유튜브 메인 페이지 “탐색/인기” 데이터 분석 및 시각화

Q1. 데이터 타입별 시각화(자유양식)

- 전체기간 카테고리->채널->비디오 개수
- 월별 카테고리->채널->비디오 개수
- 월별 TOP10 채널 (분류 기준은 비디오 개수)
- 주별 TOP5 채널 (분류 기준은 비디오 개수)
- 월별 카테고리별 태그 키워드 순위

Q1-1. 전체기간 카테고리->채널->비디오 개수

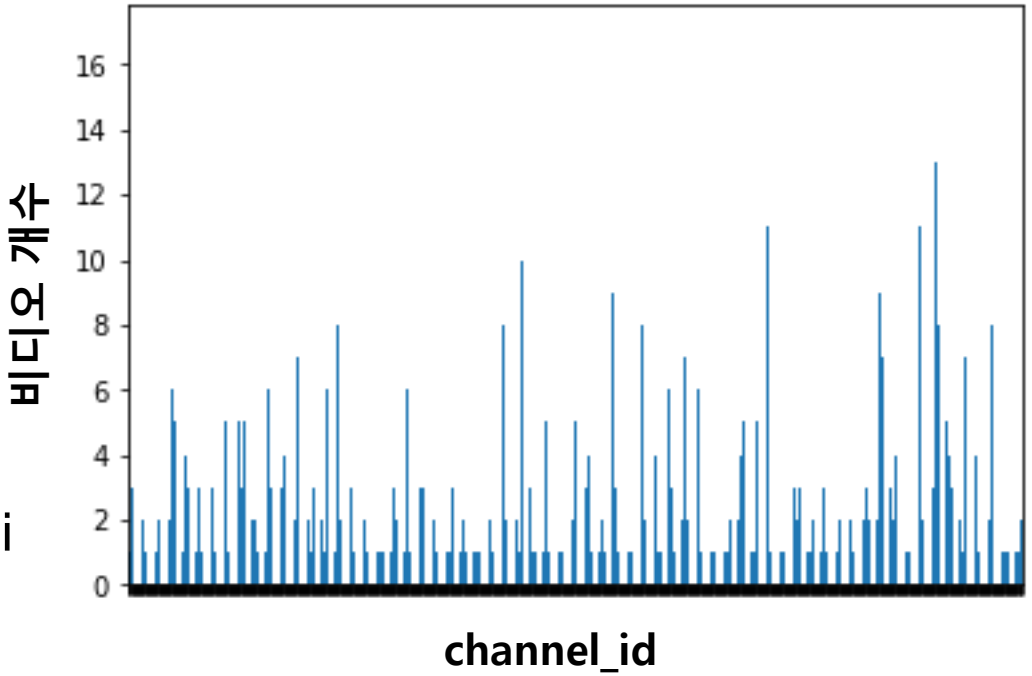
a) 제공된 자료의 전체 기간을 기준

b) "channel_id"를 기준으로 그룹화

c) 각 채널의 비디오 보유 현황

d) 비디오 개수는 1~17 사이의 값

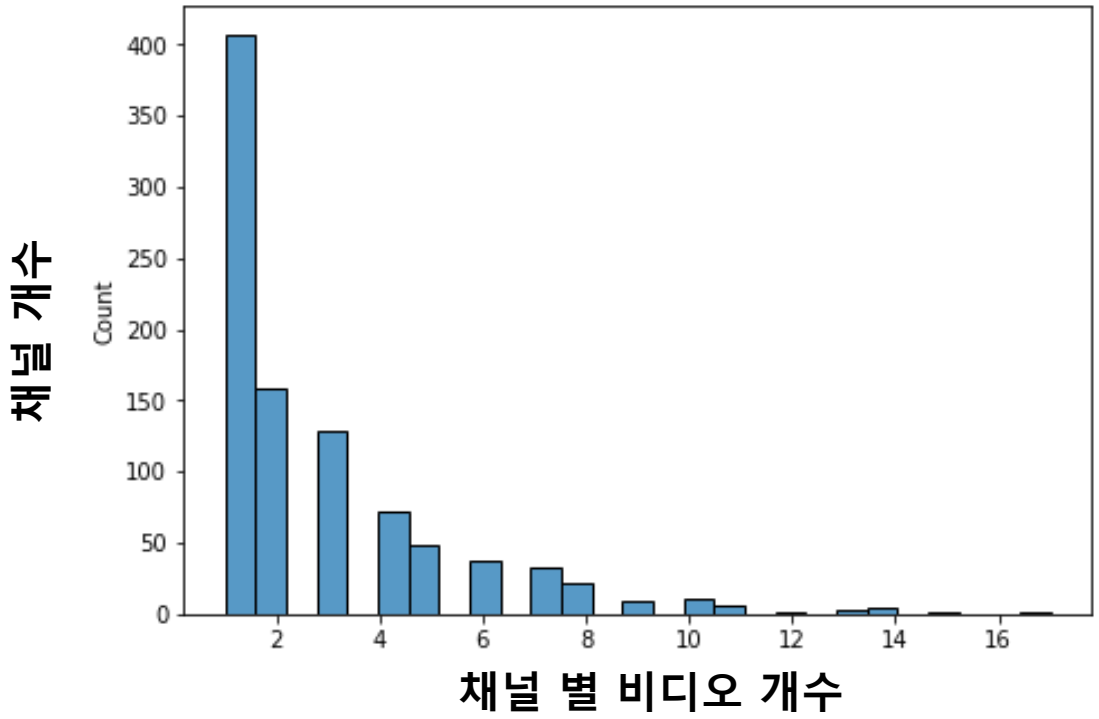
e) 인기동영상을 가장 많이 보유한 채널: CHQ2O-i

