NETFLIX

Netflix Data Analysis

Cleaning, Analysis and Visualization

Koki venu gopal reddy

About Dataset:

Netflix is a popular streaming service that offers a vast catalog of movies, TV shows, and original contents. This dataset is a cleaned version of the original version which can be found here. The data consist of contents added to Netflix from 2008 to 2021. The oldest content is as old as 1925 and the newest as 2021.This dataset will be cleaned with PostgreSQL and visualized with Tableau. The purpose of this dataset is to test my data cleaning and visualization skills. The cleaned data can be found below and the Tableau dashboard can be found here.

https://drive.google.com/file/d/1cWcK8cddROe_DSv5zH5Fk7od32tK3ftf/view

This project involves loading, cleaning, analyzing, and visualizing data from a Netflix dataset. We'll use Python libraries like Pandas, Matplotlib, and Seaborn to work through the project. The goal is to explore the dataset, derive insights, and prepare for potential machine learning tasks.

Data Cleaning:

We are going to:

- ✓ Treat the Nulls
- ✓ Treat the duplicates
- ✓ Populate missing rows
- ✓ Drop unneeded columns
- ✓ Split columns

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

data = pd.read_csv('/content/drive/MyDrive/unified projects/netlfix/netflix.csv')

data.head(10)
```

	show_id	type	title	director	country	date_added	release_year	rating	duration	listed_in
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	United States	25-09-2021	2020	PG-13	90 min	Documentaries
1	s3	TV Show	Ganglands	Julien Leclercq	France	24-09-2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...
2	s6	TV Show	Midnight Mass	Mike Flanagan	United States	24-09-2021	2021	TV-MA	1 Season	TV Dramas, TV Horror, TV Mysteries
3	s14	Movie	Confessions of an Invisible Girl	Bruno Garotti	Brazil	22-09-2021	2021	TV-PG	91 min	Children & Family Movies, Comedies
4	s8	Movie	Sankofa	Haile Gerima	United States	24-09-2021	1993	TV-MA	125 min	Dramas, Independent Movies, International Movies
5	s9	TV	The Great British	Andy	United	24-09-2021	2021	TV-14	9	British TV Shows,

Next steps: Generate code with data | View recommended plots | New interactive sheet

```
data.shape
```

```
(8790, 10)
```

```
data.info()          #checking null values
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8790 entries, 0 to 8789 Data
columns (total 10 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   show_id       8790 non-null   object
 1   type          8790 non-null   object
 2   title         8790 non-null   object
 3   director      8790 non-null   object
 4   country       8790 non-null   object
 5   date_added    8790 non-null   object
 6   release_year  8790 non-null   int64
 7   rating        8790 non-null   object
 8   duration      8790 non-null   object
 9   listed_in     8790 non-null   object
dtypes: int64(1), object(9) memory
usage: 686.8+ KB
```

```
# Convert 'date_added' to a standard datetime format
data['date_added'] = pd.to_datetime(data['date_added'], errors='coerce')
# Preview the standardized date column
print(data[['date_added']].head())
```

```
        date_added
0 2021-09-25
1 2021-09-24
2 2021-09-24
3 2021-09-22
4 2021-09-24
<ipython-input-43-0211676aedec>:2: UserWarning: Parsing dates in %d-%m-%Y format when dayfirst=False (the default) was specified. P
data['date_added'] = pd.to_datetime(data['date_added'], errors='coerce')
```

```
# Check for rows with invalid dates
invalid_dates = data[data['date_added'].isna()] print("Invalid Dates:", invalid_dates)
```

```
Invalid Dates: Empty DataFrame
Columns: [show_id, type, title, director, country, date_added, release_year, rating, duration, listed_in] Index: []
```

```
data = data.drop_duplicates() #droping duplicates
```

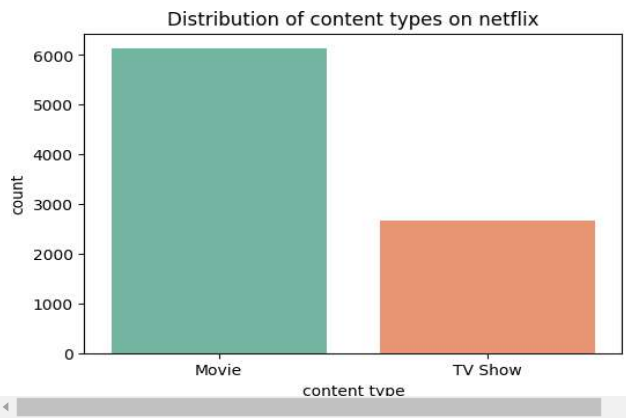
```
type_counts = data['type'].value_counts()
```

```
type_counts
```

type	count
Movie	6126
TV Show	2664

```
plt.figure(figsize=(6,4))
sns.barplot(x=type_counts.index, y=type_counts.values, palette='Set2')      #palette is for colors
plt.title('Distribution of content types on netflix')
plt.xlabel('content type')
plt.ylabel('count')
plt.show()
```

```
<ipython-input-48-ad2a103cb7a7>:2: FutureWarning:
Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `l
sns.barplot(x=type_counts.index, y=type_counts.values, palette='Set2')                  #palette is for colors
```




```
data['country'].value_counts()
```

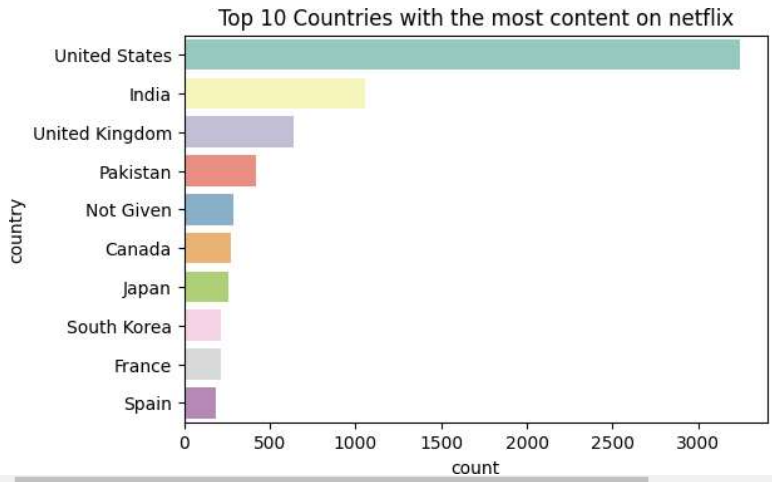
country	count
United States	3240
India	1057
United Kingdom	638
Pakistan	421
Not Given	287
...	...
Iran	1
West Germany	1
Greece	1
Zimbabwe	1
Soviet Union	1

86 rows × 1 columns

```
top_10_countries = data['country'].value_counts().head(10)
```

```
plt.figure(figsize=(6,4))
sns.barplot(x=top_10_countries.values,y=top_10_countries.index, palette='Set3')
plt.title('Top 10 Countries with the most content on netflix')
plt.xlabel('count')
plt.ylabel('country')
plt.show()
```

```
<ipython-input-51-6d7d42819e7e>:2: FutureWarning:
Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `y` variable to `hue` and set `l
sns.barplot(x=top_10_countries.values,y=top_10_countries.index, palette='Set3')
```




