TD3

Colin Coërchon

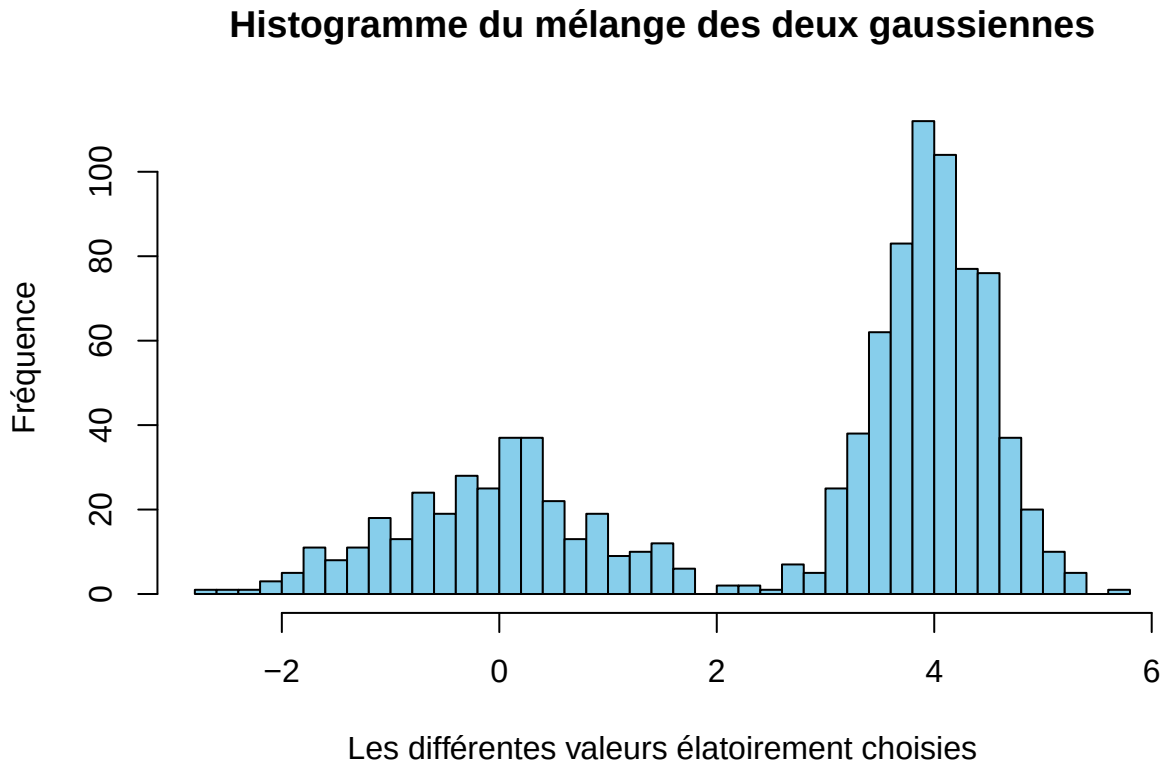
2023-11-06

Contents

Exercice 1

Question 1

```r
mu1 <- 0
mu2 <- 4
sigma1 <- 1
sigma2 <- 1/2
pi1 <- 1/3
pi2 <- 2/3 # 1 - pi1

n <- 1000

labels <- sample(c(1, 2), size = n, replace = TRUE, prob = c(pi1, pi2))

# Si label = 1, on simule un x suivant la première loi normale, sinon il suit la deuxième loi normale.
melange <- ifelse(labels == 1,
                          rnorm(n, mean = mu1, sd = sigma1),
                          rnorm(n, mean = mu2, sd = sigma2))

hist(melange, breaks = 50, col = 'skyblue', border = 'black',
     main = 'Histogramme du mélange des deux gaussiennes',
     xlab = 'Les différentes valeurs élatoirement choisies', ylab = 'Fréquence')
```



