Cluster Analysis

7109018032謝綺珊

先使用原始資料進行分析

```
data<-read.csv("insect.csv")
data.clust<-data[,-(1:4)]      #移除不是用來記錄昆蟲數量的欄位
```

階層式分群

```
E.dist <- dist(data.clust, method="euclidean")     #歐式距離
M.dist <- dist(data.clust, method="manhattan")     #曼哈頓距離
```

使用anges()函數，此函數跟hclust()很類似，可用於計算聚合係數(agglomerative coefficient)來比較多組分群連結演算法的效果，函數中的method參數不包含“ward.D2”。
聚合係數是衡量群聚結構被辨識的程度，聚合係數越接近1代表有堅固的群聚結構(strong clustering structure)。

```
E.dist <- dist(data.clust, method="euclidean") #歐式距離
M.dist <- dist(data.clust, method="manhattan") #曼哈頓距離
library(cluster)
m=c("single","complete","average", "ward")
E.ac<-function(x){agnes(E.dist, method = x)$ac}
M.ac<-function(x){agnes(M.dist, method = x)$ac}
library(purrr)
```

```
## Warning: package 'purrr' was built under R version 4.0.5
```

```
map_dbl(m, E.ac)
```

```
## [1] 0.9456646 0.9423731 0.9421172 0.9503782
```

```
map_dbl(m, M.ac)
```

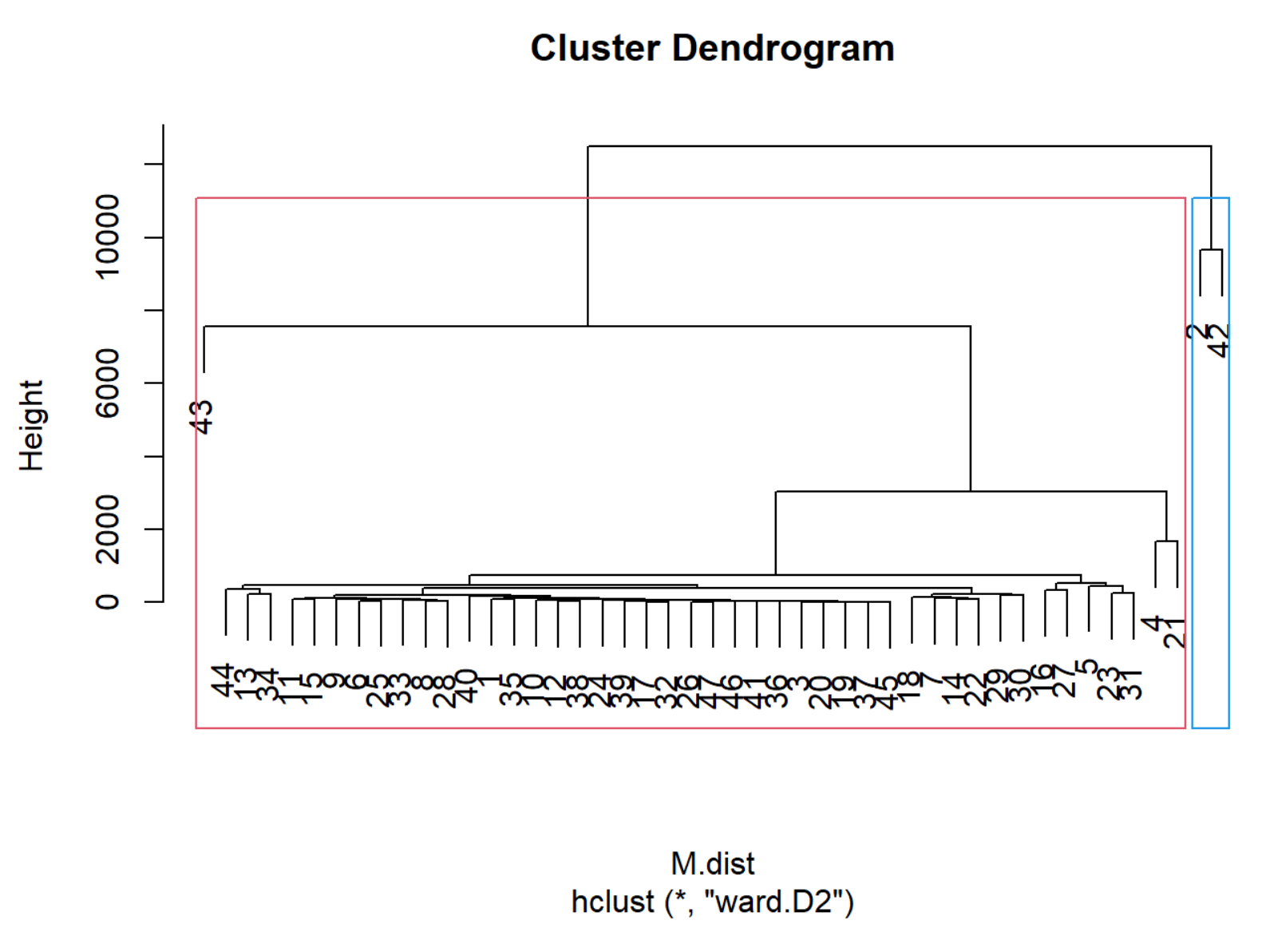
```
## [1] 0.9229604 0.9286213 0.9257725 0.9407650
```

