군집화 발표 자료

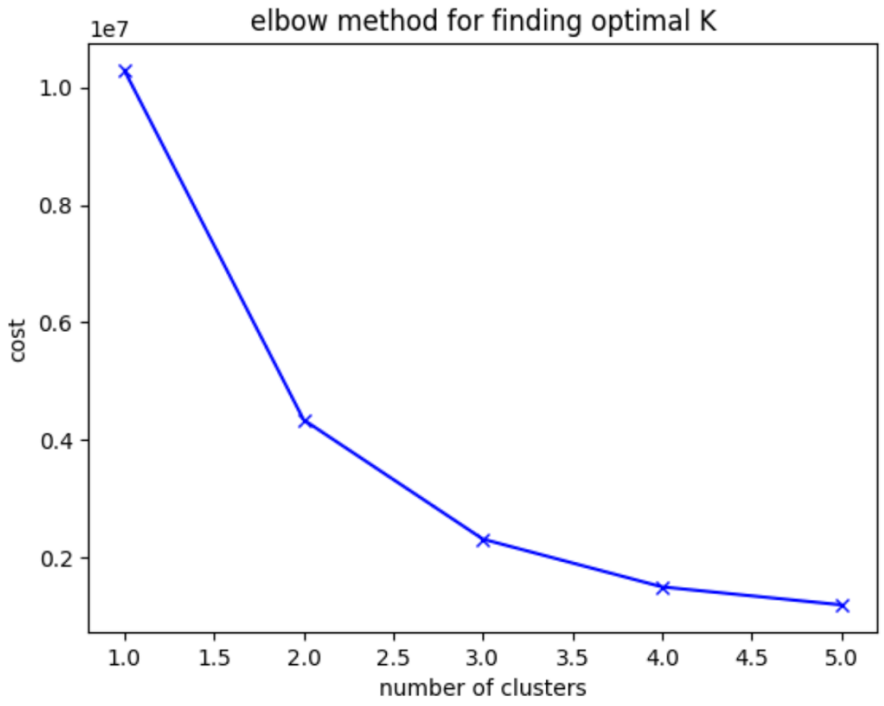
- **군집화를 한 이유!** : 머신러닝에 들어가기에 앞서 군집 별로 특성을 파악하기 위해

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

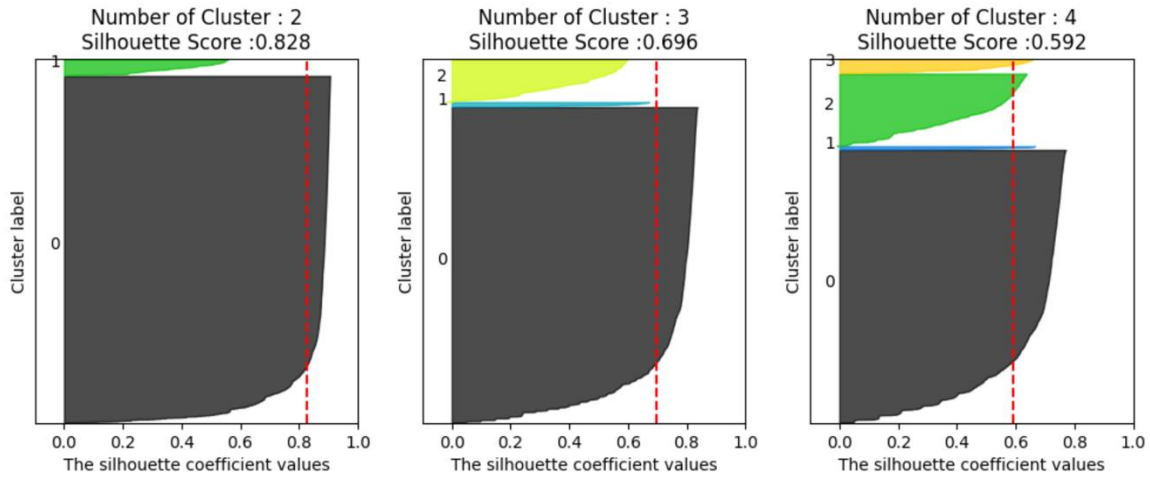
```
# KPrototype 엘보우 커브 시각화
cost = []
K = range(1, 6)
for num in K:
    kprototype = KPrototypes(n_clusters = num, n_init=10,  max_iter=500, random_state=0)
    kprototype.fit_predict(CS5, categorical=[1,2,3,4,5,6,7])
    cost.append(kprototype.cost_)

plt.plot(K, cost, 'bx-')
plt.xlabel('number of clusters')
plt.ylabel('cost')
plt.title('elbow method for finding optimal K')
plt.show()
```



→ 엘보우 커브 시각화 결과 'K가 2, 3, 4일 때 군집화에 효과적일 것' 이라는 결론 도출.

3. K가 2, 3, 4일 때의 실루엣 분석 시각화

```
# KPrototypes 군집화 - 실루엣 계수 시각화
visualize_silhouette([2, 3, 4], CS5)
```



→ 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

```
# 1번 군집의 문화예술행사 관람횟수
CS5[CS5['cluster_3']==1][['view_cnt']].min()
CS5[CS5['cluster_3']==1][['view_cnt']].max()
CS5[CS5['cluster_3']==1][['view_cnt']].mean()
```

```
view_cnt    0
dtype: int64

view_cnt    44
dtype: int64

view_cnt    13.43228
dtype: float64
```

```
# 2번 군집의 문화예술행사 관람횟수
CS5[CS5['cluster_3']==2][['view_cnt']].min()
CS5[CS5['cluster_3']==2][['view_cnt']].max()
CS5[CS5['cluster_3']==2][['view_cnt']].mean()
```

```
view_cnt    45
dtype: int64

view_cnt    173
dtype: int64

view_cnt    75.414449
dtype: float64
```

```
# 0번 군집의 문화예술행사 관람횟수
CS5[CS5['cluster_3']==0][['view_cnt']].min()
CS5[CS5['cluster_3']==0][['view_cnt']].max()
CS5[CS5['cluster_3']==0][['view_cnt']].mean()
```

```
view_cnt    178
dtype: int64

view_cnt    672
dtype: int64

view_cnt    275.39759
dtype: float64
```

→ 관람횟수(view_cnt) 기준으로 최소, 최댓값을 추출해본 결과,

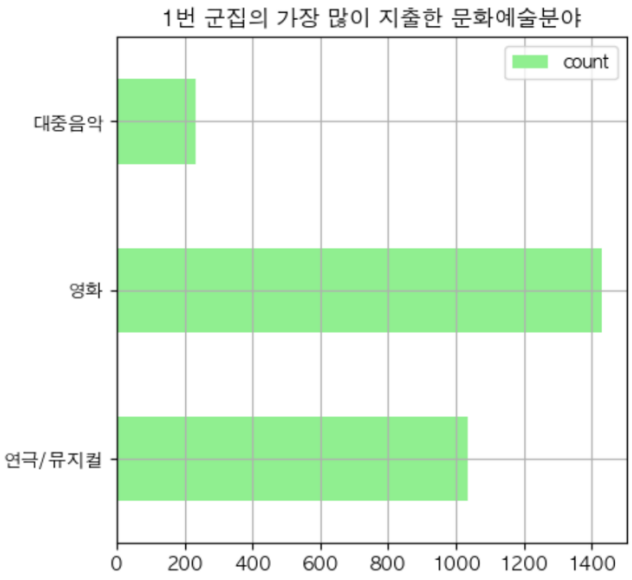
1번 군집은 0부터 44까지,

2번 군집은 45에서 173까지,

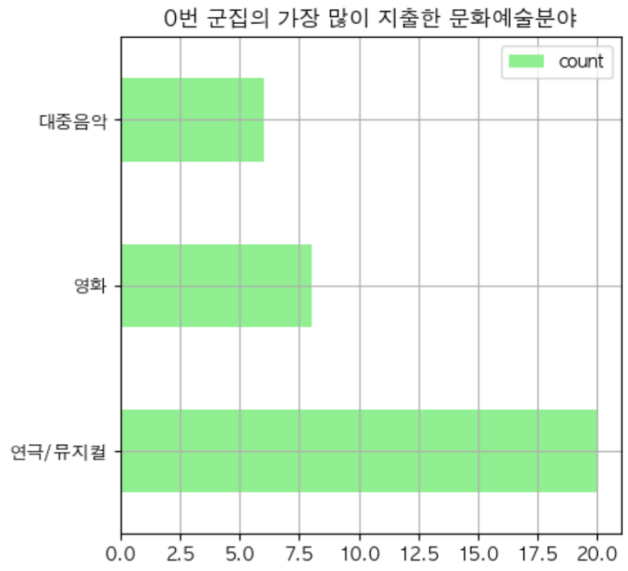
0번 군집은 178에서 672까지로

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

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



→ 관람횟수가 적은 1번 군집의 경우 가장 많이 소비한 문화예술분야가 '영화'인 반면,



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