Netflix Movies & TV Shows Analysis using Tableau & Python

The dataset contains 11 attributes and 8807 rows. Each row represents a TV show or movie available on Netflix. (We have integrated two datasets)

Variables used in the dataset:

Column names: Cast, Show Id, Type, Title, Director, Country, Date added, Release Year, Rating, Duration, Genre.

Details about the dataset:

Column names and description are:

- Show Id: Unique ID for each show.
- Type: Type of the show (TV Show & Movie).
- Title: Name of the show.
- Director: Name of the director(s) of the show.
- Cast: Names of the cast members.
- Country: Country where the show was produced.
- Date added: Date when the show was added on Netflix.
- Release year: Year of release.
- Rating: Rating of the show.
- Duration: Duration of the show (in minutes for movies or number of seasons for TV shows).
- Genre: Genre are category that shares common characteristics of Movies & TV Shows.

Objective:

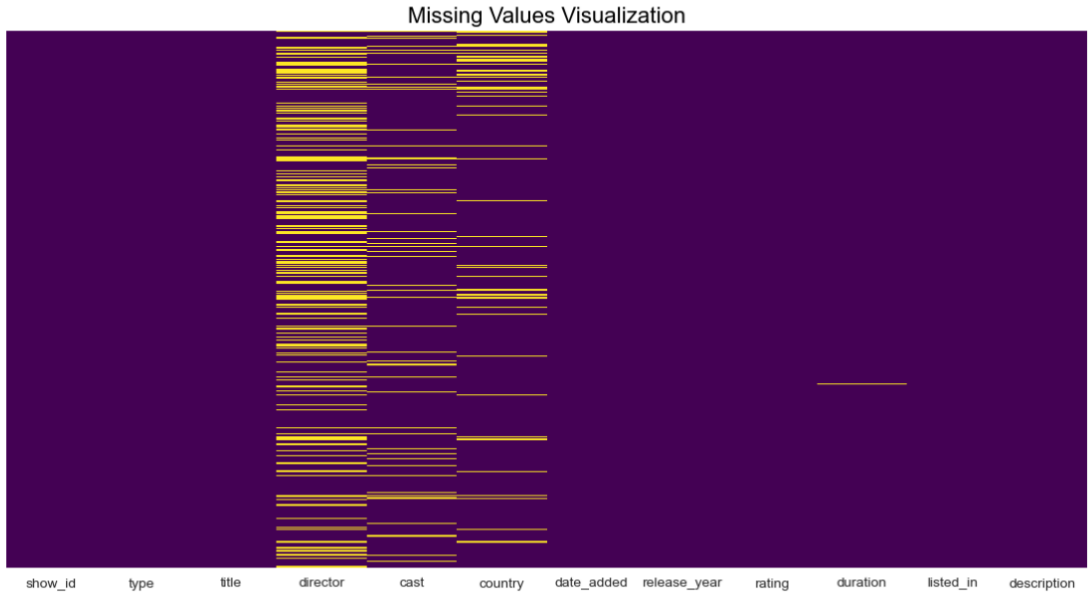
This project has four main objectives:

1. Develop an interactive dashboard in Tableau and Python to analyse the content available on Netflix in different countries, focusing on trends in TV shows and movies. This will involve analysing a comprehensive dataset of Netflix content to identify patterns in the types of shows available in different countries and presenting the findings through interactive visualizations such as heatmaps, bar charts, and scatterplots.
2. Use text-based features such as title, cast, and genre to identify similar content on Netflix and present the findings through a user-friendly interface that allows users to search for shows based on their preferred criteria.
3. Conduct network analysis of actors and directors associated with Netflix content to identify patterns and trends in their collaborations and relationships. The insights will be presented through visualizations such as network graphs, highlighting the most influential actors and directors and their impact on the popularity and success of specific shows.

4. Determine whether Netflix has shifted its focus towards TV shows in recent years by analysing trends in the number of TV shows and movies released each year. The goal is to provide insights into the strategic direction of Netflix and to identify any shifts in its content priorities over time. The findings will be presented through visualizations such as line graphs and stacked bar charts, comparing the number of TV shows and movies released each year and highlighting any significant trends or changes over time.

Exploratory Data Analysis

```
In [5]: f,ax=plt.subplots(figsize=(14,7))
        sns.heatmap(df.isnull(),yticklabels=False,cbar=False,cmap='viridis',ax=ax)
        ax.set_title('Missing Values Visualization',fontsize=16,color='black')
        plt.show()
```



The yellow horizontal lines in a column means that there are some missing values in that column. So, we “director, cast, country, duration” that have some missing values
The presence of missing values in the dataset can have an impact on the analysis and modeling results. Therefore, it is important to handle missing values appropriately, either by removing them, imputing them with appropriate values, or using algorithms that can handle missing values.

In [4]:
```
# checking missing data in stack data
df_clean = df.copy()
total = df_clean.isnull().sum().sort_values(ascending = False)
percent = (df_clean.isnull().sum()/df_clean.isnull().count()*100).sort_values(ascending = False)
missing_df_clean = pd.concat([total, percent], axis=1, keys=['Total', 'Percent'])
missing_df_clean
```

Out[4]:

	Total	Percent
director	2634	29.908028
country	831	9.435676
cast	825	9.367549
date_added	10	0.113546
rating	4	0.045418
duration	3	0.034064
show_id	0	0.000000
type	0	0.000000
title	0	0.000000
release_year	0	0.000000
listed_in	0	0.000000
description	0	0.000000

This information is useful for identifying which columns in the dataset have the most missing values and how much data is missing. This information can be used to decide how to handle missing values in the dataset, such as by imputing missing values or dropping rows or columns with a high percentage of missing values.

In [7]:
```
# number of unique values for each variable
df.nunique(axis=0)
```

Out[7]:
```
show_id         8807
type               2
title           8807
director        4528
cast            7692
country          748
date_added      1767
release_year      74
rating            17
duration         220
listed_in        514
description     8775
dtype: int64
```

In [8]:
```
# statistical summary of numeric variables
df.describe()
```

Out[8]:

	release_year
count	8807.000000
mean	2014.180198
std	8.819312
min	1925.000000
25%	2013.000000
50%	2017.000000
75%	2019.000000
max	2021.000000

The Year variable ranged from 1925 to 2021.

