2024년 유튜브 댓글과 주요 변수의 상관관계 분석

최민준

목차

- 연구 배경 및 가설
- 데이터 수집 및 전처리
- 전체 데이터 개요
- 데이터 분석
- 결론
- 한계
- 추가 탐구

연구 배경 및 질문

- 유튜브에서 선플과 악플이 조회수와 관련이 있을까?
- 핵심 질문:

1. 조회수가 높은 영상일수록 악플 비율이 높을까, 선플 비율이 높을까?

2. 카테고리별로 조회수와 선플/악플의 상관관계는 어떻게 될까?

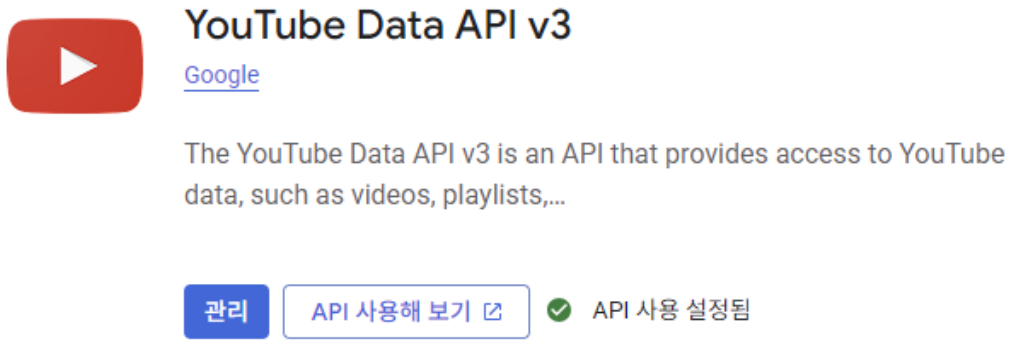
3. 영상 길이, 제목 길이와 조회수, 좋아요 수, 댓글 수 사이의 상관관계가 존재할까?

데이터 수집 및 전처리

```
# 한국 기준
REGION_CODE = "KR"

# 검색 키워드
QUERIES = [
    "음악", "kpop", "뮤직비디오",
    "게임", "롤", "배그", "minecraft",
    "뷰티", "메이크업", "패션",
    "공부", "수능", "강의", "과학",
    "축구", "야구", "스포츠",
    "예능", "코미디", "먹방", "시사",
    "뉴스", "기사", "언론"
]
```

- Google Cloud에서 제공하는 Youtube Data API v3 활용
- 검색만 가능한 API의 한계점을 여러 Queries를 둠으로써 해결
- 총 30개의 카테고리당 최대 100개의 영상 수집-> 조건(삭제/비공개/댓글 비활성화/채널 제한)을 만족하는 상위 1000개의 영상만 샘플링-> 각 영상당 최대 500개 댓글 샘플링
- 긍정적인 단어, 부정적인 단어 임의로 지정-> 해당 단어의 개수로 선플/악플 판단
- 데이터의 다양성을 위해 채널당 최대 15개의 영상만 포함되도록 설정
- 영상이 10개 이하인 카테고리는 제외



데이터 수집 및 전처리

- ID: video_id
- 명목형 변수: category_name, channel_id, channel_title, is_short
- 수치형 변수: view_count, like_count, comment_count, duration_sec, ratio_positive, ratio_negative, ratio_neutral, like_view_ratio(like_count/view_count), comment_view_ratio(comment_count/view_count)
- duration_raw -> duration_sec(초 단위)로 처리
- is_short: 60초 이하-> 1, 아니면 0
- log_view_count = log_{10} (view_count), log_comment_count = log_{10} (comment_count), log_like_count = log_{10} (like_count) (숫자를 작게 만들어 분석을 편하게 하기 위함)

데이터 전처리-감성 분석

- 감성 분석 시행

➔ n_positive/n_negative/n_neutral/ratio_positive/ratio_negative/ratio_neutral

- pos_count = 긍정 단어 개수, neg_count = 부정/욕설 단어 개수
- sentiment_score = pos_count – neg_count > 0 : positive / < 0 : negative / = 0 : neutral

```
# 긍정 단어
positive_ko = [
    "좋다", "좋아요", "감동", "감사", "행복", "기쁨", "웃긴", "웃기다",
    "예쁨", "예쁘다", "예쁘네요", "아름답네요", "아름답다",
    "재밌다", "재미있다", "재미있어요", "최고", "대박", "귀엽다", "멋있다",
    "멋있어요", "멋있네요", "멋져요",
    "사랑", "고맙다", "힐링", "감사합니다", "응원", "축하", "즐겁다",
    "예뻐요", "이쁘다", "이뻐요", "이쁘네요", "잘",
    "웃음", "감격", "감탄", "존경", "대단하다",
    "귀여워", "ㄱㅇㅇ", "귀여워요", "굿", "재밌어요", "재밌당", "즐겁당",
    "힘내세요", "화이팅", "ㅎㅇㅌ", "고마워요",
    "재밌음", "재밌네", "재밌네요", "재밌넹",
    "존잼", "꿀잼", "짱이야", "짱이다", "최고다", "최곱니다",
    "완전좋아요", "완전좋다",
    "존좋", "마음이따뜻해짐", "따뜻", "미쳤다", "고트"
]
```

```
positive_en = [
    "good", "great", "awesome", "amazing", "nice", "cool", "love",
    "funny", "hilarious", "happy", "beautiful", "wonderful", "perfect",
    "respect", "fantastic", "brilliant", "thank", "thanks", "loving",
    "cute", "enjoy", "cutie", "adorable", "wholesome",
    "lit", "dope", "fire",
    "insane", "legend", "goat", "wellmade", "welldone"
]
```

	video_id	title	published_at	duration_raw	duration_sec	view_count	like_count	comment_count	category_id
1	ekr2nlex040	ROSÉ & Bruno Mars - APT. (Official Music Video)	2024-10-18 04:00:07+00:00	PT2M54S	174.0	2176762509	17277441.0	821881.0	10
2	hUnb-mclq6k	ToRung comedy: 😂magic box📦	2024-10-10 12:12:53+00:00	PT1M	60.0	1016388151	33439342.0	43678.0	23
3	9rVBlgGFaD8	싸움이 났을 때 이 방법을 써보세요 | 달님이 | 시즌2 | 노래 | 율동 | 키즈 | 뮤지컬 | 반짝반짝	2024-06-19 08:00:05+00:00	PT1M	60.0	696912997	8596836.0	0.0	1
4	l9mw5UIDyPI	This Game Is Wild...	2024-11-12 17:00:00+00:00	PT19S	19.0	660856676	13278406.0	8650.0	24
5	3cWm9B_0_kI	How Many People To Stop Ronaldo?	2024-12-27 19:00:00+00:00	PT21S	21.0	518918612	11770078.0	27308.0	24

category_name	channel_id	channel_title	hour	weekday	is_weekend	is_short	like_view_ratio	comment_view_ratio
Music	UCBo1hnzxV9rz3WVsv__Rn1g	ROSÉ	4	4	0	0	0.007937219117182067	0.00037757035808999226
Comedy	UCXbYIU08sOTBktOtjVsvR6w	ToRung	12	3	0	1	0.0329001690614947	4.2973739862105104e-05
Film & Animation	UCXB3QDMncTKZFsGrd11zviw	반짝반짝 달님이 Dalimi_Animation	8	2	0	1	0.01233559430948595	0.0
Entertainment	UCX6OQ3DkcsbYNE6H8uQQuVA	MrBeast	17	1	0	1	0.020092716745135825	1.3089071071137973e-05
Entertainment	UCX6OQ3DkcsbYNE6H8uQQuVA	MrBeast	19	4	0	1	0.022681934561252545	5.2624822792056645e-05

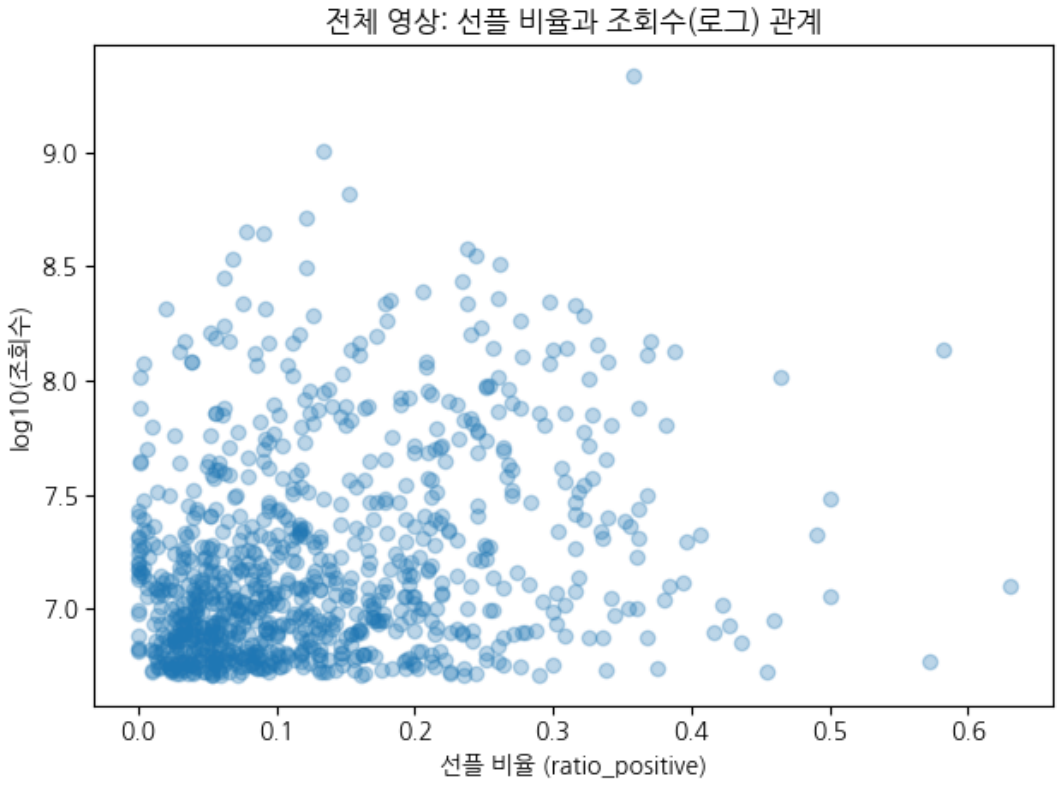
n_positive	n_negative	n_neutral	n_comments_sampled	ratio_positive	ratio_negative	ratio_neutral
179	9	312	500	0.358	0.018	0.624
67	0	433	500	0.134	0.0	0.866
0	0	0	0	0.0	0.0	0.0
76	6	418	500	0.152	0.012	0.836
61	3	436	500	0.122	0.006	0.872

