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


from google.colab import drive
drive.mount('/content/drive')

# Access and read the CSV
import pandas as pd

data_path = '/content/drive/My Drive/USvideos.csv'  # Replace with your file path
df = pd.read_csv(data_path)
```

```
Mounted at /content/drive
```

```
df.head()
```

	video_id	trending_date	title	channel_title	category_id	publish_t
0	2kyS6SvSYSE	17.14.11	WE WANT TO TALK ABOUT OUR MARRIAGE	CaseyNeistat	22	2017 13T17:13:01.0
1	1ZAPwfrtAFY	17.14.11	The Trump Presidency: Last Week Tonight with J...	LastWeekTonight	24	2017 13T07:30:00.0
2	5qpjK5DgCt4	17.14.11	Racist Superman | Rudy Mancuso, King Bach & Le...	Rudy Mancuso	23	2017 12T19:05:24.0
3	puqaWrEC7tY	17.14.11	Nickelback Lyrics: Real or Fake?	Good Mythical Morning	24	2017 13T11:00:04.0
4	d380meD0W0M	17.14.11	I Dare You: GOING BALD!?	nigahiga	24	2017 12T18:01:41.0

```
df.shape
```

```
(40949, 16)
```

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

```
(40901, 16)
```

```
df.describe()
```

	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

```
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
```

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

```
<class 'pandas.core.frame.DataFrame'>
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
```

```
from datetime import datetime
```

```
import datetime
```

```
df["trending_date"] = df["trending_date"].apply(lambda x: datetime.datetime.strptime(x, "%y.%d.%m"))
df.head(3)
```

	video_id	trending_date	title	channel_title	category_id	publish_tim
0	2kyS6SvSYSE	2017-11-14	WE WANT TO TALK ABOUT OUR MARRIAGE	CaseyNeistat	22	2017-1 13T17:13:01.000
1	1ZAPwfrtAFY	2017-11-14	The Trump Presidency: Last Week Tonight with J...	LastWeekTonight	24	2017-1 13T07:30:00.000
2	5qpjK5DgCt4	2017-11-14	Racist Superman | Rudy Mancuso, King Bach & Le...	Rudy Mancuso	23	2017-1 12T19:05:24.000

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

	video_id	trending_date	title	channel_title	category_id	publish_time
0	2kyS6SvSYSE	2017-11-14	WE WANT TO TALK ABOUT OUR MARRIAGE	CaseyNeistat	22	2017-11-13 17:13:01+00:00
1	1ZAPwfrtAFY	2017-11-14	The Trump Presidency: Last Week Tonight with J...	LastWeekTonight	24	2017-11-13 07:30:00+00:00

```
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)
```

	video_id	trending_date	title	channel_title	category_id	publish_time
0	2kyS6SvSYSE	2017-11-14	WE WANT TO TALK ABOUT OUR MARRIAGE	CaseyNeistat	22	2017-11-13 17:13:01+00:00
1	1ZAPwfrtAFY	2017-11-14	The Trump Presidency: Last Week Tonight with J...	LastWeekTonight	24	2017-11-13 07:30:00+00:00

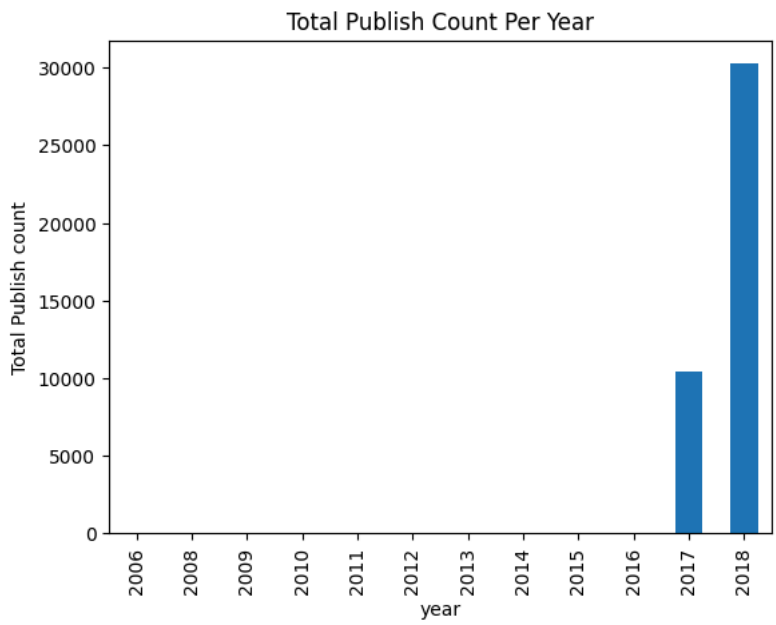
```
print(sorted(df["category_id"].unique()))
[1,2,10,15,17,19,20,222,23,24,25,26,27,28,29,30,43]
```

