

# YOUTUBE VIDEO ANALYSIS

Abhilash Antony

We have a YouTube video data set of the videos uploaded at the last week of the month of May, 2024. Each row in the table contains information about a single video, including the video ID, title, description, publish date, channel title, channel ID, category, tags, duration, definition, caption, view count, likes, dislikes, comment count, and favourites count.

Here are some specific details about the various features in the dataset:

- Video IDs: The video IDs appear to be alphanumeric strings, likely unique identifiers assigned by YouTube to each video.
- Titles: The titles of the videos are included in the dataset.
- Descriptions: The descriptions of the videos are also included, which can provide some context about the content of the video.
- Publish Date: The date the video was published on YouTube is included.
- Channel Title: The title of the channel that uploaded the video is listed.
- Channel ID: There is also a channel ID included, which could be a unique identifier for the channel.
- Category: The category of the video is listed. This could be a general category like "music" or "entertainment," or a more specific category like "video games" or "comedy."
- Tags: The tags associated with the video are also included. Tags are keywords that creators can add to their videos to help people find them.
- Duration: The duration of the video is listed in minutes.
- Definition: It appears there is a data point for video definition, but it is not clear from this sample what the values mean ("TRUE" or "FALSE").
- Caption: There is also a data point for caption, but it is not clear from this sample what the values mean ("TRUE" or "FALSE").
- View Count: The number of times the video has been viewed is listed.
- Likes: The number of likes the video has received is listed.
- Dislikes: The number of dislikes the video has received is listed.
- Comment Count: The number of comments on the video is listed.
- Favorites Count: The number of times the video has been added to a user's favorites list is listed.

This dataset can be used for understanding what kind of content is popular on YouTube. By analyzing the views, likes, dislikes, and comments on different videos, we can get a sense of what kind of content resonates with viewers. This information can be valuable for content creators, who can use it to inform their video strategy.

```
# import the libraries
import pandas as pd
import matplotlib.pyplot as plt

# import the data
df = pd.read_csv("/content/trending_videos.csv")
```

