COMPLETE DATA ANALYSIS ON YOUTUBE DATASET

IMPORTING MODULES

In [22]:
```
import pandas as pd
import numpy as np
import seaborn as sns
from matplotlib import cm
import warnings
warnings.filterwarnings("ignore")
import matplotlib.pyplot as plt
```

LOADING THE DATASET

In [23]:
```
Youtube_df = pd.read_csv('youtube.csv')
```

In [24]:
```
Youtube_df
```

Out[24]:

	index	video_id	trending_date	title	channel_title	categ
0	0	2kyS6SvSYSE	17.14.11	WE WANT TO TALK ABOUT OUR MARRIAGE	CaseyNeistat	
1	1	1ZAPwfrtAFY	17.14.11	The Trump Presidency: Last Week Tonight with J...	LastWeekTonight	
2	2	5qpjK5DgCt4	17.14.11	Racist Superman | Rudy Mancuso, King Bach & Le...	Rudy Mancuso	
3	3	puqaWrEC7tY	17.14.11	Nickelback Lyrics: Real or Fake?	Good Mythical Morning	
4	4	d380meD0W0M	17.14.11	I Dare You: GOING BALD!?	nigahiga	
...	...	...	...	...	...	
161465	161465	sGolxsMSGfQ	18.14.06	HOW2: How to Solve a Mystery	Annoying Orange	
161466	161466	8HNuRNi8t70	18.14.06	Eli Lik Lik Episode 13 Partie 01	Elhiwar Ettounsi	
161467	161467	GWIKEM3m2EE	18.14.06	KINGDOM HEARTS III â€“ SQUARE ENIX E3 SHOWCASE...	Kingdom Hearts	
161468	161468	lbMKLzQ4cNQ	18.14.06	Trump Advisor Grovels To Trudeau	The Young Turks	
161469	161469	POTgw38-m58	18.14.06	ã€□å®Œæ•´ç‰ˆã€‘é□‡å^°æ□□æ€–æƒ… ä°°è©²æ€Žé°¼è¾¦ï¼...	æ^'æ„›å°□æ˜Žæ˜Ÿå¤§è·Ÿç□	

161470 rows × 18 columns

VIEWING DATA TO UNDERSTAND THE VARIABLES AND LABELS

In [25]:
```
Youtube_df.head()
```

Out[25]:

	index	video_id	trending_date	title	channel_title	category_id	publish_date	time
0	0	2kyS6SvSYSE	17.14.11	WE WANT TO TALK ABOUT OUR MARRIAGE	CaseyNeistat	22	13/11/2017	1
1	1	1ZAPwfrtAFY	17.14.11	The Trump Presidency: Last Week Tonight with J...	LastWeekTonight	24	13/11/2017	7:00
2	2	5qpjK5DgCt4	17.14.11	Racist Superman | Rudy Mancuso, King Bach & Le...	Rudy Mancuso	23	12/11/2017	1
3	3	puqaWrEC7tY	17.14.11	Nickelback Lyrics: Real or Fake?	Good Mythical Morning	24	13/11/2017	1
4	4	d380meD0W0M	17.14.11	I Dare You: GOING BALD!?	nigahiga	24	12/11/2017	1

In [26]:
```
Youtube_df.tail()
```

Out[26]:

	index	video_id	trending_date	title	channel_title	categc
161465	161465	sGolxsMSGfQ	18.14.06	HOW2: How to Solve a Mystery	Annoying Orange	
161466	161466	8HNuRNi8t70	18.14.06	Eli Lik Lik Episode 13 Partie 01	Elhiwar Ettounsi	
161467	161467	GWlKEM3m2EE	18.14.06	KINGDOM HEARTS III â€" SQUARE ENIX E3 SHOWCASE...	Kingdom Hearts	
161468	161468	lbMKLzQ4cNQ	18.14.06	Trump Advisor Grovels To Trudeau	The Young Turks	
161469	161469	POTgw38-m58	18.14.06	ã€Œå®Œæ•´ç‰ˆã€‘é‡å°æ□□æ€–æƒ… ä°°è©²æ€Žé°¼è¾¦ï...	æˆ'æ„›å°□æ˜Žæ˜Ÿå¤§è·Ÿç□	

