****

**NETFLIX**

Instructor:

**Gahangir Hossain**

Submitted by :

**Sri Chandra Dronavalli**

**11585371**

**Introduction:**

Netflix is the most popular video streaming site contributing to nearly a quarter of global video traffic. Understanding how individual Netflix customers consume content is important to network operators, content producers and creators, consumers, advertising, and the research community because Netflix dominates Internet traffic. In this case, we gather Netflix viewing data from 1060 users over the course of a year, totalling over 1.7 million episodes and films. Based on the users' degree of activity, we categorize them and offer important data about their watch patterns, watch session length, preferences, predictability, and watch behaviour continuance inclinations. Additionally, we construct and assess classifiers that forecast a user's participation in a series based on previous behavioural patterns.

**Dataset:**

We are going to work on the Netflix movies and TV Show dataset you can find this dataset on Kaggle. If you like to see the whole notebook you can visit it here(<https://www.kaggle.com/code/manuelcabrerag/netflix-data-cleaning-analysis-and-visualisation/data?select=netflix1.csv> ). Netflix is an application that keeps growing exponentially whole around the world and it is the most famous streaming platform.

**Discussing the dataset:**

The variables of this data set are:

* show\_id: Netflix ID of the media.
* Type: Movie or TV Show.
* title: Title of the media.
* director: Director of the media.
* country: Country in which the movie was made.
* date\_added: Date in which the media was added.
* release\_year: Year in which the media was released.
* rating: Age rating of the media.
* duration: Duration of the media.
* listen\_in: Classification given by Netflix.

**Tools:**

• Python

• Tableau

Let us create an EDA through this data with beautiful charts and visuals to gain some insights.

**Dataset Importing and Querying:**

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

**Load and read data**

In this section of the report, we will gather the dataset and turn it into a DataFrame and check the shape.

Importing the data from a csv file to a DataFrame

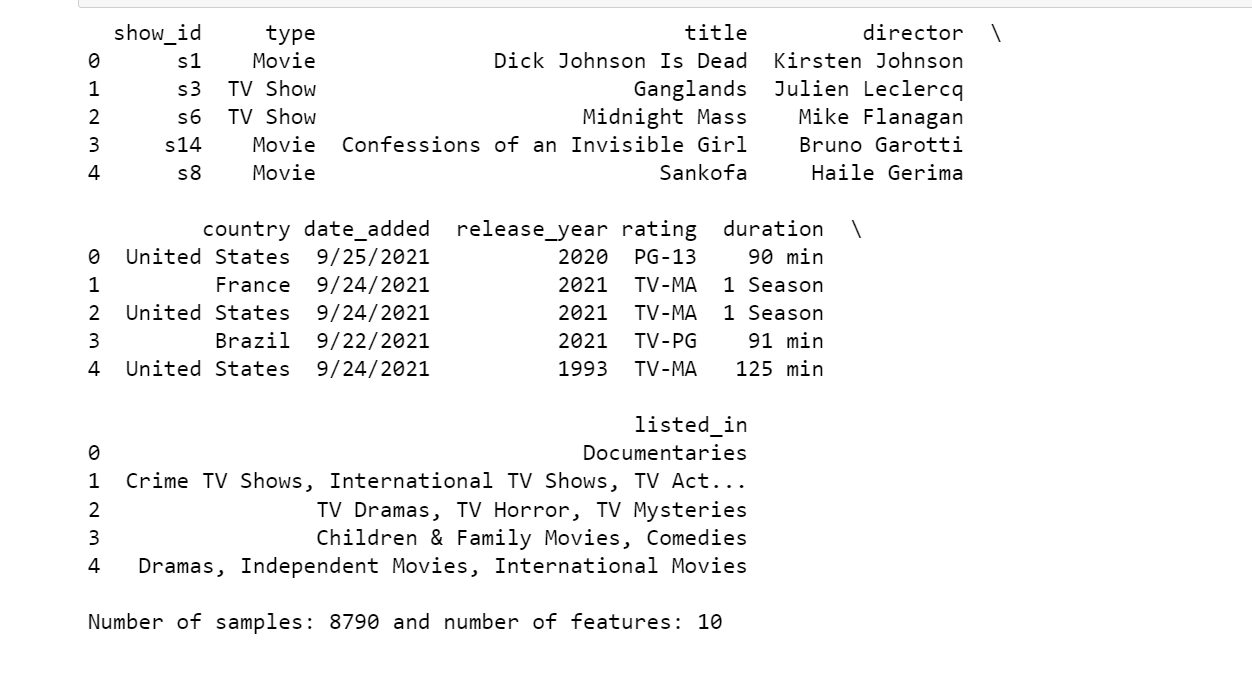
df = pd.read\_csv("netflix1.csv")

# Showing the first five values of the DataFrame

print(df.head())

#Check the size of the dataset

print("\nNumber of samples:",df.shape[0],"and number of features:",df.shape[1],"\n")

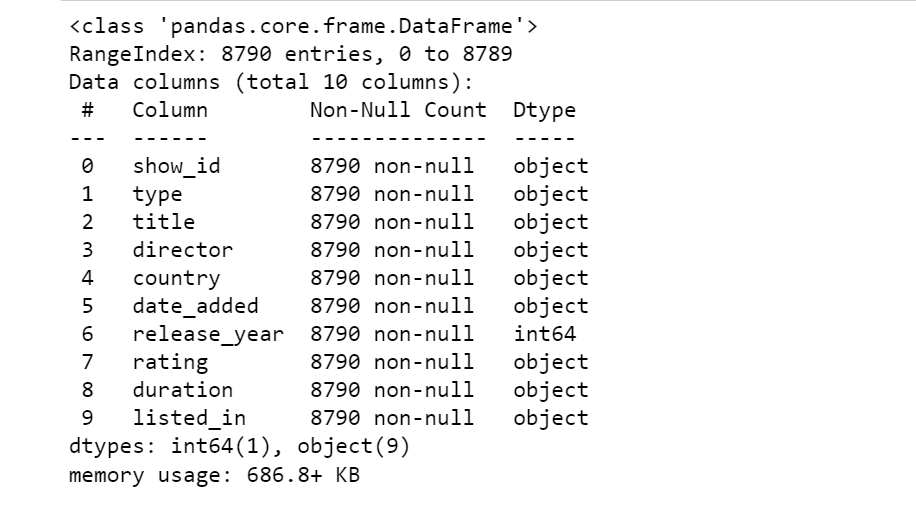


**Assessing**

This section of the report we will assess any issues the data may have.

