In [ ]:
```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import datetime
```

In [ ]:
```python
data = pd.read_csv("C:\\Users\hp\Downloads\\USvideos.csv")
data.head(3)
```

```
<>:1: SyntaxWarning: invalid escape sequence '\h'
<>:1: SyntaxWarning: invalid escape sequence '\h'
C:\Users\hp\AppData\Local\Temp\ipykernel_27160\3906100858.py:1: SyntaxWarning: invalid escape sequence '\h'
  data = pd.read_csv("C:\\Users\hp\Downloads\\USvideos.csv")
```

Out[ ]:

	video_id	trending_date	title	channel_title	category_id	publish_time	tags	views	likes	dislikes	comment_count	thumbnail_link	comments_disabled	ratings_disabled	video_error_or_removed
0	2kyS6SvSYSE	17.14.11	WE WANT TO TALK ABOUT OUR MARRIAGE	CaseyNeistat	22	2017-11-13T17:13:01.000Z	SHANtell martin	748374	57527	2966	15954	https://i.ytimg.com/vi/2kyS6SvSYSE/default.jpg	False	False	False
1	1ZAPwfrtAFY	17.14.11	The Trump Presidency: Last Week Tonight with J...	LastWeekTonight	24	2017-11-13T07:30:00.000Z	last week tonight trump presidency|"last week ...	2418783	97185	6146	12703	https://i.ytimg.com/vi/1ZAPwfrtAFY/default.jpg	False	False	False
2	5qpjK5DgCt4	17.14.11	Racist Superman | Rudy Mancuso, King Bach & Le...	Rudy Mancuso	23	2017-11-12T19:05:24.000Z	racist superman|"rudy"|"mancuso"|"king"|"bach"...	3191434	146033	5339	8181	https://i.ytimg.com/vi/5qpjK5DgCt4/default.jpg	False	False	False

In [ ]:
```python
data.shape
```

Out[ ]:
```
(40949, 16)
```

In [ ]:
```python
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 40949 entries, 0 to 40948
Data columns (total 16 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   video_id                40949 non-null  object
 1   trending_date           40949 non-null  object
 2   title                   40949 non-null  object
 3   channel_title           40949 non-null  object
 4   category_id             40949 non-null  int64
 5   publish_time            40949 non-null  object
 6   tags                    40949 non-null  object
 7   views                   40949 non-null  int64
 8   likes                   40949 non-null  int64
 9   dislikes                40949 non-null  int64
 10  comment_count           40949 non-null  int64
 11  thumbnail_link          40949 non-null  object
 12  comments_disabled       40949 non-null  bool
 13  ratings_disabled        40949 non-null  bool
 14  video_error_or_removed  40949 non-null  bool
 15  description             40379 non-null  object
dtypes: bool(3), int64(5), object(8)
memory usage: 4.2+ MB
```

In [ ]:
```python
data = data.drop_duplicates()
data = data.dropna()
data.shape
```

Out[ ]:
```
(40332, 16)
```

In [ ]:
```python
#removing outlires
q1 = data.views.quantile(0.25)
q3 = data.views.quantile(0.75)
iqr = q3 - q1

print(iqr)
```

```
1585507.25
```

In [ ]:
```python
upper_bound = q3 + (1.5*iqr)
lower_bound = q1 - (1.5*iqr)
print(upper_bound,lower_bound)
```

```
4210435.125 -2131593.875
```

In [ ]:
```python
outlires = data[(data.views < lower_bound) | (data.views > upper_bound)]
outlires.head()
```

Out[ ]:

	video_id	trending_date	title	channel_title	category_id	publish_time	tags	views	likes	dislikes	comment_count	thumbnail_link	comments_disabled	ratings_disabled	vide
32	n1WpP7iowLc	17.14.11	Eminem - Walk On Water (Audio) ft. Beyoncé	EminemVEVO	10	2017-11-10T17:00:03.000Z	Eminem|"Walk"|"On"|"Water"|"Aftermath/Shady/In...	17158531	787419	43420	125882	https://i.ytimg.com/vi/n1WpP7iowLc/default.jpg	False	False	
53	9t9u_yPEidY	17.14.11	Jennifer Lopez - Amor, Amor, Amor (Official Vi...	JenniferLopezVEVO	10	2017-11-10T15:00:00.000Z	Jennifer Lopez ft. Wisin|"Jennifer Lopez ft. W...	9548677	190083	15015	11473	https://i.ytimg.com/vi/9t9u_yPEidY/default.jpg	False	False	
69	Jw1Y-zhQURU	17.14.11	John Lewis Christmas Ad 2017 - #MozTheMonster	John Lewis	26	2017-11-10T07:38:29.000Z	christmas|"john lewis christmas"|"john lewis"|...	7224515	55681	10247	9479	https://i.ytimg.com/vi/Jw1Y-zhQURU/default.jpg	False	False	
70	2Vv-BfVoq4g	17.14.11	Ed Sheeran - Perfect (Official Music Video)	Ed Sheeran	10	2017-11-09T11:04:14.000Z	edsheeran|"ed sheeran"|"acoustic"|"live"|"cove...	33523622	1634124	21082	85067	https://i.ytimg.com/vi/2Vv-BfVoq4g/default.jpg	False	False	
104	pz95u3UVpaM	17.14.11	Camila Cabello - Havana (Vertical Video) ft. Y...	CamilaCabelloVEVO	10	2017-11-10T05:01:00.000Z	camila cabello|"camila"|"young thug"|"havana"|...	5476737	286268	4083	12254	https://i.ytimg.com/vi/pz95u3UVpaM/default.jpg	False	False	

In [ ]:
```python
df = data[(data.views > lower_bound) & (data.views < upper_bound)] #data with no outliers
df.shape
```

Out[ ]:
```
(35931, 16)
```

In [ ]:
```python
#deleting extra columns
df = df.drop(columns=['thumbnail_link','description'])
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 35931 entries, 0 to 40946
Data columns (total 14 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   video_id                35931 non-null  object
 1   trending_date           35931 non-null  object
 2   title                   35931 non-null  object
 3   channel_title           35931 non-null  object
 4   category_id             35931 non-null  int64
 5   publish_time            35931 non-null  object
 6   tags                    35931 non-null  object
 7   views                   35931 non-null  int64
 8   likes                   35931 non-null  int64
 9   dislikes                35931 non-null  int64
 10  comment_count           35931 non-null  int64
 11  comments_disabled       35931 non-null  bool
 12  ratings_disabled        35931 non-null  bool
 13  video_error_or_removed  35931 non-null  bool
dtypes: bool(3), int64(5), object(6)
memory usage: 3.4+ MB
```

In [ ]:
```python
#changing column datatype
df['publish_time'] = pd.to_datetime(df['publish_time'], dayfirst= True )
df.head(2)
```

```
C:\Users\hp\AppData\Local\Temp\ipykernel_27160\47464983.py:2: UserWarning: Parsing dates in %Y-%m-%dT%H:%M:%S.%f%z format when dayfirst=True was specified. Pass `dayfirst=False` or specify a format to silence this warning.
  df['publish_time'] = pd.to_datetime(df['publish_time'], dayfirst= True )
```

Out[ ]:

	video_id	trending_date	title	channel_title	category_id	publish_time	tags	views	likes	dislikes	comment_count	comments_disabled	ratings_disabled	video_error_or_removed
0	2kyS6SvSYSE	17.14.11	WE WANT TO TALK ABOUT OUR MARRIAGE	CaseyNeistat	22	2017-11-13 17:13:01+00:00	SHANtell martin	748374	57527	2966	15954	False	False	False
1	1ZAPwfrtAFY	17.14.11	The Trump Presidency: Last Week Tonight with J...	LastWeekTonight	24	2017-11-13 07:30:00+00:00	last week tonight trump presidency|"last week ...	2418783	97185	6146	12703	False	False	False

In [ ]:
```python
df['publish_year'] = df['publish_time'].dt.year
df['publish_month'] = df['publish_time'].dt.month
df['publish_date'] = df['publish_time'].dt.day
df.head(2)
```

Out[ ]:

	video_id	trending_date	title	channel_title	category_id	publish_time	tags	views	likes	dislikes	comment_count	comments_disabled	ratings_disabled	video_error_or_removed	publish_year	publish_month	publish_date
0	2kyS6SvSYSE	17.14.11	WE WANT TO TALK ABOUT OUR MARRIAGE	CaseyNeistat	22	2017-11-13 17:13:01+00:00	SHANtell martin	748374	57527	2966	15954	False	False	False	2017	11	13
1	1ZAPwfrtAFY	17.14.11	The Trump Presidency: Last Week Tonight with J...	LastWeekTonight	24	2017-11-13 07:30:00+00:00	last week tonight trump presidency|"last week ...	2418783	97185	6146	12703	False	False	False	2017	11	13

