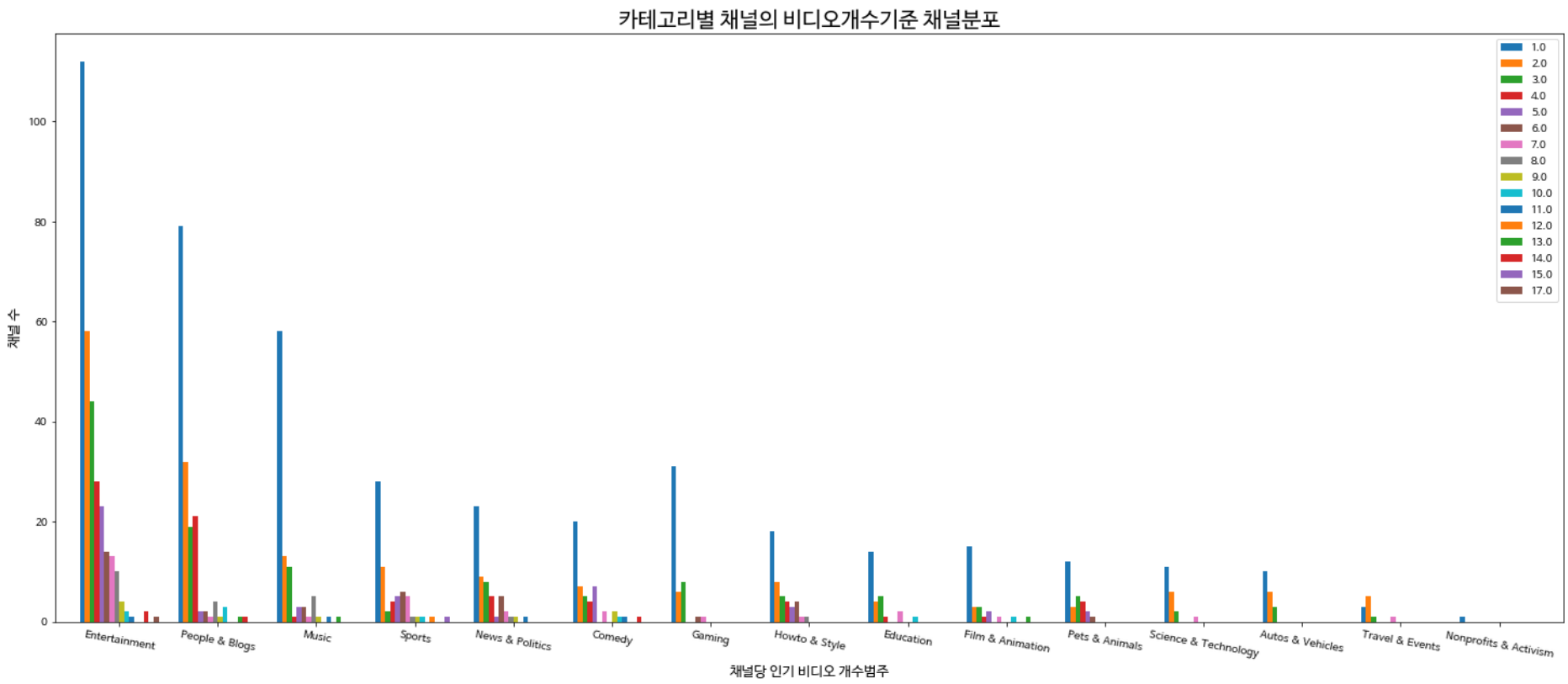
데이터 분석 및 시각화 [6팀_정태호]

