Capstone Project

NETFLIX MOVIES AND TV SHOWS CLUSTERING

Team Members
Ajinkya Dakhale
Harshjot Singh
Suvir Kapse

Contents

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.
- In 2018, they released an interesting report which shows that the number of TV shows on Netflix has nearly tripled since 2010. The streaming service's number of movies has decreased by more than 2,000 titles since 2010, while its number of TV shows has nearly tripled. It will be interesting to explore what all other insights can be obtained from the same dataset.
- Integrating this dataset with other external datasets such as IMDB ratings, rotten tomatoes can also provide many interesting findings.

- **In this project, you are required to do**
  - Exploratory Data Analysis
  - Understanding what type content is available in different countries
  - Is Netflix has increasingly focusing on TV shows rather than movies in recent years.
  - Clustering similar content by matching text-based features

Data Summary

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	s1	TV Show	3%	NaN	João Miguel, Bianca Comparato, Michel Gomes, R...	Brazil	August 14, 2020	2020	TV-MA	4 Seasons	International TV Shows, TV Dramas, TV Sci-Fi &...	In a future where the elite inhabit an island ...
1	s2	Movie	7:19	Jorge Michel Grau	Demián Bichir, Héctor Bonilla, Oscar Serrano, ...	Mexico	December 23, 2016	2016	TV-MA	93 min	Dramas, International Movies	After a devastating earthquake hits Mexico Cit...
2	s3	Movie	23:59	Gilbert Chan	Tedd Chan, Stella Chung, Henley Hii, Lawrence ...	Singapore	December 20, 2018	2011	R	78 min	Horror Movies, International Movies	When an army recruit is found dead, his fellow...
3	s4	Movie	9	Shane Acker	Elijah Wood, John C. Reilly, Jennifer Connelly ...	United States	November 16, 2017	2009	PG-13	80 min	Action & Adventure, Independent Movies, Sci-Fi...	In a postapocalyptic world, rag-doll robots hi...
4	s5	Movie	21	Robert Luketic	Jim Sturgess, Kevin Spacey, Kate Bosworth, Aar...	United States	January 1, 2020	2008	PG-13	123 min	Dramas	A brilliant group of students become card-coun...

Dataset given to us consists of 7787 rows and 12 columns.

Data Summary

- 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 of movie / show

Null Value

- **Director** feature have more than 30.68% of null values. Filling null values by ‘unknown’.
- **Country** feature have 6.51% of null values. Filling null values by ‘unknown’.
- **Cast** feature have 9.22% of null values. Filling null values by ‘unknown’.
- **Rating** feature have 0.09% of null values. Dropping rows corresponding to null values.
- **Date_added** feature have 0.13% of null values. Dropping rows corresponding to null values.

DATA MANIPULATION

Added year and month column obtained from year date added column. Assigned the Ratings into grouped categories.

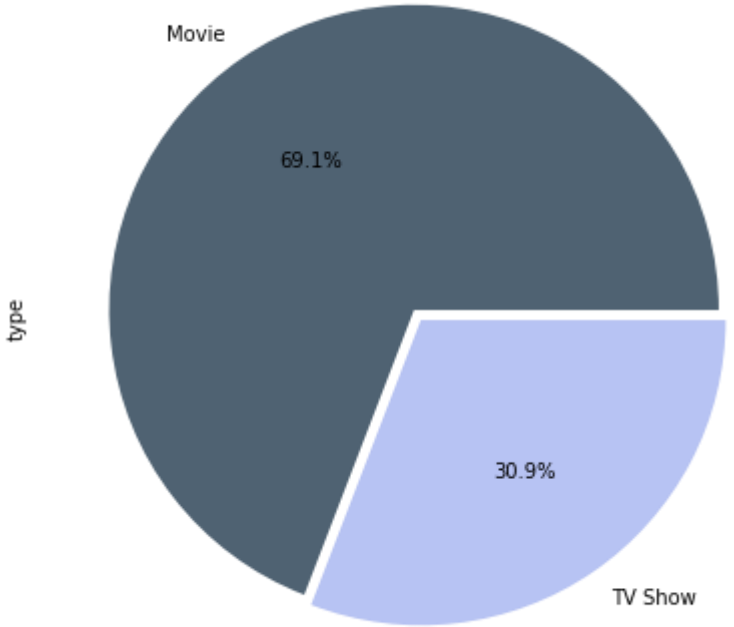
- Such as –

- ratings_ages = {'TV-PG': 'Older Kids', 'TV-MA': 'Adults', 'TV-Y7-FV': 'Older Kids', 'TV- Y7': 'Older Kids', 'TV-14': 'Teens', 'R': 'Adults', 'TV-Y': 'Kids', 'NR': 'Adults', 'PG-13': 'Teens', 'TV-G': 'Kids', 'PG': 'Older Kids', 'G': 'Kids', 'UR': 'Adults', 'NC-17': 'Adults'}

Exploratory Data Analysis

Type of content available on Netflix

- It is evident that there are more movies on Netflix than TV shows.
- Netflix has 5377 movies, which is more than double the quantity of TV shows.



