· DSL 12th EDA PROJECT ·

# 성격 특성 기반 고객 타겟팅 전략

## : Big5 성격 요인과 FoMO를 기반으로 한 영화 제작 접근법

11기 백두형, 김여원, 12기 복지민, 이정우

00

# INDEX

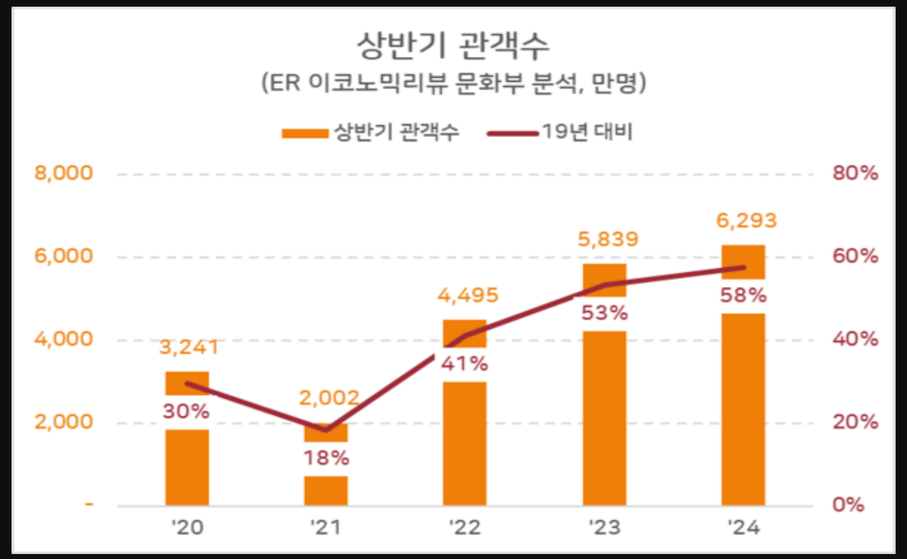
# 01 프로젝트 소개

다시 살아난 극장가?...풍악 울리지 못하는 이유는 [무비 인사이드]

김예랑 기자

입력 2024.08.04 22:20 수정 2024.08.04 22:20 가가

상반기 천만 영화 두편 '최초'
매출액 팬데믹 이전 72% 수준으로 회복




영화 제작사의 입장에서…

## 수익을 극대화할 수 있는 영화는 무엇일까?

고려할 수 있는 요소 → 고객 선호도

시장 트렌드

영화 콘텐츠

https://www.hankyung.com/article/2024080277867, https://www.econovill.com/news/articleView.html?idxno=659133

01

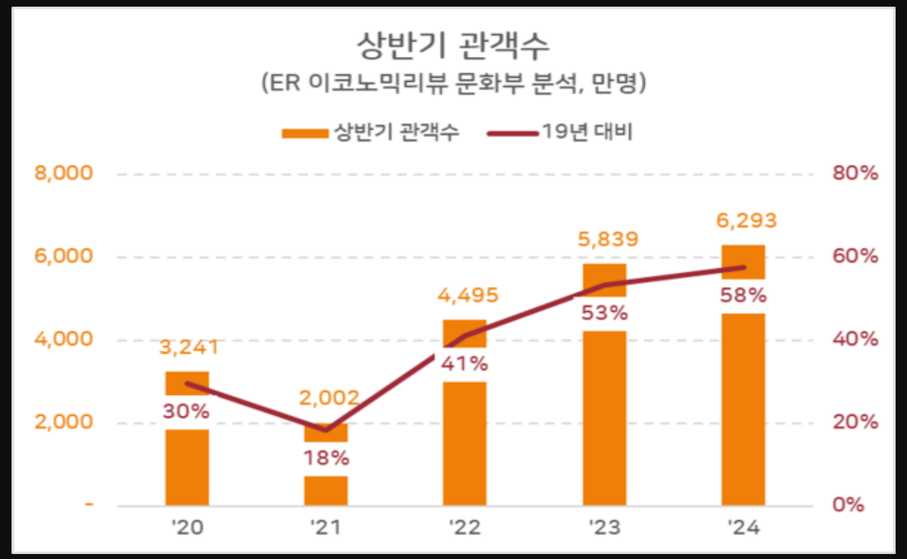
# 프로젝트 소개

다시 살아난 극장가?...풍악 울리지 못하는 이유는 [무비 인사이드]

김예랑 기자

입력 2024.08.04 22:20 수정 2024.08.04 22:20 가가

상반기 천만 영화 두편 '최초'
매출액 팬데믹 이전 72% 수준으로 회복




영화 제작사의 입장에서…

## 수익을 극대화할 수 있는 영화는 무엇일까?

“입소문” 고객 선호도

고려할 수 있는 요소 → 시장 트렌드

영화 콘텐츠

https://www.hankyung.com/article/2024080277867, https://www.econovill.com/news/articleView.html?idxno=659133

# 01 프로젝트 소개

**여름 영화 재밌다는데…왜 '대작'·'대박' 없을까 [N초점]**

정유진 기자

2024.07.14 오전 08:00

팬데믹 이후 관객들의 변화에 발맞춰 변화하고 있는 배급 전략을 주목해 볼 필요가 있다. 코로나19 팬데믹 이후의 관객들은 '입소문'의 영향을 많이 받으며 '꼭 봐야할 영화'로 여겨지는 작품들을 선택하는 경향이 큰 것으로 인식되고 있다. 과거에는 극장에 가서 볼 영화를 고르는 관객도 많았지만, 요즘에는 먼저 볼 영화를 고르고 극장에 가는 관객들이 많다고 보는 것이 업계의 중론이다. 그에 따라 각 배급사에

Our Goal…

## 고객이 심리적으로 '입소문' 낼 만한 영화를 찾아보자!

우리에게 필요한 정보 → 고객의 심리정보

고객이 본 영화

고객의 평가

https://www.news1.kr/entertain/movie/5477448

02 데이터셋

# < User personality dataset >

| | userid | openness | agreeableness | emotional_stability | conscientiousness | extraversion | movie_id | rating | title | directedBy | starring | avgRating | imdbId |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 5.0 | 2.0 | 3.0 | 2.5 | 6.5 | 1 | 5.0 | Toy Story (1995) | John Lasseter | Tim Allen, Tom Hanks, Don Rickles, Jim Varney,... | 3.89146 | 114709 |
| 1 | 0 | 5.0 | 2.0 | 3.0 | 2.5 | 6.5 | 5265 | 4.0 | Death to Smoochy (2002) | Danny DeVito | Danny DeVito, Edward Norton, Robin Williams, C... | 3.15655 | 266452 |
| 2 | 0 | 5.0 | 2.0 | 3.0 | 2.5 | 6.5 | 5258 | 3.0 | George Washington (2000) | David Gordon Green | Curtis Cotton III, Donald Holden, Damian Jewan... | 3.58985 | 262432 |
| 3 | 0 | 5.0 | 2.0 | 3.0 | 2.5 | 6.5 | 5256 | 4.0 | Stolen Summer (2002) | Pete Jones | Brian Dennehy, Kevin Pollak, Aidan Quinn, Bonn... | 2.98246 | 286162 |
| 4 | 0 | 5.0 | 2.0 | 3.0 | 2.5 | 6.5 | 97752 | 3.5 | Cloud Atlas (2012) | Tom Tykwer, Andy Wachowski, Lana Wachowski | Tom Hanks, Halle Berry, Jim Broadbent, Hugo We... | 3.56958 | 1371111 |

- row 1,067,068 / column 13
- user 1,820 / movie 34,816

+ 리뷰 데이터 분석을 위한 IMDB 리뷰 크롤링

- 각 user의 Big5 score
- user가 준 영화별 평점

데이터 출처 : https://www.kaggle.com/datasets/arslanali4343/top-personality-dataset?select=2018-personality-data.csv

