Analysis of YouTube Trending Videos

[Team 12] 20170459 Seono Lee, 20170616 Heejin Jeong

https://github.com/suno10/2020_CS492_Project

Team 12

Introduction

R-based analysis of YouTube, especially on its ranking factors and statistics

Goal Describe characteristics of videos drawing large amount of viewers

1) Which variables are highly influential on the matter in number of views?
2) What is the better way to get more view counts?

Dataset

R-based analysis of YouTube, especially on its ranking factors and statistics

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

Dimension

34,567 rows, 16 columns

Percentage of null values

3,163 / 553,032 = 0.572%

Variable

There are total 16 variables (except index)

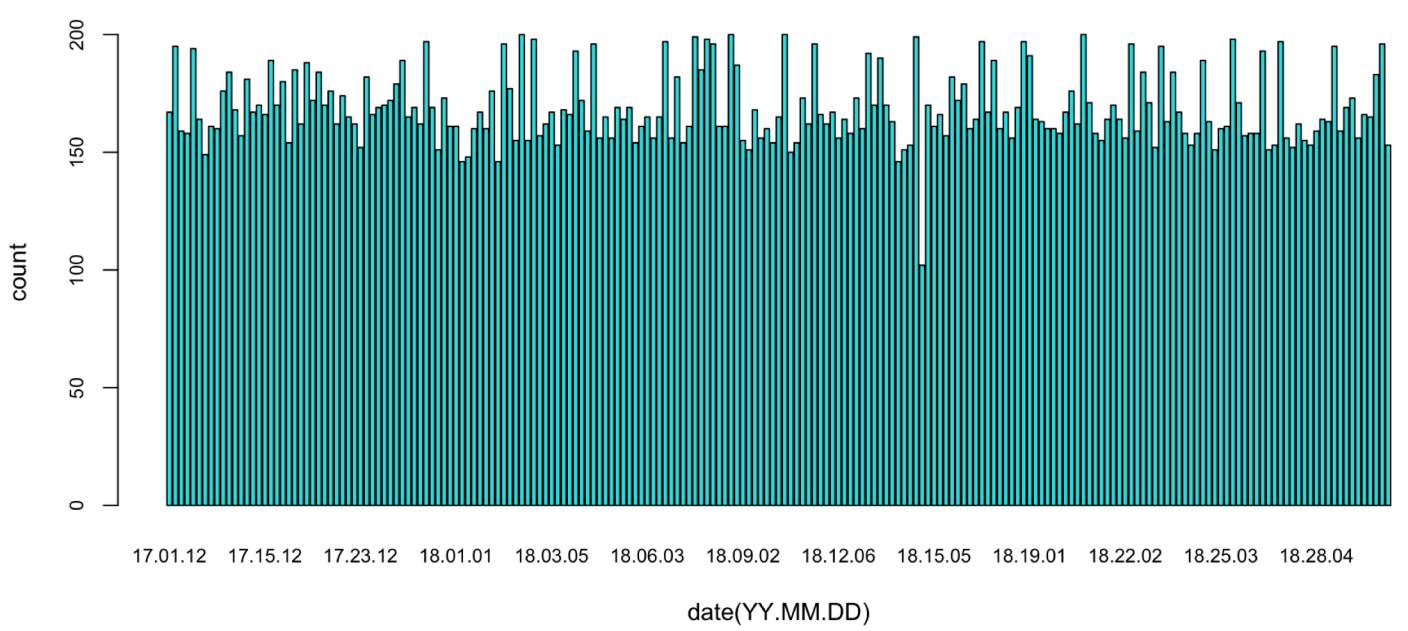
```
> colnames(KRvideos)
 [1] "video_id"            "trending_date"       "title"                  "channel_title"
 [5] "category_id"         "publish_time"        "tags"                   "views"
