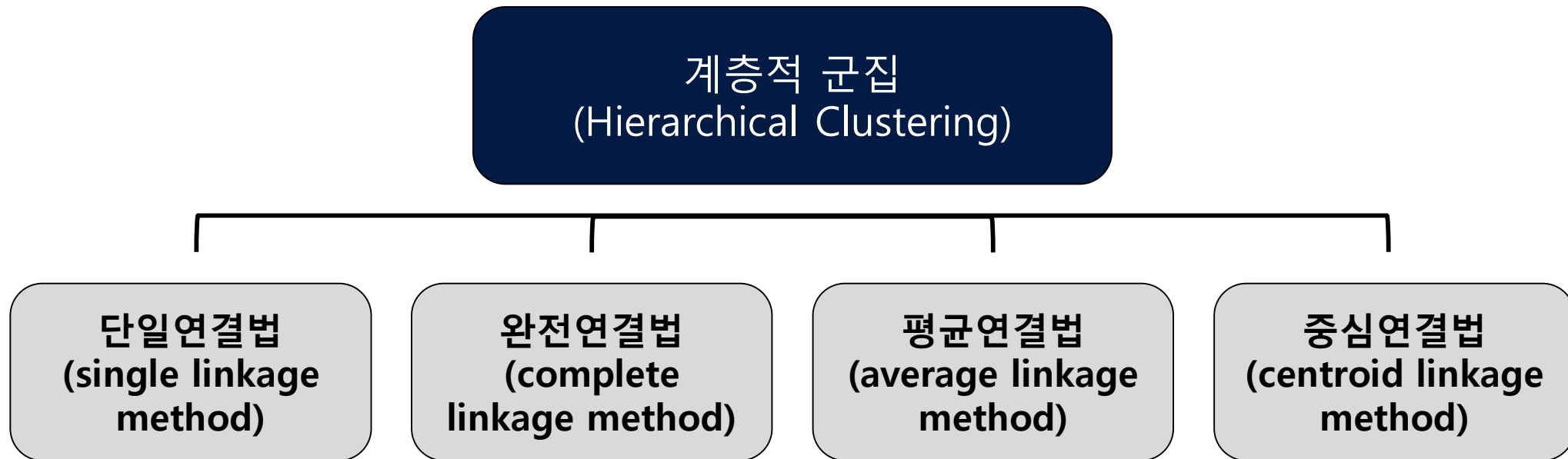

Wk13-2 : 군집분석

- 계층적 군집분석 -

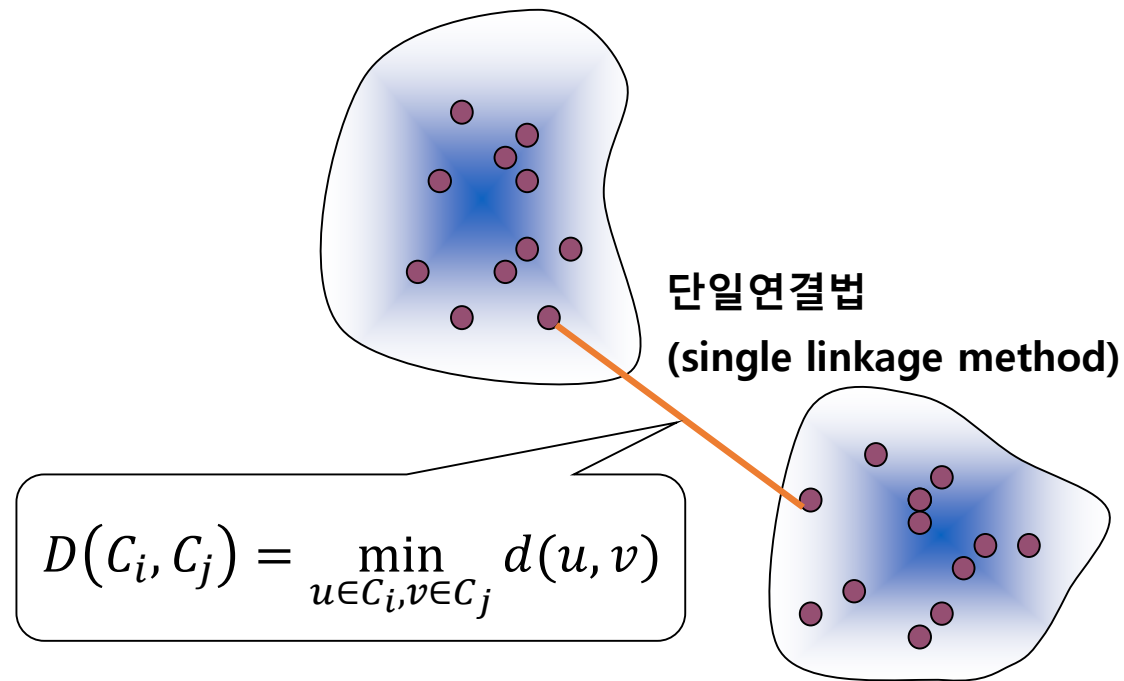
1. 계층적 군집분석

- 사전에 **군집 수 k 를 정하지 않고** 단계적으로 군집을 형성한다.
 - 유사한 객체들을 군집으로 묶고, 그 군집을 기반으로 그와 유사한 군집을 새로운 군집으로 묶어가면서 군집을 계층적으로 구성함



2. 단일연결법

- 군집 i 와 군집 j 의 유사성 척도로 두 군집의 모든 객체 쌍의 거리 중 가장 가까운 거리를 사용
 - 객체 쌍의 가장 짧은 거리가 작을수록 두 군집이 더 유사하다고 평가

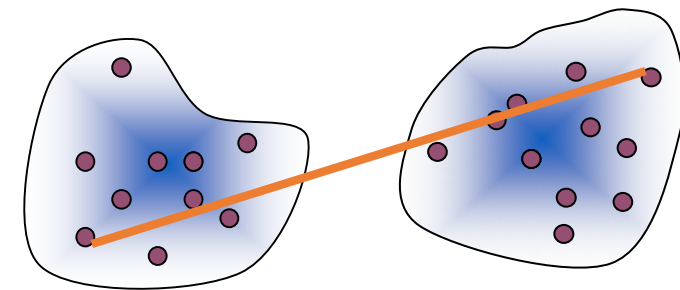


3. 그 외 연결법

• 완전연결법 (complete linkage method)