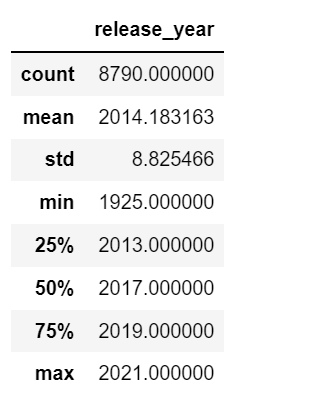
# Let's check the status of the data

df.info()



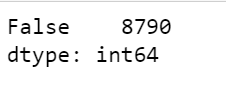
# Describe the data

df.describe()



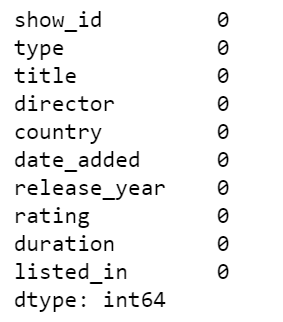
# Checking if there are any duplicates

df.duplicated().value\_counts()

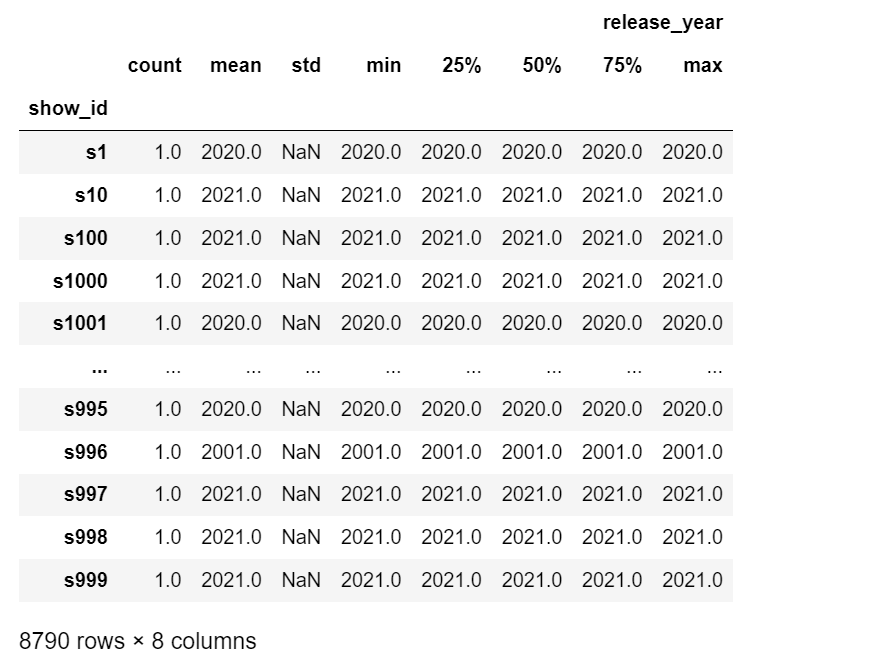


# Checking if there are any null

df.isnull().sum()

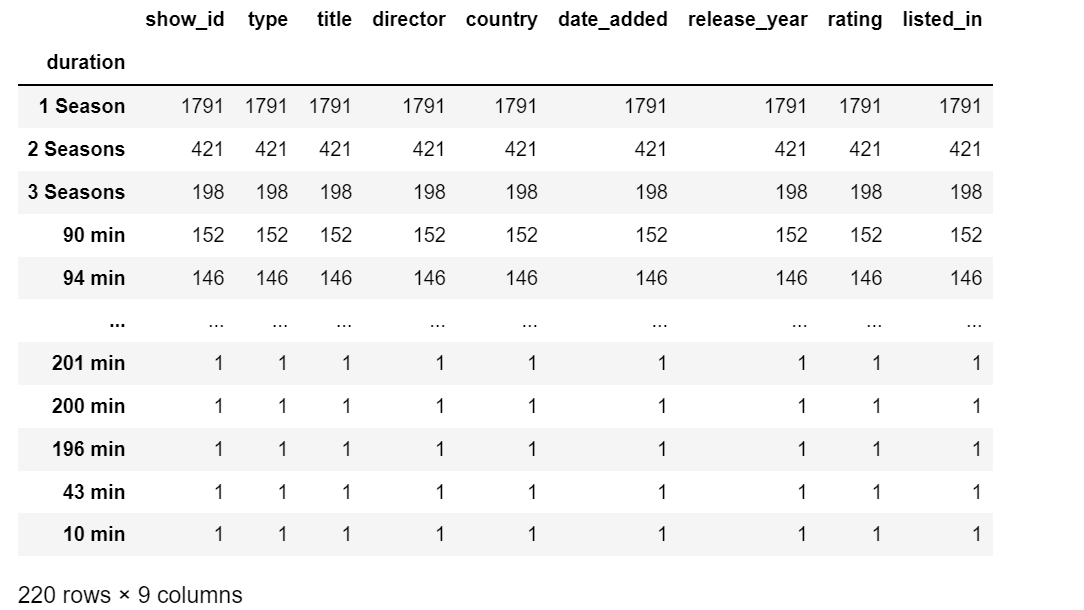


df.groupby('show\_id').describe()



# Let's perform a basic visual analysis of the data.

df.groupby('duration').count().sort\_values(by='show\_id',ascending=False)



**Assessment & Categorising**

1. Quality issue

* Variable 'date\_added' has the wrong data type.

2. Tidiness issue

* The 'listed\_in' variable has several categories in a single observation.
* There are two types of observations, TV shows and movies.

**Cleaning:**

In this section of the report we will solve the quality and tidiness issues mentioned in the assessment.

# Before cleaning, lets make a copy of the dataframe.

df\_clean = df.copy()

**'date\_added' variable has wrong data type**

**1. Define**

The variable 'date\_added' has been categorised as an object (string), the most appropriate type of data for this variable would be datetime also creating the day , year, month of the date added for future analysis

**2. Implementation**

df\_clean["date\_added"] = pd.to\_datetime(df\_clean['date\_added'])

df\_clean['day\_added'] = df\_clean['date\_added'].dt.day

df\_clean['year\_added'] = df\_clean['date\_added'].dt.year

df\_clean['month\_added']=df\_clean['date\_added'].dt.month

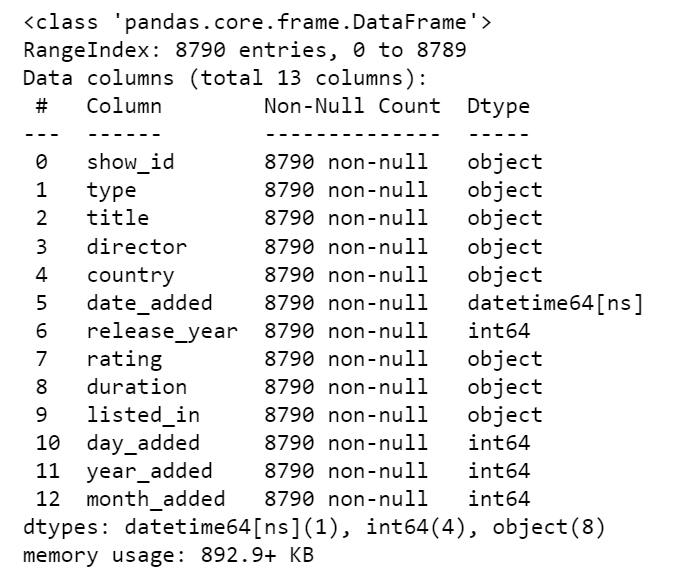
df\_clean['year\_added'].astype(int);

df\_clean['day\_added'].astype(int);

**3. Result**

Let's check the status of the data

df\_clean.info()



**'listed\_in' variable has several variables**

**1. Define**

# The 'listed\_in' variable can have several categories per media, we would like to create new variables to be able to extract this and correctly filter the data. We will assume that the first category would be the 'main' category of the movie.

