제작사 중심으로 영화의 흥행 요인 분석

김민진

SeSAC 영등포 6기 데이터 AI 개발 과정

목차

01 목적

키워드

제작사

시기

소비자

- 영화 흥행에 성공하는 요인을 파악하여 영화 제작 투자 의사 결정에 도움
- 영화 제작 과정에 흥행 가능성을 높일 수 있음

02 데이터셋 소개

이 분석에 사용된 데이터셋은 IMDb에서 제공하는 공개 데이터로, 1900년대 초부터 현재까지의 영화 정보를 포함합니다.

- credits.csv: 영화에 대한 출연진 및 제작진 정보를 포함하고 있는 파일로, JSON 형식으로 제공됩니다.
- keywords.csv: 영화의 줄거리 키워드를 포함하는 파일로, JSON 형식으로 제공됩니다.
- links.csv: 모든 영화의 TMDB 및 IMDb ID를 포함하고 있는 파일입니다.
- movies_metadata.csv: 영화 포스터, 배급사, 예산, 수익, 개봉일, 언어, 제작 국가 등의 정보가 포함되어 있습니다.
- ratings.csv: 영화에 대한 100,000개의 평점이 포함된 파일로, 영화의 대중적 평가를 분석할 수 있는 자료입니다.

03 데이터 전처리

movies_metadata.csv	release_date	연도(release_year)와 월(release_month)을 추출
	budget, revenue, popularity	수치형 데이터변환
	production_companies	제작사 이름(company_list)만 추출하여 리스트 형식으로 저장
	genres	JSON 데이터를 파싱하여 리스트 형식으로 변환
	companty_list, release_year	각 영화와 개봉연도와 제작사를 매칭한 데이터셋 생성
	runtime	결측치 처리
keywords.csv	id, keywords	JSON 데이터를 파싱하여 리스트 형식으로 변환

04 데이터 분석(EDA)

