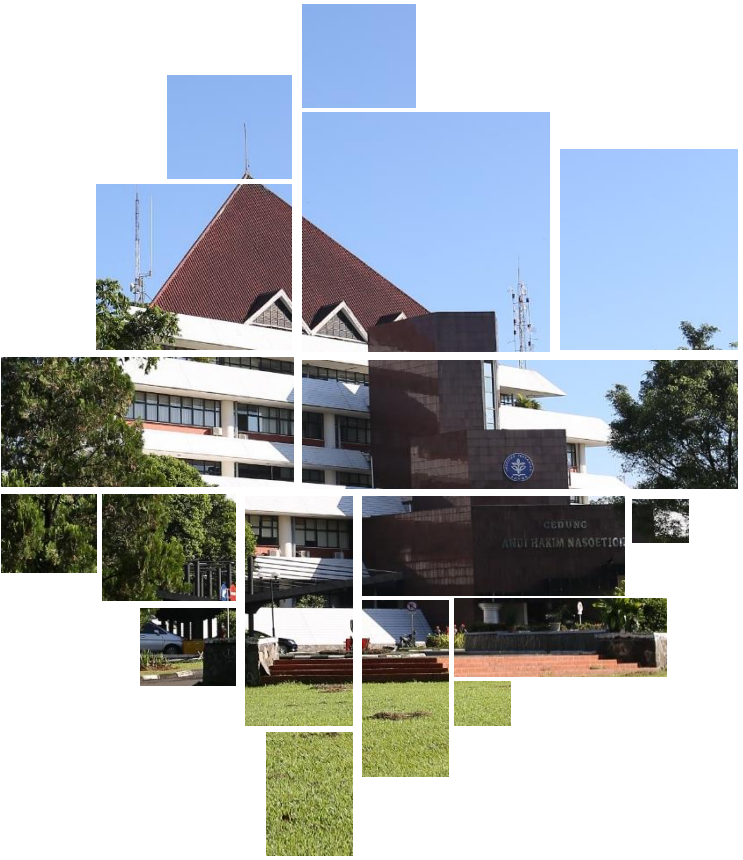
IPB University
Bogor Indonesia

Department of Statistics
Faculty of Mathematics and Natural Sciences



STA1382 Teknik Pembelajaran Mesin

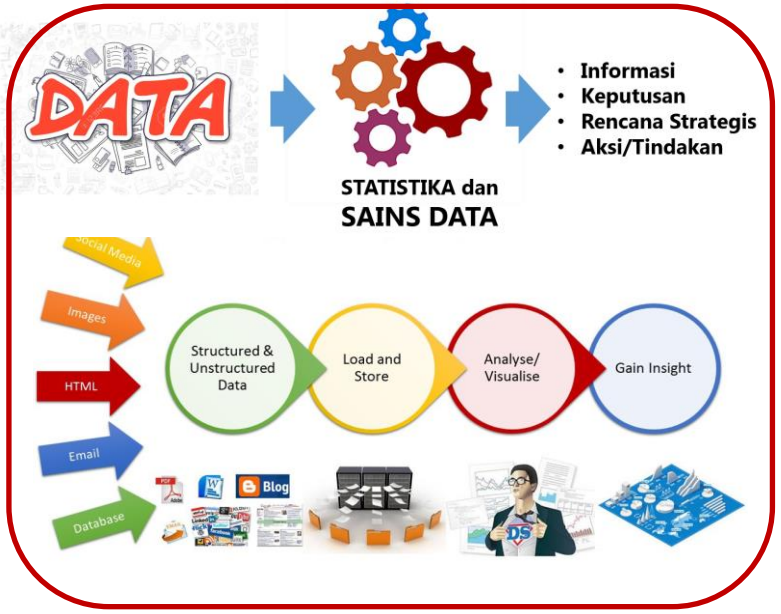
9. Penggerombolan Berhierarki dan Evaluasi

Anang Kurnia
Departemen Statistika, FMIPA - IPB
anangk[at]apps.ipb.c.id

Rencana Perkuliahan:

No	Materi
1	Penyiapan Data
2	Analisis Regresi Beserta Evaluasinya
3	Classification Tree CART (1)
4	Classification Tree CART (2)
5	Artificial Neural Network
6	Support Vector Machine (SVM)
7	Studi Kasus (Presentasi)
UTS	

No	Materi
8	Ruang Lingkup Pembelajaran Mesin Statistika
9	Penggerombolan Berhierarki dan Evaluasi
10	Penggerombolan Non-hierarki dan Evaluasi
11	Reduksi Dimensi
12	Association Rule
13	Metode Ensemble
14	Studi Kasus (Makalah)
UAS	



Process of Data Analysis

IPB University
Bogor Indonesia

Define the Problem

Identify the problems that needs to be solved through data analysis

Gather the Data

Collect data from related sources or from the Internet

Ex:- Excel files, databases, web scraping etc.

Clean & Prepare the data

Format the data correctly and make sure it's free of errors

Ex:- Cleaning Data in Excel, Python or R.

Explore the Data

Look for patterns and Trends

Ex:- using Pandas/ Numpy in python, dplyr/tidyr in R

Build models

Create models that can be used to predict future behavior or make decisions.

Ex: Linear/Logistic regression models, Clustering, K-NN etc.

Evaluate the models

Check the accuracy and performance of the model.

Methods:- F1 Score, R(squared) score, MSE(mean squared error) etc.

Communicate the results

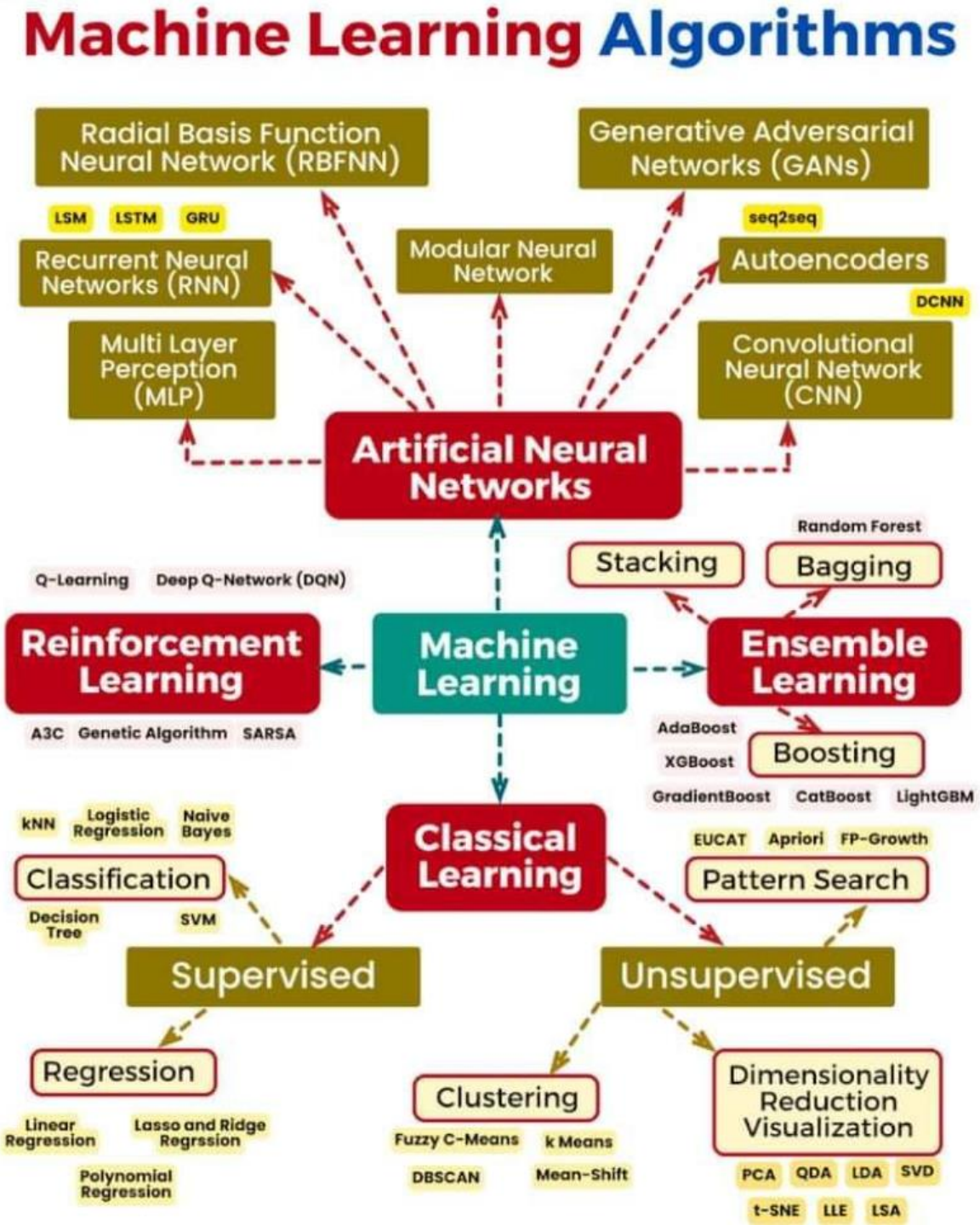
Share your findings with stakeholders.

