Capstone Project – 4

NETFLIX MOVIES AND TV SHOWS CLUSTERING

By

SHAIK AHMAD BASHA

Contents

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.

The main objective behind the project is to explore the dataset to find some key understandings and to cluster similar content by using machine learning clustering algorithms

Data Understanding :

The dataset has 7787 rows and 12 columns. It contains information of movies / TV Shows like cast, genre, description, director etc.

The features of the dataset are :

- 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 Cleaning and Manipulation :

In Data cleaning and manipulation, we will check for null values, duplicated values and manipulate the data for our need.

In the data there are 5 features with null values. They are

- director
- cast
- country
- date_added
- rating

<u>Handling of Null Values :</u>

- For director column, we can replace the null values with 'unknown'
- For cast column, we can replace the null values with 'unknown'
- For country column, we can replace null values with most frequent value (mode)
- For date_added column, we can drop the rows with null values
- For rating column, we can drop the null values.

	feature	null values	% of null values
3	director	2389	30.67934
4	cast	718	9.22050
5	country	507	6.51085
6	date_added	10	0.12842
8	rating	7	0.08989
0	show_id	0	0.00000
1	type	0	0.00000
2	title	0	0.00000
7	release_year	0	0.00000
9	duration	0	0.00000
10	listed_in	0	0.00000
11	description	0	0.00000

Created a new column ‘month_added’ from ‘data_added’ which contains month number in which the movie/TV Show released

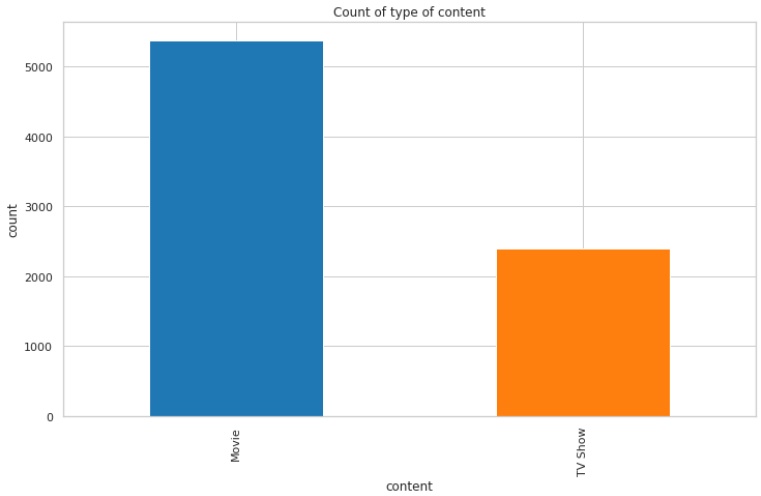
Created a new column ‘movie_duration’ from ‘duration’ column which contains the duration of movie in minutes.

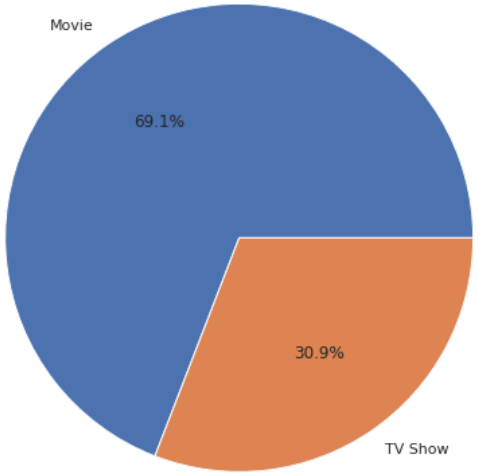
Created a new column ‘num_of_seasons’ from ‘duration’ column which contains the number of seasons of TV Show.

Created a new column ‘target_ages’ which contains age groups.

Exploratory Data Analysis :

1. Type



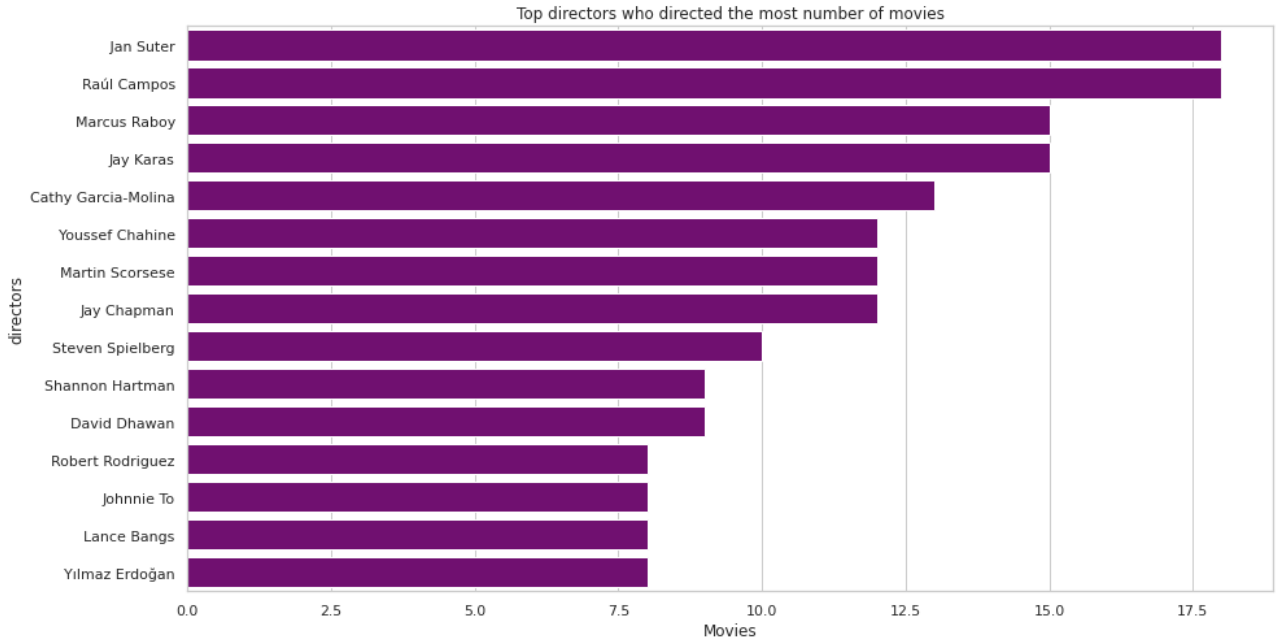


From the above visualizations,

- **we can observe that there are more movies (69.1%) that tv shows (30.9%) on Netflix**

