Capstone Project-4

NETFLIX MOVIES AND TV SHOWS CLUSTERING

By
Rajat Arya
Soumyadeep Pandit

Contents Of The Presentation

- Defining the problem statement
- Data Summary
- Data Wrangling
- Eda and Data Visualization
- Applying the Model
- Optimal value of K
- Recommender System
- Challenges
- Conclusion

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.

In this project, you are required to do

1. Exploratory Data Analysis
2. Understanding what type 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

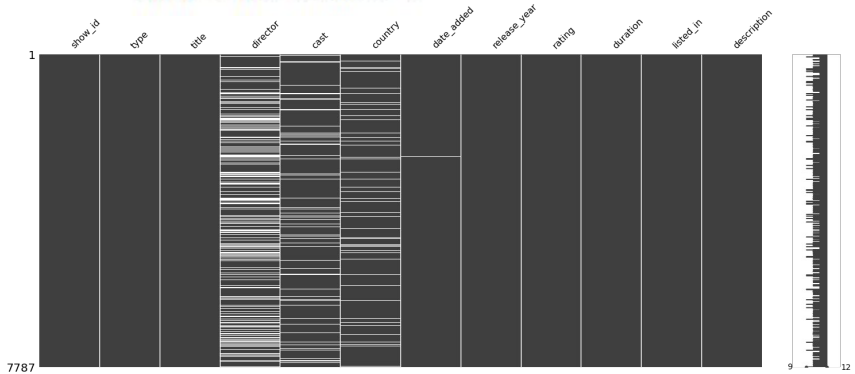
- Size of Dataset: 7787 rows, 12 columns
  - show_id : Unique ID for every Movie / Tv Show
  - type : Identifier - A Movie or TV Show
  - title : Title of the Movie / Tv Show
  - director : Director of the Movie
  - cast : Actors involved in the movie / show
  - country : Country where the movie / show 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 : Genres
  - description: The Summary description

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7787 entries, 0 to 7786
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   show_id       7787 non-null   object
 1   type          7787 non-null   object
 2   title         7787 non-null   object
 3   director      5398 non-null   object
 4   cast          7069 non-null   object
 5   country       7280 non-null   object
 6   date_added    7777 non-null   object
 7   release_year  7787 non-null   int64
 8   rating        7780 non-null   object
 9   duration      7787 non-null   object
 10  listed_in     7787 non-null   object
 11  description   7787 non-null   object
dtypes: int64(1), object(11)
memory usage: 730.2+ KB
```

Data Wrangling

- There is no duplicate observation present in our dataset.
- We will need to replace blank countries with the mode (most common) country.
- It would be better to keep director because it can be fascinating to look at a specific filmmaker's movie. As a result, we substitute the null values with the word 'unknown' for further analysis.
- It would be better to keep the cast because it can be fascinating to look at the films of a specific cast. As a result, we substitute the null values with the word 'unknown' for further analysis.
- There is no logical way to dealing with date so we remove the null values of date column.


