Analysis of Netflix Movies & TV Shows

Aayush Kumar Gupta
Computer Science & Engineering
Meghnad Saha Institute of Technology (Aff. To MAKAUT) Kolkata, India
aayushkumargupta3@gmail.com

Yash Thakur
Computer Science & Engineering
Meghnad Saha Institute of Technology (Aff. To MAKAUT) Kolkata, India
yashthakur0421@gmail.com

Anindya Chatterjee
Computer Science & Engineering
Meghnad Saha Institute of Technology (Aff. To MAKAUT) Kolkata, India
silkcandy305@gmail.com

Kaustav Lahiri
Computer Science & Engineering
Meghnad Saha Institute of Technology (Aff. To MAKAUT) Kolkata, India
kaustavlahiri5@gmail.com

Sourav Singha
Computer Science & Engineering
Meghnad Saha Institute of Technology (Aff. To MAKAUT) Kolkata, India
singhasourav632@gmail.com

Shreya Choudhury
Computer Science & Engineering
Meghnad Saha Institute of Technology (Aff. To MAKAUT) Kolkata, India
choudhuryshreya276@gmail.com

Abstract -: *Exploring datasets of Netflix for Future Release of TV shows and Movies on the Platform. In this project, we are going to explore the dataset from Kaggle and we would like to find out how long the Netflix platform takes a movie or a TV show to release on its platform, how many movies and TV shows are released in specific time frame, how many movies and TV shows are release in the recent ten years on the platform, and what were the top 10 genres that the audience of the Netflix platform liked the most. From here, we would like to apply a machine learning approach to understand the data fully and provide a great solution where the platform should be headed to. From our data analysis we conducted on R markdown, we have discovered that there were a wide variety of genres that movie directors produced worldwide and we have observed many cast members and genres they were in.*

Project Link -: *click here*

Motivation -:

Netflix is the largest online movie and TV show streaming service on the planet. Its service is widely available in many countries including but not limited to the United States, India, South Korea, Japan, and many more. The service was first introduced as a DVD rental service on the Internet and later, the founder and CEO of the company Reed Hastings transitioned to a revolutionary way of delivering movies and TV shows through its website allowing many users to directly stream their favorite contents on their Internet enabled devices including desktop computers, laptops, tablet PCs, mobile phones, and many more. With it's a whole new approach of delivering shows and movies, the sales of Netflix went up exponentially. Since then, the platform created its own recommender system to understand what types of movies and TV shows the users would like to watch,

what kind of style of cinematography they liked the most, and how they consume their favorite TV shows. With such analysis, the company released 'The House of Cards', which became a huge success in the history of streaming service providers. With the power of data analysis, more users were attracted to the platform, and many users tend to spend most of their time watching shows and movies on Netflix. With this approach, we would like explore the dataset to understand the trend of movies and TV shows on Netflix. To analyze Starbucks' stock performance, we collected historical stock data from reliable sources, ensuring data consistency and completeness. The data was then preprocessed to handle any missing values or irregularities.

Introduction -:

We first wanted to get an overview of the dataset that we were dealing with. First, we loaded up tidy verse for a simple data analysis purpose. We got the dataset from Kaggle and we are going to utilize data that the Kaggle website provides to understand the trend of movies and TV shows released on the platform. This dataset consists of 1 From the code, we could see the column names that the CSV file contains. We will utilize the following columns to understand what movies and TV shows were released in specific year, what genres they were, date when they were released and the rating the audience gave and so on. From the column names, we could observe that there are twelve columns: show_id, type, title, director, cast, country, date_added, release_year, rating, duration, listed_in, description. We will first demonstrate the overview of the dataset. Each column contains 8807 in length except for release year in interquartile range.

```
summary(df_netflix)

##    show_id             type              title             director
##  Length:8807        Length:8807        Length:8807        Length:8807
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##      cast             country           date_added          release_year
##  Length:8807        Length:8807        Min.   :2008-01-01   Min.   :1925
##  Class :character   Class :character   1st Qu.:2018-04-20   1st Qu.:2013
##  Mode  :character   Mode  :character   Median :2019-07-12   Median :2017
##                                        Mean   :2019-05-23   Mean   :2014
##                                        3rd Qu.:2020-08-26   3rd Qu.:2019
##                                        Max.   :2021-09-25   Max.   :2021
##                                        NA's   :98
##     rating            duration          listed_in          description
##  Length:8807        Length:8807        Length:8807        Length:8807
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
```

From the overview of the dataset, we could tell that the earliest year recorded as the movie or the TV show is 1925 and the latest content that is on the platform is 2021. With the information of directors, actors/actresses, movies, TV shows, rating, duration of each content and more, we can understand how long does Netflix take to upload the content on the platform, which genre is most popular on the platform, which actor/actresses are most popular and how genre popularity changed the movie or TV shows that an actor/actress appear on the genre over time. We will explore the trend of Netflix using R to figure it out.