2. Director

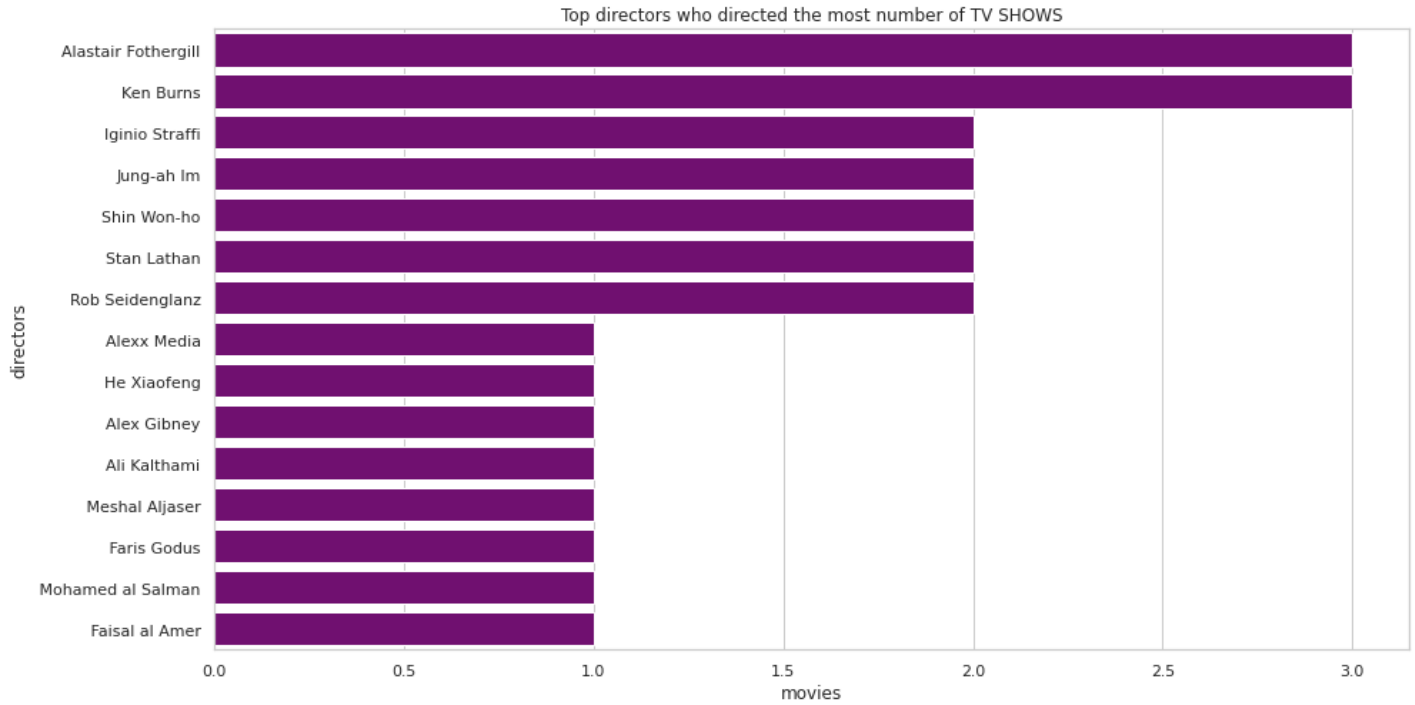
Top directors who directed most number of movies



