Capstone Project
NETFLIX MOVIES AND TV SHOWS CLUSTERING

Individual Project
Shubham Naik

Presentation Overview

- Introduction

- Problem statement

- Data Information

- EDA & feature engineering

- Preparing Data for modeling

- Implementing Model

- Model summary

- Conclusion

Introduction

Netflix, Inc. is an American technology and media services provider and production company headquartered in Los Gatos, California. Netflix was founded in 1997 by Reed Hastings and Marc Randolph in Scotts Valley, California. The company's primary business is its subscription-based streaming service, which offers online streaming of a library of films and television series, including those produced in-house.

The dataset consists of details (title,director,cast,duration,rating etc..) of tv shows and movies available on Netflix as of 2019.

Problem statement

Objective: "Which movie will you like" given that you have seen 'The Dark night, Batman Begins, Pineapple express'?

How do we figure this out?
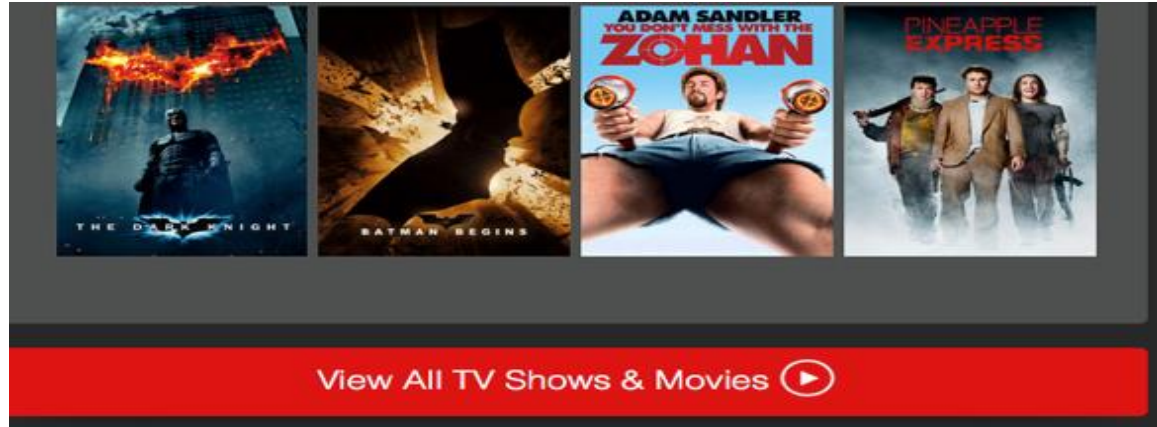
The idea of Clustering is to group the items together based on their attributes. The accuracy of predictions about how much someone is going to love a movie based on their movie preferences.

Data Information

- show_id : Unique ID for every Movie / Tv Show
- type : Identifier – A Movie or TV Show
- title : Title of the Movie / Tv Show
- director : Director of the Movie
- cast : Actors involved in the movie / show
- country : Country where the movie / show was produced
- date_added : Date it was added on Netflix
- release_year : Actual Release year of the movie / show
- rating : TV Rating of the movie / show
- duration : Total Duration – in minutes or number of seasons
- listed_in : Genre
- Description : The Summary description
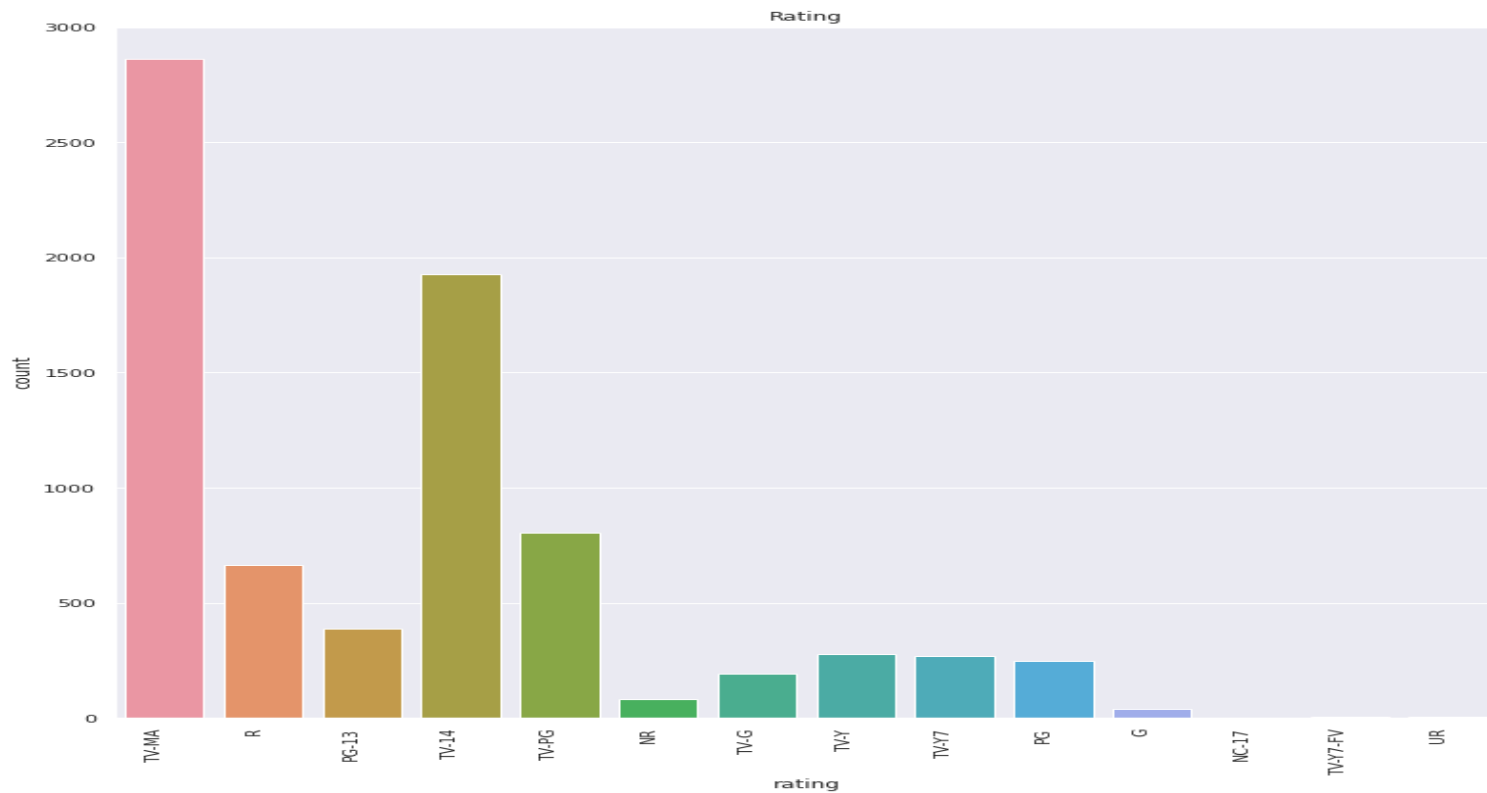
AI

