
Paramount
100 Years
A VIACOM COMPANY

투자 대비 수익률 분석을 통한 영화 투자 전략 수립을 위한 데이터 분석

프로젝트 요약

–

한국 영화산업 위기:

영화 수익률 악화
<한국 상업영화 평균 수익률>
2019년: +10.9%
2020년: –30.3%
2021년: –22.9%
2022년: –12.6%
2023년: –31%
2024년: –16.4%

목표:

팬데믹 이후 침체된 한국 영화산업의 구조적 문제 파악
평점, 예산, ROI 등 다양한 데이터 분석
산업 회복을 위한 인사이트와 영화 투자 제언

역할 및 기여:

세계 영화와 국내 영화산업 데이터 수집 및 정리 후 비교
수익률, 투자 규모 등 데이터 분석
투자·수익률 감소의 구조적 원인 진단 및 전략 수립
정부 정책 및 모태펀드 현황 정리, 비즈니스 및 정책 제안 방향성 도출

차례

1. 목표 및 문제 정의

목표

1) 투자 효율 극대화

어떤 장르, 배우, 제작비 규모가 높은 ROI를 내는가?

데이터를 통해 '투자 대비 수익'을 극대화할 수 있는 최적의 조건을 찾고자 함.

2) 리스크 관리 및 예측

제작비 대비 실패 가능성이 높은 조건(장르, 캐스팅, 러닝타임 등)을 조기 인지해 리스크를 줄임.

평점, 관객 수, 과거 수익 데이터를 통해 흥행 성공 여부를 예측하려고 함.

3) 편견 해소 및 객관적 의사결정

"주연급 배우가 있으면 무조건 성공한다"는 등 기존의 제작사 내부 편견을 데이터로 검증 및 보완.

데이터를 바탕으로 보다 과학적이고 객관적인 투자 판단 체계 구축.

1. 목표 및 문제 정의

문제정의

1) 스타 배우 효과에 대한 과대평가

주연 배우가 유명하면 반드시 높은 수익을 보장할까?

2) 장르별 수익률에 대한 편견

전통적으로 인기 있다고 알려진 장르(예: 로맨스, 액션)가 항상 높은 ROI를 낼까?

3) 평점과 흥행의 관계 과신

좋은 평점이 높은 수익으로 연결될까?

2. 데이터 수집

데이터 출처 MovieLens

MovieLens는 미국 미네소타 대학교의 'GroupLens 연구소'에서 운영하는
영화 추천 시스템 연구를 위한 '공개 데이터셋 프로젝트'

전 세계 연구자들이 '영화 추천 알고리즘 연구'에 사용함.

주요 파일

`movies-metadata.csv` – 영화 정보 (제목, 장르 등),
총 영화 수 45460개, 개봉년도 1974~2020
`ratings.csv` – 사용자-영화 간 평점 (2천만 개 이상)
`links.csv` – MovieLens movieId ↔ TMDB/IMDB ID 매핑

https://grouplens.org/datasets/movielens/

EDA HISTORY - 결측지 확인

```
print("\n! 결측치가 있는 컬럼:")
missing = metadata.isnull().sum()
missing = missing[missing > 0]
print(missing.sort_values(ascending=False))
```

```
! 결측치가 있는 컬럼:
belongs_to_collection    40972
homepage                 37684
tagline                  25054
main_country              6288
overview                   954
poster_path                386
runtime                    263
release_date                90
year                        90
status                      87
imdb_id                     17
original_language           11
revenue                      6
video                        6
title                        6
vote_count                   6
spoken_languages             6
vote_average                 6
popularity                   5
production_companies         3
production_countries         3
dtype: int64
```

EDA HISTORY - ROI 계산

ROI = (총수익 - 투자 비용) / 투자 비용 × 100%

```
print(metadata['budget'].dtype)  # object
print(metadata['revenue'].dtype)  # float64
```

