In [1]:

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

In [2]:

```python
df = pd.read_csv('C:\\Users\91778\\Downloads\\USvideos.csv')
# Display the first few rows of the dataset
print("First few rows of the dataset:")
print(df.head())
```

```
First few rows of the dataset:
      video_id trending_date  \
0  2kyS6SvSYSE      17.14.11   
1  1ZAPwfrtAFY      17.14.11   
2  5qpjK5DgCt4      17.14.11   
3  puqaWrEC7tY      17.14.11   
4  d380meD0W0M      17.14.11   

                                                title          channel_
title  \
0                  WE WANT TO TALK ABOUT OUR MARRIAGE            CaseyNe
istat   
1  The Trump Presidency: Last Week Tonight with J...         LastWeekTo
night   
2  Racist Superman | Rudy Mancuso, King Bach & Le...            Rudy Ma
ncuso   
3                    Nickelback Lyrics: Real or Fake?  Good Mythical Mo
rning   
4                            I Dare You: GOING BALD!?                nig
ahiga   

   category_id              publish_time  \
0           22  2017-11-13T17:13:01.000Z   
1           24  2017-11-13T07:30:00.000Z   
2           23  2017-11-12T19:05:24.000Z   
3           24  2017-11-13T11:00:04.000Z   
4           24  2017-11-12T18:01:41.000Z   

                                                 tags    views   likes  
\
0                                     SHANtell martin   748374   57527   
1  last week tonight trump presidency|"last week ...  2418783   97185   
2  racist superman|"rudy"|"mancuso"|"king"|"bach"...  3191434  146033   
3  rhett and link|"gmm"|"good mythical morning"|"...   343168   10172   
4  ryan|"higa"|"higatv"|"nigahiga"|"i dare you"|"...  2095731  132235   

   dislikes  comment_count                                  thumbnail_
link  \
0      2966          15954  https://i.ytimg.com/vi/2kyS6SvSYSE/defaul
t.jpg (https://i.ytimg.com/vi/2kyS6SvSYSE/default.jpg)   
1      6146          12703  https://i.ytimg.com/vi/1ZAPwfrtAFY/defaul
t.jpg (https://i.ytimg.com/vi/1ZAPwfrtAFY/default.jpg)   
2      5339           8181  https://i.ytimg.com/vi/5qpjK5DgCt4/defaul
t.jpg (https://i.ytimg.com/vi/5qpjK5DgCt4/default.jpg)   
3       666           2146  https://i.ytimg.com/vi/puqaWrEC7tY/defaul
t.jpg (https://i.ytimg.com/vi/puqaWrEC7tY/default.jpg)   
4      1989          17518  https://i.ytimg.com/vi/d380meD0W0M/defaul
t.jpg (https://i.ytimg.com/vi/d380meD0W0M/default.jpg)   

   comments_disabled  ratings_disabled  video_error_or_removed  \
0              False             False                   False   
1              False             False                   False   
2              False             False                   False   
3              False             False                   False   
4              False             False                   False   

                                         description  
0  SHANTELL'S CHANNEL - https://www.youtube.com/s... (https://www.yout
ube.com/s...)  
1  One year after the presidential election, John...  
2  WATCH MY PREVIOUS VIDEO ▶ \n\nSUBSCRIBE ► http...  
```

```
3   Today we find out if Link is a Nickelback amat...
4   I know it's been a while since we did this sho...
```

In [4]:

```
df.shape
```

Out[4]:

```
(40949, 16)
```

In [5]:

```
df=df.drop_duplicates()
df.shape
```

Out[5]:

```
(40901, 16)
```

In [7]:

```
# Summary statistics
print("\nSummary statistics:")
df.describe()
```

```
Summary statistics:
```

Out[7]:

	category_id	views	likes	dislikes	comment_count
count	40901.000000	4.090100e+04	4.090100e+04	4.090100e+04	4.090100e+04
mean	19.970588	2.360678e+06	7.427173e+04	3.711722e+03	8.448567e+03
std	7.569362	7.397719e+06	2.289999e+05	2.904624e+04	3.745139e+04
min	1.000000	5.490000e+02	0.000000e+00	0.000000e+00	0.000000e+00
25%	17.000000	2.419720e+05	5.416000e+03	2.020000e+02	6.130000e+02
50%	24.000000	6.810640e+05	1.806900e+04	6.300000e+02	1.855000e+03
75%	25.000000	1.821926e+06	5.533800e+04	1.936000e+03	5.752000e+03
max	43.000000	2.252119e+08	5.613827e+06	1.674420e+06	1.361580e+06

