ML Unsupervised – Capstone Project

Abhishek Kumar
Mohita Rathour

Introduction

Netflix is the world's leading streaming platform with over 232 million paid memberships in over 190 countries enjoying TV series, documentaries and feature films across a wide variety of genres and languages. Members can play, pause and resume watching as much as they want, anytime, anywhere, and can also change their plans at any time.

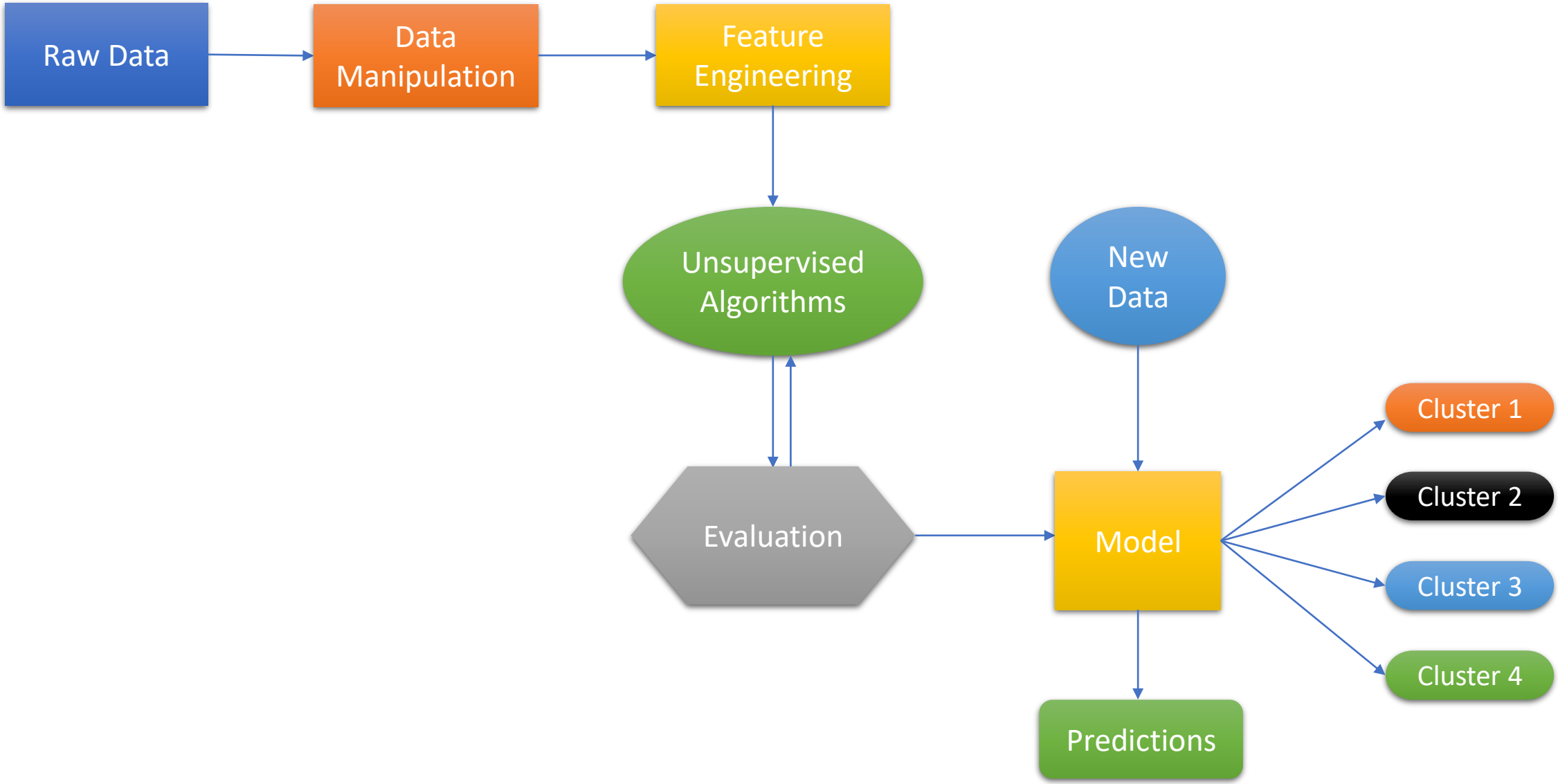
This dataset consists of TV shows and Movies available on Netflix as of 2019. It is collected from Flixable which is a third-party Netflix search engine.

In 2018, they released an interesting report which shows that the number of TV shows on Netflix has nearly tripled since 2010. The streaming service's number of movies has decreased by more than 2,000 titles since 2010, while its number of TV shows has nearly tripled.

In this project, we are required to do

1. Exploratory Data Analysis
2. Understanding what type content is available in different countries
3. Is Netflix has increasingly focus on TV rather than movies in recent years.
4. Clustering similar content by matching text-based features

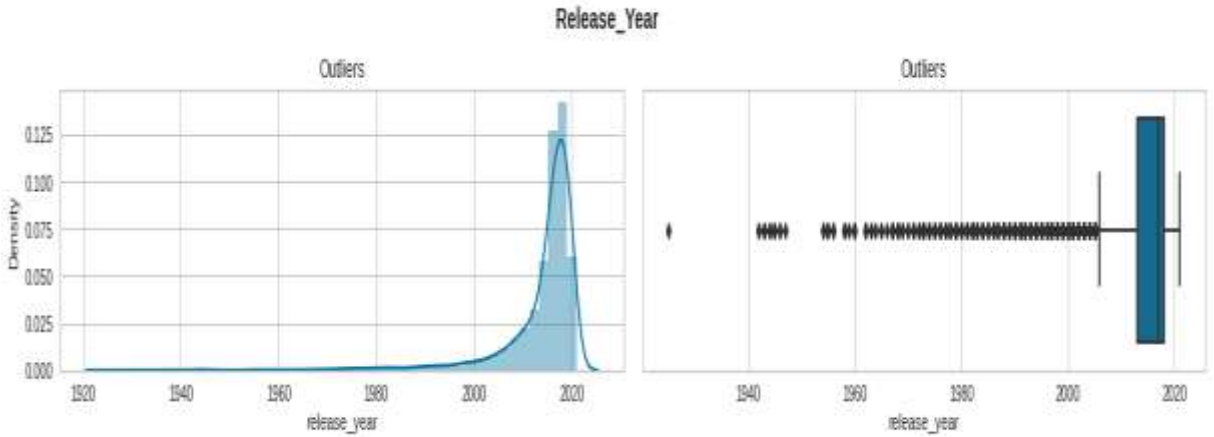
Workflow

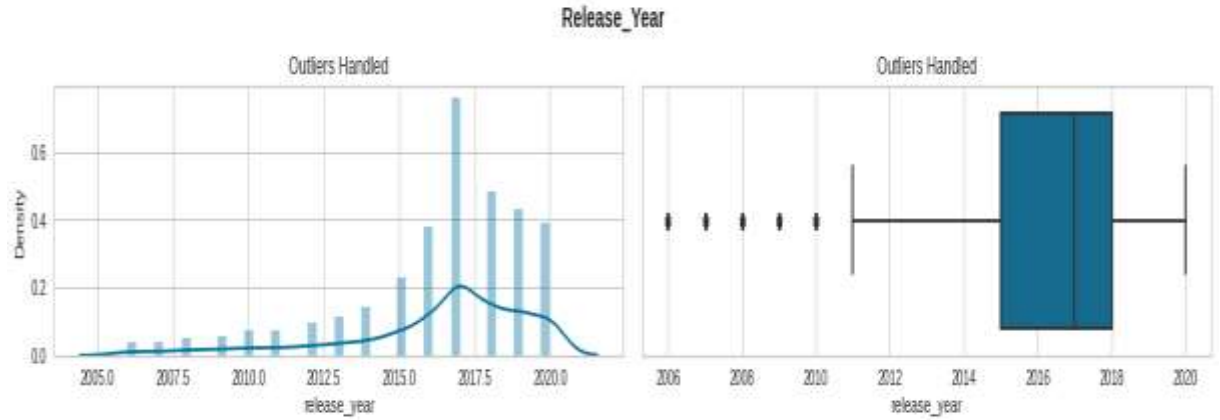

Raw Data
Data Manipulation
Feature Engineering
Unsupervised Algorithms
New Data
Evaluation
Model
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Predictions

