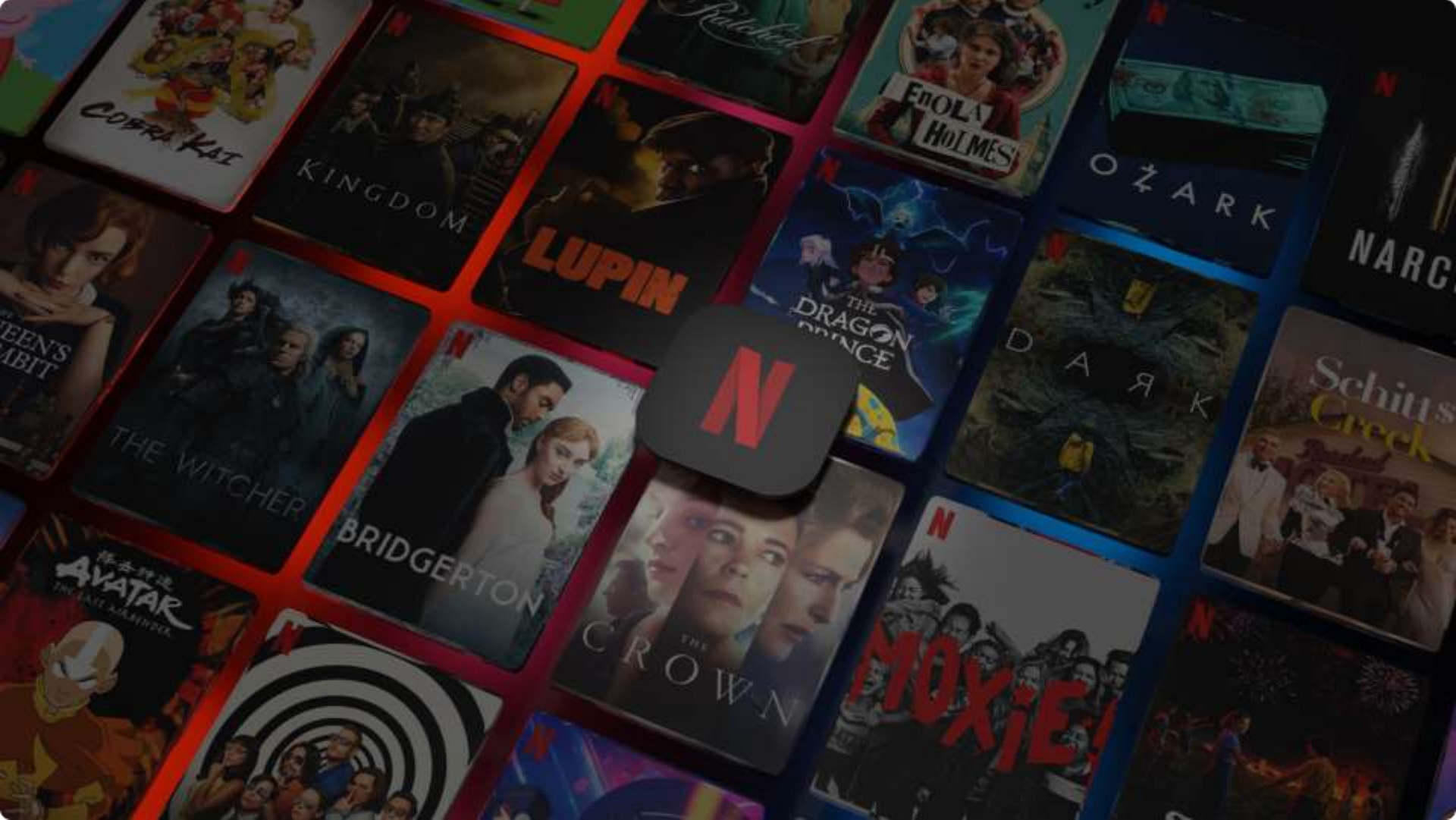

COBRA KAI
KINGDOM
LUPIN
ENOLA HOLMES
OZARK
THE WITCHER
THE DRAGON PRINCE
DARK
Schitt's Creek
AVATAR
BRIDGERTON
THE CROWN
MOXIE

# Capstone Project

## Netflix Movies and TV Shows Clustering

Team member
Mayur S. Marathe
Sanket S. Gawali
Yash

# Content

- Introduction
- Data Summary
- Problem Statement
- Data Description
- EDA
- Feature Engineering
- Text Processing
- Topic Modelling
- Feature Selection
- 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.

# Data Summary

- **show_id :** Unique ID for every Movie / Tv Show
- **type :** 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 :** Generes
- **description:** The Summary description

# Problem Statement


This dataset consists of tv shows and movies available on Netflix as of 2019. The dataset is collected from Flexible 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.

# Data Description

- The Netflix Content dataset contains data of 7,787 video content listed on the platform collected from Flexible, 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.

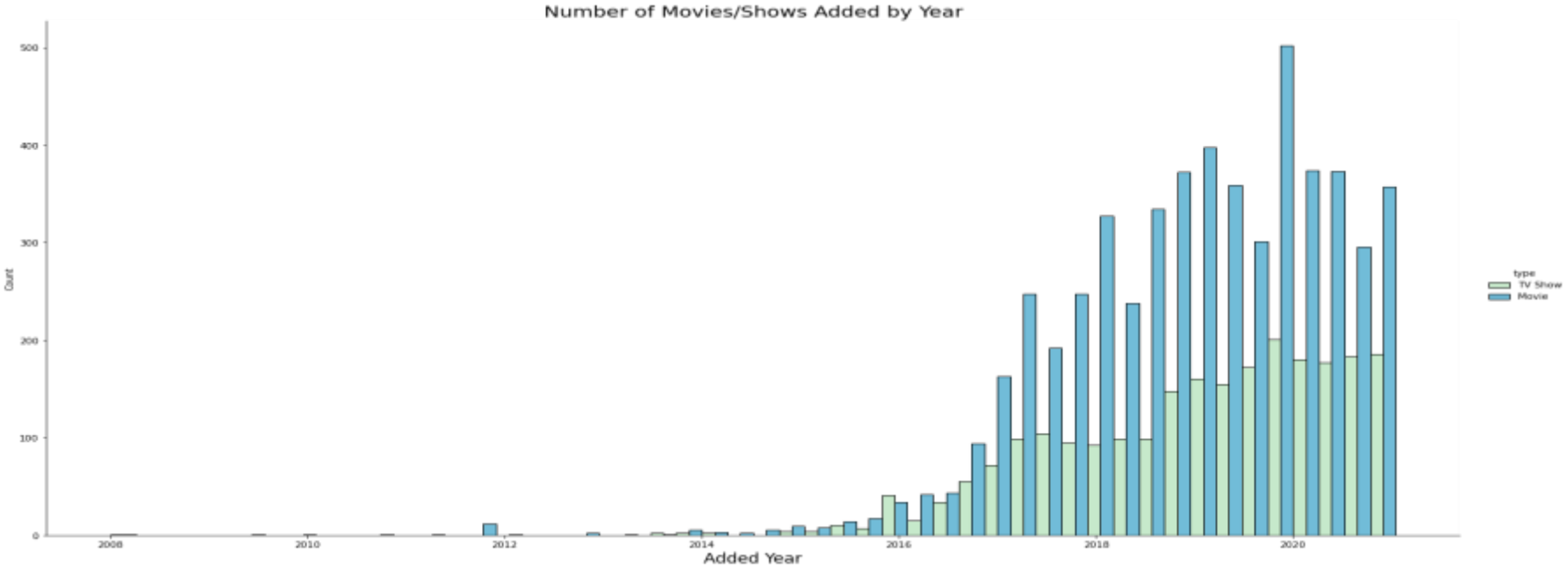
- Attributes also provides site details like signing date, listed description and topics the content is being listed under.

# 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

# Adding Dates of Movies and TV Shows

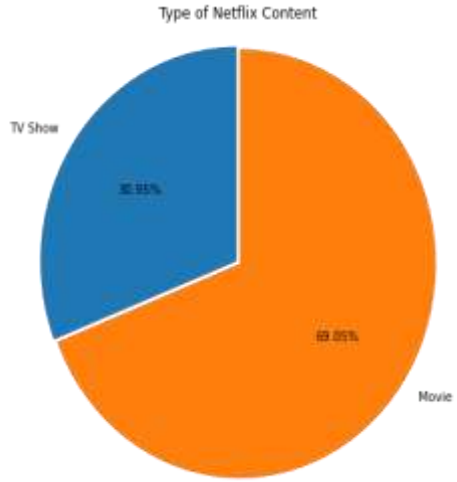


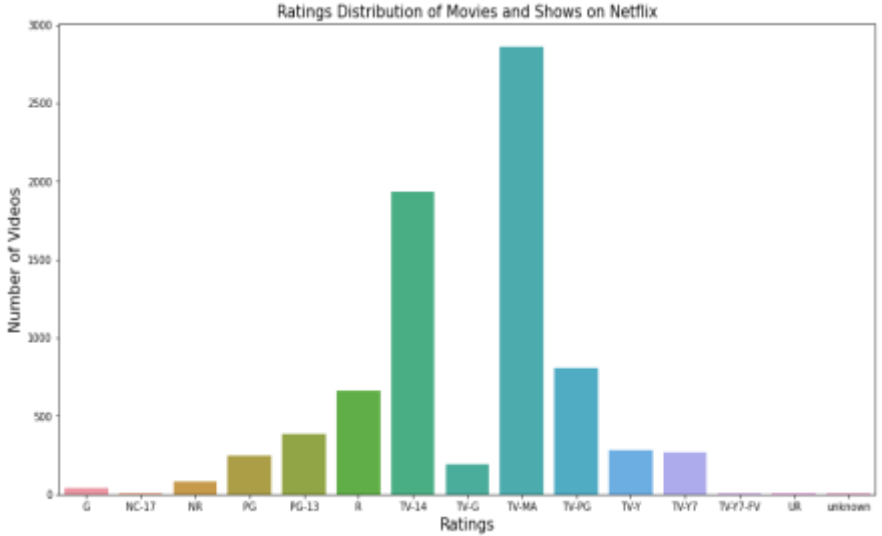

- Netflix began adding videos to its platform in 2008. This trend started increasing rapidly from 2017.
- More stand-alone movies were added per year as compared to TV shows.

# 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.

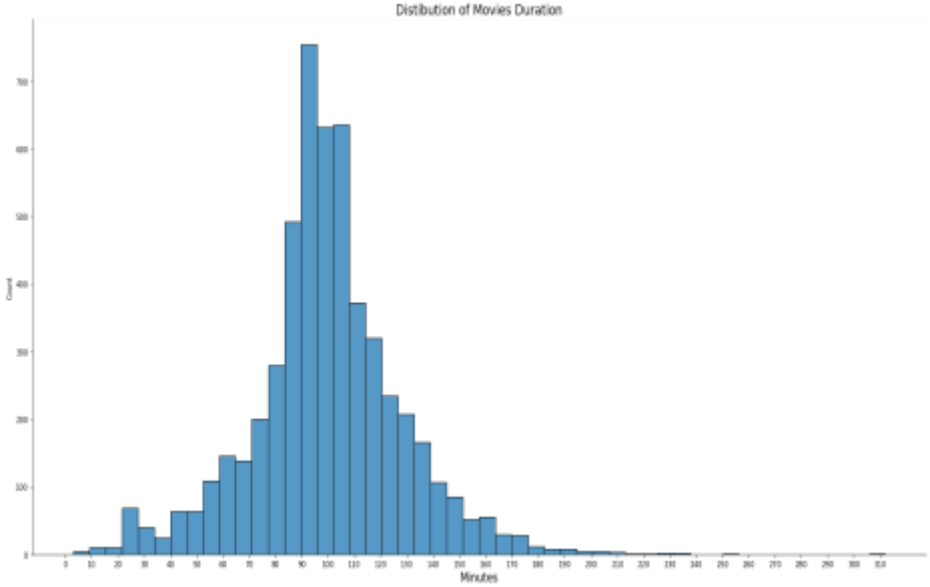
# Distribution of Video Type and Ratings on Netflix

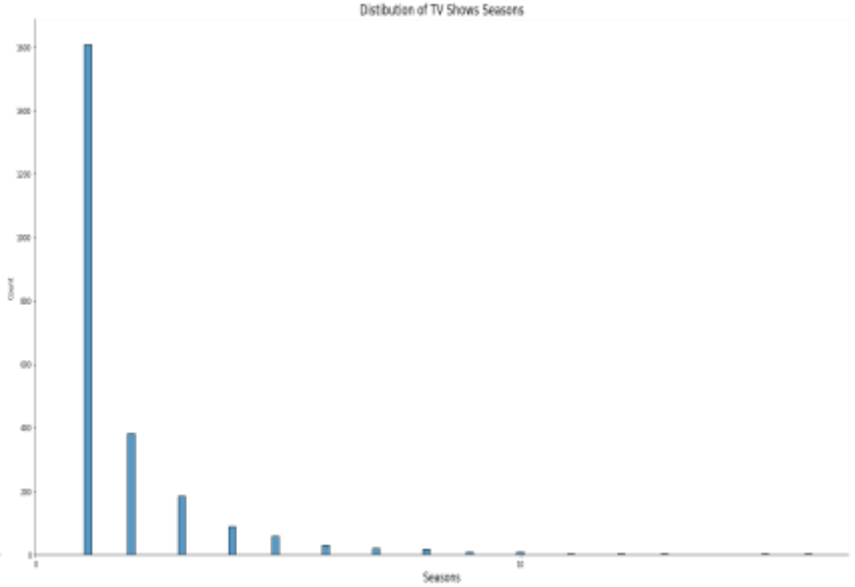



- There are almost twice as many movies as TV shows on Netflix
- Most of the contents are Movies
- Less than ⅓ content are Tv Shows
- Majority of the video content is rated for Mature Audiences and for audiences over 14 years old.

# Video Duration Distribution on Netflix



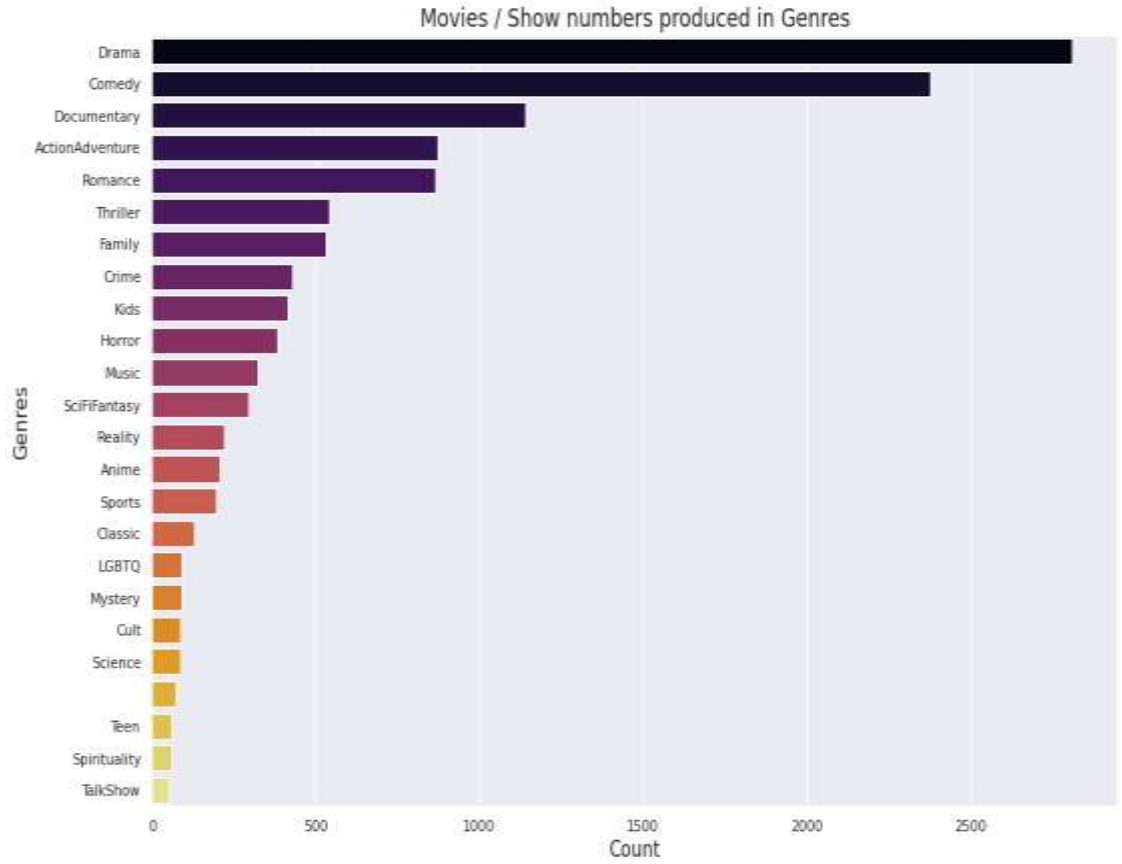




- Most movies on Netflix have duration ranging from 90 to 110 minutes
- The tenure of most TV shows on Netflix is only one season.

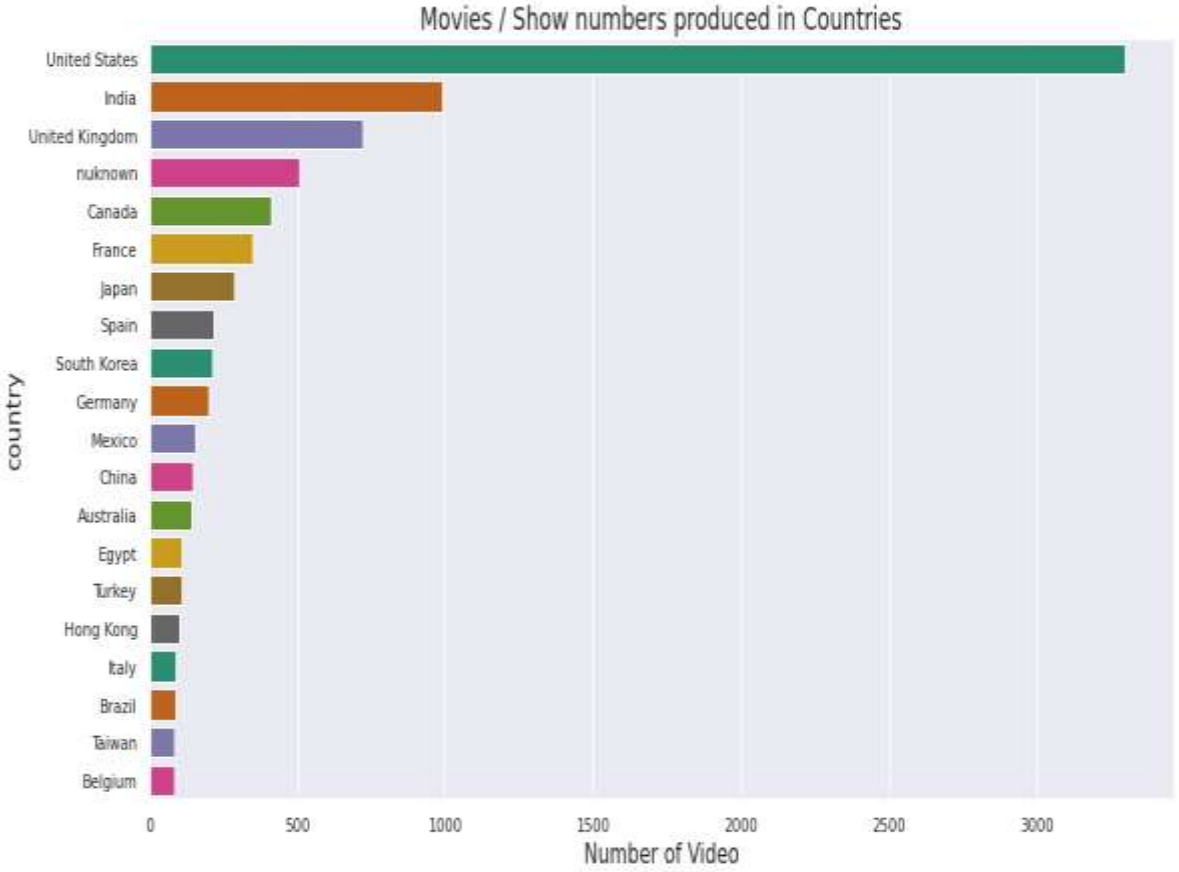
# Top Genres on Netflix

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

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



# Top Netflix video content producing Countries



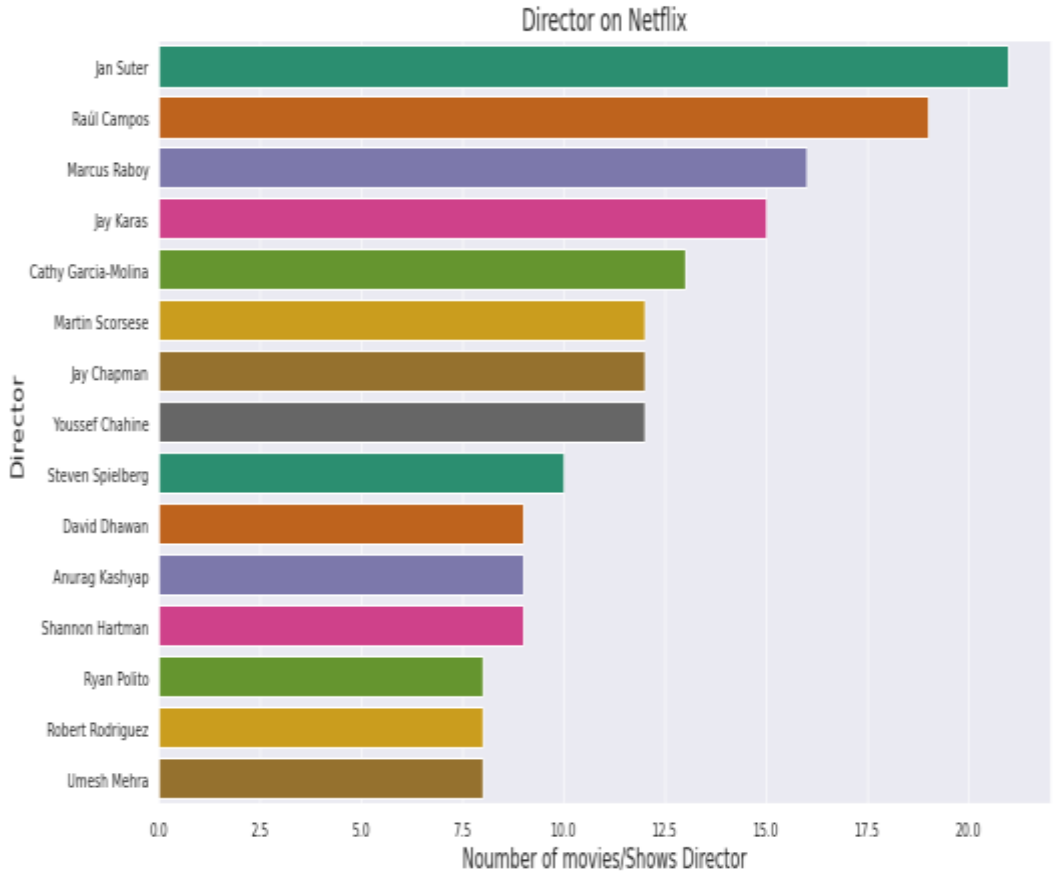

The top five biggest video content producers are

- United States of America
- India
- United Kingdom
- Canada
- France

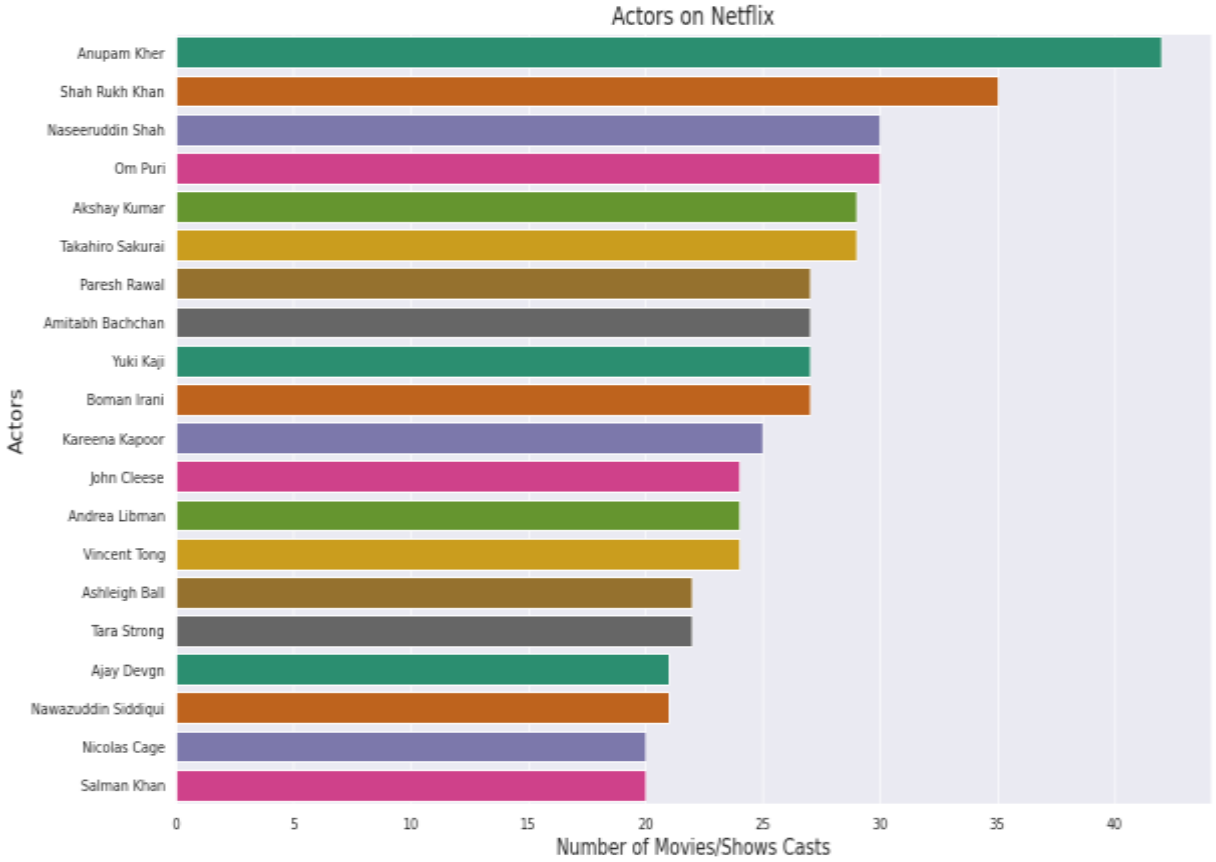
# Top Directors on Netflix

Top Directors on Netflix are:

1. Jan Suter
2. Raul Campos
3. Marcus Raboy
4. Jay Karas
5. Cathy Garcia-Molina



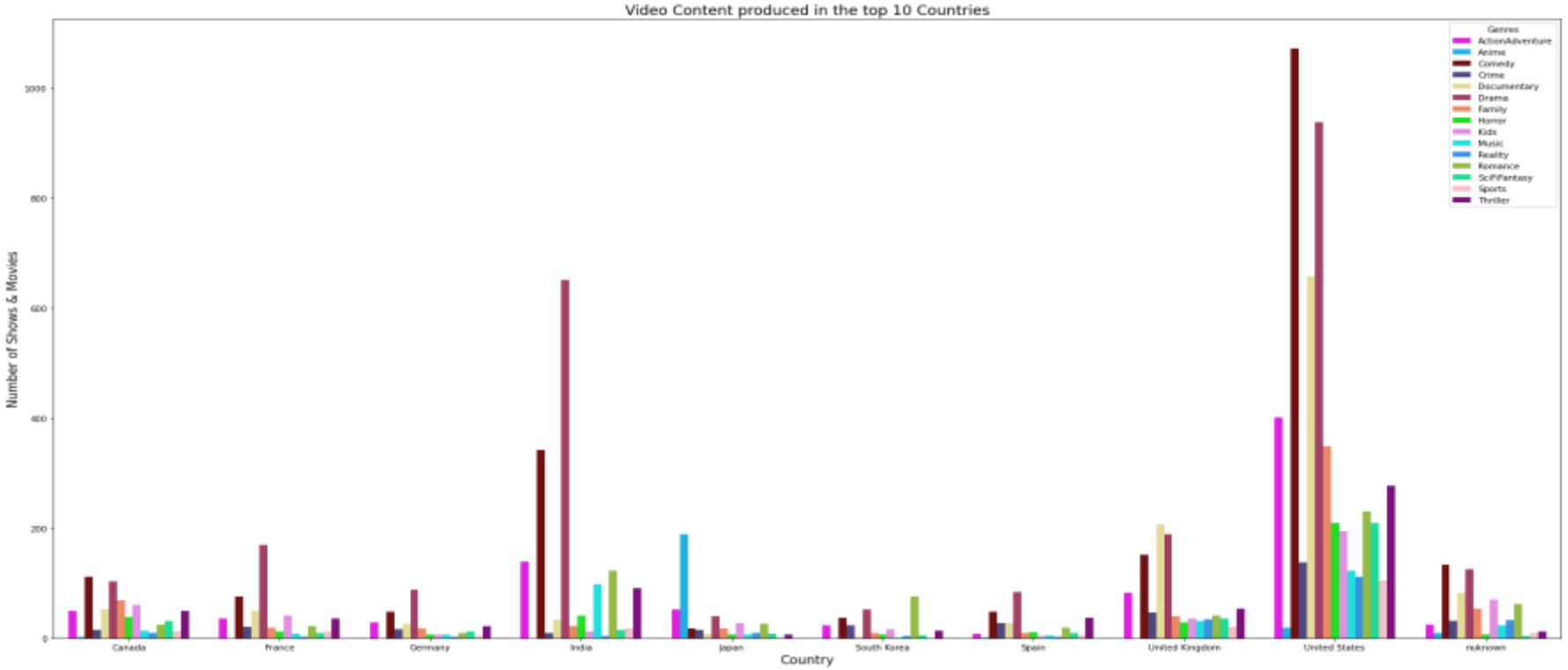
# Top Cast on Netflix




Top Actors on Netflix are:

1. Anupam Kher
2. Shah Rukh Khan
3. Naseeruddin Shah
4. Om Puri
5. Akshay Kumar

# Video Content in Top Countries



# Video Content in Top Countries

- We can observe that Drama [maroon bar] is the most produced genre in the top non-English speaking countries with exception of Japan and South Korea.

- Japan is the biggest producer of Anime [blue] across the platform. Japan is also the leading producer of the genre.

- Romance [green] is the most produced genre in South Korea.

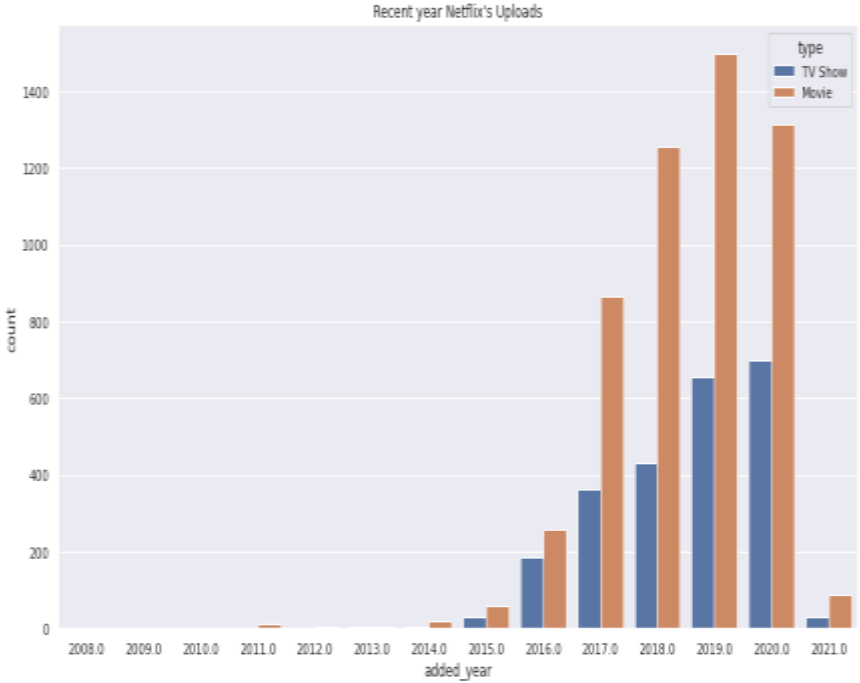
- It is noted that Comedy was a top genre in English-speaking countries like the United States of America, the United Kingdom and Canada.

- Documentaries are predominantly produced in the United Kingdom and the United States of America

# 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

- 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.



# 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

# 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.

# WordCloud

## What Is a Word Cloud?

A word cloud (also known as a tag cloud) is a visual representation of words. Cloud creators are used to highlight popular words and phrases based on frequency and relevance. They provide you with quick and simple visual insights that can lead to more in-depth analyses.

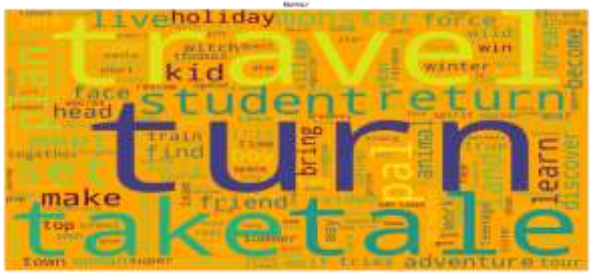
**Example :**

# Word Clouds for Top 9 Genres











# Bigram Word Clouds for Top 9 Genres

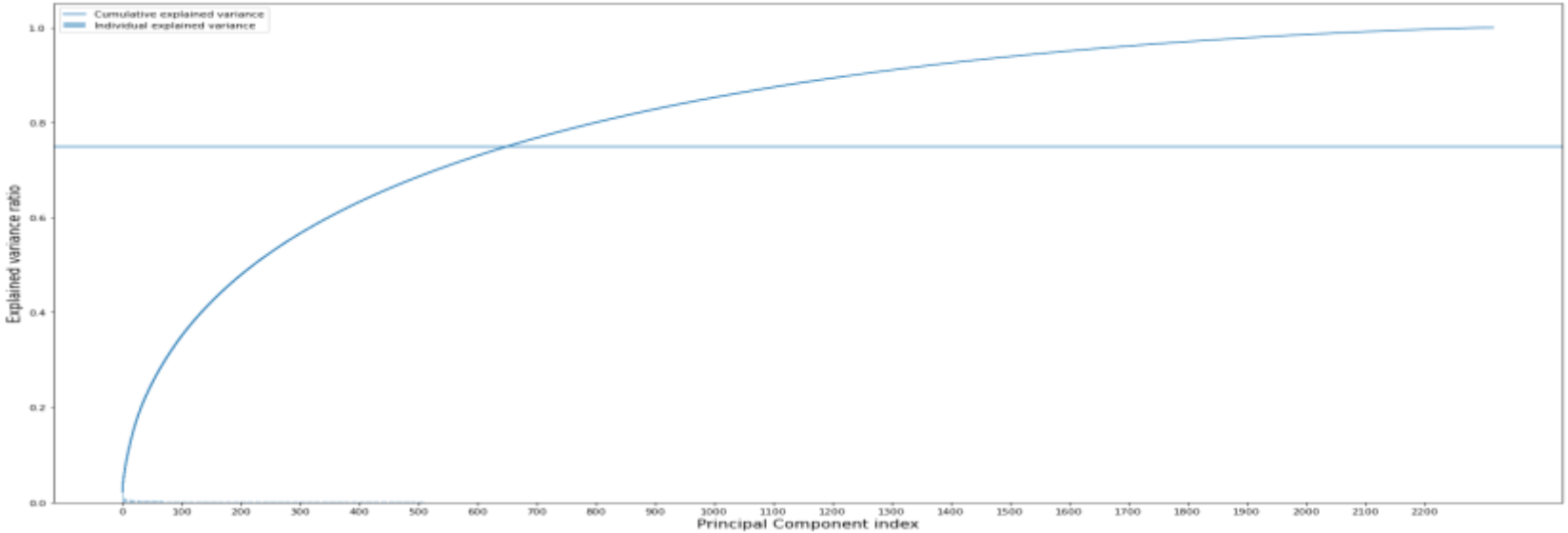








# Feature Engineering: Topic Modelling - Intuition

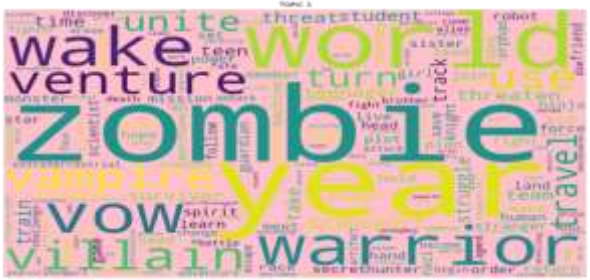



- Vectorising preprocessed attributes contributed 2,318 dimensions.
- For computational ease, these dimensions will have to be reduced using PCA.
- Upon plotting the cumulative explained variance chart, it was found that 700 components were required to explain at least 75 % of the variance.
- This was not a fair compromise as it was still computationally taxing with a 25 % loss in information.

# Most Relevant words for modeled Topics

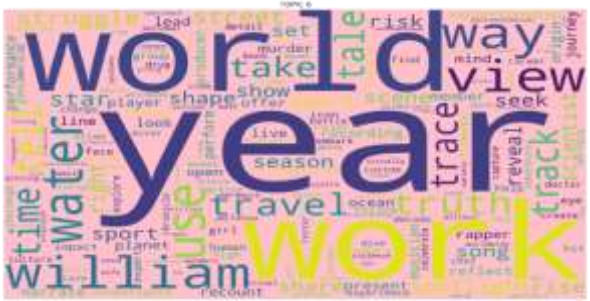
| Topic: 1 | Topic: 2 | Topic: 3 | Topic: 4 | Topic: 5 | Topic: 6 | Topic: 7 | Topic: 8 | Topic: 9 |
|---|---|---|---|---|---|---|---|---|
| drama | drama | actionadventure | comedy | crime | documentary | comedy | drama | comedy |
| comedy | horror | scififantasy | drama | drama | documentari | drama | comedy | kids |
| famili | thriller | drama | school | thriller | music | romance | romance | reality |
| romance | romance | sports | family | actionadventure | seri | new | love | friend |
| find | school | anime | high | murder | stori | find | famili | family |
| young | comedy | team | new | investig | explor | year | man | comedi |
| woman | young | world | teen | cop | life | life | young | special |
| life | teen | kids | world | polic | film | love | woman | show |
| love | student | power | student | detect | world | get | life | stand |
| man | life | save | scififantasy | drug | follow | friend | two | comedian |

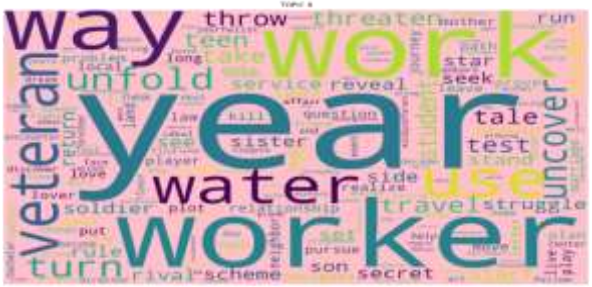
# TOPIC WISE WORD CLOUDS

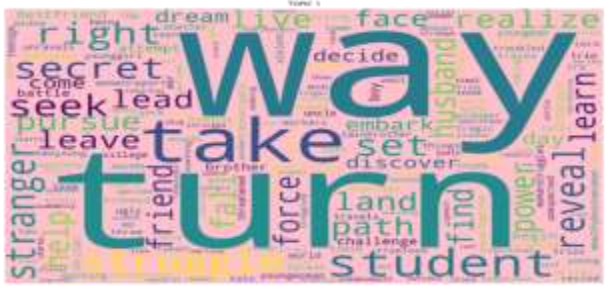







# Feature Selection & ML algorithem used

- Only selected 3 features , to do clustering
  - no_of_category
  - Length(description)
  - Length(listed-in)

- Using StandardScaler

- Used 5 algo to find out best k value
  - 1. Silhouette score
  - 2. Elbow Method
  - 3. DBSCAN
  - 4. Dendrogram
  - 5. Agglomerative Clustering

# 1. Silhouette Score

***Silhouette Coefficient Formula***

$$S = \frac{(b-a)}{max(a,b)}.$$

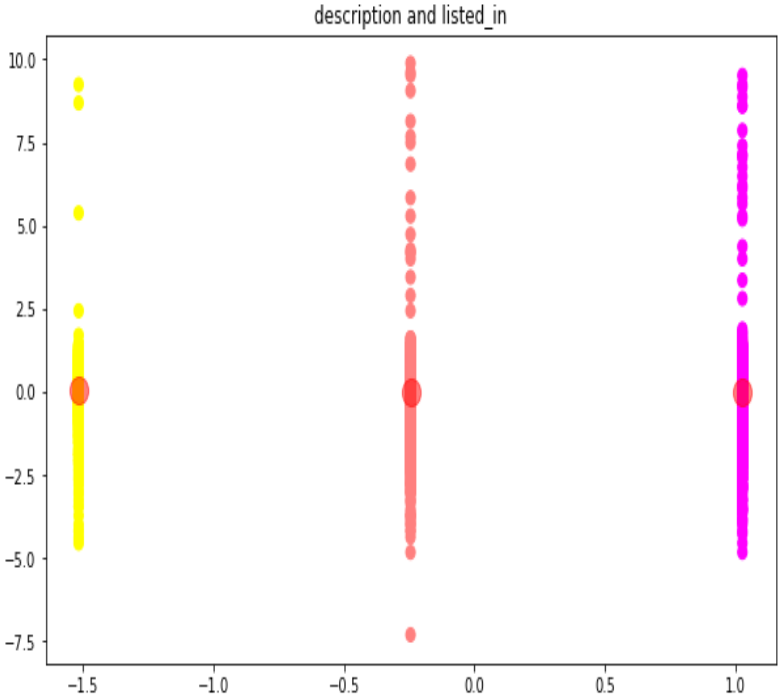
- **mean intra-cluster distance(a)**:- Mean distance between the observation and all other data points in the same cluster.
- **mean nearest-cluster distance (b)** :- Mean distance between the observation and all other data points of the next nearest cluster. This distance can also be called a.

The value of the silhouette coefficient is between [-1, 1]

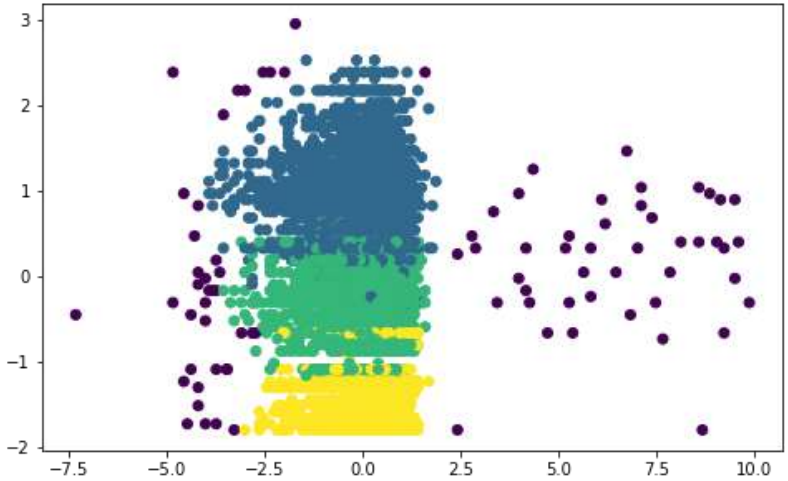
- If score is **1 denotes the best** meaning that the data point i is very compact within the cluster to which it belongs and far away from the other clusters.
- *The worst value is -1*
- If score is 0 denotes overlapping clusters

| | n clusters | silhouette score |
|---|---|---|
| 3 | 5 | 0.479 |
| 4 | 6 | 0.473 |
| 2 | 4 | 0.467 |
| 0 | 2 | 0.464 |
| 5 | 7 | 0.451 |
| 1 | 3 | 0.447 |
| 6 | 8 | 0.419 |
| 7 | 9 | 0.400 |
| 8 | 10 | 0.382 |
| 9 | 11 | 0.363 |
| 10 | 12 | 0.361 |
| 13 | 15 | 0.353 |
| 11 | 13 | 0.350 |
| 12 | 14 | 0.349 |

# 2. 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.

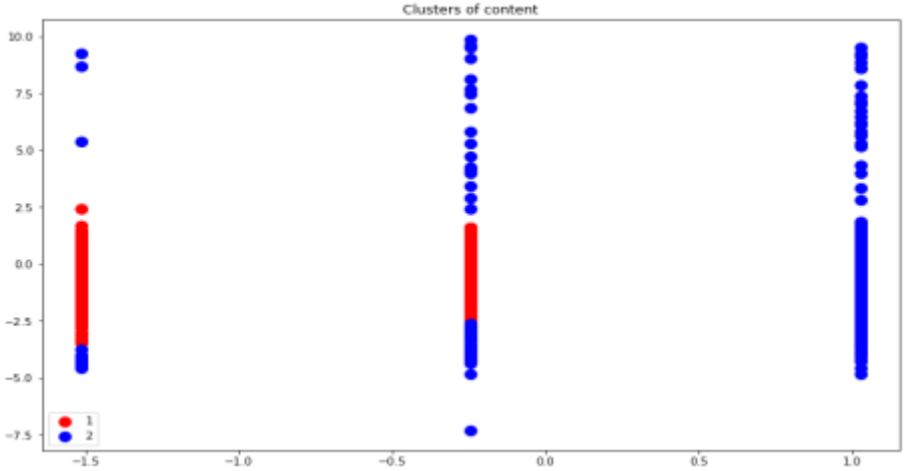



## 3) DBSCAN :-

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a base algorithm for density-based clustering. It can discover clusters of different shapes and sizes from a large amount of data, which is containing noise and outliers.

## 4) Agglomerative Clustering :-

Agglomerative Clustering is a bottom-up strategy in which each data point is originally a cluster of its own, and as one travels up the hierarchy, more pairs of clusters are combined. In it, two nearest clusters are taken and joined to form one single cluster.



# 5) Dendrograms


## 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.

# Conclusion

1)There are about 70% of movies and 30% of TV shows on Netflix.
2)Data set contains 7787 rows and 12 columns in that cast and director features contain a large number of missing values so we can drop it and we have 10 features for the further implementation
3)The United States has the highest number of content on Netflix by a huge margin followed by India.
4)Raul Campos and Jan Sulter collectively have directed the most content on Netflix.
5)Anupam Kher has acted in the highest number of films on Netflix. Drama is the most popular genre followed by comedy.
6)More of the content is released in the holiday season - October, November, December, and January.
7)The number of releases has significantly increased since 2015 and has dropped in 2021 because of Covid 19.

# Conclusion

1) Most films were released in the years 2018, 2019, and 2020.
2)The months of October, November, December and January had the largest number of films and television series released.

2. We started by removing nan values and converting the Netflix added date to year, month, and day using date time format.
3. For the clustering algorithm, we utilised type, director, nation, released year, genre, and year.
4. The final model we used was k-means clustering, which consisted of 2,3,4,5,6 clusters. Score of silhoutte in k-measn clustering :

    1. n_clusters = 2 The average silhouette_score is : 07049787496083262
    2. n_clusters = 3 The average silhouette_score is : 05882004012129721
    3. n_clusters = 4 The average silhouette_score is : 06505186632729437
    4. n_clusters = 5 The average silhouette_score is : 0.56376469026194
    5. n_clusters = 6 The average silhouette_score is : 04504666294372765


Thank
You