herichal clusteringالعنقدة التراتيبية

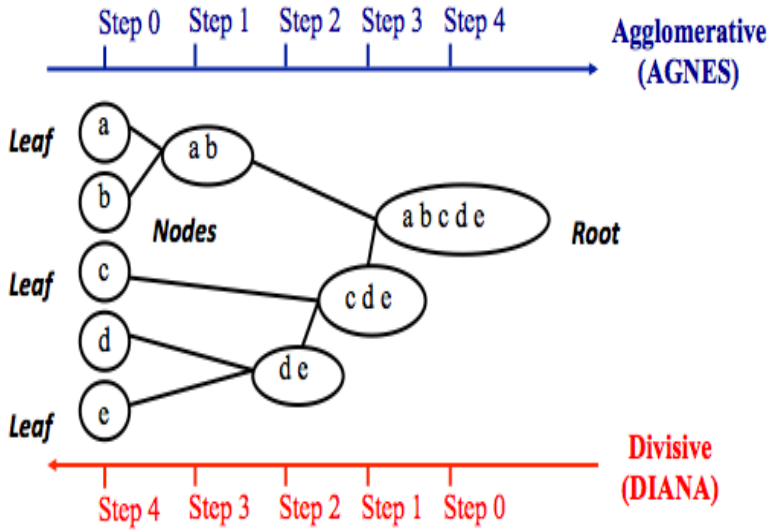
خوارزميات التجميع الهرمية

يمكن تقسيم المجموعات الهرمية إلى نوعين رئيسيين: التكتل والانقسام.

العنقدة الهرمية التجميعية: يُعرف أيضًا باسم (AGNESالتعشيش التجميعي). إنه يعمل بطريقة من القاعدة إلى القمة. بمعنى ، يتم اعتبار كل كائن مبدئيًا ككتلة أحادية العنصر (ورقة). في كل خطوة من الخوارزمية ، يتم دمج المجموعتين الأكثر تشابهاً في كتلة أكبر جديدة (العقد). يتم تكرار هذا الإجراء حتى تصبح جميع النقاط عضوًا في مجموعة كبيرة واحدة فقط (الجذر) والنتيجة هي الشجرة التي يمكن رسمها على أنها dendrogram

العنقدة الهرمية للقسمة: يُعرف أيضًا باسم (DIANA "Divise Analysis")ويعمل بطريقة من أعلى إلى أسفل. الخوارزمية هي ترتيب عكسي لـ AGNES يبدأ بالجذر ، حيث يتم تضمين جميع الكائنات في كتلة واحدة. في كل خطوة من خطوات التكرار ، يتم تقسيم الكتلة الأكثر تجانسًا إلى قسمين. يتم تكرار العملية حتى تكون جميع الكائنات في المجموعة الخاصة بها

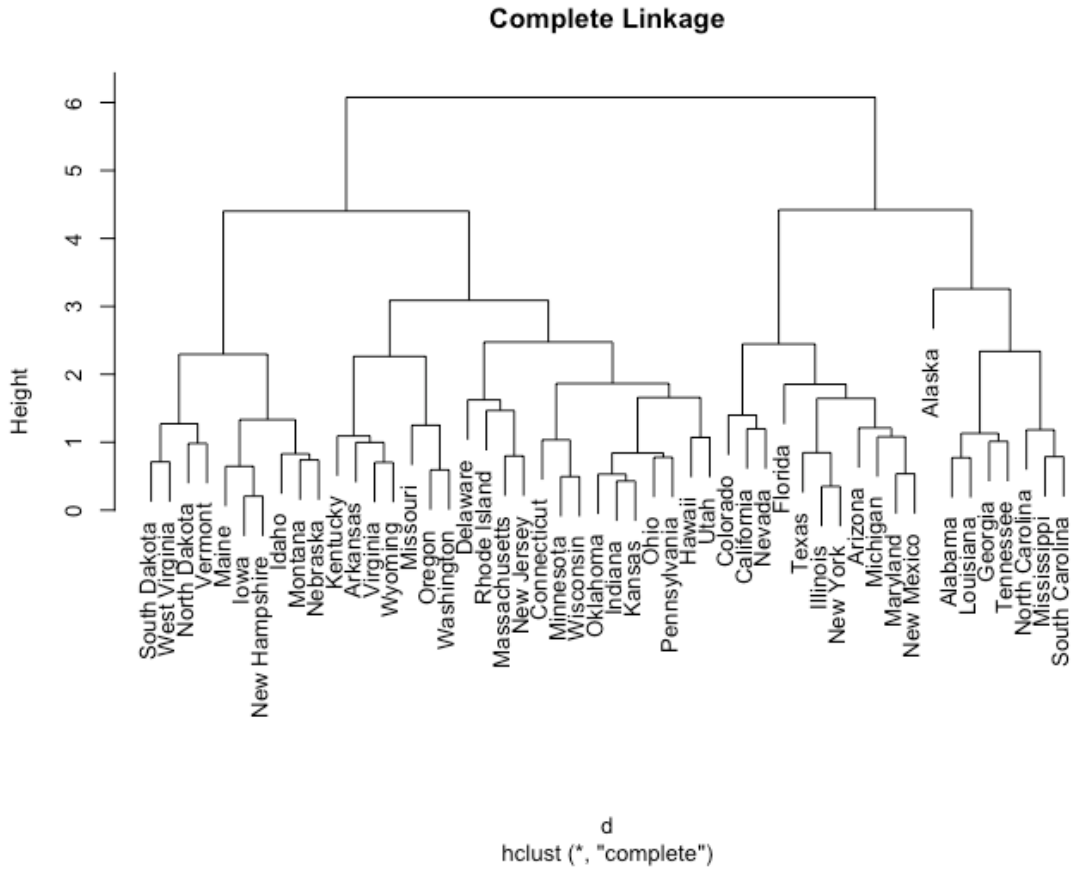
لاحظ أن العنقدة الهرمية التجميعية في تحديد التجمعات الصغيرة. العنقدة الهرمية للقسمة هو جيد في تحديد المجموعات الكبيرة



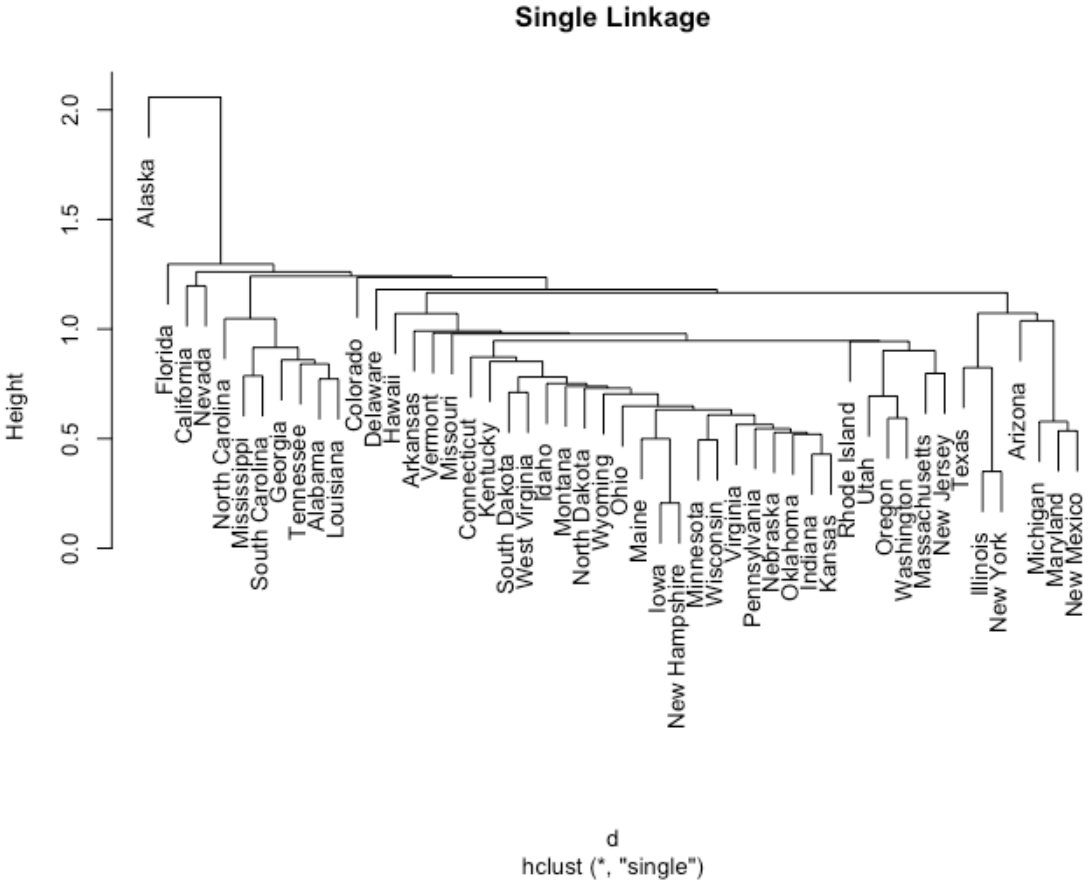
كما تعلمنا في خوارزميةk-mean ، نقيس التشابه بين المشاهدات باستخدام قياسات المسافة (أي المسافة الإقليدية ، مسافة مانهاتن ،...) في R تُستخدم المسافة الإقليدية افتراضيًا لقياس التباين بين كل زوج من المشاهدات. كما نعلم بالفعل ، من السهل حساب مقياس التباين بين اثنين من أزواج المشاهدات مع التابع get_dist()