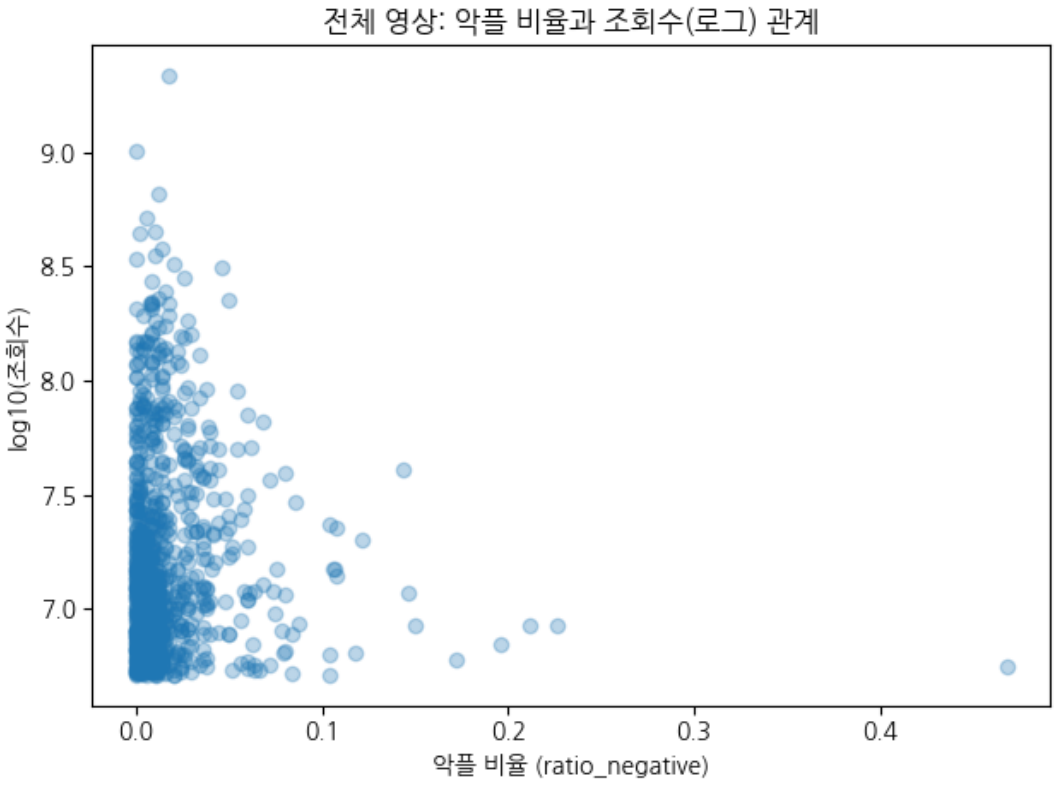
데이터 개요

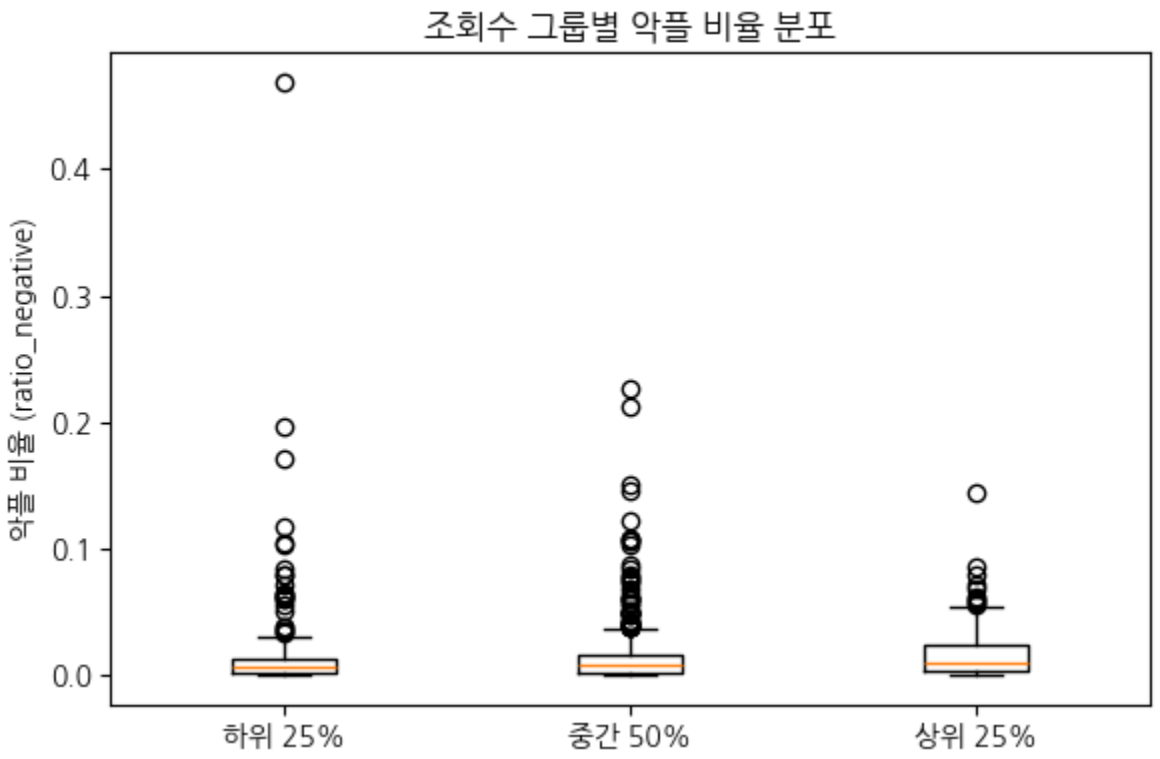
- 영상별 감성 통계 개수: 938개 (62개 영상은 감성 댓글이 없어 빠짐)
- 카테고리: 10개 (영상 10개 이하인 카테고리는 제외)
- 평균 선플 비율: 약 12.74%
- 평균 악플 비율: 약 1.56%
- 샘플링된 댓글 수: 428,784개
- 카테고리별 비교-> 10개 카테고리, 927개 영상으로 비교
- 전체 영상 비교-> 938개 영상으로 비교

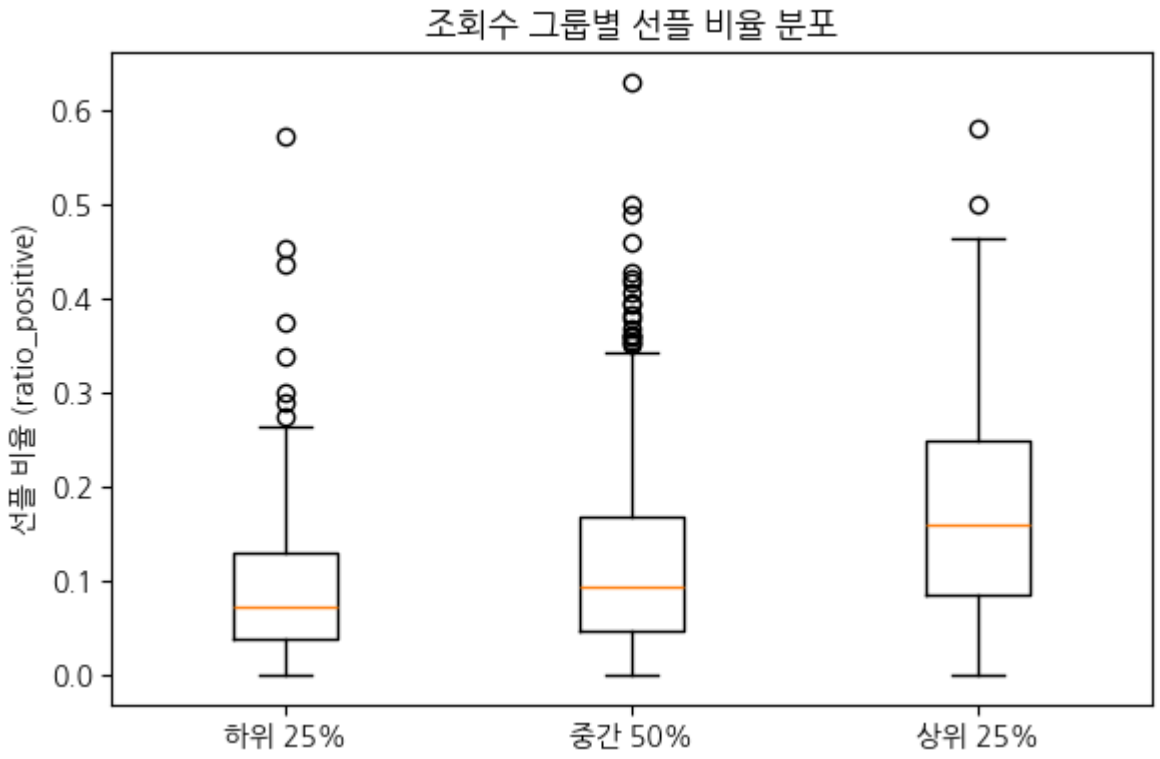
...	n_videos
category_name	
Music	94
Comedy	47
Science & Technology	8
People & Blogs	194
Entertainment	240
Education	21
Gaming	122
Autos & Vehicles	1
Film & Animation	22
Howto & Style	43
Sports	123
Pets & Animals	1
Travel & Events	1
News & Politics	21

