Chapter
9

비지도학습 – 군집분석

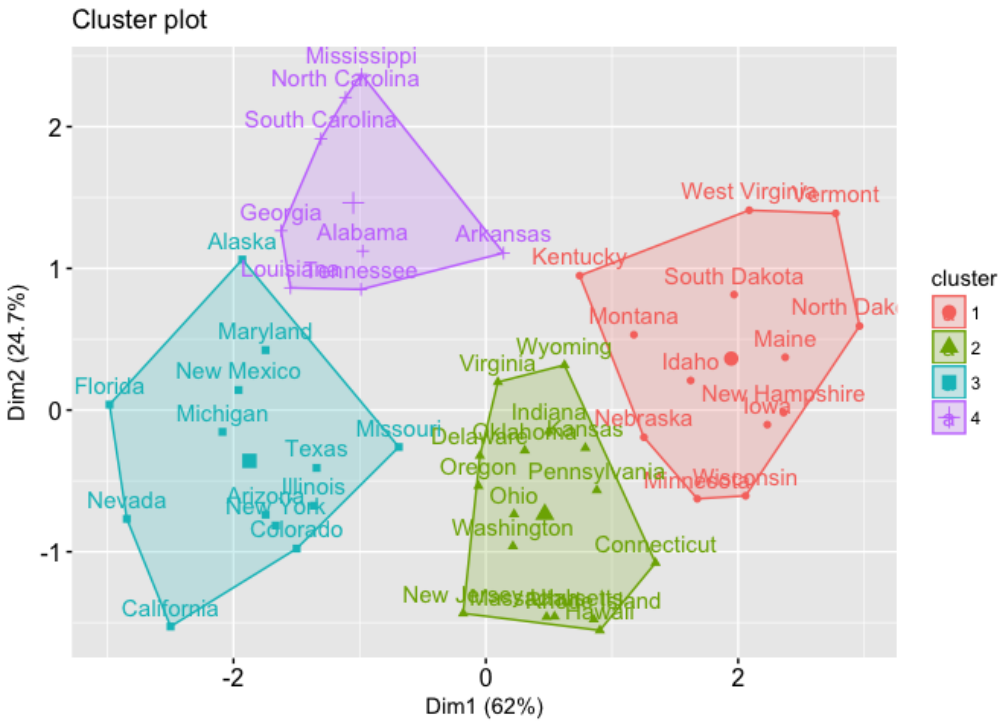
1. 군집분석이란?
2. K-means 군집분석의 원리
3. 적정 K의 값
4. 계층적 군집분석
5. 코드 실습

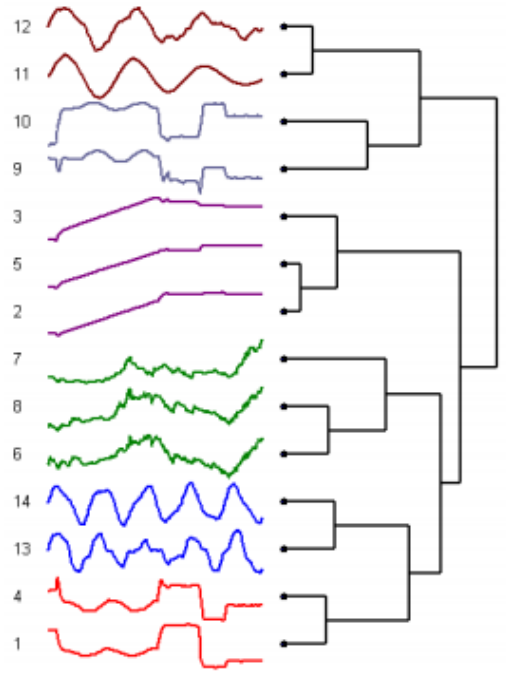
1

군집분석이란? (Clustering)

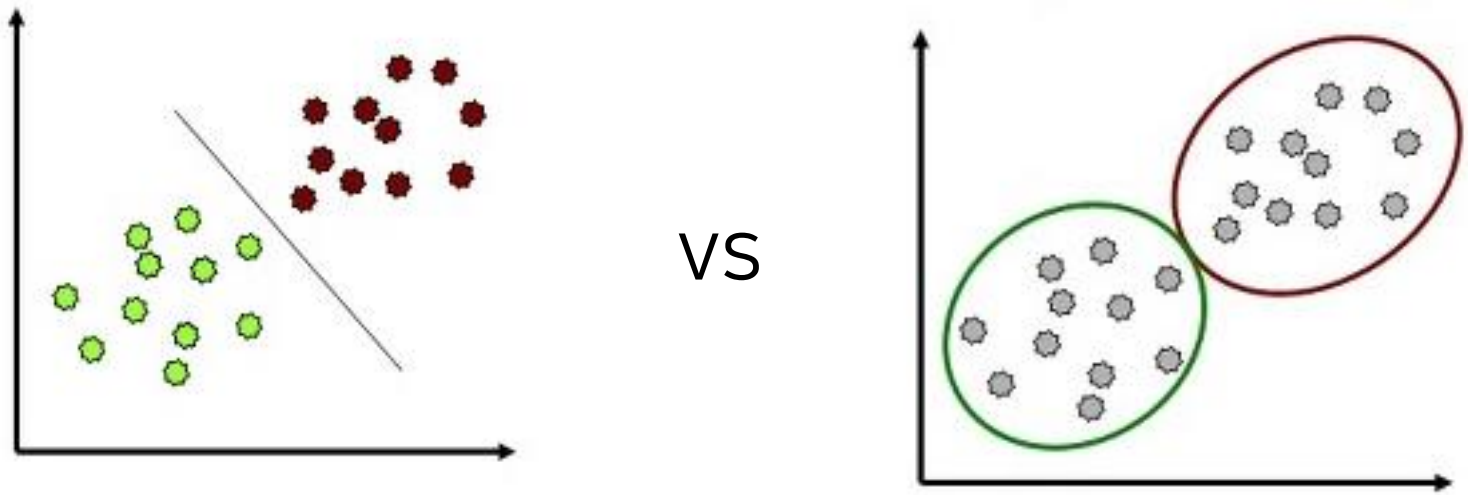
군집 분석(Clustering)

전체 데이터를 보다 일관성과 응집성이 높은 세부 그룹으로 나누는 분석이다.





분류(Classification) vs 군집화(Clustering)



- 분류(Classification): 범주의 수 및 각 개체의 범주 정보를 사전에 알 수 있으며, 입력 변수 값으로부터 범주 정보를 유추하여 새로운 개체에 대해 가장 적합한 범주로 할당하는 문제 (지도 학습)
- 군집화(Clustering): 군집의 수, 속성 등이 사전에 알려져 있지 않으며 최적의 구분을 찾아가는 문제 (비지도 학습)

군집분석(Clustering)이란?

정의

각 개체의 동질성을 측정하여 **동질성이 높은** 대상 **군집을 탐색**하고, 군집에 속한 개체들의 **동질성**과 서로 다른 군집에 속한 개체간의 **이질성을 규명**하는 통계 분석 방법

목적

군집 분석을 이해하고, R을 활용하여 군집 분석을 **실습**해보자

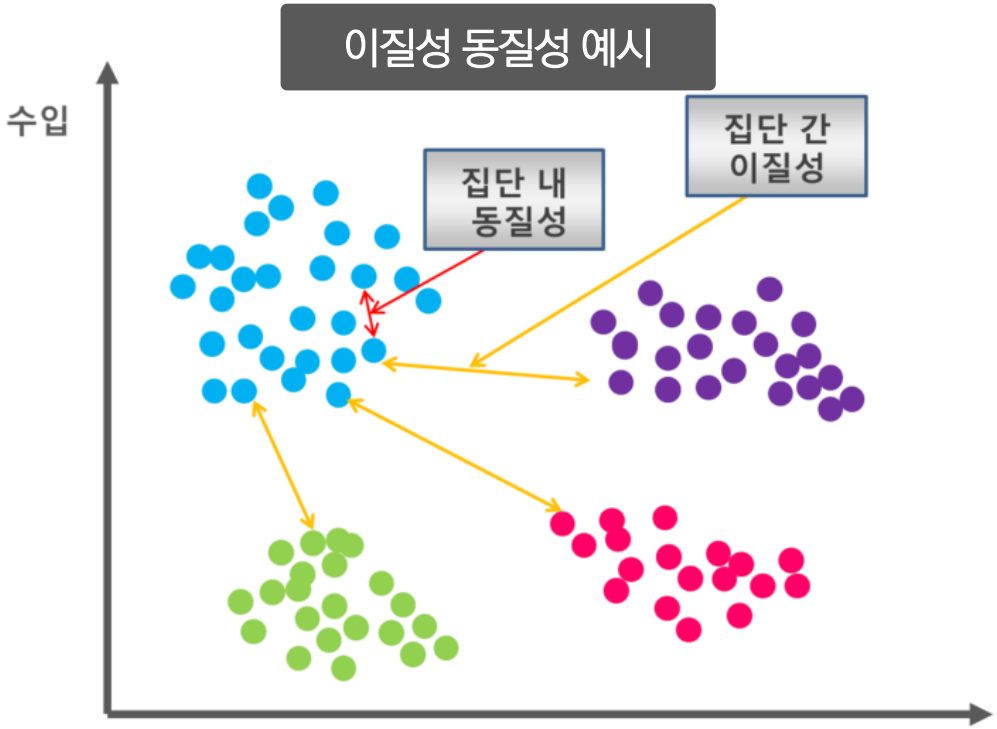
Goal!

1. **k-means 군집분석**이란?
2. k-means 군집분석의 **원리**를 알아보자
3. **적정 k**의 값은 어떻게 구하는가?
4. **계층적 군집분석**(Hierarchical Clustering)이란?
5. 군집분석을 **실습**해보자

Goal 1. k-means 군집분석이란?

k-means 군집분석이란?

각 개체의 **동질성을 거리를 이용해 측정**하여 동질성이 높은(서로 비슷한) 대상 **군집을 판별**하고, 군집에 속한 개체들의 동질성과 서로 다른 군집에 속한 개체간의 이질성을 규명하는 분석 방법



Source : https://m.blog.naver.com/PostView.nhn?blogId=gkenq&logNo=10188552802&proxyReferer=https%3A%2F%2Fwww.google.co.kr%2F

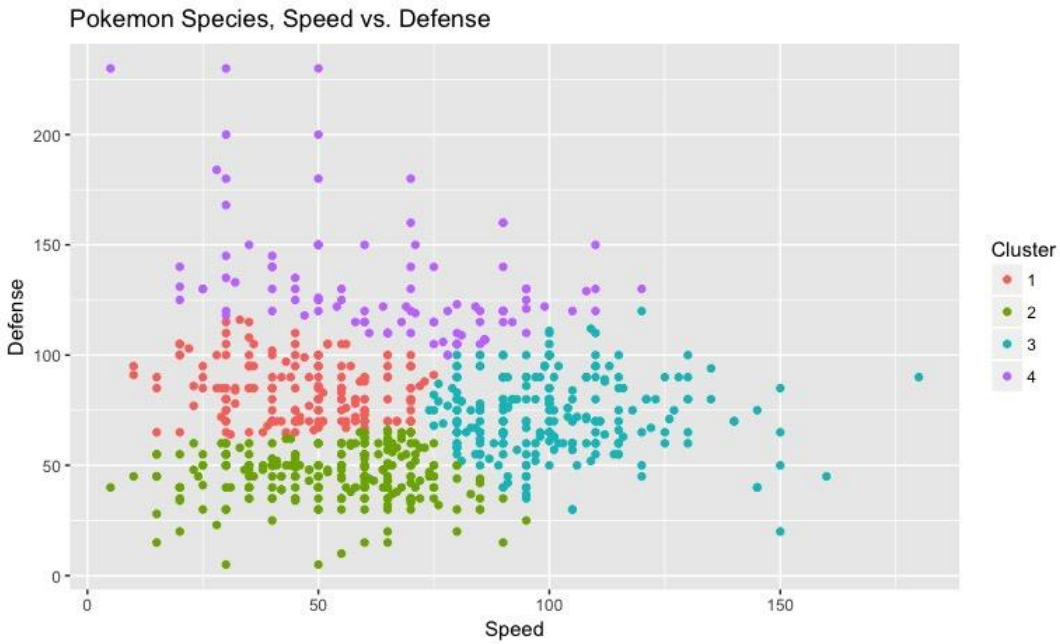
Goal 1. k-means 군집분석이란?

주요 특징

- 각 군집은 하나의 중심(Centroid)을 가진다.
- 각 개체는 가장 가까운 중심에 할당되며, 같은 중심에 할당된 개체들이 모여서 하나의 군집을 이룬다.
- 군집의 수 k가 정해져야 알고리즘을 실행할 수 있다.

k-means 군집 분석 활용 예

: 고객 세분화, 이상 탐지



Source : https://sokolj.com/K-Means-Clustering-Implementation/

Goal 1. k-means 군집분석이란?

장점

① 개체간의 거리를 기반하여 군집을 분류하는 원리가 간단함

② 데이터 변환 없이 그 자체로 이용할 수 있어 데이터 구조가 간단함

③ 개체가 많은 경우에도 쉽게 사용됨 (계산 시간 짧음)

단점

① 거리를 기반하여 군집을 형성

② 초기 군집 수(k)에 따라 결과가 달라짐

③ 연구자 주관에 따라 해석이 다를 수 있음

④ 변수의 유형에 제한이 있음 (범주형 변수 사용 불가)

Goal 1. k-means 군집분석이란?

주요 예시

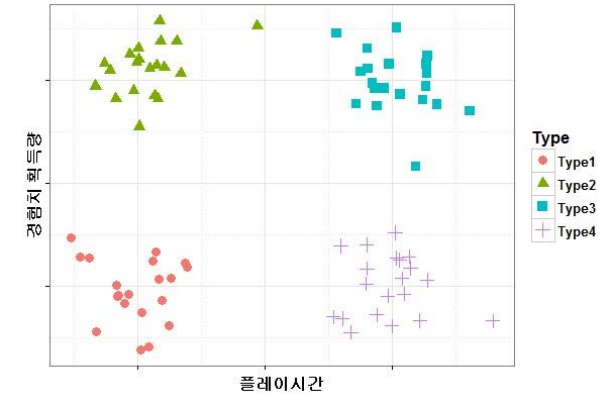
고객 세분화

우수 고객의 인구통계적 요인과 생활패턴 등을

파악함으로써 개별 고객에 대한 맞춤 관리

신상품의 판촉이나 교차판매를 위한 목표 집단을 설정

고객세분화예시



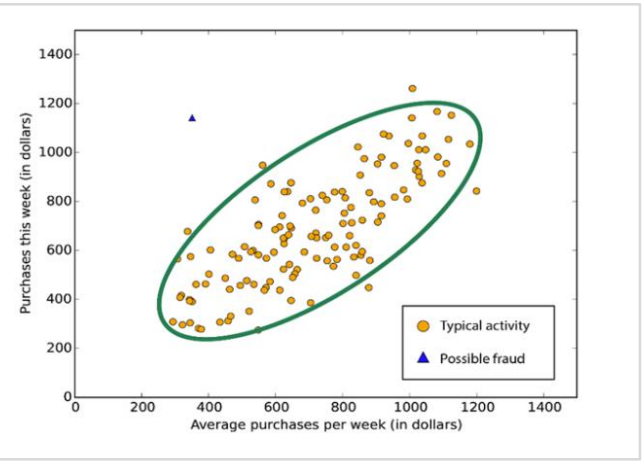
이상 탐지

정상적인 거래에 대한 데이터에 대하여 군집 분석을

실시하여 군집을 찾아서 신규 거래가 군집과의

거리가 크면 자동적으로 불량품으로 인식

이상탐지예시



Source : https://m.blog.naver.com/PostView.nhn?blogId=gkenq&logNo=10188552802&proxyReferer=https%3A%2F%2Fwww.google.co.kr%2F

Goal 1. k-means 군집분석이란?

[고객세분화]

그 결과, 화장품에 관심을 갖기 시작한 중고등학생은 다른 그룹(Group1)보다는 앱 내 활동이 구매에 미치는 영향이 컸습니다. 이들의 구매를 유도하기 위해서는 팔로워 및 팔로잉 기능을 강화하고 인적 네트워크 활성화하는 등 앱 내 활동을 활성화시키는 방안이 필요했습니다.

또 이벤트에만 관심있는 30대 이상 진짜 언니들(Group5)은 이벤트 참여를 주목적으로 언니의 파우치 앱을 사용하고 있었습니다. 이들은 10만원 이상씩 구매하는 통 큰 고객이었죠. 이들의 구매를 더욱 늘리기 위해서는 이벤트를 차별적으로 적용할 방안이 필요했습니다.

Source : http://www.edaily.co.kr/news/news_detail.asp?newsId=01521926615928592&mediaCodeNo=257

Goal 1. k-means 군집분석이란?

언니의 파우치는 30대 이상 사용자들을 위해 이벤트를 진행했습니다. 30대를 위해 안티에이징 화장품 리뷰 이벤트를 진행하고 좋은 반응을 얻었습니다.

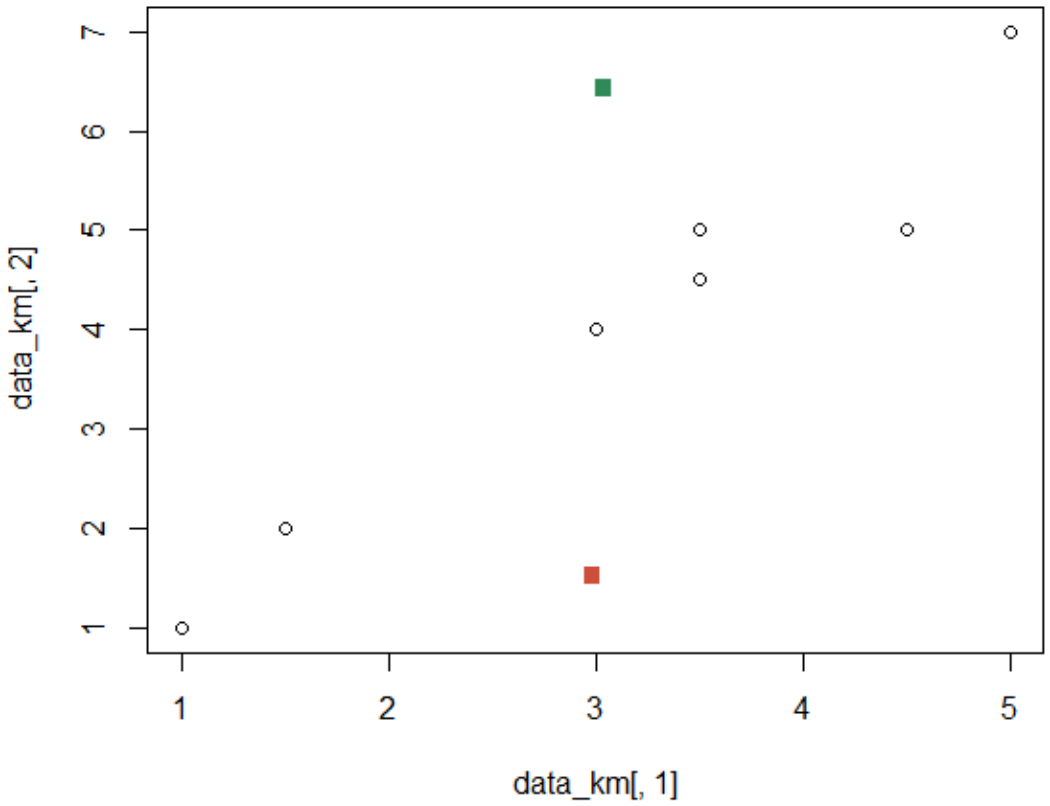
데이터분석을 통해 언니의 파우치의 주요 고객층이 10대 후반에서 20대 초반이라는 사실을 발견하고 이들을 타깃으로 하는 신제품도 개발했습니다. 고객 분석 전에는 주요 고객층이 20대 후반이라는 막연한 추정만 했었지만 데이터 분석으로 고객층을 확실하게 알 수 있었던 것입니다.

Source : http://www.edaily.co.kr/news/news_detail.asp?newsId=01521926615928592&mediaCodeNo=257

2 K-means군집분석 원리

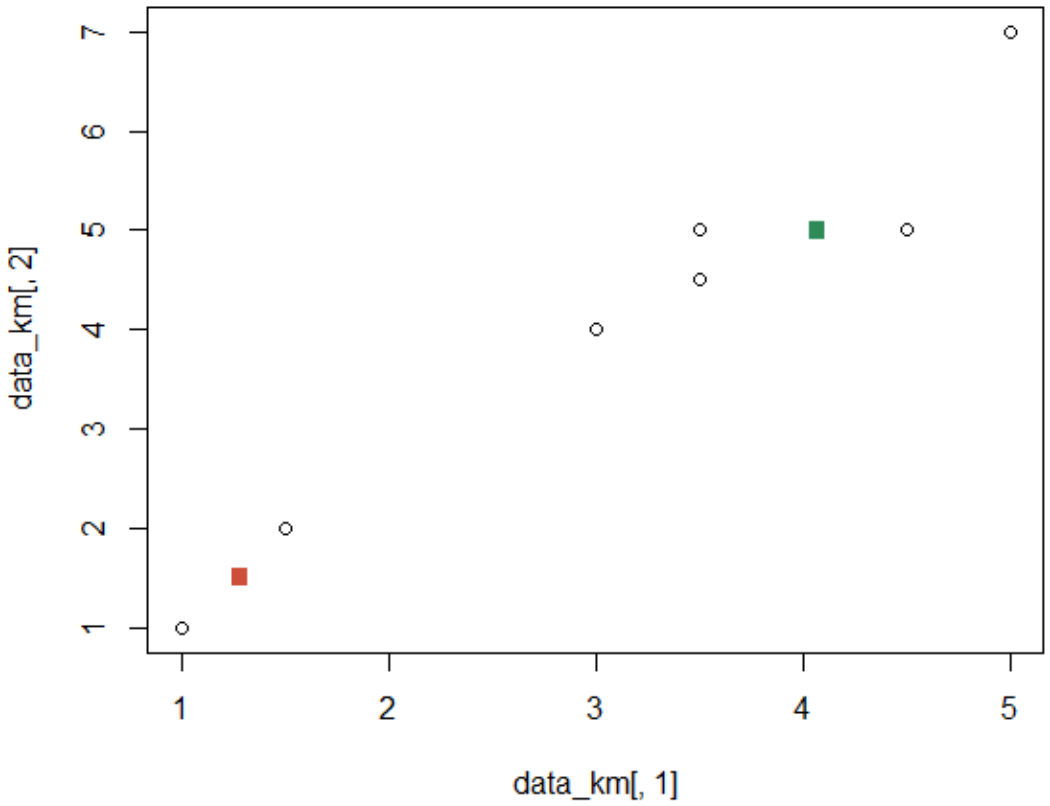
Goal 2. k-means 군집분석의 원리를 알아보자

- 1. 랜덤하게 중심값을 선택한다.(처음 중심값 선택)
  방법은 여러 가지가 있을 수 있다.



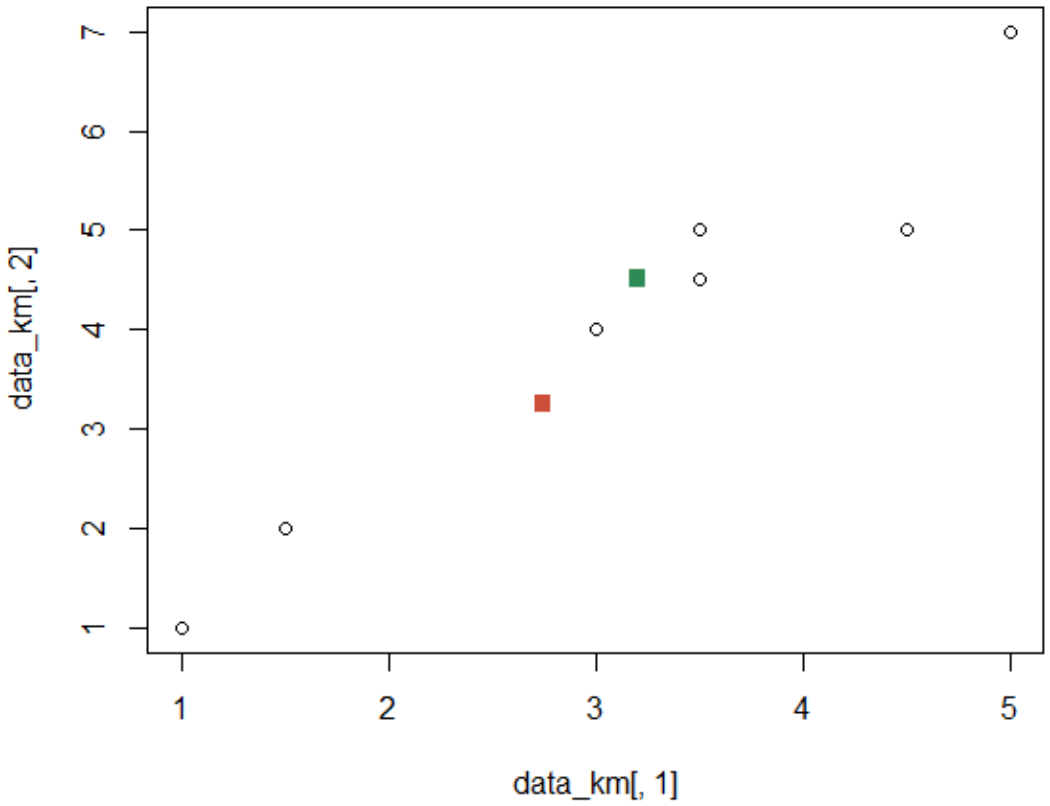
Goal 2. k-means 군집분석의 원리를 알아보자

- 1. 랜덤하게 중심값을 선택한다.(처음 중심값 선택)
  방법은 여러 가지가 있을 수 있다.



