Data Analysis of YouTube Trending Videos

CS492: Introduction to R for Data Science

Team 12

20170459 Seono Lee

20170616 Heejin Jeong

Overview

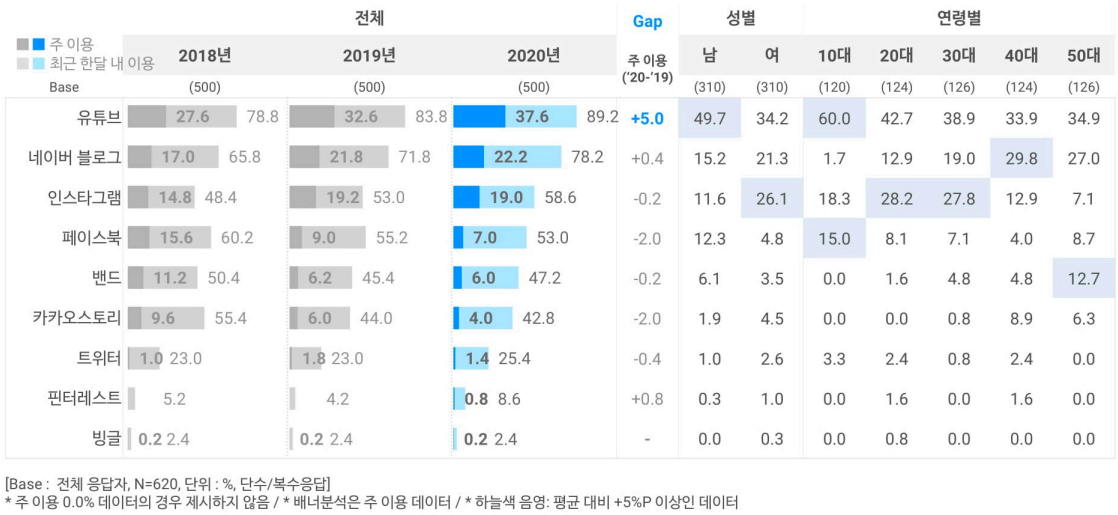
Nowadays, the online video market is starting to be more and more competitive under the broadening influence of YouTube, which is the very famous social media platform all over the world. So the creators should develop their ability to quickly analyze new trends and suggest effective strategies to survive in the market, as this change goes through.

This document contains the **R-based analysis of YouTube**, especially on its ranking factors and statistics which you can use when you manage your own video channel. Our main focus will be the questions 1) which variables are highly influential on the matter in number of views, and 2) what is the better way to get more view counts.

Also, You can see and download our full code here: https://github.com/suno10/2020_CS492_Project

Introduction

YouTube is a video sharing platform serviced by Google. As the largest video sharing site in the world, YouTube users can watch, upload, and share videos. Anyone using a computer could upload videos, making them visible to millions of people in minutes. It is used in various fields for communication in society. It is used in most fields such as marketing, politics, education, games, entertainment, music, and sports.



	전체			Gap	성별		연령별				
■ 주 이용 ■ 최근 한달 내 이용	2018년	2019년	2020년	주 이용 ('20-'19)	남	여	10대	20대	30대	40대	50대
Base	(500)	(500)	(500)		(310)	(310)	(120)	(124)	(126)	(124)	(126)
유튜브	27.6 78.8	32.6 83.8	37.6 89.2	+5.0	49.7	34.2	60.0	42.7	38.9	33.9	34.9
네이버 블로그	17.0 65.8	21.8 71.8	22.2 78.2	+0.4	15.2	21.3	1.7	12.9	19.0	29.8	27.0
인스타그램	14.8 48.4	19.2 53.0	19.0 58.6	-0.2	11.6	26.1	18.3	28.2	27.8	12.9	7.1
페이스북	15.6 60.2	9.0 55.2	7.0 53.0	-2.0	12.3	4.8	15.0	8.1	7.1	4.0	8.7
밴드	11.2 50.4	6.2 45.4	6.0 47.2	-0.2	6.1	3.5	0.0	1.6	4.8	4.8	12.7
카카오스토리	9.6 55.4	6.0 44.0	4.0 42.8	-2.0	1.9	4.5	0.0	0.0	0.8	8.9	6.3
트위터	1.0 23.0	1.8 23.0	1.4 25.4	-0.4	1.0	2.6	3.3	2.4	0.8	2.4	0.0
핀터레스트	5.2	4.2	0.8 8.6	+0.8	0.3	1.0	0.0	1.6	0.0	1.6	0.0
빙글	0.2 2.4	0.2 2.4	0.2 2.4	-	0.0	0.3	0.0	0.8	0.0	0.0	0.0

[Base : 전체 응답자, N=620, 단위 : %, 단수/복수응답]
* 주 이용 0.0% 데이터의 경우 제시하지 않음 / * 배너분석은 주 이용 데이터 / * 하늘색 음영: 평균 대비 +5%P 이상인 데이터

Figure 1

According to the 'Social media and Scanning portal Trend report 2020' of Opensurvey, which is an online-based survey firm in Korea, YouTube becomes peerless social media platform. The statistics says that it hits 1st place for 3 years. Furthermore, even the

growth in the importance of YouTube is steadily increasing. It is causing the gap to widen between social media continuously.[1] More statistics in Korean is in the Figure 1.

Then, what keeps YouTube being a largest part in this new trend in Korea? The findings of the Opensurvey suggest, the secret of its success is in **richer contents** that are provided to the users.

In the past, people use Facebook or Kakaostory to build relationships with friends through online services. But now, many people use social media to get variety of contents, for the sake of interests and needs (not for just chatting). Ratio of the users who use social media to make some communications between individuals has halved in past 4 years.[2] These changes show that how important the contents are to this business, especially in the deluge of information.

So, our goal in this project report is to describe characteristics of videos drawing large amount of viewers. Taking these features into consideration, strategic planning to get more views, subscribers, and traffic to your video will be easier.

Workflow

1. Preprocessing of the dataset & Column modifications
2. Exploratory analysis: review of data characteristics & visualization
3. Statistical analysis of ranking factors: length of the title, number of tags, common word in tags, most popular categories, etc.
4. Deep-dive for most popular category
5. Conclustion and suggestion for improvement of project

Dataset

The dataset was extracted from Kaggle, 'Trending YouTube Video Statistics'(https://www.kaggle.com/datasnaek/youtube-new).

According to the contributor of this dataset, it is a daily record of the top trending YouTube videos. They are not simply selected because of their high number of views, but are selected based on various factors. Data is included for the US, GB, DE, CA, FR, RU, MX, KR, JP, and IN regions (USA, Great Britain, Germany, Canada, France, Russia, Mexico, South Korea, Japan and India, respectively), but we are gonna be running point on data from KR(Korea).

---

[1] Opensurvey, Social media and Scanning portal Trend report 2020, https://blog.opensurvey.co.kr/trendreport/socialmedia-2020/, 2020-03-02, p.11

[2] The ratio hit 52.2% in 2016, but the ratio has dropped to 26.1% this year

The table below(Figure 2) sets out the 1 to 35 rows of 34,567 entries, and first 8 columns of this dataset.

video_id	trending_date	title	channel_title	category_id	publish_time	tags	views
jGasLbM4798	17.01.12	'어벤져스: 인피니티 워' 예고편의 티저 예고편 (한국어 자막)	천상코니	1	2017-11-28 23:01:45	[none]	325208
7FibNosIxio	17.01.12	"한 번만 만아주십시오" 문재인 대통령, 당찬 청년 창업가 요청에 '...	TVCHOSUN 뉴스	25	2017-11-30 10:20:41	TV조선|"티비조선"|"조선일보"|"종편"|"종합편성"|"티브이조선"|"...	26523
6hnuFhbMYiI	17.01.12	(심층해설) 국민 여러분! DJ, 노무현의 죄(罪) 기억하시나요? 윤창...	윤창중칼럼세상 TV	25	2017-11-30 06:12:08	윤창중|"칼럼"|"윤칼세"|"윤창중칼럼세상"|"문재인"|"노벨평화상"|"...	29701
ZFX42JkV7S8	17.01.12	(특종)귀순병사 한국TV와 거리를 바라볼 정도로 건강해진 현재모습	Sion TV	22	2017-11-28 16:50:45	[none]	113625
M5nW_-ICwWA	17.01.12	(평창롱패딩) 해바라기패러디 패바라기 ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ...	보물섬	23	2017-11-27 08:00:00	[none]	299847
eefErqeVyqI	17.01.12	[개들의 전쟁] 맨날 때리고 괴롭히던 선배 잡아서 복수하다	최딕 TV	1	2017-11-29 10:55:54	영화|"한국영화"|"개들의 전쟁"|"영화 개들의 전쟁"|"김무열"|"김성...	243824
kA4c-dGWHgQ	17.01.12	[전체] 김어준의 뉴스공장 1130(목) 권순정, 안민석, 김성태, 박...	시대정신	25	2017-11-30 00:02:27	안민석|"김성태"|"박지원"|"김진애"|"김어준의 뉴스공장"|"이영박"|...	68089
Vpt7CX0L-60	17.01.12	[11월 29일]심재철 사이버복동도 내란, 문재인 정부 내란죄 해당...	JBC 까	24	2017-11-29 10:19:48	[none]	18052
jZK48A76Ipg	17.01.12	[2017MAMA x M2] 고장환 meets 세븐틴 in MAMA	M2	24	2017-11-30 05:00:01	엠넷|"Mnet"|"엠투"|"M2"|"MPD"|"엠피디"|"고장환"|"나몰라 패...	65389
cynJAuU9WDY	17.01.12	[ARASHI] 【日本語字幕】비정상회#에서 소개된 사쿠라이쇼 韓国放...	개만후라이쇼	22	2017-11-30 14:38:17	아라시|"ARASHI"|"嵐"|"櫻井翔"|"韓国放送"|"사쿠라이쇼"	15579
_9SguIVvAaE	17.01.12	[BT21] Animated Stickers - UNIVERSTAR #2	BT21	22	2017-11-30 09:44:53	BT21|"BTS"|"방탄소년단"|"방탄"|"아미"|"ARMY"|"A.R.M.Y"|"...	139019
jZYLMjaZRfY	17.01.12	[ENG] 171129 NU'EST W Backstage 뉴이스트W 백스테이지...	ShuaShua	22	2017-11-29 14:47:46	[none]	28251
bSpEJrCeL0Y	17.01.12	[ENG] 워너원 - 강다니엘: '아이시테루' ♡ 2017 MAMA in Jap...	다닉쓰 남팬 캘빈 Kelvin	24	2017-11-29 10:30:34	프로듀스 101 시즌 2|"Wanna One"|"워너원"|"나야 나"|"강다...	14459
Xst9fL2qvkU	17.01.12	[Eng]겨울내내 완벽한베이스메이크업+미스트에 대한 오해? [백...	RISABAE	26	2017-11-28 13:49:10	이사배|"이사베메이크업"|"RISABAE"|"RISABAEART"|"TUTOR...	276235
c5_LROaHGtw	17.01.12	[EPISODE] BTS (방탄소년단) 'MIC Drop' MV Shooting	BANGTANTV	10	2017-11-29 10:00:03	방탄소년단|"BTS"|"BANGTAN"|"HIPHOP"|"랩몬스터"|"RapM...	1728678
v6_GwXU1lkg	17.01.12	[MV] JANG DEOK CHEOL(장덕철) _ Good old days(그날처럼)	1theK (원더케이)	10	2017-11-28 09:00:08	Kpop|"1theK"|"원더케이"|"loen"|"로엔"|"유비"|"티저"|"MV"|"...	50588
5paW9CN-nMM	17.01.12	[YTN LIVE] 북한, 12시 30분 '중대보도' 예고 - 미사일 발사 관...	YTN NEWS	25	2017-11-29 04:01:24	YTN|"NEWS"|"생방송"|"24시간"|"24"|"뉴스"|"생중계"|"라이브...	108943
ANhl_L14B_w	17.01.12	[가사 포함] 나얼 - 기억의 빈자리 (2017)	준영	10	2017-11-29 10:17:25	나얼|"브라운 아이드 소울"|"브아술"|"기억의 빈자리"	17960
witLhchvgtw	17.01.12	[강형욱의 소소한 Q&A] 산책 중 돌아보는 강아지, 따라오는지 확...	Bodeum official	15	2017-11-30 09:38:51	강형욱|"세상에나쁜개는없다"|"보듬"|"강아지"|"반려견"|"반려견교...	54701
qDYFjiV4vgE	17.01.12	[권창훈 볼터치] 메시 빙의한 시즌 5호골!! 원더골로 3경기 연속골...	DALMOON	17	2017-11-28 20:37:52	권창훈|"권창훈 스페셜"|"권창훈 아이맹"|"권창훈 5호골"|"권창훈 ...	249802
BYKLXTKyhdk	17.01.12	[김정민의자연사박물관]3부 (통합편)죽은 노무현이 산 문재인을 잡...	글로벌디텐스뉴스	25	2017-11-28 02:59:40	문재인|"트럼프"|"노무현"|"노건호"|"권양숙"|"박연차"|"MB"|"이명...	31832
kl7bln0_1Q4	17.01.12	[뉴스1번지] 북한, 신형ICBM 발사 성공 주장…한반도 정세 요동 /...	연합뉴스 TV	25	2017-11-29 08:33:06	Yonhapnews TV|"News"|"실시간"|"Live"|"생방송"|"뉴스"	29184
mNeEBYFoZMA	17.01.12	[뉴스타운TV] 북한 김정은과 최룡해는 이 방송을 보고나면 둘 중 ...	뉴스타운TV	25	2017-11-29 13:19:38	뉴스타운|"손상윤"|"Live"|"손상대"|"뉴스타운TV"|"라이브"|"태극...	95391
HkTIpEwn0KM	17.01.12	[방탄소년단] 엠랜쇼가서 라이브 인증 제대로하고 간 방탄 (Feat. ...	BTS_HOBBY 호비적	10	2017-11-27 08:20:39	[none]	1138521
koyP4fqMpSM	17.01.12	[법륜스님의 즉문즉설 제 1342회] 시어머니가 같은 건물로 이사 ...	법륜스님의 즉문즉설	29	2017-11-29 19:30:01	불교|"즉문즉설"|"법륜스님"|"정토회"|"buddha"|"buddhism"|"...	32455
zWfNtCIJJtw	17.01.12	[병맛더빙] 급식생 3 X 콜롬비아나	장삐쭈	24	2017-11-30 09:23:26	장삐쭈|"삐쭈"|"콜롬비아나"|"병맛더빙"|"더빙"|"ㅋㅋㅋ"|"꿀잼"|"...	994446
khkMVkUPeRA	17.01.12	[보루토 35화] 미츠키의 부모 자격으로 나뭇잎에 온 오로치마루	Naruto Video Collection	1	2017-11-29 19:19:12	나루토|"naruto"|"이타치"|"itachi"|"우치하"|"uchiha"|"sasuk...	303663
vGC4_a8pHgo	17.01.12	[서프라이즈] 캄보디아 밀림 속에서 발견된 유명 도시, 세계 7대 불...	서프라이즈	24	2017-11-28 04:55:29	서프라이즈|"서프라이즈 웹툰"|"신비한 웹툰 서프라이즈"|"서프라...	87452
TQeK1Hs95uA	17.01.12	[송은이 김숙의 비밀보장] 뮤지컬배우들의 마이크 위치가 다른 이유??	VIVO TV	23	2017-11-30 08:00:01	송은이|"김숙"|"송은이&김숙비밀보장"|"비밀보장"|"김생민"|"김생...	19170
uHBgnvwsfAs	17.01.12	[아라시] 매운 음식 먹다 이상증세를 보이는 4인(feat.리다 매운 거...	SAKU 사쿠	22	2017-11-30 10:18:39	[none]	25522
9fMjSYFEbaQ	17.01.12	[엄마가 잠든후에] 하니가 사무실에서 급 칼을 든 이유는?	피키픽처스 Piki Pictures	24	2017-11-28 13:00:17	연애초밥|"연애"|"KPOP"|"사시미"|"엄마가잠든후에"|"엄잠후"|"술...	310984
lgnysHU4gAo	17.01.12	[워너원 강다니엘 황민현] 보아&황민현 only one 리액션 부자 강...	Padi edit	22	2017-11-29 16:50:03	[none]	76697
M-rWK-vV76Y	17.01.12	[이번생은처음이라]윤보미 걸크러쉬~! 김민석에게 싸대기 한방!!!	이지핑크	22	2017-11-28 03:27:55	이번|"생은"|"처음이라"|"이번생은처음이라"|"이번생은"|"드라마"|"...	378821
_dPYNCNGkdI	17.01.12	[일곱개의 대죄] 엘리자베스의 각성과 호크의 희생	유기견	1	2017-11-29 11:53:52	일곱개의 대죄|"엘리자베스"|"호크"|"엘리오다스"	57573

Figure 2

Dimension

34,567 rows, 16 columns

Percentage of null values

3,163 / 553,032 = 0.572%

Variable

There are total 16 variables (except index)

```
> colnames(KRvideos)
 [1] "video_id"             "trending_date"        "title"                "channel_title"
 [5] "category_id"          "publish_time"         "tags"                 "views"
 [9] "likes"                "dislikes"             "comment_count"        "thumbnail_link"
[13] "comments_disabled"    "ratings_disabled"     "video_error_or_removed" "description"
```

Exploratory Analysis

Preprocessing

1. **Below columns were deleted.**
   - Video_id & video_error_or_removed: they don't provide meaningful information about our project.
   - Thumbnail_link: all video has its thumbnail link, so this variable is also meaningless.

2. **Four columns were modified for technical purpose:**
   - Trending_date: This data type was characters such as '17.30.11', so we use ydm function in lubridate library to modify them '2017-11-30'.
   - Category_id & Trending_date: We make them factor using 'as.factor' function.
   - Publish_date: This data type was characters such as '2017-11-30 10:14:31', so we seperate them as a 'publish_date' and a 'publish_time' data.

```
KRvideos$trending_date <- ydm(KRvideos$trending_date)
KRvideos$publish_date <- substr(KRvideos$publish_time, 1, 10)
KRvideos$publish_time <- substr(KRvideos$publish_time, 12, 19)

KRvideos$category_id <- as.factor(KRvideos$category_id)
KRvideos$trending_date <- as.factor(KRvideos$trending_date)
```

3. **New column:** publish_time whose data type is time, which means '__h __m __s'.

4. **The data is arranged in the order of the view counts.** The video with the highest view is at the head.

Trending_date

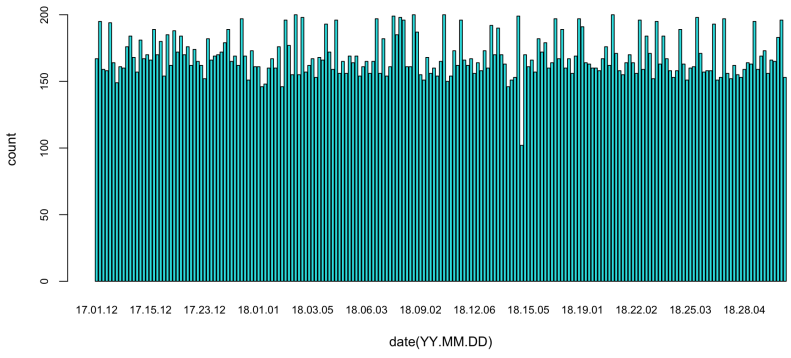


Figure 3
y-value means count of the trending videos for each date

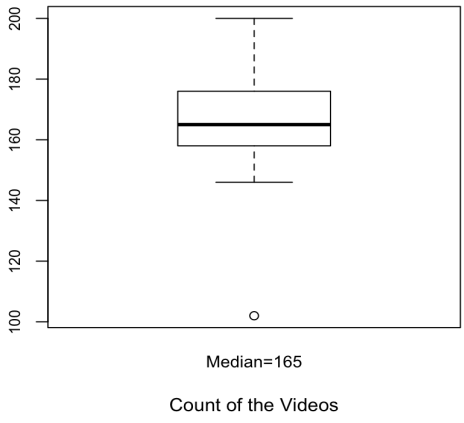


Figure 4. Boxplot
about 165 videos are listed per day

Figure 3 shows the number of trending videos by year, month, and date, and Figure 5 shows statistics of them. Figure 4 is a schematic diagram of the Figure 5 using the boxplot function.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
102	158	165	168.6	176	200

Figure 5. Summary statistics

According to the statistics, there are about 165 listed trending videos per day.

Publish_time

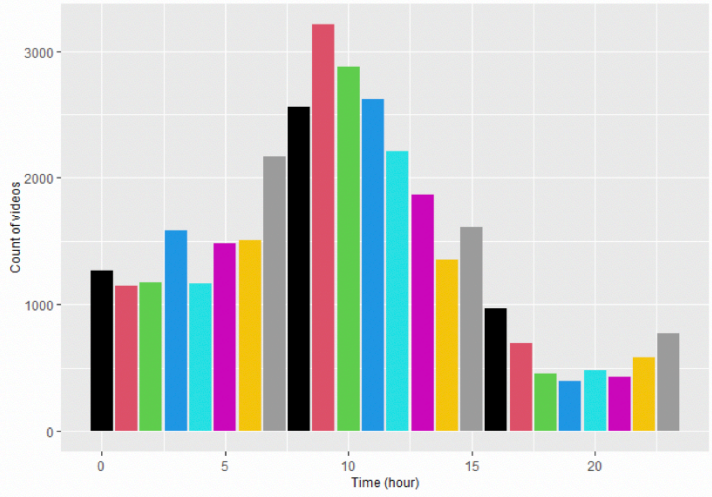


Figure 6

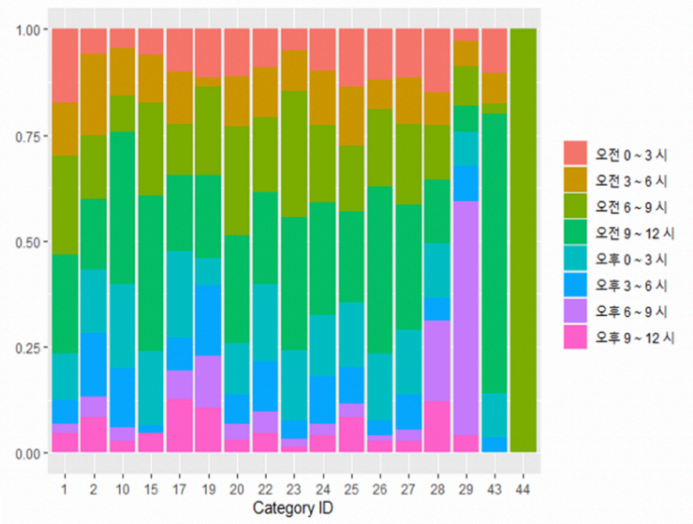


Figure 7

```
> sort(table(time_hour),decreasing=T)[1:6]
time_hour
   9   10   11    8   12    7
3214 2877 2623 2563 2206 2169
```

Figure 6 shows the number of uploaded videos according to the time of day. The interval of X data is 1 hour. The most videos were uploaded between 9 am and 10 am. Overall, the videos were uploaded the most between 7 am and 1 pm.

Also, Figure 7 shows at what time the video was uploaded according to the category. The category ID list can be seen in Figure 8. Overall, it seems that videos were uploaded the most between 6 am and 12 am. Remarkably, in Category 29, many videos were uploaded between 6 and 9 pm. And in category 44, all appeared between 6 and 12 am. This is because the number of videos is small, so the time is not distributed well and there is a tendency to be biased to one side. The number of videos by category can be seen in Figure 9.

