Capstone Project

Netflix Movies and TV Shows Clustering

by

Mahin Arvind Chanthira Sekaran

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.

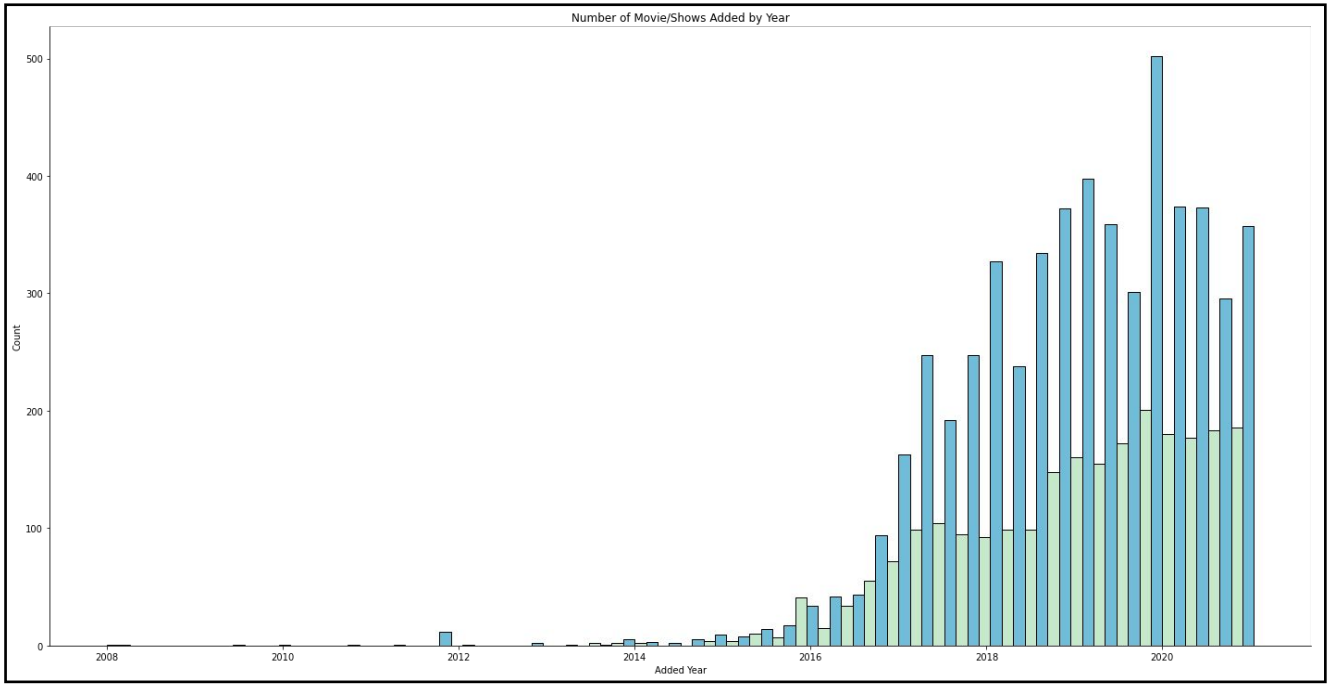
Feature	Type	Samples
show_id	Continuous	s1,s2,s3...
title	Text	[3%, Ozark,...]
type	Categorical	Movie/ TV Show
rating	Categorical	TV-MA, TV-R, R, PG-13....
director	Text	Raúl Campos, Jan Suter
cast	Text	David Attenborough
country	Categorical	United States
date added	Categorical	August 14, 2020
release 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..

Exploratory Data Analysis

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

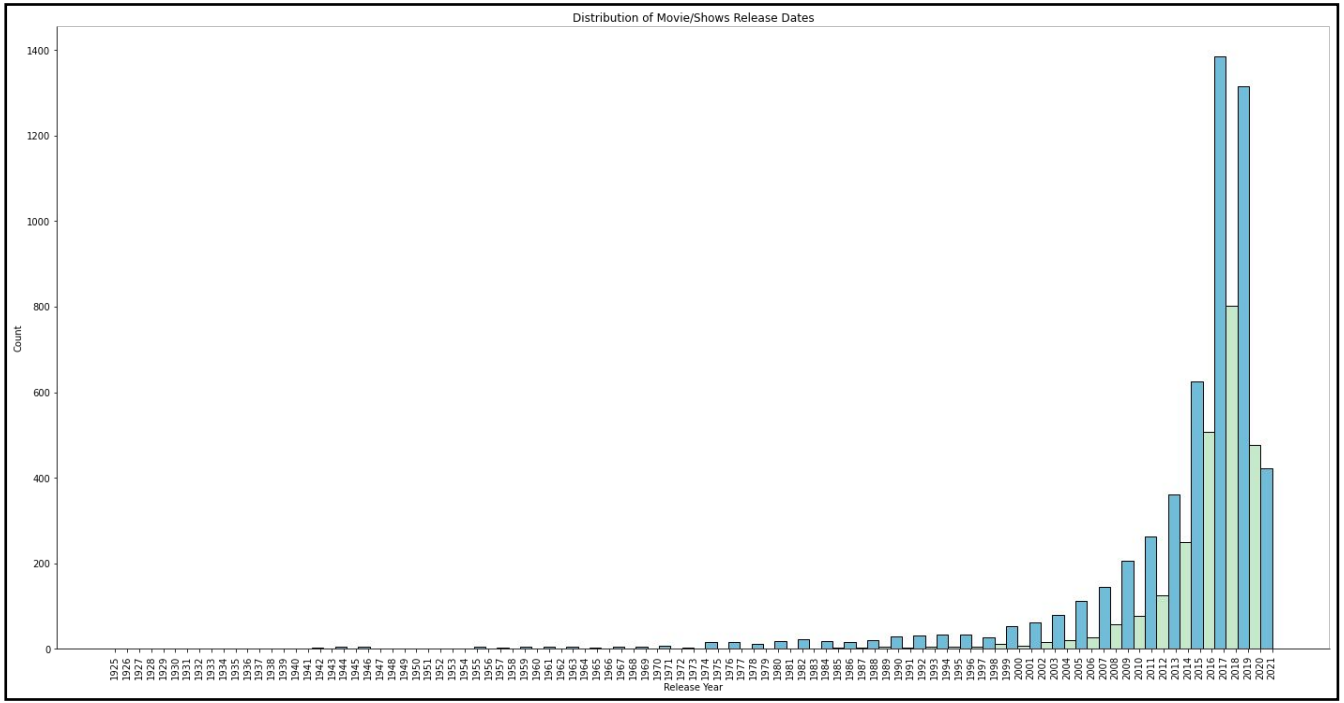
- Timelines of video content signings and releases
- Distribution of Video Content Categories on Netflix
- Type of Content Produced in the top Countries

Adding Dates of Movies and TV Shows



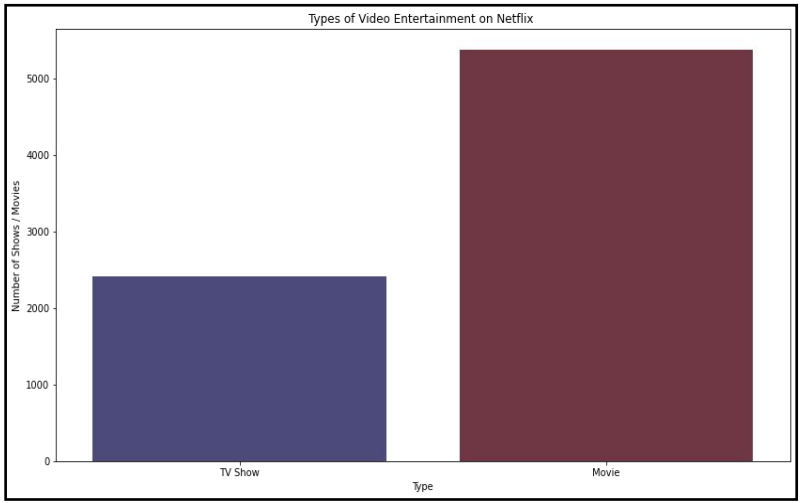
- Netflix began adding videos to its platform in 2008.This trend started increasing rapidly from 2017.
- More stand-alone movies were added per year as compared to TV shows.

