NETFLIX MOVIES & TV SHOWS CLUSTERING

Abstract

Netflix is a company that manages a large collection of TV shows and movies, streaming it anytime via online. This business is profitable because users make a monthly payment to access the platform. However, customers can cancel their subscriptions at any time. Therefore, the company must keep the users hooked on the platform and not lose their interest. This is where recommendation systems start to play an important role, providing valuable suggestions to users is essential.

INTRODUCTION

Netflix's recommendation system helps them increase their popularity among service providers as they help increase the number of items sold, offer a diverse selection of items, increase user satisfaction, as well as user loyalty to the company, and they are very helpful in getting a better understanding of what the user wants. Then it's easier to get the user to make better decisions from a wide variety of movie products. With over 139 million paid subscribers (total viewer pool -300 million) across190 countries, 15,400 titles across its regional libraries and 112 Emmy Award Nominations in 2018 — Netflix is the world's leading Internet television network and the most-valued largest streaming service in the world. The amazing digital success story of Netflix is incomplete without the mention of its

Problem Statement

This dataset consists of TV shows and movies available on Netflix as of 2019. The dataset is collected from Flexible 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.

Variables

Attribute Information:

The dataset provided contains 7787 rows and 12 columns.

The following are the columns in the dataset:

- **Show id:** Unique identifier of the record in the dataset
- **Type**: Whether it is a TV show or movie
- **Title:** Title of the show or movie
- **Director:** Director of the TV show or movie
- **Cast:** The cast of the movie or TV show
- **Country:** The list of the country in which a show/ movie is released or watched
- **Date added:** The date on which the content was on boarded on the Netflix platform
- **Release year:** Year of the release of the show/ movie
- **Rating:** The rating informs about the suitability of the content for a specific age group
- **Duration:** Duration is specified in terms of minutes for movies and in terms of the number of seasons in the case of TV shows
- **Listed in:** This columns species the category/ genre of the content
- **Description:** A short summary about the storyline of the content.

Objective

Netflix Recommender recommends Netflix movies and TV shows based on a user's favorite movie or TV show. It uses a Natural Language Processing (NLP) model and a K-Means Clustering model to make these recommendations. These models use information about movies and TV shows such as their plot descriptions and genres to make suggestions. The motivation behind this project is to develop a deeper understanding of recommender systems and create a model that can perform Clustering on comparable material by matching text-based attributes. Specifically, thinking about how Netflix create algorithms to tailor content based on user interests and behavior.

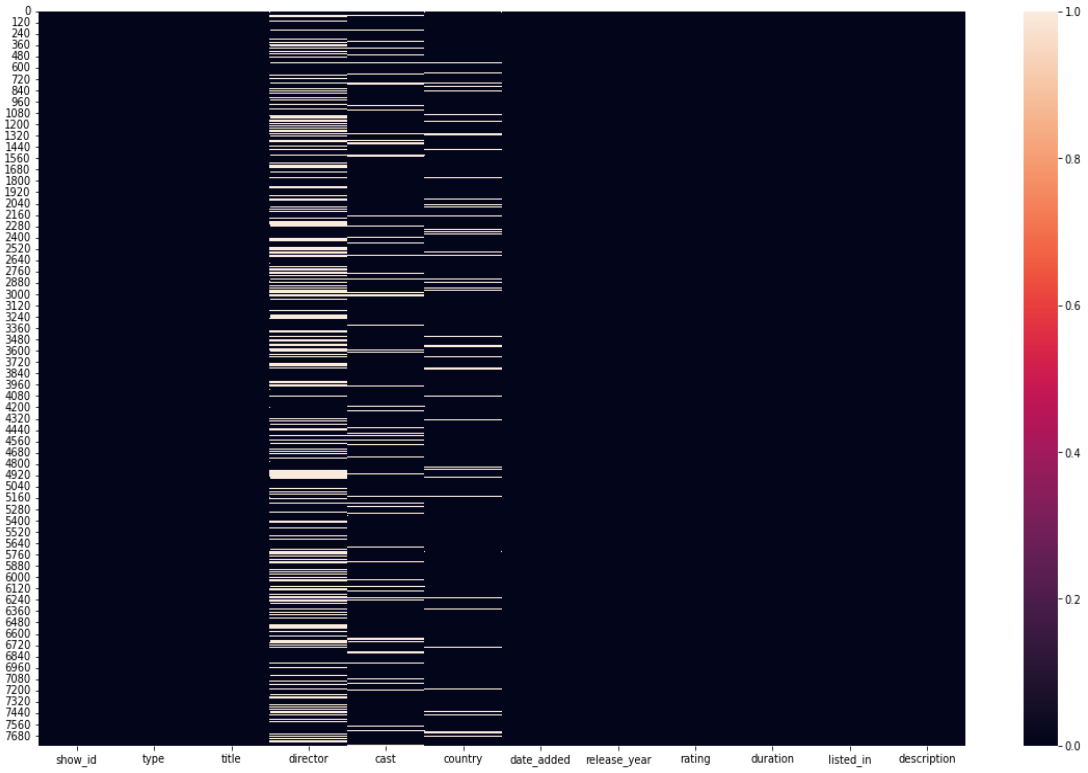
Steps involved:

- **Exploratory Data Analysis**:
  The first step of our project is performing the EDA process on the dataset so that we can get the ideaabout the dataset i.e. the number of variables, the data type of the variables , visualize the dataset for Better understanding and decide the suitable methods and algorithms that might produce desiredoutcomes

- **Data Preprocessing:**
  In EDA process we find the type of dataset and decide the approach, in this project the preprocessingsteps would removing the punctuations, stop words , generate count vectorizer and document term matrix which would help in building up the model.

- **Handling missing values:**
  We checked for null values after loading the dataset and removed the null values, as well as some Unnecessary columns.

  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.



NULL VALUE TREATMENT

1. **RATING & COUNTRY** - Replacing nulls with mode
2. **CAST**- Replacing nulls with 'unknown'.
3. **DATE** - there are few missing values for date column. so, lets drop missing value rows.
4. **DIRECTOR** - Director column has more then 30% null values so we will not use it for our model but will keep it for EDA - Replacing nulls with 'unknown

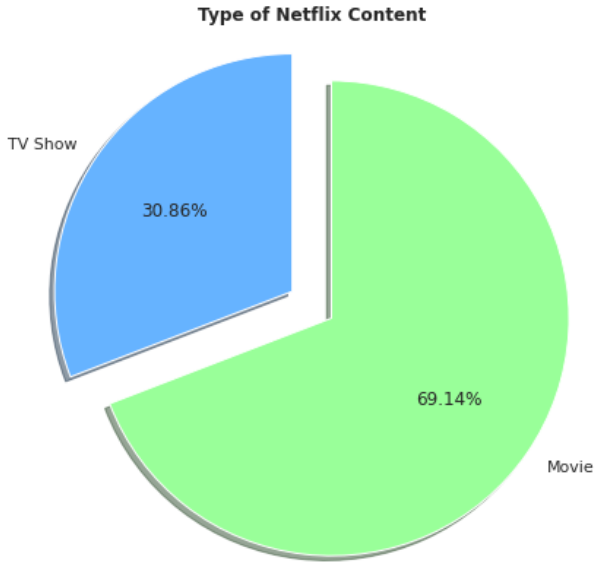
. There are very few null entries in the date_added fields thus we delete them.

Duplicate Values Treatment:

Duplicate values dose not contribute anything to accuracy of results. Our dataset dose not contains any duplicate values.

