



# 더핑크퐁컴퍼니 유튜브 데이터분석

박건웅

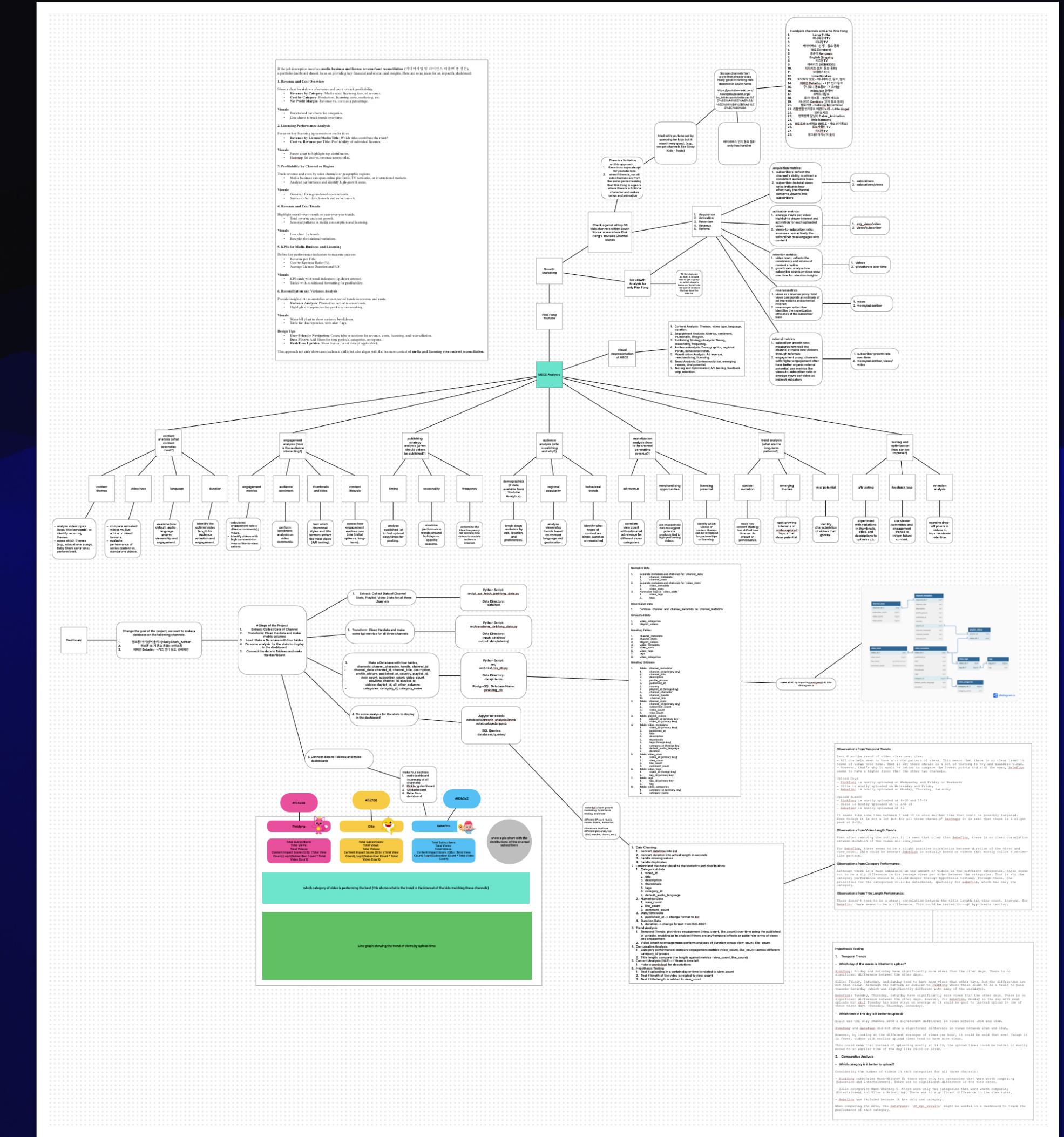
# 목차:

- 01 배경 및 문제 정의
- 02 데이터 수집
- 03 데이터 관리
- 04 탐색적 데이터분석
- 05 통계적 검정
- 06 대시보드 구축
- 07 논의

# 01. 배경 및 문제 정의

## 프로젝트 설계 및 브레인스토밍

- 더핑크퐁컴퍼니 아이덴티티: 핑크퐁, 아기상어 올리, 베베플
- 위의 대표적인 캐릭터들을 토대로 IP 컨텐츠를 조사한 결과 키즈 유튜브 채널이 있으며 상위권을 차지하고 있다.
- 유튜브 채널의 데이터를 수집하여 그로스해킹 분석 설계를 진행했다.
- 세 채널 모두 상위권을 차지하고 하고 있기에 유저 리텐션에 집중해서 분석을 진행했다.



프리폼으로 작성: [프로젝트 설계도 링크](#)

# 02: 데이터 수집

ETL: Extract, Transform, Load

- Extract:

- Youtube API를 사용해 수집한 데이터:
  - 채널 아이디
  - 채널 메타데이터 및 통계량
  - 채널 플레이리스트 아이디
  - 영상 메타데이터 및 통계량
  - 영상 카테고리

- Transform:

- 데이터베이스 설계에 맞게 전처리
- 채널과 영상 데이터의 정규화, 메타데이터와 통계량 분리
- 채널 정보와 채널 메타데이터 병합
- 영상 태그 (SEO 관련) 정규화

- Load:

- 전처리 단계에서 전처리한 테이블 PostgreSQL 데이터베이스에 저장:
  - 채널 관련: channel\_metadata, channel\_stats
  - 영상 관련: video\_metadata, video\_stats, video\_tags
  - 연결 테이블: playlist\_videos, video\_categories

모든 코드는 [깃헙 리포지토리](#)에서 확인 가능.

Extract - `src/extract_pinkfong_data.py`

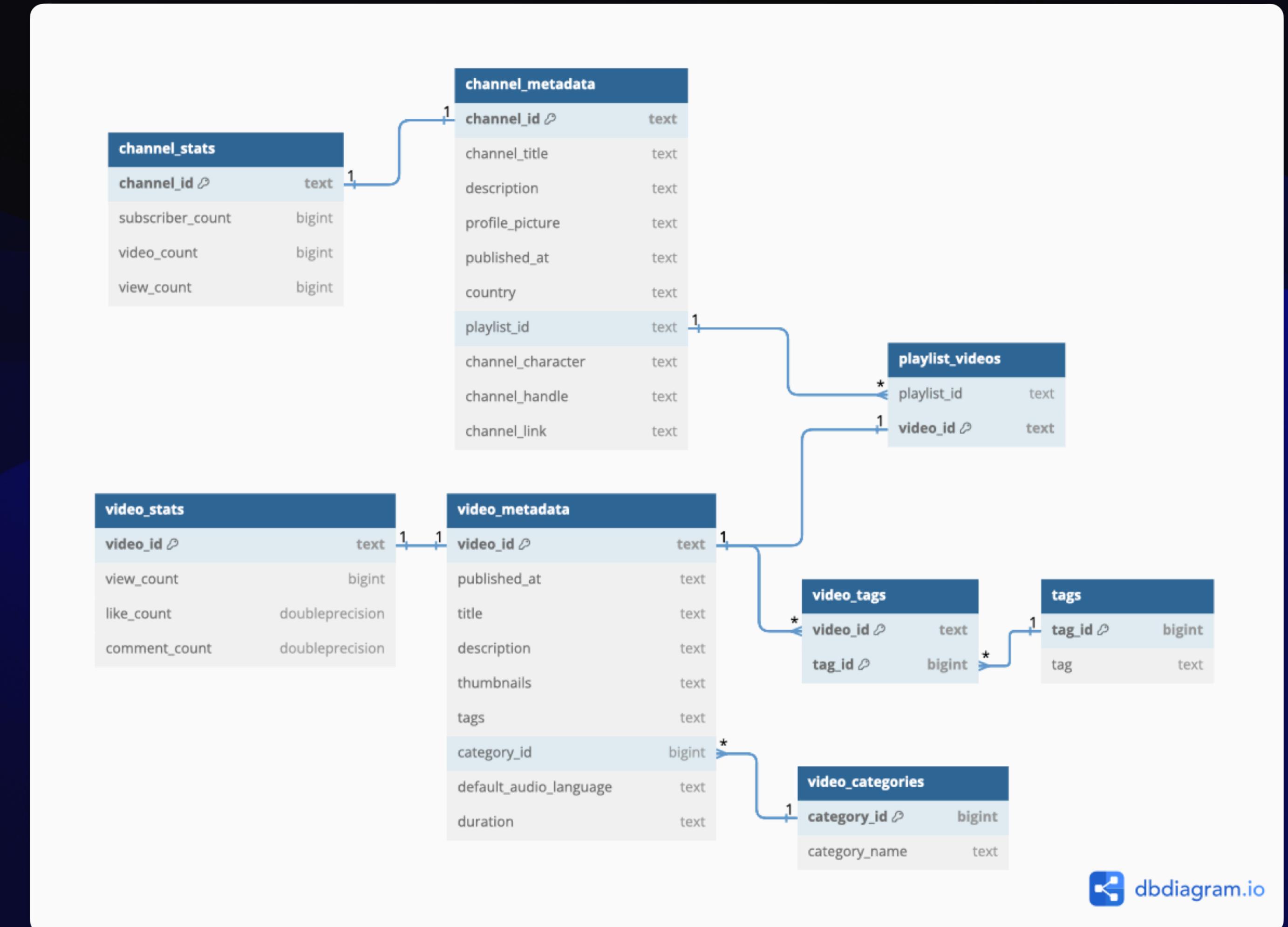
Transform - `src/transform_pinkfong_data.py`