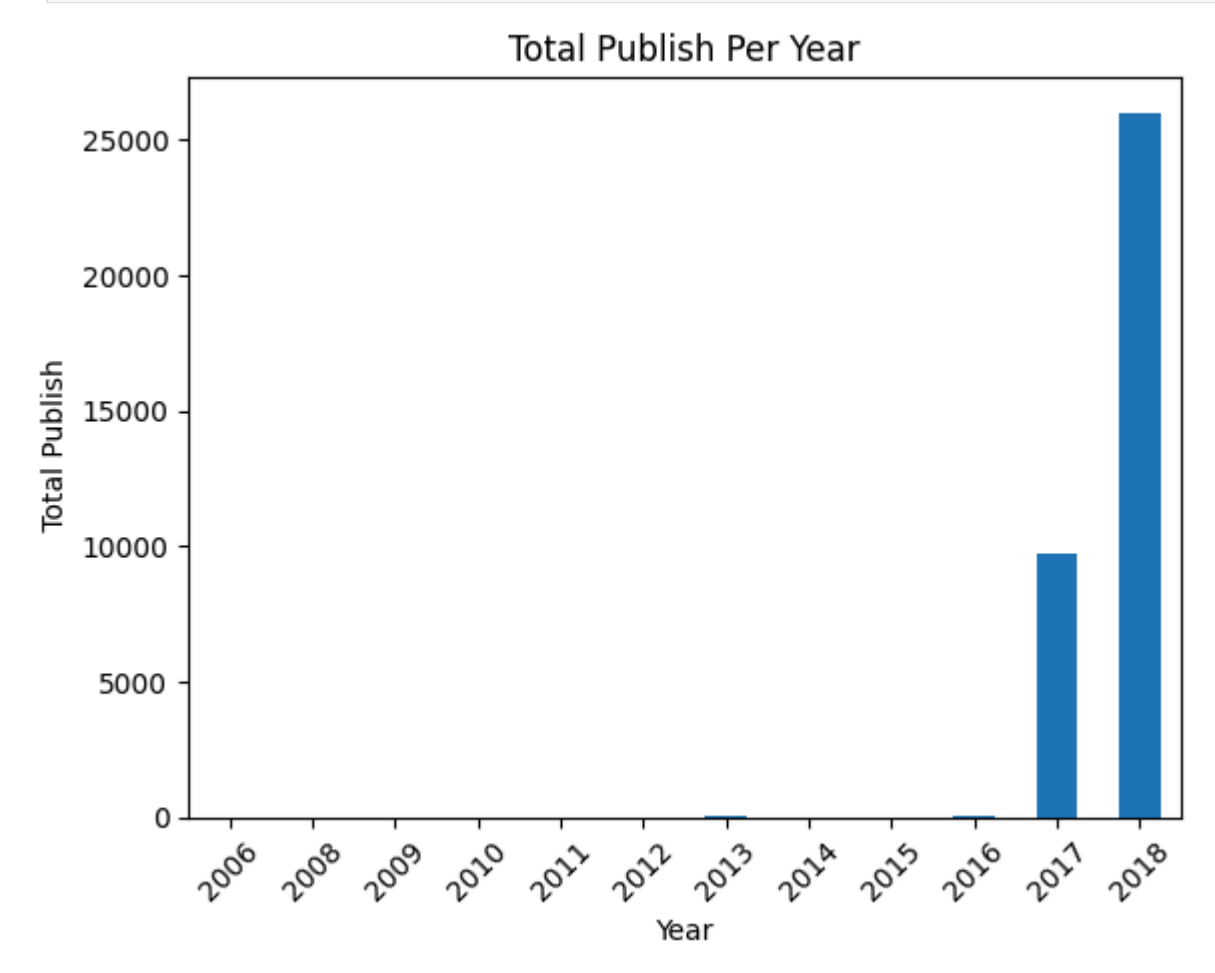
In [ ]:
```python
print(sorted(data['category_id'].unique()))
```

