CONTENTS

Chapter 1: Description of Project

Purpose

This project aims to analyze what variables in YouTube videos are more closely associated with high numbers of views per video using SQL query.

Design

YouTube began releasing Click-Through Rate (CTR) data in 2018 for its creators to better understand if their videos catch prosumers' eyeballs. This report looks at various variables like language used, duration of video, gender of YouTuber... etc. to test what matters more for views on YouTube videos.

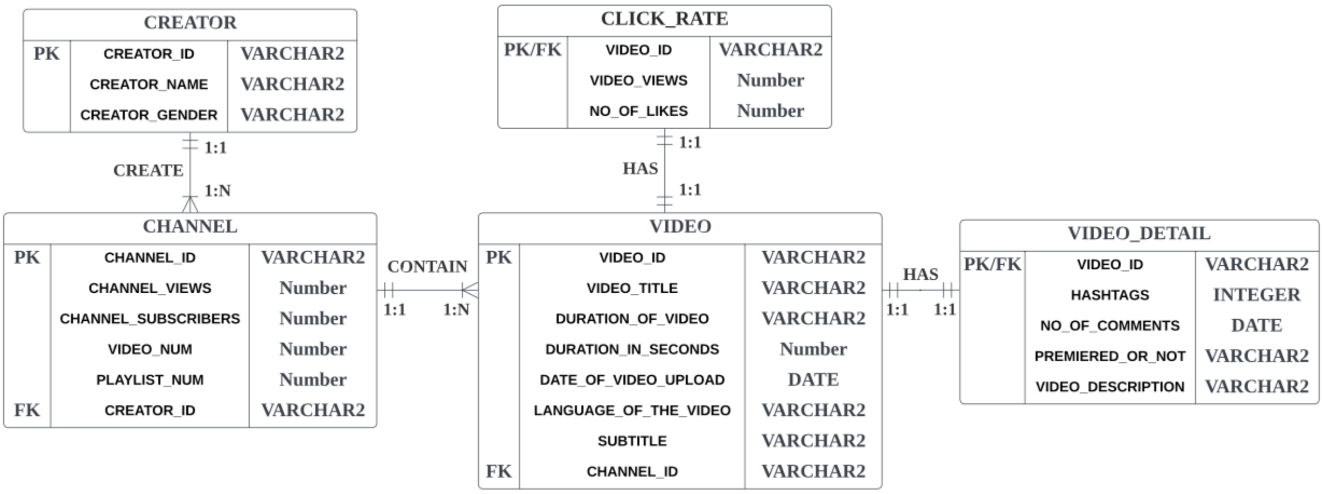
Findings

This report analyzed data containing more than 900 videos, including the first ever YouTube video and videos of 2022. This paper finds limited evidence that higher video views are associated with more numbers of Hashtags, subtitle, or the language used. Instead, videos with higher quality could have significantly more views.

Chapter 2: Conceptual Design/ Logical Data Model

The model contains 5 tables which represents different aspect of the raw dataset:

1. CREATOR table: This table contains attributes about YouTube video creators. One creator can have multiple channels.
2. CHANNEL table: It includes attributes about YouTube channels. Each channel is created by one creator and can contain multiple videos.
3. VIDEO table: It includes information related to the video. Each video can only exist in one channel and can only correspond to one click rate, as well as one video detail.
4. VIDEO_DETAIL table: This table supplements all the information related to the video but is not commonly used, for example, video description.
5. CLICK_RATE table: This table contains important attributes about video click rate. One video detail will correspond to only one video.