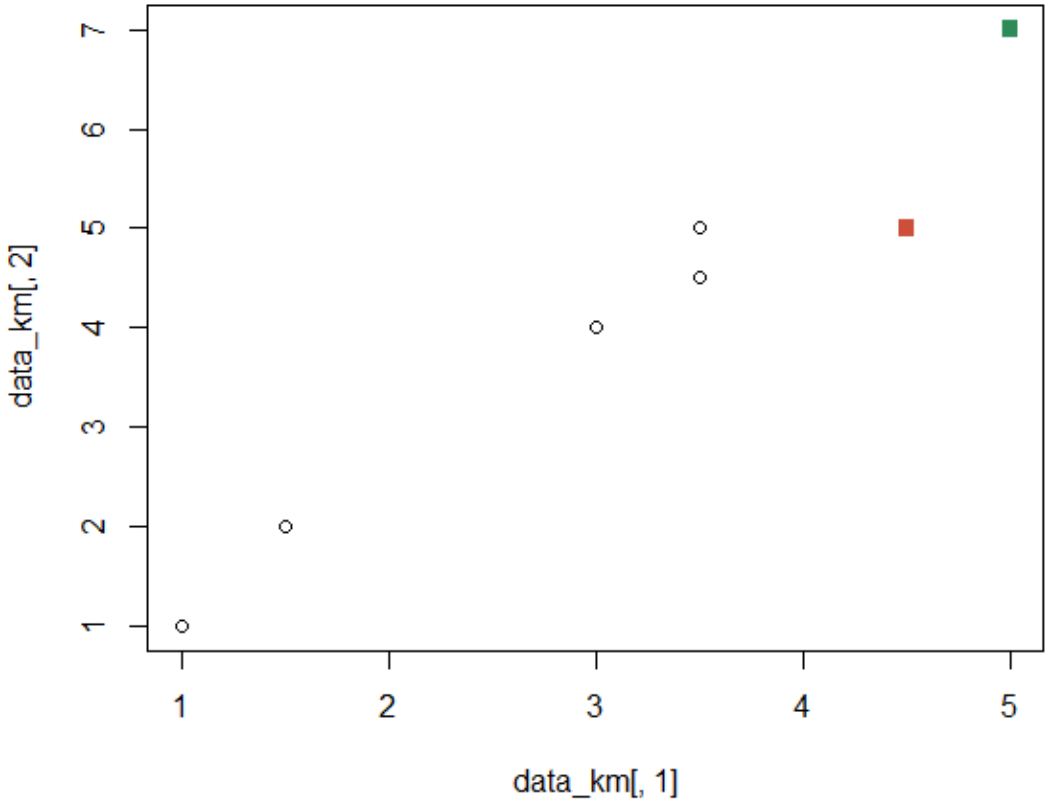
Goal 2. k-means 군집분석의 원리를 알아보자

- 1. 랜덤하게 중심값을 선택한다.(처음 중심값 선택)
  방법은 여러 가지가 있을 수 있다.



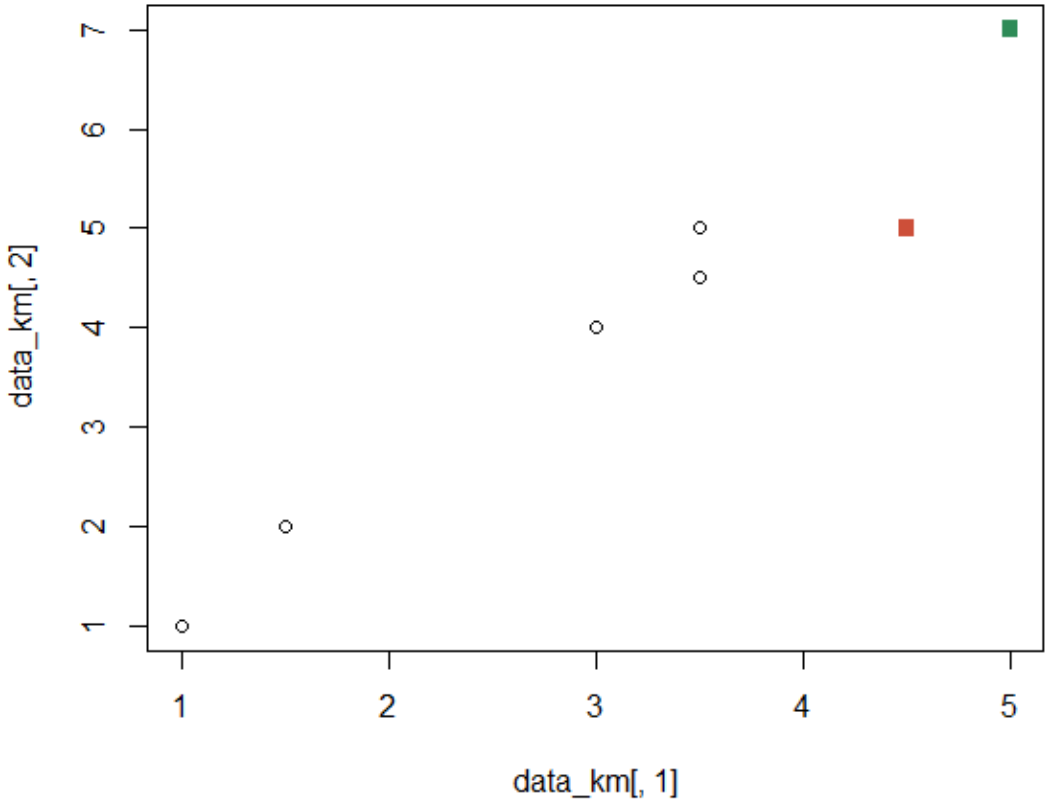
Goal 2. k-means 군집분석의 원리를 알아보자

- 1. 랜덤하게 중심값을 선택한다.(처음 중심값 선택)
  이 예시에서는 k=2라 가정하고, 처음 데이터 중 임의로 두 개를선택하였다.



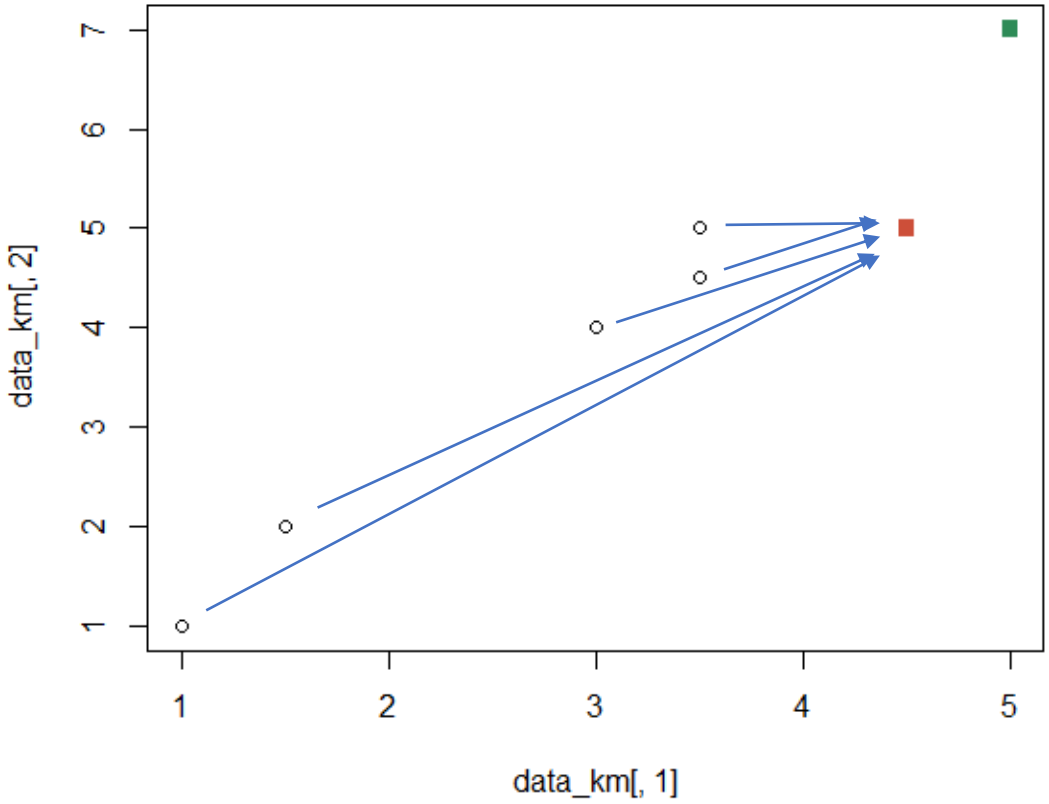
Goal 2. k-means 군집분석의 원리를 알아보자

- 2. k개의 중심값과 각 개별 데이터간의 거리를 측정한다.
  가장 가까운 군집에 해당 데이터를 할당한다.(군집 할당)



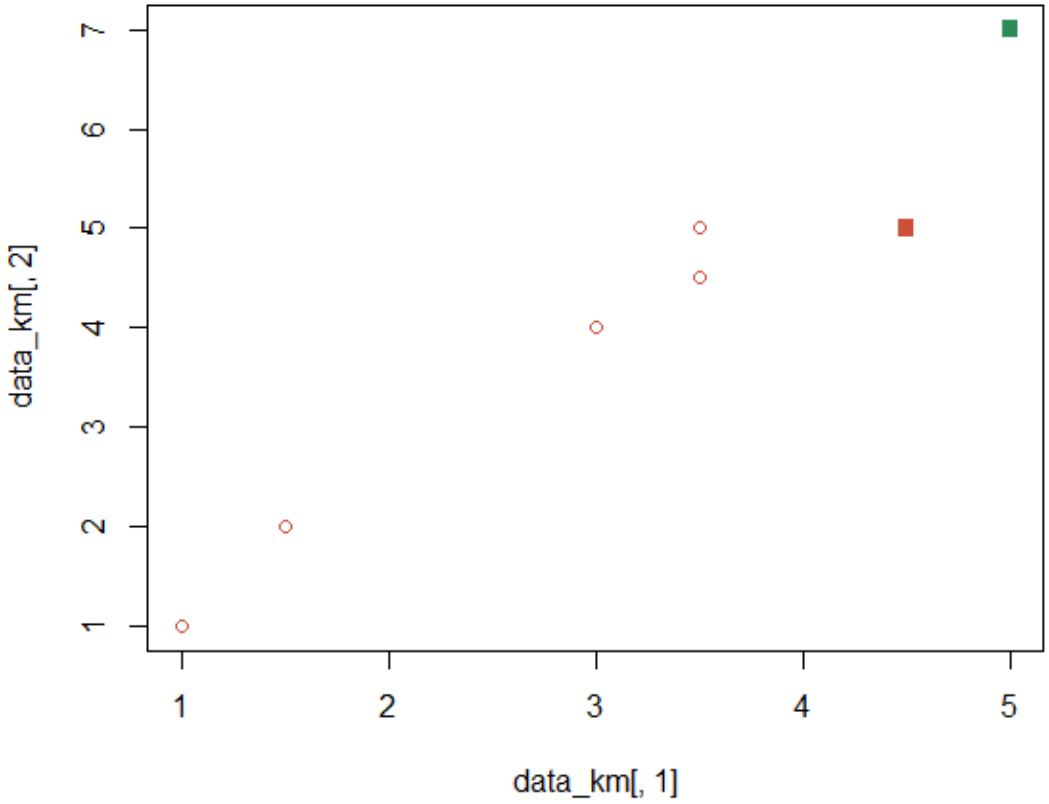
Goal 2. k-means 군집분석의 원리를 알아보자

- 2. k개의 중심값과 각 개별 데이터간의 거리를 측정한다.
  가장 가까운 군집에 해당 데이터를 할당한다.(군집 할당)



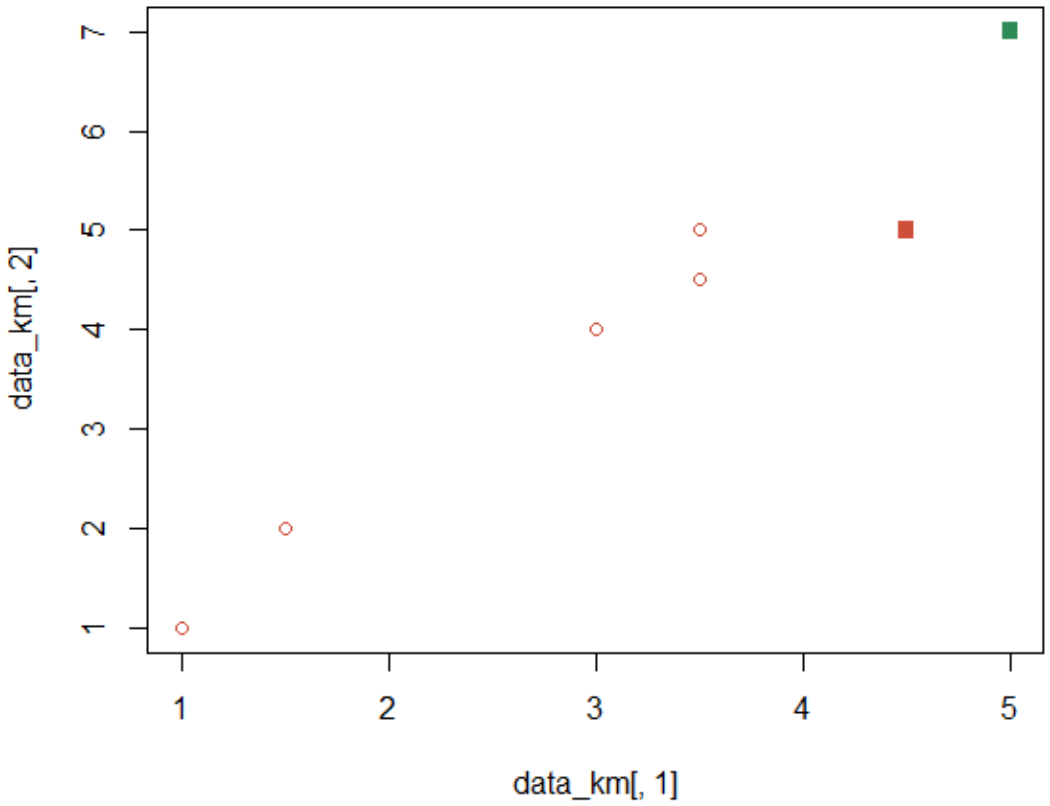
Goal 2. k-means 군집분석의 원리를 알아보자

- 2. k개의 중심값과 각 개별 데이터간의 거리를 측정한다.
  가장 가까운 군집에 해당 데이터를 할당한다.(군집 할당)



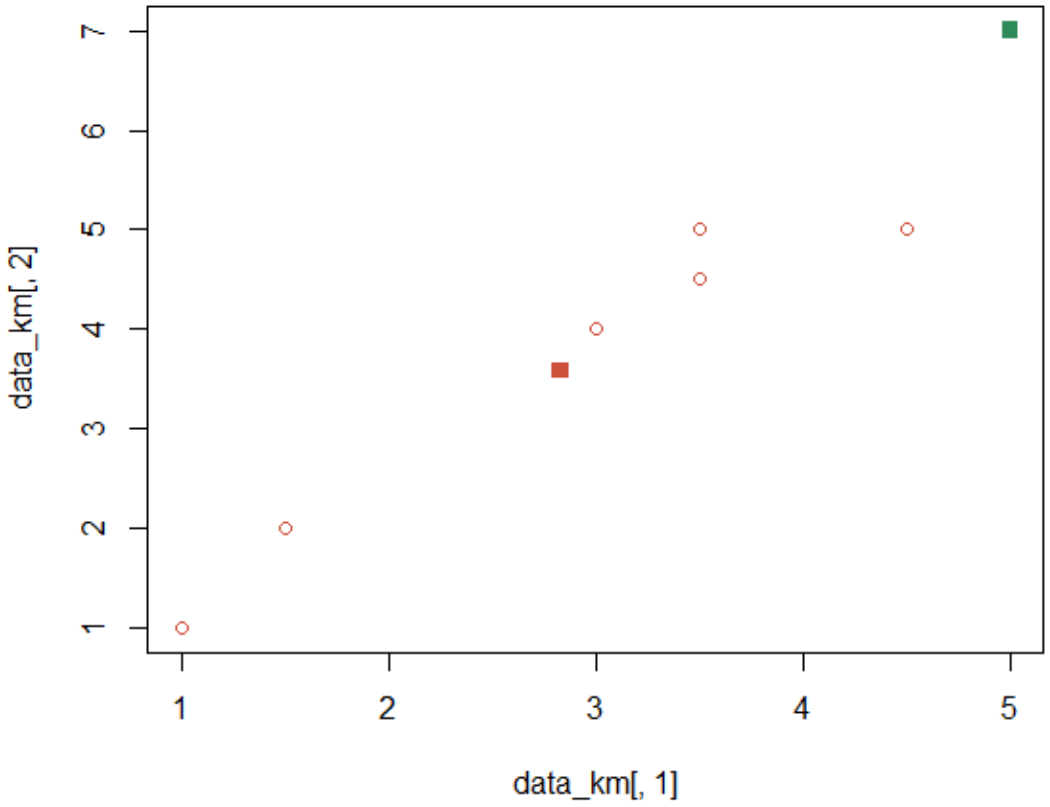
Goal 2. k-means 군집분석의 원리를 알아보자

- 3. 각 군집마다 새로운 중심값을 계산한다.



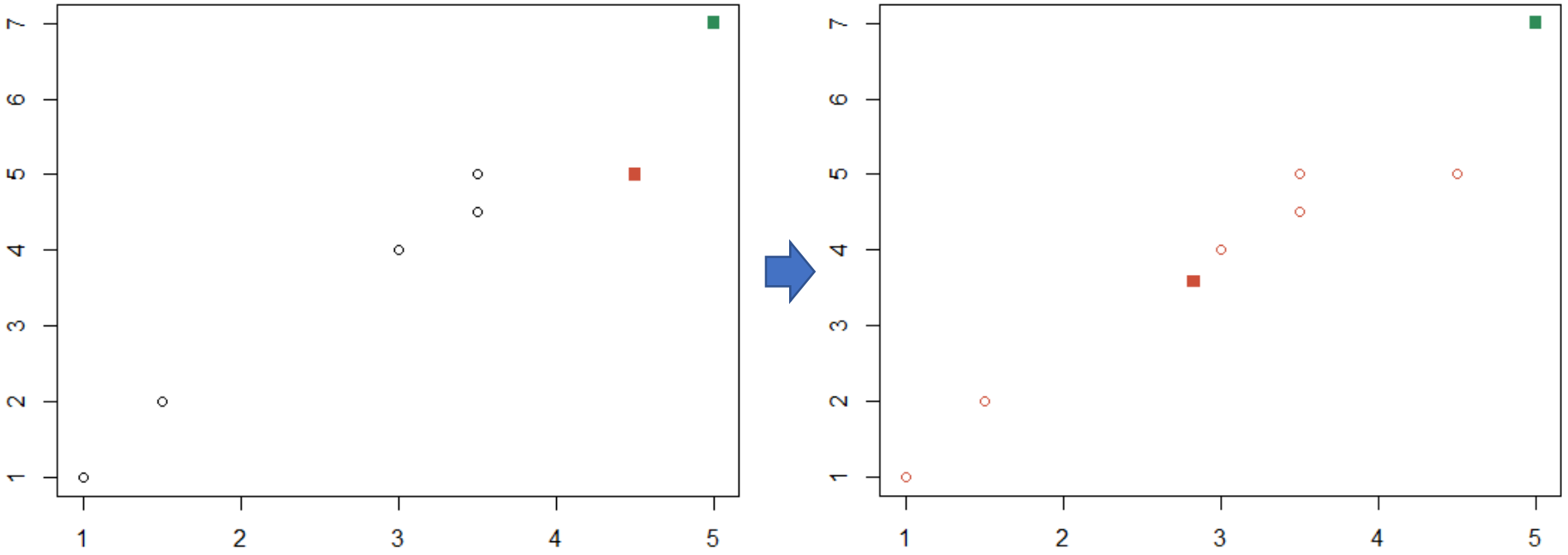
Goal 2. k-means 군집분석의 원리를 알아보자

- 현재 그룹에 속한 점들만 모아 x좌표의 평균을 그 그룹의 중심값의 x좌표로 하고, y좌표도 마찬가지로 계산한다.



Goal 2. k-means 군집분석의 원리를 알아보자

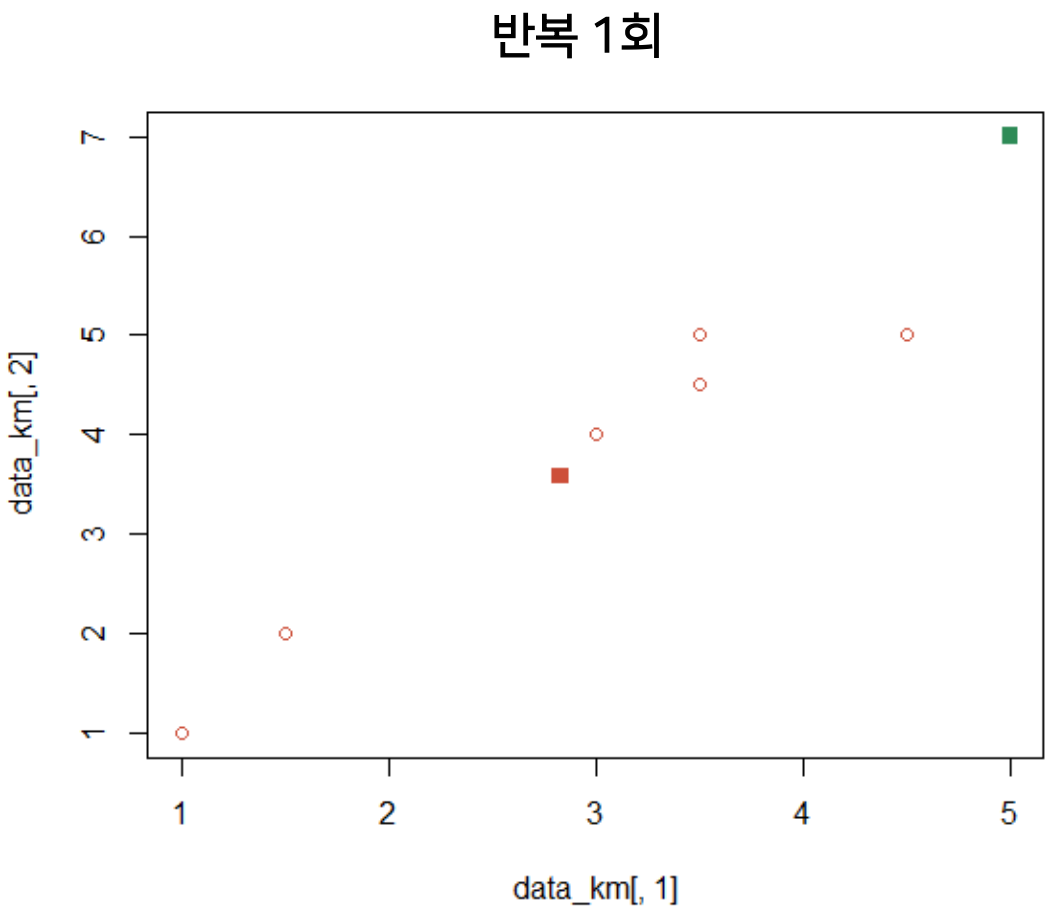
- 4. 1번의 중심값과 3번에서 선택된 중심값이 변화가 없다면 멈춘다.
  만약 데이터의 변화가 있다면 다시 1번부터 반복한다.



이 경우는 변화가 있으므로 다시 1번부터 반복하기로 한다.