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

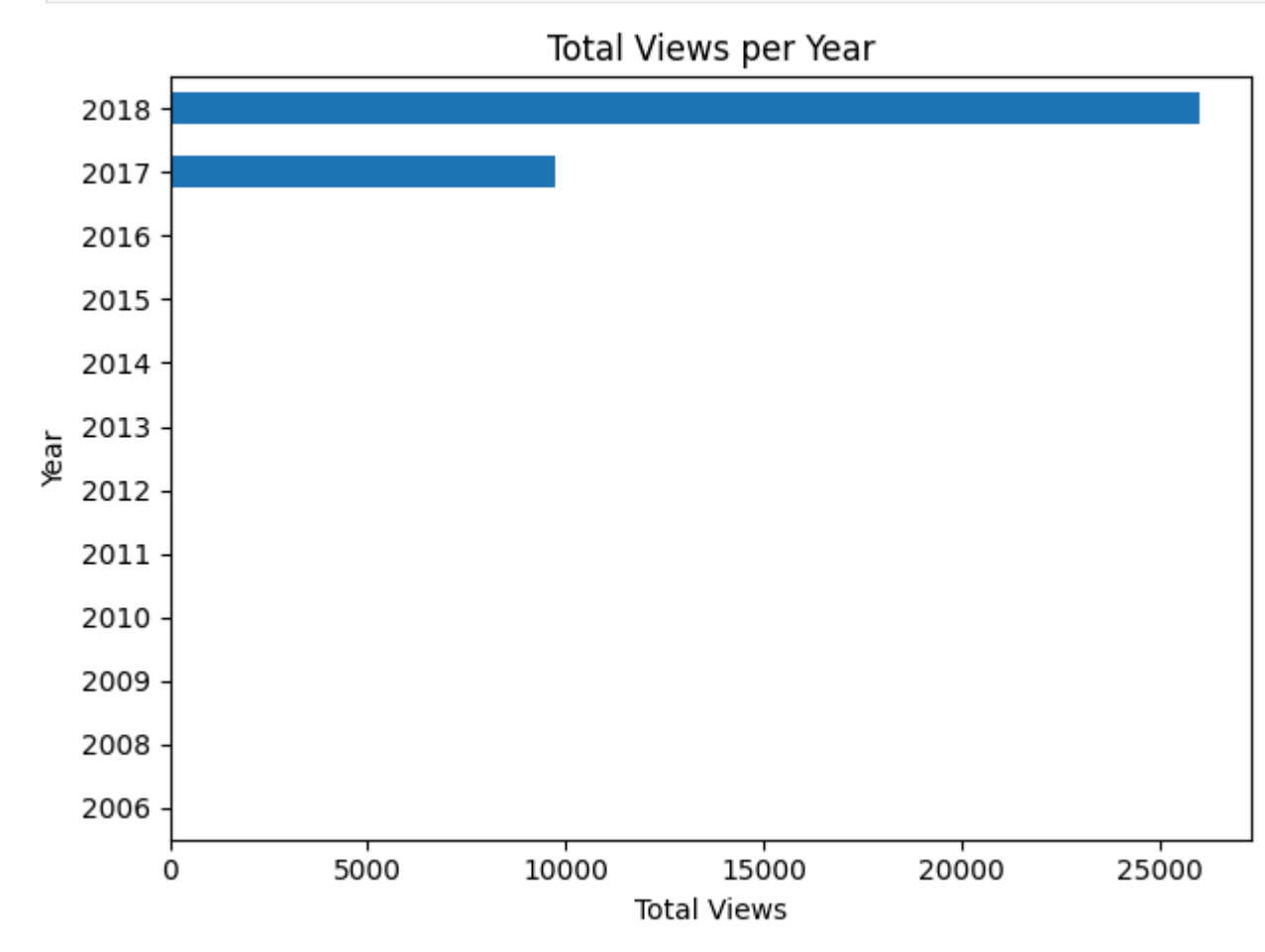
In [ ]:
```python
#adding category_name on the basis of category_id
df['category_name'] = np.nan
df.loc[(df['category_id'] == 1), "category_name"] = 'Film and Animation'
df.loc[(df['category_id'] == 2), "category_name"] = 'Autos and Vehicles'
df.loc[(df['category_id'] == 10), "category_name"] = 'Music'
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 Ploitics'
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 Activism'
df.loc[(df['category_id'] == 43), "category_name"] = 'Movies'
```

```
C:\Users\hp\AppData\Local\Temp\ipykernel_27160\4197363958.py:3: FutureWarning: Setting an item of incompatible dtype is deprecated and will raise an error in a future version of pandas. Value 'Film and Animation' has dtype incompatible with float64, please explicitly cast to a compatible dtype first.
  df.loc[(df['category_id'] == 1), "category_name"] = 'Film and Animation'
```

