유튜브 채널 성장과 영상 트렌딩 유지 기간 예측 분석

채널 특성과 영상 특성 분석 → 트렌딩 영상의 참여도 및 지속 기간에 따른 채널 성장 속도 예측

---

01. 프로젝트 개요

1. 채널 특성 분석
   - 어떤 채널이 트렌딩 영상을 많이 배출하는지 파악
   - 구독자 수, 업로드 빈도, 장르, 국가 등 채널 특성과 트렌딩 영상 발생 관계 분석석

2. 영상 특성 분석
   - 트렌딩 영상의 조회수, 좋아요 수, 댓글 기반 참여도와 트렌딩 유지 기간 예측

3. 채널 성장 예측
   - 분석 내용을 기반으로 채널 구독자 증가율 및 성장 속도 예측

---

02. 분석 목적

• 채널 전략 이해: 채널 특성이 트렌딩 영상 배출에 미치는 영향 분석
• 영상 성과 예측: 영상 특성을 기반으로 트렌딩 참여도와 지속 기간 예측
• 채널 성장 예측: 분석 내용을 기반으로 채널 성장 속도 예측 및 전략적 인사이트 도출

---

03. 사용 데이터셋

[Youtube 2025 channels]

• **column**: 채널별 구독자 수, 업로드 빈도, 장르, 국가등
• 활용: 채널 성장 패턴, 트렌딩 영상 비율 분석

[Youtube Trending Video Dataset]

• **column**: 영상별 조회수, 좋아요, 댓글, 업로드 시간, 카테고리, 해시태그등
• 활용: 트렌딩 영상 참여도 및 유지 기간 예측

---

04. 분석 포인트

[채널 분석]

• 채널별 트렌딩 영상 비율 계산: 트렌딩 영상 수 / 총 업로드 수

• 구독자 수, 업로드 빈도, 장르, 국가와 트렌딩 영상 발생 빈도 관계
• 향후 채널 성장 속도 예측 (e.g. 6개월 구독자 증가율)

[영상 분석]

• 영상별 트렌딩 유지 기간과 참여도 예측
• 독립 변수: 조회수, 좋아요, 댓글, 카테고리, 업로드 시간, 해시태그 수
• 종속 변수: 트렌딩 유지 기간 (트렌딩된 날짜 수)
• 변수 중요도 분석 (SHAP 등)

---

05. 시각화 계획

• 채널 성장 궤적: 구독자 수 vs 시간 (라인 그래프)
• 채널별 트렌딩 영상 비율: 산점도
• 트렌딩 영상 참여 예측: 라인 그래프 + 신뢰 구간
• 변수 중요도: 막대그래프 (SHAP 등)
• 트렌딩 지속 분포: 히트맵, 장르별/국가별 비교

[Week 1] 데이터 이해 · 전처리 · EDA

목표: 데이터 구조 파악 / 파일 클리닝 / 트렌딩 기간 계산 / 채널 – 영상 연결 / 기본 통계 / 결정트리 조사

---

06. 데이터셋 확보 & 구조 이해

- KaggleHub로 데이터 다운로드
- 두 데이터셋 구조 분석 (`info()`, `describe()`, `null` 체크)
- 필요한 컬럼만 필터링해서 샘플 저장

---

07. 데이터 전처리

- 날짜 타입 (`datetime`) 변환
- 영상별 `trending_days` 계산 → `video_id` 기준으로 `trending_date count`
- `channelId` 기준으로 두 데이터 연결 가능하도록 `key` 확인
- 결측치 처리, 범주형 인코딩 (단순 레이블인코딩)
- 카테고리 / 해시태그 파싱
- 채널별 트렌딩 비율 계산 → 트렌딩 영상 수 / 총 업로드 수

---

08. 기본 EDA

- 조회수·좋아요·댓글 분포 확인
- 채널별 트렌딩 영상 비율 vs 구독자 관계
- 업로드 시간대 분포
- 국가별 트렌딩 영상 차이
- 트렌딩 유지 기간의 평균·중앙값 등 기초 통계

---

09. 프로젝트 문서화 (팀장)

- 프로젝트 개요 정리
- 사용 데이터·전처리 기록
- EDA 결과 정리
- GitHub / Notion 구성 정리

---

10. 팀별 숙제 – Decision Tree 조사 & 발표 준비 (12 / 2)

- 의사결정 나무 기본 원리

- Entropy / Gini
- 하이퍼파라미터 (`max_depth` 등)
- 예시 코드
- 우리 프로젝트 적용 방안
- PPT 제작 및 발표 준비

📌 **Week 1** 산출물

- clean_channel_v1.csv, clean_trending_v1.csv
- `trending_days` 포함한 결합 데이터
- 채널별 트렌딩 비율 테이블
- Decision Tree 발표자료 준비 완료

[**Week 2**] 예측 모델링 (영상 예측 + 채널 성장 예측

목표: 영상 참여도 예측 / 트렌딩 유지 기간 예측 / 채널 성장 예측 모델 구축 + 성능 비교

---

11. 영상 분석 모델링

[목표]

- 영상별 트렌딩 유지 기간(`trending_days`) 예측

[독립변수]

- 조회수, 좋아요, 댓글, 카테고리, 업로드 시간, 해시태그 수

[모델 후보]

- Decision Tree Regressor
- Random Forest
- XGBoost

[성능 평가]

- RMSE, MAE
- SHAP 기반 변수 영향도 분석

---

12. 참여도 점수 모델링

- 참여도 = (좋아요 + 댓글) / 조회수 형태로 정의
- 회귀 모델 또는 분류 모델 실험
- “트렌딩 유지가 오래되는 영상” 특징 분석

---

13. 채널 성장 예측 모델링

[목표]

- 구독자 3개월 / 6개월 성장률 예측

[독립변수]

- 업로드 빈도, 장르, 나라, 트렌딩 영상 비율 등

[모델 후보]

- 회귀: Linear / RandomForest / XGBoost
- 분류: 성장 빠른 그룹 vs 느린 그룹

---

14. SHAP 분석

- 어떤 특징이 모델 결과에 가장 많이 기여하는지 설명
- 영상·채널 모델 각각 SHAP plot 생성

📌 Week 2 산출물

- 영상 예측 모델 코드 +성능표
- 채널 성장 예측 모델 코드+ 성능표
- SHAP 변수 중요도 시각화

[Week 3] 시각화 · 대시보드 · 발표 자료 제작

목표: 결과 시각화 · 스토리라인 제작 · 최종 발표 준비

---

15. 시각화 제작

- 채널 성장 궤적 (구독자 vs 시간)
- 채널별 트렌딩 영상 비율 산점도
- 트렌딩 유지 기간 히트맵
- 모델 성능 비교 그래프
- SHAP 중요도 막대그래프

---

16. Insight 생성

- 어떤 채널이 트렌딩을 잘 만드는가?
- 어떤 영상이 오래 트렌딩 되는가?
- 어떤 요소가 성장에 가장 영향력 있는가?
- 정책 제안 (업로드 빈도, 해시태그 수, 업로드 시간 등)

---

17. 최종 발표자료 제작

1. 프로젝트 개요
2. 데이터 설명
3. 전처리 과정
4. 모델링 접근
5. 결과 (그래프, SHAP)
6. 인사이트
7. 한계점 & 개선 방안

---

18. 발표 리허설

- 역할 분담 (데이터 / 모델 / 인사이트)
- 발표 흐름 점검
- 질의응답 준비

📌 Week 3 산출물

- 모든 그래프 및 그림
- 대시보드 또는 notebook 보고서

- 최종 발표자료 (PPT)
- 발표 스크립트

프로젝트에 필요한 컬럼 정리

[Youtube 2025 Channels]

1. 기존 컬럼 리스트

컬럼명	필요한 이유
channel_id	트렌딩 영상 데이터와 연결하는 핵심 key
channel_name	시각화·리포트용 식별자
subscriber_count	채널 규모 → 성장 예측 모델 핵심
view_count	총 조회수 → 채널 영향력 판단
video_count	전체 업로드 수 → 트렌딩 영상 비율 산출 필요
created_date	채널의 나이 (개설 연도) 파생 변수 생성 가능
category	채널 콘텐츠 성격 분석 (예: 음악/게임/뉴스 등)
country	국가별 트렌딩 패턴 차이 분석
video_last_30_days	최근 업로드 빈도 → 성장 요인
views_last_30_days	최근 채널 활성도 → 성장 예측 모델에 중요

2. 파생 컬럼 리스트

컬럼명	설명
channel_age_days	created_date 기준
upload_frequency	채널 개설 이후 하루당 평균 업로드 수
subscriber_per_view	누적 조회수 대비 구독자 비율
views_per_video	채널 영상 1개당 평균 조회수

uploads_per_subscriber	구독자 1명당 업로드 수
category_encoded	카테고리 문자열을 머신러닝 입력용으로 숫자 (Label Encoding) 로 변환
country_encoded	국가 문자열을 머신러닝 입력용 숫자로 변환

3. 파생 컬럼 생성 방식

[channel_trending_video_count]

- trending 데이터에서 채널별 video_id 수 count
- groupby(channelId).video_id.nunique()

[channel_trending_ratio]

- channel_trending_video_count/ video_count

[channel_age_days]

- (기준일자 - created_date).days

[upload_frequency]

- video_count/ channel_age_days

[subscriber_per_view]

- subscriber_count/ view_count

[views_per_video]

- view_count/ video_vount

[uploads_per_subscriber]

- video_count/ subscriber_count

[category_encoded]

- LabelEncoder().fit_transform(category)

[country_encoded]

- LabelEncoder().fit_transform(country)

[Youtube Trending Video Dataset]

4. 기존 컬럼 리스트

컬럼명	필요한 이유
video_id	트렌딩 데이터 집계 (trending_days 계산)
channelId	채널 데이터셋과 연결하는 key
title	분석 시 중요치 않지만 레코드 확인 등에 편리
publishedAt	업로드 후 트렌딩까지 걸린 시간 계산 가능
trending_date	trending_days 계산 핵심 컬럼
category_id	영상 카테고리별 트렌딩 차이 분석
tags	해시태그 수 파생변수 가능
view_count	참여도 계산 및 모델링 독립변수
likes	참여도 계산 독립변수
comment_count	참여도 계산 독립변수

5. 파생 컬럼 리스트

컬럼명	설명
trending_days	video_id 기준 등장 횟수
tags_list	태그 파싱
tags_count	태그 개수

category_name	카테고리 매핑
publish_month	월별 분석용
publish_dayofweek	요일별 분석용
days_since_publish	업로드~트렌딩까지 일수
like_ratio	likes / views
comment_ratio	comment_count / views
engagement_score	(likes + comments) / views
country	국가별 트렌딩 패턴 분석 가능

6. 파생 컬럼 생성 방식

[trending_days]

- 동일한 video_id가 여러 날짜에 등장 → video_id 별 trending_date 개수 count
- groupby(video_id).size()

[tags_list]

- '|' 기준 split
- "[ ]", "No tags" 같은 예외 처리
- tags_list = tags.split("|")

[tags_count]

- 리스트 길이 count: len(tags_list)

[publish_month]

- publish_time.dt.month

[publish_dayofweek]

- publish_time.dt.dayofweek

[days_since_publish]

- trending_date를 datetime으로 변환 후, 일(day) 단위로 계산

- days_since_publish= trending_date- publish_time

[like_ratio]

- like/ views
- `views` = 0 → 예외 처리

[comment_ratio]

- comment_count/ views

[engagement_score]

- (likes+ comment_count) / views

프로젝트 정리

---

1. 프로젝트 목적 및 필요성

요즘 누구나 유튜브에 영상을 올리면서 콘텐츠 경쟁은 더욱 치열해지고 있다. 따라서 어떤 영상과 채널이 더 주목을 받고, 그 관심이 얼마나 지속되는 지를 파악하는 것이 중요하다. 이번 프로젝트는 이러한 트렌딩 패턴을 분석하고 이를 바탕으로 채널이 성장하기 위해 필요한 실질적인 방향을 제시하는 것을 목표로 한다.

