In [1]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:

```
df = pd.read_csv(r"C:\Users\Admin\Downloads\youtube_dislike_dataset (1).csv")
```

In [4]:

```
df.head()
```

Out[4]:

	video_id	title	channel_id	channel_title	published_at	view_count	likes	dislikes	comment_count	
0	--0bCF-iK2E	Jadon Sancho Magical Skills & Goals	UC6UL29enLNe4mqwTfAyeNuw	Bundesliga	2021-07-01 10:00:00	1048888	19515	226	1319	
1	--14w5SOEUs	Migos - Avalanche (Official Video)	UCGIelM2Dj3zza3xyV3pL3WQ	MigosVEVO	2021-06-10 16:00:00	15352638	359277	7479	18729	
2	--40TEbZ9Is	Supporting Actress in a Comedy: 73rd Emmys	UClBKH8yZRcM4AsRjDVEdjMg	Television Academy	2021-09-20 01:03:32	925281	11212	401	831	
3	--4tfbSyYDE	JO1'YOUNG (JO1 ver.)' PERFORMANCE VIDEO	UCsmXiDP8S40uBeJYxvyulmA	JO1	2021-03-03 10:00:17	2641597	39131	441	3745	PF JC
4	--DKkzWVh-E	Why Retaining Walls Collapse	UCMOqf8ab-42UUQIdVoKwjlQ	Practical Engineering	2021-12-07 13:00:00	715724	32887	367	1067	Je

1. Import required libraries and read the provided dataset (youtube_dislike_dataset.csv) and retrieve top 5 and bottom 5 records.

In [5]:

```
df.head(5)
```

Out[5]:

	video_id	title	channel_id	channel_title	published_at	view_count	likes	dislikes	comment_count	
0	--0bCF-iK2E	Jadon Sancho Magical Skills & Goals	UC6UL29enLNe4mqwTfAyeNuw	Bundesliga	2021-07-01 10:00:00	1048888	19515	226	1319	
1	--14w5SOEUs	Migos - Avalanche (Official Video)	UCGIelM2Dj3zza3xyV3pL3WQ	MigosVEVO	2021-06-10 16:00:00	15352638	359277	7479	18729	
2	--40TEbZ9Is	Supporting Actress in a Comedy: 73rd Emmys	UClBKH8yZRcM4AsRjDVEdjMg	Television Academy	2021-09-20 01:03:32	925281	11212	401	831	
3	--4tfbSyYDE	JO1'YOUNG (JO1 ver.)' PERFORMANCE VIDEO	UCsmXiDP8S40uBeJYxvyulmA	JO1	2021-03-03 10:00:17	2641597	39131	441	3745	PF JC
4	--DKkzWVh-E	Why Retaining Walls Collapse	UCMOqf8ab-42UUQIdVoKwjlQ	Practical Engineering	2021-12-07 13:00:00	715724	32887	367	1067	Je

In [6]:

```
df.tail(5)
```

Out[6]:

	video_id	title	channel_id	channel_title	published_at	view_count	likes	dislikes	comment_coun
37417	zzd4ydafGR0	Lil Tjay - Calling My Phone (feat. 6LACK) [Off...	UCEB4a5o_6KfjxHwNMnmj54Q	Lil Tjay	2021-02-12 05:03:49	120408275	2180780	35871	81360
37418	zziBybeSAtw	PELICANS at LAKERS | FULL GAME HIGHLIGHTS | Ja...	UCWJ2lWNubArHWmf3FIHbfcQ	NBA	2021-01-16 05:39:05	2841917	20759	1049	2624
37419	zzk09ESX7e0	[MV] (MAMAMOO) - Where Are We Now	UCuhAUMLzJxlP1W7mEk0_6lA	MAMAMOO	2021-06-02 09:00:10	13346678	720854	4426	90616
37420	zzmQEb0Em5I	FELLIPE ESCUDERO- Master Podcast #12	UC8NjnNWMsRqq11NYvHAQb1g	Master Podcast	2020-10-20 20:59:30	252057	19198	1234	1471
37421	zzxPZwaA-8w	Gareth Bale brace secures dramatic comeback on...	UCEg25rdRZXg32iwai6N6l0w	Tottenham Hotspur	2021-05-23 21:00:31	2252090	34063	868	2004

