YouTube Trending Video Analysis (3.29 ~ 7.25)

8팀 정다현 <프리온보딩 AI/ML 기업과제>

Trending aims to surface videos that a wide range of viewers would find interesting. By combining signals such as view counts, the “temperature” - how quickly it generates views, age of the video etc., YouTube list up some of rather newly uploaded videos as ‘Trending’ and tries to catch the breadth of what’s happening in YouTube and around the world. YouTube states Trending is for showcase diversity of creators, and introduce surprising or novel contents. What stats could be additionally used, other than they are using now?

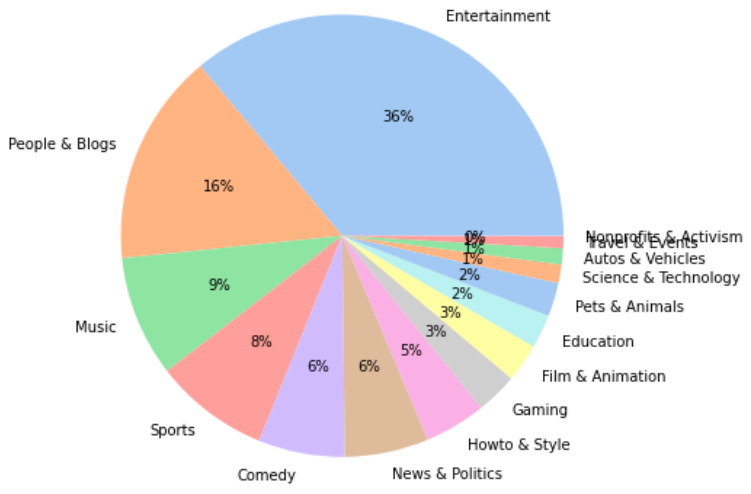
This analysis will cover channels with most trending videos by category, and also discover monthly, weekly top channels. By exploring the data, the main goal of the analysis is to develop a new index to classify certain video as Trending. What makes a video to be considered ‘hot’?

Data description & EDA

The data used in this analysis is Korea trending videos during March 25th 2021 to July 29th 2021. The data was retrieved once a day, approximately for 4 months. It constitues a total of 2,644 videos, 940 unique channels, with 25 features such as title, category, description, tags, duration. On and off trending stats are also included, such as comparing views when the video is on and off trending. The following table shows an example of the data for each video.

The categories are total 15, and the top 5 categories with the most trending videos are ‘Entertainment’, ‘People & Blogs’, ‘Music’, ‘Sports’, and ‘Comedy’. There are big differences in ratio between categories, as shown in the below chart. ‘Entertainment’ accounts for the largest portion with 36%.

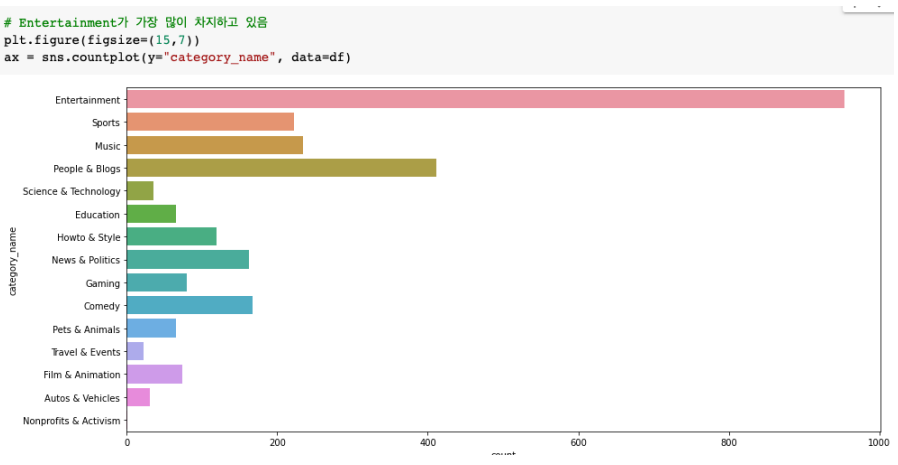
```
#create pie chart
lab = df['category_name'].value_counts().reset_index()['index']
col = df['category_name'].value_counts().reset_index()['category_name']
plt.figure(figsize=(7,7))
colors = sns.color_palette('pastel')
plt.pie(col, labels=lab, colors=colors, autopct='%.0f%%')
plt.show()
```




```
df.head(1).transpose()
```

	0
video_id	V-0db
channel_id	CH49ta0
published_date	2021-07-01
category_name	Entertainment
duration	PT8M20S
tags	SiriusXM|Sirius XM|Sirius|SXM|BIGHIT|빅히트|방탄소년단...
description	BTS performs their hit songs ‘Dynamite’ and ‘B...
on_trending_date	2021-07-03
off_trending_date	2021-07-04
on_rank	13
off_rank	28
on_views	1659484
off_views	1912983
on_likes	270004
off_likes	282204
on_dislikes	792
off_dislikes	1014
on_comments	10373
off_comments	10720
on_channel_subscribers	1080000
off_channel_subscribers	1080000
on_channel_total_views	685992413
off_channel_total_views	687485021
on_channel_total_videos	5947
off_channel_total_videos	5950

```
# Entertainment가 가장 많이 차지하고 있음
plt.figure(figsize=(15,7))
ax = sns.countplot(y="category_name", data=df)
```

