In [4]:

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

In [5]:

```
df = pd.read_csv('USvideos.csv')
```

In [6]:

```
df.head()
```

Out[6]:

	video_id	trending_date	title	channel_title	category_id	publish_tir
0	2kyS6SvSYSE	17.14.11	WE WANT TO TALK ABOUT OUR MARRIAGE	CaseyNeistat	22	2017-1 13T17:13:01.00
1	1ZAPwfrtAFY	17.14.11	The Trump Presidency: Last Week Tonight with J...	LastWeekTonight	24	2017-1 13T07:30:00.00
2	5qpjK5DgCt4	17.14.11	Racist Superman | Rudy Mancuso, King Bach & Le...	Rudy Mancuso	23	2017-1 12T19:05:24.00
3	puqaWrEC7tY	17.14.11	Nickelback Lyrics: Real or Fake?	Good Mythical Morning	24	2017-1 13T11:00:04.00
4	d380meD0W0M	17.14.11	I Dare You: GOING BALD!?	nigahiga	24	2017-1 12T18:01:41.00

In [7]:

```
df.shape
```

Out[7]:

```
(40949, 16)
```

In [8]:
```
df.describe()
```

Out[8]:

	category_id	views	likes	dislikes	comment_count
count	40949.000000	4.094900e+04	4.094900e+04	4.094900e+04	4.094900e+04
mean	19.972429	2.360785e+06	7.426670e+04	3.711401e+03	8.446804e+03
std	7.568327	7.394114e+06	2.288853e+05	2.902971e+04	3.743049e+04
min	1.000000	5.490000e+02	0.000000e+00	0.000000e+00	0.000000e+00
25%	17.000000	2.423290e+05	5.424000e+03	2.020000e+02	6.140000e+02
50%	24.000000	6.818610e+05	1.809100e+04	6.310000e+02	1.856000e+03
75%	25.000000	1.823157e+06	5.541700e+04	1.938000e+03	5.755000e+03
max	43.000000	2.252119e+08	5.613827e+06	1.674420e+06	1.361580e+06

In [9]:
```
df=df.drop_duplicates()
df.shape
```

Out[9]:
```
(40901, 16)
```

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

```
<class 'pandas.core.frame.DataFrame'>
Index: 40901 entries, 0 to 40948
Data columns (total 16 columns):
 #   Column                  Non-Null Count  Dtype 
---  ------                  --------------  ----- 
 0   video_id                40901 non-null  object
 1   trending_date           40901 non-null  object
 2   title                   40901 non-null  object
 3   channel_title           40901 non-null  object
 4   category_id             40901 non-null  int64 
 5   publish_time            40901 non-null  object
 6   tags                    40901 non-null  object
 7   views                   40901 non-null  int64 
 8   likes                   40901 non-null  int64 
 9   dislikes                40901 non-null  int64 
 10  comment_count           40901 non-null  int64 
 11  thumbnail_link          40901 non-null  object
 12  comments_disabled       40901 non-null  bool  
 13  ratings_disabled        40901 non-null  bool  
 14  video_error_or_removed  40901 non-null  bool  
 15  description             40332 non-null  object
dtypes: bool(3), int64(5), object(8)
memory usage: 4.5+ MB
```

In [11]:

```
columns_to_remove = ['thumbnail_link', 'description']
df = df.drop(columns=columns_to_remove)
df.info()
```

```
Index: 40901 entries, 0 to 40948
Data columns (total 14 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   video_id                40901 non-null  object
 1   trending_date           40901 non-null  object
 2   title                   40901 non-null  object
 3   channel_title           40901 non-null  object
 4   category_id             40901 non-null  int64
 5   publish_time            40901 non-null  object
 6   tags                    40901 non-null  object
 7   views                   40901 non-null  int64
 8   likes                   40901 non-null  int64
 9   dislikes                40901 non-null  int64
 10  comment_count           40901 non-null  int64
 11  comments_disabled       40901 non-null  bool
 12  ratings_disabled        40901 non-null  bool
 13  video_error_or_removed  40901 non-null  bool
dtypes: bool(3), int64(5), object(6)
memory usage: 3.9+ MB
```

In [13]:

```
from datetime import datetime
```

In [14]:

```
import datetime
```

In [20]:

```
df['publish_time'] = pd.to_datetime(df['publish_time'])
df.head(2)
```

