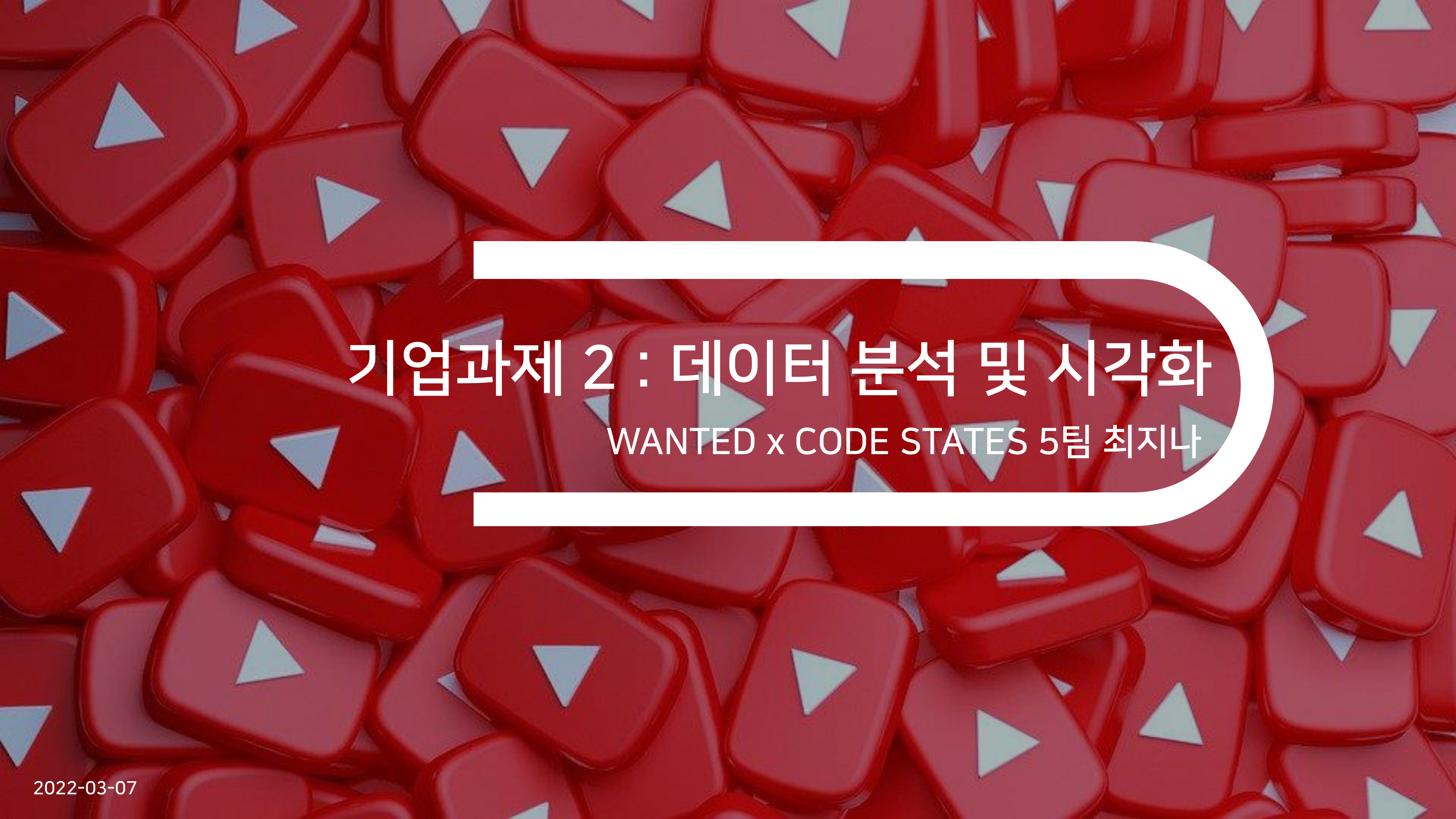


기업과제 2 : 데이터 분석 및 시각화

WANTED x CODE STATES 5팀 최지나

2022-03-07

1. Environment & Library

1. Environment

- Google Colab Pro

2. Library

- pandas, numpy, seaborn, matplotlib
- os, sys
- sklearn
- eli5

2. EDA (Exploratory Data Analysis)

Column Name	NaN	Unique
video_id	0	2643
channel_id	0	940
published_date	0	127
category_name	0	15
duration	0	1200
tags	370	1978
description	40	2492
on_trending_date	0	127
off_trending_date	0	122
on_rank	0	50
off_rank	0	50
on_views	0	2642
off_views	0	2642
on_likes	0	2479
off_likes	0	2515

NaN : Exists only in 'tags' & 'description'

The reason for counting unique values :

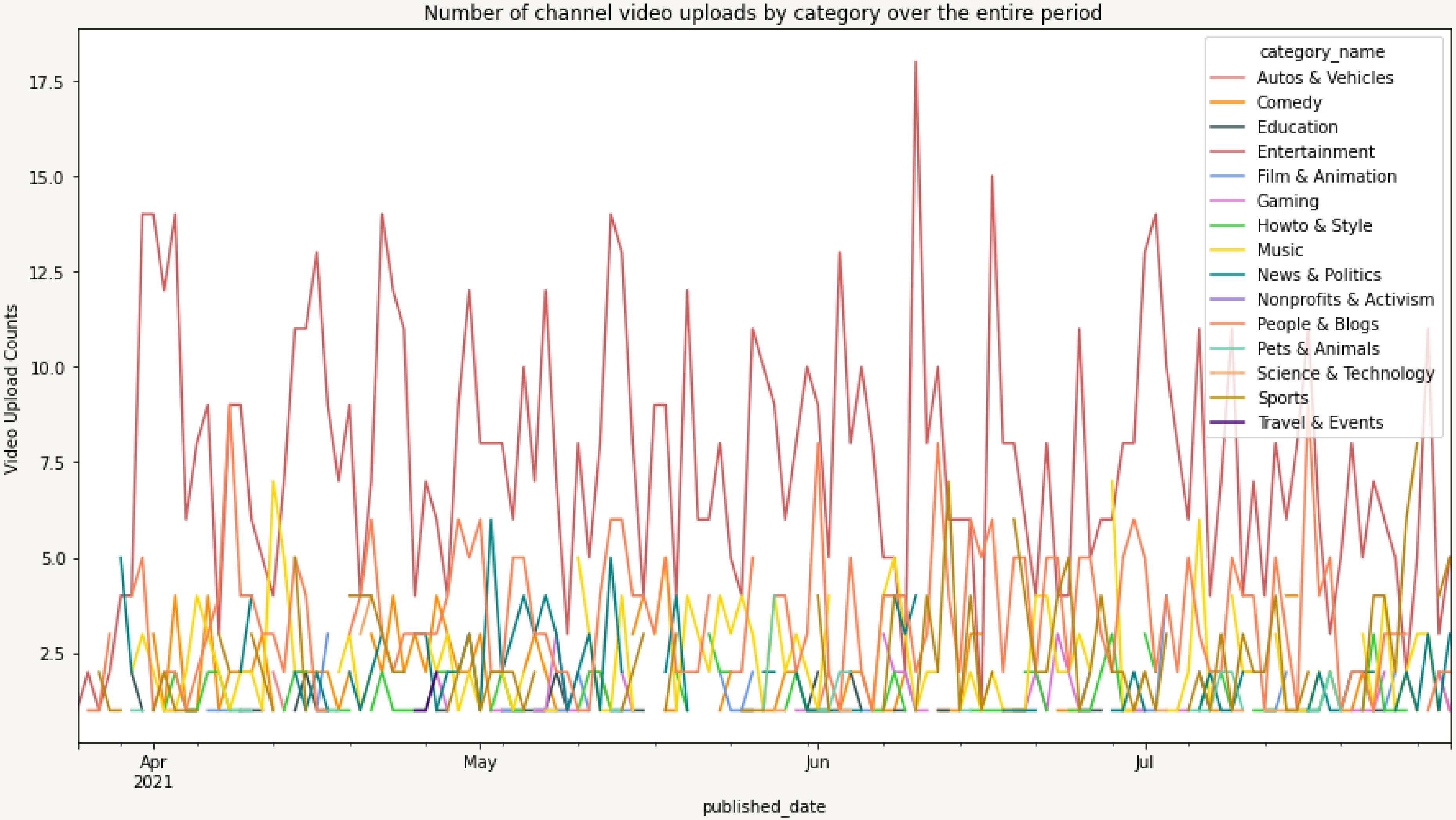
As the source of the data is based on the "Explore/Popular" section of the main page of 'YouTube', it is expected that the number of viewers or the video will be hung in that section according to social phenomena or trends.

In particular, in the case of a channel with a large number of viewers, as it is a professional channel that produces videos for the purpose of profit, it is determined that the videos uploaded here will appear in the section a lot.

Therefore, as a result of excluding duplicate channels, it is found that the videos published in the "Search/Popular" section were produced in 940 channels, which is about 35.6% of the 2,644 videos.

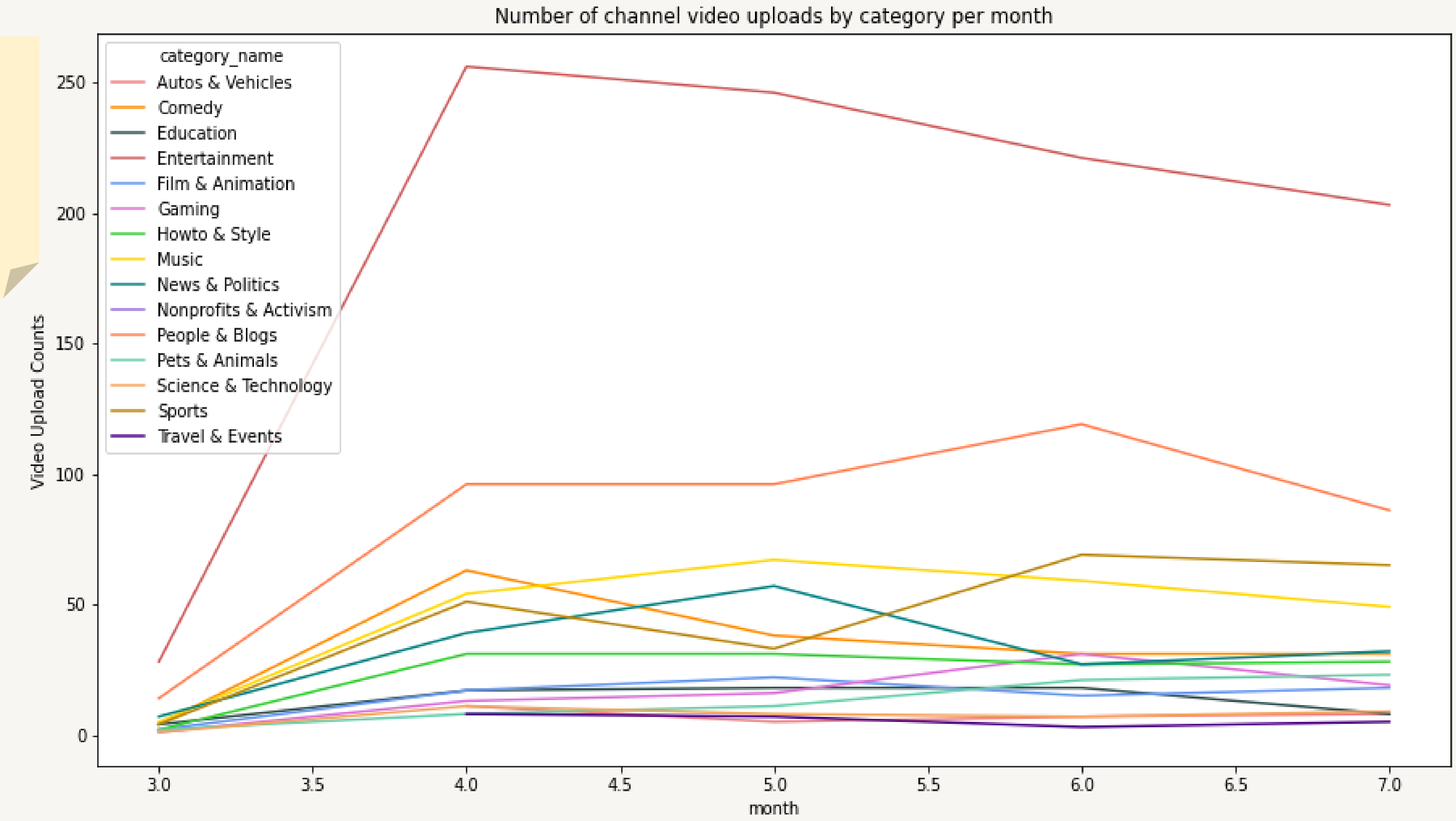
3. Visualization by Data Type (Q1)

- Videos were mainly uploaded to the categories of "Entertainment" and "people & blogs“.
- There were times when videos were not uploaded in some categories.