02 데이터셋

# < User personality dataset >

| | userid | openness | agreeableness | emotional_stability | conscientiousness | extraversion | movie_id | rating | title | directedBy | starring | avgRating | imdbId |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 5.0 | 2.0 | 3.0 | What is Big5? | 6.5 | 1 | 5.0 | Toy Story (1995) | John Lasseter | Tim Allen, Tom Hanks, Don Rickles, Jim Varney,... | 3.89146 | 114709 |
| 1 | 0 | 5.0 | 2.0 | 3.0 | 2.5 | 6.5 | 5265 | 4.0 | Death to Smoochy (2002) | Danny DeVito | Danny DeVito, Edward Norton, Robin Williams, C... | 3.15655 | 266452 |
| 2 | 0 | 5.0 | 2.0 | 3.0 | 2.5 | 6.5 | 5258 | 3.0 | George Washington (2000) | David Gordon Green | Curtis Cotton III, Donald Holden, Damian Jewan... | 3.58985 | 262432 |
| 3 | 0 | 5.0 | 2.0 | 3.0 | 2.5 | 6.5 | 5256 | 4.0 | Stolen Summer (2002) | Pete Jones | Brian Dennehy, Kevin Pollak, Aidan Quinn, Bonn... | 2.98246 | 286162 |
| 4 | 0 | 5.0 | 2.0 | 3.0 | 2.5 | 6.5 | 97752 | 3.5 | Cloud Atlas (2012) | Tom Tykwer, Andy Wachowski, Lana Wachowski | Tom Hanks, Halle Berry, Jim Broadbent, Hugo We... | 3.56958 | 1371111 |

- row 1,067,068 / column 13
- user 1,820 / movie 34,816

\+ 리뷰 데이터 분석을 위한 IMDB 리뷰 크롤링

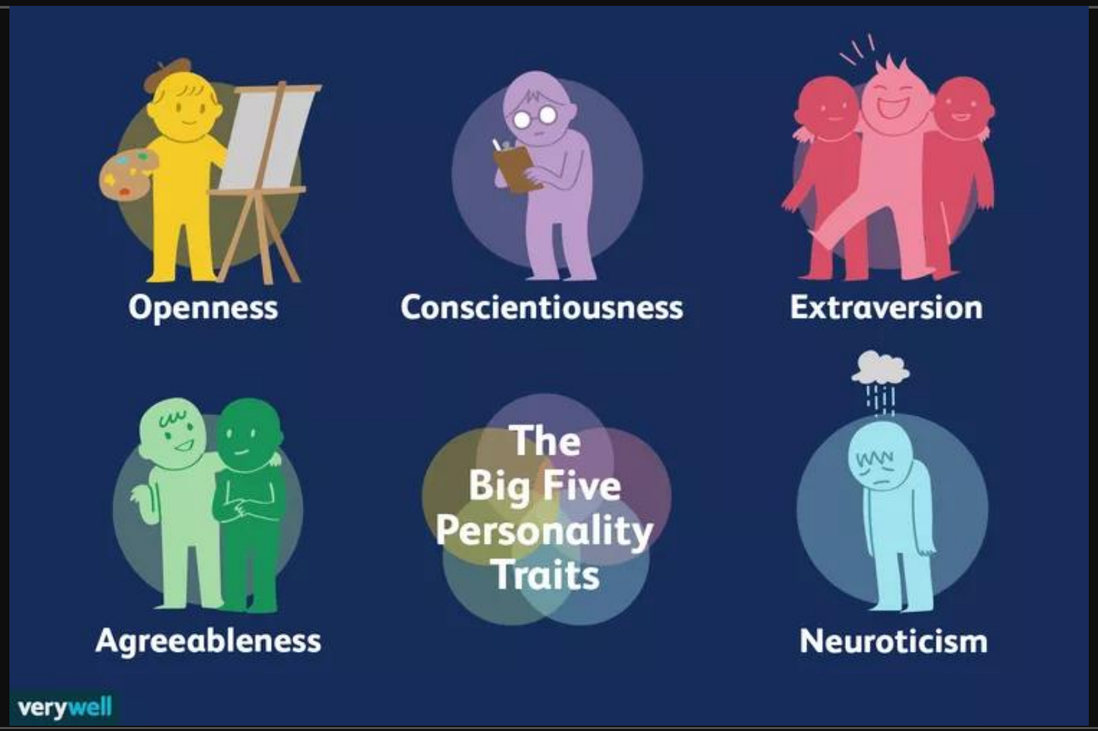
- 각 user의 Big5 score
- user가 준 영화별 평점

데이터 출처 : https://www.kaggle.com/datasets/arslanali4343/top-personality-dataset?select=2018-personality-data.csv

03 BIG 5

# < Big five Personality traits >

개인의 성격을 다섯 가지 주요 차원으로 설명하는 심리학 이론




- 개방성 (Openness to Experience)
- 성실성 (Conscientiousness)
- 외향성 (Extraversion)
- 친화성 (Agreeableness)
- 신경성 (Neuroticism)

  - 본 데이터에서는 신경성과 반대 지표인 Emotional Stability 사용

이미지 출처 : https://www.verywellmind.com/the-big-five-personality-dimensions-2795422

04

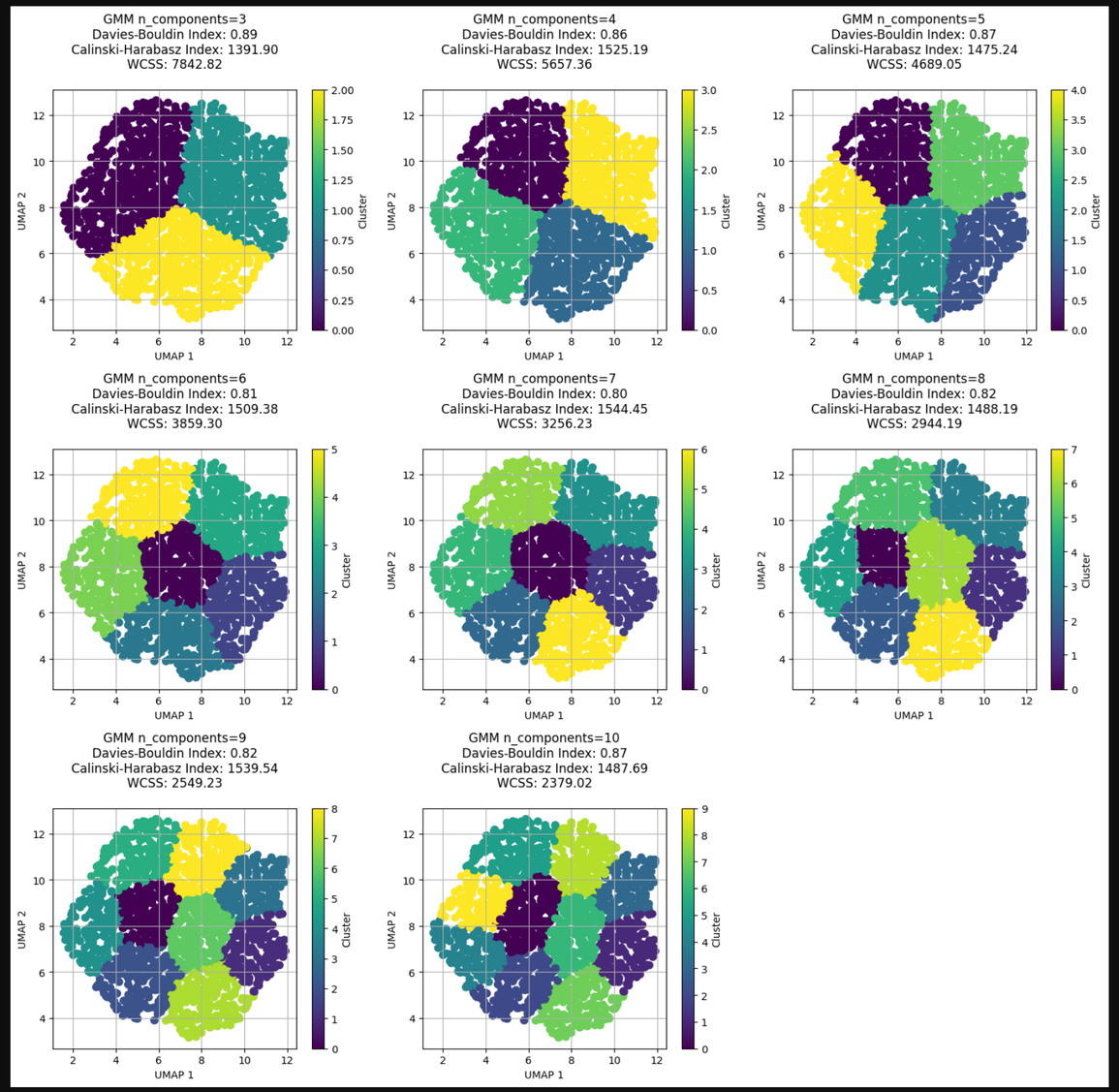
# 클러스터링

UMAP 차원축소 이후에,

Gaussian Mixture Model(GMM) 클러스터링

(n_components 는 3에서 10까지)