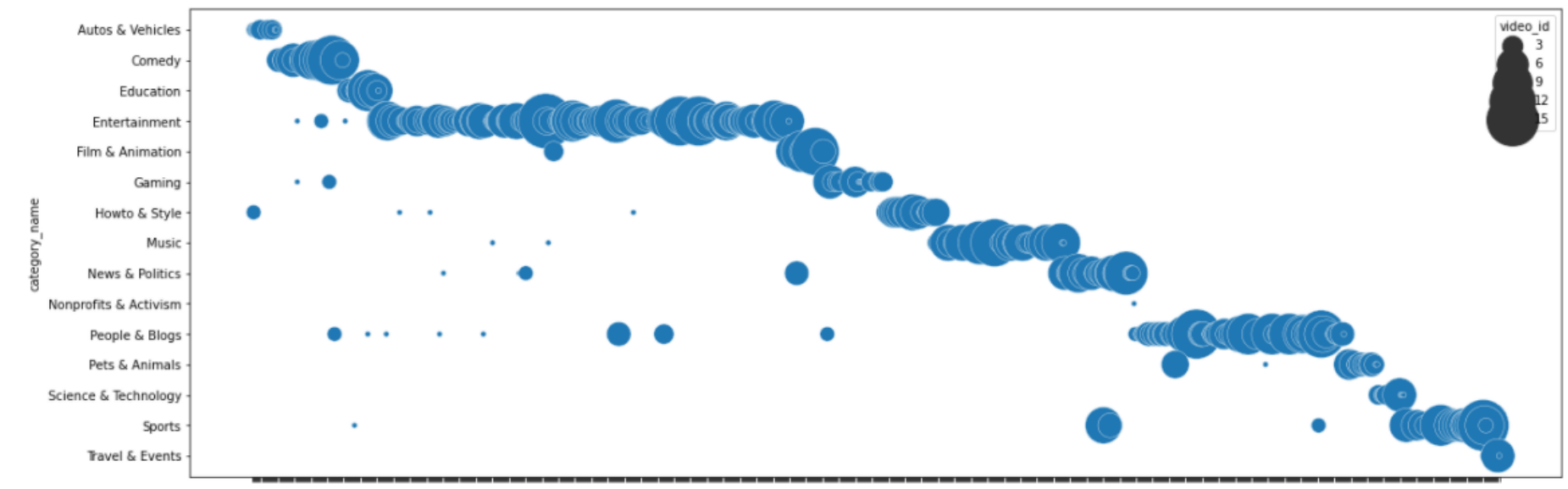


Channels with most trending videos

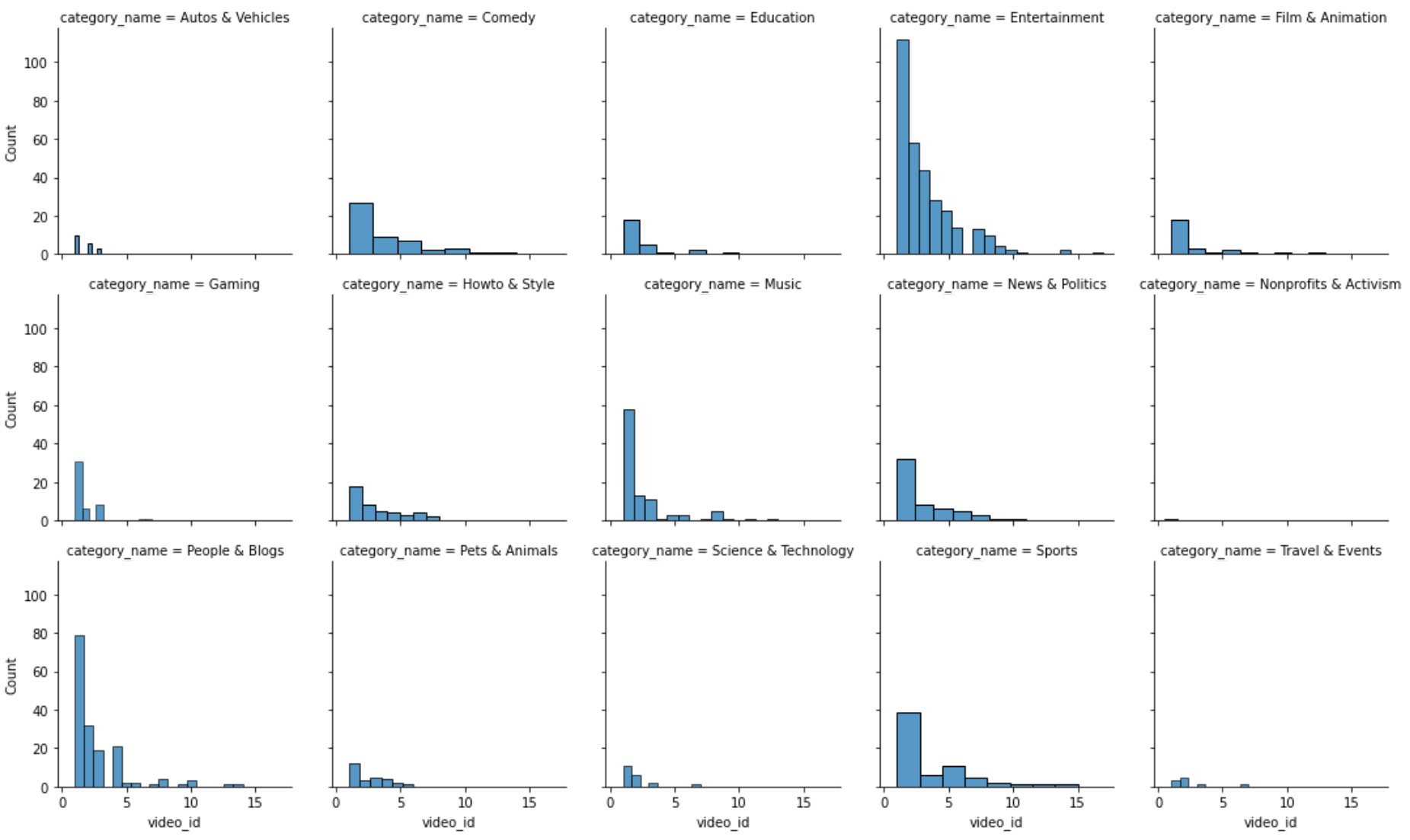
Q1-1. Entire period > trending videos per channel by category

Which channels produced more trending videos? The following bubble chart tells us the patterns of the channel by category. Xlabel is each channel, and the size of the bubble is the amount of videos that channel make it to trending. As you can see, most of the channels produce trending videos in only one category, but some channels produce in two or more categories. For example, channel id ‘CHH3mJ-’ had trending in ‘News & Politics’, ‘Film & Animation’. This channel was Joong-Ang Daily.

```
] plt.figure(figsize=(20,7))
  ax = sns.scatterplot(
      data=uploads_by_chan_cat, x="channel_id", y='category_name', size="video_id", legend=True, sizes=(20, 2000));
  ax.set(xticklabels=[]);
  ax.set(xlabel=None);
```