1. budget, revenue 타입확인, budget 컬럼을 숫자형으로 변환

2. ROI 계산 → budget이 0이거나 결측인 경우 ROI 값 제거 (0 나누기 오류 방지)

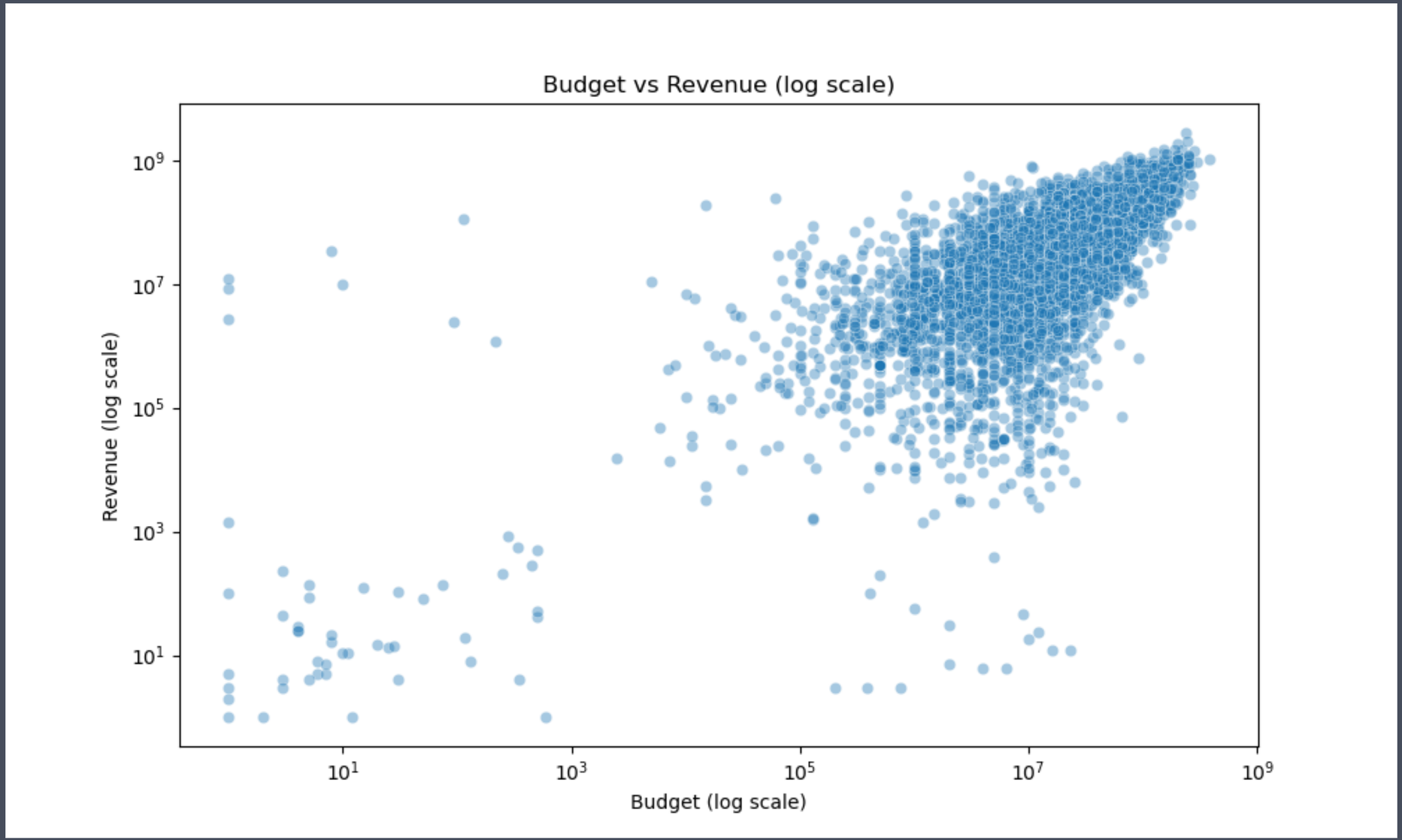
```
# budget, revenue 유효성 체크
metadata = metadata[(metadata['budget'] > 0) & (metadata['revenue'] > 0)]
```