Performing initial analysis,

```
df.head()
```

video_id	title	description	published_at	channel_id	channel_title	category_id	tags	duration	definition	caption	view_count	like_count	dislike_count	favorite_count	comment_count
2ZivWamTie8	Eminem - Houdini [Official Music Video]	Eminem - Houdini [Official Music Video] https://eminem.lnk.t...	2024-05-31T04:00:02Z	UC20vb-R_px4CguhtzBPhoyQ	EminemVEVO	10	[Eminem', 'Houdini', 'Hip Hop', '조지 마이클', '대니 데이빗 프랜시스', '데프 레퍼드', '리틀 믹스', '비스트', '블랙 알바인', '레드 제플린', '스미티', '스몰 타이머즈', '스피츠', '스튜디오 시티', '스위트 리버', '스윙 로큰롤', '스윙 밴드', '스윙 음악', '스윙 장르', '스윙 스타일', '스윙 문화', '스윙 역사', '스윙 전통', '스윙 계보', '스윙 뿌리', '스윙 본질', '스윙 정신', '스윙 영혼', '스윙 생명력', '스윙 열정', '스윙 감동', '스윙 행복', '스윙 사랑', '스윙 희망', '스윙 꿈', '스윙 미래', '스윙 가능성', '스윙 잠재력', '스윙 무한도전', '스윙 끝없는 여정', '스윙 영원불변의 클래식', '스윙 시대를 초월하는 명작', '스윙 인류 보편의 언어', '스윙 문화유산', '스윙 예술적 가치', '스윙 사회적 영향력', '스윙 경제적 파급효과', '스윙 산업적 중요성', '스윙 대중적 인기', '스윙 상업적 성공', '스윙 마케팅 전략', '스윙 홍보 효과', '스윙 팬덤 형성', '스윙 커뮤니티 구축', '스윙 브랜드 파워', '스윙 라이선싱 수익', '스윙 merchandise 판매', '스윙 live 공연 매출', '스윙 음반 판매량', '스윙 스트리밍 실적', '스윙 디지털 전환', '스윙 온라인 플랫폼 활용', '스윙 소셜 미디어 캠페인', '스윙 유튜브 알고리즘 최적화', '스윙 SEO 전략', '스윙 데이터 분석을 통한 콘텐츠 개선', '스윙 A/B 테스트를 통한 클릭률 향상', '스윙 개인화된 추천 시스템 도입', '스윙 크로스 프로모션 전략', '스윙 콜라보레이션 기회 창출', '스윙 글로벌 시장 진출', '스윙 다문화 수용성 증진', '스윙 다양성 존중', '스윙 포용적인 환경 조성', '스윙 사회적 책임 이행', '스윙 지속 가능한 발전 추구', '스윙 투명성 강화', '스윙 소비자 신뢰도 제고', '스윙 고객 만족도 향상', '스윙 재구매율 증대', '스윙 충성도 프로그램 운영', '스윙 맞춤형 서비스 제공', '스윙 실시간 피드백 반영', '스윙 유연한 대응 능력 발휘', '스윙 위기 관리 역량 강화', '스윙 평판 관리 철저', '스윙 이미지 메이킹 전략 수립', '스윙 브랜드 정체성 확립', '스윙 경쟁 우위 확보', '스윙 시장 점유율 확대', '스윙 성장 동력 발굴', '스윙 혁신 의지贯彻', '스윙 도전 정신 함양', '스윙 끈기 있는 노력', '스윙 실패를 두려워하지 않는 자세', '스윙 끊임없는 학습 태도', '스윙 협업 능력 배양', '스윙 리더십 발휘', '스윙 팀워크 강화', '스윙 소통 능력 향상', '스윙 갈등 해결 능력 키우기', '스윙 스트레스 관리 방법 습득', '스윙 긍정적인 마인셋 유지', '스윙 감사의 마음가짐 실천', '스윙 용감하게 도전하기', '스윙 새로운 시도하기', '스윙 변화에 적응하기', '스윙 불확실성 안락히 받아들이기', '스윙 리스크 관리 능력 기르기', '스윙 기회 포착 능력 높이기', '스윙 창의력 자극하기', '스윙 상상력을 현실로 만들기', '스윙 문제 해결 능력 키우기', '스윙 의사결정 능력 향상', '스윙 시간 관리 능력 배우기', '스윙 생산성 극대화하기', '스윙 효율적인 업무 프로세스 구축', '스윙 자동화 도구 활용하기', '스윙 원격 근무 능력 키우기', '스윙 유연근로제에 익숙해지기', '스윙 워라밸 지킴이 되기', '스윙 건강한 생활습관 만들기', '스윙 충분한 휴식 취하기', '스윙 긍정적인 에너지 충전하기', '스윙 삶의 질 향상시키기', '스윙 행복한 인생 설계하기', '스윙 꿈을 이루는 법 배우기', '스윙 자기계발에 투자하기', '스윙 평생학습 mindset 갖기', '스윙 다양한 경험 쌓기', '스윙 넓은 인맥 넓히기', '스윙 멘토 찾기', '스윙 롤모델 만들기', '스윙 동기 부여 요인 찾기', '스윙 목표 설정 능력 기르기', '스윙 계획 세우기', '스윙 실행력 키우기', '스윙 끈기 버리기', '스윙 좌절 극복하기', '스윙 포기하지 않기', '스윙 끝까지 밀어붙이기', '스윙 승부욕 키우기', '스윙 도전정신 불태우기', '스윙 자존감 높이기', '스윙 자신감 가지기', '스윙 긍정적 사고방식 만들기', '스윙 낙관론자 되기', '스윙 감사 일기 쓰기', '스윙 하루 소망 나누기', '스윙 감사의 편지 보내기', '스윙 봉사활동 참여하기', '스윙 기부하기', '스윙 나눔 실천하기', '스윙 사회공헌 활동 기획하기', '스윙 ESG 경영 이해하기', '스윙 친환경 제품 사용하기', '스윙 재활용품 분리수거 하기', '스윙 에너지 절약하기', '스윙 물 절약하기', '스윙 쓰레기 