Question 2

```
res.km <- kmeans(melange, 2, nstart = 10)
clusters <- res.km$cluster

clusters
```

```
##    [1] 1 1 1 1 2 2 1 1 2 1 2 2 2 1 1 1 2 1 1 2 1 1 1 2 1 1 1 1 1 2 1 1 1 1 2 2 2
##   [38] 1 1 2 2 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 2 1 1 1 1 1 2 2 1 2 1 1 1 1 1 1 1 1
##   [75] 1 2 2 2 1 1 1 2 1 2 1 2 1 1 1 1 2 1 2 1 1 1 1 1 2 1 2 2 1 1 1 1 1 1 2 1 1
##  [112] 1 2 1 1 1 1 2 2 1 1 2 1 1 2 2 1 1 1 2 2 2 1 1 2 1 2 2 1 1 1 1 1 1 1 1 1 1
##  [149] 1 1 1 1 1 2 1 1 1 2 2 1 1 1 2 1 1 1 1 2 2 2 1 2 1 1 2 1 1 1 2 1 1 2 2 1 2
##  [186] 1 1 1 2 1 1 1 2 1 1 1 1 2 1 1 1 2 2 2 1 1 1 2 1 1 1 2 2 1 1 1 2 1 1 1 1 1
##  [223] 2 1 2 1 1 1 2 1 2 1 1 2 2 1 2 1 1 2 1 1 1 2 1 2 1 2 2 2 2 2 2 2 1 2 2 2 2
##  [260] 1 2 1 1 2 1 1 1 1 1 1 2 1 2 1 1 2 1 1 1 1 1 1 2 1 2 1 1 1 1 2 2 2 1 2 1 1
##  [297] 1 1 1 2 2 1 1 2 2 1 1 1 1 2 1 1 2 1 1 1 1 1 2 1 2 1 1 2 2 1 1 1 1 1 1 1 1
##  [334] 1 1 2 2 1 1 1 2 2 1 2 1 1 2 1 1 1 1 2 1 1 1 1 2 1 1 1 2 1 1 1 1 1 2 2 2 1
##  [371] 2 1 1 1 1 1 1 2 1 2 2 1 1 2 1 1 1 1 1 1 1 1 1 1 2 2 1 1 1 1 2 1 1 1 1 1 1 2
##  [408] 1 2 1 1 1 1 2 1 1 2 1 2 2 2 2 1 2 1 1 1 2 1 1 2 1 1 1 1 1 1 2 2 1 2 1 2 1
##  [445] 1 1 2 1 1 2 2 2 1 1 1 1 2 2 2 2 2 1 1 1 1 1 1 2 1 1 1 1 1 1 2 1 1 1 2 2 2
##  [482] 1 1 2 1 1 1 1 1 1 2 1 2 1 1 1 1 1 1 1 1 1 1 1 2 2 2 1 2 2 2 1 1 1 2 2 1 2
##  [519] 1 2 2 2 2 2 2 2 1 1 1 1 2 1 1 2 2 2 1 2 1 1 2 1 2 2 1 2 1 1 2 2 1 1 2 1 1
##  [556] 2 1 2 1 1 1 1 1 1 1 1 1 1 1 1 2 1 2 1 2 1 1 1 2 1 1 1 2 1 2 2 1 2 1 1 1 1
##  [593] 1 1 1 1 2 1 1 1 2 2 2 2 2 1 1 1 2 2 1 2 2 1 1 2 1 1 1 2 2 2 2 2 1 1 1 2 1
##  [630] 1 2 1 1 2 1 1 1 1 1 1 2 1 1 1 1 1 1 2 2 1 1 1 1 1 1 1 1 2 1 1 1 2 1 1 2 1 1 1
##  [667] 1 1 1 1 1 2 1 1 1 1 1 1 2 2 1 2 1 1 1 2 1 1 2 1 2 1 2 2 1 1 2 1 2 1 1 1 2 1 1
##  [704] 1 1 1 1 1 1 1 1 1 2 1 1 2 1 2 2 1 1 1 2 1 2 1 1 2 1 1 2 2 2 1 1 1 1 2 2 1
##  [741] 1 2 2 1 1 2 2 1 1 1 2 2 1 1 1 1 1 1 2 1 1 2 1 2 2 2 2 2 1 1 2 1 1 1 1 1 2 2
##  [778] 2 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 2 1 1 1 1 1 1 2 2 1 1 2 1 2 1 1 1
##  [815] 2 2 2 2 1 2 2 1 1 1 1 1 1 2 1 1 2 2 1 1 1 1 2 1 1 2 2 1 1 1 2 2 1 1 2 2 1
##  [852] 2 1 1 2 2 1 2 1 1 1 1 1 1 1 1 2 1 1 2 1 1 1 1 2 2 1 2 1 1 2 2 1 2 1 2 1 1
##  [889] 1 1 1 1 1 2 2 1 2 1 2 1 2 2 2 2 2 1 1 2 1 1 1 1 2 1 1 1 1 1 1 1 1 2 1 1 1 1 1
##  [926] 1 1 1 1 2 2 1 2 1 1 1 1 1 1 1 1 1 2 1 1 1 2 2 1 1 1 1 2 1 1 1 1 1 1 1 1 2 2
##  [963] 1 1 2 1 1 1 2 2 1 1 1 2 1 1 1 1 2 1 1 1 1 1 1 1 1 2 1 1 1 2 2 2 1 2 2 1 2
## [1000] 2
```

Comparaison des résultats avec ceux attendus :

```
table("Pred"=clusters, "True label"=labels)
```

```
##     True label
## Pred   1   2
##    1   5 662
##    2 333   0
```

Question 3

```
# Calcul des paramètres pour le cluster 1
mu1_est <- mean(melange[clusters == 1])
sigma1_est <- sd(melange[clusters == 1])
```

```r
pi1_est <- sum(clusters == 1) / length(clusters)

# Calcul des paramètres pour le cluster 2
mu2_est <- mean(melange[clusters == 2])
sigma2_est <- sd(melange[clusters == 2])
pi2_est <- sum(clusters == 2) / length(clusters)

# Affichage des estimations
cat("Estimations pour le Cluster 1:\n")
```

```
## Estimations pour le Cluster 1:
```

```r
cat("Moyenne: ", mu1_est, "\n")
```

```
## Moyenne:  3.99053
```

```r
cat("Écart-type: ", sigma1_est, "\n")
```

```
## Écart-type:  0.5168195
```

```r
cat("Proportion: ", pi1_est, "\n\n")
```

```
## Proportion:  0.667
```

```r
cat("Estimations pour le Cluster 2:\n")
```

```
## Estimations pour le Cluster 2:
```

```r
cat("Moyenne: ", mu2_est, "\n")
```

```
## Moyenne:  -0.1009845
```

```r
cat("Écart-type: ", sigma2_est, "\n")
```

```
## Écart-type:  0.8962573
```

```r
cat("Proportion: ", pi2_est, "\n\n")
```

```
## Proportion:  0.333
```

```r
# Vraies valeurs :
cat("Vraies valeurs pour le Cluster 1 et 2 :\n")
```

```
## Vraies valeurs pour le Cluster 1 et 2 :
```

```r
cat("Moyennes: ",0,"et",4, "\n")
```

```
## Moyennes:  0 et 4
```

```r
cat("Écart-type: ",1,"et",1/2, "\n")
```

```
## Écart-type:  1 et 0.5
```

```r
cat("Proportion: ",1/3,"et",2/3, "\n")
```

```
## Proportion:  0.3333333 et 0.6666667
```

