Netflix Movies and TV Shows: Deeper Analysis in Python

In this notebook, we'll explore how Netflix's content varies across countries and genres using a global dataset.
We'll also use interactive visualizations to see how genre diversity differs between countries.

In [2]:
```
import pandas as pd
import plotly.express as px
```

Step 1: Load and Preview the Data

Let's import the dataset and take a quick look

In [3]:
```
df = pd.read_csv("netflix.csv")
df.head()
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8807 entries, 0 to 8806
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   show_id       8807 non-null   object
 1   type          8807 non-null   object
 2   title         8807 non-null   object
 3   director      6173 non-null   object
 4   cast          7982 non-null   object
 5   country       7976 non-null   object
 6   date_added    8797 non-null   object
 7   release_year  8807 non-null   int64 
 8   rating        8803 non-null   object
 9   duration      8804 non-null   object
 10  listed_in     8807 non-null   object
 11  description   8807 non-null   object
dtypes: int64(1), object(11)
memory usage: 825.8+ KB
```

Analysis

This dataset includes thousands of Netflix titles with metadata like release year, country, content type, and genres. Previewing the data allows us to check for missing values and understand the structure. We notice several null entries, especially in columns like director, cast, and country, which we'll need to address during cleaning.

Step 2: Clean the Data

We'll handle missing values and convert date columns to proper formats. We'll remove rows missing key information and create a `year_added` column to help with time-based analysis.

In [4]:
```
df = df.dropna(subset=["country", "date_added", "listed_in"])
df["date_added"] = pd.to_datetime(df["date_added"])
df["year_added"] = df["date_added"].dt.year
df["country"] = df["country"].str.strip()
df["listed_in"] = df["listed_in"].str.strip()
```

Analysis

Cleaning the dataset is critical for reliable analysis. We remove rows missing country, date_added, or listed_in since these are key to understanding content distribution, time trends, and genre variety. We also convert date_added to a datetime format and extract the year for time-based visualizations. Finally, trimming whitespace ensures values are consistent (e.g., no duplicate countries due to spacing).

Step 3. Explode the Country Column

Some titles are associated with more than one country.
We split these into separate rows so each title-country pair is counted individually.

In [5]:
```
df_exp = df.copy()
df_exp["country"] = df_exp["country"].str.split(", ")
df_exp = df_exp.explode("country")
```

