

OTT DATA ANALYSIS

EDA factory

이호원, 김도균, 오준엽



OTT DATA ANALYSIS

1. 배경 설명

2. 분석 설계

3. 탐색 질문

4. 데이터 전처리

- 데이터 셋 설명
- 데이터 전처리

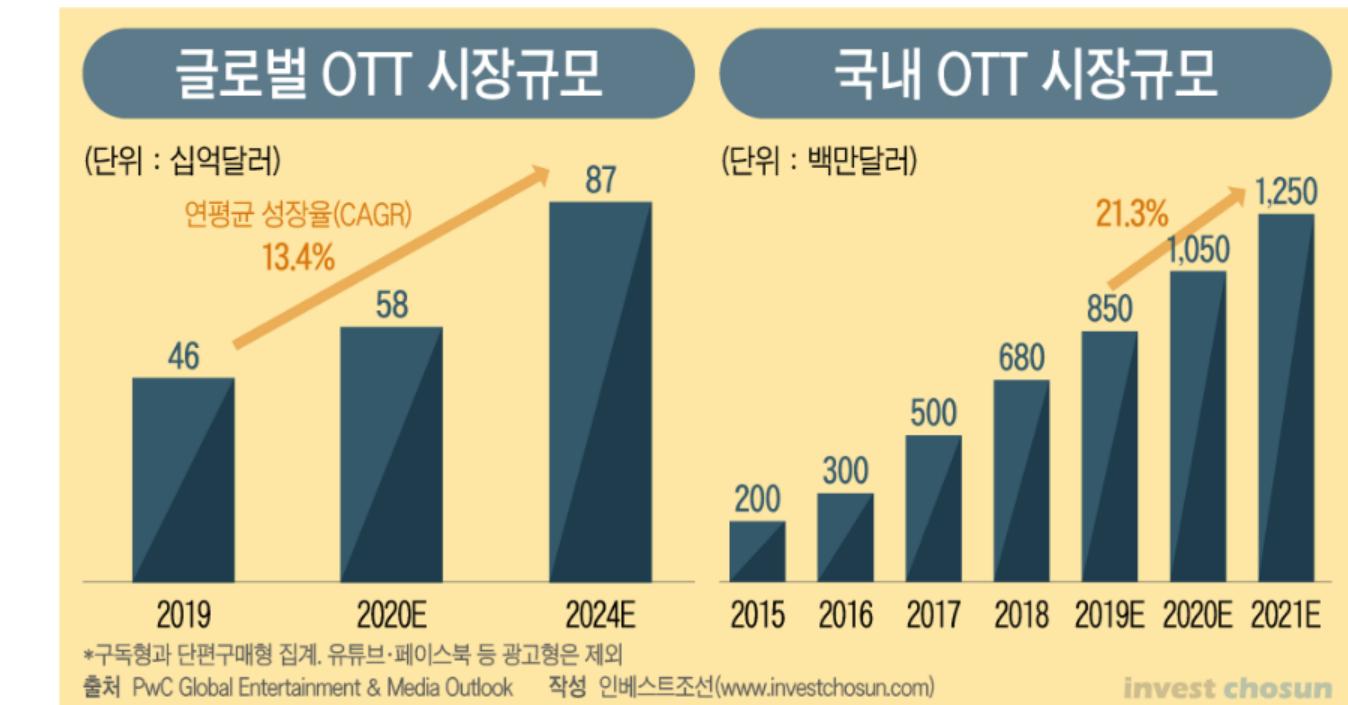
5. 데이터 분석

6. 결론

7. 심화과제

OTT (over the top media -service)

인터넷을 통해 방송 프로그램·영화·교육 등 각종 미디어 콘텐츠를 제공하는 서비스.
over-the-x는 '기존의 범위를 넘어서'라는 뜻이며, top은 TV 셋톱박스를 의미



플랫폼 규모 ↑



스코어

장르

출시일

데이터

콘텐츠 유형

분석 설계



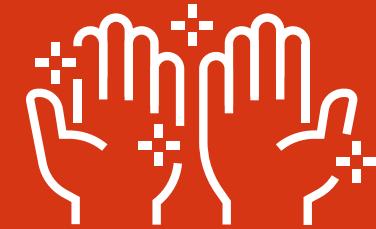
분석 목적

1. OTT 데이터의 특성
2. 변수별 상관관계 분석
3. 플랫폼 특징 파악



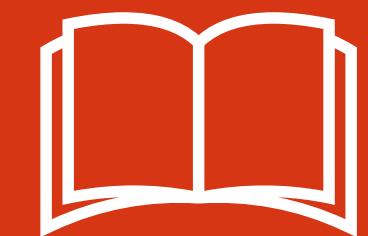
분석 방법

- OTT 데이터 특성 살펴보기
4대 플랫폼 비교 분석



탐색 질문

여러 변수들에 대해
다양한 질문을 제시



문제점

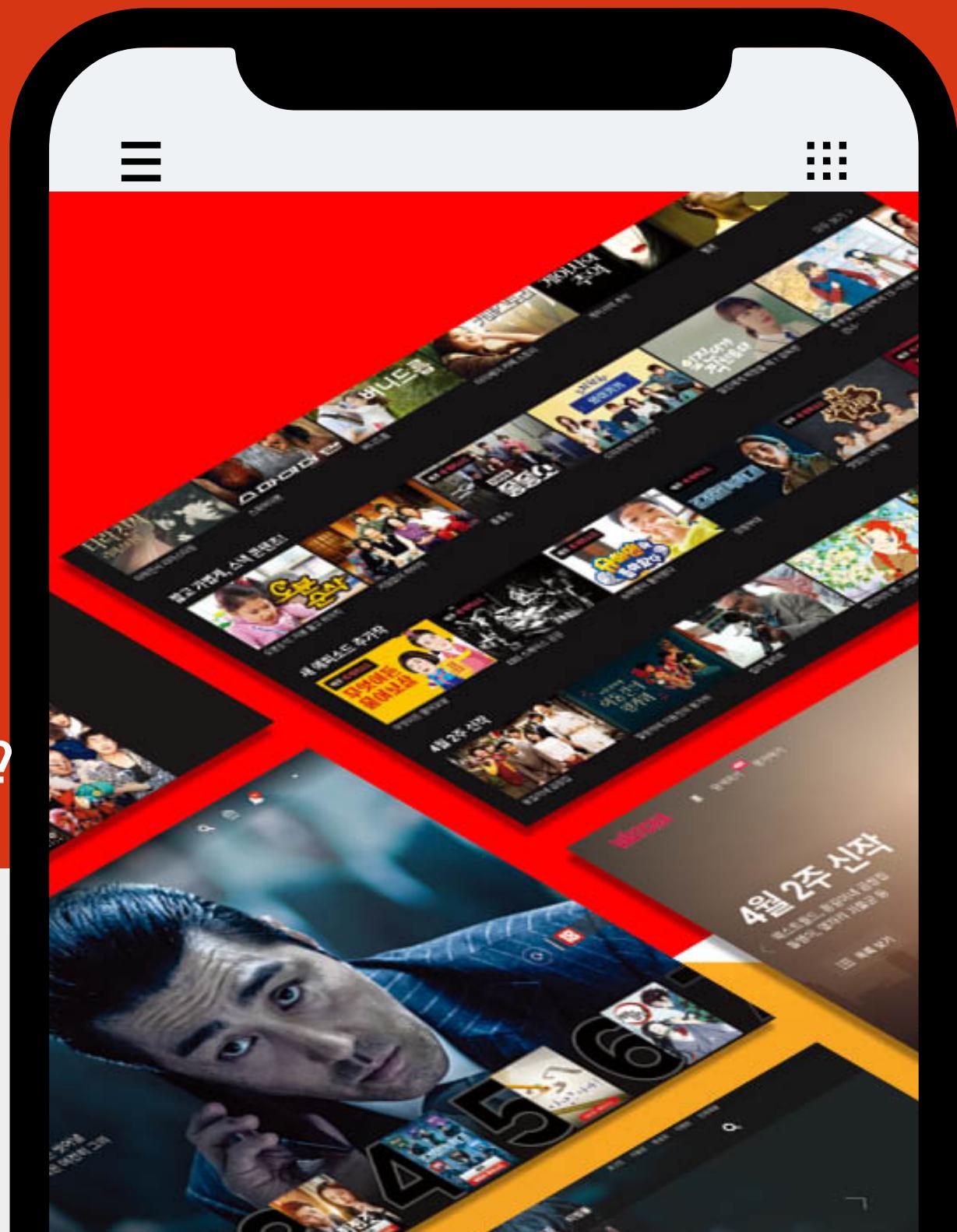
데이터 보전을 통해
결측치를 최소화

탐색 질문

어떤 콘텐츠가 많을까?

score에 영향을 미치는 요소는?

높은 score 작품을 많이 보유하고 있는 플랫폼은?



어떤 키워드가 많을까?

높은 score를 기록한 영화들의 특징

가장 많이 등장하는 배우와 감독은?

데이터셋 설명

Netflix, Amazon Prime, Disney+, Paramount 총 4개의 플랫폼

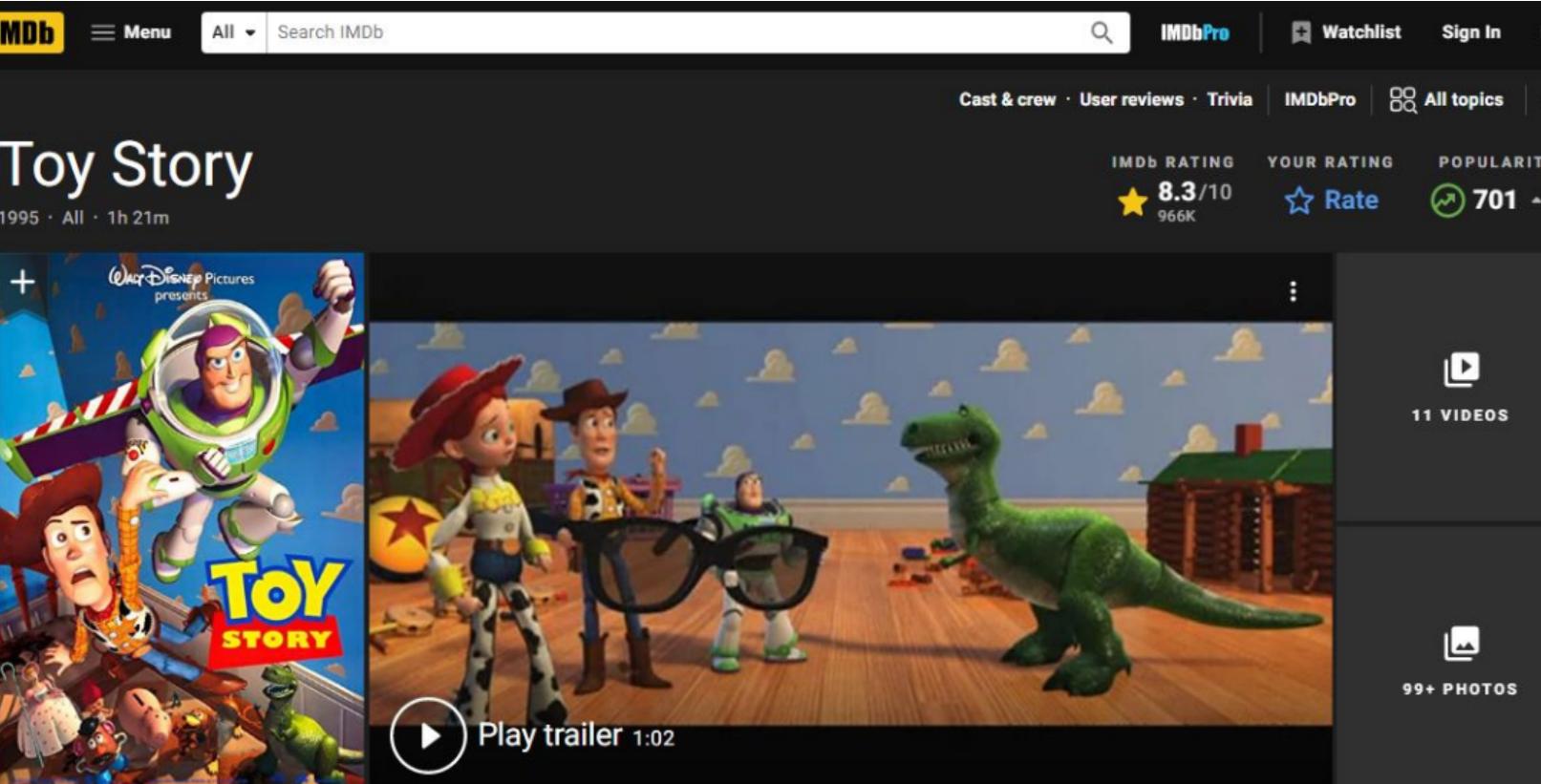
- **titles.csv** 15개의 features (id, title, score...)
- **credits.csv** 5개의 features (id, name, role...)

→ 데이터를 최대한 **보전해보자!**

데이터셋 설명

imdb_score, imdb_vote, tmdb_score, tmdb_popularity

IMDb



Toy Story

1995 · All · 1h 21m

Walt Disney Pictures presents

IMDb RATING 8.3/10 YOUR RATING Rate POPULARITY 701 ▲ 38

Cast & crew · User reviews · Trivia · IMDbPro · All topics

11 VIDEOS

99+ PHOTOS

Play trailer 1:02

Animation · Adventure · Comedy

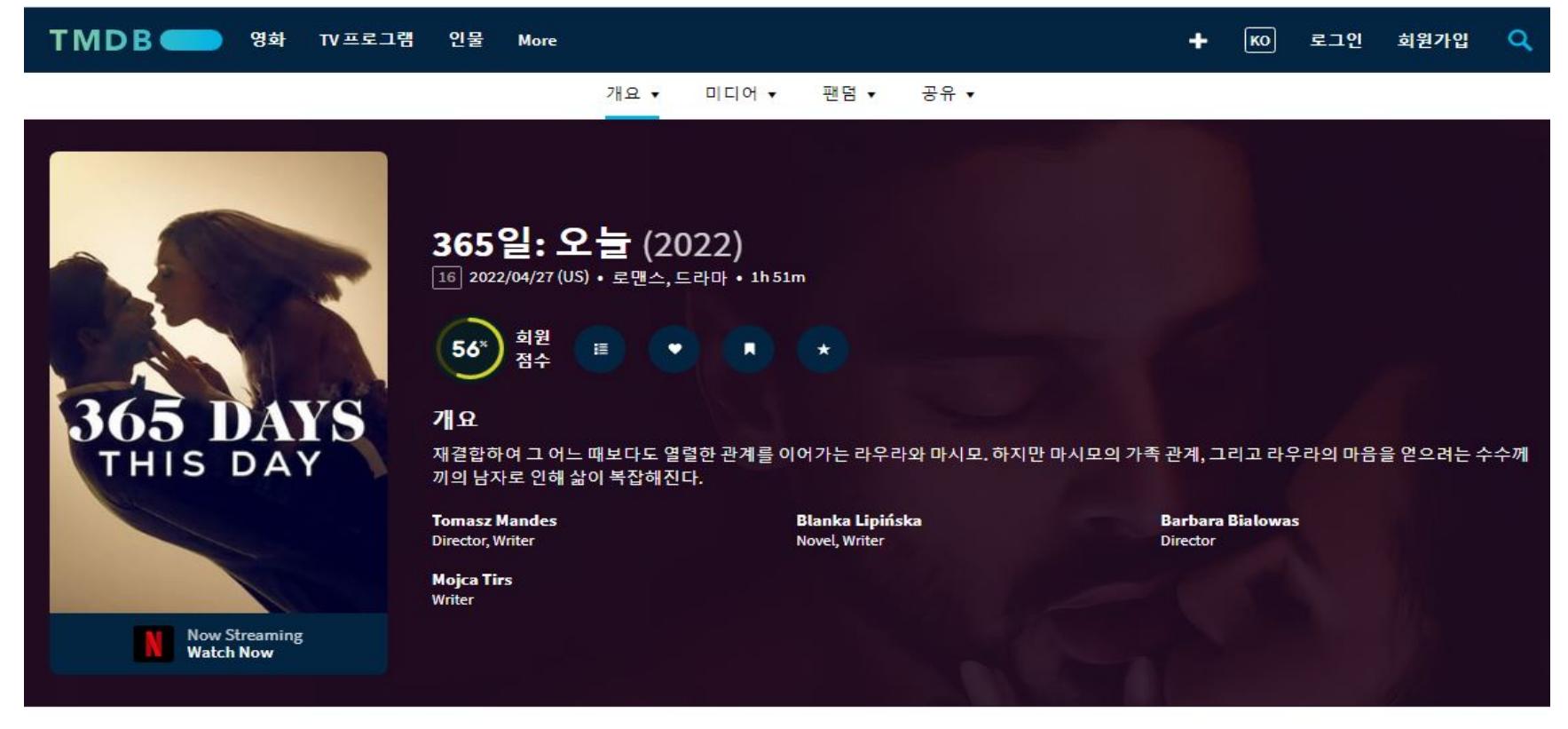
A cowboy doll is profoundly threatened and jealous when a new spaceman figure supplants him as top toy in a boy's room.