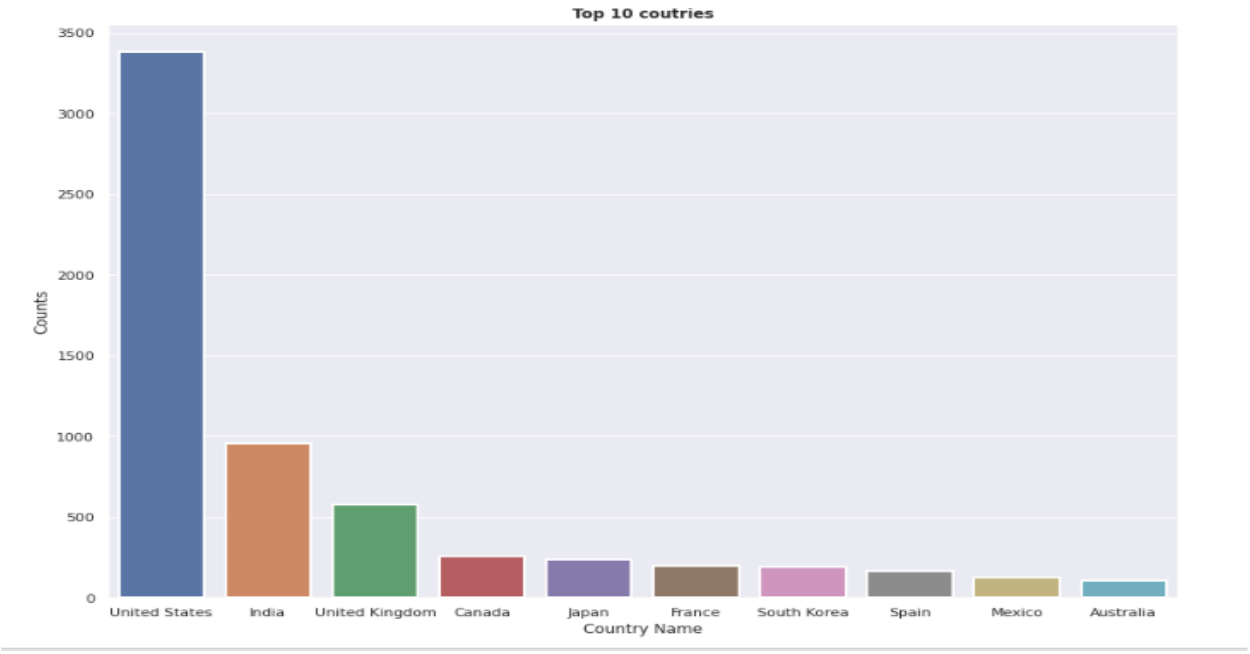
EXPLORATORY DATA ANALYSIS

Type of content



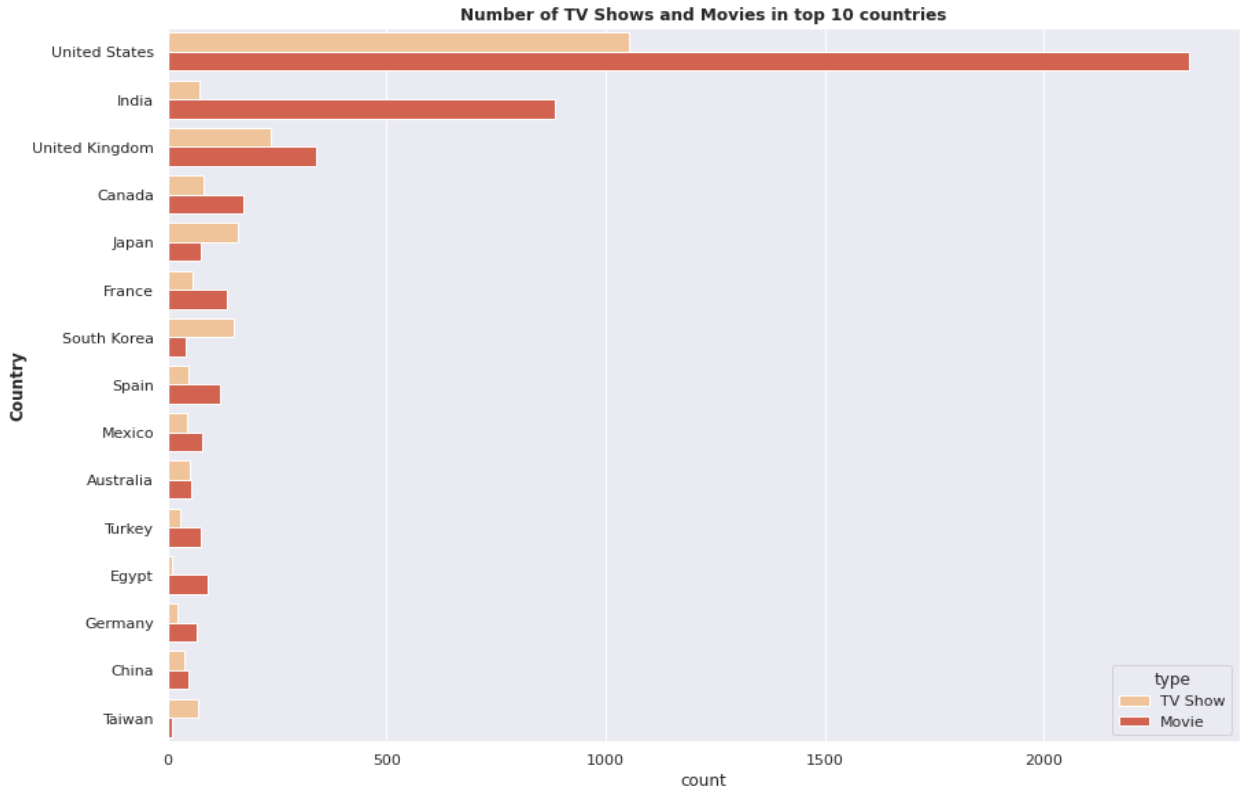
- **There are about 70% movies and 30% TV shows on Netflix.**

Top 10 Countries



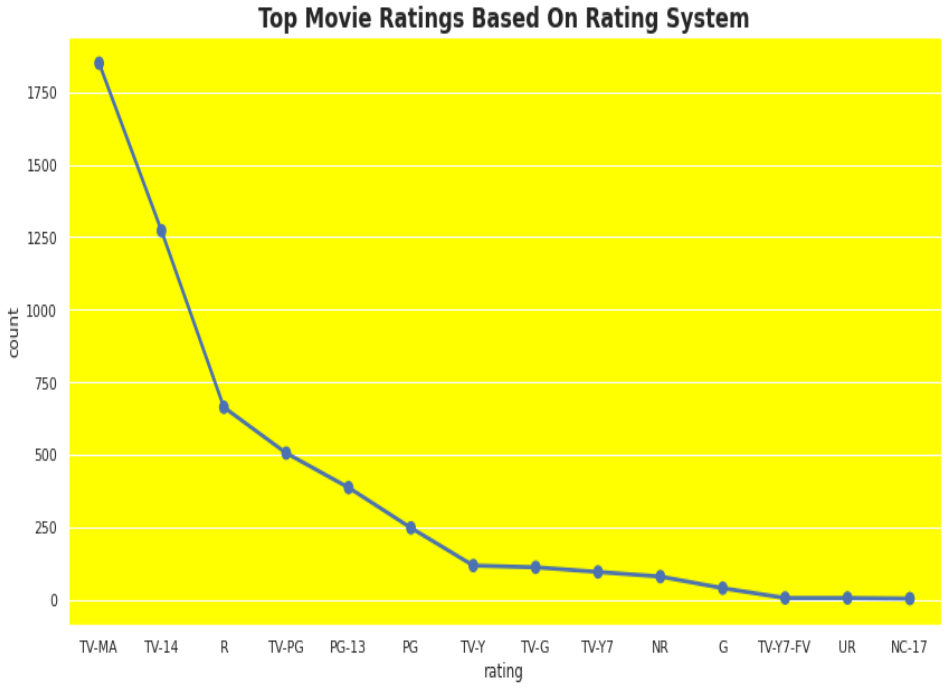
- **The United States has the most number of content on Netflix by a huge margin followed by India.**

Number Of Tv Shows And Movies In Top 10 Countries

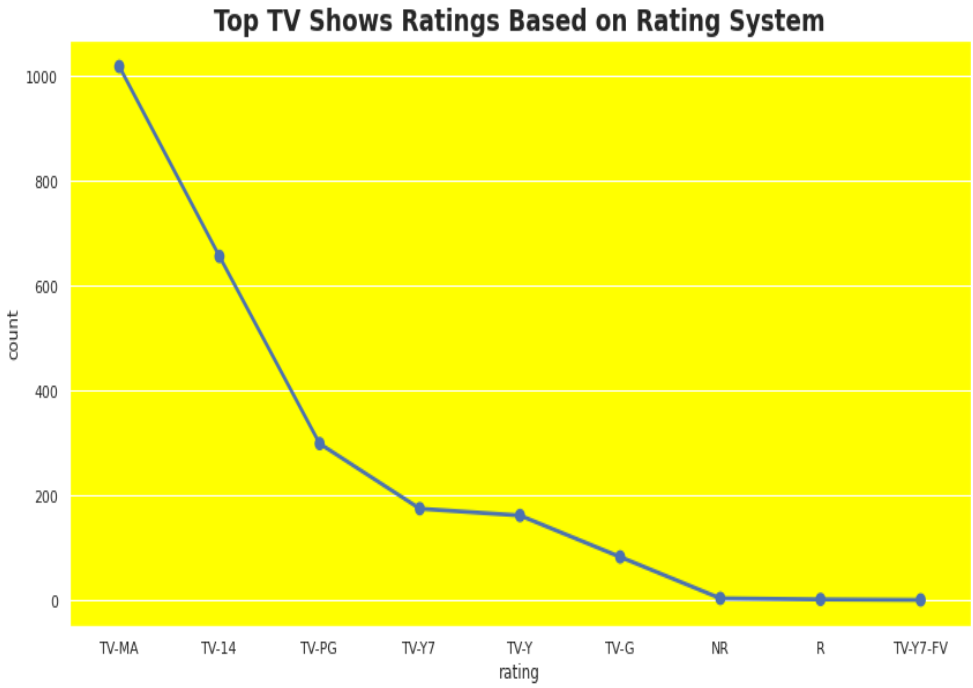


- **Most of the countries have more movies than TV shows but for South Korea and Japan it's the opposite. It maybe because K-Dramas and Anime are more popular in these two countries respectively.**