From the above visualizations,

❖ **we can say that Jan Suter and Raul Campos directed most number of movies.**

Top directors who directed most number of TV SHows

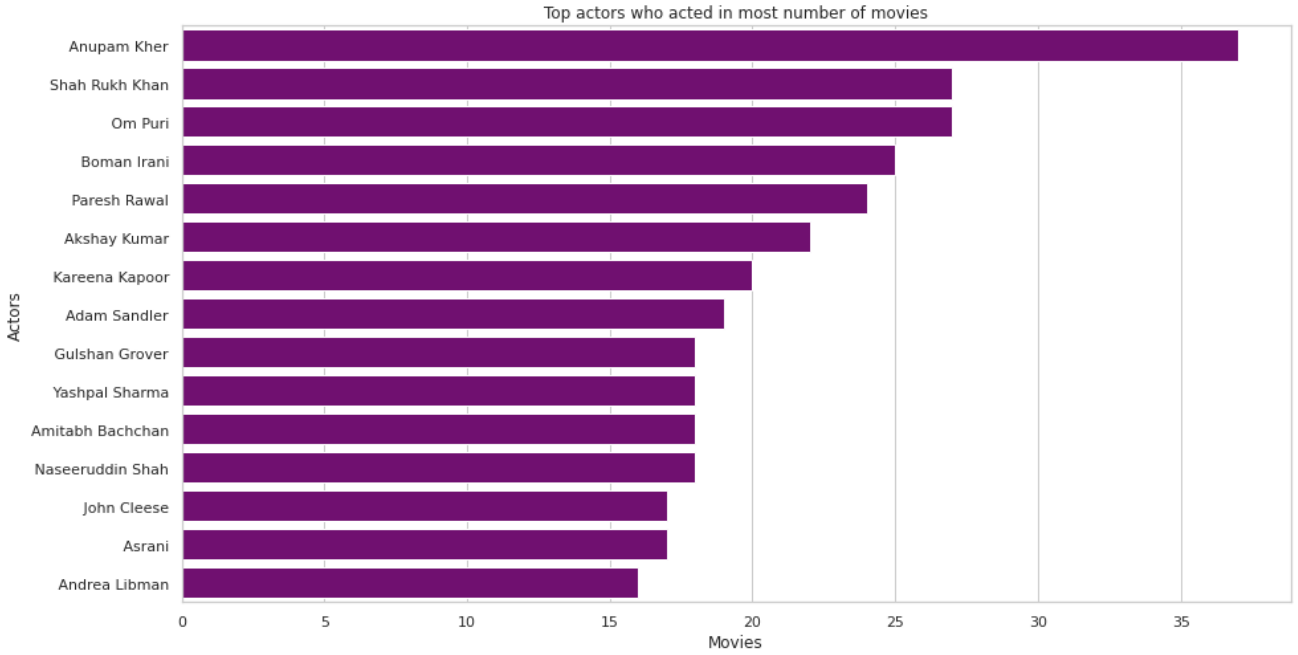


From the above visualizations, we can observe that

- **Alastair Fothergill and Ken Burns directed most number of TV Shows.**

3. Cast

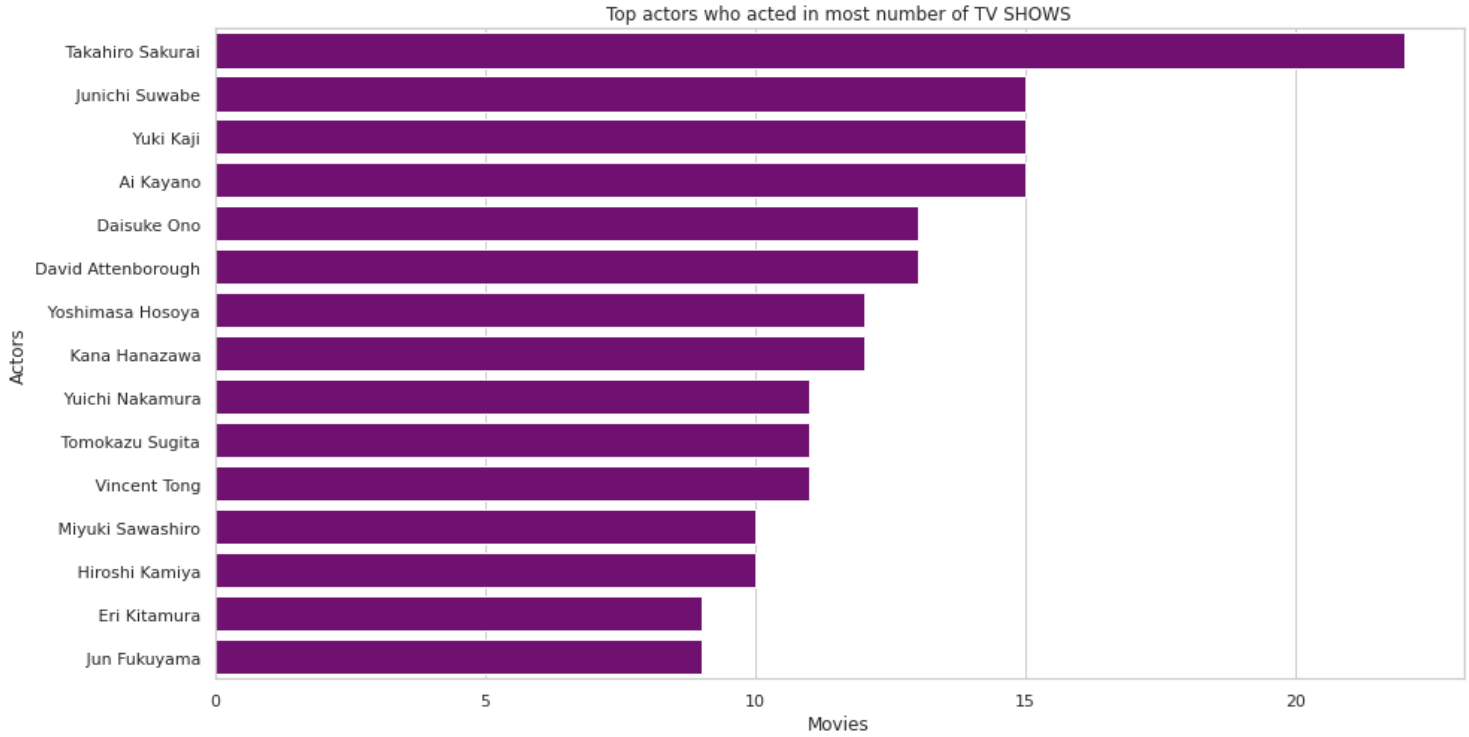
Top actors who acted in most number of movies



From the above visualizations, we can observe that

- **Anupam Kher acted in most number of movies followed by Shah Rukh Khan.**

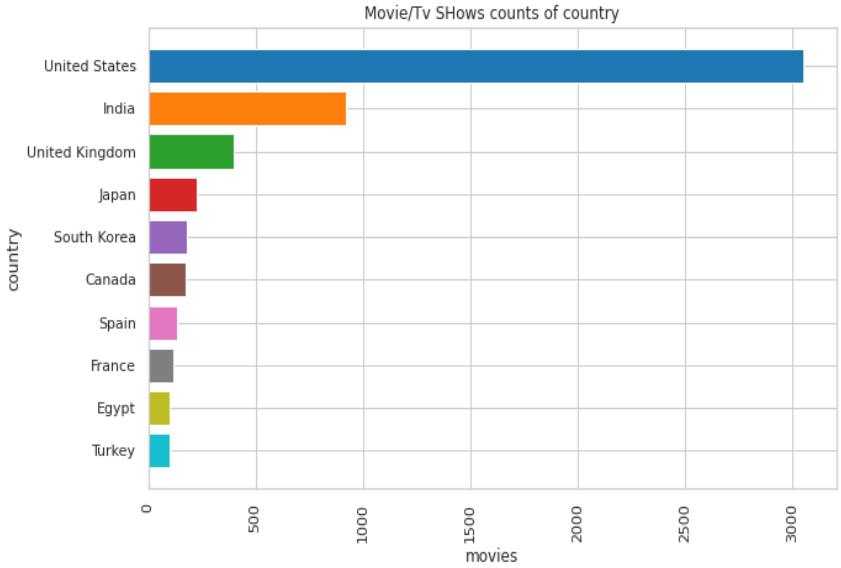
Top actors who acted in most number of TV Shows

