DS 군집화 팀과제

1. 대상이 되는 데이터는 공공 데이터입니다.
   1. 군집화에 적합한(결과가 좋은) 데이터를 선택 바랍니다.
2. 전처리
   1. 결측치 처리, 데이터 형변환, 정규화, outlier(이상치) 제거 등
3. K-mean 군집화 적용
   1. 적절한 K-군집의 수는 얼마인가?  
      아래의 결과를 참조하여 정할 수 있음.
      1. Learning curve: 군집의 수(K) vs 성능(withinss, …)
      2. Dendrogram - Hierachical clustering에서 적절한 절단선을 참조
   2. 최적화
      1. 변수선택
         1. 군집화의 목적을 임의로 설정하여 목적에 따라 변수 선택 시도
         2. 또는, 뚜렷한 군집 결과가 나오도록 최적변수집합 선택
      2. 최초의 중심 위치를 변경하여 여러 번 시도
      3. 거리 척도에 대한 선택
   3. 성능 평가 – intra(withinss), inter(betweenss)-cluster distance 등을 사용
   4. 각 군집에 대하여 labelling-설명을 시도
      1. 군집내의 사례의 일반적인 특성에 대하여 설명
4. 다른 군집화 와 비교해 볼 것
   1. K-mean 이외 한 개(이상)

- 끝 -