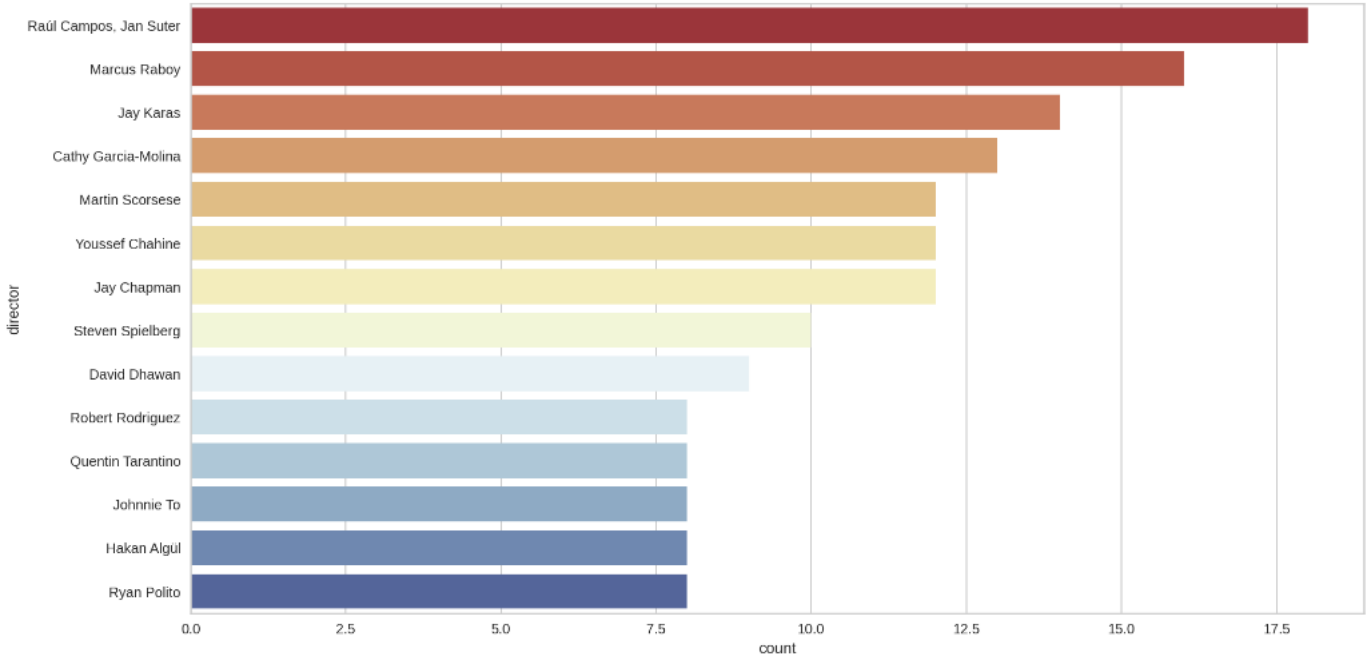
EDA

(Title)

- Christmas, love, Girl and Story are few Among the other words which are frequently used in title of Movies and TV shows

EDA

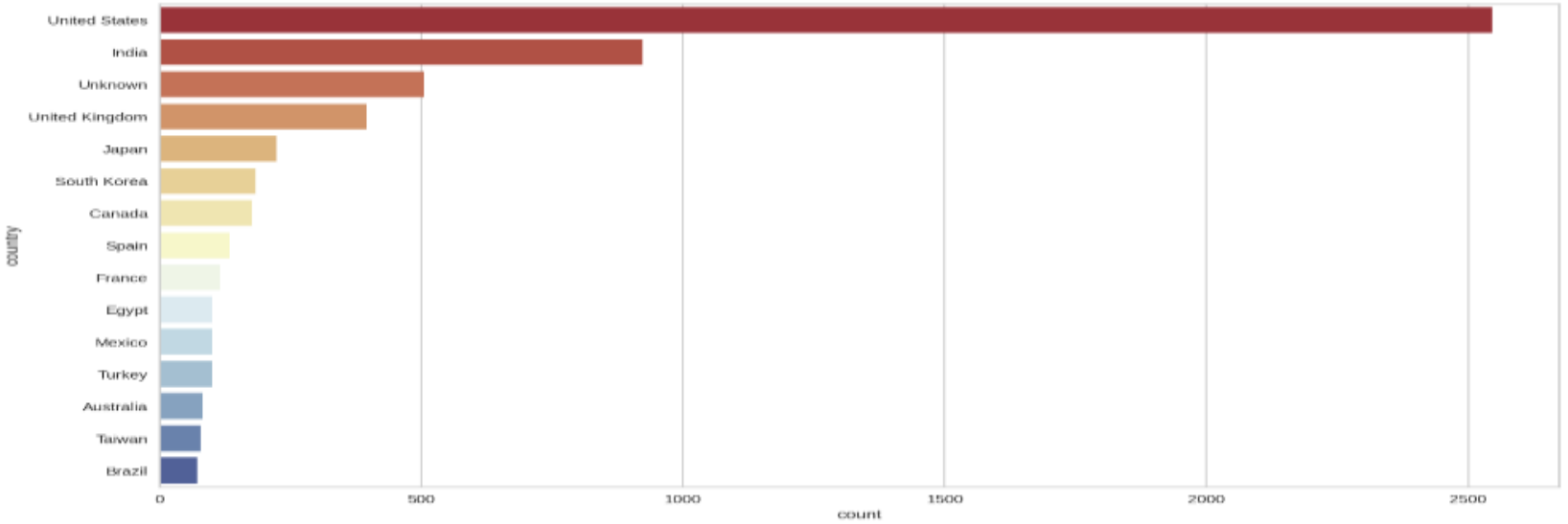
Director



Raul Campos and Jan Sulter collectively have the most content on Netflix

EDA

(Country)