Top Movies & TV Shows Rating Based

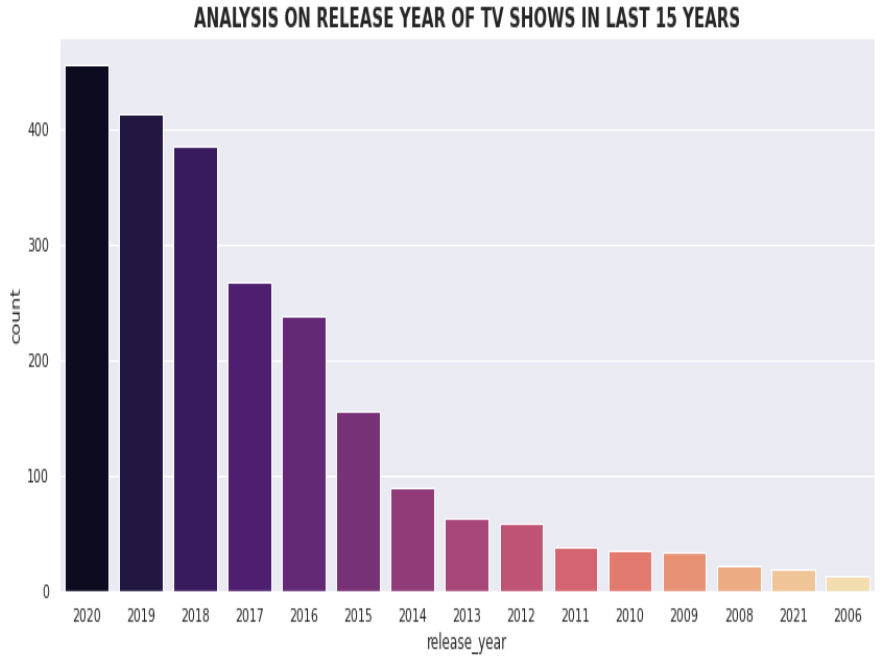


- **Most number of movies rated TV-MA i.e. Adult Rating.**

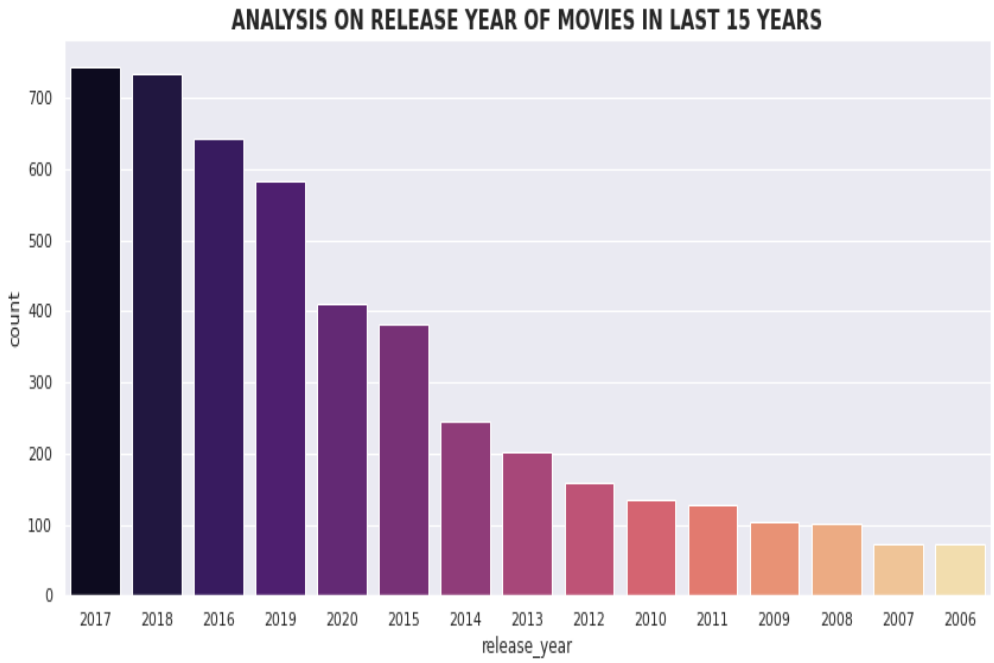


- **Most number of TV Shows rated TV- MA i.e. Adult Rating.**

Number of TV Shows & Movies Release in last 15 Year

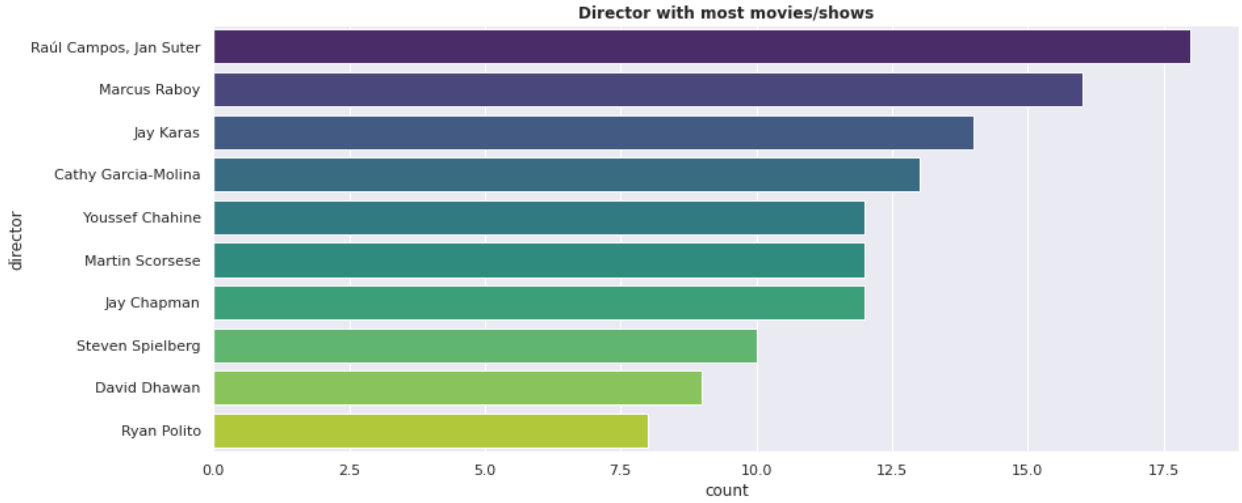


- **Year 2020 has the highest number of TV Shows release**



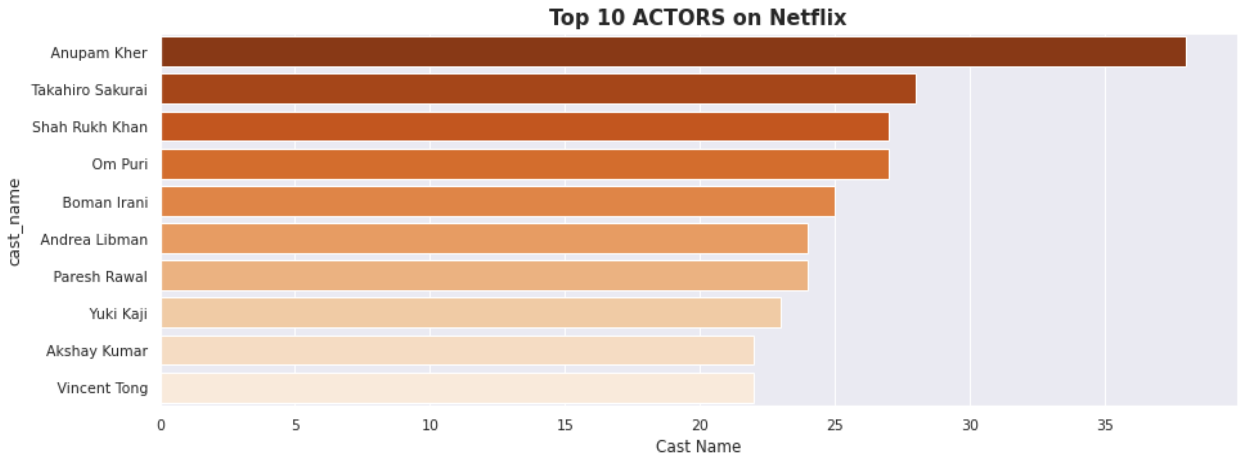
- **Year 2017 has the highest movie released.**

