# **NETFLIX MOVIES AND TV SHOWS CLUSTERING**

# **SOUMYADEEP PANDIT**

panditdeep9@gmail.com

**Data science trainee,**

Alma Better, Bangalore.

**Abstract:**

Netflix is one of the largest online streaming media providers. It began its operations in 1997. Founded by two tech entrepreneur Reed Hastings and Marc Randolph. The Company’s head office is in Los Gatos, California. Netflix’s initially started selling DVDs or provide them on a rental basis. Over the period with growth of internet users and the decline of DVD sales and rental services, it changed its business model to video on demand. From 2012 onwards, it started producing its original TV-series and movies.

Flixable is essentially a Netflix movie search engine. It makes use of a lot of the same data as Netflix's interface, allowing you to search for movies by title, genre, or person affiliated with them. Flixable's UI, on the other hand, is much easier to use. Flixable also includes some other sorting algorithms. You can arrange your results alphabetically and by year released in addition to filtering by genre and IMDB rating.

We have used 2 methods for K-means clustering, one by the normal clustering, second by the TF-IDF vectorization method, where we had used the silhouette score and elbow method to predict the optimal number of clusters i.e. K.

***Keywords: Customer segmentation, clustering, recency, frequency, monetary value.***

**1.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, we are required to do

• Exploratory Data Analysis

• Understanding what type content is available in different countries

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

• Clustering similar content by matching text-based features

**1.1. Data Set 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 Releaseyear of the movie / show

rating : TV Rating of the movie / show

duration : Total Duration - in minutes or number of seasons

listed\_in : Genere

description: The Summary description

**2. Introduction:**

Netflix is a media streaming service established in the United States. It offers movie streaming as a subscription service. Along with movies, it contains television shows and in-house generated material. Initially, Netflix sold DVDs and operated as a mail-order rental service. A year later, they stopped selling DVDs but kept their DVD rental service going. They went online in 2010 and launched a streaming service. Since then, Netflix has evolved into one of the best and most popular streaming services on the planet (Netflix,2020).

Netflix has experienced tremendous growth over the last decade, both in terms of customer numbers and income earned. Time and time again, the corporation has proven its capacity to outperform even the most ambitious sales and subscriber expectations.

Netflix is worth more than $150 billion and has over 167 million paid subscribers worldwide as of the end of 2019. Netflix is available for viewers to enjoy in 110 countries, with international subscribers fueling most of its impressive rise over the last several years.

But the question is, given all of this impressive 'bottom line' results, what is actually on Netflix?

We all know about Netflix, the worlds largest on-demand internet streaming media and online DVD movie rental service provider.

• Founded August 29, 1997 in Los Gatos, California by Marc and Reed

• It has 69 million members in over 60 countries enjoying more than 100 million hours of TV shows and movies per day.

• Most success with : House of Cards show

A "Netflix Original" is content created, co-produced, or distributed solely through Netflix's services. When Netflix signs a project, it funds it differently than other TV networks, offering the money up front and instantly ordering two seasons of most programmes.

"Bird Box," starring Sandra Bullock in 2018, and "I Care a Lot," starring Rosamund Pike in 2020, are two Netflix originals.

Netflix also put money into exclusive stand-up comedy specials from comedians like Dave Chappelle, Louis C.K., Chris Rock, Jim Gaffigan, Bill Burr, and Jerry Seinfeld, among others.

Internally, the corporation began generating its own unique material, such as The Ranch and Chelsea.

Several studios have exclusive pay TV arrangements with Netflix. The agreements provide Netflix exclusive streaming rights while maintaining standard pay TV rates. Netflix's US library includes newer releases from Relativity Media and its former subsidiary Rogue Pictures, as well as DreamWorks Animation (until May 2018, when the studio signed a new contract with Hulu), Open Road Films, Universal Animation (for animated films declined by HBO), FilmDistrict, and The Weinstein Company, which caused Netflix to pull out of hosting the 75th Golden Globe Awards.

Netflix has always been a data-driven organisation. Our analytic work provides relevant measurements, insights, projections, and analytic tools to decision-makers across the firm, allowing everyone to excel in their role. We undertake context-rich analysis in collaboration with business teams across product, content, studio, marketing, and business operations to provide insight into every area of our business, our partners, and, of course, our members' Netflix experience.

## **3. Methodology:**

Although the data used must be kept private, it is critical to compare the results with other conventional machine learning algorithms to demonstrate the significance of the positives and negatives of each method considered in this study. This section explains the algorithms used in this study in detail. We’ve done preprocessing, handled null values, created required feature, done Exploratory data analysis, done some feature engineering, created and created clusters.

