



# YOUTUBE Analysis Report

데이터 타입별 시각화와 새로운 지표의 개발

( URL ) [https://seyeon236580.notion.site/junior\\_-\\_pretest-dc0929f26d9149d793b74382eaa59798](https://seyeon236580.notion.site/junior_-_pretest-dc0929f26d9149d793b74382eaa59798)

작성자 : 정 세 연

# Q1. 데이터 타입별 시각화

```
# Numeric data에 대한 통계 확인 (datetime형 포함)
df.describe(datetime_is_numeric=True)
```

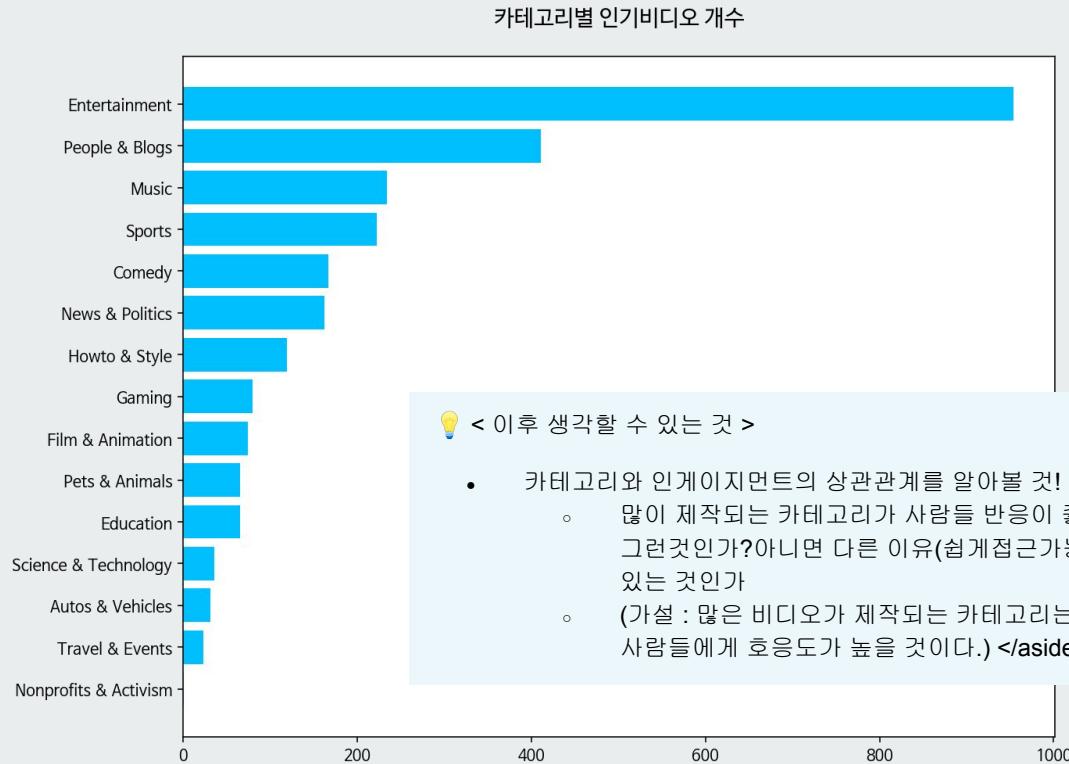
	published_date	on_trending_date	off_trending_date	on_rank	off_rank
count	2644	2644	2644	2644.000000	2644.000000
mean	2021-05-27 14:13:58.729198336	2021-05-29 15:45:28.593040896	2021-05-30 23:40:23.600605184	20.234493	35.795386
min	2021-03-25 00:00:00	2021-03-27 00:00:00	2021-04-01 00:00:00	1.000000	1.000000
25%	2021-04-27 00:00:00	2021-04-29 00:00:00	2021-04-30 00:00:00	9.000000	28.000000
50%	2021-05-28 00:00:00	2021-05-30 00:00:00	2021-05-31 00:00:00	19.000000	37.000000
75%	2021-06-27 00:00:00	2021-06-29 00:00:00	2021-07-01 00:00:00	30.000000	45.000000
max	2021-07-29 00:00:00	2021-07-31 00:00:00	2021-07-31 00:00:00	50.000000	50.000000
std	NaN	NaN	NaN	12.833115	10.376753

3월 말부터 7월 말까지의 데이터이며,  
1위부터 50위까지의 인기동영상  
데이터입니다.

먼저 어떤 카테고리에서 비디오가 많이  
제작되는지 살펴보겠습니다.

Entertainment가 압도적으로 비디오가 많이 제작되는 카테고리임을 알 수 있습니다.

	category_name	count
0	Nonprofits & Activism	1
1	Travel & Events	23
2	Autos & Vehicles	31
3	Science & Technology	36
4	Education	65
5	Pets & Animals	65
6	Film & Animation	74
7	Gaming	80
8	Howto & Style	119
9	News & Politics	162
10	Comedy	167
11	Sports	222
12	Music	234
13	People & Blogs	411
14	Entertainment	954



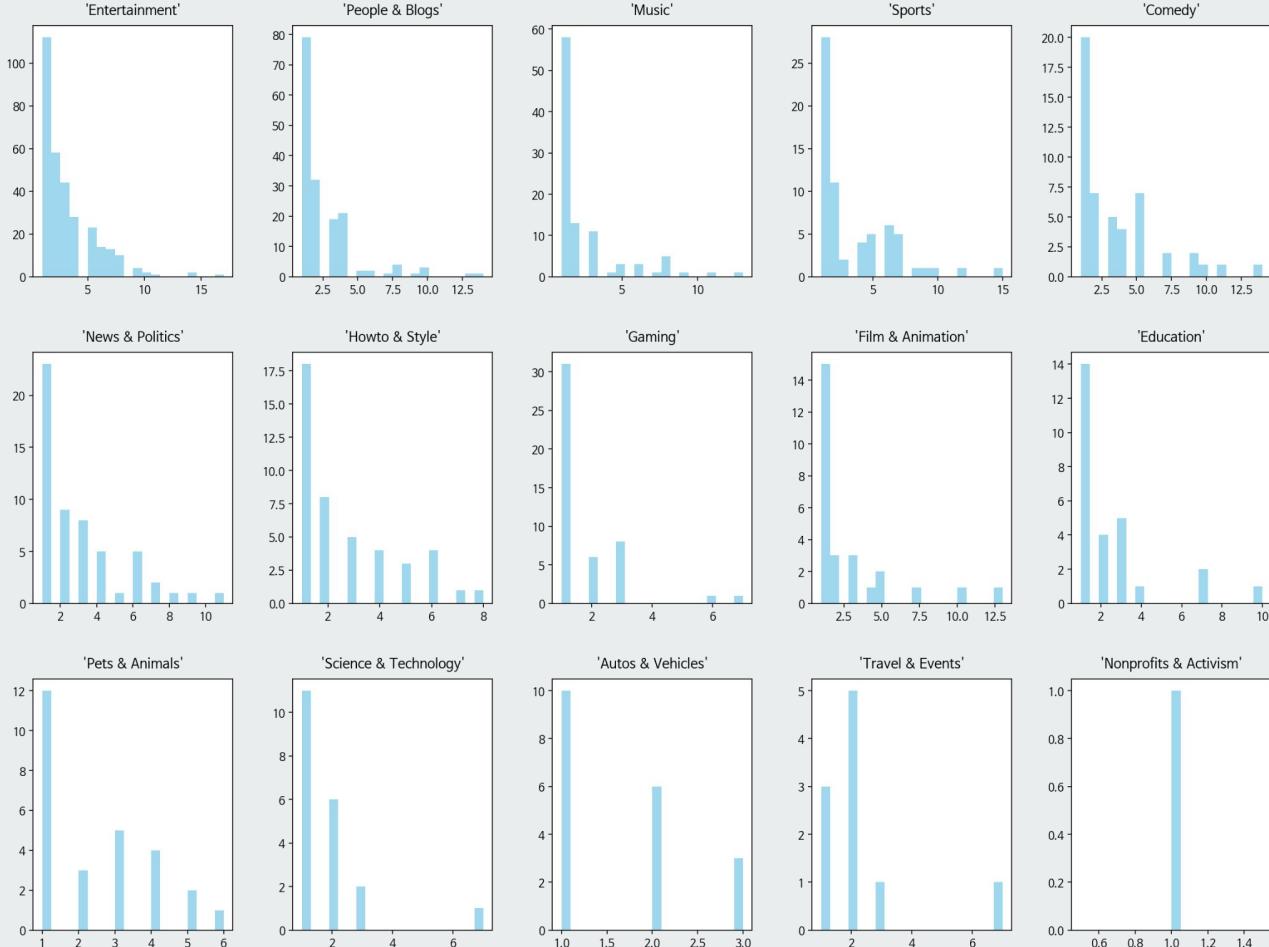


카테고리 개수는 총 15개입니다.

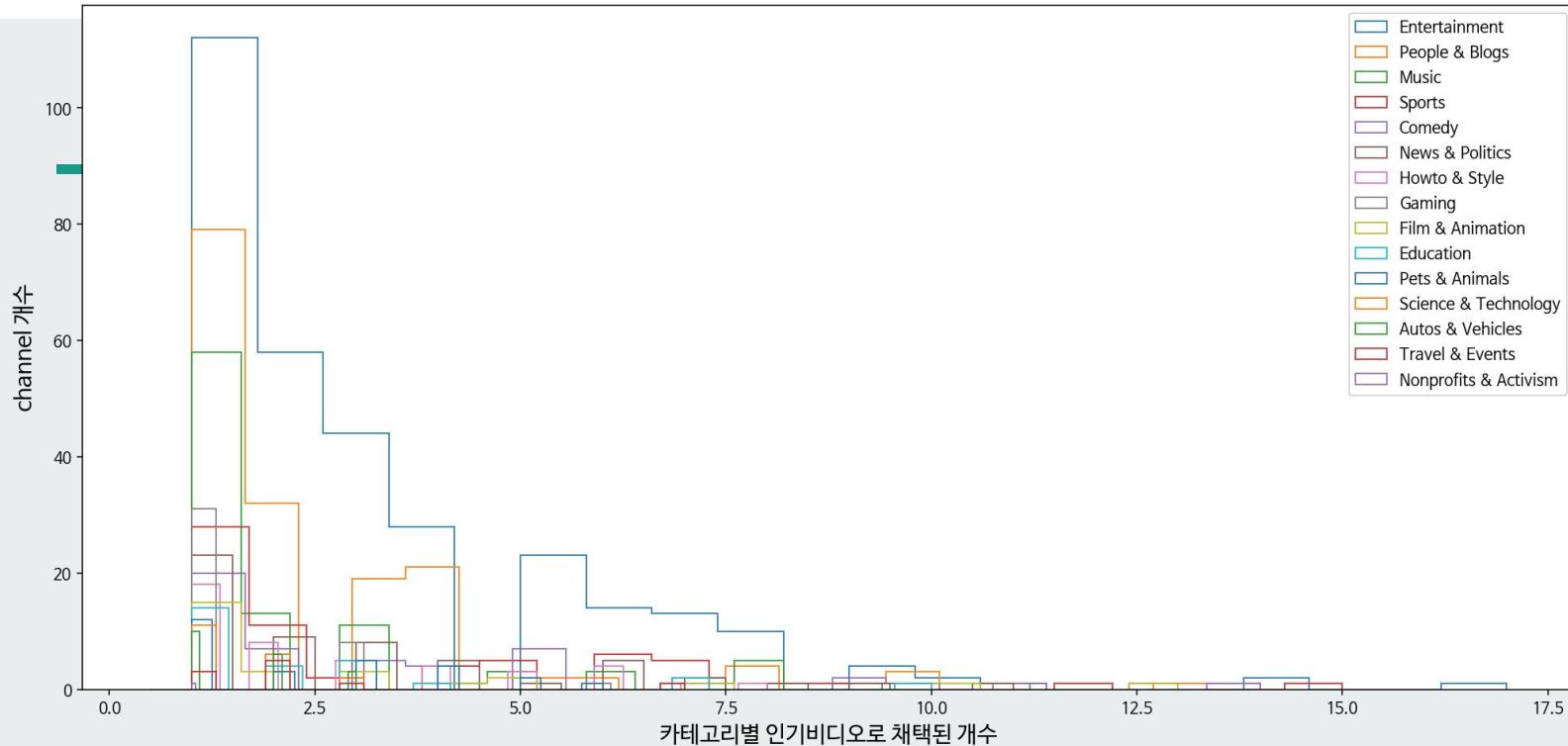
모든 카테고리에 대한 공통점은 ,

한 카테고리의 인기비디오로 선정된  
채널이 그 카테고리에 대해 1개의  
영상만 채택된 경우가 압도적으로  
많다는 것입니다.

히스토그램으로 빈도수를 표현해  
보겠습니다.



## 채널별 특정 카테고리에 대한 인기비디오 채택 개수의 경향성



모든 카테고리에서 한 카테고리에서 대부분 한 채널당 5개 이내의 영상이 인기영상으로 채택된 경우가 많은 것을 알 수 있습니다.

- 그렇다면 이 채널들의 동영상 총 제작개수는 몇 개일까요?

## 1. 전체기간 카테고리->채널->비디오 개수

---

- 인기비디오로 채택된 채널들의 전체 비디오 개수(카테고리별로 나누어 살펴보기)

인기비디오에 채택된 채널들이 제작한 전체 영상 개수에 대해 알아봅시다.

전체 비디오 개수와 관련된 컬럼은 `on_channel_total_videos`, `off_channel_total_videos`입니다.

- `on_channel_total_videos` : 인기 동영상에서 처음 기록된 채널의 비디오 개수
- `off_channel_total_videos` : 인기 동영상에서 사라지기 전 기록된 채널의 비디오 개수

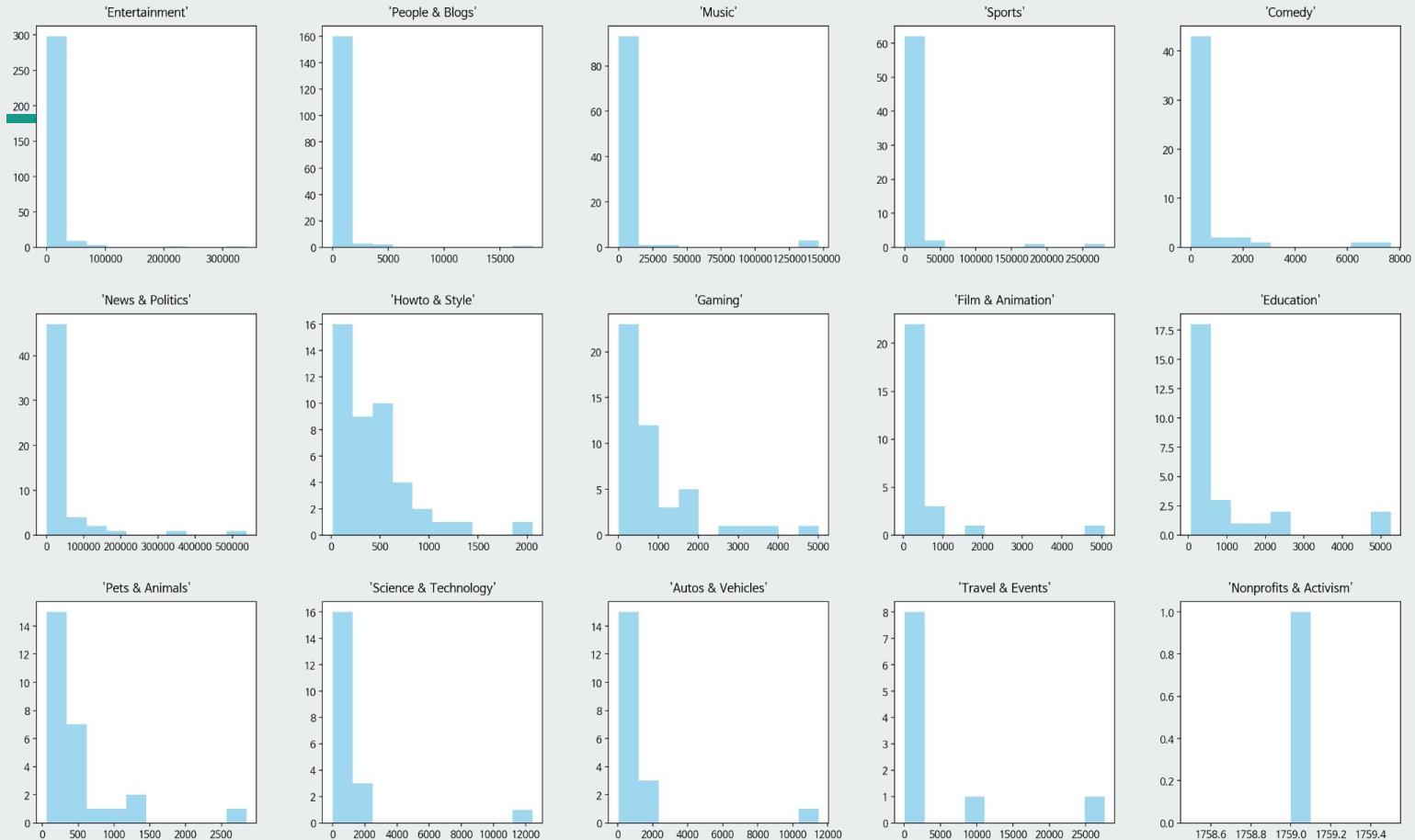
인기비디오에 채택된 채널들의 영상개수에 대해 알아보는 것이므로, `on_channel_total_videos`을 기준으로 하겠습니다.

각 카테고리로 데이터프레임을 분류하였을 때, 인기비디오로 선정된 `channel_id`에는 중복값이 있습니다. 예를 들어, Entertainment 인기비디오로 채택된 `channel`은 총 312개이지만, 한 채널당 여러개의 Entertainment 영상이 인기비디오로 채택되었기 때문에, Entertainment 카테고리에는 선정된 비디오 개수대로 `channel_id`가 여러 번 나올 수 있습니다.

`on_channel_total_videos` 컬럼은 각 비디오가 인기비디오로 채택된 시점에서 해당 채널의 총 비디오 개수를 측정한 것이므로, `channel_id`의 중복값에 대한 처리가 필요합니다.

현재 풀고자 하는 task는 채널별 비디오 개수가 인기동영상 기준(시청자 호응도)과 어떤 연관성이 있는지이므로, 각 채널별로 비디오 개수의 평균값을 구해서 비교하도록 하겠습니다. 최소값이나 최대값보다 평균값이 객관적인 지표일 것입니다.

카테고리-채널별 비디오 총보유량 경향 (x=비디오 총보유량, y=channel 개수)

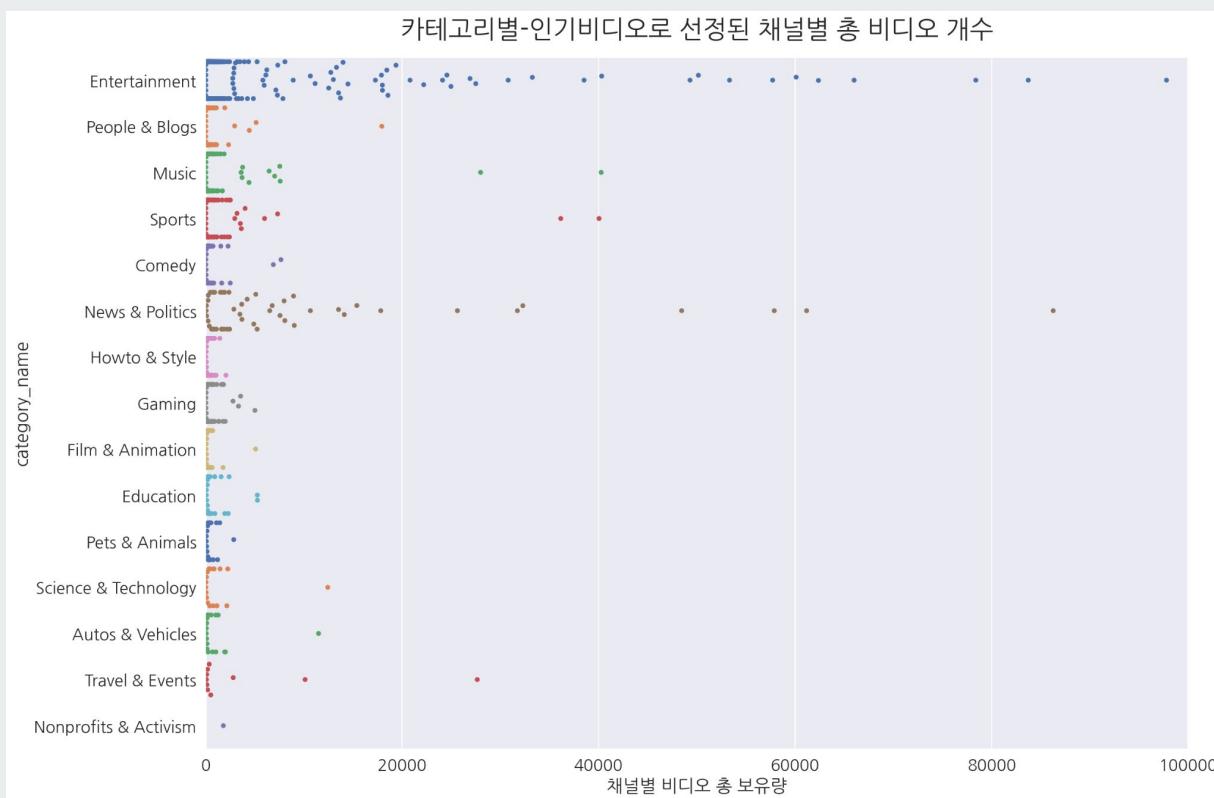


15개의 카테고리를 한 눈에 비교할 수 있도록 swarmplot과 boxplot을 이용하여 카테고리별 채널의 총 비디오 개수와 분위수를 시각화 해봅니다.

아래 분포그래프의 각 점들은 하나의 채널을 나타냅니다.

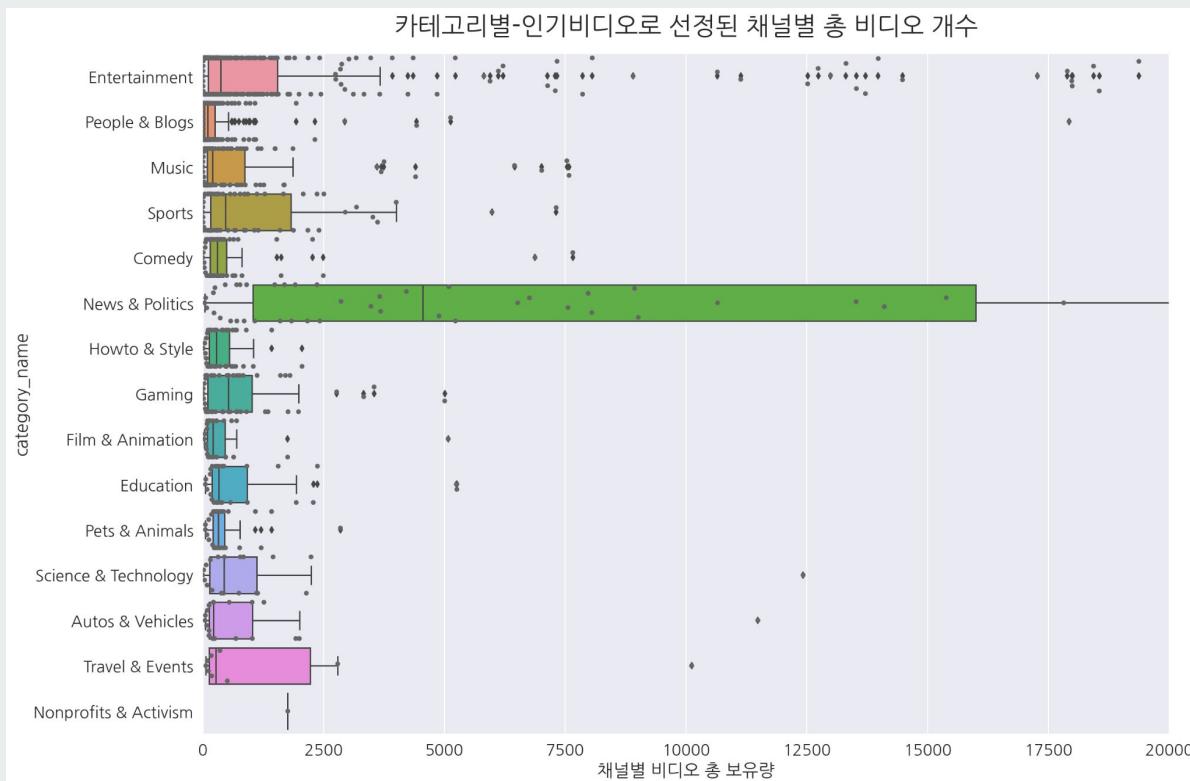


아웃라이어를 제외하고, 좀 더 많은 분포가 되어있는 1~100,000 비디오 총 보유량 지점을 확대해 보겠습니다.



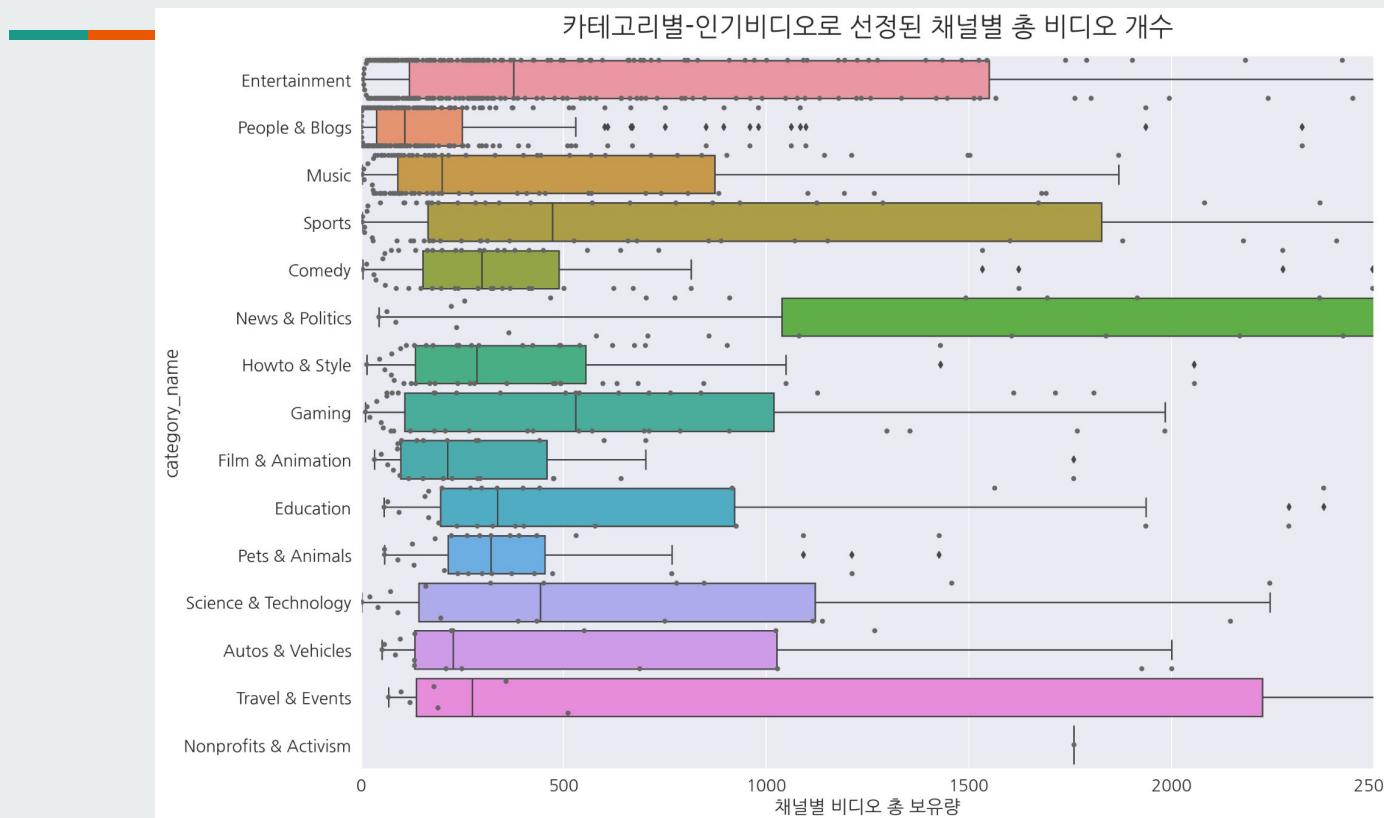
Entertainment와 News&Politics 카테고리의 채널 이외에는 비디오 총 보유량이 20,000 개 이상 되지는 않습니다.