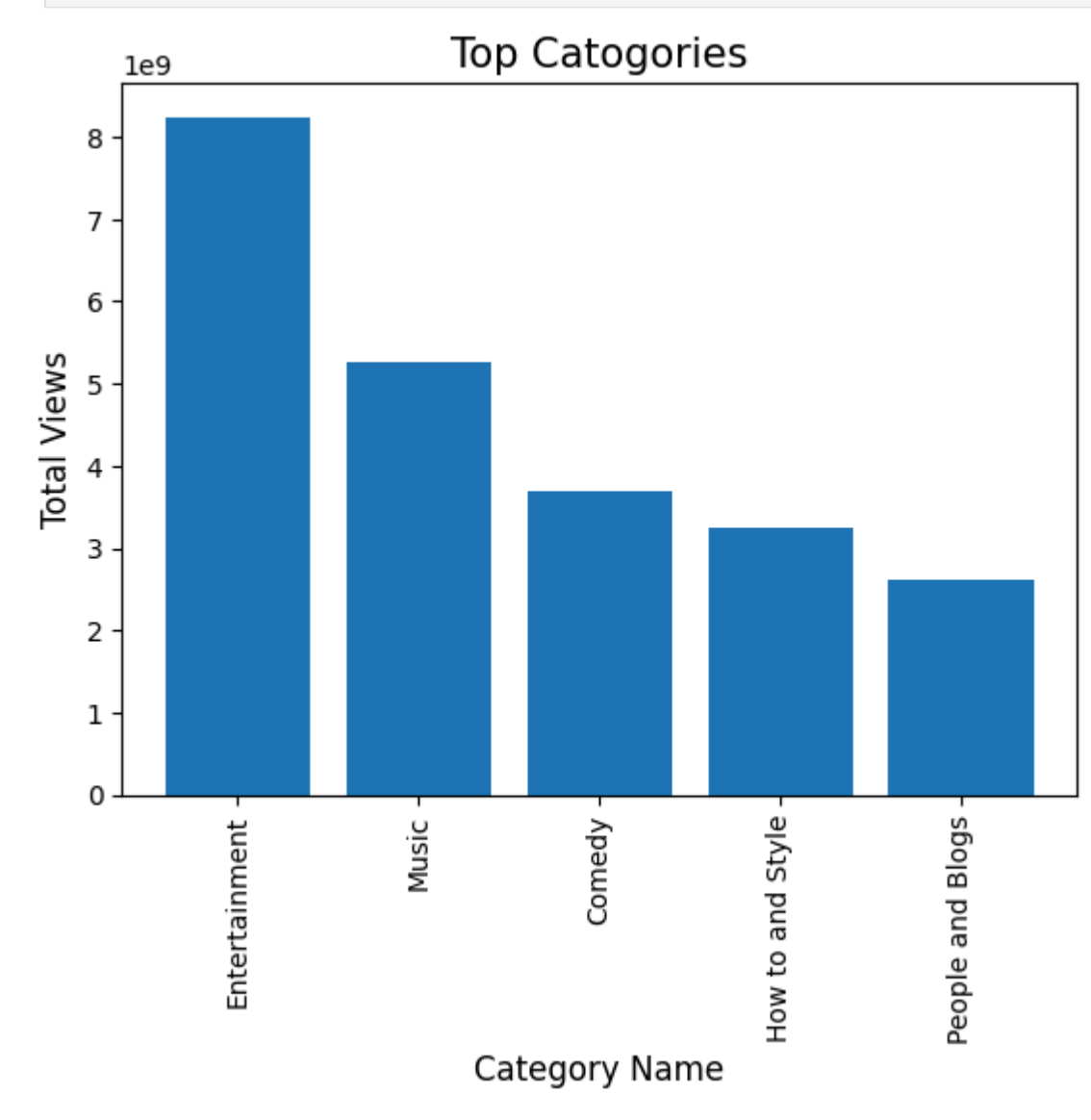
In [ ]:
```python
#creating a bar chart
grouped_years = df.groupby('publish_year')['video_id'].count()

grouped_years.plot(kind= 'bar', xlabel= 'Year', ylabel= ' Total Publish', title= 'Total Publish Per Year')
plt.xticks(rotation=45)
plt.show()
```



