

NETFLIX

Capstone Project

NETFLIX MOVIES AND TV SHOWS CLUSTERING

Submitted By

Poovarasan
shipfriend0368@gmail.com

Contents:

- Introduction.
- Problem statement.
- Data description.
- Exploratory data analysis.
- Data preprocessing.
- Modeling of Clusters.
- Data represented by each cluster.
- Recommendations System.
- Conclusions.



Introduction:

Netflix is a media distribution company. It started with DVD distribution via mail, but has evolved substantially over the course of its existence. Today, 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 today television shows are probably the more common format. Netflix works on a subscription model, where users get unlimited access to content with a paid subscription.



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.



Data description:

The dataset consists of listings of all the movies and tv shows available on Netflix, along with details such as - cast, directors, ratings, release year, duration, etc.

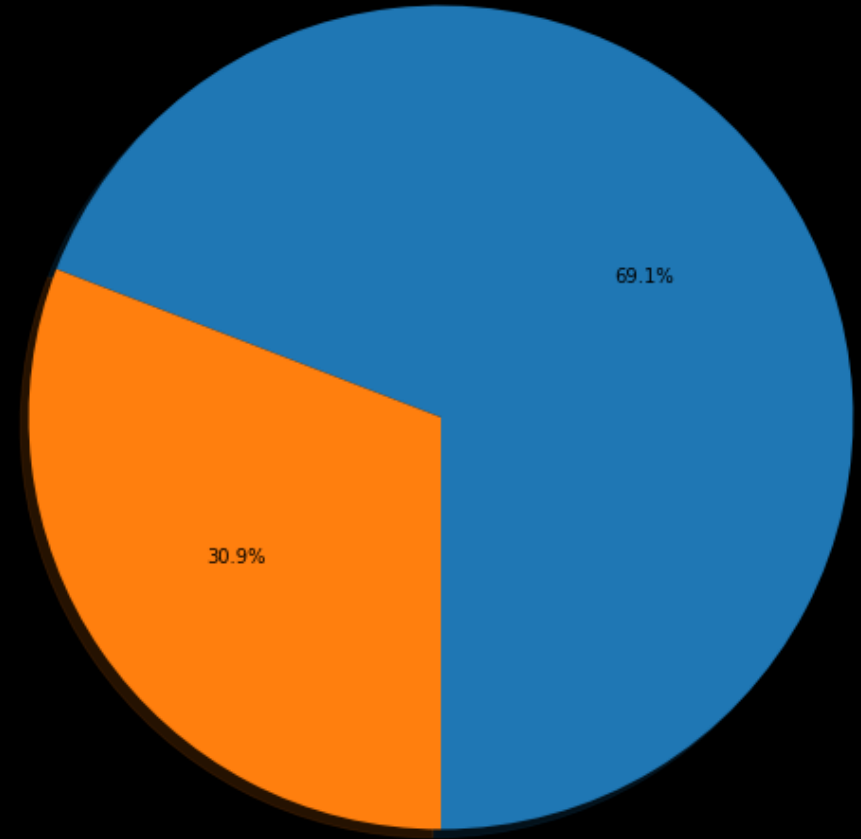
- **show_id** : Unique ID for every Movie / Tv Show
- **type** : Identifier - A Movie or TV Show
- **title** : Title of the Movie / Tv Show
- **director** : Director of the Movie
- **cast** : Actors involved in the movie / show
- **country** : Country where the movie / show was produced
- **date_added**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



Exploratory Data Analysis:

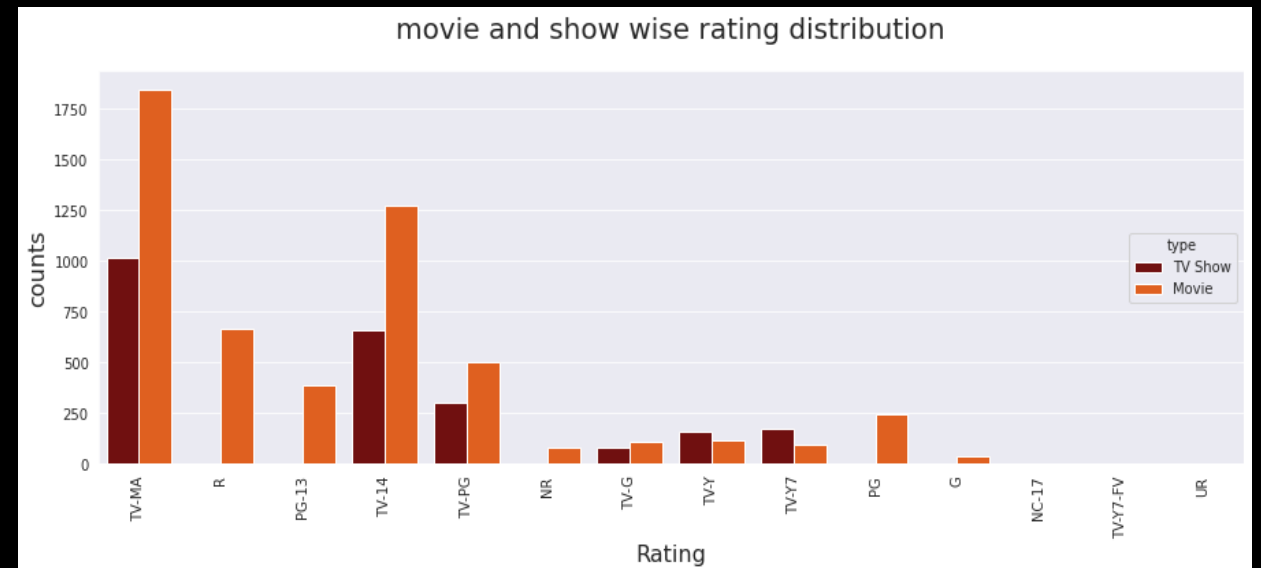
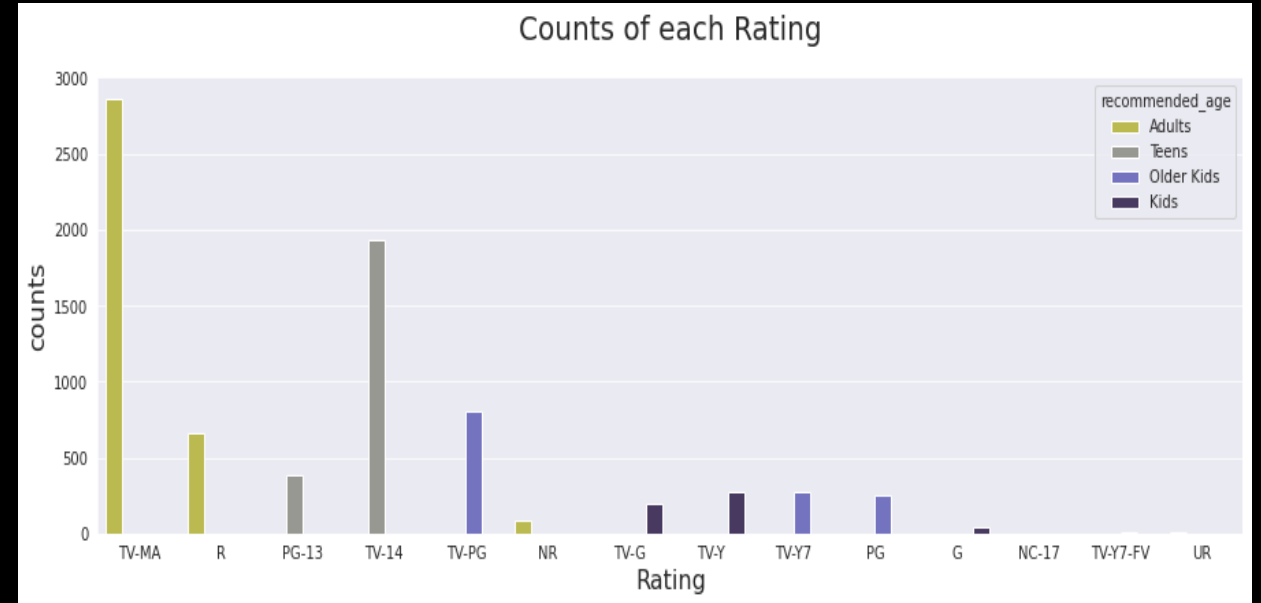
Type

- 69.1% of the content available on Netflix are movies; the remaining 30.9% are TV Shows.
- Netflix has 5372 movies and 2398 TV shows there are more number movies on Netflix than TV shows.

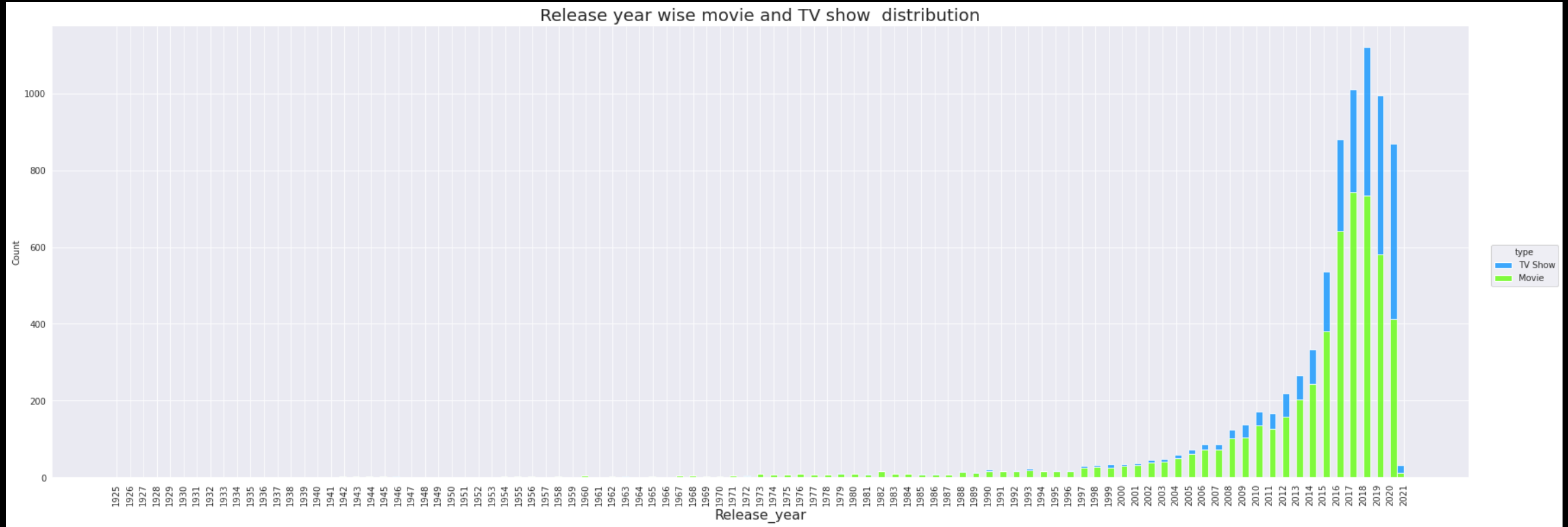


Ratings

- TV-MA and TV-14 has the highest number of ratings for tv shows, its rerecommended for Adults and Teens Respectively
- TV-MA and TV-14 has the highest number of ratings for movies , its rerecommended for Adults and Teens Respectively
- in both the cases TV-MA has the highest number of ratings



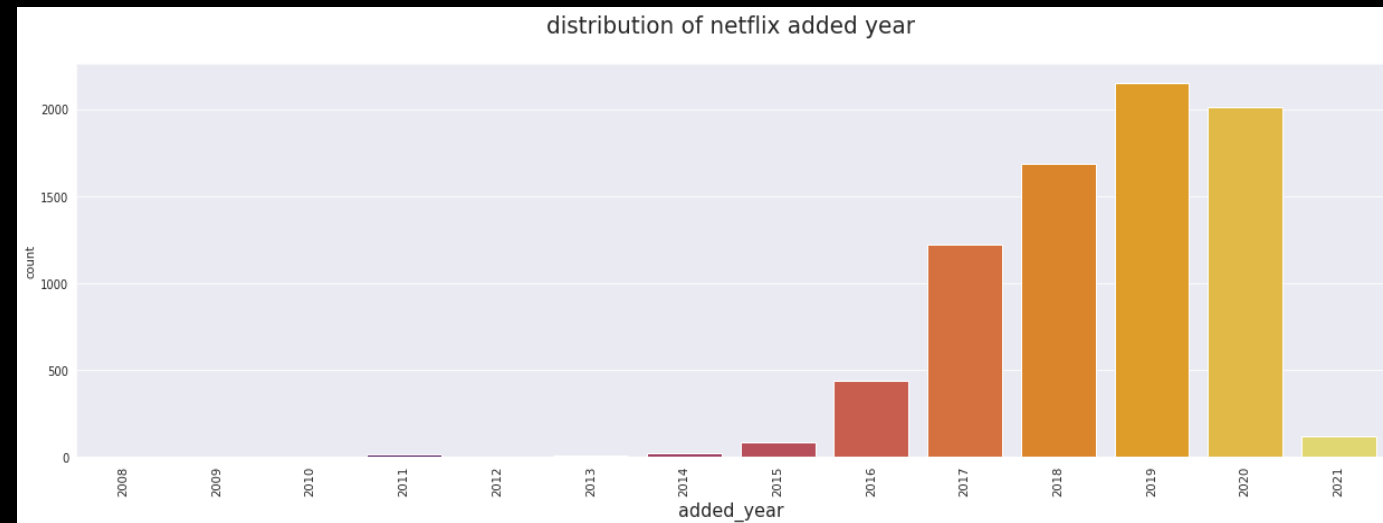
Release Year



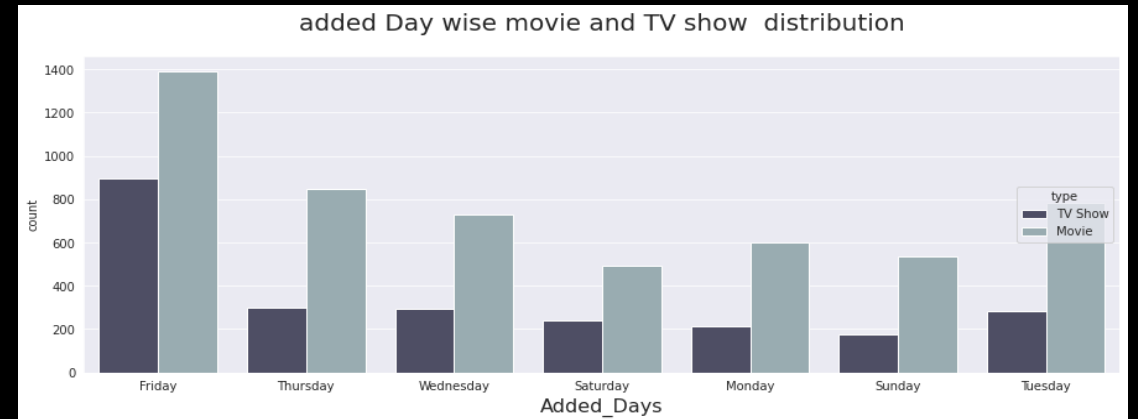
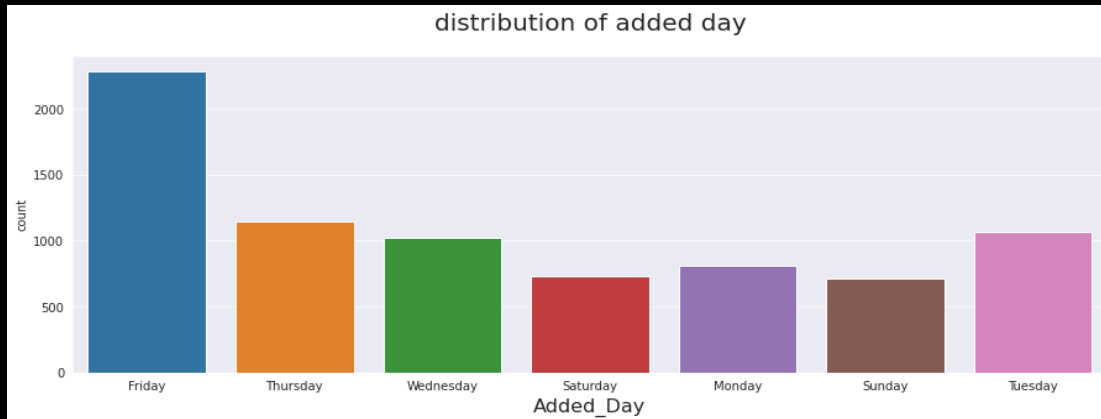
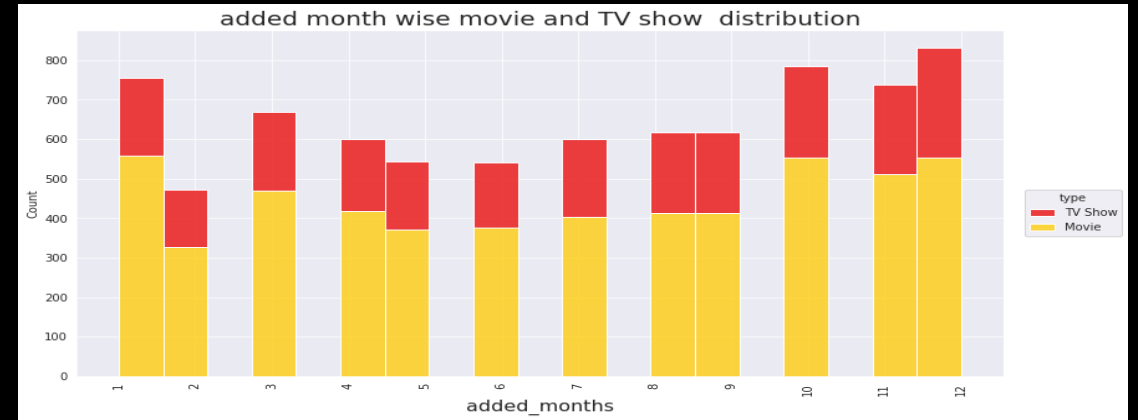
- Based on above analysis from 2000 to 2018 movies and shows are have highly released but in 2019 to 2021 its slowly get decrease may be its happened because of COVID

Year added

- Most films were released in 2017 and 2018.
- 2020 will see the most film releases.
- Netflix's movie collection is expanding much more quickly than its tv show collection.
- After 2015, we noticed a significant rise in the quantity of movies and television shows.
- After 2020, there is a drastic decrease in the volume of movies and television shows made.
- It looks that Netflix has chosen releasing more movie material over tv shows.
- The growth of movies has been far more dramatic compared to tv shows.

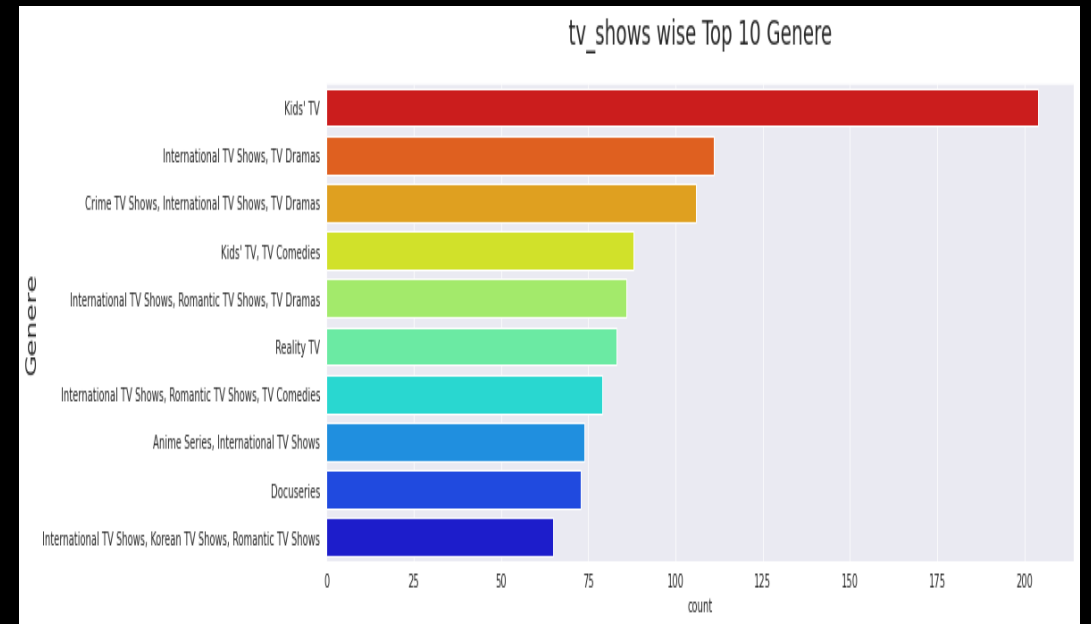
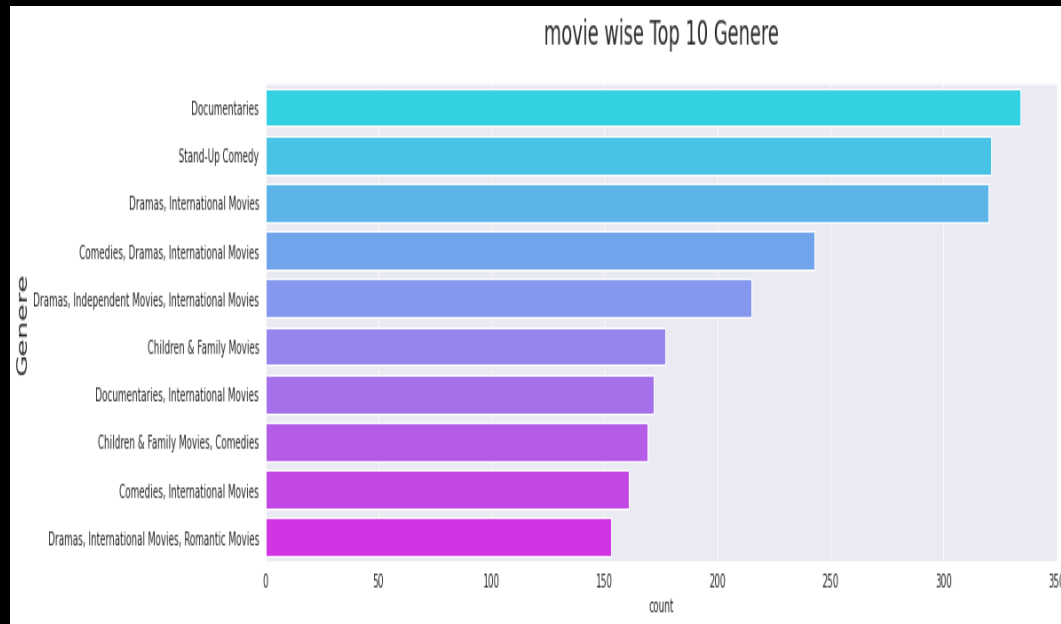


Month and Day added



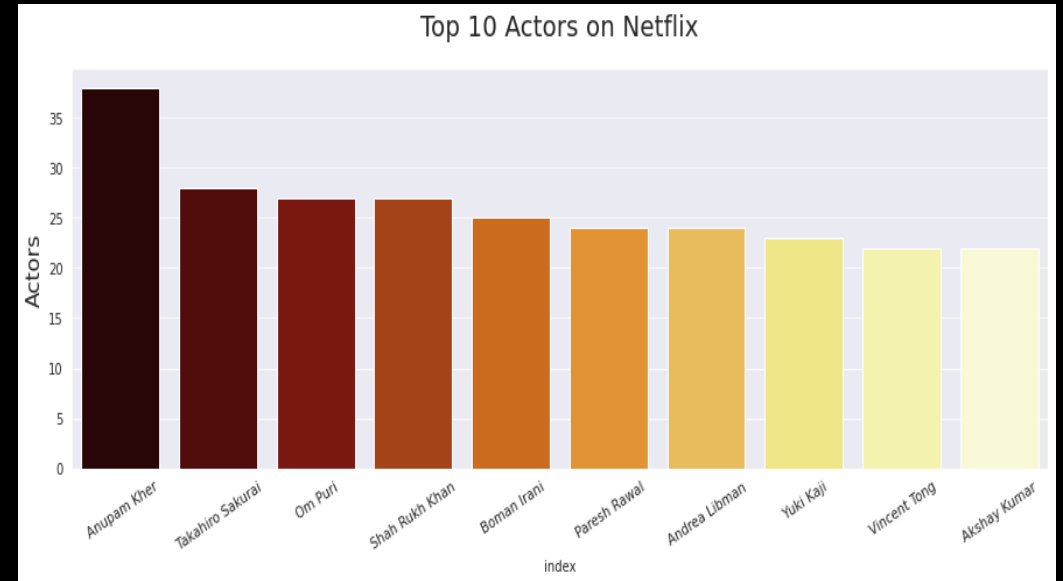
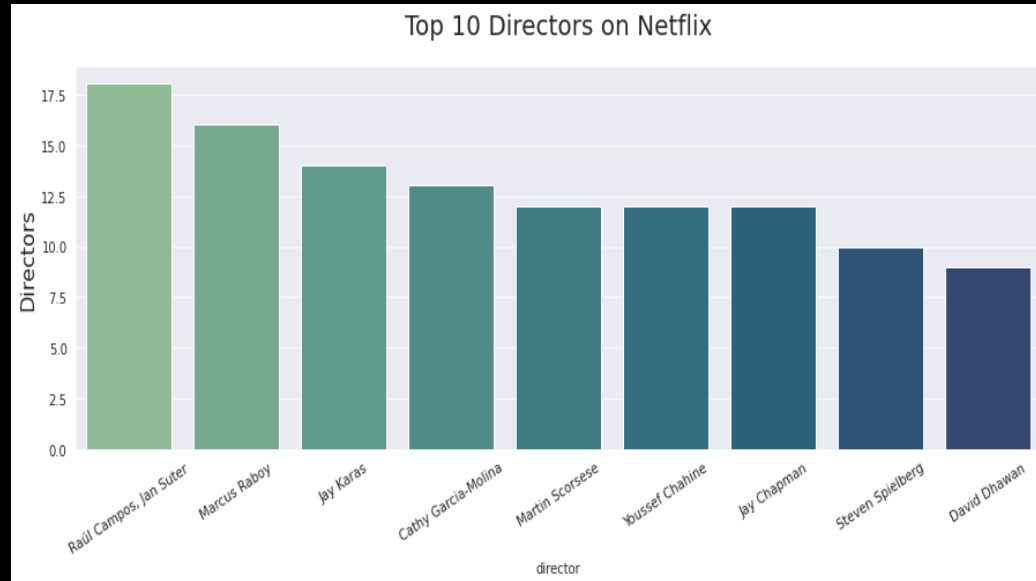
- Month wise analysis many of the movies and tv shows are released during year end and year starting It might be due to the winter, as in these months people may stay at home and watch shows and movies in their free time.
- Many of the Tv shows and Movies are released Friday, its may be Friday is week last day so many peoples can watch.

Genre



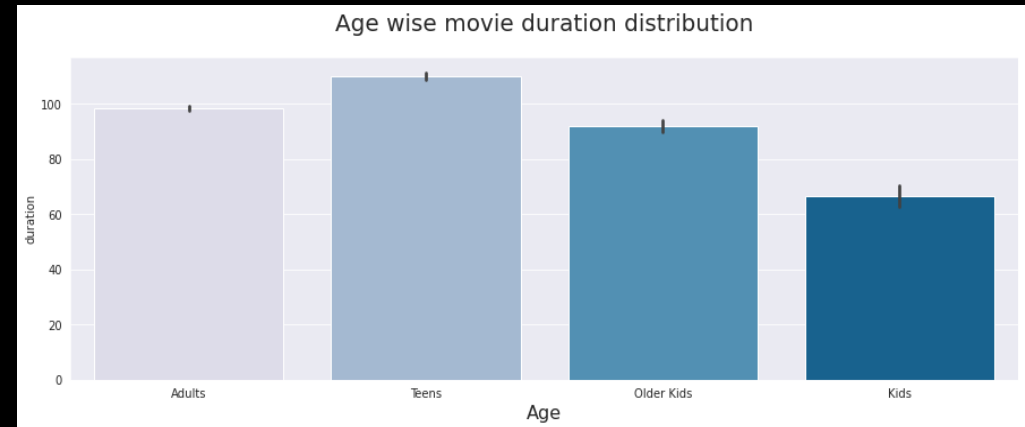
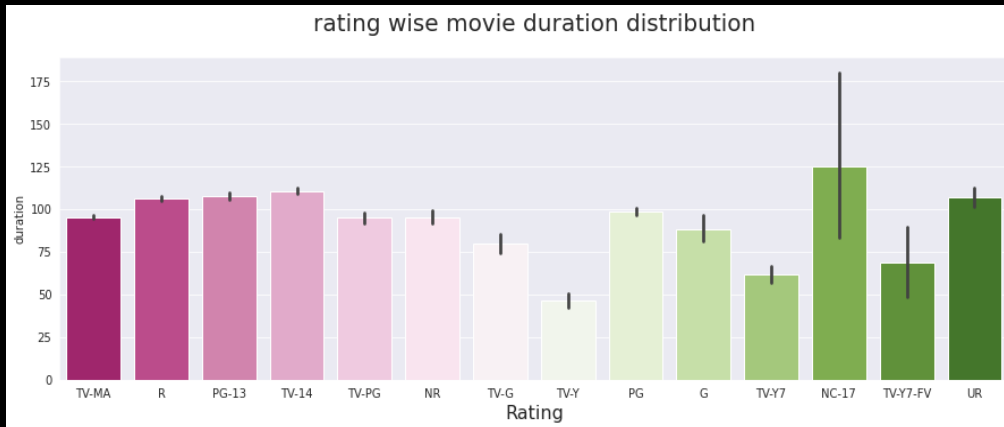
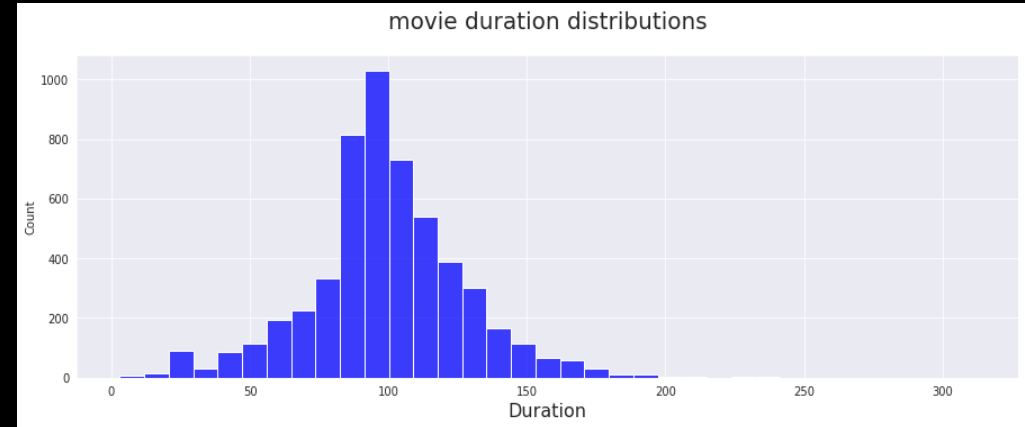
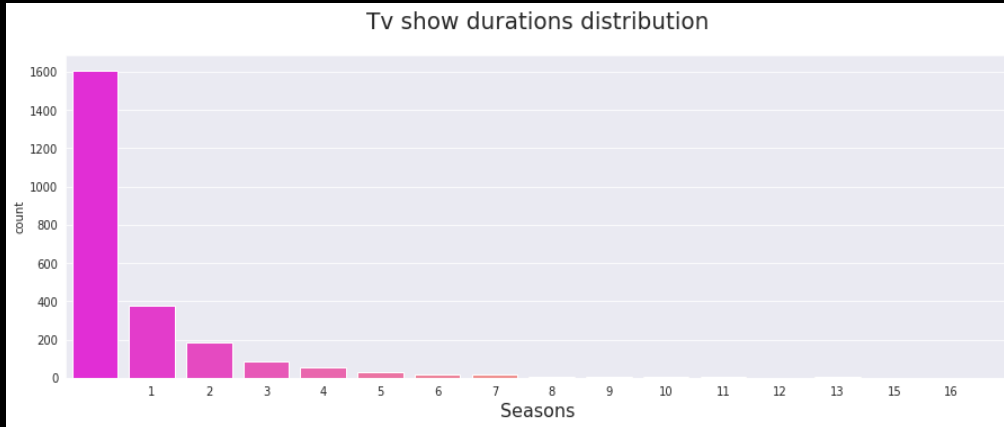
- The most popular Netflix movies category is documentaries, which are followed by standup comedy, Dramas, and foreign films.
- The most popular Netflix TV show genre is kids' television.

Director and Cast



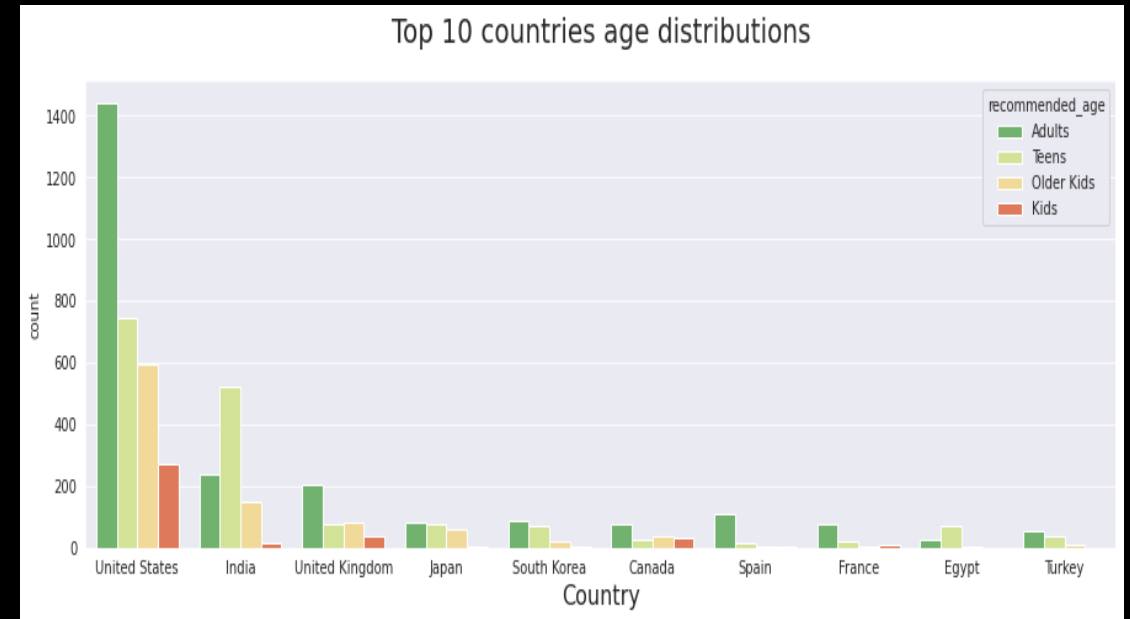
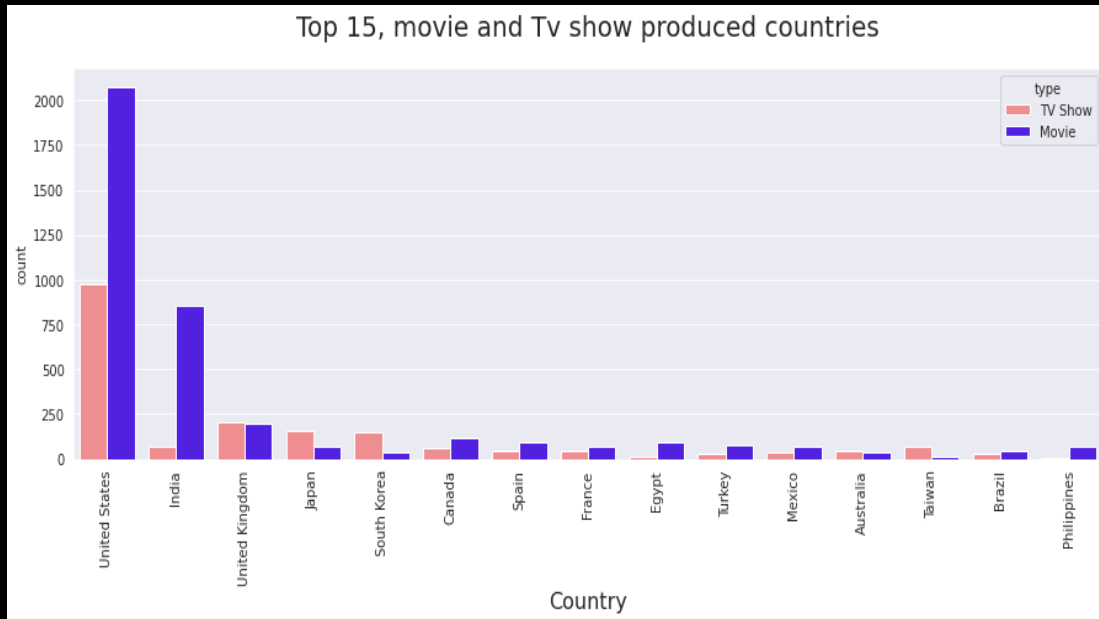
- Raul Campos, Jan Suter, Marcus Raboy, Jay Karas, Cathy Garcia-Molina, Jay Chapman are the top 5 directors with the most films and television productions, respectively.
- Six of the actors in the top ten list with most numbers tv shows and movies are from India.
- Anupam Kher captured first place in this list.

Duration



- The majority of movies run for 50 to 150 minutes and most television shows have only one season.
- The longest average is found among NC-17-rated movies.
- Teens and adults are spending more time on watching movies

Country



- Most of the Movies and Tv shows are produced in America and in India Compare Tv shows Movies are highly produced
- In USA the most produced Movies and shows are recommended for adults, but in India highly produced contents are recommended for teen age groups

Data Preprocessing:

● STEP 1

Select the require text variable from the datasets.



● STEP 2

Remove Punctuations and Stop words.

● STEP 3

Perform stemming to reduce words with similar meaning.

● STEP 4

Vectorize the text using TF-IDF vectorizer.

● STEP 5

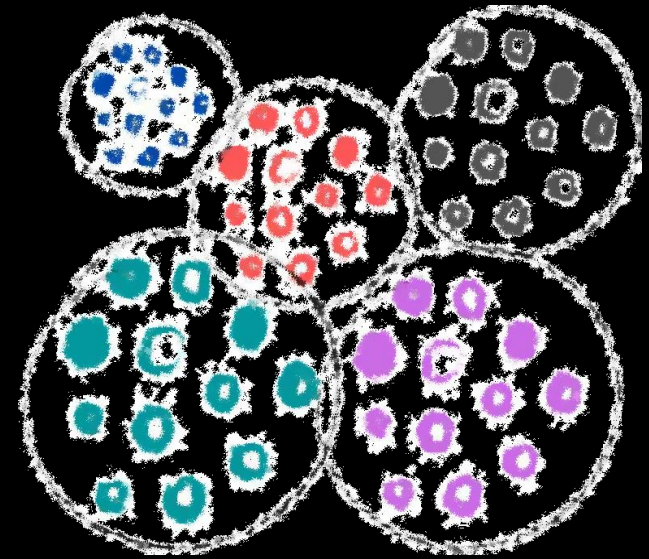
Perform PCA to reduce dimension's, select number of components that explain 99% of variance in data.



Modeling of Clusters:

What is clustering?

Clustering is the task of dividing the data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them.



K-means clustering

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. It is an iterative algorithm that divides the unlabeled dataset into k different clusters in such a way that each dataset belongs only one group that has similar properties. 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.

Agglomerative clustering

Agglomerative clustering is a type of hierarchical clustering algorithm. It is an unsupervised machine learning technique that divides the population into several clusters such that data points in the same cluster are more similar and data points in different clusters are dissimilar.

Birch clustering

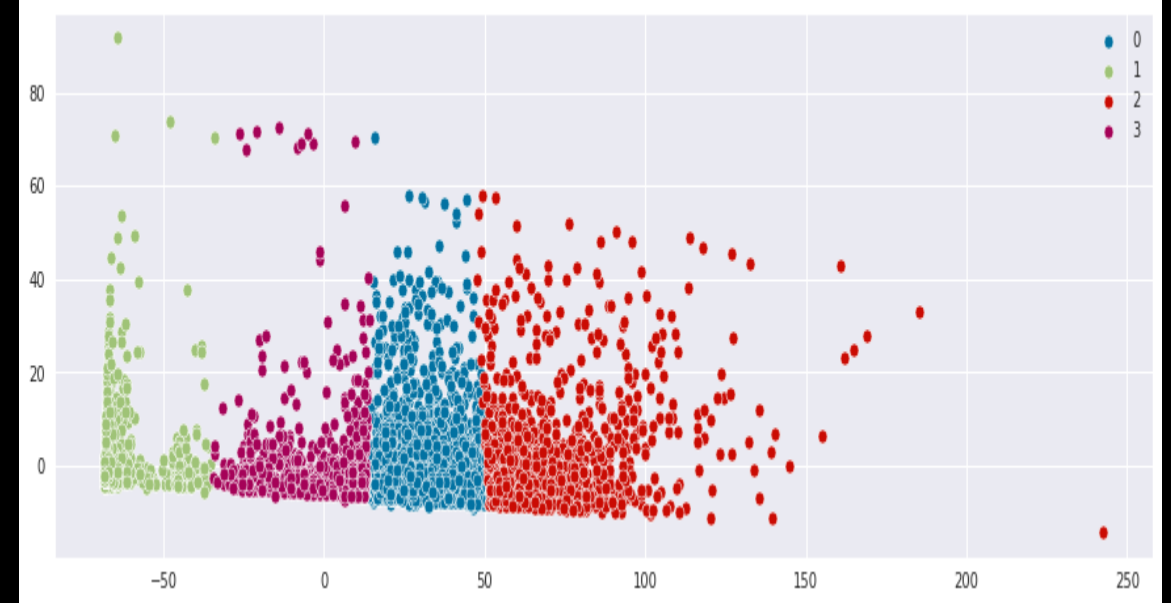
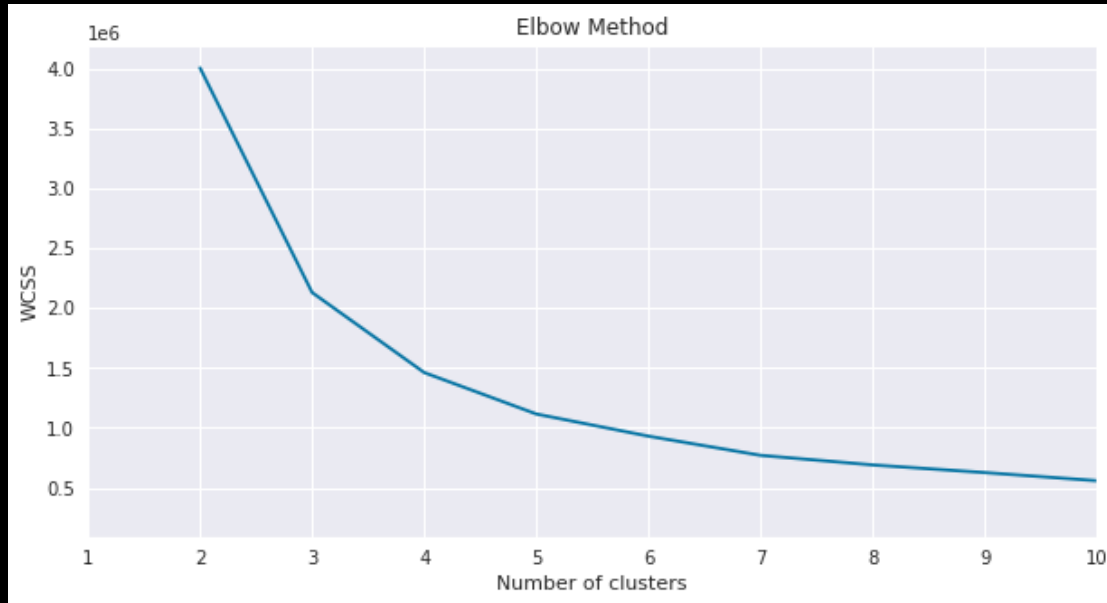
Birch (balanced iterative reducing and clustering using hierarchies) is an unsupervised data mining algorithm that performs hierarchical clustering over large data sets. with modifications, it can also be used to accelerate k-means clustering and gaussian mixture modeling with the expectation-maximization algorithm.

Evaluation criteria

Silhouette Coefficient : is a metric to evaluate the performance of clustering algorithm. It uses compactness of individual clusters(intra cluster distance) and separation amongst clusters (inter cluster distance) to measure an overall representative score of how well our clustering algorithm has performed. The value of the silhouette coefficient is between $[-1, 1]$, A score of 1 denotes the best meaning that the data point (i) is very compact within the cluster to which it belongs and far away from the other clusters. The worst value is -1 and Values near 0 denote overlapping clusters.

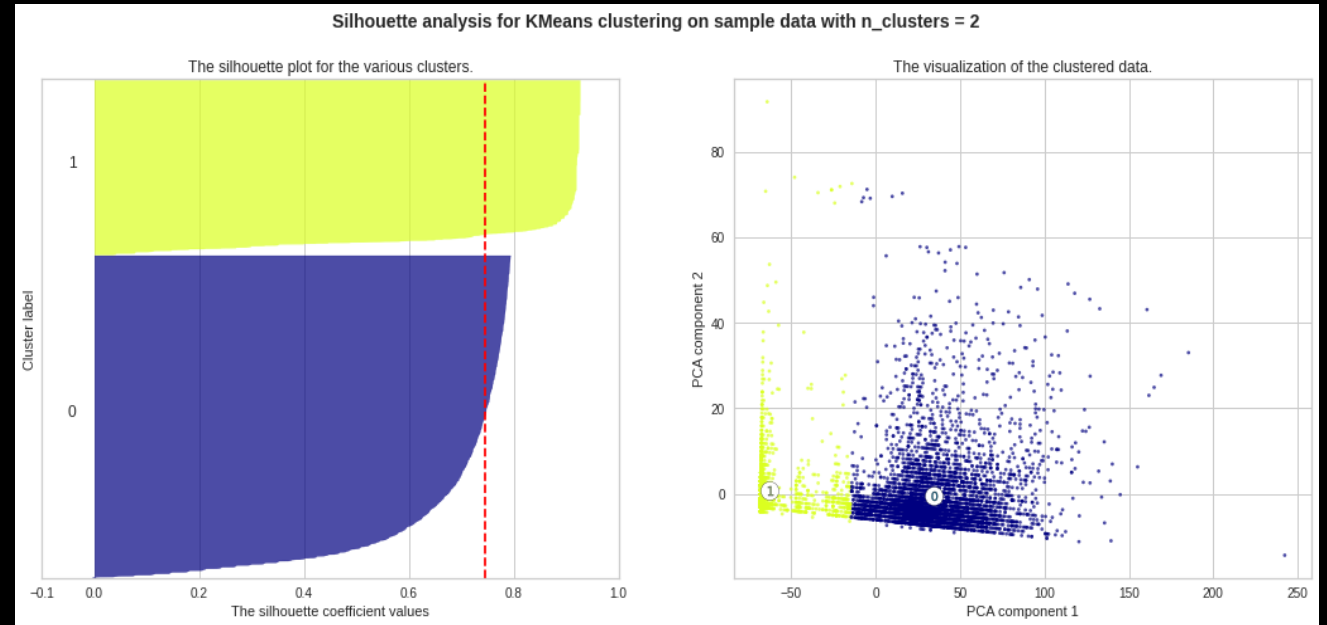
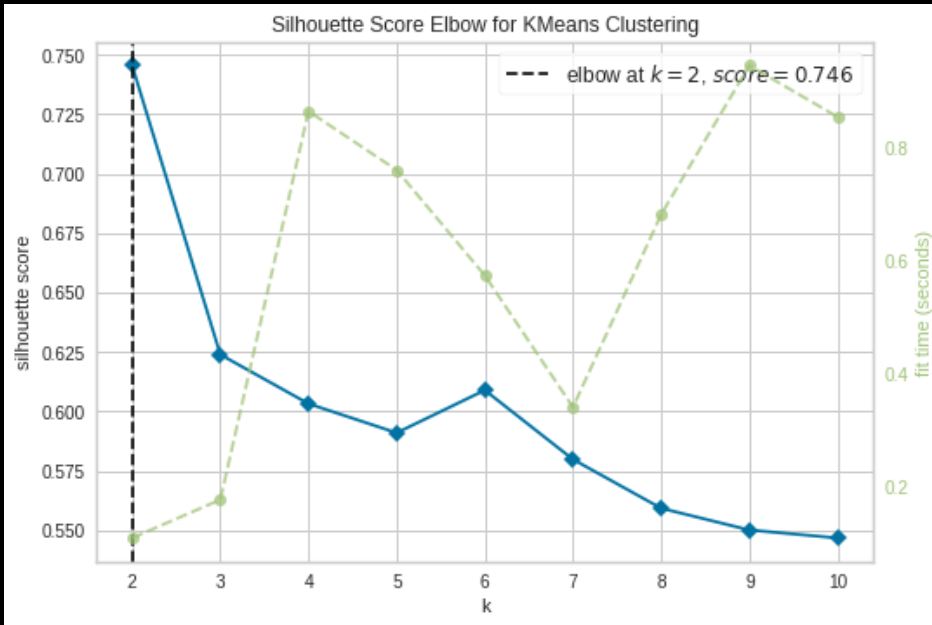
The Davies-Bouldin index : It is most commonly used to evaluate the goodness of split by a K-Means clustering algorithm for a given number of clusters. The intuition behind Davies-Bouldin index is the ratio between the within cluster distances and the between cluster distances and computing the average overall the clusters. It is therefore relatively simple to compute, bounded – 0 to 1, lower score is better.

KMeans Clustering with Tf-idf Vectorizer and Elbow method



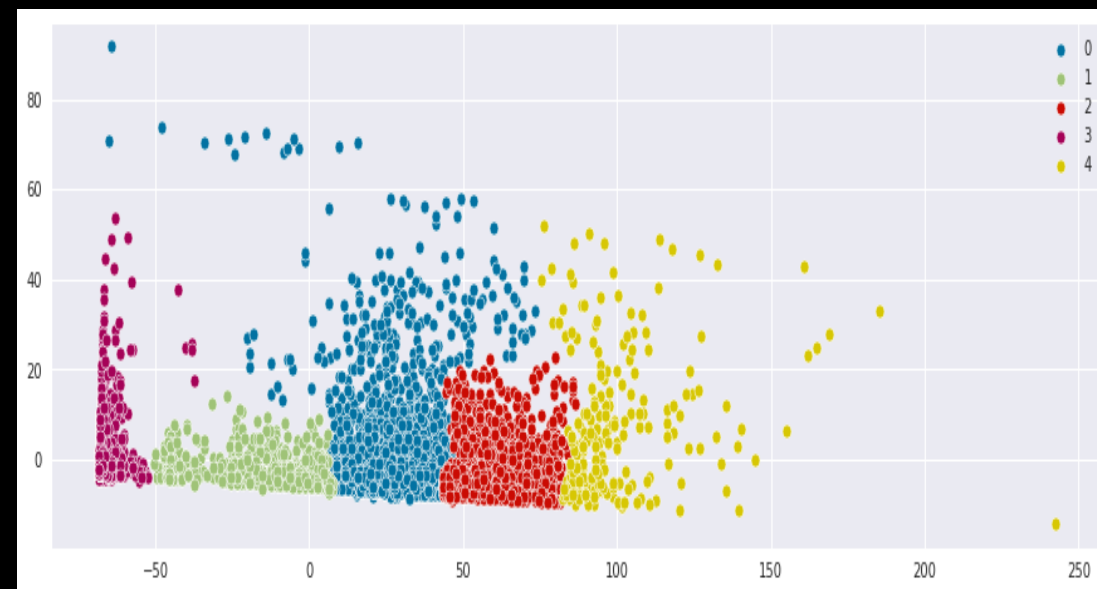
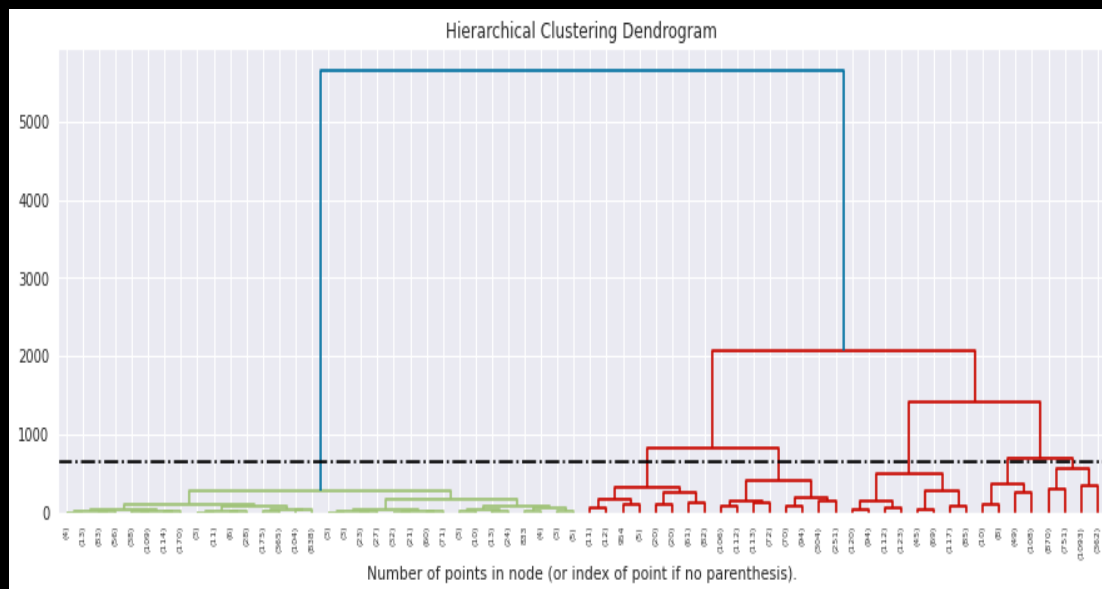
Davies_bouldin_score is 0.5976456325717732
Silhouette Coefficient is 0.6033112929703297
Optimum Clusters is 4

KMeans Clustering with Tf-idf Vectorizer and Silhouette Score



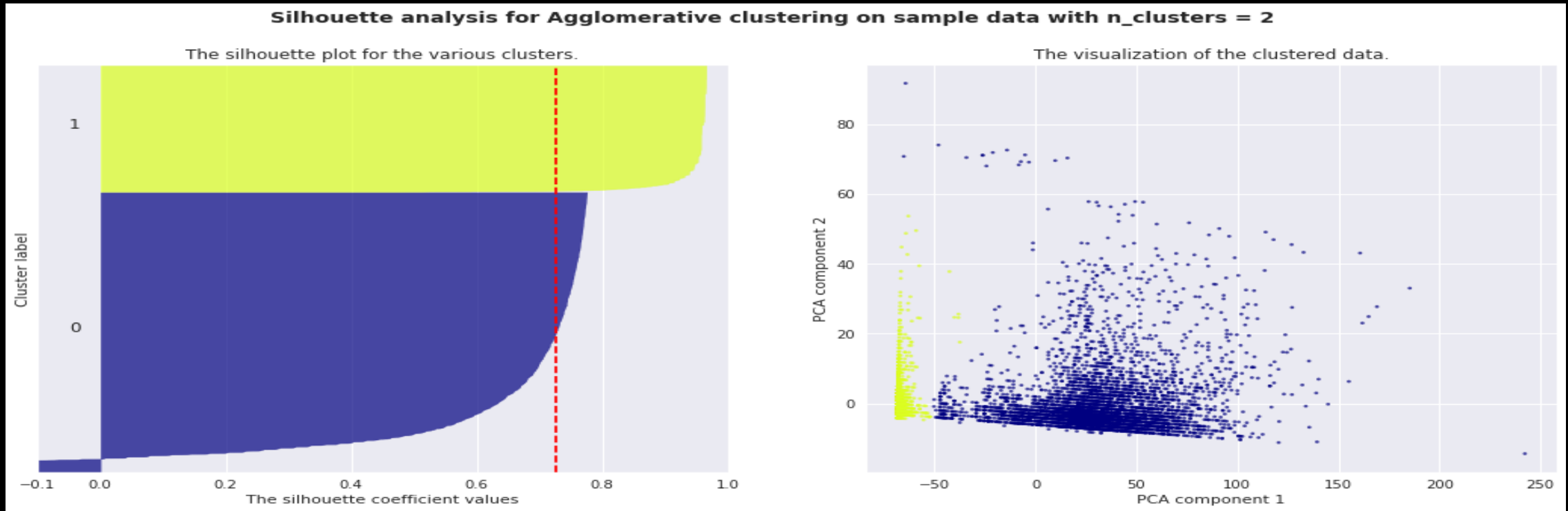
Davies_bouldin_score is 0.30702512618385613
Silhouette Coefficient is 0.7455302527893082
Optimum Clusters is 2

Agglomerative Clustering with Tf-idf Vectorizer and Dendrogram



Davies_bouldin_score is 0.6121097624841052
Silhouette Coefficient is 0.6121097624841052
Optimum Clusters is 5

Agglomerative Clustering with Tf-idf Vectorizer and Silhouette Score

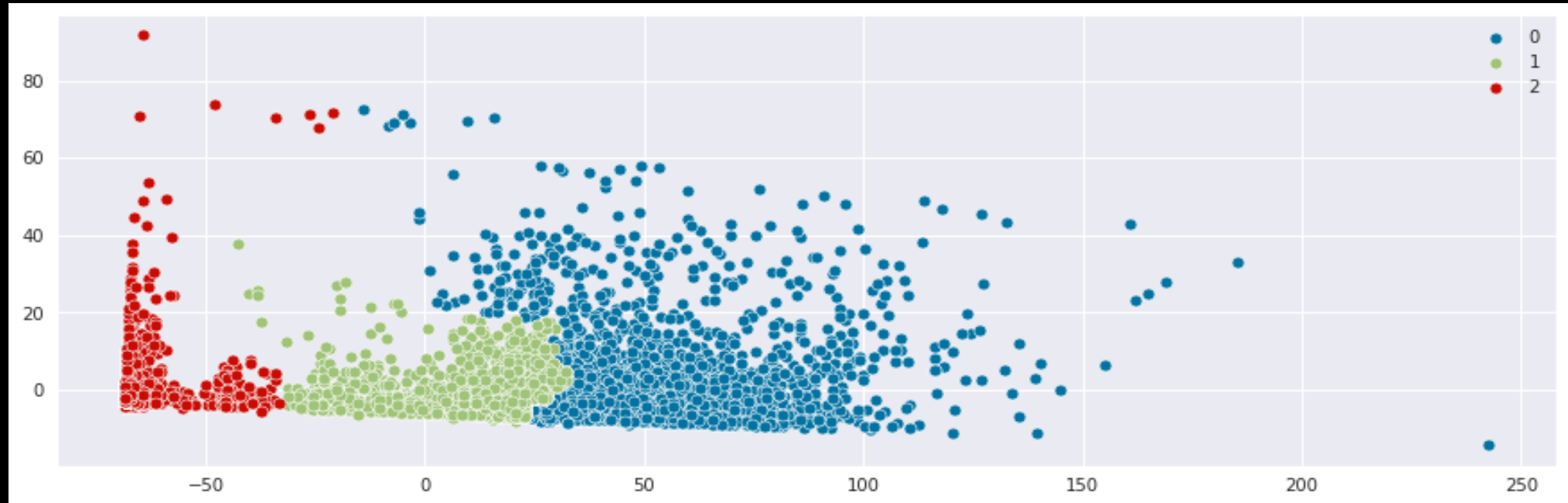


Davies_bouldin_score is 0.26837257511131596

Silhouette Coefficient is 0.7252616861641785

Optimum Clusters is 2

Birch Clustering with Tf-idf Vectorizer



Davies_bouldin_score is 0.6600927586224409

Silhouette Coefficient is 0.5326807978658717

Optimum Clusters is 3

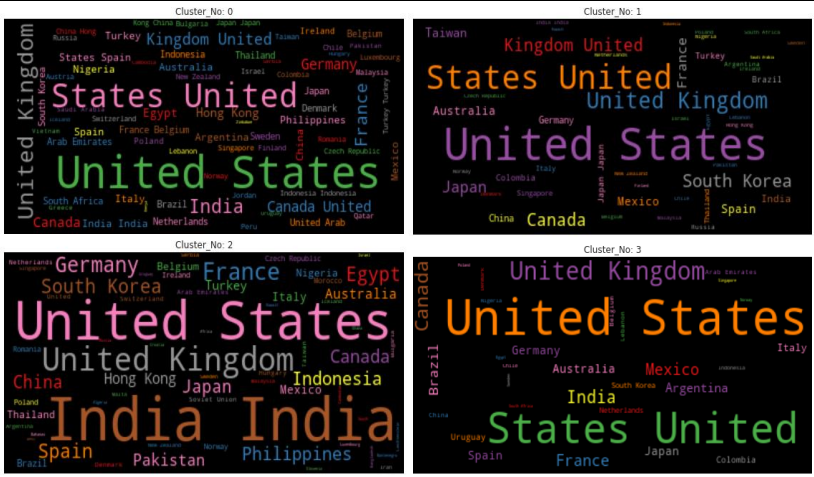
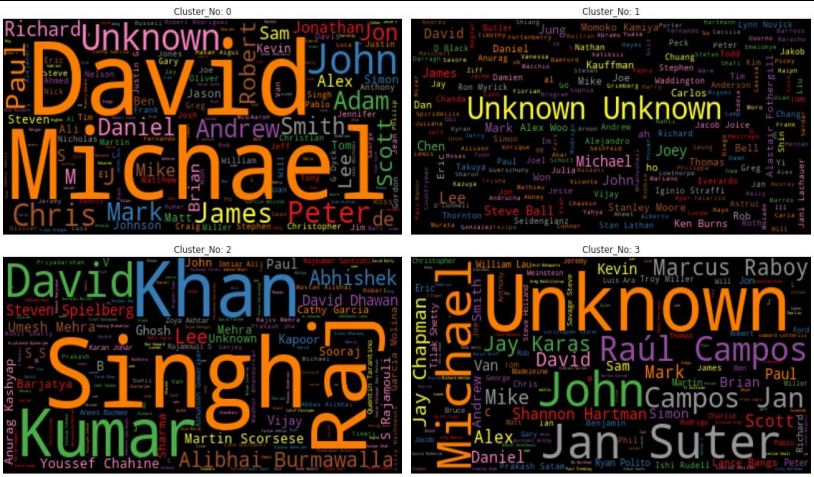
Data represented by each cluster:

C-0	C-1
C-2	C-3

Director

Country

Rating



Cast

Genre

Description



Recommendations System:

We obtained recommendations for Movies and Tv- Shows using Cosine similarity and sorting method.

Cosine similarity measures the similarity between two vectors of an inner product space. It is measured by the cosine of the angle between two vectors and determines whether two vectors are pointing in roughly the same direction. It is often used to measure document similarity in text analysis

```
[ ] # recommendations for Movies.  
movie_recommendations = pd.DataFrame(recommendations('Zulu Man in Japan'), columns=['Recommendations'])  
movie_recommendations.head(5)
```

	Recommendations
0	Zulu Man in Japan
1	Roots
2	We Are One
3	ZZ TOP: THAT LITTLE OL' BAND FROM TEXAS
4	Mexicanos de Bronce

```
[ ] # recommendations for Tv_Shows.  
Tv_Shows_recommendations = pd.DataFrame(recommendations('3%'), columns=['Recommendations'])  
Tv_Shows_recommendations.head(5)
```

	Recommendations
0	3%
1	Back with the Ex
2	Argon
3	Million Pound Menu
4	How To Ruin Christmas

```
[47] # recommendations for Director wise  
recommendations = pd.DataFrame(Recommended_Syestem('Jorge','director'), columns=['Recommendations'])  
recommendations.head(5)
```

	Recommendations
0	Los Herederos
1	7:19
2	Pablo Escobar: Angel or Demon?
3	ReMastered: Massacre at the Stadium
4	Alell

```
[48] # recommendations for Country wise  
recommendations = pd.DataFrame(Recommended_Syestem('India','country'), columns=['Recommendations'])  
recommendations.head(5)
```

	Recommendations
0	Sardaarji 2
1	Kung Fu Yoga
2	Alibaba Aur 40 Chor
3	Bombairiya
4	Romeo Ranjha


```
[49] # recommendations for genere wise  
recommendations = pd.DataFrame(Recommended_Syestem('Horror Movies','listed_in'), columns=['Recommendations'])  
recommendations.head(5)
```

	Recommendations
0	23:59
1	A.M.I.
2	1920
3	Rosemary's Baby
4	Tremors 2: Aftershocks

```
[50] # recommendations for cast wise  
recommendations = pd.DataFrame(Recommended_Syestem('Elias','cast'), columns=['Recommendations'])  
recommendations.head(5)
```

	Recommendations
0	Quién te cantará
1	Fuller House
2	Zodiac
3	Fallen
4	Teenage Mutant Ninja Turtles: The Movie

Summary:

	Model Name	Davies bouldin score	Silhouette Coefficient	Optimum Clusters	
0	KMeans Clustering with TfidfVectorizer and Elbow method	0.597646	0.603311	4	
1	Agglomerative Clustering with TfidfVectorizer and Dendrogram	0.612110	0.583070	5	
2	Birch Clustering with TfidfVectorizer	0.660093	0.532681	3	

	Model Name	Davies bouldin score	Silhouette Coefficient	Optimum Clusters
0	KMeans Clustering with Tf-idf Vectorizer and Silhouette Score	0.307025	0.745530	2
1	Agglomerative Clustering with Tf-idf Vectorizer and Silhouette Score	0.268372	0.725261	2

Based on above table we will conclude KMeans Clustering with Tf-idfVectorizer and Elbow method model and Agglomerative Clustering with Tf-idfVectorizer and Dendrogram model as our best model because it provides good overall Davies bouldin score, Silhouette Coefficient and optimum clusters.

Conclusions:

- Movies make up 69% of the content on Netflix, while TV shows make up the remaining 31%.
- This was surprising to discover that movies make up the majority of the Netflix content. But lately, it has been concentrating more on television shows.
- The majority of this content is made available either towards the end of the year or at the beginning.
- The ratings TV-MA and TV-14 have a lot of television shows and movies that are suitable for adults and teenagers, respectively.
- The majority of Netflix TV shows only have a single season, while the majority of the movies have running times between 80 and 100 minutes.
- America and India produces the majority of movies and television Shows, But in India Movies are more well produced than television shows.
- Netflix's addition of movies and television shows increased from 2016 to 2019, but from 2019 to 2021, it gradually dropped, possibly as a result of COVID.

THANKS