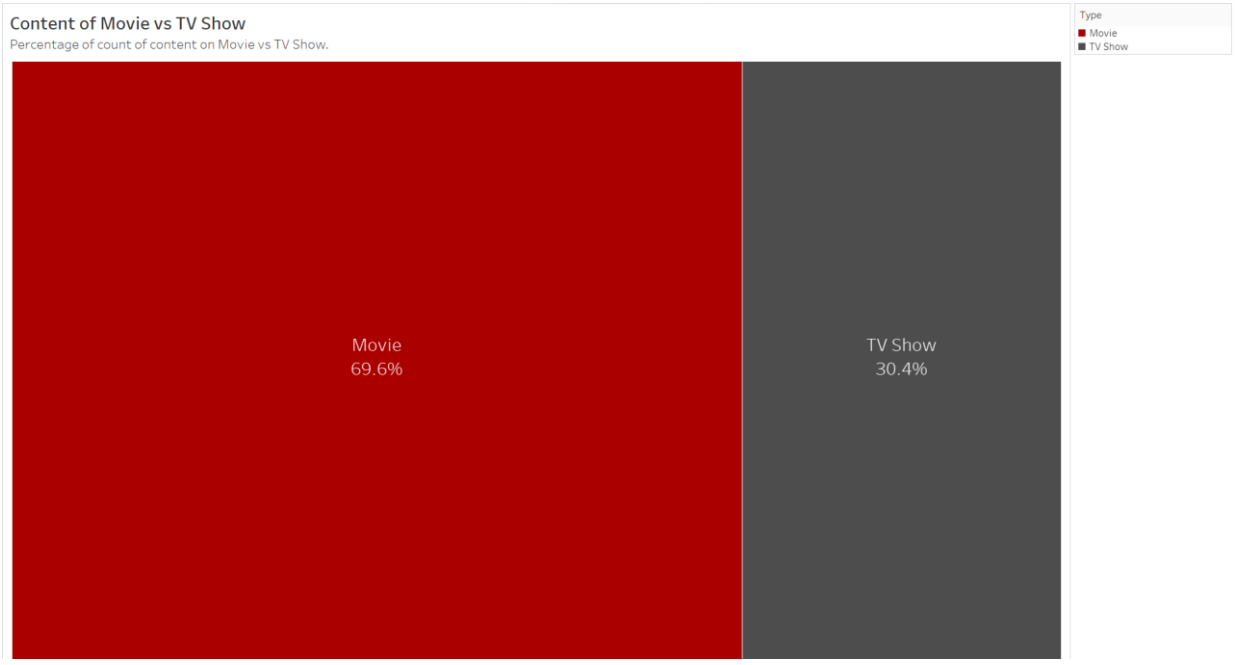
```
In [6]: df.info()
```

```
<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
```

The dataset contains 12 columns, out of which are 5 integers, 6 strings, and 1 country.

Listed below are few of the charts made in Tableau:

1. **Tree Map**



Based on the visual representation of the data, the chart displays the percentage of movies and TV shows on Netflix, where movies account for 69.6% of the content, and TV shows account for 30.4% of the content.

This insight suggests that movies make up a significant portion of the content available on Netflix, while TV shows account for a smaller portion of the content. However, it's important to note that this analysis only considers the percentage of content by type (movies vs. TV shows) and doesn't provide any information on the popularity or quality of the content. Additionally, the analysis doesn't consider any other factors that may be important to Netflix users, such as the genre of the content or the country of origin.

2. Geospatial Chart

Based on the visual representation of the data, the chart displays the number of movies and TV shows available on Netflix in various countries, categorized by region.

Some observations and insights from the chart are:

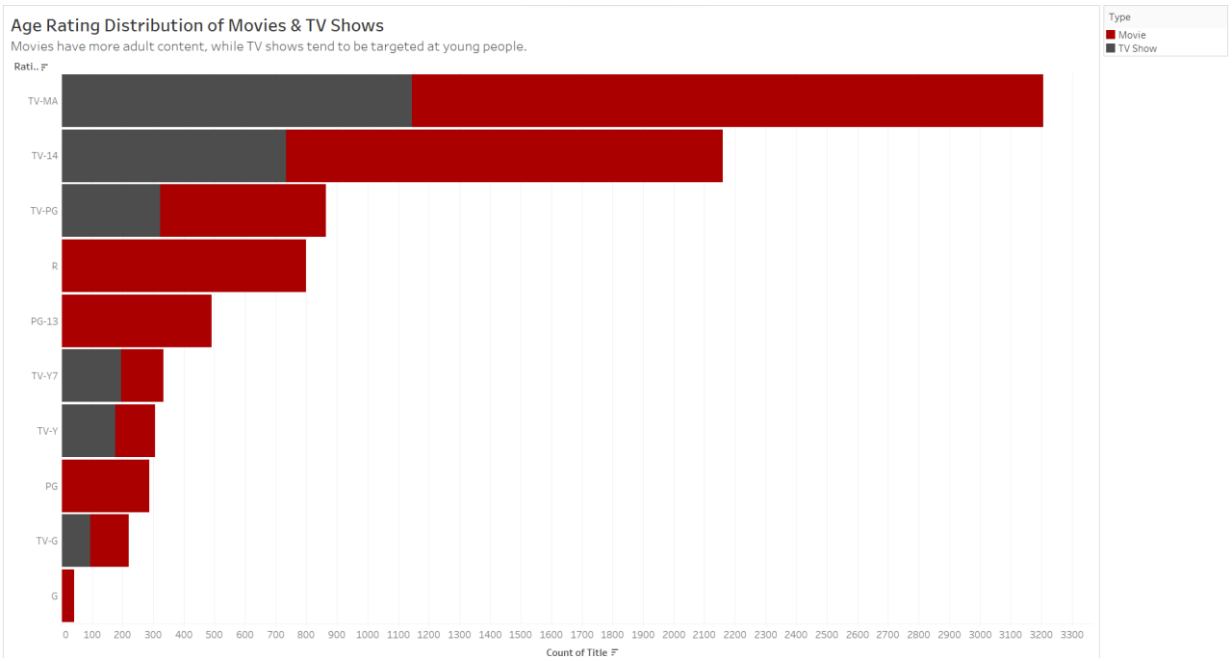
The United States has the largest number of movies and TV shows available on Netflix, with over 3,211 titles.

India has the second-largest number of movies and TV shows available on Netflix, with over 1,008 titles.

The chart also suggests that Netflix has a relatively smallest library of titles in Nigeria.

Overall, the chart provides a useful visualization of the number of titles available on Netflix across different regions and countries. It suggests that Netflix has a significant amount of content available in the United States and India, and that there may be regional variations in the amount of content available on the platform.