1. Data Wrangling.
2. Exploratory data analysis.
3. Data Pre-processing
4. Feature engineering.
5. KMeans Clustering.

**1) Data Wrangling :**

The process of cleansing and integrating chaotic and complicated data sets for easy access and analysis is known as data wrangling.

With the amount of data and data sources continually rising and expanding, it is becoming increasingly important to organise vast amounts of data for analysis. This procedure usually entails manually converting and mapping data from one raw format to another in order to facilitate data consumption and organisation.

The Goals of Data Wrangling

• Reveal a "deeper intelligence" by gathering data from multiple source

• Provide accurate, actionable data in the hands of business analysts in a timely matter

• Reduce the time spent collecting and organizing unruly data before it can be utilized

• Enable data scientists and analysts to focus on the analysis of data, rather than the wrangling

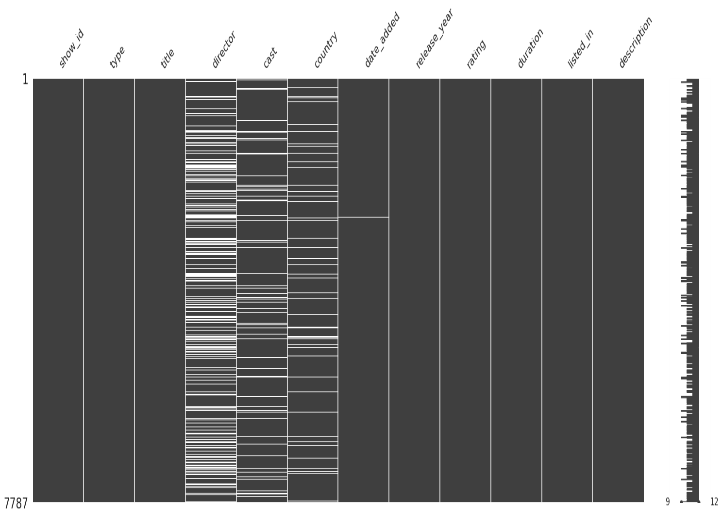
• Drive better decision-making skills by senior leaders in an organization

While wrangling this dataset, what we found out is:

•The dataset contained 7787 observations and 12 columns given.

• We had major amount of missing values in the ‘director’ column which we had filled with a random text.

• Our dataset had 0 duplicate values.



**2) Exploratory data analysis :**

Exploratory Data Analysis (EDA) is an approach to analyzing datasets to summarize their main characteristics, often with visual methods. EDA is used for seeing what the data can tell us before the modeling task. It is not easy to look at a column of numbers or a whole spreadsheet and determine important characteristics of the data. It may be tedious, boring, and/or overwhelming to derive insights by looking at plain numbers. Exploratory data analysis techniques have been devised as an aid in this situation.

Exploratory data analysis is generally cross-classified in two ways. First, each method is either non-graphical or graphical. And second, each method is either univariate or multivariate (usually just bivariate).

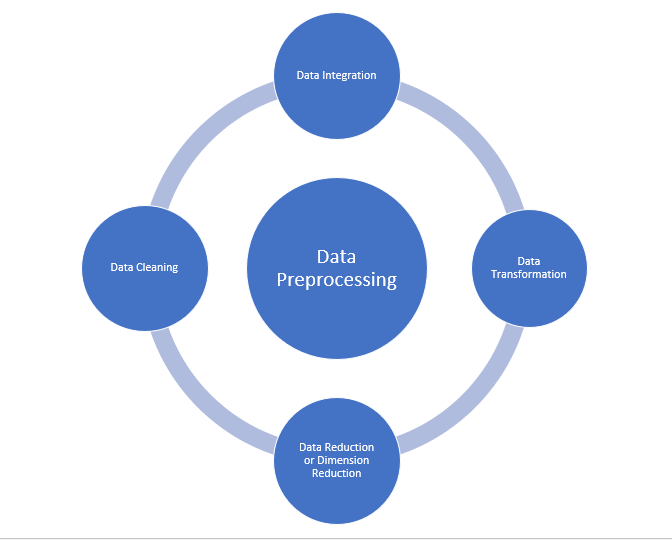
**3) Feature engineering :**

Feature engineering refers to a process of selecting and transforming variables when creating a predictive model using machine learning or statistical modeling (such as deep learning, decision trees, or regression). The process involves a combination of data analysis, applying rules of thumb, and judgement.

**4) Data Pre-processing:**

Data preprocessing is the process of transforming raw data into an understandable format. It is also an important step in data mining as we cannot work with raw data. The quality of the data should be checked before applying machine learning or data mining algorithms.