In [8]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 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 [9]:

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

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 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 [10]:

```
from datetime import datetime
import datetime
df["trending_date"]=df["trending_date"].apply(lambda x: datetime.dateti
print(df.head())
```

```
      video_id trending_date  \
0  2kyS6SvSYSE    2017-11-14
1  1ZAPwfrtAFY    2017-11-14
2  5qpjK5DgCt4    2017-11-14
3  puqaWrEC7tY    2017-11-14
4  d380meD0W0M    2017-11-14

                                               title          channel_
title  \
0                 WE WANT TO TALK ABOUT OUR MARRIAGE            CaseyNe
istat
1  The Trump Presidency: Last Week Tonight with J...         LastWeekTo
night
2  Racist Superman | Rudy Mancuso, King Bach & Le...            Rudy Ma
ncuso
3                   Nickelback Lyrics: Real or Fake?  Good Mythical Mo
rning
4                           I Dare You: GOING BALD!?                nig
ahiga

   category_id              publish_time  \
0           22  2017-11-13T17:13:01.000Z
1           24  2017-11-13T07:30:00.000Z
2           23  2017-11-12T19:05:24.000Z
3           24  2017-11-13T11:00:04.000Z
4           24  2017-11-12T18:01:41.000Z

                                                tags    views   likes
\
0                                    SHANtell martin   748374   57527
1  last week tonight trump presidency|"last week ...  2418783   97185
2  racist superman|"rudy"|"mancuso"|"king"|"bach"...  3191434  146033
3  rhett and link|"gmm"|"good mythical morning"|"...   343168   10172
4  ryan|"higa"|"higatv"|"nigahiga"|"i dare you"|"...  2095731  132235

   dislikes  comment_count  comments_disabled  ratings_disabled  \
0      2966          15954              False             False
1      6146          12703              False             False
2      5339           8181              False             False
3       666           2146              False             False
4      1989          17518              False             False

   video_error_or_removed
0                   False
1                   False
2                   False
3                   False
4                   False
```

In [11]:

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

```
      video_id trending_date  \
0  2kyS6SvSYSE    2017-11-14   
1  1ZAPwfrtAFY    2017-11-14   
2  5qpjK5DgCt4    2017-11-14   
3  puqaWrEC7tY    2017-11-14   
4  d380meD0W0M    2017-11-14   

                                               title          channel_
title  \
0                 WE WANT TO TALK ABOUT OUR MARRIAGE           CaseyNe
istat   
1  The Trump Presidency: Last Week Tonight with J...        LastWeekTo
night   
2  Racist Superman | Rudy Mancuso, King Bach & Le...           Rudy Ma
ncuso   
3                   Nickelback Lyrics: Real or Fake?  Good Mythical Mo
rning   
4                           I Dare You: GOING BALD!?               nig
ahiga   

   category_id             publish_time  \
0           22 2017-11-13 17:13:01+00:00   
1           24 2017-11-13 07:30:00+00:00   
2           23 2017-11-12 19:05:24+00:00   
3           24 2017-11-13 11:00:04+00:00   
4           24 2017-11-12 18:01:41+00:00   

                                                tags    views   likes
\
0                                    SHANtell martin   748374   57527   
1  last week tonight trump presidency|"last week ...  2418783   97185   
2  racist superman|"rudy"|"mancuso"|"king"|"bach"...  3191434  146033   
3  rhett and link|"gmm"|"good mythical morning"|"...   343168   10172   
4  ryan|"higa"|"higatv"|"nigahiga"|"i dare you"|"...  2095731  132235   

   dislikes  comment_count  comments_disabled  ratings_disabled  \
0      2966          15954              False             False   
1      6146          12703              False             False   
2      5339           8181              False             False   
3       666           2146              False             False   
4      1989          17518              False             False   

   video_error_or_removed  
0                   False  
1                   False  
2                   False  
3                   False  
4                   False  
```

