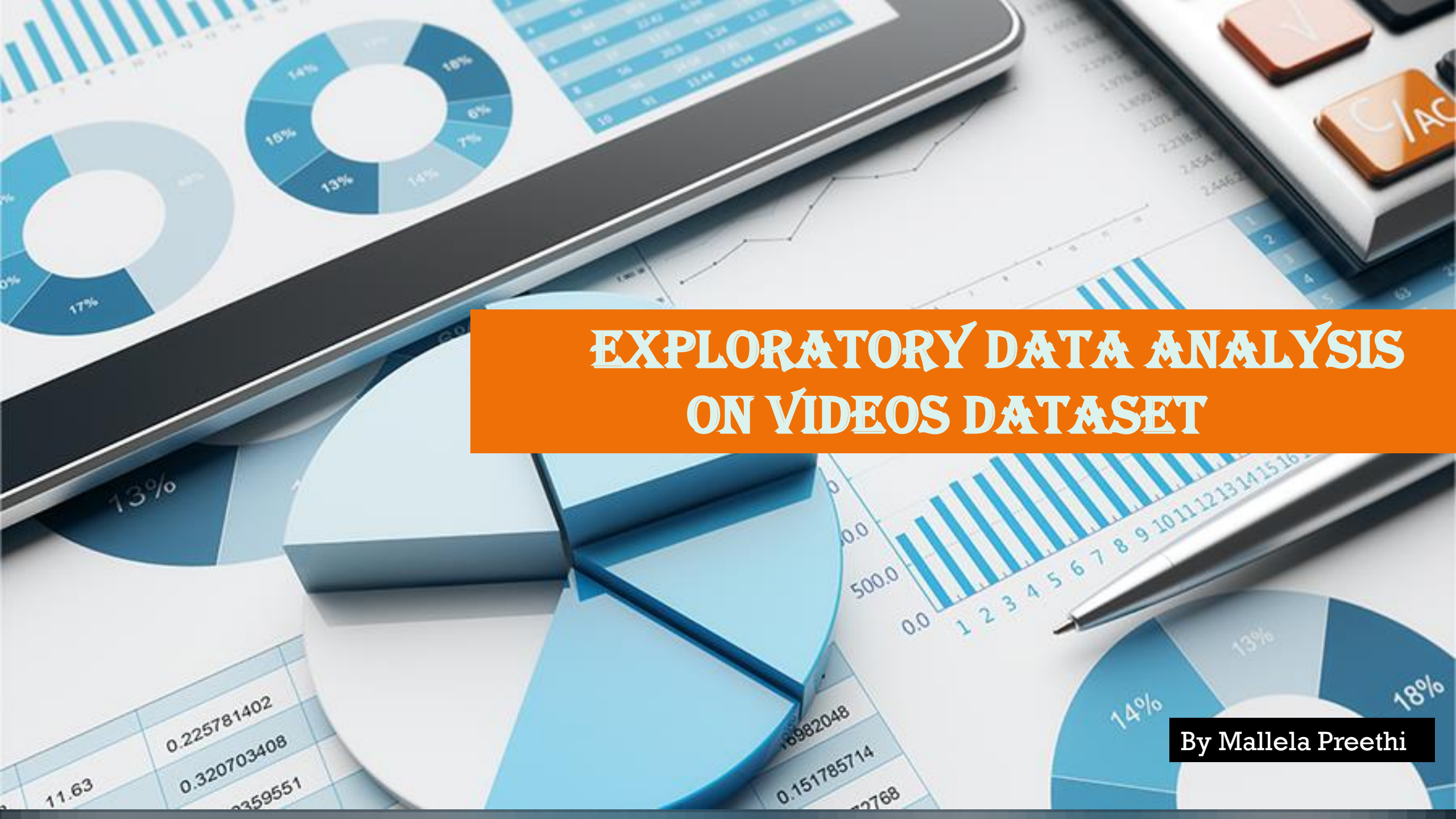


EXPLORATORY DATA ANALYSIS ON VIDEOS DATASET

By Mallela Preethi

Understanding Data

```
df.shape
```

```
(40949, 16)
```

```
df.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
```

```
df.isnull().sum()
```

```
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               570
dtype: int64
```

```
[ ]  df.duplicated().sum()
```

```
48
```

Data Cleaning and Preprocessing steps

1. Handled Missing Values
2. Handled Duplicate Records
3. Converted some columns to correct datatypes
4. Removed some columns
5. Formatted column names
6. Created new columns based on need
7. Extracted Year, Month, Hour from Published date-time and Trending date columns

After Preprocessing

```
df.shape
```

```
(40901, 19)
```

```
df.duplicated().sum()
```

```
0
```

```
df.isnull().sum()
```

```
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
Comments_Disabled         0
Ratings_Disabled          0
Video_Error_Or_Removed    0
Published_Month           0
Published_Day             0
Published_Hour            0
Published_Year            0
Category_Name             0
dtype: int64
```

```
df.info()
```

