Selvstudie 2

Nicholas Fitzhugh

22/2/2017

CH index

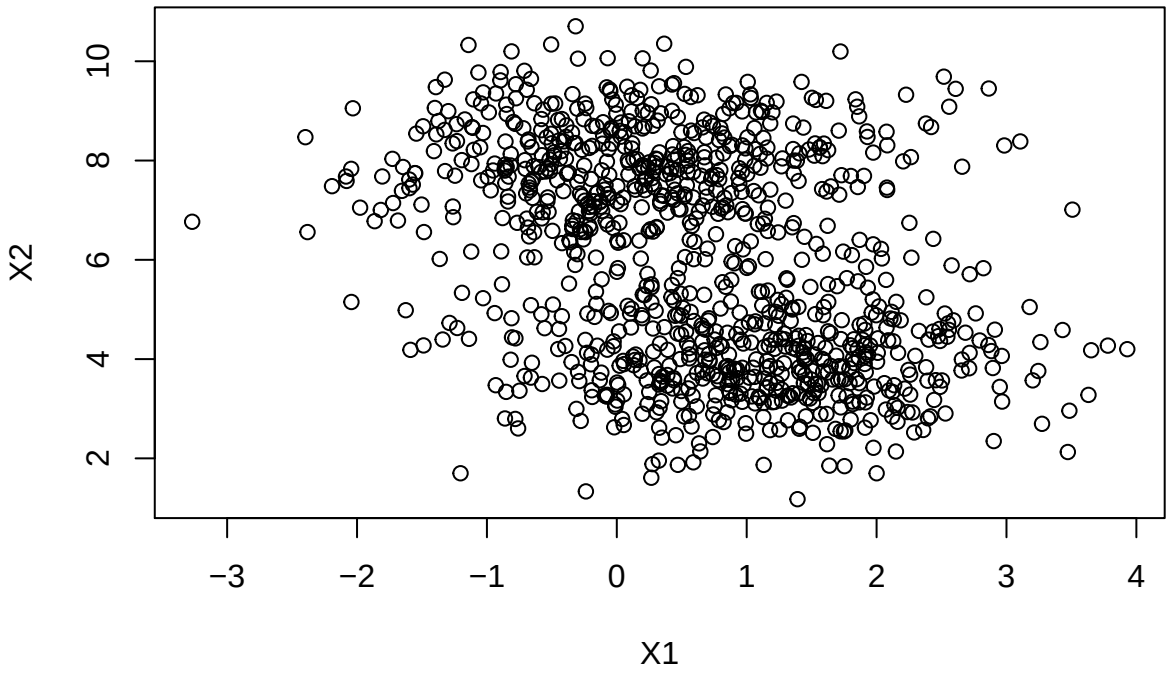
Funktionen er defineret nedenfor, og tager et kmeans objekt som input.

```
ch_index <- function(k){
  (k$betweenss / (length(k$size) - 1)) / (sum(k$withinss) / (length(k$cluster) - length(k$size)))
}
```

Simuler data

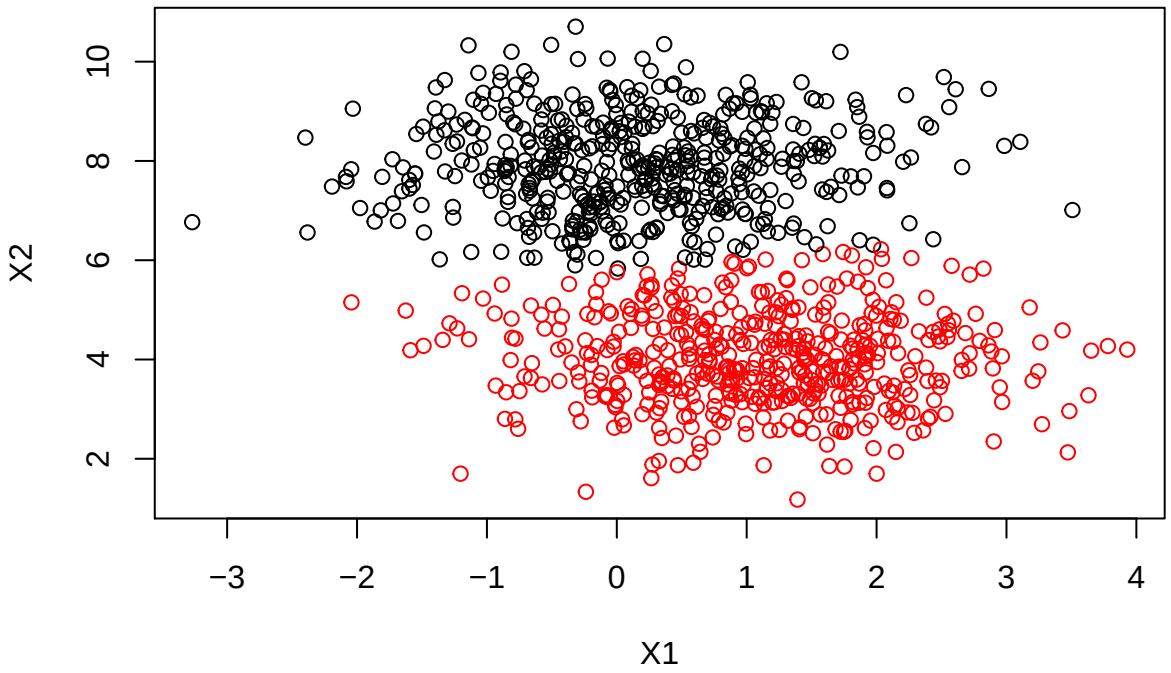
Data generation:

```
mu1 <- c(1,4)
mu2 <- c(0,8)
sigma <- matrix(ncol = 2, nrow = 2, c(1,0,0,1))
m1 <- mvrnorm(n=500, mu1, Sigma = sigma)
m2 <- mvrnorm(n=500, mu2, Sigma = sigma)
dat.data <- data.frame(rbind(m1,m2))
plot(dat.data)
```



K - Means:

```
dat.kmeans <- kmeans(dat.data, centers = 2, nstart = 10)
plot(dat.data, col = dat.kmeans$cluster)
```

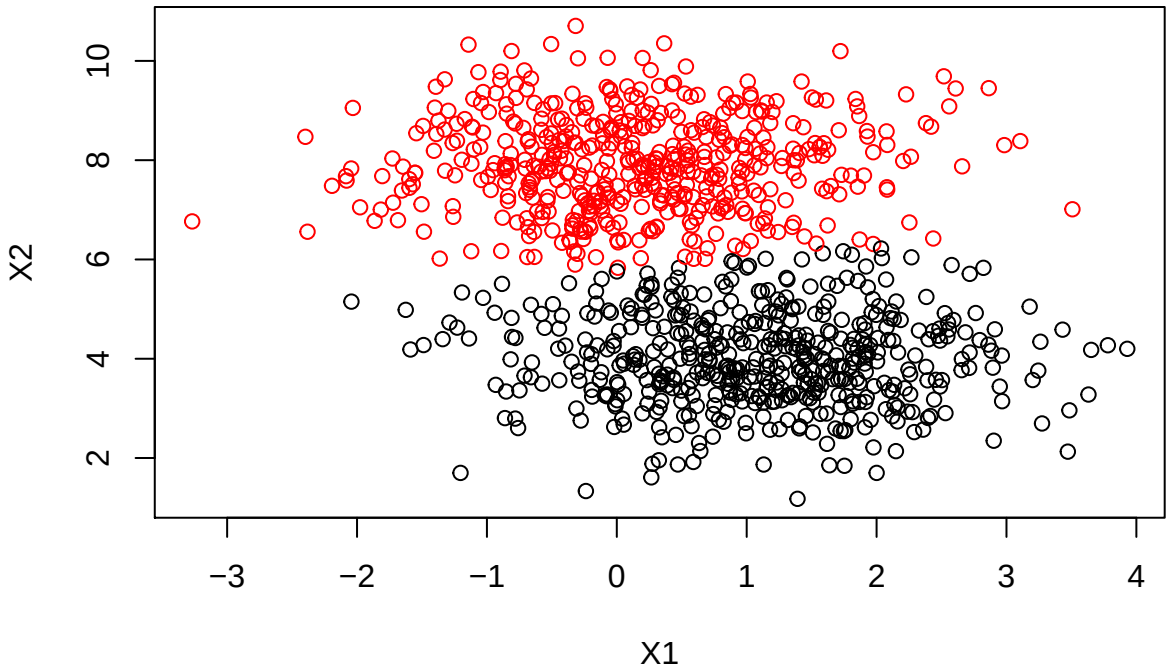


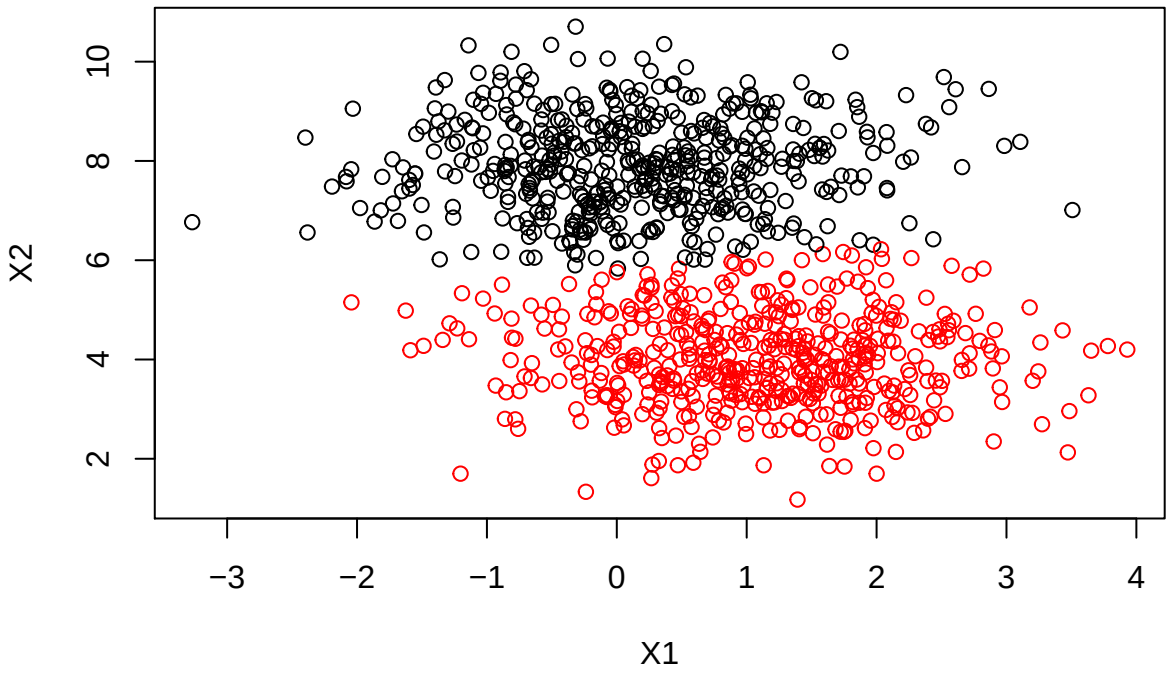

```
ch_index(dat.kmeans)
```

```
## [1] 2127.794
```

K-mediods:

```
dat.pam <- pam(dat.data, 2)
plot(dat.data, col = dat.pam$cluster); plot(dat.data, col = dat.kmeans$cluster)
```





Hclust:

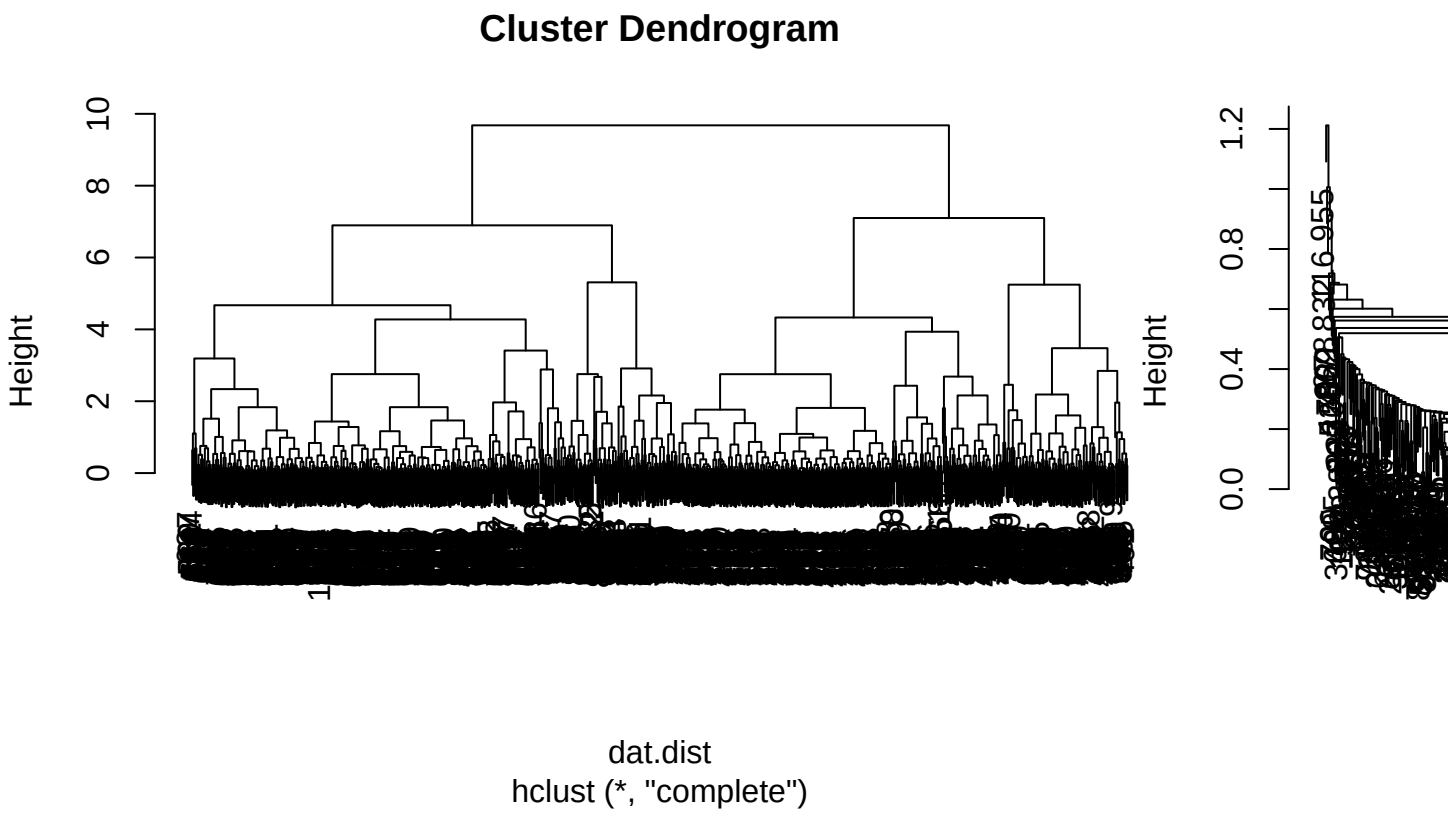
```
dat.dist <- dist(dat.data)
(dat.hclust1 <- hclust(dat.dist))
```

```
##
## Call:
## hclust(d = dat.dist)
##
## Cluster method   : complete
## Distance         : euclidean
## Number of objects: 1000
```

```
(dat.hclust2 <- hclust(dat.dist, method = "single"))
```

```
##
## Call:
## hclust(d = dat.dist, method = "single")
##
## Cluster method   : single
## Distance         : euclidean
## Number of objects: 1000
```

```
plot(dat.hclust1); plot(dat.hclust2)
```