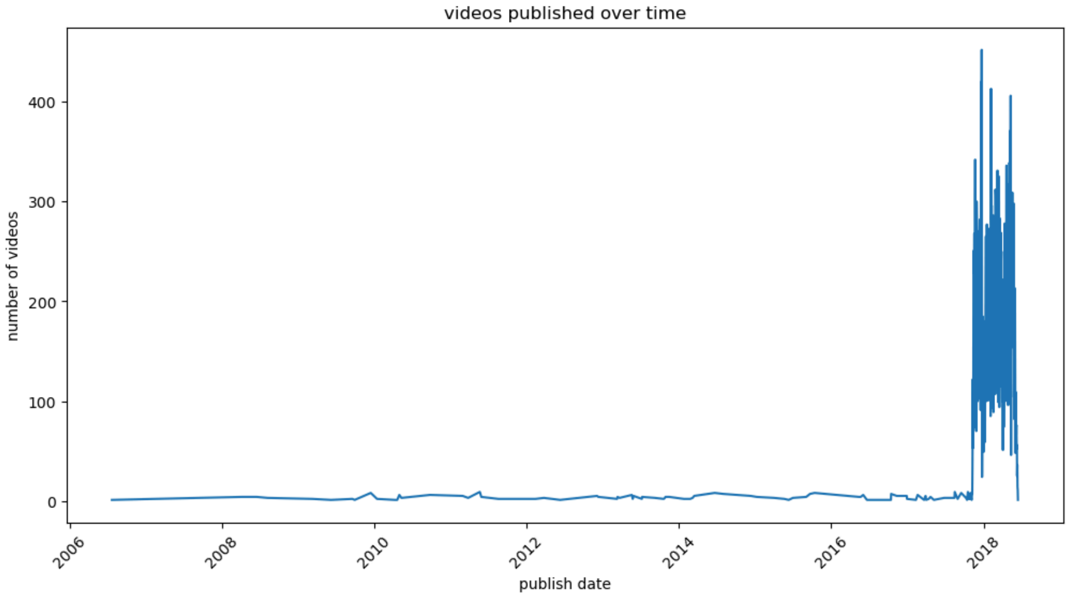
In [12]:

```python
df['publish_month']=df['publish_time'].dt.month
df['publish_day']=df['publish_time'].dt.day
df['publish_hour']=df['publish_time'].dt.hour
print(df.head(4))
```

```
      video_id trending_date  \
0  2kyS6SvSYSE    2017-11-14   
1  1ZAPwfrtAFY    2017-11-14   
2  5qpjK5DgCt4    2017-11-14   
3  puqaWrEC7tY    2017-11-14   

                                               title          channel_
title  \
0                 WE WANT TO TALK ABOUT OUR MARRIAGE           CaseyNe
istat   
1  The Trump Presidency: Last Week Tonight with J...        LastWeekTo
night   
2  Racist Superman | Rudy Mancuso, King Bach & Le...           Rudy Ma
ncuso   
3                   Nickelback Lyrics: Real or Fake?  Good Mythical Mo
rning   

   category_id              publish_time  \
0           22 2017-11-13 17:13:01+00:00   
1           24 2017-11-13 07:30:00+00:00   
2           23 2017-11-12 19:05:24+00:00   
3           24 2017-11-13 11:00:04+00:00   

                                                tags    views   likes  
\
0                                    SHANtell martin   748374   57527   
1  last week tonight trump presidency|"last week ...  2418783   97185   
2  racist superman|"rudy"|"mancuso"|"king"|"bach"...  3191434  146033   
3  rhett and link|"gmm"|"good mythical morning"|"...   343168   10172   

   dislikes  comment_count  comments_disabled  ratings_disabled  \
0      2966          15954              False             False   
1      6146          12703              False             False   
2      5339           8181              False             False   
3       666           2146              False             False   

   video_error_or_removed  publish_month  publish_day  publish_hour  
0                   False             11           13            17  
1                   False             11           13             7  
2                   False             11           12            19  
3                   False             11           13            11  
```