```
director null values percentage: 30.68%
 null value counts: 2389
cast null values percentage: 9.22%
 null value counts: 718
country null values percentage: 6.51%
 null value counts: 507
date_added null values percentage: 0.13%
 null value counts: 10
rating null values percentage: 0.09%
 null value counts: 7
```

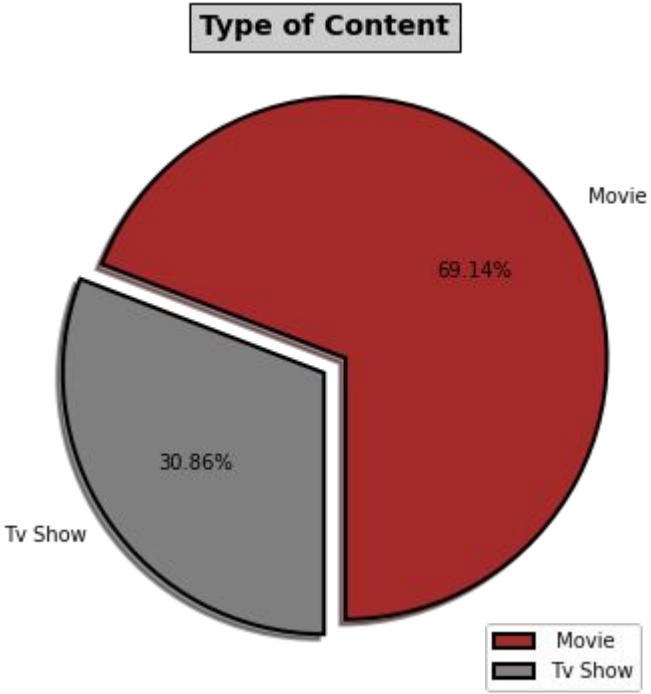


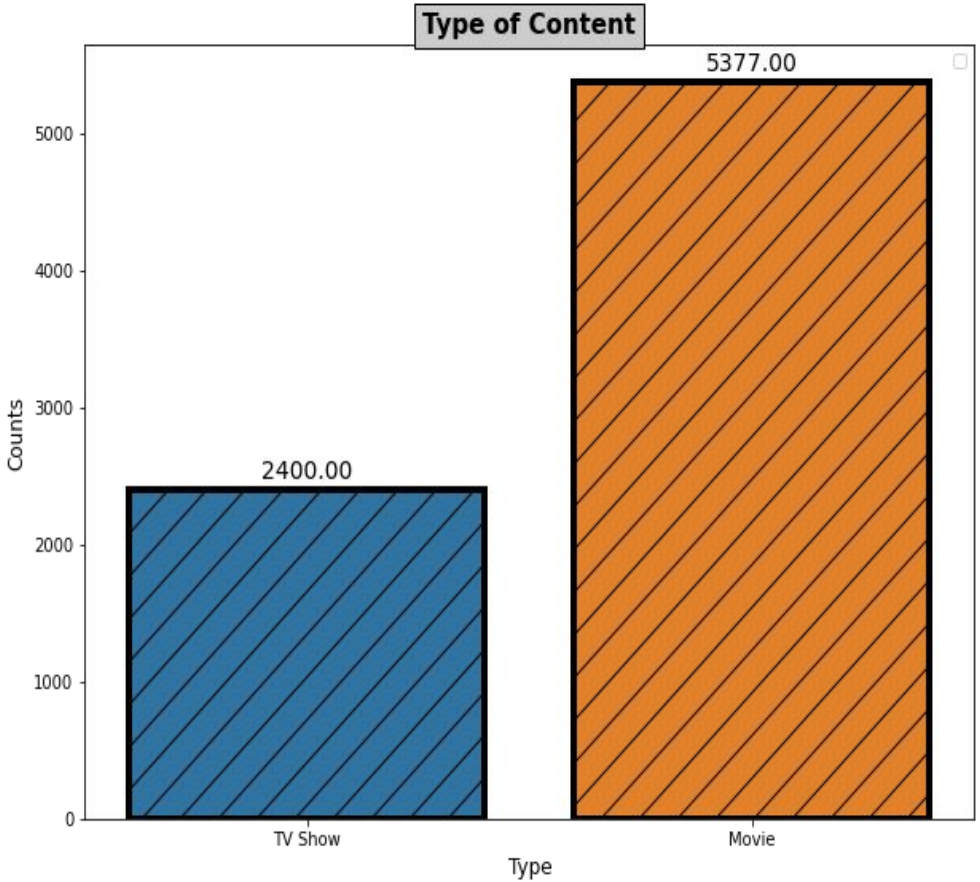
Exploratory Data Analysis(EDA) and Data Visualization

- Exploratory Data Analysis (EDA) refers to the critical process of performing initial investigations on a dataset (import from CSV, Jason, and Html files) to summarize their main characteristics, often with visual methods and check assumptions with the help of summary statistics and graphical representations. EDA is used for analyzing what the data can tell us before the modeling or by applying any set of instructions/code. When you are working with the datasets, it is not easy to determine the important characteristics of the data by looking at the column of numbers or a whole spreadsheet/dataset. It may be tedious, boring, and/or overwhelming to determine experiences by seeing plain numbers. Exploratory data analysis techniques have been devised as an aid in this situation.