Project Background

The report presents a data analysis research and coordination activity one in the Netflix dataset found on Kaggle to establish a new data exploration on the trend of movies on the platform. When Netflix was first 2 launched, it did not have any data analysis to understand the trend of audience/users using the platform as previously mentioned. As time passes by, the importance of utilizing data analysis started to emerge, and the biggest hit shows were released on the platform after scrutinizing the data the company collected from the users. In the project, the team of three is working on a same procedure to understand the trend of the platform and attempt to understand the average time the platform takes to upload the contents, the top ten film directors in top ten countries where the platform is streaming, and the top genres that the top actors/actresses are in. The team will, then, attempt to understand the overall trend of the Netflix platform according to the dataset the team got from Kaggle.

Methodology -:

1. *Dataset Description*

This dataset is from Kaggle, and Netflix is the collector of this data. Shivam Bansal created and uploaded this dataset to Kaggle, and the dataset is public domain. This dataset includes lists of all the movies and TV shows that are available on flixable, which is a third-party search engine on Netflix. This dataset has 8,807 rows and 12 variables. It includes information about the genre, cast, director, countries where the show is available to watch, and a summary of what the show is about. The columns and their descriptions are as listed below:

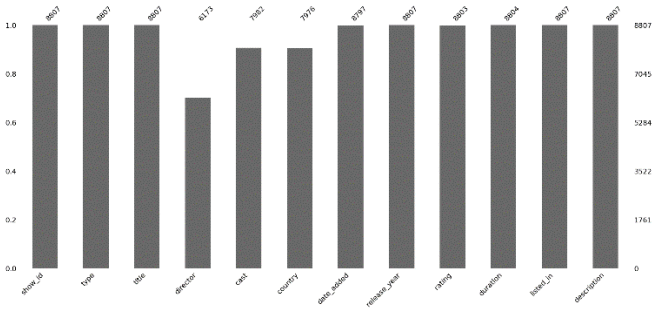
- SHOW ID: Unique ID of each show
- TYPE: Show category. Could be either a Movie or a TV Show.
- TITLE: Name of the show
- DIRECTOR: Name of the director(s) of the show
- CAST: Names of actors/actresses in the show
- COUNTRY: Countries where the show is available to watch on Netflix
- DATE ADDED: Date when the show was added on Netflix
- RATING: Show rating on Netflix
- RELEASE YEAR: Release year of the show
- DURATION: Time duration of the show
- LISTED IN: Genre of the show
- DESCRIPTION: Brief insight into what the show is about

Data Quality

The purpose of this section is to describe the quality of the dataset. Missingness, strengths, and weaknesses will be discussed. A number of tables and figures will be presented to facilitate cohort description, illustrate the distributions of key variables, and present important results.

Strengths

This data has been collected for the analysis based on the movies and series. It is a tidy dataset which means that each variable has its own column, each observation has its own row, and each value has its own cell, so it is appropriate for most analysis. Another point of strength is that the dataset has both categorical and continuous variables, so we can investigate the information in a categorized way and report the information of each category separately, making informative plots about the data. For example, the description variable can be used to find similar movies and TV shows using the text similarities for further analysis. The sample size is another strength of this dataset that makes it possible to analyze and compare the data.



