기업과제 2번 데이터 분석 및 시각화

-Junior_최진수

목차

Q1

0. 데이터 분석을 위한 전처리
1. 전체기간 카테고리
2. 월별 카테고리, 채널, 동영상 분석
3. 월별 채널 TOP 10
4. 주차별 Top 5 채널
5. 월별 카테고리 키워드 순위
6. Q1 결론

Q2

0. engagement 구하기
1. 가설 제시 및 검증
2. Q2 결론

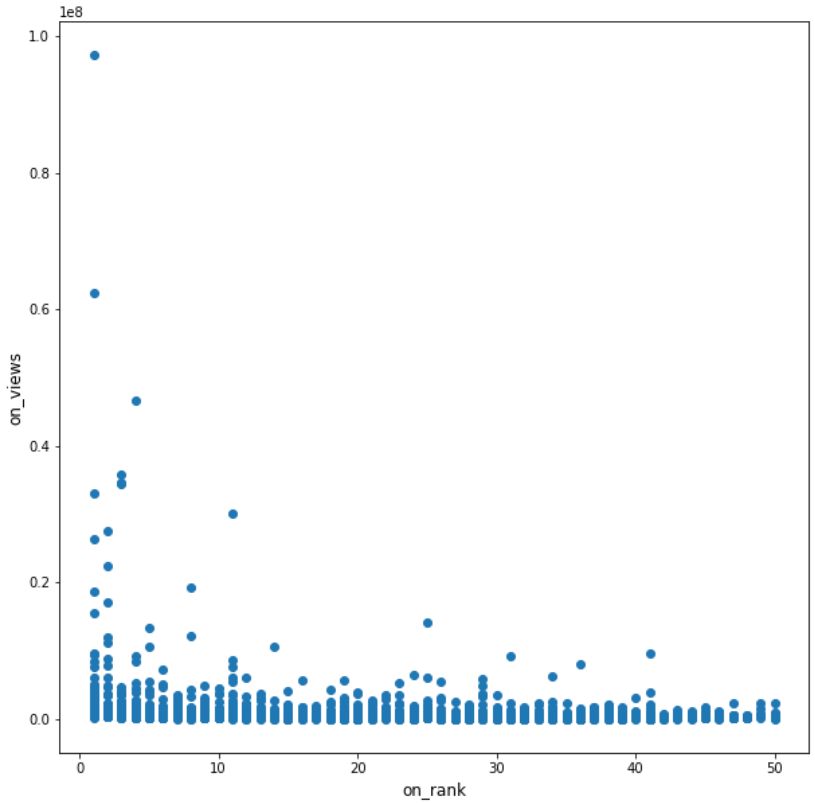
Q1-0 데이터 분석을 위한 전처리

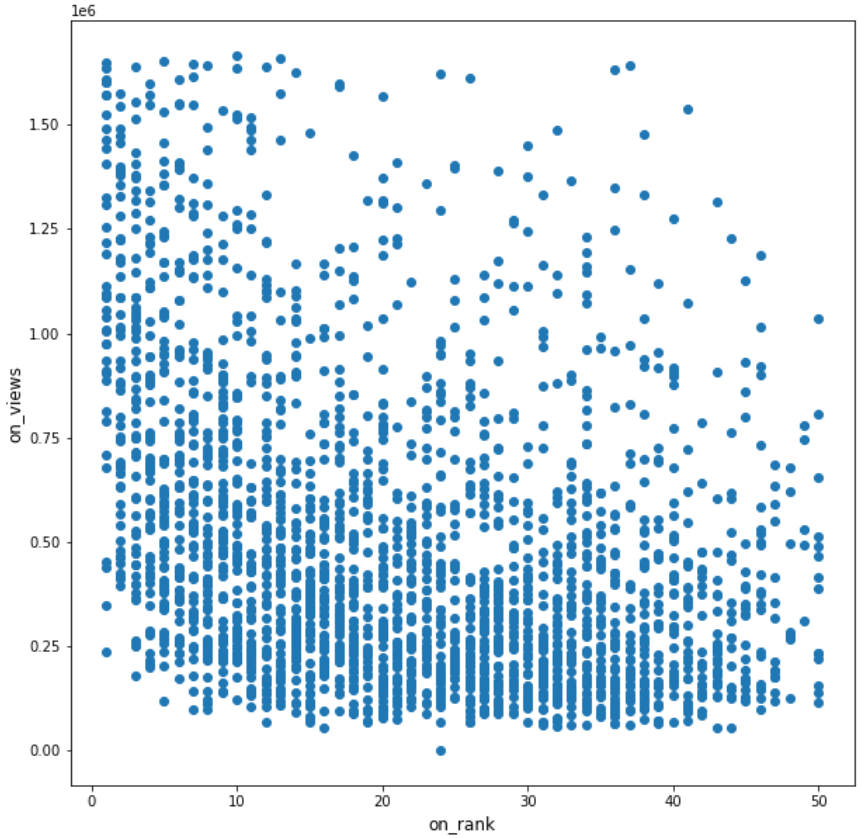
video_id	0
channel_id	0
published_date	0
category_name	0
duration	0
tags	325
description	35
on_trending_date	0
off_trending_date	0
on_rank	0
off_rank	0
on_views	0
off_views	0
on_likes	0
off_likes	0
on_dislikes	0
off_dislikes	0
on_comments	0
off_comments	0

on_channel_subscribers	0
off_channel_subscribers	0
on_channel_total_views	0
off_channel_total_views	0
on_channel_total_videos	0
off_channel_total_videos	0
published_on_date	0
on_off_date	0
engagement	0
normal_engagement	0
minmax_engagement	0