發現以歐式距離搭配華德法效果最好，因此使用ward.D2的華德法

```
h.cluster <- hclust(M.dist, method="ward.D2")      #華德法
```

將分群結果視覺化

```
plot(h.cluster)
rect.hclust(h.cluster, k=2, border=c(2,4))
```



用平均側影法(Average Silhouette method)找尋最佳分群數目

```
require(factoextra)
```

```
## Loading required package: factoextra
```

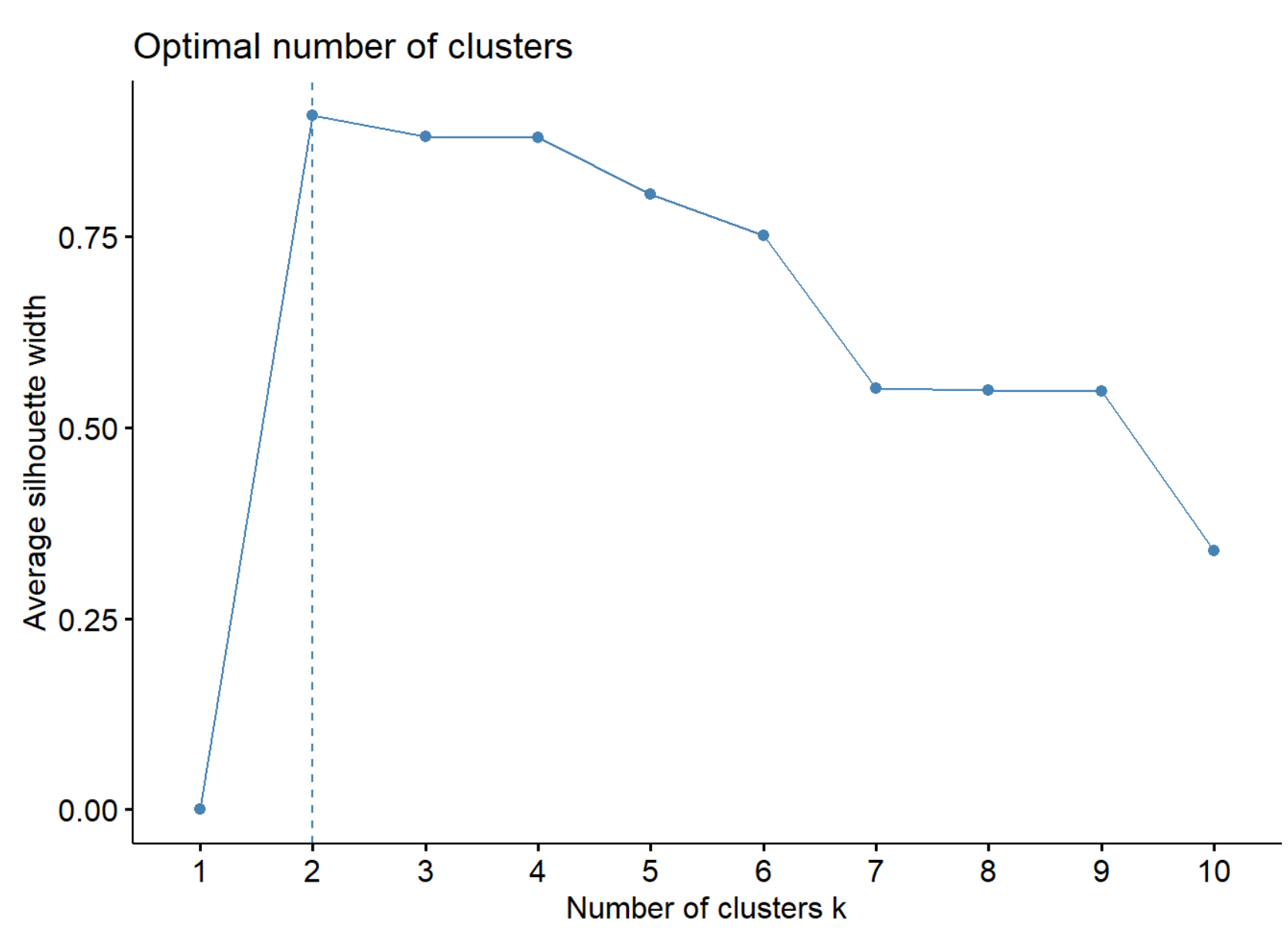
```
## Warning: package 'factoextra' was built under R version 4.0.5
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
fviz_nbclust(data.clust, hcut, method = "silhouette",
             hc_method = "ward.D2")
```

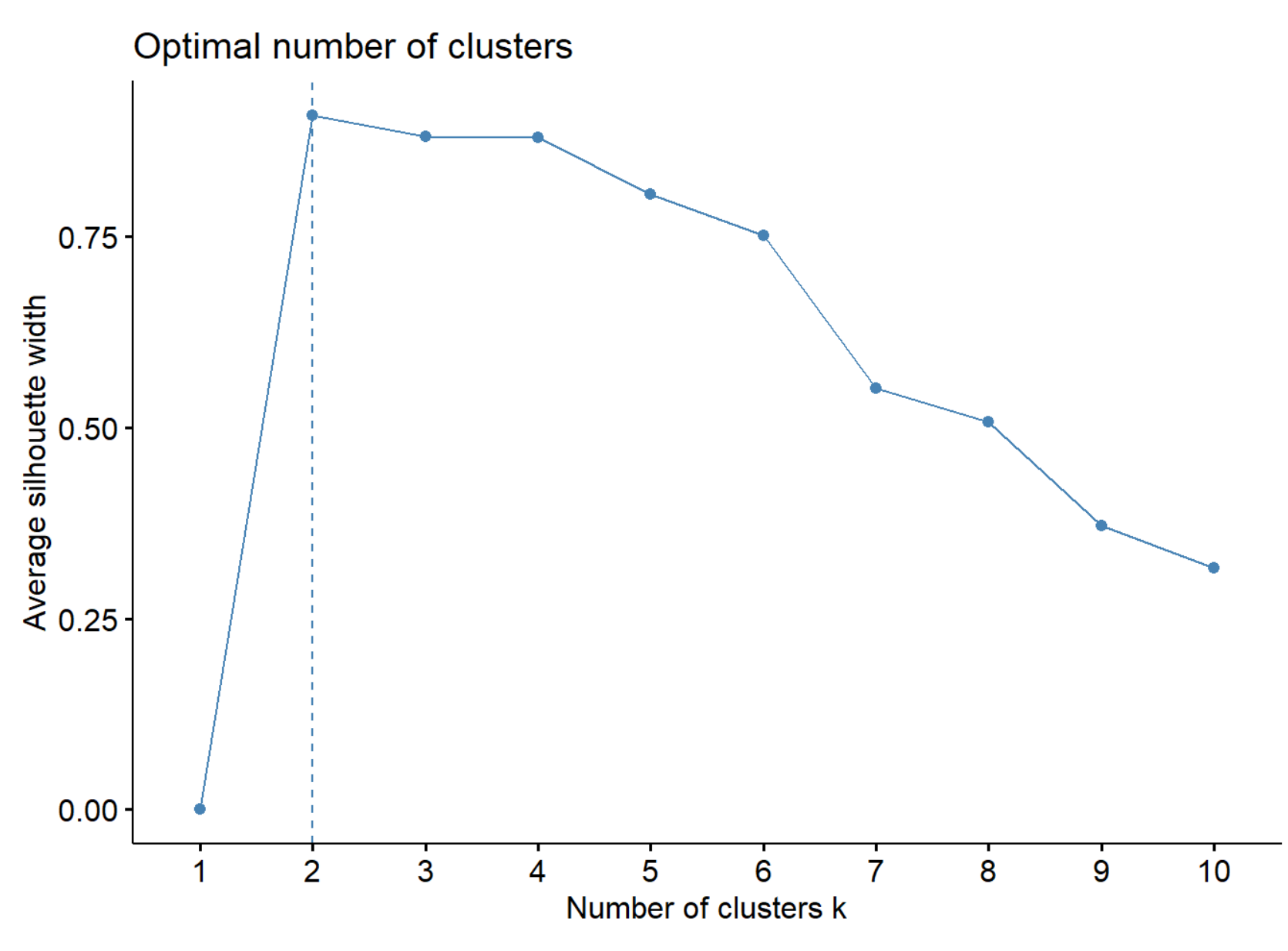


最佳分群數方法的結果為2

K-means

先試著做圖找出最佳分群數，以利套用K-means的指令

```
fviz_nbclust(data.clust, kmeans, method = "silhouette")
```



```
kmeans.fit <- kmeans(data.clust, centers = 2, nstart = 10)
```

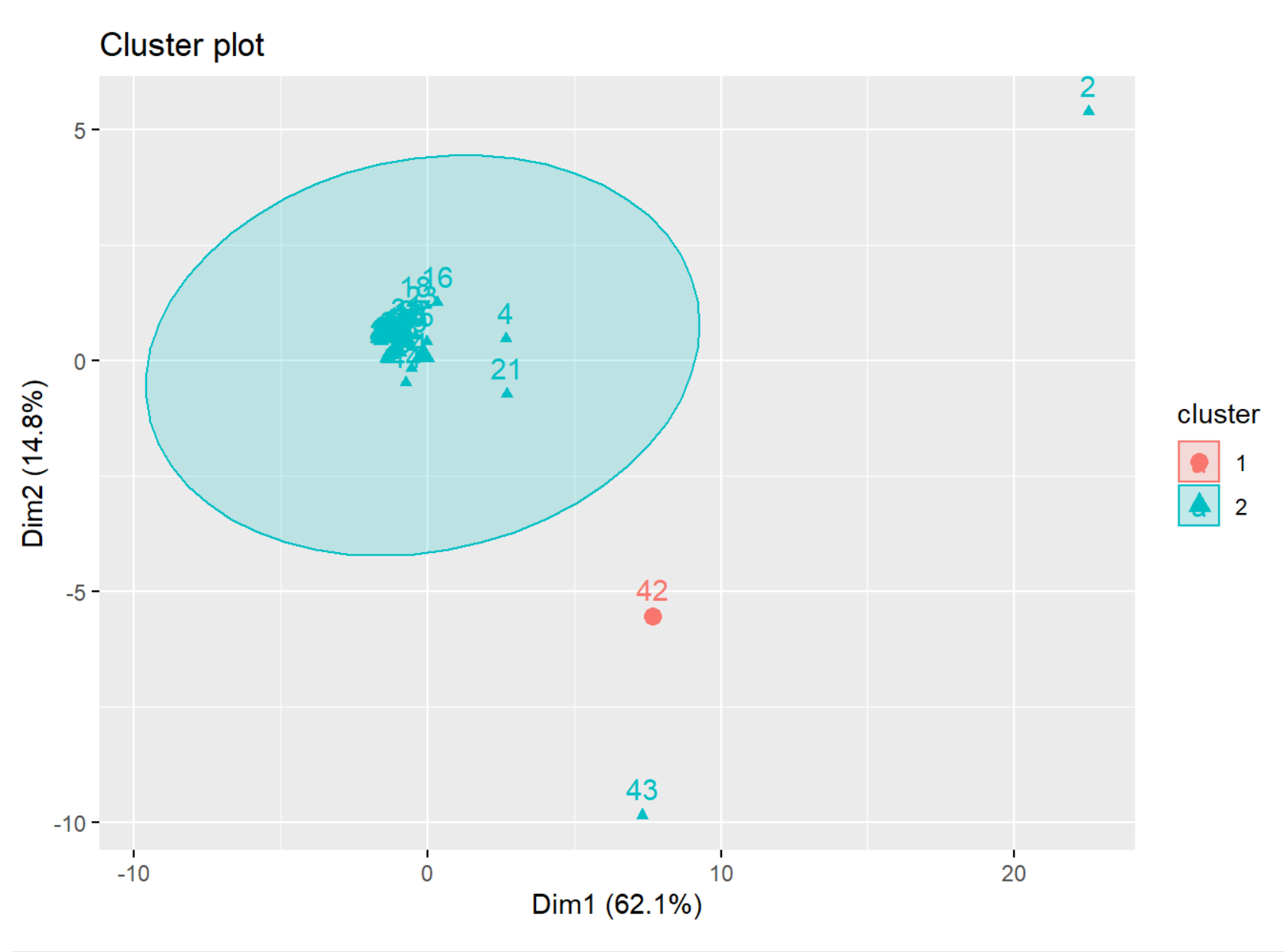
將K-means分群結果視覺化

```
fviz_cluster(kmeans.fit, data = data.clust,
             geom = c("point","text"), frame.type = "norm")
```

```
## Warning: argument frame is deprecated; please use ellipse instead.
```

```
## Warning: argument frame.type is deprecated; please use ellipse.type instead.
```

```
## Too few points to calculate an ellipse
```

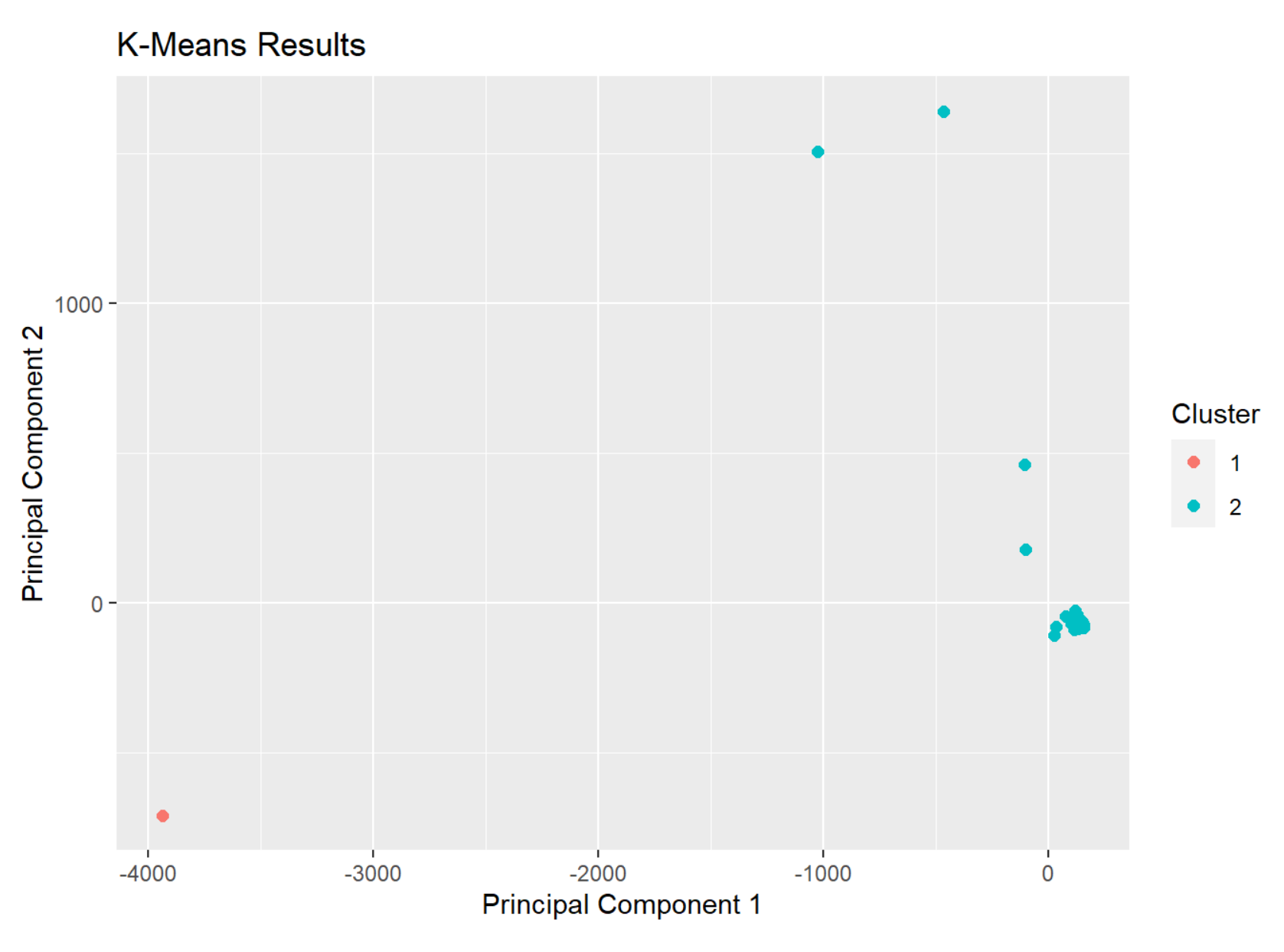


```
require(useful)
```

```
## Loading required package: useful
```

```
## Warning: package 'useful' was built under R version 4.0.5
```

```
plot(kmeans.fit,data=data.clust)
```



試著對資料進行標準化，再執行一次上述步驟，觀察其分群效果

```
means = apply(data.clust, 2, mean)
sds = apply(data.clust, 2, sd)
data.clust = scale(data.clust, center=means, scale=sds)
```

階層式分群

```
E.dist <- dist(data.clust, method="euclidean") #歐式距離
M.dist <- dist(data.clust, method="manhattan") #曼哈頓距離
library(cluster)
m=c("single","complete","average", "ward")
E.ac<-function(x){agnes(E.dist, method = x)$ac}
M.ac<-function(x){agnes(M.dist, method = x)$ac}
library(purrr)
map_dbl(m, E.ac)
```

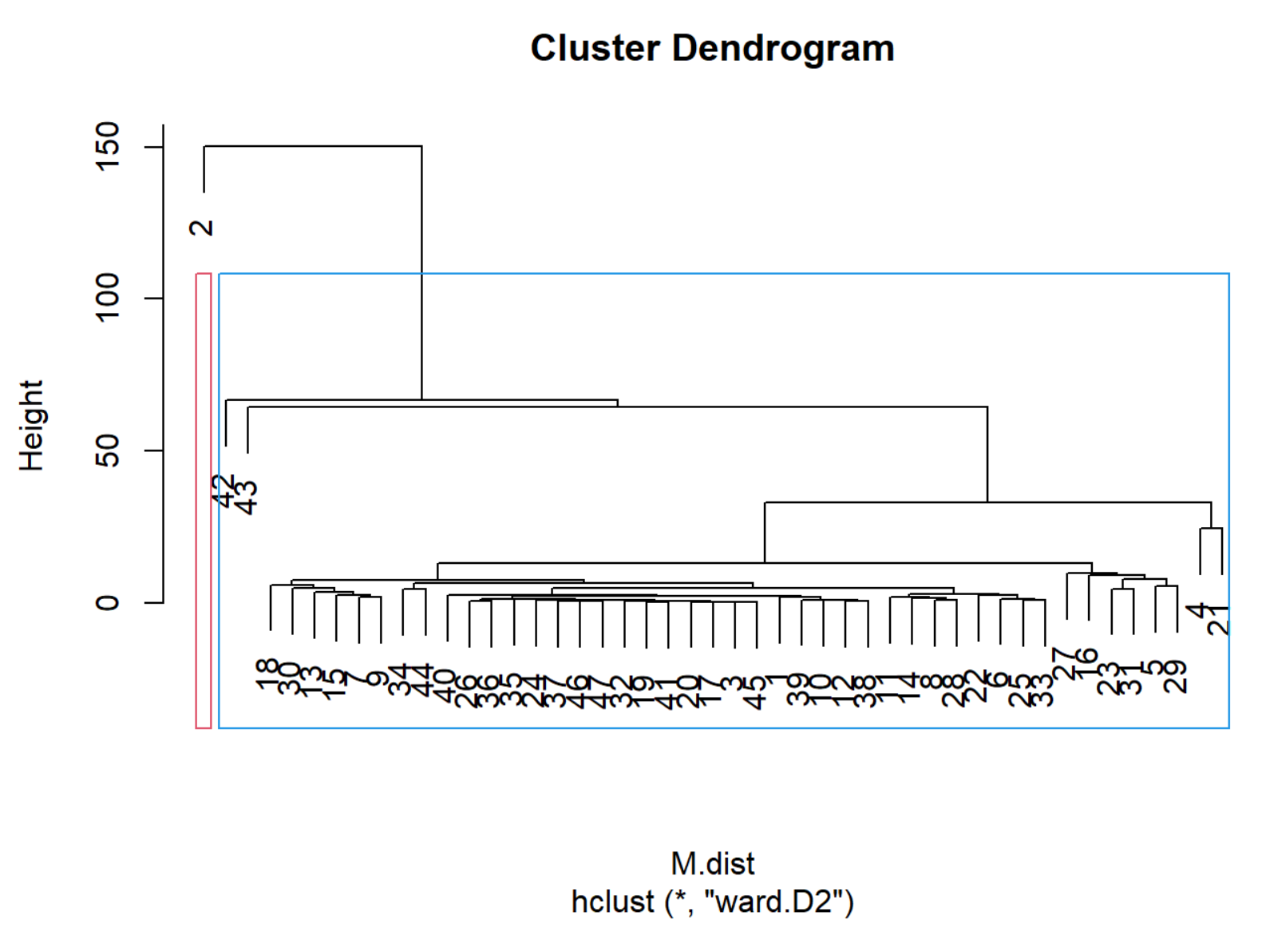
```
## [1] 0.9131797 0.9116310 0.9114719 0.9259015
```

```
map_dbl(m, M.ac)
```

```
## [1] 0.9339176 0.9294791 0.9348670 0.9404733
```

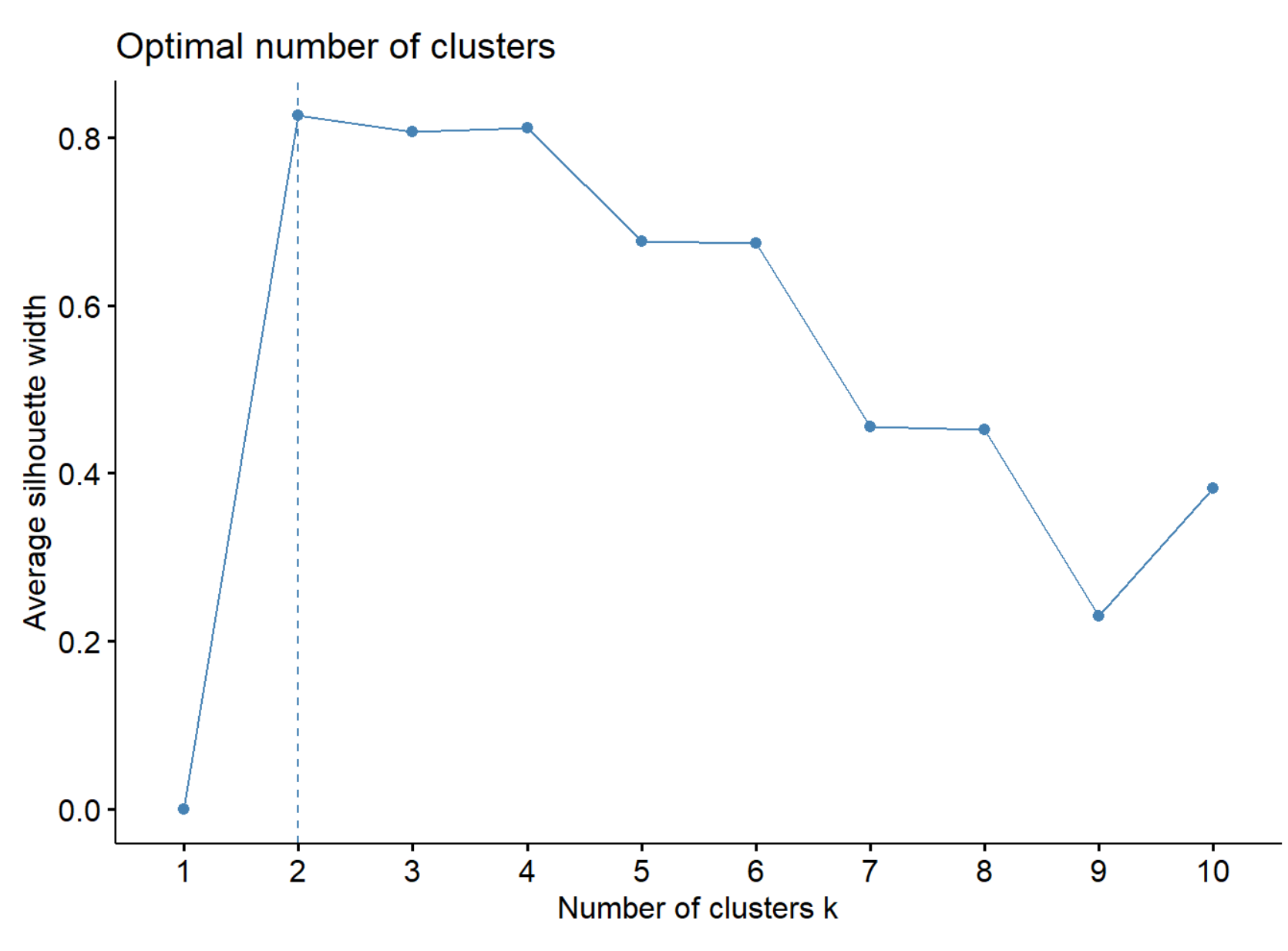
這邊可以發現跟上述情況不同，聚合係數以曼哈頓距離搭配華德法最高。

```
h.cluster <- hclust(M.dist, method="ward.D2")
plot(h.cluster)
rect.hclust(h.cluster, k=2, border=c(2,4))
```



