

[기업과제]

데이터 분석 및 시각화

원티드 프리온보딩 코스
1팀_김태연

!!!

과제를 진행하기 전, EDA 및 데이터전처리 과정이 수반되었습니다.

해당 부분 및 과제의 전체 코드가 궁금하시다면 함께 첨부드리는

'junior_김태연_pretest.ipynb' 파일을 참고 부탁드립니다.

감사합니다.

1. 데이터 타입별 시각화

먼저, 시각화에 필요한 특성만 추출하여 새로운 데이터프레임 생성

	published_date	year	month	week	category_name	tags	channel_id	video_id
0	2021-07-01	2021	7	26	Entertainment	SiriusXM Sirius XM Sirius SXM BIGHIT 빅히트 방탄소년단...	CH49ta0	V-0db
1	2021-06-24	2021	6	25	Entertainment	치킨불냉면 치킨 불냉면 냉면	CHZVD--	V-1XL
2	2021-07-17	2021	7	28	Entertainment	missing	CH9w-h_	V-4fa
3	2021-06-02	2021	6	22	Sports	News Network SBS SPORTSMUG SPORTSMUG 스포츠머그 축구 ...	CHUQVGX	V-5ip
4	2021-07-06	2021	7	27	Sports	이천수 심판도전기 축구심판	CHh13EX	V-5jn
...
2539	2021-05-09	2021	5	18	Comedy	아프리카 tv 봉준 와꾸대장봉준 B J 컨텐츠 클립	CH69uMh	VzwuB
2540	2021-07-29	2021	7	30	Comedy	장삐쭈 삐쭈 ㅋㅋㅋ 삐쭈 장삐쭈 방맛더빙 더빙 웃긴동영상 꿀잼 신병 장삐쭈 단편선 ...	CHhbE5O	VzxuL
2541	2021-04-20	2021	4	16	Science & Technology	아이패드 프로 아이패드 프로5 아이패드 프로 5세대 신형 아이맥 iMac 에...	CHO4RG1	VzygR
2542	2021-04-26	2021	4	17	Entertainment	고요 속의 외침 뽕송아학당 슬기로운캠핑생활 아는형님 미스터트롯 임영웅 영탁 장민호 ...	CHYeeEw	Vzz6W
2543	2021-06-02	2021	6	22	Music	MAMAMOO 마마무 WAW 마마무 WAW MAMAMOO WAW Where Are ...	ChuhAUM	Vzzk0

2544 rows × 8 columns

1) (전체기간) 카테고리별 > 채널별 > 비디오 개수 시각화

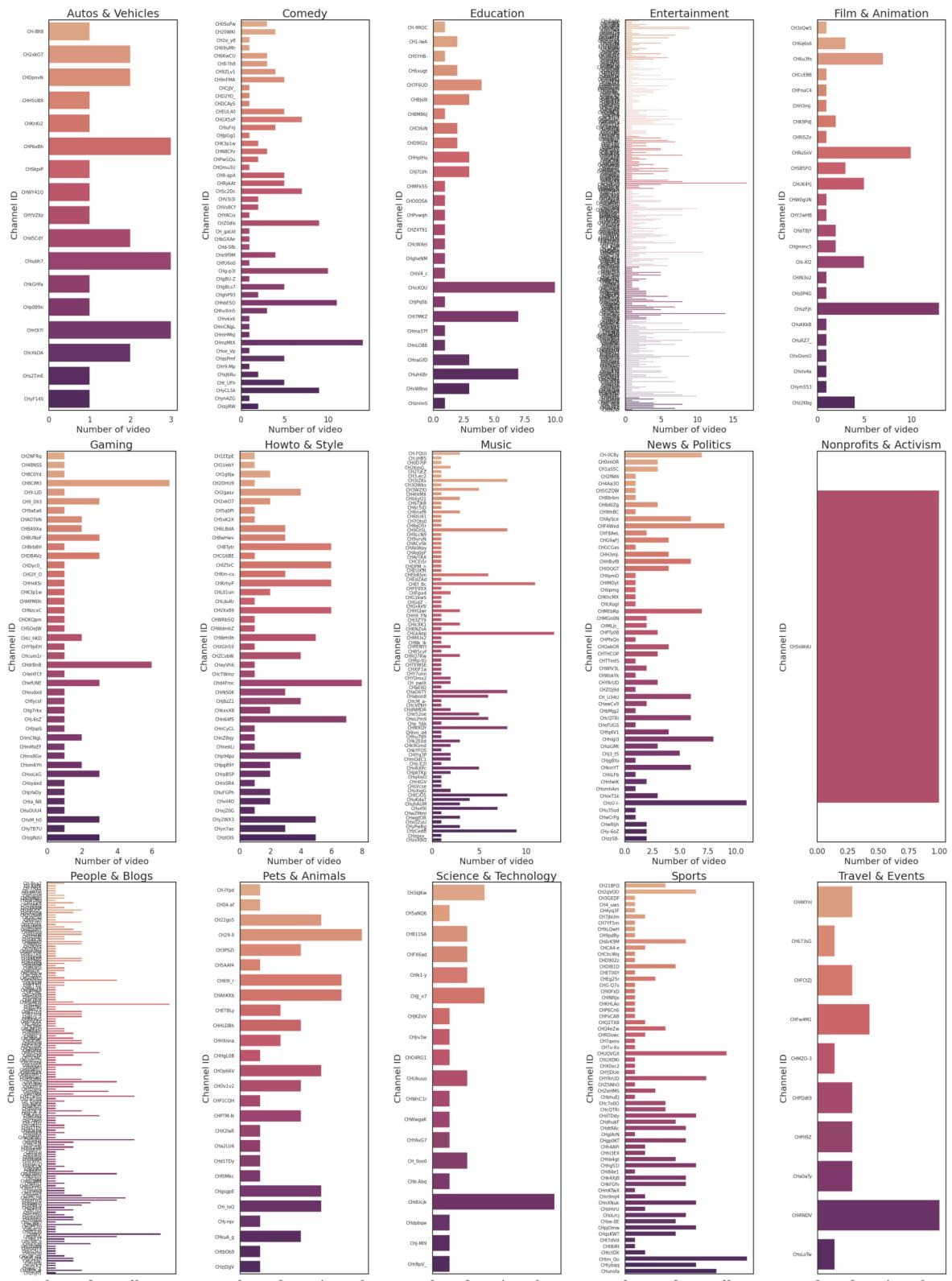
(1) Plot 버전

- melted 형태의 데이터프레임 생성

	category_name	channel_id	video_id
0	Autos & Vehicles	CH-IBt8	1
1	Autos & Vehicles	CH2xkO7	2
2	Autos & Vehicles	CHDpnvN	2
3	Autos & Vehicles	CHH5U89	1
4	Autos & Vehicles	CHKnKi2	1
...
915	Travel & Events	CHPDdt9	2
916	Travel & Events	CHPII9Z	2
917	Travel & Events	CHaOaTy	2
918	Travel & Events	CHIRNDV	7
919	Travel & Events	CHsLoTw	1

