🎬 한국 영화 데이터 분석 프로젝트 보고서 (2005–2019)

날짜	@2025년 8월 18일 → 2025년 8월 29일

1. 서론

본 프로젝트의 목표는 **2005년부터 2019년까지의 한국 영화 데이터를 분석**하여 흥행에 영향을 미치는 주요 요인을 규명하고, 이를 통해 영화 산업 전반에 의미 있는 인사이트를 도출하는 것이다.

분석 기간을 2019년까지로 한정한 이유는, **코로나19 팬데믹 이후 영화 시장의 왜곡된 특수 상황을 배제**하기 위함이며, 동시에 **2000년대 중반부터 본격적으로 자리 잡은 한류 확산기**를 반영해 한국 영화 산업의 온전한 흐름을 포착하려는 목적이었다.

사실 프로젝트는 **TMDB API 데이터**로 시작했다. 하지만 한국 영화 데이터가 충분하지 않아 **표본 부족 문제**에 직면했다. 특히 TMDB에는 예산·흥행 수익 정보가 미비했고, 한국 영화의 평점 데이터도 확보하기 어려웠다. 이를 보완하려고 **IMDB 평점 데이터**를 시도했지만, 한국 영화 데이터는 사실상 없어서 실패로 끝났다.

당시 TMDB의 `overview` 컬럼을 활용해 **키워드 분석**을 시도했다. 흥행한 영화의 개요에서 특정 키워드(예: 가족, 전쟁, 사랑)가 얼마나 자주 등장하는지 확인하며 흥행 요인을 간접적으로 파악했는데, 최종 데이터셋(KOFIC+네이버)에는 이런 텍스트 데이터가 없어 재현하지 못한 아쉬움이 있었다.

이런 여러 시행착오를 거치면서, 나는 **데이터 분석은 "어떤 데이터를 확보했느냐"에서 이미 절반이 결정된다**는 교훈을 얻었다. 즉, 분석의 성패는 알고리즘이나 모델링 이전에, 분석 목적에 필요한 데이터를 얼마나 정확히 정의하고 수집했는가에 달려 있었다.

---

2. 데이터 수집 및 전처리

2.1 데이터 수집

- **TMDB API (초기 시도)**: 기본 메타데이터 확보. (한계: 한국 영화 표본 부족, 평점/예산 부재)
- **IMDB (실패)**: 평점 보완 목적. (한국 영화 데이터 거의 없음)
- **KOFIC API (핵심)**: 관객 수·수익 데이터 제공. (**예산 데이터는 제공하지 않음**)
- **네이버 영화 크롤링**: 네티즌 평점, 비평가 평점 확보.
- **수동 입력**: 크롤링 불가 약 200편 직접 입력으로 품질 보강.

2.2 데이터 전처리

- **결측치 처리**: 예산은 전부 결측 처리. 수익/관객 데이터 없는 영화는 분석 제외.
- **텍스트 파싱**: 장르/감독/배우 문자열을 리스트로 변환.
- **파생변수 생성**: 개봉월·계절 변수, 런타임 구간 변수(≤100, 100–120, ≥120분).
- **데이터 통합**: TMDB·KOFIC·네이버를 `movie_id`로 병합 → 최종 약 900편 데이터셋.

2.3 한계와 문제의식

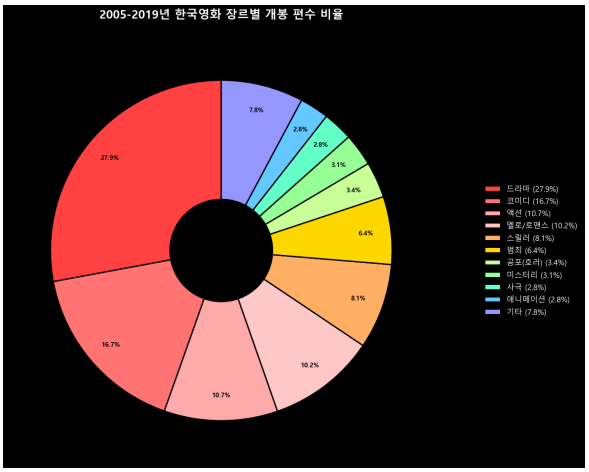
- **ROI 분석 불가**: 예산 데이터 부재로 수익성 분석 제한.
- **산업적 불투명성**: 한국 영화는 박스오피스 데이터는 공개되지만, 제작비·투자 내역은 불투명해 투자자·정책 입안자 모두에게 구조적 리스크.

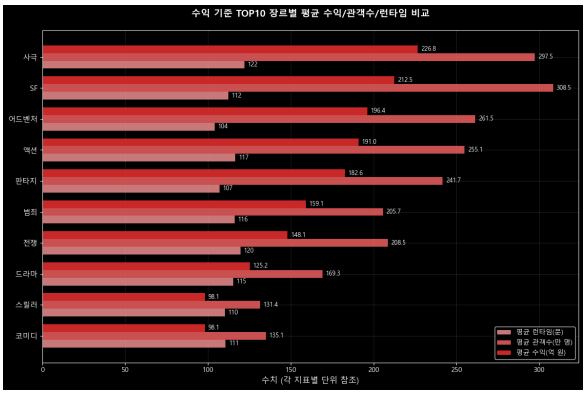
3. 분석 결과 및 시각화

과제 1️⃣ 장르별 영화 트렌드 및 흥행 성공 요인

장르 분포 및 흥행 강세

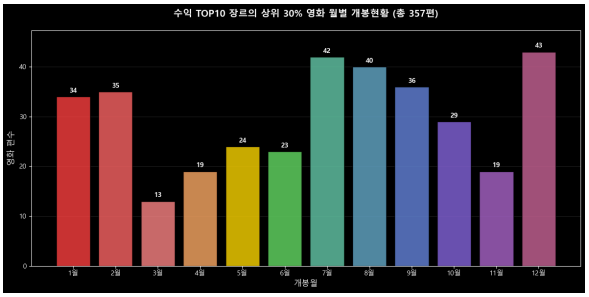
- 드라마(27.9%), 코미디(16.7%), 액션(10.7%), 멜로/로맨스(10.2%), 스릴러(8.1%) → 상위 5개 장르가 전체 절반 이상 차지.
- 평균 수익 상위 장르: 사극(226억), SF(212억), 어드벤처(196억).

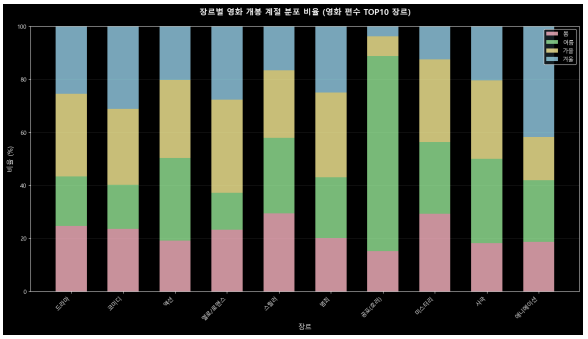




개봉 시기 특징

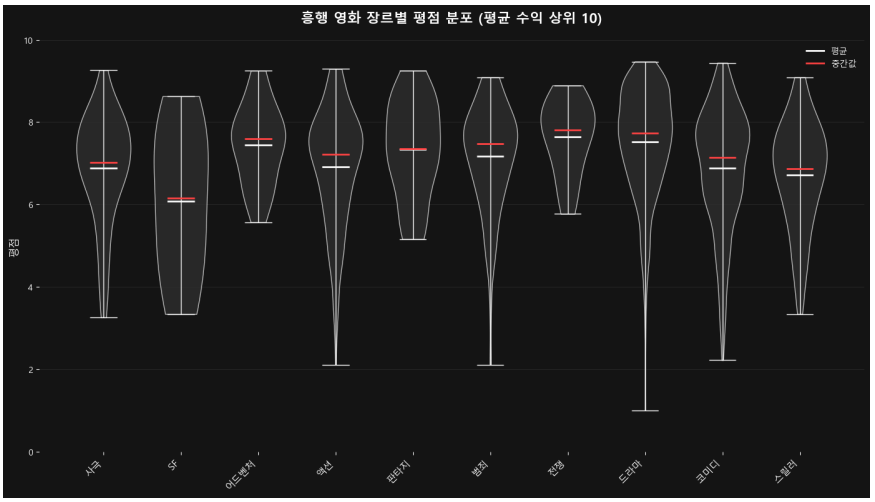
- 공포: 여름 집중 개봉 → ROI 극대화.
- 코미디·판타지: 겨울·연말 → 가족 단위 수요 집중.
- 드라마: 사계절 고른 분포 → 안정적 수익 구조.





장르별 수익 분포 (바이올린 플롯)

- 장르별 수익의 **분포와 편차**를 시각화한 결과:
  - **사극·액션·재난**: 상위 꼬리가 길게 뻗어 있음 → **대박 가능성 크지만, 편차도 크다**. 즉, 흥행이 성공하면 엄청나지만 실패 리스크도 높음.
  - **드라마·코미디**: 분포가 비교적 낮은 구간에 몰림 → **평균은 낮지만 안정적**. 큰 성공보다는 꾸준히 관객을 확보하는 장르.
  - **공포·애니메이션**: 전체적으로 분포가 좁음 → 시장 규모는 작지만, ROI 관점에서 효율적일 수 있음.



한국적 특수성: 사극

- 사극은 한국 영화만의 독특한 장르.
- 조선시대 역사·실존 인물 기반 서사가 대중적 흡인력과 문화적 정체성을 제공.
- 대규모 제작비·세트·스타 캐스팅이 결합해 반복적으로 흥행 성공.
- 대표작: <명량>, <광해>, <관상>, <사도>.
- → 사극은 **한국 영화 산업에서 문화적 자산이자 흥행 카드**.

인사이트 (과제 1)

- 장르별로 서로 다른 **흥행 공식** 존재 (사극=대규모 제작+명절/여름, 공포=여름 전략, 코미디=연말 가족 타깃).
- **장르별 수익 분포**를 보면, 일부 장르는 **리스크·리턴이 크고**, 일부는 **안정적 성과**를 내는 특징을 가짐.
- 사극은 한국 영화의 **문화적 정체성과 상업적 흥행 카드**라는 이중적 의미를 지님.
- 산업 전반적으로 특정 장르(사극, 액션, 재난)에 제작비가 집중 → **다양성 확보 필요**.

과제 2️⃣ 영화 평점과 흥행 수익 상관관계

상관계수 결과

- 네티즌 평점–수익: 0.22 / 비평가 평점–수익: 0.19 / 수익–관객수: 0.99

📊 변수 간 상관관계 행렬:

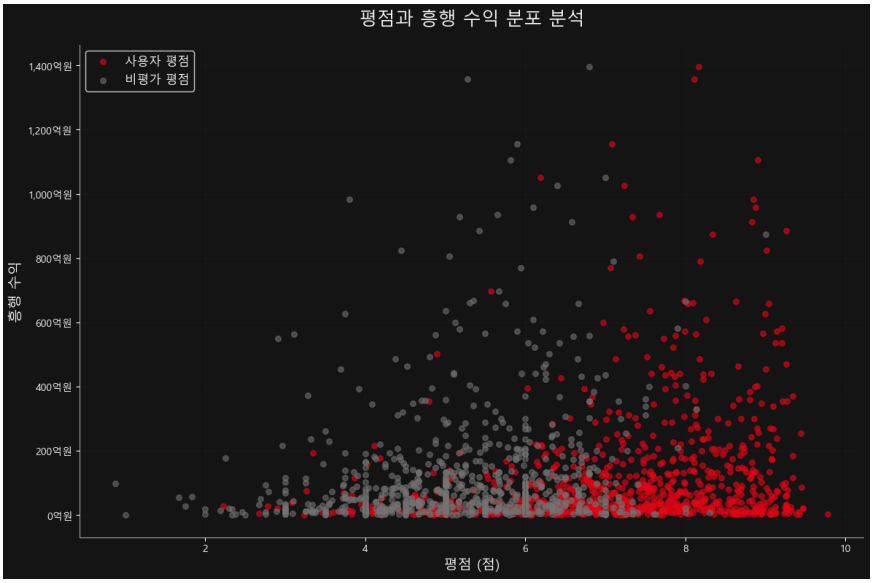
	vote_average_naver	critic_average	revenue	audience_total
vote_average_naver	1.0000	0.3696	0.2214	0.2331
critic_average	0.3696	1.0000	0.1970	0.1969
revenue	0.2214	0.1970	1.0000	0.9948
audience_total	0.2331	0.1969	0.9948	1.0000

📈 주요 상관관계 분석:

	변수1	변수2	상관계수
1	사용자 평점	흥행 수익	0.2214
2	비평가 평점	흥행 수익	0.1970
3	사용자 평점	비평가 평점	0.3696
4	흥행 수익	총 관객수	0.9948

💡 상관관계 분석 결과 해석:

	변수 조합	상관계수	해석
1	사용자 평점 - 흥행 수익	0.2214	약한 상관관계
2	비평가 평점 - 흥행 수익	0.1970	약한 상관관계
3	사용자 평점 - 비평가 평점	0.3696	보통 양의 상관관계
4	흥행 수익 - 총 관객수	0.9948	강한 양의 상관관계



산점도 분포 해석

- 비평가 평점(회색): **왼쪽(낮은 값)**에 치우침 → 박한 평가.
- 네티즌 평점(빨강): **오른쪽 분포** → 관대.
- 흥행 상위 영화: 네티즌 평점은 일정 수준 유지, 비평가 평점은 큰 차이 없음.

그룹별 특성 분석

1) 저평점·고흥행 그룹

- **대표작**: 《투사부일체》, 《군함도》, 《7광구》 등

🎬 저평점 & 고흥행 그룹 영화 리스트:

	영화제목	평점	흥행수익	장르	개봉월	제작사
1	마파도	6.141	166.3억원	코미디	3	코리아 엔터테인먼트, 씨제이엔터테인먼트, 씨제이엔터테인먼트, 무비클로저, 세방현상(주)
2	가문의 위기	6.330	284.9억원	코미디	9	(주)태원엔터테인먼트, (주)쇼박스, (주)쇼박스, (주)쇼박스, 세방현상(주), ...
3	작업의 정석	6.541	116.6억원	코미디	12	영화사청어람(주), (주)쇼박스, 영화사청어람(주), (주)쇼박스, 영화사청어람(주...
4	태풍	6.131	221.1억원	드라마, 액션	12	씨제이엔터테인먼트, 씨제이엔터테인먼트, 씨제이엔터테인먼트, ㈜라이브톤, 세방현상(주...
5	투사부일체	3.271	314.7억원	코미디	1	(주)시네마제니스, 씨제이엔터테인먼트, 씨제이엔터테인먼트, (사)부산영상위원회
...	...	...	...	...	...	...

- **주요 제작사 분포**

🏢 저평점 & 고흥행 영화의 주요 제작사:

	제작사	빈도수	비율(%)
1	씨제이엔터테인먼트	35	64.81
2	(주)씨제이이엔엠	30	55.56
3	(사)한국농아인협회	24	44.44
4	한국시각장애인연합회	24	44.44
5	(주)쇼박스	21	38.89
6	(주)넥스트엔터테인먼트월드(NEW)	19	35.19
7	롯데쇼핑㈜롯데엔터테인먼트	12	22.22
8	(주)콘텐츠판다	10	18.52
9	영화사청어람(주)	6	11.11
10	롯데컬처웍스(주)롯데엔터테인먼트	6	11.11

  - CJ ENM, 롯데, 쇼박스 등 **대기업 배급사** 중심 → 자본·스크린 장악력

- **주요 장르 비율**

🎭 저평점 & 고흥행 영화의 주요 장르:

	장르	빈도수	비율(%)
1	드라마	24	44.44
2	코미디	17	31.48
3	액션	17	31.48
4	범죄	9	16.67
5	멜로/로맨스	8	14.81
6	스릴러	8	14.81
7	사극	4	7.41
8	공포(호러)	3	5.56
9	SF	2	3.70
10	어드벤처	2	3.70

  - 액션, 범죄, 재난물이 압도적 → **볼거리·스케일 중심 장르**

- **개봉월 분포**

📅 저평점 & 고흥행 영화의 개봉월 분포:

	개봉월	빈도수	비율(%)
1	1	3	5.56
2	2	6	11.11
3	3	3	5.56
4	4	1	1.85
5	5	3	5.56
6	7	3	5.56
7	8	8	14.81
8	9	7	12.96
9	10	6	11.11
10	11	5	9.26
11	12	9	16.67

  - 여름(7–8월), 명절 시즌(설·추석)에 집중 → **성수기 흥행 전략**

- **관객 패턴**
  - 개봉 초반 폭발적 스크린 점유로 단기간 수익 확보, 그러나 장기 평가는 저조.
- **인사이트**
  - 평점이 낮아도 **규모의 힘(자본+배급+마케팅)**으로 흥행 가능.
  - 단기적 성과는 뛰어나지만, 장기적으로 브랜드/IP 가치 훼손 리스크.

2) 고평점·저흥행 그룹

- **대표작**: 《벌새》, 《윤희에게》, 《어린 의뢰인》 등

🎬 고평점 & 저흥행 그룹 영화 리스트:

	영화제목	평점	흥행수익	장르	개봉월	제작사
1	미스터 주부퀴즈 왕	8.001	19.7억원	코미디	9	폴스타엔터테인먼트, (주)쇼박스, (주)쇼박스, (주)쇼박스, 세방현상(주), (사...
2	사랑해, 말순씨	8.051	20.3억원	드라마	11	블루스톰(주), 쇼이스트(주), 쇼이스트(주), 엠라인디스트리뷰션(주), ㈜라이브톤
3	가족의 탄생	8.591	11.5억원	드라마	5	블루스톰(주), 롯데쇼핑㈜롯데엔터테인먼트, 롯데쇼핑㈜롯데엔터테인먼트, 블루스톰(주)...
4	국경의 남쪽	8.491	14.2억원	드라마	5	(주)싸이더스, 씨제이엔터테인먼트, 씨제이엔터테인먼트, ㈜라이브톤, 창고사람들, 세...
5	파이스토리	8.691	12.9억원	애니메이션, 가족	7	(주)에펙스디지탈, 원더월드 LLC, ㈜디지아트프로덕션, 씨제이엔터테인먼트, 소빅창...
...	...	...	...	...	...	...

- **주요 제작사 분포**

🏢 고평점 & 저흥행 영화의 주요 제작사:

	제작사	빈도수	비율(%)
1	인디스토리	15	20.27
2	영화사 진진	12	16.22
3	엣나인필름	9	12.16
4	리틀빅픽처스	7	9.46
5	시네마달	6	8.11
6	영화사 그램	5	6.76
7	아트하우스 모모	4	5.41
8	콘텐츠판다	3	4.05
9	무브먼트	3	4.05
10	영화사 오드	2	2.70

- 인디·중소 제작사 중심, **대기업 제작 거의 없음** → 상영관 확보/배급력 취약

- **주요 장르 비율**

🎭 고평점 & 저흥행 영화의 주요 장르:

	장르	빈도수	비율(%)
1	드라마	46	62.16
2	멜로/로맨스	12	16.22
3	다큐멘터리	8	10.81
4	가족	5	6.76
5	예술/실험	5	6.76
6	스릴러	4	5.41
7	코미디	3	4.05
8	애니메이션	2	2.70
9	음악	1	1.35
10	판타지	1	1.35

- 드라마·예술성 강한 장르 중심 → 작품성은 높지만 흥행성 한계
- **개봉월 분포**

📅 고평점 & 저흥행 영화의 개봉월 분포:

	개봉월	빈도수	비율(%)
1	1	2	2.70
2	2	4	5.41
3	3	6	8.11
4	4	7	9.46
5	5	5	6.76
6	6	4	5.41
7	7	3	4.05
8	8	2	2.70
9	9	6	8.11
10	10	8	10.81
11	11	14	18.92
12	12	11	14.86

- 봄·가을·겨울 등 **비성수기 개봉**이 다수 → 상업 영화와의 경쟁에서 불리
- **관객 패턴**
  - 관객 수는 적지만 **관람객 만족도·충성도 높음**
  - 해외 영화제 수상·비평가 찬사 다수 → 작품성 인정
  - 극장에서는 흥행 약세, OTT/해외 플랫폼에서는 재발견 가능성
- **인사이트**
  - *품질(작품성)**은 높으나 상업적 성과는 제한적
  - 극장 시장 구조상 불리, 그러나 OTT 확산은 **롱테일 흥행 기회** 제공