- 두 군집의 모든 객체 쌍의 거리 중 **가장 먼 거리**를 사용

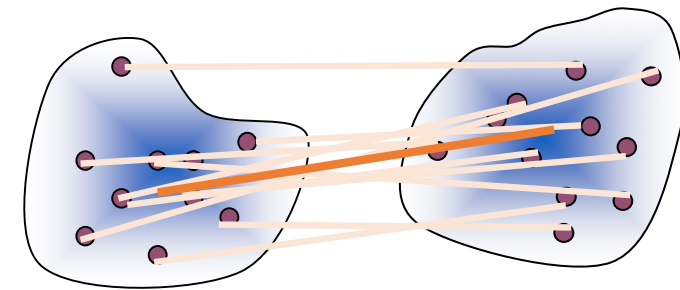
$$D(C_i, C_j) = \max_{u \in C_i, v \in C_j} d(u, v)$$



• 평균연결법 (average linkage method)

- 두 군집의 **모든 객체 쌍의 평균 거리**를 사용

$$D(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{u \in C_i, v \in C_j} d(u, v) \quad (|C_i|: \text{군집 } i \text{의 객체 수})$$

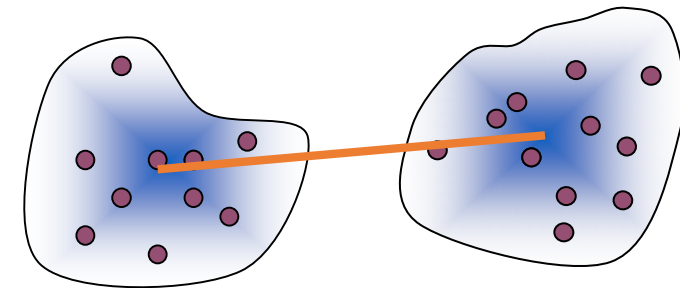


• 중심연결법 (centroid linkage method)