Load - `src/load_pinkfong_data.py`

# 03: 데이터 관리

## 데이터베이스 설계

- 우측 표와 같이 데이터베이스를 설계했다.
- 총 8개의 테이블이 있으며 **분석에서 사용된 테이블은:**
  - 채널관련:
    - channel\_metadata**
    - channel\_stats**
  - 영상관련:
    - video\_metadata**,
    - video\_stats**,
    - video\_categories**



# 04-1: 탐색적 데이터 분석

## 데이터 소개

**데이터셋:** 핑크퐁, 아기상어 올리, 베베핀 유튜브 데이터

**출처:** YouTube API

**설명:** 이 데이터셋은 YouTube 비디오의 메타데이터를 포함하며, 조회수, 좋아요 수, 댓글 수, 비디오 설명 및 게시 정보와 같은 세부 사항을 제공합니다. 이 데이터셋은 트렌드 분석, 콘텐츠 성과 평가, 그리고 잠재적으로 콘텐츠 추천 엔진 구축과 같은 다양한 분석에 사용될 수 있습니다.

**크기:** 5893 행 × 8 열

**타겟 변수:** view\_count.

	변수 이름	변수 설명	예시	타입
1	channel_character	채널 캐릭터 명	Pinkfong	str
2	published_at	영상 업로드 일시	2024-12-27T06:24:03Z	datetime
3	duration	영상 길이	PT2H8M53S	float
4	category_name	카테고리 명	Education	str
5	view_count	영상 조회 수	9152	int

**시간 범위:** 데이터는 2014-03-12부터 2024-12-27까지 포함됩니다.

# 04-2: 탐색적 데이터 분석

## 데이터 전처리

### • 데이터타입 변환

- 영상 업로드 컬럼 (published\_at):
  - T와 Z 같은 문자열을 정규표현식을 통해 제거
  - UTC에서 KST(한국 시간)으로 변환
  - 문자열에서 날짜 데이터타입 변환
- 영상 길이 (duration):
  - PTMS와 같은 문자열을 정규표현식을 통해 제거
  - PT는 날 수 MS는 시간을 뜻함
  - 문자열에서 float으로 데이터타입 변환

### • 결측치 처리

- 영상 조회 수 (view\_count):
  - 결측값인 경우 0으로 대체
- 영상 길이 (duration):
  - 0인 경우 행을 삭제

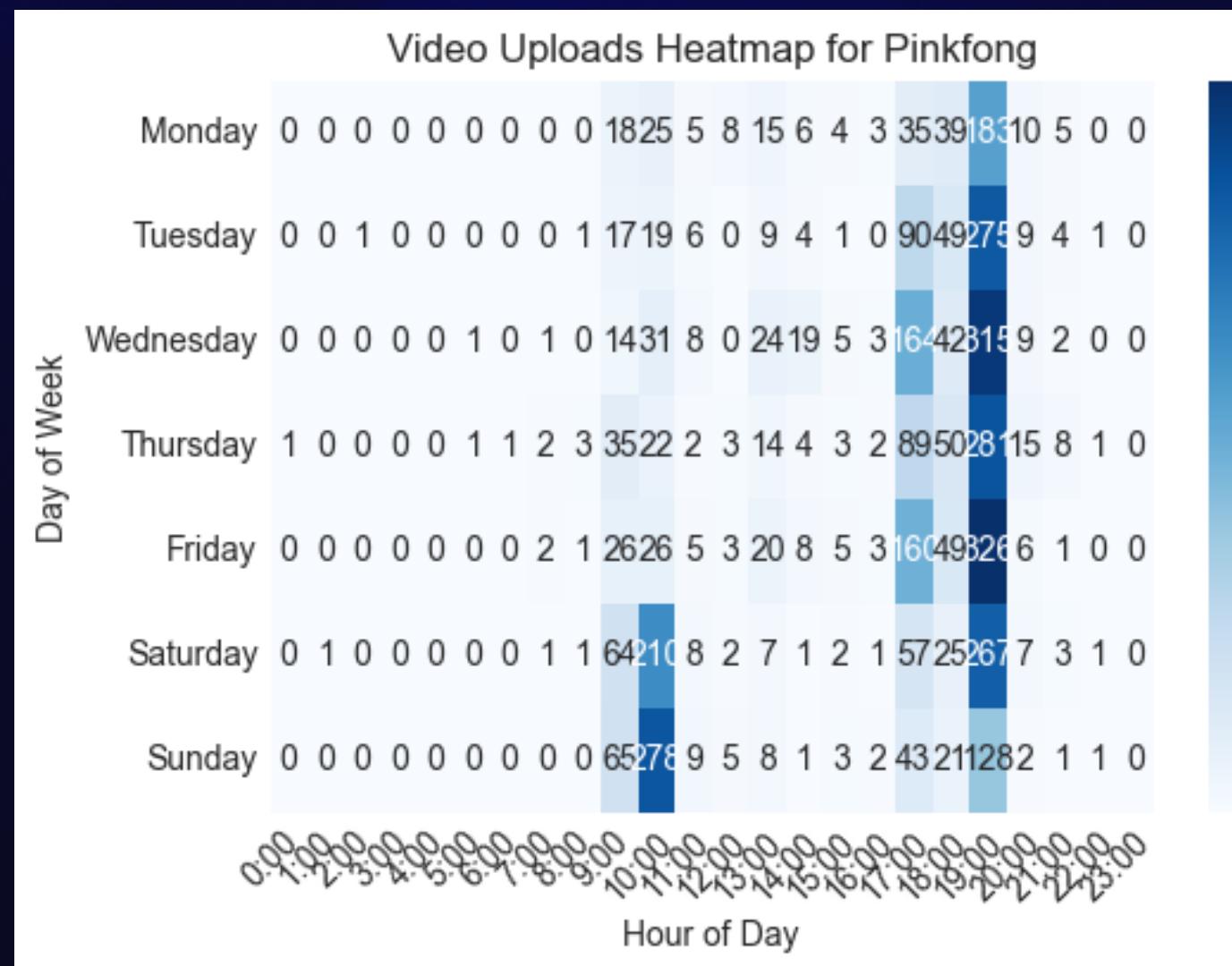
### • 이상치 처리

- 영상 조회 수 (view\_count):
  - 데이터 분석을 진행할 시 여러 카테고리와 시간대의 영상 조회 수를 비교하는데 바이럴한 영상의 영향을 방지하기 위해 IQR을 사용한 이상치 제거.

