CAPSTONE PROJECT - 4

Netflix Movies & TV Shows Clustering

COHORT - OSLO

TEAM MEMBER:

ANUP A. JAMBULKAR
VIBHU SHARMA
GAURAV MALAKAR
ANKIT WALDE
ANIL BHATT

Contents:

- Introduction.
- Dataset Preview.
- Exploratory Data Analysis.
- Data Preprocessing.
- Creating Clusters.
- Conclusions.

Introduction:

Netflix is a media distribution company. It started with DVD distribution via mail, but has evolved substantially over the course of its existence. Today, Netflix is focused on streaming video. Some of its content is licensed, and some of the content is produced in-house.

Netflix originally focused on movies, but today television shows are probably the more common format. Netflix works on a subscription model, where users get unlimited access to content with a paid subscription.

Dataset Preview:

This dataset consists of tv shows and movies available on Netflix as of 2019. The dataset is collected from Flixable which is a third-party Netflix search engine.

Attribute Information

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

Dataset summary:

The dataset contains 12 columns and 7787 rows.

There also exist some null values in our data:

Percentage of null values in director: **30.68%**

Percentage of null values in cast : **9.22%**

Percentage of null values in country : **6.51%**

Percentage of null values in date_added : **0.13%**

Percentage of null values in rating : **0.089%**

Since there are null values in the above variables *we will replace them with 0.*

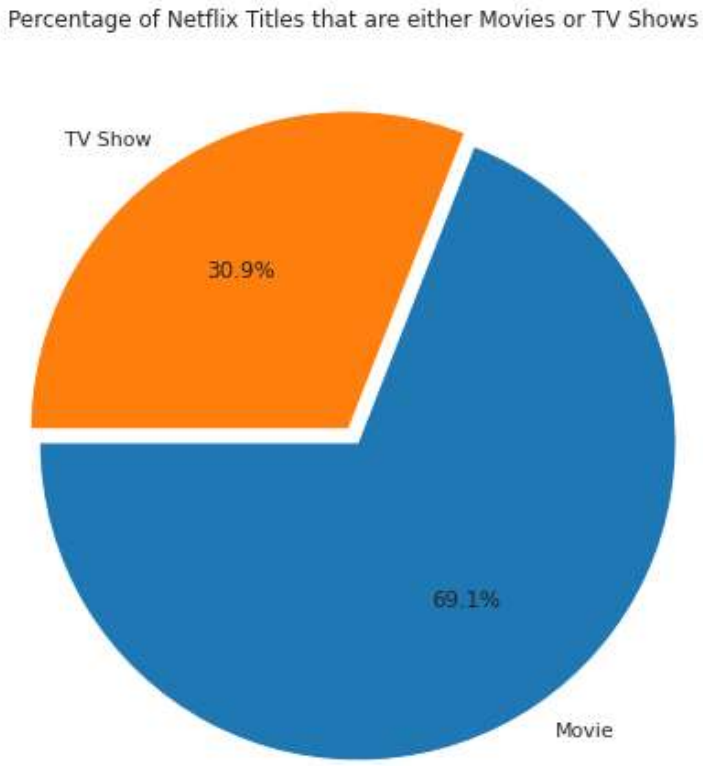
Finally, we will do some feature engineering to create few new variables:

- *Compute* ***year_added****,* ***month_added*** *and* ***day_added*** *from* ***date_added*** *after converting it into a datetime variable.*

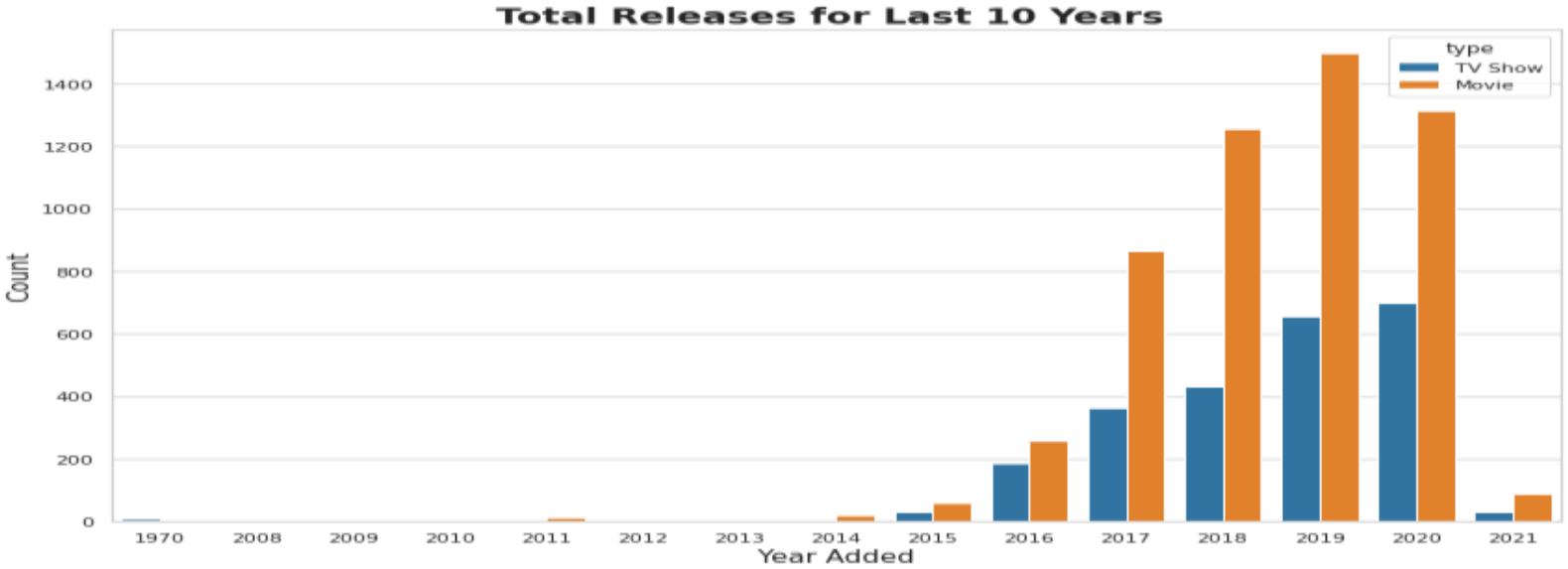
Exploratory Data Analysis:

Type:

69.1% of the content available on Netflix are movies; the remaining 30.9% are TV Shows.



Year_added:

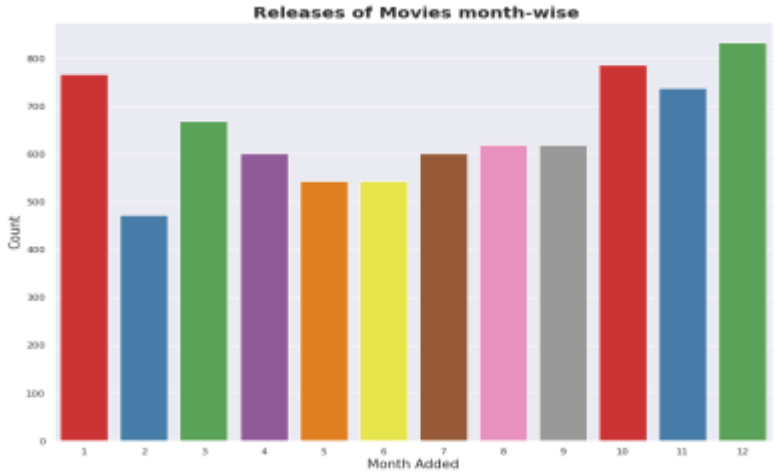


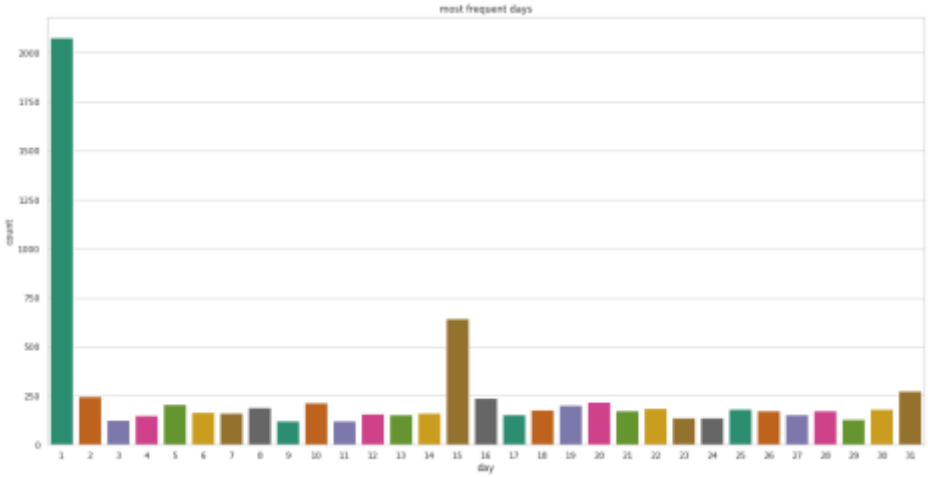
- Growth in the number of movies on Netflix is much higher than tv shows.
- From 2015 we can see a noticeable addition in the number of movies and tv shows uploaded by Netflix on its platform.
- The highest number of movies and tv shows got added in 2019 and 2020.
- The line plot shows very few movies, and tv shows got added in 2021. It is due to very little data collected from the year 2021.

Month_added and Day_added:

- Most of the content is uploaded either by year ending or beginning.
- October, November, December, and January are months in which many shows and movies get uploaded to the platform.
- It might be due to the winter, as in these months people may stay at home and watch shows and movies in their free time.

- Most of the content is uploaded at the beginning, middle, or the end of a month.
- Which makes 1st, 15th or 31st of a month more prominent in getting new tv shows and movies.





Country:

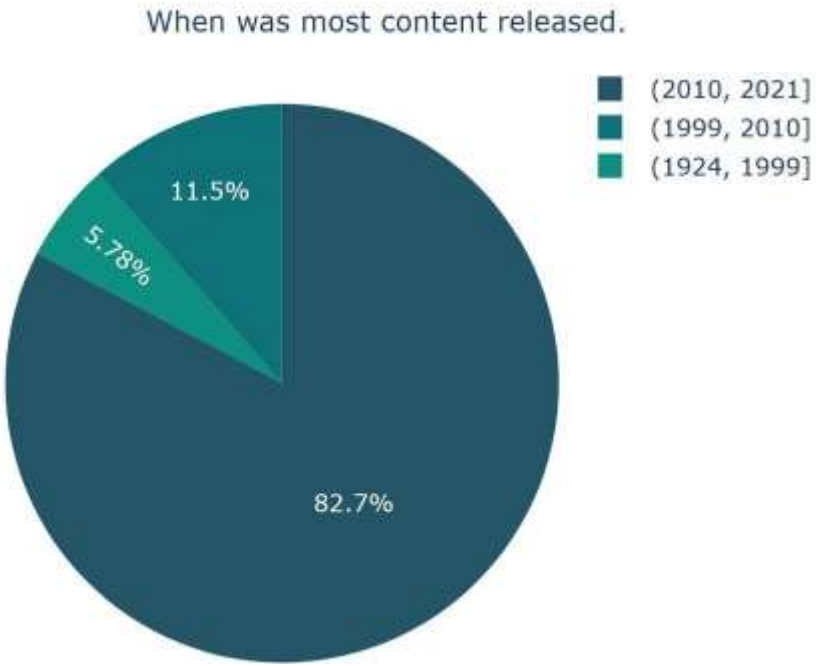
	country	count
0	United States	2555
1	India	923
2	0	507
3	United Kingdom	397
4	Japan	226
5	South Korea	183
6	Canada	177
7	Spain	134
8	France	115
9	Egypt	101
10	Mexico	100

<Figure size 1440x864 with 0 Axes>

- The majority of the content providers are in the above top-ten countries.