**2. Implementation**

df\_clean['listed\_in1'] = 0

df\_clean['listed\_in2'] = 0

df\_clean['listed\_in3'] = 0

temp\_cat = df\_clean.listed\_in.str.split(',')

i=0

for i in range (8790):

    t\_cat = temp\_cat[i]

    if len(t\_cat) == 1:

        df\_clean['listed\_in1'][i] = temp\_cat[i][0]

        df\_clean['listed\_in2'][i] = 0

        df\_clean['listed\_in3'][i] = 0

    if len(t\_cat) == 2:

        df\_clean['listed\_in1'][i] = temp\_cat[i][0]

        df\_clean['listed\_in2'][i] = temp\_cat[i][1]

        df\_clean['listed\_in3'][i] = 0

    if len(t\_cat) == 3:

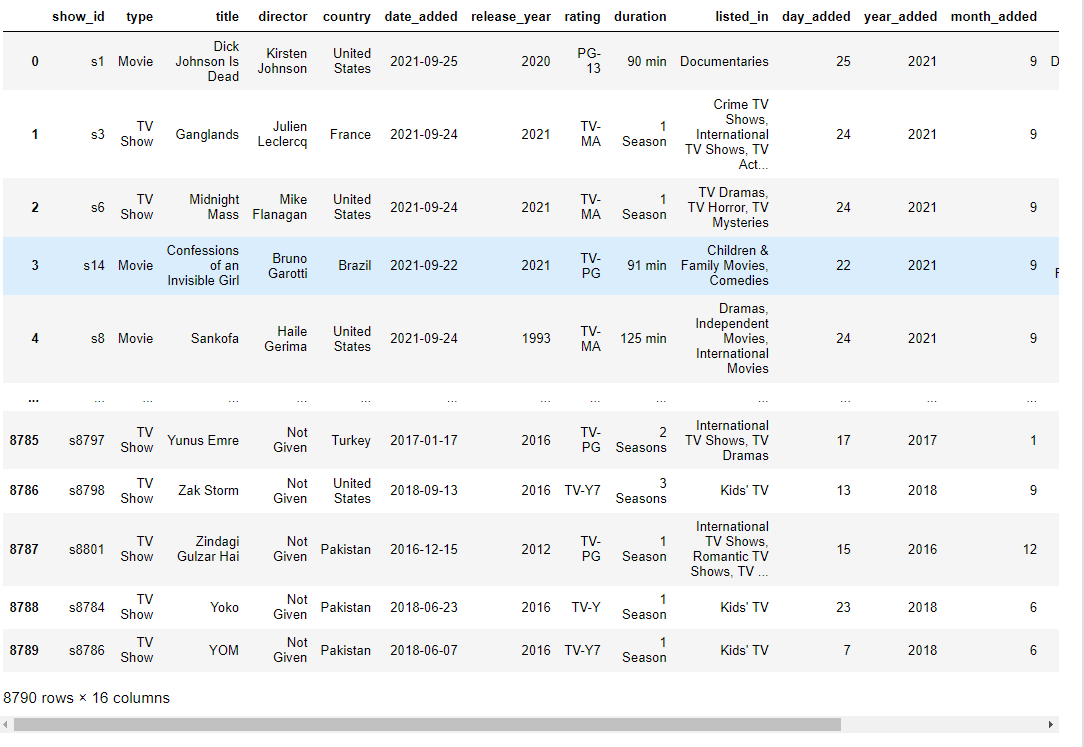
        df\_clean['listed\_in1'][i] = temp\_cat[i][0]

        df\_clean['listed\_in2'][i] = temp\_cat[i][1]

        df\_clean['listed\_in3'][i] = temp\_cat[i][2]

**3. Result**

df\_clean



**Two different observation on a single dataset**

**1. Define**

There are two different types of observations in a single data set: TV Shows and Movies. The solution would be to split the dataset into two.

**2. Implementation**

df\_tv = df\_clean[df\_clean.type == 'TV Show']

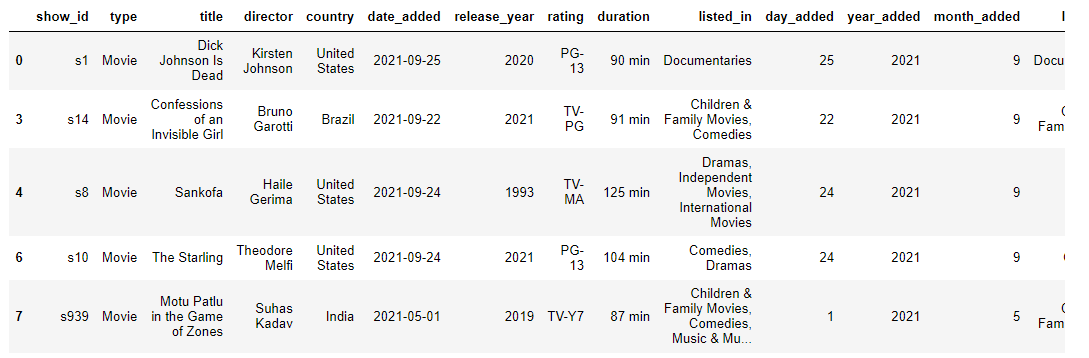
df\_movie = df\_clean[df\_clean.type == 'Movie']

**3. Result**

df\_tv.head()



df\_movie.head()



**Variable 'duration' has the wrong data type**

**1. Define**

While movies and TV shows were combined into a single dataframe, it was not possible to easily compare the length of these medias. However, now that they are separated each of these variables are not required to be kept as an object; but instead, they should be integers.

**2. Implementation**

temp\_dur = df\_tv.duration.str.split(' ',expand=True)

df\_tv['duration\_seasons'] = temp\_dur[0]

df\_tv.duration\_seasons = pd.to\_numeric(df\_tv.duration\_seasons)

temp\_dur = df\_movie.duration.str.split(' ',expand=True)

df\_movie['duration\_minutes'] = temp\_dur[0]

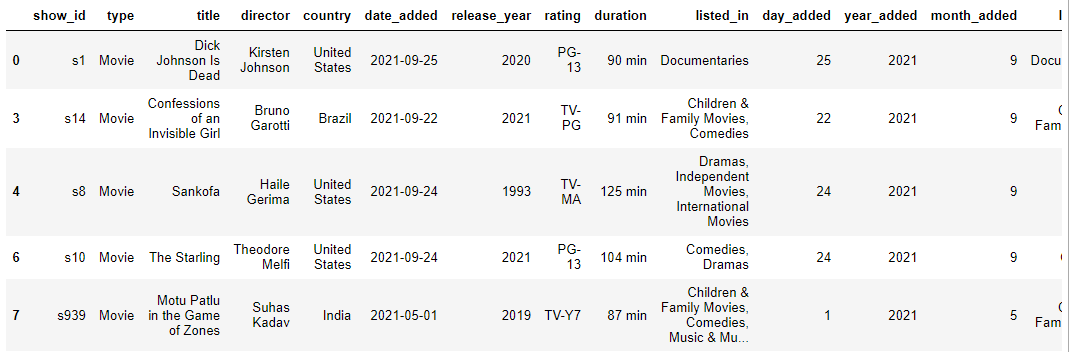
df\_movie.duration\_minutes = pd.to\_numeric(df\_movie.duration\_minutes)