- Data visualization is the process of translating large data sets and metrics into charts, graphs and other visuals. The resulting visual representation of data makes it easier to identify and share real-time trends, outliers, and new insights about the information represented in the data.

Type Of Content

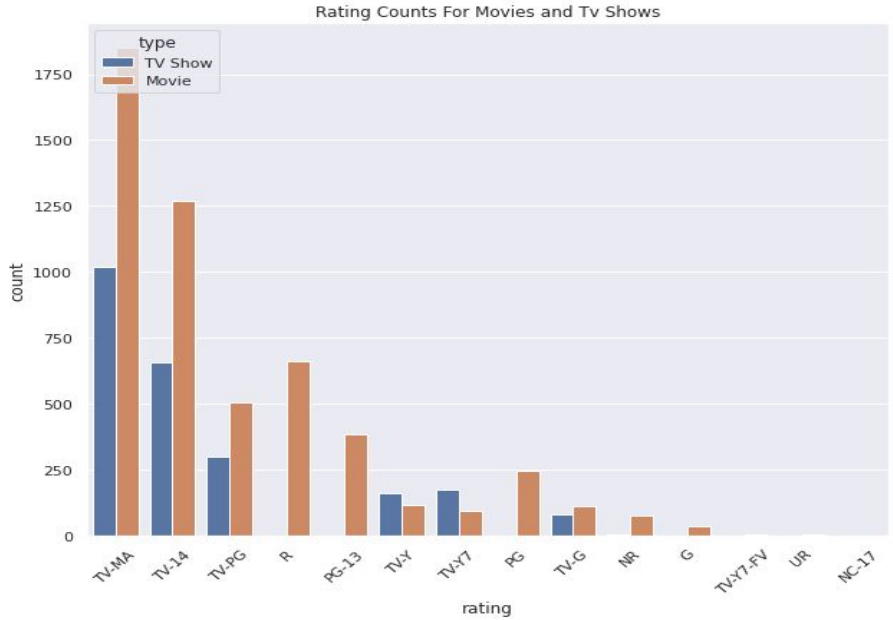
- It is evident that there are more movies on Netflix than TV shows.
- Netflix has 5377 movies, which is more than double the quantity of TV shows.

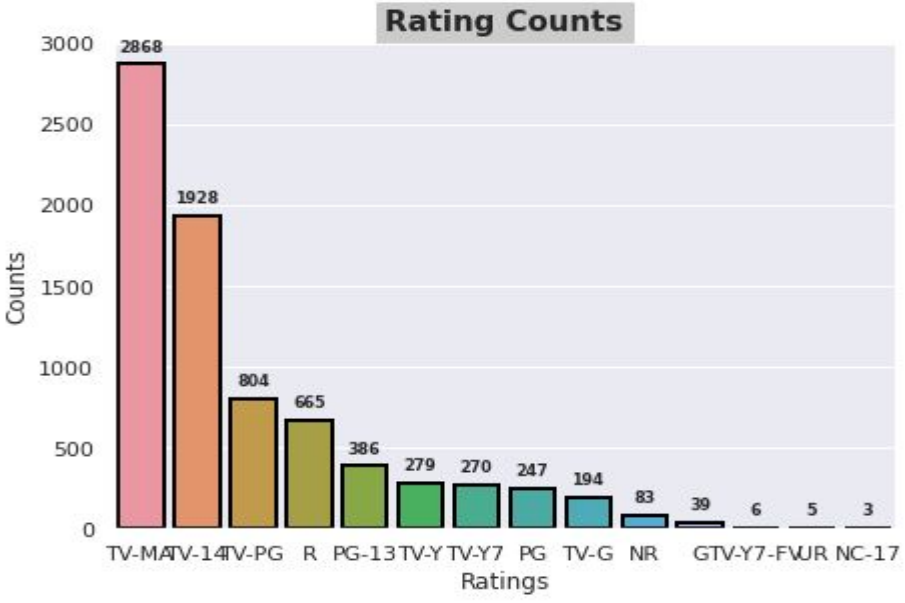




Movie Ratings Analysis

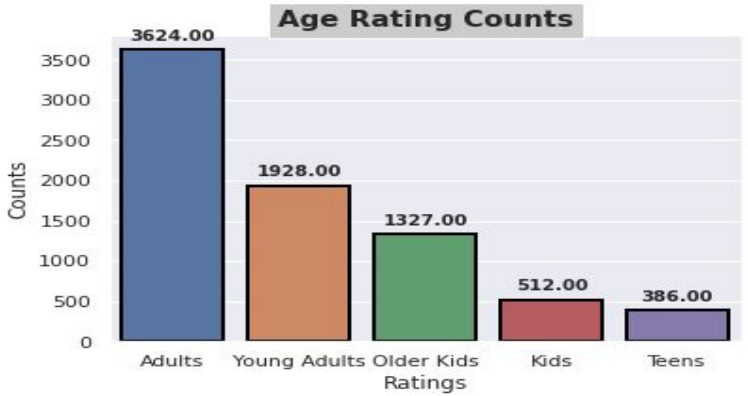
- The 'TV-MA' rating is used in the majority of the film. The TV Parental Guidelines provide a "TV-MA" classification to a television programme that is intended solely for mature audiences
- The second largest is 'TV-14,' which stands for content that may be inappropriate for minors under the age of 14.
- The third most common is the extremely popular 'R' rating. The Motion Picture Association of America defines an R-rated film as one that contains material that may be inappropriate for children under the age of 17; the MPAA states that "Under 17 requires accompanying parent or adult guardian."

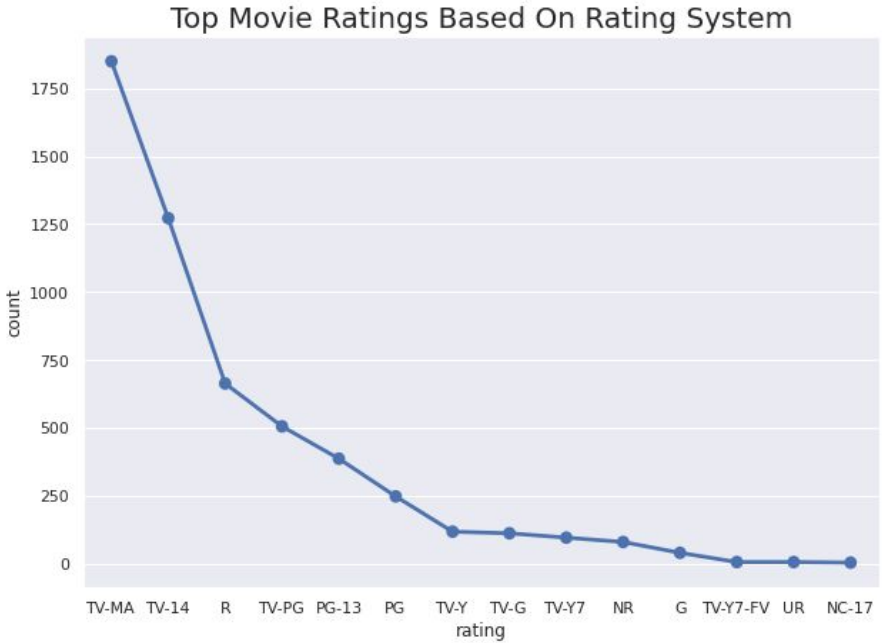


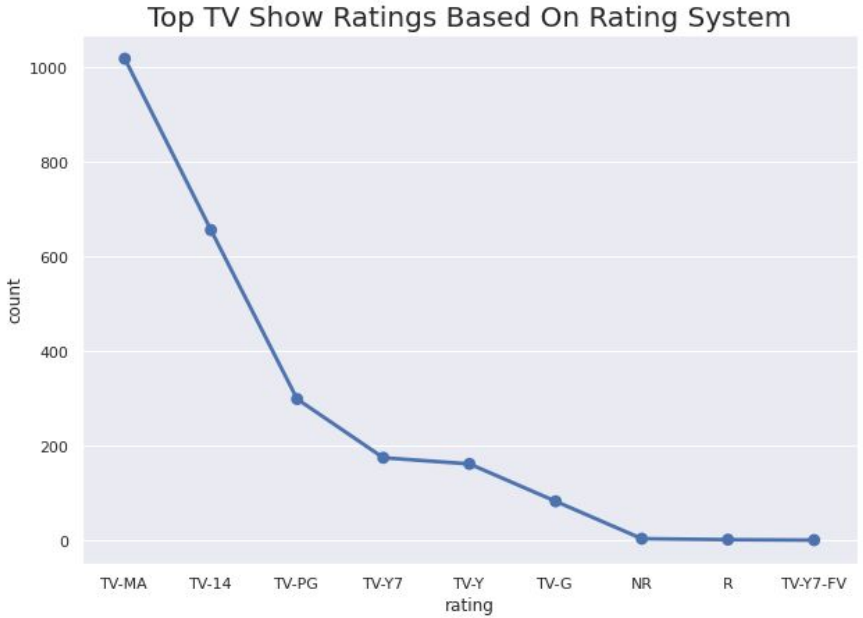


Movie Rating Analysis

- We can observe from the above count plot that the majority of Netflix material is intended for adults. There is very little content available for teens and kids.