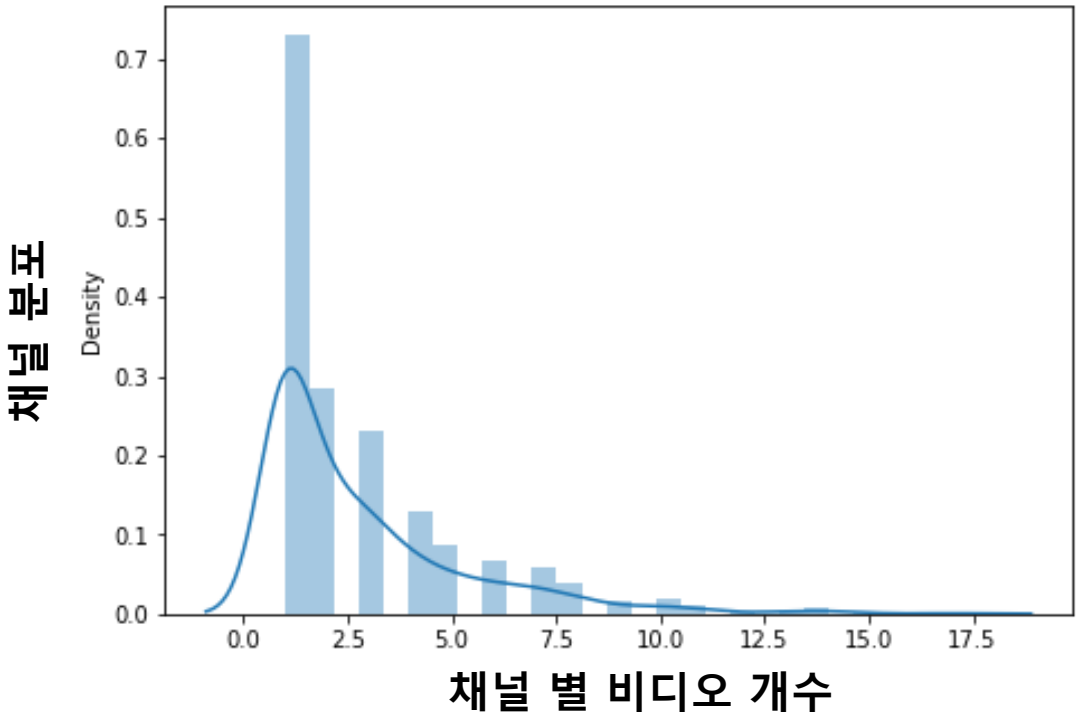


Q1-1. 전체기간 카테고리->채널->비디오 개수

a) 채널 별 비디오 개수의 분포를 파악하기 위한 그래프

b) 1~3를 보유한 채널이 가장 많음





Q1-2. 월별 카테고리->채널->비디오 개수

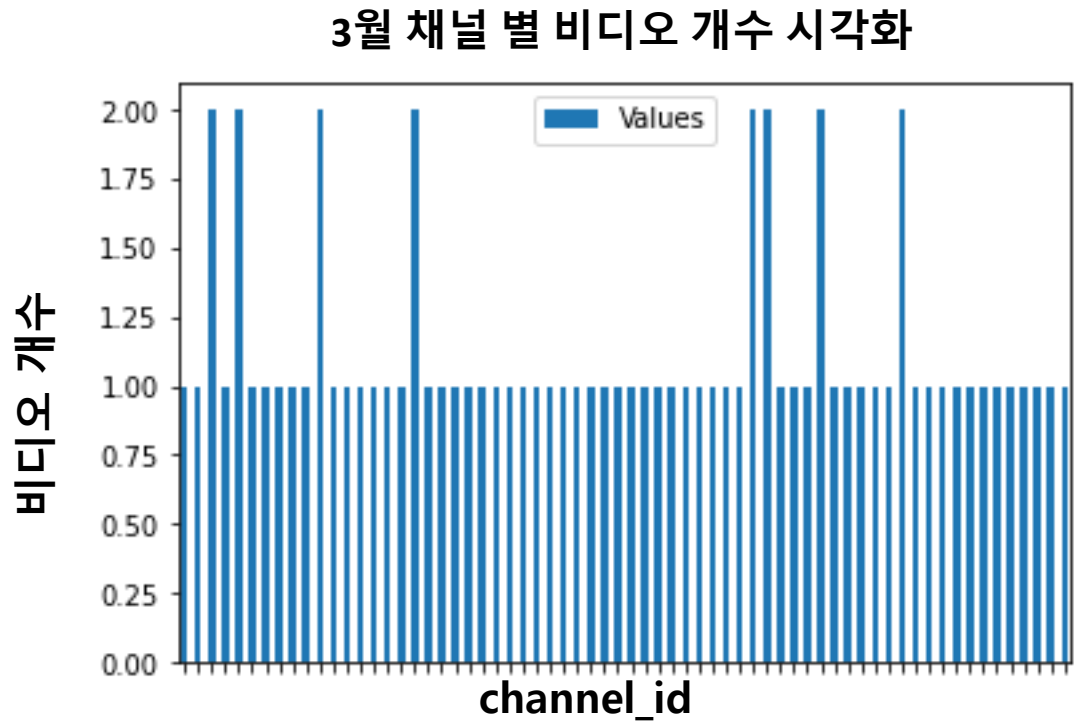
a) 비디오를 월 단위로 구별하기 위해 변수 생성

b) 기존 변수 "published_date " 를 사용하여 "published_month " 를 생성

c) 자료는 3월부터 7월까지의 유튜브 인기 동영상 자료

d) 월 단위로 채널 별 비디오 개수 파악

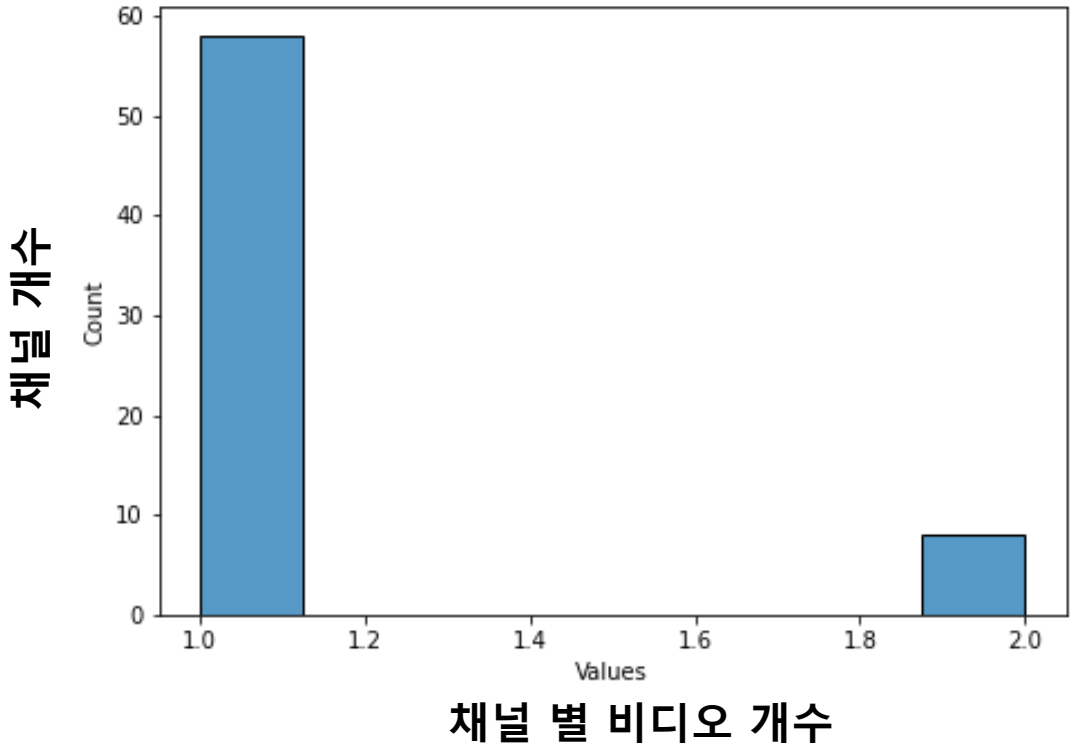
e) 오른쪽 그래프는 3월 채널 별 인기 동영상

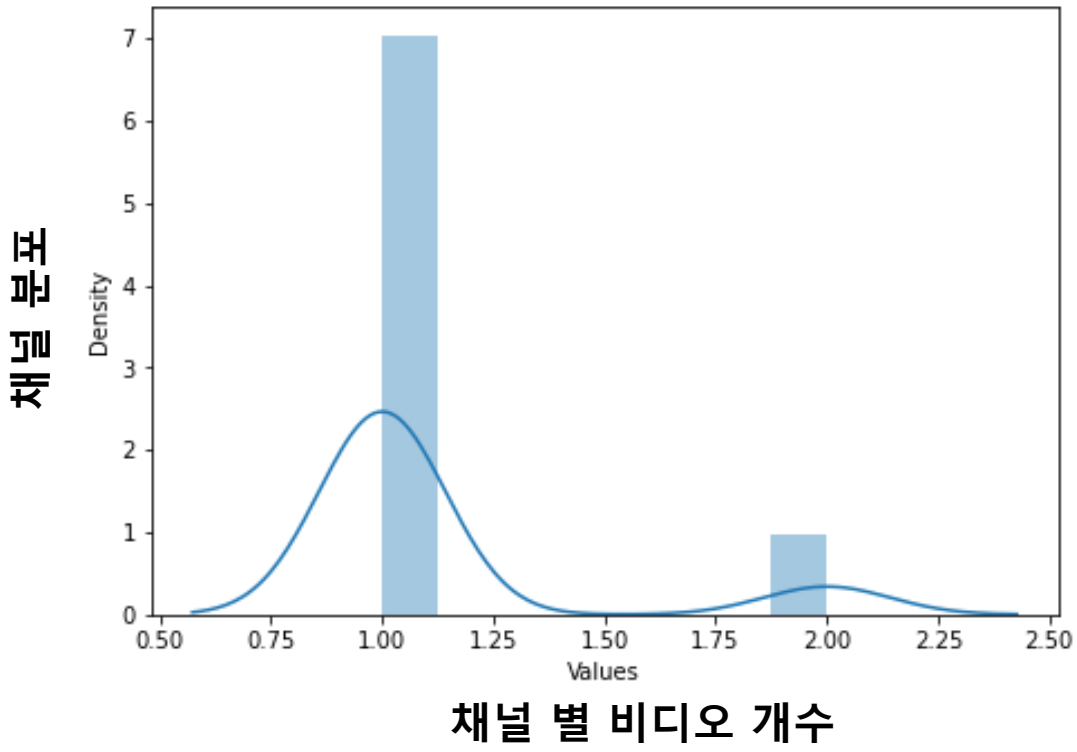