Out[20]:

	video_id	trending_date	title	channel_title	category_id	publish_time
0	2kyS6SvSYSE	17.14.11	WE WANT TO TALK ABOUT OUR MARRIAGE	CaseyNeistat	22	2017-11-13 17:13:01+00:00
1	1ZAPwfrtAFY	17.14.11	The Trump Presidency: Last Week Tonight with J...	LastWeekTonight	24	2017-11-13 07:30:00+00:00

In [21]:

```
df['publish_month'] = df['publish_time'].dt.month
df['publish_day'] = df['publish_time'].dt.day
df['publish_hour'] = df['publish_time'].dt.hour
df.head(2)
```

Out[21]:

	video_id	trending_date	title	channel_title	category_id	publish_time
0	2kyS6SvSYSE	17.14.11	WE WANT TO TALK ABOUT OUR MARRIAGE	CaseyNeistat	22	2017-11-13 17:13:01+00:00
1	1ZAPwfrtAFY	17.14.11	The Trump Presidency: Last Week Tonight with J...	LastWeekTonight	24	2017-11-13 07:30:00+00:00

In [22]:

```
print(sorted(df["category_id"].unique()))
[1,2,10,15,17,19,20,22,23,24,25,26,27,28,29,30,43]
```

```
[1, 2, 10, 15, 17, 19, 20, 22, 23, 24, 25, 26, 27, 28, 29, 43]
```

Out[22]:

```
[1, 2, 10, 15, 17, 19, 20, 22, 23, 24, 25, 26, 27, 28, 29, 30, 43]
```

In [23]:

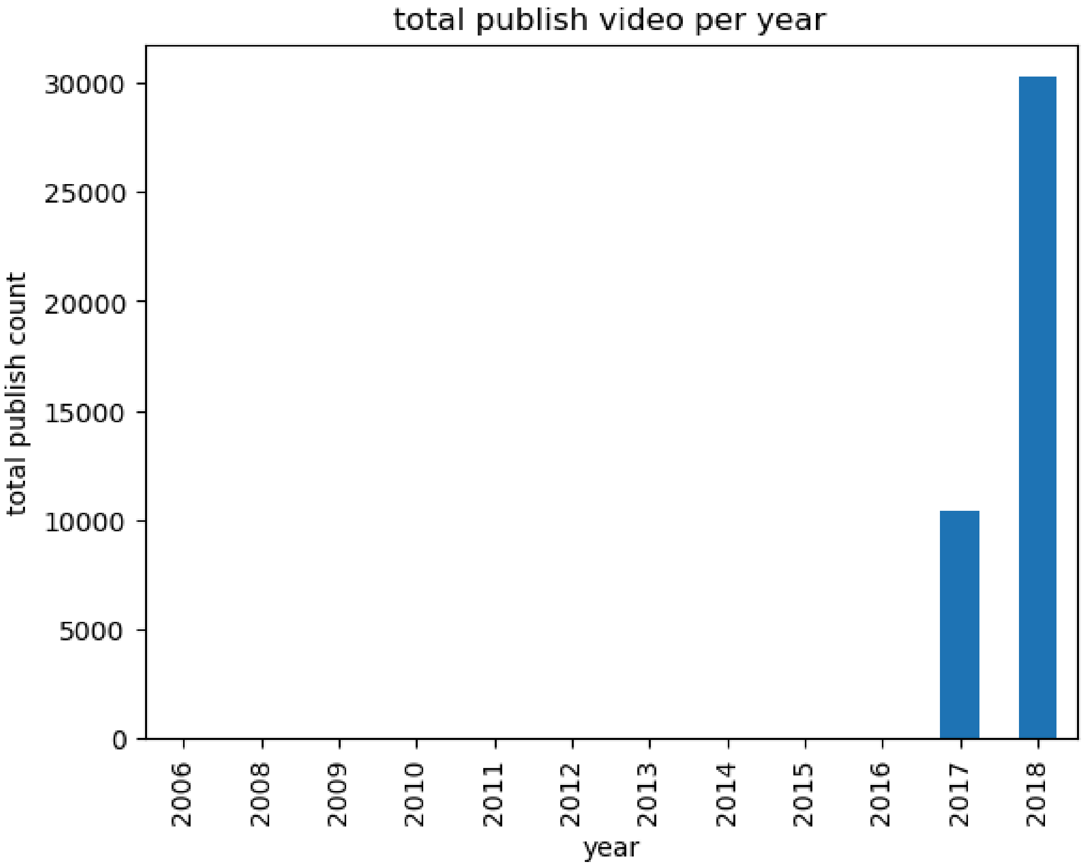
```
df['category_name'] = np.nan
df.loc[(df["category_id"] ==1), "category_name"] = 'film and animation'
df.loc[(df["category_id"] ==2), "category_name"] = 'autos and vehices'
df.loc[(df["category_id"] ==10), "category_name"] = 'music'
df.loc[(df["category_id"] ==15), "category_name"] = 'pets and animals'
df.loc[(df["category_id"] ==17), "category_name"] = 'sports'
df.loc[(df["category_id"] ==19), "category_name"] = 'travel  and events'
df.loc[(df["category_id"] ==20), "category_name"] = 'gaming'
df.loc[(df["category_id"] ==22), "category_name"] = 'people and blogs'
df.loc[(df["category_id"] ==23), "category_name"] = 'comedy'
df.loc[(df["category_id"] ==24), "category_name"] = 'entertainment'
df.loc[(df["category_id"] ==25), "category_name"] = 'news and politics'
df.loc[(df["category_id"] ==26), "category_name"] = 'how to and style'
df.loc[(df["category_id"] ==27), "category_name"] = 'education'
df.loc[(df["category_id"] ==28), "category_name"] = 'science and technology'
df.loc[(df["category_id"] ==29), "category_name"] = 'non profits and activities'
df.loc[(df["category_id"] ==30), "category_name"] = 'movies'
df.loc[(df["category_id"] ==43), "category_name"] = 'shows'

df.head()
```