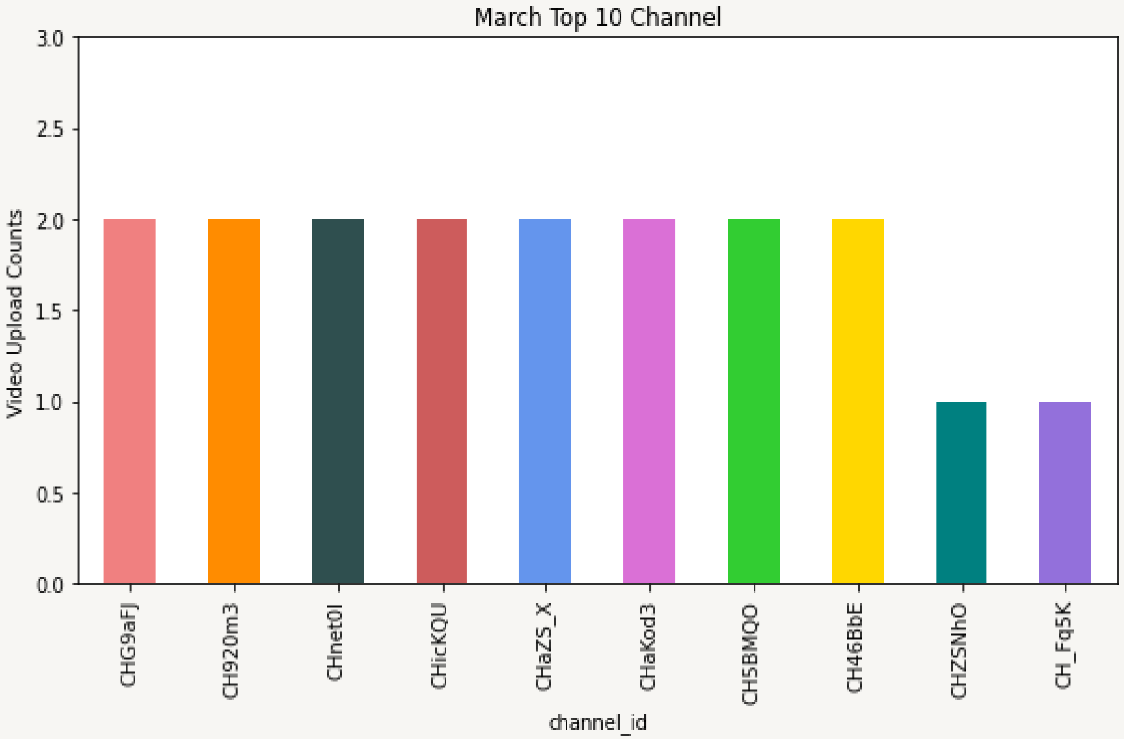


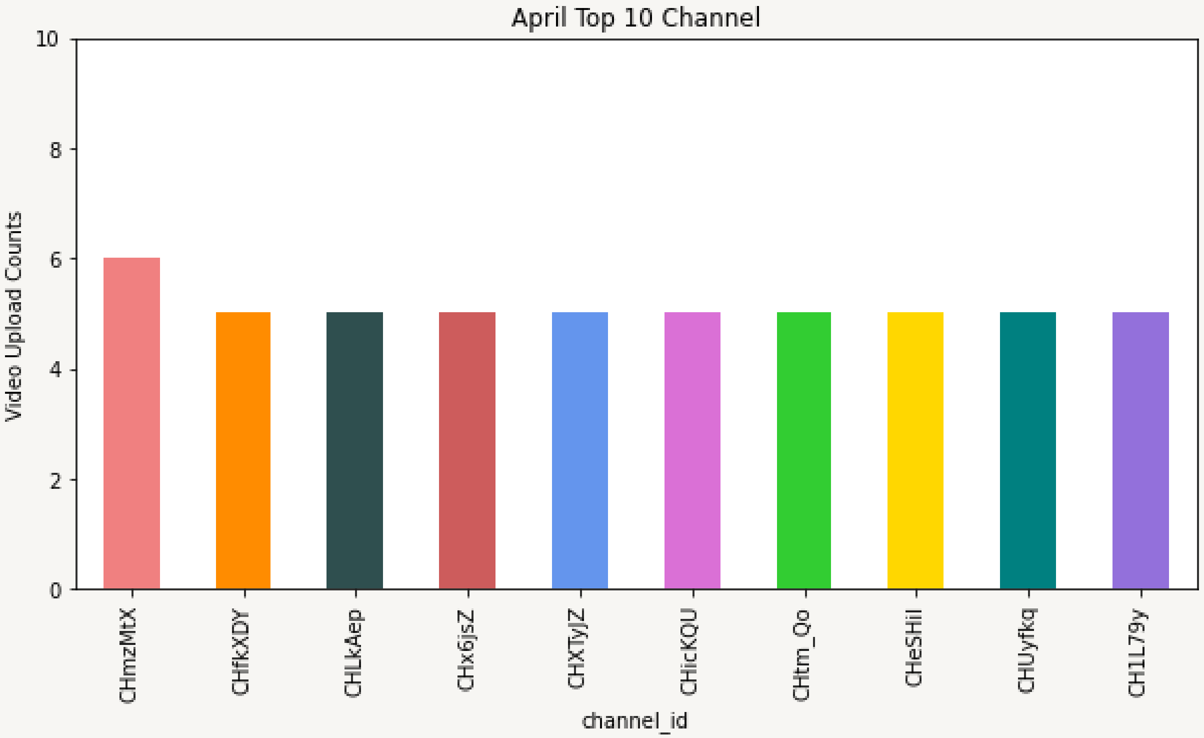
3. Visualization by Data Type (Q1)

- "Entertainment" and "Comedy" categories uploaded the most.



3. Visualization by Data Type (Q1)

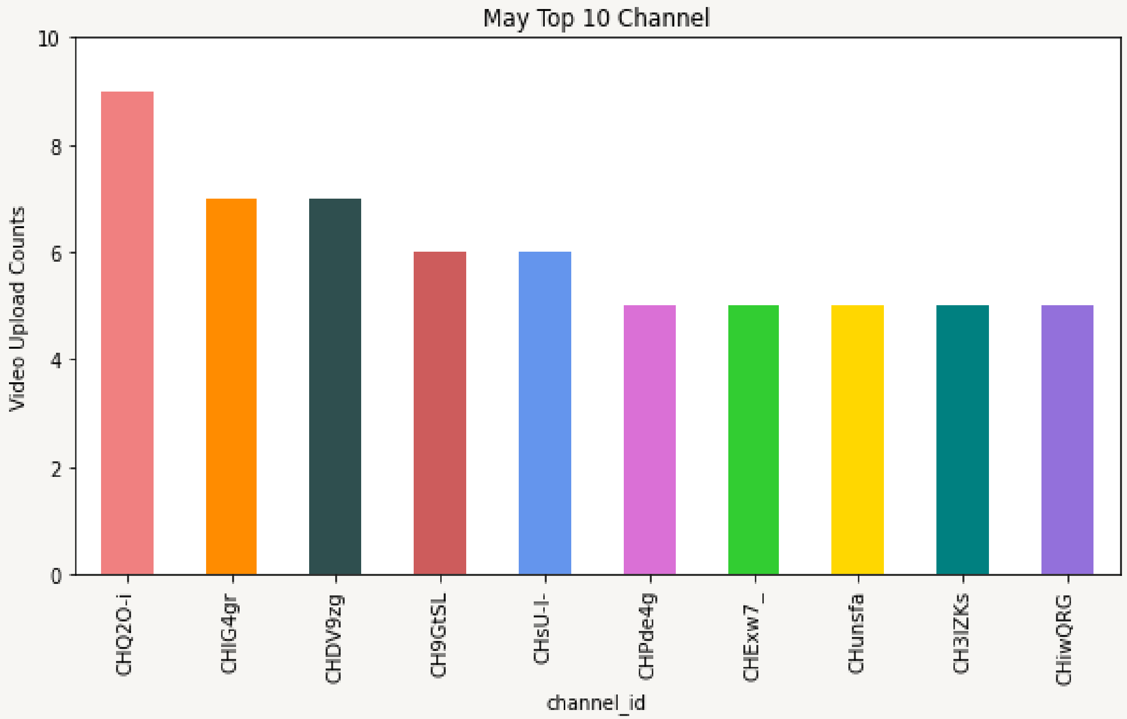

March Top 10 Channel

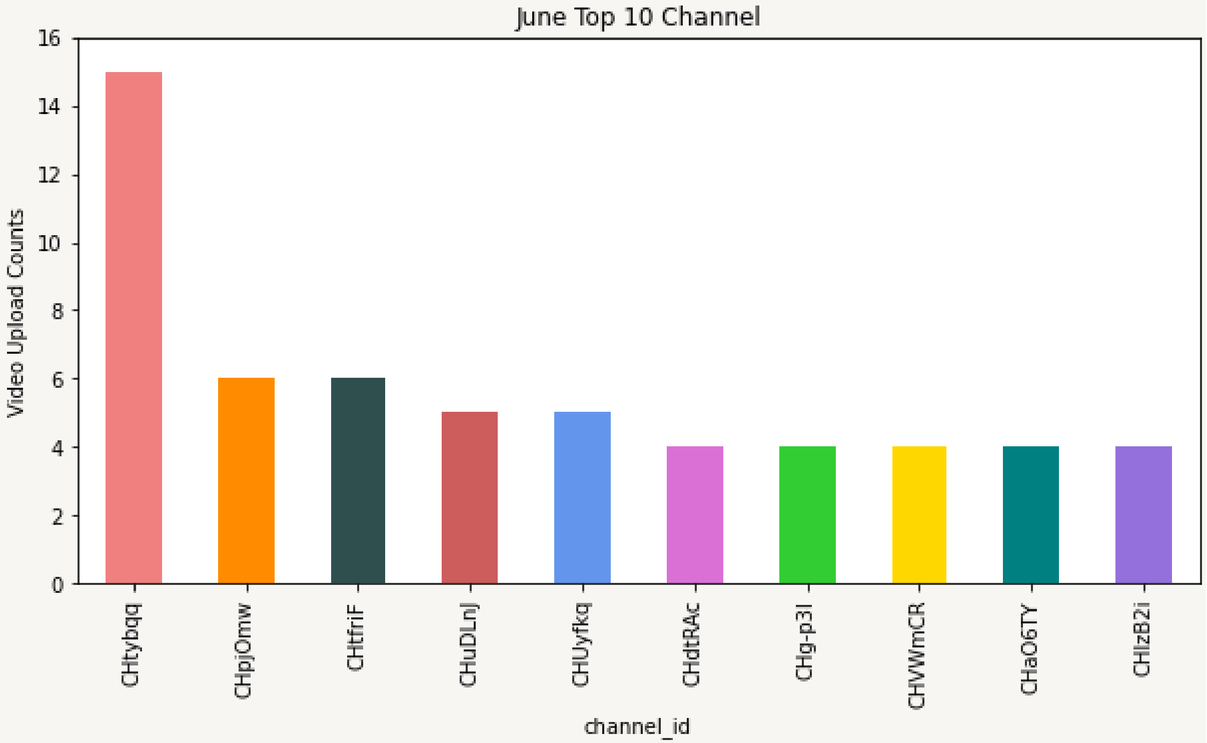

April Top 10 Channel

- In March, data was collected from the end of the month, and two or one upload became the maximum number of uploads.