1. 시장 점유율 분석

2. Performance 분석

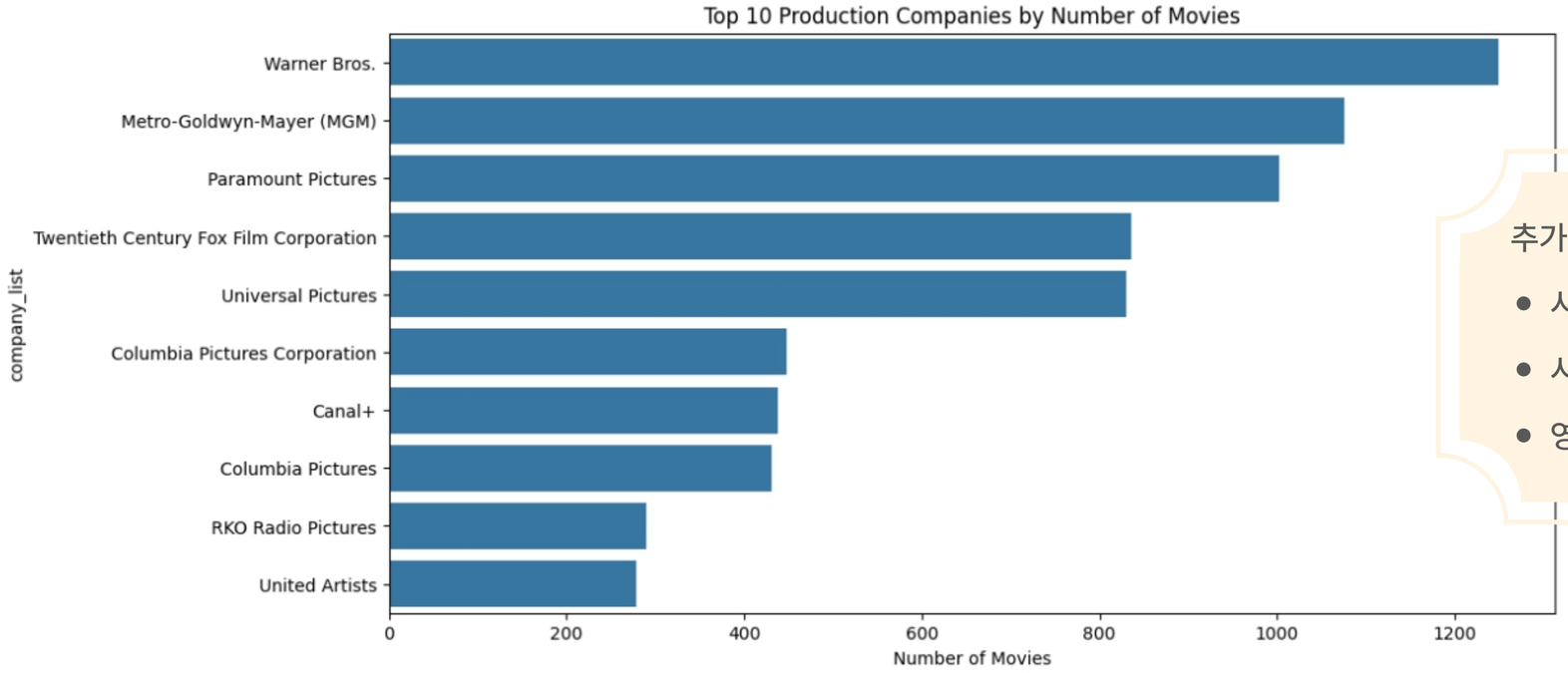
3. 제작사별 장르 분석

4. 시간별 제작 추이 분석

5. Company Collaborations 분석

1. 시장 점유율 분석

- 영화 제작사 별 영화 제작 횟수를 분석하여 가장 많은 영화를 제작한 상위 10개 제작사를 시각적으로 표현
- 어떤 제작사 들이 영화 산업에서 가장 활발히 활동하고 있는지를 시각적으로 확인 가능. (Warner Bros., MGM, Paramount 순)
- (x축 : 영화 수, y축 : 제작사)