줄이기', '스윙 탄소 발자국 줄이기', '스윙 기후위기 대응하기', '스윙 SDGs 실현하기', '스윙 UN 지속가능발전목표(SDGs) 이해하기', '스윙 평화와 정의 실현하기', '스윙 인권 존중하기', '스윙 차별 금지법 지지하기', '스윙 장애인 권리 옹호하기', '스윙 노인 복지 증진하기', '스윙 아동 보호하기', '스윙 청소년 권리 옹호하기', '스윙 동물권 존중하기', '스윙 식물권 존중하기', '스윙 생태계 보호하기', '스윙 생물다양성 보전하기', '스윙 산림 자원 관리하기', '스윙 해양 자원 관리하기', '스윙 대기환경 오염 방지하기', '스윙 수질오염 방지하기', '스윙 토양오염 방지하기', '스윙 방사능 안전관리하기', '스윙 화학물질 안전하게 사용하기', '스윙 전기안전사고 예방하기', '스윙 화재예방하기', '스윙 범죄예방하기', '스윙 재난대비 훈련하기', '스윙 비상탈출 경로 숙지하기', '스윙 소화기 사용법 익히기', '스윙 심폐소생술 배우기', '스윙 응급처치 방법 배우기', '스윙 교통안전 지키기', '스윙 보행자 안전 지키기', '스윙 자전거 안전 지키기', '스윙 운전 안전 지키기', '스윙 항공 안전 지키기', '스윙 선박 안전 지키기', '스윙 유람선 안전 지키기', '스윙 호텔 안전 지키기', '스윙 식당 안전 지키기', '스윙 카페 안전 지키기', '스윙 도서관 안전 지키기', '스윙 박물관 안전 지키기', '스윙 미술관 안전 지키기', '스윙 체육관 안전 지키기', '스윙 수영장 안전 지키기', '스윙 놀이공원 안전 지키기', '스윙 축제 안전 지키기', '스윙 콘서트 안전 지키기', '스윙 행사 안전 지키기', '스윙 집안 안전 지키기', '스윙 직장 안전 지키기', '스윙 학교 안전 지키기', '스윙 공공장소 안전 지키기', '스윙 해외여행 안전 지키기', '스윙 낯선 곳에서의 안전 지키기', '스윙 혼자 여행할 때의 안전 지키기', '스윙 야간 외출할 때의 안전 지키기', '스윙 새벽 출근할 때의 안전 지키기', '스윙 늦게 잠들었을 때의 안전 지키기', '스윙 아침 기상할 때의 안전 지키기', '스윙 샤워할 때의 안전 지키기', '스윙 목욕할 때의 안전 지키기', '스윙 화장할 때의 안전 지키기', '스윙 옷 입을 때의 안전 지키기', '스윙 신발을 신고 걷기', '스윙 가방을 메고 다니기', '스윙 스마트폰을 들고 다니기', '스윙 현금과 카드 지갑 정리하기', '스윙 신분증 잘 보관하기', '스윙 여권 잘 보관하기', '스윙 항공권 잘 보관하기', '스윙 호텔 예약 확인하기', '스윙 렌터카 계약서 꼼꼼히 읽기', '스윙 보험 가입하기', '스윙 건강보험 잘 챙기기', '스윙 국민연금 납부 확인하기', '스윙 세금 신고하기', '스윙 연말정산 준비하기', '스윙 소득세 납부하기', '스윙 재산세 납부하기', '스윙 자동차세 납부하기', '스윙 주민등록세 납부하기', '스윙 종합소득세 납부하기', '스윙 상속세 납부하기', '스윙 증여세 납부하기', '스윙 법인세 납부하기', '스윙 법인 설립하기', '스윙 법인 전환하기', '스윙 법인 합병하기', '스윙 법인 분할하기', '스윙 법인 해산하기', '스윙 법인 청산하기', '스윙 법인 회계처리하기', '스윙 법인 인사관리하기', '스윙 법인 급여 지급하기', '스윙 법인 퇴직금 지급하기', '스윙 법인 연봉 인상하기', '스윙 법인 주주총회 개최하기', '스윙 법인 이사회 구성하기', '스윙 법인 대표이사 선임하기', '스윙 법인 감사인 선임하기', '스윙 법인 변호사 선임하기', '스윙 법인 회계법인 선정하기', '스윙 법인 세무 컨설팅 받기', '스윙 법인 재무제표 작성하기', '스윙 법인 손익분기점 계산하기', '스윙 법인 현금흐름표 작성하기', '스윙 법인 자산負債比率 관리하기', '스윙 법인 신용등급 관리하기', '스윙 법인 대출 신청하기', '스윙 법인 투자 결정하기', '스윙 법인 인수합병(M&A) 추진하기', '스윙 법인 IPO 추진하기', '스윙 법인 상장하기', '스윙 법인 사모펀드 유치하기', '스윙 법인 벤처캐пит얼 유치하기', '스윙 법인 angel 투자 유치하기', '스윙 법인 seed 투자 유치하기', '스윙 법인 시리즈 A 투자 유치하기', '스윙 법인 시리즈 B 투자 유치하기', '스윙 법인 시리즈 C 투자 유치하기', '스윙 법인 시리즈 D 투자 유치하기', '스윙 법인 시리즈 E 투자 유치하기', '스윙 법인 시리즈 F 투자 유치하기', '스윙 법인 시리즈 G 투자 유치하기', '스윙 법인 시리즈 H 투자 유치하기', '스윙 법인 시리즈 I 투자 유치하기', '스윙 법인 시리즈 J 투자 유치하기', '스윙 법인 시리즈 K 투자 유치하기', '스윙 법인 시리즈 L 투자 유치하기', '스윙 법인 시리즈 M 투자 유치하기', '스윙 법인 시리즈 N 투자 유치하기', '스윙 법인 시리즈 O 투자 유치하기', '스윙 법인 시리즈 P 투자 유치하기', '스윙 법인 시리즈 Q 투자 유치하기', '스윙 법인 시리즈 R 투자 유치하기', '스윙 법인 시리즈 S 투자 유치하기', '스윙 법인 시리즈 T 투자 유치하기', '스윙 법인 시리즈 U 투자 유치하기', '스윙 법인 시리즈 V 투자 유치하기', '스윙 법인 시리즈 W 투자 유치하기', '스윙 법인 시리즈 X 투자 유치하기', '스윙 법인 시리즈 Y 투자 유치하기', '스윙 법인 시리즈 Z 투자 유치하기', '스윙 법인 시리즈 AA 투자 유치하기', '스윙 법인 시리즈 AB 투자 유치하기', '스윙 법인 시리즈 AC 투자 유치하기', '스윙 