---

2. 데이터 구성 및 주요 변수

2-1. 원본 트렌딩 영상 데이터셋 (**Youtube Trending Video Dataset**)

[주요 컬럼]

- `video_id`: 영상 고유 ID
- `title`: 영상 제목
- `channel_title`: 채널명
- `category_id`: 카테고리 ID
- `publish_time`: 업로드 시간
- `trending_date`: 트렌딩 날짜
- `tags`: 태그 정보
- `views`, `likes`, `dislikes`, `comment_count`: 성과 지표
- `description`: 영상 설명

[활용 목적]

- 트렌딩 유지 기간 예측
- 영상 참여도 예측 모델 학습용

2-2. 트렌딩 영상 전처리 데이터 (**`clean_trending.csv`**)

[생성 과정]

- 원본 트렌딩 영상 데이터 → 전처리 + 파생변수 생성

[전처리 내용]

4. 날짜 데이터 `datetime` 변환
   - publish_time→ publish_datetime
   - trending_date→ trending_datetime

5. 영상별 트렌딩 유지 기간 생성
   - `trending_days` = `video_id` 기준 트렌딩 날짜 수

6. 태그/해시태그 파생변수 생성
   - tag_count, hashtag_count

7. 참여도 지표 생성
   - engagement_rate= (likes+ comment_count) / views

8. 수치형 컬럼 변환 및 결측치 처리

2-3. 원본 채널 데이터셋 (**Youtube 2025 Channels Dataset**)

[주요 칼럼]

- `channel_id`: 채널널 고유 ID
- `channel_name` : 채널명
- `channel_title` : 채널명
- `subcriber_count` : 구독자 수
- `view_count` : 누적 조회수
- `video_count`: 총 업로드 영상 수
- `category` : 채널 카테고리
- `country` : 국가
- `created_date` : 채널 개설일

[활용 목적]

- 채널 특성(규모, 장르, 국가 등) 분석
- 채널 성장 예측 모델을 위한 기본 메타데이터 제공

2-4. 채널 전처리 데이터 (**`clean_channel.csv`**)

[생성 과정]

- 원본 채널 데이터 → 필수 컬럼 정제 → 파생변수 생성 → 범주형 인코딩

[전처리 내용]

- 필요한 컬럼만 선택
- 구독자 수, 조회수, 영상 수 → 숫자형 변환
- `created_date` → `datetime` 변환
- `channel_age_days` = 채널 개설 후 경과 일수 생성
- `upload_frequency` = 총 영상 수 / 채널 나이
- 범주형 변수(`category`, `country`) 인코딩

3. 분석 방법 소개

3-1. 의사결정트리 **(Decision Tree)**란?

- 데이터를 특정 조건에 따라 분기하면서 결과를 예측하는 규칙 기반**(machine learning)** 분석 모델

3-2. Decision Tree 예시

- “좋아요 10만 이상 → 분기 A”, “해시태그 5개 이하 → 분기 B” 와 같이 조건을 나누며 예측 규칙을 자동으로 학습
- 트리 구조로 시각화되므로 결과 해석이 매우 직관적

3-3. Decision Tree 기반 모델 종류

[결정 트리 회귀 분석 – **Decision Tree Regressor**]

- 단일 트리를 사용하여 연속형 목표 변수(트렌딩 유지기간)를 예측
- 가장 기본적인 의사결정트리 모델로 규칙을 직관적으로 확인 가능
- 영상의 조회수, 좋아요, 댓글, 해시태그 수, 채널 구독자 수, 업로드 시간대 등 입력 변수로 학습

[Random Forest]

- 여러 개의 Decision Tree를 결합한 앙상블 모델
- 과적합(overfitting) 위험 감소, 예측 정확도 상승
- 트리별 평균값(회귀) 또는 투표(분류) 방식으로 결과 산출
- 영상 단위 트렌딩 유지기간 예측 + 채널 성장 예측

💡 앙상블 모델: 여러 개의 모델을 조합해 하나의 예측 결과를 만드는 기법

[XGBoost]

- Gradient Boosting 기반 트리 앙상블
- 이전 트리의 오차를 보완하며 순차적으로 학습
- 정확도는 높지만 해석은 Random Forest보다 다소 복잡
- 영상 단위 트렌딩 유지기간 예측

💡 **Gradient Boosting**: 이전 모델의 오류를 보정하기 위해 약한 모델을 순차적으로 결합하여 더 강력한 모델을 만드는 앙상블 기법

3-4. 사용 배경 및 장점

[사용 배경]

- 트렌딩 유지기간에 영향을 주는 요인이 다양하고 범주형·수치형 변수가 혼합되어 있어 규칙 기반으로 패턴을 발견하는 데 적합
- "어떤 조건 조합이 트렌딩 유지기간을 늘리는가?"라는 질문에 명확한 규칙 형태로 대답 가능
- 복잡한 전처리 없이 바로 적용 가능하며, 변수 importance(중요도) 확인이 쉬움

[장점]

- 규칙이 눈에 보여 해석 용이
- 변수 중요도 파악 가능
- 범주형/연속형 변수 모두 처리 가능
- 실제 의사결정 구조와 유사하여 전략 도출에 유리

---

4. 예측 모델 구축 과정

4-1. 데이터 수집 및 정제

- 결측치 처리, 이상치 제거
- 날짜와 같은 문자열 정보를 수치형으로 변환
- 범주형 변수 인코딩 (one-hot encoding)

💡 **one-hot encoding**: 범주형 (categorical) 데이터를 0과 1로 이루어진 숫자 벡터로 변환하는 방법 (텍스트나 범주 정보를 수치화할 때 사용)

4-2. 변수 선택

- 상관관계 분석
- 중요 변수 후보 탐색 (조회수, 좋아요, 해시태그 등)

4-3. 학습 데이터 / 검증 데이터 분리

- Train / Test = 8 : 2 비율 (예시) → 전체 데이터 80%로는 학습, 20%로는 평가

💡 **Train** 데이터 (학습용)
  - 모델이 규칙과 패턴을 배우는 데이터
  - e.g. 유튜브 영상 데이터 80%를 학습용으로 사용

💡 **Test** 데이터 (검증용)
  - 모델이 배우지 않은 새로운 데이터에 얼마나 잘 맞는지 평가하는 데이터
  - 이 데이터로 모델의 성능 (정확도, RMSE 등)을 측정
  - e.g. 나머지 20%를 검증용으로 사용

4-4. 의사결정트리 모델 학습

- 목표 변수: 트렌딩 유지기간
- 입력 변수: 채널·영상 특성

💡 목표 변수 (**Target Variable**)
- 모델이 예측하려고 하는 값
- 트렌딩 유지기간은 연속적인 숫자 값이므로 회귀 문제로 분류

💡 입력 변수 (**Features**)
- 모델이 목표 변수를 예측할 때 참고하는 정보 (질문에 대한 단서)

4-5. 성능 평가

- 예측 정확도
- MAE, RMSE 등 회귀 평가 지표 활용

💡 **MAE** (**Mean Absolute Error**, 평균 절대 오차)
- 예측값과 실제값의 차이를 절댓값으로 바꿔 평균한 값
- 차이의 부호(+ / -)를 무시하고, 평균적으로 얼마나 벗어났는지 직관적으로 보여줌

💡 **RMSE** (**Root Mean Squared Error**, 평균 제곱근 오차)
- 예측값과 실제값의 차이를 제곱하고 평균한 뒤, 제곱근을 취한 값
- 큰 오차에 더 민감함 → 큰 오차가 있으면 RMSE가 크게 증가

4-6. 결과 시각화

- 트리 시각화
- 변수 중요도 그래프

---

5. 결론 및 한계점

5-1. 결론

- 의사결정트리를 활용해 유튜브 트렌딩 유지기간을 예측하고 영향 요인을 규칙 형태로 파악 가능
- 이를 통해 실제 유튜버나 기업 채널의 성장 전략 수립에 유용한 인사이트 제공 가능

5-2. 한계점

- 데이터 기간/표본이 제한적일 경우 일반화가 어려움
- 유튜브 추천 알고리즘은 공개되지 않아 완벽한 예측 불가능
- 의사결정트리는 과적합 위험이 있어 다른 모델과 비교 보완 필요

공통 전제

[공통 독립변수 (**Features**)]

- `view_count` : 조회수
- `likes` : 좋아요 수
- `comment_count` : 댓글 수
- `categoryId` : 영상 카테고리
- `publish_dayofweek` : 업로드 요일
- `tags_count` : 태그(해시태그) 개수

[타깃 (목표 변수)]

8. `trending_days` : 영상이 며칠간 트렌딩에 머물렀는가 (회귀)
9. `engagement_score` : 참여도 점수 **= (**좋아요 **+** 댓글**) /** 조회수 (회귀)
10. `high_engagement` : 참여도 상위 **20%** 영상 여부**(0/1)** (분류)

1. 트렌딩 유지기간 모델(trending_days)

1-1. Decision Tree 결과

변수	중요도
comment_count	0.508
likes	0.218
view_count	0.175
categoryId	0.067
tags_count	0.024
publish_dayofweek	0.007

[결과 분석]

1. 댓글 수 (0.51)
   → 트렌딩을 오래 유지하는 가장 강력한 조건: **"얼마나 많은 사람이 반응하며 토론했는지"**
2. 좋아요 (0.22), 조회수 (0.18)
   → 단순 노출보다 능동적 반응**(**댓글·좋아요**)**이 더 중요
3. 카테고리·태그·요일은 영향 거의 없음
   → 트렌딩 유지에는 콘텐츠 속성보다 '반응 강도'가 핵심

1-2. Random Forest 결과

변수	중요도
comment_count	0.489
likes	0.214
view_count	0.175
categoryId	0.072
tags_count	0.038

publish_dayofweek	0.013

[결과 분석]

- 댓글 **>** 좋아요 **>** 조회수 순서가 그대로유지됨
- 상위 3개 (댓글+좋아요+조회수) 합 **≈0.88**
  → 트렌딩 유지기간의 거의 모든 설명력이 '시청자 반응 지표'에 집중

1-3. XGBoost 결과

변수	중요도
view_count	0.364
comment_count	0.306
likes	0.147
categoryId	0.098
tags_count	0.044
publish_dayofweek	0.041

[결과 분석: **Tree/RF**와 다른 점]

- 1위가 조회수 **(view_count)**로 바뀜
- 댓글·좋아요도 여전히 상위권
- 카테고리, 태그, 요일의 비중이 Tree보다 조금커짐

XGBoost는 "복합적인 잔차 보정 모델"이라 단일 핵심변수(댓글)보다
조회수 **+** 댓글 **+** 좋아요 **+** 카테고리까지 세밀하게 조합해서 사용

1-4. 최종 정리

- 트렌딩 유지기간은 단순히 "조회수가 많은 영상"이 아니라,
  댓글과 좋아요가 많이 달리는 '활발한 반응형 영상'일수록 오래 유지됨
- 특히 Decision Tree와 Random Forest에서는 댓글 하나만으로도
  전체 영향도의 약 50%를 설명할 정도로 중요하게 작용

2. 참여도 점수 회귀 모델 (engagement_score)

2-1. Decision Tree 결과

변수	중요도
likes	**0.852**
view_count	0.147
comment_count	0.0006
categoryId	거의 0
tags_count	0
publish_dayofweek	0

[결과 분석]

- 참여도 스코어에서 좋아요가 대부분 **(85%)**
- 댓글이 거의 안 나오는 이유
  → likes와 comment가 강한 상관관계
  → 트리는 이미 likes로 대부분의 패턴을 다 설명해버림

2-2. Random Forest 결과

변수	중요도
categoryId	**0.404**
view_count	0.386
comment_count	0.165
likes	0.023

tags_count	0.012
publish_dayofweek	0.010

[결과 분석]

- 카테고리 **1**위
- Random Forest는 **"**장르별 참여 문화 차이**"**를 잡아낸 것
  - 예시1: 게임·연예 → 좋아요·댓글 활발
  - 예시2: 뉴스·정보 → 조회수는 높지만 참여는 낮음
- 참여도는 단순 수치 문제가 아니라 **'**카테고리별 시청자 성향**'**의 함수

2-3. XGBoost 결과

변수	중요도
likes	**0.842**
publish_dayofweek	0.053
tags_count	0.048
comment_count	0.029
categoryId	0.020
view_count	0.007

[결과 분석]

- 다시 likes 중심 구조로 회귀
- XGBoost는 수식 구조를 가장 정직하게 반영
- 요일, 태그 개수도 미세한 영향 요인으로 활용

2-4. 최종 정리

- 참여도 점수는 정의식 자체가 (좋아요 + 댓글) / 조회수 이기 때문에
  Tree와 XGBoost에서는 좋아요가 전체 영향도의 약 **85%**를 차지하는 핵심 변수로 나타남
- 반면 Random Forest에서는 카테고리와 조회수가 가장 중요한 변수로 나타나,
  장르에 따라 시청자의 반응 문화가 구조적으로 다르다는 점을 보여줌

---

3. 참여도 상위 20% 분류 (high_engagement)

3-1. Random Forest 분류 결과

변수	중요도
view_count	**0.495**
likes	0.417
comment_count	0.065
categoryId	0.014
tags_count	0.006
publish_dayofweek	0.002

3-2. XGBoost 분류 결과

변수	중요도
likes	**0.457**
view_count	0.430
comment_count	0.063
categoryId	0.034
tags_count	0.012
publish_dayofweek	0.006

3-3. 분류 모델 결과 분석

11. **`view_count` + `likes` 합 ≈ 0.90** 이상
   → `high_engagement` 여부는
   "얼마나 많이 봤는가?", "그중에서 좋아요가 얼마나 나왔는가?"로 대부분 결정됨
12. `comment_count`는 보조 역할
   → 참여가 이미 높아진 이후의 강화 신호
13. 카테고리·태그·요일은 결정 변수 아님
   → 상위 20% 참여 영상은 장르보다 '폭발적 반응'이 더 중요

3-4. 최종 정리

- 참여도 상위 20% 영상을 분류하는 모델에서는
  조회수와 좋아요 두 변수만으로 전체 중요도의 **90%** 이상이 설명됨
- 고참여 영상은 "많이 본 영상 중에서도 특히 좋아요가 많이 눌린 영상"이라는 구조가
  매우 뚜렷하게 나타남

---

3. 전체 종합 인사이트

3. 트렌딩을 길게 만드는 핵심

- 댓글 > 좋아요 > 조회수
- 트렌딩은 "바이럴"이 아니라 지속적인 참여의 결과

4. 참여도를 만드는 핵심

- 기본 구조는 **`likes`** 중심
- 단, Random Forest에서는 카테고리별 문화 차이가 뚜렷함

1. 트렌딩 유지기간(trending_days) 모델 심층 분석

1-1. comment_count가 모든 모델에서 최상위인 이유

💡 트렌딩 구조 자체가 "활동 신호 기반"

유튜브 트렌딩 알고리즘은 단순 조회수보다 "단시간 안에 얼마나 많은 '상호작용'이 발생했는가"를 핵심 신호로 사용한다.

- 조회수: 수동적 소비
- 댓글: 능동적 참여 + 논쟁 + 확산 가능성

즉, 댓글이 많다는 것은 "사람들이 단순히 본 것이 아니라, 남에게 공유하고 싶을 만큼 반응했다" 는 신호이기 때문에 플랫폼 입장에서는 트렌딩 유지 조건에 가장 강한 지표가 된다.

따라서 트리 모델은 "댓글 수가 많은 영상 → 트렌딩에 오래 남을 확률이 급격히 증가" 라는 규칙을 가장 먼저 분기 조건으로 사용하게 된다.

💡 좋아요와 조회수보다 댓글이 더 "선별적 변수"

- 조회수: 광고·추천에 의해 단기간 폭증 가능
- 좋아요: 비교적 쉽게 누를 수 있음
- 댓글: 시간과 인지적 노력이 필요한 행동

따라서 댓글은 "가짜 트래픽"이나 "일시적 노출"의 영향을 상대적으로 덜 받는 변수이기 때문에 모델이 댓글을 가장 신뢰도 높은 설명 변수로 판단한 것이다.

1-2. XGBoost에서 view_count 비중이 커진 이유

💡 Decision Tree · Random Forest vs XGBoost

1. Tree / RF

- 단순 분기 + 평균화 구조
- 댓글처럼 "강한 단일 신호"를 우선 사용

2. XGBoost

- 이전 예측 오차를 계속 보정하는 잔차 기반 모델
- 단일 변수보다 여러 약한 신호를 조합해서 사용

따라서 XGBoost는 댓글, 좋아요뿐 아니라 “영상이 얼마나 대중적으로 노출되었는지(조회수)”까지 함께 반영하여 더 정교한 보정용 설명 변수로 활용한 것

- Tree/RF: 댓글 중심 설명
- XGBoost: 조회수 + 댓글 + 좋아요 종합 판단

1-3. 트렌딩 모델 요약

- 트렌딩 유지기간은 단순 인기(`view`)보다 ‘참여 기반 인기(`comment`, `likes`)’에 더 민감하게 반응
- 특히 댓글은 “장기 트렌딩의 핵심 원인 변수” 로 작동
- 이는 유튜브 트렌딩이 단기 바이럴이 아니라, 지속적 반응을 유도하는 콘텐츠를 우선 추천한다는 구조적 특성 반영

---

2. 참여도 점수(`engagement_score`) 회귀 모델 심층 분석

💡 `engagement_score` = (`likes` + `comments`) / `view_count` 구조이기 때문에, 모델 해석은 “수식 구조 + 변수 상관 + 모델 특성” 3가지 관점에서 봐야 한다.

2-1. Tree / XGBoost에서 `likes`가 0.84 이상인 이유

💡 **`engagement_score`** 수식 구조의 직접적 영향

engagement_score= (likes+ comments) / view_count

→ 분자의 `likes`가 가장 직선적인 영향력을보임

→ Tree, XGBoost는 이런 직접적인 비선형 패턴을 매우 잘 학습함

따라서 **`likes`** 증가 → **`engagement_score`** 급증 이라는 가장 단순하고 강한 규칙이 그대로 모델 구조에 반영된다.

💡 **`comment_count`**의 중요도가 낮게 나온 이유

`comment_count`는 `likes`와 강한 상관 관계를 가진다.

- 좋아요가 많은 영상 → 댓글도 많은 경향을 보임
- 모델 입장에서 이미 `likes`로 설명이 되면 `comment_count`를 추가로 쓸 필요가 줄어듦듦

이 현상을 **"다중공선(Multicollinearity)"** 이라고 한다. 이에 따라 Tree/XGB에서는 `comment_count`의 중요도가 거의 0에 가까워진 것이다.

2-2. Random Forest에서 `categoryId`가 1위가 된 진짜 이유

Random Forest에서는 다음 패턴이 나왔다

변수	중요도
`categoryId`	0.403
`view_count`	0.386
`comment_count`	0.165
`likes`	0.023

→ 단순 오류가 아닌 매우 의미 있는 결과

💡 "장르별 참여 문화 차이"를 잡아낸 것

Random Forest는 수많은 트리 평균 모델이기 때문에 단일 숫자 변화보다 **"집단적 패턴 차이 (군집 특성)"** 를 잘 잡아낸다.
이 모델은 다음과 같은 패턴을 학습했다

- 게임, 스포츠, 연예: 조회수 대비 좋아요·댓글이 매우 활발
- 뉴스, 정보: 조회수는 많지만 좋아요·댓글 비율은 낮음

그래서 카테고리 (**`categoryId`**) 자체가 참여도를 좌우하는 구조적 변수로 작동한 것이다.

2-3. 참여도 회귀 모델 심층 결론

- Tree/XGB 관점: 참여도는 수식 구조상 **`likes`** 중심 현상
- Random Forest 관점: 참여도는 **"카테고리별 반응 문화 + 조회수 규모"**의 함수

따라서 종합 결론은 참여도는 단순히 "좋아요 수"의 문제가 아니라, 어떤 장르의 영상이 어떤 규모로 노출되느냐에 따라 구조적으로 달라진다.