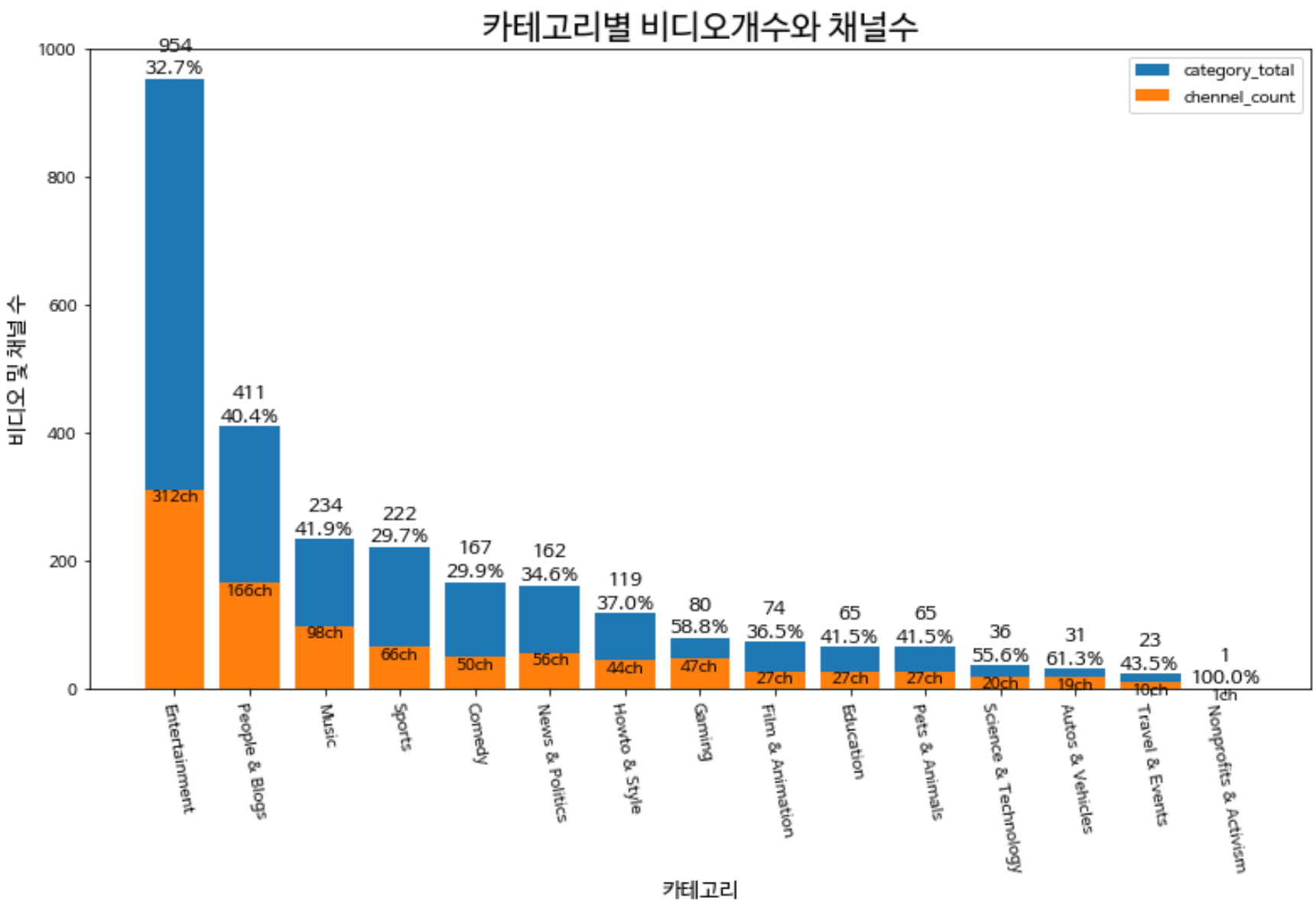
Q1. 데이터 타입별 시각화 (자유양식)

전체기간 카테고리 → 채널 → 비디오 개수



전체기간 동안 각 카테고리별 인기동영상을 보유한 개수에 따른 채널 분포를 시각화

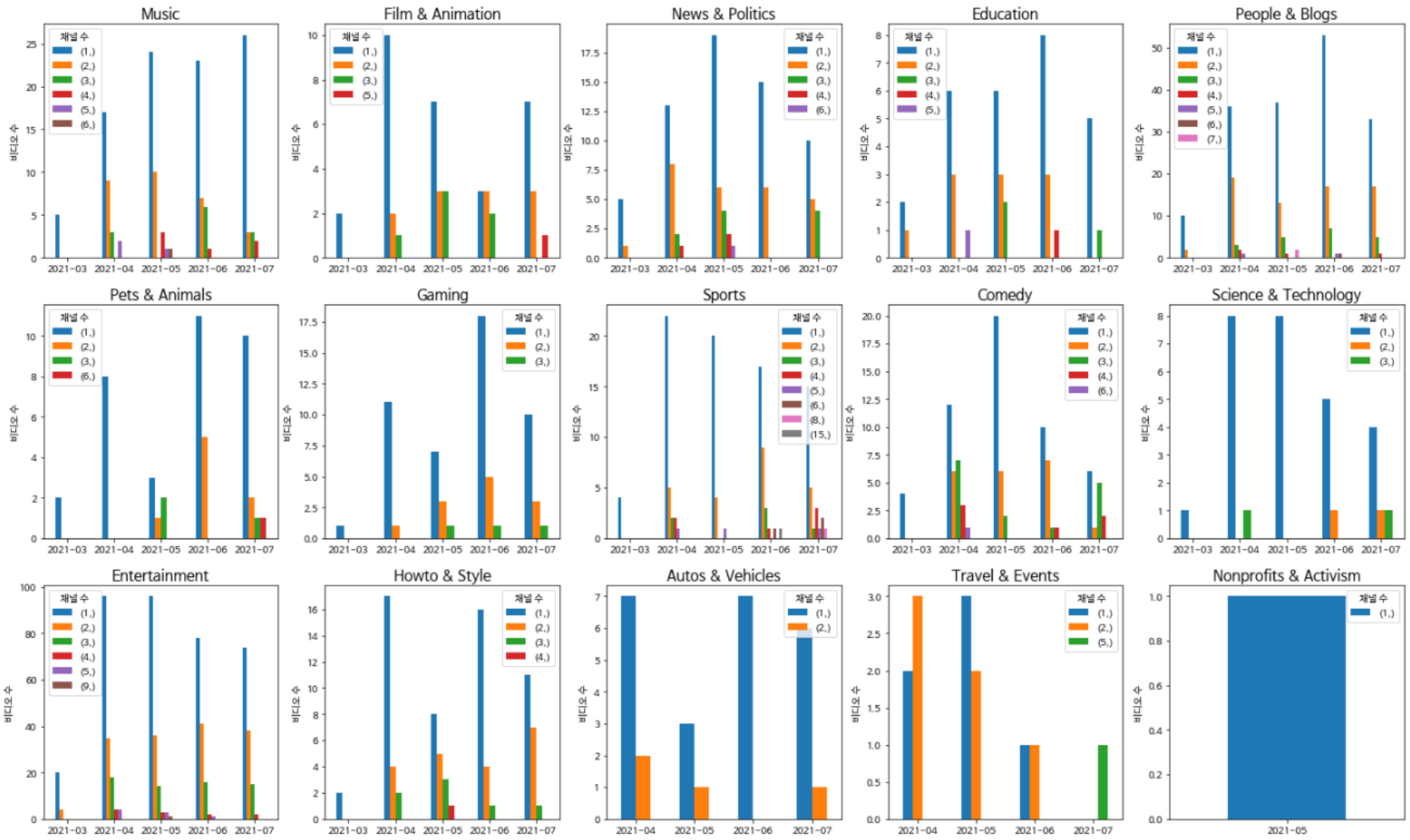
- 카테고리별 인기동영상을 보유한 채널들의 분포를 확인 가능하며 지속적인 인기동영상으로 선정되는 채널 수를 확인가능합니다.