Question 4

```r
mclust_E <- Mclust(data = melange, G = 2, modelNames = 'E')

cat("Mclust pour la méthode 'E' :\n")
```

```
## Mclust pour la méthode 'E' :
```

```r
cat("Moyennes: ", mclust_E$parameters$mean, "\n")
```

```
## Moyennes:  -0.1083719 3.98515
```

```r
cat("Variances: ", mclust_E$parameters$variance$sigmasq, "\n")
```

```
## Variances:  0.4492343
```

```r
cat("Proportion: ", mclust_E$parameters$pro, "\n\n")
```

```
## Proportion:  0.3315225 0.6684775
```

Attention, avec cette méthode, il y a logiquement qu'une seule variance (cf. cours), on est en méthode "Equal" donc il considère qu'il y a une unique variance pour chacun des groupes.

```r
mclust_V <- Mclust(data = melange, G = 2, modelNames = 'V')

cat("Mclust pour la méthode 'E' :\n")
```

```
## Mclust pour la méthode 'E' :
```

```r
cat("Moyennes: ", mclust_V$parameters$mean, "\n")
```

```
## Moyennes:  -0.06832514 4.001868
```

```r
cat("Variances: ", mclust_V$parameters$variance$sigmasq, "\n")
```

```
## Variances:  0.8697248 0.248985
```

```r
cat("Proportion: ", mclust_V$parameters$pro, "\n\n")
```

```
## Proportion:  0.3375301 0.6624699
```

Ici, en méthode “variable”, on a bien 2 variances pour nos groupes voulus.

Exercice 2

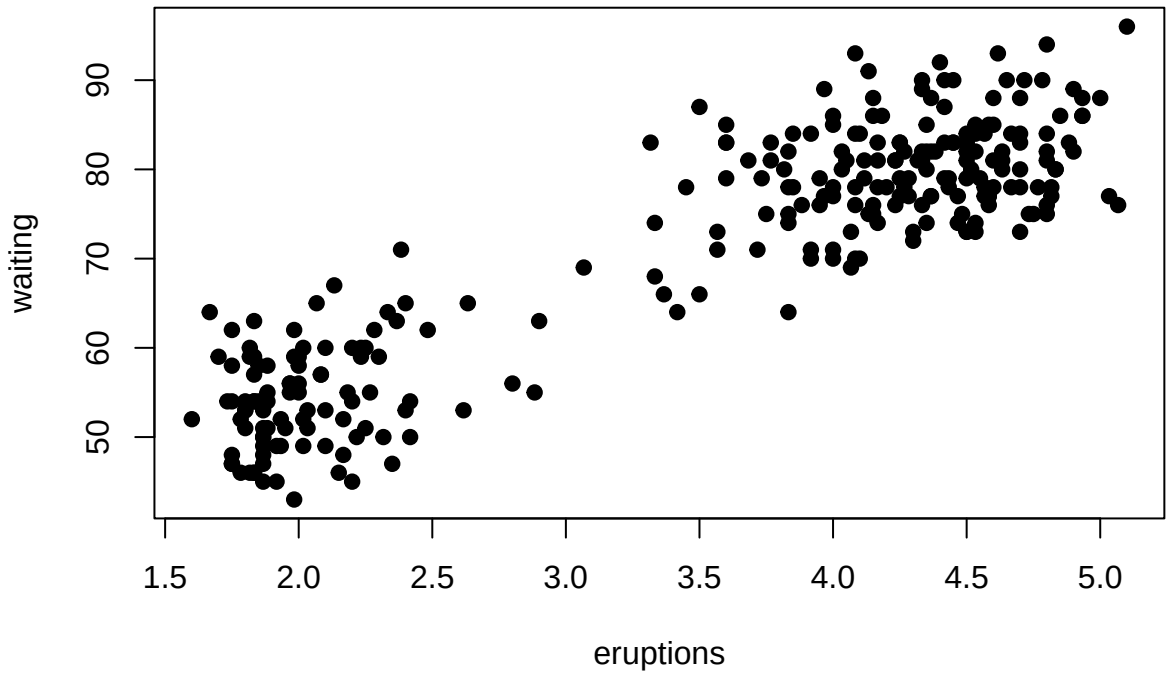
Question 1

```
data(faithful)
head(faithful)
```

```
##   eruptions waiting
## 1     3.600      79
## 2     1.800      54
## 3     3.333      74
## 4     2.283      62
## 5     4.533      85
## 6     2.883      55
```