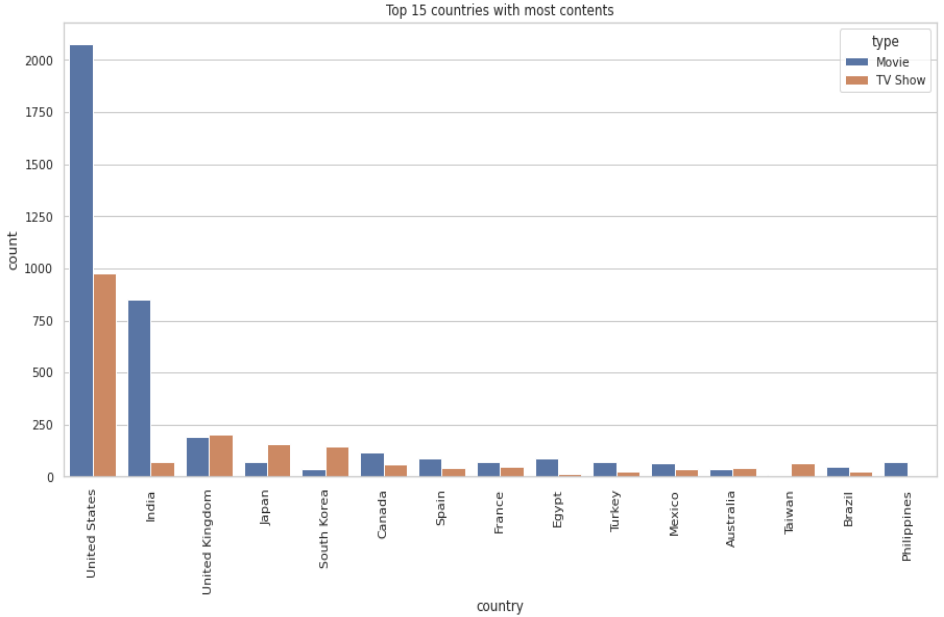


From the above visualizations, we can observe that

❖ **Takahiro Sakurai acted in most number of TV Shows followed by Junichi Suwabe**