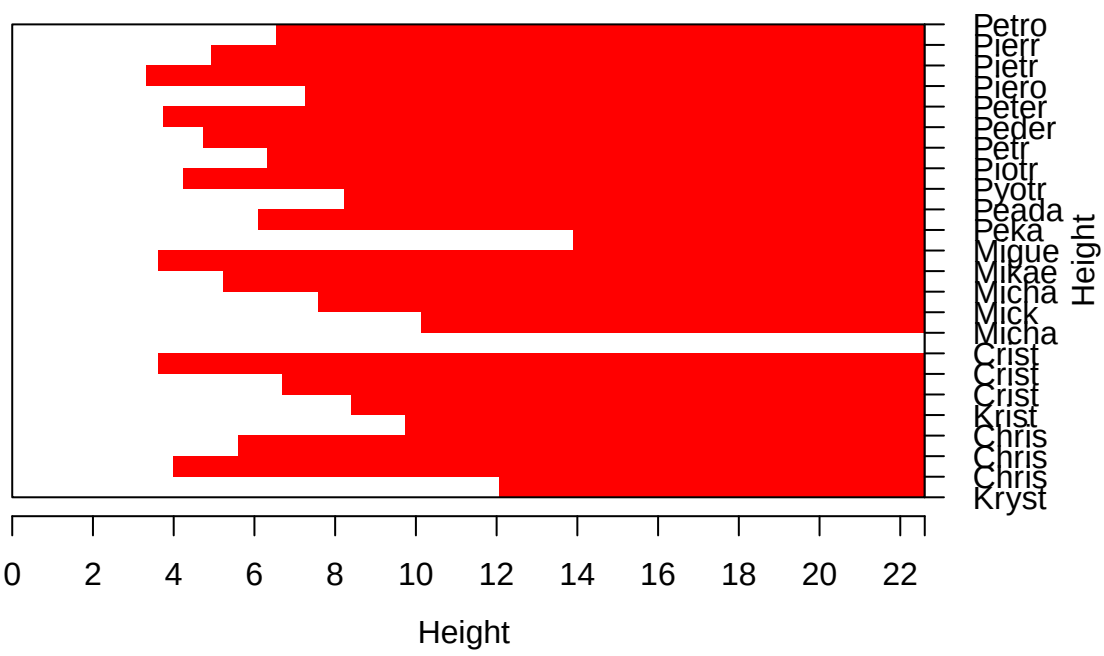


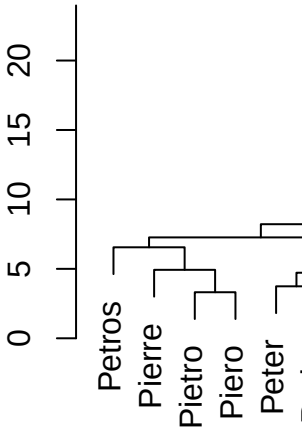
Diana og agnes giver morderiske plot.

drengenavn.csv

Agnes metoden:

```
navne.dist <- adist(drengenavne[,1])
rownames(navne.dist) <- drengenavne$name
navne.agnes <- agnes(navne.dist)
plot(navne.agnes)
```





Agglomerative Coefficient = 0.76

Hclust:

```
#navne.hclust <- hclust(navne.dist)
```

Bruger cutree til at opdele i 3 klynger:

```
cutree(navne.agnes, k=3)
```

```
##  [1] 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 3 3 3 3 3
```

Denne gruppering er den forventede.

```
navne.pam <- pam(navne.dist, 3)
navne.pam$clustering
```

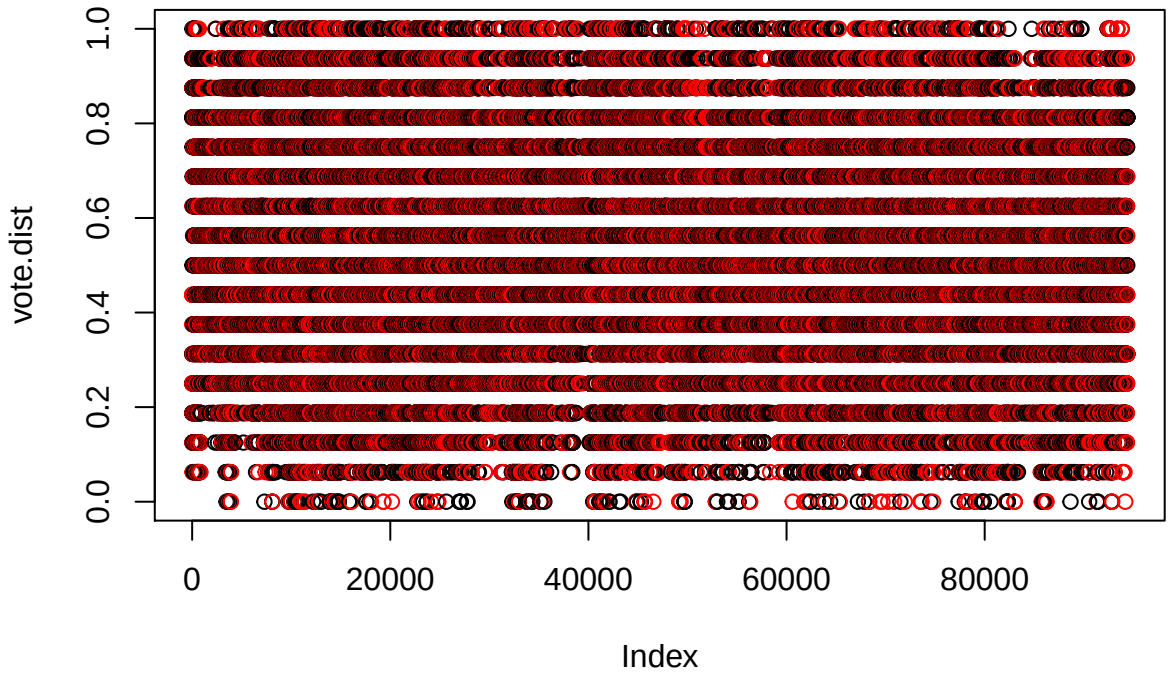
```
##      Petros       Peter       Piotr      Peadar      Pierre       Peder
##           1           1           1           1           1           1
##        Peka      Pietro       Piero        Petr       Pyotr   Cristovao
##           1           1           1           1           1           2
##   Christoph  Christophe   Cristobal  Cristoforo  Kristoffer     Krystof
##           2           2           2           2           2           2
## Christopher      Miguel    Michalis     Michael      Mikael        Mick
##           2           3           3           3           3           3
```