920 rows × 3 columns

- seaborn barplot(barh 형태)을 활용하여 카테고리별로 구분된 채널별 비디오 개수 시각화 진행



(2) DataFrame Styler 적용 버전

- Pivot table 형식의 데이터프레임에 pandas style 을 활용하여 시각화
- 카테고리별 - 채널별 비디오 개수를 희소한 형태로 확인 가능
- 한번에 보기 힘든 형태지만, 채널 ID 를 알고 있을 경우 검색하면 그 위치로 바로 이동하여 비디오 개수를 확인할 수 있다는 장점이 있음

channel_id	CH-OC8y	CH-9RQC	CH-BqPA	CH-Baa2	CH-FQZI	CH-ITpd	CH-Jbic	CH-KaFr	CH-VG66	CH-VsPg	CH-YRx5	CH-gIR4	CH-JHBS	CH-swID	CH04-aE	CHOD7tP	CHOPeUG	CHOSoPw	CHOVr2v	CHimOR	CHOsaa8	CHOwnau	CHOwnThg	CH1-lwA	CH12YJZ	CH18g7g	CH1EEpE	CH1EiHI	CH1GRhg	C
category_name																														
Autos & Vehicles	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Comedy	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	
Education	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	
Entertainment	0	0	2	0	0	0	0	1	0	1	3	0	2	0	0	0	9	0	9	0	0	0	0	1	1	1	0	0	0	
Film & Animation	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Gaming	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Howto & Style	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	
Music	0	0	0	0	3	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	
News & Politics	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	
Nonprofits & Activism	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
People & Blogs	0	0	0	2	0	0	0	1	0	0	1	0	0	1	0	0	1	0	0	0	2	2	1	0	0	0	0	3	1	
Pets & Animals	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Science & Technology	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Sports	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Travel & Events	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

2) (월별) 카테고리별 > 채널별 > 비디오 개수 시각화

(1) Plot 버전

- melted 형태의 데이터프레임 생성

month	category_name	channel_id	video_id
0	3	Comedy	CHEUL40
1	3	Comedy	CHV3i3l
2	3	Comedy	CHgBLs7
3	3	Comedy	CHyCL5k
4	3	Education	CH5YHB-
...
1608	7	Sports	CHpjOmw
1609	7	Sports	CHqsKWT
1610	7	Sports	CHtm_Qo
1611	7	Sports	CHunsfa
1612	7	Travel & Events	CHIRNDV

1613 rows × 4 columns

- category_month 형태의 변수명으로 데이터프레임 생성

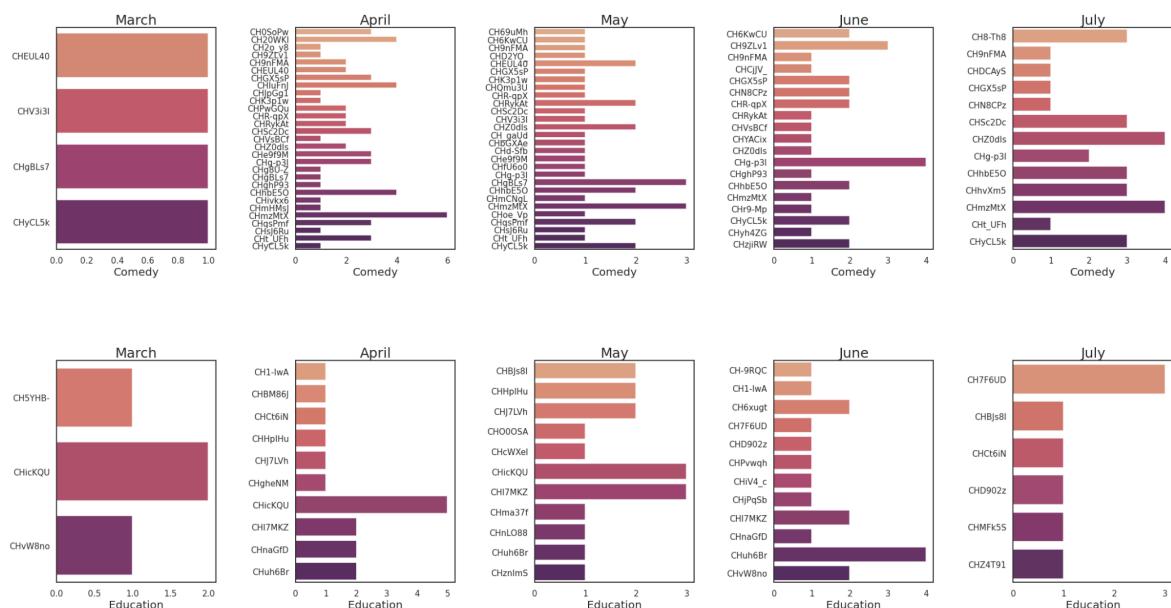
```
for category in categories:
    for idx in range(len(months)):
        globals()['{}_{}`'.format(category, months[idx])] =
            grouped_month[(grouped_month['category_name'] == category) &
                           (grouped_month['month'] == months[idx])]
```