---

3. 참여도 상위 20% 분류(high_engagement) 심층 해석

여기서는 RF, XGB 모두에서

- view_count≈ 0.50
- likes≈ 0.42
- 합 ≈ 0.90 이상

이라는 매우 극단적인 집중 현상이 나타났다.

3-1. 분류에서 조회수가 다시 최고가 된 이유

회귀와 분류의 문제 정의 차이 때문이다.

- 회귀: "참여도 점수를 정확히 몇 점으로 맞출 것인가"
- 분류: "상위 20%에 들어갈 것인가 / 아닌가"

이때 상위 20%는 대부분 조회수 분포의 극단 영역 (상위 꼬리 부분)에 위치한다.
즉, 구조적으로 조회수가 일정 기준 이상 → 상위 **20%** 진입 확률 급상승이 되는 구간이 존재한다.

그래서 분류에서는 "많이 본 영상인지"가 1차 필터, 그 안에서 "좋아요가 많은 영상인지"가 2차 필터로 작동한다.

3-2. comment_count가 0.06~0.07 수준에 머무는 이유

댓글은 참여도의 질적 신호이다. 하지만 상위 20%를 결정적으로 가르는 '경계 변수'는 아니다.

즉, 고참여 영상이 되기 위한 필요조건은 '조회수 + 좋아요' 이며, 댓글은 "이미 고참여가 된 이후의 강화 요소" 이다.

그래서 분류에서는 댓글 = 서포트용 변수, 조회수·좋아요 = 결정 변수 로 역할이 분리된다.

1. SHAP

1-1. SHAP (SHapley Additive exPlanations)

원래는 게임이론**(Shapley value)**에서 나온 개념으로 모든 feature가 함께 예측값을 만드는 플레이어라고 생각하고, 각 피처가 예측값에 얼마나 기여했는지 공정하게 나누는 방식

[해석 포인트]

- SHAP 값 > 0: 그 피처가 예측값을 평균보다 크게 만드는 방향으로 작용
- SHAP 값 < 0: 예측값을 평균보다 작게 만드는 방향으로 작용
- |SHAP| 값 ∝ 영향력

즉, 이 피처가 이 샘플에서 예측값을 얼마나, 어느 방향으로 밀어 올렸는지/내렸는지를 수치로 보여줌

1-2. SHAP 플롯 읽는 법

[SHAP Bar Plot (mean |SHAP| bar)]

- X축: 평균 **|SHAP|** 값 (절댓값의 평균)
- Y축: 변수 이름

"전체 데이터를 기준으로, 어느 변수가 가장 영향을 많이 미쳤나?"에 관한 순위로 feature_importance랑 비슷한 느낌이지만, SHAP는 기여도를 기반으로 해서 더 이론적으로 깔끔함

1-3. SHAP Beeswarm Plot (summary plot)

- Y축: 변수 (위에 있을수록 중요)
- X축: SHAP 값 (왼쪽: 예측값 낮추는 방향, 오른쪽: 예측값 높이는 방향)

[점 색]

- 빨강 = 그 샘플에서 해당 변수 값이 크다**(High)**
- 파랑 = 해당 변수 값이 작다**(Low)**

[해석 포인트]

- `likes`에서 오른쪽에 빨간 점들이 몰려있음
  → "좋아요 수가 클수록 target이 커지는 방향으로 영향을 준다"
- `view_count`에서 왼쪽에 빨간 점들이 몰려있음
  → "조회수가 큰 일부 샘플에서는 예측값을 낮추는 방향으로 작용하는 경우도 관찰될 수 있음"

즉, **"값이 큰 샘플(빨강)이 오른쪽에 많으면 → 값이 클수록 예측을 키우는 변수", "값이 큰 샘플이 왼쪽에 많으면 → 값이 클수록 예측을 줄이는 변수"**라는 포인트만 잡으면 모든 beeswarm 플롯 해석 가능

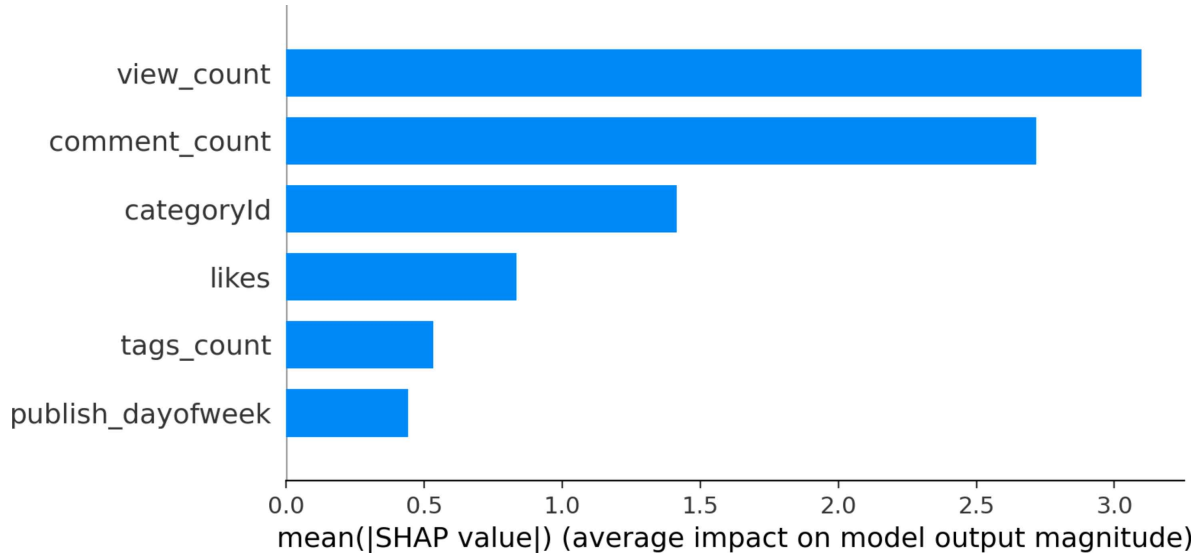
단, categoryId와 같이 숫자로 인코딩된 범주형 변수의 경우, beeswarm plot의 색상은 코드값의 크기를 의미할 뿐 해석에는 주의가 필요함

---

2. SHAP 분석 결과 해석

2-1. Trending_days 예측 모델 SHAP 분석

[전역 변수 중요도 (**Bar Plot**)]

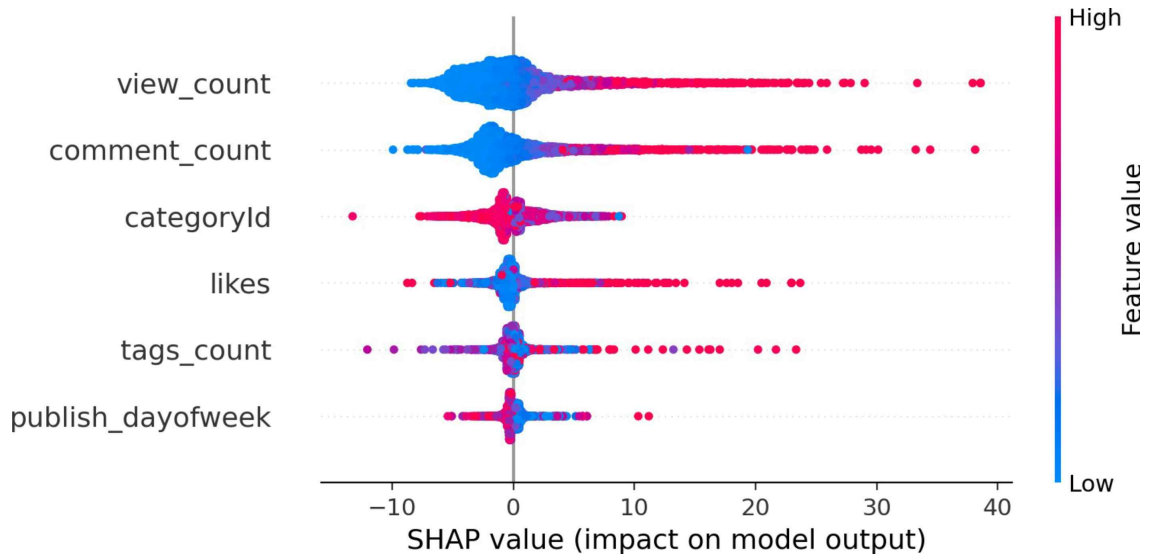


SHAP bar plot 결과, **trending_days** 예측에 가장 큰 영향을 미치는 변수는 **view_count**와 **comment_count**로 나타났다.
그 다음으로는 **categoryId, likes, tags_count, publish_dayofweek** 순으로 영향력이 감소하는 경향을 보였다.

이는 트렌딩 유지기간이 단순 조회수뿐만 아니라 시청자의 적극적 참여(댓글)와도 밀접한 관련이 있음을 시사한다. 또한 영상이 속한 카테고리(categoryId) 역시 트렌딩 지속 여부에 일정 수준의 영향을 미치는 요인으로 확인되었다.

본 분석에서는 bar plot 기준의 평균 절대 SHAP 값(mean(|SHAP|))을 기준으로 변수 중요도를 해석하였다.

