Capstone Project-4

Netflix TV Shows and Movies Clustering

TEXT BASED CLUSTERING

Palash Pathak

OVERVIEW

1. Defining Problem Statement
2. Data Cleaning
3. Exploratory Data Analysis
4. Country based Analysis
5. Movies vs TV Shows- recent focus
4. Text processing -NLP
5. Optimal number of Clusters
6. Applying Model
7. Model Validation

Overview of the Problem Statement

- 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.
- Netflix, Inc. is an American subscription streaming service and production company. Launched on August 29, 1997.
- Tasks for the Project:

1. Exploratory Data Analysis
2. Understanding what type of content is available in different countries.
3. Is Netflix has increasingly focusing on TV rather than movies in recent years.
4. Clustering similar content by matching text-based features.

Data Summary:

- No of movies/shows & no of attributes : (7787,12)

Column Name	Nan %	Data Type	Unique Values	Description
Show id	0.0	Object	7787	Unique id for every Movie / TV Show
Type	0.0	Object	2	A Movie or a TV Show
Title	0.0	Object	7787	Title of the Movie/TV Show
Director	30.67	Object	4050	Director/s of the Movie/TV Show
Cast	9.22	Object	6832	Cast of the Movie/TV Show
Country	6.51	Object	682	Country where produced
Date added	0.13	Date-time	1512	Date it was added on Netflix
Release year	0.0	Int64	73	Actual release year of the Movie/TV Show
Rating	0.09	Object	14	TV rating of the Movie/TV Show
Duration	0.0	Object	216	Title of the Movie/TV Show
Listed-in	0.0	Object	492	Genre
Description	0.0	Object	7769	Summary, description

Data Pipeline

Data Cleaning :
-Missing values : Director, Cast, Country, Date Added, Rating.

EDA : Some exploratory data analysis(EDA), Country-based Analysis, Movies vs TV Shows analysis for recent years.
IMDB dataset is also used to get interesting insights.

Text processing : Removing stop-words, punctuations, stemming, extracting important words from description feature & using TF-id vectorizer. Using PCA to reduce dimension.

Number of Clusters: Selecting optimal number of Clusters. Techniques used Elbow method, Silhouette analysis, Dendogram.

Clustering: K-means clustering.

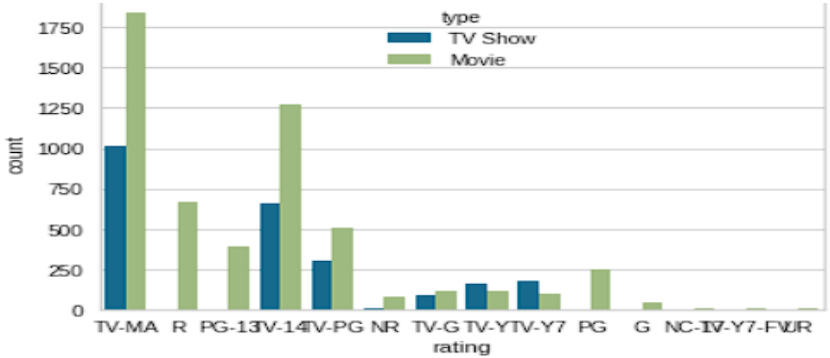
Evaluation : Checking if the clusters make sense and summary.

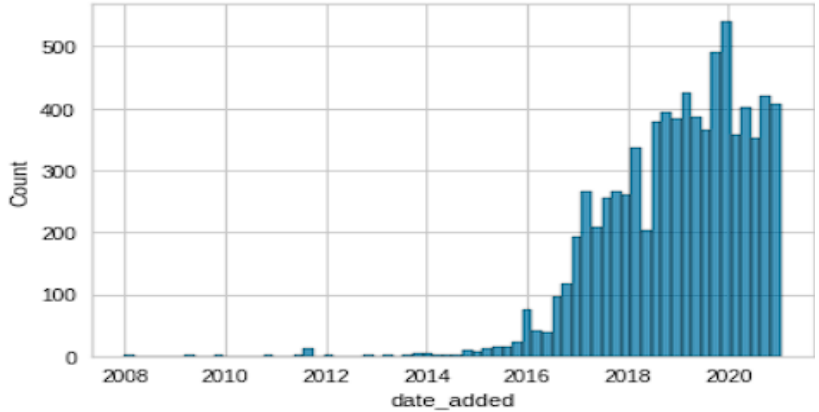
Dealing With Missing Values -

1. Rating
TV rating of the Movie/TV Show – 0.09% nulls

2. Date-added
Date on which the content was added on Netflix – 0.13% nulls

Let's first have a look at the distribution





Lets replace all these nulls for Rating & Date-added with the mode –most frequent value.
Mode for Rating : TV-MA
Mode for Date-added : 1st Jan 2020

3. Director
Director of the Movie/TV Show – 30.69% nulls

4. Cast
Cast of the Movie/TV Show – 9.22% nulls

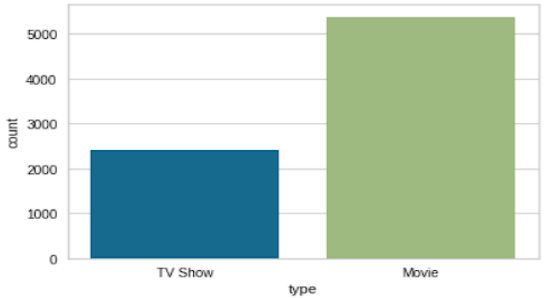
5. Country
Country where produced – 6.51% nulls

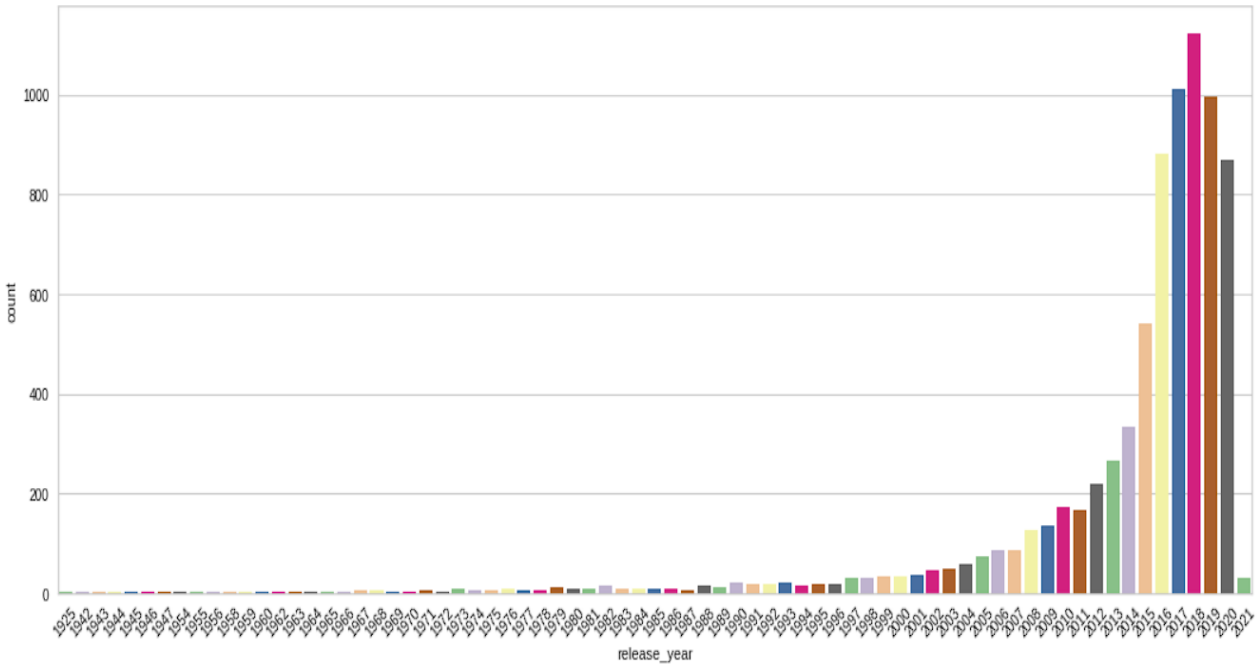
Country, Cast and Director have high percentage of nulls values.
Lets replace nulls with empty string for further use for text processing.
Before doing that lets store the non-null entries separately for EDA.

EDA

Netflix Content Type : Movie vs TV Show

Majority of content on Netflix are Movies.