EDA

- <u>Distribution of Type</u>

So there are about 4,000+ movies and almost 2,000 TV shows, with movies being the majority. There are far more movie titles (69%) that TV shows titles (31%) in terms of title.
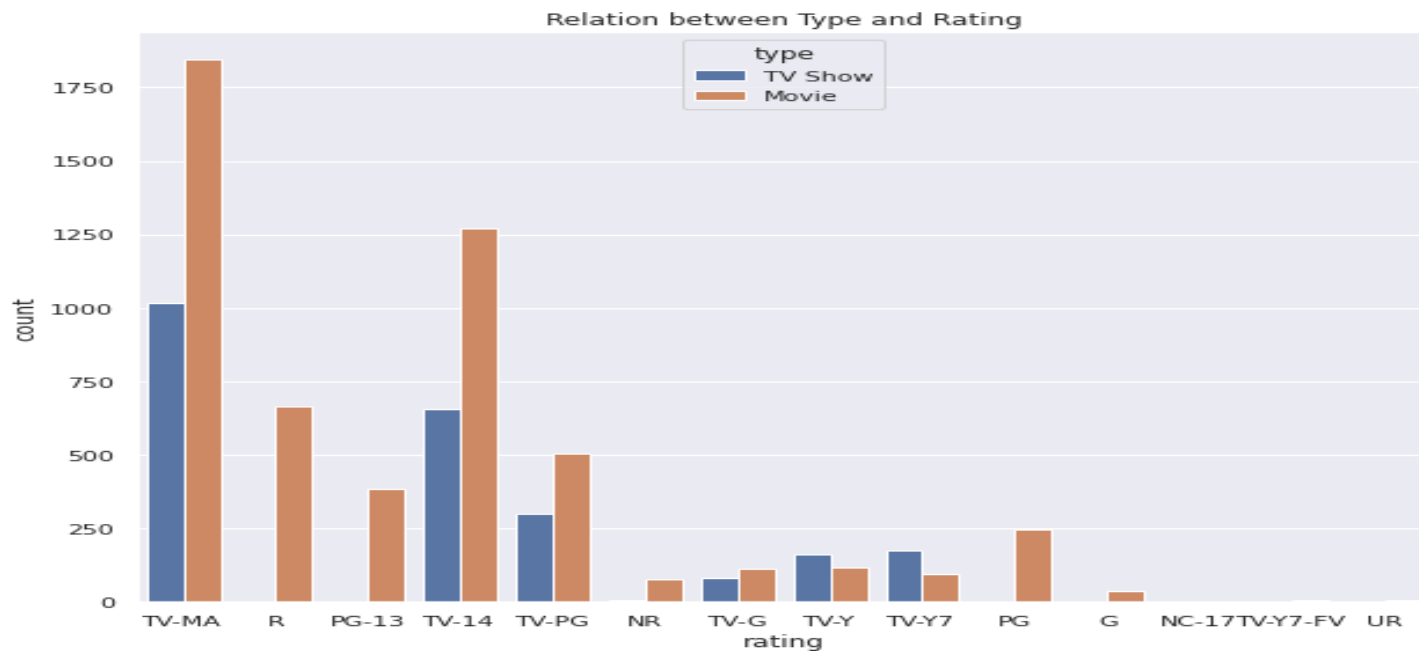

Type of Netflix Content
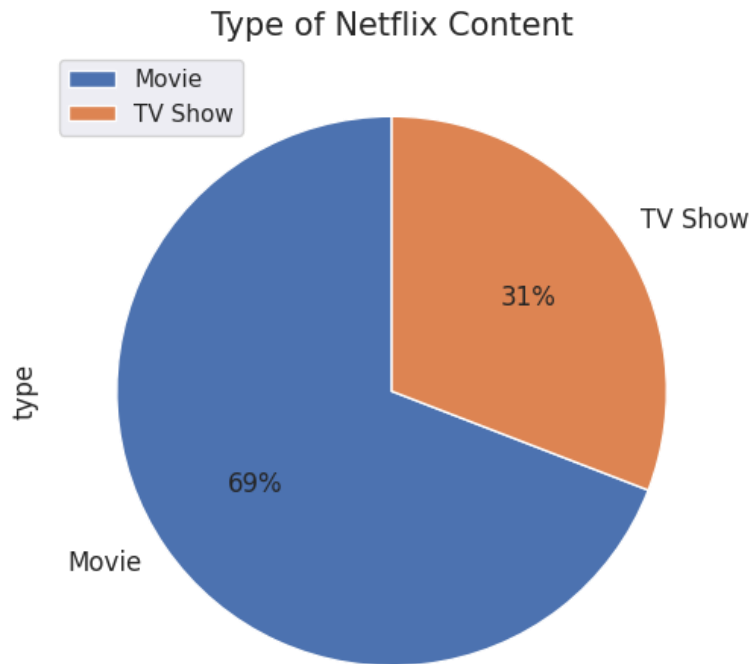
EDA

- <u>Rating of shows and movies</u>

EDA
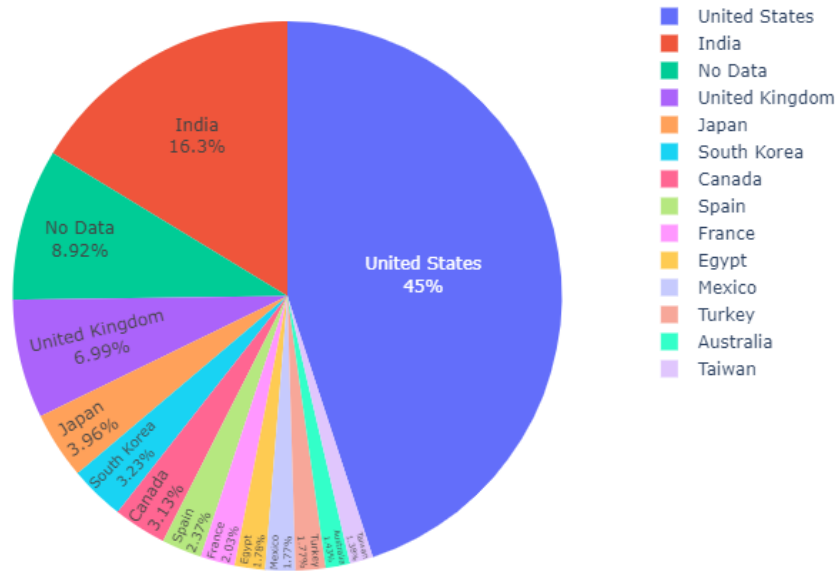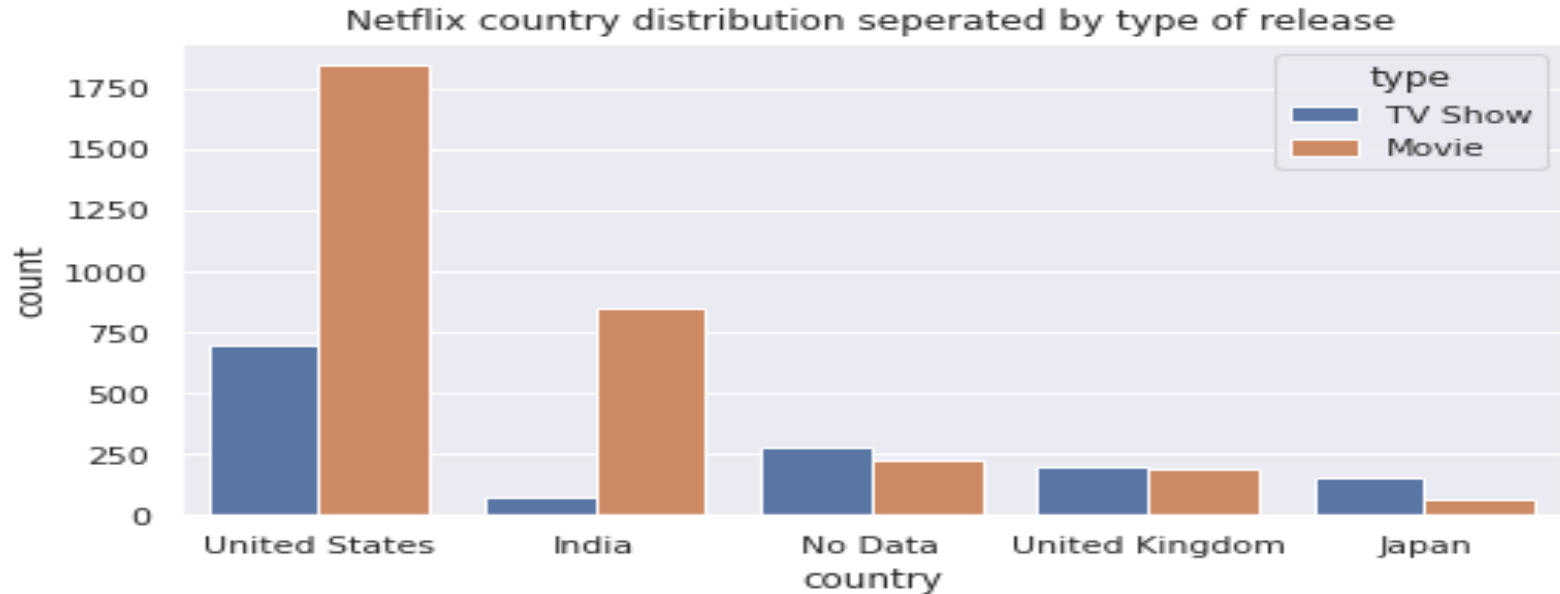
- ● <u>Relation between Type and Rating</u>

