Capstone Project

Netflix Movies and TV Shows Clustering

By Nayana Pradeep

PROBLEM STATEMENT

- Given is a dataset of the Netflix movies and TV Shows with details of rating, cast, director, genre, description etc.
- Explore and analyze the dataset to arrive at interesting insights
- Cluster similar content through text based clustering

OVERVIEW

Netflix is a streaming service that offers a wide variety of award-winning TV shows, movies, anime, documentaries and more – on thousands of internet-connected devices.

Whenever you access the Netflix service, their recommendations system strives to help you find a show or movie to enjoy with minimal effort.

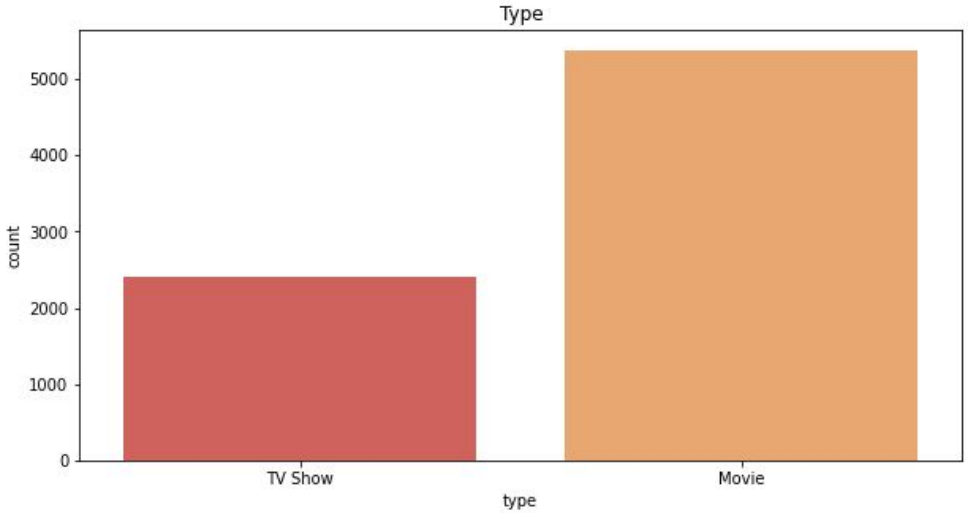
DATA SUMMARY

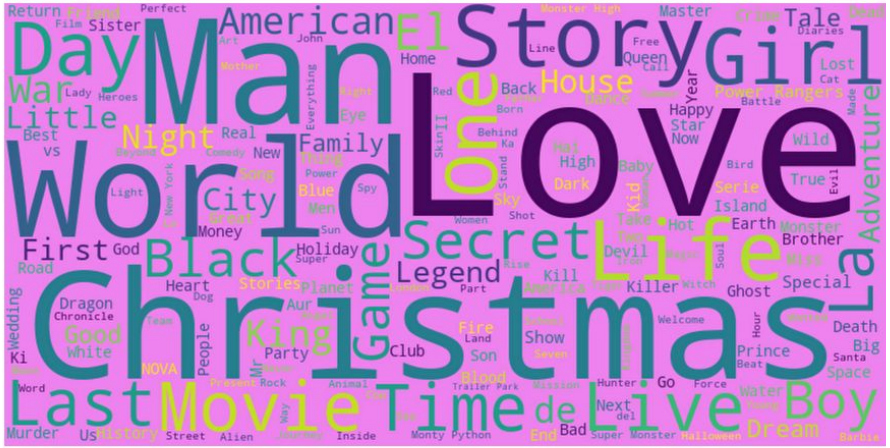
- show_id : Unique ID for every Movie / TV Show
- type : Identifier whether it is a Movie or a TV Show
- Title : Title of the Movie / TV Show
- Director : Director/s of the Movie
- cast : Cast members of the production
- country : Country where the movie / show was produced
- date_added : Date it was added on Netflix
- release_year : Actual Release year of the movie / show
- rating : Content Rating of the movie / show
- duration : Total Duration - in minutes or number of seasons
- listed_in : Genre
- description : The Summary description

EXPLORATORY DATA ANALYSIS

- 7787 rows and 12 columns
- No duplicate values present
- Null Values were present for some of the columns
- TV-MA, R, PG-13, TV-14, TV-PG, NR, TV-G, TV-Y, TV-Y7, PG, G, NC-17, TV-Y7-FV, UR are the available content ratings
- Data extraction performed on year_added column.
- Number of movies are more than the number of TV Shows.

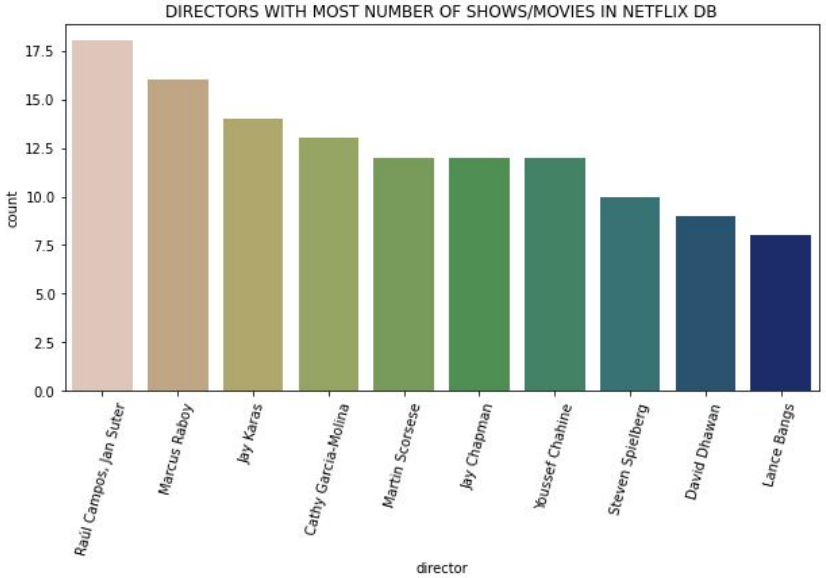
DATA VISUALIZATION

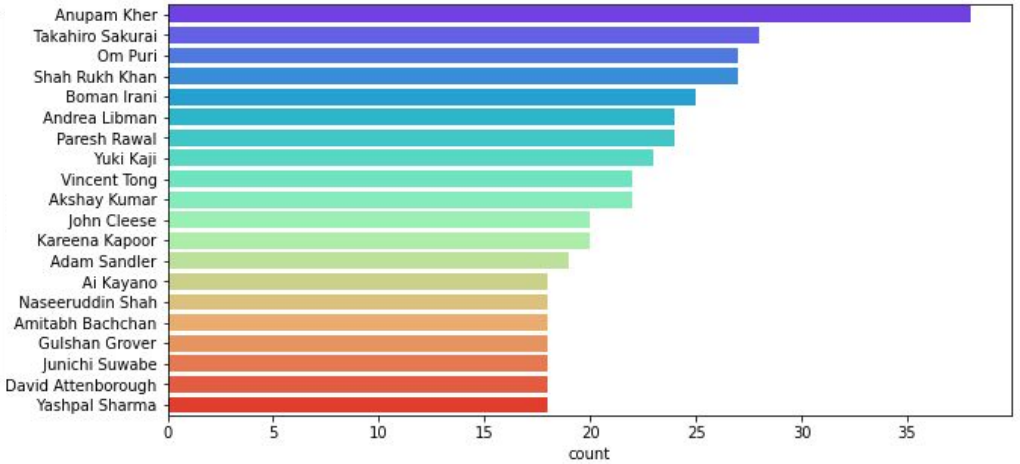




- Number of movies > Number of TV Shows
- Most used words in the title are Christmas, Love, World and Man

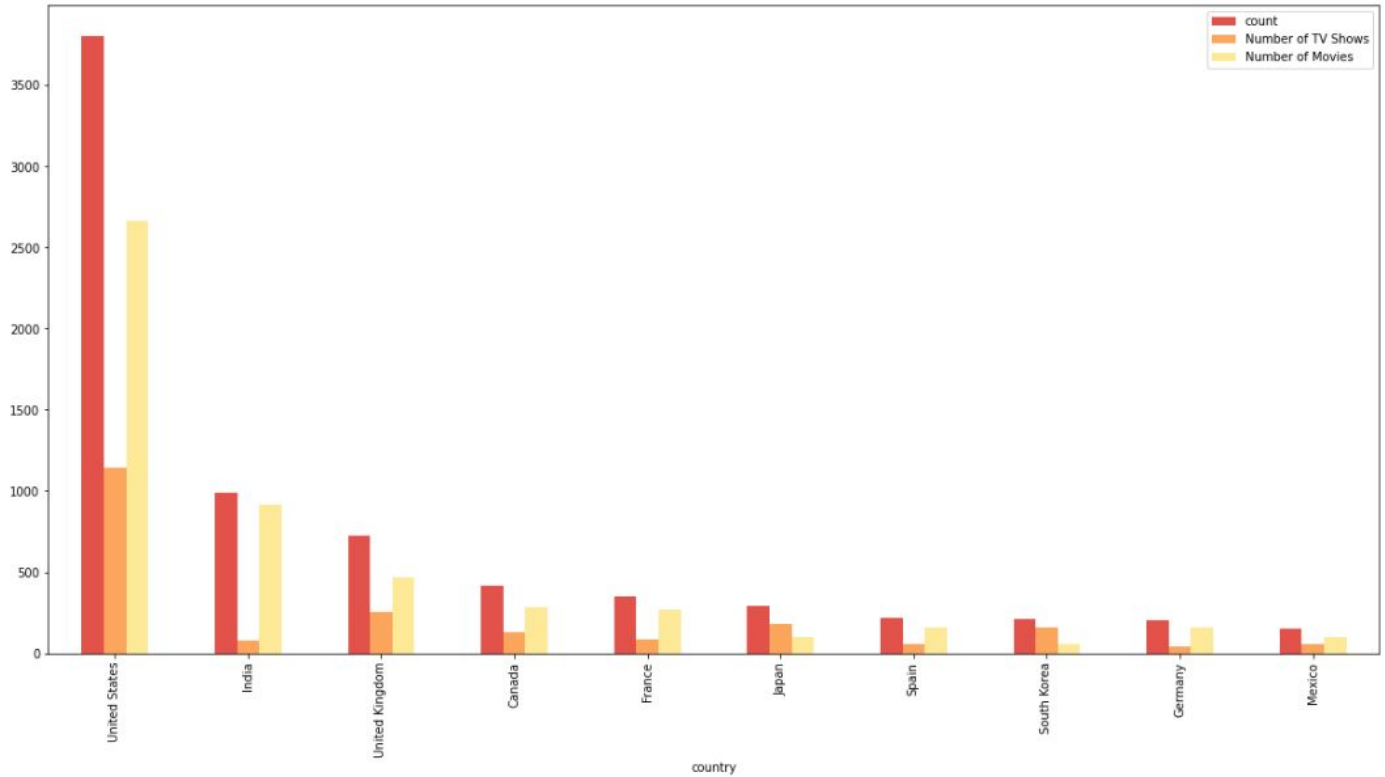
DATA VISUALIZATION





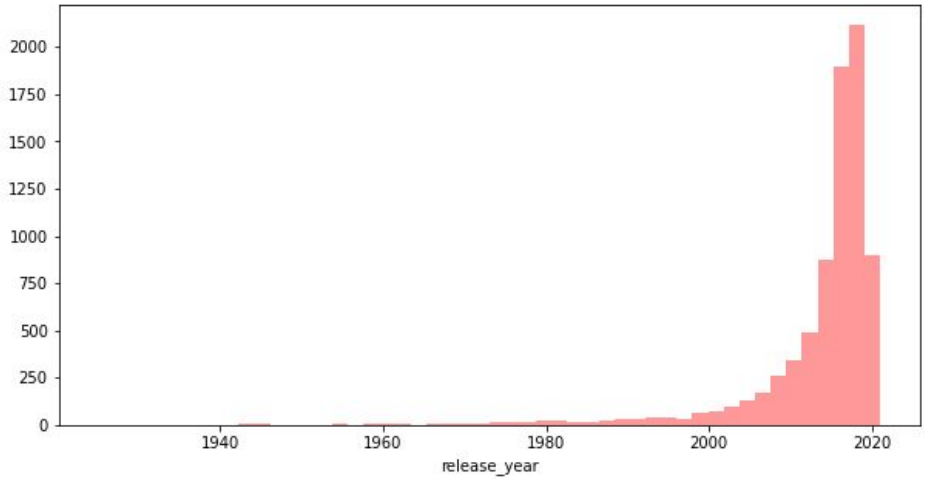
- Director with most number of movies – Jan Suter
- Actor with most number of movies – Anupam Kher

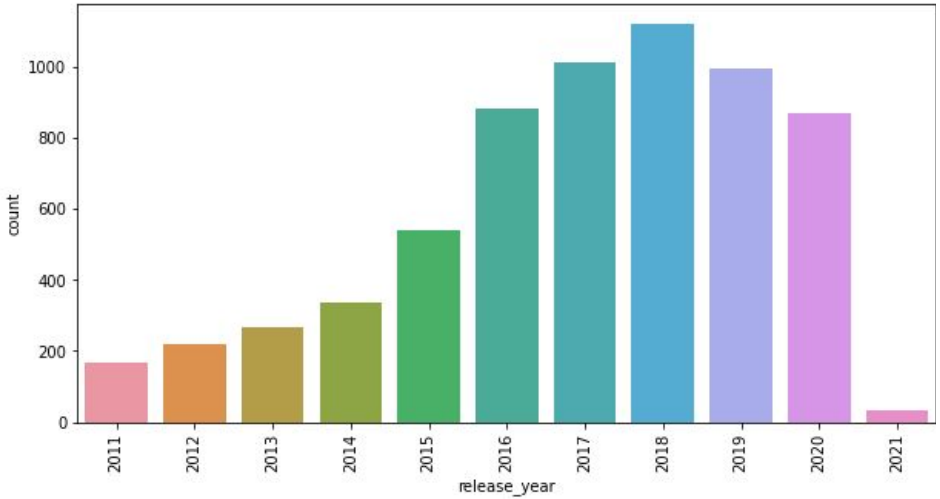
TOP COUNTRIES AND PRODUCTION TYPE



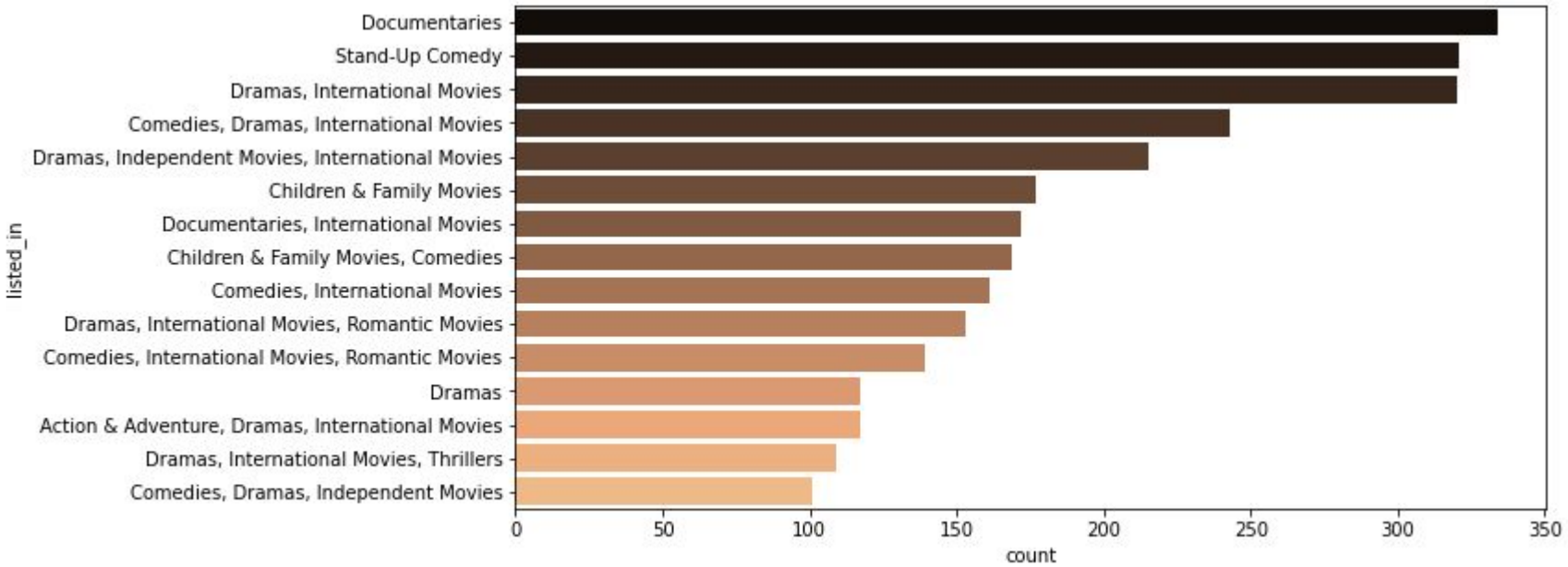
- USA
- INDIA
- UK
- CANADA
- FRANCE
- JAPAN
- SPAIN
- SOUTH KOREA
- GERMANY
- MEXICO

NETFLIX CONTENT AND RELEASED YEARS

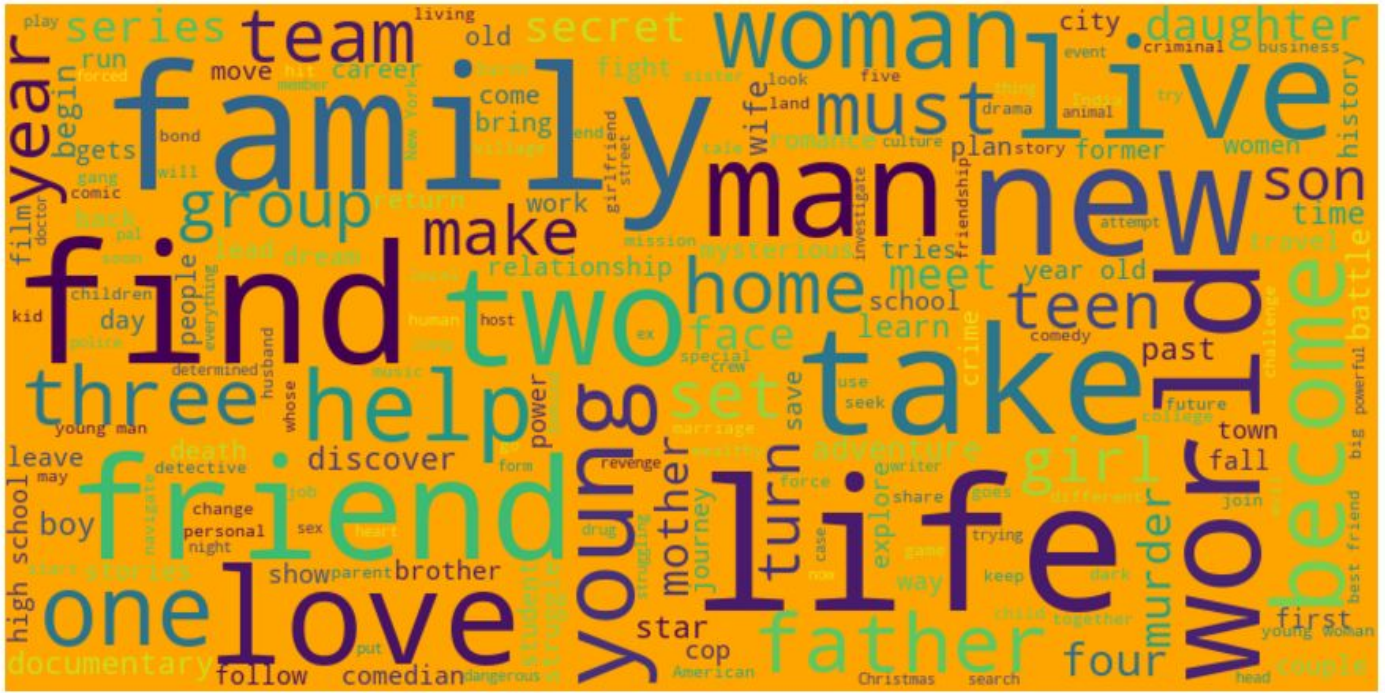




NETFLIX CONTENT GENRE

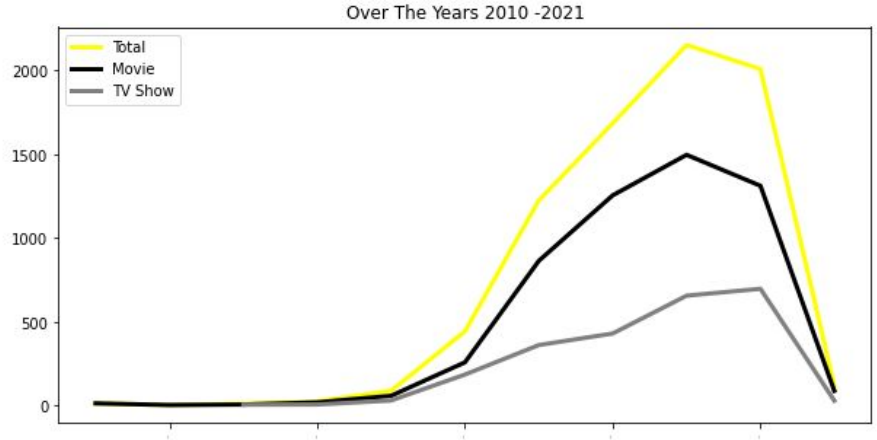


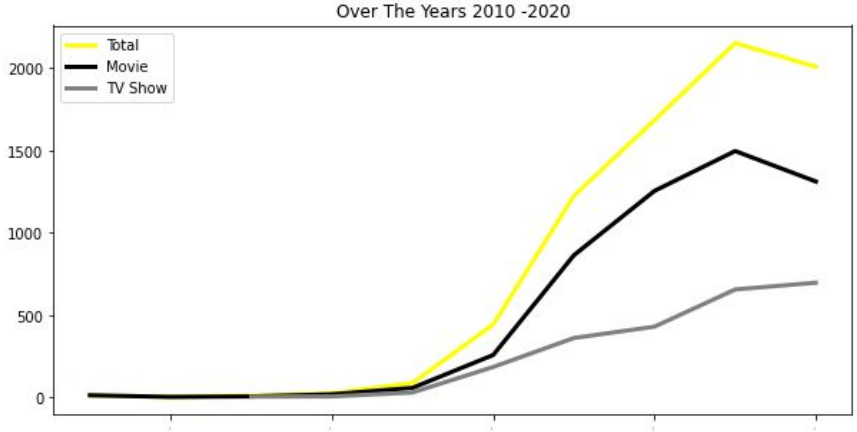
DESCRIPTION-MOST COMMON WORDS

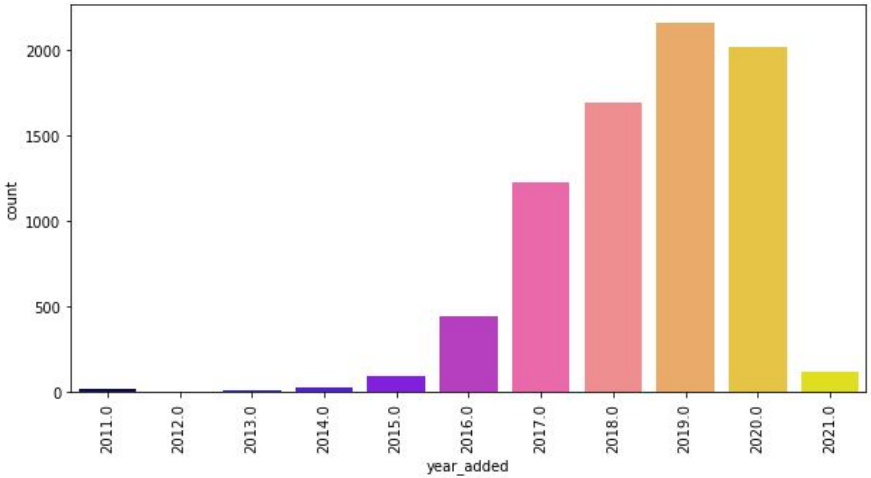


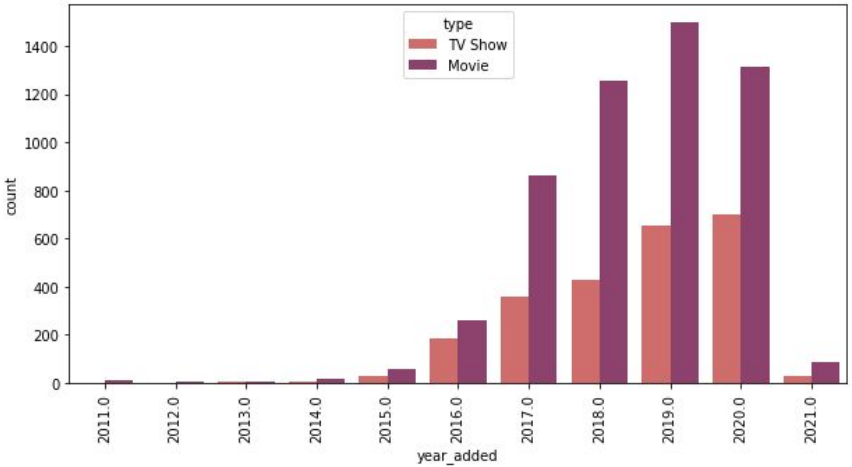
COUNTRIES AND CONTENT RATINGS

	country	count	Number of TV Shows	Number of Movies	Units with TV-MA rating	Units with R rating	Units with PG-13 rating	Units with TV-14 rating	Units with TV-PG rating	Units with NR rating	Units with TV-G rating	Units with TV-Y rating	Units with TV-Y7 rating	Units with PG rating	Units with G rating	Units with NC-17 rating	Units with TV-Y7-FV rating	Units with UR rating
0	United States	3804	1143	2661	1193	541	343	600	354	43	101	185	187	210	37	1	3	2
1	India	990	75	915	246	5	9	542	142	8	10	7	14	5	0	0	1	1
2	United Kingdom	723	256	467	237	120	66	97	95	12	23	27	11	31	3	0	0	1
3	Canada	412	126	286	100	68	28	47	39	5	17	45	35	23	2	1	1	1
4	France	349	84	265	152	49	22	43	11	4	6	18	21	18	2	1	0	2


Over The Years 2010 -2021
Total
Movie
TV Show
2000
1500
1000
500
0


Over The Years 2010 -2020
Total
Movie
TV Show
2000
1500
1000
500
0


count
2000
1500
1000
500
0
2011.0
2012.0
2013.0
2014.0
2015.0
2016.0
2017.0
2018.0
2019.0
2020.0
2021.0
year_added


type
TV Show
Movie
count
1400
1200
1000
800
600
400
200
0
2011.0
2012.0
2013.0
2014.0
2015.0
2016.0
2017.0
2018.0
2019.0
2020.0
2021.0
year_added

DATA CLEANING AND FEATURE ENGINEERING

- Textual features are combined to one column ‘combined_text_feat’ column
- Text cleaning was performed on this column .
- Converted to lower case
- Removed symbols, special character and stopwords
- Removed 2 letter words
- Stemming is also performed on the column text

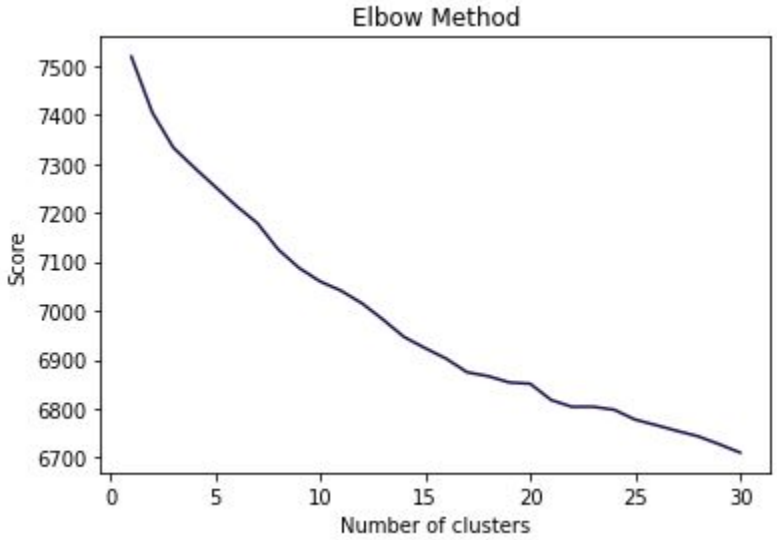
TEXT BASED CLUSTERING

- TF –IDF Vectorization is performed on the data to perform clustering
- Kmeans Clustering is used here
- Optimal number of cluster is 25.
- Concluded after the Elbow and Silhouette test
- Silhouette score with 25 clusters in 0.0279
- Clusters are identified for all the movies and TV shows in the dataset.
- Additionally the recommendation algorithm was also performed on the dataset using the cosine similarity method on same vectorized data