Tags, description 결측치가 있지만
구체적인 수치 데이터가 아니므로 유지

Q1-0 데이터 분석을 위한 전처리





X축: on_rank, y축:on_views 값들이 편차가 너무 크므로 각 rank마다 on_views top5 전처리

상대적으로 고른 분포를 보이게 됨

Q1-0 데이터 분석을 위한 전처리

```python
## 조회수 0인데 인기  1 순위 결측치
print(f'조회수 결측치 총: {len(df[df.on_views == 0])} 개')
df = df[df.on_views != 0]

#좋아요, 싫어요 놓은 결측치 제거
print(f'좋아요 결측치 총: {len(df[df.on_likes == 0])} 개')
df = df[df.on_likes!=0]

## 채널 구독자 비공개 결측치 제거
print(f'구독자수 결측치 총: {len(df[df.on_channel_subscribers == 0])} 개')
df = df[df.on_channel_subscribers!=0]

## 댓글 결측치 (인기 동영상일 때는 공개하고 인기동영상 끝날때 비공개 댓글도 결측치로봄)
print(f'ON 댓글 결측치 총: {len(df[df.off_comments == 0])} 개')
print(f'OFF 댓글 결측치 총: {len(df[df.on_comments == 0])} 개')
df = df[df.off_comments!=0]
df = df[df.on_comments!=0]

print()
print(f'현재 결측치 제거 후 데이터 수: {len(df)} 개')
```

```
조회수 결측치 총: 1 개
좋아요 결측치 총: 19 개
구독자수 결측치 총: 55 개
ON 댓글 결측치 총: 12 개
OFF 댓글 결측치 총: 12 개

현재 결측치 제거 후 데이터 수: 2306 개
```

기타 결측치 제거 후 총 데이터 수 2306개

Q1-0 데이터 분석을 위한 전처리

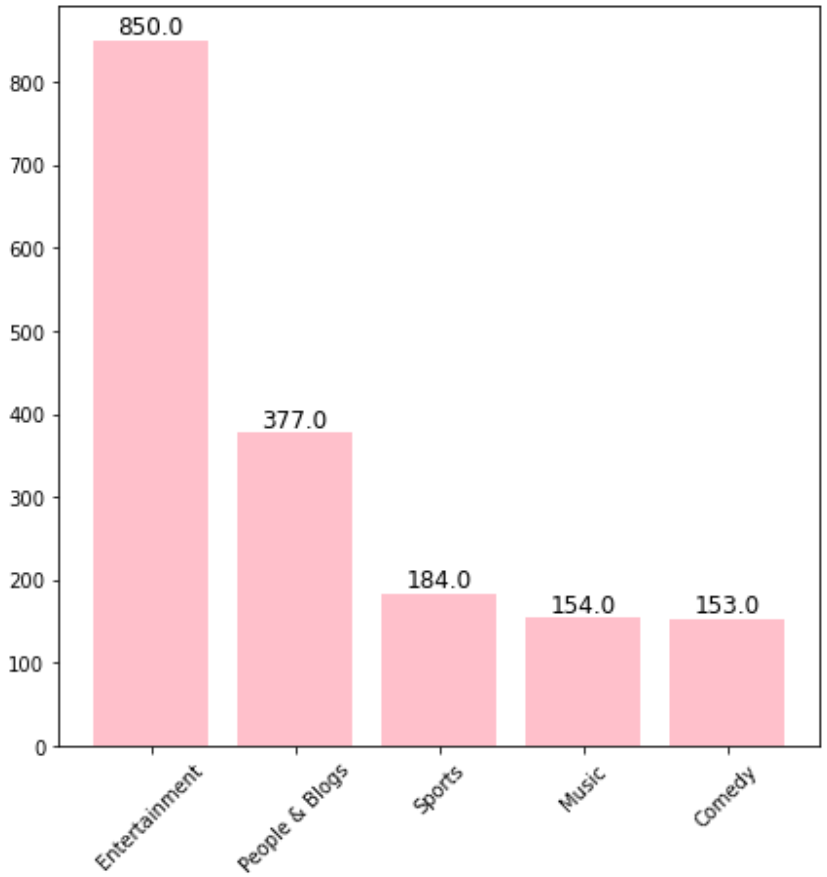
```
1  print('인기동영상이 되기 위해서는')
2  print(f'평균 {df.on_views.mean().round()} views 필요하다.')
3  print(f'최소: {df.on_views.min().round()} views 가 필요하다')
```

```
인기동영상이 되기 위해서는
평균 496290.0 views 필요하다.
최소: 53297 views 가 필요하다
```

인기 동영상은 평균 496,000 views 최소 53,000 views 필요

Q 1-1 전체기간 카테고리

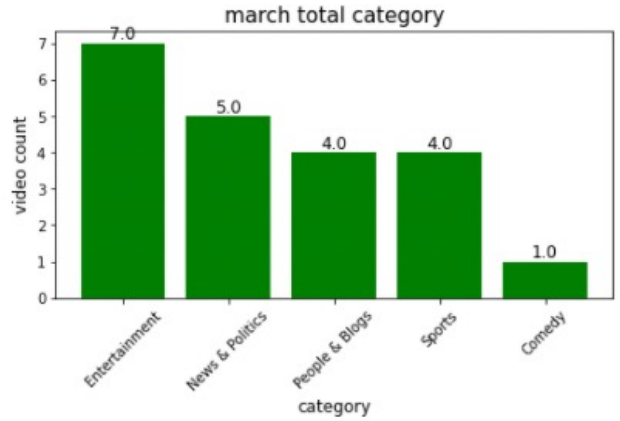
```
['Entertainment' 'Sports' 'Music' 'People & Blogs' 'Science & Technology'
 'Education' 'Howto & Style' 'News & Politics' 'Gaming' 'Comedy'
 'Pets & Animals' 'Travel & Events' 'Film & Animation' 'Autos & Vehicles'
 'Nonprofits & Activism']

카테고리 개수: 15
```