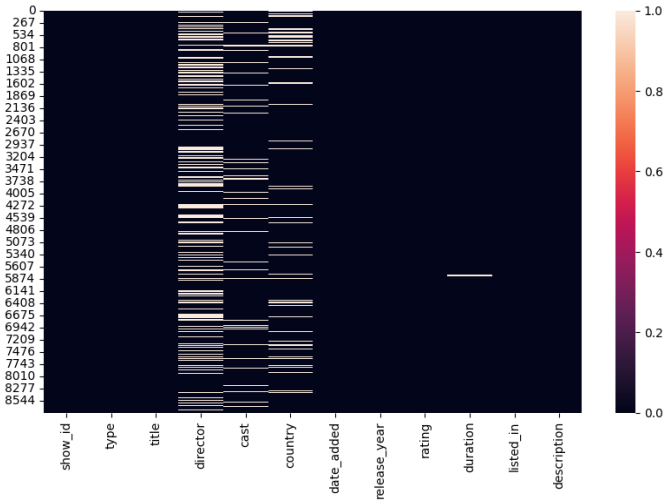
Weaknesses

There are some missingness in the data set, as seen in **Table 1**, requiring some data cleaning. The data cleaning process involved identifying incorrect, incomplete, inaccurate, irrelevant, or missing pieces of data and modifying, replacing, or deleting them as needed. This table shows that only the categories of director, cast, country, date added, rating, and duration have missing values and the category of director has the most missing values. There are 2,634 cases where director has missing values. There are 831 cases where country has missing values. There are 825 cases where cast has missing values. There are 17 cases where date added, rating, and duration have missing values. Another weakness in this dataset is data duplication. However, this could be fixed by merging duplicate data into a single value.

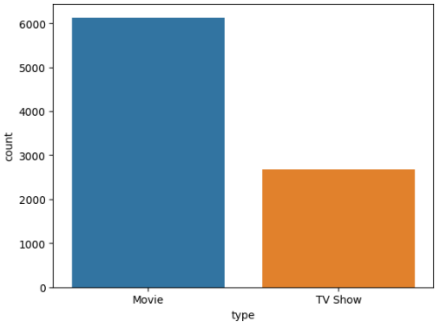
```
[ ]  #Checking for null values in dataset
     netflix.isnull().sum()

     show_id          0
     type             0
     title            0
     director      2634
     cast           825
     country        831
     date_added      10
     release_year     0
     rating           4
     duration         3
     listed_in        0
     description      0
     dtype: int64
```

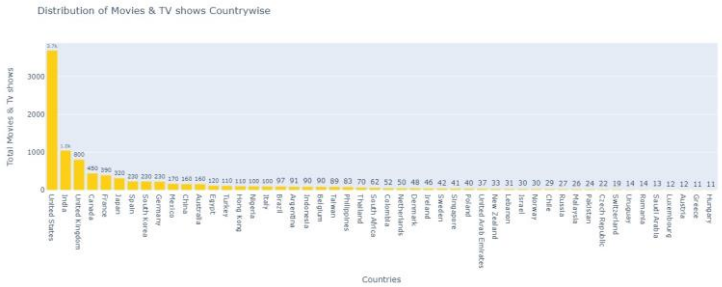
As we can see that in the Director, cast, Country, Date_added, Rating and Duration columns has null values so, we will show this with the help of heatmap.



As we mentioned before Netflix is the largest online movie and TV show streaming service on the planet. Its service is widely available in many countries including but not limited to the United States, India, South Korea, Japan, adman more.

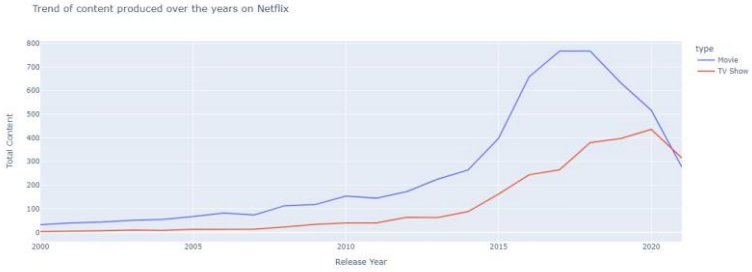


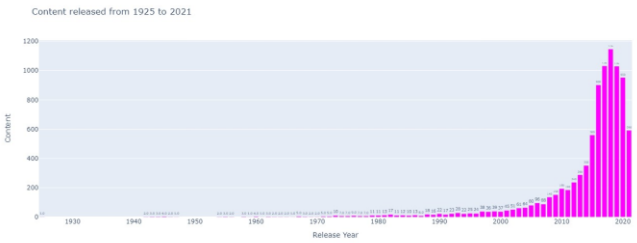
We visuliased top 10 countries. **The below figure** the amount of media (including TV shows and movies) in different countries. We can see that the United States is the #1 leader in the amount of content on Netflix, and India is in 2nd place. For this plot we use the fct_reorder() to reorder factor levels by sorting another variable. Also we use the fct_lump() to Lump together factor levels into "other" and finally we use the fct_explicit_na()for the missing values as we mentioned before we have some missingness. The reason that I chose bar charts because they are good choice for showing the relationship between numeric and categorical variables.