- 두 군집의 **중심 좌표**

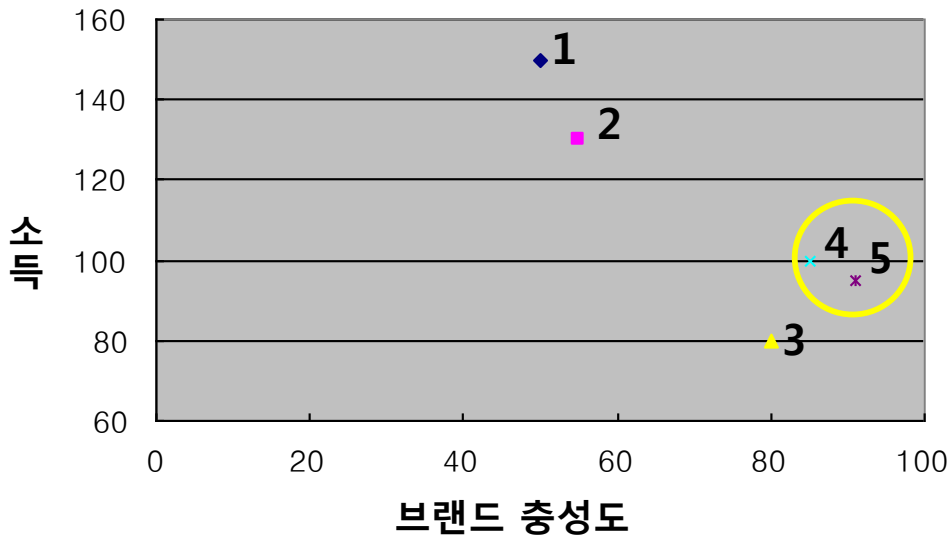
$$D(C_i, C_j) = d(c_i, c_j) \quad c_i = (\bar{X}_1^{(i)}, \bar{X}_2^{(i)}, \dots, \bar{X}_p^{(i)})$$

$$\bar{X}_k^{(i)} = \frac{1}{|C_i|} \sum_{j \in C_i} X_{kj}, (k = 1, \dots, p)$$



4. 단일연결법 예제

- 단일연결법을 사용한 군집화 과정 (유클리디안 거리 사용)



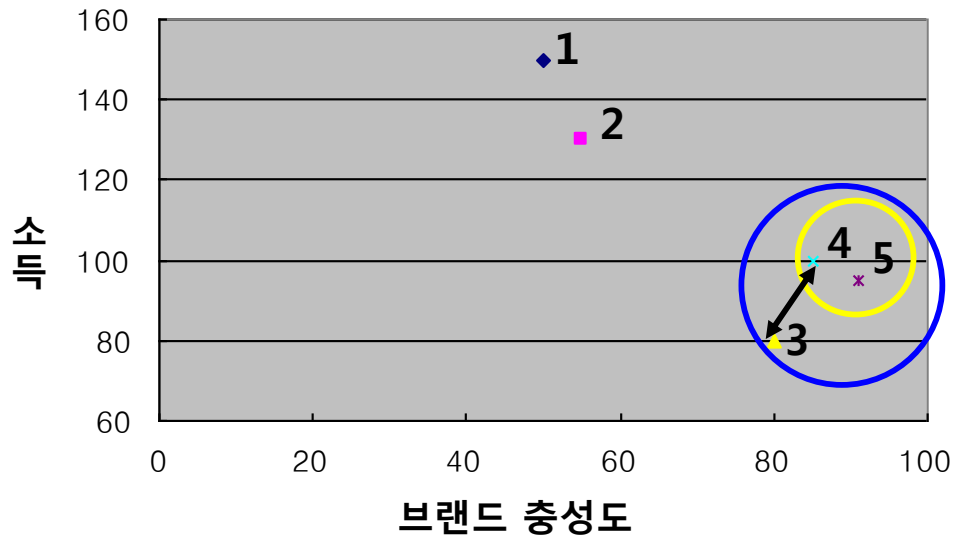
ID	소득	브랜드 충성도
1	150	50
2	130	55
3	80	80
4	100	85
5	95	91

ID	1	2	3	4	5
1	0.0				
2	20.6	0.0			
3	76.2	55.9	0.0		
4	61.0	42.4	20.6	0.0	
5	68.6	50.2	18.6	7.8	0.0

(1단계) 군집 사이 거리가 최소인 두 군집 4와 5를 묶어 하나의 군집으로 만든다.

