In [9]:
```
import zipfile
import pandas as pd
import os

zip_files = ['USvideos.csv.zip', 'GBvideos.csv.zip', 'FRvideos.csv.zip', 'DEvide

all_dfs = []

for zip_file in zip_files:
    region = zip_file[:2]
    with zipfile.ZipFile(zip_file, 'r') as z:
        file_name = z.namelist()[0]
        with z.open(file_name) as f:
            df = pd.read_csv(f)
            df['region'] = region
            all_dfs.append(df)

combined_df = pd.concat(all_dfs, ignore_index=True)

print(combined_df['region'].value_counts())
combined_df.head()
```

```
region
US    40949
CA    40881
DE    40840
FR    40724
GB    38916
Name: count, dtype: int64
```

Out[9]:

	video_id	trending_date	title	channel_title	category_id	publish_
0	2kyS6SvSYSE	17.14.11	WE WANT TO TALK ABOUT OUR MARRIAGE	CaseyNeistat	22	2017 13T17:13:01.(
1	1ZAPwfrtAFY	17.14.11	The Trump Presidency: Last Week Tonight with J...	LastWeekTonight	24	2017 13T07:30:00.(
2	5qpjK5DgCt4	17.14.11	Racist Superman | Rudy Mancuso, King Bach & Le...	Rudy Mancuso	23	2017 12T19:05:24.(
3	puqaWrEC7tY	17.14.11	Nickelback Lyrics: Real or Fake?	Good Mythical Morning	24	2017 13T11:00:04.(
4	d380meD0W0M	17.14.11	I Dare You: GOING BALD!?	nigahiga	24	2017 12T18:01:41.(

In [11]:
```
combined_df['trending_date'] = pd.to_datetime(combined_df['trending_date'], form
combined_df['publish_time'] = pd.to_datetime(combined_df['publish_time'])
```

In [16]:
```
from textblob import TextBlob

combined_df['title_sentiment'] = combined_df['title'].apply(lambda x: TextBlob(s
combined_df['tags_sentiment'] = combined_df['tags'].apply(lambda x: TextBlob(str
def classify_sentiment(p):
    if p > 0.1:
        return 'Positive'
    elif p < -0.1:
        return 'Negative'
    else:
        return 'Neutral'

combined_df['title_sentiment_label'] = combined_df['title_sentiment'].apply(clas
```

In [13]:
```
missing_summary = combined_df.isnull().sum()
print("Missing values per column:\n", missing_summary)

combined_df[combined_df.isnull().any(axis=1)]
```

```
Missing values per column:
 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                6942
region                        0
title_sentiment               0
tags_sentiment                0
title_sentiment_label         0
dtype: int64
```

Out[13]:

	video_id	trending_date	title	channel_title	category_id	publ
42	NZFhMSgbKKM	2017-11-14	Dennis Smith Jr. and LeBron James go back and ...	Ben Rohrbach	17	201 15:11:0
47	sbcbvuitiTc	2017-11-14	Stephon Marbury and Jimmer Fredette fight in C...	NBA Highlights · YouTube	17	201 18:23:0
175	4d07RXYLsJE	2017-11-14	Sphaera - demonstrating interaction	Jenny Hanell	28	201 20:48:1
267	NZFhMSgbKKM	2017-11-15	Dennis Smith Jr. and LeBron James go back and ...	Ben Rohrbach	17	201 15:11:0
312	sbcbvuitiTc	2017-11-15	Stephon Marbury and Jimmer Fredette fight in C...	NBA Highlights · YouTube	17	201 18:23:0
...	...	...	...	...	...	
202222	xq9PVtS2TYo	2018-06-14	Taylor Hatala , Josh Beauchamp , world of danc...	All things Hatala/FAN ACCOUNT	10	201 01:56:3
202243	z9BDS6s5Cw4	2018-06-14	BTS (방탄소년단) - JIMIN &JUNGKOOK - Black Or Whit...	2018 BTS FESTA	22	201 13:36:3
202275	crJ4yqYQ_qw	2018-06-14	My F ing Tourettes Family 2018 720p	Eternally profound	24	201 22:34:3
202288	ccYF1Tv301g	2018-06-14	SHAHRZAD SERIES - Season 3 - Episode 16 || سر...	Tasvir Gostar Pasargad	24	201 03:30:0
202301	uTGsMan-6Ss	2018-06-14	NCT 미니게임 천국 #3: 최강 손가락 컨트롤러 (Professional Finge...	SMTOWN	10	201 10:00:0