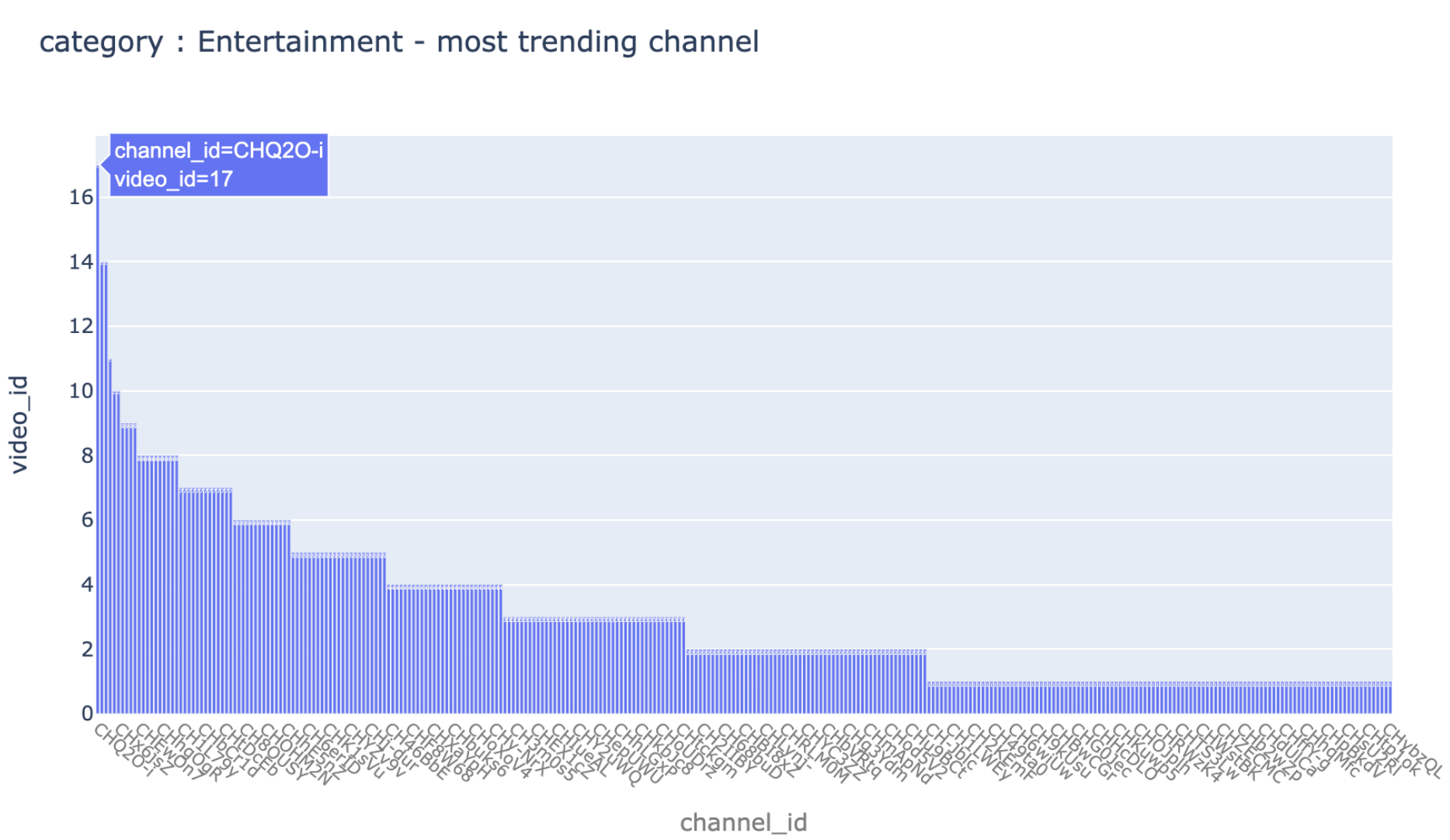


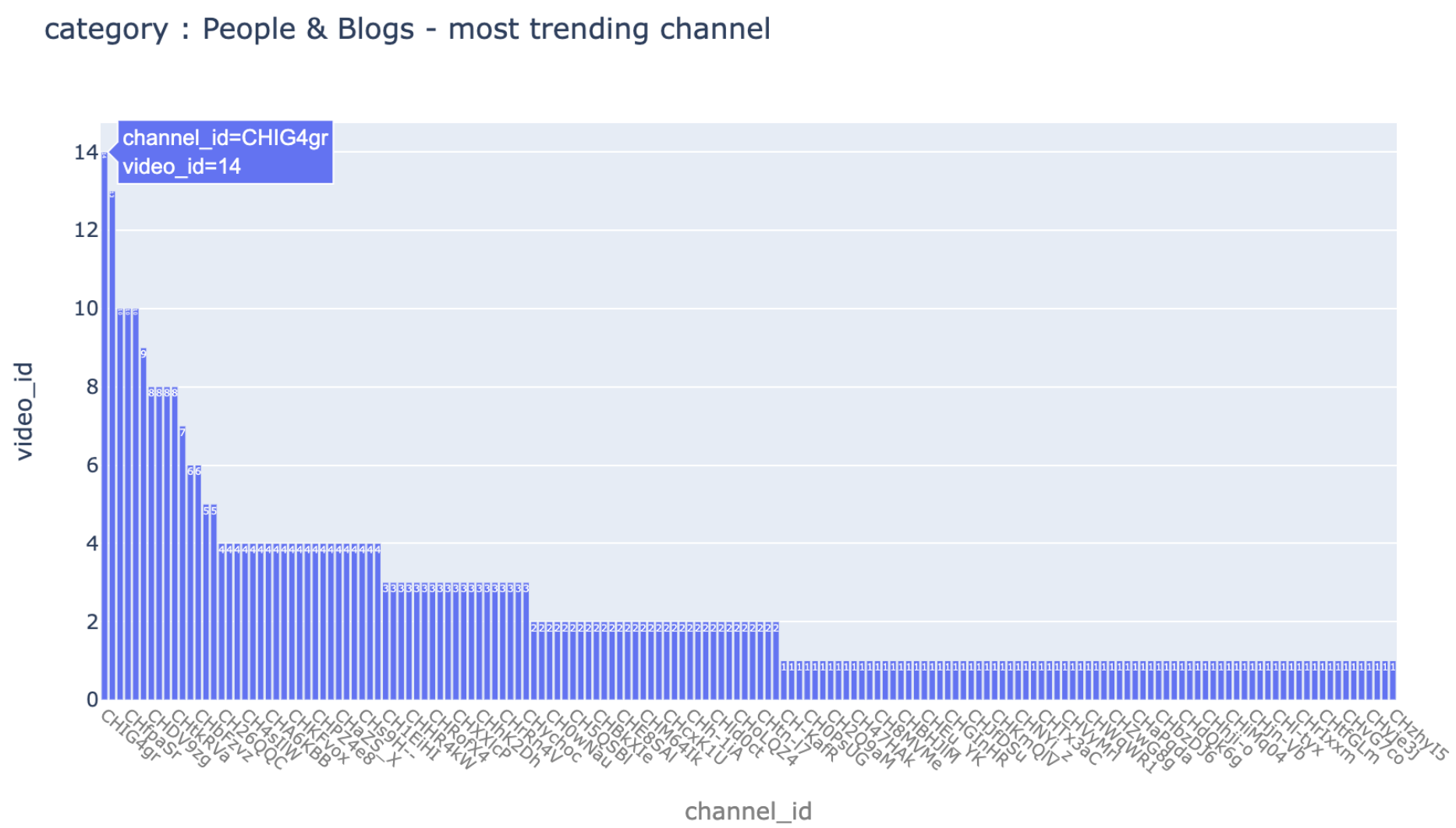
The following grid chart is a histogram per categories, how the videos per channels are distributed. as you can see there are single channels that produce more than 10 videos, which could be considered super popular channels. These super channels belongs to such categories like Comedy, Entertainment, Film&Animation, Music, News&Politics, People&Blogs, and Sports.



Now let’s look at the categories individually. I grouped the 15 categories by their ratio.

1. Entertainment (36%)
2. People & Blogs (16%)
3. Music, Sports (9,8%)
4. Comedy, News & Politics, Howto & Style (6,6,5%)
5. Gaming, Film & Animation, Education, Pets & Animals, Science & Technology, Autos & Vehicles, Travel & Events, Nonprofits & Activism





Most trending channel by Category

Most trending channel by Category

Howto & Style

Most trending channel by Category

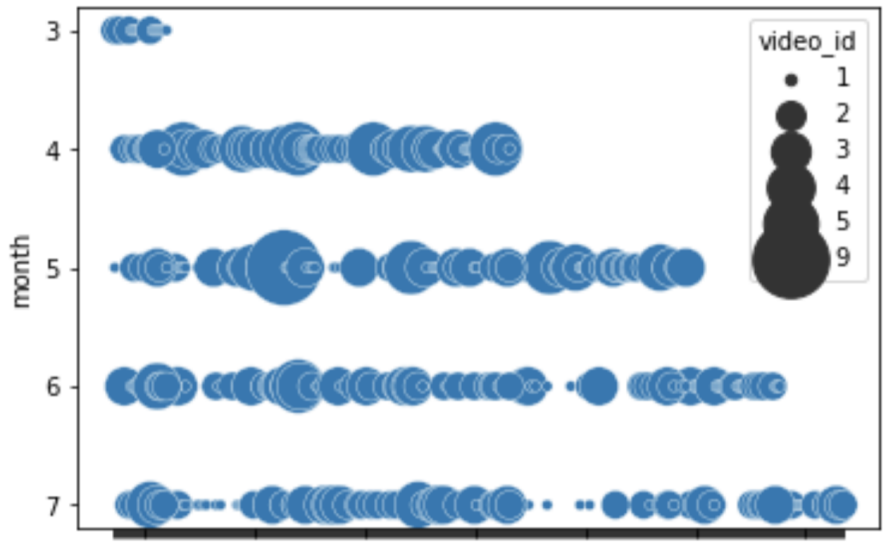
The categories share similar patterns. Each category has 2,3 top channels that produces the most trending videos. Average videos per Categories range in 2 to 3 videos. The channel that produce the most is `CHQ2O-i`, with 17 videos during 4 months in entertainment category.

Q1-2. Monthly > trending videos per channel by category

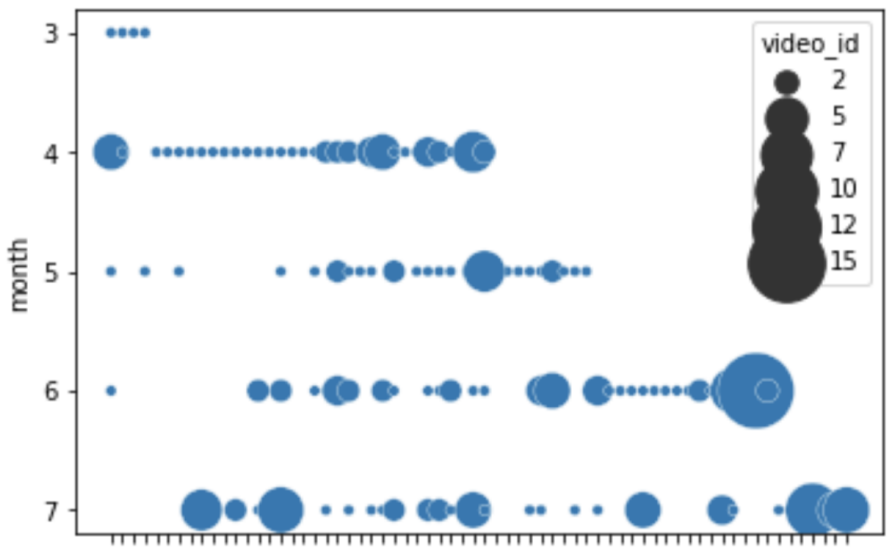
Following Bubble chart tells us which channels has the most trending videos per month. Channel Id is on the x-axis, Month is on the y-axis. The size represent how many videos each channel produce. Overall, it seems popular channels consistently produce trending videos. But there are some interesting channels which videos are focused on certain period of time. For example, the big bubble on ‘Sports’ at june, is the channel that produced all its trending videos only at june. We can infer that channel maybe covered a special event, such as olympics or so.