4. 단일연결법 예제

- 단일연결법을 사용한 군집화 과정 (유클리디안 거리 사용)



(2단계) 군집 사이 거리가 최소인 두 군집 3과 (4,5)를 묶어 하나의 군집으로 만든다.

$$D\{(1), (4,5)\} = \min\{d_{14}, d_{15}\} = d_{14} = 61.0$$

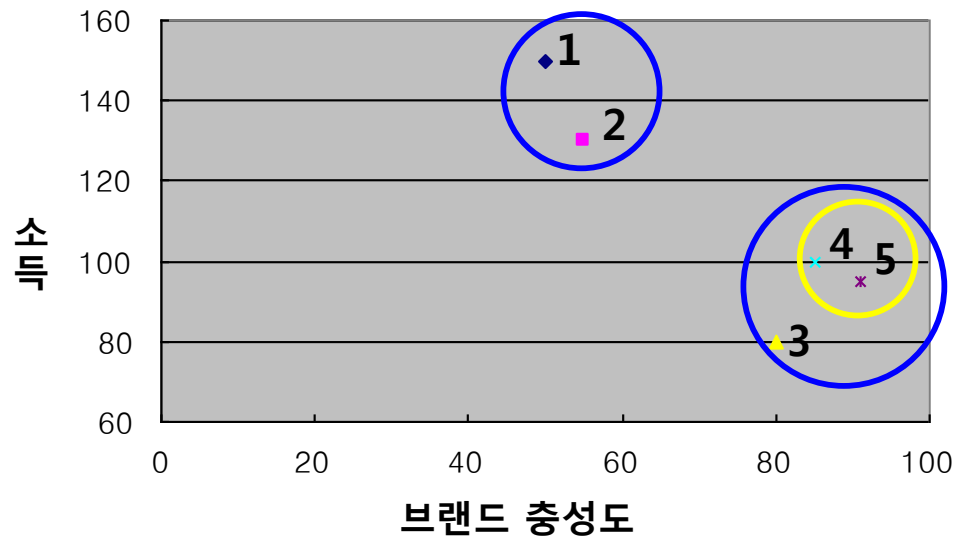
$$D\{(2), (4,5)\} = \min\{d_{24}, d_{25}\} = d_{24} = 42.4$$

$$D\{(3), (4,5)\} = \min\{d_{34}, d_{35}\} = d_{35} = 18.6$$

ID	1	2	3	4	5
1	0.0				
2	20.6	0.0			
3	76.2	55.9	0.0		
4	61.0	42.4	20.6	0.0	
5	68.6	50.2	18.6	7.8	0.0

4. 단일연결법 예제

- 단일연결법을 사용한 군집화 과정 (유클리디안 거리 사용)