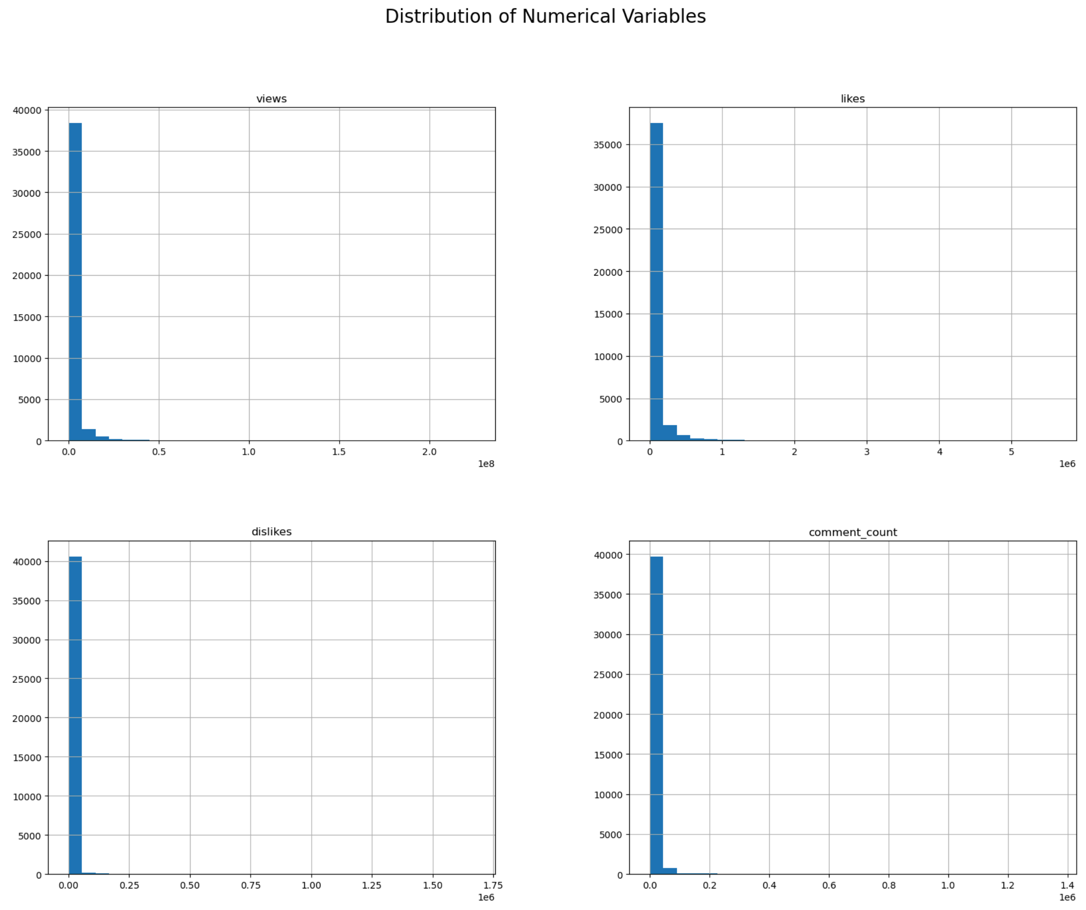
In [13]:

```
df['publish_time'] = pd.to_datetime(df['publish_time'])
df['publish_date'] = df['publish_time'].dt.date
video_count_by_date = df.groupby('publish_date').size()
plt.figure(figsize = (12,6))
sns.lineplot(data=video_count_by_date)
plt.title("videos published over time")
plt.xlabel("publish date")
plt.ylabel("number of videos")
plt.xticks (rotation = 45)
plt.show()
```