3. Distributed Bar Chart



The chart displays the distribution of age ratings for movies and TV shows available on Netflix.

Some observations and insights from the chart are:

The majority of movies and TV shows on Netflix have an age rating of TV-MA, which indicates that the content is intended for mature audiences only.

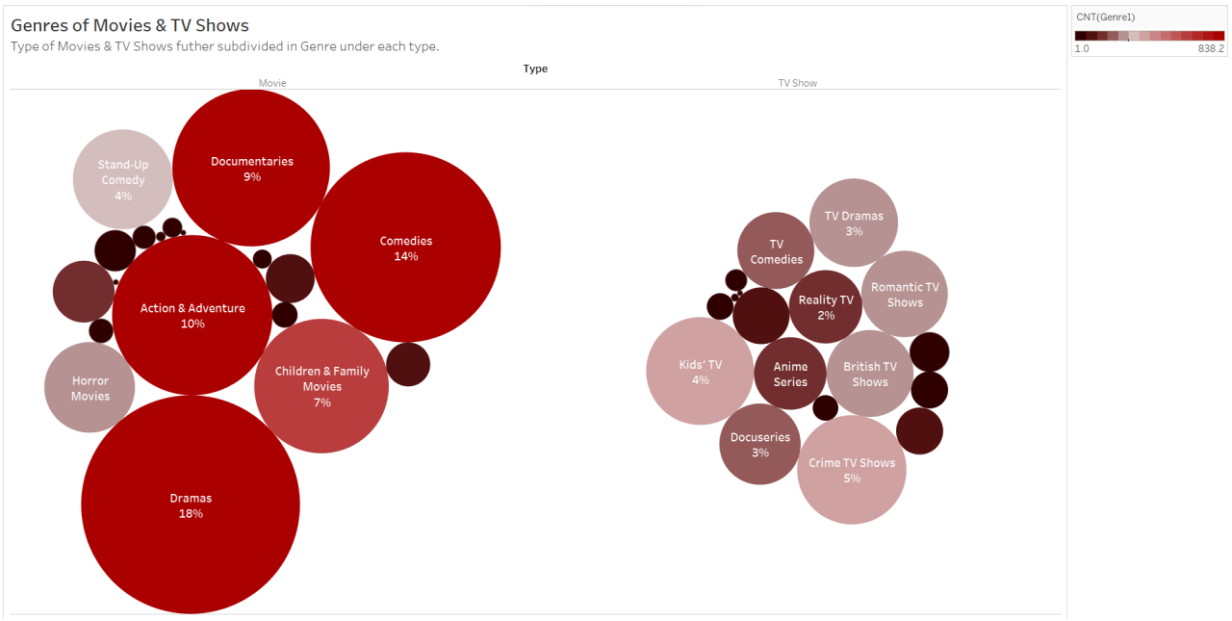
Other common age ratings for movies and TV shows on Netflix include TV-14, PG-13, and TV-R, each of which has several hundred titles available on the platform.

Certain age ratings, such as R and G, are primarily associated with movies, while others, such as TV-Y and TV-G, are primarily associated with TV shows.

The chart also suggests that Netflix offers a relatively small number of titles with a TV-PG and G, rating compared to other age ratings.

Overall, the chart provides a useful visualization of the distribution of age ratings for movies and TV shows available on Netflix. It suggests that Netflix offers a wide range of content intended for various age groups, but that most of the content is intended for mature audiences.

4. Bubble Chart



The chart represents the distribution of genres for movies and TV shows available on Netflix. Each bubble in the chart represents a different genre, and the size of the bubble corresponds to the number of titles in that genre.

Some insights from the chart include:

The most common genre for movies on Netflix is dramas, which make up 18% of all movie titles. Comedies and action & adventure movies are also relatively common, comprising 14% and 10% of movie titles, respectively.

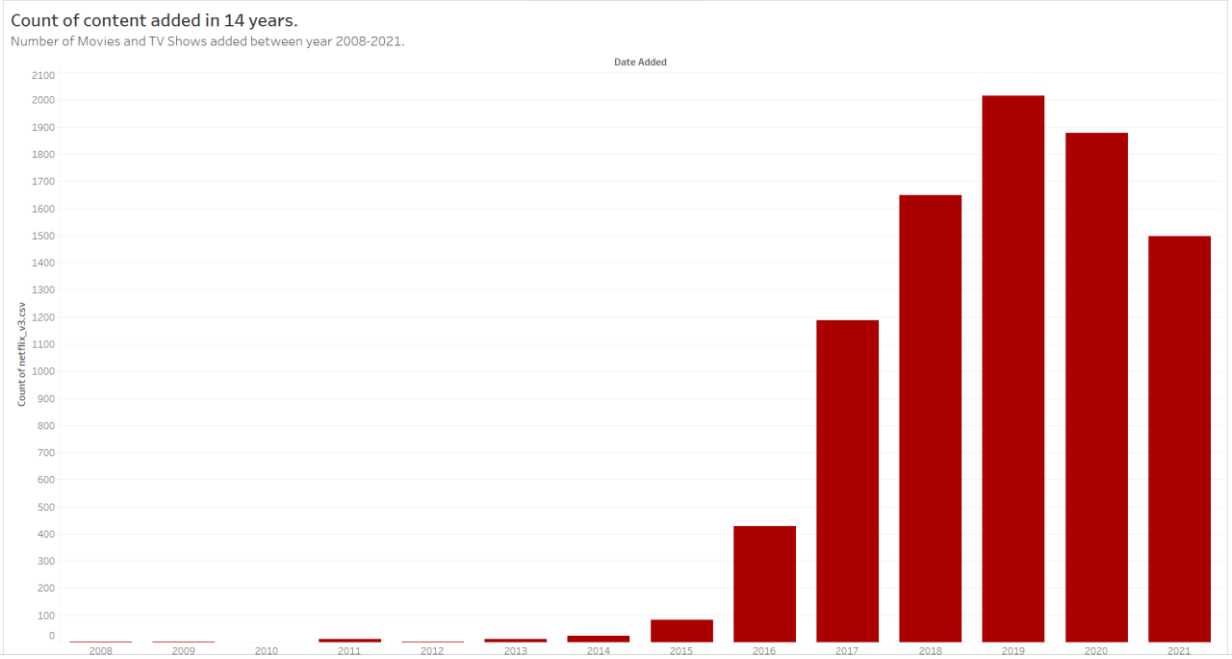
For TV shows, the most common genres are crime TV shows and kids' TV shows, which make up 5% and 4% of all TV show titles, respectively. Docuseries and TV dramas are also relatively common genres for TV shows on Netflix.