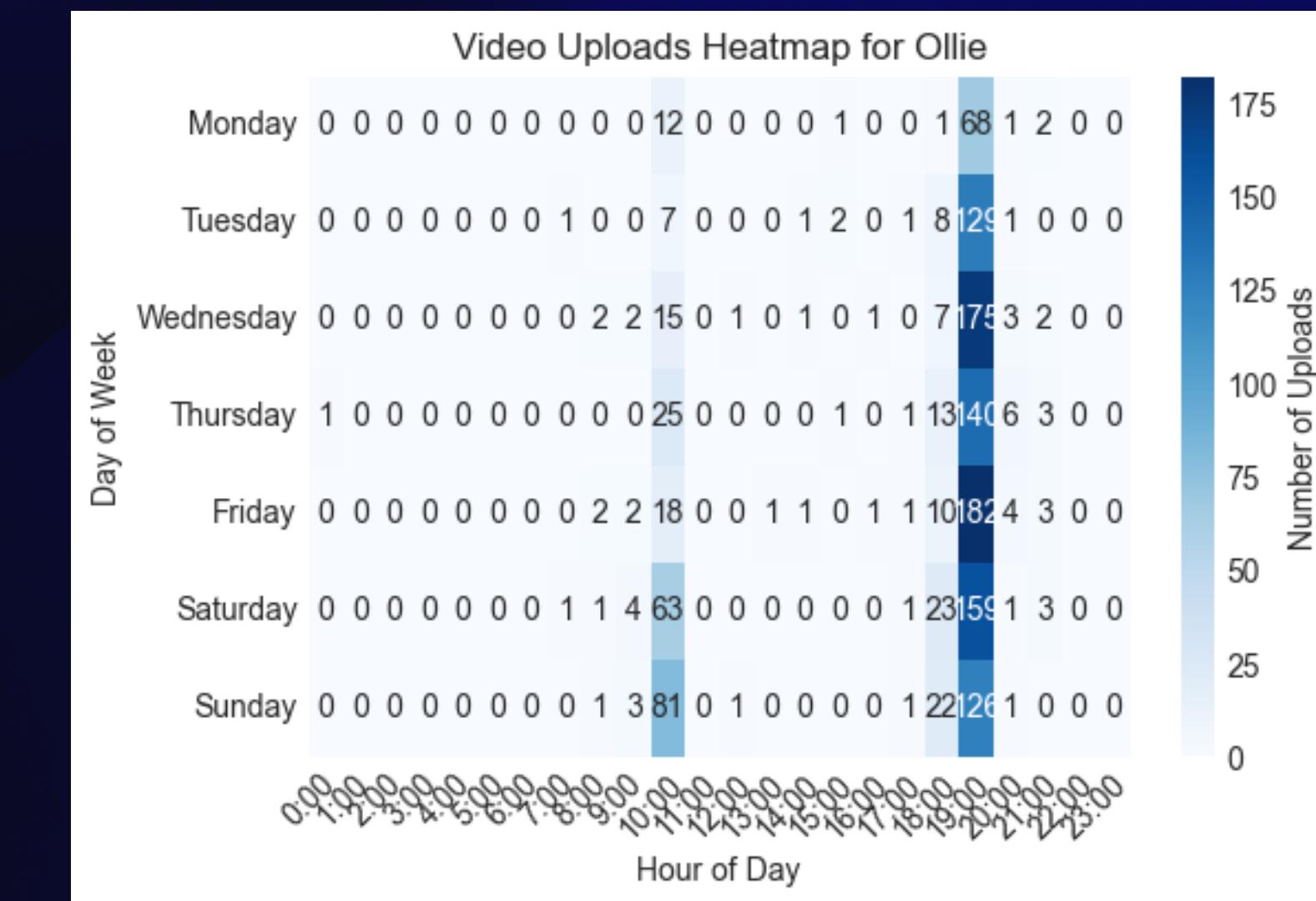
# 04-3: 탐색적 데이터 분석

## 데이터 인사이트: 업로드 시간

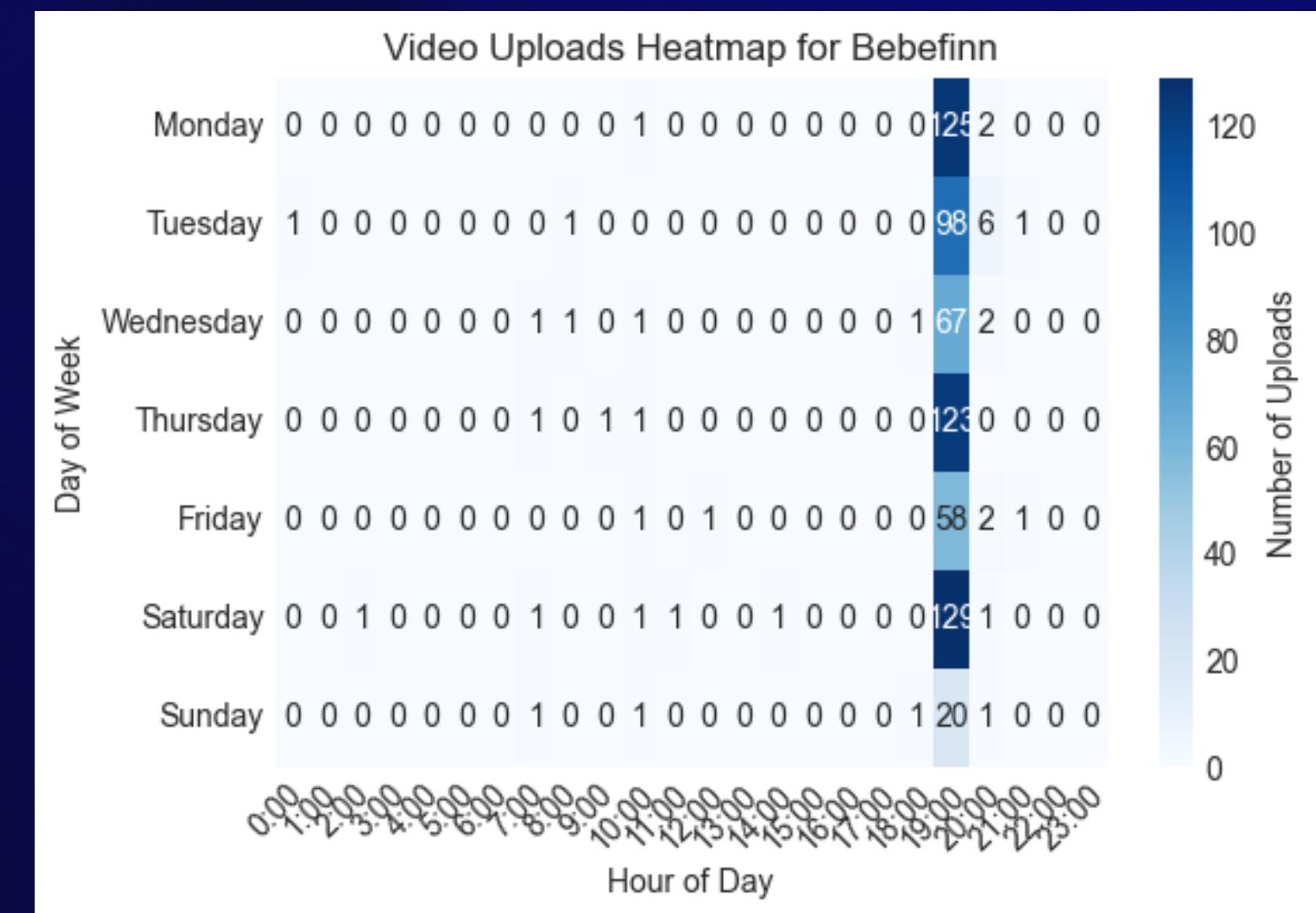
- 핑크퐁, 아기상어 올리, 베베플 세 채널 모두:
- 주말 보다는 평일에 업로드 하는 경향
- 시간대는 19시에 업로드 하는 패턴이 강하게 드러났다.



핑크퐁 채널 업로드 주기



아기상어 올리 채널 업로드 주기

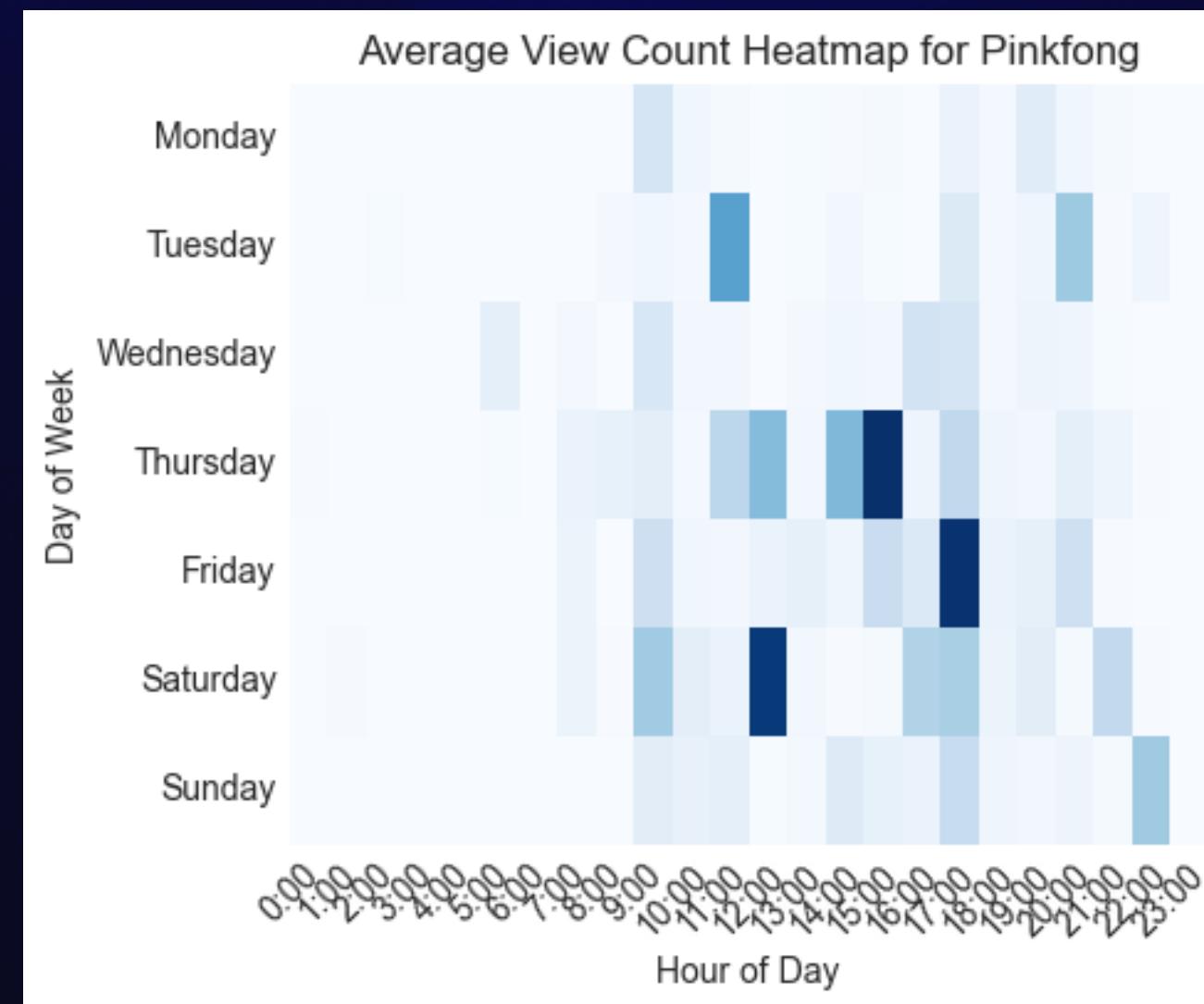


베베플 채널 업로드 주기

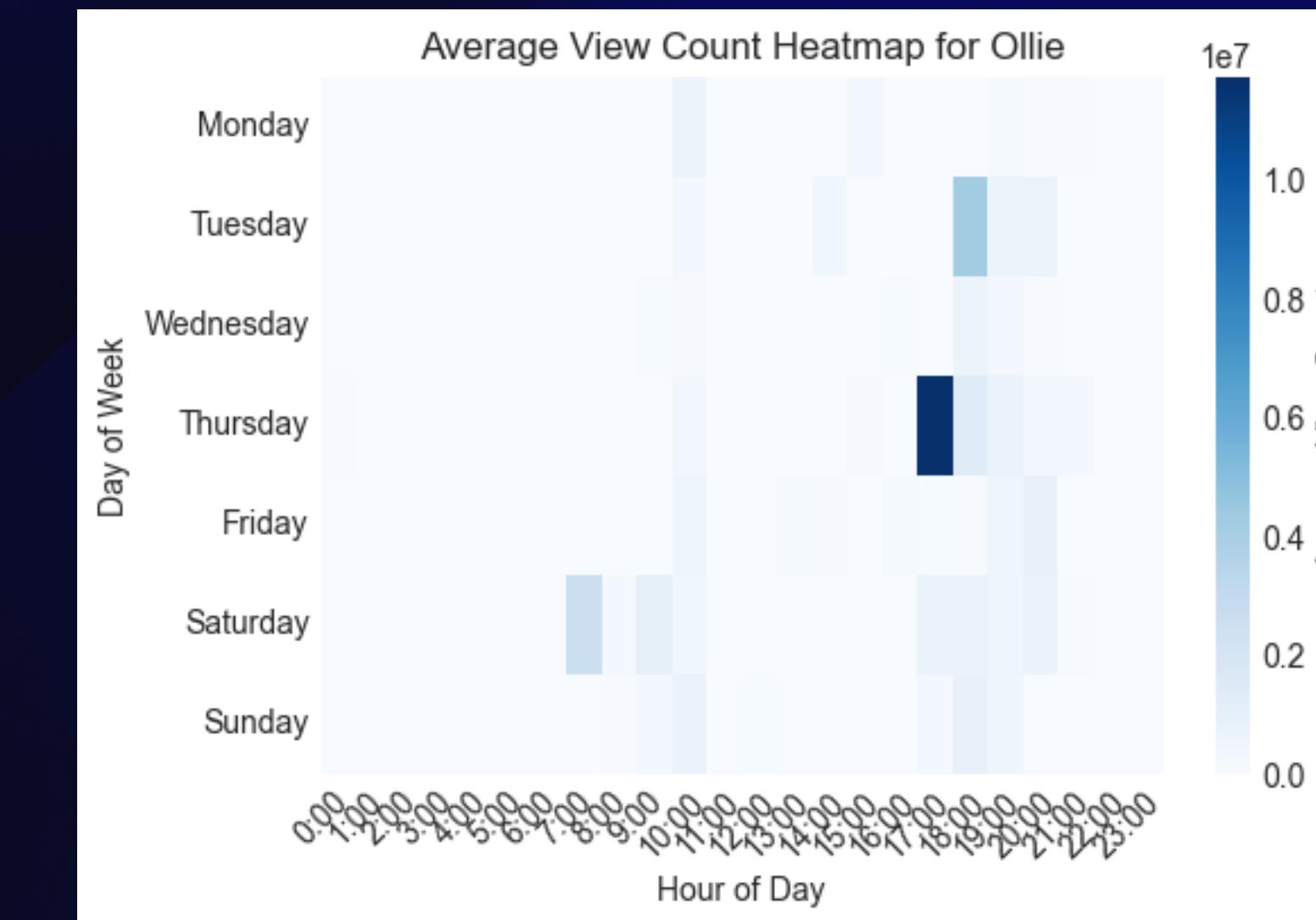
# 04-4: 탐색적 데이터 분석

## 데이터 인사이트: 업로드 주기의 평균 