 [9] "likes"               "dislikes"            "comment_count"          "thumbnail_link"
[13] "comments_disabled"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

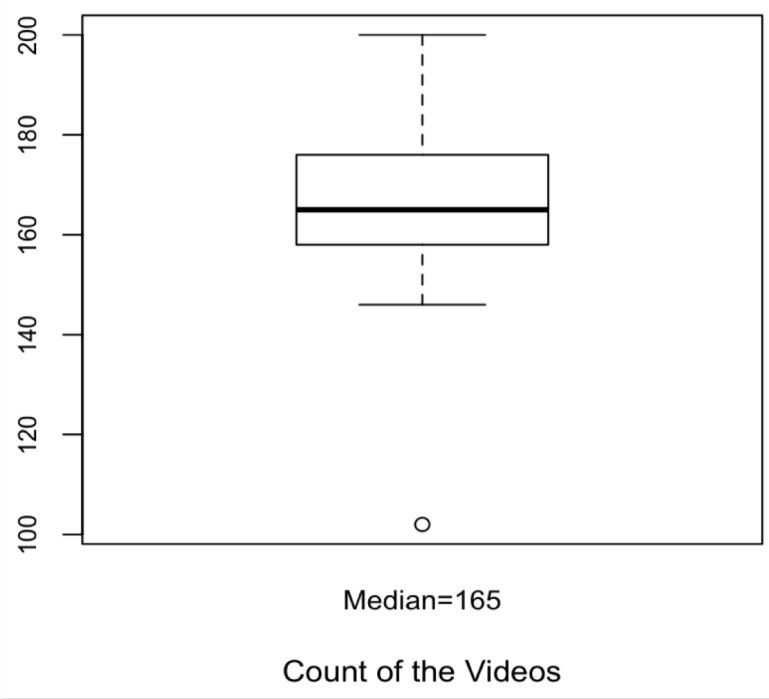
Team 12

Exploration of the Dataset

Pre-processing and review of data characteristics

Trending date





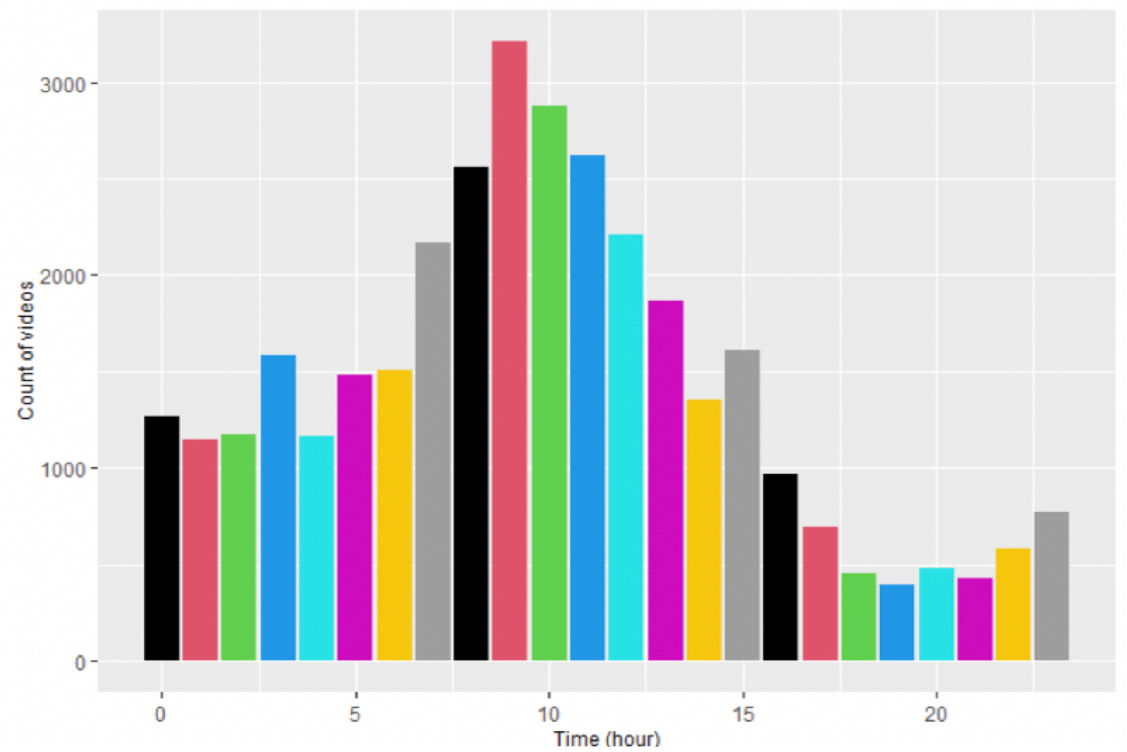
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
102	158	165	168.6	176	200

Team 12

Exploration of the Dataset

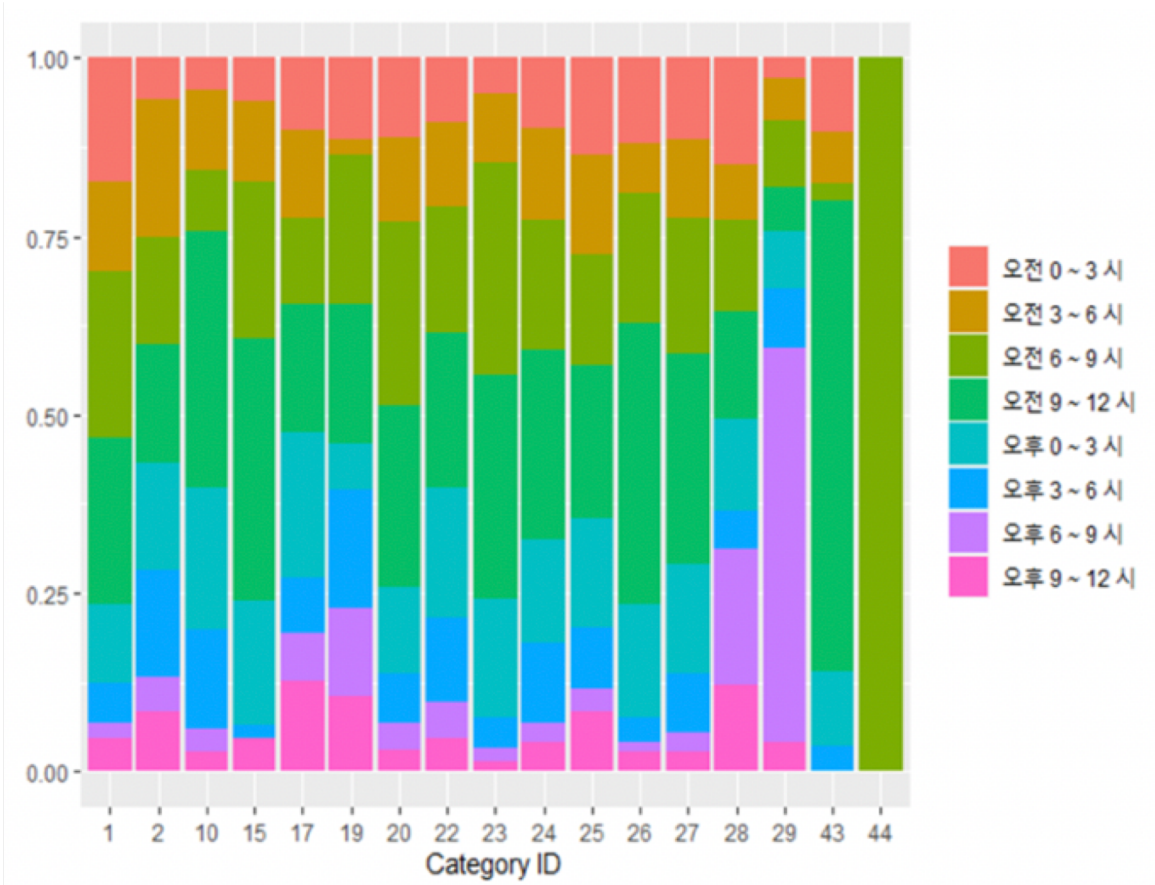
Pre-processing and review of data characteristics

Publish time



9	10	11	8	12	7
3214	2877	2623	2563	2206	2169

sort(table(time_hour), decreasing=true)[1:6]

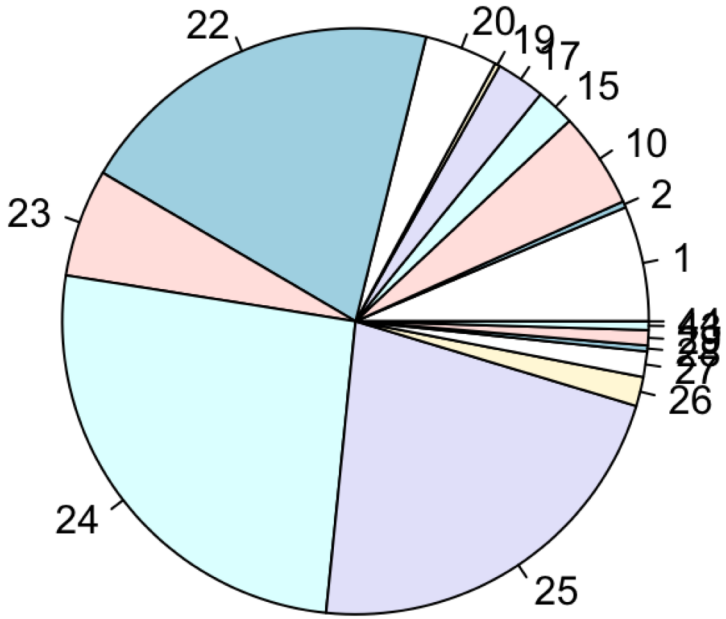


Team 12

Exploration of the Dataset

Pre-processing and review of data characteristics

Category ID



"Entertainment" occupies

the largest percentage of 25.91%

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

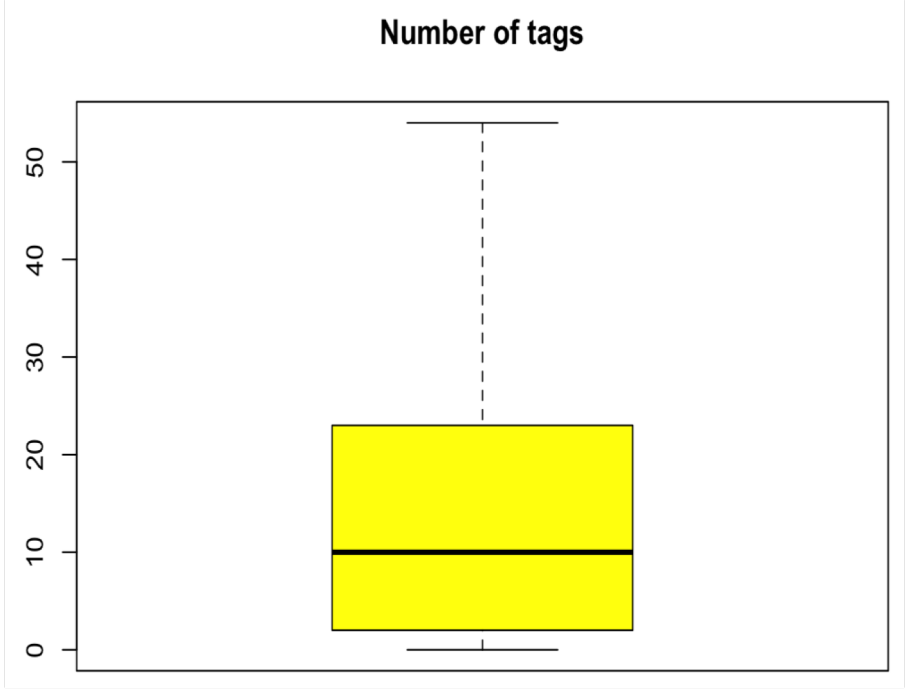
Team 12

Exploration of the Dataset

Pre-processing and review of data characteristics

Tags

word-cloud



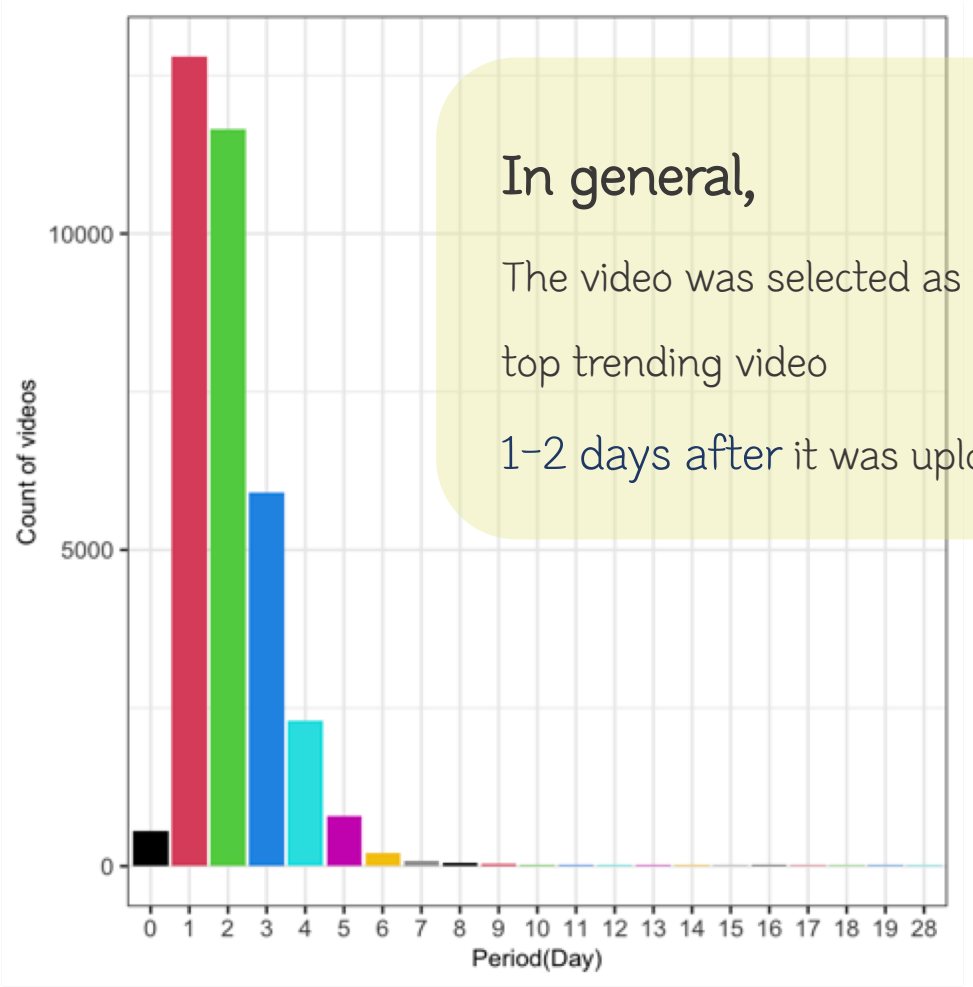
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	2	10	16.22	23	152

Team 12

Exploration of the Dataset

Pre-processing and review of data characteristics

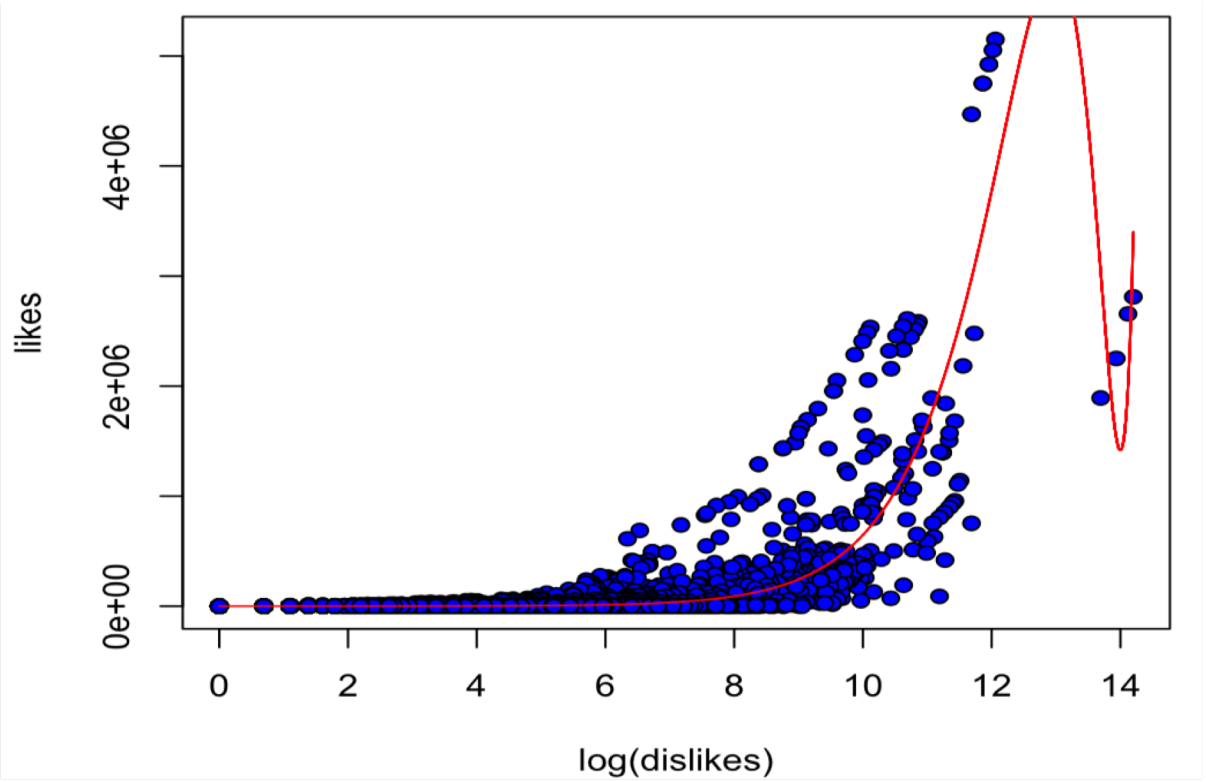
Time interval between Trending & Publishing



In general,

The video was selected as the top trending video 1-2 days after it was uploaded.

Likes & Dislikes



The red line follows polynomial regression of order 3

Team 12

Analysis of ranking factors

Detailed ranking factors with some statistical schemes

Problem Statement

We will make a subset of videos from the whole dataset for each problem statements, and analyze them with statistical methods to find answer for the questions :

- What is the length of the title that increases the exposure most?
- Does the exposure increase when the tags contain the most frequent words?
- What is the number of the tags that increases the exposure most?
- What are popular and unpopular categories?
- All of those factors are statistically meaningful? If they are, how much meaningful?

Team 12

Analysis of ranking factors

Detailed ranking factors with some statistical schemes

Approach

- **Determine the normality**

  - Use shapiro test or cvm test(Cramer-vom Mises test)

- If it follows normal distribution: **T-test or ANOVA**(analysis of variance) test
- If it doesn't follow normal distribution: Take **Kruskal-Wallis rank sum test**

You can get further information here:

T-test as a parametric statistics: https://ekja.org/upload/pdf/kjae-68-540_ko.pdf
ANOVA - The fundamental concepts: https://www.researchgate.net/publication/272311020
Methodology and Application of the Kruskal-Wallis Test: https://www.researchgate.net/publication/289442433

Team 12

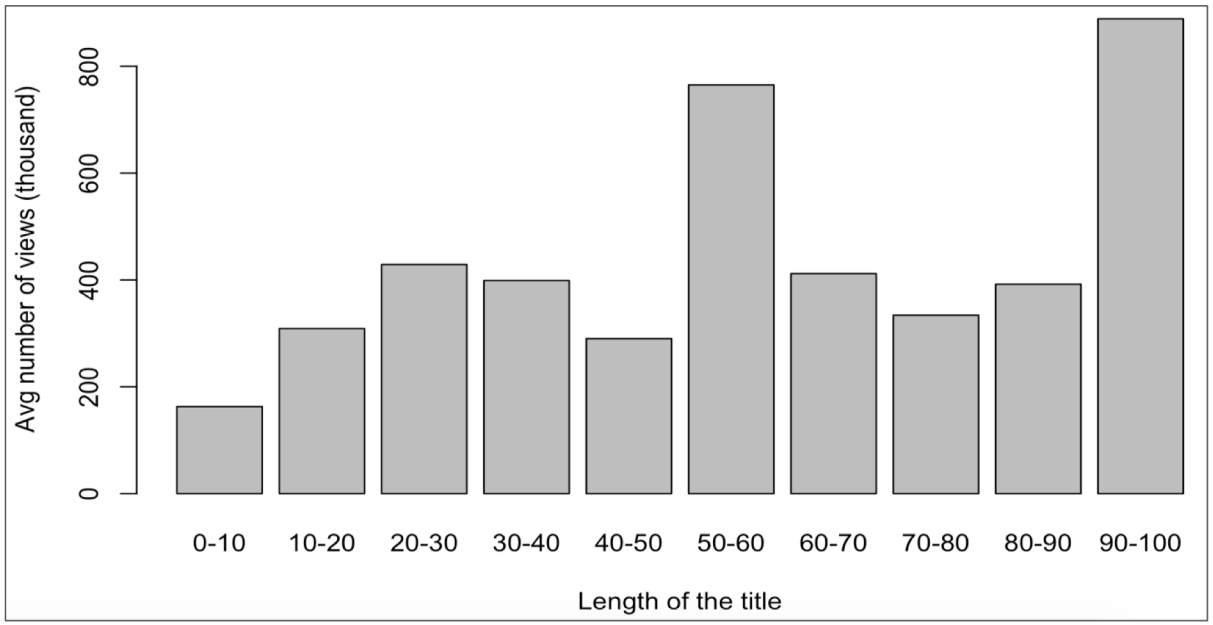
Analysis of ranking factors

Detailed ranking factors with some statistical schemes

Length of Title