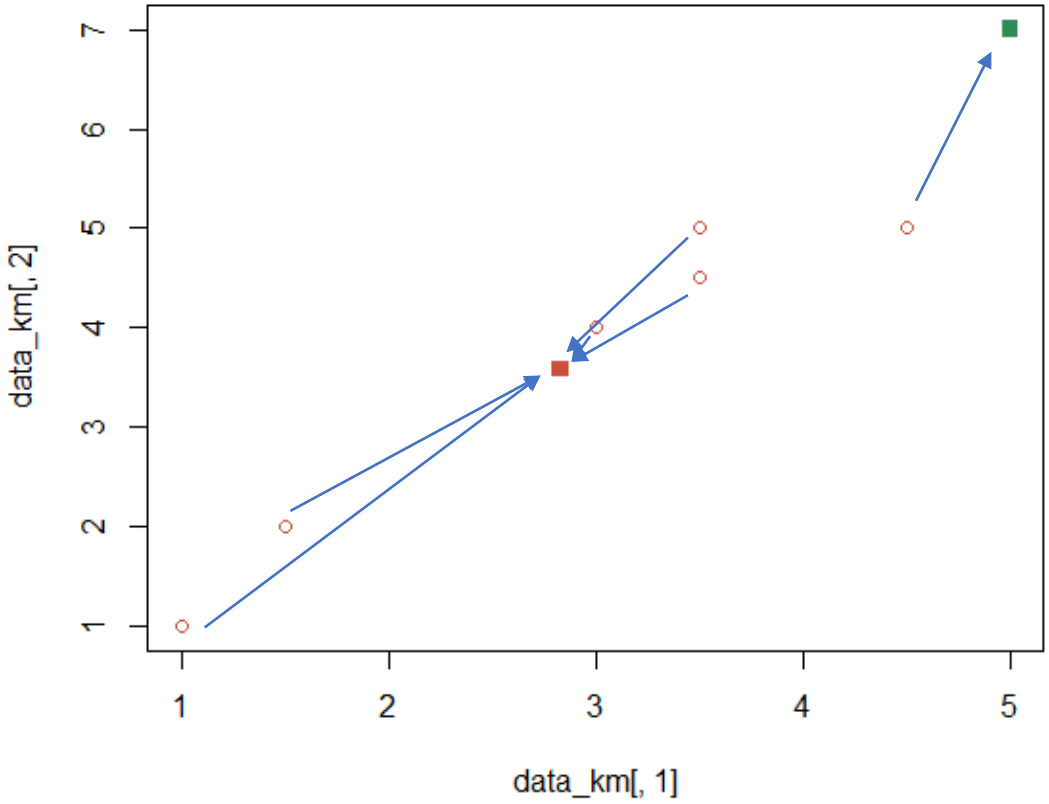
Goal 2. k-means 군집분석의 원리를 알아보자

- 4. 1번의 중심값과 3번에서 선택된 중심값이 변화가 없다면 멈춘다.
  만약 데이터의 변화가 있다면 다시 1번부터 반복한다.



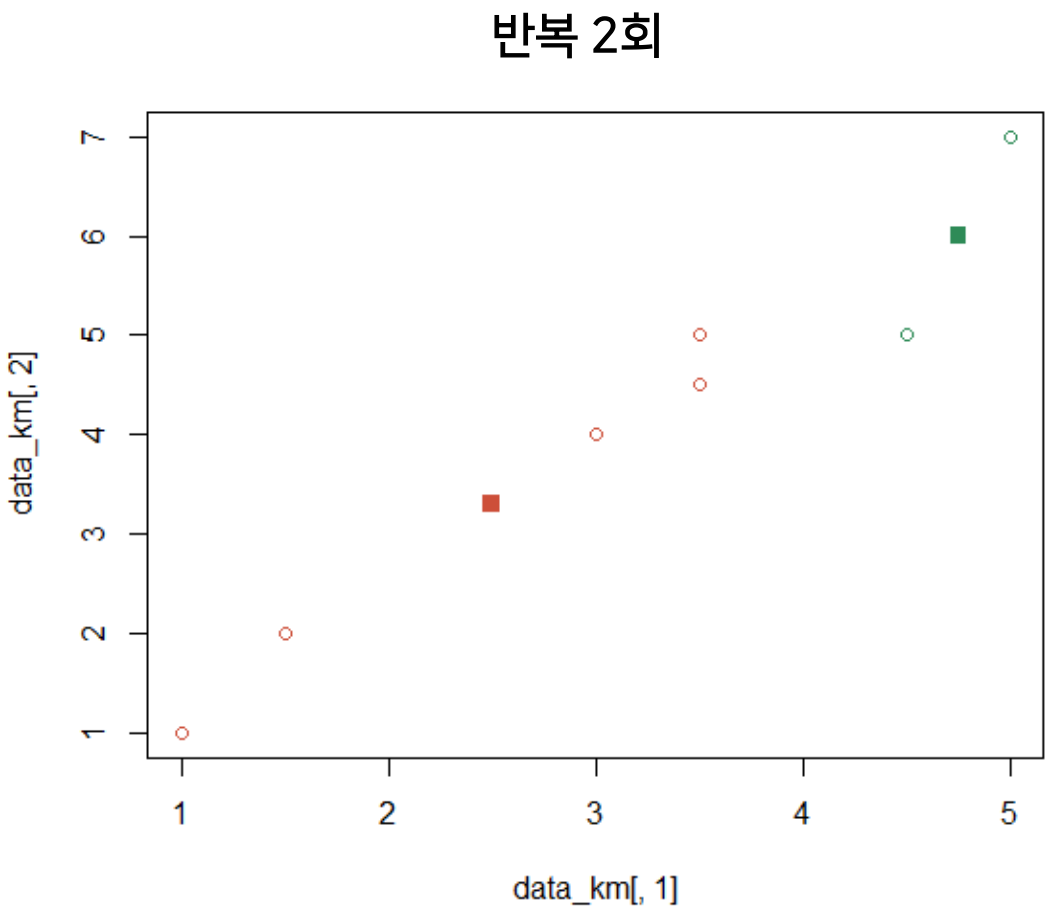
Goal 2. k-means 군집분석의 원리를 알아보자

- 4. 1번의 중심값과 3번에서 선택된 중심값이 변화가 없다면 멈춘다.
  만약 데이터의 변화가 있다면 다시 1번부터 반복한다.



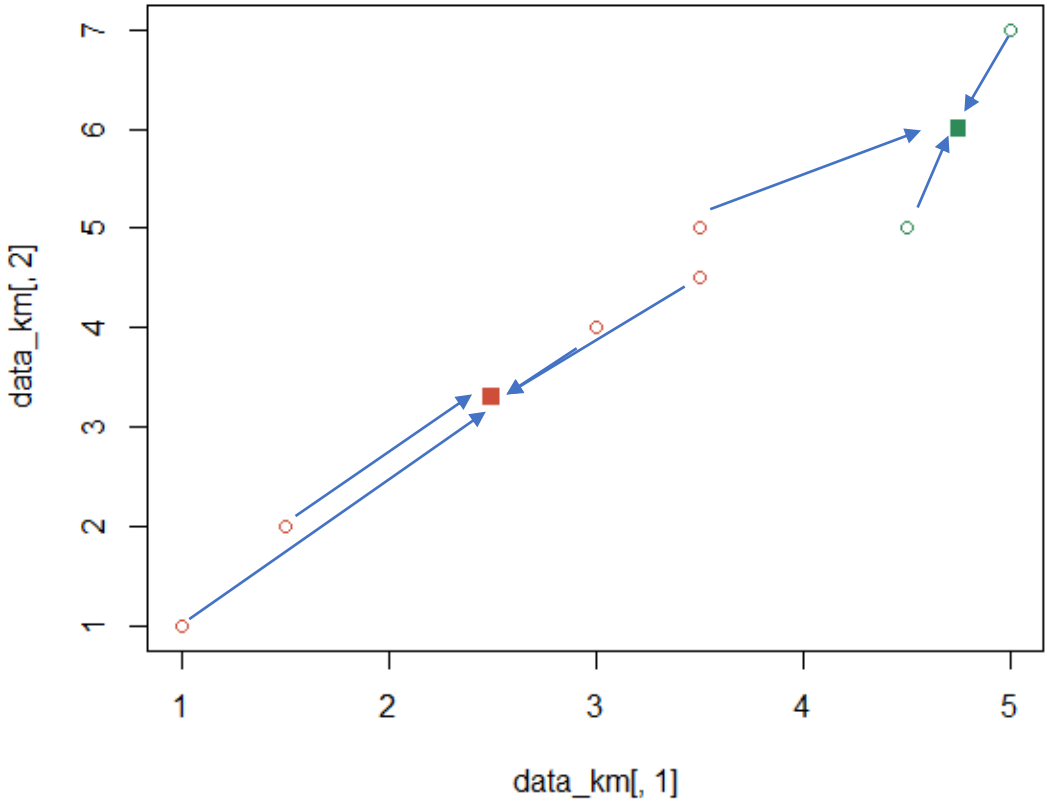
Goal 2. k-means 군집분석의 원리를 알아보자

- 4. 1번의 중심값과 3번에서 선택된 중심값이 변화가 없다면 멈춘다.
  만약 데이터의 변화가 있다면 다시 1번부터 반복한다.



Goal 2. k-means 군집분석의 원리를 알아보자

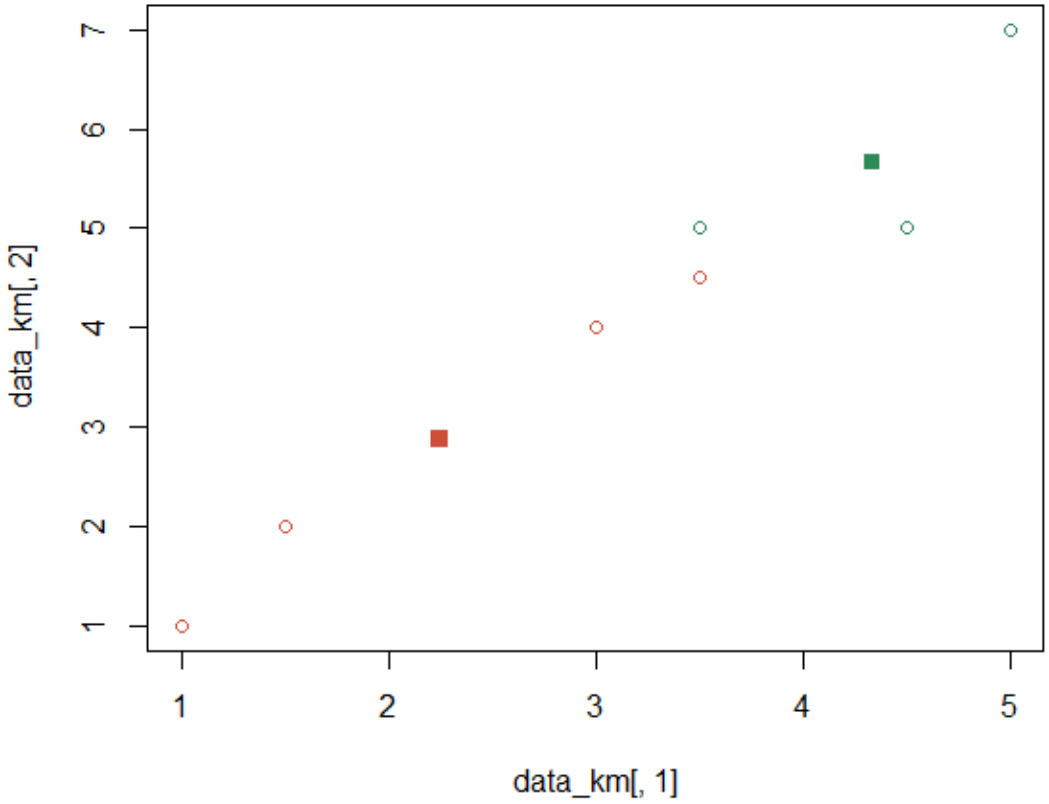
- 4. 1번의 중심값과 3번에서 선택된 중심값이 변화가 없다면 멈춘다.
  만약 데이터의 변화가 있다면 다시 1번부터 반복한다.



Goal 2. k-means 군집분석의 원리를 알아보자

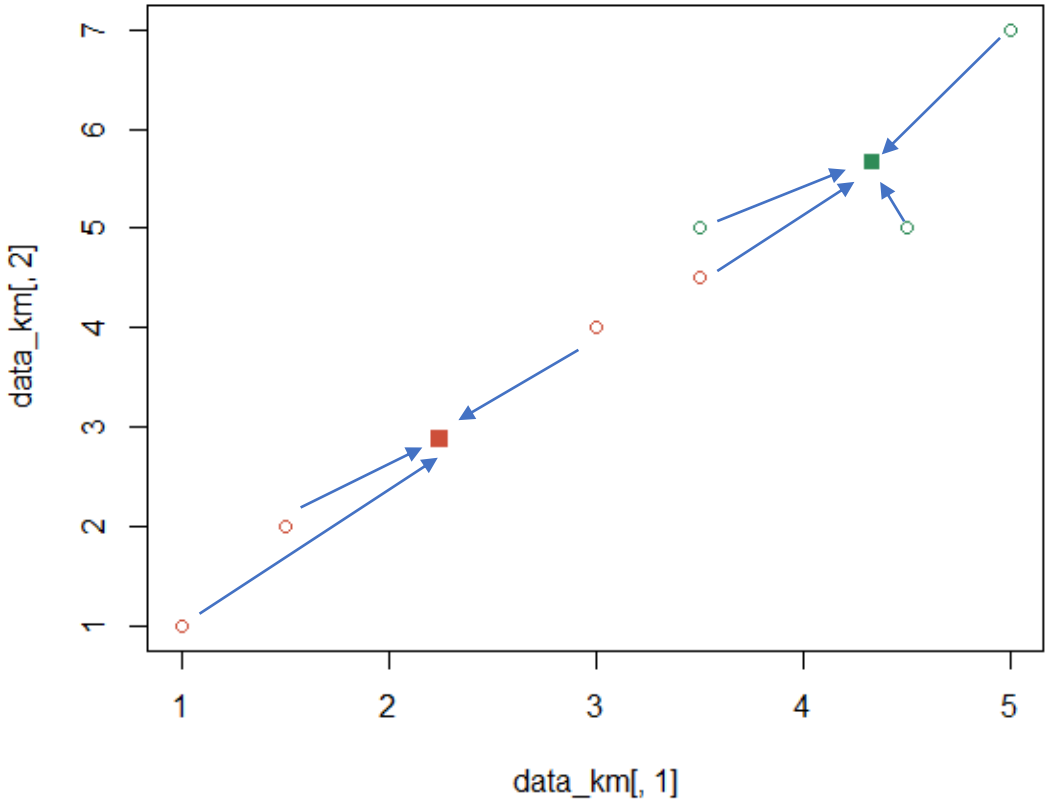
- 4. 1번의 중심값과 3번에서 선택된 중심값이 변화가 없다면 멈춘다.
  만약 데이터의 변화가 있다면 다시 1번부터 반복한다.

반복 3회



Goal 2. k-means 군집분석의 원리를 알아보자

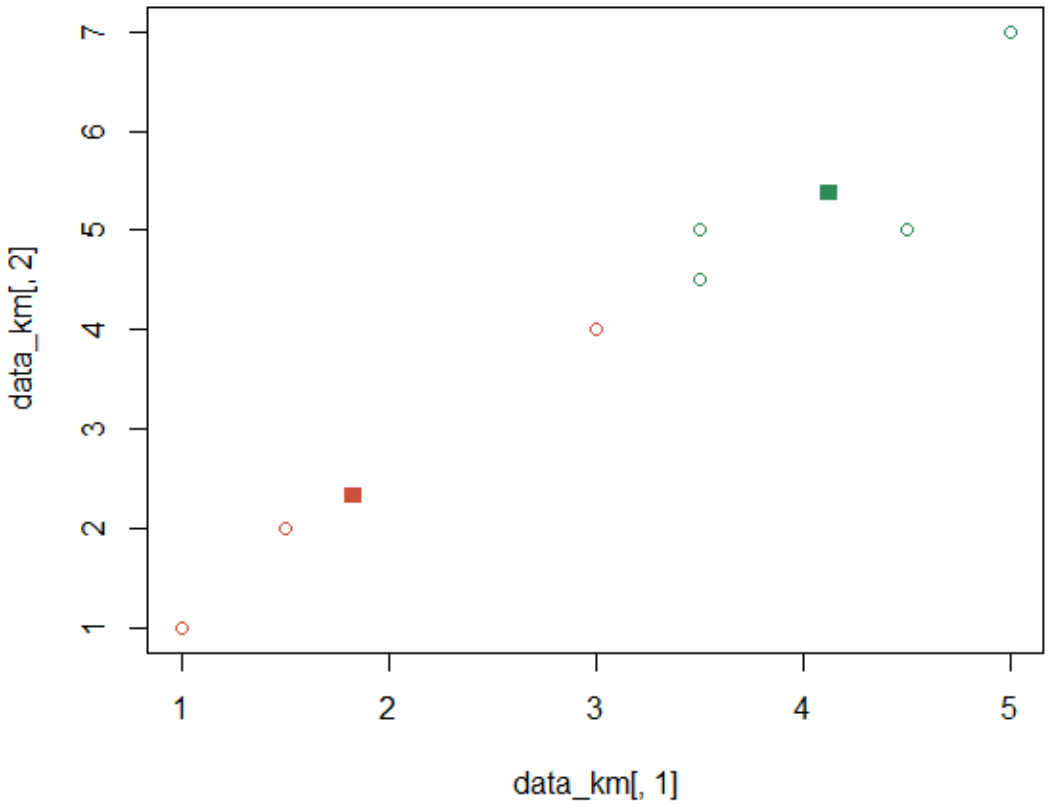
- 4. 1번의 중심값과 3번에서 선택된 중심값이 변화가 없다면 멈춘다.
  만약 데이터의 변화가 있다면 다시 1번부터 반복한다.



Goal 2. k-means 군집분석의 원리를 알아보자

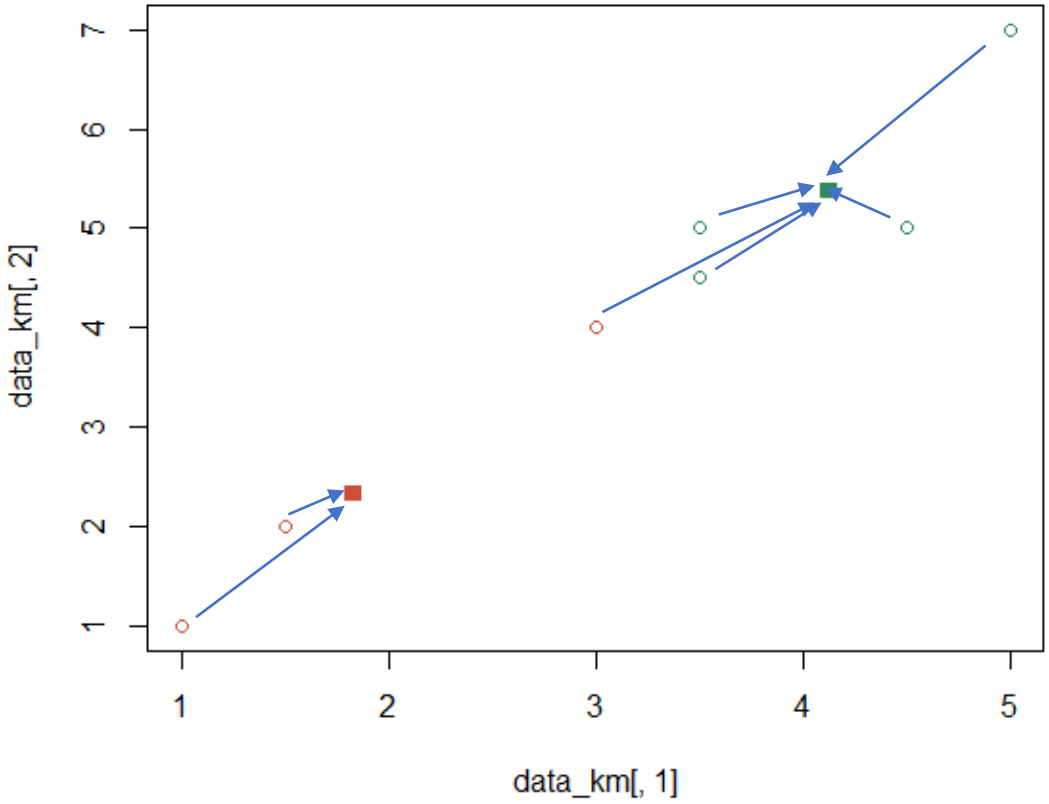
- 4. 1번의 중심값과 3번에서 선택된 중심값이 변화가 없다면 멈춘다.
  만약 데이터의 변화가 있다면 다시 1번부터 반복한다.

반복 4회



Goal 2. k-means 군집분석의 원리를 알아보자

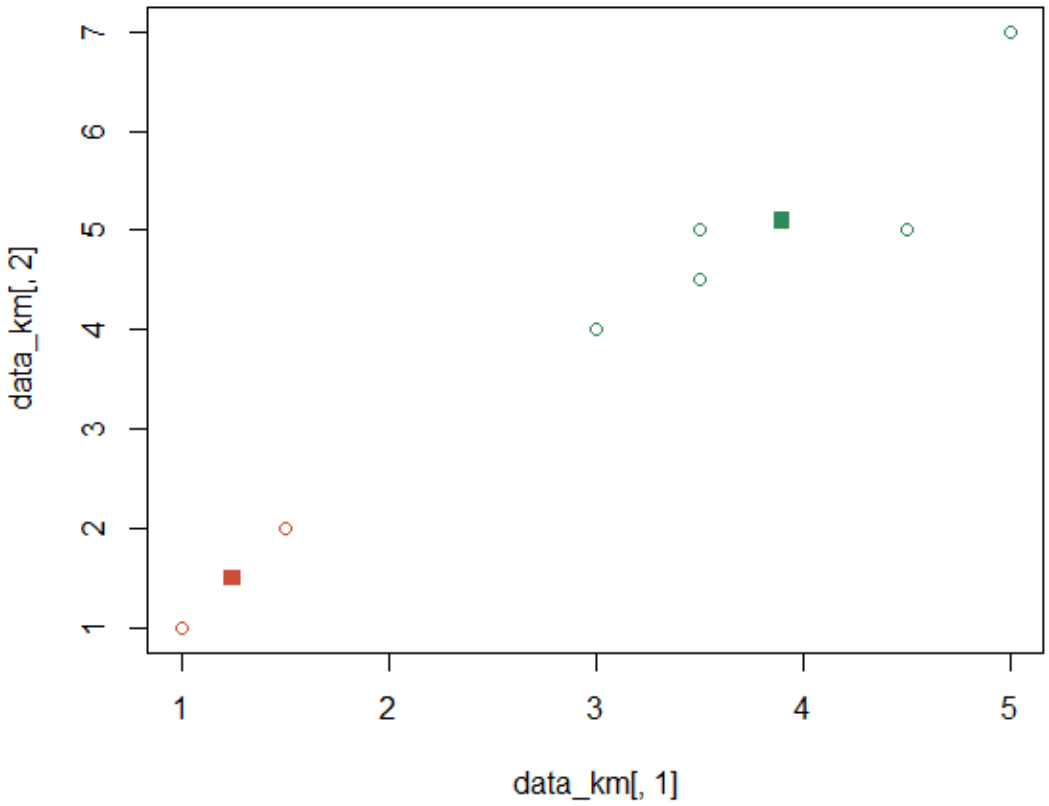
- 4. 1번의 중심값과 3번에서 선택된 중심값이 변화가 없다면 멈춘다.
  만약 데이터의 변화가 있다면 다시 1번부터 반복한다.



Goal 2. k-means 군집분석의 원리를 알아보자

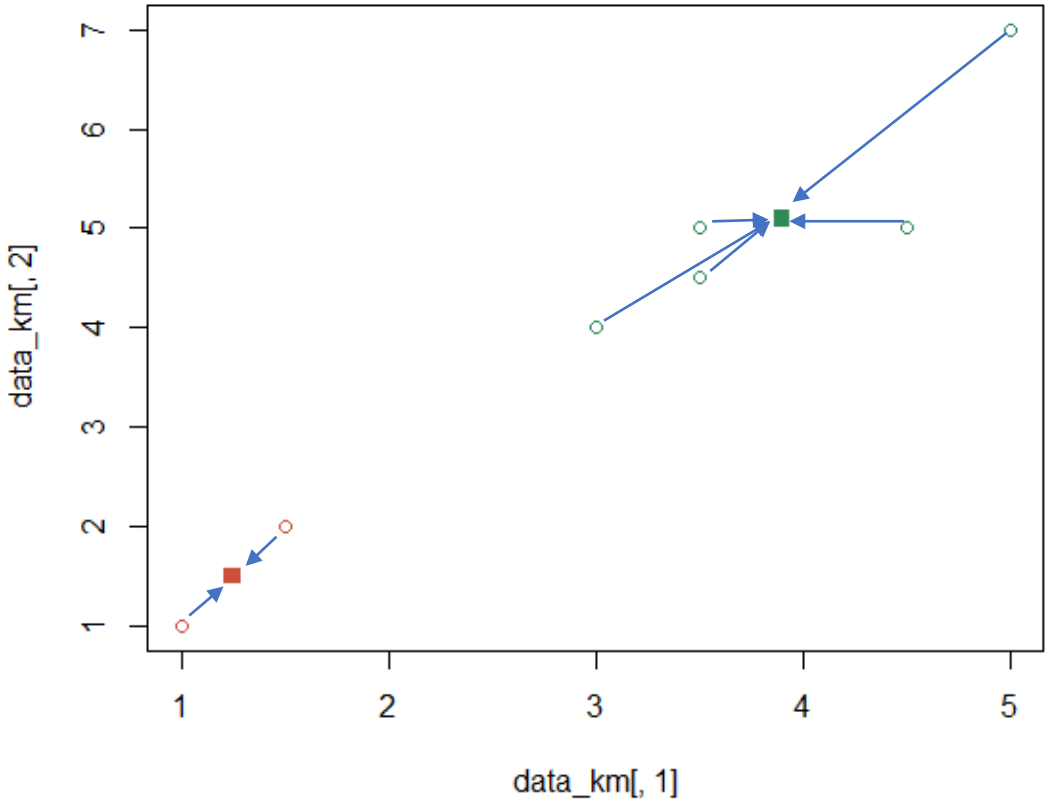
- 4. 1번의 중심값과 3번에서 선택된 중심값이 변화가 없다면 멈춘다.
  만약 데이터의 변화가 있다면 다시 1번부터 반복한다.