X축의 범위를 20,000 까지 축소해 보겠습니다.



4분위를 표현가능한 boxplot까지 함께 표기해 보았습니다. News&Politics 카테고리를 제외하면 대부분 채널별 비디오 총보유량이 2500개 이하입니다.

이 부분을 다시 확대해서 보겠습니다.



category\_df\_concat

	channel_id	channel_videos_mean	category_name
117	CHNI-TU	5	Entertainment
32	CH4JAFO	7	Entertainment
79	CHF8W68	7	Entertainment
96	CHIY5oU	8	Entertainment
140	CHTmrCB	9	Entertainment
...	...	...	...
5	CHPDdt9	511	Travel & Events
9	CHsLoTw	2795	Travel & Events
3	CHFw4M1	10116	Travel & Events
2	CHFCtZJ	27639	Travel & Events
0	CHSsWdU	1759	Nonprofits & Activism

970 rows × 3 columns

카테고리별로 채널별 비디오 총보유량을 봤을 때,

### Entertainment와 News & Politics

카테고리에는 아웃라이어가 많은 편이고  
채널별 비디오 보유량 수가 2500개 이상으로  
매우 많은 경우가 많습니다.

그러나 이외의 카테고리에서는 대부분  
2500개 이하의 보유량을 가진편이며,  
중앙값으로는 대체로 300~400개 전후에  
머무르고 있습니다.(아웃라이어 수가  
많기때문에 평균값이 무의미)

시각화한 부분을 구체적인 수치로 좀 더  
정확하게 우측과 같이 알아볼 수 있습니다.

인기비디오로 채택된 채널들을 카테고리별로 나누어 총 비디오 개수를 고려해 보았습니다.

최소 개수는 **Comedy** 카테고리에서 4개의 영상만을 제작한 채널도 인기비디오로 채택될 수 있었고, 최대로는 **News&politics** 카테고리에서 538034개의 영상을 제작한 채널이 채택되었습니다.

채널별 제작비디오의 총 개수가 인기비디오 선정과 큰 연관이 있다고 보기는 어렵습니다.

그러나 카테고리별로 제 1사분위(Q1)값이 110내외, 중앙값이 300~400개 내외에 있다는 점을 고려했을 때,

(즉, 인기비디오로 선정된 채널의 75% 가량이 약 110개 이상의 비디오를 제작함.)

보다 안정적인 인기비디오 채택을 위해서 고려할 수 있는 채널별 총 비디오 개수는 약 100개 이상이라고 할 수 있습니다.

## 2. 월별 카테고리->채널->비디오 개수

---

그렇다면 위에서 살펴봤던 채널들은 얼마나 열심히 비디오를 제작하고 있을까요? 채널들이 얼마나 많은 비디오를 제작하고 있을지 월별 비디오 보유량을 알아보도록 하겠습니다.

이 데이터는 3월부터 7월까지 제작된 인기비디오에 대해 다루고 있습니다. 따라서 5개월에 대해 개월별로 나누어 카테고리 별로 각 채널의 비디오 보유량에 대해 알아보겠습니다.

날짜를 나타내는 컬럼은 `published_date`, `on_trending_date`, `off_trending_date` 3가지가 있지만, 기준으로 `published_date`('영상이 유튜브에 업로드된 날짜')를 선정하겠습니다. 왜냐하면, 지금부터 알아볼 것은 인기비디오로 채택된 채널들이 월평균 몇개의 비디오를 제작하느냐이기 때문입니다.

데이터셋은 총 2644행으로 구성되어 있고, 분석해야 할 채널의 총 개수는 940개입니다.

위의 문제해결과 마찬가지의 이유로, 중복된 `channel_id`에 대한 비디오 개수에 대해 평균값을 취하도록 하겠습니다.

```

for i in range(len(category)):
    print(f'{category[i]} : {category_df3[i].shape[0]}개')

```

3월 카테고리별 channel\_id 개수

-----

Entertainment : 24개  
 People & Blogs : 12개  
 Music : 5개  
 Sports : 4개  
 Comedy : 4개  
 News & Politics : 6개  
 Howto & Style : 2개  
 Gaming : 12개  
 Film & Animation : 2개  
 Education : 3개  
 Pets & Animals : 2개  
 Science & Technology : 1개  
 Autos & Vehicles : 0개  
 Travel & Events : 0개  
 Nonprofits & Activism : 0개

월별로 나누어 카테고리별로 채널의 총 비디오 보유량을 우측와 같이 표 형태로 알아볼 수도 있습니다.

카테고리의 개수는 총 15개이기 때문에, 가장 비디오 보유량이 많은 Entertainment 카테고리 기준으로 3월 채널별 비디오 보유량을 추출해 보았는데요,

하지만, 단순히 채널의 비디오보유량을 알아보는 것은 큰 의미가 없습니다.

왜냐하면 첫번째 분석에서 인기비디오로 선정되는 것과 비디오보유량 간에 절대적인 상관관계는 없는 것으로 결론지어졌기 때문입니다. (대략적으로 100개 이상의 비디오 보유량 시에, 인기비디오 선정과 큰 상관 없음) 예를 들어 CH6erID라는 아이디의 소유자가 340724개의 비디오 보유량을 7월에 가졌다 한들, 그것이 인기비디오로 채택되는 이유는 아니라는 의미입니다.

```

print(f'3월: {category[0]} 카테고리에서 채널별 비디오 평
category_df3[0]

```

3월: Entertainment 카테고리에서 채널별 비디오 평균보유량

channel\_id channel\_videos\_mean

10	CH1Y5oU	4
9	CHEwOn7	15
8	CHDb1t5	40
11	CHJmcPV	76
18	CHjGoJb	76
6	CH7Krez	86
4	CH68buD	101
22	CHriPmQ	132
12	CHLVwgJ	167
0	CH12YJZ	169
17	ChdULCa	387
21	CHrhH1d	425
1	CH46BbE	655

이번에는 카테고리별로 나누어 각 채널의 월별 비디오 보유량을 시각화해 보겠습니다.

- 이 분석을 통해 인기비디오로 선정된 채널의 월별 비디오 제작량을 알 수 있게 됩니다. 결과적으로 우리는 인기비디오로 선정되기 위해 1개월을 기준으로 몇 개의 비디오를 제작해야 하는지 알 수 있습니다.

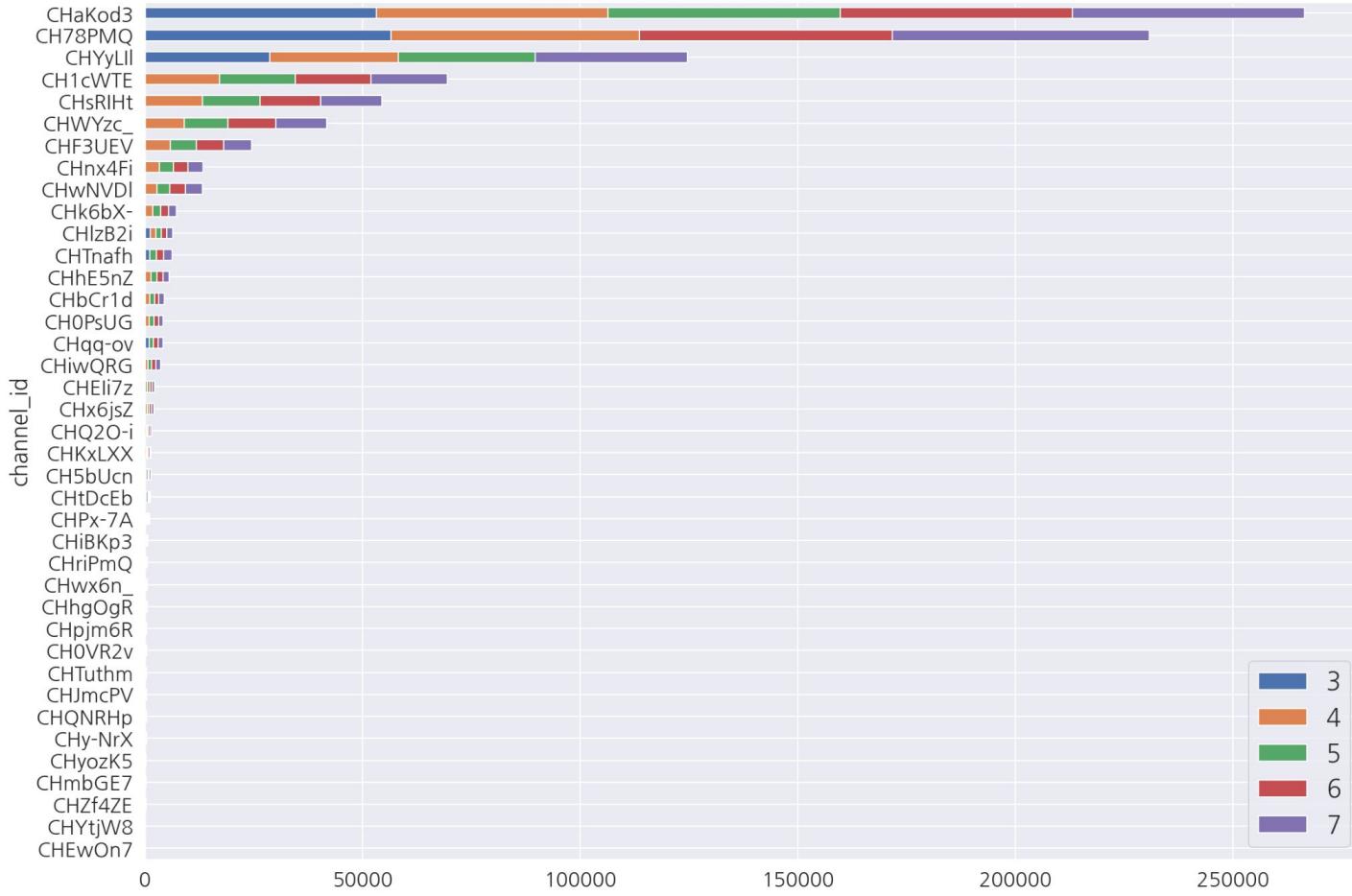
## 0. Entertainment

Entertainment 카테고리에 속한 'channel\_id'의 개수는 312개입니다.

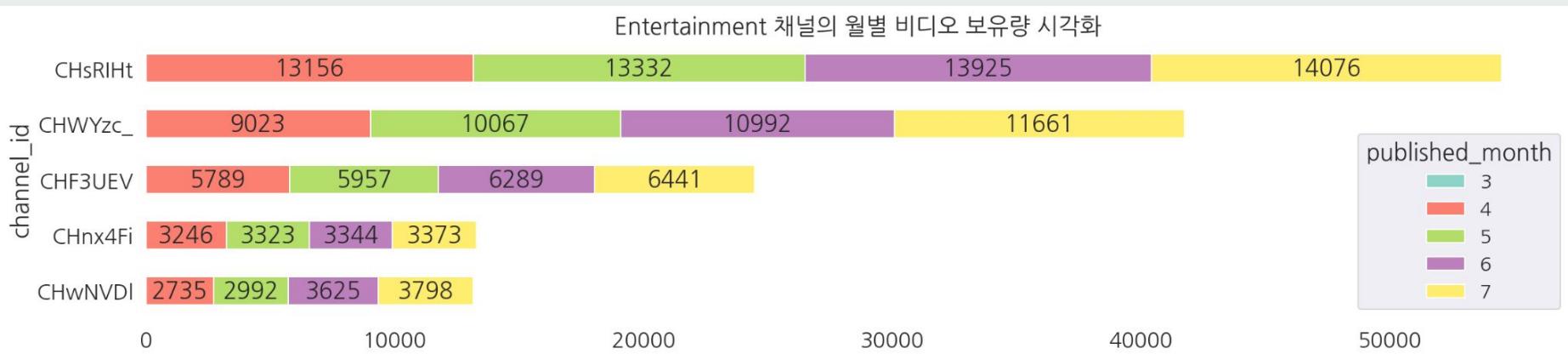
312개의 channel\_id를 모두 시각화하기에는 너무 많습니다.

312개의 channel\_id가 매월 인기비디오로 채택된 것은 아닙니다. 알고자 하는 것은 채널의 월별 비디오 제작량이므로, 월별 비디오 보유량의 추이를 볼 수 있어야 합니다. 따라서 총 5개월 중 4개월 이상 꾸준히 Entertainment 인기비디오로 채택된 channel\_id 40개만 추출해 보겠습니다. 그리고 심한 아웃라이어인 'CHmjNKit'는 제외합니다.

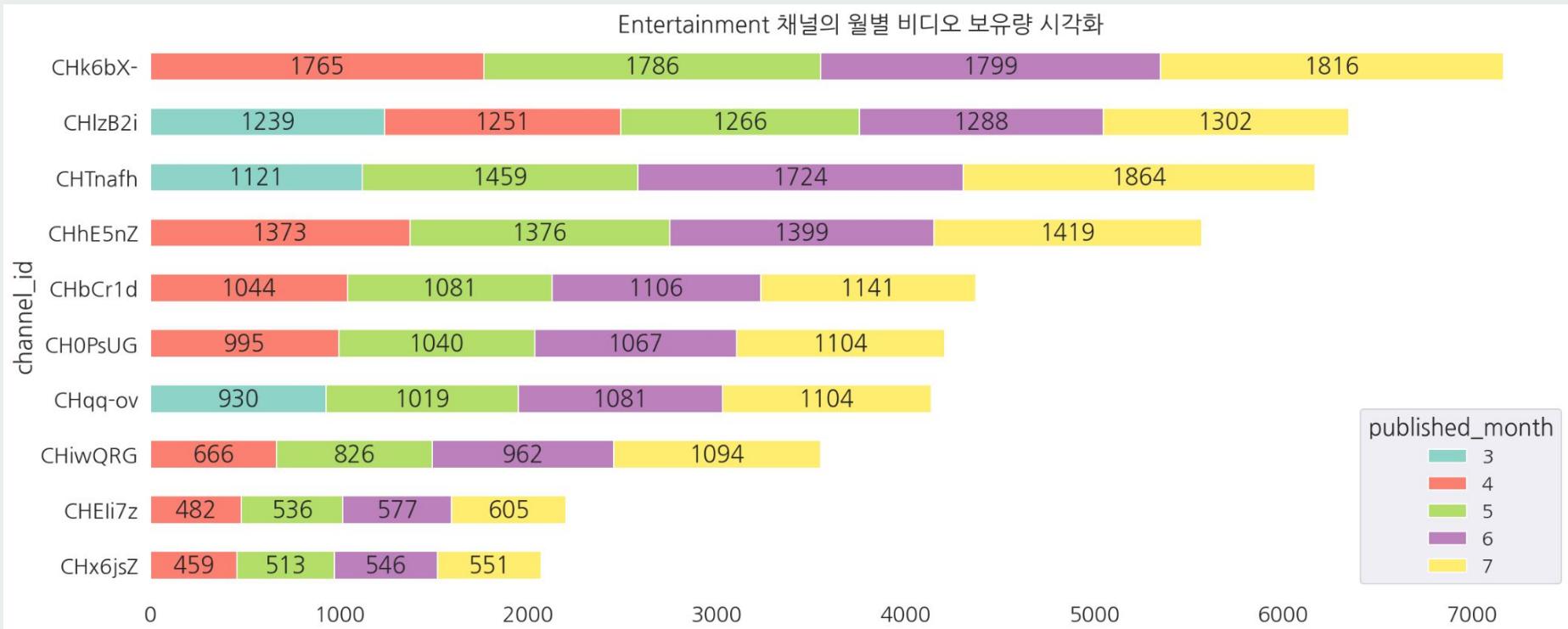
Entertainment 채널의 월별 비디오 보유량 시각화



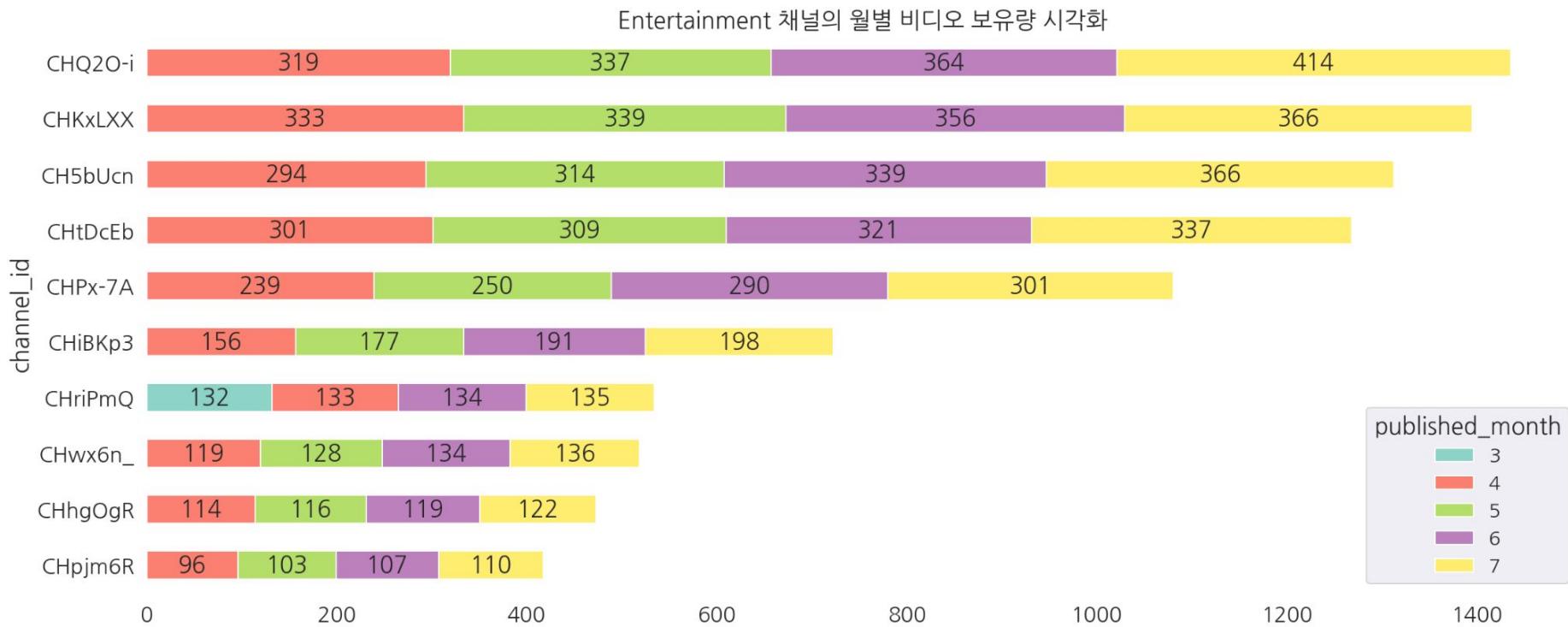
범위가 너무 커서  
채널들의 수치가 한눈에  
보이지 않습니다.  
  
