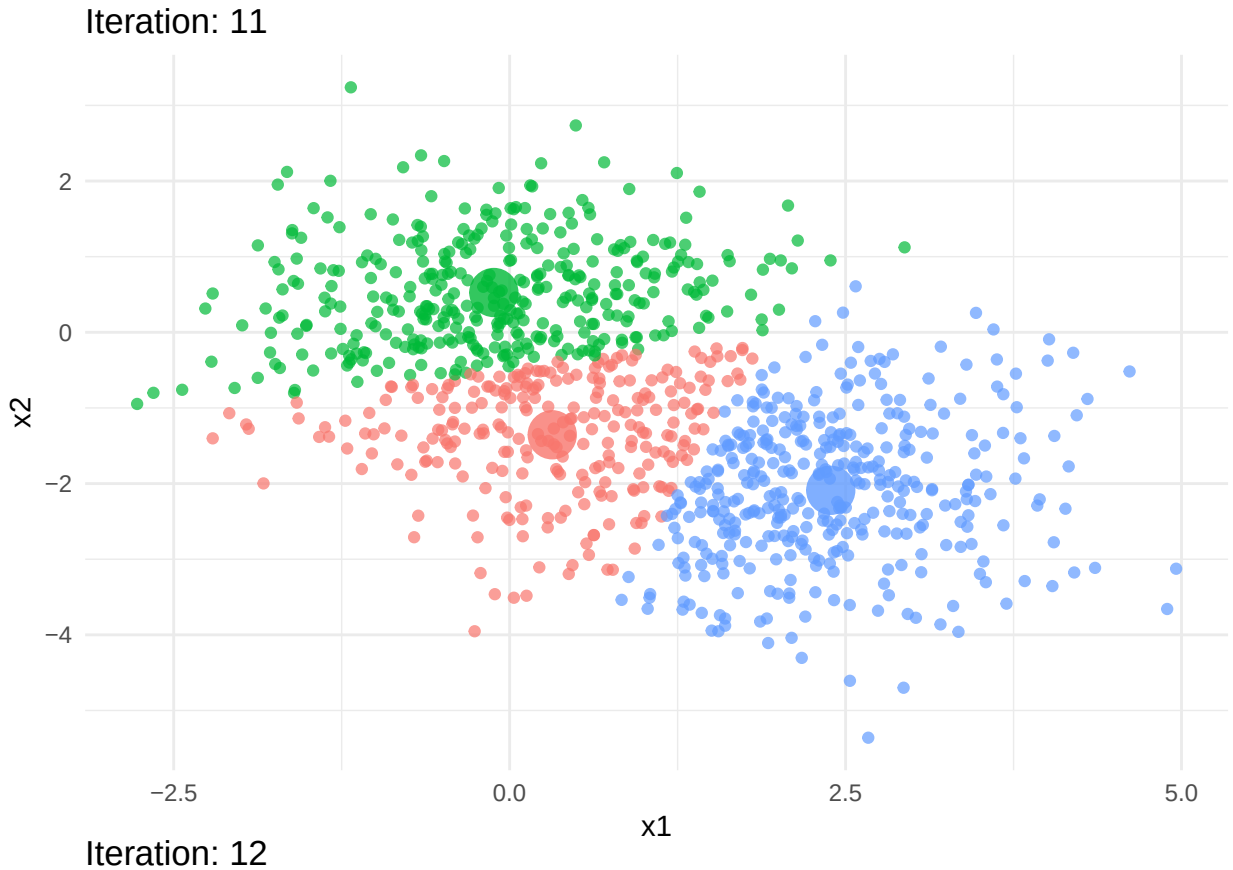

Iteration: 6
2
0
−2
−4
x2
−2.5
0.0
2.5
5.0
x1

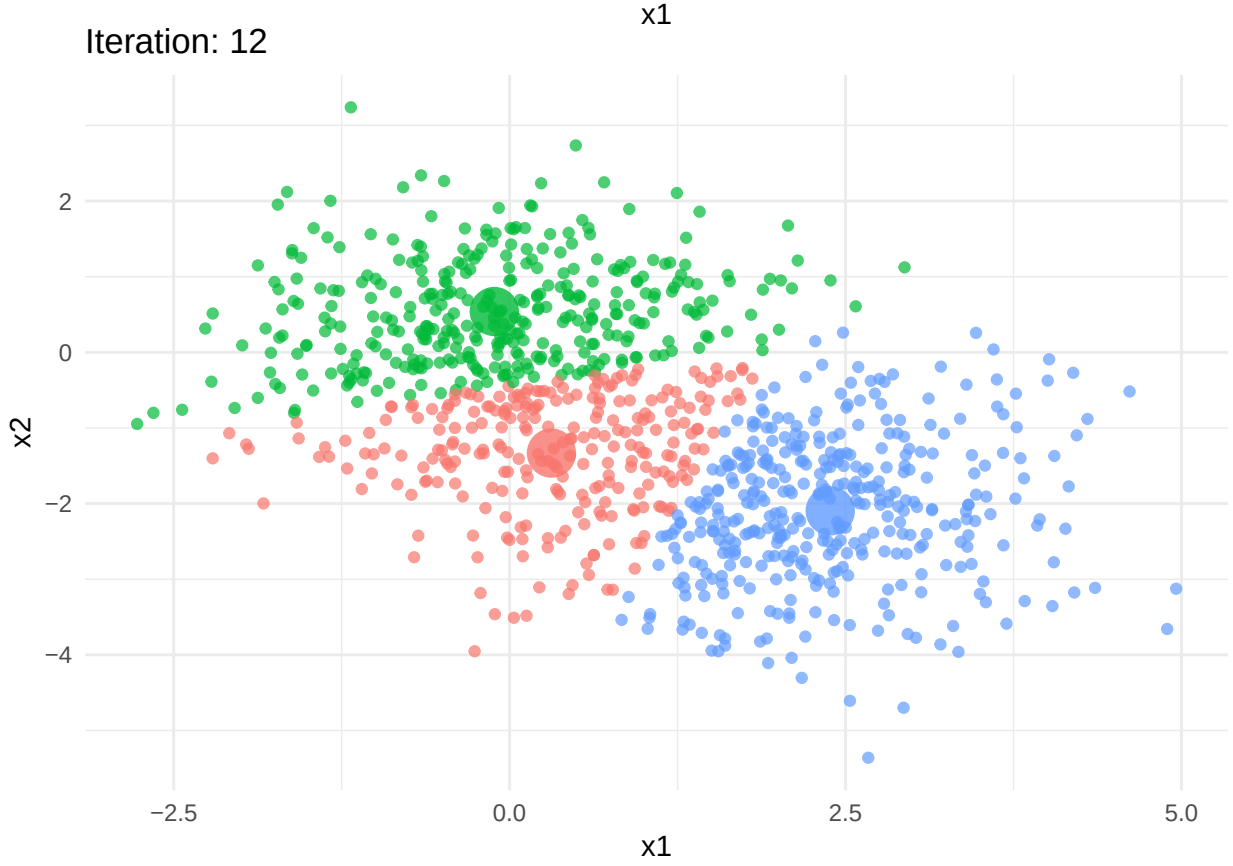

Iteration: 7
2
0
−2
−4
x2
−2.5
0.0
2.5
5.0
x1

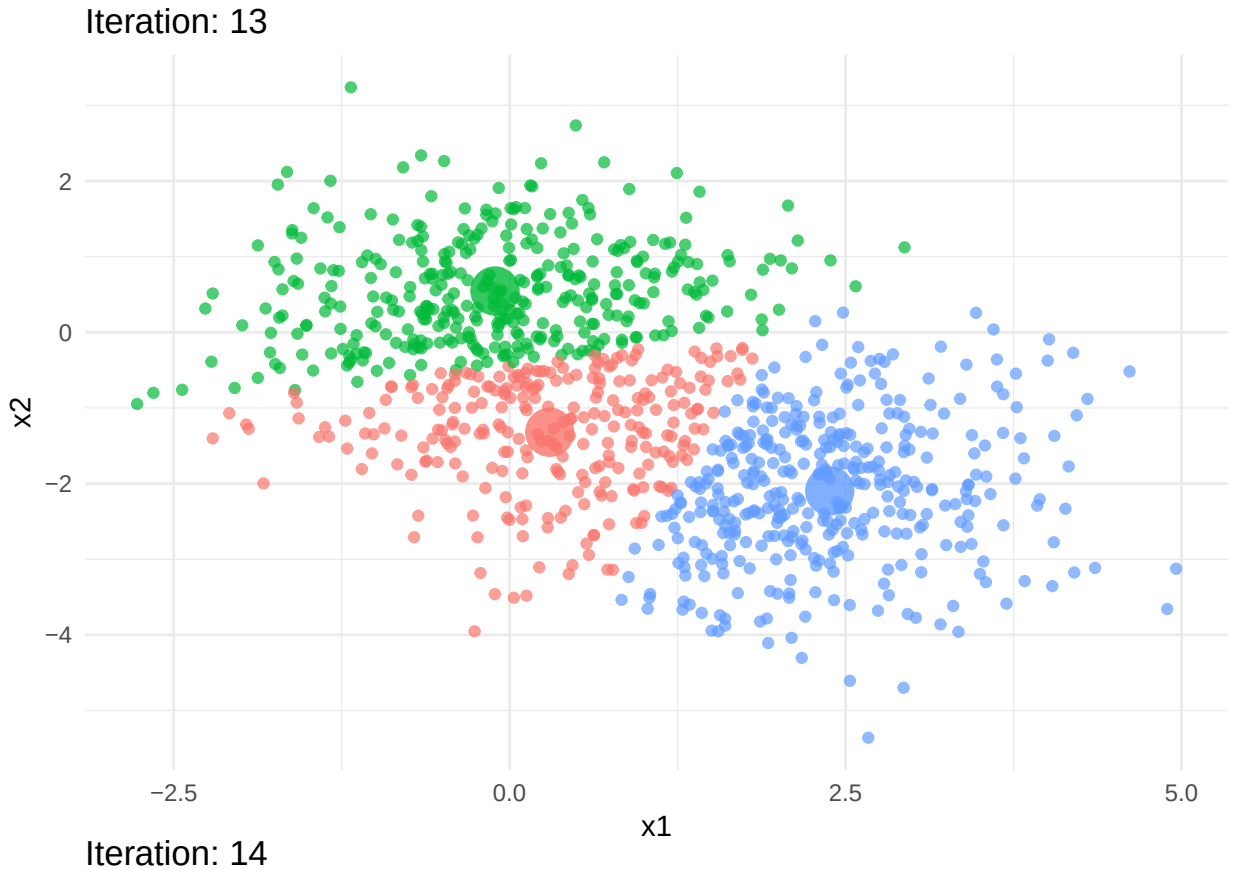

Iteration: 8