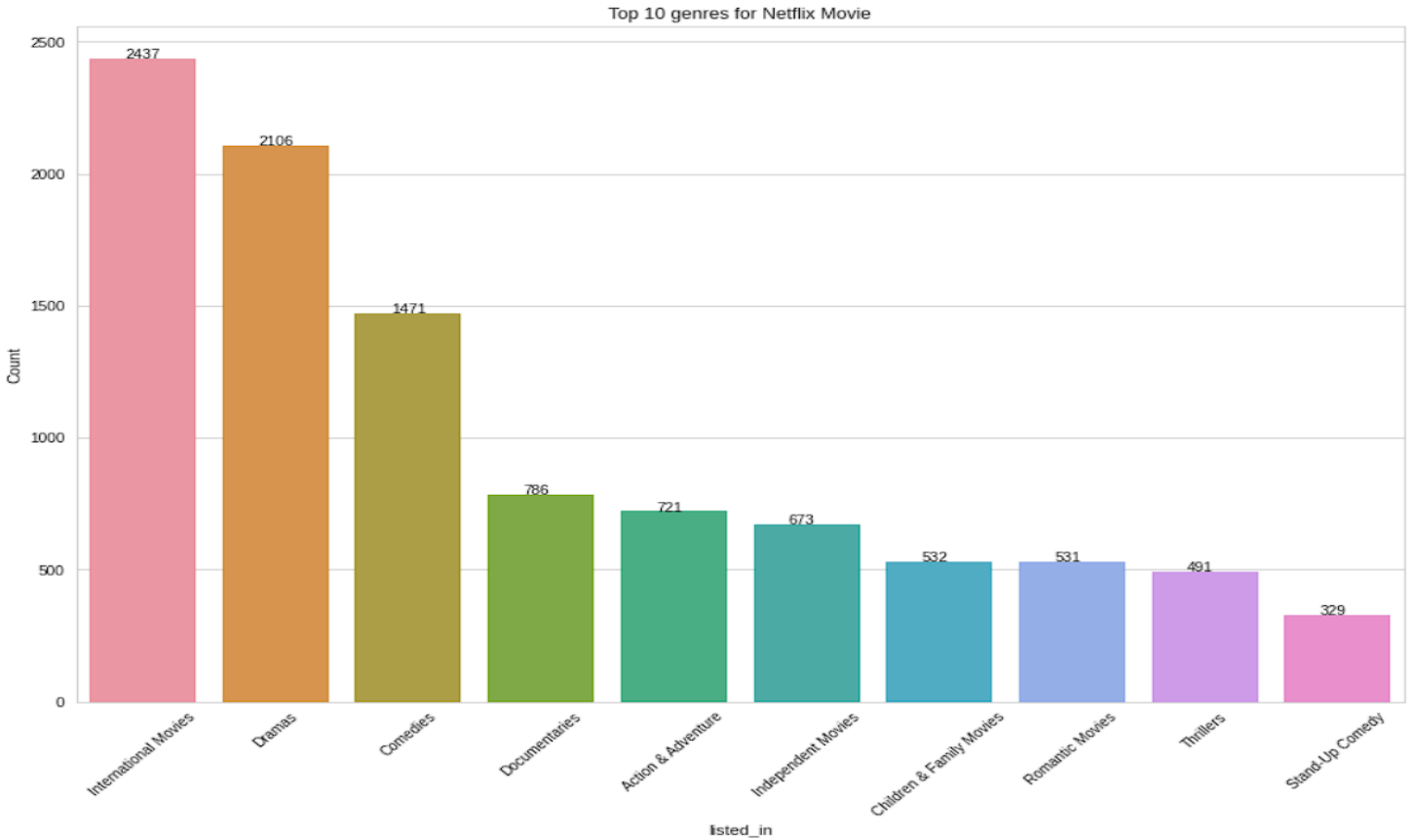




Netflix Content Release year :

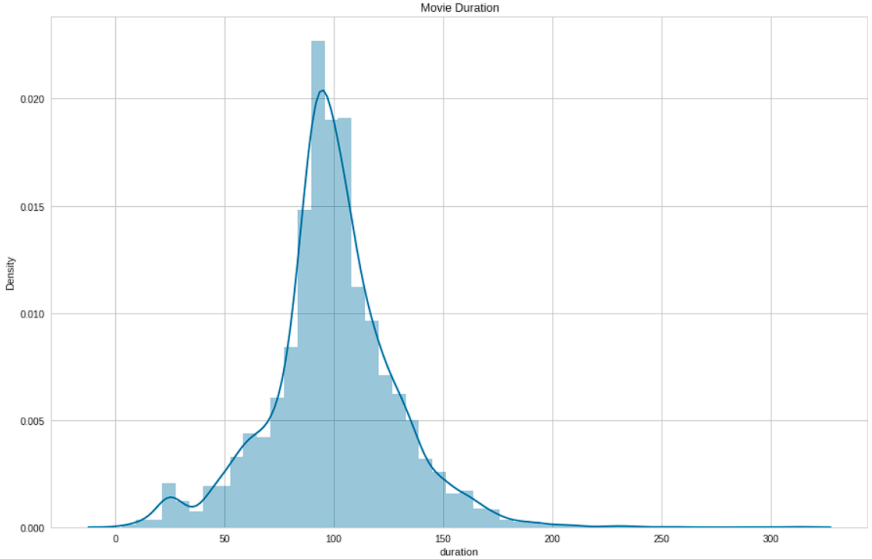
Majority of content on Netflix are have actual release date in the years from 2015-2020

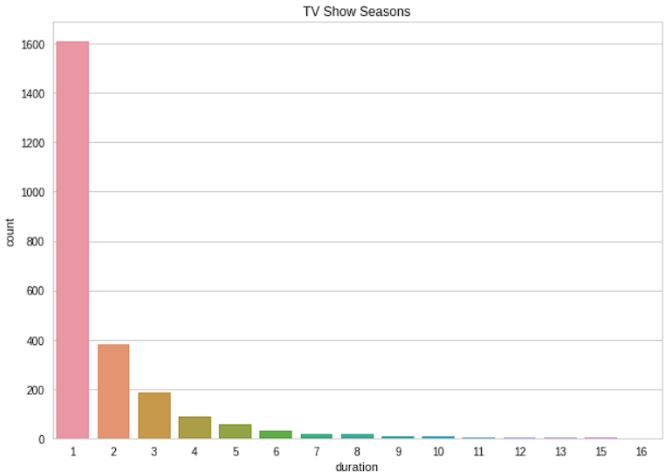
Netflix Content Genre : Movie & TV Show Genres



Duration of Movies on Netflix

Average duration is around 100mins

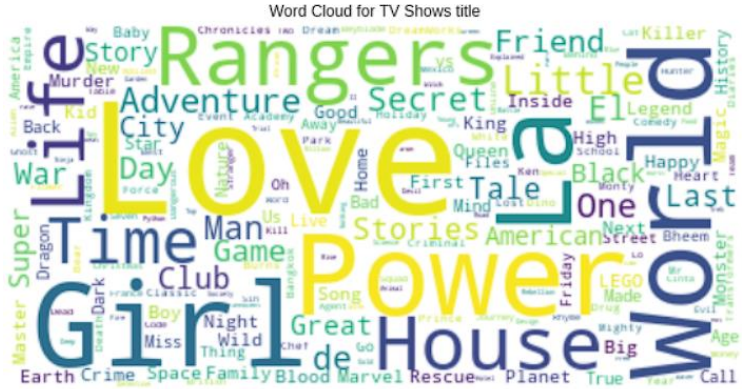




Season length of TV Shows on Netflix

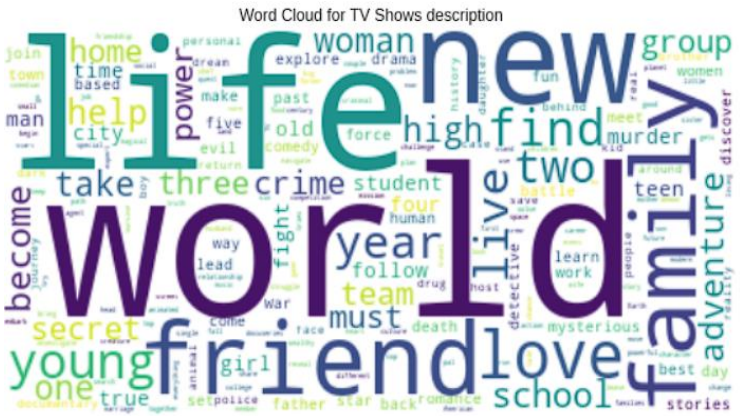
Most of the shows have 1-2 seasons.

Word-Cloud for TV Show Titles



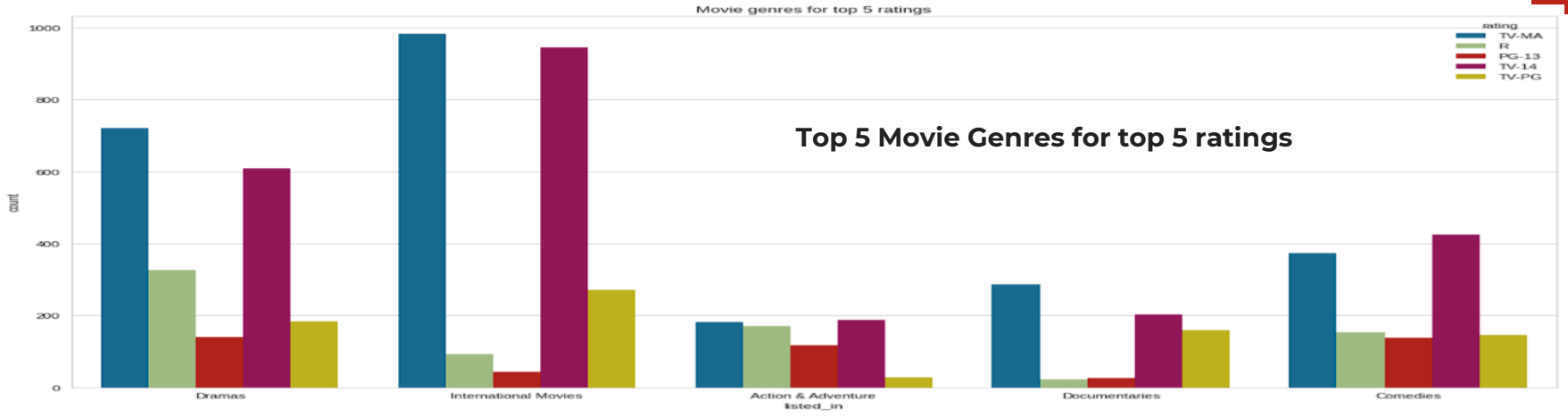
Word-Cloud for Movie Titles

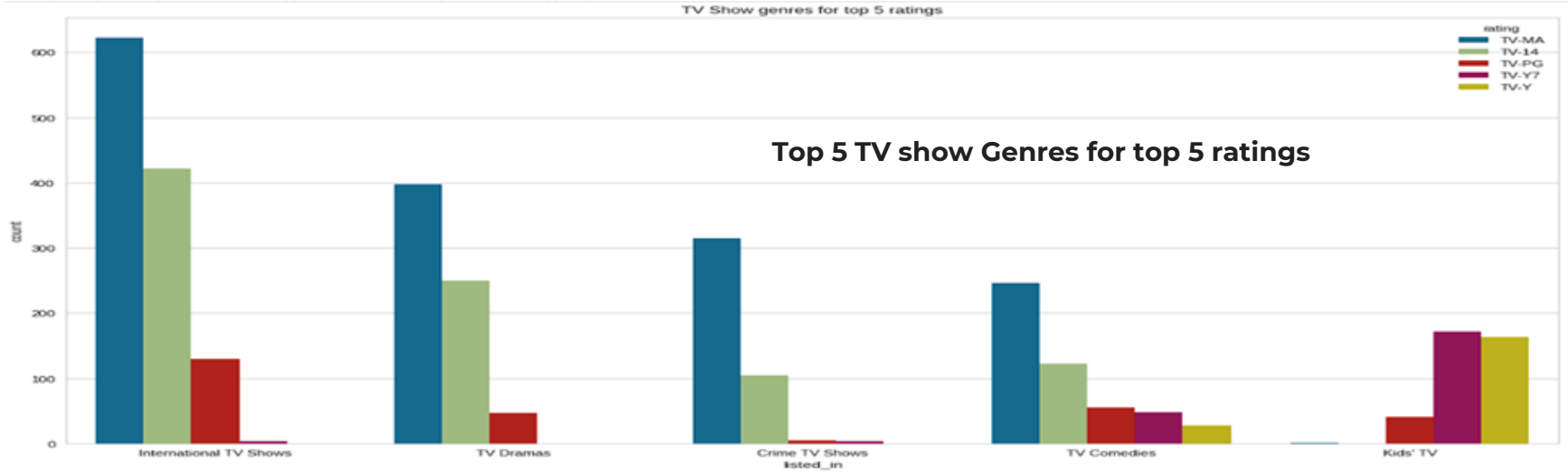
Word-Cloud for TV Show Descriptions



Word-Cloud for Movie Descriptions

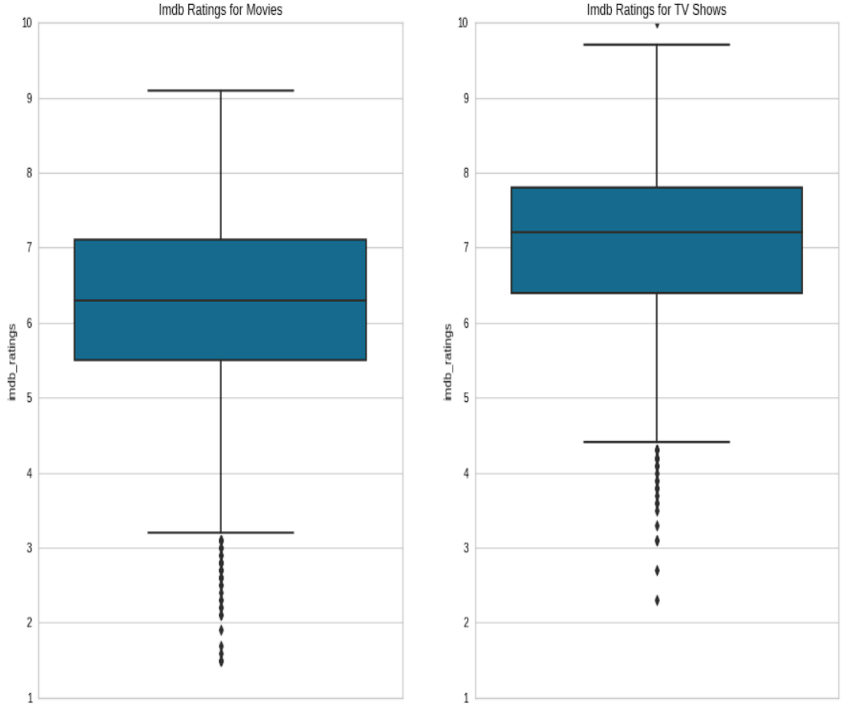
Top 5 Movie/TV Show Genre for top 5 ratings



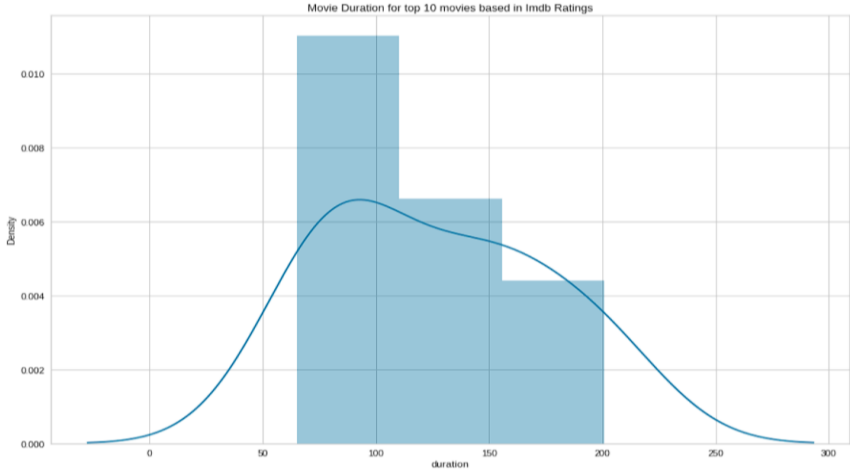


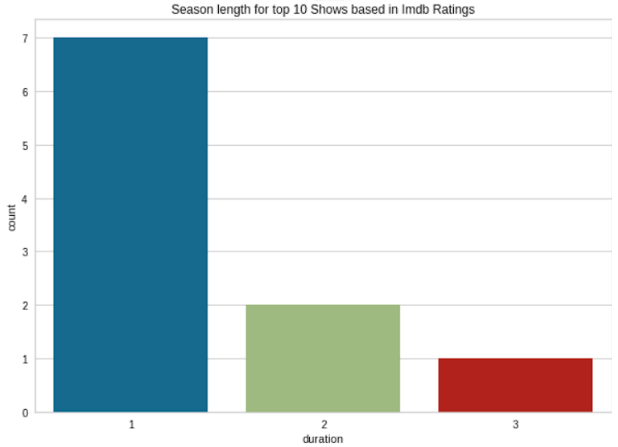
Using IMDB dataset with Netflix contents

IMDB ratings for Movies & TV Shows



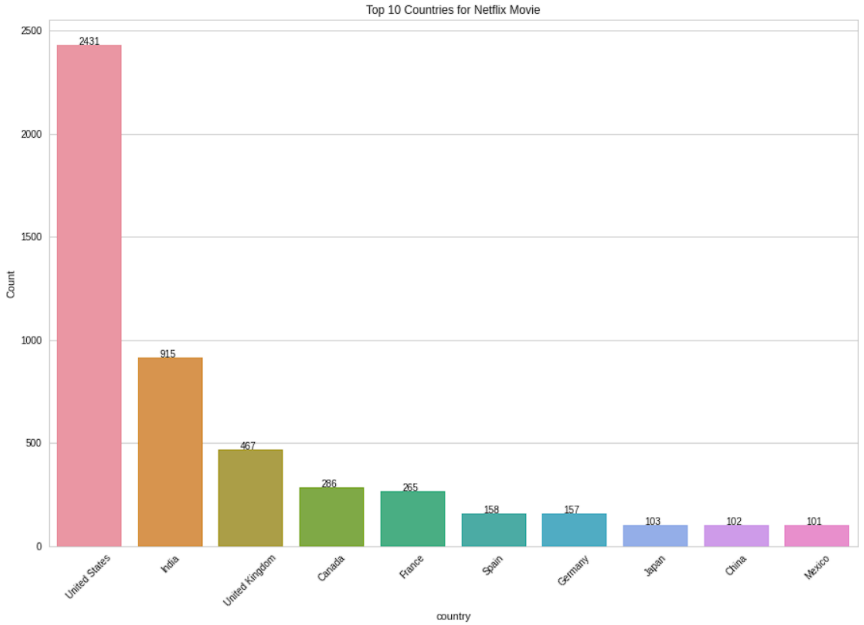
Movie duration & TV Show Season length for top 10 Netflix Movie, Show based on IMDB user ratings



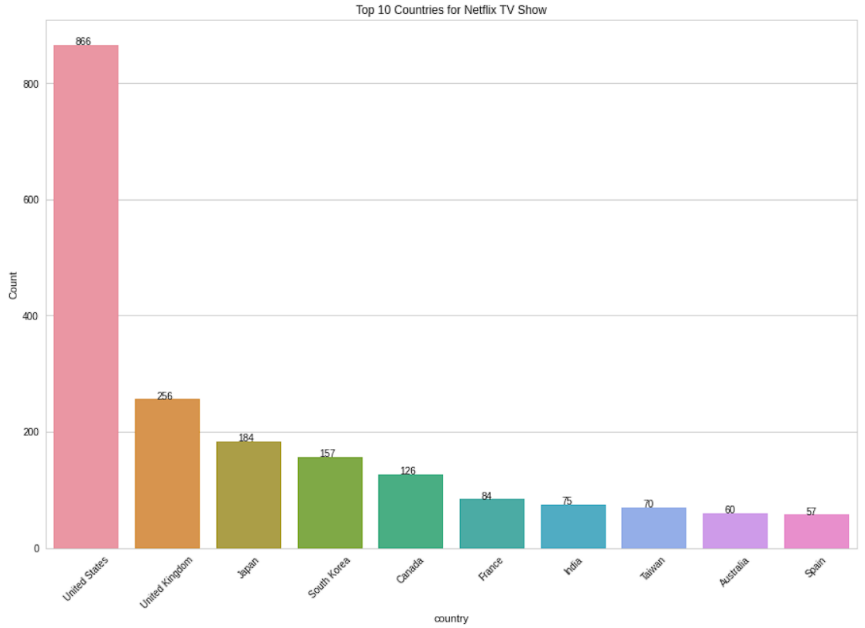


Countries based on Netflix contents

Countries based on Movies

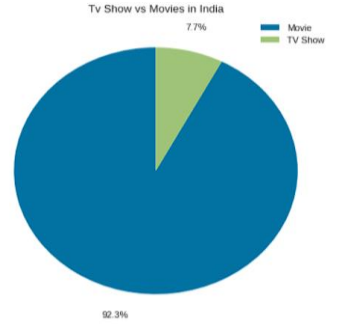


Countries based on TV Shows



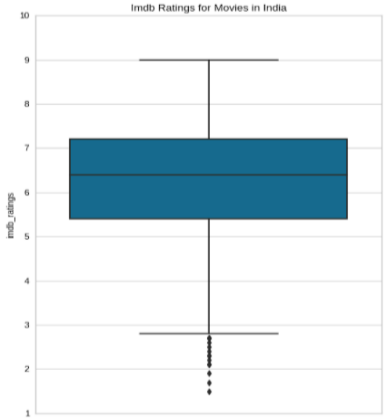
Contents from top countries - INDIA

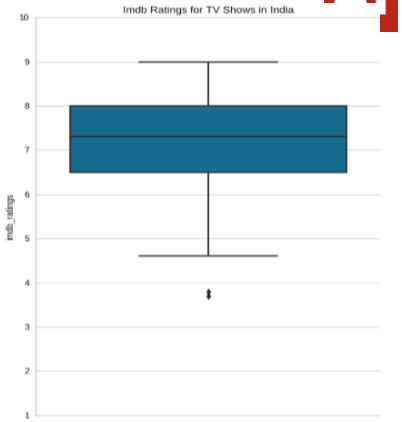
Contents produced



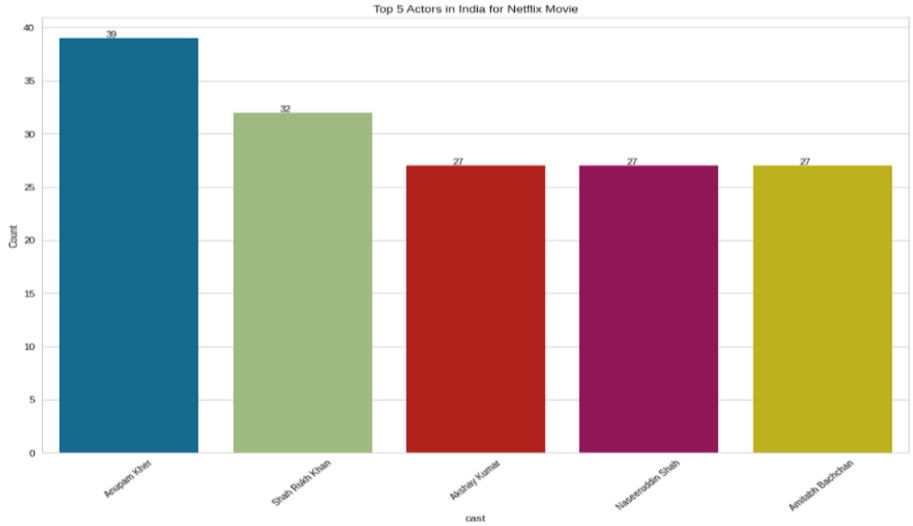
Word-Cloud for movie title

Imdb User Ratings





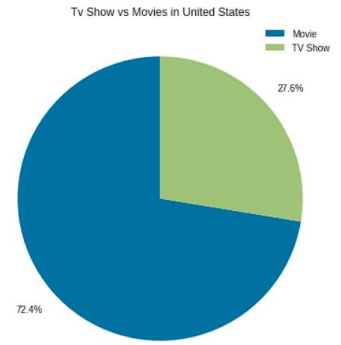
Top Movie Actors from India



Top Movie Genres for Indian

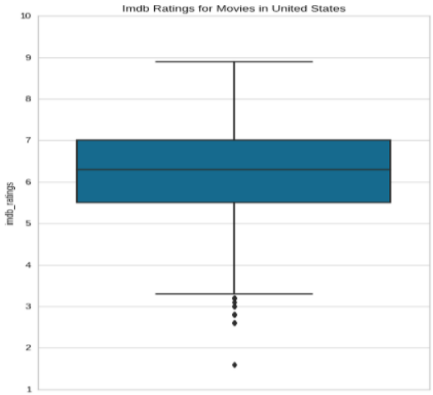
Contents from top countries – United States

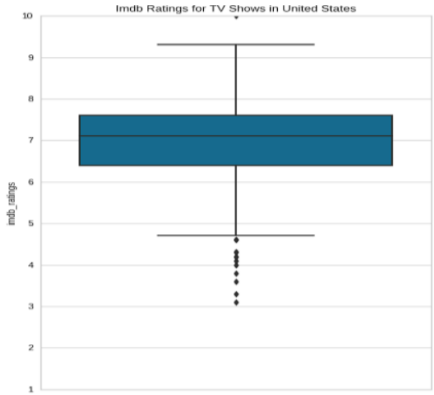
Contents produced



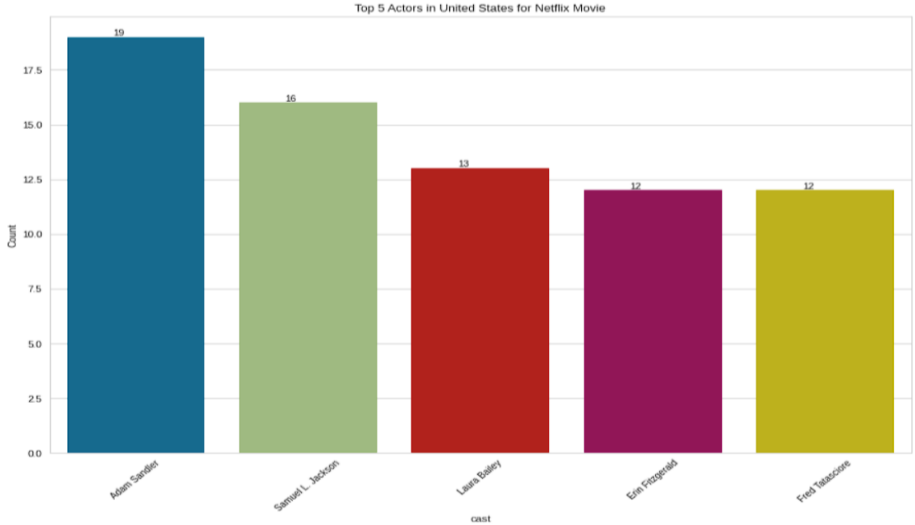
Word-Cloud for movie title

Imdb User Ratings

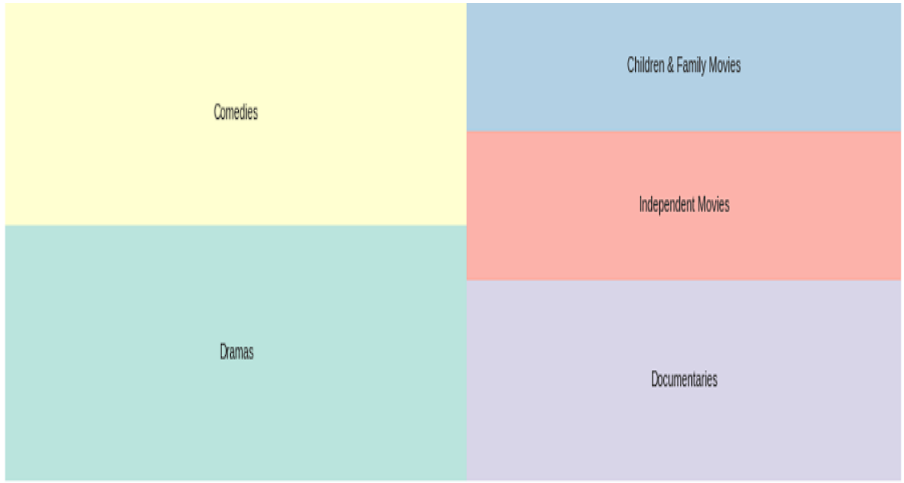




Top Movie Actors from U.S.

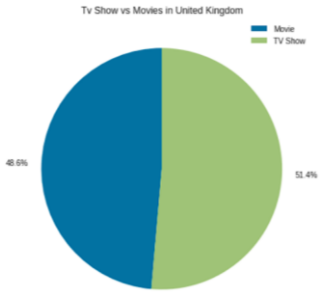


Top Movie Genres for U.S.



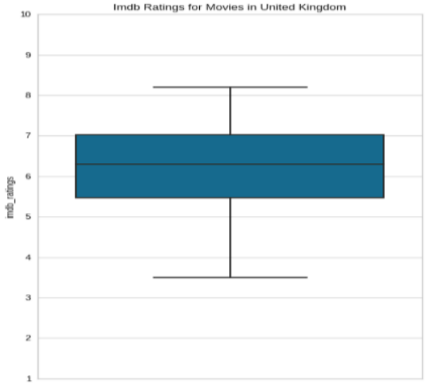
Contents from top countries – United Kingdom

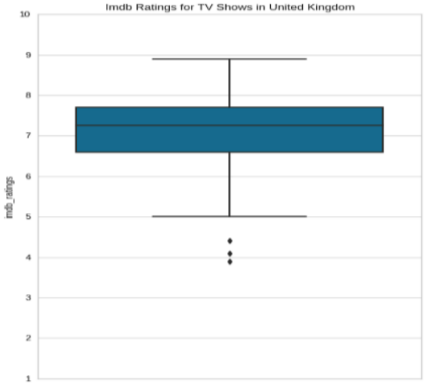
Contents produced



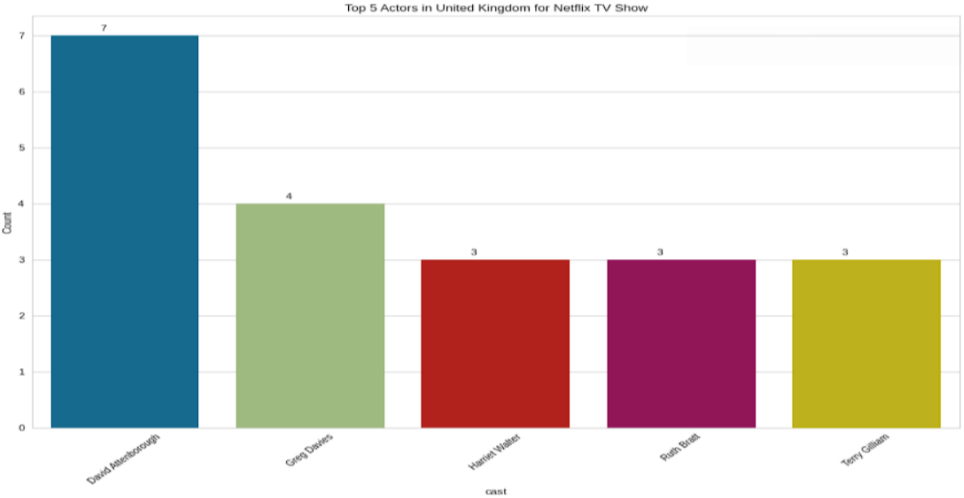
Word-Cloud for tv show title

Imdb User Ratings

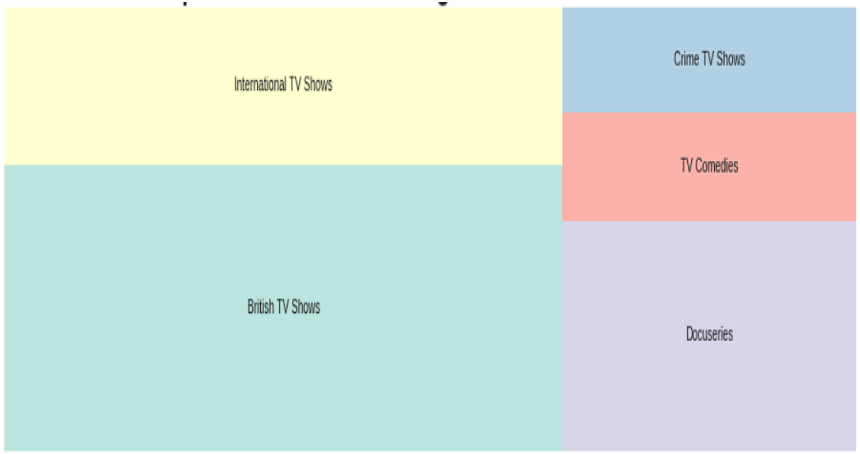




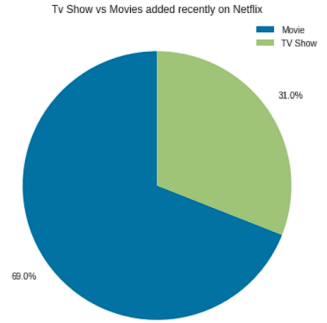
Top TV Show Actors from U.K.

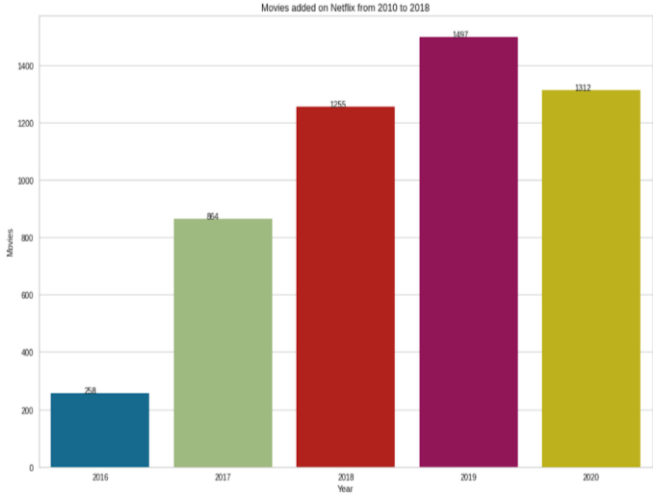


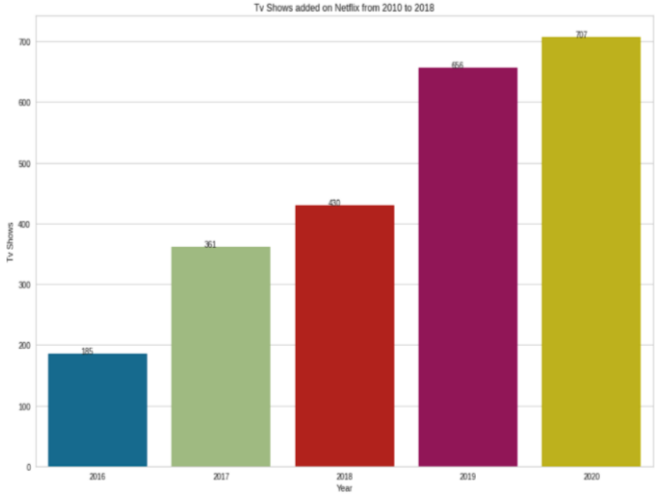
Top TV Show Genres for U.K.

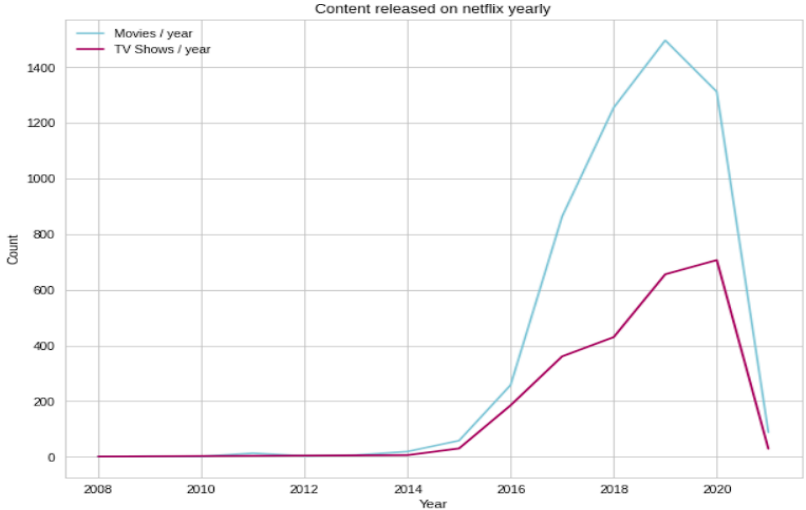


Movies/TV Show in recent years









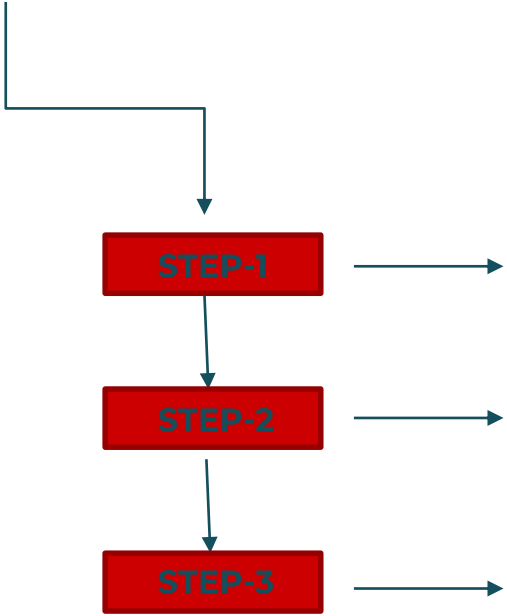
Netflix is increasing both TV show & Movies contents. Also, they have focused on movies in recent year not the other way round. We should ignore the year 2021 as we have data for contents added on or before 16/01/21

Netflix has added 5186 movies in last 5 years(2016-20)

Netflix has added 2339 Tv shows during this period

So, overall Netflix has added 7525 contents in this period.

Text Processing:



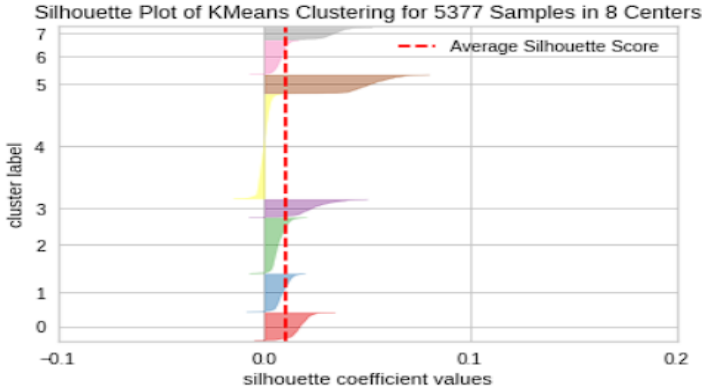
Selecting relevant text based features

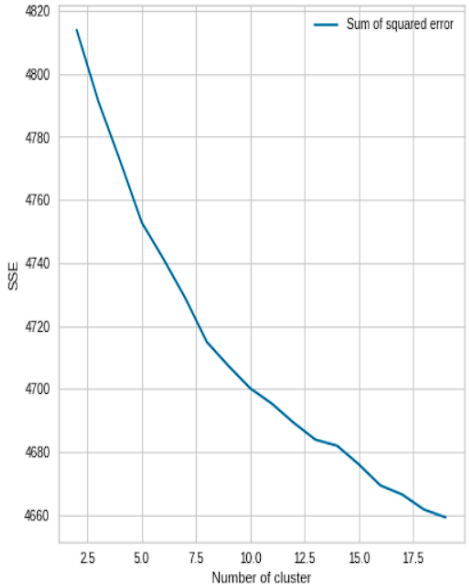
Extracting important words from description

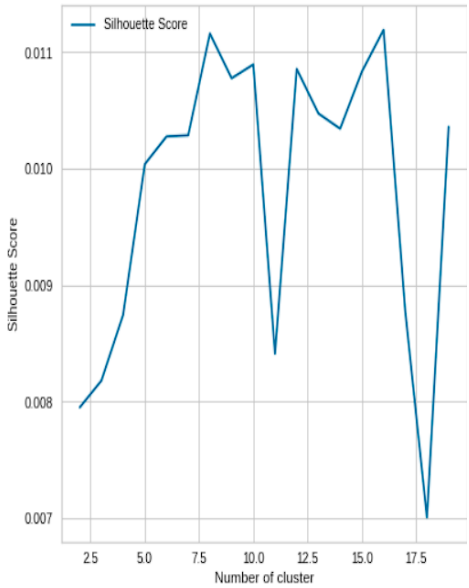
- Removing blank spaces, punctuations, stop-words
- Stemming
- Converting all text to lower case
- For names merging together first & last name to avoid mix up between people
- Tfid vectorizer
- Using PCA to reduce dimension

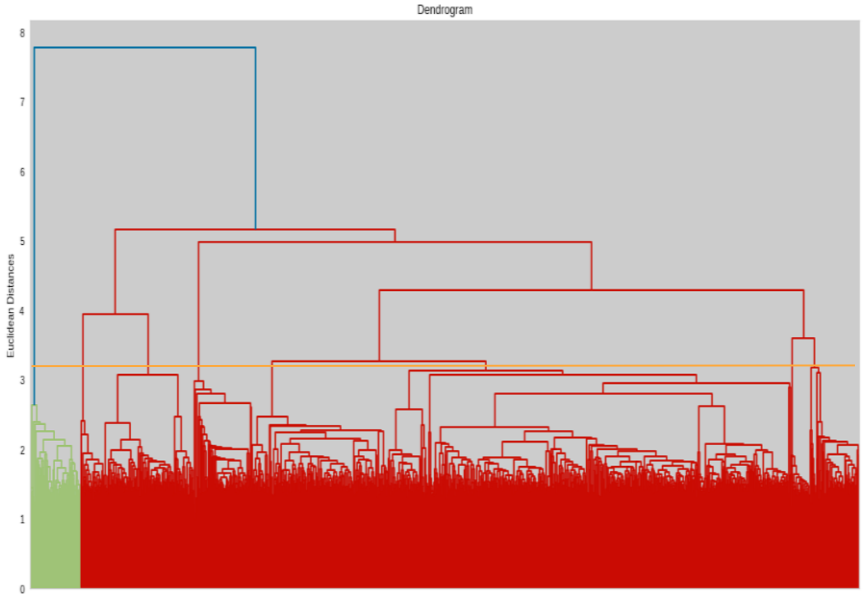
Movies -
Determine Optimal number of Clusters :

Using Elbow method, Silhouette Analysis & Dendogram,
Lets select k=8 as optimal number of clusters





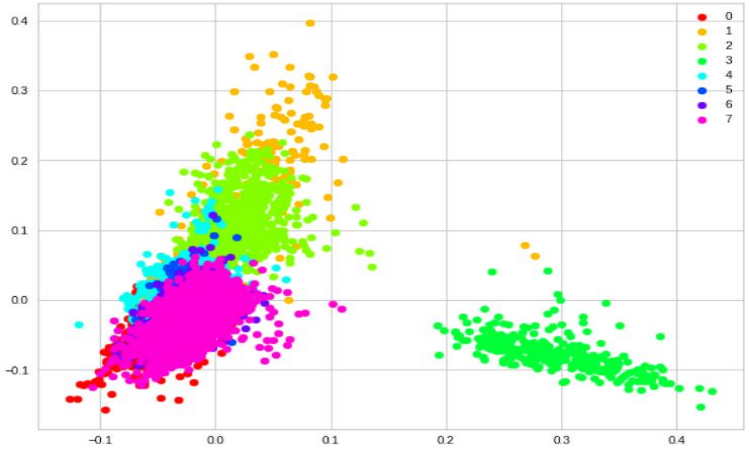




Movies-
K-means Clustering :

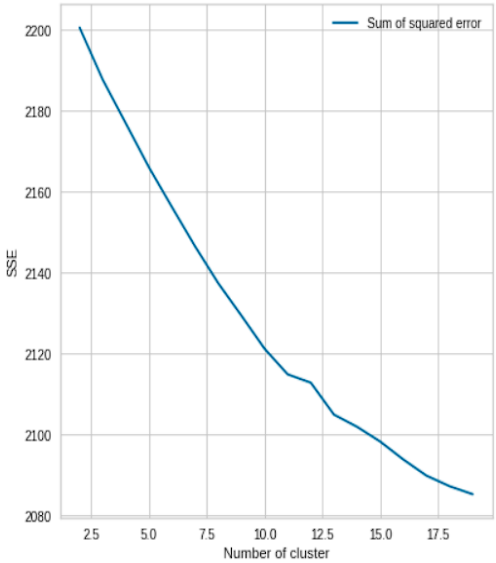
With k=8

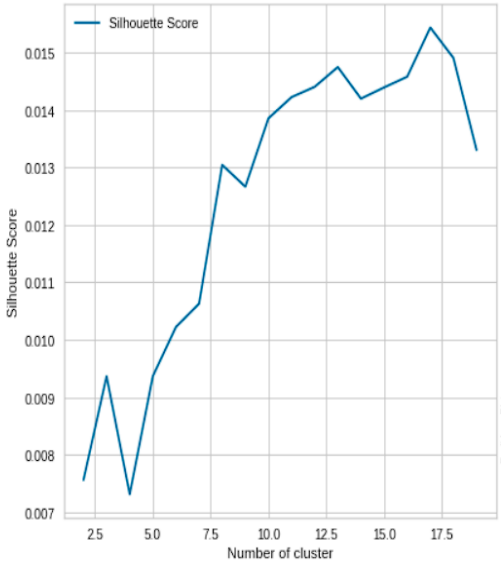
Cluster	Cluster_Lable	Top_Words_Title	Top_Words_Description
0	Romantic,Drama,Comedy	[love,wedding,christmas]	[love,life,young]
1	Musicals	[sessions,remastered,music,concert]	[music,band,rock]
2	Documentaries,Sports	[world,story,nova,life]	[explore,journey,chronicle]
3	Stand Up	[live,special,jeff,russel]	[special,comic,show]
4	Children,Family	[little,monster,christmas]	[friend,new,save]
5	Sci-Fi,Fantasy	[dragonheart,black,hulk,star]	[life,earth]
6	Action,Adventure	[man,kill, vangeance, cop]	[agent,mission,crime]
7	Horror,Thriller	[house, last, ghost,night]	[young,find,killer]

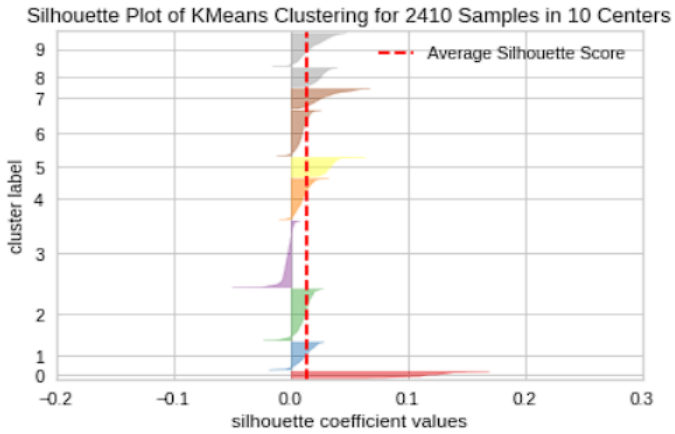


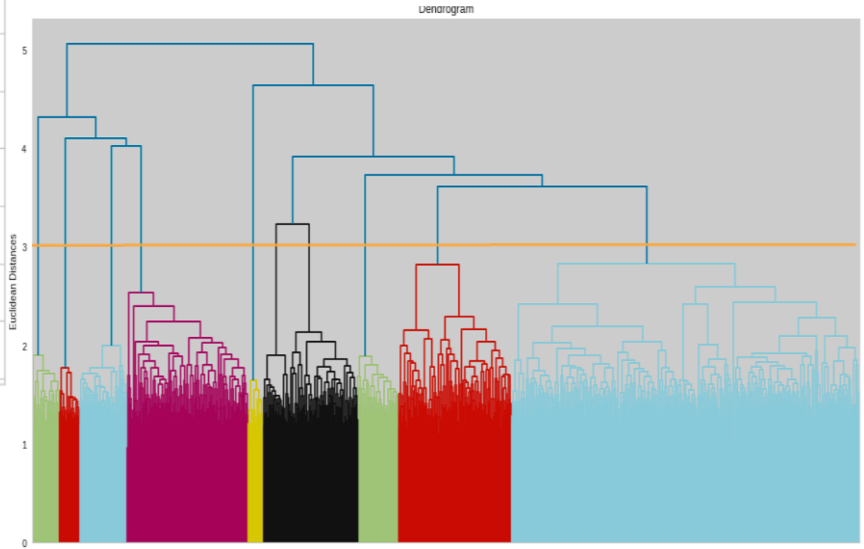
TV Shows-
Determine Optimal number of Clusters :

Using Elbow method, Silhouette Analysis & Dendogram,
Lets select k=10 as optimal number of clusters





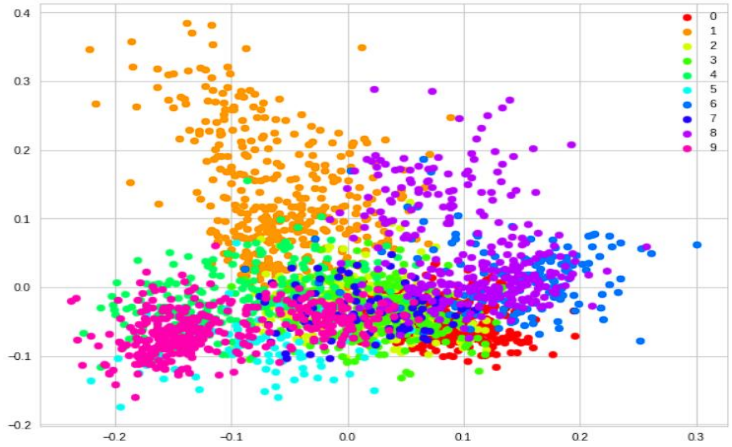


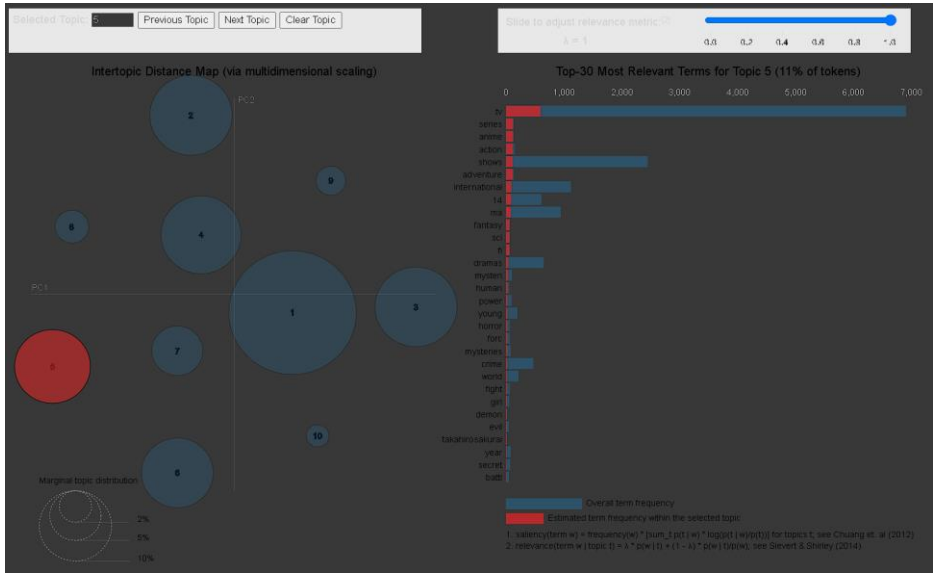


TV Shows-
K-means Clustering :

With k=10

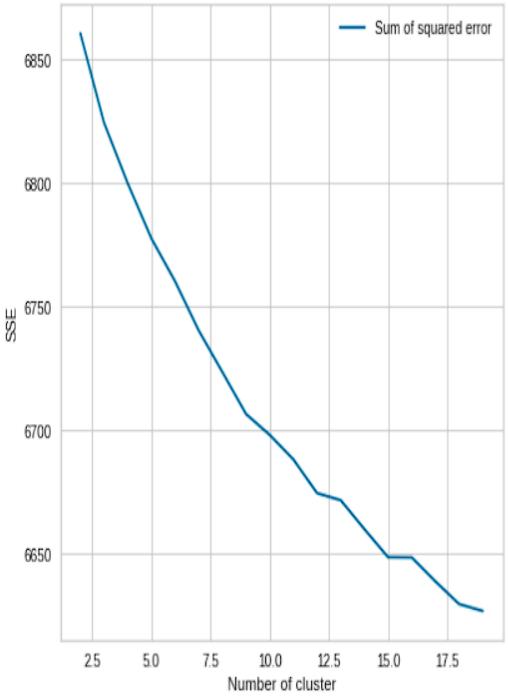
Cluster	Cluster_Lable	Top_Words_Title	Top_Words_Description
0	Korean,Romantic,Drama	[love,club,miss,beautiful]	[love,life,woman,romance]
1	Docuseries,Science,Reality	[planet,world,nature]	[world,explore,history]
2	British,Comedy,Drama	[family,young,house]	[life,friend,struggle]
3	International Dramas	[girl,queen,american,rebellion]	[school,high,find]
4	Stand Up Comedy	[show,comedy,dreamworks]	[comedy,show,fun]
5	Sci-Fi,Fantasy	[star,trek,dragon,arcadia]	[evil,power,human]
6	Spanish,Crime,Romantic	[mexico,spain,de,El,La]	[drug,mexico,life]
7	Horror,Mysteries	[haunting,anjaan,darr,house]	[past,secret,dark]
8	Docuseries,Crime	[killer,murder,drug]	[detective,police,case]
9	Kids,Anime	[power,rangers,super]	[friend,adventure,world]