```
[1, 2, 10, 15, 17, 19, 20, 22, 23, 24, 25, 26, 27, 28, 29, 43]
[1, 2, 10, 15, 17, 19, 20, 222, 23, 24, 25, 26, 27, 28, 29, 30, 43]
```

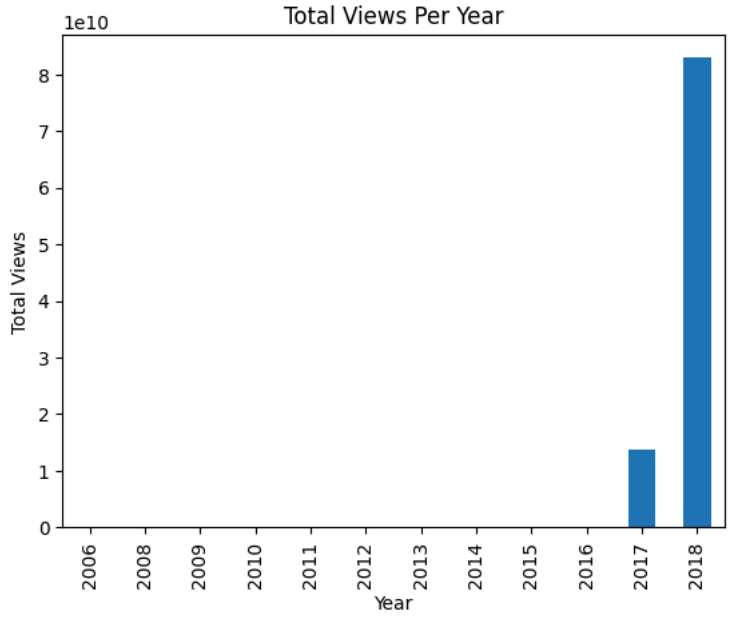
```
df['category_name'] = np.nan
df.loc[(df["category_id"]==1),"category_name"] = "Film & Animation"
df.loc[(df["category_id"]==2),"category_name"] = "Autos & Vehicles"
df.loc[(df["category_id"]==10),"category_name"] = "Music"
df.loc[(df["category_id"]==15),"category_name"] = "Pets & Animals"
df.loc[(df["category_id"]==17),"category_name"] = "Sports"
df.loc[(df["category_id"]==19),"category_name"] = "Travel & Events"
df.loc[(df["category_id"]==20),"category_name"] = "Gaming"
df.loc[(df["category_id"]==22),"category_name"] = "People & 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 & Politics"
df.loc[(df["category_id"]==26),"category_name"] = "Howto & Style"
df.loc[(df["category_id"]==27),"category_name"] = "Education"
df.loc[(df["category_id"]==28),"category_name"] = "Science & Technology"
df.loc[(df["category_id"]==29),"category_name"] = "Nonprofits & Activism"
df.loc[(df["category_id"]==30),"category_name"] = "Movies"
df.loc[(df["category_id"]==43),"category_name"] = "Shows"
df.head()
```

	video_id	trending_date	title	channel_title	category_id	publish_tim
0	2kyS6SvSYSE	2017-11-14	WE WANT TO TALK ABOUT OUR MARRIAGE	CaseyNeistat	22	2017-11-1 17:13:01+00:0
1	1ZAPwfrtAFY	2017-11-14	The Trump Presidency: Last Week Tonight with J...	LastWeekTonight	24	2017-11-1 07:30:00+00:0
2	5qpjK5DgCt4	2017-11-14	Racist Superman | Rudy Mancuso, King Bach & Le...	Rudy Mancuso	23	2017-11-1 19:05:24+00:0
3	puqaWrEC7tY	2017-11-14	Nickelback Lyrics: Real or Fake?	Good Mythical Morning	24	2017-11-1 11:00:04+00:0
4	d380meD0W0M	2017-11-14	I Dare You: GOING BALD!?	nigahiga	24	2017-11-1 18:01:41+00:0

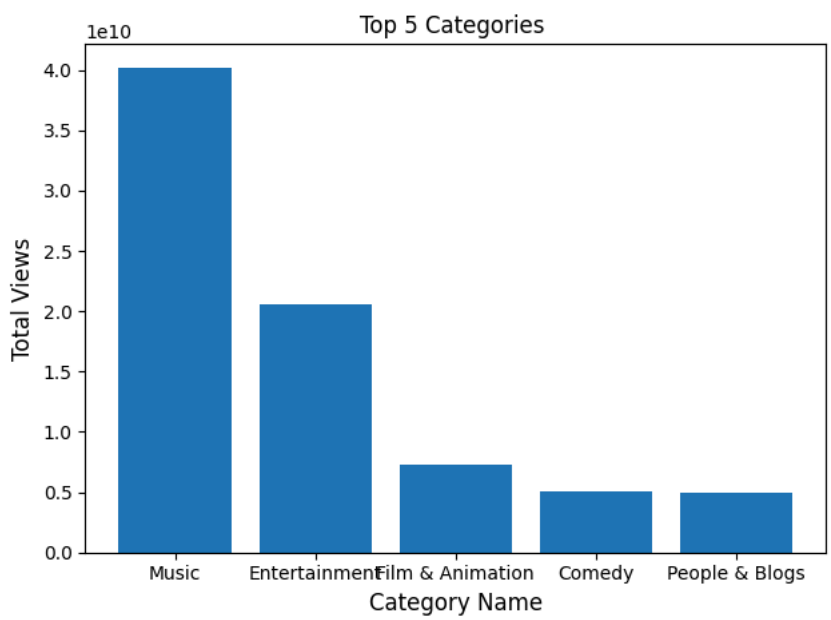
```
df['year']=df['publish_time'].dt.year
yearly_counts = df.groupby('year')['video_id'].count()