2
0
−2
−4
x2
−2.5
0.0
2.5
5.0
x1

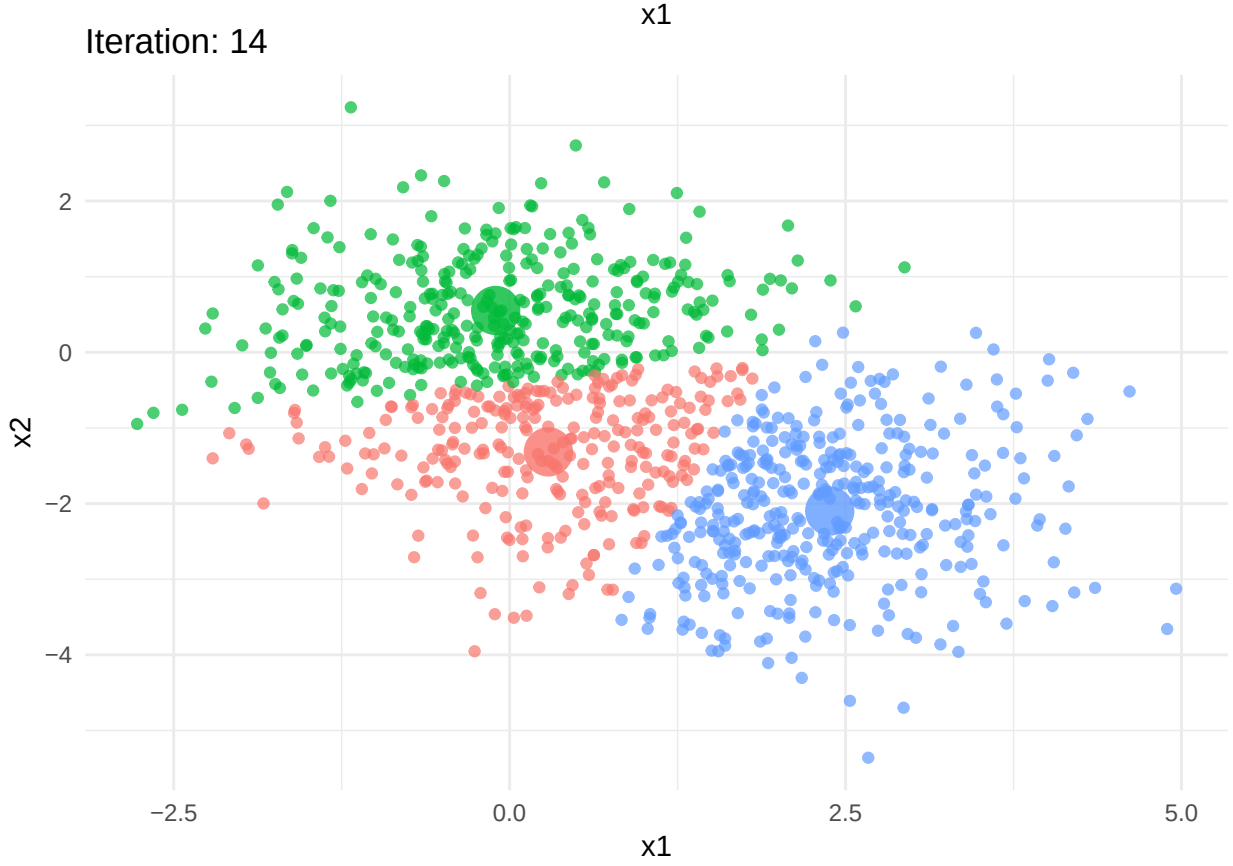

Iteration: 9
2
0
−2
−4
x2
−2.5
0.0
2.5
5.0
x1

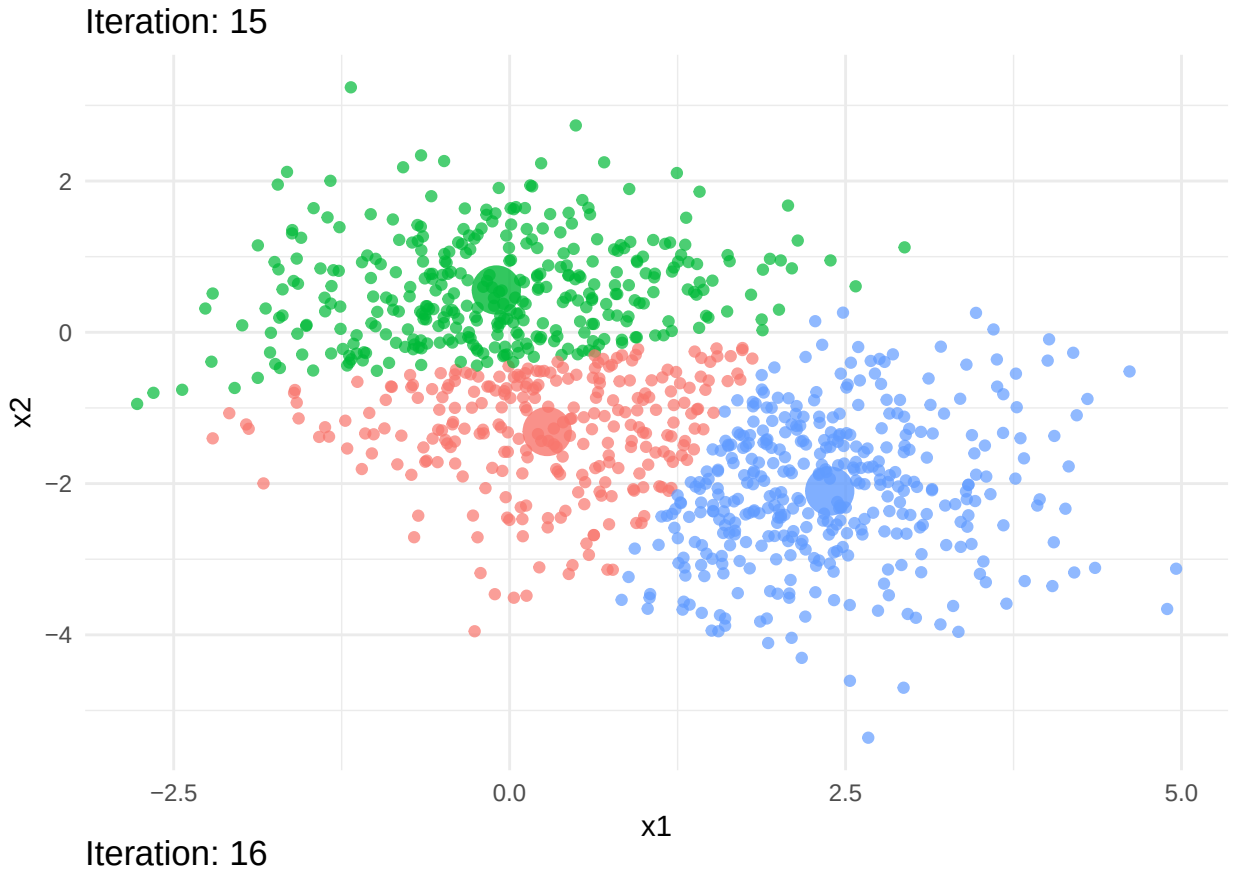

Iteration: 10
2
0
−2
−4
x2
−2.5
0.0
2.5
5.0
x1