CHECKING NULL VALUES AND DATA TYPES

In [ ]:
```
Youtube_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 161470 entries, 0 to 161469
Data columns (total 18 columns):
 #   Column                  Non-Null Count   Dtype
---  ------                  --------------   -----
 0   index                   161470 non-null  int64
 1   video_id                161470 non-null  object
 2   trending_date           161470 non-null  object
 3   title                   161470 non-null  object
 4   channel_title           161470 non-null  object
 5   category_id             161470 non-null  int64
 6   publish_date            161470 non-null  object
 7   time_frame              161470 non-null  object
 8   published_day_of_week   161470 non-null  object
 9   publish_country         161470 non-null  object
 10  tags                    161470 non-null  object
 11  views                   161470 non-null  int64
 12  likes                   161470 non-null  int64
 13  dislikes                161470 non-null  int64
 14  comment_count           161470 non-null  int64
 15  comments_disabled       161470 non-null  bool
 16  ratings_disabled        161470 non-null  bool
 17  video_error_or_removed  161470 non-null  bool
dtypes: bool(3), int64(6), object(9)
memory usage: 18.9+ MB
```

In [ ]:
```
Youtube_df.describe()
```

Out[ ]:

	index	category_id	views	likes	dislikes	comment_count
count	161470.00000	161470.000000	1.614700e+05	1.614700e+05	1.614700e+05	1.614700e+05
mean	80734.50000	19.461151	2.419854e+06	6.566194e+04	3.490153e+03	7.035494e+03
std	46612.51832	7.432001	1.043749e+07	2.260617e+05	3.114779e+04	3.404121e+04
min	0.00000	1.000000	2.230000e+02	0.000000e+00	0.000000e+00	0.000000e+00
25%	40367.25000	15.000000	1.015382e+05	1.975000e+03	8.500000e+01	2.790000e+02
50%	80734.50000	23.000000	3.847395e+05	9.840000e+03	3.480000e+02	1.144000e+03
75%	121101.75000	24.000000	1.339528e+06	4.006275e+04	1.350000e+03	4.144750e+03
max	161469.00000	44.000000	4.245389e+08	5.613827e+06	1.944971e+06	1.626501e+06

REMOVING UNWANTED COLUMNS

In [ ]:
```
del_cols = ['comments_disabled','ratings_disabled', 'video_error_or_removed' , 'time_f
Youtube_df = Youtube_df.drop(del_cols, axis=1)
```

In [ ]:
```
Youtube_df
```

Out[ ]:

	index	video_id	trending_date	title	channel_title	categ
0	0	2kyS6SvSYSE	17.14.11	WE WANT TO TALK ABOUT OUR MARRIAGE	CaseyNeistat	
1	1	1ZAPwfrtAFY	17.14.11	The Trump Presidency: Last Week Tonight with J...	LastWeekTonight	
2	2	5qpjK5DgCt4	17.14.11	Racist Superman | Rudy Mancuso, King Bach & Le...	Rudy Mancuso	
3	3	puqaWrEC7tY	17.14.11	Nickelback Lyrics: Real or Fake?	Good Mythical Morning	
4	4	d380meD0W0M	17.14.11	I Dare You: GOING BALD!?	nigahiga	
...	...	...	...	...	...	
161465	161465	sGolxsMSGfQ	18.14.06	HOW2: How to Solve a Mystery	Annoying Orange	
161466	161466	8HNuRNi8t70	18.14.06	Eli Lik Lik Episode 13 Partie 01	Elhiwar Ettounsi	
161467	161467	GWIKEM3m2EE	18.14.06	KINGDOM HEARTS III â€“ SQUARE ENIX E3 SHOWCASE...	Kingdom Hearts	
161468	161468	lbMKLzQ4cNQ	18.14.06	Trump Advisor Grovels To Trudeau	The Young Turks	
161469	161469	POTgw38-m58	18.14.06	ã€□å®Œæ•´ç‰ˆâ€'é□‡å^°æ□□æ€–æƒ... ä°°è©²æ€Žé°¼è¾¦ï¼...	æ^'æ„›å°□æ˜Žæ˜Ÿå¤§è·Ÿç□	

161470 rows × 14 columns

CHECKING IF THERE ANY DUPLICTAE VALUES PRESENT IN THE CURRENT DATASET

In [ ]:
```
Youtube_df['duplicated'] = Youtube_df.duplicated()
result = Youtube_df['duplicated'].value_counts()
print(result)
```

```
False    161470
Name: duplicated, dtype: int64
```