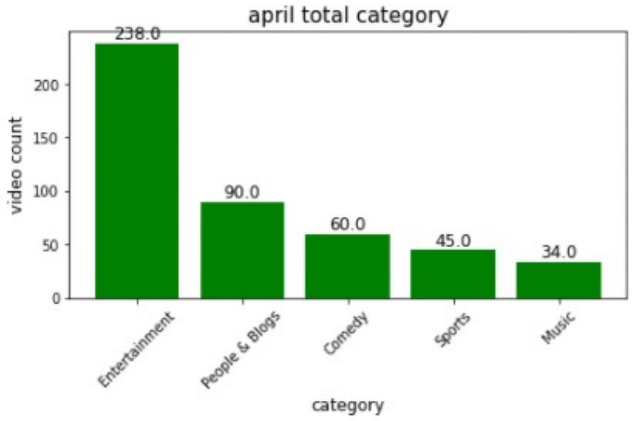


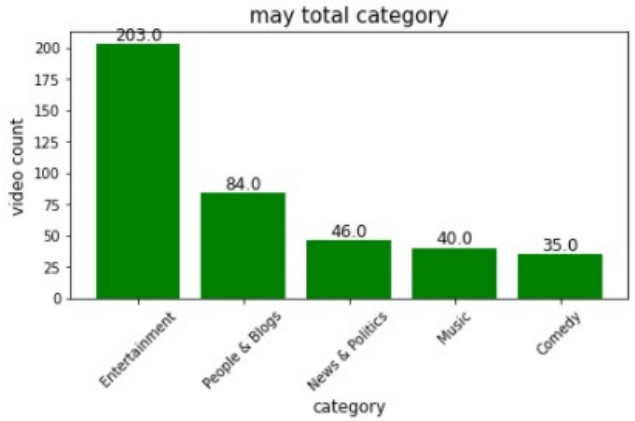
총 15개의 카테고리가 존재함

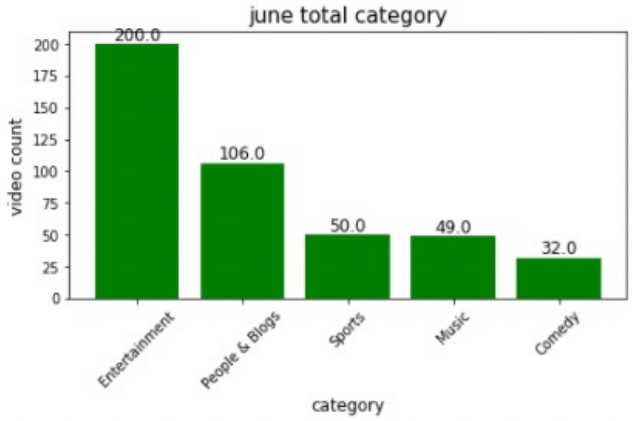
entertainment의 압도적인 동영상 개수

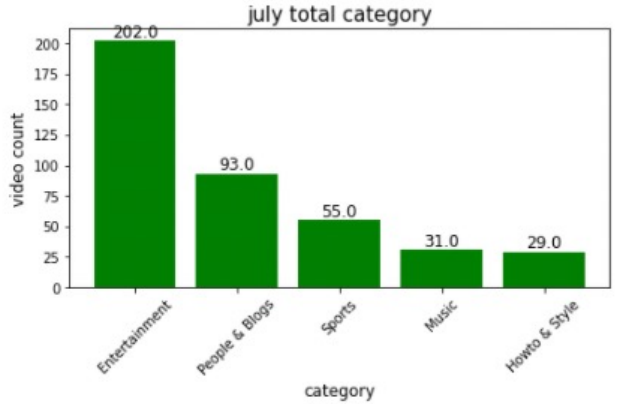
Q1-2 월별 카테고리별 채널 개수





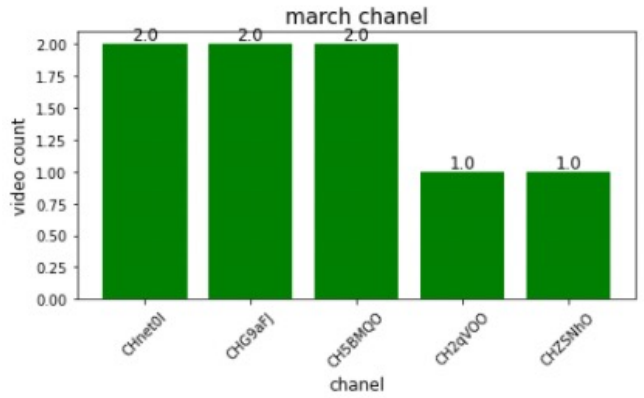


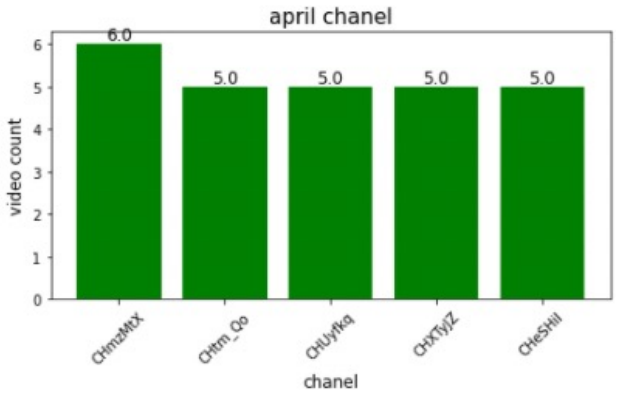


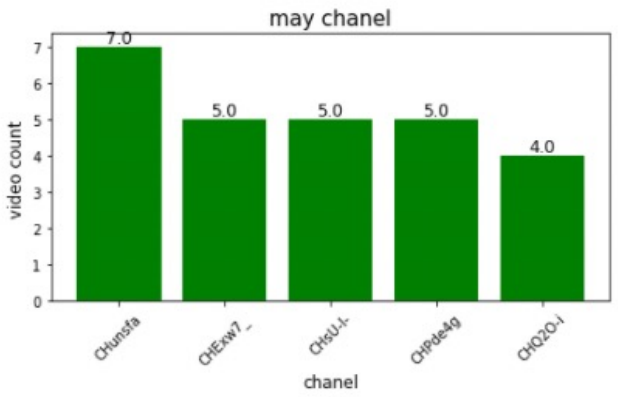


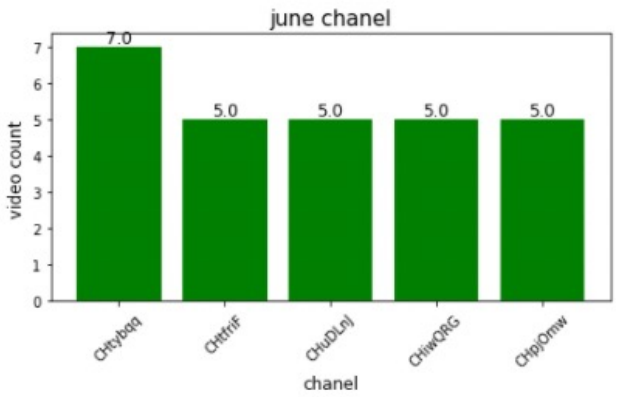
전반적으로 entertainment 카테고리 채널이
많은 인기동영상을 가지고 있음