2. Check the info of the dataframe and write your inferences on data types and shape of the dataset.

In [7]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 37422 entries, 0 to 37421
Data columns (total 12 columns):
 #   Column         Non-Null Count  Dtype 
---  ------         --------------  ----- 
 0   video_id       37422 non-null  object
 1   title          37422 non-null  object
 2   channel_id     37422 non-null  object
 3   channel_title  37422 non-null  object
 4   published_at   37422 non-null  object
 5   view_count     37422 non-null  int64 
 6   likes          37422 non-null  int64 
 7   dislikes       37422 non-null  int64 
 8   comment_count  37422 non-null  int64 
 9   tags           37422 non-null  object
 10  description    37422 non-null  object
 11  comments       37264 non-null  object
dtypes: int64(4), object(8)
memory usage: 3.4+ MB
```

In [8]:

```
df.shape
```

Out[8]:

```
(37422, 12)
```

3. Check for the Percentage of the missing values and drop or impute them.

In [9]:

```
(df.isnull().sum()/df.shape[0])*100
```

Out[9]:

```
video_id         0.000000
title            0.000000
channel_id       0.000000
channel_title    0.000000
published_at     0.000000
view_count       0.000000
likes            0.000000
dislikes         0.000000
comment_count    0.000000
tags             0.000000
description      0.000000
comments         0.422212
dtype: float64
```

4. Check the statistical summary of both numerical and

categorical columns and write your inferences.

In [10]:
```
df.head()
```

Out[10]:

	video_id	title	channel_id	channel_title	published_at	view_count	likes	dislikes	comment_count	
0	--0bCF-iK2E	Jadon Sancho Magical Skills & Goals	UC6UL29enLNe4mqwTfAyeNuw	Bundesliga	2021-07-01 10:00:00	1048888	19515	226	1319	
1	--14w5SOEUs	Migos - Avalanche (Official Video)	UCGIelM2Dj3zza3xyV3pL3WQ	MigosVEVO	2021-06-10 16:00:00	15352638	359277	7479	18729	
2	--40TEbZ9Is	Supporting Actress in a Comedy: 73rd Emmys	UClBKH8yZRcM4AsRjDVEdjMg	Television Academy	2021-09-20 01:03:32	925281	11212	401	831	
3	--4tfbSyYDE	JO1'YOUNG (JO1 ver.)' PERFORMANCE VIDEO	UCsmXiDP8S40uBeJYxvyulmA	JO1	2021-03-03 10:00:17	2641597	39131	441	3745	PF J(
4	--DKkzWVh-E	Why Retaining Walls Collapse	UCMOqf8ab-42UUQIdVoKwjlQ	Practical Engineering	2021-12-07 13:00:00	715724	32887	367	1067	Je

In [11]:
```
df.columns
```

Out[11]:
```
Index(['video_id', 'title', 'channel_id', 'channel_title', 'published_at',
       'view_count', 'likes', 'dislikes', 'comment_count', 'tags',
       'description', 'comments'],
      dtype='object')
```

In [13]:
```
df.describe(include="all")
```

Out[13]:

	video_id	title	channel_id	channel_title	published_at	view_count	likes	dislikes	comment_count
count	37422	37422	37422	37422	37422	3.742200e+04	3.742200e+04	3.742200e+04	3.742200e+04
unique	37422	37113	10961	10883	36772	NaN	NaN	NaN	NaN
top	--0bCF-iK2E	www	UCNAf1k0yIjyGu3k9BwAg3lg	Sky Sports Football	2020-10-16 04:00:10	NaN	NaN	NaN	NaN
freq	1	21	533	533	6	NaN	NaN	NaN	NaN
mean	NaN	NaN	NaN	NaN	NaN	5.697838e+06	1.668147e+05	4.989862e+03	9.924930e+03
std	NaN	NaN	NaN	NaN	NaN	2.426622e+07	5.375670e+05	3.070824e+04	1.171003e+05
min	NaN	NaN	NaN	NaN	NaN	2.036800e+04	0.000000e+00	0.000000e+00	0.000000e+00
25%	NaN	NaN	NaN	NaN	NaN	5.122970e+05	1.323350e+04	2.810000e+02	9.000000e+02
50%	NaN	NaN	NaN	NaN	NaN	1.319078e+06	4.233050e+04	7.960000e+02	2.328000e+03
75%	NaN	NaN	NaN	NaN	NaN	3.670231e+06	1.304698e+05	2.461750e+03	6.184000e+03
max	NaN	NaN	NaN	NaN	NaN	1.322797e+09	3.183768e+07	2.397733e+06	1.607103e+07

5. Convert datatype of column published_at from object to pandas datetime.

In [14]:
```
pd.DataFrame(pd.to_datetime(df['published_at']))
```

Out[14]:

	published_at
0	2021-07-01 10:00:00
1	2021-06-10 16:00:00
2	2021-09-20 01:03:32
3	2021-03-03 10:00:17
4	2021-12-07 13:00:00
...	...
37417	2021-02-12 05:03:49
37418	2021-01-16 05:39:05
37419	2021-06-02 09:00:10
37420	2020-10-20 20:59:30
37421	2021-05-23 21:00:31

37422 rows × 1 columns

6. Create a new column as 'published_month' using the column published_at (display the months only)

In [15]:

```
df['published_month']=df['published_at'].str[5:7]
df[['published_month']]
```

Out[15]:

	published_month
0	07
1	06
2	09
3	03
4	12
...	...
37417	02
37418	01
37419	06
37420	10
37421	05

37422 rows × 1 columns

7. Replace the numbers in the column published_month as names of the months i,e., 1 as 'Jan', 2 as 'Feb' and so on.....

In [16]:

```
month={'01':'Jan','02':'Feb','03':'Mar','04':'Apr','05':'May','06':'Jun','07':'Jul','08':'Aug','09':'Sep','10':
```

In [17]:

```
month
```

Out[17]:

```
{'01': 'Jan',
 '02': 'Feb',
 '03': 'Mar',
 '04': 'Apr',
 '05': 'May',
 '06': 'Jun',
 '07': 'Jul',
 '08': 'Aug',
 '09': 'Sep',
 '10': 'Oct',
 '11': 'Nov',
 '12': 'Dec'}
```

In [18]:

```
df['published_month']=df['published_month'].map(month)
df['published_month']
```

Out[18]:

```
0        Jul
1        Jun
2        Sep
3        Mar
4        Dec
        ...
37417    Feb
37418    Jan
37419    Jun
37420    Oct
37421    May
Name: published_month, Length: 37422, dtype: object
```

8. Find the number of videos published each month and arrange the months in a decreasing order based on the video count.

In [19]:

```
pd.DataFrame(df.groupby('published_month')['video_id'].count().sort_values(ascending=False))
```

Out[19]:

	video_id
published_month	
Oct	4991
Sep	4880
Nov	4851
Aug	4262
Dec	3072
Jul	2340
Jun	2316
Mar	2258
Feb	2137
Apr	2126
Jan	2108
May	2081

9. Find the count of unique video_id, channel_id and channel_title

In [20]:

```
len(df['video_id'].unique()),len(df['channel_id'].unique()),len(df['channel_title'].unique())
```

Out[20]:

```
(37422, 10961, 10883)
```

10. Find the top 10 channel names having the highest number of videos in the dataset and the bottom10 having lowest number of videos.

In [21]:

```
pd.DataFrame(df.groupby('channel_title')['video_id'].count().sort_values(ascending=False).head(10))
```

Out[21]:

	video_id
channel_title	
Sky Sports Football	533
The United Stand	301
BT Sport	246
NBA	209
NFL	162
WWE	122
SSSniperWolf	99
SSundee	98
FORMULA 1	87
NHL	86

In [22]:

```
pd.DataFrame(df.groupby('channel_title')['video_id'].count().sort_values(ascending=False).tail(10))
```

Out[22]:

	video_id
channel_title	
Karchez	1
Karate Combat	1
Kaptain Kuba	1
Kanye West	1
Kannur kitchen	1
Kannada Cinema	1
KanalD	1
Kanak News	1
Kamille Ramos	1
zoom	1

11. Find the title of the video which has the maximum number of likes and the title of the video having minimum likes and write your inferences

In [23]:

```python
pd.DataFrame(df.groupby('title')['likes'].max().sort_values(ascending=False).head(1))
```

Out[23]:

	likes
title	
BTS () 'Dynamite' Official MV	31837675

In [26]:

```python
pd.DataFrame(df.groupby('title')['likes'].max().sort_values(ascending=False).tail(1))
```

Out[26]:

	likes
title	
Kim Kardashian's Must-See Moments on "Saturday Night Live" | E! News	0

12. Find the title of the video which has the maximum number of dislikes and the title of the video having minimum dislikes and write your inferences.

In [48]:

```python
pd.DataFrame(df.groupby('title')['dislikes'].max().sort_values(ascending=False).tail(1))
```

Out[48]:

	dislikes
title	
Kim Kardashian's Must-See Moments on "Saturday Night Live" | E! News	0

In [49]:

```python
pd.DataFrame(df.groupby('title')['dislikes'].max().sort_values(ascending=False).head(1))
```

Out[49]:

	dislikes
title	
Cuties | Official Trailer | Netflix	2397733

13. Does the number of views have any effect on how many people disliked the video? Support your answer with a metric and a plot.

In [29]:

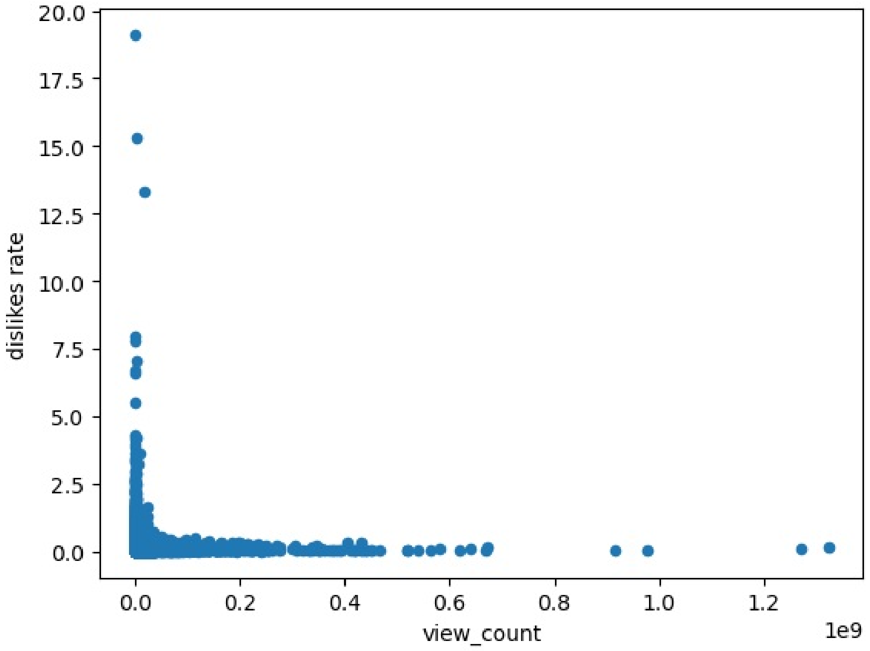
```python
df['dislikes rate']=df['dislikes']/df['view_count']*100
```

In [31]:

```python
pd.DataFrame(df[['dislikes rate','view_count']]).plot(x='view_count',y='dislikes rate',kind='scatter')
```

Out[31]:

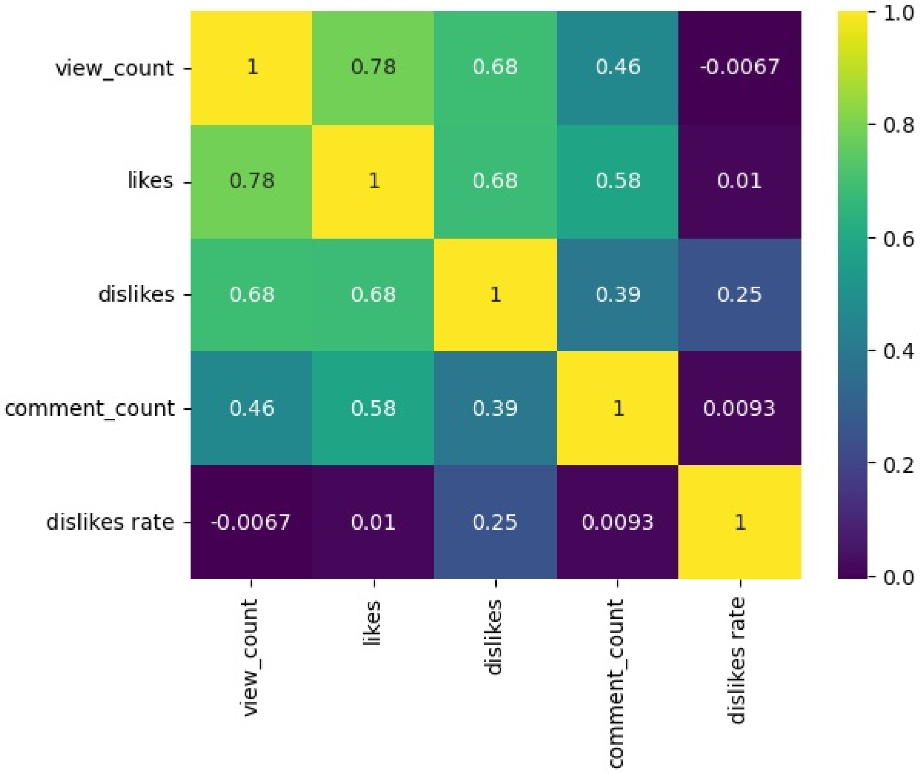
```
<Axes: xlabel='view_count', ylabel='dislikes rate'>
```