The chart suggests that there is a greater variety of genres available for movies compared to TV shows. There are several genres with a significant number of movie titles, while TV shows are relatively concentrated in a few genres.

Interestingly, the chart also indicates that there is only one LGBTQ movie available on Netflix, which suggests a potential area for improvement in terms of diversity and representation in the content offered by the platform.

Overall, the chart provides a useful visualization of the distribution of genres for movies and TV shows available on Netflix, and highlights some of the most common and less common genres represented on the platform.

5. Histogram



From the given bar chart, we can see that the number of movies and TV shows added on Netflix has been consistently increasing from the year 2008 to 2020, with a peak in 2019 with 2016 count.

However, there is a dip in 2020 with 1879 count. The dip in the number of new movies and TV shows added to Netflix in 2020 could be attributed to several factors. One factor could be the impact of the COVID-19 pandemic, which led to disruptions in the film and television production industry, resulting in delays in releasing new content.

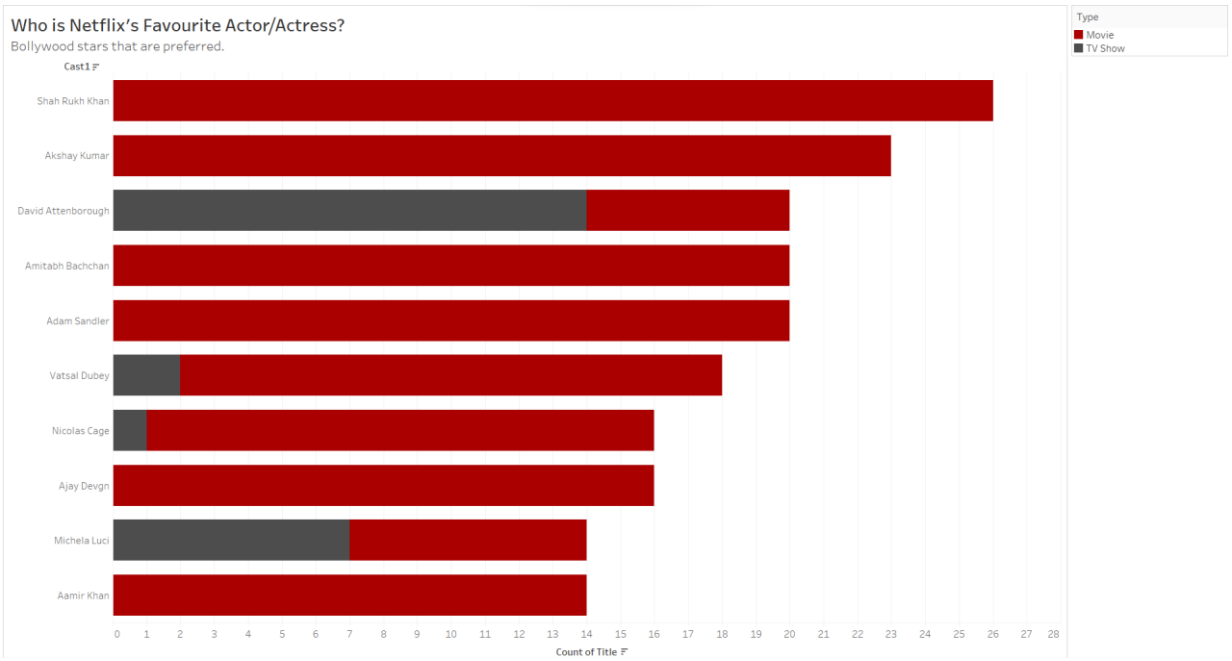
Another factor could be the increasing competition from other streaming services like Amazon Prime, Hulu, and Disney+. These services have been growing in popularity and are also investing heavily in producing original content, which could be drawing viewers away from Netflix.

It's worth noting that Netflix remains the dominant player in the streaming industry with a vast content library and a large subscriber base. However, as more competitors enter the market, it will be interesting to see how Netflix adapts and continues to attract new viewers.

We can also see that the number of movies and TV shows added in the year 2016 and 2017 shows a significant spike in comparison to the previous years. This could be due to the fact that Netflix was expanding globally during this time, leading to an increased demand for content.

The trend of increasing content on Netflix is consistent with the company's strategy of investing heavily in original programming, which has contributed to its massive popularity and success in the entertainment industry. Overall, the bar chart provides insights into the growth of Netflix's content library and its evolving business strategy.

6. Distributed Bar Chart



This chart shows that Netflix seems to prefer Bollywood stars as the favorite actors/actresses on its platform. Shah Rukh Khan and Akshay Kumar are the top two actors, with 26 and 23 movies respectively. This indicates that Bollywood movies are popular among Netflix users.

David Attenborough, who has 14 TV shows and 6 movies, is the only non-Bollywood star to make it to the top of the list. This shows that Netflix is not just limited to Bollywood content and also values non-fiction programming.

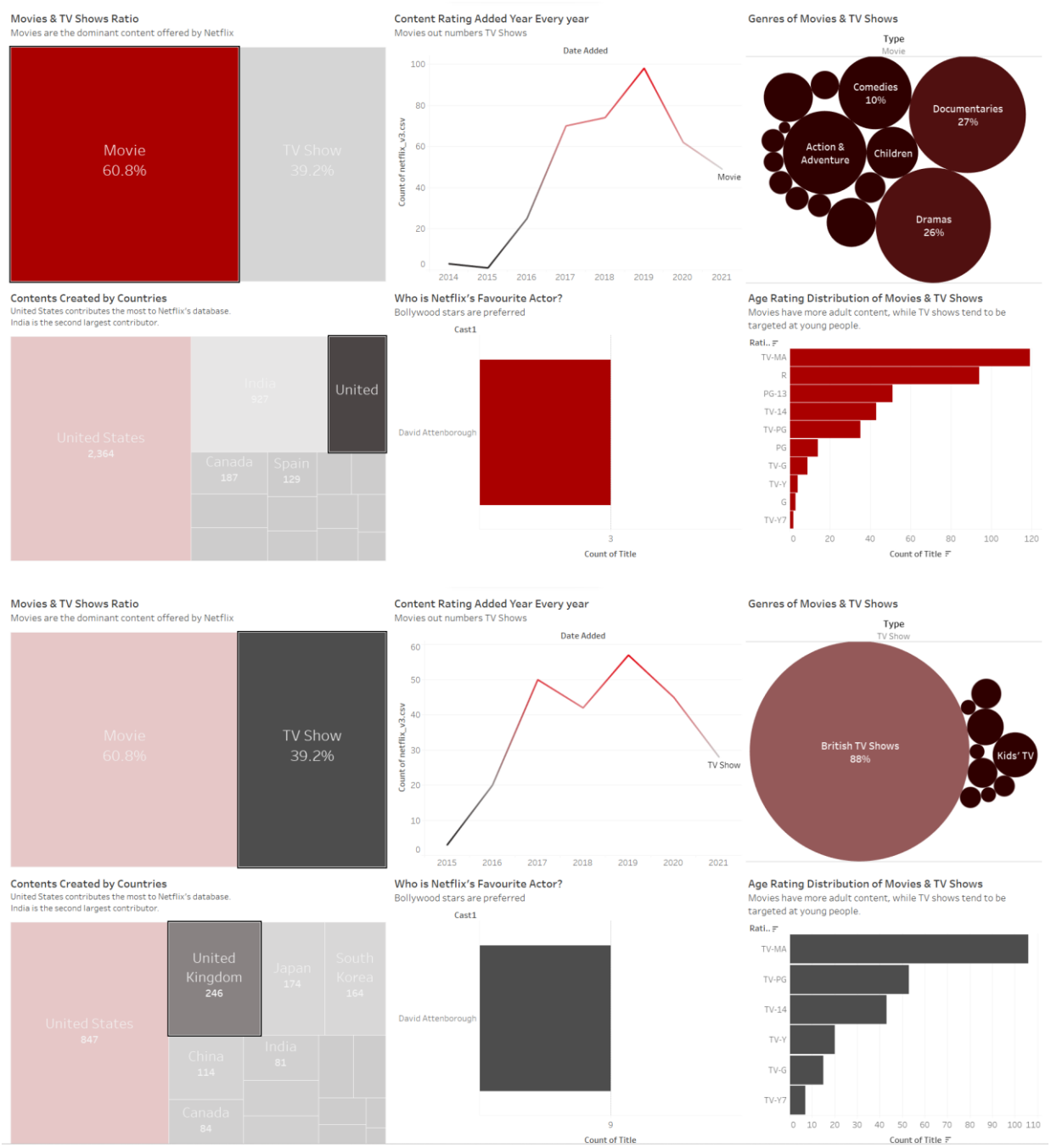
Amitabh Bachchan and Adam Sandler both have 20 movies each, which demonstrates that Netflix caters to diverse audiences and their preferences.

Nicolas Cage, and Ajay Devgn follow with 16 movies each, indicating that these actors are also popular among Netflix users.

Finally, Michela Luci has 7 TV shows and 7 movies, and Vatsal Dubey has 16 movies, both indicating that Netflix also provides a wide range of content for children and international audiences followed by Aamir Khan with 14 movies.

Overall, this chart suggests that Netflix's content library is diverse, catering to a variety of interests and preferences, including Bollywood movies, non-fiction programming, and international content.

7. Dashboard 1

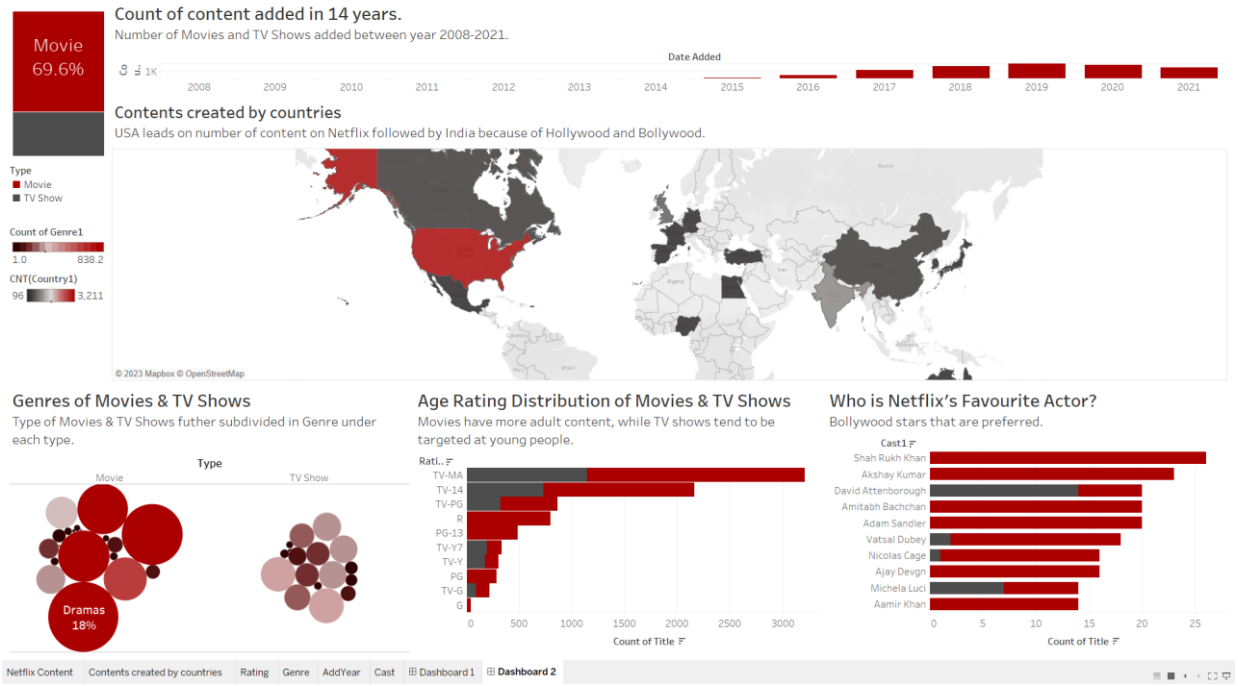


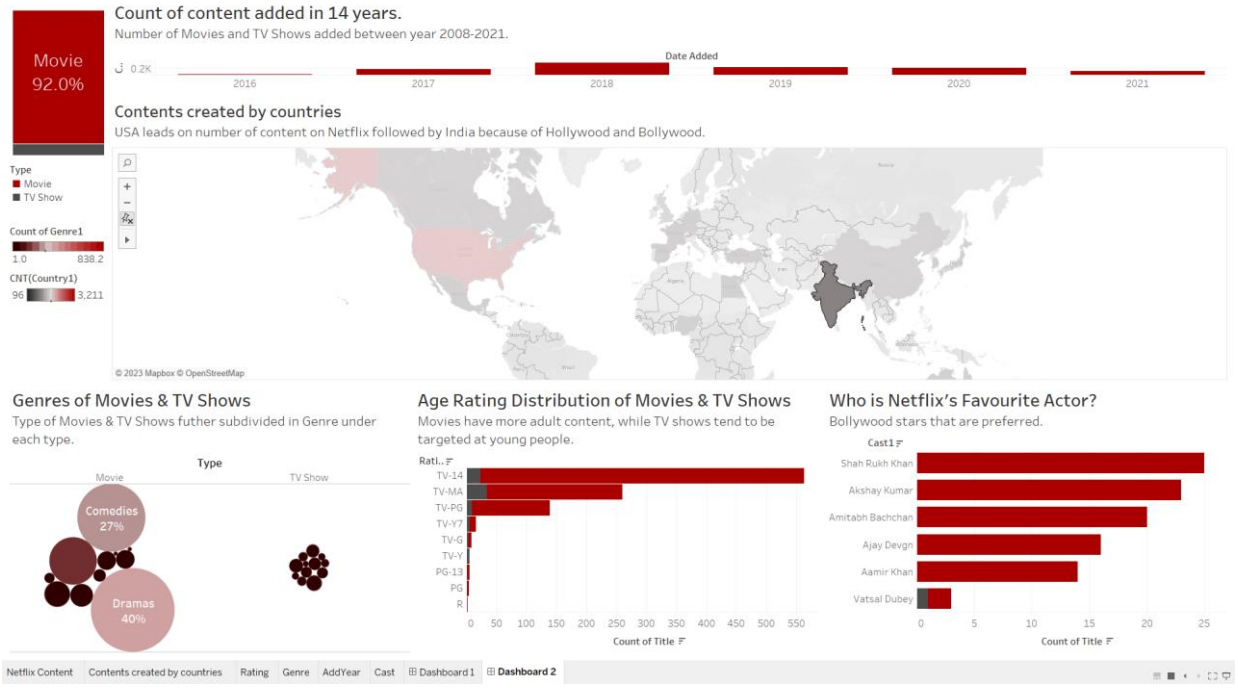
The above dashboard represents Netflix and its trends all over the world. This includes the Movies and TV Shows Ratio, Content Rating Added every Year, Genres of Movies and TV Shows, Contents Created by different Countries and its content, Netflix's Favourite Actor according to the movies released and finally, Age Rating Distribution of the Movies and TV Shows.

All the titles are graphically represented using various graphs and maps with the help of Tableau and Python Programming. The dashboard is constructed in a way, that when you click on one of the charts, it automatically displays the other charts with the variation.

For example, if you click on United States in the Tree Map created, one gets to see the Movie and TV Shows Ratio for the United States, the Content Rating Added every year so on and so forth.

8. Dashboard 2



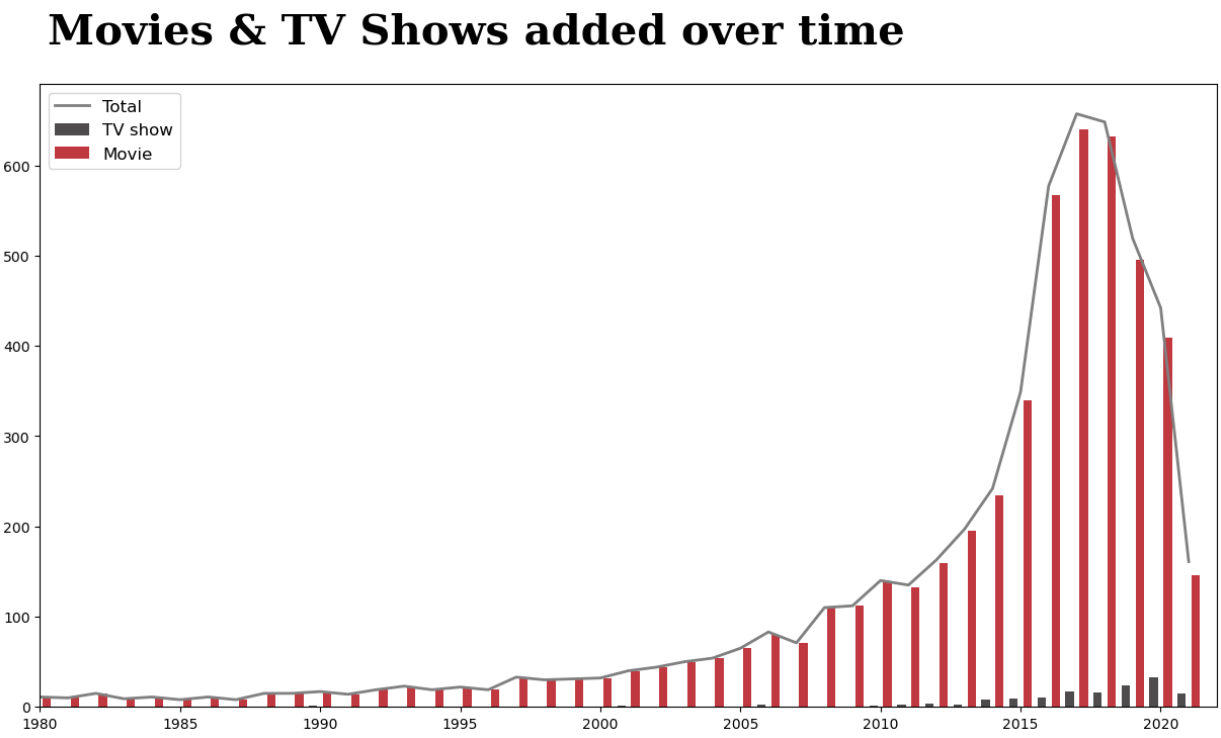


The dashboard 2 is constructed in kind of a similar way as of the first dashboard. This dashboard is interconnected with the other graphs plotted. Here, clicking on a particular country on the Geospatial Map, will show the projections of that country in the other charts and graphs outlined.

For example, if you click on India, it will show the different types of Genres of Movies and TV shows viewed in India. Similarly, one can check out all the other charts made and generate insights based on the observations made.

We have created visualizations using Python with a dataset similar to the one at hand, presented separately below.

9. Histogram Chart

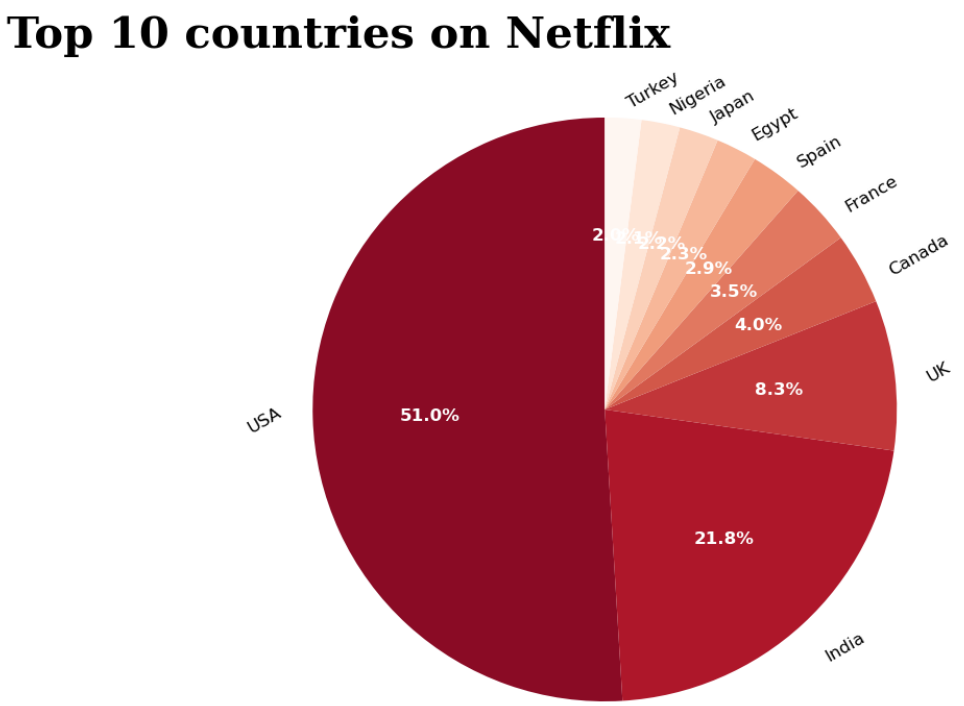


The histogram with years on the x-axis and count on the y-axis shows the number of movies and TV shows added to Netflix over time. The graph has three bars: one for the total count of both movies and TV shows, one for TV shows, and one for movies. The graph also shows a total line on the bar.

From the graph, we can see that the number of movies and TV shows from 1980s added to Netflix has been steadily increasing, with a significant increase in the number of TV shows from 2010 that are added. However, there appears to be a dip in the number of movies and TV shows added in 2020, which could be due to the COVID-19 pandemic or increased competition from other streaming services.

The graph also shows that the number of TV shows added to Netflix has surpassed the number of movies added in recent years. In addition, the graph shows that Netflix has been adding more TV shows and movies each year, indicating the company's continued growth in the streaming industry.

10. Pie-Chart

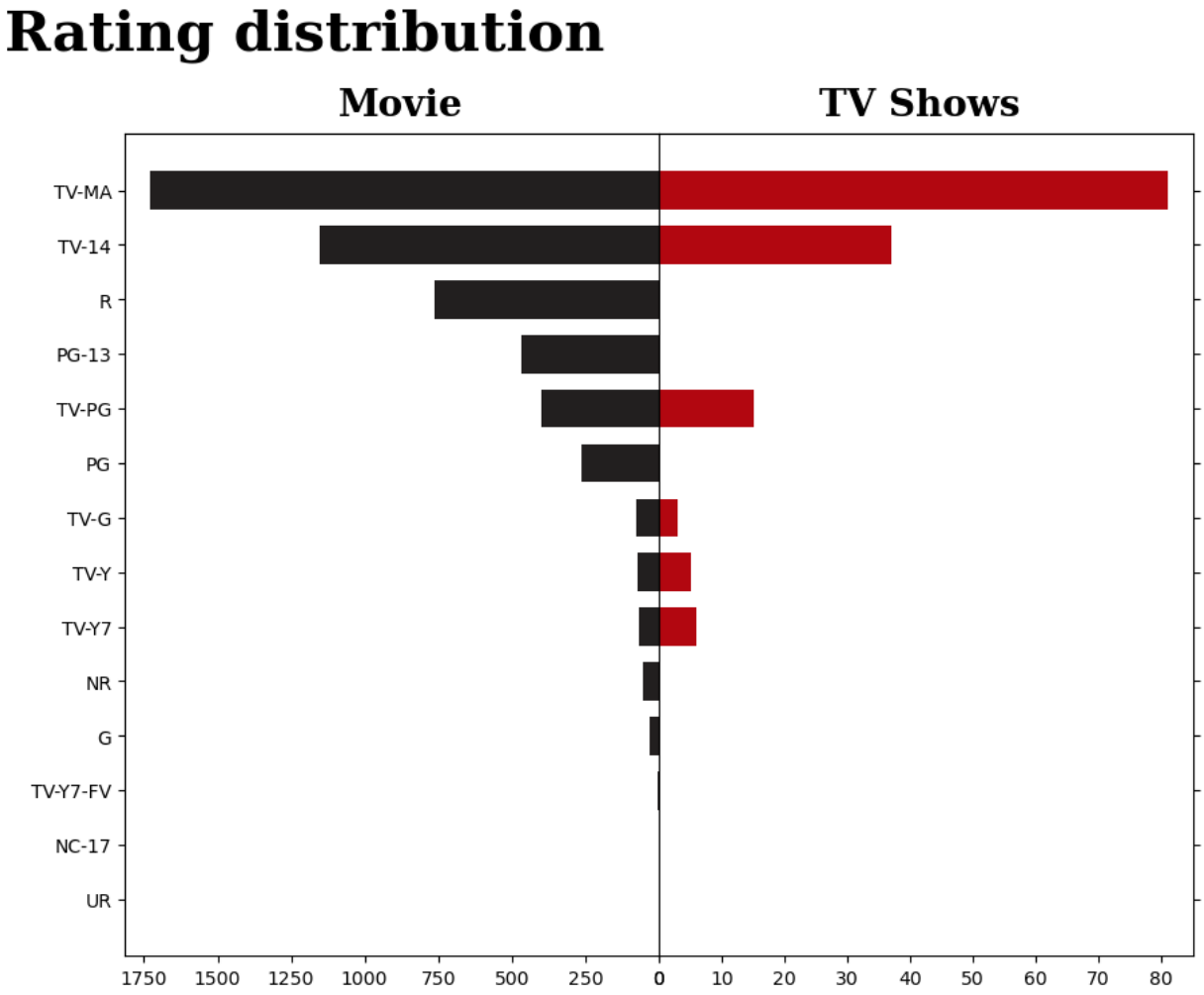


The pie chart shows the distribution of the top 10 countries on Netflix based on the number of titles available in each country. The highest percentage is in the USA with 51%, followed by India with 21.8%. One possible reason for the high percentage of titles in the USA could be that Netflix was launched there in 2007, and it is also the company's headquarters, so it is natural that it has a large library of content. As for India, it has a huge population and a thriving film industry, so it makes sense that there is a significant amount of content available for Indian audiences.

The UK has 8.3% of the total titles, which is relatively low compared to the USA and India. This may be due to the fact that Netflix was launched in the UK in 2012, and it faces stiff competition from established streaming services like Amazon Prime and the BBC's iPlayer. Canada and France have slightly lower percentages than the UK, but they are still significant. On the other hand, countries like Turkey and Nigeria have the lowest percentage of titles available, which may be due to their smaller populations and lesser demand for streaming services.

Looking at the Python code, it seems that the data has been grouped by the first country column, and the top 10 countries have been selected based on the number of titles available in each country. Then, the pie chart has been created using the Matplotlib library, with the labels, colors, and text properties customized for better visibility. The title has also been added to the chart to give context to the viewer.

11. Butterfly Chart



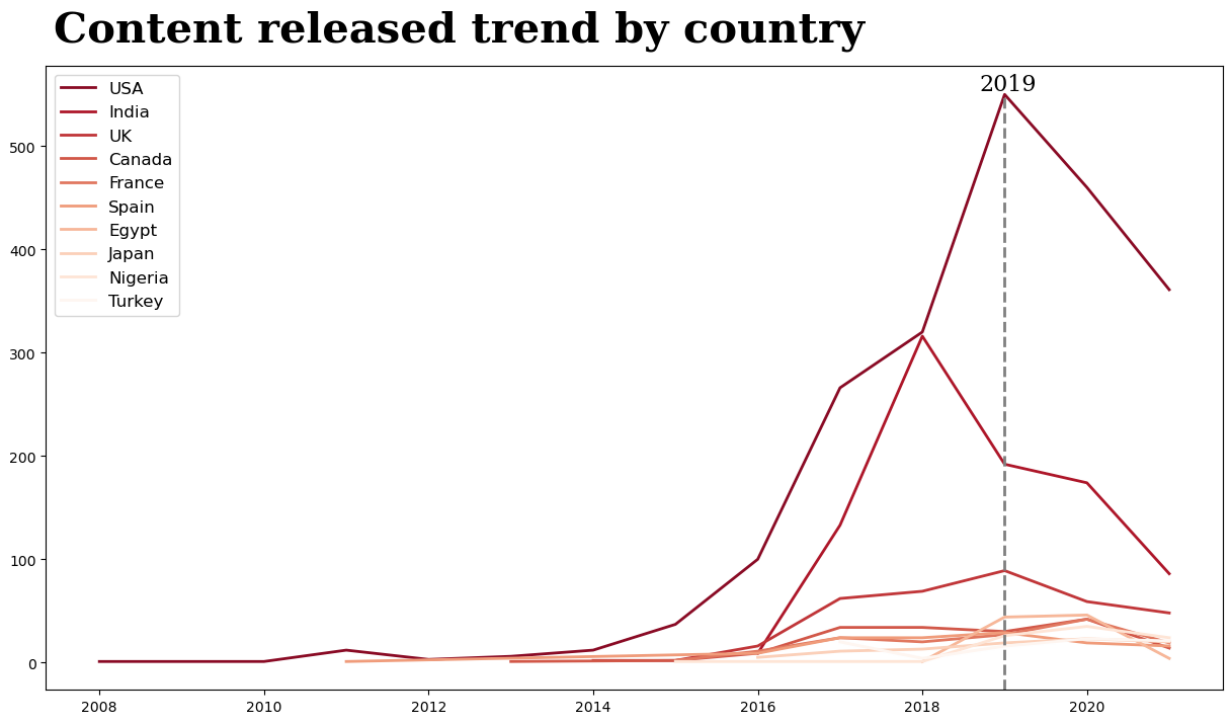
The butterfly chart compares the rating distribution of Movies and TV Shows on Netflix. The chart shows the count of movies and TV shows on the y-axis and the different ratings on the x-axis.

The left-hand side of the chart shows the distribution of movie ratings, with the TV-MA rating having the highest count and TV-Y7-FV having the lowest count of movies. TV-MA stands for "Television Mature Audience," and it indicates that the content is intended for mature audiences and may not be suitable for children under 17. R stands for Restricted, indicating that children under 17 require an accompanying parent or adult guardian. PG-13 means Parental Guidance is recommended for children under 13. PG is Parental Guidance suggested. G stands for General Audiences and is suitable for all ages. The NR rating means that no rating information is available for the content.

The right-hand side of the chart shows the distribution of TV show ratings, with TV-MA again having the highest count and TV-G having the lowest count. TV-Y stands for Television for Young Children, and it indicates that the content is suitable for children of all ages. TV-Y7 means that the program is suitable for children aged 7 and up. TV-G means General Audience and is suitable for all ages. TV-PG means Parental Guidance suggested, and TV-14 means that parents may find the program unsuitable for children under 14.

The chart shows that TV-MA is the most common rating for both movies and TV shows, indicating that mature content is in high demand on Netflix. It is interesting to note that while the distribution of movie ratings is relatively evenly spread out across different ratings, the distribution of TV show ratings is more skewed towards TV-MA and TV-14. Overall, the chart provides valuable insights into the types of content that are popular on Netflix and the preferences of its audience.

12. Line Chart



The chart shows the trend in content released by country over the years. The x-axis represents the years, and the y-axis represents the count of content released. The chart is represented by trend lines for each country with different colours.

The code first creates a list of top 10 countries based on the count of content added to Netflix. It then creates a color palette of 20 colors. The code then plots a line chart for each country in the list using the data for the year added and country. The color of the line is assigned based on the index of the country in the list of countries and the color palette. The code also adds a vertical line to indicate the year 2019 and a label for the same.

From the chart, we can observe that the USA has consistently been the highest content contributor over the years, with a peak in 2019. India is the second-highest content contributor, with a steep increase in content release from 2019. Other countries in the list such as the UK, Canada, France, and Japan have shown a relatively stable trend in content release. On the other hand, countries like Turkey and Nigeria have shown a slow and steady increase in content release over the years. The vertical line at 2019 indicates a significant increase in content release for many countries, including the USA and India. This may be due to increased investment in original content and a focus on global expansion by Netflix