데이터 분석
-전체 영상의 선플/악플과 조회수 관계

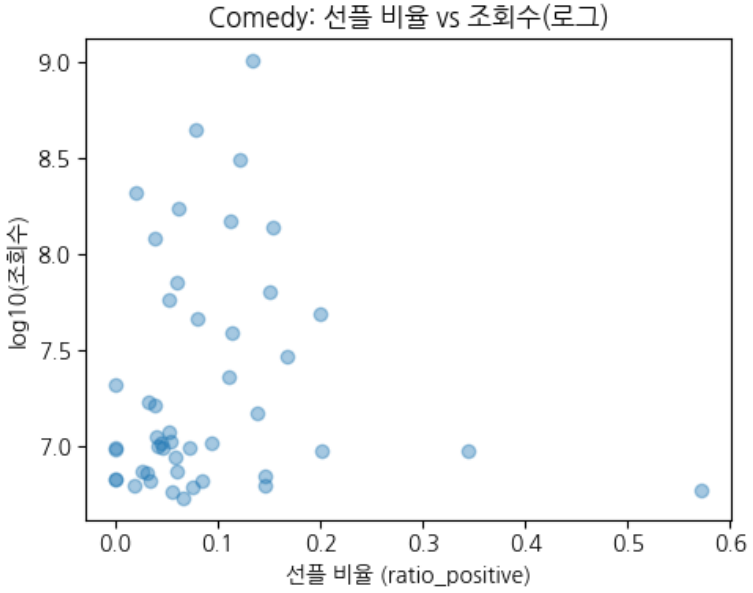


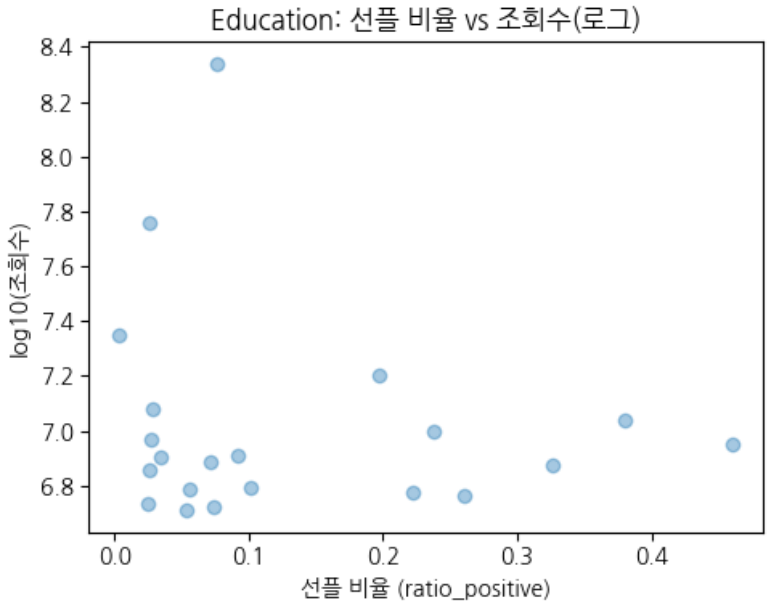


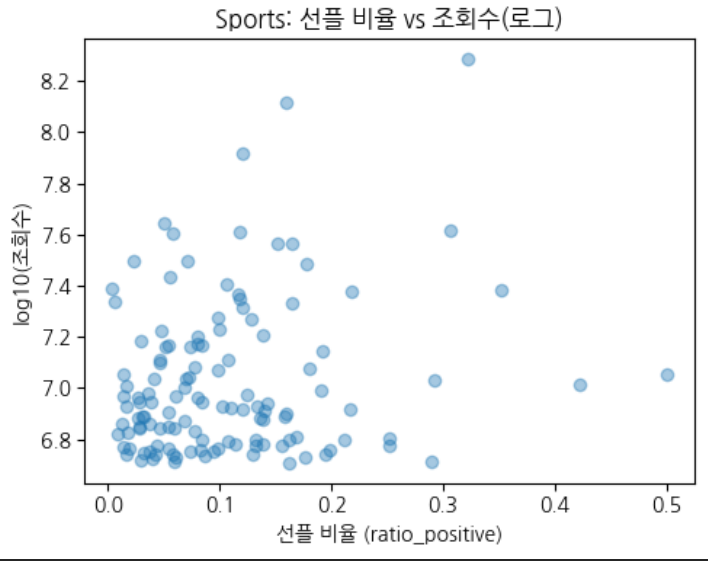

조회수 그룹별 악플 비율 분포
악플 비율 (ratio_negative)
0.0
0.1
0.2
0.3
0.4
하위 25%
중간 50%
상위 25%

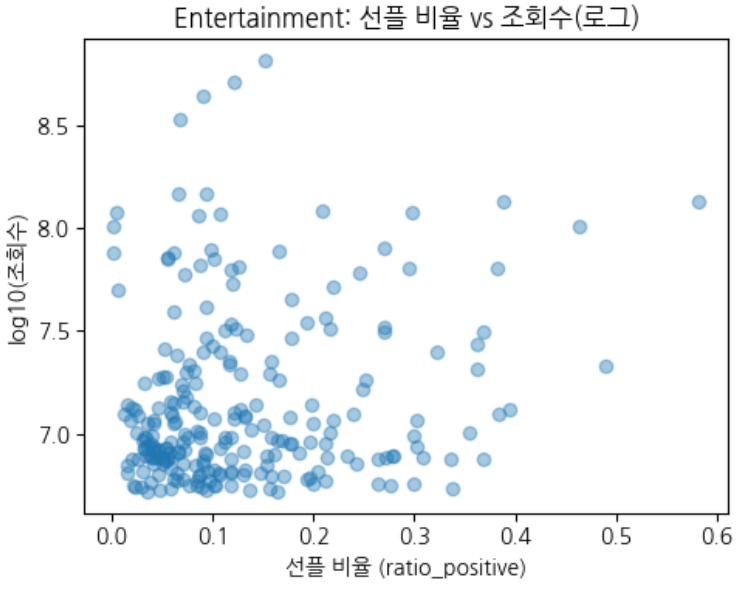

조회수 그룹별 선플 비율 분포
선플 비율 (ratio_positive)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
하위 25%
중간 50%
상위 25%