[개별 변수 영향 방향성 (**Beeswarm Plot**)]



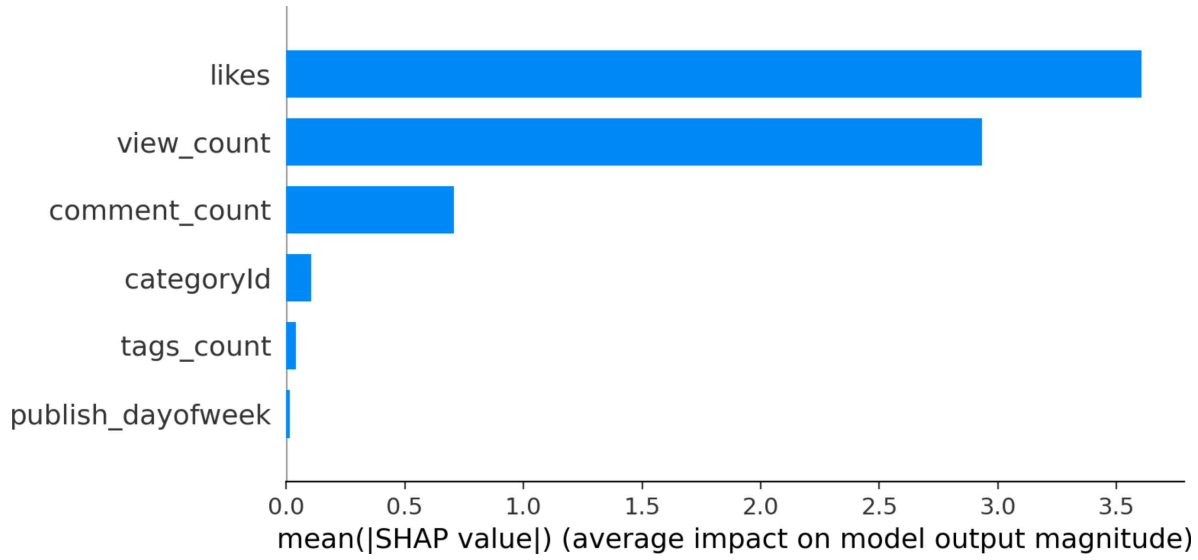
- **`view_count`**
  전반적으로 조회수가 높을수록 trending_days 예측값을 증가시키는 방향의 SHAP 값을 보였다. 다만 일부 구간에서는 높은 조회수임에도 불구하고 예측값을 감소시키는 사례가 관찰되었으며, 이는 비선형 효과나 다른 변수(예: likes, comment_count)와의 상호작용에 따른 결과일 가능성이 있다.
- **`comment_count`**
  댓글 수가 많은 영상일수록 trending_days 예측값을 증가시키는 경향이 비교적 일관되게 나타났다. 이는 시청자 참여도가 트렌딩 유지에 중요한 요인임을 의미한다.
- **`likes`**
  좋아요 수 역시 전반적으로 긍정적인 영향을 보였으나, view_count나 comment_count에 비해 영향력의 크기는 상대적으로 작게 나타났다.
- **categoryId (범주형 변수)**
  categoryId는 범주형 변수이므로 beeswarm plot에서의 색상(High/Low)은 단순한 코드값의 크기를 의미하며, 값의 대소 자체에는 해석적 의미가 제한적이다. 따라서 본 결과는 “특정 카테고리가 trending_days 예측에 긍정적 또는 부정적으로 작용할 수 있음” 정도로 해석하는 것이 적절하며, 카테고리별 효과를 정밀하게 비교하기 위해서는 원-핫 인코딩 후 더미 변수 단위의 SHAP 분석이 필요하다.

[해석 시 유의사항 (타깃 누수 가능성)]

본 모델에서 사용된 view_count, likes, comment_count와 같은 참여 지표는 시간에 따라 누적되는 변수이므로, 예측 시점 이후의 정보가 포함될 경우 타깃 누수(**target leakage**)가 발생할 가능성이 있다.
따라서 본 결과는 "해당 지표들이 예측 시점 이전에 관측 가능한 값(예: 트렌딩 진입 초기 또는 업로드 직후 일정 시간 내 값)"이라는 가정하에 해석되어야 하며, 향후 분석에서는 시점 제한(feature cutoff)을 명확히 적용할 필요가 있다.

2-2. High_engagement 분류 모델 SHAP 분석

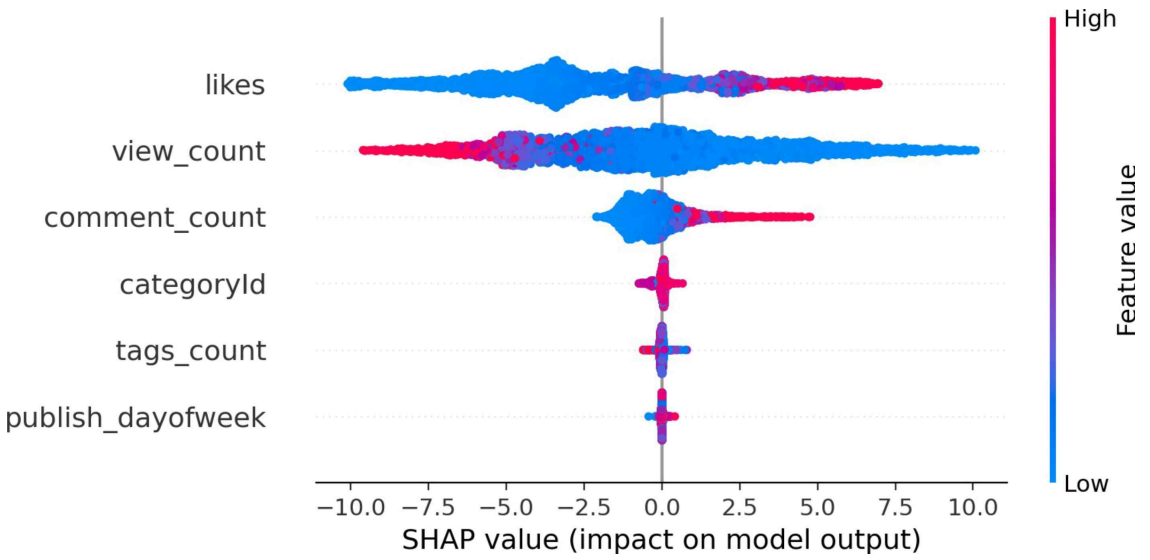
[전역 변수 중요도 (**Bar Plot**)]



SHAP bar plot 결과, **High_engagement**(참여도 상위 **20%**) 분류에 가장 큰 영향을 미치는 변수는 **likes**와 **view_count**로 나타났다.
두 변수의 평균 절대 SHAP 값은 다른 변수들에 비해 현저히 크게 나타났으며, 그 외 변수들의 영향력은 상대적으로 매우 제한적인 수준에 머물렀다.

이는 영상의 참여도 수준을 구분하는 데 있어 시청자의 직접적인 반응 지표가 핵심적인 역할을 수행함을 의미한다.

[개별 변수 영향 방향성 (**Beeswarm Plot**)]



Beeswarm plot을 통해 변수 값의 크기에 따른 예측 기여 방향을 살펴본 결과, 다음과 같은 경향이 관찰되었다.

- **`likes`**
  좋아요 수가 높은 영상일수록 High_engagement로 분류될 확률을 증가시키는 방향의 SHAP 값이 주로 관찰되었다. 이는 좋아요 수가 시청자의 적극적 반응을 가장 직접적으로 반영하는 지표임을 시사한다.
- **`view_count`**
  조회수 역시 전반적으로 High_engagement 예측값을 증가시키는 방향으로 작용하였다. 다만 likes에 비해 SHAP 분포의 폭이 넓어, 조회수 단독보다는 다른 참여 지표와의 결합 효과가 함께 작용했을 가능성을 시사한다.

그 외 변수들은 SHAP 값의 분포가 0 근처에 밀집되어 있어, High_engagement 분류에 미치는 영향이 제한적인 것으로 나타났다.

[해석 시 유의사항 (순환 정의 가능성)]

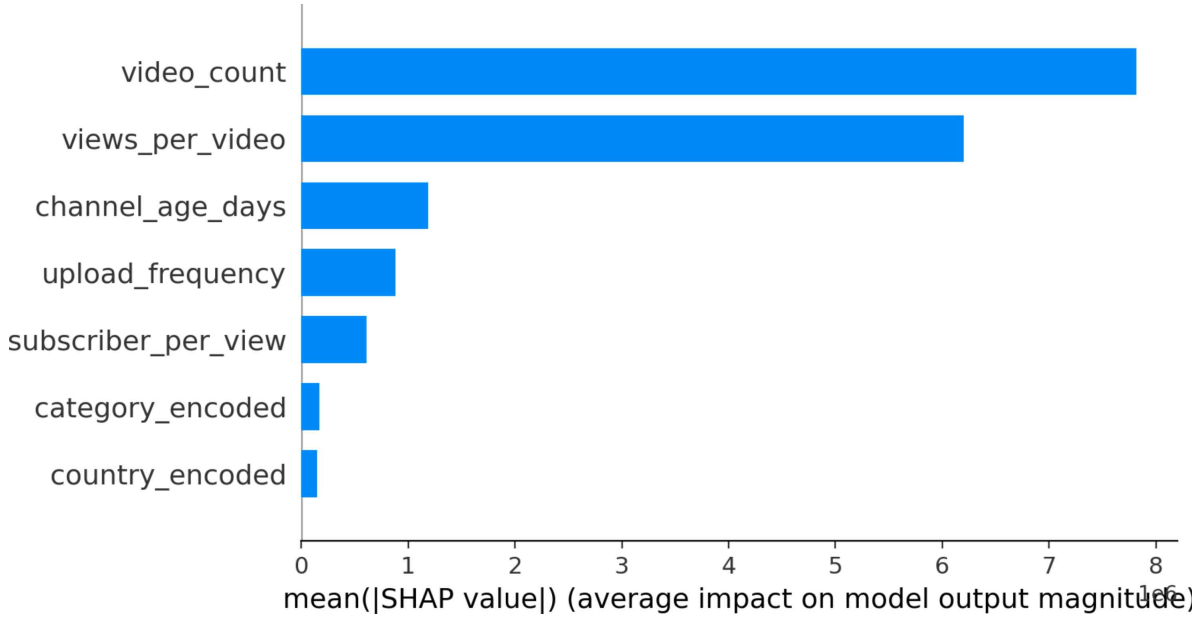
High_engagement 라벨이 likes, view_count, comment_count 등의 참여 지표를 기반으로 정의되었을 경우, 해당 변수들을 입력 피처로 사용하는 것은 예측하려는 '정답(label)'을 만들 때 쓴 정보가, 다시 입력 변수(feature)로 들어가 있는 상태의 순환 정의**(circularity)** 위험을 내포할 수 있다.
따라서 본 분석 결과는 "참여도 지표가 참여도 분류에 강하게 반영됨"을 확인하는 설명적 분석 수준에서 해석되어야 하며, 향후 연구에서는 라벨 정의에 포함되지 않은 외생적 변수 중심의 예측 모델링이 필요하다.

High_engagement(참여도 상위 20%) 분류 모델의 SHAP 분석 결과, **likes**와 **view_count**가 가장 높은 영향력을 보이는 변수로 나타났다.
그 외 변수들의 영향력은 상대적으로 매우 제한적인 수준으로 확인되었다.

이는 시청자의 직접적 반응 지표가 영상 참여도 수준을 구분하는 데 핵심적인 역할을 함을 의미한다.