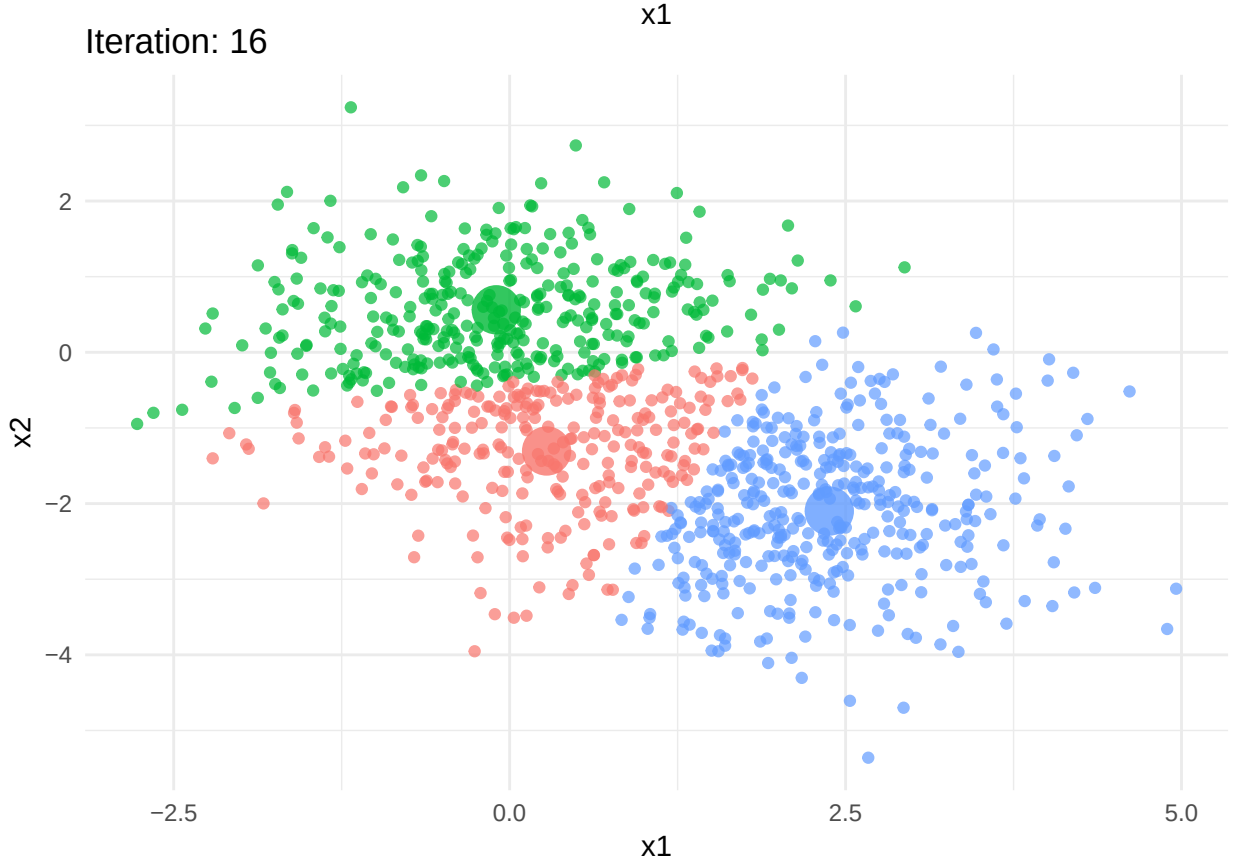

Iteration: 11
2
0
−2
−4
x2
−2.5
0.0
2.5
5.0
x1

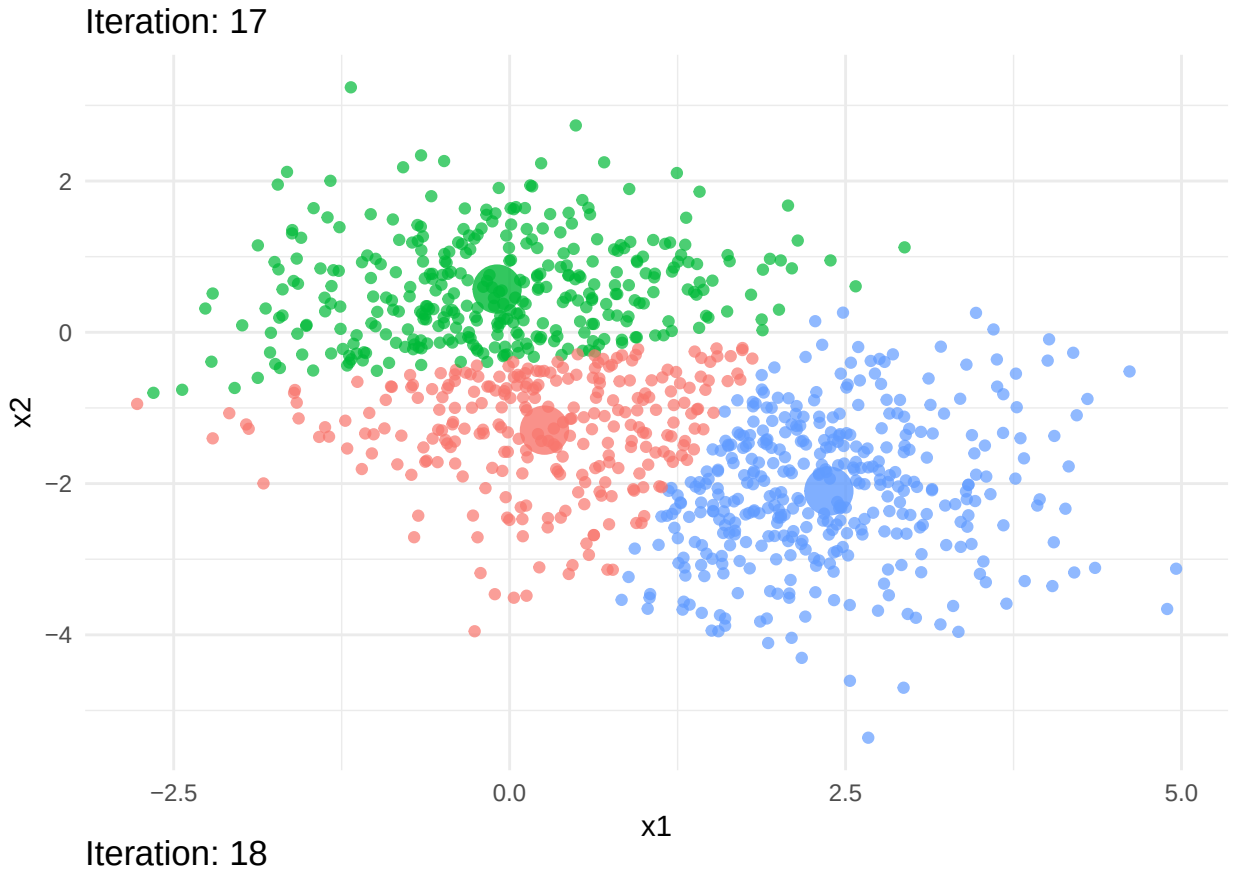

Iteration: 12
2
0
−2
−4
x2
−2.5
0.0
2.5
5.0
x1

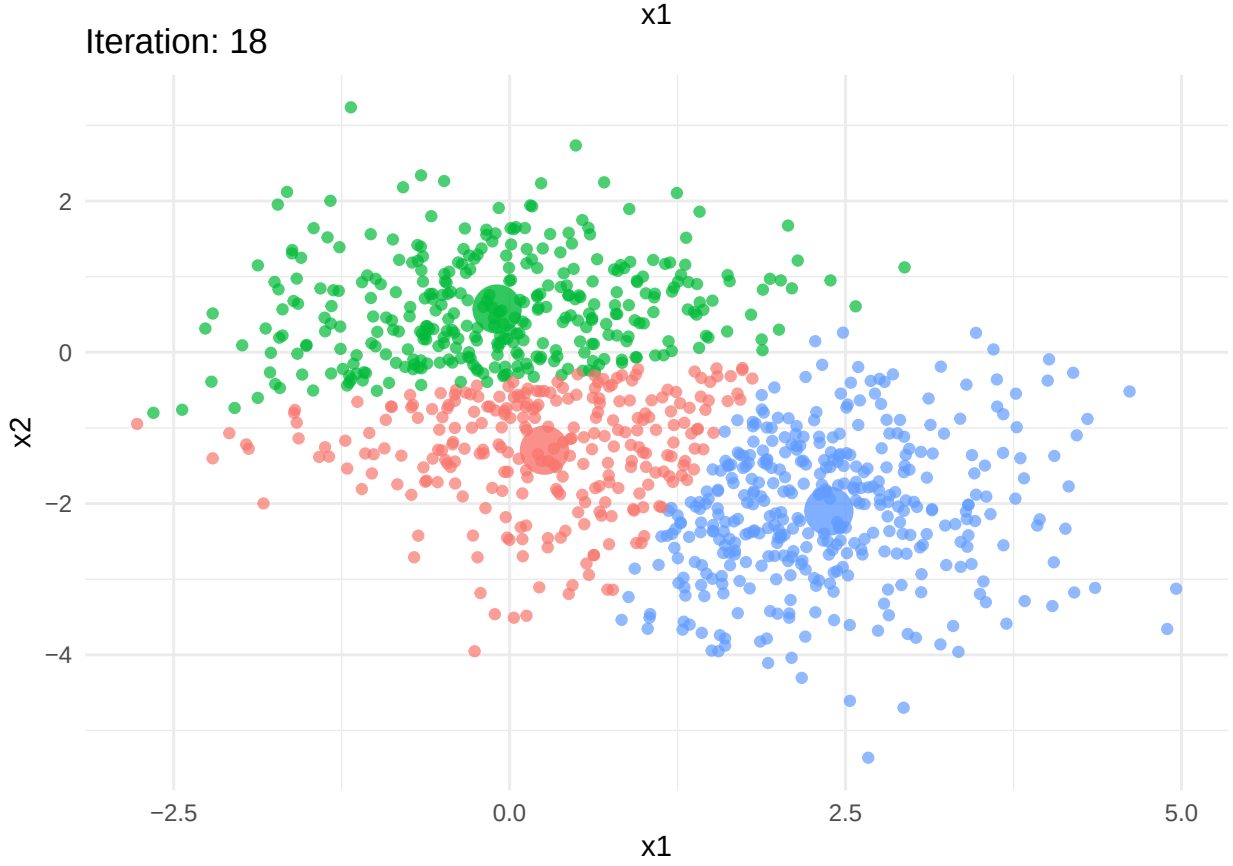

Iteration: 13
2
0
−2
−4
x2