반복 5회



Goal 2. k-means 군집분석의 원리를 알아보자

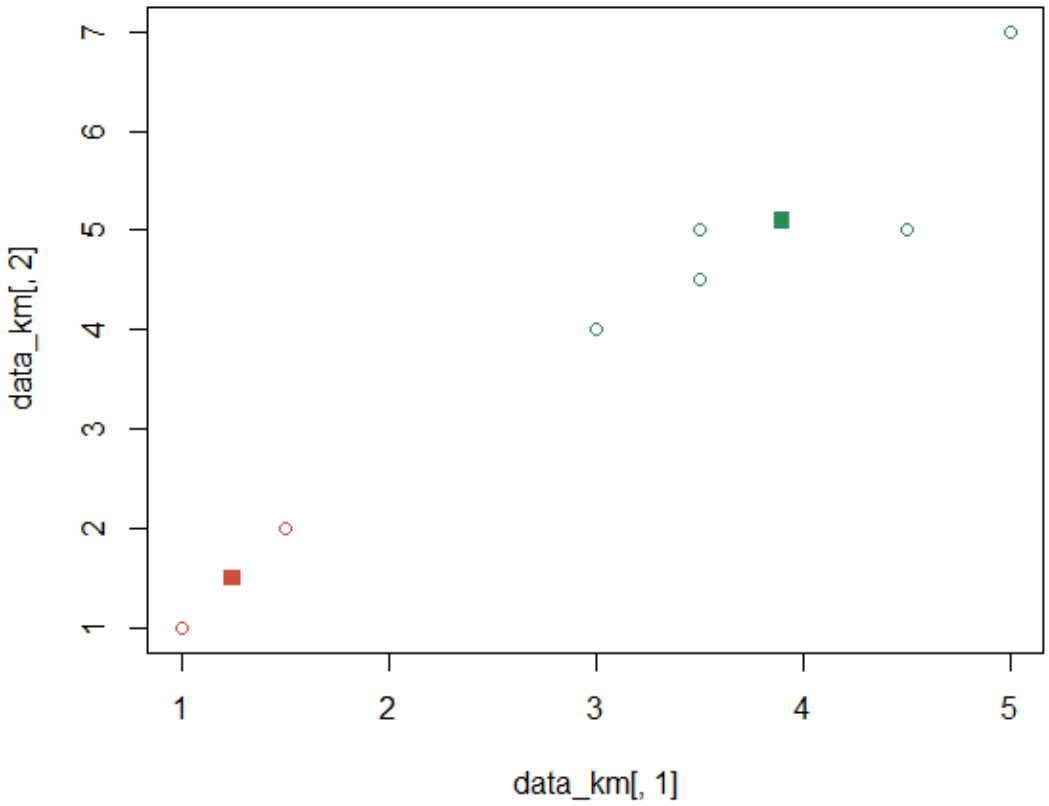
- 4. 1번의 중심값과 3번에서 선택된 중심값이 변화가 없다면 멈춘다.
  만약 데이터의 변화가 있다면 다시 1번부터 반복한다.



Goal 2. k-means 군집분석의 원리를 알아보자

- 4. 1번의 중심값과 3번에서 선택된 중심값이 변화가 없다면 멈춘다.
  만약 데이터의 변화가 있다면 다시 1번부터 반복한다.

반복 6회: 종료

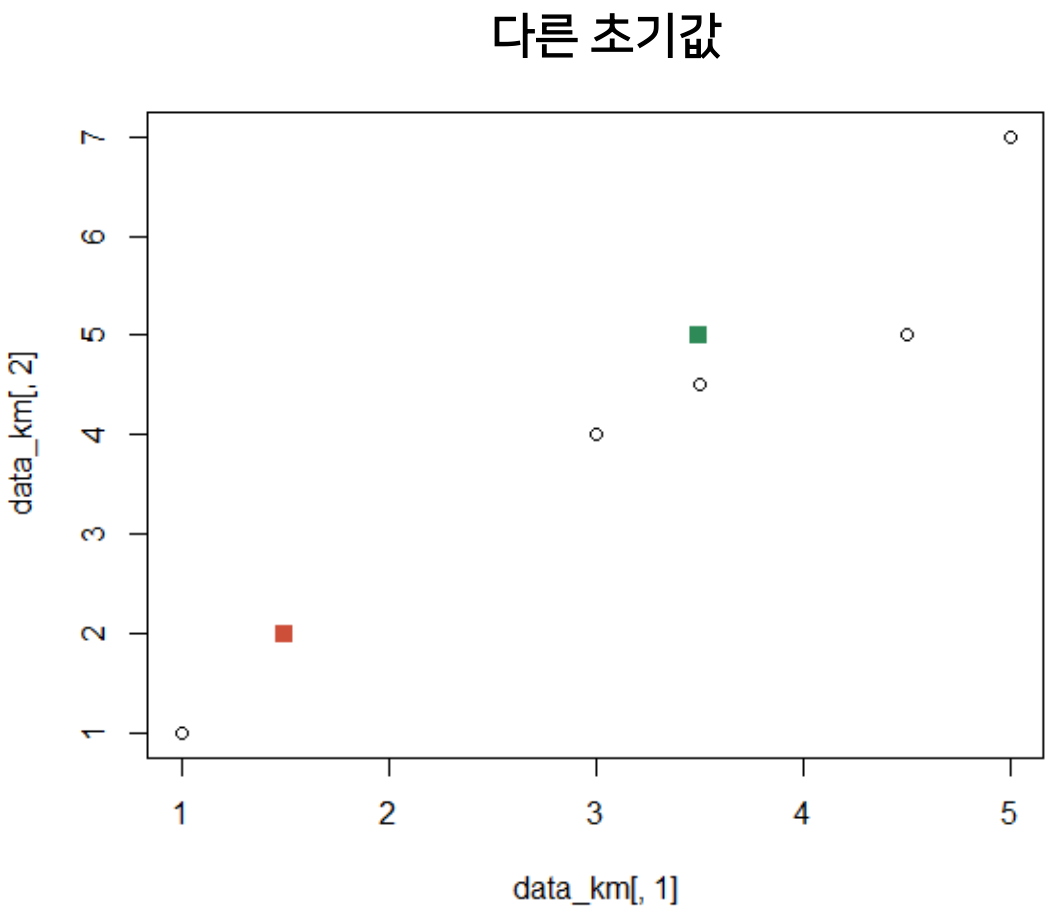


Goal 2. k-means 군집분석의 원리를 알아보자

- 5. k-means 군집분석은 초기값에 따라 수렴속도가 달라질 수 있다.

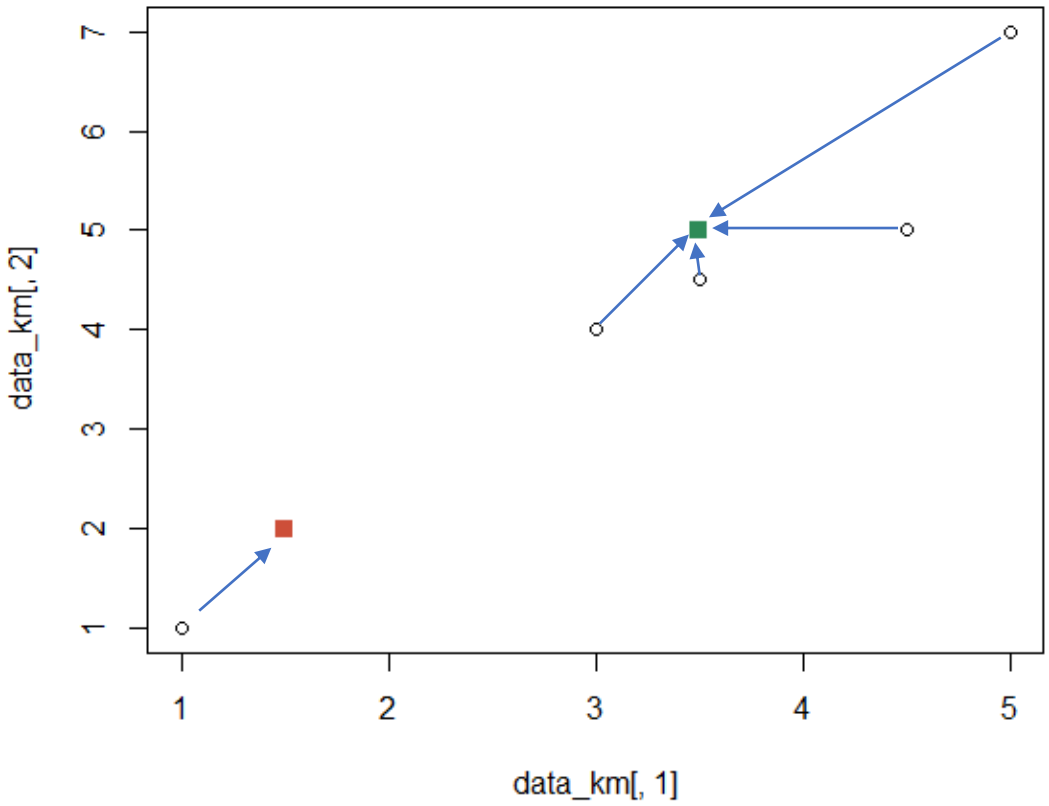
Goal 2. k-means 군집분석의 원리를 알아보자

- 5. k-means 군집분석은 초기값에 따라 수렴속도가 달라질 수 있다.



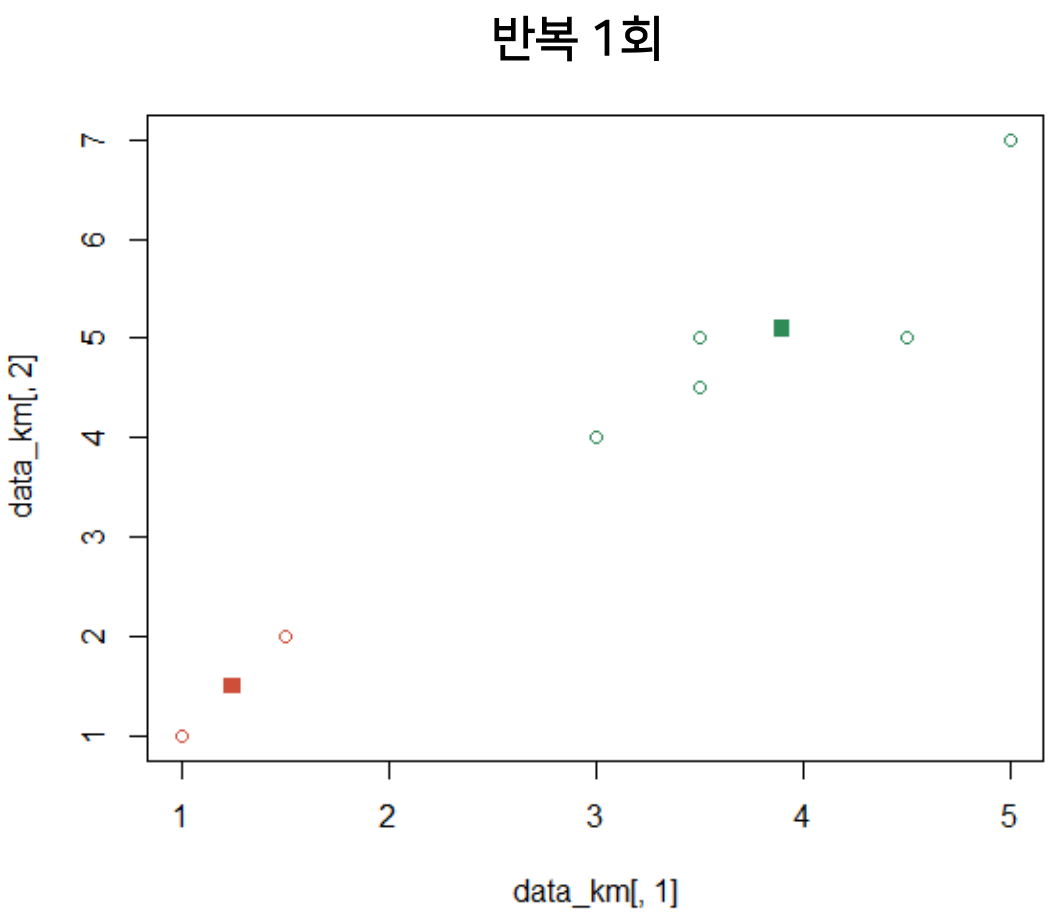
Goal 2. k-means 군집분석의 원리를 알아보자

- 5. k-means 군집분석은 초기값에 따라 수렴속도가 달라질 수 있다.



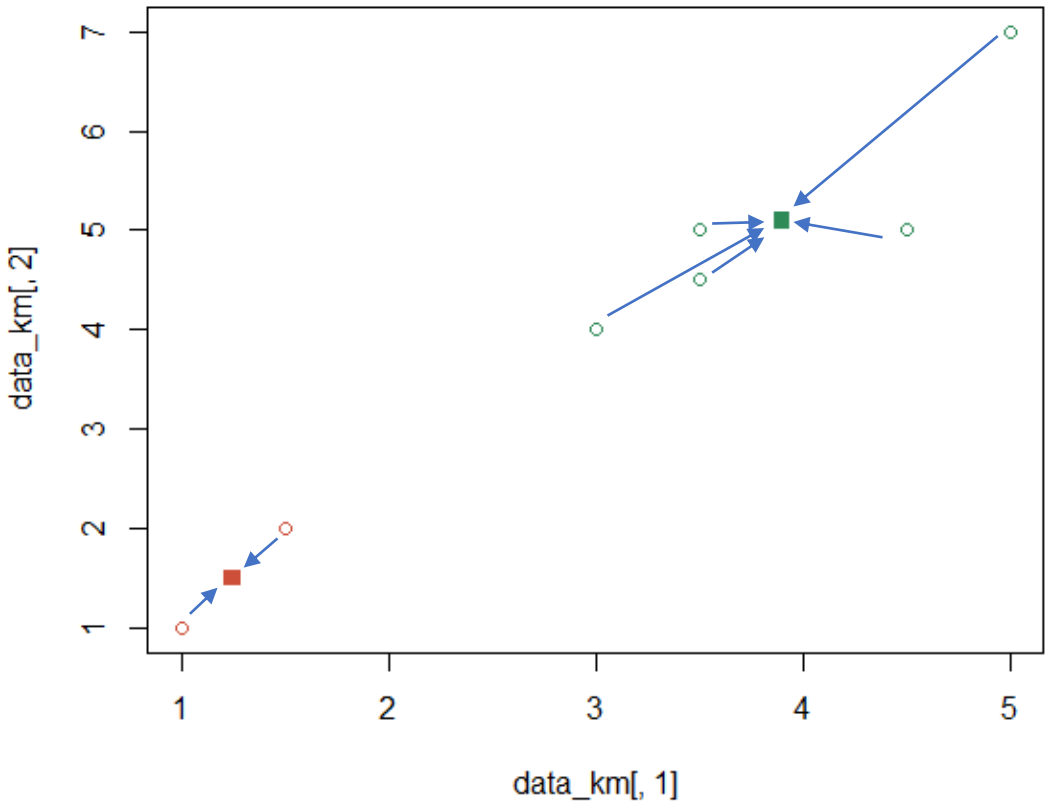
Goal 2. k-means 군집분석의 원리를 알아보자

- 5. k-means 군집분석은 초기값에 따라 수렴속도가 달라질 수 있다.



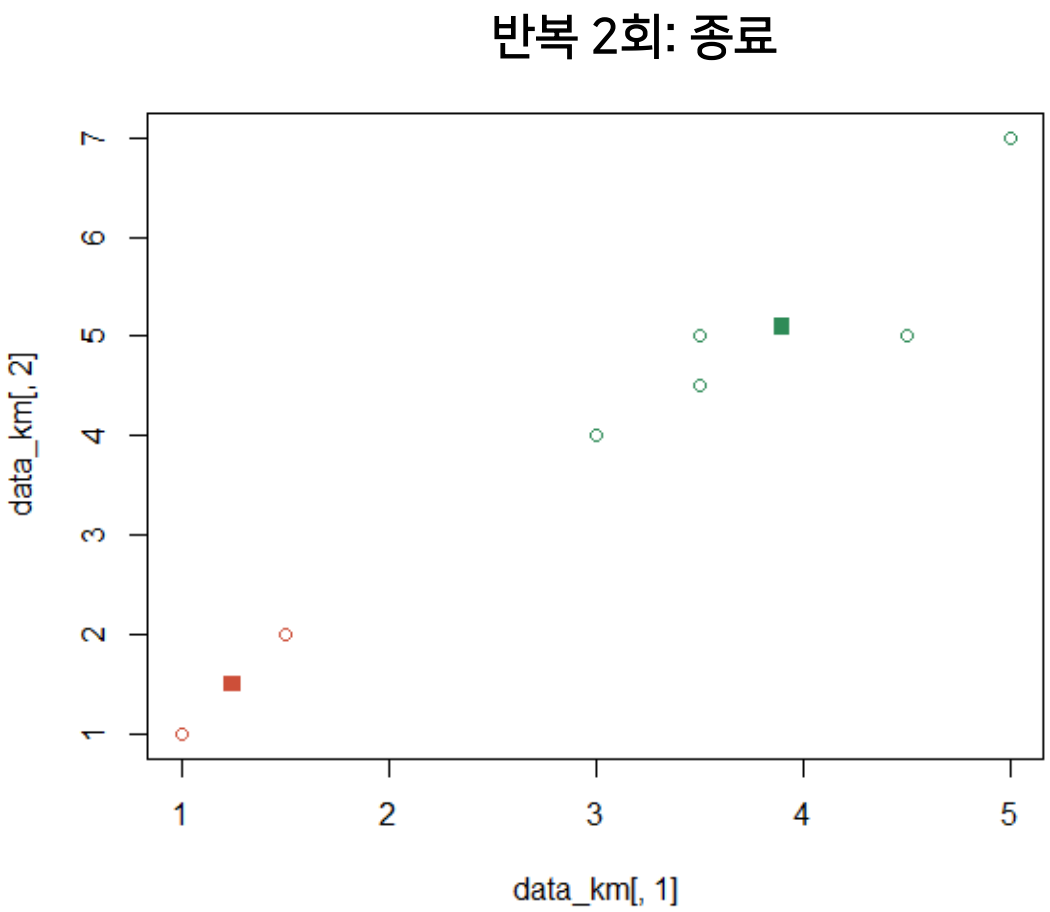
Goal 2. k-means 군집분석의 원리를 알아보자

- 5. k-means 군집분석은 초기값에 따라 수렴속도가 달라질 수 있다.



Goal 2. k-means 군집분석의 원리를 알아보자

- 5. k-means 군집분석은 초기값에 따라 수렴속도가 달라질 수 있다.

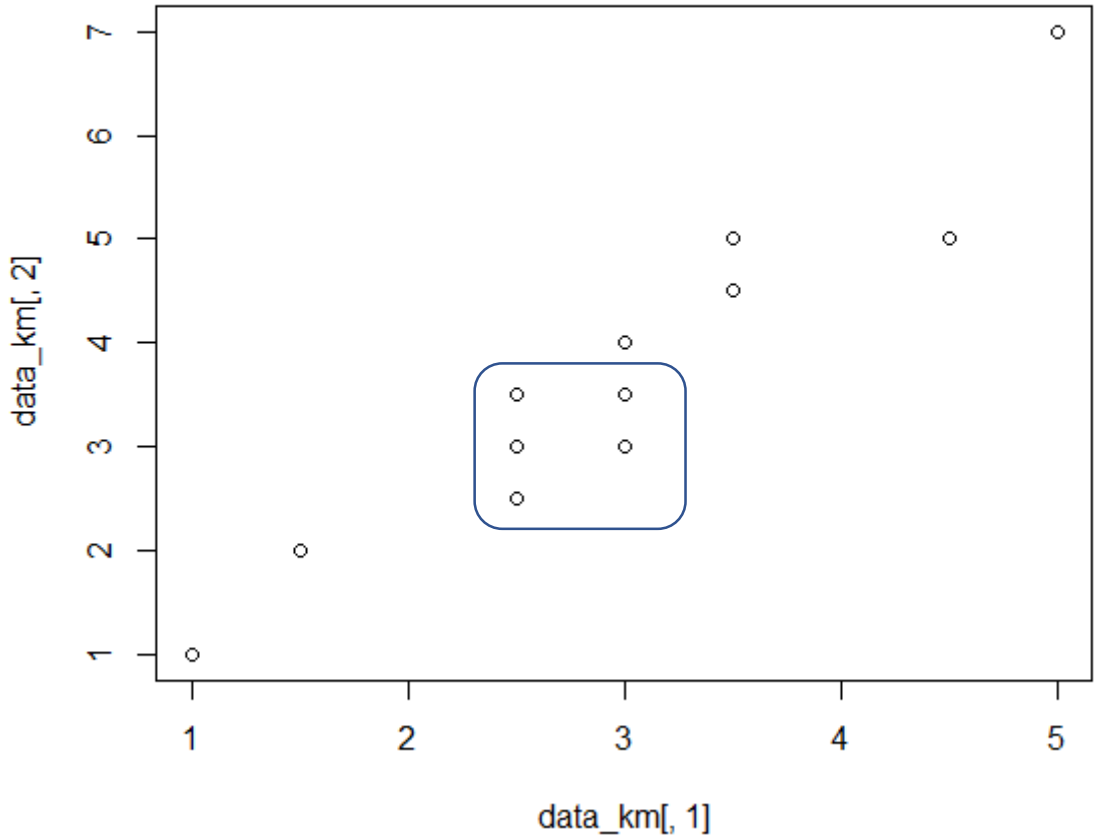


Goal 2. k-means 군집분석의 원리를 알아보자

- 6. 데이터가 정확하게 분리되기 힘든 경우(비슷한 위치에 모여 있는 경우) 초기값에 따라 분석 결과가 달라질 수 있다.

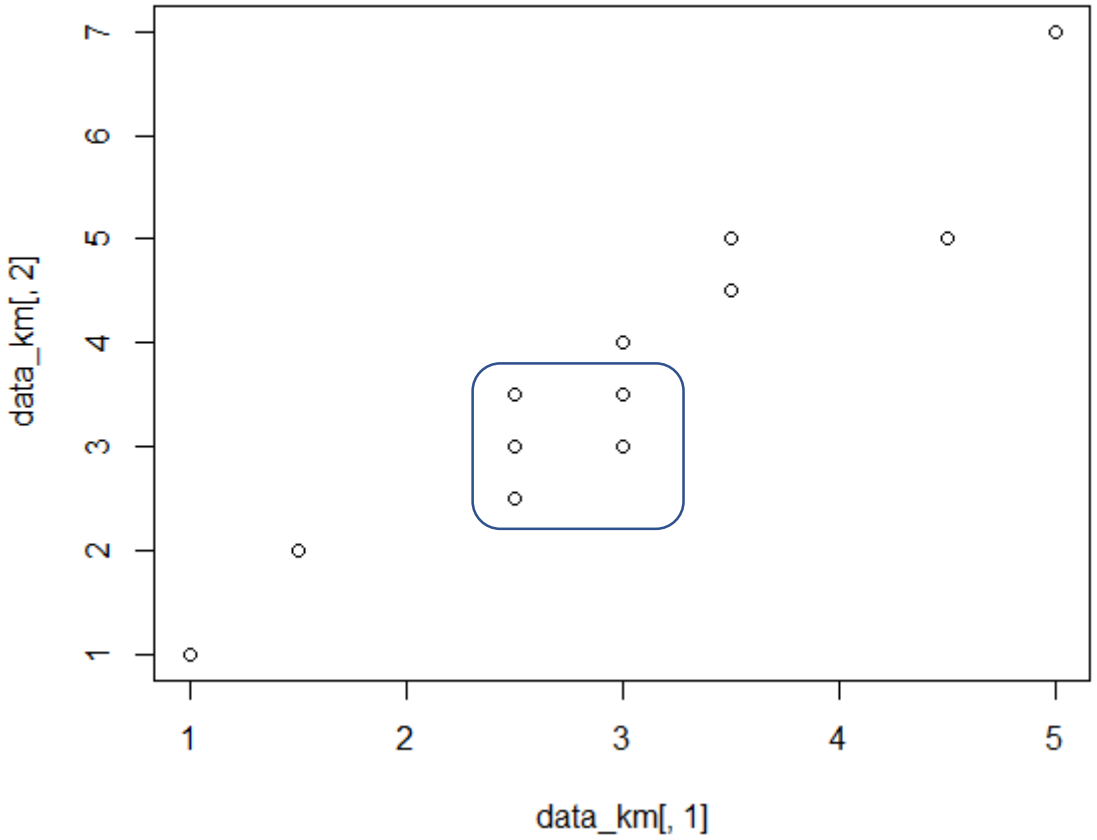
Goal 2. k-means 군집분석의 원리를 알아보자

- 6. 데이터가 정확하게 분리되기 힘든 경우(비슷한 위치에 모여 있는 경우) 초기값에 따라 분석 결과가 달라질 수 있다.