4. Country

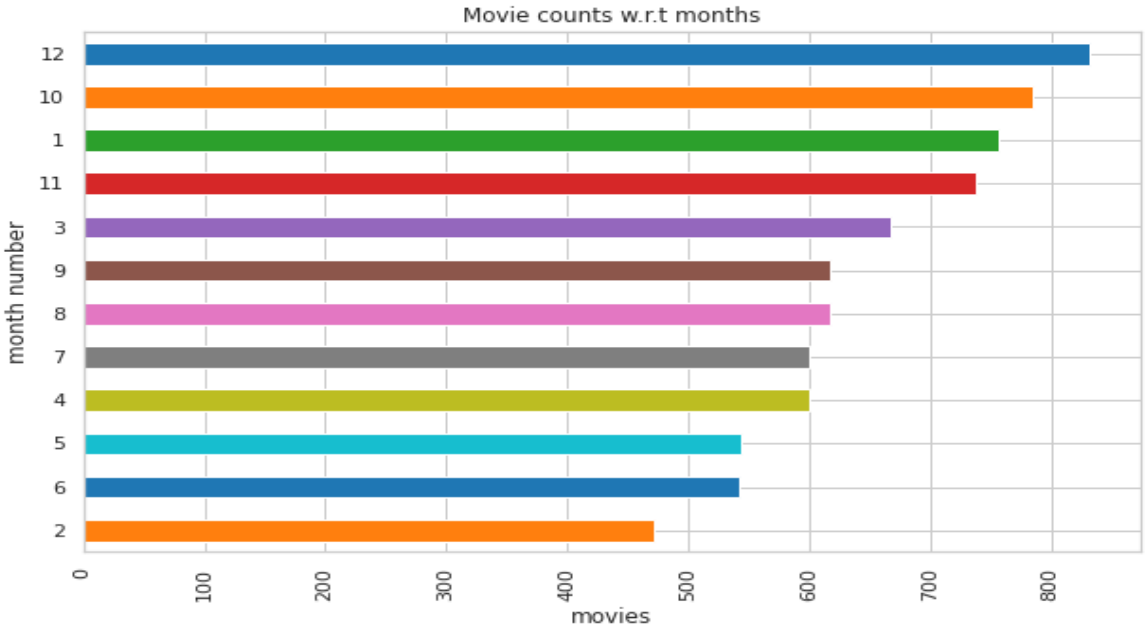




From the above visualizations,

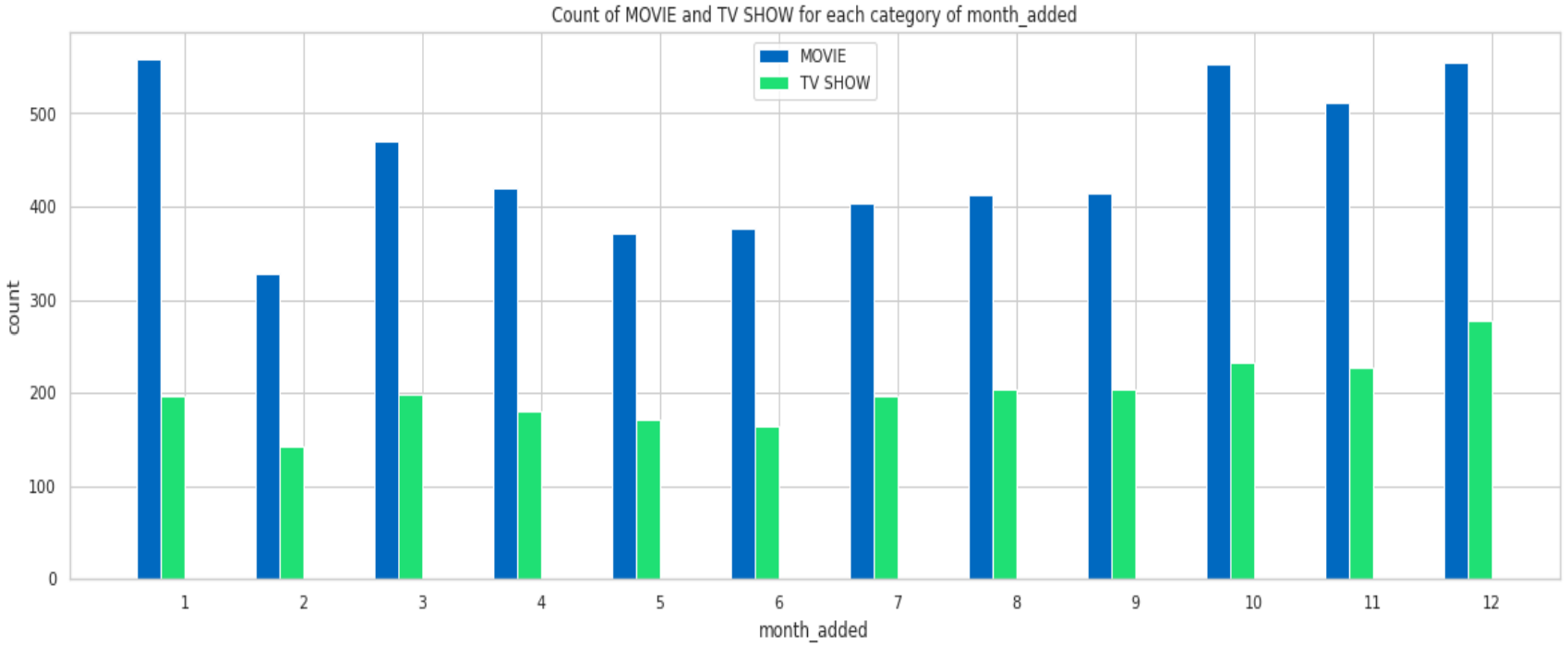
- **we can say the United Stated produced most number of movies/Tv Shows followed by India.**

5. month_added



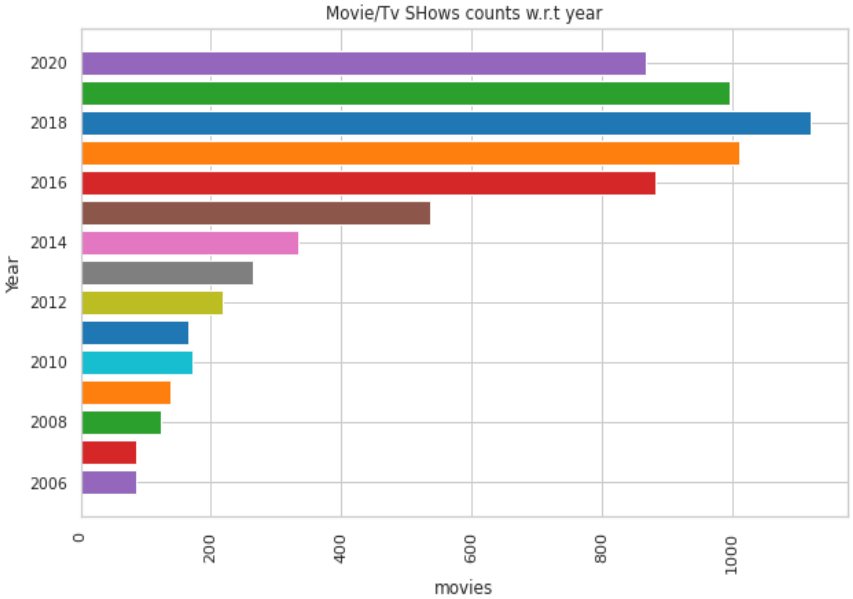
From the above visualizations,

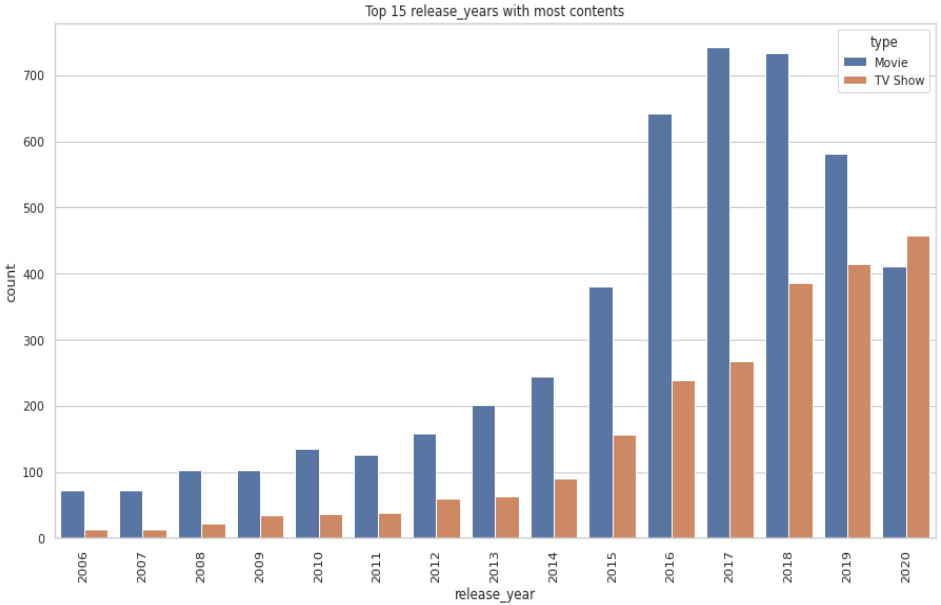
- **we can observe that most number of movies/tv shows are released in the month of december.**



From the above observations, we can observe that most number of movies are released in the month of December and January.

6. Release Year

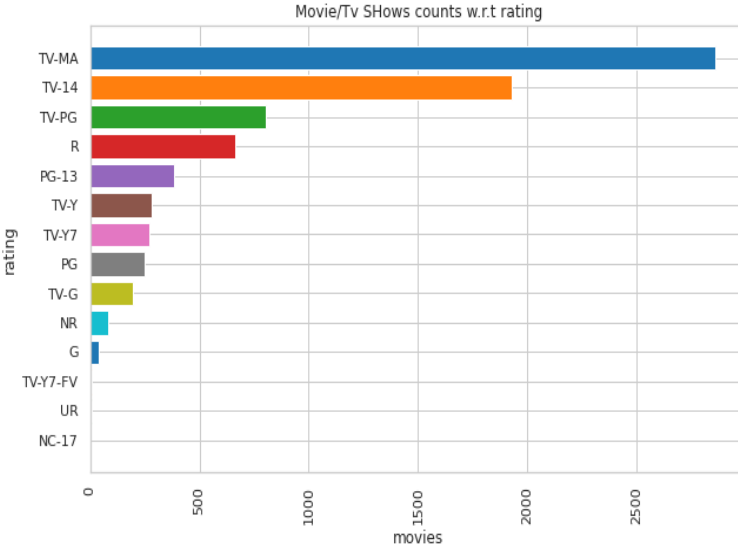


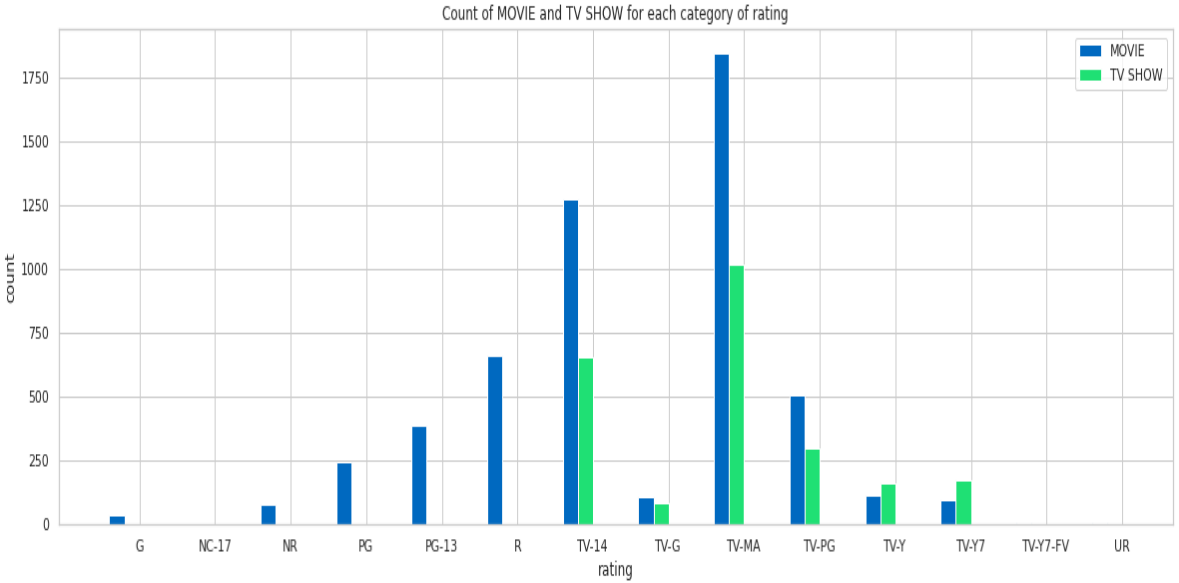


From the above Visualizations we can say that

- **In the year 2018 most number of movies and TV Shows are released.**
- **In the recent years, netflix has started to increase the TV Shows content as you can see the graph is increasing gradually.**