Q1-2. 월별 카테고리->채널->비디오 개수

a) 3월의 채널 별 인기 동영상 보유 개수 시각화

b) 3월 ~7월 까지 월 별로 시각화 가능





Q1-3. 월별 TOP10 채널 (분류 기준은 비디오 개수)

a) 월 단위 인기 동영상의 개수를 기준으로 상위 10개의 채널 선정

b) “published_month” 월 단위 정렬을 위한 변수를 사용

c) ” channel_id”를 기준으로 개수 파악

d) 채널 간 보유 동영상의 개수가 같을 경우: 조회수가 더 많은 동영상을 보유한 채널이 높은 순위에 배정

e) 정렬을 위하여 (“ published_month”, ” channel_id”, 비디오 개수, 'on_views ' 의 max값)을 가지는 튜플변수를 생성하여 리스트 할당

Q1-3. 월별 TOP10 채널 (분류 기준은 비디오 개수)

a) 3월을 기준으로 TOP10 채널 출력 결과

b) 변경되는 기준에 따른 재정렬이 가능한 변수

c) 월 단위로 상위 채널을 출력이 가능하도록 코드 작성

d) 이후 다른 분석 자료와 함께 사용이 가능하도록 변수에 저장

```
[('2021-03', 'CH920m3', 2, 1409041),
 ('2021-03', 'CHicKQU', 2, 1262201),
 ('2021-03', 'CH5BMQO', 2, 1176510),
 ('2021-03', 'CHG9aFJ', 2, 932967),
 ('2021-03', 'CHaKod3', 2, 852156),
 ('2021-03', 'CHaZS_X', 2, 466251),
 ('2021-03', 'CH46BbE', 2, 226941),
 ('2021-03', 'CHnet0I', 2, 152606),
 ('2021-03', 'CHEf_Bc', 1, 9205421),
 ('2021-03', 'CHweOkP', 1, 8553414)]
```

3월 TOP10 채널 출력

Q1-4. 주별 TOP5 채널 (분류 기준은 비디오 개수)

a) 주 단위 인기 동영상의 개수를 기준으로 상위 5개의 채널 선정

b) 주 단위 구분을 위하여 “published_week”의 변수 생성

c) 동영상 업로드 날짜를 기준으로 주 단위 구분을 하기 위하여 datetime의 strptime, isocalendar 함수 사용

d) 채널 간 보유 동영상의 개수가 같을 경우: 조회수가 더 많은 동영상을 보유한 채널이 높은 순위에 배정

e) 정렬을 위하여 (“ published_week”, ” channel_id”, 비디오 개수, 'on_views ' 의 max값)을 가지는 튜플 변수를 생성하여 리스트 할당