**2. Result**

df\_tv.head()



df\_movie.head()



**Storing:**

In this step we will store the dataframes into CSV files.

df\_clean.to\_csv('Netflix\_cleaned.csv', index=False)

df\_tv.to\_csv('Netflix\_TV\_cleaned.csv', index=False)

df\_movie.to\_csv('Netflix\_Movie\_cleaned.csv', index=False)

**Analysing and Visualisation of Data**

In this section of the report we will explore the answers to the following questions:

- Which country produced the most of Netflix's media?

- Which Director produced the most of Netflix's media?

- what Listings produced the most of Netflix's media?

- What length of the movies and seasons are high?

- What type of media has Netflix produced the most?

- Which month the media is high for releases?

- Which listings are high for different media types?

- Which day the media is high for releases?

- Which ratings stood high in what media?

- Which year the media is high for releases?

- Has Netflix's media classification changed?

- What are the most popular genres for countries that produced media?

**Country:**

from wordcloud import WordCloud

plt.subplots(figsize=(25,15))

wordcloud = WordCloud(

                          background\_color='white',

                          width=1920,

                          height=1080

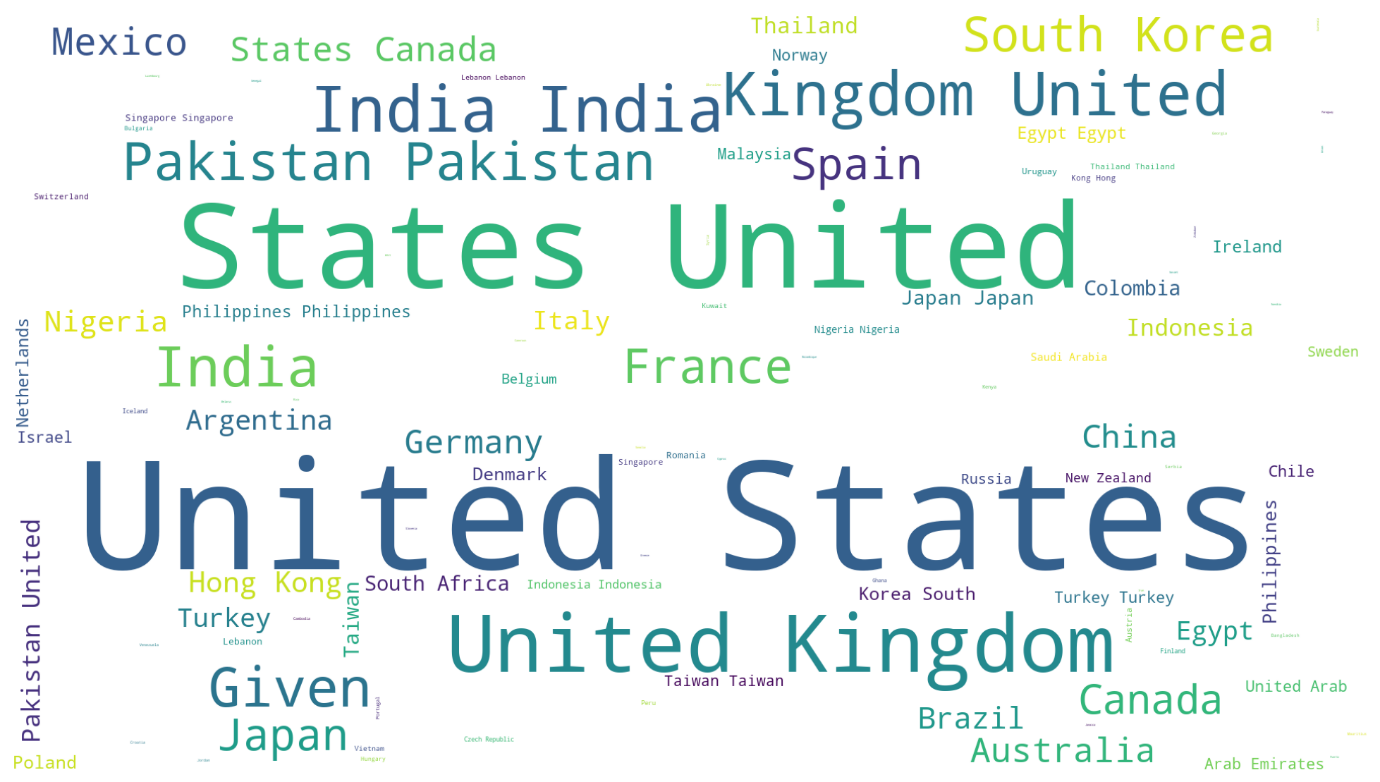
                         ).generate(" ".join(df\_clean.country))

plt.imshow(wordcloud)

plt.axis('off')

plt.savefig('country.png')

plt.show()



**Inference:** United States has the highest media contribution.

**Director**

plt.subplots(figsize=(25,15))