6942 rows × 20 columns

In [14]:
```
print("Data types:\n", combined_df.dtypes)

for col in combined_df.columns:
    unique_types = combined_df[col].map(type).nunique()
    if unique_types > 1:
        print(f"Inconsistent types in column: {col}")
```

```
Data types:
 video_id                                object
trending_date                 datetime64[ns]
title                                   object
channel_title                           object
category_id                              int64
publish_time              datetime64[ns, UTC]
tags                                    object
views                                    int64
likes                                    int64
dislikes                                 int64
comment_count                            int64
thumbnail_link                          object
comments_disabled                         bool
ratings_disabled                          bool
video_error_or_removed                    bool
description                             object
region                                  object
title_sentiment                        float64
tags_sentiment                         float64
title_sentiment_label                   object
dtype: object
Inconsistent types in column: description
```

In [15]:
```
text_cols = ['title', 'tags', 'description']

for col in text_cols:
    try:
        combined_df[col].astype(str).apply(lambda x: x.encode('utf-8').decode('u
    except UnicodeDecodeError as e:
        print(f"Encoding issue in column: {col}", e)
```

In [21]:
```
import sqlite3

conn = sqlite3.connect('youtube_trending.db')  # Creates file if it doesn't exis

combined_df.to_sql('trending_videos', conn, if_exists='replace', index=False)

test_query = pd.read_sql_query("SELECT region, COUNT(*) as count FROM trending_v
print(test_query)
```

```
  region  count
0     CA  40881
1     DE  40840
2     FR  40724
3     GB  38916
4     US  40949
```

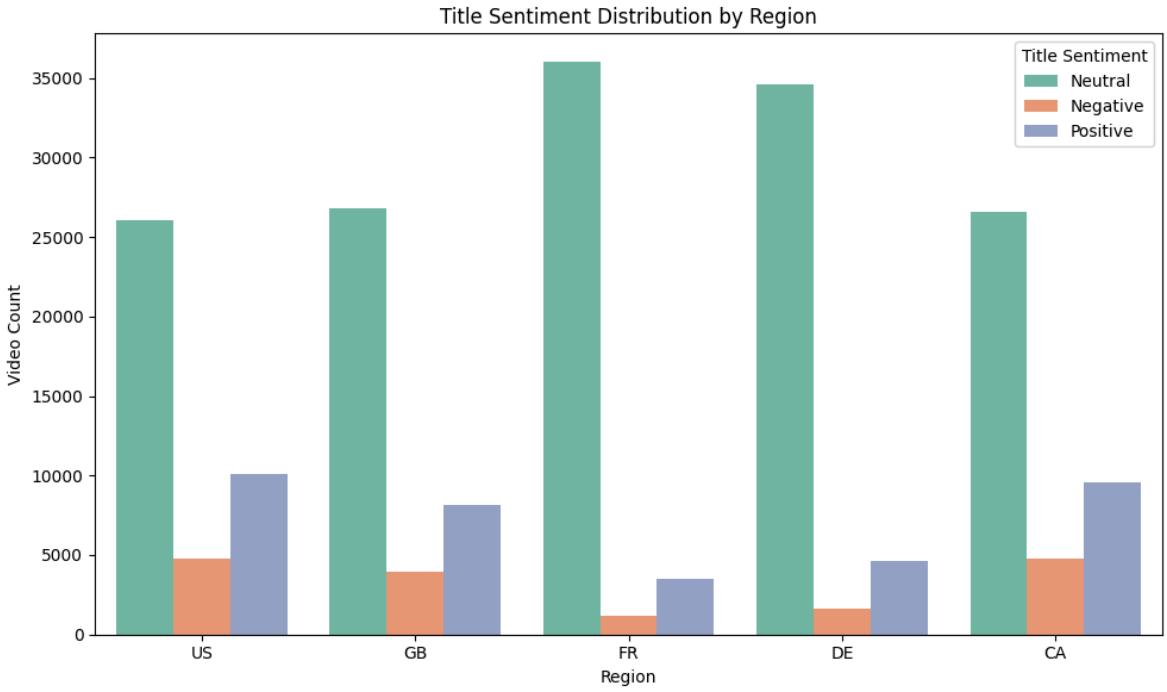
In [22]:
```
pd.read_sql_query("SELECT category_id, AVG(views) as avg_views FROM trending_vid
```