*표준화 스케일링 진행

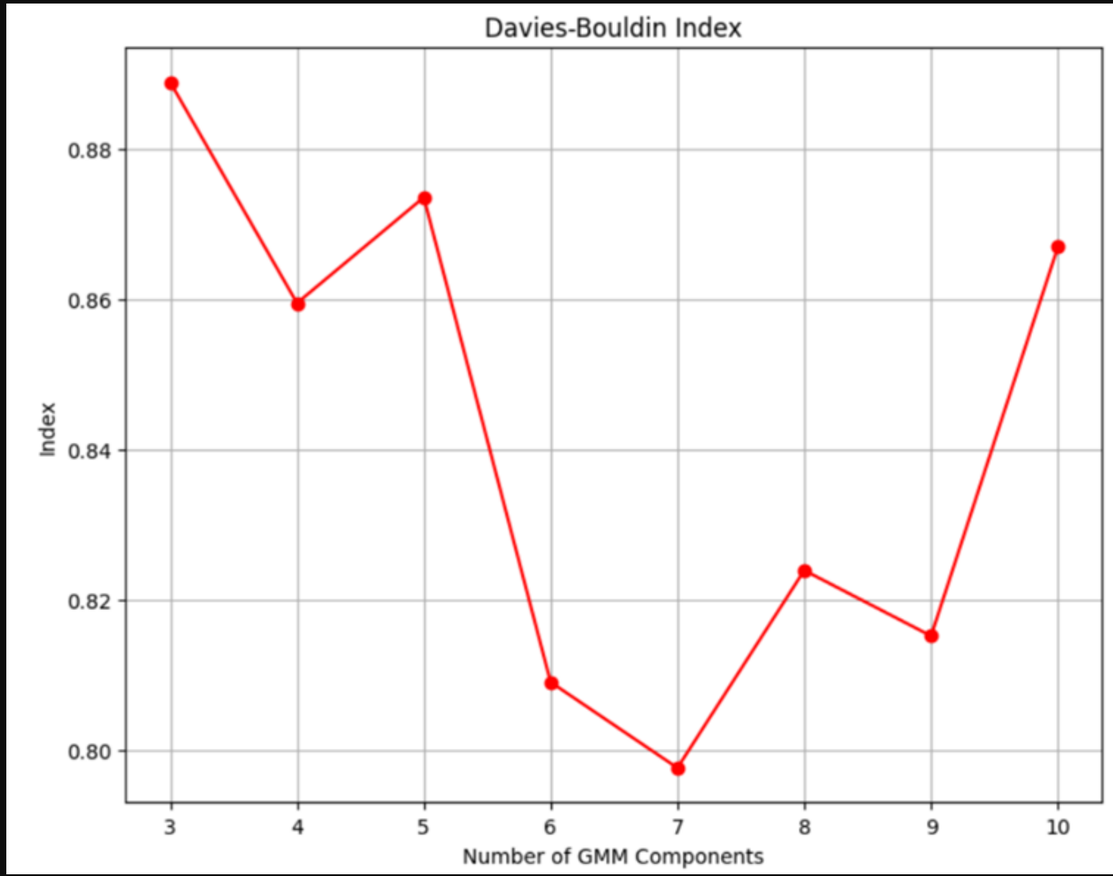


04

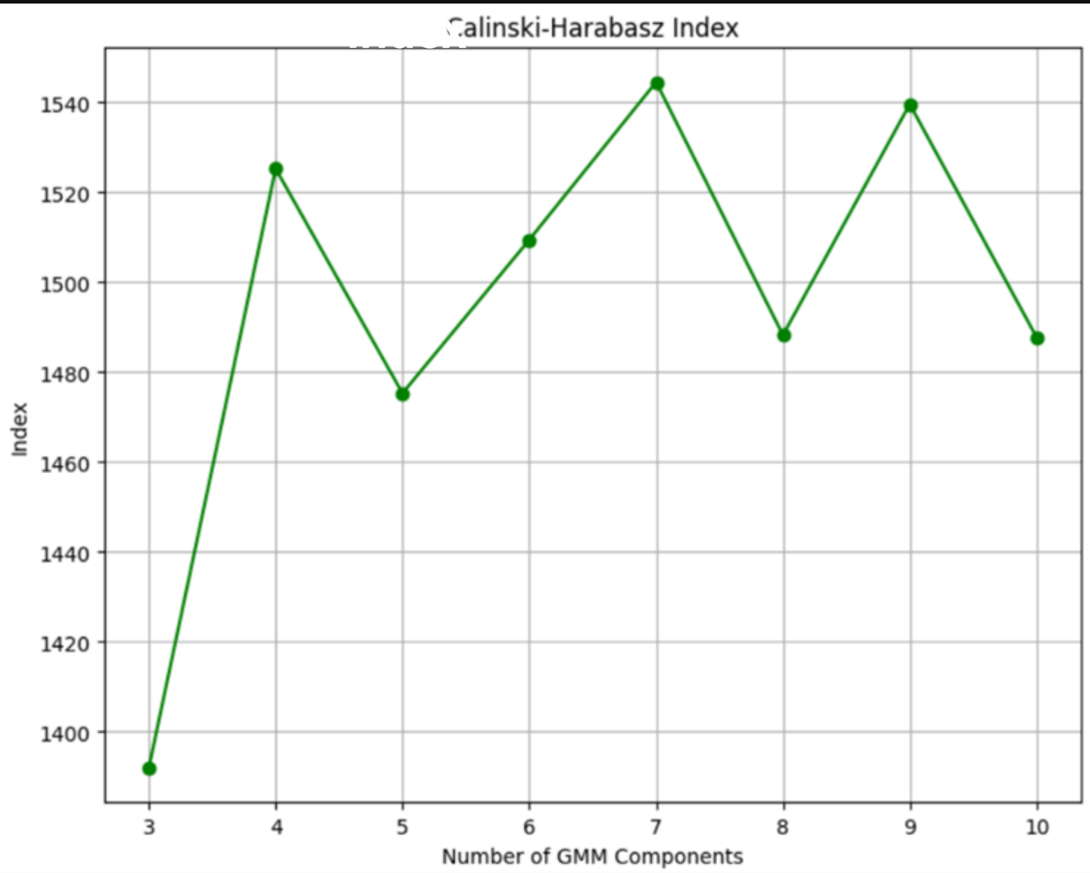
# 클러스터링

## Cluster 개수에 따른 Cluster 평가 지표

### Davies-Bouldin Index

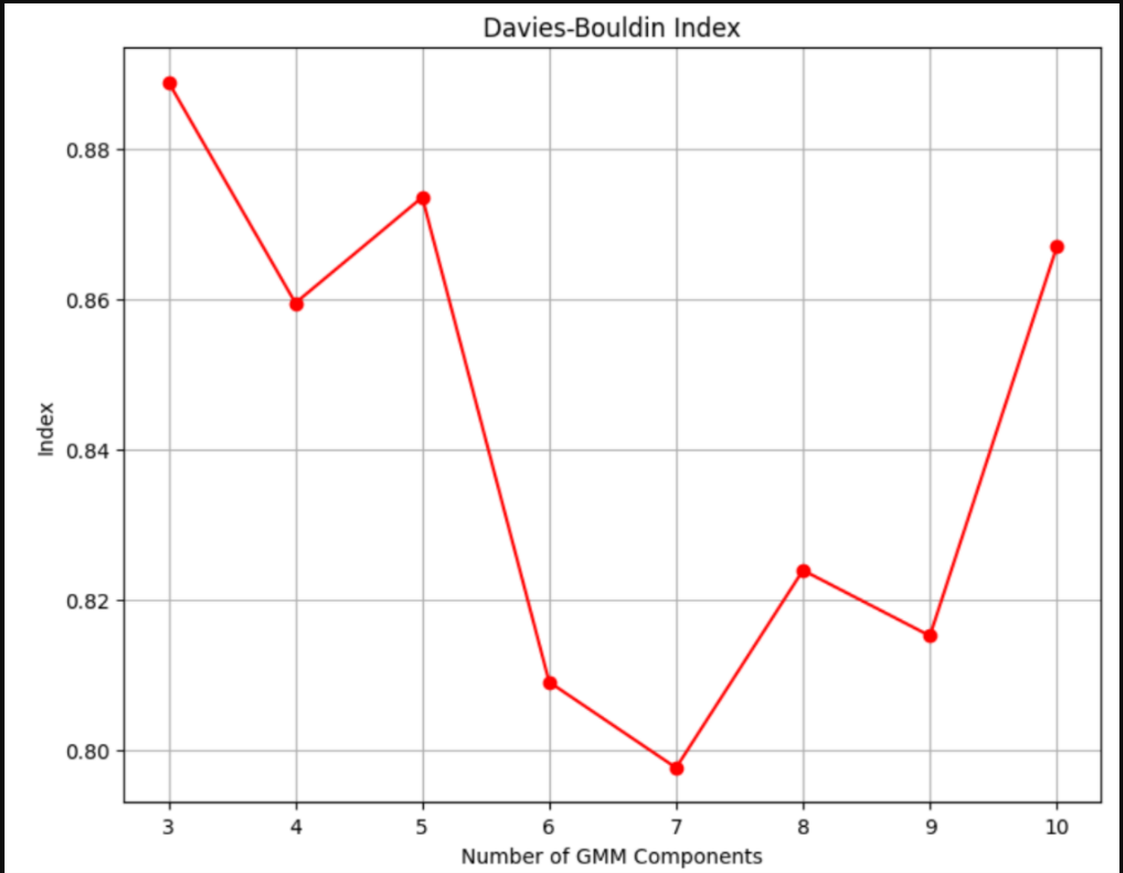



### Calinski-Harabasz



# 04 클러스터링

## Davies-Bouldin Index




This index signifies the average 'similarity' between clusters, where the similarity is a measure that compares the distance between clusters with the size of the clusters themselves.

Zero is the lowest possible score. Values closer to zero indicate a better partition.

**클러스터링 잘 되었을수록 점수 낮아짐**

The index is defined as the average similarity between each cluster $C_i$ for $i = 1, \ldots, k$ and its most similar one $C_j$. In the context of this index, similarity is defined as a measure $R_{ij}$ that trades off:

- $s_i$, the average distance between each point of cluster $i$ and the centroid of that cluster – also know as cluster diameter.
- $d_{ij}$, the distance between cluster centroids $i$ and $j$.

A simple choice to construct $R_{ij}$ so that it is nonnegative and symmetric is:

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}$$