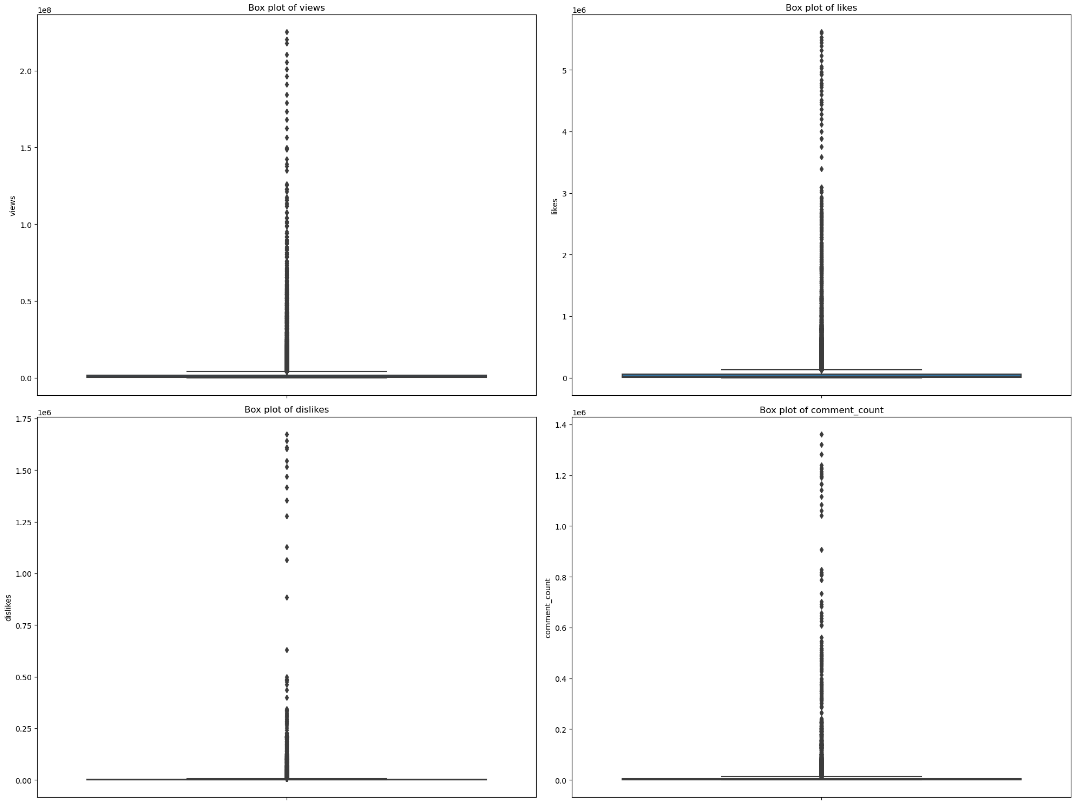
In [14]:

```
# Distribution of numerical variables: Histograms
numerical_columns = ['views', 'likes', 'dislikes', 'comment_count']
df[numerical_columns].hist(bins=30, figsize=(20, 15))
plt.suptitle('Distribution of Numerical Variables', fontsize=20)
plt.show()
```



In [15]:

```
# Identify outliers: Box plots
plt.figure(figsize=(20, 15))
for i, column in enumerate(numerical_columns):
    plt.subplot(2, 2, i + 1)
    sns.boxplot(data=df, y=column)
    plt.title(f'Box plot of {column}')
plt.tight_layout()
plt.show()
```