EDA HISTORY - 대표 장르 추출

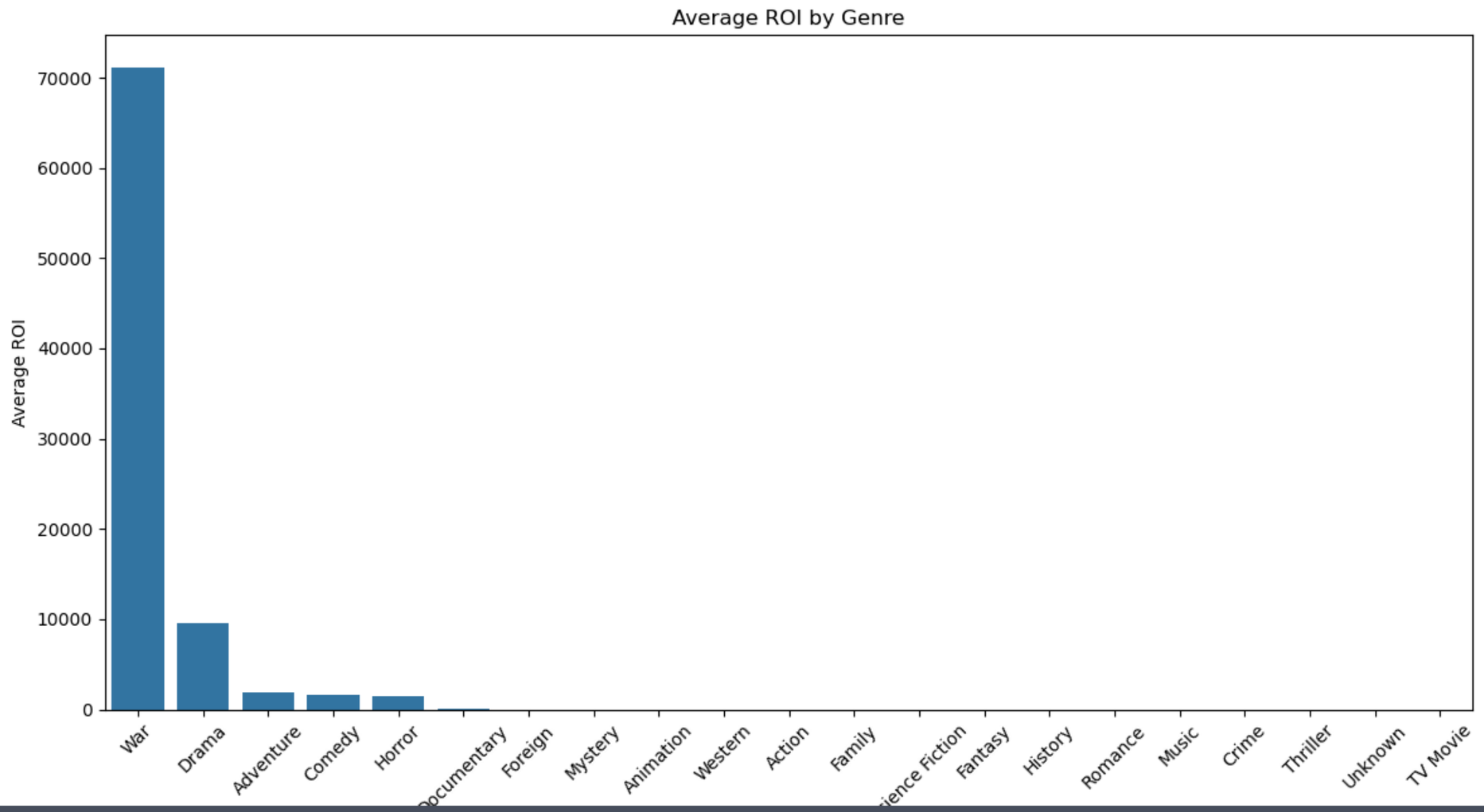
```python
# 장르 데이터 처리 (문자열을 리스트로 변환)
def parse_genres(genres_str):
    try:
        genres_list = ast.literal_eval(genres_str)
        return [genre['name'] for genre in genres_list]
    except:
        return []

movies['genres_parsed'] = movies['genres'].apply(parse_genres)

# 가장 대표 장르 하나만 선택 (복수 장르 영화는 첫 번째 장르로 대표)
movies['main_genre'] = movies['genres_parsed'].apply(lambda x: x[0] if len(x) > 0 else 'Unknown')
```