- In June, compared to other periods, there is a channel that posted the most prominently the video.

3. Visualization by Data Type (Q1)

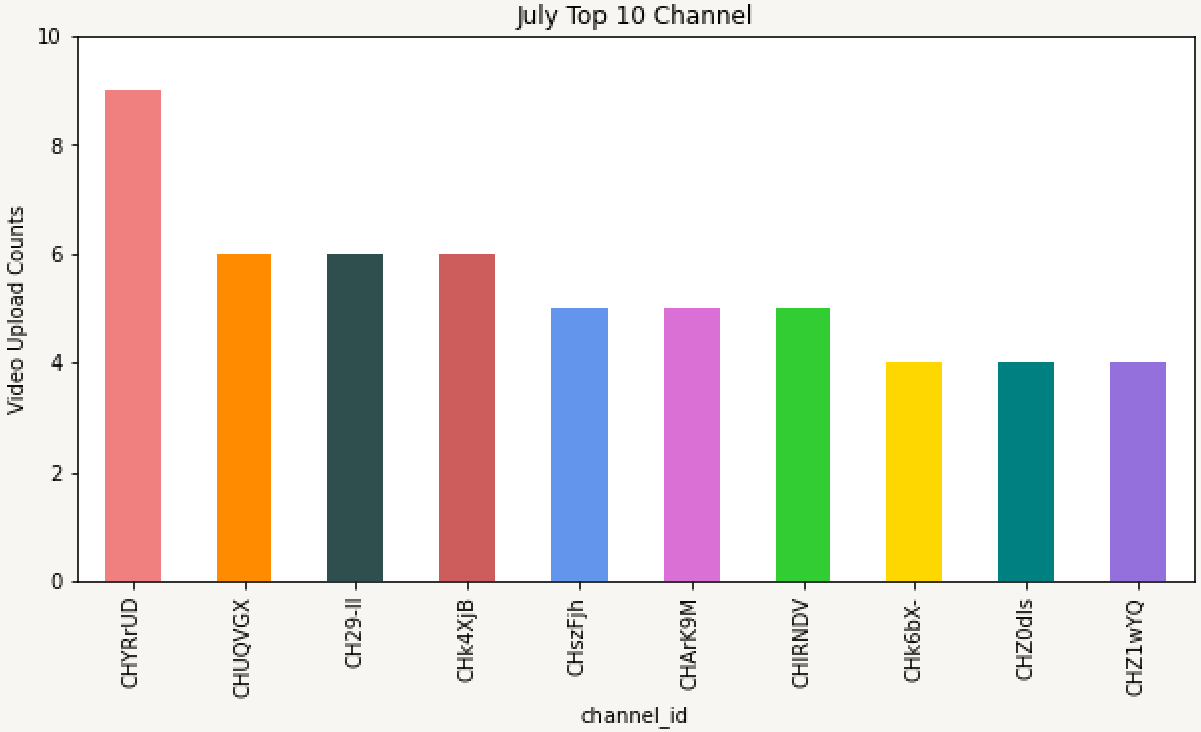




- It was confirmed that there was a trend of uploading slightly more in May than in April.

- In June, compared to other periods, there is a channel that posted the most prominently the video.

3. Visualization by Data Type (Q1)



- July is similar to May.
- Generally, between 4 and 6 are uploaded.

Top 5 Channels by Week

Weekly	Channel	Num of Videos
2021-03-22	CHnet0	2
	CH2qVOO	1
	CH7Krez	1
	CHCA4-e	1
	CHGiqkg	1
2021-03-29	CHaKod3	3
	CHaZS_X	3
	CH0PsUG	2
	CHcTWmz	2
	CHmzMtX	2
2021-04-05	CHmzMtX	3
	CHtm_Qo	3
	CHIA-LP	3
	CHUyfkq	2
	CHFL1sC	2

3. Visualization by Data Type (Q1)

Top 5 Channels by Week

Weekly	Channel	Num of Videos
2021-04-12	CHx6jsZ	5
	CHeSHil	3
	CHMEbRp	3
	CHmONdw	3
	CH-BqPA	2
2021-04-19	CH5lda8	3
	CHiwQRG	3
	CHGsJRp	3
	CHe9f9M	2
	CHoPTla	2
2021-04-26	CHQ2O-i	5
	CHlG4gr	4
	CHkxbPw	3
	CHunsfa	3
	CHLJNGm	2

Weekly	Channel	Num of Videos
2021-05-03	CHDV9zg	4
	CHQ2O-i	3
	CHWlV3L	3
	CHl7MKZ	3
	CHM31rB	2
2021-05-10	CHoXoV4	3
	CH_Fxf0	3
	CHQ2O-i	3
	CHlG4gr	2
	CHYSjF7	2
2021-05-17	CH3lZKs	3
	CH4DnB5	3
	CH9GtSL	2
	CHzlOlS	2
	CH8WoHU	2

3. Visualization by Data Type (Q1)

Top 5 Channels by Week

Weekly	Channel	Num of Videos
2021-05-24	CHLkAep	3
	CHPde4g	3
	CH-VbFg	2
	CHIUfR-	2
	CHld0ct	2
2021-05-31	CH78PMQ	3
	CHuDLnJ	3
	CHpjOmw	3
	CHIA-LP	3
	CHOHM2N	2
2021-06-07	CHUyfkq	5
	CHtybqq	4
	CHviI9l	3
	CHpjOmw	3
	CHkinYT	2

Weekly	Channel	Num of Videos
2021-06-14	CHtybqq	4
	CHXTyJZ	3
	CH6lNIb	2
	CHPx-7A	2
	CHdhukF	2
2021-06-21	CHtybqq	5
	CHdtRAc	2
	CHweOkP	2
	CHPhHBE	2
	CHmQdC1	2
2021-06-28	CHuKdaT	3
	CHnx4Fi	2
	CHtybqq	2
	CHk6bX-	2
	CHOHM2N	2

3. Visualization by Data Type (Q1)

Top 5 Channels by Week

Weekly	Channel	Num of Videos
2021-07-05	CH29-Il	3
	CHtm_Qo	3
	CHd1TDy	2
	CHnLeqv	2
	CHoxT1k	2
2021-07-12	CH8-Th8	3
	CHLkAep	3
	CHoLrcj	2
	CH0imOR	2
	CH29-Il	2
2021-07-19	CHYRrUD	4
	CHArK9M	4
	CHk4XjB	3
	CHaKod3	2
	CHdWgRS	2

Weekly	Channel	Num of Videos
2021-07-26	CHYRrUD	5
	CH-FQUI	3
	CHcQTRi	3
	CHk4XjB	2
	CHkinYT	2

3. Visualization by Data Type (Q1)

Tag keyword ranking by category by March

Category	Keyword Rankikng
Comedy	('웃긴영상', '시트콤', '몰카', 2), ('#깨방정', '#정승빈', '#몰카', '#미녀', '#개그맨', '#존잘남’, '#존예녀’ ,1)
Education	('조승연’, '조승연의 탐구생활', '조승연작가’, '조승연 작가', '럭키', '럭키 인디아’, '럭키인디아', '럭키 인디아 레스토랑', '조승연 럭키’, ’채널 354’, 1)
Entertainment	('유재석', 'KBS', 4), ('아이유', 'kbs', 'eng', 'idol', 3), ('라일락', 'IU', 'LILAC', '런닝맨', 2)
Film & Animation	(‘고민툰’, ‘사연툰‘, ‘썰툰’, ‘사이다툰’, ’영상툰’, ’animation’, ’animations’,’Cartoon’, ’Korean animation’. ’comics’ ,1)
Gaming	(‘리그오브레전드’,’리그 오브 레전드‘, ‘LoL’, ‘’Leagueoflegends’, ‘’League of Legends’, ‘’Riotgames’, ‘라이엇 게임즈’ ,1)
Howto & Style	(‘사나고’,’3D펜‘, ‘3Dpen’,’만들기’, ‘making’,’3d프린터‘, ‘3Dprinting’,’계란부침‘, ‘계란부침 만들기‘,’계란부침 레시피‘,1)
Music	('music', '아이돌', 2), ('BAEKHYUN', 'Bambi - The 3rd Mini Album', 'Bambi', '딩고뮤직', 'dingo’, 'dingomusic', 'kpop', 'live', 1)
News & Politics	('박수홍', 4), ('출연료', 2), ('횡령', 2), ('MBN', '오열', '다홍이', '뉴스파이터', '김명준앵커', 'SNS', '가족’, 1)
People & Blogs	('강철부대', '특수부대’, '육군특수전사령부', '특전사', 'UDT해군특수전전단', 'UDT', '제707특수임무단', '707’, '대테러 부대', '해병대’, 2)
Pets & Animals	('고양이', 'cat', 2), ('하하하', 'hahaha', 'haha ha', '하하 하', '무', 'Mu', '강아지', 'puppy’ ,1)
Science & Technology	('YTN사이언스', '사이언스투데이’, '과학','뉴스', 사이언스TV’, 1)
Sports	('골프', '골프레슨', 'golf', 'golf lesson', '드라이버', '아이언', '골프맨', '조윤성프로', '골프스윙’, '비거리’ ,1)

- The reason why I did not delete the spaces in the tag is to focus on the keyword of the tag itself.

3. Visualization by Data Type (Q1)

Tag keyword ranking by category by April

Category	Keyword Rankikng
Comedy	('몰카', 12), ('몰래카메라', 8), ('개그맨', '레전드', '웃소', 6), ('보물섬', '장난', 'prank', '동네놈들', '웃긴영상', 5)
Education	('주식', '투자', 3), ('아이템', '아이템의인벤토리', '드립', '유래', '어디서', '재테크', '카카오', '삼성전자', 2)
Entertainment	('예능', 24), ('먹방', 22), ('브레이브걸스', 19), ('롤린', '유재석', 18), ('유나', 17), ('쁘걸', '유정’, '민영', 16), ('은지', 15)
Film & Animation	('만화', '더빙', '병맛더빙', 4), ('영상툰', '짤툰', '애니메이션', '웹툰', '병맛', 'yt:cc=on', 3), ('사연툰', 2)
Gaming	('먹방', 3), ('리그오브레전드', '얼공’, '#리그오브레전드', '#롤', '#괴물쥐', '#원딜', '#트위치', 2), ('넷마블', '모바일’, 1)
Howto & Style	('사나고', '만들기', 4), ('3D펜', '3Dpen', 'making', '3d프린터', '3Dprinting', 3), ('계란', '달걀', '계란먹는법', 2)
Music	('SEVENTEEN', 'BTS', 8), ('セブチ', '세븐틴', 7), ('방탄소년단', 'BANGTAN', 6), ('HIPHOP’, '알엠', 'RM', '슈가', 5)
News & Politics	('뉴스', 8), ('news', '주식', 'ETF', 5), ('윤여정', 'News Network', 'SBS VIDEOMUG', 'VIDEOMUG', '비디오머그', 'KBS’, 4)
People & Blogs	('먹방', 13), ('브이로그', 10), ('vlog', 8), ('korean', 7), ('맛집', 7), ('요리', 6), ('mukbang', 5), ('중소기업드라마', 'streetfood’, 'korean street food', 4)
Pets & Animals	('브이로그', '일상', '동물병원', 2), ('애니멀봐’, '동물농장', 'TV동물농장', '동물농장 애니멀봐', '애니멀봐 동물농장', 'sbs animal farm', 'sbs animal eng sub', 1)
Science & Technology	('아이맥’, '애플', 4), ('Apple', '에어태그’, '아이패드 프로 5세대', 3), ('Apple Event', 'iMac’, '아이패드', '아이패드프로', '아이패드프로5세대', 2)
Sports	('손흥민', 11), ('토트넘', 10), ('축구', 7), ('스포티비’, 'SPOTV', 6), ('맨유', '무리뉴', '하이라이트', 5), ('김국진골프', '황의조', 4)
Autos & Vehicles	('K8', 5), ('자동차', '기아', '그랜저', '시승기', 4), ('Kia', '기아K8', 'K7', '비린내PD', '솔님', 2)
Travel & Events	('베트남', '국제커플', '에티오피아’, '에티오피아 여행’, '아프리카 여행', '필리핀', '세부', '막탄', 2), ('여행', '꾸이년', 1)

3. Visualization by Data Type (Q1)

Tag keyword ranking by category by May

Category	Keyword Rankikng
Comedy	('몰카', 9), ('피식대학', '더블비', '참교육', 5), ('웃긴영상', '몰래카메라', '핫소스', '보물섬', '개그맨', '아프리카tv', 4)
Education	('복지정보', '유용한정보', '한시생계지원금', 2), ('아이템', '아이템의인벤토리', '드립', '유래', '어디서', '알테니스킵', '풀메로병', 1)
Entertainment	('tvN', 19), ('먹방', 16), ('방탄소년단', 14), ('피오', 12), ('패션', '라면', '이수근', 11), ('BTS', '놀면뭐하니', '나PD', 10)
Film & Animation	('영상툰', 6), ('썰툰', 5), ('애니메이션', 4), ('호돌이영상툰', '일상', '마블', '마블 영화', '블랙위도우’, ’짤툰', '만화', 3)
Gaming	('먹방', 5), ('웃긴', '재밌는', 4), ('아프리카TV', '혜안', '병맛', '배틀그라운드', 3), ('취한무드등', '취무등', '취한무드등 썰', 2)
Howto & Style	('요리', 4), ('간단요리', '양배추', '꿀팁', '쉬운요리', '초보요리', '함께해요 맛나요리', '한식', 3), ('간단한 요리', '하체비만', 2)
Music	('BTS', 15), ('방탄소년단', 13), ('aespa', 11), ('BANGTAN', 10), ('음악', '라이브', '에스파', 9), ('kpop', '아이돌', '방탄', 7)
News & Politics	('뉴스', 20), ('손정민', 13), ('의대생', 7), ('CCTV’, '코로나19', 'JTBC', 'JTBC NEWS', '뉴스룸', '손석희’, 'newsroom', 6)
People & Blogs	('먹방', 16), ('머니게임', 'mukbang', 6), ('진용진', '핫소스', '브이로그', 5), ('쿡방', '요리용디', '웅이', '푸드파이터', 4)
Pets & Animals	('animals’, 'animal farm', '애니멀봐', '동물농장', '동물영상', 6), ('강아지', 'cute dogs', 4), ('진솔쓰', 'jinsolss', 'jinsols', 3)
Science & Technology	('애플', 3), ('아이맥', 'apple', '아이패드 프로 12.9', 2), ('귀상어', '머리’, '상어', ‘진화', '부력', '로렌치니기관', 1)
Sports	('메이저리그', '류현진', '토론토', '토론토블루제이스', 6), ('mlb', '조미예’, '블루제이스', '조미예의 MLB현장', 5), ('축구', '맨시티', 4)
Autos & Vehicles	('한문철', '블랙박스', 3), ('과실비율', '몇대몇', '사고’, '교통사고', '영상', '블박', '블박세', '블랙박스로 본 세상', 2)
Travel & Events	('길거리음식', '맛집', '한국 길거리음식', 'shorts', 'waffle', 'street food', 'korean street food', ‘EBS', '해외여행', '관광', 2)
Nonprofits & Activism	('불교', '즉문즉설', '법륜스님', '정토회’, 'buddha', 'buddhism', 'pomnyun', '스님의주례사', '엄마수업', '깨달음', 1)

3. Visualization by Data Type (Q1)

Tag keyword ranking by category by June

Category	Keyword Rankikng
Comedy	('반응', 6), ('라면', 5), ('Korean', '영국남자', '영국', '조쉬', '올리', 'Josh', 'Ollie', 'KoreanEnglish', 4)
Education	('아이템', '아이템의인벤토리', '드립', '유래', '어디서', 3), ('조승연', '조승연의 탐구생활', '조승연 작가', '과학실험', '건나물TV', 2)
Entertainment	('먹방', 27), ('예능', 20), ('레전드', 'SBS', 15), ('유재석', 13), ('kpop', 12), ('개그', 11), ('아이돌', 10), ('하하', 10), ('브레이브걸스', 9)
Film & Animation	('병맛더빙’, '만화', '더빙', 5), ('웃긴영상', '뚜식이', '병맛애니', '웃긴만화', '꿀잼', '볼만한애니', '재밌는애니', 3)
Gaming	('프나펌', '프라이데이 나이트 펌킨', 7), ('fnf', 'friday night funkin', 6), ('프나펑', 5), ('프라이데이 나이트 펑킨', '모바일게임', 4), ("friday night funkin'", '시엘', '프나펌 한글자막', 3)
Howto & Style	('한식', ‘밑반찬', 4), ('반찬', '영자씨의 부엌', 'Korean recipes', 'Korean mom', 'K-food', 'Korean style food', 'banchan', '자취음식', 3)
Music	('음악', '트와이스', 7), ('kpop', 'TWICE', 6), ('dingo', '딩고', '알콜프리', 5), ('Brave Girls', '브레이브걸스', 'KPOP', 4)
News & Politics	('뉴스', 6), ('MBC뉴스', '뉴스데스크', ‘뉴스투데이', 'news', '경찰', 'News Network', '광주', 3), ('KBS’, '인터뷰', 2)
People & Blogs	('먹방', 15), ('브이로그', 12), ('맛집', 9), ('일상', 8), ('vlog', 'mukbang', 7), ('이과장', '중낳괴', '좋좋소', '중소기업', 6)
Pets & Animals	('고양이', 6), ('강아지', 5), ('pet', '아리랑', ‘포메라니안', '동물', 'cat’, 3), ('Shorts', 'shorts', 'dog video', 2)
Science & Technology	('Apple', 'Apple Event', 'Apple Keynote', 'Apple Special Event', 'Apple WWDC', 'June', 'Developers’, 'World Wide Developers Conference', '2021', 'WWDC', 1)
Sports	('축구', 30), ('손흥민', 25), ('football', 유로', '이동국', 16), ('EURO', 'EURO2020', 'tvn', 'xtvn', '메시', 15)
Travel & Events	('자동차', 3), ('자동차꿀팁', ‘운전', '초보운전', '기아', 2), ('자동차리뷰', 꿀팁', '썬팅', '방어운전', '안전운전', 1)
Nonprofits & Activism	('캠핑카', '캠핑', '카라반', '모터홈', 2), ('차박', '농막', ‘이동주택', '트럭캠퍼', '캠퍼', '봉고캠핑카', 1)

