DBSCAN 보고서

Tags

DBSCAN 군집화는 특정 공간 내 데이터의 밀도 차이를 기반으로 한 알고리즘이다

DBSCAN은 복잡한 기하학적 분포도를 가진 데이터에 대해서도 군집화를 잘 수행한다

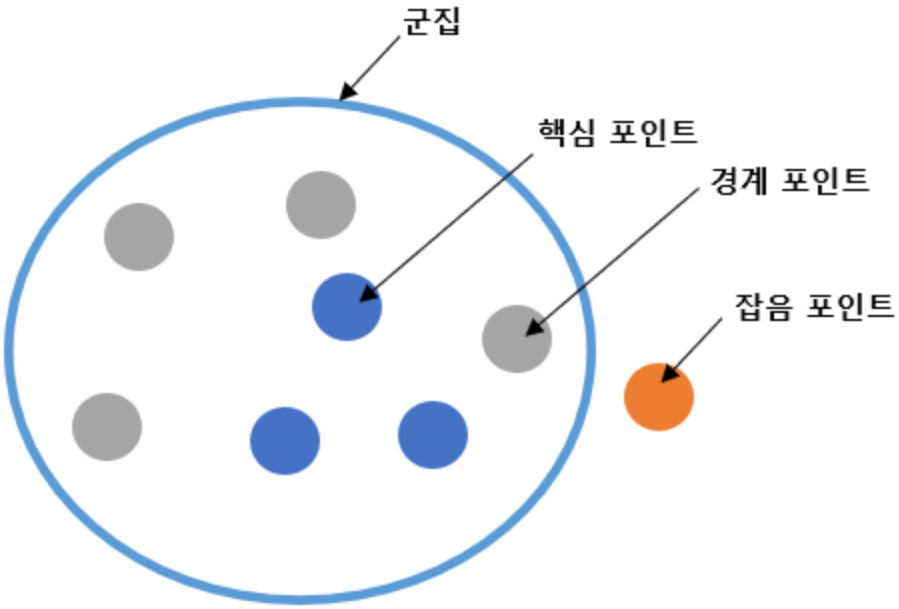
DBSCAN은 K-평균과 GMM과 달리 클러스터의 개수를 미리 지정할 필요가 없으며 어떤 군집에도 속하지 않는 포인트를 구분할 수 있다

- 데이터가 모여있는 밀도를 기반으로 클러스터를 형성
- 고밀도 지역과 저밀도 지역을 이용해 군집화를 진행
- 이상치 탐지에서 좋은 성능을 보여준다

- 장점: 클러스터 개수를 설정할 필요가 없다.
- 단점: 모든 데이터를 봐야하기 때문에 시간 복잡도가 높아서 대용량 데이터에는 처리하는데 문제가 있다.

[주요 파라미터]

- **입실론 주변 영역(epsilon)** : 개별 데이터를 중심으로 입실론 반경을 가지는 원형의 영역 → eps
- **최소 데이터 개수(min points)** : 핵심 포인트가 되기 위해 입실론 주변 영역에 포함되는 타 데이터의 최소 개수 → min_samples



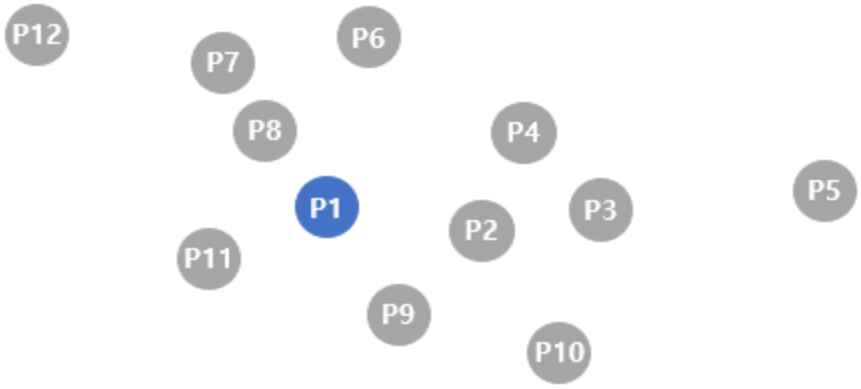
- **핵심 포인트(Core Point)** : 주변 영역 내 최소 데이터 개수 이상의 타 데이터를 가지고 있을 경우
- **이웃 포인트(Neighbor Point)** : 주변 영역 내 위치한 타 데이터
- **경계 포인트(Border Point)** : 주변 영역 내 최소 데이터 개수 이상의 이웃 포인트를 가지고 있지 않지만

