NETFLIX TV SERIES AND MOVIES EXPLORATORY DATA ANALYSIS USING PYTHON AND TABLEAU

BY GAUTAM KHARBANDA AND PAWAN THAKUR

Netflix Inc. is a pay television over-the-top media service and American production company that offers subscription-based video on demand from a library of films and television series, 40% of which is original programming produced in-house.

Netflix was founded in 1997 by Reed Hastings and Marc Randolph in Scotts Valley, USA. Netflix initially both sold and rented DVDs by mail. The company expanded to Canada in 2010, followed by Latin America and the Caribbean. Netflix entered the content production industry in 2013, debuting its first series *House of Cards*. In January 2016, it expanded to an additional 130 countries and then operated in 190 countries.

It all began in April 1998, when Netflix started renting out DVDs by mail. Only a year later Netflix changed its pay-for-use model into a subscription model. Nearly a decade later, Netflix changed their proposition to a streaming service, which changed the way millions of people spend their free time.

ABOUT THE DATASET

We have taken the dataset from Kaggle which describes the TV Shows and movies listed on Netflix along with *details such as - cast, directors, ratings, release year, duration, etc.* as of September 2021.

IMPORTING THE LIBRARIES AND UNDERSTANDING THE DATASET

```
In [1]: import numpy as np
        import pandas as pd
        import seaborn as sns
        import matplotlib.pyplot as plt
```

GETTING STARTED WITH OUR ANALYSIS

READING THE DATA

```
In [2]: df=pd.read_csv('D:/EDA Datasets/netflix_titles.csv')
```

PRINTING THE FIRST FIVE ROWS OF THE DATASET

```
In [3]: df.head()
```

Out[3]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Documentaries	As her father nears the end of his life, filmm...
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	After crossing paths at a party, a Cape Town t...
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	To protect his family from a powerful drug lor...
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA	1 Season	Docuseries, Reality TV	Feuds, flirtations and toilet talk go down amo...
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...	In a city of coaching centers known to train I...

We then imported our dataset using pd.read csv. The data was in csv format. Then by using df.head() argument we printed the first five rows of the dataset. We see that the dataset is mostly categorical with features like showid which is a serial number for the data, type i.e. either Movie or TV series, title, cast, country of origin, date added on Netflix or date when it was released on Netflix. The release year corresponds to the actual year it released i.e. if it was a movie the year it was released theatrically or Tv series that aired. The rating columns describe the target audience the show is meant for like PG-13 is for People above 13 years of age, TV-MA is TV show for mature audiences. The duration column describes the movie length or the number of seasons of the show that aired. The listed in feature tells us the genre a particular movie or TV series belongs to i.e. Documentries, Standup Comedy, Anime, Romantic comedy, Drama etc. The description column is a short summary about the show or the movie.

After understanding the dataset we then checked for the unique values in the dataset, The different columns had different unique values. For our Analysis we checked for unique columns as well.

CHECKING THE DIFFERENT COLUMNS IN THE DATASET

```
In [5]: df.columns

Out[5]: Index(['show_id', 'type', 'title', 'director', 'cast', 'country', 'date_added',
               'release_year', 'rating', 'duration', 'listed_in', 'description'],
              dtype='object')
```

CHECKING THE UNIQUE VALUES IN THE DATASET

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

Out[6]: 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
```

After the above step we checked for any Null Values in the dataset. The director, Cast, Country, Date Added, rating and duration had Null values. We handled Null values by Filling the null values with "Unavailable" because the given dataset is categorical and each show and films are unique It was the most appropriate method to fill it using Unavailable using fillna() function in Pandas

CHECKING FOR NULL VALUES

```
In [7]: df.isnull().sum()

Out[7]: 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
```

FILLING THE NULL VALUES