CHANGING THE DATA TYPE IN THE CORRECT FORMAT FOR THE DATE DATA

TYPE

In [ ]:
```
Youtube_df['trending_date'] = pd.to_datetime(Youtube_df['trending_date'], format='%y.%
Youtube_df['publish_date'] = pd.to_datetime(Youtube_df['publish_date'], format='%d/%m/
```

In [ ]:
```
Youtube_df
```

Out[ ]:

	index	video_id	trending_date	title	channel_title	categ
0	0	2kyS6SvSYSE	2017-11-14	WE WANT TO TALK ABOUT OUR MARRIAGE	CaseyNeistat	
1	1	1ZAPwfrtAFY	2017-11-14	The Trump Presidency: Last Week Tonight with J...	LastWeekTonight	
2	2	5qpjK5DgCt4	2017-11-14	Racist Superman | Rudy Mancuso, King Bach & Le...	Rudy Mancuso	
3	3	puqaWrEC7tY	2017-11-14	Nickelback Lyrics: Real or Fake?	Good Mythical Morning	
4	4	d380meD0W0M	2017-11-14	I Dare You: GOING BALD!?	nigahiga	
...	...	...	...	...	...	
161465	161465	sGolxsMSGfQ	2018-06-14	HOW2: How to Solve a Mystery	Annoying Orange	
161466	161466	8HNuRNi8t70	2018-06-14	Eli Lik Lik Episode 13 Partie 01	Elhiwar Ettounsi	
161467	161467	GWIKEM3m2EE	2018-06-14	KINGDOM HEARTS III â€" SQUARE ENIX E3 SHOWCASE...	Kingdom Hearts	
161468	161468	lbMKLzQ4cNQ	2018-06-14	Trump Advisor Grovels To Trudeau	The Young Turks	
161469	161469	POTgw38-m58	2018-06-14	ã€□å®Œæ•´ç‰ˆã€'é□‡å^°æ□□æ€–æƒ... ä°°è©²æ€Žé°¼è¾¦ï...	æ^'æ„›å°□æ˜Žæ˜Ÿå¤§è·Ÿç□	

161470 rows × 15 columns

In [ ]:
```
Youtube_df = Youtube_df.drop("duplicated", axis=1)
Youtube_df
```

Out[ ]:

	index	video_id	trending_date	title	channel_title	categ
0	0	2kyS6SvSYSE	2017-11-14	WE WANT TO TALK ABOUT OUR MARRIAGE	CaseyNeistat	
1	1	1ZAPwfrtAFY	2017-11-14	The Trump Presidency: Last Week Tonight with J...	LastWeekTonight	
2	2	5qpjK5DgCt4	2017-11-14	Racist Superman | Rudy Mancuso, King Bach & Le...	Rudy Mancuso	
3	3	puqaWrEC7tY	2017-11-14	Nickelback Lyrics: Real or Fake?	Good Mythical Morning	
4	4	d380meD0W0M	2017-11-14	I Dare You: GOING BALD!?	nigahiga	
...	...	...	...	...	...	
161465	161465	sGolxsMSGfQ	2018-06-14	HOW2: How to Solve a Mystery	Annoying Orange	
161466	161466	8HNuRNi8t70	2018-06-14	Eli Lik Lik Episode 13 Partie 01	Elhiwar Ettounsi	
161467	161467	GWIKEM3m2EE	2018-06-14	KINGDOM HEARTS III â€“ SQUARE ENIX E3 SHOWCASE...	Kingdom Hearts	
161468	161468	lbMKLzQ4cNQ	2018-06-14	Trump Advisor Grovels To Trudeau	The Young Turks	
161469	161469	POTgw38-m58	2018-06-14	ã€□å®Œæ•´ç‰ˆâ€'é□‡å^°æ□□æ€–æƒ… äº°è©²æ€Žé°¼è¾¦ï¼...	æˆ'æ„›å°□æ˜Žæ˜Ÿå¤§è·Ÿç□	

161470 rows × 14 columns

CREATING A FUNCTION FOR VIEWING TOP 5 AND LAST 5 VALUES

In [ ]:
```
def print_top5(column_of_interest,column_stats):
    df = Youtube_df.groupby(column_of_interest)['views', 'likes','comment_count', 'dis
    return df.sort_values(by=column_stats, ascending=False).head()
```

