Capstone Project

Netflix Movies and TV Shows Clustering

Ashik Kumar

Contents

Introduction

- As the world’s leading Internet television network with over 160 million members in over 190 countries, our members enjoy hundreds of millions of hours of content per day, including original series, documentaries and feature films. We invest heavily in machine learning to continually improve our member experience and optimize the Netflix service end to end.

Problem Statement

- The goal of this project is to find out similarity within groups in people to build a movie recommendation system for users. We are going analyze a dataset from Netflix database to explore the characteristics that people share in movies taste.
- We have experienced it ourselves or have been in the room is the endless scrolling of selecting what to watch. Users spend more time deciding what to watch than watching their movie.

Data Summary

This dataset consists of tv shows and movies available on Netflix as of 2019. The dataset is collected from Fixable 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. It will be interesting to explore what all other insights can be obtained from the same dataset.

Integrating this dataset with other external datasets such as IMDB ratings, rotten tomatoes can also provide many interesting findings.

- show_id : Unique ID for every Movie / Tv Show
- title : Title of the Movie / Tv Show
- cast : Actors involved in the movie / show produced
- date_added : Date it was added on Netflix show
- rating : TV Rating of the movie / show seasons
- listed_in : Genere

type : Identifier - A Movie or TV Show

director : Director of the Movie

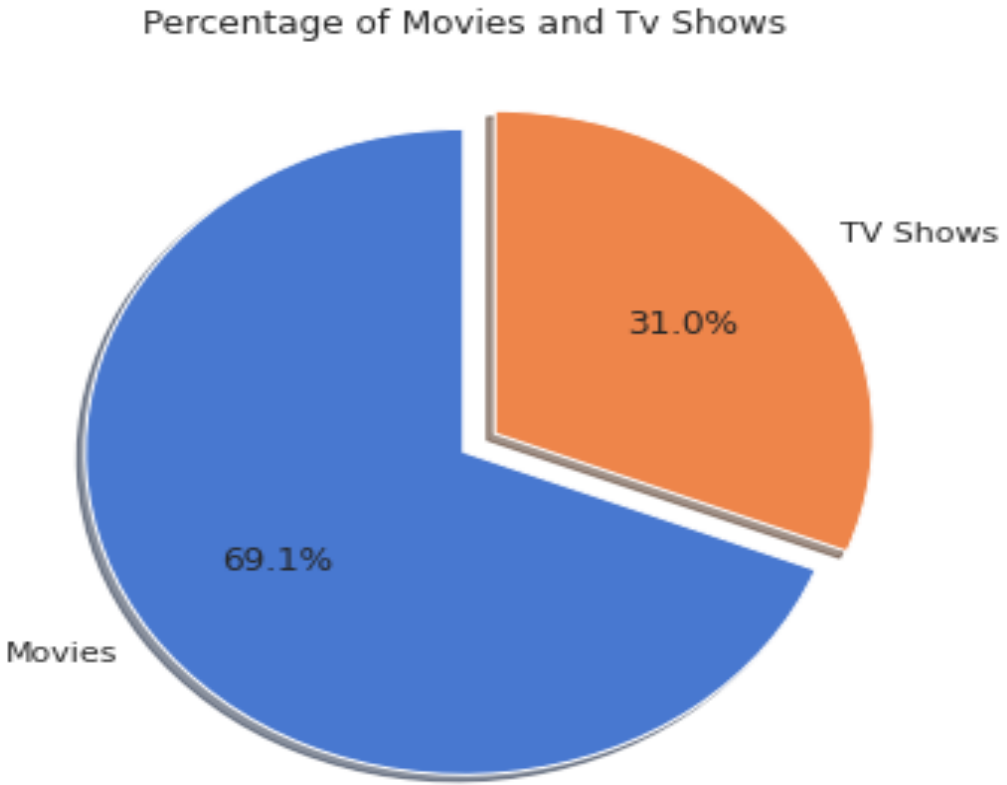
country : Country where the movie / show was

release_year : Actual Releaseyear of the movie /

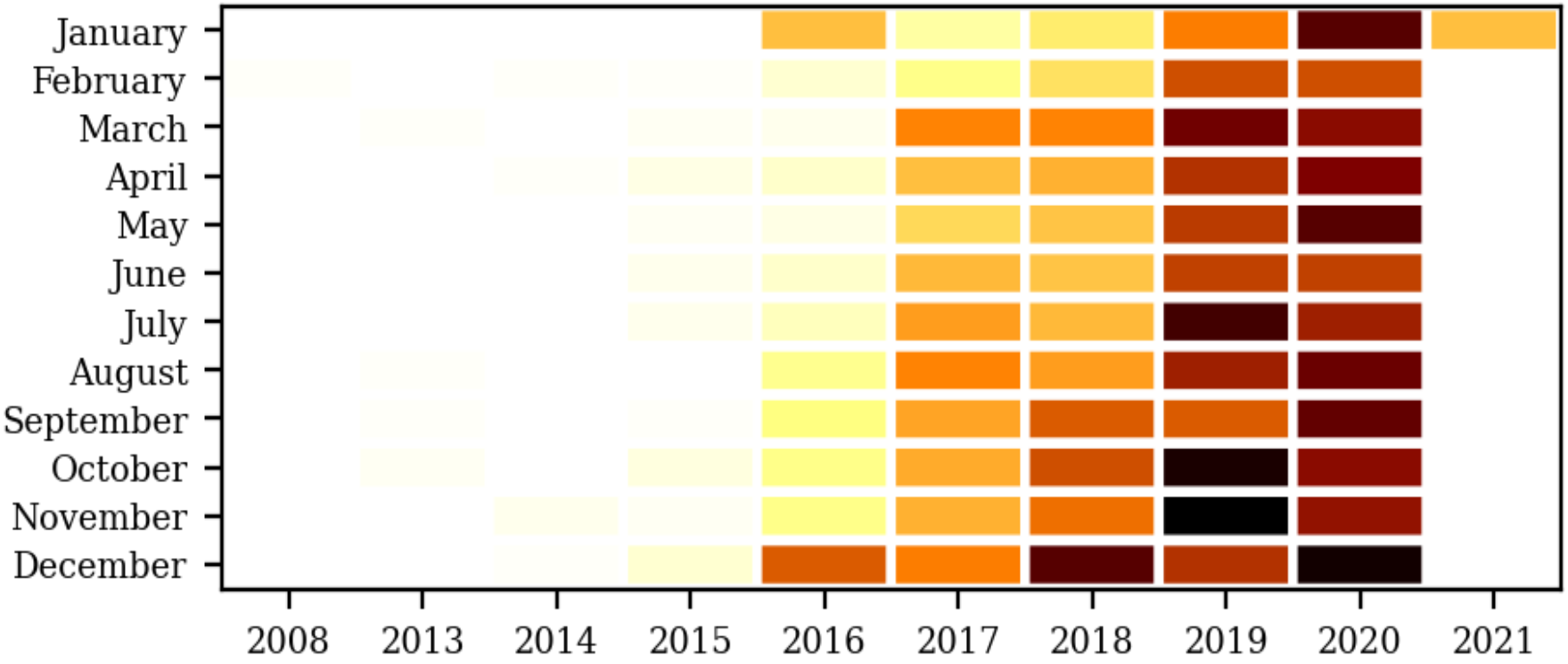
duration : Total Duration - in minutes or number of