비디오 개수가 많은  
순으로 잘라서 차례대로  
보겠습니다



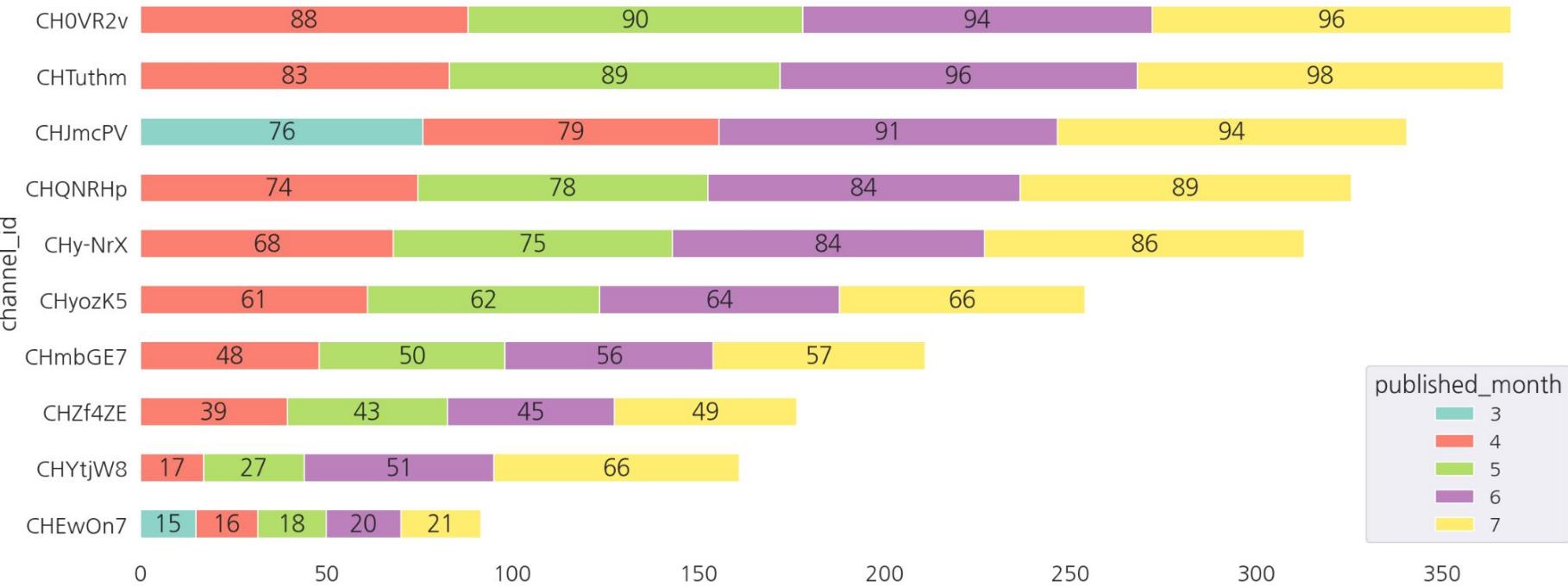
### Entertainment 채널의 월별 비디오 보유량 시각화



Entertainment 채널의 월별 비디오 보유량 시각화

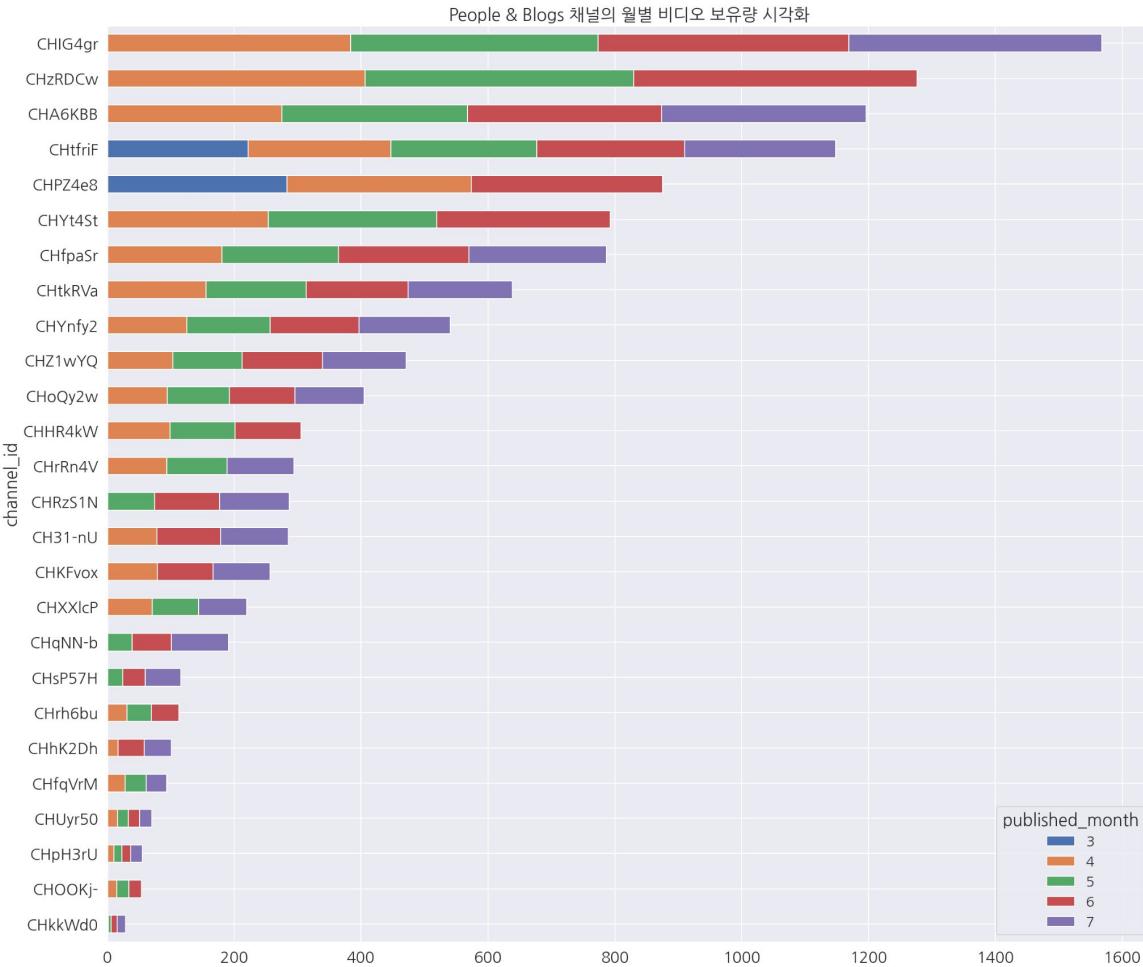


Entertainment 채널의 월별 비디오 보유량 시각화

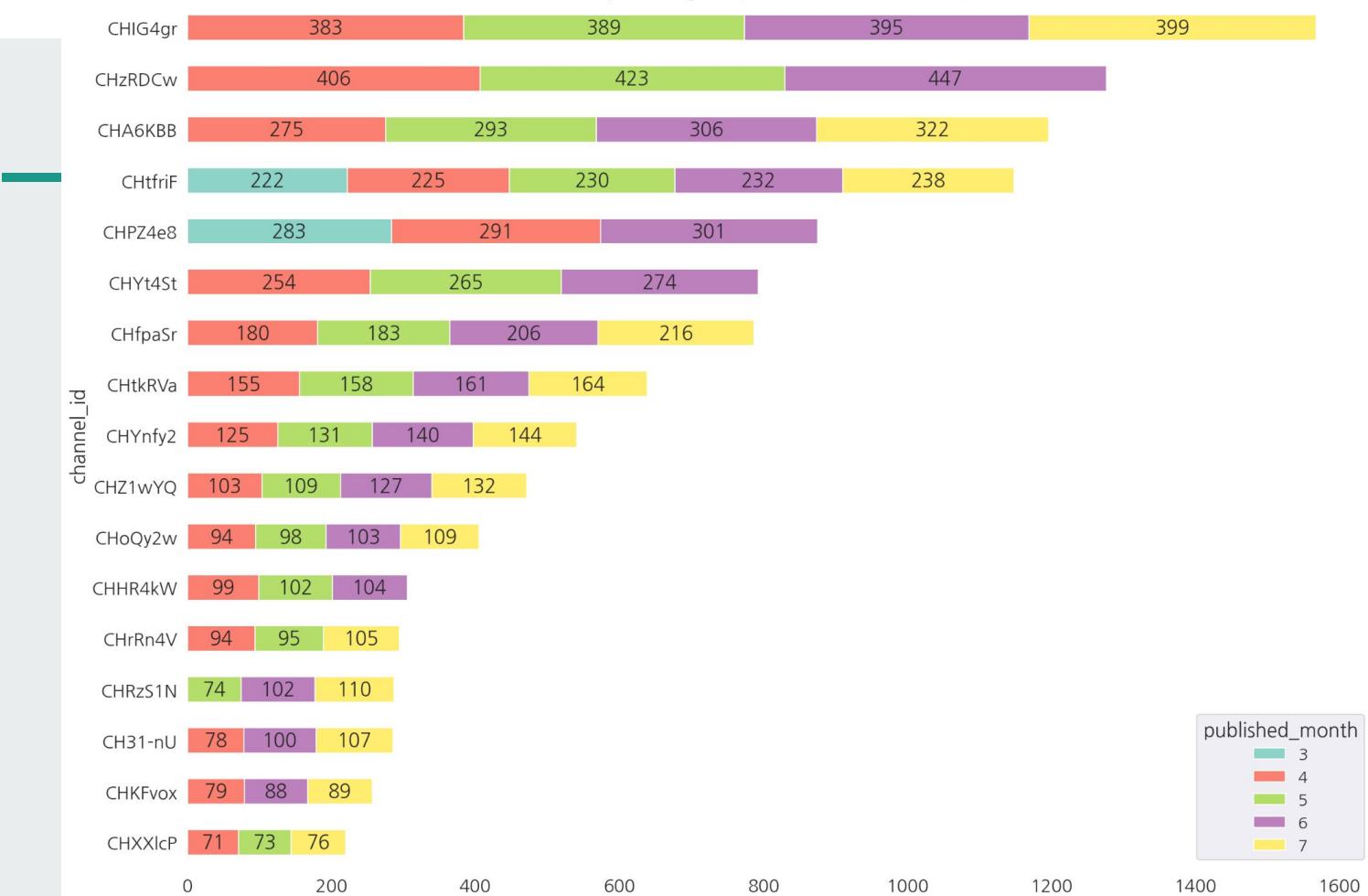


People & Blogs

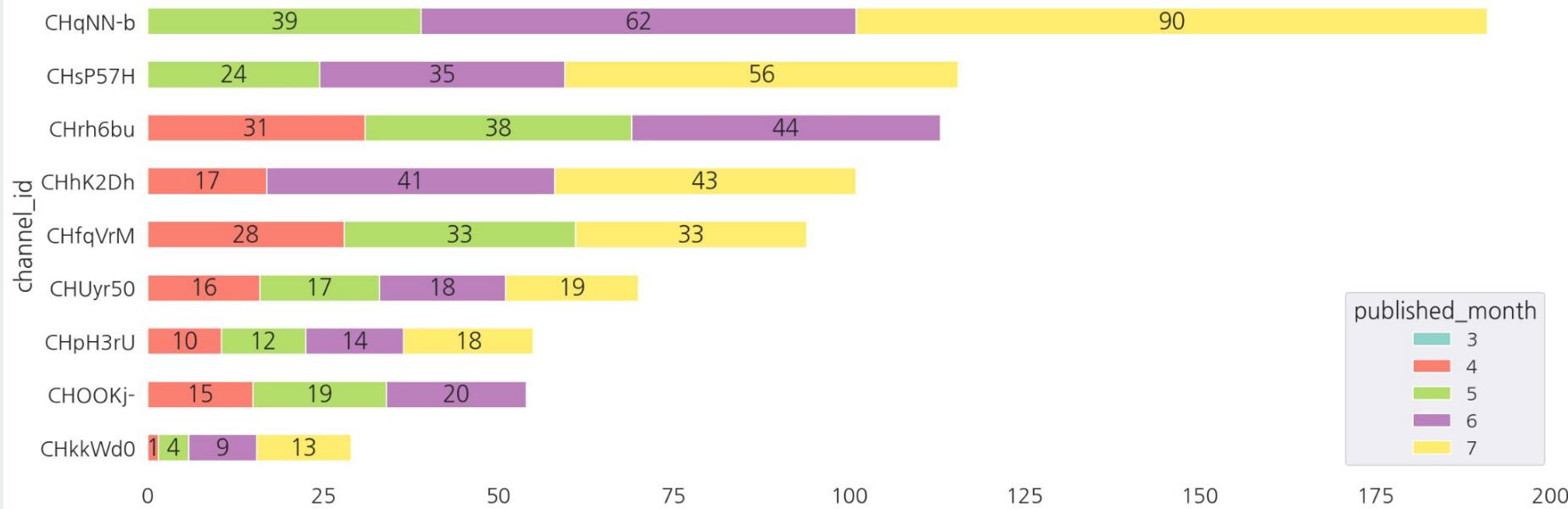
166개의 channel\_id가 있습니다. 5개월 중 3개월 이상 꾸준히 선정된 채널 28개의 channel\_id만 추출합니다.  
아웃라이어인 'CHbFzvz', 'CHaZS\_X'를 제외합니다.



People &amp; Blogs 채널의 월별 비디오 보유량 시각화



People & Blogs 채널의 월별 비디오 보유량 시각화





## 2.Music

98개의 channel\_id가 있습니다.

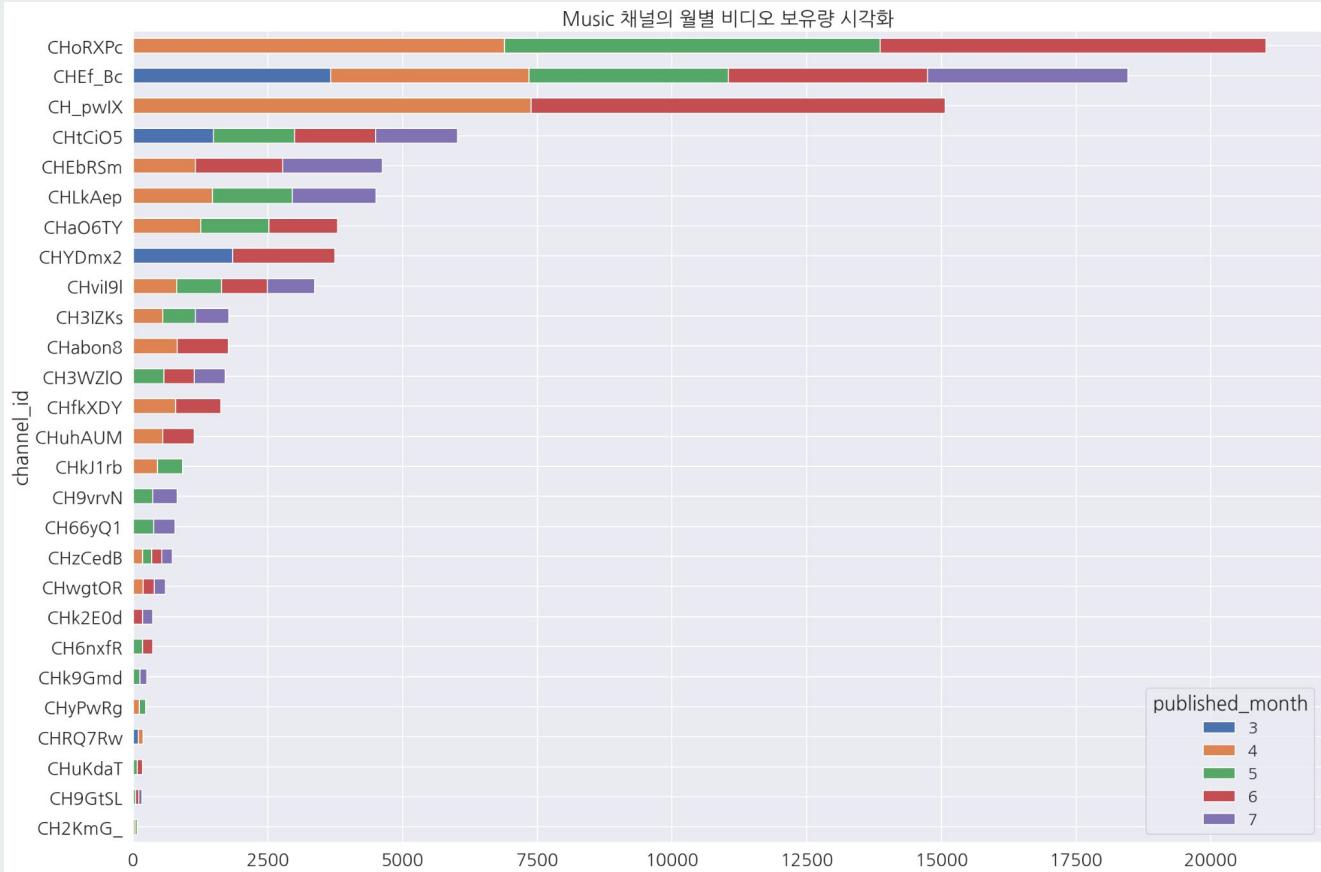
5개월 중 2개월 이상 선정된

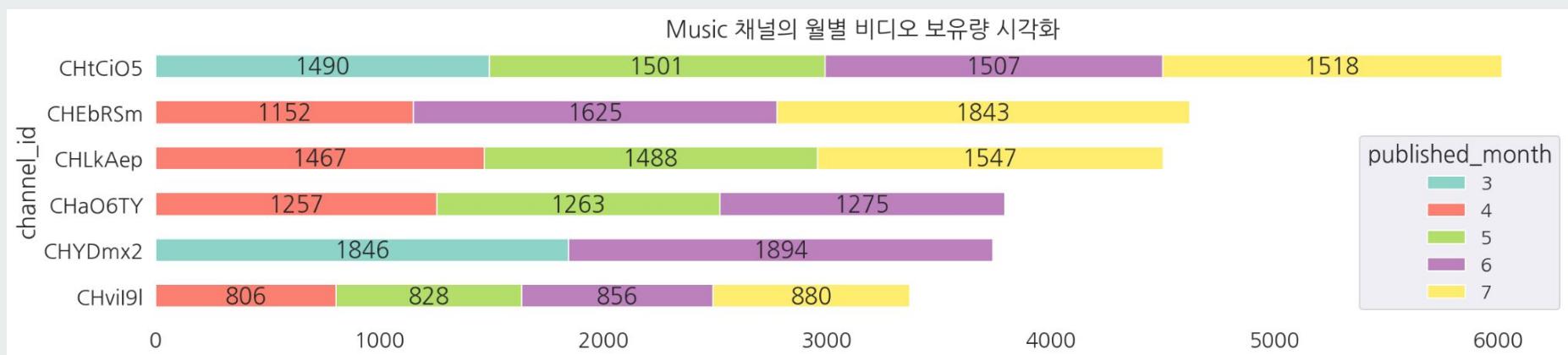
채널 29개의 channel\_id만

추출합니다. 웃라이어인

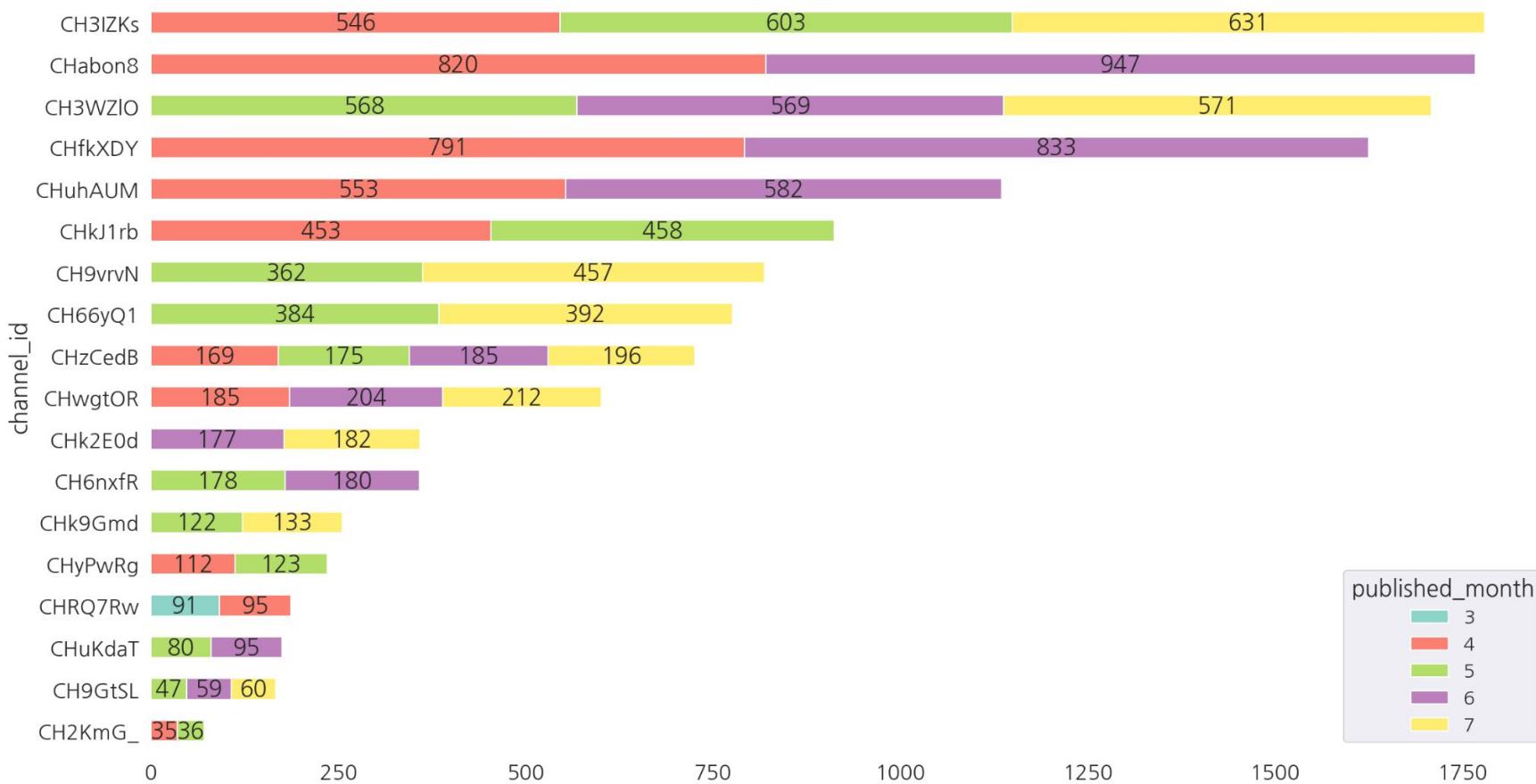
'CHeLPm9', 'CHe52oe'를

제외합니다.



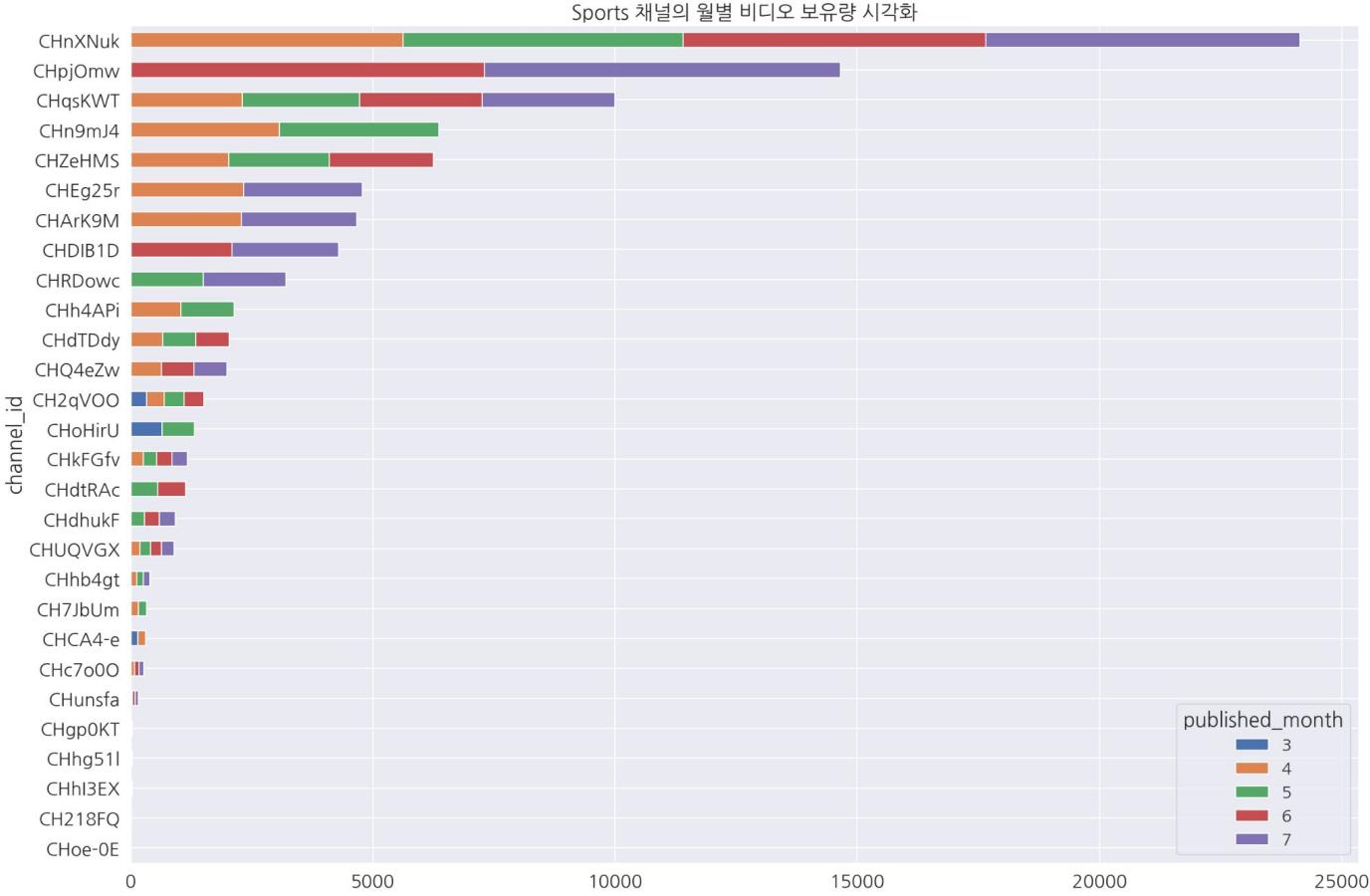


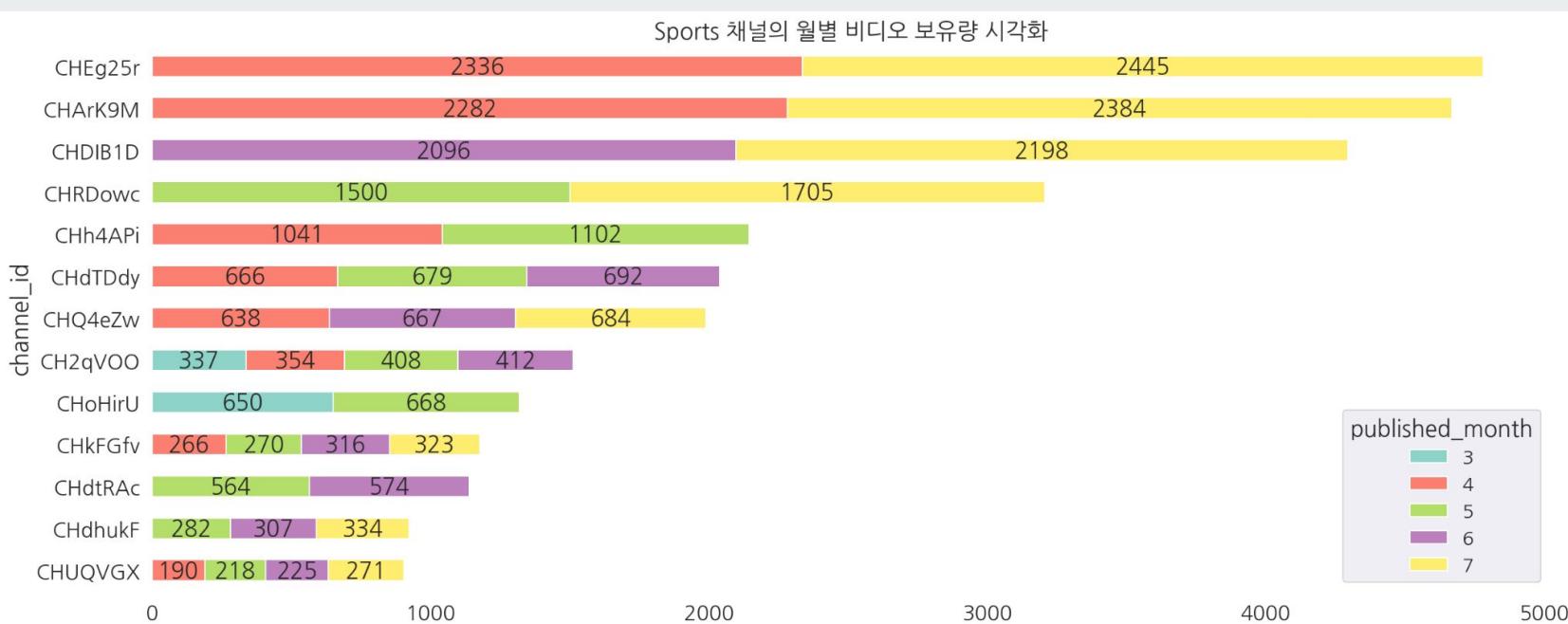
Music 채널의 월별 비디오 보유량 시각화



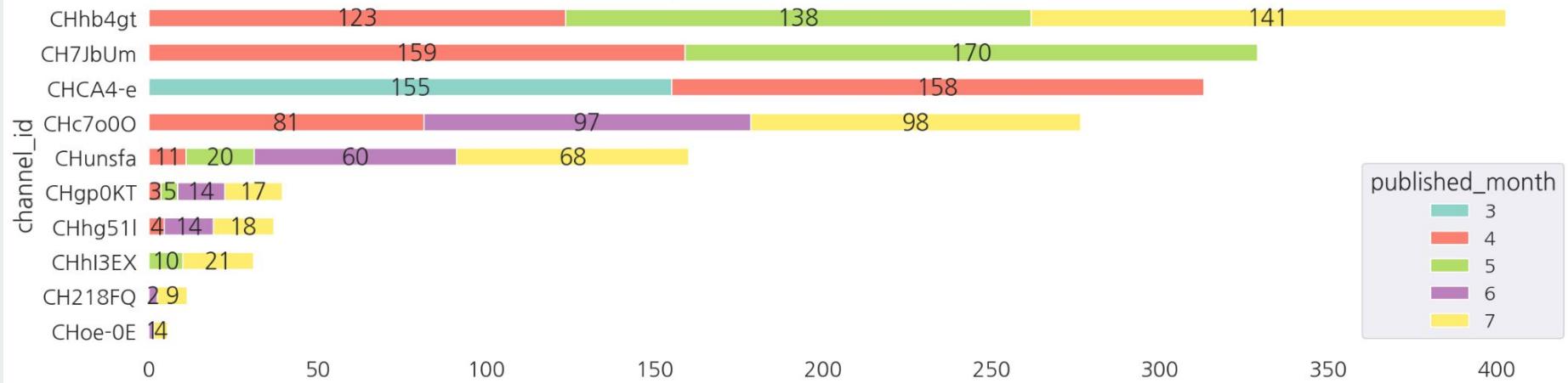
### 3.Sports

66개의 channel\_id가 있습니다.  
5개월 중 2개월 이상 선정된  
채널 30개의 channel\_id만  
추출합니다. 아웃라이어인  
'CHtm\_Qo', 'CHoLrcj'를  
제외합니다.