CREATOR

PK	CREATOR_ID	VARCHAR2
	CREATOR_NAME	VARCHAR2
	CREATOR_GENDER	VARCHAR2

CLICK_RATE

PK/FK	VIDEO_ID	VARCHAR2
	VIDEO_VIEWS	Number
	NO_OF_LIKES	Number

CHANNEL

PK	CHANNEL_ID	VARCHAR2
	CHANNEL_VIEWS	Number
	CHANNEL_SUBSCRIBERS	Number
	VIDEO_NUM	Number
	PLAYLIST_NUM	Number
FK	CREATOR_ID	VARCHAR2

VIDEO

PK	VIDEO_ID	VARCHAR2
	VIDEO_TITLE	VARCHAR2
	DURATION_OF_VIDEO	VARCHAR2
	DURATION_IN_SECONDS	Number
	DATE_OF_VIDEO_UPLOAD	DATE
	LANGUAGE_OF_THE_VIDEO	VARCHAR2
	SUBTITLE	VARCHAR2
FK	CHANNEL_ID	VARCHAR2

VIDEO_DETAIL

PK/FK	VIDEO_ID	VARCHAR2
	HASHTAGS	INTEGER
	NO_OF_COMMENTS	DATE
	PREMIERED_OR_NOT	VARCHAR2
	VIDEO_DESCRIPTION	VARCHAR2

A Creator could have multiple Channels; one Channel can only has one Creator.
A Channel can contain multiple videos; one video can only exist in one Channel.
Each Video will only have one set of Click Rate; each Click Rate can only correspond to one Video.
Each Video can only has one Video Detail; each Video Detail only belongs to one Video.

Chapter 3: Physical Data Model

CREATOR	CHANNEL
CREATE TABLE CREATOR(CREATOR_ID varchar2(4) PRIMARY KEY, CREATOR_NAME varchar2(100), CREATOR_GENDER varchar2(10));	CREATE TABLE CHANNEL(CHANNEL_ID varchar2(10) PRIMARY KEY, CHANNEL_VIEWS number, CHANNEL_SUBSCRIBERS number, VIDEO_NUM number, PLAYLIST_NUM number, CREATOR_ID varchar2(4), CONSTRAINT CHANNEL_FK FOREIGN KEY (CREATOR_ID) REFERENCES CREATOR(CREATOR_ID));
VIDEO_DETAIL	CLICK_RATE
create TABLE VIDEO_DETAIL(VIDEO_ID VARCHAR(4) PRIMARY KEY, VIDEO_DESCRIPTION VARCHAR(4) NOT NULL, HASHTAGS INTEGER, NO_OF_COMMENTS INTEGER, PREMIERED_OR_NOT VARCHAR(4) NOT NULL, CONSTRAINT VIDEO_DETAIL_FK FOREIGN KEY (VIDEO_ID) REFERENCES VIDEO(VIDEO_ID));	CREATE TABLE CLICK_RATE(VIDEO_ID varchar2(4) PRIMARY KEY, VIDEO_VIEWS number, NO_OF_LIKES number, CONSTRAINT CLICK_RATE_FK FOREIGN KEY (VIDEO_ID) REFERENCES VIDEO(VIDEO_ID));
VIDEO	
CREATE TABLE VIDEO(VIDEO_ID VARCHAR(10) PRIMARY KEY, VIDEO_TITLE VARCHAR(200), DURATION_OF_VIDEO VARCHAR(100), DURATION_IN_SECONDS NUMBER, DATE_OF_VIDEO_UPLOAD DATE, LANGUAGE_OF_THE_VIDEO VARCHAR(15), SUBTITLE VARCHAR(5), CHANNEL_ID VARCHAR(10), CONSTRAINT VIDEO_FK FOREIGN KEY (CHANNEL_ID) REFERENCES CHANNEL(CHANNEL_ID));	

Chapter 4: Data Loading Concept Used

Pre-processing

During the data loading process, we encountered several tasks regarding data cleaning, including dealing with missing data, smoothing out noisy data and correcting inconsistent data. The key concept here is to maintain data consistency and data integrity.

Data Consistency

When loading data, our priority is to make sure that the format within each column is consistent. Since there are a huge amount of inconsistent values in our raw dataset, especially the date values, including different order of year, month and date, and the text/ date format inconsistency. Therefore, we correct them in order to maintain data consistency.