Director John Lasseter

735 User reviews 159 Critic reviews 95 Metascore

Add to Watchlist

TMDB



365일: 오늘 (2022)

16 2022/04/27 (US) · 로맨스, 드라마 · 1h 51m

56 회원 점수

개요

재결합하여 그 어느 때보다도 열렬한 관계를 이어가는 라우라와 마시모. 하지만 마시모의 가족 관계, 그리고 라우라의 마음을 염으려는 수수께끼의 남자로 인해 삶이 복잡해진다.

Tomasz Mandes
Director, Writer

Blanka Lipińska
Novel, Writer

Barbara Bialowas
Director

Mojca Tirs
Writer

주요 출연진

원제
365 Days: This Day

상태
개통됨

원어

데이터 전처리

None age 설정

불필요 컬럼 drop

Platform 컬럼 추가
NF, DS, AM, PM

Movie Season 0삽입

imdb, tmdb 연관성 평균 0.9의 보정계수
score 들 간 보정계수 삽입

중복 결측인 경우 삭제
votes, popularity

5806

score 이상치 제거

리스트 언패킹

id title type description release_year age_certification runtime genres production_countries seasons imdb_id imdb_score tmdb_votes tmdb_popularity tmdb_score Platform



데이터 전처리

결과

넷플릭스: 5806 → 5062

아마존: 9871 → 8189

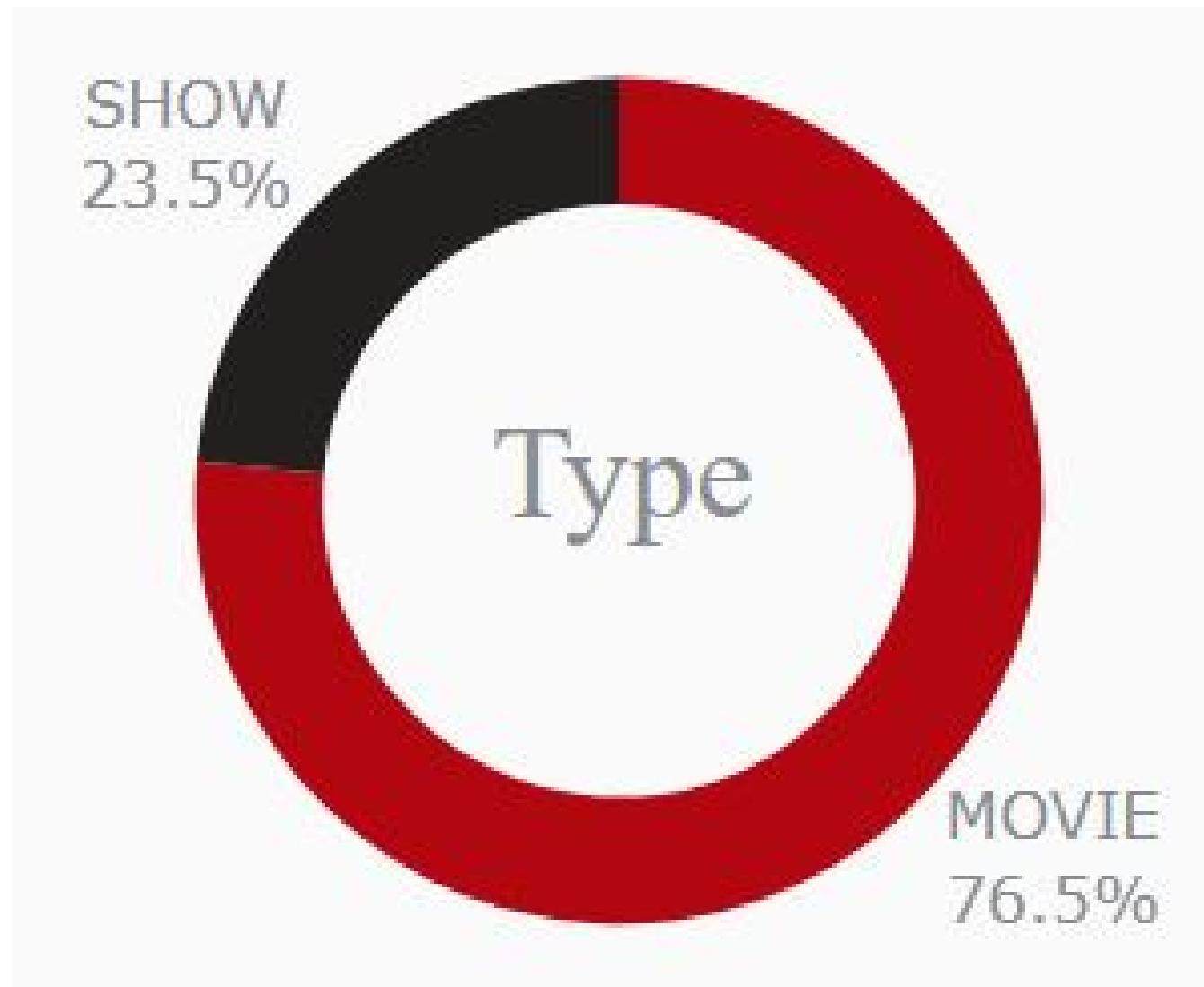
디즈니: 1535 → 1085

파라마운트: 2825 → 2544

총 16880 → 14983 로 1,897개의 데이터만 제거하며 **데이터 소실 최소화**

데이터 분석

어떤 콘텐츠가 많을까?

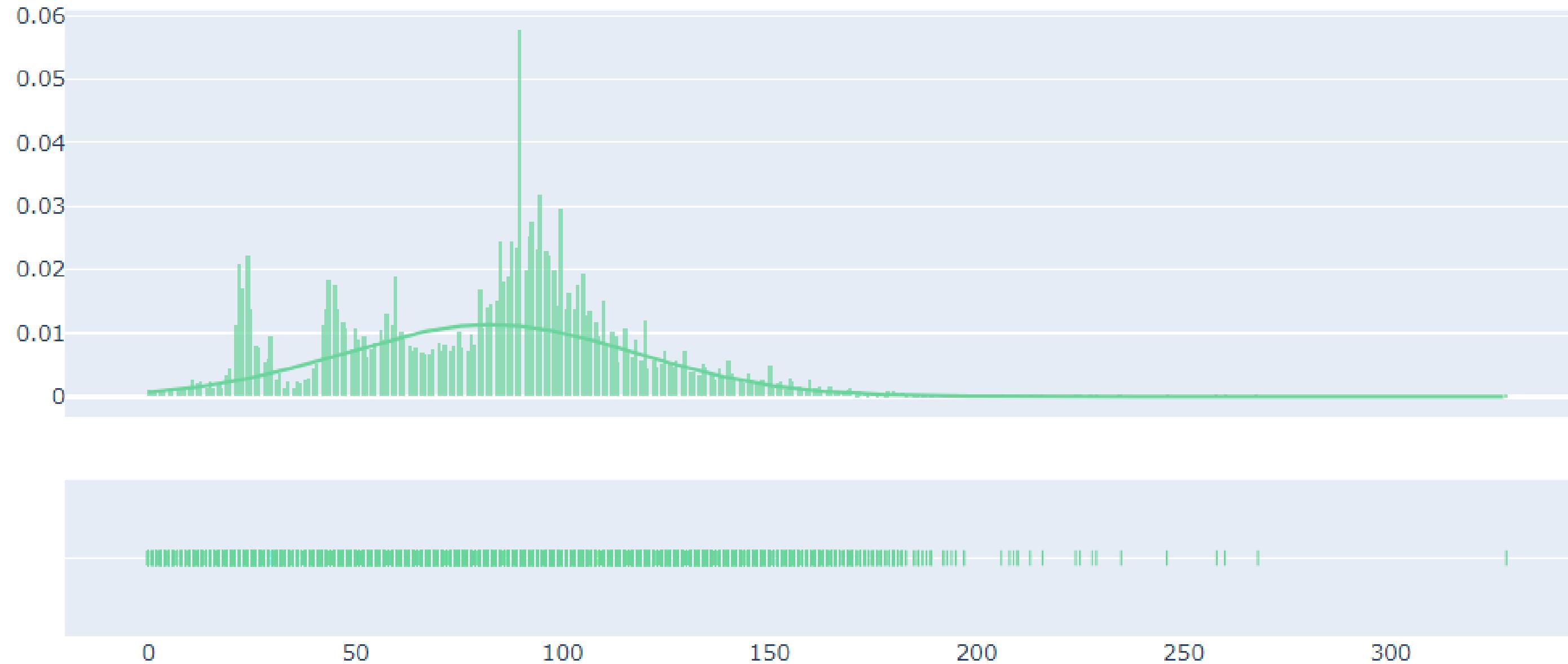


영화의 비중이 생각보다 더 높다

데이터 분석

어떤 콘텐츠가 많을까?

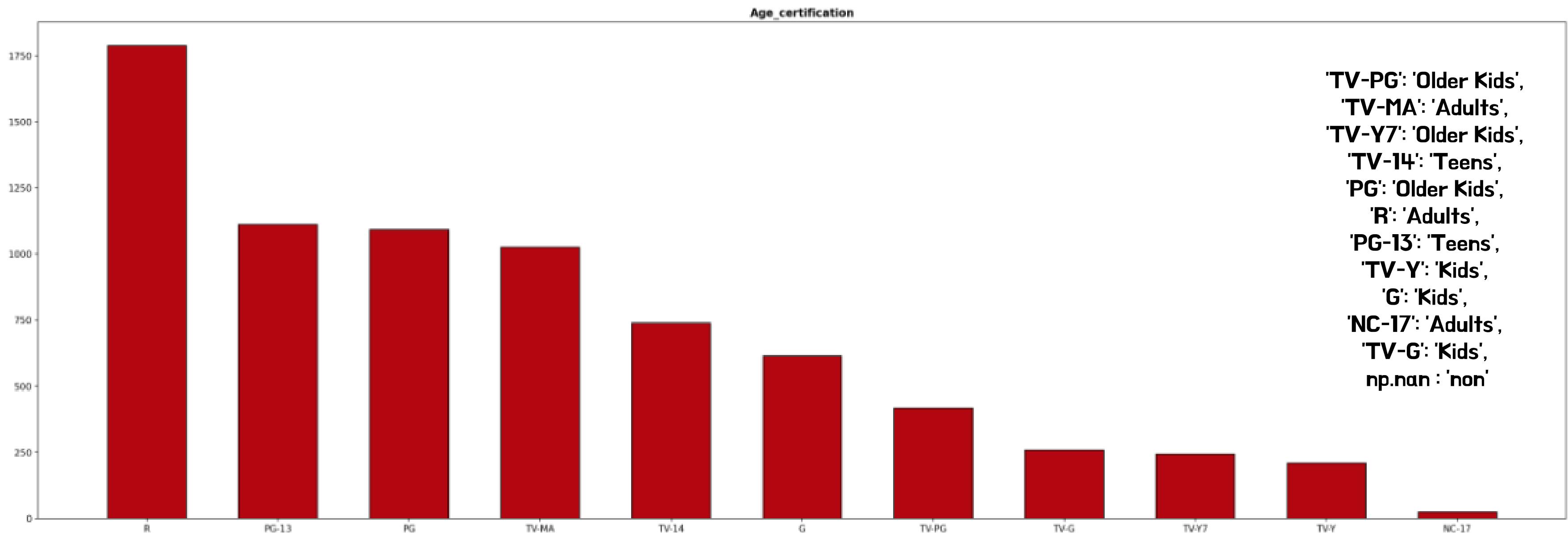
Distplot with Normal Distribution



100분 내외의 콘텐츠가 많음을 알 수 있다

데이터 분석

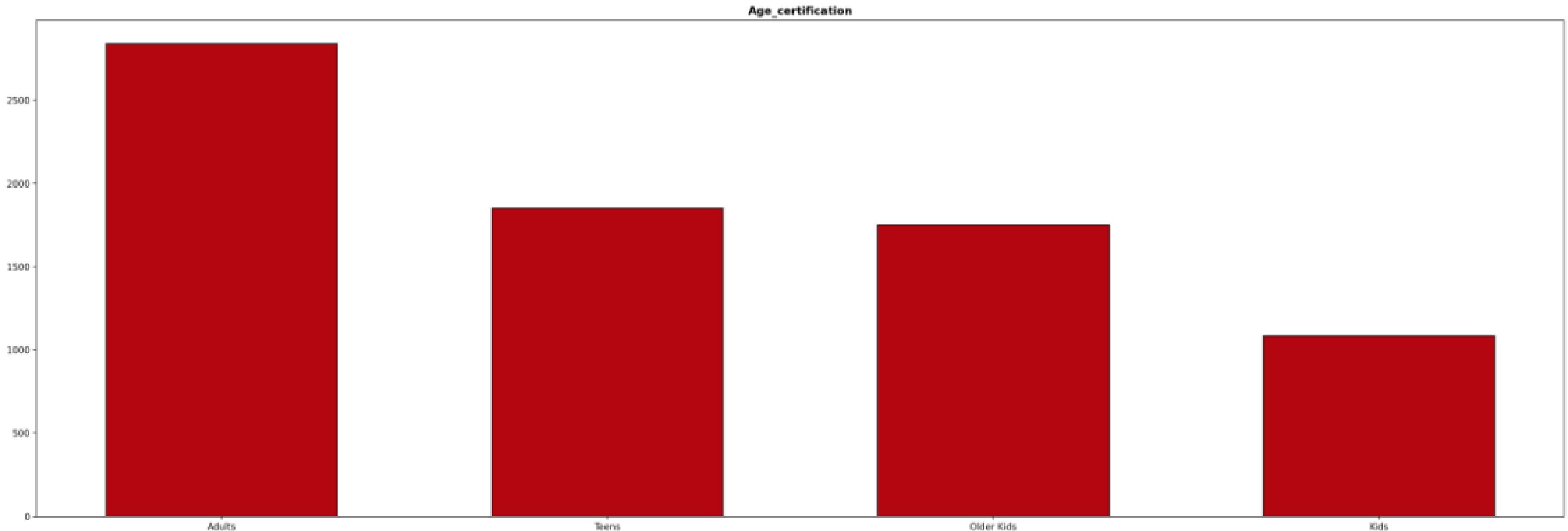
어떤 콘텐츠가 많을까?



자료 그대로 확인하기 어려움 → age 구간 나누기 실행

데이터 분석

어떤 콘텐츠가 많을까?

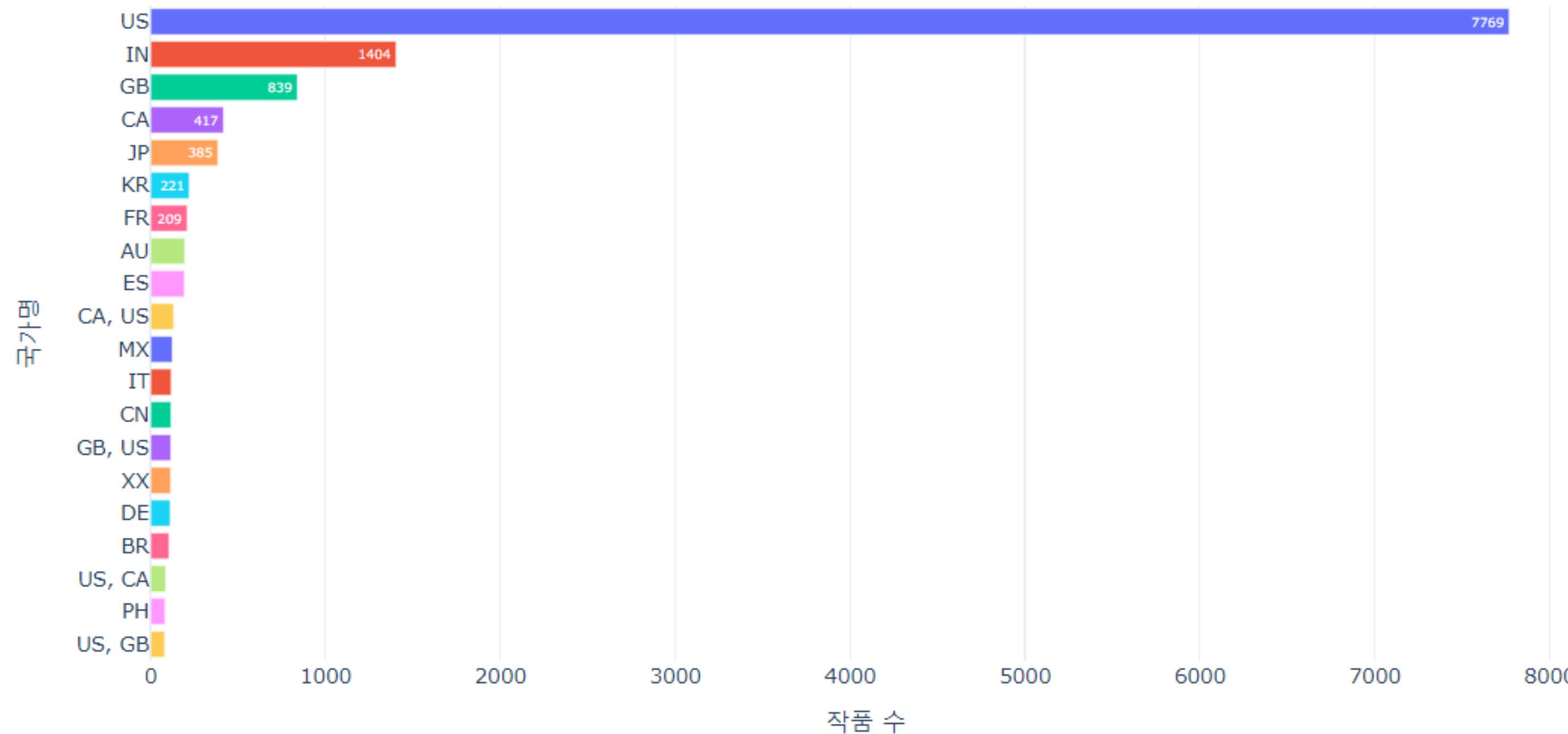


Adults > Teens > Older Kids > Kids

데이터 분석

어떤 콘텐츠가 많을까?

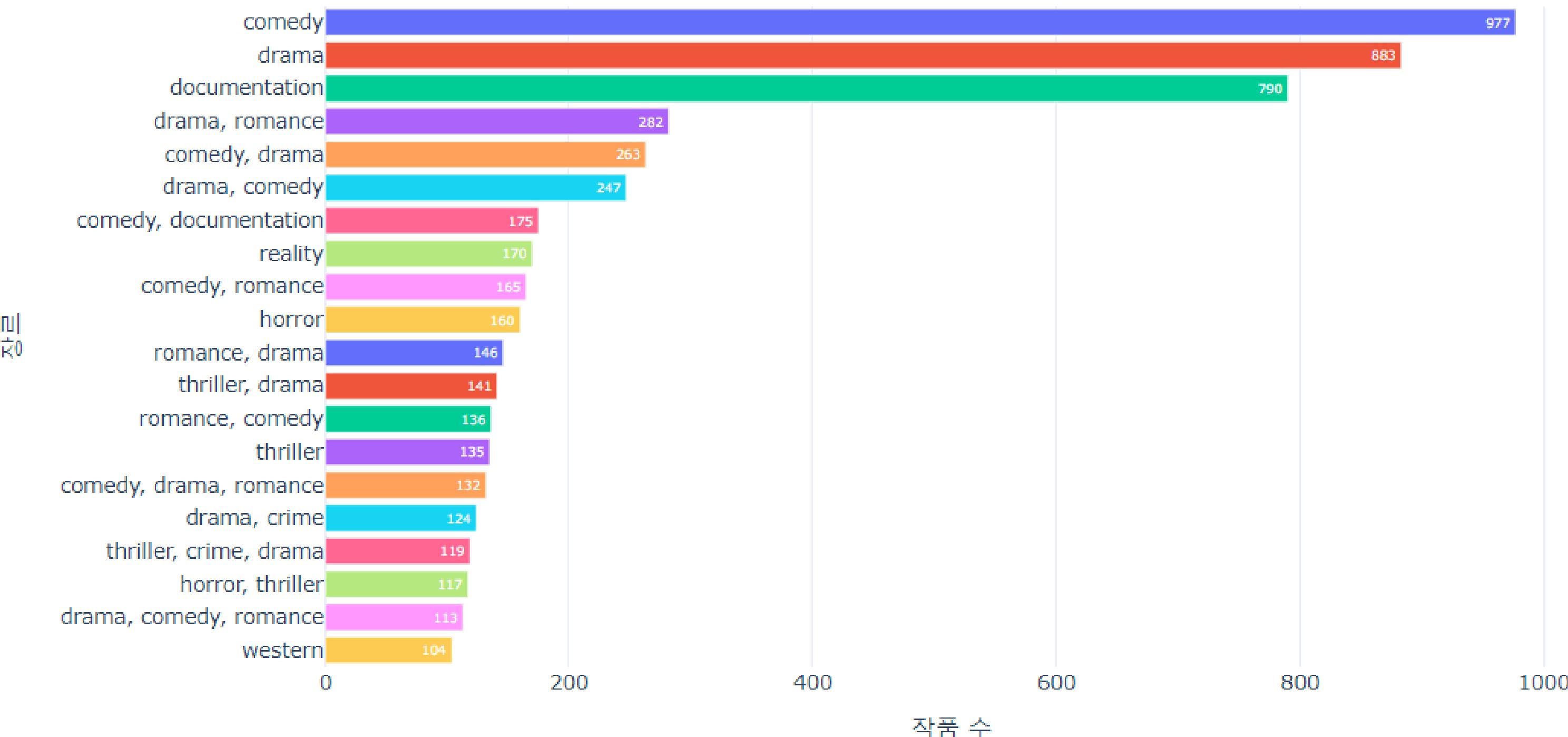
제작 국가별 작품 수 Top20



데이터 분석

어떤 콘텐츠가 많을까?

장르별 작품 수 Top20



데이터 분석

어떤 키워드가 많을까?



Title

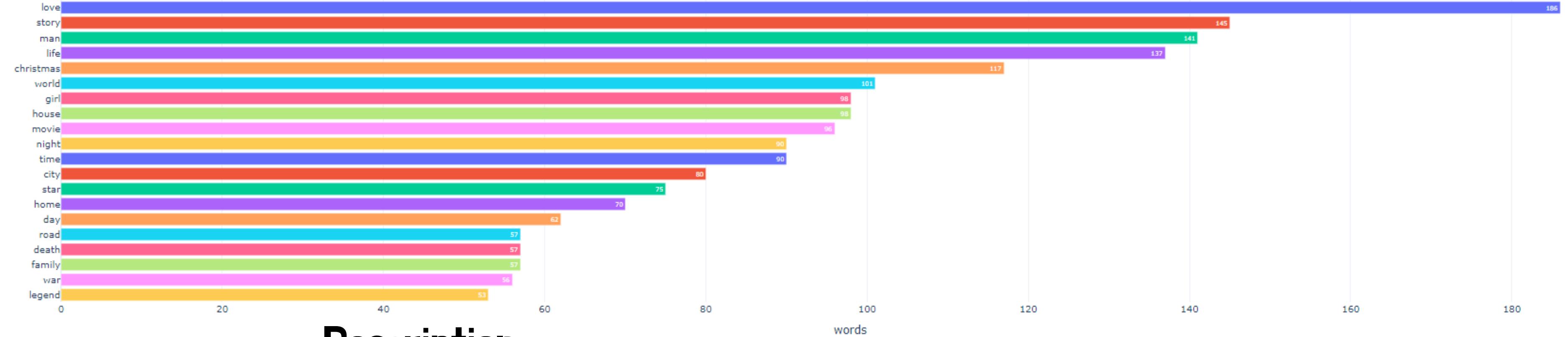


Description

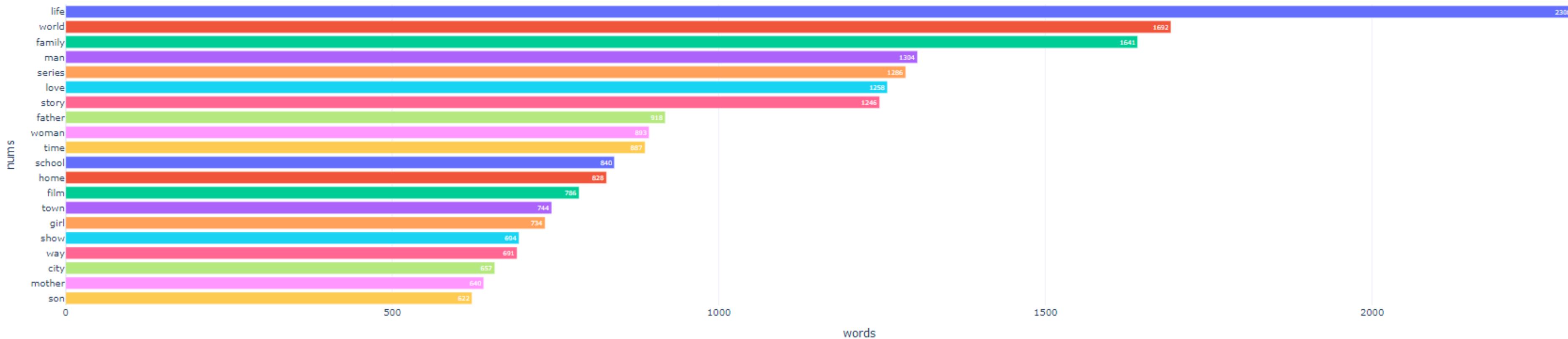
데이터 분석

어떤 키워드가 많을까?

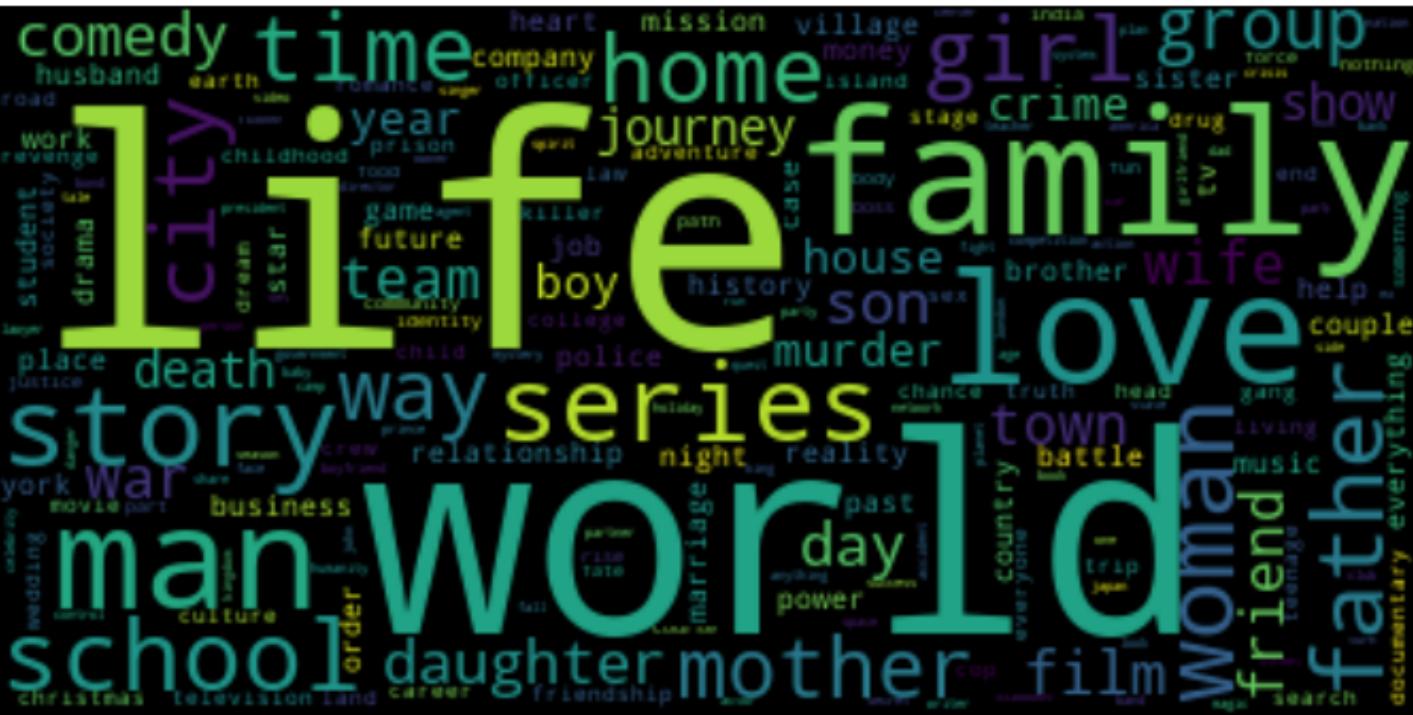
NN_title words Top20 **Title**



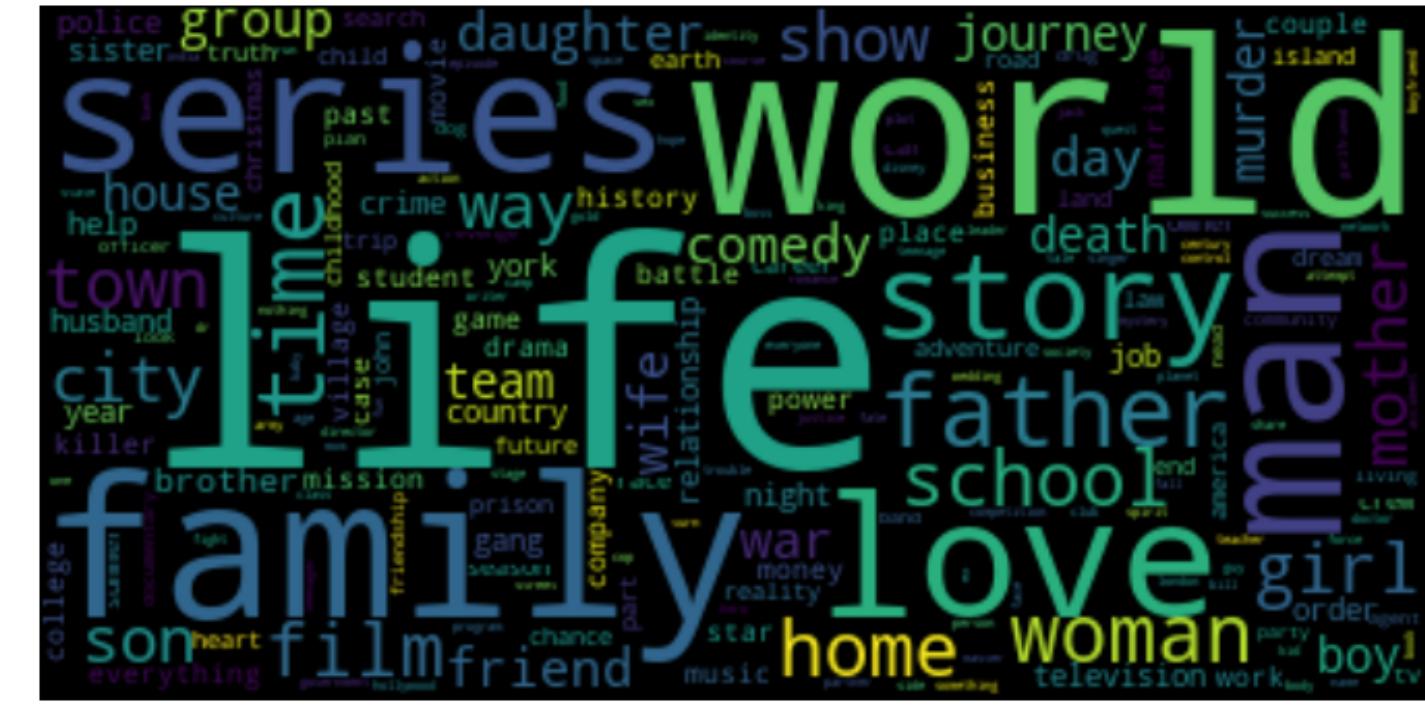
NN_description words Top20 **Description**