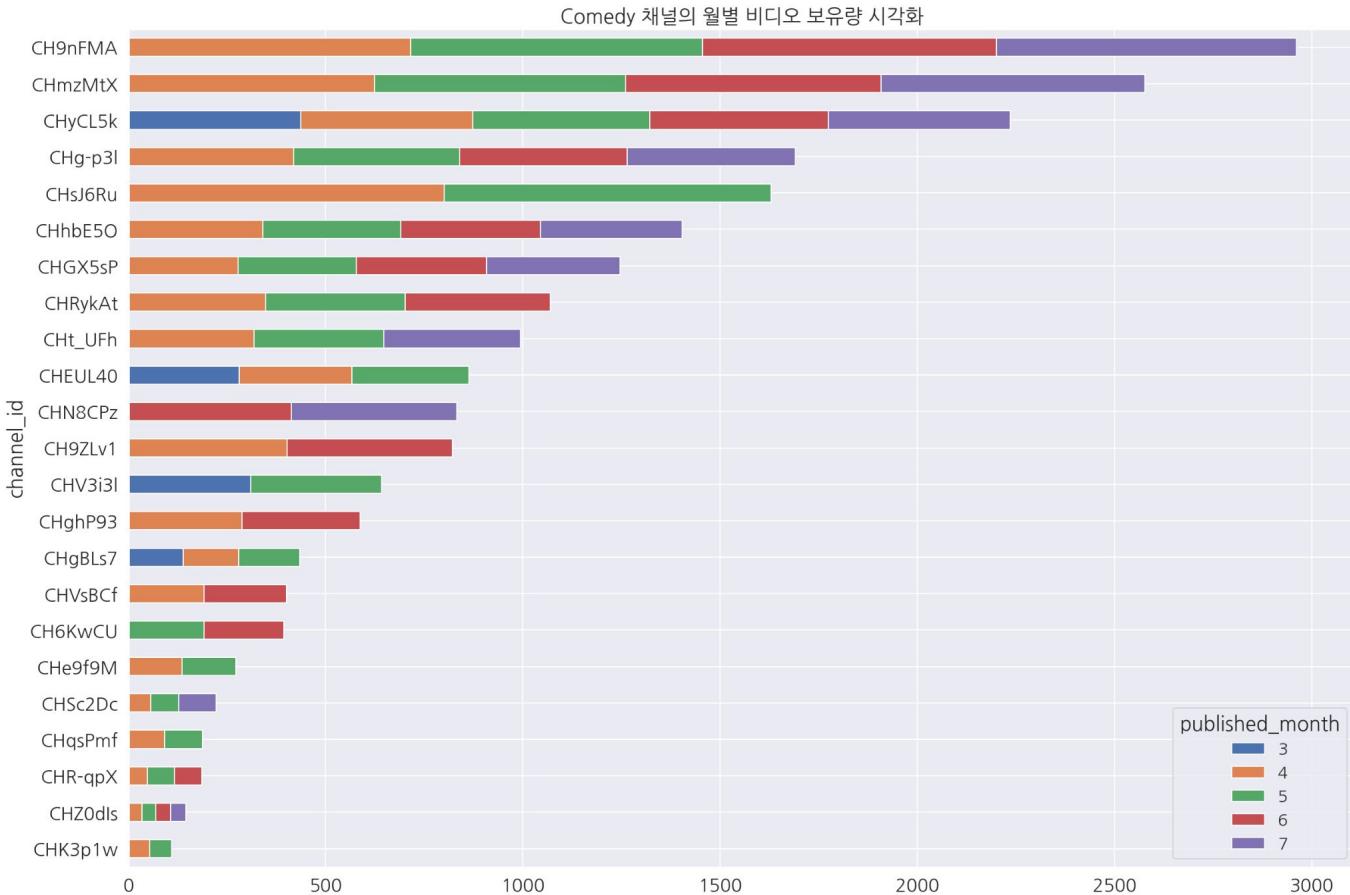


Sports 채널의 월별 비디오 보유량 시각화

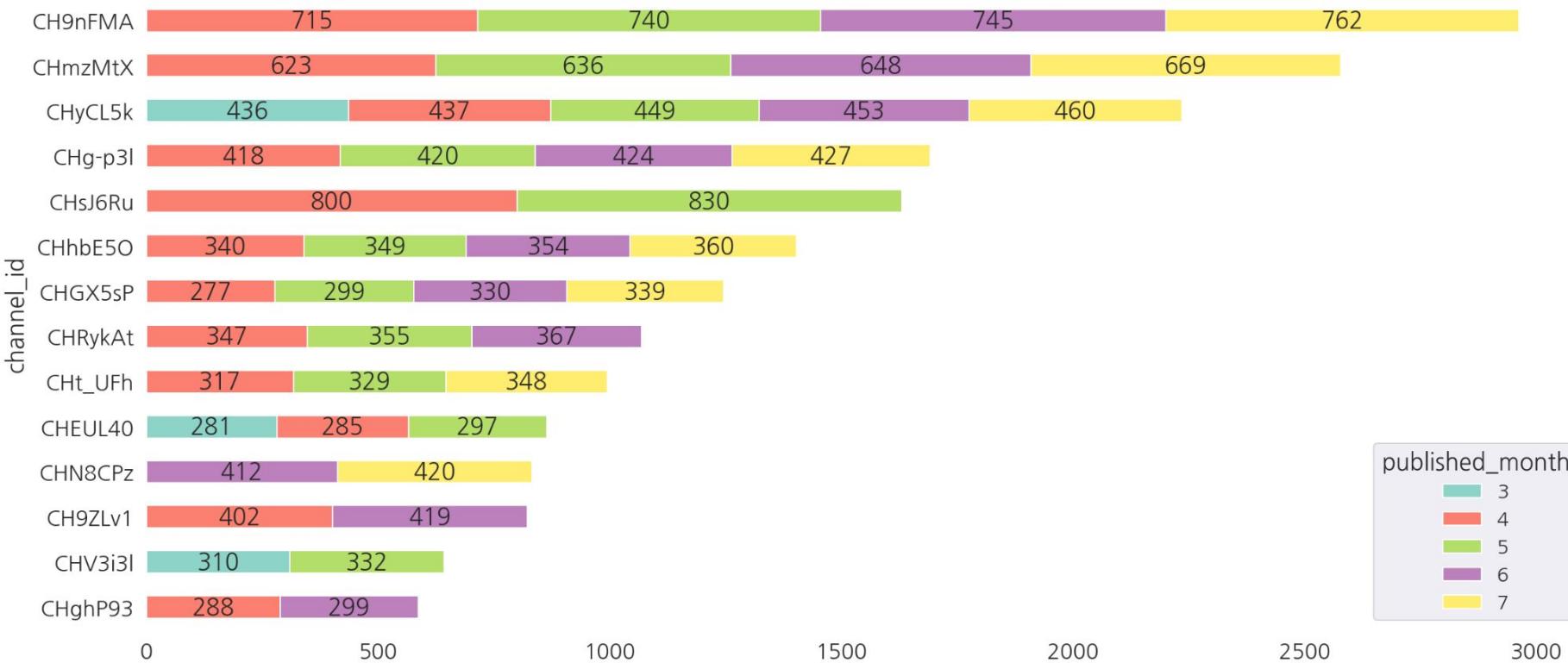


## 4.Comedy

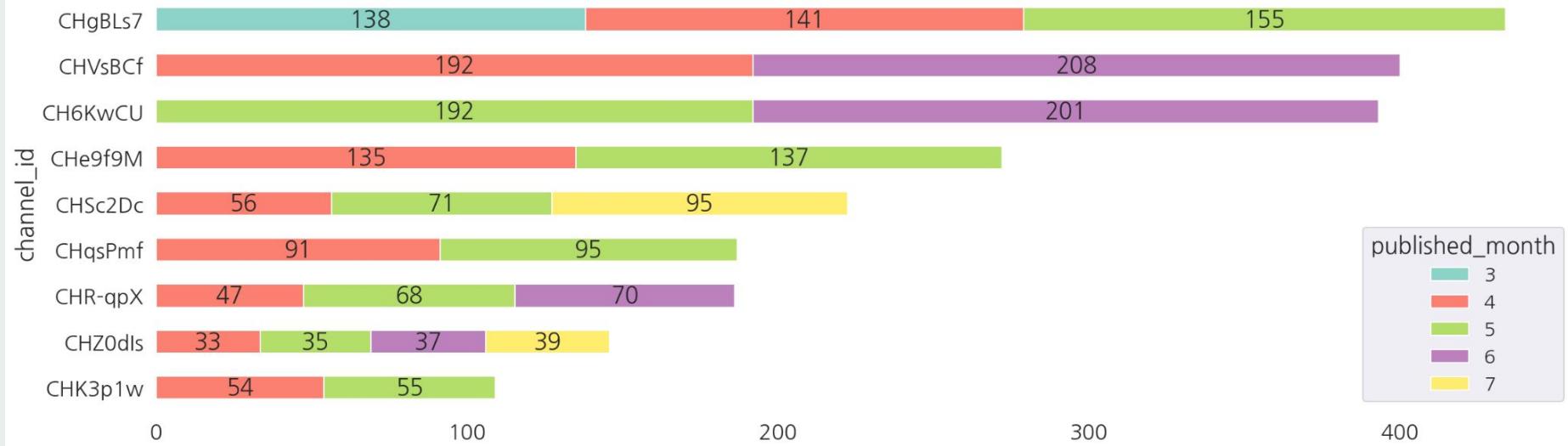
50개의 channel\_id가 있습니다. 5개월 중 2개월 이상 선정된 채널 23개의 channel\_id만 추출합니다.



Comedy 채널의 월별 비디오 보유량 시각화



Comedy 채널의 월별 비디오 보유량 시각화

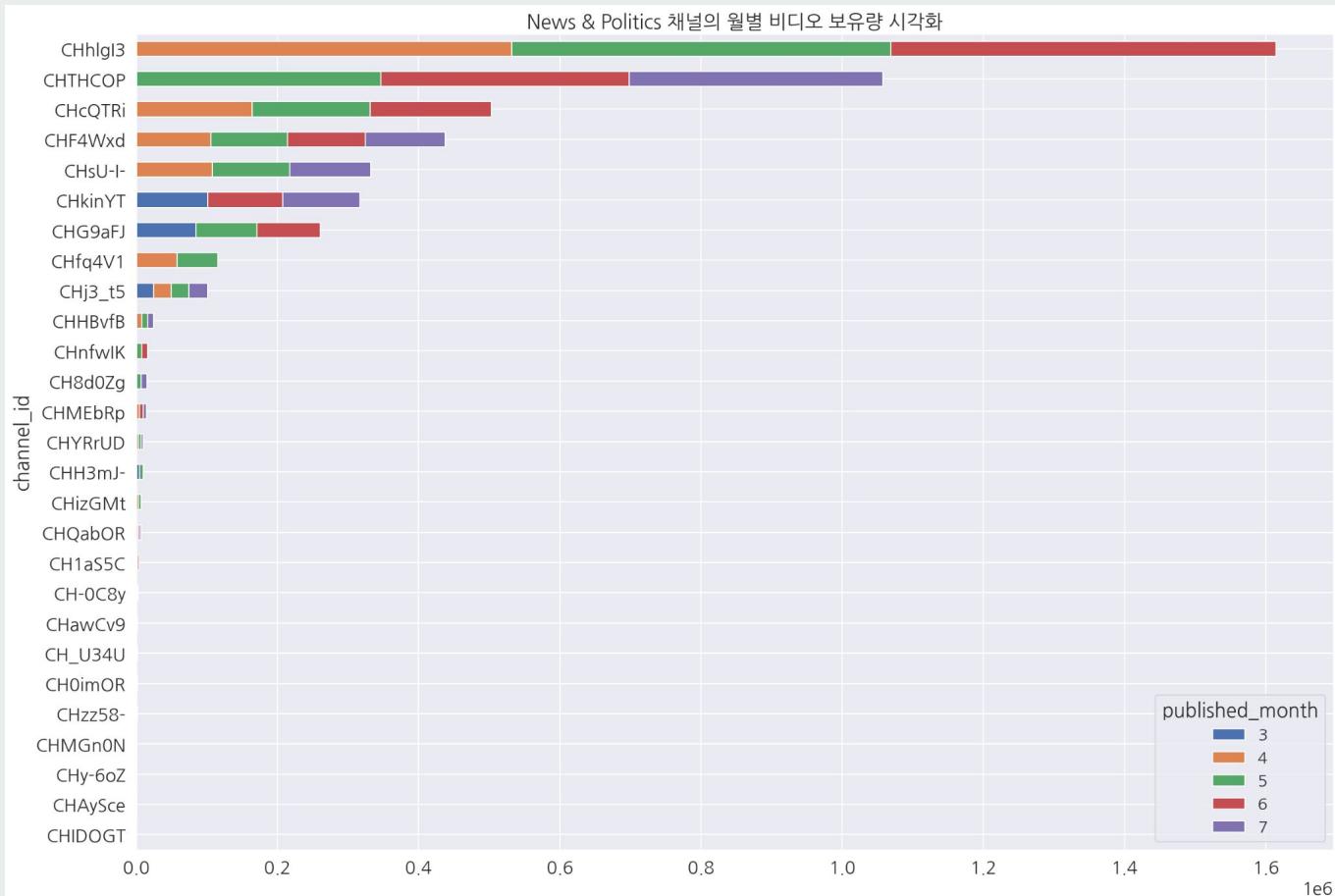


## 5.News & Politics

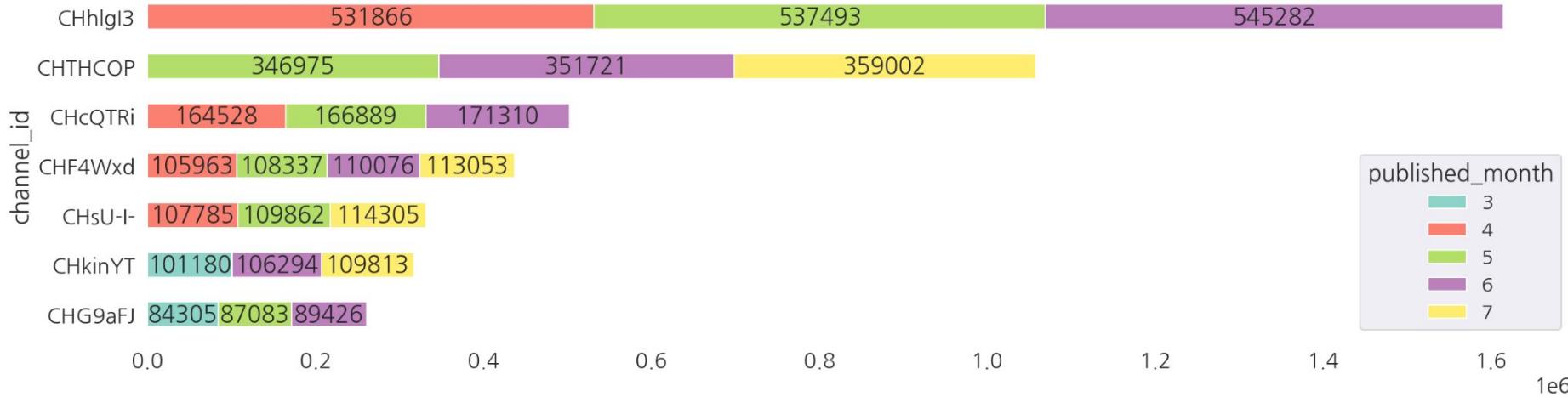
56개의 channel\_id가 있습니다.

5개월 중 2개월 이상 선정된 채널

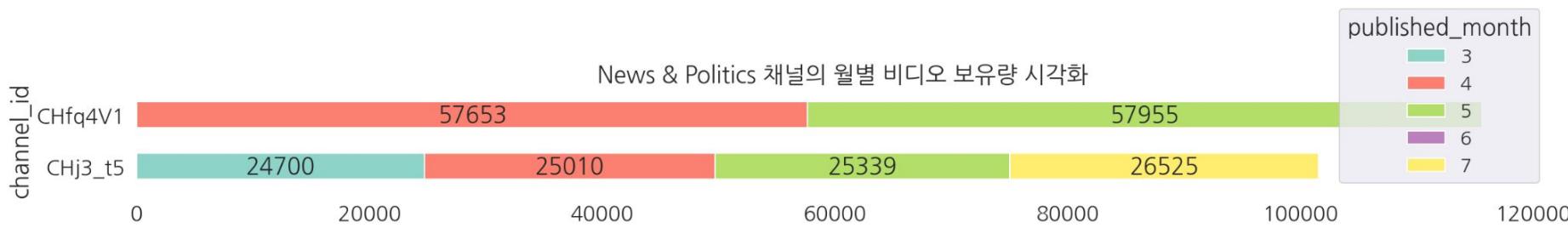
27개의 channel\_id만 추출합니다.



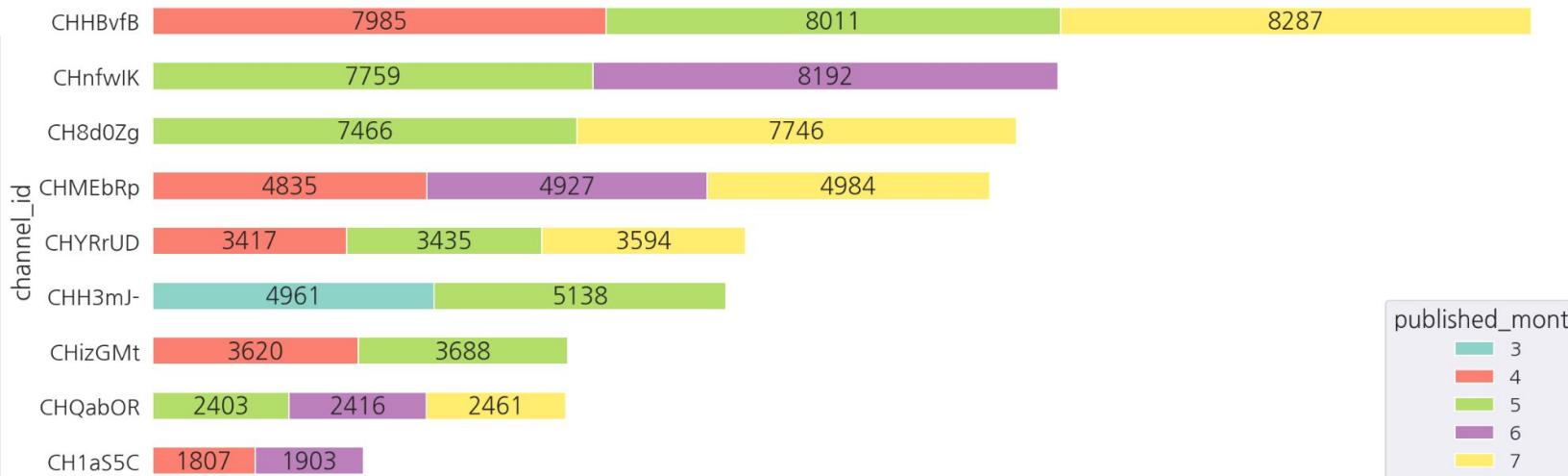
### News & Politics 채널의 월별 비디오 보유량 시각화



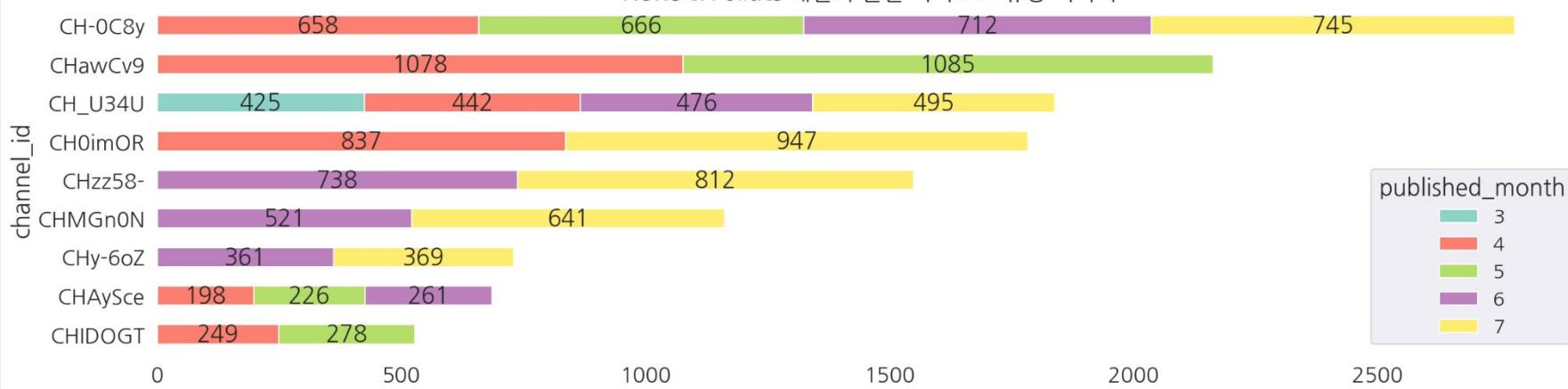
### News & Politics 채널의 월별 비디오 보유량 시각화



News &amp; Politics 채널의 월별 비디오 보유량 시각화



News &amp; Politics 채널의 월별 비디오 보유량 시각화



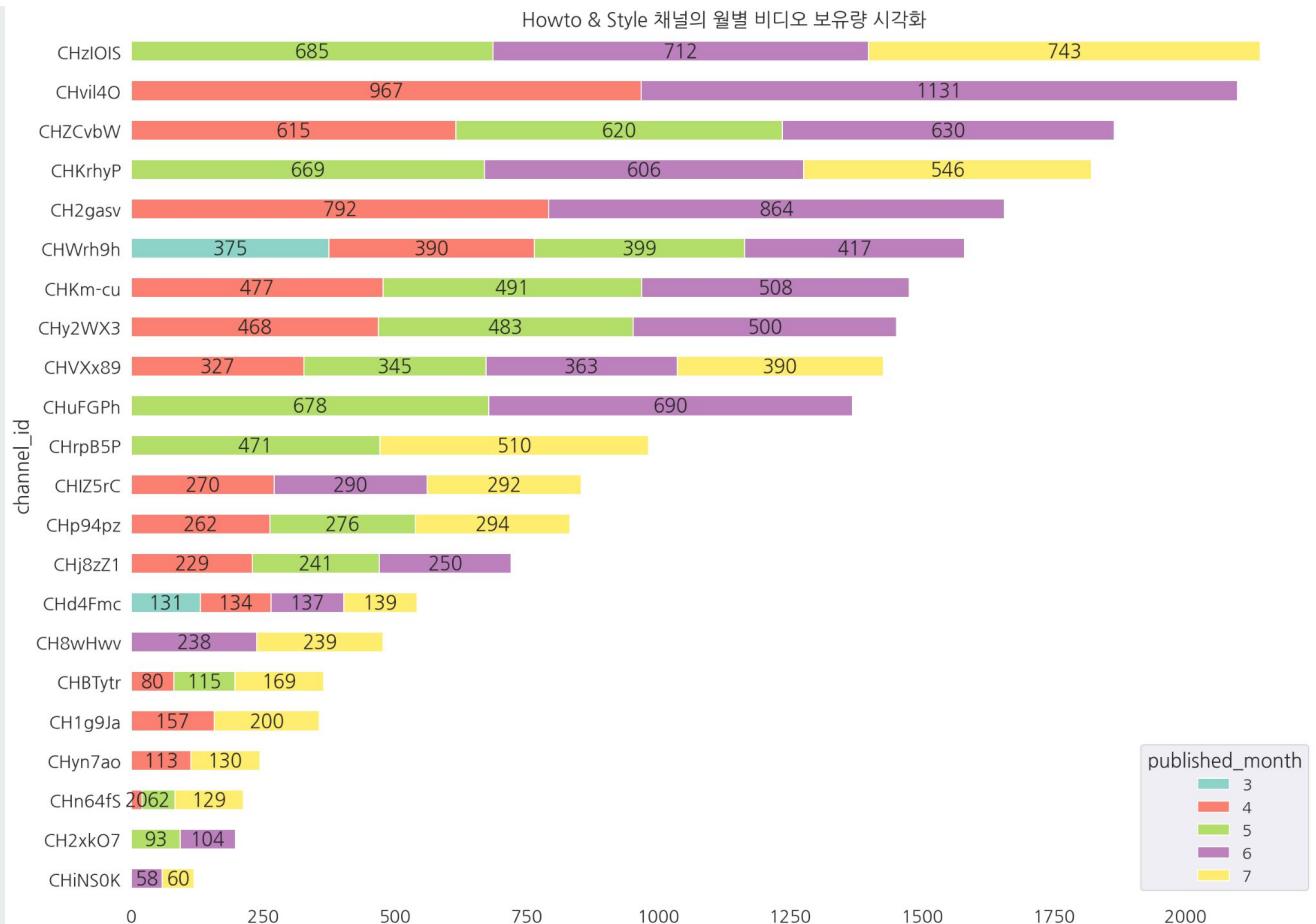


## 6. Howto & Style

44개의 channel\_id가 있습니다.

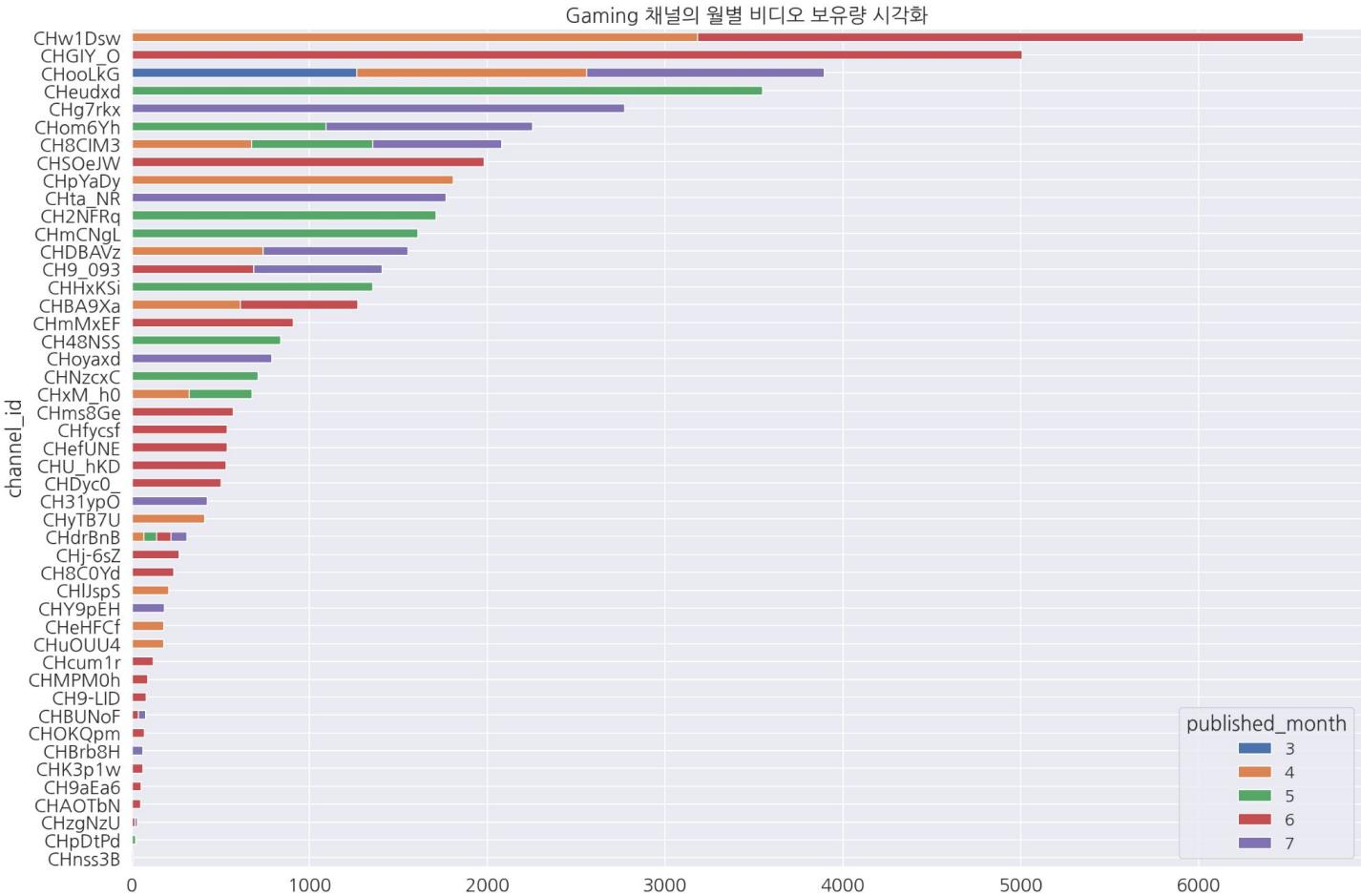
5개월 중 2개월 이상 선정된 채널

22개의 channel\_id만 추출합니다.