전체기간 카테고리별 인기동영상 개수 및 채널수(비율)를 시각화

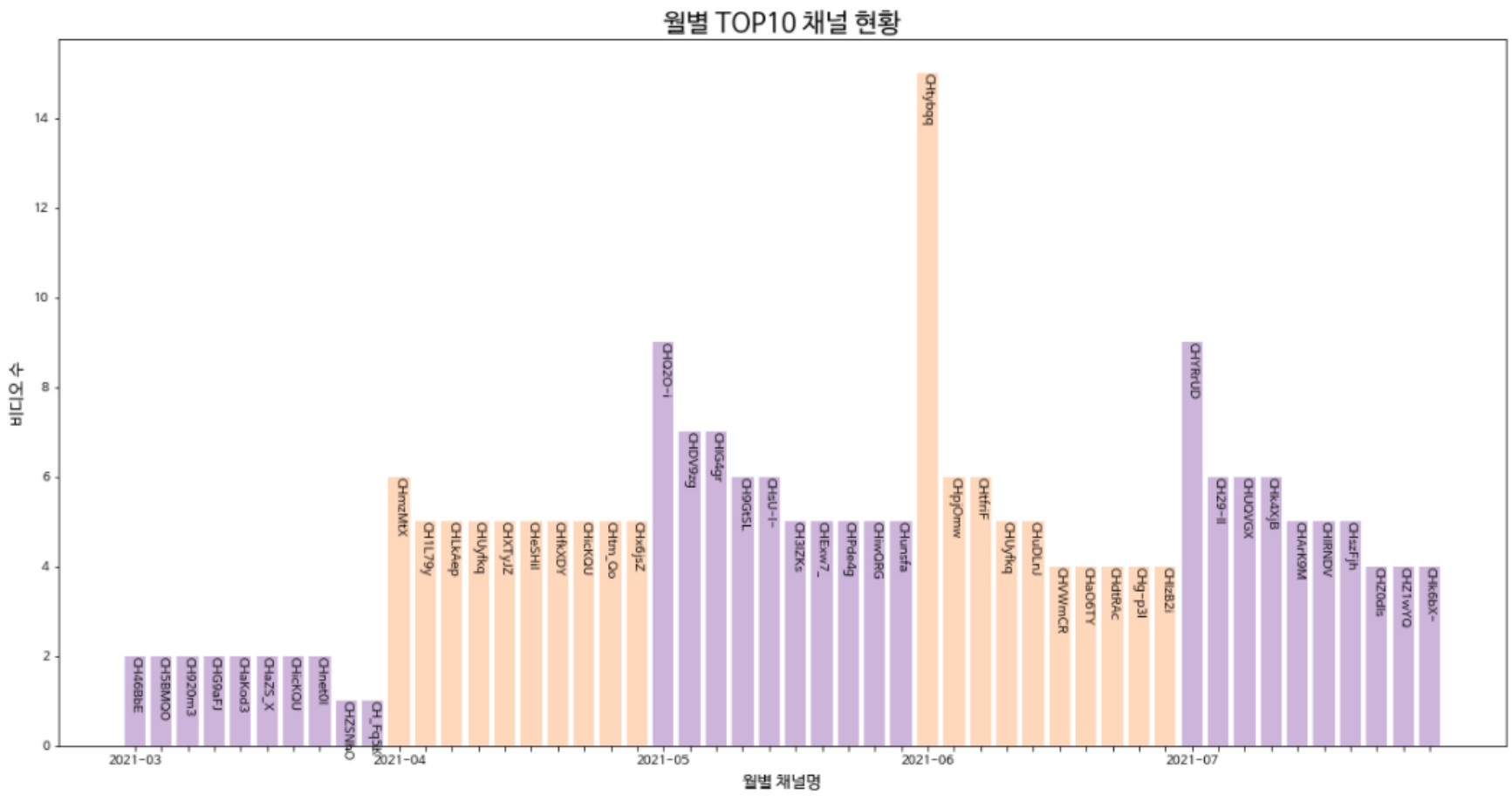
- 카테고리별 인기동영상 비율을 확인하며 비디오당 채널수의 비율을 통해 일부 채널의 독점도를 확인해볼 수 있습니다.