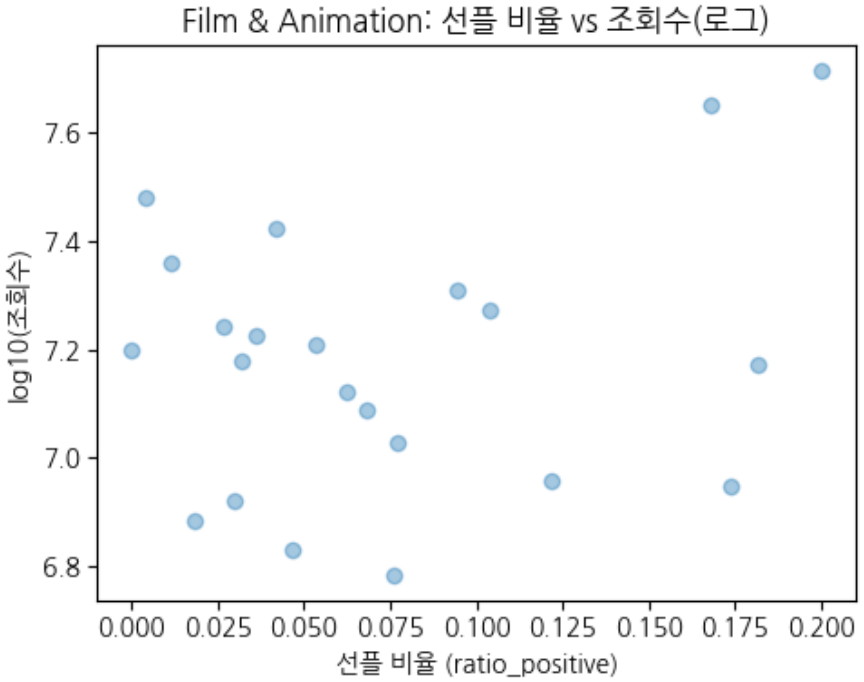
데이터 분석
-카테고리별 선플과 조회수 관계

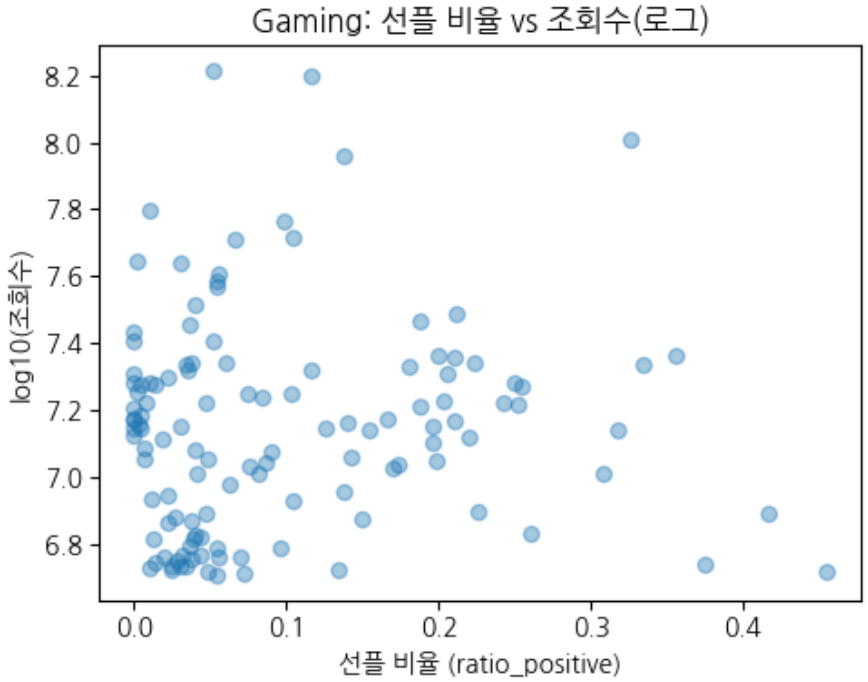


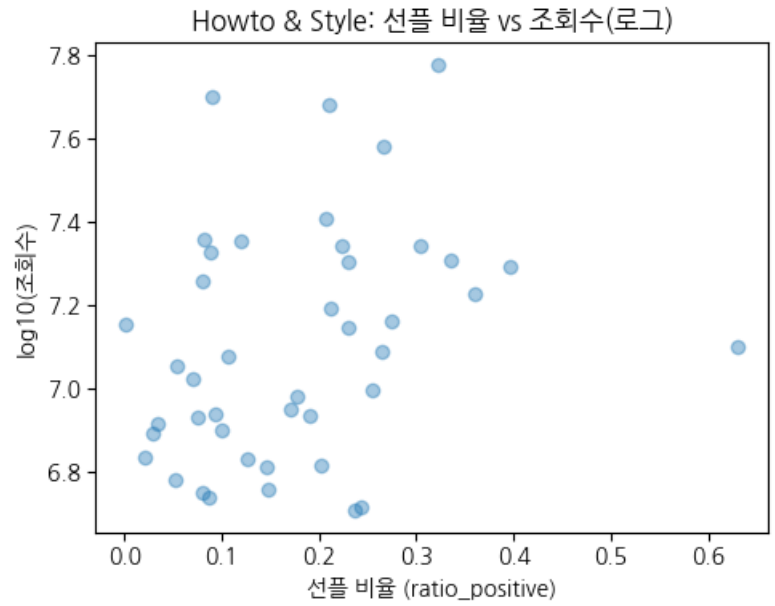


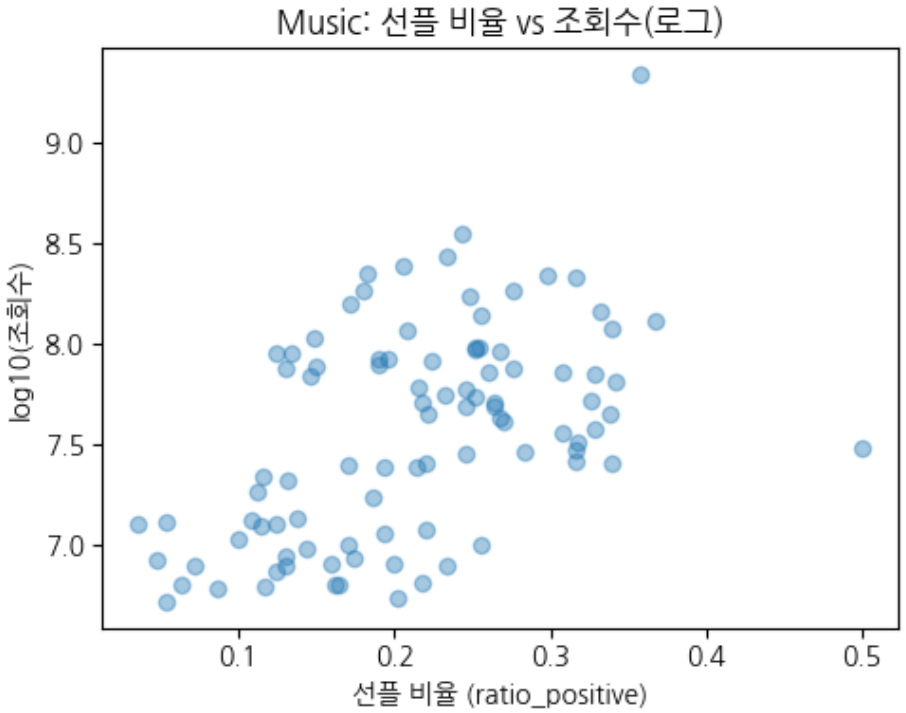


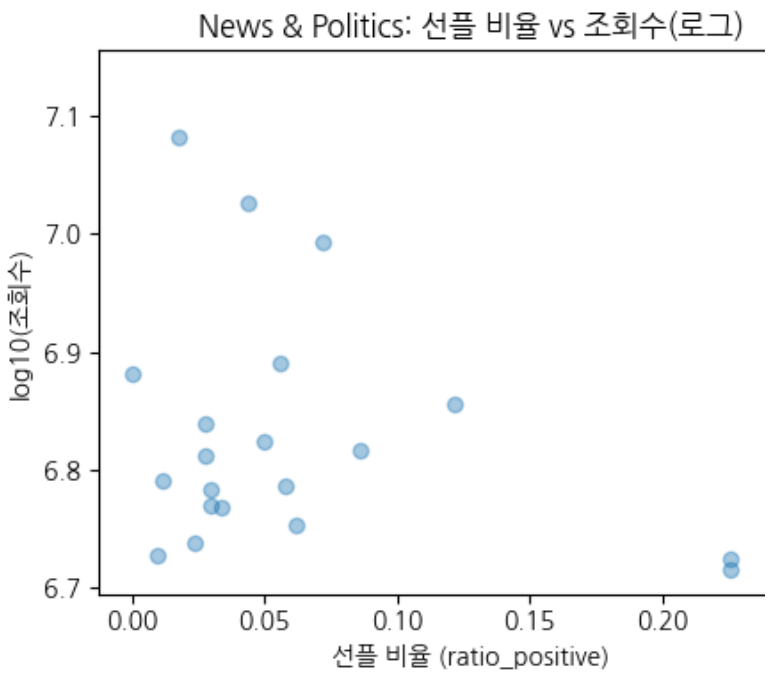


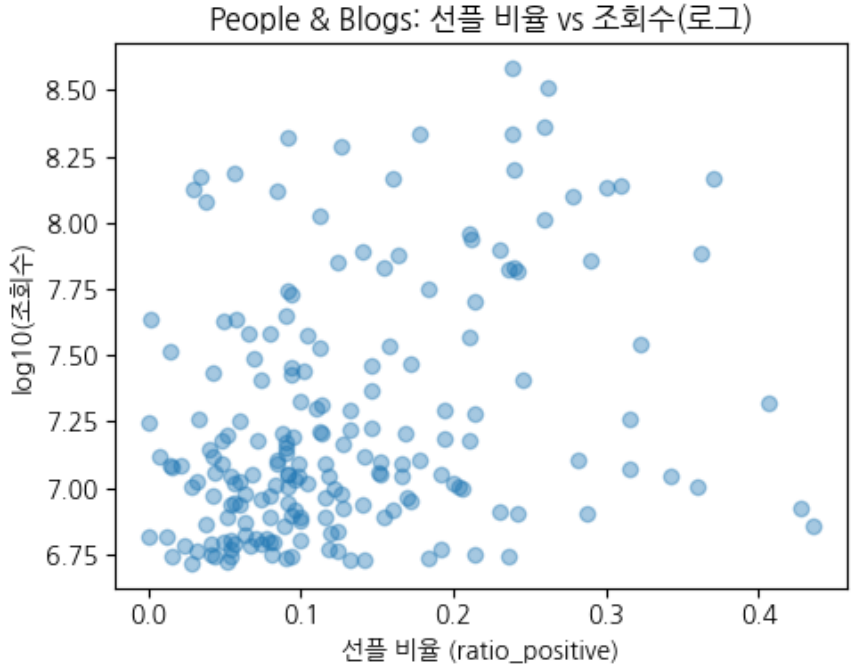

Film & Animation: 선플 비율 vs 조회수(로그)
log10(조회수)
선플 비율 (ratio_positive)


Gaming: 선플 비율 vs 조회수(로그)
log10(조회수)
선플 비율 (ratio_positive)

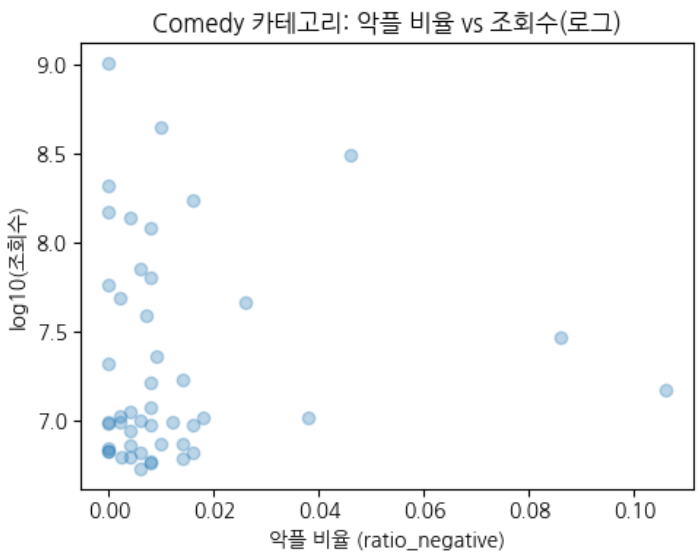

Howto & Style: 선플 비율 vs 조회수(로그)
log10(조회수)
선플 비율 (ratio_positive)

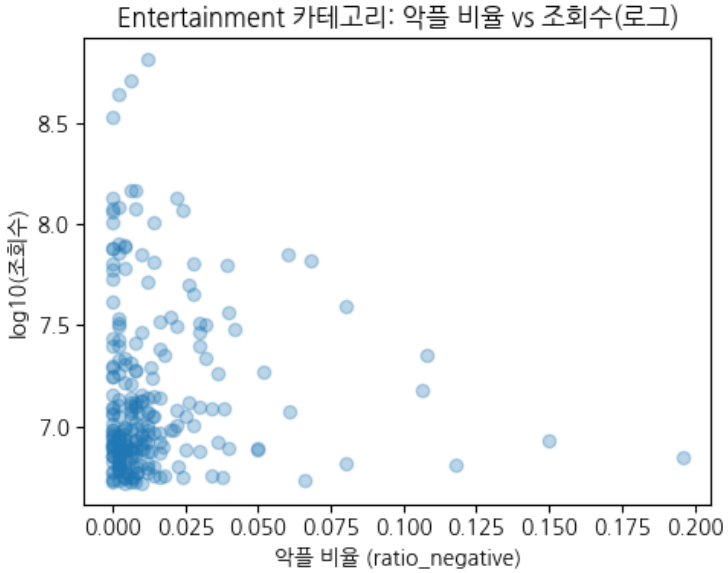

Music: 선플 비율 vs 조회수(로그)
log10(조회수)
선플 비율 (ratio_positive)

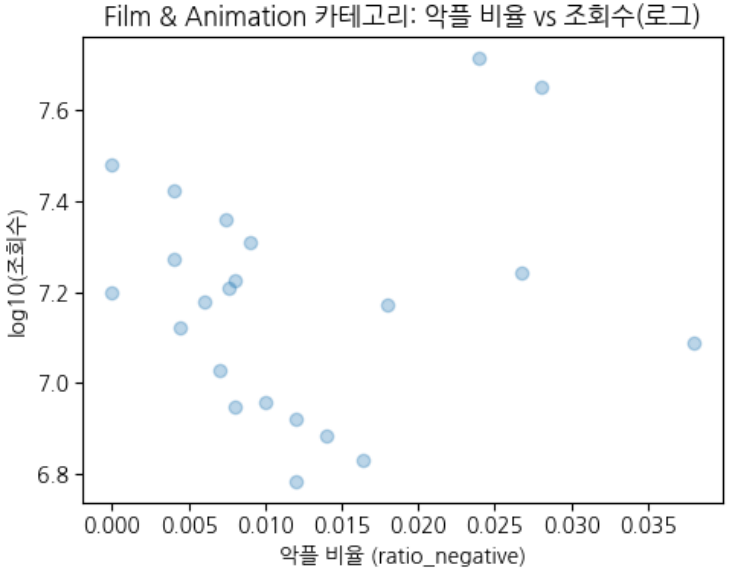

News & Politics: 선플 비율 vs 조회수(로그)
log10(조회수)
선플 비율 (ratio_positive)

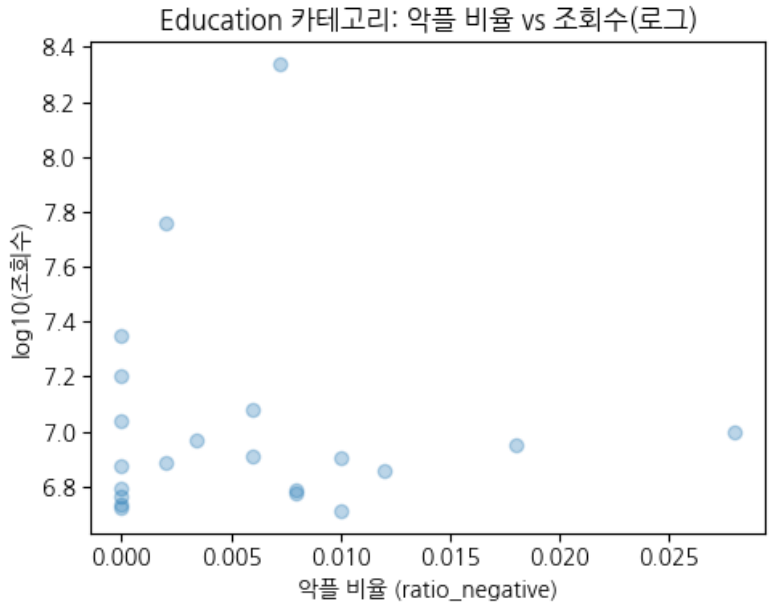

People & Blogs: 선플 비율 vs 조회수(로그)
log10(조회수)
선플 비율 (ratio_positive)