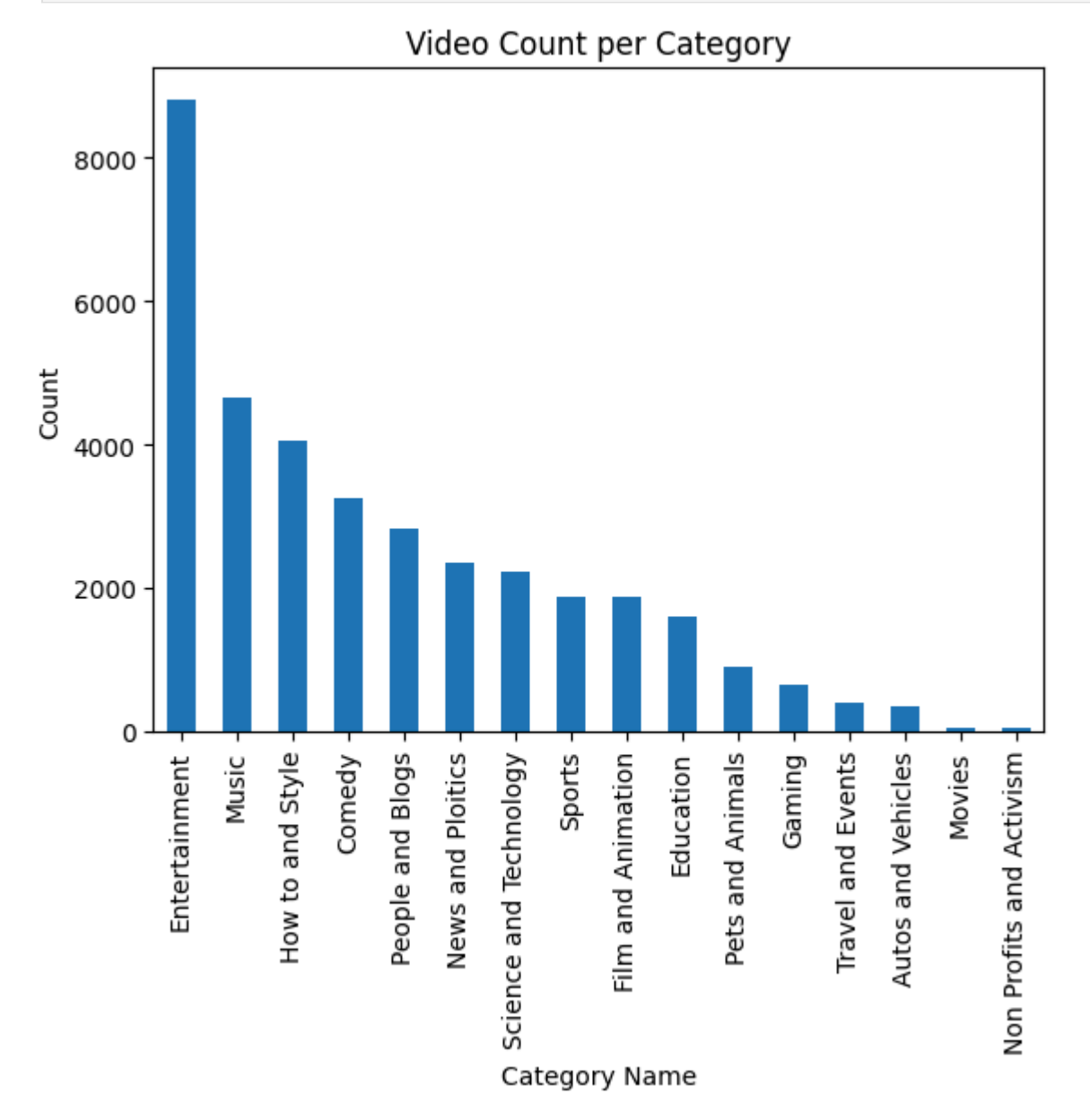
In [ ]:
```python
views_per_year = df.groupby('publish_year')['views'].count()

views_per_year.plot(kind='barh', xlabel= 'Total Views', ylabel='Year', title='Total Views per Year')
plt.xticks(rotation=0)
plt.tight_layout()

plt.show()
```



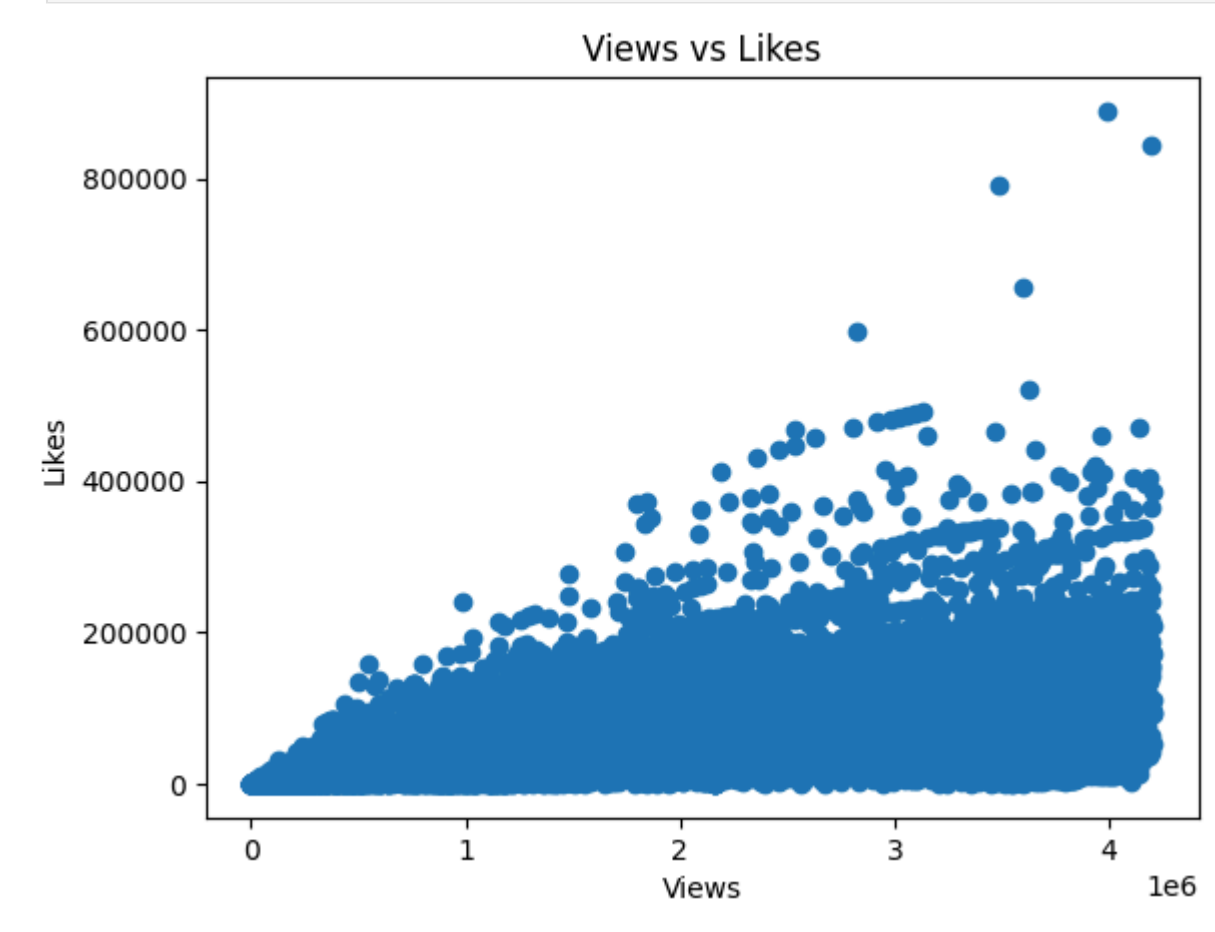
In [ ]:
```python
category_views = df.groupby('category_name')['views'].sum().reset_index()

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

#creating visualisation by using matplotlib
plt.bar(top_categories['category_name'], top_categories['views'])
plt.xlabel('Category Name', fontsize = 12)
plt.xticks(rotation=90)
plt.ylabel('Total Views', fontsize = 12)
plt.title('Top Catogories', fontsize = 15)

plt.show()
```



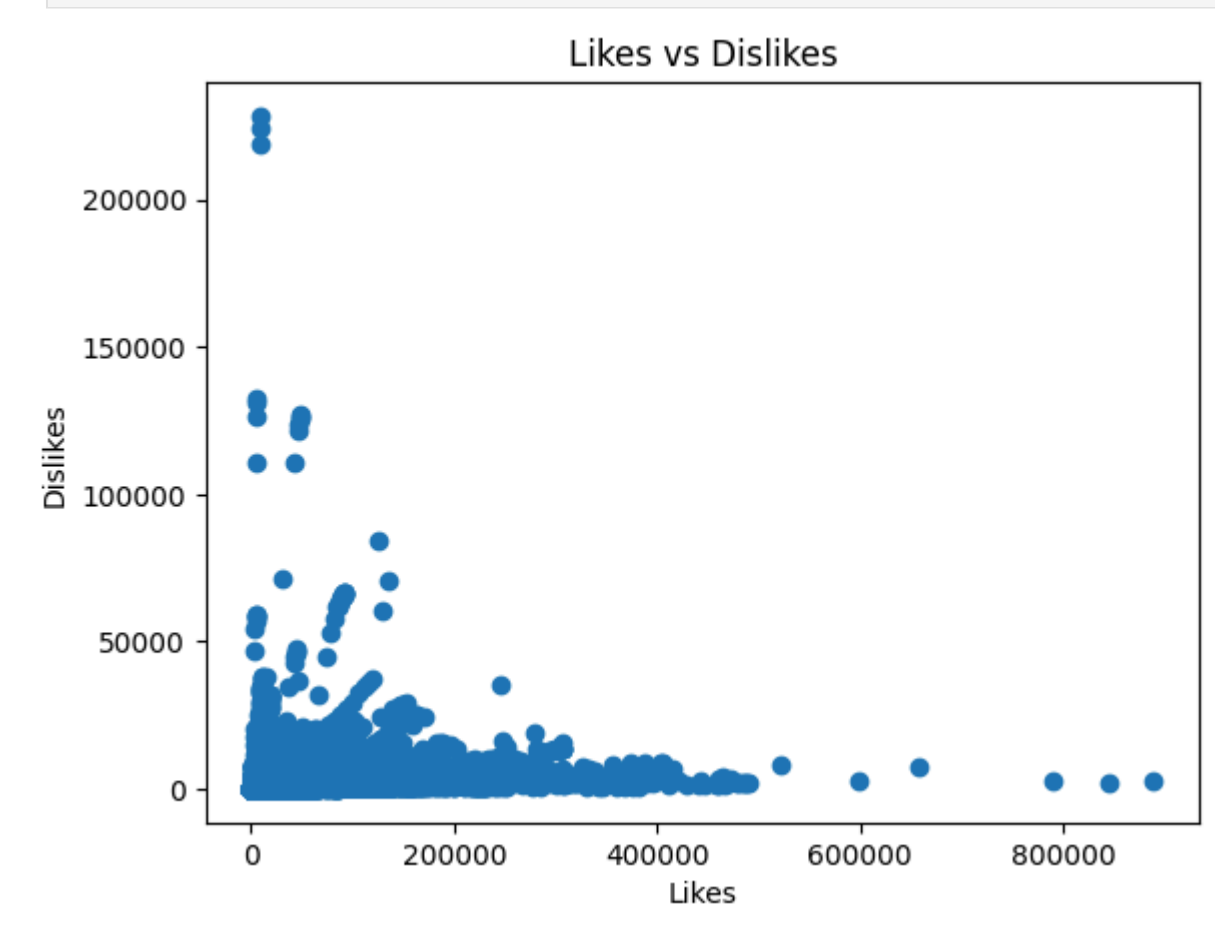
In [ ]:
```python
videos_per_category = df.groupby('category_name')['video_id'].count().sort_values(ascending=False)

videos_per_category.plot(kind='bar', xlabel='Category Name', ylabel='Count', title='Video Count per Category' )

plt.show()
```



In [ ]:
```python
plt.scatter(data=df , x='views', y='likes')
plt.title('Views vs Likes')
plt.xlabel('Views')
plt.ylabel('Likes')
plt.show()
```



In [ ]:
```python
plt.scatter(data=df , x='likes', y='dislikes')
plt.title('Likes vs Dislikes')
plt.xlabel('Likes')
plt.ylabel('Dislikes')
plt.show()
```



In [ ]:
```python
corr_views_likes = df['views'].corr(df['likes']) #co-relation b/w views and likes
corr_views_likes
```

Out[ ]:
```
0.7094102582069112
```

In [ ]:
```python
corr_likes_comments = df['likes'].corr(df['comment_count']) #co-relation b/w likes and comments_count
corr_likes_comments
```

Out[ ]:
```
0.628280125628402
```