Out[23]:

	video_id	trending_date	title	channel_title	category_id	publish_time
0	2kyS6SvSYSE	17.14.11	WE WANT TO TALK ABOUT OUR MARRIAGE	CaseyNeistat	22	2017-11-13 17:13:01+00:00
1	1ZAPwfrtAFY	17.14.11	The Trump Presidency: Last Week Tonight with J...	LastWeekTonight	24	2017-11-13 07:30:00+00:00
2	5qpjK5DgCt4	17.14.11	Racist Superman | Rudy Mancuso, King Bach & Le...	Rudy Mancuso	23	2017-11-12 19:05:24+00:00
3	puqaWrEC7tY	17.14.11	Nickelback Lyrics: Real or Fake?	Good Mythical Morning	24	2017-11-13 11:00:04+00:00
4	d380meD0W0M	17.14.11	I Dare You: GOING BALD!?	nigahiga	24	2017-11-12 18:01:41+00:00

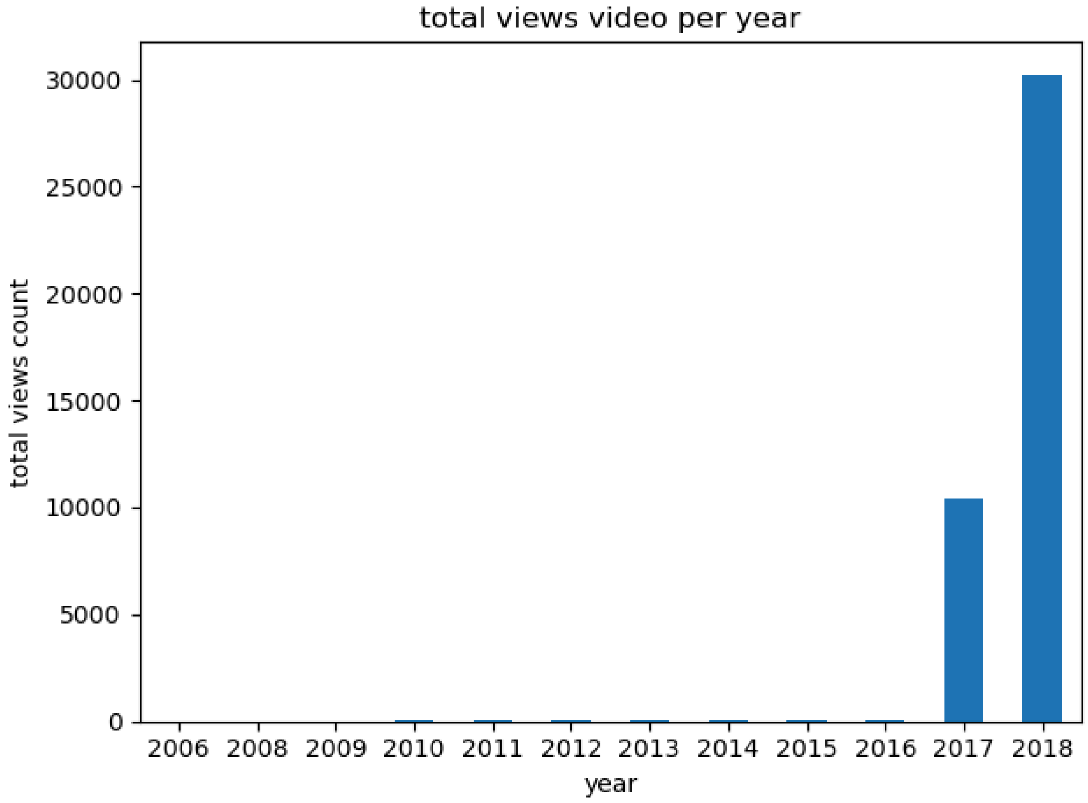
In [24]:

```
df['year'] = df['publish_time'].dt.year
yearly_counts = df.groupby('year')['video_id'].count()
yearly_counts.plot(kind='bar', xlabel='year', ylabel = 'total publish count', title = 'to
plt.show()
```