법인 시리즈 AD 투자 유치하기', '스윙 법인 시리즈 AE 투자 유치하기', '스윙 법인 시리즈 AF 투자 유치하기', '스윙 법인 시리즈 AG 투자 유치하기', '스윙 법인 시리즈 AH 투자 유치하기', '스윙 법인 시리즈 AI 투자 유치하기', '스윙 법인 시리즈 AJ 투자 유치하기', '스윙 법인 시리즈 AK 투자 유치하기', '스윙 법인 시리즈 AL 투자 유치하기', '스윙 법인 시리즈 AM 투자 유치하기', '스윙 법인 시리즈 AN 투자 유치하기', '스윙 법인 시리즈 AO 투자 유치하기', '스윙 법인 시리즈 AP 투자 유치하기', '스윙 법인 시리즈 AQ 투자 유치하기', '스윙 법인 시리즈 AR 투자 유치하기', '스윙 법인 시리즈 AS 투자 유치하기', '스윙 법인 시리즈 AT 투자 유치하기', '스윙 법인 시리즈 AU 투자 유치하기', '스윙 법인 시리즈 AV 투자 유치하기', '스윙 법인 시리즈 AW 투자 유치하기', '스윙 법인 시리즈 AX 투자 유치하기', '스윙 법인 시리즈 AY 투자 유치하기', '스윙 법인 시리즈 AZ 투자 유치하기', '스윙 법인 시리즈 BA 투자 유치하기', '스윙 법인 시리즈 BB 투자 유치하기', '스윙 법인 시리즈 BC 투자 유치하기', '스윙 법인 시리즈 BD 투자 유치하기', '스윙 법인 시리즈 BE 투자 유치하기', '스윙 법인 시리즈 BF 투자 유치하기', '스윙 법인 시리즈 BG 투자 유치하기', '스윙 법인 시리즈 BH 투자 유치하기', '스윙 법인 시리즈 BI 투자 유치하기', '스윙 법인 시리즈 BJ 투자 유치하기', '스윙 법인 시리즈 BK 투자 유치하기', '스윙 법인 시리즈 BL 투자 유치하기', '스윙 법인 시리즈 BM 투자 유치하기', '스윙 법인 시리즈 BN 투자 유치하기', '스윙 법인 시리즈 BO 투자 유치하기', '스윙 법인 시리즈 BP 투자 유치하기', '스윙 법인 시리즈 BQ 투자 유치하기', '스윙 법인 시리즈 BR 투자 유치하기', '스윙 법인 시리즈 BS 투자 유치하기', '스윙 법인 시리즈 BT 투자 유치하기', '스윙 법인 시리즈 BU 투자 유치하기', '스윙 법인 시리즈 BV 투자 유치하기', '스윙 법인 시리즈 BW 투자 유치하기', '스윙 법인 시리즈 BX 투자 유치하기', '스윙 법인 시리즈 BY 투자 유치하기', '스윙 법인 시리즈 BZ 투자 유치하기', '스윙 법인 시리즈 CA 투자 유치하기', '스윙 법인 시리즈 CB 투자 유치하기', '스윙 법인 시리즈 CC 투자 유치하기', '스윙 법인 시리즈 CD 투자 유치하기', '스윙 법인 시리즈 CE 투자 유치하기', '스윙 법인 시리즈 CF 투자 유치하기', '스윙 법인 시리즈 CG 투자 유치하기', '스윙 법인 시리즈 CH 투자 유치하기', '스윙 법인 시리즈 CI 투자 유치하기', '스윙 법인 시리즈 CJ 투자 유치하기', '스윙 법인 시리즈 CK 투자 유치하기', '스윙 법인 시리즈 CL 투자 유치하기', '스윙 법인 시리즈 CM 투자 유치하기', '스윙 법인 시리즈 CN 투자 유치하기', '스윙 법인 시리즈 CO 투자 유치하기', '스윙 법인 시리즈 CP 투자 유치하기', '스윙 법인 시리즈 CQ 투자 유치하기', '스윙 법인 시리즈 CR 투자 유치하기', '스윙 법인 시리즈 CS 투자 유치하기', '스윙 법인 시리즈 CT 투자 유치하기', '스윙 법인 시리즈 CU 투자 유치하기', '스윙 법인 시리즈 CV 투자 유치하기', '스윙 법인 시리즈 CW 투자 유치하기', '스윙 법인 시리즈 CX 투자 유치하기', '스윙 법인 시리즈 CY 투자 유치하기', '스윙 법인 시리즈 CZ 투자 유치하기', '스윙 법인 시리즈 DA 투자 유치하기', '스윙 법인 시리즈 DB 투자 유치하기', '스윙 법인 시리즈 DC 투자 유치하기', '스윙 법인 시리즈 DD 투자 유치하기', '스윙 법인 시리즈 DE 투자 유치하기', '스윙 법인 시리즈 DF 투자 유치하기', '스윙 법인 시리즈 DG 투자 유치하기', '스윙 법인 시리즈 DH 투자 유치하기', '스윙 법인 시리즈 DI 투자 유	PT4M57S	hd	True	14736971	1306831	0	0	105793
Kt56x8P9m90	College Football 25   Gameplay Deep Dive	Bring Glory Home. Pre-order EA SPORTS College Deep Dive	2024-05-31T14:55:06Z	UCT4wAMwETXqDf-U_DVUqabA	EA SPORTS College	20	[college football', 'college football 25', 'c...	PT4M52S	hd	False	1079642	50259	0	0	6936
mtz-Zzk88s	ILLEGAL builds in LEGO...	50+ secret ways to build in Lego you probably ...	2024-05-31T15:30:36Z	UCUUsGdGuQshZFRGnxAPBf_w	TD BRICKS	24	[lego', 'lego set', 'lego sets', 'lego movie'...	PT9M7S	hd	True	1064281	24723	0	0	2690
/GrOpZnsPk4	ATEEZ (에이티즈) 'WORK' WORK! MV	[GOLDEN HOUR : Part.1]Release Date: 2024. 5...	2024-05-31T04:00:01Z	UCQd3_qPEq_zY_wP_kuVB9Q	KQ ENTERTAINMENT	10	[KQ', '에이티즈', 'ATEEZ', '								