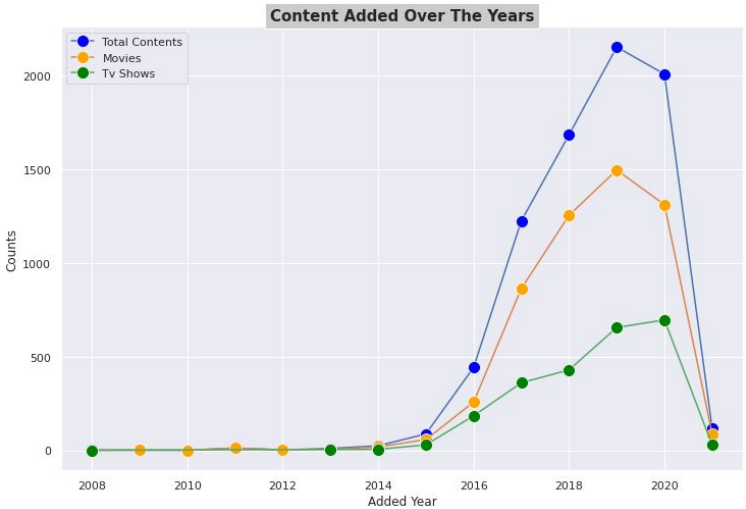


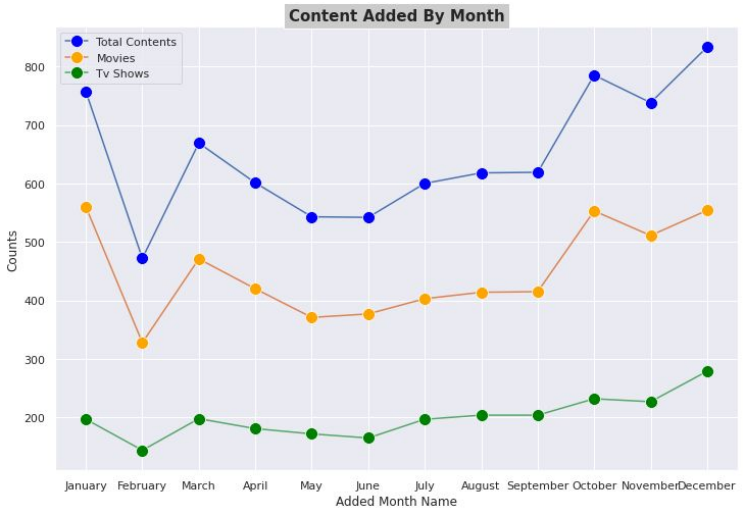




Content Growth

- The number of movies on Netflix is growing significantly faster than the number of TV shows.
- In both 2018 and 2019, approximately 1200 new movies were added.
- We saw a huge increase in the number of movies and television episodes after 2014.
- Because of covid-19, there is a significant drop in the number of movies and television episodes produced after 2019.
- The above graph shows that the most content is added to Netflix in December.
- In February, Netflix adds extremely few new movies and television episodes.

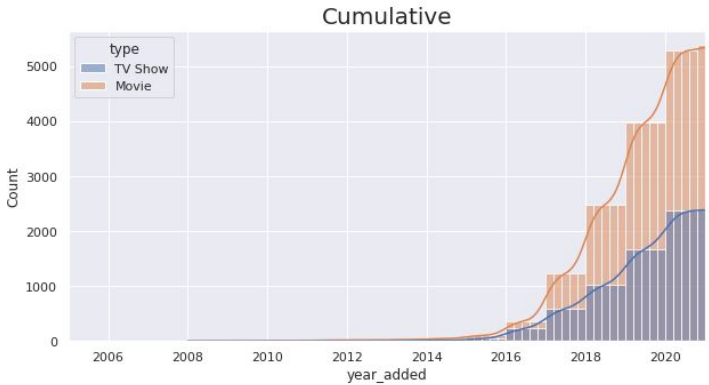


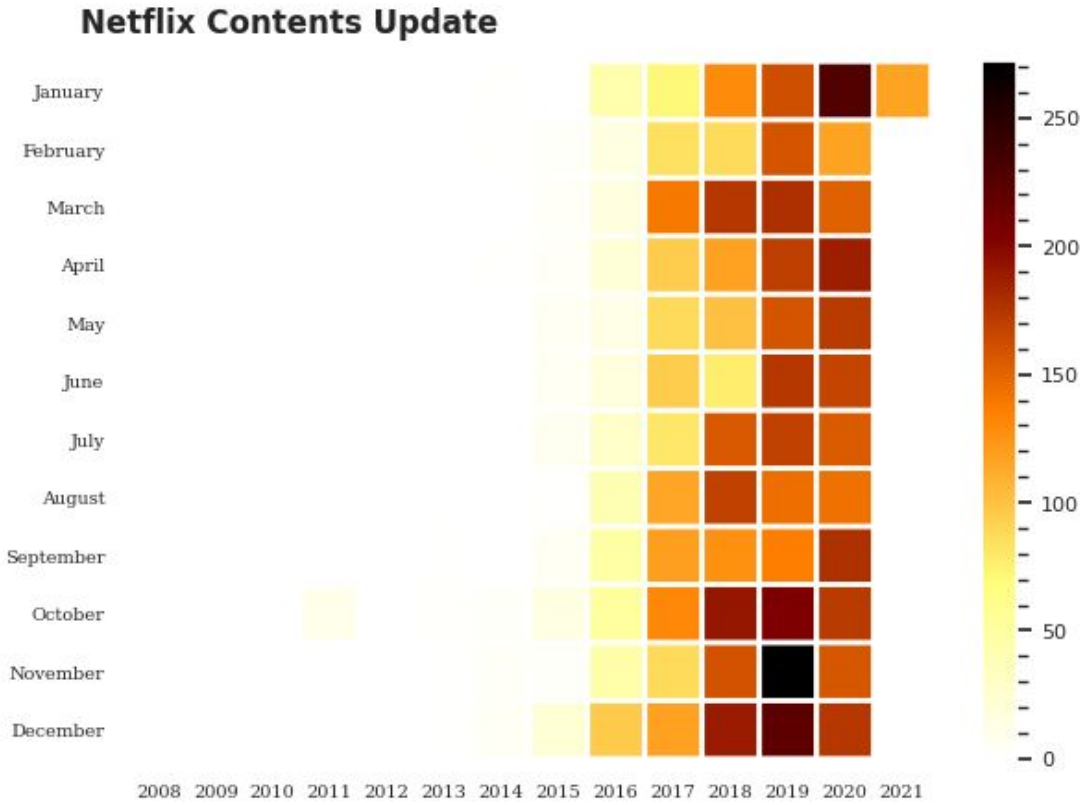


Content Growth

If the latest year 2019 is considered, January and December were the months when comparatively much less content was released.Therefore, these months may be a good choice for the success of a new release!

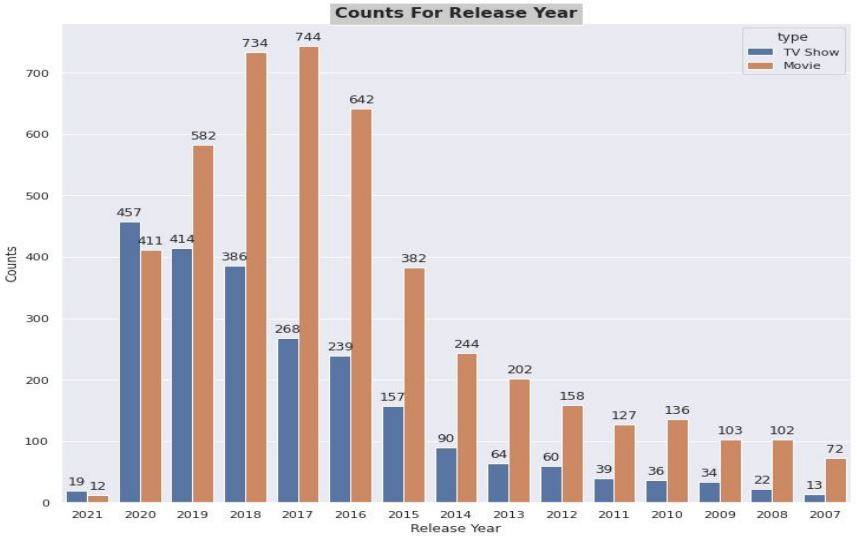
It appears that Netflix has focused more attention on increasing Movie content that TV Shows. Movies have increased much more dramatically than TV shows.





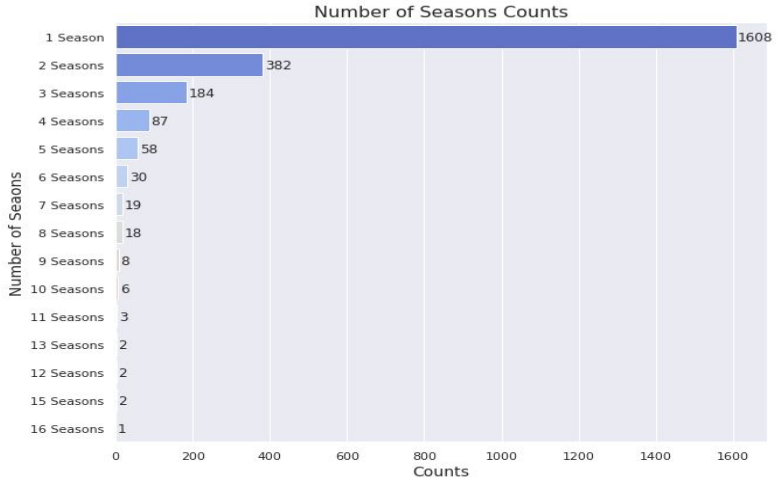
Release Year Analysis

- According to the following countplot, 774 movies were released in 2017, which is the most number of releases in any year.