Category_id

Category ID	Title	Category ID	Title
1	Film & Animation	24	Entertainment
2	Autos & Vehicles	25	News & Politics
10	Music	26	Howto & Style
15	Pets & Animals	27	Education
17	Sports	28	Science & Technology
19	Travel & Events	29	Nonprofits & Activism
20	Gaming		
22	People & Blogs	43	Shows
23	Comedy	44	Trailers

Figure 8. Youtube video category name and ID list

Figure 9 (below pictures) show the percentage of each category. "Entertainment" occupies the largest percentage of 25.91%, "News & Politics" occupies 21.93%, and "People & Blogs" occupies 20.41%.

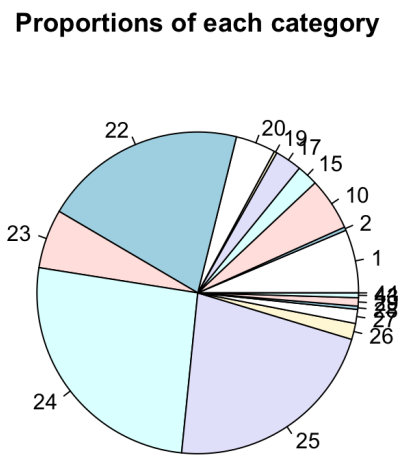


Figure 9-1.
Pie-chart for category ID

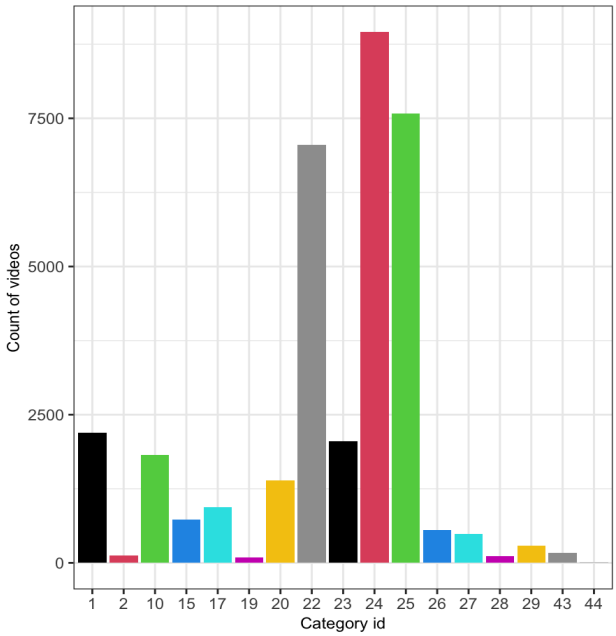


Figure 9-2. Barplot for category ID

Tags

```
> summary(num_tag)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    2.00   10.00   16.22   23.00  152.00
```

Figure 10 below shows word-cloud[3] of tags. The most frequent tags are '문재인(Moon Jae-in)' and '먹방(mukbang)'. Also, Figure 11 shows the number of tags with the boxplot function. In the boxplot, the outliners were excluded.

Figure 10

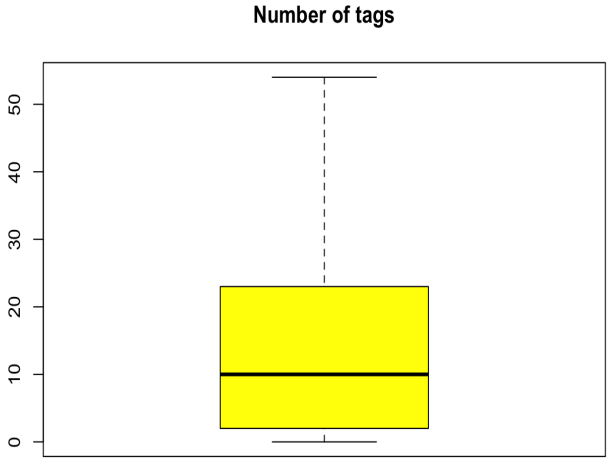


Figure 11

Top 10 Videos

Figure 12 is a selection of 10 videos with the most views among the videos uploaded on top trending videos of YouTube. At the left figure, there were many duplicates because some videos were uploaded to the top trending videos several times. The right one shows unique title of the top 10 videos, not to be duplicated.

```
> head(KRvideos[,'title'], 10)
 [1] "YouTube Rewind: The Shape of 2017 | #YouTubeRewind"
 [2] "YouTube Rewind: The Shape of 2017 | #YouTubeRewind"
 [3] "Marvel Studios' Avengers: Infinity War Official Trailer"
 [4] "BTS (방탄소년단) 'FAKE LOVE' Official MV"
 [5] "Marvel Studios' Avengers: Infinity War Official Trailer"
 [6] "BTS (방탄소년단) 'FAKE LOVE' Official MV"
 [7] "Marvel Studios' Avengers: Infinity War Official Trailer"
 [8] "YouTube Rewind: The Shape of 2017 | #YouTubeRewind"
 [9] "Marvel Studios' Avengers: Infinity War Official Trailer"
[10] "BTS (방탄소년단) 'FAKE LOVE' Official MV"
```

```
> head(unique(KRvideos[,'title']), 10)
 [1] "YouTube Rewind: The Shape of 2017 | #YouTubeRewind"
 [2] "Marvel Studios' Avengers: Infinity War Official Trailer"
 [3] "BTS (방탄소년단) 'FAKE LOVE' Official MV"
 [4] "Childish Gambino - This Is America (Official Video)"
 [5] "VENOM - Official Trailer (HD)"
 [6] "Marvel Studios' Avengers: Infinity War - Official Trailer"
 [7] "TWICE What is Love? M/V"
 [8] "TWICE Heart Shaker M/V"
 [9] "Bruno Mars - Finesse (Remix) [Feat. Cardi B] [Official Video]"
[10] "To Our Daughter"
```

Figure 12. Left one allows duplicated values

Correlation between Publish date and Trending date

Period = Trending date - Publish date

```
summary(period)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.000   1.000   2.000   2.112   3.000  30.000
```

[3] For reproducibility of word cloud, set.seed(20170459)

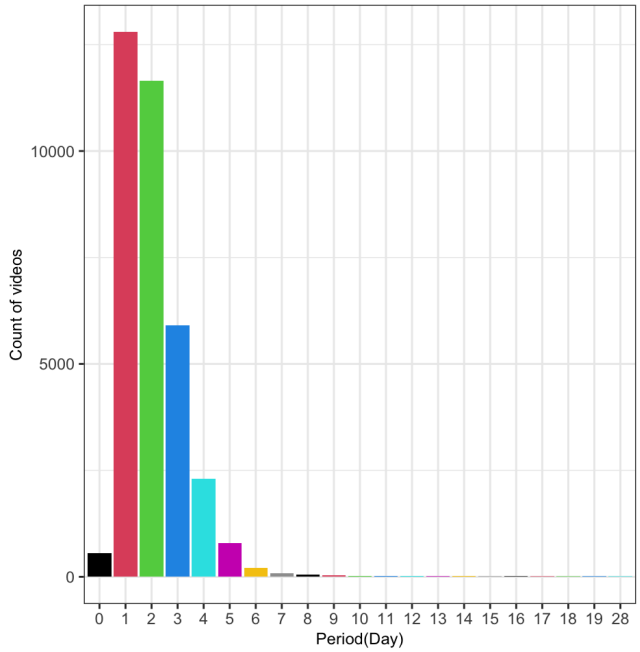


Figure 13

Figure 13 show how long after the video was uploaded it was selected as the top trending video. In general, it was often selected after 1–2 days.

Correlation between Likes and Dislikes

Figure 14 shows the relationship between likes and dislikes. To better representation, the log command was used in plot function.