Goal 2. k-means 군집분석의 원리를 알아보자

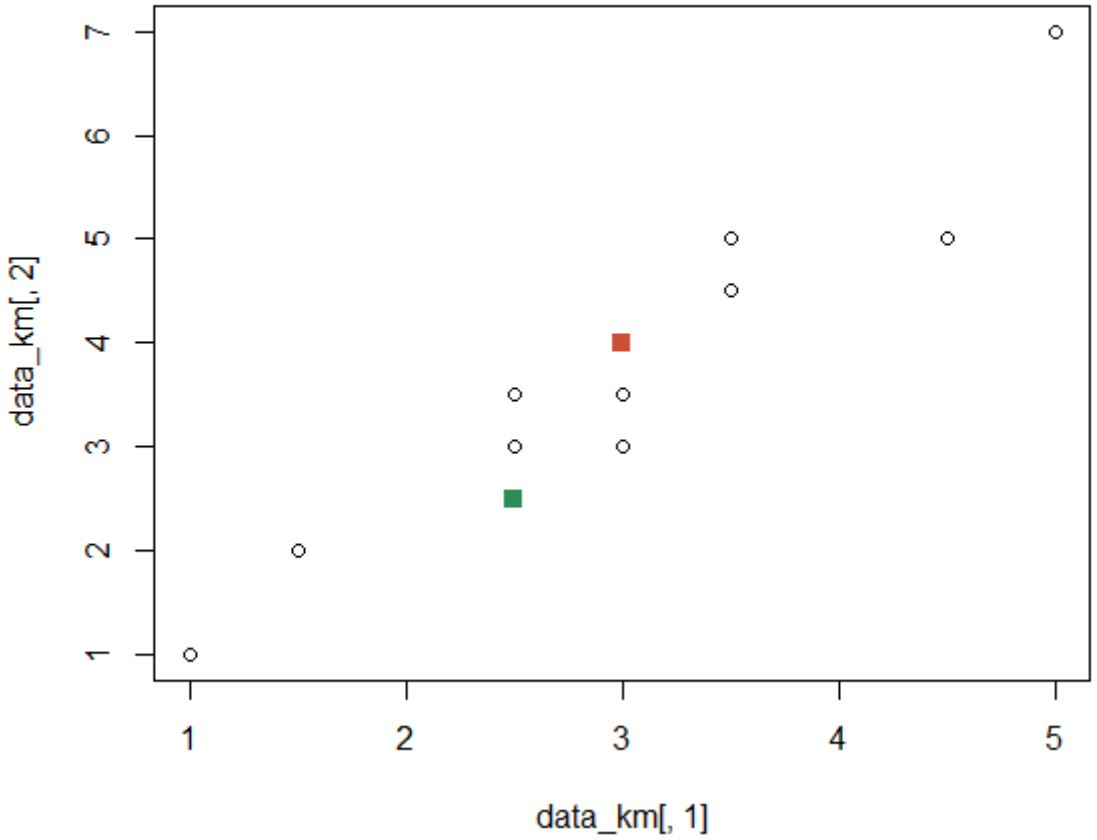
- (앞 예시의 데이터에서 총 5개의 점을 추가하였다.
  x=2.5일 때 y=2.5, 3.0, 3.5
  x=3.0일 때 y=3.0, 3.5)



Goal 2. k-means 군집분석의 원리를 알아보자

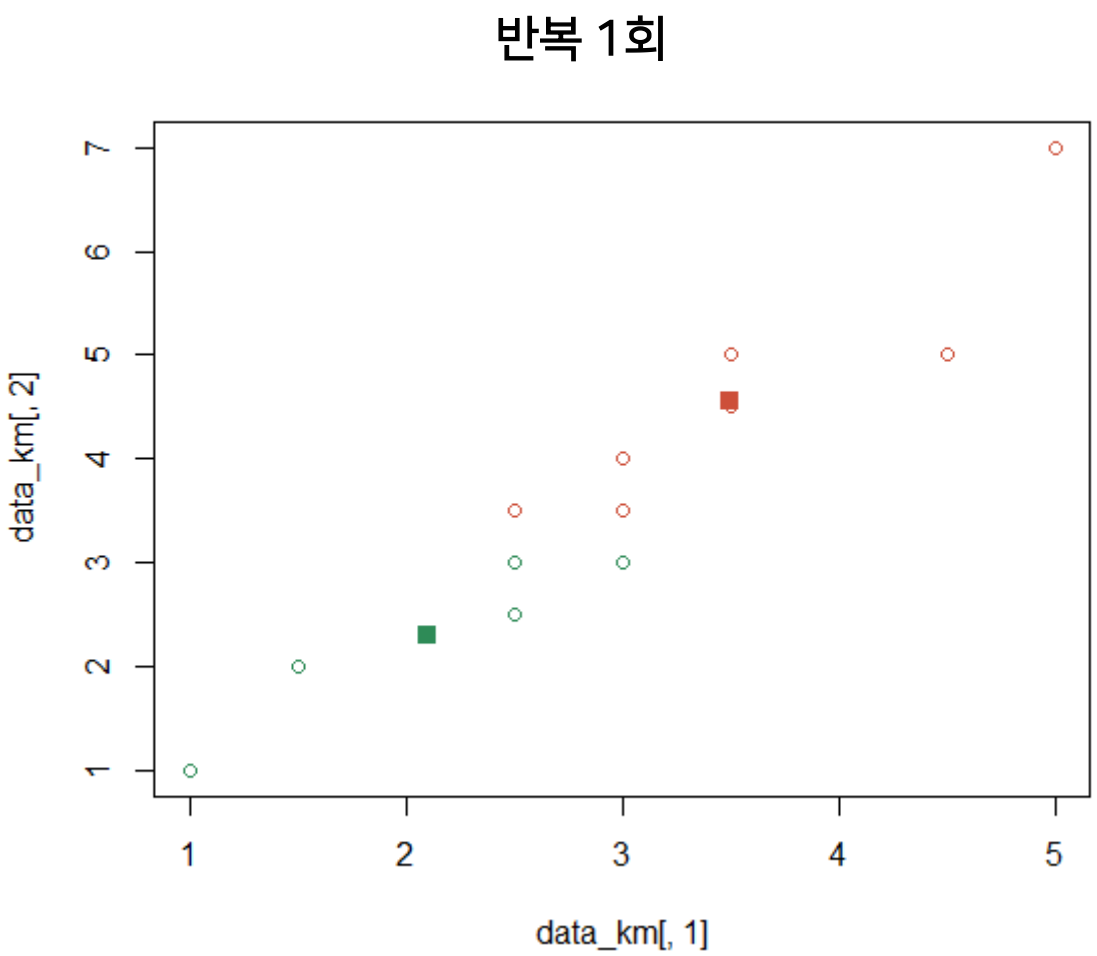
- 6. 데이터가 정확하게 분리되기 힘든 경우(비슷한 위치에 모여 있는 경우) 초기값에 따라 분석 결과가 달라질 수 있다.

초기값: 경우 1



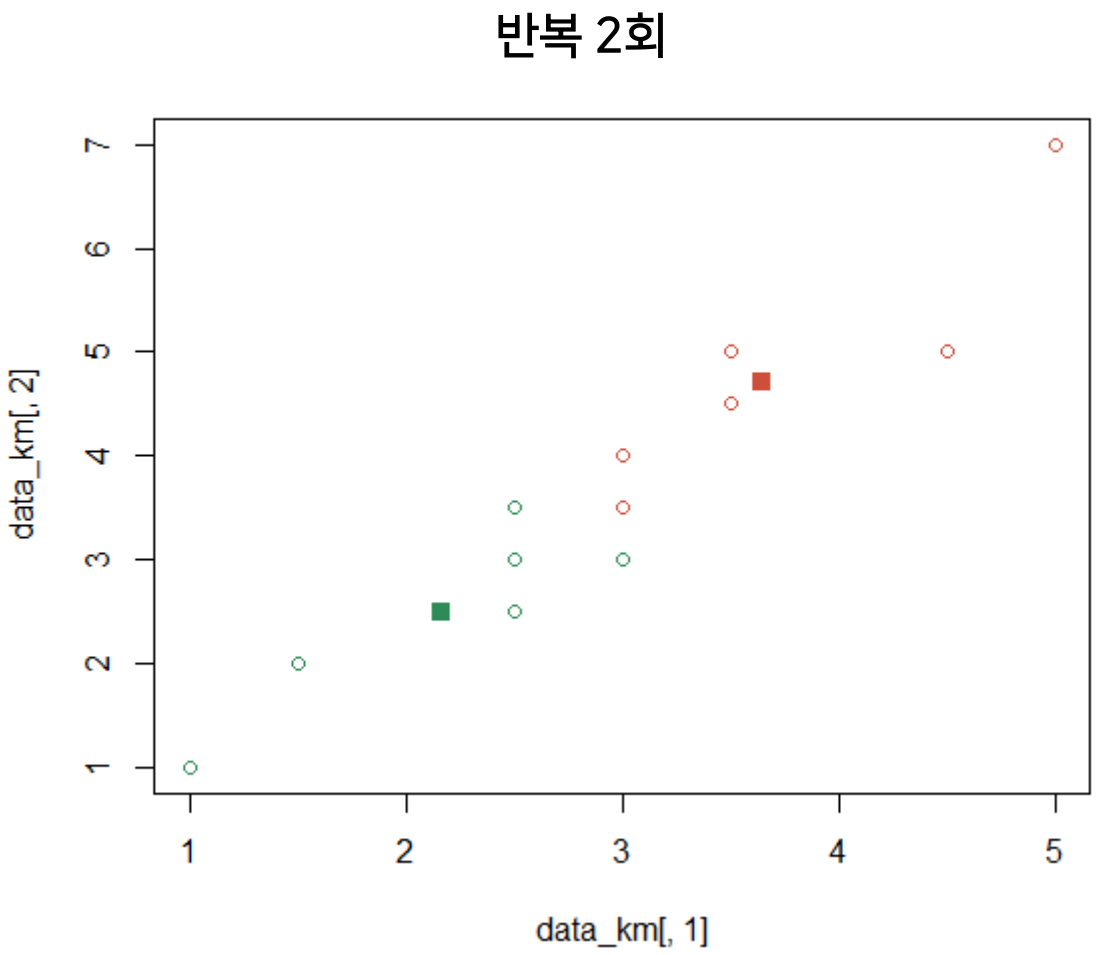
Goal 2. k-means 군집분석의 원리를 알아보자

- 6. 데이터가 정확하게 분리되기 힘든 경우(비슷한 위치에 모여 있는 경우) 초기값에 따라 분석 결과가 달라질 수 있다.



Goal 2. k-means 군집분석의 원리를 알아보자

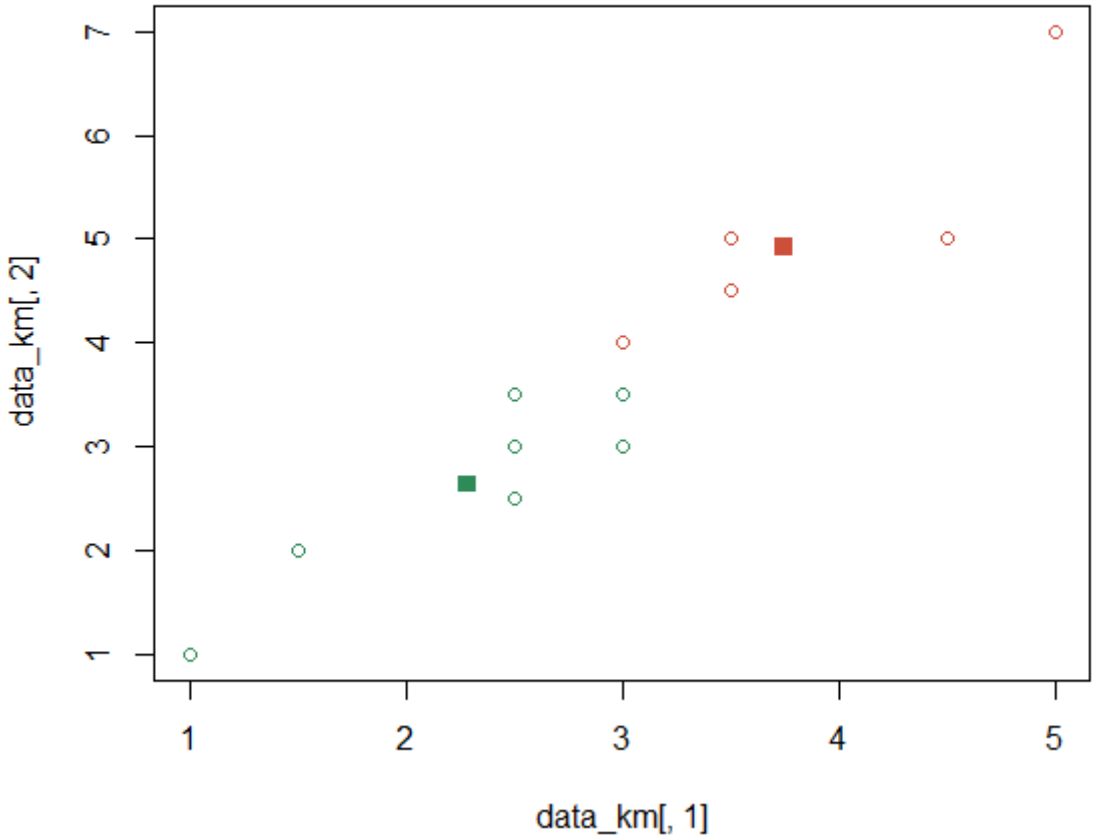
- 6. 데이터가 정확하게 분리되기 힘든 경우(비슷한 위치에 모여 있는 경우) 초기값에 따라 분석 결과가 달라질 수 있다.



Goal 2. k-means 군집분석의 원리를 알아보자

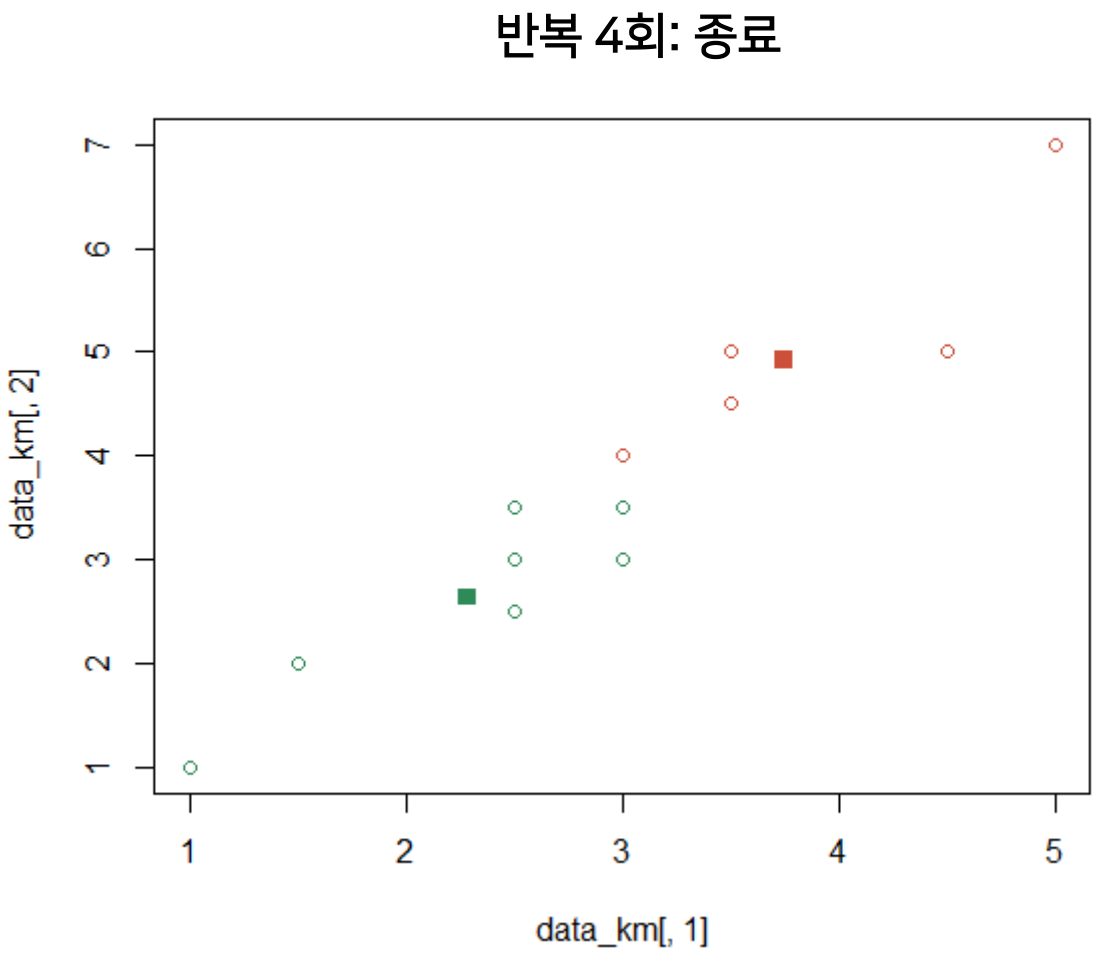
- 6. 데이터가 정확하게 분리되기 힘든 경우(비슷한 위치에 모여 있는 경우) 초기값에 따라 분석 결과가 달라질 수 있다.

반복 3회



Goal 2. k-means 군집분석의 원리를 알아보자

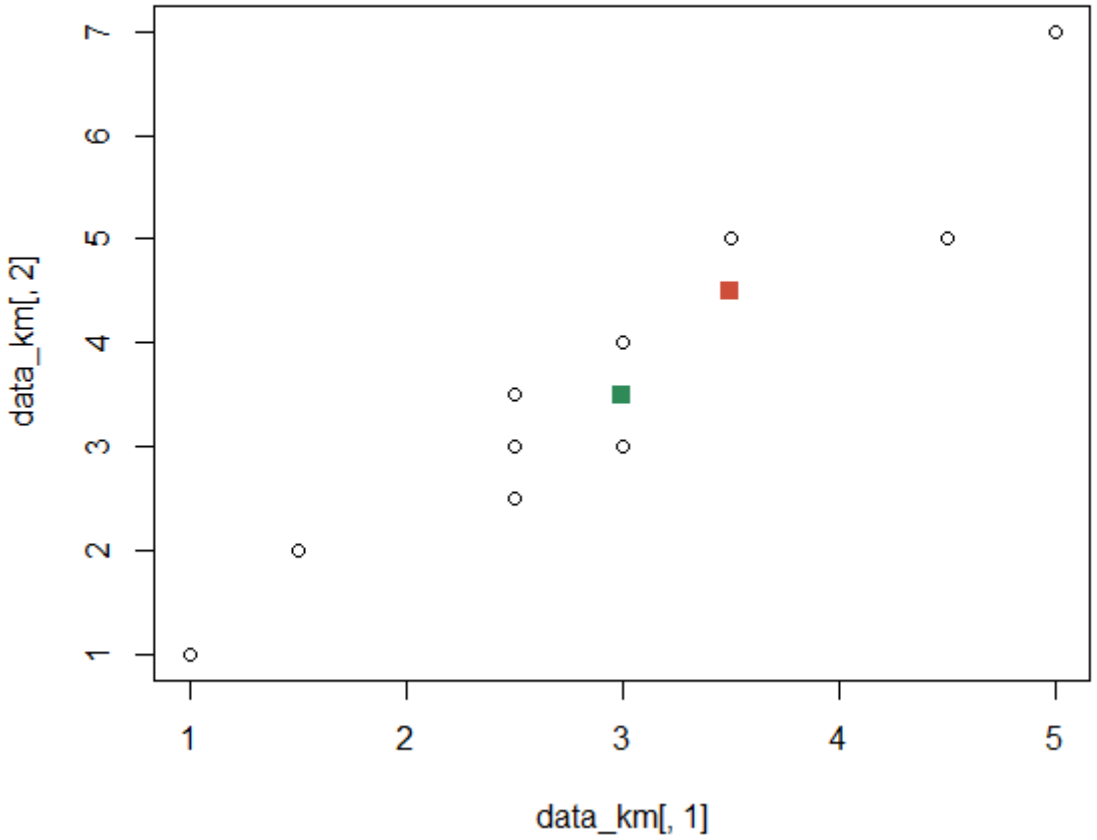
- 6. 데이터가 정확하게 분리되기 힘든 경우(비슷한 위치에 모여 있는 경우) 초기값에 따라 분석 결과가 달라질 수 있다.



Goal 2. k-means 군집분석의 원리를 알아보자

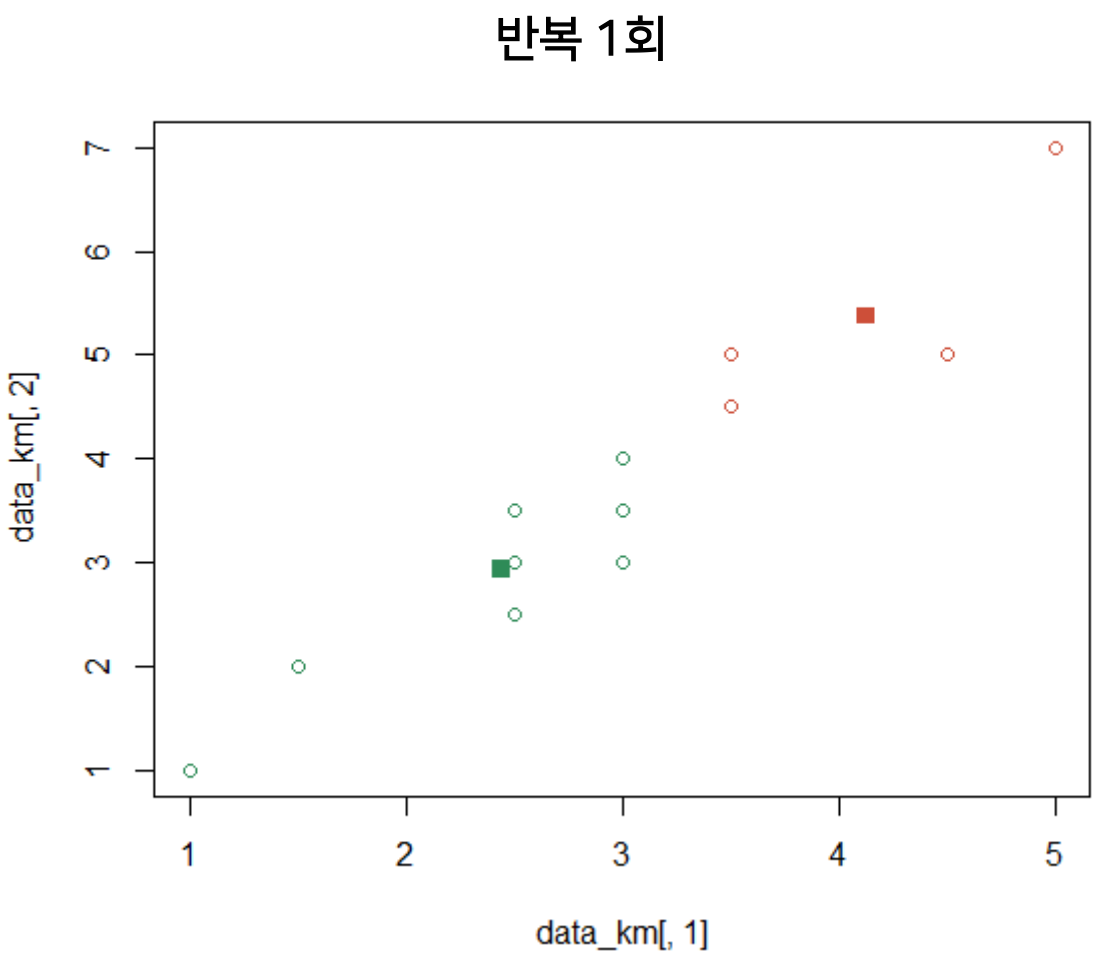
- 6. 데이터가 정확하게 분리되기 힘든 경우(비슷한 위치에 모여 있는 경우) 초기값에 따라 분석 결과가 달라질 수 있다.

