import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from wordcloud import WordCloud

# 读取三个CSV文件

us\_data = pd.read\_csv(r"C:\Users\22183\Desktop\USvideos.csv", encoding='ISO-8859-1')

jp\_data = pd.read\_csv(r"C:\Users\22183\Desktop\JPvideos.csv", encoding='ISO-8859-1')

gb\_data = pd.read\_csv(r"C:\Users\22183\Desktop\GBvideos.csv", encoding='ISO-8859-1')

# 添加国家列，并合并成一个DataFrame

us\_data['country'] = 'US'

jp\_data['country'] = 'JP'

gb\_data['country'] = 'GB'

data = pd.concat([us\_data, jp\_data, gb\_data], ignore\_index=True)

# 统计不同国家的视频类别数量

category\_count = data.groupby(['country', 'category\_id']).size().unstack(fill\_value=0)

# 绘制柱状图

plt.figure(figsize=(12, 6))

sns.set\_style("whitegrid")

category\_count.plot(kind='bar', stacked=True, width=0.9)

plt.title("Video Category Count by Country")

plt.xlabel("Category ID")

plt.ylabel("Count")

plt.legend(loc='center left', bbox\_to\_anchor=(1.0, 0.5))

plt.xticks(rotation=45, ha='right') # 调整 x 轴标签旋转角度和对齐方式

plt.tight\_layout()

plt.show()

# 统计不同国家的视频点赞数、踩数平均值

likes\_dislikes = data.groupby('country')[['likes', 'dislikes']].mean()

# 绘制饼图

likes\_dislikes.T.plot(kind='pie', subplots=True, figsize=(15, 5), autopct='%1.1f%%')

plt.suptitle("Average Likes and Dislikes by Country")

plt.tight\_layout() # 调整布局，避免标签遮挡

plt.show()

# 统计不同国家的视频评论数中位数

comments = data.groupby('country')['comment\_count'].median()

# 绘制饼图

comments.plot(kind='pie', labels=['US', 'JP', 'GB'], autopct='%1.1f%%')

plt.axis('equal')

plt.title("Median Comment Count by Country")

plt.tight\_layout() # 调整布局，避免标签遮挡

plt.show()

# 统计不同国家视频的发布时间（年份）分布

year\_count = data.groupby(['country', data['trending\_date'].str[:4]])['video\_id'].count().unstack(fill\_value=0)

# 绘制条形图，并将X轴标签旋转45度

year\_count.plot(kind='bar', figsize=(10, 6))

plt.title('Video Publish Year Distribution by Country')

plt.xlabel('Year')

plt.ylabel('Count')

plt.legend(loc='upper right') # 调整图例位置

plt.xticks(rotation=45, ha='right') # 调整 x 轴标签旋转角度和对齐方式

plt.tight\_layout() # 调整布局，避免标签遮挡

plt.show()

# 统计不同国家视频的标题长度分布

title\_len = data.groupby('country')['title'].apply(lambda x: x.str.len())

# 绘制箱线图

title\_len.plot(kind='box')

plt.title("Video Title Length Distribution by Country")

plt.ylabel("Title Length")

plt.tight\_layout() # 调整布局，避免标签遮挡

plt.show()

# 绘制标签词云图

tags\_text = ' '.join(data['tags'].str.lower().str.replace('|', ' '))

wordcloud = WordCloud(width=800, height=400, background\_color='white').generate(tags\_text)

plt.figure(figsize=(10, 5))

plt.imshow(wordcloud, interpolation='bilinear')

plt.axis('off')

plt.title('Tag Word Cloud')

plt.tight\_layout() # 调整布局，避免标签遮挡

plt.show()

# 统计不同国家视频的标签数量分布

tag\_count = data.groupby('country')['tags'].apply(lambda x: x.str.count('|')+1)

# 绘制密度图

tag\_count.plot(kind='kde')

plt.title("Video Tag Count Distribution by Country")

plt.xlabel("Tag Count")

plt.ylabel("Density")

plt.tight\_layout() # 调整布局，避免标签遮挡

plt.show()

# 绘制频道类别分布图

channel\_categories = data['channel\_title'].value\_counts().head(10)

channel\_categories.plot(kind='barh')

plt.title('Top 10 Channel Categories')

plt.xlabel('Count')

plt.ylabel('Channel Category')

plt.tight\_layout() # 调整布局，避免标签遮挡

plt.show()