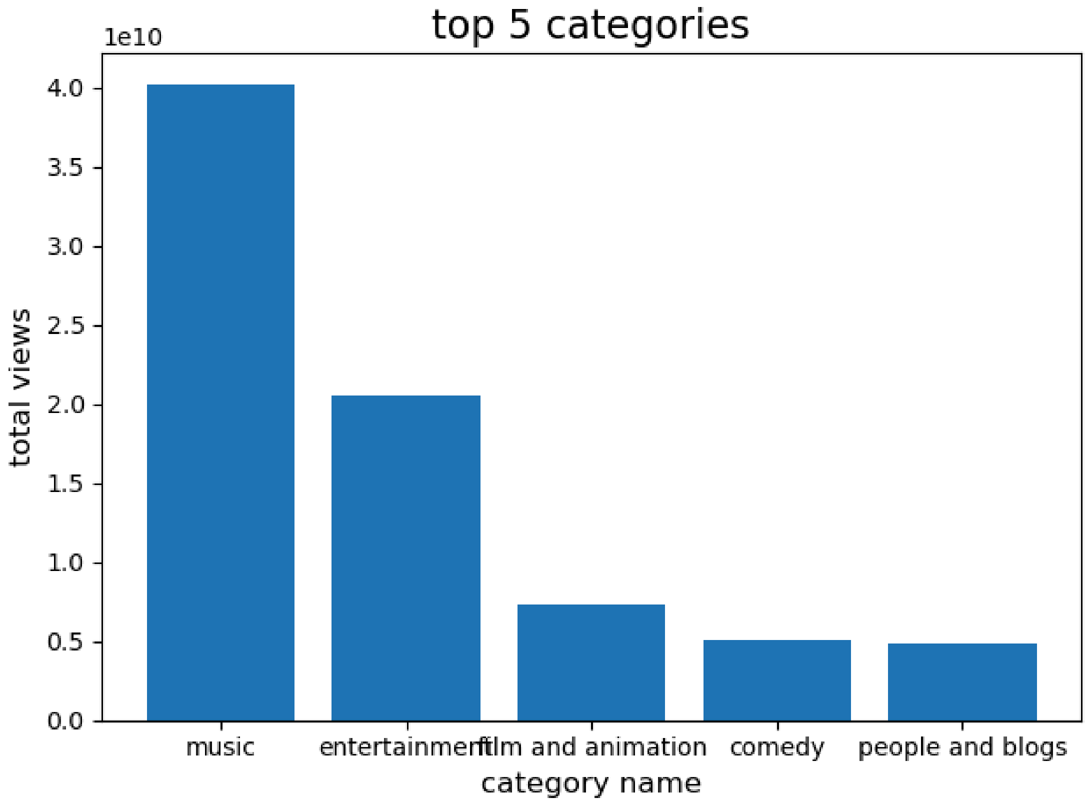
In [25]:

```
yearly_views = df.groupby('year')['views'].count()
yearly_views.plot(kind='bar', xlabel='year', ylabel = 'total views count', title = 'total
plt.xticks(rotation=0)
plt.tight_layout()
plt.show()
```



