In [1]:

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

In [2]:

```
india_df=pd.read_csv('INDIAvi.csv')
```

```
In [3]: india_df
```

Out[3]:

	video_id	trending_date	title	channel_title	category_id	publish_time
0	kzwfHumJyYc	17.14.11	Sharry Mann: Cute Munda (Song Teaser) | Parmi...	Lokdhun Punjabi	1	2017-11-12T12:20:39.000Z
1	zUZ1z7FwLc8	17.14.11	पीरियड्स के समय, पेट पर पति करता ऐसा, देखकर दं...	HJ NEWS	25	2017-11-13T05:43:56.000Z
2	10L1hZ9qa58	17.14.11	Stylish Star Allu Arjun @ ChaySam Wedding Rece...	TFPC	24	2017-11-12T15:48:08.000Z
3	N1vE8iiEg64	17.14.11	Eruma Saani | Tamil vs English	Eruma Saani	23	2017-11-12T07:08:48.000Z
4	kJzGH0PVQHQ	17.14.11	why Samantha became EMOTIONAL @ Samantha naga ...	Filmylooks	24	2017-11-13T01:14:16.000Z
...	...	...	...	...	...	...
37347	iNHecA3PJCo	18.14.06	फेकू आशिक़ - राजस्थान की सबसे शानदार कॉमेडी | ...	RDC Rajasthani	23	2018-06-13T08:01:11.000Z
37348	dpPmPbhcsIM	18.14.06	Seetha | Flowers | Ep# 364	Flowers TV	24	2018-06-13T11:30:04.000Z
37349	mV6aztP58f8	18.14.06	Bhramanam I Episode 87 - 12 June 2018 I Mazhav...	Mazhavil Manorama	24	2018-06-13T05:00:02.000Z
37350	qxqDNP1bDEw	18.14.06	Nua Bohu | Full Ep 285 | 13th June 2018 | Odia...	Tarang TV	24	2018-06-13T15:07:49.000Z
37351	wERgpPK44w0	18.14.06	Ee Nagaraniki Emaindi Trailer | Tharun Bhascke...	Suresh Productions	24	2018-06-10T04:29:54.000Z

37352 rows × 16 columns

In [4]:
```
india_df.head()
```

Out[4]:

	video_id	trending_date	title	channel_title	category_id	publish_time	
0	kzwfHumJyYc	17.14.11	Sharry Mann: Cute Munda (Song Teaser) | Parmi...	Lokdhun Punjabi	1	2017-11-12T12:20:39.000Z	sha son
1	zUZ1z7FwLc8	17.14.11	पीरियड्स के समय, पेट पर पति करता ऐसा, देखकर दं...	HJ NEWS	25	2017-11-13T05:43:56.000Z	पीरि
2	10L1hZ9qa58	17.14.11	Stylish Star Allu Arjun @ ChaySam Wedding Rece...	TFPC	24	2017-11-12T15:48:08.000Z	Stylis @ Ch
3	N1vE8iiEg64	17.14.11	Eruma Saani | Tamil vs English	Eruma Saani	23	2017-11-12T07:08:48.000Z	Eru Video
4	kJzGH0PVQHQ	17.14.11	why Samantha became EMOTIONAL @ Samantha naga ...	Filmylooks	24	2017-11-13T01:14:16.000Z	F

In [5]:
```
type(india_df)
```

Out[5]:
```
pandas.core.frame.DataFrame
```

In [6]:
```
india_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 37352 entries, 0 to 37351
Data columns (total 16 columns):
 #   Column                  Non-Null Count  Dtype 
---  ------                  --------------  ----- 
 0   video_id                37352 non-null  object
 1   trending_date           37352 non-null  object
 2   title                   37352 non-null  object
 3   channel_title           37352 non-null  object
 4   category_id             37352 non-null  int64 
 5   publish_time            37352 non-null  object
 6   tags                    37352 non-null  object
 7   views                   37352 non-null  int64 
 8   likes                   37352 non-null  int64 
 9   dislikes                37352 non-null  int64 
 10  comment_count           37352 non-null  int64 
 11  thumbnail_link          37352 non-null  object
 12  comments_disabled       37352 non-null  bool  
 13  ratings_disabled        37352 non-null  bool  
 14  video_error_or_removed  37352 non-null  bool  
 15  description             36791 non-null  object
dtypes: bool(3), int64(5), object(8)
memory usage: 3.8+ MB
```