Ex:- PowerBI/Tableau Dashboards

Build models

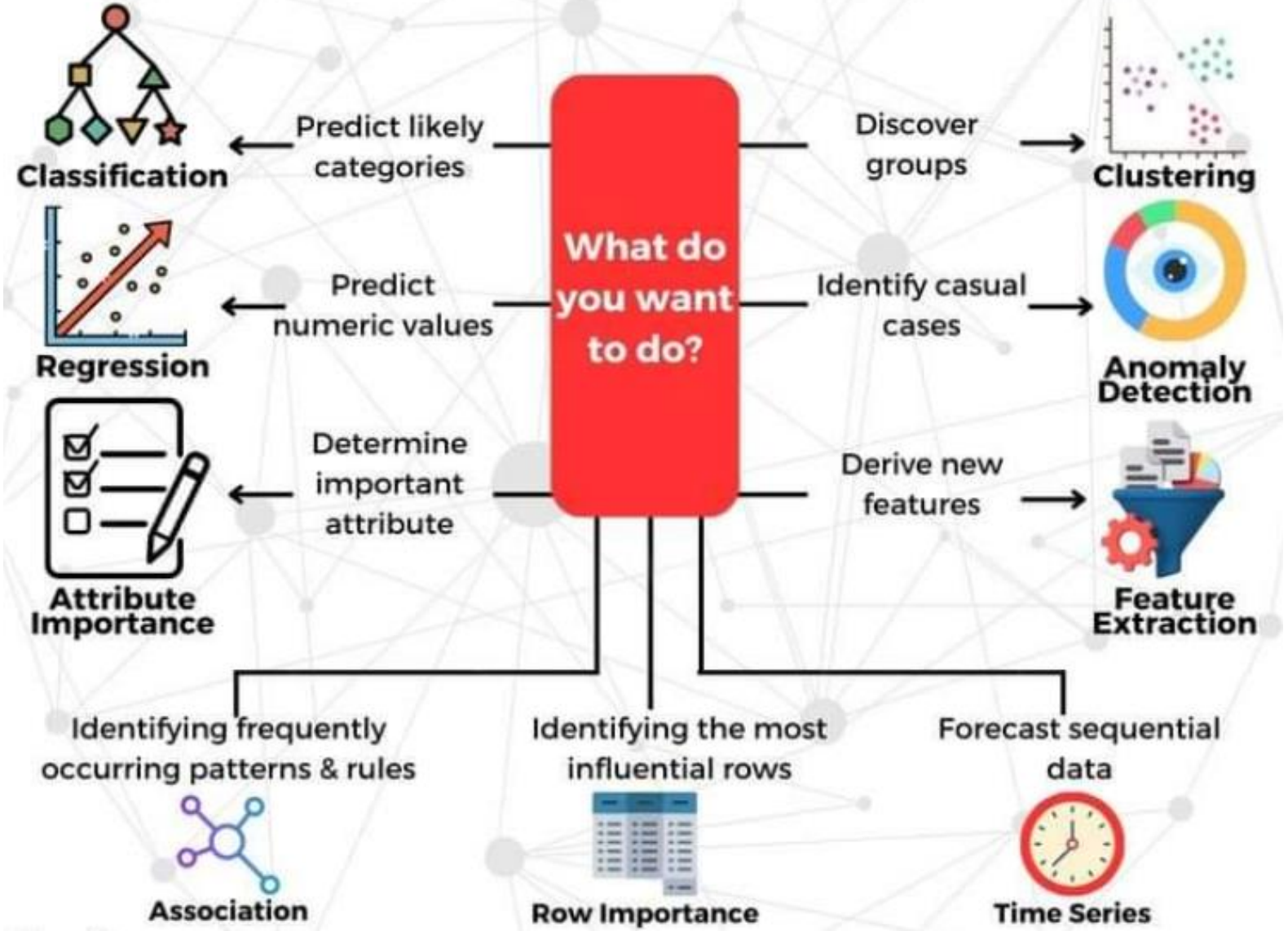
Create models that can be used to predict future behavior or make decisions.

Ex: Linear/Logistic regression models, Clustering, K-NN etc.



MACHINE LEARNING

ALGORITHM USAGE



Analisis Gerombol Berhierarki

- Salah satu kelemahan dari penggerombolan K-means adalah bahwa kita harus menentukan terlebih dahulu banyaknya gerombol K.

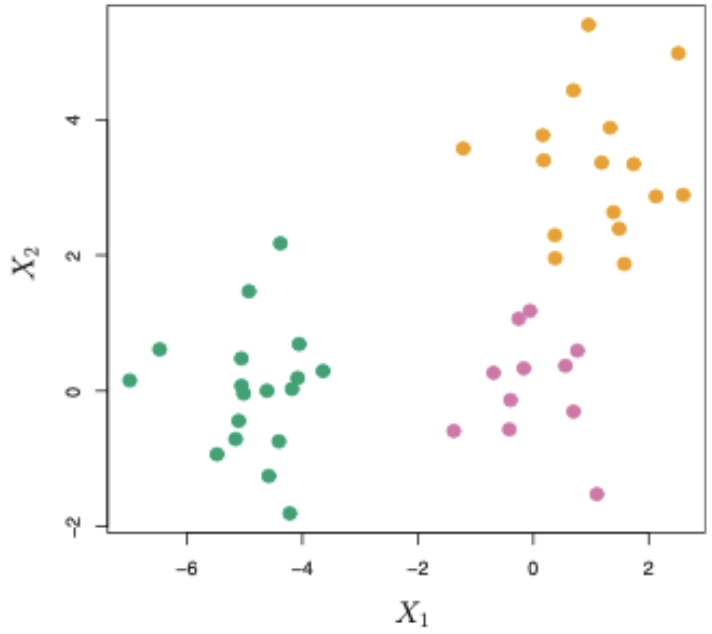
- Penggerombolan berhierarki adalah pendekatan alternatif yang tidak mengharuskan kita berkomitmen pada pilihan K tertentu.

- Penggerombolan berhierarki memiliki nilai tambah keunggulan dibandingkan penggerombolan K-means karena menghasilkan representasi pengamatan berbasis pohon yang menarik, yang disebut dendrogram.

- Pada bagian ini, akan dijelaskan penggerombolan bottom-up atau agglomerative.

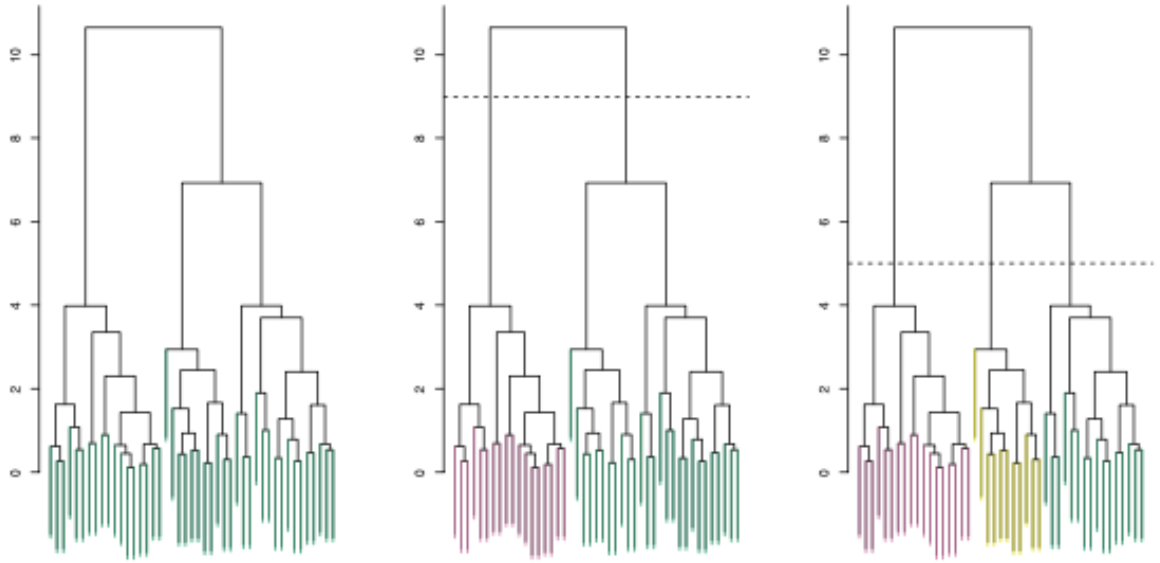
- Ini adalah jenis penggerombolan berhierarki yang paling umum, dan mengacu pada fakta bahwa dendrogram dibangun mulai dari daun dan menggabungkan kelompok hingga ke batang.

- Kita akan mulai dengan diskusi tentang bagaimana menginterpretasikan dendrogram dan kemudian mendiskusikan bagaimana sebenarnya pengelompokan hierarki dilakukan—yaitu, bagaimana dendrogram dibangun.



Data simulasi

Metode complete linkage dan Euclidean distance.



ketinggian potongan ke dendrogram memiliki peran yang sama dengan K dalam penggerombolan K-means: ia mengontrol jumlah gerombol yang diperoleh.