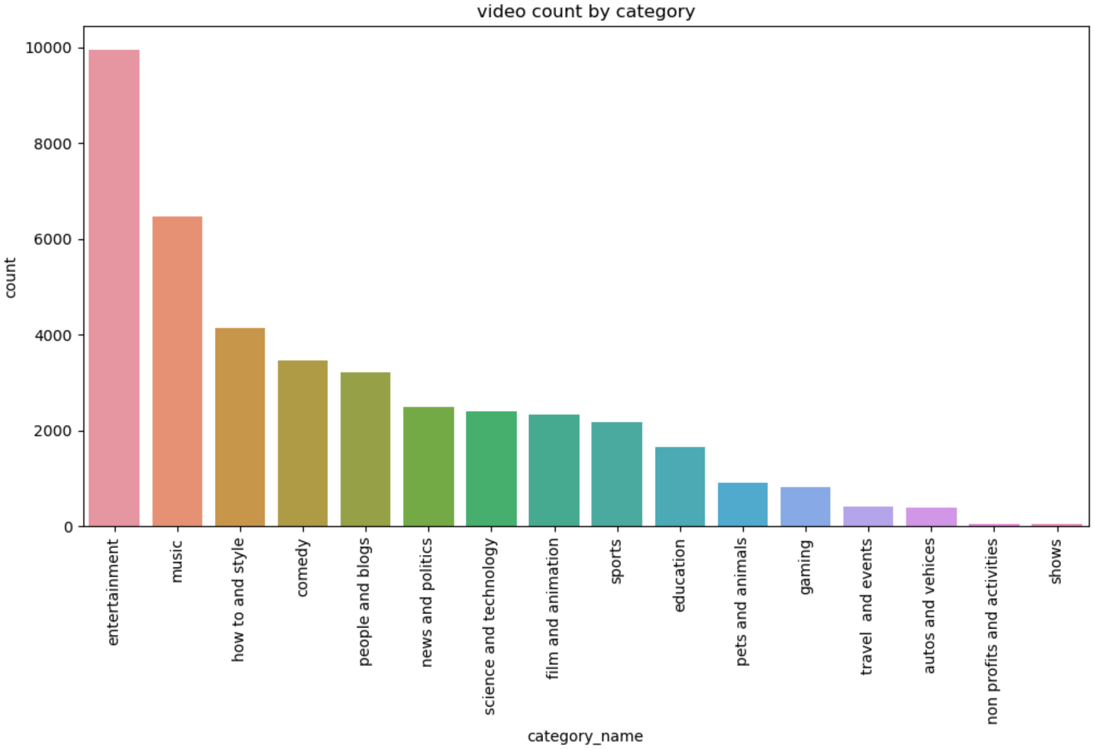
In [26]:

```
category_views = df.groupby('category_name')['views'].sum().reset_index()
top_categories = category_views.sort_values(by='views', ascending=False).head(5)
plt.bar(top_categories['category_name'], top_categories['views'])
plt.xlabel('category name', fontsize = 12)
plt.ylabel('total views', fontsize=12)
plt.title('top 5 categories', fontsize = 16)
plt.tight_layout()
plt.show()
```