```
In [10]: df.director.fillna('Unavailable',inplace=True)
         df.cast.fillna('Unavailable',inplace=True)
         df.country.fillna('Unavailable',inplace=True)
         df.date_added.fillna('Unavailable',inplace=True)
         df.rating.fillna('Unavailable',inplace=True)
         df.duration.fillna('Unavailable',inplace=True)
```

After the above step we checked the first 5 values of the dataset using df.head() to check whether the null values have been properly filled or not.

We can see that the appropriate value has been filled in the Null values place.

After the above we dropped unnecessary columns from the dataset which were of no use to us for our Analysis like description. After this we again check by printing first five rows of the dataset.

## GETTING STARTED WITH OUR VISUALIZATIONS

The first graph given below is a line diagram of the number of Shows and Movies added in Netflix in the last decade. We see overall that number of Movies is more than the number of Tv series. We also see that from 2010 to 2016 that the trend of films increased and remained constant during 2017-2018 and started decreasing from 2019 and dropped considerably in 2020 up to September 2021. The Tv series line shows an increase from 2010 to 2019 and started dropping from 2020 up-to September 2021. This could be due to the Covid 19 Pandemic

RELEASE YEAR OF DIFFERENT CONTENTS
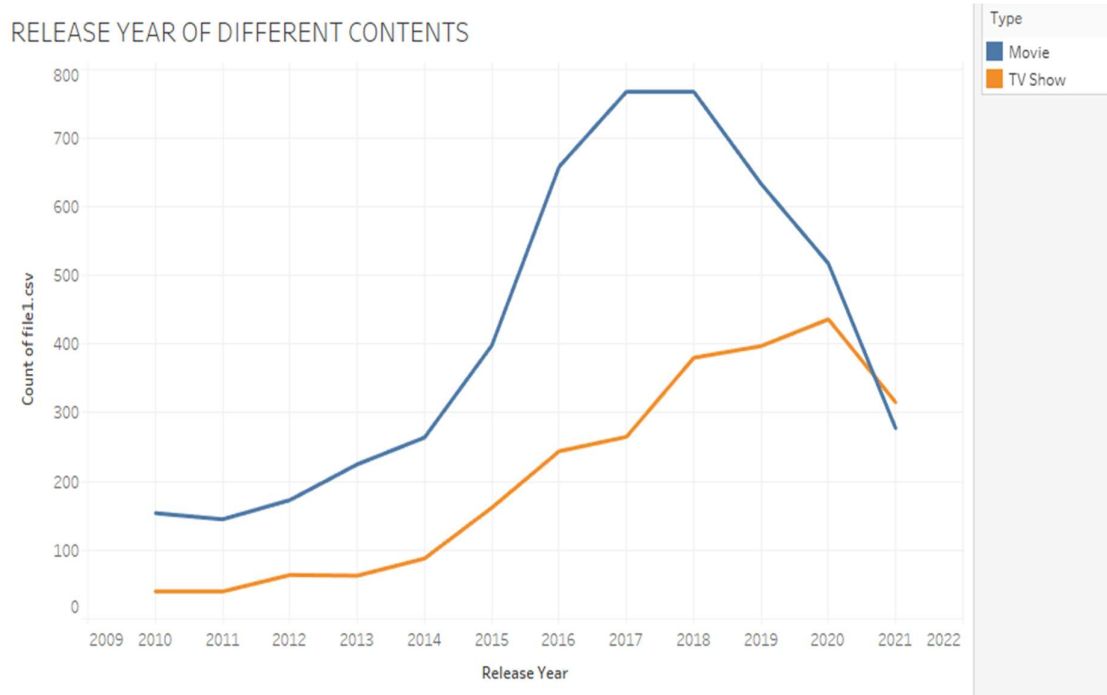
Due to Covid-19 pandemic production of shows and films suffered greatly that is why the effect has been shown in 2021 as well as the content addition was also below expected.

Next Graph is Ratings Tree Map which shows that the most number of content that was added in Netflix during the last decade was of TV-MA, TV-14, TV-PG, PG-13, from which we can conclude that Netflix generally adds content for Mature Audiences or Content for children under parental guidance, or Children of 14 years and above or Children above 13 years requiring Parental supervision. Netflix already has a special section that has content only for kids.

RATINGS HEAT MAP
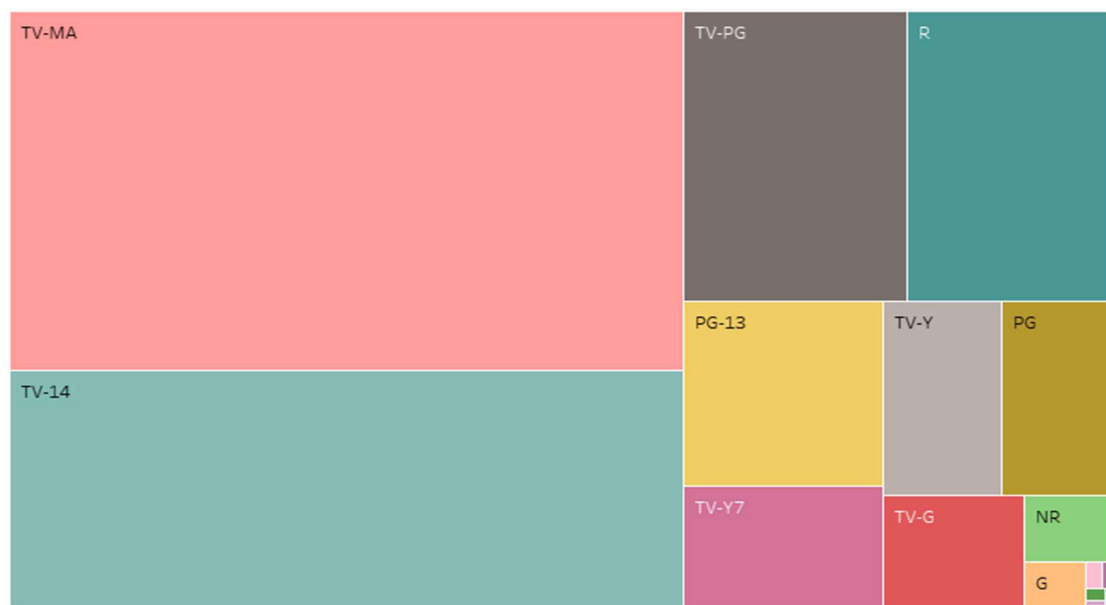
The Next Graph shows the top 10 countries' content available on Netflix. These are USA, India, UK, Canada, South Korea, Spain, Mexico, Japan, Nigeria and France. From India Netflix has content from Bollywood and Regional Cinema and also a lot of Independent Films. India also has English speaking TV- Series, Indian content from Youtube and Regional shows. This is the reason why India Ranks at 2nd Position with the number of Content available.

TOP 10 COUNTRIES CONTENT AVAILABLE ON NETFLIX
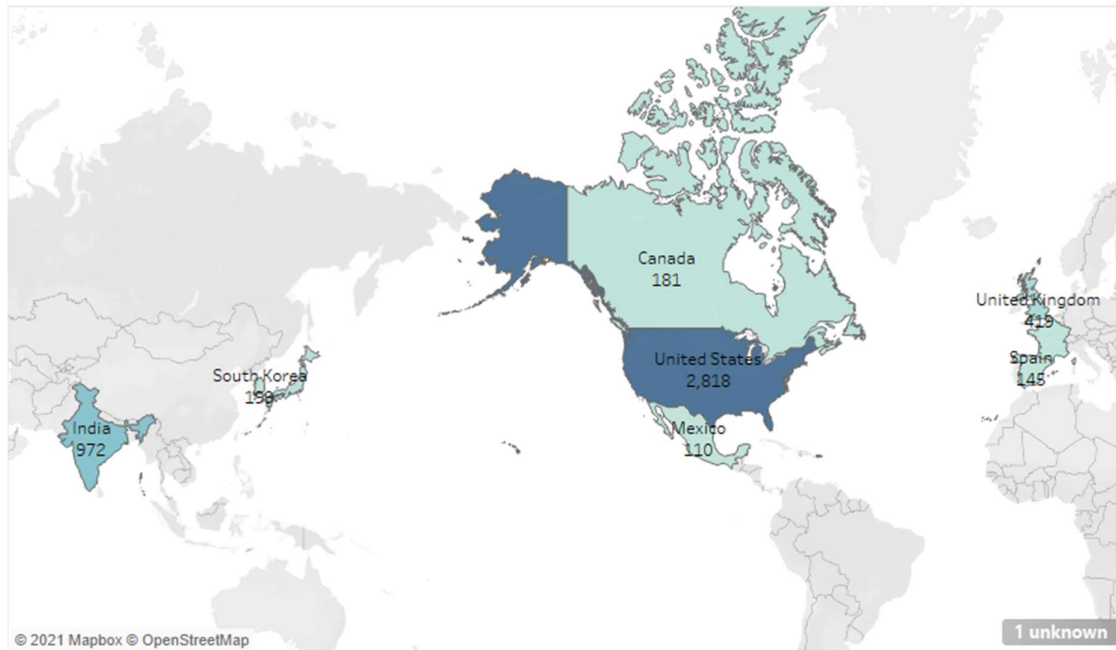