Director's With Most Movies & Tv Shows



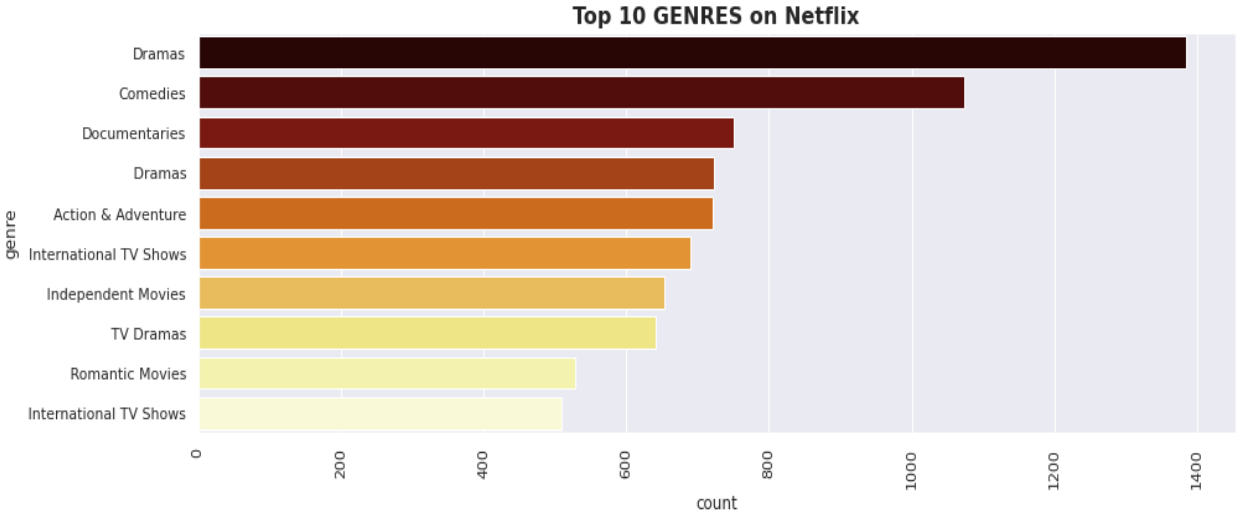
- **Raul Campos and Jan Sulter collectively have the most content on Netflix**
- **Followed by Marcus Rayboy, Jay karas and Others.**

Top 10 Actor's On Netflix



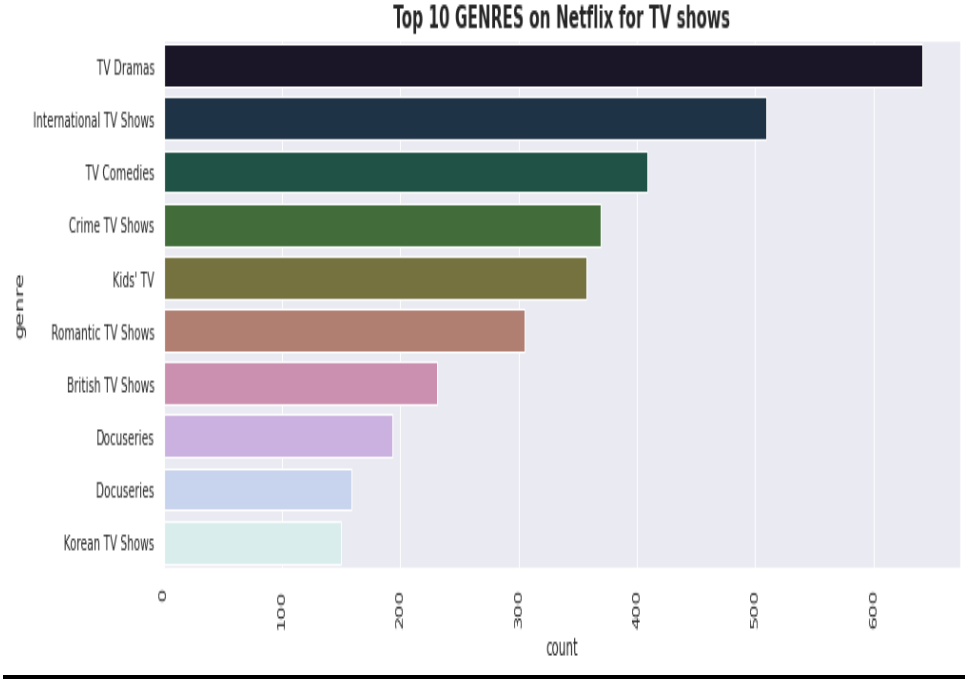
- **Anupam Kher Have the most number of films on Netflix.**
- **Followed by Takahiro Sakurai, Shahrukh Khan, Om Puri & Others.**

Top 10 Genre's On Netflix



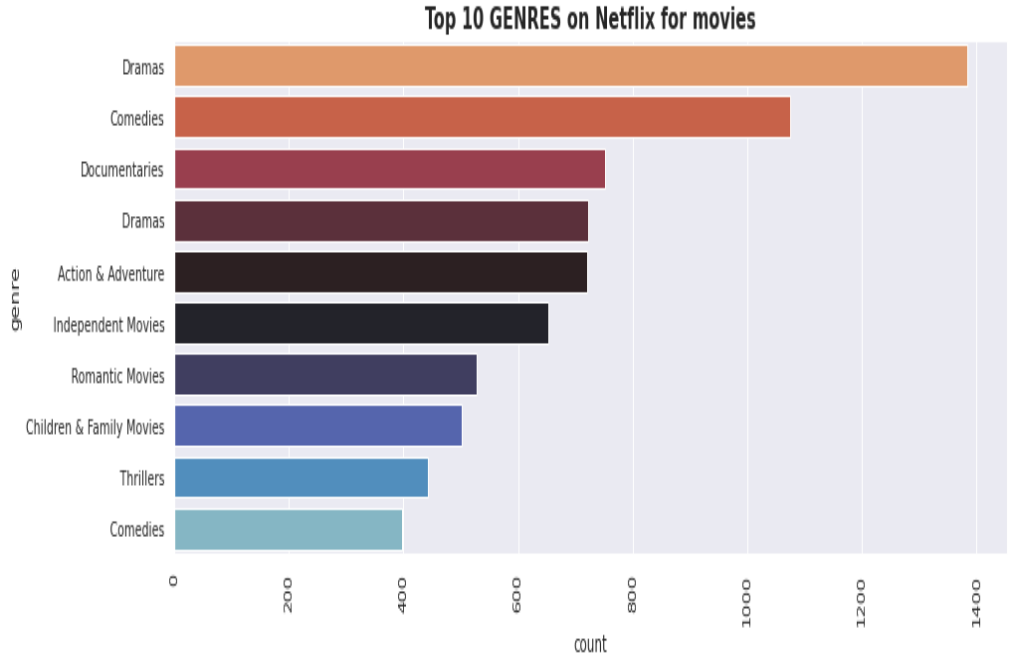
- **Drama is the most popular genre followed by comedy.**
- **Romantic Movies & International Tv shows have least popularity.**