genres 문자열을 리스트로 변환 후 대표 장르 선정 → 다중 장르 영화는 첫 번째 장르로 대표

전세계 데이터 - 분석결과 1



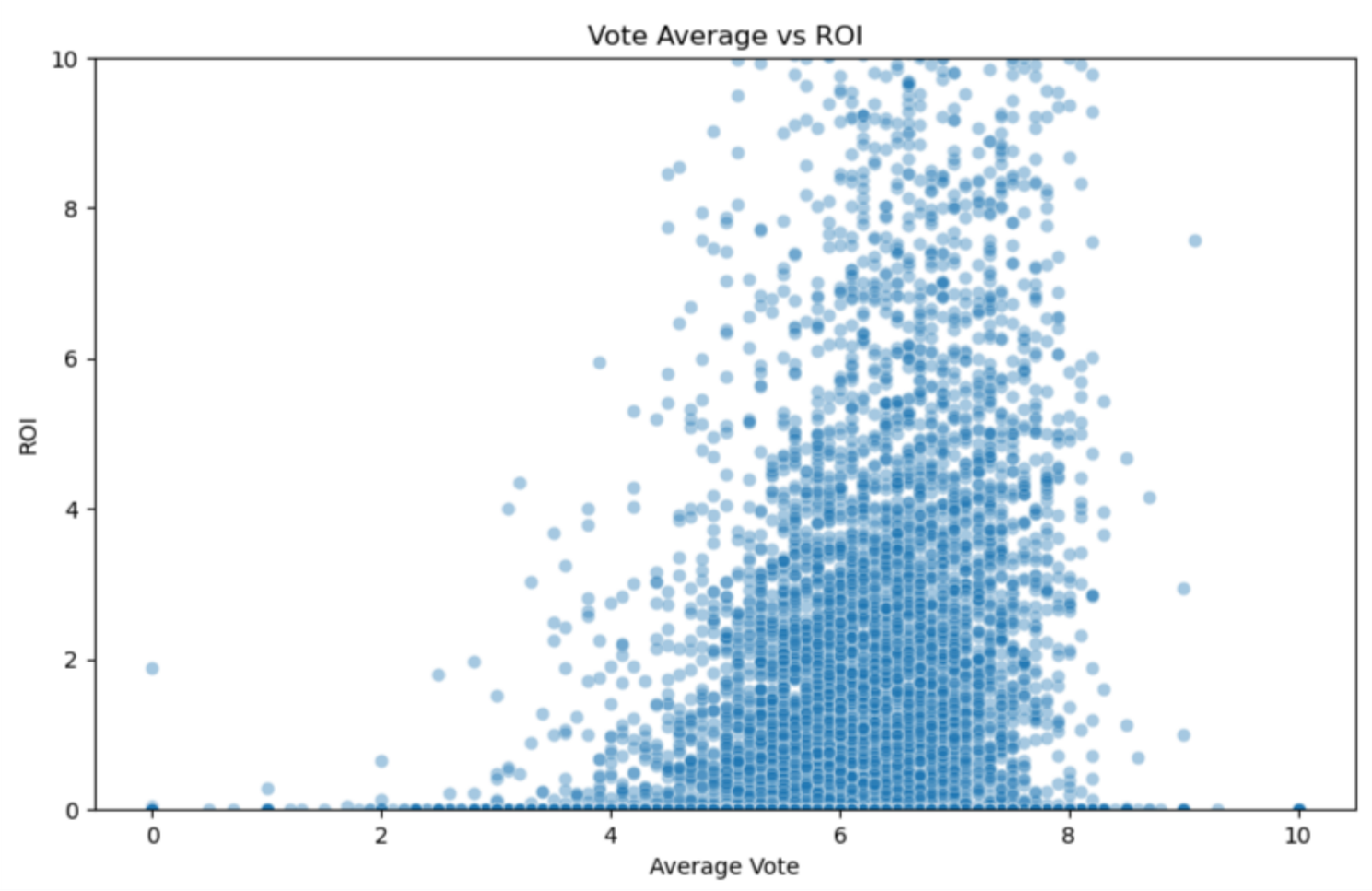
전세계 데이터 - 분석결과 2



전세계 데이터 - 분석결과 3

평점과 ROI 간에 뚜렷한 상관관계 없음

평점과 ROI 관계 산점도



국내 데이터 전처리 한국 영화 458개

```python
def extract_country(countries_str):
    if pd.isna(countries_str): return 'Unknown'
    try:
        countries = ast.literal_eval(countries_str)
        if countries: return countries[0]['name']
    except:
        pass
    return 'Unknown'

metadata['main_country'] = metadata['production_countries'].apply(extract_country)
metadata['is_korean'] = metadata['main_country'].apply(lambda x: 'South Korea' in x or 'Korea' in x)
```

1. 제작 국가 정보가 문자열로 되어 있어서 파싱

production_countries 컬럼은 문자열 형태로 리스트가 있음.
예) "[{'iso_3166_1': 'KR', 'name': 'South Korea'}]"
→ 이걸 리스트+딕셔너리 형태로 바꿔야 국가명을 꺼낼 수 있음.