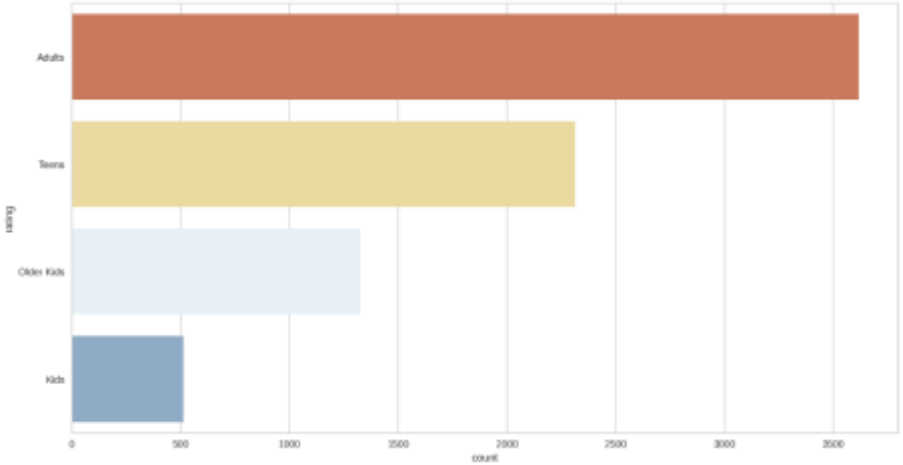
Top 20 Countries with most Content

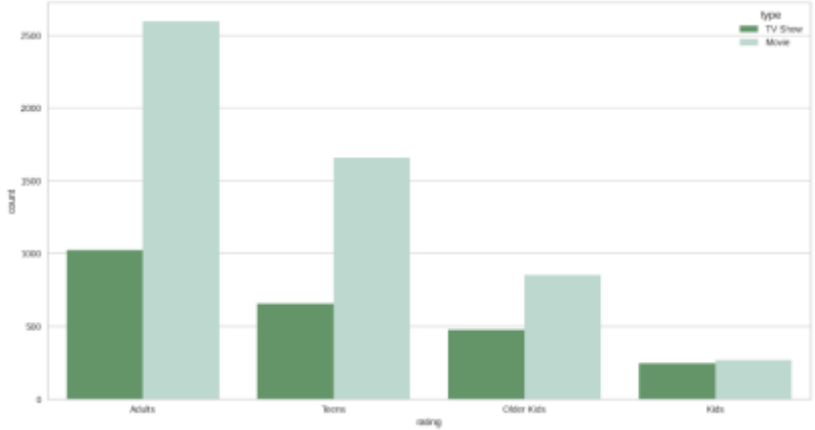
The United States has the most number of content on Netflix by a huge margin followed by India.

EDA

(Rating)