7. Rating

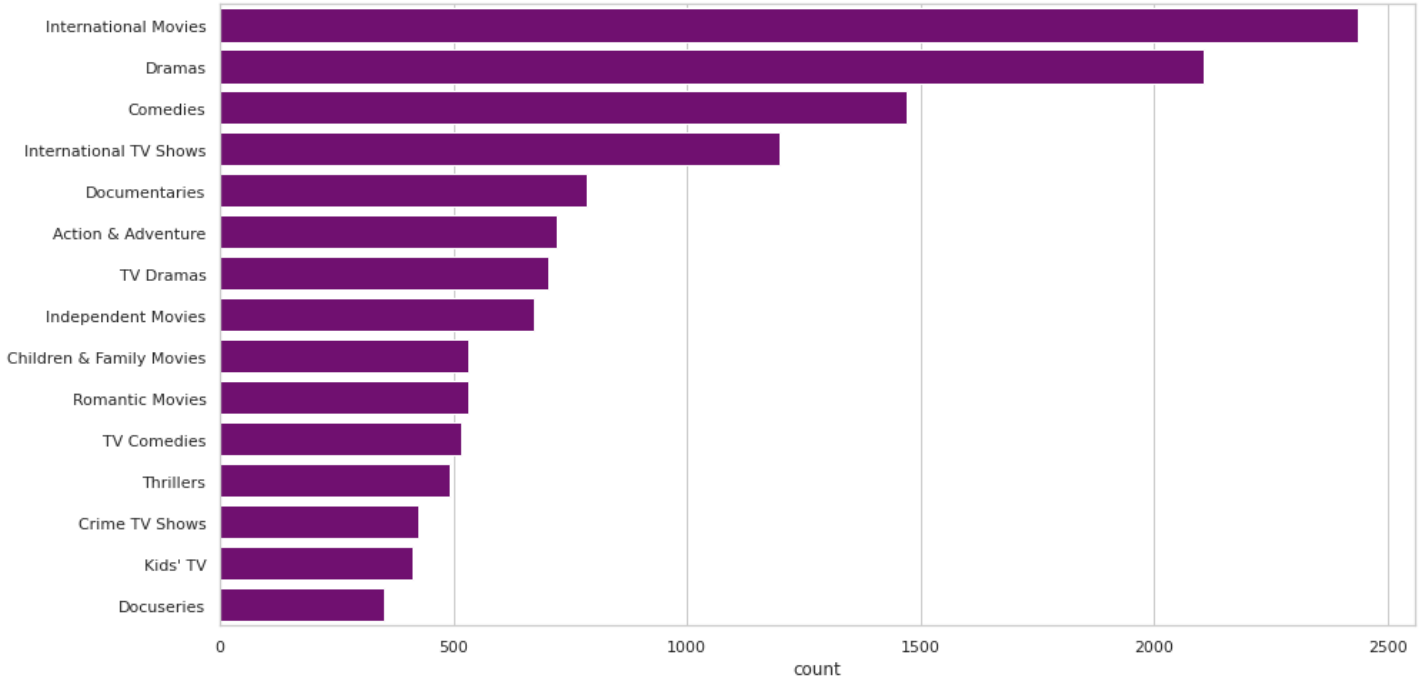




From the above visualizations,

- **We can observe that most number of movies and TV SHows belongs the rating category of 'TV-MA'**

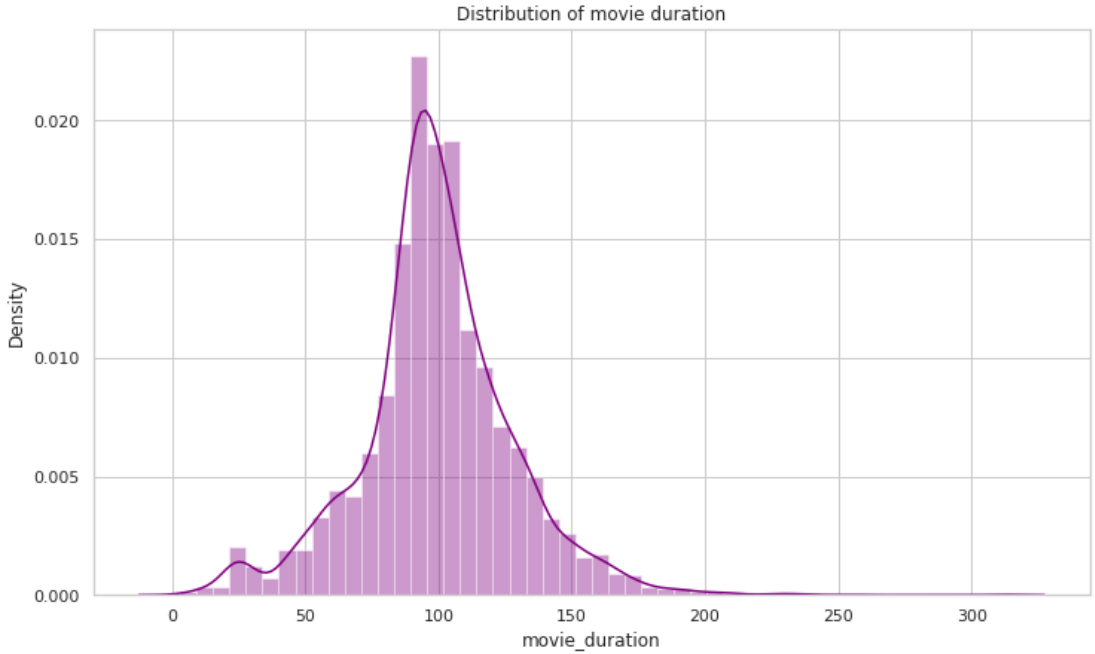
8. Listed in



From the above visualization,

- **we can observe that most number of movies / TV Shows are from genre of 'International Movies' followed by dramas.**

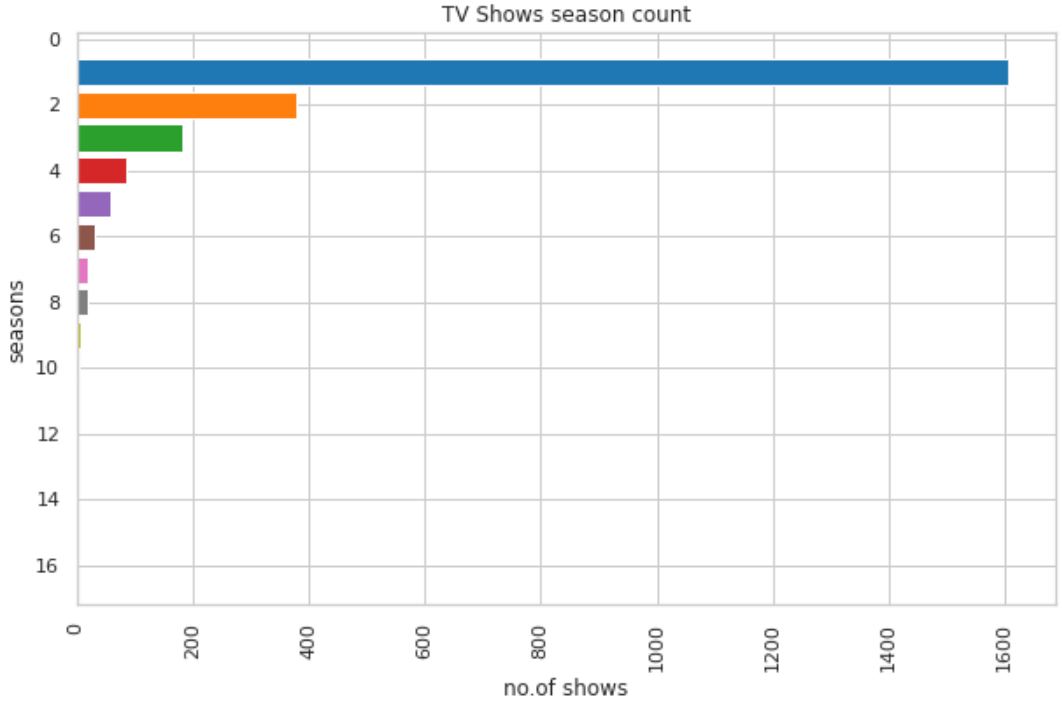
9. Movie Duration



From the above visualizations,

- **We can observe that most number of movie duration is in the range of 80 to 125 min.**