2. 한국 영화 True/False로 구분

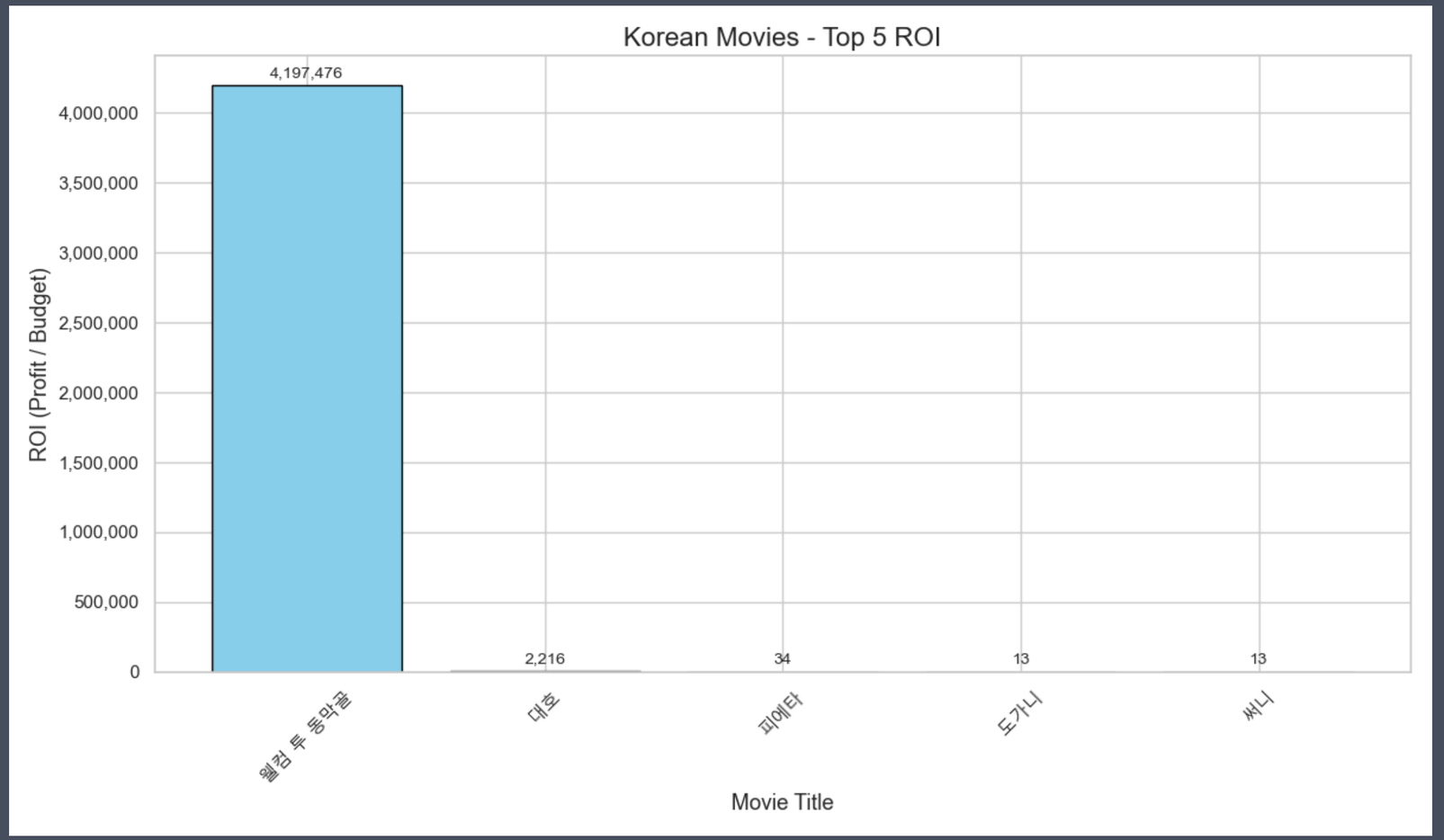
'is_korean' 플래그 생성, 필터링 용이

국내 데이터 분석결과 1