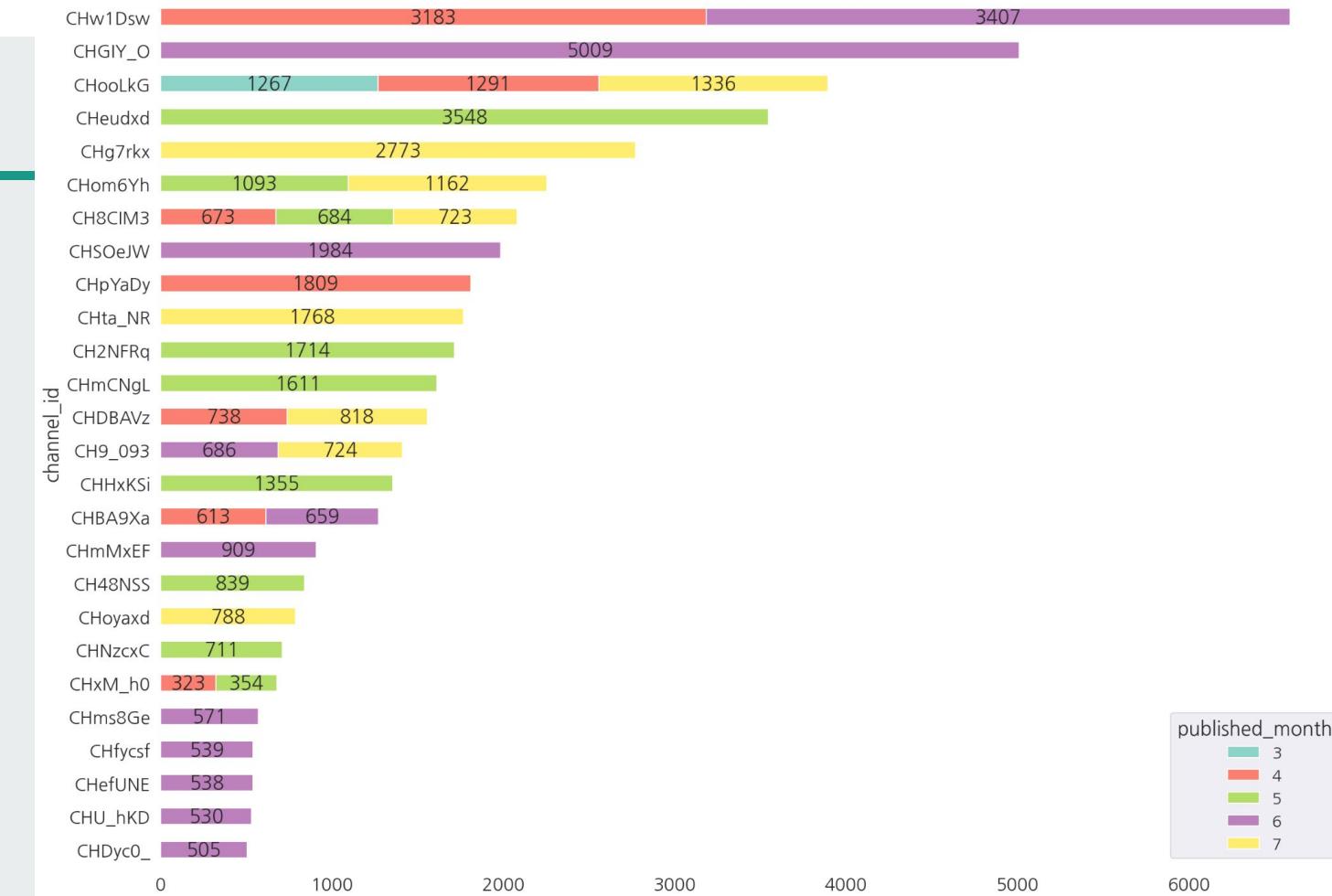


## 7.Gaming

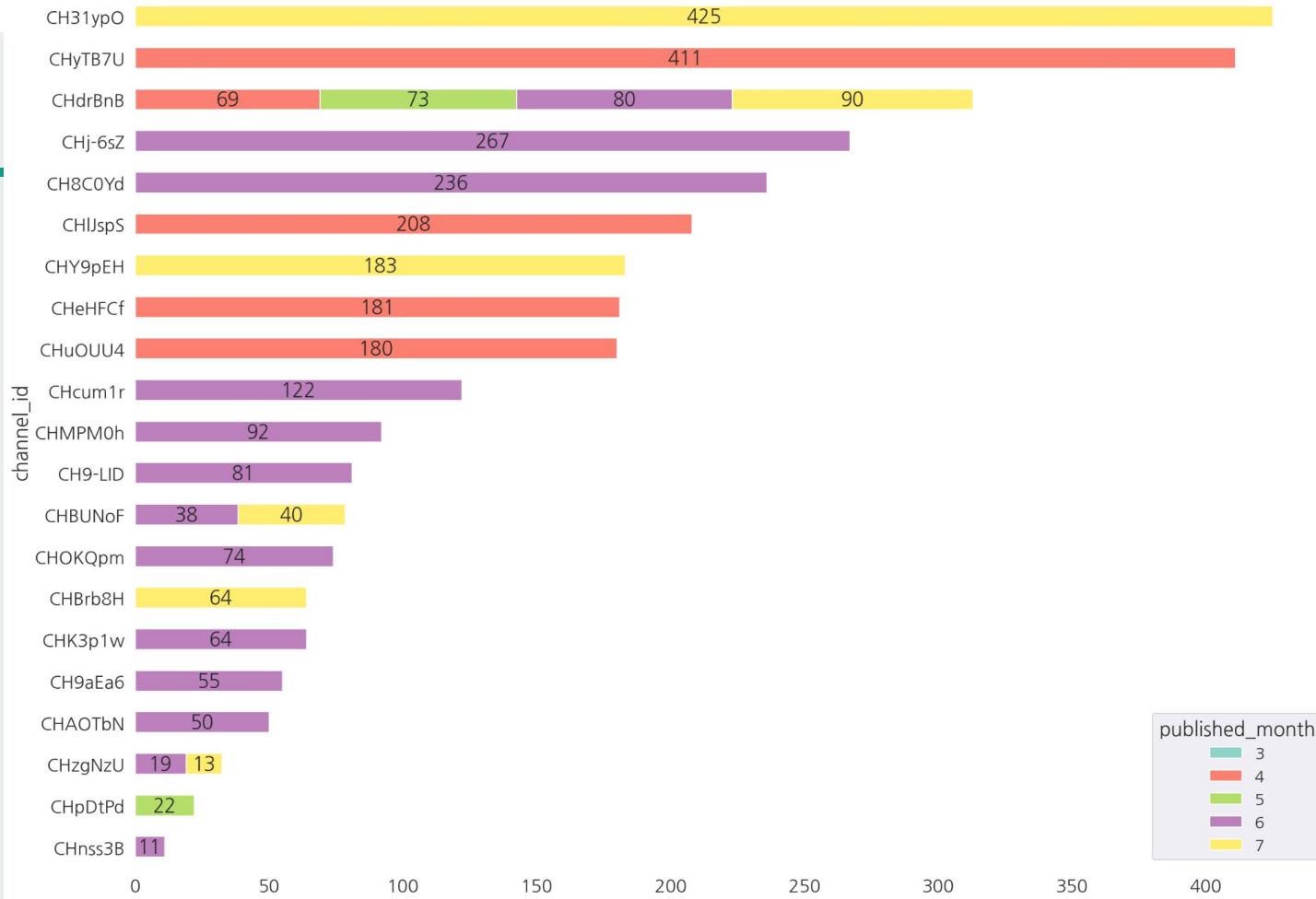
47개의 channel\_id가 있습니다. 2개월 이상 꾸준히 인기비디오로 채택된 채널수가 너무 적어서(11개), 47개 그대로 시각화를 진행합니다.



Gaming 채널의 월별 비디오 보유량 시각화



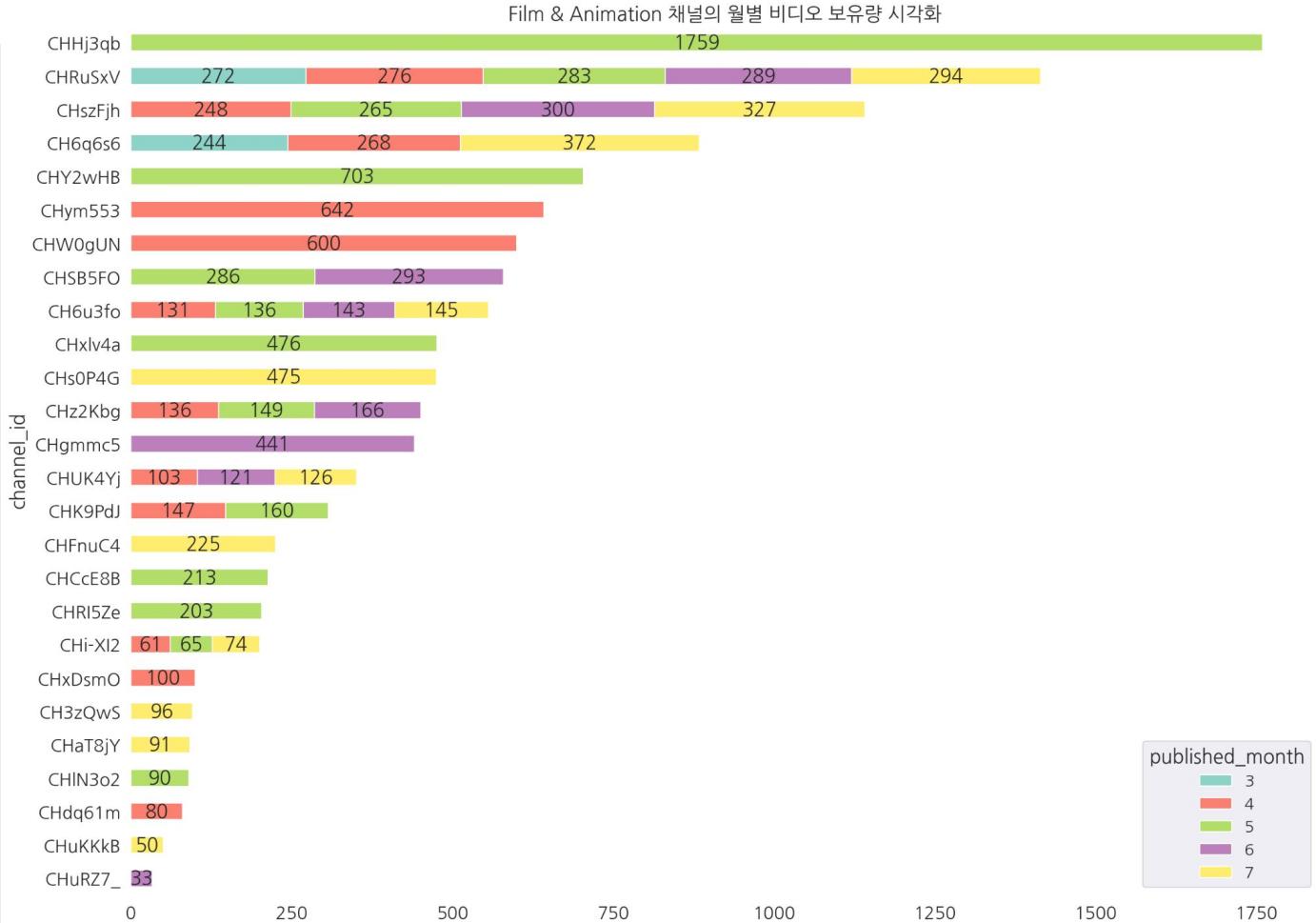
Gaming 채널의 월별 비디오 보유량 시각화



1759

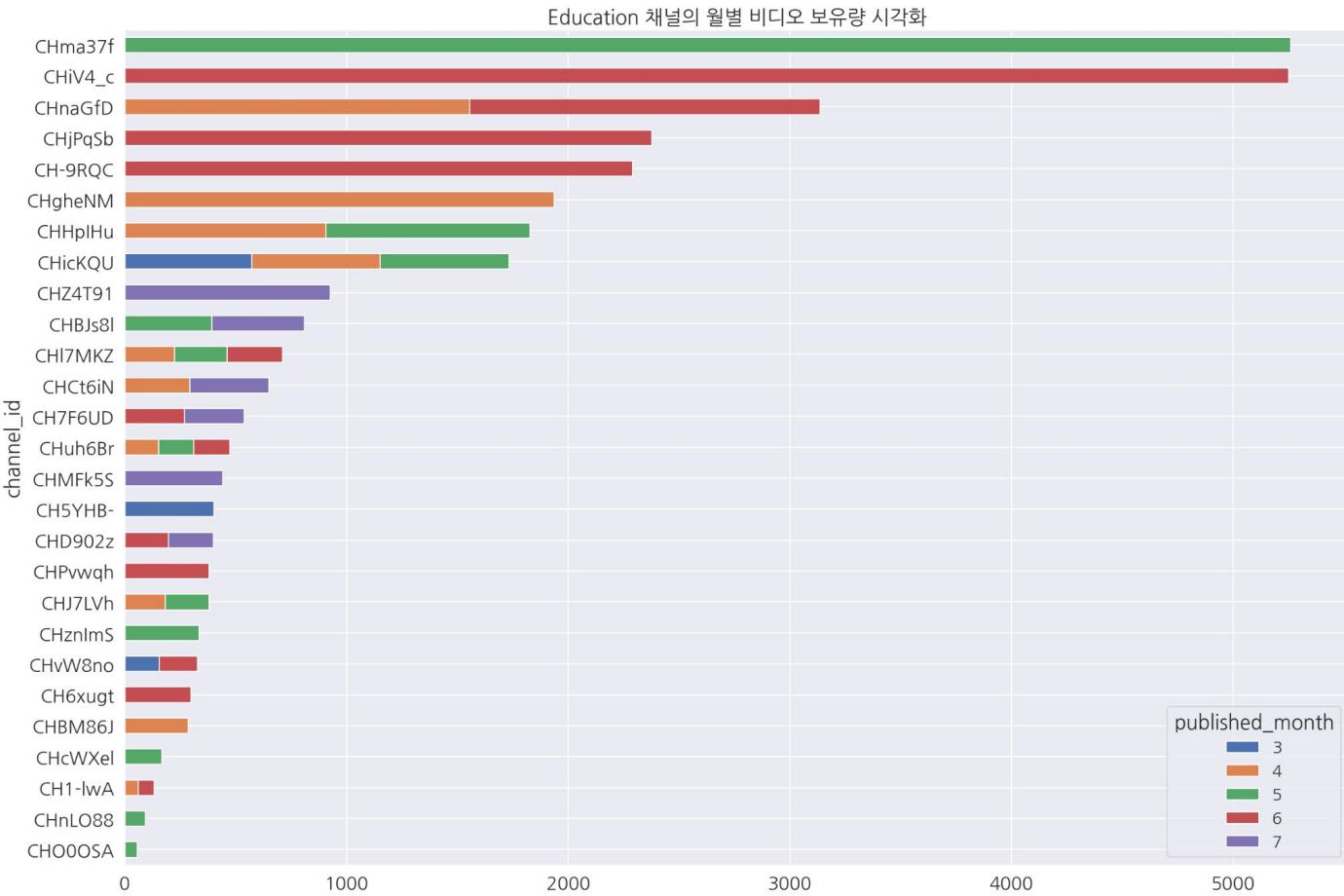
## 8.Film & Animation

27개의 channel\_id가 있습니다. 시각화할 채널수가 27밖에 되지 않아서 그대로 진행합니다. 이하 채널은 모두 같은 형태로 진행합니다.