Then the Davies-Bouldin index is defined as:

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{i \neq j} R_{ij}$$

https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation

# 04 클러스터링

## Calinski-Harabasz Index

The score is higher when clusters are dense and well separated, which relates to a standard concept of a cluster.

**클러스터링 잘 되었을수록 점수 높아짐**

For a set of data $E$ of size $n_E$ which has been clustered into $k$ clusters, the Calinski-Harabasz score $s$ is defined as the ratio of the between-clusters dispersion mean and the within-cluster dispersion:

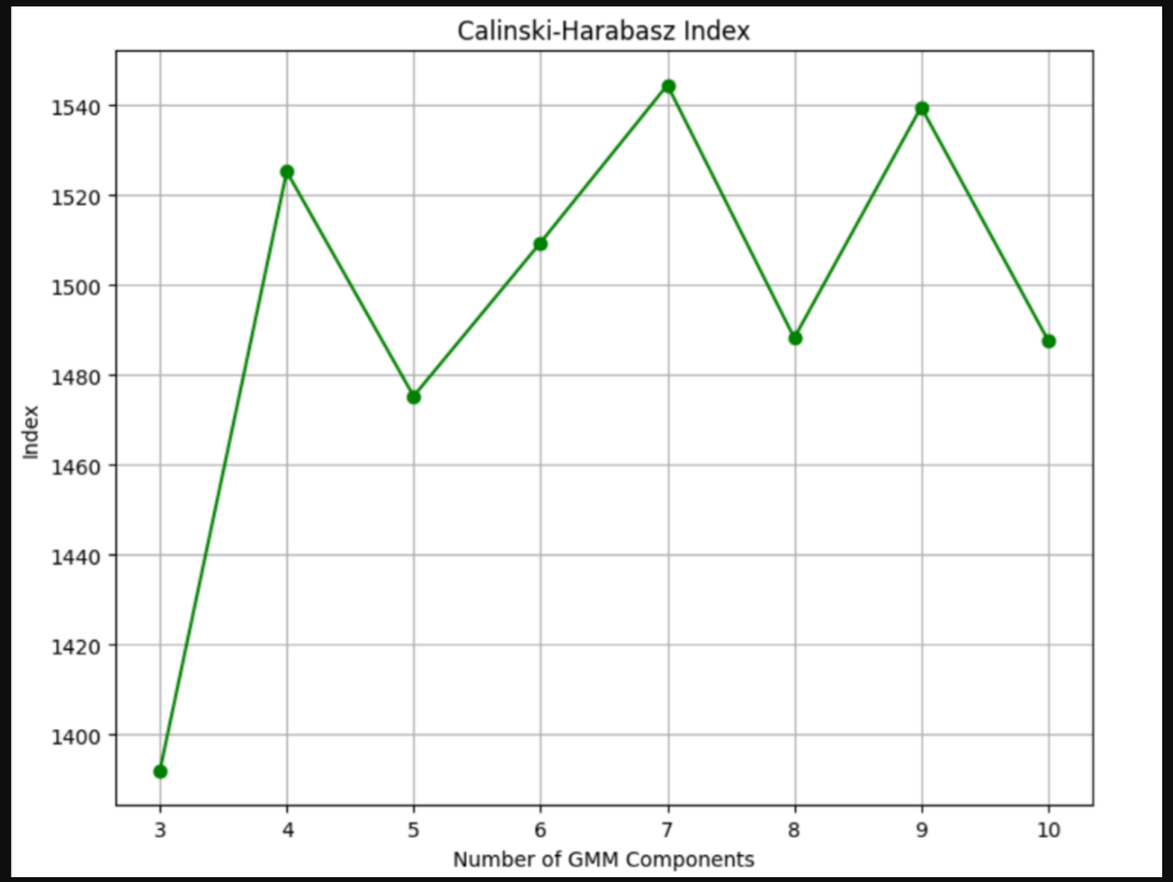
$$s = \frac{\mathrm{tr}(B_k)}{\mathrm{tr}(W_k)} \times \frac{n_E - k}{k - 1}$$

where $\mathrm{tr}(B_k)$ is trace of the between group dispersion matrix and $\mathrm{tr}(W_k)$ is the trace of the within-cluster dispersion matrix defined by:

$$W_k = \sum_{q=1}^{k} \sum_{x \in C_q} (x - c_q)(x - c_q)^T$$

$$B_k = \sum_{q=1}^{k} n_q (c_q - c_E)(c_q - c_E)^T$$

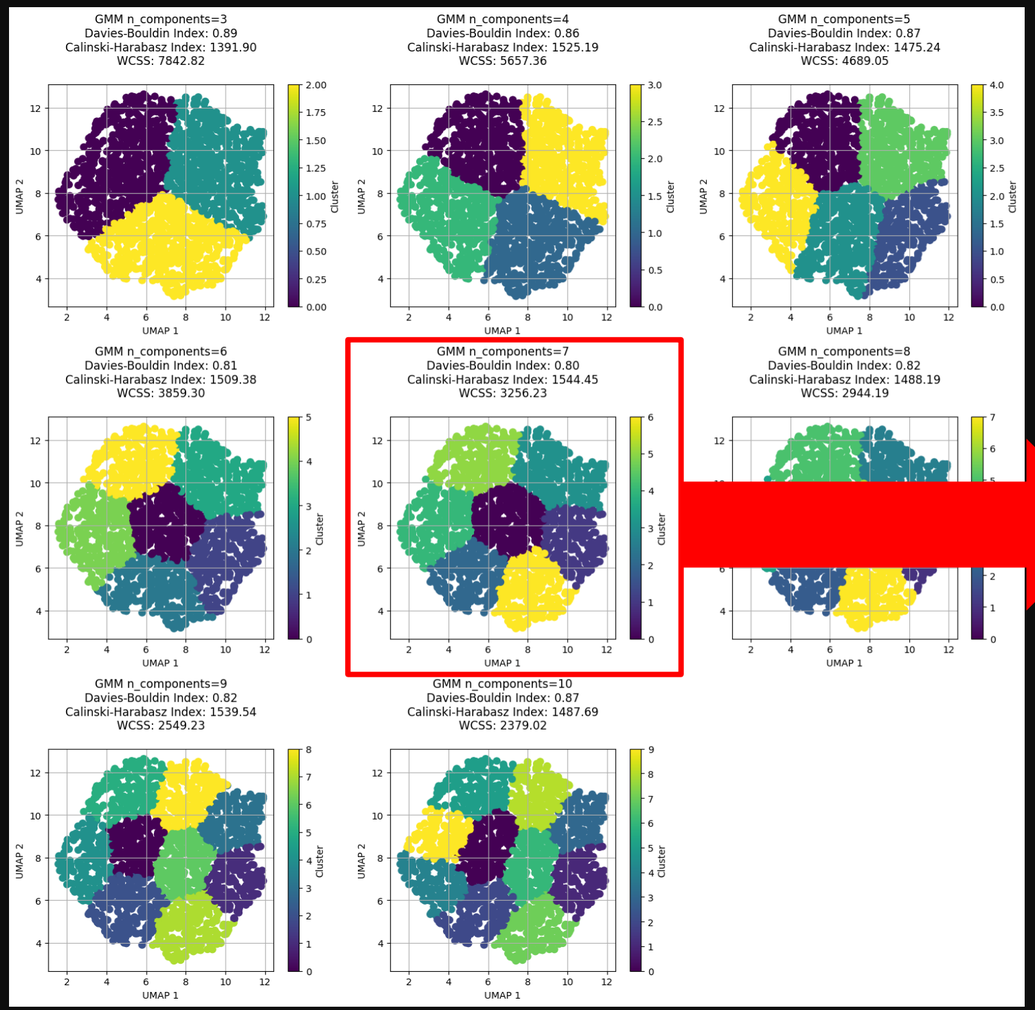
with $C_q$ the set of points in cluster $q$, $c_q$ the center of cluster $q$, $c_E$ the center of $E$, and $n_q$ the number of points in cluster $q$.

