In [1]:

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

Load dataset

In [2]:

```
df = pd.read_csv('youtube.csv')
```

Show first few rows

In [3]:

```
df.head()
```

Out[3]:

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

In [7]:

```
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 [9]:

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

Out[9]:

```
index                     0
video_id                  0
trending_date             0
title                     0
channel_title             0
category_id               0
publish_date              0
time_frame                0
published_day_of_week     0
publish_country           0
tags                      0
views                     0
likes                     0
dislikes                  0
comment_count             0
comments_disabled         0
ratings_disabled          0
video_error_or_removed    0
dtype: int64
```

In [11]:

```
df.duplicated()
```

```
Out[11]: 0         False
         1         False
         2         False
         3         False
         4         False
                   ...
         161465    False
         161466    False
         161467    False
         161468    False
         161469    False
         Length: 161470, dtype: bool
```

```
In [13]: df.duplicated().sum()
```

```
Out[13]: 0
```

```
In [15]: df.describe()
```

Out[15]:

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

Remove duplicate rows

```
In [17]: df.drop_duplicates(inplace = True)
```

Handle missing values (fill or remove)

```
In [20]: df.dropna(subset=['title', 'channel_title'])
         df.fillna({'tags': '', 'description': ''})
```

Out[20]:

	index	video_id	trending_date	
0	0	2kyS6SvSYSE	17.14.11	WE WANT TO T
1	1	1ZAPwfrtAFY	17.14.11	The Trump Presidency
2	2	5qpjK5DgCt4	17.14.11	Racist Superman | Rudy
3	3	puqaWrEC7tY	17.14.11	Nick
4	4	d380meD0W0M	17.14.11	
...	...	...	...	
161465	161465	sGolxsMSGfQ	18.14.06	HO'
161466	161466	8HNuRNi8t70	18.14.06	Eli
161467	161467	GWIKEM3m2EE	18.14.06	KINGDOM HEARTS III â€“ SQU
161468	161468	lbMKLzQ4cNQ	18.14.06	Trump
161469	161469	POTgw38-m58	18.14.06	ã€�å®Œæ•´ç‰ˆã€‘é�‡åˆ°æ��æ€

161470 rows × 18 columns

Convert column names to lowercase and uniform style

In [24]:
```python
df.columns.str.strip().str.lower().str.replace(' ', '_')
```

Out[24]:
```
Index(['index', 'video_id', 'trending_date', 'title', 'channel_title',
       'category_id', 'publish_date', 'time_frame', 'published_day_of_week',
       'publish_country', 'tags', 'views', 'likes', 'dislikes',
       'comment_count', 'comments_disabled', 'ratings_disabled',
       'video_error_or_removed'],
      dtype='object')
```

Convert publish time to datetime if available

In [26]:
```python
if 'publish_time' in df.columns:
    df['publish_time'] = pd.to_datetime(df['publish_time'], errors='coerce')
```

Standardize country code column (if exists)

In [28]:
```python
if 'country' in df.columns:
    df['country'] = df['country'].str.upper().str.strip()
```

Remove special characters or extra spaces from text columns

```
In [30]: text_columns = ['title', 'channeltitle', 'tags']
         for col in text_columns:
             if col in df.columns:
                 df[col] = df[col].astype(str).str.replace(r'[^A-Za-z0-9\s]+', '', rege
```

Save clean data

```
In [32]: df.to_csv("youtube_cleaned.csv", index=False)

         print("\nAfter cleaning:")
         print(df.head())
         print("\nCleaned data saved as youtube_cleaned.csv ✅")
```

```
After cleaning:
   index     video_id trending_date  \
0      0  2kyS6SvSYSE      17.14.11
1      1  1ZAPwfrtAFY      17.14.11
2      2  5qpjK5DgCt4      17.14.11
3      3  puqaWrEC7tY      17.14.11
4      4  d380meD0W0M      17.14.11

                                               title          channel_title  \
0                 WE WANT TO TALK ABOUT OUR MARRIAGE           CaseyNeistat
1  The Trump Presidency Last Week Tonight with Jo...        LastWeekTonight
2  Racist Superman  Rudy Mancuso King Bach  Lele ...           Rudy Mancuso
3                     Nickelback Lyrics Real or Fake  Good Mythical Morning
4                              I Dare You GOING BALD               nigahiga

   category_id publish_date      time_frame published_day_of_week  \
0           22   13/11/2017  17:00 to 17:59                Monday
1           24   13/11/2017    7:00 to 7:59                Monday
2           23   12/11/2017  19:00 to 19:59                Sunday
3           24   13/11/2017  11:00 to 11:59                Monday
4           24   12/11/2017  18:00 to 18:59                Sunday

