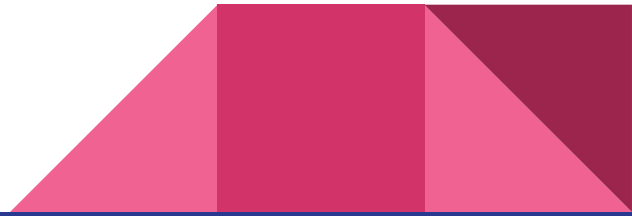
영화데이터셋 분석

박소연

Sasac 영등포 6기 데이터 AI 개발 과정

목차

1. 분석 주제
2. 사용한 데이터셋 소개
3. 데이터 분석, 인사이트 도출
4. 데이터 분석 결과 정리



분석 주제

결과를 활용하여 향후 (영화수익을 높이기 위한 목적)으로,

소비자의 관점에서 **영화 인기도에 영향을 미치는 요인을 분석**

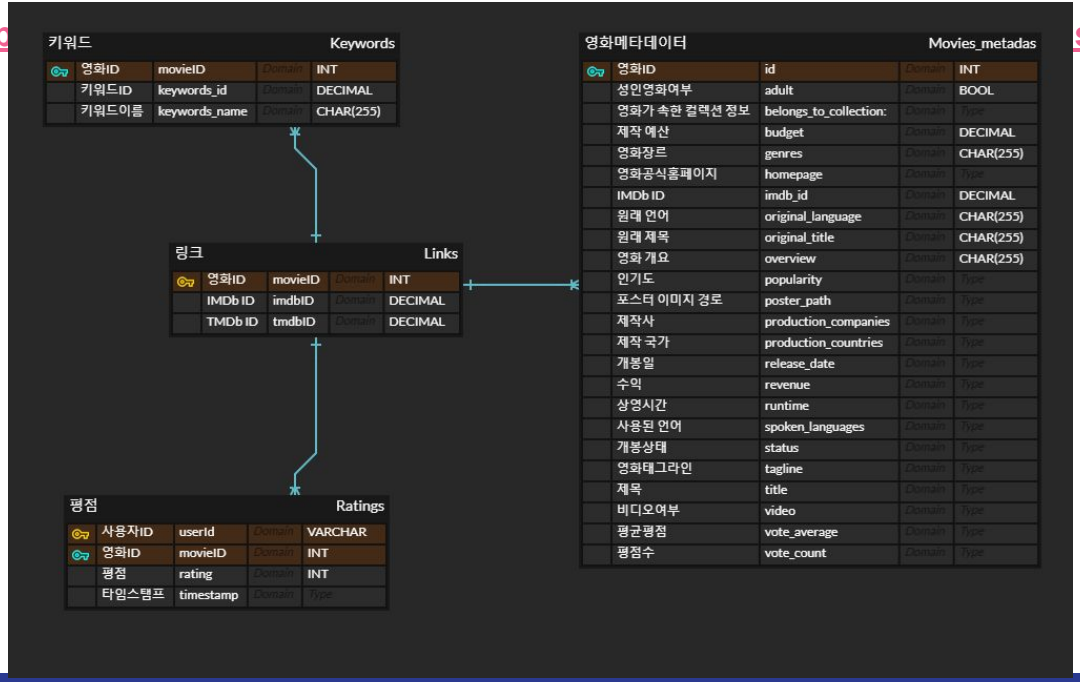
=> 다양한 요인과 인기도를 비교하여 패턴을 찾는 상관 연구.