Most of the content made are for adult followed by teens

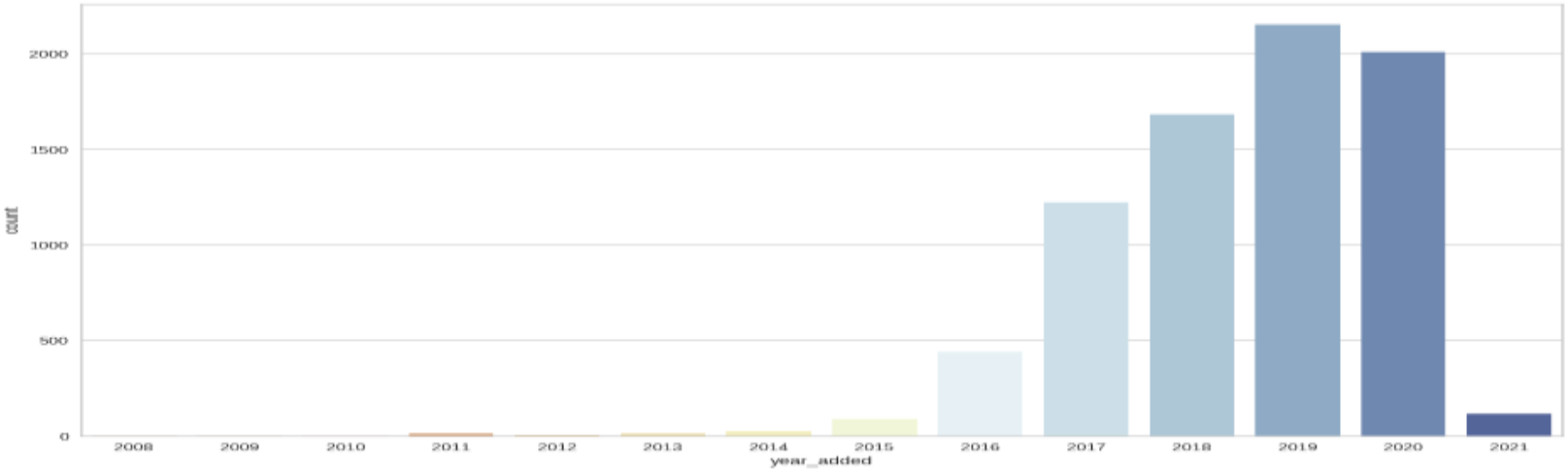




'TV-PG': 'Older Kids', 'TV-MA': 'Adults', 'TV-Y7-FV': 'Older Kids', 'TV-Y7': 'Older Kids', 'TV-14': 'Teens', 'R': 'Adults', 'TV-Y': 'Kids',‘ NR': 'Adults', 'PG-13': 'Teens‘, 'TV-G': 'Kids', 'PG': 'Older Kids', 'G': 'Kids', 'UR': 'Adults', 'NC-17': 'Adults'

EDA

(year_added)

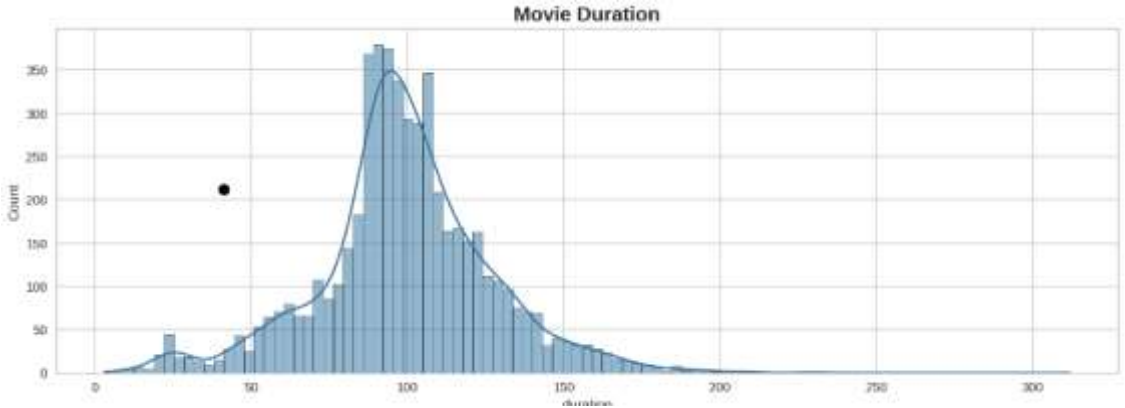


The number of release have significantly increased after 2015 and have dropped in 2021 because of Covid 19.

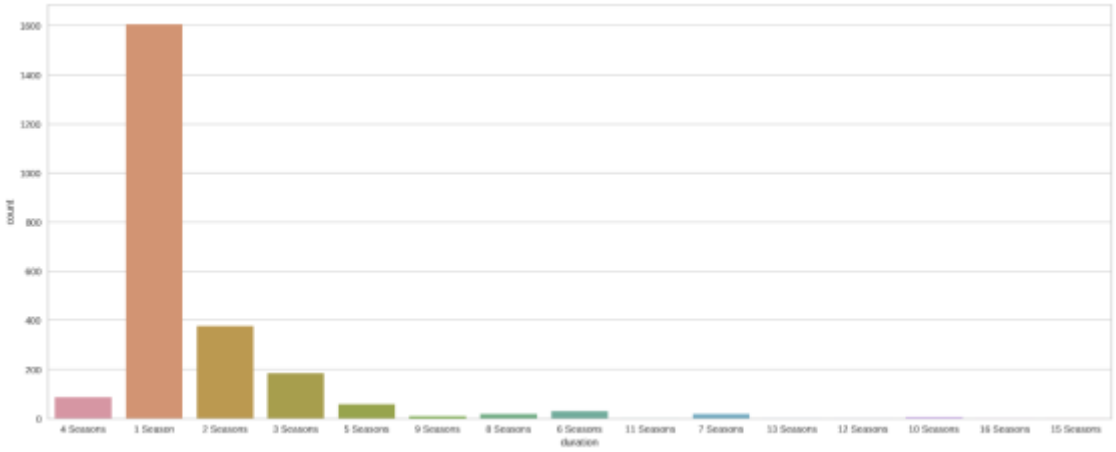
EDA

(Duration)

Movie Duration in (min)
Average movie length is 75-120 Min.

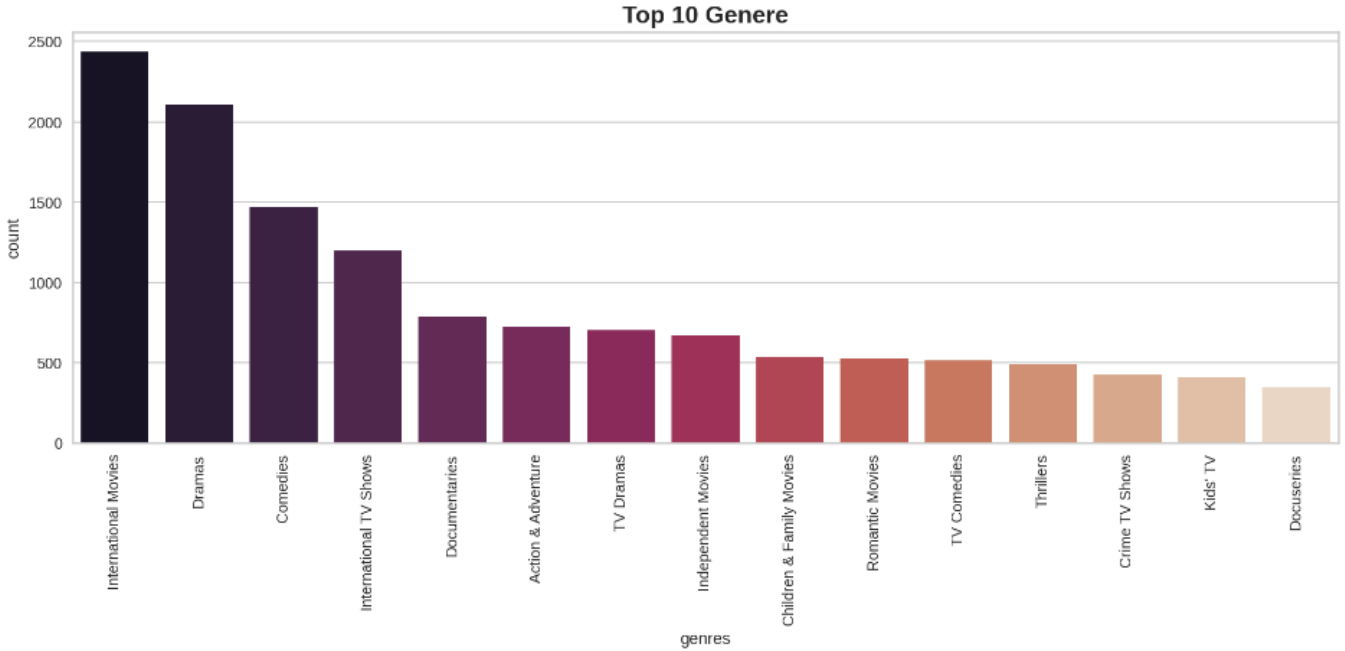


TV shows Duration in Seasons
Majority TV shows has 1 season



EDA

Genres



Top 15 genres to which Netflix content belongs. International movies & Dramas at top.