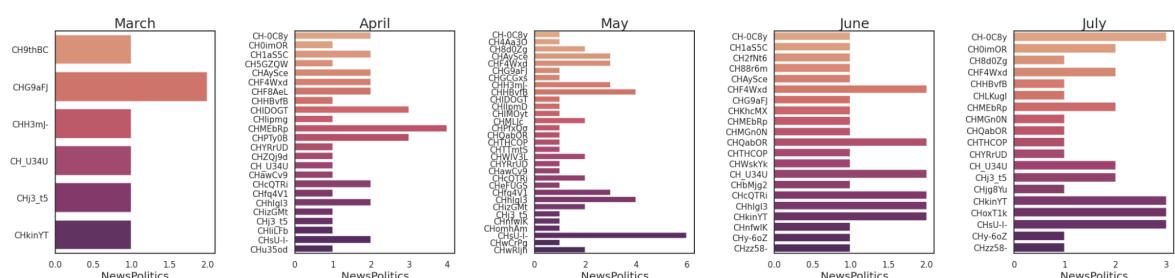
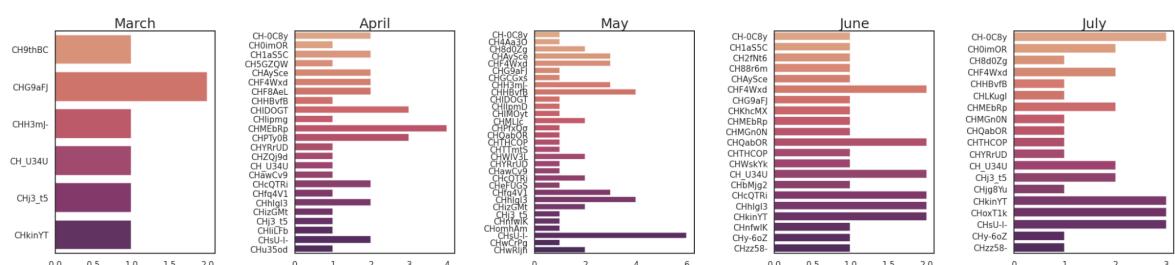
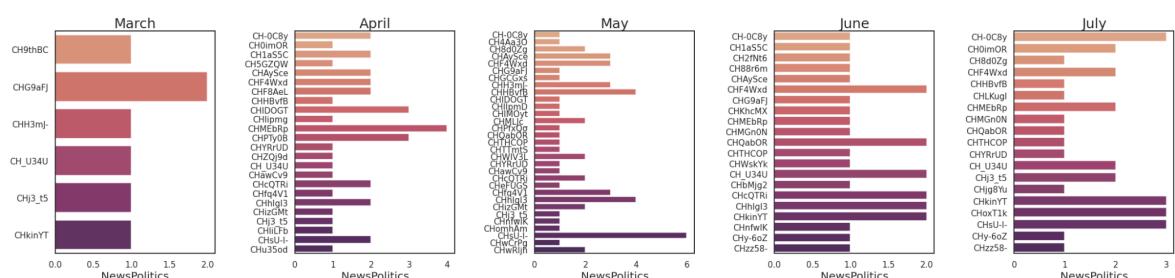
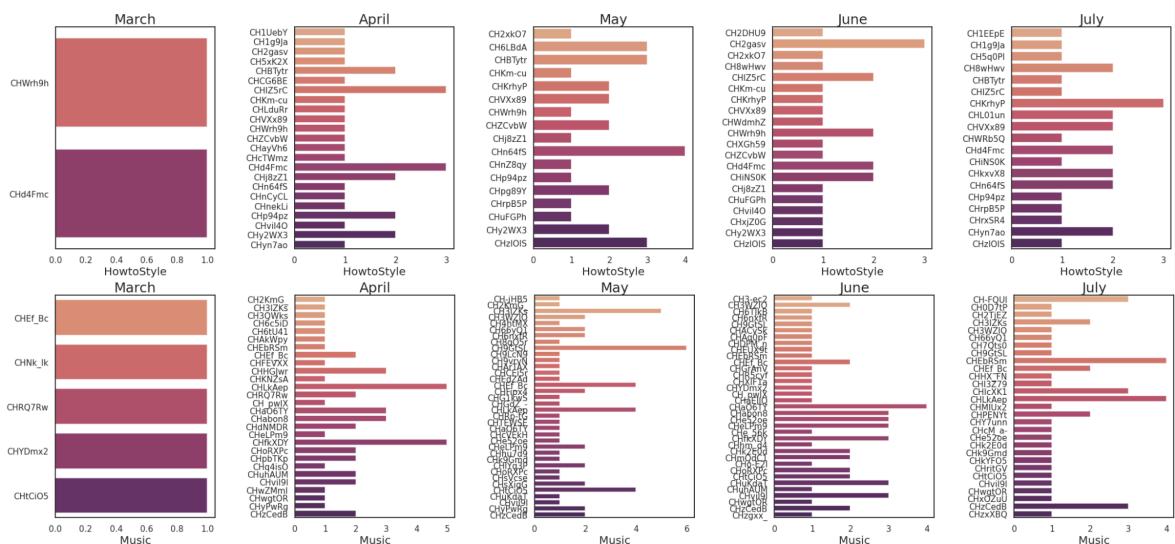
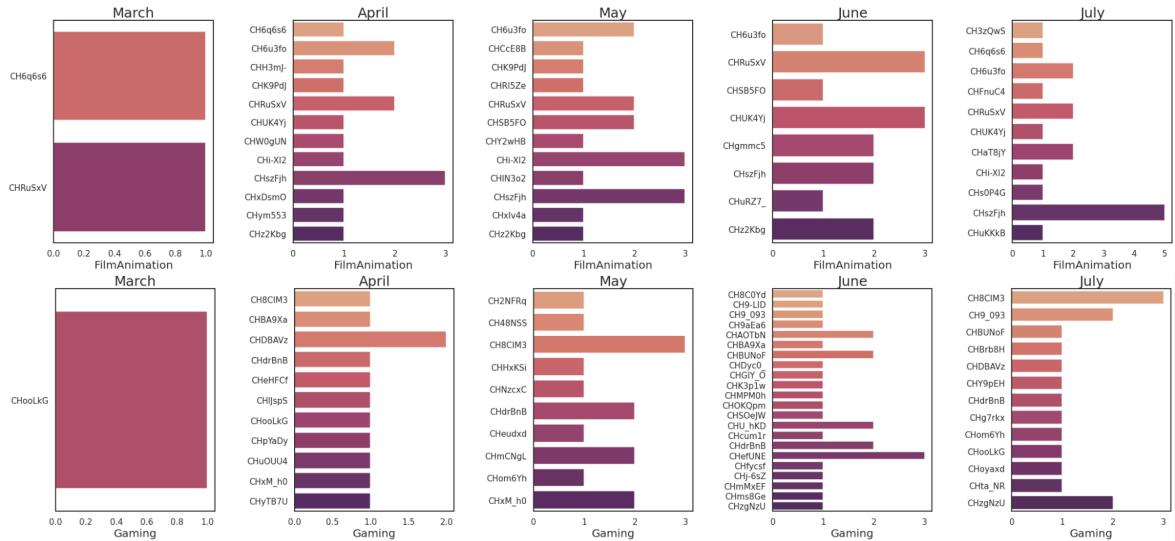
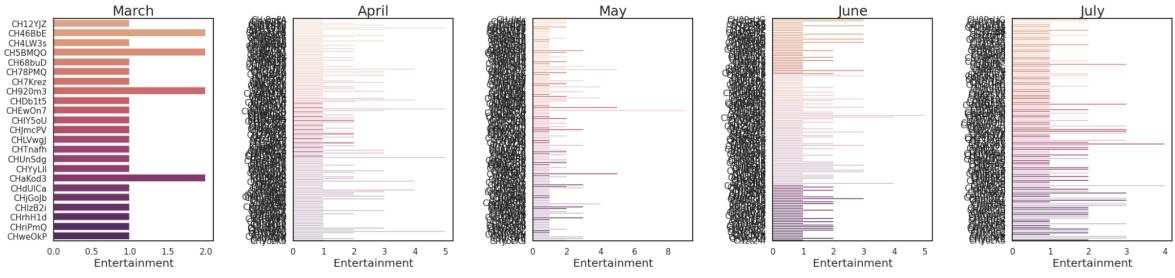
(변수명 예시) Entertainment_3 / Sports_5 / ...
 Entertainment_7 데이터프레임 출력 예시