초기값: 경우 2



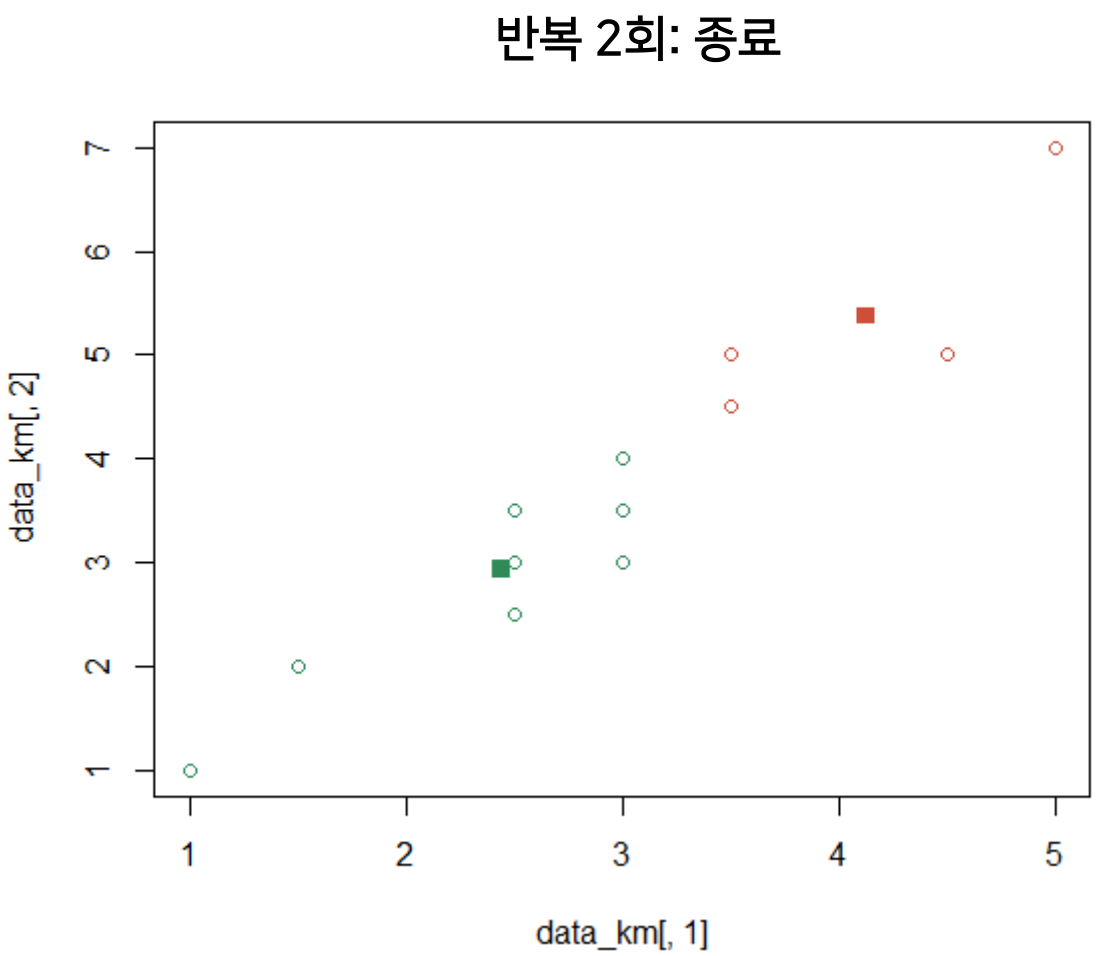
Goal 2. k-means 군집분석의 원리를 알아보자

- 6. 데이터가 정확하게 분리되기 힘든 경우(비슷한 위치에 모여 있는 경우) 초기값에 따라 분석 결과가 달라질 수 있다.



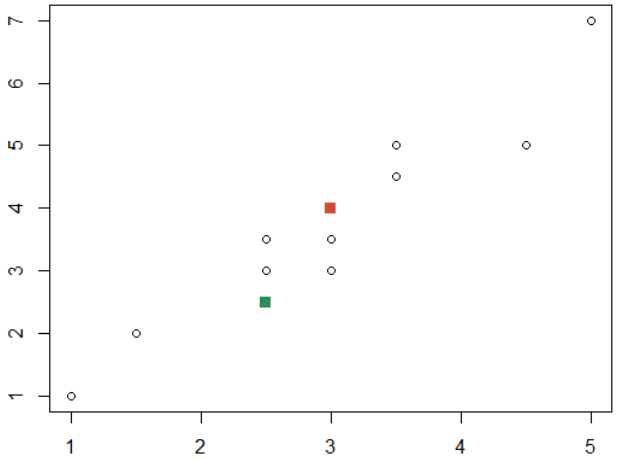
Goal 2. k-means 군집분석의 원리를 알아보자

- 6. 데이터가 정확하게 분리되기 힘든 경우(비슷한 위치에 모여 있는 경우) 초기값에 따라 분석 결과가 달라질 수 있다.

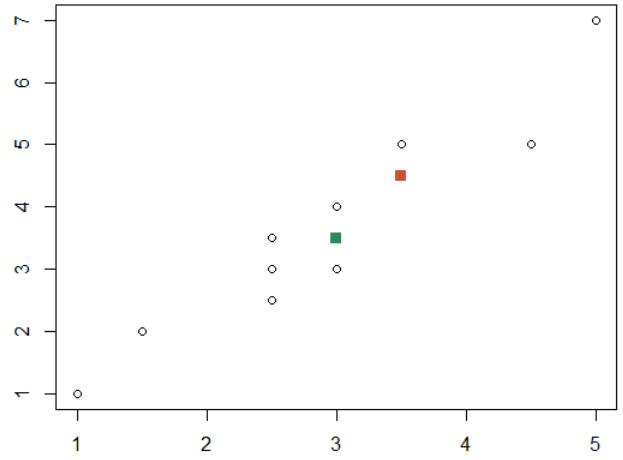


Goal 2. k-means 군집분석의 원리를 알아보자

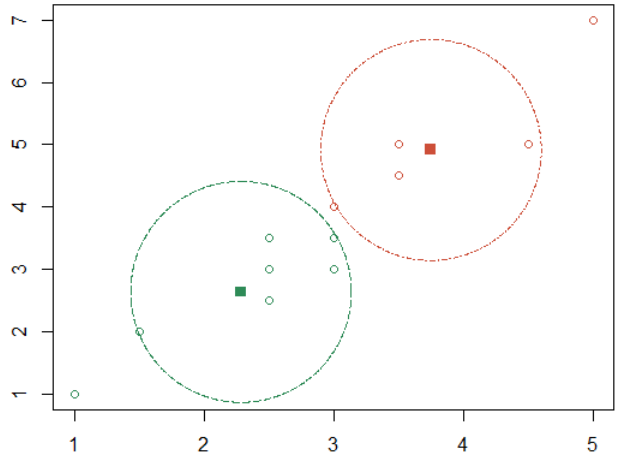
- 6. 데이터가 정확하게 분리되기 힘든 경우(비슷한 위치에 모여 있는 경우) 초기값에 따라 분석 결과가 달라질 수 있다.

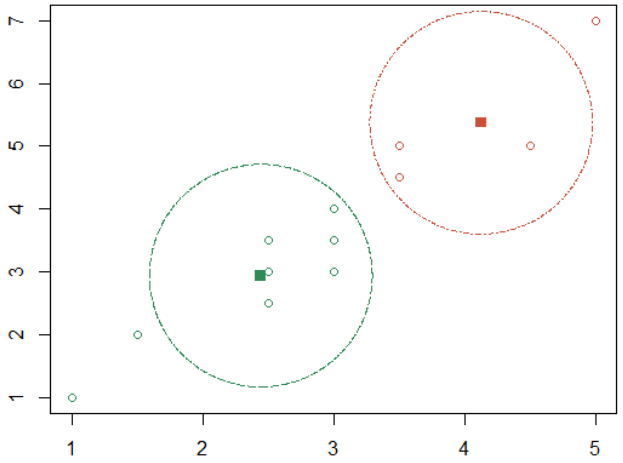


초기값: 경우 1



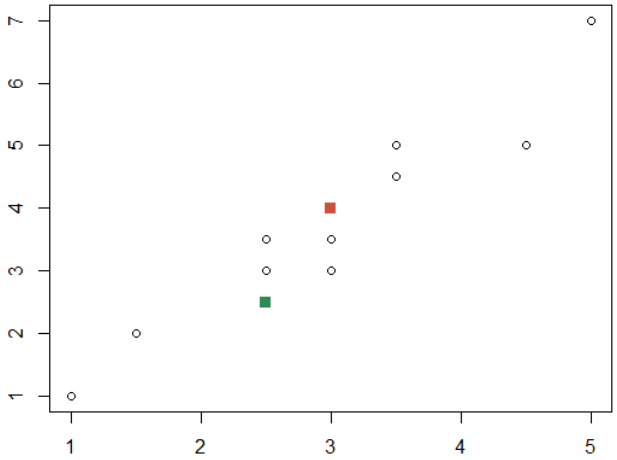
초기값: 경우 2



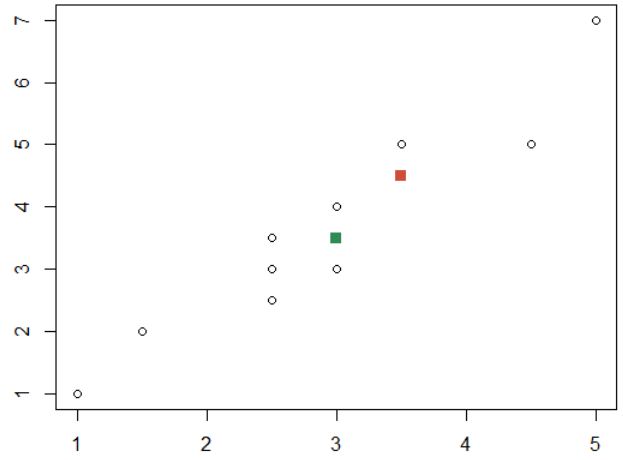


Goal 2. k-means 군집분석의 원리를 알아보자

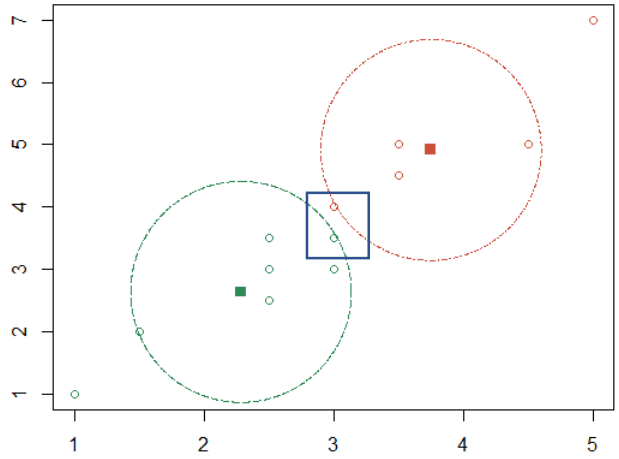
- 6. 데이터가 정확하게 분리되기 힘든 경우(비슷한 위치에 모여 있는 경우) 초기값에 따라 분석 결과가 달라질 수 있다.

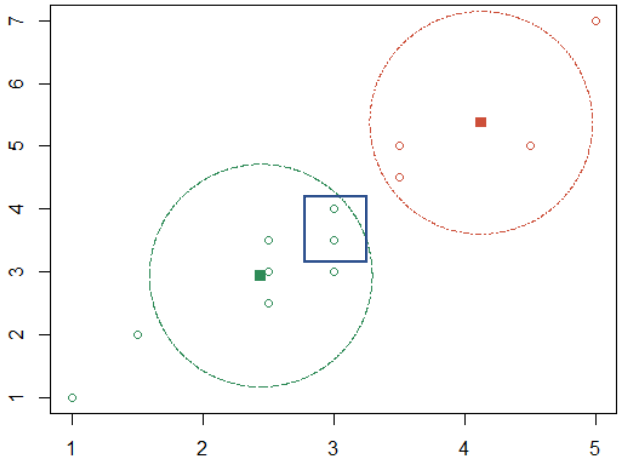


초기값: 경우 1



초기값: 경우 2





Goal 2. k-means 군집분석의 원리를 알아보자

k-means 군집분석의 수행절차

1. k값을 초기값으로 먼저 받고, k개의 초기 군집의 임의 중심점을 설정
2. 각 개체와 군집 중심점 사이의 거리를 계산
3. 가장 가까운 중심점 군집으로 재할당
4. 변경된 군집을 기준으로 개체와 군집 중심점 사이의 거리를 다시 계산
5. 3 ~ 4단계를 군집의 변동이 없을 때까지 반복

3

적정 K의 값

Goal 3. 적정 k의 값은 어떻게 구하는가?

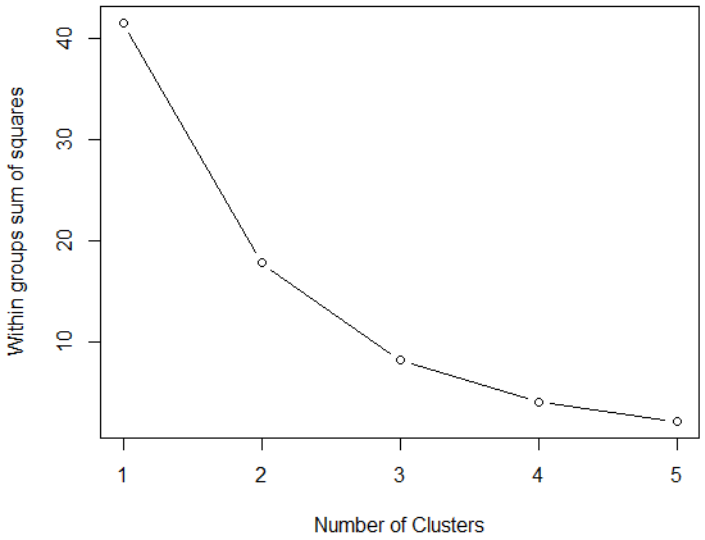
k-means 군집 분석에서 군집 수(k) 결정 방법 (1/2)

- 분석가 결정 : 데이터에 대한 **사전 정보가 있거나** 분석 목표에 의해 군집 개수가 **확실히 정해진** 경우
- 일반적으로 k-means 군집 분석을 하는 상황에서는 k를 미리 결정하기가 힘들다.

Goal 3. 적정 k의 값은 어떻게 구하는가?

엘보우(elbow) 기법

여러 가지 k값에 대한 군집 분석 결과의 군집 내에 얼마나 밀집되어 있는가를 나타내는 통계량의 합을 기준으로 **크기의 변화가 작아지는 지점**을 적정 군집 수로 결정하는 방법.

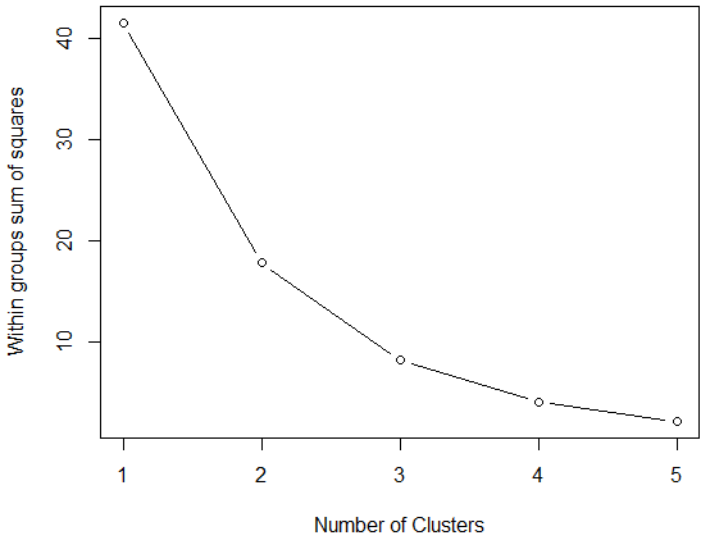


Goal 3. 적정 k의 값은 어떻게 구하는가?

Within group sum of square

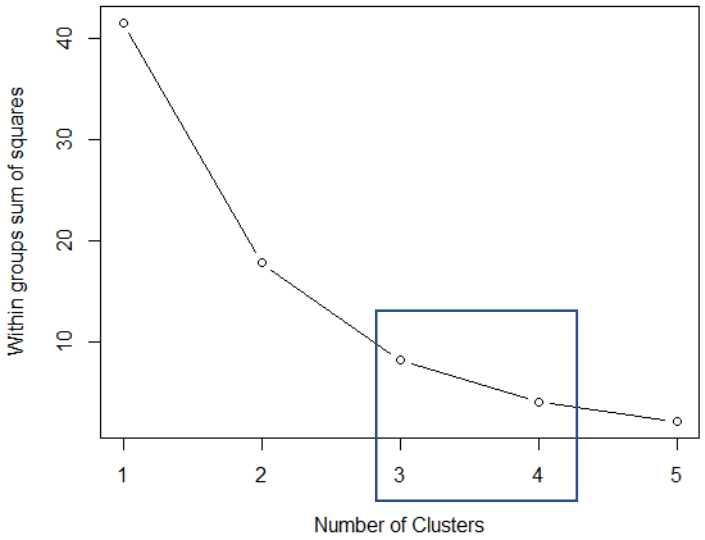
- 군집 내 동질성(같은 군집 내에서의 분산)을 나타내는 측도
- 개별 데이터별로 중심(centroid)과의 거리를 계산해 모두 더한 것

  -> 군집화가 잘 됐으면 데이터들이 중심 근처에 모여 있을 것이므로 값이 작을 것이다.



Goal 3. 적정 k의 값은 어떻게 구하는가?

- 군집 수 k가 증가하면 within group sum of square는 항상 감소한다.
  하지만 무작정 k를 늘리는 것은 좋지 않다.
- k가 증가했는데 within group sum of square가 별로 차이 나지 않는다면
  (기울기가 **급격하게 완만**해진다면) 그 부분이 elbow point가 된다.

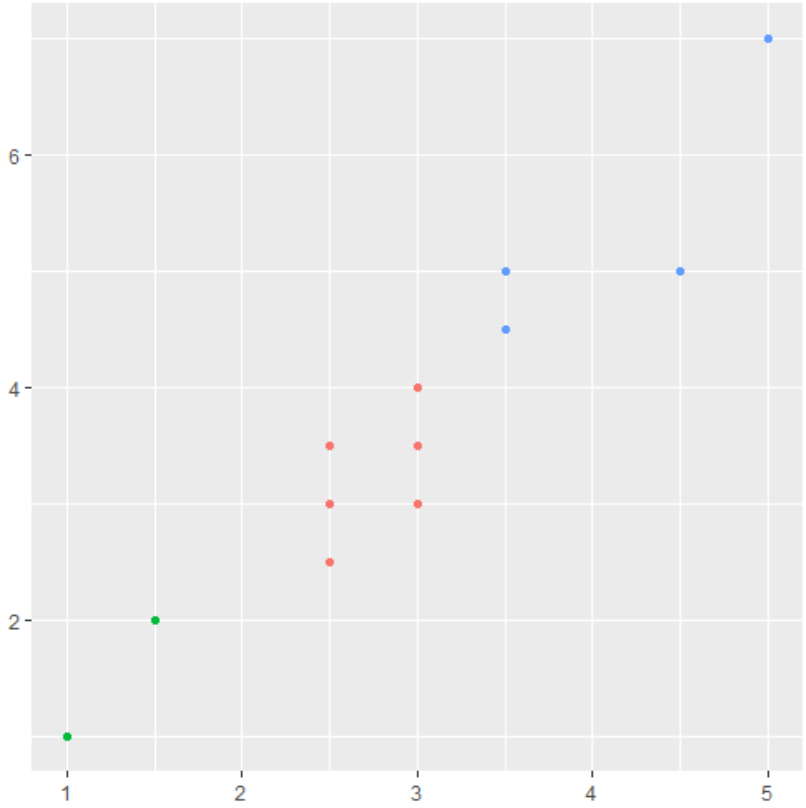


세분화 예시

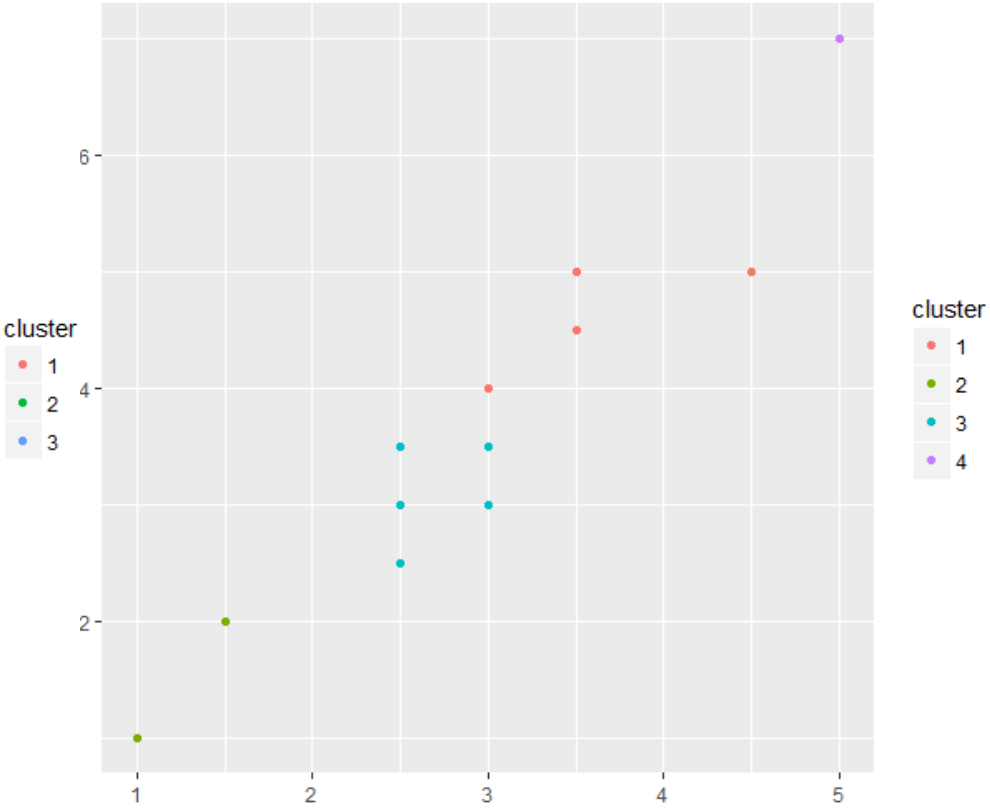
좌측 예시의 경우, 군집 수를 3~4개로
하는게 적정하다고 판단됨
군집 수 결정시, 군집 수를 매우 큰 값으로 설정
하면 군집 내 동질성은 향상되지만
데이터에 대한 **과적합**이 될 위험이 있음

Goal 3. 적정 k의 값은 어떻게 구하는가?

- k=4와 같이 적정 군집 수가 정해졌다고 하더라도 비슷한 군집 수인 k=3,5 등의 경우에 대해 비교해보고 군집의 특성을 확인해보는 것이 바람직하다.