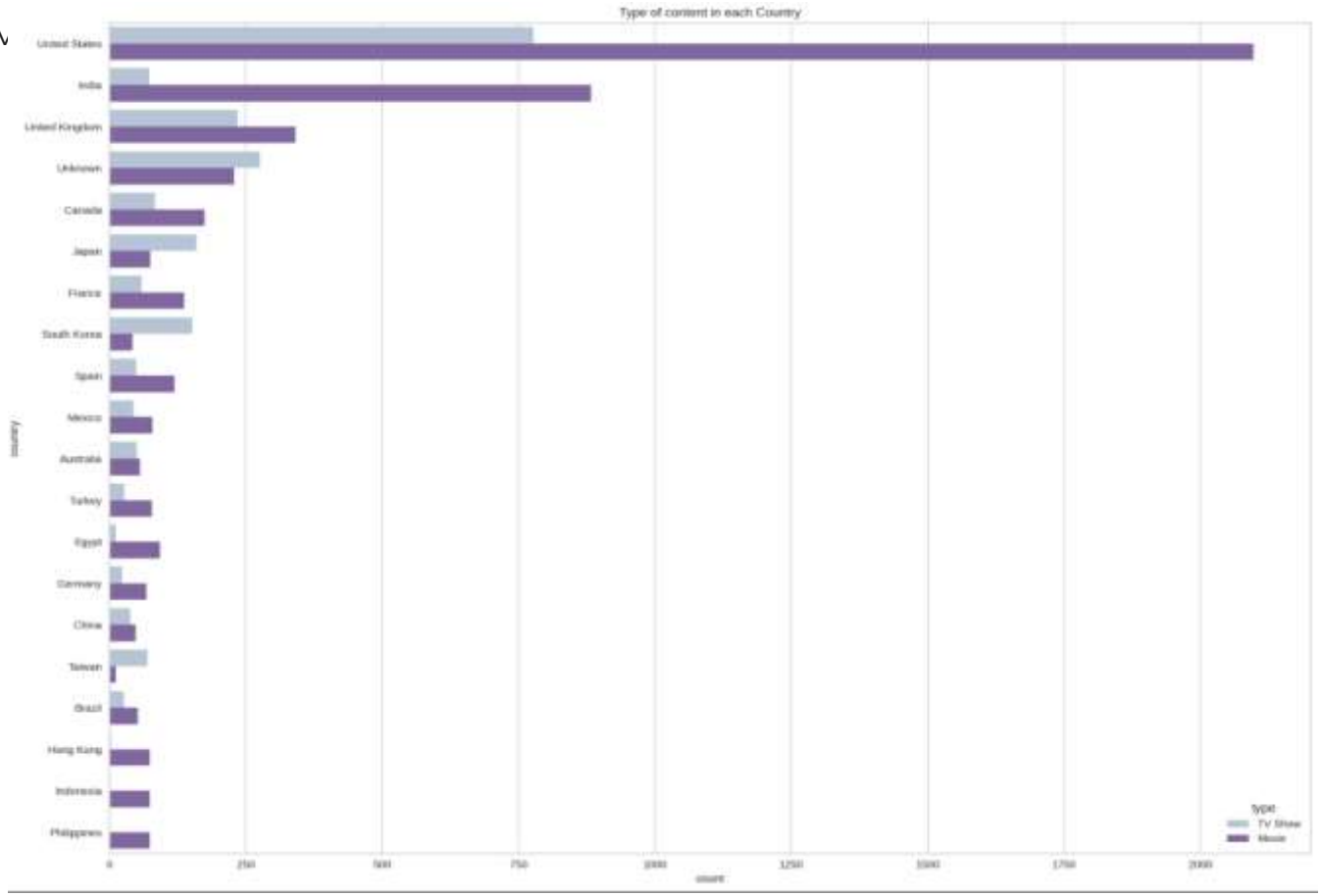
EDA

Description

Family, Life, Find, World are few among the other words, which are used frequently in the Description of Netflix content (Tv shows and movies).

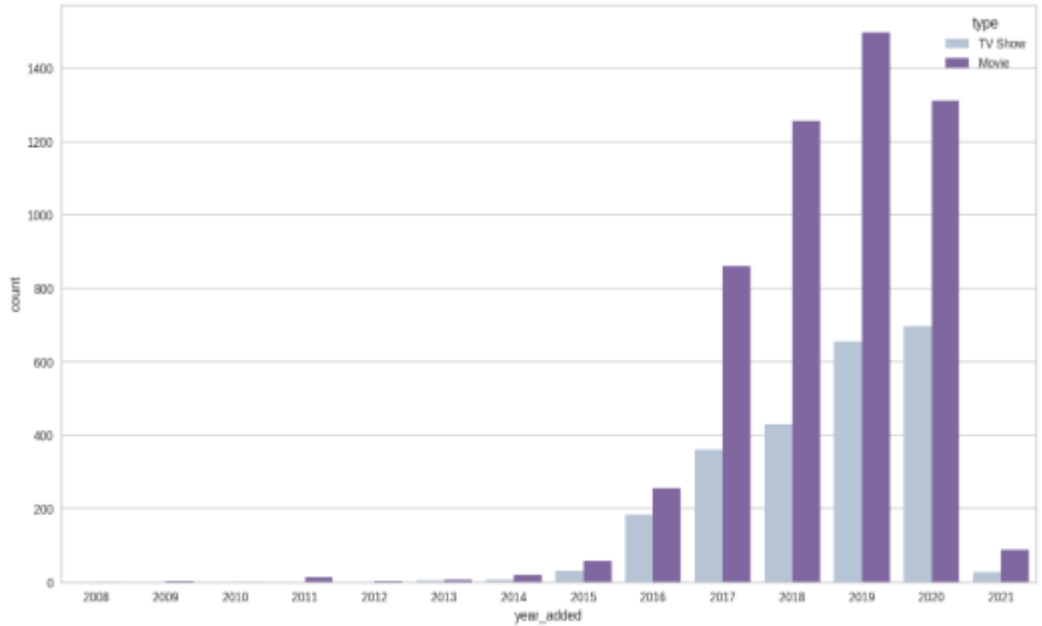
Understanding what type content is available in different countries

- Most of the countries hav more movies than TV Shows.
- The United States is a leading producer of both types of content; this makes sense since Netflix is a US company.
- The influence of Bollywood in India explains the type of content available, and perhaps the main focus of this industry is Movies and not TV Shows.
- For South Korea and Japan it's the opposite. It Maybe because K-Dramas And Anime are more Popular in these two countries respectively.



Is Netflix has increasingly focusing on TV rather than movies in recent years ?

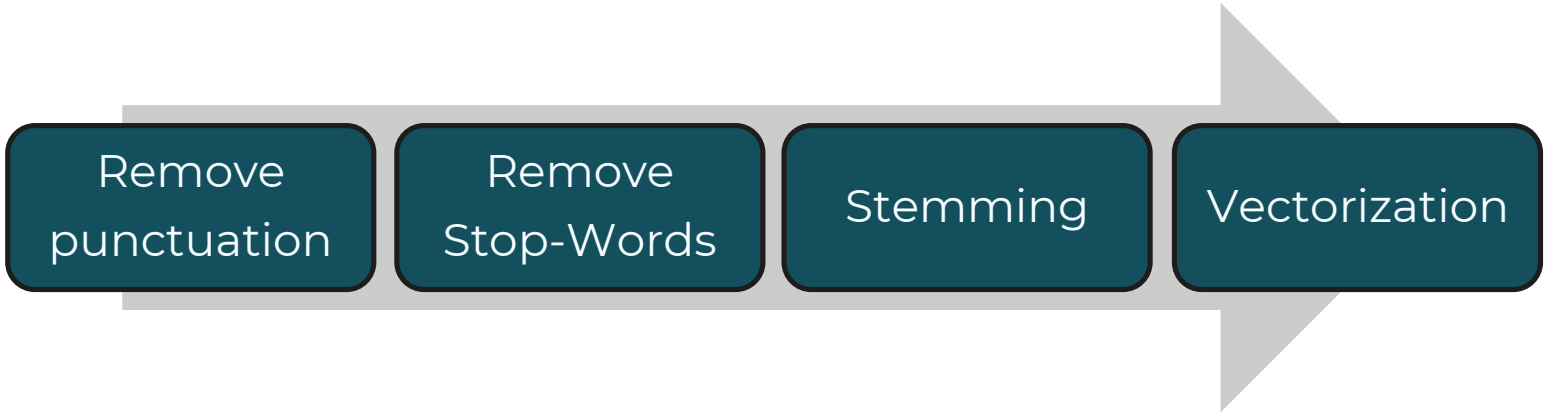
- Till 2019 we can see there is gradual rise in both movies and TV show but in 2020 we can see that is a bit drop in movies whereas the growth of TV shows remains the same.



So yes we can say Netflix has increasingly focusing on TV rather than movies in recent years

Data Pre-processing

We cannot go straight from raw text to fitting a machine learning model. We must clean text first, which means splitting it into words and handling punctuation.
For clustering we choose “description” ,”title” and “genres” variables. Before clustering we need to pre process the data. So that we filtered data with following steps:

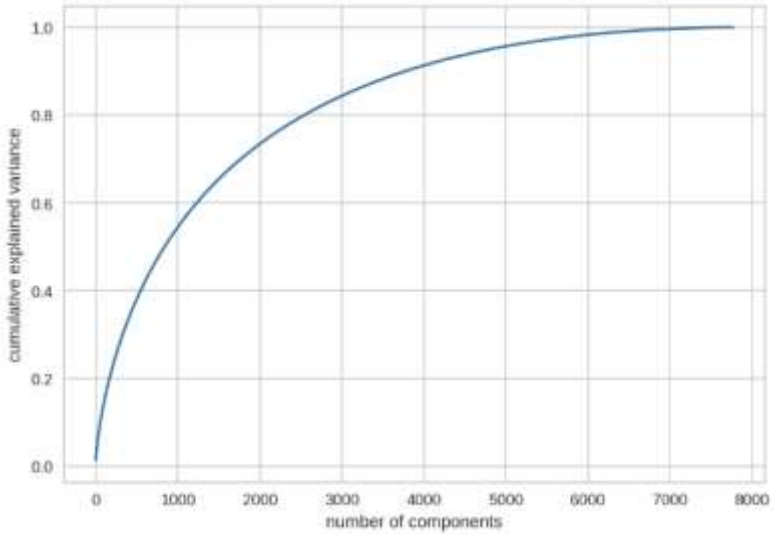


TfidfVectorizer And PCA

- TfidfVectorizer - Transforms text to feature vectors that can be used as input to estimator.
- TF-IDF stands for Term Frequency — Inverse Document Frequency and is a statistic that aims to better define how important a word is for a document, while also taking into account the relation to other documents from the same corpus.
- PCA is dimensionality reduction algorithm . It can be used for data visualization, noise filtering, for feature engineering.

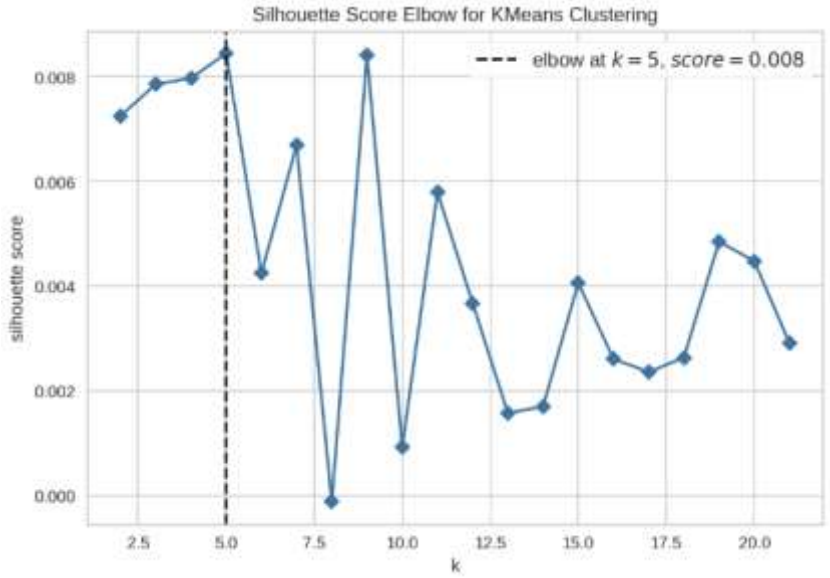
```
[ ]  #using tfidf transforming data

     tfidf = TfidfVectorizer(max_df = 0.9,min_df = 1,max_features=12000)
     x= tfidf.fit_transform(filtered_text)
```