Preprocessing of data is mainly to check the data quality. The quality can be checked by the following

* **Accuracy**: To check whether the data entered is correct or not.
* **Completeness**: To check whether the data is available or not recorded.
* **Consistency:** To check whether the same data is kept in all the places that do or do not match.
* **Timeliness**: The data should be updated correctly.
* **Believability**: The data should be trustable.
* **Interpretability**: The understandability of the data.
* 

**5) KMeans Clustering :**

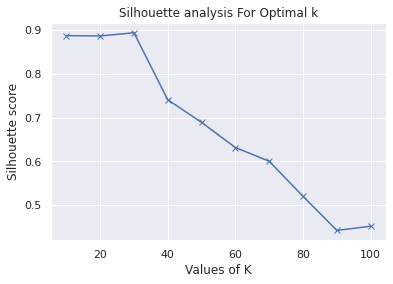
Clustering is defined as the process which divides the whole data into groups or clusters supporting the patterns within the data. To process the training data, the K-means algorithm in data processing starts with a primary group of randomly selected centroids, which are used because the beginning points for each cluster, then performs iterative calculations to optimize the positions of the centroids.

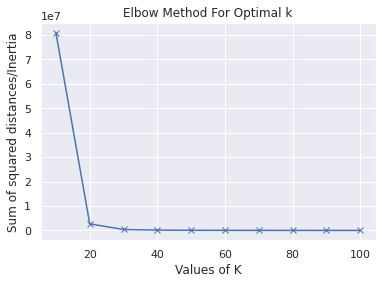
* Implementation of KMeans clustering :

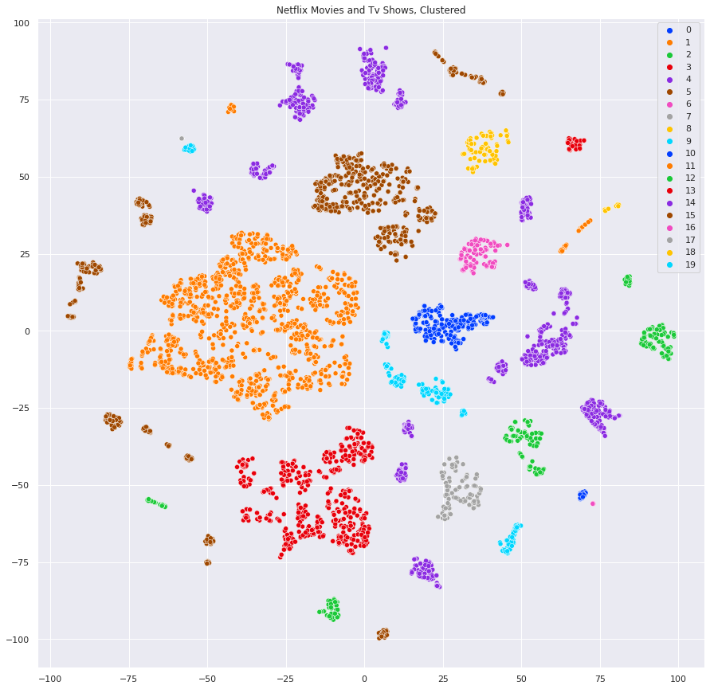
In this project, for creation of Netflix clustering using K-Means algorithm it is essential to decide the number of clusters to form i.e. the value of K. For deciding the value of k Elbow Technique is used.

Elbow technique: The elbow method runs k-means clustering on the dataset for a range of values for k (say from 1-10) and then for each value of k computes an average score for all clusters. By default, the distortion score is computed, the sum of square distances from each point to its assigned center and picking the elbow of the curve as the number of clusters to use.

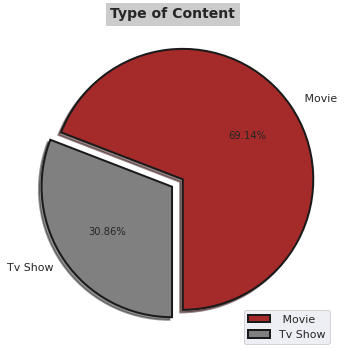
We can observe, as the number of clusters increases the sum of square distances are becoming lesser. And will take the count of clusters where this elbow is bending. In our case, the sum of square distance is dramatically decreasing at K = 3, so this is the optimal value to choose for no of clusters.



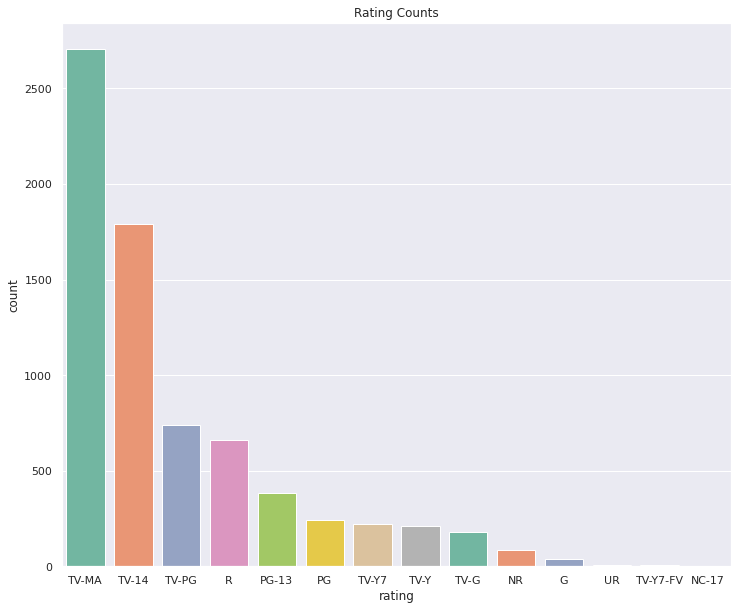


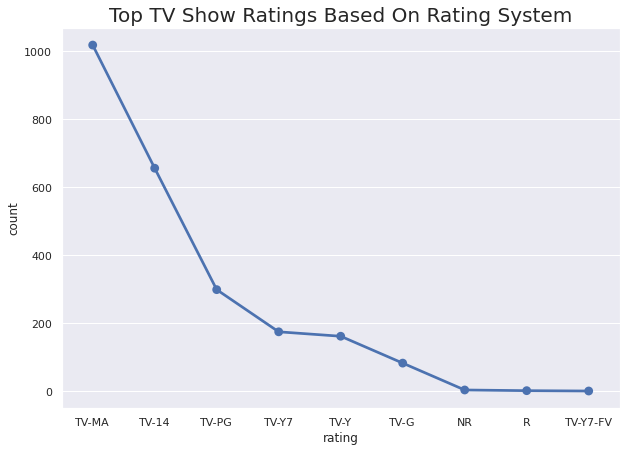
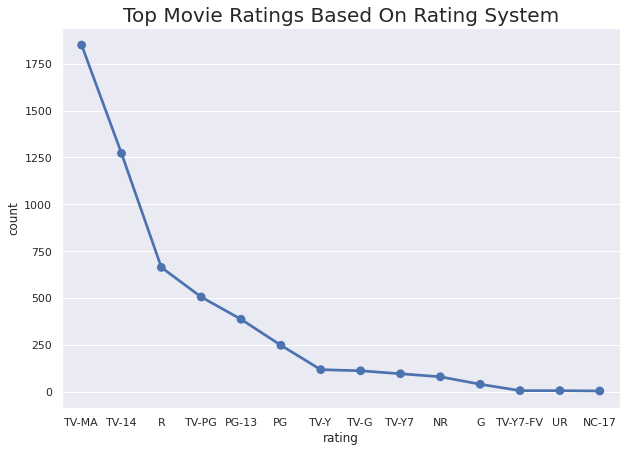


**4. Trends:**

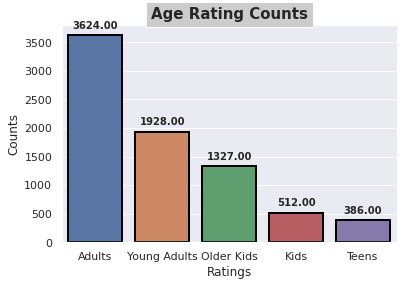


* It is evident that there are more movies on Netflix than TV shows.
* Netflix has 5377 movies, which is more than double the quantity of TV shows.