From here on We take two countries from North America. One English Speaking and one Non-English speaking country to see whether English speaking countries have same consumption pattern or not. And within a continent 2 countries One English Speaking and other Non English speaking have different content consumption pattern.

Also, We check whether countries which are developed have a different watching pattern than the countries which are poor or have a lower standard of living or developing countries.

Firstly We check the consumption pattern in the United States, an English Speaking Country the Top Programs listed in are documentaries, Stand- up comedy and Children and Family Movies.

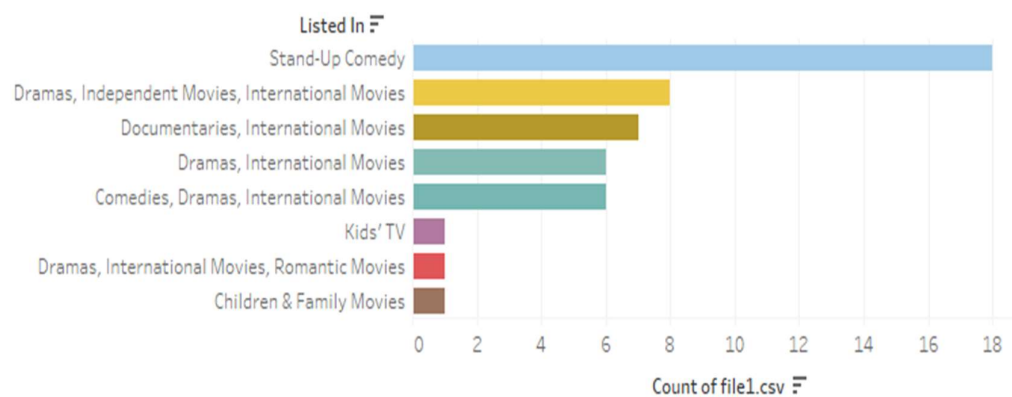UNITED STATES CONTENT CONSUMPTION PATTERN
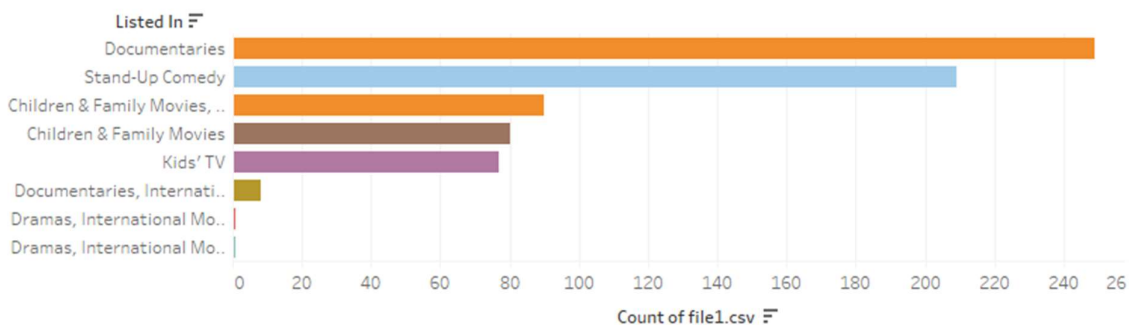
Watching Documentaries is the most preferred in the United States, followed by Stand up comedy. Netflix has a lot of American comedy specials like Dave Chapelle, Bert Kreischer, Tom Segura etc to name a few. Also Children's and Kid Friendly films are quite common

Next, we check the consumption pattern of Mexico which is a Non-English Speaking country of North America but poorer than united states and having social inequality greater than that of the United States.

We find that Stand up comedy is the most watched content followed by International content like International Drama, International Documentaries. The consumption pattern is significantly different in a continent as the language changes.