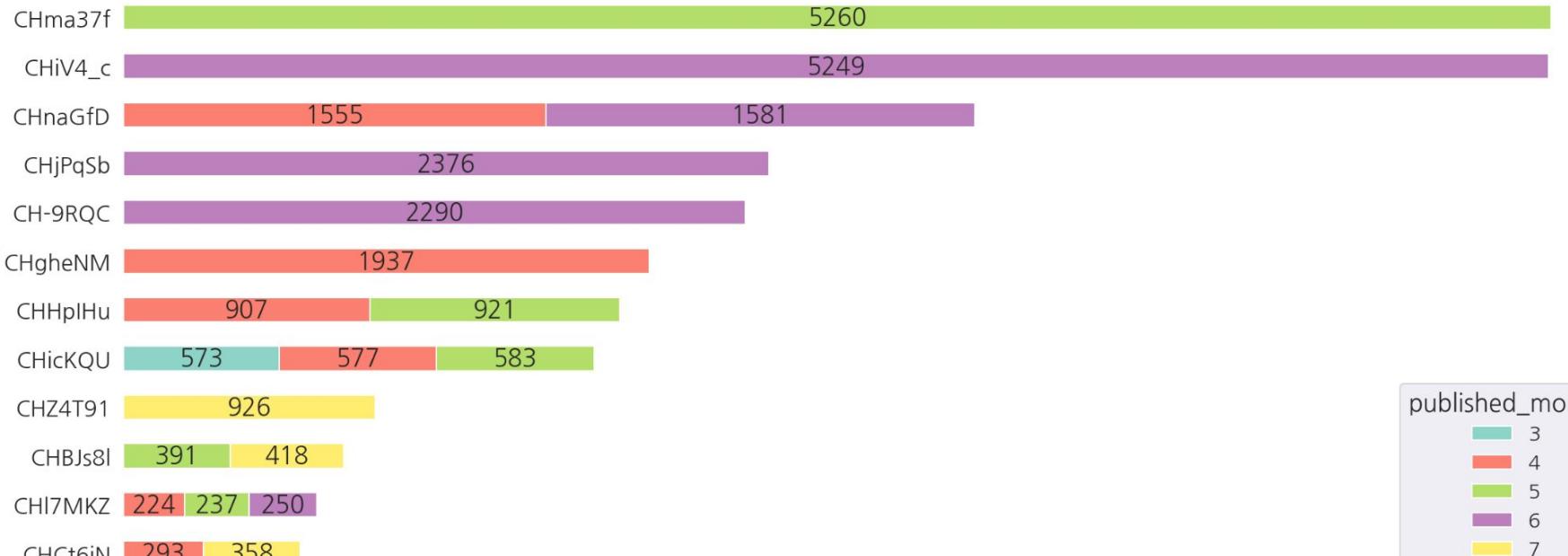


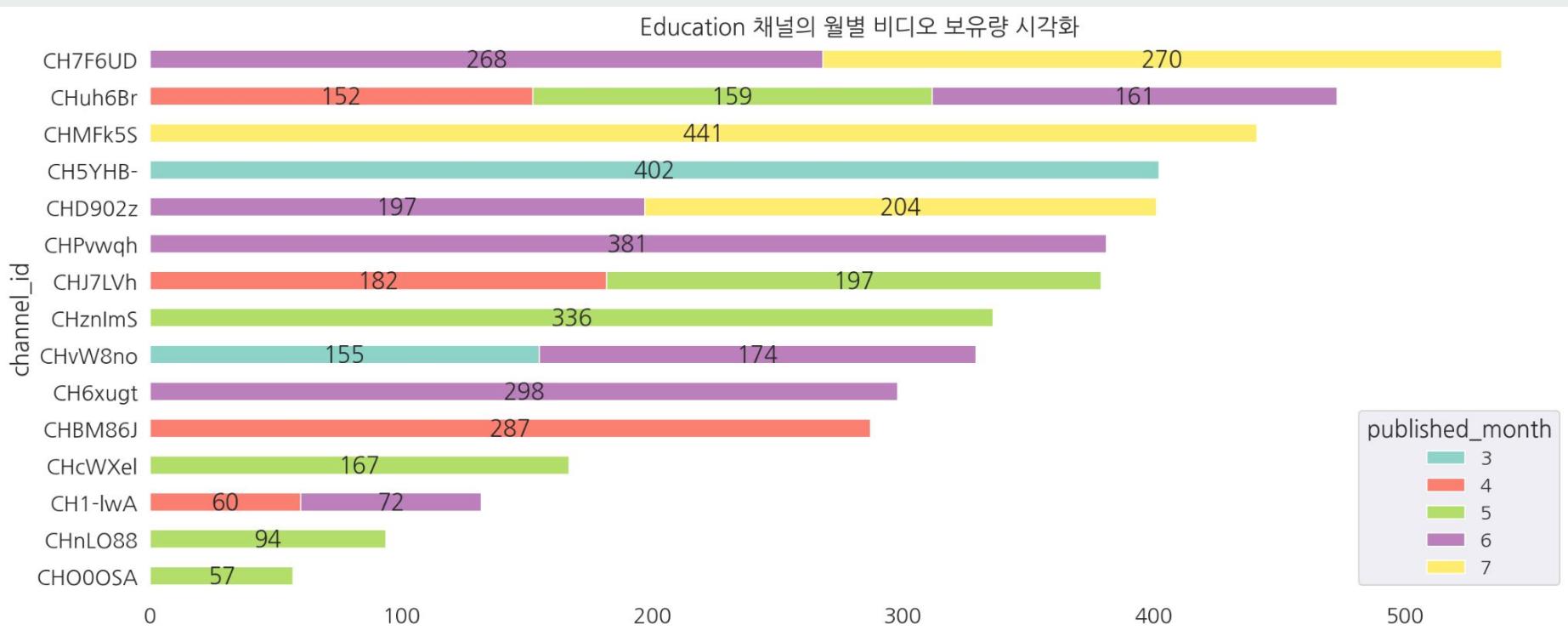
## 9. Education

'channel\_id'의 개수 : 27개



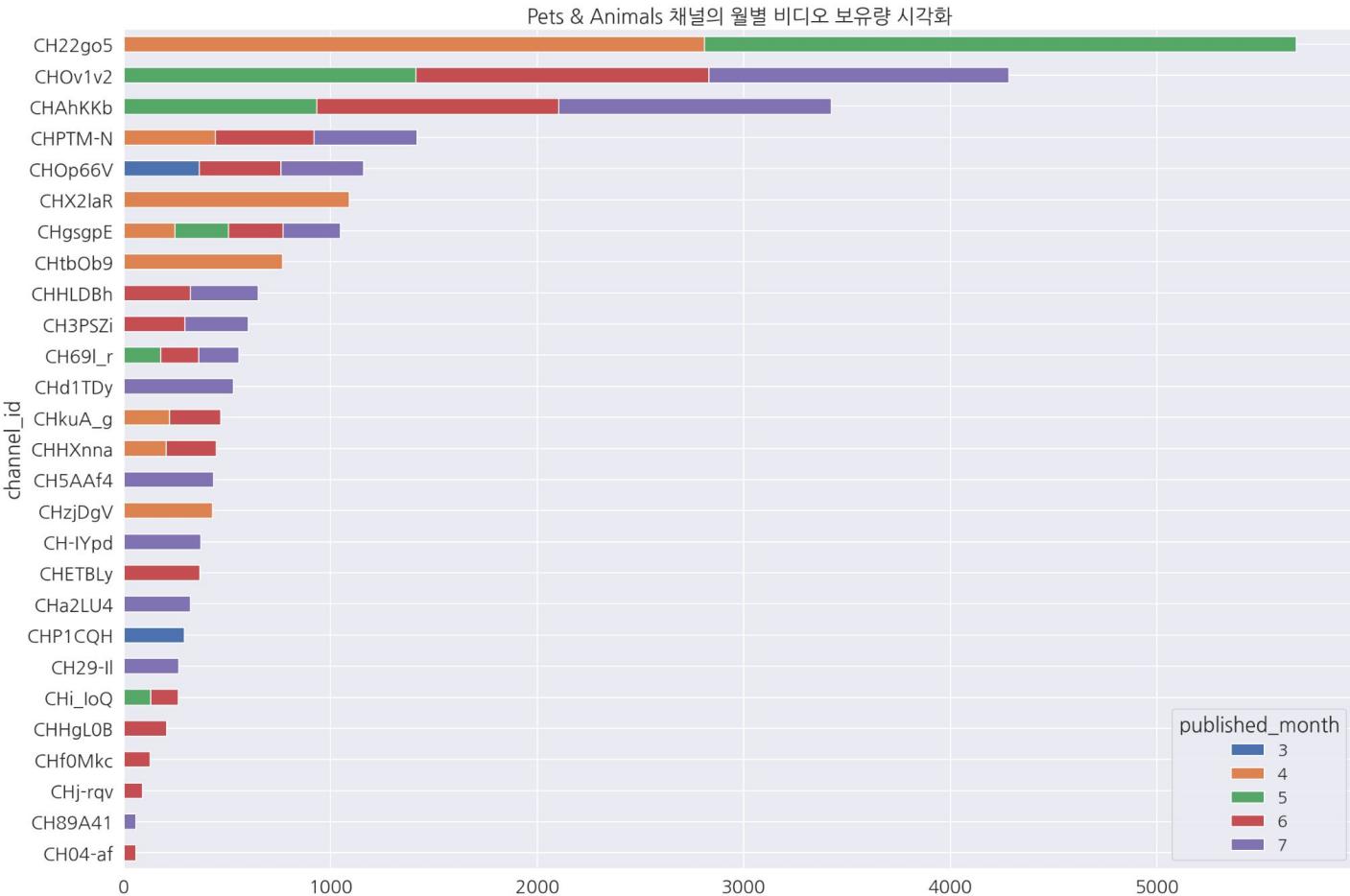
Education 채널의 월별 비디오 보유량 시각화

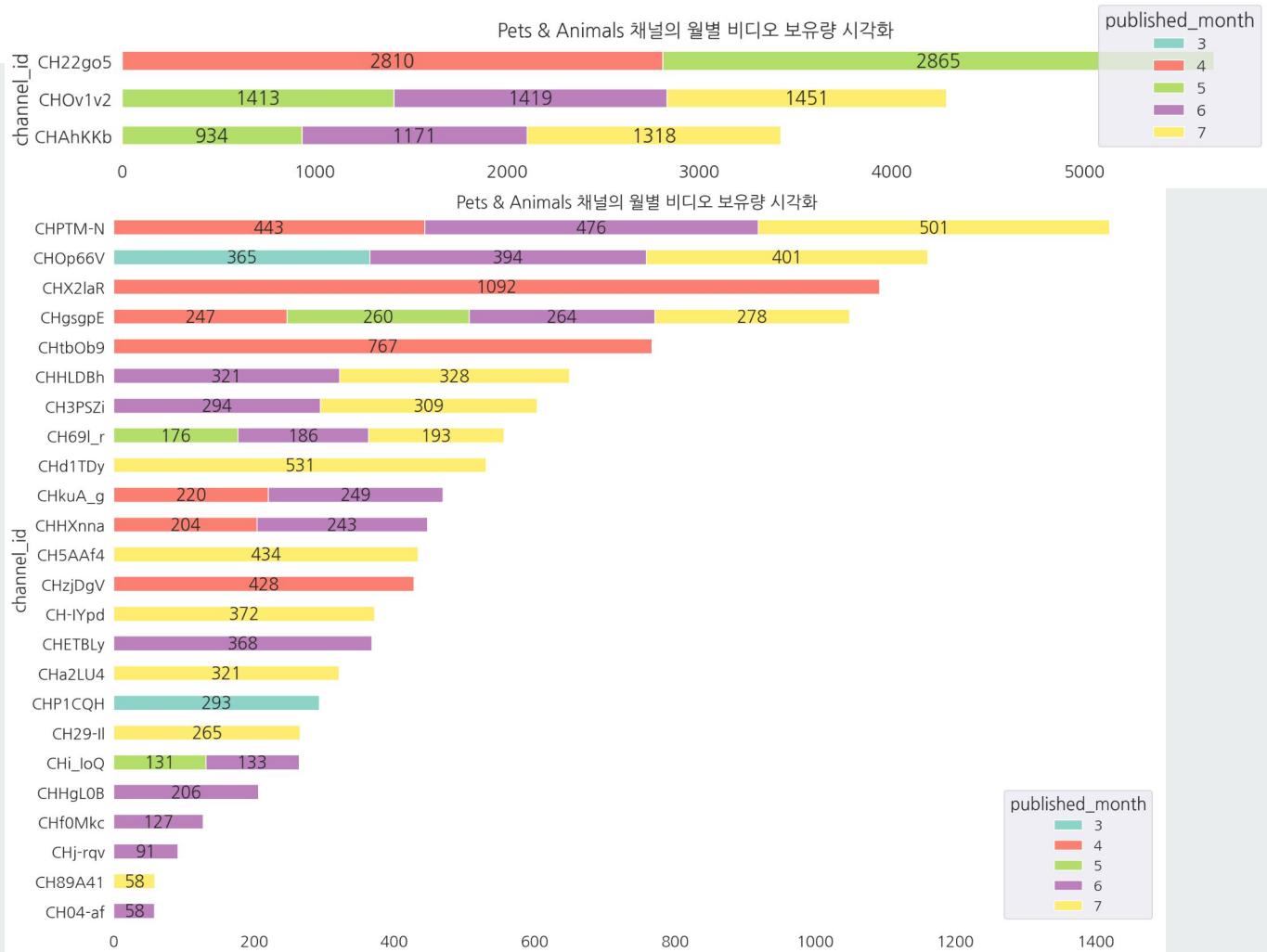




## 10.Pets & Animals

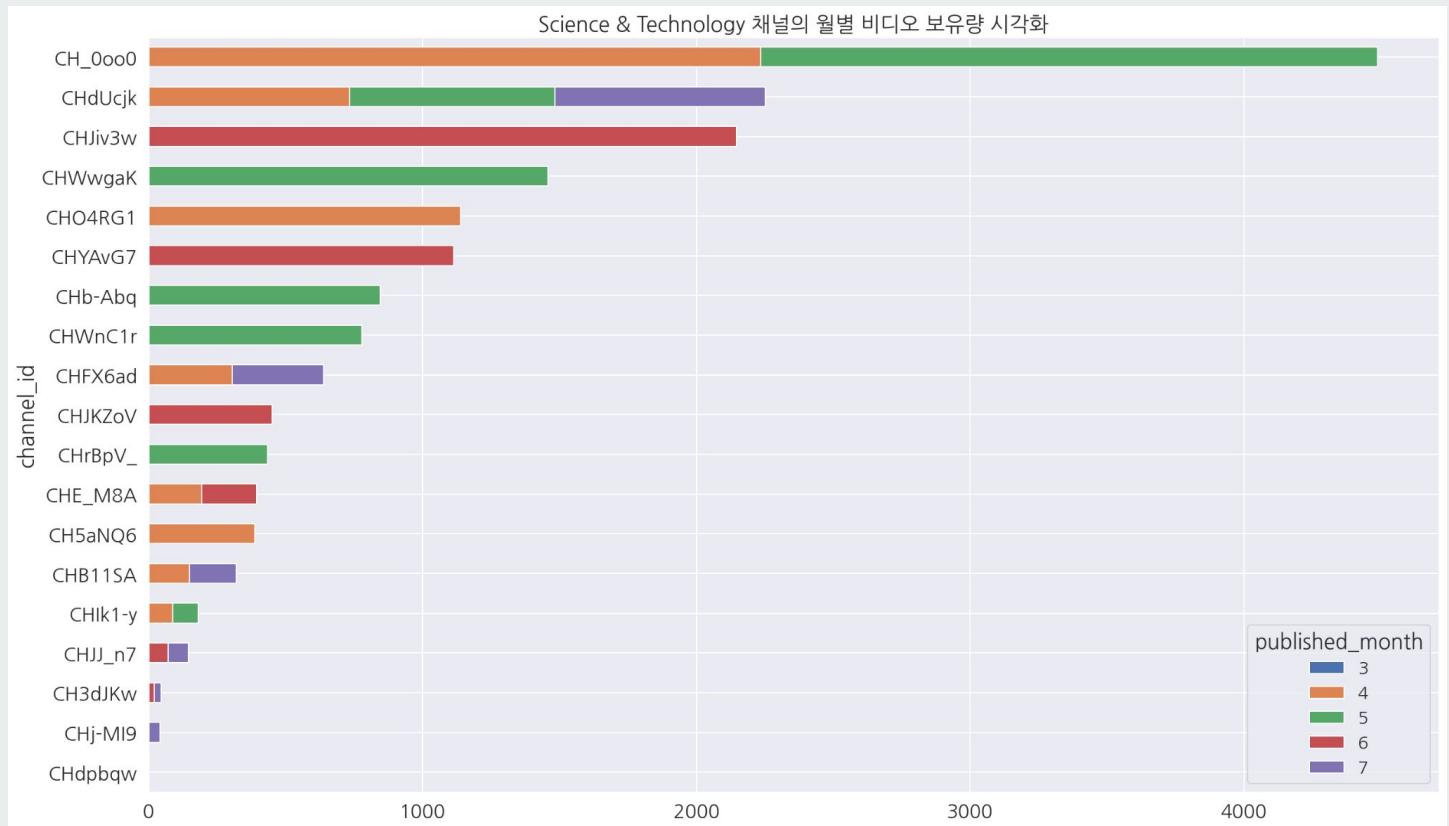
'channel\_id'의 개수 : 27개



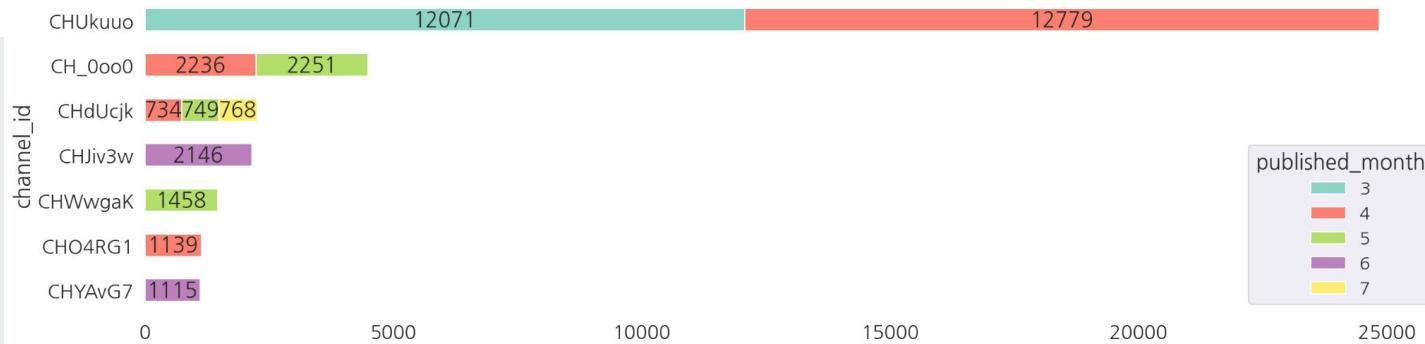


## 11.Science & Technology

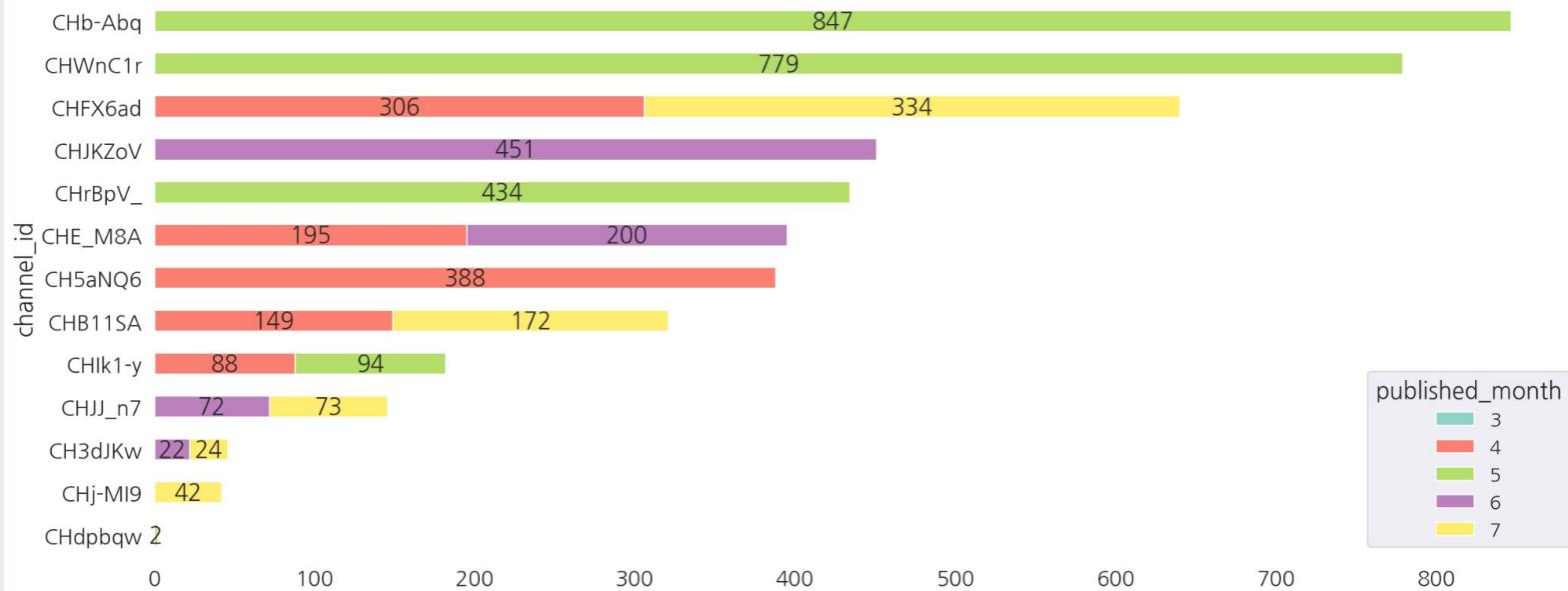
'channel\_id'의 개수 :  
20개 , 아웃라이어  
CHUkuuo 는 제외 후  
시각화합니다.



Science &amp; Technology 채널의 월별 비디오 보유량 시각화

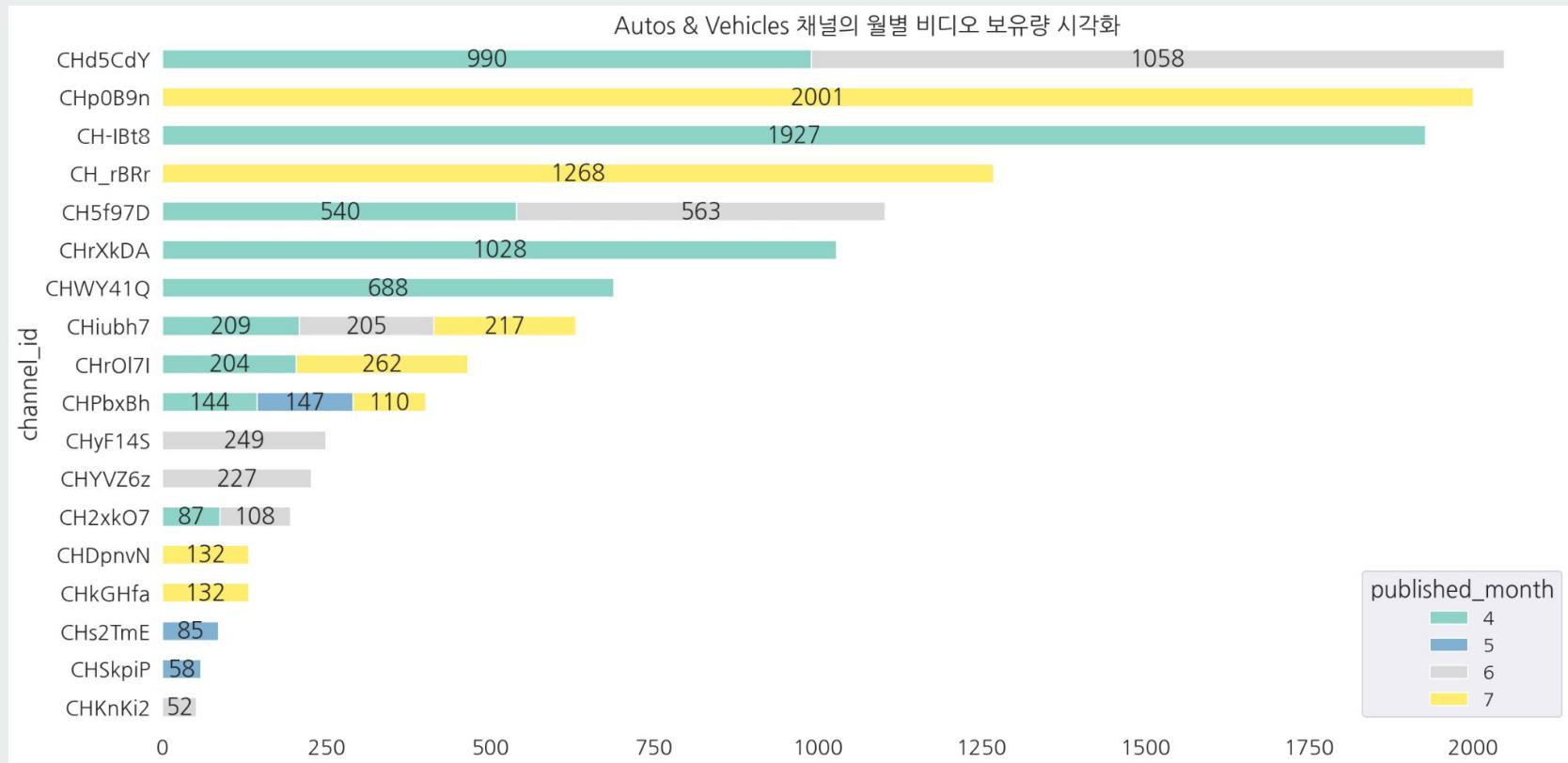


Science &amp; Technology 채널의 월별 비디오 보유량 시각화



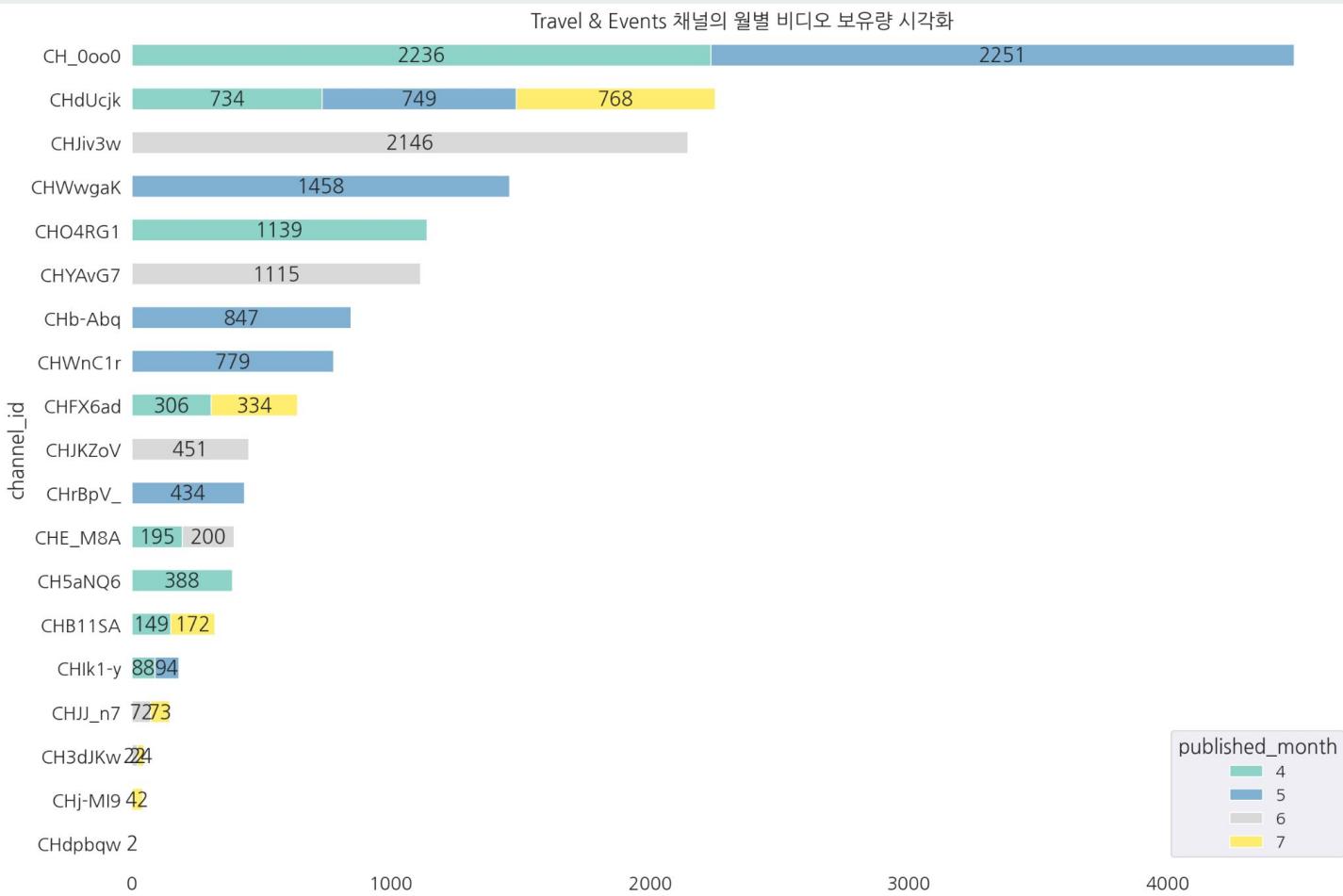
## 12.Autos & Vehicles

'channel\_id'의 개수 : 19개, 아웃라이어 CHH5U89 는 제외 후 시각화합니다.



## 13.Travel & Events

'channel\_id'의 개수 :  
10개



## 14. Nonprofits & Activism

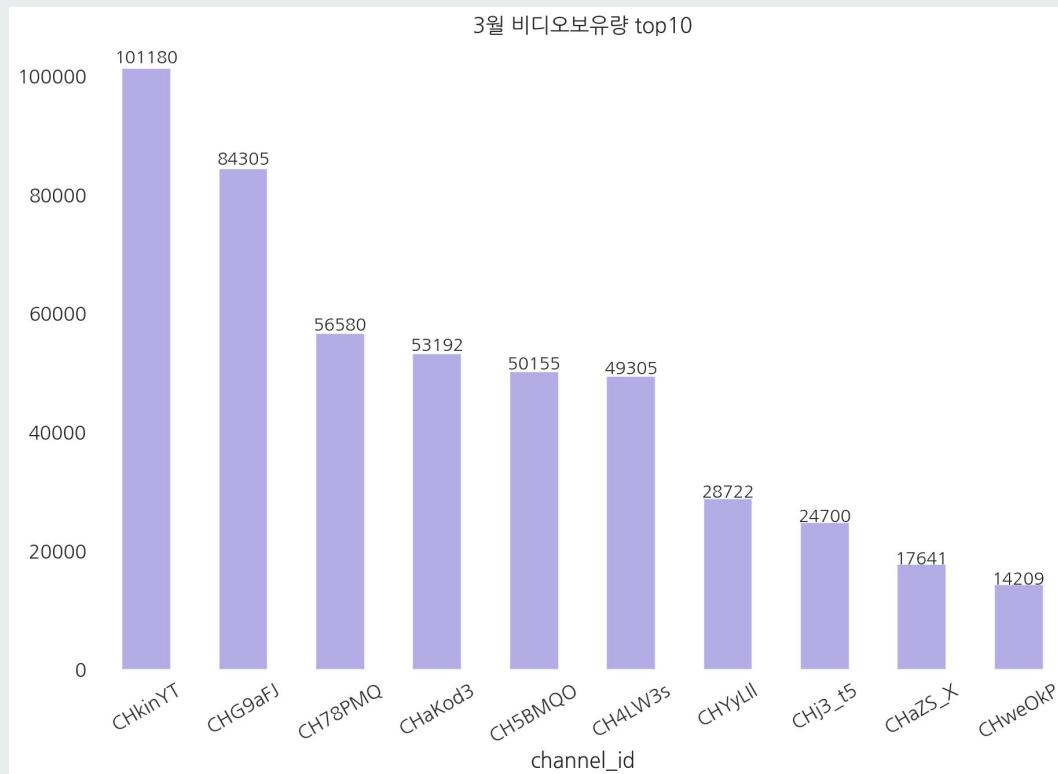
'channel\_id'의 개수 : 1개



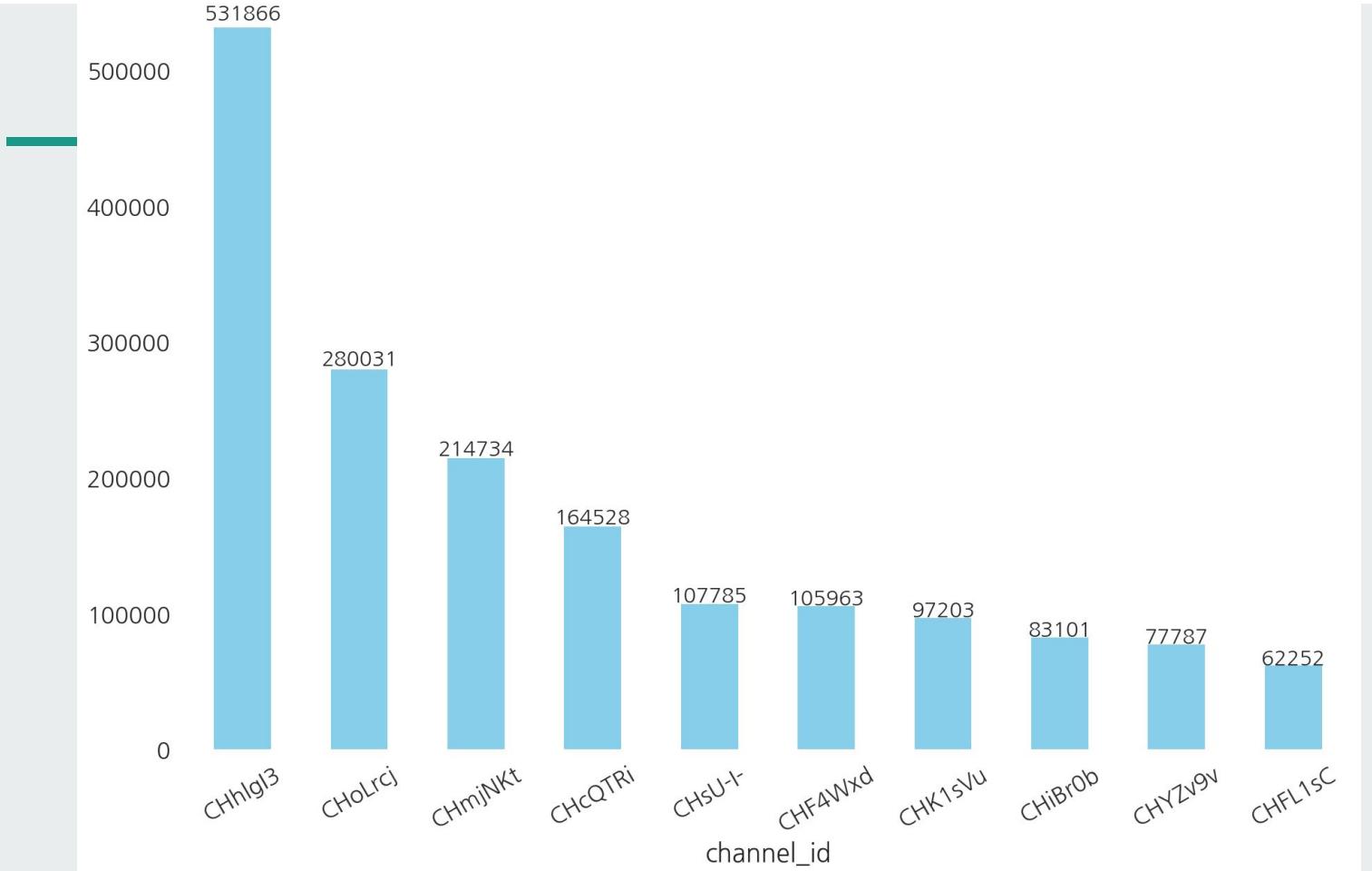
### 💡 더 알아볼 문제

1. 각 카테고리 인기비디오 선정 채널들의 월별 평균 비디오 제작량은 어떻게 될까요?
2. 전체 카테고리에서, 인기비디오 선정 채널들의 월별 평균 비디오 제작량은 어떻게 될까요?