Top 10 Genres For Tv Shows & Movies



TV Shows

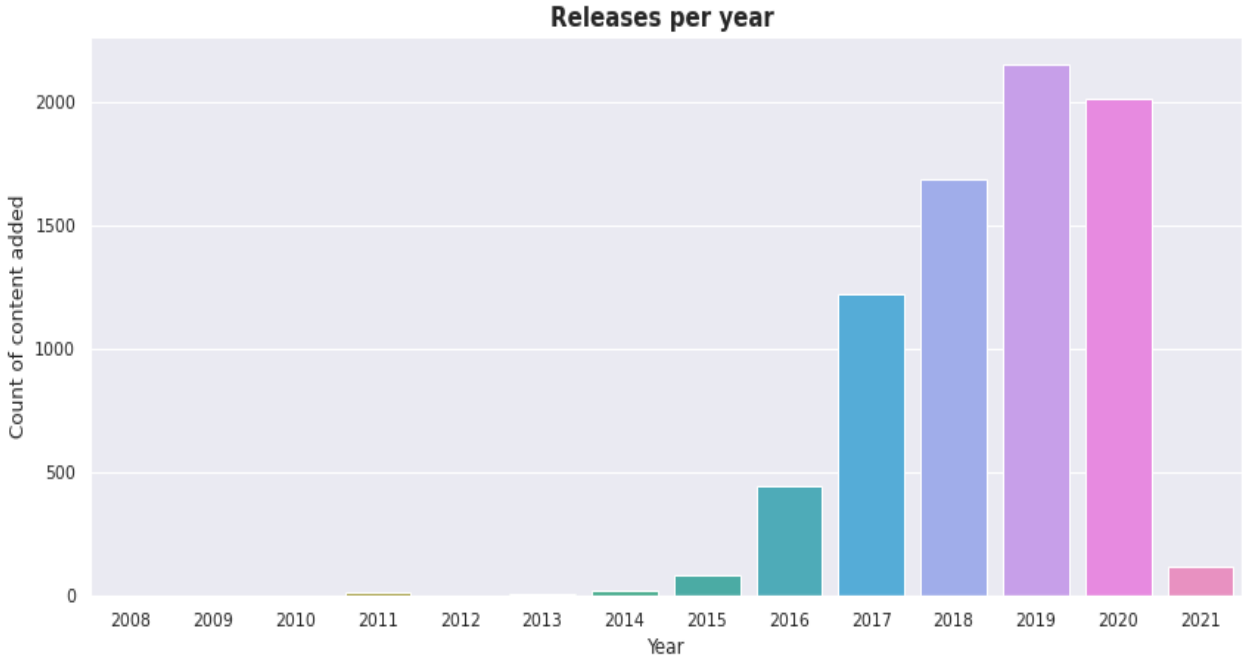
- **Drama is the most popular genre followed by International TV shows for movies.**



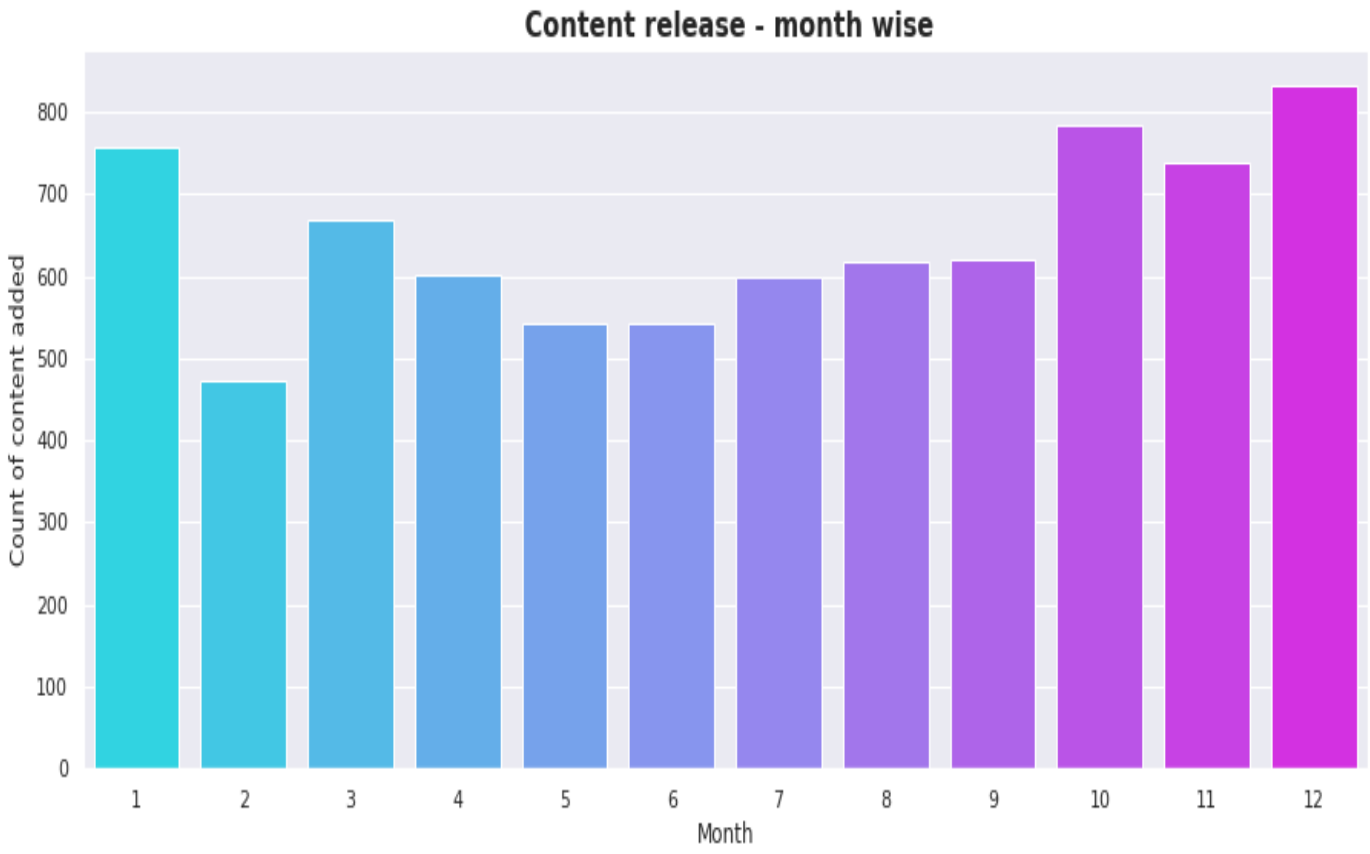
Movies

- **Drama is the most popular genre followed by comedy for movies.**

Year & Month wise Analysis

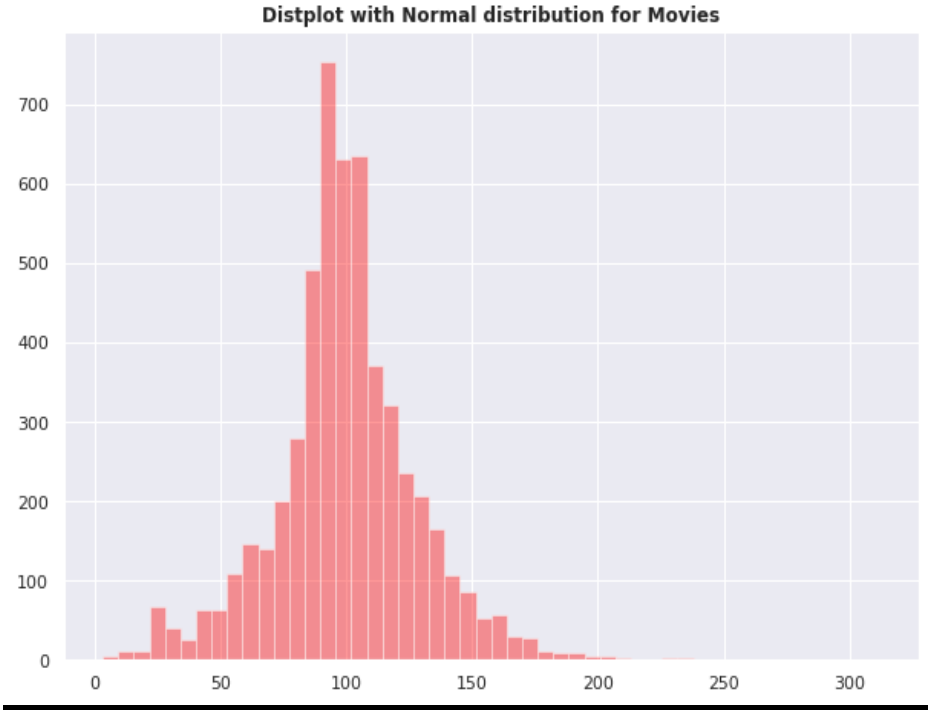


- **The number of release have significantly increased after 2015 and have dropped in 2021 because of Covid 19.**