EDA

- **Country-wise Content Production**

The majority of the content providers are in the above top-ten countries. Among which USA, India, and UK create more than half of the tv shows and movies on the platform.


Legend:
- United States
- India
- No Data
- United Kingdom
- Japan
- South Korea
- Canada
- Spain
- France
- Egypt
- Mexico
- Turkey
- Australia
- Taiwan

Pie chart values:
- United States 45%
- India 16.3%
- No Data 8.92%
- United Kingdom 6.99%
- Japan 3.96%
- South Korea 3.23%
- Canada 3.13%
- Spain 2.37%
- France 2.03%
- Egypt 1.78%
- Mexico 1.77%

EDA

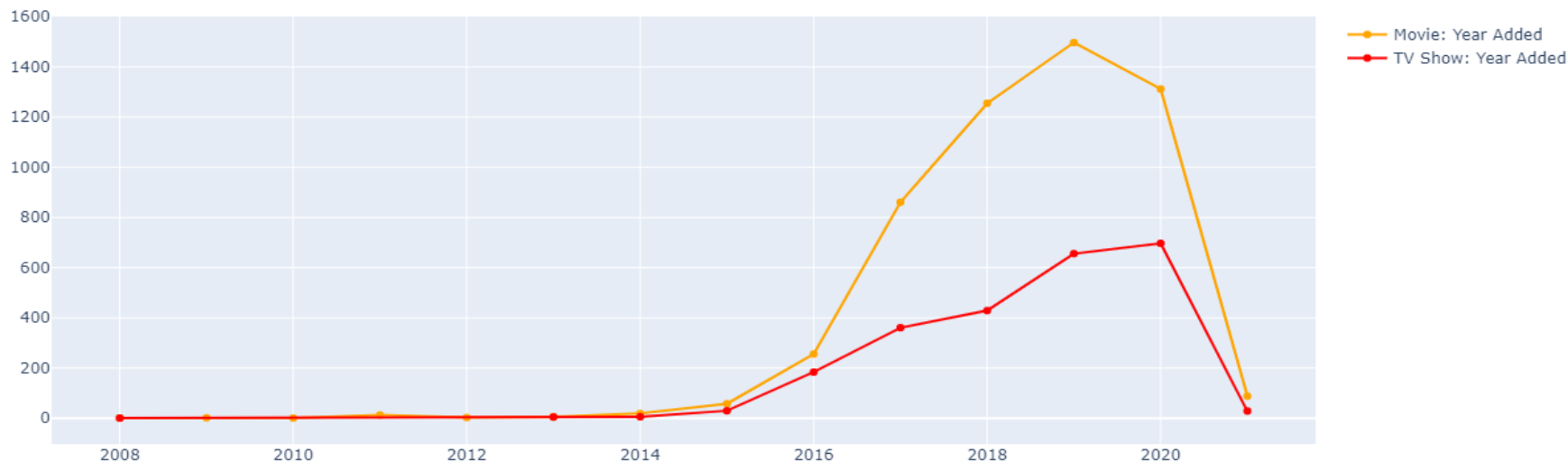- Top 5 countries separated by type of release

EDA

- **<u>Content released over the years</u>**
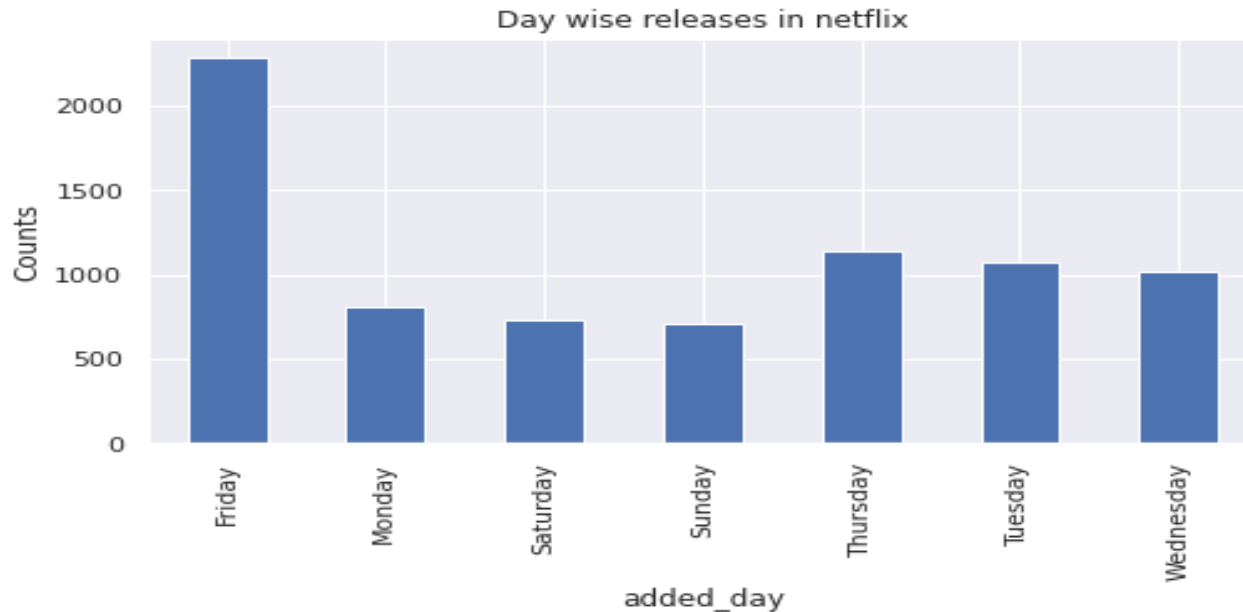
Maximum movies are released in 2017.

