

한국 유튜브 데이터 분석을 통한

Team.밥그릇 👛



최동규 · 김지연 · 우재원 · 조재상 · 최동주



개요

• 주제 선정 이유

02

데이터 정보

- 데이터 소개
- 분석 툴 소개
- 변수 설명

03

주제별 데이터 분석

- 시간대별 트렌드 분석
- 계절별 트렌드 분석
- 연도별 트렌드 분석 1
- 연도별 트렌드 분석 2
- 카테고리별 성장성 분석

04

결론

- 요약 및 결론
- 한계점
- 향후 분석 계획

01 <mark>개요</mark>: 주제 선정 이유

Why?

01 2005년, 유튜브 설립 이후 수많은 크리에이터들이 유튜브 시장에 뛰어들며 높은 진입장벽이 생겼음.

Need.

02 신규 유튜버는 기존 크리에이터들과 경쟁하며, 성공 전략을 가지고 치열한 경쟁에서 생존하고 성장할 방법이 필요함

So, We Do!

누적된 유튜브 트랜딩 데이터를 활용하여 성장하는 분야 및 트랜드를 확인하여 신규 유튜버에게 유의미한 정보를 전달하고자 해당 주제를 선정하였음.

02 데이터 정보



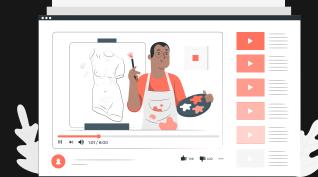
1) 데이터 소개

본 데이터는 캐글에서 제공하는 데이터셋으로, 수개월의 유튜브 트렌드 자료(인기 동영상)를 취합.
 캐글에서는 다양한 국가의 데이터셋을 제공하나, 본 분석과제에서는
 한국의 유튜브 데이터를 분석하여 결론을 도출하였음.

 해당 자료는 유튜브가 일일 인기동영상의 조회수, 댓글, 좋아요 수 등의 자료를 제공하여 사용자와의 상호작용을 측정한 자료가 포함되어 있어 트렌드를 파악하는데 용이함.
 2020년 8월부터 2024년 4월까지의 자료가 수집되어 있어 분석시 유의미한 결과를 도출하리라 기대됨.

2) 분석 툴 소개

- 데이터 전처리 : EXCEL, MYSQL
- 시각화 툴 : EXCEL, 구글 스프레드 시트, Python & Originlab



02 데이터 정보



3) 변수 설명

- 데이터는 video_id, title, publishedAt, channelId, channelTitle, categoryId, trending_date, tags, view_count, likes, dislikes, comment_count, thumbnail_link, comments_disabled, ratings_disabled, description의 총 16개 열로 구성되어 있음
- kaggle에서 데이터를 수집하며 중복 데이터가 발견되어 해당 데이터는 전처리를 거쳐 데이터를 활용함. 또한 각 데이터를 실제로 분석하는 과정에서 팀원별로 사용한 칼럼이 다르며, 전처리 방식에 차이가 있어, 각 주제별 데이터 분석란에 기술하였음.

칼럼	의미	칼럼	의미
Video_id	개별 영상의 고유 이름	Title	영상의 이름
publishedAt	영상 발행 시점	channelld	크리에이터의 이름
categoryld	영상의 카테고리	Trending_date	인기영상 순위에 오른 시점
Tags	영상의 태그	View_count	조회수
Likes	좋아요 수	Comment_count	댓글 수