description: The Summary description

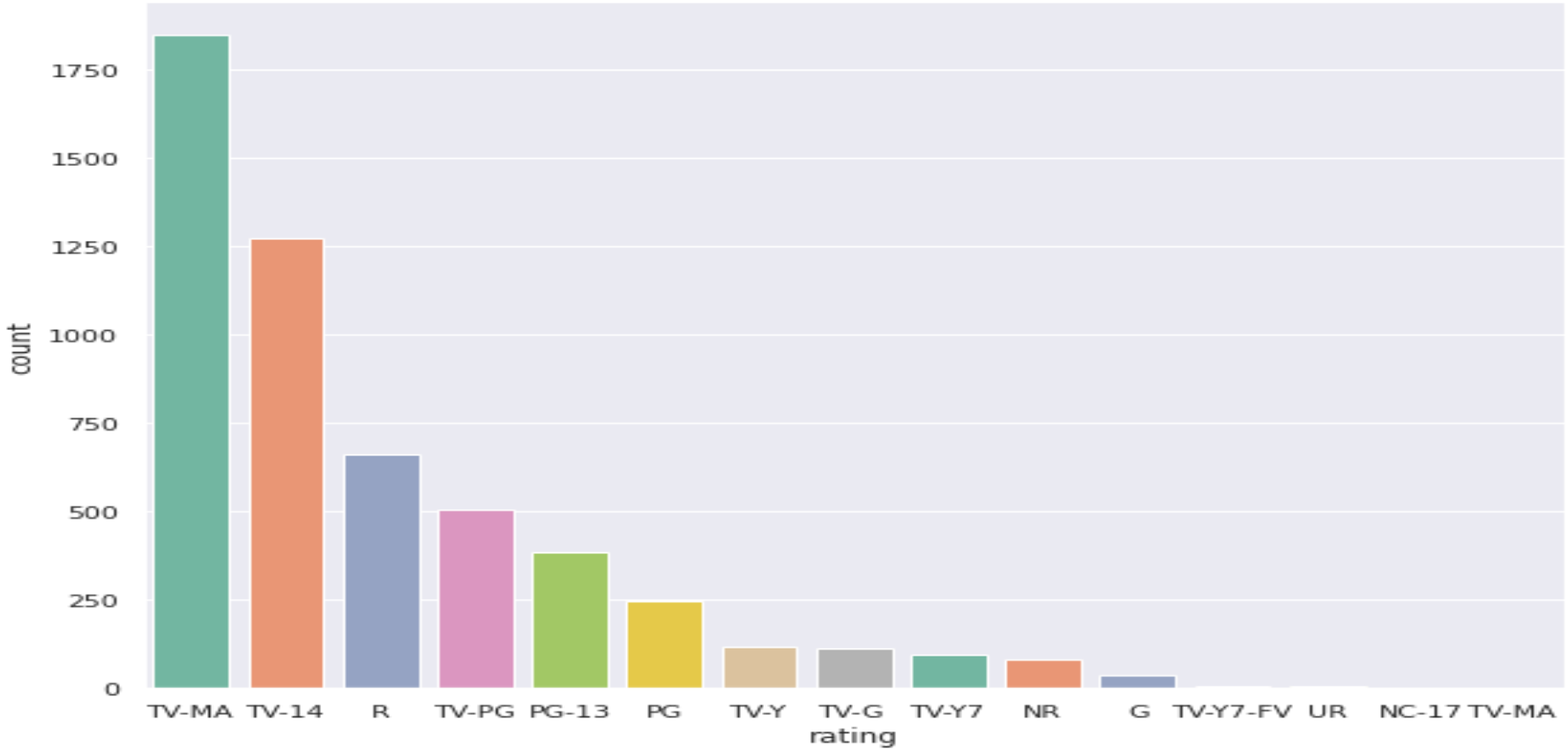
Explanatory Data Analysis



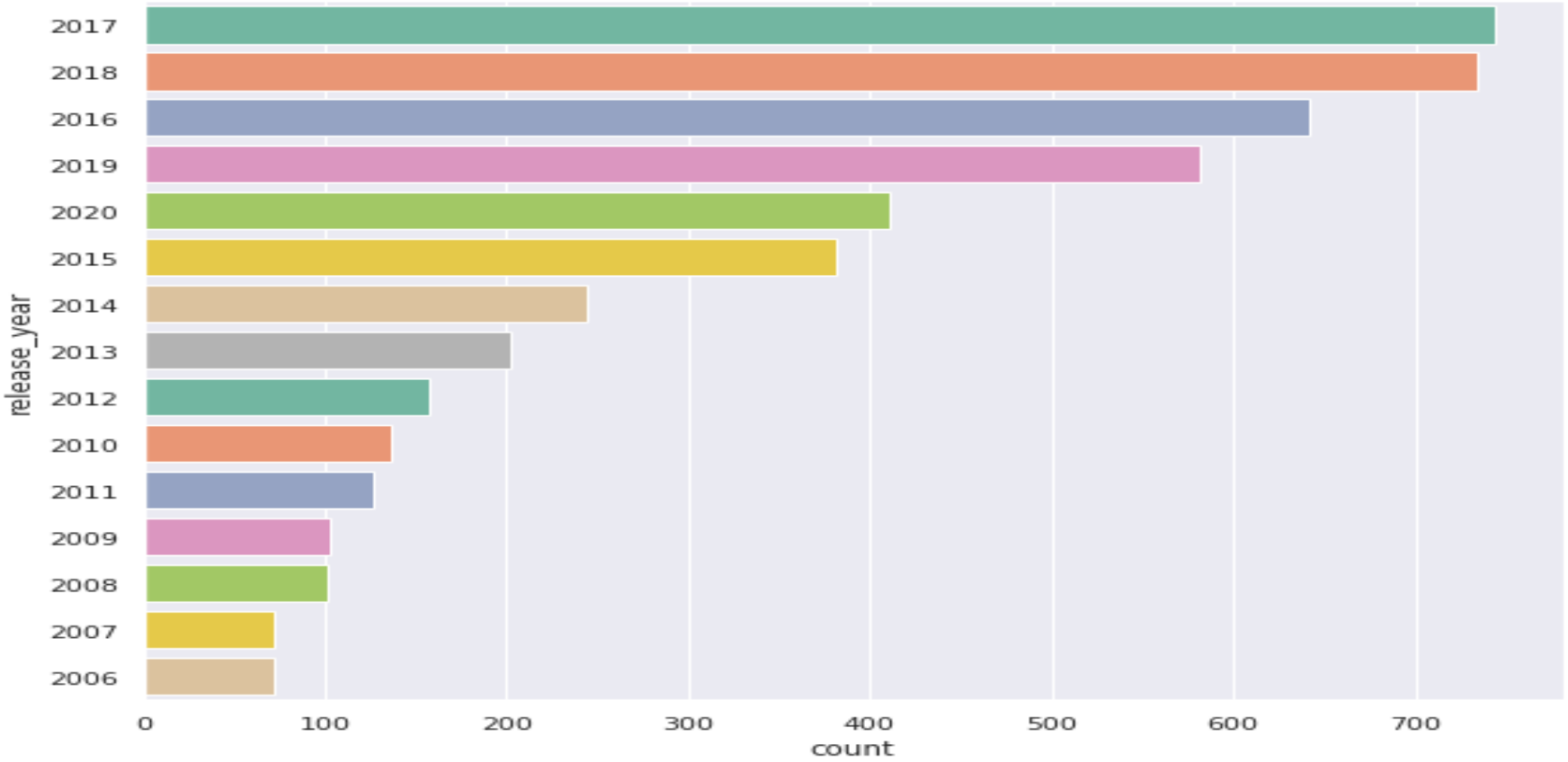
Monthly Content Release