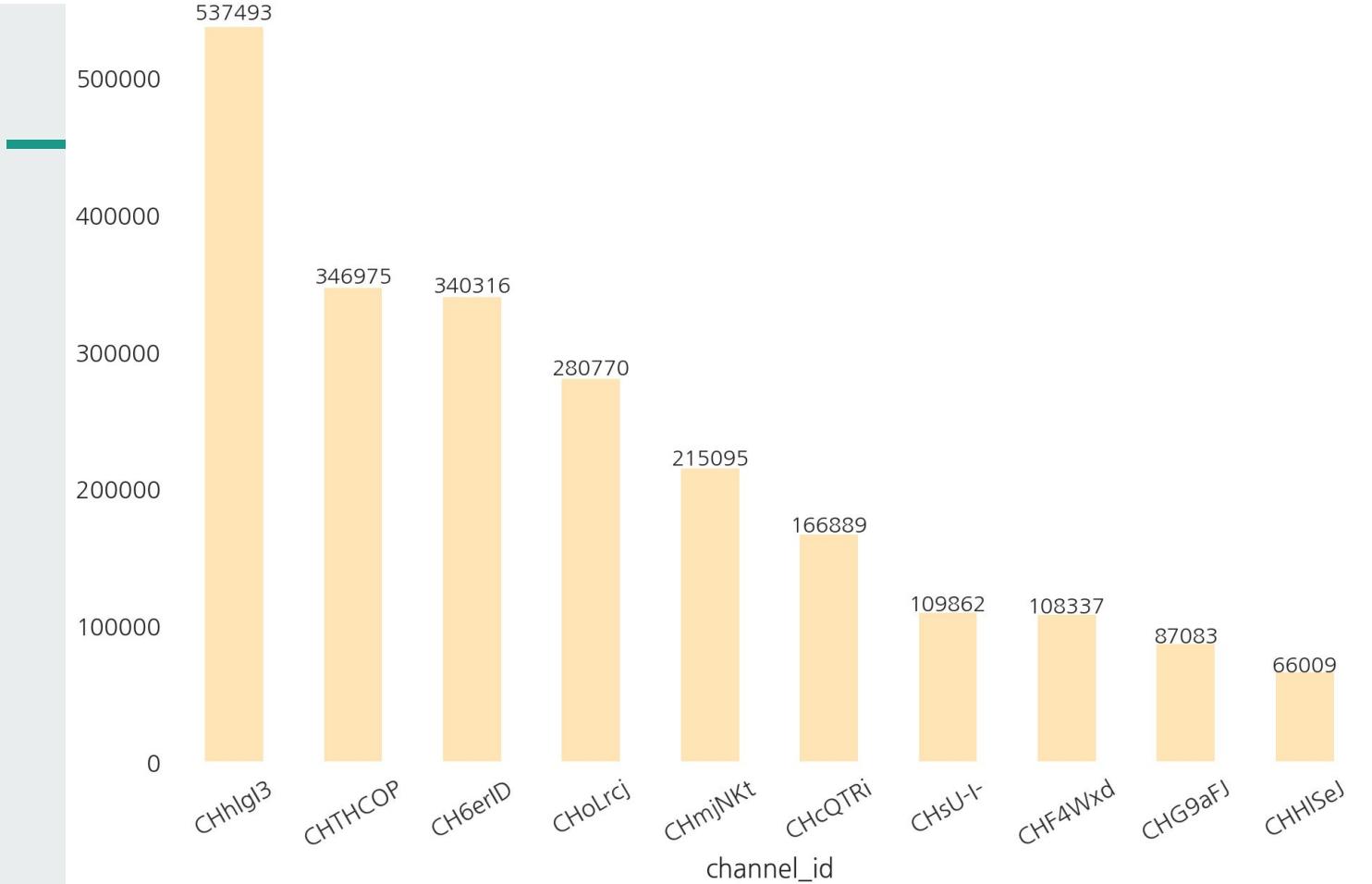
### 3. 월별 TOP10 채널 (분류기준 : 비디오 개수)



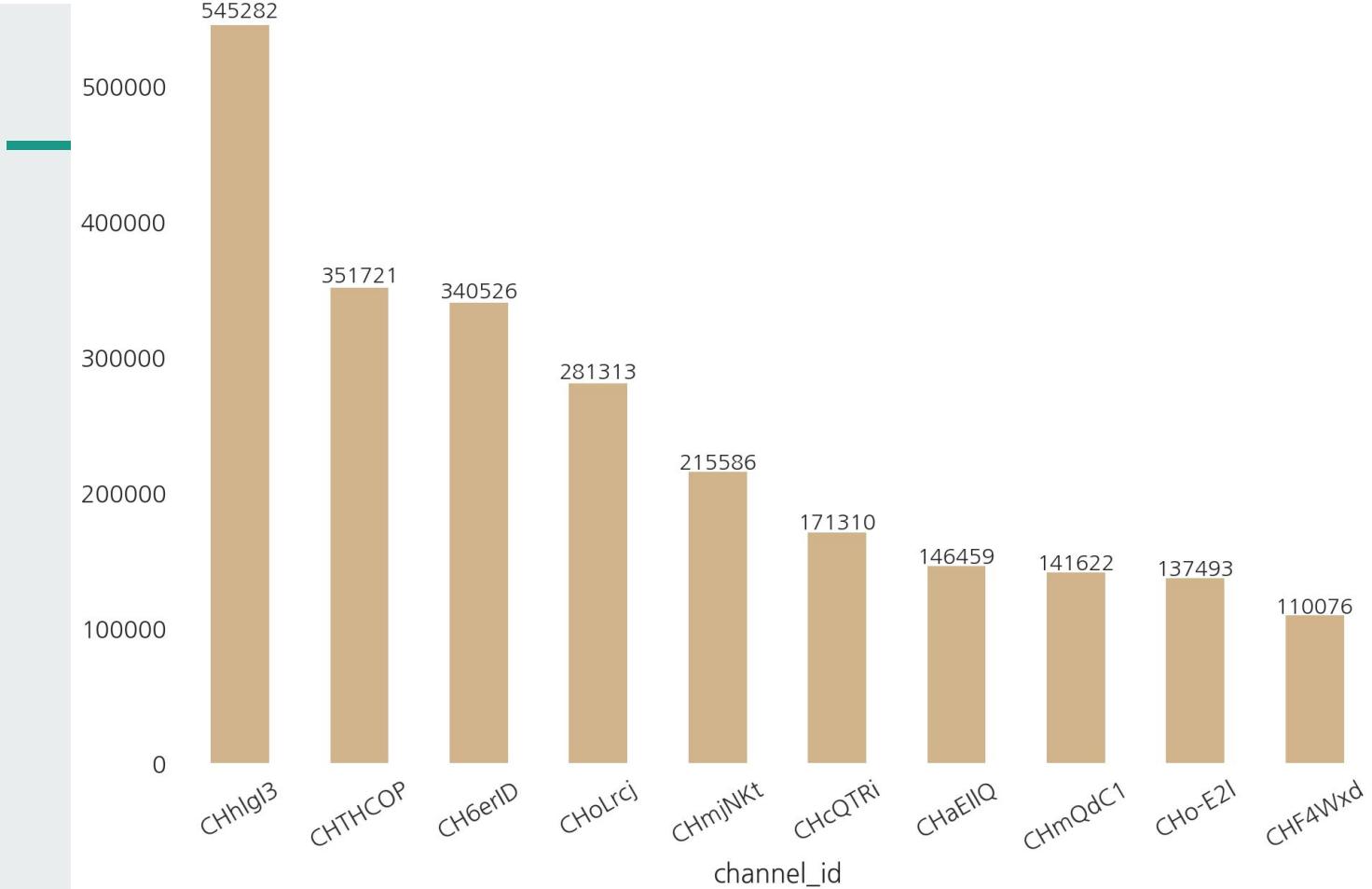
4월 비디오보유량 top10



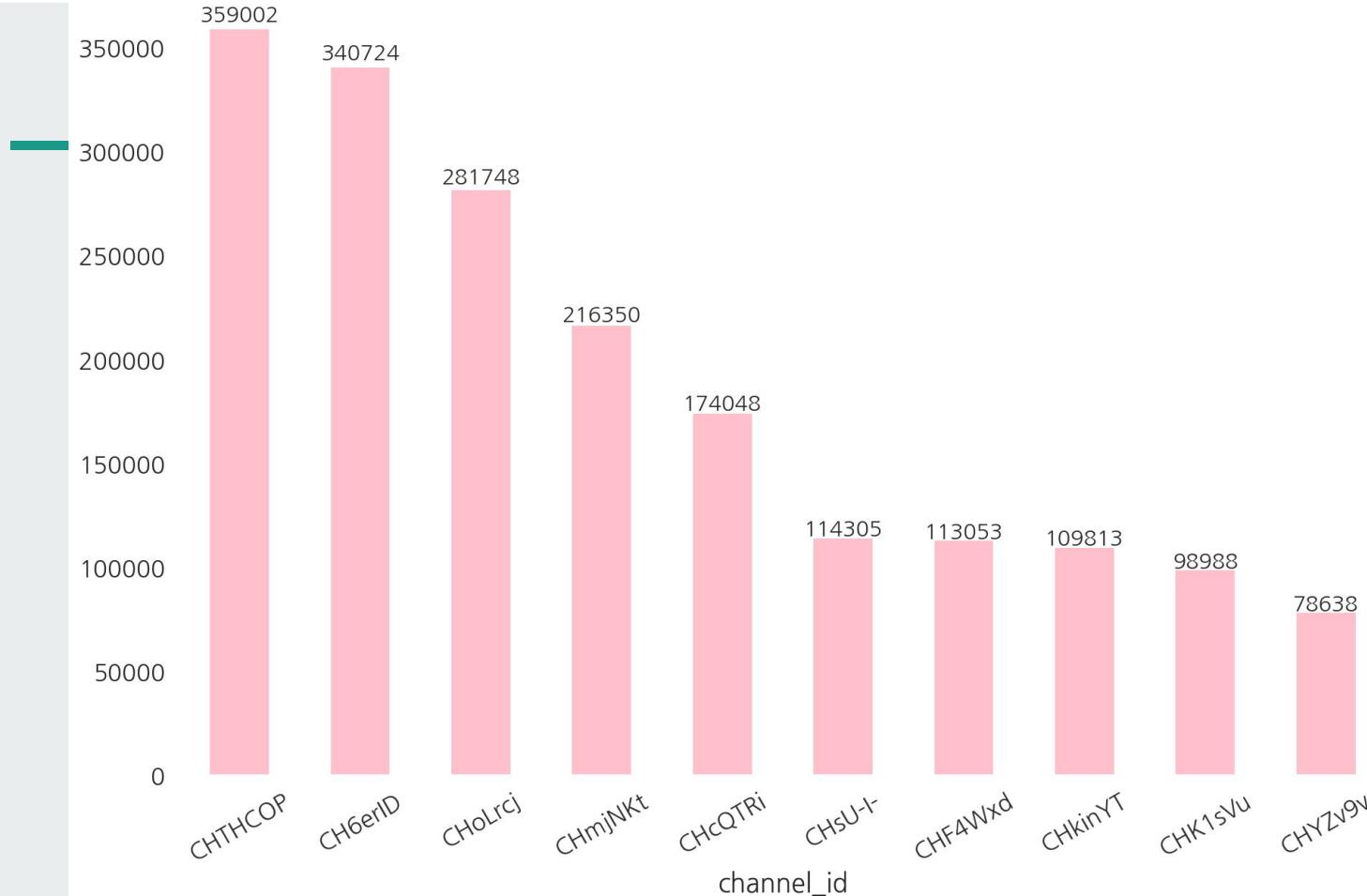
5월 비디오보유량 top10



6월 비디오보유량 top10



7월 비디오보유량 top10

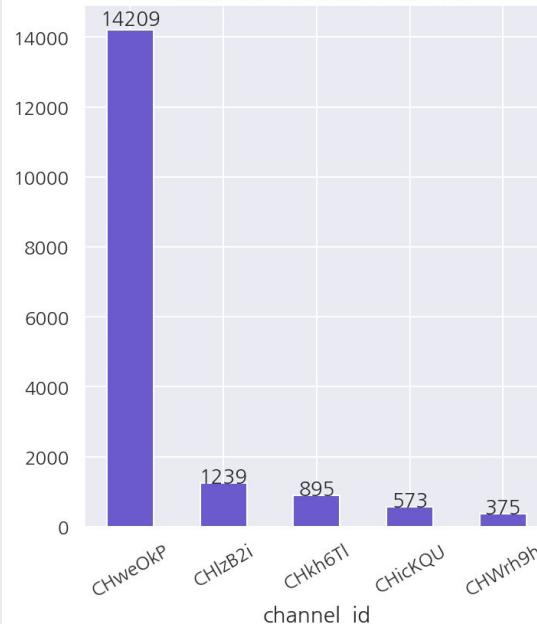


## 4. 주별 TOP5 채널 (분류기준 : 비디오 개수)

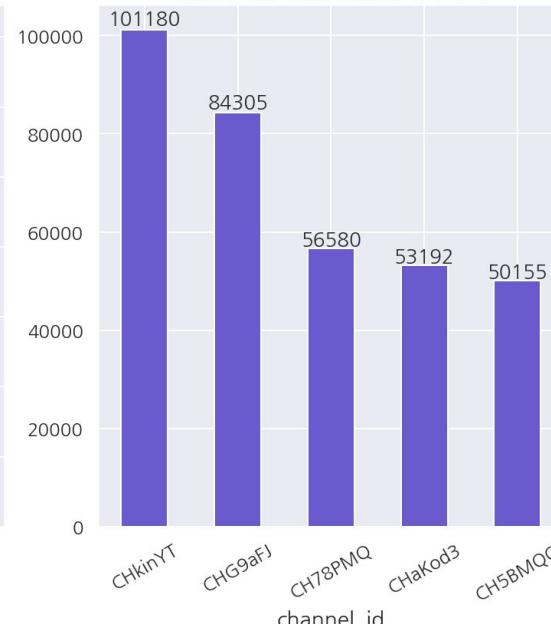


3월 주별 Top5

2021-03-22/2021-03-28

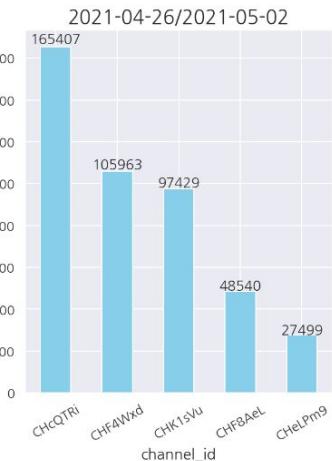
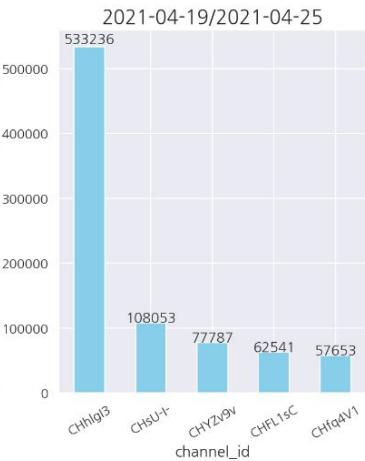
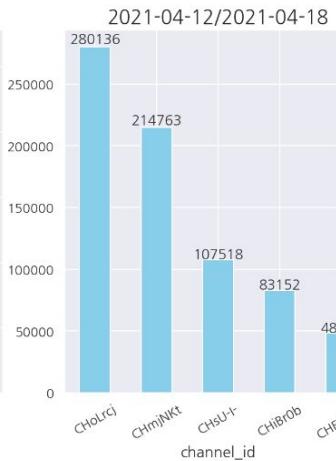
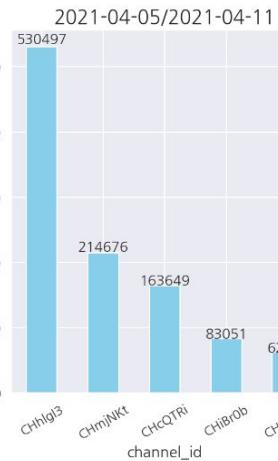
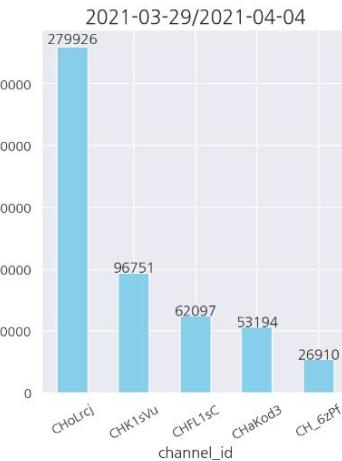


2021-03-29/2021-04-04

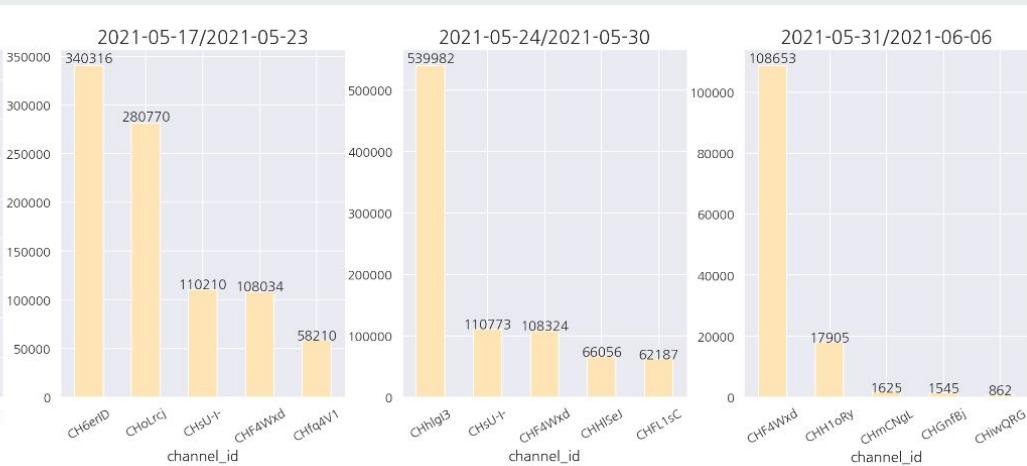
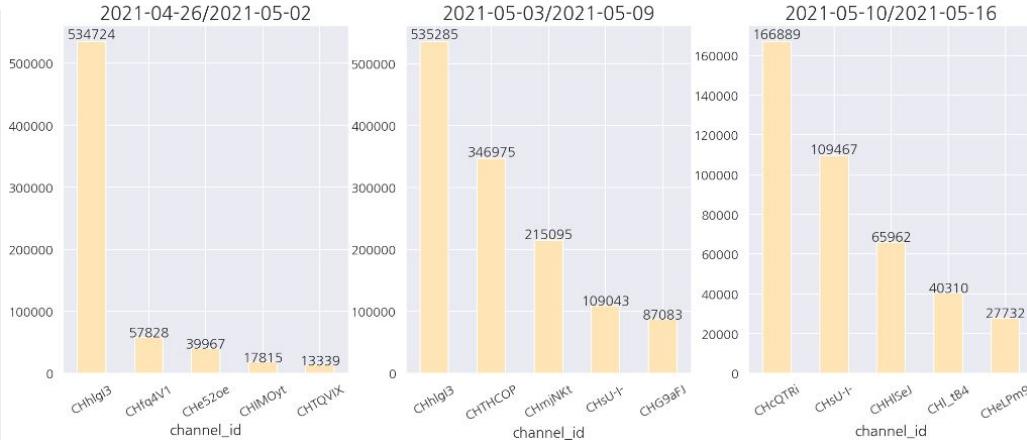




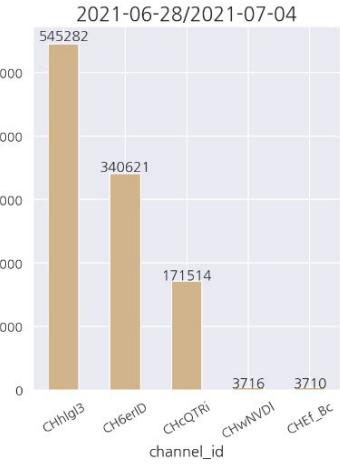
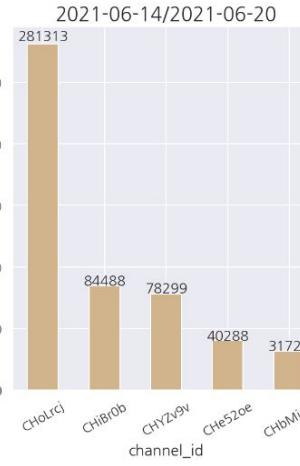
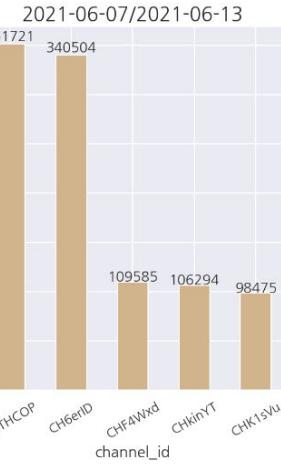
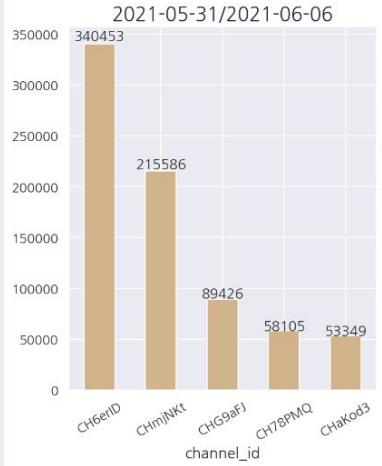
## 4월 주별 Top5



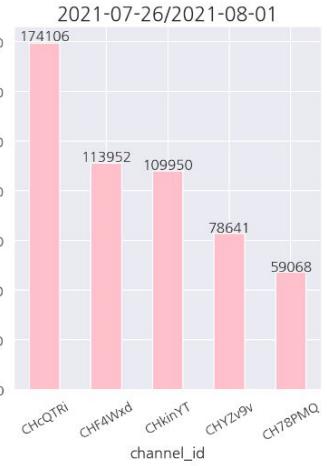
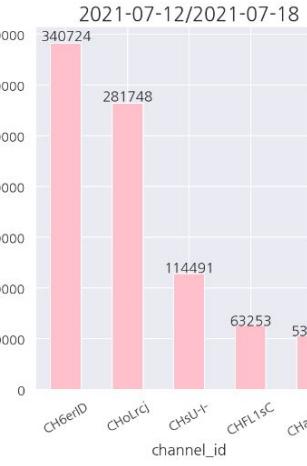
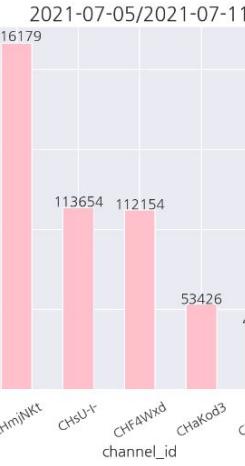
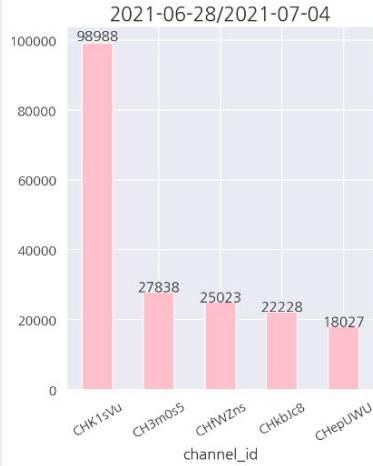
## 5월 주별 Top5



## 6월 주별 Top5



## 7월 주별 Top5

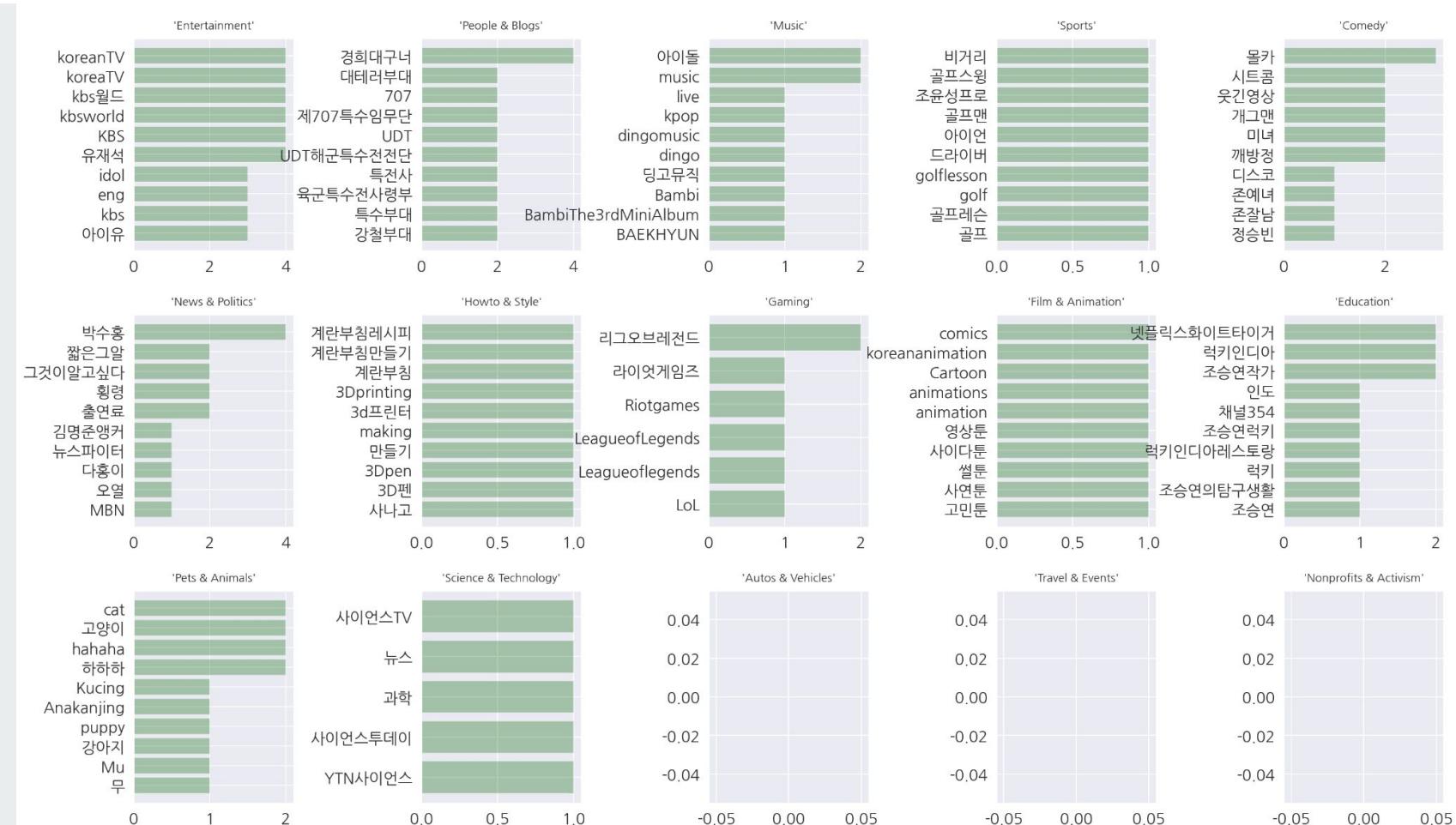




## 5. 월별 카테고리별 태그 키워드 순위

월별 카테고리별로 많이 쓰인 태그 키워드 순위를 1위에서 10위까지 선정하였습니다.