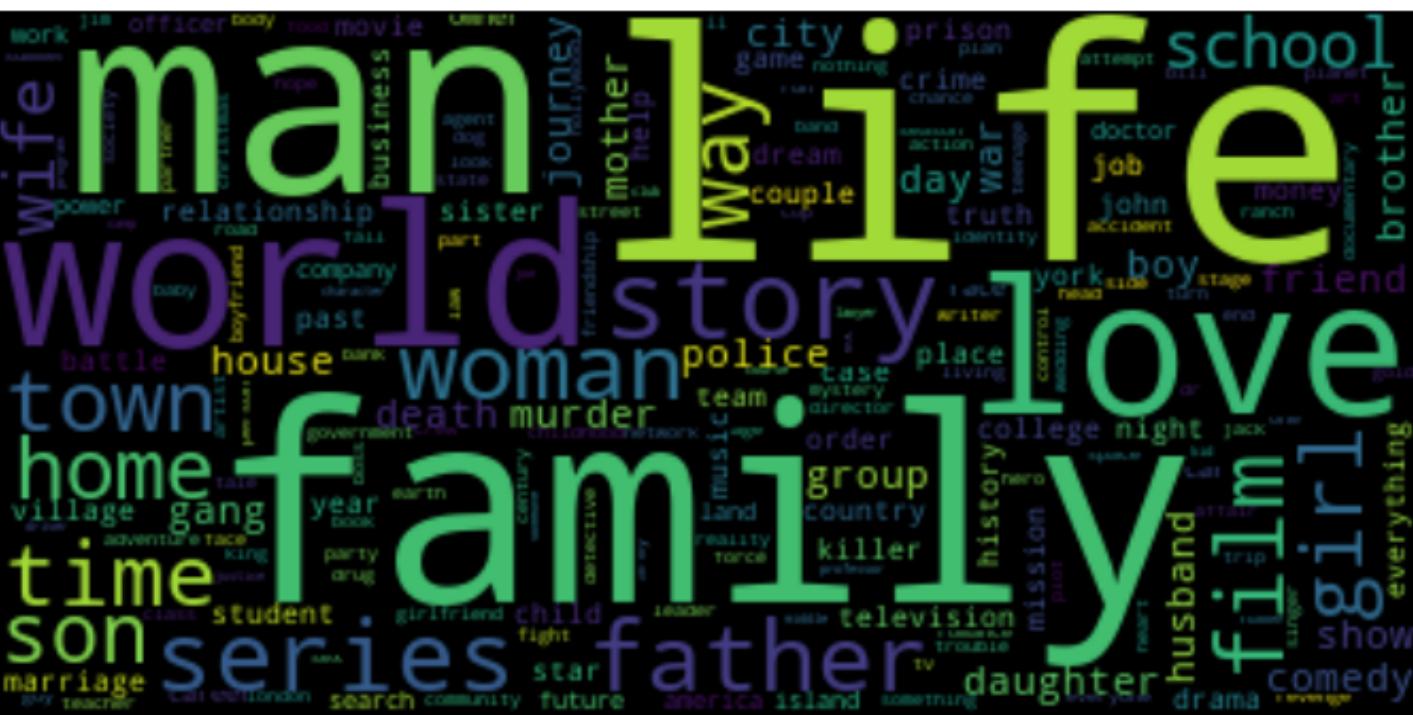
데이터 분석 어떤 키워드가 많을까?



Netflix



Disney+



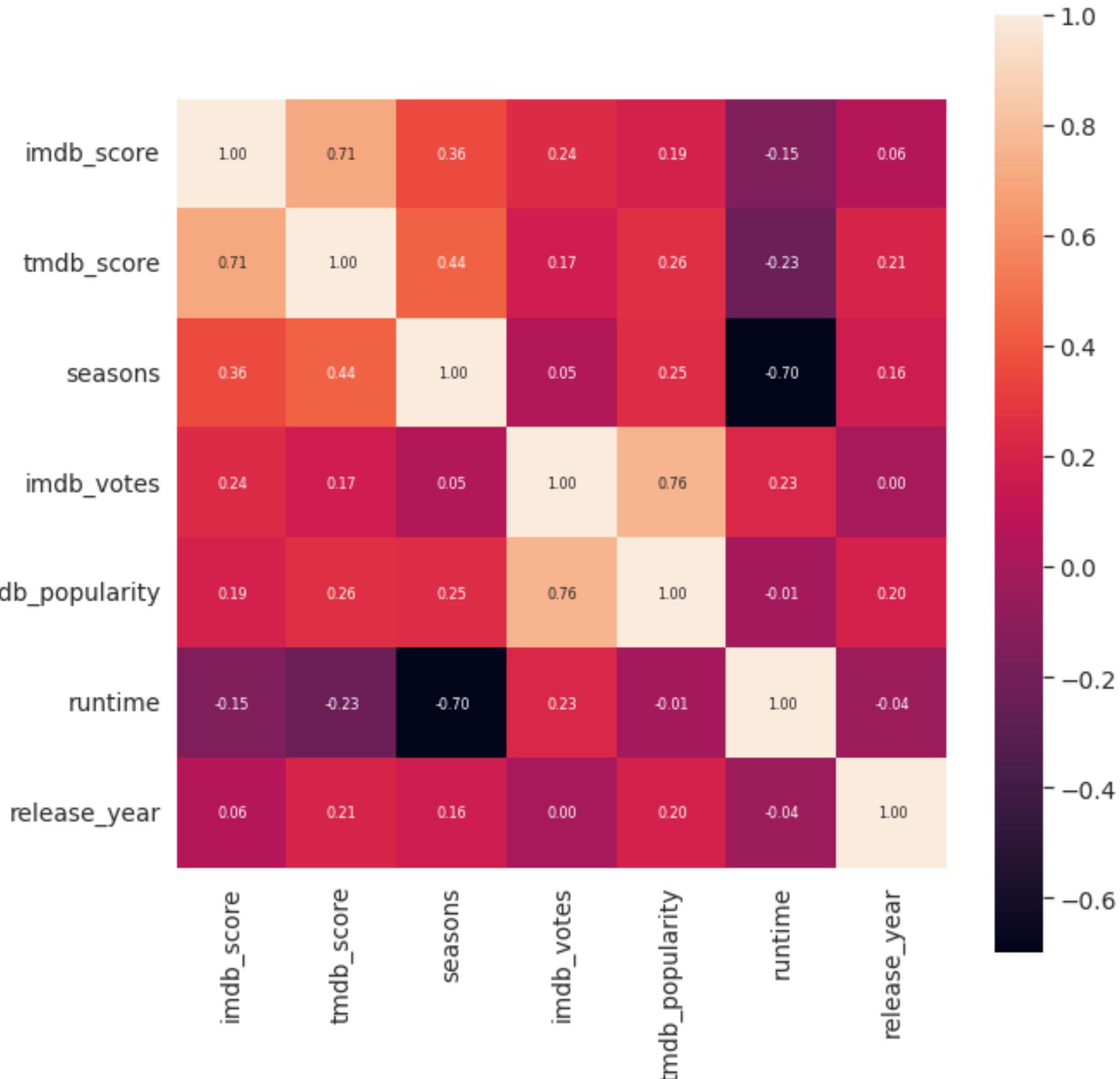
Amazon Prime



Paramount

데이터 분석

score에 영향을 미치는 요소는?

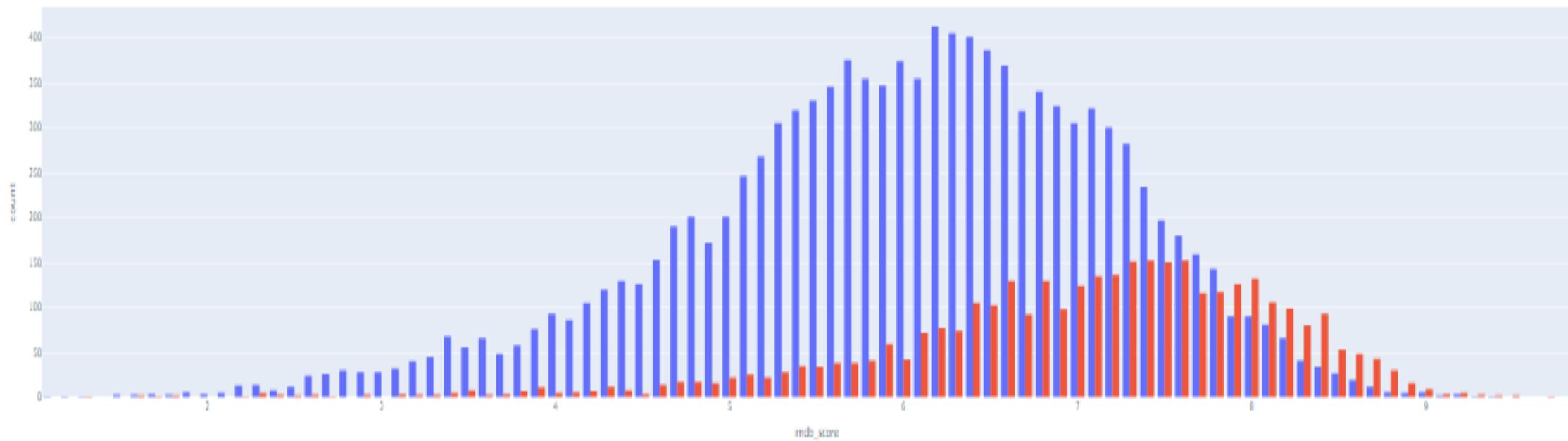


**Spearmen 상관관계를 활용한
Heatmap 분석을 통해
상관도가 높다고 판단되는 항목에 대해 분석 진행**

**tmdb_score, imdb_votes, tmdb_popularity
runtime, release_year**

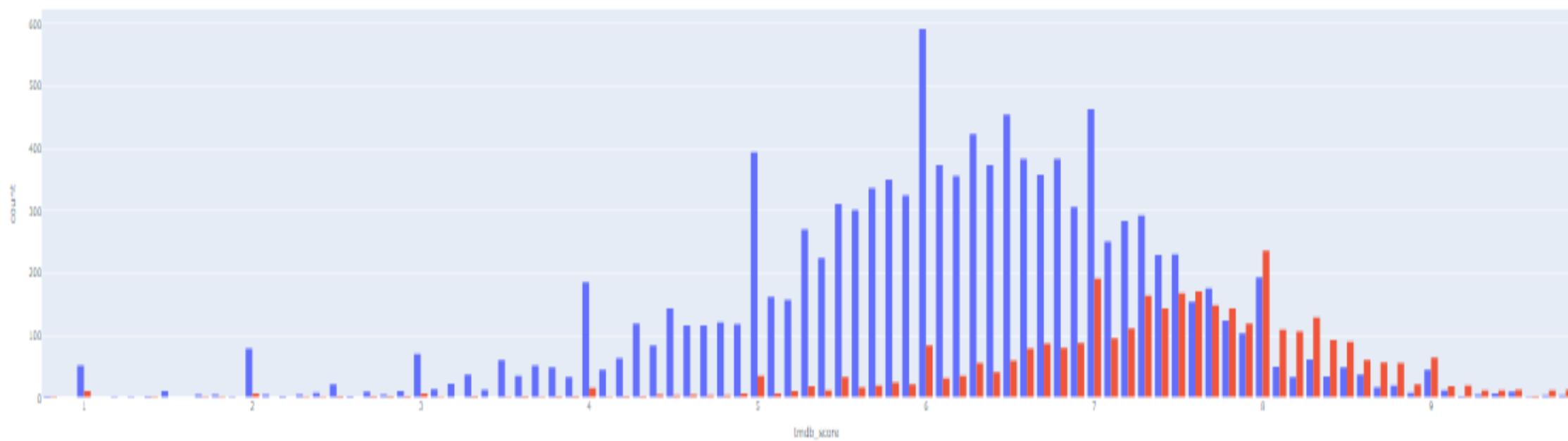
데이터 분석

score에 영향을 미치는 요소는?



영화인가, TV물인가에 따라
imdb_score에 영향을 미칠까?

연속형 변수와 달리 이분형 데이터이기
때문에 label encoding을 통해
'점이연 상관계수'
(point biserial correlation coefficient)
활용



데이터 분석

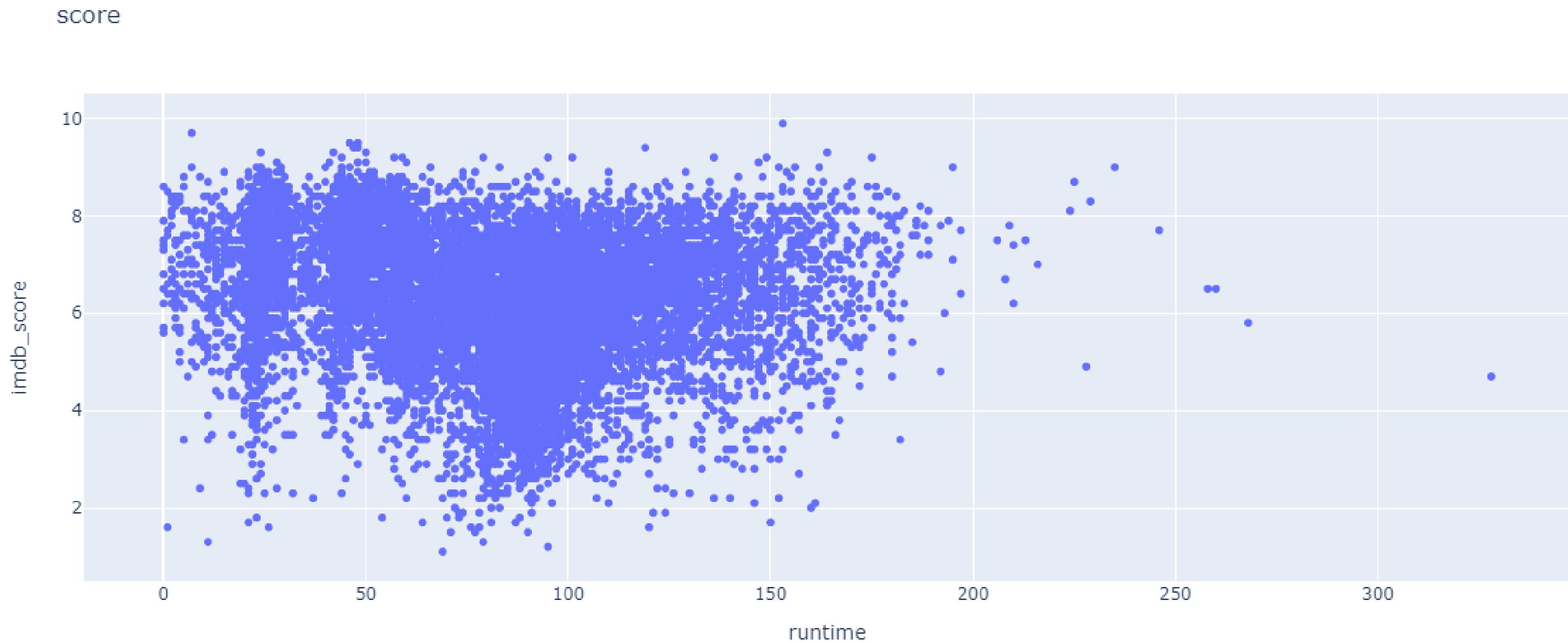
score에 영향을 미치는 요소는?



→ type은 imdb_score에
영향을 미치는 유의미한 변수가 아님을 확인

데이터 분석

score에 영향을 미치는 요소는?