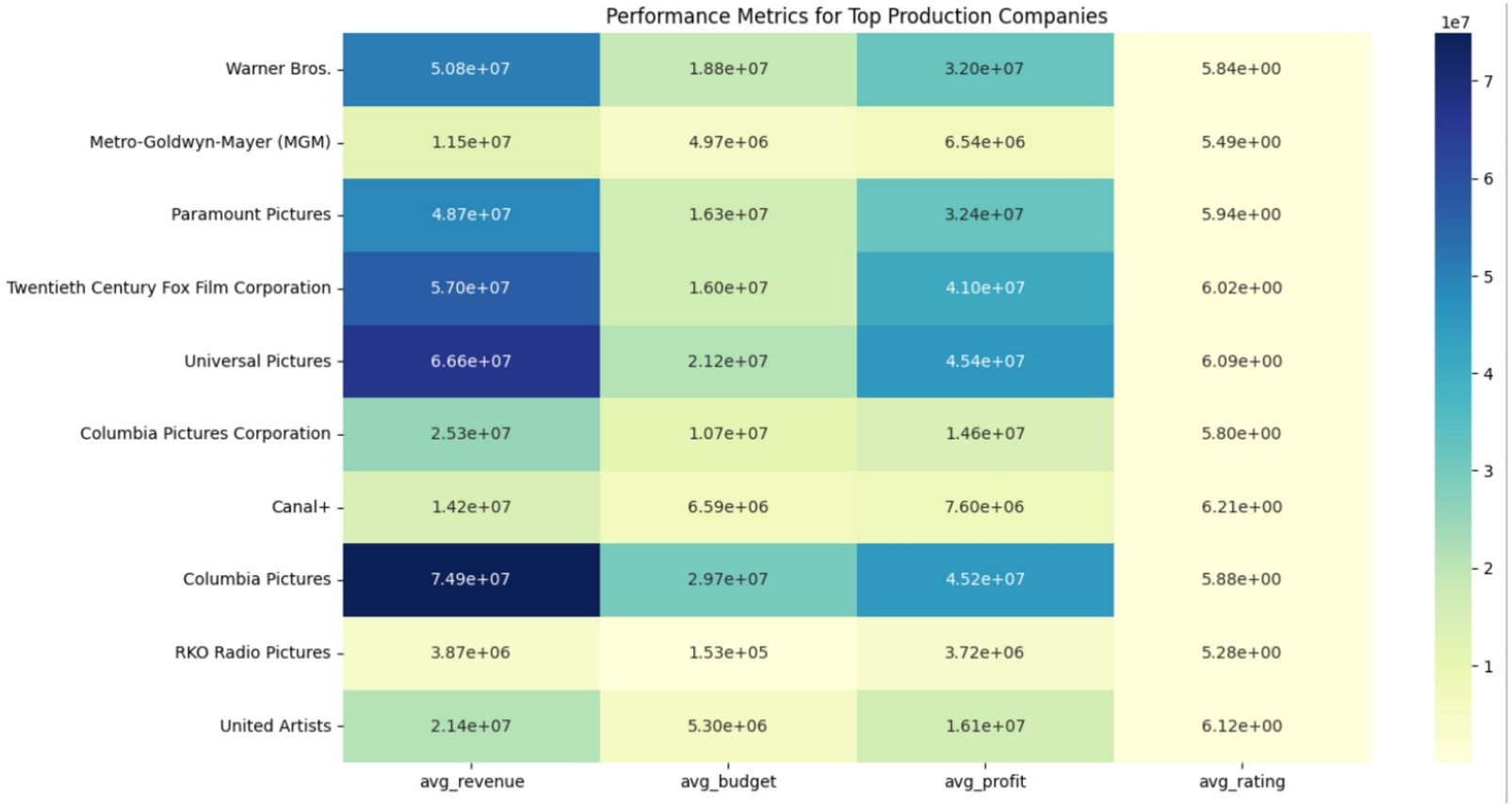


추가 분석 방향:

- 시장 점유율 백분율 계산
- 시간에 따른 시장 점유율 변화 (e.g.,5년 마다)
- 영화 예산 범위(저예산, 중예산, 고예산)별 시장 점유율 분석.

2. Performance 분석

- 각 상위 회사의 평균 수익, 예산, 이익, 평점을 계산
- 히트맵을 통해 각 제작사의 성과를 한눈에 비교할 수 있음.