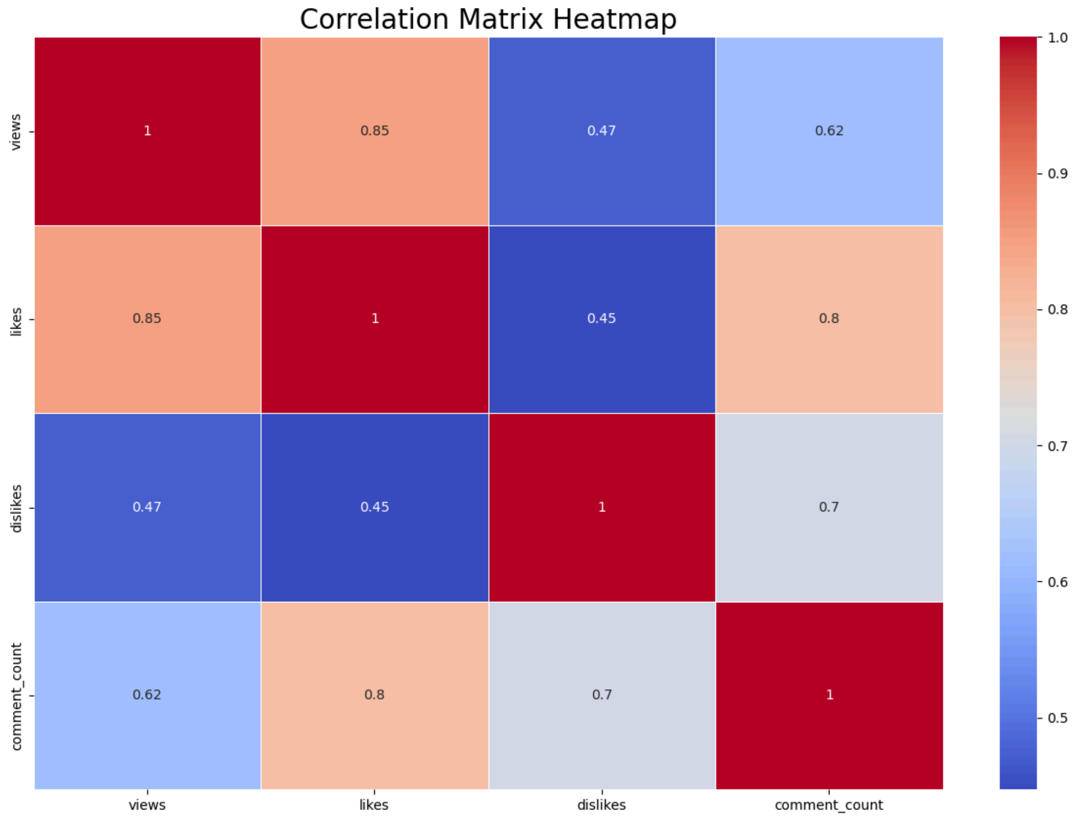


In [16]:

```
# Correlation matrix
correlation_matrix = df[numerical_columns].corr()
```

In [17]:

```
# Correlation matrix heatmap
plt.figure(figsize=(15, 10))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', linewidths
plt.title('Correlation Matrix Heatmap', fontsize=20)
plt.show()
```



In [18]:

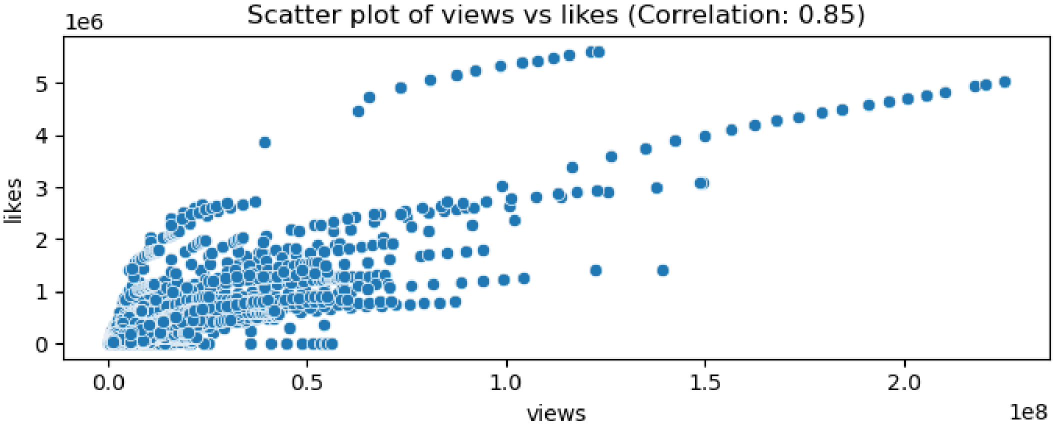
```
# Scatter plots for top correlated pairs
plt.figure(figsize=(20,25))  # Adjust size for combined layout
top_correlations = correlation_matrix.unstack().sort_values(ascending=F
```

```
<Figure size 2000x2500 with 0 Axes>
```

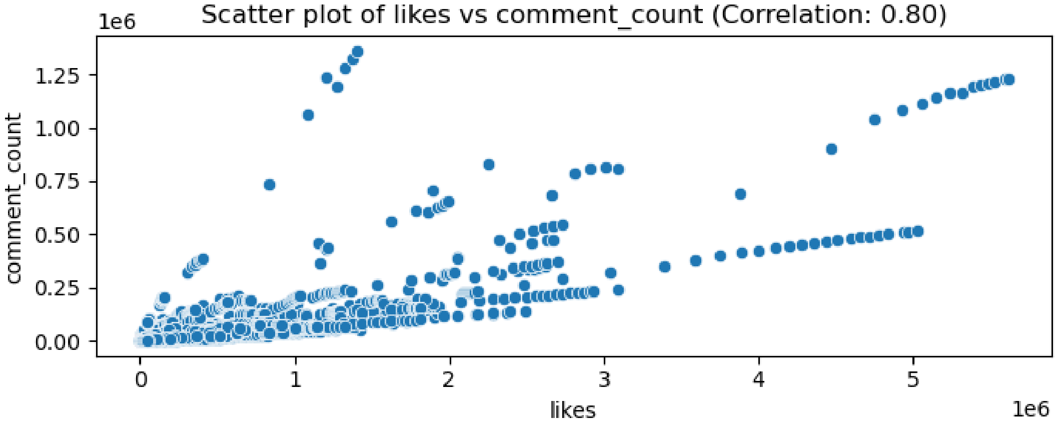
In [19]:

```
 #Select the top 10 correlation pairs (excluding self-correlation)
top_pairs = top_correlations[(top_correlations != 1) & (top_correlation
In [20]:  # Plot each scatter plot separately
for i, (pair, corr) in enumerate(top_pairs.items()):
    col1, col2 = pair
    plt.figure(figsize=(7, 3))
    sns.scatterplot(data=df, x=col1, y=col2)
    print(plt.title(f'Scatter plot of {col1} vs {col2} (Correlation: {c
    print(plt.xlabel(col1))
    print(plt.ylabel(col2))
    print(plt.tight_layout())
    print(plt.show())
```