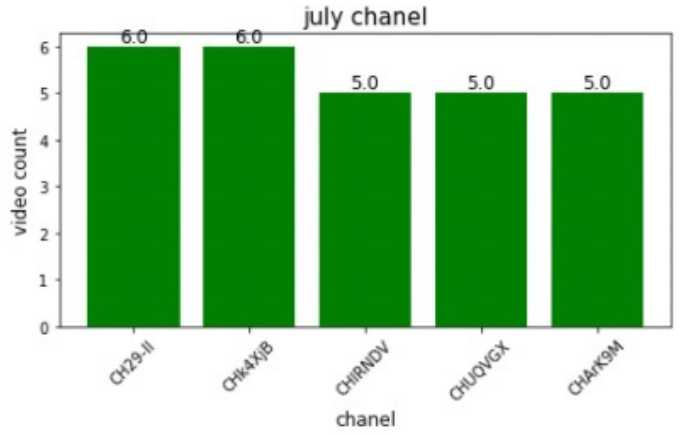
Q1-2 월별 채널 동영상 개수





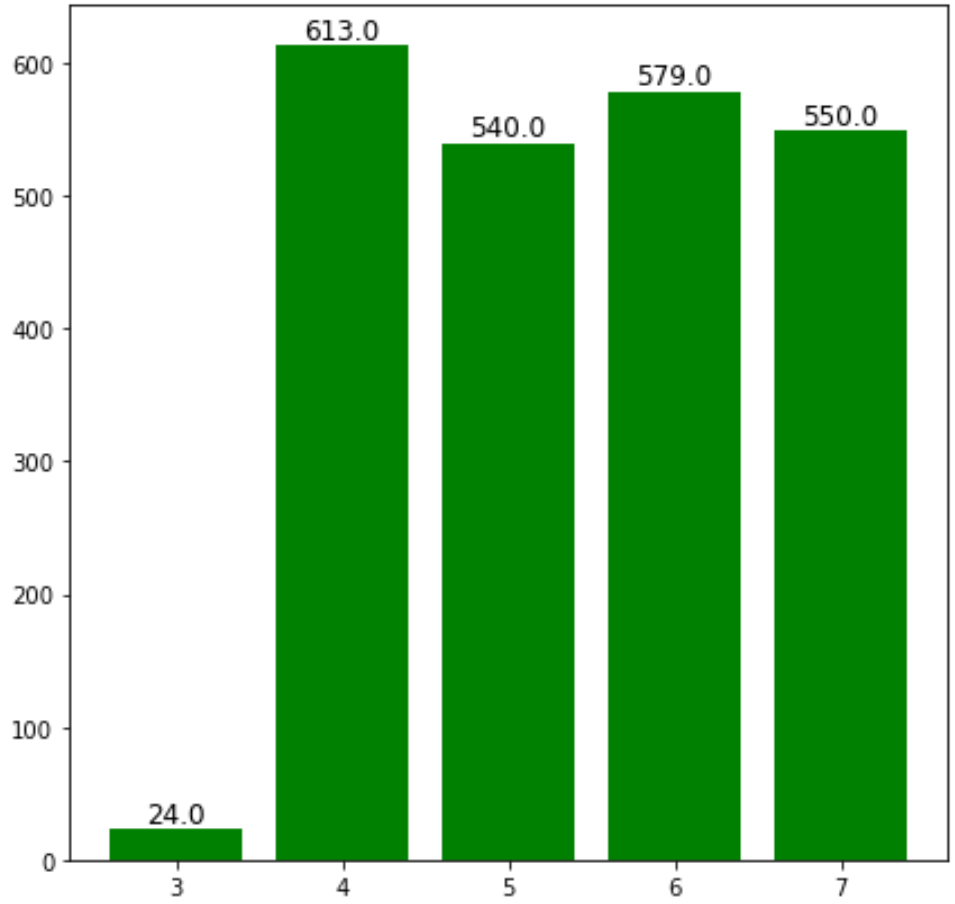






여러 인기동영상을 가지고 있는 상위 채널은
보통 5개 정도를 가지고 있는 것으로 보임

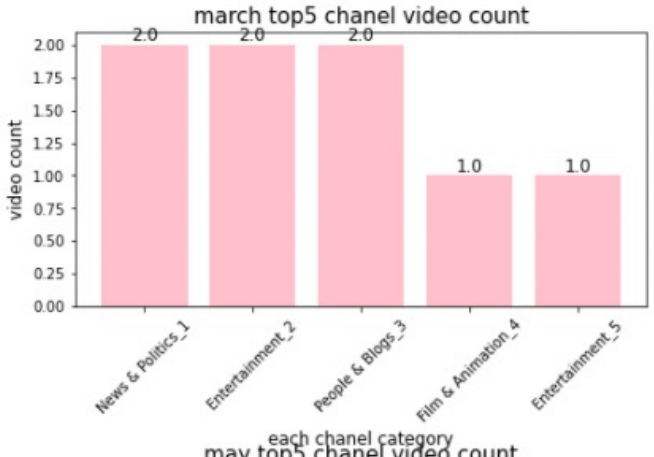
Q1-2 월별 동영상 개수

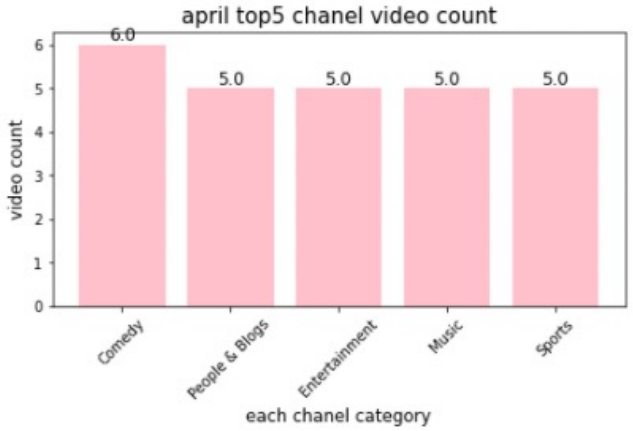


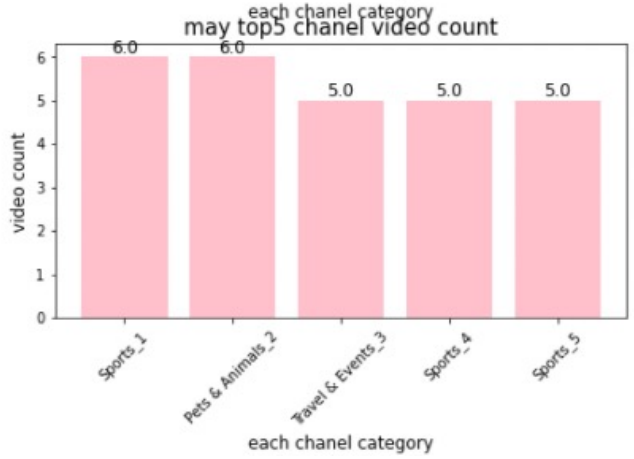
3월 24개, 4월 613개, 5월 540개,6월 579개 7월 550개이며

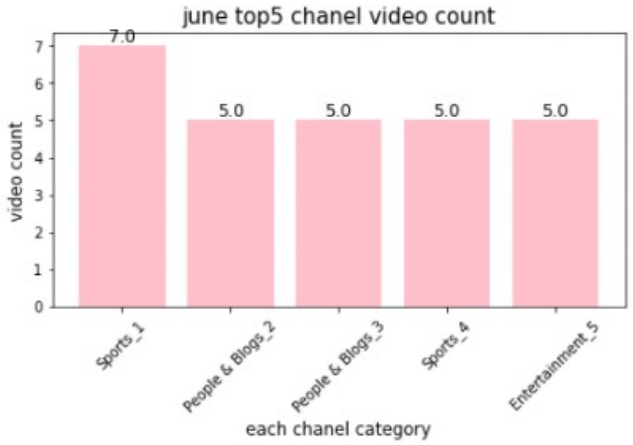
2021년도 데이터이며 3월달을 제외한 모든 데이터가 비교적 균등하게 있습니다.

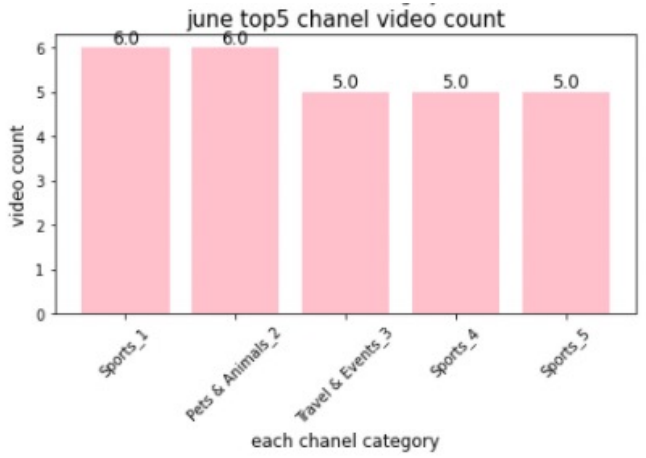
Q 1-3 월별 채널 TOP 10











월별 채널 top 5 카테고리 별로 분석해 본 결과

3월은 entertainment 채널이 압도적으로 많은 분포를 보였으며
4~7월에는 sport 채널이 많이 분포 되어있었습니다

top채널은 최소 2개~9개 사이에 인기동영상이 있었습니다.

Q 1-4 주차별 Top 5 채널

12 Week부터 30 week까지 채널 분석 결과 최대 5개까지
인기동영상을 가질 수 있는 것으로 나타났다.

Q1-5월별 카테고리 키워드 순위

박수홍	3
Korean	3
먹방	3
가족	2
뉴스	2

3월 카테고리 키워드는
동영상이 적어 tag 카운트
작은 분포가 보였다.