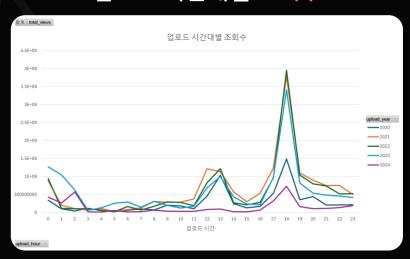
1) 시간대별 트렌드 분석 - 데이터 분석 방법



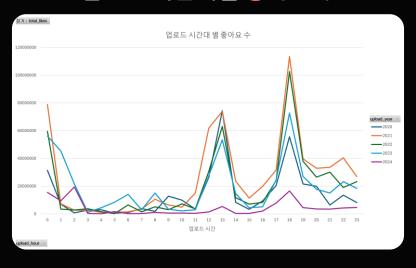
- 분석에 사용한 변수
 publishedAt(업로드 날짜), view_count(조회수), likes(좋아요 수), comment_count(댓글 수)
- 분석 설계
 - 1. 업로드 시간대 별 영상 조회수, 사용자 수 댓글 수를 확인하여 최적의 업로드 시간대를 확인
 - 2. 업로드 요일별 영상 조회수, 사용자 수, 댓글 수를 확인하여 최적의 업로드 요일을 확인
 - 3. 사용자 상호작용이 좋은 업로드 시간대별 영상의 카테고리를 확인
 - 4. 최적의 업로드 시간, 요일, 시간대별 카테고리 추천

1) 시간대별 트렌드 분석 - 데이터 분석 및 시각화

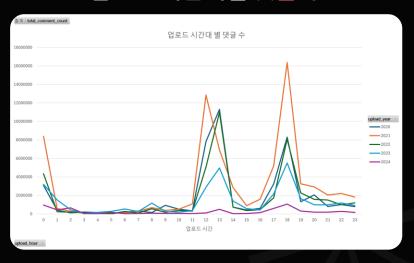
업로드 시간대별 조회수



업로드 시간대별 좋아요 수



업로드 시간대별 댓글 수

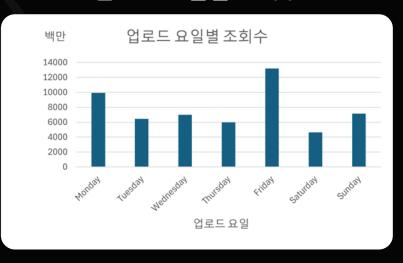


Analysis

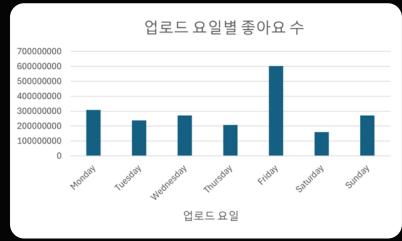
평균적으로 18시에 가장 많은 상호작용이 일어나는 것을 확인할 수 있음 12 -13시에도 상호작용이 활발함