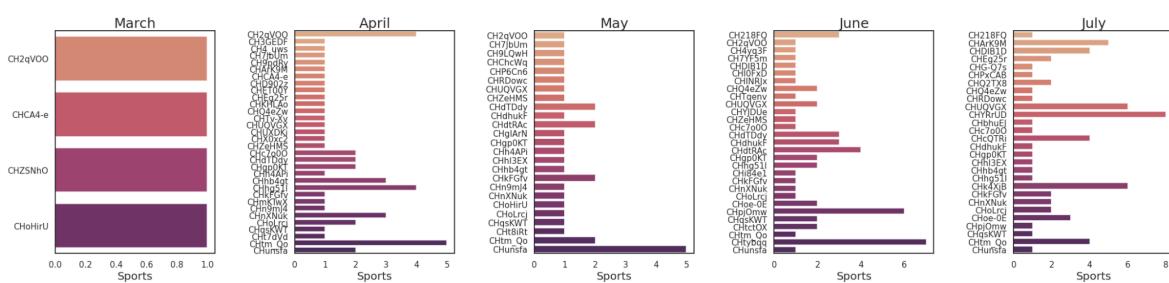
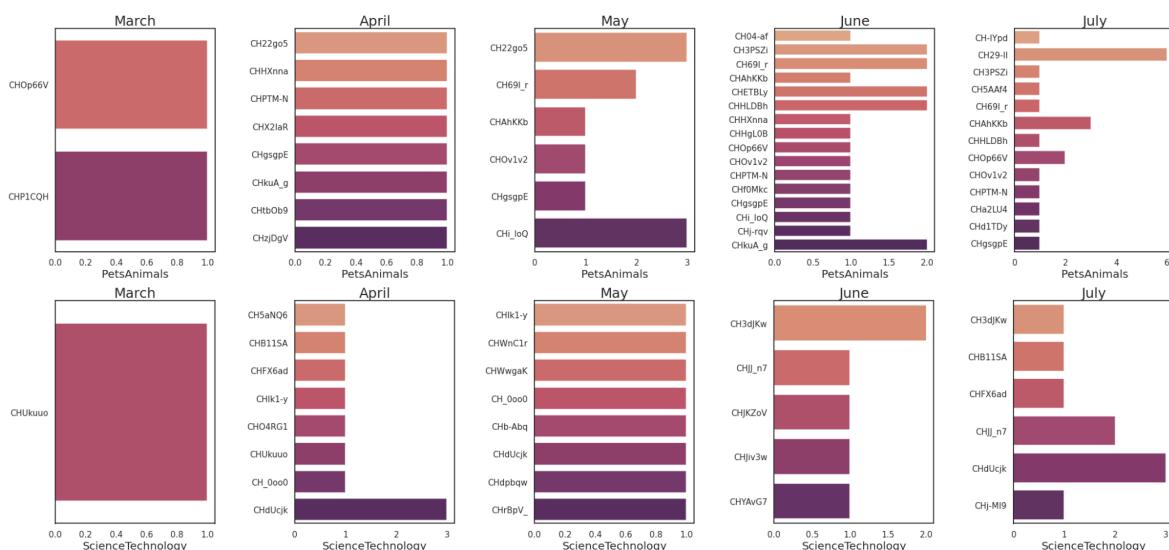
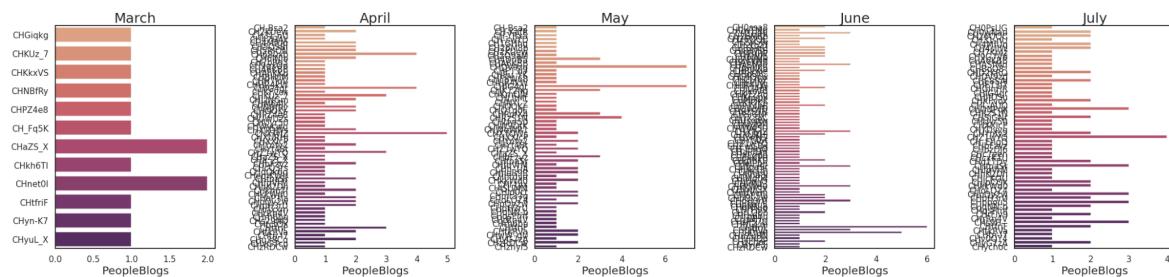
month	category_name	channel_id	video_id
1296	7	Entertainment	CHOPsUG
1297	7	Entertainment	CHOVR2v
1298	7	Entertainment	CH1L79y
1299	7	Entertainment	CH1cWTE
1300	7	Entertainment	CH2DHU9
...
1414	7	Entertainment	CHxLcOz
1415	7	Entertainment	CHy-NrX
1416	7	Entertainment	CHy-swB
1417	7	Entertainment	CHybPxZ
1418	7	Entertainment	CHyoZK5

123 rows × 4 columns

- 시각화 결과 (barh 형태)





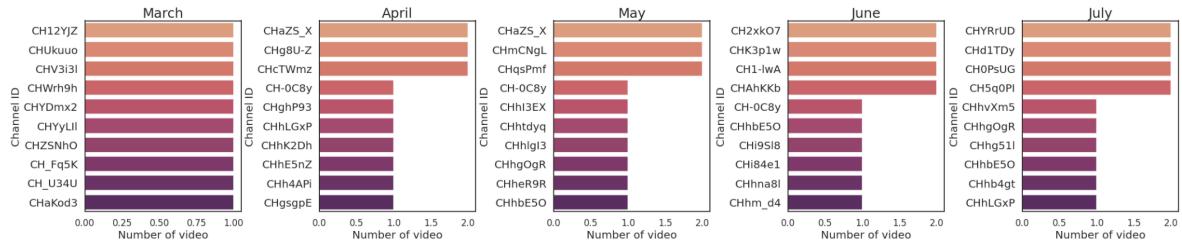


(2) Dataframe Styler 적용 버전

- (위에서부터) 3 월 ~ 7 월

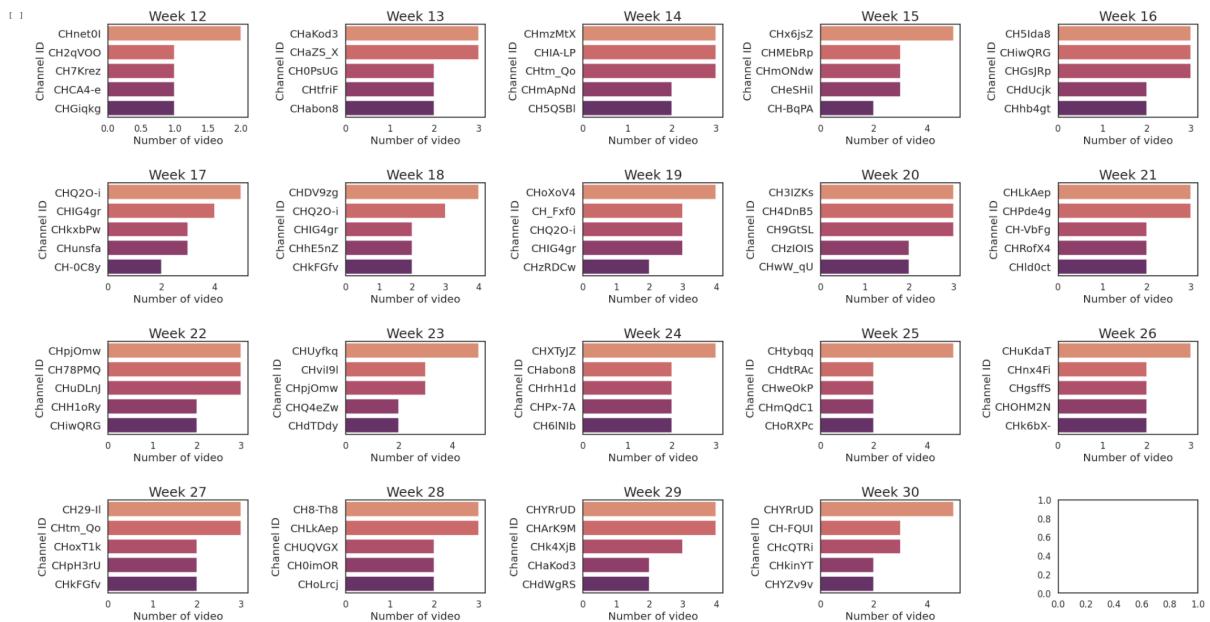
3) 월별 TOP10 채널 (분류기준 : 비디오 개수)