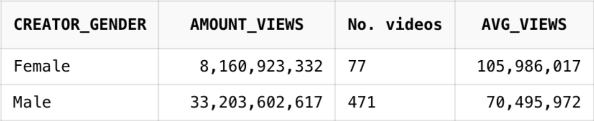
Data Integrity

In order to minimize the negative impacts of uploading tables that could trigger anomalies in the dataset, we have split and joined tables, and identified the entities that rely on one other for existence.

Chapter 5: Ten Insights and Visualization Charts

1. Are female YouTubers more popular than male?

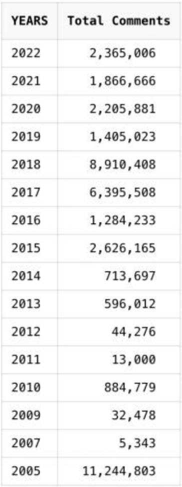
After comparing the average video view of female YouTubers to male YouTubers, we found that, on average, female creators get more clicks than male creators and the difference between the two is up to almost 1.5 times.



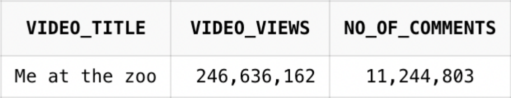
CREATOR_GENDER	AMOUNT_VIEWS	No. videos	AVG_VIEWS
Female	8,160,923,332	77	105,986,017
Male	33,203,602,617	471	70,495,972

2. The number of comments on all videos uploaded each year

The result shows the number of comments on all videos that were uploaded in different years. Worth-notedly, the number in 2005 is distinguished among others. We take a closer look at this anomaly, and found that this 2005 video, *me at the zoo*, is the first ever YouTube video.



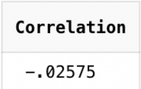
YEARS	Total Comments
2022	2,365,006
2021	1,866,666
2020	2,205,881
2019	1,405,023
2018	8,910,408
2017	6,395,508
2016	1,284,233
2015	2,626,165
2014	713,697
2013	596,012
2012	44,276
2011	13,000
2010	884,779
2009	32,478
2007	5,343
2005	11,244,803



VIDEO_TITLE	VIDEO_VIEWS	NO_OF_COMMENTS
Me at the zoo	246,636,162	11,244,803

3. Does the duration of video affect the video views?

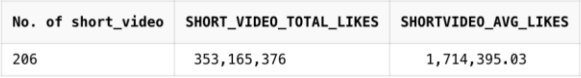
We use the SQL correlation function to get the association between video duration and video views: -0.0257. The result indicates that these two columns are slightly negatively correlated, which means that the longer the video is doesn’t mean the more views it can have.



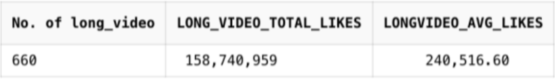
Correlation
-.02575

4. Will short videos have more number of likes?

We define short videos as videos less than three minutes. As the result shows, short videos are more favored than long videos by having 1,722,751 likes on average.



No. of short_video	SHORT_VIDEO_TOTAL_LIKES	SHORTVIDEO_AVG_LIKES
206	353,165,376	1,714,395.03

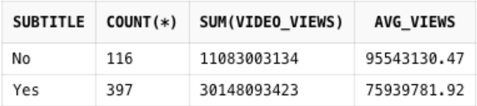


No. of long_video	LONG_VIDEO_TOTAL_LIKES	LONGVIDEO_AVG_LIKES
660	158,740,959	240,516.60

5. Does the existence of subtitles affect video views?

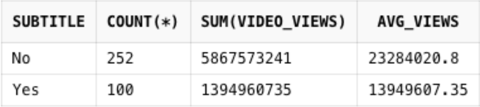
Since English dominates the world, we separate English videos from other languages when analyzing the effect of subtitle. The result shows that the average number of views is higher without subtitle, no matter English or other languages.