Question 2

```
plot(faithful, pch = 19)
```




```
mclust_surprise <- Mclust(data = faithful)

cat("Mclust :\n")
```

```
## Mclust :
```

```
cat("Nombre de clusters :",mclust_surprise$G, "\n")
```

```
## Nombre de clusters : 3
```

```
cat("Moyennes: ", mclust_surprise$parameters$mean, "\n")
```

```
## Moyennes:  3.793066 77.52105 2.037596 54.49116 4.463245 80.83344
```

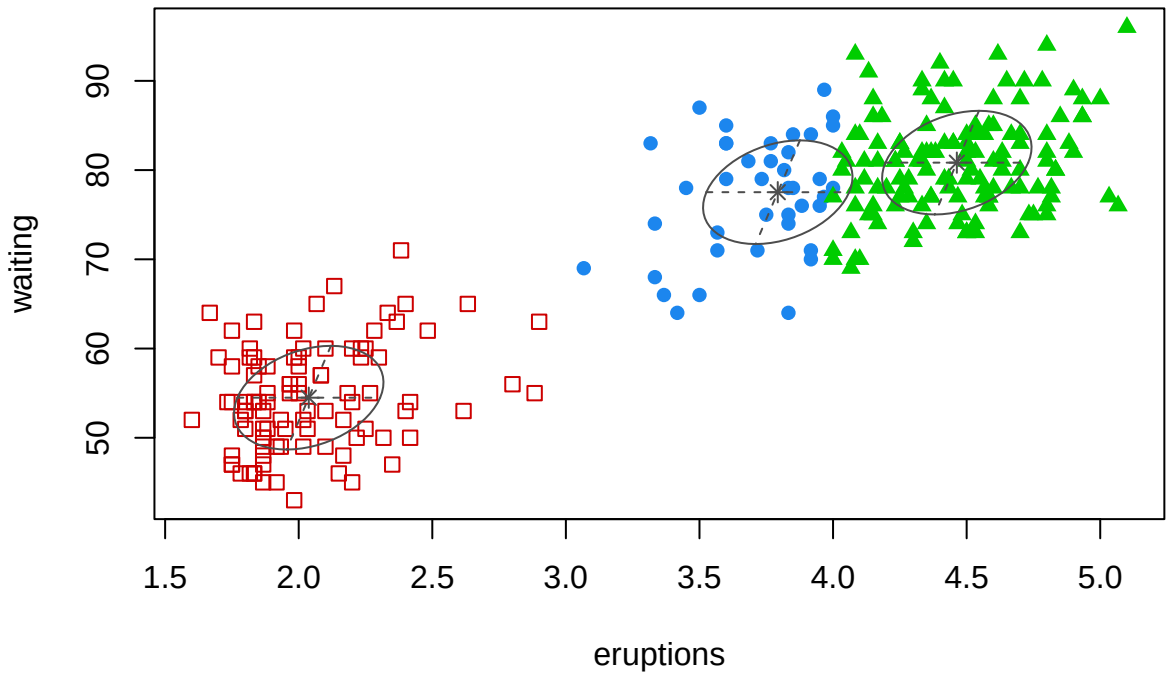
```
cat("Variances: ", mclust_surprise$parameters$variance$sigmasq, "\n")
```

```
## Variances:
```

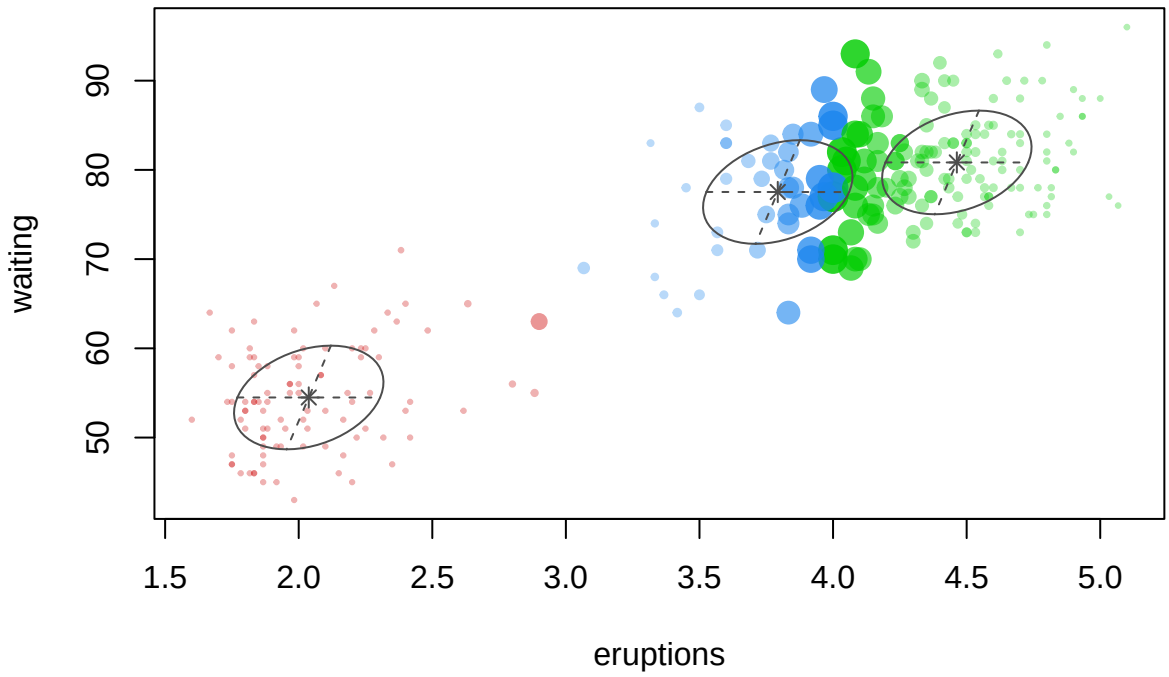
```
cat("Proportion: ", mclust_surprise$parameters$pro, "\n\n")
```

```
## Proportion:  0.1656784 0.3563696 0.477952
```