−2.5
0.0
2.5
5.0
x1


Iteration: 14
2
0
−2
−4
x2
−2.5
0.0
2.5
5.0
x1


Iteration: 15
2
0
−2
−4
x2
−2.5
0.0
2.5
5.0
x1


Iteration: 16
2
0
−2
−4
x2
−2.5
0.0
2.5
5.0
x1


Iteration: 17
2
0
−2
−4
x2
−2.5
0.0
2.5
5.0
x1


Iteration: 18
2
0
−2
−4
x2
−2.5
0.0
2.5
5.0
x1

```
d <- d %>%
    crossing(centroids) %>%
    mutate(euc_dist = sqrt((x1_cent - x1)^2 + (x2_cent - x2)^2)) %>%
    group_by(x1,x2) %>%
    filter(euc_dist == min(euc_dist)) %>%
    select(x1, x2, assign)
```

Kmeans lab

```
d_mat <- d %>%
  select(-assign) %>%
  as.matrix()

centroids
```

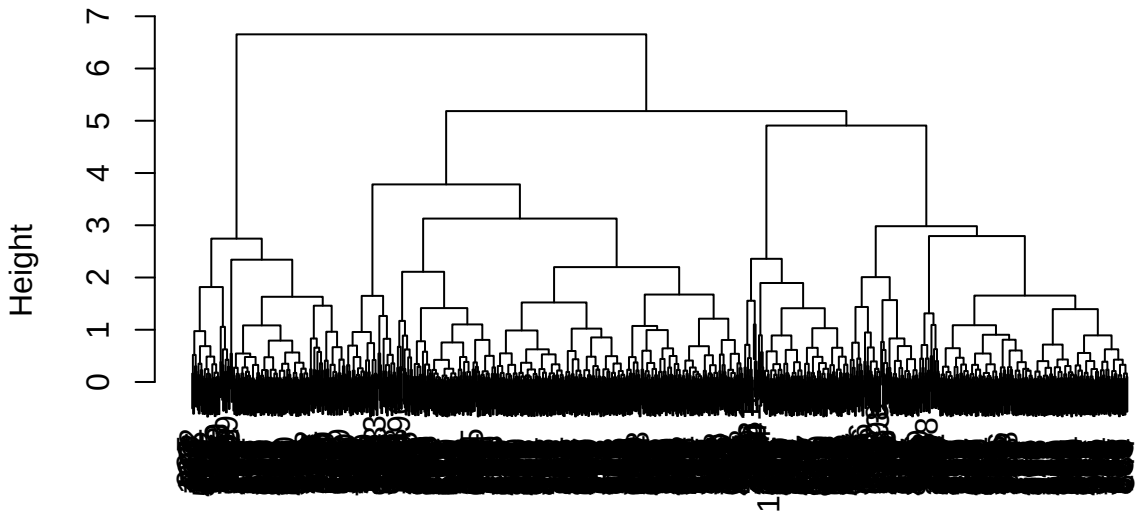
```
## # A tibble: 3 x 3
##   assign x1_cent x2_cent
##    <int>   <dbl>   <dbl>
## 1      1  0.259   -1.29
## 2      2 -0.0931   0.572
## 3      3  2.38    -2.10
```

```
kmeans(d_mat, centers = k)$centers
```

```
##           x1         x2
## 1  0.1052287  0.6929296
## 2 -0.1645216 -1.0569640
## 3  2.2499632 -2.0654722
```

Hierarchichcal Clustering

```
h_clust <- hclust(dist(scale(d_mat)), method = "complete")

h_clust %>%
  plot()
```




```
# method: average, single

cutree(h_clust, 2)
```

```
##   [1] 1 1 1 1 1 1 1 2 1 1 2 1 1 1 1 1 2 1 1 2 1 2 1 2 2 1 1 2 1 1 1 1 1 1
##  [35] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 2 1 1 1 1 2 2 1 1 1 1 1 2
##  [69] 1 1 1 2 1 1 1 1 1 1 1 1 1 2 1 1 1 1 2 1 1 1 1 1 1 2 1 1 1 1 1 1 1 2
## [103] 2 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 2 2 1 1
## [137] 1 1 1 1 1 1 1 1 2 1 2 1 1 1 1 1 2 1 1 2 2 1 1 1 1 1 2 1 1 1 1 1 1 1 1
## [171] 1 1 1 1 1 1 1 1 1 1 2 1 2 1 1 1 2 1 1 2 1 1 1 1 1 1 1 1 1 1 1 2 2 1 1
## [205] 1 2 1 1 1 1 1 1 1 1 2 2 1 2 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 2 1 1 1
## [239] 1 1 2 1 1 2 1 1 1 1 1 1 1 1 1 2 1 1 2 1 1 1 1 2 1 1 1 2 1 1 2 1 1 1
## [273] 1 1 1 1 1 1 1 1 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 1 1 1 1
## [307] 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 2 1 1 1 1 1 1 2 1 2 1
## [341] 1 2 2 1 1 2 1 1 1 2 1 1 1 1 1 1 1 1 1 1 2 1 2 1 1 1 1 1 1 1 1 2 1 2
## [375] 1 1 1 1 1 1 1 1 2 1 2 1 2 1 1 1 1 1 1 1 2 1 1 1 2 1 1 1 1 1 1 2 2 1 1
## [409] 1 1 1 1 1 2 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1
## [443] 1 1 2 2 1 1 1 1 1 2 1 1 1 2 1 1 1 1 1 2 1 1 2 1 2 1 1 2 1 1 1 1 2 1
## [477] 1 1 1 1 1 1 2 1 1 1 2 1 1 1 1 1 1 1 1 2 1 1 1 1 1 2 1 1 1 2 2 1 1 2 1
## [511] 1 1 1 1 1 1 1 1 1 1 1 2 1 1 2 1 1 1 1 1 1 2 1 1 2 2 1 1 1 2 1 1 1 2 1
## [545] 1 1 1 1 1 1 1 1 1 1 2 1 1 2 1 1 2 1 2 1 1 1 1 2 1 2 1 1 1 2 1 1 1 1 2
## [579] 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 2 2 1 1 1 1 1 1 2 1 1 1 1
## [613] 1 1 1 1 1 1 2 1 1 1 1 2 1 1 1 1 1 1 1 1 1 2 2 1 1 1 1 1 1 1 1 1 1 1 1
## [647] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 2 1 1 1 1 2
## [681] 1 2 1 1 1 1 2 1 1 2 1 1 1 2 2 1 1 1 1 1 1 1 2 1 1 2 1 2 1 1 1 1 1 2
## [715] 1 1 1 1 2 1 1 1 1 2 1 1 2 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [749] 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 2 2 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1
## [783] 1 1 1 1 1 1 1 2 1 1 1 1 2 2 2 1 1 1 1 2 1 1 1 2 1 1 1 1 1 1 1 1 1 2 1
```

```
##  [817] 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 2 1 1 1
##  [851] 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 2 1 1 1 2 2 1 2 1 1 1 1 1 2 1 1 2 1 2
##  [885] 1 1 1 1 1 1 1 2 1 1 1 1 1 2 2 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 2 1 1 1
##  [919] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1
##  [953] 1 2 1 1 1 1 1 1 1 1 1 2 1 2 1 1 1 1 2 1 1 1 1 2 1 1 1 2 1 1 1 1 1 1
##  [987] 1 1 1 1 2 1 1 1 1 1 1 1 1 1
```