- According to the above countplot, there will be 457 television shows released in 2020, which is the most of any year.



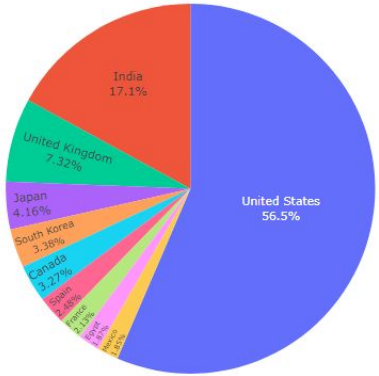
Seasons Count

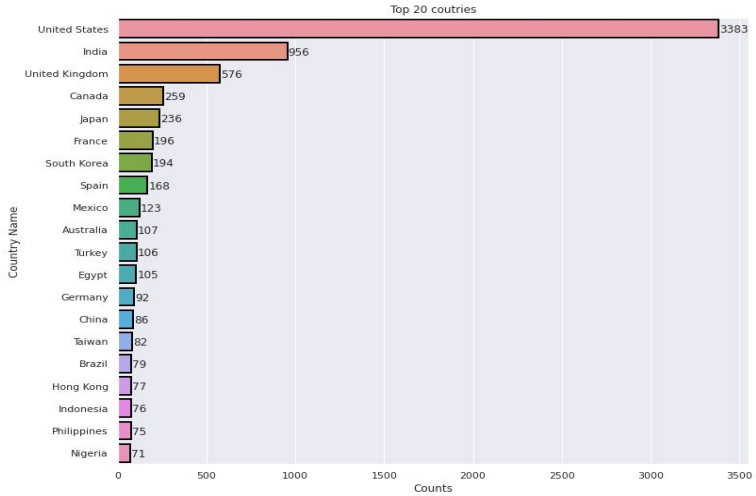
- According to the above counterplot, each of the 1608 television shows had only one season.
- There were extremely few television shows that had more than six seasons.

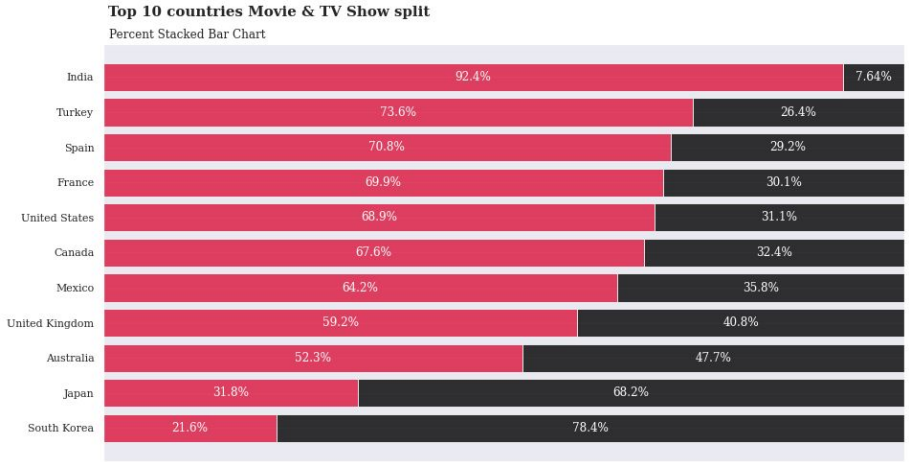


Country Analysis

- The United States is the most prolific generator of Netflix content, with India and the United Kingdom trailing far behind.
- The majority of the content on Netflix in India is comprised of movies.
- Bollywood is a significant business, and movies, rather than TV shows, may be the industry's major focus.
- South Korean Netflix on the other hand is almost entirely TV Shows.
- The fundamental reason for the variation in content must be due to market research undertaken by Netflix.

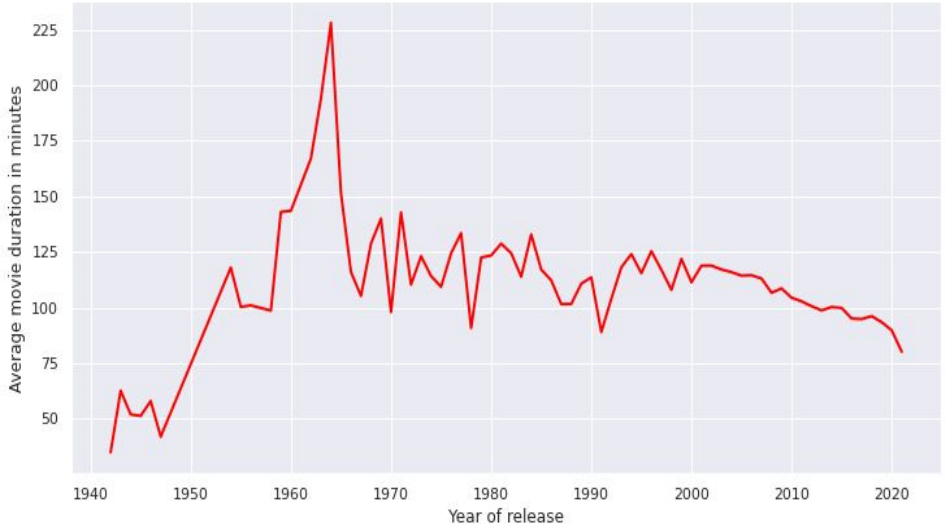


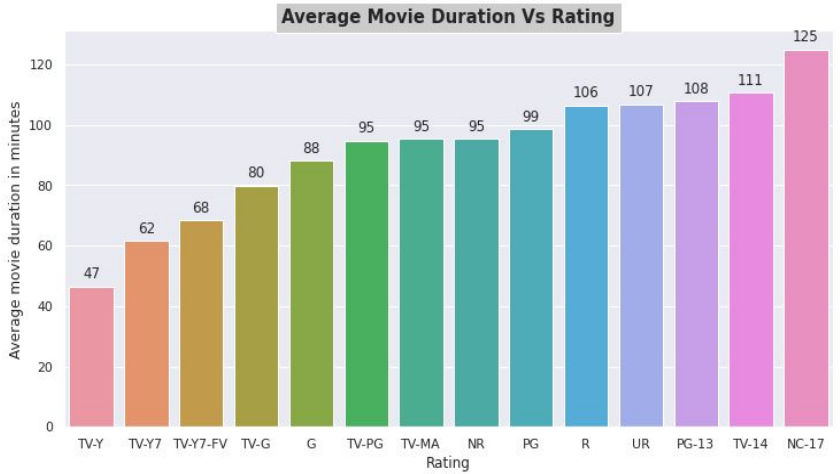




Movie Duration Trends

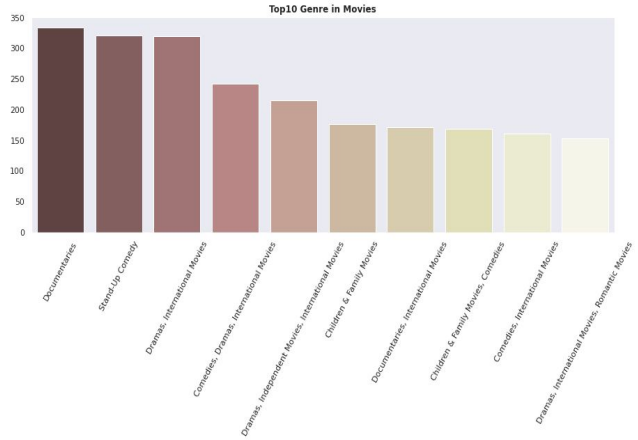
- Movies made before 1948 have a short average duration, compared to those released after 1948.
- The average movie duration released during the 1960 era was the longest.
- At the beginning of the twentieth century, the average length of a film was decreasing over time.
- Those movies that have a rating of NC-17 have the longest average duration.