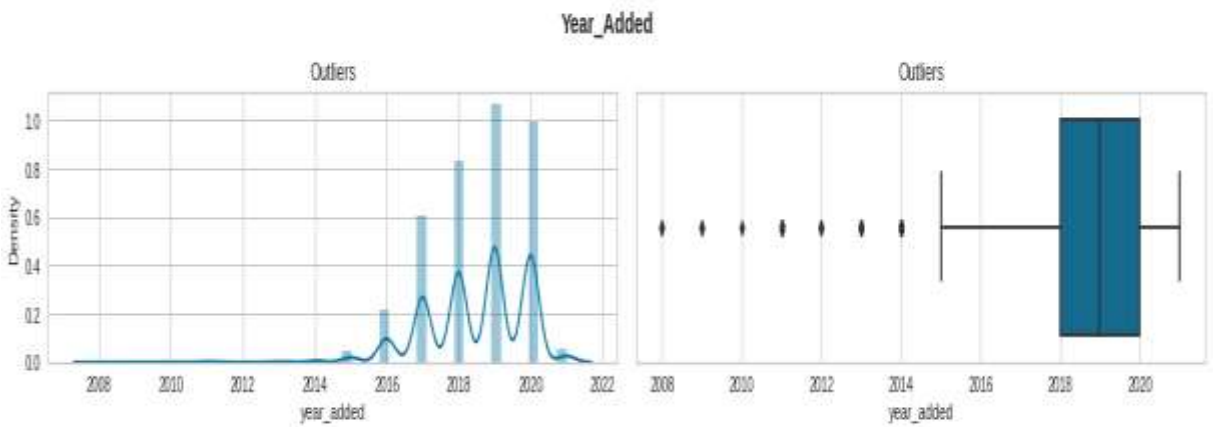
Attributes

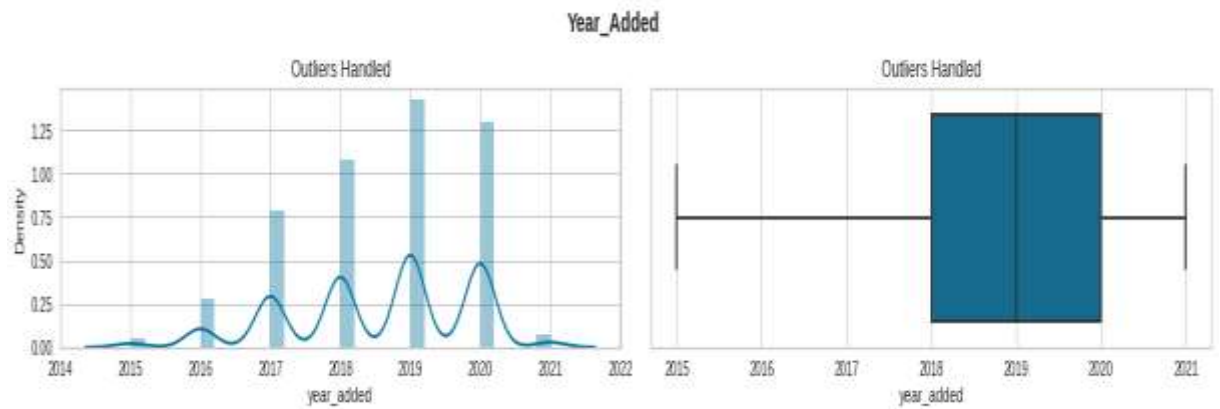
1. show_id : Unique ID for every Movie / Tv Show
2. type : Identifier - A Movie or TV Show
3. title : Title of the Movie / Tv Show
4. director : Director of the Movie
5. cast : Actors involved in the movie / show
6. country : Country where the movie / show was produced
7. date_added : Date it was added on Netflix
8. release_year : Actual Releaseyear of the movie / show
9. rating : TV Rating of the movie / show
10. duration : Total Duration - in minutes or number of seasons
11. listed_in : Genere
12. description: The Summary description

Handling Outliers

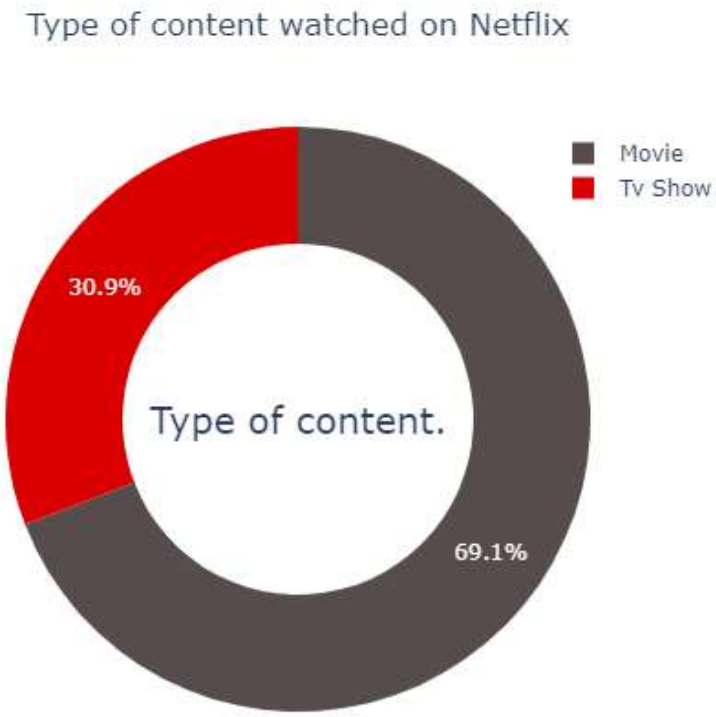






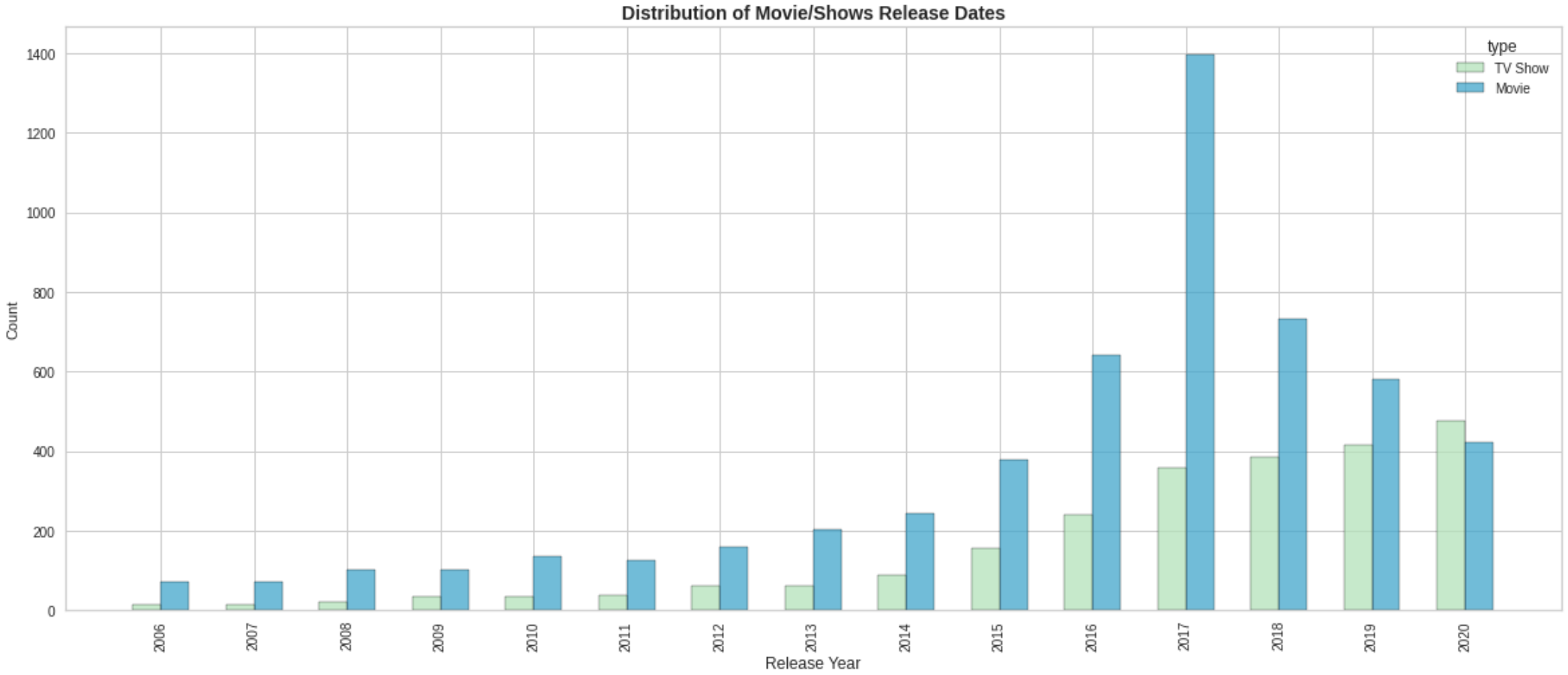


Type of Content on Netflix



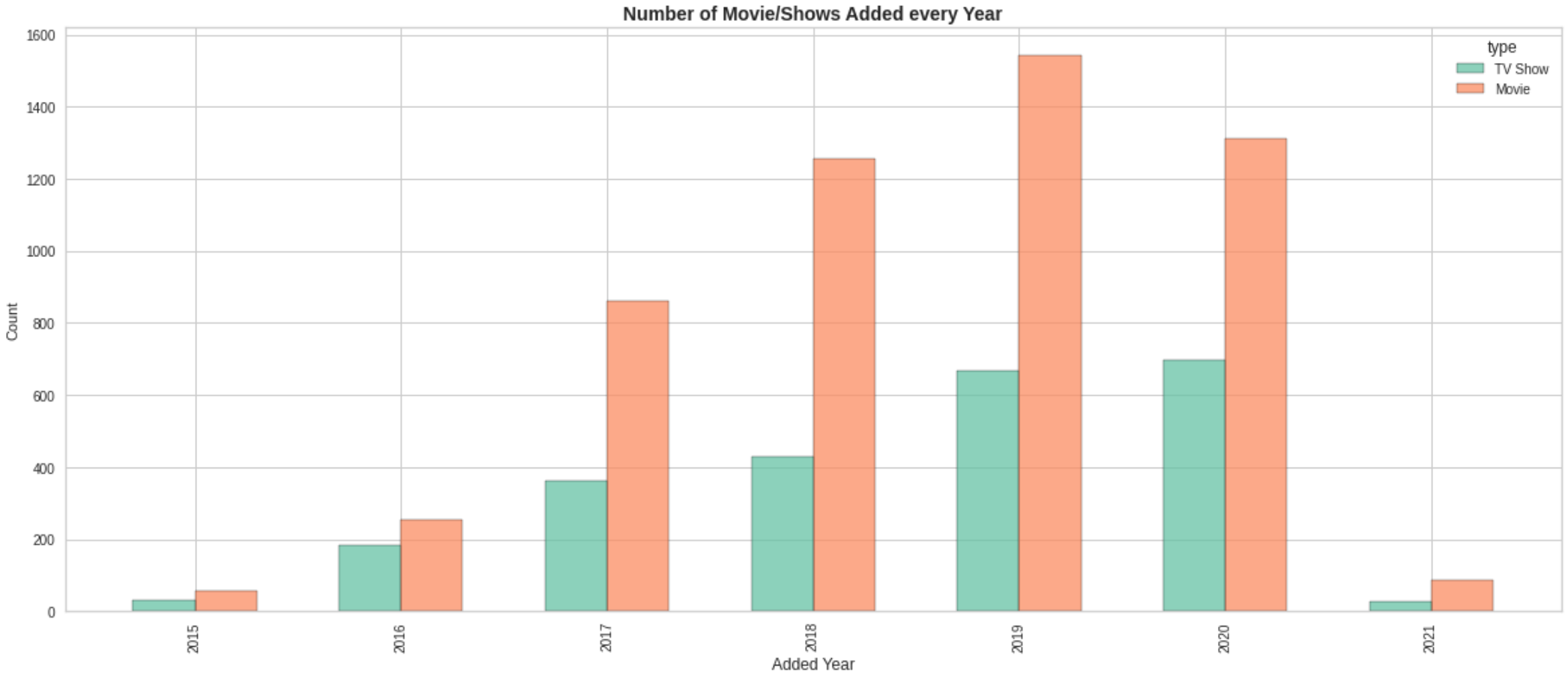
69.1% of the content available on Netflix are movies and the remaining 30.9% are TV Shows

Release year of Movies/TV Shows



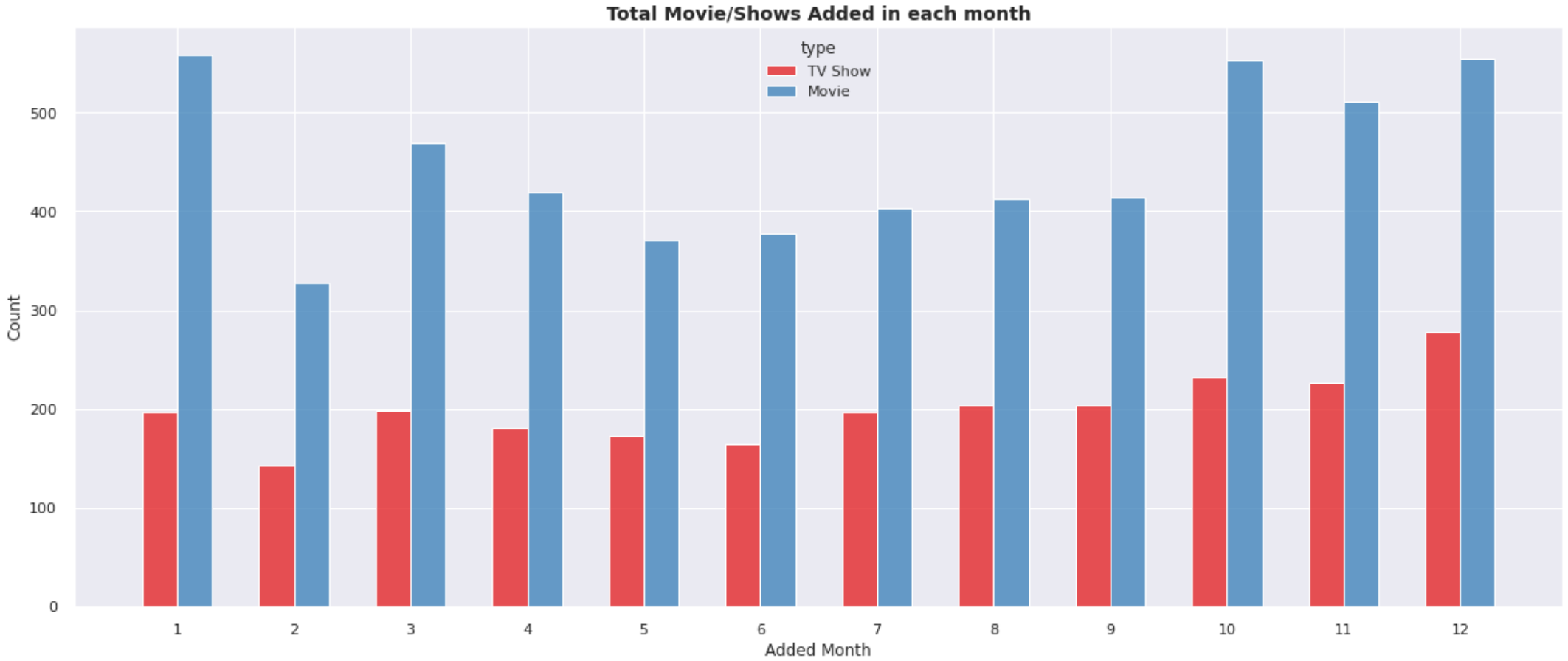
- Maximum number of Movies streaming on the platform were released in 2017.
- Most TV Shows streaming on the platform were released after 2015
- Since the number of movies releasing each year has started decreasing after 2017 whereas number of TV Shows have increased gradually after 2015.

Number of Movies/TV Shows added per Year



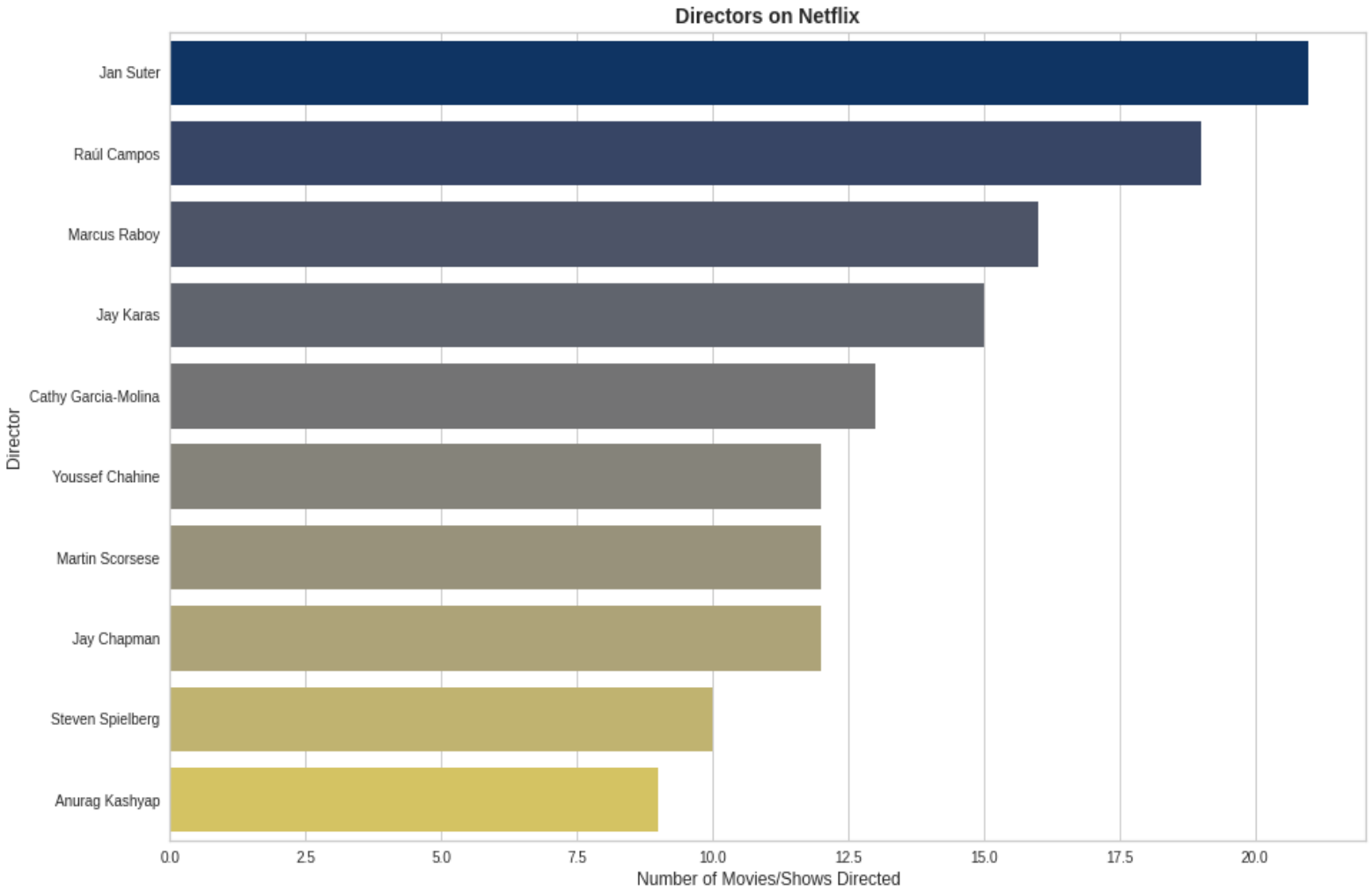
- Number of movies added to the platform showed a deliberate increase after 2017 to 2019 and has been decreased after that.
- Whereas TV Shows have been added continuously from 2015 and its number been increased every year.

Number of Movies/TV Shows added per Month



Most number of Movies and TV Shows are added between October and January

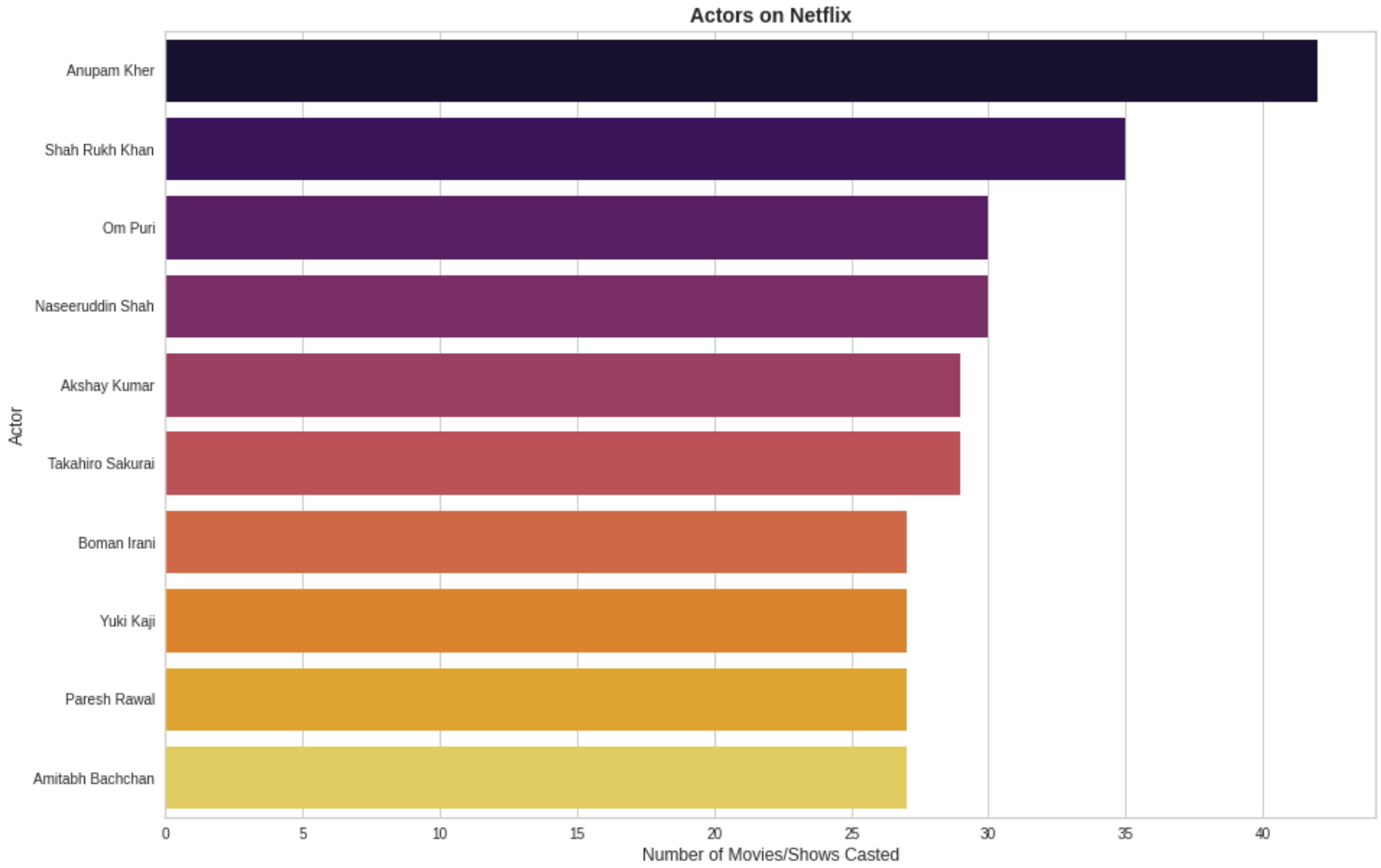
Top 10 Directors



Top 10 Directors are-

1. Jan Suter
2. Raul Campos
3. Marcus Raboy
4. Jay Karas
5. Cathy Garcia Molina
6. Youssef Chahine
7. Martin Scorsese
8. Jay Chapman
9. Steven Spielberg
10. Anurag Kashyap

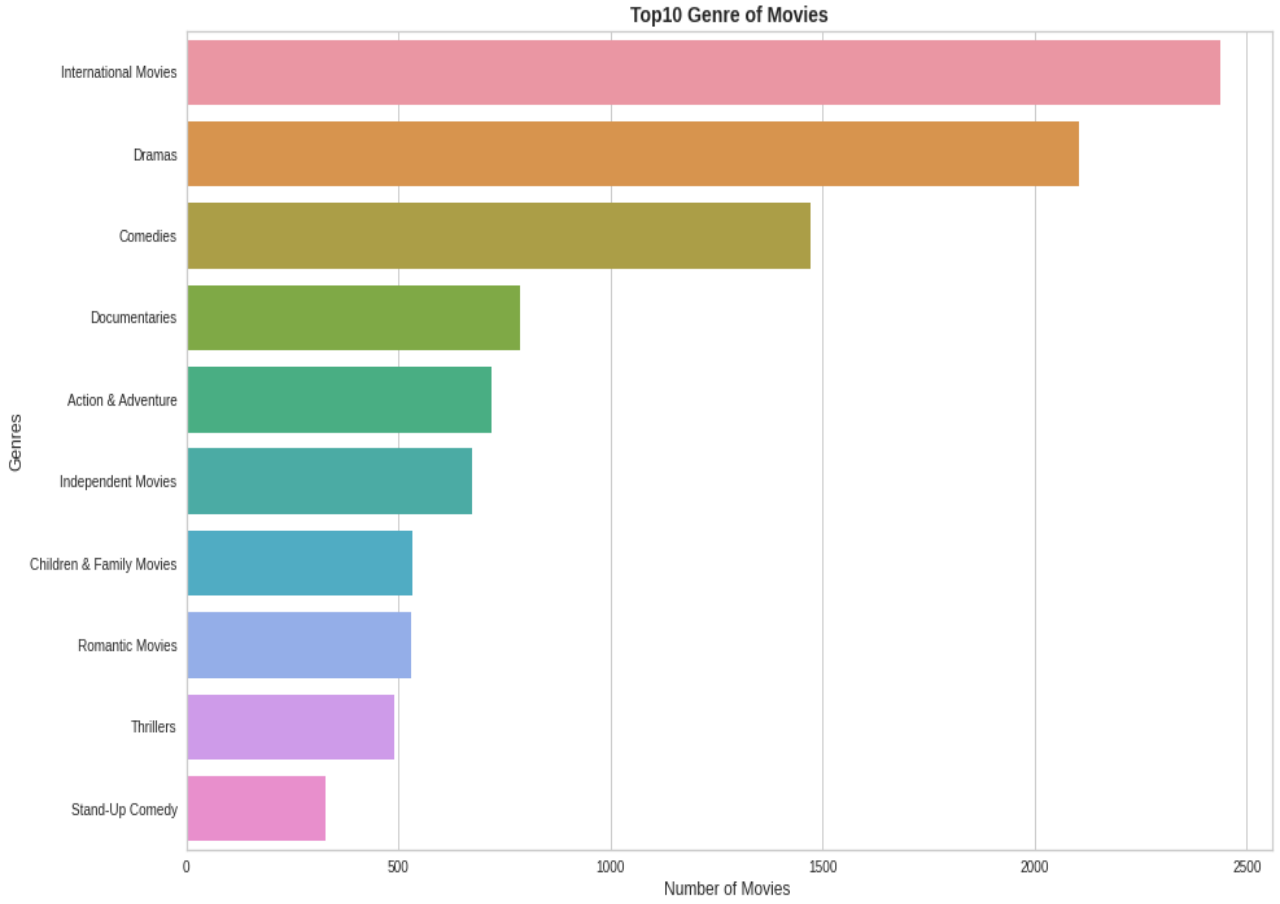
Top 10 Actors

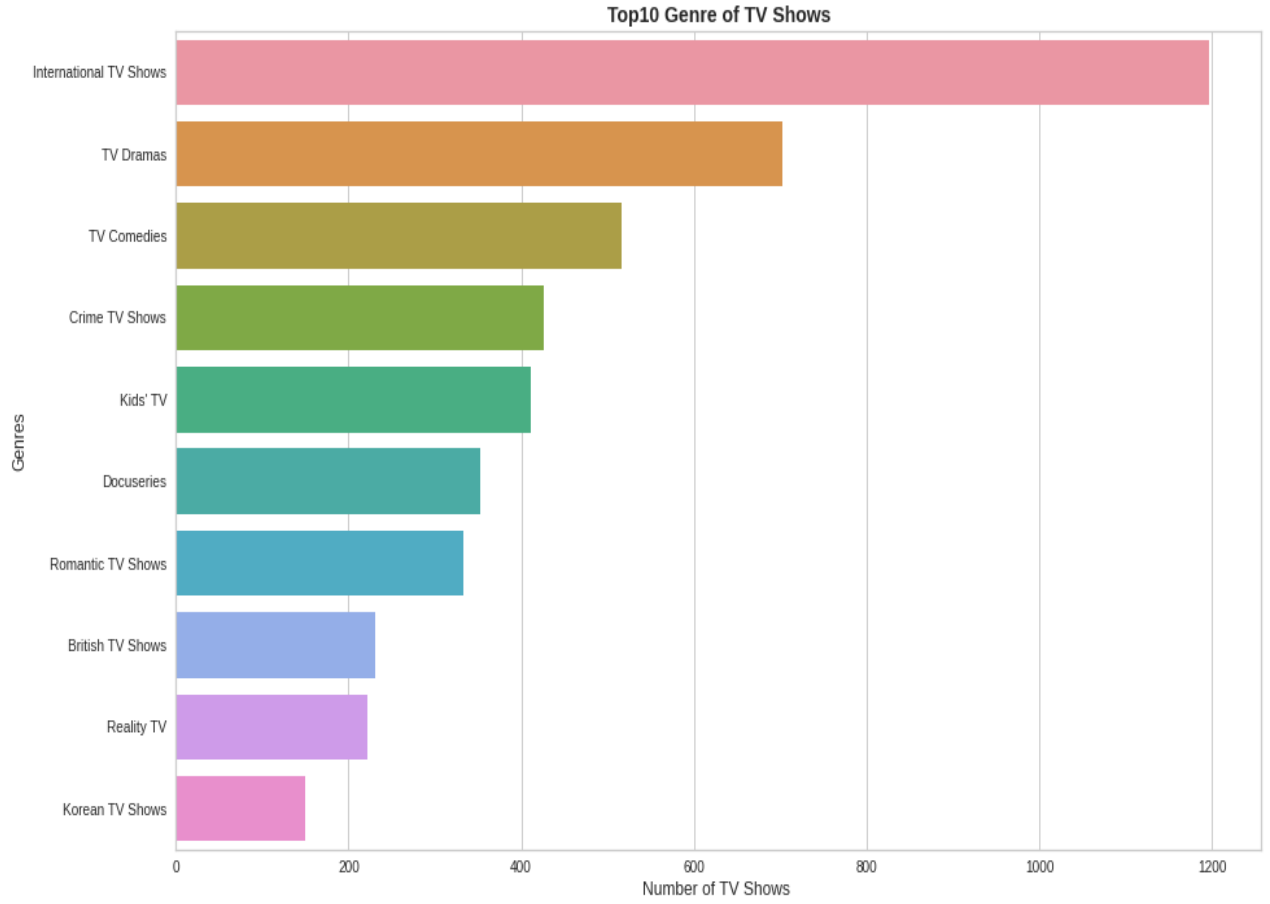


Top 10 Actors are-

1. Anupam Kher
2. Shah Rukh Khan
3. Om Puri
4. Naseeruddin Shah
5. Akshay Kumar
6. Takahiro Sakurai
7. Boman Irani
8. Yuki Kaji
9. Paresh Rawal
10. Amitabh Bachchan

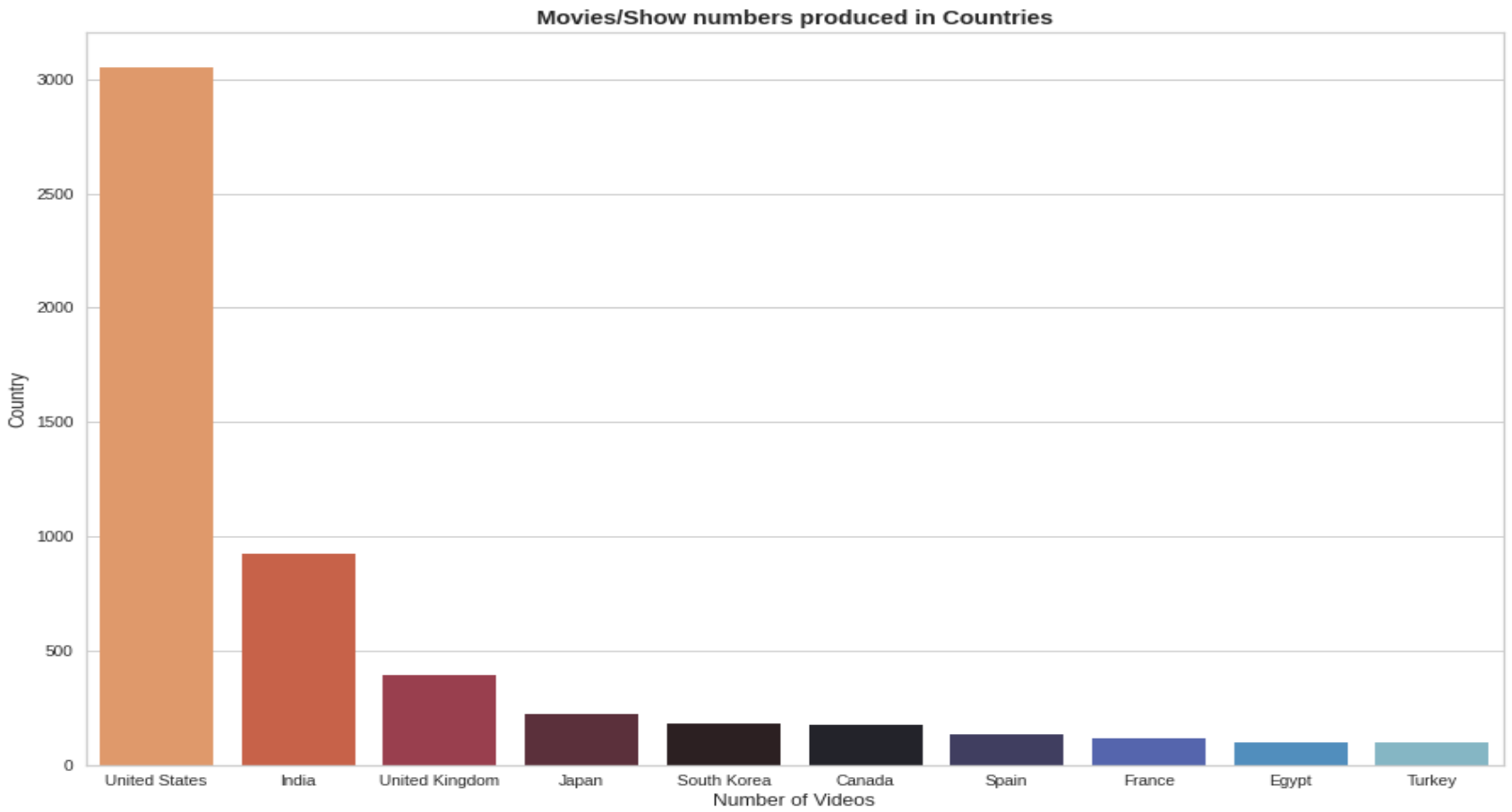
Top Genres of Netflix





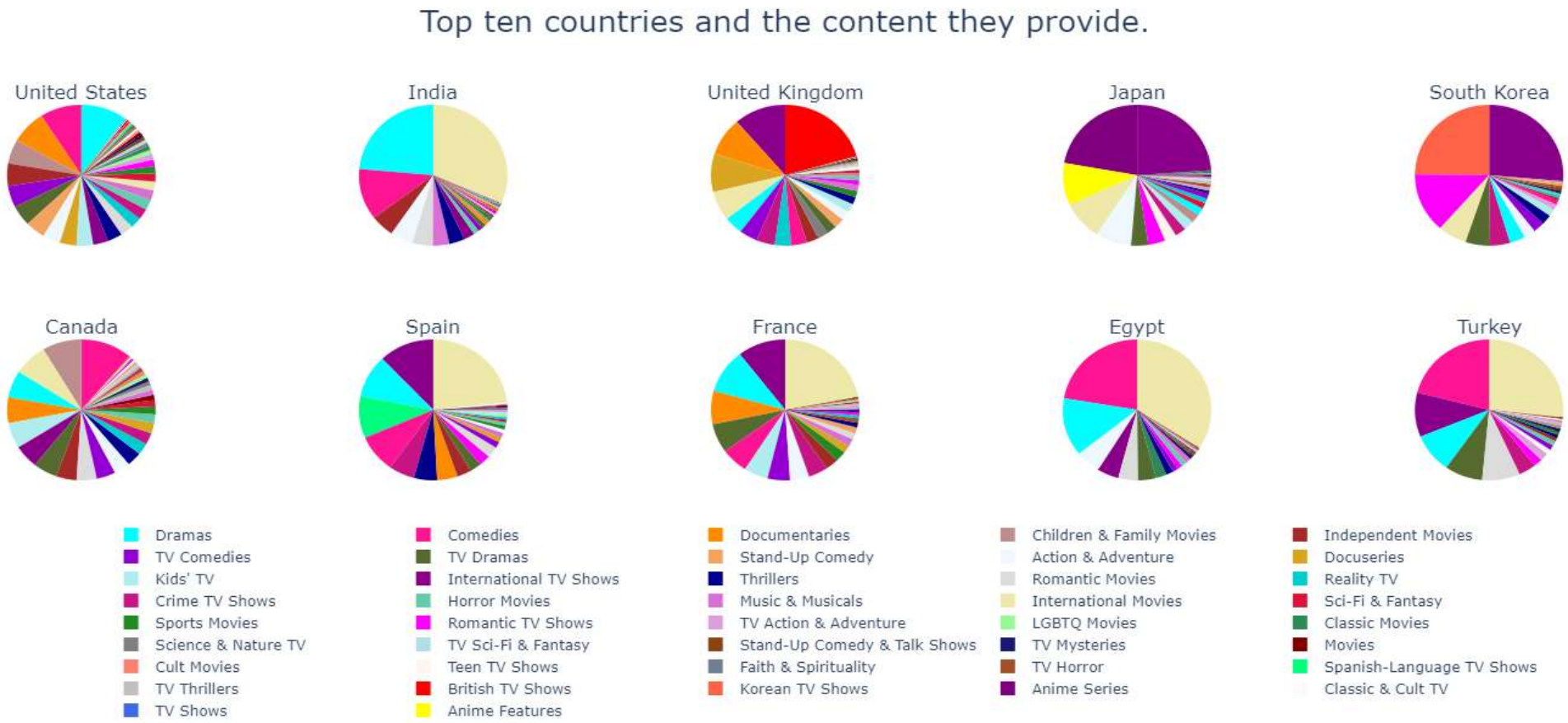
In both Movies and TV Shows top genres are International Movies/Shows, Dramas and Comedies.

Top 10 Countries producing content on Netflix



United States is the country producing maximum content on Netflix followed by India and UK.

What Type of content produced by Top 10 Countries

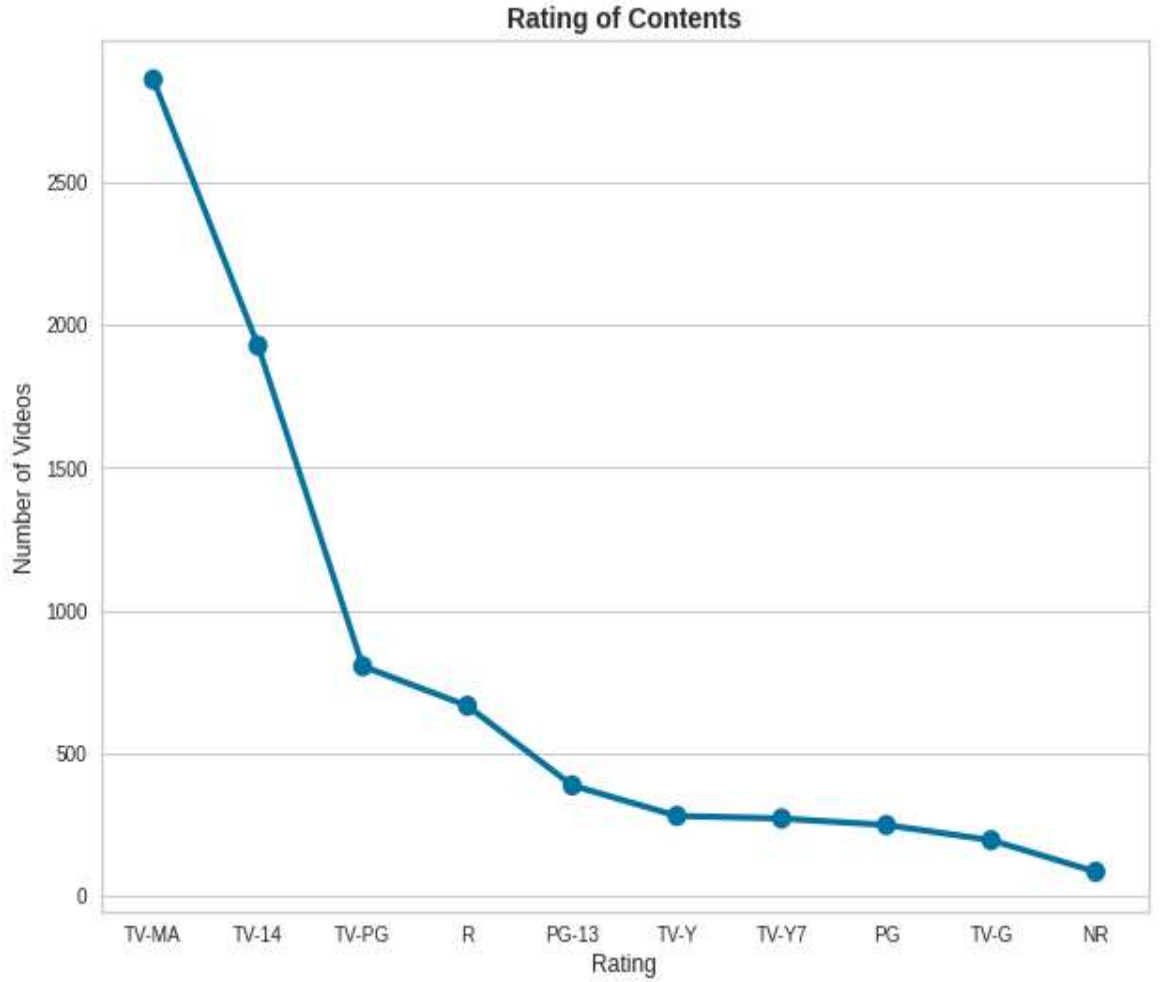


- *Drama, International Movies*, and *Comedies* seem popular choices in most countries.
- *British* and *International Tv Shows* dominate in the United Kingdom.
- Regional specialties such as *Anime* in Japan, *Spanish TV Shows* in Spain and *Korean Tv Shows* in South Korea are more prominent in these countries.
- It is also observed that in the countries where the regional language is not English, *International Tv Shows/Movies* are more in demand.

Ratings of content

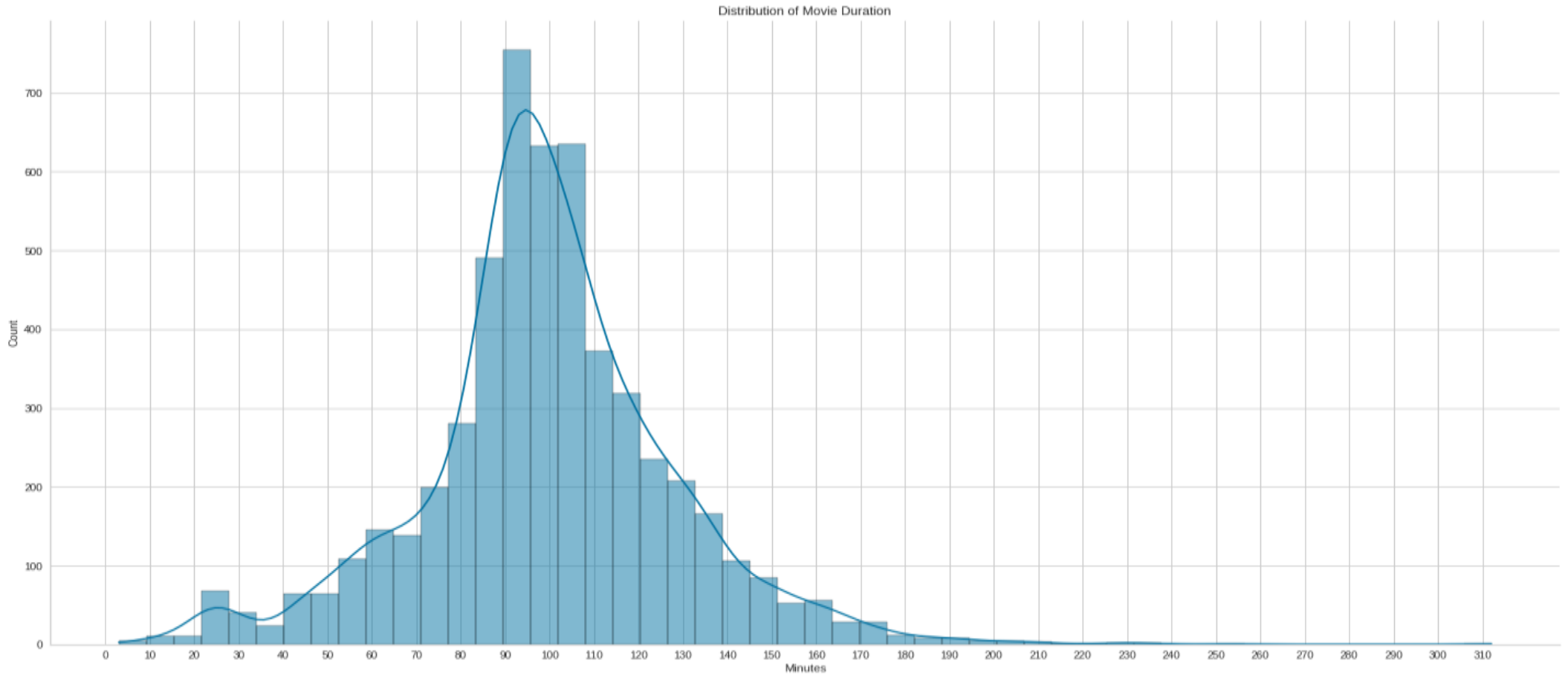
Netflix Rating of Movies/TV Shows based on content:-

- TV-MA : for Mature Audiences
- R : Restricted*
- PG-13 : Parents guidance required. May be Inappropriate for ages 12 and under
- TV-14 : Parents strongly cautioned. May not be suitable for ages 14 and under
- TV-PG : Parental Guidance suggested
- NR : Not Rated
- TV-G : Suitable for General Audiences
- TV-Y : Designed to be appropriate for all children
- PG : Parental Guidance suggested
- G : Suitable for General Audiences
- NC-17 : the content isn't suitable for children under 17 and younger
- TV-Y7-FV : Suitable for ages 7 and up



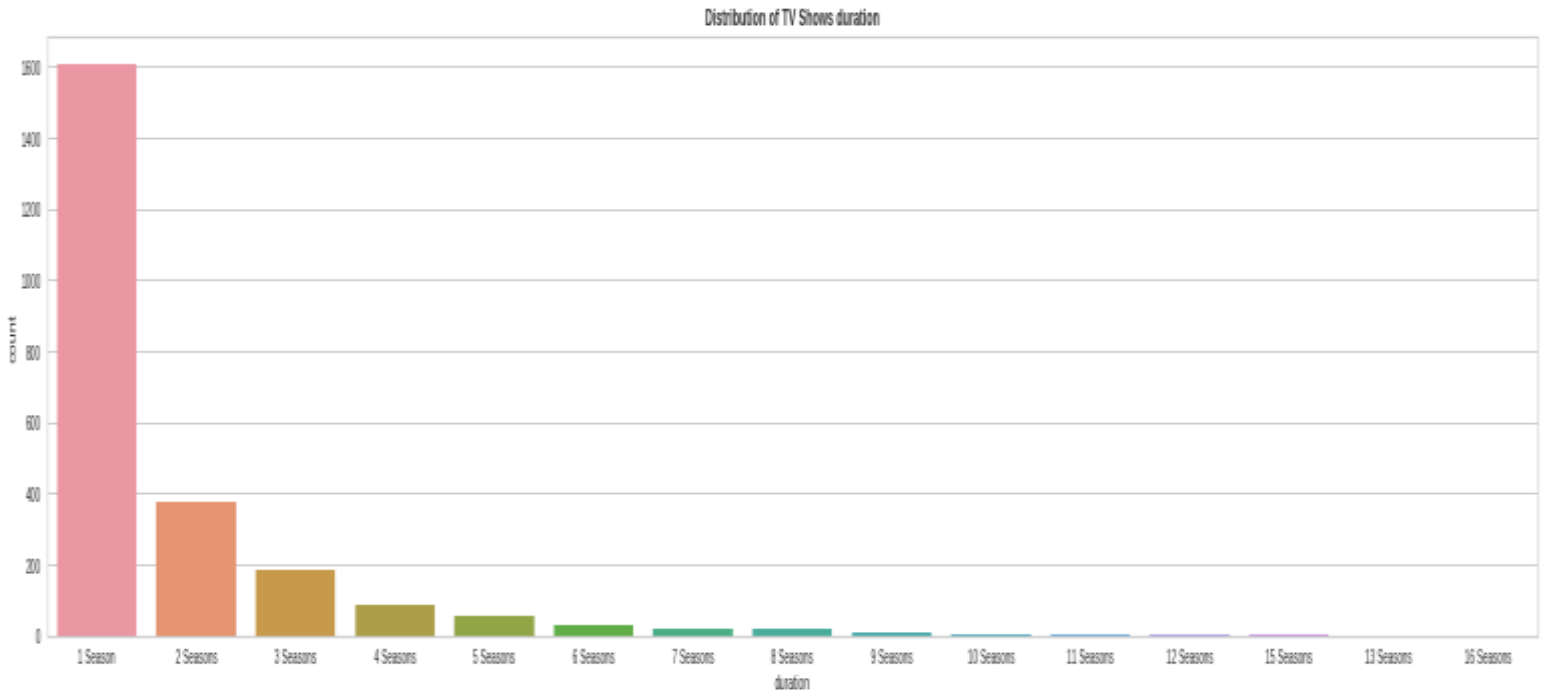
Most content on Netflix is rated for Mature Audiences(MA) and over 14 years old

Duration of Movies



Most movies on Netflix have a duration range from 80 to 120 minutes

Seasons of TV Shows



More number of TV shows are having single season

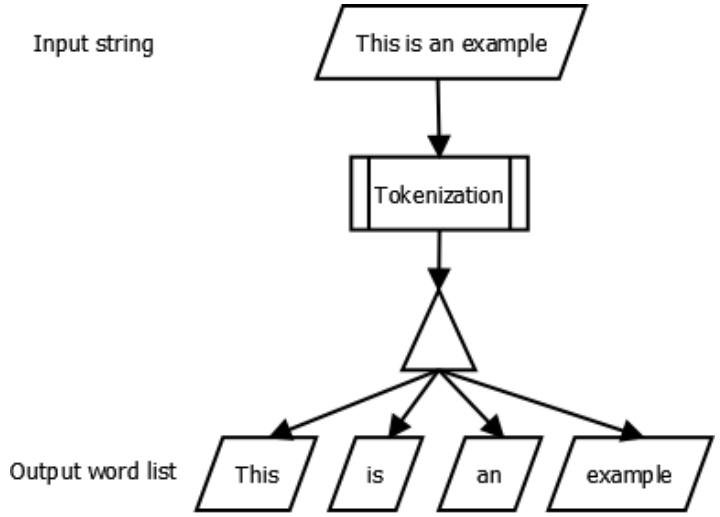
Longest TV Shows

	title	duration
2538	Grey's Anatomy	16
4438	NCIS	15
5912	Supernatural	15
1471	COMEDIANS of the world	13
1537	Criminal Minds	12
7169	Trailer Park Boys	12
1300	Cheers	11
2678	Heartland	11
1577	Dad's Army	10
1597	Danger Mouse: Classic Collection	10
3592	LEGO Ninjago: Masters of Spinjitzu	10
5538	Shameless (U.S.)	10
5795	Stargate SG-1	10

Grey's Anatomy is the longest TV Show with 16 Seasons

Feature Engineering

- **Tokenization-** Tokenization is method used for converting a large amount of textual data into parts to perform an analysis of the character of the text.



- **Lemmatization-** Lemmatization is a text normalization technique used in Natural Language Processing (NLP), that switches any kind of a word to its base root mode.

- **Stop Words Removal-** Stop words removal is the data pre-processing step in which we remove all the highly frequent words from the text that doesn't add any valuable information to understand the text better resulting in the NLP model dealing with less number of features.

Sample text with Stop Words	Without Stop Words
GeeksforGeeks – A Computer Science Portal for Geeks	GeeksforGeeks , Computer Science, Portal ,Geeks
Can listening be exhausting?	Listening, Exhausting
I like reading, so I read	Like, Reading, read

- **TF-IDF(Term Frequency-Inverse Document Frequency)-** Term Frequency - Inverse Document Frequency is a 2 dimensional data matrix where each term denotes the relative frequency of a particular word in a particular document as compared to other documents.

TF-IDF

$$\text{tfidf}(t, d, D) = \text{tf}(t, d) \cdot \text{idf}(t, D)$$

term frequency

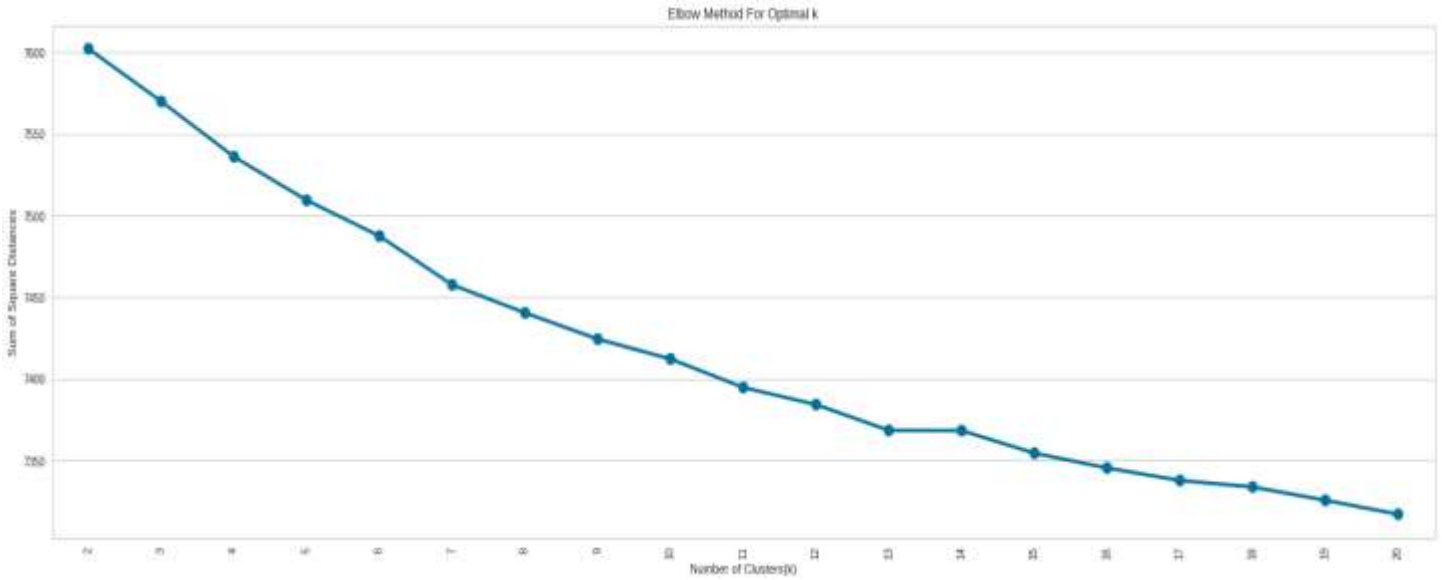
$$\text{count}(t, d) \div |d|$$

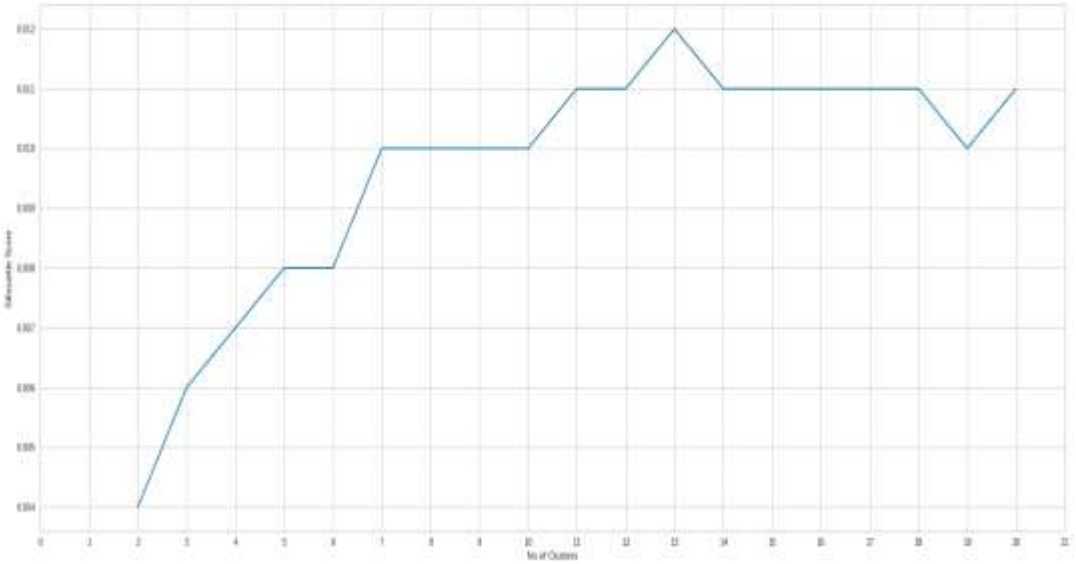
inverse document frequency

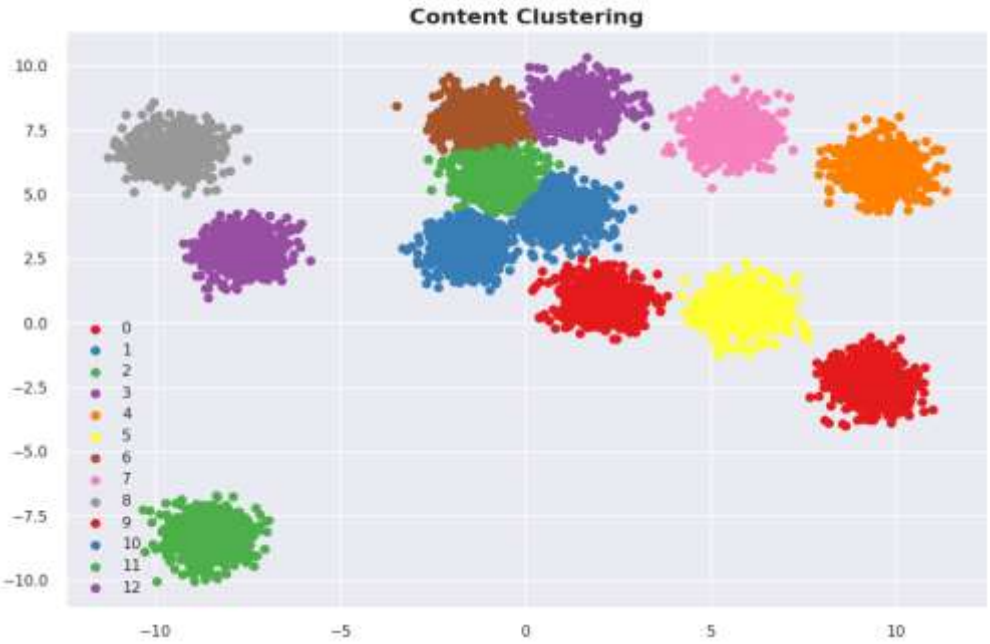
$$\log(|D| \div |\{d \in D : t \in d\}|)$$

K-Means Clustering

K-means Clustering is a centroid-based algorithm, where we calculate the distances to assign a point to a cluster. The main objective of the K-Means algorithm is to minimize the sum of distances between the points and their respective cluster centroid.



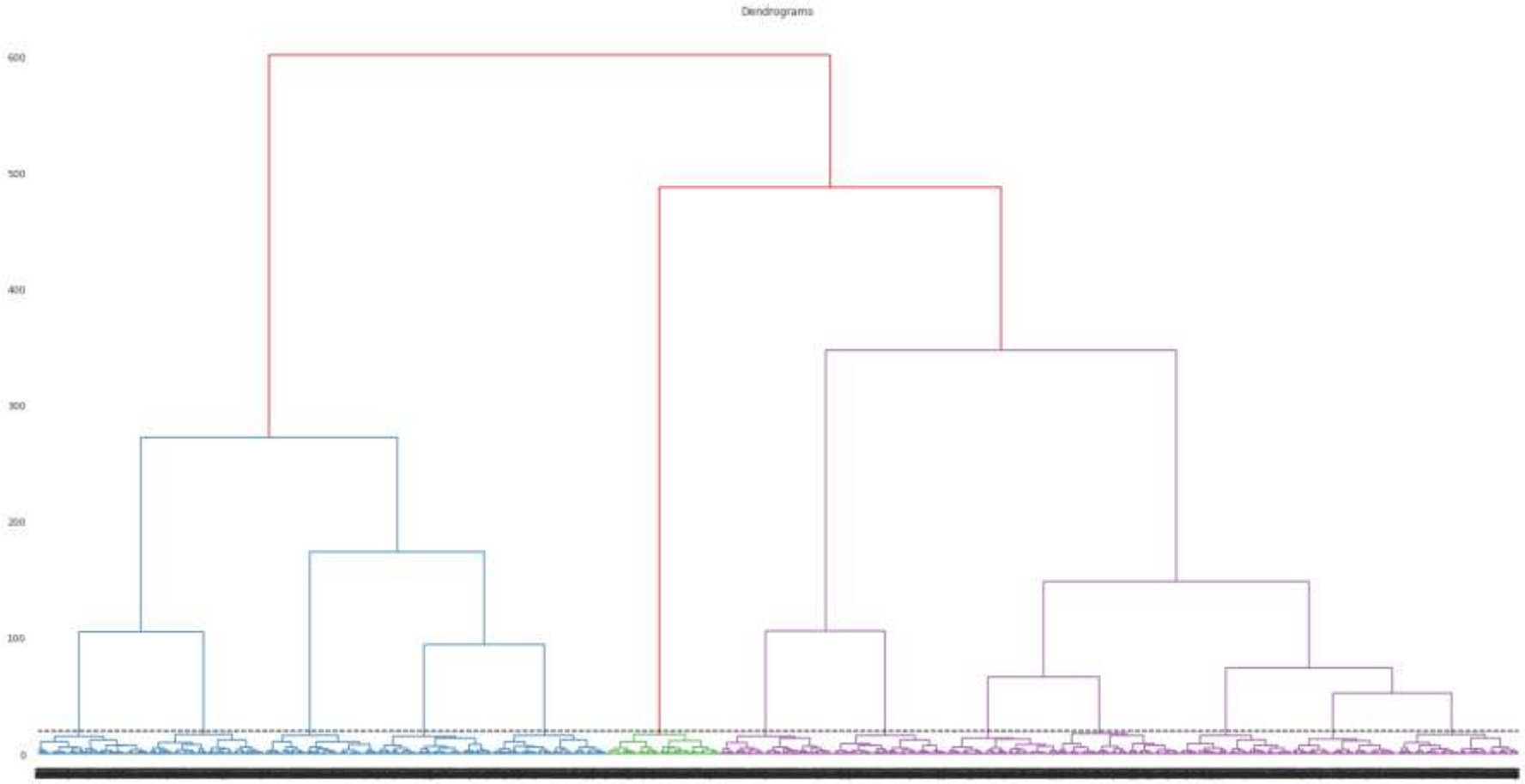




From Silhouette Analysis and Elbow method, the optimal cluster is 13. This gives a clustering score of 0.012.

Hierarchical Clustering

Hierarchical clustering is a method of creating groups so that objects within a group are similar to each other and different from objects in other groups. Clusters are visually represented in a hierarchical tree called a dendrogram.

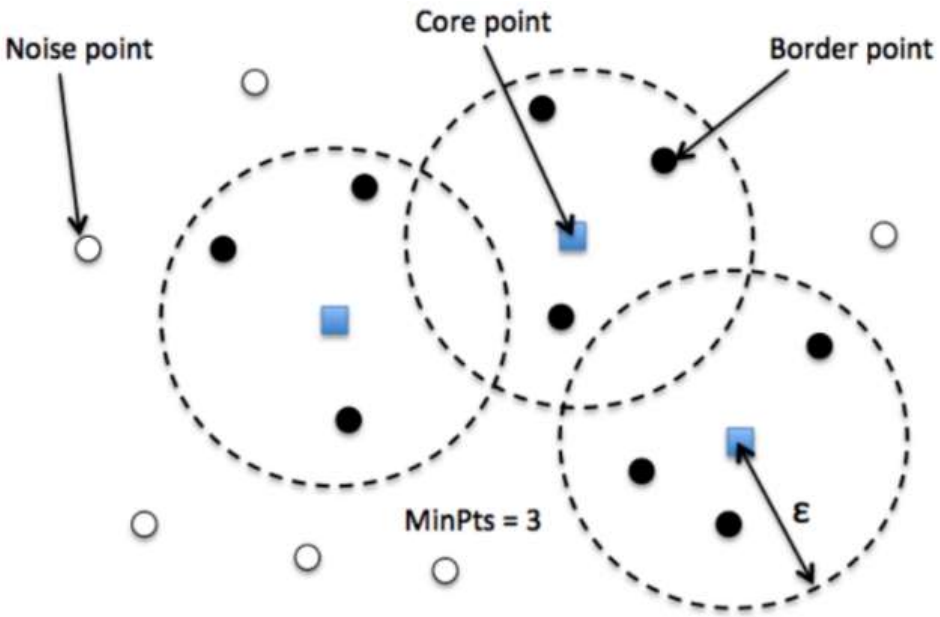


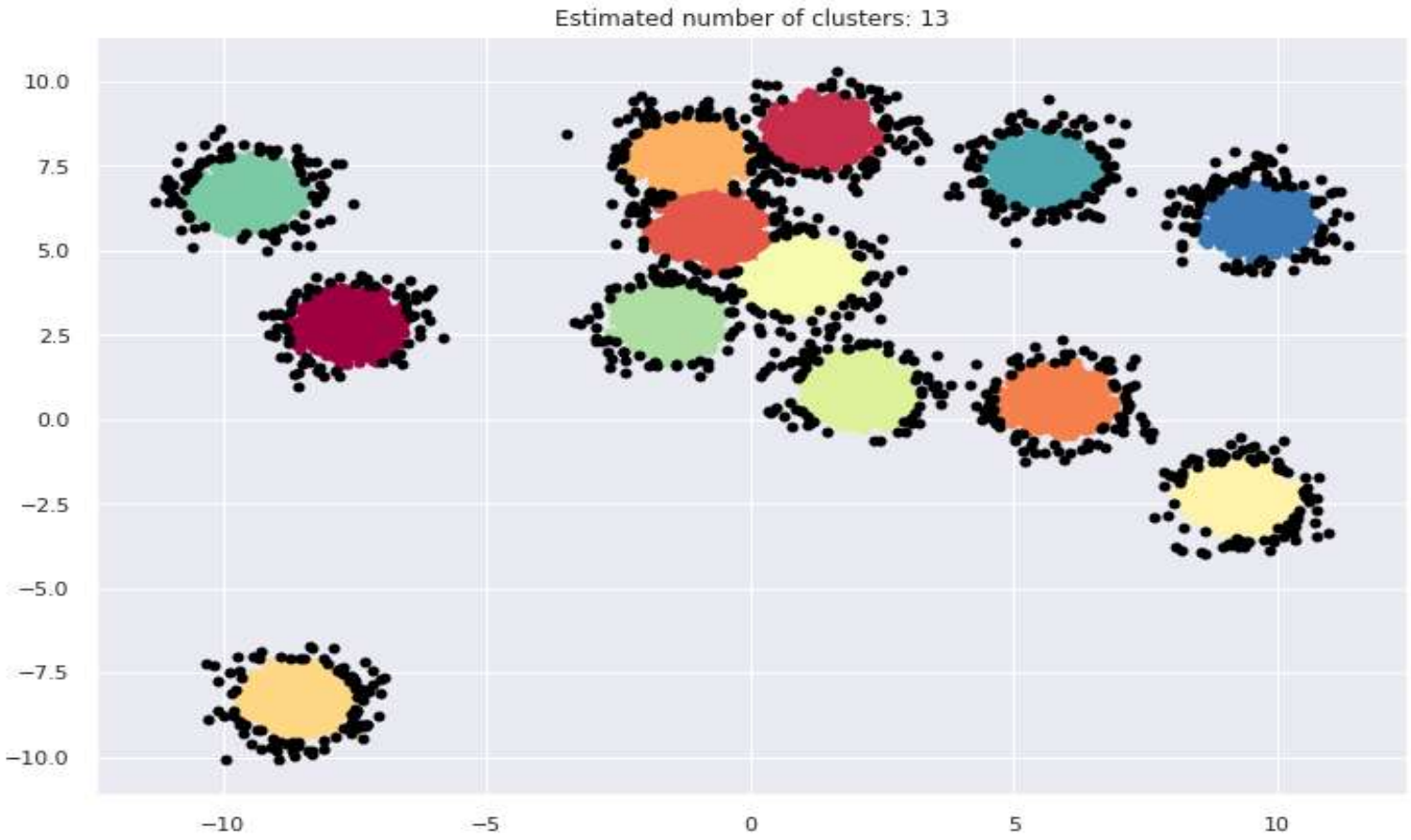
	Clusters	Silhouette Scores	Distance
3	13	0.671327	20
4	13	0.671327	25
5	13	0.671327	30
6	13	0.671327	35
7	13	0.671327	40
10	12	0.655458	55
11	12	0.655458	60
12	12	0.655458	65
13	11	0.643528	70
2	25	0.320521	15
0	95	0.286928	5
1	45	0.286070	10

Highest Silhouette Score of 0.671327 achieved at distance = 20 with 13 clusters

DB SCAN

DBSCAN stands for Density-Based Spatial Clustering of Applications with Noise. It is a density-based clustering algorithm. The algorithm increases regions with sufficiently high density into clusters and finds clusters of arbitrary architecture in spatial databases with noise.





Estimated number of clusters: 13
Estimated number of noise points: 1039

Conclusion

- Majority of content on Netflix are movies. Netflix has 5372 movies and 2398 TV shows.
- The number of movies on Netflix is growing significantly faster than the number of TV shows.
- We saw a huge increase in the number of movies and TV Shows after 2015. Highest number of movies released in 2017.
- Less Number of movies released after 2017 whereas more number of TV shows were released in this period.
- Most of these contents are released either in the year ending or in the beginning.
- International Movies/TV Shows are the top most genre in Netflix which is followed by Drama and Comedy movies/TV shows.
- United States is the major content producing country on the platform followed by India, UK, Japan, South Korea.
- Jan Suter and Raul Campos have directed the most content on Netflix.
- Also 6 of the actors among the top ten actors with maximum content are from India. Anupam Kher, Shah Rukh Khan, Om Puri are top 3 Actors.
- TV-MA tops the rating chart, indicating that mature content is more popular on Netflix.
- Most of the movies have duration between 80 to 120 minutes.
- Most number of TV shows are having single season. Grey's Anatomy is the longest TV Show with 16 Seasons.
- k=13 was found to be an optimal value for clusters using which we grouped our data into 13 distinct clusters.


THANK
YOU