Q1-5 월별 카테고리 키워드 순위

먹방	34
예능	24
레전드	23
몰카	22
브레이브걸스	22

4월은 먹방, 예능 등
tag가 높은 순위를 차지했다.

Q1-5 월별 카테고리 키워드 순위

먹방	38
예능	19
뉴스	19
아이돌	17
라이브	13

5월 또한 먹방, 예능 tag가 상위권을 차지했으며 뉴스 아이돌이 뒤를 이었다.

Q1-5 월별 카테고리 키워드 순위

먹방	45
예능	28
축구	21
브이로그	20
mukbang	18

6월 또한 먹방, 예능 tag가
높은 tag 순위를 차지했으며
축구, 브이로그가 뒤를 이었다.

Q1-5 월별 카테고리 키워드 순위

먹방	28
도쿄올림픽	27
브이로그	21
유재석	20
요리	18

7월 먹방과 도쿄올림픽이 상위권을
차지했으며 브이로그, 유재석이 뒤를 이었다.

Q 1-5 월별 카테고리 키워드 순위

3 월	
박수홍	3
Korean	3
먹방	3
가족	2
뉴스	2

4 월	
먹방	34
예능	24
레전드	23
몰카	22
브레이브걸스	22

5 월	
먹방	38
예능	19
뉴스	19
아이돌	17
라이브	13

6 월	
먹방	45
예능	28
축구	21
브이로그	20
mukbang	18

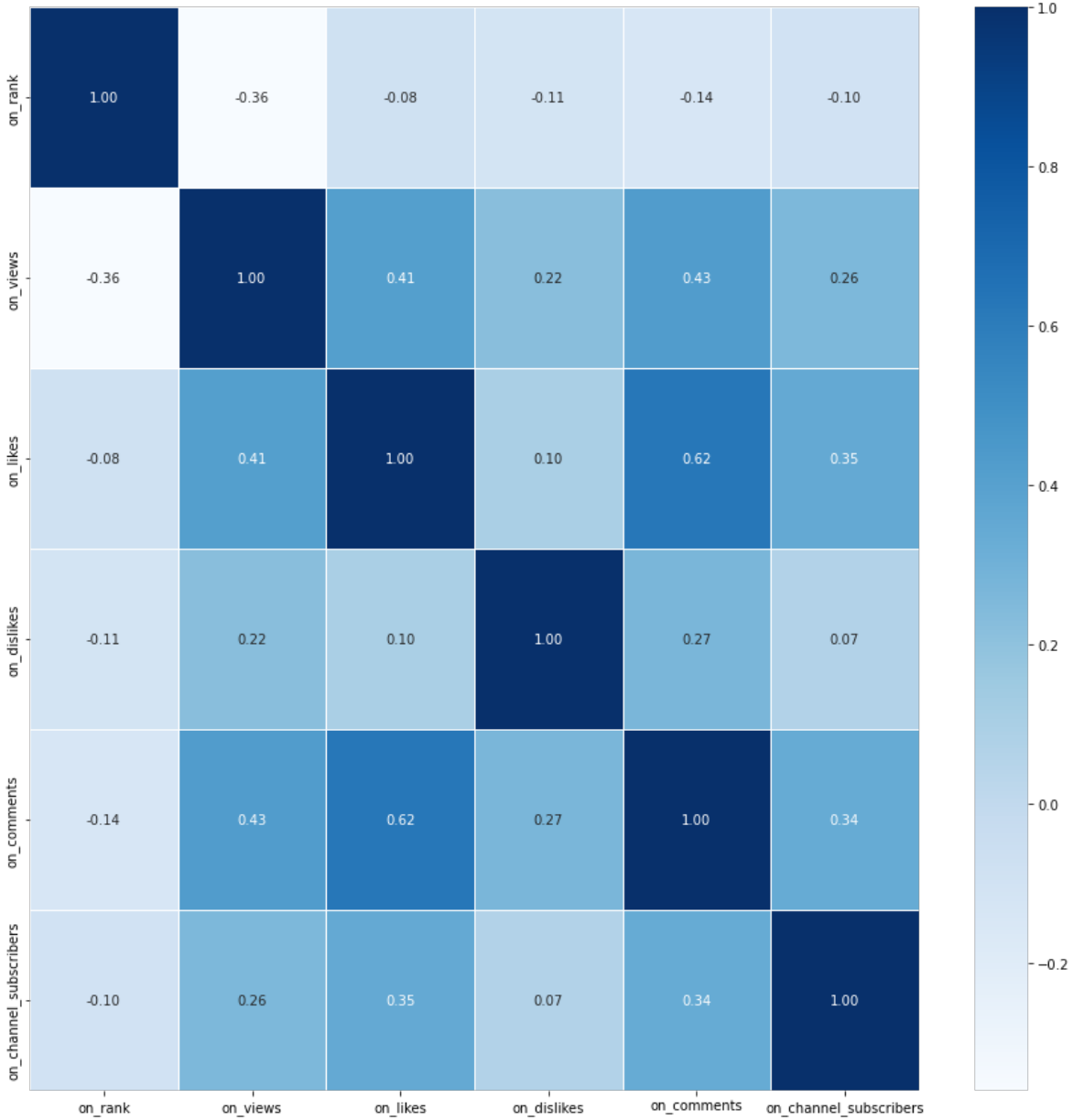
7 월	
먹방	28
도쿄올림픽	27
브이로그	21
유재석	20
요리	18

3월부터 7월까지 상위 tag 분석 결과 먹방 예능이 높은 순위를 보였으며
Entertain 카테고리 관련 카테고리가 높은 순위권인 것을 볼 수 있었다.

Q1-6 결론

1. entertainment, people & blog 카테고리 그리고 먹방, 예능 tag 소재 동영상 인기동영상이 되기 유리하다.
2. 매월마다 인기동영상은 600개 정도가 올라온다.
3. 인기동영상이 되기위해서는 최소 약53,000view가 필요하며 평균 약496,000 view로 인기동영상이 된다.
4. 같은 채널의 동영상은 최대5번까지 인기동영상이 될 수 있다.

Q2-0 engagement 구하기



대표 지표들에 대한 상관관계 조사를 해보았다.

대표 지표: on_trending_date, off_trending_date, on_rank on_views, on_likes,on_dislike, on_comments, on_channel_subscribers

on comment가 on likes와 0.62의 상관관계를 보이는 것으로 나타났다. 따라서 comment남기면 like를 누르는데 연관이 있다고 볼 수 있다.

하지만, 가장 중요한 views와 상관관계가 높은 지표를 찾을 수 없었다.

Q2-1 가설 제시 및 검증

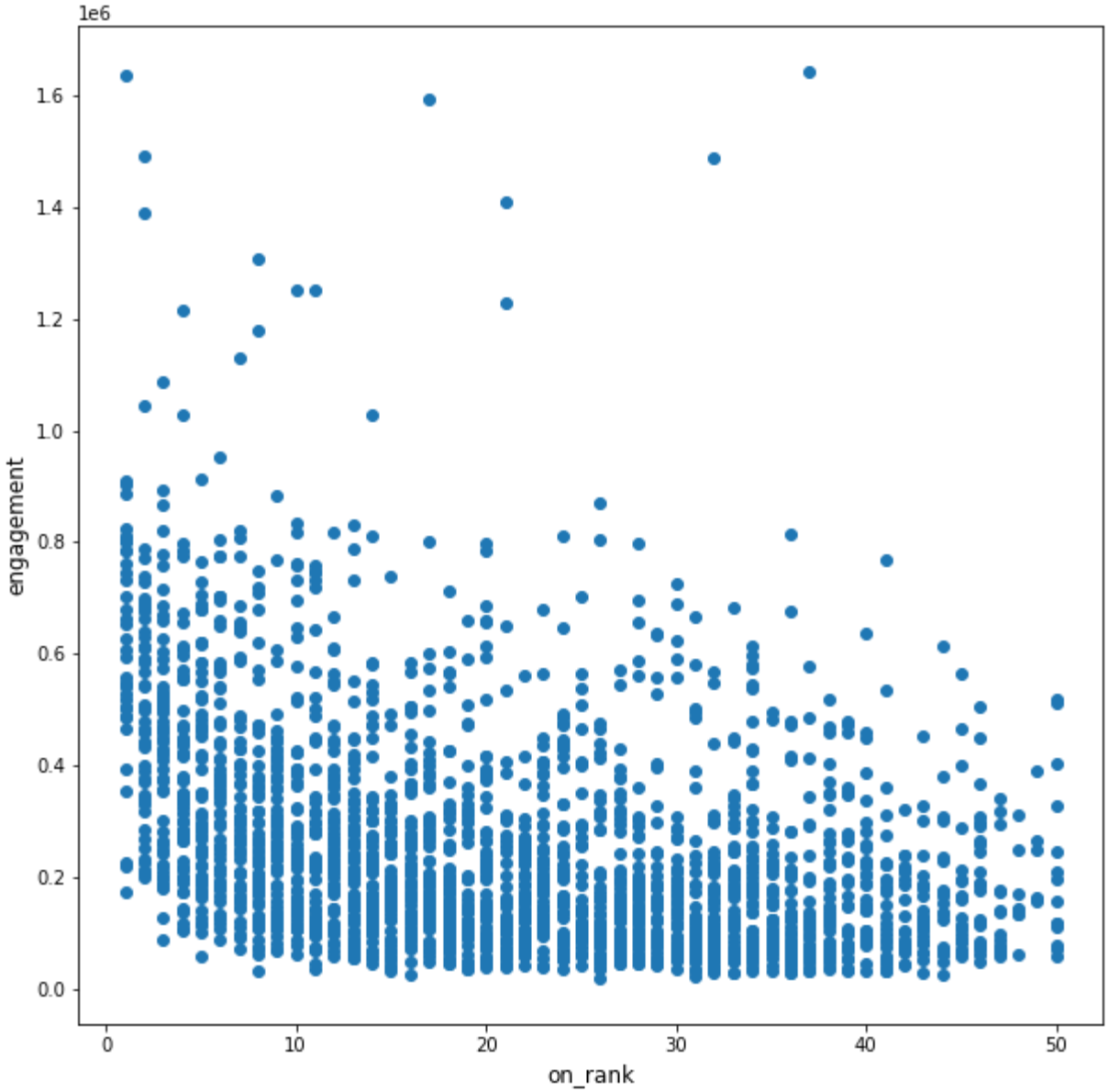
<Engagement 가설>

가설: on_views/(on_trending_date-published_date)는 rank와 상관관계가 있을 것이다.

on_view가 가장 큰 영향을 줄 것이라고 생각했으며
published_date와 on_trending_date시기와 관계를 주목해보고자 했다.

가설을 검증하기 published_on_date(on_trending_date-published_date)를 구해보았다.