1) 시간대별 트렌드 분석 - 데이터 분석 및 시각화

업로드 요일별 조회수



업로드 요일별 좋아요 수



업로드 요일별 댓글 수

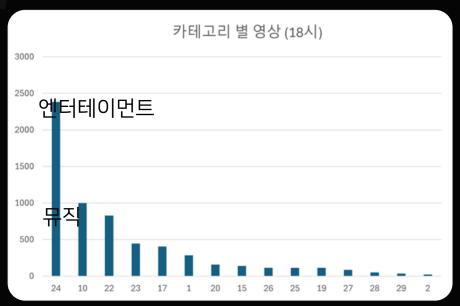


Analysis

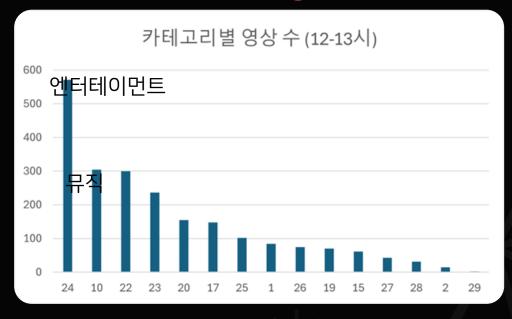
평균적으로 금요일에 가장 많은 상호작용이 일어나는 것을 확인할 수 있음

1) 시간대별 트렌드 분석 - 데이터 분석 및 시각화

업로드 요일별 조회수



업로드 요일별 좋아요 수



Analysis

가장 많은 상호작용이 일어나는 시간대들의 카테고리에는 큰 차이가 보이지 않음 빠른 시간에 소비할 수있는 컨텐츠들이 인기가 많은 것으로 보임

1) 시간대별 트렌드 분석



- 분석 결과
 - 1. 기본적으로 18시, 금요일에 업로드한 영상이 가장 조회수,좋아요 수, 댓글수가 많음
 - 2. 12-13시에 업로드한 영상도 높은 수준의 상호작용이 일어남
 - 3. 퇴근시간, 점심시간, 휴식시간
 - 4. 기분좋고 여유로운 금요일



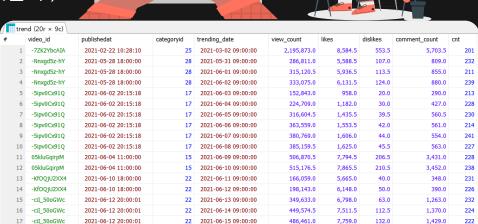
Conclusion

✓ 사람들이 여유를 가지고 유튜브를 이용할 수 있는 시간대를 노리는 것이 중요함 영상 업로드 최적의 시간은 18시, 최적의 요일은 금요일

2) 계절별 트렌드 분석 - 데이터 추출 방법



- 사용 변수 추출
 - 1. video_id(영상 ID)
 - 2. publishedat(업로드 일자)
 - 3. categoryid (카테고리 ID)
 - 4. trending_date(트렌드된 일자)
- 5. view_count(조회수)
- 6. likes(좋아요 수)
- 7. comment_count(댓글 수)
- 트렌드된 일자별 영상 개수 확인
 - Kaggle 데이터 설명에 따르면 트랜드된 일자별로 하루 최대 200개 영상을 수집하였음
 - 중복 제거 : 모든 칼럼의 값이 같을 경우
 - 평균값 : 사용자 상호작용 변수가 다른 경우
 - 전처리 후에도 <u>영상 개수가 200개가 넘는 날이 있었으나,</u> 모두 다른 영상으로 판단되어 최종 데이터셋으로 사용



< 전처리 쿼리 실행 후 결과 >

2021-06-16 09:00:00

524,662.0

553,164.5

1,497.0

2021-06-12 20:00:01

2021-06-12 20:00:01

19 -ciI_50oGWc

2) 계절별 트렌드 분석 - 데이터 시각화



■ 트렌드 기간2021.03.01~2024.02.29

● 계절 구분

■ 봄: 3월, 4월, 5월 가을: 9월, 10월, 11월

■ 여름: 6월, 7월, 8월 겨울: 12월, 1월, 2월

● 분석 변수

■ 조회수 : 가을에 가장 높게 나타났으며, <u>봄에 가장 낮게</u> 나타남

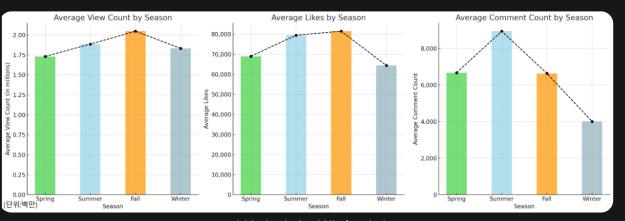
■ 좋아요 : 여름과 가을에 가장 높은 경향을 보였으나, <u>봄과 겨울에 낮게</u>나타남

■ 댓글수 : 여름에 가장 높고

<u>겨울에 가장 낮은</u> 경향을 보임

trending_date (4r × 5c)					
#	season_index	season	avg_view	avg_likes	avg_comment
1	1	spring	1,729,930	68,984	6,670
2	2	summer	1,885,067	79,450	8,941
3	3	fall	2,046,058	81,457	6,634
4	4	winter	1,831,305	64,405	4,003