#create a bar charrt
yearly_counts.plot(kind='bar',xlabel='year',ylabel='Total Publish count',title='Total Publish Count Per Year')

#Show the chart
plt.show()
```




```
#Group by the year and sum the views for each year
yearly_views = df.groupby('year')['views'].sum()


#Create a bar chart
yearly_views.plot(kind='bar',xlabel='Year',ylabel='Total Views',title='Total Views Per Year')

#Show the chart
plt.show()
```

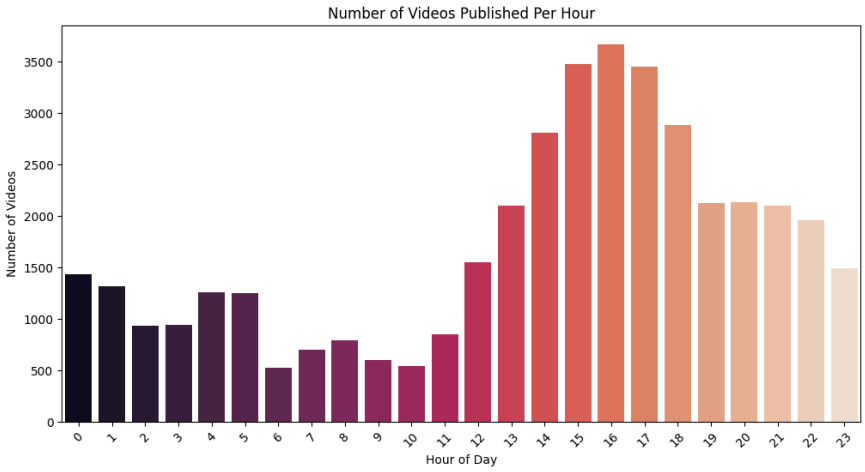



```
#Group the data by 'category_name' and calculate the sum of the 'views' in each category
category_views = df.groupby('category_name')['views'].sum().reset_index()

#sort the categories by views in decending order
top_categories = category_views.sort_values(by='views',ascending=False).head(5)

#create a bar plot to vishualize the top 5 categories
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=12)
plt.tight_layout()
plt.show()
```

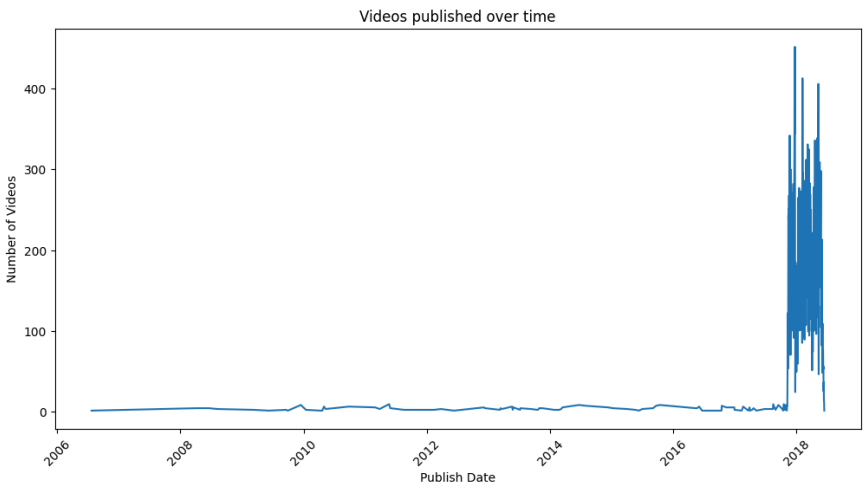



```
#Count the number of videos published per hour
videos_per_hour = df['publish_hour'].value_counts().sort_index()

#Create a bar plot
plt.figure(figsize=(12,6))
sns.barplot(x=videos_per_hour.index,y=videos_per_hour.values,palette='rocket')
plt.title('Number of Videos Published Per Hour')
plt.xlabel('Hour of Day')
plt.ylabel('Number of Videos')
plt.xticks(rotation=45)
plt.show()
```

```
<ipython-input-22-36513ce13ed0>:6: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.

  sns.barplot(x=videos_per_hour.index,y=videos_per_hour.values,palette='rocket')
```

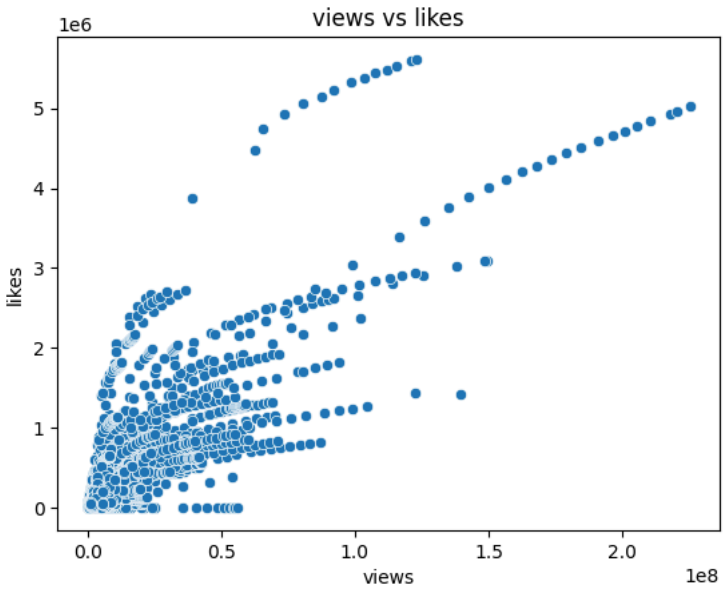



```
df['publish_time'] = pd.to_datetime(df['publish_time'])
df['publish_date'] = df['publish_time'].dt.date
video_count_by_date =df.groupby('publish_date').size()
plt.figure(figsize=(12,6))
sns.lineplot(data=video_count_by_date)
plt.title('Videos published over time')
plt.xlabel('Publish Date')
plt.ylabel('Number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




```
#Scatter plot between 'views' and 'likes'
sns.scatterplot(data=df,x='views',y='likes')
plt.title('views vs likes')
plt.xlabel('views')
```

```
plt.ylabel('likes')
plt.show()
```




```
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)
g=sns.countplot(x='ratings_disabled',data=df)
g.set_title("ratings_disabled",fontsize=16)
plt.subplot(2,2,3)
g=sns.countplot(x='video_error_or_removed',data=df)
g.set_title("video_error_or_removed",fontsize=16)
plt.show()
```