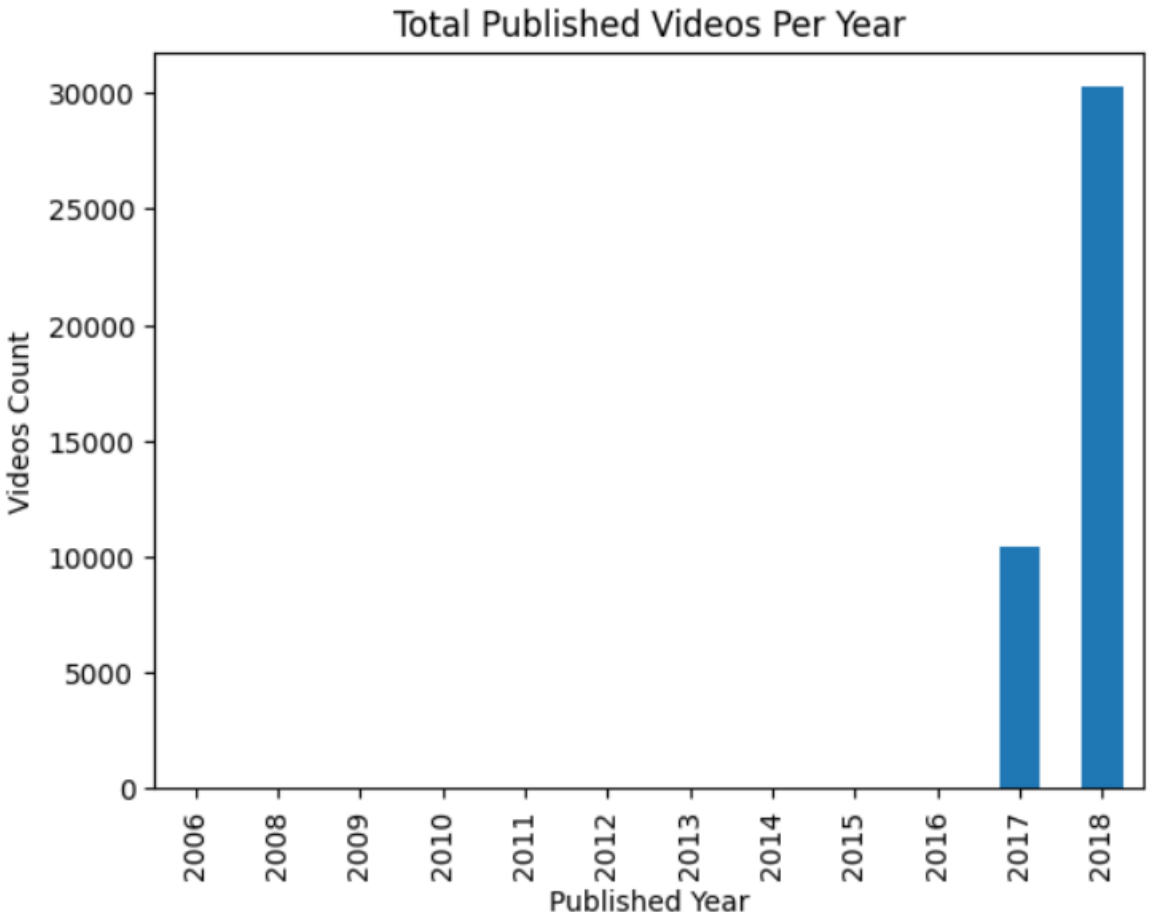
```
<class 'pandas.core.frame.DataFrame'>
Index: 40901 entries, 0 to 40948
Data columns (total 19 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Video_Id                40901 non-null  object
 1   Trending_Date           40901 non-null  datetime64[ns]
 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  datetime64[ns, UTC]
 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
 14  Published_Month         40901 non-null  int32
 15  Published_Day           40901 non-null  int32
 16  Published_Hour          40901 non-null  int32
 17  Published_Year          40901 non-null  int32
 18  Category_Name           40901 non-null  object
dtypes: bool(3), datetime64[ns, UTC](1), datetime64[ns](1), int32(4), int64(5), object(5)
memory usage: 5.8+ MB
```

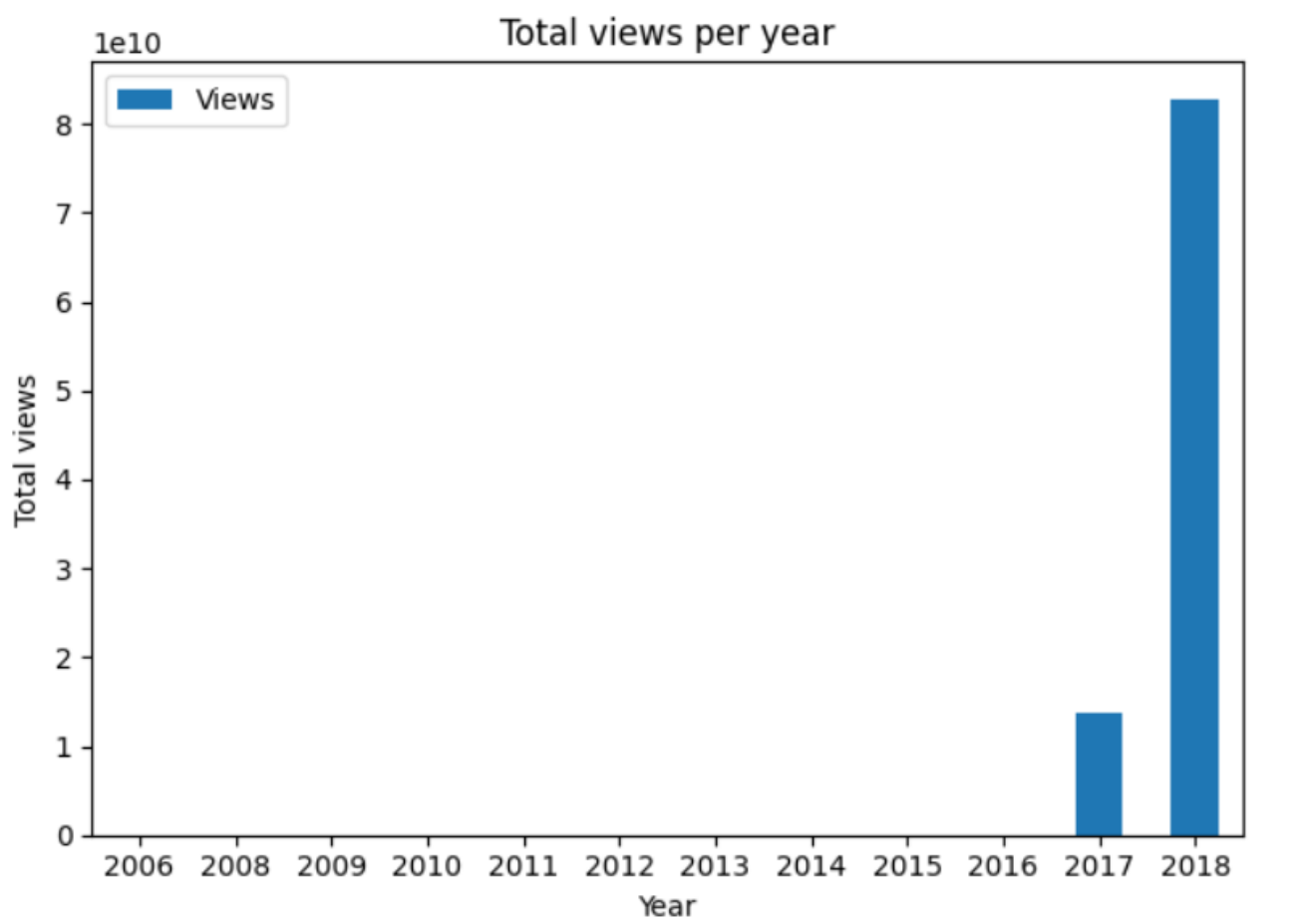
Summary Statistics

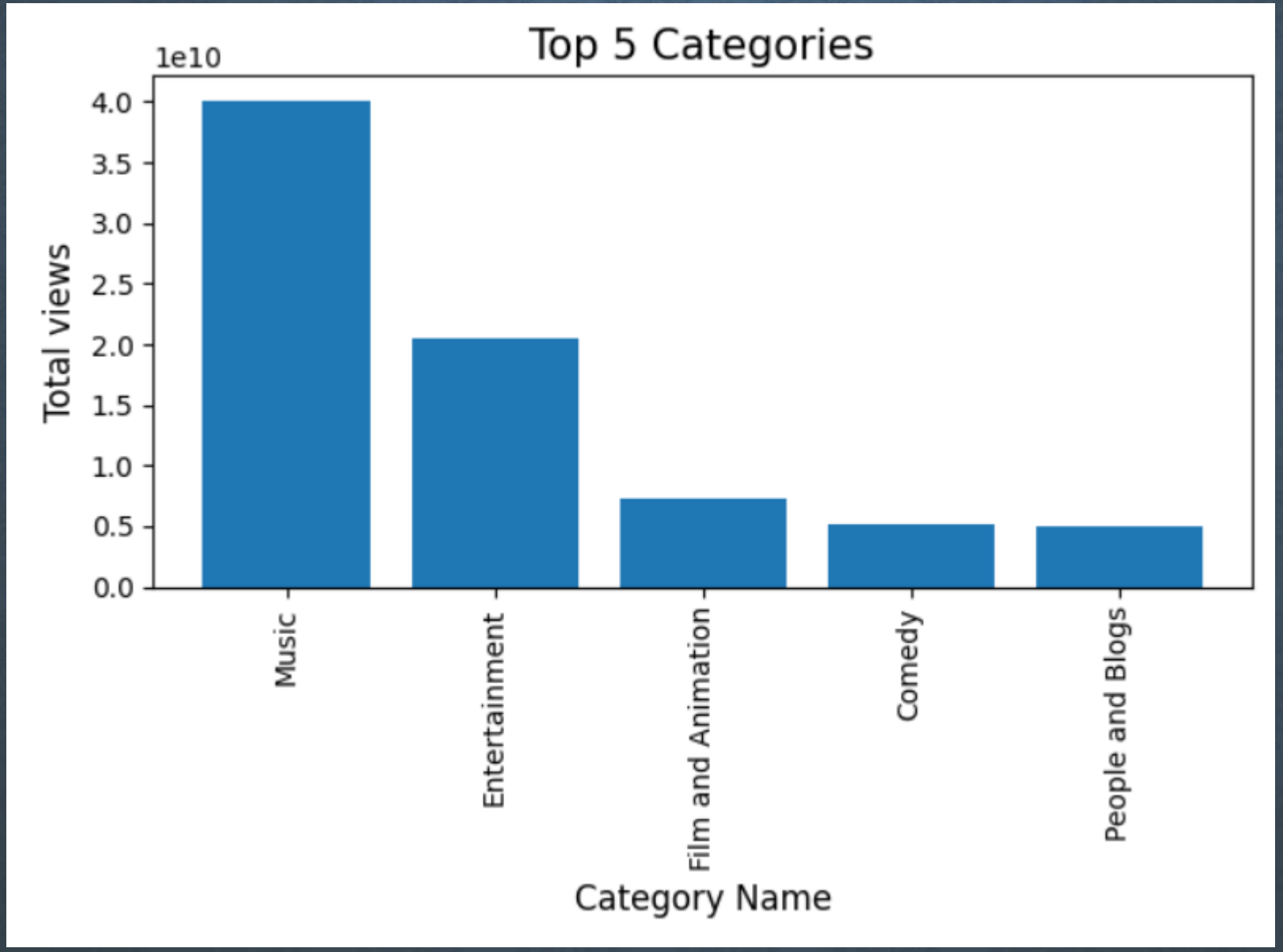
```
df.describe()
```

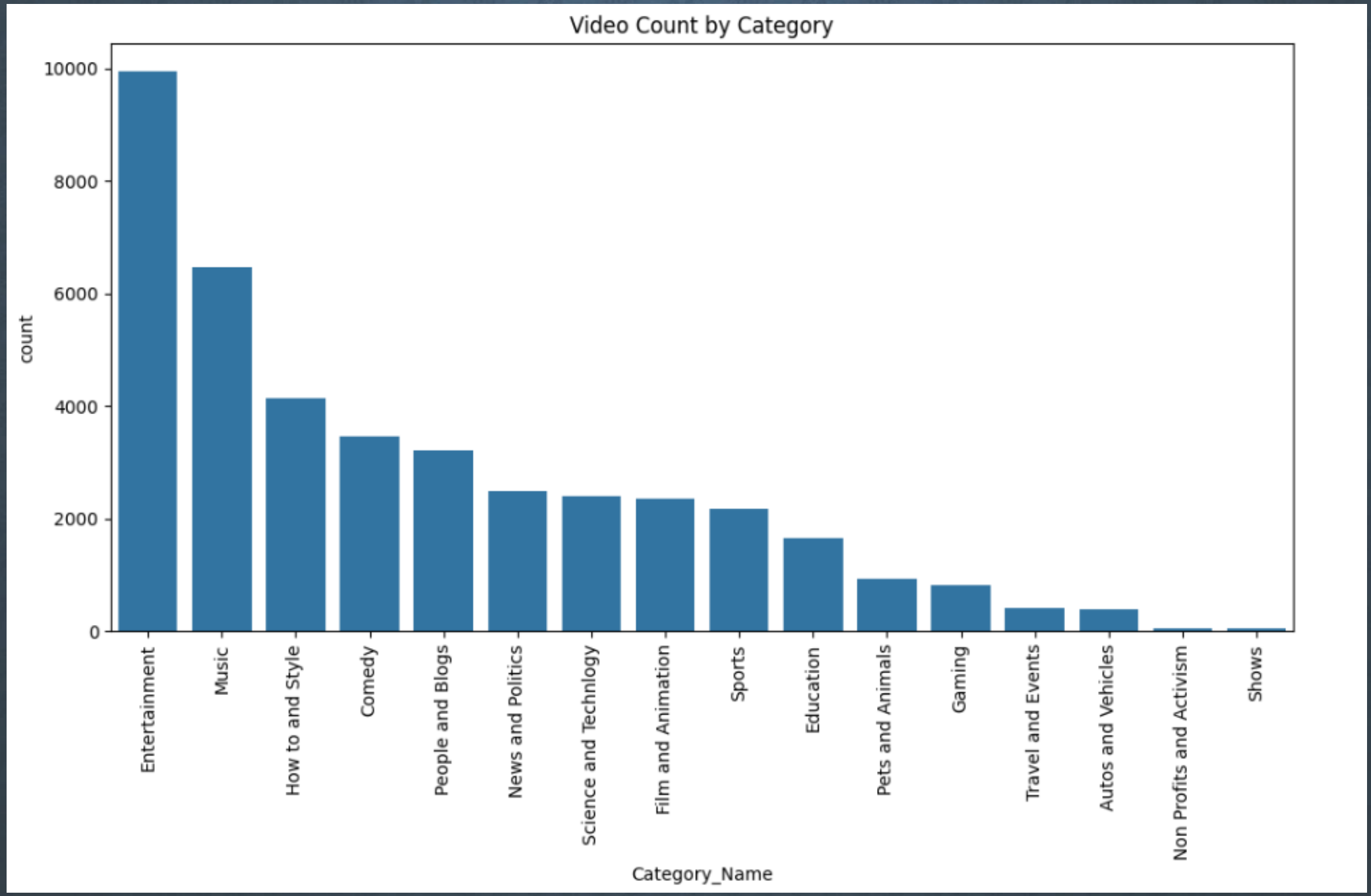
	category_id	views	likes	dislikes	comment_count
count	40901.000000	4.090100e+04	4.090100e+04	4.090100e+04	4.090100e+04
mean	19.970588	2.360678e+06	7.427173e+04	3.711722e+03	8.448567e+03
std	7.569362	7.397719e+06	2.289999e+05	2.904624e+04	3.745139e+04
min	1.000000	5.490000e+02	0.000000e+00	0.000000e+00	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