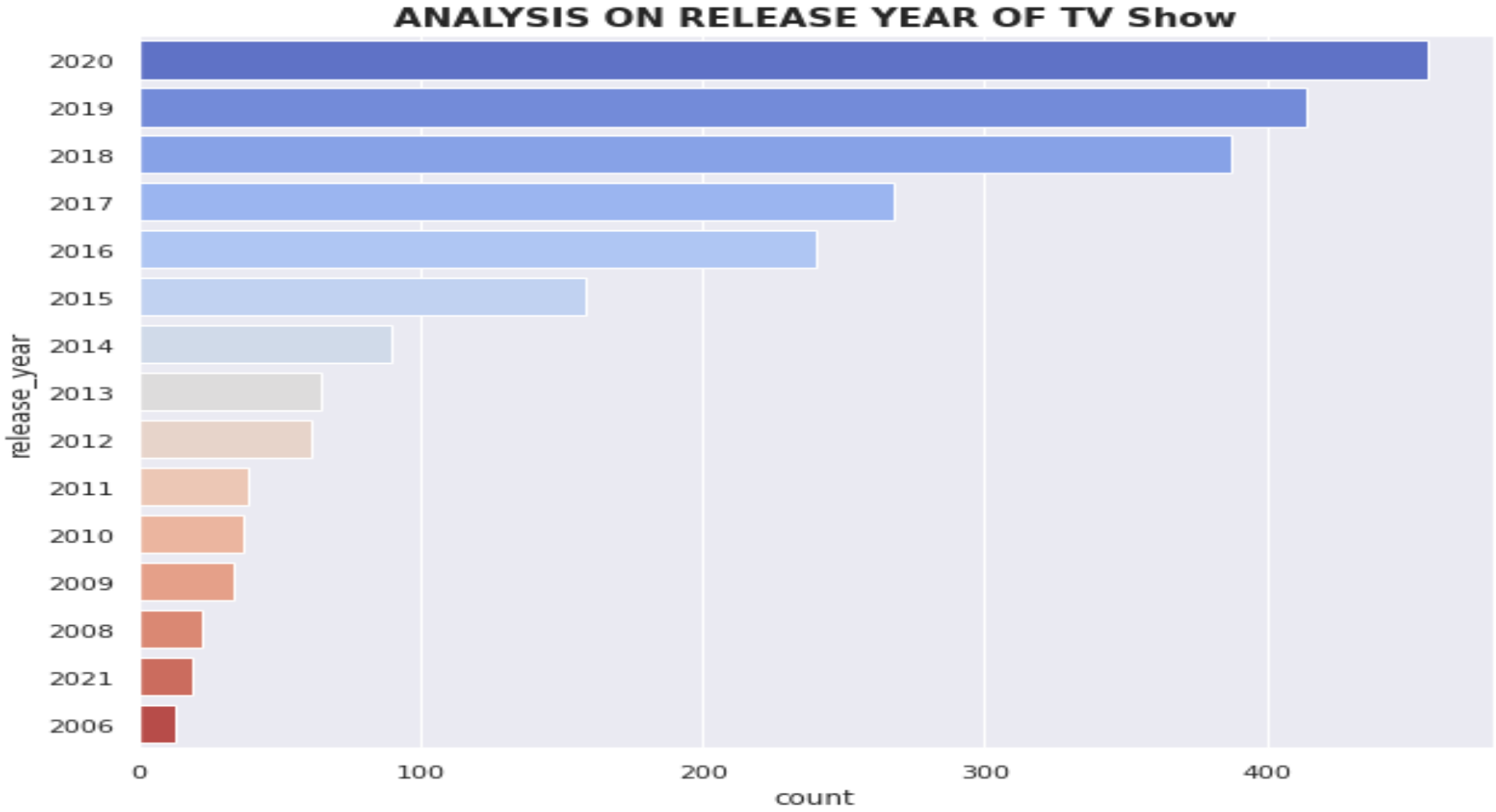


Rating

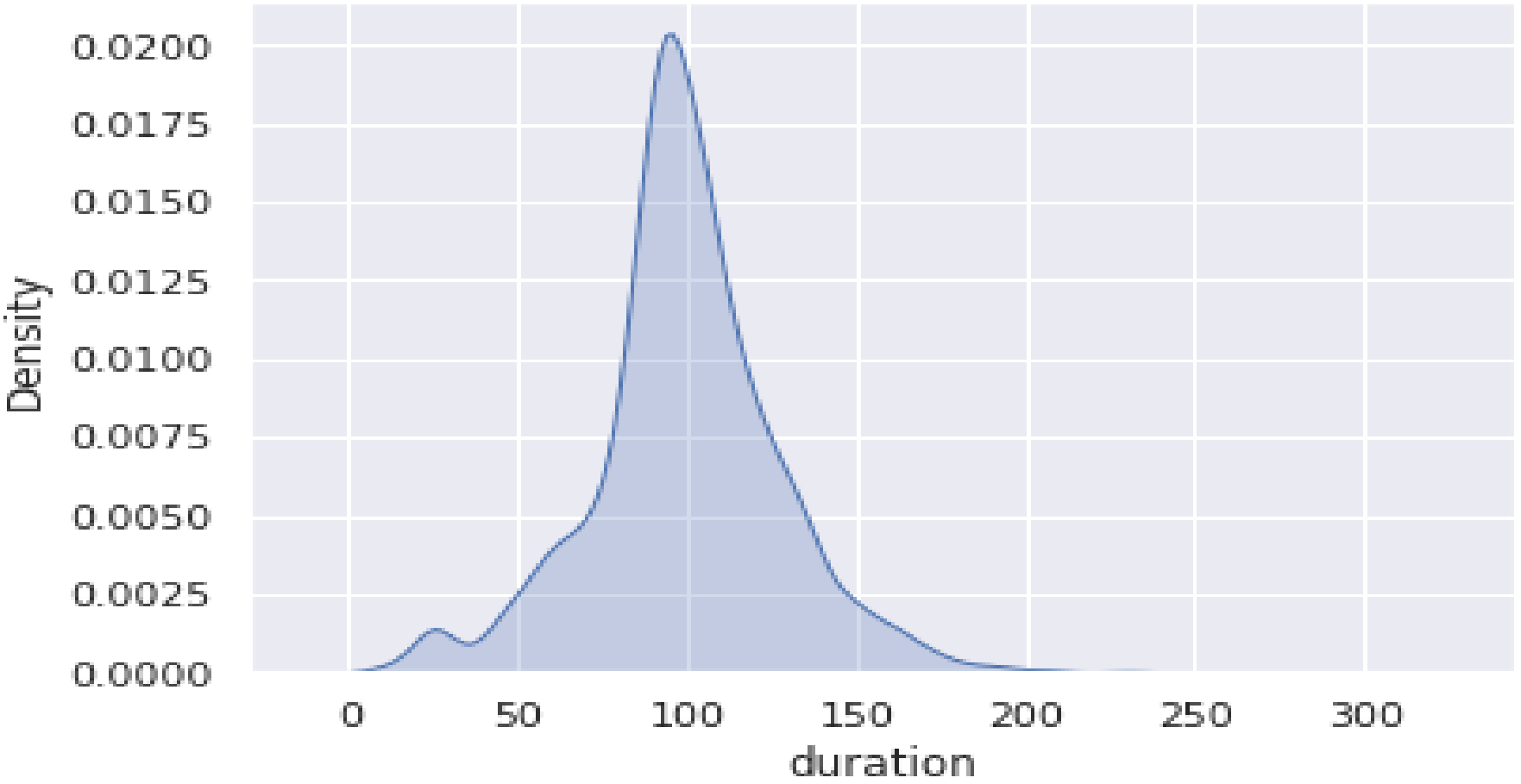

1750
1500
1250
1000
750
500
250
0
count
TV-MA
TV-14
R
TV-PG
PG-13
PG
TV-Y
TV-G
TV-Y7
NR
G
TV-Y7-FV
UR
NC-17
TV-MA
rating

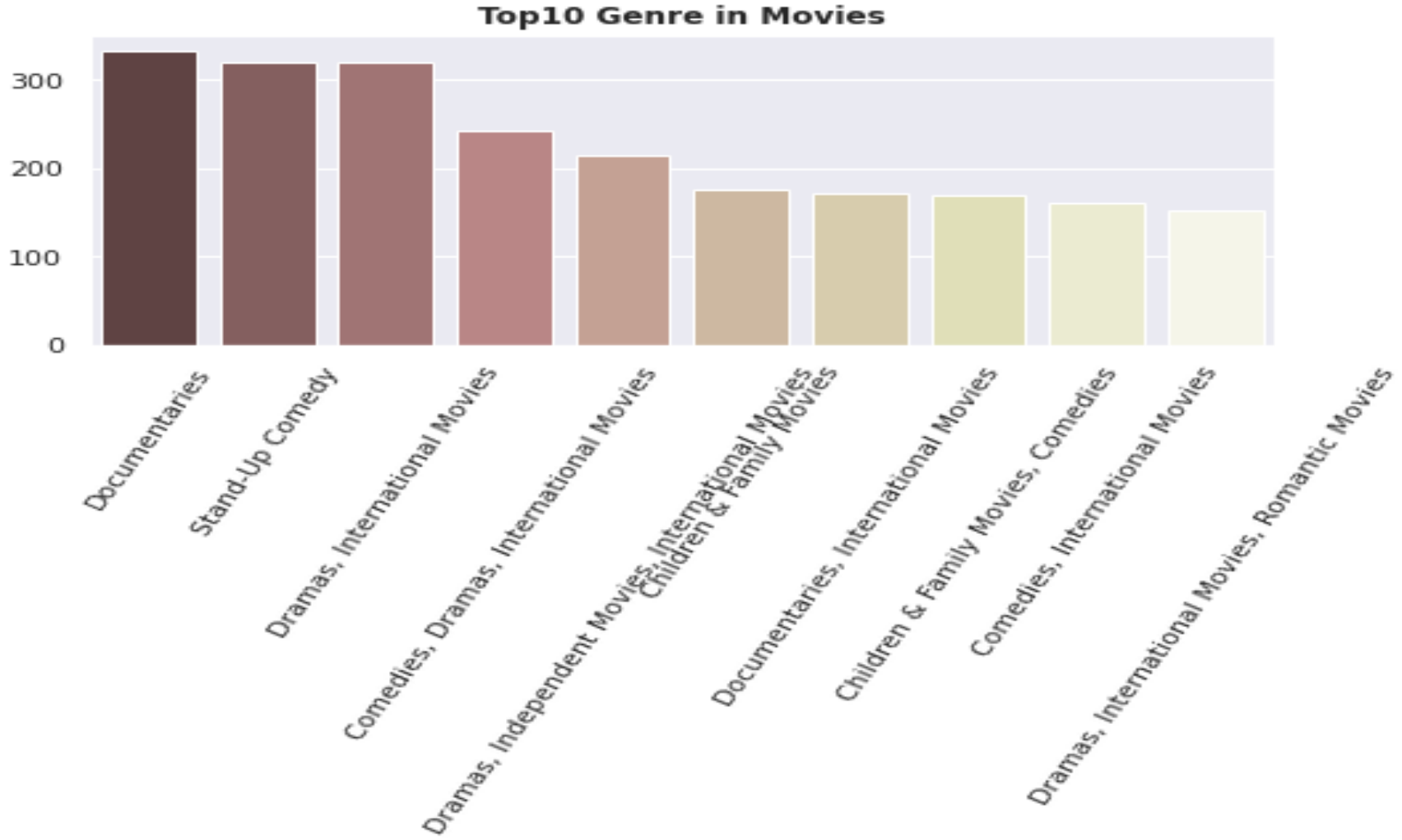
Number of content released per year

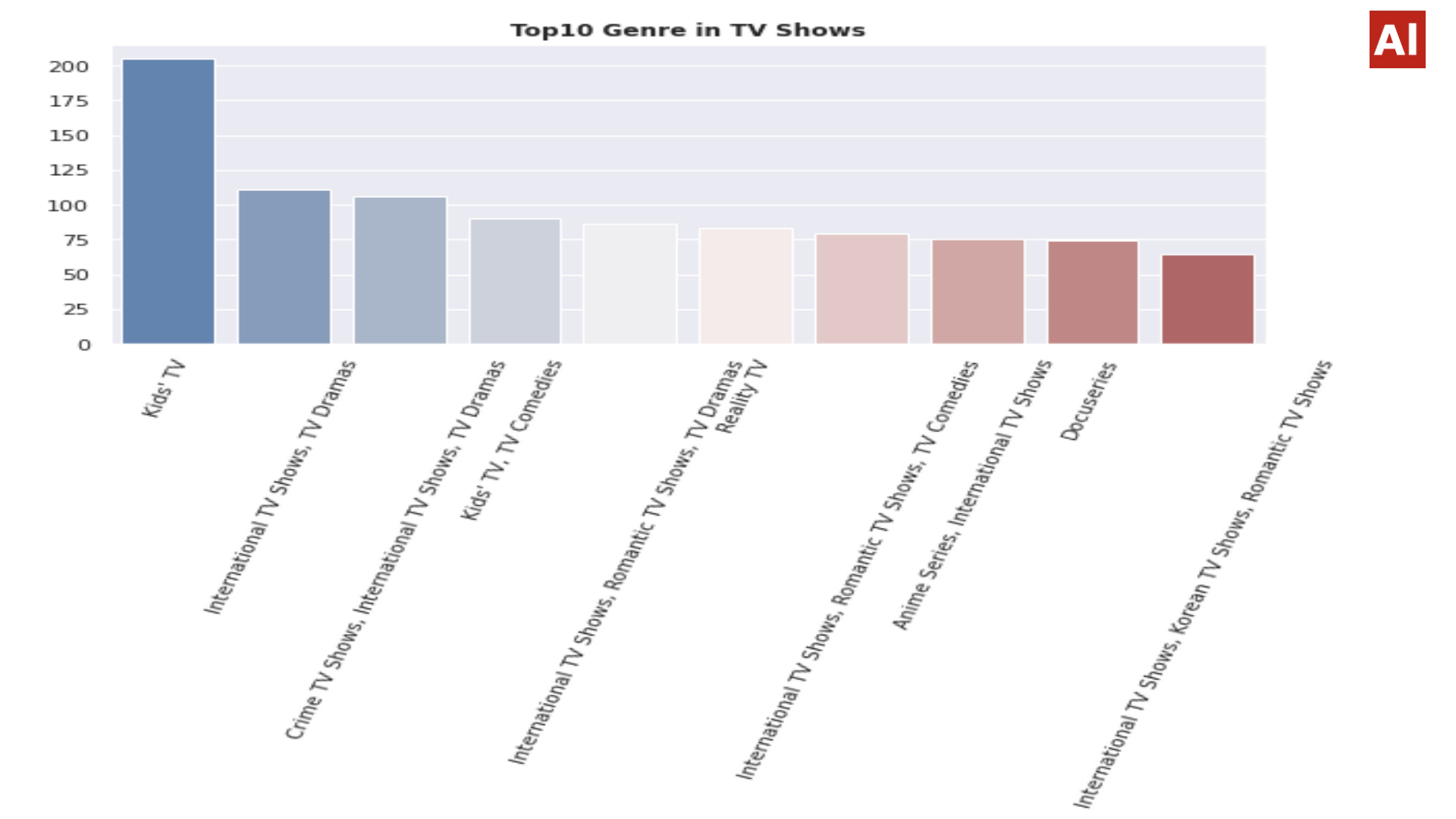


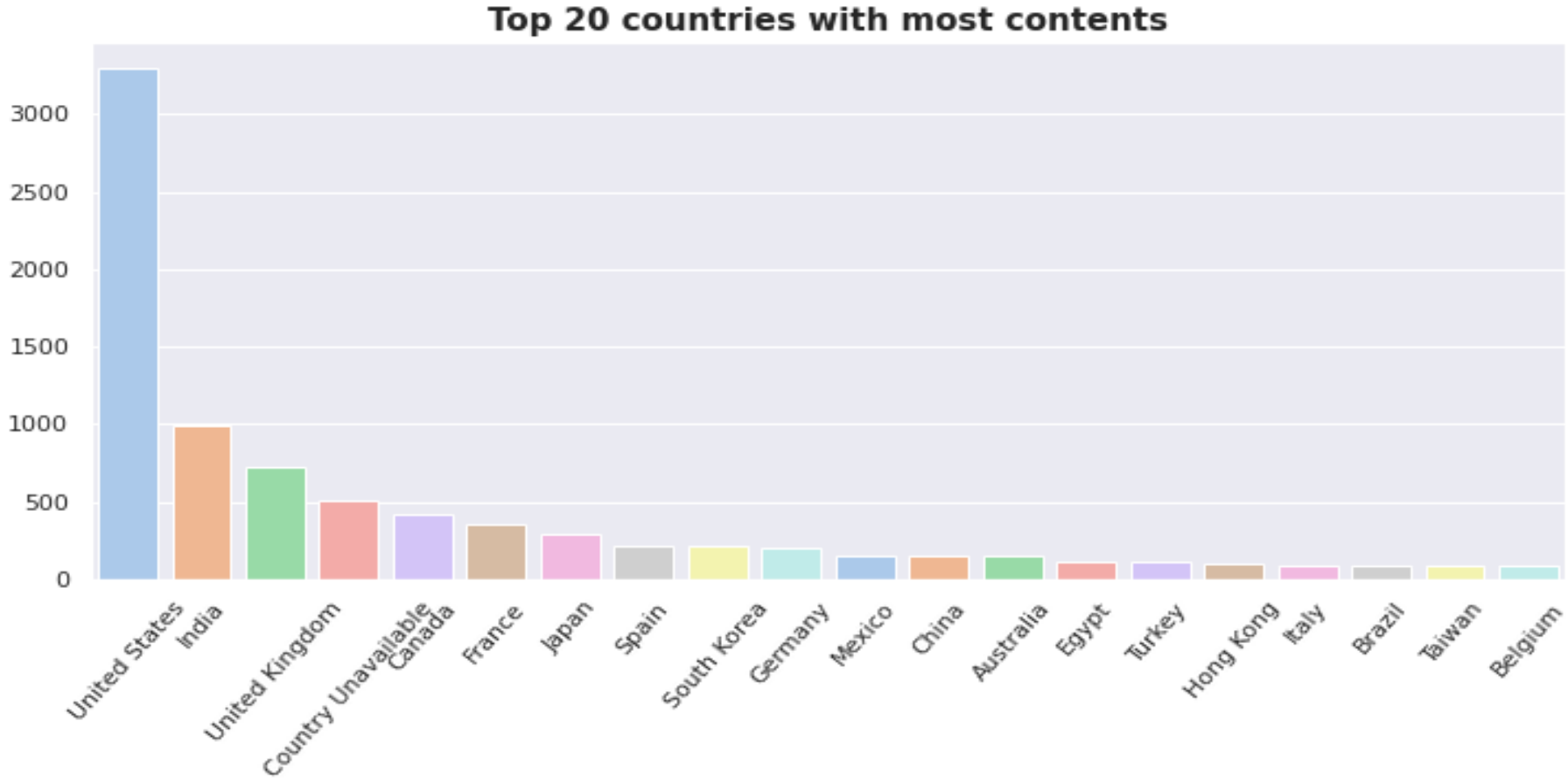

ANALYSIS ON RELEASE YEAR OF TV Show
release_year
2020
2019
2018
2017
2016
2015
2014
2013
2012
2011
2010
2009
2008
2021
2006
0
100
200
300
400
count

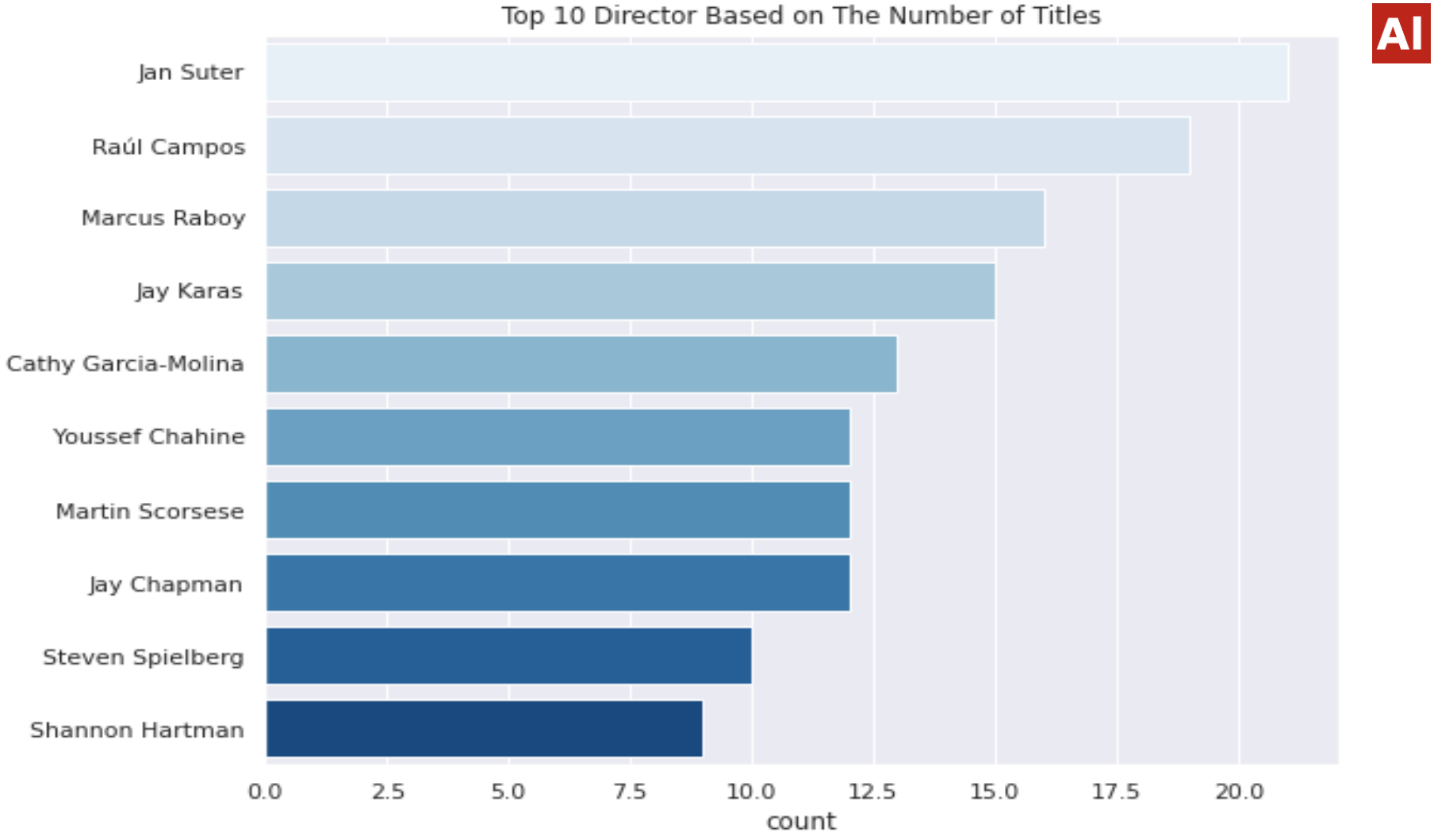
DURATION

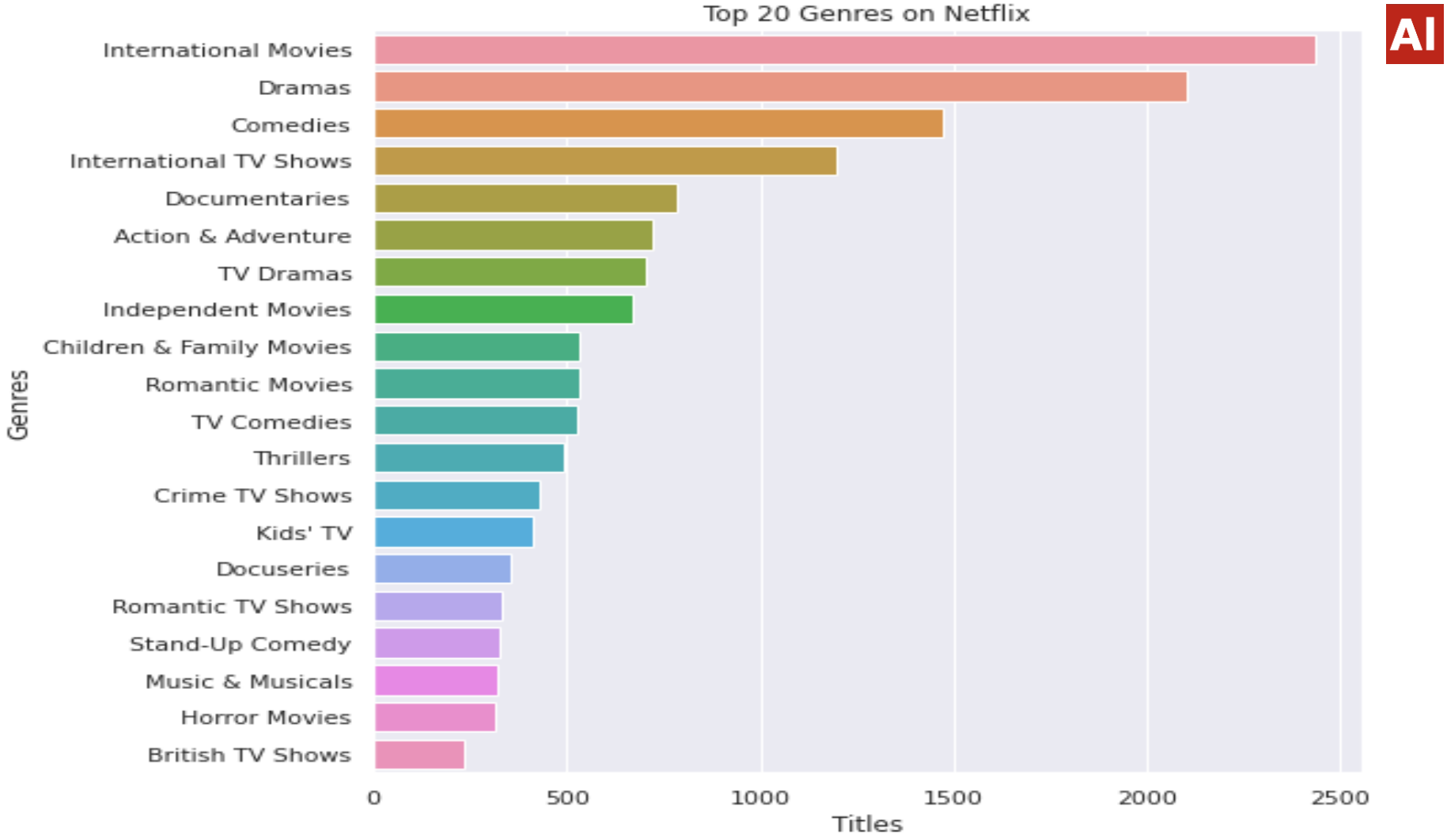

0.0200
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
Density
0
50
100
150
200
250
300
duration

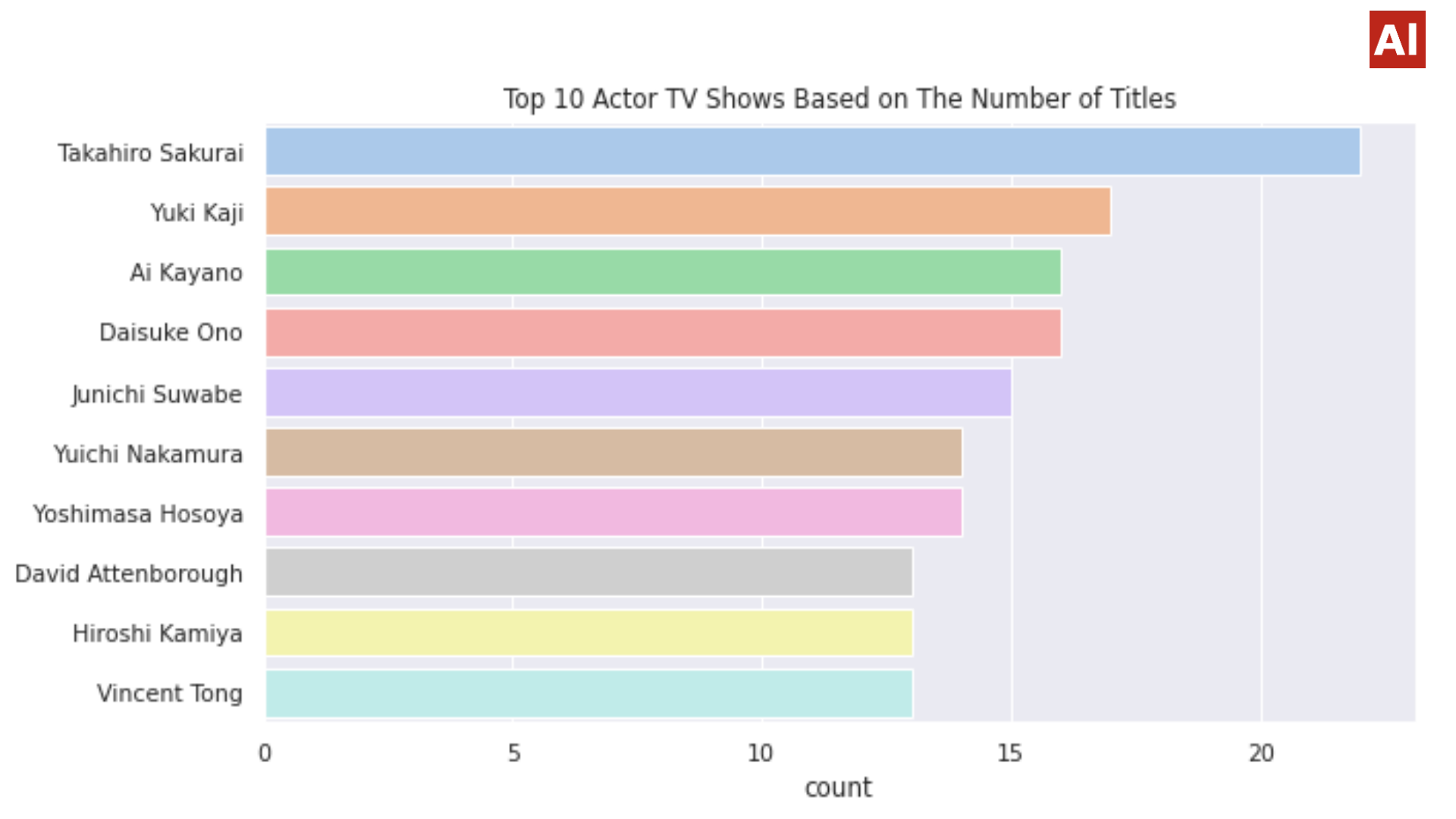

Top10 Genre in Movies
300
200
100
0
Documentaries
Stand-Up Comedy
Dramas, International Movies
Comedies, Dramas, International Movies
Dramas, Independent Movies, International Movies
Children & Family Movies
Documentaries, International Movies
Children & Family Movies, Comedies
Comedies, International Movies
Dramas, International Movies, Romantic Movies

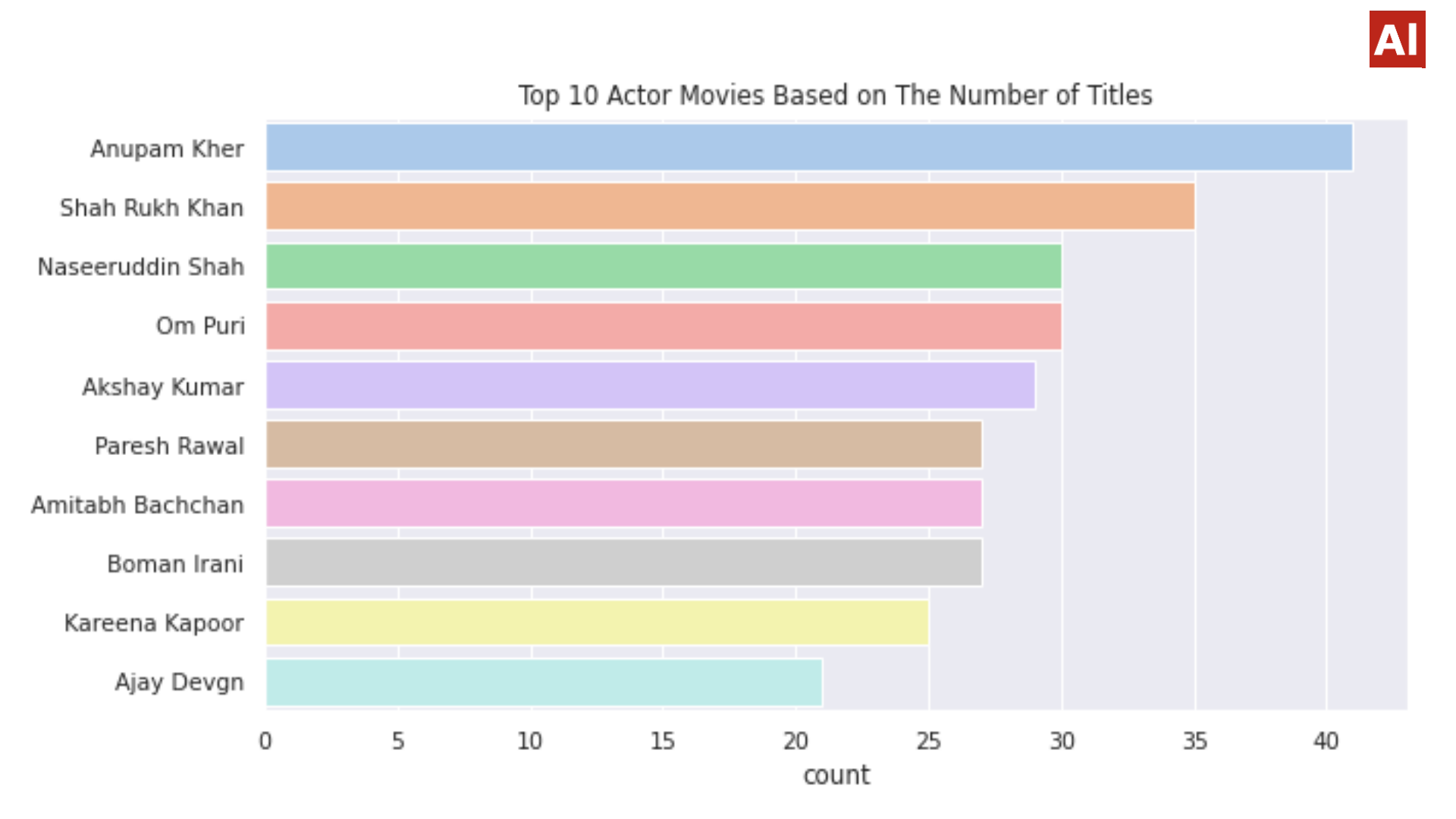

Top10 Genre in TV Shows
200
175
150
125
100
75
50
25
0
Kids' TV
International TV Shows, TV Dramas
Crime TV Shows, International TV Shows, TV Dramas
Kids' TV, TV Comedies
International TV Shows, Romantic TV Shows, TV Dramas
Reality TV
International TV Shows, Romantic TV Shows, TV Comedies
Anime Series, International TV Shows
Docuseries
International TV Shows, Korean TV Shows, Romantic TV Shows


Top 20 countries with most contents
3000
2500
2000
1500
1000
500
0
United States
India
United Kingdom
Country Unavailable
Canada
France
Japan
Spain
South Korea
Germany
Mexico
China
Australia
Egypt
Turkey
Hong Kong
Italy
Brazil
Taiwan
Belgium


Top 10 Director Based on The Number of Titles
Jan Suter
Raúl Campos
Marcus Raboy
Jay Karas
Cathy Garcia-Molina
Youssef Chahine
Martin Scorsese
Jay Chapman
Steven Spielberg
Shannon Hartman
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
count


Top 20 Genres on Netflix
International Movies
Dramas
Comedies
International TV Shows
Documentaries
Action & Adventure
TV Dramas
Independent Movies
Children & Family Movies
Romantic Movies
TV Comedies
Thrillers
Crime TV Shows
Kids' TV
Docuseries
Romantic TV Shows
Stand-Up Comedy
Music & Musicals
Horror Movies
British TV Shows
Genres
0
500
1000
1500
2000
2500
Titles


Top 10 Actor TV Shows Based on The Number of Titles
Takahiro Sakurai
Yuki Kaji
Ai Kayano
Daisuke Ono
Junichi Suwabe
Yuichi Nakamura
Yoshimasa Hosoya
David Attenborough
Hiroshi Kamiya
Vincent Tong
0
5
10
15
20
count


Top 10 Actor Movies Based on The Number of Titles
Anupam Kher
Shah Rukh Khan
Naseeruddin Shah
Om Puri
Akshay Kumar
Paresh Rawal
Amitabh Bachchan
Boman Irani
Kareena Kapoor
Ajay Devgn
0
5
10
15
20
25
30
35
40
count

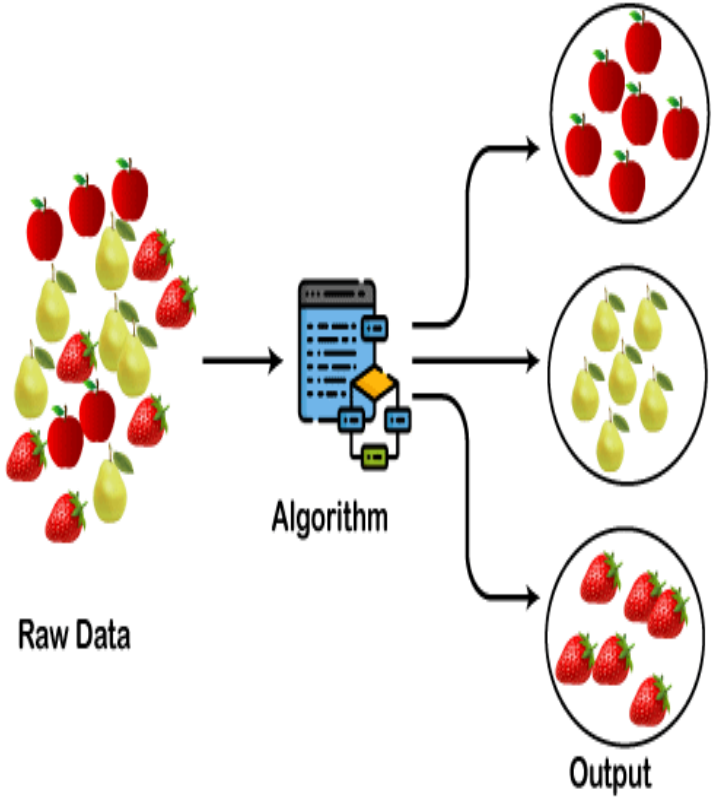
Dataset

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	s1	TV Show	3%	NaN	João Miguel, Bianca Comparato, Michel Gomes, R...	Brazil	August 14, 2020	2020	TV-MA	4 Seasons	International TV Shows, TV Dramas, TV Sci-Fi &...	In a future where the elite inhabit an island ...
1	s2	Movie	7:19	Jorge Michel Grau	Demián Bichir, Héctor Bonilla, Oscar Serrano, ...	Mexico	December 23, 2016	2016	TV-MA	93 min	Dramas, International Movies	After a devastating earthquake hits Mexico Cit...
2	s3	Movie	23:59	Gilbert Chan	Tedd Chan, Stella Chung, Henley Hii, Lawrence ...	Singapore	December 20, 2018	2011	R	78 min	Horror Movies, International Movies	When an army recruit is found dead, his fellow...
3	s4	Movie	9	Shane Acker	Elijah Wood, John C. Reilly, Jennifer Connelly...	United States	November 16, 2017	2009	PG-13	80 min	Action & Adventure, Independent Movies, Sci-Fi...	In a postapocalyptic world, rag-doll robots hi...
4	s5	Movie	21	Robert Luketic	Jim Sturgess, Kevin Spacey, Kate Bosworth, Aar...	United States	January 1, 2020	2008	PG-13	123 min	Dramas	A brilliant group of students become card-coun...

Row-7787, Columns-12

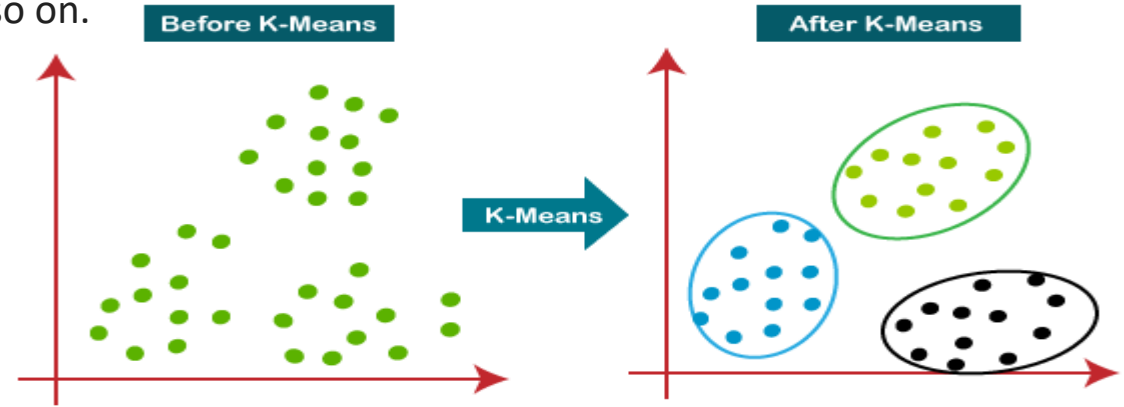
Clustering

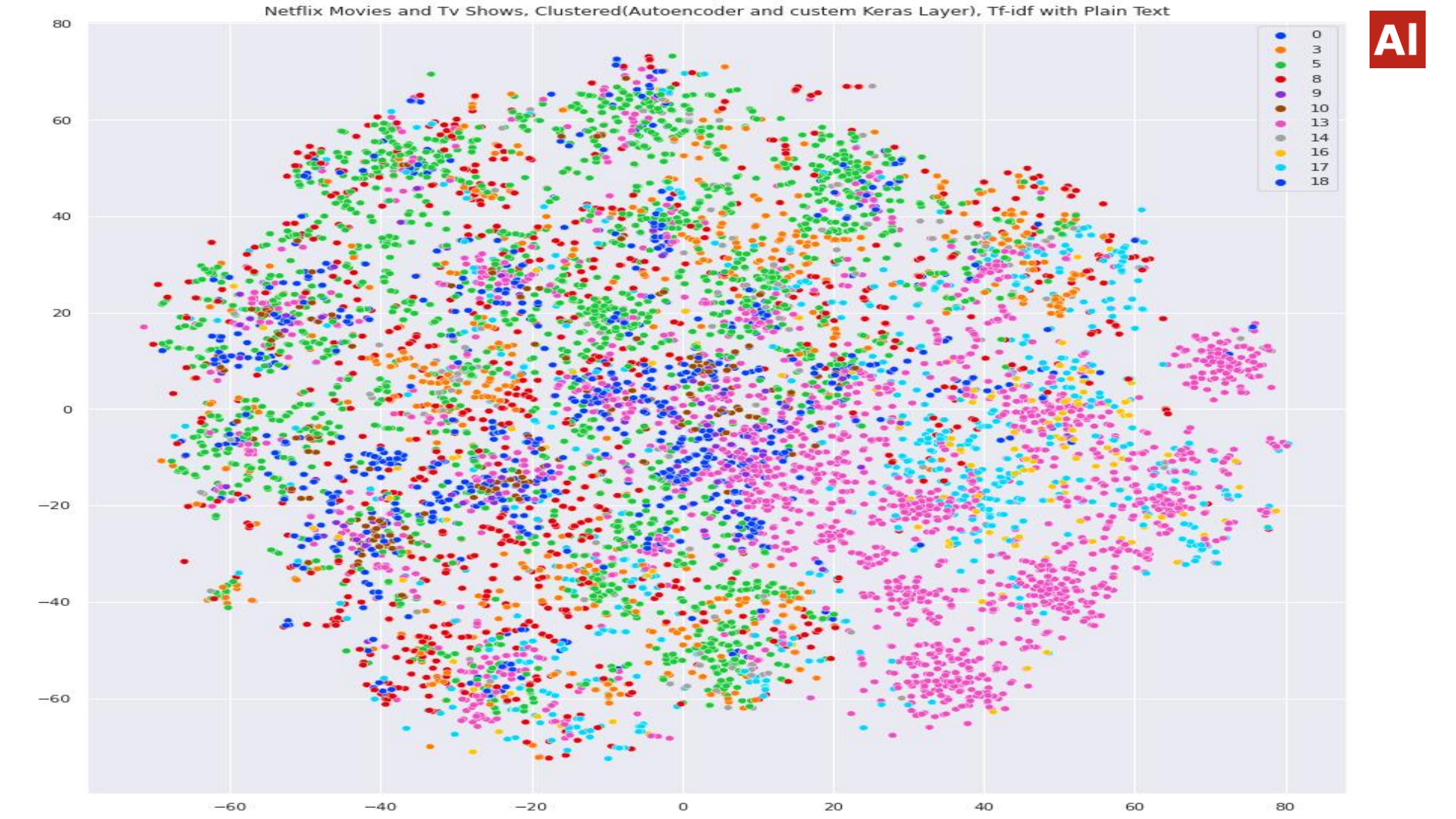
- Clustering or cluster analysis is a machine learning technique, which groups the unlabelled dataset. It can be defined as *"A way of grouping the data points into different clusters, consisting of similar data points. The objects with the possible similarities remain in a group that has less or no similarities with another group."*
- It does it by finding some similar patterns in the unlabelled dataset such as shape, size, color, behavior, etc., and divides them as per the presence and absence of those similar patterns.
- It is an unsupervised learning method, hence no supervision is provided to the algorithm, and it deals with the unlabeled dataset.



K-Means Clustering Algorithm

- The k-means algorithm is one of the most popular clustering algorithms. It classifies the dataset by dividing the samples into different clusters of equal variances. The number of clusters must be specified in this algorithm. It is fast with fewer computations required, with the linear complexity of O(n). K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.




Netflix Movies and Tv Shows, Clustered(Autoencoder and custem Keras Layer), Tf-idf with Plain Text
0
3
5
8
9
10
13
14
16
17
18
80
60
40
20
0
−20
−40
−60
−60
−40
−20
0
20
40
60
80

Conclusions:

1. The most content type on Netflix is movies. But Netflix has increasingly focusing on TV rather than movie in recent year.
2. The popular streaming platform started gaining traction after 2014. Since then, the amount of content added has been increasing significantly,
3. The country by the amount of the produces content is the United States,
4. The most popular director on Netflix , with the most titles, is Jan Suter
5. International Movies is a genre that is mostly in Netflix
6. The largest count of Netflix content is made with a “TV-14” rating
7. The most popular actor on Netflix TV Shows based on the number of titles is Takahiro Sakurai,
8. The most popular actor on Netflix movie, based on the number of titles, is Anupam Kher.
9. Documentaries is most famous genre in movie.
10. kid's TV is most famous genre in TV shows.
11. MostTV Shows end by season 3.
12. Mostly movies on Netflix are among the duration of 75-120 mins. It is considering the fact that a fair amount of the audience cannot watch a 3 hour movie in one sitting.
13. Most of the TV Shows were released in 2020 followed by 2019 and 2018
14. Most of the movies were released in the year 2017 followed by 2018 and 2016.

Thank You……