3. Visualization by Data Type (Q1)

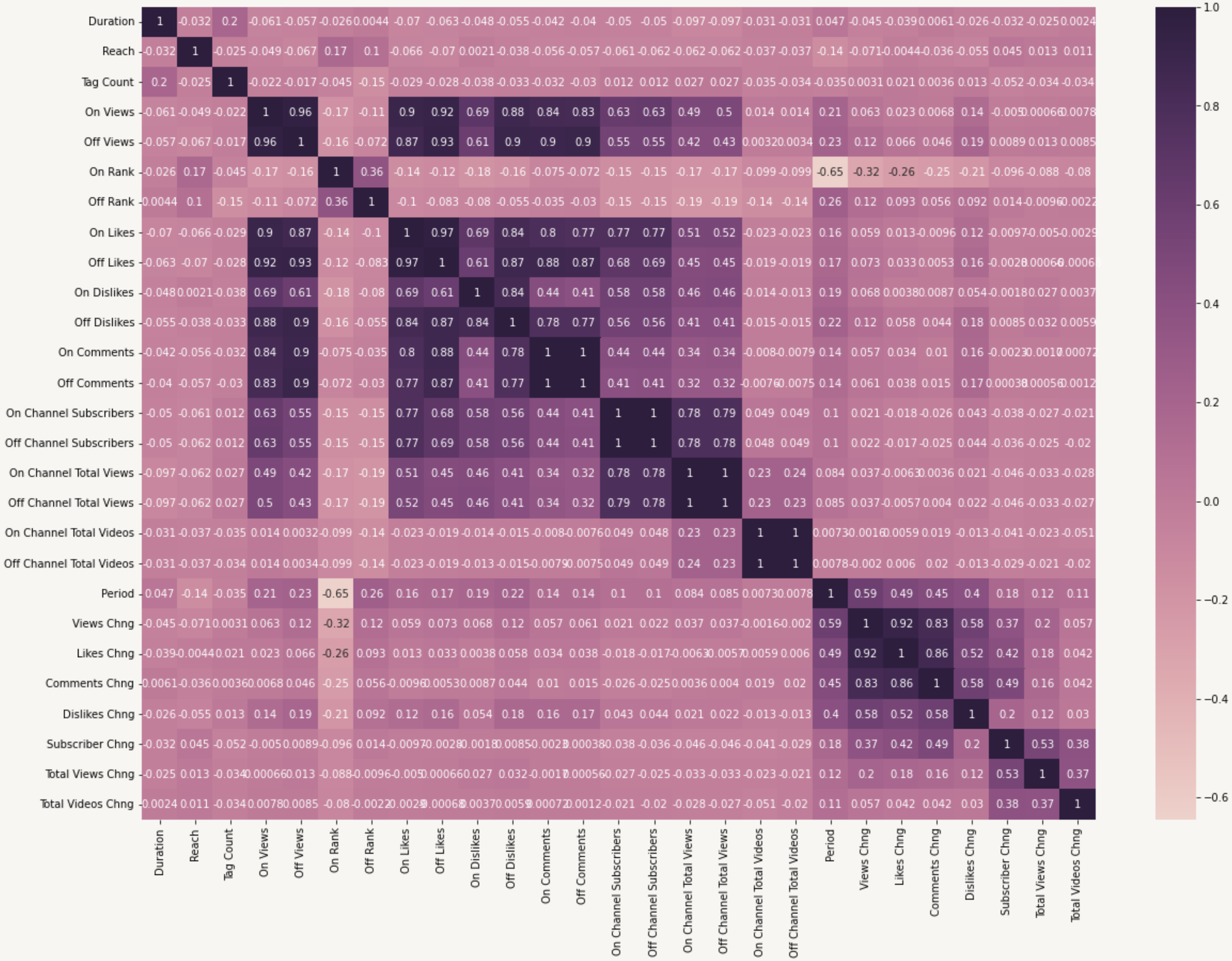
Tag keyword ranking by category by July

Category	Keyword Rankikng
Comedy	('어몽어스애니', '어몽어스애니메이션', 'among us animation', '웃소', 4), ('꼰대희', '김대희', '밥묵자', '장동민', '신봉선', '유세윤', 3)
Education	('사물궁이', '호기심', '궁금증', '잡학', '지식', 3), ('과학', 2), ('매일경제', '부동산', '매부리TV', '재테크', 1)
Entertainment	('유재석', 19), ('예능', 17), ('SBS', 14), ('먹방', 12), ('레전드', 11), ('kpop', 10), ('하하', 10), ('tvN', '런닝맨', 9), ('Diggle', 8)
Film & Animation	('만화', 10), ('애니메이션', 9), ('병맛더빙', '더빙', 7), ('웹툰', 6), ('짤툰', '병맛', 'yt:cc=on', 5), ('animation', 4), ('ㅋㅋㅋ', 3)
Gaming	('먹방', 4), ('혜안', '마인크래프트', 3), ('kbs', 'KBS', 'kbs esports', 'esports kbs', '이스포츠', '이스포츠 케이비에스', '케이비에스', 2)
Howto & Style	('반찬', '요리', 8), ('한식', 7), ('레시피', 6), ('간단요리', 'recipe', '만들기', 4), ('함께해요 맛나요리', '사나고', '간식', 3)
Music	('BTS', 10), ('방탄소년단', 7), ('알엠', 'RM', '지민', '정국', 'JIN', 'AKMU', 'NEXT EPISODE', 'mv', 4)
News & Politics	('도쿄올림픽', 10), ('뉴스', 9), ('올림픽', 6), ('news', 'News Network', 4), ('영상', 'SBS', 'SBS NEWS', '에스비에스', '특파원보고 세계는 지금', 3)
People & Blogs	('먹방', 14), ('브이로그', 12), ('vlog', 9), ('국제커플', '자취', 'mukbang', '일상', 5), ('real sound', '리얼사운드', '배말랭', 4)
Pets & Animals	('cat', 10), ('냥줍', 7), ('고양이', '새끼 길고양이', 'chat', 'alley cat', 6), ('dog', '포메라니안', '매탈남', '매탈남 고양이', 5)
Science & Technology	('unboxing', '해부', 2), ('4K', '4K 모니터', '주연테크', '가성비 모니터', '갓성비', '4K 모니터 입문용', '입문용', '언박싱', 1)
Sports	('도쿄올림픽', 22), ('올림픽', '스포츠', 14), ('SPOTV', '스포티비', '축구', '이강인', 'KBS', 8), ('금메달', '김제덕', 7)
Travel & Events	('자동차', 4), ('리뷰', '시승기', '벤츠', '벤츠튜닝', '중고차관리', '수입차정비', '벤츠 가솔린', 수입차수리', '메르카바', 2)
Nonprofits & Activism	('러시아 여행', 4), ('러시아', 3), ('모스크바 여행', '세계여행', 2), ('러시아 백신', '러시아 백신 후기', '스푸트니크', '여행유튜버', '모스크바', '이즈마일로보 시장', 1)

4. Categorize Popular Videos (Q2)

- Before deciding which column to use as the degree of reactivity, the correlation between the columns of numeric data was checked.
- Added Column :
  - 'reach' : The date reached from upload to section entry
  - 'period' : How long was in the section
  - '…_chng' : Growth rate while in section

- Correlation between the added column and the existing columns does not appear.
- Therefore, the 'engagement' will be 'on_views', which shows a strong correlation with the most columns among the existing columns.



	Duration	Reach	Tag Count	On Views	Off Views	On Rank	Off Rank	On Likes	Off Likes	On Dislikes	Off Dislikes	On Comments	Off Comments	On Channel Subscribers	Off Channel Subscribers	On Channel Total Views	Off Channel Total Views	On Channel Total Videos	Off Channel Total Videos	Period	Views Chng	Likes Chng	Comments Chng	Dislikes Chng	Subscriber Chng	Total Views Chng	Total Videos Chng
Duration	1	-0.032	0.2	-0.061	-0.057	-0.026	0.0044	-0.07	-0.063	-0.048	-0.055	-0.042	-0.04	-0.05	-0.05	-0.097	-0.097	-0.031	-0.031	0.047	-0.045	-0.039	0.0061	-0.026	-0.032	-0.025	0.0024
Reach	-0.032	1	-0.025	-0.049	-0.067	0.17	0.1	-0.066	-0.07	0.0021	-0.038	-0.056	-0.057	-0.061	-0.062	-0.062	-0.062	-0.037	-0.037	-0.14	-0.071	-0.0044	-0.036	-0.055	0.045	0.013	0.011
Tag Count	0.2	-0.025	1	-0.022	-0.017	-0.045	-0.15	-0.029	-0.028	-0.038	-0.033	-0.032	-0.03	0.012	0.012	0.027	0.027	-0.035	-0.034	-0.035	0.0031	0.021	0.0036	0.013	-0.052	-0.034	-0.034
On Views	-0.061	-0.049	-0.022	1	0.96	-0.17	-0.11	0.9	0.92	0.69	0.88	0.84	0.83	0.63	0.63	0.49	0.5	0.014	0.014	0.21	0.063	0.023	0.0068	0.14	-0.005	0.00066	0.0078
Off Views	-0.057	-0.067	-0.017	0.96	1	-0.16	-0.072	0.87	0.93	0.61	0.9	0.9	0.9	0.55	0.55	0.42	0.43	0.0032	0.0034	0.23	0.12	0.066	0.046	0.19	0.0089	0.013	0.0085
On Rank	-0.026	0.17	-0.045	-0.17	-0.16	1	0.36	-0.14	-0.12	-0.18	-0.16	-0.075	-0.072	-0.15	-0.15	-0.17	-0.17	-0.099	-0.099	-0.65	-0.32	-0.26	-0.25	-0.21	-0.096	-0.088	-0.08
Off Rank	0.0044	0.1	-0.15	-0.11	-0.072	0.36	1	-0.1	-0.083	-0.08	-0.055	-0.035	-0.03	-0.15	-0.15	-0.19	-0.19	-0.14	-0.14	0.26	0.12	0.093	0.056	0.092	0.014	-0.0096	-0.0022
On Likes	-0.07	-0.066	-0.029	0.9	0.87	-0.14	-0.1	1	0.97	0.69	0.84	0.8	0.77	0.77	0.77	0.51	0.52	-0.023	-0.023	0.16	0.059	0.013	-0.0096	0.12	-0.0097	-0.005	-0.0029
Off Likes	-0.063	-0.07	-0.028	0.92	0.93	-0.12	-0.083	0.97	1	0.61	0.87	0.88	0.87	0.68	0.69	0.45	0.45	-0.019	-0.019	0.17	0.073	0.033	0.0053	0.16	-0.0028	0.00066	0.00068
On Dislikes	-0.048	0.0021	-0.038	0.69	0.61	-0.18	-0.08	0.69	0.61	1	0.84	0.44	0.41	0.58	0.58	0.46	0.46	-0.014	-0.013	0.19	0.068	0.0038	0.0087	0.054	-0.0018	0.027	0.0037
Off Dislikes	-0.055	-0.038	-0.033	0.88	0.9	-0.16	-0.055	0.84	0.87	0.84	1	0.78	0.77	0.56	0.56	0.41	0.41	-0.015	-0.015	0.22	0.12	0.058	0.044	0.18	0.0085	0.032	0.0059
On Comments	-0.042	-0.056	-0.032	0.84	0.9	-0.075	-0.035	0.8	0.88	0.44	0.78	1	1	0.44	0.44	0.34	0.34	-0.008	-0.0079	0.14	0.057	0.034	0.01	0.16	-0.0023	-0.0017	0.00072
Off Comments	-0.04	-0.057	-0.03	0.83	0.9	-0.072	-0.03	0.77	0.87	0.41	0.77	1	1	0.41	0.41	0.32	0.32	-0.0076	-0.0075	0.14	0.061	0.038	0.015	0.17	0.00038	0.00056	0.0012
On Channel Subscribers	-0.05	-0.061	0.012	0.63	0.55	-0.15	-0.15	0.77	0.68	0.58	0.56	0.44	0.41	1	1	0.78	0.79	0.049	0.049	0.1	0.021	-0.018	-0.026	0.043	-0.038	-0.027	-0.021
Off Channel Subscribers	-0.05	-0.062	0.012	0.63	0.55	-0.15	-0.15	0.77	0.69	0.58	0.56	0.44	0.41	1	1	0.78	0.78	0.048	0.049	0.1	0.022	-0.017	-0.025	0.044	-0.036	-0.025	-0.02
On Channel Total Views	-0.097	-0.062	0.027	0.49	0.42	-0.17	-0.19	0.51	0.45	0.46	0.41	0.34	0.32	0.78	0.78	1	1	0.23	0.24	0.084	0.037	-0.0063	0.0036	0.021	-0.046	-0.033	-0.028
Off Channel Total Views	-0.097	-0.062	0.027	0.5	0.43	-0.17	-0.19	0.52	0.45	0.46	0.41	0.34	0.32	0.79	0.78	1	1	0.23	0.23	0.085	0.037	-0.0057	0.004	0.022	-0.046	-0.033	-0.027
On Channel Total Videos	-0.031	-0.037	-0.035	0.014	0.0032	-0.099	-0.14	-0.023	-0.019	-0.014	-0.015	-0.008	-0.0076	0.049	0.048	0.23	0.23	1	1	0.0073	-0.0016	0.0059	0.019	-0.013	-0.041	-0.023	-0.051
Off Channel Total Videos	-0.031	-0.037	-0.034	0.014	0.0034	-0.099	-0.14	-0.023	-0.019	-0.013	-0.015	-0.0079	-0.0075	0.049	0.049	0.24	0.23	1	1	0.0078	-0.002	0.006	0.02	-0.013	-0.029	-0.021	-0.02
Period	0.047	-0.14	-0.035	0.21	0.23	-0.65	0.26	0.16	0.17	0.19	0.22	0.14	0.14	0.1	0.1	0.084	0.085	0.0073	0.0078	1	0.59	0.49	0.45	0.4	0.18	0.12	0.11
Views Chng	-0.045	-0.071	0.0031	0.063	0.12	-0.32	0.12	0.059	0.073	0.068	0.12	0.057	0.061	0.021	0.022	0.037	0.037	-0.0016	-0.002	0.59	1	0.92	0.83	0.58	0.37	0.2	0.057
Likes Chng	-0.039	-0.0044	0.021	0.023	0.066	-0.26	0.093	0.013	0.033	0.0038	0.058	0.034	0.038	-0.018	-0.017	-0.0063	-0.0057	0.0059	0.006	0.49	0.92	1	0.86	0.52	0.42	0.18	0.042
Comments Chng	0.0061	-0.036	0.0036	0.0068	0.046	-0.25	0.056	-0.0096	0.0053	0.0087	0.044	0.01	0.015	-0.026	-0.025	0.0036	0.004	0.019	0.02	0.45	0.83	0.86	1	0.58	0.49	0.16	0.042
Dislikes Chng	-0.026	-0.055	0.013	0.14	0.19	-0.21	0.092	0.12	0.16	0.054	0.18	0.16	0.17	0.043	0.044	0.021	0.022	-0.013	-0.013	0.4	0.58	0.52	0.58	1	0.2	0.12	0.03
Subscriber Chng	-0.032	0.045	-0.052	-0.005	0.0089	-0.096	0.014	-0.0097	-0.0028	-0.0018	0.0085	-0.0023	0.00038	-0.038	-0.036	-0.046	-0.046	-0.041	-0.029	0.18	0.37	0.42	0.49	0.2	1	0.53	0.38
Total Views Chng	-0.025	0.013	-0.034	0.00066	0.013	-0.088	-0.0096	-0.005	0.00066	0.027	0.032	-0.0017	0.00056	-0.027	-0.025	-0.033	-0.033	-0.023	-0.021	0.12	0.2	0.18	0.16	0.12	0.53	1	0.37
Total Videos Chng	0.0024	0.011	-0.034	0.0078	0.0085	-0.08	-0.0022	-0.0029	0.00068	0.0037	0.0059	0.00072	0.0012	-0.021	-0.02	-0.028	-0.027	-0.051	-0.02	0.11	0.057	0.042	0.042	0.03	0.38	0.37	1

4. Categorize Popular Videos (Q2)

Goals

1. Analysis of factors that influence video’s entry into the Top Video Section.
2. Analysis of factors affecting the number of views during the period in which a section was entered.

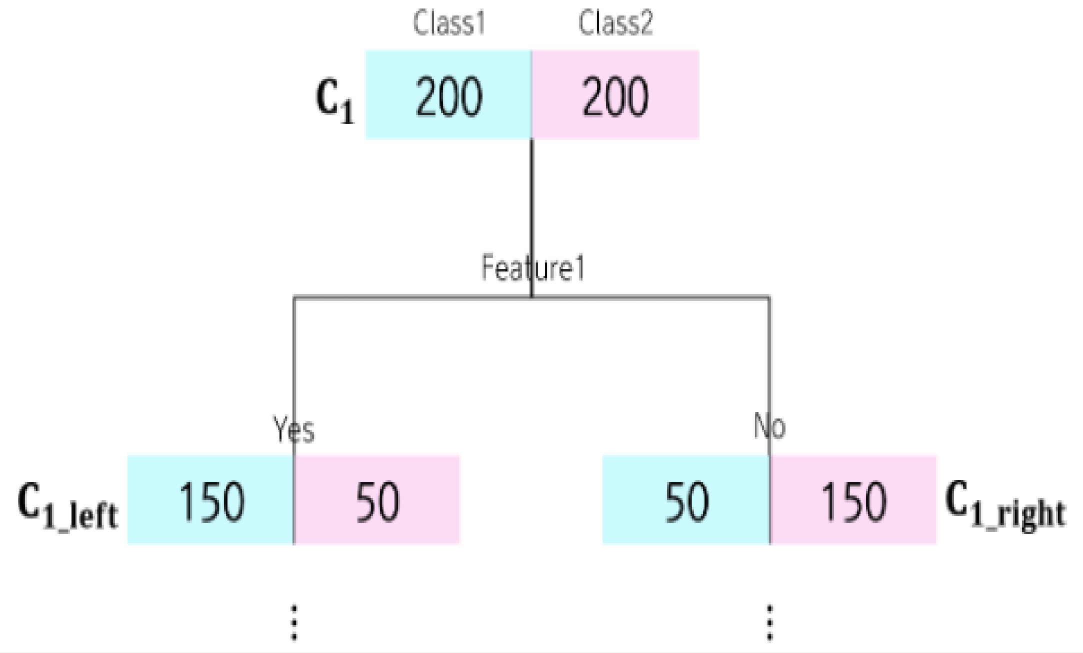
Steps

1. A model with good performance is found using various classification or regression models.
2. 'Feature Selection' is performed using the importance used in the calculation of each column.
3. The importance is recalculated using the 'Permutation Importance' method to secure validity.

4-2. Feature Selection

Default Feature Importance by ‘Gini’

1 Feature importance is calculated as the decrease in node impurity weighted by the probability of reaching that node. The node probability can be calculated by the number of samples that reach the node, divided by the total number of samples. **The higher the value the more important the feature.** It is usually calculated using the concept of **'Gini Impurity'**. In regression analysis, 'Information Gain' is the highest branching by using 'Mean Squared Error'.



$$G(N_j) = \sum_{i=1}^{K} p_i(1-p_i) = 1 - \sum_{i=1}^{K} p_i^2$$

Formula

$$G(C_1) = 1 - \{(\frac{200}{400})^2 + (\frac{200}{400})^2\} = 0.5$$
$$G(C_{1\_left}) = 1 - \{(\frac{150}{200})^2 + (\frac{50}{200})^2\} = 0.25$$
$$G(C_{1\_right}) = 1 - \{(\frac{50}{200})^2 + (\frac{50}{200})^2\} = 0.25$$

Gini Impurity of each node

$$\begin{aligned} I(C_j) &= 1 \cdot G(C_j) - \frac{200}{400} \cdot G(C_{j\_left}) - \frac{200}{400} \cdot G(C_{j\_right}) \\ &= 0.5 - 0.5 \cdot 0.25 - 0.5 \cdot 0.25 \\ &= 0.25 \end{aligned}$$

Importance of node

4-2. Feature Selection

Problems with the default feature importance mechanism

1 The mean decrease in impurity importance of a feature is computed by measuring how effective the feature is at reducing uncertainty (classifiers) or variance (regressors) when creating decision trees within RFs. The problem is that this mechanism, while fast, does not always give an accurate picture of importance.

2 It tends to inflate the importance of continuous or high-cardinality categorical variables. In 2007 Strobl et al pointed out that "*the variable importance measures of Breiman's original Random Forest method ... are not reliable in situations where potential predictor variables vary in their scale of measurement or their number of categories.*". (Bias in random forest variable importance measures: Illustrations, sources and a solution)

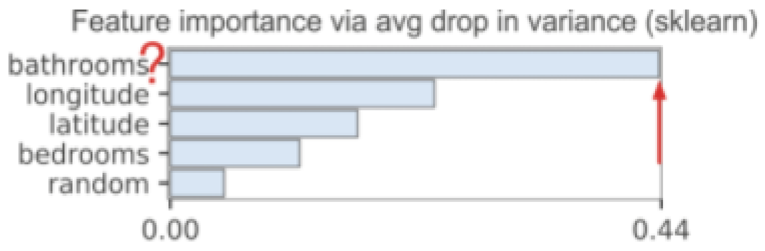


Figure 1(a). scikit-learn default importances for Random Forest **regressor** predicting apartment rental price from 4 features + a column of random numbers. Random column is last, as we would expect but the importance of the number of bathrooms for predicting price is highly suspicious.

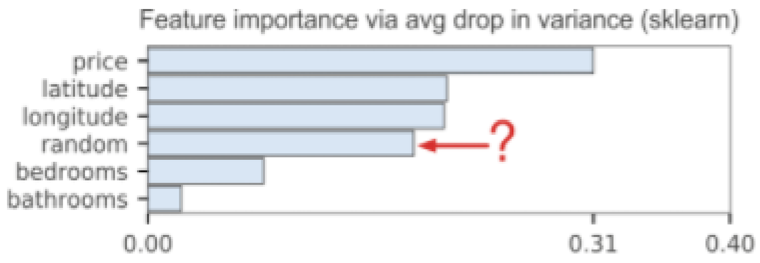


Figure 1(b). scikit-learn default importances for Random Forest **classifier** predicting apartment interest level (low, medium, high) using 5 features + a column of random numbers. Highly suspicious that random column is much more important than the number of bedrooms.

4-2. Feature Selection

Solution : Permutation Importance

1. Method : Record a baseline accuracy (classifier) or R2 score (regressor) by passing a validation set or the out-of-bag (OOB) samples through the Random Forest. Permute the column values of a single predictor feature and then pass all test samples back through the Random Forest and recompute the accuracy or R2. The importance of that feature is the difference between the baseline and the drop in overall accuracy or R2 caused by permuting the column.

2. Advantage :The permutation mechanism is much more computationally expensive than the mean decrease in impurity mechanism, but the results are more reliable.
The permutation importance strategy does not require retraining the model after permuting each column.

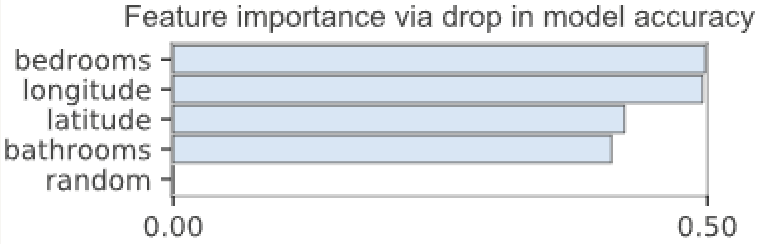


Figure 2(a). Importances derived by permuting each column and computing change in out-of-bag R^2 using `scikit-learn` **regressor**. Predicting apartment rental price from 4 features + a column of random numbers.

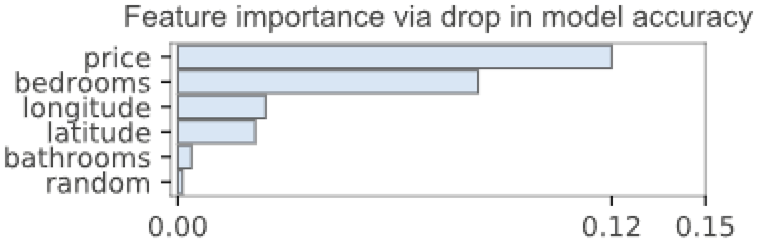


Figure 2(b). Importances derived by permuting each column and computing change in out-of-bag accuracy using `scikit-learn` Random Forest **classifier**.