*추후 더나아가,**=>**인기를 기능으로 사용하여 향후 흥행 성공이나 스트리밍 수를 예측하는 예측 모델링 생성 예정

사용한 영화데이터셋

1 데이터 수집: Kaggle의 영화 메타데이터 또는 IMDb, TMDb와 같은 데이터베이스.

http scussion?sort=hotness



인기도에 영향을 미치는 요인:

"평균평점"으로 추측

상관계수가 0.15

상관계수가 0.15인 경우, 이는 두 변수 간의 선형 상관관계가 매우 약하다는 것을 의미합니다.
상관계수의 해석은 다음과 같은 기준을 따릅니다:

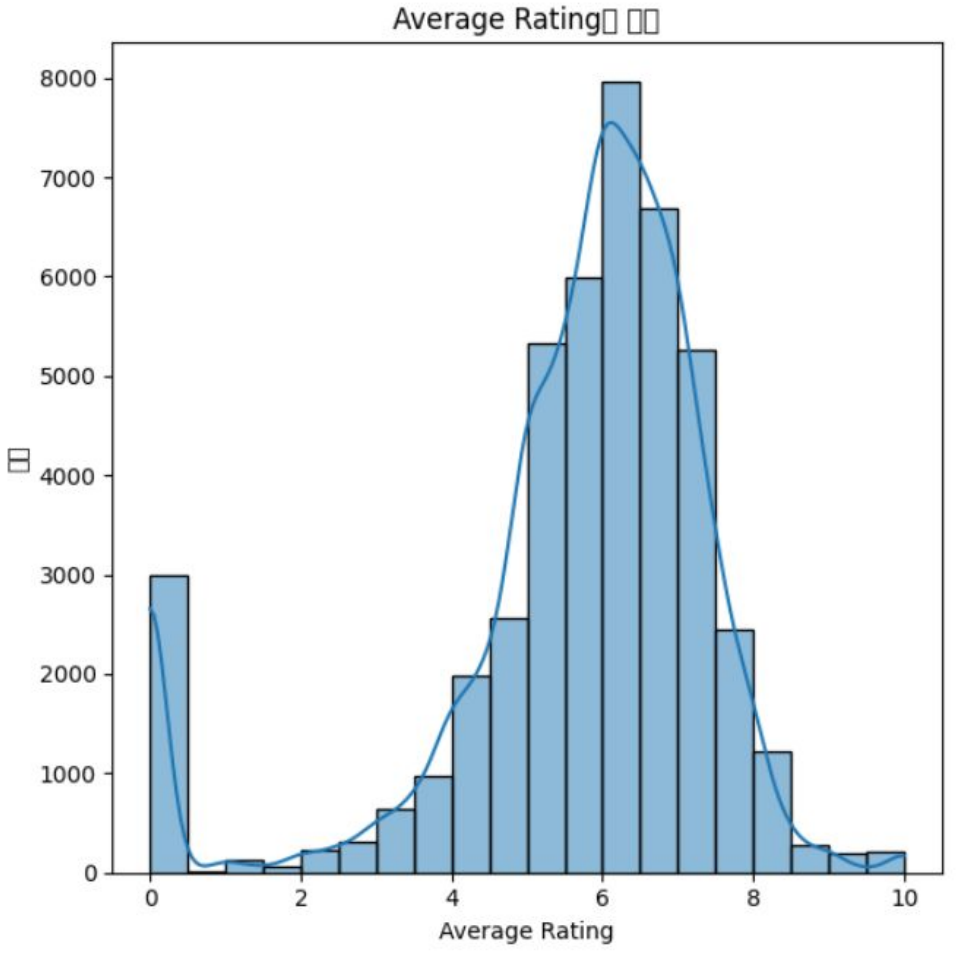
1 또는 -1: 완벽한 양의(또는 음의) 선형 상관관계

0.7 ~ 1 또는 -0.7 ~ -1: 강한 양의(또는 음의) 선형 상관관계

0.3 ~ 0.7 또는 -0.3 ~ -0.7: 중간 정도의 양의(또는 음의) 선형 상관관계

0.1 ~ 0.3 또는 -0.1 ~ -0.3: 약한 양의(또는 음의) 선형 상관관계

0: 상관관계 없음



결론 도출:

평균평점이 높아진다고 무조건 인기도가 높아지는 것은 아니었으나,

중간(5점)이상의 평점 6점~7점대에서 인기도가 가장 높은 것을 보아

어느정도 평점이 높은것이 인기도가 높아지는데 상관관계가 없진 않아 보인다고 할 수 있음.

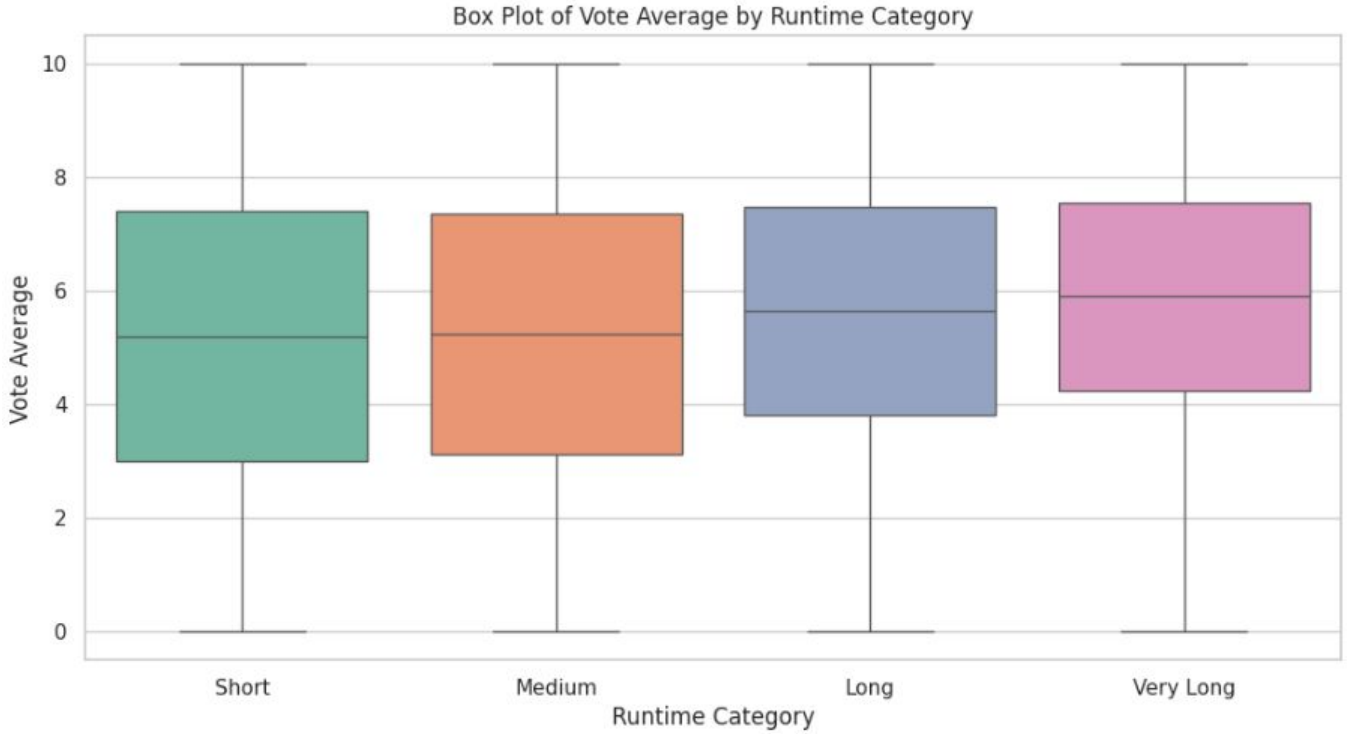
결과 원인 예측:

평점을 평가하지 않은 사람들의 점수는 0점으로 집계되서 0점의 인기도가 높을수도 있을 것 같고,

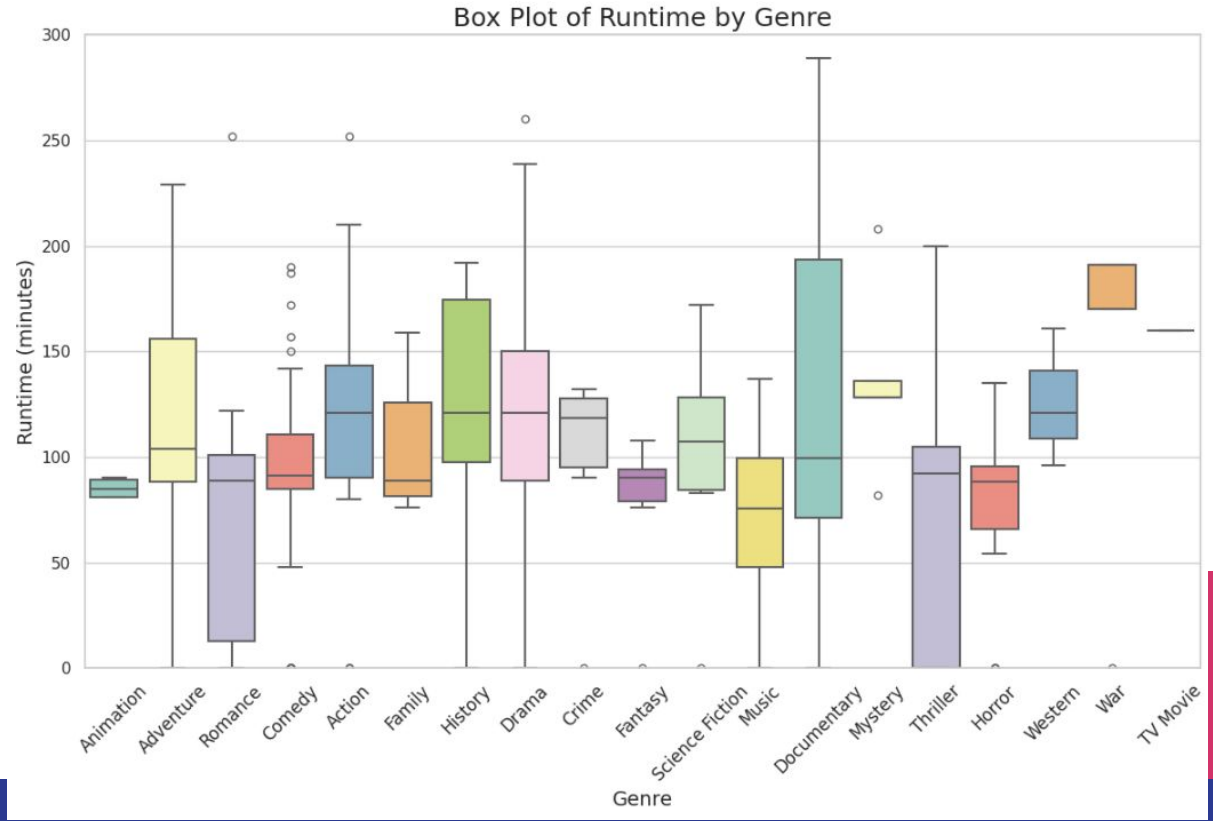
평점측정 시스템을 하지않은 인기영화들은 0점으로 집계되서 이런 결과가 나왔을수도 있을 것 같다.

인기도가 높을수록 많은 사람들이 평점을 투표해서, 정확하지 않은 평점의 집계가 많이 됐을 수도 있을 것 같다.

런타임과 평점간의 상관관계분석

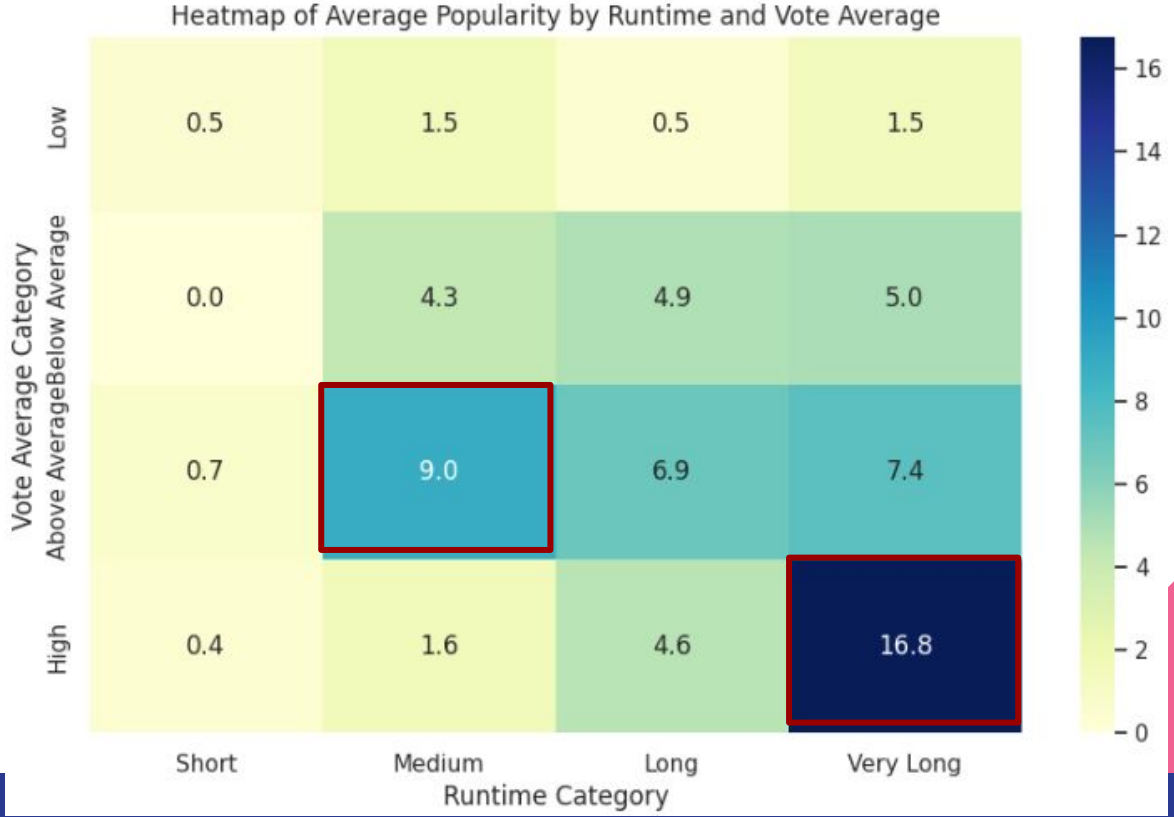


장르별 런타임 비교


Box Plot of Runtime by Genre
Runtime (minutes)
300
250
200
150
100
50
0
Animation
Adventure
Romance
Comedy
Action
Family
History
Drama
Crime
Fantasy
Science Fiction
Music
Documentary
Mystery
Thriller
Horror
Western
War
TV Movie
Genre

런타임,평균평점이 인기도에 미치는 상관관계

```
bins=[0, 60, 100, 120, 300], labels=['Short', 'Medium', 'Long', 'Very Long'])   런타임분류
bins=[0, 4, 6, 8, 10], labels=['Low', 'Below Average', 'Above Average', 'High'])평균평점분류
```



(런타임과 평균평점),인기도의 상관관계 결론설명

1. **Medium 런타임 & Above Average**: 이 조합에서 평균 인기도가 가장 높습니다(9.0). 이는 **중간 런타임을 가진 영화들**이 관객에게 **긍정적인 평가를 받는 경향**이 있음을 시사합니다.
2. **Very Long 런타임 & High**: 이 조합에서도 인기도가 높고(16.8), 이는 **긴 런타임을 가진 고평가 영화들이 존재함**을 나타냅니다.
3. **Short 런타임 카테고리**: **모든 투표 평균 카테고리에서 상대적으로 낮은 인기도**를 보이고 있습니다. 이는 **짧은 런타임의 영화들이 인기도가 낮은 경향**이 있음을 의미할 수 있습니다.