- When it comes to movies having a TV-Y rating, they have the shortest runtime on average.

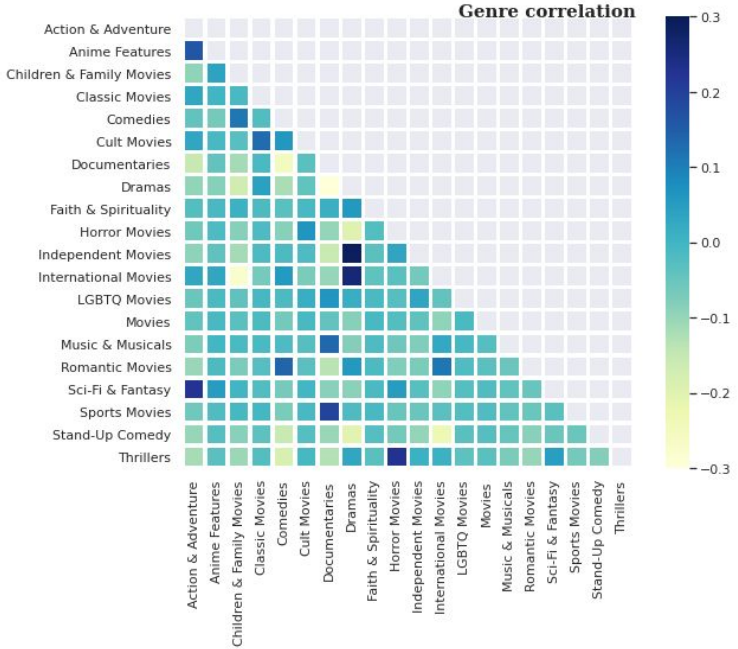


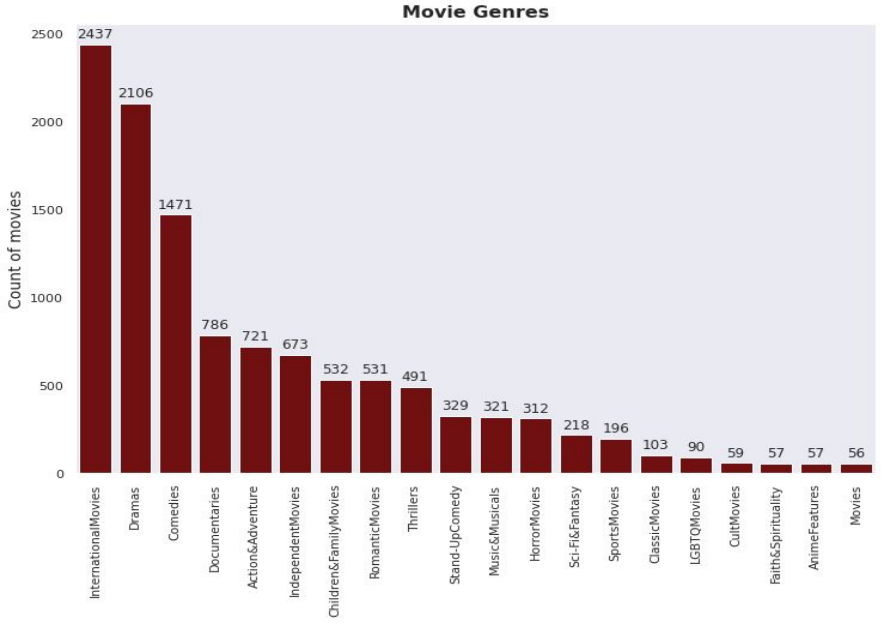


Analysis of Movie Genres

- As a result, it is evident that international movies, dramas, and comedies are the top three genres with the most content on Netflix.
- It is interesting that International Movies tend to be Dramas.

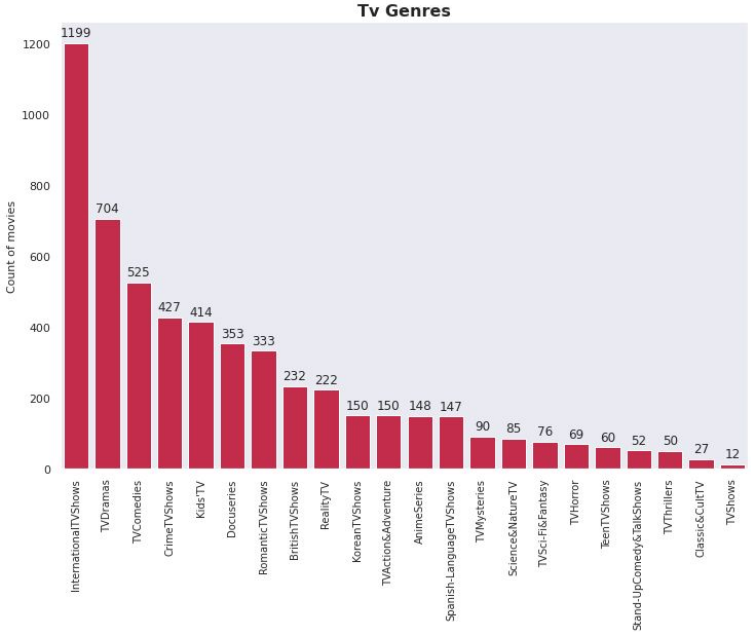


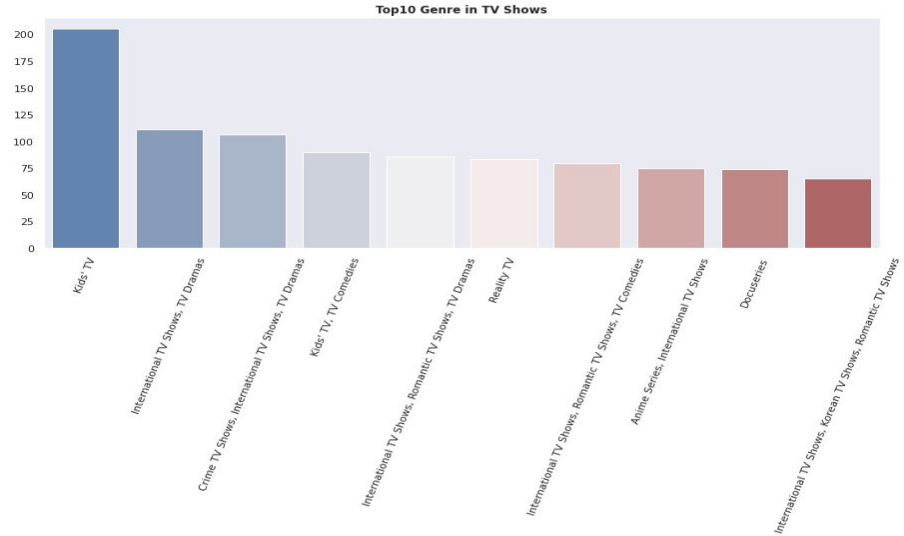




Analysis of Tv Show Genres

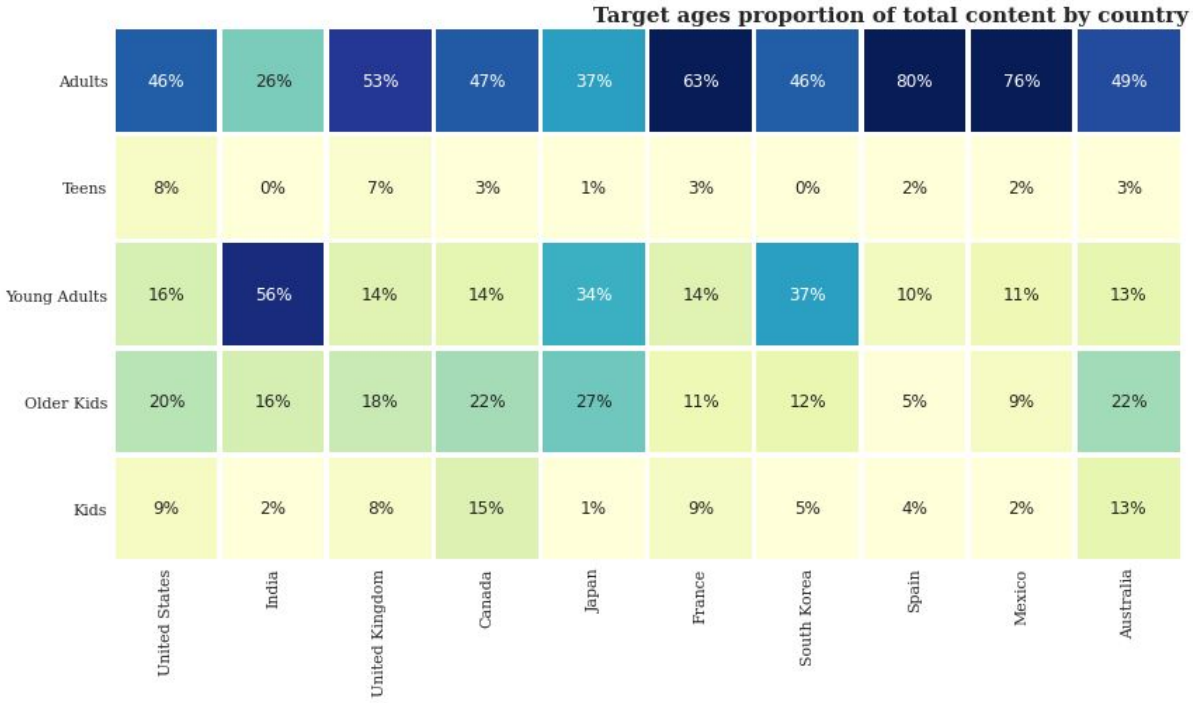
- As a result, it is evident that international tv shows, tv dramas, and tv comedies are the top three genres with the most content on Netflix.





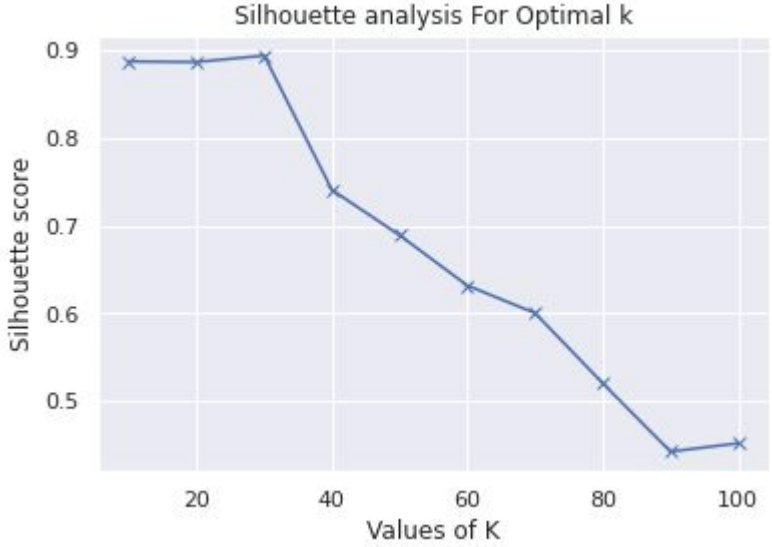
Netflix Content for different age groups in top 10 countries

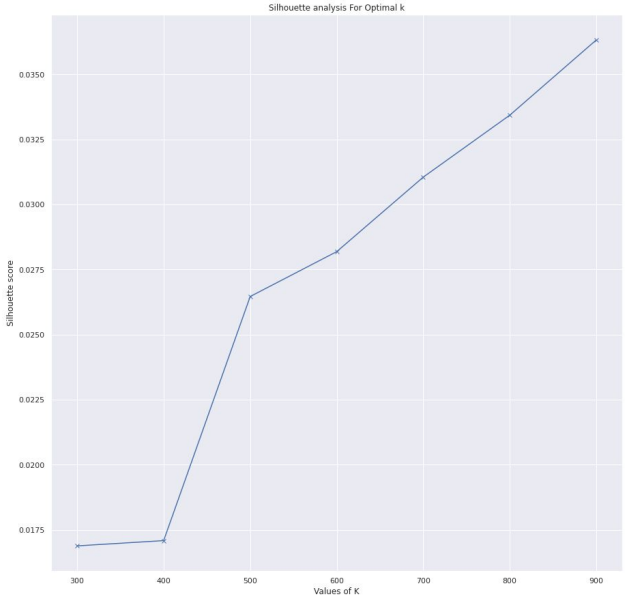
- It is also interesting to see parallels between culturally comparable nations - the US and UK are closely aligned with their Netflix target ages, but radically different from, example, India or Japan!
- Also, Mexico and Spain have similar content on Netflix for different age groups.

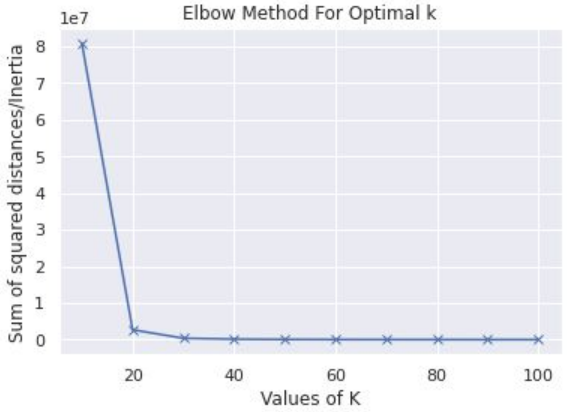


Optimal Value of k

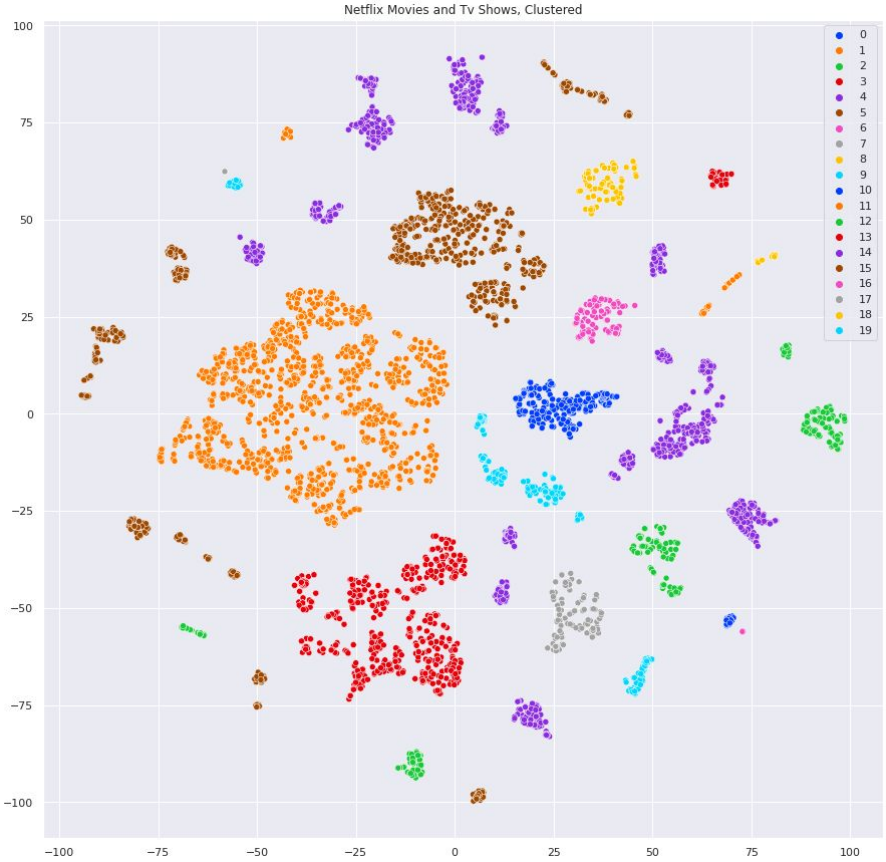
- The optimal value of K is 20.
- For the alternate method, the optimal value of k is 800.







Cluster