ومع ذلك ، فإن السؤال الأكبر هو: كيف نقيس التباين بين مجموعتين من الملاحظات؟ تم تطوير عدد من طرق العنقدة الهرمية التجمعية(أي طرق الربط) للإجابة على هذا السؤال فإن أنواع الطرق الأكثر شيوعًا هي:

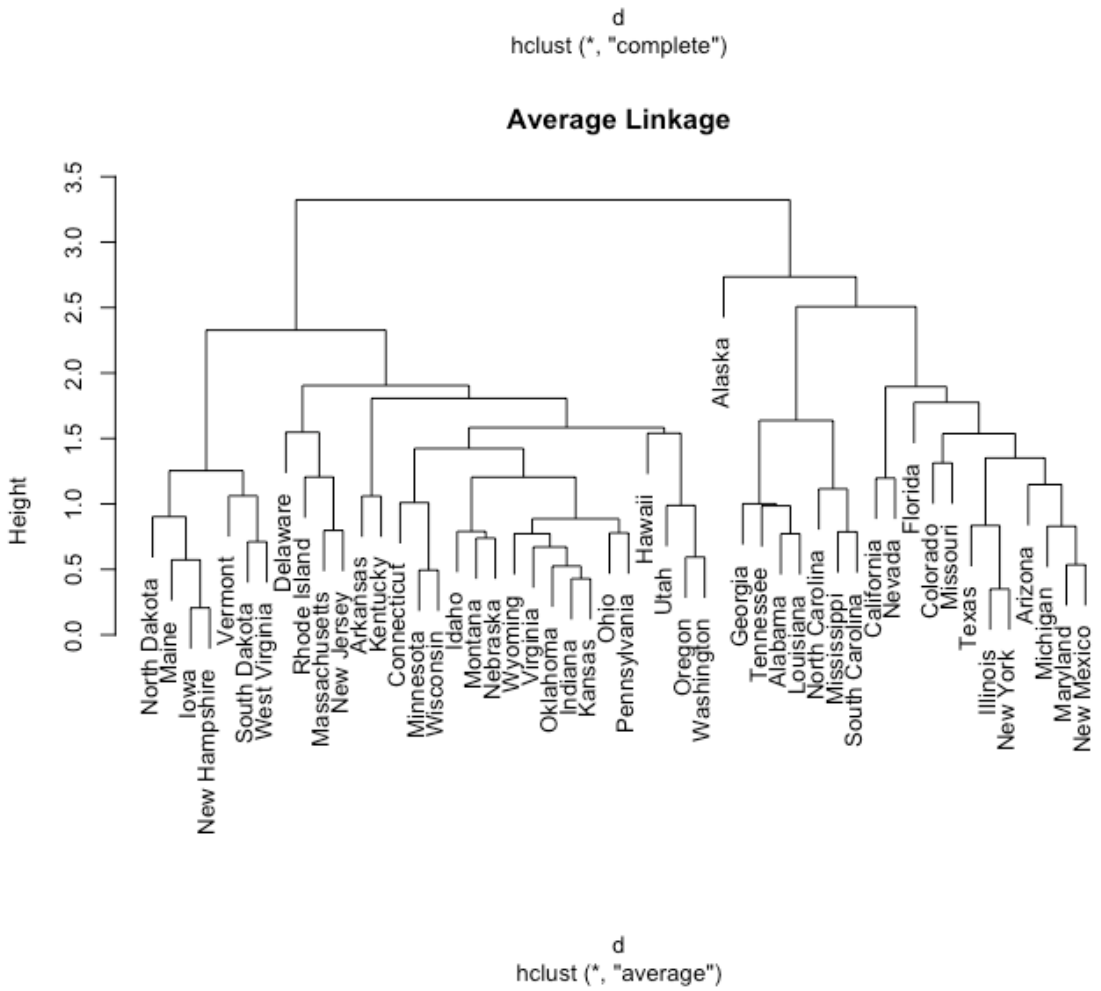
Maximum or complete linkage clustering يقوم بحساب جميع أوجه التباين الزوجي بين العناصر في العنقود 1 والعناصر في العنقود 2 ، ويعتبر أكبر قيمة (على سبيل المثال ، القيمة القصوى) لهذه الاختلافات كمسافة بين العنقودين. تميل إلى إنتاج المزيد من العناقيد المدمجة.



Minimum or single linkage clustering: تحسب جميع أوجه التباين الزوجي بين العناصر في العنقود 1 والعناصر في العنقود 2 ، وتعتبر أصغر هذه الاختلافات كمعيار للربط. تميل إلى إنتاج مجموعات طويلة غير متداخلة

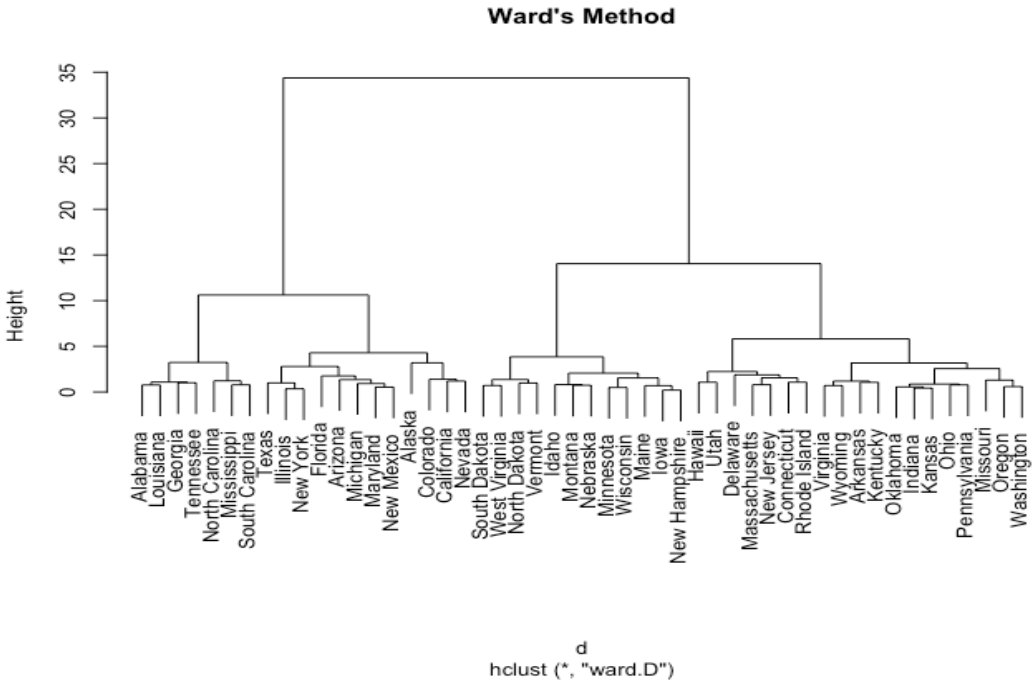


Mean or average linkage clustering :يحسب جميع أوجه التباين الزوجي بين العناصر في العنقود 1 والعناصر في العنقود 2 ، ويعتبر متوسط هذا التباين كمسافة بين العنقودين