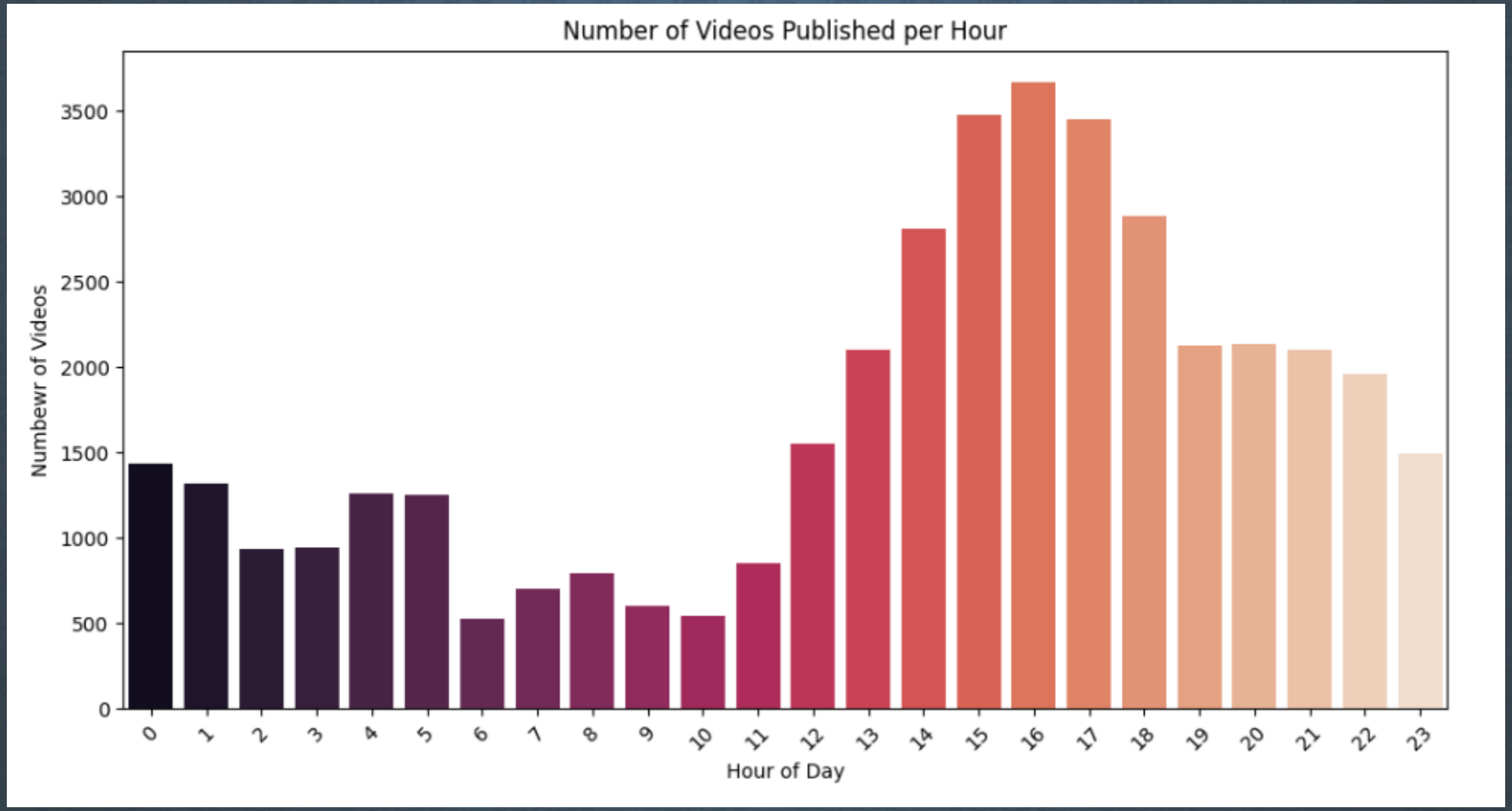
Visualizations

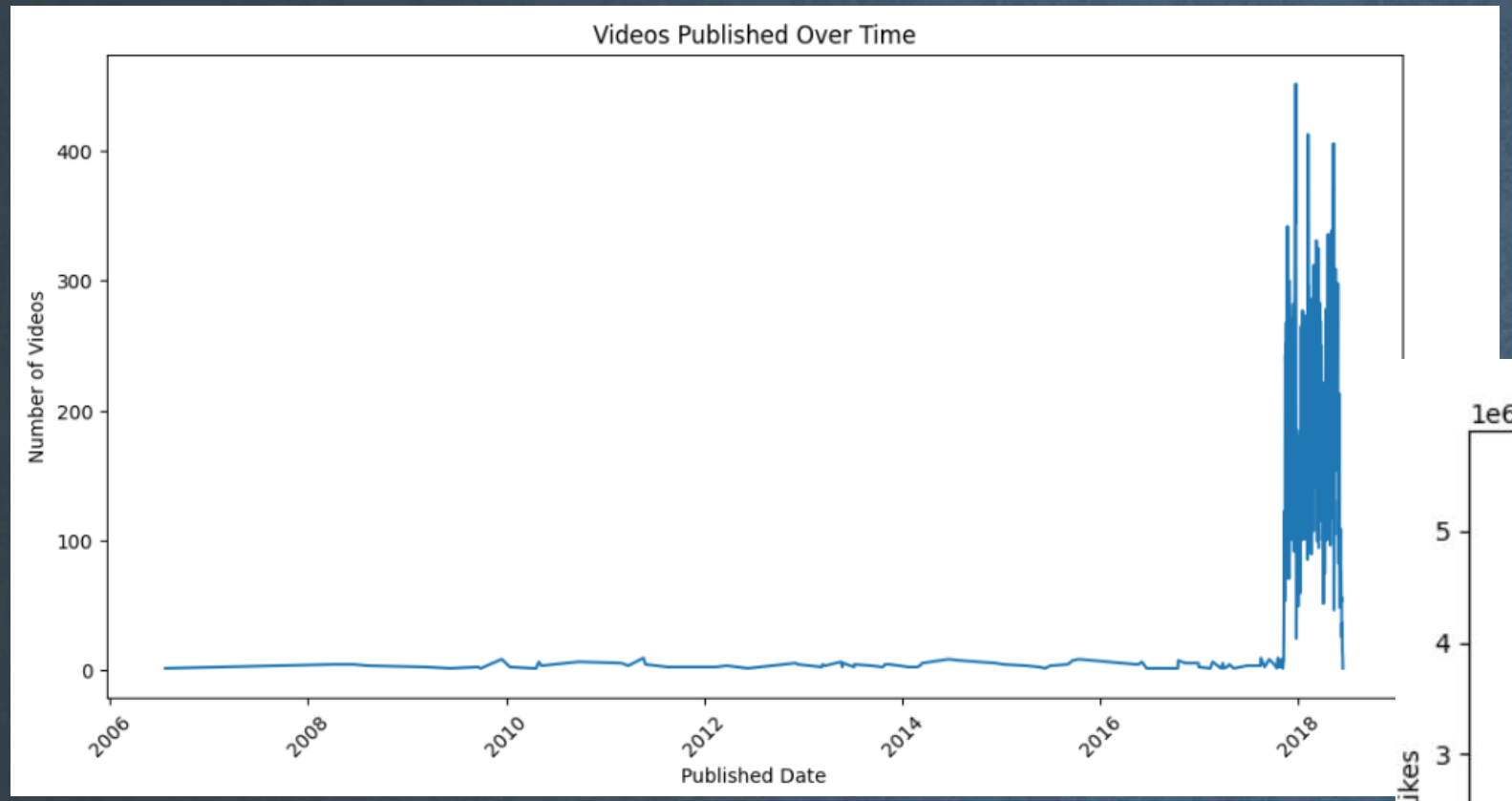


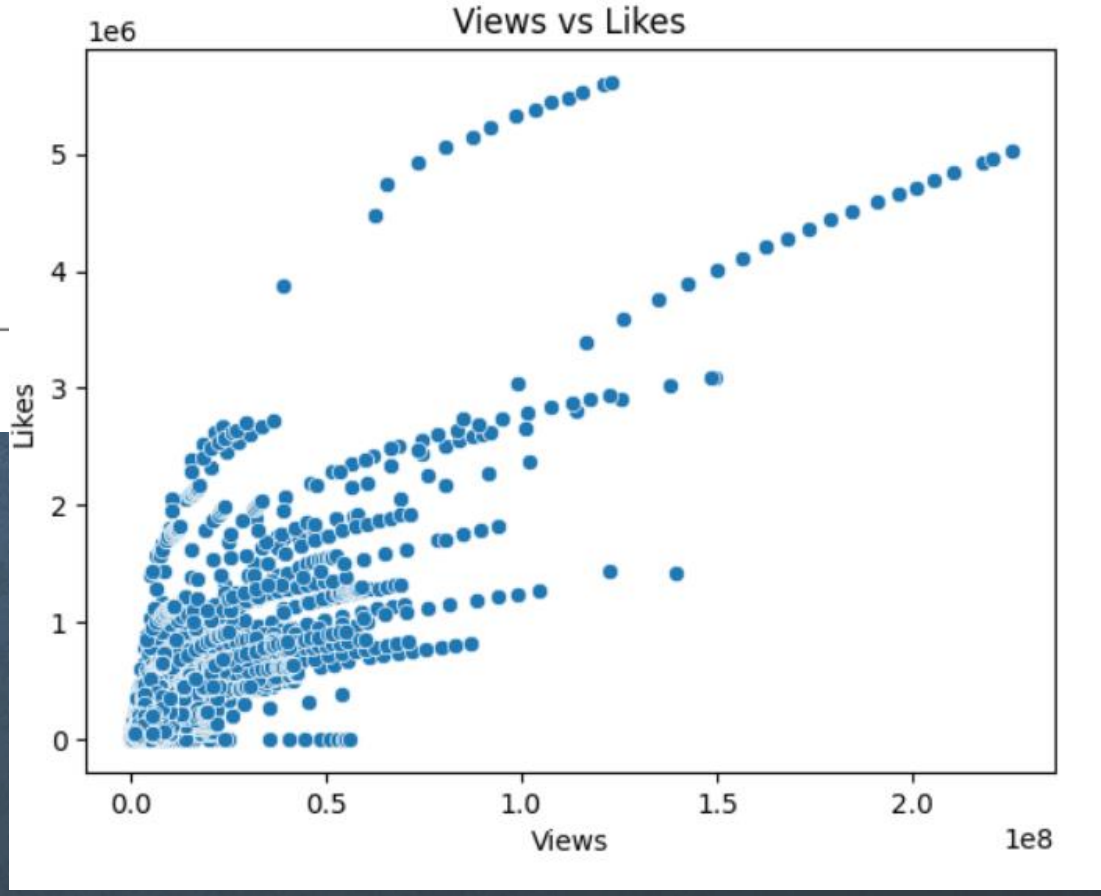


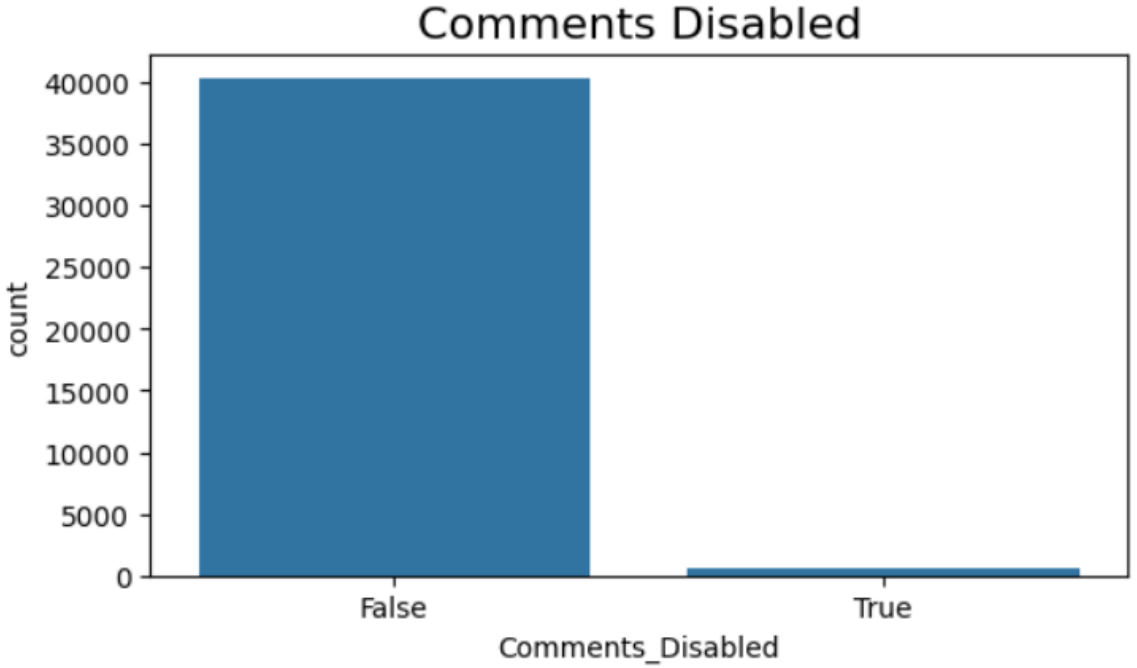

Top 5 Categories
1e10
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Total views
Music
Entertainment
Film and Animation
Comedy
People and Blogs
Category Name