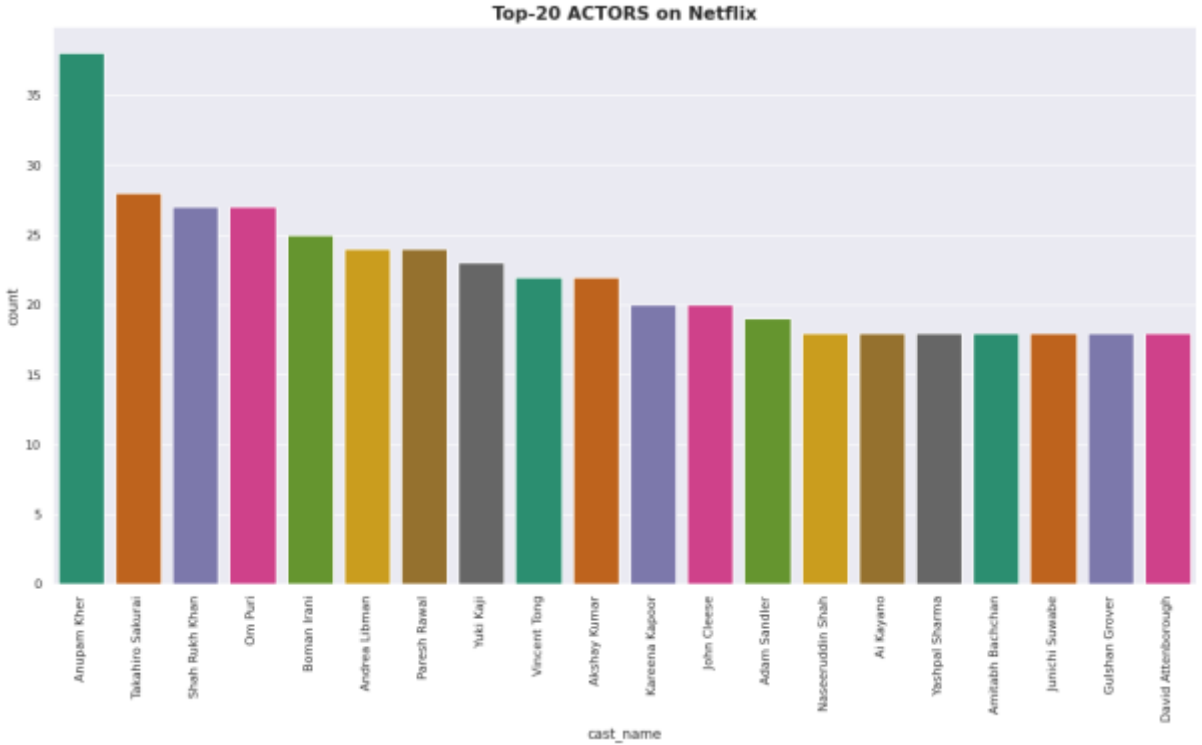
- Among which USA, India, and Uk create more than half of the tv shows and movies on the platform.

Release_year:



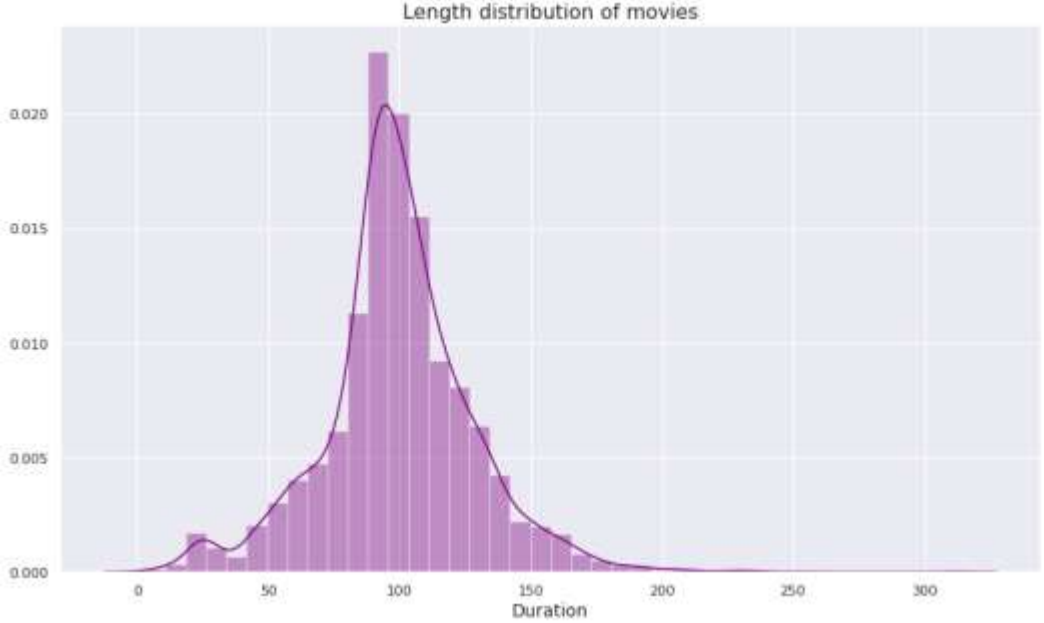
- 82% of the content available was released between 2010 and 2021.
- 17.28% of the content available was released before 2010.

Cast:



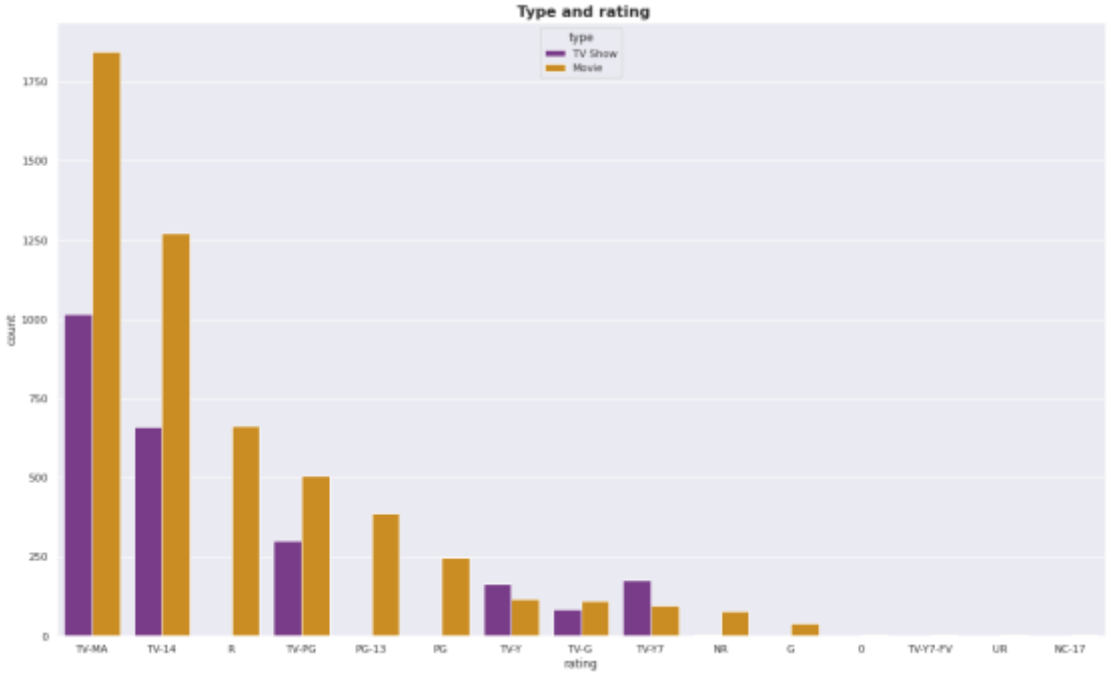
- Six of the actors in the top ten list with most numbers tv shows and movies are from India.
- With Anupam Kher at the top with 38 tv shows and movies in total.

Duration:



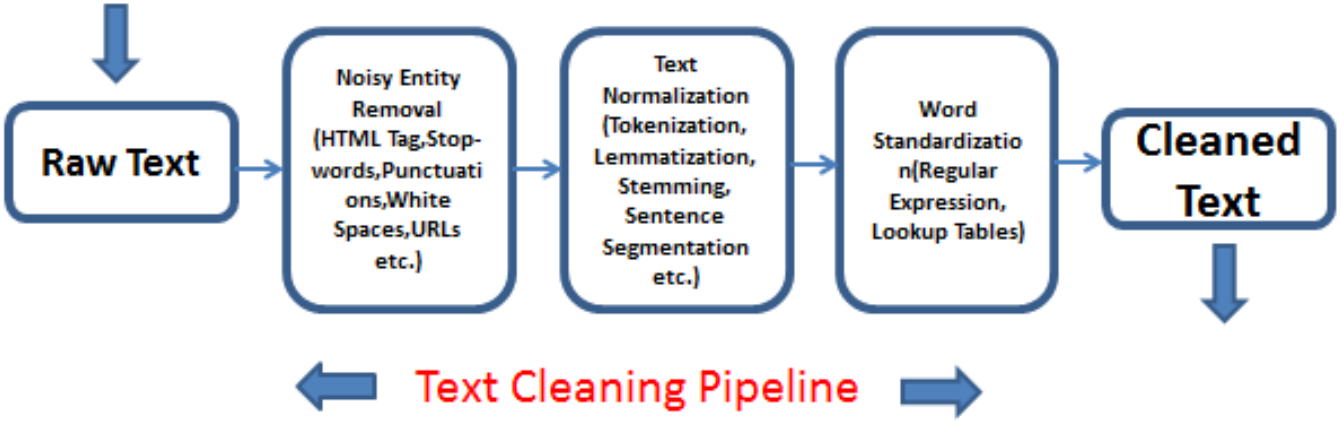
- Most of the movies last for 90 to 120 minutes.

Ratings:



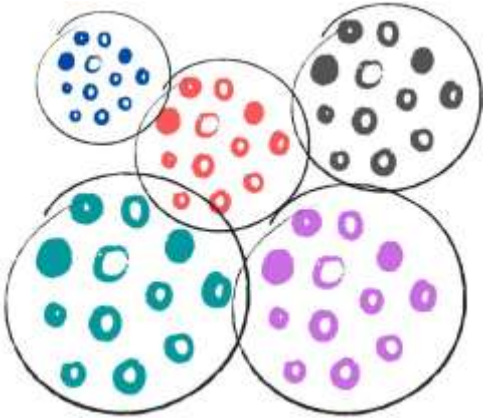
- TV-MA tops the charts, indicating that mature content is more popular on Netflix.
- This popularity is followed by TV-14 and TV-PG, which are Shows focused on Teens and Older kids.
- Very few titles with a rating NC-17 exist. It can be understood since this type of content is purely for the audience above 17.

Data Preprocessing.



Creating Clusters:

What is clustering?

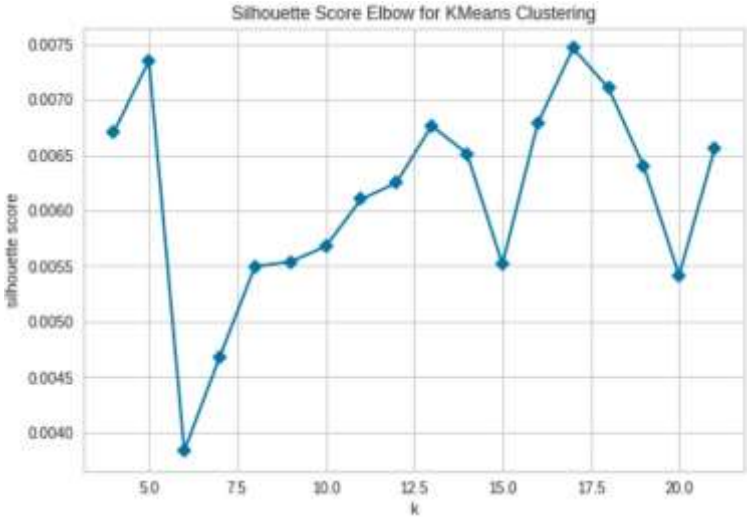


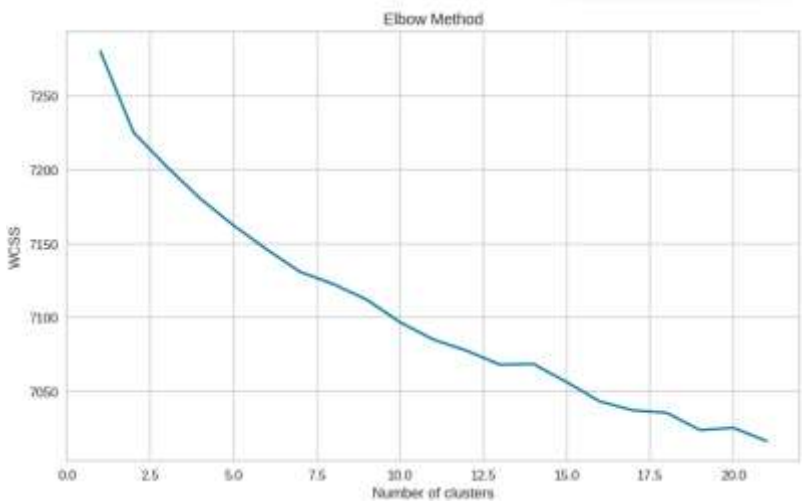
Clustering is the task of dividing the data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them.

How to cluster similar data?

To create clusters we will use the K-Means Clustering; which is an iterative process in which the dataset is grouped into k number of predefined non-overlapping clusters or subgroups, making the inner points of the cluster as similar as possible while trying to keep the clusters at distinct space it allocates the data points to a cluster so that the sum of the squared distance between the clusters centroid and the data point is at a minimum.

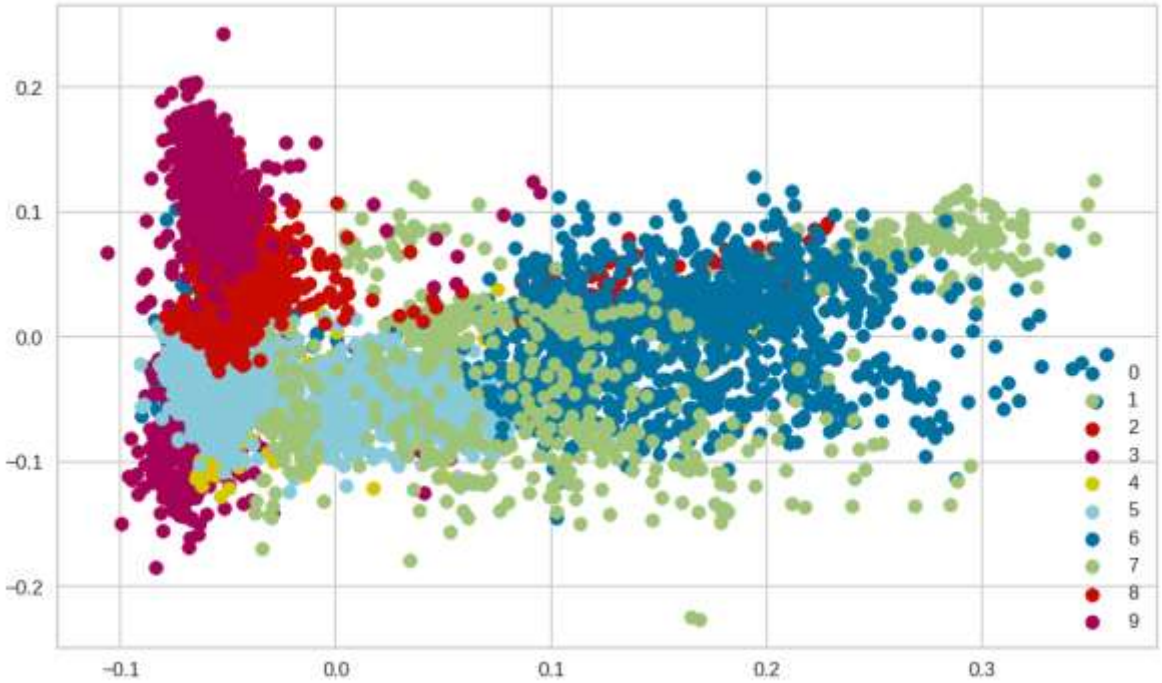
Determining optimal value for k:





- Using the Silhouette Score and Elbow Method we select the optimal number of clusters to be 10.