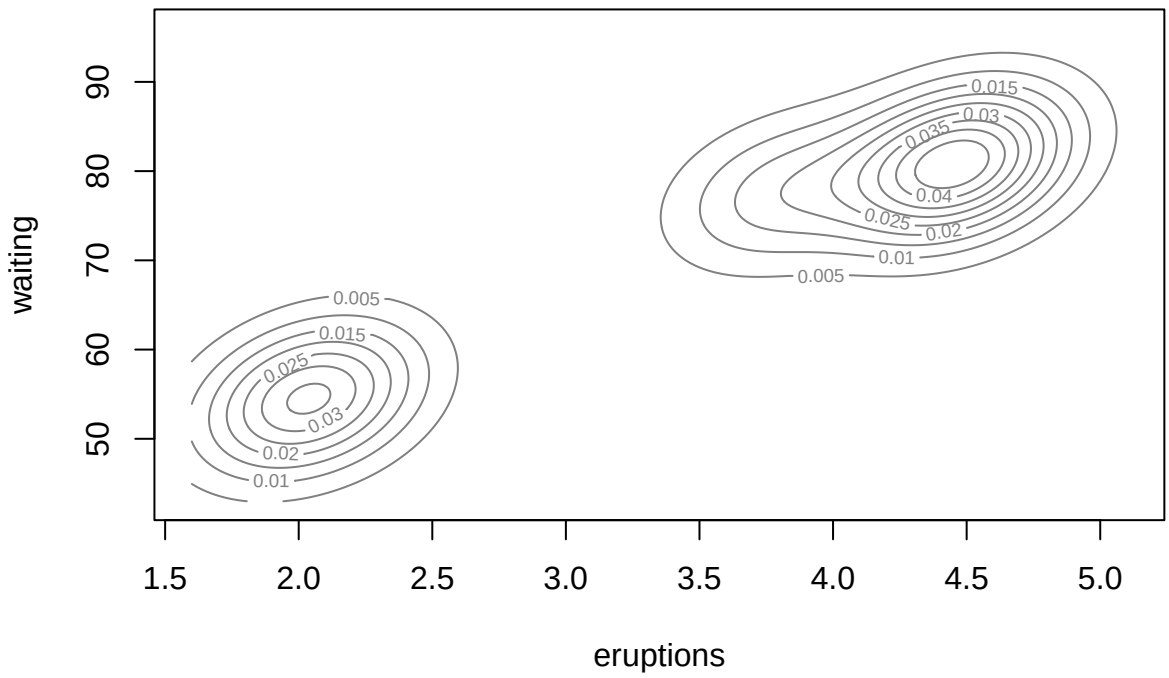
```
plot(mclust_surprise, what = "classification")
```



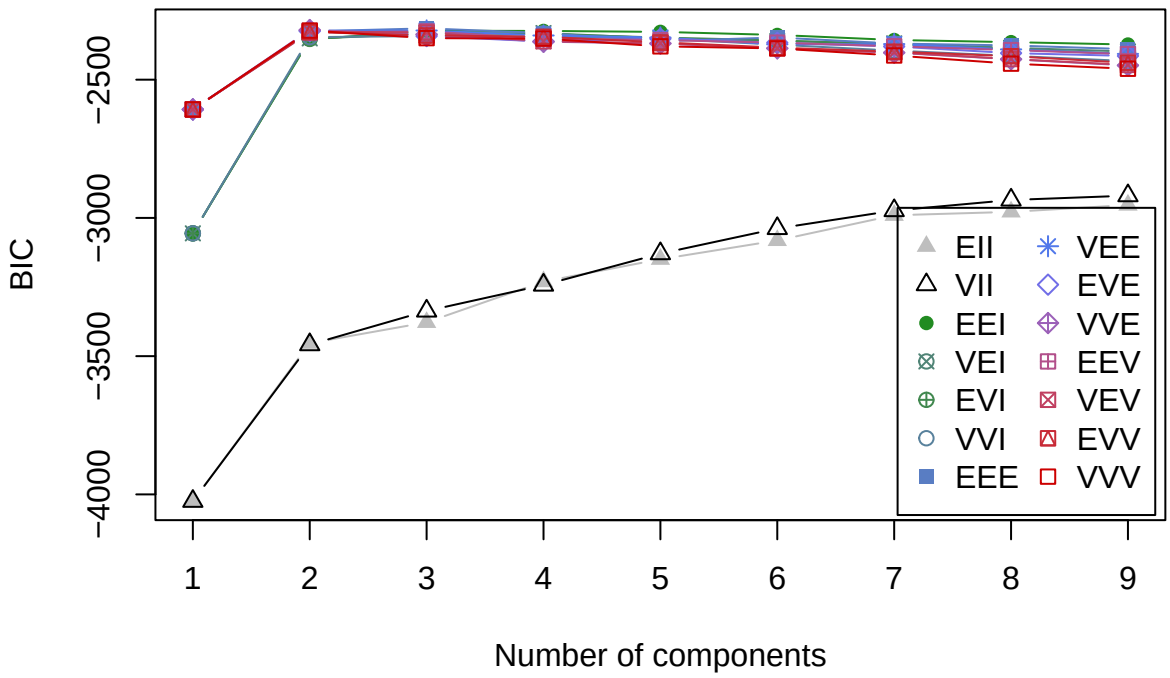
```
plot(mclust_surprise, what = "uncertainty")
```