< 계절별 상호작용 변수 평균값 >



< 전처리 쿼리 실행 후 결과 >

2) 계절별 트렌드 분석



- 분석 결과
 - 1. 전반적으로 여름과 가을에는 사용자 상호작용 변수인 조회수, 좋아요, 댓글수가 높게 나타나며, 봄과 겨울에는 낮게 나타남
 - 야외 활동량이 높고 즐길 컨텐츠가 많은 여름과 가을에 여러 정보들을 얻기 위해 영상을 보고, 댓글을 작성하면서 자신이 궁금한 추가 정보를 탐색하면서 조회수 및 댓글수가 높게 나타났을 것으로 예상됨.

Conclusion

✓ 신규 유튜버의 영상이 트랜드 영상에 올라가기 위해서는 사용자 상호작용 변수의 평균이 높은 여름 또는 가을에 업로드하는 것이 유리할 것으로 판단됨

3) 연도별 트렌드 분석 1 - 콘텐츠 발행과 조회수 분석



- 변수 선정
 - 1. publishedat(업로드 일자)
 - 2. categoryid (카테고리 ID)
 - 3. View_count(조회수)
- 데이터 전처리
 - PublishedAt 기준으로 연도별로 정리
 - 중복된 데이터의 경우 Tranding_date 기준으로 가장 최신의 날짜를 골라 사용
 - PublishedAt 기준으로 매년 1월 부터 4월 까지의 데이터를 추출 하여 사용

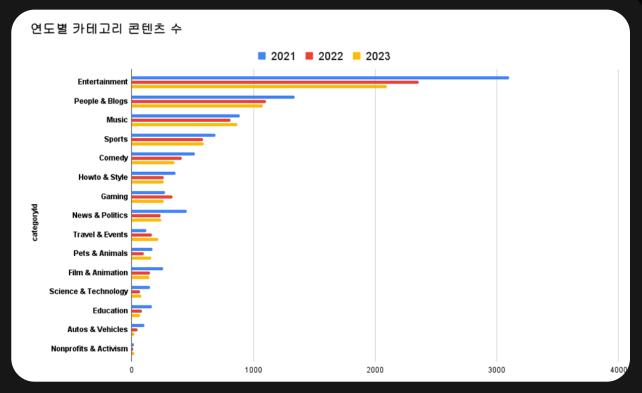
	publishedA	t	categoryId	category_cnt
•	2021-01-01	14:00:00	1	96
	2021-01-01	00:15:00	2	48
	2021-01-01	00:25:00	10	304
	2021-01-01	18:00:00	15	63
	2021-01-02	10:00:00	17	247
	2021-01-07	20:00:00	19	33
	2021-01-02	10:00:00	20	82
	2021-01-01	15:02:00	22	429
	2021-01-01	11:00:00	23	247
	2021-01-01	00:00:00	24	1139
	2021-01-03	00:33:00	25	175
	2021-01-01	18:44:00	26	122
	2021-01-01	23:10:00	27	77
	2021-01-01	18:00:00	28	53
	2021-01-01	00:22:00	29	5

< 전처리 쿼리 실행 후 결과 >

3) 연도별 트렌드 분석 1 - 콘텐츠 발행과 조회수 분석

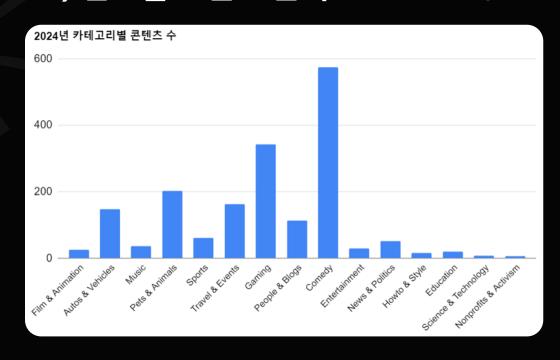


- 1년 단위 카테고리 분석
 - Entertainment, people & Blogs,
 Music 순서로 콘텐츠가 가장 많음.
 - Entertainment 카테고리가 압도적 으로 콘텐츠 수가 많으나 21년도부터 시간이 지날수록 점점 감소하는 경향 을 보이고 있음.