MEXICO CONTENT CONSUMPTION PATTERN

Next, We check the consumption Pattern in United Kingdom We find a consumption pattern which is similar to United States i.e. we find that documentaries followed by International Movies and Stand up comedy. This shows that different countries speaking the same language have similar content consumption.

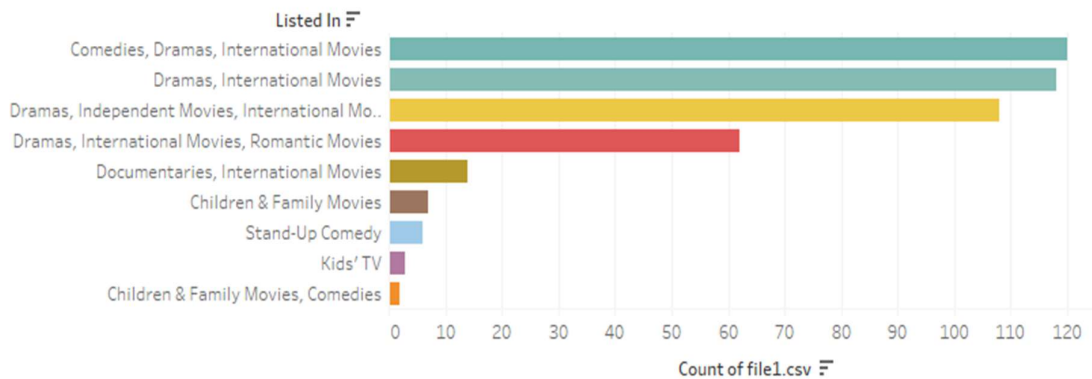## UNITED KINGDOM CONTENT CONSUMPTION PATTERN
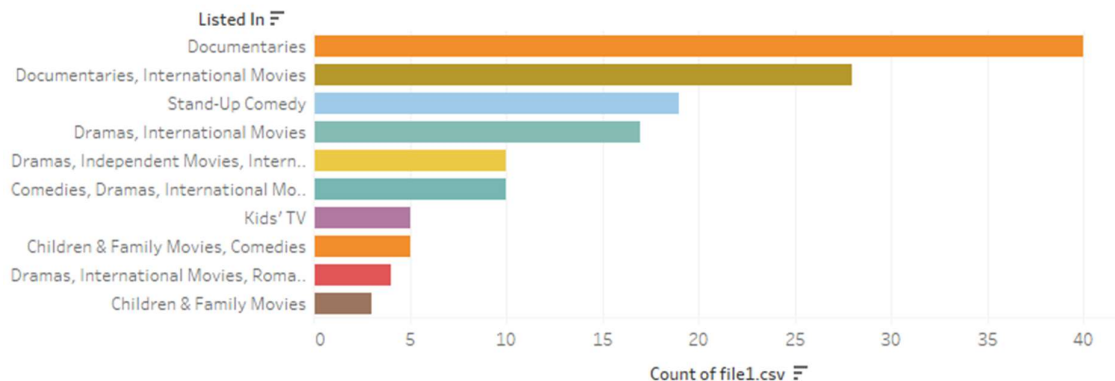
Next we check the content consumption of India We find that International films (Comedies and dramas are most watched, followed by children and family movies. This could be due to

people watching International content because it is very different from Bollywood and Netflix as opposed to communal viewing in cinemas offers a more personalized experience of watching content that one might not be able to watch with Kids or Adults with their Parents.
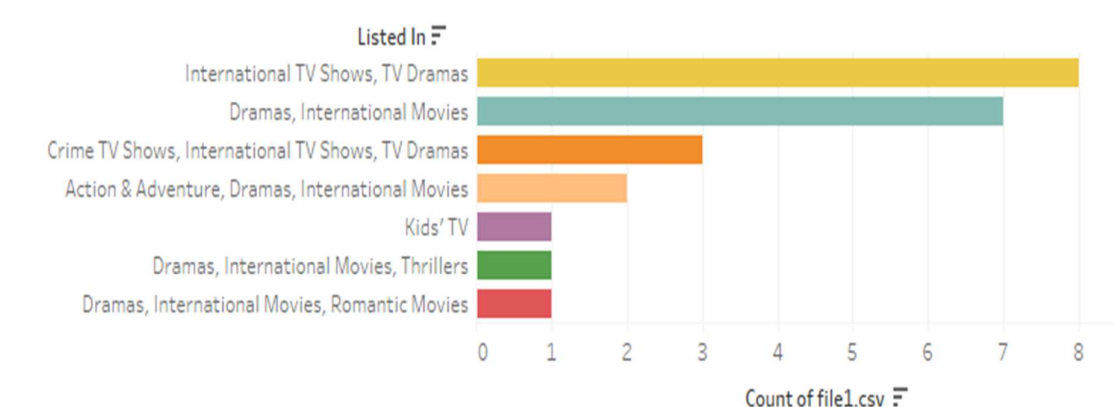
Secondly India having the highest population of millennials in the urban, semi urban parts watch international content to learn spoken English like watching your favourite sitcom from America (Modern Family, Friends) which have a high number of seasons help people to communicate in English. The same could also be said about Mexico whose people watch International content as well.