```
        Kruskal-Wallis rank sum test

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



- Length range of All vidoes

```
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
```

- Length range 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
```

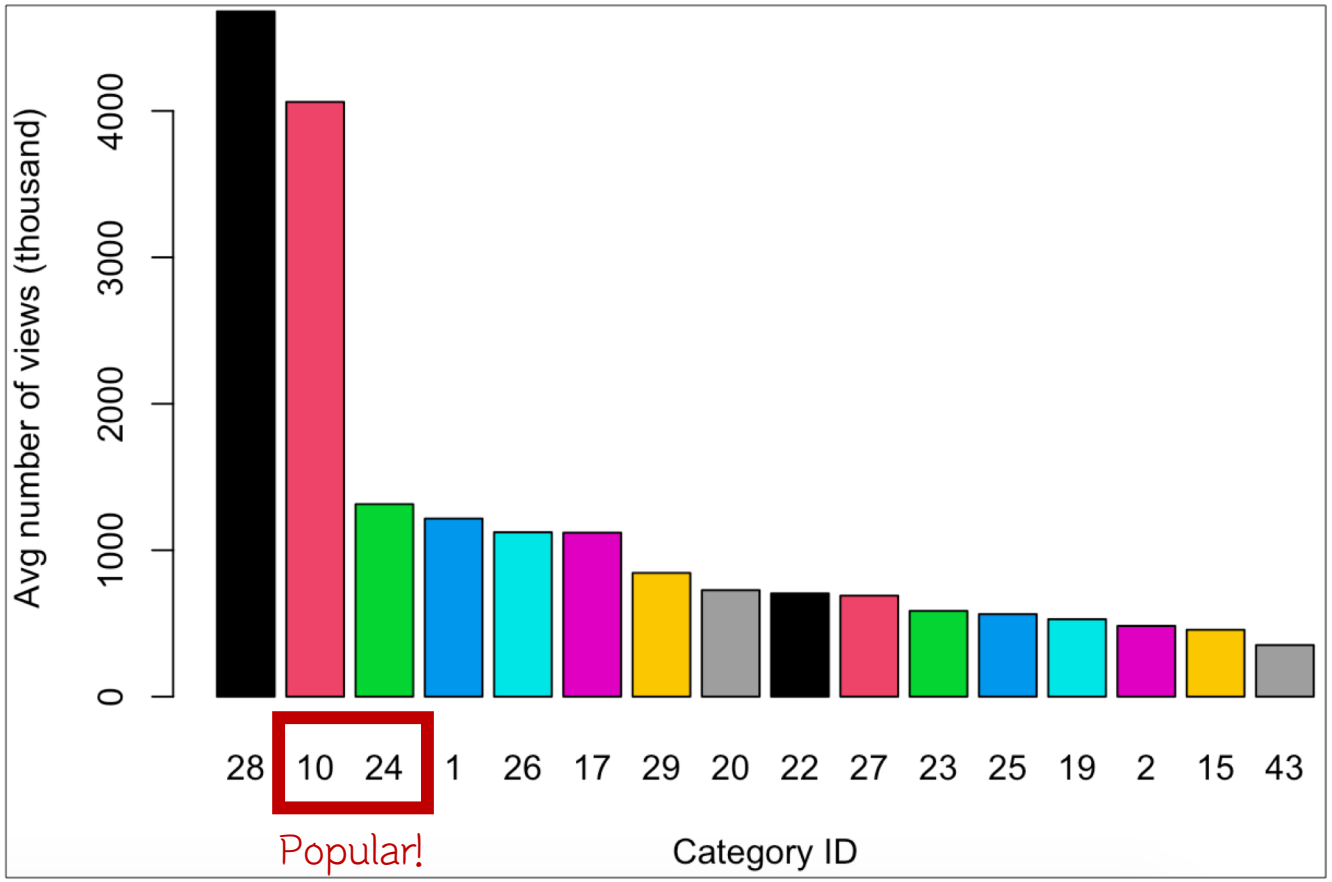
Videos that have 50-60 characters of title have the highest percentage of top 500 videos.