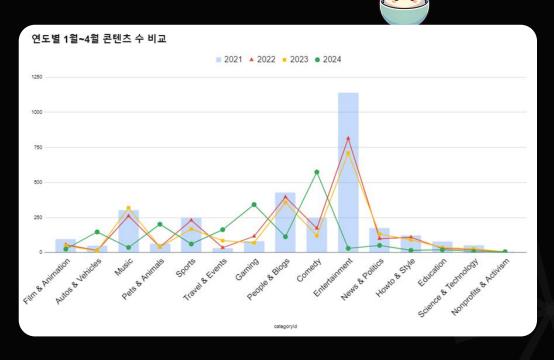


< 연도별 카테고리 콘텐츠 수 >

3) 연도별 트렌드 분석 1 - 콘텐츠 발행과 조회수 분석



 2024년은 비록 4개월간의 데이터만 있지만, 이전 연도와 다른 양상을 보임. 이를 확인하기 위해 매년 1월부터 4월까지의 데이터를 추출해 비교



• 비교 결과 2024년에만 발생하는 현상임을 확인

3) 연도별 트렌드 분석 1 - 콘텐츠 발행과 조회수 분석



- 분석 결과
 - 1. 21년도의 인기 카테고리들이 시간이 지날수록 하향하는 추세를 보임.
 - 2. 21~23년도 사이의 1~4월 카테고리별 콘텐츠 수를 비교해 보았을 때 24년도1~4월의 카테고리별 콘텐츠 수와 정반대의 양상을 띄고 있어 트렌드가 변화하고 있다고 해석할 수 있음
 - 3. 트렌드의 큰 변화가 24년에만 나타나는 경향을 보이며 해당 변화가 일시적인지 지켜볼 필요가 있음

Conclusion

✓ 각 연도별 1~4월까지의 데이터만 모아 비교하는 것은 위험한 접근일 수 있으나 비교적 가장 안정적인 성장폭을 보였던 Travel & Events를 추천.

3) 연도별 트렌드 분석 2 - 조회수와 카테고리 내부 성장률 분석



- 데이터 전처리
 - video_id, tags, view_count, publish, categoryld 열을 csv로 저장/활용.
- 데이터 추출
 - mysql을 활용하여 조회수를 기준으로 카테고리별 순위 추출.



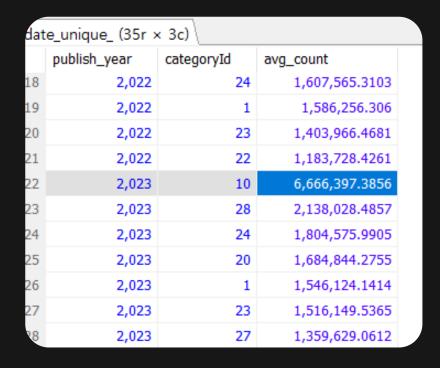


```
WITH RankedCategories AS (
    SELECT
        publish_year,
        categoryId,
        avg(view_count) AS avg_count,
        ROW_NUMBER() OVER (PARTITION BY publish year ORDER BY avg(view count) DESC) AS rn
    FROM
        date_unique_
    WHERE
        publish year BETWEEN 2020 AND 2024
    GROUP BY
        publish_year, categoryId
SELECT
    publish_year,
    categoryId,
    avg_count
FROM
    RankedCategories
WHERE
    rn <= 7
ORDER BY
    publish year ASC, avg count DESC;
```

3) 연도별 트렌드 분석 2 - 조회수와 카테고리 내부 성장률 분석



A. 데이터 추출 및 분석



B. 연도별 조회수 상위 5개 카테고리.

평균 : 평균 조회수	열 레이블 🧊				
행 레이블 🧊	Entertainment	Film & Animation	Gaming	Music	Science & Technology
2021	1,730,773	1,267,271	1,387,264	6,396,663	1,597,244
2022	1,607,565	1,586,256	1,853,393	5,976,133	3,257,385
2023	1,804,576	1,546,124	1,684,844	6,666,397	2,138,028
평균 : 평균 조회수	1,714,305	1,466,551	1,641,834	6,346,398	2,330,886

Music, Science & Technology, Entertainment 순서대로 평균 조회수가 높음.