월별 카테고리 → 채널 → 비디오 개수



- 각 카테고리별 월간 인기비디오의 보유개수에 따른 채널분포를 변화를 세분화하여 확인할 수 있습니다.

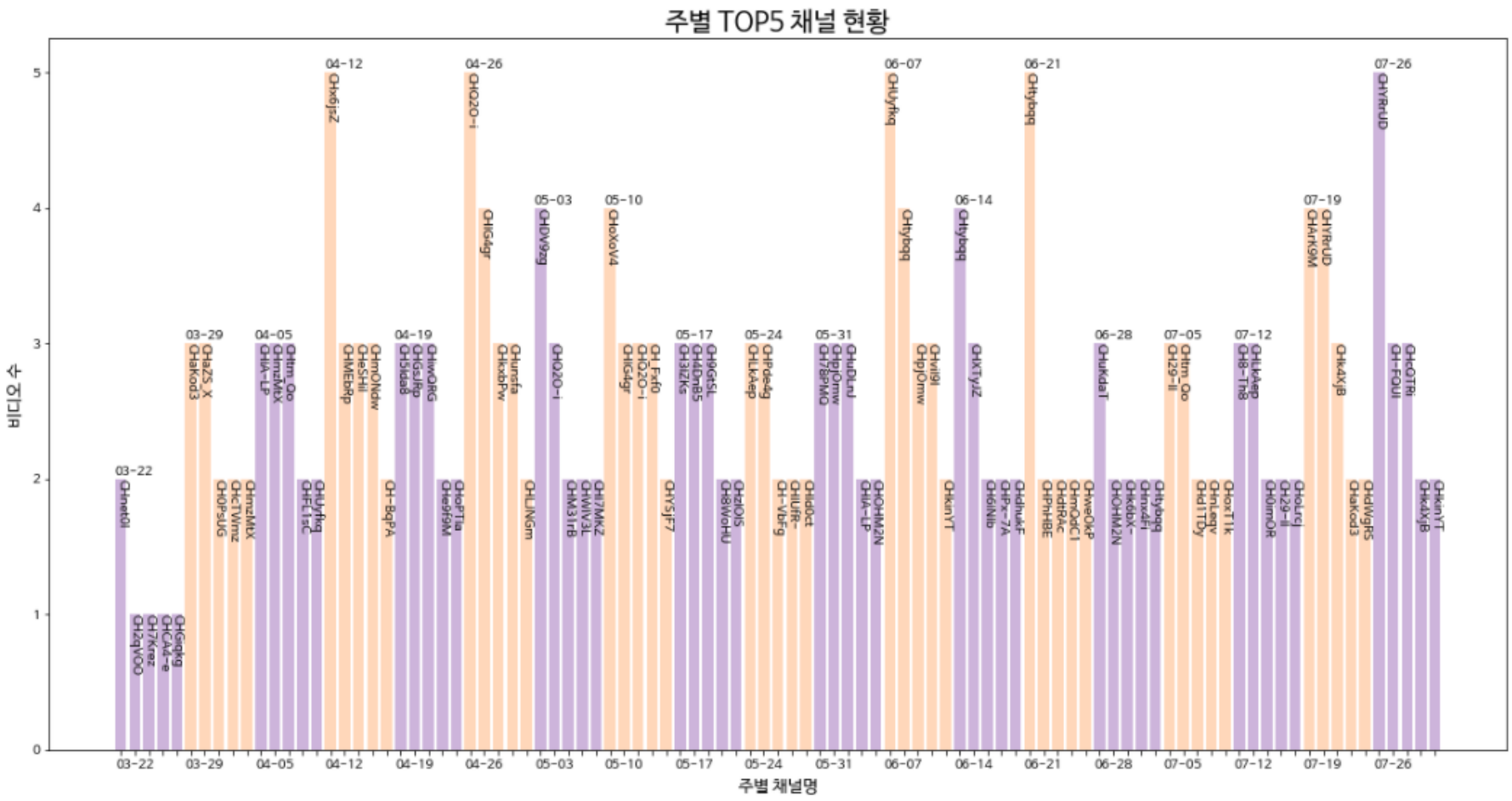
월별 TOP10 채널 (분류 기준은 비디오 개수)



월별 인기동영상을 가장 많이 보유한 채널 TOP 10을 시각화

- 월별 현황을 한번에 확인하는 것으로 특정 채널의 인기동영상 보유 추이와 변동 추이를 확인할 수 있습니다.