```
In [33]: sns.heatmap(df.corr(),annot=True,cmap='viridis')
```

```
C:\Users\Admin\AppData\Local\Temp\ipykernel_11376\3199758246.py:1: FutureWarning: The default value of numeric_
only in DataFrame.corr is deprecated. In a future version, it will default to False. Select only valid columns
or specify the value of numeric_only to silence this warning.
  sns.heatmap(df.corr(),annot=True,cmap='viridis')
```

```
Out[33]: <Axes: >
```



14. Display all the information about the videos that were published in January, and mention the count of videos that were published in January.

```
In [34]: df[df['published_month']=='Jan']
```

Out[34]:

	video_id	title	channel_id	channel_title	published_at	view_count	likes	dislikes	comment_count
27	-2Gwm7QfBnE	Q&A With Naisha	UCYwNMbogQFzMccPSuy-pPWg	MianTwins	2021-01-21 00:05:47	872372	38626	239	621
48	-4sfXSHSxzA	SURPRISING BRENT WITH HIS TIKTOK CRUSH!!	UCPpATKqmMV-CNRNWYaDUwiA	Alexa Rivera	2021-01-16 21:40:04	6504784	262477	5779	7907
95	-AJD1Fc5rpQ	WE ARE HAVING A BABY! | finding out i'm pregna...	UCVsTboAhpnuL6j-tDePvNwQ	Tess Christine	2021-01-03 21:53:48	533084	38965	119	1650
103	-AuJiwjsmWk	Do Ugly Foods Taste Worse? Taste Test	UCzpCc5n9hqiVC7HhPwcIKEg	Good Mythical MORE	2021-01-19 11:00:01	1057077	22526	531	773
182	-JhqO2KWr5U	Schlatt gets fit	UCWZp4y1jqBuvLtiyxSs_ZBw	Big guy	2021-01-24 22:50:57	1724965	119431	325	1578
...	...	...	...	...	...	...	...	...	...
37300	zmzFL5bG-jc	DEVINE MON PERSONNAGE AVANT AKINATOR ! (c'est ...	UCIlr3byh6wmXgcPx_Tm9Ocw	Piwerre	2021-01-16 16:12:19	670357	54462	832	1249
37329	zpzjex7qwrA	Lampard Sacked Within Days Rorys Misery | Chel...	UCkD-ZOixI0a9FjIExDsHsbg	The Kick Off	2021-01-03 20:13:49	428646	12060	296	1505
37345	zqyv-B6mnBM	Lil Wayne - Ain't Got Time (Audio)	UCO9zJy7HWrIS3ojB4Lr7Yqw	Lil Wayne	2021-01-21 05:00:10	2238244	58925	2365	5539
37383	zwfu1-24T7Q	PRADA Cup Day 1 | Full Race Replay | PRADA Cup...	UCo15ZYO_XDRU9LI30OPtxAg	America's Cup	2021-01-15 04:07:55	317382	2008	83	192
37418	zziBybeSAtw	PELICANS at LAKERS | FULL GAME HIGHLIGHTS | Ja...	UCWJ2lWNubArHWmf3FIHbfcQ	NBA	2021-01-16 05:39:05	2841917	20759	1049	2624

2108 rows × 14 columns

In [36]:

```
df[df['published_month']=='Jan']['video_id'].count()
```

Out[36]:

```
2108
```

In [ ]:

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js