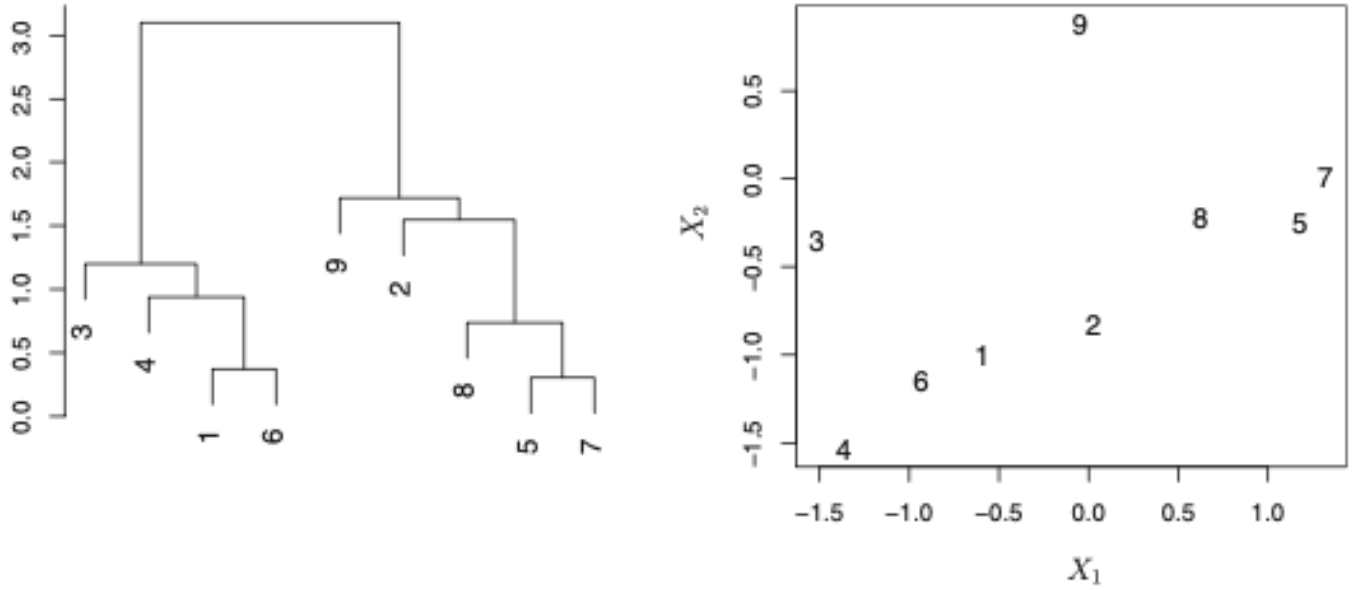


FIGURE 10.10. *An illustration of how to properly interpret a dendrogram with nine observations in two-dimensional space.* Left: *a dendrogram generated using Euclidean distance and complete linkage. Observations* 5 *and* 7 *are quite similar to each other, as are observations* 1 *and* 6. *However, observation* 9 *is* no more similar to *observation* 2 *than it is to observations* 8, 5, *and* 7, *even though observations* 9 *and* 2 *are close together in terms of horizontal distance. This is because observations* 2, 8, 5, *and* 7 *all fuse with observation* 9 *at the same height, approximately* 1.8. Right: *the raw data used to generate the dendrogram can be used to confirm that indeed, observation* 9 *is no more similar to observation* 2 *than it is to observations* 8, 5, *and* 7.

Catatan:

- Dalam praktiknya, orang sering melihat dendrogram dan memilih sejumlah gerombol yang masuk akal, berdasarkan ketinggian penggabungan (fusion) dan jumlah gerombol yang diinginkan.

- Namun, seringkali pilihan tempat pemotongan dendrogram tidak begitu jelas.

- Istilah hierarki mengacu pada fakta bahwa gerombol yang diperoleh dengan memotong dendrogram pada ketinggian tertentu harus bersarang di dalam gerombol yang diperoleh dengan memotong dendrogram pada ketinggian yang lebih tinggi. Namun, pada kumpulan data secara umum, asumsi struktur hierarki ini mungkin tidak realistis.

- Misalnya, anggaplah bahwa pengamatan kita berhubungan dengan sekelompok orang dengan pembagian pria dan wanita 50–50, terbagi rata antara orang Amerika, Jepang, dan Prancis. Kita dapat membayangkan sebuah skenario di mana pembagian terbaik menjadi dua gerombol mungkin membagi orang-orang ini berdasarkan jenis kelamin, dan pembagian terbaik menjadi tiga gerombol mungkin memisahkan mereka berdasarkan kebangsaan. Dalam hal ini, cluster yang sebenarnya tidak bersarang, dalam arti bahwa pembagian terbaik menjadi tiga gerombol tidak dihasilkan dari mengambil pembagian terbaik menjadi dua gerombol dan memisahkan salah satu dari gerombol tersebut.

- Akibatnya, situasi ini tidak dapat terwakili dengan baik oleh penggerombolan berhierarki. Karena situasi seperti ini, penggerombolan berhierarki terkadang dapat menghasilkan hasil yang lebih buruk (kurang akurat) daripada penggerombolan K-means untuk sejumlah kelompok tertentu.

Algoritma Penggerombolan Berhierarki

- Dendrogram penggerombolan berhierarki diperoleh melalui algoritma yang sangat sederhana.

- Kita mulai dengan mendefinisikan semacam ukuran perbedaan (dissimilarity) antara masing-masing pasangan pengamatan. Paling sering, jarak Euclidean digunakan.

- Algoritma berjalan secara iteratif. Dimulai dari bagian bawah dendrogram, masing-masing n pengamatan diperlakukan sebagai gerombol tersendiri. Dua gerombol yang paling mirip satu sama lain kemudian digabungkan sehingga menjadi $n-1$ gerombol. Selanjutnya dua gerombol yang paling mirip satu sama lain dilebur kembali, sehingga menjadi $n-2$ gerombol. Algoritma berjalan dengan cara ini sampai semua pengamatan termasuk dalam satu gerombol tunggal, dan dendrogram selesai.

- Konsep perbedaan (dissimilarity) antara sepasang pengamatan perlu diperluas menjadi sepasang kelompok pengamatan. Perluasan ini dicapai dengan mengembangkan gagasan keterkaitan (linkage), yang mendefinisikan perbedaan (dissimilarity) antara dua kelompok pengamatan. Umumnya, jenis linkage: complete, average, single, and centroid

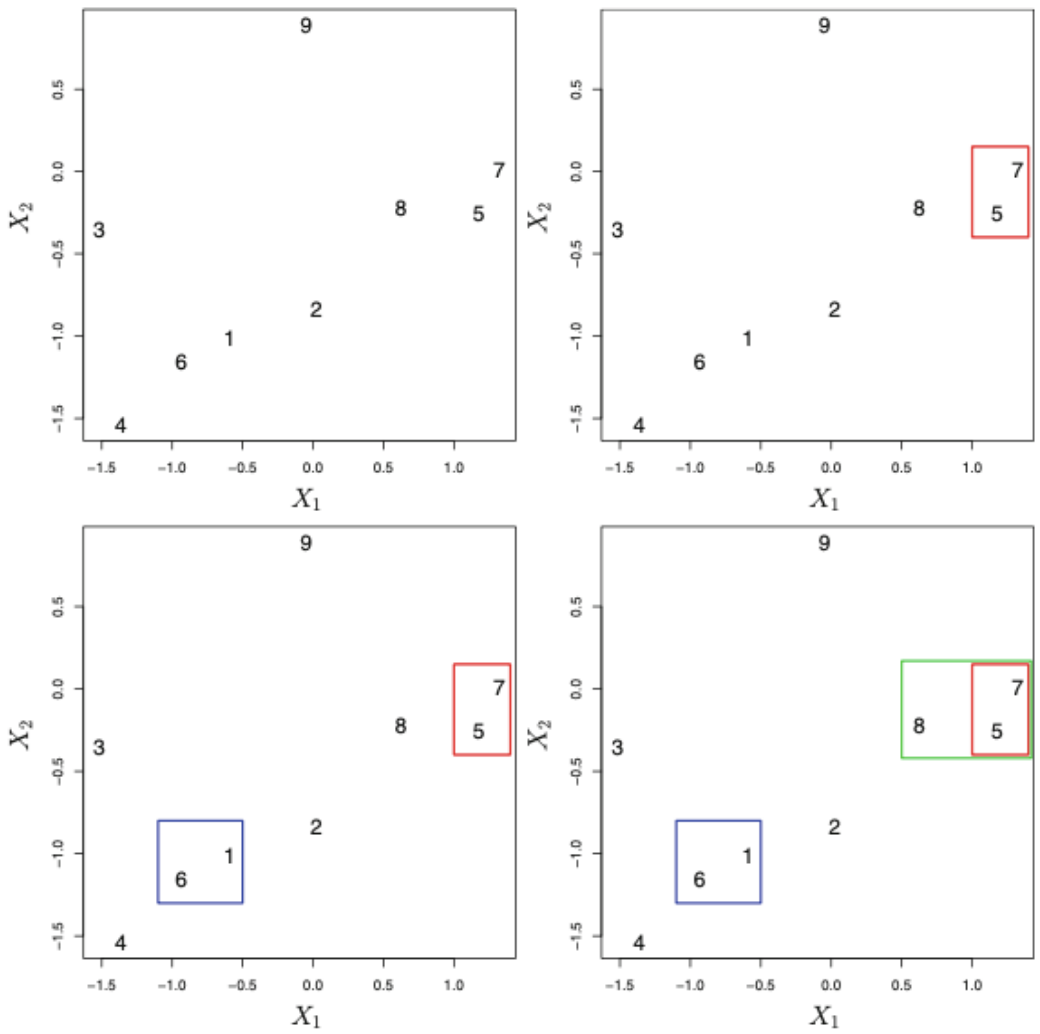
Algorithm 10.2 *Hierarchical Clustering*

1. Begin with n observations and a measure (such as Euclidean distance) of all the $\binom{n}{2} = n(n-1)/2$ pairwise dissimilarities. Treat each observation as its own cluster.

2. For $i = n, n-1, \ldots, 2$:

   (a) Examine all pairwise inter-cluster dissimilarities among the i clusters and identify the pair of clusters that are least dissimilar (that is, most similar). Fuse these two clusters. The dissimilarity between these two clusters indicates the height in the dendrogram at which the fusion should be placed.