wordcloud = WordCloud(

                          background\_color='white',

                          width=1920,

                          height=1080

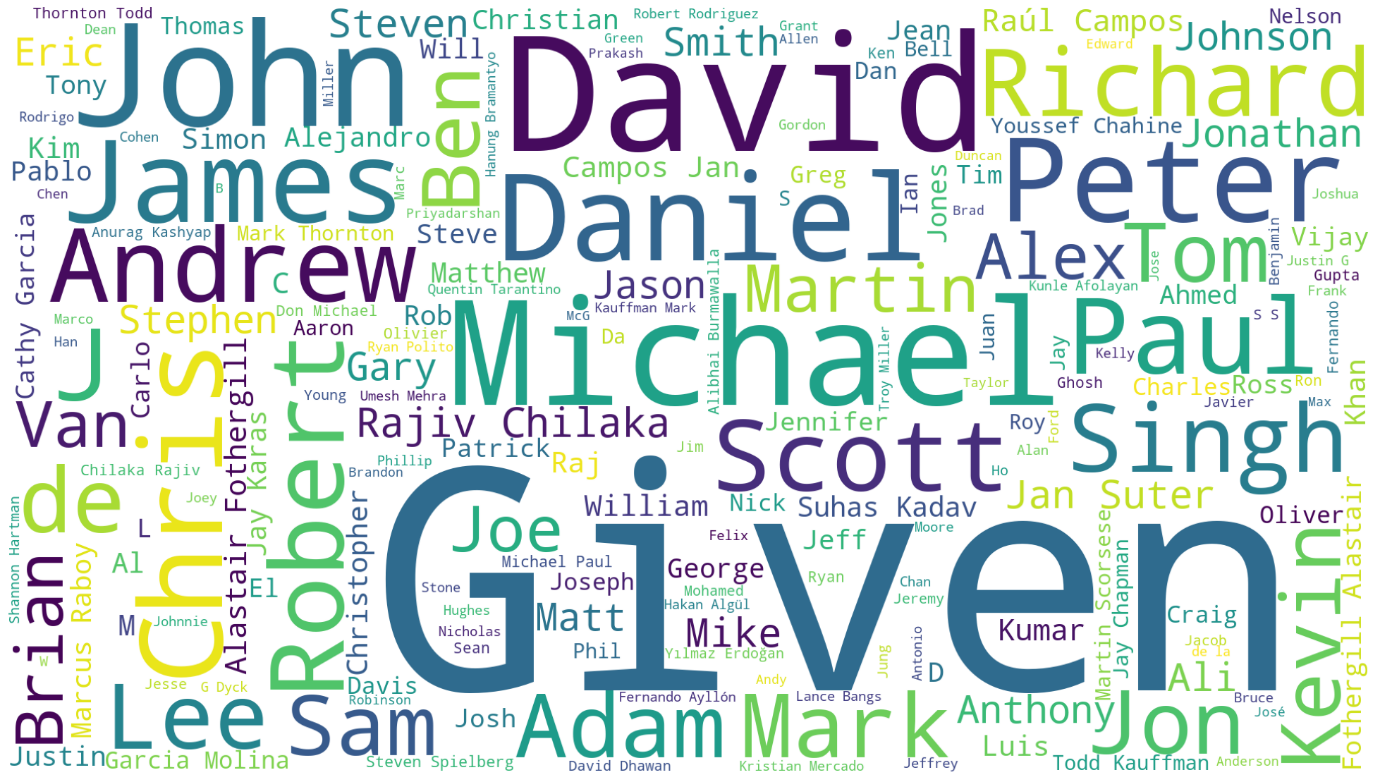
                         ).generate(" ".join(df\_clean.director))

plt.imshow(wordcloud)

plt.axis('off')

plt.savefig('director.png')

plt.show()



**Inference:** Given contribution is high in Netflix media**.**

**Listed In:**

plt.subplots(figsize=(25,15))

wordcloud = WordCloud(

                          background\_color='white',

                          width=1920,

                          height=1080

                         ).generate(" ".join(df\_clean.listed\_in))

plt.imshow(wordcloud)

plt.axis('off')

plt.savefig('listed\_in.png')

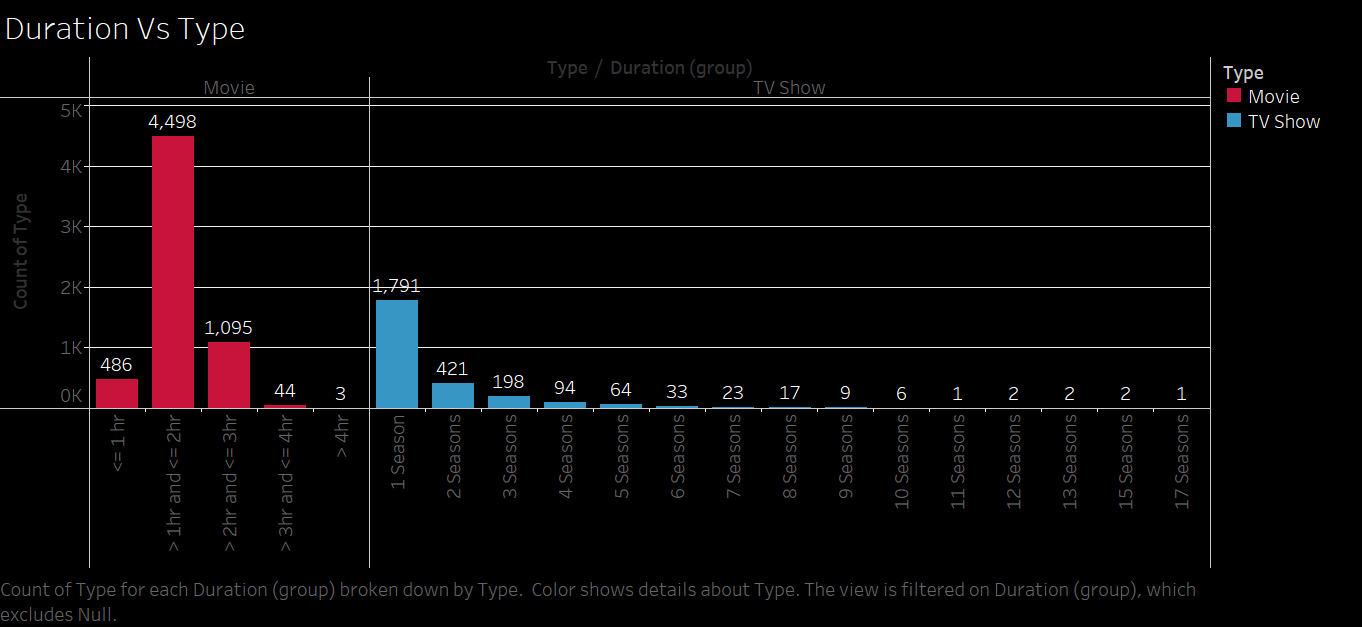
plt.show()



**Inference:** TV Shows are listed high in overall listings.

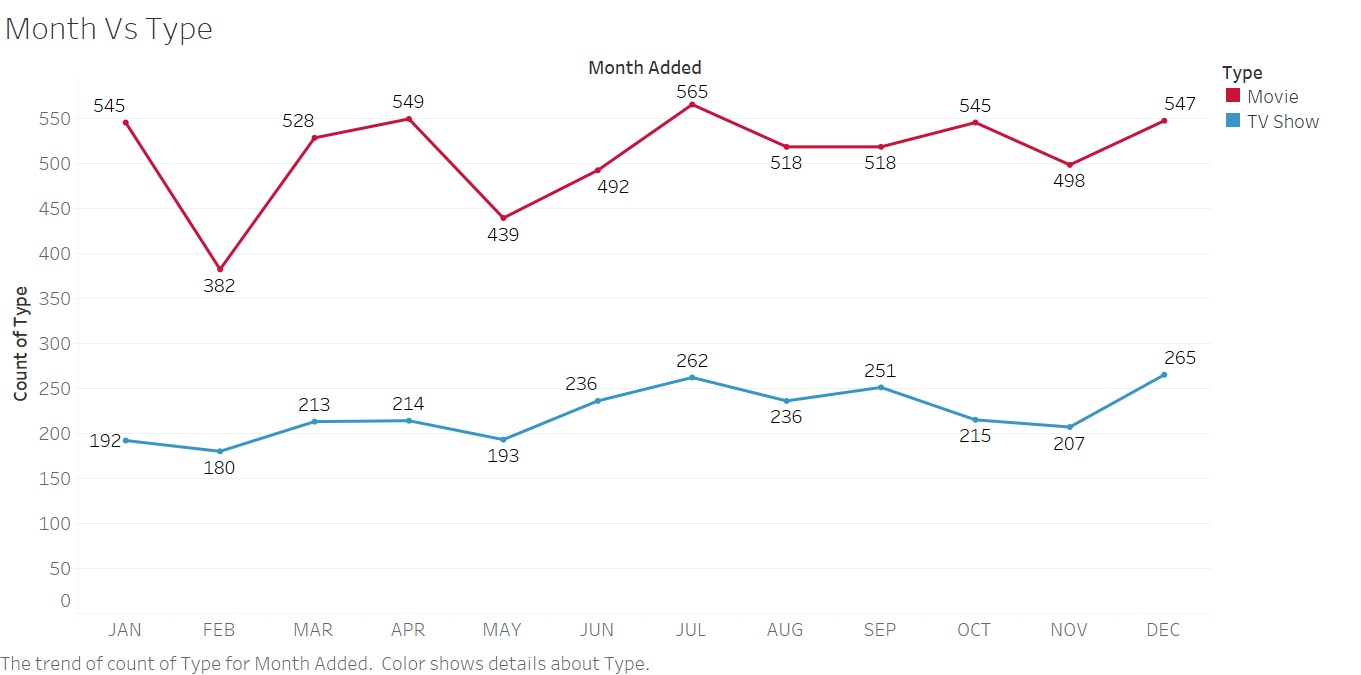
**Visualisation in Tableau:**

**Duration Vs Type:**



**Inference :** Most of the movies are in duration of 1 to 2 hours and Fewer TV shows are recorded from 7 seasons.

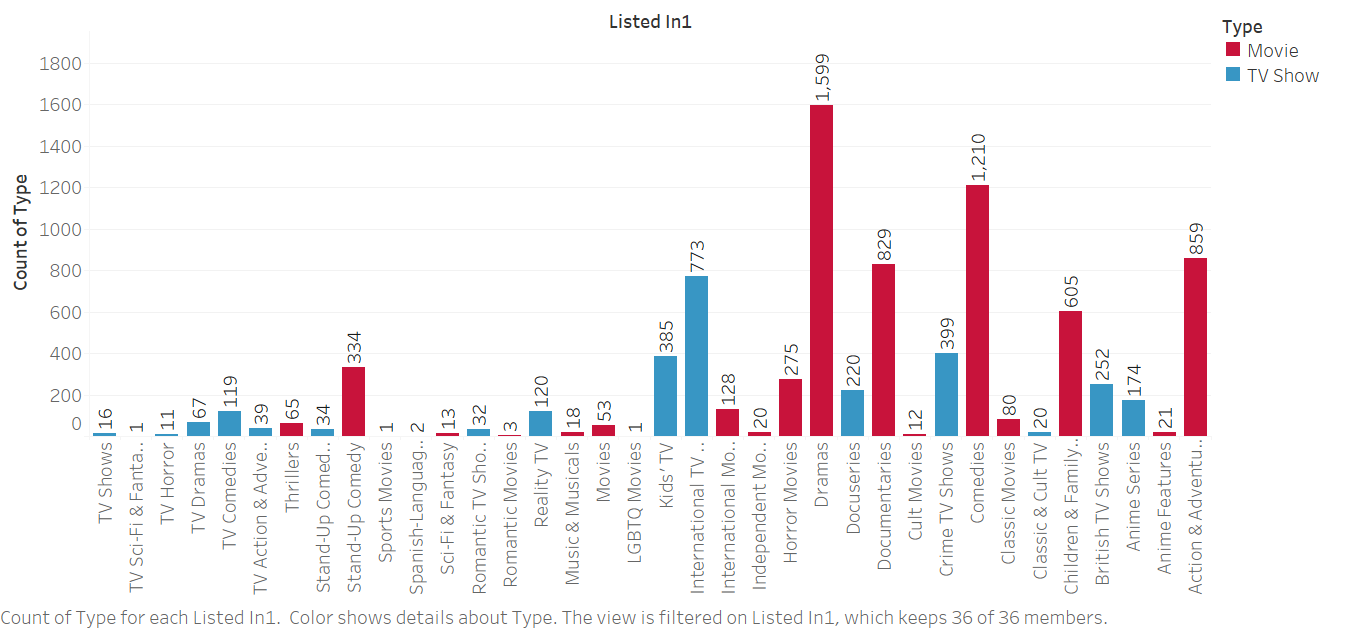
**Month Vs Type :**



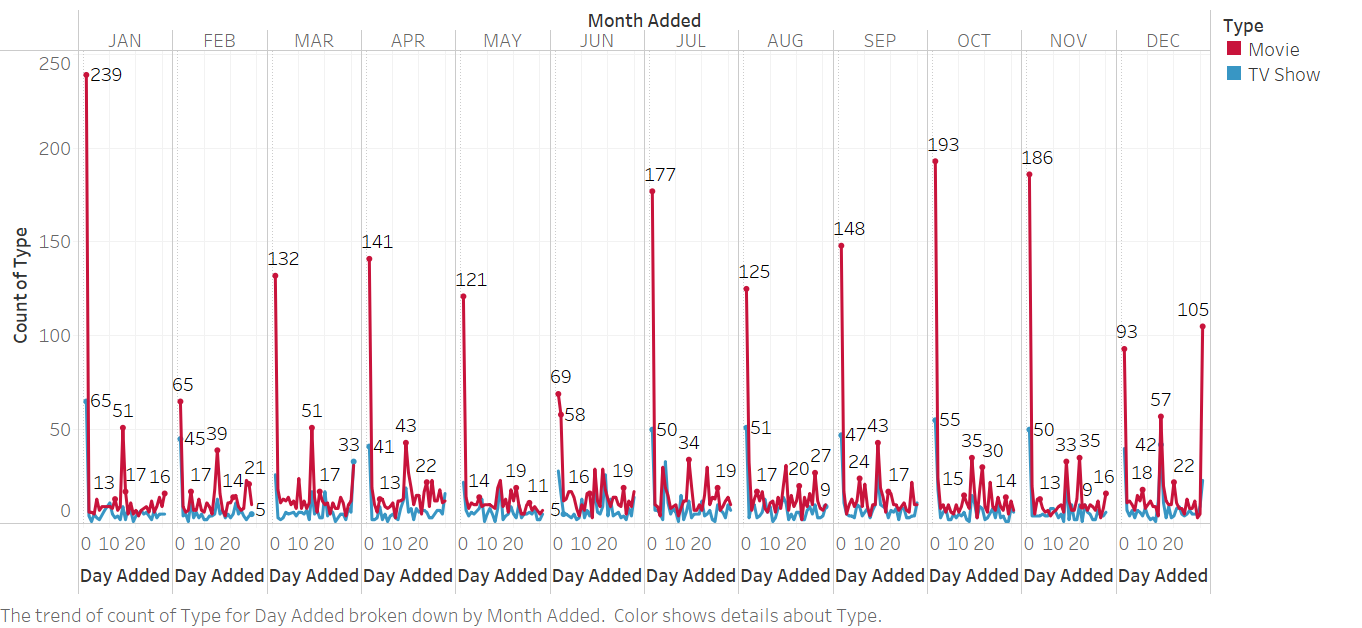
**Inference:** Highest count of movies released in the month of July (565) and lowest recorded in February (382).