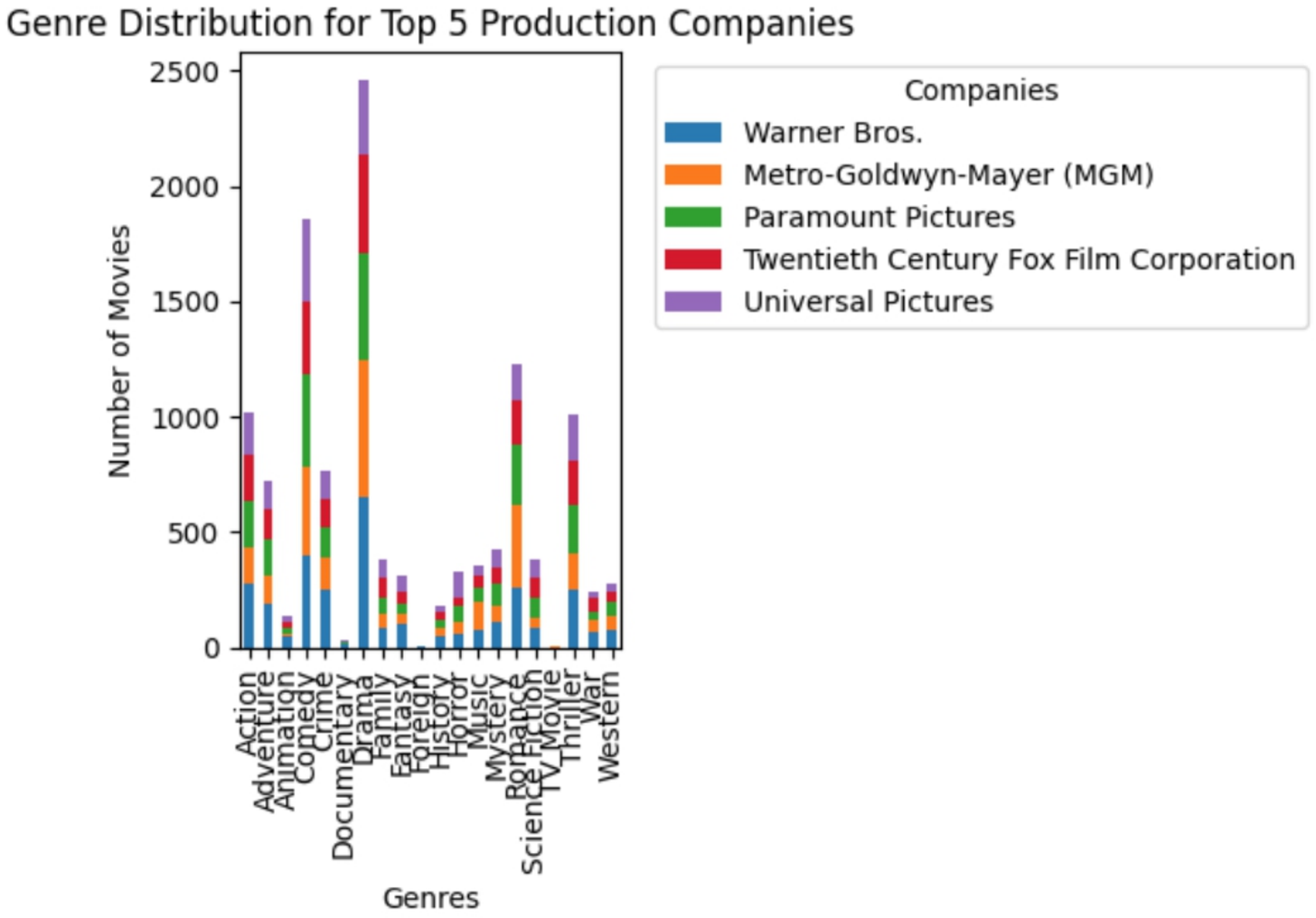


추가 분석 방향:

- ROI(투자 수익률) 계산
- 국내 vs. 국외 성과 비교.