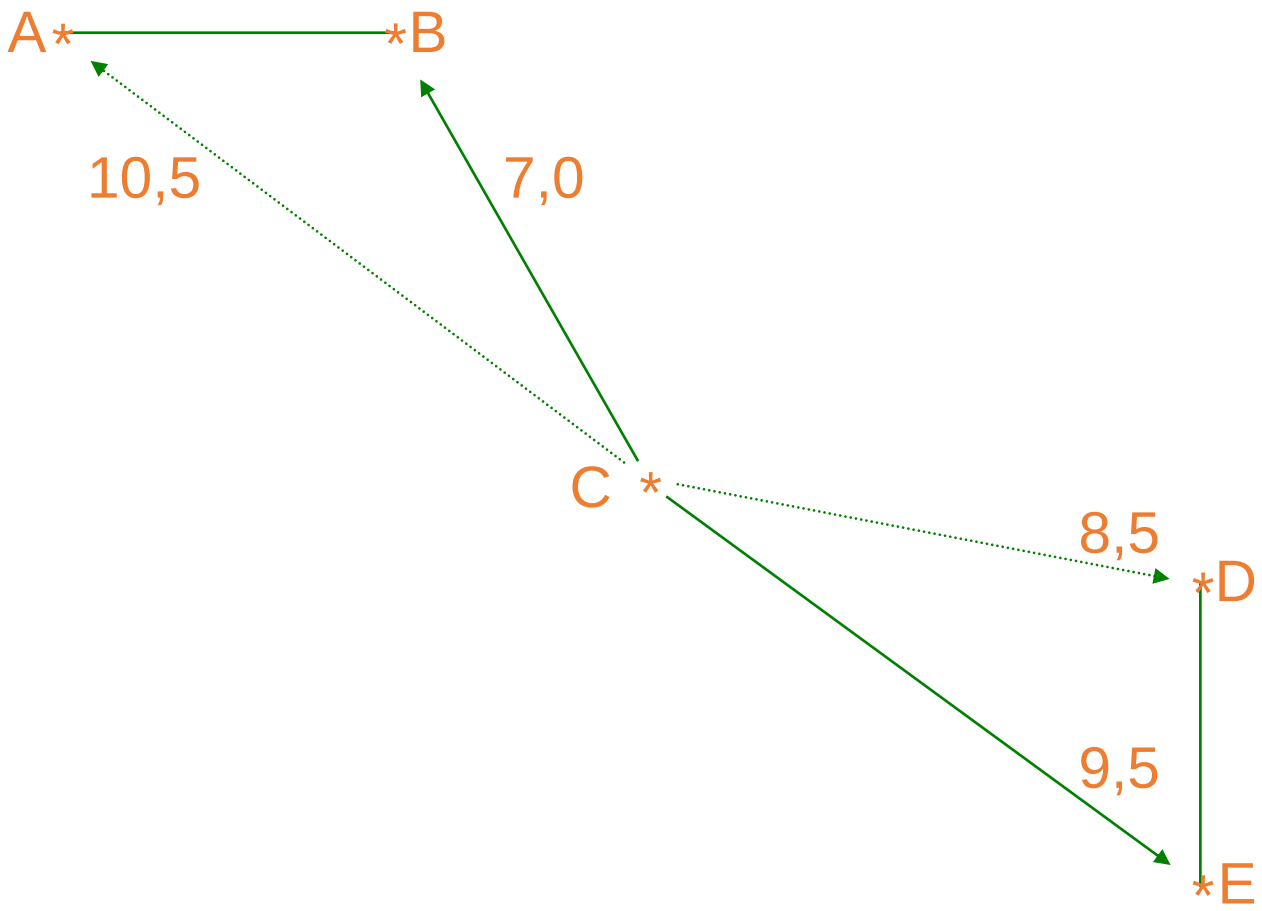
   (b) Compute the new pairwise inter-cluster dissimilarities among the $i-1$ remaining clusters.



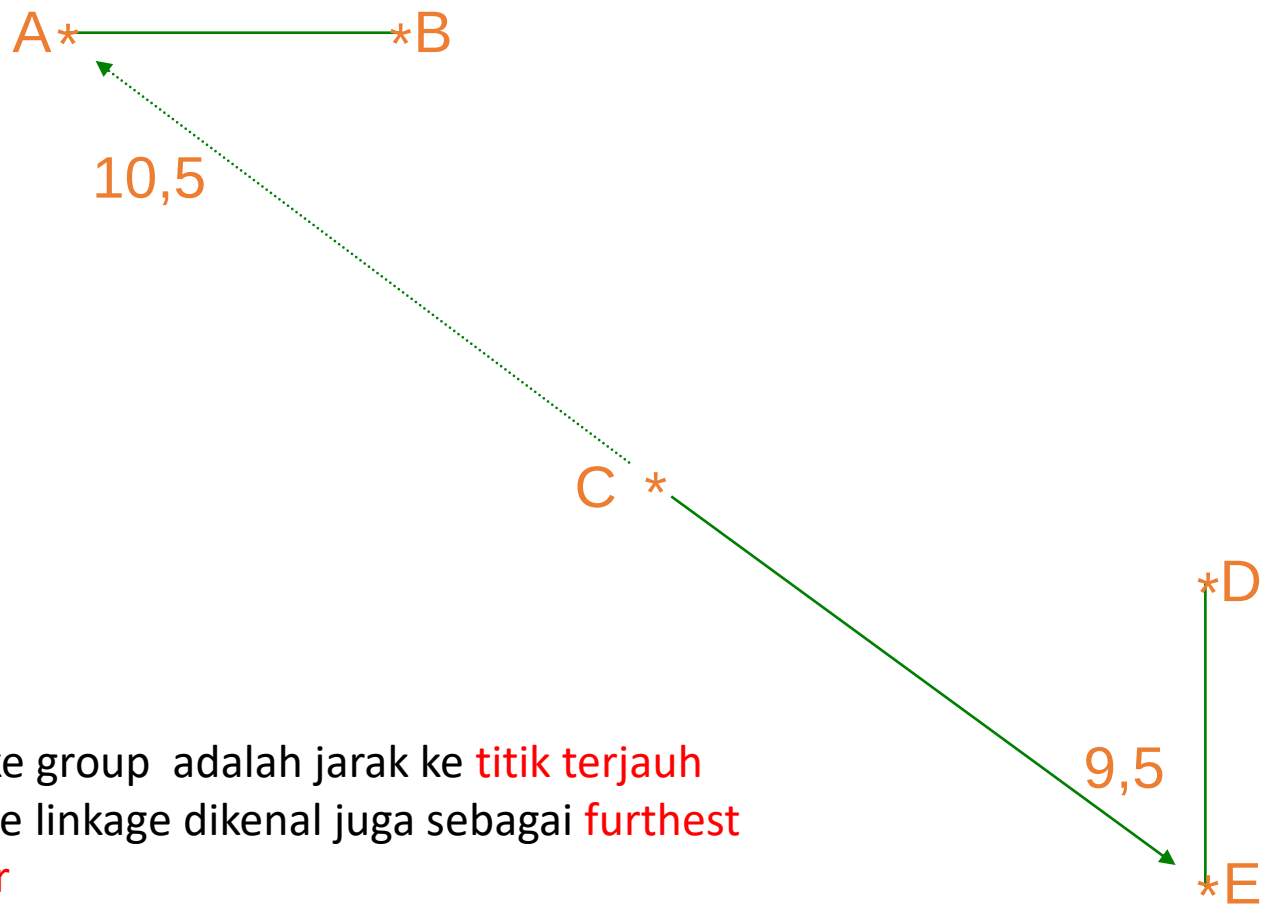
Complete linkage and Euclidean distance

Linkage	*Description*	
Complete	Maximal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *largest* of these dissimilarities.	$d(i \cup j, k) = \max\{d(i,k), d(j,k)\}$
Single	Minimal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *smallest* of these dissimilarities. Single linkage can result in extended, trailing clusters in which single observations are fused one-at-a-time.	$d(i \cup j, k) = \min\{d(i,k), d(j,k)\}$
Average	Mean intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *average* of these dissimilarities.	$d(i \cup j, k) = (1/2)\{d(i,k) + d(j,k)\}$
Centroid	Dissimilarity between the centroid for cluster A (a mean vector of length p) and the centroid for cluster B. Centroid linkage can result in undesirable *inversions*.	$d(i \cup j, k) = \{\bar{x}_{ij}, \bar{y}_k\}$

Linkage Method?