Q1-3. 월별 TOP10 채널 (분류 기준은 비디오 개수)

a) 3월을 첫주 기준으로 TOP5 채널 출력 결과

b) 변경되는 기준에 따른 재정렬이 가능한 변수

```
[(12, 'CHnet0I', 2, 152606),
 (12, 'CHweOkP', 1, 8553414),
 (12, 'CHicKQU', 1, 1262201),
 (12, 'CH7Krez', 1, 1121118),
 (12, 'CH2qVOO', 1, 1101249)]
```

3월 1주차 TOP5 채널 출력

Q1-5. 월별 카테고리별 태그 키워드 순위

a) 두 가지 기준을 사용하여 데이터를 분류 및 정렬

b) 월 단위 / 카테고리 단위

   a) 월 단위로 카테고리에 따른 상위 키워드 추출

   b) 월 단위의 전체 키워드 순위 추출: 카테고리 별 키워드 추출의 경우 분야에 따라 상위 키워드를 추출하기 때문에, 해당 월의 전체적인 인기 키워드를 파악하기 위하여 추출

Q1-5. 월별 카테고리별 태그 키워드 순위

a) 월 단위로 카테고리에 따른 상위 키워드 추출

b) 예시: “News & Politics”의 6월의 상위 20개의 키워드와 빈도수

c) 카테고리와 월을 변경하여 다른 결과도 추출 가능

d) 카테고리 혹은 특정 월에 따라서 분석이 가능한 자료 형태로 출력

```
[('뉴스', 6),
 ('뉴스투데이', 3),
 ('뉴스데스크', 3),
 ('광주', 3),
 ('경찰', 3),
 ('news', 3),
 ('News Network', 3),
 ('MBC뉴스', 3),
 ('투자', 2),
 ('철거', 2),
 ('주식', 2),
 ('제보영상', 2),
 ('제보', 2),
 ('정오뉴스', 2),
 ('재테크', 2),
 ('장갑차', 2),
 ('인터뷰', 2),
 ('유상철', 2),
 ('에스비에스 뉴스', 2),
 ('에스비에스', 2)]
```

“News & Politics”의 6월의 상위 20개의 키워드

Q1-5. 월별 카테고리별 태그 키워드 순위

a) 월 단위의 전체 키워드 순위 추출

b) 예시: 3월의 상위 20개의 키워드와 빈도수

c) 월에 따라서 추출 가능

```
[('박수홍', 5),
 ('몰카', 5),
 ('유재석', 4),
 ('웃긴영상', 4),
 ('먹방', 4),
 ('고양이', 4),
 ('idol', 4),
 ('eng', 4),
 ('KBS', 4),
 ('횡령', 3),
 ('특전사', 3),
 ('특수부대', 3),
 ('참호격투', 3),
 ('예능', 3),
 ('아이유', 3),
 ('아이돌', 3),
 ('밀리터리', 3),
 ('런닝맨', 3),
 ('뉴스', 3),
 ('꿀잼', 3)]
```

3월 상위 20개의 키워드

Q2. 인기 동영상 분류를 위한 새로운 지표 개발

- 각각의 비디오는 시청자의 호응도(engagement)를 판단할 수 있는 객관적인 지표 들이 있음
  - views, likes, dislikes, comments, ...
- 비디오를 인기 동영상 기준에 부합하도록 분류할 수 있는 새로운 지표를 개발하고
- 이 지표를 사용하여 engagement 와 어떤 상관관계가 있는지 설명