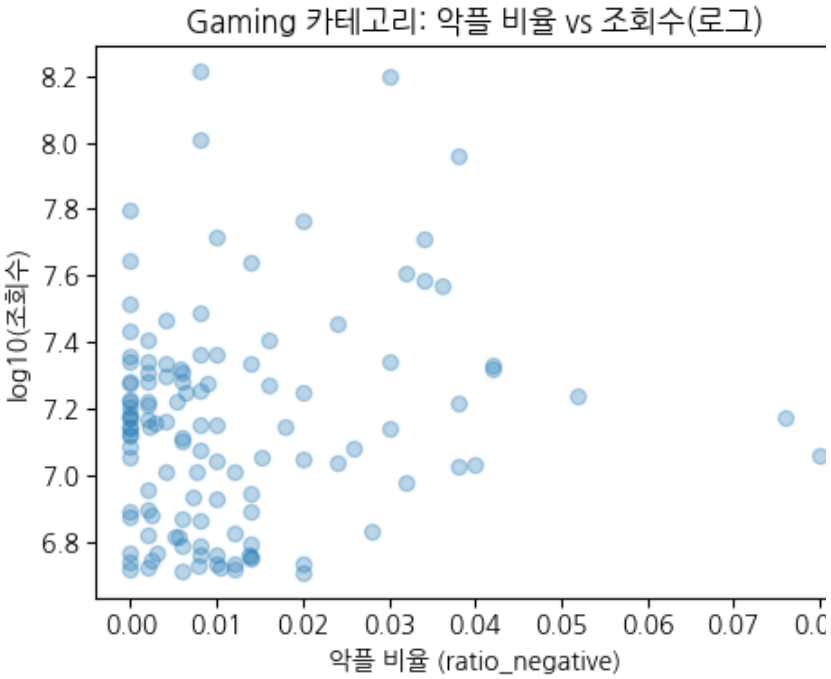
데이터 분석
-카테고리별 악플과 조회수 관계

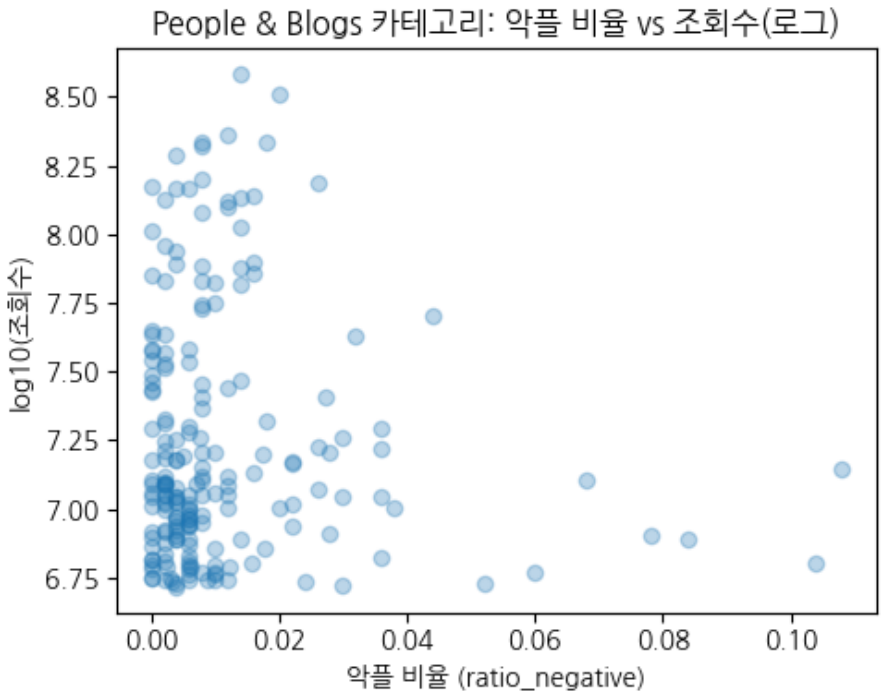


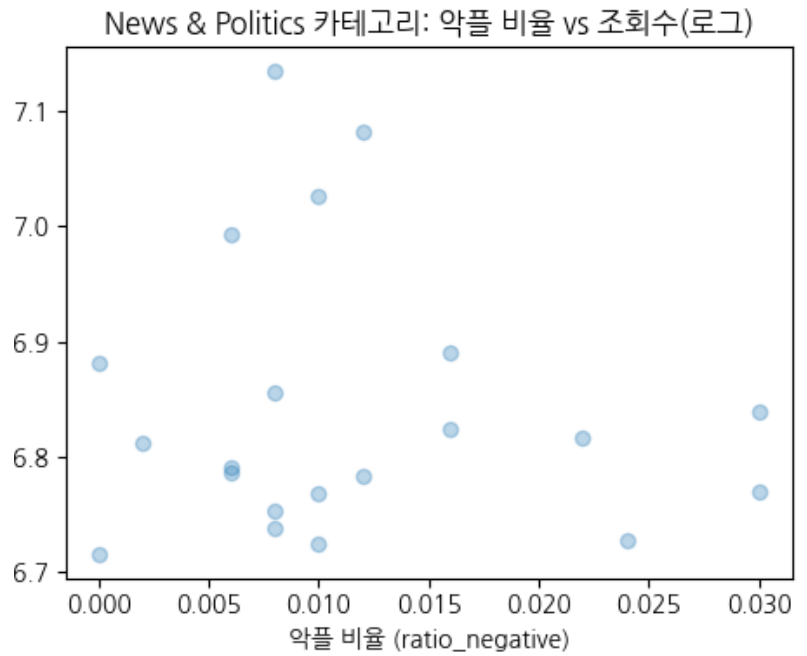


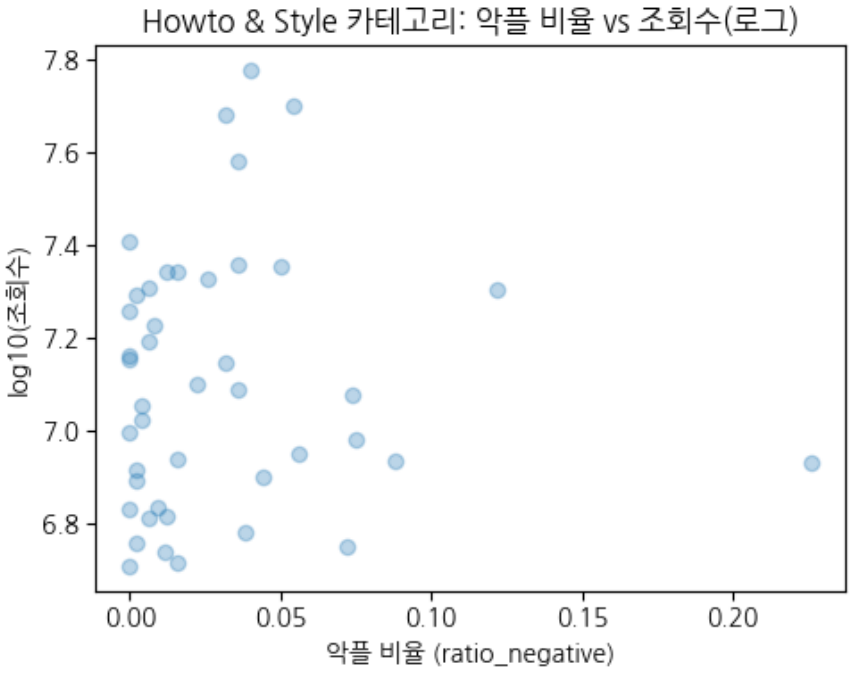


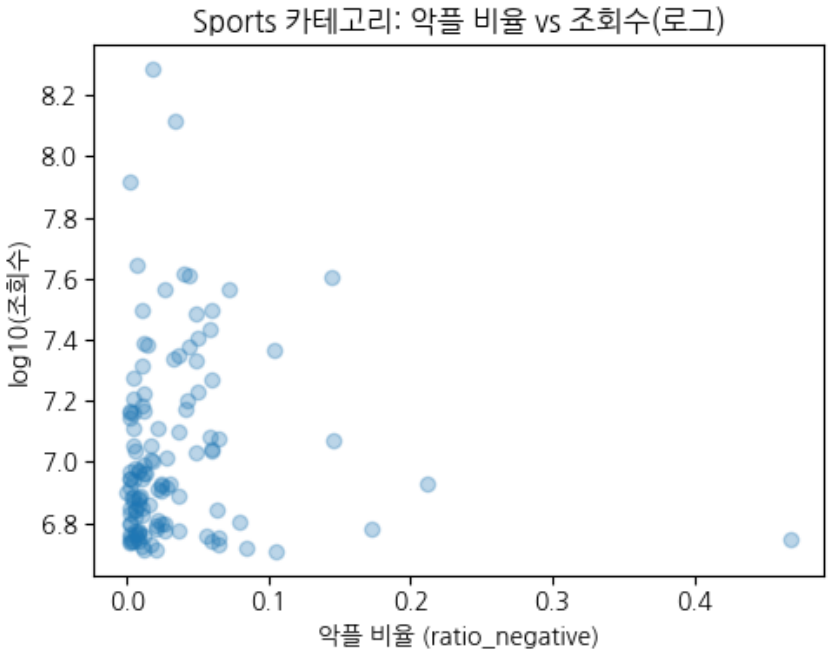


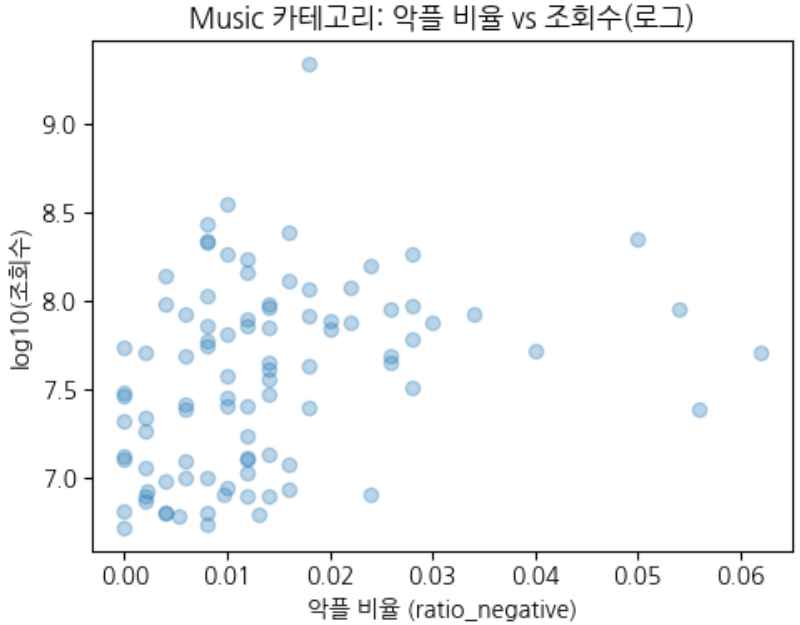

Gaming 카테고리: 악플 비율 vs 조회수(로그)
log10(조회수)
악플 비율 (ratio_negative)


People & Blogs 카테고리: 악플 비율 vs 조회수(로그)
log10(조회수)
악플 비율 (ratio_negative)

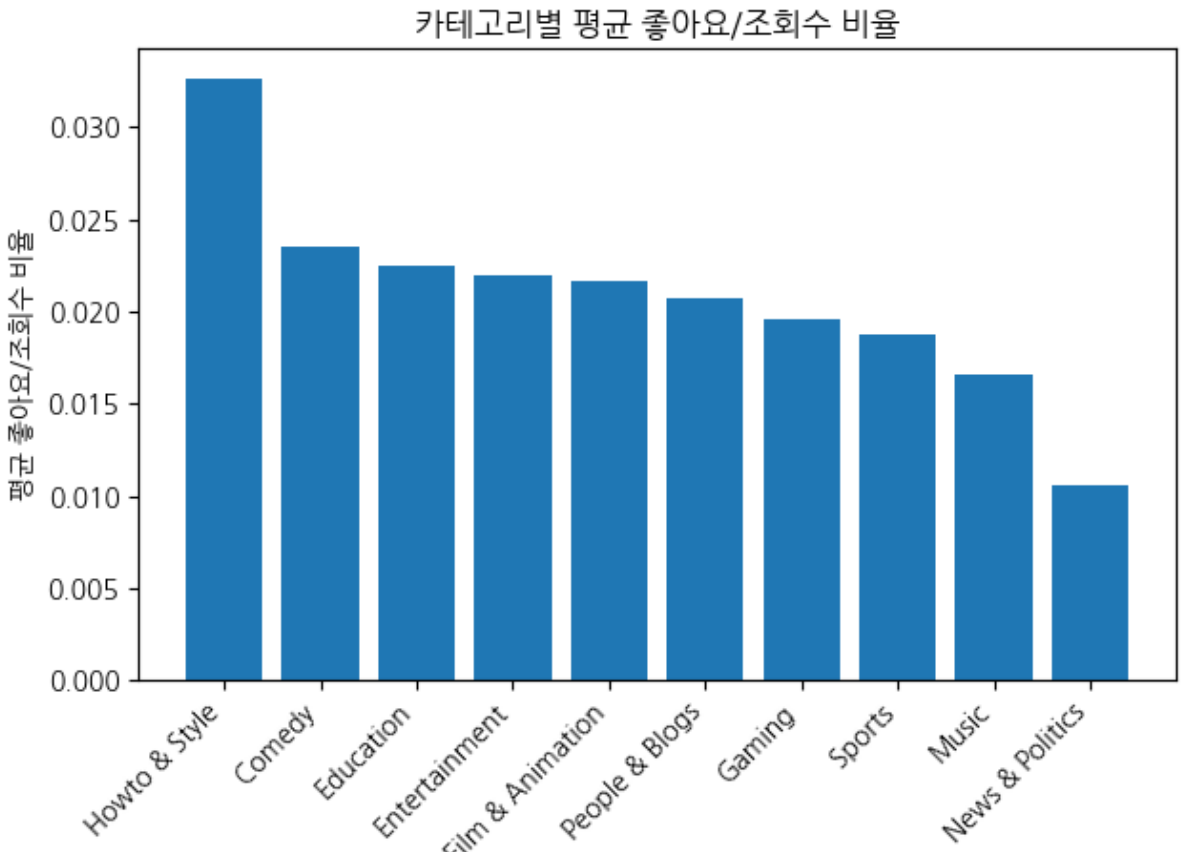

News & Politics 카테고리: 악플 비율 vs 조회수(로그)
악플 비율 (ratio_negative)

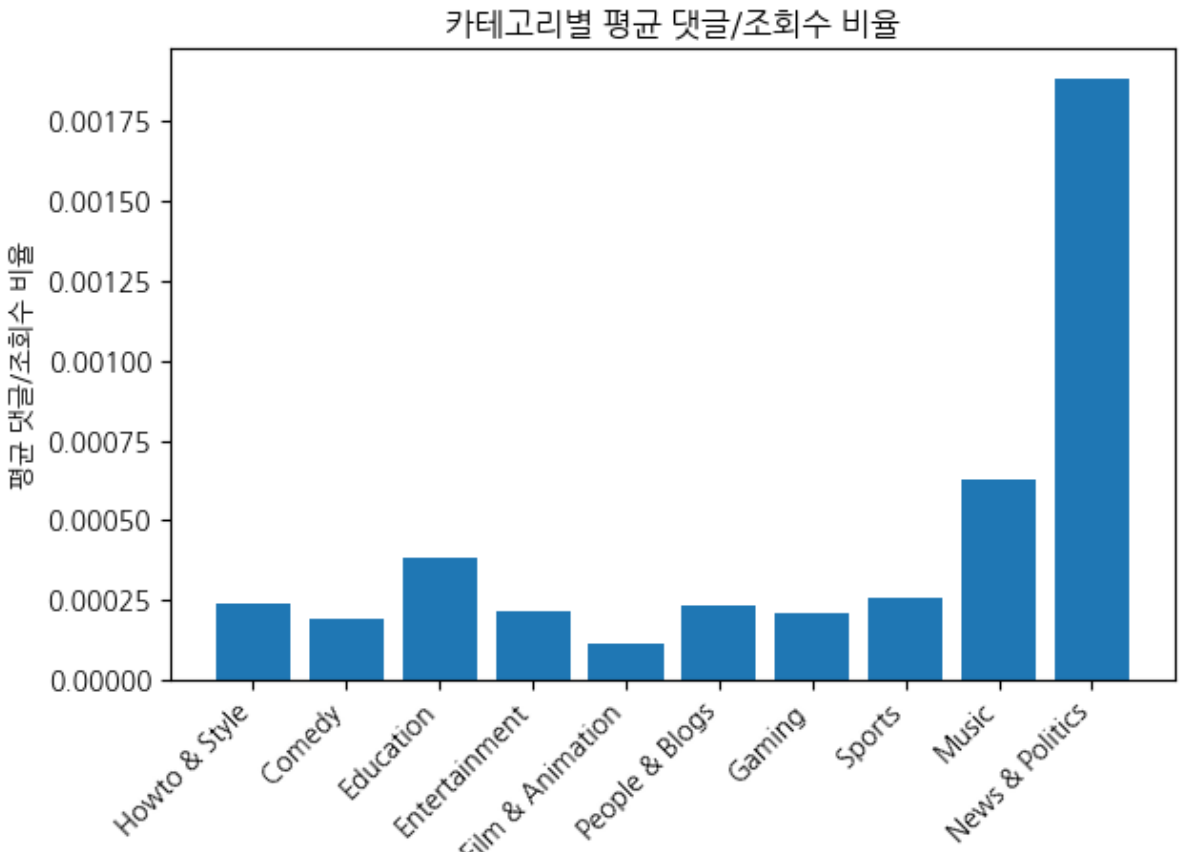

Howto & Style 카테고리: 악플 비율 vs 조회수(로그)
log10(조회수)
악플 비율 (ratio_negative)


Sports 카테고리: 악플 비율 vs 조회수(로그)
log10(조회수)
악플 비율 (ratio_negative)

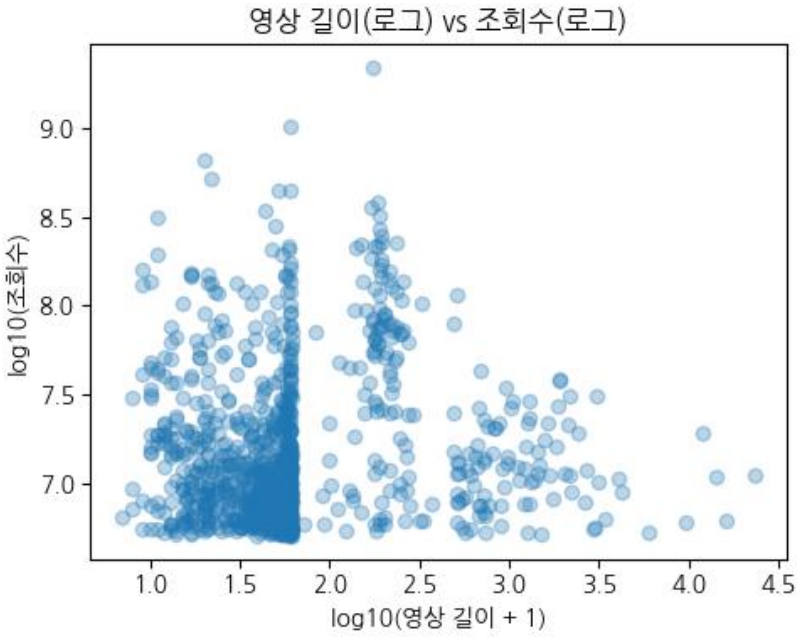

Music 카테고리: 악플 비율 vs 조회수(로그)
log10(조회수)
악플 비율 (ratio_negative)