In [7]:
```
india_df.describe()
```

Out[7]:

	category_id	views	likes	dislikes	comment_count
count	37352.000000	3.735200e+04	3.735200e+04	3.735200e+04	37352.00000
mean	21.576596	1.060478e+06	2.708272e+04	1.665082e+03	2676.99743
std	6.556593	3.184932e+06	9.714510e+04	1.607617e+04	14868.31713
min	1.000000	4.024000e+03	0.000000e+00	0.000000e+00	0.00000
25%	23.000000	1.239155e+05	8.640000e+02	1.080000e+02	81.00000
50%	24.000000	3.045860e+05	3.069000e+03	3.260000e+02	329.00000
75%	24.000000	7.992912e+05	1.377425e+04	1.019250e+03	1285.00000
max	43.000000	1.254322e+08	2.912710e+06	1.545017e+06	827755.00000

In [8]:
```
# looking for unique values
```

In [9]:
```
india_df.nunique()
```

Out[9]:
```
video_id                  16307
trending_date               205
title                     16721
channel_title              1426
category_id                  17
publish_time              16339
tags                      12578
views                     32136
likes                     15529
dislikes                   5079
comment_count              6027
thumbnail_link            16523
comments_disabled             2
ratings_disabled              2
video_error_or_removed        2
description               13992
dtype: int64
```

In [10]:
```
# cleaning the data
```

In [11]:
```
# checking for null values
```

In [12]:
```
india_df.isnull().sum()
```

Out[12]:
```
video_id                    0
trending_date               0
title                       0
channel_title               0
category_id                 0
publish_time                0
tags                        0
views                       0
likes                       0
dislikes                    0
comment_count               0
thumbnail_link              0
comments_disabled           0
ratings_disabled            0
video_error_or_removed      0
description               561
dtype: int64
```

In [13]:
```
# removing duplicates
```

In [14]:
```
df=india_df.drop_duplicates()
```

In [15]:
```
df.shape
```

Out[15]:
```
(33089, 16)
```

In [16]:
```
df=df.drop(['video_id','title','tags','thumbnail_link','comments_disabled','ra
```

In [17]:
```
df
```

Out[17]:

	trending_date	channel_title	category_id	publish_time	views	likes	dislikes	cc
0	17.14.11	Lokdhun Punjabi	1	2017-11-12T12:20:39.000Z	1096327	33966	798	
1	17.14.11	HJ NEWS	25	2017-11-13T05:43:56.000Z	590101	735	904	
2	17.14.11	TFPC	24	2017-11-12T15:48:08.000Z	473988	2011	243	
3	17.14.11	Eruma Saani	23	2017-11-12T07:08:48.000Z	1242680	70353	1624	
4	17.14.11	Filmylooks	24	2017-11-13T01:14:16.000Z	464015	492	293	
...	...	...	...	...	...	...	...	
37300	18.14.06	The Timeliners	24	2018-06-08T13:54:39.000Z	2675706	96485	4181	
37301	18.14.06	WWE	17	2018-06-13T03:09:21.000Z	770873	13316	552	
37302	18.14.06	Dharma Productions	1	2018-06-11T06:50:41.000Z	27696924	468472	60025	
37319	18.14.06	Angry Prash	23	2018-06-11T08:37:21.000Z	1214423	85601	4677	
37330	18.14.06	Warangal Diaries	23	2018-06-13T10:16:21.000Z	132055	11170	393	

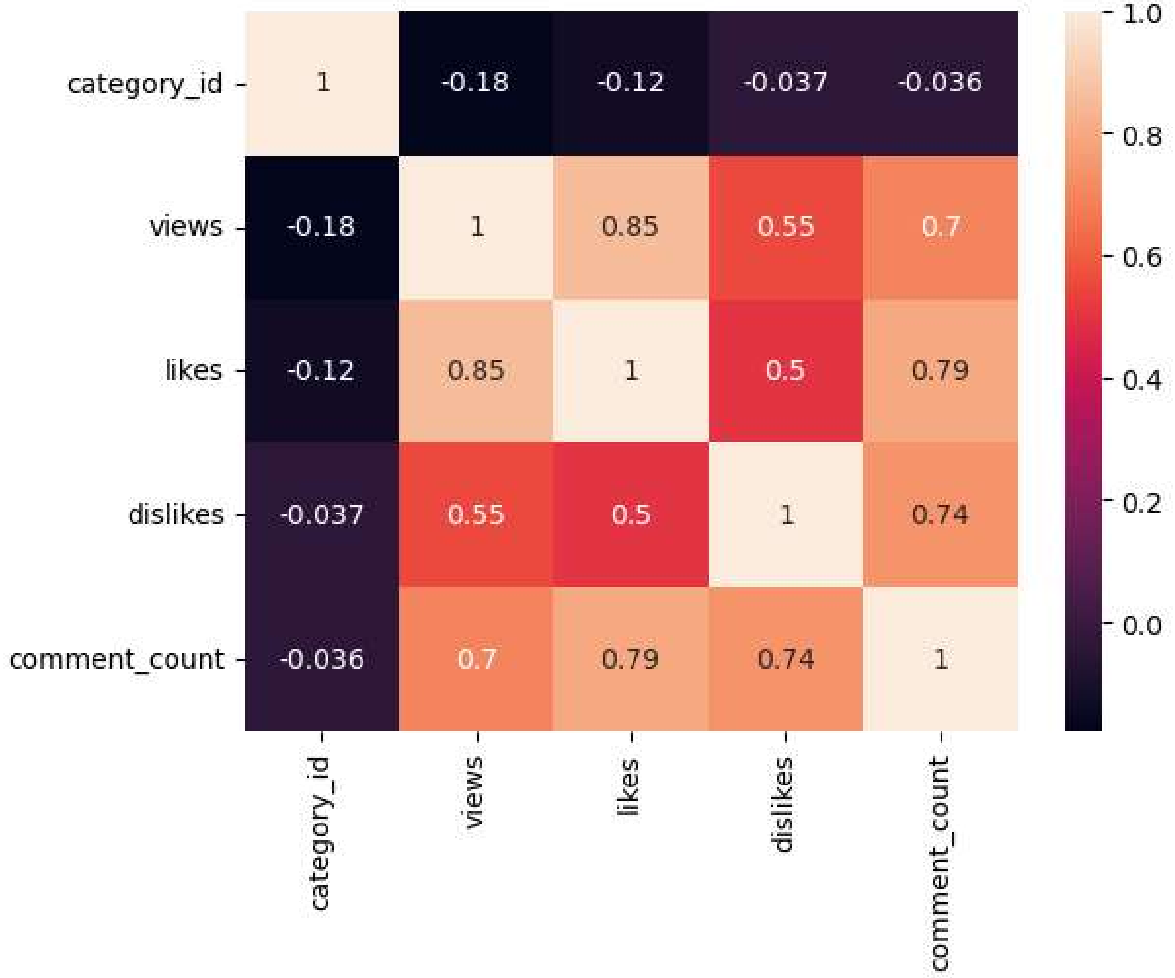
33089 rows × 8 columns

relationship analysis

In [18]:
```
corelation=df.corr()
```

```
C:\Users\pooja sharma\AppData\Local\Temp\ipykernel_6716\3476424618.py:1: Futu
reWarning: 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 s
pecify the value of numeric_only to silence this warning.
  corelation=df.corr()
```

In [19]:
```
sns.heatmap(corelation, xticklabels=corelation.columns, yticklabels=corelation
```



In [20]:
```
#taking a random sample
```

In [21]:
```
df.sample(10)
```

Out[21]:

	trending_date	channel_title	category_id	publish_time	views	likes	dislikes	com
4234	17.06.12	NewsGlitz - Next Generation Tamil News Channel	25	2017-12-04T13:44:31.000Z	257097	2638	553	
3024	17.29.11	Times Music Tamil	22	2017-11-24T14:30:59.000Z	1832370	96991	3210	
15614	18.07.02	Tarang TV	24	2018-02-06T05:24:10.000Z	39469	89	3	
46	17.14.11	Vikram Aditya	24	2017-11-13T03:04:30.000Z	127517	3676	381	
24954	18.02.04	Muzik247	1	2018-03-31T11:21:41.000Z	173565	5335	73	
18755	18.25.02	Hyderabad Diaries	24	2018-02-23T11:40:02.000Z	122312	13443	251	
18806	18.25.02	TsMadaan	27	2018-02-22T03:53:25.000Z	238552	8878	457	
19531	18.02.03	Mazhavil Manorama	24	2018-03-01T05:00:01.000Z	443952	2057	402	
329	17.15.11	Troom Troom	26	2017-11-12T15:00:05.000Z	3897195	31125	2771	
9745	18.04.01	NDTV	25	2018-01-02T10:34:15.000Z	331059	483	96	

exploring the variables

In [22]:
```
import matplotlib

sns.set_style('darkgrid')
matplotlib.rcParams['font.size']=14
matplotlib.rcParams['figure.figsize']=(9,5)
matplotlib.rcParams['figure.facecolor']='#00000000'
```

In [23]:
```
# identifying top channels
```

In [24]:
```
top_channels=df.channel_title.value_counts().head(15)
top_channels
```

Out[24]:
```
VikatanTV              208
SAB TV                 206
ETV Plus India         206
etvteluguindia         205
Study IQ education     202
Flowers Comedy         202
SET India              199
Tarang TV              199
Mazhavil Manorama      196
RadaanMedia            193
V6 News Telugu         190
Technical Guruji       189
T-Series               188
ETV Jabardasth         185
mallemalatv            184
Name: channel_title, dtype: int64
```