K-means

```
fviz_nbclust(data.clust, kmeans, method = "silhouette")
```

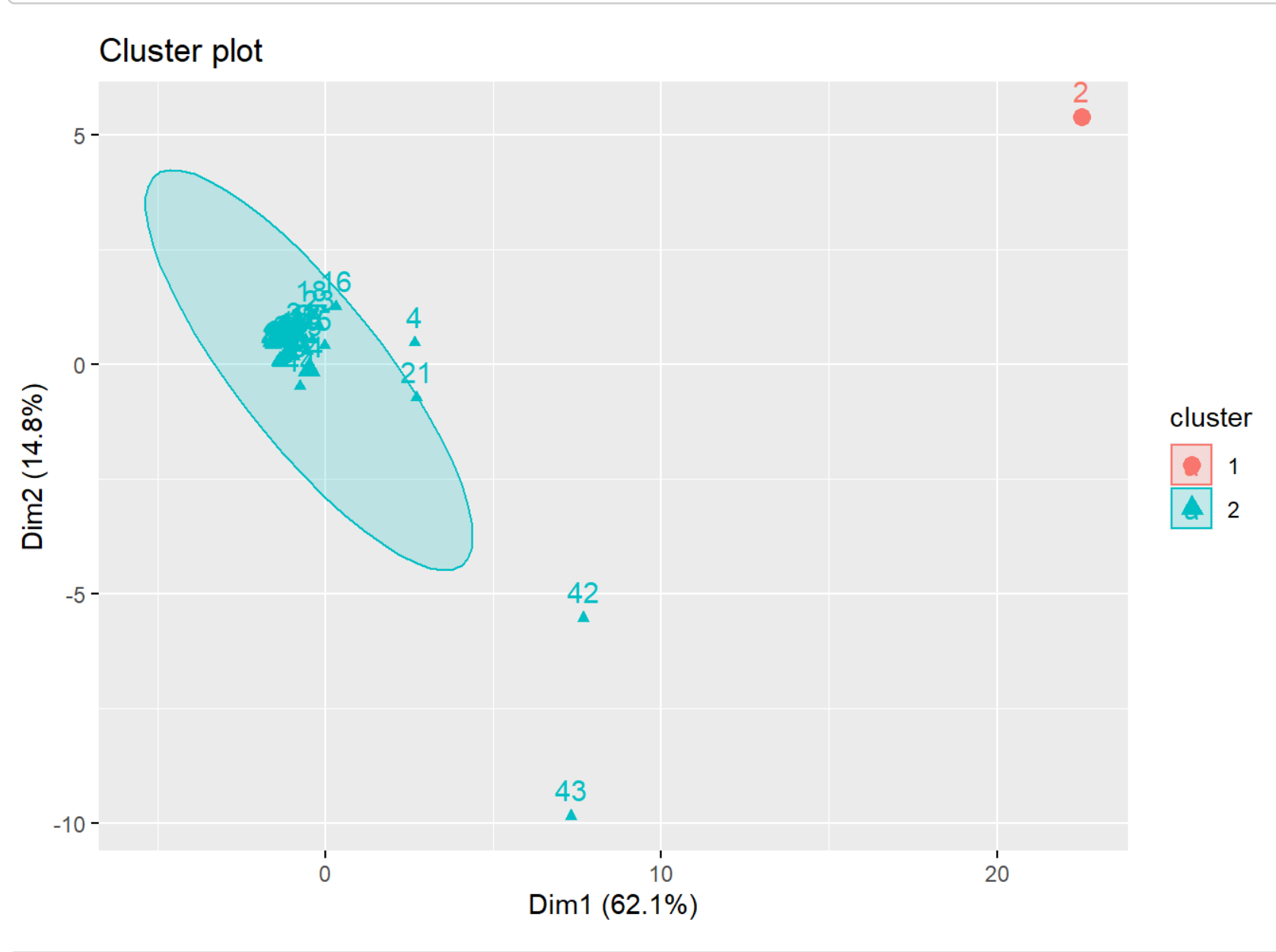


```
kmeans.fit <- kmeans(data.clust, centers = 2, nstart = 10)
fviz_cluster(kmeans.fit, data = data.clust,
             geom = c("point","text"), frame.type = "norm")
```