2-3. Channel Growth (views_last_30_days) 예측 모델 SHAP 분석

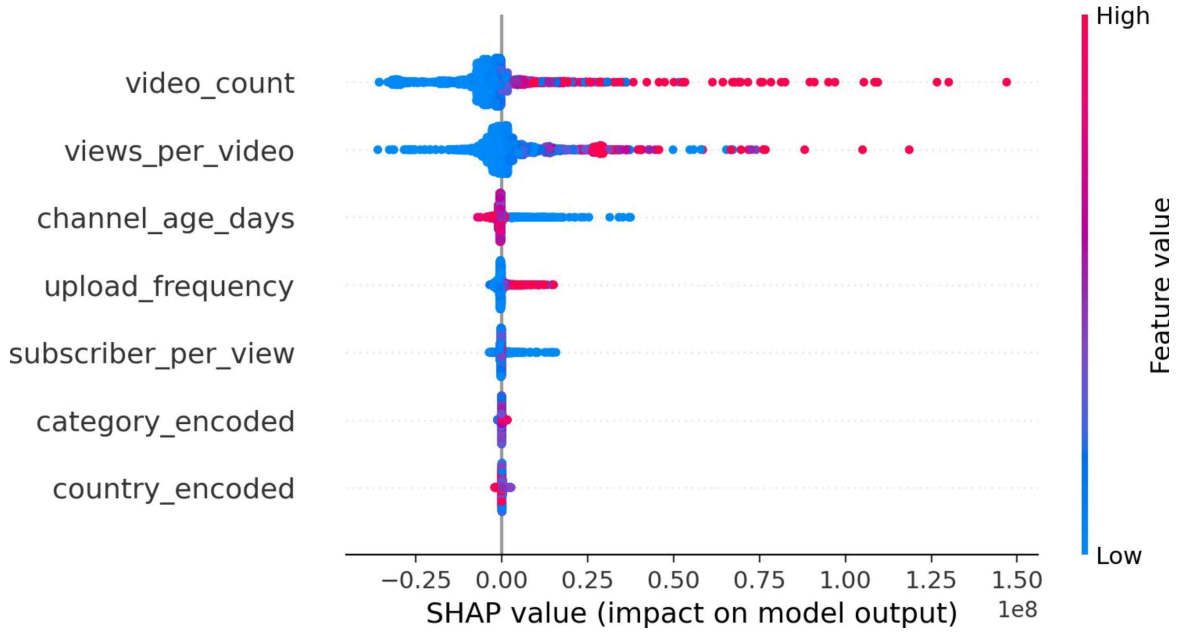
[전역 변수 중요도 (**Bar Plot**)]



Channel growth 예측 모델의 SHAP bar plot 결과, **video_count**(총 업로드 수)와 **views_per_video**(영상당 평균 조회수)가 가장 중요한 변수로 나타났다.
그 다음으로는 **channel_age_days, upload_frequency, subscriber_per_view** 순으로 영향력이 감소하였다.

이는 채널 성장에 있어 단기 성과 지표뿐만 아니라 지속적인 콘텐츠 생산량과 채널 운영 이력이 중요한 역할을 함을 시사한다.

[개별 변수 영향 방향성 (**Beeswarm Plot**)]



- **video_count**
  업로드 영상 수가 많을수록 views_last_30_days 예측값을 증가시키는 경향이 뚜렷하게 나타났다.
- **views_per_video**
  영상당 평균 조회수가 높은 채널일수록 채널 성장 예측값이 증가하는 방향으로 작용하였다.
- **channel_age_days**
  채널 연령이 증가할수록 성장 예측값에 미치는 영향은 완만하며, 신규 채널과 장기 운영 채널 간 차이가 비선형적으로 나타나는 경향을 보였다.

2-4. 종합 해석

SHAP 분석 결과,

- 트렌딩 유지기간은 조회수와 댓글 수를 중심으로 한 시청자 참여 지표의 영향을 크게 받는다.
- 참여도 분류는 라벨 정의 특성상 설명적 분석의 성격이 강하며, 예측 성능보다는 변수 반영 구조를 해석하는 데 의의를 둔다.
- 채널 성장은 단기 성과 지표와 함께 업로드 규모 및 채널 운영 특성이 함께 작용하는 구조임이 확인되었다.

본 분석은 SHAP을 활용하여 머신러닝 모델의 예측 구조를 해석 가능하게 제시하였다는 점에서 의의가 있으며, 향후에는 변수 시점 통제 및 범주형 변수의 정밀 분석을 통해 해석의 신뢰도를 더욱 제고할 수 있을것이다.

유튜브 채널 및 트렌딩 데이터 **EDA**

1. 결측치 분석

1-1. 채널 데이터 결측치치

컬럼명	결측치 개수	이유
created_date	1130	채널 생성일 정보가 없는 채널이 있음
channel_age_days	1130	생성일이 없어서 계산 불가
upload_frequency	1130	생성일이 없어서 계산 불가
category	108	일부 채널의 카테고리 정보 없음
country	1800	국가 정보 비공개 채널
subscriber_per_view	70	계산 불가/이상값 제거
views_per_video	74	계산 불가/이상값 제거

1-2. 트렌딩 데이터 결측치

- 트렌딩 데이터는 모든 값이 채워져 있어 결측치 없음

1-3. 병합 데이터 결측치

- 영상 단위 트렌딩 변수에서는 결측치 발생 없음
- 채널 단위 변수에서 대규모 결측치관찰

[원인]

- 트렌딩 영상은 존재하지만 해당 영상의 채널 정보가 채널 데이터에 포함되지 않아 병합 실패
- 일부 변수는 채널 데이터 원천 결측치가 병합 과정에서 중첩되어 결측치 규모 증가

1-4. 결측치가 대량 발생한 컬럼

컬럼명	결측치 개수
channel_id	1,918,962
channel_name	1,918,962
subscriber_count	1,918,962
view_count_y	1,918,962
video_count	1,918,962
videos_last_30_days	1,918,962
views_last_30_days	1,918,962
uploads_per_subscriber	1,918,962

1-5. 결측치가 가장 많이 발생한 컬럼

컬럼명	결측치 개수	결측 발생 이유
created_date	1,993,728	병합 실패 + 채널 데이터 원천 결측
channel_age_days	1,993,728	생성일 결측으로 파생 변수 계산 불가
upload_frequency	1,993,728	채널 나이 정보 부재로 계산 불가
category	1,920,028	병합 실패 + 일부 채널의 카테고리 미기재
country_y	2,076,465	병합 실패 + 국가 정보 비공개 채널 존재
subscriber_per_view	1,920,832	병합 실패 + 분모 0 또는 이상값 처리
views_per_video	1,921,094	병합 실패 + 분모 0 또는 이상값 처리

---

2. 구독자 수 분포분석

[로그 스케일 적용 이유]

- 구독자 수 범위가 매우 넓어 일반 스케일에서 분포 해석 어려움

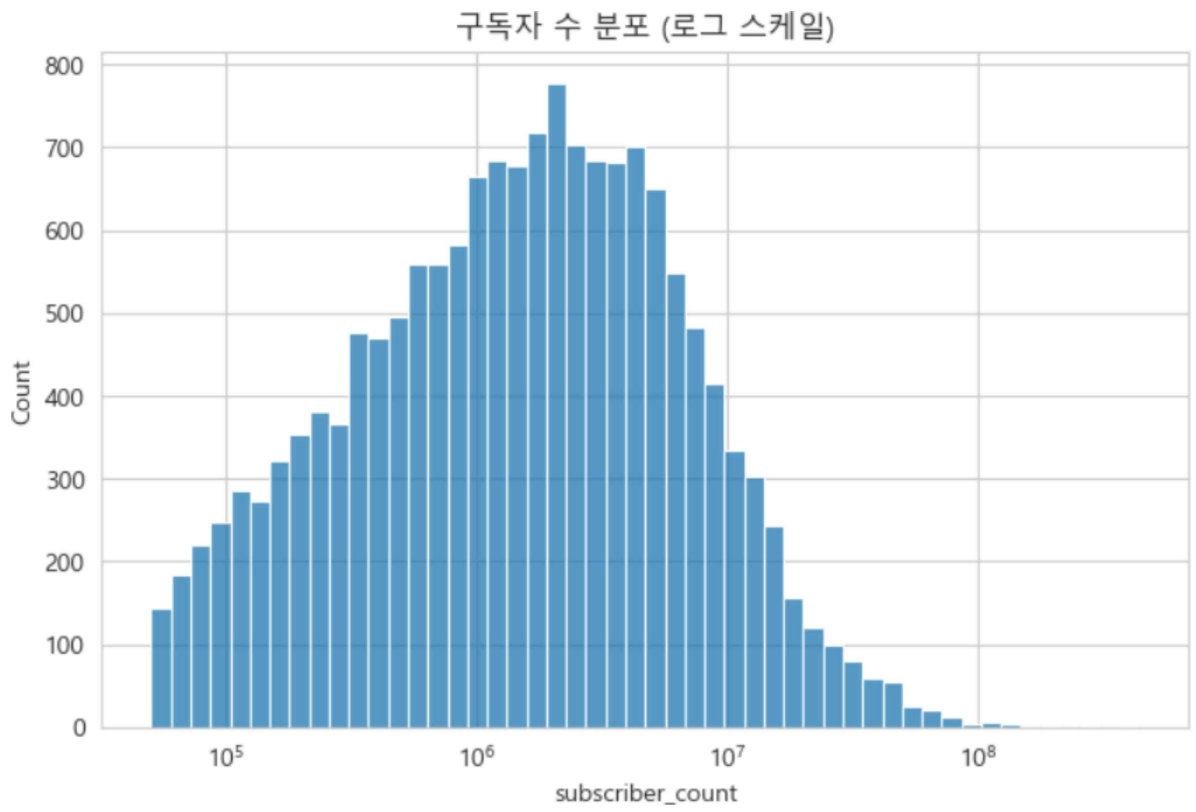
💡 로그스케일: 값의 차이를 배수(몇 배) 기준으로 보여주는 축

[로그 스케일 기준 구독자 수 분포]