4-3. Important Factors Before Entering the Section

Dataset

X

['reach','duration', 'tag_count', 'on_views', 'on_likes', 'on_dislikes', 'on_comments','on_channel_subscribers', 'on_channel_total_views', 'on_channel_total_videos’, ’Columns that one-hot encoded the “Category”’]

Y

Column labeled by categorizing the 'on_ranks' column into upper and lower ranks

	Random Forest Classifier	Logistic Regression	SVC	Decision Tree	KNN	Gaussian NB
Accuracy	0.675	0.665	0.597	0.560	0.624	0.371
Precision	0.607	0.656	0.470	0.449	0.481	0.371
Recall	0.347	0.204	0.689	0.837	0.194	1.000

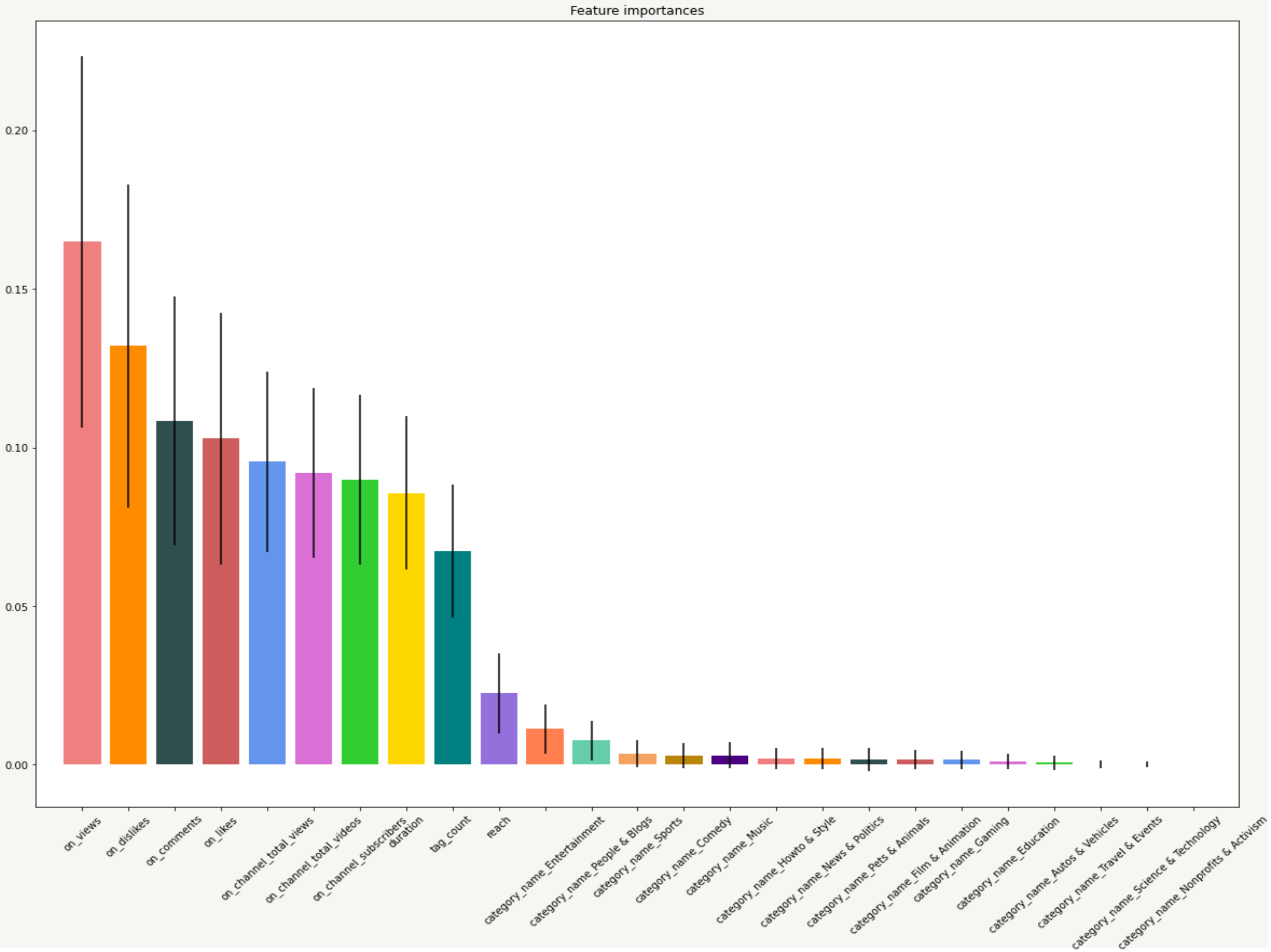
- For 'SVC', the accuracy was 0.6 when 'kernel=sigmoid'. However, for ‘Feature Selection' calculation, 'kernel=linear' must be set, so it is excluded as a final model. The above result is the result when it is 'linear’.
- Final Choice : Random Forest Classifer

4-3. Important Factors Before Entering the Section

Default Feature Importance

Default Feature Rankings:

1. on_views (0.165)
2. on_dislikes (0.132)
3. on_comments (0.108)
4. on_likes (0.103)
5. on_channel_total_views (0.095)
6. on_channel_total_videos (0.092)
7. on_channel_subscribers (0.090)
8. duration (0.086)
9. tag_count (0.067)
10. reach (0.023)
11. category_name_Entertainment (0.011)
12. category_name_People & Blogs (0.008)
13. category_name_Sports (0.004)
14. category_name_Comedy (0.003)
15. category_name_Music (0.003)
16. category_name_Howto & Style (0.002)
17. category_name_News & Politics (0.002)
18. category_name_Pets & Animals (0.002)
19. category_name_Film & Animation (0.002)
20. category_name_Gaming (0.002)
21. category_name_Education (0.001)
22. category_name_Autos & Vehicles (0.001)
23. category_name_Travel & Events (0.000)
24. category_name_Science & Technology (0.000)
25. category_name_Nonprofits & Activism (0.000)



4-3. Important Factors Before Entering the Section

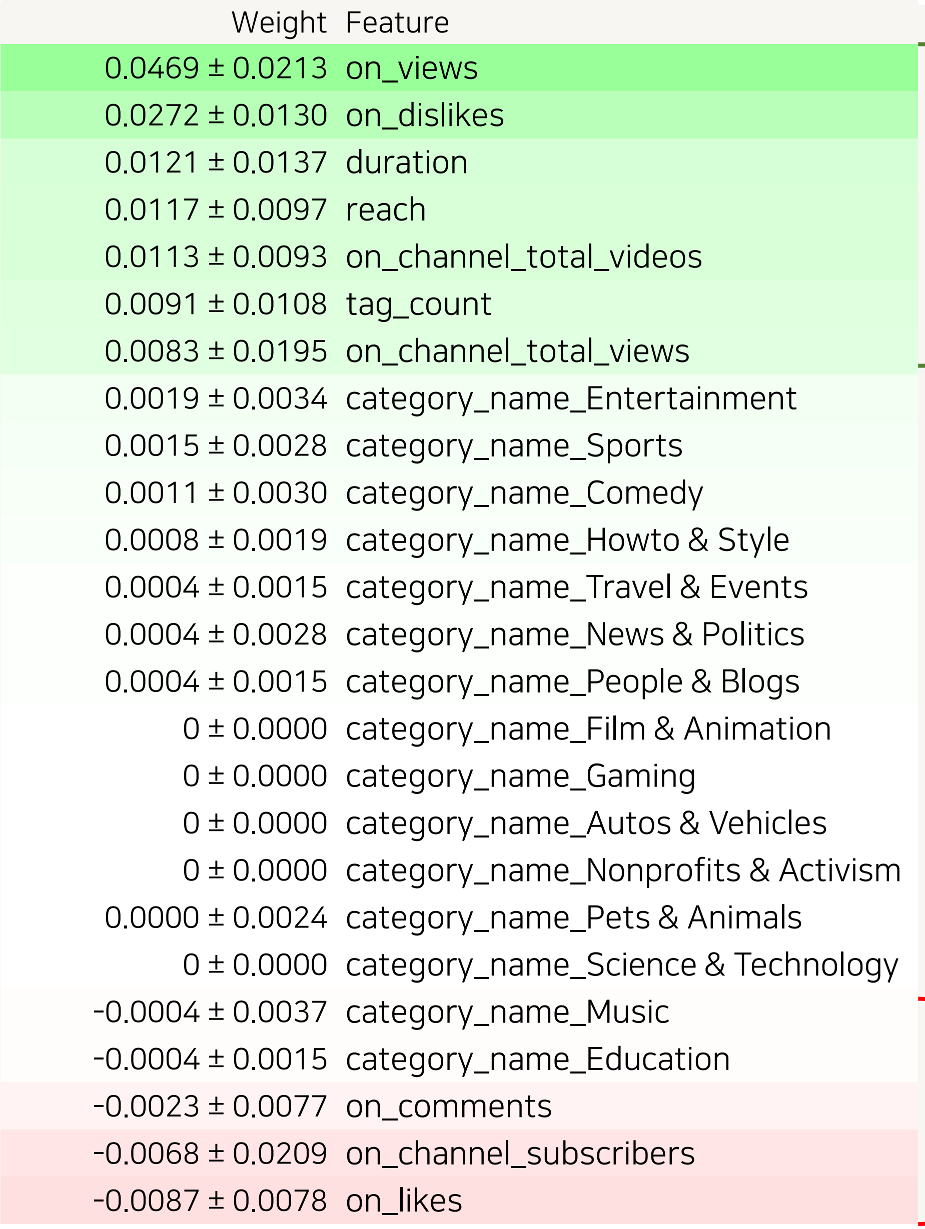
Permutation Importance

- Contrary to the previous results, . **'on_comments'** was a column that had no influence, and it can be seen that **'duration'** and **'reach'** are more influential.
- Also, it seems that **'on_comments'** is not related to the entry into popular videos rather than the one-hot encoded 'category' columns.

Defaualt Feature Rankings:

1. on_views (0.165)
2. on_dislikes (0.132)
3. on_comments (0.108)
4. on_likes (0.103)
5. on_channel_total_views (0.095)
6. on_channel_total_videos (0.092)
7. on_channel_subscribers (0.090)
8. duration (0.086)
9. tag_count (0.067)
10. reach (0.023)

23. category_name_Travel & Events (0.000)
24. category_name_Science & Technology (0.000)
25. category_name_Nonprofits & Activism (0.000)



Weight	Feature
0.0469 ± 0.0213	on_views
0.0272 ± 0.0130	on_dislikes
0.0121 ± 0.0137	duration
0.0117 ± 0.0097	reach
0.0113 ± 0.0093	on_channel_total_videos
0.0091 ± 0.0108	tag_count
0.0083 ± 0.0195	on_channel_total_views
0.0019 ± 0.0034	category_name_Entertainment
0.0015 ± 0.0028	category_name_Sports
0.0011 ± 0.0030	category_name_Comedy
0.0008 ± 0.0019	category_name_Howto & Style
0.0004 ± 0.0015	category_name_Travel & Events
0.0004 ± 0.0028	category_name_News & Politics
0.0004 ± 0.0015	category_name_People & Blogs
0 ± 0.0000	category_name_Film & Animation
0 ± 0.0000	category_name_Gaming
0 ± 0.0000	category_name_Autos & Vehicles
0 ± 0.0000	category_name_Nonprofits & Activism
0.0000 ± 0.0024	category_name_Pets & Animals
0 ± 0.0000	category_name_Science & Technology
-0.0004 ± 0.0037	category_name_Music
-0.0004 ± 0.0015	category_name_Education
-0.0023 ± 0.0077	on_comments
-0.0068 ± 0.0209	on_channel_subscribers
-0.0087 ± 0.0078	on_likes

Significant Features

Features that have no effect at all

4-4. Factors affecting the number of views after entering a section

Dataset

X

After removing outliers ['duration','on_views', 'off_views','on_likes', 'off_likes','on_dislikes','off_dislikes', 'on_comments', 'off_comments','on_channel_subscribers', 'off_channel_subscribers','on_channel_total_views','off_channel_total_views', 'period', 'on_channel_total_videos','off_channel_total_videos','Columns that one-hot encoded the "Category"']