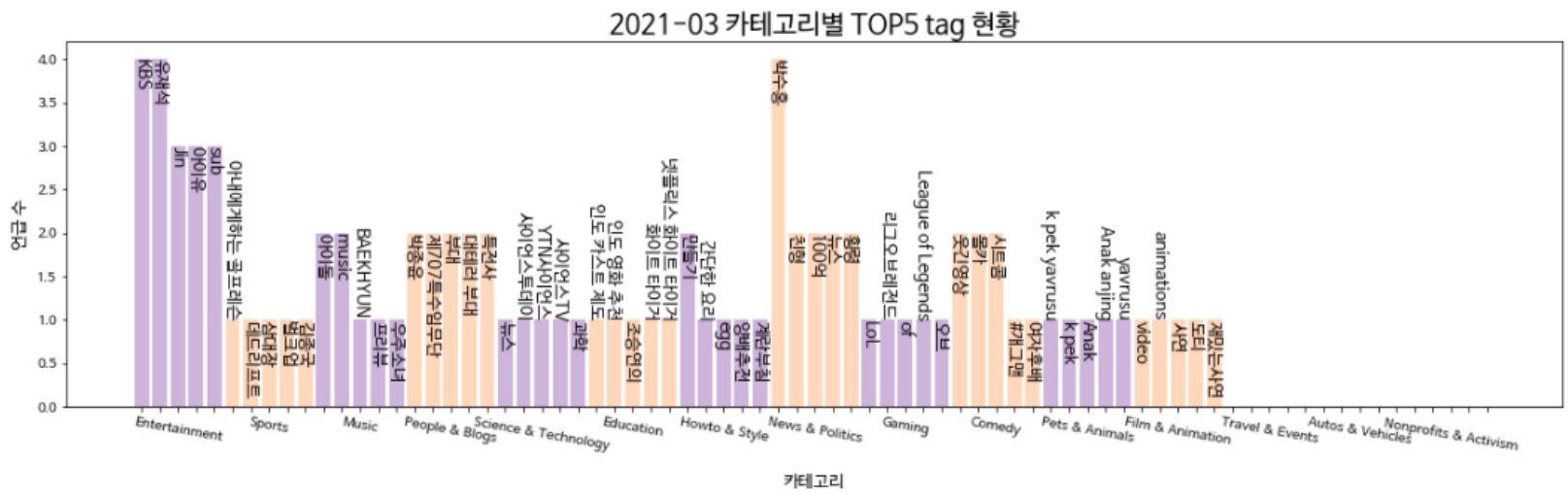
주별 TOP5 채널 (분류 기준은 비디오 개수)

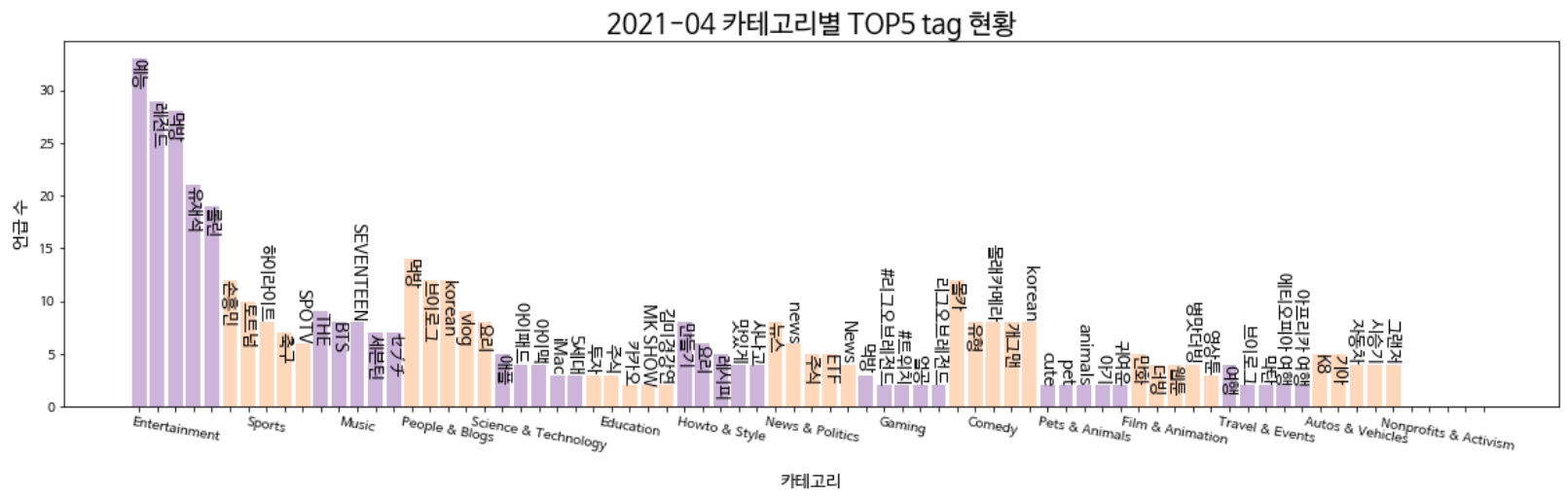


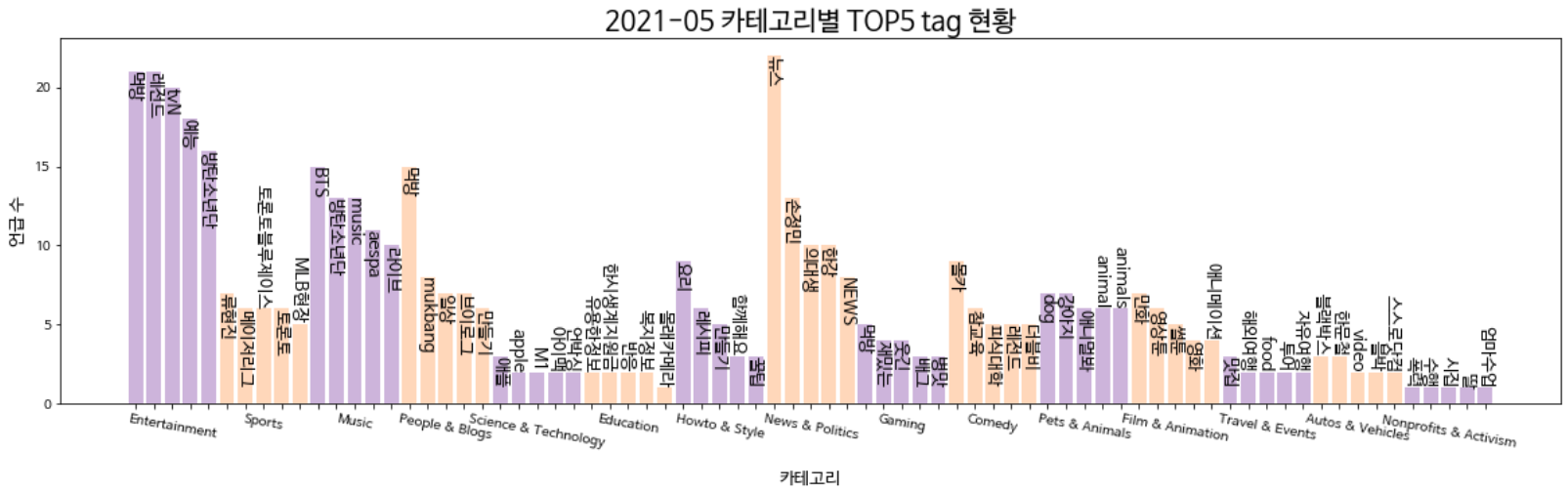
주별 인기동영상을 가장 많이 보유한 채널 TOP 5을 시각화

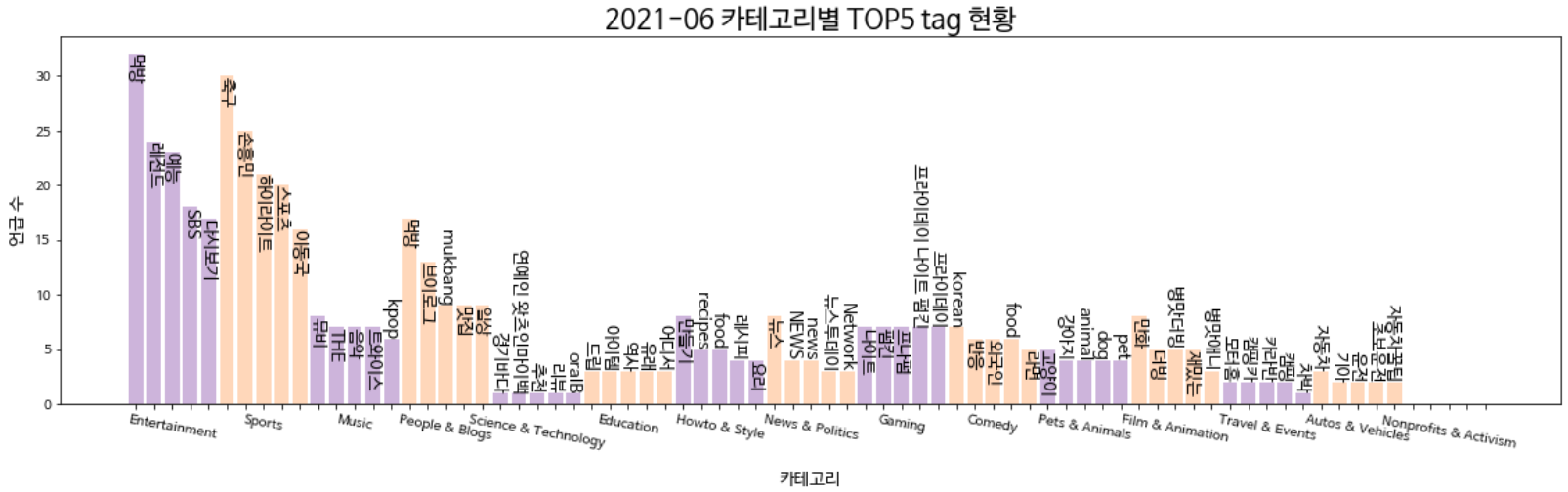
- 특정한 채널의 인기동영상 보유 개수의 변동을 주간으로 추적이 가능하며 순위권 등장 시기와 사라진 시기 등의 세부적인 흐름의 확인이 가능합니다.