Entertainment

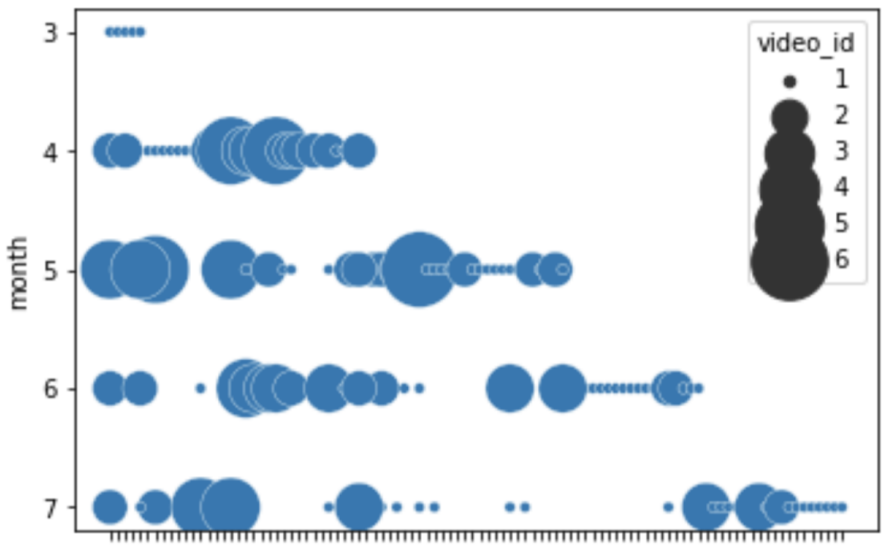
Sports



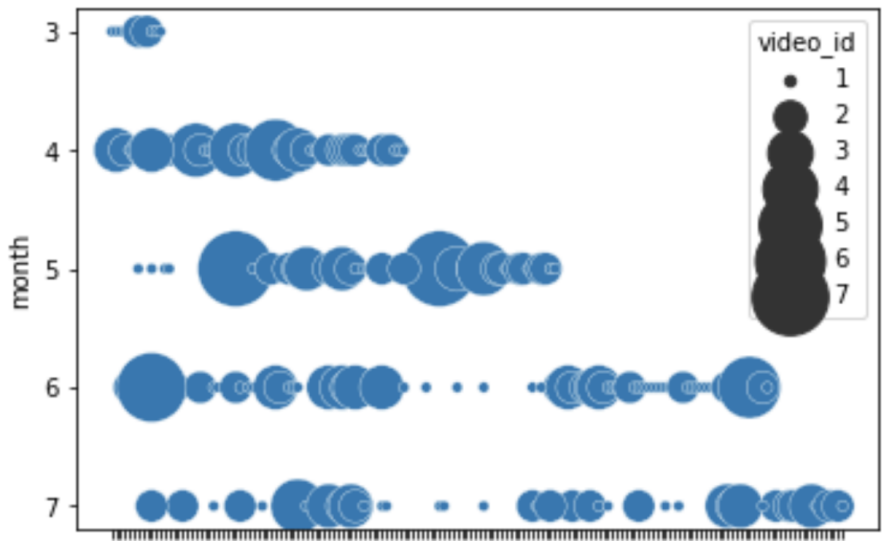
Music



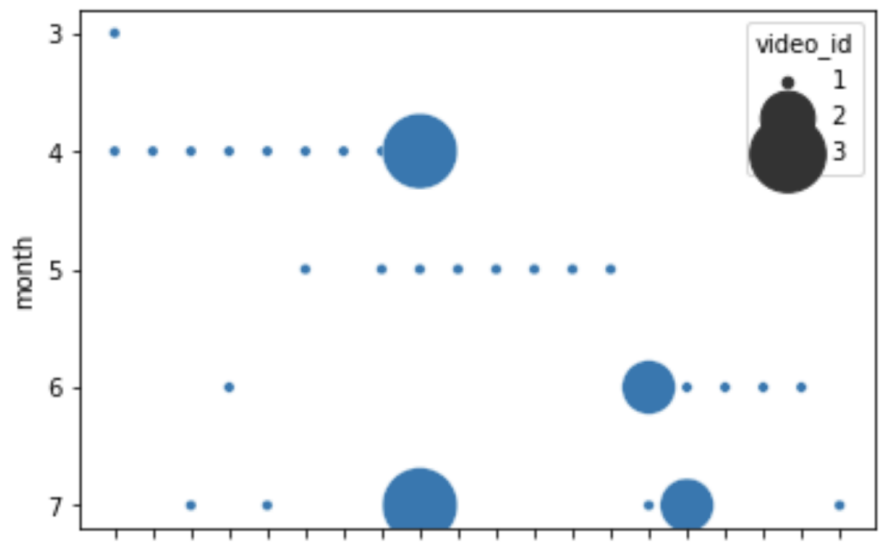
People & Blogs



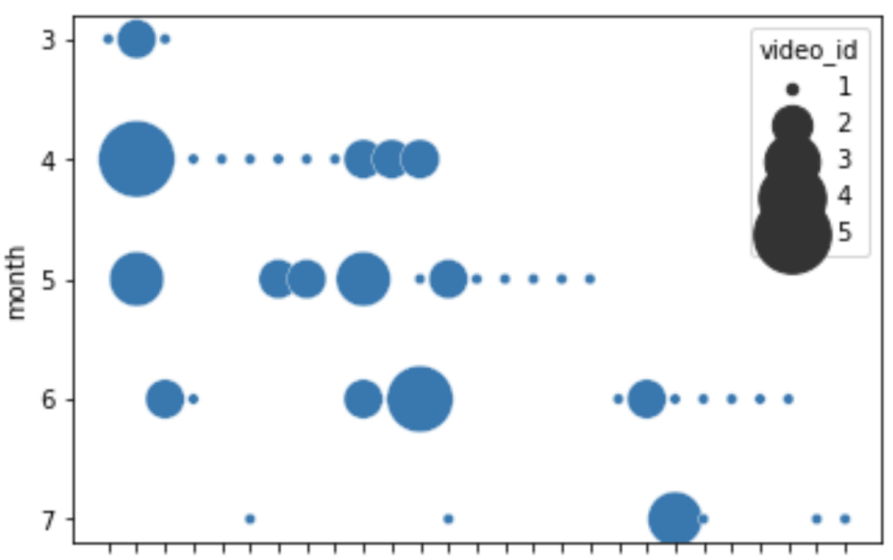
Science & Technology



Education

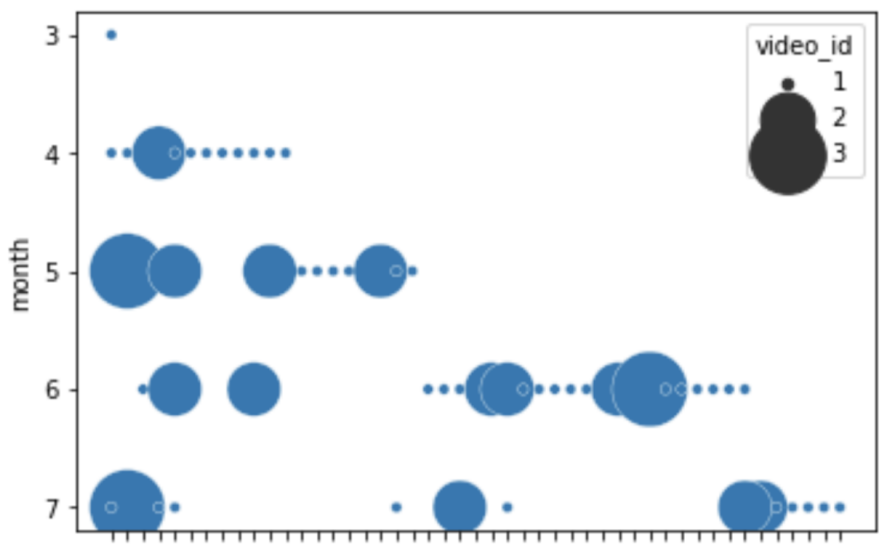


Howto & Style

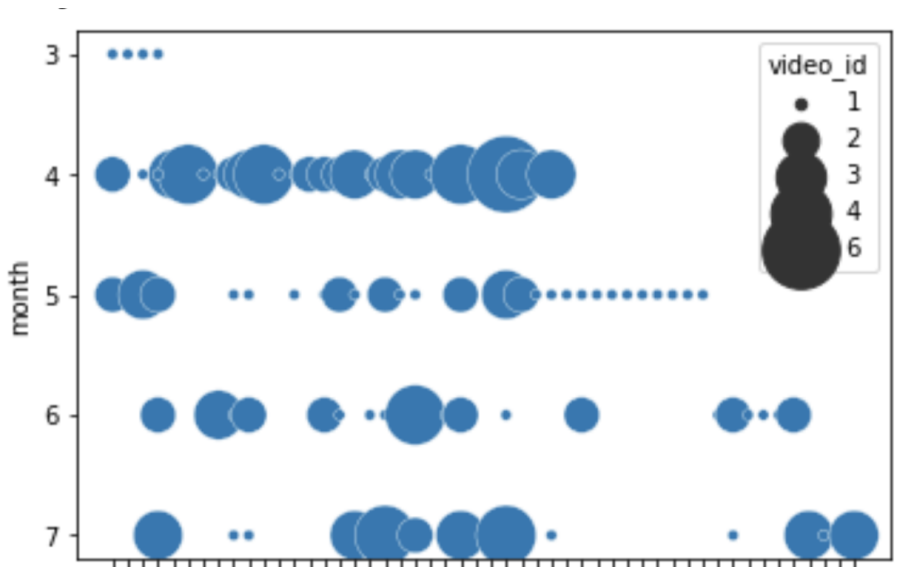


News & Politics

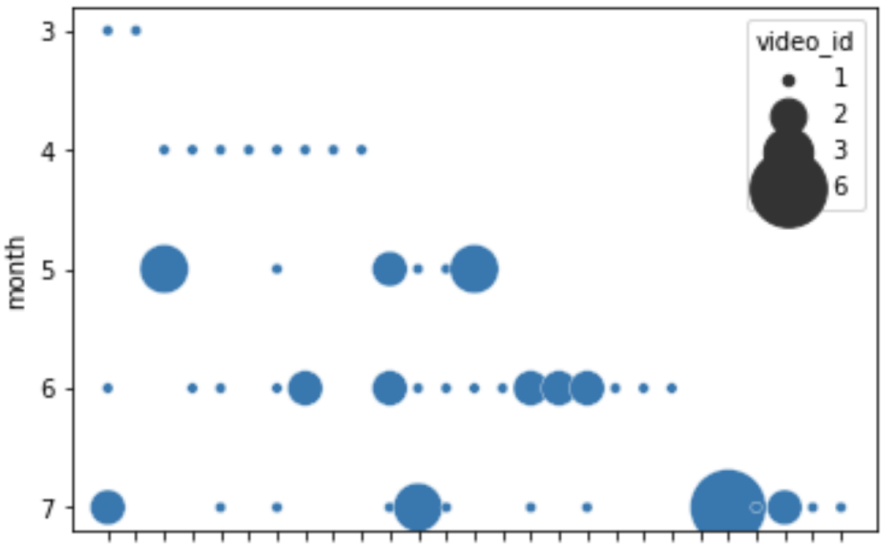
Gaming



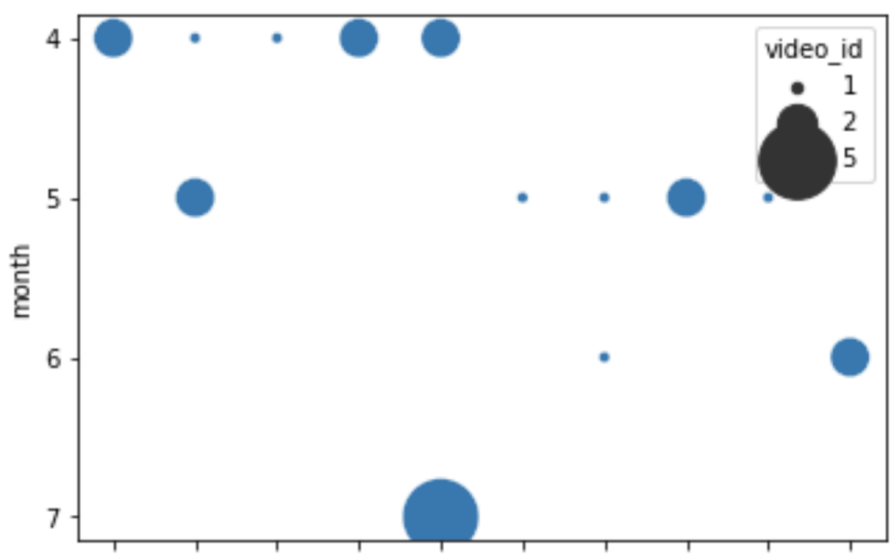
Comedy



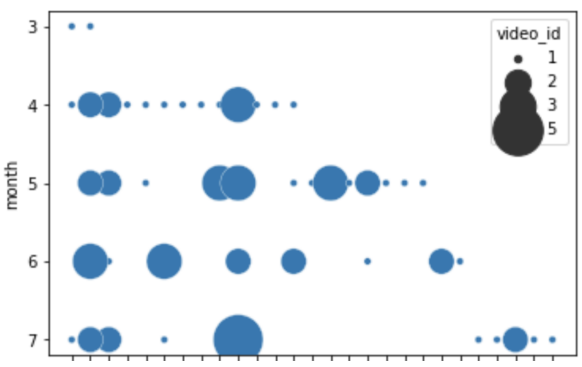
Pets & Animals



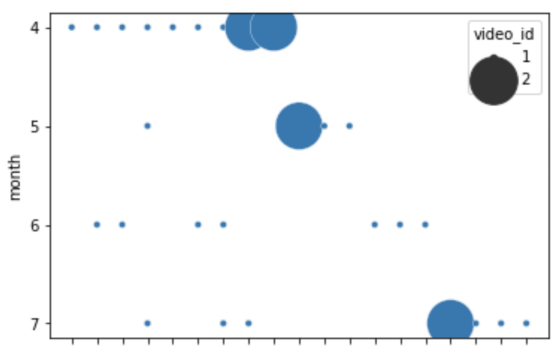
Travel & Events



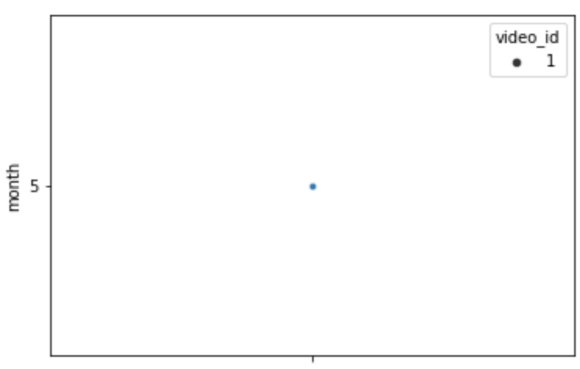
Film & Animation



Autos & Vehicles

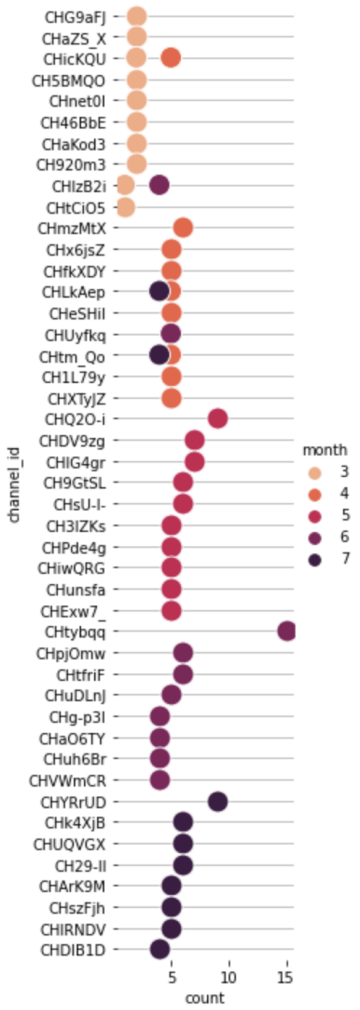


Nonprofits & Activism



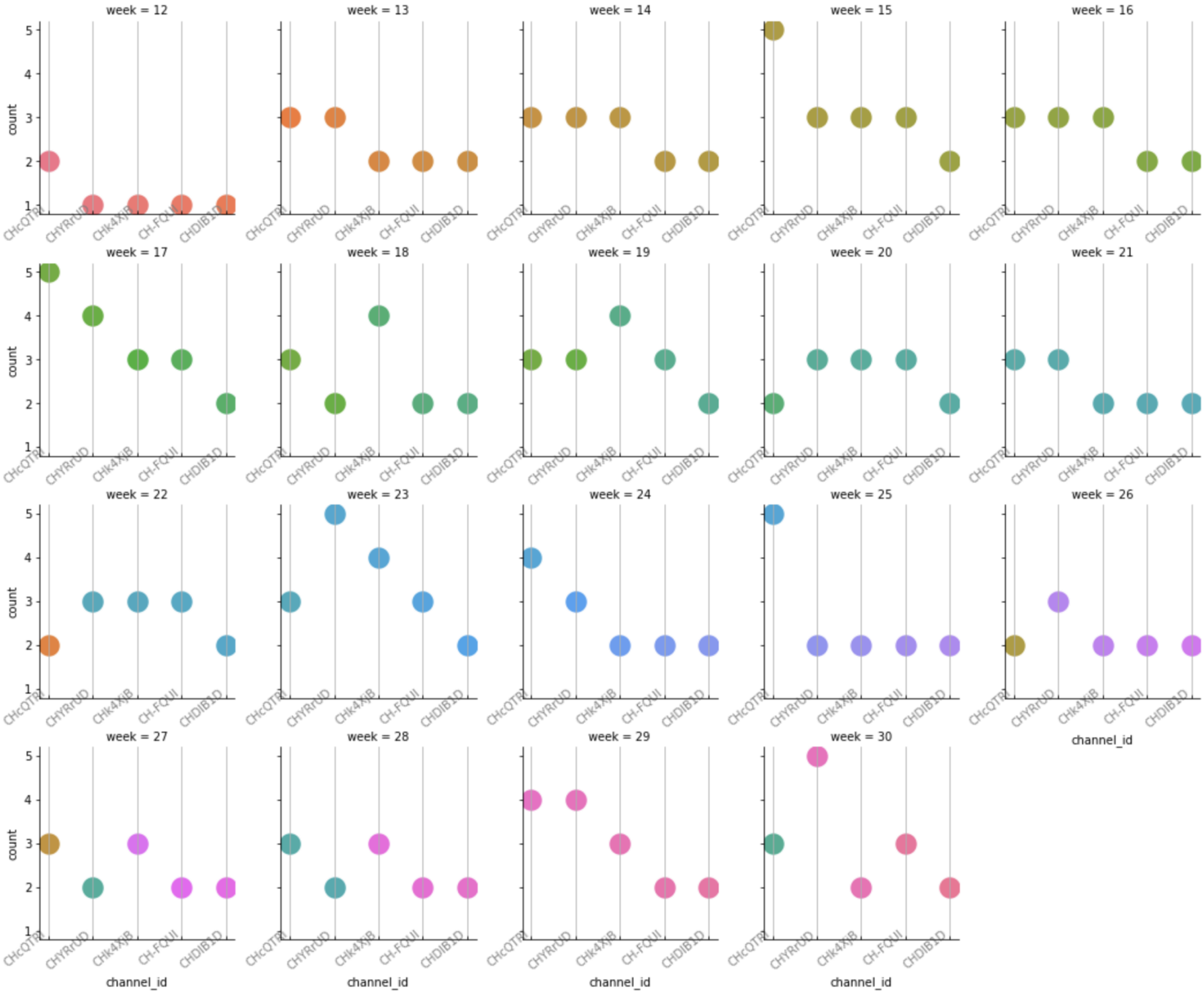
Q1-3. Monthly Top 10 Channels

The next stripplot shows monthly top 10 channels based on the number of videos. The colors vary by month. Most of the top channels rank approximately 10 videos or below. Only one channel listed 16 videos as the highest record. Interesting part is that there are fewer channels than expected that listed several months. On the graph, only 4 channels listed their videos trending twice.