In the case of TV Shows, Highest in December (265) and the lowest in February (180).

**Listedin Vs Type:**

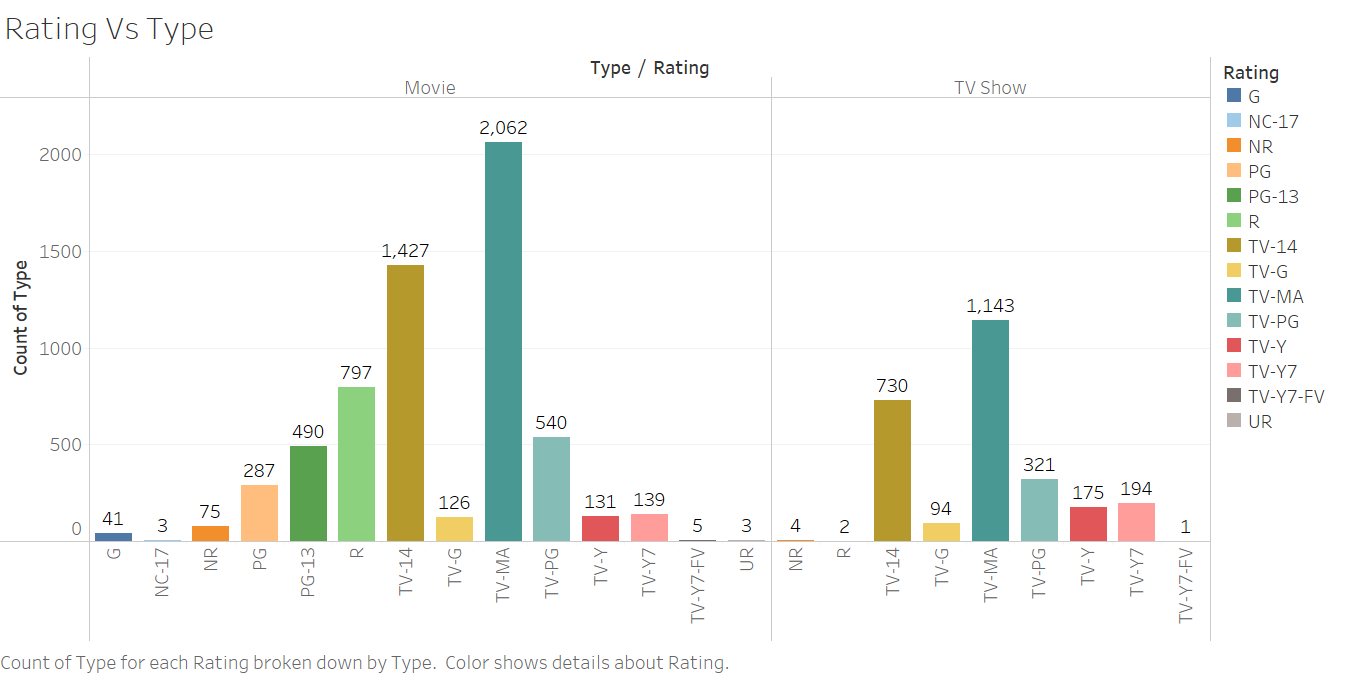
  
 **Inference:** Highest movies listed in Dramas for TV shows in International TV.

**Day Vs Month Vs Type:**



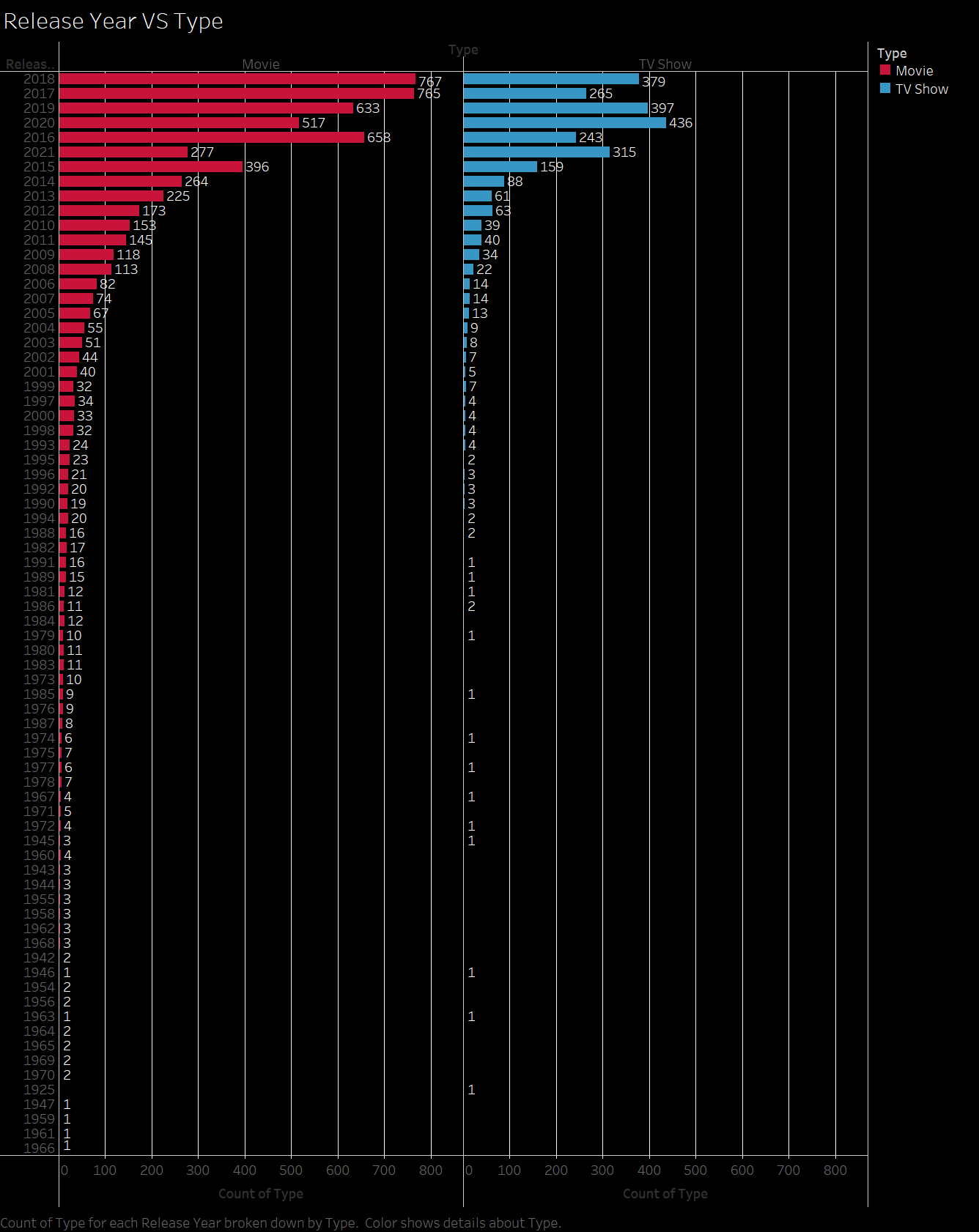
**Inference:** In the month of January(239) during Days 0 – 10 approximately the highest no of movies were released, whereas in the month of January(65) during Days 0-10 approximately the highest no of TV Shows.

**Rating Vs Type:**

****

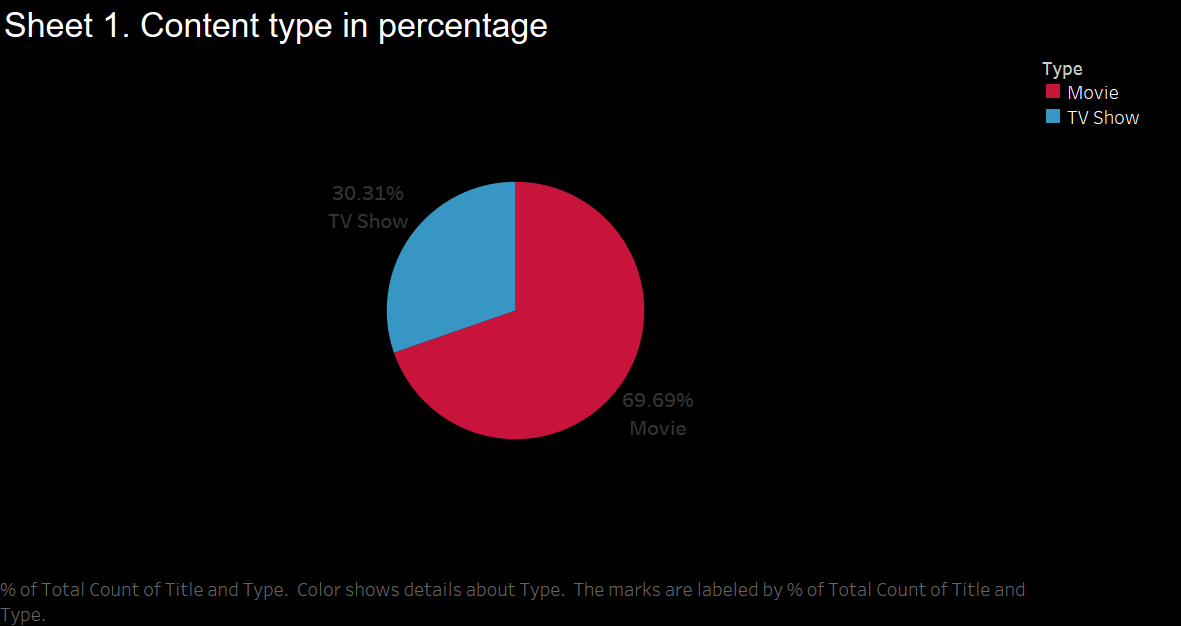
**Inference:** TV-MA rating is recorded as the highest in Movie and TV Show, whereas the lowest rating is in NC-17 and UR in Movie type also TV-Y7-FV in TV Show type.

**Release Year Vs Type:**

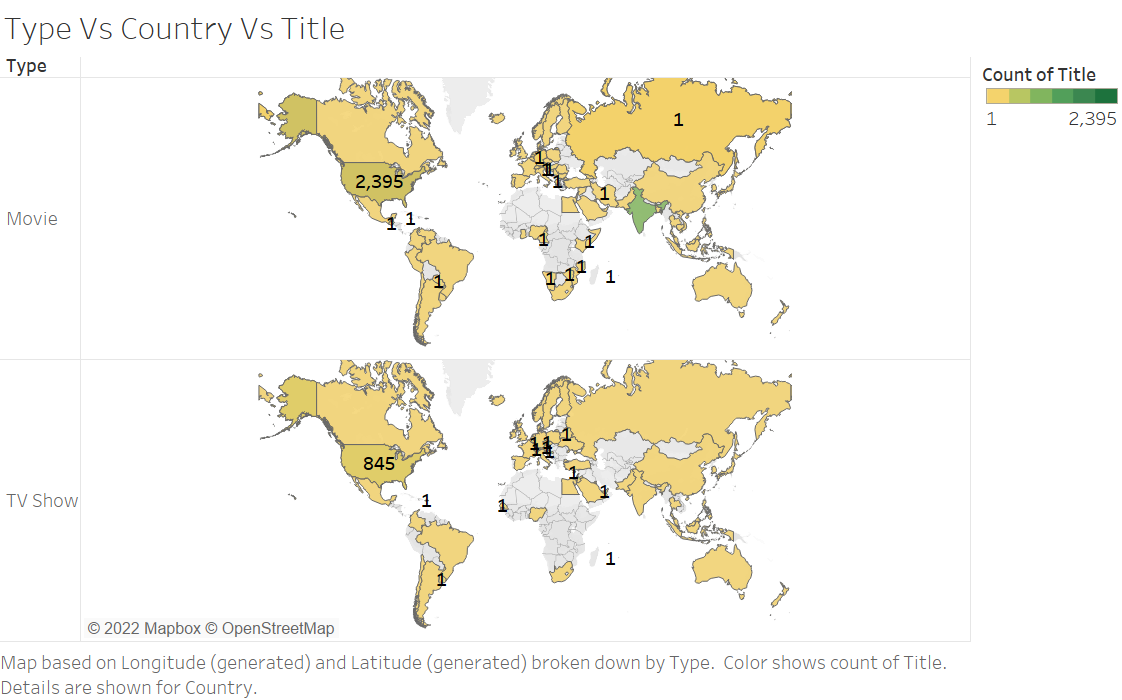


**Inference:** highest movies released in the year 2017 and 2018 approx and 2020 for TV Shows.

**Content-Type:**

****

**Inference:** highest percentage of movies released in Netflix with 69.69%.

****

**Inference:** highest percentage of movies released in Netflix with 69.69%.

**References:**

[**https://wuraolaifeoluwa.medium.com/basic-data-wrangling-and-visualization-of-netflix-data-8b9609328f8c**](https://wuraolaifeoluwa.medium.com/basic-data-wrangling-and-visualization-of-netflix-data-8b9609328f8c)

[**https://deepnote.com/@a\_mas/Investigating-Netflix-Movies-Dataset-f0f270c8-c7e2-46e4-a149-786408ec0804**](https://deepnote.com/@a_mas/Investigating-Netflix-Movies-Dataset-f0f270c8-c7e2-46e4-a149-786408ec0804)

**https://www c.kaggle.com/datasets/shivamb/netflix-shows**

**Conclusion:**

The United States has the highest contribution to Netflix’s media. TV shows with 7 seasons in the month of December (265) has the highest count and the lowest count in the month of February (180) whereas the highest percentage of movies are released with 69.69%.