Capstone Project-4

Project Title

NETFLIX MOVIES AND TV SHOWS CLUSTERING

By-Bhushan Patil


NETFLIX
LAW SCHOOL

Problem Statement

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. It will be interesting to explore what all other insights can be obtained from the same dataset. This dataset consists of TV shows and movies available on Netflix as of 2019. The dataset is collected from Flixable which is a third-party Netflix search engine.

Dataset

- There are 7787 entries and 12 columns.
- 11 columns present in object type and 1 column present in int type.
- There are a total of 3,631 null values across the entire dataset with 2,389 missing points under "director", 718 under "cast", 507 under "country", 10 under "date_added", and 7 under "rating". We will have to handle all null data points before we can dive into EDA and modeling.

```
RangeIndex: 7787 entries, 0 to 7786
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   show_id       7787 non-null   object
 1   type          7787 non-null   object
 2   title         7787 non-null   object
 3   director      5398 non-null   object
 4   cast          7069 non-null   object
 5   country       7280 non-null   object
 6   date_added    7777 non-null   object
 7   release_year  7787 non-null   int64 
 8   rating        7780 non-null   object
 9   duration      7787 non-null   object
 10  listed_in     7787 non-null   object
 11  description   7787 non-null   object
dtypes: int64(1), object(11)
```

Data Profiling & Cleaning

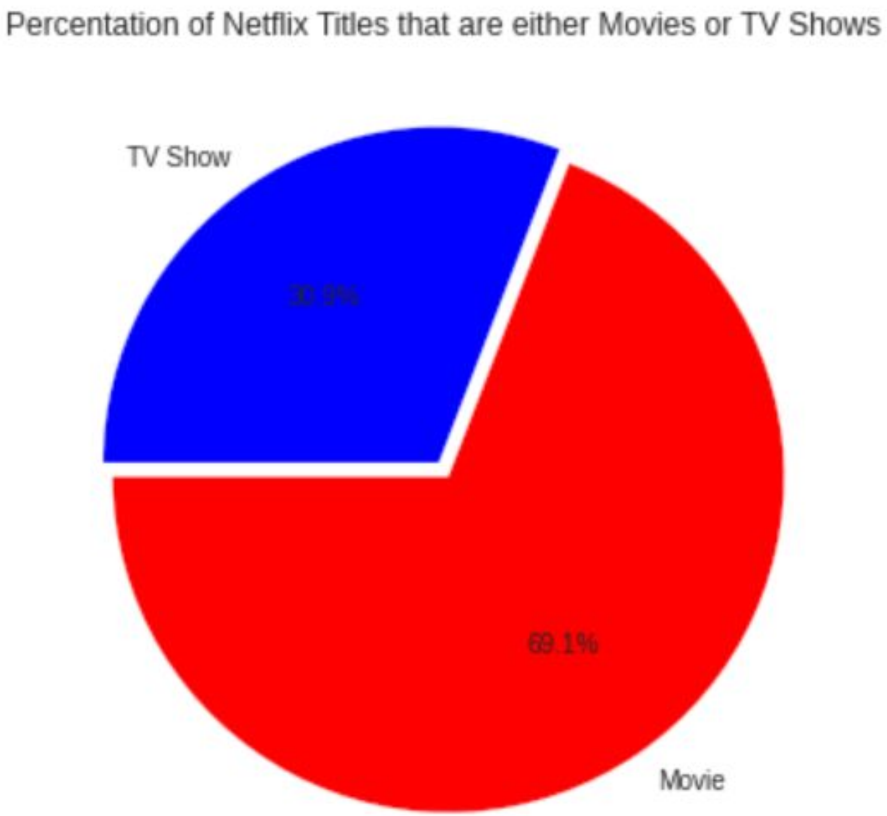
- There are a total of 3,631 null values across the entire dataset
- The easiest way to get rid of them would be to delete the rows with the missing data for missing values. This wouldn’t be beneficial to our EDA since it is a loss of information.
- Since “director” ,“cast”, and “country” contain the majority of null values, we chose to treat each missing value is unavailable.
- The other two label “date_added” and “rating” contain an insignificant portion of the data, so it drops from the dataset.
- Finally, we can see that there are no more missing values in the data frame.

```
show_id         0
type            0
title           0
director        0
cast            0
country         0
date_added      0
release_year    0
rating          0
duration        0
listed_in       0
description     0
dtype: int64
```

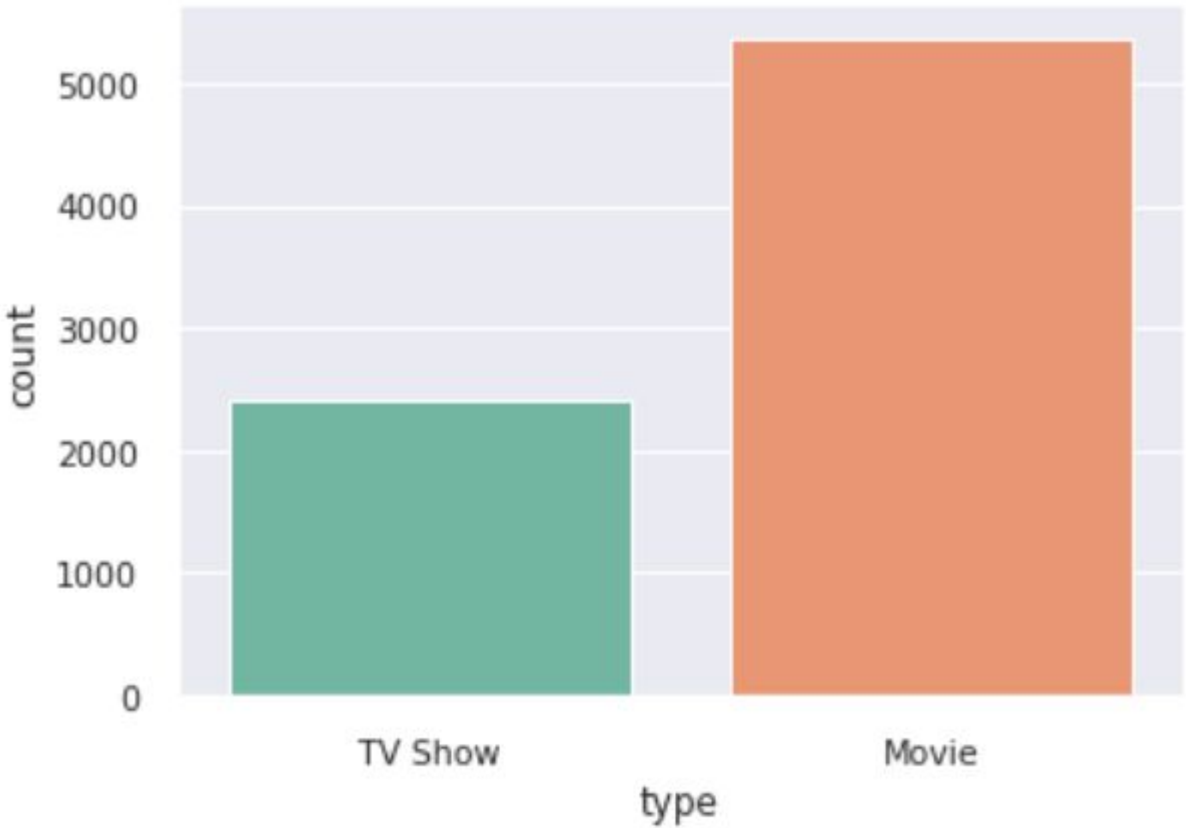
Exploratory Data Analysis with Python

Netflix Content Analysis

Movie present in higher percentage than TV shows



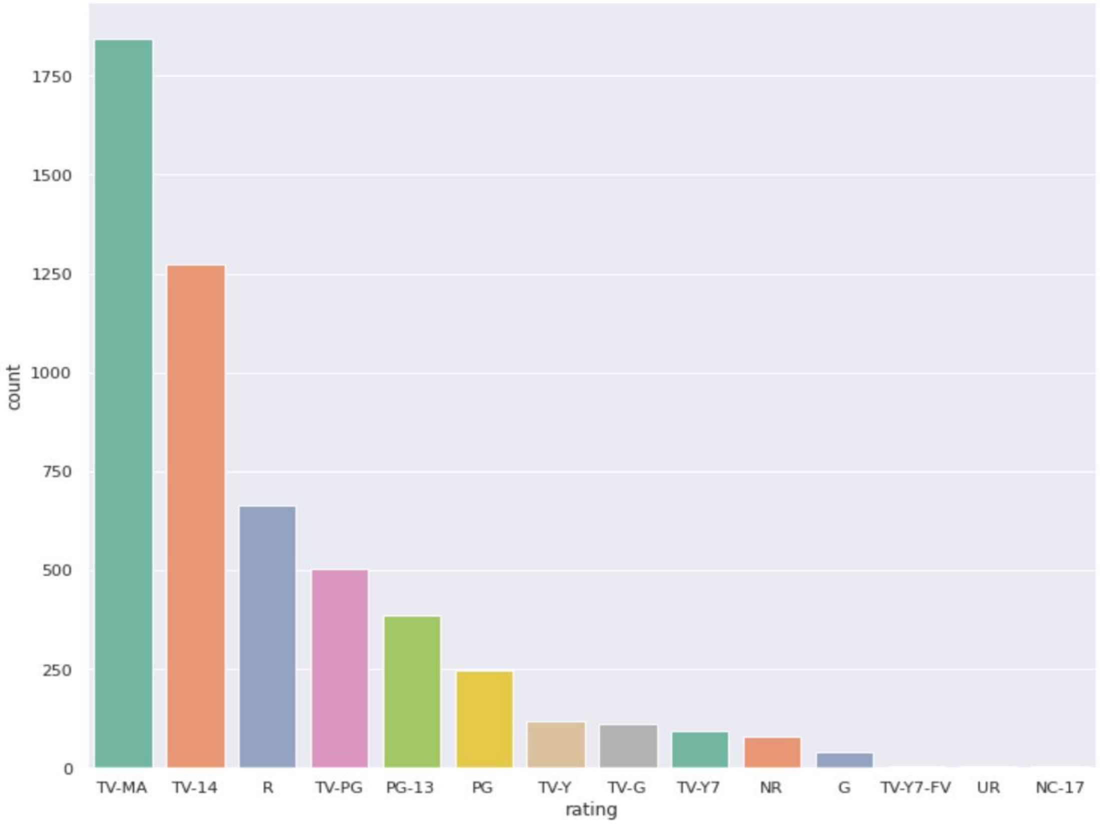
From this bar graph, it is evident that there are more Movies on Netflix than TV shows.



Netflix Rating Analysis

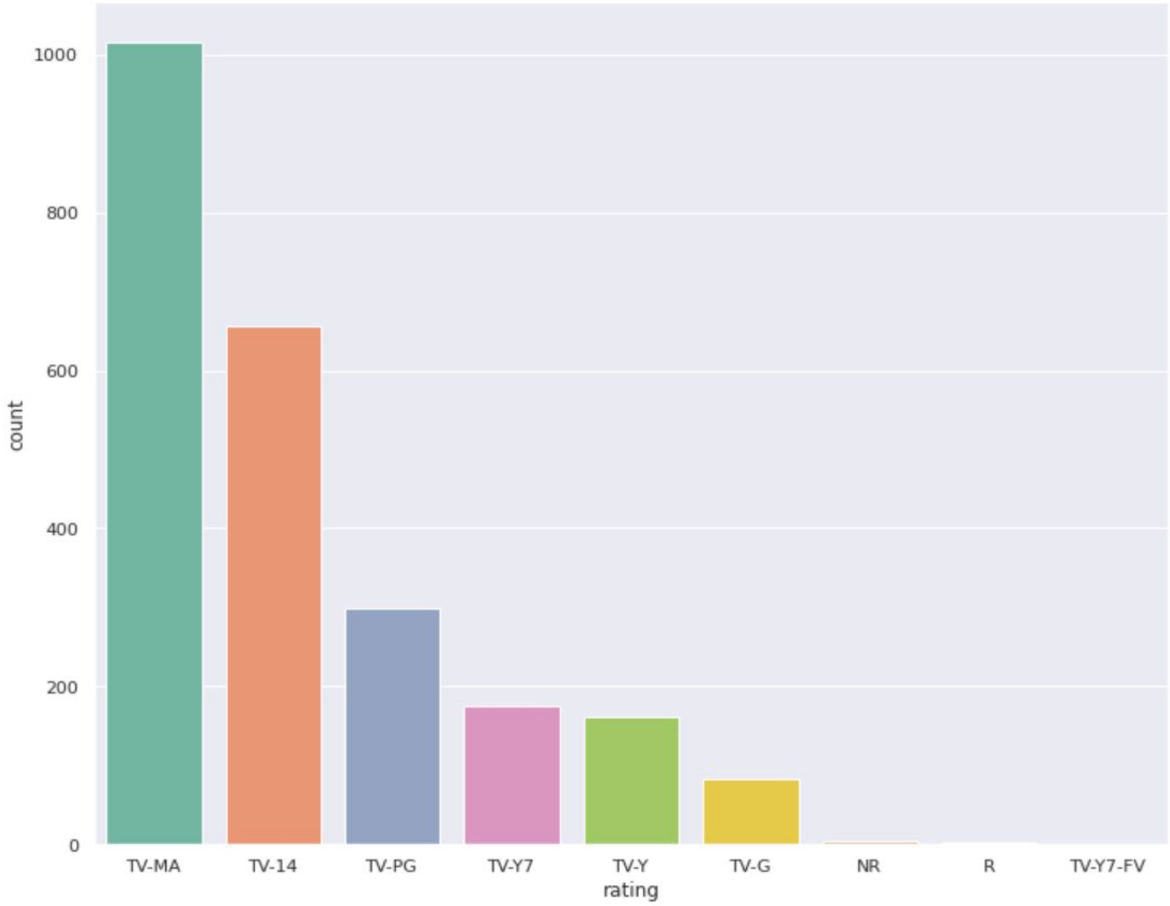
Movies

	rating	title
0	G	39
1	NC-17	3
2	NR	79
3	PG	247
4	PG-13	386
5	R	663
6	TV-14	1272
7	TV-G	111
8	TV-MA	1845
9	TV-PG	505
10	TV-Y	117
11	TV-Y7	95
12	TV-Y7-FV	5
13	UR	5