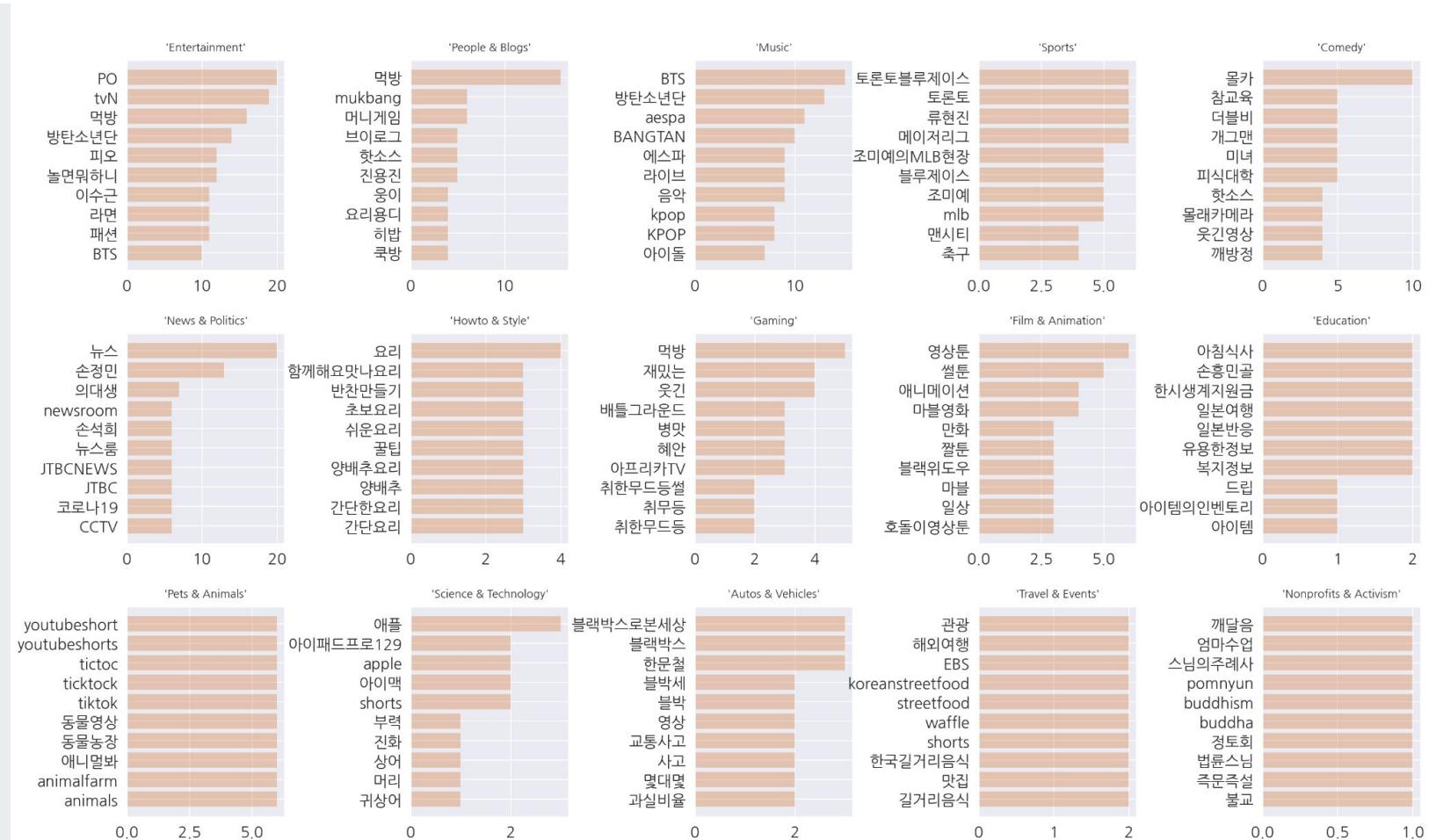
### 3월 카테고리별 Tag 키워드 순위



## 4월 카테고리별 Tag 키워드 순위



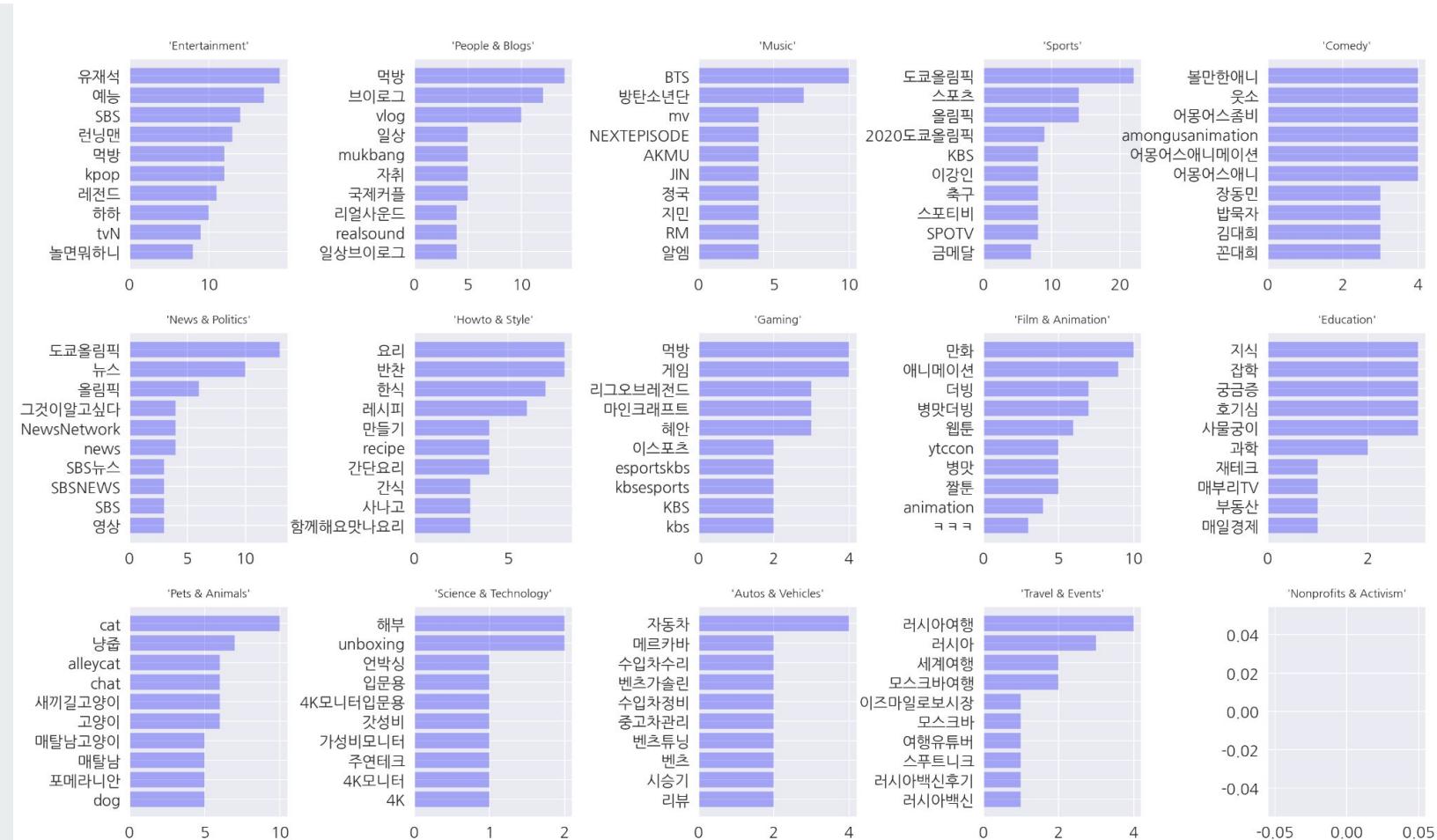
## 5월 카테고리별 Tag 키워드 순위



## 6월 카테고리별 Tag 키워드 순위



## 7월 카테고리별 Tag 키워드 순위



## Q2. 새로운 지표의 개발

---

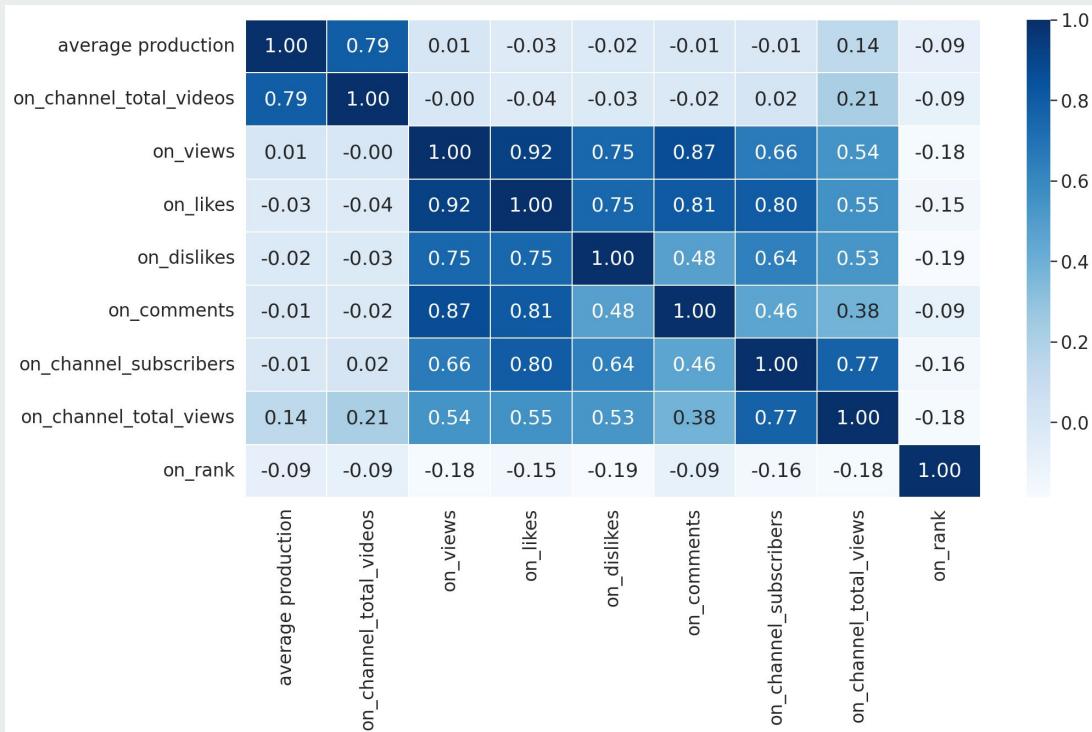
각각의 비디오는 시청자의 호응도(engagement)를 판단할수 있는 객관적인 지표 들이 있음. ex) views, likes, dislikes, comments,... 비디오를 인기 동영상 기준에 부합하도록 분류할수 있는 새로운 지표를 개발하고, 이 지표를 사용하여 engagement 와 어떤 상관관계가 있는지 설명하시오.

 문제의 해석방향을 크게 2가지로 선정하였습니다. 상관계수로 engagement와의 상관관계를 분석하였습니다.

## 문제해석1) 해당 비디오가 ‘인기’ 비디오가 될 것인지 판별하기 위한 지표

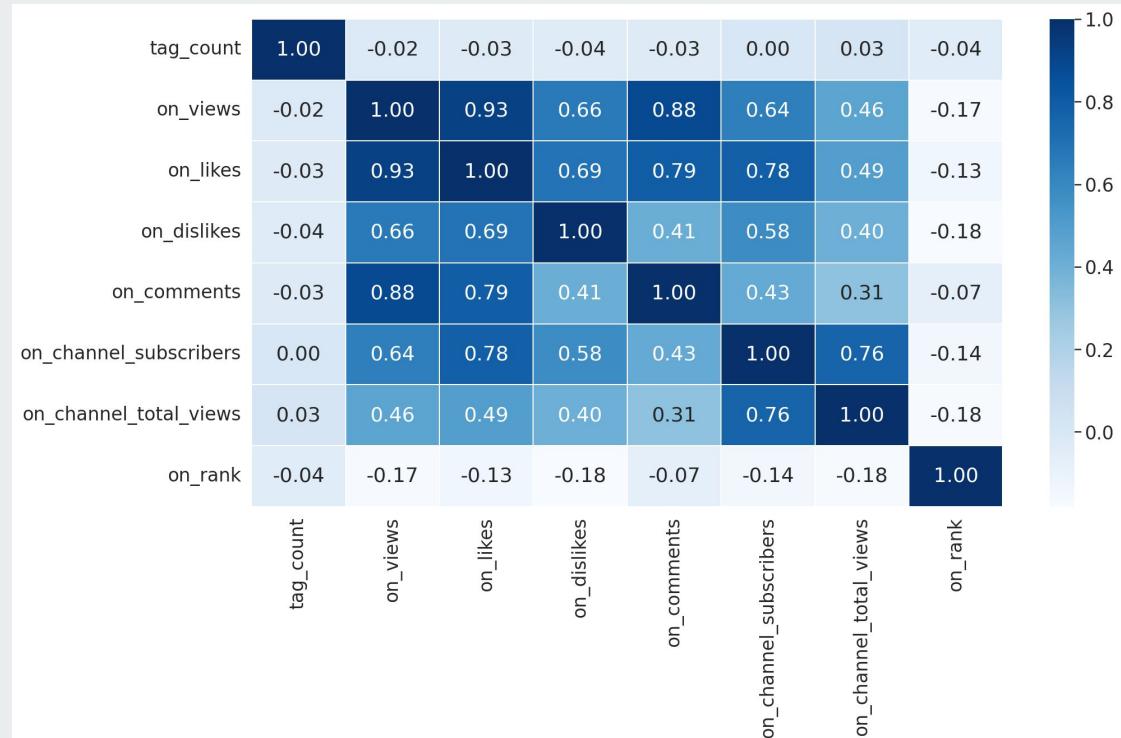
가설1. 인기비디오 선정 채널들의 월평균  
비디오 제작수가 많을수록 인게이지먼트와의  
상관관계가 높을 것이다.(카테고리별 분석)

- 생성한 컬럼 : 인기비디오 선정  
채널들의 월평균 비디오 제작수  
(컬럼명 : average production)
- 결론 : 인기비디오로 선정될 당시  
채널의 전체 조회수와 양의 상관관계를  
가지고 있긴 하지만, 인기비디오 선정  
채널들의 원평균 비디오 제작수가  
높다고 해서 시청자의 호응도가  
높아지는 것은 아니다.(대부분 음의  
상관관계를 지니고 있음.)



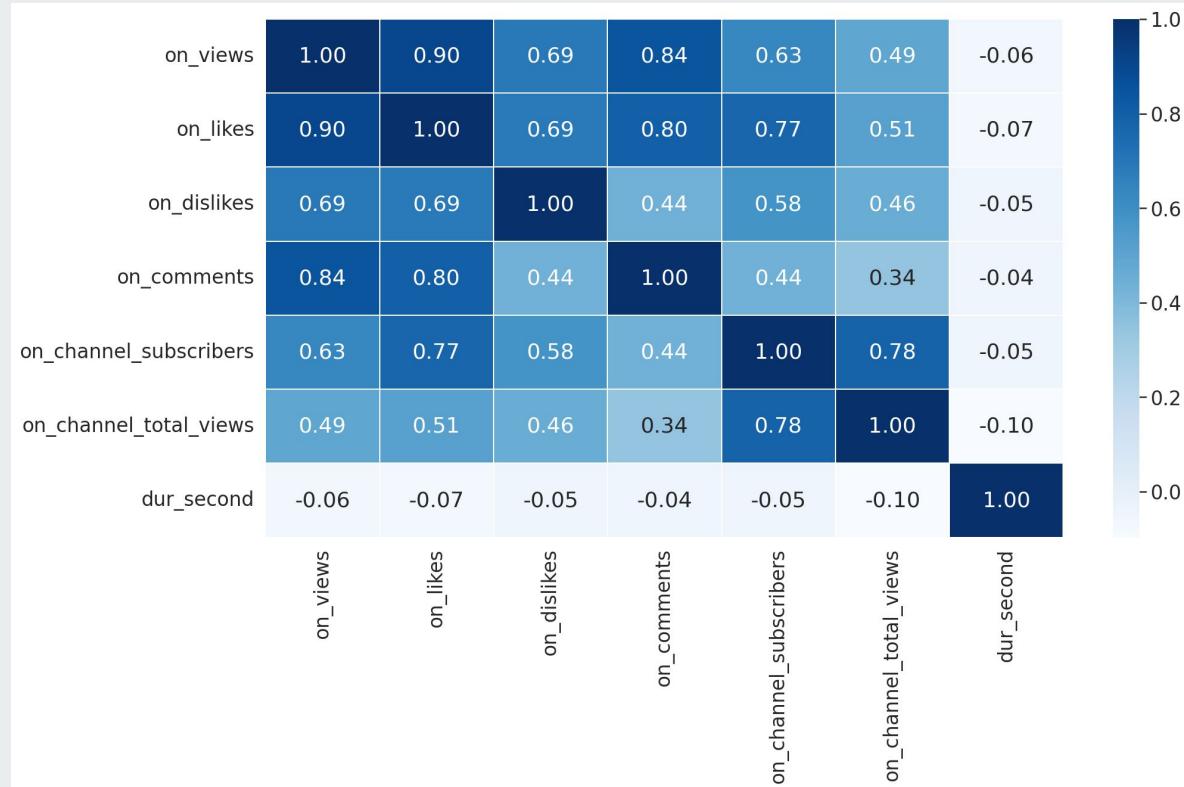
가설2. 해시태그가 상세할 수록  
인게이지먼트와의 연관성이 높을  
것이다.(해시태그 개수와의 연관성)

- 생성한 컬럼 : 해시태그의 개수  
(컬럼명 : tag\_count)
- 결론 : 해시태그를 많이 단다고 해서  
시청자의 호응도가 높아지는 것은  
아니다.



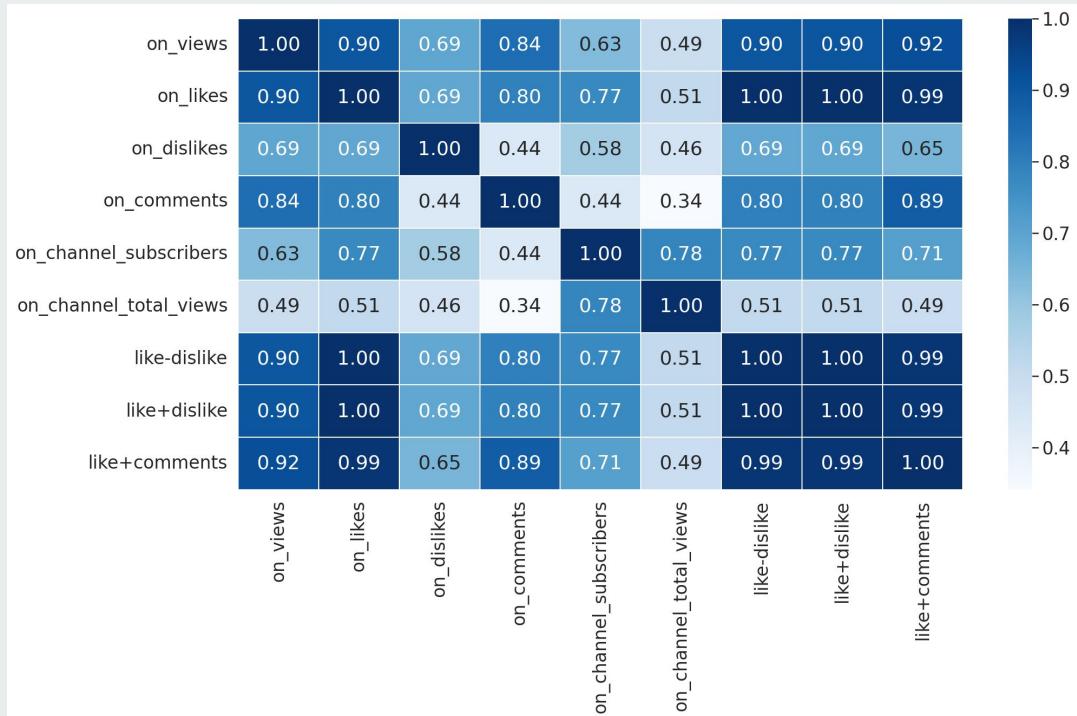
### 가설3. 영상길이와 인게이지먼트와의 연관성

- 생성한 컬럼 : **영상길이**  
(컬럼명 : dur\_second)
- 결론 : 영상의 길이가 길거나 짧은 것은 시청자의 호응도와 별 상관이 없다.



## 문제해석2) 인기비디오를 선정하는 새로운 기준 만들기

- 인기비디오가 된 후 받은 좋아요 수와 싫어요 수의 차이 & 합
- 좋아요 수와 커맨수 수의 합.



 주어진 컬럼들을 활용하여 비디오를 인기 동영상 기준에 부합하도록 분류할수 있는 새로운 지표를 2가지 만들 수 있었습니다.



1. 인기비디오로 처음 기록된 좋아요수 + 싫어요수를 합하거나 뺀 것(결과가 동일하였습니다.)
2. 인기비디오로 처음 기록된 좋아요수 + 코멘트수

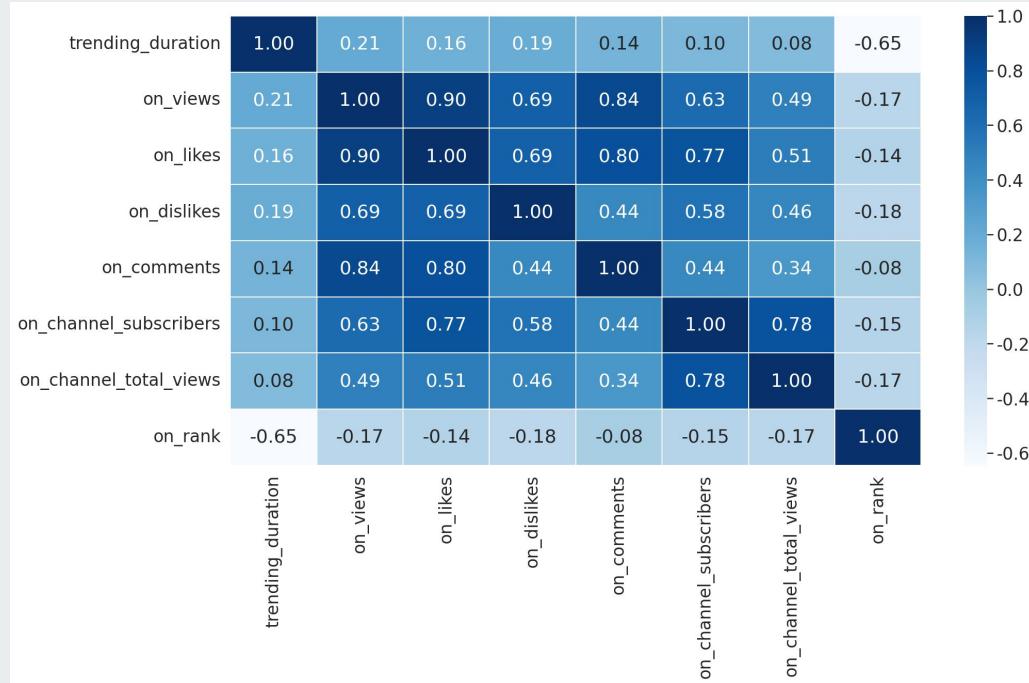
2가지 지표를 상관계수를 통해 기존의 인게이지먼트 지표와 비교하였습니다.

상관계수란 두 변수 중에서 한 변수의 변화가 다른 변수의 변화에 따라 어떤 변화가 일어나는지를 보여주는 지표입니다. 변수간의 관계의 정도와 방향을 하나의 수치로 요약해 주는 지수로.  $-1.00$ 에서  $+1.00$  사이의 값을 가지게 됩니다. 양의 상관관계일 경우에는 (+) 값이 나타나고, 음의 상관관계의 경우에는 (-)값이 나타나는데, 상관계수의 절대값이 높을수록 두 변수간의 관계가 높다고 할 수 있습니다.

`corr()` 메서드를 통해 상관계수를 분석한 결과, 새롭게 만든 두 지표는 인기 동영상에서 처음 기록된 채널의 전체 비디오 조회수의 합 ('on\_channel\_total\_views') 지표를 제외한 나머지 지표에서 0.7이상의 높은 상관계수로 나타났습니다. 기존의 지표와 양의 상관관계를 가지고 있다고 판단되므로, 새로 만든 지표는 인기비디오를 선정하는데 유의미하게 사용될 수 있습니다.

다른 가설1. 인기비디오로 게재되어 있는 시간이 길수록 인게이지먼트와의 상관관계가 높을 것이다.

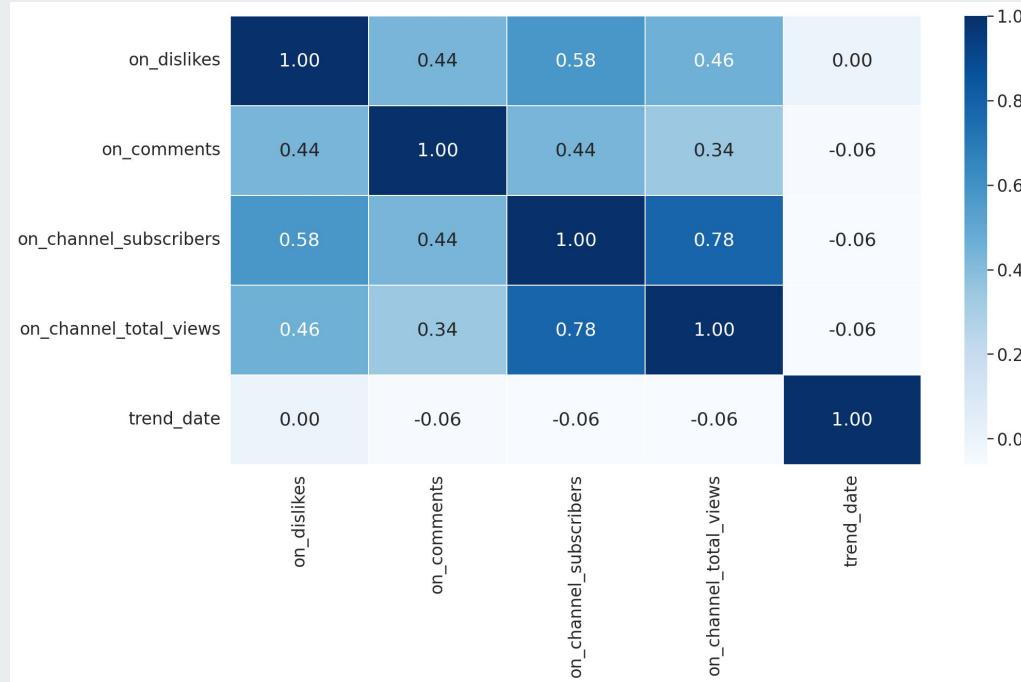
- 생성한 컬럼 : 인기비디오로 게재되어 있는 시간 (컬럼명 : trending\_duration)



- 결론 : 인기비디오로 게재되어 있는 시간이 긴 것은 시청자의 호응도와 미약한 상관이 있다. 인게이지먼트를 나타내는 새로운 지표로 쓸 수는 있겠지만 설득력이 크지는 않다.

## 다른 가설 2. 영상제작일로부터 인기비디오 선정일까지의 기간

- 생성한 컬럼 : 영상제작일로부터 인기비디오 선정일까지의 기간(컬럼명 : trend\_date)



- 결론 : 영상제작일부터 인기비디오 선정일까지의 기간은 시청자의 호응도와 관계가 거의 없는 지표이다.