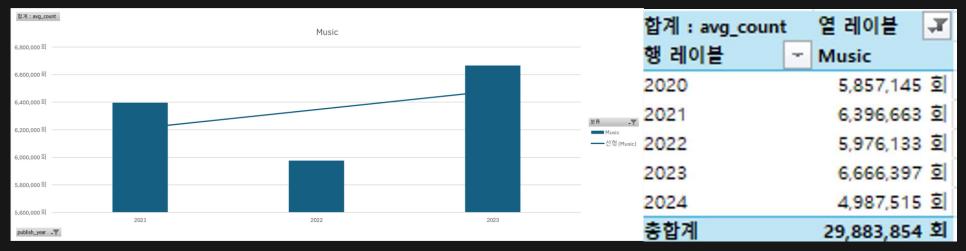
C. 연도별 조회수 상위 5개 카테고리.

3) 연도별 트렌드 분석 2 - 조회수와 카테고리 내부 성장률 분석



C. 카테고리별 성장 추이 분석

a. music

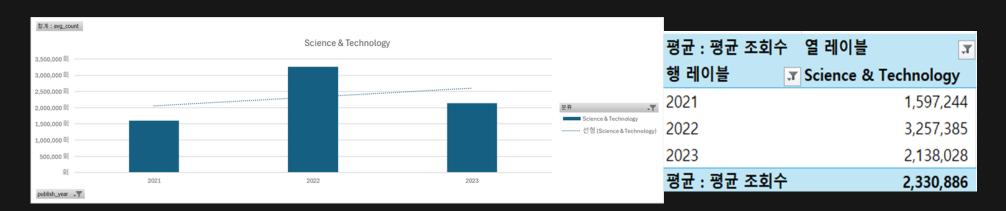


2024 예측 조회수	성장률	
6616132.085		2.09%

3) 연도별 트렌드 분석 2 - 조회수와 카테고리 내부 성장률 분석



- C. 카테고리별 성장 추이 분석
 - b. Science & Technology



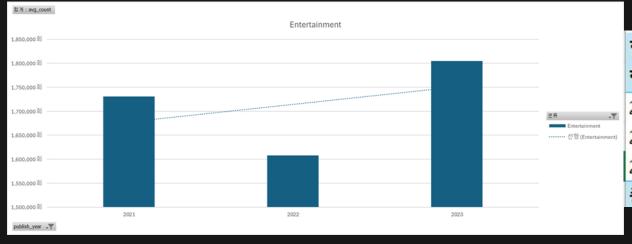
2024년 예측치	성장률	
2,871,670		16%

3) 연도별 트렌드 분석 2 - 조회수와 카테고리 내부 성장률 분석



C. 카테고리별 성장 추이 분석

c. Entertainment



합계 : avg_count	열 레이블 🛒
행 레이블	▼ Entertainment
2021	1,730,773 회
2022	1,607,565 회
2023	1,804,576 회
총합계	5,142,914 회

2024년 예측치	성장률	
1,788,108	2.11%	

3) 연도별 트렌드 분석 2 - 조회수와 카테고리 내부 성장률 분석



- 분석 결과
 - 1. 조회수 기준 상위 5개 카테고리를 나열함으로써 한국 유튜브 생태계에서 가장 조회수가 잘 나오 는 카테고리를 다음 순서로 제안할 수 있음. (Music, Science & Technology, Entertainment, Gaming, Film & Animation)
- 분석의 한계
 - 1. 파워 크리에이터에 의한 왜곡 발생 우려
 - 2. 데이터 수집 기간이 짧아 추세분석에 한계가 있음
 - 3. 현재 성장중인 카테고리에 대한 분석을 하지 못하였음.
 - 4. 어떤 세부 테마로 유튜브를 시작할지 제안하기 어려움.