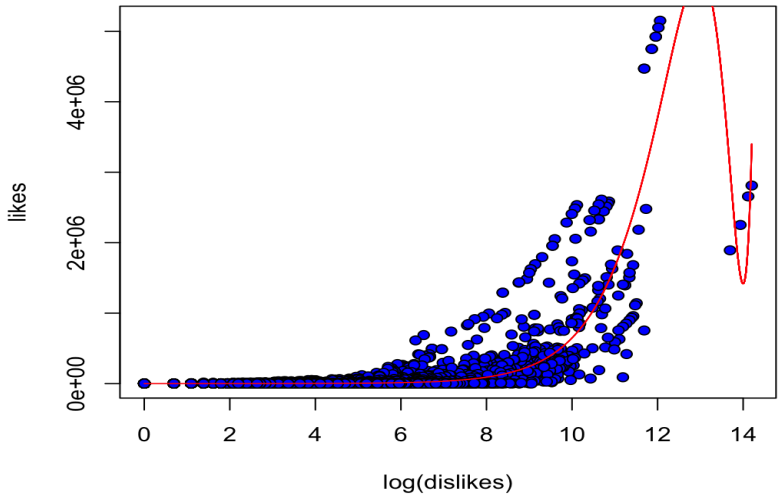


Figure 14. The red line follows *lm(likes~poly(dislikes, degree=3))*

Comments_disabled & Ratings_disabled

```
> table(comments_disabled) > table(ratings_disabled)
comments_disabled          ratings_disabled
FALSE  TRUE                FALSE  TRUE
34056   511                33167  1400
```

Figure 15

Figure 15 show the distribution of videos that enable commenting and ranking. Each relative ratio is 98.52%, and 95.95%. Comments and rating of most of the videos are abled.

Summary

Here are 5 facts about YouTube trending videos, which can be indicators of references if you want your video to garner great popularity.

1. Most of the trending videos were uploaded between **9 am and 10 am.**
2. The most frequent tags are **'문재인(Moon Jae-in)', '먹방(Mukbang)', and '뉴스(News)'.**
3. Overall, **"Entertainment"** occupies the largest percentage in categories
4. On average, one video has **10~15 tags.**
5. Comments and rating of most of the videos are abled.

The detailed ranking factors will be addressed and analyzed in the next chapter, with some statistical schemes.

Statistical analysis of ranking factors

Problem Statement

1. What is the **length of the title** that increases the exposure most?
2. Does the exposure increase when the **tags contain** the most **frequent words**?
3. What is the **number of tags** that increases the exposure most?
4. What are popular and unpopular **categories**?
5. All of those factors are statistically meaningful?

Approach

We will make a subset of videos from the whole dataset for each problem statements, and analyze them with statistical methods.

First, to determine the normality, we use *shapiro.test()* or *cvm.test()* in *nortest* library. If it follows normal distribution, we do T-test[4] or ANOVA-analysis of variance[5] test to determine whether each variables show significant different statistically. If it doesn't follow normal distribution, we take the way of Kruskal-Wallis rank sum test[6]. We can reject null hypothesis which means those variables are the same across view counts, when p-values ≤ 0.05. So, it means that variable can be one of the ranking factors.

If it shows significant differences, we will do visualization which seems to be best suited to each of them. The detailed steps are described in the coding document.

Length of the title

1. Data pre-processing

   There was no video whose length of the title is more than 100 characters. So, we separated data into 10 levels by length of the title. The level was broken down for each based on the double digits, from 0 to 100.

   The first row of the Figure 16 means the range of the length of title, and its second

[4] You can get further information here: https://ekja.org/upload/pdf/kjae-68-540_ko.pdf (Tae Kyun Kim, "T test as a parametric statistics", Korean Journal of Anesthesiology)

[5] https://www.researchgate.net/publication/272311020 (Steven Sawyer, "Analysis of Variance: The fundamental concepts", The Journal of manual & manipulative therapy)

[6] https://www.researchgate.net/publication/289442433 (Eva Ostertagova., et al. "Methodology and Application of the Kruskal-Wallis Test", Applied Mechanics and Materials, August 2014)

row means number of videos in each range.

```
table(title_df$length_range)

0-10  10-20  20-30  30-40  40-50  50-60  60-70  70-80  80-90 90-100
 200   3167   7207   7979   6515   3777   2363   1521    990    848
```

Figure 16

2. Cramer-von Mises test for normality

We performed the Cramer-von Mises test(cvm test) for the composite hypothesis of normality, by range of the length.

```
> by(title_df$views, title_df$length_range, cvm.test)
```

- null hypothesis: It follows normal distribution.
- alternative hypothesis: It doesn't follow normal distribution.

3. Kruskal-Wallis Rank Sum Test

At the cvm test, the p-value is smaller than 0.05, so we couldn't execute ANOVA test. So we performed the Kruskal-Wallis rank sum test.

```
        Kruskal-Wallis rank sum test

data:  views by length_range
Kruskal-Wallis chi-squared = 283.22, df = 9, p-value < 2.2e-16
```

It shows a very small p-value. So we can conclude that length of the title has a significant effect on view counts.

4. Visualization and Conclusion

4-1. Average view for each range of the title length

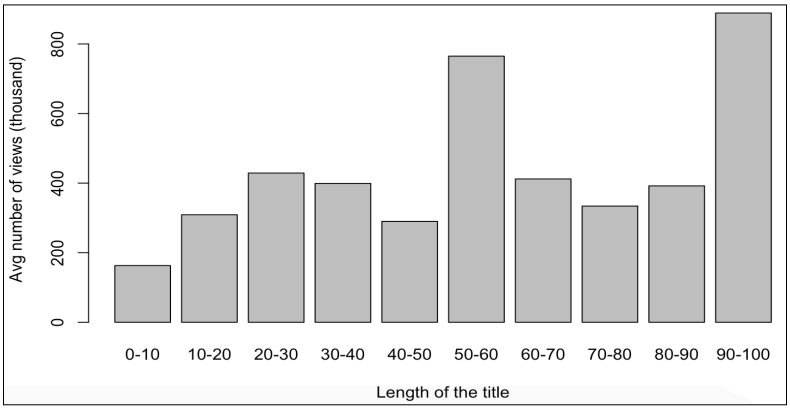


Figure 17. barplot
y-value means average number of view counts (divided by 1,000)

4-2. Analysis of top 500 videos in range of the title length

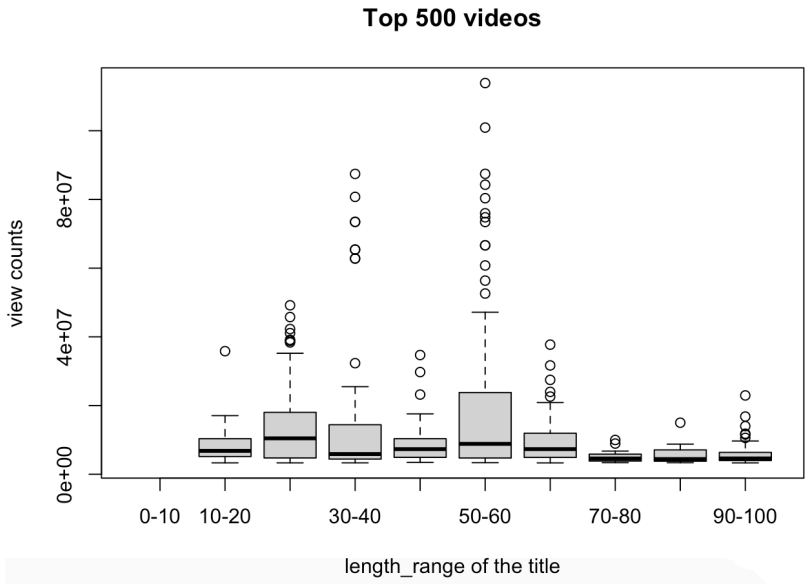


Figure 18. Picked up top 500 videos by view counts

Figure 19 shows that, the videos which have 50-60 characters of title length had the highest percentage of top 500 videos.

```
> table(test.set$length_range)

  0-10  10-20  20-30  30-40  40-50  50-60  60-70  70-80  80-90 90-100
     0     21     89     87     52    106     40     21     21     63
```

Figure 19

4-3. Conjecture of the conclusion

- If the length of the title is too short(less then 20 characters), view counts will get lower because of the paucity of information.

- If the length of the title is too long(between 60-90 characters), view counts will get lower because of the paucity of focusing on the subject.

- If the length of the title is very long, between 90-100 characters, view counts will get little higher. However, as it shows in Figure 16, there aren't as much videos in 90-100 levels. Maybe there are people who watch that videos because of the key word stuffing (Spamdexing).

Category ID

1. Data pre-processing

We filtering out the videos whose category ID is 44. Number 44 means 'trailers' category, and there are only two videos among 34,567 test-sets. So we decided to eliminate these 2 videos in test set for more accurate test.

2. CVM test & Kruskal test[7]

```
> by(test.set$views, test.set$category_id, cvm.test)
```

At the cvm test, the p-value is smaller than 0.05, so we couldn't execute ANOVA test. So we performed the Kruskal-Wallis rank sum test.

```
        Kruskal-Wallis rank sum test

data:  views by category_id
Kruskal-Wallis chi-squared = 3014, df = 15, p-value < 2.2e-16
```

It shows a very small p-value. So we can conclude that category ID has a significant effect on view counts.

3. Visualization and Conclusion

According to the Figure 20 below, videos in 'Science & Technology' category has largest average view counts.

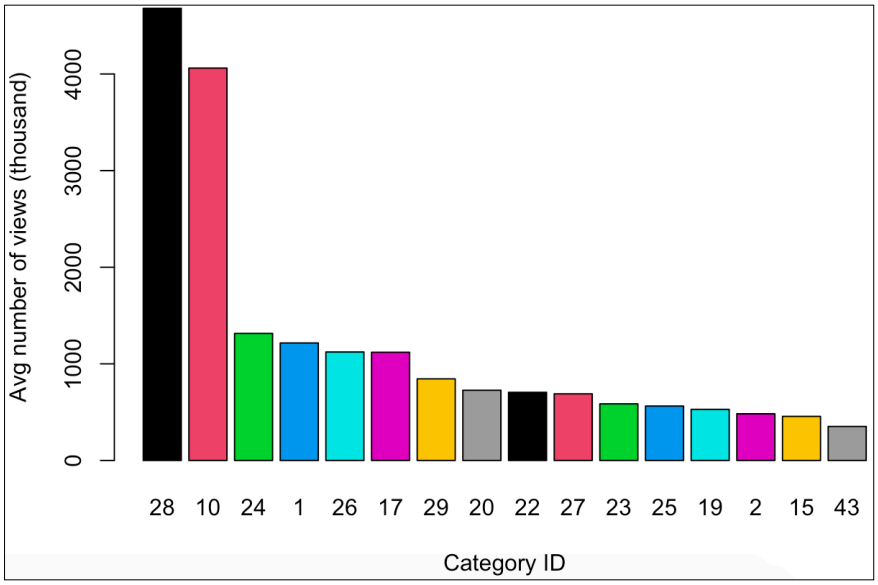


Figure 20

[7] From that part, we dispensed with the result of CVM and Kruskal test in this document, because it is just a repetition of the front part. If you want full descriptions of workflow, please see our codes in github.

However, as it described in Figure 9, only 115 videos are in 'Science & Technology' category. So the data can be biased to one side. Therefore, we concluded that 'Music(number 10)' and 'Entertainment(number 24)' are the most popular category between YouTube trending videos.

Number of the tags

1. Data Pre-processing

We made num_tag variables as a factor, separated into 8 levels. The levels and statistics of the factor are shown in Figure 21.

```
> levels(tag_df$num_tag)
[1] "0-5"    "5-10"   "10-20"  "20-30"  "30-50"  "50-80"  "80-100" "100~"

> table(tag_df$num_tag)

   0-5   5-10  10-20  20-30  30-50  50-80 80-100   100~
 11382   5848   7065   4041   4065   1428    436    302
```

Figure 21

2. Visualization and Conclusion

According to Figure 22, there was a tendency to increase at 0~30, and after that, the number of views decreases slightly as the number of tags increases. If the number of the tags is too much(more than 50), view counts get lower because of the paucity of focusing on the subject.

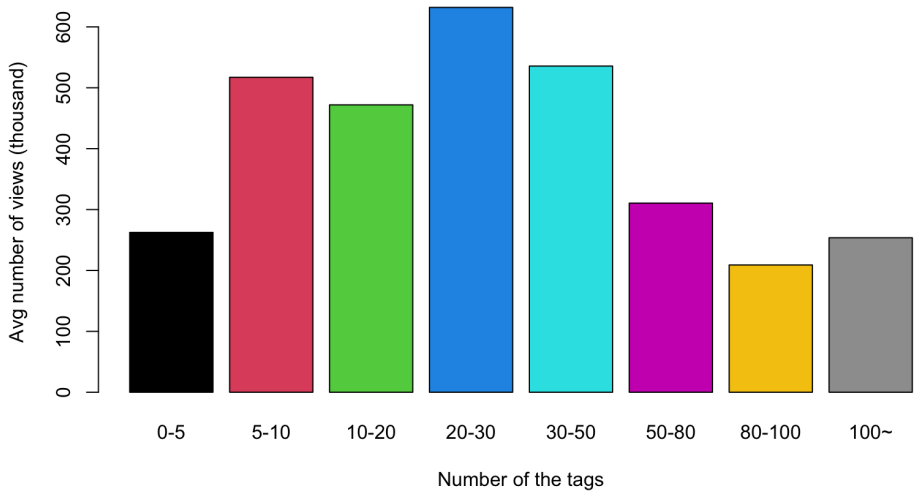


Figure 22

Common tags

1. Data pre-processing

Data analysis was performed using only the top 1000 most used tags obtained in the Exploration-Tags step.

First, to see how the number of views fluctuates as tags with higher rankings are included, we analyzed 100 tags each, for a total of 1000 tags. We selected videos with one or more tags in each group and selected the 100 most viewed videos from among them.

Second, the number of views was analyzed by dividing the videos with the most used tag top 1000 and videos that do not. At this time, since the number of videos in each two groups is different, 100 videos with the highest views were selected and analyzed.

2. Visualization and Conclusion

2-1. Changes in views according to common tags ranking

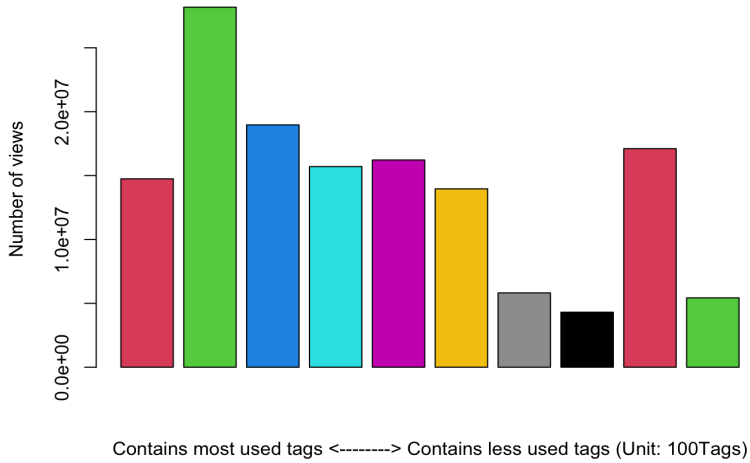
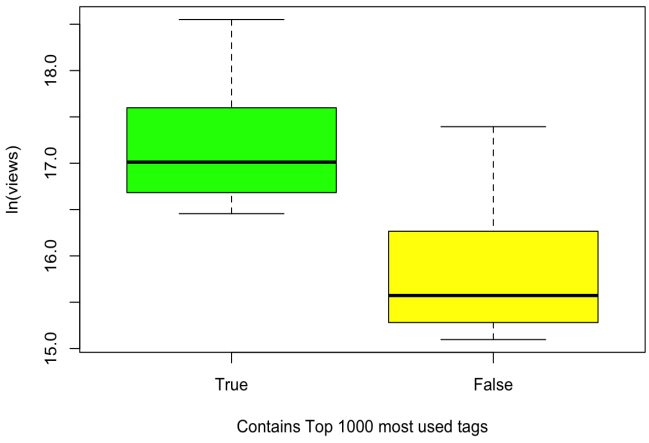


Figure 23

2-2. Analysis of views of videos with and without top 1000 tags




```
> summary(views[c][1:100])
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
 14013696  17603335  24445522  34956794  43412762 113876217
> summary(views[!c][1:100])
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
 3603101  4342583  5795039  9320700 11453298 35832484
```

Figure 24

2-3. Conjecture of the conclusion

- In Figure 23, it appears that the number of views increased as the most used tags were included. Exceptional fluctuations, such as Top 1~10 and Top 81~90, are expected to be affected by videos with very high views.

- In Figure 24, it appears that the number of views increased as the tag that is used most in general is included.

Deep-dive: Tags in the Entertainment category

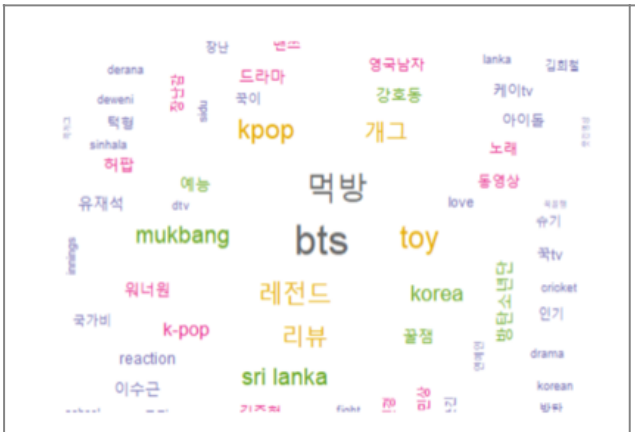
We studied more about 'Entertainment' category, because it seems to be the most popular category. In this chapter, we will see how the flow of YouTube trends of this category goes, and how can we explain it.

1. Data pre-processing

Data were separated at intervals of one month in order to analyze the change of tags over time in the Entertainment category. The videos were distributed from November 2017 to June 2018.

2. Visualization

2-1. Word cloud showing the most distributed tags in the category

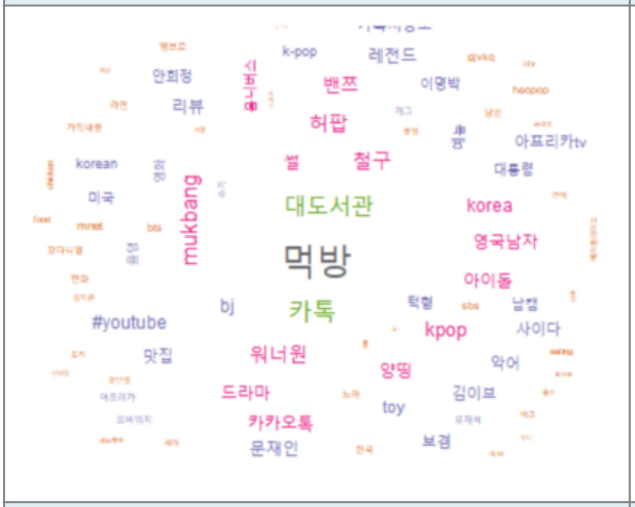


November 2017

December 2017

January 2018

February 2018



March 2018

April 2018

May 2018

June 2018

Figure 25

2-2. Top 10 most distributed tags in the category

Year	2017		2018					
Month	11	12	1	2	3	4	5	6
1st	bts	먹방	먹방	평창올림픽	먹방	먹방	먹방	먹방
2nd	먹방	리뷰	toy	평창	대도서관	철구	bts	bts
3rd	레전드	대도서관	레전드	올림픽	카톡	bj	방탄소년단	장난감
4th	toy	윰댕	muknang	먹방	워너원	드라마	mukbang	toy
5th	리뷰	bj	대도서관	아이돌	철구	워너원	장난감	방탄소년단
6th	개그	허팝	korea	철구	허팝	허팝	레전드	배그
7th	kpop	mukbang	아프리카tv	대도서관	mukbang	김이브	꿀잼	k-pop
8th	mukbang	샤이니	리뷰	무한도전	드라마	리뷰	철구	레전드
9th	sri lanka	웃긴영상	양띵	레드벨벳	썰	영국남자	k-pop	배틀그라운드
10th	korea	꿀잼	kpop	방탄소년단	양띵	강다니엘	kpop	개그

Figure 26

3. Conclusion

- There are distributed tags related to Internet broadcasters who were famous at that time such as 대도서관, 윰댕, 허팝, 양띵, 철구, 김이브, 영국남자.

- 먹방(mukbang) was always picked as one of the most tags regardless of the season.

- There were a lot of tags when there was a special event. (ex. , SHINee 종현 suicide in Dec. 2017, PyeongChang Olympic in Feb. 2018, 무한도전 토토가3 H.O.T in Feb. 2018)

- There were a lot of tags related to famous Kpop star comebacks. (ex. BTS 'MIC Drop' in Nov. 2017, Red Velvet 'Bad boy' in Feb. 2018, Wanna One 'BOOMERANG' in Mar. 2018, BTS 'FAKE LOVE' (the top of the Billboard chart) in May 2018)

Conclusions with Clustering

What makes the number of views high?

1. Summary of the results

According to the statistical analysis, common features of the most popular videos are as follows.

1) Title length: 50-60 characters
2) Number of tags: 20
3) Category id: 10(Music), 24(Entertainment)
4) Including tags that is used most in general

2. Clustering

1) Context

To sum-up the results, we will use clustering algorithm and analyze how many or which classes the videos are divided into. After clustering, the data points that are in the same group have similar properties and features, while data points in different groups have highly dissimilar properties and features. So we can gain some valuable insights from our data by seeing what features the videos have in the each groups, when we apply a clustering algorithm.

2) Data Pre-processing

We made a subset of videos from the whole dataset that have the following characteristics.

- head(KRvideos, 100): Top 100 videos in views
- tail(KRvideos, 100): Bottom 100 videos in views
- KRvideos[17231:17330]: Middle 100 videos in views

To adjust all features to calculate a distance that better aligns with our expectations, we converted the features to be on a similar scale with one another using scale() function.

3) K-medoids clustering (PAM; Partitioning Around Medoids)

Although K-means is probably the most well-known clustering algorithm, it is very sensitive to ouliers, so we chose PAM(Partitioning Around Medoids) method. Because of using the medoid rather than mean of the data points, PAM is more robust and less

sensitive to outliers.[8] But it is much slower for larger dataset as sorting is required on each iteration when computing the median vector, we selected only 300 videos for clustering.

4) Result of the clustering

The videos are divided into three classes: The high-view-group(cluster 2), middle-view-group(cluster 1), and low-view-group(cluster 3).

The result shows the fact that the more view, the more likes. So we can also write them as high-likes-group(cluster 2), middle-likes-group(cluster 1), and low-likes-group(cluster 3).

```
Numerical information per cluster:
     size max_diss  av_diss diameter separation
[1,]   90 40.11840 9.795611 49.09175   2.449493
[2,]  103 10.60069 5.456167 18.05547   2.000000
[3,]  107 20.07486 6.511234 27.94638   2.000000
```

```
> res_clst %>%
+   select(views, num_tags, title_length, cluster, is_pm, common_tag,likes) %>%
+   group_by(cluster) %>%
+   summarise(mean_views=mean(views, na.rm = TRUE), mean_numTag=mean(num_tags),
+             mean_titleLength=mean(title_length),
+             is_pm=mean(is_pm), common_tag=mean(common_tag),
+             mean_likes=mean(likes))
`summarise()` ungrouping output (override with `.groups` argument)
```

```
# A tibble: 3 x 7
  cluster mean_views mean_numTag mean_titleLength is_pm common_tag mean_likes
    <int>      <dbl>       <dbl>            <dbl> <dbl>      <dbl>      <dbl>
1       1     0.0485        33.8             33.2 0.517      0.224    -0.193
2       2      0.265        10.8             56.5 0.569      0.246     0.249
3       3     -0.261         3.42            25.0 0.364      0.195    -0.0648
```

Figure 27. Numerical information into three classes

For the each groups, we analyzed seven characteristics as follows.

- views: number of view counts
- numTag: number of tags
- titleLength: length of the title
- is_pm: if this value is 1, the video is published in the afternoon(post-mortem). if this value is 0, the video is published in the morning(anno mundi).
- common_tag: if the video is including top 100 commonly used tags, value of the common_tag is 1. if not, value of it is 0.
- likes: number of like
- category ID

[8] Preeti Arora., et al. "Analysis of K-Means and K-Medoids Algorithm For Big Data", Procedia Computer Science, Vol.78, 2016, page 508

5) Visualization

It seems that five variables are significant: numTag, titleLength, common_tag, and category ID. So we visualized differences of those five variables by clusters.

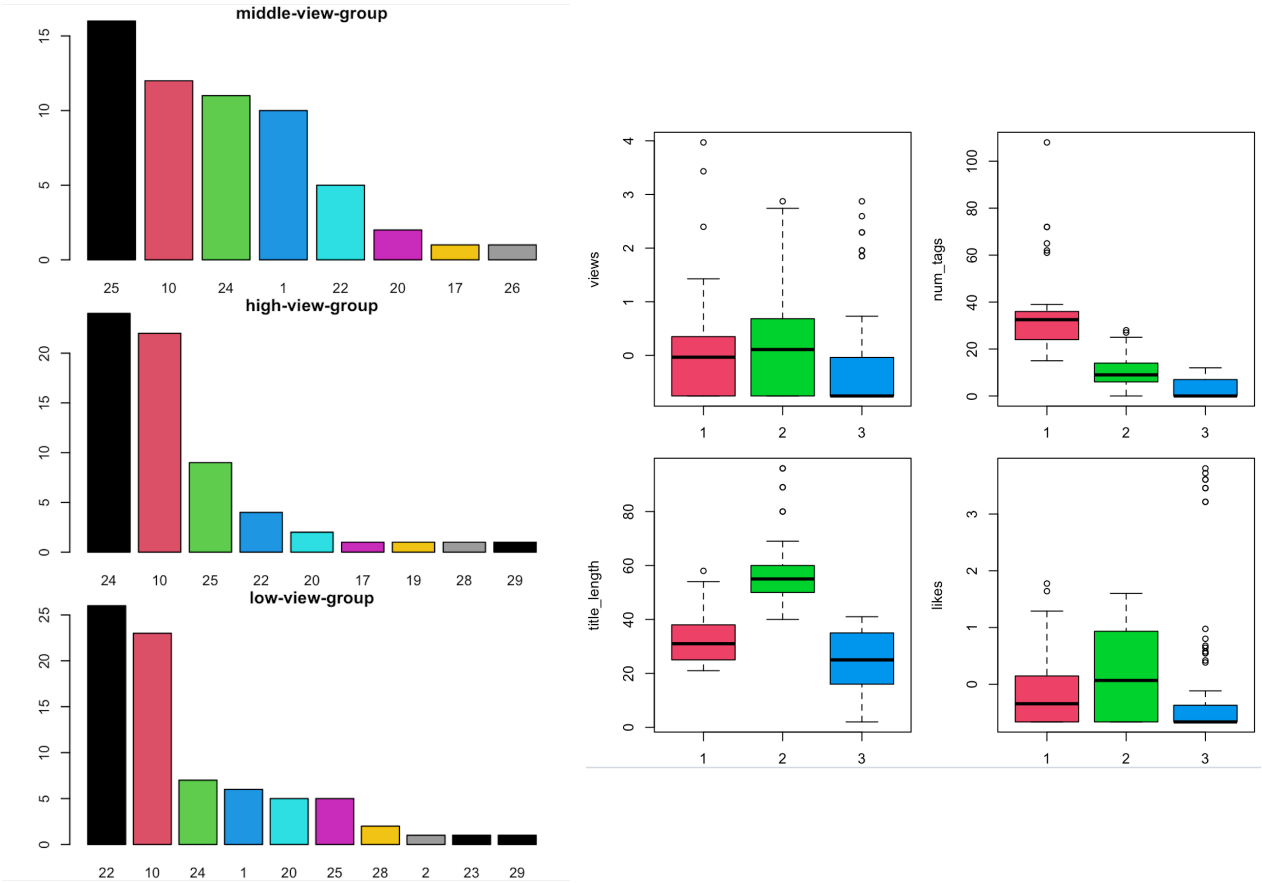


Figure 28

According to the Figure 27, the high-view-group(cluster 2) has 0~30 number of tags while other groups has more or less tags. This facts do stack up with our analysis for number of tags on 14 page: there was a tendency to increase of the view at 0~30, and after that, the number of views decreases slightly as the number of tags increases.

Length of the title is also concordant with result shown in Figure 17. Videos have 50-60 of title length are getting higher view counts. We can find that the value of 'common_tag' of cluster 2 and 1 are higher than cluster 3, but the range is not that big.

Besides, we can find tendencies of category ID among different groups:

• High-view-group: Entertainment and music has larger proportion.

- Middle-view-group: News & politics and music has larger proportion.
- Low-view-group: People & Blogs and music has larger proportion.

So we can successfully conclud that 'Music(number 10)' and 'Entertainment(number 24)' are the most popular category between YouTube trending videos, as it has been shown at Figure 20.

Conclusionally, clustering shows highly similar results with statistical analysis.

What could have been better?

That's all for our projects, but there are some further works for improvement of project. If we have more time, we are going to study about:

1. NLP: Tokenizing Korean words

   We've got rid of two problem statements about title and description: which could be very indicative but were not easy to treat. At first, we tried to use KoNLP packages to tokenize Korean words, but there are installation issues so we can't go with it. And many other popular NLP API, like googleLanguageR, doesn't support Korean.

   When problems of KoNLP are resolved, we will try again once to analyze 1)does the exposure increase when the description contains the most frequent words, and 2)does the exposure increase when the title contains the most frequent words.

2. Eject outliers & Selection algorithm

   The way we achieved better performance in the analysis was by ejecting outliers and selecting what videos should be included in test-set. We are looking forward to study more in depth those algorithms.

3. Larger dataset: beyond Korean videos

   There're not only Korean dataset, but also many other datasets of different countries' YouTube trending video at Kaggle. Working with those huge, big-dataset will be challenging, but interesting project.

References

1. Opensurvey, "Social media and Scanning portal Trend report 2020", https://blog.opensurvey.co.kr/trendreport/socialmedia-2020/

2. Kaggle, "Trending YouTube Video Statistics", https://www.kaggle.com/datasnaek/youtube-new

3. Tae Kyun Kim, "T test as a parametric statistics", Korean Journal of Anesthesiology, https://ekja.org/upload/pdf/kjae-68-540_ko.pdf

4. Steven Sawyer, "Analysis of Variance: The fundamental concepts", The Journal of manual & manipulative therapy, https://www.researchgate.net/publication/272311020

5. Eva Ostertagova., et al. "Methodology and Application of the Kruskal-Wallis Test", Applied Mechanics and Materials, https://www.researchgate.net/publication/289442433, August 2014

6. Thode Jr., H.C. (2002): Testing for Normality. Marcel Dekker, New York.

7. Preeti Arora., et al. "Analysis of K-Means and K-Medoids Algorithm For Big Data", Procedia Computer Science, Vol.78, 2016