INDIA CONTENT CONSUMPTION PATTERN

Listed In
Comedies, Dramas, International Movies
Dramas, International Movies
Dramas, Independent Movies, International Mo..
Dramas, International Movies, Romantic Movies
Documentaries, International Movies
Children & Family Movies
Stand-Up Comedy
Kids' TV
Children & Family Movies, Comedies

0 10 20 30 40 50 60 70 80 90 100 110 120

Count of file1.csv

Next we analyse content from Japan a thriving country in Asia. We find that people apart from watching International TV also watch a lot of Crime Drama. Japanese society is considered to be one of the most peaceful societies in the world with a very low crime rate. This shows that people are more interested in watching content that are different from their day to day lives. People want to watch things that they don't encounter in their day to day lives.

JAPAN CONTENT CONSUMPTION

Listed In
International TV Shows, TV Dramas
Dramas, International Movies
Crime TV Shows, International TV Shows, TV Dramas
Action & Adventure, Dramas, International Movies
Kids' TV
Dramas, International Movies, Thrillers
Dramas, International Movies, Romantic Movies

0 1 2 3 4 5 6 7 8

Count of file1.csv

Next, we analyse the trend of content release year. We see that highest number of contents on Netflix was released in year 2018. Year 2018 has been the peak year till date, for overall content release on Netflix.