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 16 columns):
#   Column             Non-Null Count  Dtype
---  -
0   video_id           200 non-null    object
1   title              200 non-null    object
2   description         196 non-null    object
3   published_at        200 non-null    object
4   channel_id          200 non-null    object
5   channel_title       200 non-null    object
6   category_id         200 non-null    int64
7   tags                200 non-null    object
8   duration            200 non-null    object
9   definition          200 non-null    object
10  caption             200 non-null    bool
11  view_count          200 non-null    int64
12  like_count          200 non-null    int64
13  dislike_count       200 non-null    int64
14  favorite_count       200 non-null    int64
15  comment_count       200 non-null    int64
dtypes: bool(1), int64(6), object(9)
memory usage: 23.8+ KB
```

```
# Check for missing values
print(df.isnull().sum())
```

```

video_id      0
title         0
description    4
published_at   0
channel_id    0
channel_title  0
category_id   0
tags          0
duration      0
definition    0
caption       0
view_count    0
like_count    0
dislike_count 0
favorite_count 0
comment_count 0

```

We see here that the variable description has 4 missing values. Usually we impute or remove them. Here, we remove the tuples with missing values.

```
# Drop the missing values
df = df.dropna()
```

```
# Get descriptive statistics
df.describe()
```

	category_id	view_count	like_count	dislike_count	favorite_count	comment_count
count	200.000000	2.000000e+02	2.000000e+02	200.0	200.0	200.000000
mean	18.835000	2.296781e+06	9.129304e+04	0.0	0.0	8131.505000
std	6.585943	5.992482e+06	2.397322e+05	0.0	0.0	28670.786143
min	1.000000	5.526100e+04	1.430000e+02	0.0	0.0	0.000000
25%	17.000000	3.462905e+05	1.472700e+04	0.0	0.0	1010.000000
50%	20.000000	7.330895e+05	2.795400e+04	0.0	0.0	2046.000000
75%	24.000000	1.386557e+06	6.148650e+04	0.0	0.0	4197.000000
max	28.000000	6.643700e+07	2.535500e+06	0.0	0.0	279003.000000

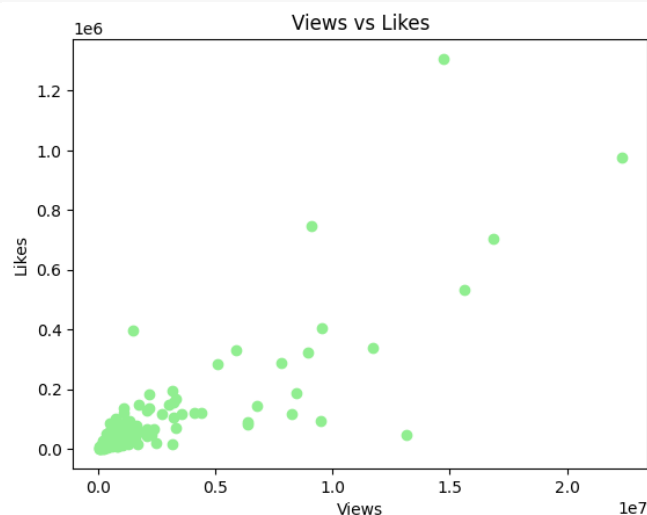
```
# find the channel_id that published the most videos
most_videos_channel = df['channel_id'].value_counts().idxmax()
print(f"Channel with the most videos: {most_videos_channel}")
```

Channel with the most videos: UCWJ2IWNubArHWmf3FIHbfcQ