Q1-4. Weekly Top 10 Channels

The following chart shows weekly top 10 channels based on the number of videos. Each channel has each specific color, which shows that some channels appear again later on the top 5 list.



Q1-5. What are the Most Common Words in Video Tags?

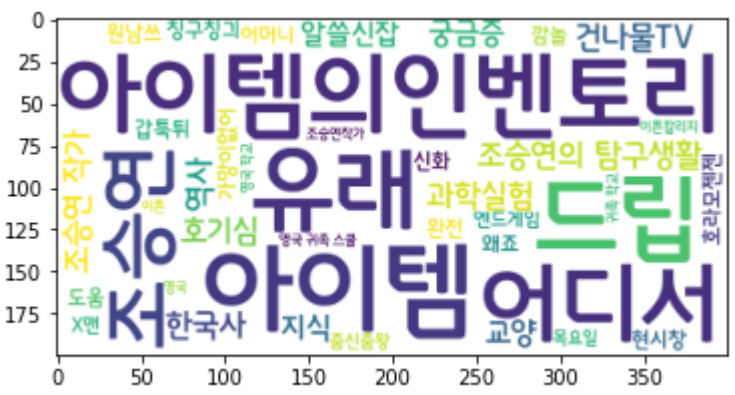
Are there some words that occur in trending video tags more than others? Below wordcloud is monthly top occurrences of words in tags. The size of the word reflects how common it is:

March

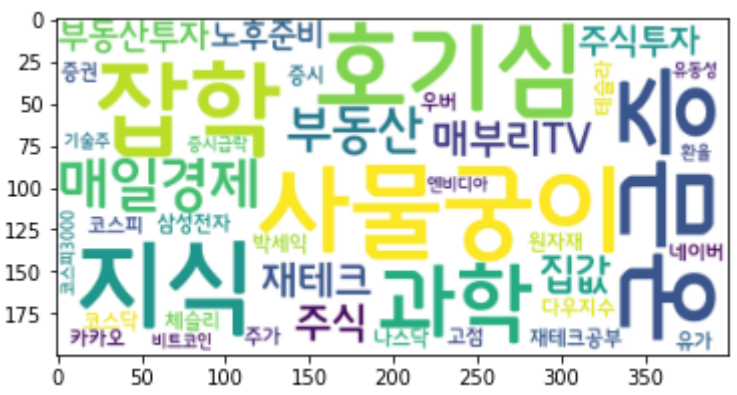
April

May

June



July



You can see tags represent directly channel's name, such as '사물궁이', '아이템의 인벤토리'. Also, viewers like informative videos, usually focused on investments.

Q2. New index to determine Trending : dislike is also a reaction

Youtube aims 'Trending' to be appealing to a wide range of viewers. In order to do so, they combine signals such as view counts, temperature of videos(how quickly it generates views), age of the video, and how it perform compared to other recent uploads from the same channel. This is why the most view count video are not always the most trending video. But what else index could lead to consider 'hotness'?

