**군집화 발표 자료**

* **군집화를 한 이유!** : 머신러닝에 들어가기에 앞서 군집 별로 특성을 파악하기 위해

**1. K-Prototype이란?**

* 범주형과 연속형 데이터가 혼재되어있을 때 사용할 수 있는 군집화 알고리즘.
* 수업시간에 배웠던 K-Means는 수치형 데이터만 사용가능
* K-Modes는 범주형 데이터만 사용가능
* K-Modes, K-Means, K-Prototype 세 가지 알고리즘 모두 적용해본 결과 가장 결과가 좋았던 K-Prototype 모델 채택

**2. K-Prototype 모델에 쓸 최적 군집의 개수를 찾기 위해 엘보우 커브 시각화**

텍스트, 폰트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

텍스트, 도표, 라인, 그래프이(가) 표시된 사진

자동 생성된 설명

🡪 엘보우 커브 시각화 결과 ‘K가 2, 3, 4일 때 군집화에 효과적일 것’ 이라는 결론 도출.

**3. K가 2, 3, 4일 때의 실루엣 분석 시각화**

텍스트, 스크린샷, 도표, 그래프이(가) 표시된 사진

자동 생성된 설명

🡪 실루엣 분석 결과, 군집의 개수가 2개일 때 실루엣 스코어가 0.828으로 가장 높고, 군집의 개수가 4개일 때 실루엣 스코어가 0.592로 가장 낮지만, 개별 군집의 실루엣 계수 평균값과 전체 실루엣 계수의 평균값의 편차가 크지 않고, 즉 빨간 점선을 관통할 경우에 군집화가 잘 되어있다고 판단할 수 있으므로, 두 가지 기준을 미루어봤을 때 그 중간값인 군집의 개수가 3인 경우를 채택한다는 결론 도출.

**4.** **군집의 개수가 3일 때의 군집화 결과**

텍스트, 스크린샷, 폰트, 라인이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

🡪 관람횟수(view\_cnt) 기준으로 최소, 최댓값을 추출해본 결과,

1번 군집은 0부터 44까지,

2번 군집은 45에서 173까지,

0번 군집은 178에서 672까지로

관람횟수가 적은 그룹, 보통인 그룹, 많은 그룹으로 군집화가 잘되었다는 결론 도출.

**5. 군집 별 가장 많이 지출한 항목(expense) feature의 특성 파악**

텍스트, 스크린샷, 라인, 도표이(가) 표시된 사진

자동 생성된 설명

🡪 관람횟수가 적은 1번 군집의 경우 가장 많이 소비한 문화예술분야가 ‘영화’인 반면,

텍스트, 스크린샷, 도표, 라인이(가) 표시된 사진

자동 생성된 설명

🡪 관람횟수가 많은 0번 군집의 경우 가장 많이 소비한 문화예술분야가 ‘영화’에 비해 ‘연극/뮤지컬’의 경우가 월등히 많음을 파악할 수 있었다.