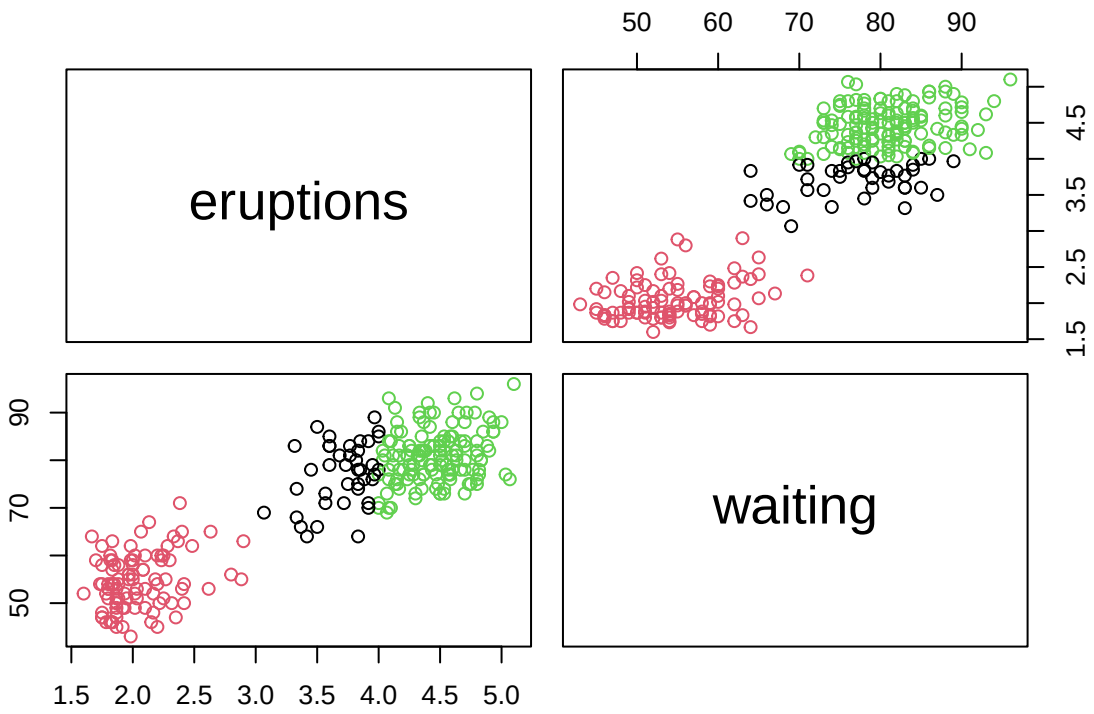

```
plot(mclust_surprise, what = "density")
```




```
plot(mclust_surprise, what = "BIC")
```




```
pairs(faithful, col = mclust_surprise$classification)
```

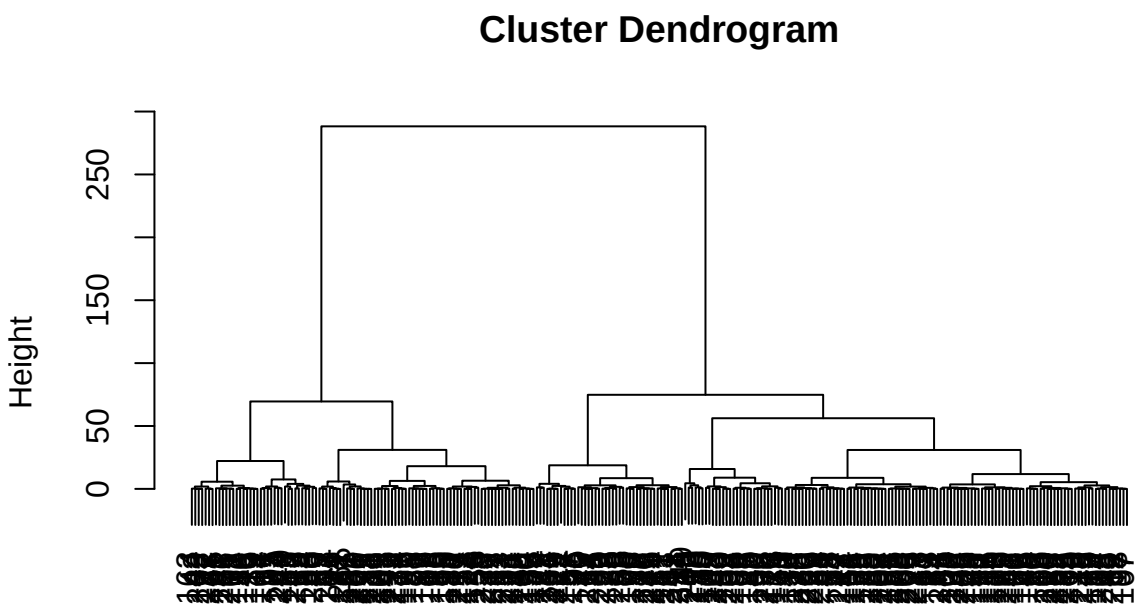



```
# plot(silhouette(mclust_surprise))
```