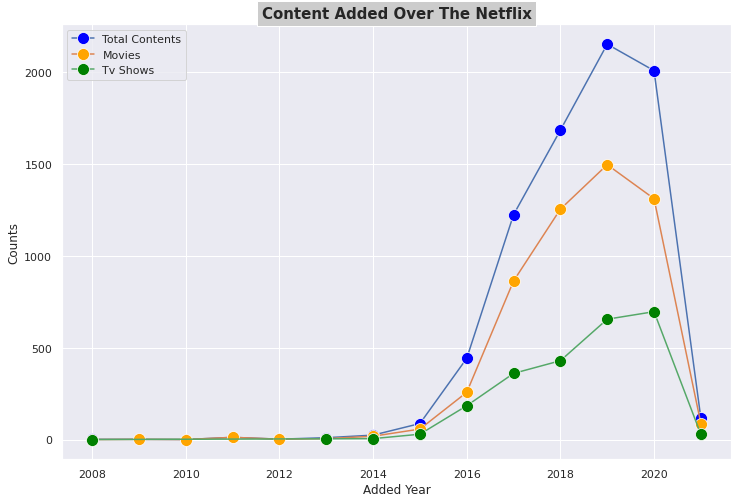


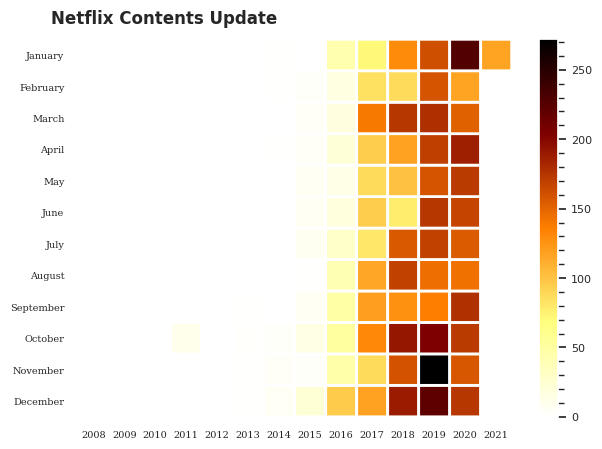


* The 'TV-MA' rating is used in the majority of the film. The TV Parental Guidelines provide a "TV-MA" classification to a television programme that is intended solely for mature audiences.
* The second largest is 'TV-14,' which stands for content that may be inappropriate for minors under the age of 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."



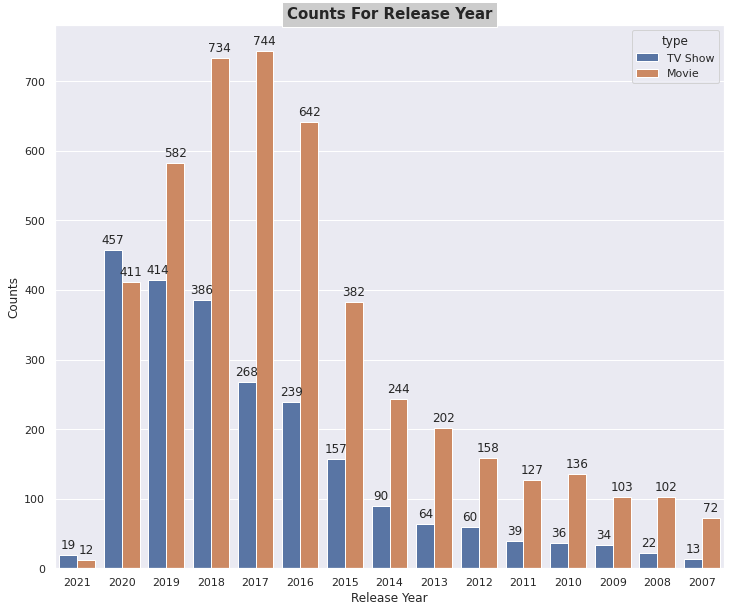
* We can observe from the above count plot that the majority of Netflix material is intended for adults. There is very little content available for teens and kids.



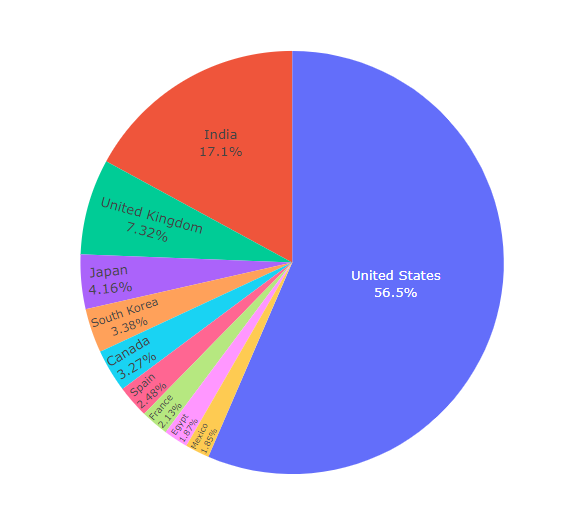


* The number of movies on Netflix is growing significantly faster than the number of TV shows.
* In both 2018 and 2019, approximately 1200 new movies were added.
* We saw a huge increase in the number of movies and television episodes after 2014.
* Because of covid-19, there is a significant drop in the number of movies and television episodes produced after 2019.
* It appears that Netflix has focused more attention on increasing Movie content that TV Shows. Movies have increased much more dramatically than TV shows.
* 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.