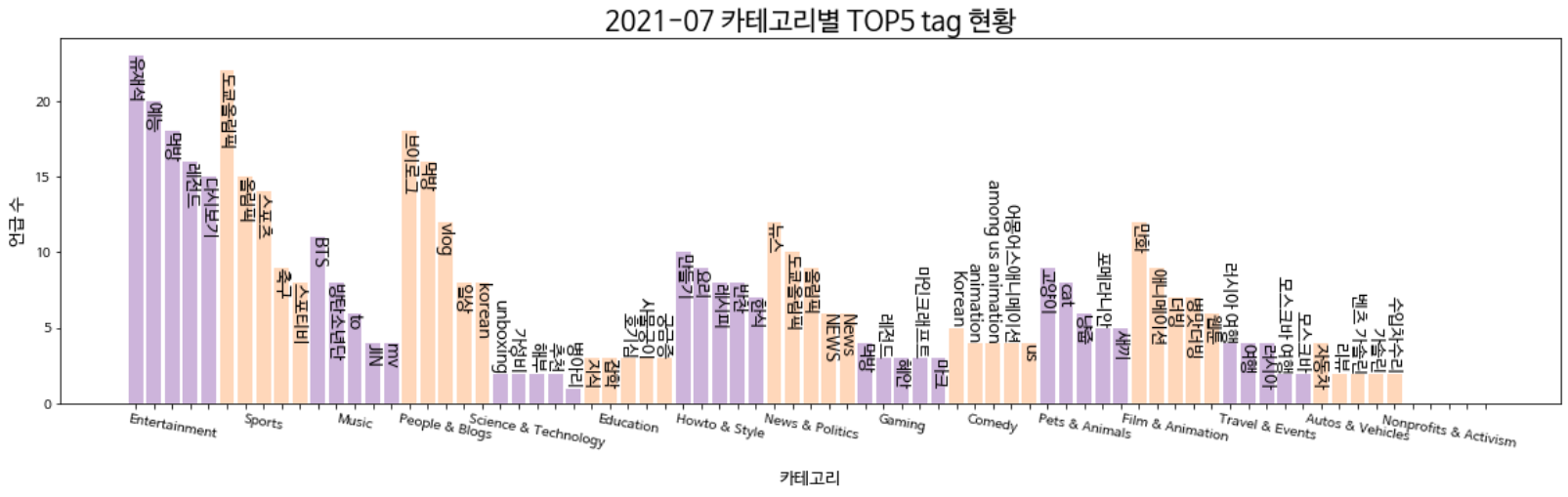
월별 카테고리별 태그 키워드 순위











- 월별 카테고리내 인기동영상에 tag된 단어들의 순위를 확인 가능하며, 이를 통해 사람들이 궁금해 하거나 흥미를 가지고 있는 키워드의 흐름을 확인할 수 있습니다.
- 또한 카테고리별 언급된 수를 통해 사람들의 시선이 어느 카테고리로 이동하는지 확인이 가능합니다.

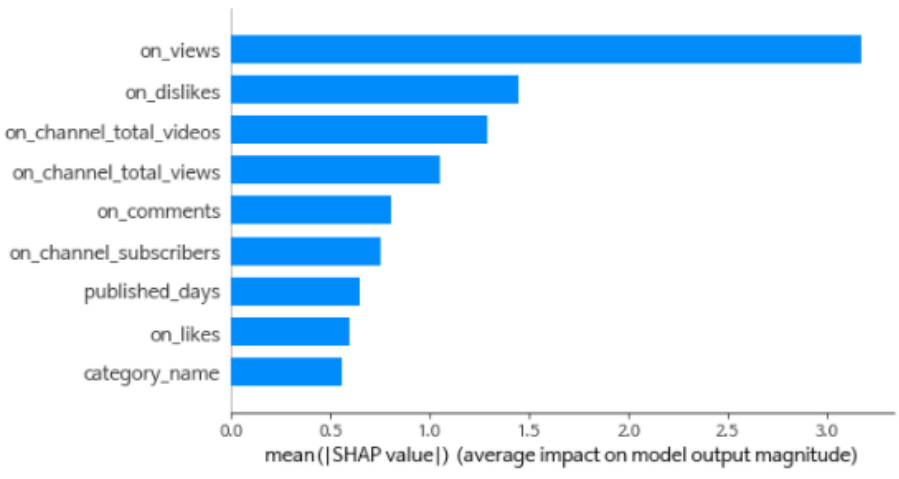
Q2. 각각의 비디오는 시청자의 호응도(engagement)를 판단할수 있는 객관적인 지표들이 있음

