Capstone Project

Netflix Movies and TV Shows Clustering

By

Nidhi Pandey

AlmaBetter

Content

- Introduction
- Problem Statement
- Data Description
- EDA
- Feature Engineering
- Text Processing
- Topic Modelling
- Feature Selection
- Performance Metrics
- Observations
- Conclusion

Introduction

- Netflix began experimenting with data since 2006 when they attempted to predict how much a viewer would like a movie based on existing preferences.

- The Netflix Recommendation Engine’s precise recommendations account for 80% of the Netflix viewer activity.

- The NRE has an estimated worth of a billion dollars.

- Clustering plays a significant role in building recommendation engines helping group similar content and similar users together to predict user preferences accordingly.

Problem Statement

In this project we’ll be using the Netflix Content Data to :

1. Understand trends and gain insights on the content listed on Netflix
2. Understand the type of content available in different countries
3. Find if Netflix has been focusing increasingly on TV shows as compared to movies
4. Cluster similar content based on textual features.

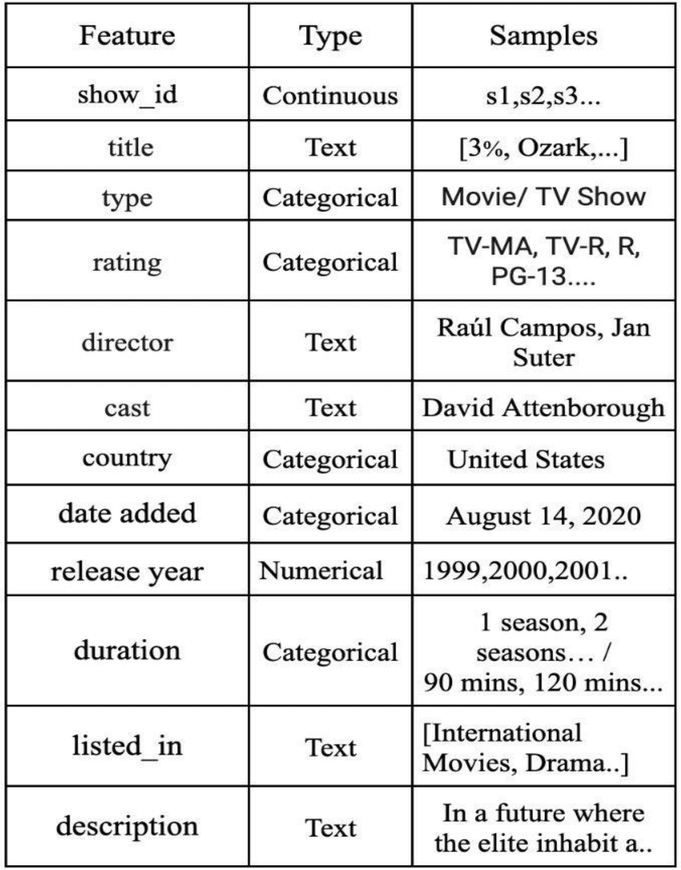
Data Description

- The Netflix Content dataset contains data of 7,787 video content listed on the platform collected from Flixable, a third-party Netflix search engine.

- This dataset consists of 12 attributes.

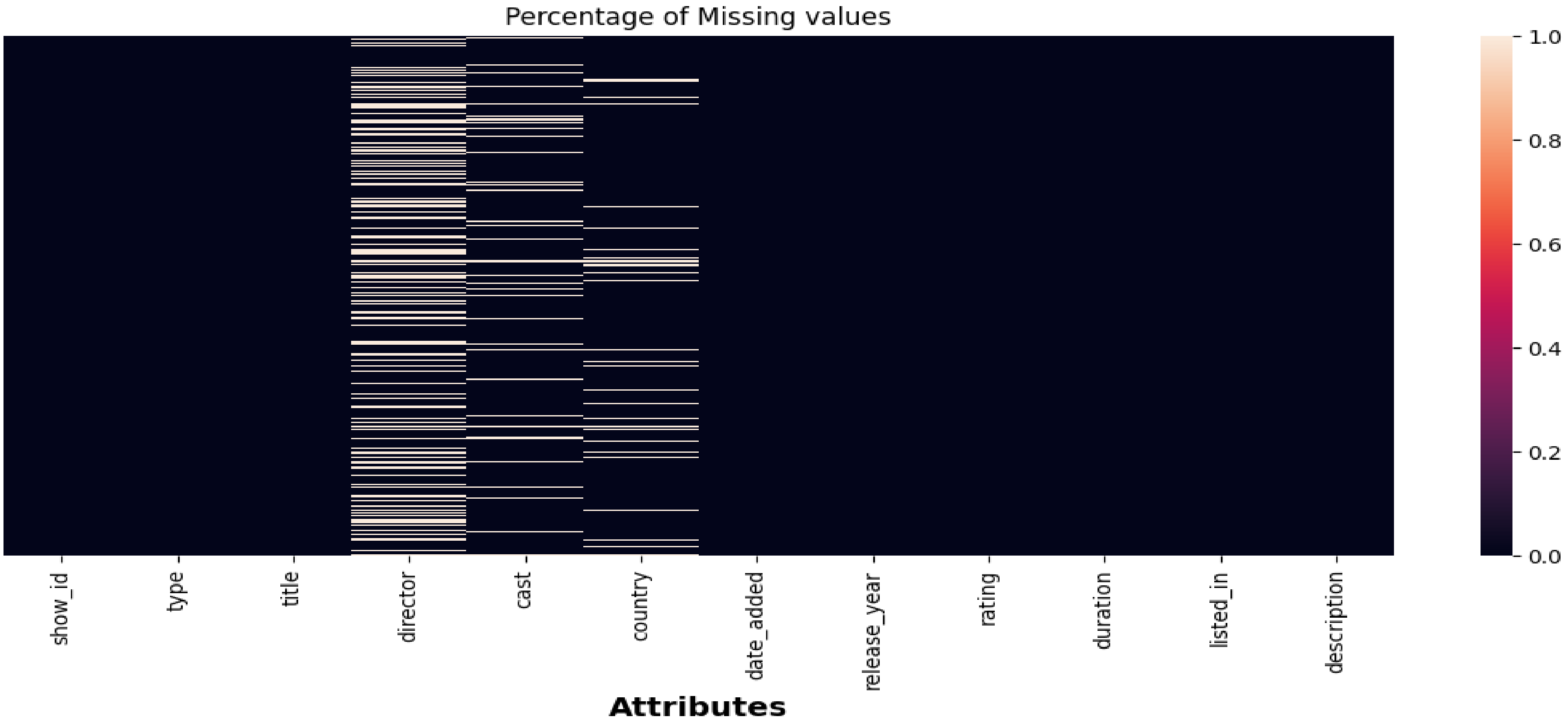
- Attributes providing video details about the video cast, director, duration and countries the content was produced in.

- Attributes also provides site details like signing date, listed description and topics the content is being listed under.



Feature	Type	Samples
show_id	Continuous	s1,s2,s3...
title	Text	[3%, Ozark,...]
type	Categorical	Movie/ TV Show
rating	Categorical	TV-MA, TV-R, R, PG-13....
director	Text	Raúl Campos, Jan Suter
cast	Text	David Attenborough
country	Categorical	United States
date added	Categorical	August 14, 2020
release year	Numerical	1999,2000,2001..
duration	Categorical	1 season, 2 seasons… / 90 mins, 120 mins...
listed_in	Text	[International Movies, Drama..]
description	Text	In a future where the elite inhabit a..

VISUALISING MISSING VALUES

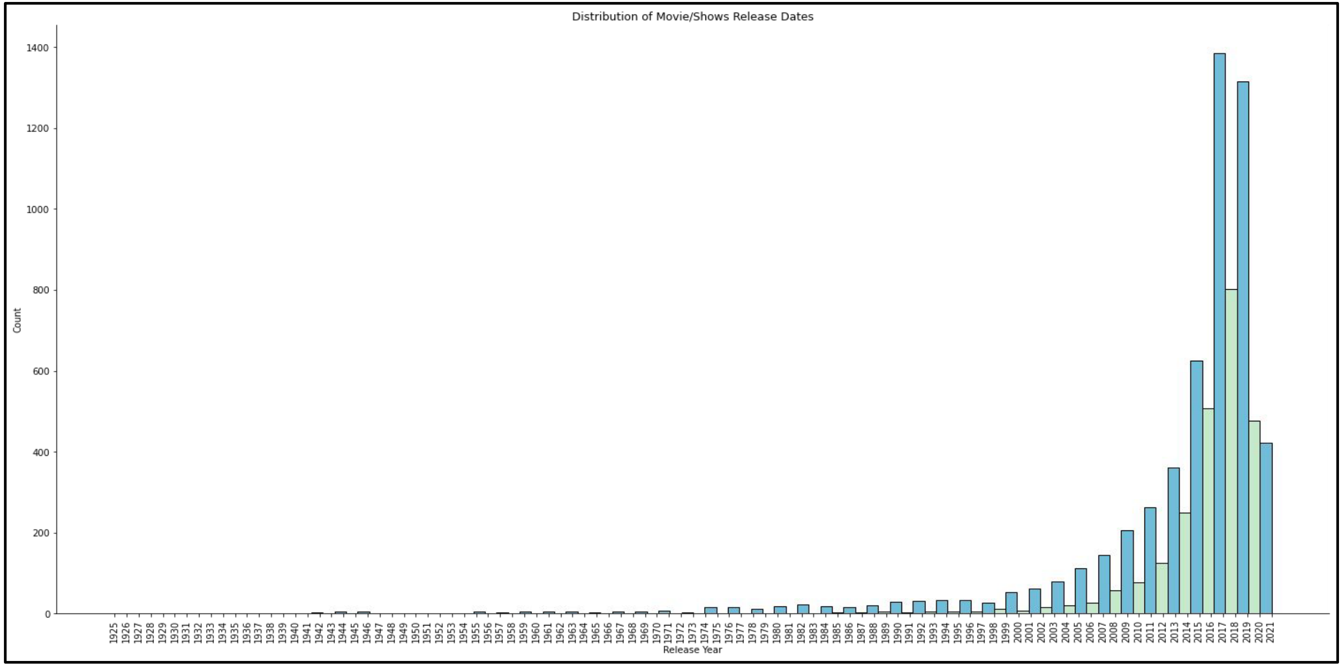


Exploratory Data Analysis

In this part of the project, we inspected and explored:

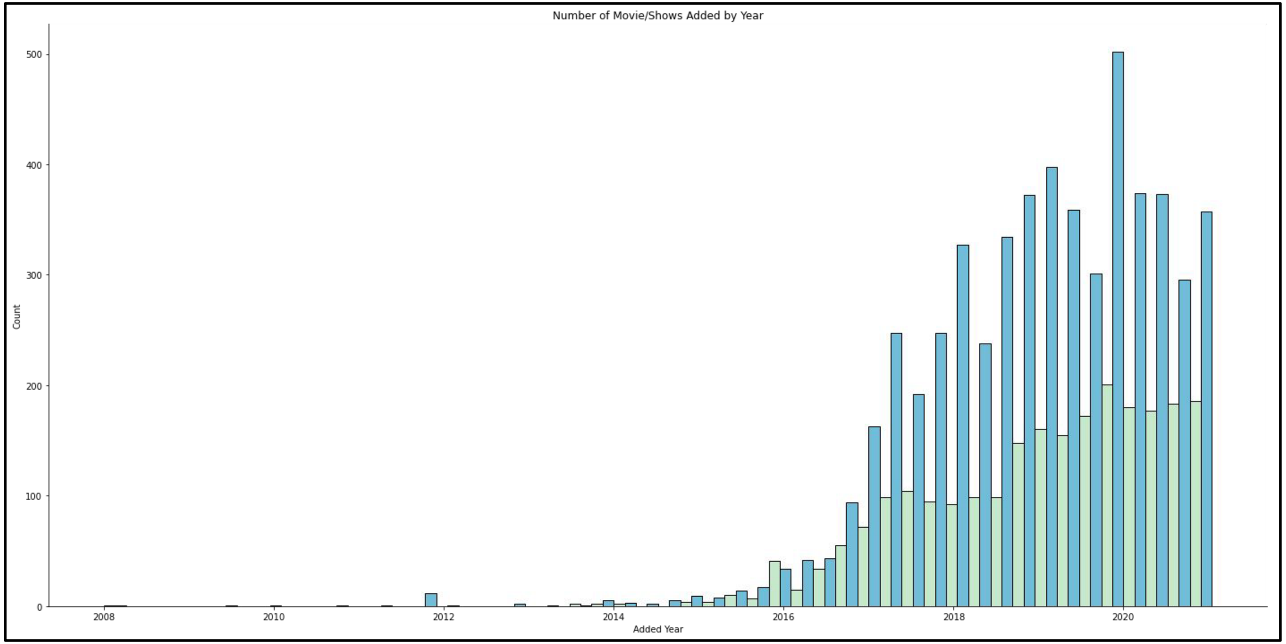
- Timelines of video content signings and releases
- Distribution of Video Content Categories on Netflix
- Type of Content Produced in the top Countries

Release Dates of Movies and TV Shows



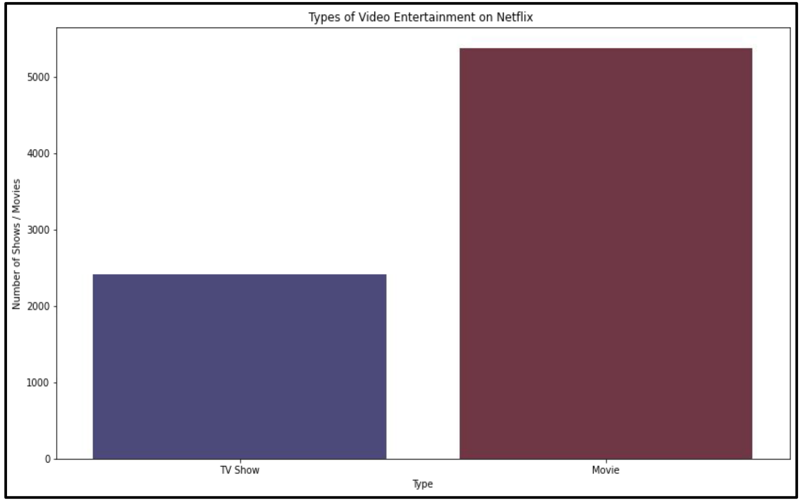
- Large portion of movies streaming on the platform were released after 2010.
- Most TV Shows streaming on the platform was released after 2015.

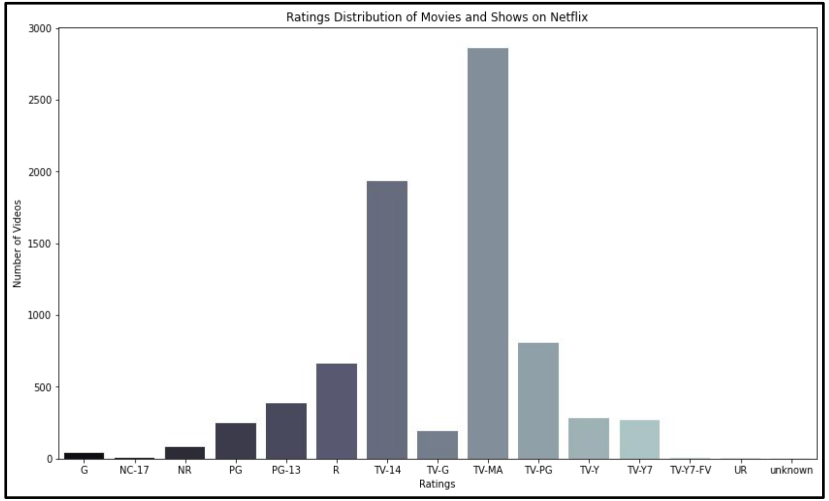
Adding Dates of Movies and TV Shows



- Netflix began adding videos to its platform in 2008.This trend started increasing rapidly from 2017.
- More stand-alone movies were added per year as compared to TV shows.

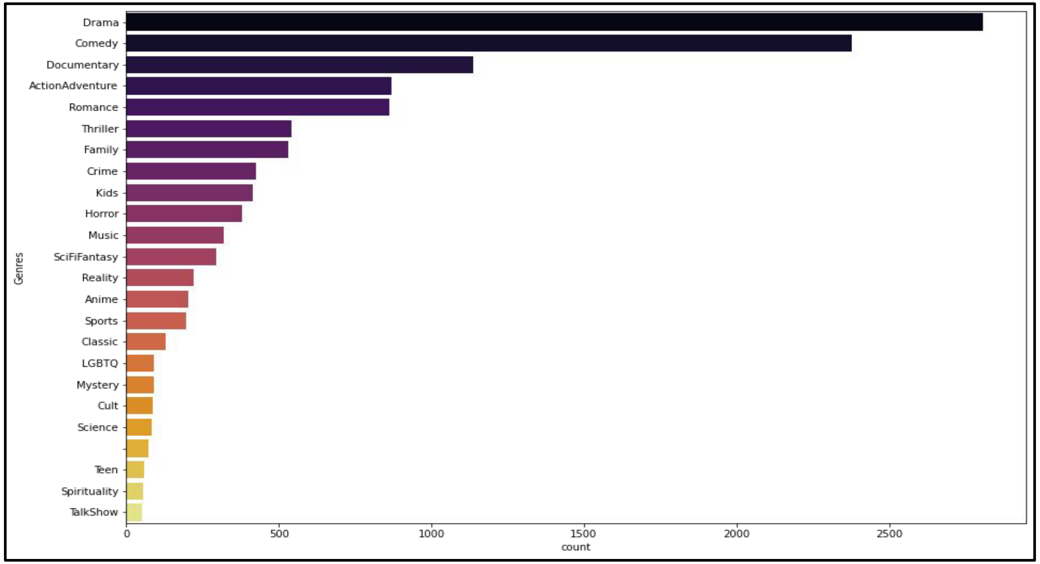
Distribution of Video Type and Ratings on Netflix





- There are almost twice as many movies as TV shows on Netflix
- Majority of the video content is rated for Mature Audiences and for audiences over 14 years old.

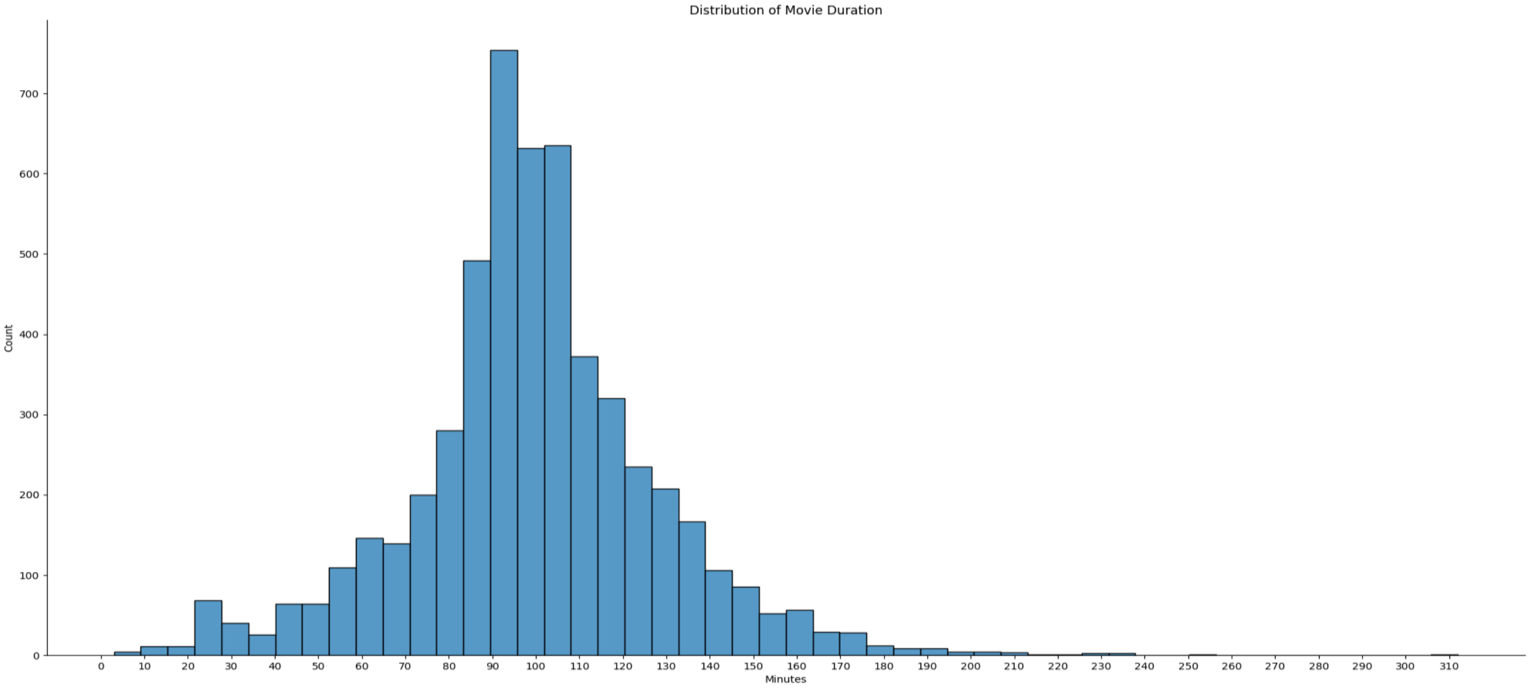
Top Genres on Netflix



It is observed that the top video content genres on Netflix are

- Drama
- Comedy
- Documentary
- Action and Adventure
- Romance

DURATION OF MOVIES ON NETFLIX



- Most movies on Netflix have duration ranging from 90 to 110 minutes

Top Directors on Netflix

Top Directors on Netflix are:

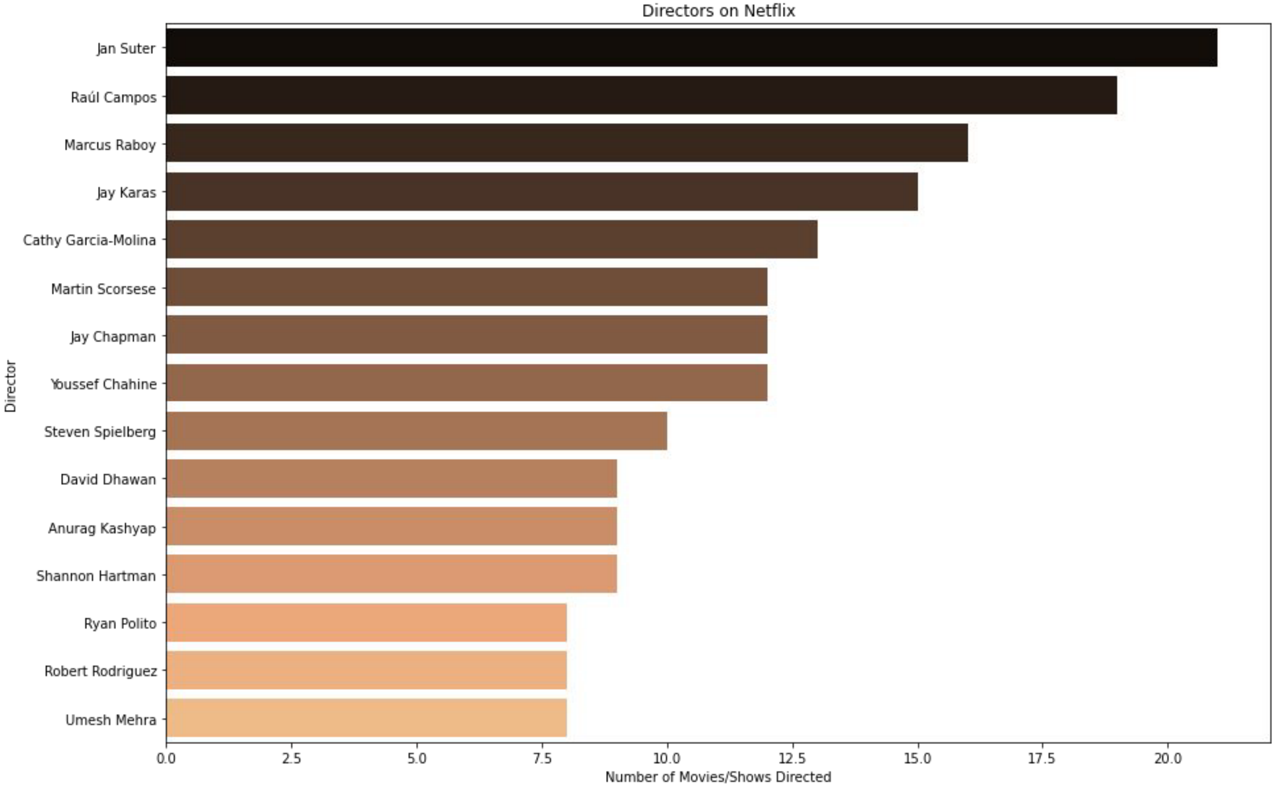
1. Jan Suter

2. Raul Campos

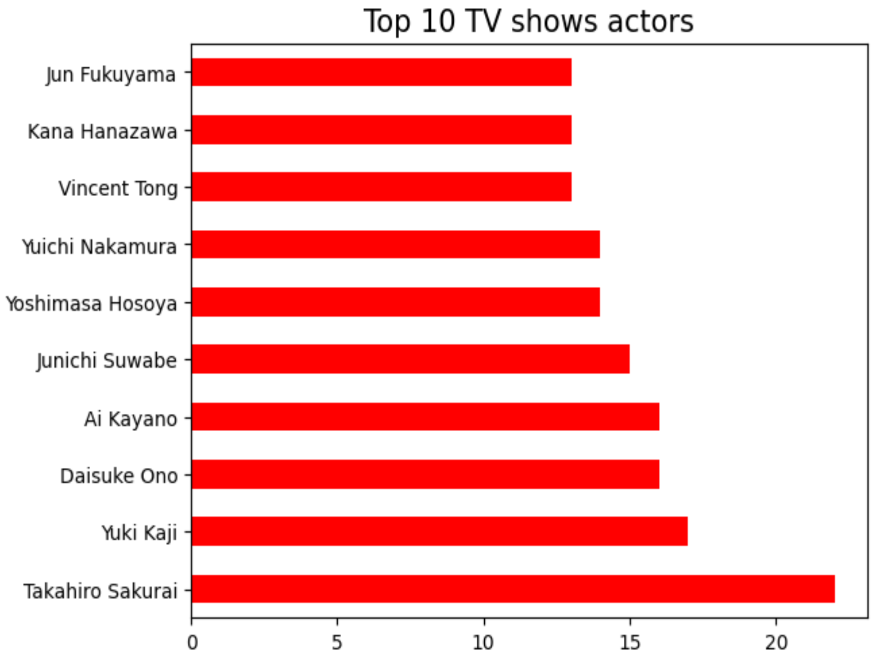
3. Marcus Raboy

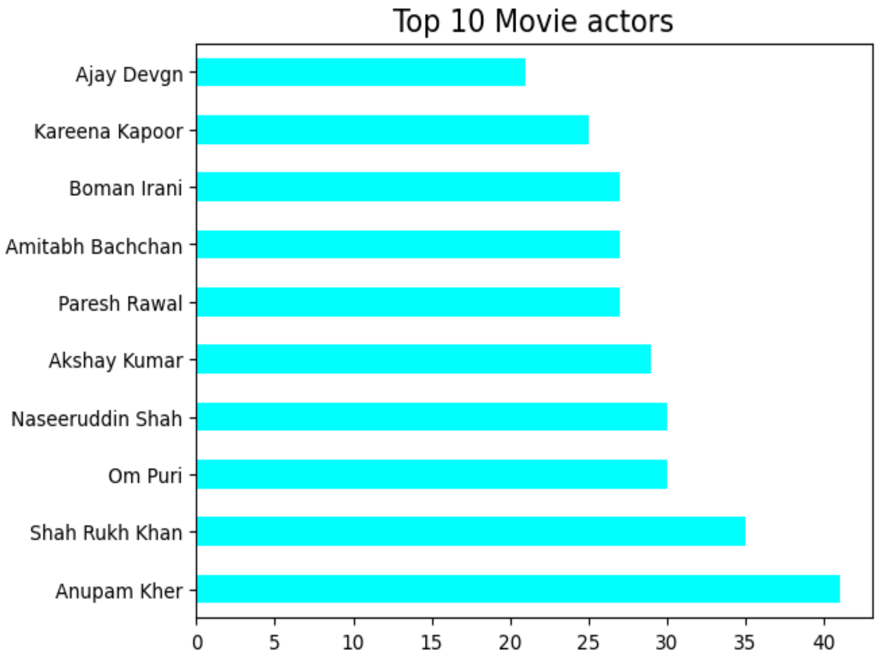
4. Jay Karas

5. Cathy Garcia-Molina



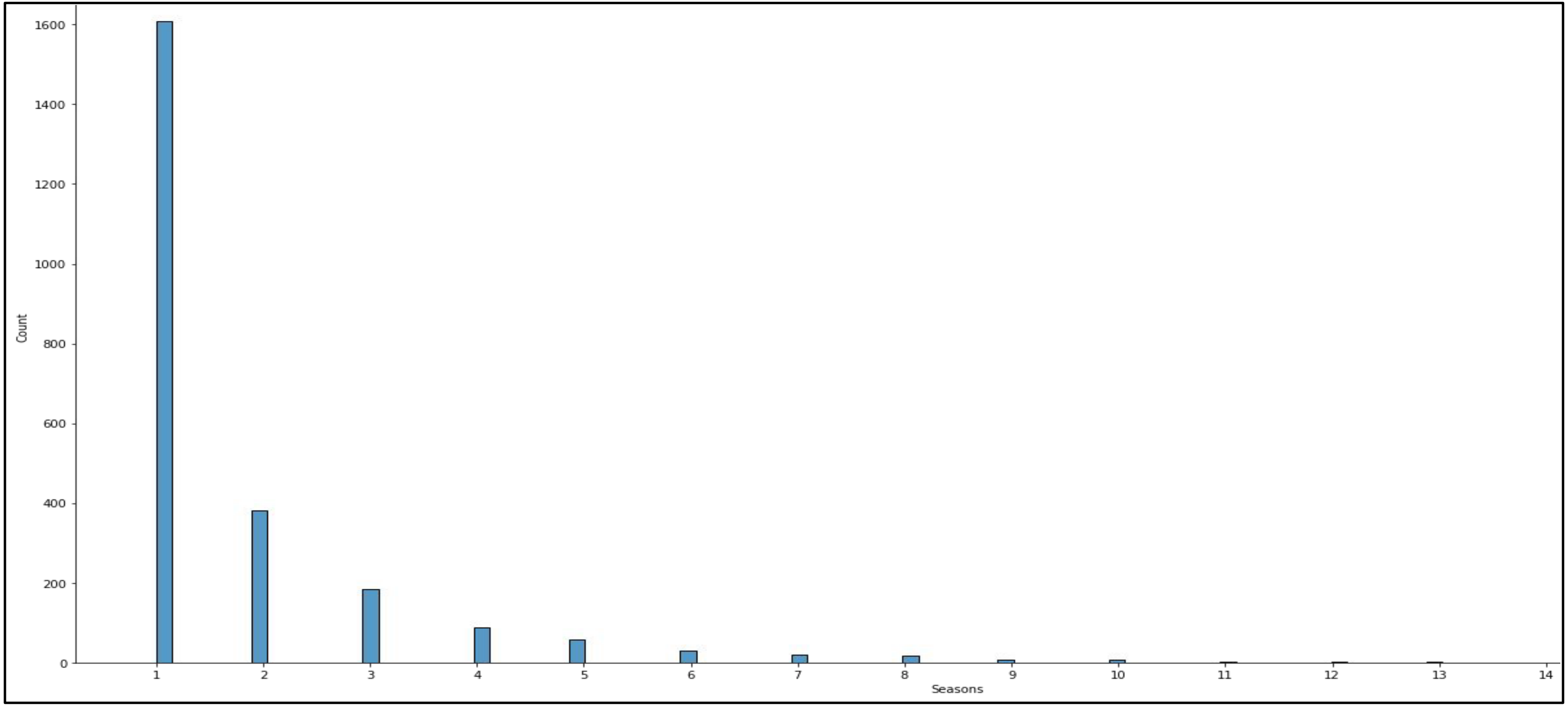
Top Cast on Netflix


Top 10 TV shows actors


Top 10 Movie actors

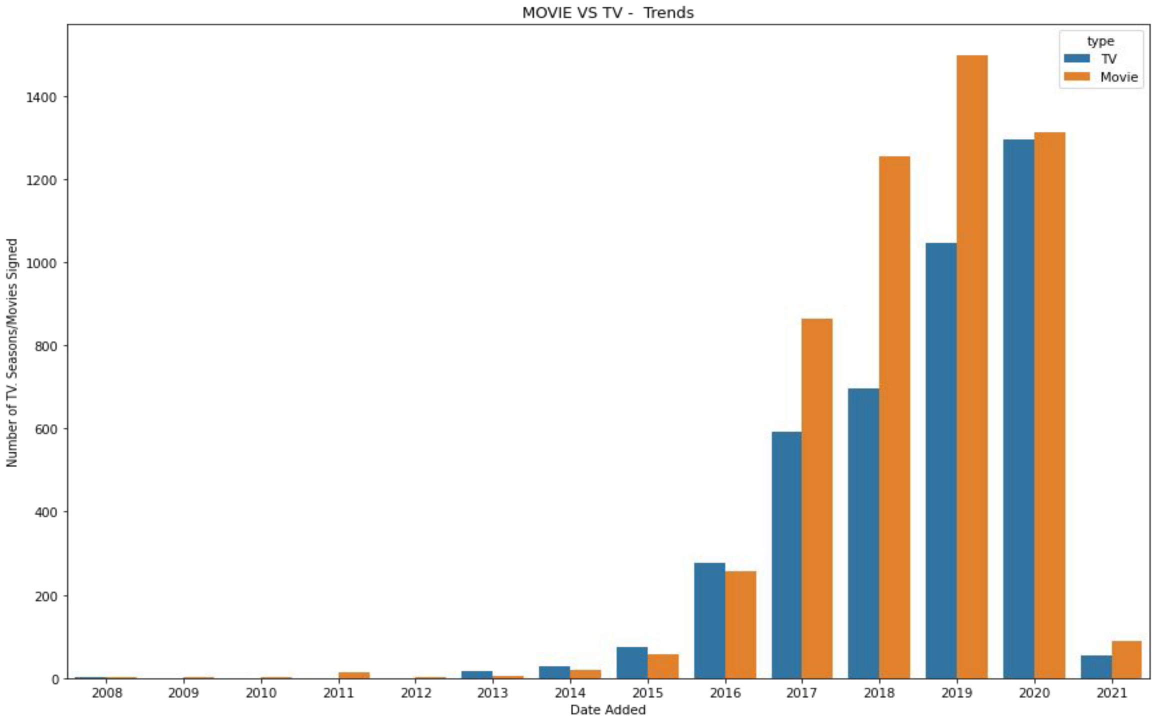
- **Takahiro Sakurai is the actor with the most appearances in the TV programmes category. Anupam Kher is the actor with the highest appearance in the category of films.**

Video Duration Distribution on Netflix



- The tenure of most TV shows on Netflix is only one season.

Adjusted TV Show and Movie Added Plot



- In order to reflect the long-term commitment for TV shows, duplicates of TV shows are made corresponding to their seasons

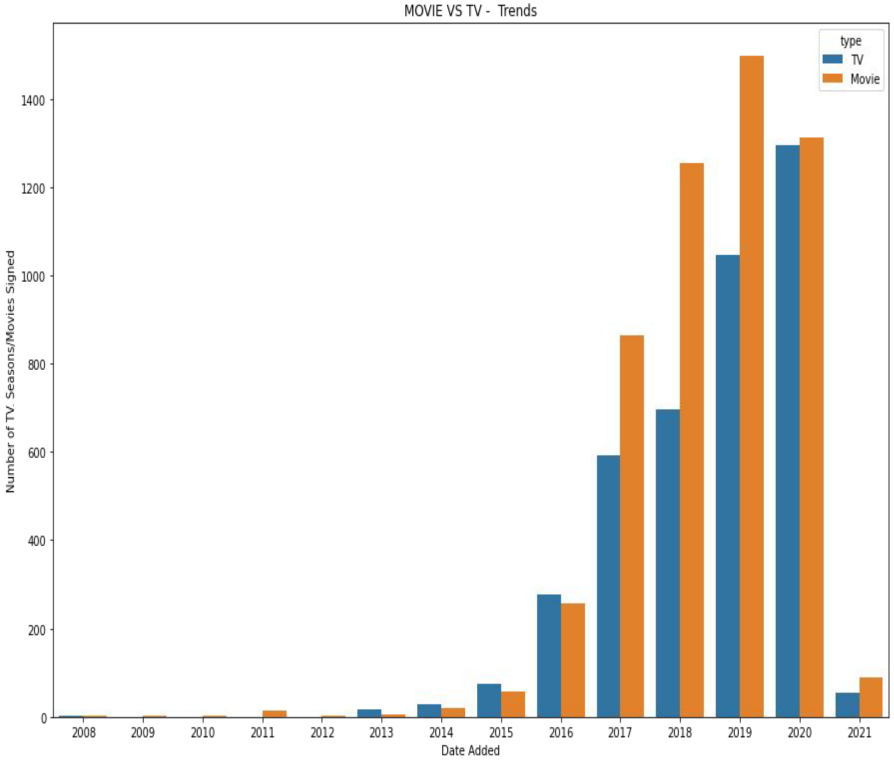
Adjusted TV Show and Movie Added Plot Observations

- We can observe that the TV shows [blue] signed have been higher than the movies[orange] in 2016

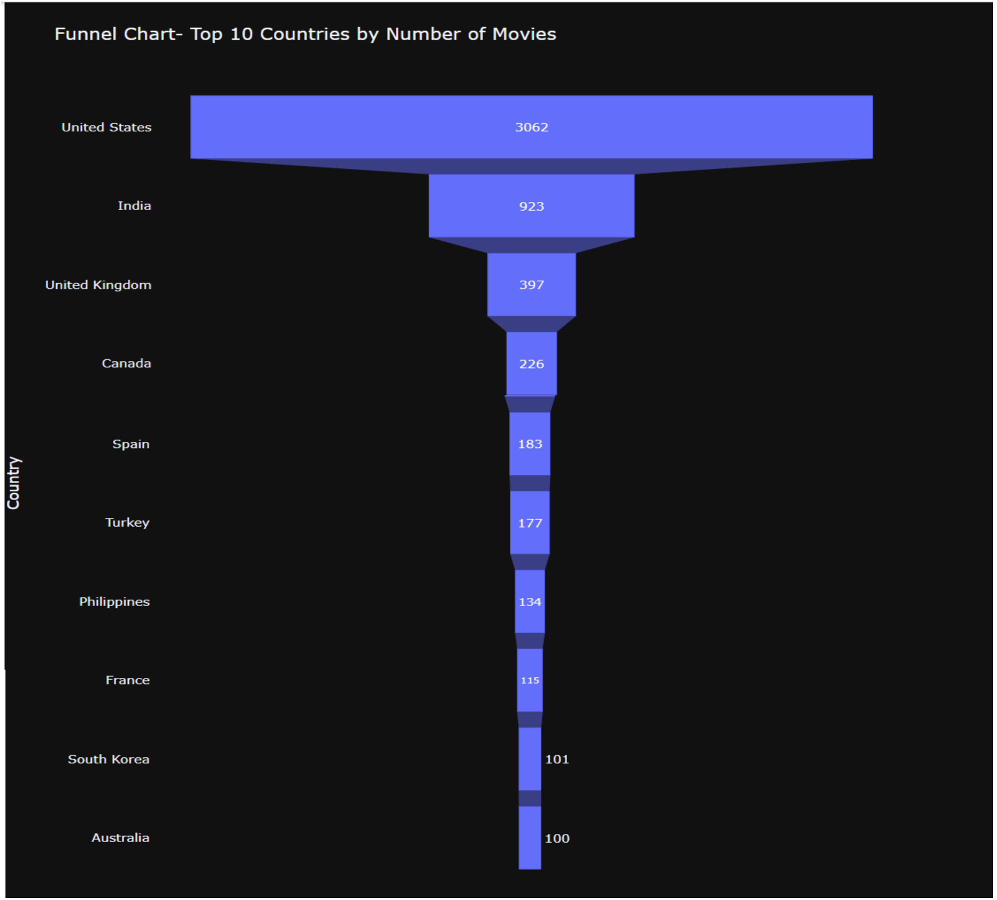
- The movies signed have been higher ever since.

- It can also be observed that TV shows signed annually are catching up to the movies signed per year

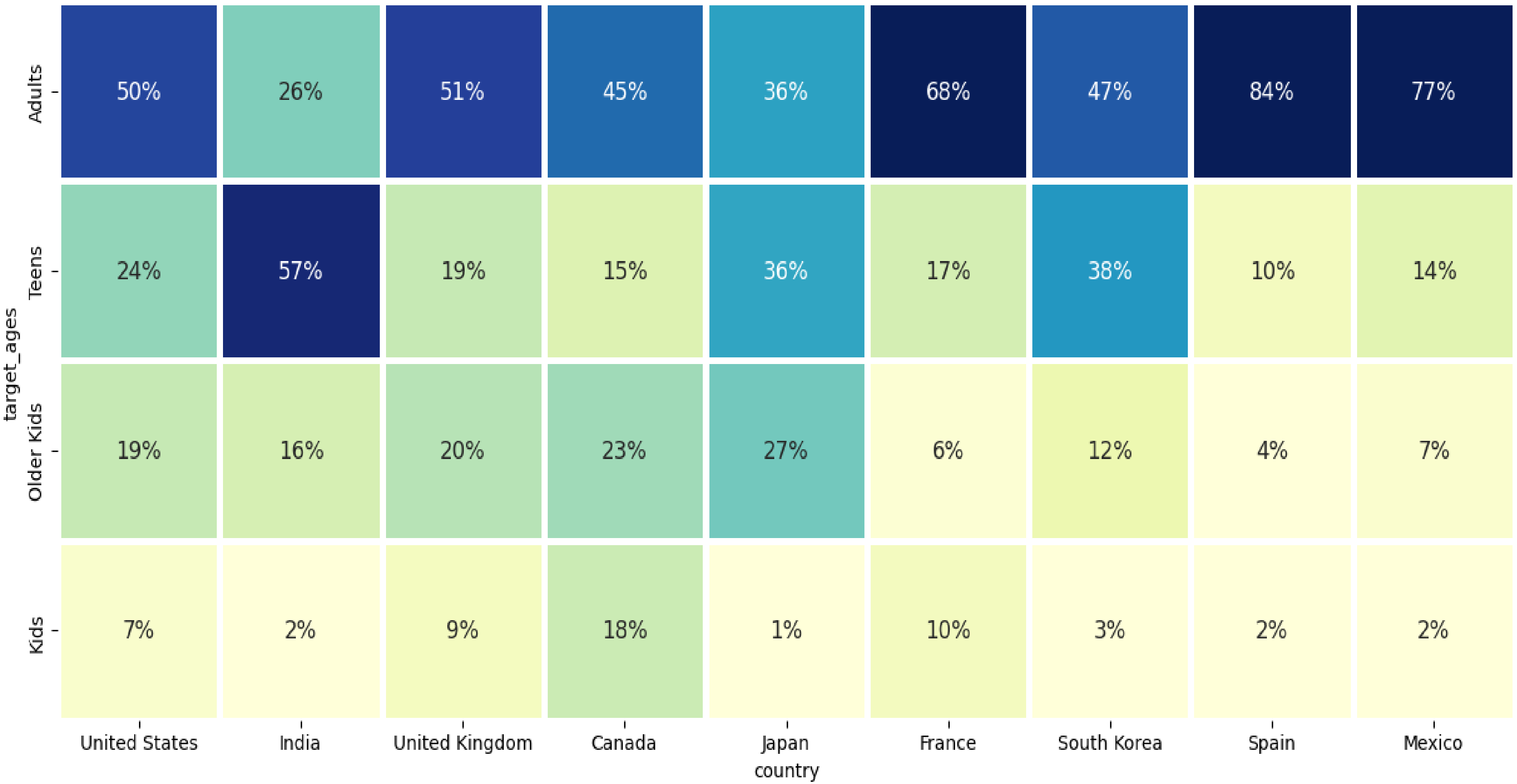
- Hence, we can say that it is true that Netflix has been showing more interest in TV shows as compared to movies.



Top Netflix video content producing Countries



CORRELATION



	United States	India	United Kingdom	Canada	Japan	France	South Korea	Spain	Mexico
Adults	50%	26%	51%	45%	36%	68%	47%	84%	77%
Teens	24%	57%	19%	15%	36%	17%	38%	10%	14%
Older Kids	19%	16%	20%	23%	27%	6%	12%	4%	7%
Kids	7%	2%	9%	18%	1%	10%	3%	2%	2%

CORRELATION CONTD..

In summary, the data provided suggests that the level of interest in the subject varies across different countries and target age groups. Here are the overall conclusions:
Among the countries listed, Spain stands out with the highest percentage of adults showing interest at 84%. This indicates a strong interest in the subject among adults in Spain.

1. France- Follows closely with 68% of adults expressing interest, demonstrating a significant level of engagement in the subject.
2. India- It has the highest percentage of interest among teenagers, with 57% showing interest. This suggests a notable interest among the younger population in India.
3. United Kingdom -It has a relatively high level of interest among adults, with 51% expressing interest.
4. Mexico-Here ,also demonstrates a substantial level of interest, with 77% of adults showing interest in the subject.
5. South Korea , United States- Both have 47% of adults showing interest, indicating a moderate level of engagement in these countries.
6. Japan- It shows a moderate level of interest among both adults and teens, with 36% of each group expressing interest.
7. Canada- It has the lowest percentage of interest among the listed countries, with 45% of adults showing interest.

Overall, these conclusions highlight the varying levels of interest in the subject among different countries and target age groups. The data indicates that Spain, France, India, and Mexico have higher levels of interest in the adults, while Canada has relatively lower interest compared to the other countries

Feature Engineering

- Null values were observed in attributes ‘director’, ‘cast’, ‘rating’ and ‘country’.
- As these values were text-based, the null values were replaced with the label ‘unknown’.
- Attribute ‘released year’ was converted from string to date-time type.
- The year of release was extracted from this feature and was binned by the decade to perform effective clustering.
- The attribute ‘ratings’ contained age-based ratings for Movies and TV shows.
- These movie and TV show ratings were merged based on age using the maturity rating guidelines provided by Netflix and Amazon.

Feature Engineering (contd.)

- The attribute ‘listed_in’ provided genres for TV shows and movies separately.
- The common genres from both content types were combined and the non-genres like ‘International Movie’ and ‘Independent Film’ were removed.
- Non-plot-related text attributes like director name, lead cast and country of production were merged with the genres they were listed under into a single text.
- This text was treated as an attribute providing text insight for clustering in the future

Feature Engineering: Topic Modelling - Intuition

- Alternatively, it was decided that the two attributes should be used to model video content into topics using Latent Dirichlet Allocation and used as inputs for clustering.

- This would make sure that all the topical information about video content is captured without putting any available information to waste.

- Topic modelling also entertains the possibility of a video exhibiting multiple themes at different extents by expressing the probability of a document belonging to a given topic.

.

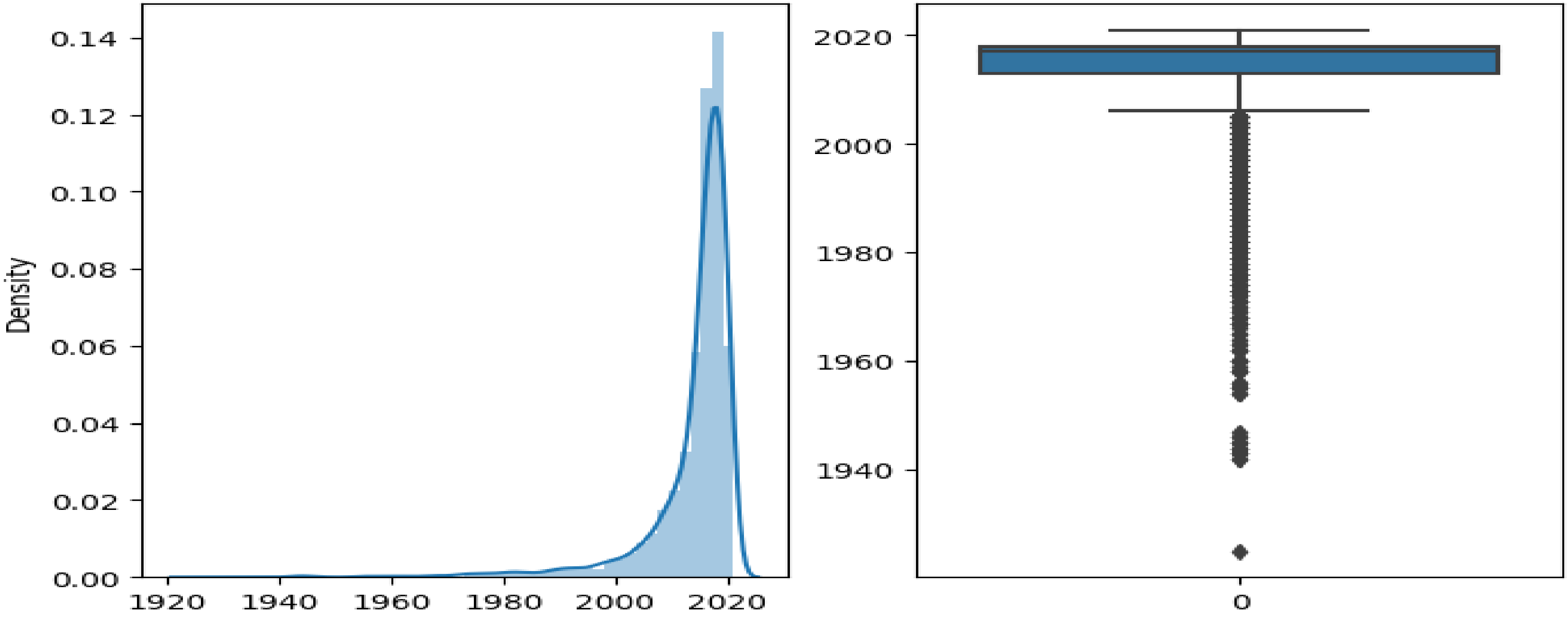
Feature Selection

- Relevant non-text attributes describing the content’s maturity ratings, duration, year of release and type of content are taken.

- The attributes exempted are ‘show id’, ‘title’ and ‘added date’ as they add little to no substance in the qualitative and quantitative characterization of the video itself.

- To feed in information about the video content’s plot, directors, cast and genres, we will be using the preprocessed and topic modelled version of text attributes ‘description’ and ‘Movie Deets’.

HANDLING OUTLIERS



- ❖ **The figure (release_year less than 2009) are being displayed as outliers.**

Text Processing

The steps involved in text preprocessing are :

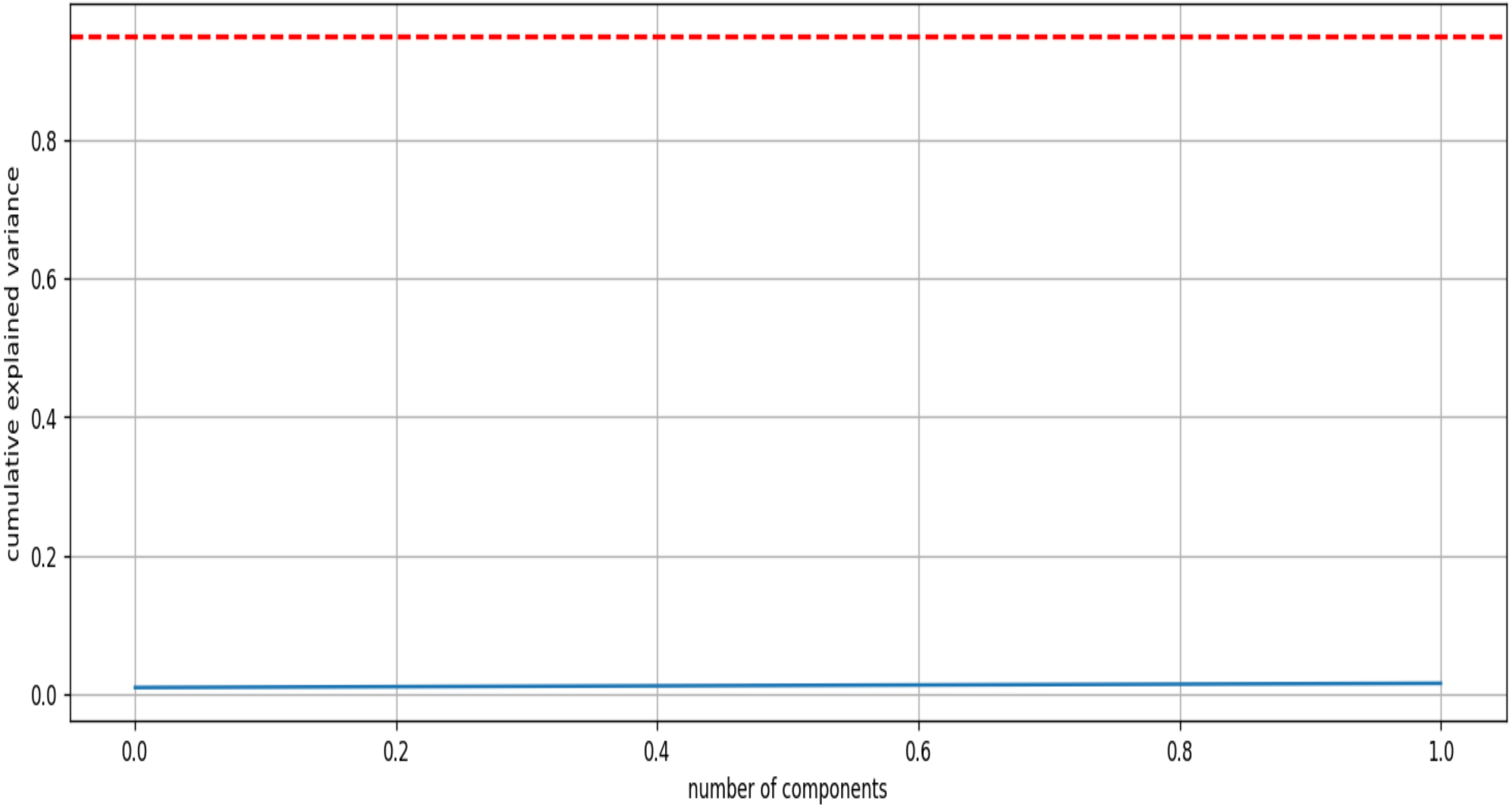
- **Tokenization:** Involves breaking of natural language text into chunks of information that can be considered as discrete elements. The token occurrences in a document can be used directly as a vector representing that document.

- **Punctuation Removal:** All the punctuations from the text are removed.

- **Stopword Removal:** Common words that add very little or no significant insight to the text being processed are removed beforehand. This reduces time and computational complexity.

- **Stemming Words:** Stemming is the process of reducing inflected words to their word stem, base or root form—generally a written word form. This reduces different forms of the same word carrying the same base meaning. It should be noted that stemming does not remove synonyms.

DIMENSIONALITY REDUCTION BY TRUNCATEDSVD METHOD



Methods to Choose Optimal Cluster Numbers

- **Elbow Method:** The elbow method plots the value of the cost function produced by different values of clusters, k, in K-means clustering.

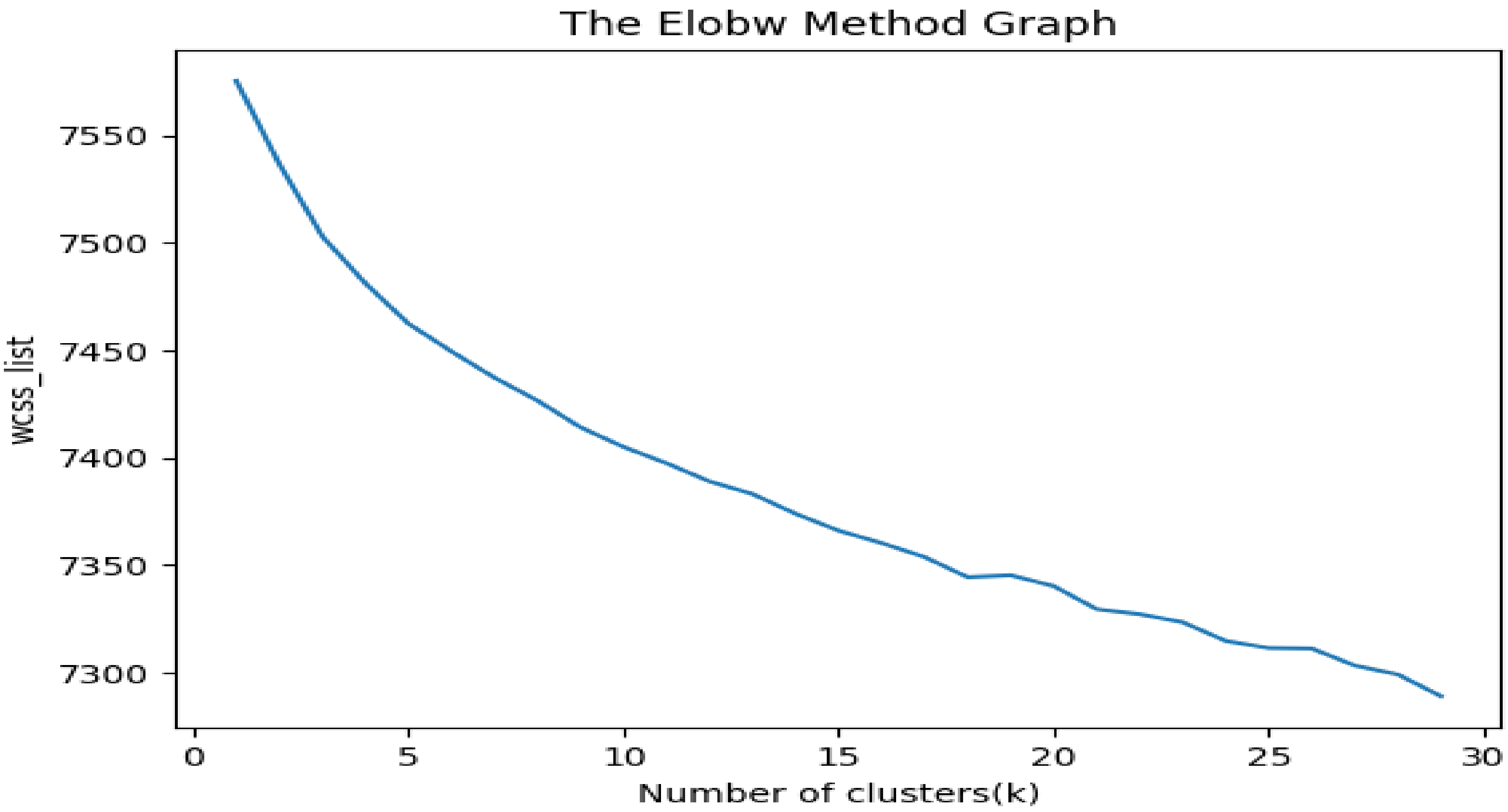
- The value of k at which improvement in distortion declines the most is called the elbow, at which we should stop dividing the data into further clusters.

- **Dendrogram Method:** Dendrograms are a diagrammatic representation of the hierarchical relationship between the data points.

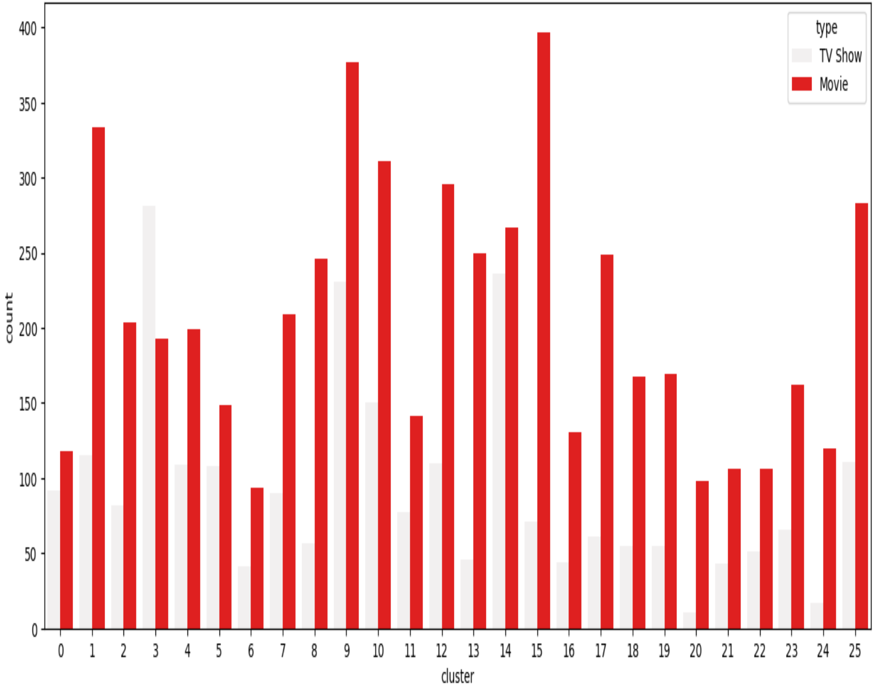
- These are used to observe the output of hierarchical agglomerative clustering.

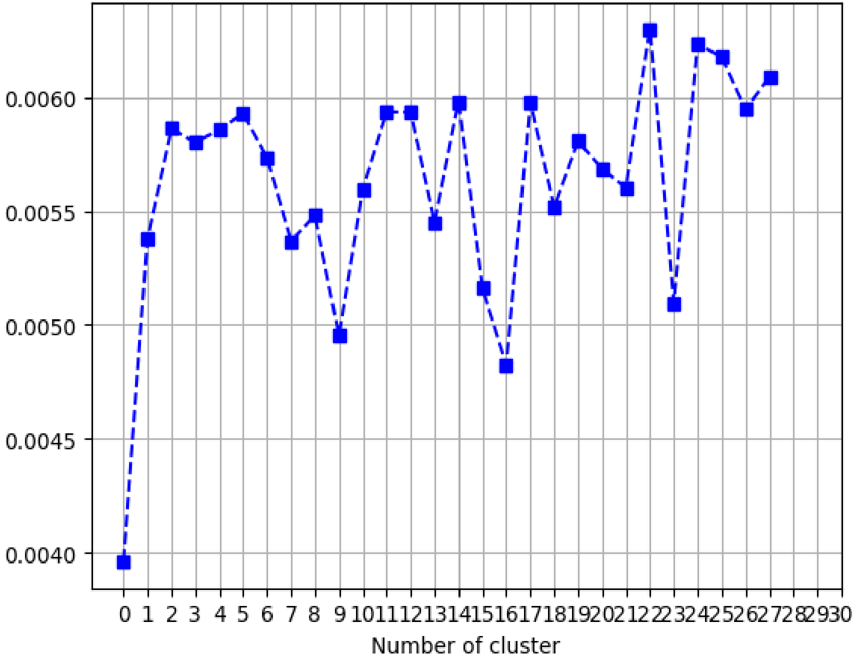
- The number of clusters is determined by slicing the dendrogram horizontally. All the resulting child branches formed below the horizontal cut represent an individual cluster at the highest level in the system

THE ELBOW METHOD GRAPH



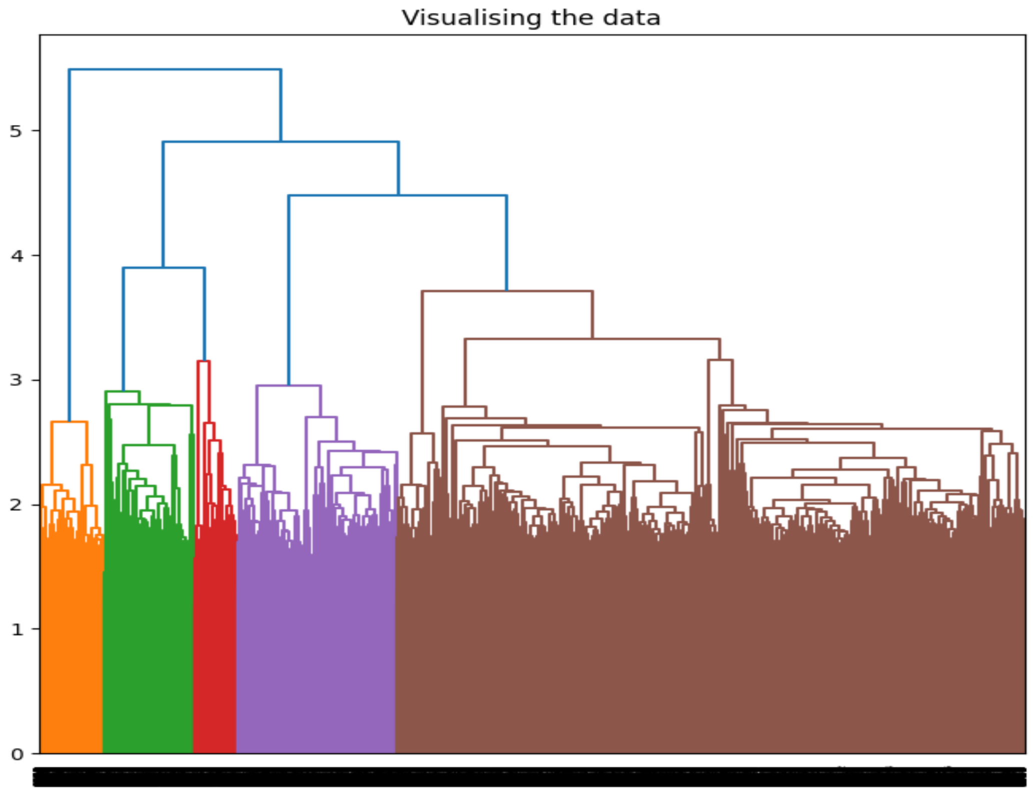
Plots to Choose Optimal Clustering





Hierarchical Cluster Visualization DENDOGRAM

A dendrogram is a diagram that depicts the relationship between things in terms of hierarchy. It is frequently produced as a byproduct of hierarchical clustering. A dendrogram is mostly used to determine how to assign objects to clusters.



K - Means

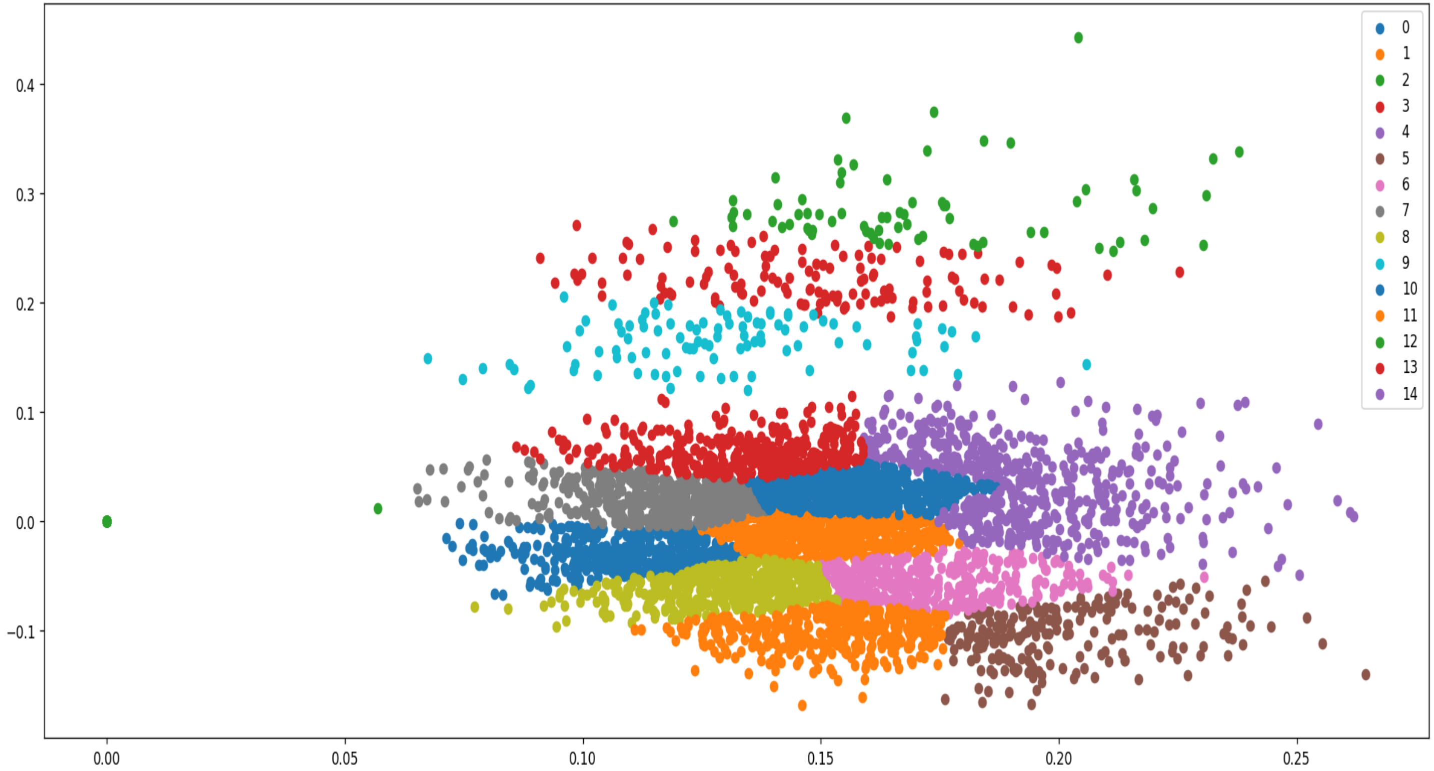
- It is an iterative algorithm that divides the unlabeled dataset into K unique clusters where each dataset belongs to only one group having similar properties.

- This algorithm aims to minimise the sum of distances between the data point and their corresponding clusters.

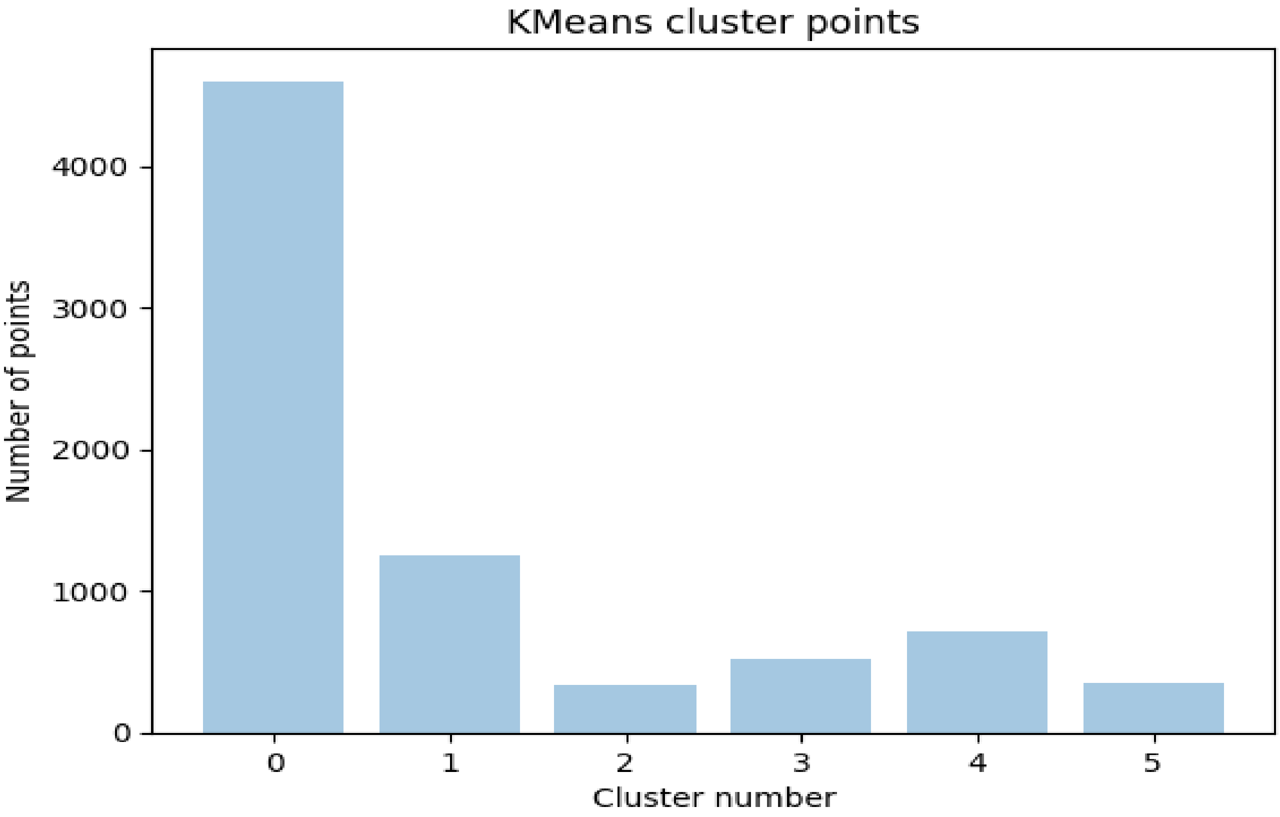
- The algorithm takes the unlabeled dataset as input, divides the dataset into K number of clusters, and repeats the process until it can’t find better clusters.

- For K-Means clustering the elbow and optimal silhouette score were found at 8 clusters with a silhouette score of 0.4686, Davies-Bouldin Index of 0.887 and Calinski-Harbaz Score of 2901.84.

K MEANS CLUSTER

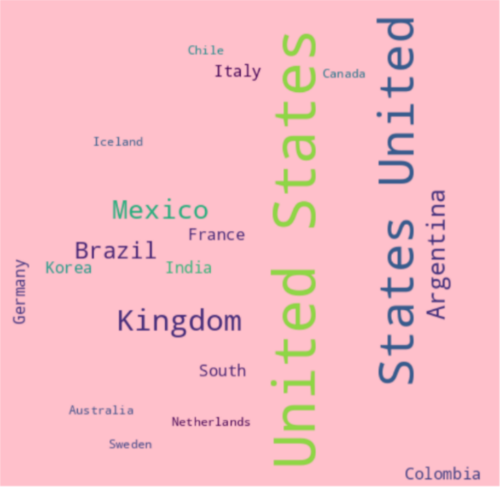


K-Means Cluster Visualization



BAG OF WORDS

MA
TV
G
TV
PG
MA
NR
TV
TV

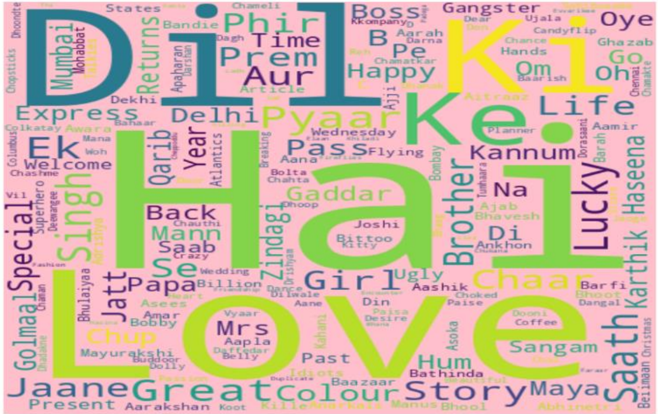

Chile
Italy
Canada
Iceland
Mexico
France
Brazil
Korea
India
Germany
Kingdom
South
United States
States United
Argentina
Australia
Netherlands
Sweden
Colombia

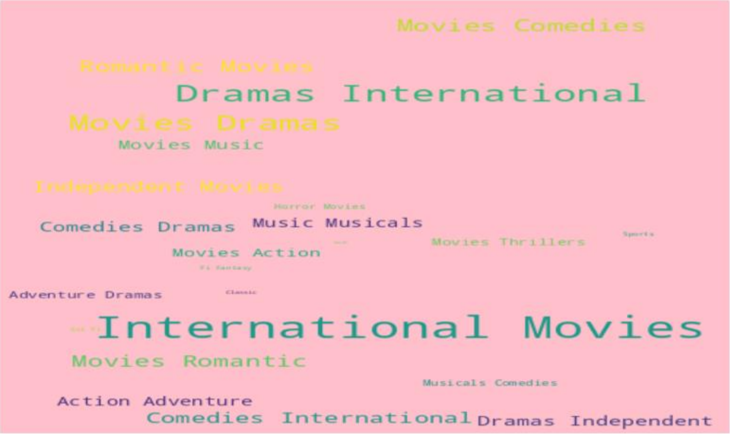
Birbiglia
king
Cars
Amy
Amani
Comedy
Mittal
Padilha
Pain
Peters
Sam
Ali
Larry
Let
One
Tim
Gadsby
Jani
Inferior
Dunham
Sorry
Whitehall
Minchin
Caldi
Wong
Kinison
Soy
Dany
Brian
Sanint
Arnold
Rada
Tom
Peretti
de
Bill
Gone
Getting
Father
Rancour
Temi
Martin
Best
Edoardo
Mother
Business
Ríase
amor
Batra
Icelandic
Hard
Gad
Classless
Bienvenido
Ricardo
Jack
Escamilla
Rossello
Chris
Carlos
y
Guy
Way
Wife
Berlin
Notas
Disculpe
show
World
Special
Night
Stand
Coffee
Fire
Arango
Schumer
Abrazo
Alex
Carey
Teeuwen
Duenas
Hungry
Dana
Elder
Hicks
Gabriel
O'Farrill
Enissa
Russell
Agustin
Hannah
Hauts
Fernández
Afonso
Elefanten
White
el
Knock
Jeff
Live
Say
Mike
Furia
Rodeo
Iliza
Aristarán
D'Elia
Cable
Shlesinger
Christmas
Douglas
Burr
Greats
Demetri
Brad
Ari
Said
Comedian
Nuhr
Segura
Ballarta
David
Black
Christina
Relentless
Garfunkel
Growing
Fernando
Paisley
Pardon
Eldjárn
Ferrario
Feel
Trying

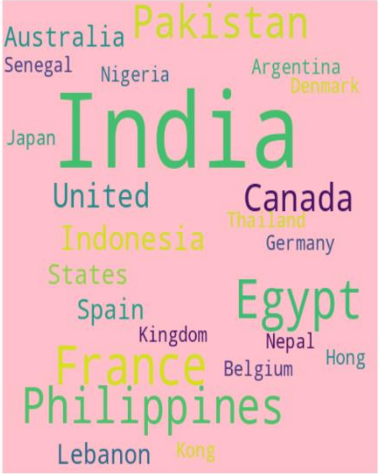
Documentaries
Comedies
Children
Stand
Shows
Movies
Comedy
Stand
Music
Musicals
Comedy
Family
TV
Sports
Talk

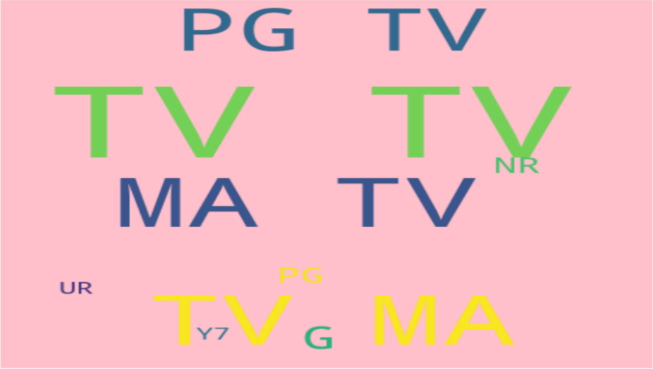
Russell
Brazilian
traditions
awkward
may
performs
silly
comedic
stage
man
show
find
explore
old
Larry
O'Farrill
American
documentary
End
late
bring
relationship
touches
men
celebrity
dives
women
media
details
age
signature
explaining
gender
truths
topics
share
comic
race
Brooklyn
America
escapes
Dunham
fun
traveled
special
German
love
audience
personal
parenthood
gets
dolphins
Two
even
big
modern
humor
take
struggles
featuring
live
sex
uses
make
parenting
star
journey
reality
say
style
raunchy
songs
stories
let
power
tale
funny
Bill
differences
Christmas
language
value
night
things
Jeff
Deadpan
talk
time
series
raising
evening
life
online
set
comedy
entertainer
concert
variety
deeper
Seattle
history
thoughts
reason
marriage
political
woman
hit
Tim
music
Italian
leader
laughs
one
Argentine
road
children
comedy special
hilarious
performance
Minneapolis
French
Ricardo
come
digs
pokes
tour
sport
new
jokes
comedian
offer
childhood
self
yet
daily
including
issues
storm
crowd
baby
ventriloquist
talent
kids
social
public
host
everything
highlights
regales
reflections
country
father
favorite
three
Birbiglia

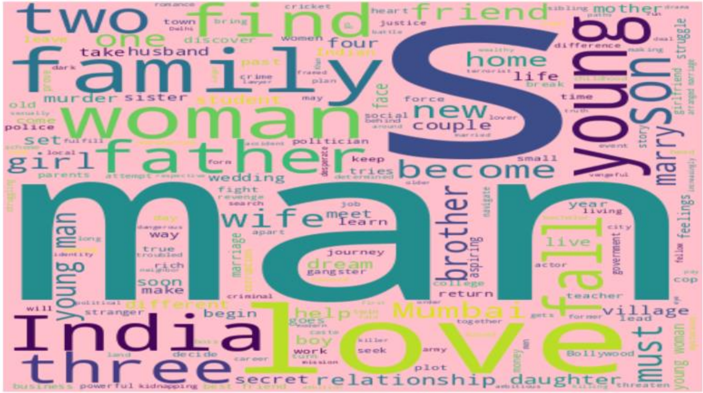
BAG OF WORDS CONTD…











Most Relevant words for modelled Topics

Cluster 9 in a dataset contains a total of 232 words. The most frequently occurring words in this cluster are as follows:
Type - Movie & Tv shows
Title - Broadway, Remastered, Christmas, Friends Orchestra
Country- United Kingdom, Argentina, United States,India
Rating -TV-MA, PG-TV
Listed_in - Dramas International, Musical Dramas, Musicial Documentaries, Comedies International
Description- Documentary, Music, One, Bad, Tour, Love.

Cluster 11 in a dataset contains a total of 410 words. The most frequently occurring words in this cluster are as follows:
Type - Movie & Tv shows
Title - Special, America,Time,Live,Comedy, Netflix Alive, Martin
Country - United States,Brazil,Mexico,Italy
Rating -TV-MA,TV-PG
Listed_in - Tv-Comedies, Comedy Stand, Talk shows
Description- Stand Comedy, Comic, Take, Life, Live, Share,Stories.

Conclusions

EDA

Release dates of shows/movies on Netflix

- **Most Movies streaming on the platform were released after 2010.**
- **Most TV Shows streaming on the platform were released after 2015.**
- **The year 2017 had highest number of Movie and TV show releases on the platform.**
- **Number of shows/movies added by the Streaming giant**
- **Netflix began adding videos to the platform from 2008**
- **The streaming giant started aggressively adding movies and TV shows from 2017**
- **More movies are added as compared to TV shows**
- **Type of content watched on Netflix.**
- **TV shows constitute the majority, accounting for 69.1% of the content watched on Netflix, while movies make up a smaller percentage of 30.9%.**
- **Type of Videos on Netflix**
- **There are almost twice as many movies as TV shows on Netflix.**
- **Content added over the year**
- **The trend in the visualization indicates that between 2008 and 2022, there were relatively fewer TV shows and movies added to Netflix. However, starting from 2016, there was a slight increase in content additions. In 2019, there was a significant peak in the number of movies added, while TV shows experienced a similar trend but with a lesser increase compared to movies.**

Top Genres in Netflix are:

1. **Drama**
2. **Comedy**
3. **Documentary**
4. **Action and Adventure**
5. **Romance**

Conclusions (contd.)

Has Netflix Been Focusing Increasingly on TV Shows as compared to movies

- The above graph depicts seasons of TV shows signed vs the movies signed
- This distinction gives contacts as TV shows require recurring investment for each seasons. So the TV numbers have been increased in accordance to the seasons. As they were considered as one entity earlier
- We can observe that TV shows signed have been higher than movies in 2016. While the the movies signed have been higher, it is blatantly visible that the TV shows signed per year is catching up to the movies signed by the year
- Top 10 countries for movies
- The United States has the highest number of movies, with 3062 films, indicating a dominant presence in the film industry. India is the second-highest contributor with 923 movies, demonstrating a significant presence in the global Movies/Tv shows market.

CONCLUSION FOR ML

- It's noteworthy to notice that films make up the majority of the content offered by Netflix. However, the platform has been concentrating more on TV shows in recent years.
- The majority of these shows debut at the end of the year or the beginning.
- Among the top five nations that create all of the content that is made available on the site are the United States and India. Furthermore, six of the top ten actors with the most content hail from India.
- According to content ratings, TV-MA is at the top of the list, showing that Netflix users prefer mature fare.
- Since k=15 was discovered to be the best value for clustering the data, it was utilised to divide the content into ten distinct clusters.
- Cosine similarity was used to develop a content-based recommender system using this data, which offered suggestions for films and TV series.

Thank You