Out[22]:

	category_id	avg_views
0	10	7.768255e+06
1	1	1.846429e+06
2	29	1.448382e+06
3	28	1.335567e+06
4	24	1.317329e+06
5	30	1.116827e+06
6	17	1.051343e+06
7	23	1.024560e+06
8	20	8.957245e+05
9	22	7.916058e+05
10	15	7.256611e+05
11	26	6.480032e+05
12	43	6.248203e+05
13	19	5.643929e+05
14	27	5.133146e+05
15	2	4.782935e+05
16	25	4.010503e+05
17	44	7.248000e+03

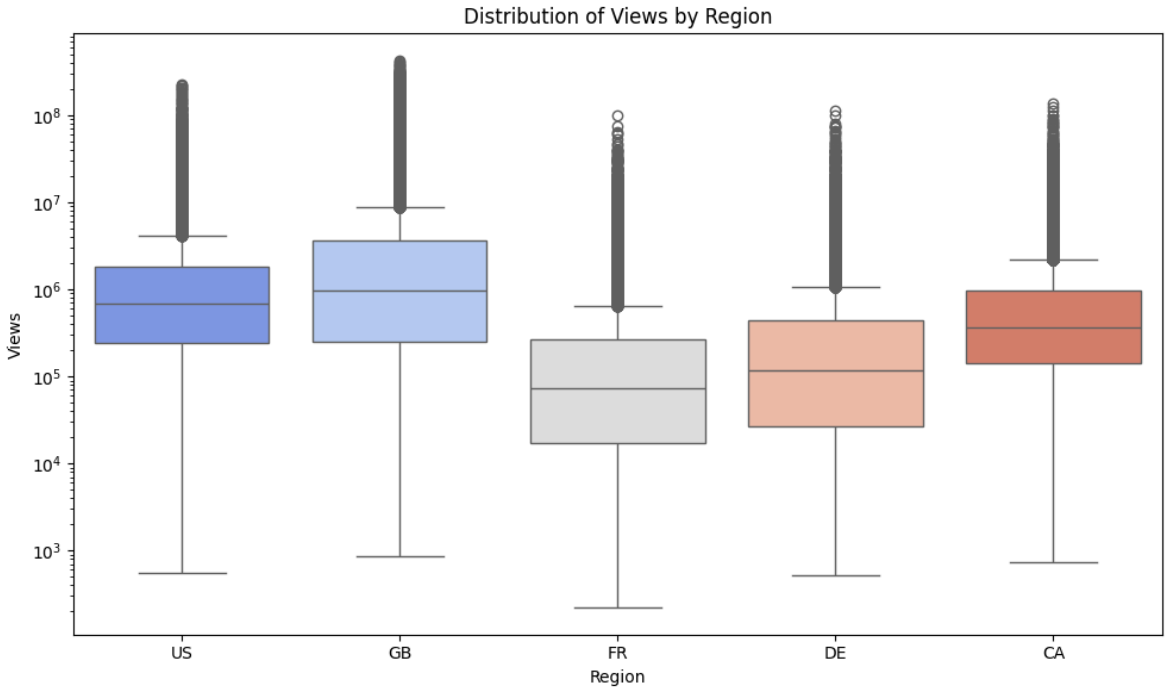
In [25]:

```
import matplotlib.pyplot as plt
import seaborn as sns

plt.figure(figsize=(10, 6))
sns.countplot(data=combined_df, x='region', hue='title_sentiment_label', palette
plt.title("Title Sentiment Distribution by Region")
plt.xlabel("Region")
plt.ylabel("Video Count")
plt.legend(title="Title Sentiment")
plt.tight_layout()
plt.show()
```



In [26]:
```
plt.figure(figsize=(10, 6))
sns.boxplot(data=combined_df, x='region', y='views', palette='coolwarm')
plt.title("Distribution of Views by Region")
plt.xlabel("Region")
plt.ylabel("Views")
plt.yscale('log')
plt.tight_layout()
plt.show()
```

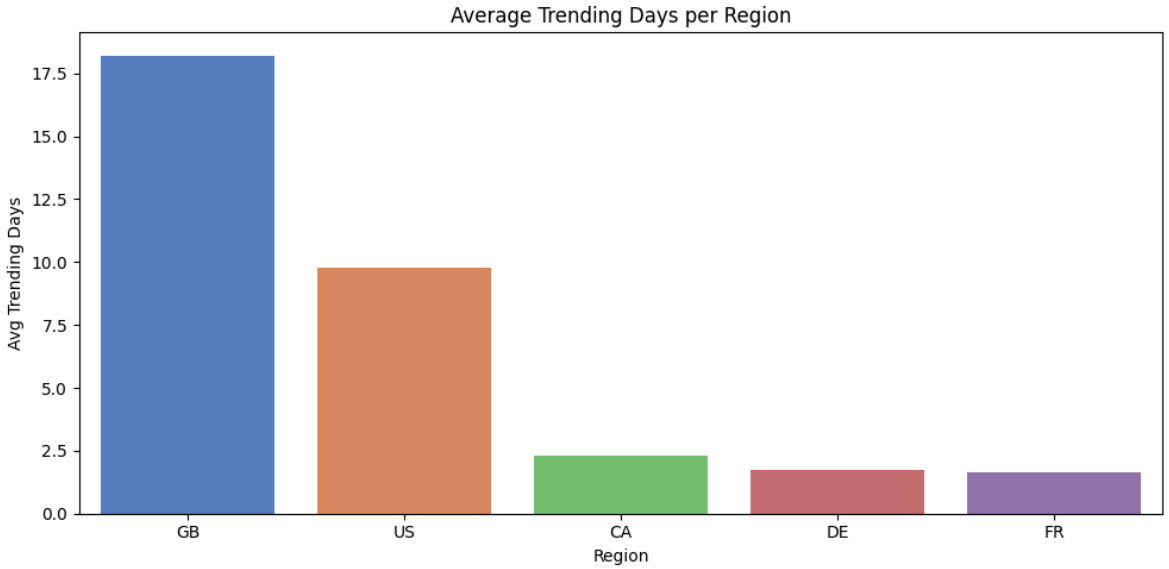
```
/tmp/ipython-input-26-1364954578.py:2: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v
0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the same effe
ct.

  sns.boxplot(data=combined_df, x='region', y='views', palette='coolwarm')
```



In [36]:
```
trending_duration = (
    combined_df.groupby(['video_id', 'region'])['trending_date_x']
    .nunique()
    .reset_index(name='trending_days1')
)
combined_df = pd.merge(combined_df, trending_duration, on=['video_id', 'region']

avg_trend_days = (
    combined_df.groupby('region')['trending_days1']
    .mean()
    .sort_values(ascending=False)
)
```

In [40]:
```
plt.figure(figsize=(10, 5))
sns.barplot(x=avg_trend_days.index, y=avg_trend_days.values, palette='muted')
plt.title("Average Trending Days per Region")
plt.xlabel("Region")
plt.ylabel("Avg Trending Days")
plt.tight_layout()
plt.show()
```

```
/tmp/ipython-input-40-3175453495.py:2: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v
0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the same effe
ct.

  sns.barplot(x=avg_trend_days.index, y=avg_trend_days.values, palette='muted')
```



In [38]:
```
combined_df.to_csv('youtube_cleaned.csv', index=False)
```