10 Distinct clusters created using kMeans Clustering



- The numbers 0 to 9 represent 10-distinct clusters formed by K-means clustering.
- Each cluster contains data points similar to those in the same groups but varies from other groups.

Data represented by each cluster:

Cluster 0: Family and Children Movies.

Cluster 1: Korean and Romantic Tv Shows.

Cluster 2: International Movies and Tv Shows.

Cluster 3: Musical Movies and Documentaries.

Cluster 4: Stand Up Comedy and Comedy Shows.

Cluster 5: Action, Adventure, and Independent Movies.

Cluster 6: Science, Nature, Reality.

Cluster 7: Crime Tv Shows and Docuseries,

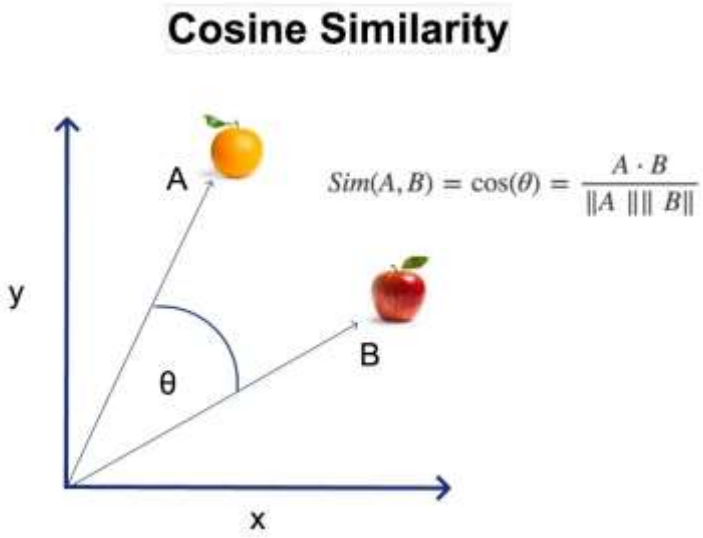
Anime Series.

Cluster 8: International Comedy Shows.

Cluster 9: International Tv Shows.

Getting Recommendations:

We obtained recommendations for Movies and Tv- Shows using Cosine similarity.




```
# Lets try getting recommendations for Movies
movie_recommendations = pd.DataFrame(recommen
movie_recommendations.head(11)
```

	Recommendations
0	Bad Boys II
1	GoldenEye
2	Tortilla Soup
3	Martin Lawrence Live: Runteldat
4	War on Everyone
5	Madam Secretary
6	Slow West
7	Tremors 5: Bloodline
8	Dollar
9	Operation Odessa

```
# Lets try getting recommendations for Tv-Shows.
tvshows_recommendations = pd.DataFrame(recommendations('13 Re
tvshows_recommendations.head(11)
```

	Recommendations
0	13 Reasons Why: Beyond the Reasons
1	Mind Game
2	The Sinner
3	Disappearance
4	Unsolved Mysteries
5	Anjaan: Special Crimes Unit
6	Frequency
7	Re:Mind
8	Gigantosaurus
9	The Staircase

Conclusions:

- It was interesting to find that majority of the content available on Netflix is Movies.
- But in the recent years it has been focusing more on Tv-Shows.
- Most of these contents are released either in the year ending or the beginning.
- United States and India are among the top 5 countries that produce all of the available content on the platform.
- Also 6 of the actors among the top ten actors with maximum content are from India.
- TV-MA tops the charts, indicating that mature content is more popular on Netflix.
- k=10 was found to be an optimal value for clusters using which we grouped our data into 10 distinct clusters.
- Using the given data a simple recommender system was created using cosine_similarity and recommendations for Movies and Tv Shows were obtained..

Future Scope:

- Integrating this dataset with other external datasets such as IMDB ratings, rotten tomatoes can also provide many interesting findings.
- More time could be given into building a better recommender system, which later can be deployed on web for usage.


thank you