TRENDING YOUTUBE VIDEO ANALYTICS

In [1]:
```
import pandas as pd
import os

#LOADING THE FOLDER CONATINING DIFFERNET COUNTRY DATASETS
folder_path=(r"C:\Users\Rakshitha K B\Documents\ELEVATE LABS INTERNSHIP\DATASETS\Youtube trending video analysis

# List all the CSV files in the folder
csv_files = [f for f in os.listdir(folder_path) if f.endswith('.csv')]

# List to store dataframes
dataframes = []

# Loop through each file, read the data, and add a new column for country
for file in csv_files:

    country_name = file.split('videos')[0]
    # Create the full path to the file
    file_path = os.path.join(folder_path, file)

    df = pd.read_csv(file_path, encoding='ISO-8859-1')

    # Add a new column 'Country' for the respective country
    df['Country'] = country_name

    # Append the DataFrame to the list
    dataframes.append(df)

# Concatenate all the DataFrames into one
combined_df = pd.concat(dataframes, ignore_index=True)

# Remove leading and trailing spaces in column names
combined_df.columns = combined_df.columns.str.strip()

# Preview the result
print(combined_df.head())

# save the combined DataFrame to a new CSV file
combined_df.to_csv(r"C:\Users\Rakshitha K B\Documents\ELEVATE LABS INTERNSHIP\Project_Trending_Youtubevideos_ana
```

```
      video_id trending_date  \
0  n1WpP7iowLc      17.14.11
1  0dBIkQ4Mz1M      17.14.11
2  5qpjK5DgCt4      17.14.11
3  d380meD0W0M      17.14.11
4  2Vv-BfVoq4g      17.14.11

                                          title channel_title  \
0       Eminem - Walk On Water (Audio) ft. BeyoncÃ©    EminemVEVO
1                     PLUSH - Bad Unboxing Fan Mail     iDubbbzTV
2  Racist Superman | Rudy Mancuso, King Bach & Le...  Rudy Mancuso
3                          I Dare You: GOING BALD!?      nigahiga
4       Ed Sheeran - Perfect (Official Music Video)    Ed Sheeran

   category_id              publish_time  \
0           10  2017-11-10T17:00:03.000Z
1           23  2017-11-13T17:00:00.000Z
2           23  2017-11-12T19:05:24.000Z
3           24  2017-11-12T18:01:41.000Z
4           10  2017-11-09T11:04:14.000Z

                                               tags     views    likes  \
0  Eminem|"Walk"|"On"|"Water"|"Aftermath/Shady/In...  17158579   787425
1  plush|"bad unboxing"|"unboxing"|"fan mail"|"id...   1014651   127794
2  racist superman|"rudy"|"mancuso"|"king"|"bach"...   3191434   146035
3  ryan|"higa"|"higatv"|"nigahiga"|"i dare you"|"...   2095828   132239
4  edsheeran|"ed sheeran"|"acoustic"|"live"|"cove...  33523622  1634130

   dislikes  comment_count                                thumbnail_link  \
0     43420         125882  https://i.ytimg.com/vi/n1WpP7iowLc/default.jpg
1      1688          13030  https://i.ytimg.com/vi/0dBIkQ4Mz1M/default.jpg
2      5339           8181  https://i.ytimg.com/vi/5qpjK5DgCt4/default.jpg
3      1989          17518  https://i.ytimg.com/vi/d380meD0W0M/default.jpg
4     21082          85067  https://i.ytimg.com/vi/2Vv-BfVoq4g/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 Country
0  Eminem's new track Walk on Water ft. BeyoncÃ© ...      CA
1  STill got a lot of packages. Probably will las...      CA
2  WATCH MY PREVIOUS VIDEO â¶ \n\nSUBSCRIBE â º ...      CA
3  I know it's been a while since we did this sho...      CA
4  ð§: https://ad.gt/yt-perfect\nð°: https://...      CA
```

In [2]:
```python
#DROPPING THE COLUMNS

columns_to_remove = ['video_id ','channel_title','description','thumbnail_link','comments_disabled','ratings_di
combined_df.drop(columns=columns_to_remove, inplace=True, errors='ignore')  # errors='ignore' prevents errors i
# Preview the result
print(combined_df.head())

# save to CSV file
combined_df.to_csv(r"C:\Users\Rakshitha K B\Documents\ELEVATE LABS INTERNSHIP\Project_Trending_Youtubevideos_ana
```

```
     video_id trending_date  \
0  n1WpP7iowLc      17.14.11
1  0dBIkQ4Mz1M      17.14.11
2  5qpjK5DgCt4      17.14.11
3  d380meD0W0M      17.14.11
4  2Vv-BfVoq4g      17.14.11

                                          title  category_id  \
0       Eminem - Walk On Water (Audio) ft. BeyoncÃ©           10
1                    PLUSH - Bad Unboxing Fan Mail           23
2  Racist Superman | Rudy Mancuso, King Bach & Le...           23
3                          I Dare You: GOING BALD!?           24
4       Ed Sheeran - Perfect (Official Music Video)           10

              publish_time  \
0  2017-11-10T17:00:03.000Z
1  2017-11-13T17:00:00.000Z
2  2017-11-12T19:05:24.000Z
3  2017-11-12T18:01:41.000Z
4  2017-11-09T11:04:14.000Z

                                          tags     views    likes  \
0  Eminem|"Walk"|"On"|"Water"|"Aftermath/Shady/In...  17158579   787425
1  plush|"bad unboxing"|"unboxing"|"fan mail"|"id...   1014651   127794
2  racist superman|"rudy"|"mancuso"|"king"|"bach"...   3191434   146035
3  ryan|"higa"|"higatv"|"nigahiga"|"i dare you"|"...   2095828   132239
4  edsheeran|"ed sheeran"|"acoustic"|"live"|"cove...  33523622  1634130

   dislikes  comment_count Country
0     43420         125882      CA
1      1688          13030      CA
2      5339           8181      CA
3      1989          17518      CA
4     21082          85067      CA
```

In [3]:
```
# Dictionary mapping short codes to full country names
country_rename_map = {
    'US': 'USA',
    'GB': 'Great Britain',
    'DE': 'Germany',
    'CA': 'Canada',
    'FR': 'France',
    'RU': 'Russia',
    'MX': 'Mexico',
    'KR': 'South Korea',
    'JP': 'Japan',
    'IN': 'India'
}

# Replace short names with full country names
combined_df['Country'] = combined_df['Country'].replace(country_rename_map)

# Preview to confirm changes
print(combined_df['Country'].unique())

# save to CSV file
combined_df.to_csv(r"C:\Users\Rakshitha K B\Documents\ELEVATE LABS INTERNSHIP\Project_Trending_Youtubevideos_ana
```

```
['Canada' 'Germany' 'France' 'Great Britain' 'India' 'Japan' 'South Korea'
 'Mexico' 'Russia' 'USA']
```

In [4]:
```
combined_df.isnull().values.any()
```

Out[4]:
```
np.False_
```

In [5]:
```
combined_df.duplicated()
```

Out[5]:
```
0         False
1         False
2         False
3         False
4         False
          ...
375937    False
375938    False
375939    False
375940    False
375941    False
Length: 375942, dtype: bool
```

Sentiment analysis on titles and tags

```
```

In [6]:
```
import pandas as pd
from textblob import TextBlob

#cleaning and processing the title and tags
combined_df['text_combined'] = combined_df['title'].astype(str) + " " + combined_df['tags'].astype(str)


#Apply Sentiment Analysis Using TextBlob:
def get_sentiment(text):
    return TextBlob(text).sentiment.polarity  # ranges from -1 to +1

combined_df['sentiment_score'] = combined_df['text_combined'].apply(get_sentiment)


#labelling the sentiment

def label_sentiment(score):
    if score > 0.1:
        return 'Positive'
    elif score < -0.1:
        return 'Negative'
    else:
        return 'Neutral'

combined_df['sentiment_label'] = combined_df['sentiment_score'].apply(label_sentiment)

print(combined_df[['title', 'tags', 'sentiment_score', 'sentiment_label']].head())
```

```
                                          title  \
0       Eminem - Walk On Water (Audio) ft. BeyoncÃ©
1                      PLUSH - Bad Unboxing Fan Mail
2  Racist Superman | Rudy Mancuso, King Bach & Le...
3                          I Dare You: GOING BALD!?
4       Ed Sheeran - Perfect (Official Music Video)

                                          tags  sentiment_score  \
0  Eminem|"Walk"|"On"|"Water"|"Aftermath/Shady/In...         0.000000
1  plush|"bad unboxing"|"unboxing"|"fan mail"|"id...        -0.133333
2  racist superman|"rudy"|"mancuso"|"king"|"bach"...         0.111111
3  ryan|"higa"|"higatv"|"nigahiga"|"i dare you"|"...        -0.350000
4  edsheeran|"ed sheeran"|"acoustic"|"live"|"cove...         0.568182

  sentiment_label
0         Neutral
1        Negative
2        Positive
3        Negative
4        Positive
```

In [7]:
```
#changing date fromate

# Convert 'trending_date' from 'YY.DD.MM' to 'YYYY-MM-DD'
combined_df['trending_date'] = pd.to_datetime(
    combined_df['trending_date'],
    format='%y.%d.%m'
).dt.strftime('%Y-%m-%d')
```

In [8]:
```
combined_df.head()
```

Out[8]:

	video_id	trending_date	title	category_id	publish_time	tags	views
0	n1WpP7iowLc	2017-11-14	Eminem - Walk On Water (Audio) ft. BeyoncÃ©	10	2017-11-10T17:00:03.000Z	Eminem|"Walk"|"On"|"Water"|"Aftermath/Shady/In...	1715857
1	0dBIkQ4Mz1M	2017-11-14	PLUSH - Bad Unboxing Fan Mail	23	2017-11-13T17:00:00.000Z	plush|"bad unboxing"|"unboxing"|"fan mail"|"id...	101465
2	5qpjK5DgCt4	2017-11-14	Racist Superman | Rudy Mancuso, King Bach & Le...	23	2017-11-12T19:05:24.000Z	racist superman|"rudy"|"mancuso"|"king"|"bach"...	319143
3	d380meD0W0M	2017-11-14	I Dare You: GOING BALD!?	24	2017-11-12T18:01:41.000Z	ryan|"higa"|"higatv"|"nigahiga"|"i dare you"|"...	209582
4	2Vv-BfVoq4g	2017-11-14	Ed Sheeran - Perfect (Official Music Video)	10	2017-11-09T11:04:14.000Z	edsheeran|"ed sheeran"|"acoustic"|"live"|"cove...	3352362

In [9]:
```
#changing the formate of publishing date

combined_df['publish_time'] = pd.to_datetime(combined_df['publish_time'], errors='coerce')

# Format it as 'YYYY-MM-DD'
combined_df['publish_time'] = combined_df['publish_time'].dt.strftime('%Y-%m-%d')
```

In [10]:
```
combined_df.head()
```

Out[10]:

	video_id	trending_date	title	category_id	publish_time	tags	views	
0	n1WpP7iowLc	2017-11-14	Eminem - Walk On Water (Audio) ft. BeyoncÃ©	10	2017-11-10	Eminem|"Walk"|"On"|"Water"|"Aftermath/Shady/In...	17158579	7
1	0dBIkQ4Mz1M	2017-11-14	PLUSH - Bad Unboxing Fan Mail	23	2017-11-13	plush|"bad unboxing"|"unboxing"|"fan mail"|"id...	1014651	1
2	5qpjK5DgCt4	2017-11-14	Racist Superman | Rudy Mancuso, King Bach & Le...	23	2017-11-12	racist superman|"rudy"|"mancuso"|"king"|"bach"...	3191434	1
3	d380meD0W0M	2017-11-14	I Dare You: GOING BALD!?	24	2017-11-12	ryan|"higa"|"higatv"|"nigahiga"|"i dare you"|"...	2095828	1
4	2Vv-BfVoq4g	2017-11-14	Ed Sheeran - Perfect (Official Music Video)	10	2017-11-09	edsheeran|"ed sheeran"|"acoustic"|"live"|"cove...	33523622	16

Using SQL to rank categories by avg views

In [11]:
```
import pandas as pd
import sqlite3

# Create an in-memory SQLite database
conn = sqlite3.connect(":memory:")

# Write the DataFrame to a SQL table
combined_df.to_sql("youtube_data", conn, index=False, if_exists='replace')

# Run SQL to rank categories by average views
query = """
```

```
SELECT
    category_id,
    AVG(views) AS avg_views,
    RANK() OVER (ORDER BY AVG(views) DESC) AS rank
FROM youtube_data
GROUP BY category_id
ORDER BY rank;
"""

# Execute and get results
ranked_categories = pd.read_sql_query(query, conn)

# Show the result
print(ranked_categories)
```

```
    category_id     avg_views  rank
0            10  6.020772e+06     1
1            30  1.954438e+06     2
2             1  1.319480e+06     3
3            28  1.125286e+06     4
4            24  9.588231e+05     5
5            23  8.176072e+05     6
6            17  8.010651e+05     7
7            20  6.723543e+05     8
8            26  5.181922e+05     9
9            43  4.559184e+05    10
10           22  4.366234e+05    11
11           29  4.364434e+05    12
12           15  4.130114e+05    13
13           19  4.091638e+05    14
14           27  3.511609e+05    15
15            2  3.510464e+05    16
16           25  2.795136e+05    17
17           44  1.100860e+04    18
```

In [13]:
```
# save to CSV file
combined_df.to_csv(r"C:\Users\Rakshitha K B\Documents\ELEVATE LABS INTERNSHIP\Project_Trending_Youtubevideos_ana
```

In [14]:
```
combined_df.isnull()
```

Out[14]:

	video_id	trending_date	title	category_id	publish_time	tags	views	likes	dislikes	comment_count	Country	text_cor
0	False	False	False	False	False	False	False	False	False	False	False	
1	False	False	False	False	False	False	False	False	False	False	False	
2	False	False	False	False	False	False	False	False	False	False	False	
3	False	False	False	False	False	False	False	False	False	False	False	
4	False	False	False	False	False	False	False	False	False	False	False	
...	...	...	...	...	...	...	...	...	...	...	...	
375937	False	False	False	False	False	False	False	False	False	False	False	
375938	False	False	False	False	False	False	False	False	False	False	False	
375939	False	False	False	False	False	False	False	False	False	False	False	
375940	False	False	False	False	False	False	False	False	False	False	False	
375941	False	False	False	False	False	False	False	False	False	False	False	

375942 rows × 14 columns

In [ ]:
```
```