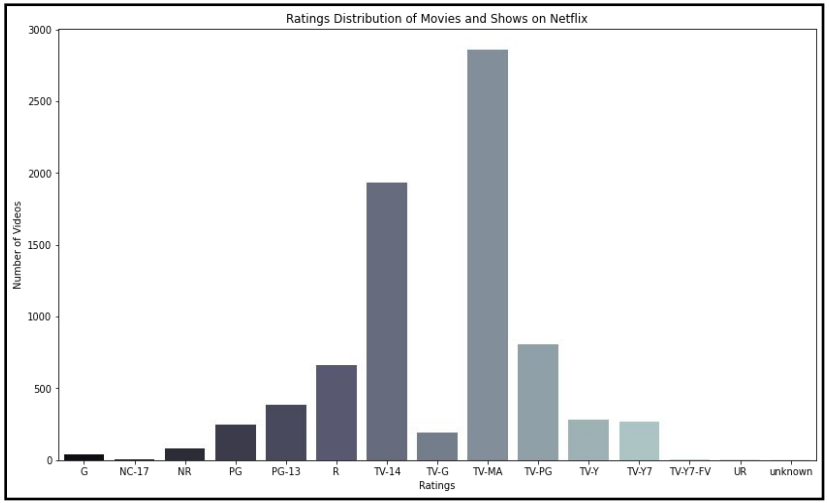
Release Dates of Movies and TV Shows



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

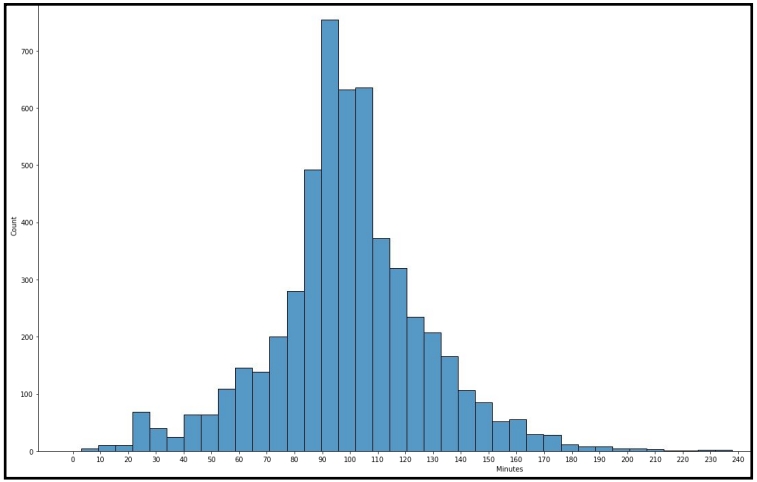
Distribution of Video Type and Ratings on Netflix

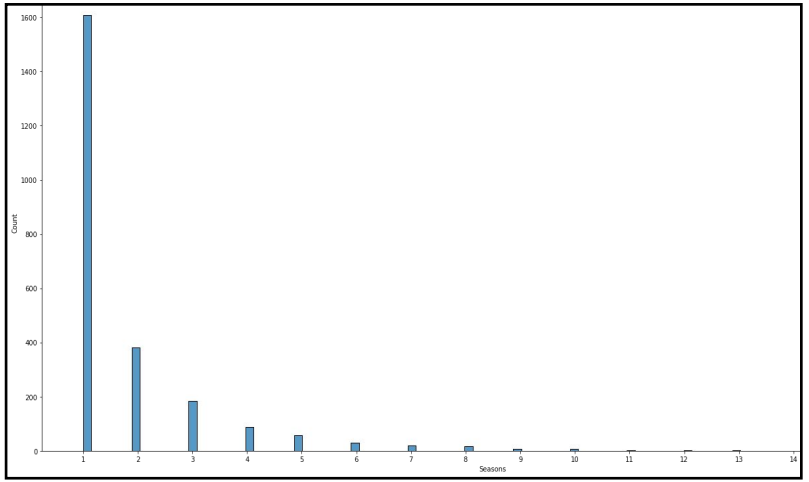




- There are almost twice as many movies as TV shows on Netflix
- Majority of the video content is rated for Mature Audiences and for audiences over 14 years old.

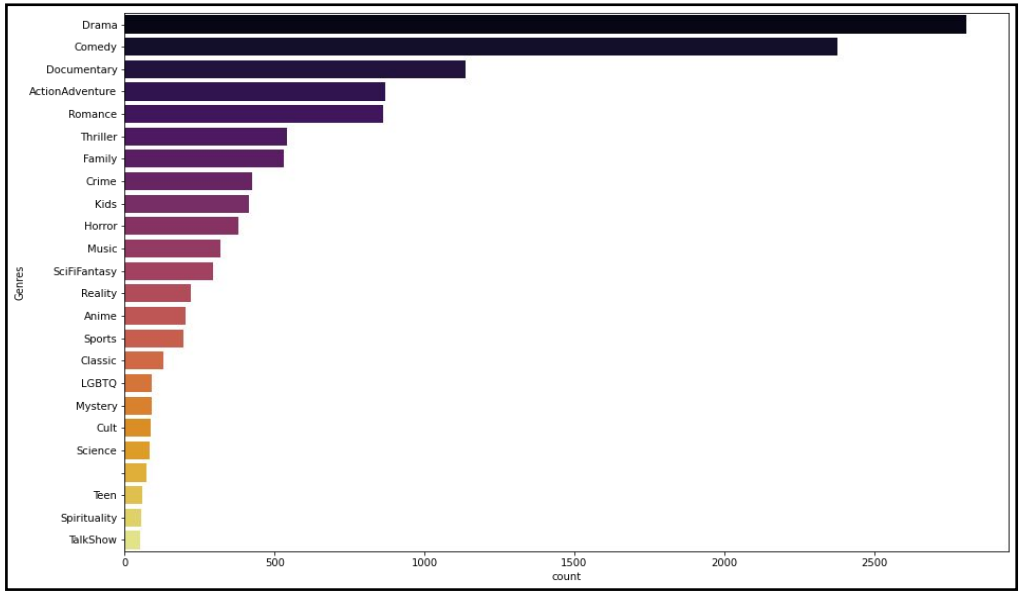
Video Duration Distribution on Netflix





- 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



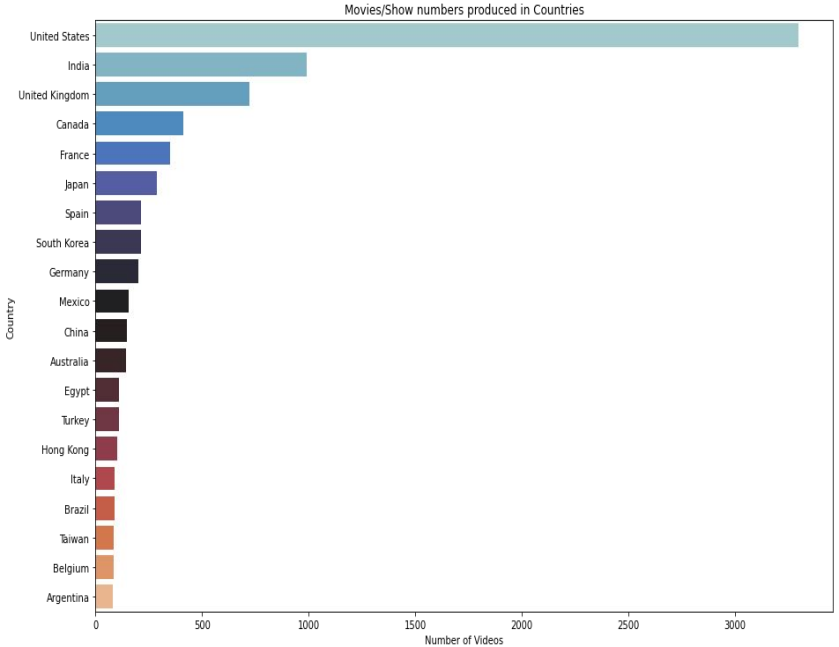
It is observed that the top video content genres on Netflix are