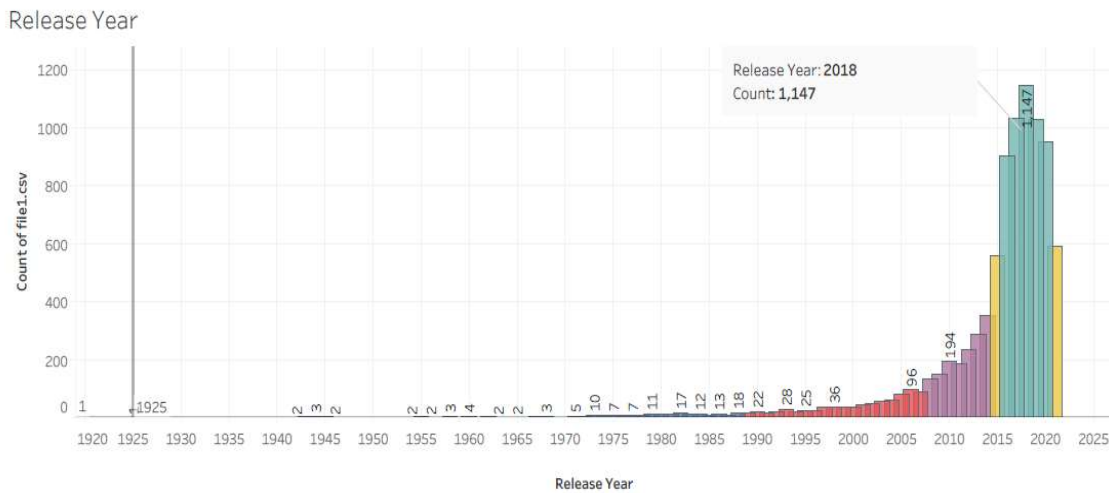
Next, we analyse Top 5 Directors with numbers of released content from inception till date, where we find Director Marcus Raboy leading the chart with 16 releases, Director Jay Karas with 14 releases, Director Martin Scorsese and Jay Chapman with Tie of 12 releases and Steven Spielberg with 11 releases, on Netflix.
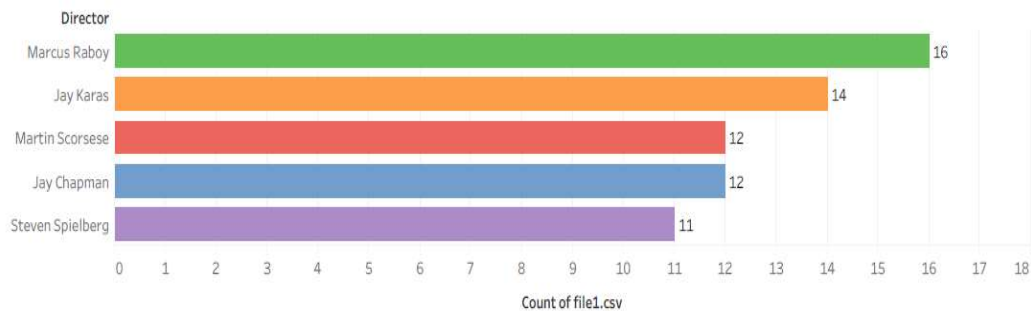
Next Graph is Content Tree Map which shows that the most number of content that was added in Netflix during the last decade was in Q3 of 2020 and 2021 This may be the outcome of post COVID impact where people are more prone to watch Netflix and due to increase in demand of OTT platform.

Movies / TV Show per Year

Legend — Quarter of Date Added:
- Q1
- Q2
- Q3
- Q4

Q3 (pink):
- 2021 Movie 404
- 2021 TV Show 214
- 2018 Movie 336
- 2019 Movie 271
- 2019
- 2020 Movie 300
- 2020 TV
- 2017 Movie

Q2 (dark red/magenta):
- 2021 Movie 353
- 2020 Movie 347
- 2021 TV Show
- 2020
- 2019 Movie 332
- 2018 Movie 207
- 2019
- 2018
- 2017 Movie

Q1 (purple):
- 2019 Movie 338
- 2020 Movie 317
- 2019
- 2020
- 2018 Movie 306
- 2021 Movie 236
- 2021
- 2017 Movie

Q4 (gray/brown):
- 2019 Movie 483
- 2020 Movie 320
- 2017 Movie 232
- 2018 Movie 388
- 2020
- 2017
- 2016
- 2016

----End of Project Report----