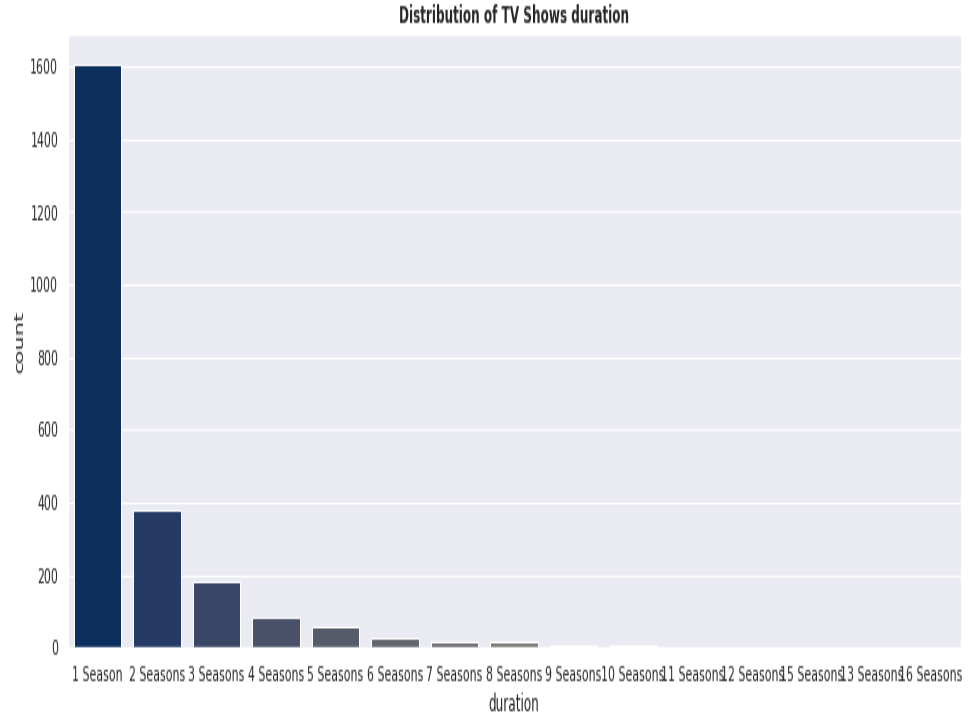


- **More of the content is released in holiday season - October, November, December and January.**

Distribution of Movies & TV Shows Duration



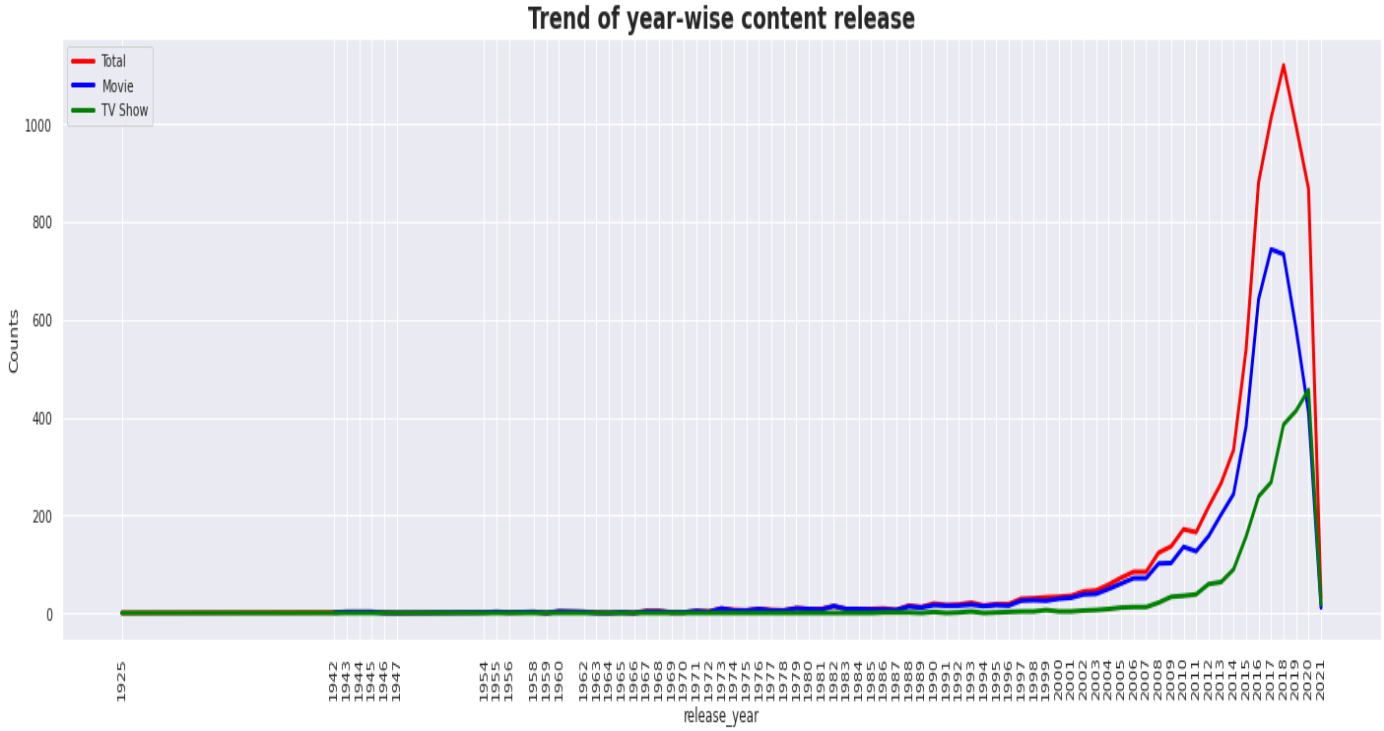
- **Mainly the movie duration is in b/w 55 to 150 minutes.**

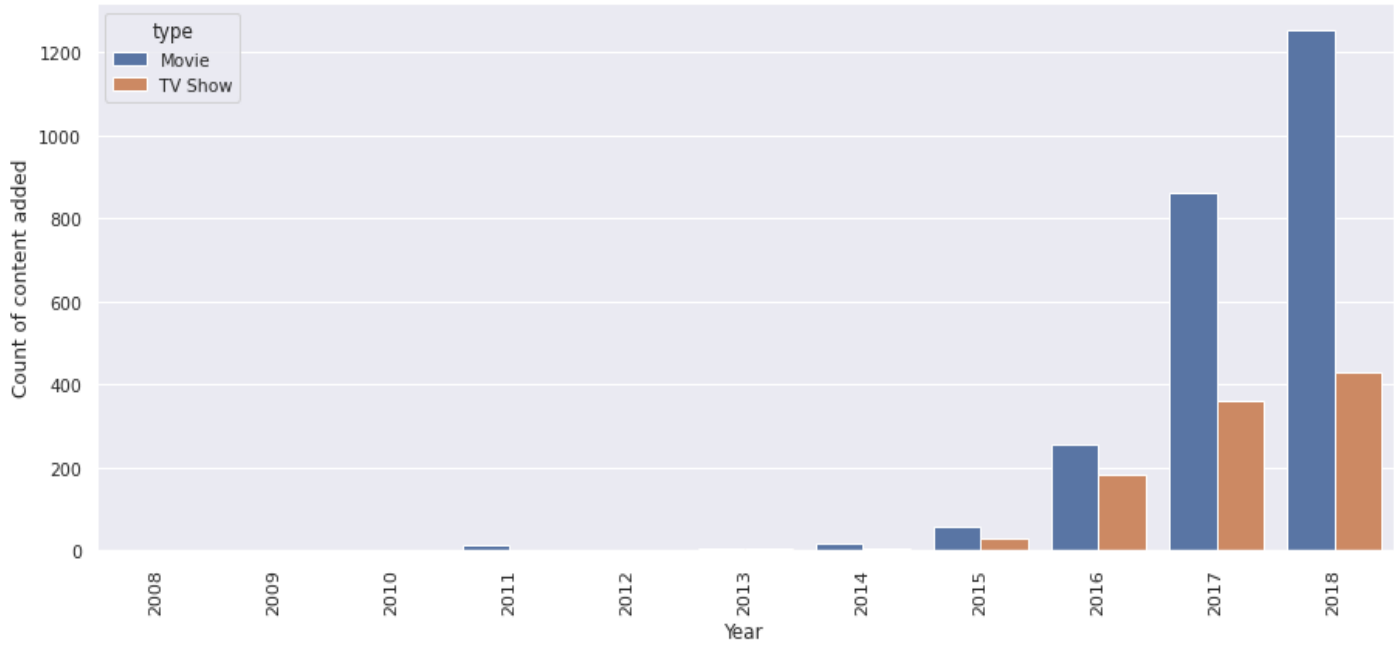


- **Mostly every TV Shows has atleast 3 seasons.**

HYPOTHESIS TESTING

Year Wise Trend





- **RESULT: Irrespective of the release years, there is no decline in the number of movies. Also number if movies added has always been more than the number of TV shows added. So with this information, we hereby reject our Hypothesis.**

Data Preprocessing

- We have made some changes in data just for EDA so we will start our clustering analysis with fresh data and do the manipulations again.
- Since number of unique actors are more than our number of rows, it's not going to help much for our analysis, hence we will not use this feature. Director has 30% null values and will not be used by us.

Then, We Apply**:**

Removing stop-words:

- Stop-words are basically a set of commonly used words in any language, not just in English. If we remove the words that are very commonly used in a given language, we can focus on the important words instead.

Stemming:-

- Stemming is the process of removing a part of a word, or reducing a word to its stem or root. Applying stemming to reduce words to their basic form or stem, which may or may not be a legitimate word in the language.

Vectorization:

- TF-IDF is an abbreviation for Term Frequency Inverse Document Frequency. This is a very common algorithm to transform text into a meaningful representation of numbers which is used to fit a machine learning algorithm for prediction.
- We have also utilized the PCA because it can help us improve performance at a very low cost of model accuracy. Other benefits of PCA include reduction of noise in the data, feature selection (to a certain extent), and the ability to produce independent, uncorrelated features of the data.
- So, it's essential to transform our text into tfidf vectorizer, then convert it into an array so that we can fit into our model.

- ## Finding number of clusters

- The goal is to separate groups with similar characteristics and assign them to clusters.
- We used the Elbow method and the Silhouette score to do so, and we have determined that 26 clusters should be an optimal number of clusters.

- **Fitting into model**

  - In this task, we have implemented a K means clustering algorithm. K-means is a technique for data clustering that may be used for unsupervised machine learning. It is capable of classifying unlabeled data into a predetermined number of clusters based on similarities (k).

K-Means Clustering

K-Means Clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science. In this topic, we will learn what is K-means clustering algorithm, how the algorithm works, along with the Python implementation of k-means clustering.
K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on

It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training. It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.

The algorithm takes the unlabeled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.

The k-means clustering algorithm mainly performs two tasks:

- Determines the best value for K center points or centroids by an iterative process.
- Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.

Hence each cluster has datapoints with some commonalities, and it is away from other clusters.

The working of the K-Means algorithm is explained in the below steps:

Step-1: Select the number K to decide the number of clusters.

Step-2: Select random K points or centroid. (It can be other from the input dataset).

Step-3: Assign each data point to their closest centroid, which will form the predefined K clusters.

Step-4: Calculate the variance and place a new centroid of each cluster.

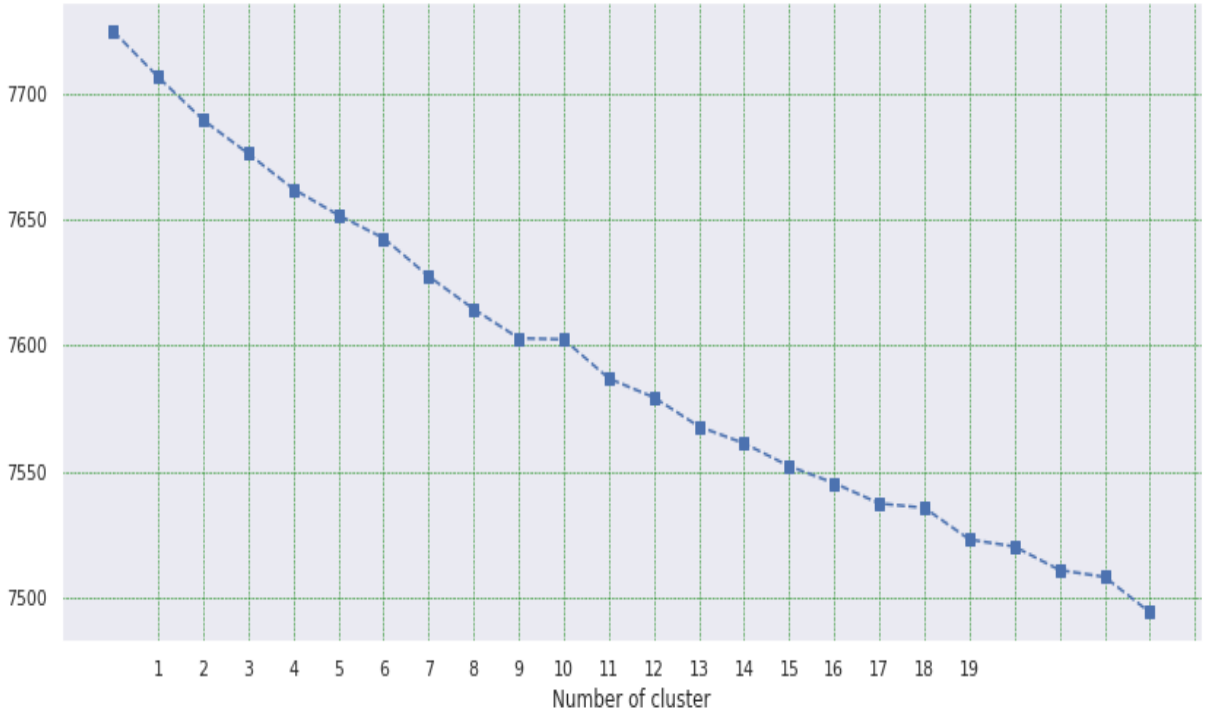
Step-5: Repeat the third steps, which means reassign each datapoint to the new closest centroid of each cluster.

Step-6: If any reassignment occurs, then go to step-4 else go to FINISH.

Step-7: The model is ready.

Finding Number Of Clusters Using Elbow Method

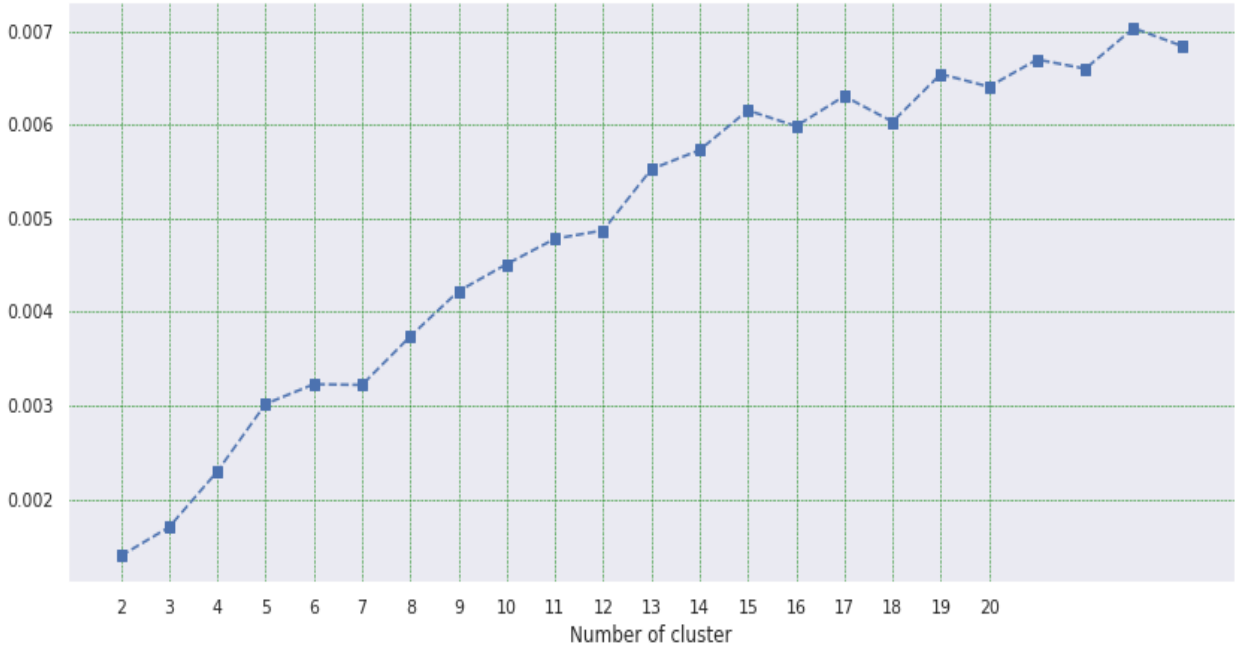
```
cluster: 1       SSE: 7724.51
cluster: 2       SSE: 7706.33
cluster: 3       SSE: 7689.28
cluster: 4       SSE: 7675.98
cluster: 5       SSE: 7661.96
cluster: 6       SSE: 7651.62
cluster: 7       SSE: 7642.53
cluster: 8       SSE: 7627.41
cluster: 9       SSE: 7614.44
cluster: 10      SSE: 7602.91
cluster: 11      SSE: 7602.42
cluster: 12      SSE: 7586.94
cluster: 13      SSE: 7579.45
cluster: 14      SSE: 7567.92
cluster: 15      SSE: 7561.26
cluster: 16      SSE: 7552.22
cluster: 17      SSE: 7545.50
cluster: 18      SSE: 7537.46
cluster: 19      SSE: 7535.74
cluster: 20      SSE: 7523.18
cluster: 21      SSE: 7520.39
cluster: 22      SSE: 7511.10
cluster: 23      SSE: 7508.36
cluster: 24      SSE: 7494.37
```



- **Looks like we can go with 20 clusters from the visualizations.**

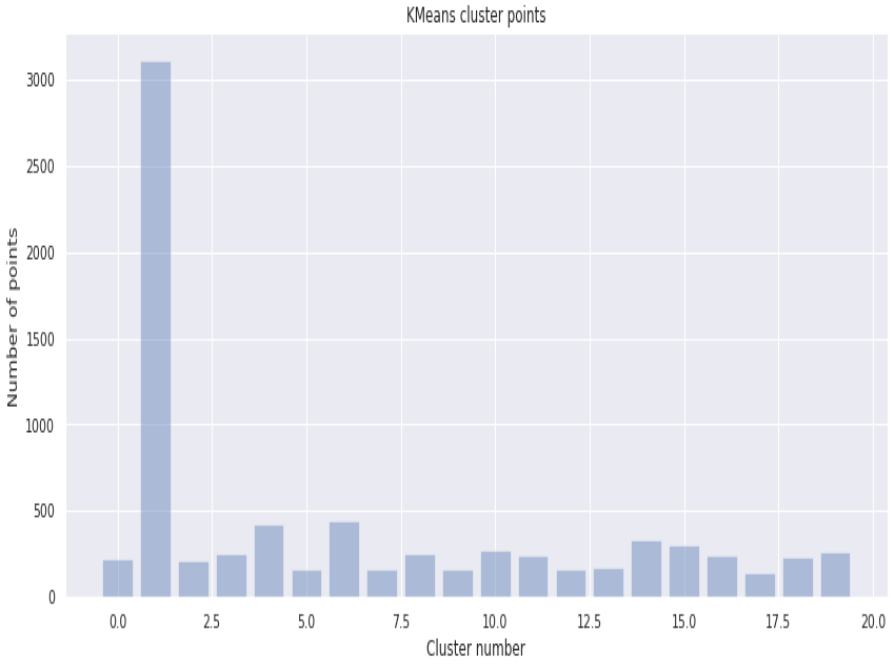
Finding number of Clusters from Sillhoute's score

```
cluster: 2      Sillhoute: 0.0014
cluster: 3      Sillhoute: 0.0017
cluster: 4      Sillhoute: 0.0023
cluster: 5      Sillhoute: 0.0030
cluster: 6      Sillhoute: 0.0032
cluster: 7      Sillhoute: 0.0032
cluster: 8      Sillhoute: 0.0037
cluster: 9      Sillhoute: 0.0042
cluster: 10     Sillhoute: 0.0045
cluster: 11     Sillhoute: 0.0048
cluster: 12     Sillhoute: 0.0049
cluster: 13     Sillhoute: 0.0055
cluster: 14     Sillhoute: 0.0057
cluster: 15     Sillhoute: 0.0062
cluster: 16     Sillhoute: 0.0060
cluster: 17     Sillhoute: 0.0063
cluster: 18     Sillhoute: 0.0060
cluster: 19     Sillhoute: 0.0065
cluster: 20     Sillhoute: 0.0064
cluster: 21     Sillhoute: 0.0067
cluster: 22     Sillhoute: 0.0066
cluster: 23     Sillhoute: 0.0070
cluster: 24     Sillhoute: 0.0068
```



- **Looks like we can go with 20 clusters from both the visualizations.**

Implementing K-means



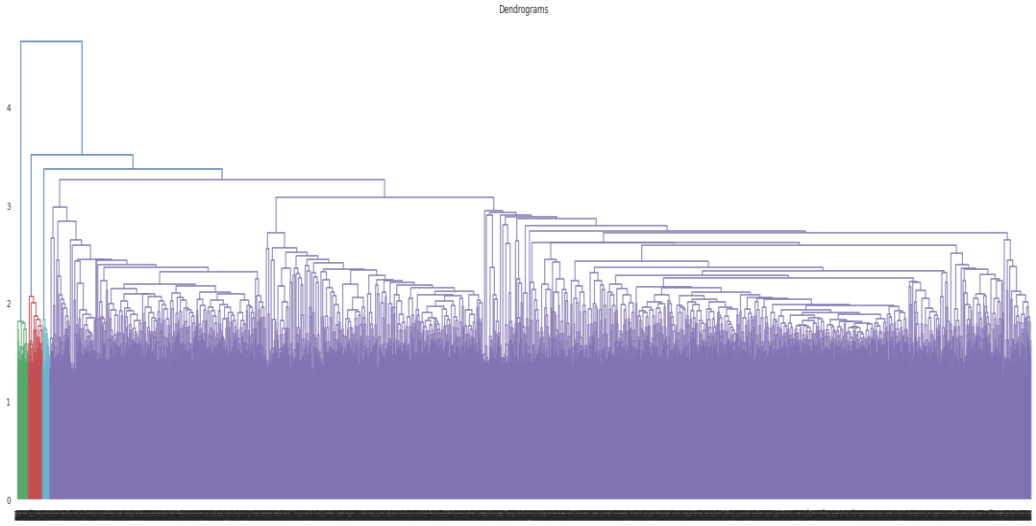
- **Cluster 0 have highest number of cluster points.**

	title	listed_in
29	#blackAF	[TV Comedies]
65	13 Sins	[Horror Movies, Thrillers]
148	A Bad Moms Christmas	[Comedies]
174	A Futile and Stupid Gesture	[Comedies]
197	A Little Help with Carol Burnett	[Stand-Up Comedy & Talk Shows, TV Comedies]

Evalution

Silhouette Coefficient: 0.00618
Davies-Bouldin index: 9.650555
Calinski-Harabasz index: 10.351

Hierarchical Clustering
Finding number of clusters from Dendogram



- **By using Sillhouette's score and Elbow method , we generated optimal of 20 clusters for K Means and from Dendogram , 6 clusters were generated.**

Evaluation:

- **Silhouette Coefficient: -0.002**
- **Calinski-Harabasz index: 6.6459**
- **Davies-Bouldin index: 19.0527**

Challenges

- Reading the dataset and understanding the problem statement.
- Designing multiple visualizations to summarize the Data points in the dataset and effectively communicating the results and insights to the reader.
- Data preprocessing – Remove stop words, Stemming and vectorization.
- Careful tuning of hyper parameters as it affects accuracy.
- Computation time was a big challenge for us.

Conclusion

- There are about 70% movies and 30% TV shows on Netflix.
- The United States has the highest number of content on Netflix by a huge margin
- followed by India.
- Raul Campos and Jan Sulter collectively have directed the most content on Netflix.
- Anupam Kher has acted in the highest number of films on Netflix.
- Drama is the most popular genre followed by comedy.
- More of the content is released in holiday season - October, November, December
- and January.
- The number of releases have significantly increased after 2015 and have dropped
- in 2021 because of Covid 19.
- NULL HYPOTHESIS -The number of TV shows on Netflix have tripled and number
- of movies have reduced by 2000 between 2010 and 2018. (REJECTED)
- By using Sillhouette's score and Elbow method , we generated optimal of 20 clusters for K Means and from Dendogram , 6 clusters were generated.
- In both the cases, one cluster accounts more than 3000 points whereas in other clusters the points were unevenly distributed. 3.For Tfidf K Means is best for identification than Hierarchical as the evaluation metrics also indicates the same.