  핵심 포인트를 이웃 포인트로 가지고 있는 데이터
- **잡음 포인트(Noise Point)** : 최소 데이터 개수 이상의 이웃 포인트를 가지고 있지 않으며, 핵심 포인트도 이웃 포인트로 가지고 있지 않는 데이터

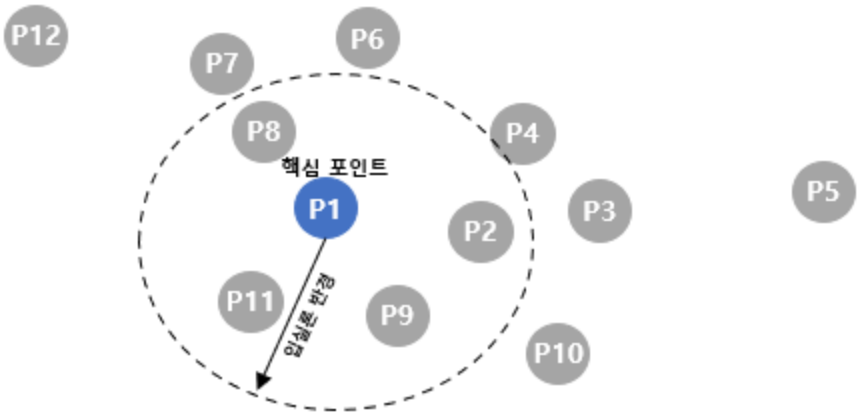
---

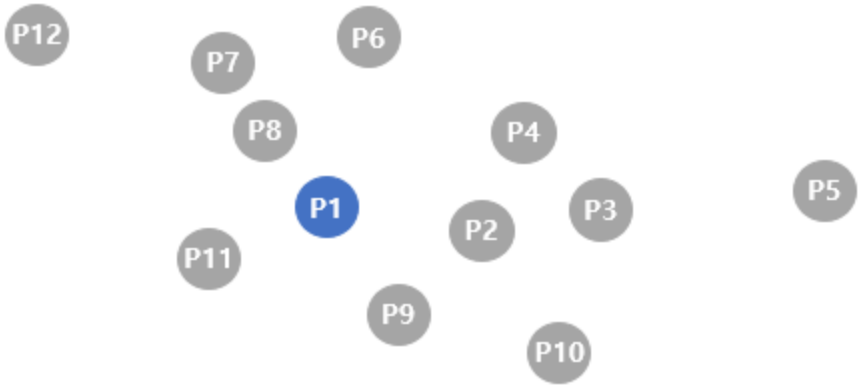
2) DBSCAN 군집화 원리

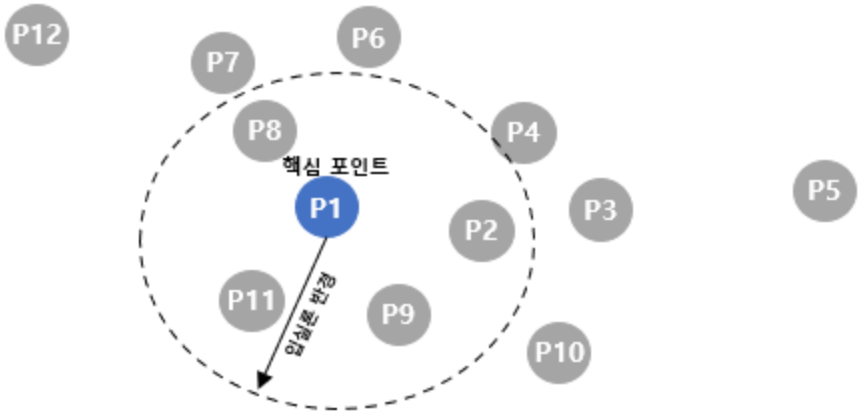
1. 특정 입실론 반경 내 포함될 최소 데이터 세트를 자기 자신을 포함한 5개로 가정하겠습니다.



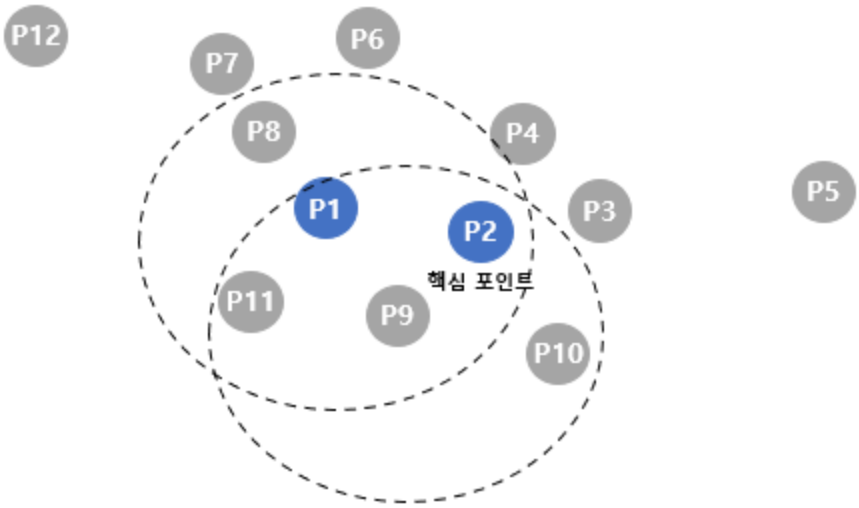
2. **P1** 데이터를 중심으로 입실론 반경 내 포함된 데이터가 5개로 최소 데이터 4개 이상을 만족하므로 P1 데이터는 **핵심 포인트**입니다.

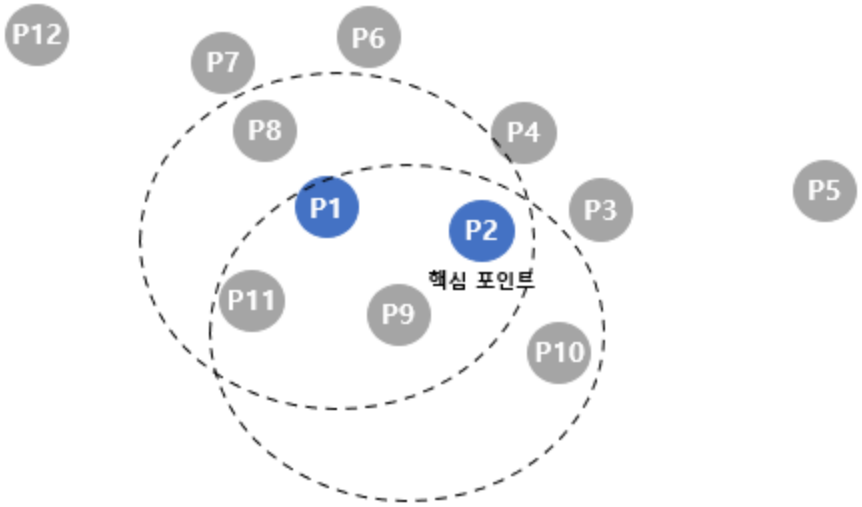




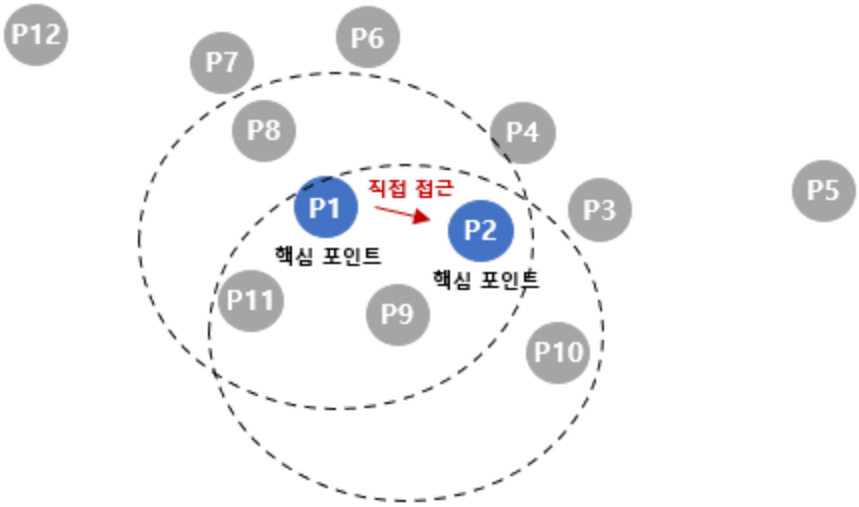


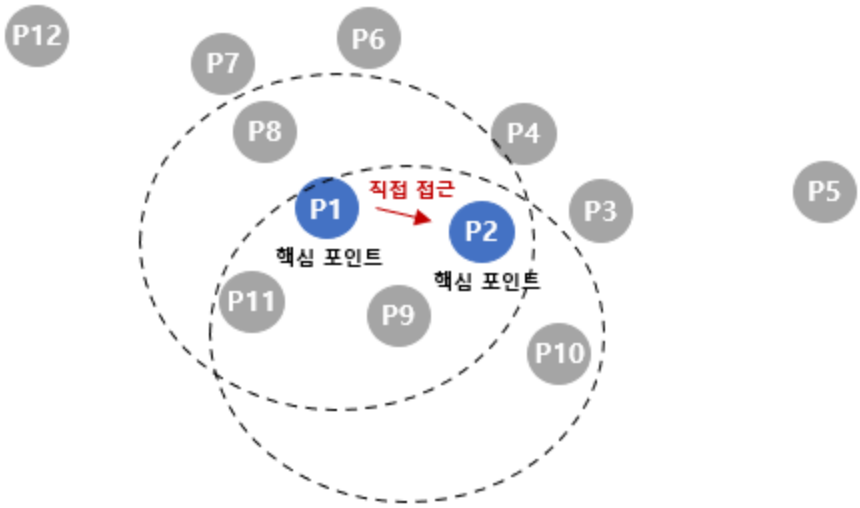
3. **P2**도 마찬가지로 반경 내 5개의 데이터를 가지고 있으므로 **핵심 포인트** 입니다.



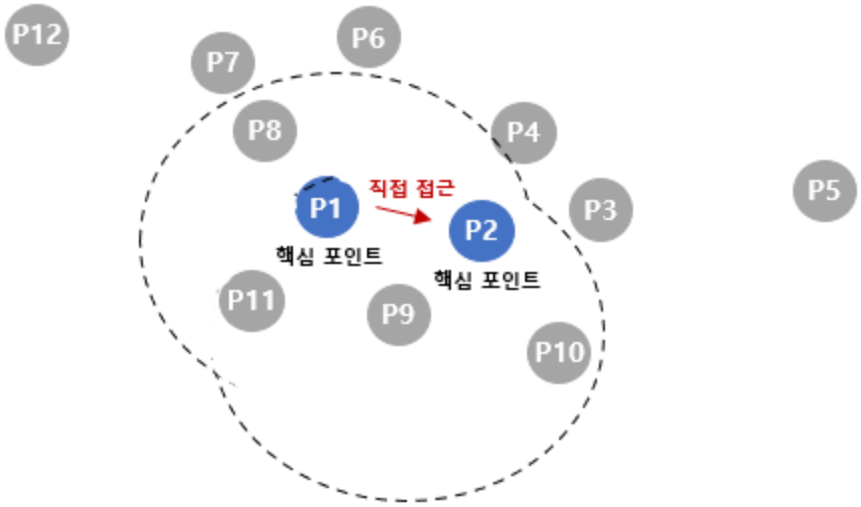


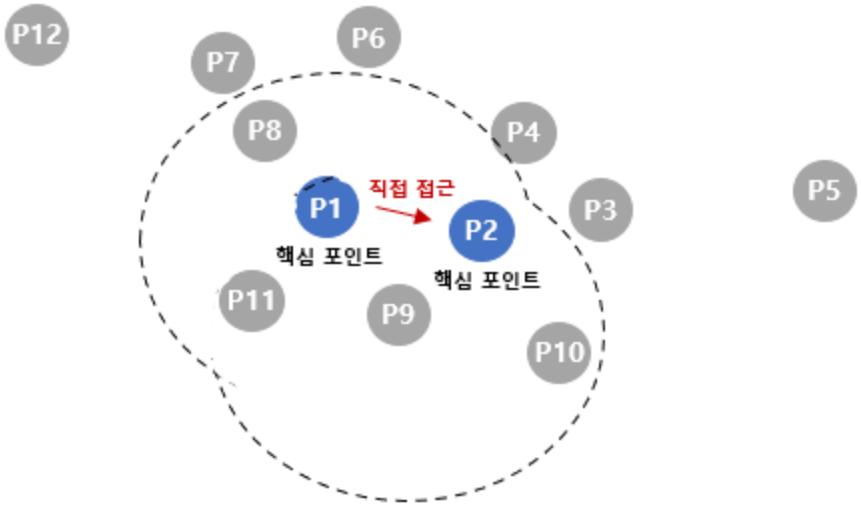
4. **핵심 포인트 P1의 이웃 데이터 포인트 P2** 역시 **핵심 포인트**일 경우 P1에서 P2로 연결해 **직접 접근**이 가능합니다.



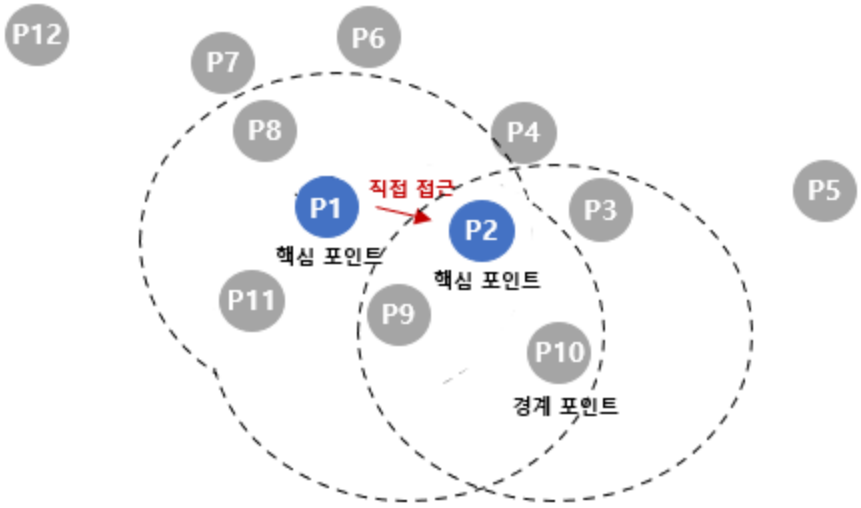


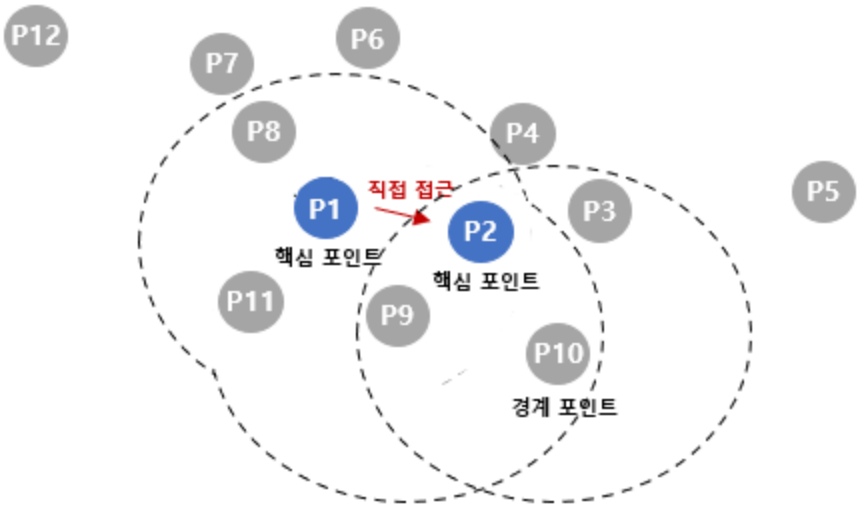
5. 특정 핵심 포인트에서 **직접 접근**이 가능한 다른 핵심 포인트를 서로 연결하면서 군집화를 구성합니다.
   이러한 방식으로 점진적으로 군집 영역을 확장해 나가는 것이 DBSCAN 군집화 방식입니다.



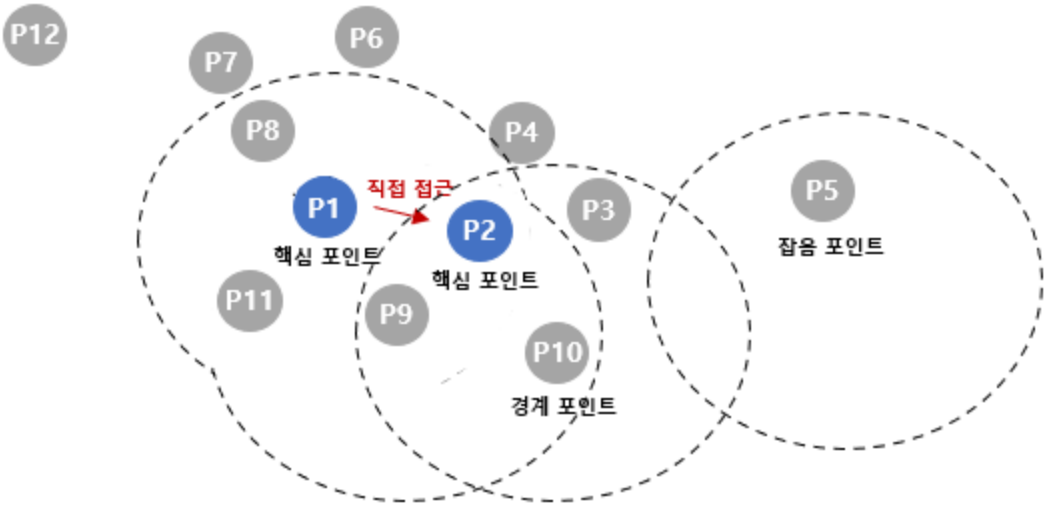


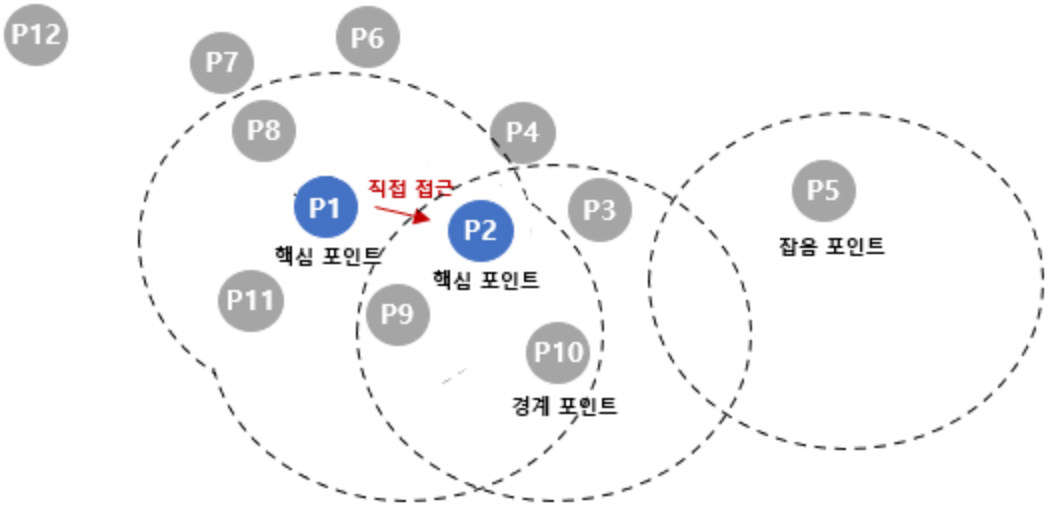
6. P10 의 경우 반경 내 포함되는 이웃 데이터가 3개 이므로 핵심 포인트가 될 수 없습니다. 그러나 이웃 데이터 중 핵심 포인트인 P2를 가지고 있기 때문에 경계 포인트에 해당합니다. 경계 포인트는 군집의 외곽을 형성합니다.

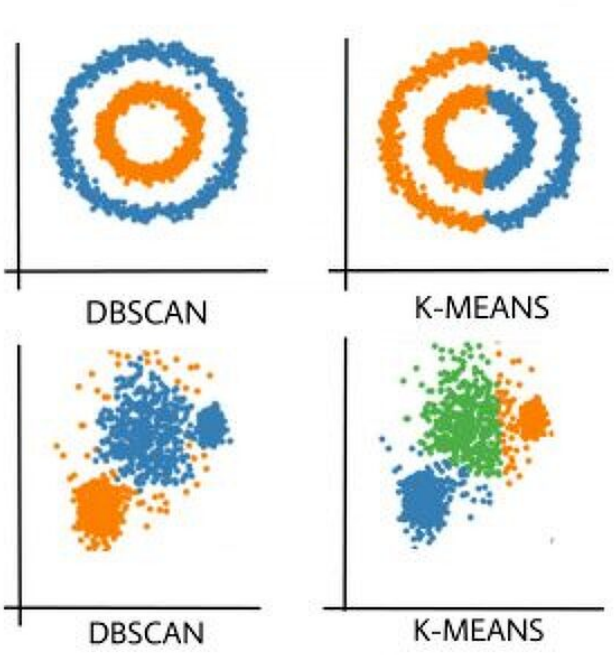




7. P5와 같이 반경 내 최소 데이터를 가지고 있지도 않고 핵심 포인트 또한 이웃 데이터로 가지고 있지 않는 데이터를 **잡음 포인트**라고 합니다.








```
dbscan = DBSCAN(eps=110, min_samples=10)
dbscan.fit(df_rfm2)
```

스케일링을 안 하고 시각화