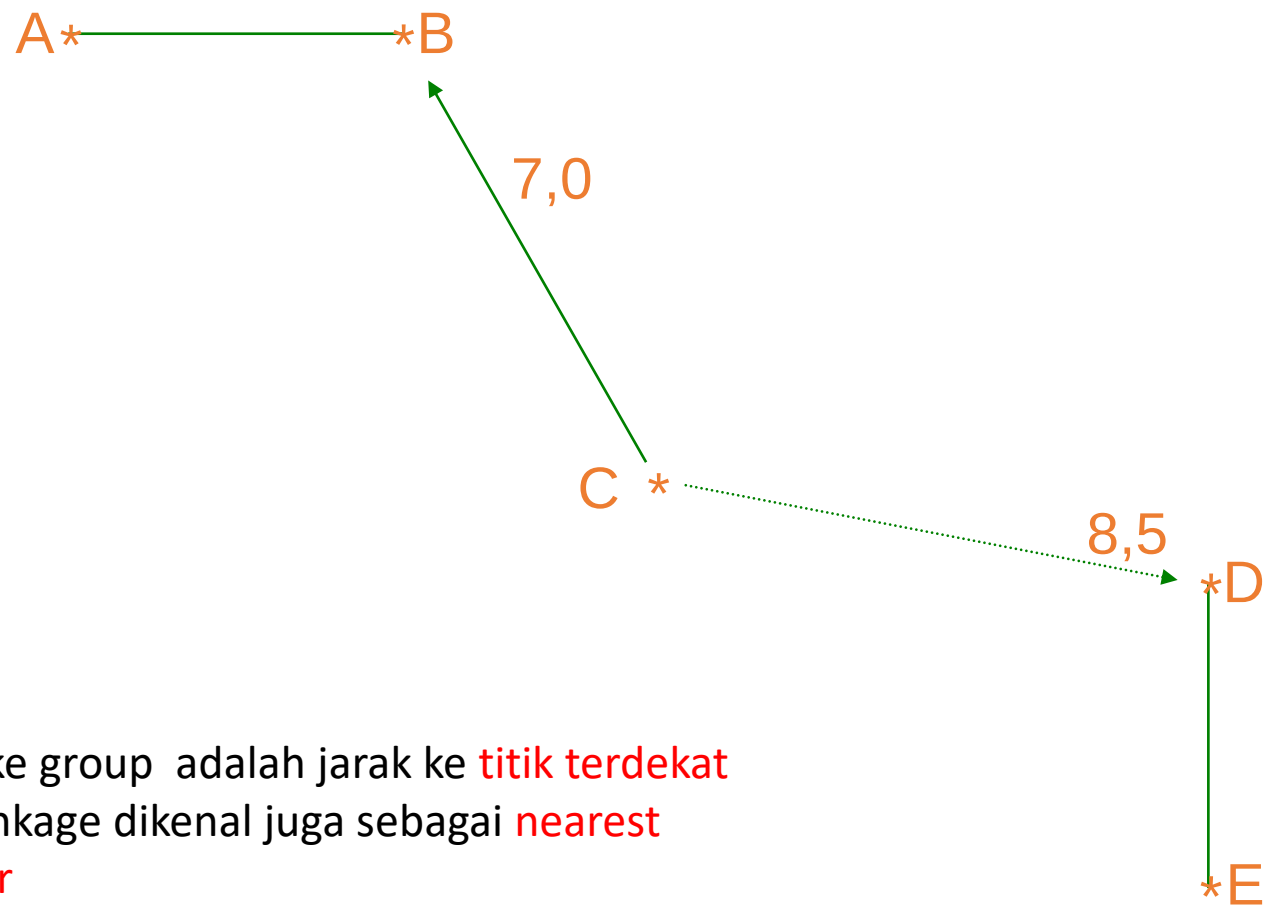


Complete linkage



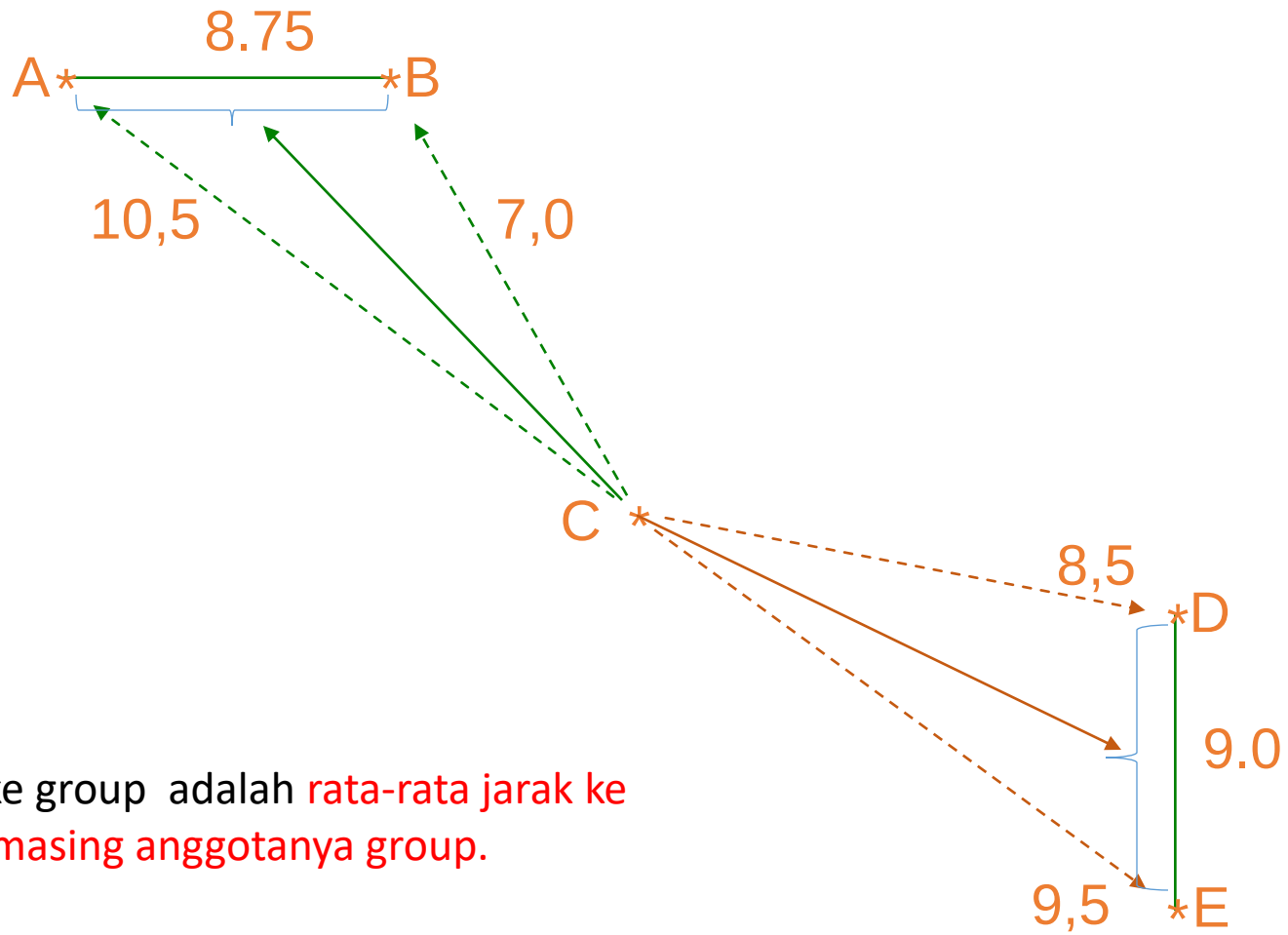
- Jarak C ke group adalah jarak ke titik terjauh
- Complete linkage dikenal juga sebagai furthest neighbor

Single linkage



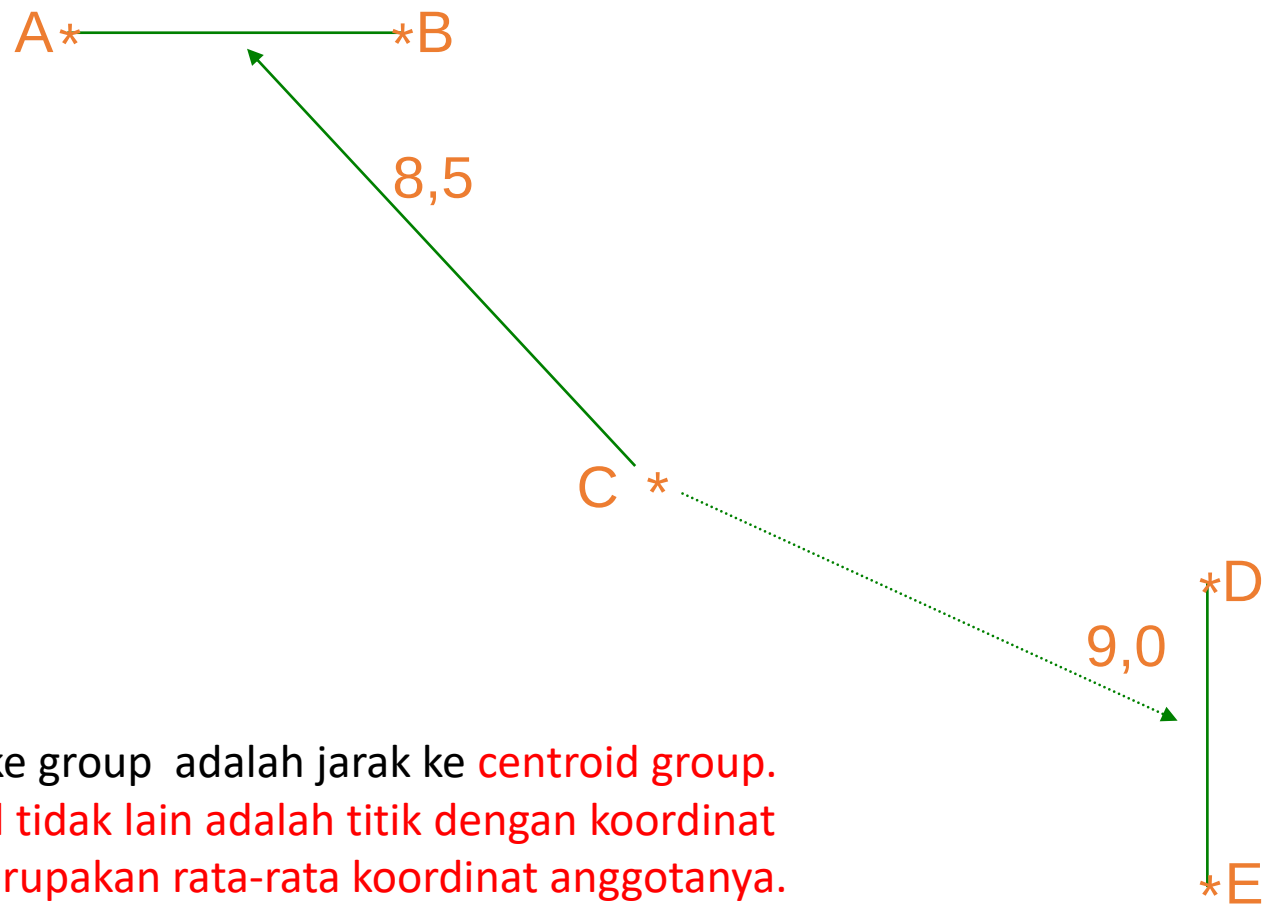
- Jarak C ke group adalah jarak ke titik terdekat
- Single linkage dikenal juga sebagai nearest neighbor

Average linkage



- Jarak C ke group adalah rata-rata jarak ke masing-masing anggotanya group.