$$D\{(1), (2)\} = \min\{d_{12}\} = d_{12} = 20.6$$

$$D\{(1), (3,4,5)\} = \min\{d_{13}, d_{14}, d_{15}\} = d_{14} = 61.0$$

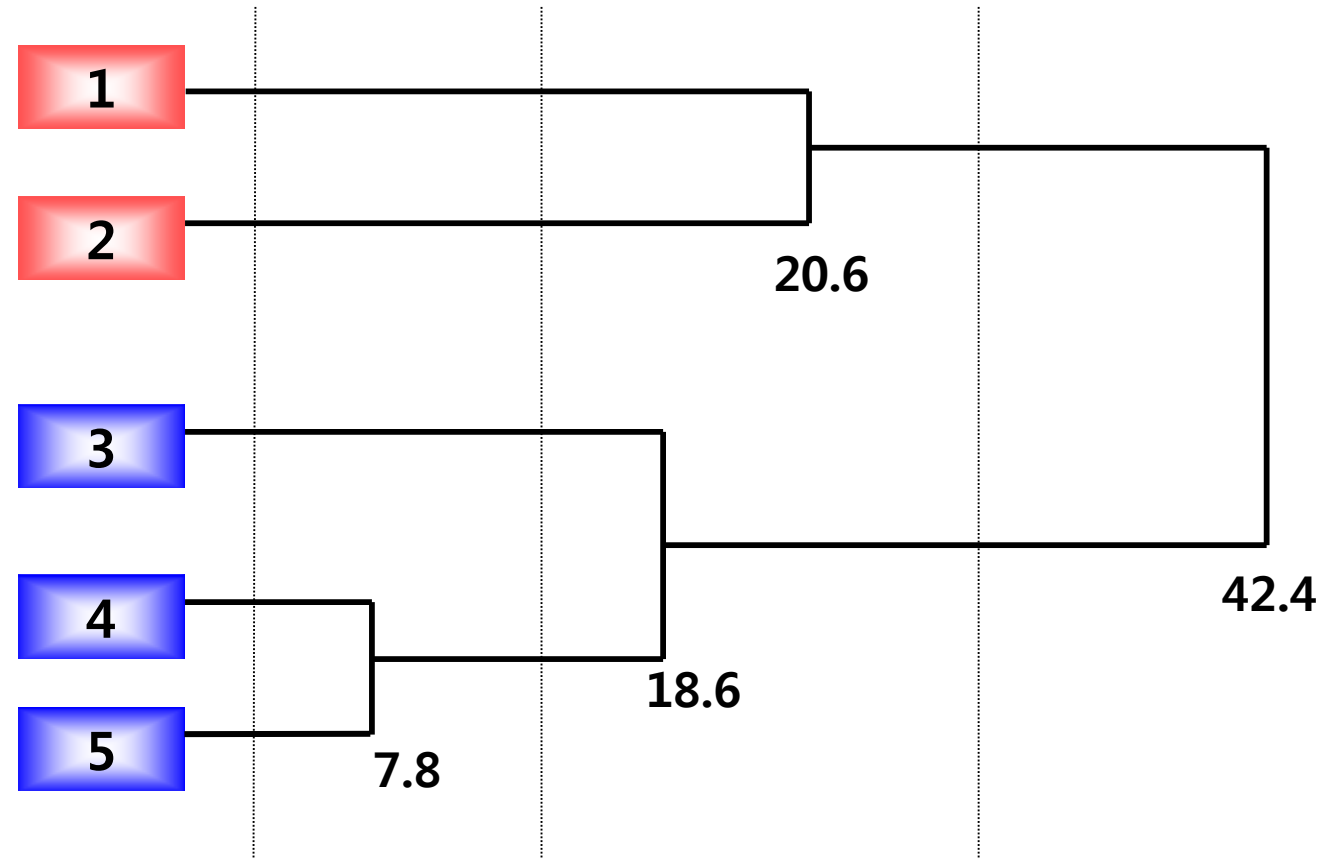
$$D\{(2), (3,4,5)\} = \min\{d_{23}, d_{24}, d_{25}\} = d_{24} = 42.4$$

ID	1	2	3	4	5
1	0.0				
2	20.6	0.0			
3	76.2	55.9	0.0		
4	61.0	42.4	20.6	0.0	
5	68.6	50.2	18.6	7.8	0.0

(3단계) 군집 사이 거리가 최소인 두 군집 1과 2를 묶어 하나의 군집으로 만든다.

4. 단일연결법 예제

- 덴드로그램은 군집 그룹과 유사성 수준을 표시하는 트리 다이어그램
 - 군집이 어떻게 형성되는지 확인하고 형성된 군집의 유사성 수준을 평가



5. 완전연결법 vs 평균연결법

• 데이터 설명

- 1833년 영국 Lancashire 방직 공장 임금
- DAAG package built in 데이터
- 총 51개의 객체
- 객체별 5개의 속성

1) 나이(age)

2) 남성 근로자 수(mnum)

3) 남성 근로자 평균 임금(mwage)

4) 여성 근로자 수(fnum)