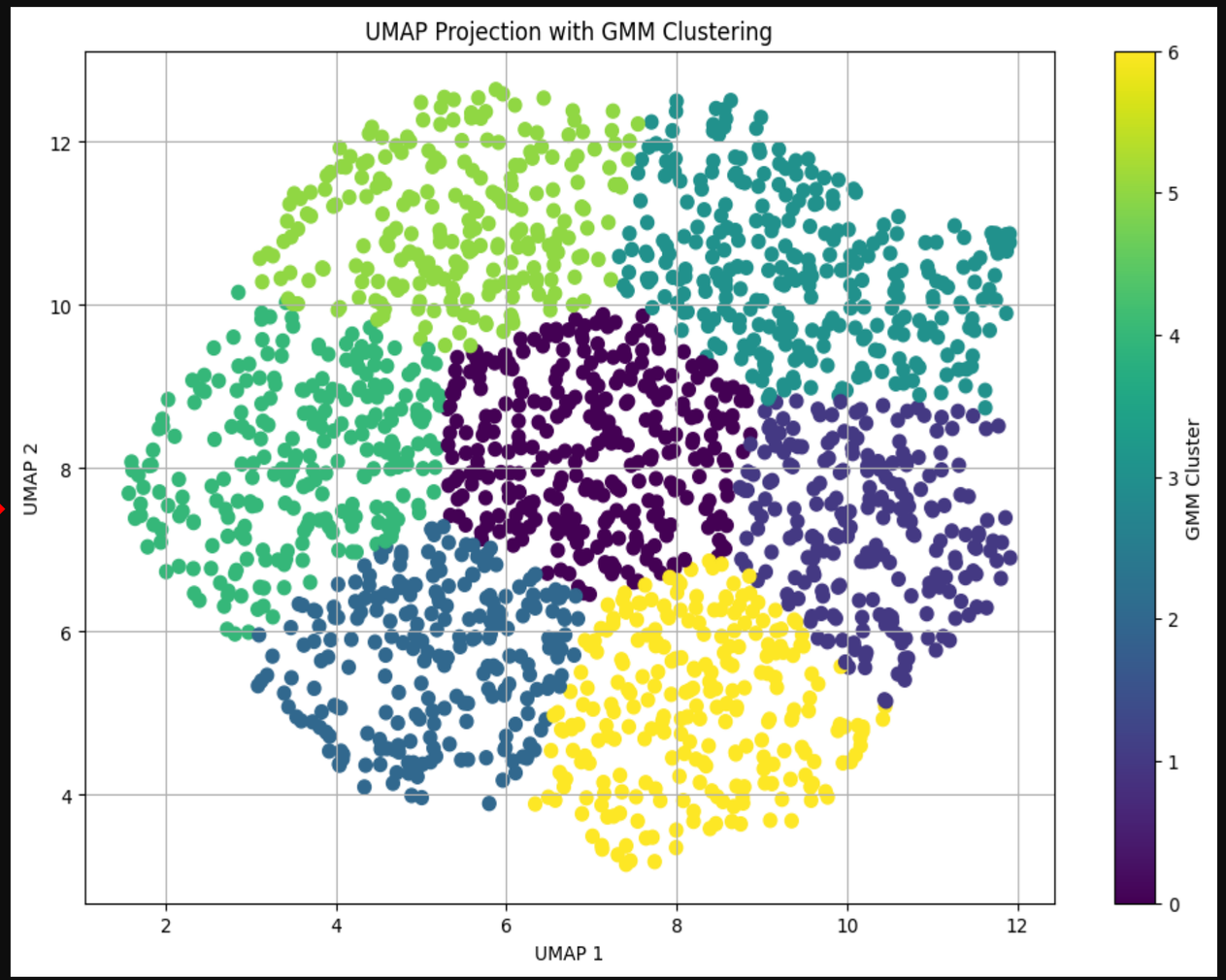



https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation

# 04 클러스터링

## 따라서 Davies-Bouldin Index가 가장 낮고, Calinski-Harabasz Index가 가장 높은 7개의 클러스터로 진행 !



# 05 클러스터별 분석

영화 제작사 : “어떤 클러스터의 고객을 타겟팅 해야 할까?”

**FoMO : Fear of Missing Out**

**(유행에 뒤쳐지는 것 같아 두려움과 스트레스 받는 상태)**

> 영화 관람 후 소셜 미디어에 경험 공유를 예측(Tefertiller, Maxwell & Morris, 2020)

FoMO가 높은 사람은 영화 관람 후 소셜 미디어에 활발히 공유하는 경향이 있을 것

Big5와의 상관 메타연구(Zhang et al., 2024)

> 신경성 : 0.325
> 성실성 : -0.107
> 외향성 : 0.061

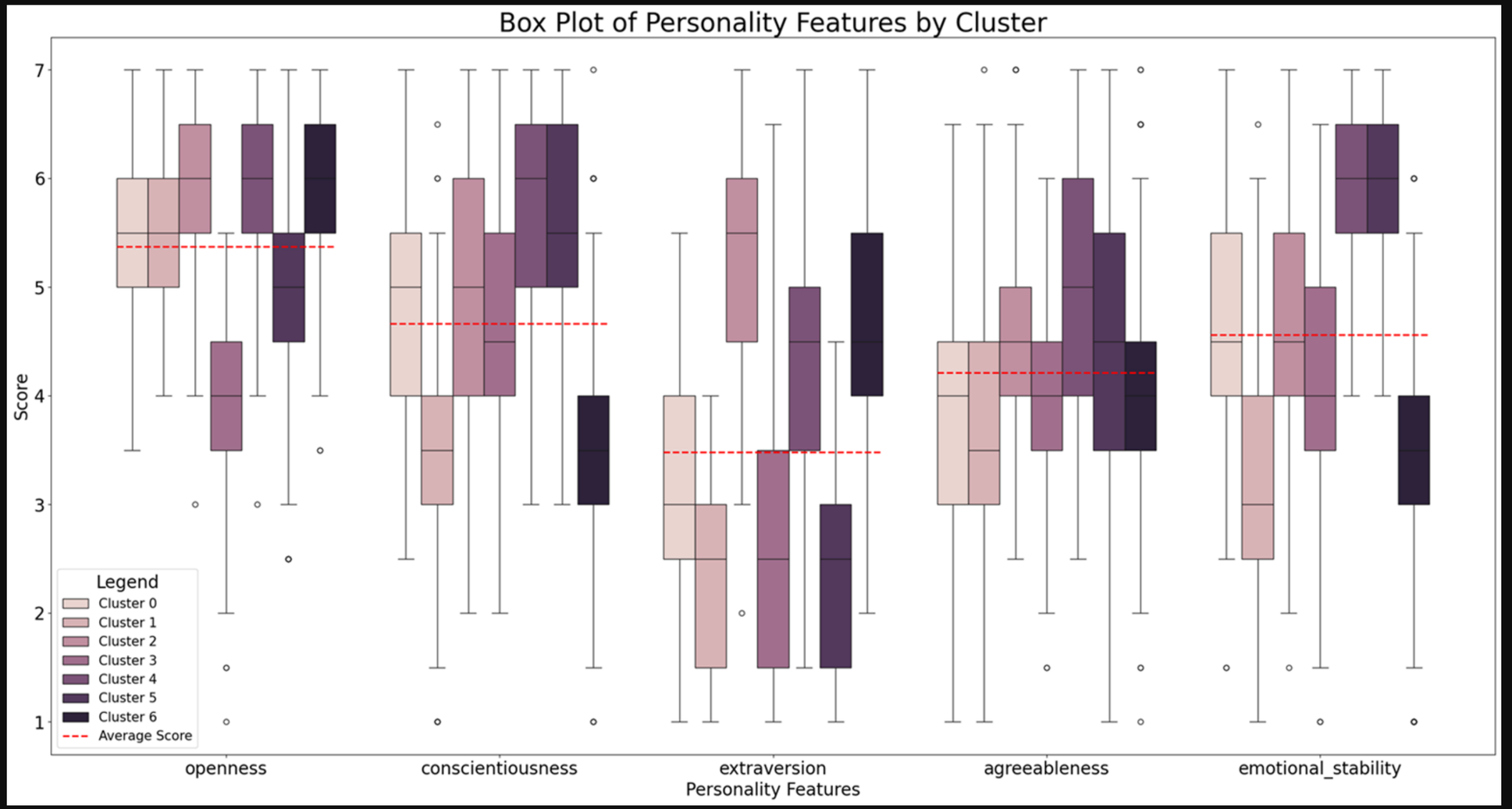
**FoMO가 높은 관객이 입소문의 주역!**

= FoMO 성향이 높은 관객을 타겟팅한 영화를 만들자

05

# 클러스터별 분석

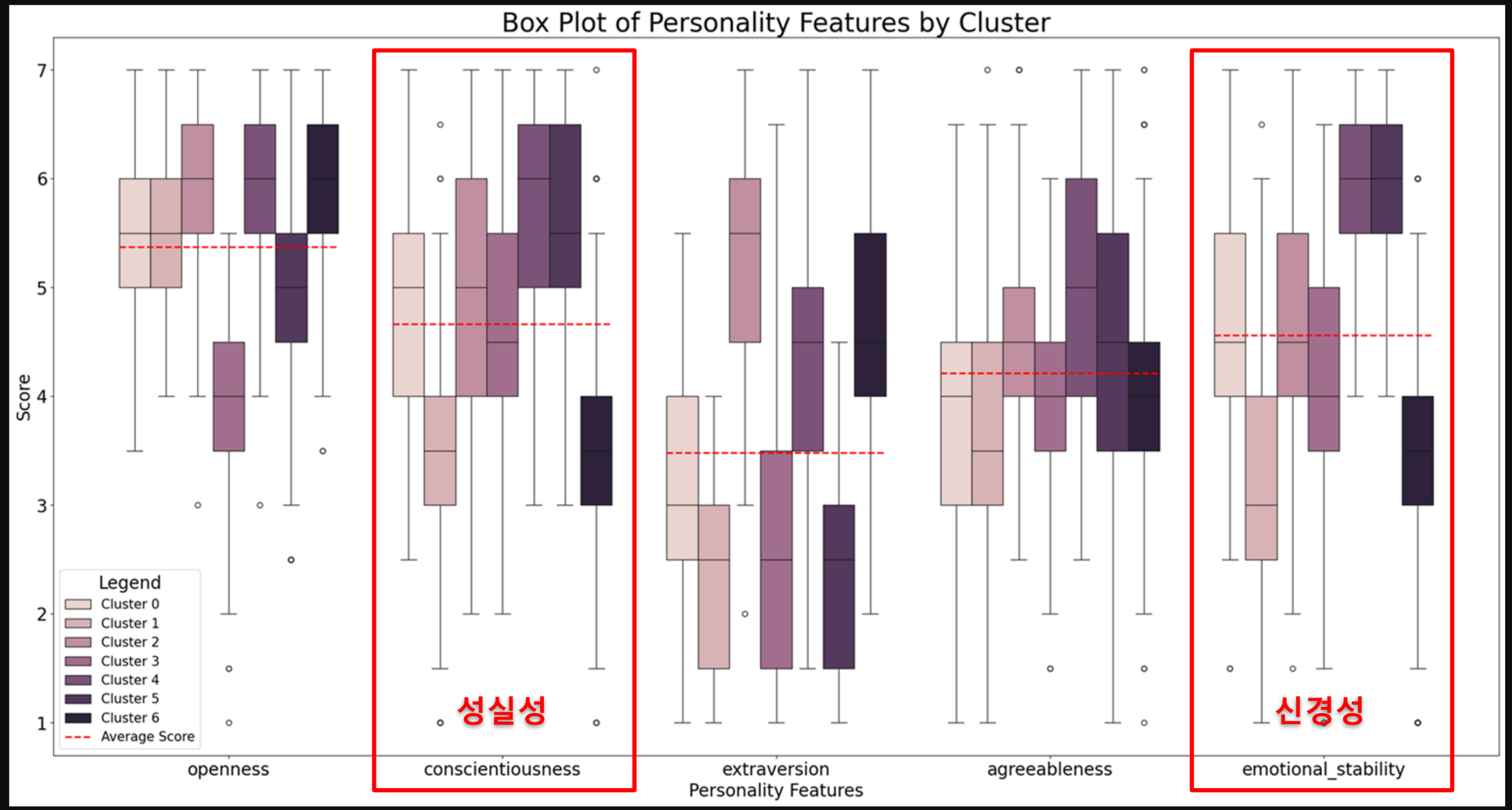
## 클러스터별 BIG5 score Boxplot 시각화



05

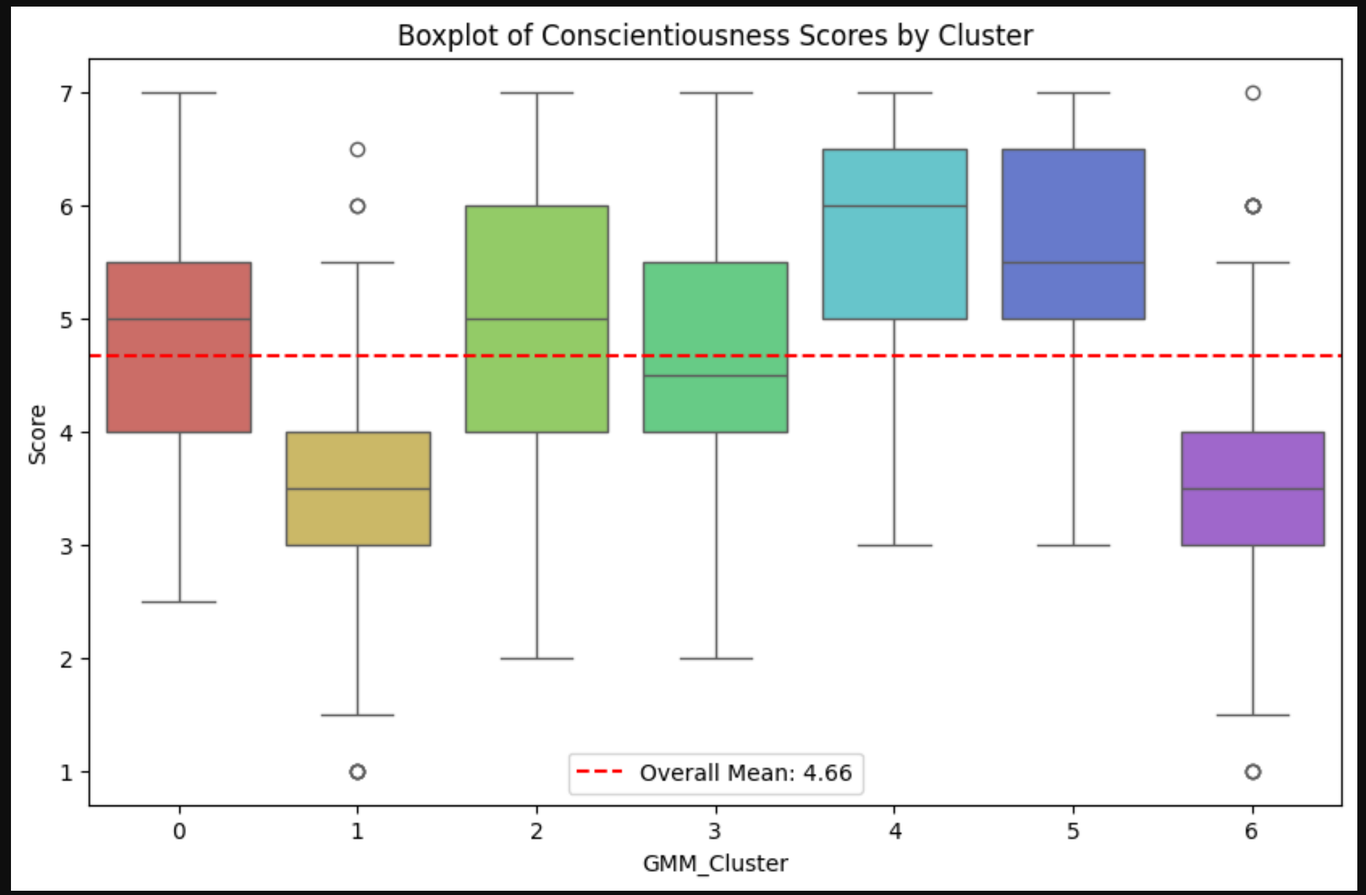
# 클러스터별 분석

## 클러스터별 BIG5 score Boxplot 시각화



# 05 클러스터별 분석

## 클러스터별 성실성(Conscientiousness) score

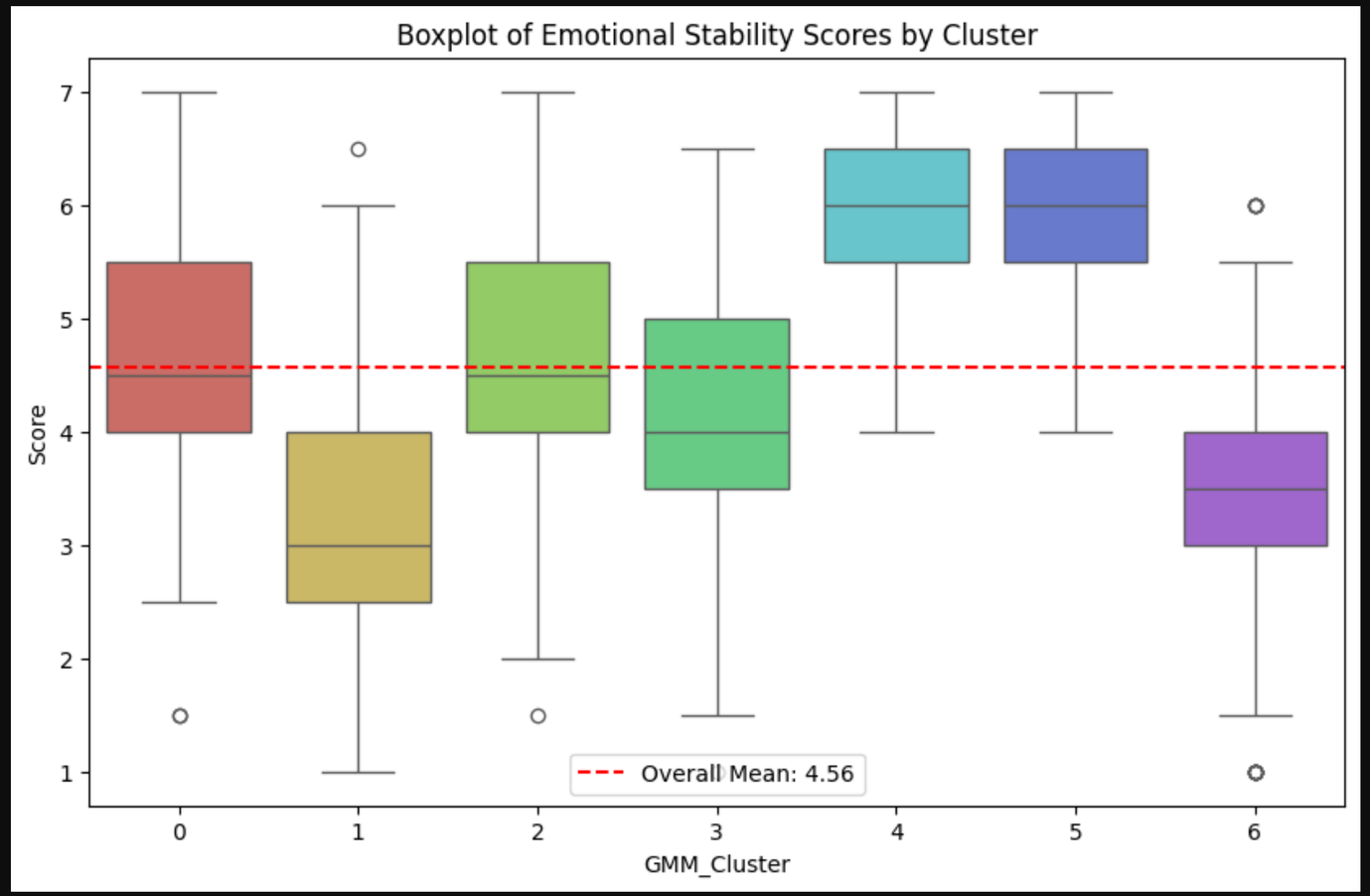



Multiple Comparison of Means - Tukey HSD, FWER=0.05

| group1 | group2 | meandiff | p-adj | lower | upper | reject |
|---|---|---|---|---|---|---|
| 0 | 1 | -1.438 | 0.0 | -1.6976 | -1.1784 | True |
| 0 | 2 | -0.0001 | 1.0 | -0.2565 | 0.2563 | False |
| 0 | 3 | -0.3975 | 0.0 | -0.6386 | -0.1563 | True |
| 0 | 4 | 0.8401 | 0.0 | 0.5886 | 1.0916 | True |
| 0 | 5 | 0.9142 | 0.0 | 0.6587 | 1.1697 | True |