Centroid linkage



- Jarak C ke group adalah jarak ke centroid group. Centroid tidak lain adalah titik dengan koordinat yang merupakan rata-rata koordinat anggotanya.

Pemilihan Ukuran dissimilarity

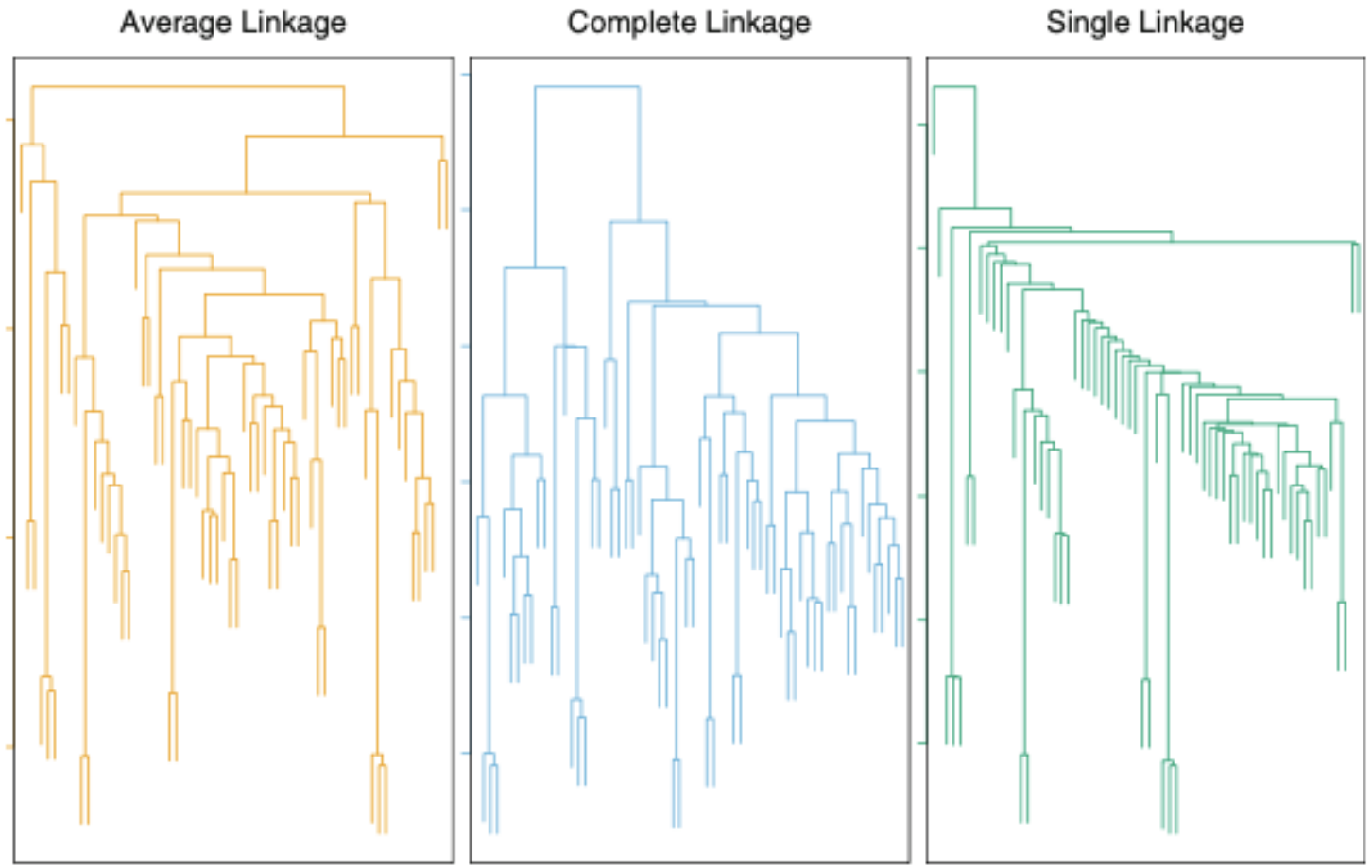


FIGURE 10.12. *Average, complete, and single linkage applied to an example data set. Average and complete linkage tend to yield more balanced clusters.*

- Pemilihan ukuran dissimilarity sangat penting, karena memiliki efek yang kuat pada dendrogram yang dihasilkan.
- Secara umum, perhatian yang cermat harus diberikan pada jenis data yang digerombolkan dan pertanyaan ilmiah yang ada.
- Pertimbangan ini harus menentukan jenis ukuran ketidaksamaan (dissimilarity) apa yang digunakan untuk penggerombolan berhierarki.

Isu Praktis dalam Analisis Gerombol

- Keputusan Kecil dengan Konsekuensi Besar
  - Haruskah pengamatan atau peubah perlu distandarisasi?
  - Pada kasus penggerombolan berhierarki: ukuran dissimilarity apa yang harus digunakan? Jenis linkage apa yang digunakan? Dimana kita harus memotong dendogram untuk memperoleh sejumlah gerombol?
  - Pada kasus penggerombolan K-means, berapa banyak gerombol yang akan kita tentukan pada data?

- Memvalidasi Gerombol yang diperoleh
  - Kita benar-benar ingin tahu apakah gerombol yang ditemukan mewakili subgrup sebenarnya dalam data, atau apakah itu hanya hasil dari penggerombolan noise.

- Pertimbangan Lain dalam Penggerombolan
  - metode penggerombolan umumnya tidak terlalu kuat terhadap gangguan pada data (hasil gerombol bisa sangat berbeda dengan data sama namun banyaknya berbeda)

- Menafsirkan Hasil Penggerombolan
  - Yang terpenting, kita harus berhati-hati tentang bagaimana hasil analisis penggerombolan diinterpretasikan. Hasil ini tidak boleh dianggap sebagai kebenaran mutlak tentang kumpulan data. Sebaliknya, mereka harus merupakan titik awal untuk pengembangan hipotesis ilmiah dan studi lebih lanjut.

Penggunaan R

- Diketahui data Mall_Customers.csv yang ingin dikelompokkan berdasarkan peubah Age, Annual.Income, dan Spending.Score, untuk dapat memberikan gambaran strategi marketing yang baik untuk dilakukan.

- Tentukanlah gerombol customer tersebut dengan menggunakan:
  - Analisis gerombol berhierarkhi

Persiapan data

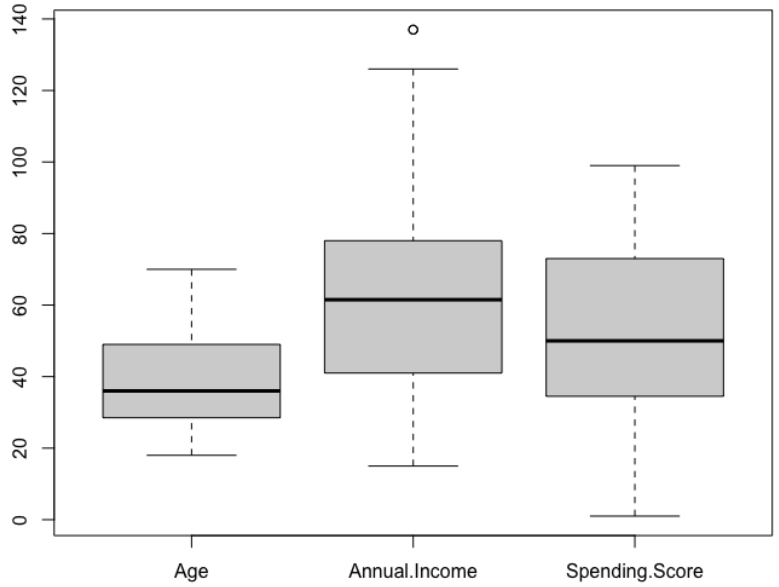
```
#persiapan data
data.mall <- read.csv("Mall_Customers.csv")
str(data.mall)
head(data.mall)
```

```
> data.mall <- read.csv("Mall_Customers.csv")
> str(data.mall)
'data.frame':   200 obs. of  5 variables:
 $ CustomerID    : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Genre         : chr  "Male" "Male" "Female" "Female" ...
 $ Age           : int  19 21 20 23 31 22 35 23 64 30 ...
 $ Annual.Income : int  15 15 16 16 17 17 18 18 19 19 ...
 $ Spending.Score: int  39 81 6 77 40 76 6 94 3 72 ...
> head(data.mall)
  CustomerID  Genre Age Annual.Income Spending.Score
1          1   Male  19             15              39
2          2   Male  21             15              81
3          3 Female  20             16               6
4          4 Female  23             16              77
5          5 Female  31             17              40
6          6 Female  22             17              76
```

```
#data yang digunakan
data.mall.OK <- data.mall[,3:5]
str(data.mall.OK)
```

```
> data.mall.OK <- data.mall[,3:5]
> str(data.mall.OK)
'data.frame':   200 obs. of  3 variables:
 $ Age            : int  19 21 20 23 31 22 35 23 64 30 ...
 $ Annual.Income : int  15 15 16 16 17 17 18 18 19 19 ...
 $ Spending.Score: int  39 81 6 77 40 76 6 94 3 72 ...
```

```
boxplot(data.mall.OK)
```



```
#standarisasi peubah
data.mall.stdz <- scale(data.mall.OK)
apply(data.mall.stdz,2,mean)      #rataan 0
apply(data.mall.stdz,2,sd)        #sd 1
```

```
> data.mall.stdz <- scale(data.mall.OK)
> apply(data.mall.stdz,2,mean)
           Age  Annual.Income Spending.Score
 -1.016906e-16  -8.144310e-17  -1.096708e-16
> apply(data.mall.stdz,2,sd)
           Age  Annual.Income Spending.Score
             1              1              1
```

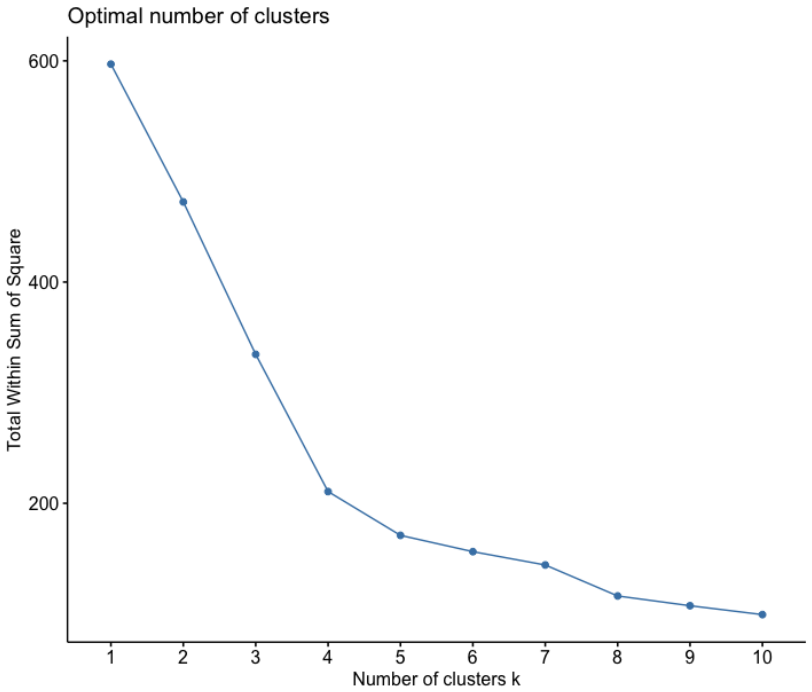
Analisis gerombol berhierarkhi

Memilih metode linkage dan banyaknya cluster

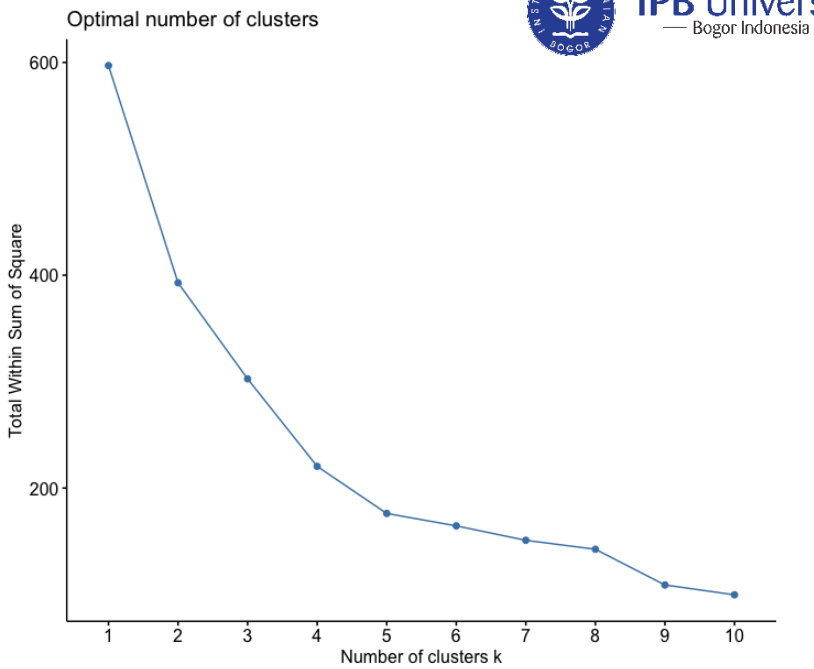
Pada kasus ini menggunakan package “factoextra”

Menggunakan kriteria wss yang sudah relative tidak berubah

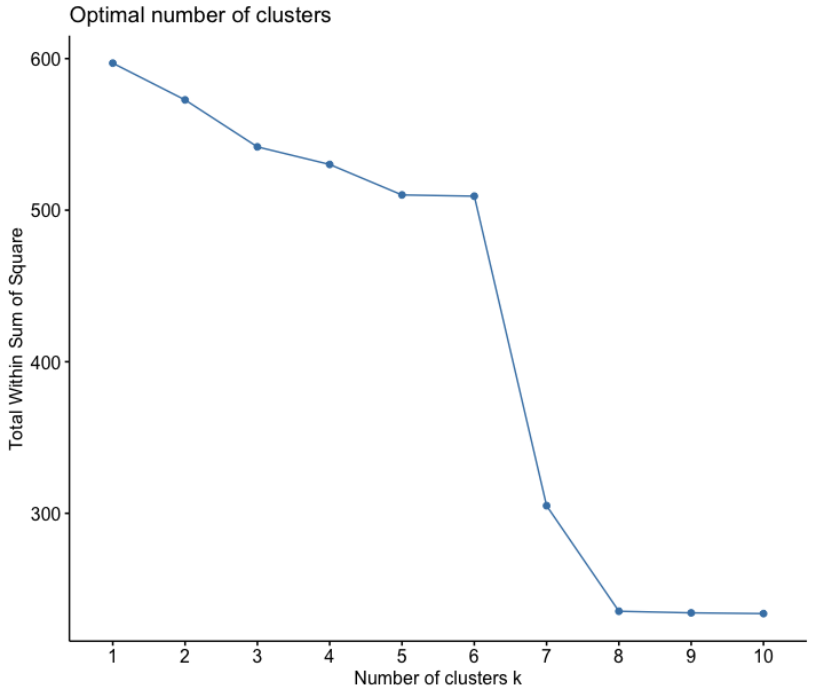
```
#Analisis gerombol berhierarkhi
##Memilih metode linkage dan banyaknya cluster
install.packages("factoextra")
library(factoextra)

###complete
a1 <- fviz_nbclust(data.mall.stdz,FUNcluster =
hcut,method = "wss",hc_method = "complete",hc_metric =
"euclidean")
a1
a1$data
```



```
a2 <- fviz_nbclust(data.mall.stdz,FUNcluster =
hcut,method = "wss",hc_method = "average",hc_metric
= "euclidean")
a2
a2$data
```




```
###centroid
a3 <- fviz_nbclust(data.mall.stdz,FUNcluster =
hcut,method = "wss",hc_method = "centroid",hc_metric
= "euclidean")
a3
a3$data
```

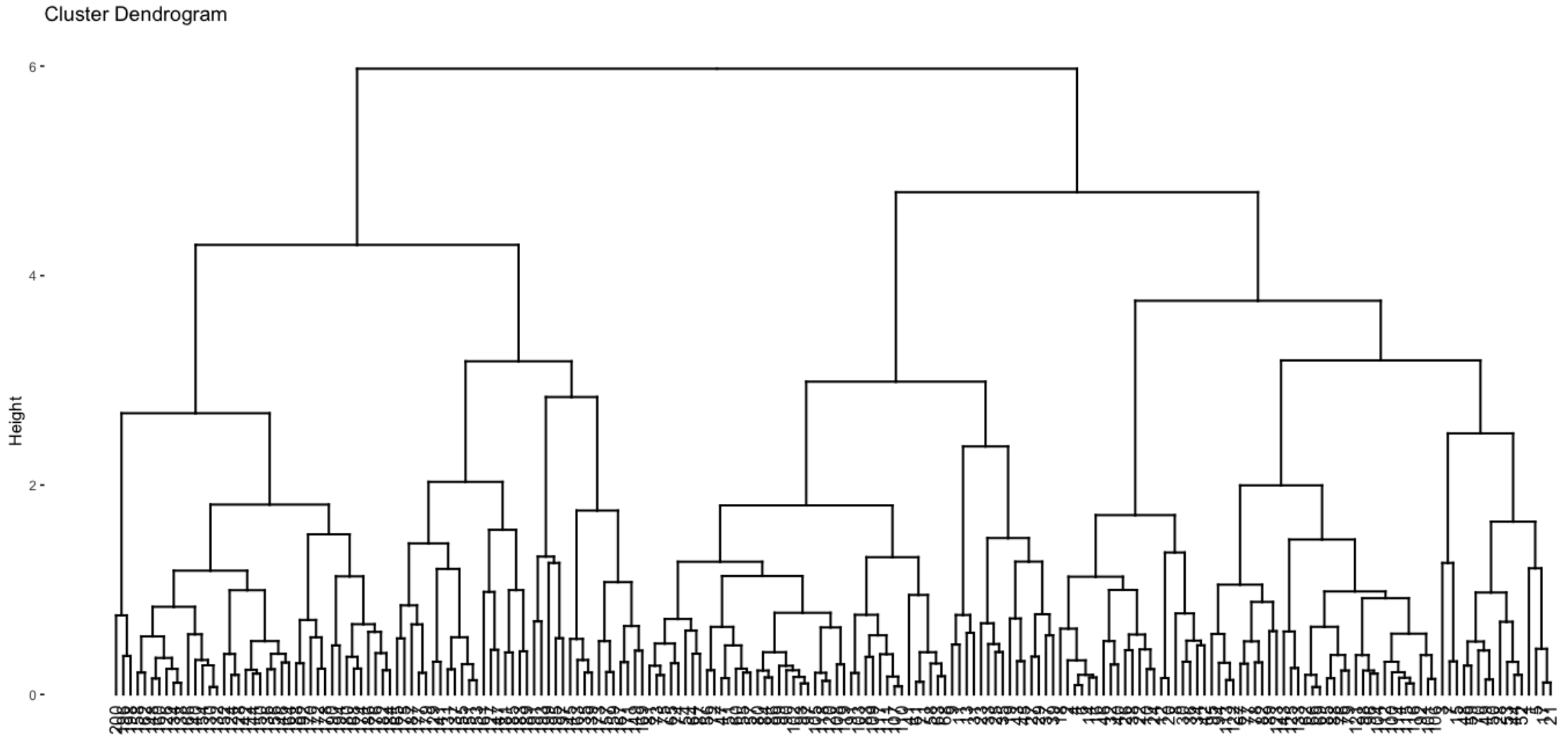


```
cbind(clusters=a1$data[,1],complete=a1$data[,2],average=a2$data[,2],
centroid=a3$data[,2])
```

```
      clusters  complete   average centroid
 [1,]        1 597.00000 597.00000 597.0000
 [2,]        2 472.44364 392.89818 572.7536
 [3,]        3 334.64043 302.62546 541.7832
 [4,]        4 210.61574 220.40815 530.1611
 [5,]        5 171.00969 176.13829 510.0553
 [6,]        6 156.28341 164.47095 509.2355
 [7,]        7 144.22360 150.90845 305.0162
 [8,]        8 116.22958 142.48203 235.3848
 [9,]        9 107.39757 108.93658 234.3236
[10,]       10  99.36778  99.67111 233.8285
```

Dendogram

```
###misalkan dipilih metode linkage = complete
fviz_dend(hclust(dist(data.mall.stdz,method = 'euclidean'),method = "complete"))
```



Interpretasi Gerombol yang terbentuk

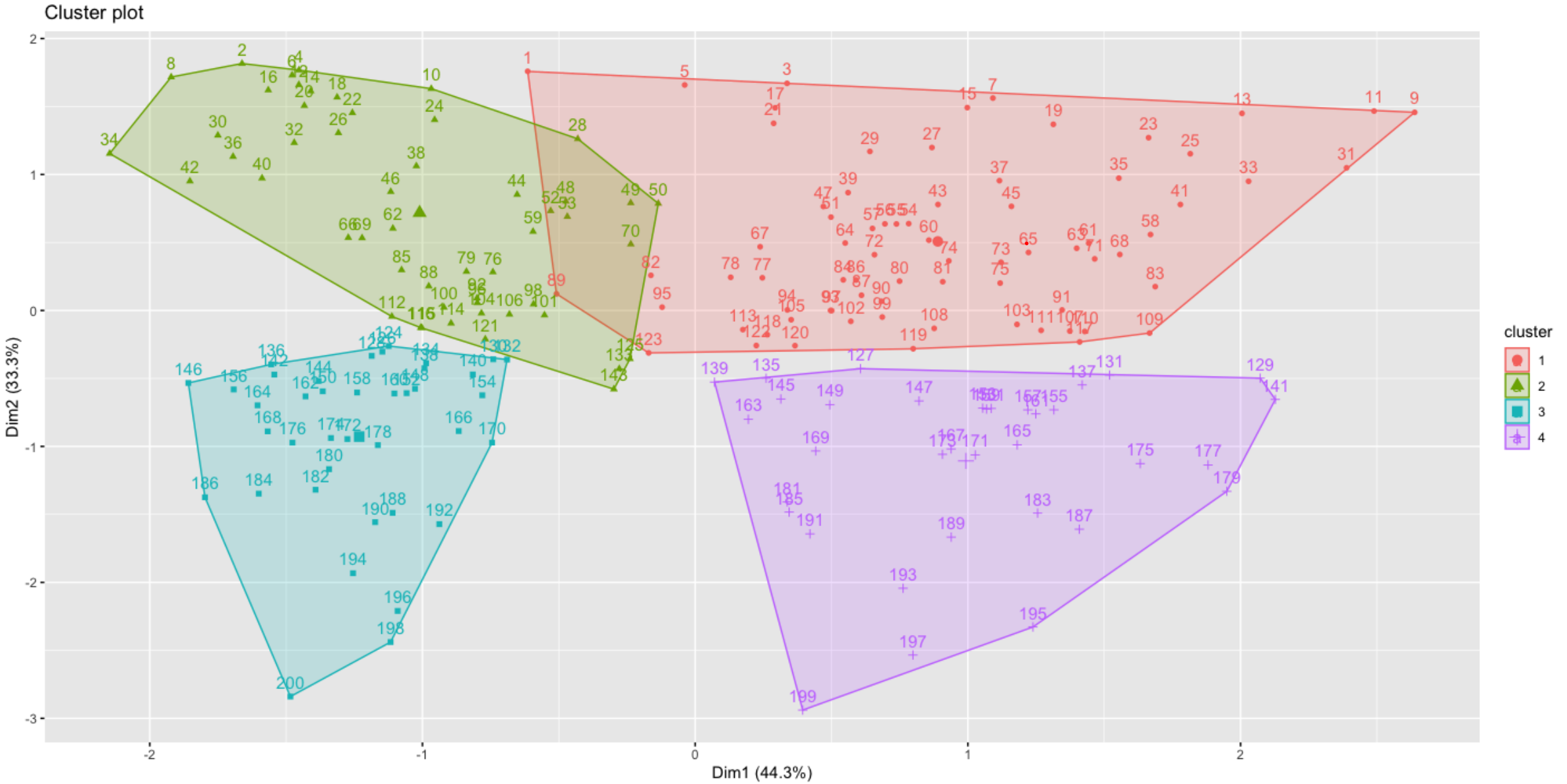
```
#interpretasi
hc.mall <- eclust(data.mall.OK,stand = TRUE,FUNcluster = "hclust",k=4,hc_method
= "average",hc_metric = "euclidean",graph = F)
hc.mall$cluster
```

```
> hc.mall$cluster
  [1] 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2
 [35] 1 2 1 2 1 2 1 2 1 2 1 2 1 2 2 2 1 2 2 1 1 1 1 1 2 1 1 2 1 1 1 2 1 1
 [69] 2 2 1 1 1 1 1 2 1 1 2 1 1 1 1 1 2 1 1 2 1 1 1 2 1 1 1 2 1 2 1 2 2 1
[103] 1 2 1 2 1 1 1 1 1 2 1 2 2 2 1 1 1 1 2 1 1 3 2 3 4 3 4 3 4 3 2 3 4 3
[137] 4 3 4 3 4 3 2 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3
[171] 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3
```

```
aggregate(data.mall.OK,by=list(cluster=hc.mall$cluster),FUN = mean)
```

	cluster	Age	Annual.Income	Spending.Score
1	1	50.72973	46.16216	40.59459
2	2	24.65385	42.94231	62.07692
3	3	32.69231	86.53846	82.12821
4	4	41.68571	88.22857	17.28571

```
#scatterplot
fviz_cluster(hc.mall)
```



Bersambung …….

IPB University

Bogor Indonesia

Inspiring Innovation with Integrity

in Agriculture, Ocean and Biosciences for a Sustainable World

anang kurnia (anangk[at]apps.ipb.ac.id)

Aplikasi di R

```
#Hierarki
hc.complete=hclust(dist(x), method="complete")
hc.average=hclust(dist(x), method="average")
hc.single=hclust(dist(x), method="single")

par(mfrow=c(1,3))
plot(hc.complete,main="Complete Linkage", xlab="", sub="", cex =.9)
plot(hc.average , main="Average Linkage", xlab="", sub="", cex =.9)
plot(hc.single , main="Single Linkage", xlab="", sub="", cex =.9)

###2 cluster
cutree(hc.complete, 2)
cutree(hc.average, 2)
cutree(hc.single, 2)
cutree(hc.single , 4)
```

```
df <- USArrests
df <- na.omit(df)
df <- scale(df)
head(df)

# Dissimilarity matrix
d <- dist(df, method = "euclidean")

# Hierarchical clustering using Complete Linkage
hc1 <- hclust(d, method = "complete" )

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

# Ward's method
hc2 <- hclust(d, method = "ward.D2" )
plot(hc2, cex = 0.6)
rect.hclust(hc2, k = 3, border = 2:5)
```