```
# Get cluster number for TV show
cluster_num = df2[df2.title=='Breaking Bad'].cluster.item()

# View cluster the TV show belongs to
df2[df2.cluster == cluster_num]
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description	cluster
565	s566	TV Show	Aquarius	NaN	David Duchovny, Gethin Anthony, Grey Damon, Em...	United States	June 16, 2017	2016	TV-MA	2 Seasons	Crime TV Shows, TV Dramas	Amid the turmoil of 1960s LA, two cops and a p...	243
761	s762	TV Show	Bates Motel	NaN	Vera Farmiga, Freddie Highmore, Max Thieriot, ...	United States	February 20, 2018	2017	TV-MA	5 Seasons	Crime TV Shows, TV Dramas, TV Horror	When his father dies, Norman Bates and his mot...	243
858	s859	TV Show	Better Call Saul	NaN	Bob Odenkirk, Jonathan Banks, Michael McKean, ...	United States	February 9, 2020	2018	TV-MA	4 Seasons	Crime TV Shows, TV Comedies, TV Dramas	This Emmy-nominated prequel to "Breaking Bad" ...	243
1089	s1090	TV Show	Breaking Bad	NaN	Bryan Cranston, Aaron Paul, Anna Gunn, Dean No...	United States	August 2, 2013	2013	TV-MA	5 Seasons	Crime TV Shows, TV Dramas, TV Thrillers	A high school chemistry teacher dying of cance...	243
1584	s1585	TV Show	Damnation	NaN	Logan Marshall-Green, Killian Scott, Sarah Jon...	United States	November 7, 2018	2017	TV-MA	1 Season	Crime TV Shows, TV Dramas	During the Great Depression, a stranger with a...	243
1608	s1609	TV Show	Dare Me	NaN	Willa Fitzgerald, Herizen Guardiola, Marlo Kel...	United States	December 30, 2020	2019	TV-MA	1 Season	Crime TV Shows, TV Dramas, TV Thrillers	Relationships topple and loyalties flip when a...	243

