# Capstone Project

## Netflix Movies & TV Shows Clustering

## By –Amogha K

## CONTENTS:

- Introduction
- Problem statement
- Data summary
- Data Preparation
- Exploratory Data analysis
- Clustering
- Conclusion

# INTRODUCTION:

- ✓ Netflix, Inc. is an American company based in Los Gatos, California. Founded in 1997 by Reed Hastings and Marc Randolph in Scotts Valley, California,
- ✓ Netflix is a media distribution company. It started with DVD distribution via mail, but has evolved substantially over the course of its existence.
- ✓ Netflix is focused on streaming video. Some of its content is licensed, and some of the content is produced in-house.
- ✓ Netflix originally focused on movies, but television shows are probably the more common format today.
- ✓ Netflix works on a subscription model, where users get unlimited access to content with a paid subscription.
- ✓ Netflix can be accessed via web browsers or via application software installed on smart TVs, tablet computers, smartphones. It is available in 4K resolution.

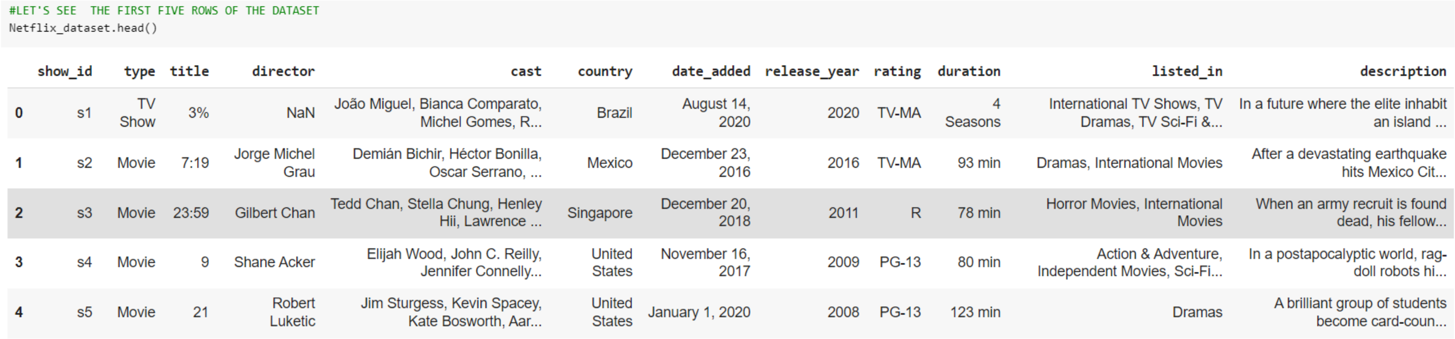
# PROBLEM STATEMENT

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

# DATA SUMMARY

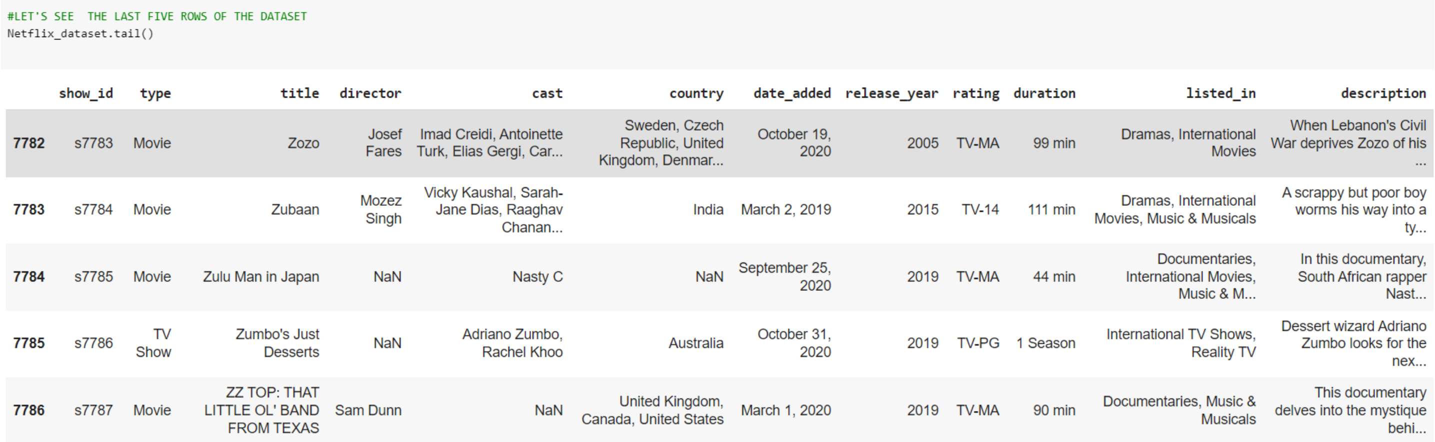
- show_id : Unique ID for every Movie / Tv Show
- type : Identifier - A Movie or TV Show
- title : Title of the Movie / Tv Show
- director : Director of the Movie
- cast : Actors involved in the movie / show
- country : Country where the movie / show was produced
- date_added : Date it was added on Netflix
- 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 description

# DATA PREPARATION:



```
#LET'S SEE  THE FIRST FIVE ROWS OF THE DATASET
Netflix_dataset.head()
```

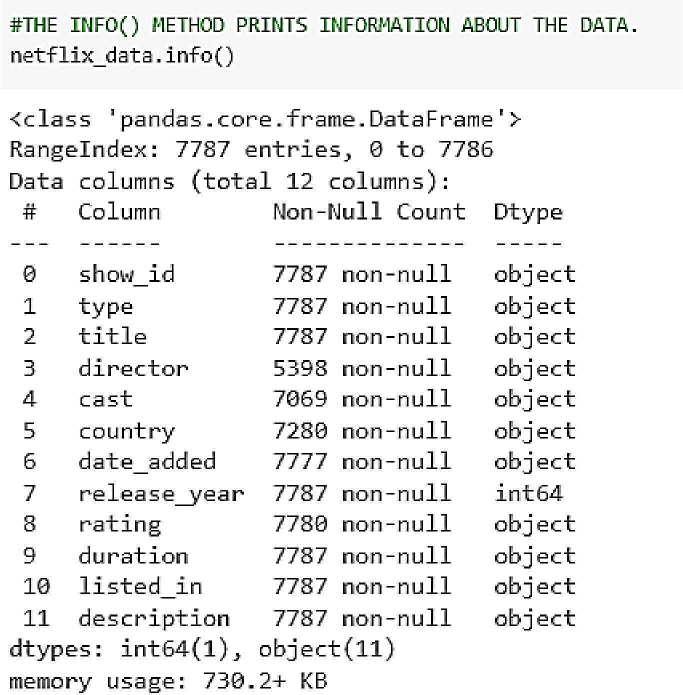
| | show_id | type | title | director | cast | country | date_added | release_year | rating | duration | listed_in | description |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | s1 | TV Show | 3% | NaN | João Miguel, Bianca Comparato, Michel Gomes, R... | Brazil | August 14, 2020 | 2020 | 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... |



```
#LET'S SEE  THE LAST FIVE ROWS OF THE DATASET
Netflix_dataset.tail()
```

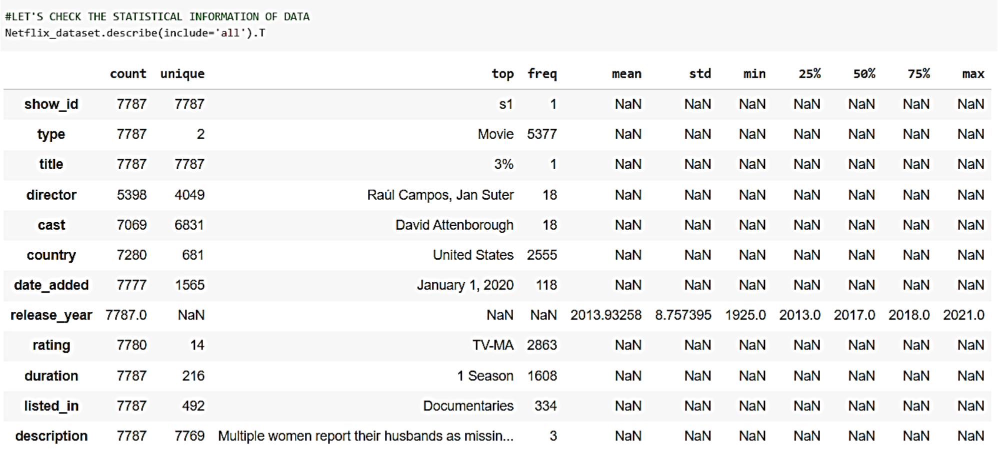
| | show_id | type | title | director | cast | country | date_added | release_year | rating | duration | listed_in | description |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **7782** | s7783 | Movie | Zozo | Josef Fares | Imad Creidi, Antoinette Turk, Elias Gergi, Car... | Sweden, Czech Republic, United Kingdom, Denmar... | October 19, 2020 | 2005 | TV-MA | 99 min | Dramas, International Movies | When Lebanon's Civil War deprives Zozo of his ... |
| **7783** | s7784 | Movie | Zubaan | Mozez Singh | Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan... | India | March 2, 2019 | 2015 | TV-14 | 111 min | Dramas, International Movies, Music & Musicals | A scrappy but poor boy worms his way into a ty... |
| **7784** | s7785 | Movie | Zulu Man in Japan | NaN | Nasty C | NaN | September 25, 2020 | 2019 | TV-MA | 44 min | Documentaries, International Movies, Music & M... | In this documentary, South African rapper Nast... |
| **7785** | s7786 | TV Show | Zumbo's Just Desserts | NaN | Adriano Zumbo, Rachel Khoo | Australia | October 31, 2020 | 2019 | TV-PG | 1 Season | International TV Shows, Reality TV | Dessert wizard Adriano Zumbo looks for the nex... |
| **7786** | s7787 | Movie | ZZ TOP: THAT LITTLE OL' BAND FROM TEXAS | Sam Dunn | NaN | United Kingdom, Canada, United States | March 1, 2020 | 2019 | TV-MA | 90 min | Documentaries, Music & Musicals | This documentary delves into the mystique behi... |

❑ Head and Tail is  used to view a small sample of a Series or the Data frame from our data



```
#THE INFO() METHOD PRINTS INFORMATION ABOUT THE DATA.
netflix_data.info()
```

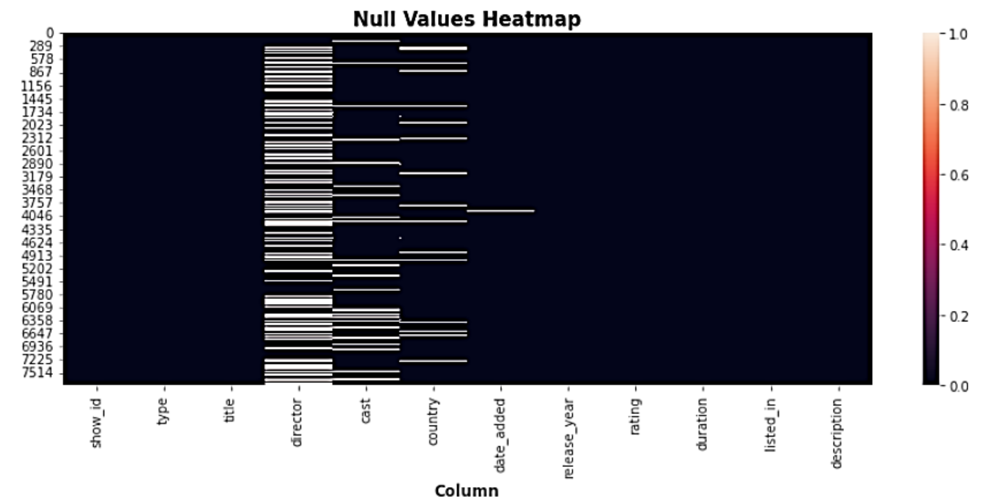
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7787 entries, 0 to 7786
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   show_id       7787 non-null   object
 1   type          7787 non-null   object
 2   title         7787 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
```

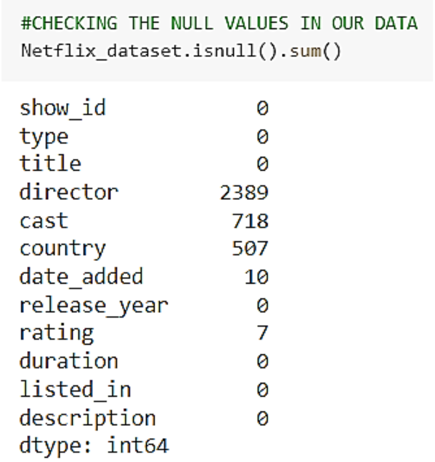


```
#LET'S CHECK THE STATISTICAL INFORMATION OF DATA
Netflix_dataset.describe(include='all').T
```

| | count | unique | top | freq | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **show_id** | 7787 | 7787 | s1 | 1 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **type** | 7787 | 2 | Movie | 5377 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **title** | 7787 | 7787 | 3% | 1 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **director** | 5398 | 4049 | Raúl Campos, Jan Suter | 18 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **cast** | 7069 | 6831 | David Attenborough | 18 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **country** | 7280 | 681 | United States | 2555 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **date_added** | 7777 | 1565 | January 1, 2020 | 118 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **release_year** | 7787.0 | NaN | NaN | NaN | 2013.93258 | 8.757395 | 1925.0 | 2013.0 | 2017.0 | 2018.0 | 2021.0 |
| **rating** | 7780 | 14 | TV-MA | 2863 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **duration** | 7787 | 216 | 1 Season | 1608 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **listed_in** | 7787 | 492 | Documentaries | 334 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **description** | 7787 | 7769 | Multiple women report their husbands as missin... | 3 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |

- Our data has 7877 rows and 12 columns.
- Our data includes only one numerical type of data and the rest is all categorical data.





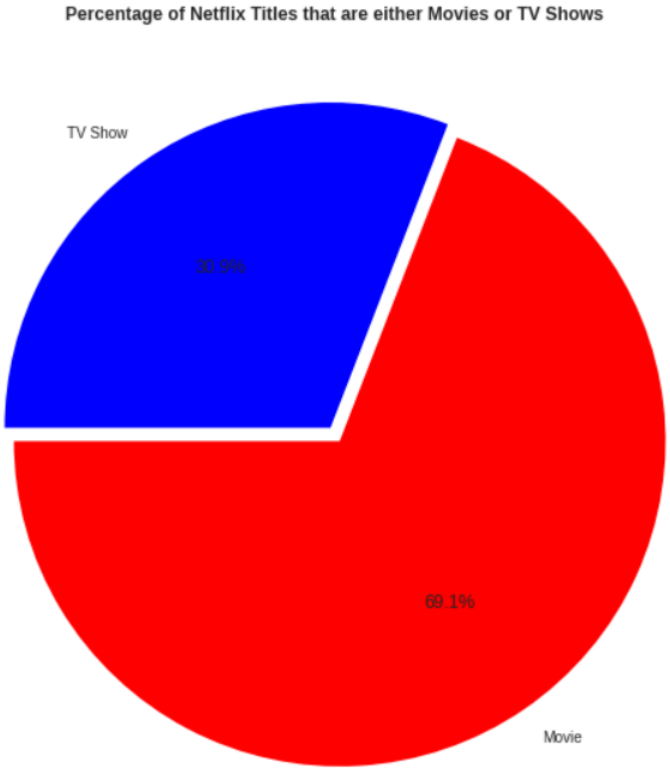
```
#CHECKING THE NULL VALUES IN OUR DATA
Netflix_dataset.isnull().sum()
```

```
show_id            0
type               0
title              0
director        2389
cast             718
country          507
date_added        10
release_year       0
rating             7
duration           0
listed_in          0
description        0
dtype: int64
```

- There are 3631 null values in the dataset, 2389 null values in the director column, 718 null values in the cast column, 507 null values in the country column, 10 in date_added, and 7 in rating. so we need to handle the null values.
- For the Director column, we filled in null values as 'No director', and for the Cast column, we filled in 'No cast' and country as 'country unavailability’.
- The other two columns, "date_added" and "rating," contain an insignificant portion of the data, so we can drop them from the dataset.

# EXPLORATORY DATA ANALYSIS:

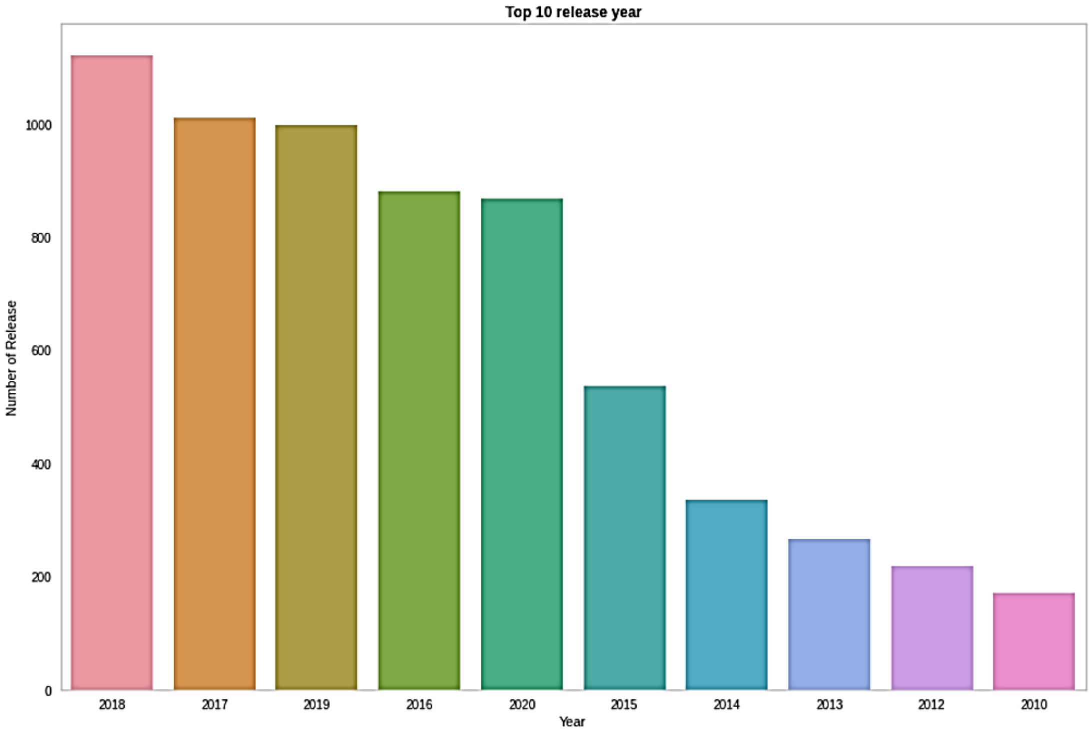
## TYPE OF CONTENT ON NETFLIX



- ❖ It's 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. There are about 69.1% movies and 30.9% TV shows on Netflix.
- ❖ There are more than twice as many movies uploaded to Netflix as there are TV shows. This does not imply that movies are more indulgent than TV shows. Because TV shows may have several seasons, which consist of a number of episodes, TV shows have a much longer run time than movies.

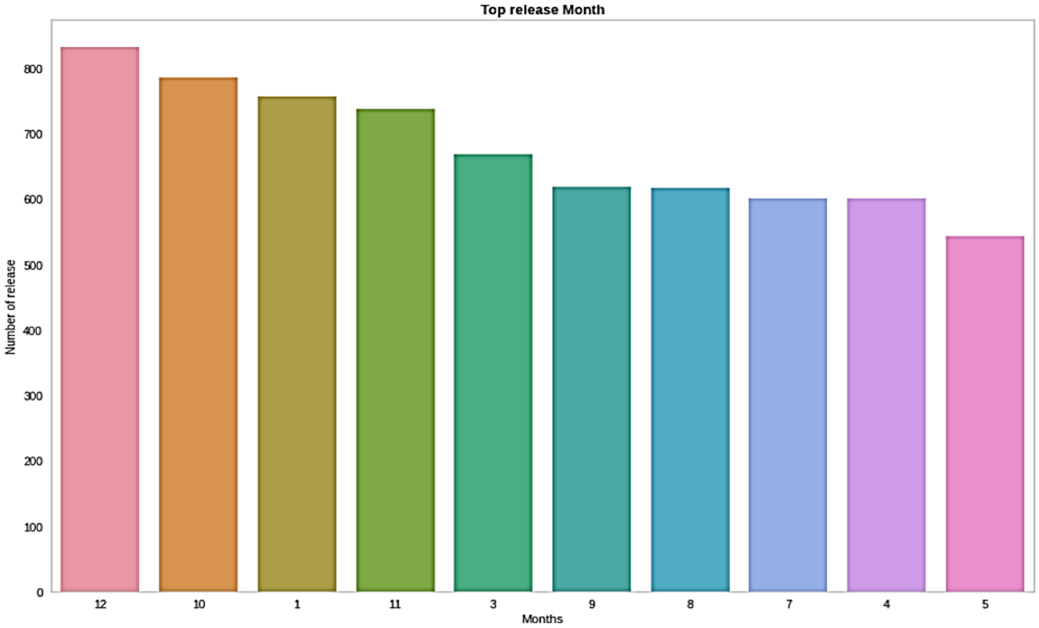
# RELEASE YEAR AND RELEASE MONTH WITH CONTENT

## YEAR:



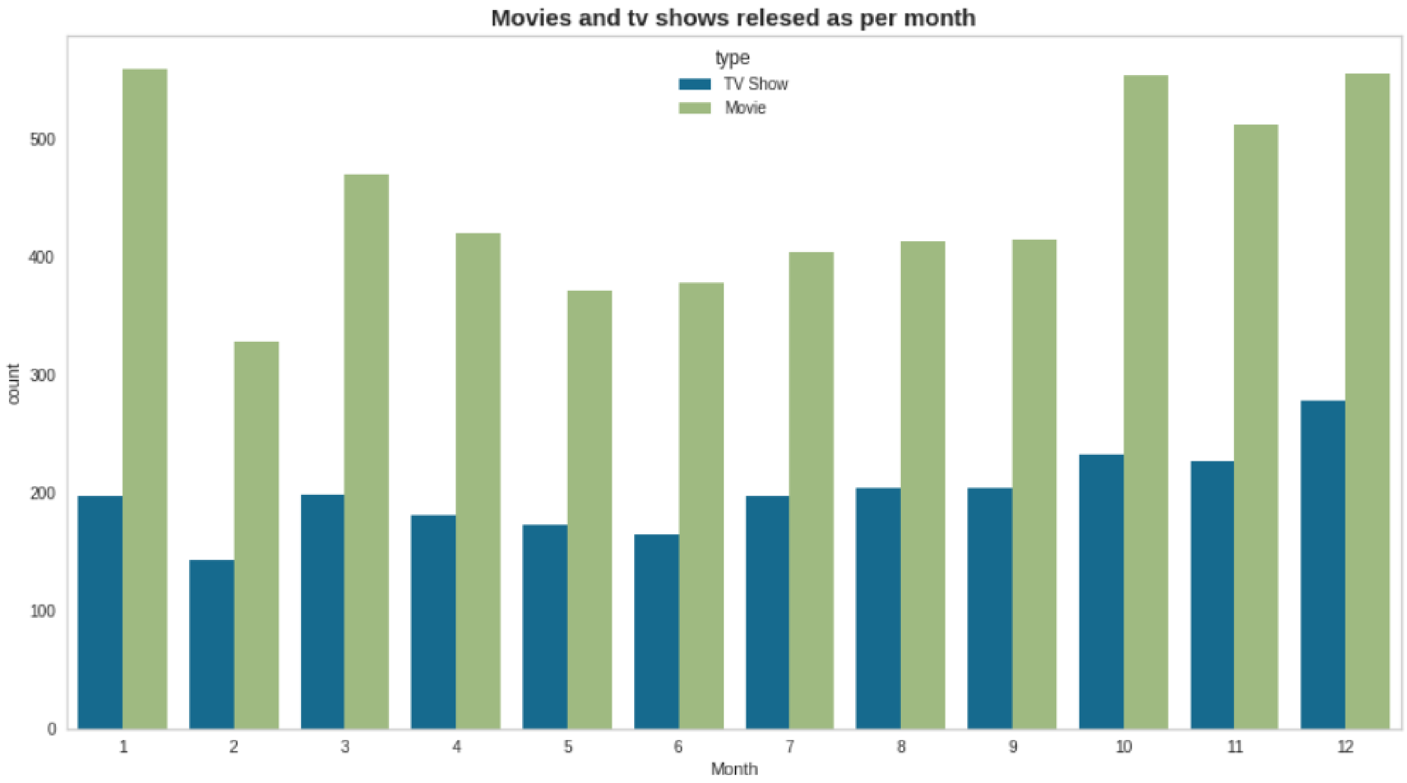
- ✓ Netflix produced the most content in 2018 followed by 2017.
- ✓ Because of the Corona virus (Pandemic), less TV shows and movies were released in 2020 and 2021.

## MONTH:



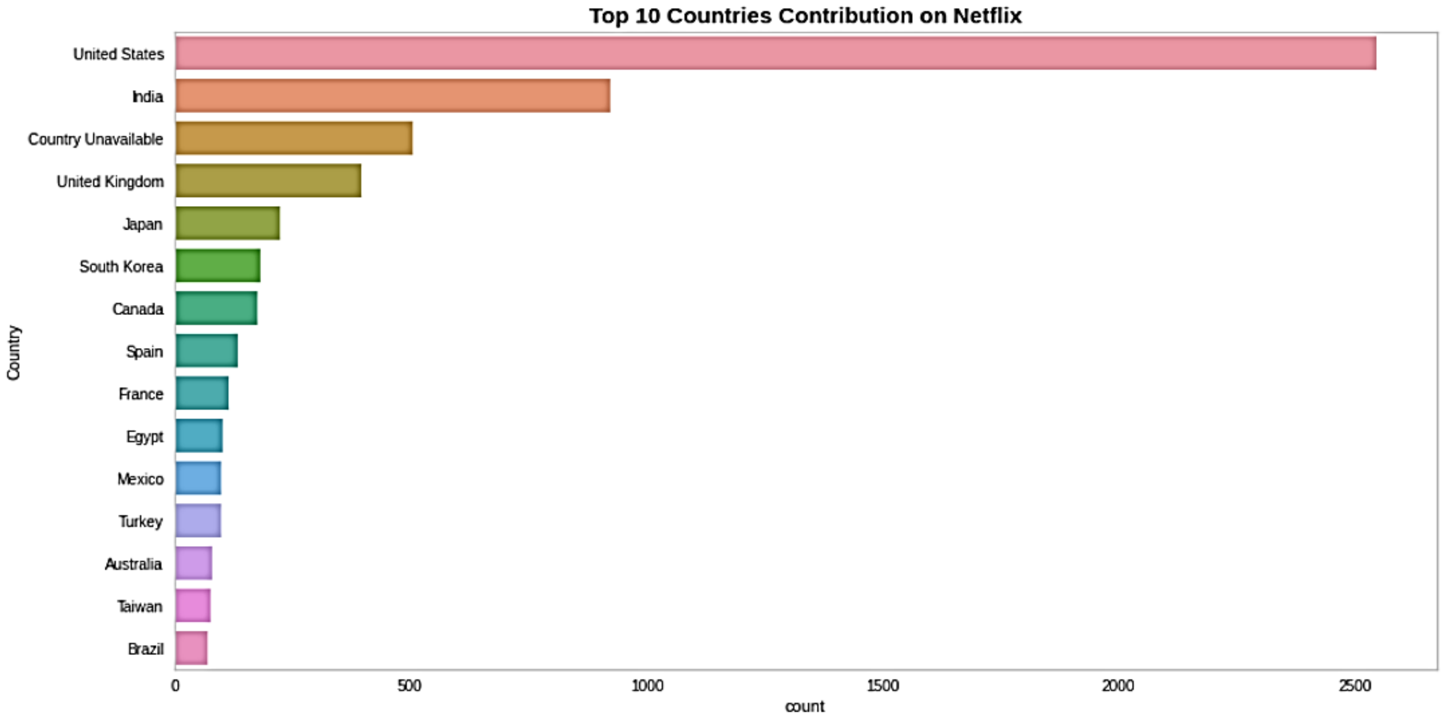
- Most of the holidays came in December and January in the US and Europe. Most of the holidays came to India in October month. So releasing a movie or TV show between October and January is the best way to earn a lot of profit as the whole family will be spending time with each other and watching shows.
- The best 4 months to release content are October, November, December, and January.

# MOVIES AND TV SHOWS RELESED AS PER MONTH



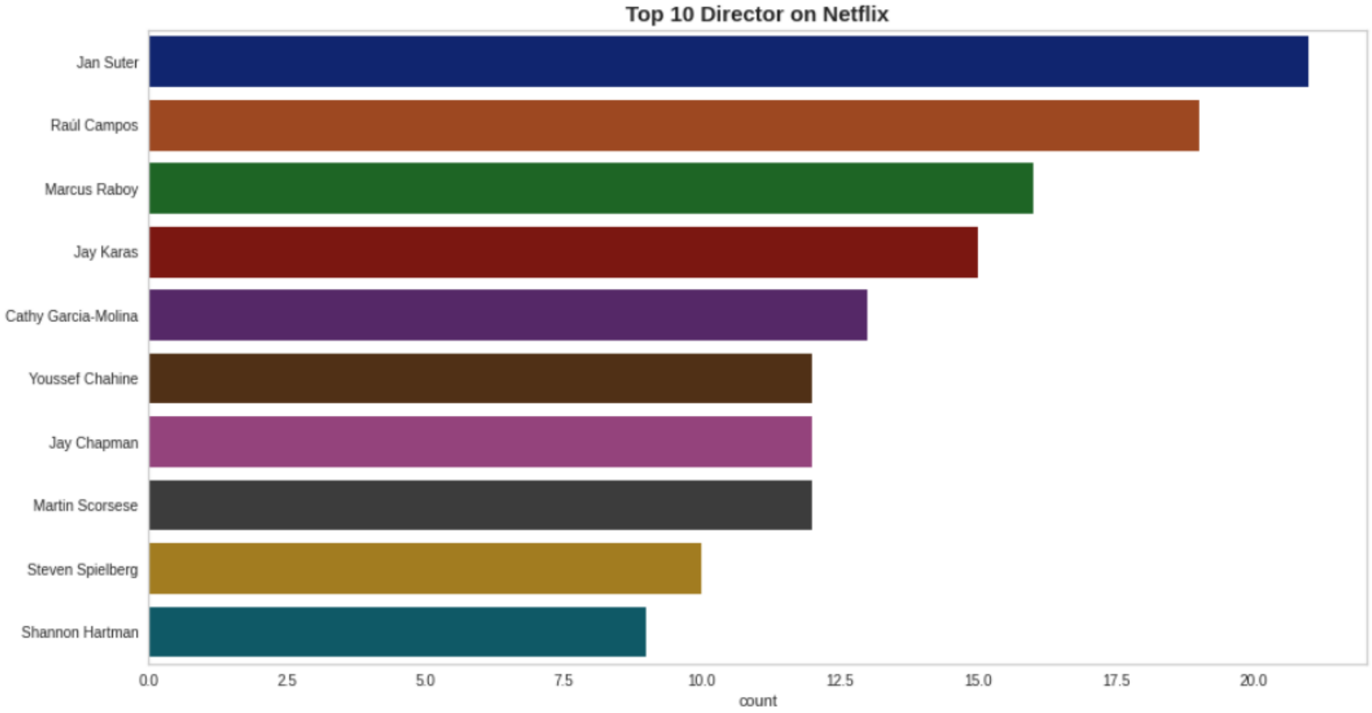
✓ According to the above graph, movie releases are higher than TV show releases in all month.

# TOP CONTRIBUTING COUNTRIES ON NETFLIX



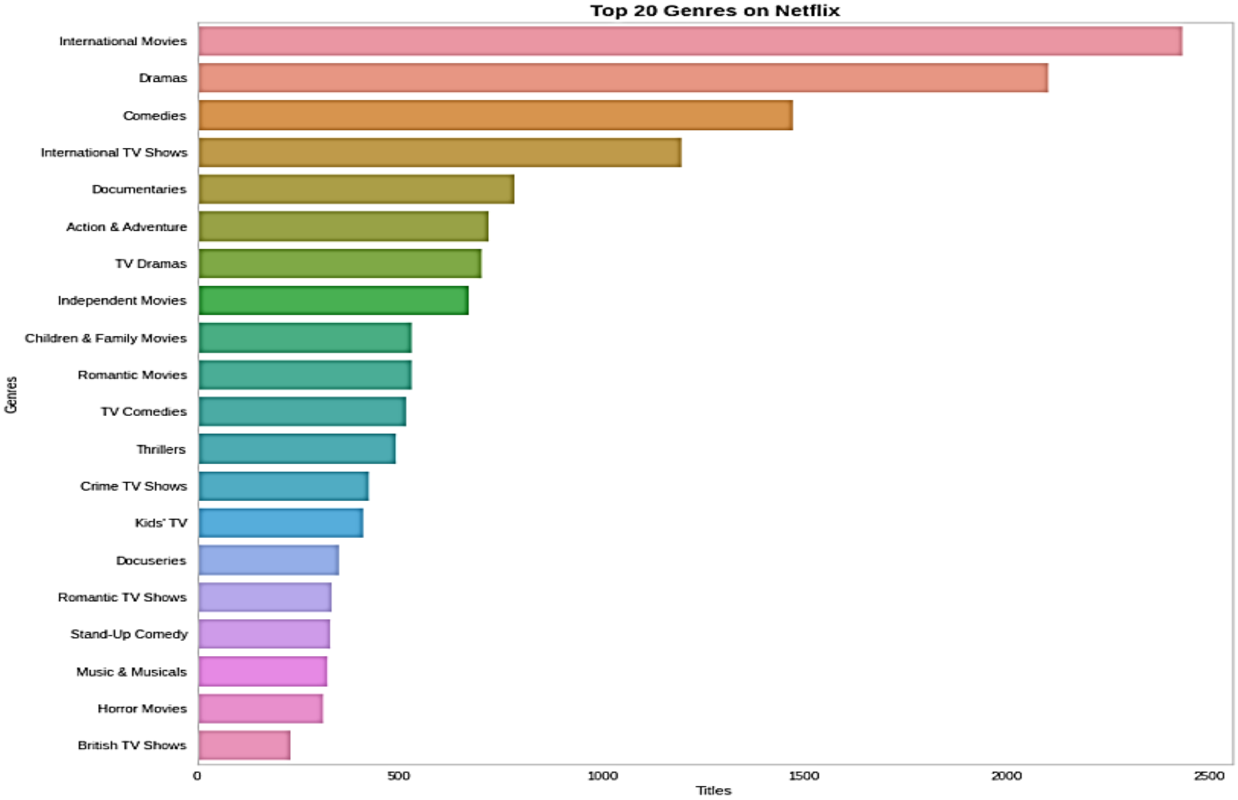
- From the above graph, we can see the top 15 countries that contribute the most to Netflix. The United States produces the most content in terms of quantity.
- India is the second top country contributing to Netflix.

# TOP 10 DIRECTORS ON NETFLIX



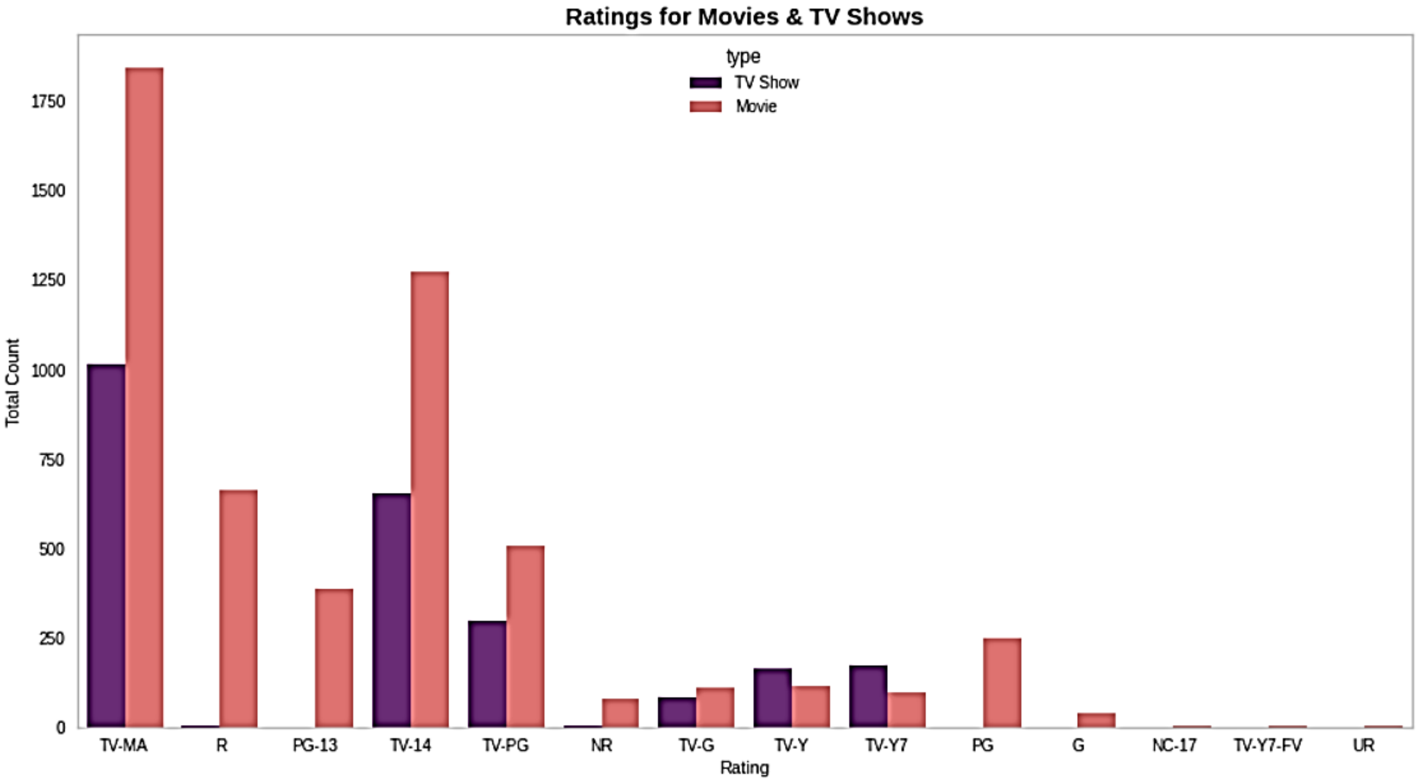
- ❖ The most popular director on Netflix, with the most titles, is Jan Suter, and next is Raul Compos.

# GENRES ON NETFLIX



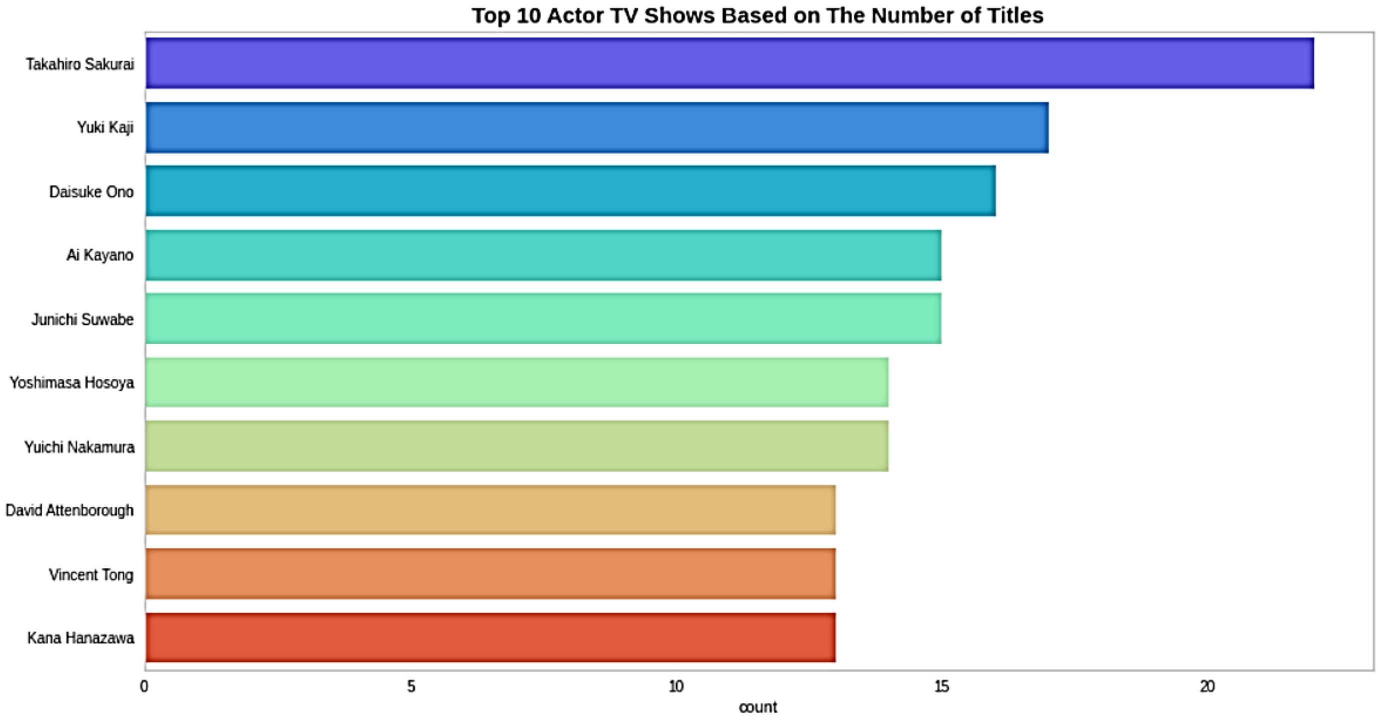
- From the graph, we know that international movies take the first place, followed by dramas and comedies.

# RATINGS FOR MOVIES & TV SHOWS



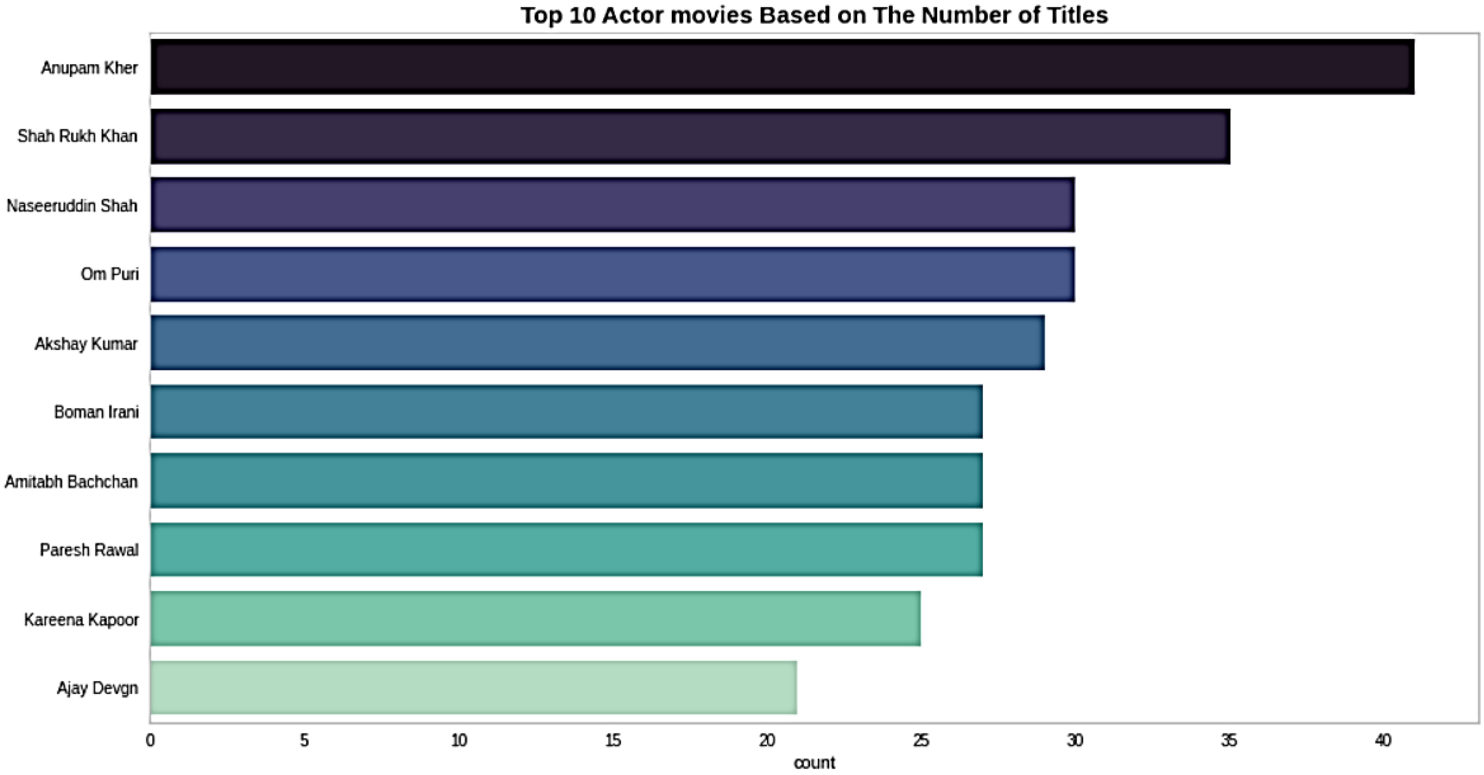
- The largest number of TV shows have a "TV-MA" rating. "TV-MA" is a rating assigned by the TV Parental Guidelines to a television program designed for mature audiences only.

# TOP ACTORS IN TV SHOWS



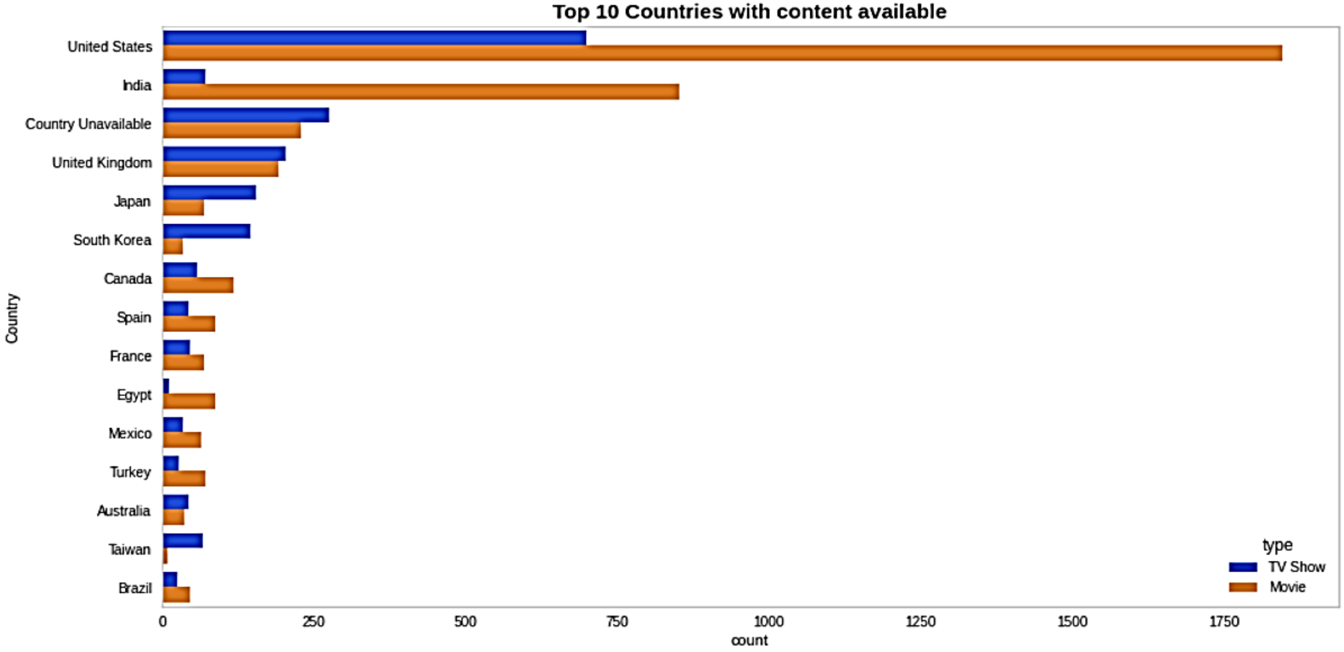
❖ Based on the number of titles, Takahiro Sakurai is the top actor on Netflix TV shows.

# TOP ACTORS IN MOVIES



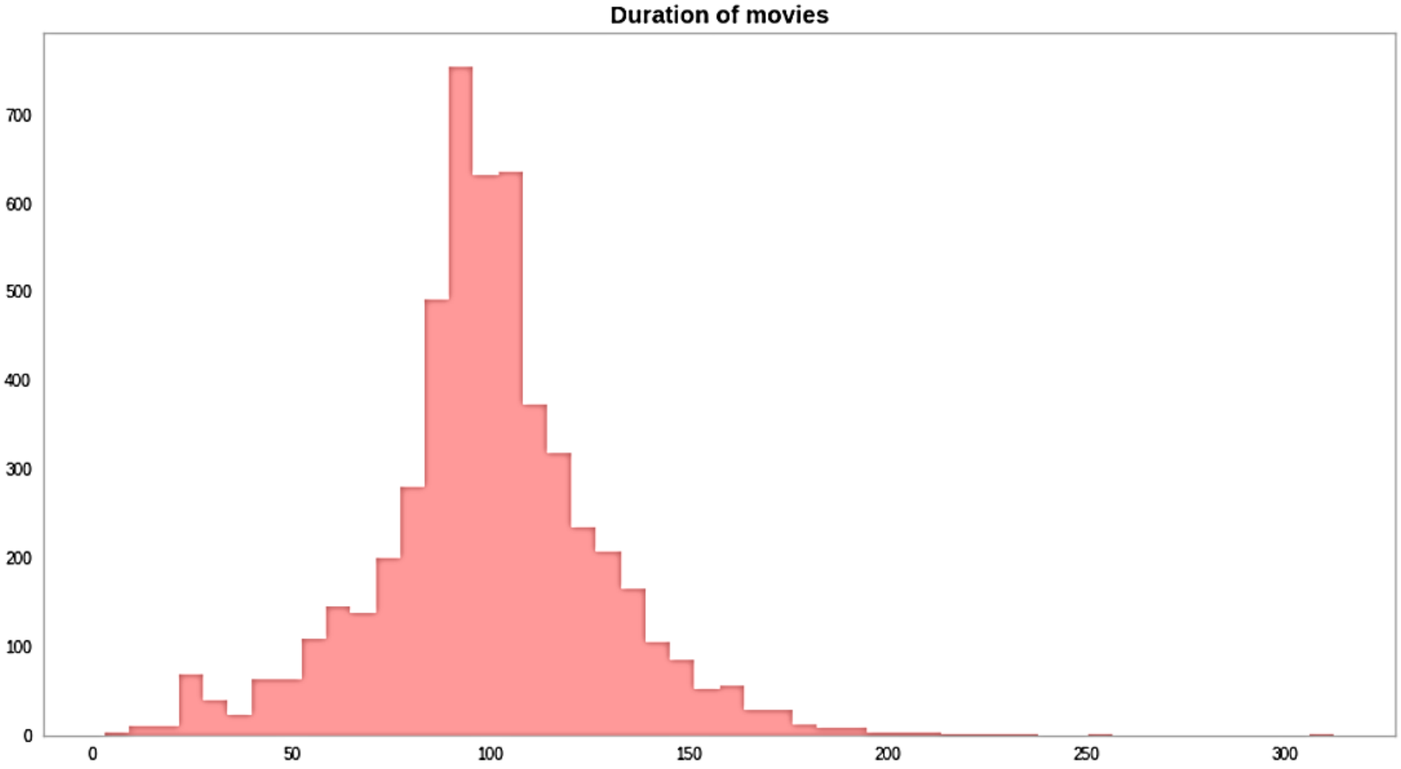
- The top actor on Netflix Movies, based on the number of titles, is Anupam Kher.

# COUNTRIES WITH CONTENTS AVAILABLE



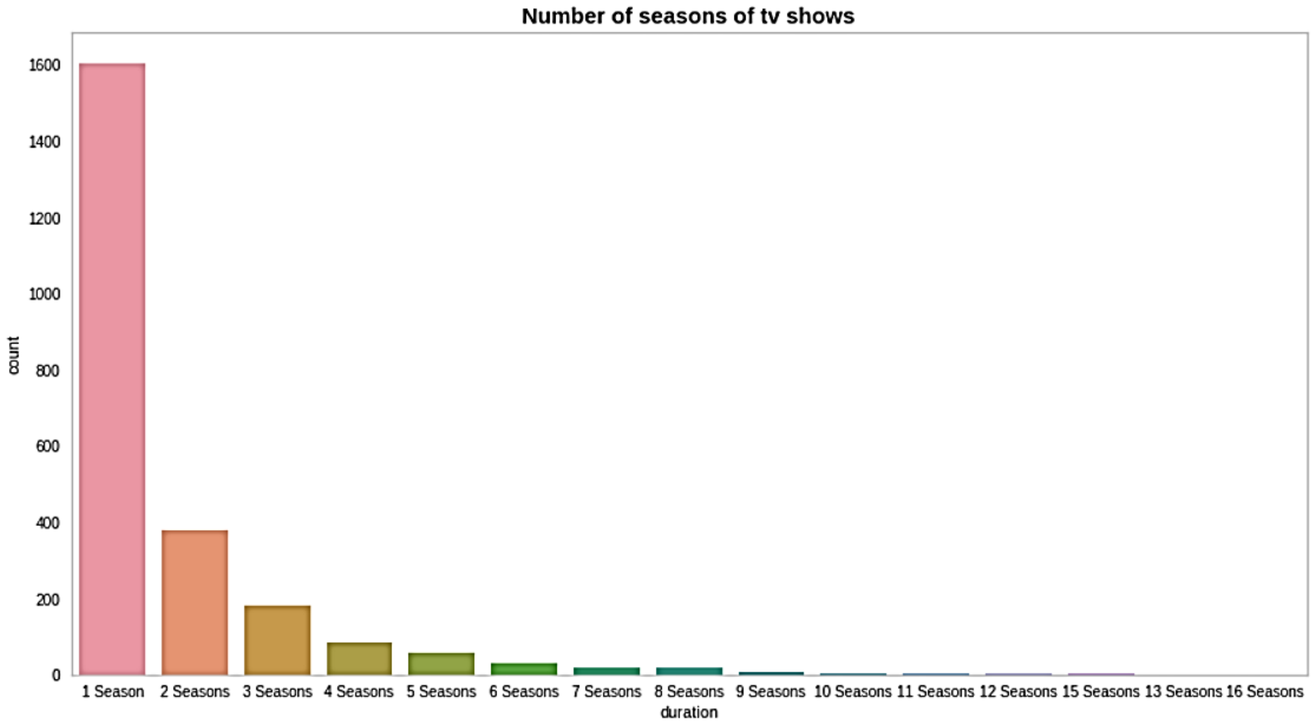
- ❑ The United States is a leading producer of both types of content; this makes sense since Netflix is a US company.
- ❑ The majority of Netflix content available in India, Canada, Spain, France, Egypt, and Turkey consists of movies.
- ❑ South Korea has a higher frequency of television shows, which explains the current KDrama culture.

# DISTRIBUTION OF DURATION OF MOVIES



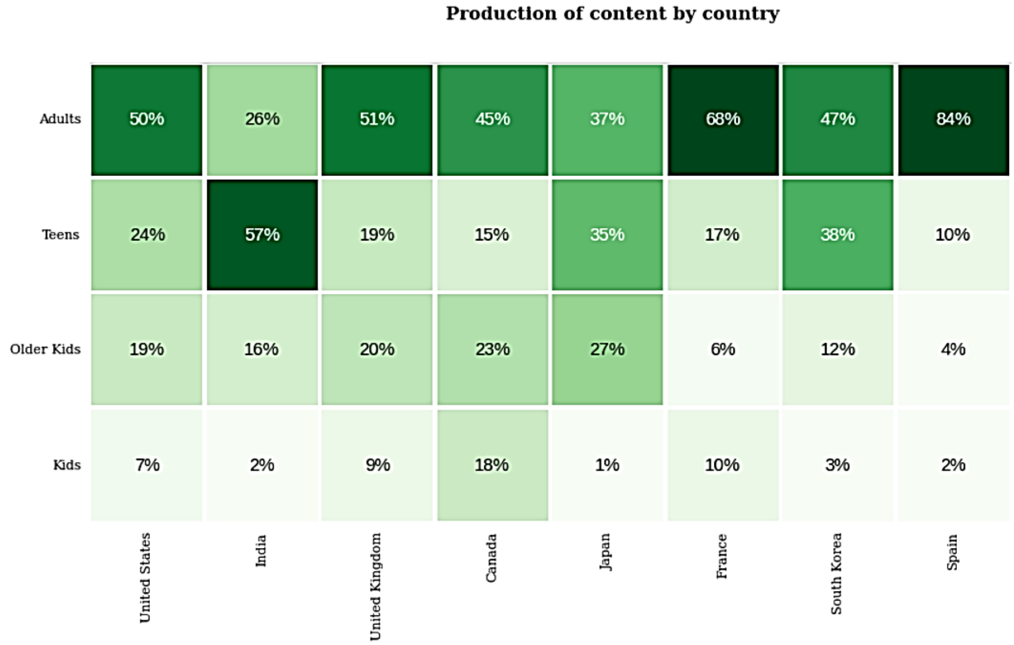
- ❖ The majority of the Movies run between 90 and 110 minutes.

# NUMBER OF SEASONS OF TV SHOWS



- It has been noted that 1608 television shows only have one season. There are only a few television shows that have the longest running times.

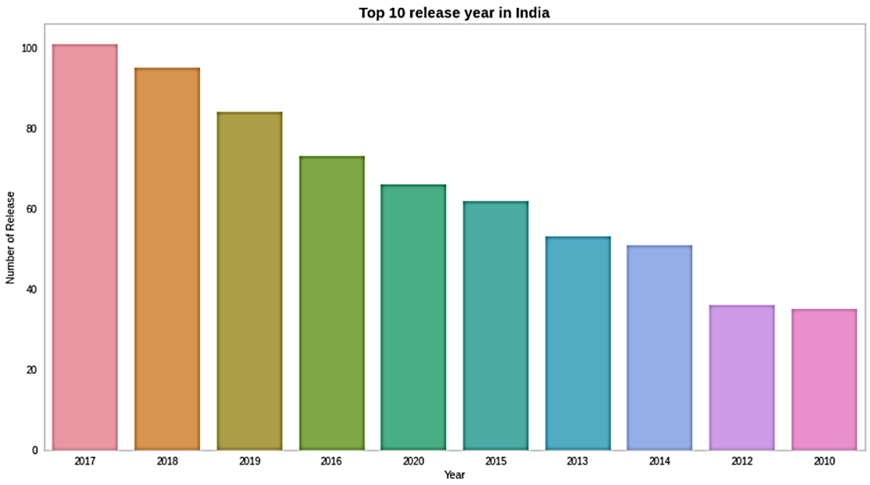
# PRODUCTION OF CONTENT BY COUNTRY



- ➢ According to the correlation graph, adults prefer to watch movies and TV shows in Spain, France, the United Kingdom, and the United States.
- ➢ In India, 57 percent of teens watch movies and TV shows, while only 26 percent of adults do that means in India Netflix need to more concentrate to Teens.
- ➢ Spain is producing the most adult content on Netflix, at 84%.

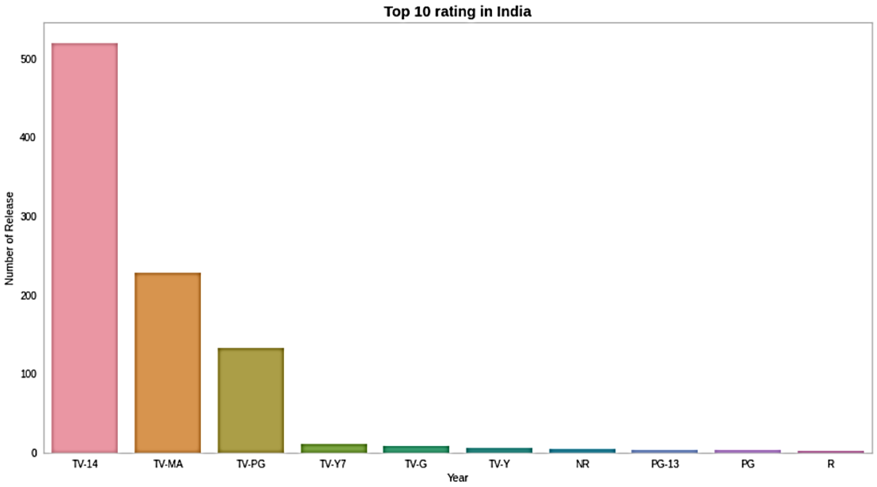
# ANALYSIS WITH COUNTRY INDIA

## TOP CONTENT RELEASE YEAR IN INDIA



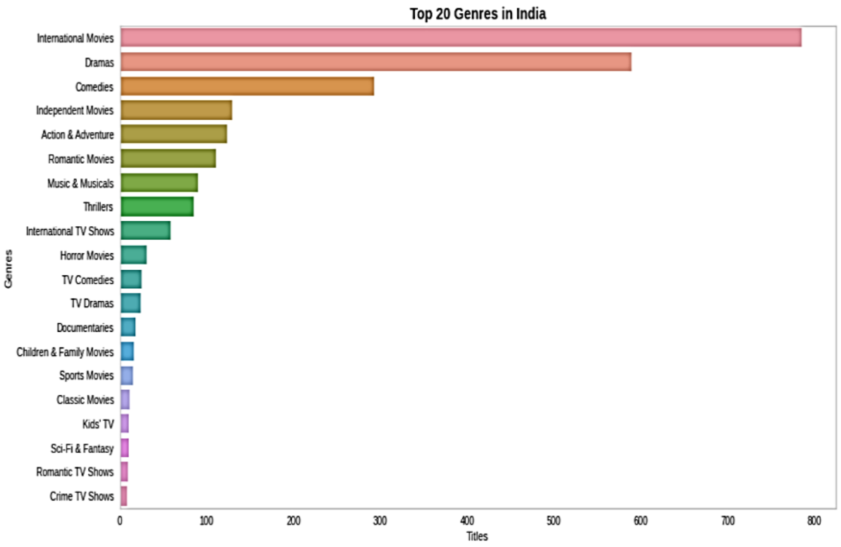
- ❑ The most popular content released in 2017 according to India.

## TOP RATINGS IN INDIA

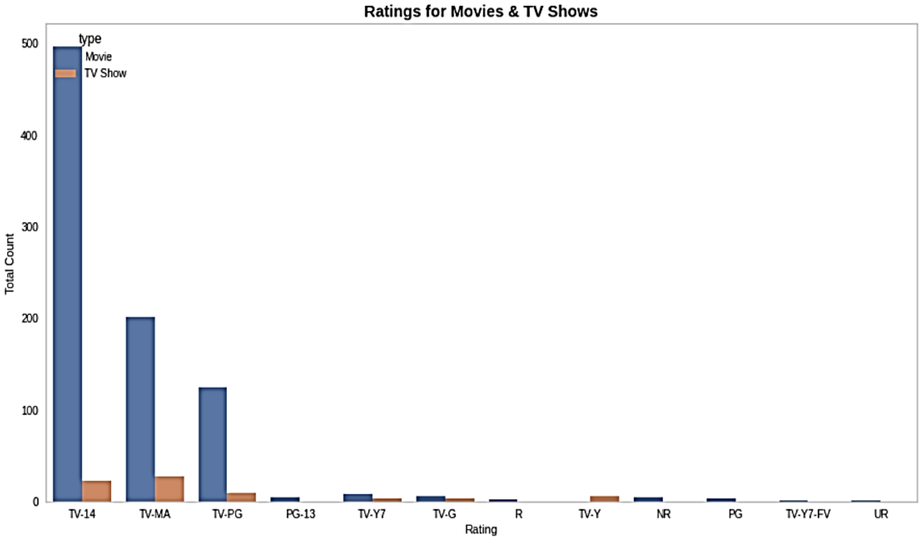


- ❑ Teenage content has the highest rating in India.

## GENRES IN INDIA



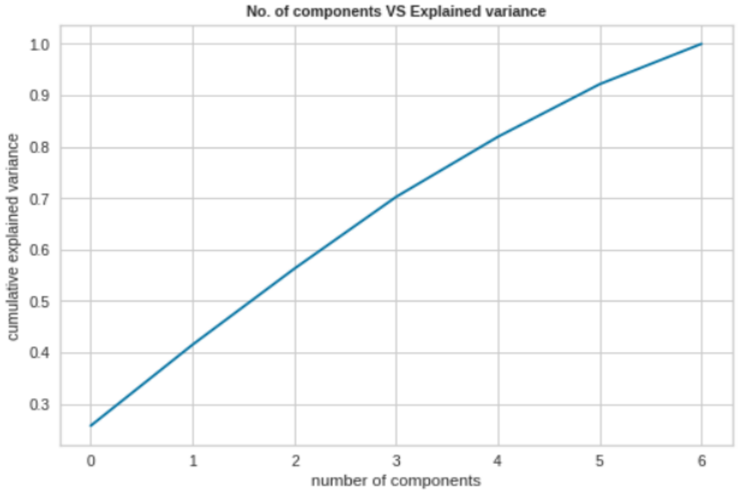
## COMPARISOIN RATINGS FOR MOVIES AND TV SHOWS IN INDIA



- ✓ International content is in high demand in India.
- ✓ Teenagers have the highest rating in India.

# PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis (PCA) is a linear dimensionality reduction technique that can be utilized for extracting information from a high-dimensional space by projecting it into a lower-dimensional sub-space. It tries to preserve the essential parts that have more variation of the data and remove the non-essential parts with fewer variation



Typically, we want the explained variance to be between 95 and 99%. In this case, 7 principal components are required to explain 99% (most approximate is 100%) of the variance.

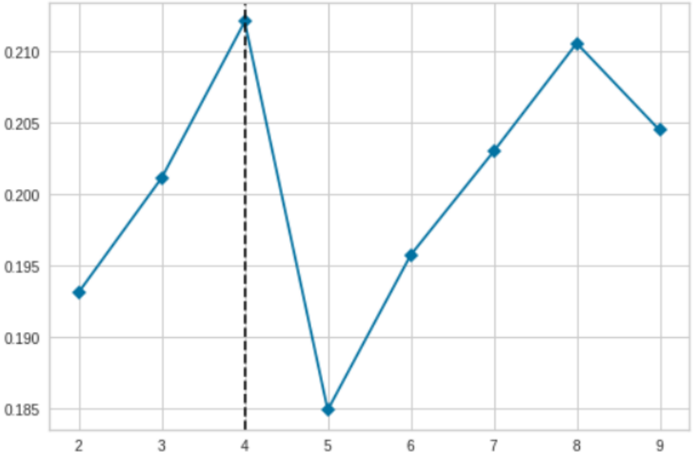
# CLUSTERING

## KMEANS CLUSTERING

K-means Clustering is a type of unsupervised learning that 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.

## K-ELBOW METHOD

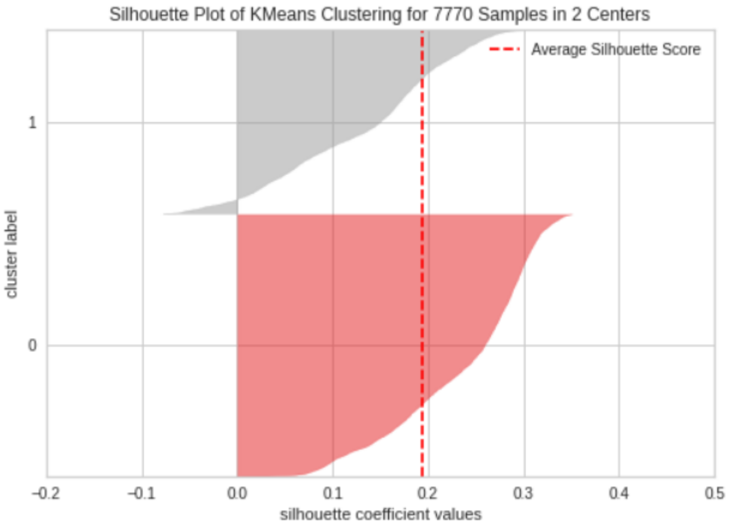
The Elbow Method is one of the most popular methods to determine this optimal value of k.

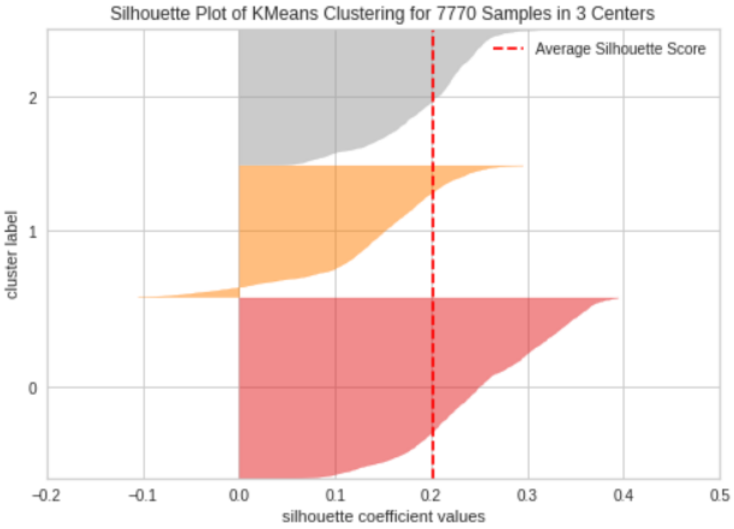


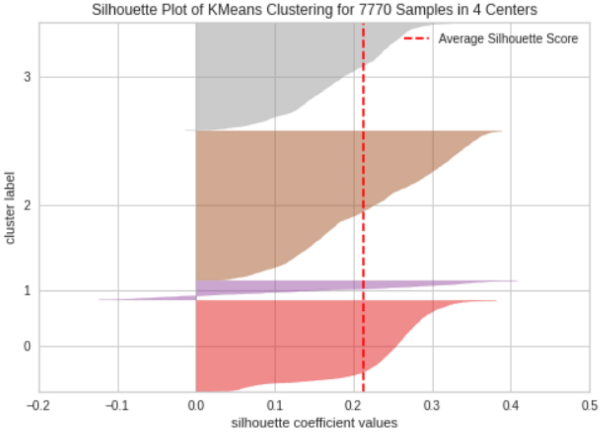
To determine the optimal number of clusters, we have to select the value of k at the "elbow," i.e., the point after which the distortion/inertia start decreasing in a linear. Thus, for the given data, we conclude that the optimal number of clusters for the data is 4.

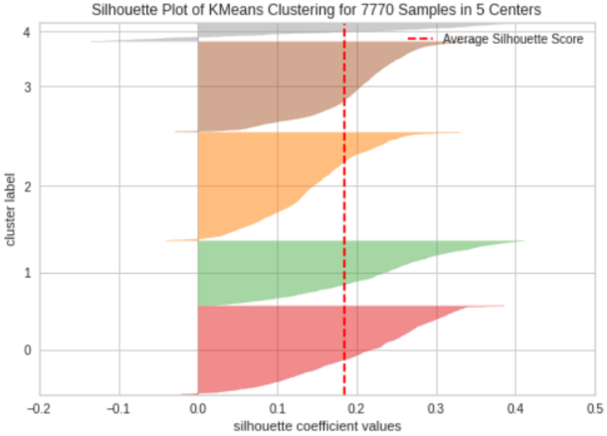
## SILHOUETTE SCORE

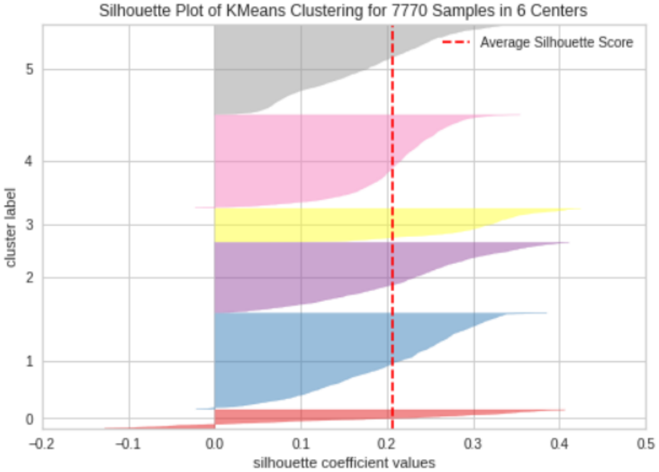
Silhouette Coefficient or silhouette score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1.

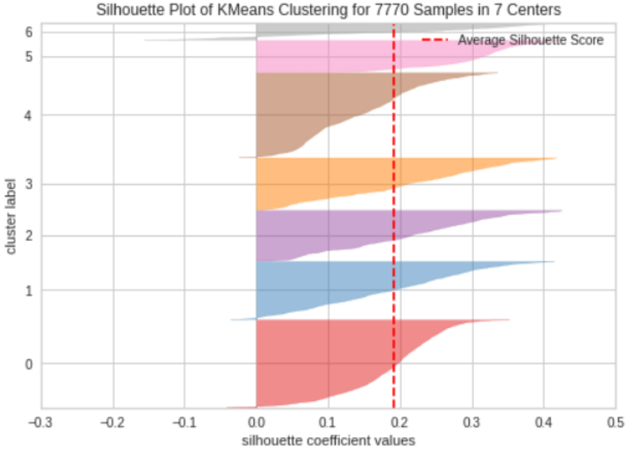





Silhouette Plot of KMeans Clustering for 7770 Samples in 4 Centers

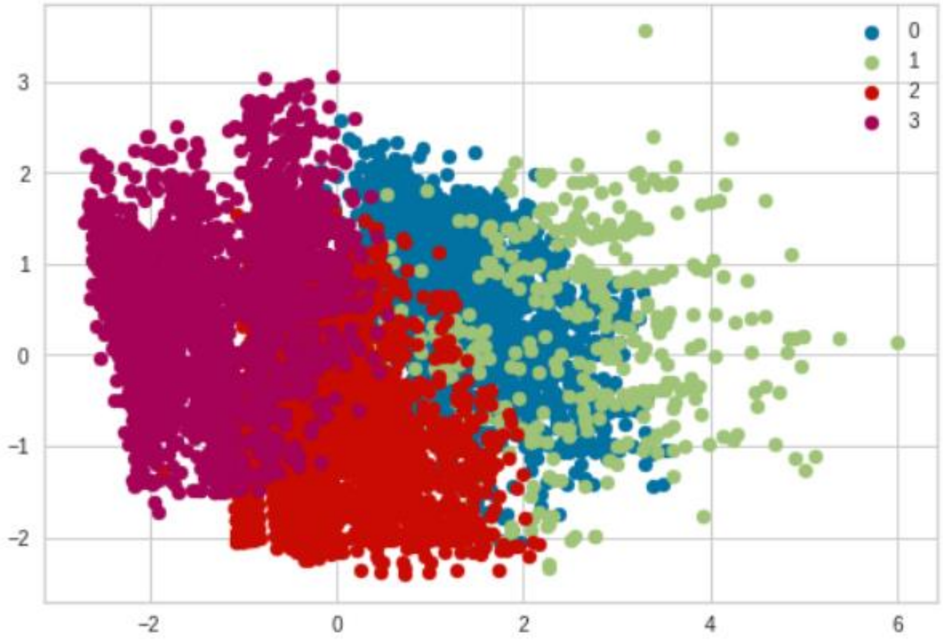

Silhouette Plot of KMeans Clustering for 7770 Samples in 5 Centers


Silhouette Plot of KMeans Clustering for 7770 Samples in 6 Centers


Silhouette Plot of KMeans Clustering for 7770 Samples in 7 Centers

- Here is the silhouette analysis done on the above plots to select an optimal value for n_clusters.The value of 4 for n_clusters looks to be the optimal one. The silhouette score is 0.21.

## K MEANS CLUSTERING

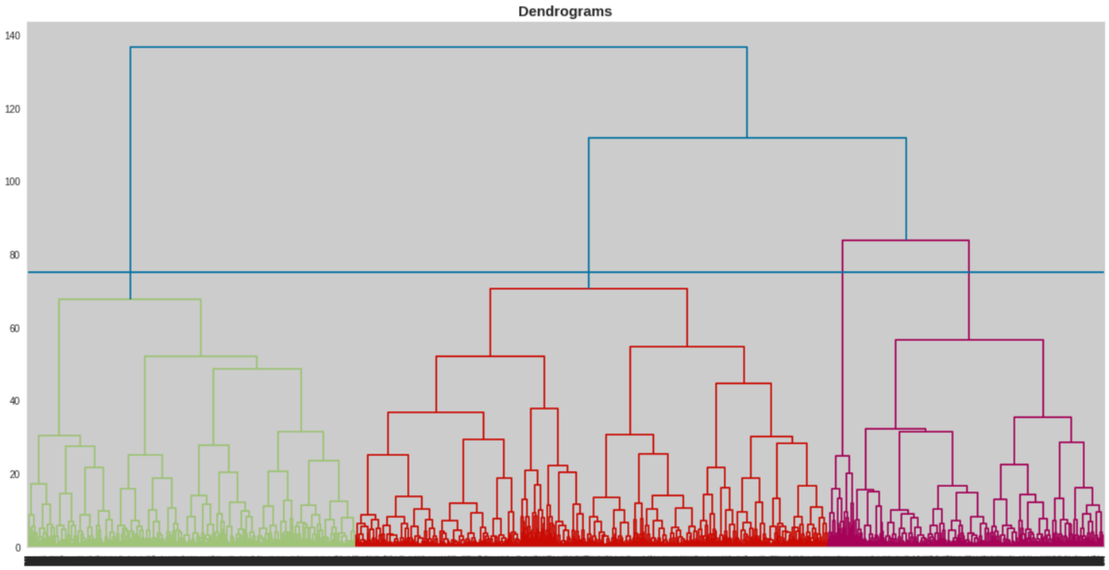


- The numbers 0 to 3 represent 4-distinct clusters formed by K-means clustering.

# HIRARCHICAL CLUSTERING

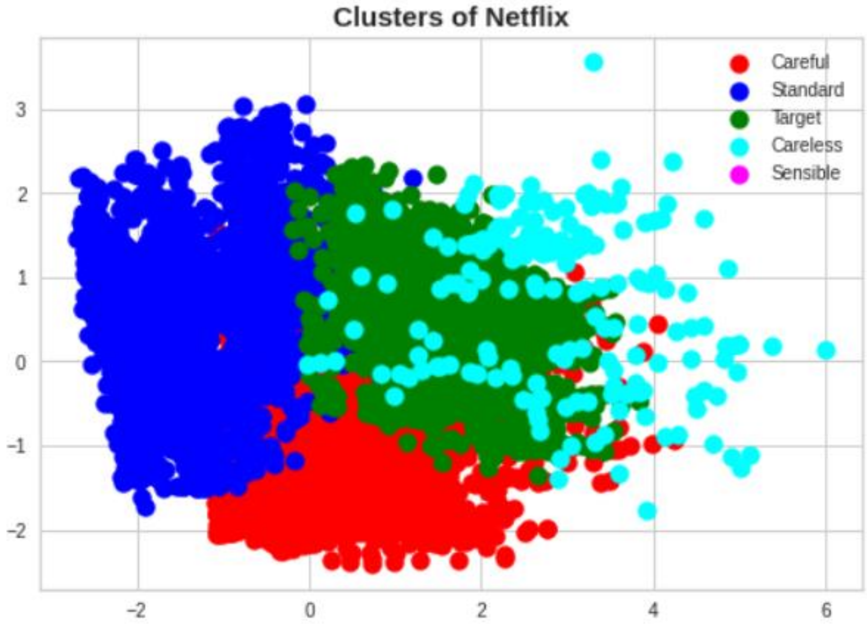
Hierarchical clustering is another unsupervised machine learning algorithm, which is used to group the unlabeled datasets into a cluster and also known as hierarchical cluster analysis or HCA.

In this algorithm, we develop the hierarchy of clusters in the form of a tree, and this tree-shaped structure is known as the dendrogram.



# AGGLOMERATIVE CLUSTERING

The agglomerative hierarchical clustering algorithm is a popular example of HCA. To group the datasets into clusters, it follows the bottom-up approach. It means, this algorithm considers each dataset as a single cluster at the beginning, and then start combining the closest pair of clusters together. It does this until all the clusters are merged into a single cluster that contains all the datasets.



# CONCLUSION

We have done EDA and clustering , and we have drawn many interesting inferences from the Netflix title dataset; here’s a summary of a few of them:

- Movies are the most popular type of content on Netflix. It appears that Netflix has focused more attention on increasing movie content than TV shows. Movies have increased much more dramatically than TV shows.
- There are about 70% movies and 30% TV shows on Netflix.
- Most films were released in the years 2018, 2019, and 2020.
- The number of releases significantly increased after 2015 and dropped in 2021 because of COVID 19.
- The months of October, November, December, and January had the largest number of films and television series released.
- More of the content is released during the holiday season—October, November, December, and January.
- 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.
- The most popular genre on Netflix is international movies, followed by stand-up comedy and drama.

- The majority of the films run between 90 and 110 minutes.
- Highest number of TV shows consisting of a single season
- In the correlation heatmap, In India, most of the teens watch Netflix.
- TV-MA has the highest number of ratings for tv shows i,e adult ratings
- In India, teens mostly watched international movies.
- Principal Component analysis (PCA)reduced the number of components as 7 with approximately 99% of variance.
- We used the elbow and silhouette score methods for K-means clustering to determine the number of k. Using both methods, we discovered that k = 4 is the best value for clustering.
- Using the hierarchical clustering method again, we find that k = 4 is the optimal value for clustering.

# THANK YOU!