```
# find the range of date our data spans
min_date = df['published_at'].min()
max_date = df['published_at'].max()
print(f"Range of published dates: {min_date} to {max_date}")
```

Range of published dates: 2024-05-04T03:06:01Z to 2024-05-31T15:30:38Z

```
# the relationship between views and likes
plt.scatter(df['view_count'], df['like_count'], color='lightgreen')
plt.title('Views vs Likes')
plt.xlabel('Views')
plt.ylabel('Likes')
plt.show()
```



The plot above shows a positive correlation between the Views and Likes. That is, as the number of views to the video increases, the possibility of getting more likes also increases.

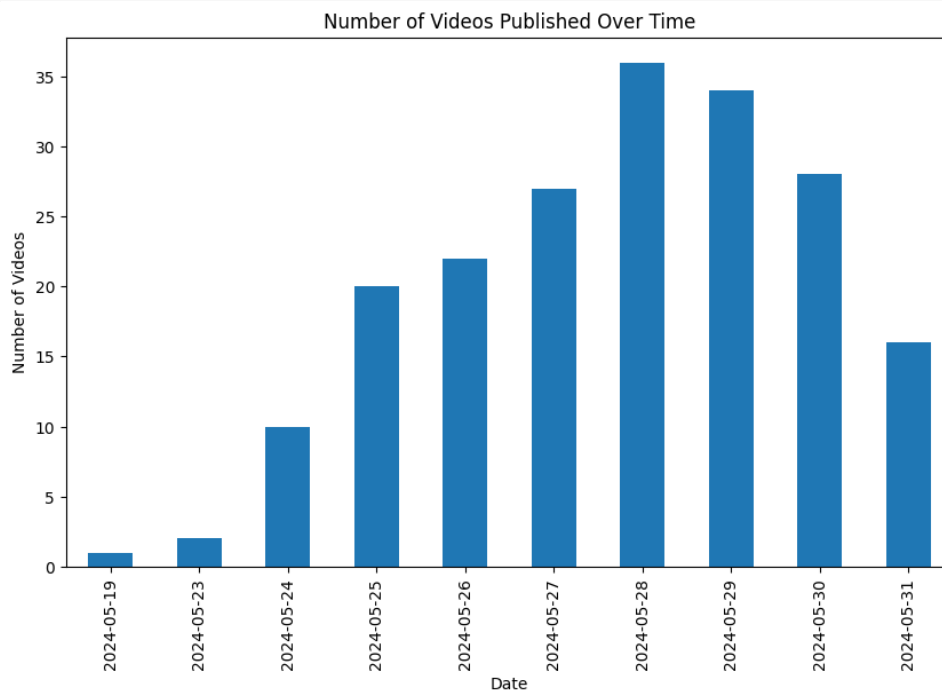
Here, all the numerical values are in a consistent format as all of them are counts. So, we opt not to perform any standardization procedures.

Now, let us look at the distribution of videos from 2024-05-04 to 2024-05-31.

```
# Create a new column with the date only
df['date'] = pd.to_datetime(df['published_at']).dt.date

# Group the data by date and count the number of videos
grouped_data = df.groupby('date')['video_id'].count()

# Plot the data
grouped_data.plot(kind='bar', figsize=(10, 6))
plt.title('Number of Videos Published Over Time')
plt.xlabel('Date')
plt.ylabel('Number of Videos')
plt.show()
```



Here, we see that most of the videos were uploaded at the end of the week. There can be other potential factors affecting this. But that would need further data.

Another thing that affects the popularity of YouTube videos is the tags that are used along with the videos. Now, let's find out the tags that are most used in the given data.

```
# Convert the tags column to a list of strings
df['tags'] = df['tags'].str.strip('[]').str.strip('"').str.split(',',
'')

# Flatten the list of lists into a single list
```

```

all_tags = [tag for sublist in df['tags'].tolist() for tag in sublist]

# Count the occurrences of each tag
tag_counts = {}
for tag in all_tags:
    if tag not in tag_counts:
        tag_counts[tag] = 0
    tag_counts[tag] += 1

# Sort the tags by their counts
sorted_tags = sorted(tag_counts.items(), key=lambda item: item[1],
reverse=True)

# Print the most used tags
print("Most used tags:")
for tag, count in sorted_tags[:10]:
    print(f"{tag}: {count}")

```

```

⇒ Most used tags:
the: 40
vs: 37
and: 34
man: 34
2: 33
of: 33
new: 29
pokemon: 29
Fortnite: 29
minecraft: 28

```

This seems to be misleading as we are not sure of the stochastic behaviour of our data. There is a possibility that the data might be skewed to videos for kids as the most used tags seem to relate to kids' videos.

Now, let's look at how tags are related to the view count.