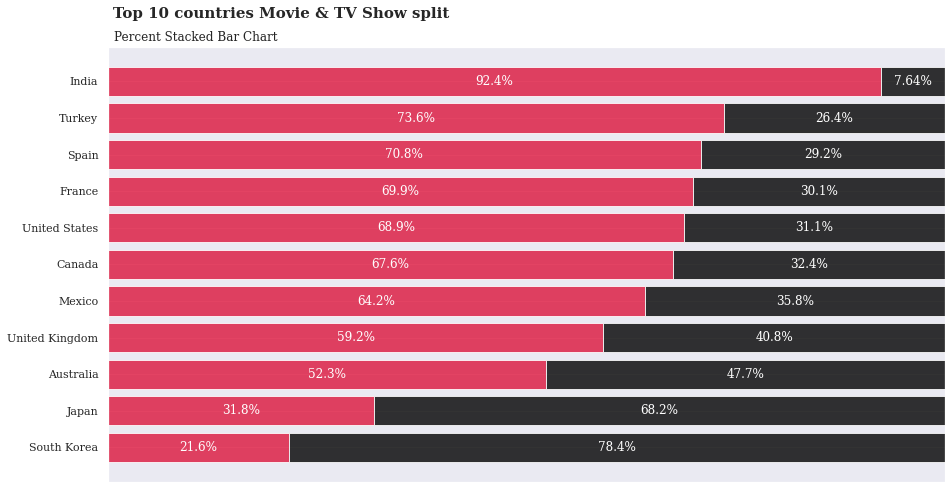
Here we can see people are buying on afternoon time period more.



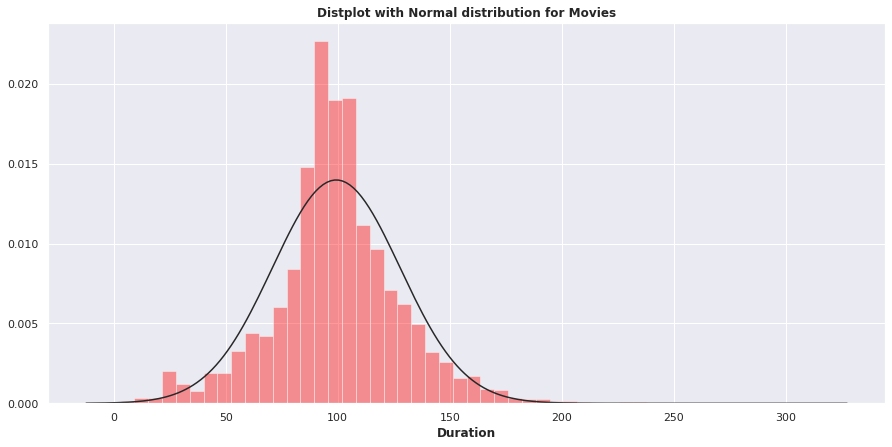
* According to the following countplot, 774 movies were released in 2017, which is the most number of releases in any year.
* According to the above countplot, there will be 457 television shows released in 2020, which is the most of any year.



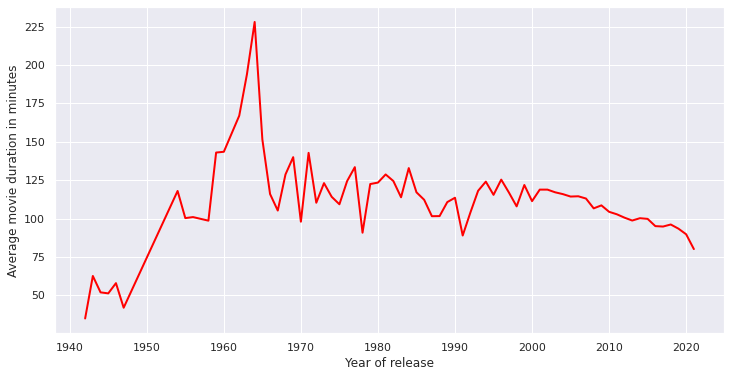
* The United States is the most prolific generator of Netflix content, with India and the United Kingdom trailing far behind.