'웰컴 투 동막골' ROI가 매우 높음. '가성비甲' 투자처

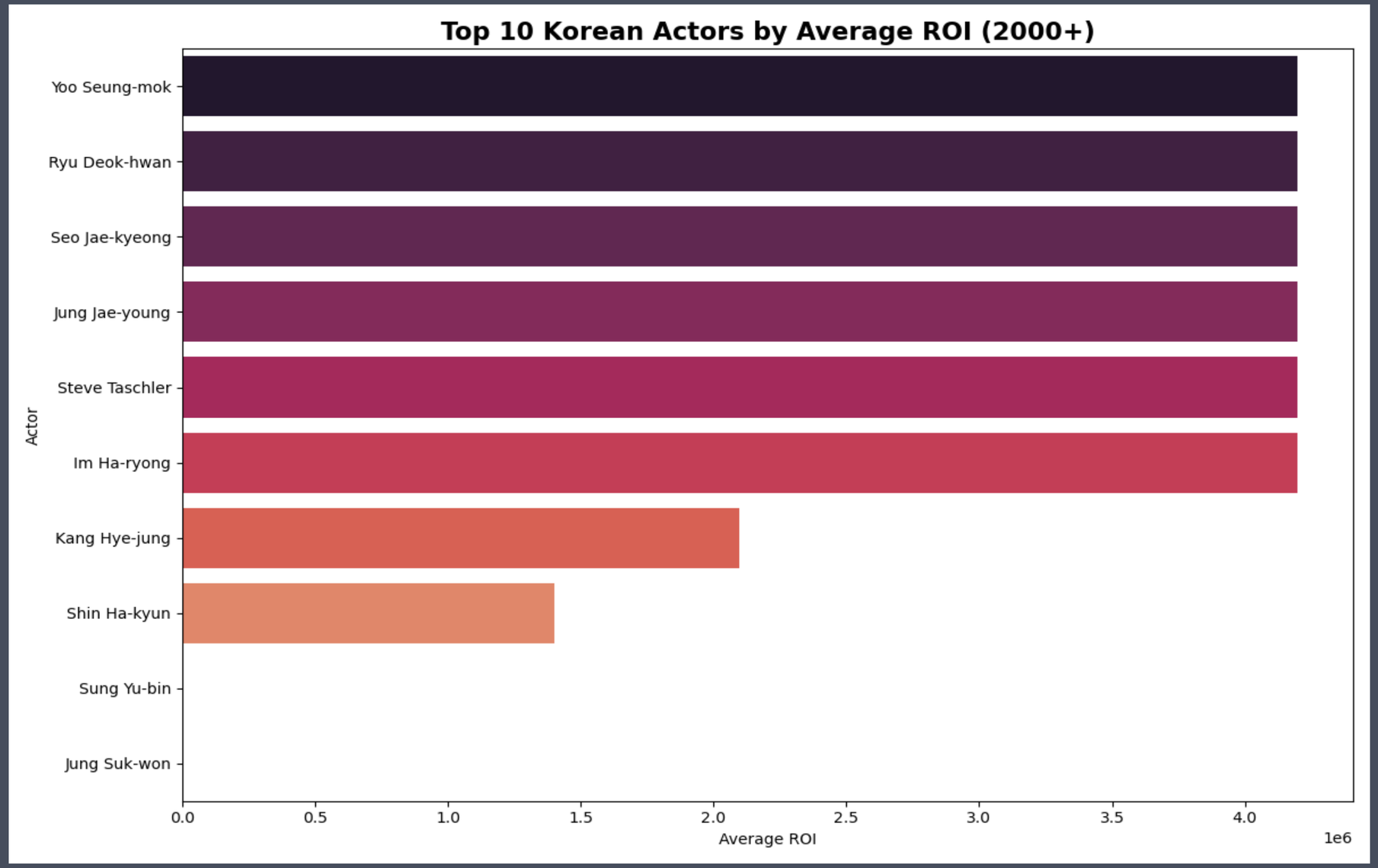
저예산 + 대중적 스토리 + 신선한 출연진 + 장르 믹스 + 타이밍 전략