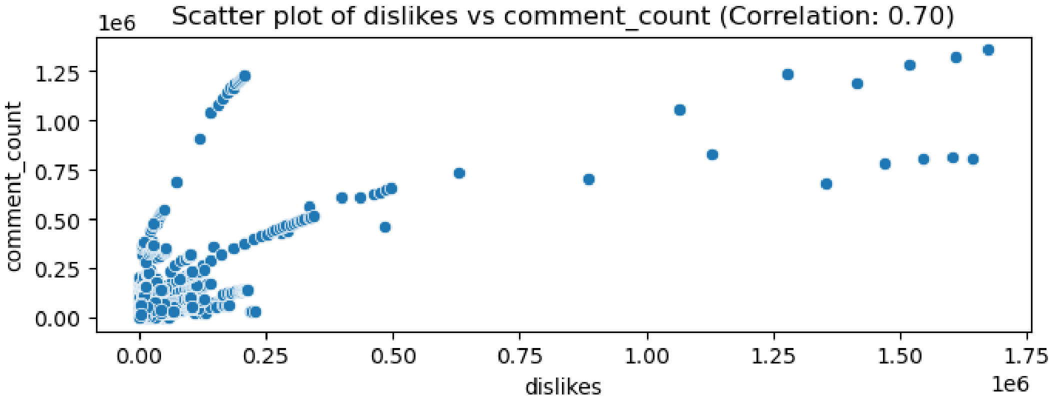
```
Text(0.5, 1.0, 'Scatter plot of views vs likes (Correlation: 0.85)')
Text(0.5, 0, 'views')
Text(0, 0.5, 'likes')
None
```



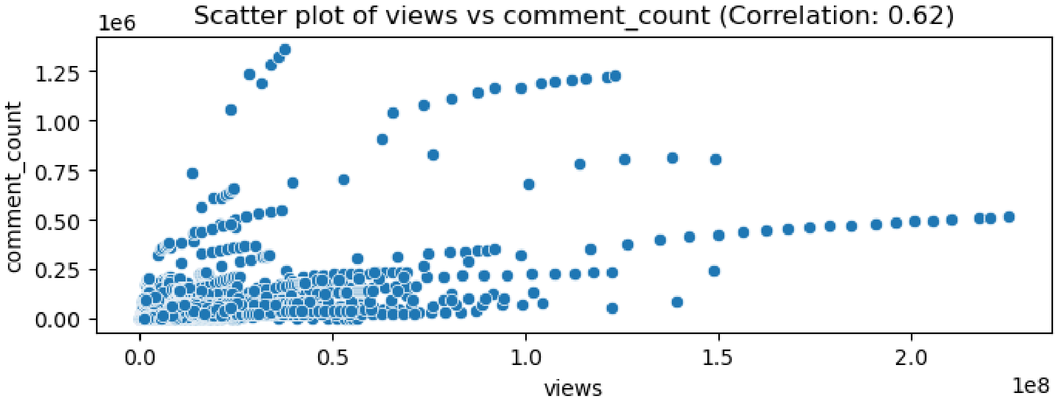
```
None
Text(0.5, 1.0, 'Scatter plot of likes vs comment_count (Correlation:
0.80)')
Text(0.5, 0, 'likes')
Text(0, 0.5, 'comment_count')
None
```



```
None
Text(0.5, 1.0, 'Scatter plot of dislikes vs comment_count (Correlatio
n: 0.70)')
Text(0.5, 0, 'dislikes')
Text(0, 0.5, 'comment_count')
None
```



```
None
Text(0.5, 1.0, 'Scatter plot of views vs comment_count (Correlation:
0.62)')
Text(0.5, 0, 'views')
Text(0, 0.5, 'comment_count')
None
```



```
None
```