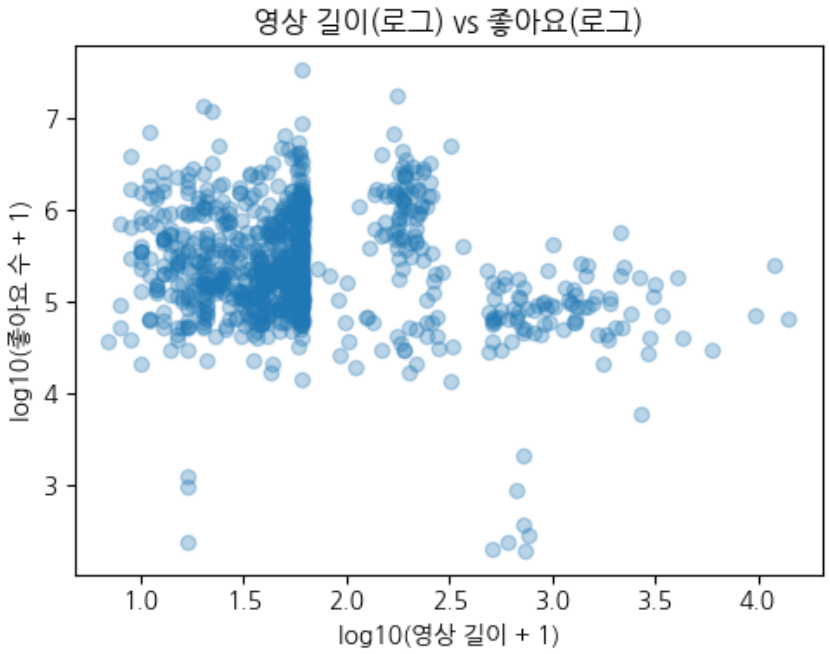
데이터 분석
-카테고리별 평균 좋아요&댓글/조회수 비율

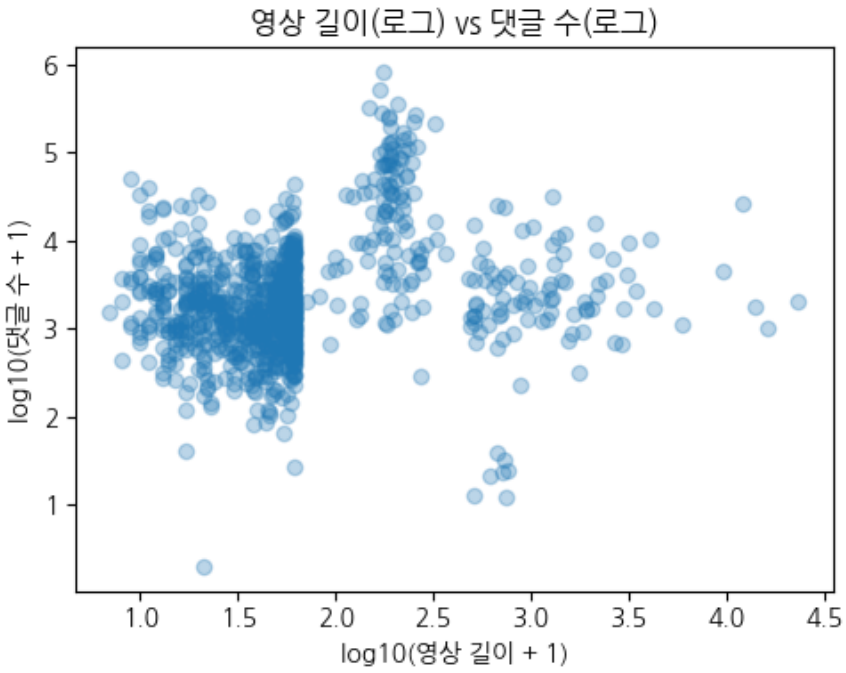




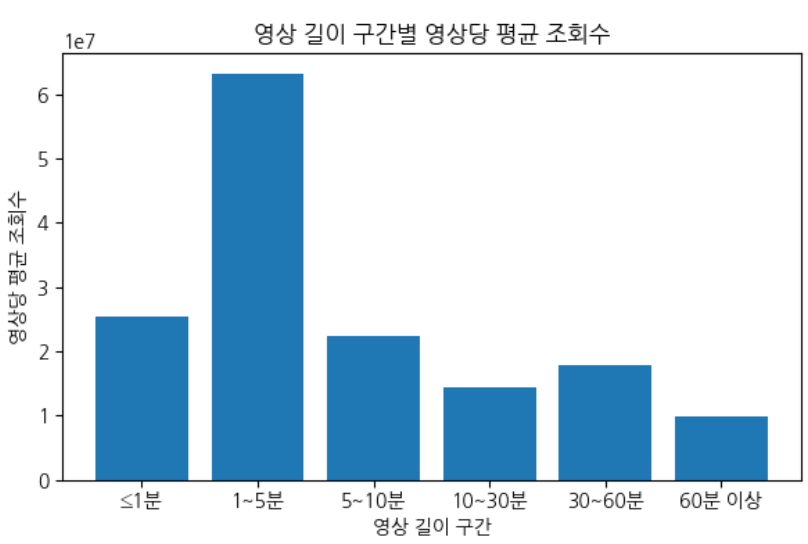
데이터 분석
-영상 길이와 조회수&좋아요 수&댓글 수 관계

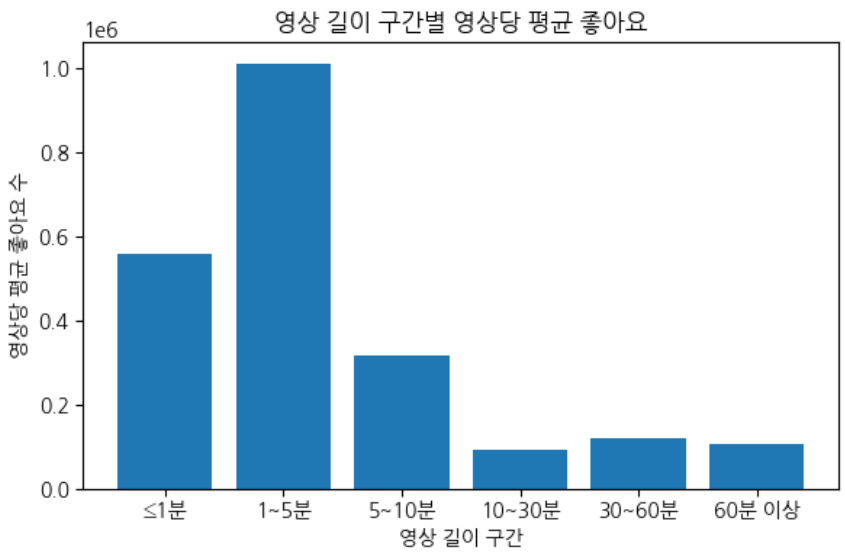


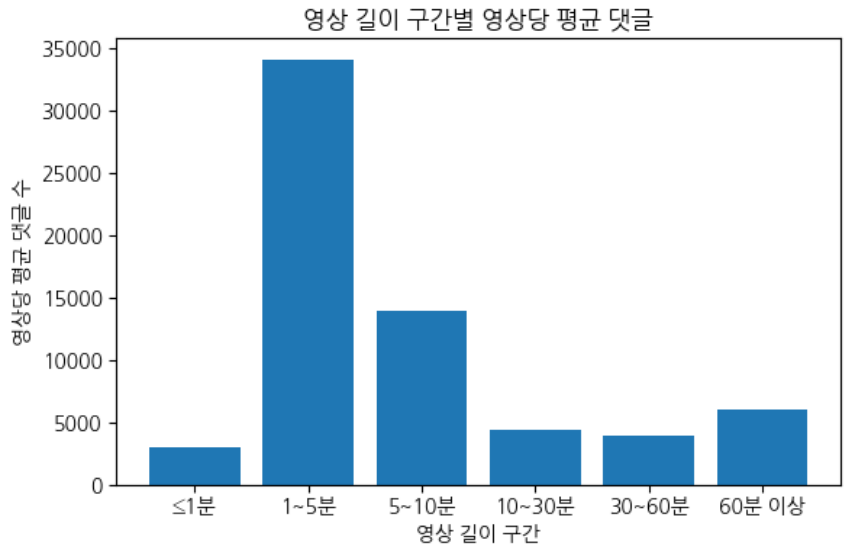




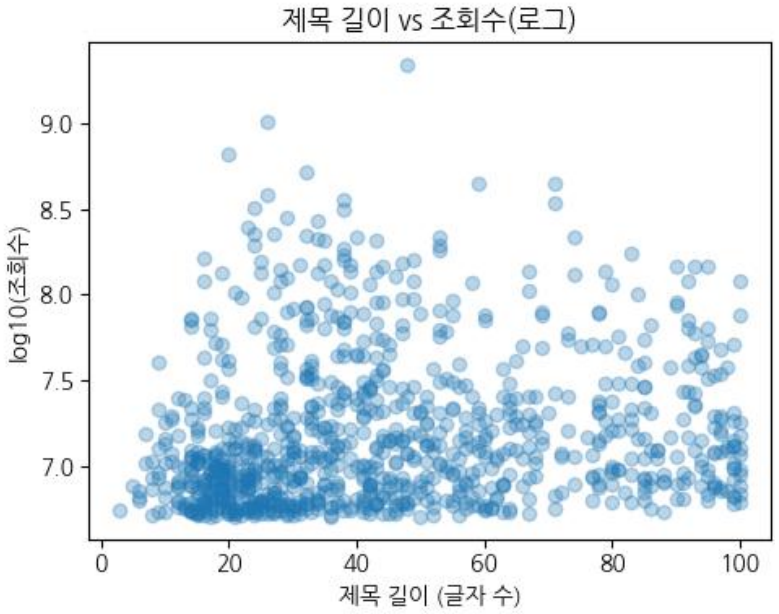
데이터 분석
-영상 길이와 조회수&좋아요 수&댓글 수 관계

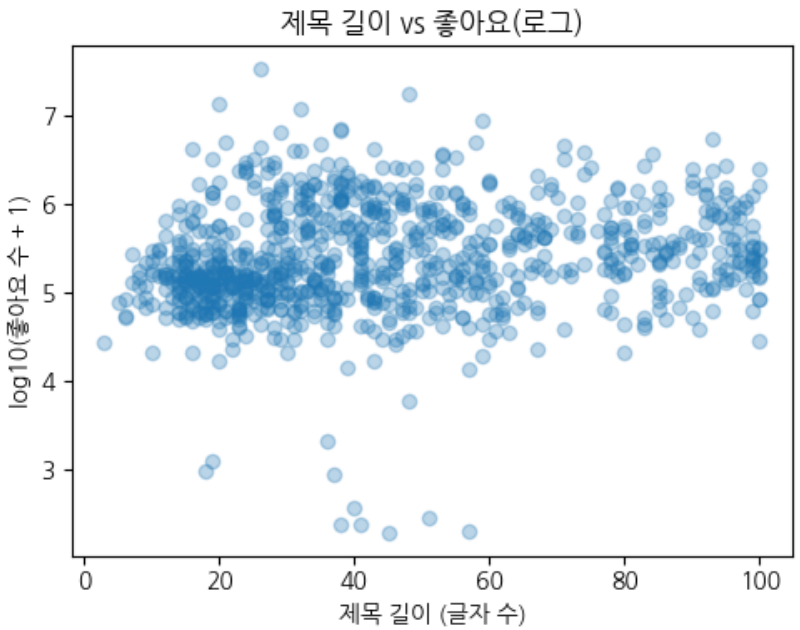


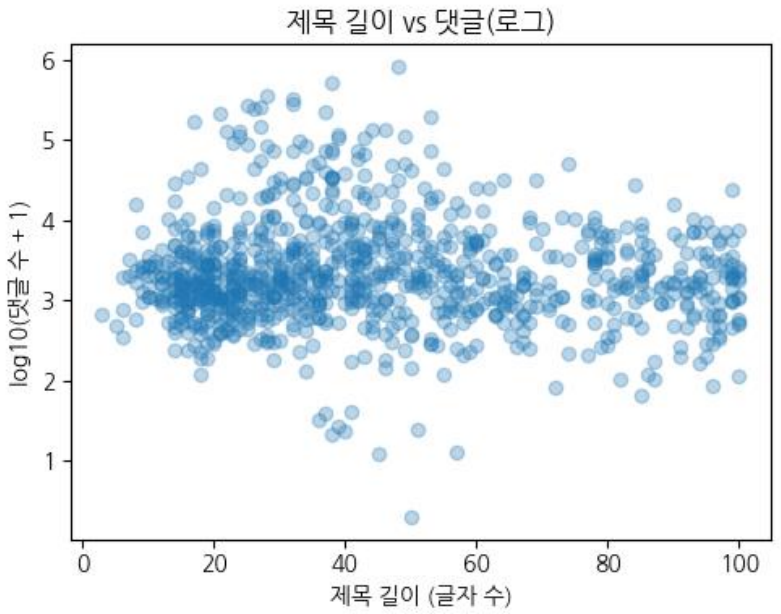




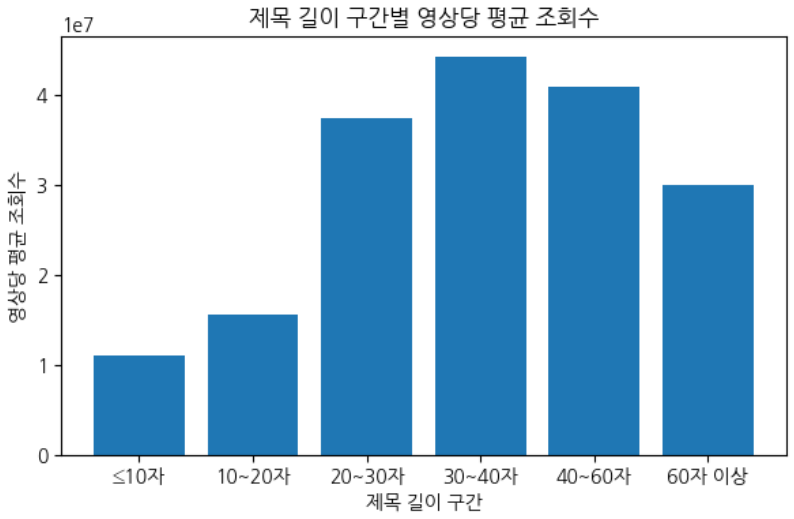
데이터 분석
-제목 길이와 조회수&좋아요 수&댓글 수 관계

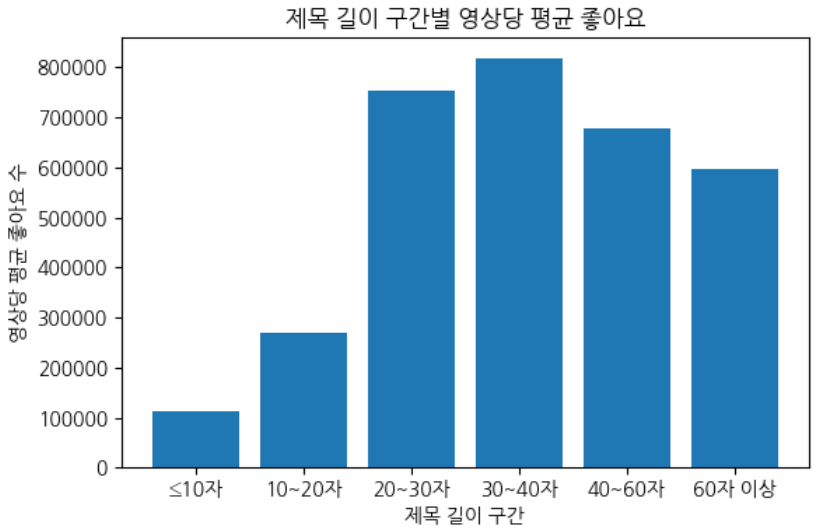


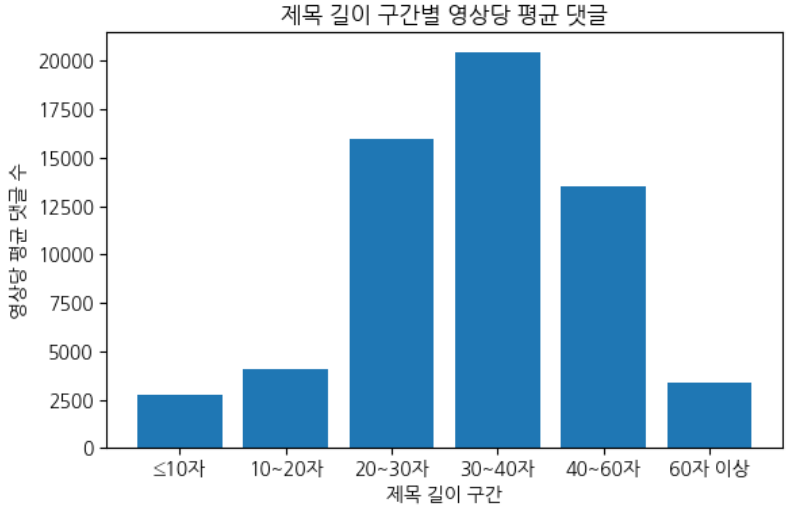




데이터 분석
-제목 길이와 조회수&좋아요 수&댓글 수 관계



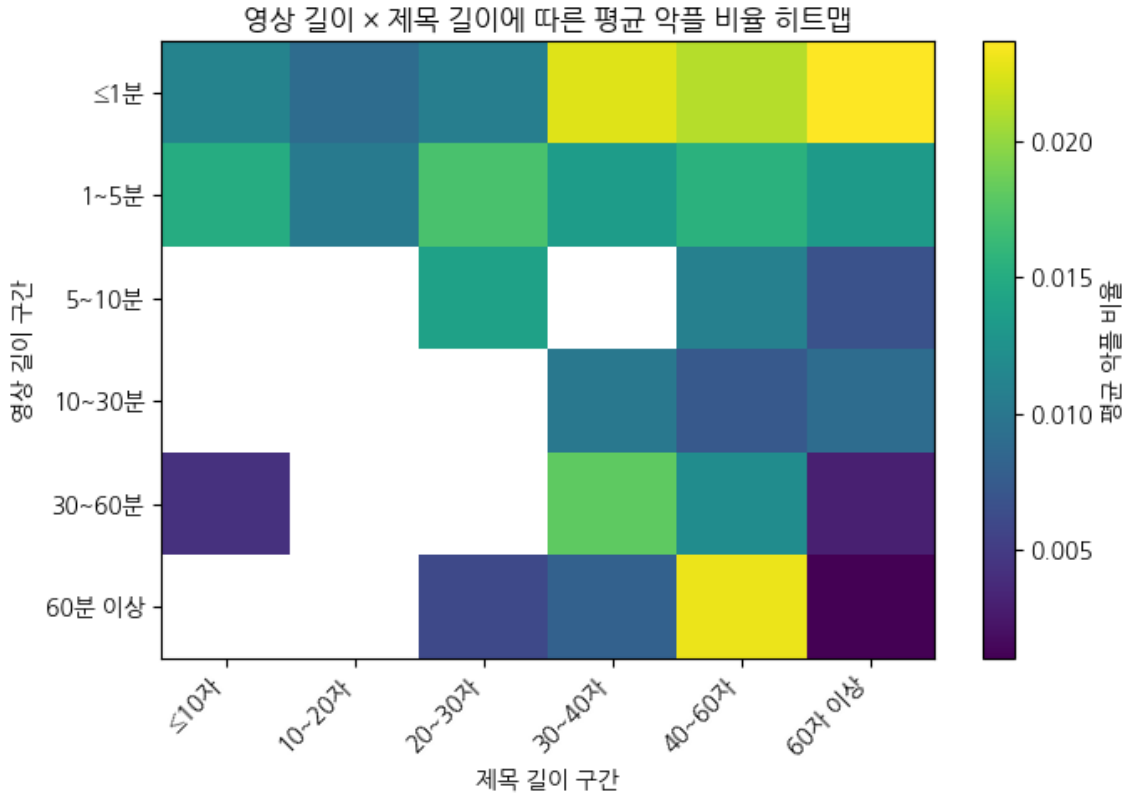


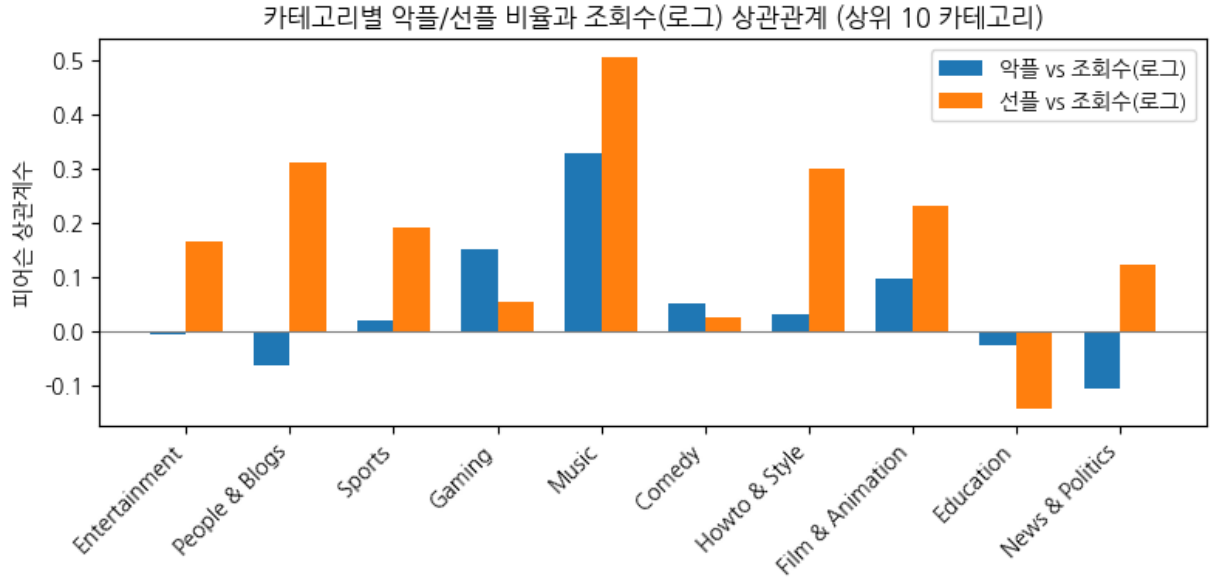


종합 데이터

	view_count	log_view_count
ratio_positive	0.157	0.268
ratio_negative	-0.016	-0.009

	category_name	n_videos	corr_neg_vs_logview	corr_pos_vs_logview
2	Entertainment	240.0	-0.005475	0.163645
8	People & Blogs	194.0	-0.065369	0.310896
9	Sports	123.0	0.020031	0.189318
4	Gaming	122.0	0.150254	0.053963
6	Music	94.0	0.326650	0.505832
0	Comedy	47.0	0.049644	0.024402
5	Howto & Style	43.0	0.029442	0.297795
3	Film & Animation	22.0	0.096635	0.231879
1	Education	21.0	-0.026585	-0.143541
7	News & Politics	21.0	-0.106803	0.121453





	view_count	log_view_count	like_count	log_like_count	comment_count	log_comment_count
duration_sec	-0.026272	-0.039870	-0.055175	-0.157978	-0.002623	0.034987
log_duration	0.037079	0.064568	-0.035225	-0.198480	0.157931	0.214070

	view_count	log_view_count	like_count	log_like_count	comment_count	log_comment_count
title_len_chars	0.016864	0.14845	0.028865	0.17721	-0.058444	-0.067789

피어슨 상관계수 활용

$$r_{xy} = \frac{\operatorname{cov}(X,Y)}{s_X\, s_Y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\,\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

=> : 추측

결론

1. 조회수와 선플/악플 개수는 상관관계가 있을까?

-조회수와 선플은 상관계수가 약 0.268으로 약한 상관관계/조회수와 악플은 상관계수가 약 -0.009로 상관관계x

=> 댓글은 인기순으로 정렬되어 있어 선플이 비교적 많이 노출됨-> 이것이 영상을 볼지 말지를 결정

2. 카테고리별로 조회수와 선플/악플의 상관관계는 어떻게 될까?

-음악 카테고리: 선플&악플과 조회수가 양의 상관관계 => 선플을 쓴 사람: 주로 팬덤, 반복 재생/악플을 쓴 사람: 감성 분석에서의 오류?

-게임 카테고리: 악플과 조회수가 약한 양의 상관관계 => 다소 과격한 표현이 많이 등장하기 때문일 것

-블로그 카테고리: 선플과 조회수가 양의 상관관계 => 내용이 소소할 가능성이 높음, 이에 따라 댓글도 순할 것

-교육 카테고리: 선플과 조회수가 약한 음의 상관관계=> 문제의 난이도 같은 걸 얘기할 때 부정적인 표현이 사용됐을 가능성이 높음

결론

3. 영상 길이, 제목 길이와 조회수, 좋아요 수, 댓글 수 사이의 상관관계가 존재할까?

-영상 길이: 좋아요 수와 약한 음의 상관관계, 댓글 수와 약한 양의 상관관계

=> 쇼츠에서는 좋아요를 누르기 더 쉬움, 긴 영상일수록 게시자의 노력이 더 느껴져 댓글을 남김

-제목 길이: 뚜렷한 상관관계X

-쇼츠가 유행을 타며, 1분보다 짧거나 비슷한 길이의 영상이 많이 등장, 주로 1-5분 길이의 영상이 성과가 좋음

-제목 길이는 30-40자의 꽤 긴 영상의 성과가 좋음 => 아마 높은 조회수를 가진 외국 영상이 많아 영어로 글자수가 세져서 길어졌을 것

한계

- 감성 분석에서 사용된 단어가 한정적임(오타, 반어법 등 포함X)

⇒ 딥러닝을 통해 인공지능이 감성 분석을 하도록 지시

- 쿼리를 직접 임의로 썼으므로, 조회수가 높거나 통계적으로 유의미한 영상이 수집되지 않았을 가능성이 높음

⇒ Kaggle 등 다른 사용자들이 이미 만들어 놓은 데이터 활용

- 영상, 댓글을 1000개, 500개씩 샘플링하여 사용해 전체 데이터의 경향을 확인하지 못함

⇒ Google Cloud 유료 결제를 통해 불러올 수 있는 데이터양 확대

추가 연구-딥러닝 감성분석

기존 감성 분석

	category_name	n_videos	corr_neg_vs_logview	corr_pos_vs_logview
2	Entertainment	240.0	-0.005475	0.163645
8	People & Blogs	194.0	-0.065369	0.310896
9	Sports	123.0	0.020031	0.189318
4	Gaming	122.0	0.150254	0.053963
6	Music	94.0	0.326650	0.505832
0	Comedy	47.0	0.049644	0.024402
5	Howto & Style	43.0	0.029442	0.297795
3	Film & Animation	22.0	0.096635	0.231879
1	Education	21.0	-0.026585	-0.143541
7	News & Politics	21.0	-0.106803	0.121453

	view_count	log_view_count
ratio_positive	0.157	0.268
ratio_negative	-0.016	-0.009

딥러닝 감성 분석

	category_name	n_videos	corr_neg_logview	corr_pos_logview	corr_neg_view	corr_pos_view
7	Music	94.0	0.394124	0.511878	0.116870	0.254865
5	Gaming	122.0	0.143870	0.050656	0.071783	0.039182
4	Film & Animation	22.0	0.121867	0.218638	0.194567	0.357174
12	Sports	123.0	0.055097	0.182901	0.021792	0.236514
6	Howto & Style	43.0	0.032678	0.277779	0.031215	0.224232
1	Comedy	47.0	0.005918	0.004593	-0.069385	0.031329
3	Entertainment	240.0	-0.005925	0.150817	-0.046563	0.057494
9	People & Blogs	194.0	-0.068397	0.296529	-0.022884	0.274801
8	News & Politics	20.0	-0.186036	0.131206	-0.199346	0.193861
2	Education	21.0	-0.196149	-0.159682	-0.139510	-0.157509

악플 비율 vs 로그 조회수: 0.0009148944211050108
선플 비율 vs 로그 조회수: 0.26036252814596317

감사합니다