Conclusion

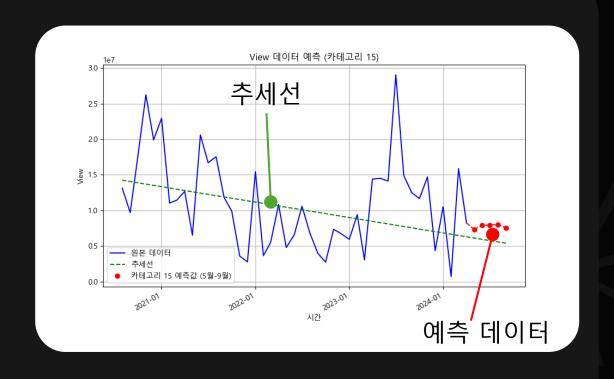
✓ 상위 5개 카테고리에서 <mark>성장세가 가장 높은 Science & Technology</mark> 카테고리에서 신규로 유튜브를 시작하도록 제안할 수 있음.

- 데이터 전처리 방법 및 변수 선택
 - 모든 데이터를 카테고리별로 분류함.
 - Trending date를 기준으로 월 별로 수집함.
 - 중복 데이터의 경우 가장 마지막의 데이터를 기준으로 정리함.
 - 2020년 8월 부터 2024년 4월 까지의 데이터를 수집.
 - 변수는 조회수, 좋아요 수, 댓글 수 3가지로 선택함.





- 추세선과 예측값 계산 방법
 - 추세선의 경우 데이터의 선형적인 변동을 분석함.
 - Linear regression 모델로 진행함.
 - 예측값의 경우 선형적인 예측에 비선형적인 분포를 합쳐 분석함.
 - Linear regression과 Random forest 두 모델의
 앙상블 모델을 사용함.
 - 2024년 5월 부터 9월까지의 값을 예측함.

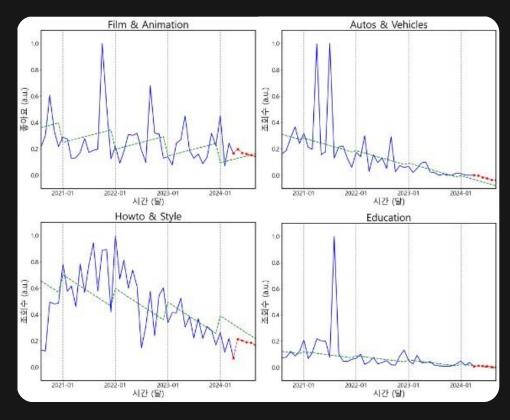


- 성장성 분석 결과
 - 총 카테고리 수는 15 개
 - 우하향 추세를 보이는 카테고리는 11 개
 : Film & Animation, Auotos & Vehicles, Music,
 Pets & Animals, Sports, Comedy, News &
 Politics, Howto & Style, Education, Science &
 Technology, Nonprofits & Activism
 - 우상향 추세를 보이는 카테고리는 4 개
 : Travel & Events, Gaming, People & Blogs,
 Entertainments