- 월 : 3월 ~ 7월
- X축 = 비디오 개수
- Y축 = 채널 ID



4) 주별 TOP5 채널 (분류기준 : 비디오 개수)

- 주차 : Week 12 ~ Week 30
- X축 = 비디오 개수
- Y축 = 채널 ID



5) 월별 카테고리별 태그 키워드 순위

(1) tags 값들 토큰나이징

month	category_name	tags	tokens
0	7 Entertainment	SiriusXM Sirius XM Sirius SXM BIGHIT 빅히트 방탄소년단...	[siriusxm, sirius, xm, sirius, sxm, bighit, 빅히...
1	6 Entertainment	치킨불냉면 치킨 불냉면 냉면	[치킨불냉면, 치킨, 불냉면, 냉면]
2	7 Entertainment	missing	[missing]
3	6 Sports	News Network SBS SPORTSMUG SPORTSMUG 스포츠머그 축구 ...	[news, network, sbs, sportsmug, sportsmug, 스포츠...
4	7 Sports	이천수 심판도전기 축구심판	[이천수, 심판도전기, 축구심판]
...
2539	5 Comedy	아프리카tv 봉준 와꾸대장봉준 BJ 컨텐츠 클립	[아프리카tv, 봉준, 와꾸대장봉준, bj, 컨텐츠, 클립]
2540	7 Comedy	장빼쭈 빼쭈 ㅋㅋㅋ 빼쮸 장빼쮸 벙앗더빙 더빙 웃긴동영상 꿀잼 신병 장빼쭈 단편선 ...	[장빼쭈, 빼쭈, 빼쮸, 장빼쮸, 벙앗더빙, 더빙, 웃긴동영상, 꿀잼, 신병, 장빼...
2541	4 Science & Technology	아이패드 프로 아이패드 프로5 아이패드 프로 5세대 신형 아이맥 아이맥 iMac 에...	[아이패드, 프로, 아이패드, 프로, 아이패드, 프로, 세대, 신형, 아이맥, 아이...
2542	4 Entertainment	고요 속의 외침 방송아학당 슬기로운캠핑생활 아는형님 미스터트롯 임영웅 영탁 장민호 ...	[고요, 속의, 외침, 뽕송아학당, 슬기로운캠핑생활, 아는형님, 미스터트롯, 임영웅...
2543	6 Music	MAMAMOO 마마무 WAW 마마무 WAW MAMAMOO WAW Where Are ...	[mamamoo, 마마무, waw, 마마무, waw, mamamoo, waw, wh...

2544 rows x 4 columns

(2) 토큰화된 문서들을 입력받아 토큰을 카운트 하고 관련된 속성을 가진 데이터프레임을 반환하는 word_count() 함수 정의

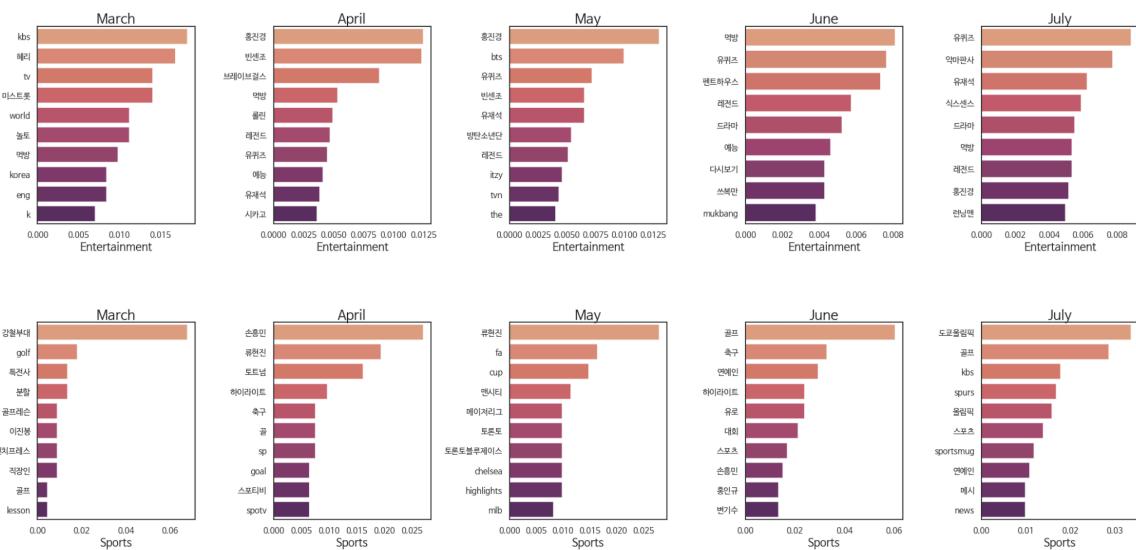
- 입력값 : 토큰화된 문서가 들어있는 list
- 출력값 : Dataframe
- 계산속성 : 단어 빈도 카운트 / 단어의 순위 / 코퍼스 내 단어의 비율 / 누적 비율 / 전체 문서 중 단어가 존재하는 비율
- word count 데이터프레임 출력 예시

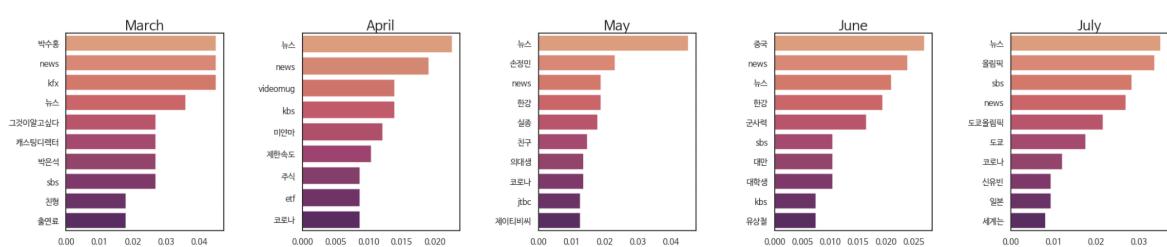
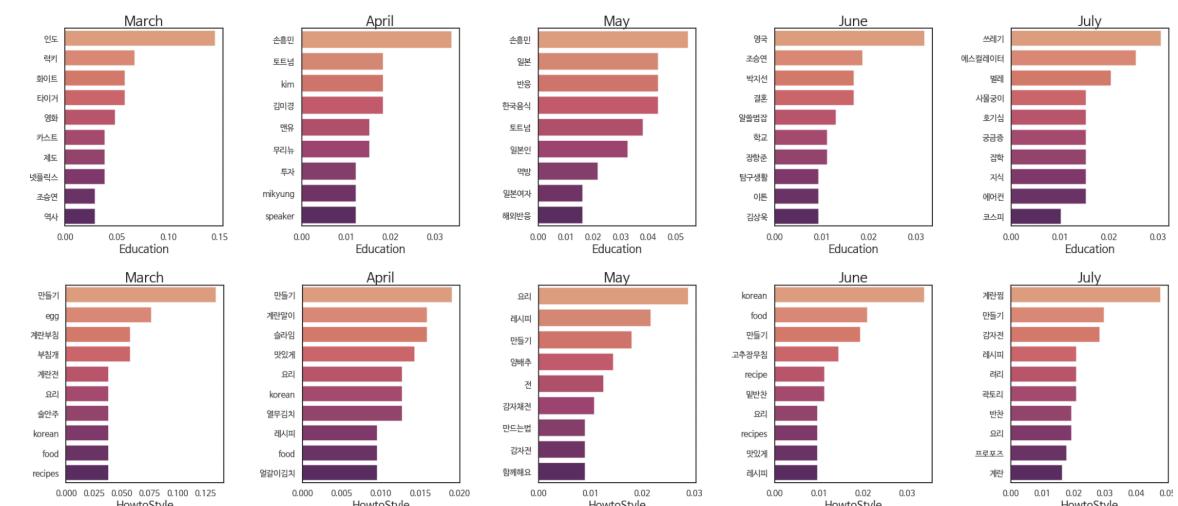
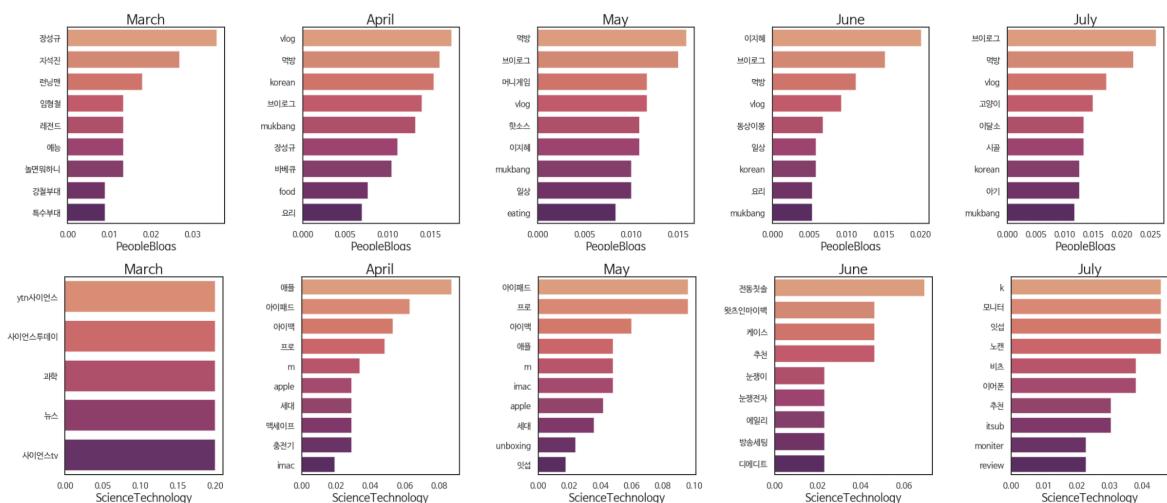
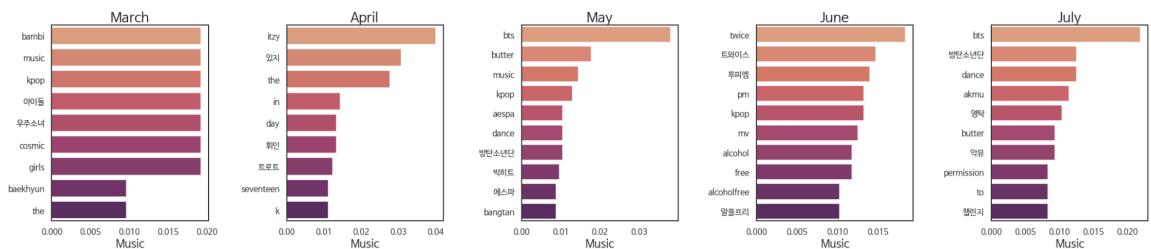
----- word count 데이터프레임 예시 -----

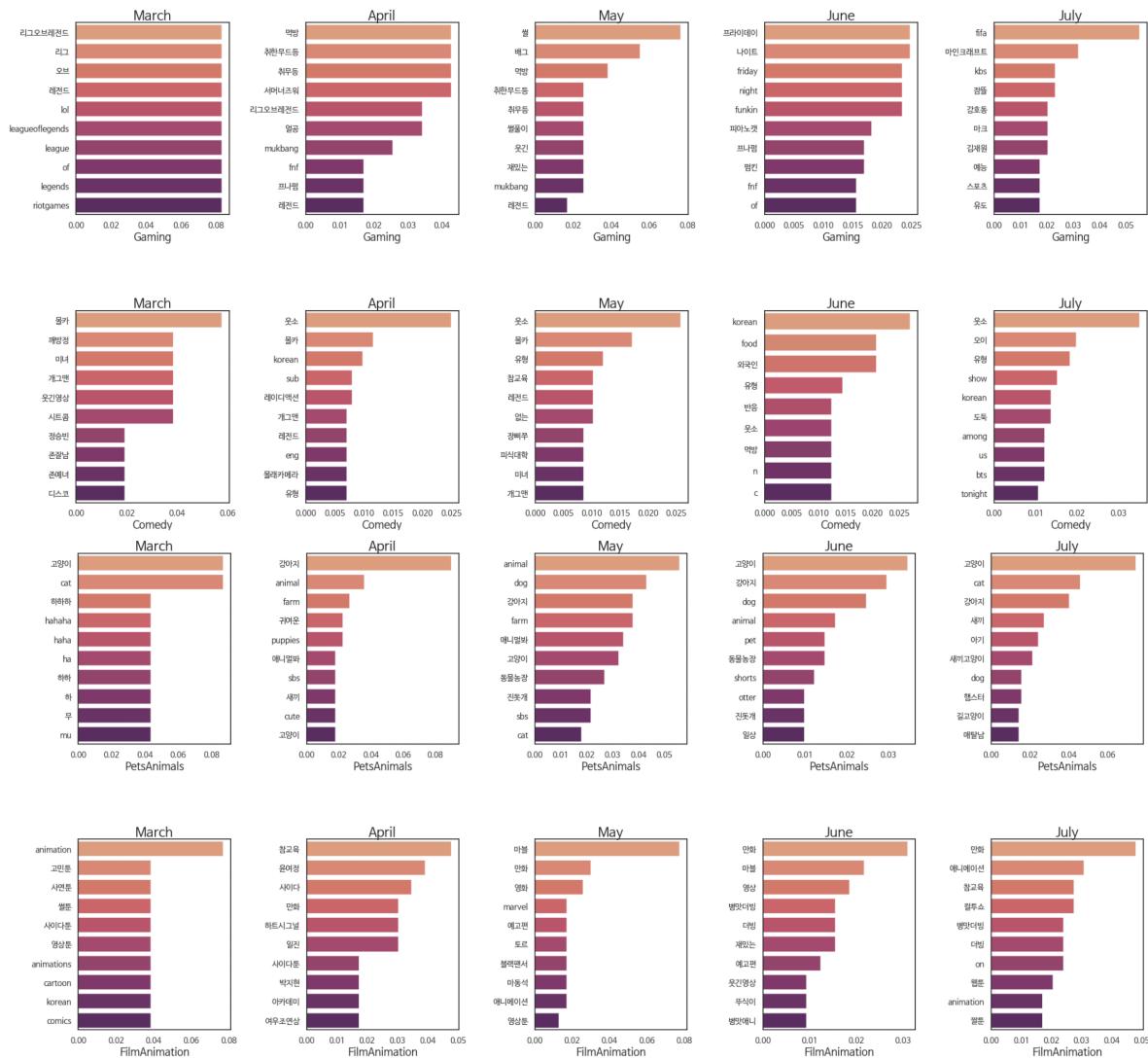
	word	word_in_docs	count	rank	percent	cum_percent
53	kbs	4	13	1.0	0.018310	0.018310
255	해리	1	12	2.0	0.016901	0.035211
100	tv	2	10	3.0	0.014085	0.049296
410	미스트롯	1	10	4.0	0.014085	0.063380
93	world	2	8	5.0	0.011268	0.074648
271	놀토	1	8	6.0	0.011268	0.085915
67	먹방	3	7	7.0	0.009859	0.095775
120	korea	2	6	8.0	0.008451	0.104225
95	eng	3	6	9.0	0.008451	0.112676
97	k	3	5	10.0	0.007042	0.119718

(3) 시각화

- (우측방향으로) 동일한 카테고리를 월별로 구분
- (하측방향으로) 동일한 월을 카테고리로 구분
- X 축 = 태그(단어)가 차지하는 퍼센티지
- Y 축 = 태그(단어)







2. 새로운 지표 개발

왜 이 동영상이 인기동영상 섹션에 올랐을까?라는 질문으로 시작한다.

1) Engagement 지표들의 통계치 확인

먼저, 인기동영상 섹션에 오른 시점의 engagement 지표들의 통계치를 확인한다.

on_views on_likes on_dislikes on_comments on_channel_subscribers on_channel_total_videos

	2544	2544	2544	2544	2544	2544
count	2544	2544	2544	2544	2544	2544
mean	956679	55827	537	9285	1792517	9608
std	3248054	329941	2098	117771	5634920	43962
min	53297	423	5	1	499	1
25%	238498	4319	89	520	165750	121
50%	429635	7915	162	1161	482500	329
75%	831281	16929	317	2720	1312500	1265
max	97276666	8097173	37349	4625133	58900000	545577

- ⇒ 각 지표들의 표준편차가 큰 값을 가지는 상황이고, 이는 데이터의 흩어짐이 크다는 의미이다.
- ⇒ 같은 인기동영상일지라도 각각의 engagement 값의 크기에는 큰 차이가 있기 때문에 이를 아무런 가공 없이 인기동영상의 기준으로 활용하기에는 무리가 있다고 판단한다.
- ⇒ 또한, 채널의 전체 비디오 수와 채널 구독자 수는 인기동영상에 오르는데 큰 영향을 주지 못한다고 말할 수 있다. 이유는, 비디오의 수가 많은 적은 혹은 구독자가 많은 적은 다양한 비디오들이 인기동영상 섹션에 올라 갈 수 있다는 것을 통계치를 바탕으로 알 수 있기 때문이다.

2) Engagement 지표 1차 가공

각 engagement 지표를 활용하여 상대성을 가지는 지표를 만들어낸다.

- (1) duration = 영상 시간 (데이터 전처리 작업에서 초단위로 환산 완료)
- (2) view_per_sub = 구독자 대비 조회수
- (3) like_per_view = 조회수 대비 좋아요수
- (4) dislike_per_like = 좋아요수 대비 싫어요수
- (5) comment_per_view = 조회수 대비 댓글수

	category_name	duration	view_per_sub	like_per_view	dislike_per_like	comment_per_view
0	Entertainment	500.0	1.53656	0.16270	0.00293	0.00625
1	Entertainment	557.0	0.46820	0.02426	0.01699	0.00369
2	Entertainment	459.0	0.07848	0.00885	0.04056	0.00154
3	Sports	400.0	2.29773	0.00535	0.02591	0.00192
4	Sports	687.0	13.26887	0.00608	0.02917	0.00109
...
2539	Comedy	154.0	1.12654	0.00293	6.84968	0.01690
2540	Comedy	399.0	0.71826	0.01953	0.01363	0.00864
2541	Science & Technology	705.0	0.80974	0.01232	0.06310	0.00740
2542	Entertainment	687.0	0.34343	0.03057	0.02141	0.00225
2543	Music	266.0	0.55418	0.12393	0.00466	0.01785

2544 rows × 6 columns

3) 차원축소 / 2 차원 공간 특징벡터

- (1) 데이터 표준화 → 벡터 표현

위에서 1차 가공된 값들을 표준화(Normalization)하여 각 샘플을(각 행) 5 차원 벡터로 만든다.

즉, 이 벡터는 다음과 같은 값들로 5 차원에 표시된다.

(duration, view_per_sub, like_per_view, dislike_per_like, comment_per_view)

	category_name	duration	view_per_sub	like_per_view	dislike_per_like	comment_per_view
0	Entertainment	-0.264874	-0.067225	3.227000	-0.131906	0.366551
1	Entertainment	-0.157293	-0.087143	-0.200484	-0.078763	-0.094257
2	Entertainment	-0.342256	-0.094409	-0.582003	0.010325	-0.481263
3	Sports	-0.453612	-0.053034	-0.668656	-0.045048	-0.412862
4	Sports	0.088068	0.151510	-0.650583	-0.032726	-0.562264
...
2539	Comedy	-0.917909	-0.074869	-0.728570	25.746952	2.283582
2540	Comedy	-0.455500	-0.082481	-0.317589	-0.091463	0.796758
2541	Science & Technology	0.122041	-0.080775	-0.496094	0.095520	0.573554
2542	Entertainment	0.088068	-0.089469	-0.044262	-0.062057	-0.353461
2543	Music	-0.706522	-0.085540	2.267136	-0.125367	2.454584

2544 rows × 6 columns

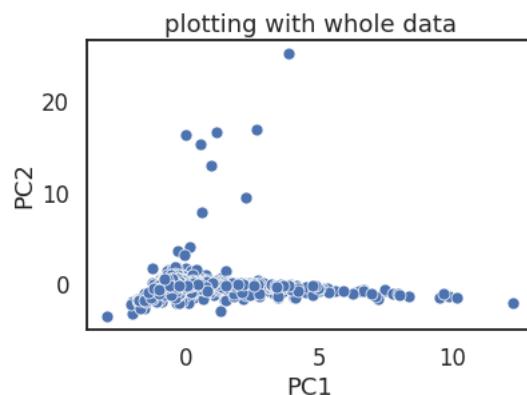
(2) PCA (Principle Component Analysis)

- PCA 를 통해 5 차원의 벡터를 2 차원으로 차원축소 진행
- (PC1, PC2)의 2 차원으로 표현되는 벡터 생성

	category_name	PC1	PC2
0	Entertainment	2.543516	-0.301186
1	Entertainment	-0.167725	-0.005480
2	Entertainment	-0.649379	0.196288
3	Sports	-0.632638	0.245661
4	Sports	-0.843040	0.180931
...
2539	Comedy	0.948711	12.970848
2540	Comedy	0.427659	0.014593
2541	Science & Technology	0.008870	-0.101485
2542	Entertainment	-0.295668	-0.084987
2543	Music	3.399126	-0.316527

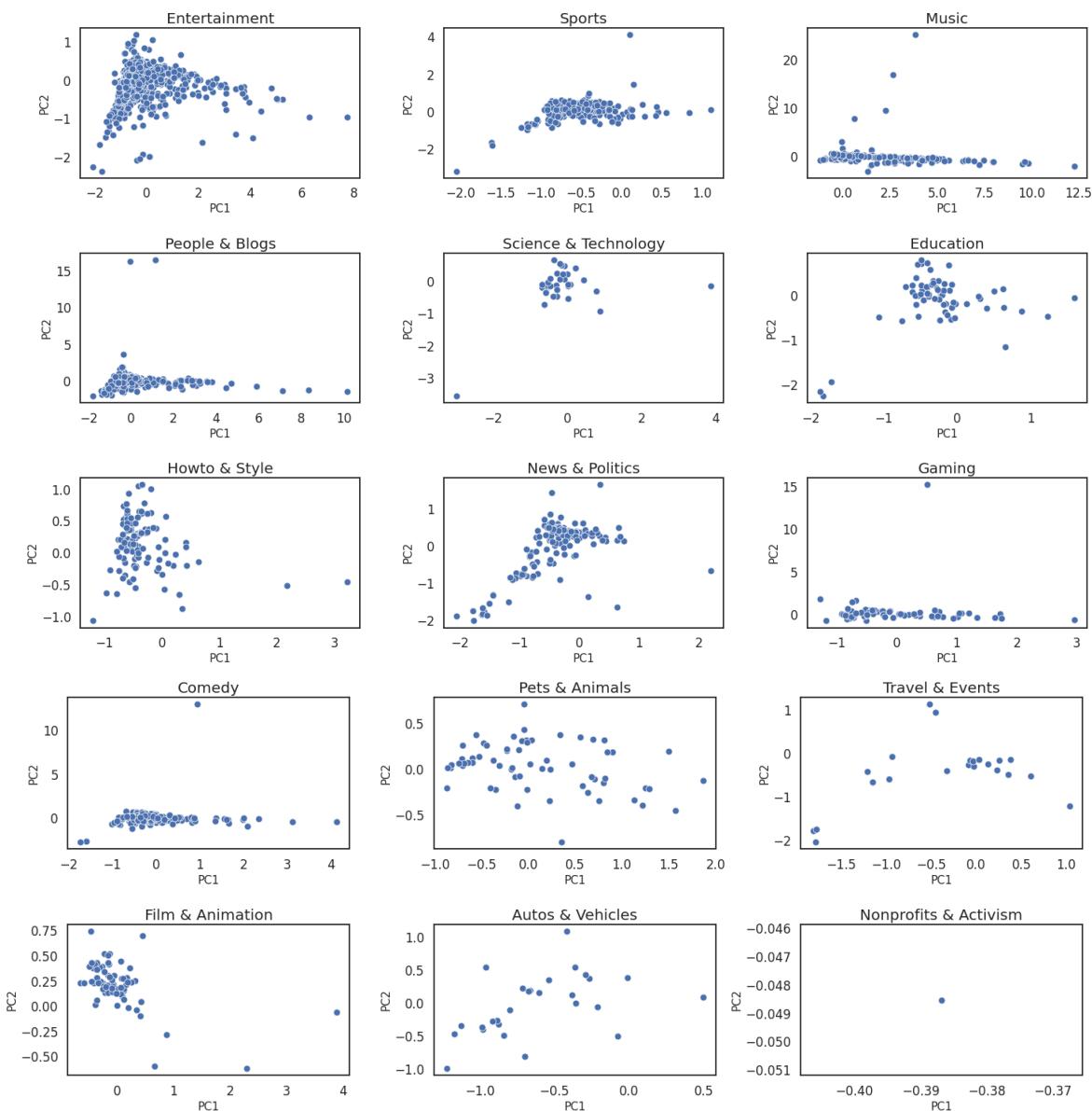
2544 rows × 3 columns

- 이를 2 차원 평면위에 나타내어 인기동영상이 가지는 공간특징을 확인



- 전체데이터를 대상으로 scatter plot 을 그려보면 위와 같이 데이터가 어느정도 밀집되어 있는 형태를 확인할 수 있다.
- 인기동영상이 아닌 다른 동영상들에 대한 데이터가 있었다면, 이를 동일한 지표로 차원축소를 진행할 경우 위와는 다른 위치에 밀집 혹은 밀집되지 못 할 것으로 ‘예상’한다.

- 카테고리별로도 2 차원 평면에 위치하는 데이터를 확인한다.



- 대부분의 카테고리에서 각각의 벡터들이 특정 위치에 밀집된 형태를 확인할 수 있다.
- 역시, 각 카테고리 별로 인기동영상이 아닌 다른 동영상에 대한 데이터가 있었다면 이들은 해당 카테고리 내의 인기동영상과는 다른 위치에 밀집 혹은 밀집되지 못하는 형태를 가질 것으로 예상할 수 있다.

3. 결론

Engagement 지표를 활용하여 인기동영상들의 공간적 특징을 알 수 있는 (2 차원의)지표를 개발하였다.