Challenges

- The first challenge we face is in the case of null values. It is not easy to decide what to do with the column director, either we remove it or keep it.
- How to perform clustering on the text-based features. It is not easy to decide which column we take or leave in clustering.
- Finding the optimal value of k is not an easy task. Here, we perform silhouette score and elbow method on different sets of values of k.

Conclusion

- First, we run Data Wrangling on our model to ensure that there are no duplicate entries in our dataset. After checking the duplicates in our dataset we perform analysis for null values in our dataset. Here, we found more than 30% null values in the director's column. Then, we take appropriate action for null values according to the circumstances. We remove null values of the added_date columns because there is no logical way to deal with the null values of the date column.
- In the second step, we perform EDA and Data Visualization on our dataset. Here, we found that the proportion of tv shows in Netflix content is very less as compared to the movies. We can observe that the majority of Netflix material is intended for adults. There is very little content available for teens and kids.

Conclusion:

- The number of movies on Netflix is growing significantly faster than the number of TV shows. Because of covid-19, there is a significant drop in the number of movies and television episodes produced after 2019. Because of covid-19, there is a significant drop in the number of movies and television episodes produced after 2019.
- The United States is the most prolific generator of Netflix content, with India and the United Kingdom trailing far behind. The majority of the content on Netflix in India is comprised of movies. The fundamental reason for the variation in content must be due to market research undertaken by Netflix. It is also interesting to see parallels between culturally comparable nations - the US and UK are closely aligned with their Netflix target ages, but radically different from, for example, India or Japan!
- It is evident that international movies/ tv shows, tv dramas, and tv comedies are the top three genres with the most content on Netflix. It is interesting that International Movies tend to be Dramas.
- Here, we perform the K-Means clustering on our dataset. Here, we find the optimal value of k is 20. But, if we want to recommend some movies and tv shows then k=20 is not good so in such a case, we take the value of k as 600.
- The silhouette score for k=20 is 0.886575253337518 which is a very good score.
- We also perform the K-means clustering using the TF-IDF. In this case, we get the optimal value of k is 800. And the silhouette score for k=800 is 0.034.