- 10만~1천만 구간 가장 많음
- 1억 명 이상 극소수

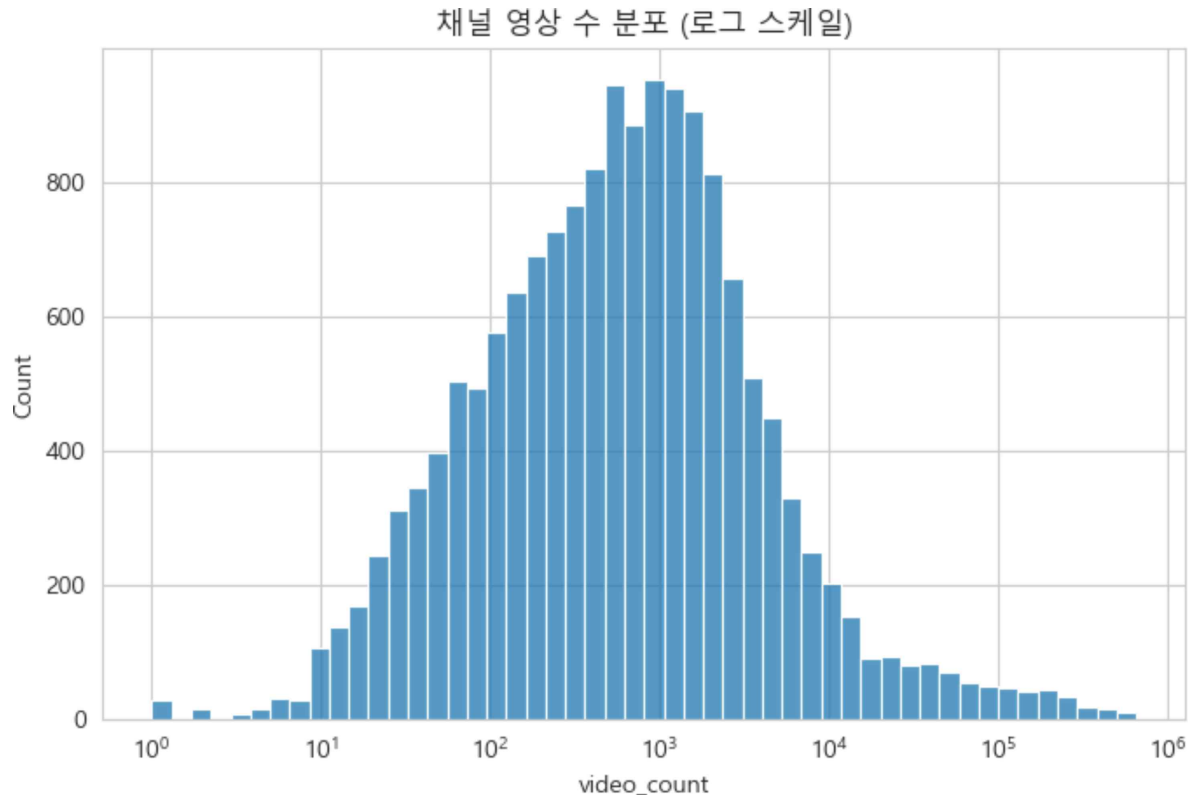
[특징]

- 대부분 중·소형 채널
- 롱테일 분포(오른쪽 꼬리 길게 이어짐)
- 유튜브 생태계는 구독자 수 기준 강한 불균형 구조
- 트렌딩 성과 분석 시 구독자 수만으로 판단하는 것은 한계, 영상 단위 성과와 연계 분석 필요



3. 채널 영상 수 분포

- 로그 스케일 기준 대부분 채널 10~1,000개 영상 구간 집중
- 1만 개 이상 영상 보유 채널 극소수
- 특징: 롱테일 분포, 채널 간 콘텐츠 생산량 편차 매우 큼
- 트렌딩 성과 분석 시 개별 영상 성과 지표 함께 고려 필요



---

4. 채널별 트렌딩 영상 수

channel_id	트렌딩 영상 수	해석
UC---IM1j0uNzsFxF0V2IZnw	4	서로 다른 영상 4개가 트렌딩에 진입한 채널로, 평균 이상의 트렌딩 성과를 보임
UC--3pN-wtQrC48GAK8wsBsg	2	트렌딩 영상이 2개로, 소수의 트렌딩 경험을 가진 일반적인 채널
UC--5TQCGo0KzYRjVi3OsPAQ	7	여러 개의 영상이 트렌딩에 진입하여 비교적 높은 트렌딩 성과를 보임

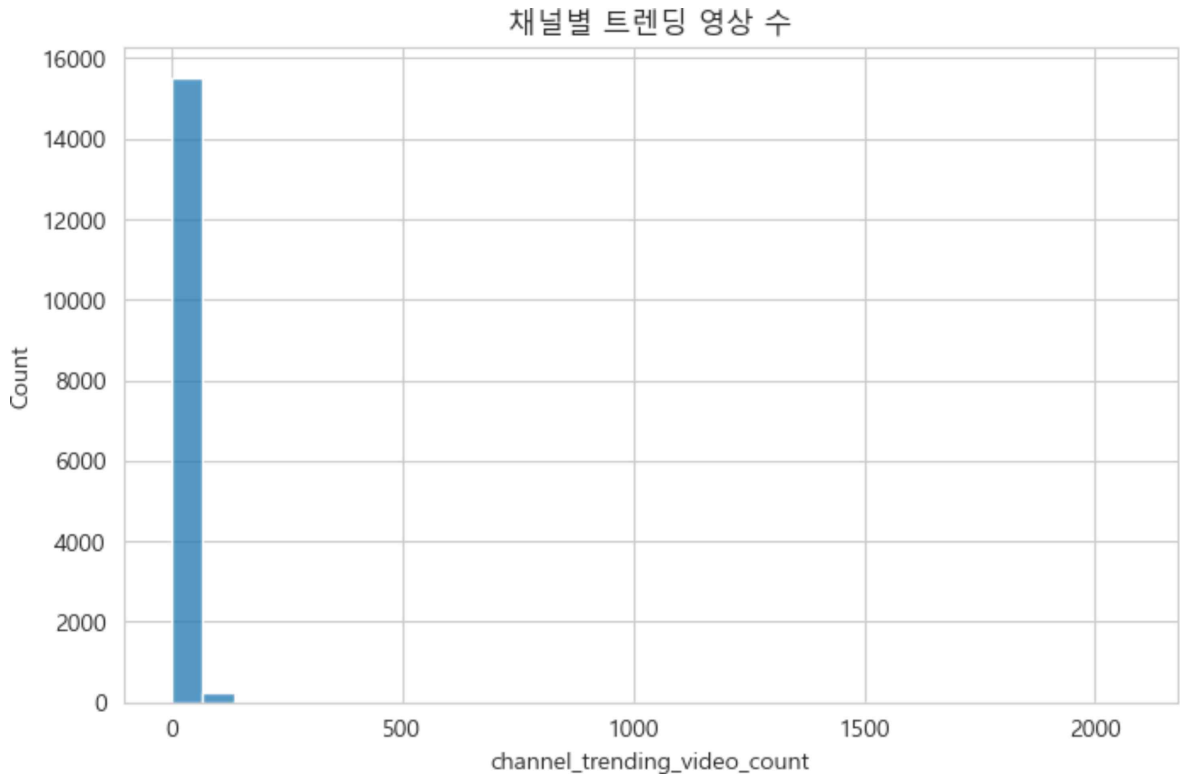
UC--FL6OwLFWGIZfLfazY4yA	24	다수의 영상이 반복적으로 트렌딩에 진입한 상위 트렌딩 채널
UC--MSNrR_zO4Zb5OvwjWF4A	1	단일 영상만 트렌딩에 진입한 채널로, 대부분 채널에 해당하는 일반적인 유형

[해석]

- 대부분 채널 1~2개 트렌딩 경험
- 일부 상위 채널은 다수 영상 반복 트렌딩
- 트렌딩은 소수 채널에 집중

4-1. 채널별 트렌딩 영상 분포 해석

- 대부분의 채널은 트렌딩에 단 한 번도 오르지 못하거나, 1~2개 영상만 트렌딩에 진입
  → 트렌딩은 대부분 채널에게 일상적인 현상이 아님
- 일부 상위 채널은 수천 개에 가까운 트렌딩 영상을 보유
  → 반복적으로 트렌딩을 만들어 내는 영향력 채널



💡 결론: 유튜브 생태계 전반에서 성과는 소수 채널에 집중되는 구조

5. 카테고리별 요약 (상위 **10%**)

- Music / Pop music / Electronic music / Film / Entertainment 포함
- 특징: 높은 평균 구독자 수, 조회수, 트렌딩 비율
- 해석: 대중적 콘텐츠 중심, 반복 소비 가능
- 주의: 대부분 카테고리 채널 수 적어 평균 지표는 대형 채널 영향 크게 받음

카테고리별 요약 (상위 10):

category	avg_subscribers	avg_views	avg_trending_ratio	count_channels
Music of Asia, Pop music, Music, Electronic music	9.900000e+07	3.996259e+10	9.900000e+07	1
Hip hop music, Music, Pop music, Electronic music, Soul music, Rhythm and blues	7.630000e+07	3.594206e+10	7.630000e+07	1
Music, Music of Asia, Film, Entertainment, Soul music	7.360000e+07	3.368513e+10	7.360000e+07	1
Lifestyle (sociology), Pop music, Music of Asia, Music, Film	6.830000e+07	3.830118e+10	6.830000e+07	1
Pop music, Music, Music of Asia, Electronic music, Hip hop music	6.415320e+07	6.416417e+10	6.415320e+07	5
Sport, Lifestyle (sociology), Entertainment, Film	6.150000e+07	1.970004e+10	6.150000e+07	1
Soul music, Pop music, Music of Asia, Music, Film	5.860000e+07	2.940033e+10	5.860000e+07	1
Soul music, Pop music, Music, Music of Asia, Film	5.840000e+07	3.899103e+10	5.840000e+07	1
Film, Pop music, Music of Asia, Music, Entertainment	5.441333e+07	3.519532e+10	5.441333e+07	3
Pop music, Hip hop music, Music, Music of Latin America, Electronic music	5.100000e+07	4.277989e+10	5.100000e+07	1

6. 국가별 요약 (상위 **10%**)

- 상위 국가: SV, JO, PR, AR, CO, MX, ID 등
- 특징: 평균 구독자 수 높은 국가일수록 평균 조회수·트렌딩 비율 높음
- 해석: 대형 채널 존재 혹은 글로벌 소비 가능한 콘텐츠 영향
- 주의: 대부분 국가 채널 수 적어 해당 결과를 국가 전체의 일반적인 특성으로 보기 힘듦

국가별 요약 (상위 10):

country	avg_subscribers	avg_views	avg_trending_ratio	count_channels
SV	2.726000e+07	9.507480e+09	2.726000e+07	2
JO	2.285000e+07	1.765409e+10	2.285000e+07	2
PR	1.916571e+07	1.366921e+10	1.916571e+07	7
AR	1.718558e+07	1.086992e+10	1.718558e+07	24
CO	1.359169e+07	9.965571e+09	1.359169e+07	16
KZ	1.319875e+07	7.845021e+09	1.319875e+07	4
MX	1.252779e+07	4.535866e+09	1.252779e+07	58
EC	1.178667e+07	1.987341e+09	1.178667e+07	3
CL	9.837645e+06	4.101842e+09	9.837645e+06	11
ID	9.704382e+06	2.990443e+09	9.704382e+06	76

---

7. 채널별 트렌딩 성과 (상위**10%**)

- 트렌딩 영상 수 많을수록 총 조회수, 총 좋아요 수, 총 댓글 수 증가
- 평균 참여도 (`avg_engagement`): 0.013~0.098

[해석]

- 트렌딩 영상 수 많다고 평균 참여도 항상 높지 않음
- 트렌딩 성과는 양**(**얼마나 자주 트렌딩**) vs** 질**(**얼마나 강한 반응**)** 분리 구조

