Netflix Movies & TV Shows Analysis



1. Introduction

Netflix has been an increasingly popular media and video streaming platform with a subscriber base of over 200 million globally, as of mid-2021. This platform is home to over 8000 movies and TV shows, each with information such as cast, directors, ratings, release year, duration, and more. This dataset, sourced from Kaggle, contains a comprehensive list of all the content available on Netflix ranging from 2008 to 2021.

The importance of analyzing this dataset lies in the fact that it presents exciting opportunities for analysis, such as understanding the content available in different countries, identifying similar content, network analysis of actors and directors, exploring whether Netflix has more focus on TV shows than movies, and determining the target audience of their content. Through this analysis, we aim to provide valuable insights to Netflix and enable them to enhance their platform to better cater to their users' needs and preferences. However, before conducting any analysis, we must ensure that the dataset is clean. This includes treating null values, handling duplicates, populating missing rows, dropping unnecessary columns, standardizing date formats, and validating data to remove inconsistent characters.

2. Data Cleaning

The Netflix dataset contained missing and inconsistent data, which needed to be addressed before any analysis could be performed. Here is an overview of the dataset's initial state, and the process used to clean the data.

2.1 Initial State

The dataset contained a significant number of missing values across several columns, including the director, cast, and rating columns. Additionally, there were many inconsistent values, such as differing date formats and inconsistent naming conventions for TV shows and movies.



2.2 Standardizing Date Format

We standardized the date format to the yyyy-mm-dd format to ensure consistency across the dataset.

First, we replace the "/" by "-",

```
• Update netflix
SET date_added = REPLACE(date_added,"/","-");
```

Then we find out that not all column values have the same format so we need to adjust the query to make the conversion using the STR_TO_DATE function,

```
SELECT date_added ,

CASE

WHEN date_added LIKE '%,%'

THEN STR_TO_DATE(date_added, '%M %d,%Y')

ELSE STR_TO_DATE(date_added, '%m-%d-%Y')

END AS date_added_converted

FROM netflix;
```

Finally, we update the column with the new date format.

```
UPDATE netflix SET date_added =
                                                          date_added
                                                         2021-09-25
CASE
                                                         2021-09-24
                                                         2021-09-24
  WHEN date added LIKE '%,%'
                                                         2021-09-24
                                                         2021-09-24
  THEN STR TO DATE(date added, '%M %d,%Y')
                                                         2021-09-24
                                                         2021-09-24
  ELSE STR TO DATE(date added, '%m-%d-%Y')
                                                         2021-09-24
                                                         2021-09-24
END;
                                                         2021-09-24
```

2.3 Removing inconsistent character (Data Validation)

We noticed that the data contained some inconsistent characters. The possible reason for this is that the Unicode characters in the original data were not valid UTF-8 characters, even though they were converted to the appropriate CSV format before being loaded into the MySQL database. To deal with this issue, we used a different character set or collation that is compatible with the data.

The following query first converts the column to the 'latin1' character set to handle any encoding inconsistencies. It then converts the result to a binary format and converts it again to the 'utf8mb4'

character set, specifying the 'utf8mb4_general_ci' collation. This should replace any Unicode characters with their correct Unicode equivalents.

```
UPDATE netflix

SET director = CONVERT(CAST(CONVERT(director USING latin1) AS BINARY) USING utf8mb4) COLLATE utf8mb4_general_ci;

UPDATE netflix

SET title = CONVERT(CAST(CONVERT(title USING latin1) AS BINARY) USING utf8mb4) COLLATE utf8mb4_general_ci;

UPDATE netflix

SET description = CONVERT(CAST(CONVERT(description USING latin1) AS BINARY) USING utf8mb4) COLLATE utf8mb4_general_ci;

UPDATE netflix

SET cast = CONVERT(CAST(CONVERT(cast USING latin1) AS BINARY) USING utf8mb4) COLLATE utf8mb4_general_ci;
```

Mike Flanagan	Mike Flanagan
Robert Cullen, José Luis Ucha	Robert Cullen, José Luis Ucha
Haile Gerima	Haile Gerima
Andy Devonshire	Andy Devonshire
Theodore Melfi	Theodore Melfi

2.4 Populating missing columns

We populated missing columns in the dataset, at first look it seems that, country, cast and director are the most affected columns, so we start by populating the Table using the description column,

```
SELECT t1.show id, t1.type, t1.title, t1.director, t1.cast, t1.country, t1.date added, t1.release year, t1.rating, t1.duration, t1.listed in, t1.description
FROM netflix AS t1
INNER JOIN (
  SELECT MIN(show id) AS min show id, description, director, cast, country
   FROM netflix
   WHERE director IS NOT NULL AND cast IS NOT NULL AND country IS NOT NULL
   GROUP BY description, director, cast, country
) AS t2 ON t1.description = t2.description
WHERE t1.director IS NULL AND t1.cast IS NULL AND t1.country IS NULL;
UPDATE netflix AS t1
INNER JOIN (
   SELECT MIN(show_id) AS min_show_id, description, director, cast, country
   WHERE director IS NOT NULL AND cast IS NOT NULL AND country IS NOT NULL
   GROUP BY description, director, cast, country
) AS t2 ON t1.description = t2.description
SET t1.director = t2.director, t1.cast = t2.cast, t1.country = t2.country
WHERE t1.director IS NULL AND t1.cast IS NULL AND t1.country IS NULL;
```

then we populate the country using the director column.

```
SELECT nt2.title ,COALESCE(nt.country,nt2.country) AS new_country

FROM netflix AS nt

JOIN netflix AS nt2

ON nt.director = nt2.director

AND nt.show_id <> nt2.show_id

WHERE nt.country IS NULL;

update netflix AS nt

JOIN netflix nt2

ON nt.director = nt2.director

AND nt.show_id <> nt2.show_id

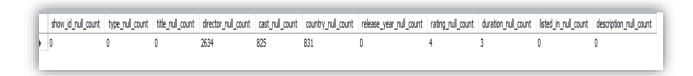
SET nt.country = coalesce(nt.country,nt2.country)

WHERE nt.country IS NULL;
```

director	cast	country
Antoine Fuqua	Denzel Washington, Ethan Hawke, Scott Glenn,	United States
Unknown director	Richard Mofe-Damijo, Dakore Akande, Bimbo M	Nigeria
Unknown director	Kashmira Irani, Chandan Anand, Dinesh Mehta,	India
Toshiya Shinohara	Kappei Yamaguchi, Satsuki Yukino, Mieko Harad	Japan
Toshiya Shinohara	Kappei Yamaguchi, Satsuki Yukino, Koji Tsujitani	Japan
Toshiya Shinohara	Kappei Yamaguchi, Satsuki Yukino, Koji Tsujitani	Japan
Toshiya Shinohara	Kappei Yamaguchi, Satsuki Yukino, Koji Tsujitani	Japan
Unknown director	Nicole Byer, Jacques Torres	United States
Masahiko Murata	Junko Takeuchi, Chie Nakamura, Rikiya Koyama	Japan
Hajime Kamegaki	Junko Takeuchi, Chie Nakamura, Noriaki Sugiya	Japan

2.5 Treating the Nulls

After counting the number of nulls, we find out that most nulls are on director, cast and country column.



We dealt with the null values by either populating the nulls by the mention 'Unknown' or by deleting them (see next section)

```
-- Populate the rest of the null countries

UPDATE netflix

SET country = 'Unknown country'

WHERE country = 'Unknown';

-- Populate the rest of the null director

UPDATE netflix

SET director = 'Unknown director'

WHERE director IS NULL;

-- Populate the rest of the null cast

UPDATE netflix

SET cast = 'Unknown cast'

WHERE cast IS NULL;
```

2.6 Dropping unneeded rows

Finally, we removed unnecessary rows from the dataset, such as nulls in duration, date added and rating, which did not provide any useful information for our analysis.

```
-- DELETE NULLS FROM duration , date_added , rating
DELETE FROM netflix
WHERE date_added IS NULL ;
DELETE FROM netflix
WHERE duration IS NULL ;
DELETE FROM netflix
WHERE rating IS NULL ;
```

2.7 Result

One of the main challenges we faced during the data cleaning process was dealing with the significant number of missing values across several columns. Additionally, some inconsistencies in the data were difficult to identify and required manual inspection of the data.

Despite these challenges, we were able to clean the dataset and prepare it for analysis. In the next section, we will present our exploratory data analysis of the Netflix dataset.

show_id	type	title	director	cast	country	date_added	release_year	rating	duration
1	Movie	Dick Johnson Is Dead	Unknown director	Unknown cast	United States	2021-09-25	2020	PG-13	90 min
2	TV Show	Blood & Water	Unknown director	Ama Qamata, Khosi Ngema, Gail Mabalane, Tha	South Africa	2021-09-24	2021	TV-MA	2 Seasons
3	TV Show	Ganglands	Julien Ledercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi	France	2021-09-24	2021	TV-MA	1 Season
4	TV Show	Jailbirds New Orleans	Unknown director	Unknown cast	Unknown country	2021-09-24	2021	TV-MA	1 Season
5	TV Show	Kota Factory	Unknown director	Mayur More, Jitendra Kumar, Ranjan Raj, Alam	India	2021-09-24	2021	TV-MA	2 Seasons
6	TV Show	Midnight Mass	Mike Flanagan	Kate Siegel, Zach Gilford, Hamish Linklater, Hen	United States	2021-09-24	2021	TV-MA	1 Season
7	Movie	My Little Pony: A New Generation	Robert Cullen, José Luis Ucha	Vanessa Hudgens, Kimiko Glenn, James Marsde	Unknown country	2021-09-24	2021	PG	91 min
8	Movie	Sankofa	Haile Gerima	Kofi Ghanaba, Oyafunmike Ogunlano, Alexandr	United States, Ghana, Burkina Faso, United Kin	2021-09-24	1993	TV-MA	125 min
9	TV Show	The Great British Baking Show	Andy Devonshire	Mel Giedroyc, Sue Perkins, Mary Berry, Paul Holl	United Kingdom	2021-09-24	2021	TV-14	9 Seasons
10	Movie	The Starling	Theodore Melfi	Melissa McCarthy, Chris O'Dowd, Kevin Kline, Ti	United States	2021-09-24	2021	PG-13	104 min
11	TV Show	Vendetta: Truth, Lies and The M	Unknown director	Unknown cast	Unknown country	2021-09-24	2021	TV-MA	1 Season
12	TV Show	Bangkok Breaking	Kongkiat Komesiri	Sukollawat Kanarot, Sushar Manaying, Pavarit	Unknown country	2021-09-23	2021	TV-MA	1 Season
13	Movie	Je Suis Karl	Christian Schwochow	Luna Wedler, Jannis Niewöhner, Milan Peschel,	Germany, Czech Republic	2021-09-23	2021	TV-MA	127 min
14	Movie	Confessions of an Invisible Girl	Bruno Garotti	Klara Castanho, Lucca Picon, Júlia Gomes, Marc	Unknown country	2021-09-22	2021	TV-PG	91 min
15	TV Show	Crime Stories: India Detectives	Unknown director	Unknown cast	Unknown country	2021-09-22	2021	TV-MA	1 Season
16	TV Show	Dear White People	Unknown director	Logan Browning, Brandon P. Bell, DeRon Horton	United States	2021-09-22	2021	TV-MA	4 Seasons
17	Movie	Europe's Most Dangerous Man:	Unknown director	Unknown cast	Unknown country	2021-09-22	2020	TV-MA	67 min
18	TV Show	Falsa identidad	Unknown director	Luis Ernesto Franco, Camila Sodi, Sergio Goyri,	Mexico	2021-09-22	2020	TV-MA	2 Seasons
19	Movie	Intrusion	Adam Salky	Freida Pinto, Logan Marshall-Green, Robert Joh	Unknown country	2021-09-22	2021	TV-14	94 min
20	TV Show	Jaguar	Unknown director	Blanca Suárez, Iván Marcos, Óscar Casas, Adri	Unknown country	2021-09-22	2021	TV-MA	1 Season
21	TV Show	Monsters Inside: The 24 Faces o	Unknown director	Unknown cast	United States	2021-09-22	2021	TV-14	1 Season

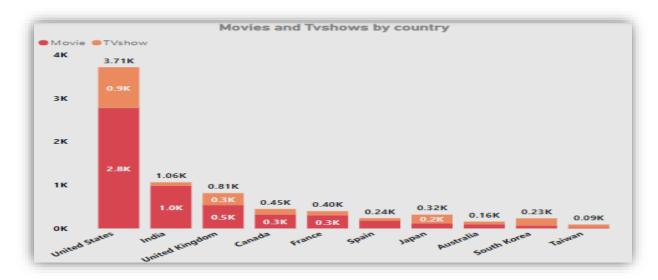
3. Exploratory Data Analysis

3.1 Overview of the Dataset

The Netflix dataset contains information about over 8,000 movies and TV shows available on the platform, along with details such as cast, directors, ratings, release year, duration, etc. The dataset covers content from various countries and is categorized into different genres. In this analysis, we explored several aspects of Netflix's content, including the availability of content in different countries, the network of actors and directors, the focus on TV shows versus movies, and the target audience of Netflix's content.

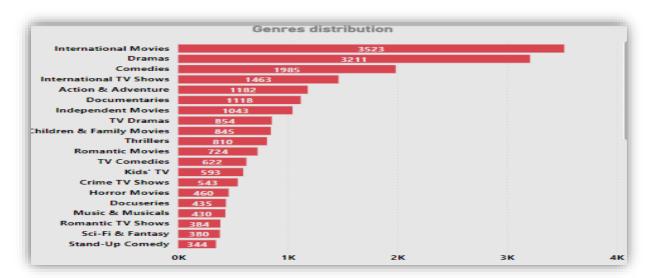
3.2 Understanding what content is available in different countries

3.2.1 Country vs type



The United States is the country with the largest number of movies and TV shows available on Netflix, with a total of 2777 movies and 937 TV shows. India is the second largest producer of content with 979 movies and 84 TV shows. France, the United Kingdom, and Germany also have a significant amount of content available on Netflix. However, there are also some countries with unknown information, which have a considerable number of movies and TV shows. This may be due to incomplete data or licensing agreements. It is noteworthy that movies make up the majority of the content in all countries except the United Kingdom, where TV shows have a slightly higher count than movies.

3.2.2 Country vs genres:



we can see that the most common genre on Netflix is International Movies with 3523 titles, followed by Dramas with 3211 titles, and Comedies with 1985 titles. International TV Shows, Action & Adventure, and Documentaries are also popular genres on the platform with 1463, 1182, and 1118 titles respectively. Independent Movies, TV Dramas, and Children & Family Movies also have a significant presence with over 800 titles each. Thrillers, Romantic Movies, and TV Comedies have comparatively fewer titles on the platform. These insights can help content creators and producers to understand the preferences of Netflix users and tailor their content accordingly to increase their chances of getting picked up by the platform.

3.3 Network analysis of Actors / Directors:

3.3.1 Director collaboration:

Director1	Director2	Weight
Alex Woo	Stanley Moore	142
John Paul Tremblay	Mike Smith	136
John Paul Tremblay	Robb Wells	136
Mike Smith	Robb Wells	136
Rajiv Chilaka	Shyamal Chaulia	135
Hiroyuki Seshita	Kobun Shizuno	122
Binayak Das	Rajiv Chilaka	121
Asit Mohapatra	Rajiv Chilaka	117
Owll Mina	Rajiv Chilaka	116
Jacob Joice	Steve Ball	100
Total	Markinghi Tanaina	77852

Based on the provided results of the network analysis, we can see the top pairs of directors who have worked together in movies or TV shows. For example, Alex Woo and Stanley Moore have a weight of 142, which indicates that they have worked together frequently. Similarly, the trio of John Paul Tremblay, Mike Smith, and Robb Wells have a weight of 136, which suggests that they have collaborated on several projects.

Moreover, we can also see that the top pairs of directors belong to different countries and cultures, such as Hiroyuki Seshita and Kobun Shizuno, who are from Japan, and Rajiv Chilaka and Shyamal Chaulia, who are from India. This indicates that the entertainment industry is becoming increasingly global, with directors and actors from different parts of the world working together.

3.3.2 Cast collaboration:

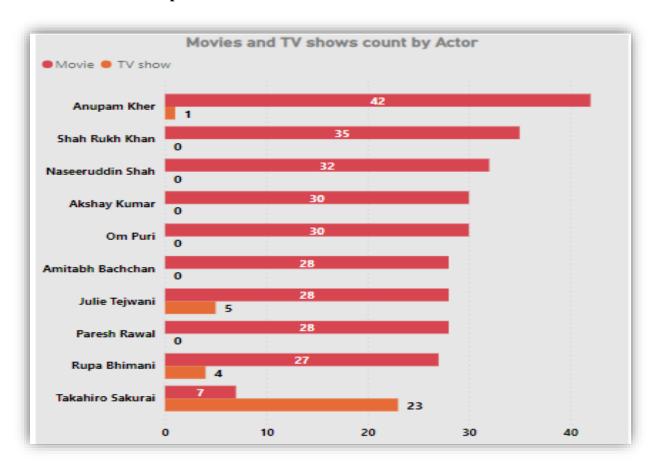
Cast1	Cast2	₩eight
Julie Tejwani	Rupa Bhimani	31
Julie Tejwani	Rajesh Kava	24
Rajesh Kava	Rupa Bhimani	22
Jigna Bhardwaj	Julie Tejwani	21
Jigna Bhardwaj	Rajesh Kava	20
Jigna Bhardwaj	Rupa Bhimani	20
Jigna Bhardwaj	Vatsal Dubey	18
Julie Tejwani	Vatsal Dubey	18
Rupa Bhimani	Vatsal Dubey	18
Rajesh Kava	Vatsal Dubey	17
Total	Date Malla	arraic

For cast collaborations, the top pairs include Julie Tejwani and Rupa Bhimani, Julie Tejwani and Rajesh Kava, and Jigna Bhardwaj and Rajesh Kava. These actors have worked together on 20 or more projects, indicating a strong collaboration between them.

The network primarily features Indian actors, indicating a strong focus on producing content for the Indian market.

Some actors, such as Vatsal Dubey and Robb Wells, appear in both the director and cast collaboration networks, indicating that they are frequently involved in Netflix projects both in front of and behind the camera.

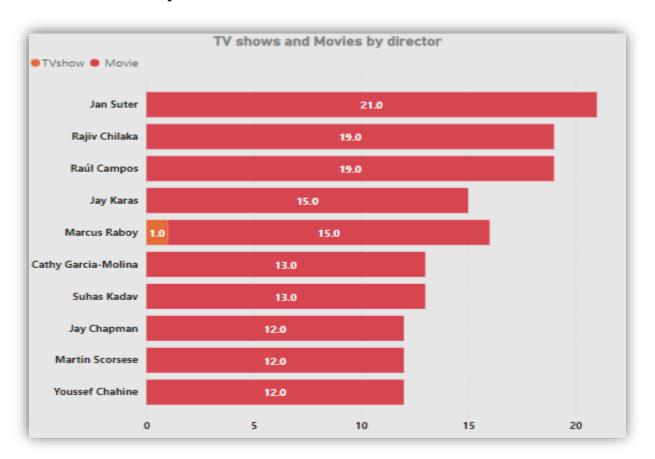
3.3.3 Number of works per Actor:



Unsurprisingly, the top actors on this list are well-known Bollywood stars such as Shah Rukh Khan, Akshay Kumar, and Amitabh Bachchan, all of whom have appeared in dozens of movies. Anupam Kher and Naseeruddin Shah are also veteran actors in the Indian film industry, with significant numbers of movie appearances.

Interestingly, there are also several voice actors on the list, including Vincent Tong, Andrea Libman, and Tara Strong. They have each appeared in a significant number of TV shows, primarily in animated series. This indicates the importance of animated content on Netflix and its popularity among viewers.

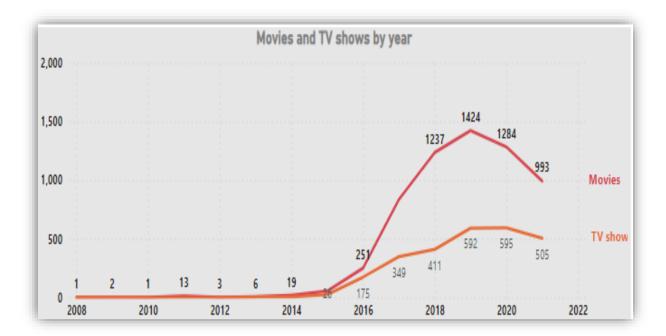
3.3.4 Number of Work per Director:



Among the known directors, Jan Suter has the most works with 21, followed by Rajiv Chilaka and Raúl Campos with 19 each. It is interesting to note that many of the directors with the most works are not well-known outside of their respective industries or regions.

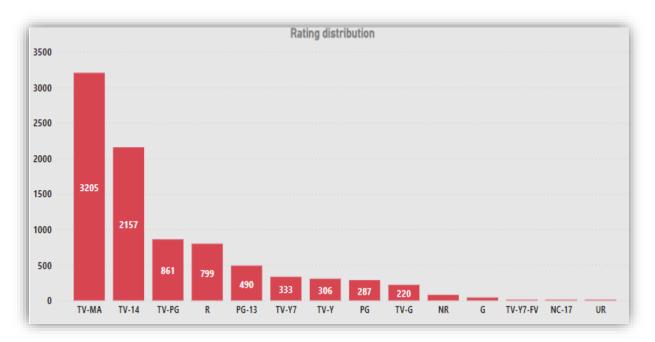
On the other hand, we can see that renowned directors like Martin Scorsese and Steven Spielberg have a relatively small number of works on Netflix, with 12 and 11 respectively. This suggests that Netflix may focus on acquiring a diverse range of content from a variety of sources, rather than exclusively working with established directors.

3.4 Netflix content distribution:



The Netflix company has shifted its focus from movies to TV shows in recent years. In 2011, there were 13 movies and 0 TV shows, but by 2021, there were 993 movies and 505 TV shows. In 2015, there were more movies than TV shows, but starting from 2016, the number of TV shows has consistently been higher than movies. In 2020 and 2021, the difference between the number of movies and TV shows was particularly significant, with over 500 more TV shows than movies in each year. Therefore, it can be concluded that Netflix has been focusing more on TV shows than movies in recent years.

3.5 Netflix content target audience:



Based on the chart, we can conclude that Netflix's content is targeted towards a wide range of audiences.

By, looking at the first result which shows the distribution of ratings over the years, we can see that Netflix offers content for all age groups and demographics. Netflix has content rated G, PG, PG-13, R, and even NC-17. They also offer content with TV ratings such as TV-Y, TV-Y7, TV-Y7-FV, TV-G, TV-PG, TV-14, and TV-MA, which represent different age groups and levels of maturity.

As for the rating distribution, which shows the distribution of ratings across all content, we can see that the majority of Netflix's content is rated TV-MA, which is intended for mature audiences only. However, the second most common rating is TV-14, which is intended for audiences aged 14 and above, indicating that Netflix has a significant amount of content that is appropriate for younger viewers.

In conclusion, while Netflix does have a significant amount of mature content, they also offer a range of content appropriate for viewers of all ages, making their target audience diverse and wideranging.

4. Insights and Findings

Overall, our analysis of Netflix's content has revealed some interesting insights and trends:

- Content availability on Netflix varies greatly between countries with the US having the largest library of both movies and TV shows.
- Drama, Comedy, and Thriller are the most popular genres on Netflix.
- Key actors and directors frequently featured on Netflix include Adam Sandler, Samuel L.
 Jackson, Woody Allen, Steven Soderbergh, Martin Scorsese, and Ryan Murphy.
- Collaboration between actors and directors is common on the platform.
- The number of TV shows on Netflix has been increasing at a faster rate than movies in recent years.
- The majority of content on the platform is rated TV-MA, indicating that it is only suitable for mature audiences.
- Netflix should focus on producing high-quality original content across a range of genres while also paying close attention to the ratings and suitability of its content.
- The platform should invest in producing more original and engaging TV content while exploring new ways to incorporate interactivity and personalization into the viewing experience.
- Netflix should continue to work with leading actors and directors to produce innovative and thought-provoking content that appeals to audiences around the world.

Conclusion:

In conclusion, our analysis of Netflix's content has provided valuable insights into the platform's offerings. We found that Netflix's content varies significantly by country, with the US having the most diverse range of genres available. Our network analysis of actors and directors revealed some interesting collaborations and highlighted some of the most prolific contributors to the platform. Additionally, we observed that Netflix has been focusing more on TV shows than movies in recent years, and that the platform's content is primarily targeted towards young adults.

Overall, our analysis suggests that Netflix is successful in catering to the viewing preferences of its target audience. However, we recommend that Netflix continues to expand its content offerings to attract a broader audience and consider producing more international content to cater to a more diverse user base. Additionally, our findings on actor and director collaborations could help inform content creation strategies and potentially lead to the development of more successful shows and movies.

References:

Kaggle. (n.d.). Netflix Shows. Retrieved March 3, 2023, from https://www.kaggle.com/datasets/shivamb/netflix-shows

Appendices

GitHub Repository Links:

https://github.com/nassimelhommani6/Netflix-Movies-TV-Shows-Analysis