(English)



SUBTITLE	COUNT(*)	SUM(VIDEO_VIEWS)	AVG_VIEWS
No	116	11083003134	95543130.47
Yes	397	30148093423	75939781.92

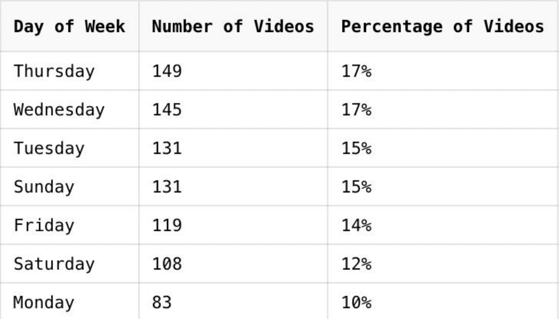
(No English)



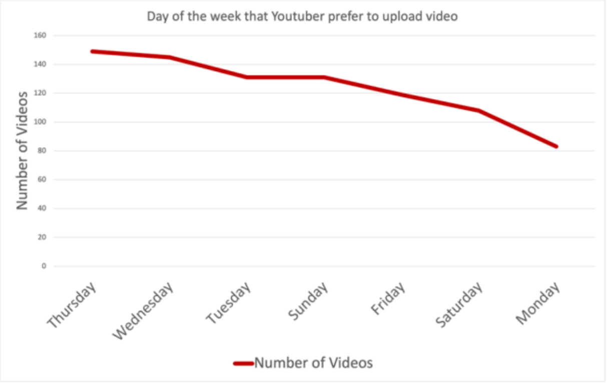
SUBTITLE	COUNT(*)	SUM(VIDEO_VIEWS)	AVG_VIEWS
No	252	5867573241	23284020.8
Yes	100	1394960735	13949607.35

6. What day of the week do Youtubers prefer uploading videos?

The result shows the percentage of each day of a week that YouTubers upload videos. We found that Youtubers prefer posting videos on both Thursday (17%) and Wednesday (17%), while on other days, the percentages only drop slightly. Although there is no big difference in the ratio of other days, Monday has the least percentage of video uploads in a week.

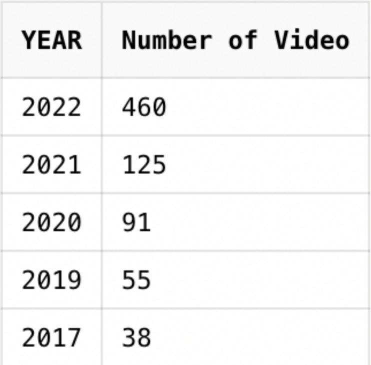


Day of Week	Number of Videos	Percentage of Videos
Thursday	149	17%
Wednesday	145	17%
Tuesday	131	15%
Sunday	131	15%
Friday	119	14%
Saturday	108	12%
Monday	83	10%

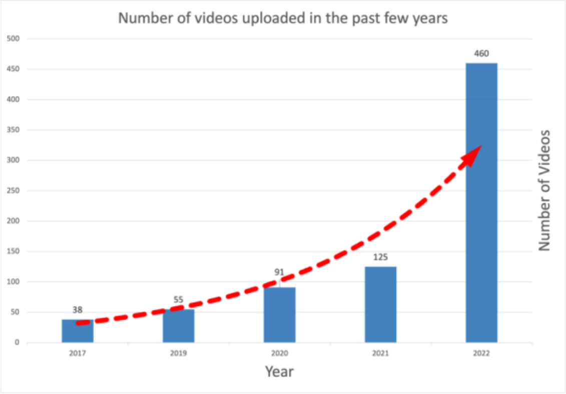


7. The trend of the number of videos uploaded in the past five years.

It can be seen that the number of videos uploaded has increased significantly year by year which we could assume that posting videos on YouTube has become a trend in recent years.