K=3



K=4

k-means 군집 분석에서 군집 수(k) 결정 방법 (2/2)

- 실루엣 계수(Silhouette)

  군집 내 동질적인 정도와 군집 간 이질적인 정도를 수치화한 값
- 실루엣 계수가 1에 가까울수록 군집화가 **잘 된 것**,

  -1에 가까울수록 제대로 되지 않은 것

- 실루엣 계수의 정의 자체는 개별 데이터에 대해 각각 계산,

  보통 실루엣 계수를 평균내서 평균 수치를 보고 군집화가 잘 됐는지 판단

Goal 3. 적정 k의 값은 어떻게 구하는가?

k-means 군집 분석에서 군집 수(k) 결정 방법 (2/2)

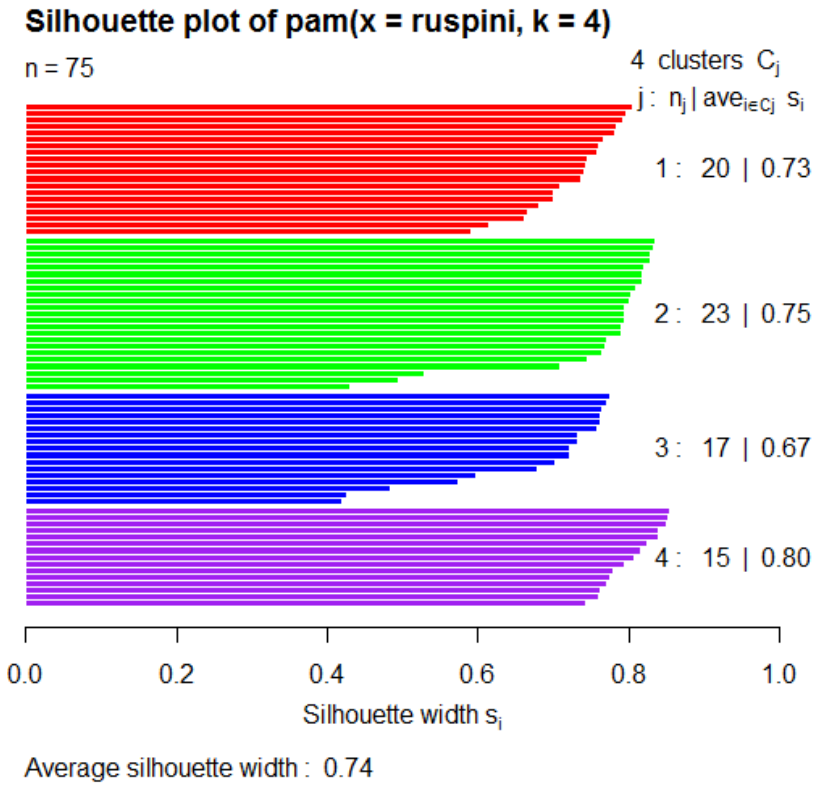
- 실루엣 계수: 군집 내 동질적인 정도와 군집 간 이질적인 정도를 수치화한 값
- 실루엣 계수가 1에 가까울수록 군집화가 잘 된 것, -1에 가까울수록 제대로 되지 않은 것
  - 개별 데이터 i에 대해, 군집 내 동질성을 의미하는 a(i)와 군집 간 이질성을 의미하는 b(i)로 구성
  - a(i): i가 속한 군집에 있는 다른 데이터들과의 평균 거리 -> a(i)가 작을수록 좋음
  - b(i): 가장 가까운 이웃 군집과의 평균 거리 -> b(i)가 클수록 좋음

- $s(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))} = \begin{cases} 1 - \frac{a(i)}{b(i)} & if\ a(i) < b(i) \\ 0 & if\ a(i) = b(i) \\ \frac{b(i)}{a(i)} - 1 & if\ a(i) > b(i) \end{cases}$
  - a(i) << b(i) 이면 s(i) -> 1 이므로 s(i)가 1에 가까우면 좋음

  - 실제로는 모든 데이터에 대해 s(i)를 구해서 그 평균을 이용한다.
  - 군집화에 대한 지표가 될 수 있다.

Goal 3. 적정 k의 값은 어떻게 구하는가?

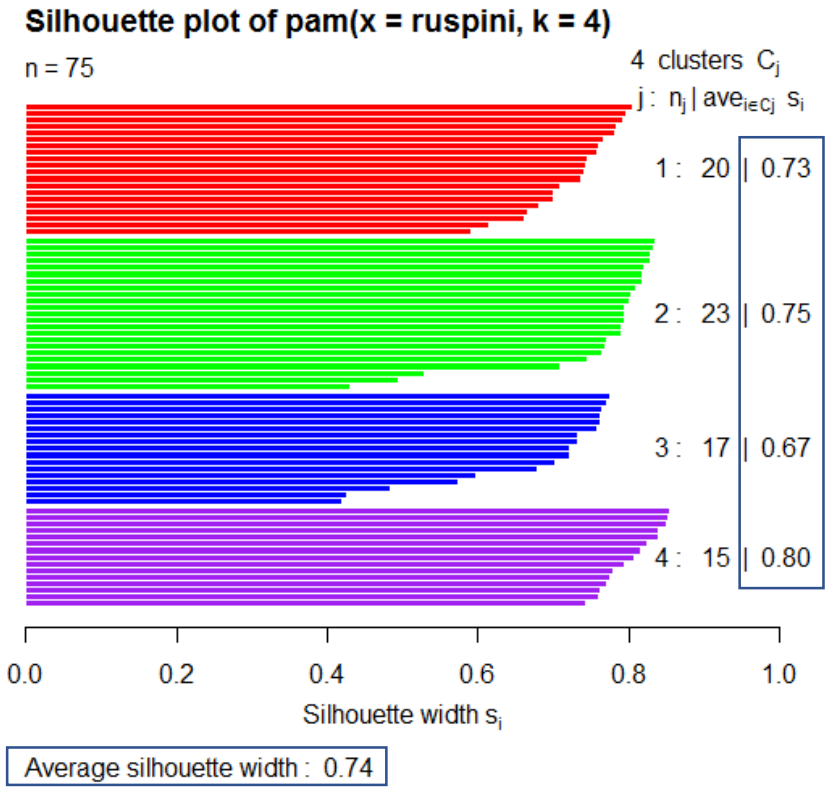
실루엣 그래프 예시



Goal 3. 적정 k의 값은 어떻게 구하는가?

평균 실루엣 계수가 0.5보다 크면 좋은 k라고 판단

각 군집별 실루엣 계수들이 높은 경우도 적절한 k라고 판단 가능



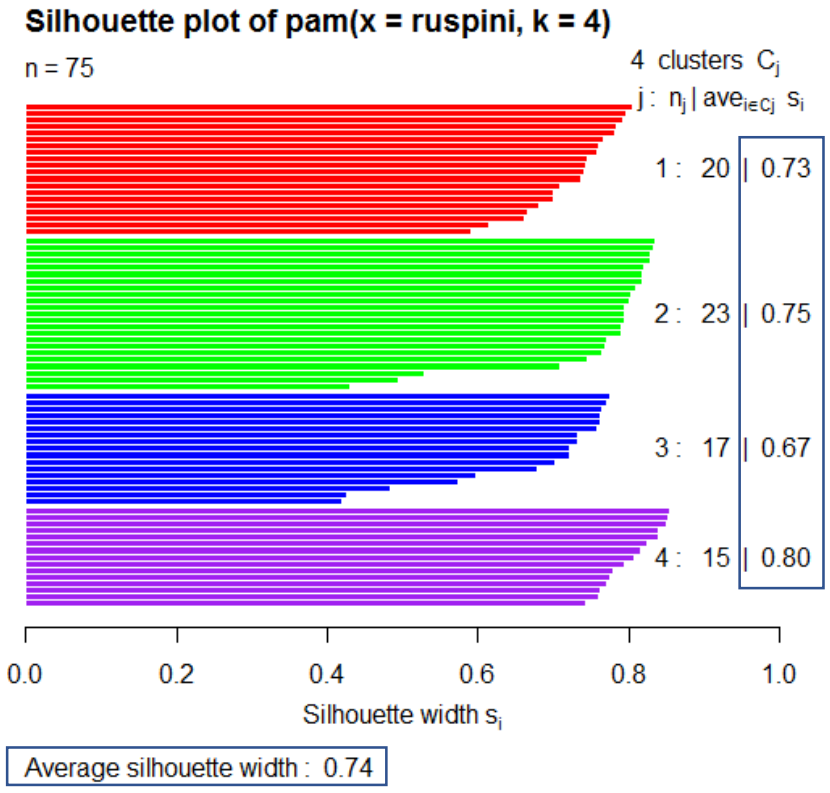
각 군집별 실루엣 계수 평균

데이터 전체의 실루엣 계수 평균

Goal 3. 적정 k의 값은 어떻게 구하는가?

평균 실루엣 계수가 0.5보다 크면 좋은 k라고 판단

각 군집별 실루엣 계수들이 높은 경우도 적절한 k라고 판단 가능



데이터 전체의 실루엣 계수 평균

각 군집별 실루엣 계수 평균

실루엣 계수 평균

-1.0 ≤ 실루엣 계수 < 0.3 : 나쁨

0.3 ≤ 실루엣 계수 ≤ 0.5 : 보통

0.5 < 실루엣 계수 ≤ 1.0: 좋음

4

계층적 군집분석

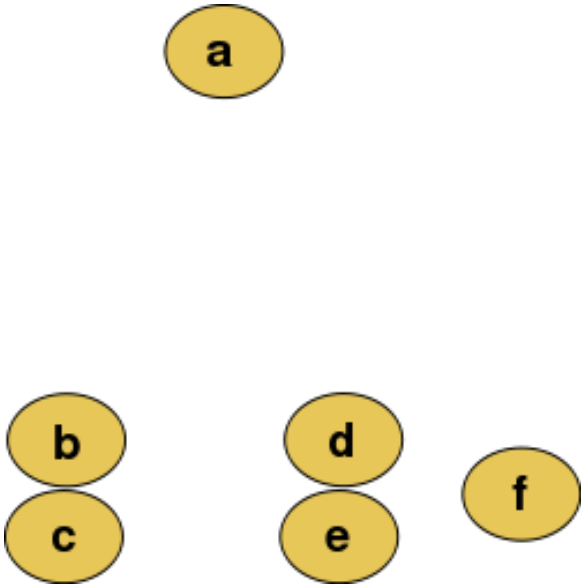
Goal 4. 계층적 군집분석이란?

- 계층적 군집화(Hierarchical Clustering)란?
- 가까운 관측 값들끼리 묶는 병합과 먼 관측 값들을 나누어 가는 분할에 의해 전체 군집들간 **구조적 관계**를 분석하는 기법
- 덴드로그램(Dendrogram)을 통해 시각화 가능

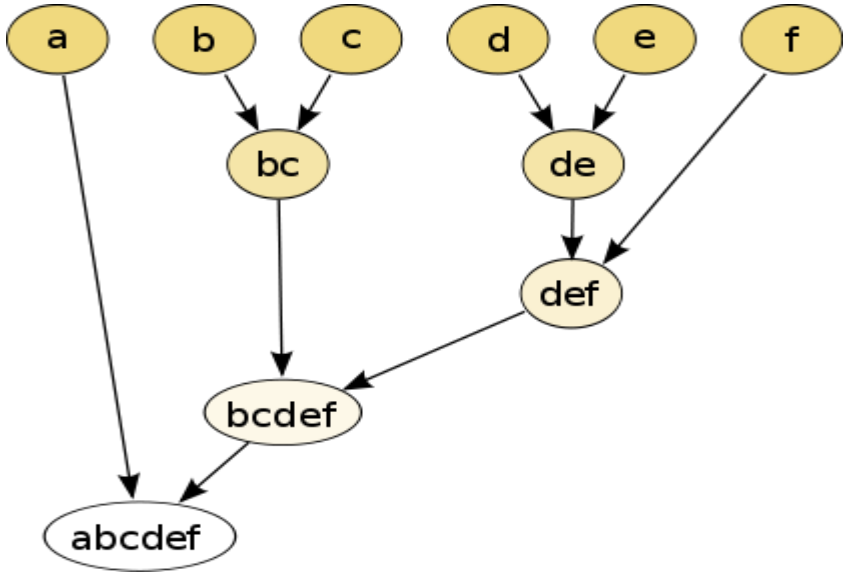
덴드로그램: 계층적 군집화에 의해 생성된 트리 다이어그램

Goal 4. 계층적 군집분석이란?

- 계층적 군집화(Hierarchical Clustering)란?
- 가까운 관측 값들끼리 묶는 병합과 먼 관측 값들을 나누어가는 분할에 의해 전체 군집들간 **구조적 관계**를 분석하는 기법
- 덴드로그램(Dendrogram)을 통해 시각화 가능



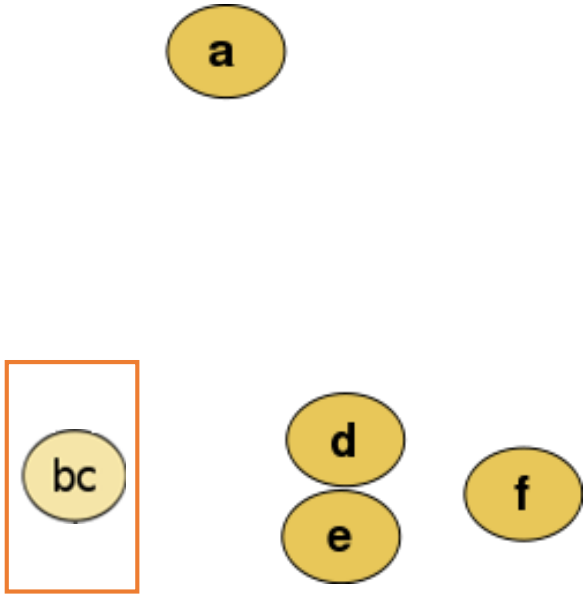
< 원시 데이터 >



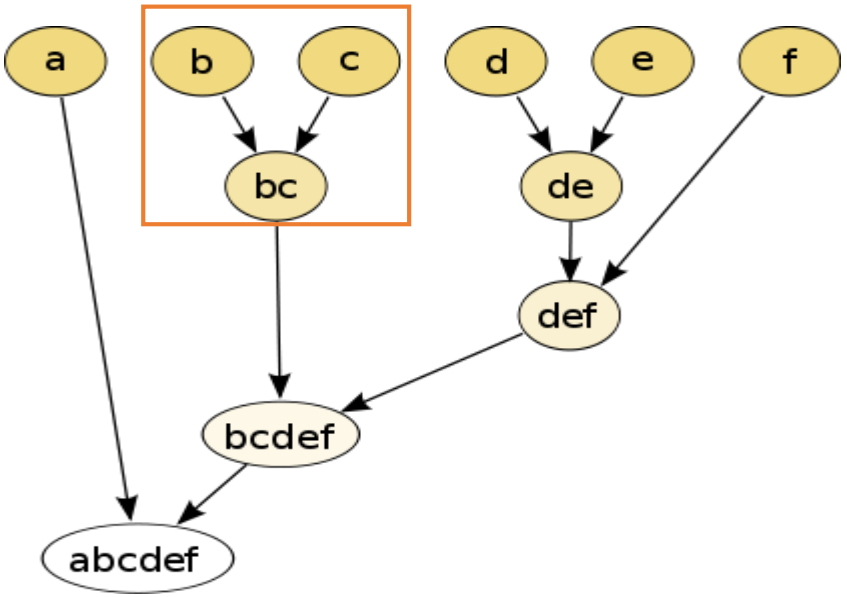
< 덴 드로 그램 >

Goal 4. 계층적 군집분석이란?

- 계층적 군집화(Hierarchical Clustering)란?
- 가까운 관측 값들끼리 묶는 병합과 먼 관측 값들을 나누어가는 분할에 의해 전체 군집들간 **구조적 관계**를 분석하는 기법
- 덴드로그램(Dendrogram)을 통해 시각화 가능



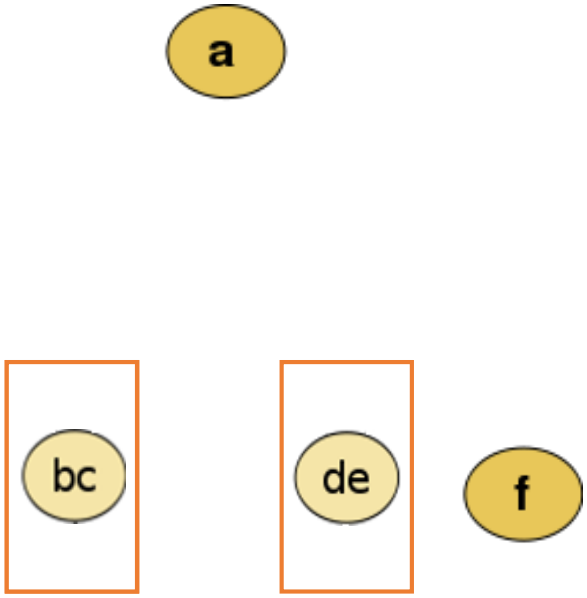
< 원시 데이터 >



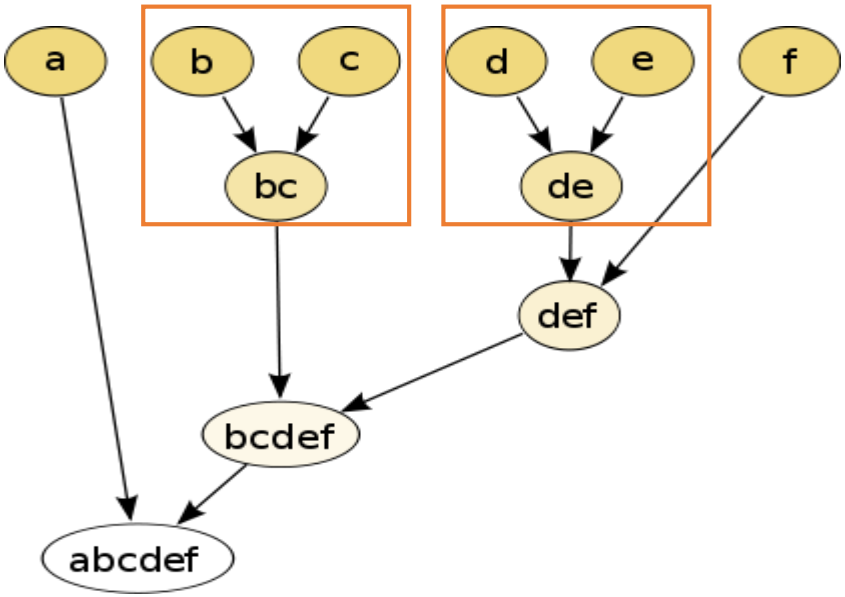
< 덴 드로 그램 >

Goal 4. 계층적 군집분석이란?

- 계층적 군집화(Hierarchical Clustering)란?
- 가까운 관측 값들끼리 묶는 병합과 먼 관측 값들을 나누어가는 분할에 의해 전체 군집들간 **구조적 관계**를 분석하는 기법
- 덴드로그램(Dendrogram)을 통해 시각화 가능



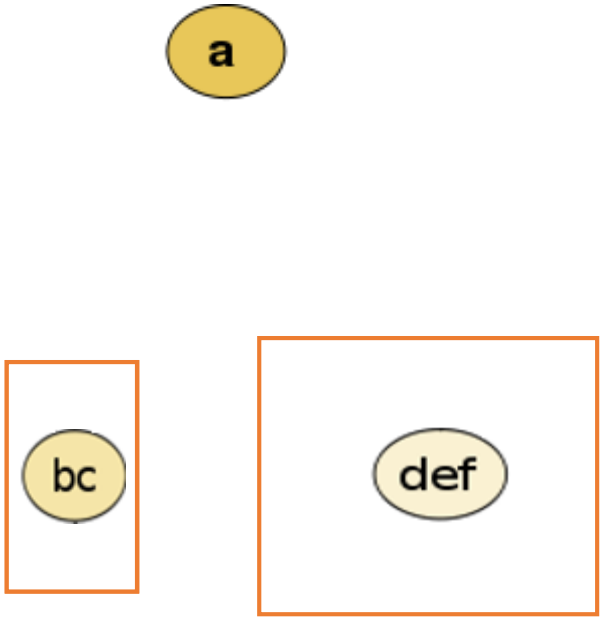
< 원시 데이터 >



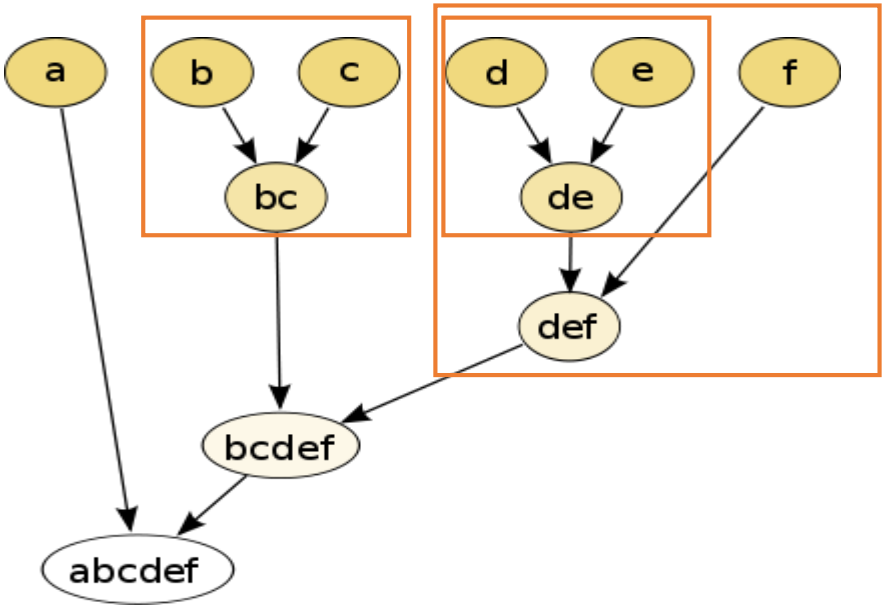
< 덴 드로 그램 >

Goal 4. 계층적 군집분석이란?

- 계층적 군집화(Hierarchical Clustering)란?
- 가까운 관측 값들끼리 묶는 병합과 먼 관측 값들을 나누어가는 분할에 의해 전체 군집들간 **구조적 관계**를 분석하는 기법
- 덴드로그램(Dendrogram)을 통해 시각화 가능



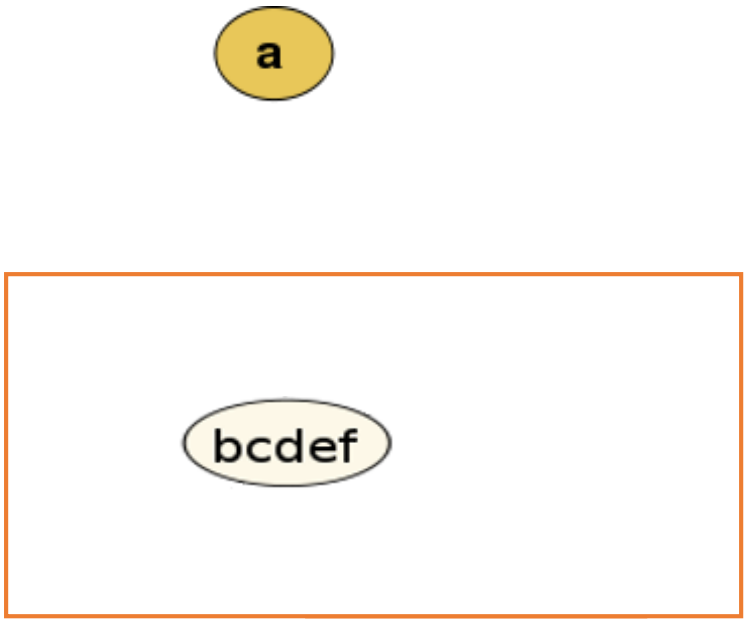
< 원시 데이터 >



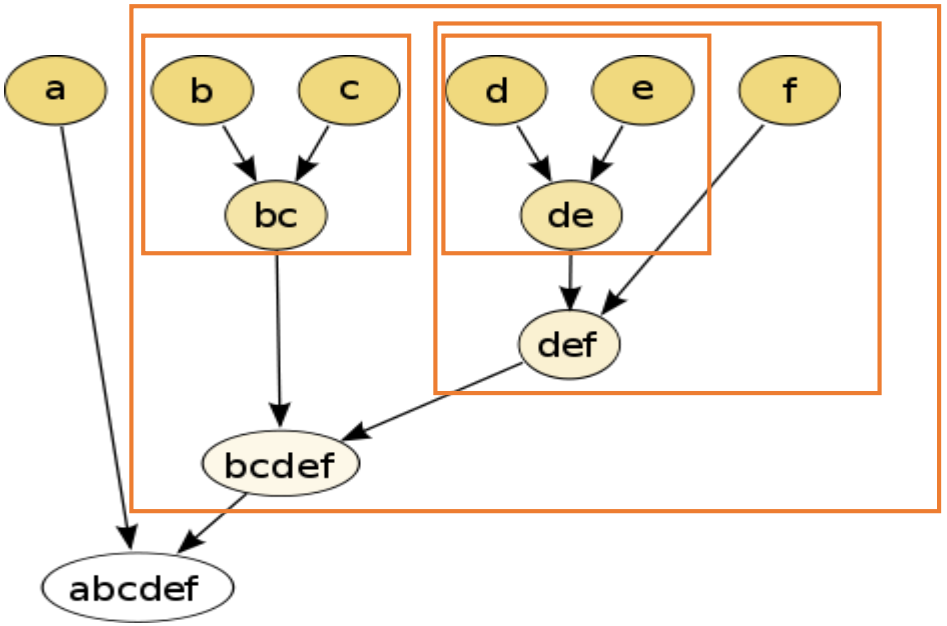
< 덴 드로 그램 >

Goal 4. 계층적 군집분석이란?