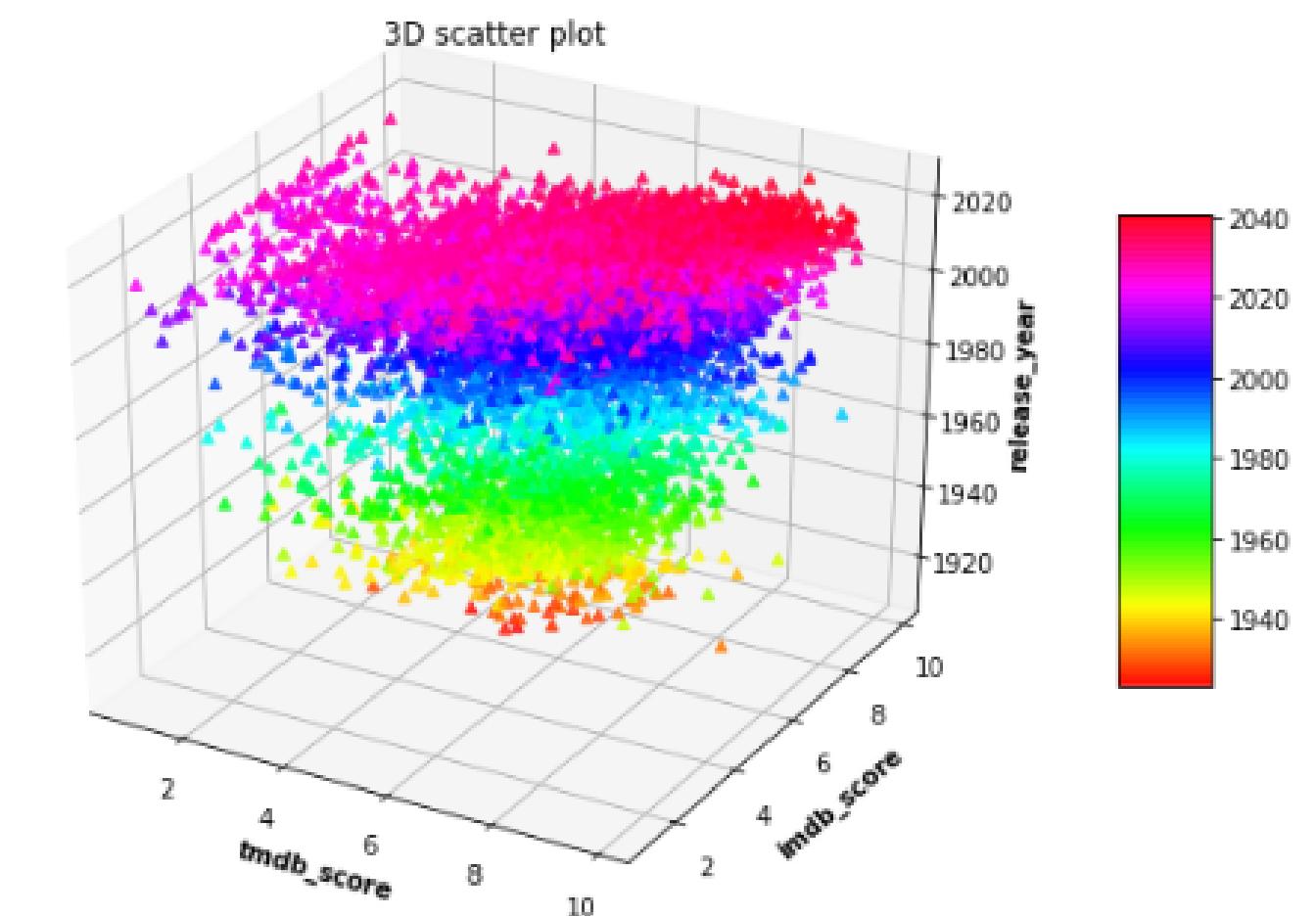
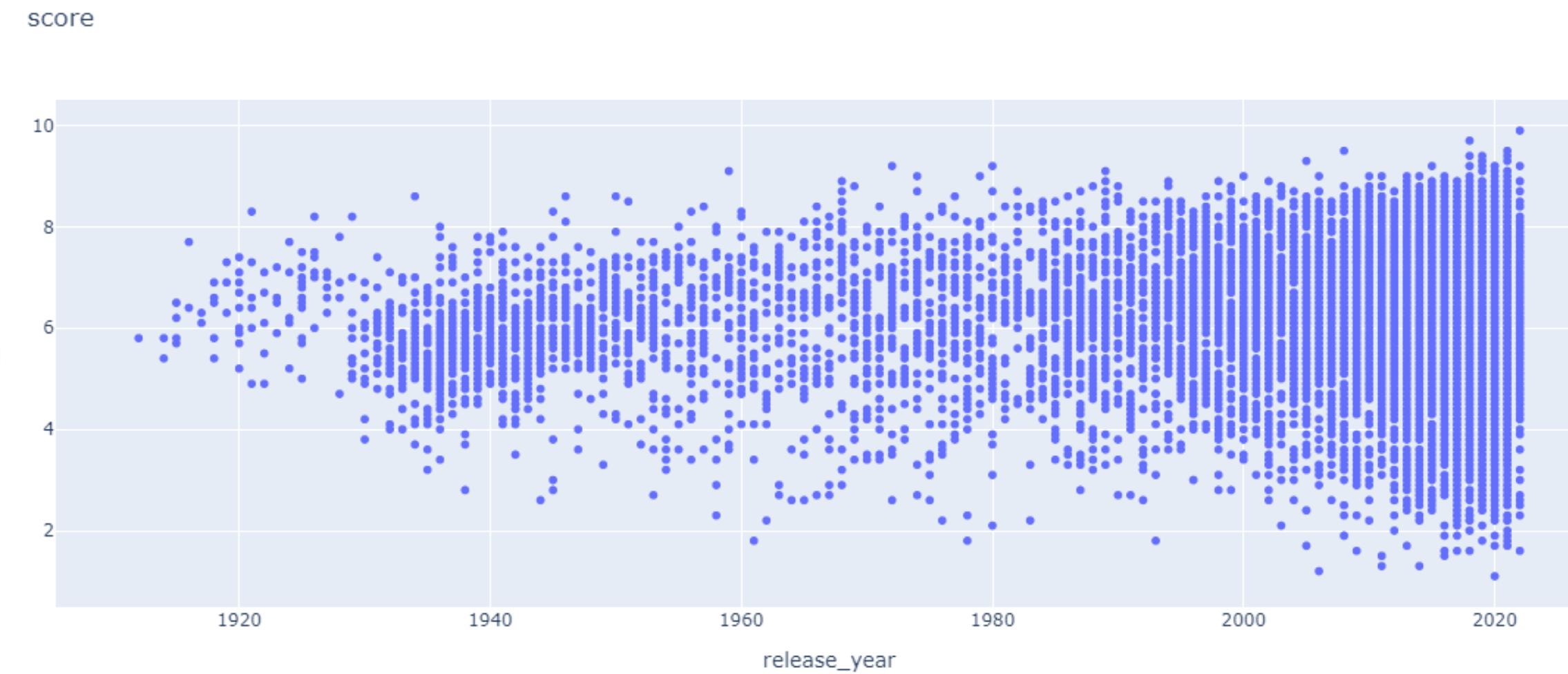


Runtime이 길거나 짧을 수록 imdb_score가 높을 것이다 **기각**

→ 상영시간은 score에 영향을 주지 않는다

데이터 분석

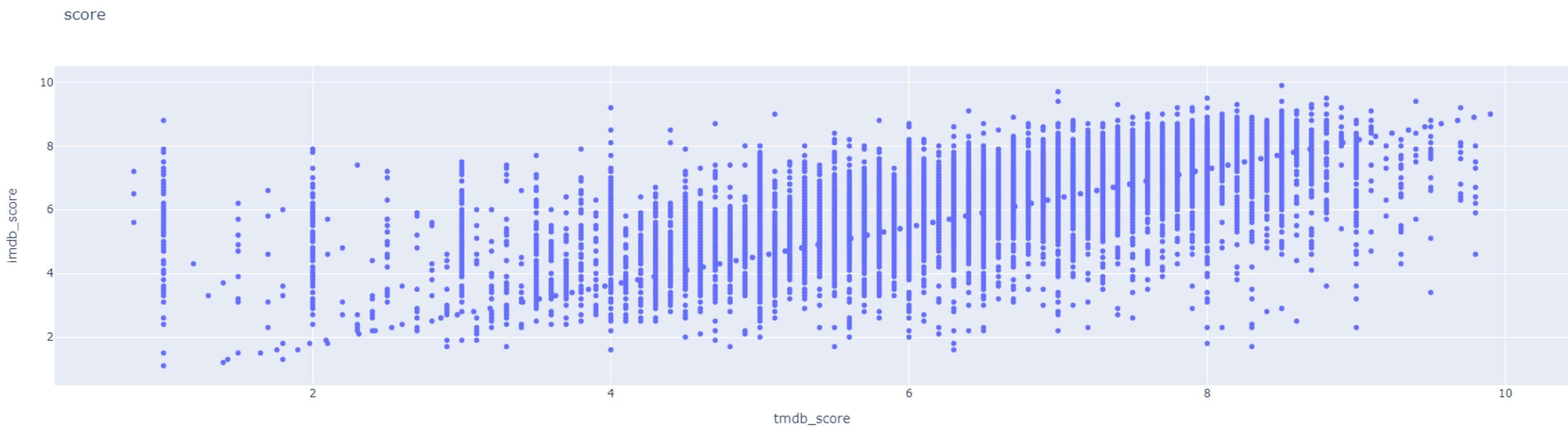
score에 영향을 미치는 요소는?



최신영화일수록 imdb_score가 높을 것이다 **기각**
→ 출시일과 인기도는 연관성이 없다

데이터 분석

score에 영향을 미치는 요소는?



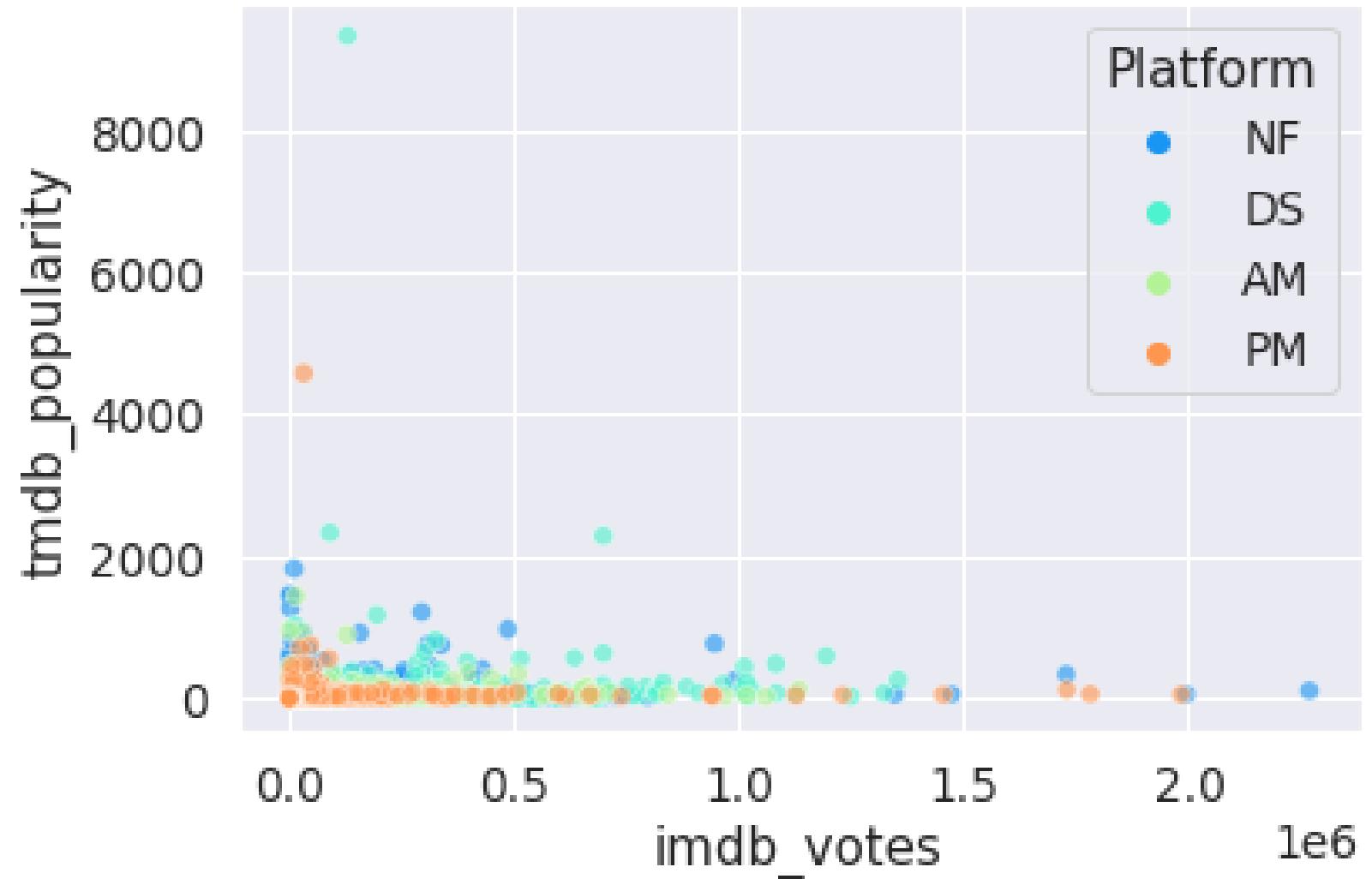
IMDB Score와 TMDB Score는 높은 상관관계

Spearman 상관계수의 경우 약 0.707

→ 어떤 db사이트라도 사람의 평가가 비슷한 경향을 보인다

데이터 분석

score에 영향을 미치는 요소는?



`imdb_votes`, `tmdb_popularity`는 높은 상관관계

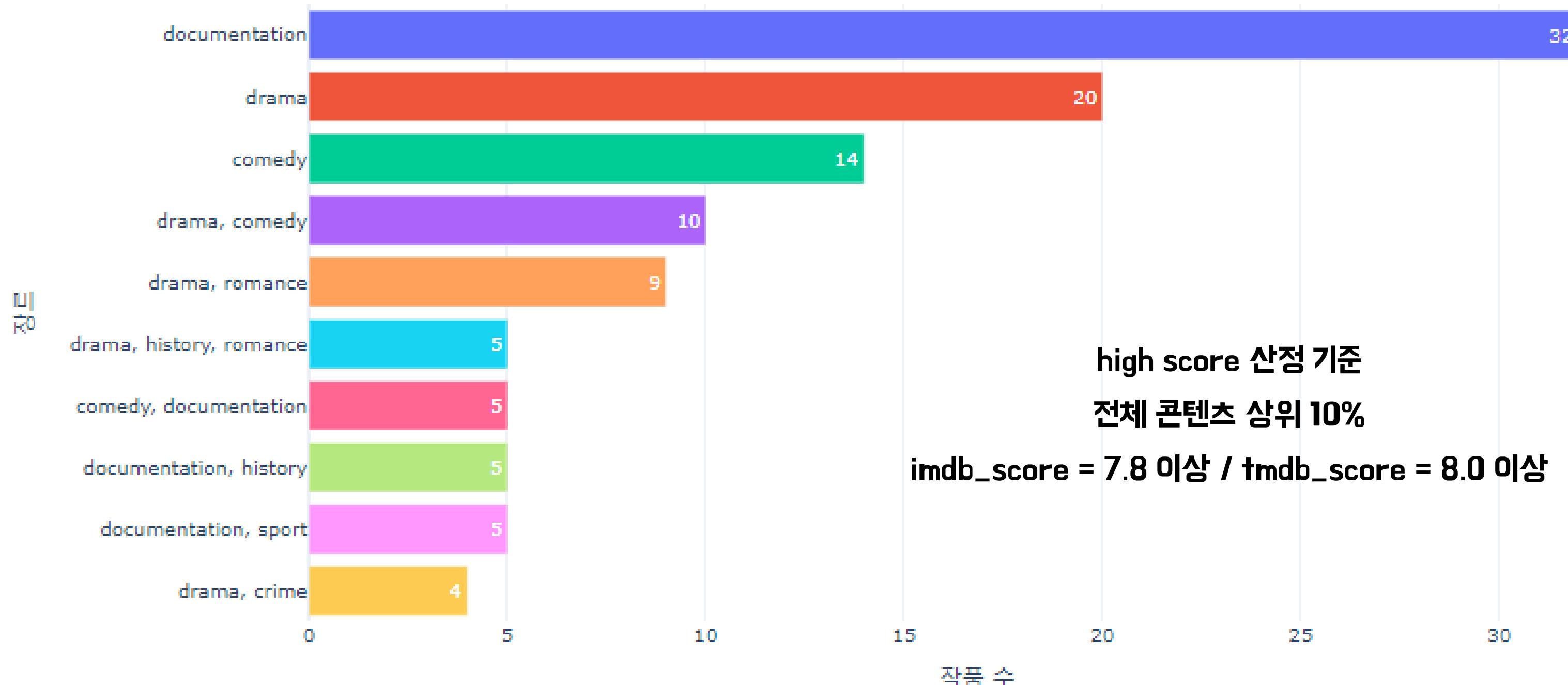
Spearmann 상관계수 약 0.7604

→ 두 db 사이트의 인기 척도가 상관성을 갖고 있다

데이터 분석

높은 score를 받은 영화의 특징?