  publish_country                                              tags    views
\
0              US                                   SHANtell martin   748374
1              US  last week tonight trump presidencylast week to...  2418783
2              US  racist supermanrudymancusokingbachracistsuperm...  3191434
3              US  rhett and linkgmmgood mythical morningrhett an...   343168
4              US  ryanhigahigatvnigahigai dare youidyrhpcdaresno...  2095731

    likes  dislikes  comment_count  comments_disabled  ratings_disabled  \
0   57527      2966          15954              False             False
1   97185      6146          12703              False             False
2  146033      5339           8181              False             False
3   10172       666           2146              False             False
4  132235      1989          17518              False             False

   video_error_or_removed
0                   False
1                   False
2                   False
3                   False
4                   False

Cleaned data saved as youtube_cleaned.csv ✅
```

In [63]:
```python
from textblob import TextBlob
print("TextBlob installed successfully ✅")
```

```
TextBlob installed successfully ✅
```

In [69]:
```python
!pip install textblob
from textblob import download_corpora
download_corpora.download_all()
```

```
Requirement already satisfied: textblob in c:\users\bses\appdata\local\program
s\python\python313\lib\site-packages (0.19.0)
Requirement already satisfied: nltk>=3.9 in c:\users\bses\appdata\local\program
s\python\python313\lib\site-packages (from textblob) (3.9.2)
Requirement already satisfied: click in c:\users\bses\appdata\local\programs\py
thon\python313\lib\site-packages (from nltk>=3.9->textblob) (8.1.8)
Requirement already satisfied: joblib in c:\users\bses\appdata\local\programs\p
ython\python313\lib\site-packages (from nltk>=3.9->textblob) (1.5.2)
Requirement already satisfied: regex>=2021.8.3 in c:\users\bses\appdata\local\p
rograms\python\python313\lib\site-packages (from nltk>=3.9->textblob) (2025.1
0.23)
Requirement already satisfied: tqdm in c:\users\bses\appdata\local\programs\pyt
hon\python313\lib\site-packages (from nltk>=3.9->textblob) (4.67.1)
Requirement already satisfied: colorama in c:\users\bses\appdata\local\program
s\python\python313\lib\site-packages (from click->nltk>=3.9->textblob) (0.4.6)
```

```
[notice] A new release of pip is available: 25.0.1 -> 25.3
[notice] To update, run: python.exe -m pip install --upgrade pip
[nltk_data] Downloading package brown to
[nltk_data]     C:\Users\BSES\AppData\Roaming\nltk_data...
[nltk_data]   Unzipping corpora\brown.zip.
[nltk_data] Downloading package punkt_tab to
[nltk_data]     C:\Users\BSES\AppData\Roaming\nltk_data...
[nltk_data]   Unzipping tokenizers\punkt_tab.zip.
[nltk_data] Downloading package wordnet to
[nltk_data]     C:\Users\BSES\AppData\Roaming\nltk_data...
[nltk_data] Downloading package averaged_perceptron_tagger_eng to
[nltk_data]     C:\Users\BSES\AppData\Roaming\nltk_data...
[nltk_data]   Unzipping taggers\averaged_perceptron_tagger_eng.zip.
[nltk_data] Downloading package conll2000 to
[nltk_data]     C:\Users\BSES\AppData\Roaming\nltk_data...
[nltk_data]   Unzipping corpora\conll2000.zip.
[nltk_data] Downloading package movie_reviews to
[nltk_data]     C:\Users\BSES\AppData\Roaming\nltk_data...
[nltk_data]   Unzipping corpora\movie_reviews.zip.
```

In [10]:

```python
import pandas as pd
import sqlite3
from textblob import TextBlob

# 1️⃣ Load cleaned YouTube dataset
df = pd.read_csv("youtube_cleaned.csv")

# 2️⃣ Create (or connect to) SQLite database
conn = sqlite3.connect("youtube.db")

# 3️⃣ Store the DataFrame in the database
df.to_sql("youtube_data", conn, if_exists="replace", index=False)

# 4️⃣ Load only needed columns
query = "SELECT title, tags, views FROM youtube_data"
data = pd.read_sql_query(query, conn)

# 5️⃣ Define a simple sentiment function
```

```python
def get_sentiment(text):
    try:
        return TextBlob(str(text)).sentiment.polarity
    except:
        return 0.0

# 6⃣ Apply sentiment analysis
data["title_sentiment"] = data["title"].apply(get_sentiment)
data["tags_sentiment"] = data["tags"].apply(get_sentiment)