| 0 | 6 | -1.3707 | 0.0 | -1.6233 | -1.1182 | True |
| 1 | 2 | 1.4379 | 0.0 | 1.161 | 1.7149 | True |
| 1 | 3 | 1.0406 | 0.0 | 0.7777 | 1.3035 | True |
| 1 | 4 | 2.2781 | 0.0 | 2.0058 | 2.5505 | True |
| 1 | 5 | 2.3522 | 0.0 | 2.0761 | 2.6283 | True |
| 1 | 6 | 0.0673 | 0.991 | -0.2061 | 0.3407 | False |
| 2 | 3 | -0.3974 | 0.0001 | -0.6571 | -0.1376 | True |
| 2 | 4 | 0.8402 | 0.0 | 0.5709 | 1.1095 | True |
| 2 | 5 | 0.9143 | 0.0 | 0.6412 | 1.1873 | True |
| 2 | 6 | -1.3706 | 0.0 | -1.641 | -1.1003 | True |
| 3 | 4 | 1.2375 | 0.0 | 0.9827 | 1.4924 | True |
| 3 | 5 | 1.3116 | 0.0 | 1.0528 | 1.5705 | True |
| 3 | 6 | -0.9733 | 0.0 | -1.2292 | -0.7173 | True |
| 4 | 5 | 0.0741 | 0.9835 | -0.1943 | 0.3425 | False |
| 4 | 6 | -2.2108 | 0.0 | -2.4765 | -1.9452 | True |
| 5 | 6 | -2.2849 | 0.0 | -2.5544 | -2.0154 | True |