EDA
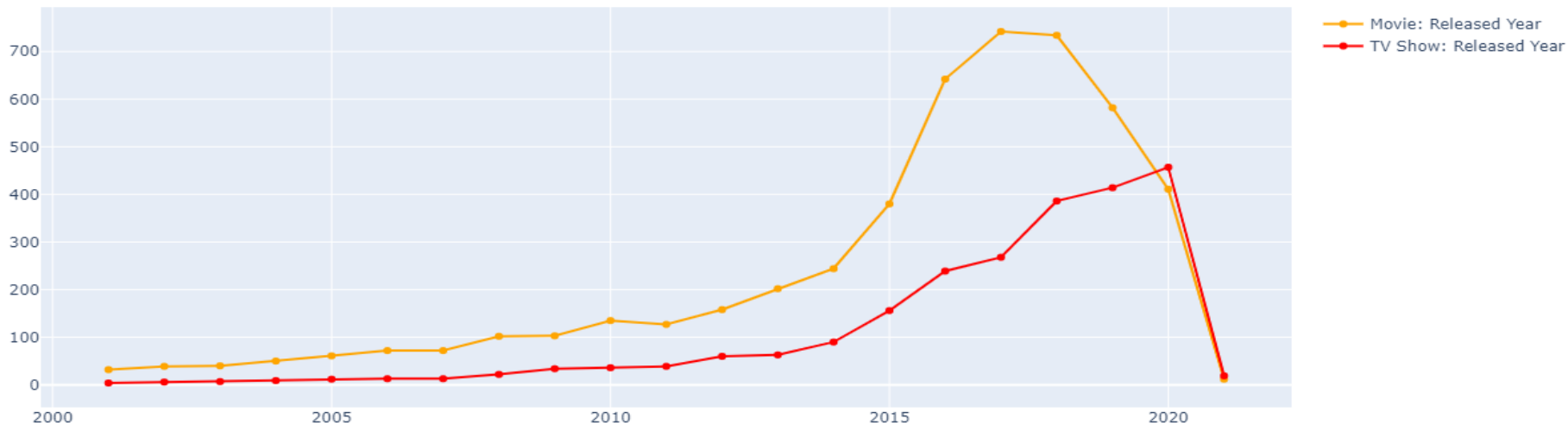
- **Content added over the years**

From 2016 we can see a noticeable addition in the number of movies and tv shows uploaded by Netflix on its platform.
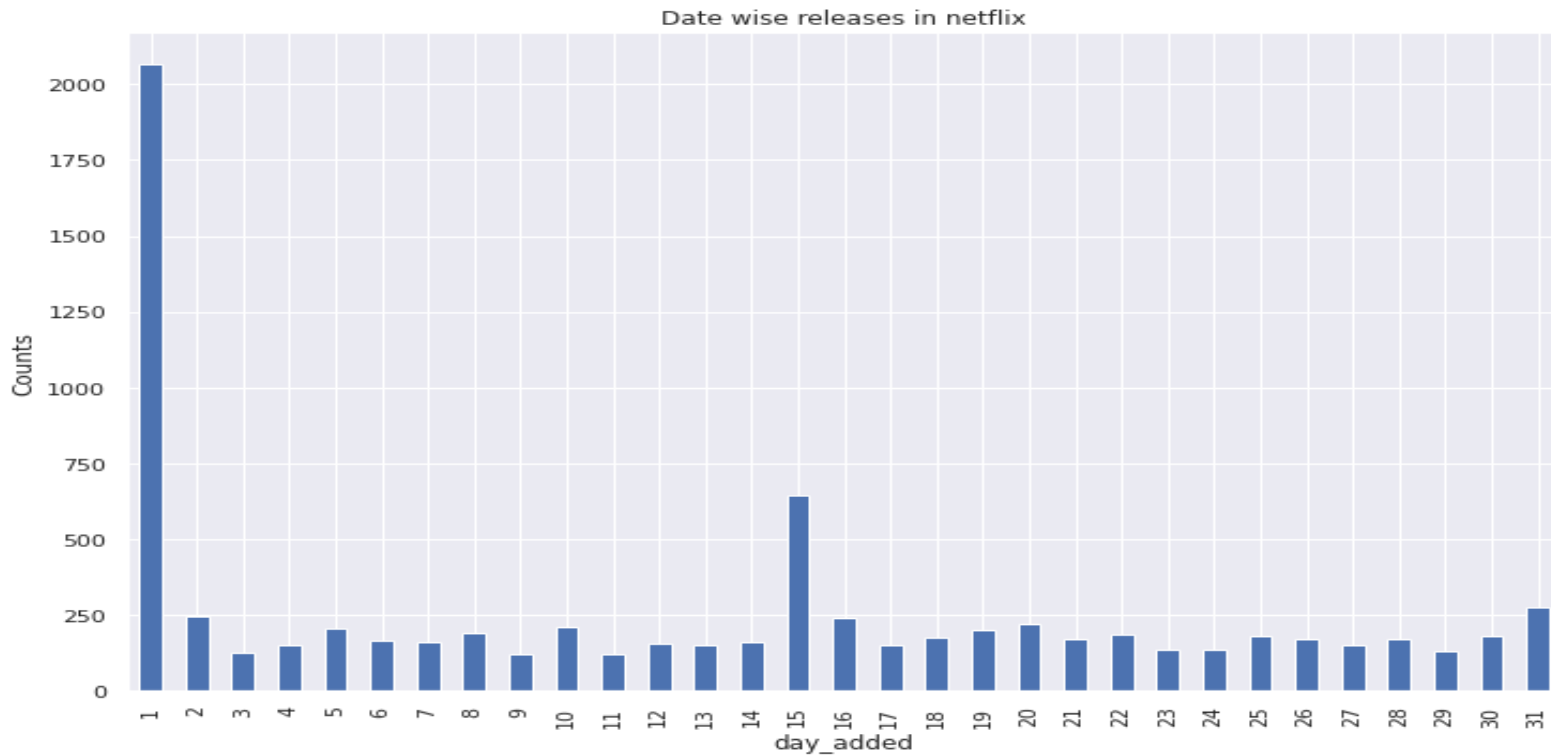
EDA

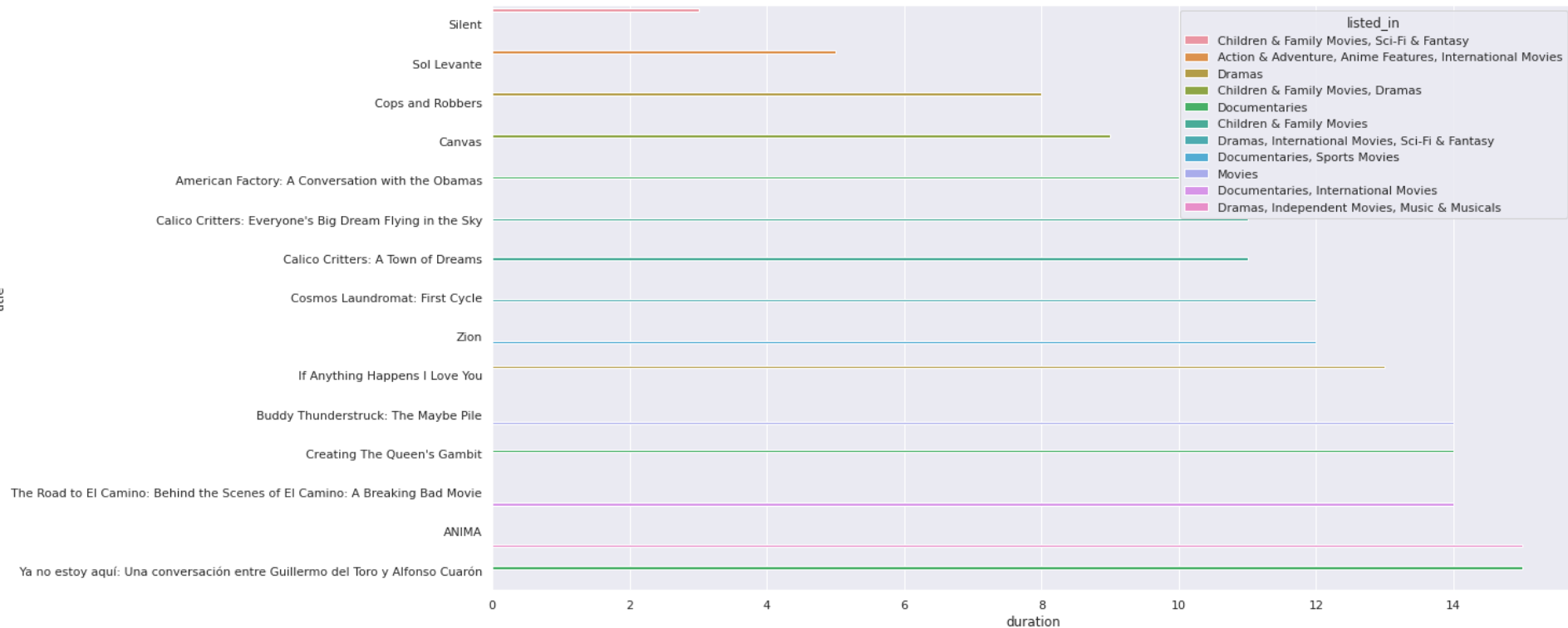- **Day wise addition of movies and shows to the platform**