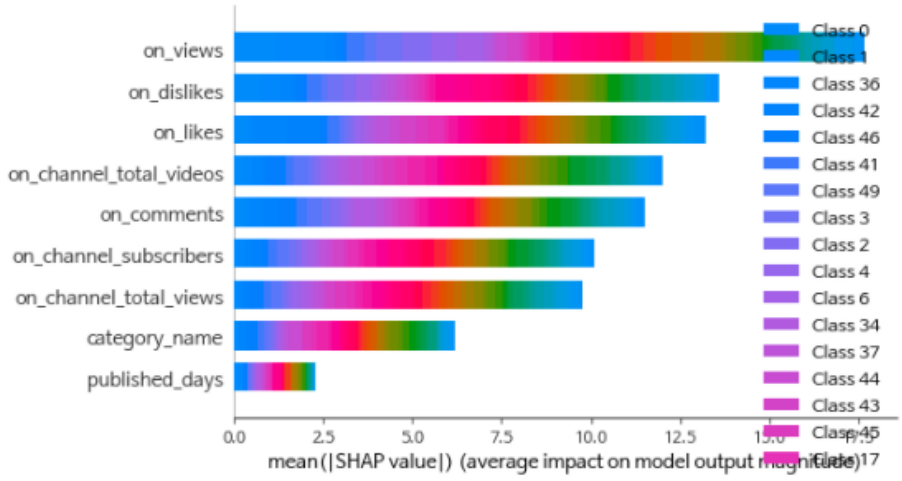
ex) views, likes, dislikes, comments,...

- 비디오를 인기 동영상 기준에 부합하도록 분류할수 있는 새로운 지표를 개발하고
- 이 지표를 사용하여 engagement 와 어떤 상관관계가 있는지 설명하시오.

	category_name	on_rank	on_views	on_likes	on_dislikes	on_comments	on_channel_subscribers	on_channel_total_views	on_channel_total_videos	published_days	engagement_rank
2021-03-25	1	1	8553414	825421	9096	47712	22200000	19311654452	14209	2	2.121344
2021-03-26	1	5	572113	9566	228	522	419000	31302574	76	2	6.655338
2021-03-26	2	19	275624	15272	35	2233	105000	5562950	19	2	19.233964
2021-03-26	3	10	1262201	69464	271	14729	1830000	380182956	573	2	10.246160
2021-03-26	1	28	1121118	31266	270	3499	1110000	126619182	86	2	23.558397
2021-03-27	1	4	599075	19047	415	6278	893000	196496420	132	2	6.938976
2021-03-27	4	13	1101249	3503	285	474	138000	173481297	337	2	12.530604
2021-03-27	2	28	106471	2433	22	780	109000	13299559	153	2	27.430343
2021-03-27	4	28	139114	2036	34	127	254000	48264340	283	3	27.633644
2021-03-27	5	16	445927	4930	296	565	555000	150769748	310	2	16.406825

- 인기비디오 지정 조건은 동영상을 시청하는 시청자의 다양성(타 카테고리 중심 시청자)과 채널을 최근 평균 조회수 및 24시간 이내의 조회수 증가 속도와 동영상 자체의 심의 준수 정도에 따른 실시간 시계열 데이터를 기반으로 선정되는 것을 확인 하였습니다.
- 그렇기 때문에 인기동영상 선정 시점과 탈락시점의 단편적인 데이터를 기준으로 해당 동영상의 호응도를 판단하기 위한 수식적인 정립에 어려움을 느꼈습니다.
- 제한적인 데이터를 통한 충분한 정확도 달성과 시간에 흐름에 따른 기준의 정량적 변화를 감안 할 수 있도록 XGBoost의 머신러닝 모델을 사용하였습니다.
- 2개의 서로다른 결과산정 방식을 가지는 모델을 통해 해당 동영상이 인기동영상의 몇 순위에 랭크될 정도의 호응도(engagement)를 지니는지 계산할 수 있도록 하였습니다.





- 결과 on_views의 값이 가장 큰 영향을 끼치는 것을 확인할 수 있으며 on_likes의 값보다 on_dislikes의 값이 영상의 engagement_rank의 계산에 큰 영향을 끼치는 것을 확인 할 수 있습니다.
- 이를 바탕으로 시간에 따른 지속적인 학습을 통해 일정수준의 정확도를 유지하는 조건 하에서 인기동영상으로 선정되기 위해 신경써야할 지수의 선정에 도움을 줄 수 있을 것으로 판단할 수 있습니다.