# 05 클러스터별 분석

## 클러스터별 Emotional Stability score




Multiple Comparison of Means - Tukey HSD, FWER=0.05

| group1 | group2 | meandiff | p-adj | lower | upper | reject |
|---|---|---|---|---|---|---|
| 0 | 1 | -1.628 | 0.0 | -1.8774 | -1.3786 | True |
| 0 | 2 | -0.2293 | 0.0874 | -0.4755 | 0.017 | False |
| 0 | 3 | -0.6616 | 0.0 | -0.8933 | -0.43 | True |
| 0 | 4 | 1.1927 | 0.0 | 0.9512 | 1.4342 | True |
| 0 | 5 | 1.166 | 0.0 | 0.9206 | 1.4114 | True |
| 0 | 6 | -1.4058 | 0.0 | -1.6484 | -1.1632 | True |
| 1 | 2 | 1.3987 | 0.0 | 1.1327 | 1.6647 | True |
| 1 | 3 | 0.9664 | 0.0 | 0.7139 | 1.2189 | True |
| 1 | 4 | 2.8207 | 0.0 | 2.5591 | 3.0823 | True |
| 1 | 5 | 2.794 | 0.0 | 2.5288 | 3.0592 | True |
| 1 | 6 | 0.2222 | 0.1605 | -0.0404 | 0.4848 | False |
| 2 | 3 | -0.4324 | 0.0 | -0.6818 | -0.1829 | True |
| 2 | 4 | 1.4219 | 0.0 | 1.1633 | 1.6806 | True |
| 2 | 5 | 1.3952 | 0.0 | 1.1329 | 1.6575 | True |
| 2 | 6 | -1.1765 | 0.0 | -1.4362 | -0.9168 | True |
| 3 | 4 | 1.8543 | 0.0 | 1.6095 | 2.0991 | True |
| 3 | 5 | 1.8276 | 0.0 | 1.579 | 2.0762 | True |
| 3 | 6 | -0.7442 | 0.0 | -0.99 | -0.4983 | True |
| 4 | 5 | -0.0267 | 0.9999 | -0.2846 | 0.2311 | False |
| 4 | 6 | -2.5985 | 0.0 | -2.8536 | -2.3433 | True |
| 5 | 6 | -2.5717 | 0.0 | -2.8306 | -2.3129 | True |

06

# 리뷰 태그 추출

“ 클러스터 1과 6이 FoMO 성향이
높게 나타났다 ! (FoMO 클러스터) ”

“그렇다면 그들이 평균보다 높게
평가한 영화들의 **특징**은 어떻게
파악할 수 있을까 ? ”

# 06 리뷰 태그 추출

## ABSA (Aspect Based Sentiment Analysis)

Why? 단순한 키워드 추출이 아닌,
리뷰에서 **‘감정이 담긴 특징’**을 파악하기 위함

06

# 리뷰 태그 추출

## Step 1.
## Pre-trained Longformer

긴 리뷰에 대해 최대 512 토큰의 길이로 요약 진행
Evaluation : 0.78 By BERTScore

## Step 2.
## GPT API for ABSA

ABSA에서 GPT의 성능을 검증한, Simmering & Huoviala(2023)에 기반
Langchain 라이브러리에서 GPT 4o-mini를 활용해 진행

06

# 리뷰 분석

## Chi-Square 검정

**Why?**
**다른 클러스터에 비해, FoMO에서 유의하게 더 많이 언급된 태그 찾기 위함.**

**가정 : 긍정/부정 관계 없이, 특정 태그의 빈도가 유의하게 높다는 것은**
**클러스터 1과 6이 해당 태그의 중요성을 높게 평가함을 함축!**

| | 태그 | Z-통계량 | p-값 | total |
|---|---|---|---|---|
| 4 | horror | 2.930412 | 0.003385 | 16 |
| 8 | visuals | 3.329503 | 0.000870 | 15 |
| 25 | fantasy | 1.987892 | 0.046824 | 4 |

| | 태그 | Z-통계량 | p-값 | total |
|---|---|---|---|---|
| 3 | humor | 2.531504 | 0.011357 | 44 |
| 50 | franchise | 2.540121 | 0.011081 | 12 |
| 41 | screenplay | 2.518533 | 0.011784 | 7 |

# 06 리뷰 분석

| | 태그 | Z-통계량 | p-값 | total |
|---|---|---|---|---|
| 4 | horror | 2.930412 | 0.003385 | 16 |
| 8 | visuals | 3.329503 | 0.000870 | 15 |
| 25 | fantasy | 1.987892 | 0.046824 | 4 |

# FoMO 클러스터가 다른 클러스터들에 비해 **평점을 높게 준** 영화들에서, 빈도 수가 유의미하게 높은 태그들

| | 태그 | Z-통계량 | p-값 | total |
|---|---|---|---|---|
| 3 | humor | 2.531504 | 0.011357 | 44 |
| 50 | franchise | 2.540121 | 0.011081 | 12 |
| 41 | screenplay | 2.518533 | 0.011784 | 7 |

# FoMO 클러스터가 다른 클러스터들에 비해 **평점을 낮게 준** 영화들에서, 빈도 수가 유의미하게 높은 태그들

# 07 결론

## Our Goal

영화 제작사의 입장에서 수익을 극대화할 수 있는 영화, 즉 고객의 심리를 분석하여 '입소문' 낼 만한 영화를 찾자

## 클러스터링

User 클러스터링 진행한 후, FoMO(Fear of Missing Out) 특성을 가지는 클러스터를 선정

## 영화 리스트업

FoMO에 해당하는 클러스터 내 User들의 rating 정보를 통해 '입소문' 날 만한 영화를 리스트업

## 리뷰 태그 추출

앞선 리스트업 영화의 리뷰로부터 ABSA를 활용하여 태그를 추출 및 분석

**" 시각적 요소가 풍부한 공포 장르를 만들어, 입소문을 노려보자 "**

# 07 결론

FORBES > INNOVATION > GAMES

## The Best Horror Movie Of 2024 Just Broke A Box Office Record

**Paul Tassi** Senior Contributor ⓘ

*News and opinion about video games, television, movies and the internet.*

Follow

With $3 million in Thursday previews, that makes it a record for Neon, surpassing Sydney Sweeney's Immaculate from just a few months ago. The highest preview gross ever indicates it may be headed for a huge opening weekend thanks to good review and buzzy word of mouth. Estimates are around $20 million right now, which would double its supposedly $10 million budget immediately.

https://www.forbes.com/sites/paultassi/2024/07/13/the-best-horror-movie-of-2024-just-broke-a-box-office-record/

## 'A Quiet Place' Rockets to $50 Million Opening at Box Office

John Krasinski's horror film gets huge word of mouth, while "Blockers" continues Universal's success with R-rated comedies

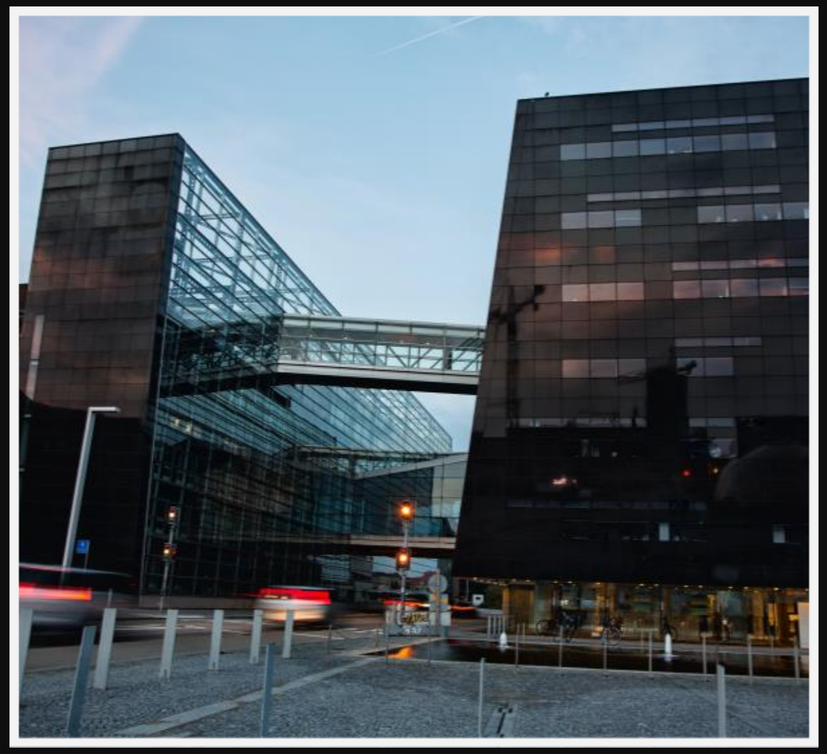
https://www.thewrap.com/quiet-place-rockets-50-million-opening-box-office/

07

# 한계점

## 한계점

1. 2018년에 만든 데이터셋 -> COVID 이후 영화의 트렌드는 반영 X
2. 해외 영화 위주 -> 국내 영화 제작에 바로 적용하기 어려움
3. BIG5 score 및 rating이 연속형 아닌, 이산형 데이터였다는 점



# 경청해주셔서

# 감사합니다!

| 11기 백두형 | 11기 김여원 | 12기 복지민 | 12기 이정우 |
| --- | --- | --- | --- |