→ 단순 스타 배우나 고예산 의존 없이도 대박 가능성을 극대화한 전략적 투자 사례



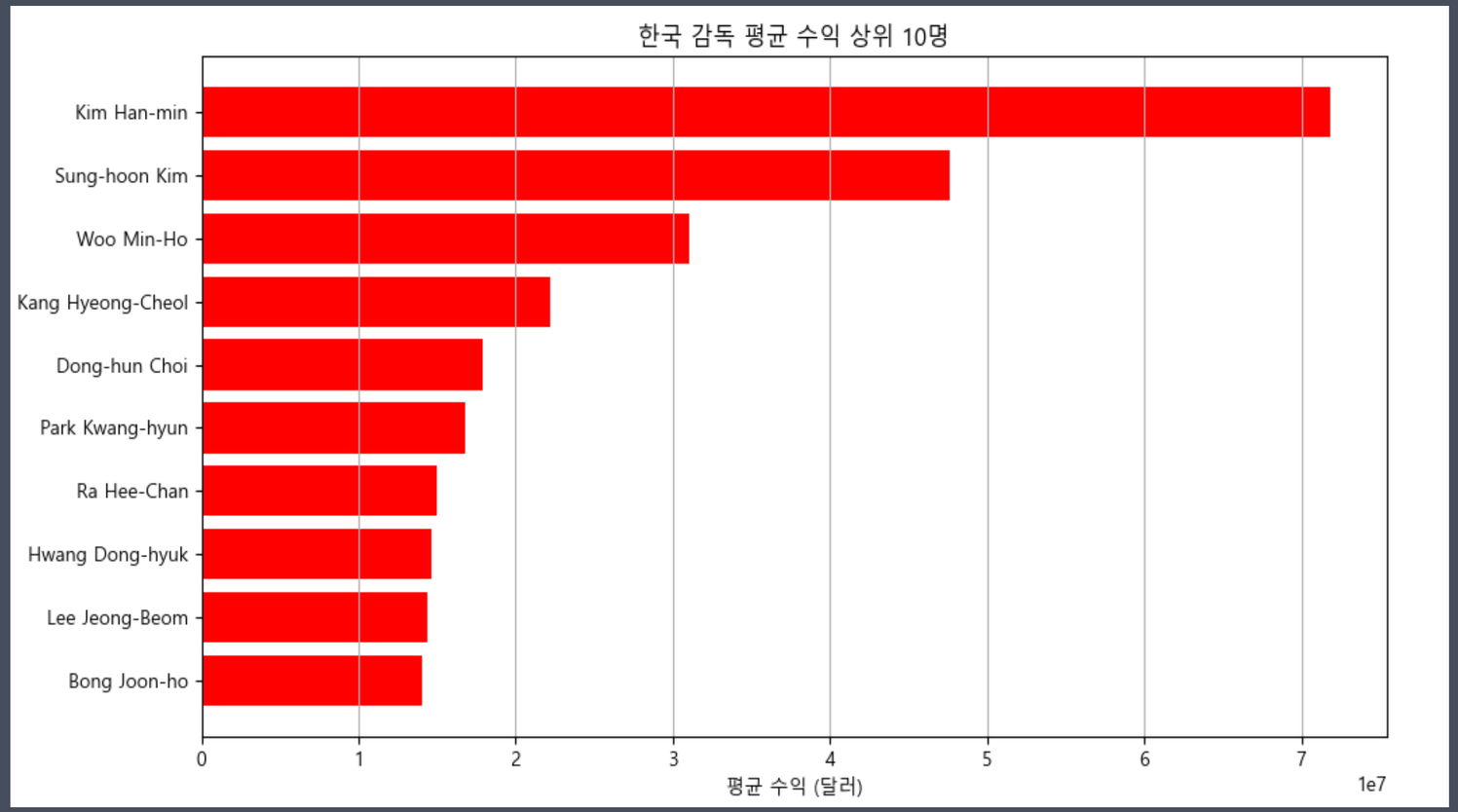
국내 데이터 분석결과 2

웰컴투 동막골 출연 배우가 순위권



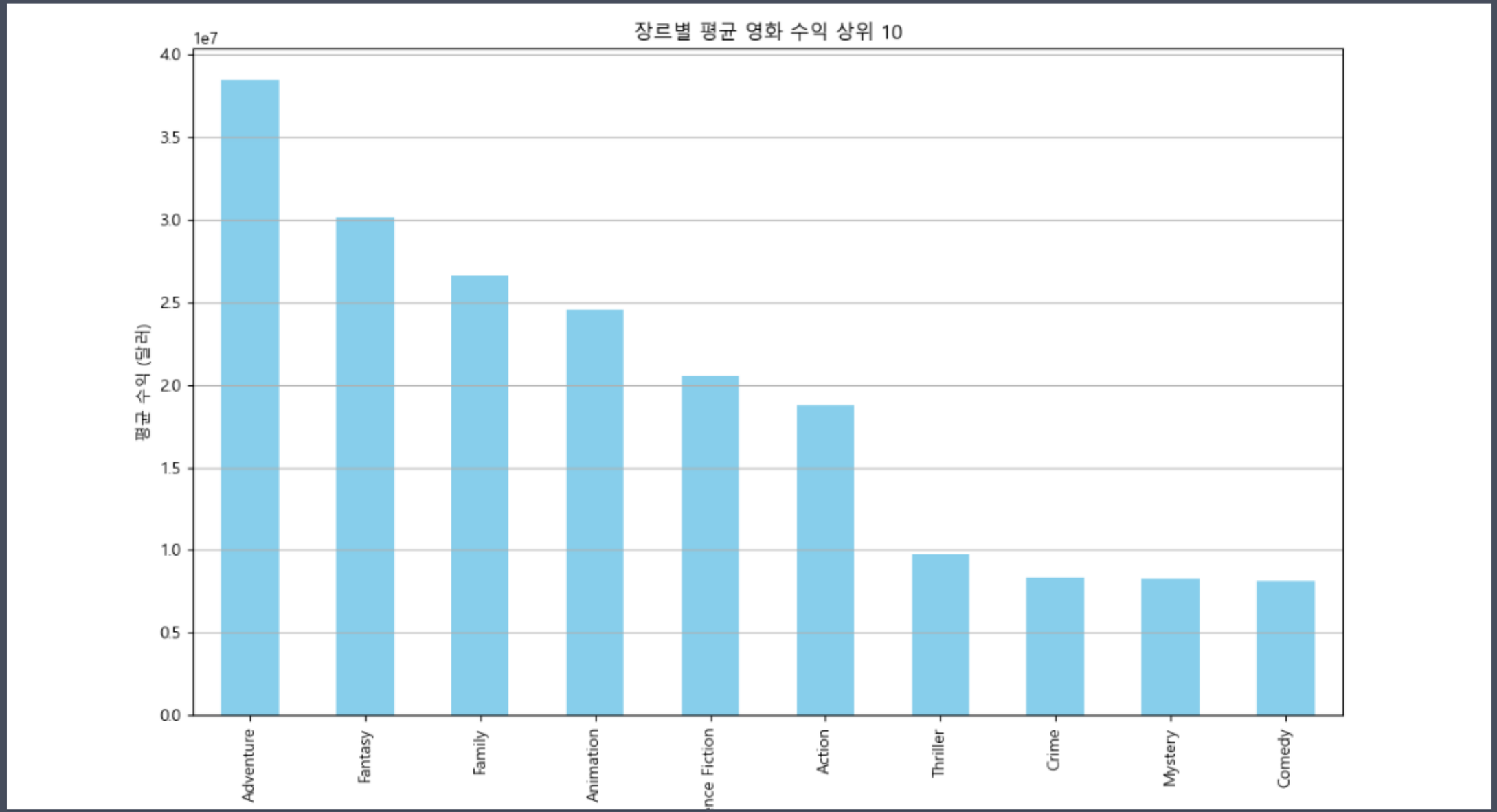
국내 데이터 분석결과 3

각 감독이 만든 모든 영화의 총 수익(revenue)



감독 이름	대표작	특징 및 설명
Kim Han-min (김한민)	《명량》(2014), 《한산: 용의 출현》(2022)	역사 대작
Sung-hoon Kim (김성훈)	《터널》(2016), 《특별시민》(2017), 《PMC: 더 벙커》(2018)	스릴러, 액션 장르
Woo Min-Ho (우민호)	《내부자들》(2015), 《더 킹》(2017)	정치 스릴러 장르
Kang Hyeong-Cheol (강형철)	《과속스캔들》(2008), 《럭키》(2016)	코미디 영화
Dong-hun Choi (최동훈)	《암살》(2015), 《도둑들》(2012), 《전우치》(2009)	액션, 흥행 대작
Park Kwang-hyun (박광현)	《연가시》(2012), 《타짜: 신의 손》(2014)	다양한 장르
Ra Hee-Chan (라희찬)	《바르게 살자》(2007)	코믹
Hwang Dong-hyuk (황동혁)	《기생충》(제작 참여), 《도가니》(2021), 《수상한 그녀》(2014)	오징어게임 연출