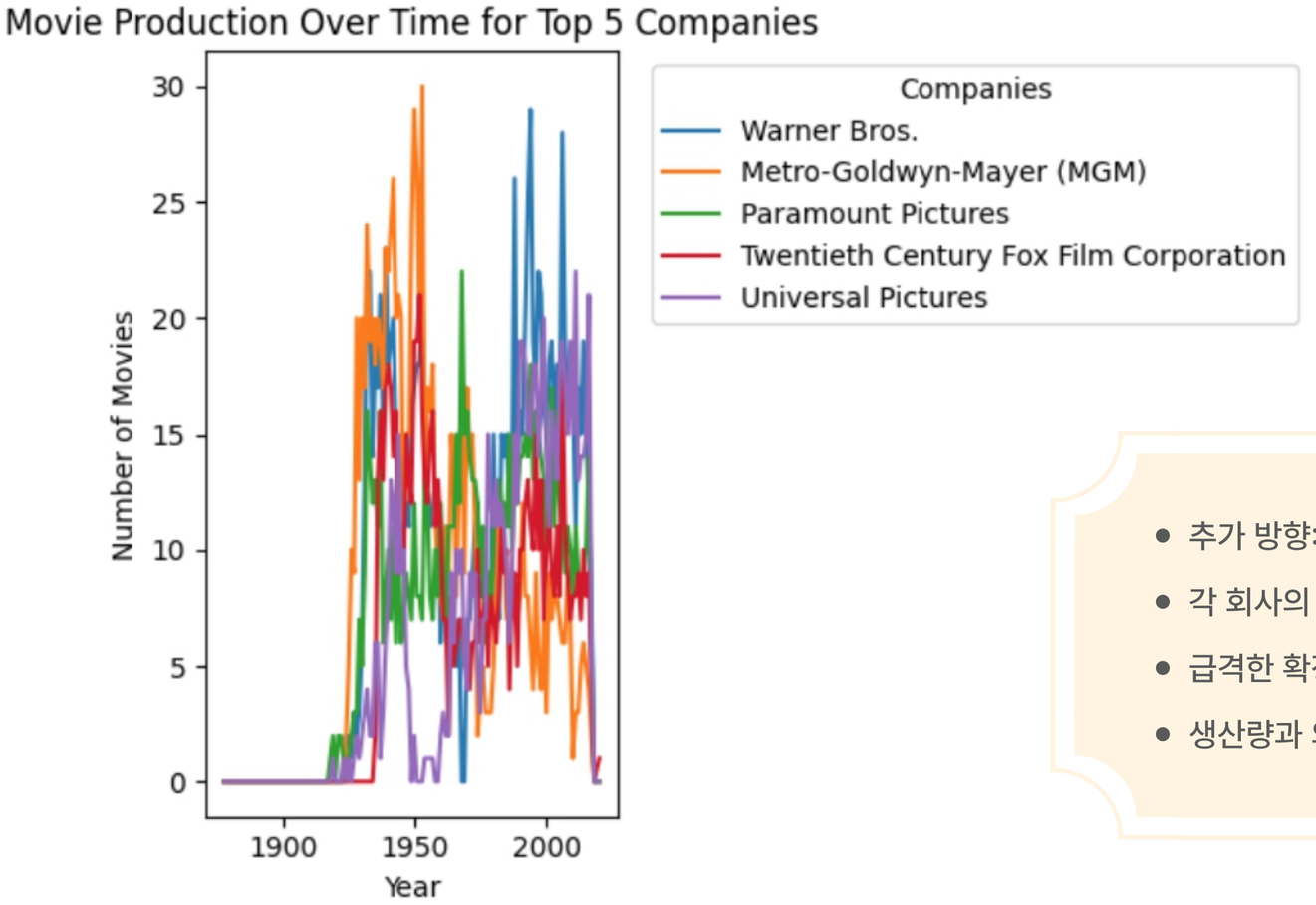
3. 제작사별 장르 분석

- 상위 5개 영화 제작사의 장르별 특성을 분석
- 각 제작사가 주로 어떤 장르의 영화를 제작하는지 시각적으로 보여줌



4. 시간별 제작 추이 분석

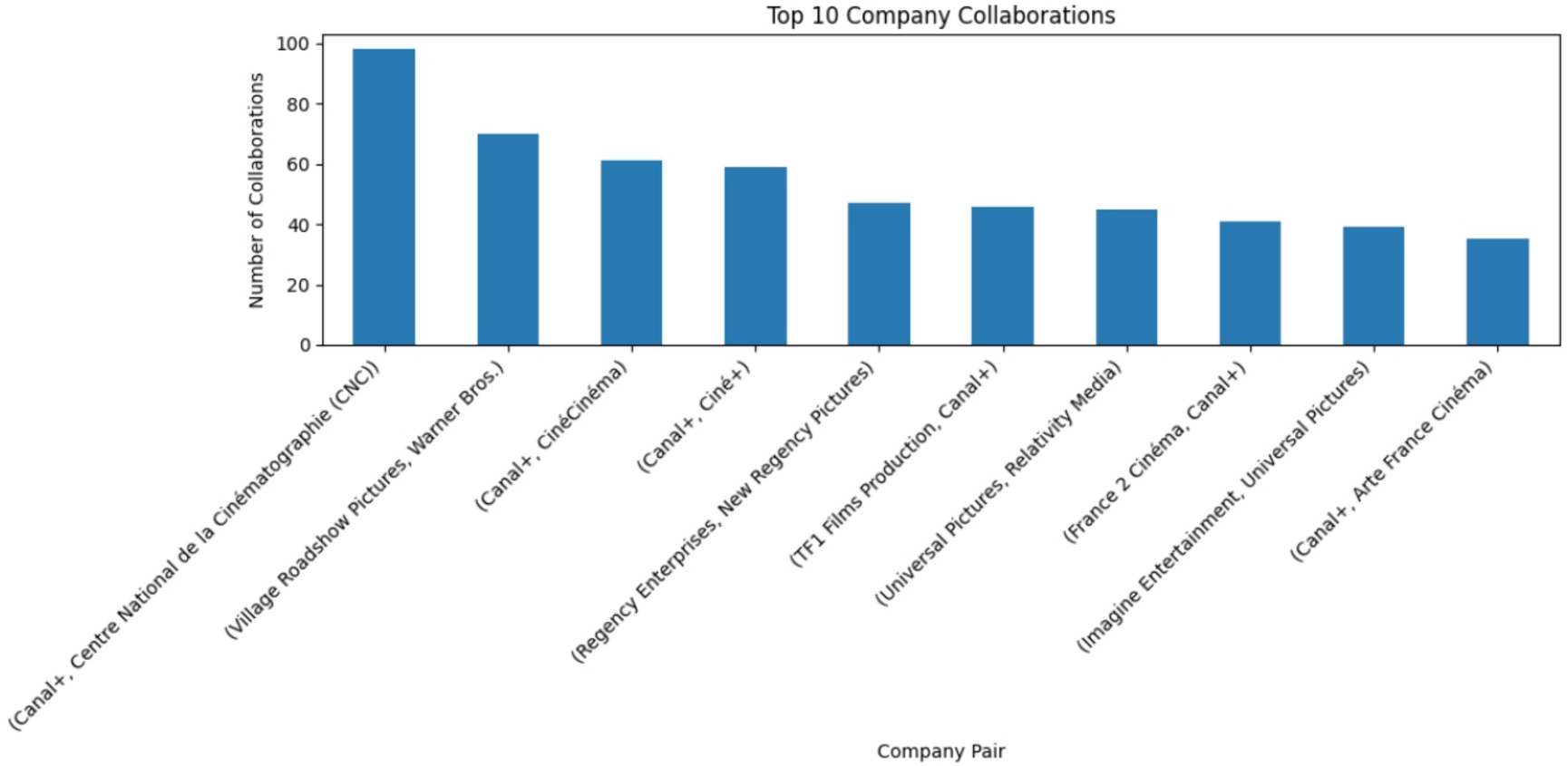
- 시간에 따라 각 제작사의 영화 제작 활동이 어떻게 변화했는지



- 추가 방향:
- 각 회사의 연도별 성장률 계산.
- 급격한 확장 또는 축소 시기 판별.
- 생산량과 외부 요인(예: 경제 지표, 주요 산업 변화) 간의 상관 관계 분석.

5. Company Collaborations 분석

- 각 영화의 제작사 목록에서 협업한 제작사 쌍을 추출한 후, 그 빈도를 계산해 가장 많이 협력한 제작사 쌍을 시각화.
- 어느 제작사들이 자주 협력하는지를 한눈에 확인 가능.
- 영화 산업에서 특정 제작사들 간의 협력 관계를 파악할 수 있음.



05 평점 예측 모델

Linear Regression	RMSE : 1.8469	1. 예측된 평점이 실제 값과 크게 차이가 남. 2. RMSE 값이 가장 높다. R^2 값이 0.0407로 매우 낮다.
	R^2 Score : 0.0407	