Question 5

```
# Calculer les distances
distances <- dist(faithful)

# Appliquer hclust avec le critère de Ward
hc <- hclust(distances, method = "ward.D2")
```

```
plot(hc)
```




```r
# Couper l'arbre pour obtenir deux clusters avec hclust
clusters_hc_2 <- cutree(hc, k = 2)

# Clusters avec Mclust pour 2 clusters
mclust_2 <- Mclust(data = faithful, G=2)
clusters_mc_2 <- mclust_2$classification

table(clusters_mc_2, clusters_hc_2)
```

```
##              clusters_hc_2
## clusters_mc_2   1   2
##             1 171   4
##             2   1  96
```

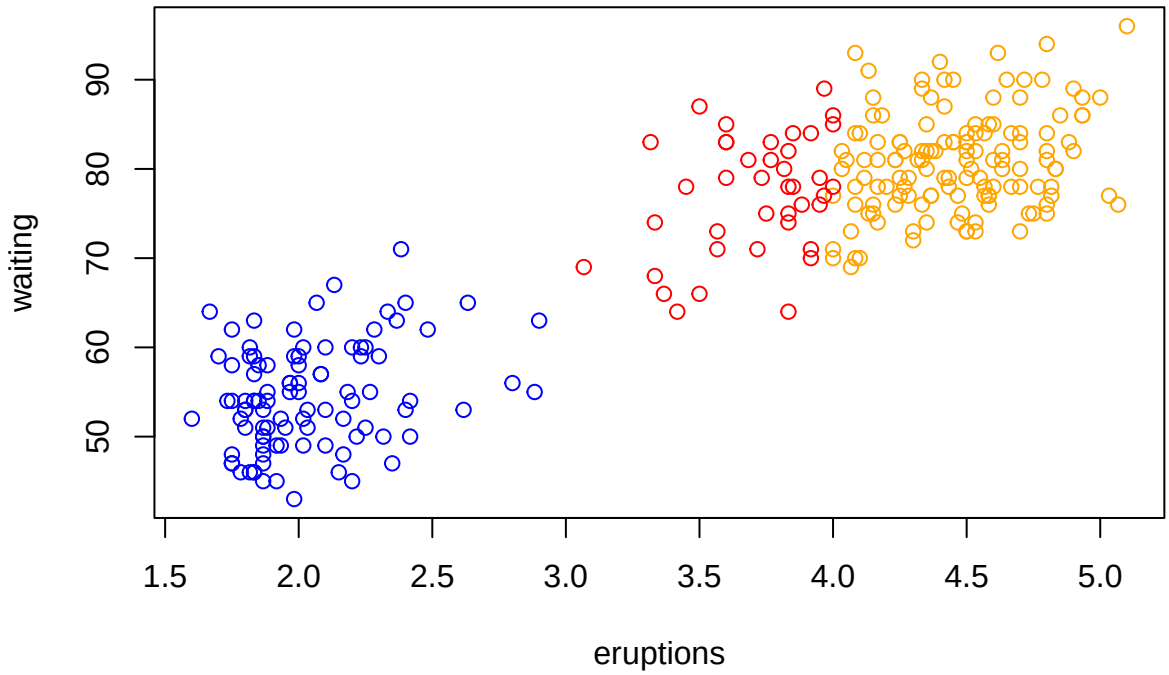
Question 6

```r
# Couper l'arbre pour obtenir trois clusters avec hclust
clusters_hc_3 <- cutree(hc, k = 3)

# Clusters avec Mclust pour trois clusters
mclust_3 <- Mclust(data = faithful, G=3)
clusters_mc_3 <- mclust_3$classification

table(clusters_mc_3, clusters_hc_3)
```