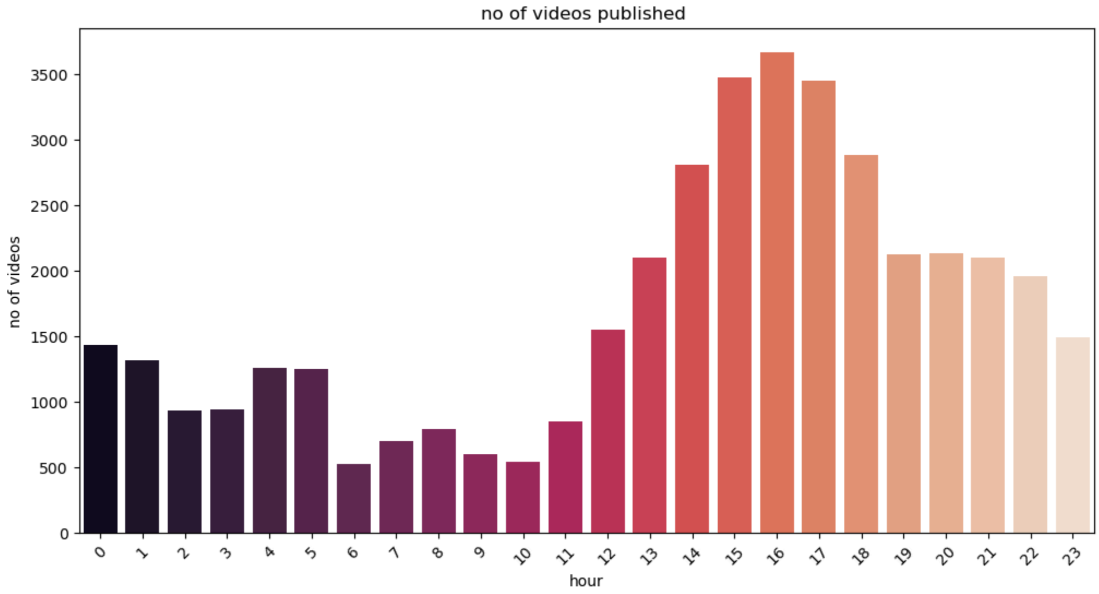
In [33]:

```
plt.figure(figsize=(12,6))
sns.countplot(x='category_name', data=df, order=df['category_name'].value_counts(
plt.xticks(rotation=90)
plt.title('video count by category')
plt.show()
```



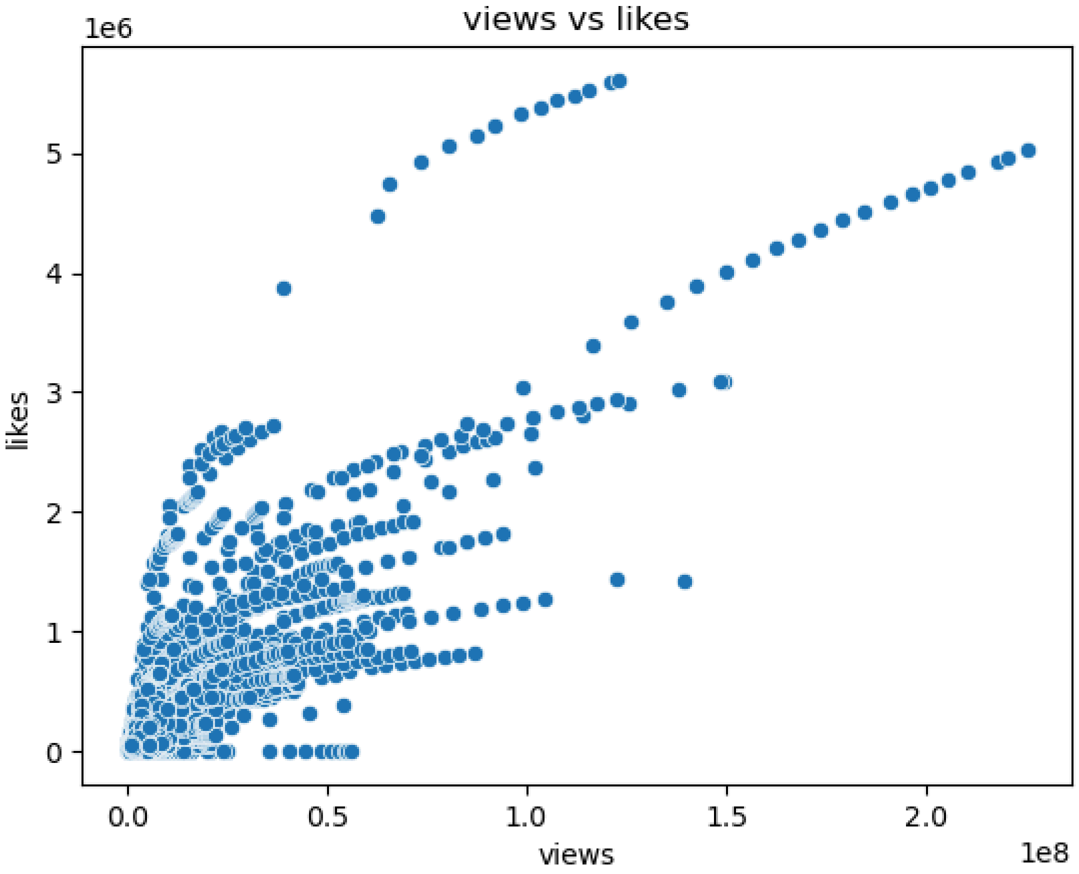
In [35]:

```
videos_per_hour = df['publish_hour'].value_counts().sort_index()
plt.figure(figsize=(12,6))
sns.barplot(x=videos_per_hour.index, y=videos_per_hour.values, palette ='rocket')
plt.title('no of videos published ')
plt.xlabel('hour')
plt.ylabel('no of videos')
plt.xticks(rotation=45)
plt.show()
```



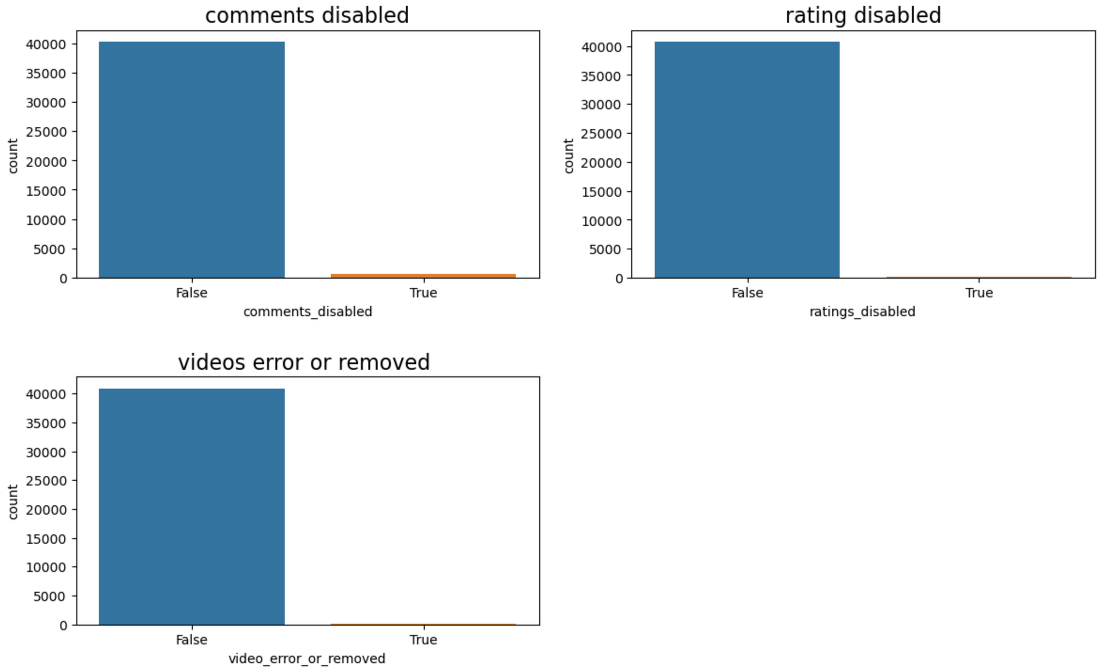
In [36]:

```
sns.scatterplot(data=df, x='views', y='likes')
plt.title('views vs likes')
plt.xlabel('views')
plt.ylabel('likes')
plt.show()
```



In [40]:

```
plt.figure(figsize = (14,8))
plt.subplots_adjust(wspace =0.2, hspace =0.4, top =0.9)
plt.subplot(2,2,1)
g = sns.countplot(x='comments_disabled', data = df)
g.set_title("comments disabled", fontsize=16)
plt.subplot(2,2,2)
g1 = sns.countplot(x= 'ratings_disabled', data=df)
g1.set_title("rating disabled", fontsize =16)
plt.subplot(2,2,3)
g2 = sns.countplot(x='video_error_or_removed', data=df)
g2.set_title("videos error or removed ", fontsize =16)
plt.show()
```



In [42]:

```
corr_matrix = df['views'].corr(df['likes'])
corr_matrix
```

Out[42]:

```
0.8491785476230509
```

In [ ]:

```
```