조회수

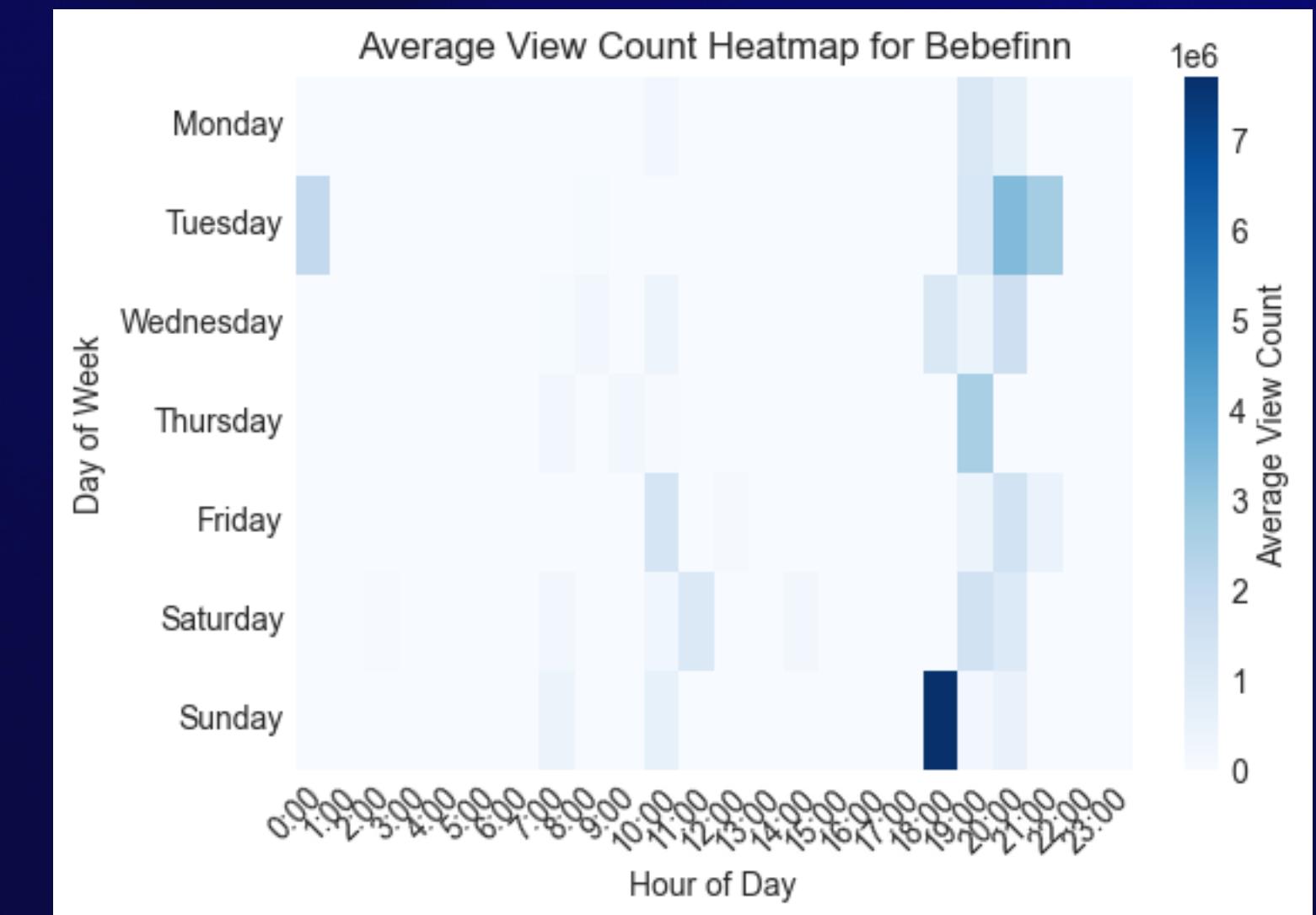
- 핑크퐁, 아기상어 올리, 베베플 세 채널 모두:
- 19시에 업로드하는 경향의 반면에 이른 시간에 업로드 한 영상들이 오히려 높은 조회 수를 얻는 현상을 발견
- 연하지만 세 채널 모두 10시에 조회수가 높게 나오는 것을 발견