# 7⃣ Save results to database
data.to_sql("youtube_sentiment", conn, if_exists="replace", index=False)

# 8⃣ Calculate average sentiment
avg_query = """
SELECT
    ROUND(AVG(title_sentiment), 2) AS avg_title_sentiment,
    ROUND(AVG(tags_sentiment), 2) AS avg_tags_sentiment
FROM youtube_sentiment;
"""
result = pd.read_sql_query(avg_query, conn)

print("📊 Average Sentiment (Overall):")
print(result)

# 9⃣ Find Top 10 Positive & Negative Titles
top_positive = data.sort_values(by="title_sentiment", ascending=False).head(10
top_negative = data.sort_values(by="title_sentiment", ascending=True).head(10)

print("\n🌞 Top 10 Positive Titles:")
for i, row in top_positive.iterrows():
    print(f"{i+1}. {row['title']} (Sentiment: {row['title_sentiment']:.2f})")

print("\n🌧️ Top 10 Negative Titles:")
for i, row in top_negative.iterrows():
    print(f"{i+1}. {row['title']} (Sentiment: {row['title_sentiment']:.2f})")

# 🔟 Save all results
result.to_csv("avg_sentiment_overall.csv", index=False)
top_positive.to_csv("top10_positive_titles.csv", index=False)
top_negative.to_csv("top10_negative_titles.csv", index=False)

print("\n✅ Results saved as:")
print(" - avg_sentiment_overall.csv")
print(" - top10_positive_titles.csv")
print(" - top10_negative_titles.csv")

# 11⃣ Close the connection
conn.close()
```

```
📊 Average Sentiment (Overall):
   avg_title_sentiment  avg_tags_sentiment
0                 0.04                0.04

🌞 Top 10 Positive Titles:
88075. BEST OF 2017 LEGRANDJD (Sentiment: 1.00)
44122. Best Friend From Heaven  Trailer 2017 (Sentiment: 1.00)
4731. Ed Sheeran  Perfect Duet with Beyonc Official Audio (Sentiment: 1.00)
123876. CAN WE TRUST OUR BEST FRIEND (Sentiment: 1.00)
44547. Ed Sheeran  Perfect Duet with Beyonc Official Audio (Sentiment: 1.00)
123940. Londons Best Burger (Sentiment: 1.00)
123945. Perfect  Ed Sheeran Lyrics (Sentiment: 1.00)
123951. Pitch Perfect 3  RiffOff Clip HD (Sentiment: 1.00)
124025. Perfect  Ed Sheeran Lyrics (Sentiment: 1.00)
44354. John Boyega Shows Off His Best Michael Jackson Dance Moves (Sentiment: 
1.00)

🌧️ Top 10 Negative Titles:
52392. Resident Evil 7 Biohazard  Carcinogen  AGDQ 2018  In 14927  HD (Sentimen
t: -1.00)
14430. Metro Boomin Shows Off His Insane Jewelry Collection  GQ (Sentiment: 
-1.00)
145000. The Shocking Truth about Stephen Hawking (Sentiment: -1.00)
145017. Terrifying Ski Lift Malfunction Caught On Camera  NBC News (Sentiment: 
-1.00)
14291. usa gymnastics  larry nassar  i am disgusted (Sentiment: -1.00)
44985. Jesse Lingards INSANE solo Goal vs Watford (Sentiment: -1.00)
145156. Fortnite on an INSANE 20000 Gaming PC (Sentiment: -1.00)
37422. Terrible Magicians  Rudy Mancuso  Juanpa Zurita (Sentiment: -1.00)
52226. THE WORST GIFTS OF 2017 YIAY 389 (Sentiment: -1.00)
144816. This V12 Mercedes CL65 AMG Is an Insane 30000 Used Car (Sentiment: -1.0
0)

✅ Results saved as:
 - avg_sentiment_overall.csv
 - top10_positive_titles.csv
 - top10_negative_titles.csv
```

In [1]:

```python
import pandas as pd
import sqlite3

# 1⃣ Connect to your existing database
conn = sqlite3.connect("youtube.db")

# 2⃣ Check if category and views columns exist (optional safety)
data = pd.read_sql_query("SELECT * FROM youtube_data LIMIT 5;", conn)
print("Sample data:\n", data.head())

# 3⃣ SQL query to rank categories by average views
query = """
SELECT
    category,
    ROUND(AVG(views), 2) AS avg_views
FROM youtube_data
```

```
GROUP BY category
ORDER BY avg_views DESC;
"""

ranked_categories = pd.read_sql_query(query, conn)

# 4� Display results
print("\n🏆 Categories Ranked by Average Views:")
print(ranked_categories)

# 5� Save to CSV
ranked_categories.to_csv("categories_ranked_by_avg_views.csv", index=False)
print("\n✅ Results saved as 'categories_ranked_by_avg_views.csv'")