장르별 작품 수 Top20

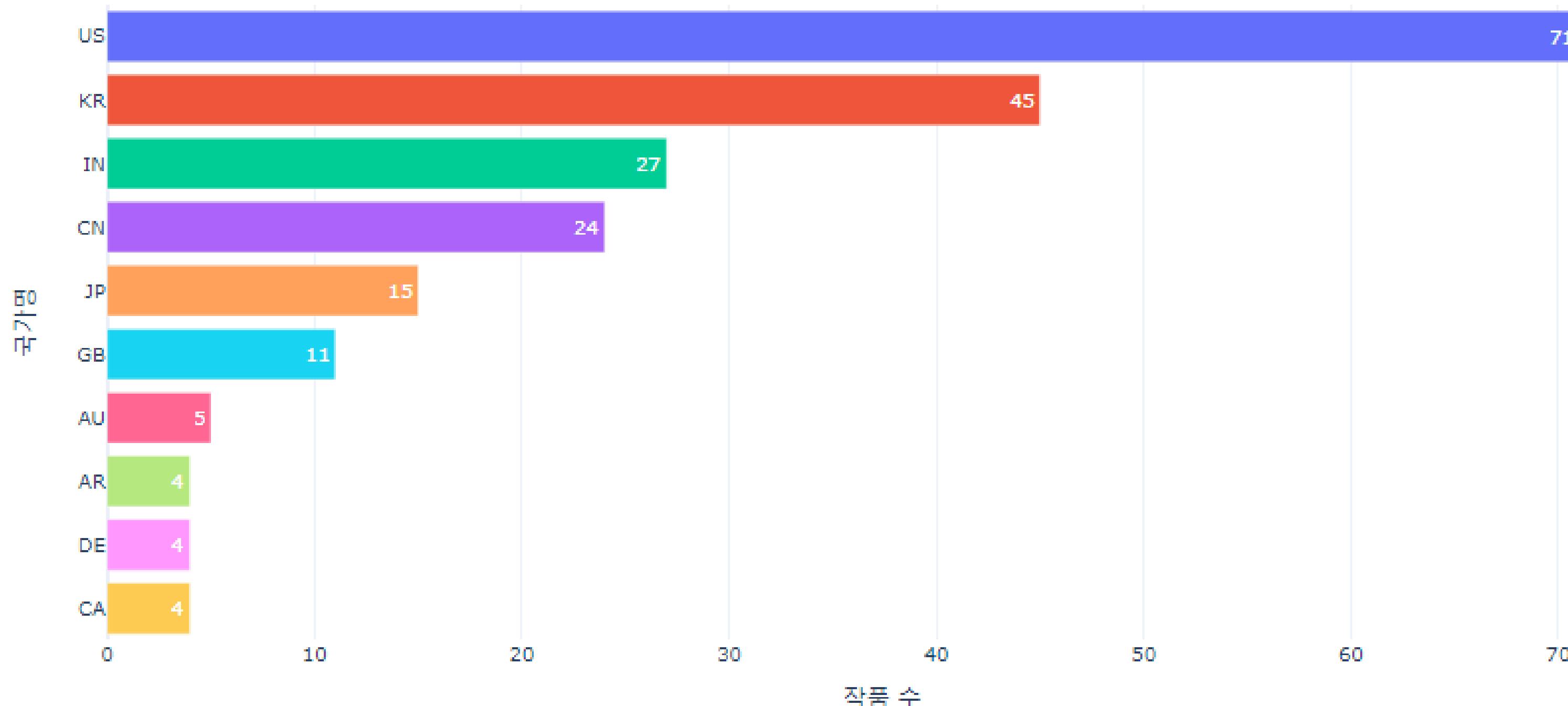


중복장르로 drama, documentation 순

데이터 분석

높은 score를 받은 영화의 특징?

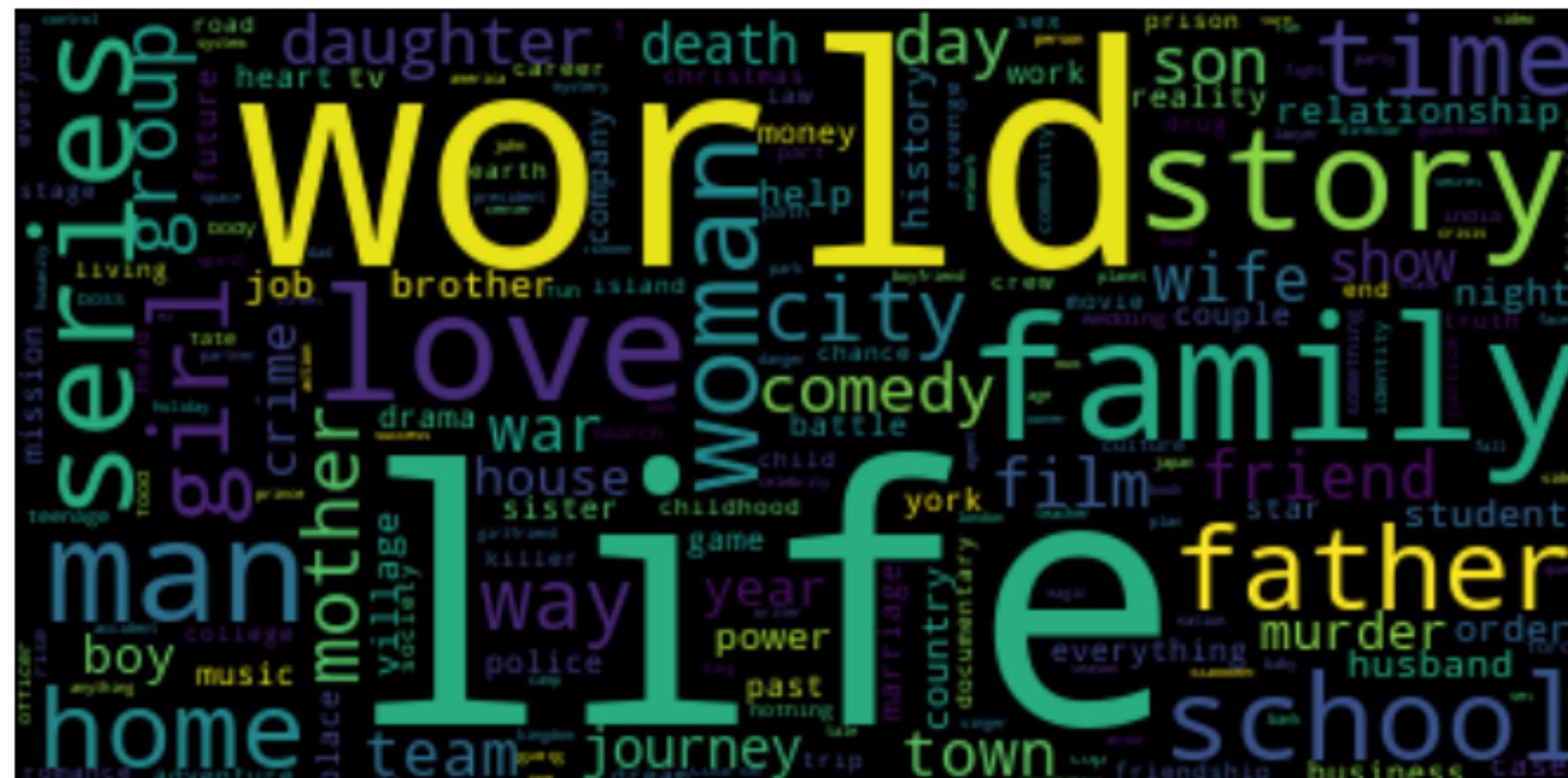
제작 국가별 작품 수 Top20



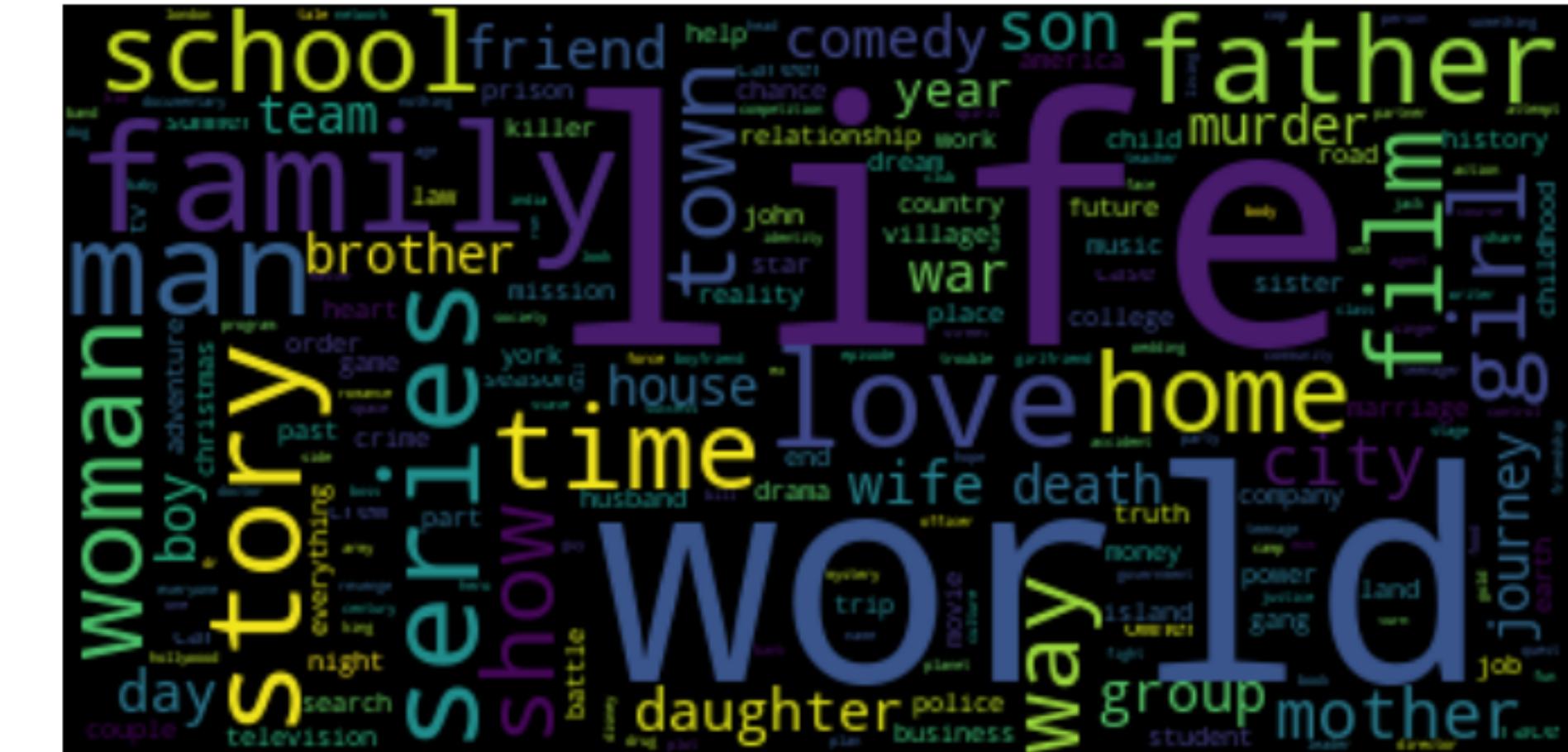
K-컨텐츠 전체 비중 6위 → High score 작품 비중 2위

데이터 분석

높은 score를 받은 영화의 특징?



high-score

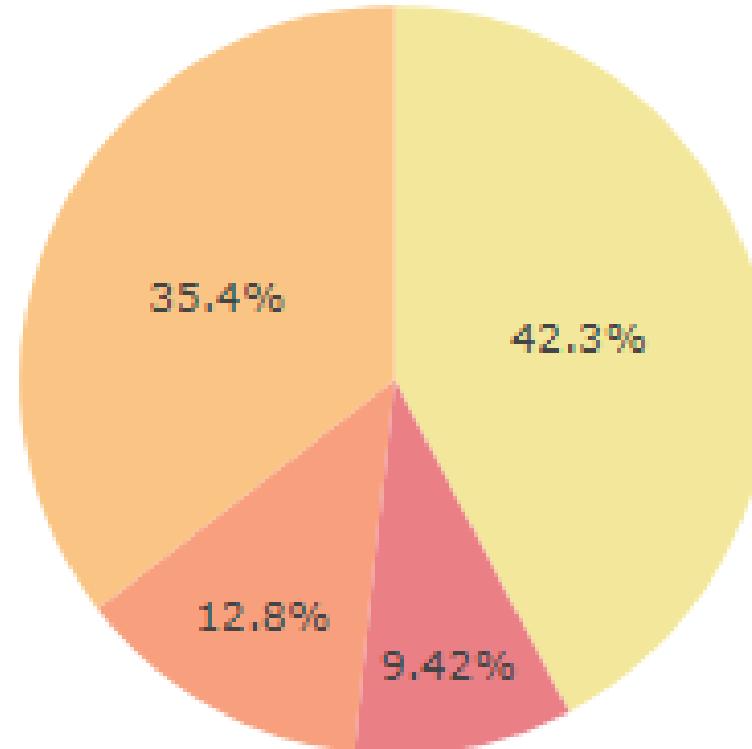


original

데이터 분석

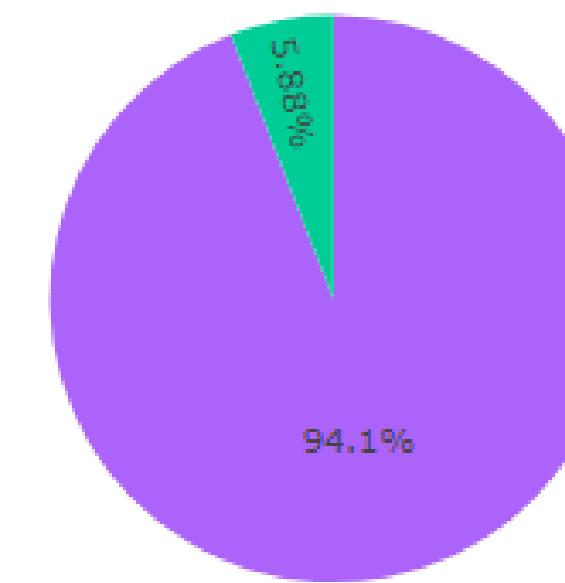
높은 score를 받은 컨텐츠를 많이 보유하고 있는 곳은?

High_score_pie_chart



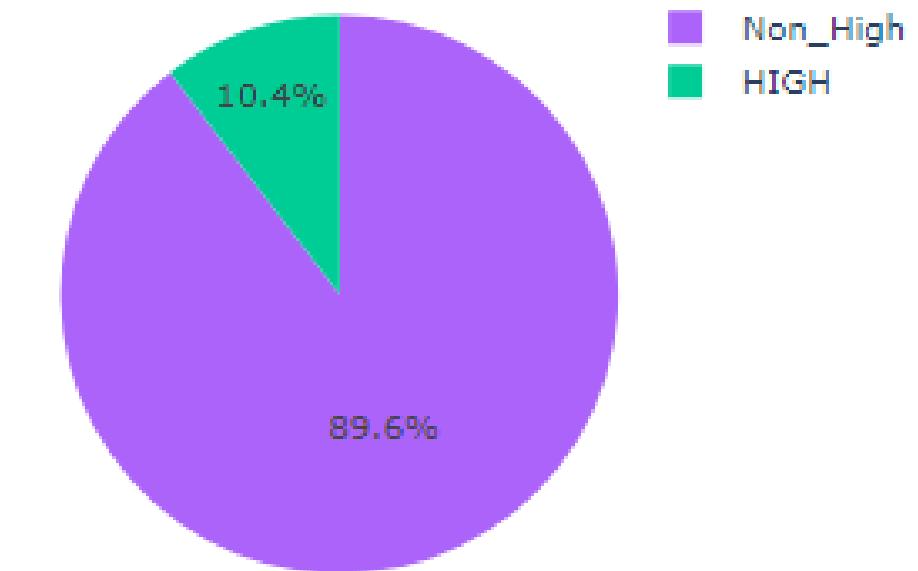
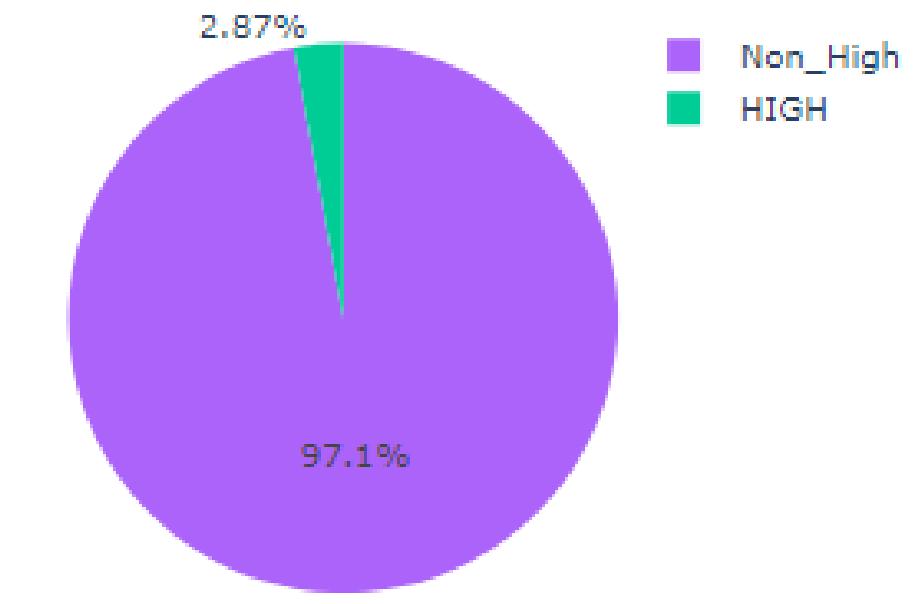
Netflix

DISNEY



AMAZON

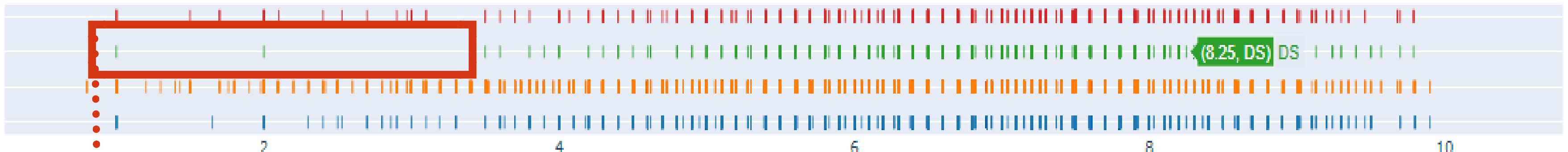
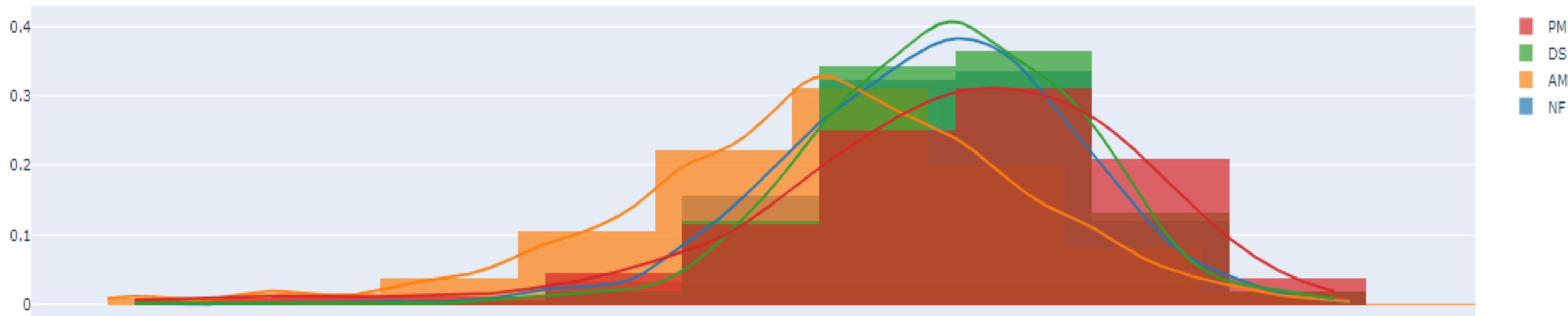
PARAMOUNT



전체 컨텐츠 대비 high score 컨텐츠 수는 Paramount 가 가장 많이 보유한 것으로 확인 됨

데이터 분석

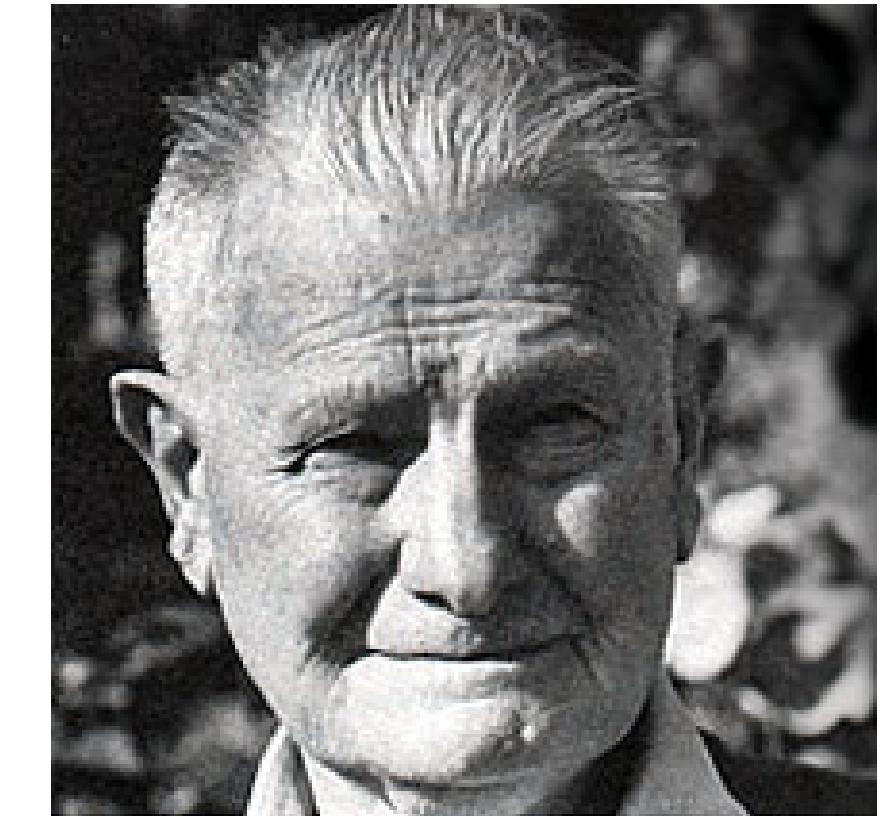
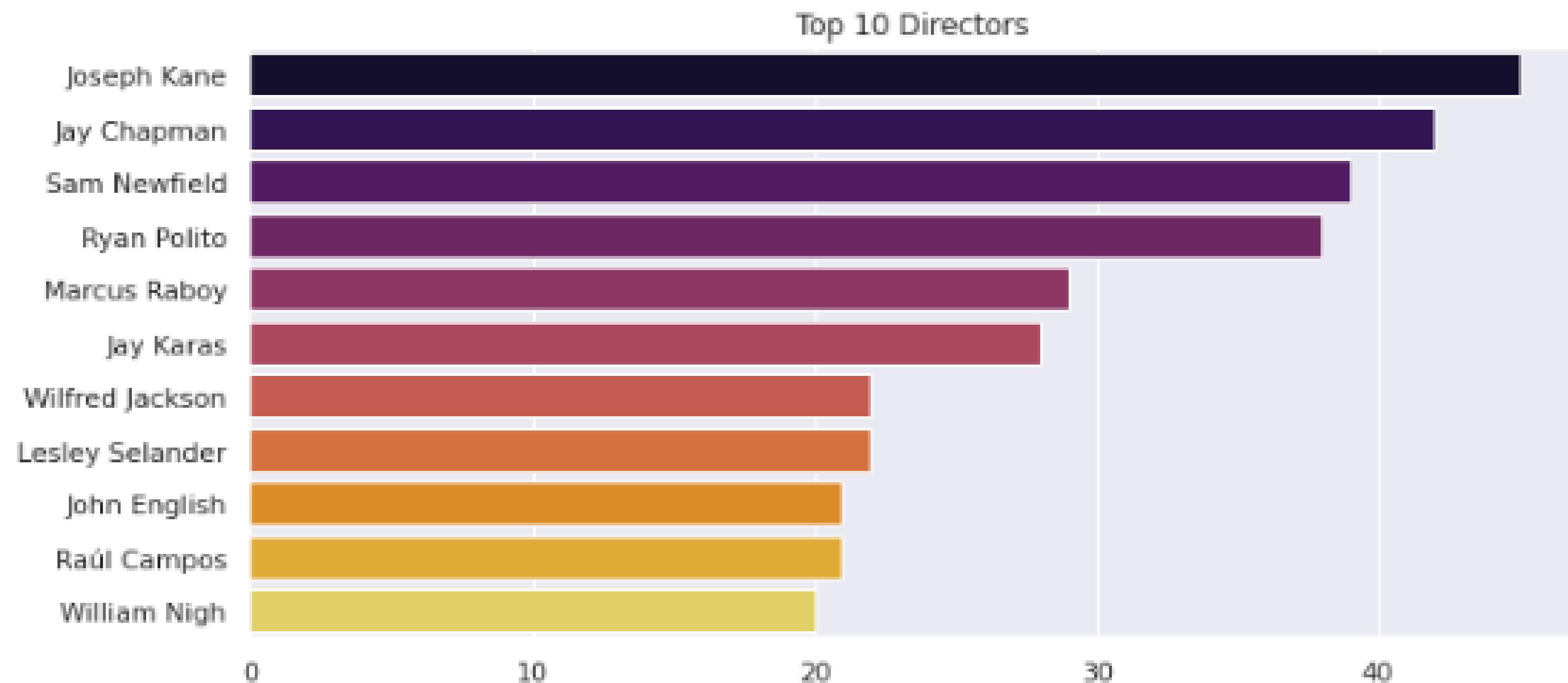
높은 score를 받은 컨텐츠를 많이 보유하고 있는 곳은?



• tmdb_score 분포를 확인했을 때, Disney+가 낮은 점수의 콘텐츠가 적은 편이란 걸 알 수 있음

데이터 분석

가장 많은 작품을 만든 감독은?

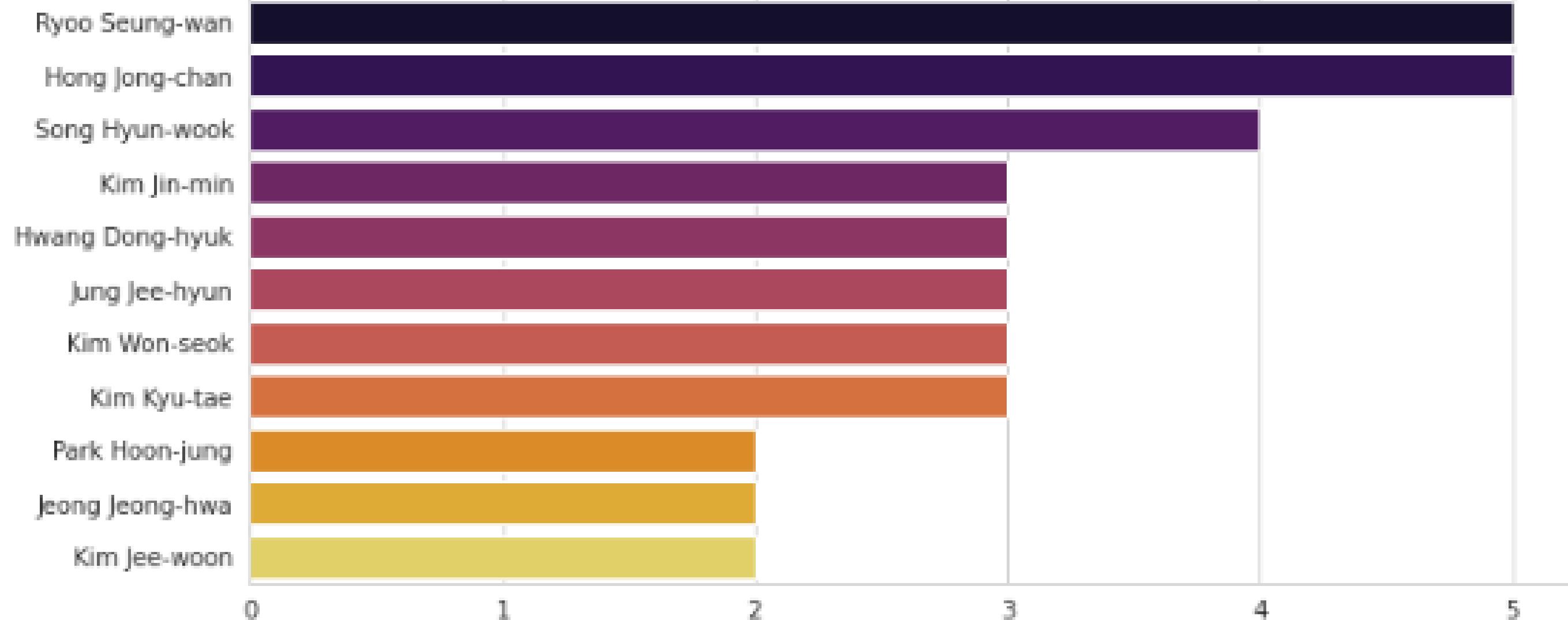


재스퍼 조셉 인만 케인 Jasper Joseph Inman Kane
(1894년 3월 19일 - 1975년 8월 25일)
무법 90년대, 바바리 해안의 불꽃 등

데이터 분석

가장 많은 작품을 만든 감독은?

Korea Top 10 Directors



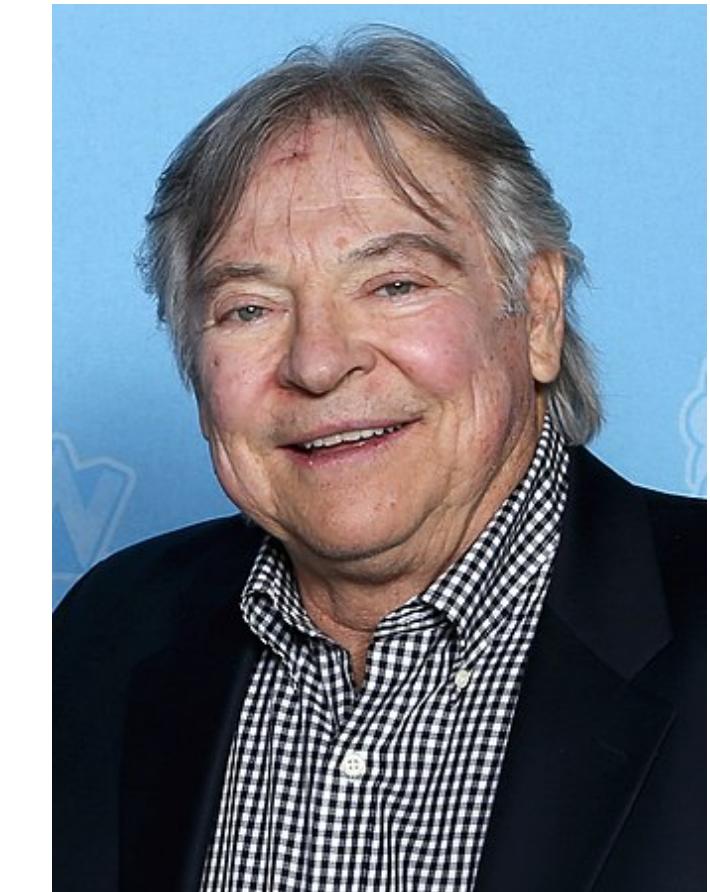
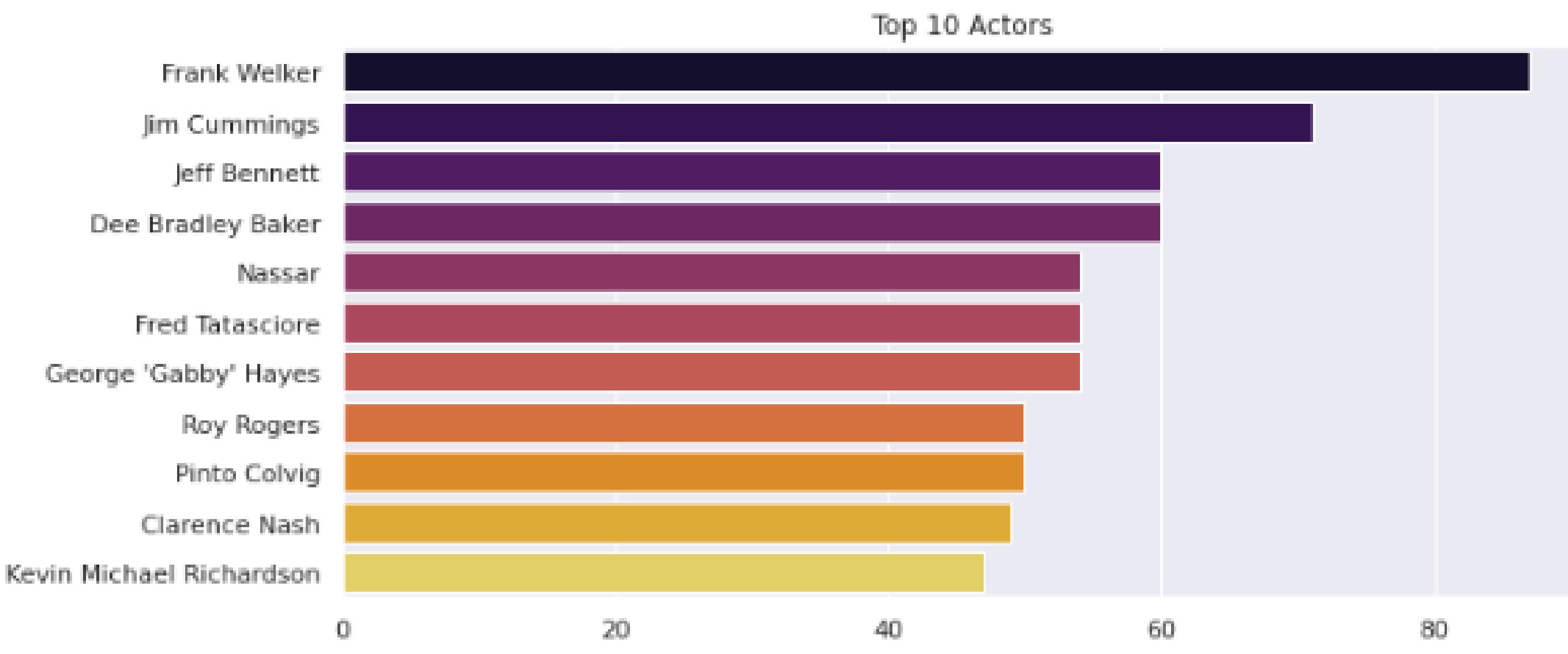
류승완

(1973년 12월 15일 ~)

모가디슈, 베테랑, 베를린 등

데이터 분석

가장 많은 작품에 출연한 배우는?



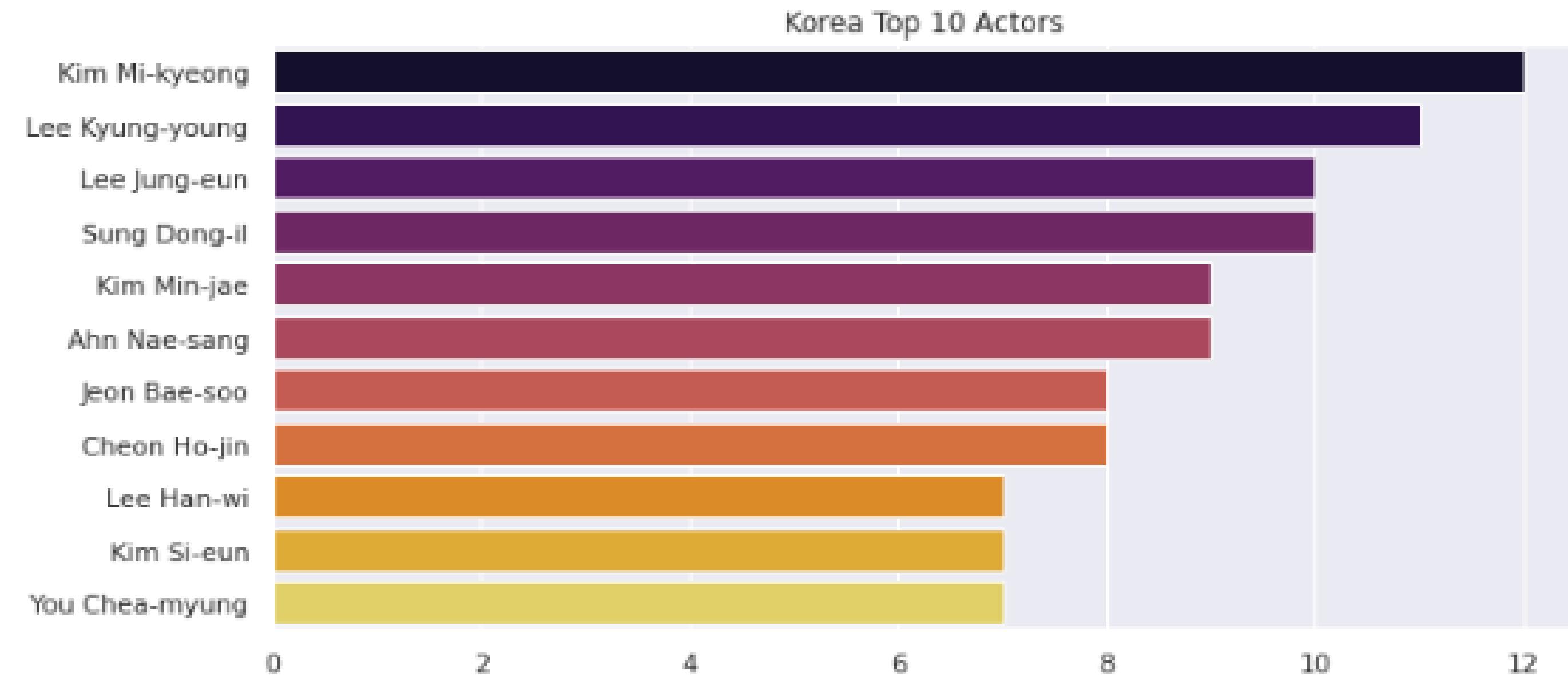
프랭클린 웬델 웰커 Franklin Wendell Welker

(1946년 3월 12일 ~)

성우 트랜스포머 메가트론, 가필드 등

데이터 분석

가장 많은 작품에 출연한 배우는?

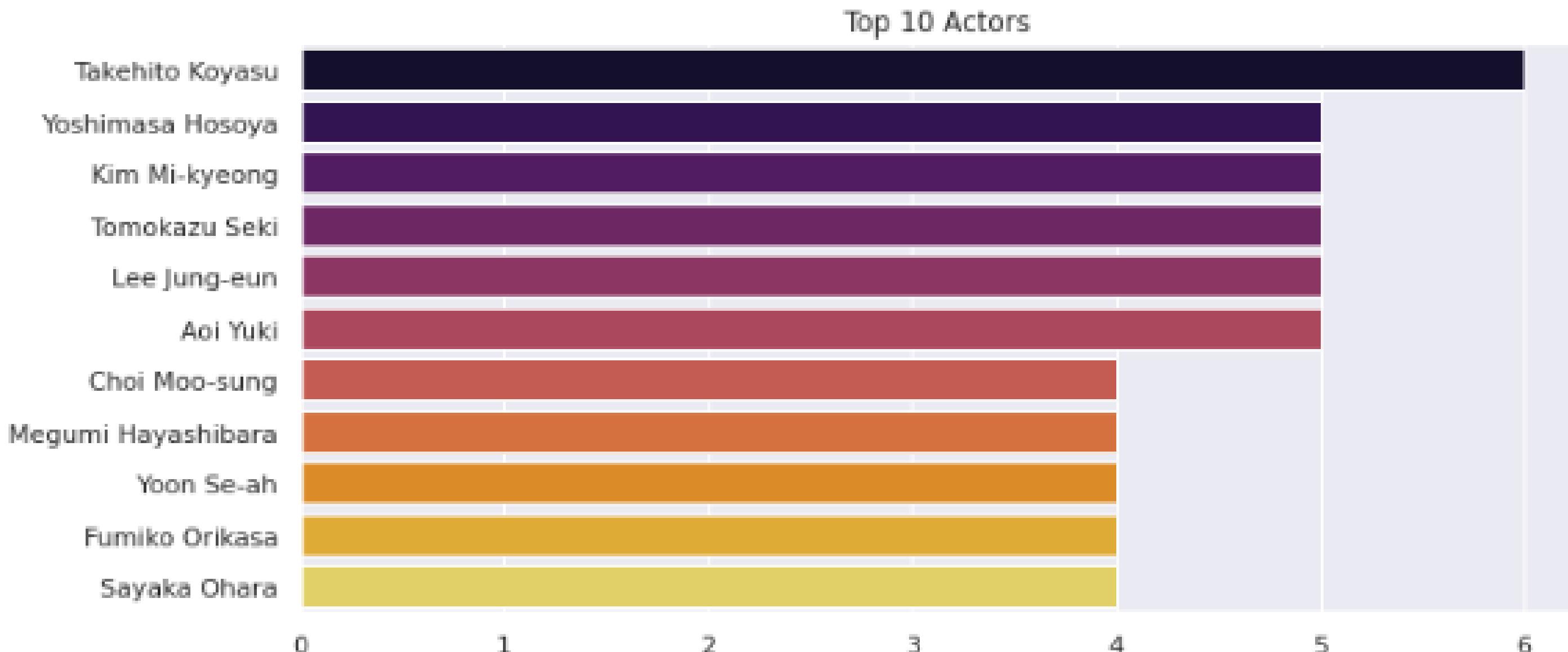


(1963년 10월 14일 ~)

기상청 사람들, 사이코지만 괜찮아, VIP 등

데이터 분석

가장 많은 high-score 작품에 출연한 배우는?



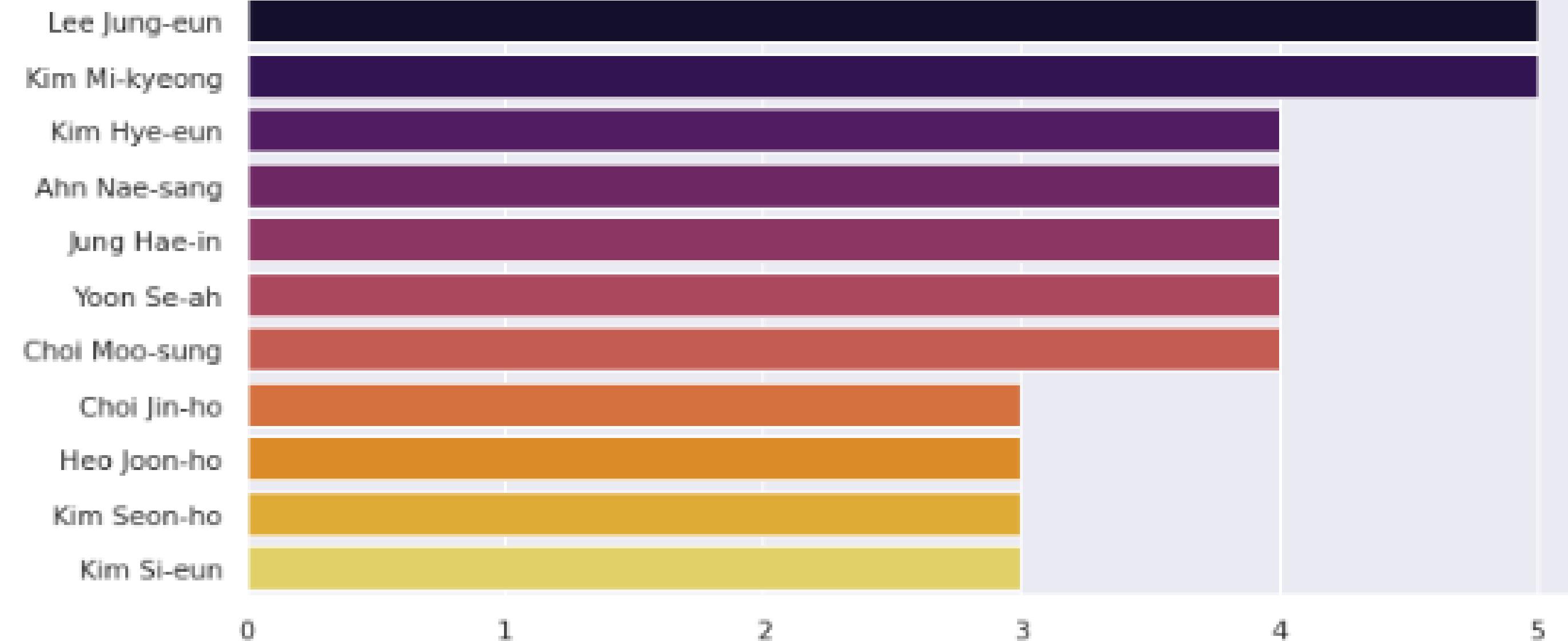
고야스 타케히토
(1967년 5월 5일 ~)

진격의 거인 비스트 타이탄, 포켓몬스터 등

데이터 분석

가장 많은 high-score 작품에 출연한 배우는?

Korea Top 10 Actors



이정은

(1970년 1월 23일 ~)

기생충, 옥자, 우리들의 블루스 등

결론

1. 어떤 콘텐츠가 많을까?

#Movie #100분 #Adults #US #Comedy #Drama

2. 어떤 키워드가 많을까?

Title #Love #Man Description #Life #World

3. Score에 영향을 미치는 요소는?

Tmdb_score - Imdb_score, vote - popularity

4. 높은 score를 기록한 콘텐츠의 특징은?

US, KR 콘텐츠, 비슷한 키워드

5. 높은 score 작품을 많이 보유하고 있는 플랫폼

보유수: Netflix, 비중: Paramount, 평균 score ↑ : Disney+

6. 가장 많이 등장하는 배우와 감독은?

Jasper Joseph Inman Kane, Franklin Wendell Welker, KR) 류승완, 김미경

결론

참고자료

	Netflix	Amazon Prime	Disney+	Paramount
월 요금	9,500원	5,500원	9,900원	6,200원
점유율	1 (48.3%)	2 (12.7%)	3 (7.3%)	4 (2.7%)
low score ↓	3	4	1	2
high score 보유 비율	3	4	2	1

심화 과제

추천 서비스를 만들어보자

영화명

콘텐츠

유저 ID + rating

시청한 콘텐츠에 대한 스코어 정보

+

무비 ID

original_title	!Women Art Revolution	'Gator Bait	'Twas the Night Before Christmas
userId			
1	NaN	NaN	NaN
2	NaN	NaN	NaN
3	NaN	NaN	NaN

피벗테이블 생성

기준 데이터셋



피어슨 상관계수

	Title	Correlation	Genre
0	Wild Wild West	1.00	[Action, Adventure, Comedy, Science Fiction, Western]
1	Population 436	0.88	[Drama, Horror, Mystery, Thriller]
2	The Dark Knight	0.87	[Drama, Action, Crime, Thriller]
3	Midnight in the Garden of Good and Evil	0.77	[Crime, Drama, Mystery, Thriller]
4	The Sentinel	0.75	[Horror, Drama, Mystery]
5	Godzilla	0.73	[Science Fiction, Action, Thriller]
6	Deadlier Than the Male	0.68	[Action, Comedy, Thriller]
7	A Kiss Before Dying	0.61	[Drama, Thriller, Crime, Mystery, Romance]
8	Bushwhacked	0.61	[Adventure, Action, Comedy, Crime, Family]
9	Wet Hot American Summer	0.61	[Comedy]

역할 분배

이호원

시각화
결측치 처리
분석 진행
가설 검증
방향 설정

추천 서비스 구현

김도균

시각화
결측치 처리
분석 진행
가설 검증
방향 설정

코드 취합 및 점제

오준엽

시각화
결측치 처리
분석 진행
가설 검증
방향 설정

인사이트 취합 및 가이드

**Thank you
for watching!**