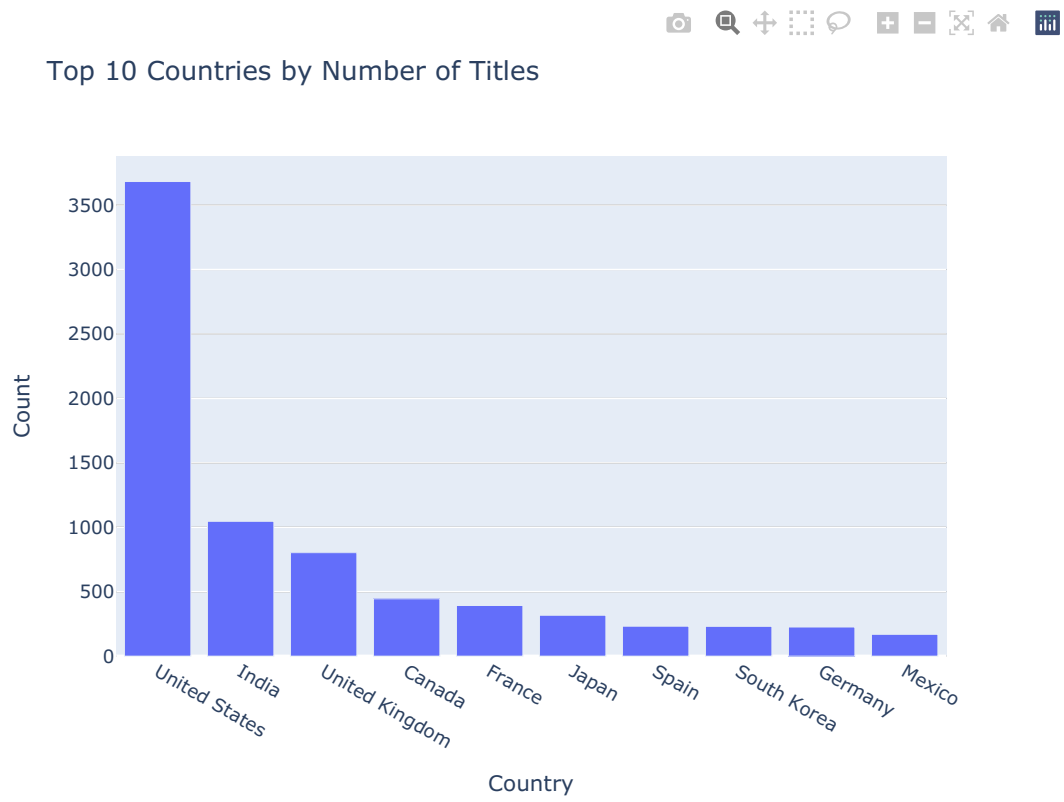
Analysis

Many titles are listed under multiple countries, especially in co-productions or shared distribution rights. By splitting these into individual rows, we avoid undercounting shared content and get a true sense of how geographically distributed Netflix's catalog is. This transformation is essential for accurate country-level insights later on.

Step 4. Identify Top 10 Countries

We'll now find the countries that appear most often in the dataset.

In [6]:
```
top_countries = df_exp["country"].value_counts().nlargest(10).reset_index()
top_countries.columns = ["Country", "Count"]
top_countries
px.bar(top_countries, x='Country', y='Count', title='Top 10 Countries by Number of Titles')
```



Analysis

This bar chart highlights the countries most represented in Netflix's catalog. Unsurprisingly, the U.S. leads with a large margin due to its vast entertainment industry and Netflix's home base. India, the UK, and South Korea follow — reflecting growing investments in Bollywood, K-dramas, and international licensing. These figures also suggest Netflix's prioritization of certain markets.

Step 5. Explode the Genre Column

Like countries, titles often belong to multiple genres.
We'll separate them out so we can analyze genre frequency by country.

In [7]:
```
df_genre = df_exp.copy()
df_genre["genre"] = df_genre["listed_in"].str.split(", ")
df_genre = df_genre.explode("genre")
```

Analysis

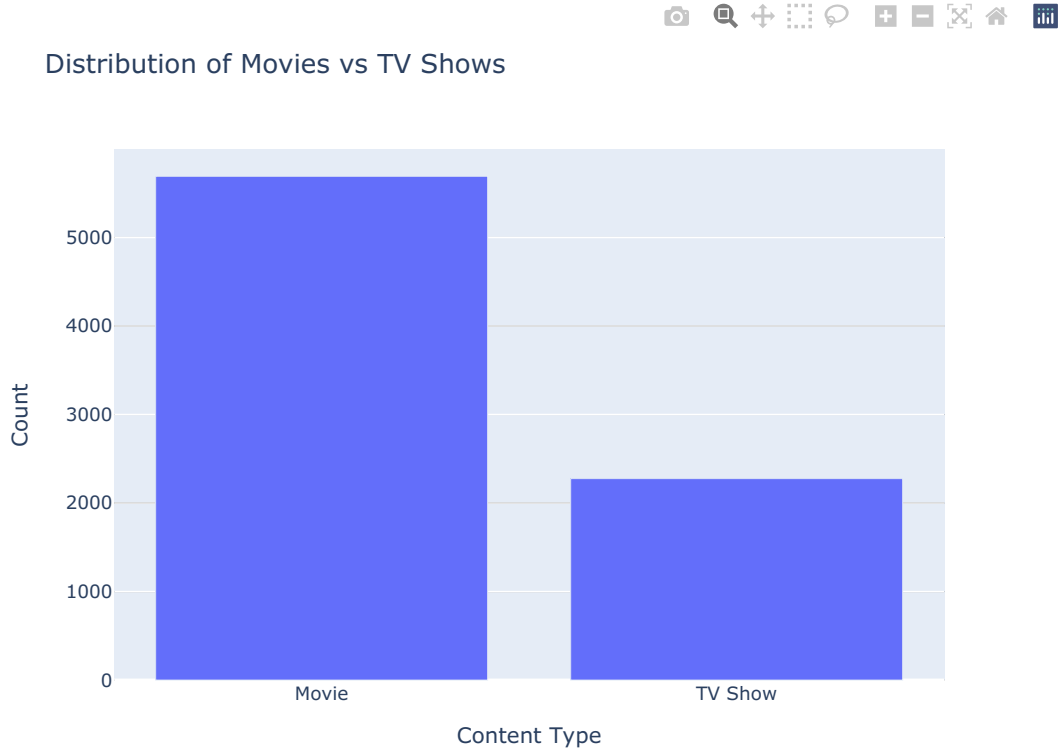
Just as with countries, many titles span multiple genres. A show might be labeled both "Drama" and "Romantic," or a documentary might also fall under "Politics." Exploding the genre column helps us treat each genre equally in our analysis, which is especially useful when analyzing preferences or diversity by country.

Step 6: Movies vs TV Shows

This chart shows the overall distribution of Movies and TV Shows on Netflix.

In [8]:
```
type_counts = df['type'].value_counts().reset_index()
type_counts.columns = ['Content Type', 'Count']
```

```
px.bar(type_counts, x='Content Type', y='Count', title='Distribution of Movies vs TV Shows')
```



Analysis

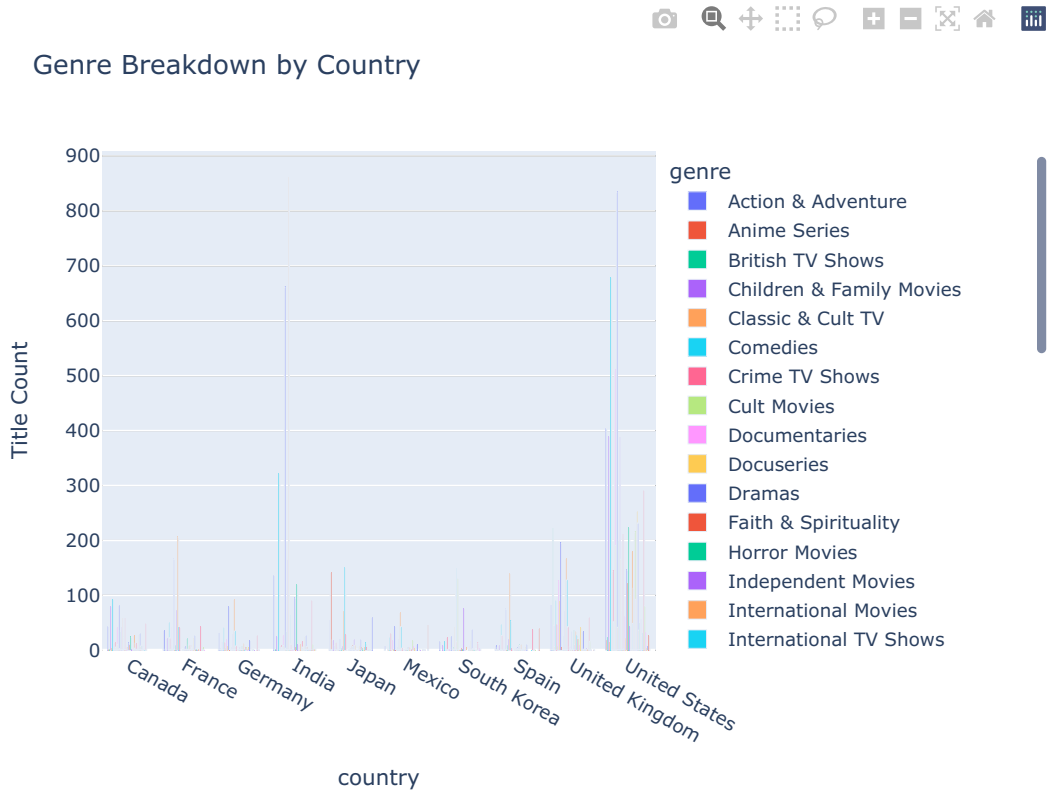
This plot shows the breakdown between movies and TV shows in the catalog. While movies dominate in quantity, TV shows have been increasingly emphasized by Netflix due to their binge-worthy appeal and subscriber retention potential. Over time, Netflix's investments in long-form series across various languages have also contributed to this shift.

Step 7. Group Genre Counts by Country

We'll group our exploded data to count how many times each genre appears in each country. Then we filter to just the top 10 countries to keep the visualization clear.

In [14]:
```
genre_counts = df_genre.groupby(["country", "genre"]).size().reset_index(name="count")
top_genre_counts = genre_counts[genre_counts["country"].isin(top_countries["Country"])]
top_10 = df_exp['country'].value_counts().nlargest(10).index
filtered_genre_counts = genre_counts[genre_counts['country'].isin(top_10)]

# Plot grouped bar chart
fig = px.bar(filtered_genre_counts,
             x='country',
             y='count',
             color='genre',
             title='Genre Breakdown by Country',
             labels={'count': 'Title Count'},
             barmode='group')
fig.show()
```



Analysis

Grouping genre counts by country lets us examine content tendencies. For example, romantic comedies may be more prevalent in India, while South Korea might skew toward drama and thrillers. These trends reflect regional audience tastes and production cultures, and help Netflix optimize content recommendations and investments per market.

Step 10: Diversity by Genre and Country (New Angle)

Why it matters: Are certain genres more international than others? How we'll do it:

Explode both country and listed_in columns

Bar chart: Top genres per country or vice versa

In [12]:
```
df_genre = df_exp.copy()
df_genre['genre'] = df_genre['listed_in'].str.split(', ')
df_genre = df_genre.explode('genre')

genre_counts = df_genre.groupby(['country', 'genre']).size().reset_index(name='count')
top = genre_counts[genre_counts['country'].isin(top_countries['Country'])]

px.sunburst(top, path=['country', 'genre'], values='count',
            title='Genres by Country (Sunburst View)')
```

Analysis

The sunburst visualization allows us to explore genre distribution hierarchically — starting from countries down to specific genres. Countries with wide genre variation like the U.S. and UK show Netflix's full catalog spread, while more niche-producing countries might focus on a few categories. This helps evaluate how balanced or specialized a country's Netflix offering is.

10. Country Content Growth Over Time

This line chart shows how Netflix has increased the number of titles from top countries year-over-year.

In [13]:
```
# Filter to top countries
df_years = df_exp[df_exp["country"].isin(top_countries["Country"])]

# Count titles added by year per country
country_growth = df_years.groupby(["year_added", "country"]).size().reset_index(name="count")

# Plot
px.line(country_growth, x="year_added", y="count", color="country",
        title="Netflix Content Growth by Country Over Time",
        labels={"year_added": "Year", "count": "Titles Added"})
```

Analysis

This line chart illustrates how content from each country has grown on Netflix year by year. After around 2015, there's a noticeable surge in titles from all countries, especially the U.S., India, and South Korea. This aligns with Netflix's global expansion, where they invested in local productions to attract regional audiences while also making them available internationally.

Conclusion

Netflix's catalog has grown significantly in both volume and global diversity. While the U.S. still dominates, countries like India and South Korea are contributing more each year. Genre trends reflect both global favorites and regional preferences. Overall, Netflix is clearly evolving into a more internationally balanced platform that serves a wide range of audiences.

In [ ]: