# Capstone Project-4

## Netflix Movies and TV Show Clustering

By-Md.ImranHaji

# Points to discuss:-

- Problem Statement.
- Data summary.
- Data Analysis and understand the relationships between variables
- Hypothesis Testing
- Text processing.
- Model Implementation and evaluation.
- 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. 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**

Along with Recommendation building, will look in some behavior.

- Exploratory Data Analysis
- what content is available by different countries?
- Is Netflix has increasingly focusing on TV rather than movies in recent years?
- Clustering similar content by matching text-based features

# Data summary.

**As i was provided with data set of a OTT platform Netflix, which is one of the on-demand internet streaming media and online movie rental service provider. It has million's of subscribed members acrose the gloab enjoying more than 100 million hours of TV shows and movies per day, enjoying TV series, documentaries, and feature films across a wide variety of genres and languages.**

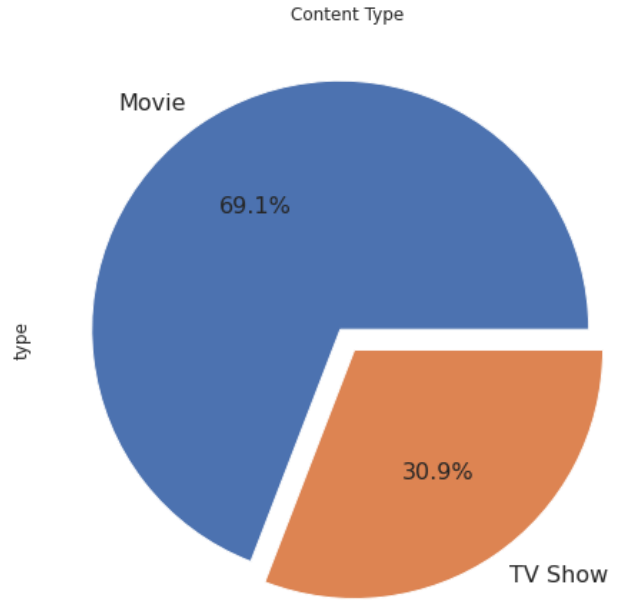
**For which Data Description is:-**

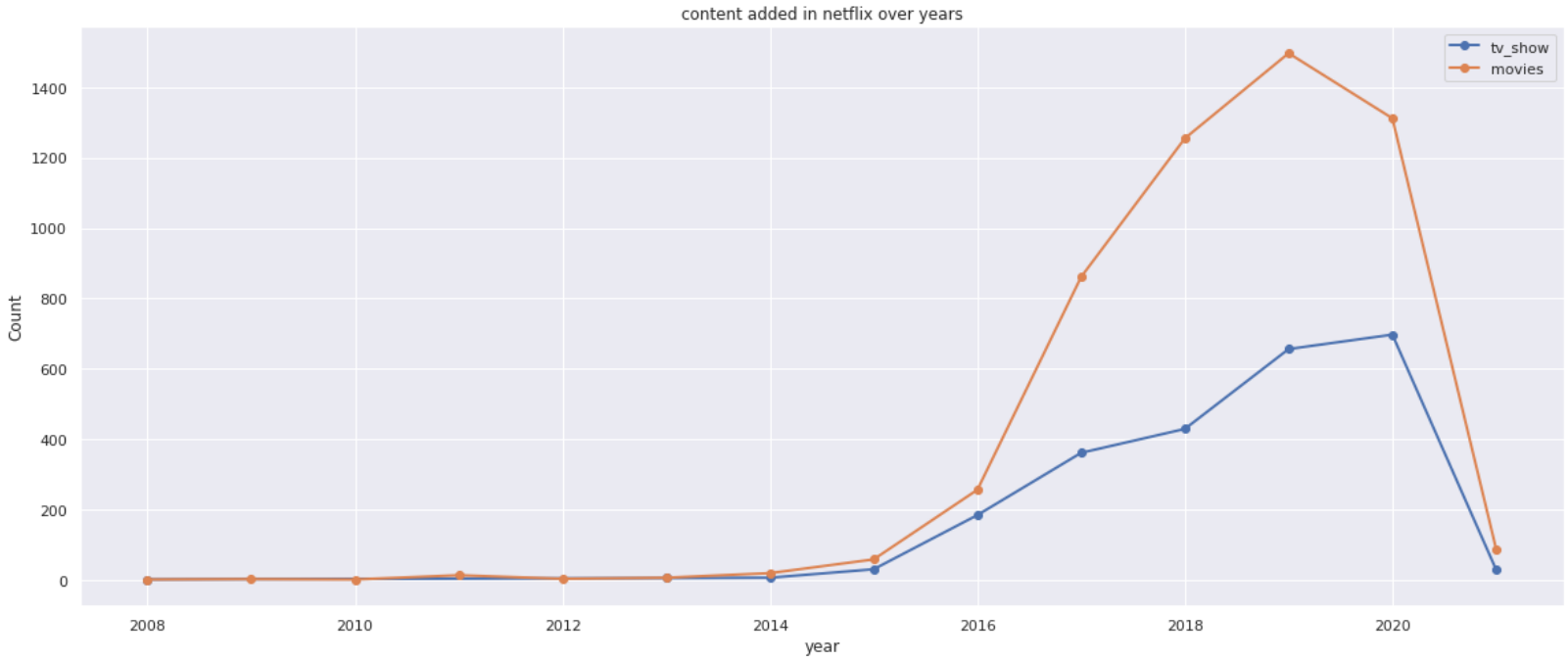
- show_id : Unique ID for every Movie / Tv Show
- type : Identifier - A Movie or TV Show
- title : Title of the Movie / Tv Show
- director : Director of the Movie
- cast : Actors involved in the movie / show
- country : Country where the movie / show was produced
- date_added : Date it was added on Netflix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

# Data Analysis and understand the relationships between variables

It can be seen that

- Movie are with 5372 count which is 69.1%
- TV Show are with 2398 count which is 30.9%

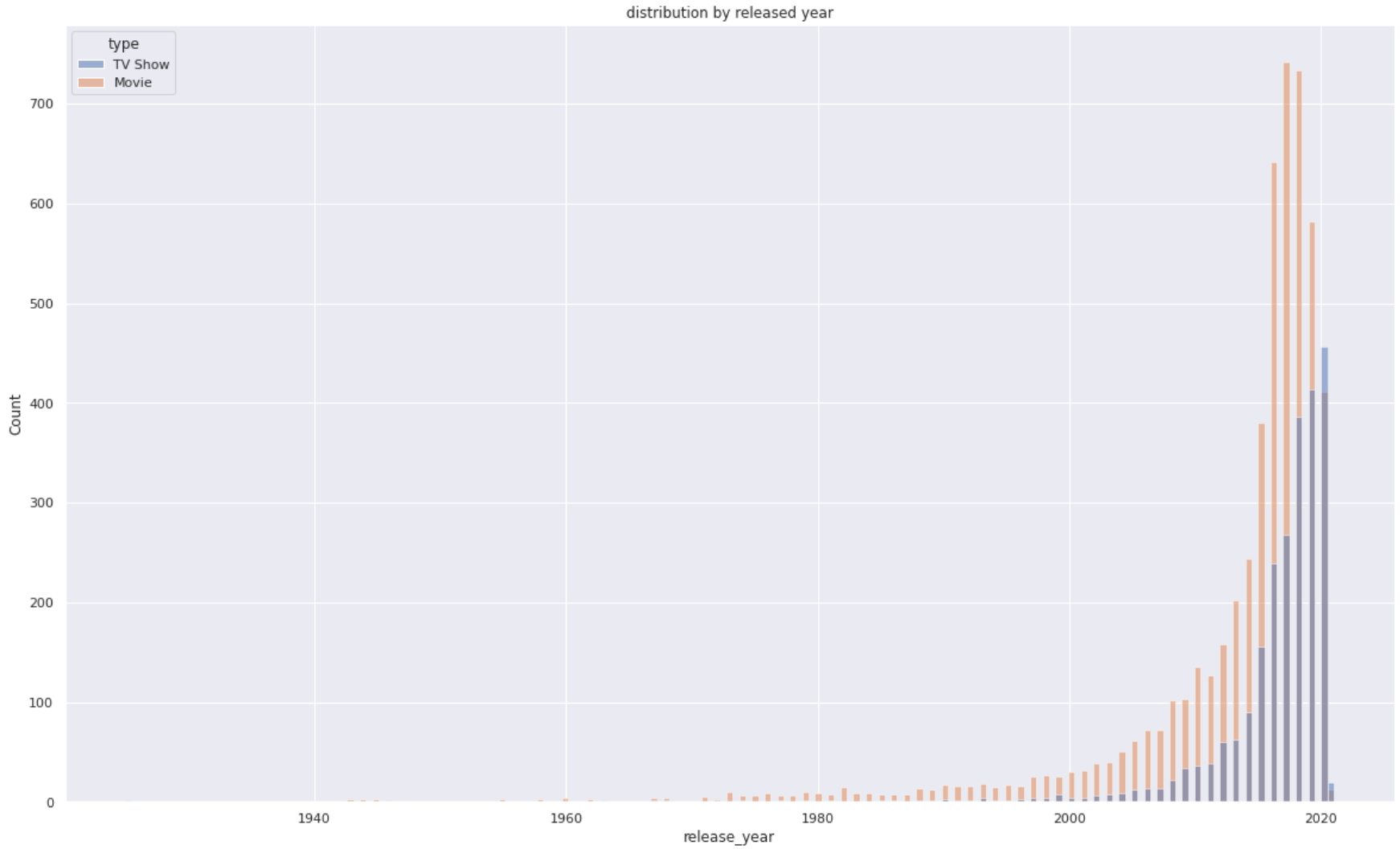



- Movies added in netflix are more compared to tv shows.
- And year 2019 had large amount of content added.

# Data Analysis

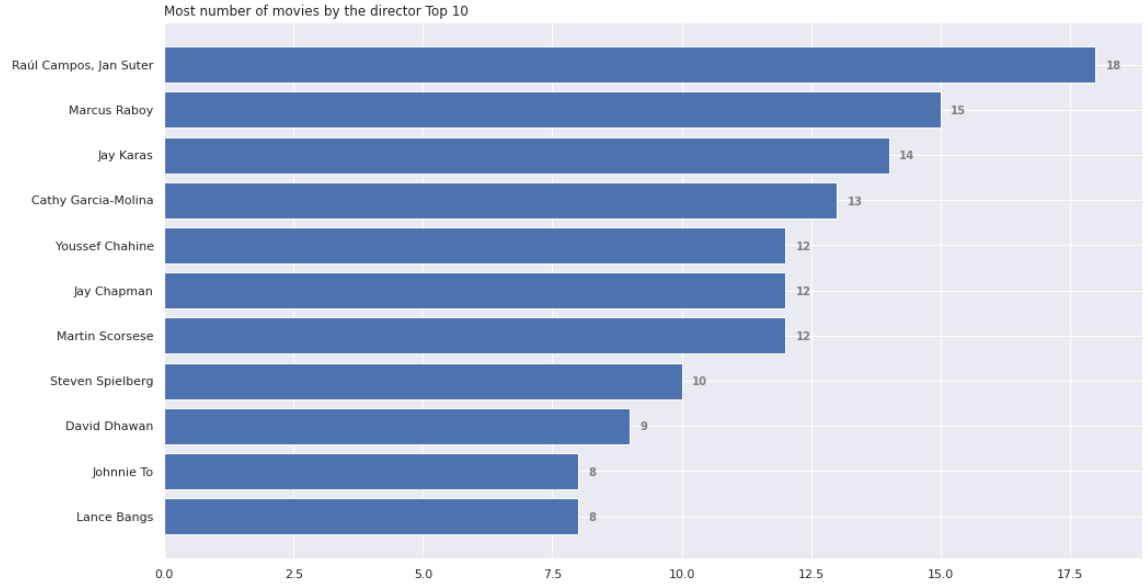
- Movies released over years are more compared to tv shows.
- Early 2000s had good start for increase of tv show production

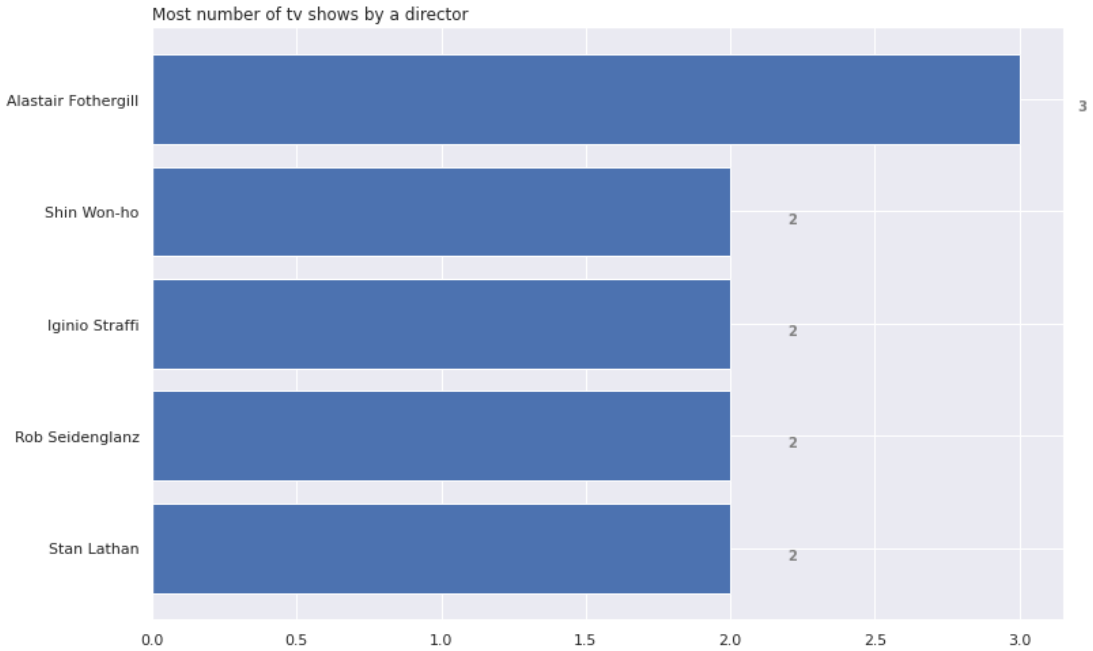


# Data Analysis

**Top 5 movie making Directors are:**

- Raúl Campos, Jan Suter with 18
- Marcus Raboy with 15
- Jay Karas with 14
- Cathy Garcia-Molina with 13
- Youssef Chahine with 12




**Top 5 Tvshow making Directors are:**

- Alastair Fothergill with 3
- Shin Won-ho with 2
- Iginio Straffi with 2
- Rob Seidenglanz with 2
- Stan Lathan with 2

# Data Analysis

Bar chart to define the top among the list

**Top 5 movie making Countries are**

1)United States with 1847

2)India with 852

3)United Kingdom with 193

4)Canada with 118

5)Egypt with 89

**Top 5 TV show making Countries are**

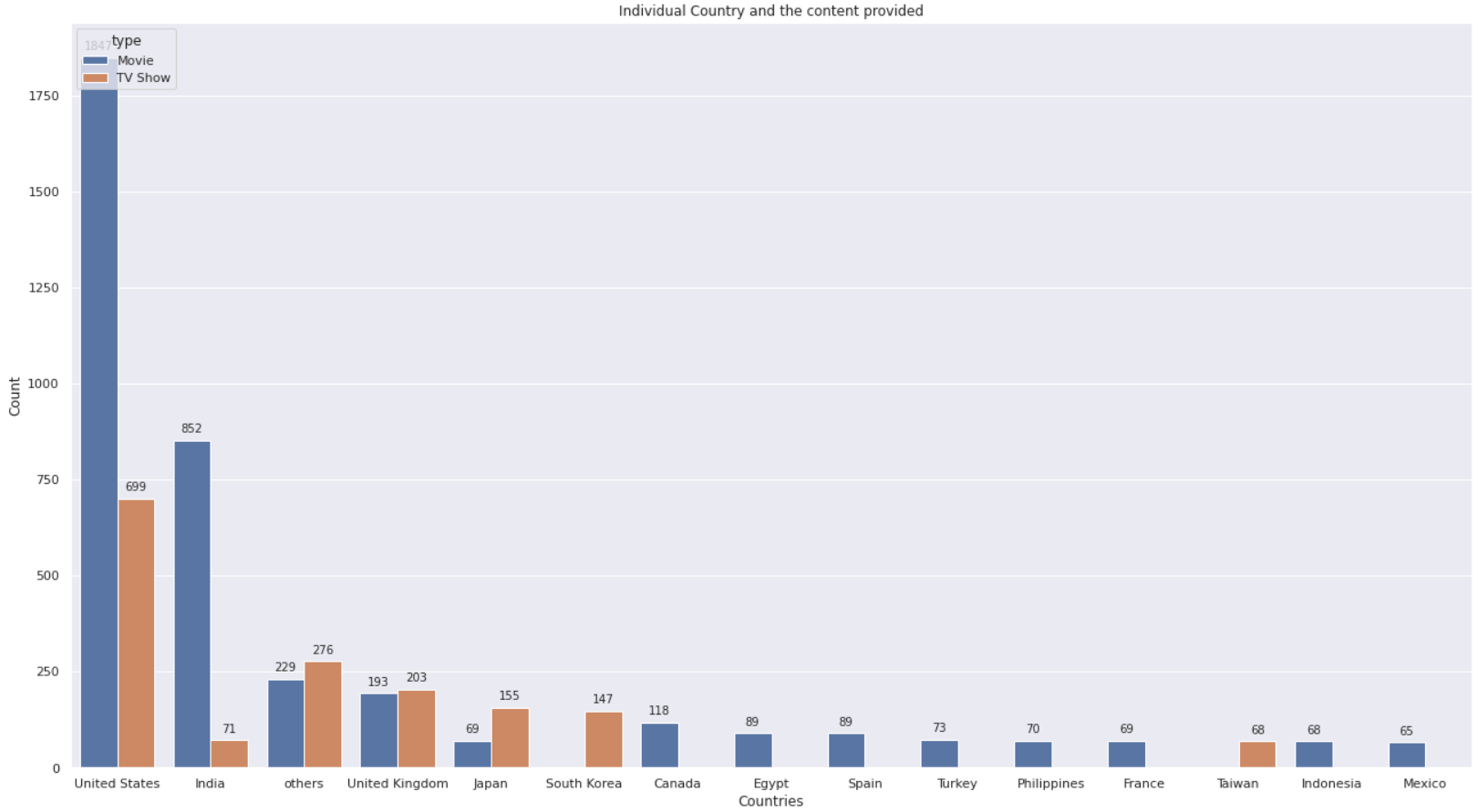
1)United States with 699

2)United Kingdom with 203

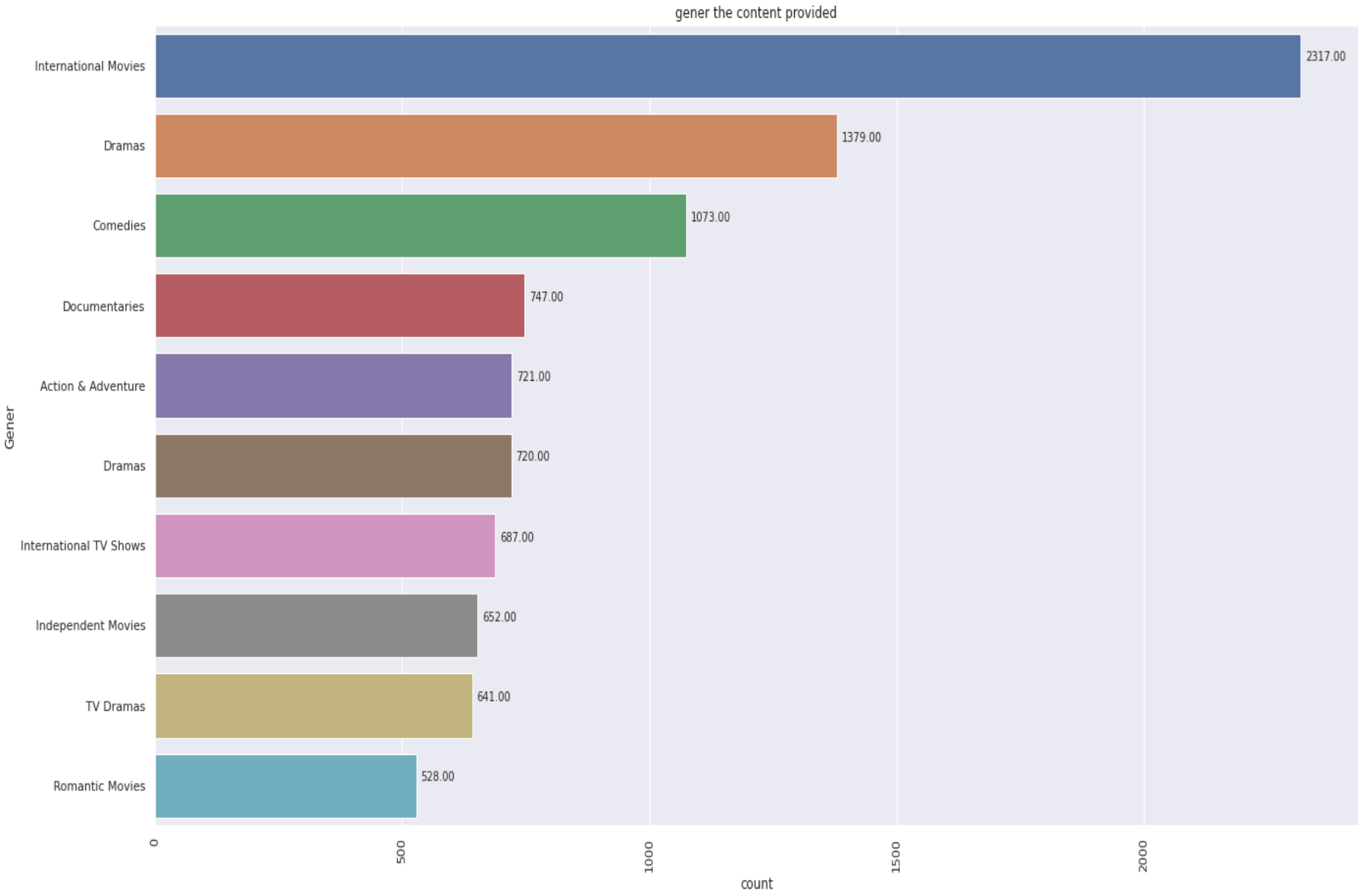
3)Japan with 155

4)South Korea with 147

5)India with 71



# Data Analysis




**Top 5 Geners in netflix are**

1) International Movies 2317
2) Dramas 1379
3) Comedies 1073
4) Documentaries 747
5) Action & Adventure 721

# Data Analysis

**Top 5 most appeared actor are**

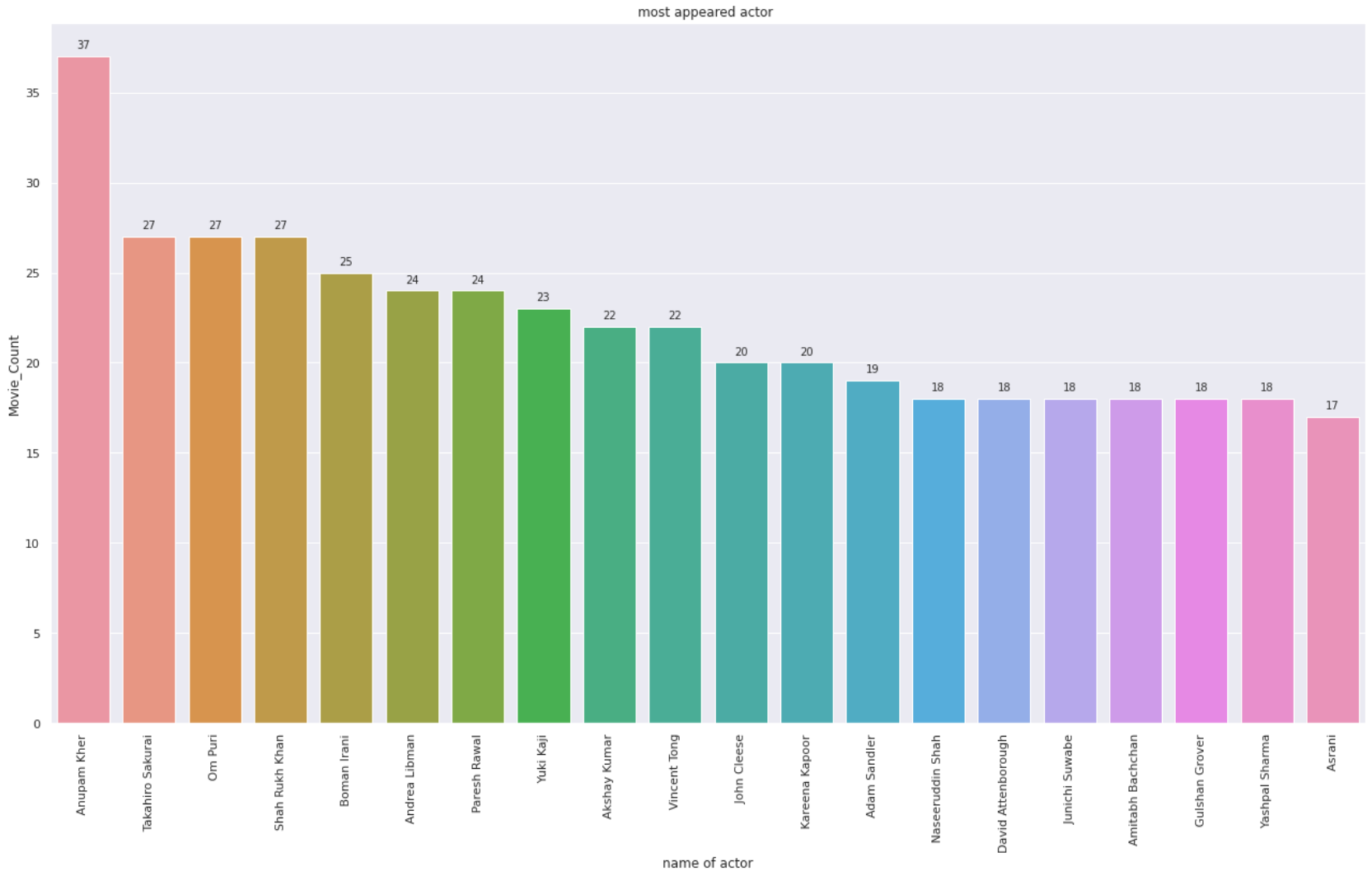
1)Anupam Kher 37

2) Takahiro Sakurai 27

3) Om Puri 27

4) Shah Rukh Khan 27

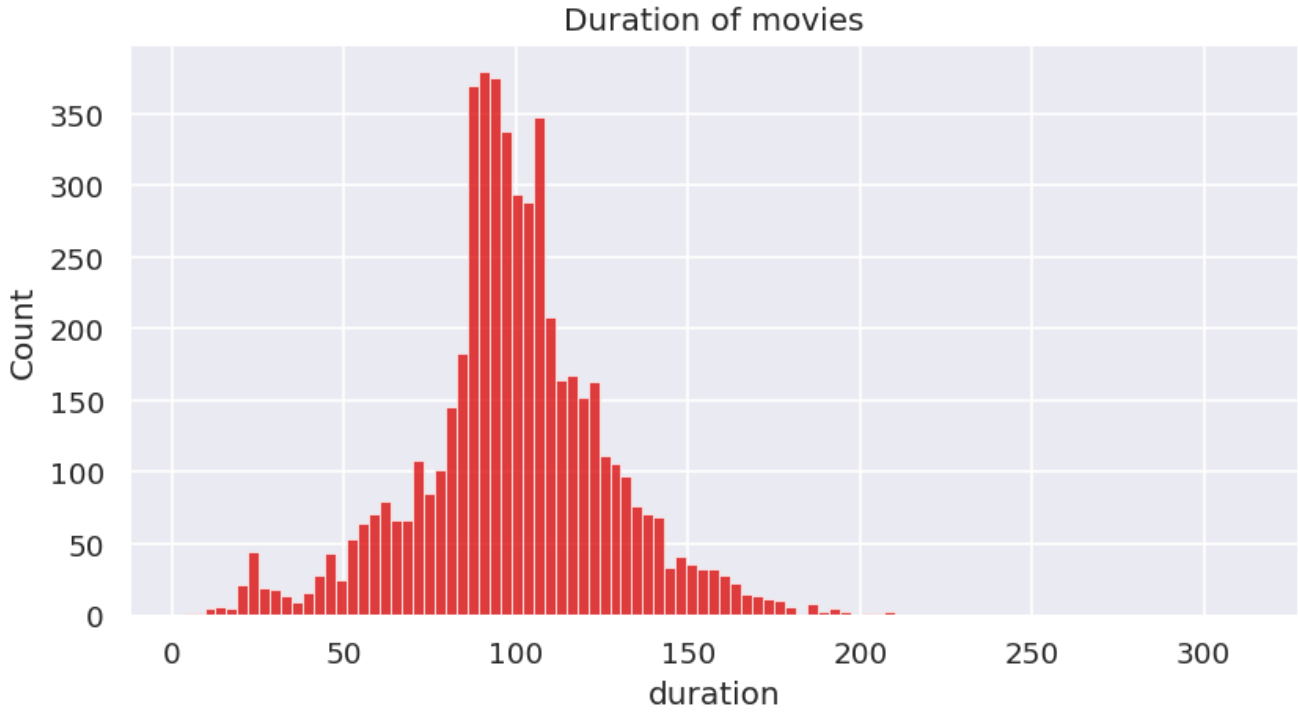
5) Boman Irani 25

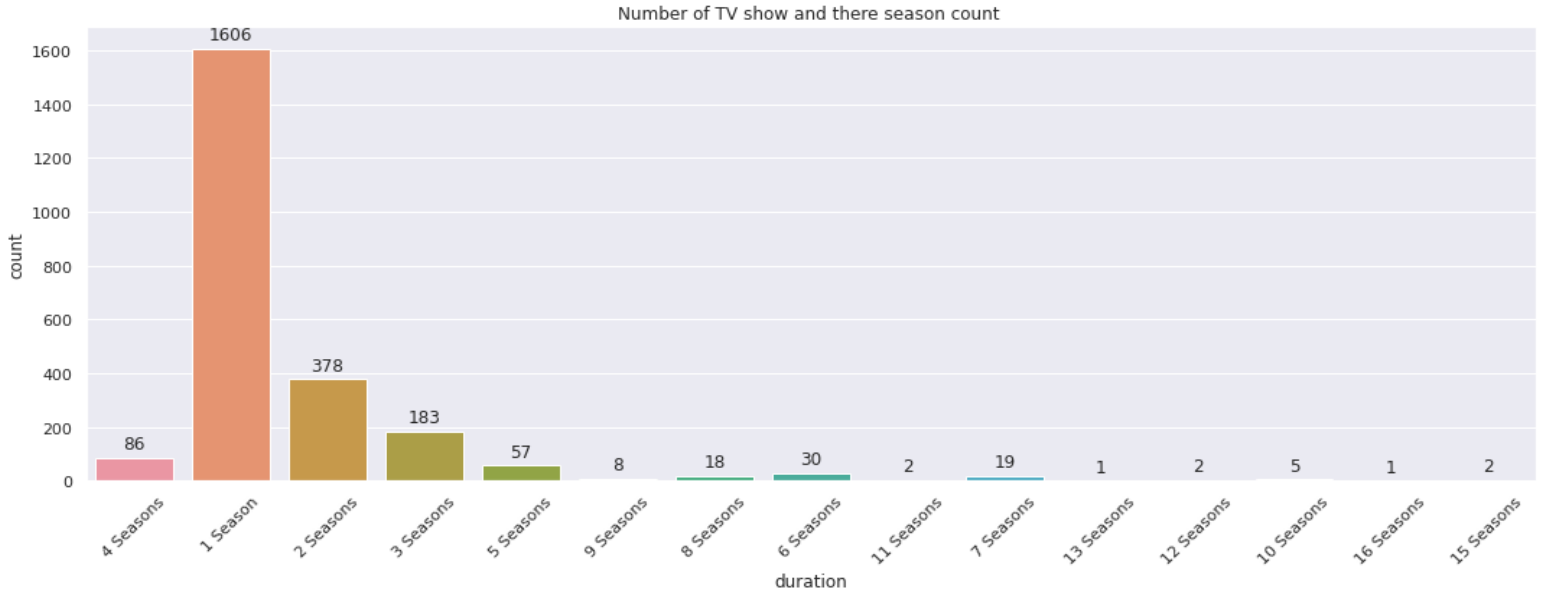


# Data Analysis

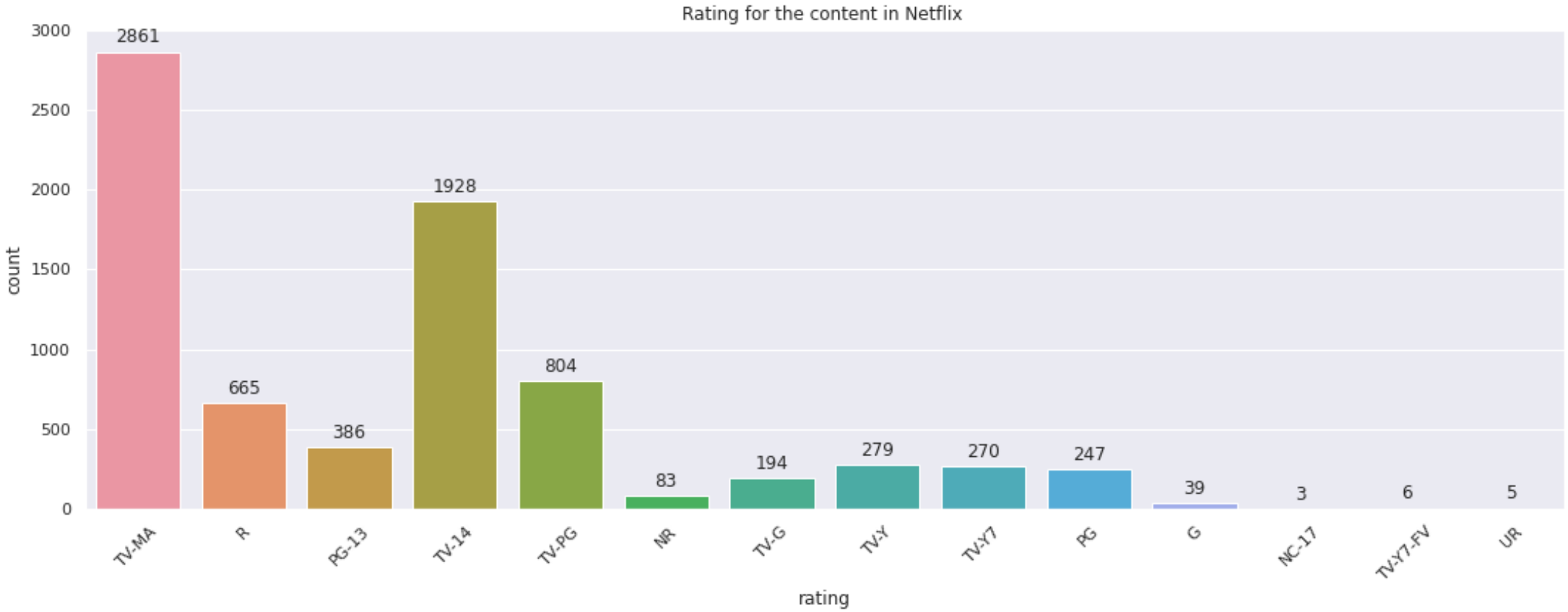
1)Movie Duration is between 10 to 200,where as most of the movies are in between 60 to 150 minutes.

2)Tv shows with only season 1 are most, following season 2 and 3.Very few tv shows ran for longer seasons



# Data Analysis :

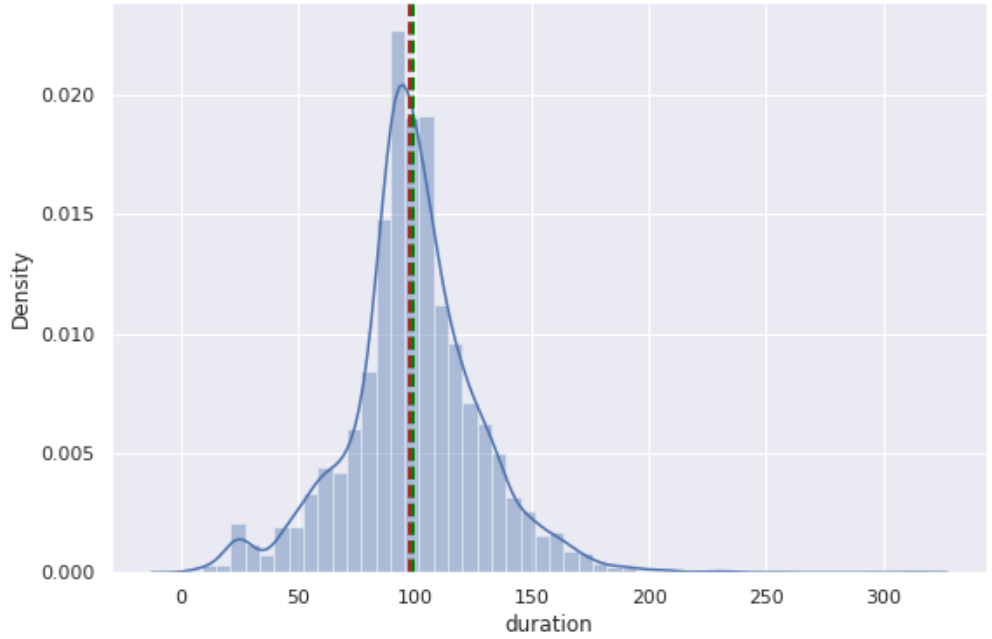



**Observation:**

Content with rating TV-MA are highest with 2861 count, following TV-14 with 1928 and TV-PG with 804 whereas content with rating UR is 5 and G is 3 recorded low count.

# Hypothesis Testing

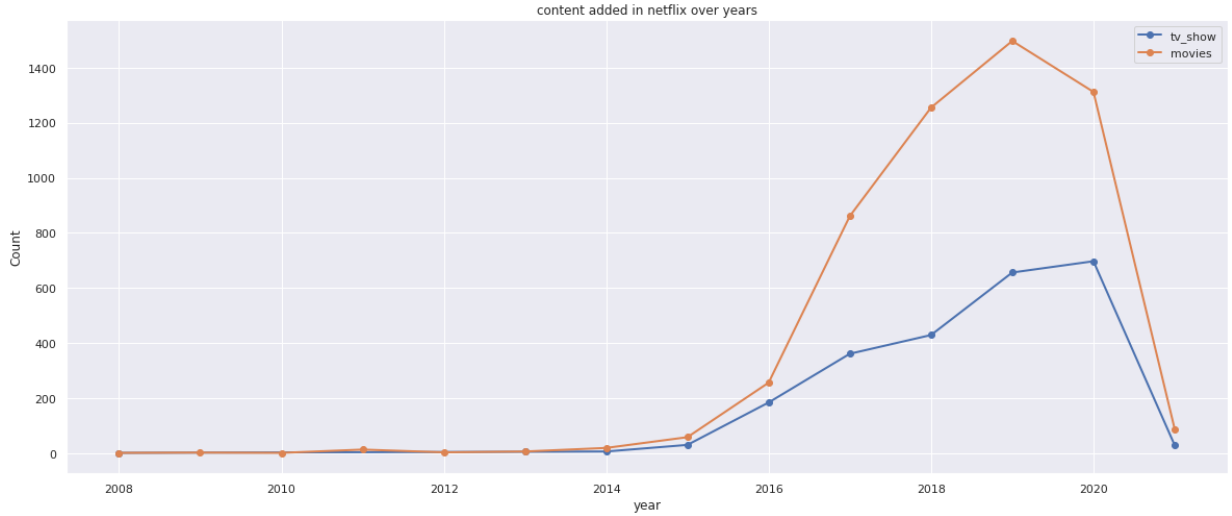
Null Hypothesis = Average movie duration is greater than 90 Minutes
Alternative Hypothesis = Average movie duration not the given values.




Since our data is normally distributed, Right tailed Z-test

Failed to reject the Null Hypothesis for p = 2.84356613164316 7 . .

Null Hypothesis = year added has no impact on type of content that gets added to the platform.
Alternative Hypothesis = year added has impact on type of content that gets added to the platform.




Chi-Squared testing is used when testing statistical independence or associaton between categorical variables

P-value: 7.47833695275089 9e-11 Reject the null hypothesis,and thereis a relatiionship between year added and type of content

# Implementation

## Textual Data Pre-processing:

- Removing space
- Removing Punctuations and stop words
- Stemming/lemmatization
- Vectorization
- PCA

## Model Implementation

- Since our project is unsupervised, I used :
    a. Kmeans clustering,
    b. Hierarchical clustering
    c. DBSCAN

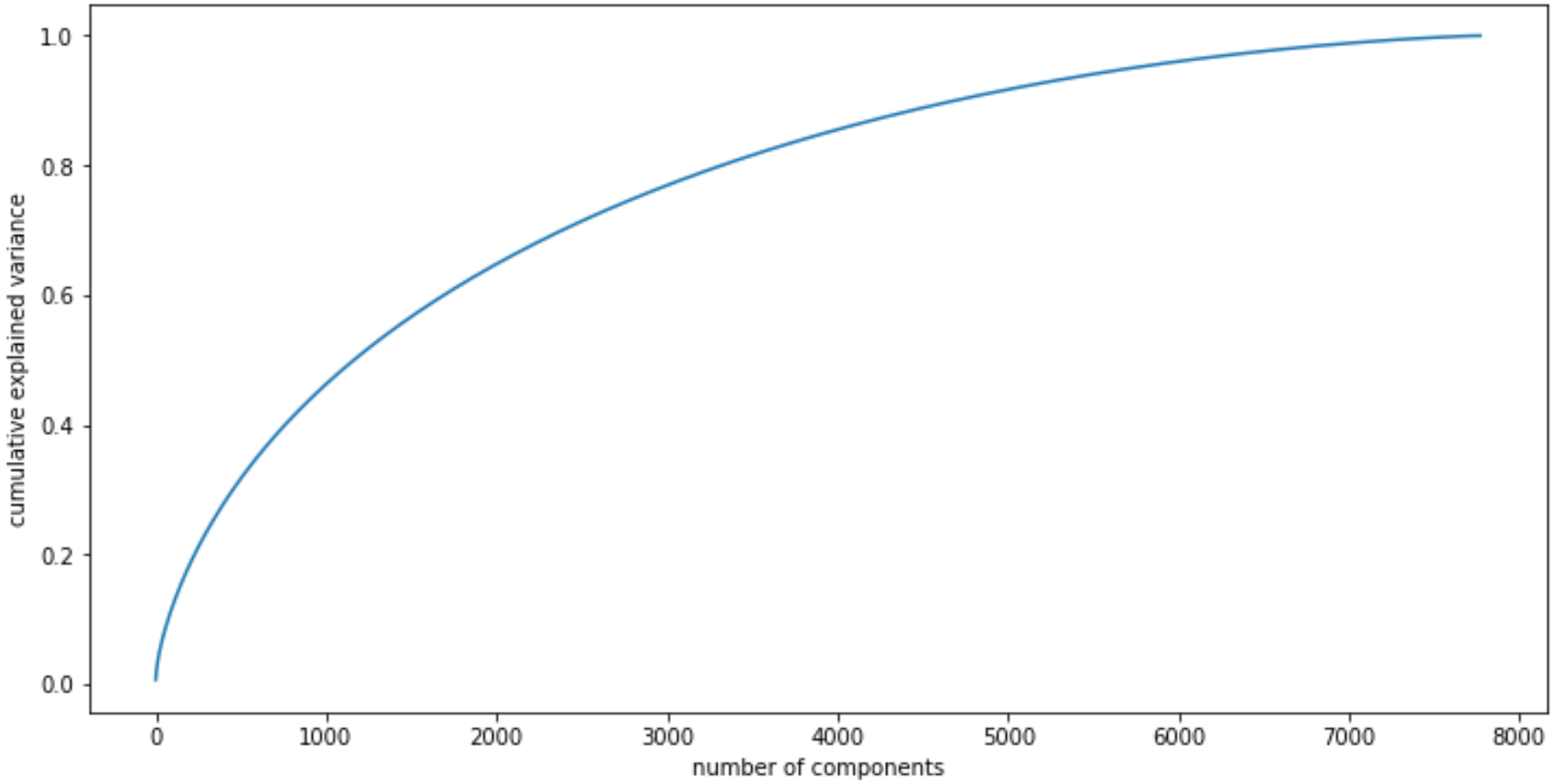
## Model Evaluation:

- silhouette score.

# Implementation

## PCA **Principal Component Analysis**

PCA is an unsupervised statistical technique used to examine the interrelations among a set of variables



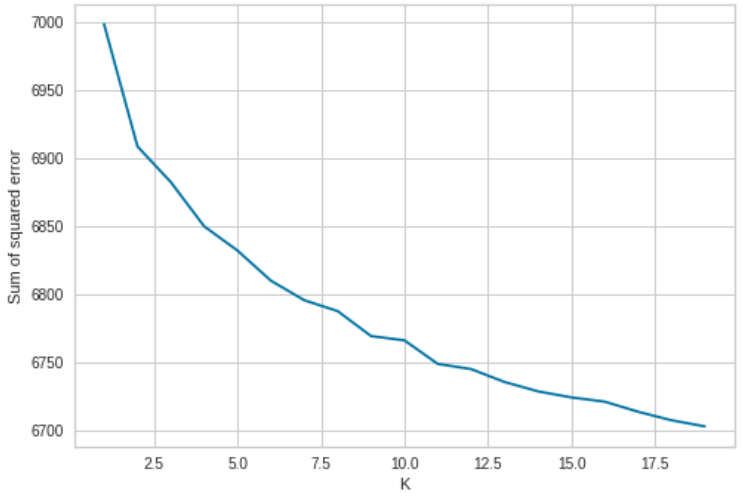

- PCA for reducing dimension for lesser variable with variance maximizing.
- Here I maximized the overall variance to 90%
- Which was obtained by 5000 features, That is data with 20000 features are reduced to 5000.

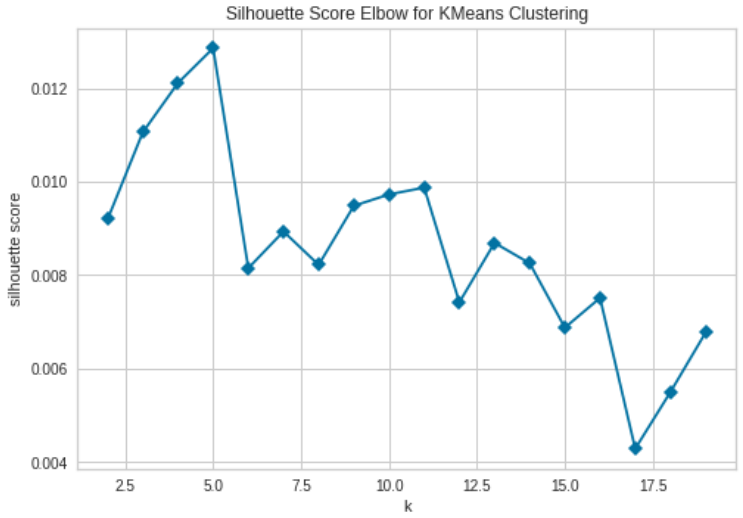
# Implementation

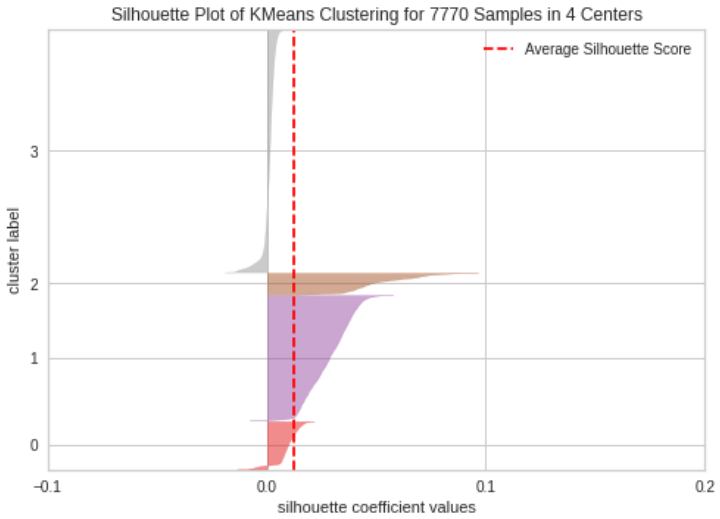
## Model 1 K Means clustering

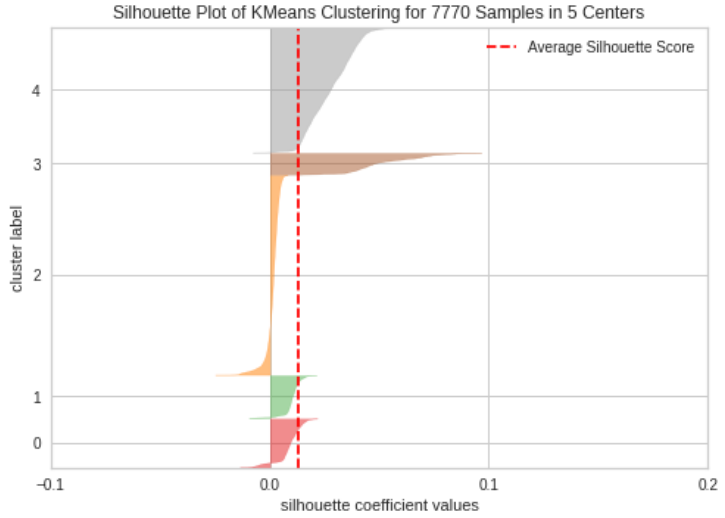
- Elbow method is used to find the optimum number of clusters

From the plot it is seen that either 4 or 5 are good for clustering





For n_clusters = 4, silhouette score is 0.0121151167167034 56

For n_clusters = 5, silhouette score is 0.0128598854864752 52

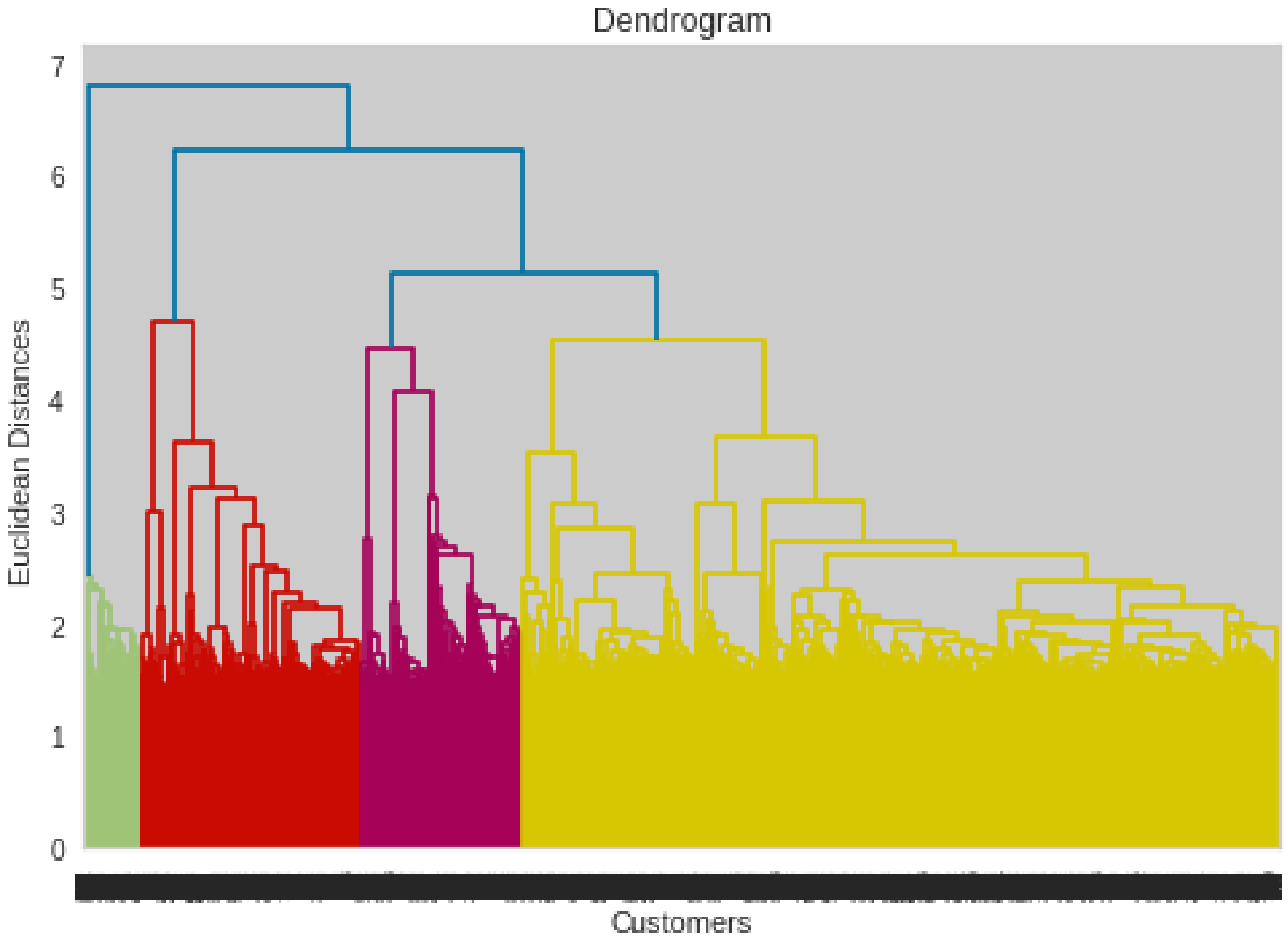
For n_clusters = 6, silhouette score is 0.0081509795384251 58

## Model 2 Hierarchical clustering

Dendrogram to find the optimum number of clusters , More the distance of the vertical lines in the dendrogram, the more the distance between those clusters

AgglomerativeClustering,with wards method to make 4 clusters is implemented .

For clusters = 4,
silhouette score is
0.0027066154436784 68

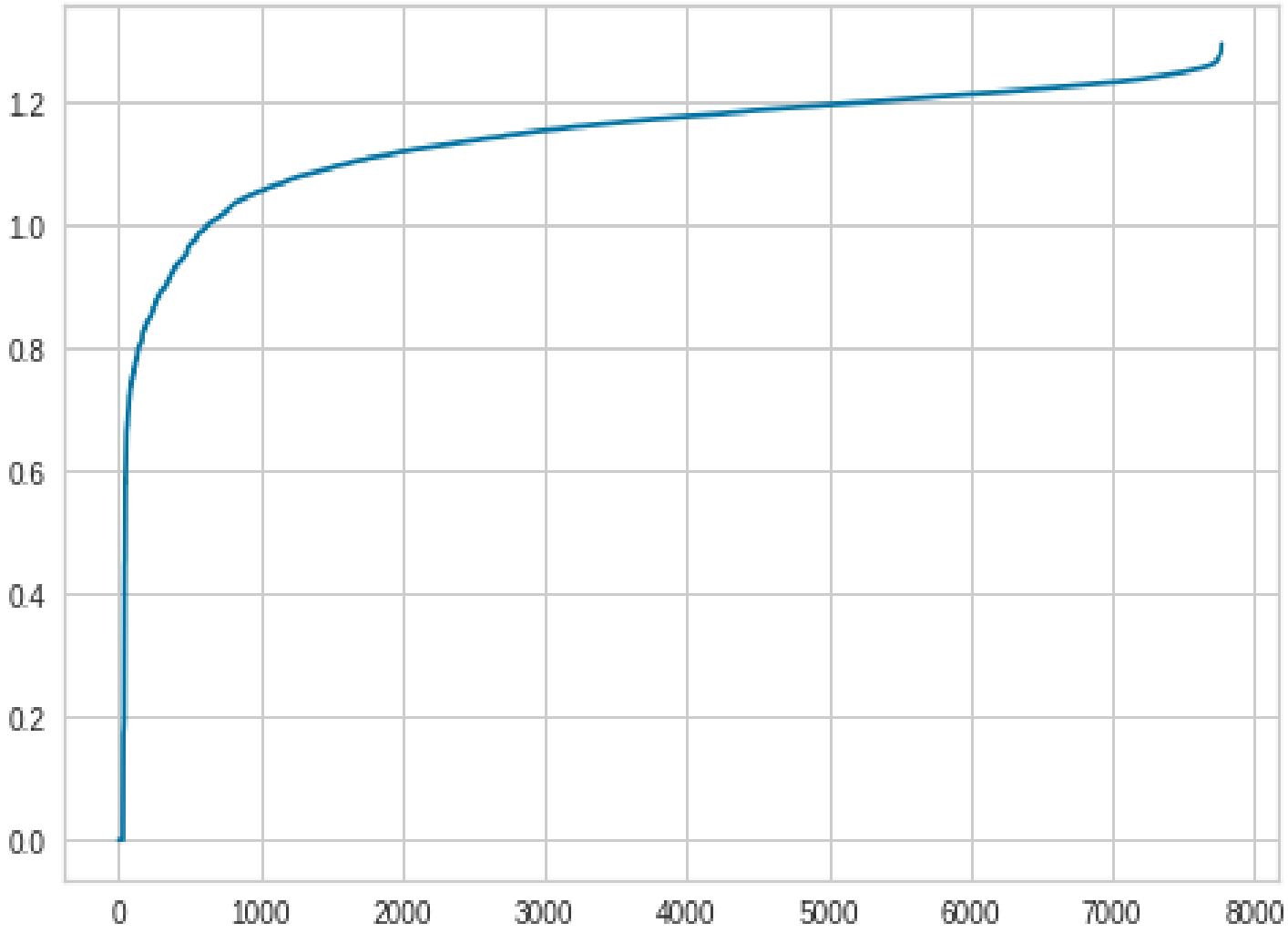


## Model 3 DBSCAN

**Density-based spatial clustering of applications with noise**
**Parameter:**

**eps** : It defines the neighborhood around a data point, One way to find the eps value is based on the ***k-distance graph. Here*** it found to be as 0.9

And it gave 4 clusters and silhouette score is – 0.004455436295342932 5



# Recommendations:

I used cosine similarity to generate recommendation system for a given movie.
For example:
if movie with title 9 is given as input then the recommendation are as following:

1. Cirque du Freak: The Vampire's Assistant
2. The River Wild
3. Real Steel
4. The Lord of the Rings: The Return of the King
5. Extinction
6. Small Soldie`rs
7. The Dark Crystal: Age of Resistance
8. The Book of Eli
9. Tremors 2: Aftershocks
10. The Imaginarium of Doctor Parnassus

# Conclusion.

- Majority of the content available on Netflix is Movies and amount of shows added on Netflix is growing exponentially.
- United States and India are among the top 5 countries that produce majority of the available content on the platform. But United States took the NO.1 spot in both movie making and tv show.
- Movie making directors are larger than TV Shows.
- Also 6 of the actors among the top ten actors with maximum content are from India.
- TV-MA tops the charts, indicating that mature content is more popular on Netflix.
- Kmeans clustering was doing better than other models,even tough score is week but K= 4 or 5 was found to be an optimal value for clusters
- Recommender system was created using cosine similarity and top 10 recommendations for Movies and Tv Shows were obtained

## References :-


1. mygreatlearning.com
2. GeeksforGeeks
3. Analytics Vidhya
4. Almabetter Notes.