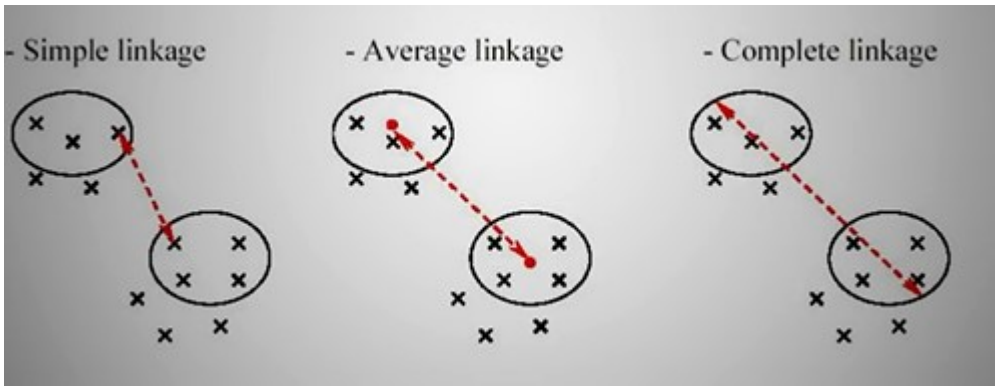


Centroid linkage clustering: يحسب التباين بين مركز العنقود 1ومركز العنقود 2

Ward's minimum variance method:تقلل من التباين الكلي داخل المجموعة. في كل خطوة ، يتم دمج زوج العناقيد من خلال المسافة الدنيا بين العنقودين



اختصار الأنماط الثلاثة



إعداد البيانات

لإجراء العنقدة في R ، بشكل عام ، يجب إعداد البيانات على النحو التالي:

1-الصفوف عبارة عن مشاهدات (أفراد) والأعمدة متغيرات
2-يجب إزالة أو تقدير أي قيمة مفقودة في البيانات.
3-يجب توحيد البيانات (على سبيل المثال ، تقييسها) لجعل المتغيرات قابلة للمقارنة.
بتتحويل المتغيرات بحيث يكون لها صفر وانحراف معياري واحد [scale]
سنقوم لاجراء عنقدة علي بيانات الزهور
بداية إزالة العمود الخامس لانه يحتوي على محارف

```
df<-iris[,1:4]
```

لإزالة أي قيمة مفقودة قد تكون موجودة في البيانات ، اكتب هذا

```
df <- na.omit(df)
```

نبدأ بتوسيع / توحيد البيانات(تقييس)في ال R:

```
df <- scale(df)
head(df)
```

```
> head(df)
     Sepal.Length Sepal.Width Petal.Length Petal.Width
[1,]   -0.8976739  1.01560199    -1.335752   -1.311052
[2,]   -1.1392005 -0.13153881    -1.335752   -1.311052
[3,]   -1.3807271  0.32731751    -1.392399   -1.311052
[4,]   -1.5014904  0.09788935    -1.279104   -1.311052
[5,]   -1.0184372  1.24503015    -1.335752   -1.311052
[6,]   -0.5353840  1.93331463    -1.165809   -1.048667
```

العنقدة الهرمية في R:

هناك توابع مختلفة متوفرة في R لحساب العناقيد الهرمية. توابع شائعة الاستخدام:

1-hclust [in stats package] and agnes [in cluster package] for agglomerative hierarchical clustering (HC)

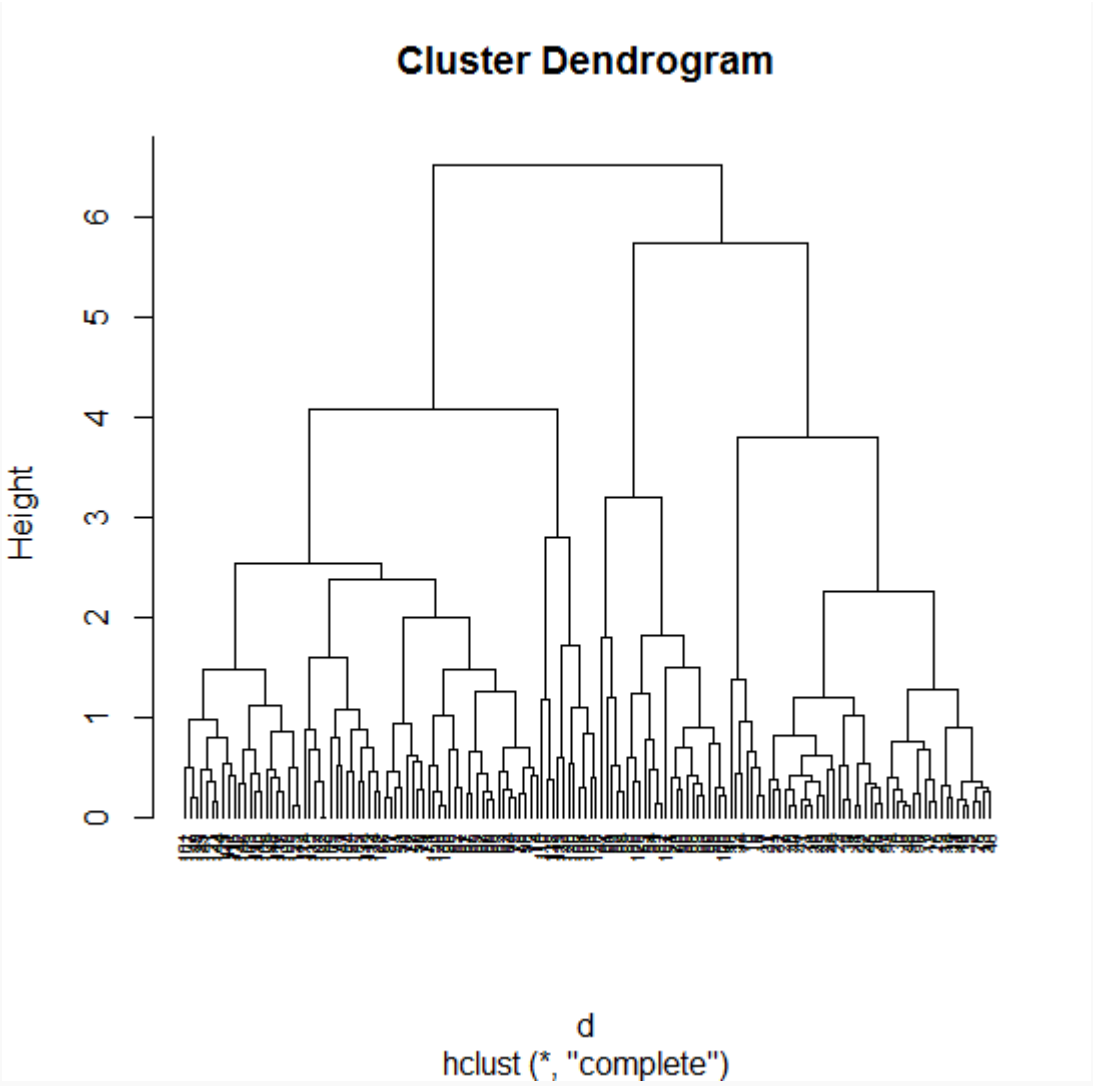
2-diana [in cluster package] for divisive HC

أي ان البند 1 من أجل العنقدة التجميعية التكتلية والبند 2 من اجل العنقدة التجميعية للقسمة

Agglomerative Hierarchical Clustering العنقدة التجميعية التكتلية:

حيث اننا يمكن القيام بالعنقدة ولكن بداية يجب القيام بحساب المسافة dist بين الأزواج من خلال المسافة ثم نحدد طريقة التجميع ضمن التابع hclust("complete") ("average", "single", "ward.D) ثم نرسم مخطط التجميع dendrogram

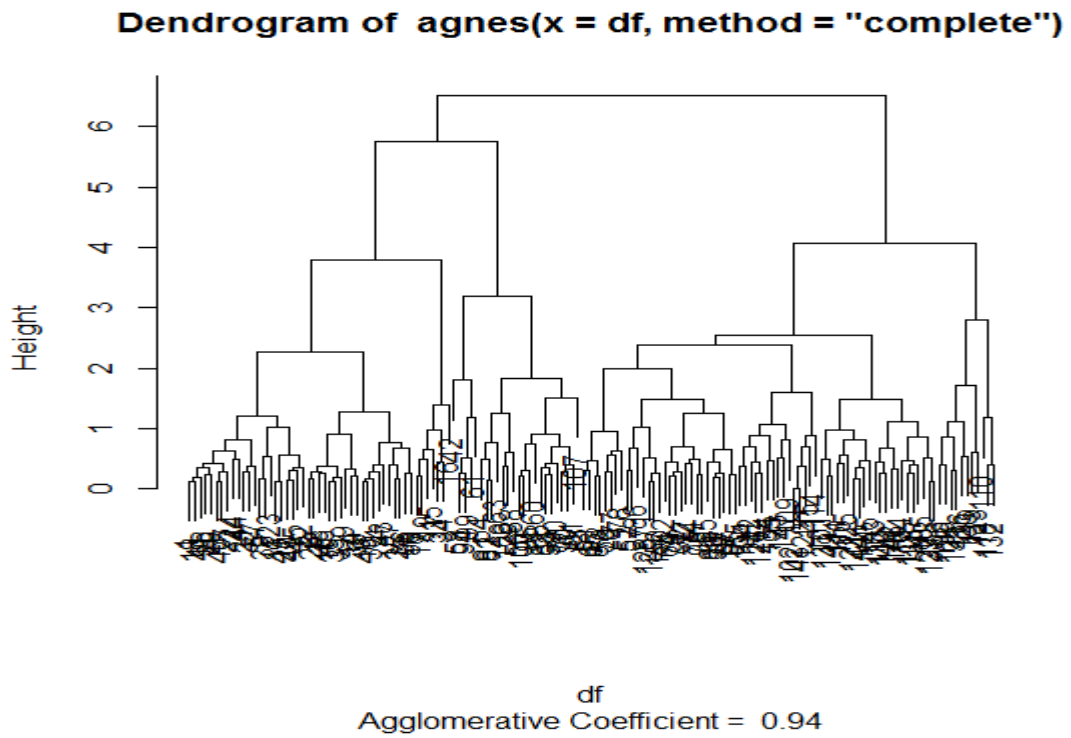
```
# Dissimilarity matrix
d <- dist(df, method = "euclidean")
# Hierarchical clustering using Complete Linkage
hc1 <- hclust(d, method = "complete" )
> hc1 <- hclust(d, method = "complete
> hc1

Call:
hclust(d = d, method = "complete")

Cluster method   : complete
Distance         : euclidean
Number of objects: 150

# Plot the obtained dendrogram
plot(hc1, cex = 0.6, hang = -1)
```



بدلا من ذلك ، يمكننا استخدام تابع `agnes` . هذه الوظائف تتصرف بالمثل. ومع ذلك ، مع تابع `agnes` ،الموجود ضمن مكتبة cluster يمكنك أيضًا الحصول على معامل التكتل ، الذي يقيس مقدار بنية العناقيد الموجودة (تشير القيم الأقرب إلى 1 إلى بنية تجميع قوية)

```
# Compute with agnes
# Agglomerative coefficient
hc2$ac
## [1] 0.9438858
Plot(hc2)
```

```
> hc2 <- agnes(df, method = "complete")
> hc2
Call:     agnes(x = df, method = "complete")
Agglomerative coefficient:  0.9438858
Order of objects:
  [1]   1  18  41  28  29   8  40  27  24  44  21  32  37   5  38  23  11  49
 [19]  20  47  45  22   2  26  13  46  10  35  31   9  14  39   3  48  30   4
 [37]  43   7  12  25  36  50   6  17  19  15  33  34  16  42  58  94  99  61
 [55]  54  81  82  63  69 120  88  56 100  95  60  68  83  93  80  70  90  91
 [73] 107  51  53  66  87  59  76  77  78  52  57  86  71 128 139 150  62  64
 [91]  79  92  72  74  75  98  65  89  96  97  67  85  55 134 112 124 127  84
[109] 135  73 147 109 102 143 122 115 114 101 137 149 111 116 125 121 144 141
[127] 145 103 113 140 142 146 104 117 138 148 105 129 133 106 136 108 131 126
[145] 130 119 123 110 118 132
Height (summary):
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.2858  0.4613  0.7550  0.8539  6.5075

Available components:
[1] "order"  "height" "ac"     "merge"  "diss"   "call"   "method" "data"
```



وبامكاننا تطبيق طرق التجميع الهرمية التي يمكنها تحديد هياكل أقوى للتكتل. هنا نرى أي طريقة واردة تحدد أقوى بنية عنقودية في الطرق الأربعة التي تم تقييمها:

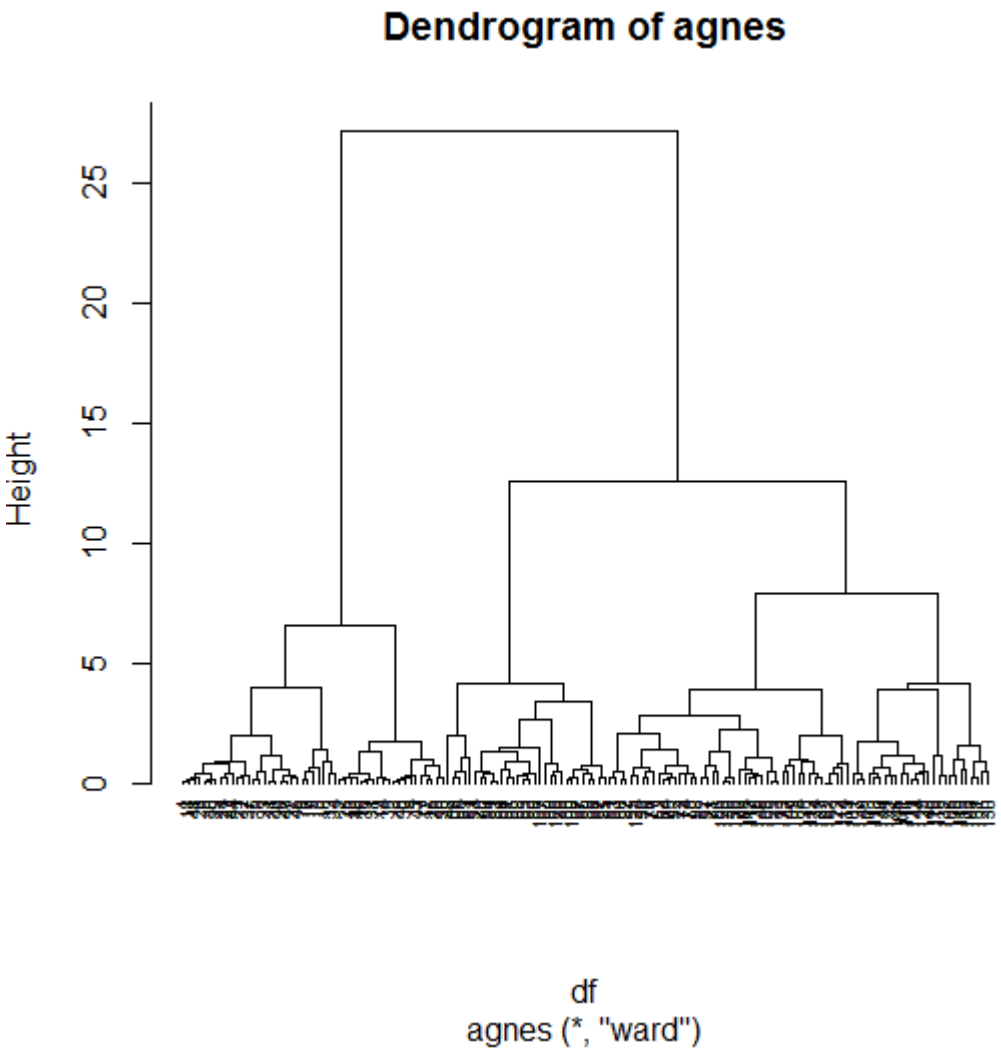
```
# methods to assess
m <- c( "average", "single", "complete", "ward")
names(m) <- c( "average", "single", "complete", "ward")

# function to compute coefficient
ac <- function(x) {
```

```
  agnes(df, method = x)$ac
}

map_dbl(m, ac)
```

```
  average    single  complete      ward
```

وبطريقة مشابهة:

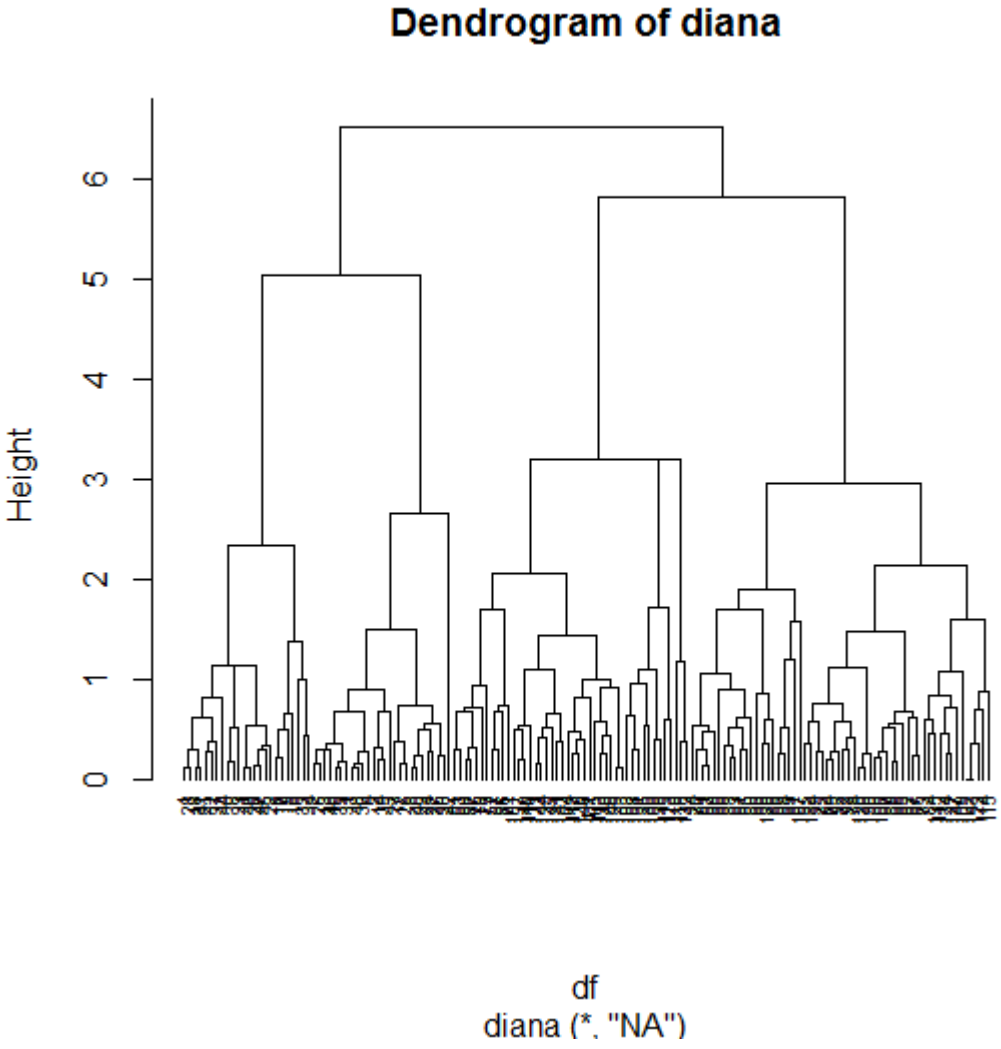
```
hc3 <- agnes(df, method = "ward")
pltree(hc3, cex = 0.6, hang = -1, main = "Dendrogram of
agnes")
```



Divisive Hierarchical Clustering

```
# compute divisive hierarchical clustering
```

```
hc4 <- diana(df)

# Divise coefficient; amount of clustering structure
found
hc4$dc
## [1] 0.8514345

# plot dendrogram
pltree(hc4, cex = 0.6, hang = -1, main = "Dendrogram of
diana")
```

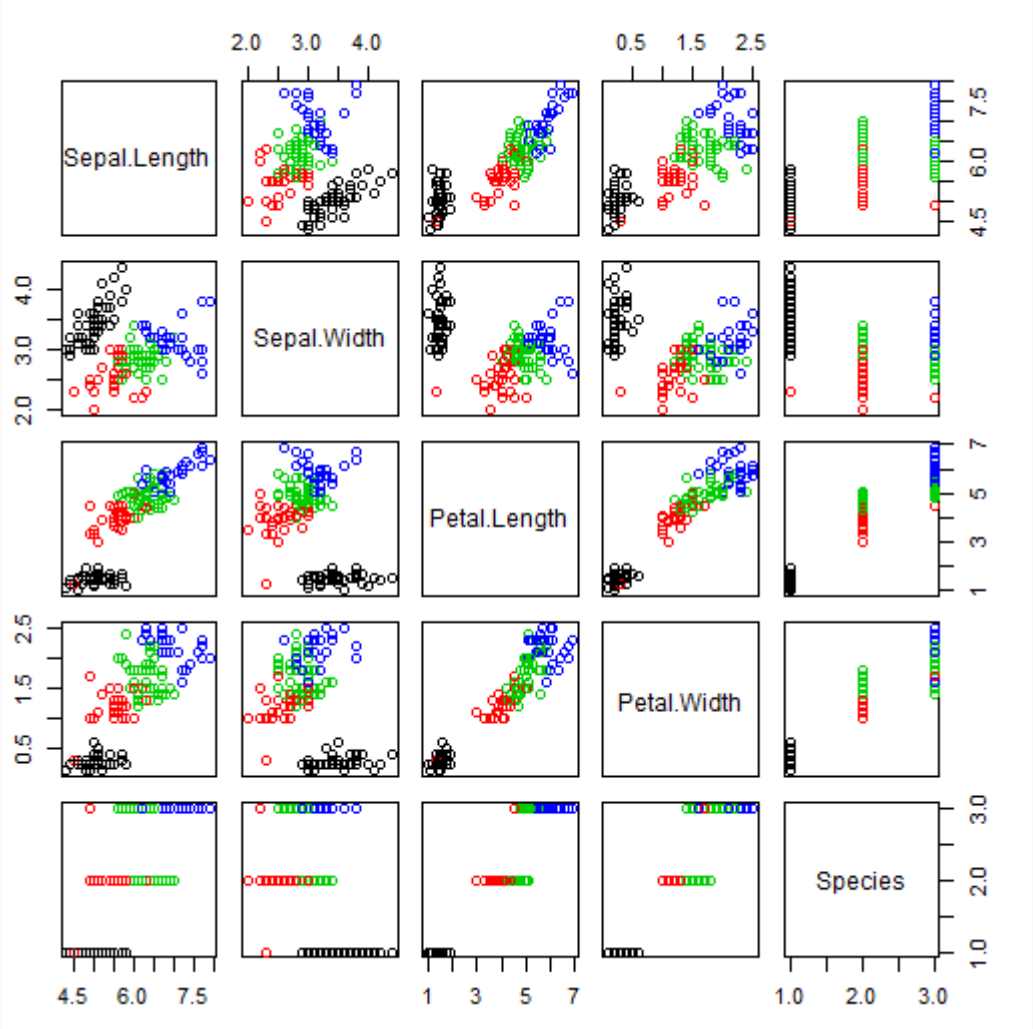


العمل مع المخطط العنقودي:
في الشكل السابق نجد أن كل ورقة عبارة عن مشاهدة من المشاهدات التي تشكل الشجري وكل مجموعة من المشاهدات متشابهة في خصائص معينة تشكل عصنا أو فرعا من الشجرة التي يتم دمجها في ارتفاع اعلى، ارتفاع الدمج يتم تمثيله على المحور

العمودي الذي يشير الى التباين أو التشابه بين المشاهدات.وعندما يكون التشابه قليل المشاهدات يكون ارتفاع الدمج أقل.

يمكننا تجزيء العنقود الواحد ونقطع منه شجيرات باستخدام التابعcutree للحصول على مجموعات(عناقيد) ثانوية:

```
# Ward's method
hc5 <- hclust(d, method = "ward.D2" )

# Cut tree into 4 groups
sub_grp <- cutree(hc5, k = 4)

# Number of members in each cluster
table(sub_grp)
table(sub_grp)
sub_grp
 1  2  3  4
49 30 45 26
plot(iris,col=sub grp)
```



إن خرج الcutree هو تحديد العناصر لأي عنقود تنتمي

sub_grp

```
[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 1 1 1 1 1 1 1 1
1 1 1
[38] 1 1 1 1 2 1 1 1 1 1 1 1 1 3 3 3 2 3 2 3 2 3 2 2 3 2 3 2 3 2 2 2 2 3
3 3 3
[75] 3 3 3 3 3 2 2 2 2 3 2 3 3 2 2 2 2 3 2 2 2 2 2 3 2 2 4 3 4 3 3 4 2 4
3 4 4
[112] 3 4 3 3 4 3 4 4 2 4 3 4 3 4 4 3 3 3 4 4 4 3 3 3 4 4 3 3 4 4 4 3 4 4
4 3 3
[149] 4 3
```

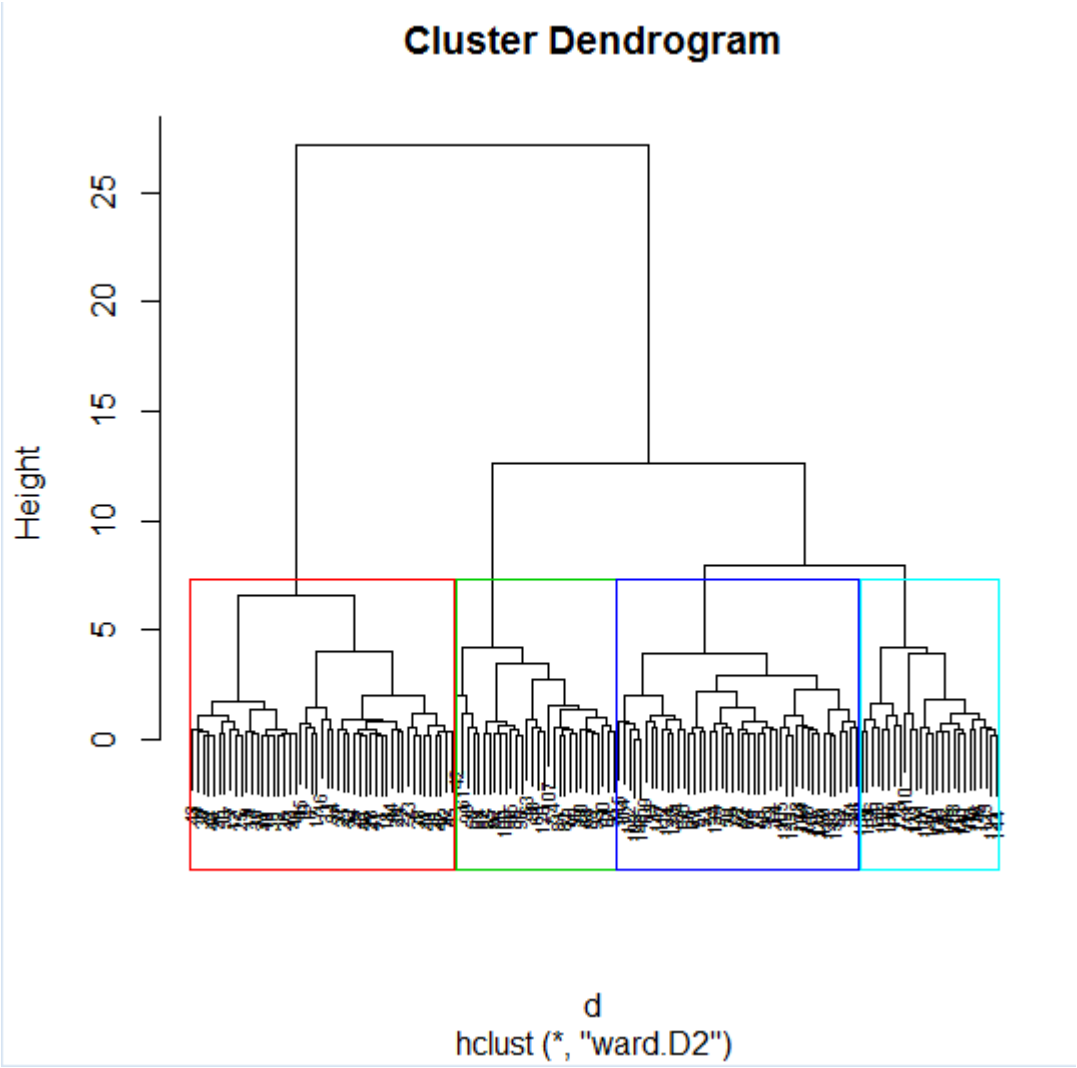
ويمكن إضافة رقم العنقود لكل مشاهدات قاعدة البيانات الاصلية من خلال التابعcbind

```
l<-cbind(df,sub_grp)
```

```
> head(l)
     Sepal.Length Sepal.Width Petal.Length Petal.Width sub_grp
[1,]   -0.8976739  1.01560199    -1.335752   -1.311052       1
[2,]   -1.1392005 -0.13153881    -1.335752   -1.311052       1
[3,]   -1.3807271  0.32731751    -1.392399   -1.311052       1
[4,]   -1.5014904  0.09788935    -1.279104   -1.311052       1
[5,]   -1.0184372  1.24503015    -1.335752   -1.311052       1
[6,]   -0.5353840  1.93331463    -1.165809   -1.048667       1
```

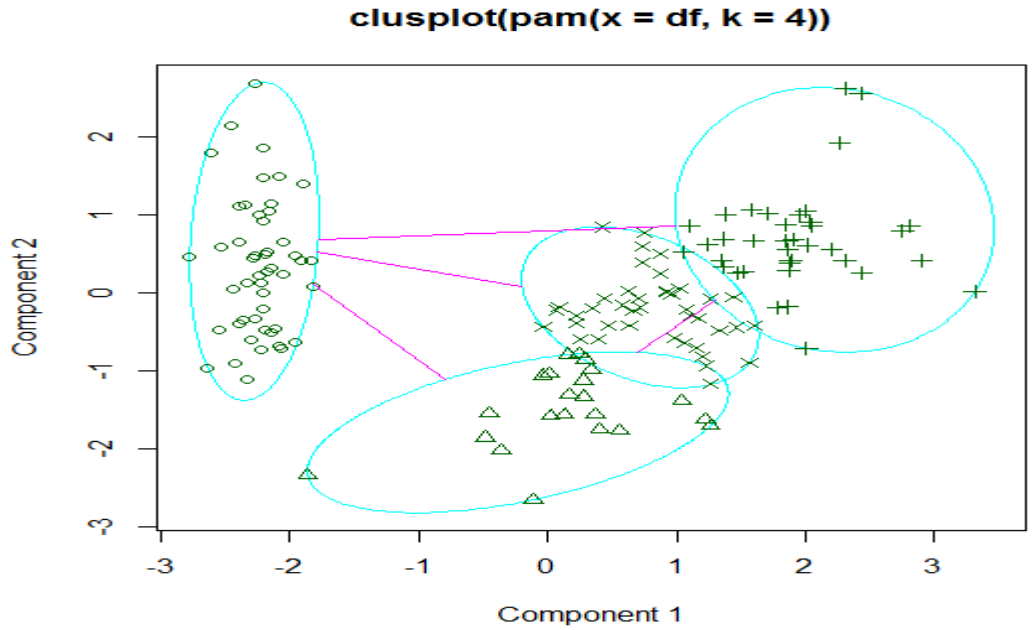
يمكن أيضا أن نحدد العناقيد الفرعية ضمن مستطيلات حيث يرمز kالى عدد المستطيلات ويمكن تلوينها بالوان مختلفة ويتم ذلك من خلالborderالذي يمكن أن يأخذ قيمة عددية او اللون المحدد

```
plot(hc5, cex = 0.6)
rect.hclust(hc5, k = 4, border = 2:5)
or one color
#rect.hclust(hc5, k = 4, border ='red')
```



ويمكننا رسم العناقيد المجمعة:

clusplot(pam(df,4))



لاستخدام cutree مع agensوdiana يمكنك تنفيذ ما يلي:

```
# Cut agnes() tree into 4 groups
hc_a <- agnes(df, method = "ward")
cutree(as.hclust(hc_a), k = 4)
```

```
Call:     agnes(x = df, method = "ward")
Agglomerative coefficient:  0.9867558
Order of objects:
  [1]   1  18  41  28   8  40  29  24  27  44  21  32  37   5  38  23  11  49
 [19]  20  47  22  45   6  17  19  15  16  33  34   2  26  13  46  10  35  31
 [37]   9  14  39   3  48  30   4  43   7  12  25  36  50  42  58  94  99  61
 [55]  54  70  90  81  82  60  91  95  68  83  93  80 107  63  69 120  88  56
 [73] 100  97  65  89  96  67  85  51  53  66  87  55 134  59  76  77  62  64
 [91]  79  92  72  74  75  98  52  57  71  86 128 139 150  78 104 117 138 148
[109] 105 129 133  73 147 109  84 135 112 124 127 102 143 122 114 115 101 137
[127] 149 103 113 140 141 142 146 111 116 121 144 125 145 110 118 132 106 136
[145] 119 123 108 131 126 130
Height (summary):
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.2729  0.4651  1.1058  0.9255 27.1589

Available components:
[1] "order"  "height" "ac"     "merge"  "diss"   "call"   "method" "data"
> cutree(as.hclust(hc_a), k = 4)
  [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 1 1 1 1 1 1 1 1 1
 [38] 1 1 1 1 2 1 1 1 1 1 1 1 3 3 3 2 3 2 3 2 3 2 2 3 2 3 2 3 2 2 2 2 3 3 3 3
 [75] 3 3 3 3 3 2 2 2 2 3 2 3 3 2 2 2 2 3 2 2 2 2 2 3 2 2 4 3 4 3 3 4 2 4 3 4 4
[112] 3 4 3 3 4 3 4 4 2 4 3 4 3 4 4 3 3 3 4 4 4 3 3 3 4 4 3 3 4 4 4 3 4 4 4 3 3
[149] 4 3
```

```
# Cut diana() tree into 4 groups
hc_d <- diana(df)
cutree(as.hclust(hc_d), k = 4)
```

```
 [64] 0.5415773 1.1032774 0.1655887 0.4240501 0.5247707 0.6732405 0.3946778
 [71] 1.4368123 0.4807928 0.2592702 0.4061376 0.8253124 1.0029209 0.5954263
 [78] 0.2653865 0.4389240 0.9318676 0.1311927 3.2054985 0.6507573 0.3118788
 [85] 0.9596625 0.5405840 1.1121328 0.4045414 1.7174836 0.6150517 3.2054985
 [92] 1.1765724 0.3950466 5.8085556 0.5451495 0.3118788 0.1429002 0.4862320
 [99] 1.0637495 0.9017960 0.3531180 0.2363181 0.5306769 0.3118788 0.6221790
[106] 1.7050054 0.8667085 0.3722352 0.6068441 1.9059847 0.2592702 0.5207927
[113] 1.1949691 1.5794859 2.9595367 0.3722352 0.5843290 0.2953231 0.7727037
[120] 0.2112608 0.2960431 0.5859036 0.3100027 0.4176726 1.1303779 0.1333894
[127] 0.2667788 1.4915616 0.2265906 0.2855134 0.5258100 0.1870942 0.5719672
[134] 0.6797003 0.6272758 0.2415266 2.1424883 0.6097954 0.4650710 0.8401728
[141] 0.4565170 0.2653865 0.7199768 1.0766734 1.6022309 0.0000000 0.3755259
[148] 0.7010908 0.8757354
Divisive coefficient:
[1] 0.9397208

Available components:
[1] "order"  "height" "dc"     "merge"  "diss"   "call"   "data"
> cutree(as.hclust(hc_d), k = 4)
  [1] 1 2 2 2 1 1 2 2 2 2 1 2 2 2 1 1 1 1 1 1 1 1 1 2 2 2 2 1 2 2 2 1 1 1 2 2 1
 [38] 1 2 2 1 2 2 1 1 2 1 2 1 2 3 3 3 4 4 4 3 4 4 4 4 4 4 4 3 4 4 4 4 3 4 4 4
 [75] 4 3 3 3 4 4 4 4 4 4 4 3 3 4 4 4 4 4 4 4 4 4 4 4 4 3 4 3 3 3 3 4 3 4 3 3
[112] 4 3 4 4 3 3 3 3 4 3 4 3 4 3 3 4 4 3 3 3 3 3 4 4 3 3 3 4 3 3 3 4 3 3 3 4 3
[149] 3 4
```

وأخيرا ، يمكننا أيضا مقارنة اثنين من dendrograms وهنا نقارن المجموعات الهرمية(بطريقةward&complete). يرسم تابع tanglegram اثنين من dendrograms ، جنبًا إلى جنب ، مع تسمياتهما متصلة بخطوط:

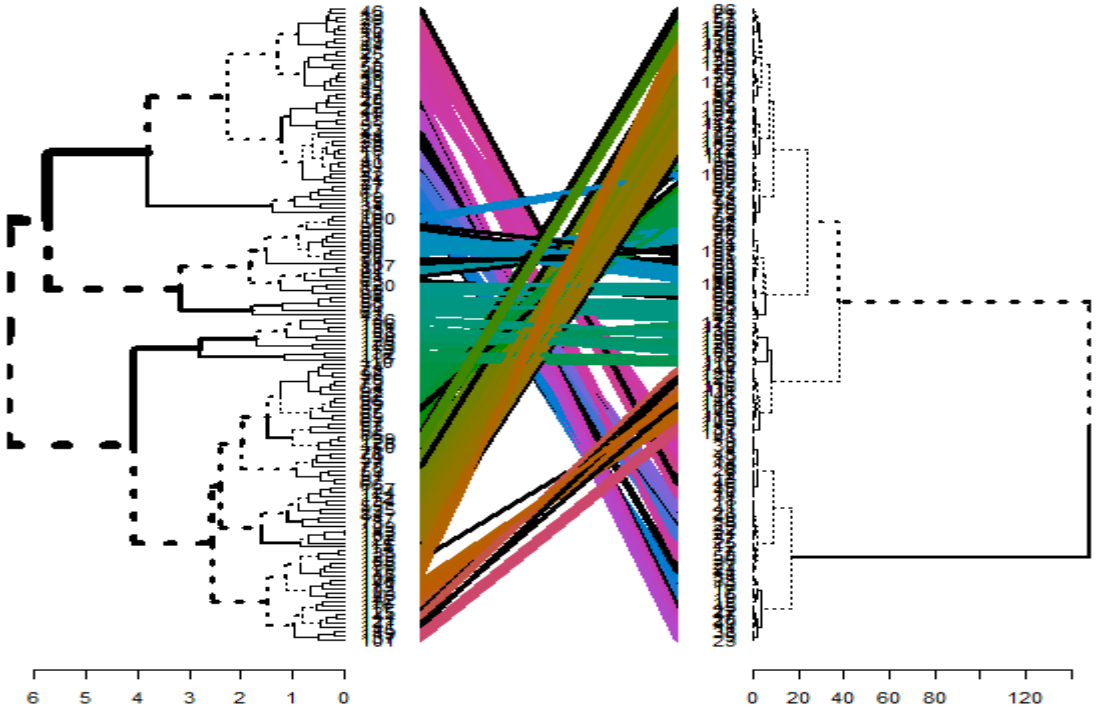
```
# Compute distance matrix
res.dist <- dist(df, method = "euclidean")

# Compute 2 hierarchical clusterings
hc1 <- hclust(res.dist, method = "complete")
hc2 <- hclust(res.dist, method = "ward.D2")
```

```
> hc1 <- hclust(res.dist, method = "complete")
> hc1

Call:
hclust(d = res.dist, method = "complete")

Cluster method   : complete
Distance         : euclidean
Number of objects: 150

> hc2 <- hclust(res.dist, method = "ward.D")
> hc2

Call:
hclust(d = res.dist, method = "ward.D")

Cluster method   : ward.D
Distance         : euclidean
Number of objects: 150
```

```
# Create two dendrograms
dend1 <- as.dendrogram (hc1)
dend2 <- as.dendrogram (hc2)
```

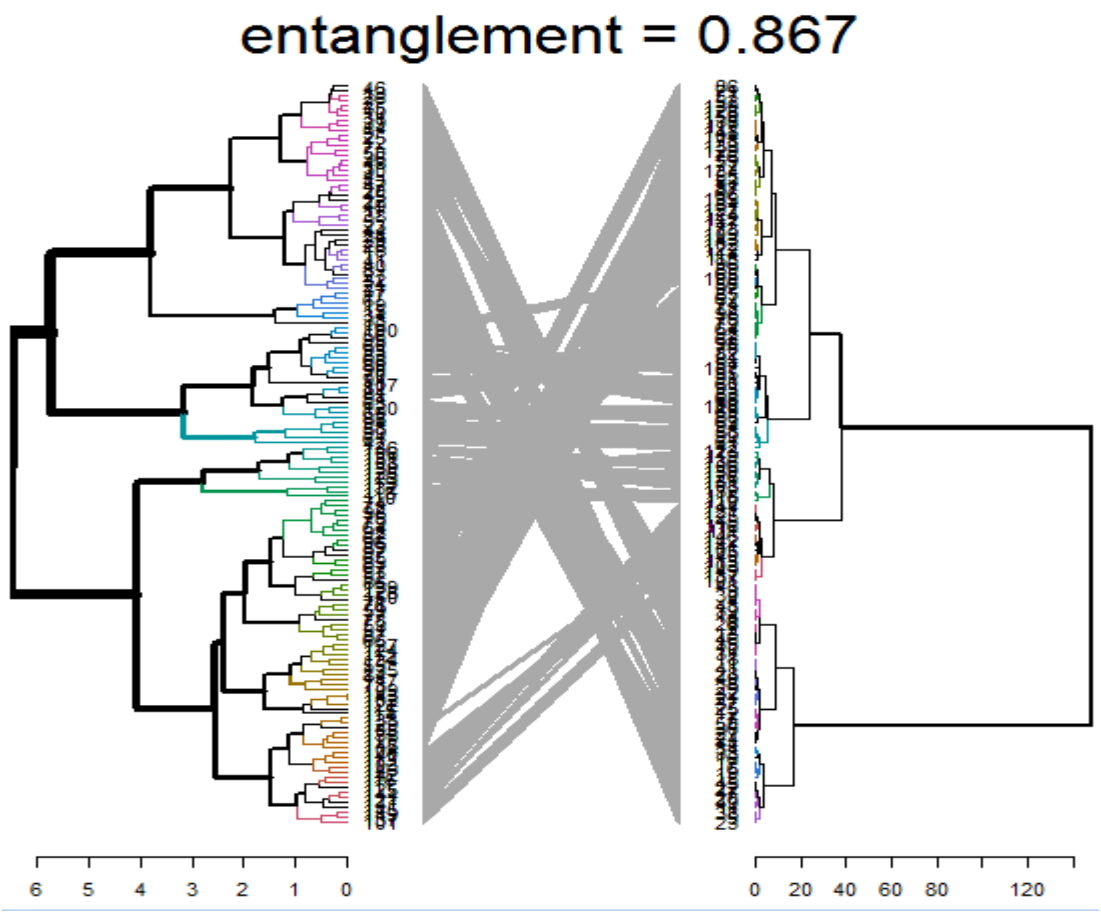
```
> dend1 <- as.dendrogram (hc1)
> dend1
'dendrogram' with 2 branches and 150 members total, at height 6.507523
> dend2 <- as.dendrogram (hc2)
> dend2
'dendrogram' with 2 branches and 150 members total, at height 148.0957
```

```
library(dendextend)
tanglegram(dend1, dend2)
```



يعرض الخرج العقد "الفريدة" ، مع مجموعة من الاسماء / العناصر غير الموجودة في الشجرة الأخرى ، مع إبراز الخطوط المتقطعة. يمكن قياس جودة محاذاة الشجرتين باستخدام تابعentanglement التشابك هو مقياس بين 1 (تشابك كامل) و 0 (بدون تشابك). معامل التشابك المنخفض يتوافق مع محاذاة جيدة. يمكن تخصيص خرج tanglegram باستخدام العديد من الخيارات الأخرى على النحو التالي:

```
dend_list <- dendlist(dend1, dend2)
> dend_list <- dendlist(dend1, dend2)
> dend_list
[[1]]
'dendrogram' with 2 branches and 150 members total, at height 6.507523

[[2]]
'dendrogram' with 2 branches and 150 members total, at height 148.0957

attr(,"class")
[1] "dendlist"
tanglegram(dend1, dend2,
  highlight_distinct_edges = FALSE, # Turn-off dashed
lines
  common_subtrees_color_lines = FALSE, # Turn-off line
colors
  common_subtrees_color_branches = TRUE, # Color common
branches
  main = paste("entanglement =",
round(entanglement(dend_list), 3))
  )
```


entanglement = 0.867
6 5 4 3 2 1 0
0 20 40 60 80 120