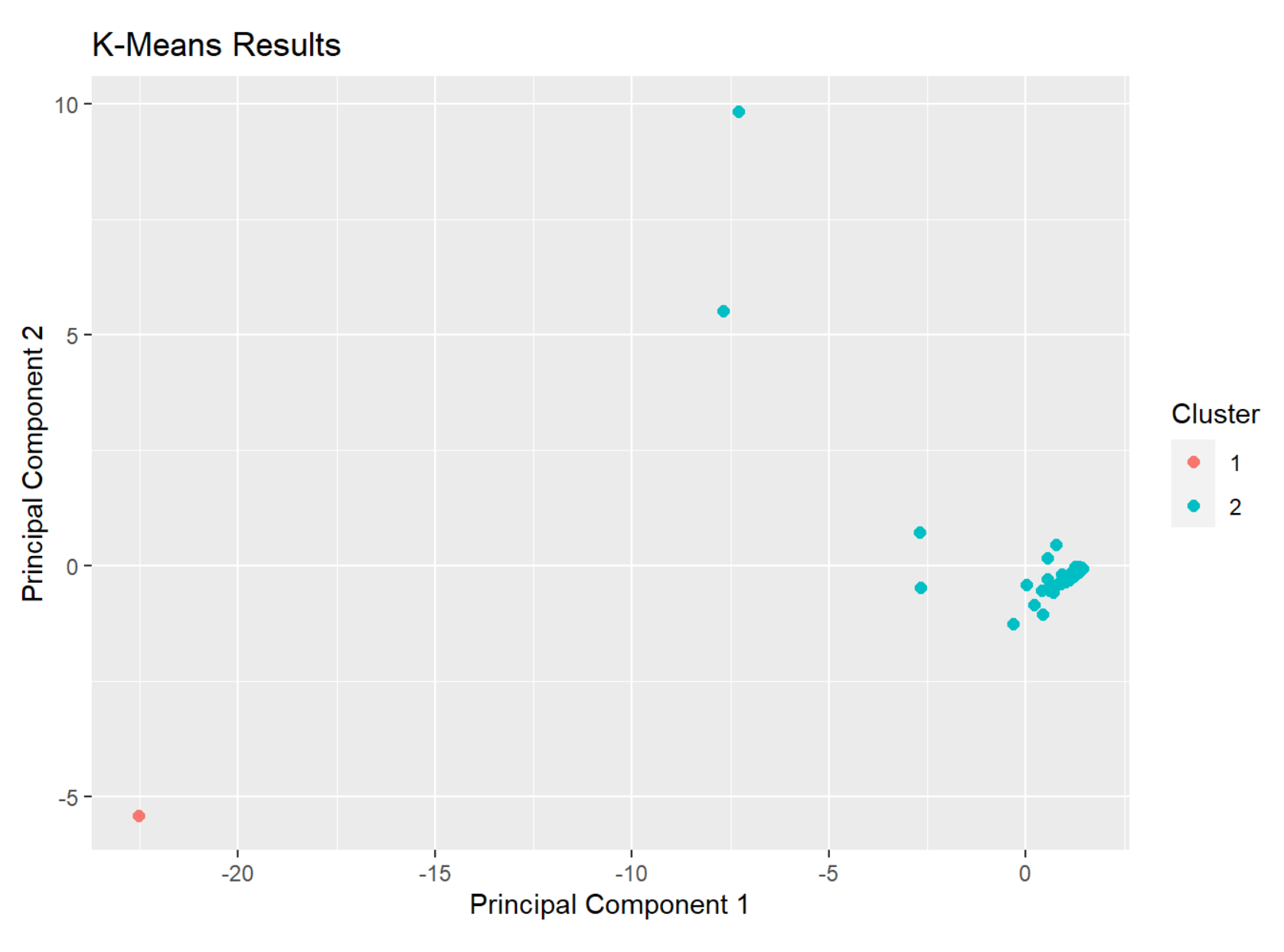
```
## Warning: argument frame is deprecated; please use ellipse instead.
```

```
## Warning: argument frame.type is deprecated; please use ellipse.type instead.
```

```
## Too few points to calculate an ellipse
```



```
require(useful)
plot(kmeans.fit,data=data.clust)
```



結論

目前看出來對這筆資料而言，分析過程中有無對資料使用標準化，其影響似乎不大，但可以發現目前分不好的點是id編號2、42、43，觀察原始資料發現這三種昆蟲在四年間補捉到的總數量是五千隻以上，相較於其他而言，總數量差距甚大，推測因差距太多，以致於分群效果可能不佳