```

# Explode the tags column to get individual tag-video pairs
df_exploded = df.explode('tags')

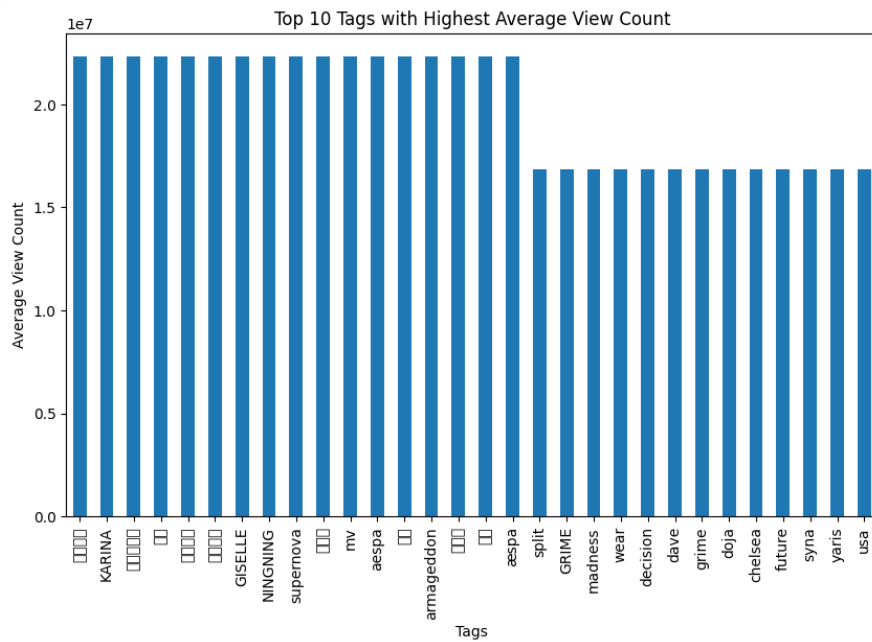
# Group the exploded data by tags and calculate the average view count
avg_views_by_tag = df_exploded.groupby('tags')['view_count'].mean()

# Sort the tags by average view count
sorted_tags_by_views = avg_views_by_tag.sort_values(ascending=False)

# Plot the top 10 tags with the highest average view count
plt.figure(figsize=(10, 6))
sorted_tags_by_views[:20].plot(kind='bar')
plt.title('Top 10 Tags with Highest Average View Count')
plt.xlabel('Tags')
plt.ylabel('Average View Count')