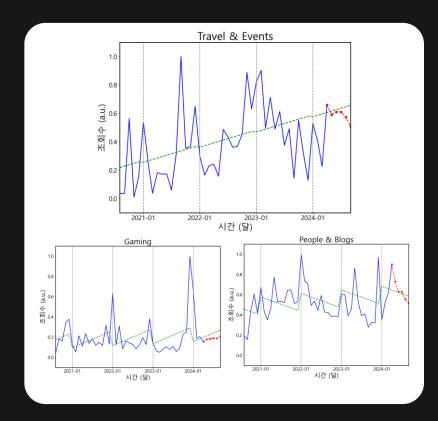


- 성장성 분석 결과 (우하향 예시)
 - Film & Animation, Music 등의 카테고리는 방대한 양의 콘텐츠가 과포화 되어, 개별 콘텐츠가 주목받기어려울 가능성이 있음.
 - Autos & Vehicles 등의 카테고리는 단 기간에 인기를
 끌던 주제들이 관심을 잃어 점차 관심이 감소할 수 있음.
 - Howto & Style, Education 등의 카테고리는 더 창의 적이고 혁신적인 콘텐츠를 제공하는 타 카테고리로 인해 상대적으로 인기가 줄어들 가능성 존재.



- 성장성 분석 결과 (우상향 예시)
 - Travel & Events의 경우 2021년과 2022년에 코로나 해제와 맞물려 전체적인 시청자의 수가 증가함을 확인함.
 - Gaming의 경우 성장하기는 하나 다소 완만한 성장세임,
 이는 특정 게임에 콘텐츠가 집중되고 이미 포화 상태에 들어간 것으로 추측함.
 - People & Blogs는 지속적으로 성장하고 있으나 속도가 더딘 편, 매우 광범위한 콘텐츠를 포함하고 있어, 영상의 주제나 스타일이 통일되지 않고 조회수가 다양하게 분산되는 것을 원인으로 추측함.





3) 카테고리별 성장성 분석



- 분석 결과
 - 1. 여행 및 이벤트 관련 콘텐츠는 지속적으로 성장 가능성이 높음.
 - 코로나가 끝난 이후 다양한 여행지 소개, 이벤트 현장 스케치, 여행 팁 등 많은 사람들이 관심을 가 질 것으로 예상됨.



Conclusion

✓ 성공 가능성이 높은 카테고리를 고른다면 Travel & Events 를 추천함.

04 결론



1) 전반적인 요약 및 결론

- ① 본 보고서는 한국 유튜브 데이터 분석을 통한 신규 유튜버 성공 전략을 돕고자 다각적으로 데이터 분석을 하여 결론을 도출하였음.
- ② 시계열적 분석을 통하여 여름과 가을에 영상을 올리는 것이 조회수, 좋아요와 유의미한 상관관계를 가지고 있었으며, 구체적 시간은 12-13시 또는 18시에 가장 사용자와의 상 호작용이 활발하였음.
- ③ 전체적으로 카테고리에서 Gaming, Travel & Event, Music, Science & Technology가 컨텐츠의 수와 조회수의 측면에서 상위권에 있음을 확인하였으며, 마지막으로 회귀분석과 랜덤포레스트 모델을 조합하여 각 카테고리별 추세를 분석했을 때, 가장 성장세가 두드러진 Travel & Event 카테고리를 추천하게 되었음.

04 결론

2) 한계점

- ① 제안된 카테고리 내 세부적인 테마를 제안하기 어려움
- ② 데이터 외부의 요인에 의해 변동하는 이상치를 설명하기 어려움
- ③ Trending_date기준 영상의 누적 조회수 등 자료를 수집하여 전체적인 유튜브의 동향 파악이 어려움



3) 향후 분석 계획

- ① 기존 분석 내용에 더해 카테고리 내 태그를 활용하거나, 제목에 대한 분석 등을 활용해 트렌드를 구체적으로 제안.
- ② 데이터에 대한 원인 분석을 명확히 하기 위해 계절별 시간대별 분석 결과에 따른 요인에 대한 분석이 필요.
- ③ 본 보고서의 결과와 실제 크리에이터의 구독자수, 조회수 등의 변화를 비교하여 결론에서 도출한 결과값이 2024년 4월 이후의 실제 데이터와 유사한지 비교/평가 분석이 필요.