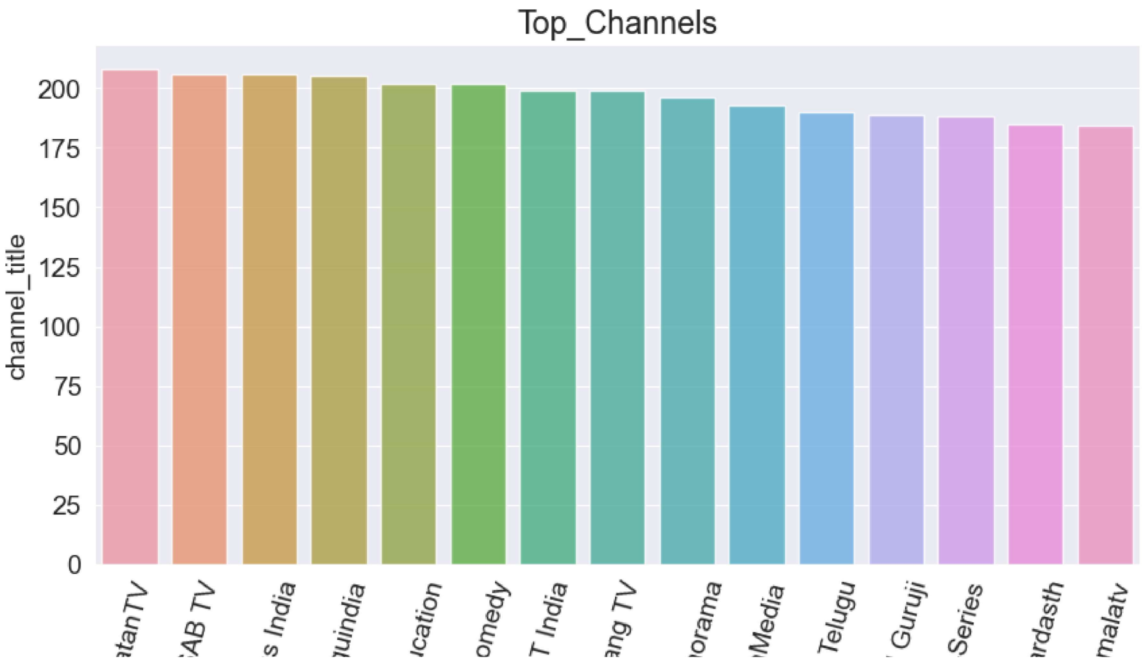
In [25]:
```
# bottom channels
```

In [26]:
```
bottom_channels=df.channel_title.value_counts().tail(15)
bottom_channels
```

Out[26]:
```
BapaoGiri               1
Top 5                   1
Kerala Fans Club        1
Telugu Trending         1
Netflix India           1
Illumination            1
gallinews               1
Jaaz Multimedia         1
BLUSH                   1
Charan TV Online        1
All Updates             1
Challenge Mantra        1
Alpha Digitech          1
YouTube Got Talent      1
PropheC Productions     1
Name: channel_title, dtype: int64
```

visualisisng top channels using bar cart

In [27]:
```
plt.figure(figsize=(10,5))
plt.xticks(rotation=75)
plt.title('Top_Channels')
sns.barplot(x=top_channels.index, y=top_channels, alpha=0.8);
```



hypothesis testing using Z test for likes

In [28]:
```
import math
```

In [29]:
```
# population mean
```

In [30]:
```
pop_mean=df.likes.mean()
pop_mean
```

Out[30]:
```
25587.621052313458
```

In [31]:
```
# creating the hypothesis
```

In [32]:
```
#      Null hypothesis H0: μ=25587.6
# Alternate hypothesis H1: μ!=25587.6
```

In [33]:
```
# taking α=0.05, Z=±1.96
```

In [34]:
```
# taking a random sample of 1000 and getting sample mean
```

In [35]:
```
sample_mean=df.sample(1000).likes.mean()
sample_mean
```

Out[35]:
```
30934.174
```

```
In [36]: # standard deviation

In [37]: std=np.std(df.likes)

In [38]: std

Out[38]: 96471.73722117719

In [39]: # calculating for z

In [40]: (pop_mean-sample_mean)/(std/math.sqrt(1000))

Out[40]: -1.7525635416530168

In [42]: # calculated z score -1.75 is more than -1.96, so we do not reject the null hyp
         # observed z score is -1.75
         # critical value is -1.96

In [43]: df.to_csv("INDIAvi.csv")

In [ ]:
```