Day wise releases in netflix

EDA

- **Month wise addition of movies and shows to the platform**

Month wise releases in netflix

EDA

- <u>**Date wise addition of movies and shows to the platform**</u>

Date wise releases in netflix

EDA
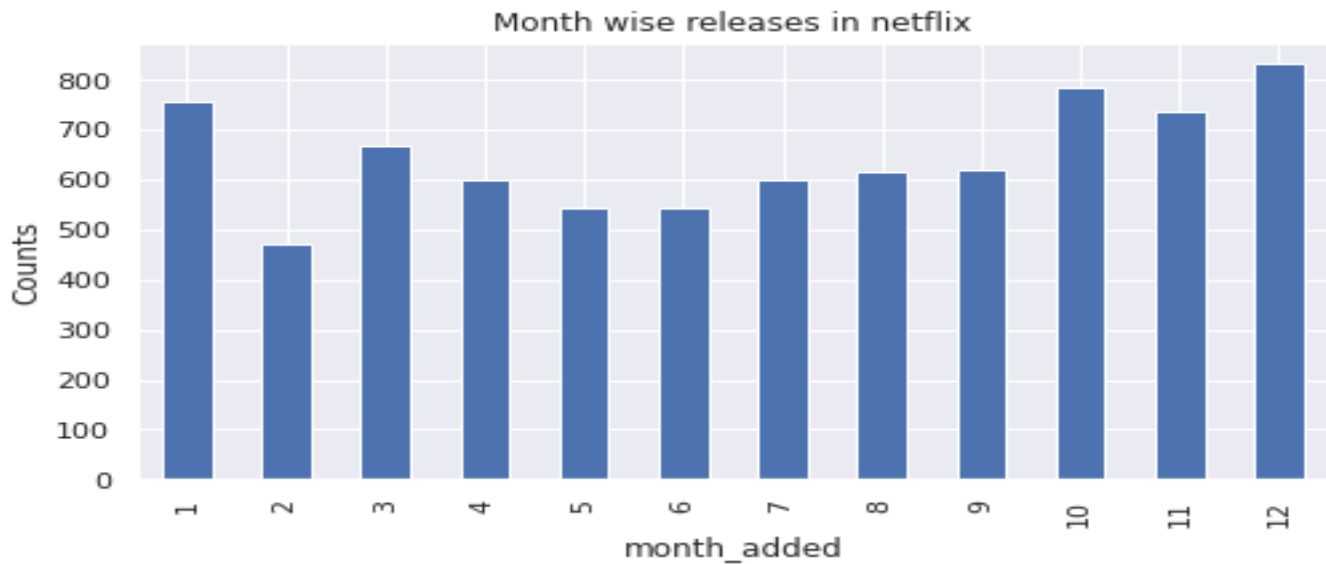
- <u>Movies that take less amount of time</u>