핑크퐁 채널 평균 조회 수



아기상어 올리 채널 평균 조회 수

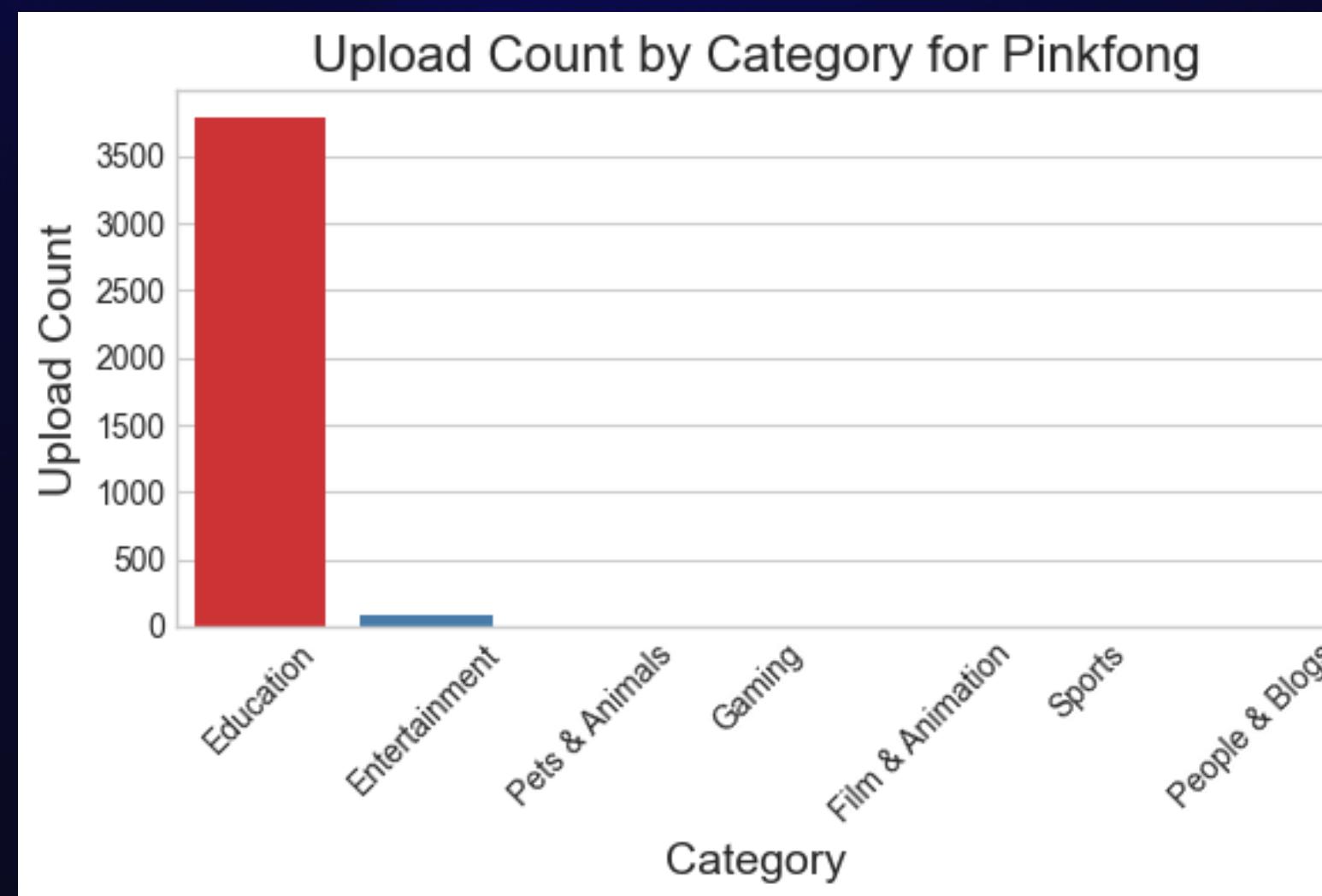


베베플 채널 평균 조회 수

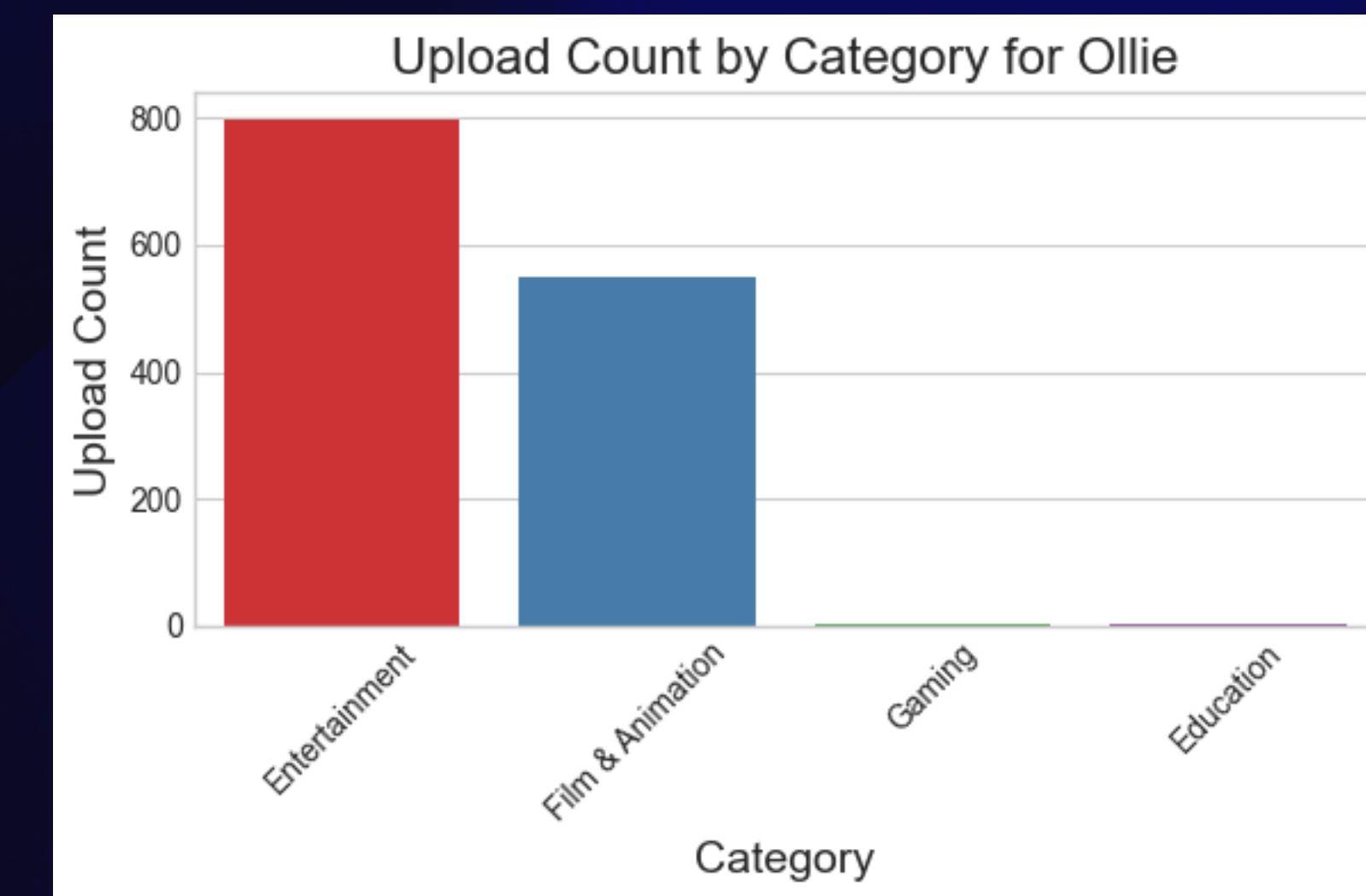
# 04-5: 탐색적 데이터 분석

## 데이터 인사이트: 업로드 카테고리

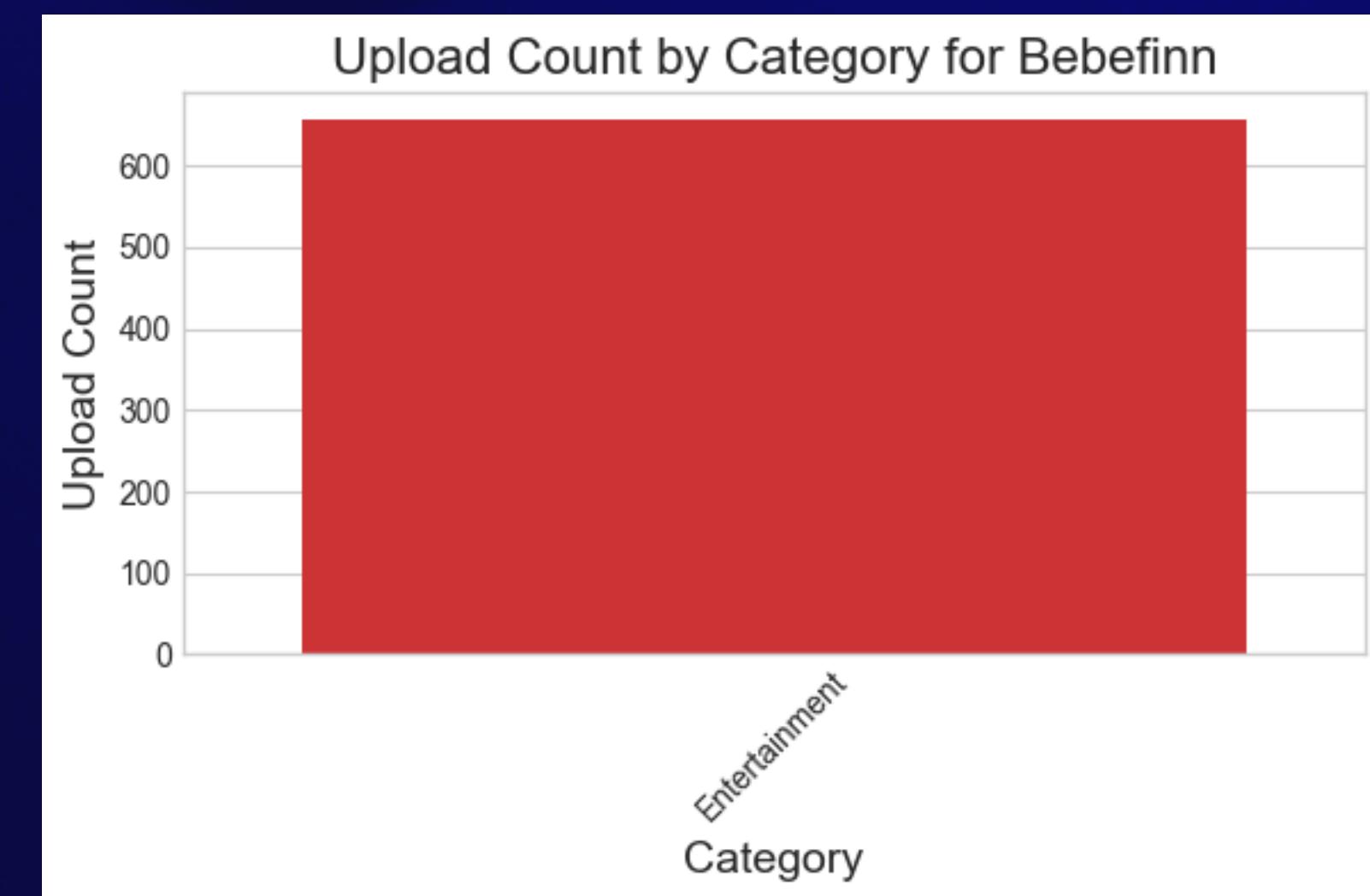
- 핑크퐁 업로드 카테고리는 다양하지만 Education으로 매우 치우쳐져 있다.
- 아기상어 올리 업로드 카테고리는 Education과 Film & Animation을 주로 업로드 한다.
- 베베플은 하나의 카테고리 밖에 없다.