Q2. 인기 동영상 분류를 위한 새로운 지표 개발

- "published_date " 와 “on_trending_date”을 사용하여 인기 동영상을 분류하는 새로운 지표 생성
- **새로운 지표: 동영상 업로드 이후의 기간**
- **인기 동영상이 되기까지는 기간을 파악**
- 총 2644개의 인기 동영상 중 2390개의 동영상이 **이틀** 안에 인기 동영상으로 선정
- **90% 이상의 동영상**이 이틀 안에 인기 동영상으로 선정
- **이틀을 초과**하여 선정되는 동영상은 역주행 혹은 성지순례 같은 **외적인 요인으로 인기 동영상 선정**
- 분석을 위하여 제공 된 자료는 모두 인기 동영상에 선정된 동영상

Q2. 인기 동영상 분류를 위한 새로운 지표 개발

- 자료의 90% 이상의 자료들이 이틀 안에 선정되었으므로 동영상 업로드 후 이틀을 인기 동영상 판별 기준으로 사용이 가능함
- 이틀 안에 인기 동영상이 된 자료들의 기존 지표를 활용 가능
- 혹은 이틀 안에 인기동영상으로 선정된 동영상의 데이터를 사용하여 예측 모델을 만드는 연구의 가능성을 고려

Q2. 인기 동영상 분류를 위한 새로운 지표 개발

- 그러나 인기에 관련된 대부분의 지표(on_views, on_likes, on_dislikes, on_comments)들의 평균 혹은 기준을 설정하기 어려움
- 조회수를 기준으로 boxplot을 그린 결과를 보았을 때,

다수의 이상치가 존재함

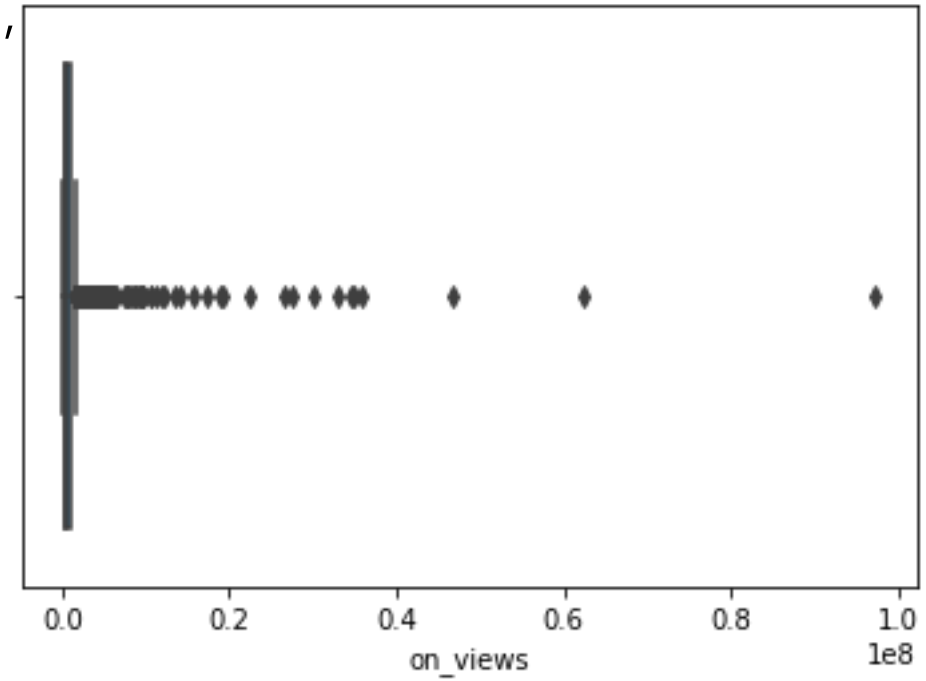
- 채널 혹은 동영상에 따라서 **편차**가 큼

따라서 인기 동영상의 분류 모델을 만들기 위해서는

데이터의 후처리가 필요

- 평균점을 찾기 보다는 인기 동영상의 되기 위한

최소한의 기준을 만드는 것이 실현 가능해 보임



조회수를 기준으로 boxplot을 그린 결과

Thank you