EDA

- **Duration of Tv Shows**

Duration of Tv Shows.
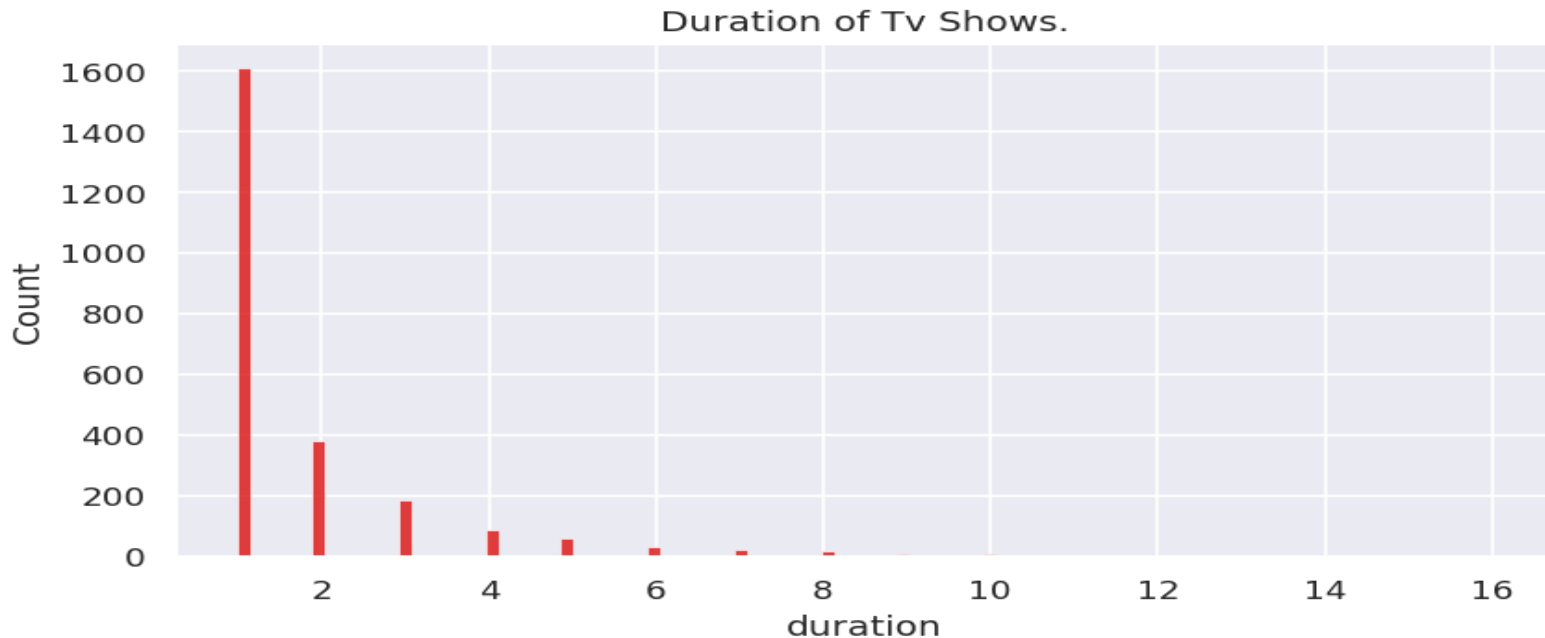
Most of the Tv Shows last for 1 or 2 seasons
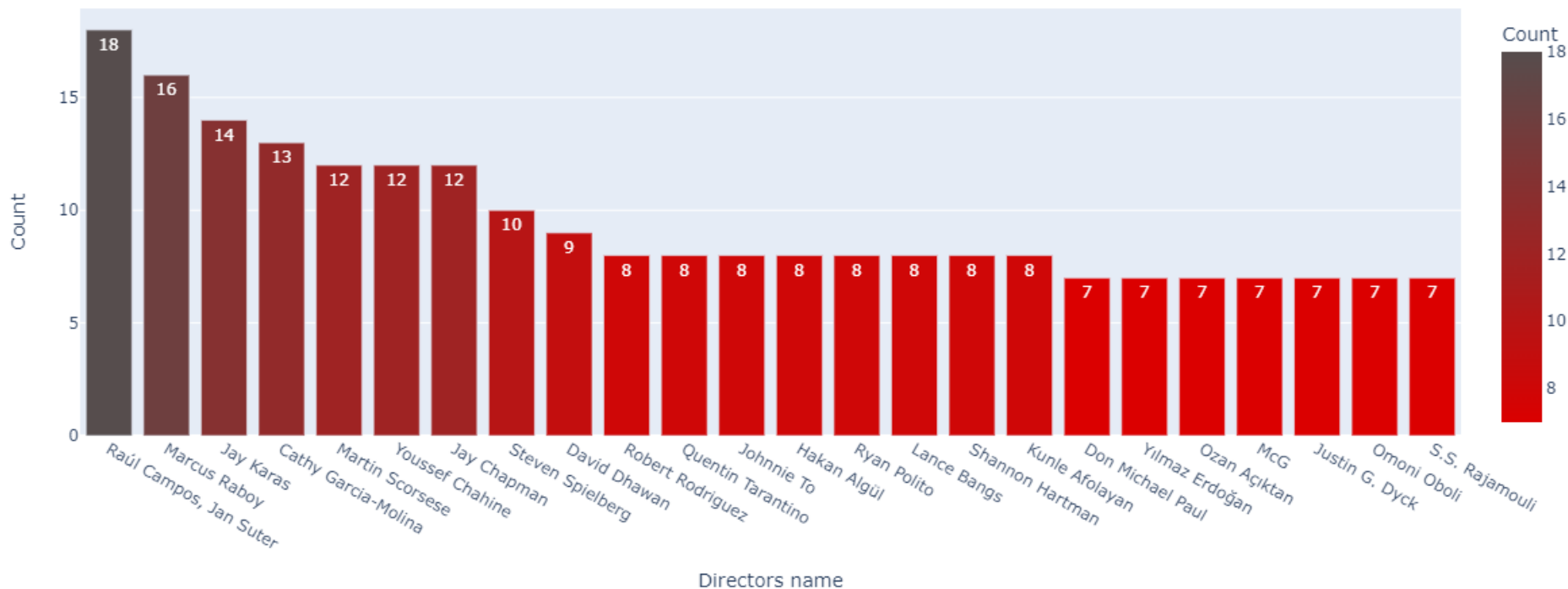
EDA