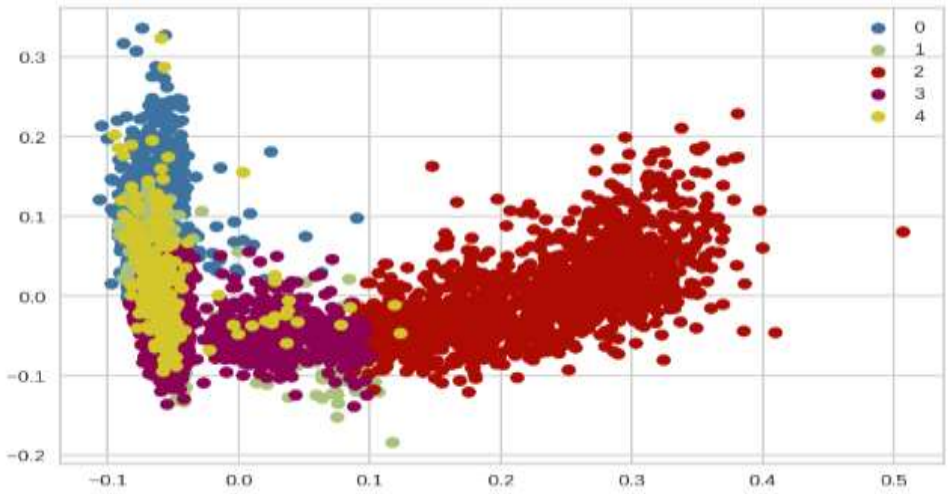


K- Means Clustering

K-means algorithm is an iterative algorithm that tries to partition the dataset into K pre defined distinct non overlapping subgroups where each data point belongs to only one group.



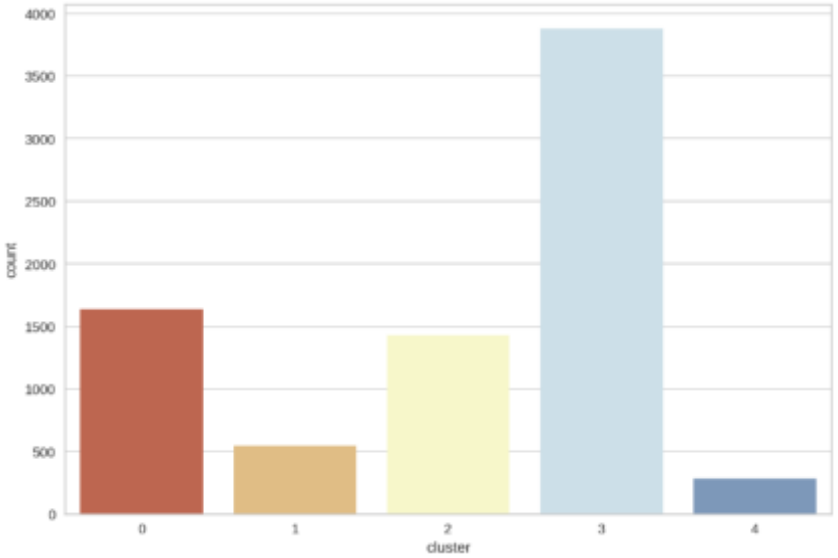
Clusters for k-means clustering

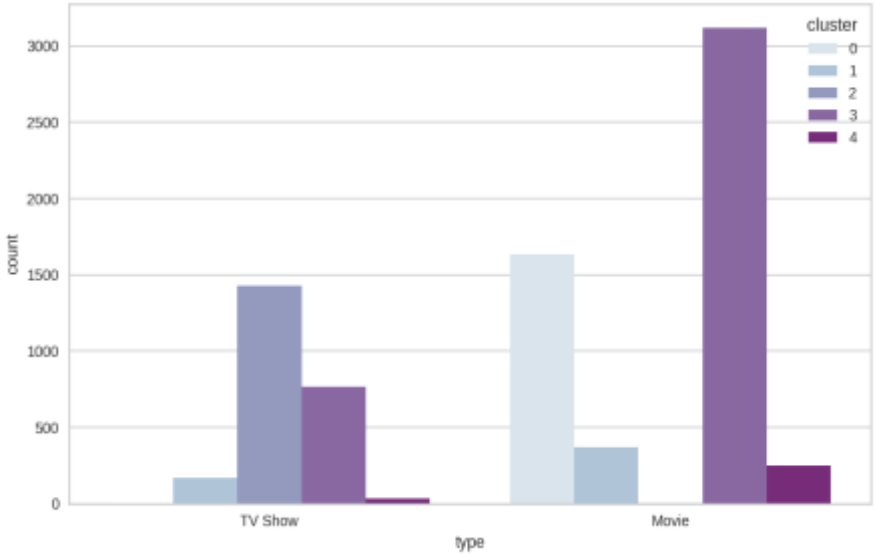


- The numbers 0 to 4 represent 5-distinct clusters formed by K-means clustering.
- Each cluster contains data points similar to those in the same groups but varies from other groups.

Analysis on Cluster

From below graph we can infer that cluster #3 contains most amount of words followed by cluster #0.





- Cluster #3 contains most data on movies while cluster #2 contains least/none.
- Cluster #2 contains most data on TV shows while cluster #0 contains least/none

Conclusion

- The Netflix has been increasingly focusing on TV show rather than movies in recent year, as we observed in year 2020 with decline in growth of movies while the growth of TV shows remains the same.
- USA and India are two countries producing the maximum number of content.
- International movies, Drama and Comedy are top genre on Netflix platform. Further we found number of movies on Netflix outnumbers TV-shows..
- Most TV shows end by 1st season and duration of Most movies on Netflix is in 75-120 mins range.
- We Performed K-Means Clustering to create clusters and used silhouette score and elbow curve to find optimal number of clusters k=5.
- We've defined 5 clusters and implemented the KMEANS clustering algorithm. And then we determined that cluster number 3 covers most data than any other.