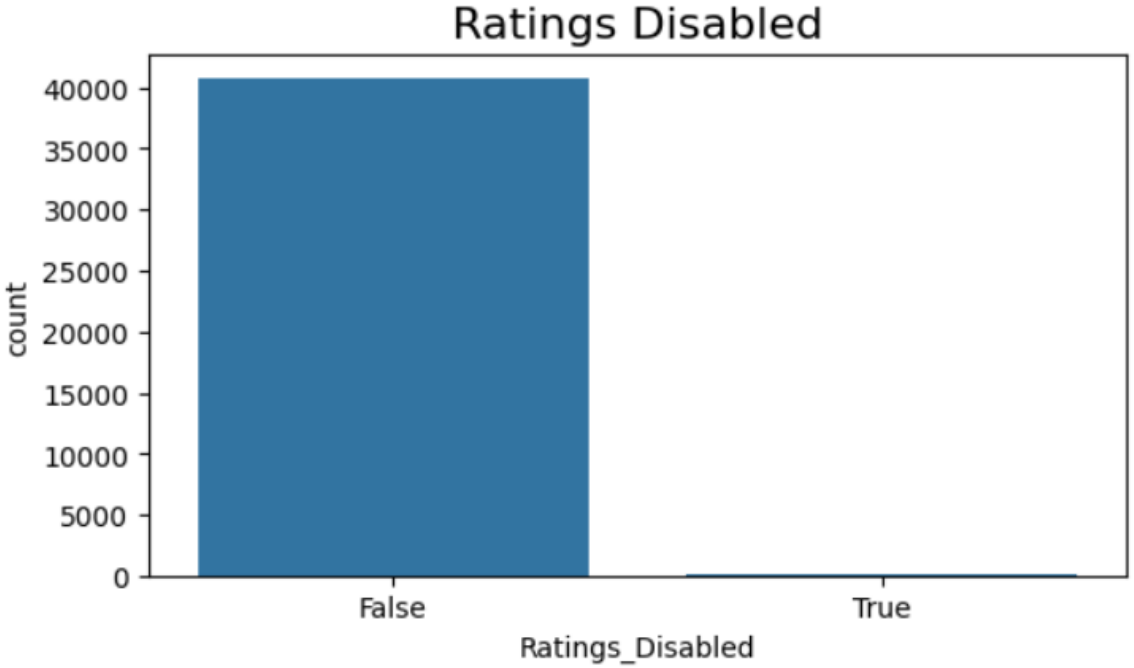

Video Count by Category
count
10000
8000
6000
4000
2000
0
Entertainment
Music
How to and Style
Comedy
People and Blogs
News and Politics
Science and Technlogy
Film and Animation
Sports
Education
Pets and Animals
Gaming
Travel and Events
Autos and Vehicles
Non Profits and Activism
Shows
Category_Name

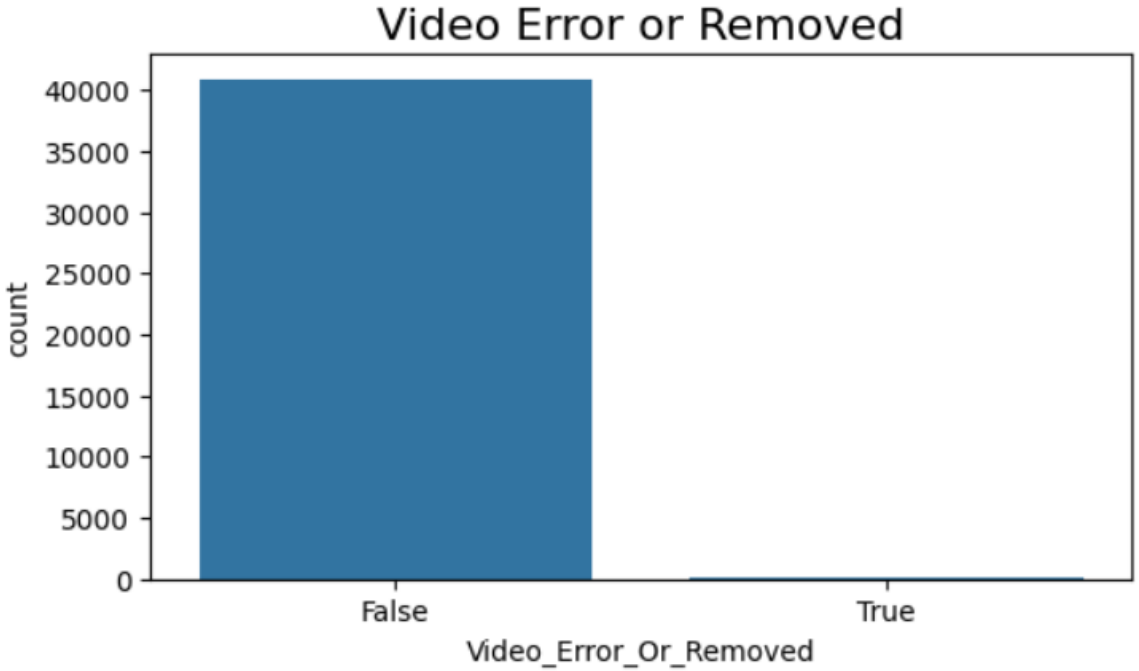

Number of Videos Published per Hour
Numbewr of Videos
0
500
1000
1500
2000
2500
3000
3500
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
Hour of Day