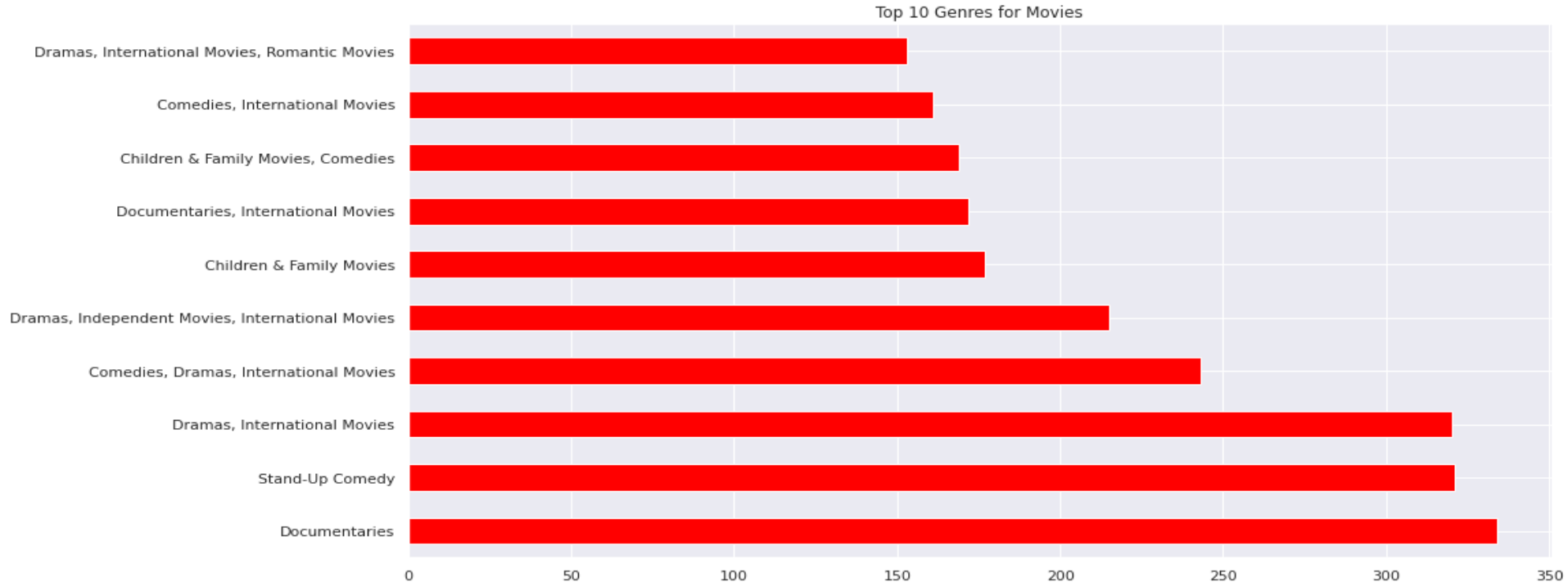- <u>Duration of movies</u>

Duration of movies
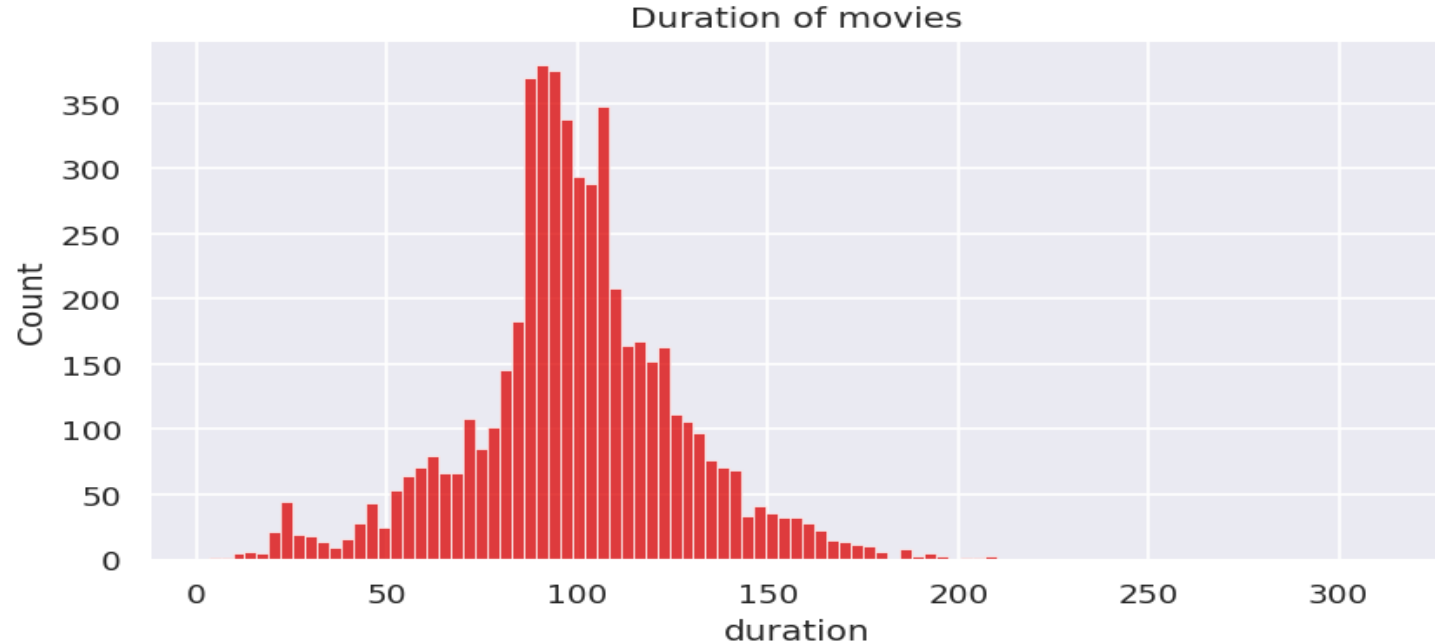
Most of the movies last for 90 to 120 minutes.

EDA

- <u>**Top 25 directors**</u>

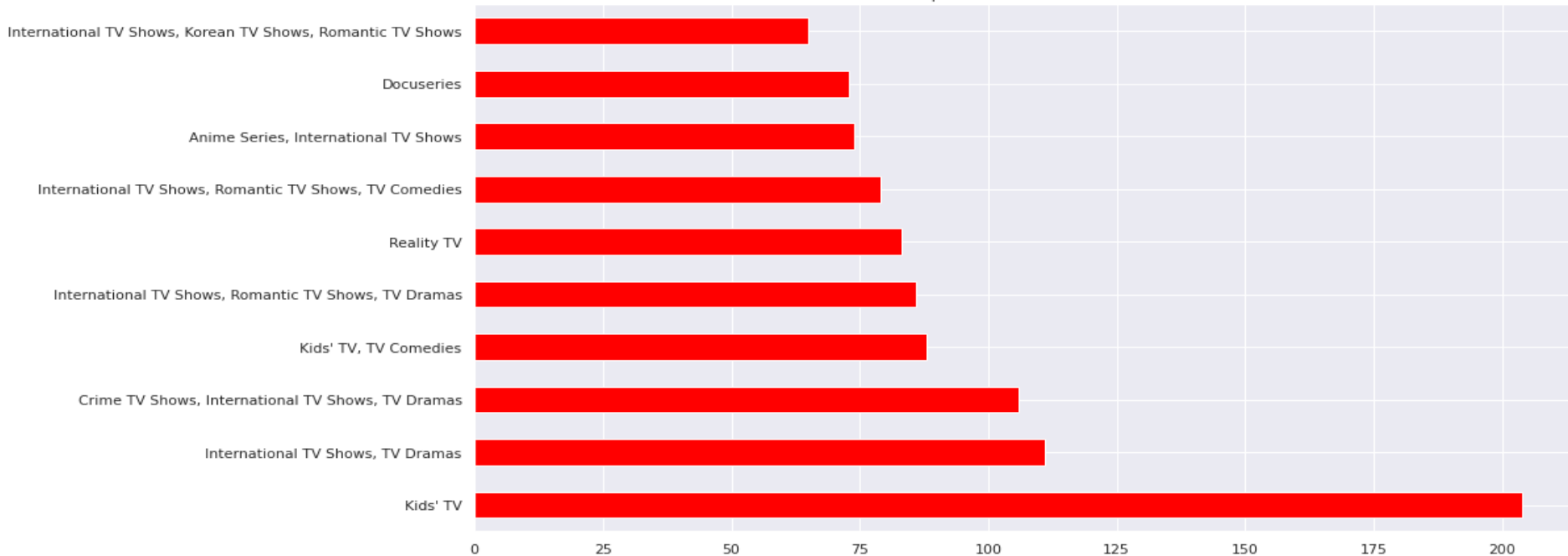
Top 25 directors with highest number of Movies and Tv Shows.
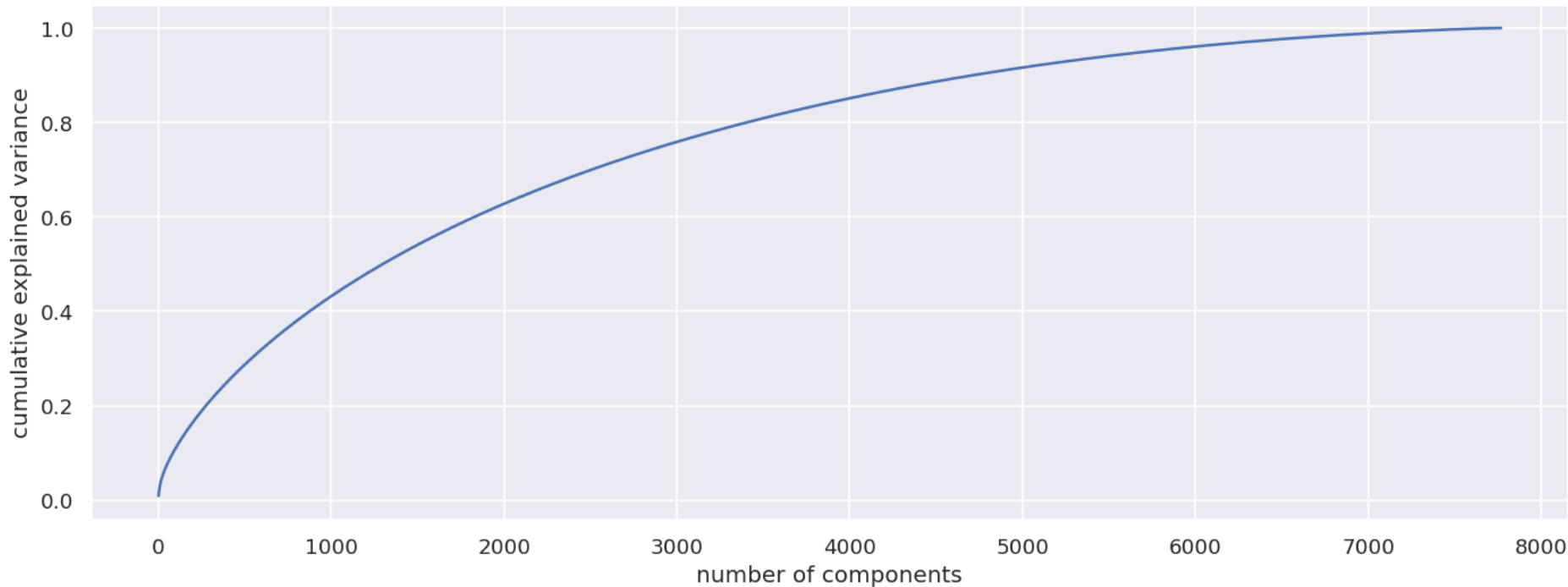
EDA

- [Genres for Movies](#)

Top 10 Genres for Movies

EDA

- [Genres for TV shows](#)

Top 10 Genres for TV shows

EDA

● <u>Most used words in title</u>

Most repeated words in title include
Christmas, Love, World, Man, and
Story.
We saw that most of the movies and
tv shows got added during the
winters, which tells why Christmas
appeared many times in the titles.

Dimensionality Reduction

We can see from the above plot almost 90% of the variance can be explained by 5000 components. Since choosing 5000 could be tricky we will set the value to be 90% in sklearn.
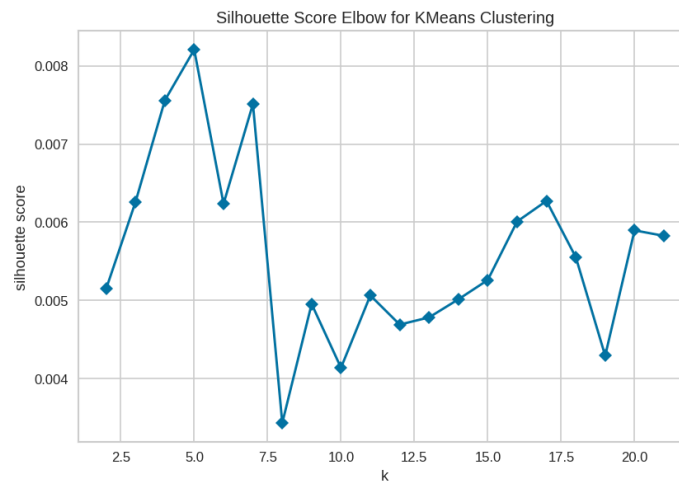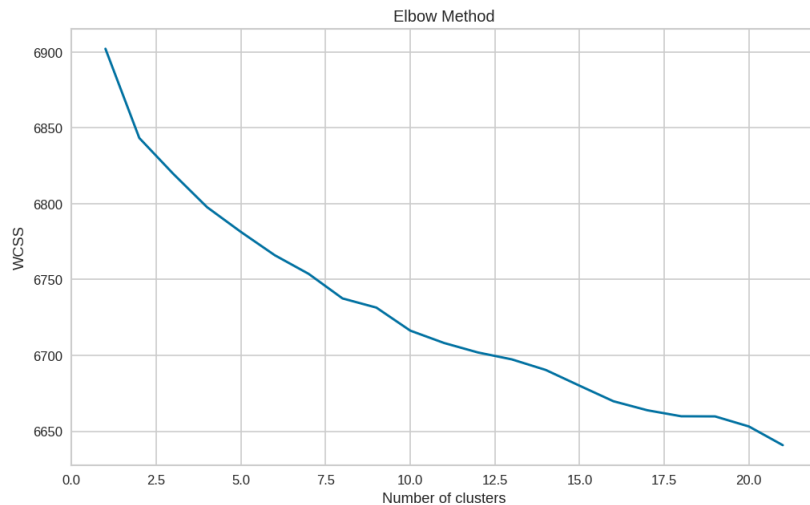
K-Means Clustering with Silhouette

- For n_clusters = 5, silhouette score is 0.00820712709740042

Silhouette Plot of KMeans Clustering for 7770 Samples in 5 Centers

K-Means Clustering with Elbow method

Elbow Method

Silhouette Score Elbow for KMeans Clustering

Hyper parameter
{ n_clusters=[2,22], init='k-means++',
max_iter=300, n_init=10
random_state=0}

Clustering

Hyper parameter
{n_clusters=5, init= 'k-means++', random_state=0}
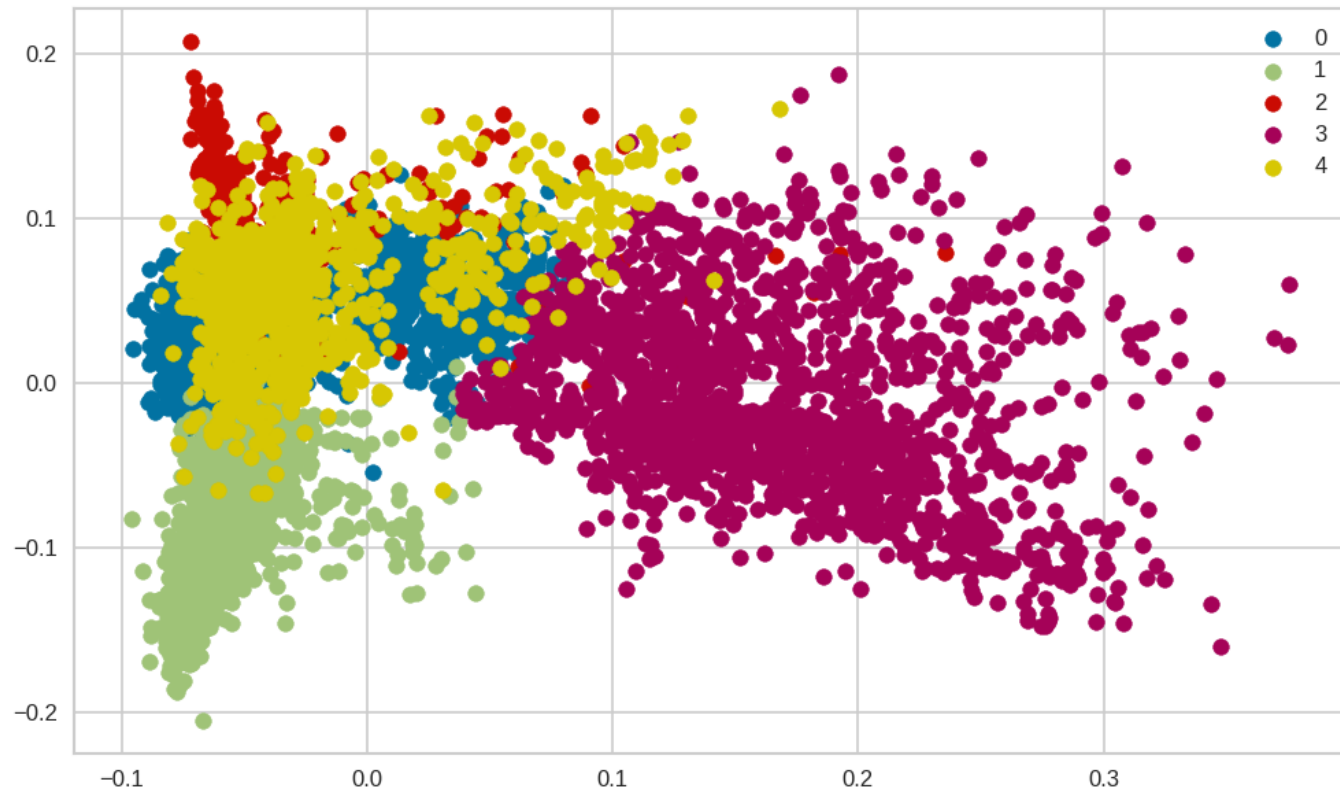
Conclusion

- Nearly 70 percent are movies and 30 percent are TV Shows
- After 2016 Movies and Tv Shows added maximum
- There are two different type of time durations for Movies it's in minutes and for TV Shows it's in season
- Maximum movies are in the range of 90 to 120 minutes
- Most of the children, sci-fi & Fantasy movies and documentaries take less amount of time
- K-Means Clustering with Silhouette gives the highest score of 82% for number of clusters 5

THANK YOU