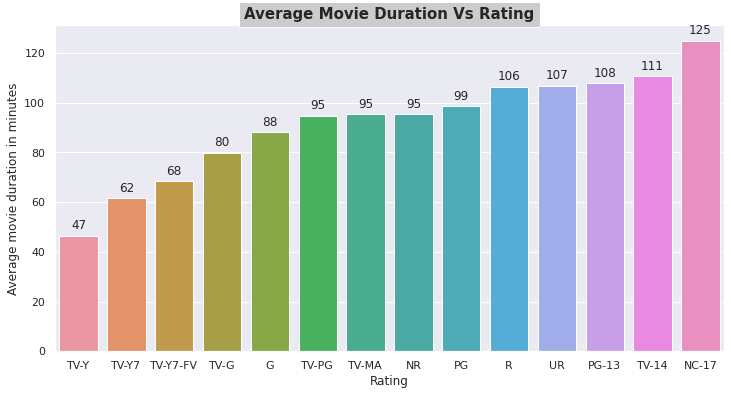
* The majority of the content on Netflix in India is comprised of movies.
* Bollywood is a significant business, and movies, rather than TV shows, may be the industry's major focus.
* South Korean Netflix on the other hand is almost entirely TV Shows.
* The fundamental reason for the variation in content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



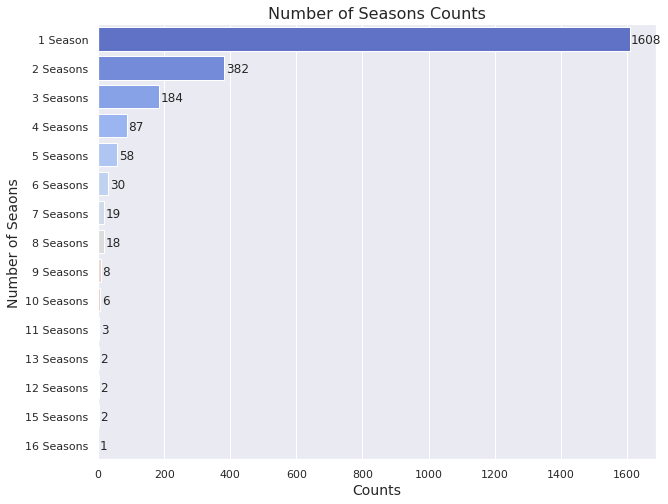
* The majority of the films are between 85 and 120 minutes long, which is appropriate.



* Movies made before 1948 have a short average duration, compared to those released after 1948.
* The average movie's 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.



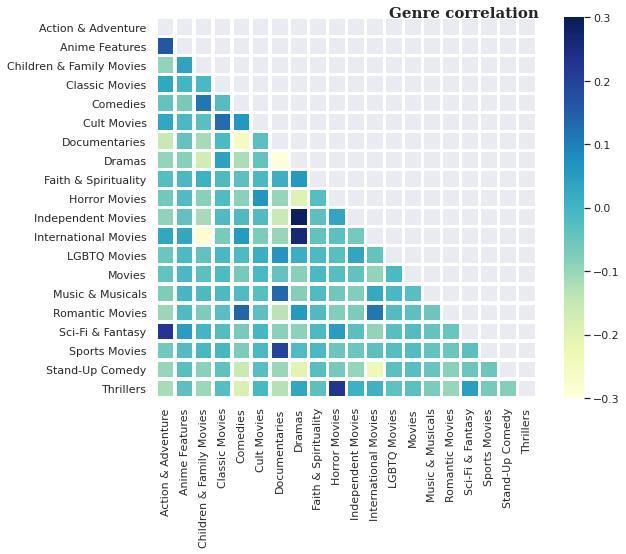
* According to the above counterplot, each of the 1608 television shows had only one season.
* There were extremely few television shows that had more than six seasons.



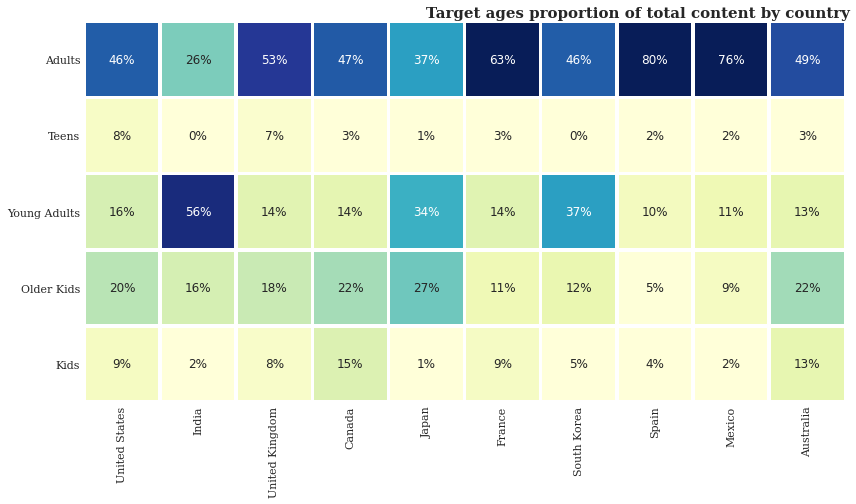
* As a result, it is evident that international movies, dramas, and comedies are the top three genres with the most content on Netflix.