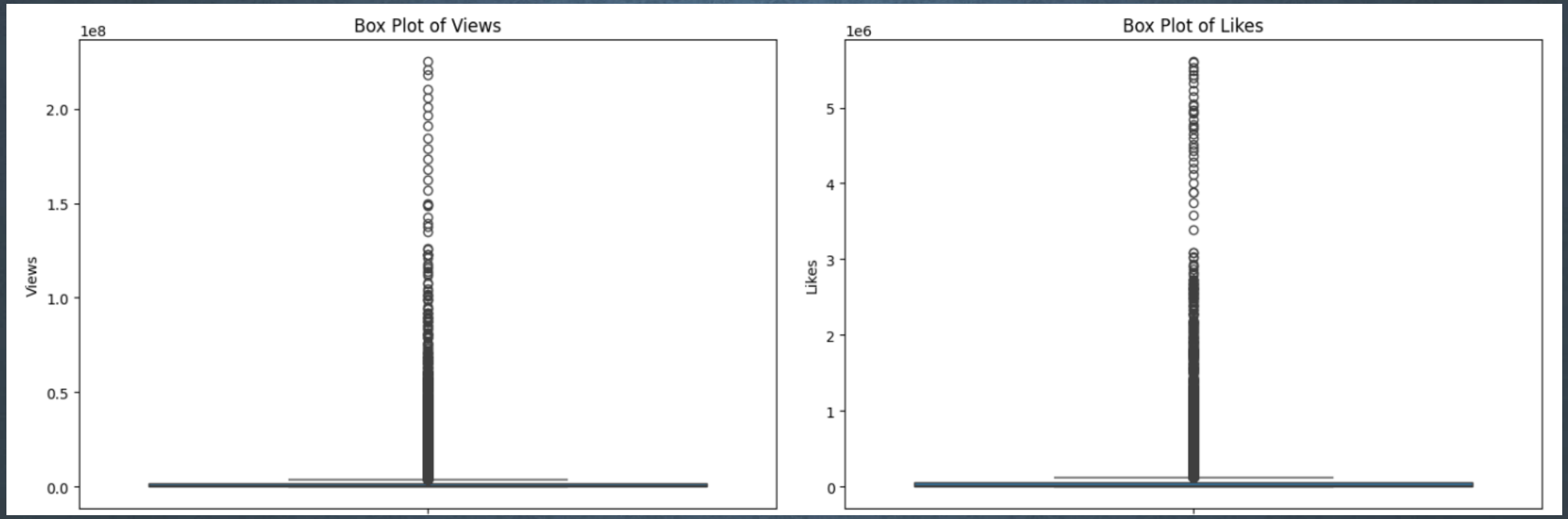

Videos Published Over Time
Number of Videos
0
100
200
300
400
2006
2008
2010
2012
2014
2016
2018
Published Date

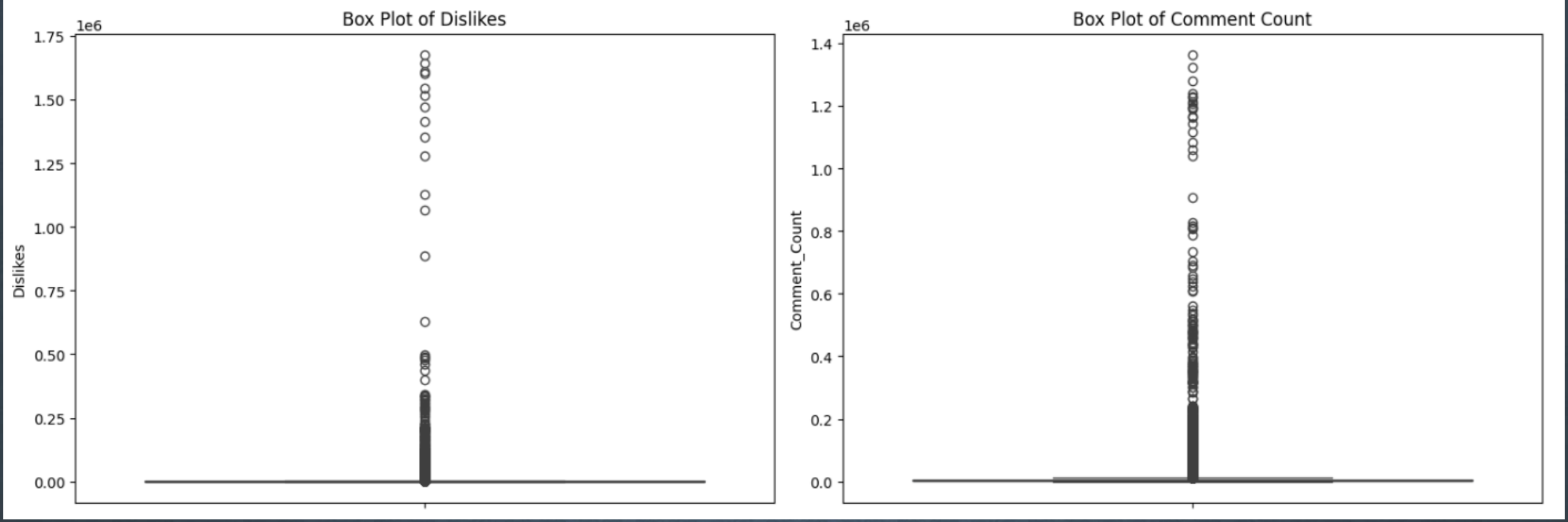

Views vs Likes
1e6
Likes
0
1
2
3
4
5
0.0
0.5
1.0
1.5
2.0
Views
1e8

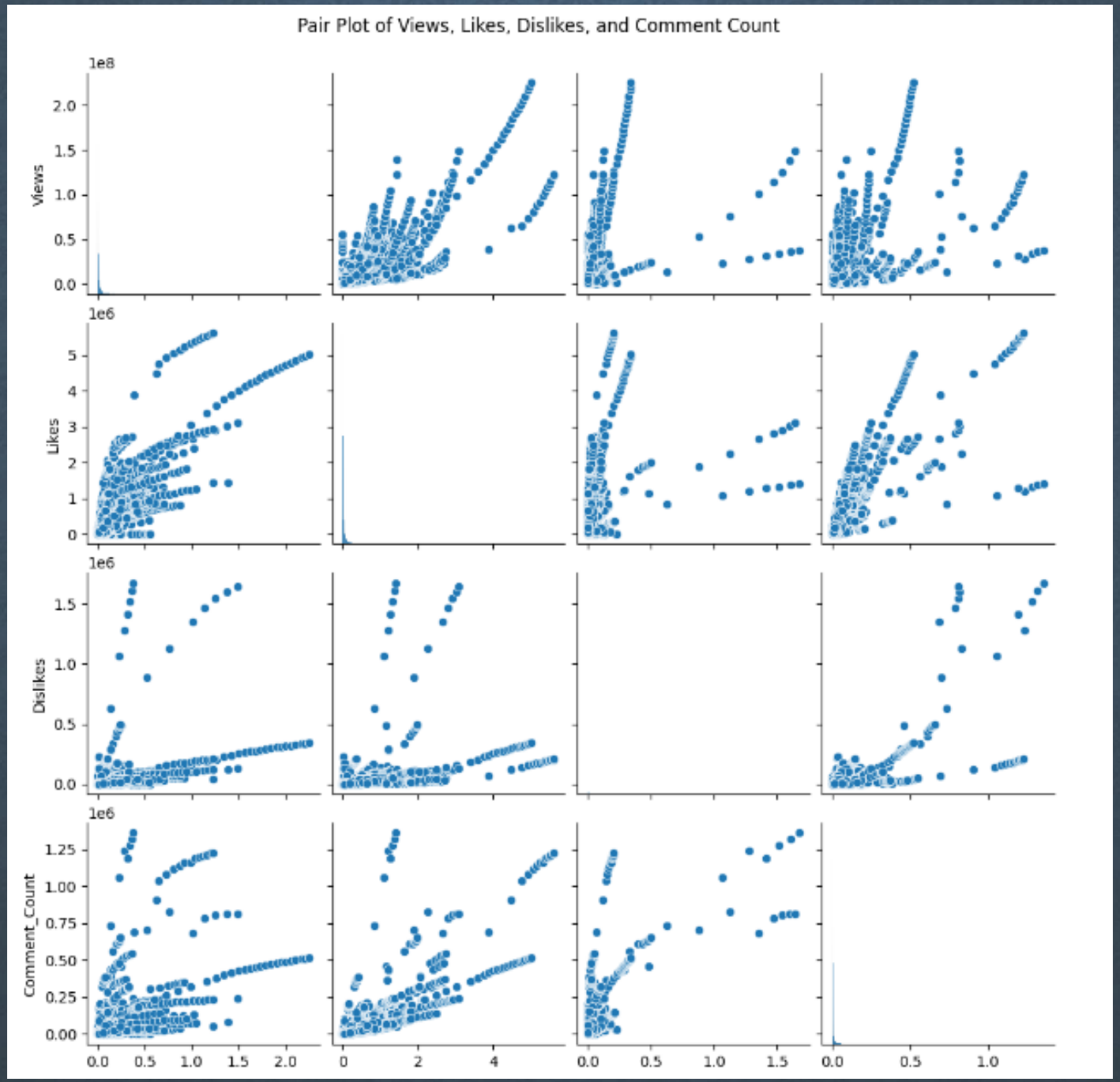

Comments Disabled
40000
35000
30000
25000
20000
15000
10000
5000
0
count
False
True
Comments_Disabled

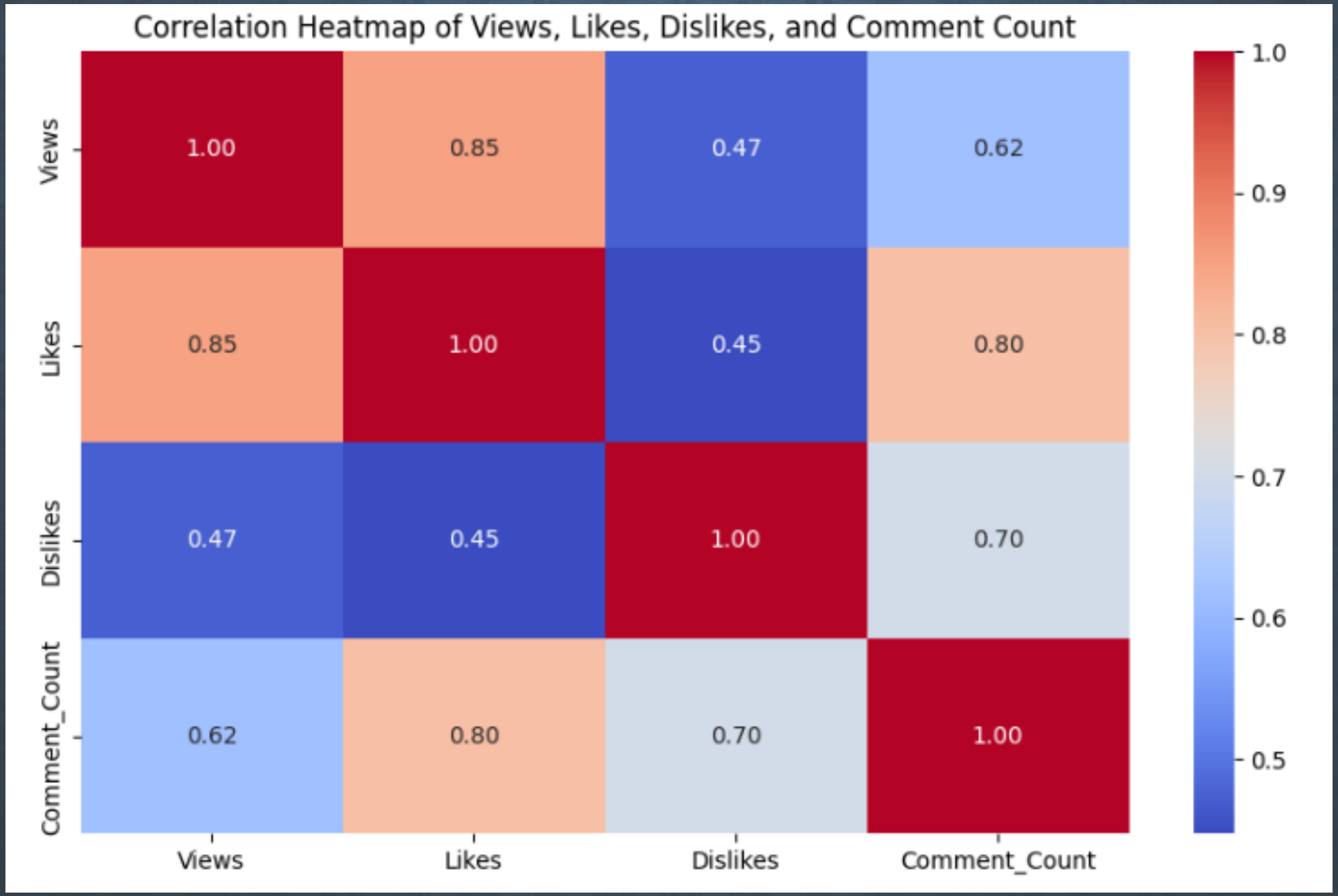

Ratings Disabled
40000
35000
30000
25000
20000
15000
10000
5000
0
count
False
True
Ratings_Disabled


Video Error or Removed
40000
35000
30000
25000
20000
15000
10000
5000
0
count
False
True
Video_Error_Or_Removed


Box Plot of Views
1e8
Views
2.0
1.5
1.0
0.5
0.0
Box Plot of Likes
1e6
Likes
5
4
3
2
1
0


Box Plot of Dislikes
1e6
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Dislikes
Box Plot of Comment Count
1e6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Comment_Count


Pair Plot of Views, Likes, Dislikes, and Comment Count
Views
Likes
Dislikes
Comment_Count
1e8
1e6
1e6
1e6


Correlation Heatmap of Views, Likes, Dislikes, and Comment Count
Views
Likes
Dislikes
Comment_Count
1.00
0.85
0.47
0.62
0.85
1.00
0.45
0.80
0.47
0.45
1.00
0.70
0.62
0.80
0.70
1.00
Views
Likes
Dislikes
Comment_Count
1.0
0.9
0.8
0.7
0.6
0.5


Thank you!