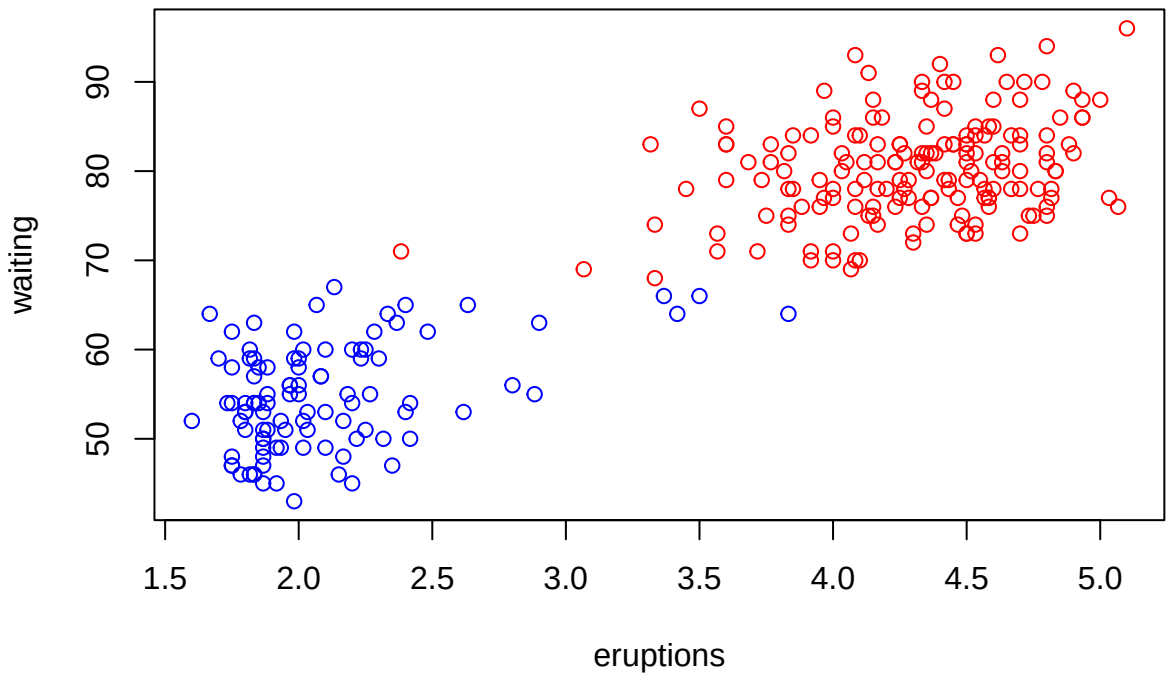
```
##               clusters_hc_3
## clusters_mc_3   1   2   3
##             1  23   4  13
##             2   0  96   1
##             3 106   0  29
```

Et en terme d'affichage :

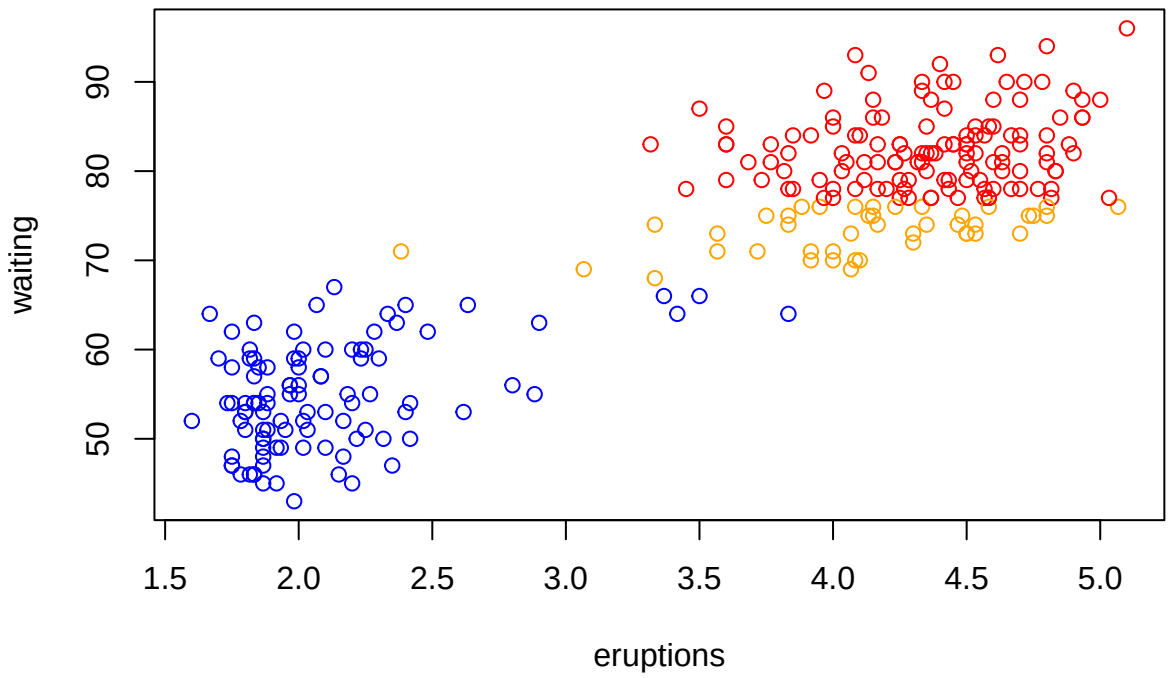
```
plot(faithful, col = c("Red","Blue","Orange")[mclust_surprise$classification])
```




```
plot(faithful, col = c("Red", "Blue")[clusters_hc_2])
```




```
plot(faithful, col = c("Red", "Blue","Orange")[clusters_hc_3])
```


waiting
50
60
70
80
90
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
eruptions