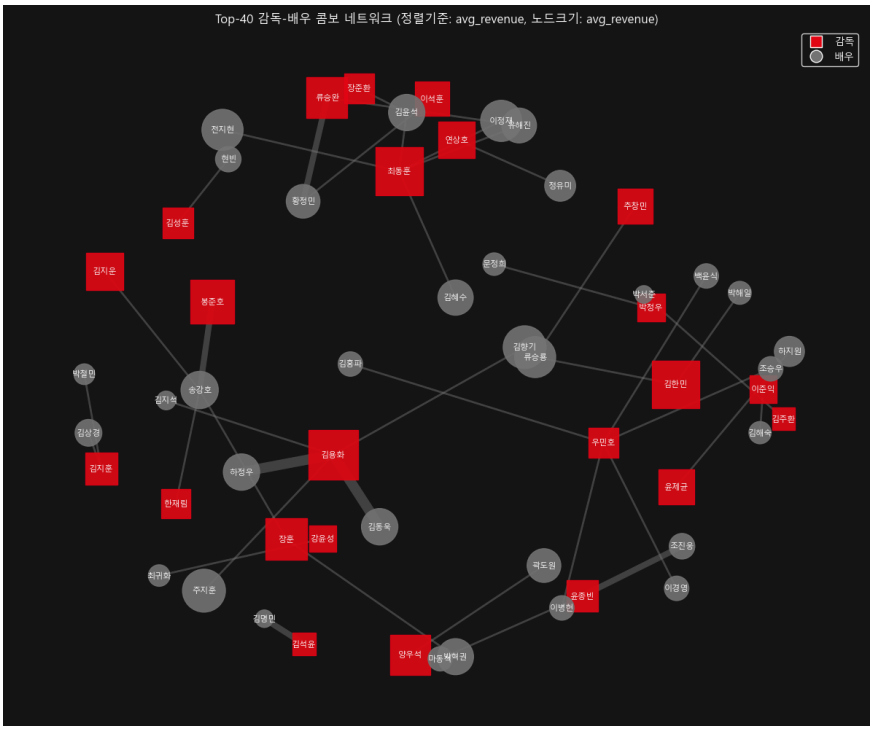
그룹 비교 종합 인사이트

- 저평점·고흥행 = **규모의 힘** (자본·배급·마케팅).
- 고평점·저흥행 = **품질의 힘** (작품성·비평적 인정).
- 한국 극장 시장에서는 규모가 작품성보다 더 직접적 영향, 하지만 OTT 시대에는 작품성 있는 영화가 재조명될 가능성.

과제 3️⃣ 배우/감독 네트워크 분석

데이터 및 분석

- 총 927편, 협업 관계 4,552건 → 유효 협업 178건.
- 상위 40개 콤보를 기준으로 네트워크 시각화. (노드 수 56개, 엣지 수 40개)



감독	배우	협업횟수	평균수익
1	김용화	김향기	2
2	김용화	주지훈	2
3	최동훈	이정재	2
4	최동훈	전지현	2
5	김한민	류승룡	2
6	봉준호	송강호	3
7	김용화	김동욱	4
8	김용화	하정우	4
9	최동훈	김윤석	2
10	장훈	박혁권	2

주요 콤보

- 김용화–하정우/주지훈/김향기 → 평균 수익 1,091억
- 최동훈–이정재/전지현 → 959억
- 봉준호–송강호 → 736억

성공 패턴

- 반복 협업 → 관객에게 신뢰(브랜드 효과).
- 네트워크 중심성이 높은 감독 → 장르 확장과 안정적 흥행.
- 브리지 역할의 배우 → 새로운 관객층 유입.

인사이트 (과제 3)

- 흥행은 개별 스타보다 **네트워크 구조**에서 비롯.
- 반복 협업은 안정적이지만, **편중 현상**으로 신인 진입 장벽 발생.
- 캐스팅 전략은 **황금 조합 활용 + 신인 발굴**이 병행되어야 함.

---

4. 결론 및 회고

4.1 종합 인사이트

1. **사극**은 한국 영화 산업의 정체성과 상업성이 결합된 장르로 반복적 흥행 성공.
2. **개봉 타이밍·런타임 최적화**가 장르별 흥행의 중요한 변수.
3. **평점과 흥행의 괴리**: 네티즌 평점은 흥행과 부분 연결되지만, 비평가 평점은 작품성 지표로 기능.
4. **네트워크 효과**: 감독–배우 협업 네트워크가 흥행을 구조적으로 강화.
5. **데이터 공백 리스크**: 예산 부재로 ROI 분석 불가, 산업 투명성 부족은 구조적 한계.

4.2 회고

이번 프로젝트에서 가장 크게 배운 점은 **데이터 수집의 중요성**이었다. TMDB→IMDB 실패, KOFIC·네이버·수동 입력을 거치면서, **분석의 성패는 데이터 정의와 확보에서 갈린다**는 사실을 뼈저리게 느꼈다.

데이터가 불완전했기 때문에, 분석가는 단순히 코딩하는 사람이 아니라 **현실의 제약을 이해하고 극복하는 사람**이라는 것도 배웠다.

4.3 향후 계획

- **ROI 추정 모델**: 대체 변수 활용.
- **OTT 데이터 결합**: 극장+OTT 통합 성과 분석.
- **네트워크 기반 추천**: 감독–배우–장르 조합의 성공 확률 예측.