- 계층적 군집화(Hierarchical Clustering)란?
- 가까운 관측 값들끼리 묶는 병합과 먼 관측 값들을 나누어가는 분할에 의해 전체 군집들간 **구조적 관계**를 분석하는 기법
- 덴드로그램(Dendrogram)을 통해 시각화 가능



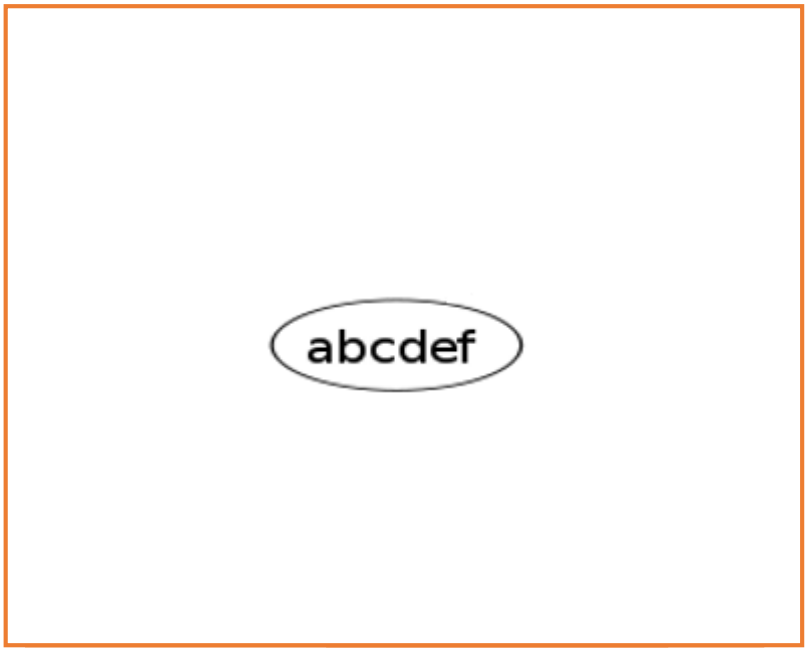
< 원시 데이터 >



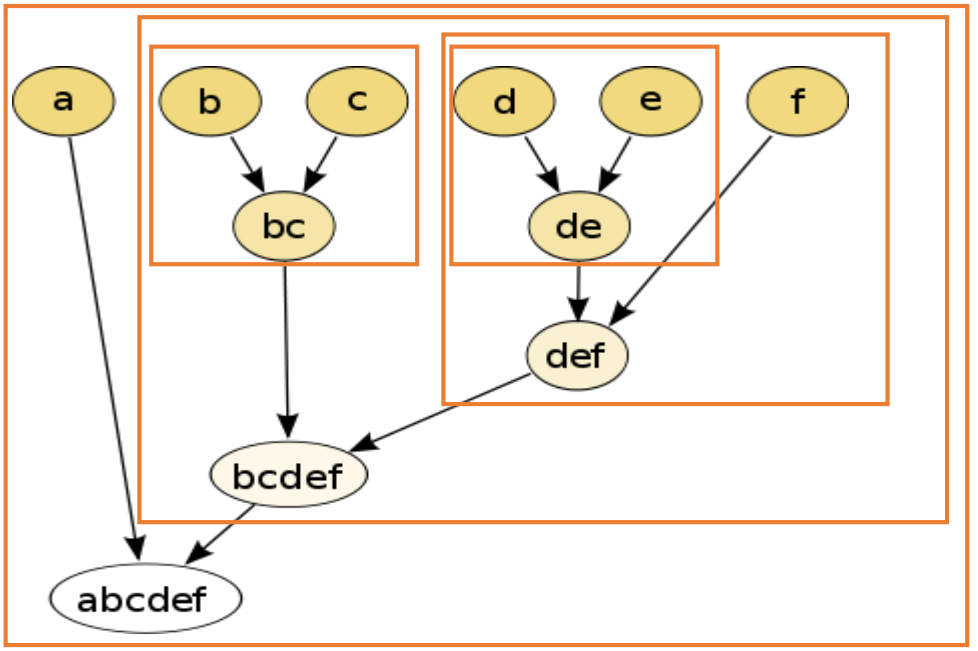
< 덴 드로 그램 >

Goal 4. 계층적 군집분석이란?

- 계층적 군집화(Hierarchical Clustering)란?
- 가까운 관측 값들끼리 묶는 병합과 먼 관측 값들을 나누어가는 분할에 의해 전체 군집들간 **구조적 관계**를 분석하는 기법
- 덴드로그램(Dendrogram)을 통해 시각화 가능



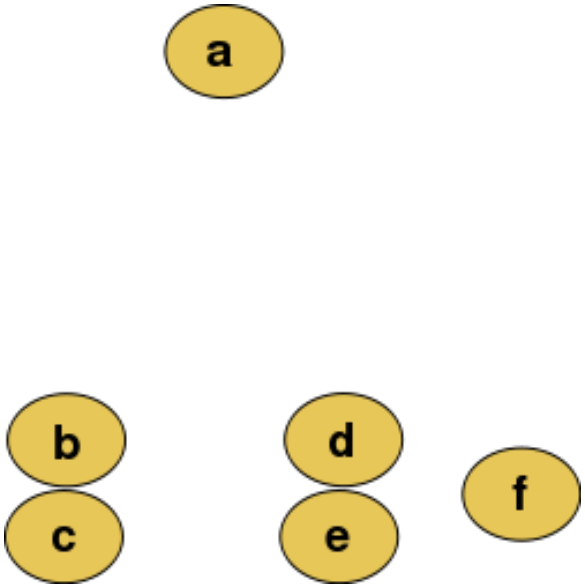
< 원시 데이터 >



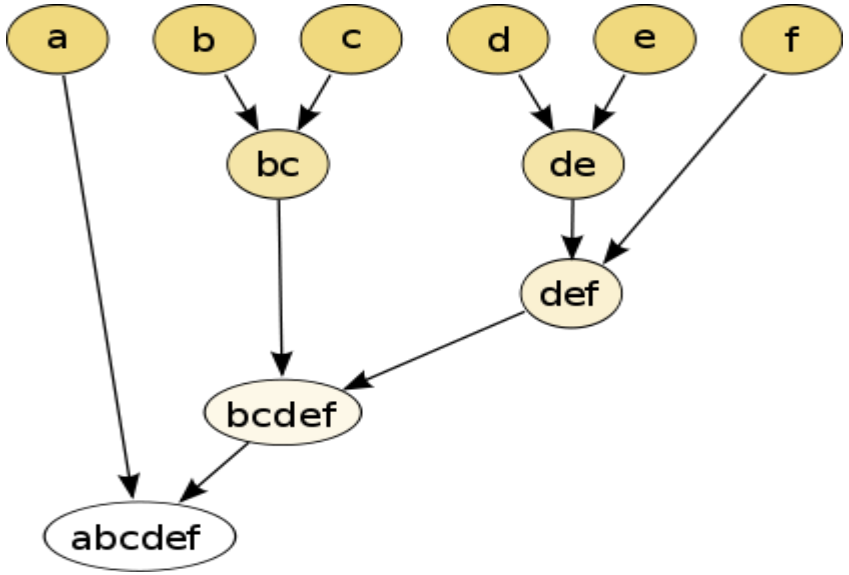
< 덴 드로 그램 >

Goal 4. 계층적 군집분석이란?

- 계층적 군집화(Hierarchical Clustering)란?
- 가까운 관측 값들끼리 묶는 병합과 먼 관측 값들을 나누어가는 분할에 의해 전체 군집들간 **구조적 관계**를 분석하는 기법
- 덴드로그램(Dendrogram)을 통해 시각화 가능



< 원시 데이터 >

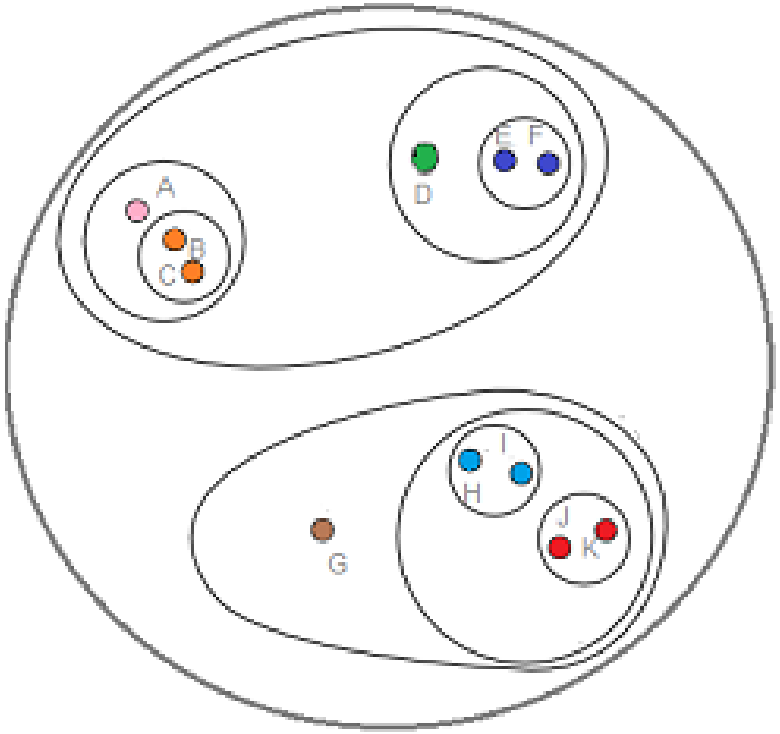


< 덴 드로 그램 >

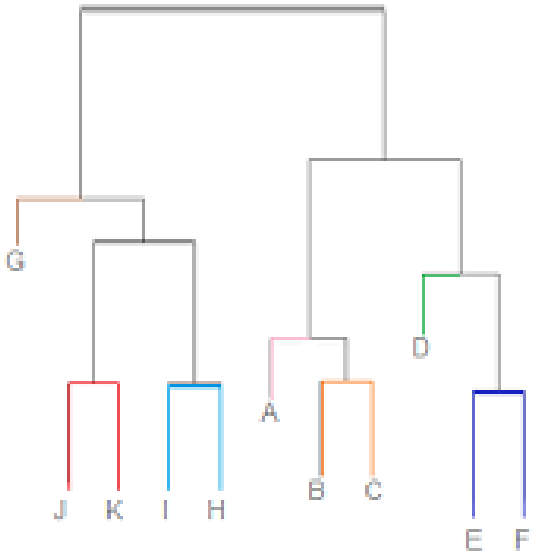
Goal 4. 계층적 군집분석이란?

계층적 군집화(Hierarchical Clustering) 예시

- 계층적 트리모형을 구축하여 개별 개체들을 유사한 군집과 통합하는 분석



< 중첩된 군집 >



< 덴 드로 그램 >

Goal 4. 계층적 군집분석이란?

주요 특징

- 덴드로그램이 생성된 후 적절한 수준에서 자르면 군집화 결과 생성

  계층적 군집분석에선 주로 병합 방법 사용

- 병합계층 군집화(Agglomerative Hierarchical Method)

  n개의 군집으로 시작해서 하나의 군집이 남을 때까지

  순차적으로 유사한 군집들을 병합

- 분할계층 군집화(Divisive Hierarchical Method)

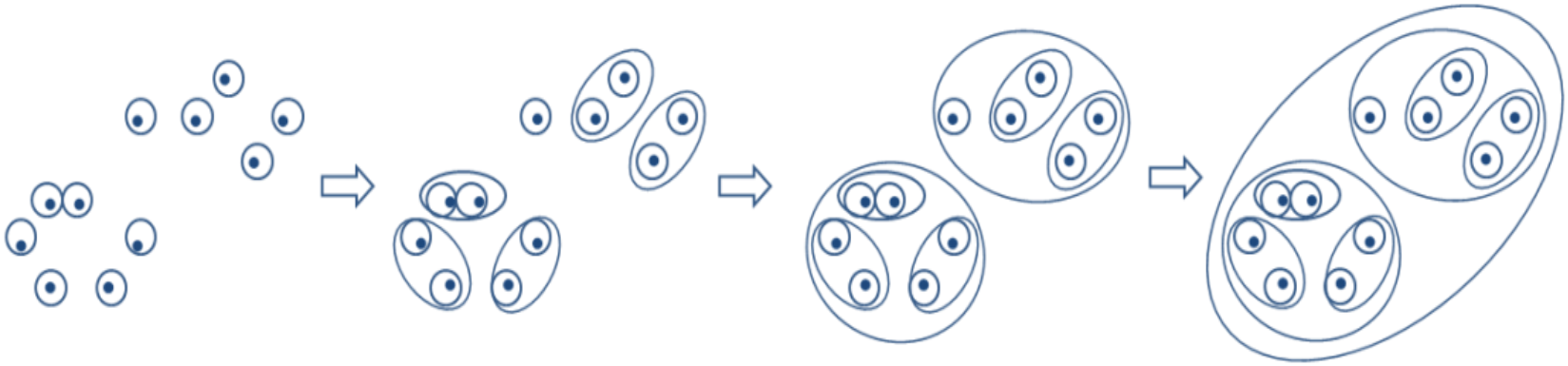
  병합방법과는 반대로 모든 개체들을 포함하고 있는

  하나의 군집에서 출발

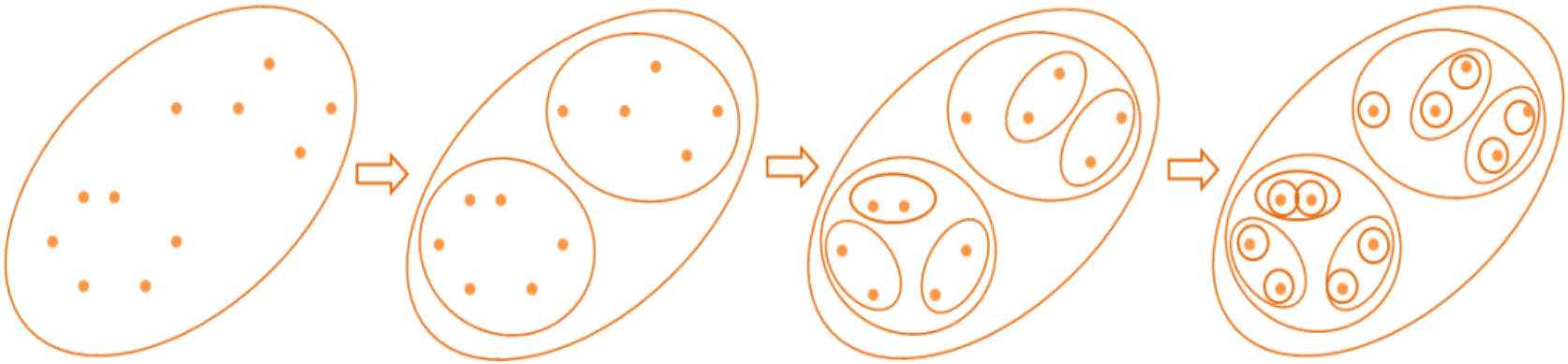
Goal 4. 계층적 군집분석이란?

계층적 군집분석의 두 가지 방식

Agglomerative Hierarchical Clustering < 병합계층 군집화 >



Divisive Hierarchical Clustering < 분할계층 군집화 >

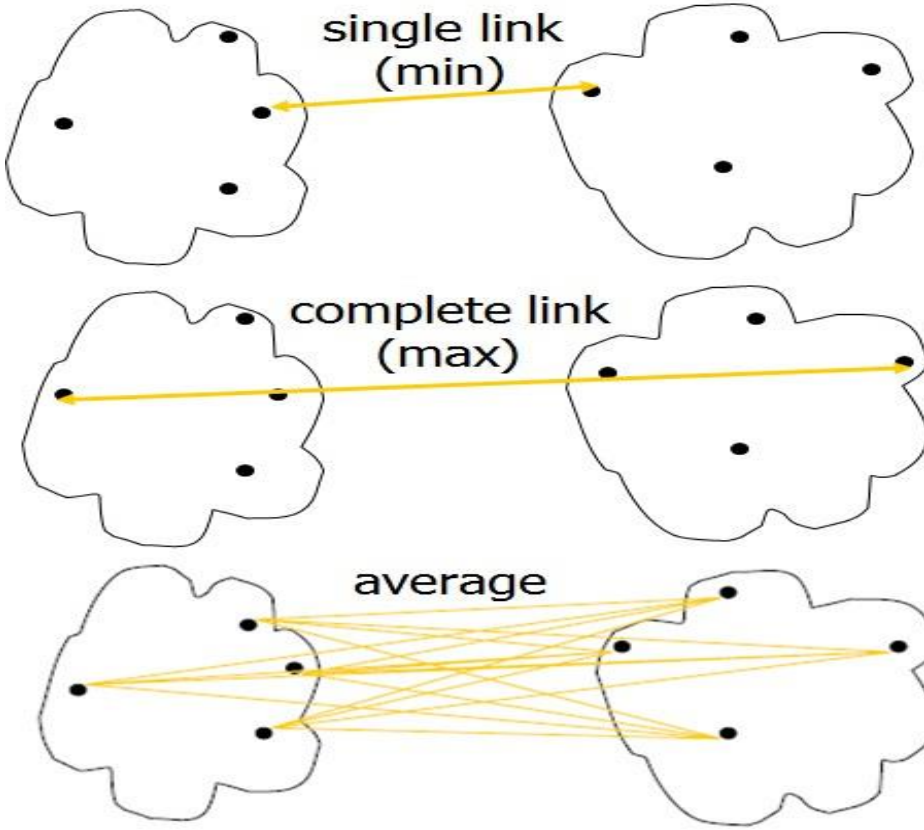


Goal 4. 계층적 군집분석이란?

병합계층 군집화 알고리즘

- 핵심 수행 절차 : 두 군집 사이의 유사도/거리 측정

  Ex) Single Linkage, Average linkage, Complete Linkage, Centroid Linkage, etc



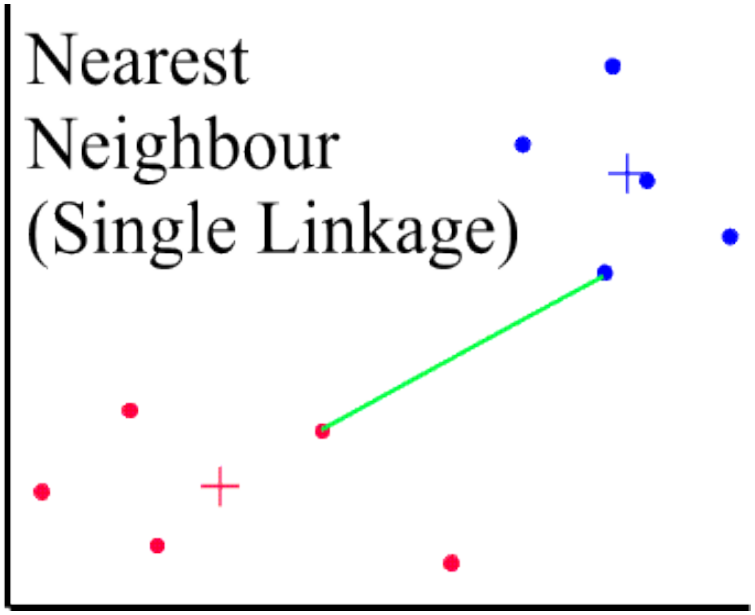
각 군집에 속한 개체들 사이의 거리 중 가장 가까운 값을 군집간 거리로 정의

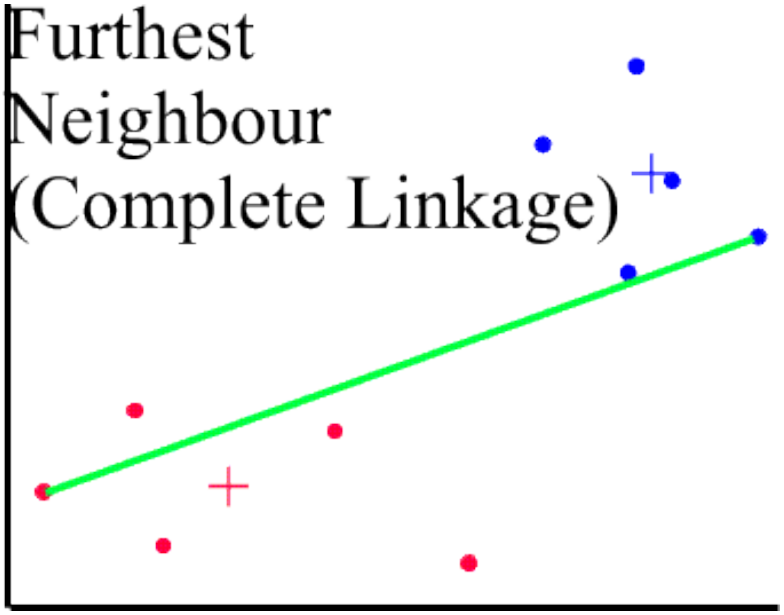
각 군집에 속한 개체들 사이의 거리 중 가장 먼 값을 군집간 거리로 정의

각 군집에 속한 개체들 사이의 거리의 평균값을 군집간 거리로 정의

Goal 4. 계층적 군집분석이란?

병합계층 군집화 알고리즘





- Single Linkage(Minimum Distance)

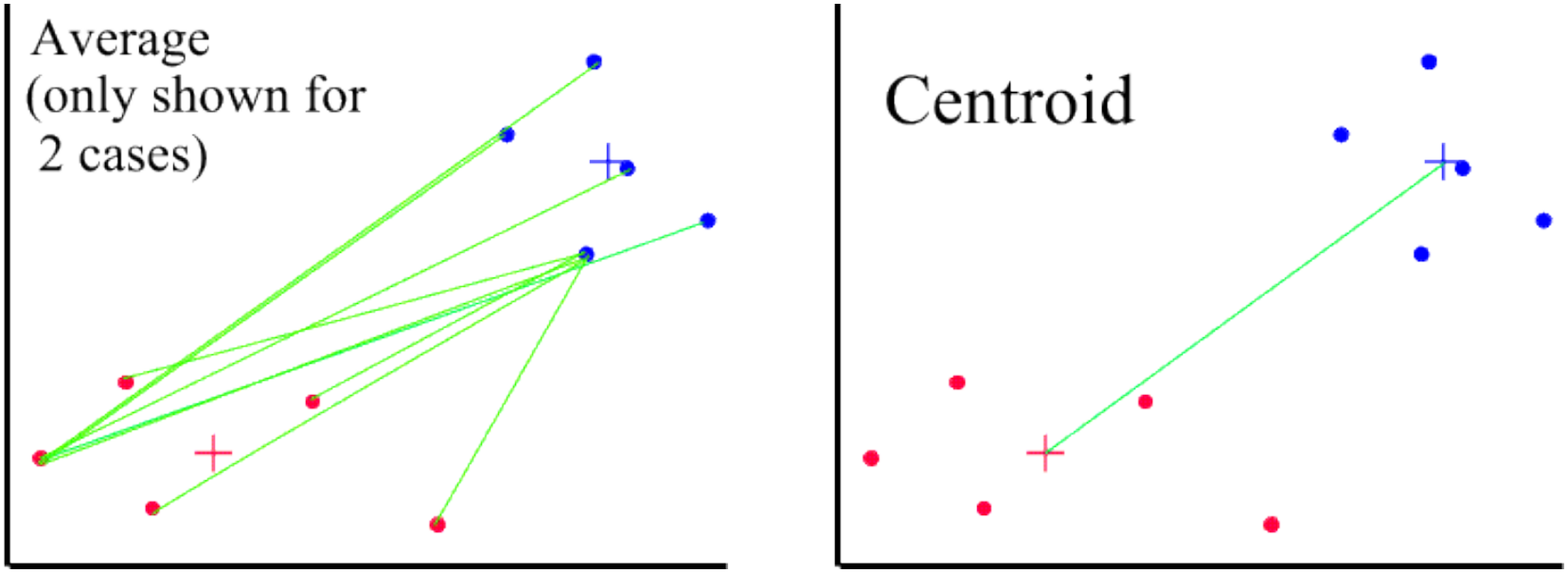
  각 군집에 속한 개체들 사이의 거리 중 가장 가까운 값을 군집간 거리로 정의

- Complete Linkage(Distance Between Centroids)

  각 군집에 속한 개체들 사이의 거리 중 가장 먼 값을 군집간 거리로 정의

Goal 4. 계층적 군집분석이란?

병합계층 군집화 알고리즘



- Average linkage(Mean distance)

  각 군집에 속한 개체들 사이의 거리 평균값을 군집간 거리로 정의

- Centroid linkage(Distance between centroids)

  각 군집의 중심간 거리로 정의

Goal 4. 계층적 군집분석이란?

병합계층 군집화 알고리즘 수행절차

1. 모든 개체를 개별 군집으로 정의하고 군집간 거리 행렬 계산
2. 가장 가까운 두 개의 군집을 하나의 군집으로 통합
3. 군집간 거리 행렬 업데이트
4. 2, 3단계를 반복
5. 모든 개체가 하나의 군집으로 통합되면 종료

Goal 5. 예제 R 코드를 통해 군집분석을 실습해보자

k-means 군집분석 실습

데이터: iris dataset

목표: 꽃받침과 꽃잎에 대한 정보를 이용해 데이터를 군집화

- k-means clustering 자체는 비지도학습이지만, 실습에 사용하는 데이터인 iris data는 실제로 '종(species)'이라는 label이 존재한다.
- clustering 결과가 species에 따라 나눈 결과와 비슷하게 나오는지 확인해 보기 위하여 iris dataset을 사용한다.