5) 여성 근로자 평균 임금(fwage)

```
help("wages1833")
```

```
age
  age in years
mnum
  number of male workers
mwage
  average wage of male workers
fnum
  number of female workers
fwage
  average wage of female workers
```

```
> head(wages1833,n=10)
  age mnum mwage fnum fwage
1   10  204  30.5  122   35
2   11  195  37.8  198   38
3   12  245  43.0  241   44
4   13  233  50.5  233   46
5   14  256  56.5  236   59
6   15  240  63.0  215   68
7   16  204  83.5  256   72
8   17  141  88.5  245   78
9   18  164 141.0  279   90
10  19  135 138.3  251   98
```

5. 완전연결법 vs 평균연결법

• 데이터 불러오기 (DAAG package)

```
# lec13_2_clus.R
# Clustering
# Hierarchical Clustering
# Linkage method, Dendrogram

# needs "lattice", "DAAG" package for loading dataset
# install.packages("lattice")
install.packages("DAAG")
library(lattice)
library(DAAG)

# Load data in DAAG package
# the wages of Lancashire cotton factory workers in 1833
data("wages1833")
#help("wages1833")
head(wages1833,n=10)

# remove observations with the missing values
dat1<-wages1833
dat1<-na.omit(dat1)
str(dat1)
```

DAAG패키지 설치, 라이브러리 설정

데이터 불러오기

결측치 데이터 삭제 (전처리)

5. 완전연결법 vs 평균연결법

- 계층적 군집분석 : `hclust(거리계산결과, method=" ")`

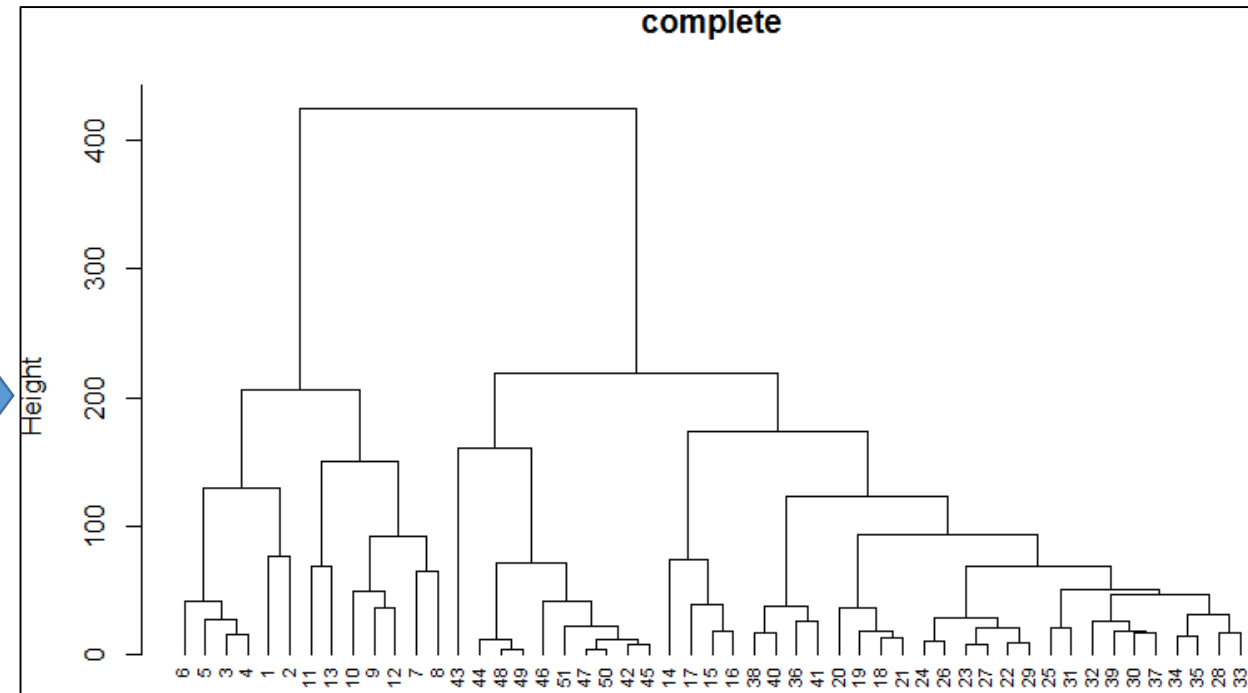
```
# calculate distance between each nodes
dist_data<-dist(dat1)
```