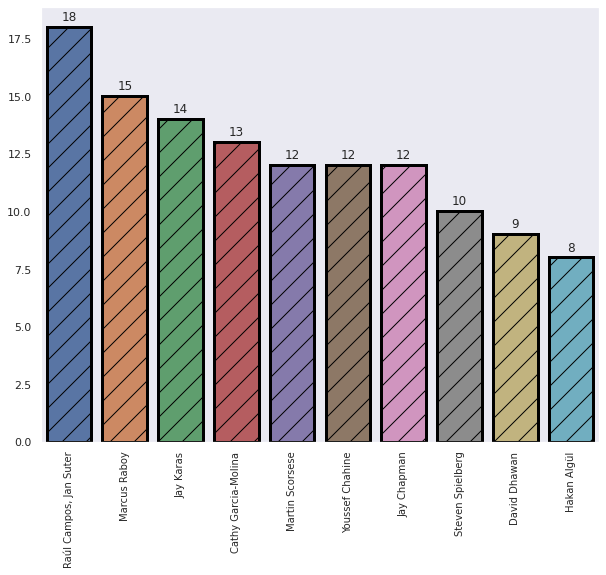
* As a result, it is evident that international tv shows, tv dramas, and tv comedies are the top three genres with the most content on Netflix.



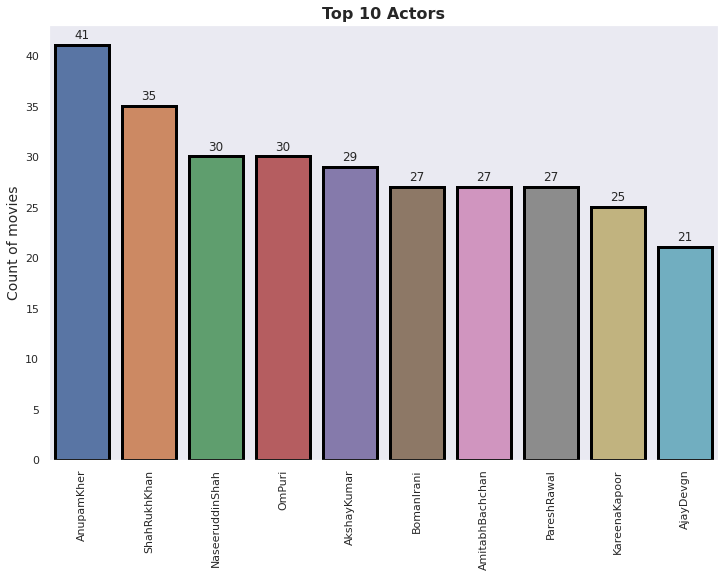
* It is interesting that International Movies tend to be Dramas.
* Another observation is that International Movies are rarely in the Children's genres.



* 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.



* The Directors who produce the most material are Raul Campos and Jan Sutler. They work in 18 movies as a director.
* David Dhawan is ranked second among top directors, having directed 15 films.



* According to the above barplot, Anupam Kher has worked in over 40 films.
* After Anupam Kher, Shahrukh Khan is ranked second, with 35 films under his belt.
* Naseeruddin Shah and Ompuri have worked in 30 films.

1. **Steps involved:**

* **Null values Treatment**

We’ve checked the null values in our dataset with isnull() function with the help of msno matrix, there was many null values present in our dataset. Null values have been removed as the null values were present in ‘director’ column and we have to do clustering as per director which is why we cannot risk to lose the column.

* **Exploratory Data Analysis**

I used this strategy after importing the dataset by comparing our different variable. This method assisted us in determining numerous characteristics and correlations between variables. It helped us understand which features behave in which ways in relation.

* **Feature Engineering**

We’ve created and selected important features which will help us in clustering.

* **Standardizaton of features**

Our main goal in this step was to scale our data into a standard format so that we could better use it for fitting and applying multiple algorithms.

The main purpose was to ensure that specific behavior or processes within the chosen environment were consistent or uniform.

We’ve used categorical encoding for the scaling purpose.

* **KMeans Clustering :**

K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. Data points are clustered based on feature similarity. With elbow method I came to know that optimal clusters is 20.

1. **Conclusion:**

So finally, we can conclude here!

* 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.
* 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.

**References-**

1. MachineLearningMastery
2. GeeksforGeeks
3. Analytics Vidhya
4. Towards DataScience
5. Data Science from scratch
6. Neptune
7. Builtin
8. Plosone
9. ijert