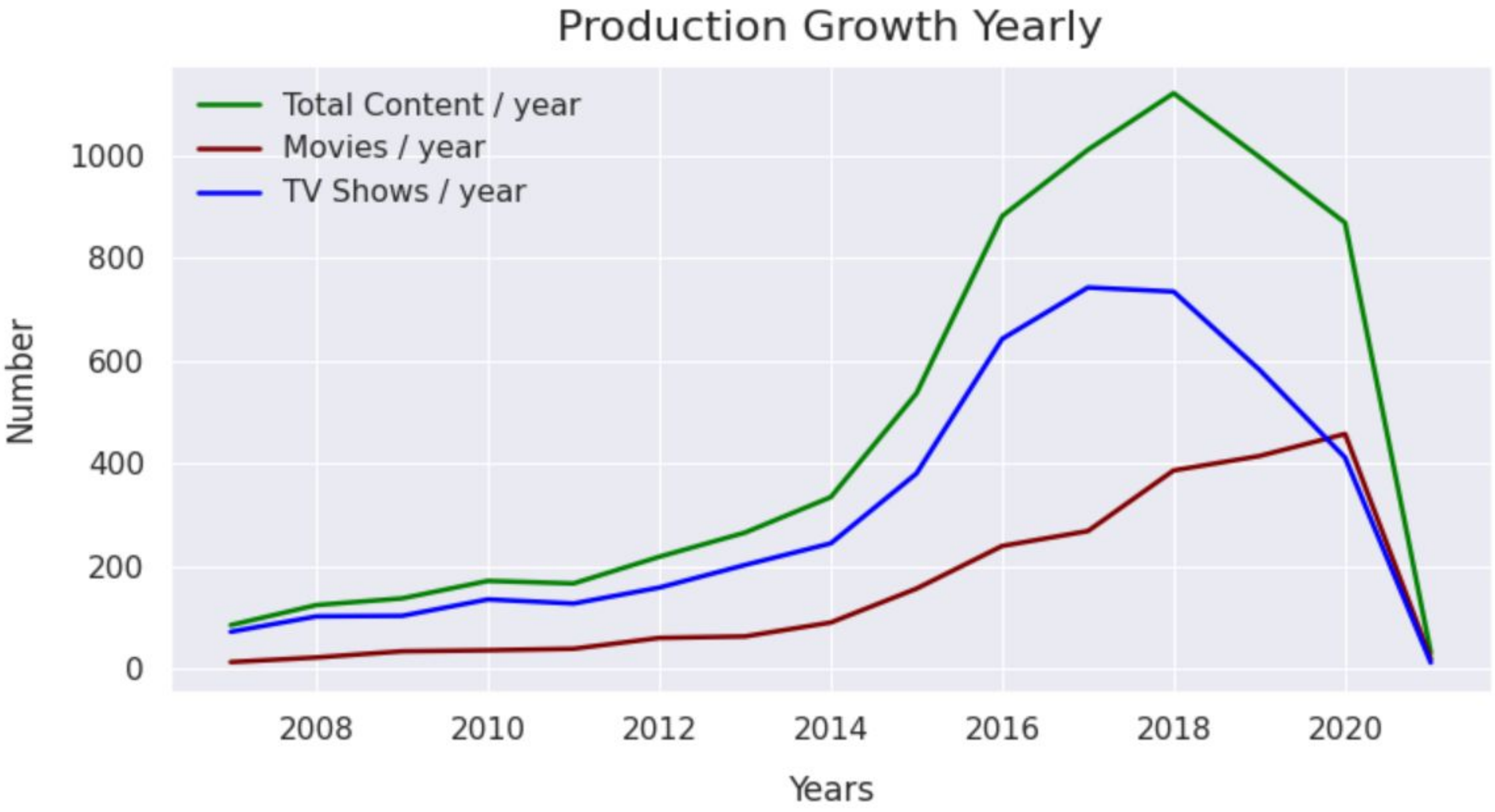


TV Shows

	rating	title
0	NR	4
1	R	2
2	TV-14	656
3	TV-G	83
4	TV-MA	1016
5	TV-PG	299
6	TV-Y	162
7	TV-Y7	175
8	TV-Y7-FV	1

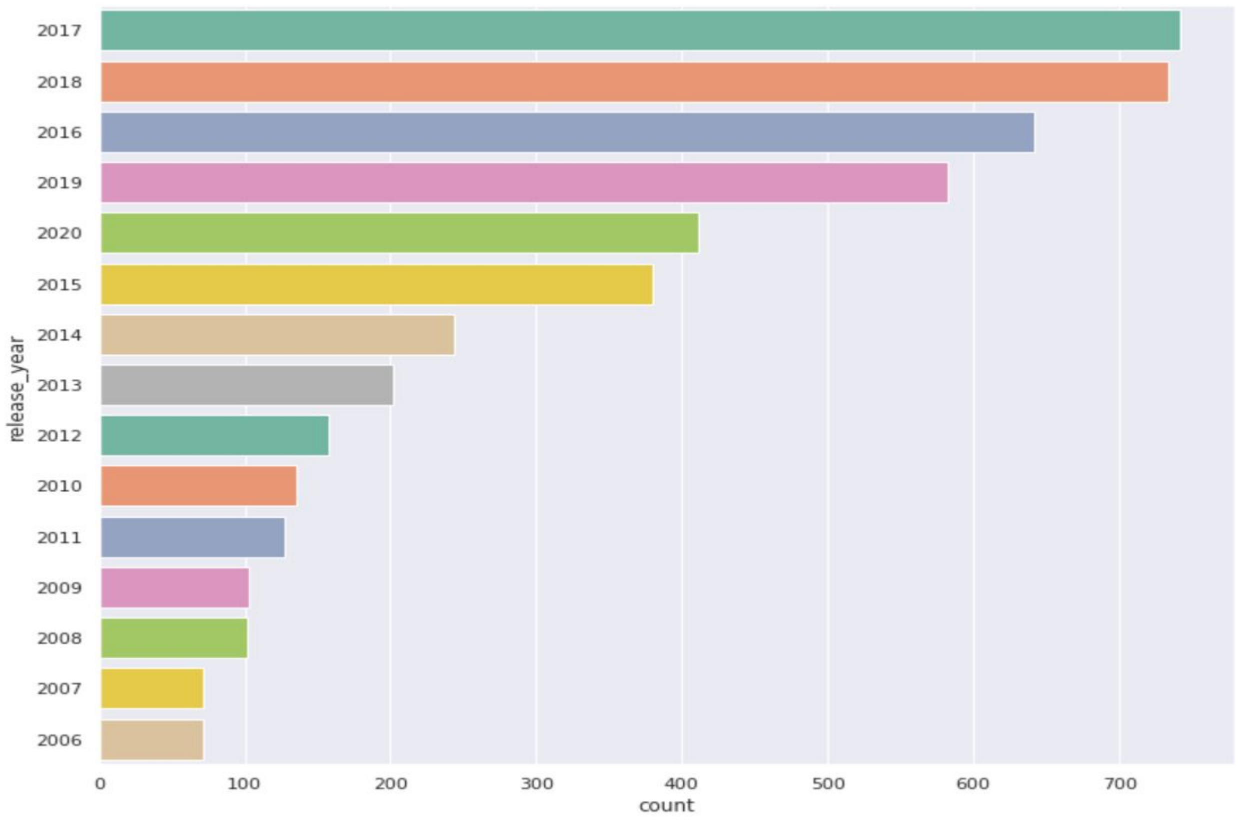


Content growth over years

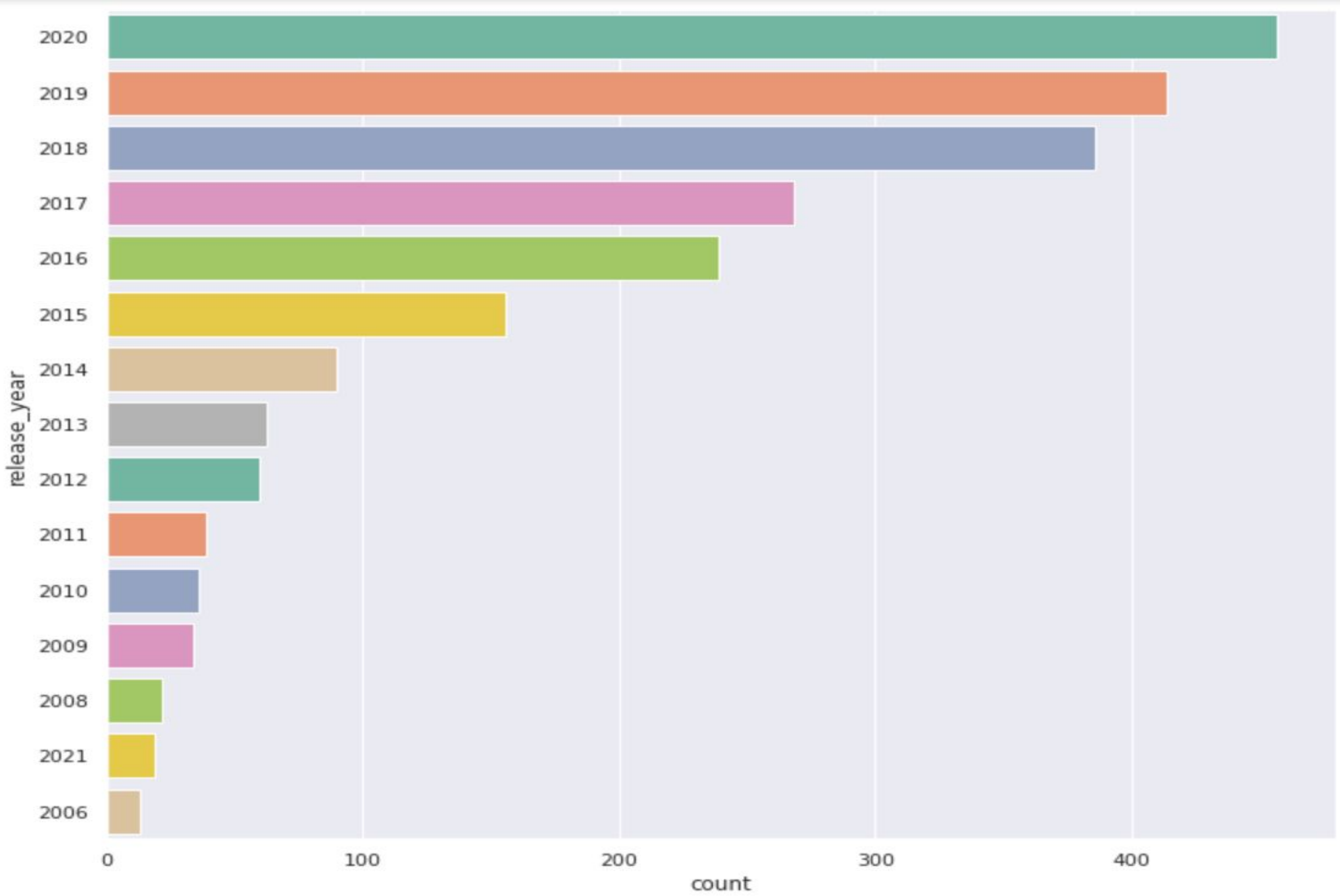


Year wise analysis

Movies

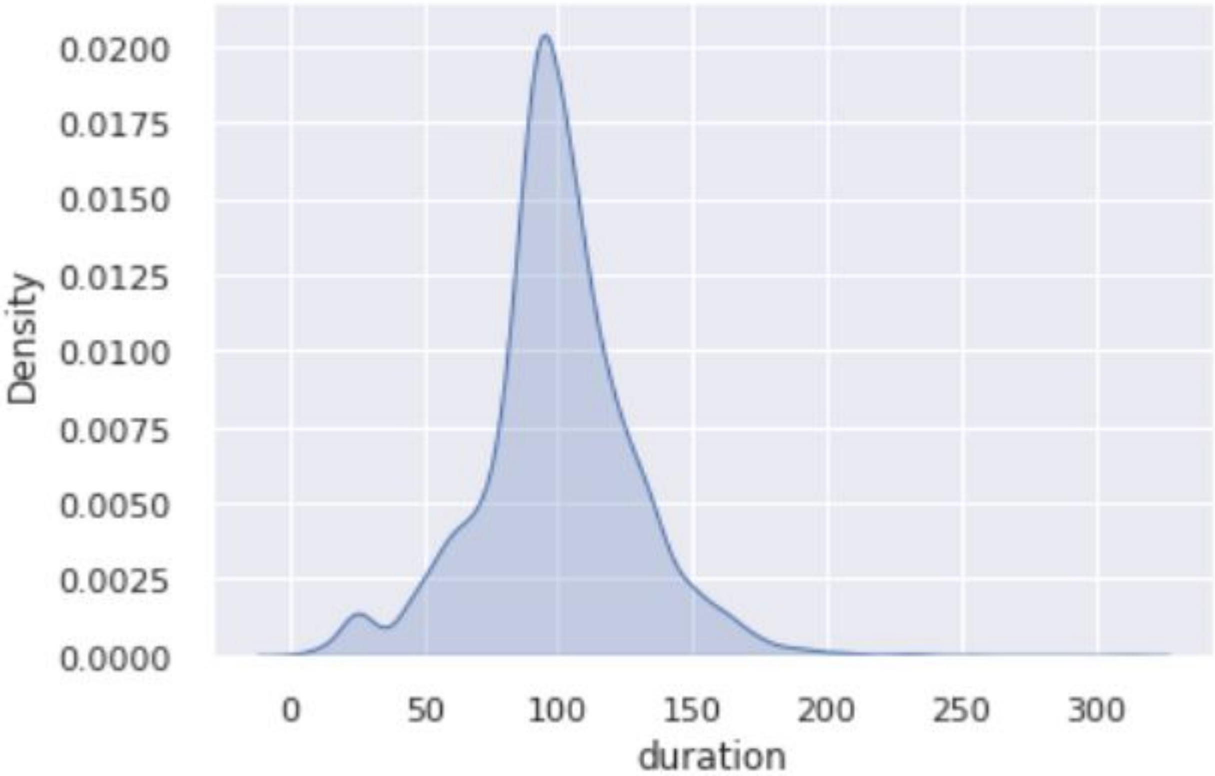


TV Shows

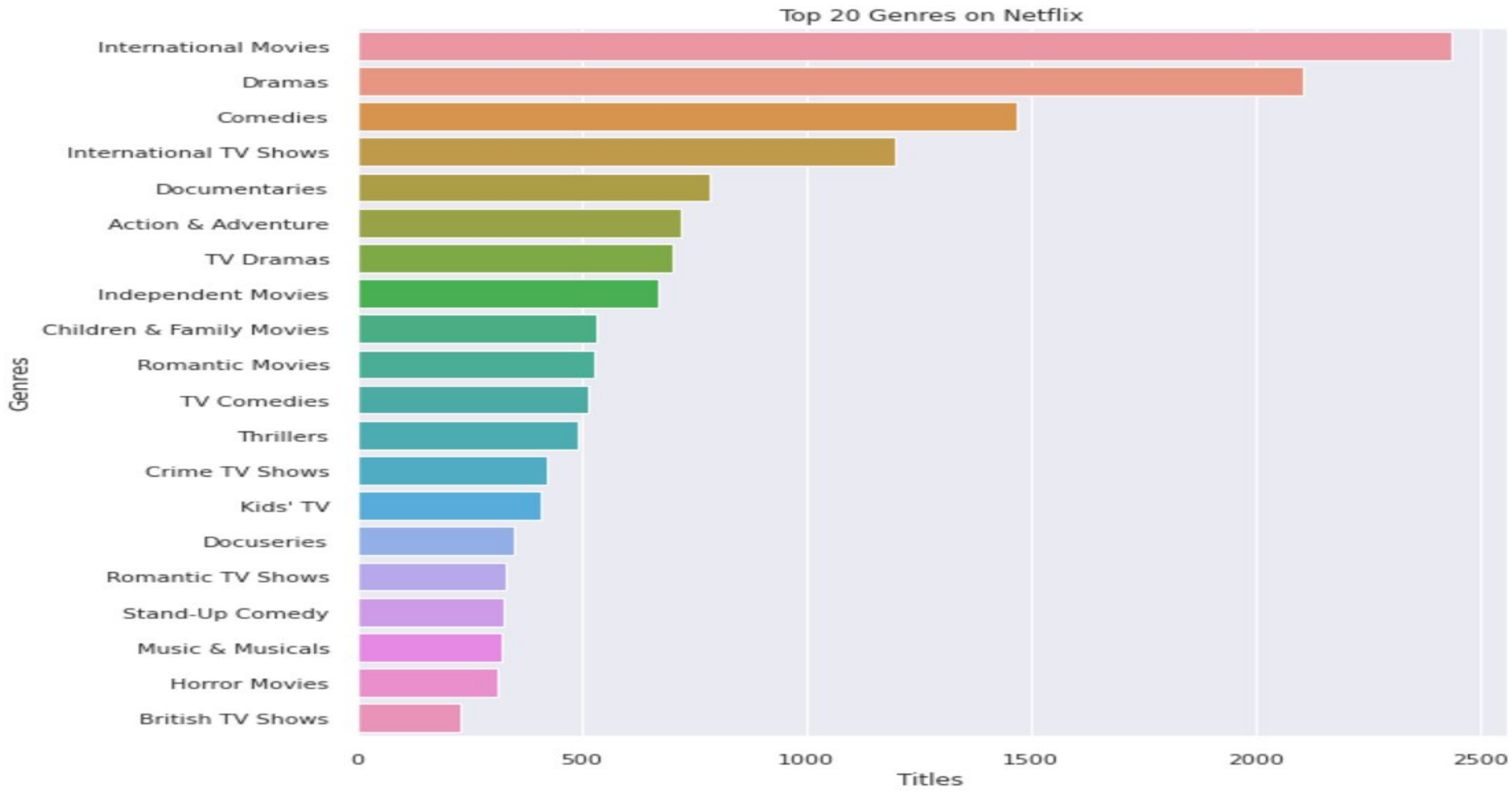

2020
2019
2018
2017
2016
2015
2014
2013
2012
2011
2010
2009
2008
2021
2006
release_year
0
100
200
300
400
count

Analysis of duration of movies

A good amount of movies on Netflix are among the duration of 75-120 mins.

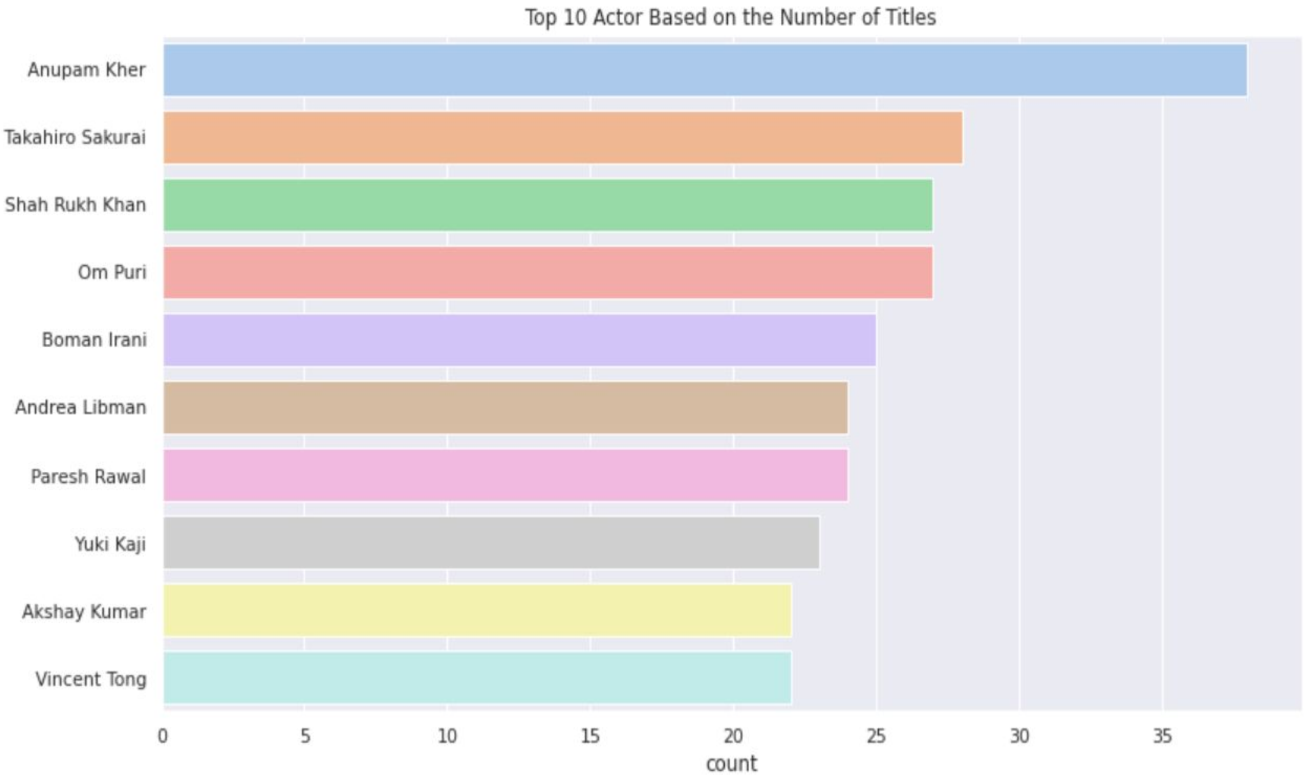


Top Genres on Netflix

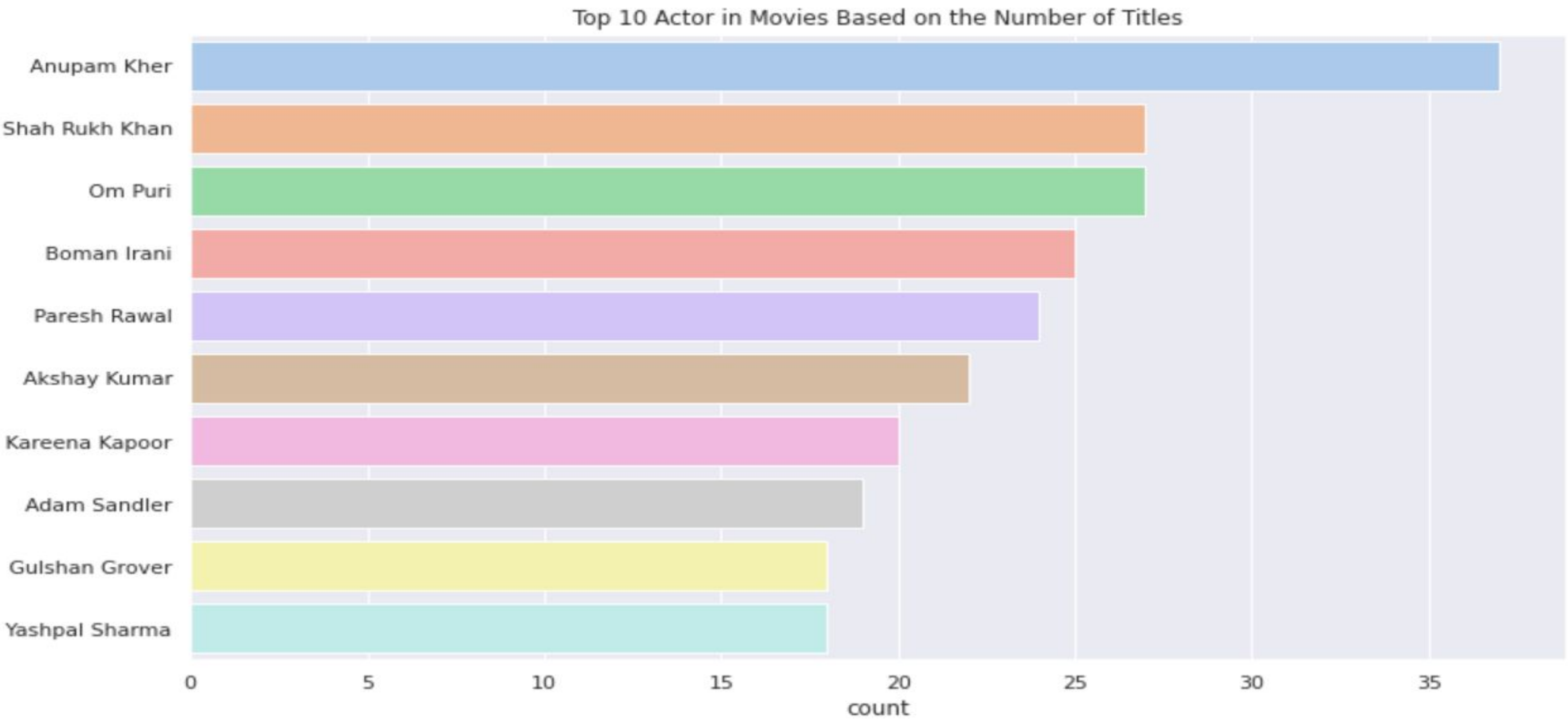


Top 10 Actor Based on the Number of Titles

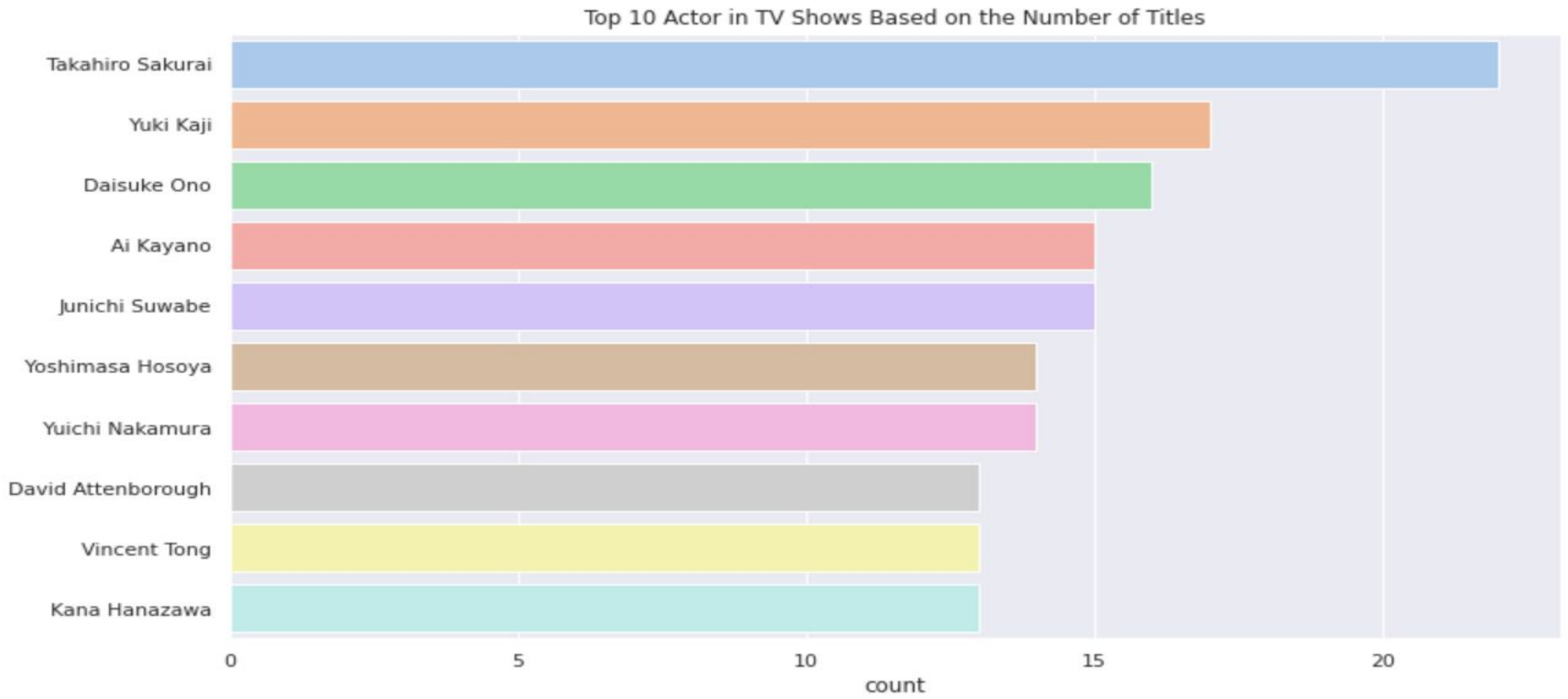
	Actors name	Count
1294	Anupam Kher	38
194	Takahiro Sakurai	28
2203	Om Puri	27
4039	Shah Rukh Khan	27
740	Boman Irani	25
2506	Paresh Rawal	24
8462	Andrea Libman	24
4368	Yuki Kaji	23
4994	Vincent Tong	22
2201	Akshay Kumar	22



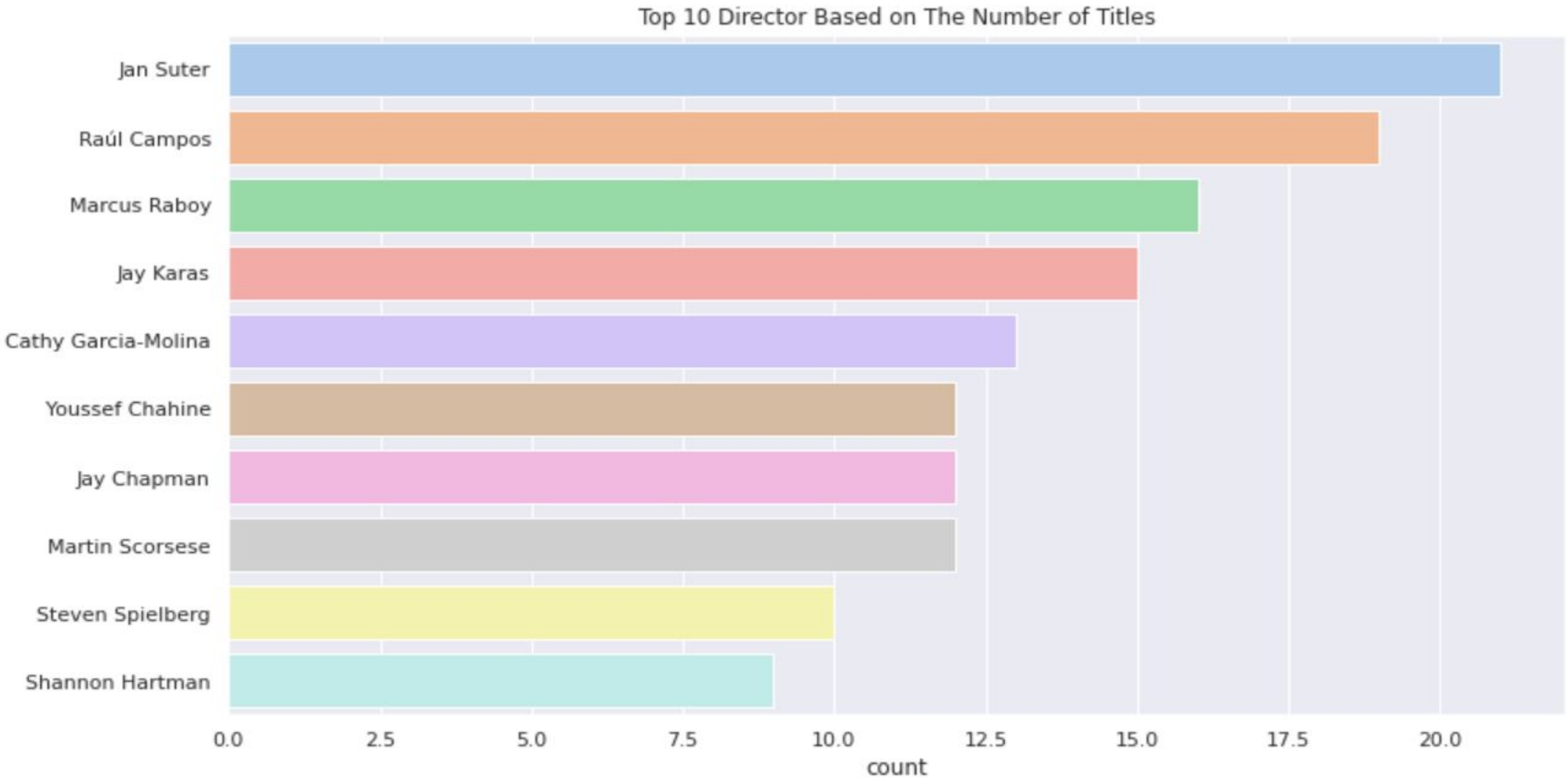
Top 10 Actor in Movies Based on the Number of Titles



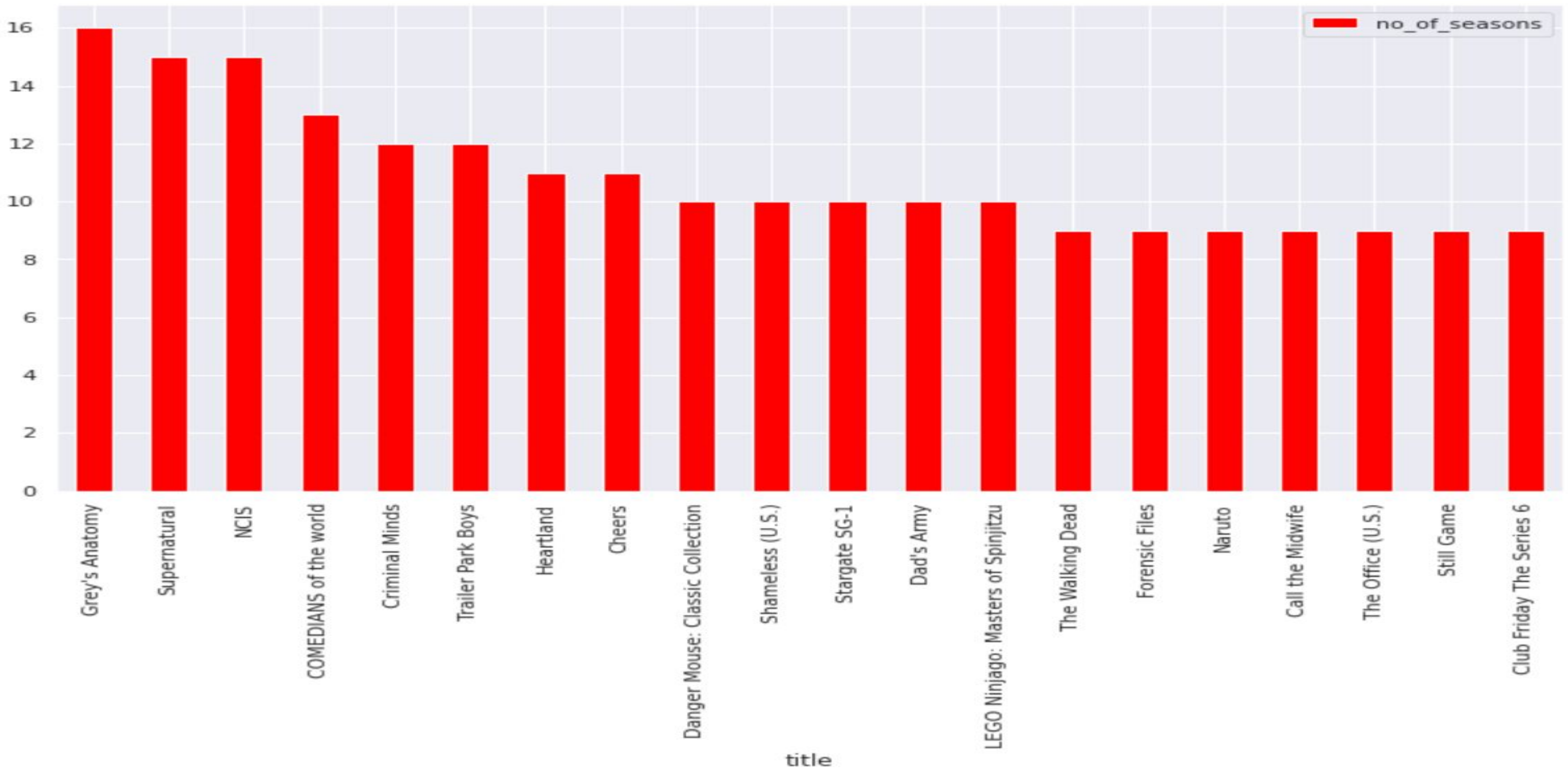
Top 10 Actor in TV Shows Based on the Number of Titles



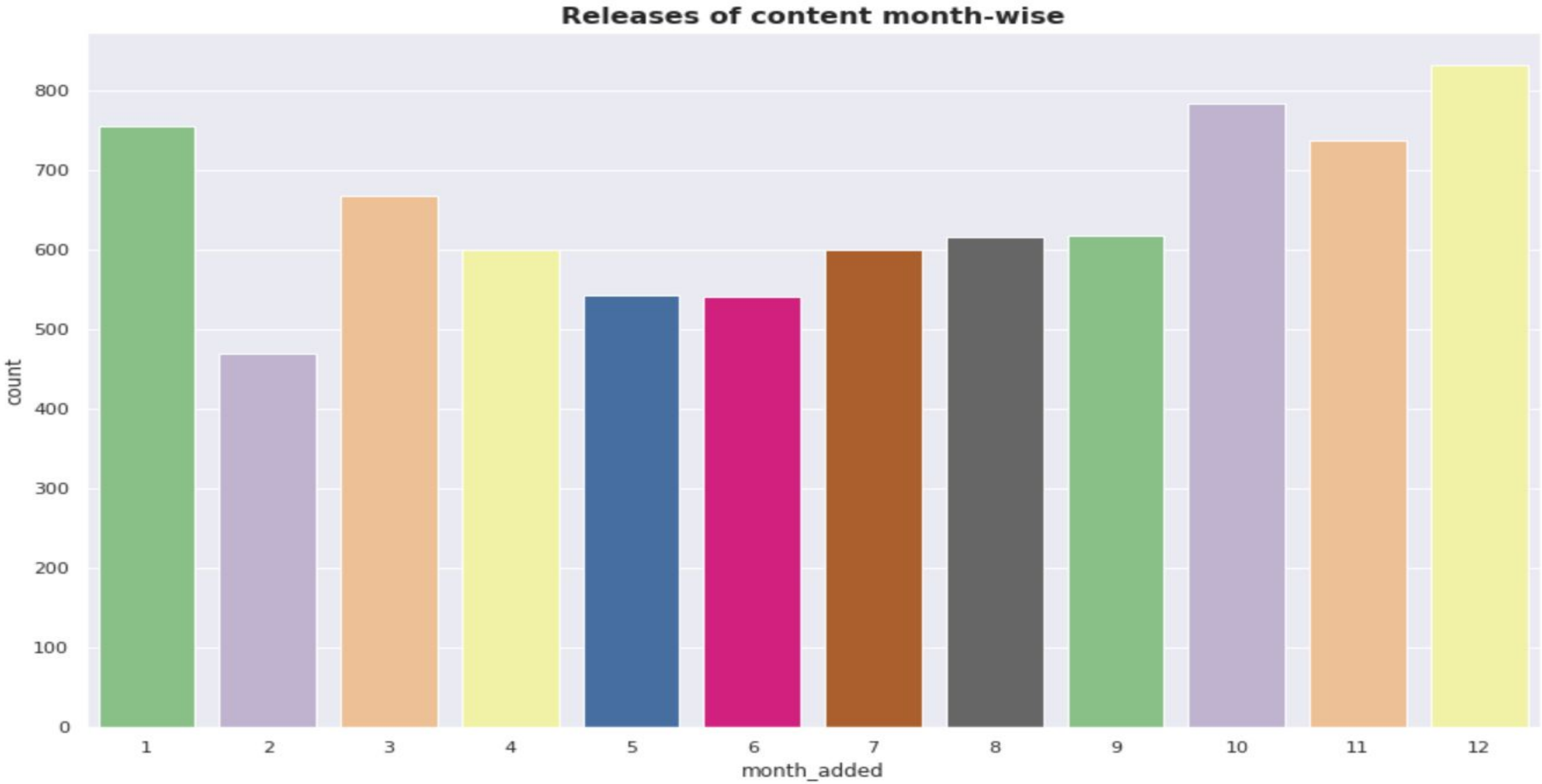
Top Directors on Netflix



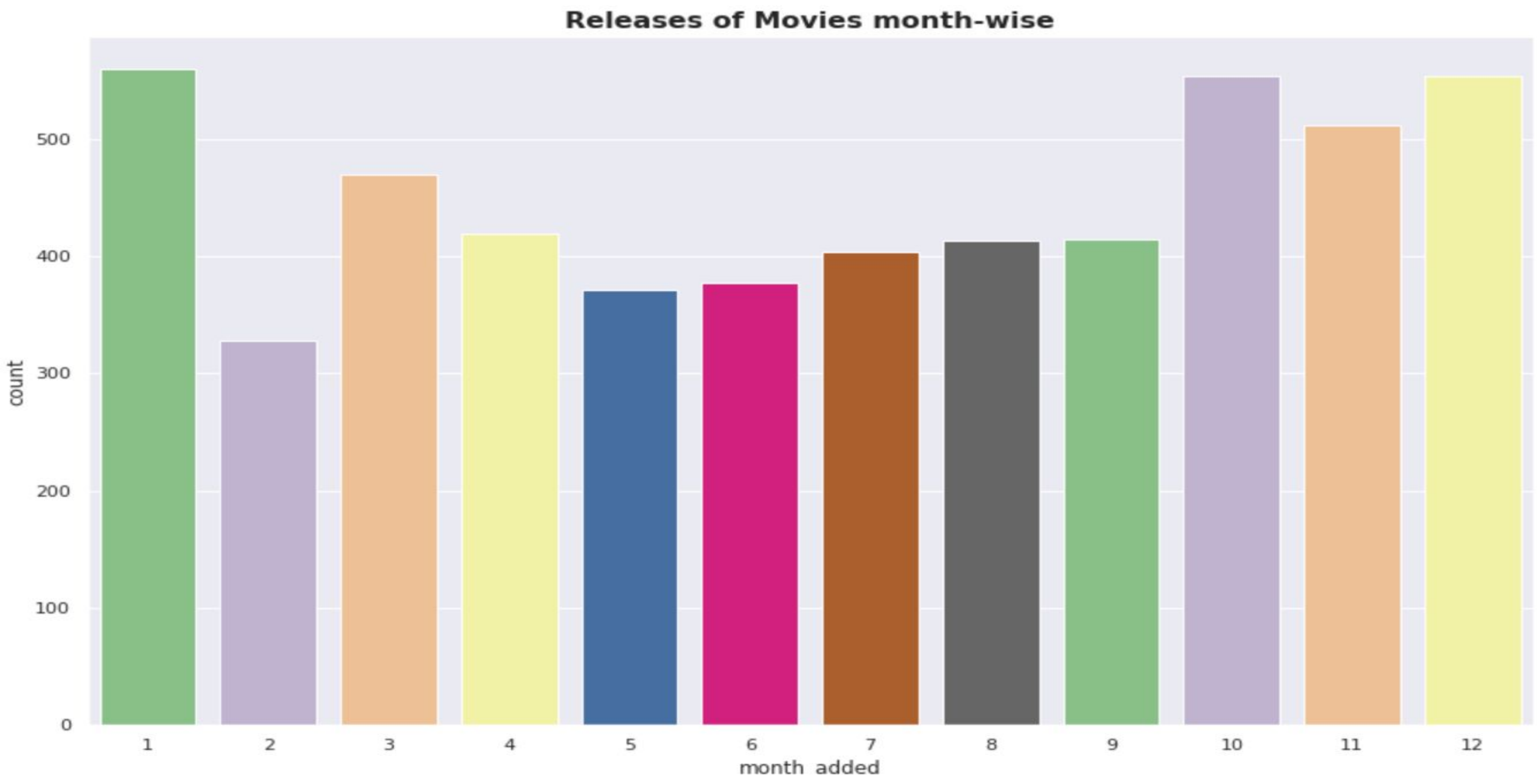
TV shows with largest number of seasons



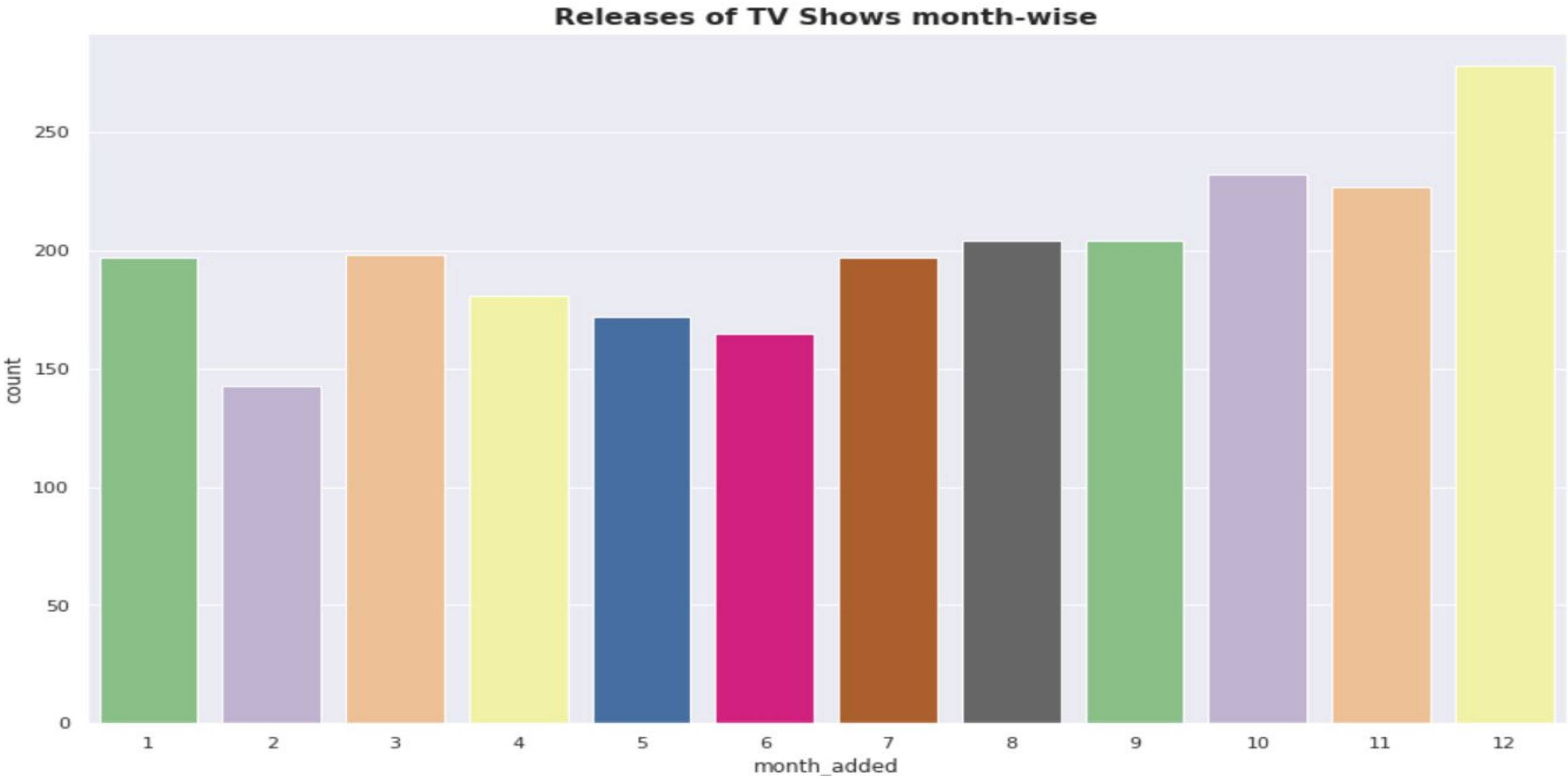
Release of content month-wise



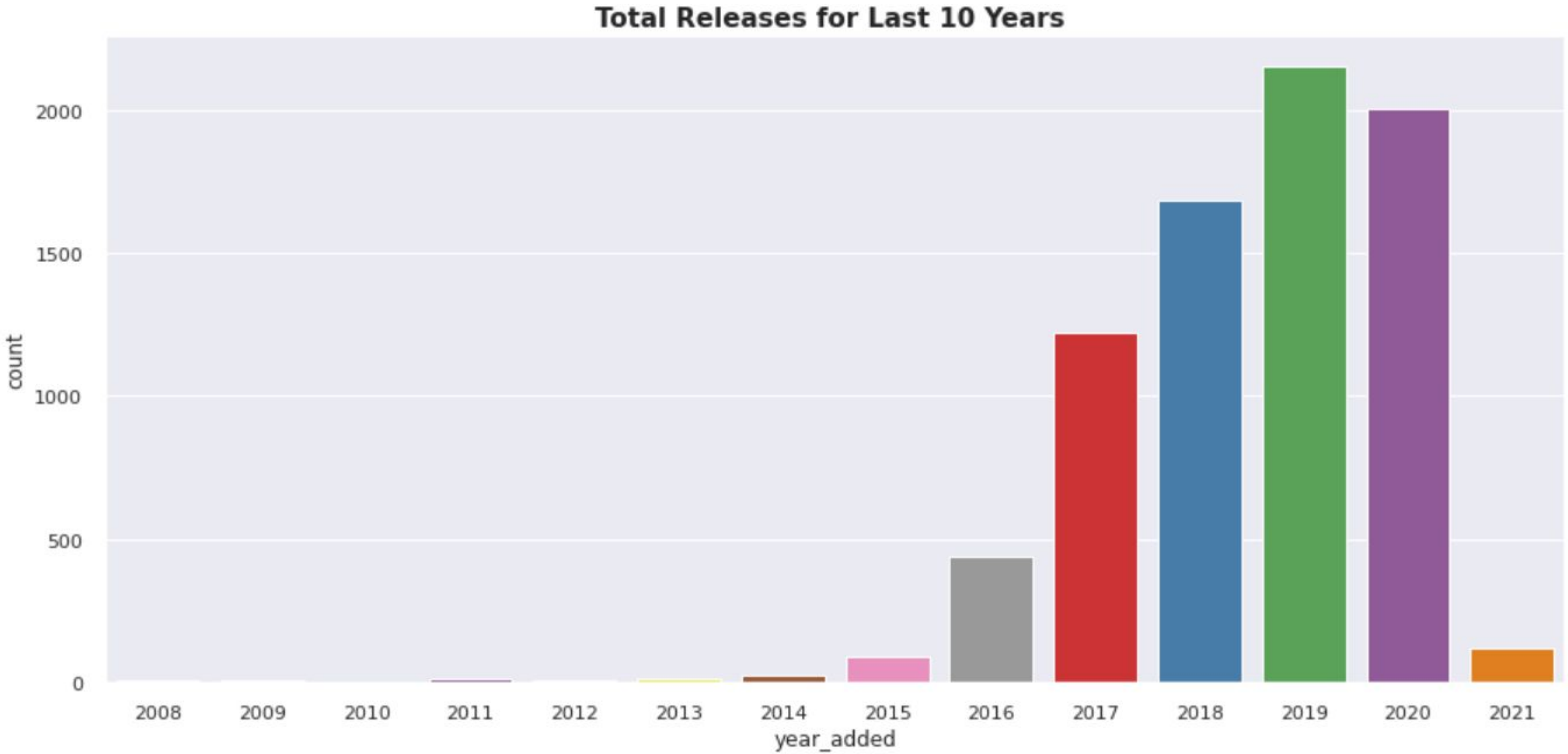
Release of movies month-wise



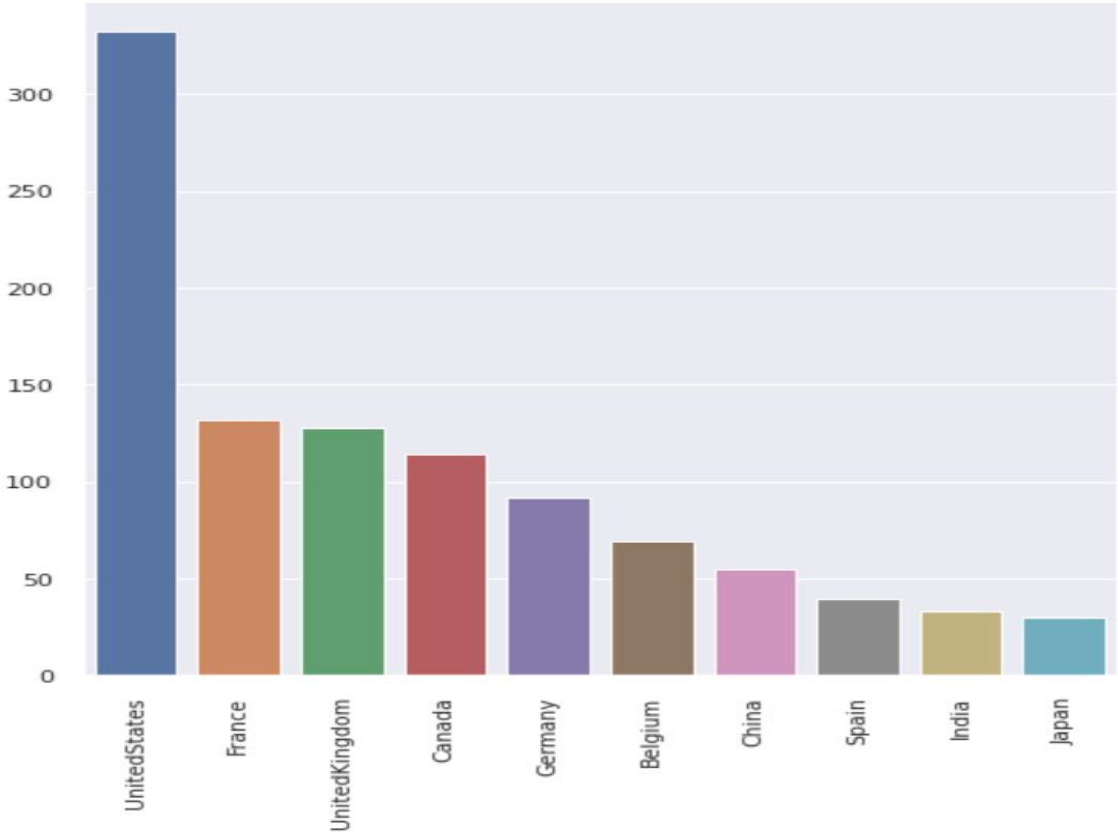
Releases of TV Shows month-wise



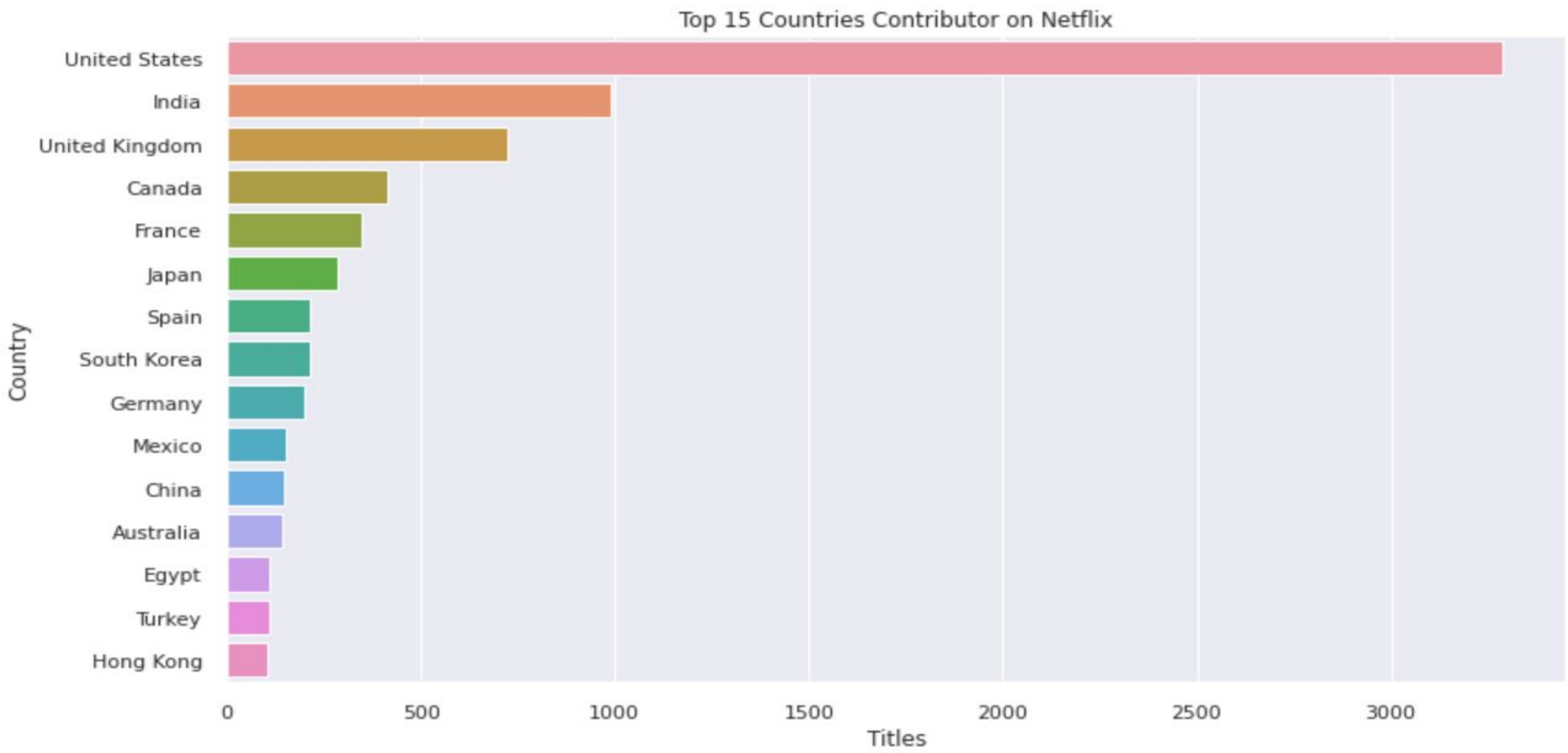
Total release for last 10 years

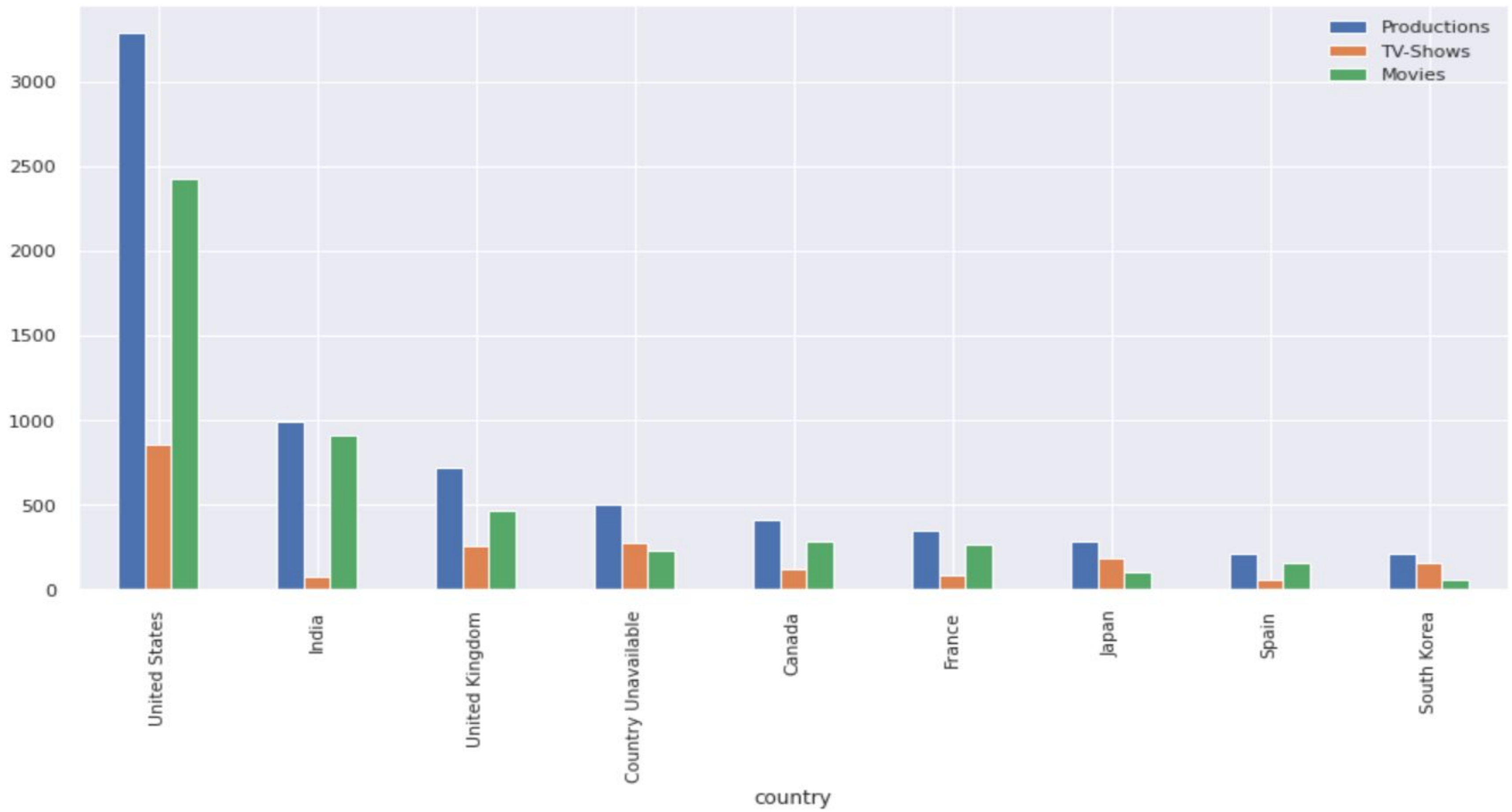


Top 10 Movie Content Creating Countries



Top 15 Countries Contributor on Netflix



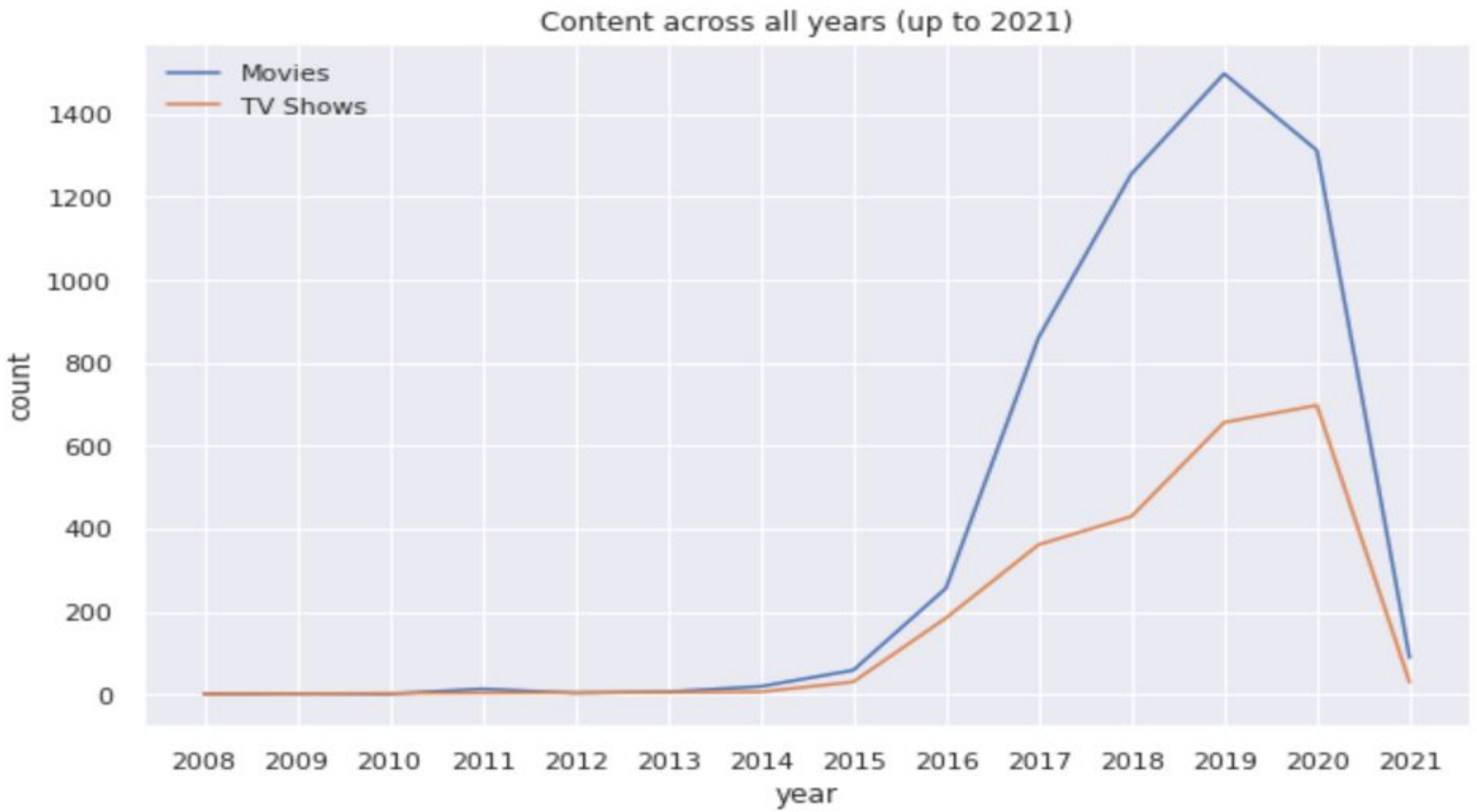


Productions
TV-Shows
Movies
3000
2500
2000
1500
1000
500
0
United States
India
United Kingdom
Country Unavailable
Canada
France
Japan
Spain
South Korea
country

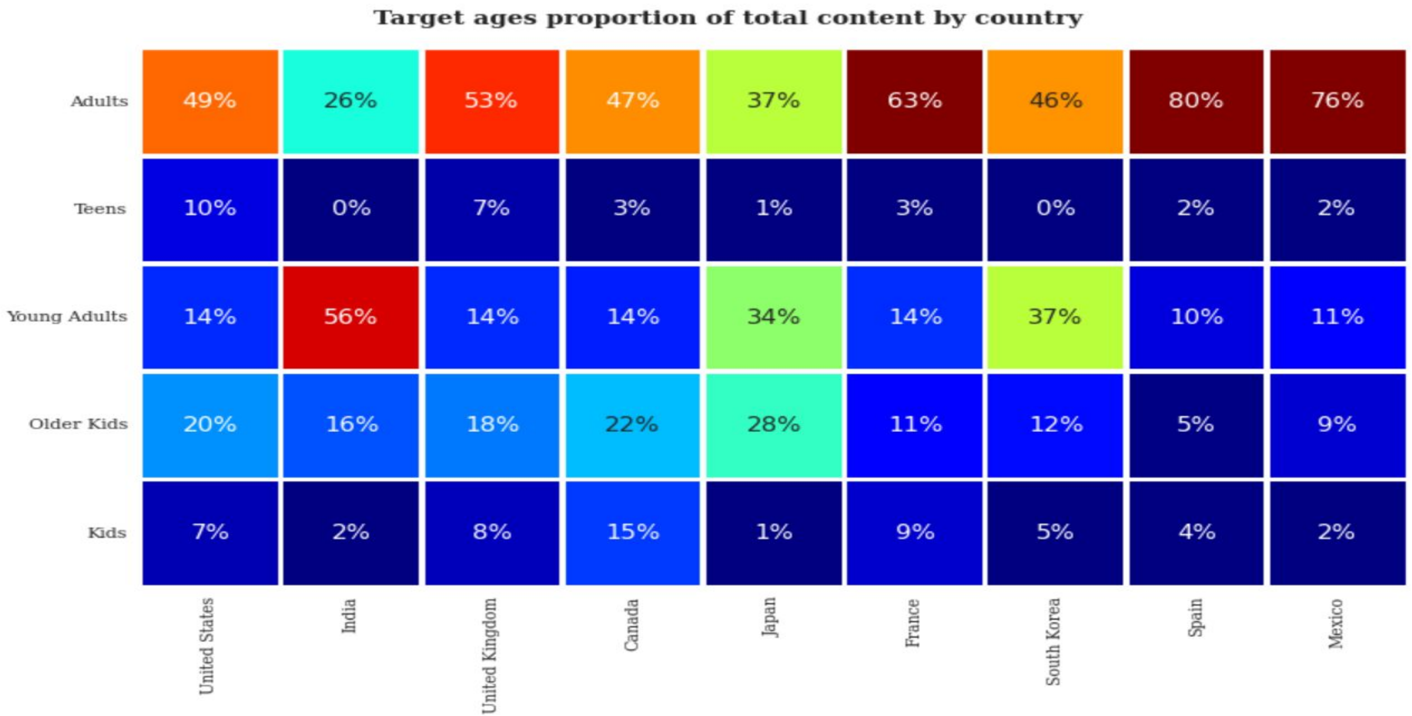
Top 5 countries

	country	Productions	TV-Shows	Movies
0	United States	3288	860	2428
1	India	990	75	915
2	United Kingdom	722	255	467
3	Country Unavailable	505	276	229
4	Canada	412	126	286

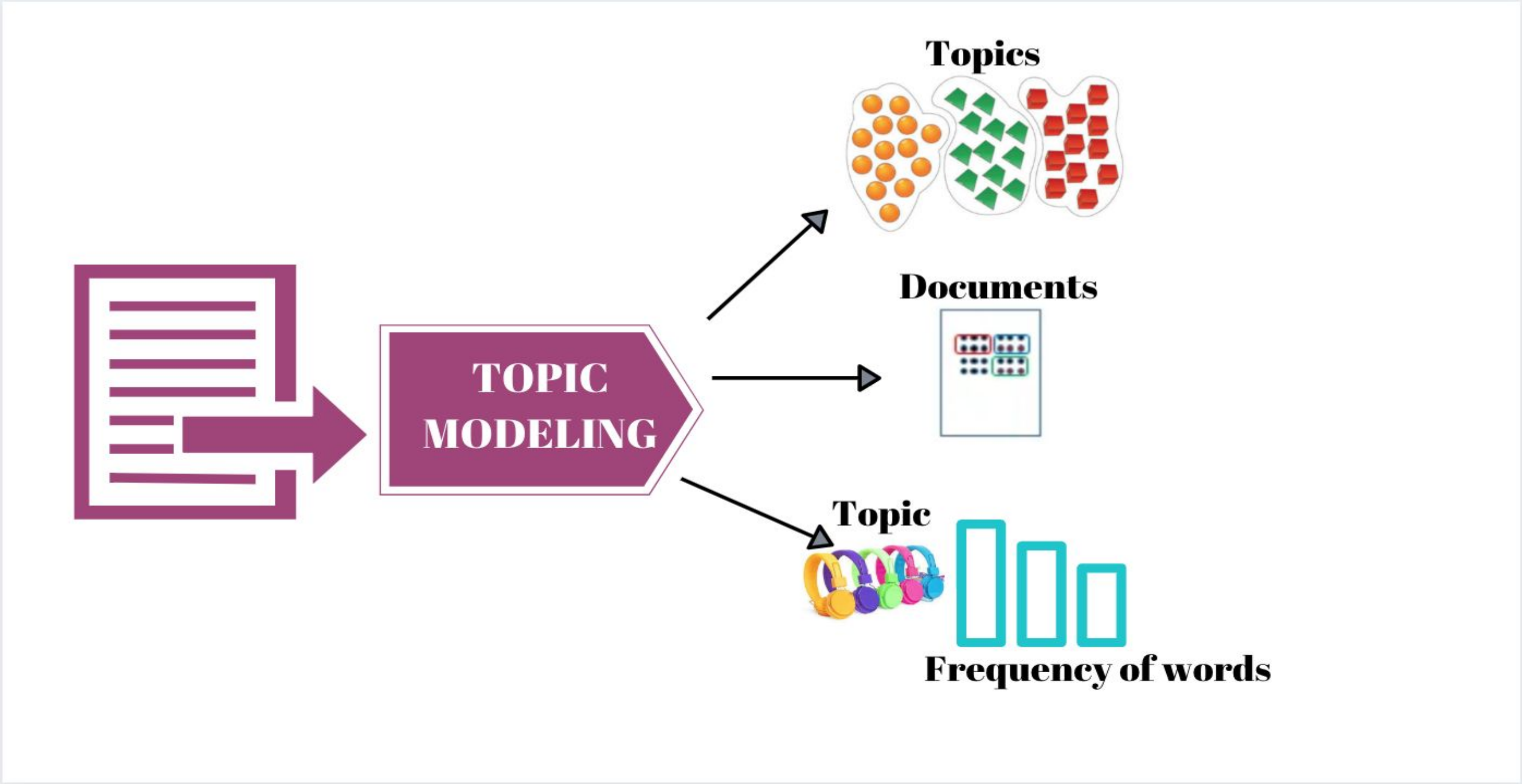
Content across all years



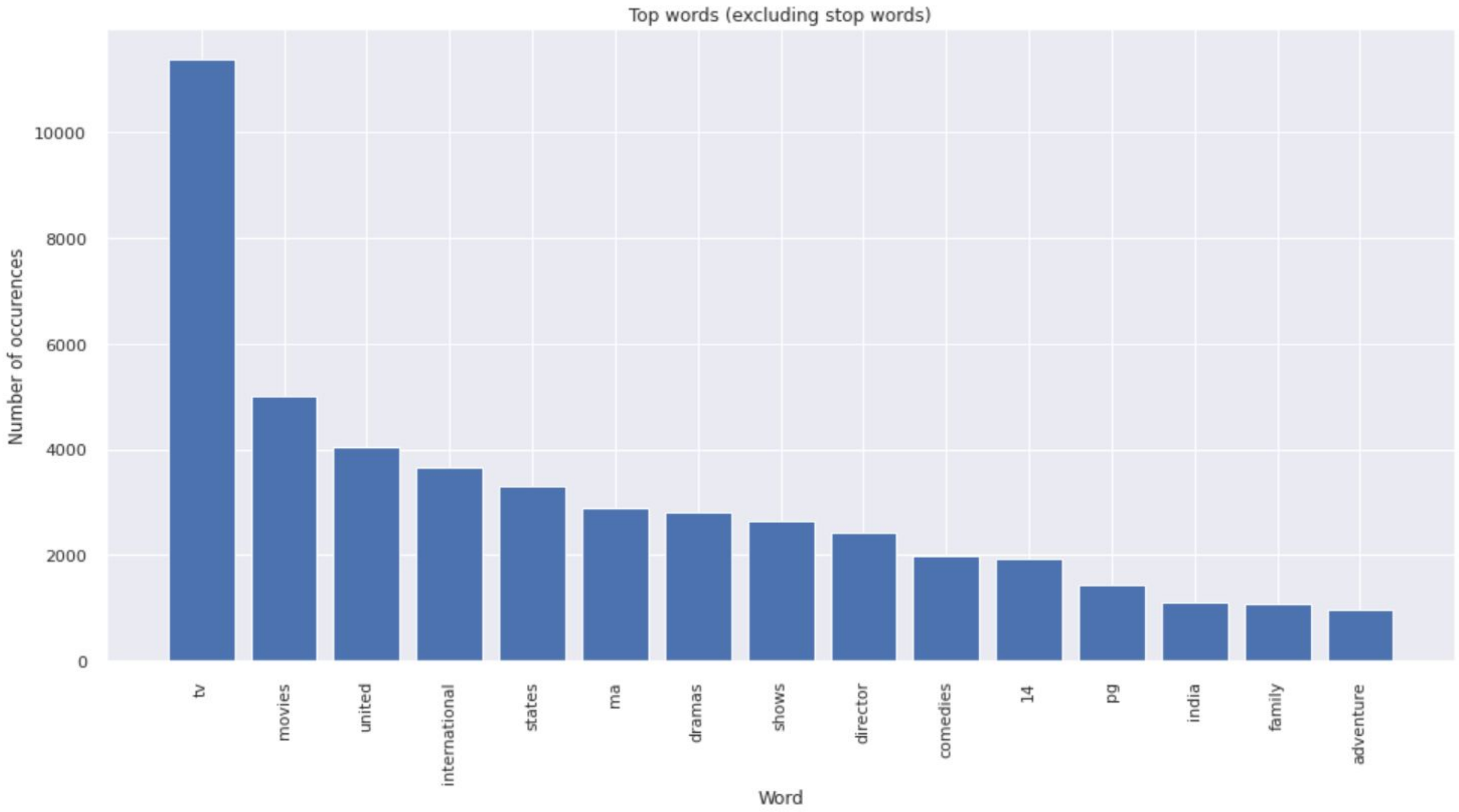
Netflix Content for different age groups in top 10 countries



Topic Modeling

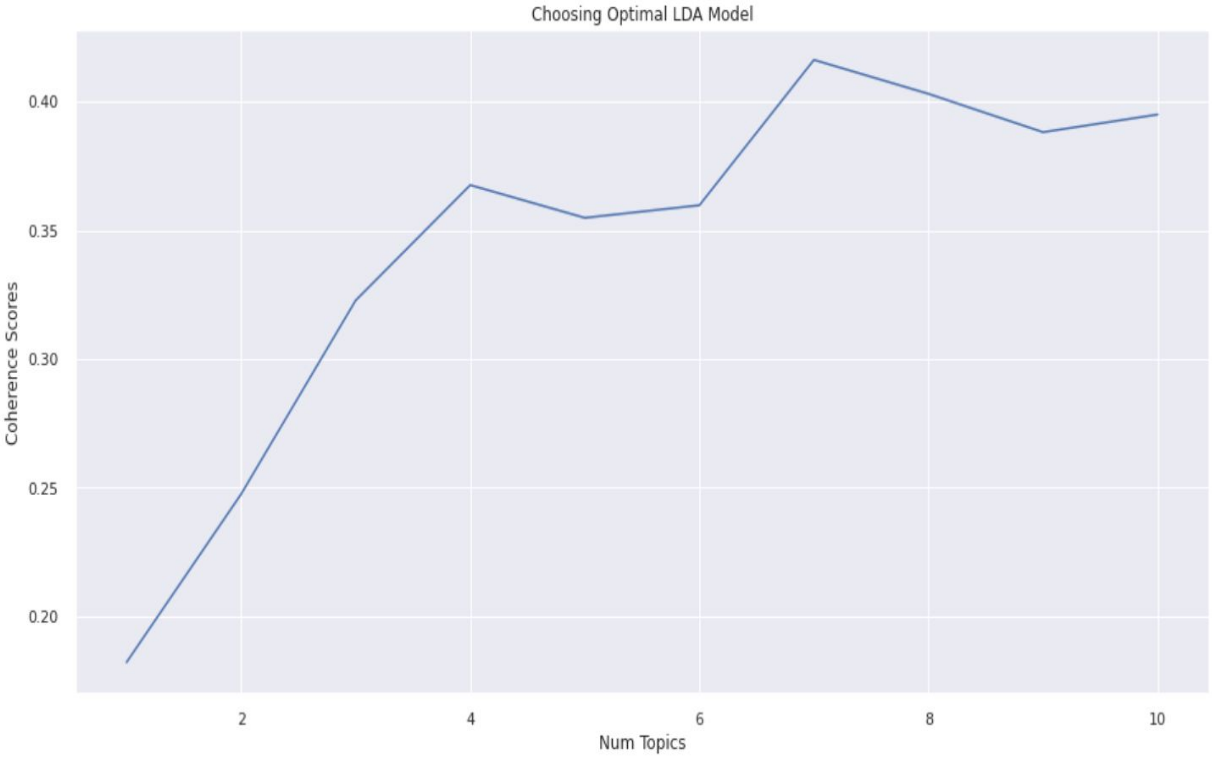


Top words (excluding stop words)

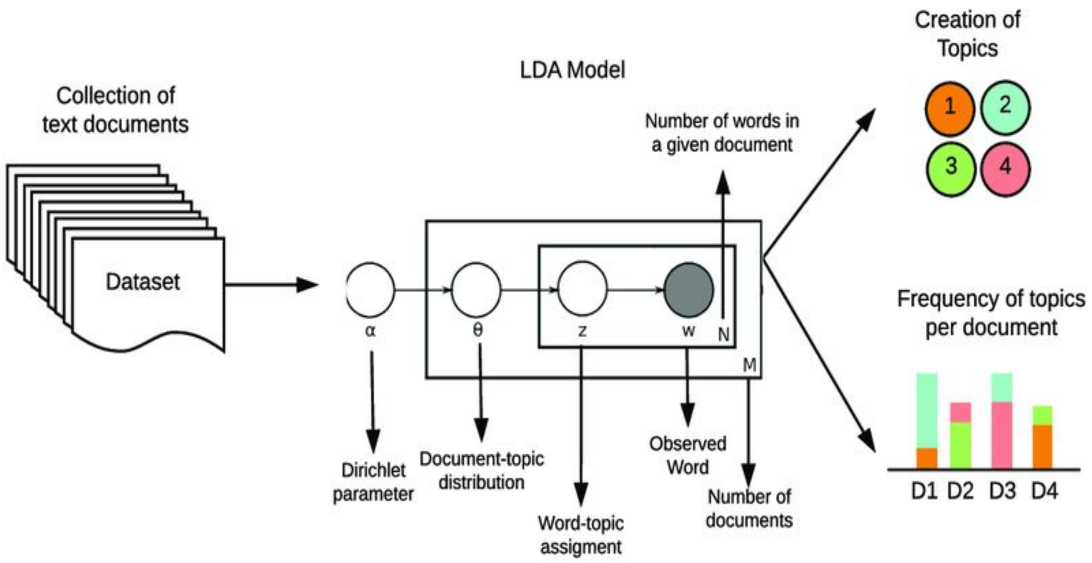


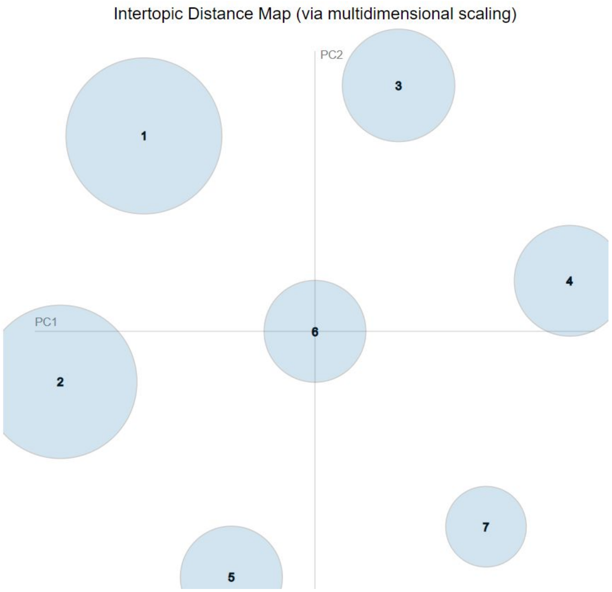
Coherence Score for Number of Topics

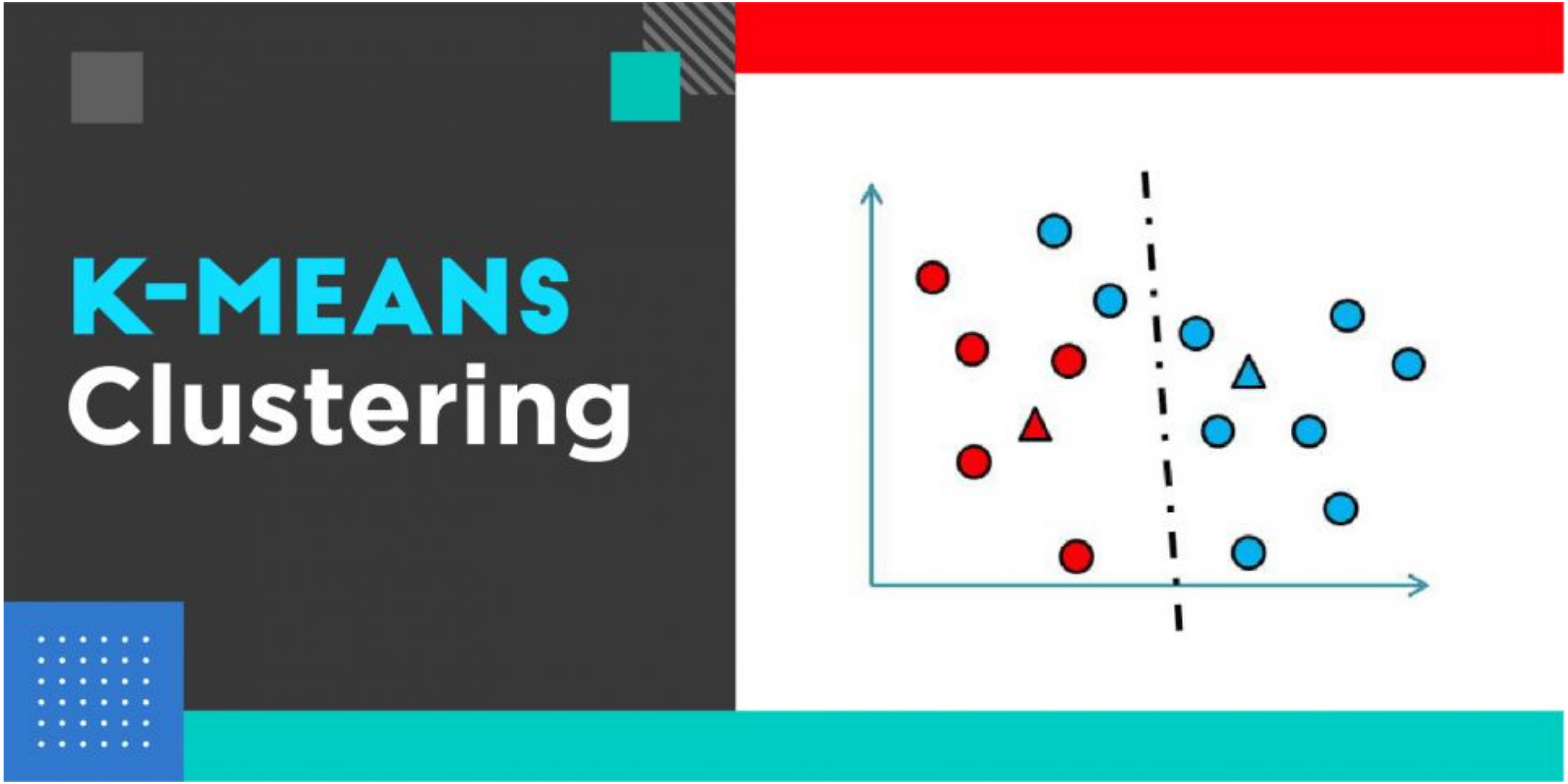
	topic_param	coherence_score
0	1	0.182079
1	2	0.247459
2	3	0.322775
3	4	0.367605
4	5	0.354870
5	6	0.359747
6	7	0.416282
7	8	0.403071
8	9	0.388175
9	10	0.395035



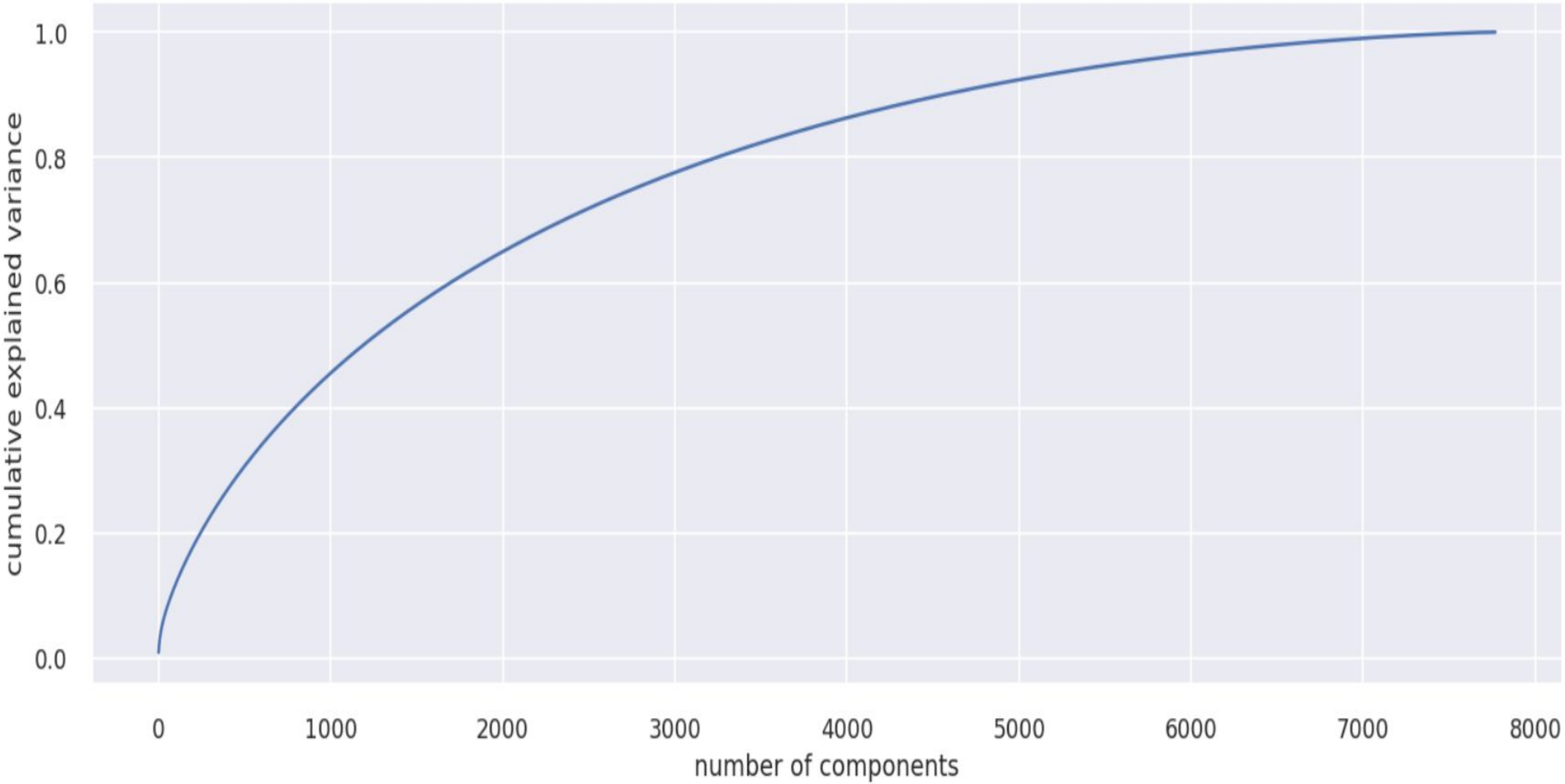
Latent Dirichlet Allocation (LDA)

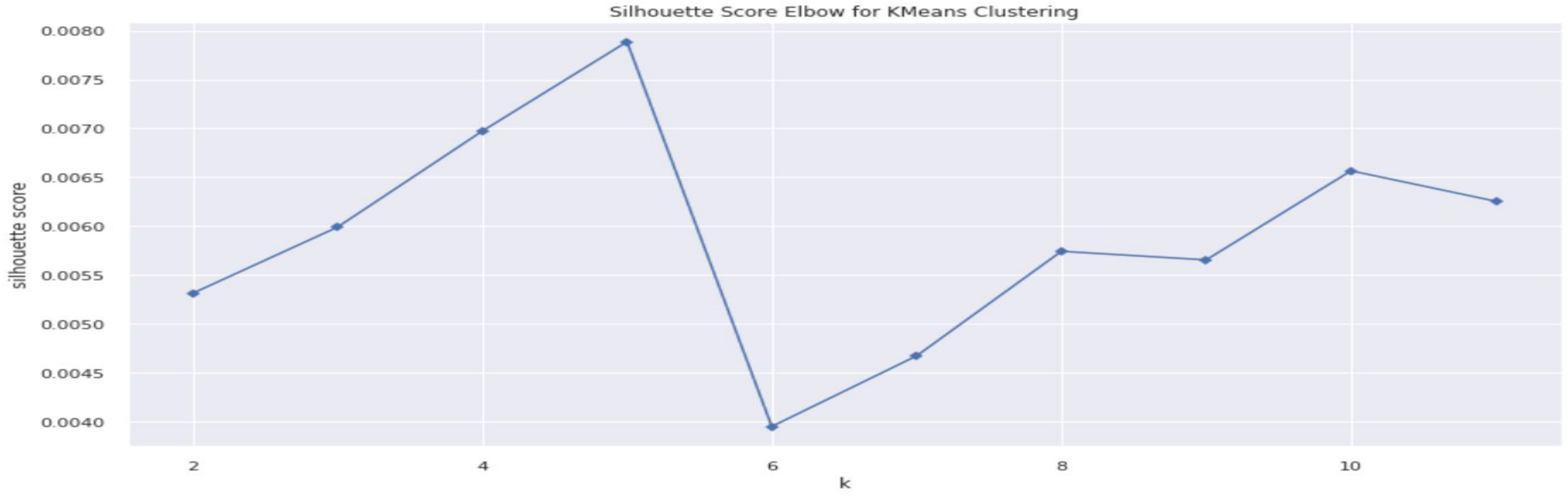


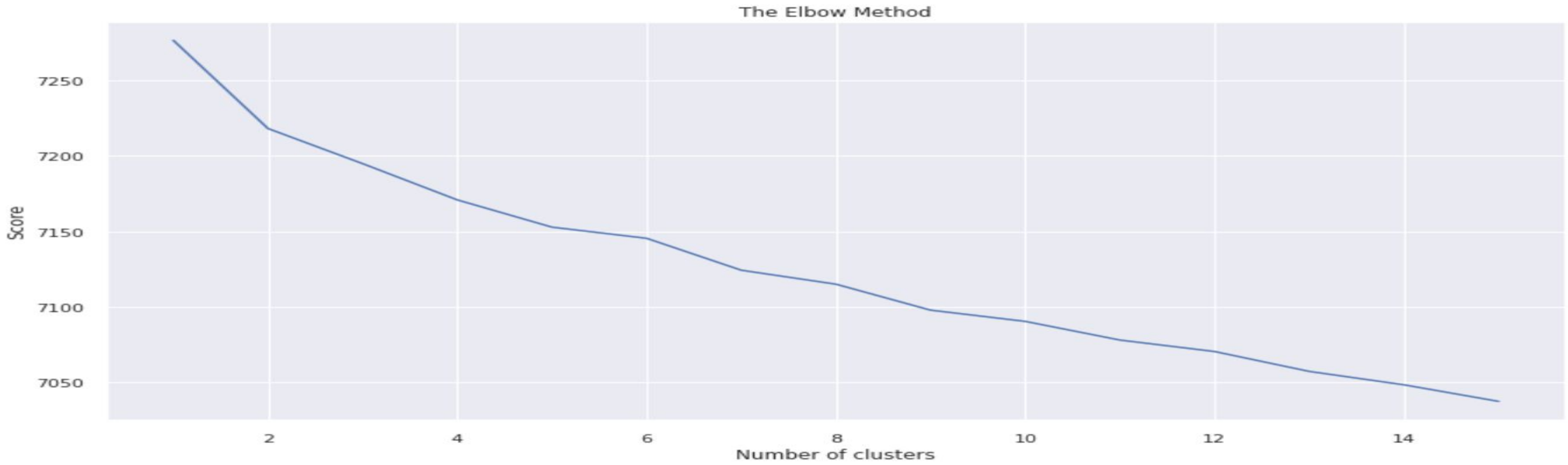



K-MEANS
Clustering

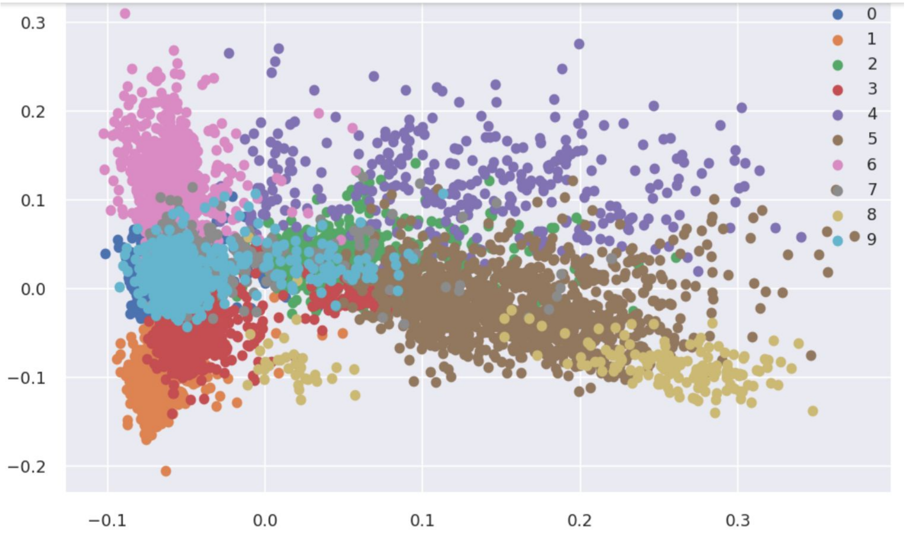
Finding number of components





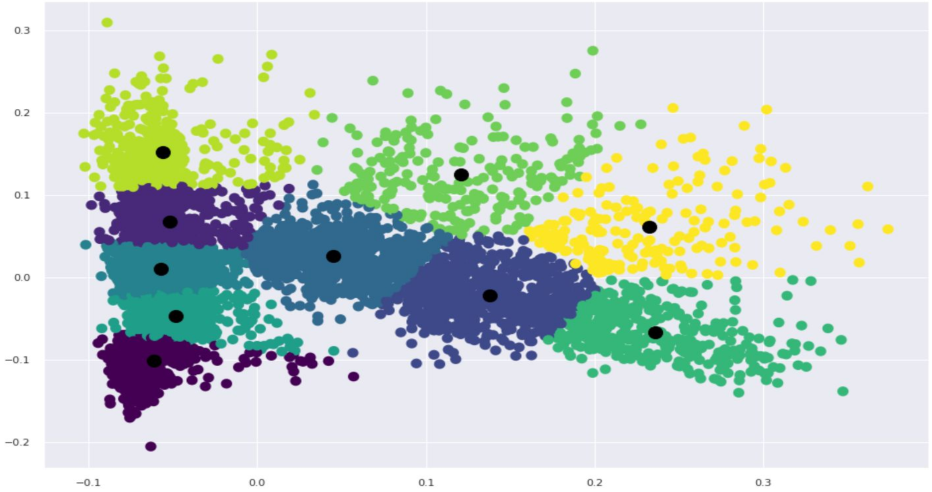


Finding number of clusters



The number of clusters = 10

Clusters Formation



Recommendations

Recommendations for Movies
Zulu Man in Japan

	Recommendations
0	ZZ TOP: THAT LITTLE OL' BAND FROM TEXAS
1	Roots
2	Nightmare Tenants, Slum Landlords
3	Bad Rap
4	We Are One
5	Rhythm + Flow
6	LA Originals
7	Emicida: AmarElo - It's All For Yesterday
8	Miss Sharon Jones!
9	Trixie Mattel: Moving Parts

Recommendations for TV-Shows
3%

	Recommendations
0	All The Reasons To Forget
1	Lo que la verdad esconde: El caso Asunta (Oper...
2	Kissing Game
3	The Chosen One
4	Back with the Ex
5	Blood Pact
6	Secrets of Great British Castles
7	Away
8	The Underclass
9	Million Pound Menu

CONCLUSION

- More Movies(69.1%) on Netflix than TV shows(30.9%).
- Growth in TV shows from 2018 to 2020 and decreases in movies from 2019 to 2020. Therefore Netflix has increasingly focusing on TV rather than movies in recent years.
- United State was highest contributor on Netflix.
- Topic modeling by using Latent Dirichlet Allocation (LDA) perform on text dataset and obtain highest coherence score on 7 number of topic. Feed the LDA model into the pyLDAvis instance and obtain intertopic distance map (via multidimensional scaling).
- k=10 was found to be an optimal value for clusters using which we grouped our data into 10 distinct clusters and obtain cluster using k=10 and found top words obtain in cluster. Using the given data a simple recommender system was created using cosine_similarity and recommendations for Movies and TV Shows were obtained.


THANK
YOU