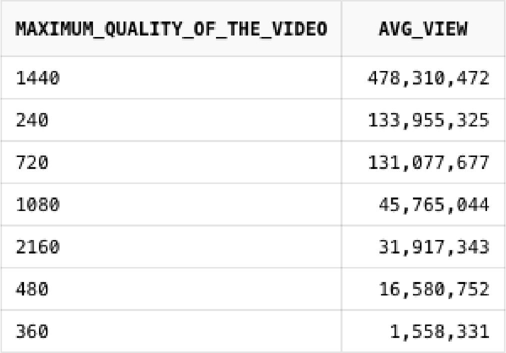
YEAR	Number of Video
2022	460
2021	125
2020	91
2019	55
2017	38



8. Does the Maximum Quality of the video affect Video View?

As technology continues to advance, people can enjoy videos with higher quality. Since the 2160 high-definition videos can only be viewed by paid Premium members, 1440 video becomes the most popular video with the highest click rate.

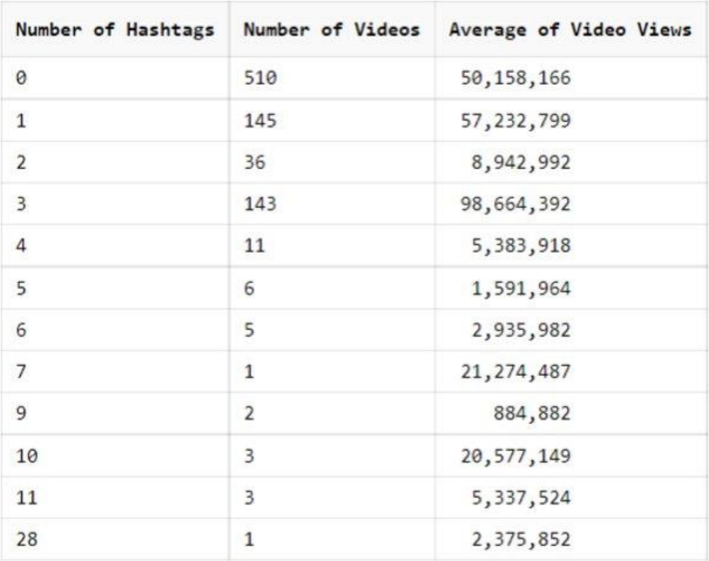
It's worth noting that there is a 240-quality video that has 100 million views, which seems to be inconsistent with our expectations. We dug deeper and found that it was the first video ever on YouTube. Although this video has only 19 seconds, it is the kickstarter of the YouTube era, symbolizing a new way how people consume the media.



MAXIMUM_QUALITY_OF_THE_VIDEO	AVG_VIEW
1440	478,310,472
240	133,955,325
720	131,077,677
1080	45,765,044
2160	31,917,343
480	16,580,752
360	1,558,331

9. Does the number of hashtag influence the video view?

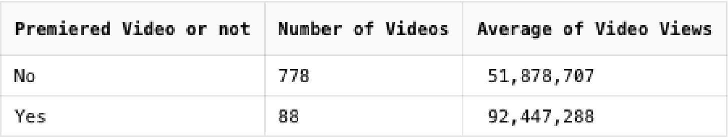
Nowadays, creators mark as many Hashtags as possible under their video to increase views and exposure. Therefore, we want to know if the number of Hashtags really affect the number of views. The result, average views of videos group by the number of Hashtags, is not aligned with Youtubers' expectation as the top 3 most viewed videos are with 3 Hashtags, 1 Hashtag, then 0 Hashtag.



Number of Hashtags	Number of Videos	Average of Video Views
0	510	50,158,166
1	145	57,232,799
2	36	8,942,992
3	143	98,664,392
4	11	5,383,918
5	6	1,591,964
6	5	2,935,982
7	1	21,274,487
9	2	884,882
10	3	20,577,149
11	3	5,337,524
28	1	2,375,852

10. Does video premiered or not influence the video view?

YouTube launched a new "Premiered" video recently, creators can set video showtime and start a countdown. This function could be important to hook viewers to watch your video. As the result shows, the average video views of Premiered video is almost 2 times than that of not Premiered video. Therefore, the data tells us that the "Premiered" function can indeed color the number of views.



Premiered Video or not	Number of Videos	Average of Video Views
No	778	51,878,707
Yes	88	92,447,288