In [ ]:
```
def print_last5(column_of_interest,column_stats):
    df = Youtube_df.groupby(column_of_interest)['views', 'likes','comment_count', 'dis
    return df.sort_values(by=column_stats, ascending=True).head()
```

Top 5 Videos with Highest Views

In [ ]:
```
print_top5('title','views')
```

Out[ ]:

title	views	likes	comment_count	dislikes
Childish Gambino - This Is America (Official Video)	10425338513	265935343	27796551	16564194
Nicky Jam x J. Balvin - X (EQUIS) | Video Oficial | Prod. Afro Bros & Jeon	8395094350	62438438	2316279	2973389
Bad Bunny - Amorfoda | Video Oficial	6970018019	111253872	6439286	5398588
Ariana Grande - No Tears Left To Cry	5872980845	162132923	13498459	6306520
Te Bote Remix - Casper, Nio GarcÃa, Darell, Nicky Jam, Bad Bunny, Ozuna | Video Oficial	5168434018	54135423	2025882	2819055

Videos With Lowest Views

In [ ]:
```
print_last5('title','views')
```

Out[ ]:

title	views	likes	comment_count	dislikes
France: Famous climber Alain Rober scales skyscraper in Paris business district	284	3	0	0
L'histoire des enfants Tinning | La mÃ¨re aux neuf enfants dÃ©cÃ©dÃ©s	365	43	14	2
Feì□vrier 2018, Performance 8 10 Ans	374	43	5	0
La chanson de nos 14 ans (15/02/2018) - Best Of de Bruno dans la Radio	385	40	3	0
2/2 â□© Coffret de communication : Les RJ45 STP avec cÃ¢ble Grade 3	404	47	5	2

Top 5 Videos with Highest likes

In [ ]:
```
print_top5('title','likes')
```

Out[ ]:

title	views	likes	comment_count	dislikes
Childish Gambino - This Is America (Official Video)	10425338513	265935343	27796551	16564194
BTS (ë°©íƒ„ì†Œë…„ë‹¨) 'FAKE LOVE' Official MV	3055662678	179849973	38614637	5768877
Ariana Grande - No Tears Left To Cry	5872980845	162132923	13498459	6306520
Drake - Godâ€™s Plan	4895081566	136390285	8772422	2560165
Bad Bunny - Amorfoda | Video Oficial	6970018019	111253872	6439286	5398588

Videos With Lowest Likes

In [ ]:
```
print_last5('title','likes')
```

Out[ ]:

title	views	likes	comment_count	dislikes
ðŸ•Œ Ramadan 2018 | Msemen Farcis bechahma Un vrai DÃ©lice | Ù…Ø³Ù…Ù† Ø¨Ø§Ù„Ø´ØÙ… Ù„Ø°ÙŠØ°	59400	0	314	0
A Hidden Warning from History!	42746	0	1053	0
à®šà®±à¯à®±à¯ à®®à¯à®©à¯ à®šà®šà®¿à®•à®²à®µà®°à¯ à®¨à®Ÿà®¿à®•à¯ à®‰à®Ÿà®²à¯à®¨à®²à¯à®®à¯à®±à¯ à®•à®¤à®±à®¿ à®…à®´à¯à®¤ à®²à®±à®¿ à®¤à¯à®²à®²à®®à¯ - Sasikala Husband Died	907048	0	182	0
Le rorqual Ã©chouÃ© en Gironde dissÃ©quÃ© par les scientifiques du Pelagis de La Rochelle	3080	0	0	1
I Like Beer (:60) | Michelob ULTRA Super Bowl	2868544	0	60	0

Top 5 Videos with Highest Dislikes

In [ ]:
```
print_top5('title','dislikes')
```

Out[ ]:

title	views	likes	comment_count	dislikes
YouTube Rewind: The Shape of 2017 | #YouTubeRewind	4070143496	96438271	27876748	49680581
So Sorry.	712094226	27689216	27271712	31575810
Childish Gambino - This Is America (Official Video)	10425338513	265935343	27796551	16564194
Suicide: Be Here Tomorrow.	547739263	49620129	16476651	11246153
Lucas Lucco e Pabllo Vittar - ParaÃso	1073294377	28344732	6861819	10006058

Videos With Lowest DisLikes

In [ ]:
```
print_last5('title','dislikes')
```

Out[ ]:

title	views	likes	comment_count	dislikes
ðŸ•Œ Ramadan 2018 | Msemen Farcis bechahma Un vrai DÃ©lice | Ù…Ø³Ù…Ù† Ø¨Ø§Ù„Ø´Ø§Ù‡ Ù„Ø°ÙŠØ°	59400	0	314	0
un crash en voiture Ã NÃ¼rburgring,	4230	3	0	0
Christmas Day 2000	24546	19	1	0
PlanÃ¨te Rap - RomÃ©o Elvis Morale2Luxe #Jeudi	2274	74	38	0
Pity x 3.S1 x SÃ©rie Sims 4	1910	0	57	0

Top 5 Channels with Highest likes

In [ ]:
```
print_top5('channel_title', 'views')
```

Out[ ]:

channel_title	views	likes	comment_count	dislikes
ChildishGambinoVEVO	10425338513	265935343	27796551	16564194
NickyJamTV	9118588957	77527596	4234962	4951650
Ozuna	8465233685	76620691	3263914	4966543
DrakeVEVO	7434941039	178049699	11146185	3975055
Marvel Entertainment	7037198870	197199206	22986630	4020864

Channels With Lowest likes

In [ ]:
```
print_last5('channel_title', 'views')
```

Out[ ]:

channel_title	views	likes	comment_count	dislikes
No Comment TV	284	3	0	0
NavylittleMonster	365	43	14	2
Melimelo78	374	43	5	0
Pix'Perience	528	67	1	0
Natsu et Lucy Heartfilia	543	56	13	0

Correlation Matrix for Identifying Relationships

In [ ]:
```
columns_of_interest = ['views', 'likes', 'dislikes','comment_count']
correlation_matrix = Youtube_df[columns_of_interest].corr()
correlation_matrix
```

Out[ ]:

	views	likes	dislikes	comment_count
views	1.000000	0.786897	0.414189	0.499026
likes	0.786897	1.000000	0.453304	0.776422
dislikes	0.414189	0.453304	1.000000	0.732969
comment_count	0.499026	0.776422	0.732969	1.000000

Visualizing Correlation Matrix With Help Of Heatmap

In [ ]:
```
plt.figure(figsize=(12, 10))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f", linewidths=.4)
plt.title('Correlation Matrix Heatmap', fontsize=16)
plt.show()
```

Channels With The Highest Number Of Videos

In [ ]:
```
Youtube_df['channel_title'].value_counts().head()
```

Out[ ]:
```
The Late Show with Stephen Colbert        653
Late Night with Seth Meyers               594
TheEllenShow                              586
The Tonight Show Starring Jimmy Fallon    569
Jimmy Kimmel Live                         560
Name: channel_title, dtype: int64
```

Channels With The Lowest Number Of Videos

In [ ]:
```
Youtube_df['channel_title'].value_counts().tail()
```

Out[ ]:
```
RapLyricsVEVO        1
Mazen Zbatata        1
Karismatik Dancer    1
Lucile Joseph        1
SevenSuperGirls      1
Name: channel_title, dtype: int64
```

Countries With Highest Number Of Views

In [ ]:
```
print_top5('publish_country','views')
```

Out[ ]:

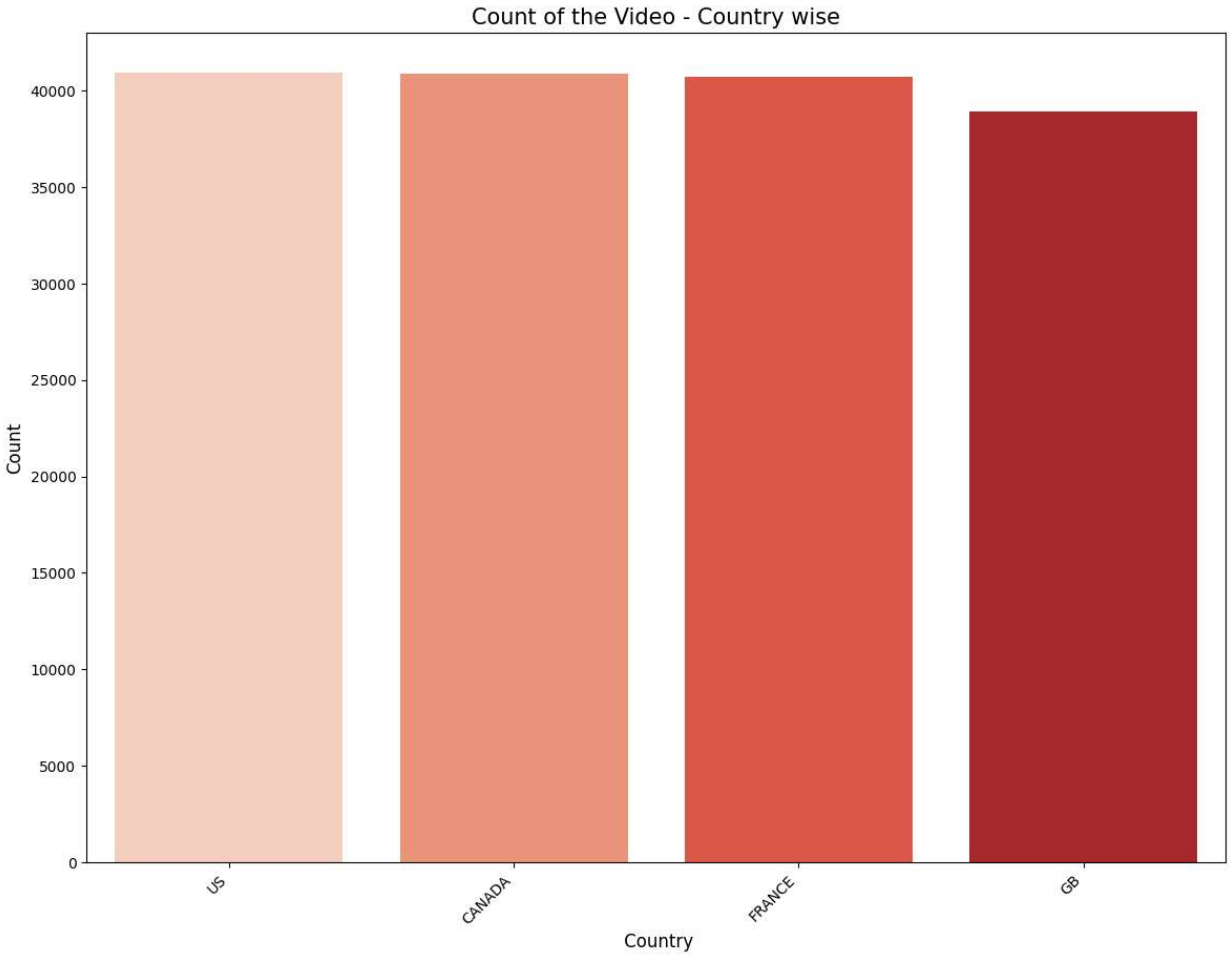
publish_country	views	likes	comment_count	dislikes
GB	230069198174	5234962944	509346351	296250384
US	96671770152	3041147198	345888164	151978155
CANADA	46891975069	1618179878	206161849	82137919
FRANCE	17100897444	708144090	74624804	33188528

Extracting The Modified Dataset

In [ ]:
```
Youtube_df.to_csv('youtube2.csv')
```

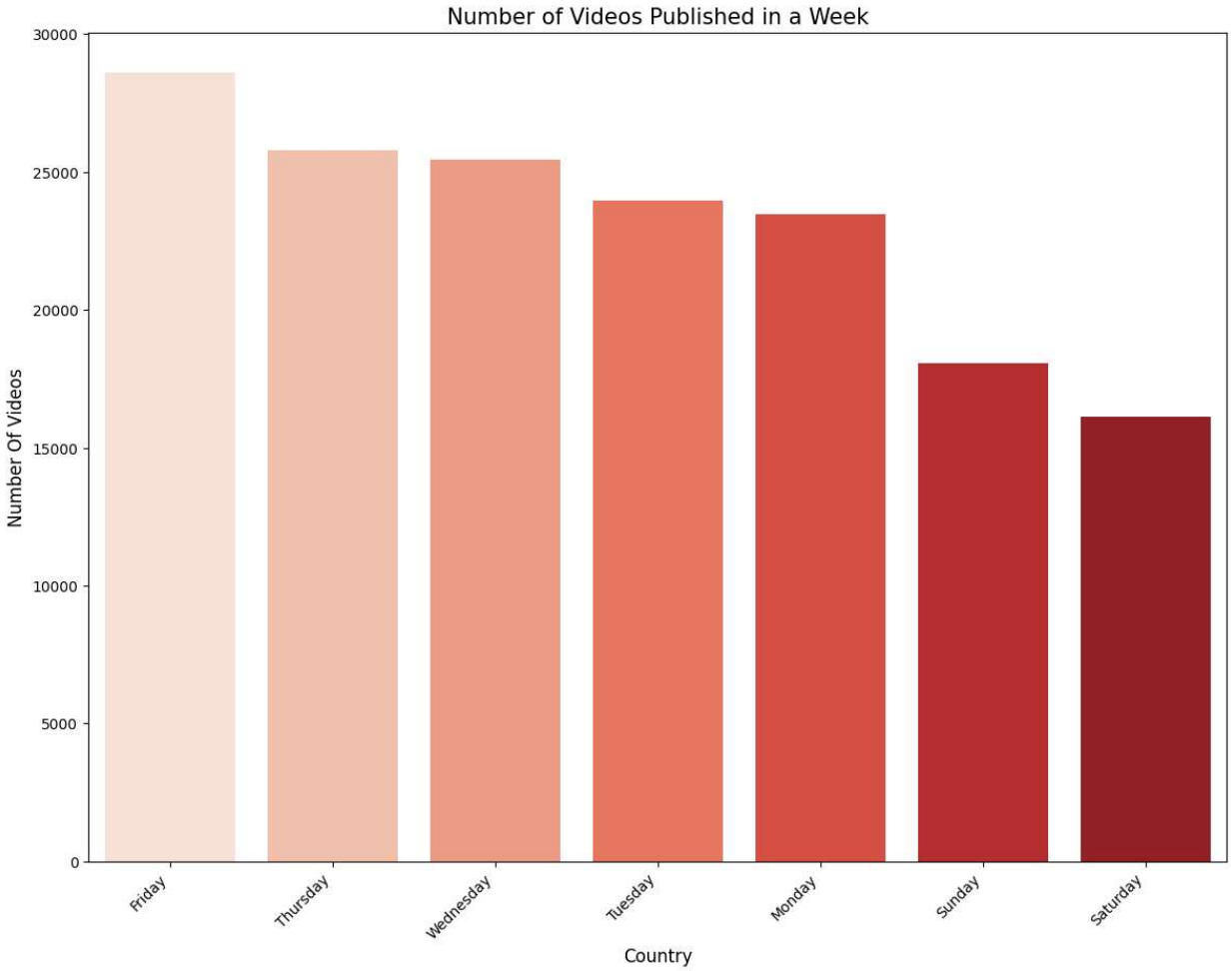
Visualizing The Count of Videos Country-wise

In [ ]:
```
plt.figure(figsize=(14, 10))
g = sns.countplot(x="publish_country", data=Youtube_df, palette="Reds", order=Youtube_
g.set_xticklabels(g.get_xticklabels(), rotation=45, ha="right")
g.set_title("Count of the Video - Country wise", fontsize=15)
g.set_xlabel(" Country", fontsize=12)
g.set_ylabel("Count", fontsize=12)
plt.subplots_adjust(wspace=0.9, hspace=0.9, top=0.9)
plt.show()
```



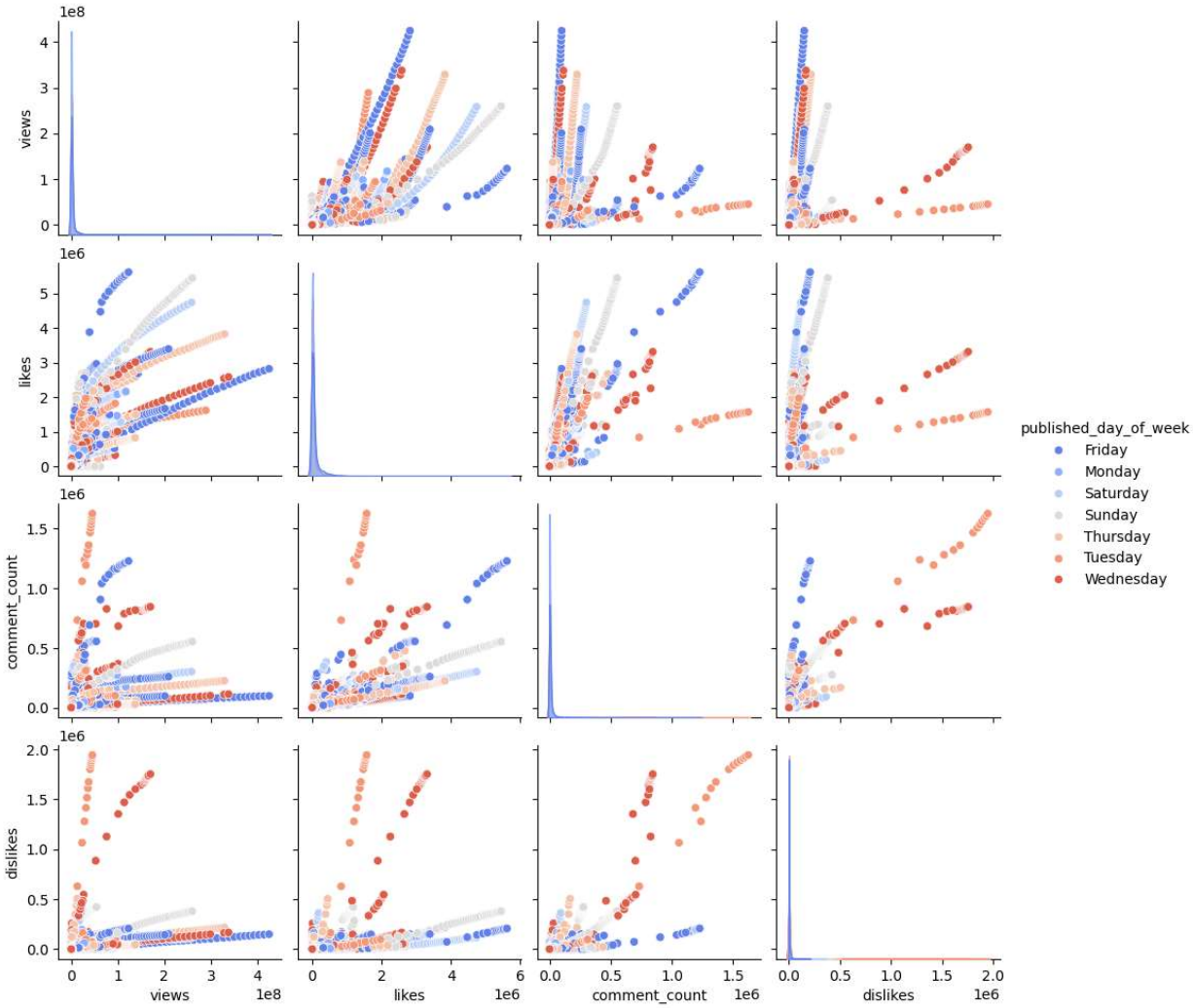
Number of Videos Published in a Week

In [ ]:
```
plt.figure(figsize=(14, 10))
g = sns.countplot(x="published_day_of_week", data=Youtube_df, palette="Reds", order=Yo
g.set_xticklabels(g.get_xticklabels(), rotation=45, ha="right")
g.set_title("Number of Videos Published in a Week", fontsize=15)
g.set_xlabel("Country", fontsize=12)
g.set_ylabel("Number Of Videos", fontsize=12)
plt.subplots_adjust(wspace=0.9, hspace=0.9, top=0.9)
plt.show()
```



Pair Plot of views,likes,comment count and dislikes with respect to published week

In [ ]:
```
data2_cols = Youtube_df[['views', 'likes', 'comment_count', 'dislikes', 'published_day
data2_cols['published_day_of_week'] = data2_cols['published_day_of_week'].astype('cate
data2_cols.dropna(subset=['published_day_of_week'], inplace=True)
sns.pairplot(data2_cols, hue='published_day_of_week', palette='coolwarm')
plt.show()
```



Identifyng the Most Used Tags with help of Word Cloud

In [ ]:
```
from wordcloud import WordCloud, STOPWORDS
plt.figure(figsize=(20, 20))

stopwords = set(STOPWORDS)
colormap = 'YlOrRd'

wordcloud = WordCloud(
    background_color='black',
    stopwords=stopwords,
    max_words=100,
    max_font_size=120,
    random_state=0,
    colormap=colormap
).generate(str(Youtube_df['tags']))

plt.imshow(wordcloud)
plt.title('WORD CLOUD for Tags', fontsize=20)
plt.axis('off')
plt.show()
```

WORD CLOUD for Tags