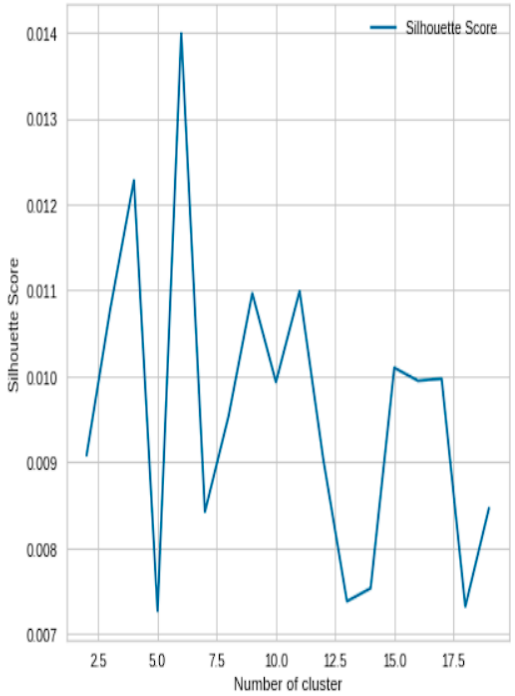


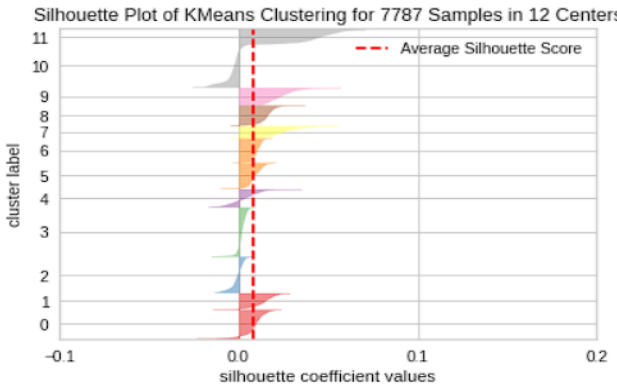


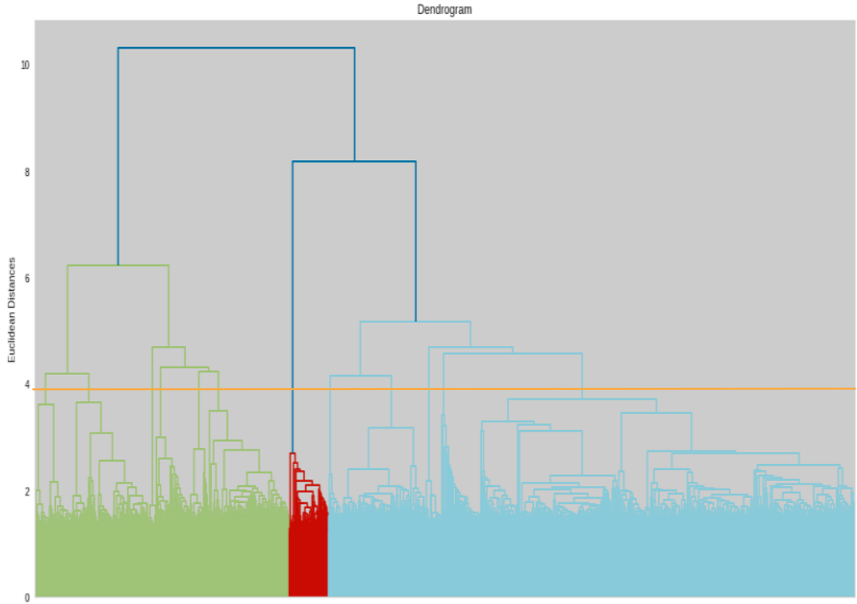
All Netflix Contents-
Determine Optimal number of Clusters :

Using Elbow method, Silhouette Analysis & Dendogram,
Lets select k=12 as optimal number of clusters





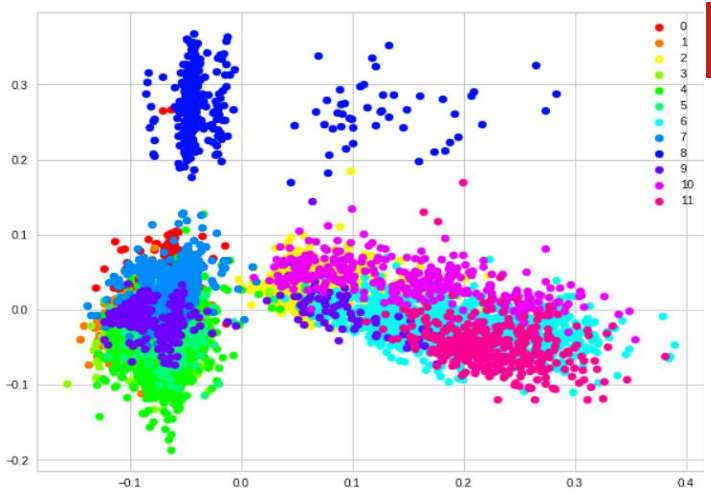




All Netflix Contents- K-means Clustering :

With k=12

Cluster	Majority_type	Cluster_Lable	Top_Words_Title	Top_Words_Description
0	Movie	Musicals,Documentaries	[lifetime, sessions, remastered]	[music, love, band]
1	Movie	Children,Family	[christmas, monster, little]	[save, young, find, new]
2	TV Show	Kids,Anime	[power, rangers, super]	[adventure, friend, family]
3	Movie	Horrors,Thrillers	[ghost, house, killer,game]	[young, find, murder]
4	Movie	Dramas,Romantic,Comedy	[love, christmas, boy, girl]	[love, find, man, woman]
5	Movie	Action,Adventure	[kill, man,cop]	[take, life, cop, agent]
6	Movie	International,Crime,Dramas	[game, girl, crime, La,El]	[family, school, murder]
7	Movie	Documentaries,Sports	[story,nova,world]	[history, explore, interview]
8	Movie	Stand Up Comedy	[live, special, jeff, bill]	[special,stand, comic, humor]
9	Movie	Sci-Fi,Fantasy	[black, star, legend, dark]	[life, earth, discover, alien]
10	TV Show	Docuseries,Science	[planet, world,nature]	[series, world,explorer]
11	TV Show	International,Romantic,Dramas	[love, friday, club,first]	[love, life, woman, young]



Summary :

Netflix dataset has Movies/Shows info for total of 7787 contents added on or before Jan 2021.

IMDB dataset has several information out of which user ratings were used here by merging with Netflix dataset using title and release year.

Part1: EDA
Following are the takeaway points from exploratory data analysis:

- Majority of contents are TV-MA : **Mature Audience** only.
- Majority of contents were added after year 2015.
- Director column has 30% null values, out of which most of them are for TV shows and very few for movies.
- Majority of movies & shows are of around 100mins length & 1 season resp.
- Majority of Netflix movies have IMDB user ratings between 5-7 while for TV Shows its 6-8.

Part2: Country level analysis

- Top 3 countries based on Netflix movies are U.S., India & U.K. and based on TV Show are U.S., U.K. & Japan.
- India has produced more movies compared to shows, majority of which fall into International, Drama & comedy Genre. Top actors from India are Anupam Kher, Shahrukh Khan, Akshay kumar. IMDB user ratings of majority movies are in range of 5-7.
- U.S. has also produced more movies compared to shows, majority of which falls into comedy, drama & documentaries. Top actors from U.S. are Adam Sandler, Samuel Jackson, Laura Bailey. IMDB user ratings of majority movies are in range of 5-7.

Summary :

Part2: Country level analysis(cont)

- U.K. has also produced more shows compared to movies, majority of which falls into international & british genre. Top actors from U.K. are David Attenborough, Greg Davies & Harriet Walter. IMDB user ratings of majority shows are in range of 6-8.
- Japan has also produced more shows compared to movies, majority of which falls into Anime & international genre. IMDB user ratings of majority shows are in range of 6-8.

Part3: Movie vs TV Show in recent years

- Netflix is increasing both TV show & Movies contents.
- Netflix has focused more on movies in recent year not the other way round.
- Netflix has added 5186 movies and 2339 shows during 2016-2020 period(last 5 years),i.e. twice number of movies were added as compared to shows.

Part4 : Clustering

- Feature selected for text based clustering : type, director, cast, rating, genre & description.
- Top words extracted from Description are Life, young, new, family, etc.
- Movies are clustered into 8 clusters.
- TV Shows are clustered into 10 clusters.
- All contents are clustered into 12 clusters which are given relevant labels.

Thank You