💡 트렌딩 성과의 양과 질이 반드시 비례하지 않음

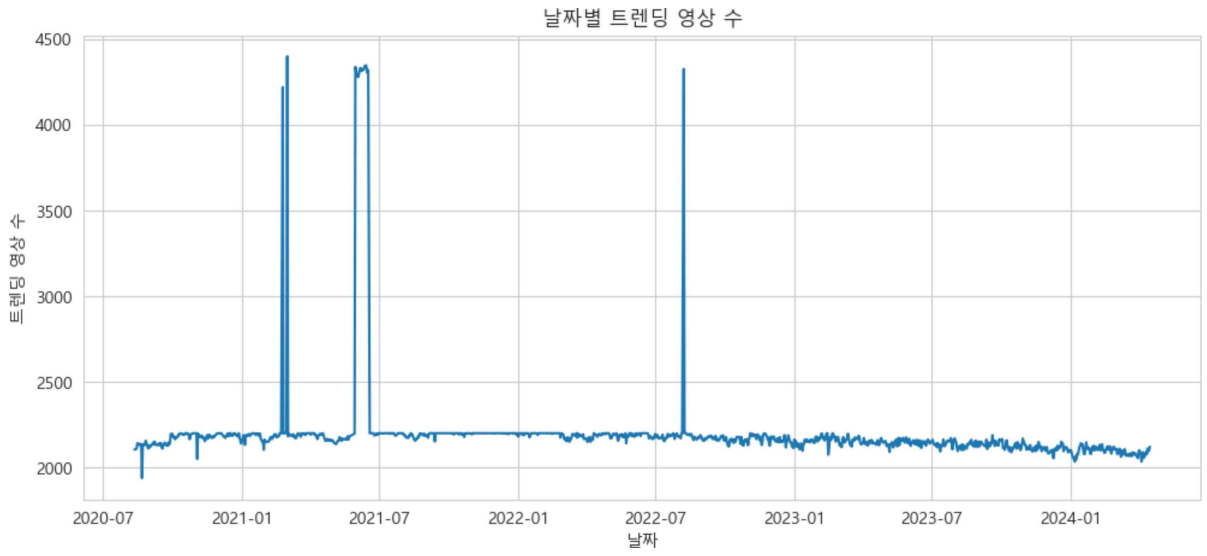
채널별 트렌딩 성과 (상위 10):

channelId	total_trending_videos	total_views	total_likes	total_comments	avg_engagement
UCvrhwpnp2DHYQ1CbXby9ypQ	2073	7401851708	150335927	6472983	0.020427
UCdubelOloxR3wzwJG9x8YqQ	1238	1097746837	38579863	4556839	0.047900
UC6-F5tO8uklgE9Zy8IvbdFw	1156	4892225970	78150393	2212593	0.016892
UCBnxEdpoZwstJqC1yZpOjRA	1082	8569018386	387320256	20189303	0.024103
UC55IWqFLDH1Xp7iu1_xknRA	1073	1817042929	27027882	634023	0.013482
UCNAf1k0yIjyGu3k9BwAg3lg	1070	3609825242	54985412	7259407	0.017181
UCtetJiD_llxdGL4Uh5-Oe2w	994	776202105	18238182	1817565	0.025236
UCjvgGbPPn-FgYeguc5nxG4A	983	11761288222	1097856941	43443568	0.097862
UCgCKagVhzGnZcuP9bSMgMCg	947	2920436249	99176557	8143096	0.037253
UC_IEcnNeHc_bwd92Ber-lew	879	529211945	7550518	2604490	0.019777

---

8. 시계열 분석 (최근트렌딩 추세)

- 전체 기간 트렌딩 영상 수 완만한 감소 추세
- 2020~2021년 초반 비교적 높음
- 2022년 이후 평균 트렌딩 수 소폭 하락, 변동 폭 감소
- 해석: 일반적 콘텐츠 소비 변화라기보다는 데이터 수집 방식, 정책 변화, 이벤트성 트렌딩 영향 가능
- 특징: 유튜브 트렌딩 시스템 안정적, 특정 시점만 비정상적 확대 발생



---

9. 구독자 수규모별 영상 참여도

- X축: 구독자 수 그룹 (소형/중형/대형)

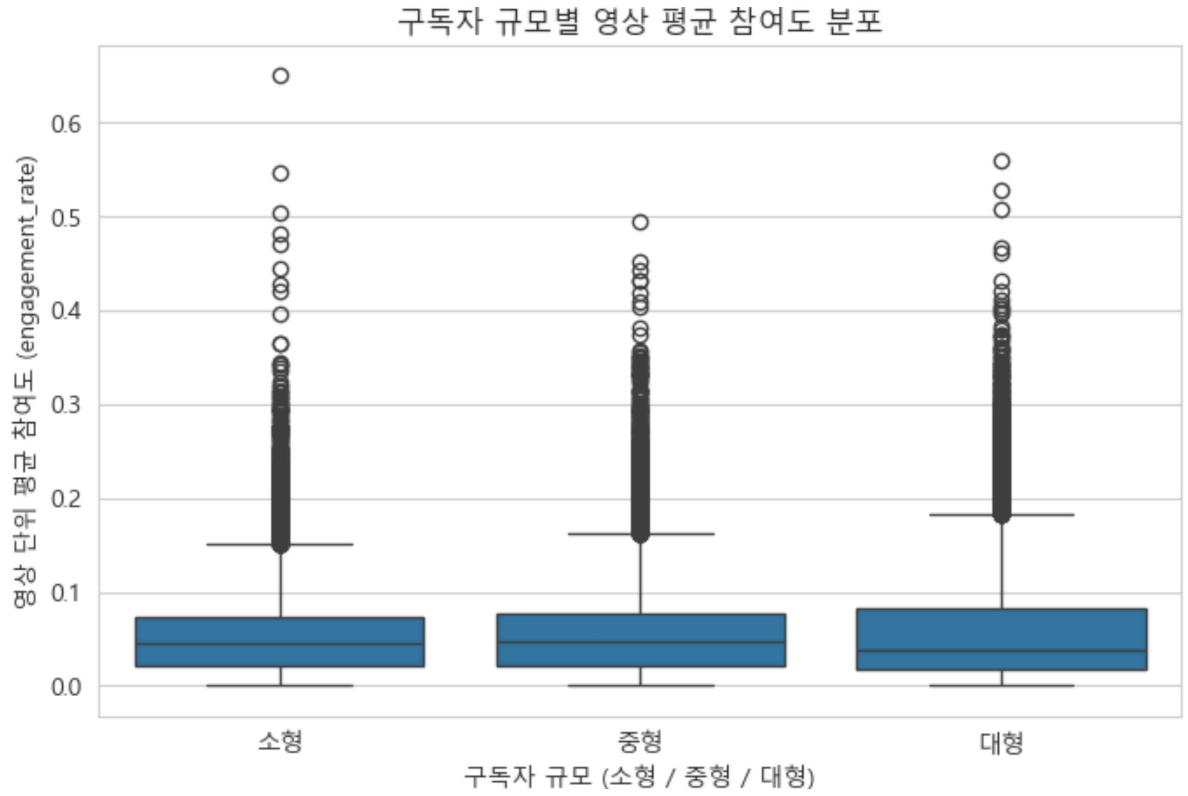
- Y축: 영상 단위 참여도 (`engagement_rate`)

[결과]

- 중앙값: 그룹 간 큰 차이 없음
- 대형 채널이라고 항상 높은 참여도 아님 (구독자 수가 영상 참여도를 보장하지 않음)
- 대형 채널일수록 참여도 분산 큼 (낮은 참여도 영상과 바이럴 영상 동시 존재)

[해석]

- 트렌딩 성과는 빈도**(양) vs** 반응 강도**(질)** 분리, 영상 단위 참여도 지표 함께 고려 필요



---

10. 종합 해석

[영상 단위 성과 지표]

	video_id	view_count	likes	comment_count	engagement_rate	like_rate	comment_rate
0	s9FH4rDMvds	263835	85095	4500	0.339587	0.322531	0.017056
1	jbGRowa5tlk	6000070	714310	31040	0.124224	0.11905	0.005173
2	3EfkCrXKZNs	2296748	39761	0	0.017312	0.017312	0.0
3	gBjox7vn3-g	300510	46222	2748	0.162956	0.153812	0.009144
4	npoUGx7UW7o	327235	22059	2751	0.075817	0.06741	0.008407

[개별 영상 성과 테이블 (트렌딩 성과 분석 시 함께 고려 필요)]

	video_id	channel_id	title	country	trending_days	max_views	avg_engagement_rate	avg_like_rate	avg_comment_rate
0	---Pt7XVuOk	UC3PGilZDRq6wzd7mURXET2g	CHÃ□ REVELAÃ□Ã□O MENINO OU MENINA? *Quarta Gra...	BR	13	520256	0.08809	0.083322	0.004768
1	--0bCF-iK2E	UC6UL29enLNe4mqwTfAyeNuw	Jadon Sancho â□¢ Magical Skills & Goals	GB	5	433340	0.029967	0.027191	0.002776
2	--0fND5o57I	UCvDpfkniGG6WkX5L7DwyGaA	On fait un tour aì□ bois XXXXL de 3 mÃ¯tres !	FR	6	81739	0.069446	0.067337	0.002109
3	--14w5SOEUs	UCGIelM2Dj3zza3xyV3pL3WQ	Migos - Avalanche	CA	51	8034045	0.055263	0.051679	0.003584
4	--1XTdLLpBo	UCcftUjNsimynCrKGFolLRKQ	Darum schlÃ¤gt China Deutschland! BYD Dolphin ...	DE	4	355979	0.022501	0.017965	0.004536

- 유튜브 트렌딩 성과는 소수 채널에 집중되는 강한 롱테일 구조
- 대부분 채널의 규모는 중·소형이며, 트렌딩 경험 제한적
- 상위 채널 다수 영상 반복 트렌딩
  → 누적 조회수, 좋아요, 댓글 압도적
  → 성과가 균등하게 분포되지 않고, 상위 채널에 편중된 구조
- 구독자 수와 영상 참여도 관계: 채널 규모와 무관하게 높은 참여도 영상 존재
- 대형 채널: 영상 참여도의 분산이 높게나타남
  → 낮은 참여도 영상과 바이럴 영상이 동시에 존재하는 경향을 보임
- 카테고리별·국가별 분석: 소수의 대형 채널 영향 강함, 일반적인 특성으로 해석하지 않게 주의
- 트렌딩 성과 구조: 얼마나 자주 트렌딩에 오르는가**(**양**)**와 얼마나 강한 반응을 얻는가**(**질**)**로 분리
  → 트렌딩 성과 분석 시 채널 규모 지표 **+** 영상 단위 참여도·반응 지표 함께 고려 필요