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

Top Netflix video content producing Countries

The top five biggest video content producers are

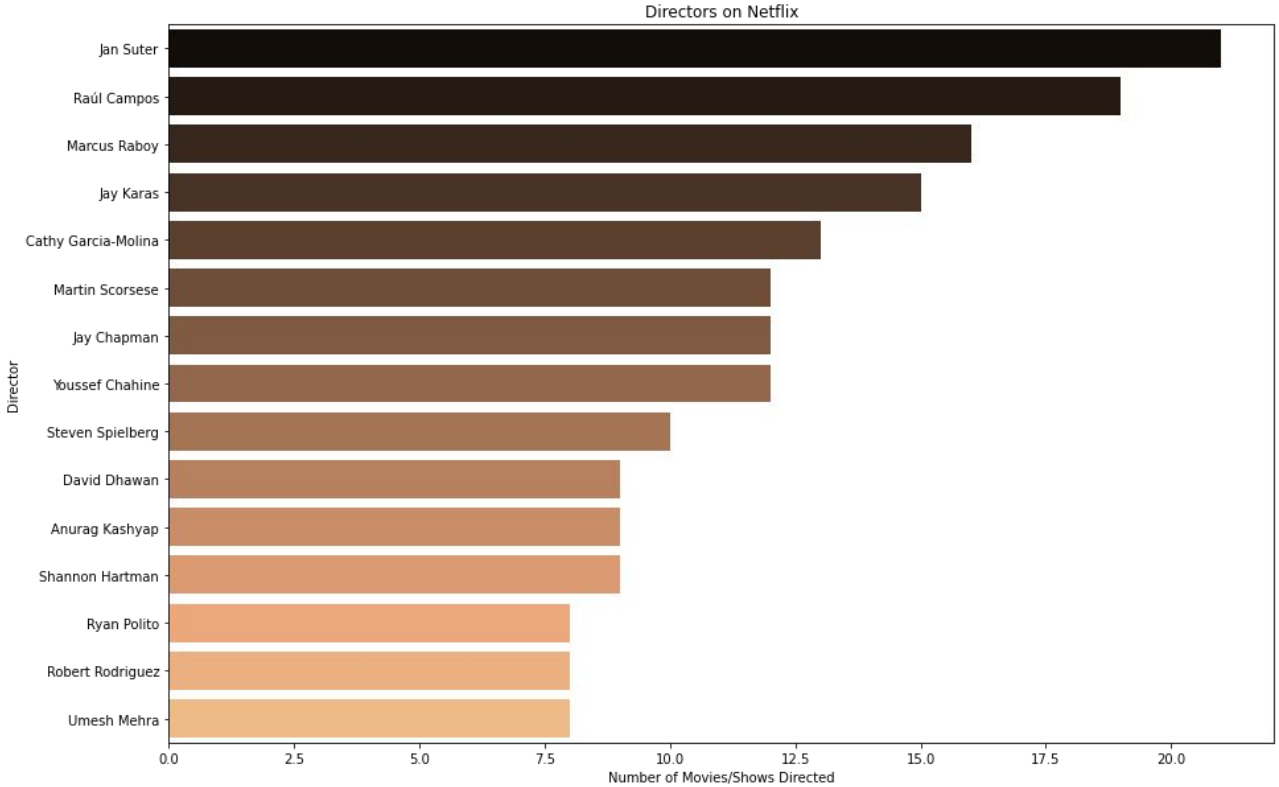
- United States of America
- India
- United Kingdom
- Canada
- France



Top Directors on Netflix

Top Directors on Netflix are:

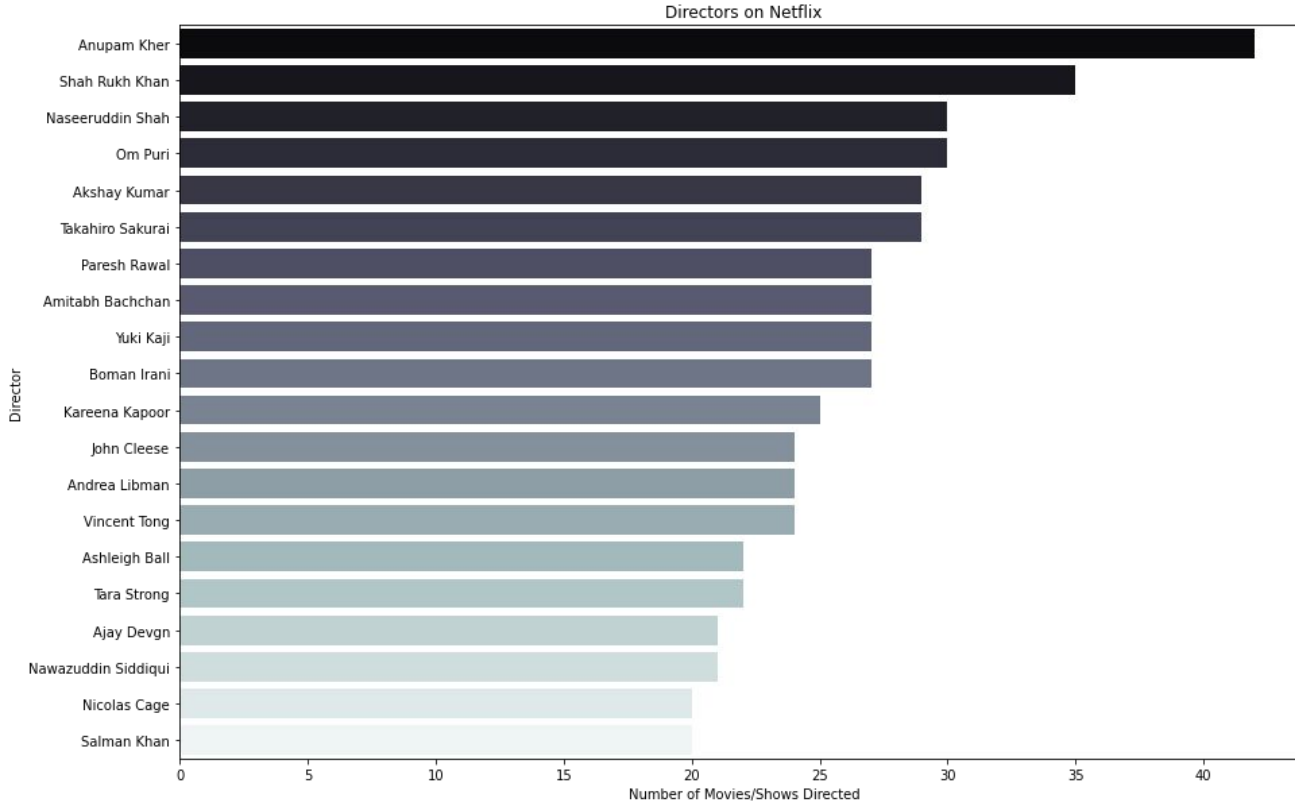
1. Jan Suter
2. Raul Campos
3. Marcus Raboy
4. Jay Karas
5. Cathy Garcia-Molina



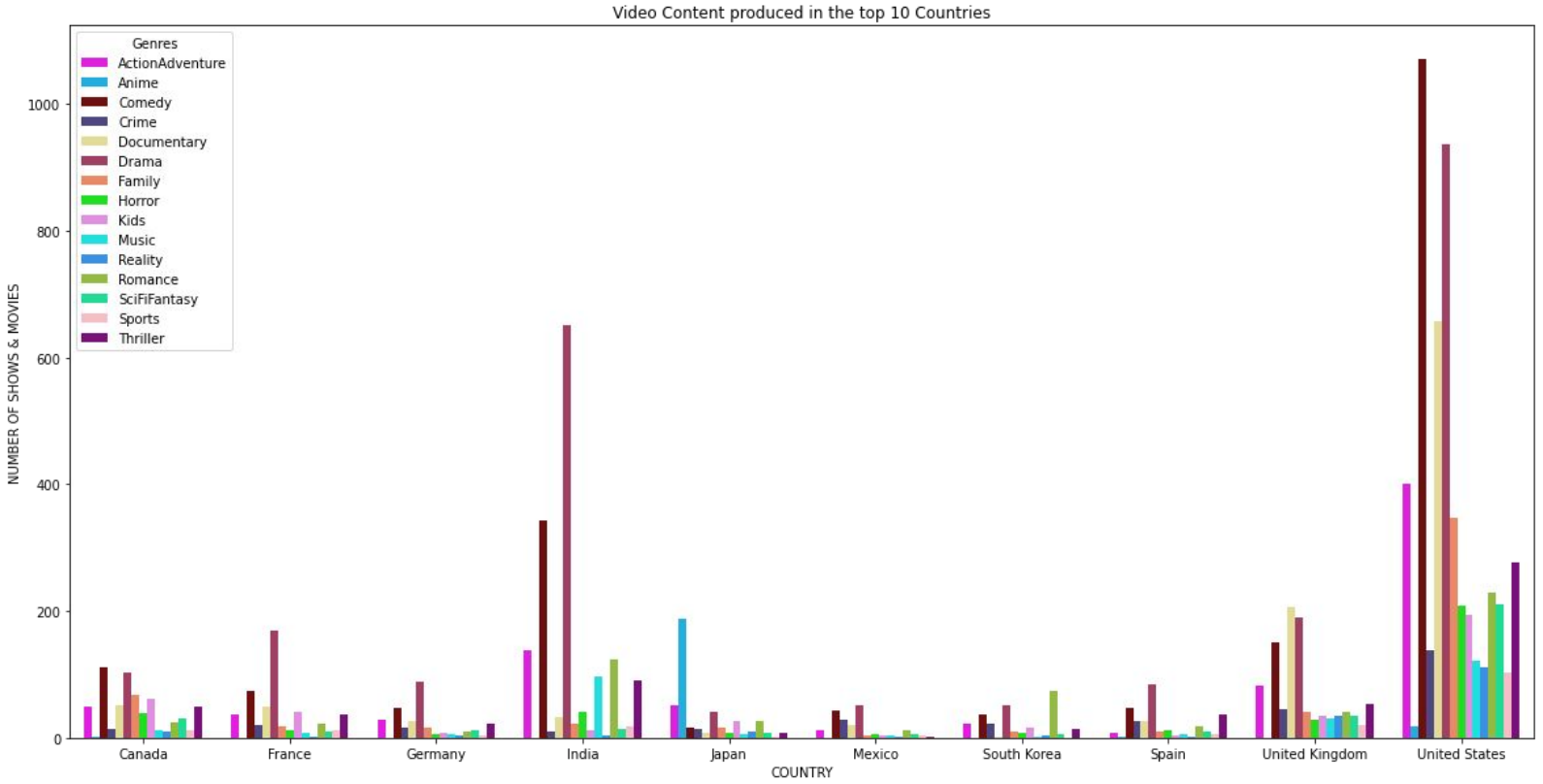
Top Cast on Netflix

Top Actors on Netflix are:

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

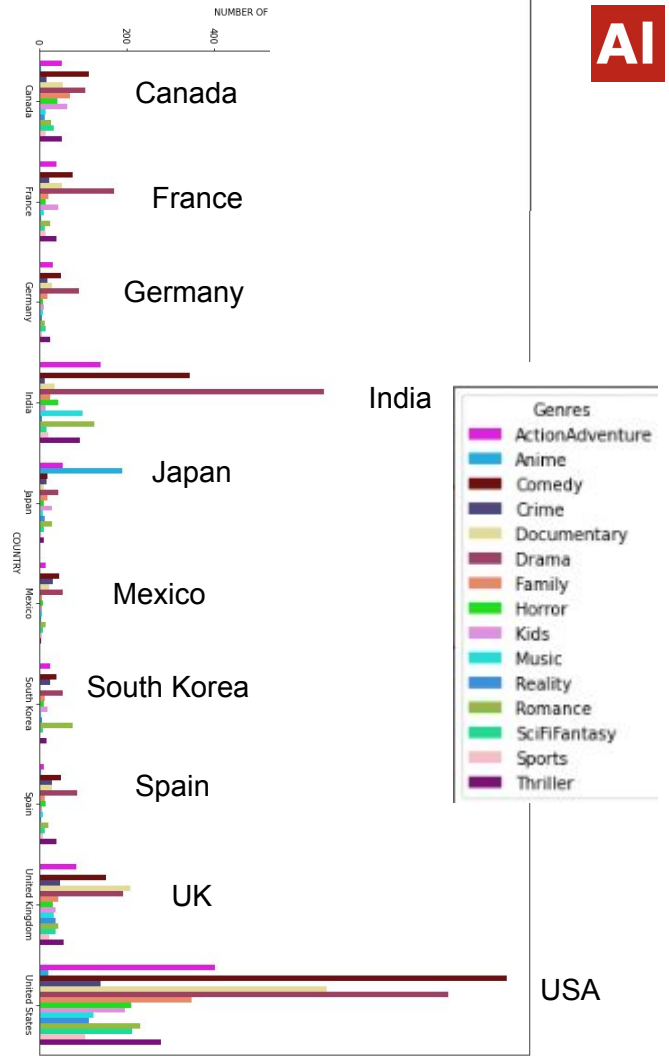


Video Content in Top Countries

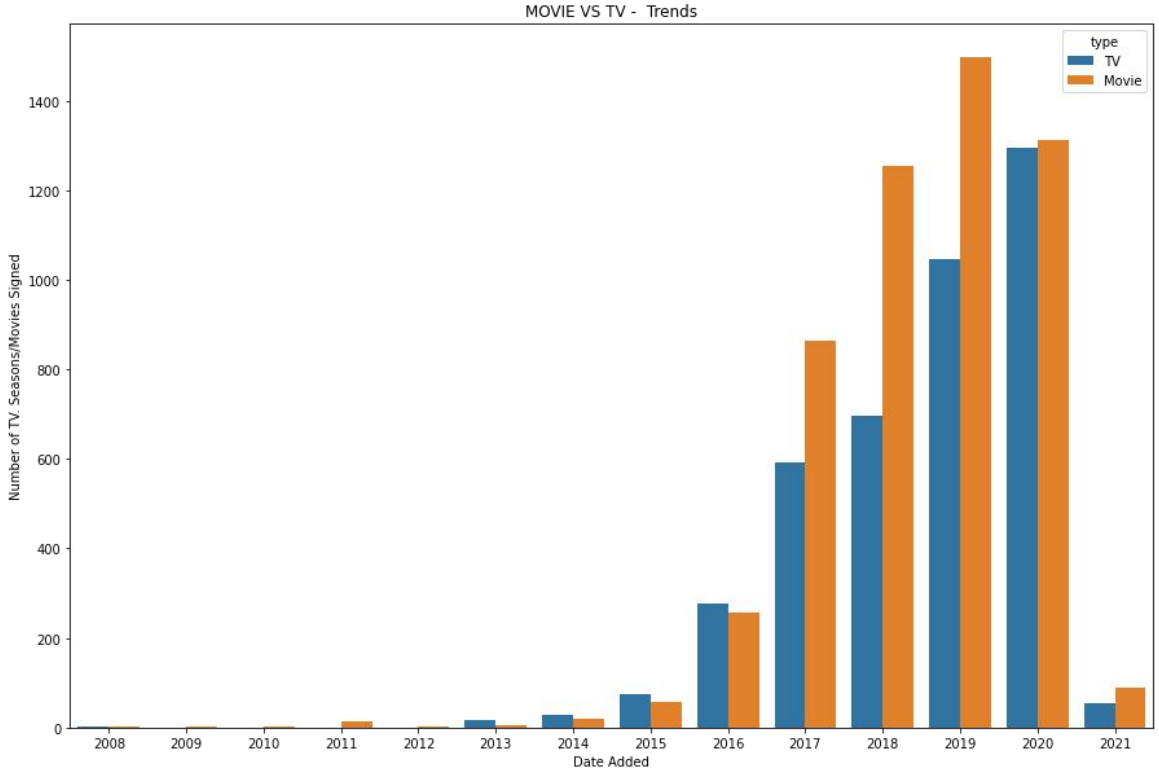


Video Content in Top Countries

- 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



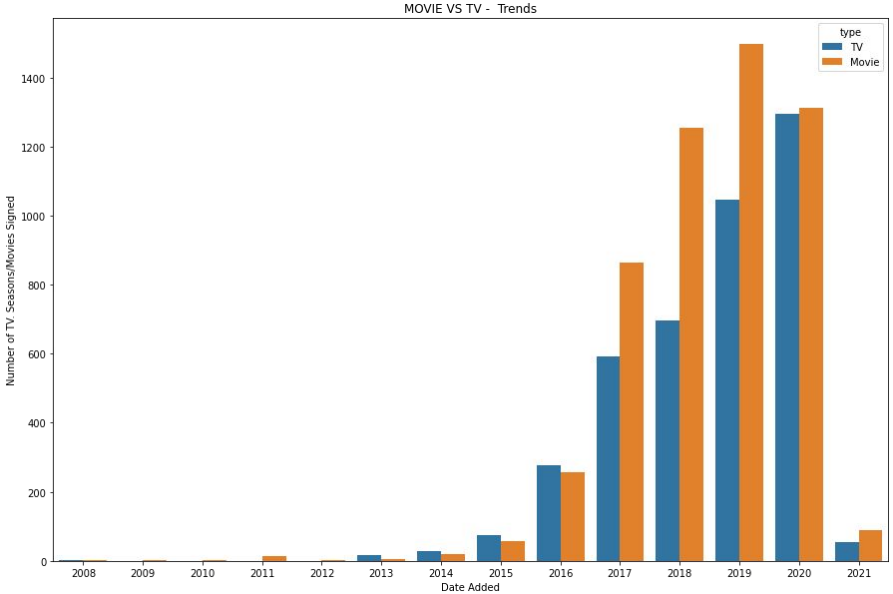
Adjusted TV Show and Movie Added Plot



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

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.



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 :

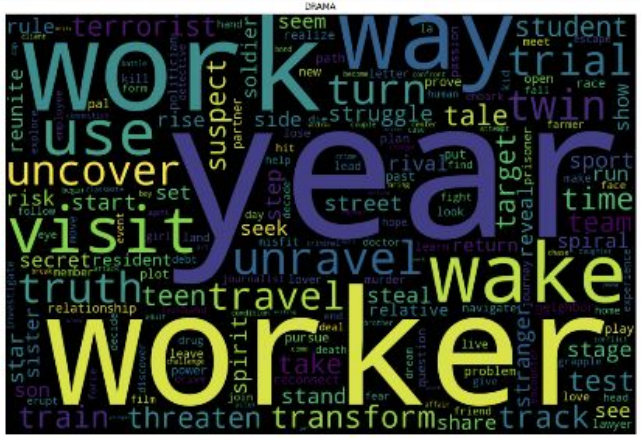
- **Tokenization:** Involves breaking of natural language text into chunks of information that can be considered as discrete elements. The token occurrences in a document can be used directly as a vector representing that document.

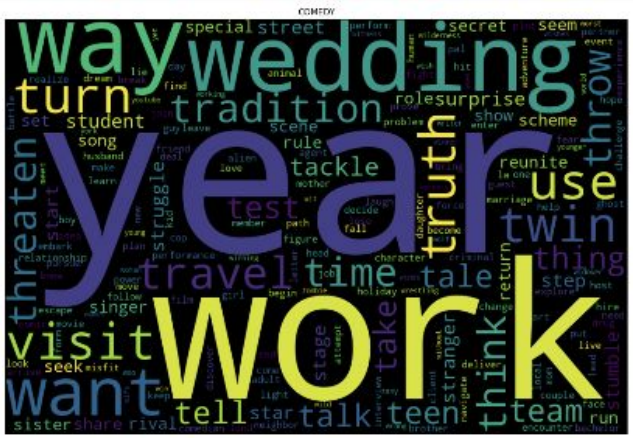
- **Punctuation Removal:** All the punctuations from the text are removed.

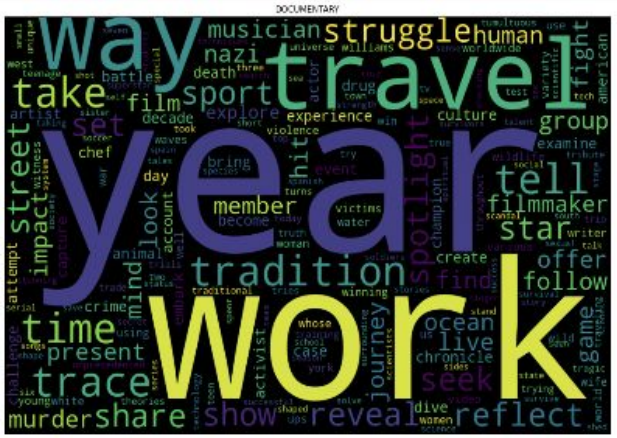
- **Stopword Removal:** Common words that add very little or no significant insight to the text being processed are removed beforehand. This reduces time and computational complexity.

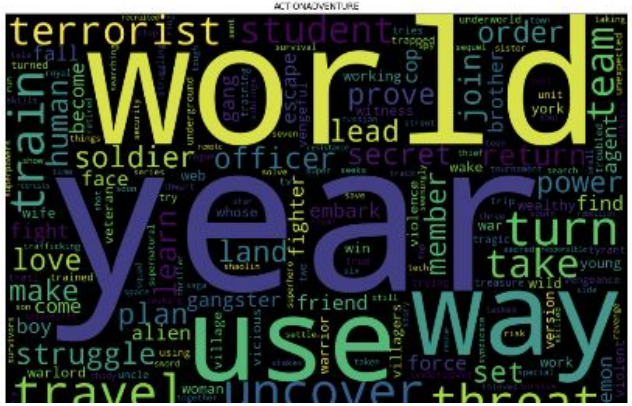
- **Stemming Words:** Stemming is the process of reducing inflected words to their word stem, base or root form—generally a written word form. This reduces different forms of the same word carrying the same base meaning. It should be noted that stemming does not remove synonyms.

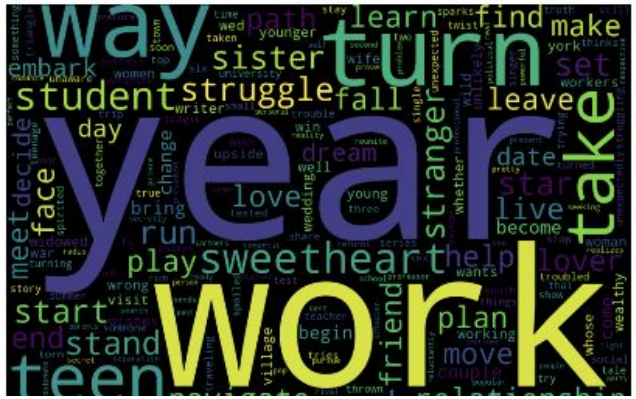
Word Clouds for Top 6 Genres

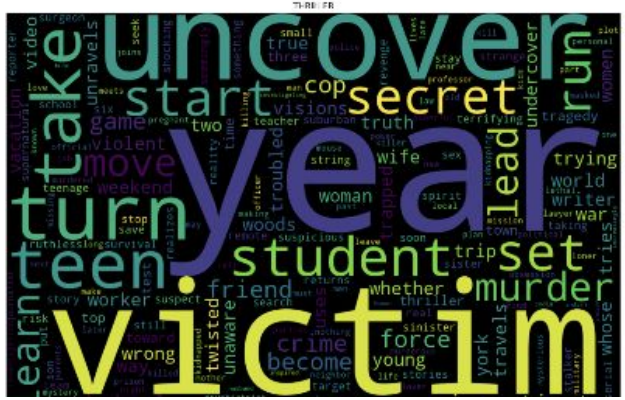




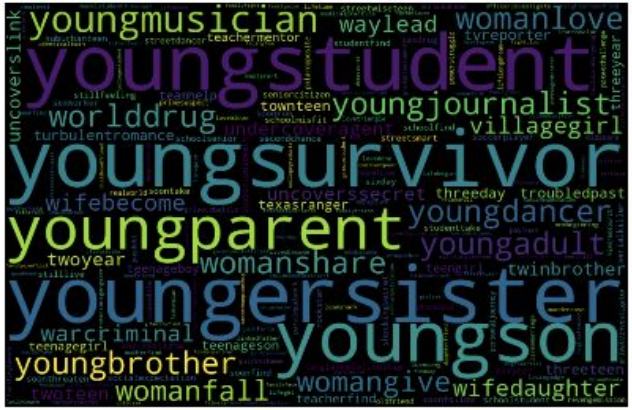


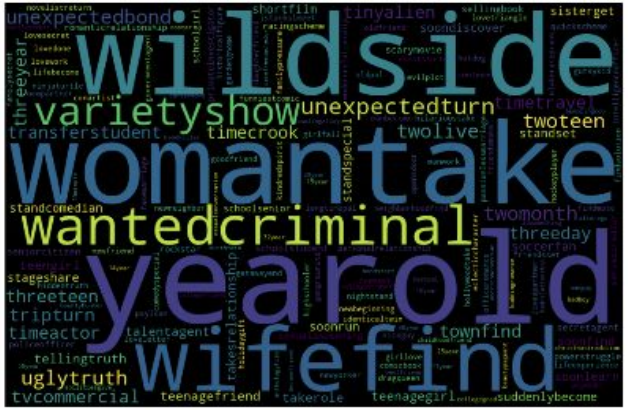


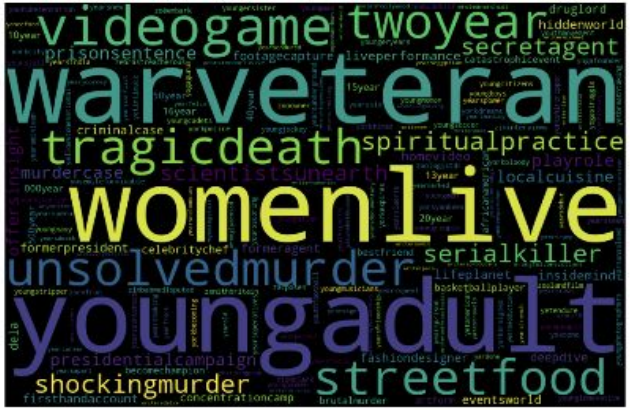


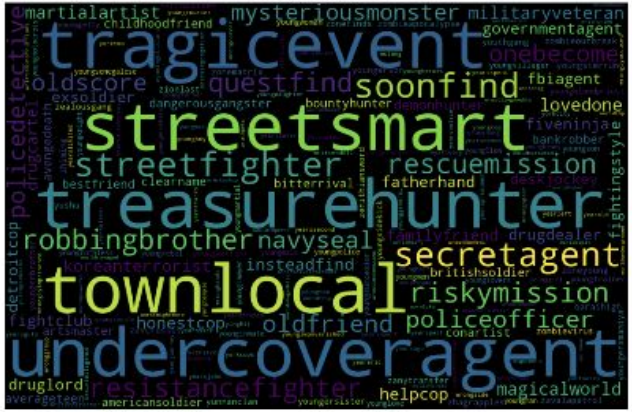


Bigram Word Clouds for Top 6 Genres

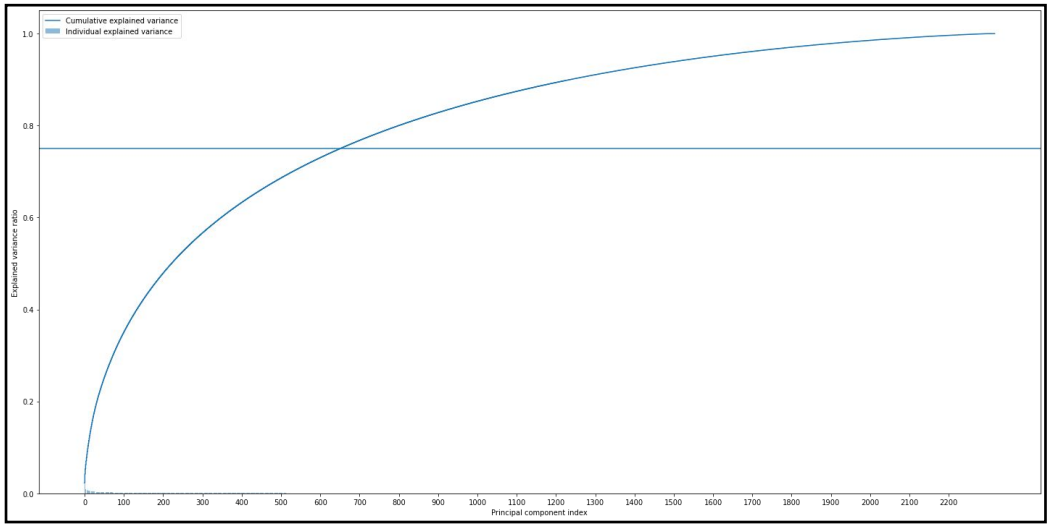








Feature Engineering: Topic Modelling - Intuition



Explained Variance Plot of Text Vector

- 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 percent of the variance.
- This was not a fair compromise as it was still computationally taxing with a 25 per cent loss in information.

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.

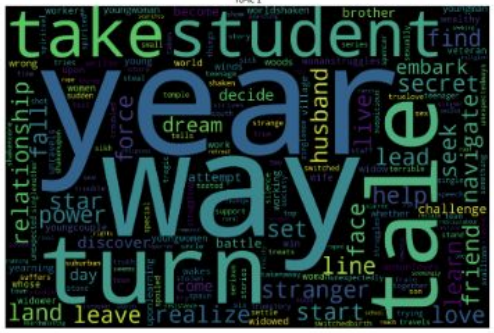
Topic Modelling - Latent Dirichlet Allocation

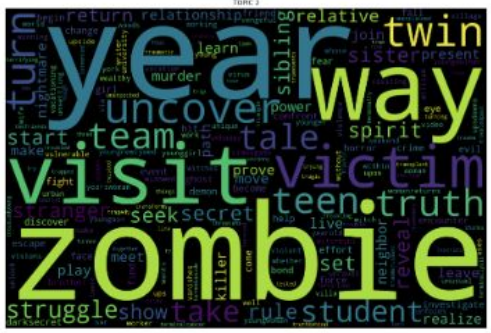
- LDA is a generative probabilistic model for collections of discrete data such as text corpora.
- It is a three-level hierarchical Bayesian model, in which each item of a collection is modelled as a finite mixture over an underlying set of topics.
- Each topic is modelled as an infinite mixture over an underlying set of topic probabilities.
- The topic probabilities provide an explicit representation of a document, in context of text modelling.
- To measure the coherence of topics modelled, Coherence Score was used to calculate the topic score by measuring the degree of semantic similarity between high-scoring words in the topic.
- This was chosen as the measure correlated well with human impression.

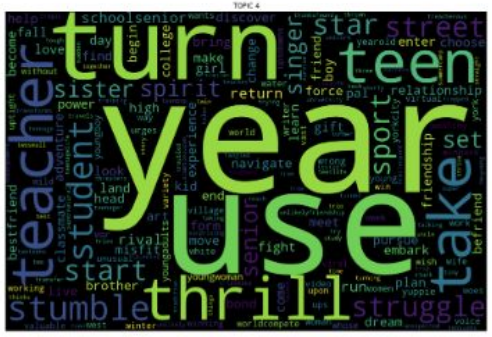
Most Relevant words for modelled Topics

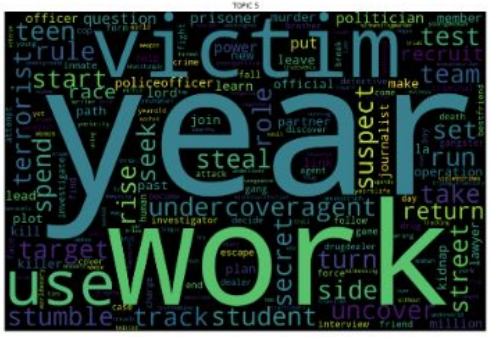
Topic: 1	Topic: 2	Topic: 3	Topic: 4	Topic: 5	Topic: 6	Topic: 7	Topic: 8	Topic: 9
drama	drama	actionadventure	comedy	crime	documentary	comedy	drama	comedy
comedy	horror	scififantasy	drama	drama	documentari	drama	comedy	kids
famili	thriller	drama	school	thriller	music	romance	romance	reality
romance	romance	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

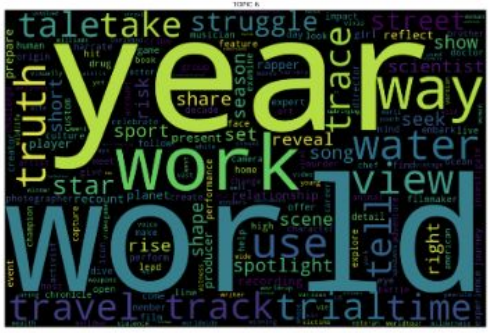
TOPIC WISE WORD CLOUDS

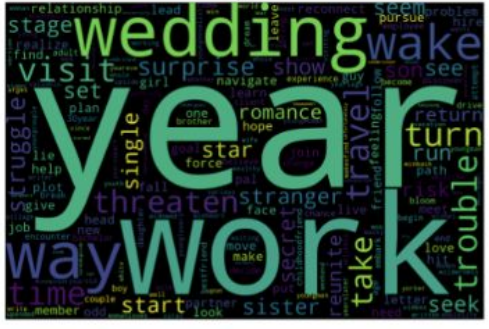


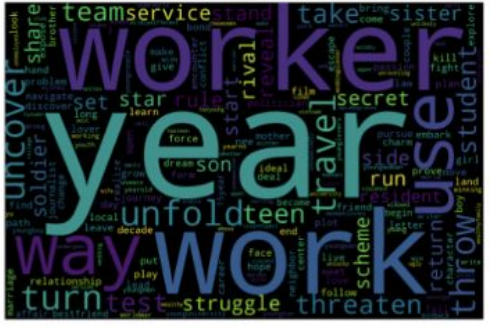


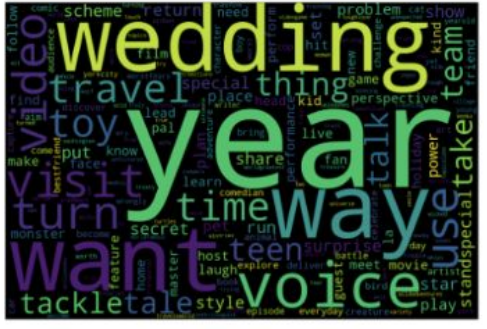












Feature Selection

- 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'.

Performance Metrics

- **Silhouette Score** is used to measure the separation distance between clusters. It displays a measure of how close each point in a cluster is to points in the neighbouring clusters.

- **Calinski-Harabasz Index** is the score is defined as the ratio between the within-cluster dispersion and the between-cluster dispersion. The higher the index, the better the performance.

- **Davies-Bouldin Index** is defined as the average similarity measure of each cluster with its own cluster. The similarity is the ratio of within-cluster distances to between-cluster distances. Closer to zero, the better.

Methods to Choose Optimal Cluster Numbers

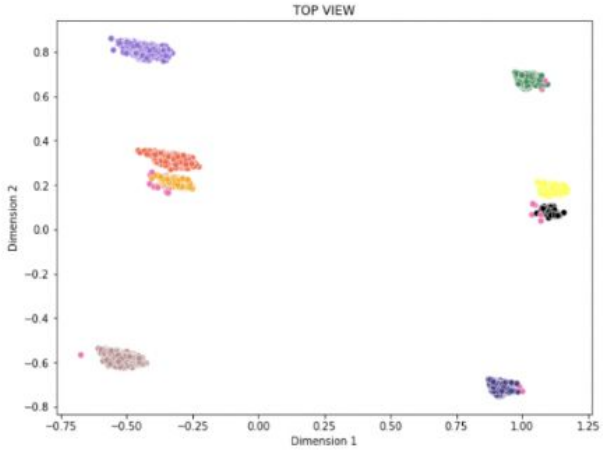
- **Elbow Method:** The elbow method plots the value of the cost function produced by different values of clusters, k, in K-means clustering.
- The value of k at which improvement in distortion declines the most is called the elbow, at which we should stop dividing the data into further clusters.
- **Dendrogram 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

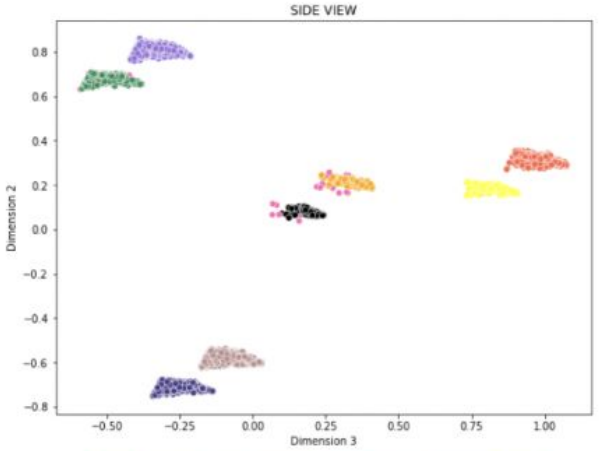
Observations

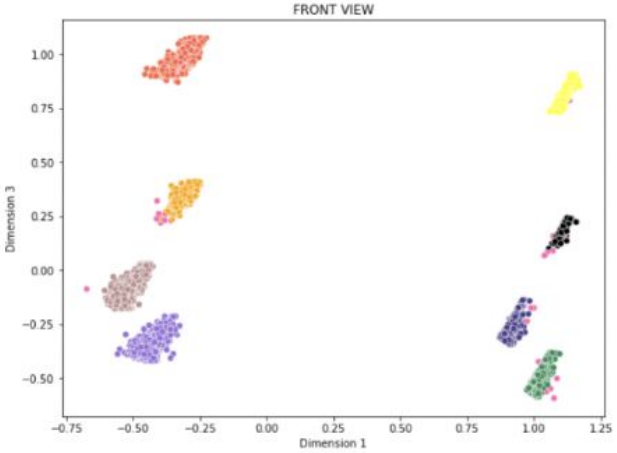
Density Based Spatial Clustering with Application of Noise

- DBSCAN is the base algorithm for density-based clustering
- It is based on the idea that a cluster in data space is a contiguous region of high point density separated from other clusters by contiguous regions of low point density.
- It can discover clusters of different shapes and sizes from a large amount of data with the presence of noise and outliers.
- DBSCAN clustered the content into 9 clusters with a silhouette score of 0.4664, Davies-Bouldin Index of 1.62 and Calinski-Harbaz Score of 2510.76.

DBSCAN Cluster Visualisation using PCA





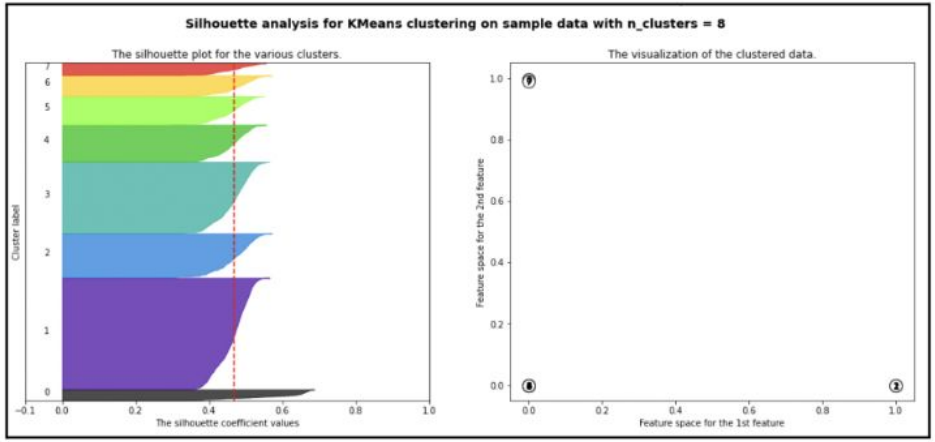


Cluster Colors

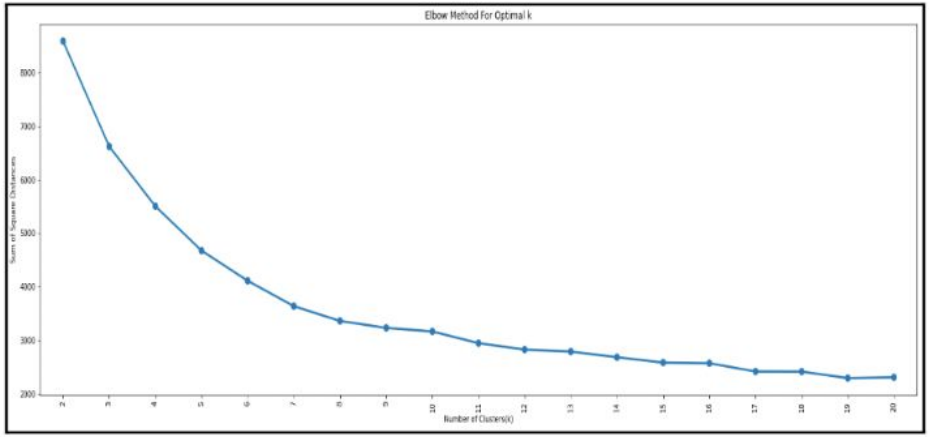
K - Means

- 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.

Plots to Choose Optimal Clustering

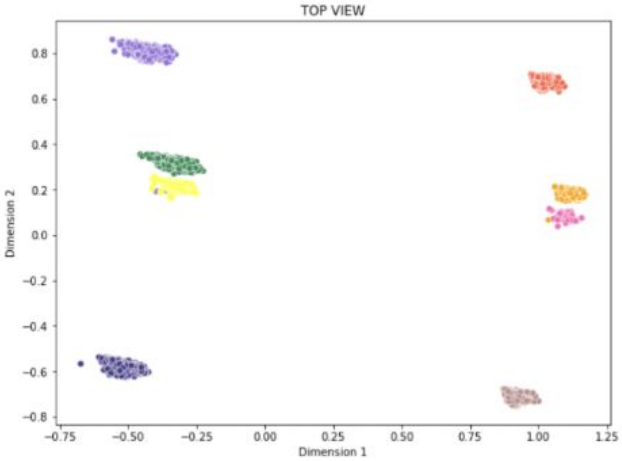


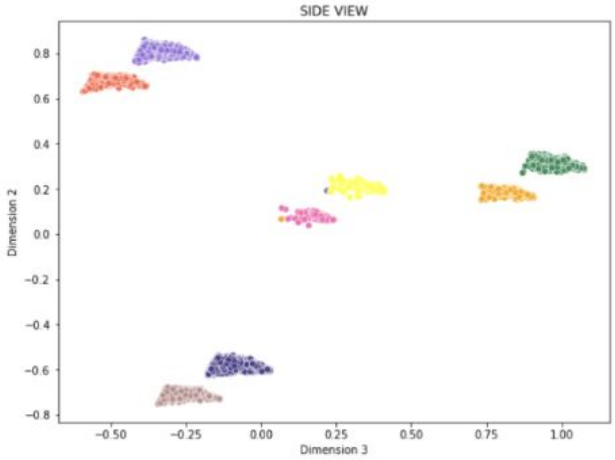
Silhouette Plot

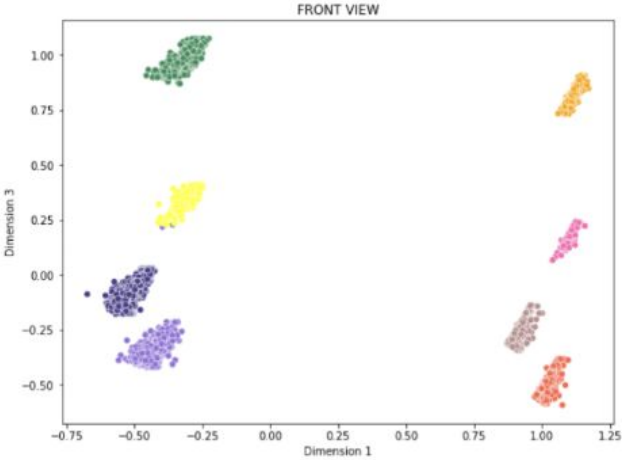


Elbow Curve

K-Means Cluster Visualisation using PCA





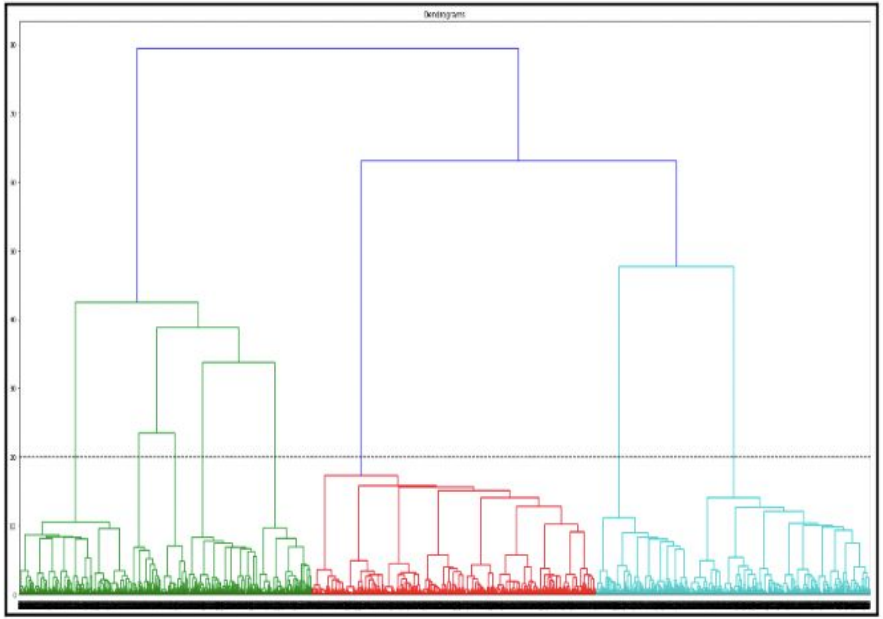


Cluster Colors

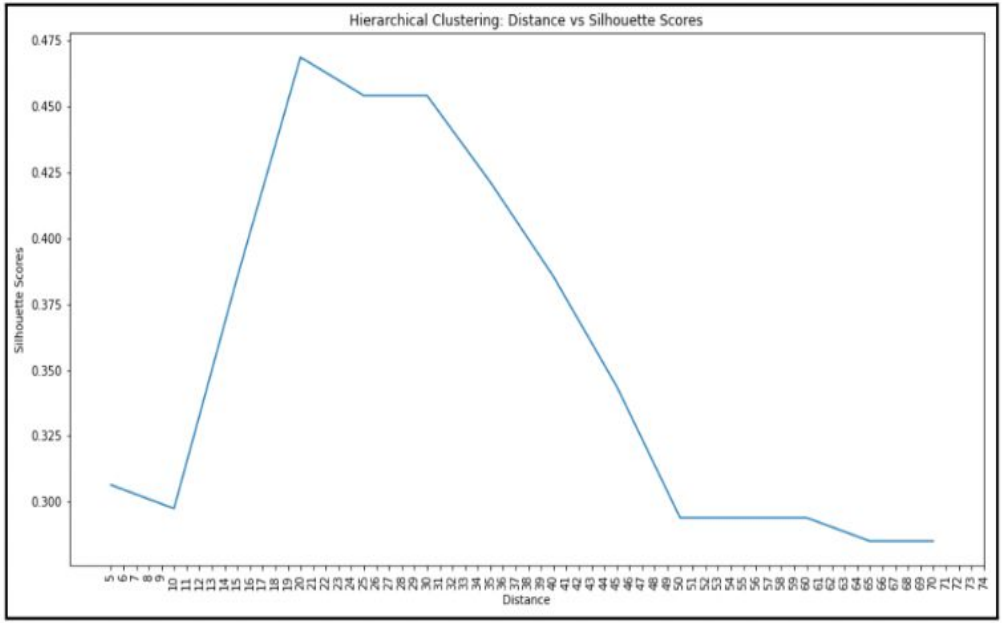
Hierarchical Clustering

- A Hierarchical clustering method works via grouping data into a tree of clusters. The aim is to produce a hierarchical series of nested clusters.
- The agglomerative algorithm starts by considering every data point as an individual cluster and calculates the similarity of one cluster with all the other clusters.
- The highly similar or close clusters are merged and the proximity matrix for each cluster is recalculated.
- These steps are repeated until only a single cluster is made
- The dendrogram distance was optimal at a distance of 20 with eight clusters producing a silhouette score of 0.46867, Davies-Bouldin Index of 0.889 and Calinski-Harbaz Score of 2900.28.

Plots to Choose Optimal Clustering

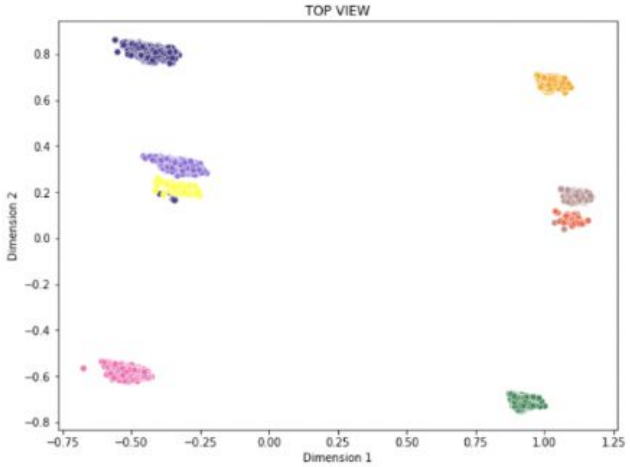


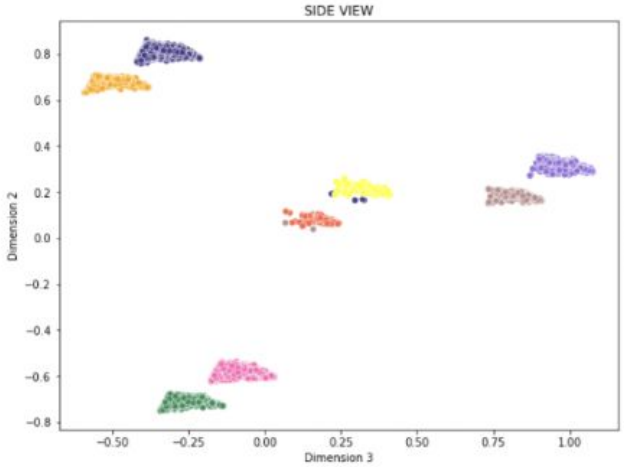
Dendrogram Plot

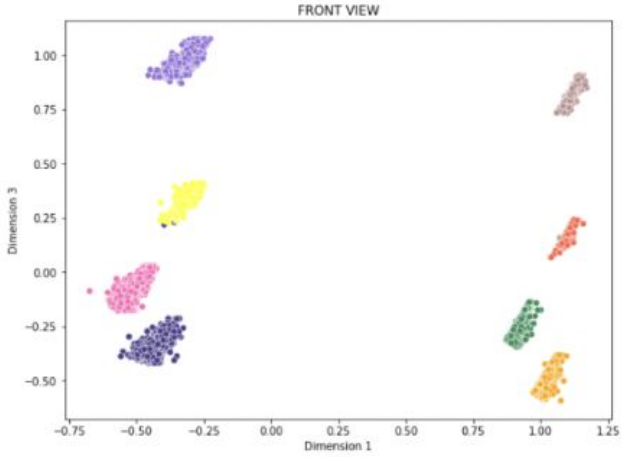


Silhouette Score vs Distance Plot

Hierarchical Cluster Visualisation using PCA



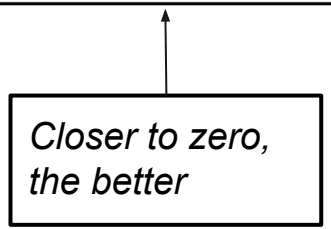




Cluster Colors

Evaluation Report

Algorithm	Clusters	Silhouette Coefficient	Davies-Bouldin Index	Calinski-Harbaz Score
DBSCAN	9	0.4664	1.6202	2510.7685
KMeans	8	0.4686	0.8896	2901.8448
Hierarchical Agglomerative	8	0.4687	0.8886	2900.2838



Conclusions

We can now address the four sections of the problem statement effectively:

- In the first section, we performed a comprehensive exploratory data analysis and found the content expansion trends and timelines, the genre and rating distribution and the average duration of the video content streaming on the platform.

- The findings from the second section were that most non-English-speaking countries predominantly produced content belonging to the genre, Drama with exception of Japan and South Korea. English-speaking countries, on the other hand, were major producers of Comedy and Documentaries.

- Upon exploring the shows and movie signing trends in the third section, we have effectively confirmed that it is true that Netflix has been focusing increasingly on TV shows as compared to movies.

Conclusions (contd.)

- The use of a combination of topic models to process text data has aided in clustering movies and TV shows on Netflix.
- The best performing models, K-Means and Hierarchical Clustering, grouped data into eight clusters with a silhouette score of 0.47.
- In addition to helping build recommendation engines, this labelled content can be studied and explored to determine the type of content on demand, potentially providing intuition to content creators about the content Netflix would be interested in signing.

References

1. DBSCAN Clustering Algorithm in Machine Learning | KDNuggets
2. Evaluate Topic Models: Latent Dirichlet Allocation (LDA) | by Shashank Kapadia | Towards Data Science.
3. Agglomerative Clustering and Dendrograms — Explained | by Satyam Kumar | Towards Data Science.
4. Understanding Topic Coherence Measures | by João Pedro | Towards Data Science
5. Hierarchical Clustering: Agglomerative and Divisive — Explained |by Satyam Kumar| Towards Data Science|

Thank You