Random Forest	RMSE : 1.5244	1. RMSE가 Linear Regression에 비해 낮아짐 2. R^2 Score가 0.3464로 개선되었으나, 예측력이 완벽하지는 않음
	R^2 Score : 0.3464	
XGBoost	RMSE : 1.4577	1. 가장 낮은 RMSE와 가장 높은 R^2 Score를 기록. R^2 Score가 0.4024로 약 40.24%의 데이터 변동성을 설명.
	R^2 Score : 0.4024	

- 가장 우수한 모델: XGBoost 모델은 가장 낮은 RMSE와 가장 높은 R^2 Score를 기록하여,
  주어진 데이터에서 평점을 예측하는 데 가장 적합한 모델로 평가됨.
- 모델 성능 개선: 세 모델 모두 R^2 Score가 1에 가깝지 않기 때문에 모델의 한계가 있음
- 추가적인 피처 엔지니어링, 하이퍼파라미터 튜닝, 또는 추가적인 데이터 수집 등을 통해 모델의 성능을 더욱 개선할 수 있음

06 결론

01 Drama, Comedy는 영화의 흥행의 지름길

02 추가적인 ROI(투자수익률) 계산 필요

03 비선형적인 관계를 잘 학습할 수 있는 XGBoost이 효과적임

결론

- 다른 팀들과의 결과를 합치면 영화 제작에 실패가 없을 것이다.
- 평점 예측 모델은 더 많은 데이터와 더 세밀한 전처리가 필요할 듯 하다.

감사합니다