Q2-1 가설 제시 및 검증

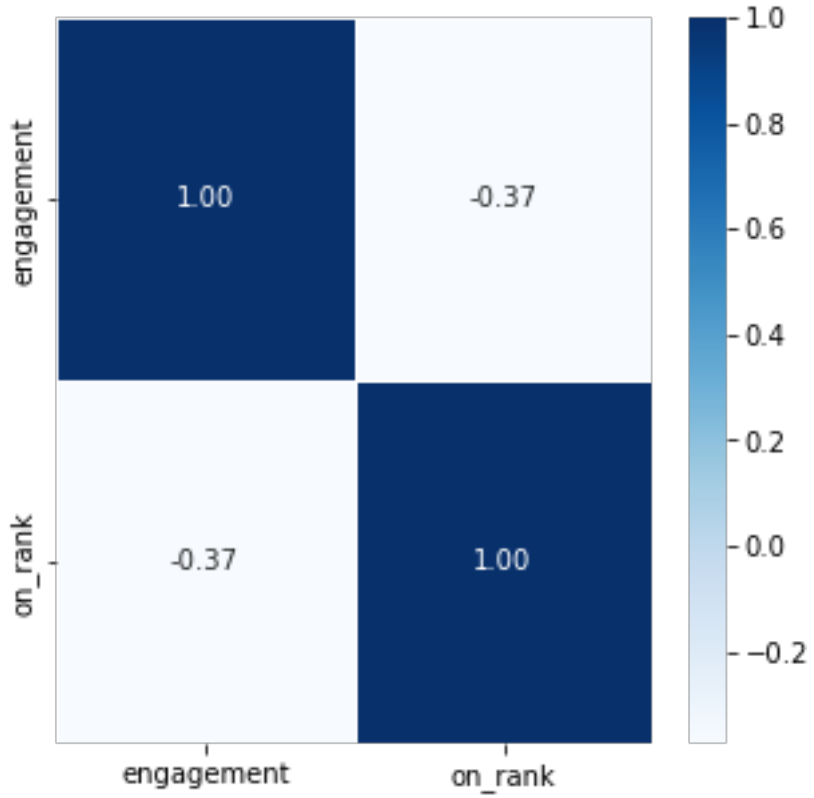



```
df['engagement'] = df['on_views']//df['published_on_date']
```

```
plt.figure(figsize=(10,10))
plt.scatter(list(df.on_rank.values),list(df.engagement.values))
plt.xlabel('on_rank', fontsize=12)
plt.ylabel('engagement', fontsize=12)
plt.show()
```

Engagement와 on_rank의 Scatter는
위와 같은 분포를 보인다.

Q2-1 가설 제시 및 검증



상관관계 검증 결과 -0.37로 음의 상관관계가 나타났다.

Views가 rank에 영향을 주는 것을 볼 수 있다.

Q2-1 가설 제시 및 검증

인기 급상승 동영상에서는 이 조건을 모두 고려하고자 노력합니다. 이를 위해 인기 급상승 동영상은 다음을 포함하여 다양한 신호를 고려합니다.

- 조회수
- 동영상 조회수 증가 속도(즉, '온도')
- YouTube 외부를 포함하여 조회수가 발생하는 소스
- 동영상 업로드 기간
- 해당 동영상을 같은 채널에 최근 업로드한 다른 동영상과 비교한 결과

채널의 영향을 고려한 engagement에 가중치를 한다.

따라서 on_channel_total_views를 통해 engagement애 가중치를 부여한다.

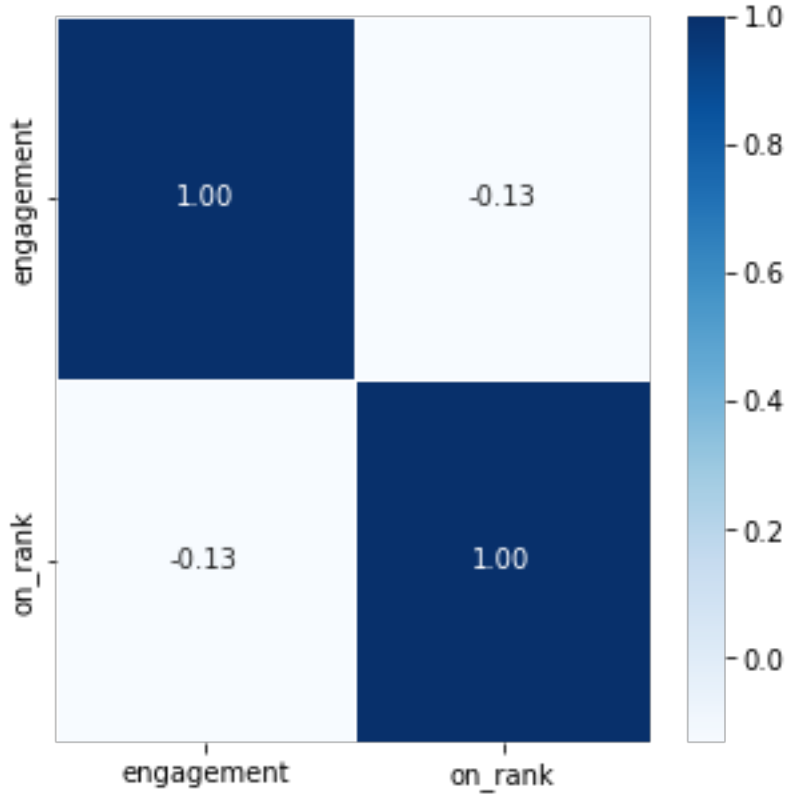
Q2-1 가설 제시 및 검증

```
df['standard']= (df['on_channel_total_views']-df['on_channel_total_views'].mean())/df['on_channel_total_views'].std
df['engagement'] = df['engagement']*df['standard']
```

기존 engagement 지표에 on_channel_total_views를
전체 데이터의 standardscale한 지표를 곱하기로 한다.

engagement는 engagement *standard(on_channel_total_views) 결과 값이다

Q2-1 가설 제시 및 검증



on_channel_total_views를 engagement에 반영한 상관관계 검증 결과 -0.13으로 오히려 낮아지는 결과를 보였다.

Q2-2 결론

결론

engagement의 가장 중요한 지표는 views이다.
on_channel_total_views에 대한 standardscale를
engagement에 반영하는 것은 좋지 못하다.