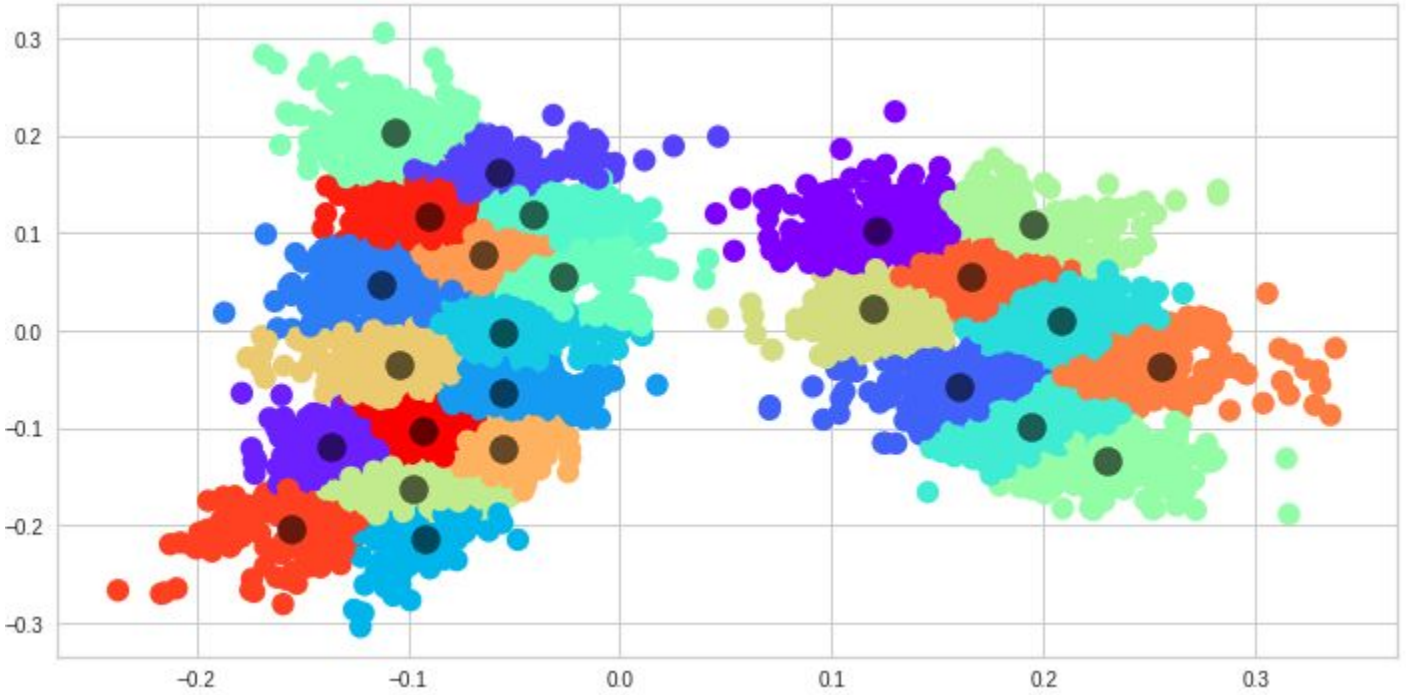
OPTIMAL CLUSTER NUMBER

- ELBOW Method - The method consists of plotting the explained variation as a function of the number of clusters, and picking the elbow of the curve as the number of clusters to use.

- Silhouette Method - The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from −1 to +1,



PRINCIPAL COMPONENT ANALYSIS



CONCLUSION

- United States provides the most number of movies and shows followed by India and United Kingdom.
- TV-MA rated content is maximum in number in the dataset. This rating indicates that the content is for mature and adult audience above the age of 17.
- There is an exponential raise in the number of TV shows and movies distributed by Netflix in the recent years.

.

- Optimal number of clusters were found out to be 25 with silhouette coefficient value of 0.0279

CHALLENGES

- Complex dataset
- Clusters are too close to each other
- As the cluster number increase, time taken to fit the cluster also increase.

THANK YOU