Lee Jeong-Beom (이정범)	《추격자》(2008), 《만추》(2011)	스릴러 장르
Bong Joon-ho (봉준호)	《기생충》(2019), 《괴물》(2006), 《설국열차》(2013)	아카데미 수상 감독

국내 데이터 분석결과 4


장르별 평균 영화 수익 상위 10
1e7
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
평균 수익 (달러)
Adventure
Fantasy
Family
Animation
Action
Thriller
Crime
Mystery
Comedy

결론

1. 유명 배우 출연 ≠ 수익 보장	실제 ROI는 배우보다는 저예산+기획력+마케팅 전략이 더 큰 영향
2.특정 장르가 항상 수익성이 높지 않음	여러 장르를 섞는 포트폴리오 전략이 유효
3. 좋은 평점이 반드시 높은 수익으로 이어지지 않음	다만 평균 이하 평점 영화는 실패 확률↑, 따라서 평점은 리스크 필터링 도구로 활용 가능

<투자 전략>

스타 배우 의존 → X
특정 장르 맹신 → X
평점 = ROI 착각 → X

→ 대신 데이터 기반, 분산 투자, ROI 중심 전략 필수

회고

데이터 한계 및 특이사항

한국 영화 데이터는 총 458개로 제한적이며, OTT(온라인 스트리밍) 관련 데이터는 포함되지 않음
따라서 OTT 시장 동향 분석 및 한국 영화 전체 트렌드 파악에는 한계가 존재