Begge metoder giver samme resultat, når der bruges 3 klynger.

vote.csv

```
?daisy
```

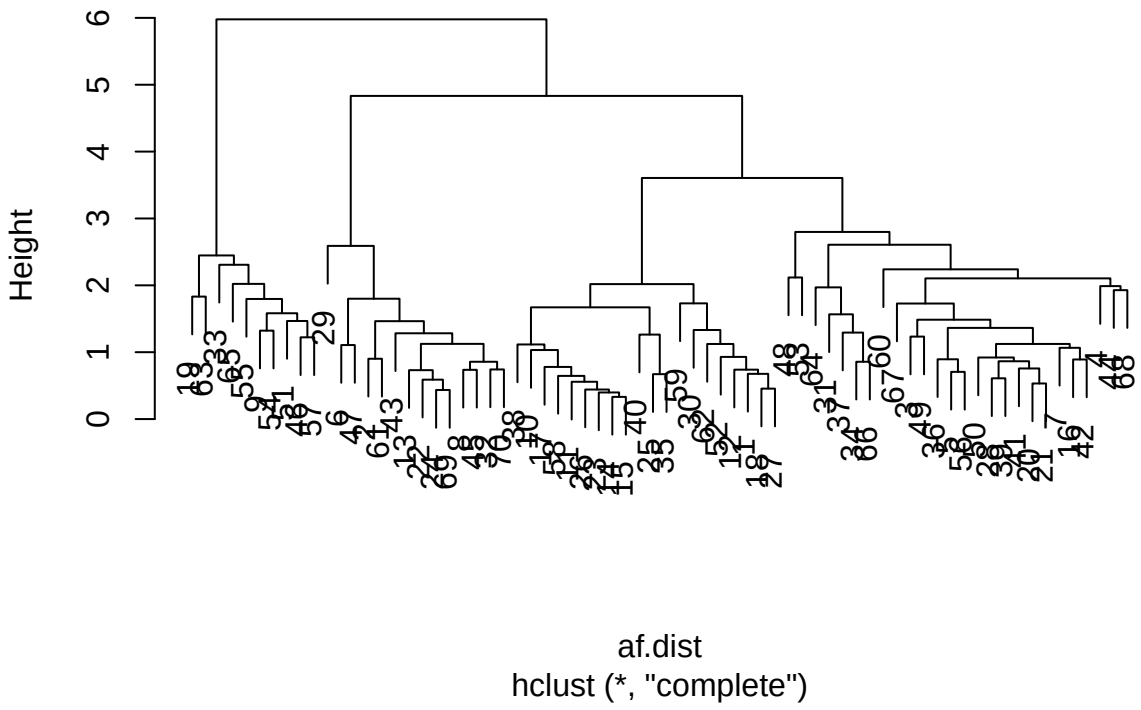
```
vote.dist <- daisy(vote[,2:17])
vote.kmeans <- kmeans(vote.dist, centers = 2, nstart = 10)
plot(vote.dist, col = vote.kmeans$cluster)
```



wine.csv

aims_freq.csv

```
af.dist <- dist(af[,4:100])
af.hclust <- hclust(af.dist)
plot(af.hclust)
```


Cluster Dendrogram
Height
0
1
2
3
4
5
6
af.dist
hclust (*, "complete")