If Youtube wants to discover a diamond in the rough, videos with not so many subscribers but goes viral, getting suddenly increased signals are the ones to look at. It should lure viewers that are not subscribers.

As shown below, there are even 64 videos that have 0 subscribers and 17 videos of 0 comments but made it to Trending. This means views are more significant feature than subscribers or comments on considering the list.

```
] df['on_channel_subscribers'].value_counts()
```

```
0          64
2070000    12
1600000    11
1540000    11
109000     11
           ..
5730000     1
21200       1
25200       1
30000       1
5810000     1
Name: on_channel_subscribers, Length: 1319, dtype: int64
```

```
df['on_comments'].value_counts()
```

```
0        17
333       7
416       6
985       6
461       5
         ..
31818     1
1554      1
1787      1
841       1
57460     1
Name: on_comments, Length: 1920, dtype: int64
```

So I calculated the differences between on and off stats to demonstrate what feature has the biggest difference after going on Trending. As below, `dislikes` stats are the most rapidly changing indicator at 40.7%. So it could be said that once the video is up in the Trending, the most affected stat is people's `dislike`. Negative actions might be the most easiest way of showing engagement. Then what about using this index and develop into something useful?

	Differences
views	37.814727
likes	22.015870
dislikes	40.719558
total_views	4.591549
total_vids	0.414059

The new index is based on `dislikes`, but it is restricted by `likes` also, to control content quality. The two measurements are added and `likes` have more weight to show it is still widely apealing content.

```
[256] weight = 1.5
      df['dislove'] = df['on_dislikes'] + df['on_likes'] * weight
```

This is a simple modification, but I believe if there is enough data to train and backpropagate to calculate weight, the new index combining solely `dislikes` and `likes` might work as a interesting index for considering the Trending section. Also, the index now had the highest score of correlation with views.

```
sns.heatmap(df_.corr(),annot=True, fmt=".3");
```