핑크퐁 채널 평균 조회 수



아기상어 올리 채널 평균 조회 수

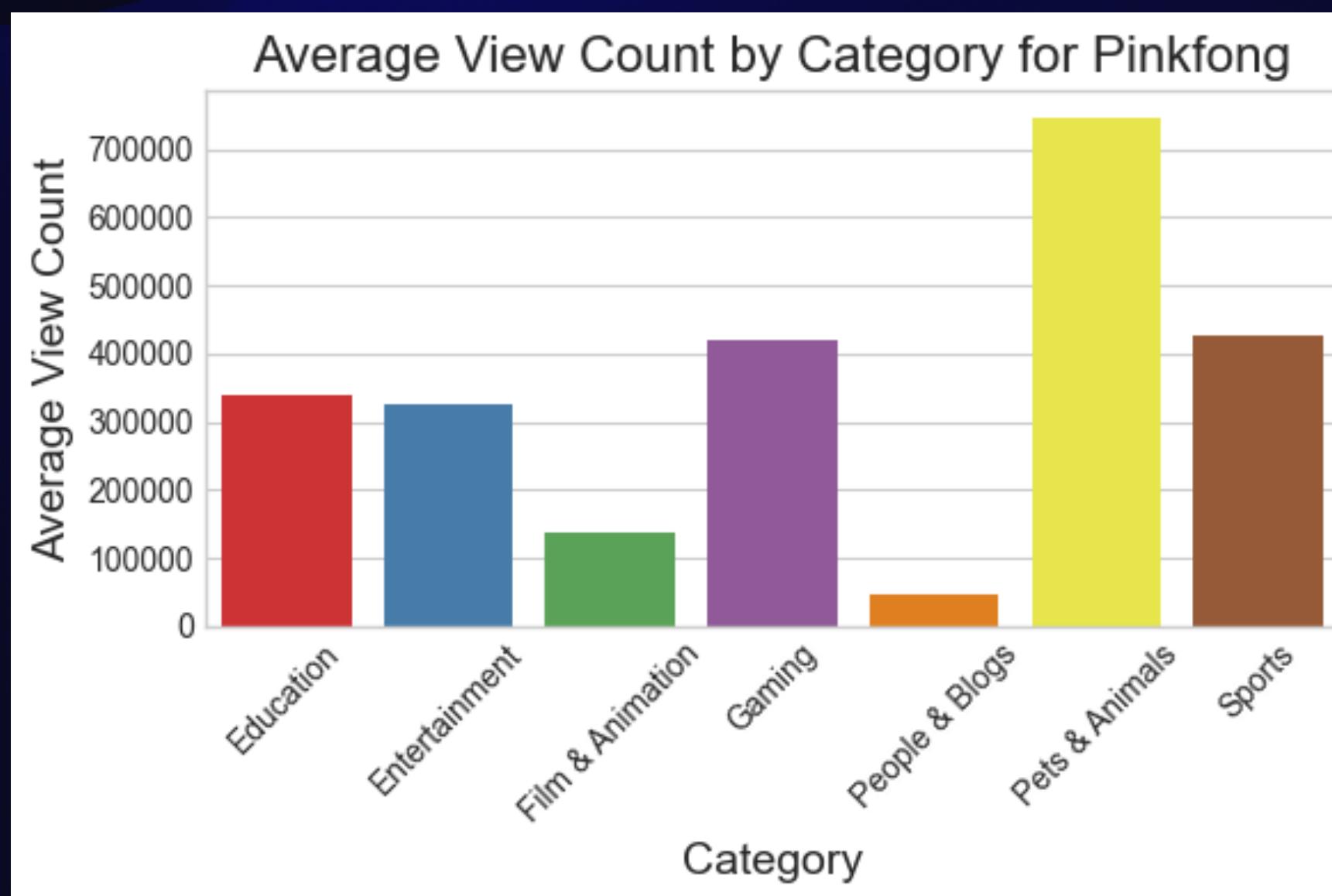


베베플 채널 평균 조회 수

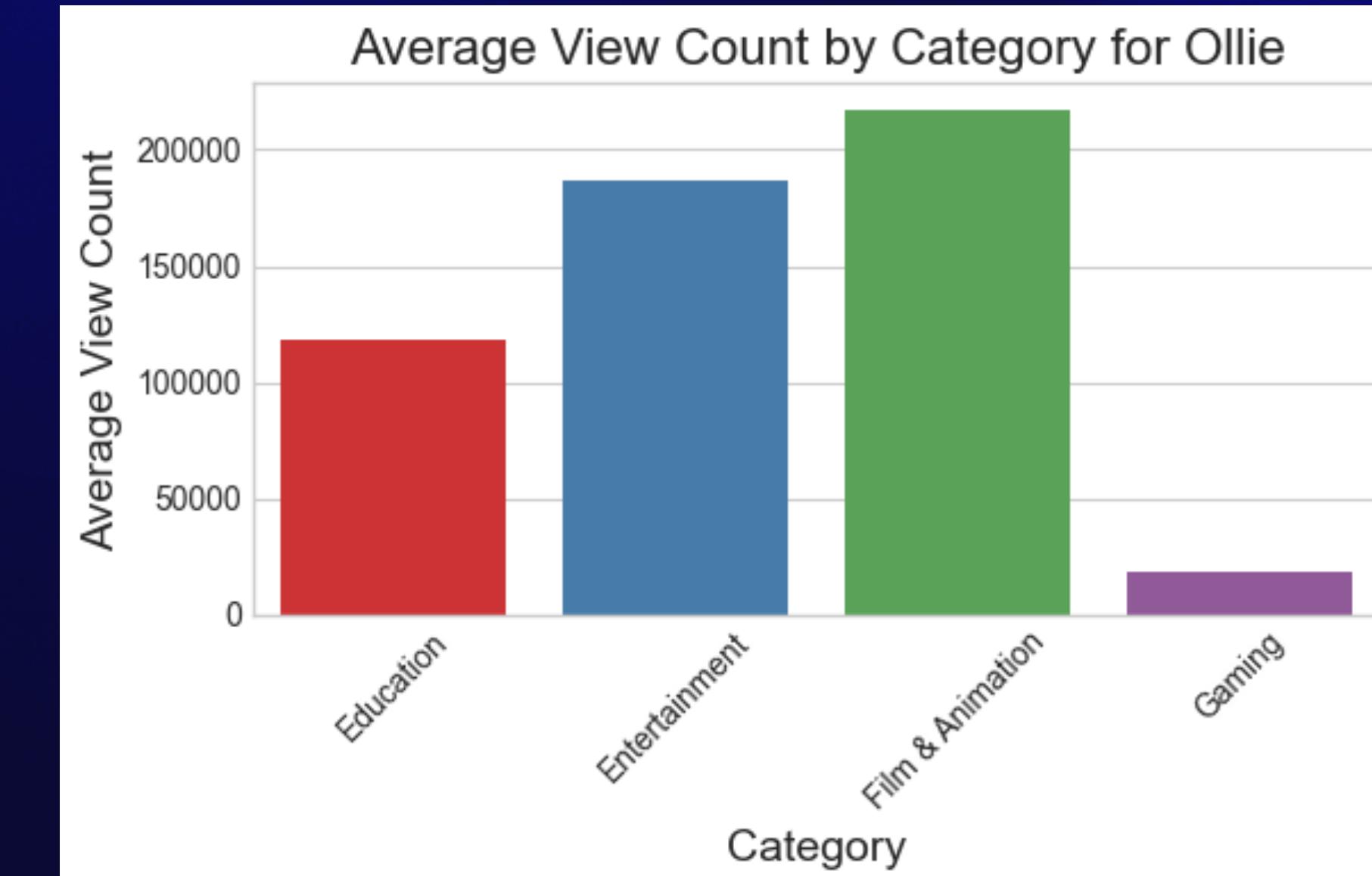
# 04-6: 탐색적 데이터 분석

## 데이터 인사이트: 업로드 카테고리별 평균 조회 수

- 핑크퐁은 Education 영상을 많이 올리는 반면에 다른 카테고리 영상들의 조회 수가 높았음을 발견 했다. 다만 Education외에 다른 카테고리의 영상 수는 매우 적기 때문에 정확한 비교가 되지는 못 한다.
- 아기상어 올리 또한 평균 조회 수는 적은 업로드 수의 카테고리도 평균 조회 수는 상당하다.