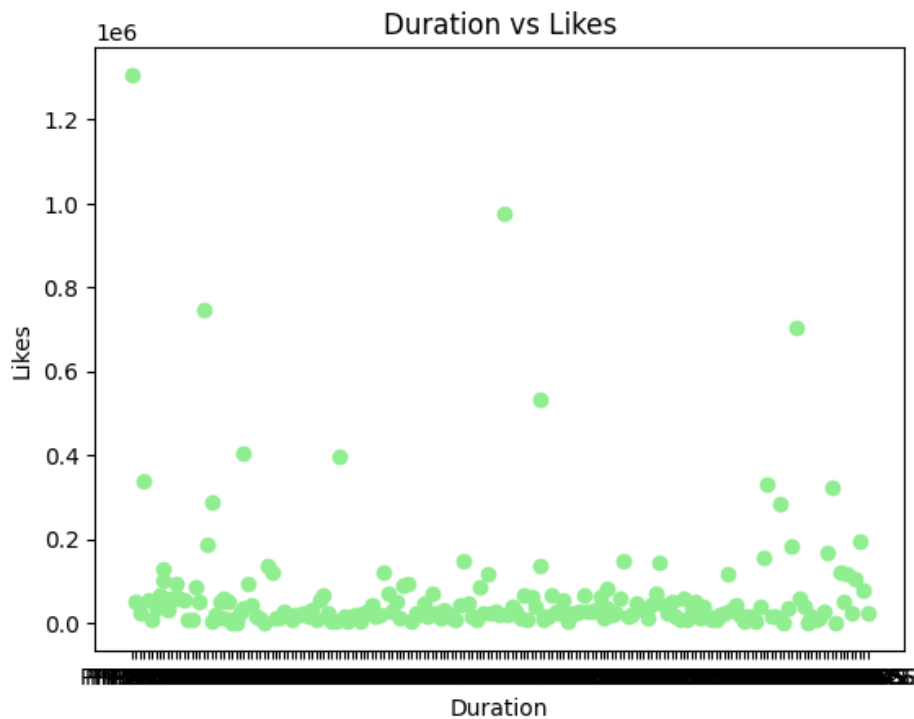
```

```
plt.show()
```



Now, we look into the relationship between the video duration and the number of likes the video received.

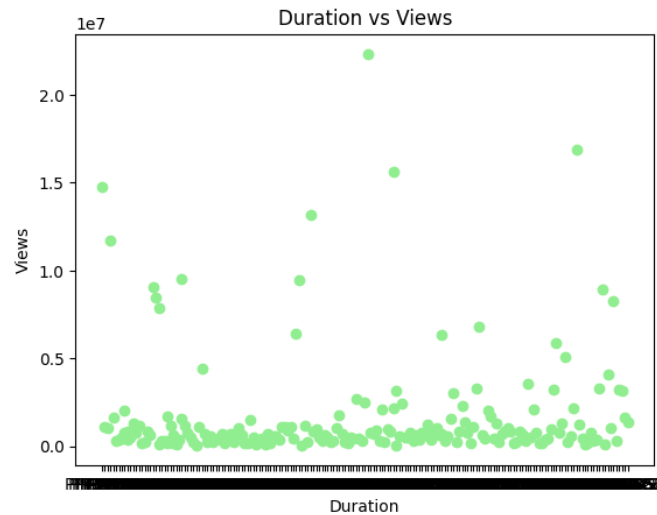
```
plt.scatter(df['duration'], df['like_count'], color='lightgreen')
plt.title('Duration vs Likes')
plt.xlabel('Duration')
plt.ylabel('Likes')
plt.show()
```



It's evident that the videos with the least length are the most likely to be watched. However, the association of duration with the number of likes seems vague.

Lets also peek into the relationship between the duration and number of views.

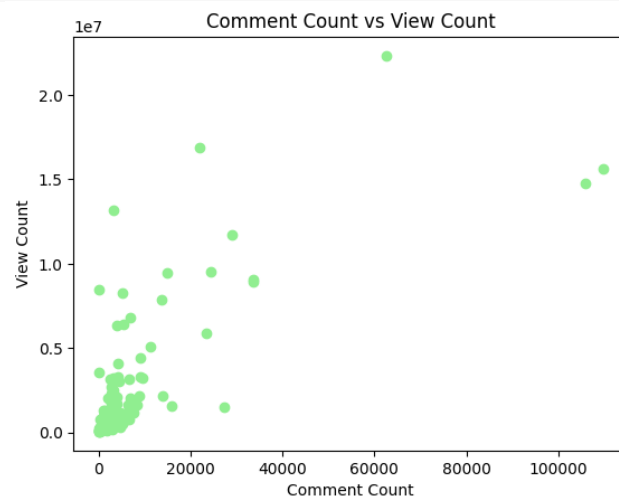
```
plt.scatter(df['duration'], df['view_count'], color='lightgreen')
plt.title('Duration vs Views')
plt.xlabel('Duration')
plt.ylabel('Views')
plt.show()
```



It seems that videos with moderate duration are more likely to be viewed more frequently.

Now, looking at the number of comments, with respect to the number of views.

```
plt.scatter(df['comment_count'], df['view_count'], color='lightgreen')
plt.title('Comment Count vs View Count')
plt.xlabel('Comment Count')
plt.ylabel('View Count')
plt.show()
```



In contrast to my expectations, videos with lesser view counts get fewer comments. And a few videos with more views have more comments.

## Conclusions...

This exploratory data analysis (EDA) provided insights into a YouTube video dataset for videos uploaded during the last week of May 2024. Here's a summary of the key findings:

- **Video characteristics:** The dataset includes video IDs, titles, descriptions (with some missing values), publish dates, channel titles and IDs, categories, tags, durations, definitions (unclear meaning), captions (unclear meaning), view counts, likes, dislikes, comment counts, and favorites counts.
- **Content popularity:** There might be a positive correlation between views and likes, suggesting viewers tend to like videos they watch more.
- **Temporal distribution:** The number of videos uploaded increased towards the end of the analyzed week. Further investigation is required to determine if this is a consistent pattern.
- **Tag analysis:** "Most used tags" might be misleading due to potential skewness towards a specific category (e.g., kids' videos) requiring further exploration. There's a possibility of identifying tags associated with higher average view counts.
- **Video duration and engagement:** There seems to be a weak association between video duration and likes/dislikes. Videos with moderate duration might be more likely to receive views. There's a need for further analysis to solidify these observations.
- **Comments and views:** Contrary to expectations, videos with fewer views have fewer comments, suggesting a low level of interaction for less popular content.

However, there are some limitations:

- The analysis is based on data for a single week, limiting the generalizability of the findings.
- The meaning of "definition" and "caption" data points is unclear and needs clarification.
- Further investigation is required to confirm the observed trends and understand the underlying reasons.

This analysis has applications to future prospects:

- Analyze data from a longer period to understand seasonal or long-term trends.
- Investigate the content of video descriptions (after dealing with missing values) to gain insights into video topics, and NLP can be applied for further investigation.
- Explore sentiment analysis of comments to understand audience reception.

This initial exploration provides a foundation for further analysis of YouTube video data. By delving deeper, you can gain valuable insights into user preferences and content creation strategies for the platform. These insights provide a comprehensive understanding of the trends and patterns in YouTube video data. Content creators can leverage these findings to optimize video uploads, tagging, and content curation strategies to enhance viewer engagement and popularity.