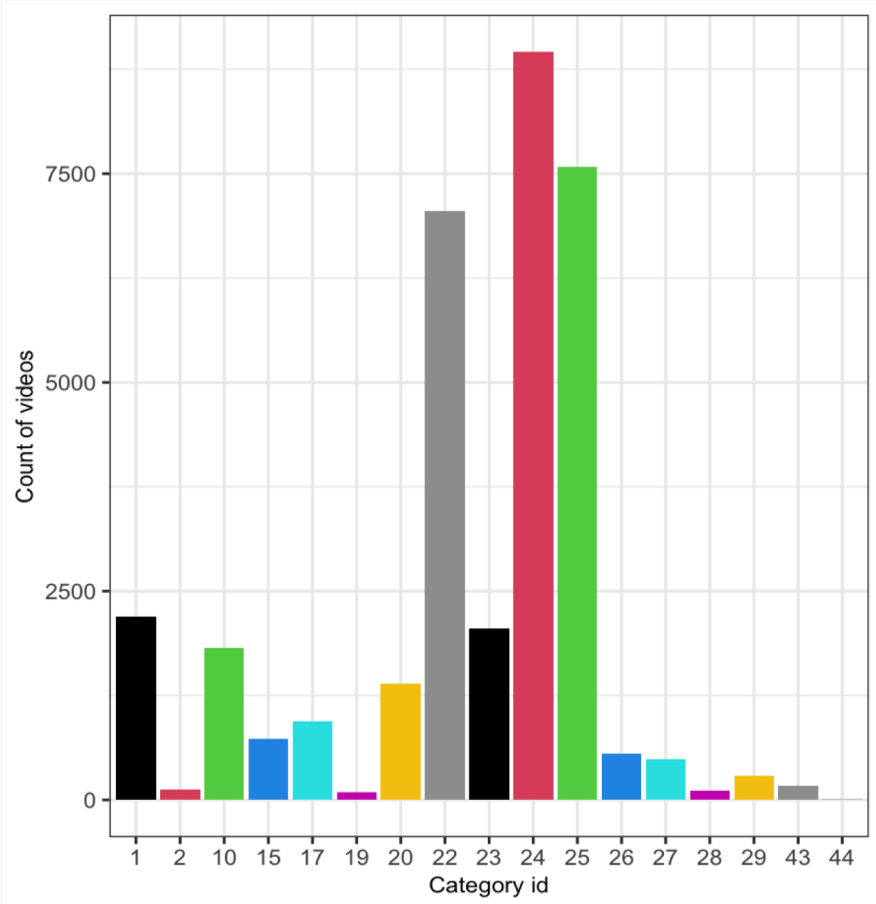
Team 12

Analysis of ranking factors

Detailed ranking factors with some statistical schemes

Category ID





Team 12

Analysis of ranking factors

Detailed ranking factors with some statistical schemes

Number of the tags

- Data pre-processing

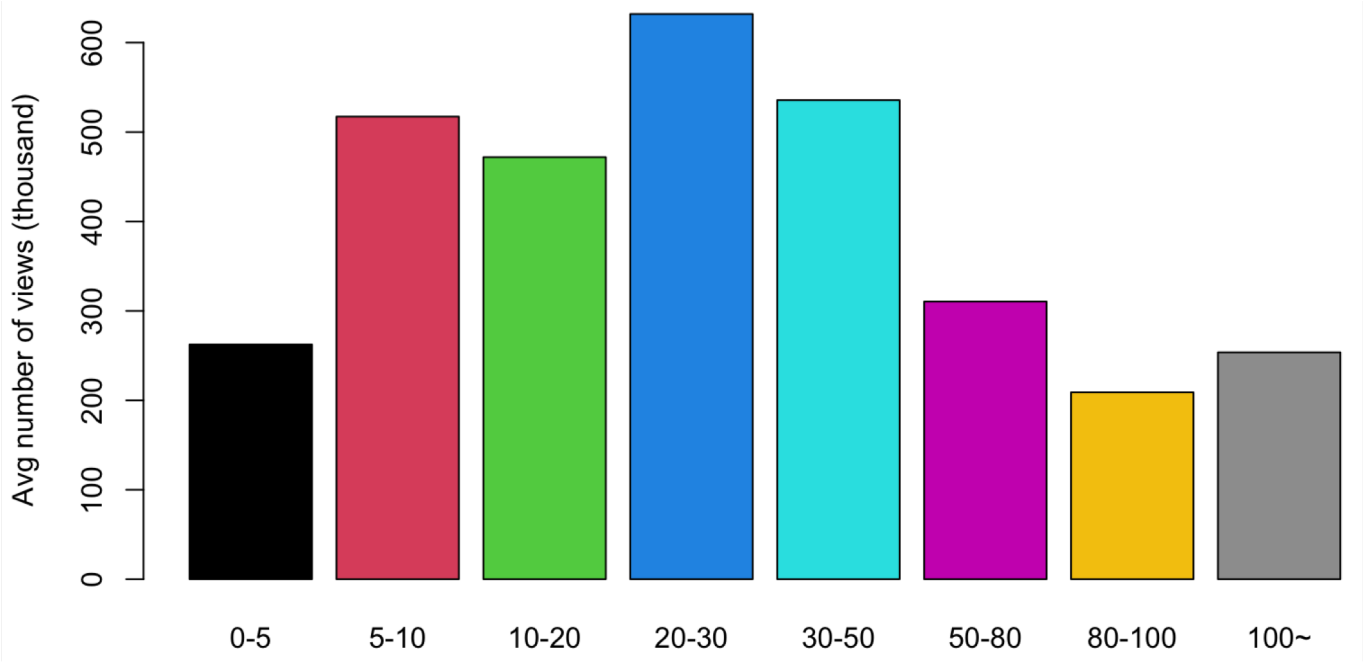
We made number of the tags as a factor, separated into 8 levels.

```
> 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
```

- Visualization



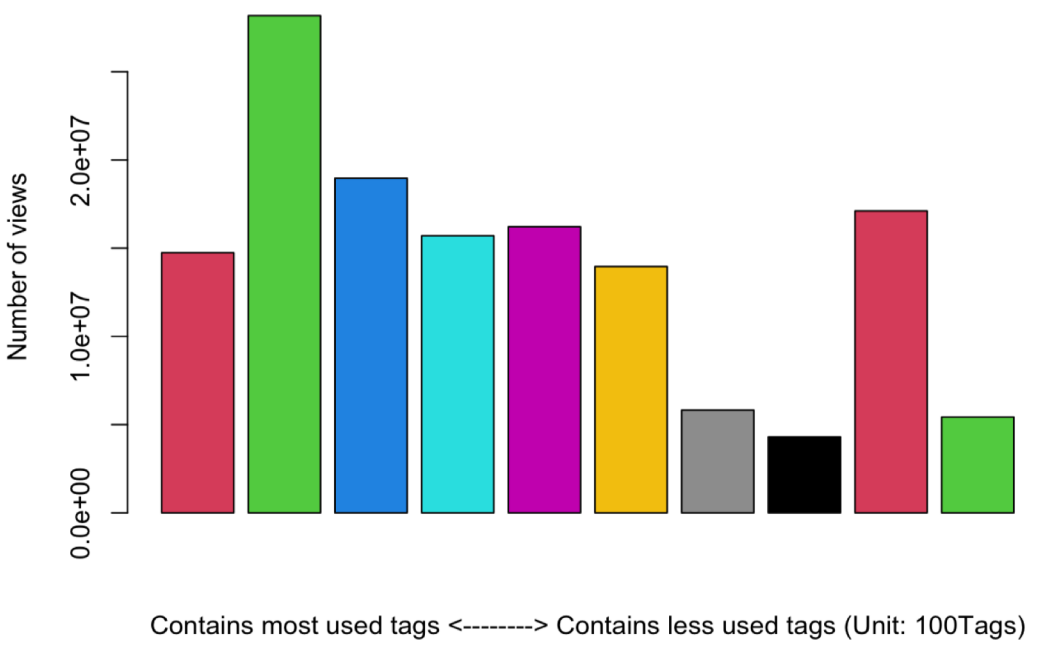
Team 12

Analysis of ranking factors

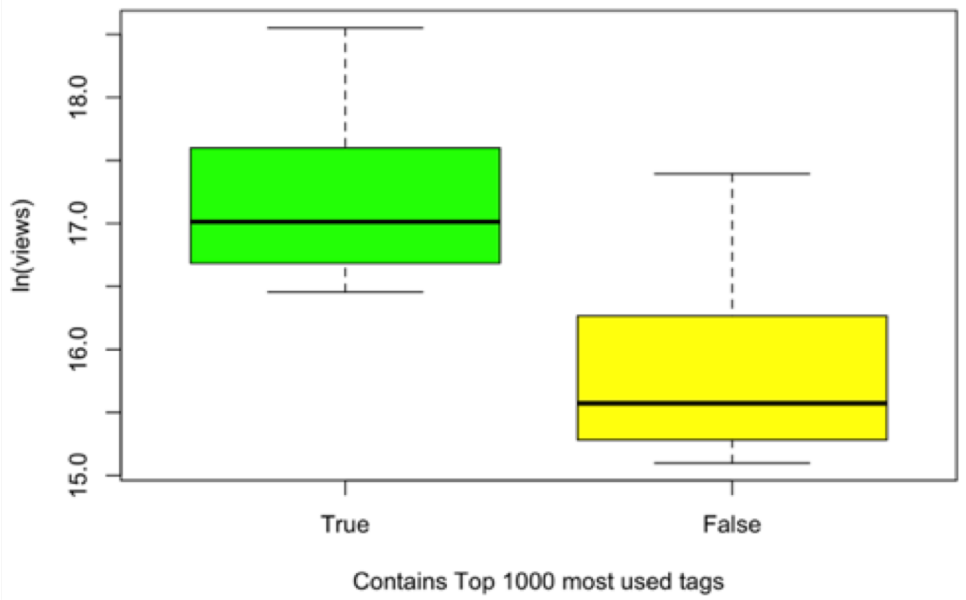
Detailed ranking factors with some statistical schemes

Common tags

- Changes in views according to common tags ranking



- Analysis of view of videos with / without top 1000 tags



Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
14013696	17603335	24445522	34956794	43412762	113876217

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
3603101	4342583	5795039	9320700	11453298	35832484

Team 12

Deep-dive for Entertainment

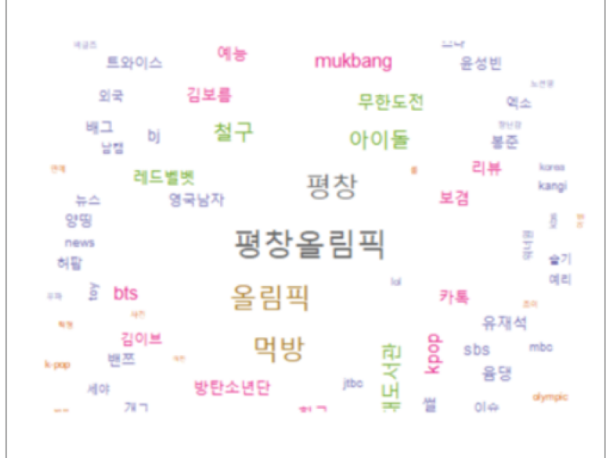
Detailed analysis for Entertainment category

- Popular tags of each month (from 2017.11. to 2018.06.)

November 2017

December 2017

January 2018



February 2018

March 2018

April 2018

May 2018

June 2018

Deep-dive for Entertainment

Detailed analysis for Entertainment category

- Top 10 tags in the category (from 2017.11. to 2018.06.)

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

- 유명 BJ
- 먹방 (mukbang)
- 샤이니, 평창 올림픽, 무한도전
- 아이돌 컴백 시즌

Team 12

Conclusion & Clustering

As a result, what makes the number of views high?

Summary

- Four Advices for getting higher views

1. Title length: 50-60 characters
2. Number of tags: 20
3. Category: Music, Entertainment
4. Include commonly used tags

Clustering

- To sum-up,

1) 조회수별 3개 그룹(high, middle, low)에서
   각각 100개씩 video 추출
2) 300개 video로 Clustering 진행
3) 나누어진 Cluster들의 characteristic이
   실제 analysis result와 일치하는지 비교

Conclusion & Clustering

As a result, what makes the number of views high?

K-medoids (PAM)

- K-means vs K-medoids

  Although K-means is the most well-known clustering algorithm, it is very sensitive to outliers.

  So, we chose PAM (Partitioning Around Medoids) method, rather than K-means.

- PAM

  using the medoid rather than mean of the data points, less sensitive but much slower for larger dataset.

  So, we selected only 300 videos for clustering

You can get further information here: Analysis of K-Means and K-Medoids Algorithm For Big Data

Team 12

Conclusion & Clustering

As a result, what makes the number of views high?

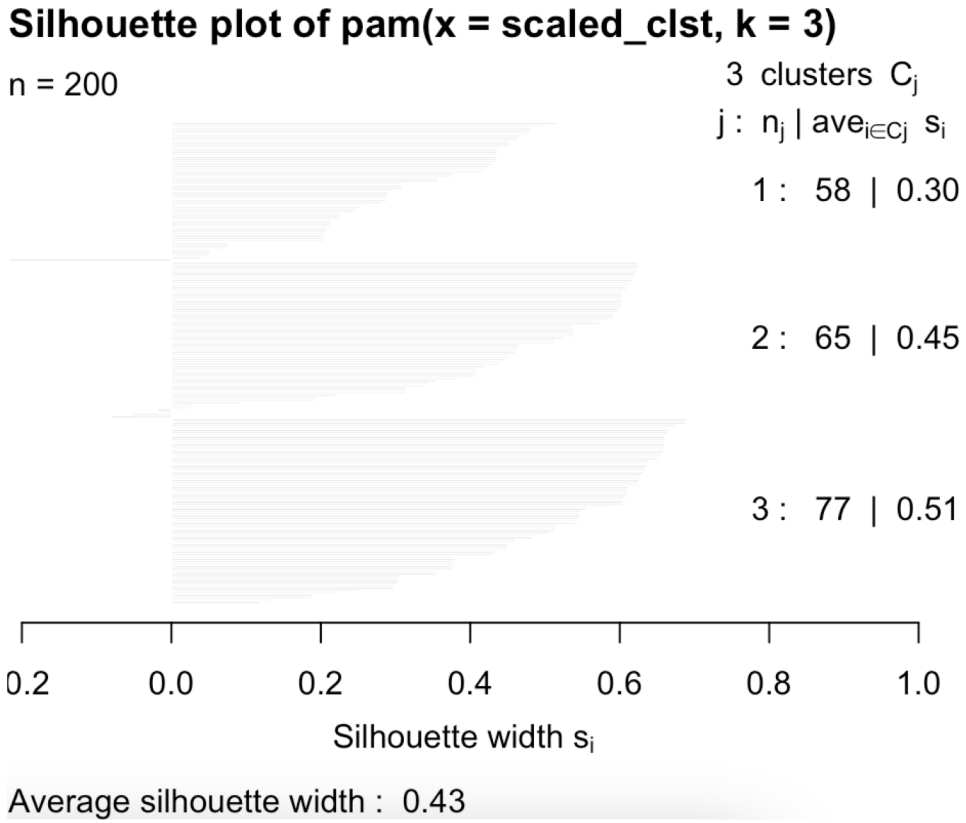
Clustering

- with scaled data using scale() function

```
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	0.0485	33.8	33.2	0.517	0.224	-0.193
2	0.265	10.8	56.5	0.569	0.246	0.249
3	-0.261	3.42	25.0	0.364	0.195	-0.0648

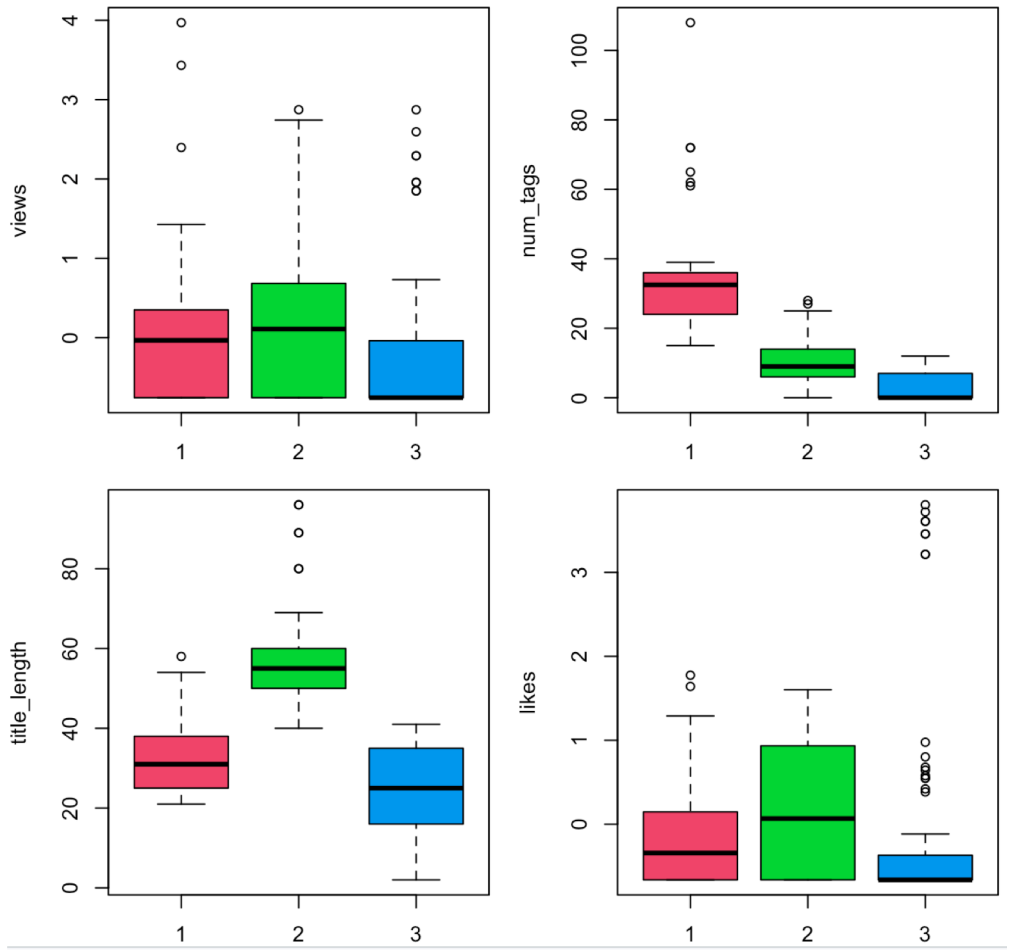


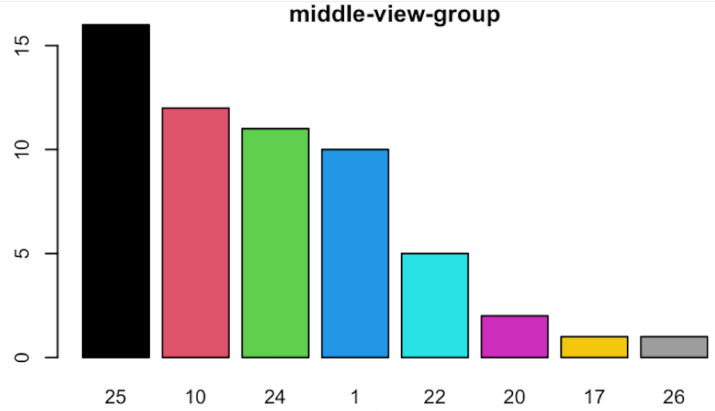
Team 12

Conclusion & Clustering

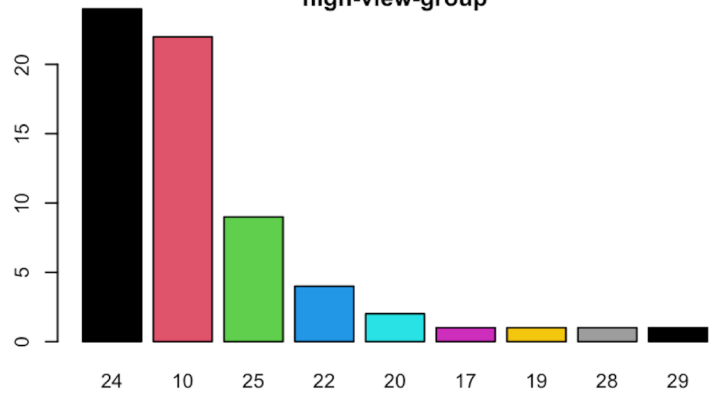
As a result, what makes the number of views high?

Clustering

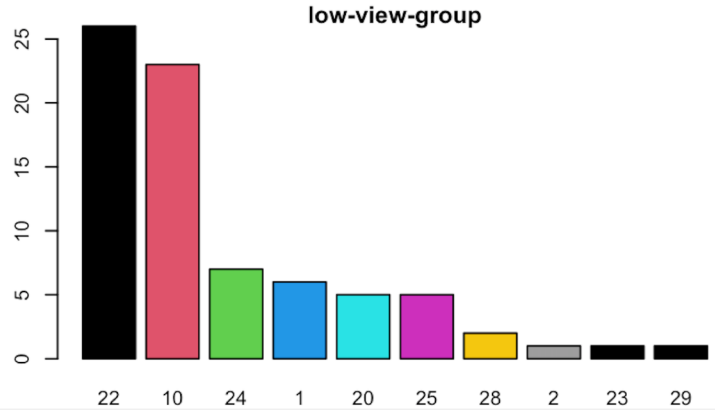




- 25: News & Politics
- 10: Music



- 24: Entertainment
- 10: Music



- 22: People & Blogs
- 10: Music

Further works

What could have been better for our project?

Significance of the project

R-based analysis on **Korean videos**, using many **statistical tools** and hypothesis

Further works to do

- **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 because:

  1) KoNLP: installation issues

  2) Other popular NLP API: not support Korean language

- **Eject outliers & Selection algorithm**
- **Larger dataset: beyond Korean videos**

References

- Opensurvey, "Social media and Scanning portal Trend report 2020", https://blog.opensurvey.co.kr/trendreport/socialmedia-2020/
- Kaggle, 'Trending YouTube Video Statistics'(https://www.kaggle.com/datasnaek/youtube-new)
- Tae Kyun Kim, "T test as a parametric statistics", Korean Journal of Anesthesiology, https://ekja.org/upload/pdf/kjae-68-540_ko.pdf
- Steven Sawyer, "Analysis of Variance: The fundamental concepts", The Journal of manual & manipulative therapy, https://www.researchgate.net/publication/272311020
- Eva Ostertagova., et al. "Methodology and Application of the Kruskal-Wallis Test", Applied Mechanics and Materials, https://www.researchgate.net/publication/289442433, August 2014
- Thode Jr., H.C. (2002): Testing for Normality. Marcel Dekker, New York.
- Preeti Arora., et al. “Analysis of K-Means and K-Medoids Algorithm For Big Data”, Procedia Computer Science, Vol.78, 2016