```
top_10_ratings = data['rating'].value_counts()
```

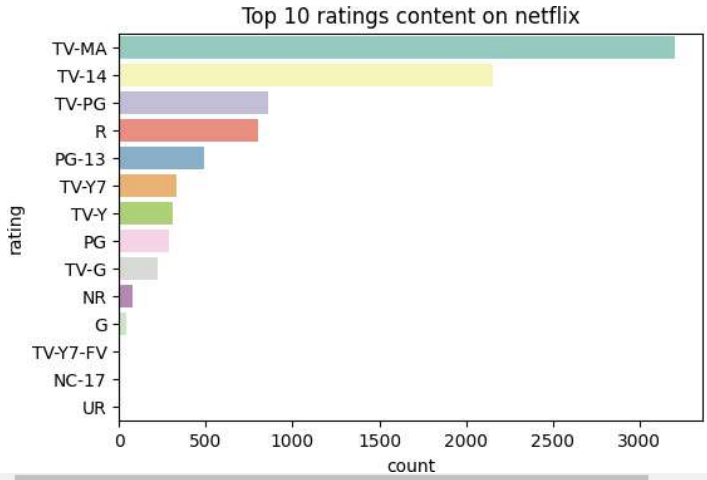
```
top_10_ratings
```

count

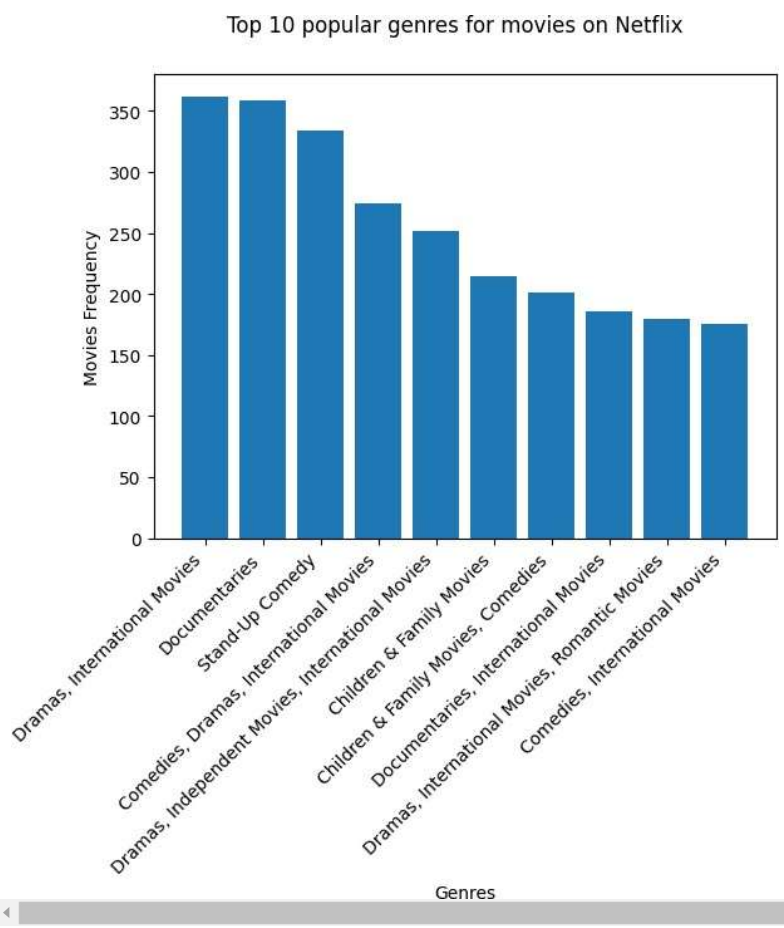
rating	
TV-MA	3205
TV-14	2157
TV-PG	861
R	799
PG-13	490
TV-Y7	333
TV-Y	306
PG	287
TV-G	220
NR	79
G	41
TV-Y7-FV	6
NC-17	3
UR	3

```
plt.figure(figsize=(6,4))
sns.barplot(x=top_10_ratings.values,y=top_10_ratings.index, palette='Set3')
plt.title('Top 10 ratings content on netflix')
plt.xlabel('count')
plt.ylabel('rating')
plt.show()
```

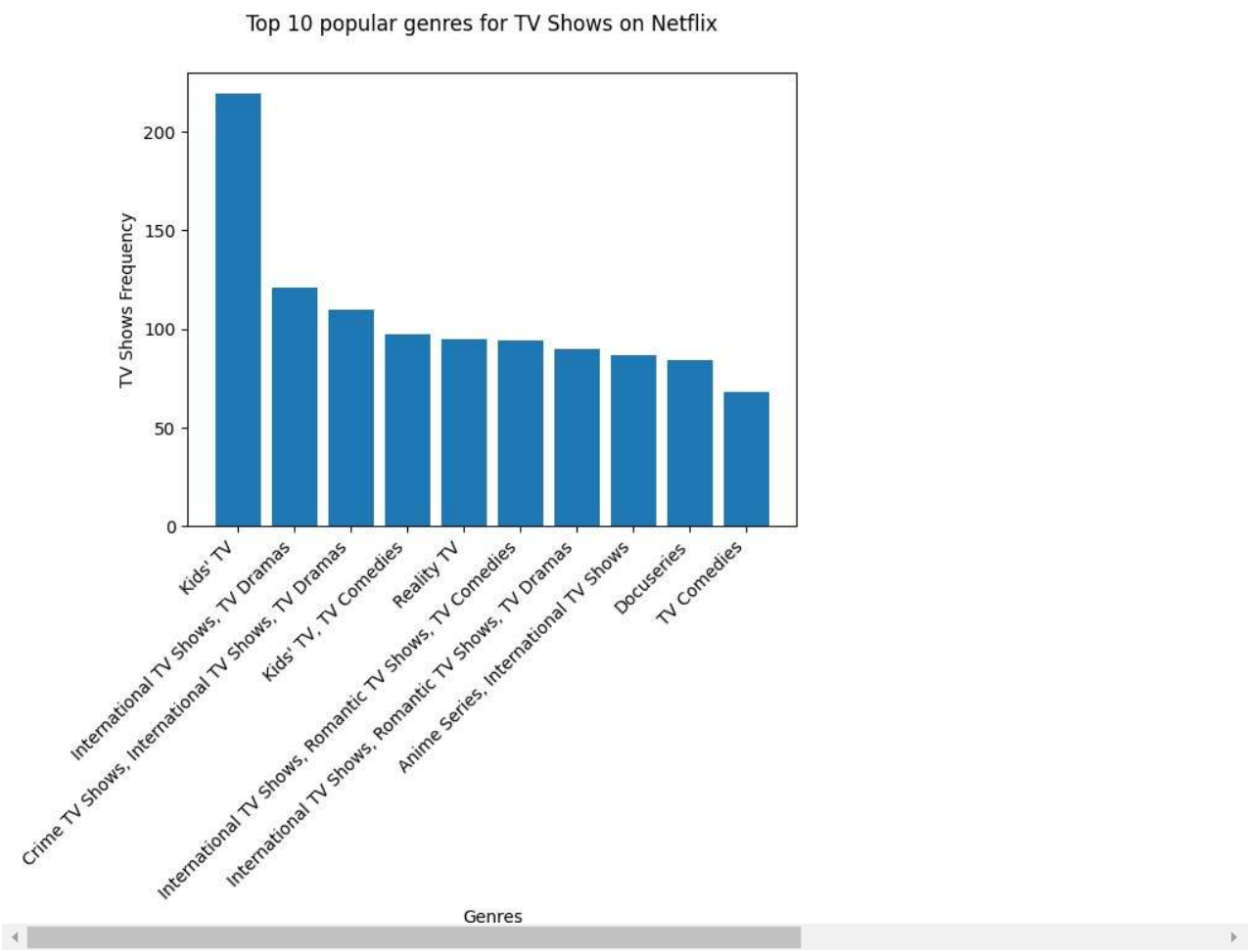
```
<ipython-input-54-a630e78014ad>:2: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `y` variable to `hue` and set `l

sns.barplot(x=top_10_ratings.values,y=top_10_ratings.index, palette='Set3')
```



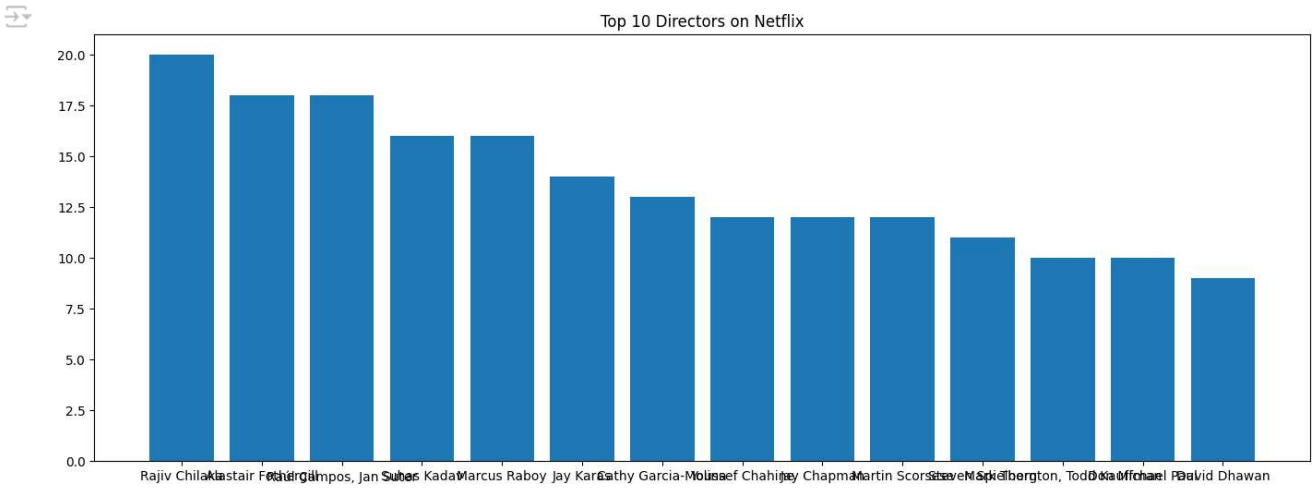

```
popular_movie_genre=data[data['type']=='Movie'].groupby("listed_in").size().sort_values(ascending=False)[:10]
popular_series_genre=data[data['type']=='TV Show'].groupby("listed_in").size().sort_values(ascending=False)[:10]
plt.bar(popular_movie_genre.index, popular_movie_genre.values)
plt.xticks(rotation=45, ha='right')
plt.xlabel("Genres")
plt.ylabel("Movies Frequency")
plt.suptitle("Top 10 popular genres for movies on Netflix")
plt.show()
```



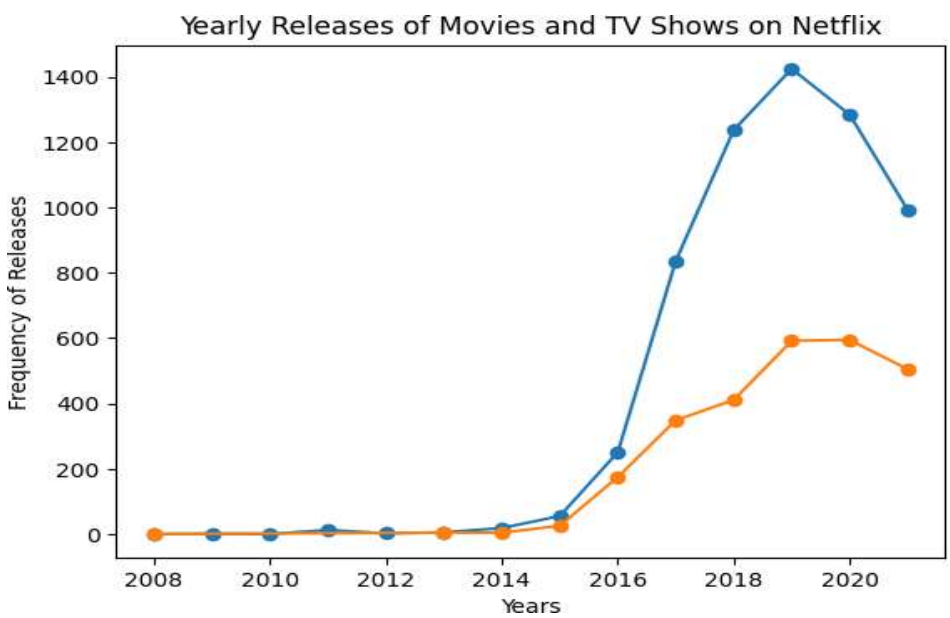
```
plt.bar(popular_series_genre.index, popular_series_genre.values)
plt.xticks(rotation=45, ha='right')
plt.xlabel("Genres")
plt.ylabel("TV Shows Frequency")
plt.suptitle("Top 10 popular genres for TV Shows on Netflix")
plt.show()
```



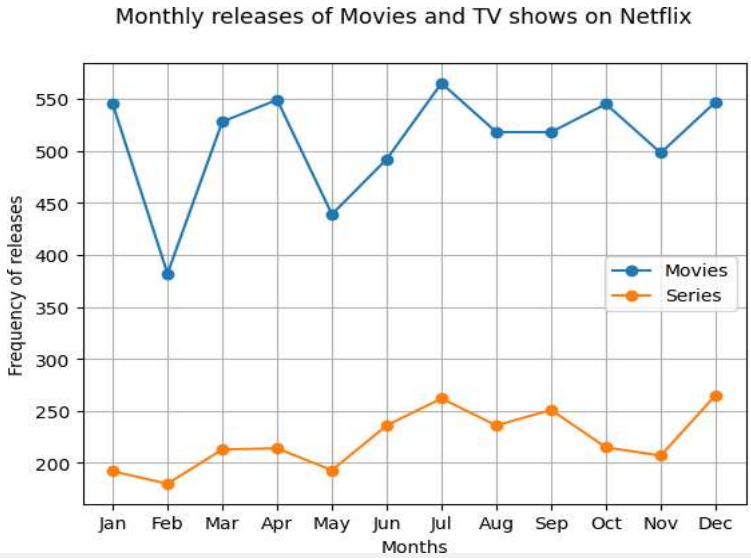

```
directors= data['director'].value_counts().reset_index().sort_values(by='count', ascending=False)[1:15]
plt.figure(figsize=(17,6))
plt.bar(directors['director'], directors['count'])
plt.title("Top 10 Directors on Netflix")
plt.show()
```



```
# Now you can extract the year
data['year_added'] = data['date_added'].dt.year

# Filter yearly release data for Movies and TV Shows
yearly_movie_releases = data.loc[data['type'] == 'Movie', 'year_added'].value_counts().sort_index()
yearly_series_releases = data.loc[data['type'] == 'TV Show', 'year_added'].value_counts().sort_index()

# Plot yearly releases
import matplotlib.pyplot as plt

plt.plot(yearly_movie_releases, marker='o')
plt.plot(yearly_series_releases, marker='o')
plt.xlabel("Years")
plt.ylabel("Frequency of Releases")
plt.title("Yearly Releases of Movies and TV Shows on Netflix")
plt.show()
```




```
data['month'] = data['date_added'].dt.month


monthly_movie_release=data[data['type']=='Movie']['month'].value_counts().sort_index()
monthly_series_release=data[data['type']=='TV Show']['month'].value_counts().sort_index()
plt.plot(monthly_movie_release.index, monthly_movie_release.values, label='Movies', marker = 'o')
plt.plot(monthly_series_release.index, monthly_series_release.values, label='Series', marker = 'o')
plt.xlabel("Months")
plt.ylabel("Frequency of releases")
plt.xticks(range(1, 13), ['Jan', 'Feb', 'Mar', 'Apr', 'May', 'Jun', 'Jul', 'Aug', 'Sep', 'Oct', 'Nov', 'Dec'])
plt.legend()
plt.grid(True)
plt.suptitle("Monthly releases of Movies and TV shows on Netflix")
plt.show()
```



data.head(15)

show_id		type	title	director	country	date_added	release_year	rating	duration	listed_in	year_added	month
	0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	United States	2021-09-25	2020	PG-13	90 min	Documentaries	2021	9
	1	s3	TV Show	Ganglands	Julien Leclercq	Crime TV Shows, France	2021-09-24	2021	TV-MA	1 Season	International TV Shows, TV Act...	2021	9
	2	s6	TV Show	Midnight Mass	Mike Flanagan	United States	2021-09-24	2021	TV-MA	1 Season	TV Dramas, TV Horror, TV Mysteries	2021	9
	3	s14	Movie	Confessions of an Invisible Girl	Bruno Garotti	Brazil	2021-09-22	2021	TV-PG	91 min	Children & Family Movies, Comedies	2021	9
	4	s8	Movie	Sankofa	Haile Gerima	United States	2021-09-24	1993	TV-MA	125 min	Dramas, Independent Movies, International Movies	2021	9
	5	s9	TV Show	The Great British Baking Show	Andy Devonshire	United Kingdom	2021-09-24	2021	TV-14	9 Seasons	British TV Shows, Reality TV	2021	9
	6	s10	Movie	The Starling	Theodore Melfi	United States	2021-09-24	2021	PG-13	104 min	Comedies, Dramas	2021	9
Motu Patlu	7	s939	Movie	in the Game of Zones	Suhas Kadav	India	2021-05-01	2019	TV-Y7	87 min	Children & Family Movies, Comedies, Music & Mu...	2021	5
	8	s13	Movie	Je Suis Karl	Christian	Germany	2021-09-23	2021	TV-MA	127 min	Dramas, International	2021	9

Next steps: Generate code with data | View recommended plots | New interactive sheet

In this project, we:

- ✓ Cleaned the data by handling missing values, removing duplicates, and converting data types.
- ✓ Explored the data through various visualizations such as bar plots and word clouds.
- ✓ Analyzed content trends over time, identified popular genres, and highlighted top directors.

Next Steps

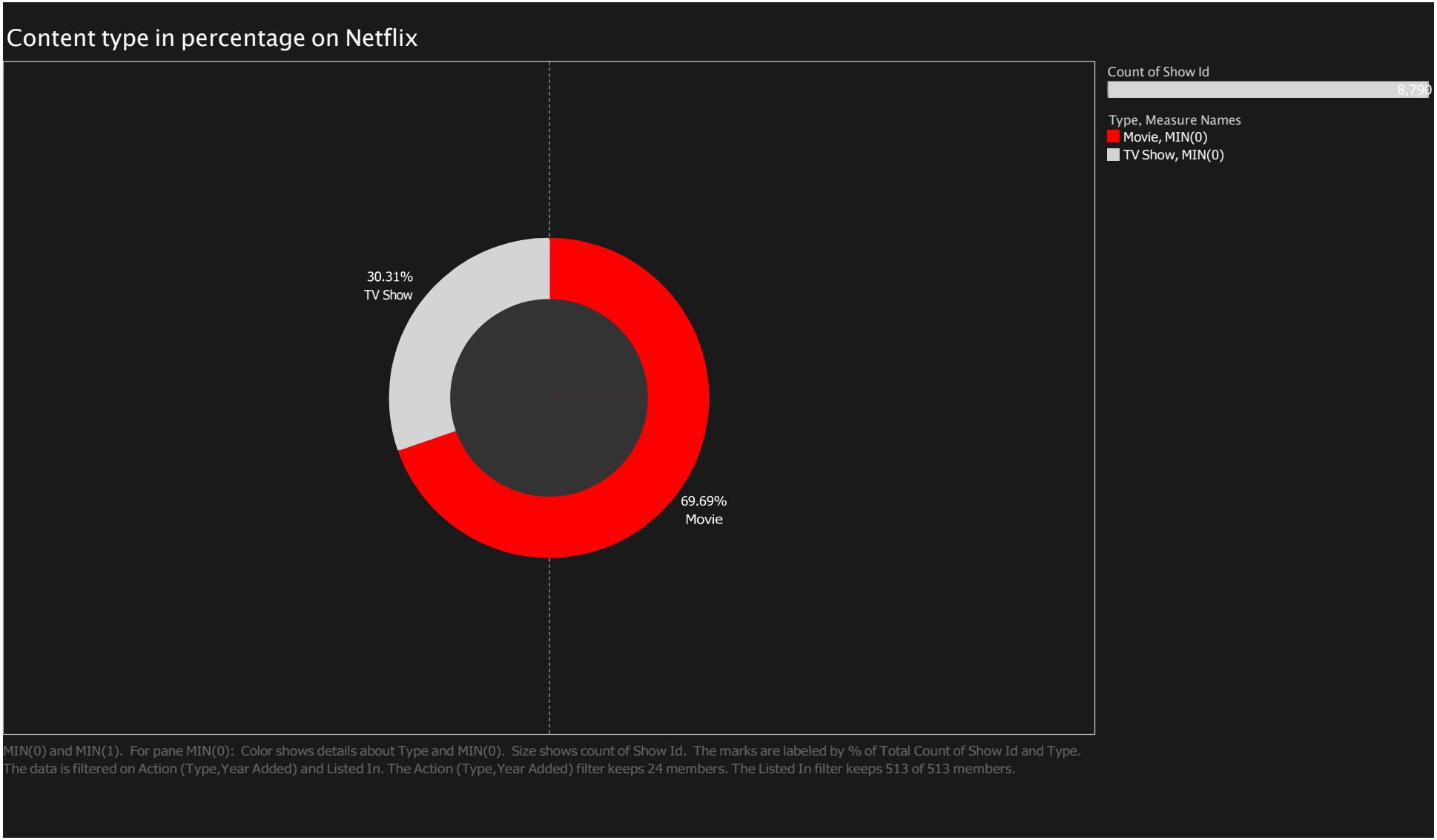
- ➢ **Advanced Visualization:** Use interactive plots or dashboards for more detailed analysis.

This project is a foundational exercise that introduces essential data analysis techniques, paving the way for more advanced projects.

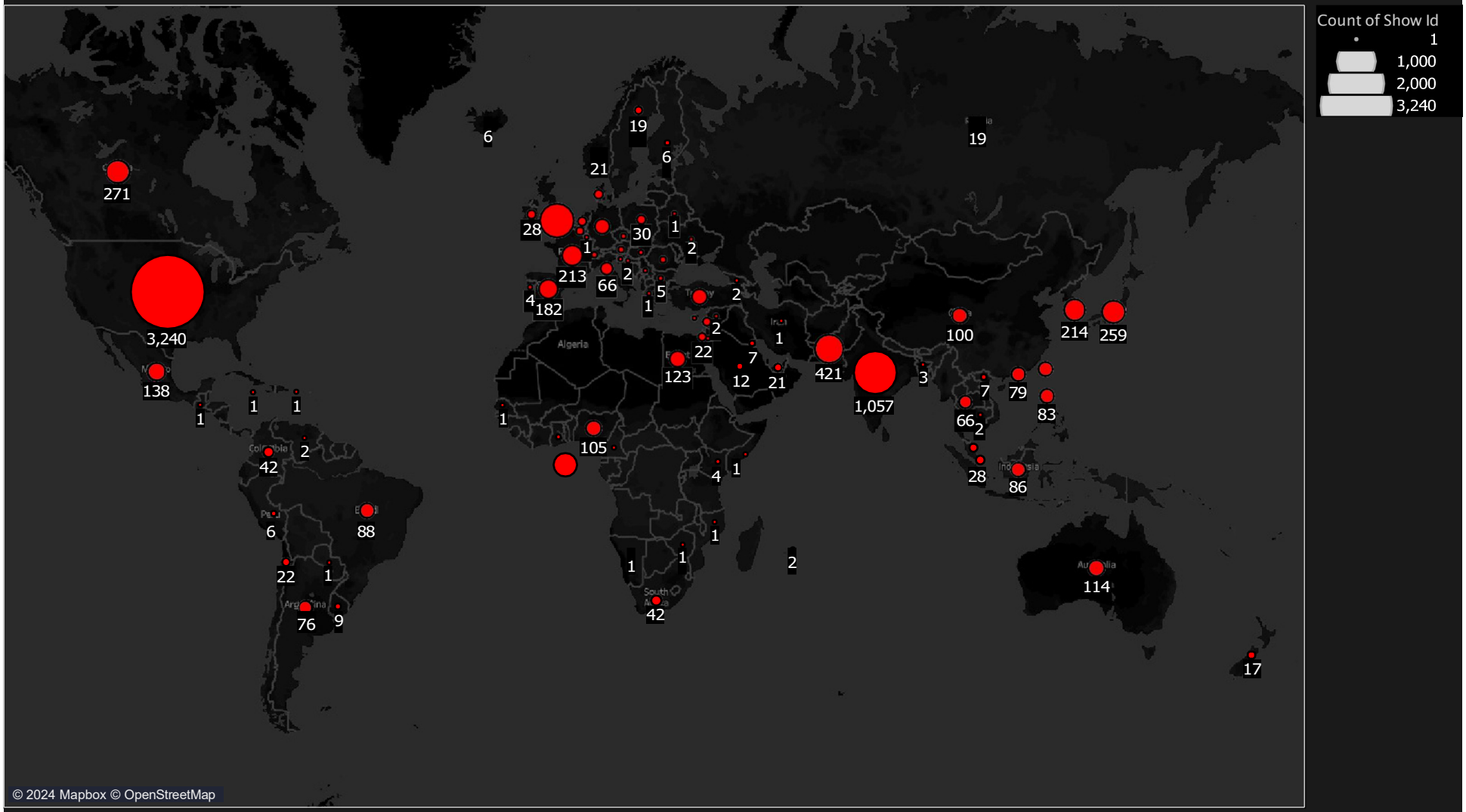
https://public.tableau.com/views/netflix_17347928383290/Dashboard2?:language=en-US&:sid=&:redirect=auth&:display_count=n&:origin=viz_share_link

Data Visualization

After cleaning, the dataset is set for some analysis and visualization with Tableau.

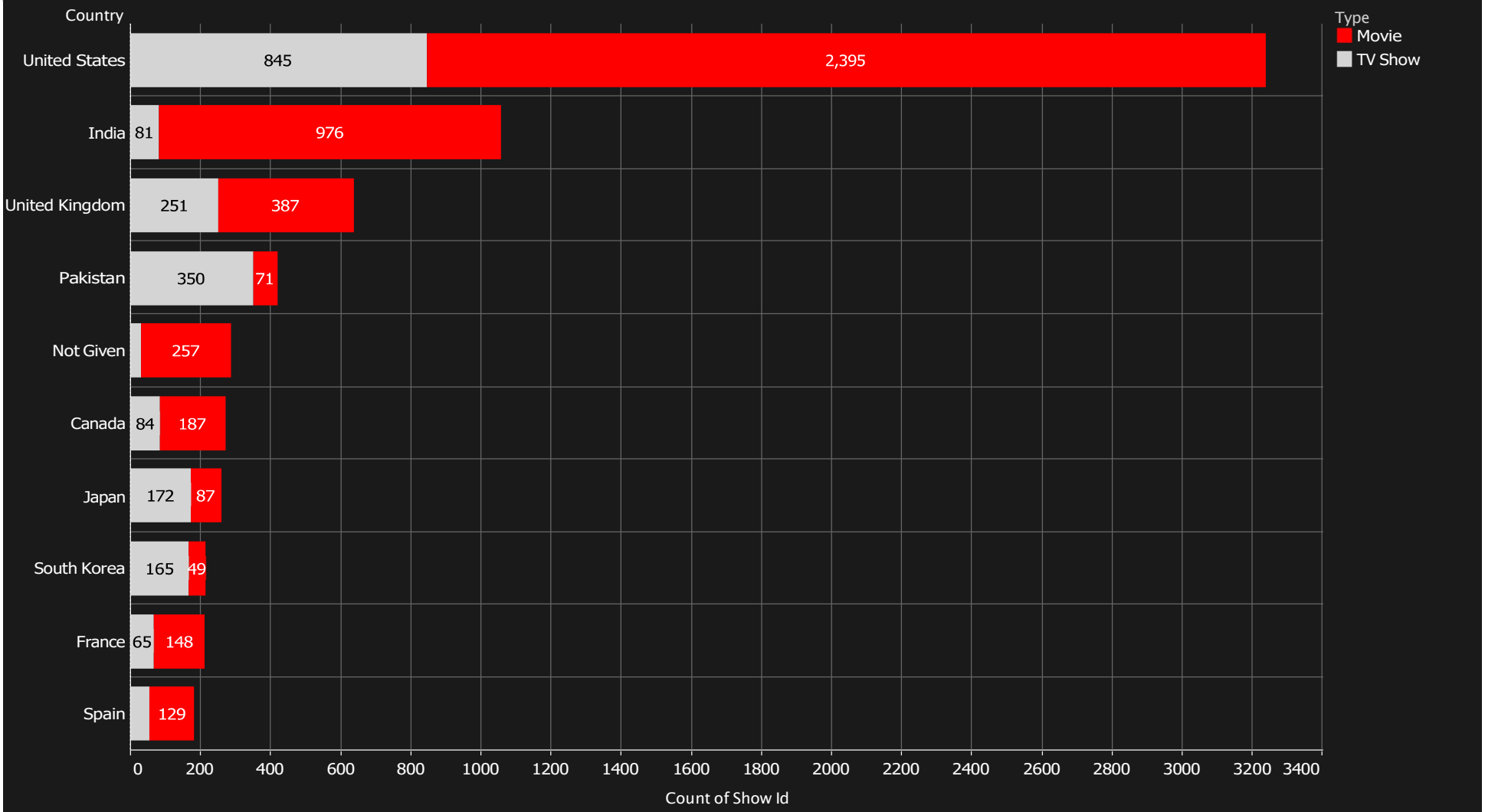


Movie & TV Show by Country on Netflix

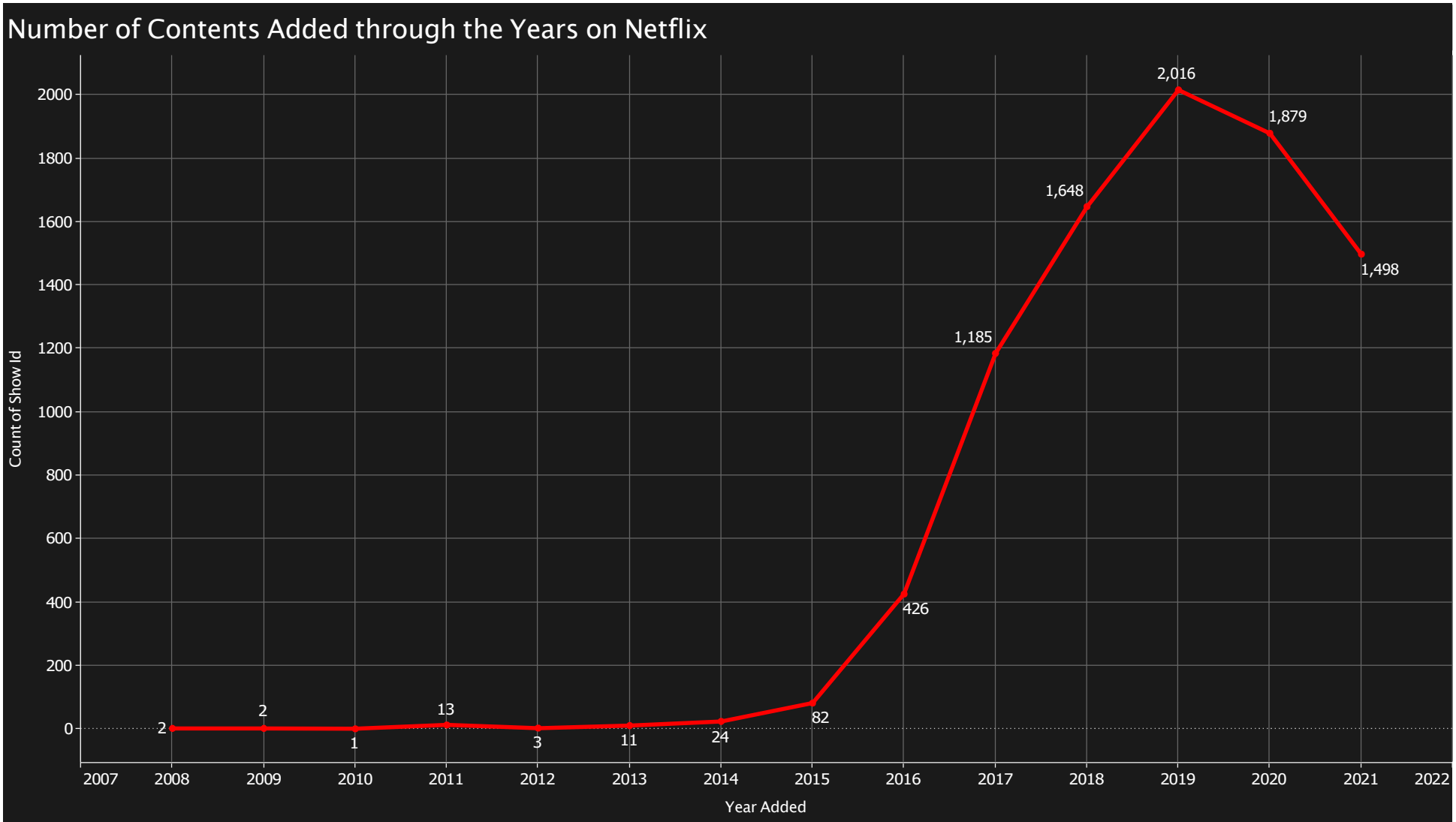


Map based on Longitude (generated) and Latitude (generated). Size shows count of Show Id. The marks are labeled by count of Show Id. Details are shown for Country. The data is filtered on Action (Type,Year Added) and Listed In. The Action (Type,Year Added) filter keeps 24 members. The Listed In filter keeps 513 of 513 members.

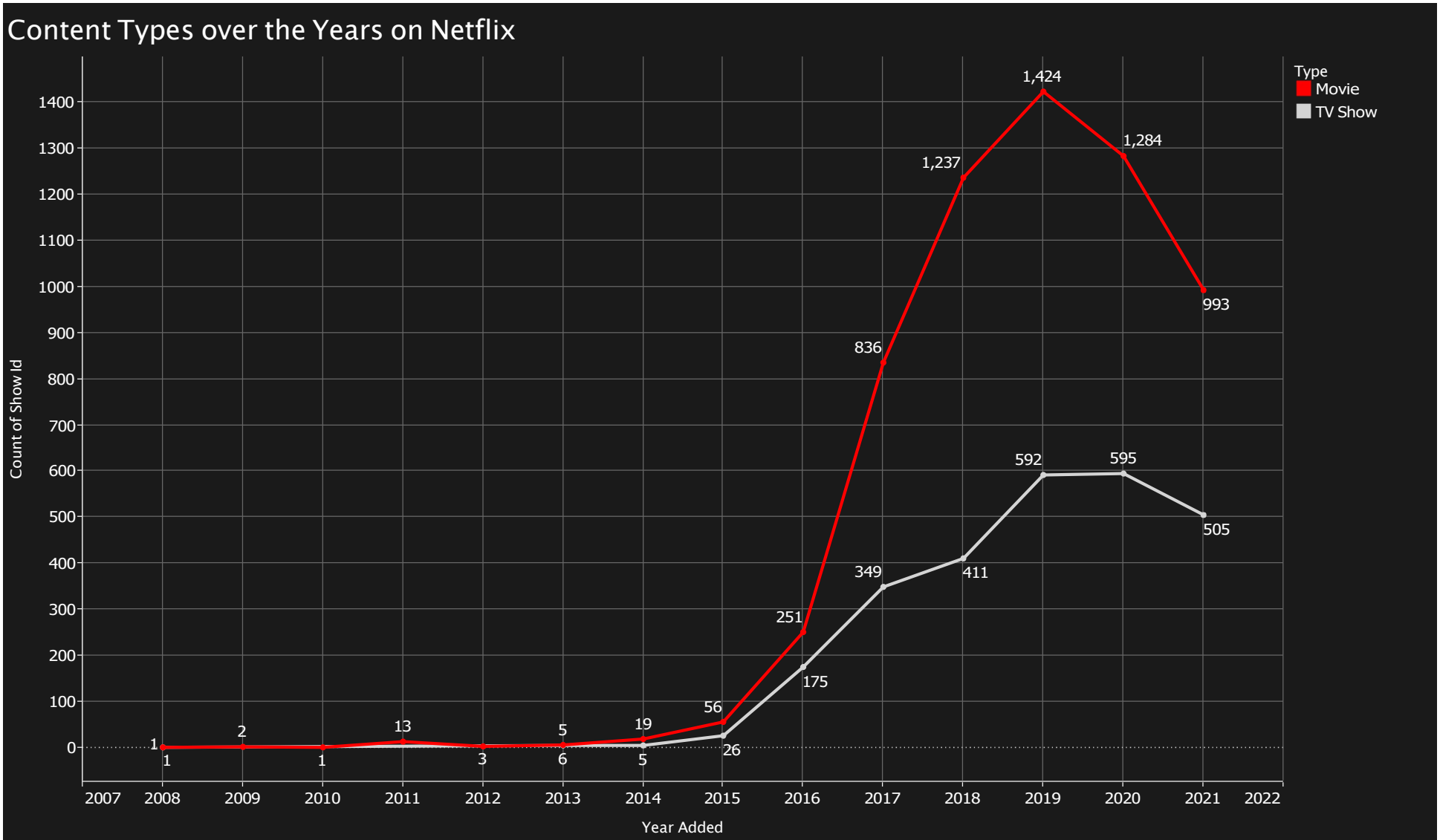
Top Movie & TV Show by Country on Netflix



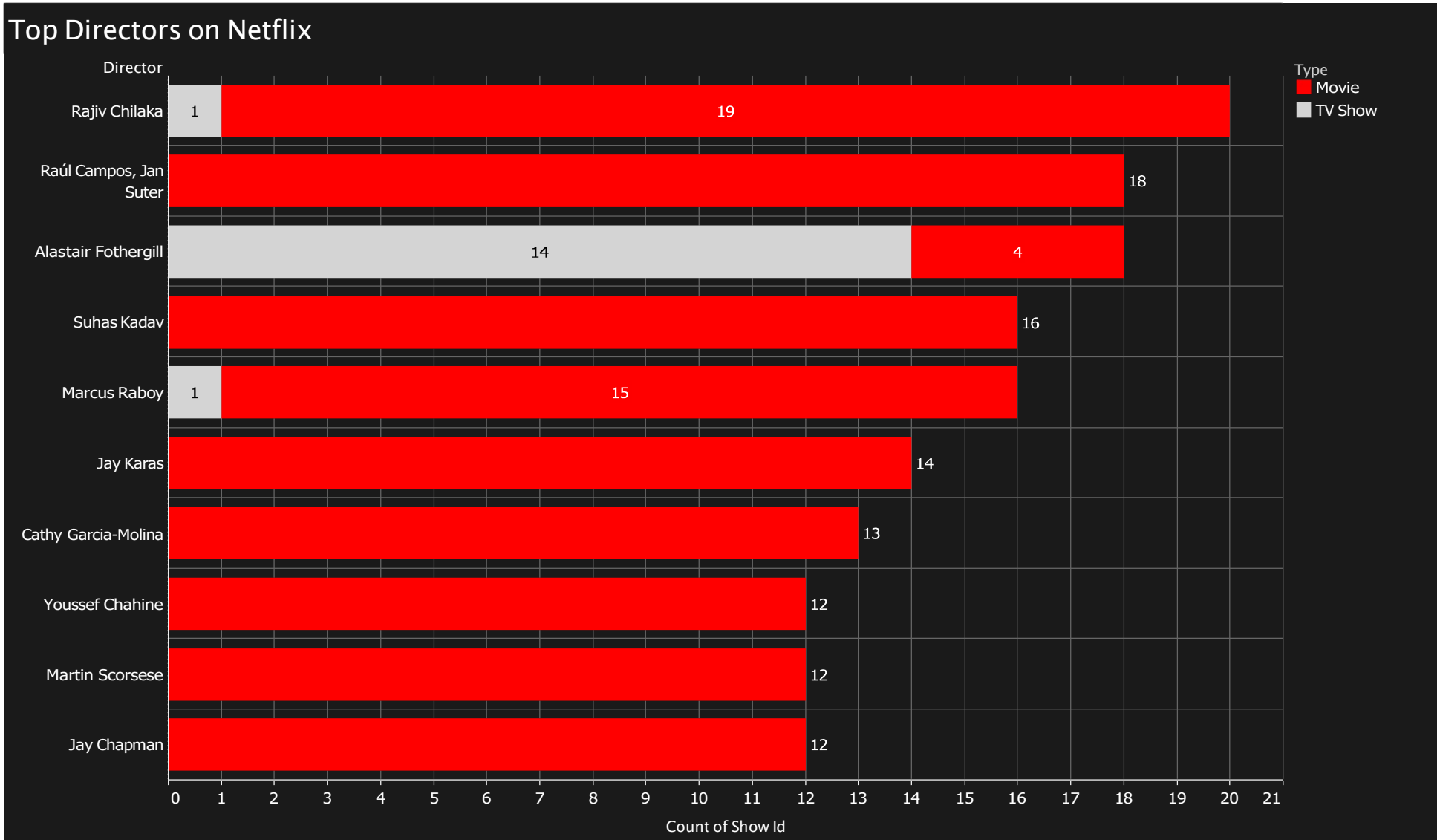
Count of Show Id for each Country. Color shows details about Type. The marks are labeled by count of Show Id. The data is filtered on Action (Country), Action (Type,Year Added) and Action (Director,Type). The Action (Country) filter keeps 86 members. The Action (Type,Year Added) filter keeps 24 members. The Action (Director,Type) filter keeps 4,581 members. The view is filtered on Country, which keeps 10 of 86 members.



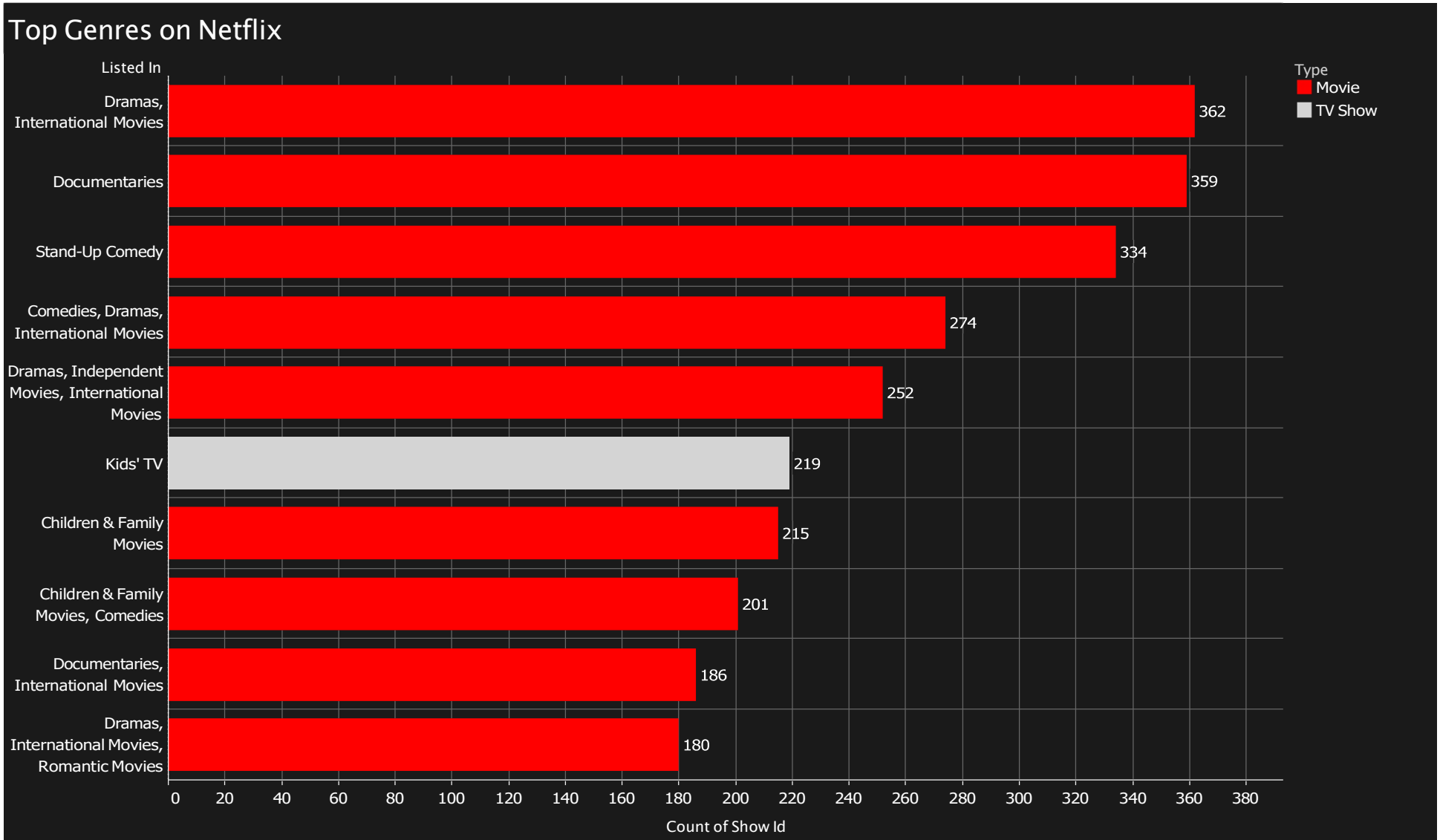
The trend of count of Show Id for Year Added. The marks are labeled by count of Show Id. The data is filtered on Action (Country), Action (Type,Year Added) and Listed In. The Action (Country) filter keeps 86 members. The Action (Type,Year Added) filter keeps 24 members. The Listed In filter keeps 513 of 513 members.



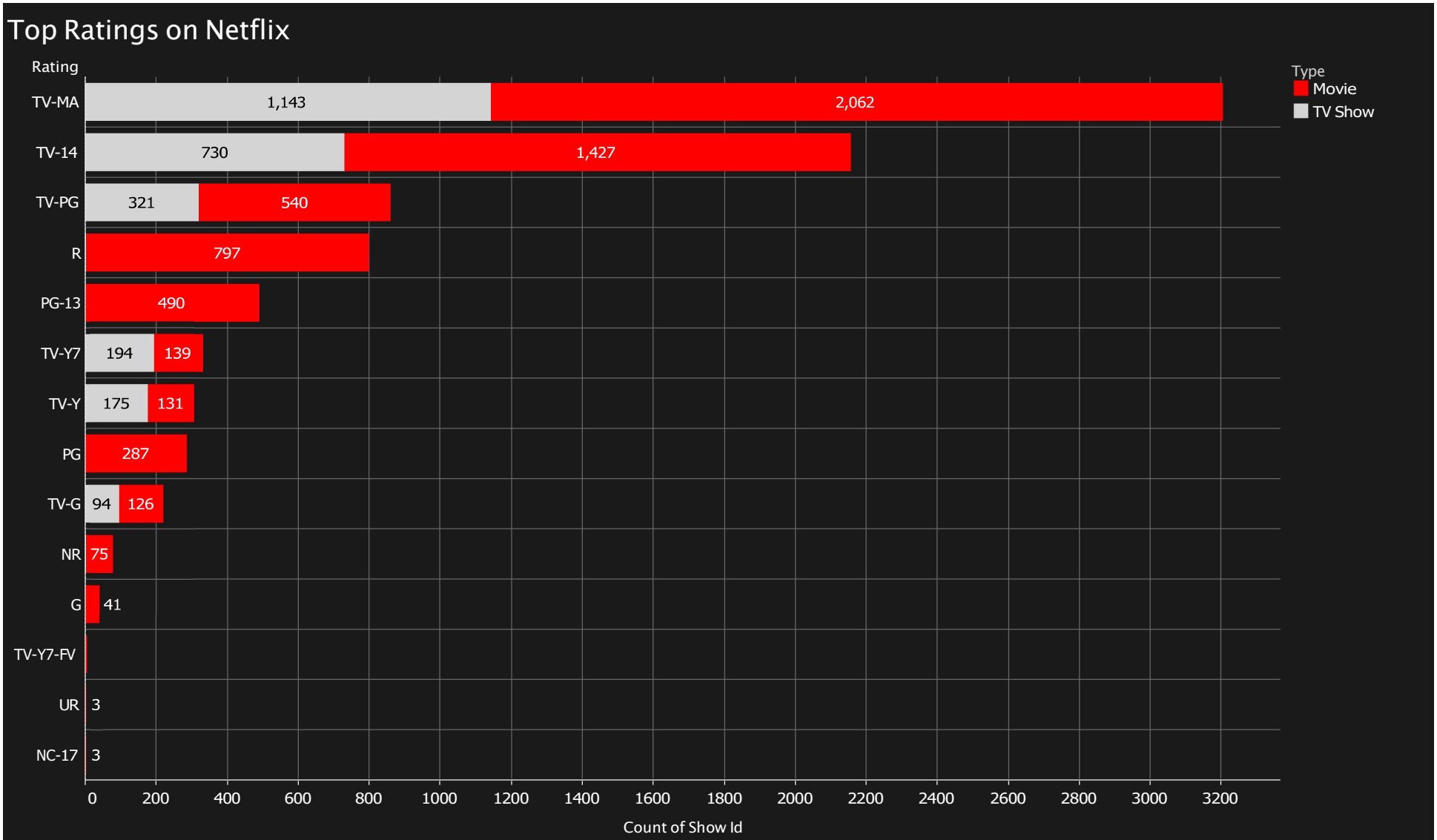
The trend of count of Show Id for Year Added. Color shows details about Type. The marks are labeled by count of Show Id. The data is filtered on Action (Country) and Listed In. The Action (Country) filter keeps 86 members. The Listed In filter keeps 513 of 513 members.



Count of Show Id for each Director. Color shows details about Type. The marks are labeled by count of Show Id. The view is filtered on Director, which keeps 10 of 4,528 members.

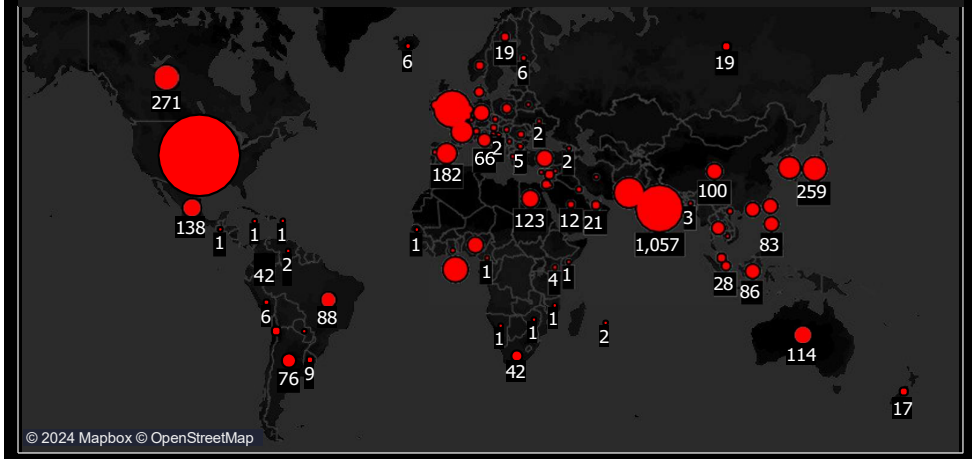


Count of Show Id for each Listed In. Color shows details about Type. The marks are labeled by count of Show Id. The data is filtered on Action (Director,Type), which keeps 4,581 members. The view is filtered on Listed In, which keeps 10 of 513 members.

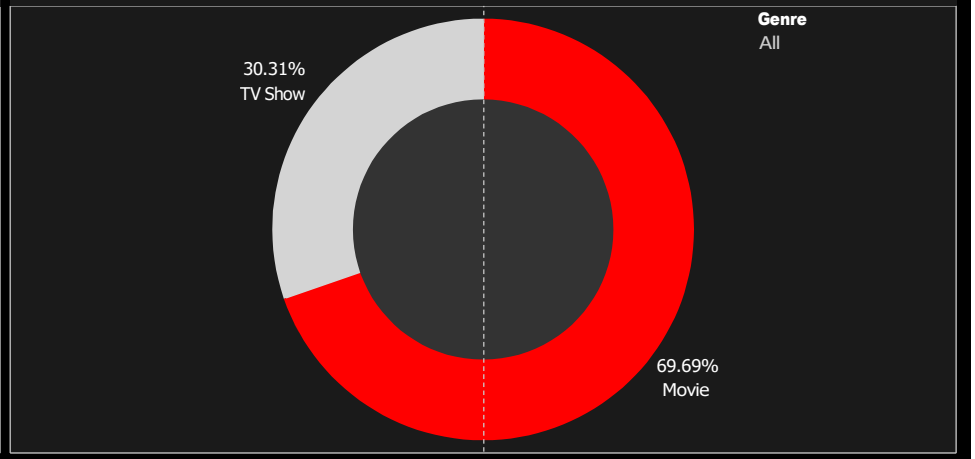


Count of Show Id for each Rating. Color shows details about Type. The marks are labeled by count of Show Id. The data is filtered on Action (Director,Type), which keeps 4,581 members.

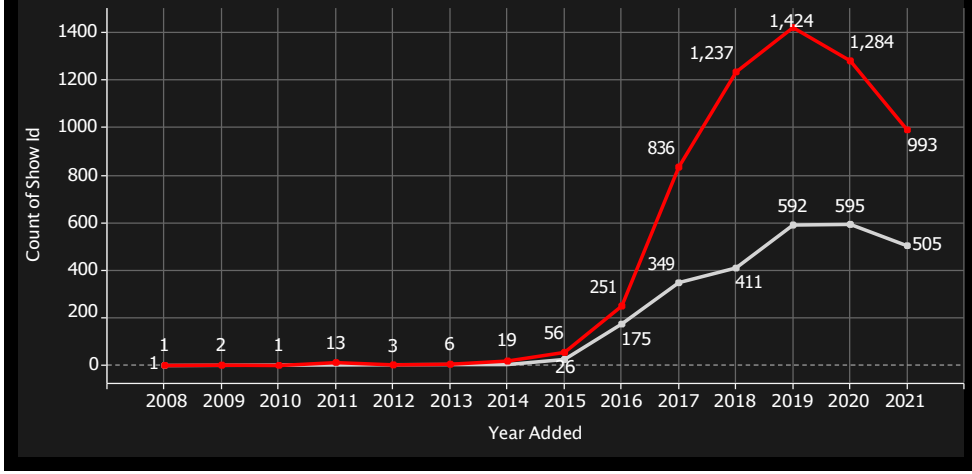
Movie & TV Show by Country on Netflix



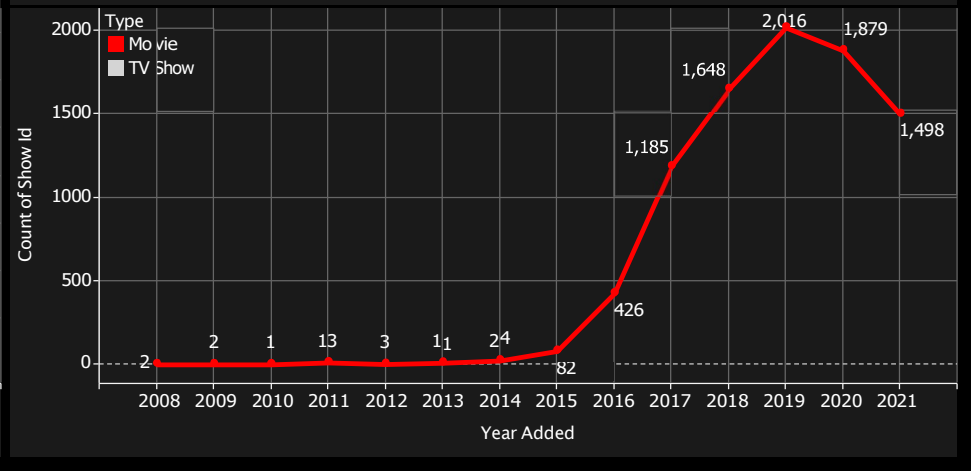
Content type in percentage on Netflix



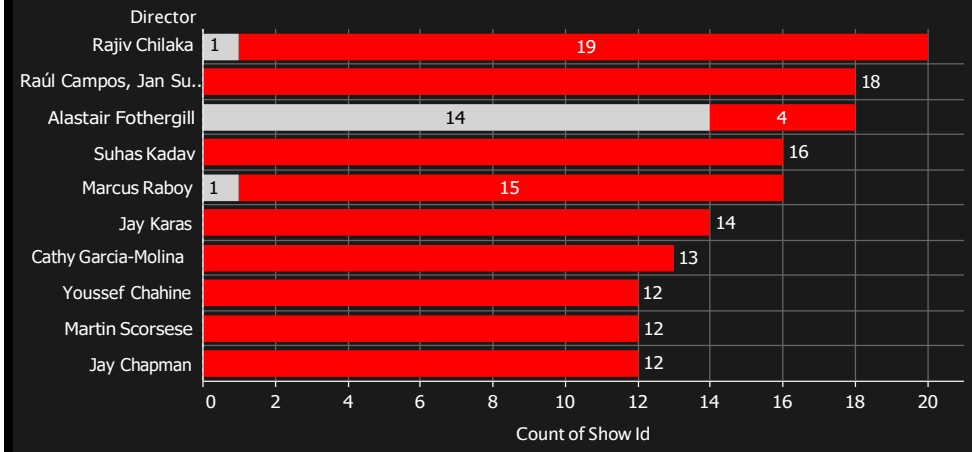
Content Types over the Years on Netflix



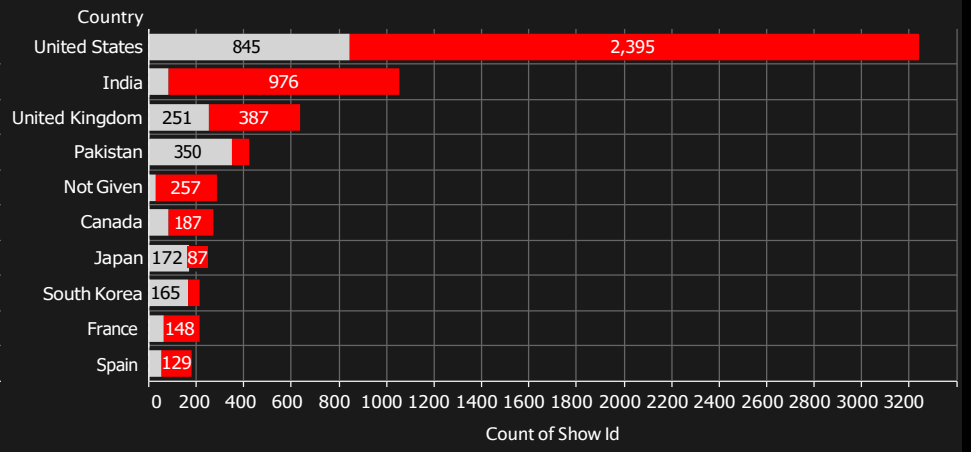
Number of Contents Added through the Years on Netflix



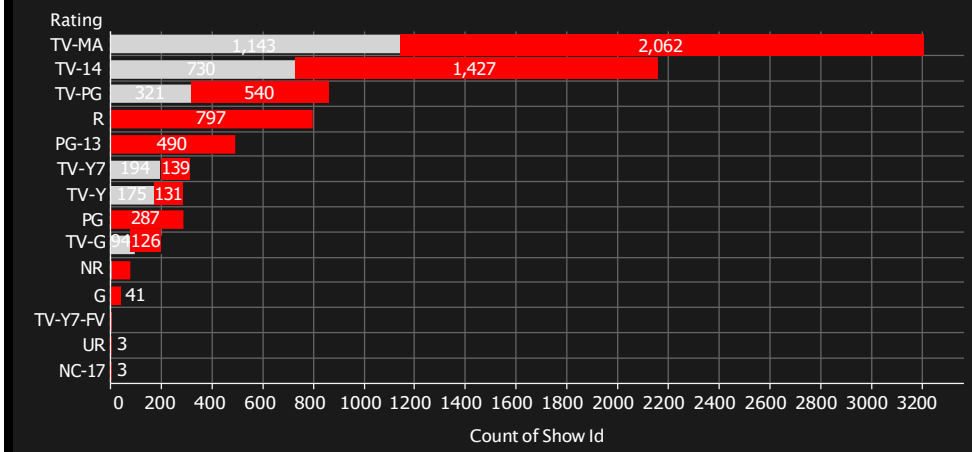
Top Directors on Netflix



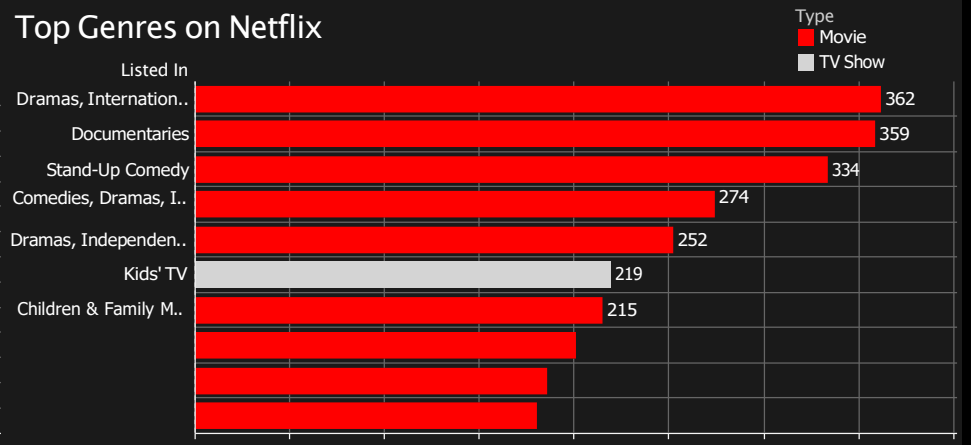
Top Movie & TV Show by Country on Netflix



Top Ratings on Netflix



Top Genres on Netflix



Observations for Each Chart/Dashboard:

Content Type Percentage:

- Movies dominate Netflix with 69.69%, while TV shows make up 30.31%.
- Indicates a larger library of movies compared to TV shows.

Movie & TV Show Distribution by Country:

- The U.S. has the highest content count (2,395), followed by India (976) and the UK (387).
- Suggests Netflix prioritizes content for the U.S. and India markets.

Year Added Trend:

- Sharp growth in content added around 2016–2018, with a peak in 2020.
- Reflects aggressive content expansion, possibly influenced by rising competition.

Content Types Over the Years:

- Movies consistently outnumber TV shows, with both growing after 2015.
- Aligns with Netflix's increasing focus on varied content types.

Top Genres:

- Dramas, international movies, and documentaries are the most frequent genres.
- Highlights Netflix's focus on diverse and globally appealing genres.

Top Ratings:

- Majority of content is TV-MA and TV-14, suitable for mature audiences.
- Reflects a strategy to target adult viewers.

Top Directors:

- Rajiv Chilaka and Raúl Campos lead in the number of directed shows.
- Indicates Netflix's investment in specific creators for targeted audiences.

Future scope:

- **Feature Engineering:** Create new features, such as counting the number of genres per movie or extracting the duration in minutes.
- **Machine Learning:** Use the cleaned and processed data to build models for recommendations or trend predictions.

Conclusion:

Netflix's content catalogue is significantly biassed towards films, accounting for approximately 70% of its offerings, indicating an interest for short-form, solo entertainment in people. The dominance of the US and Indian markets suggests a regional approach to meet high-demand areas. The increase in content additions between 2016 and 2020 demonstrates its response to competition and subscriber growth. A dedication to catering to a wide, adult-centric audience is demonstrated by Netflix's concentration on mature-rated content and globally diverse genres, such as international movies and documentaries.

Furthermore, the consistent investment in top directors and popular genres highlights its strategy to maintain quality while expanding. This data suggests Netflix's balance of global reach, market-specific tailoring, and an adaptive approach to evolving viewer preferences to maintain its leadership in the streaming industry.

koki venu gopal reddy