The trends of TV Shows and Movies over the years -:

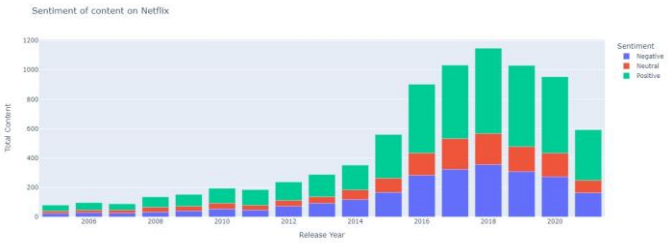
we want to know that Is there any relationship between the Country and Duration of Movies? At the first, we hypothesized that the time duration of movies are the same among the countries. I selected top 10 countries to see the relationships between movies and their duration. Below figure shows the duration of movies in 10 countries. It can be seen from this plot that Movies produced in India tend to be the longest on average with the average duration of 127 min. We created the raw points plot with the summary showing the mean duration. The purpose is showing both raw data and a summary.



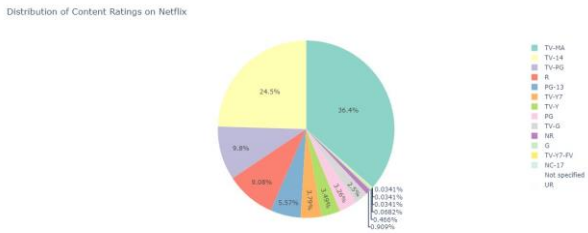


As we mentioned before in the overview of the dataset, we can see (that the earliest year recorded as the movie or the TV show is 1925 and the latest content that is on the platform is 2022. **This figure** indicates that most of the movies were released after the year 2000s. As can be seen from the two graphs, from 2016 to 2019 was the most titles were added to the streaming network the spike was in 2019 and of course, 2020 following 2019, remember here we are talking about titles added not the viewer's numbers. 2020 Covid-19 pandemic has stopped the increase, indeed due to decrease of Movies and TV-Shows production!

Let's analyze the sentiment of content on Netflix:

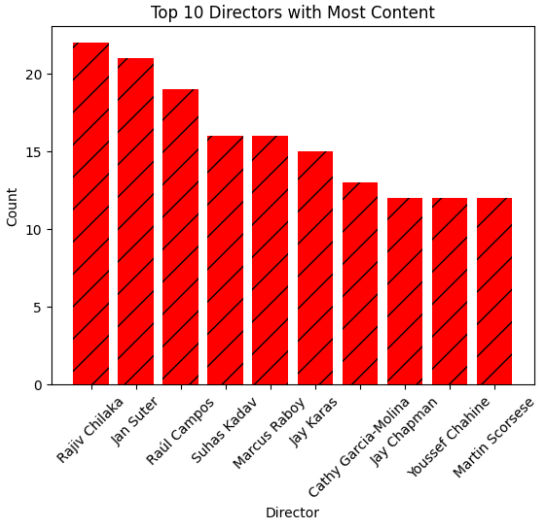


Let's look at the distribution of content ratings on Netflix:



The pie chart above shows that the majority of content on Netflix is categorized as "TV-MA", which means that most of the content available on Netflix is intended for viewing by mature and adult audiences.

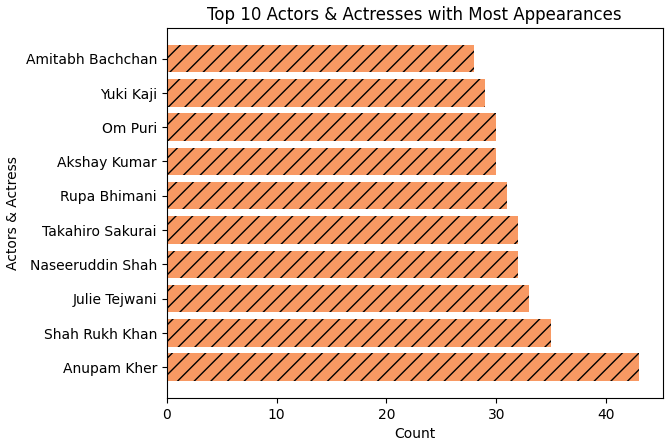
Now let's see the top 10 successful directors on this platform:



From the above graph it is derived that the top 5 directors on this platform are:

1. Rajiv Chilaka
2. Jan Suter
3. Raul Campos
4. Marcus Raboy
5. Suhas Kadav

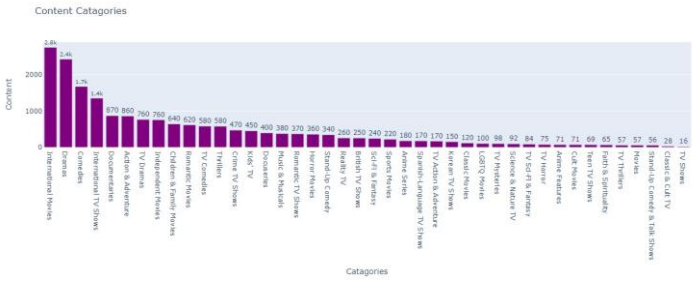
Now let's have a look at the top 10 successful actors and actress on this platform:



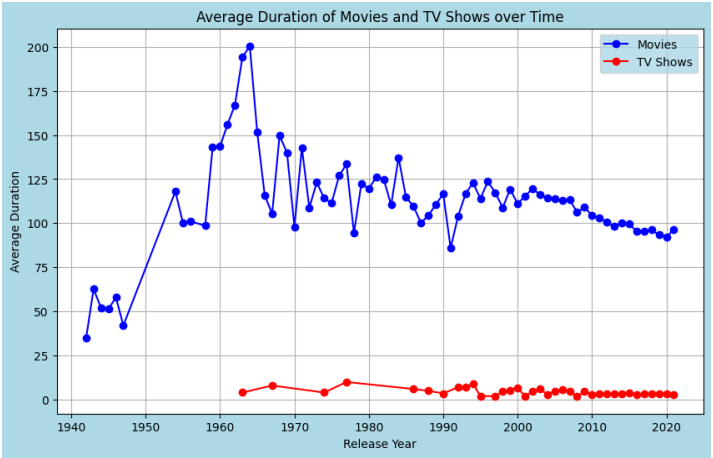
From the above graph it is derived that the top 5 actors & actress on this platform are:

1. Anupam Kher
2. Shah Rukh Khan
3. Julie Tejwani
4. Naseeruddin Shah
5. Takahiro Sakurai

Let's visualize the types of categories of contents:



The next analysis is on content durations. As we talked before, we have two different contents movies and TV shows. Since there are two groups of contents using different units.2410 of the Netflix contents use “season” as the measurement for duration. 5377 of the Netflix contents use “season” as the measurement for duration. I decided to process just movies. I used the Histogram to show the movie duration. As we can see in **Figure** Duration 90-99min are the most movies duration, then 100-109min, then 80-89min, then 110-119min.we have A continuous variable so we used histogram plot to visualize the distribution of variables. function geom_histogram() is used for showing this distribution. Also, in this figure we used gghighlight() to highlight the lines whose max values. The red line in the histogram shows that the it centered around 100 for this type of Netflix contents.



We will create a word cloud from the listed_in of the content:

We will create a word cloud from the titles of the content:

We will create a word cloud from the cast of the content:

ACKNOWLEDGMENTS

On behalf of our accomplished team, we extend our heartfelt gratitude to our dear friends whose unwavering support and encouragement have been instrumental in the success of this project. Your belief in us and constant motivation have been pivotal throughout our journey. We would also like to express our profound appreciation to our esteemed mentors and advisors for their invaluable guidance and expertise. Their wisdom and encouragement have steered us in the right direction, ensuring the quality and depth of our research. To all our team members, your dedication, expertise, and seamless collaboration have been indispensable in achieving our goals. Together, we share this success with our friends, mentors, and team members, and we are sincerely thankful for their unwavering support on this remarkable journey.

CONCLUSION

By analyzing the Netflix dataset and using different plots like histogram, boxplot and bar charts. we found that 68.4% of the content on Netflix is Movies. The top 3 countries creating TV Shows for Netflix are United States, United Kingdom and Japan and top 3 countries creating Movies are United States, India and United Kingdom. Movies that produced in India are the longest contents. Also we found that most movies duration are between 90 to 99 minutes and there has been very few movies and TV shows released before 2000 and the spike was in 2019.I have plan to do more analysis on this dataset to find more relationship between variables.

To conclude, we have constructed a relatively accurate data analysis to determine the genre of a given TV show and movie. Processing and narrowing down the features of the Netflix Dataset by identifying the top ten countries with top five genres the audience watch and feeding this data to figure out the trend over the years, we were able to identify which actors/actresses by which directors were popular in each country. Steps such as expanding the dataset and feature set and using logistic regression might have the potential to improve upon on results in the future.

REFERENCES

[1] https://www.kaggle.com/shivanirana63/netflix-eda-movie-recommendation-system/data

[2] https://www.kaggle.com/shivamb/netflix-shows

[3] https://github.com/yihui/knitr-examples/blob/master/077-wrap-output.Rmd

[4] https://www.kaggle.com/datasets/shivamb/netflix-shows

[5] https://en.wikipedia.org/wiki/Netflix

[6] https://www.kaggle.com/datasets/shivamb/netflix-shows/discussion

[7] https://coolors.co/