# 6� Close connection
conn.close()
```

```
Sample data:
    index     video_id trending_date  \
0      0  2kyS6SvSYSE      17.14.11   
1      1  1ZAPwfrtAFY      17.14.11   
2      2  5qpjK5DgCt4      17.14.11   
3      3  puqaWrEC7tY      17.14.11   
4      4  d380meD0W0M      17.14.11   

                                               title          channel_title  \
0                 WE WANT TO TALK ABOUT OUR MARRIAGE           CaseyNeistat   
1  The Trump Presidency Last Week Tonight with Jo...        LastWeekTonight   
2  Racist Superman  Rudy Mancuso King Bach  Lele ...           Rudy Mancuso   
3                     Nickelback Lyrics Real or Fake  Good Mythical Morning   
4                              I Dare You GOING BALD               nigahiga   

   category_id publish_date      time_frame published_day_of_week  \
0           22   13/11/2017  17:00 to 17:59                Monday   
1           24   13/11/2017    7:00 to 7:59                Monday   
2           23   12/11/2017  19:00 to 19:59                Sunday   
3           24   13/11/2017  11:00 to 11:59                Monday   
4           24   12/11/2017  18:00 to 18:59                Sunday   

  publish_country                                               tags    views  
\
0              US                                    SHANtell martin   748374   
1              US  last week tonight trump presidencylast week to...  2418783   
2              US  racist supermanrudymancusokingbachracistsuperm...  3191434   
3              US  rhett and linkgmmgood mythical morningrhett an...   343168   
4              US  ryanhigahigatvnigahigai dare youidyrhpcdaresno...  2095731   

    likes  dislikes  comment_count  comments_disabled  ratings_disabled  \
0   57527      2966          15954                  0                 0   
1   97185      6146          12703                  0                 0   
2  146033      5339           8181                  0                 0   
3   10172       666           2146                  0                 0   
4  132235      1989          17518                  0                 0   

   video_error_or_removed  
0                       0  
1                       0  
2                       0  
3                       0  
4                       0  
```

```
---------------------------------------------------------------------------
OperationalError                          Traceback (most recent call last)
File ~\anaconda3\Lib\site-packages\pandas\io\sql.py:2674, in SQLiteDatabase.exe
cute(self, sql, params)
   2673 try:
-> 2674     cur.execute(sql, *args)
   2675     return cur

OperationalError: no such column: category

The above exception was the direct cause of the following exception:

DatabaseError                             Traceback (most recent call last)
Cell In[1], line 21
     11 # 3️⃣ SQL query to rank categories by average views
     12 query = """
     13 SELECT
     14     category,
   (...)
     18 ORDER BY avg_views DESC;
     19 """
---> 21 ranked_categories = pd.read_sql_query(query, conn)
     23 # 4️⃣ Display results
     24 print("\n🏆 Categories Ranked by Average Views:")

File ~\anaconda3\Lib\site-packages\pandas\io\sql.py:526, in read_sql_query(sql,
con, index_col, coerce_float, params, parse_dates, chunksize, dtype, dtype_back
end)
    523 assert dtype_backend is not lib.no_default
    525 with pandasSQL_builder(con) as pandas_sql:
--> 526     return pandas_sql.read_query(
    527         sql,
    528         index_col=index_col,
    529         params=params,
    530         coerce_float=coerce_float,
    531         parse_dates=parse_dates,
    532         chunksize=chunksize,
    533         dtype=dtype,
    534         dtype_backend=dtype_backend,
    535     )

File ~\anaconda3\Lib\site-packages\pandas\io\sql.py:2738, in SQLiteDatabase.rea
d_query(self, sql, index_col, coerce_float, parse_dates, params, chunksize, dty
pe, dtype_backend)
   2727 def read_query(
   2728     self,
   2729     sql,
   (...)
   2736     dtype_backend: DtypeBackend | Literal["numpy"] = "numpy",
   2737 ) -> DataFrame | Iterator[DataFrame]:
-> 2738     cursor = self.execute(sql, params)
   2739     columns = [col_desc[0] for col_desc in cursor.description]
   2741     if chunksize is not None:
```

```
File ~\anaconda3\Lib\site-packages\pandas\io\sql.py:2686, in SQLiteDatabase.execute(self, sql, params)
   2683     raise ex from inner_exc
   2685 ex = DatabaseError(f"Execution failed on sql '{sql}': {exc}")
-> 2686 raise ex from exc

DatabaseError: Execution failed on sql '
SELECT
    category,
    ROUND(AVG(views), 2) AS avg_views
FROM youtube_data
GROUP BY category
ORDER BY avg_views DESC;
': no such column: category
```

In [ ]: