Netflix Movie and TV Show Clustering

Mahin Arvind C
Almabetter, Bangalore

Abstract

Netflix is a subscription service model that offers personalized recommendations, to help viewers find shows and movies of interest. The streaming giant uses the content details and app usage data to make recommendations to each user.
In order to build recommendation engines, grouping and analysing the content catalogue may prove to be extremely beneficial. This would aid in planning the type of content needed to be signed to effectively suit the entertainment needs of different client bases.
In this project, three clustering techniques namely DBSCAN, K-Means and Hierarchical Clustering are implemented and analyzed to efficiently cluster Netflix video content data.

1. Problem Statement

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

1. Understand distributions of different types of content categories across 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, and
4. Cluster similar content based on text-based features.

2. Introduction

Netflix began experimenting with data in 2006 when they held a competition to create an algorithm to accurately predict how much a viewer would like a movie based on their preferences. Netflix has expanded the use of data beyond rating forecasting and into a variety of areas, including personalised ranking, page generation, search, picture selection, messaging, and marketing. Netflix Recommendation Engine (NRE), is composed of algorithms that select content based on each user's unique profile. The engine filters over 3,000 titles at once utilising 1,300 recommendation clusters based on user choices. The engine's precise recommendations account for 80% of the Netflix viewer activity. According to estimates, Netflix saves over $1 billion annually thanks to the NRE.
This project aims to cluster the video content available on Netflix based on the company's site data. Apart from aiding in the development of an efficient recommendation system, clustering the video content would also provide information about the type of content the company is interested in listing on its site. Thus giving an insight to content creators and filmmakers on the type of video content in demand.
In 2018, Flixable, a third-party Netflix search engine, released a report showing the number of TV shows on Netflix tripling since 2010. Apart from clustering the video content, a comprehensive exploratory data analysis will be performed during this project to understand the trends in the diverse video content catalogue and to verify the findings of Flexible, that Netflix has been focusing on producing TV shows in recent years.

3. Data Description

3.1. Data Preparation

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, providing content details about the video cast, director and duration corresponded with the site details like signing date, countries the content was produced in and topics the content was being listed under. The provided features are displayed in **Table. 1**.

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..

Table 1: Netflix Movies and TV Show Dataset Attributes

3.2. Exploratory Data Analysis

Exploratory Data Analysis (EDA) is the most commonly used methodology to summarize and visually analyze data using different parameters.

The Pandas package has been utilized in data handling while Matplotlib, Seaborn, and Wordcloud libraries have been used for data visualization.

For the preliminary analysis, the timelines of Netflix movies and TV shows were analyzed. The histogram plot of the number of TV shows and Movies premiered and signed on each year is displayed in **Fig. 1** and **Fig. 2**. From **Fig. 1**, we can observe that most movies streaming on the platform were released after 2010 while a large portion of the TV Shows streaming on the platform was released after 2015. The year 2017 had the highest number of movie and TV show releases on the platform.

From **Fig. 2**, we can observe that Netflix began adding videos to the platform in 2008 and started aggressively adding video content in 2017. It was also found that more stand-alone movies were added as compared to TV shows.

Fig. 3 and **Fig. 4** show the categorical distribution of the type and the rating of the video content streaming on Netflix. From these plots, we can observe that there are almost twice as many movies as TV shows on Netflix and that the majority of the content is rated for Mature Audiences and for audiences over 14 years old.

Fig. 5 and **Fig. 6** show the histogram plot of the duration of movies and TV show seasons on Netflix. These histogram plots show that the movies on Netflix have an average duration ranging from 90 to 110 minutes while the TV shows on Netflix have a span of only one season.

Fig. 7 shows the different genres available on Netflix. It is observed that the top video content genres on Netflix are Drama, Comedy, Documentary, Action and Adventure and Romance.

Fig. 8 shows the top video content-producing countries on Netflix. The top five biggest video content producers are the United States of America, India, the United Kingdom, Canada and France.

The insights from **Fig. 7** and **8** were combined to understand the type of content produced in different countries. The most popular genres produced by the top countries are displayed in **Fig. 9**. We can observe that Drama 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 across the platform which is also the leading genre in Japan. Romance is the most produced genre in South Korea. It is noted that Comedy was the most produced 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

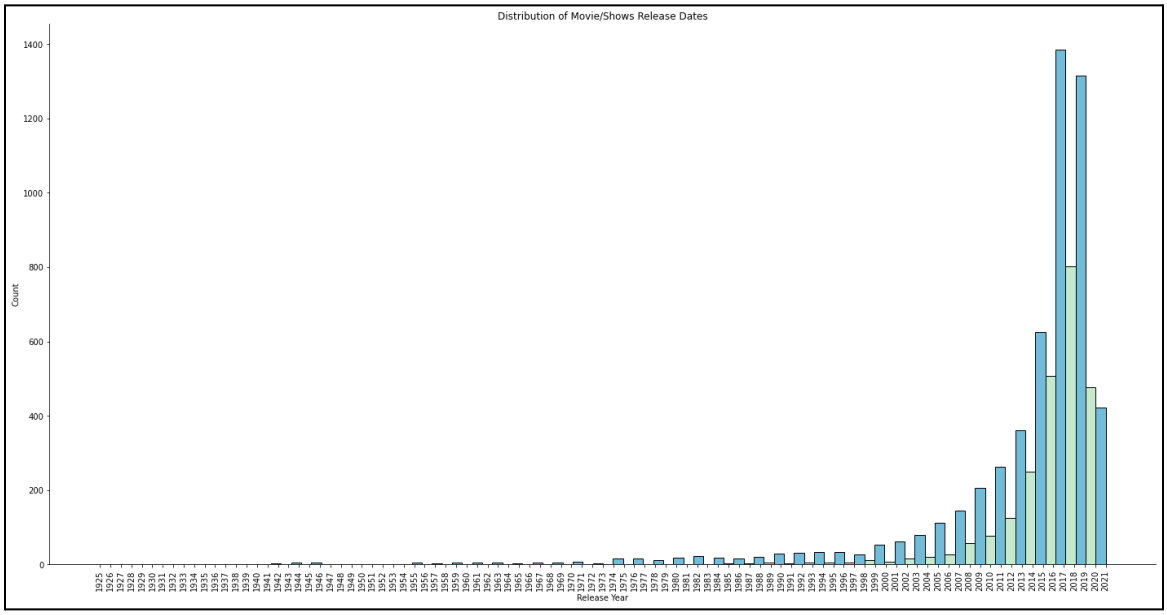

Figure 1: Number of Movies and TV Shows released by year

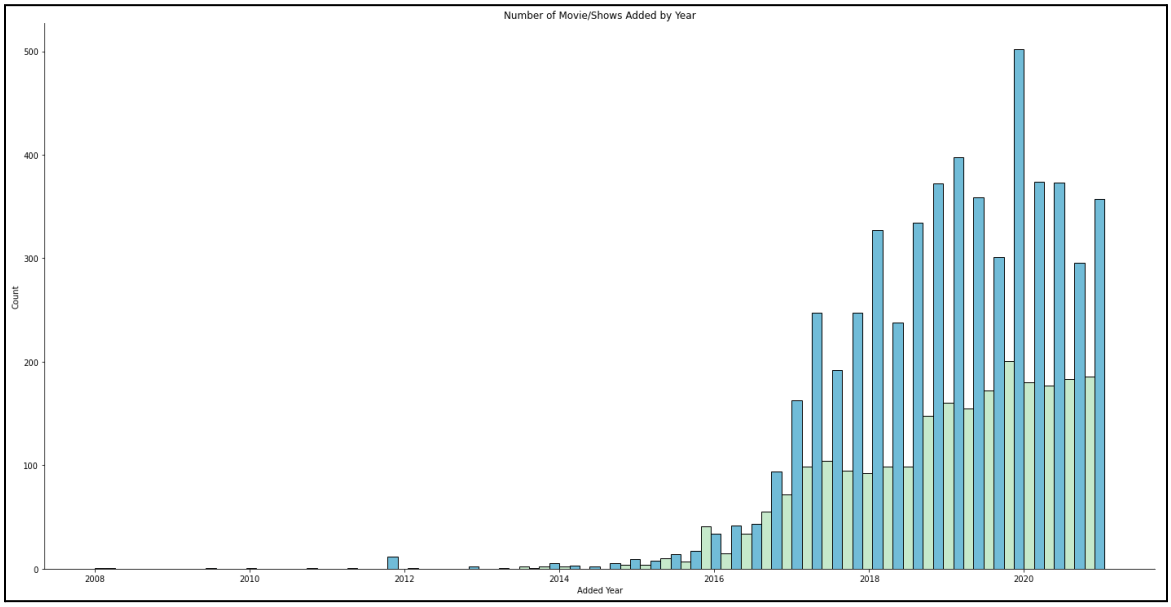

Figure 2: Number of Movies and TV Shows added by year

Fig. 1 shows the histogram plot of the number of movies and TV shows released each year (Blue depicting Movies and Green depicting TV Shows). **Fig. 2** shows the histogram plot of the number of movies and TV shows added by Netflix each year (Blue depicting Movies and Green depicting TV Shows)

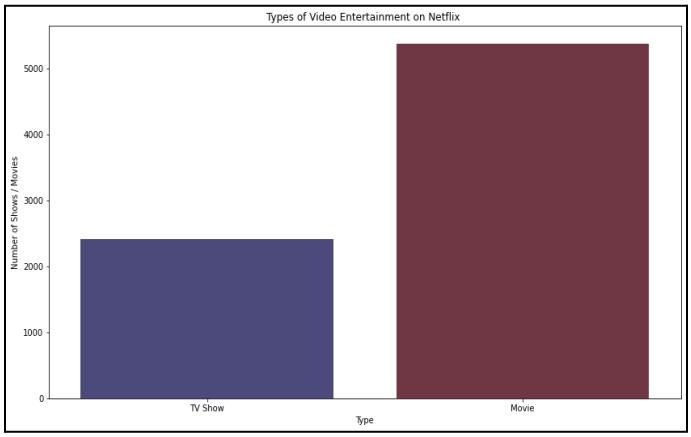

Figure 3: Types of Video Content on Netflix

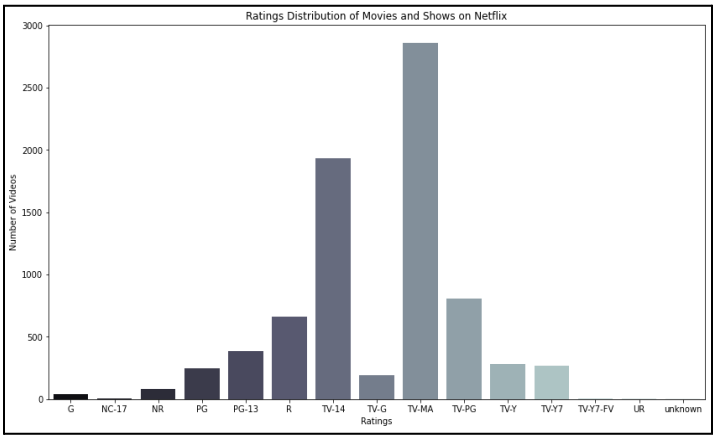

Figure 4: Video Content Ratings on Netflix

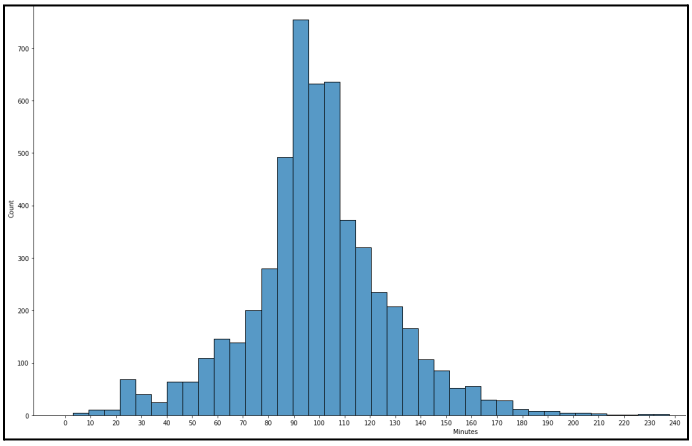

Figure 5: Histogram Plot of Movie Duration

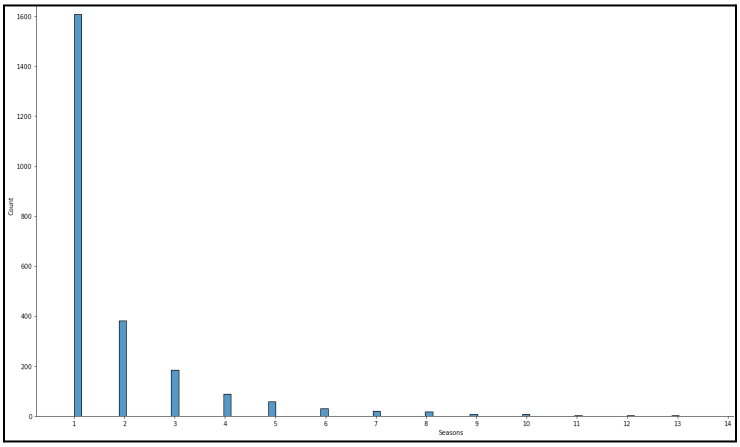

Figure 6: Histogram Plot of TV Show seasons

Fig. 3 depicts the categorical plot of video content type on Netflix. **Fig. 4** depicts the categorical plot of video content ratings on Netflix. **Fig. 5** shows the histogram plot of the duration of movies available on Netflix. **Fig. 6** shows the histogram plot of TV show seasons on Netflix.

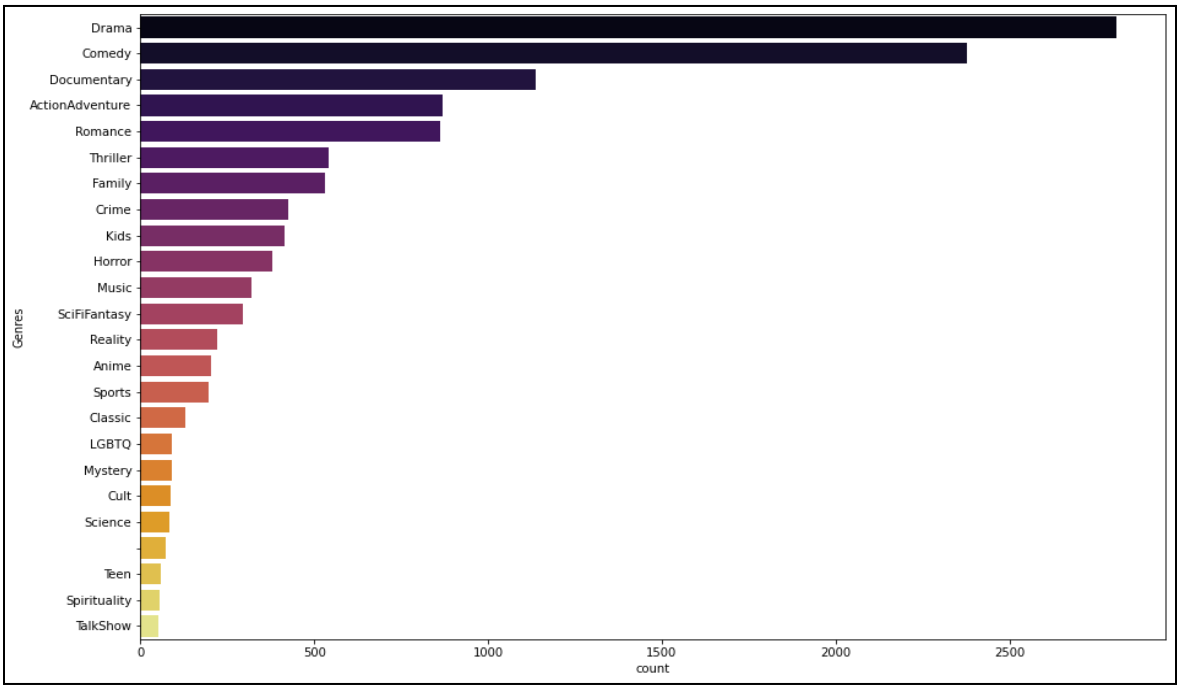

Figure 7: Video content genres available on Netflix

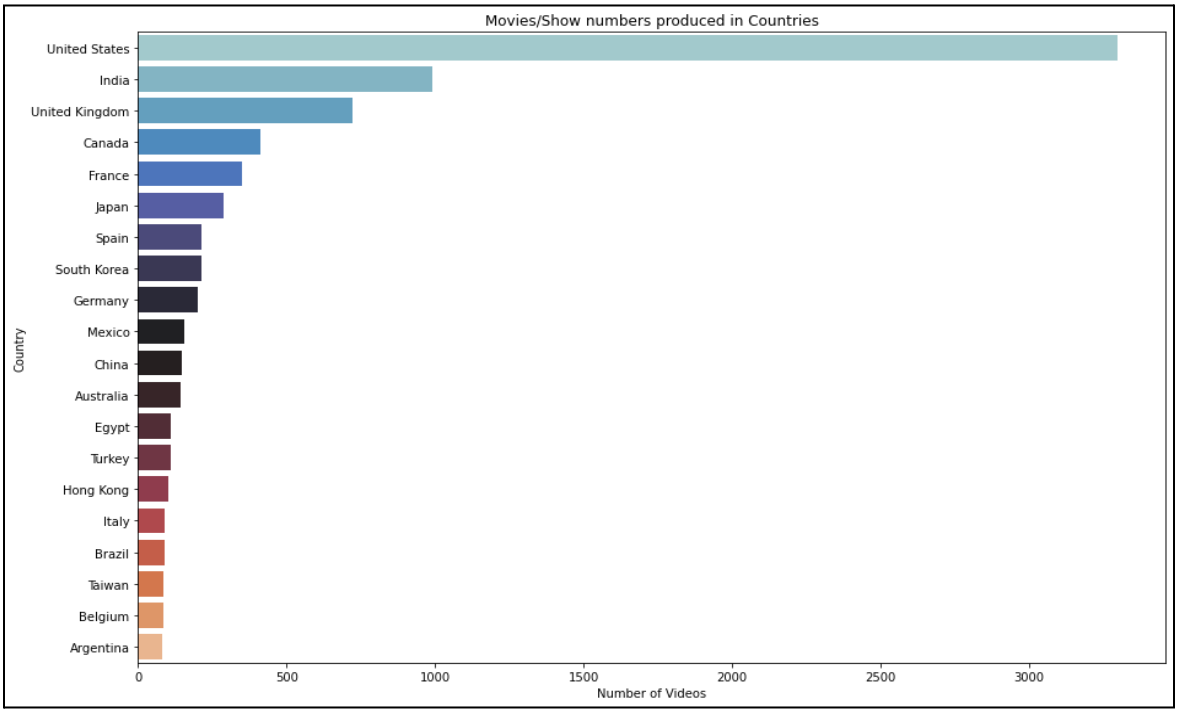

Figure 8: Top Netflix Video Content producing Countries

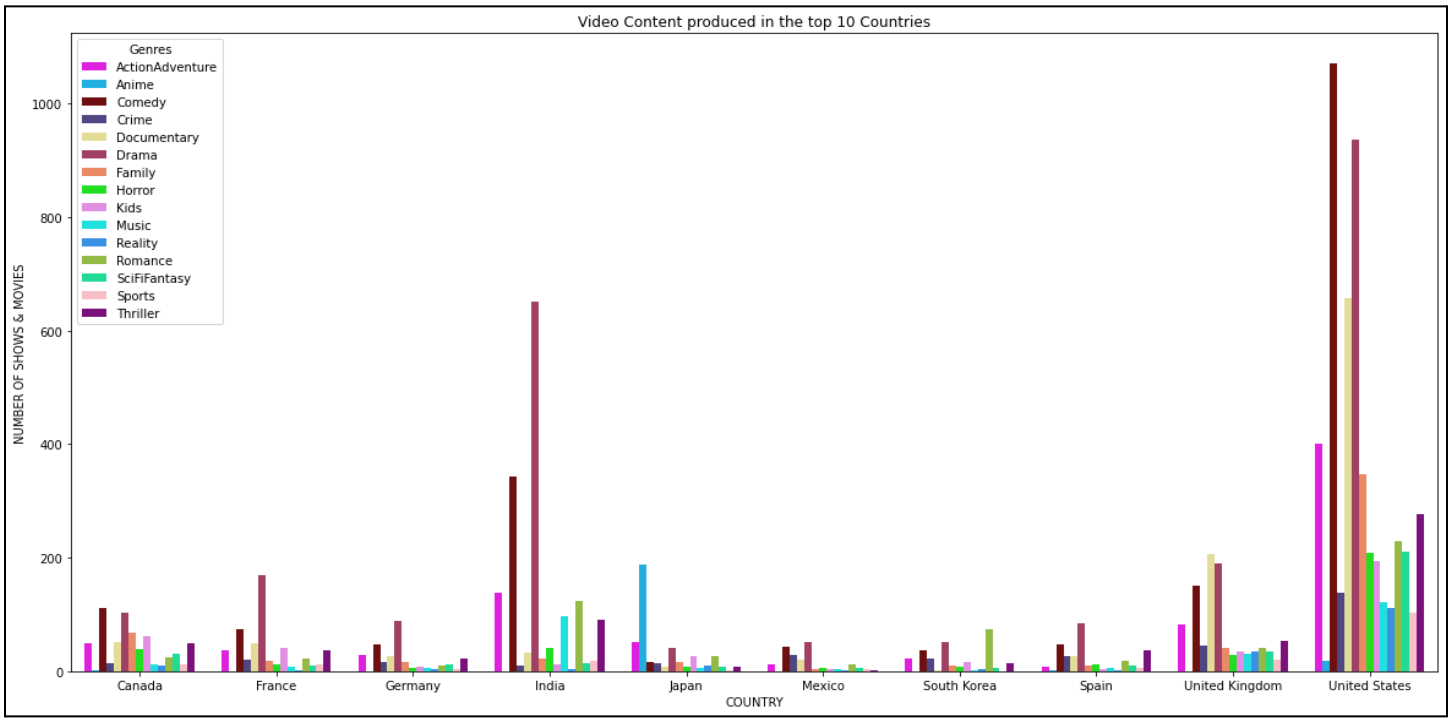

Figure 9: Top countries and their most produced genres

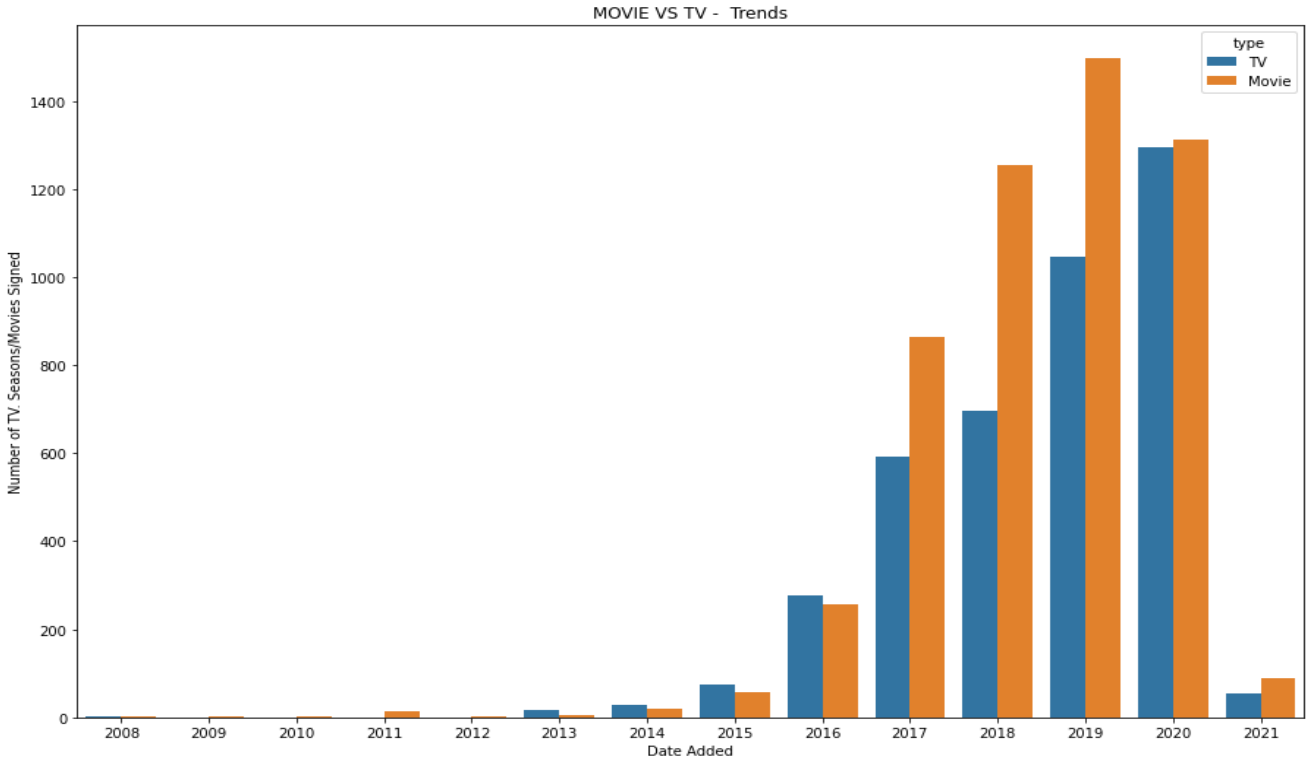

Figure 10: Adjusted TV show and movies added by year

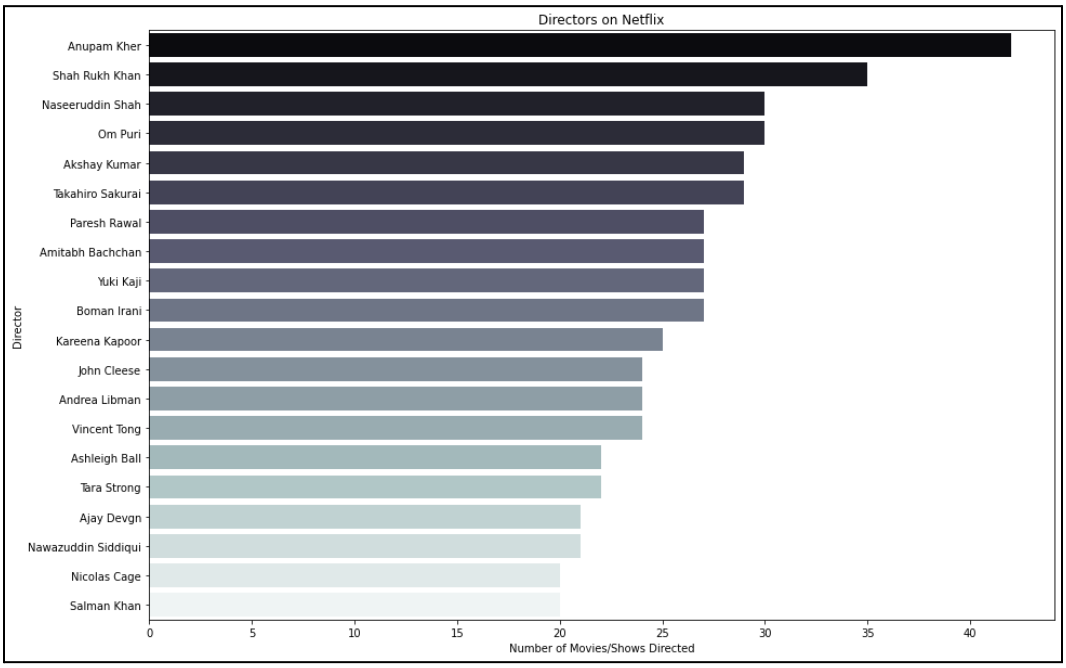

Fig. 10.2 : Most Recurring Directors on Netflix

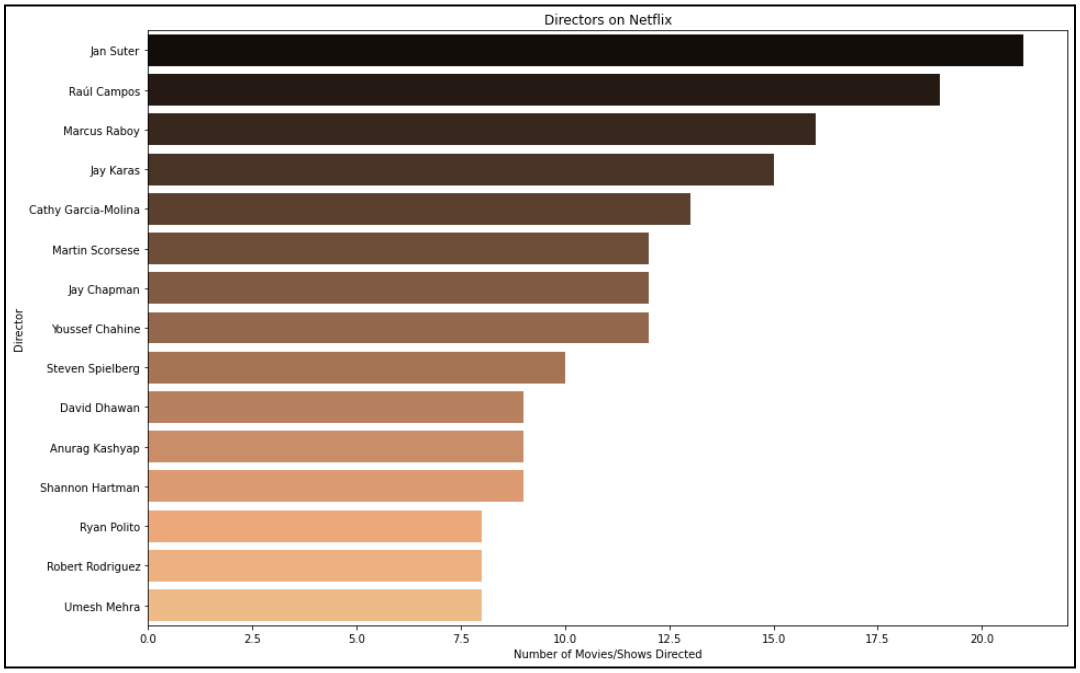

Fig. 10.3: Most Recurring Actors on Netflix

To check if Netflix has been focusing increasingly on TV shows, we'll need to consider each season of a TV show as its entity. This is because Netflix is known to renew seasons on an annual basis, which means that there is an ongoing financial commitment throughout the season's tenure.
Therefore, duplicates of TV shows are made corresponding to their seasons. Under the assumption that the filmmakers and Netflix sign their contracts with an idea of the seasonal tenure of the show, the adjusted TV Show and Movies added by the year is displayed in **Fig. 10**.
We can observe that the TV shows signed have been higher than the movies in 2016. While the movies signed have been higher ever since it is strikingly prominent that the TV shows signed per year 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.
The most recurring actors and directors on Netflix are displayed in **Fig.10.2 a** and **Fig.10.3**.

3.2. Feature Engineering

Null values were observed in '*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 of 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.
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 as a new attribute as '*Movie Deets*'.

3.3. Text Preprocessing

To handle text data, the text needs to be converted into a machine-interpretable form called a text vector. To facilitate the conversion into text vector, the text attributes are required to be cleaned to be converted efficiently. The steps involved in cleaning are

1. 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.
2. Punctuation Removal: All the punctuations from the text are removed. Popularly removed punctuations are mentioned in blue (*!'"#$%&'()*+,-./:;?@[\]^_`{|}~*).
3. 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.
4. 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.

The aforementioned preprocessing steps are implemented on the '*description*' attribute and the previously derived '*Movie Deets*' attribute. The unigram and bi-gram word clouds of content description in top genres are displayed in **Fig. 11a** and **Fig. 11b**.

3.2. Feature Selection

To cluster movies and TV shows on Netflix, 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 the name, ID and year of the signing of the video content 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 version of text attributes '*description*' and '*Movie Deets'*. Vectorising the preprocessed attributes contributes a total of 2318 dimensions. For computational ease, these dimensions will have to be reduced using PCA. Upon plotting the cumulative explained variance chart (**Fig. 11)**, it was found that 700 components were required to explain 75 per cent of the variance. This is not a fair compromise as it is still computationally taxing with a 25 per cent loss in information.

Alternatively, it was decided that the two attributes could be used to model video content into topics using Latent Dirichlet Allocation. This would make sure that all the topical information about video content is captured without putting any available information to waste. Apart from this, it also entertains the possibility of a video exhibiting multiple themes at different extents by expressing the probability a document belongs to a given topic.

4. 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, in turn, modelled as an infinite mixture over an underlying set of topic probabilities. In the context of text modelling, the topic probabilities provide an explicit representation of a document.

5. Methodology

The performance of three unsupervised machine learning algorithms is evaluated and compared to cluster Netflix movies and TV shows.

5.1. Density-Based Spatial Clustering of Applications with Noise

DBSCAN is used to identify distinctive groups/clusters in the data, based on the idea that a cluster in data space is a contiguous region of high point density, separated from other such clusters by contiguous regions of low point density. It is a base algorithm for density-based clustering. It can discover clusters of different shapes and sizes from a large amount of data with the presence of noise and outliers. DBSCAN algorithm requires two parameters:

1. *eps*: It defines the neighbourhood around a data point i.e. if the distance between two points is lower or equal to '*eps*' then they are considered neighbours. If the *eps* value is chosen too small then a large part of the data will be considered as outliers. If it is chosen very large then the clusters will merge and the majority of the data points will be in the same clusters. One way to find the *eps* value is based on the k-distance graph.
2. *MinPts*: Minimum number of neighbours (data points) within eps radius. The larger the dataset, the larger value of MinPts must be chosen. As a general rule, the minimum *MinPts* can be derived from the number of dimensions D in the dataset as, *MinPts* >= D+1. The minimum value of *MinPts* must be chosen at least 3.

The algorithm proceeds by arbitrarily picking up a point in the dataset (until all points have been visited). If there are at least '*MinPts*' points within a radius of '*eps*' to the point then we consider all these points to be part of the same cluster. The clusters are then expanded by recursively repeating neighbourhood calculations for each neighbouring point.

5.2. K-Means Clustering

K-Means clustering is an iterative algorithm that divides the unlabeled dataset into K unique clusters where each dataset belongs to only one group having similar properties. It is a centroid-based algorithm, where each cluster is associated with a centroid. The k-means clustering algorithm determines the best value for K centre points or centroids by an iterative process of assigning each data point to its closest k-centre. The data points closest to the k-centre create a cluster. Hence each cluster has data points exhibiting commonalities while being farther from other clusters.
The K-Means clustering algorithm works by

1. Assuming K points or centroids at random
2. Each data point is assigned to its closest centroids forming K clusters.
3. The variance of data points is calculated and a new centroid is placed for each cluster.
4. Steps 3 and 4 are repeated until the data points are assigned to the same K centroids.

5.3. Agglomerative 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 diagram called dendrogram is graphed to represent this hierarchy. It is an inverted tree that describes the order in which factors are merged (bottom-up view).
The algorithm for Agglomerative Hierarchical Clustering is

1. Calculate the similarity of one cluster with all the other clusters (calculate proximity matrix)
2. Consider every data point as an individual cluster
3. Merge the clusters which are highly similar or close to each other.
4. Recalculate the proximity matrix for each cluster
5. Repeat Steps 3 and 4 until only a single cluster remains.

6. Evaluation Metrics and Plots

In order to quantitatively evaluate the performance of the clustering model, clustering metrics have been used to gauge the model performance. As the data is not provided with the ground truth, we'll be using intrinsic evaluation methods to measure the clustering quality. Apart from clustering metrics, other metrics and methods performed to evaluate the Topic Model and choose optimal clusters have been discussed in this section.

6.1. Silhouette Score

The Silhouette Score and Silhouette Plot are 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.
This measure has a range of -1 to 1 and serves as a great tool to visually inspect the similarities within clusters and differences across clusters. The silhouette score is given by,

$$s(i) = \frac{b(i) - a(i)}{\max\left(a(i), b(i)\right)}$$

Where *b(i) is the* minimum average distance from the data point, '*i'* to all the other clusters and *a(i)* is the average distance between '*i'* and all other points in its cluster. The typical Silhouette Plots represent the cluster label on the y-axis, while the actual Silhouette Score is on the x-axis. The size/thickness of the silhouettes is also proportional to the number of samples inside that cluster. The higher the Silhouette Coefficients (the closer to +1), the further away the cluster's samples are from the neighbouring clusters' samples. A value of zero

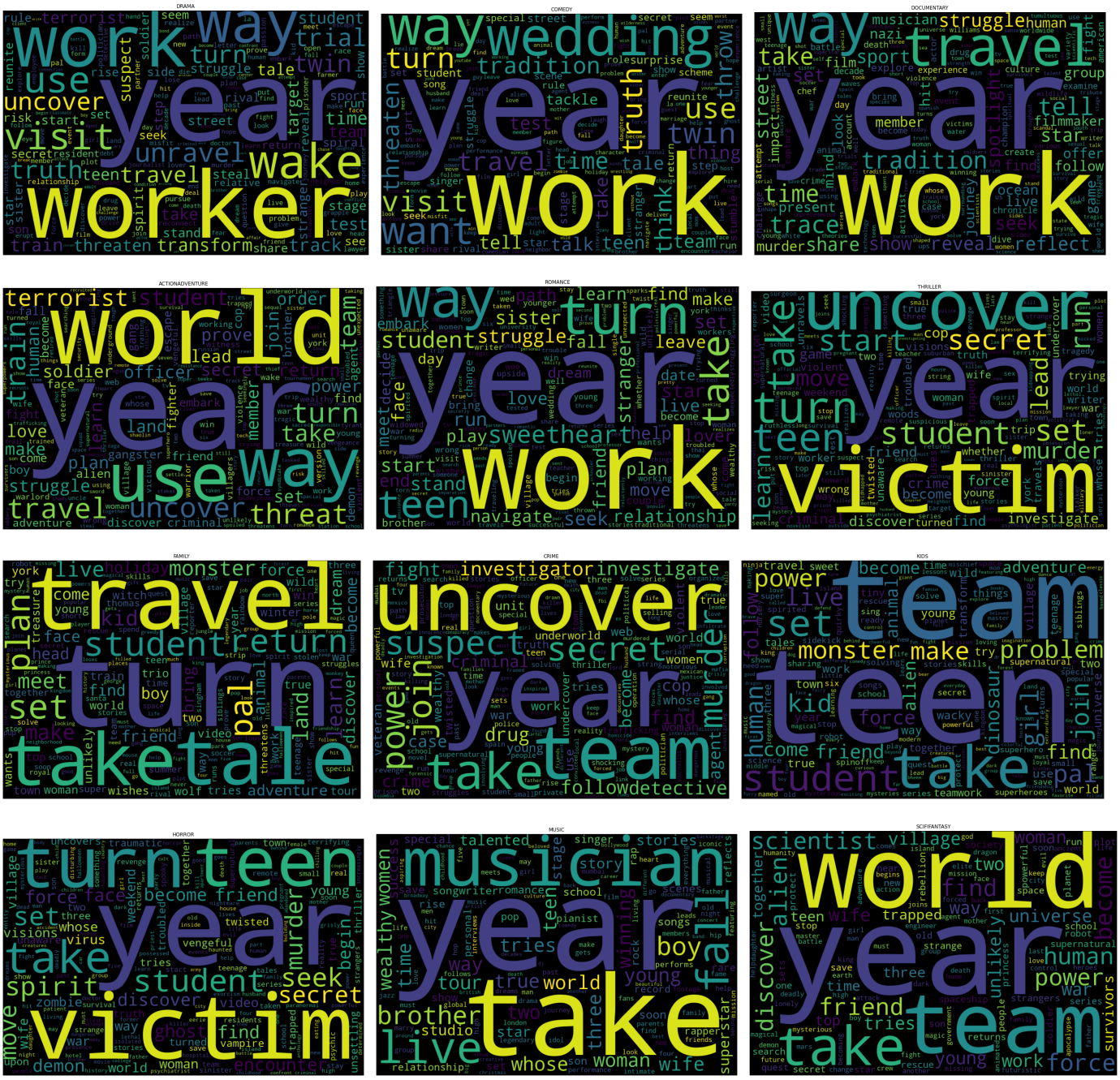

Fig 11 a: Wordclouds of the top video content genres in Netflix. (From left to right: Row 1:Drama, Crime and Comedy. Row 2: Action-Adventure, Romance and Thriller. Row 3: Family, Crime and Kids. Row 4: Horror, Music and SciFi-Fantasy)

Fig 11 b: Bigram Wordclouds of the top video content genres in Netflix. (From left to right: Row 1:Drama, Crime and Comedy. Row 2: Action-Adventure, Romance and Thriller. Row 3: Family, Crime and Kids. Row 4: Horror, Music and SciFi-Fantasy)

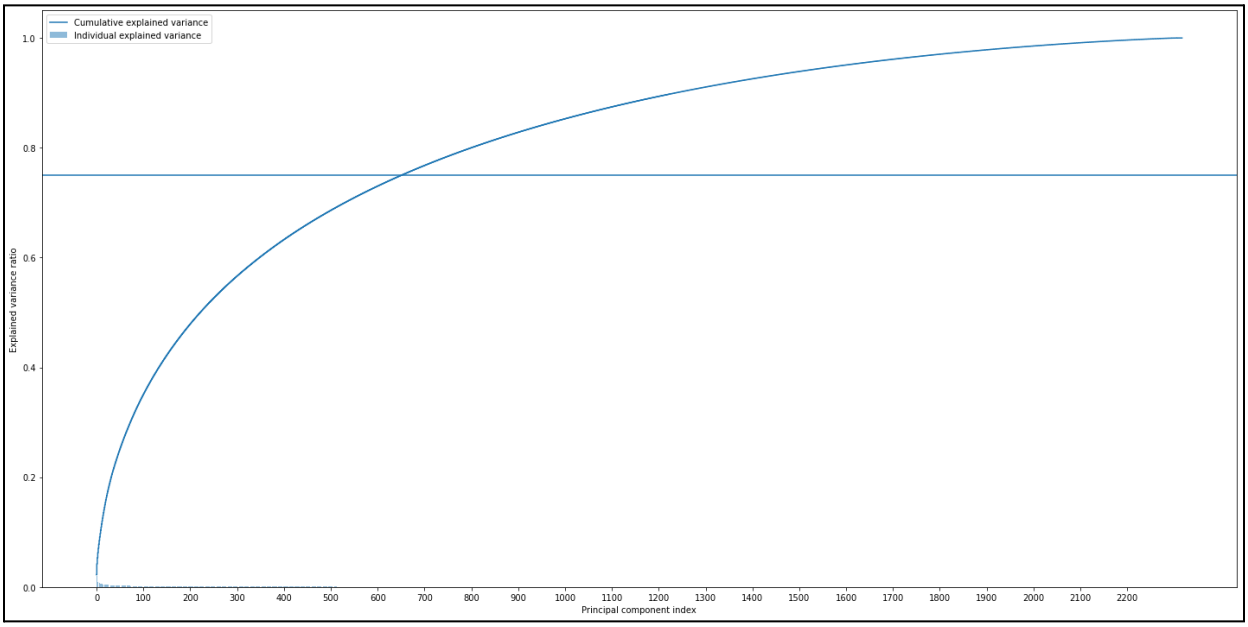

Fig. 12: Explained Variance Plot of Text Vector

indicates that the sample is on or very close to the decision boundary between two neighbouring clusters. Negative values, instead, indicate that those samples might have been assigned to the wrong cluster. Averaging the Silhouette Coefficients, we can get to a global Silhouette Score which is used to describe the entire population's performance with a single value.

6.2 Davies-Bouldin Index

The 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. In this way, clusters which are farther apart and less dispersed will lead to a better score.
The minimum score is zero, and differently from most performance metrics, the lower values the better clustering performance. Similar to the Silhouette Score, the D-B Index does not require ground-truth labels and has a simpler implementation in comparison to the Silhouette Score.

6.3 Calinski-Harabasz Index

Calinski-Harabasz Index is also known as the Variance Ratio Criterion. 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. C-H index is given by,

$$s = \frac{\mathrm{tr}(B_k)}{\mathrm{tr}(W_k)} \times \frac{n_E - k}{k - 1}$$

where tr(Bk) is the trace of the between-group dispersion matrix and *tr(Wk)* is the trace of the within-cluster dispersion matrix.

6.4 Coherence Score

Topic Coherence measures the score of a single topic by measuring the degree of semantic similarity between high-scoring words in the topic. These measurements help distinguish between topics that are semantically interpretable and topics

that are artefacts of statistical inference. This metric is chosen as it closely represents human impression unlike metrics like perplexity or log-likelihood.

6.5 Elbow Method

The elbow method is used to determine the optimal number of clusters in k-means clustering. The elbow method plots the value of the cost function produced by different values of k. If k increases, average distortion will decrease, each cluster will have fewer constituent instances, and the instances will be closer to their respective centroids. However, the improvements in average distortion will decline as k increases. 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.

6.6 Dendrogram Method

Dendrograms are a diagrammatic representation of the hierarchical relationship between the data points. They illustrate the arrangement of clusters produced by the corresponding analyses and 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 and it defines the associated cluster membership for each data sample

7. Model Development

In order to find the optimal number of clusters, a range of clusters were assigned for models and their corresponding scores were evaluated using elbow curves and Silhouette plots for K-means clustering while dendrograms and an iterative paradigm calculating silhouette scores for varying distances were used to optimise the clustering quality

Hierarchical Clustering method. As DBSCAN does not require any explicit cluster numbers to perform clustering, the default parameters were deemed to be sufficient.

8. Results and Discussion

8.1 Topic Modelling

Multicore Latent Dirichlet Allocation model offered by the Gensim library was used to model the description, genres, country, director and cast documents from the dataset. Coherence Score (*c_npmi*) was the opted metric to evaluate the topic modelling of the LDA model. The best quality of allocation was observed when modelled nine topics producing a score of 0.02. The relevant topical words for each topic are displayed in **Table 2** and their corresponding word clouds are in **Fig. 13**.

8.2 Clustering

The final dataset to perform clustering was created upon scaling non-textual data using the Min-Max scaler and concatenating the resulting table with the topic probability table obtained from the LDA model. DBSCAN clustered the content into 9 clusters with a silhouette score of 0.4664, Davies-Bouldin Index of 1.62 and Calinski-Harabasz Score of 2510.76. The visualisation of the clusters among the top three principal components of the input dataset is plotted in **Fig 14**.
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-Harabasz Score of 2901.84. The silhouette plot, elbow curve and visualisation of the top three principal components of the input dataset are plotted in **Fig 15, 16** and **17**.
For Hierarchical clustering, 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-Harbasz

Table 2: Modelled Topics and Relevant Top Words

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

Fig. 13: Word clouds of Topics 1 to 9 (from left-right)

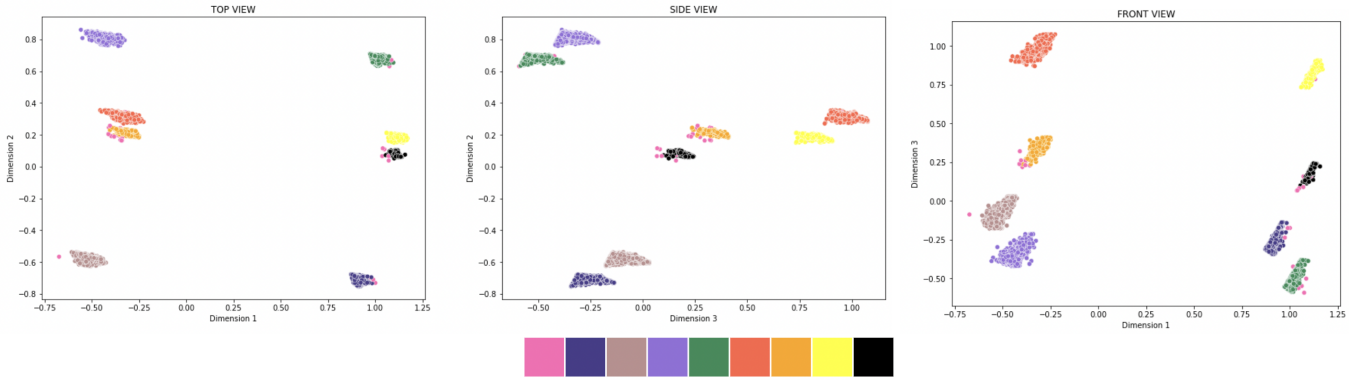

Fig. 14: DBSCAN Cluster Visualisation for 3 Principal Components (colour palette indicates legend for cluster)

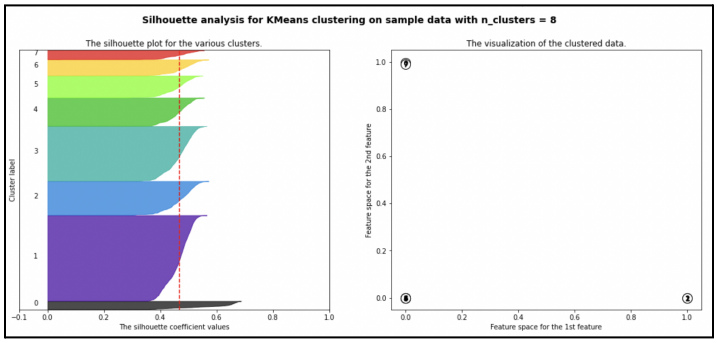

Fig. 15: K-means Silhouette Plot for eight clusters

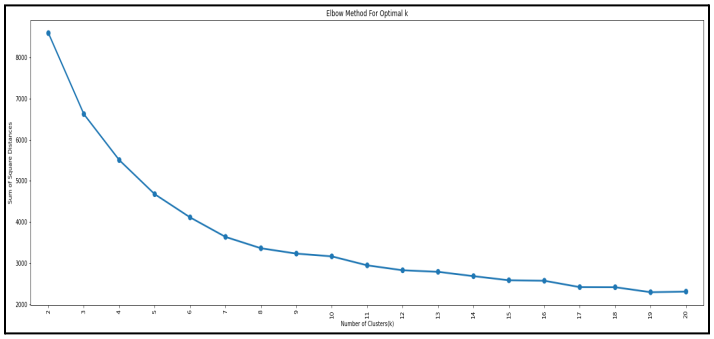

Fig. 16: K-means Elbow Curve for Optimal Clusters

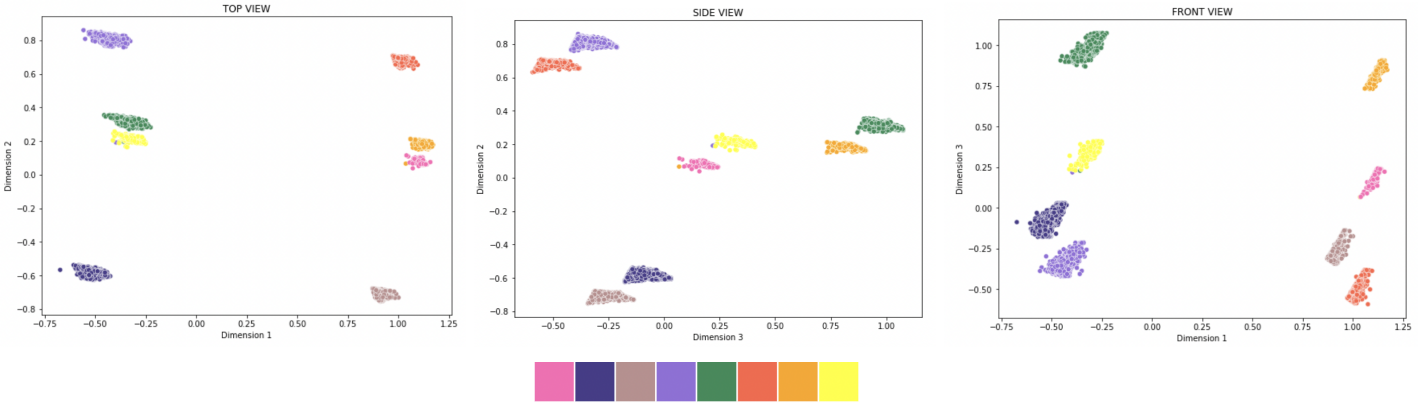

Fig. 17: K-Means Cluster Visualisation for 3 Principal Components (colour palette indicates legend for cluster)

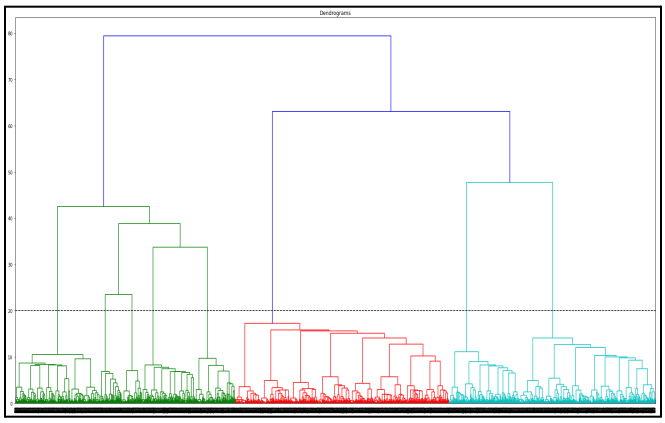

Fig. 18: Dendrogram for Hierarchical Clustering

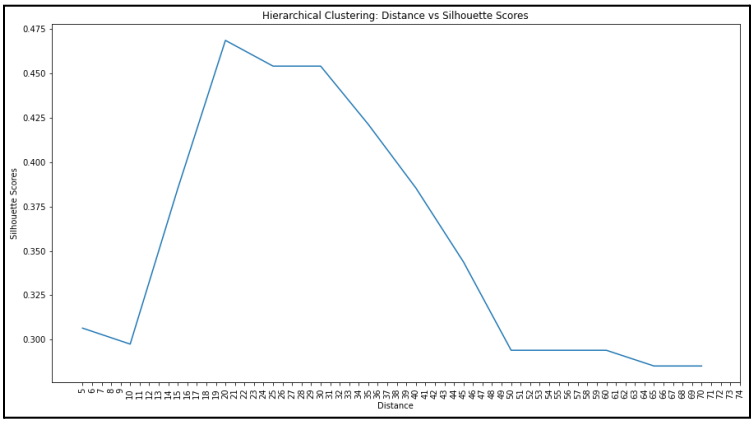

Fig. 19: Distance-Silhouette Score Line Plot

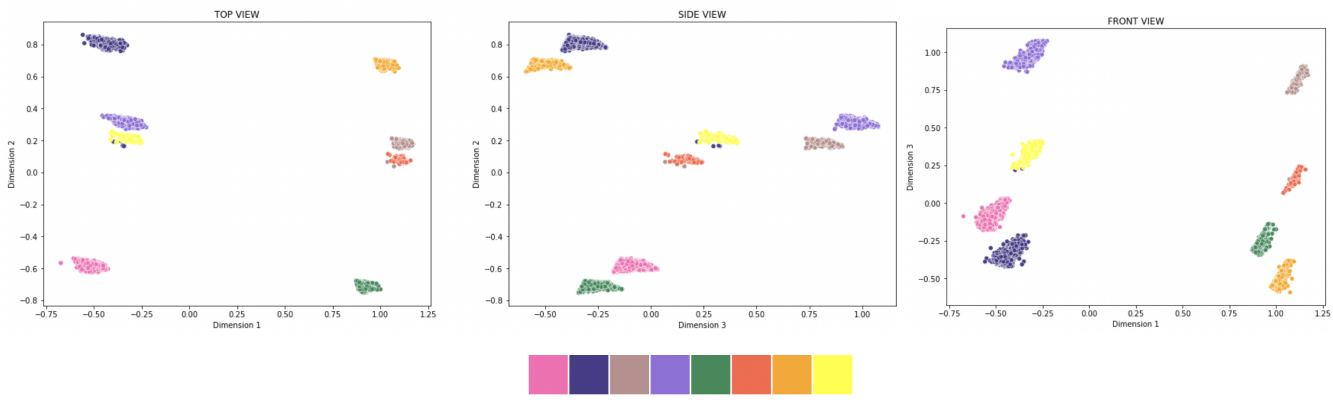

Fig. 17: Hierarchical Cluster Visualisation for 3 Principal Components (colour palette indicates clusters)

Algorithm	Clusters	Silhouette Coefficient	Davies-Bouldin Index	Calinski-Harabasz 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

Table 3: Clustering Evaluation Report for DBSCAN, K-Means and Hierarchical Clustering

Score of 2900.28. The dendrogram, distance-silhouette score graph and cluster visualisation for the three principal components are depicted in **Fig 18, 19** and **20**.

9. Conclusions

We can now address the four sections of the problem statement effectively. In the first section, we have performed a comprehensive exploratory data analysis and found the content expansion trends and timelines, the video genre and rating distribution, and the average duration of the content types 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.

The use of a combination of topic models to process text data has aided in clustering movies and TV shows on Netflix with the best performing models, K-Means and Hierarchical Clustering, producing eight clusters with a silhouette score of 0.47.

Apart from helping develop recommender systems, this labelled content can also be studied and explored to determine what type of content is on-demand, potentially providing an intuition to content creators about the type of content Netflix is interested in listing on its catalogue.

10. References

1. DBSCAN Clustering Algorithm in Machine Learning
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|