유클리디안 거리 사용

(1) 완전연결법 적용결과 (거리 계산은 유클리디안 사용)

single(단일), complete(완전)
average(평균), centroid(중심)

```
# prepare hierarchical cluster
# complete linkage method
hc_a <- hclust(dist_data, method = "complete")
plot(hc_a, hang = -1, cex=0.3, main = "complete")
```



5. 완전연결법 vs 평균연결법

(2) 평균연결법 적용결과 (거리 계산은 유클리디안)

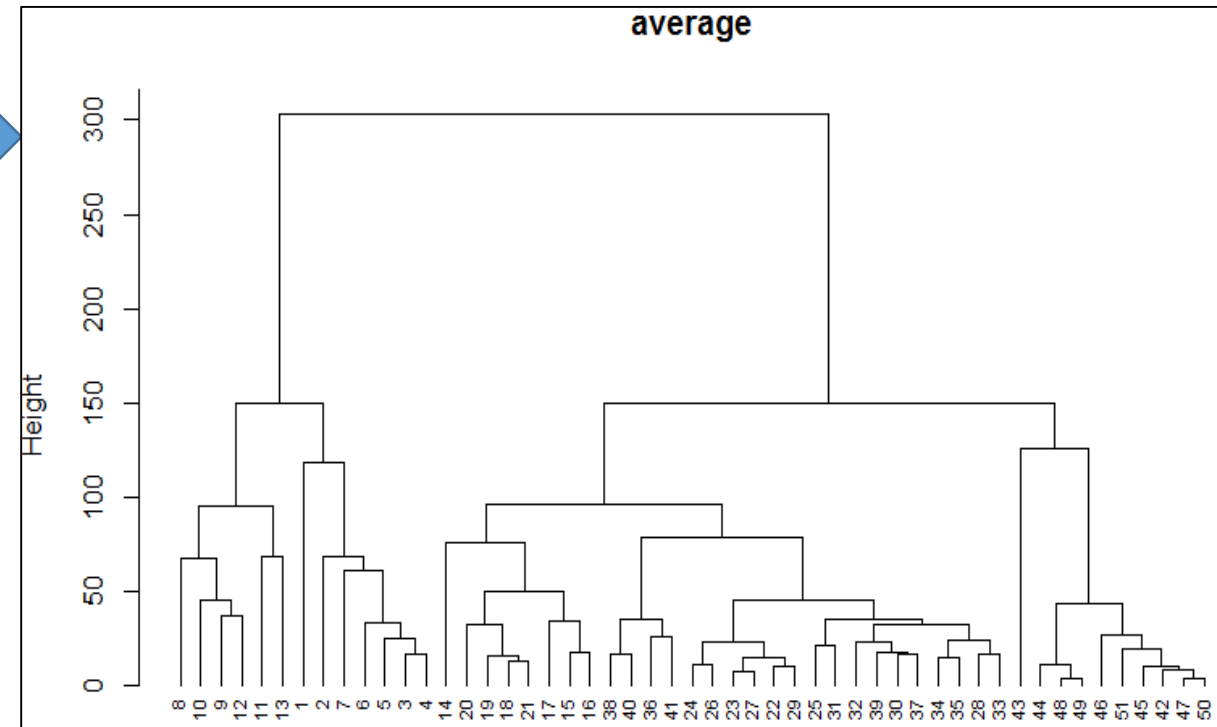
single(단일), complete(완전)
average(평균), centroid(중심)

```
# average linkage method
# check how different from complete method
hc_c <- hclust(dist_data, method = "average")
plot(hc_c, hang = -1, cex=0.3, main = "average")
```

라벨을 일정한
위치로 고정

글자 크기

메인 타이틀



6. 워드 연결방법 (Ward's method)

(3) 워드방법을 적용한 결과 (거리 계산은 유클리디안)

```
# Ward's method
hc_c <- hclust(dist_data, method = "ward.D2")
plot(hc_c, hang = -1, cex=0.7, main = "Ward's method")
```