Y

Values obtained by subtracting 'on_views' from 'off_views' and removing outliers

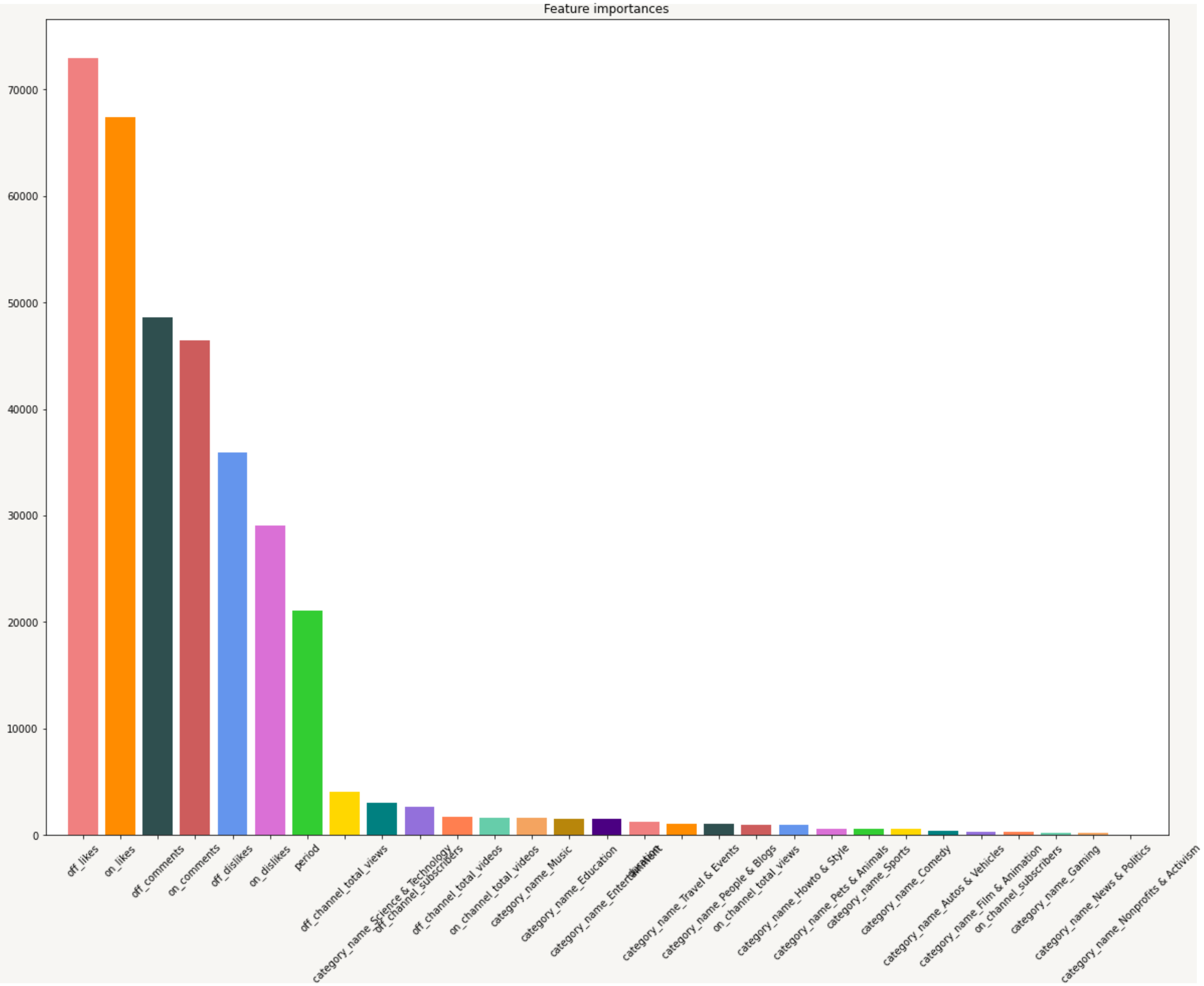
	Random Forest Regressor	Lasso	Ridge	ElasticNet
R2 score	0.716	0.992	0.992	0.961
Mean Squared Error	989,986,814.949	26,671,923.813	27,703,008.632	135,283,880.039
Mean Absolute Error	16,646.913	3,971.4936	4,023.288	7,632.152

- Lasso and Ridge have the highest explanatory coefficients, but in MSE, Lasso is rather low.
- Final Choice : Lasso

4-4. Factors affecting the number of views after entering a section

Default Feature Importance

Defaualt Feature Rankings:
1. off_likes (72986.632)
2. on_likes (67415.291)
3. off_comments (48625.184)
4. on_comments (46487.134)
5. off_dislikes (35933.920)
8. off_channel_total_views (4025.268)
9. category_name_Science & Technology (3032.550)
10. off_channel_subscribers (2653.861)
11. off_channel_total_videos (1657.764)
12. on_channel_total_videos (1598.726)
13. category_name_Music (1579.264)
14. category_name_Education (1491.767)
15. category_name_Entertainment (1469.539)
16. duration (1269.514)
17. category_name_Travel & Events (1050.670)
18. category_name_People & Blogs (1031.486)
19. on_channel_total_views (995.044)
20. category_name_Howto & Style (932.213)
21. category_name_Pets & Animals (616.910)
22. category_name_Sports (591.253)
23. category_name_Comedy (557.477)
24. category_name_Autos & Vehicles (396.215)
25. category_name_Film & Animation (330.847)
26. on_channel_subscribers (313.758)
27. category_name_Gaming (217.199)
28. category_name_News & Politics (197.342)
29. category_name_Nonprofits & Activism (0.000)

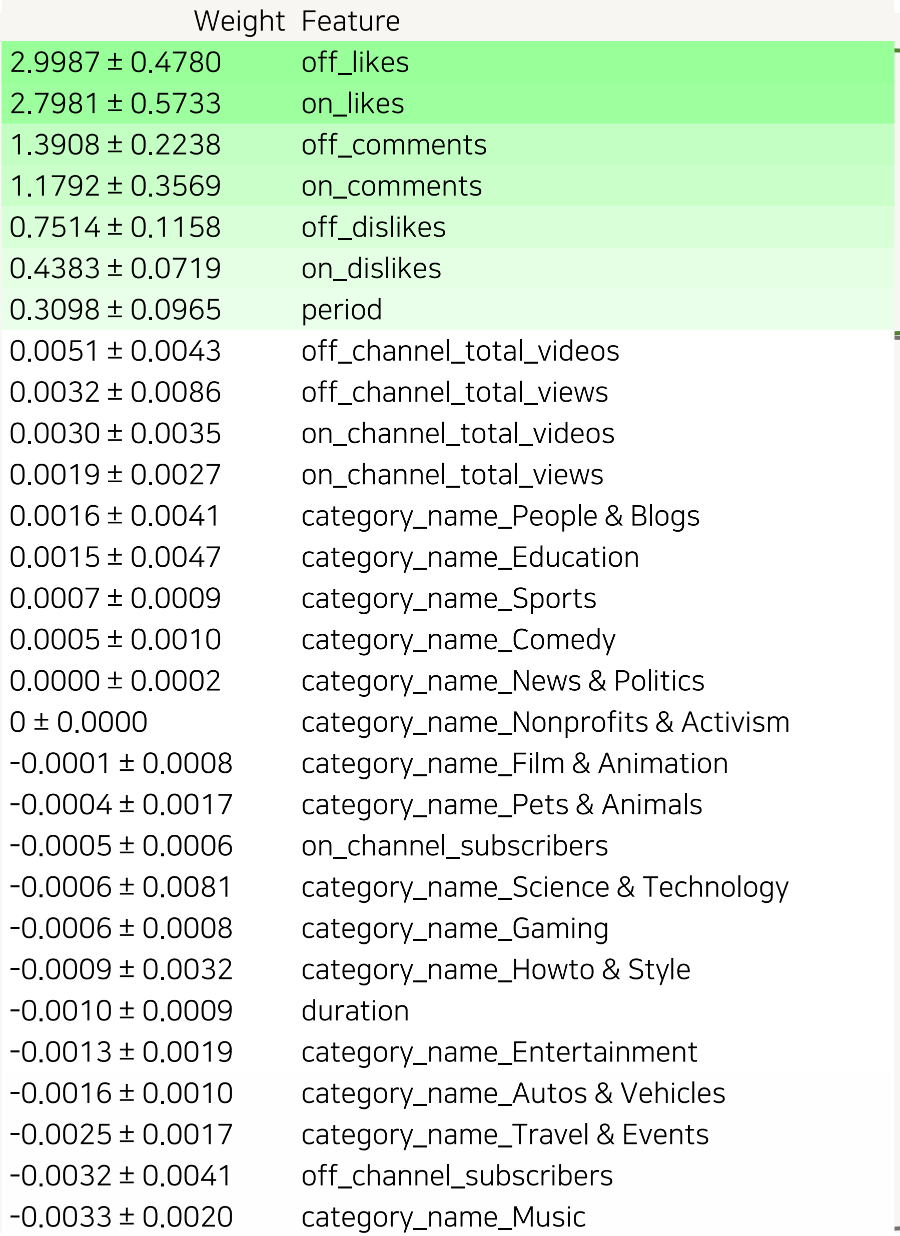


4-3. Important Factors Before Entering the Section

Permutation Importance

- Both yielded the same as the most influential factor.
- However, while 'category_name Nonprofits & Activism' was calculated as insignificant in the previous results, it was determined that there was no element to be removed in this result.

Defaualt Feature Rankings:
1. off_likes (72986.632)
2. on_likes (67415.291)
3. off_comments (48625.184)
4. on_comments (46487.134)
5. off_dislikes (35933.920)
6. on_dislikes (29008.680)
7. period (21063.150)
29. category_name_Nonprofits & Activism (0.000)



Weight	Feature
2.9987 ± 0.4780	off_likes
2.7981 ± 0.5733	on_likes
1.3908 ± 0.2238	off_comments
1.1792 ± 0.3569	on_comments
0.7514 ± 0.1158	off_dislikes
0.4383 ± 0.0719	on_dislikes
0.3098 ± 0.0965	period
0.0051 ± 0.0043	off_channel_total_videos
0.0032 ± 0.0086	off_channel_total_views
0.0030 ± 0.0035	on_channel_total_videos
0.0019 ± 0.0027	on_channel_total_views
0.0016 ± 0.0041	category_name_People & Blogs
0.0015 ± 0.0047	category_name_Education
0.0007 ± 0.0009	category_name_Sports
0.0005 ± 0.0010	category_name_Comedy
0.0000 ± 0.0002	category_name_News & Politics
0 ± 0.0000	category_name_Nonprofits & Activism
-0.0001 ± 0.0008	category_name_Film & Animation
-0.0004 ± 0.0017	category_name_Pets & Animals
-0.0005 ± 0.0006	on_channel_subscribers
-0.0006 ± 0.0081	category_name_Science & Technology
-0.0006 ± 0.0008	category_name_Gaming
-0.0009 ± 0.0032	category_name_Howto & Style
-0.0010 ± 0.0009	duration
-0.0013 ± 0.0019	category_name_Entertainment
-0.0016 ± 0.0010	category_name_Autos & Vehicles
-0.0025 ± 0.0017	category_name_Travel & Events
-0.0032 ± 0.0041	off_channel_subscribers
-0.0033 ± 0.0020	category_name_Music

Significant Features

Not enough to exclude, but not very influential

5. Conclusion

- The number of views and dislikes, the length of the video, the number of tags, the time it takes to enter the section after upload, the number of videos and the total number of views of the channel affect the section entry.
- It can be seen that during the entry period, the number of likes, comments, and dislikes is the factor that has the greatest influence on the number of views of the entered video.

6. Source

- https://soohee410.github.io/iml_tree_importance
- https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3
- https://explained.ai/rf-importance/index.html