핑크퐁 채널 평균 조회 수



아기상어 올리 채널 평균 조회 수

# 05-1: 통계적 검정

## 시계열 트랜드

- 핑크퐁, 아기상어 올리, 베베핀 채널들의 업로드 요일별 시간별 통계적 검정을 실시 했다.
- 모든 데이터가 정규성과 등분산성을 띠지 않기 때문에 비모수적 방법을 채택했다.
- 요일별 비교:
  - Kruskal-Wallis H-Test 이후 Tukey's HSD
    - **토요일이 (p-value < 0.05)으로 다른 요일들에 비해 유의미하게 평균 조회 수가 높게 나타났다.**
  - 시간별 비교 (19시 vs. 10시)
    - Mann-Whitney U Test
      - 아기상어 올리 채널은 **10시에 영상을 업로드 했을 경우 19시보다 유의미하게 평균 조회 수가 높게 나타났다.**
      - 핑크퐁과 베베핀 채널을 차이가 유의미하지 않았지만 10시 업로드가 19시보다 평균 조회 수가 높았다.

# 05-2: 통계적 검정

## 비교분석

- 핑크퐁, 아기상어 올리, 채널들의 카테고리별 통계적 검정을 실시 했다.
- 베베핀 채널은 카테고리가 한 가지여서 비교분석에서는 제외가 되었다.
- 시계열 통계 검정과 마찬가지로 모든 데이터가 정규성과 등분산성을 띠지 않기 때문에 비모수적 방법을 채택했다.
- 핑크퐁 Education vs. Entertainment
  - Mann-Whitney U Test 실시한 결과 유의미한 차이가 없었다.
- 아기상어 올리 Entertainment vs. Film & Animation
  - Mann-Whitney U Test 실시한 결과 유의미한 차이가 없었다.
- 핑크퐁 채널과 아기상어 올리 채널 모두 유의미한 차이는 없었지만, 이전 슬라이드에서 평균 조회수의 차이는 보였기 때문에 다른 카테고리 영상을 업로드하는 것은 좋은 방향일 수 있다.

# 06: 대시보드 구축

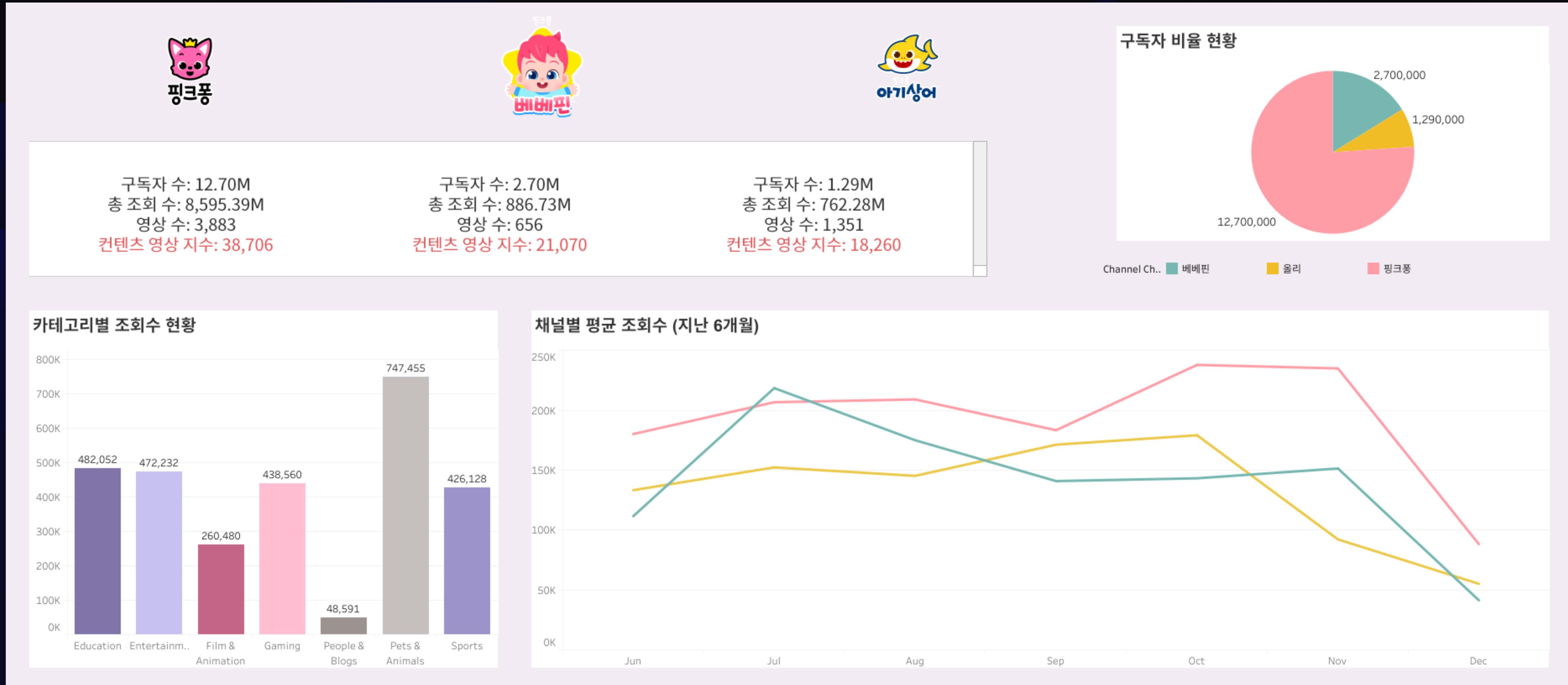
## Tableau 대시보드

아래의 대시보드는 [Tableau Public](#)에서 확인 가능.  
컨텐츠 영상 지수 (KPI) 계산 방법:

$$\text{View Rate (VR)} = \frac{\text{Total Views in Category}}{\text{Number of Videos in Category}}$$

$$\text{Proportion (P)} = \frac{\text{Number of Videos in Category}}{\text{Total Videos}}$$

$$\text{KPI} = \text{Proportion (P)} \cdot \text{View Rate}$$



# 07. 논의

## 결과와 느낀 점

- 결과:
  - 채널의 성장과 성공은 꾸준함에서 드러나기 때문에 토요일의 업로드된 영상이 유의미하게 높은 평균 조회 수가 보였더라도 토요일에만 올릴 수는 없다. 하지만 시간대는 이른 시간에 조정할 수 있을 것으로 보인다. 이는 아침 시간대(등교 전)와 점심 시간대의 시청자가 많은 것을 시사한다. 또한 시청자에게 강조하고 싶은 부분이 있다면 토요일이 적절한 날이다.
  - 유의미한 차이는 드러나지 않았지만 카테고리별 영상 업로드 수가 불균형해서 정확한 비교를 할 수 없었다. 하지만 시청자의 관심 또한 시간에 따라 변하기 때문에 다른 카테고리를 시도해 볼 수는 있다.
- 느낀 점:
  - 채널의 주인이 아니기 때문에 높은 퀄리티의 데이터를 수집하는 데에 있어서 한계가 있었다.
  - 만약에 더 자세한 데이터를 수집할 수 있다면 머신러닝으로 예측을 하거나 클러스터링을 통해 시청자의 관심사를 파악하고 그에 맞는 마케팅 전략을 세울 수 있을 것이다.

감사합니다!