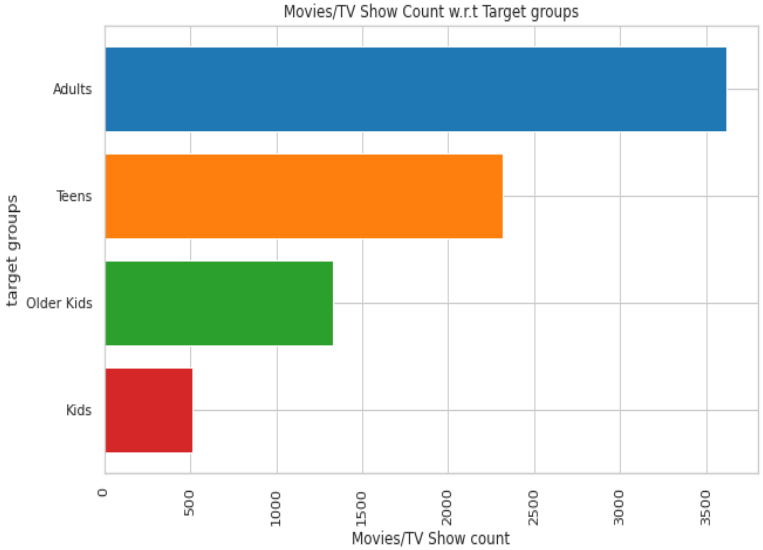
10. Number of seasons

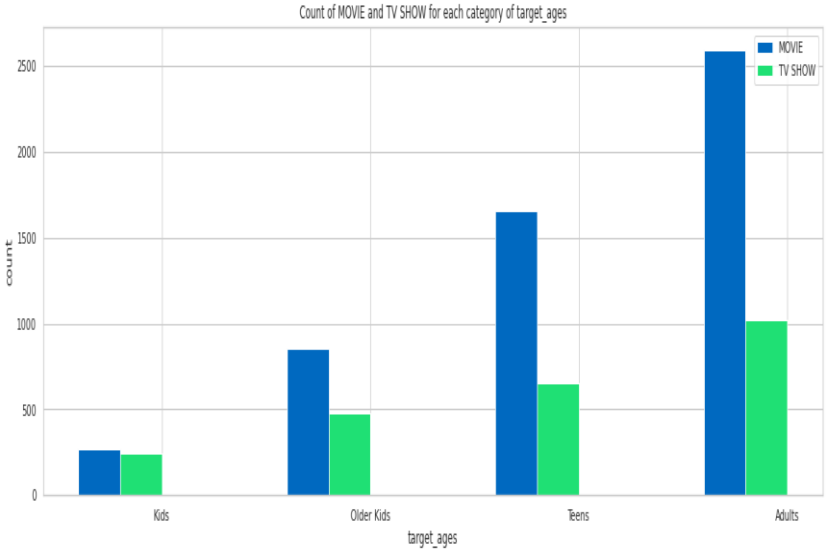


From the above visualizations,

- **we can observe that most number of TV Shows have only One season.**

11. Target Ages





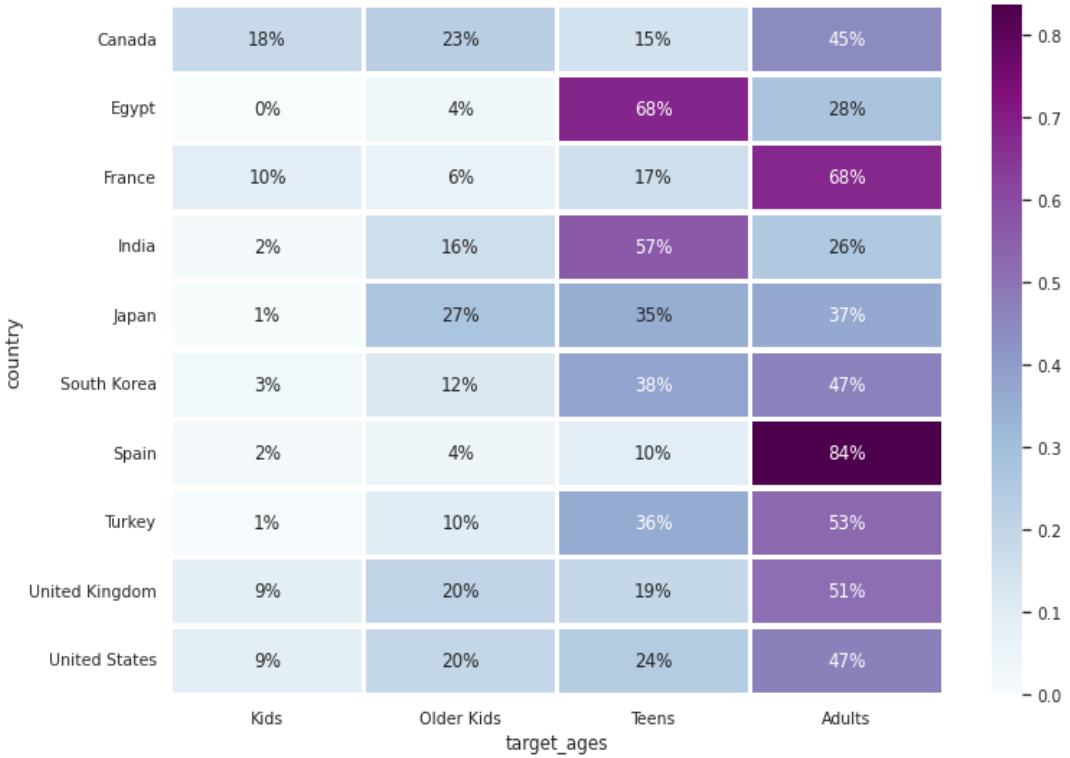
From the above visualizations,

❖ **we can observe that most number of movies and TV Shows are made for adults only.**

Lets observe how countries made content w.r.t age groups

Observations :

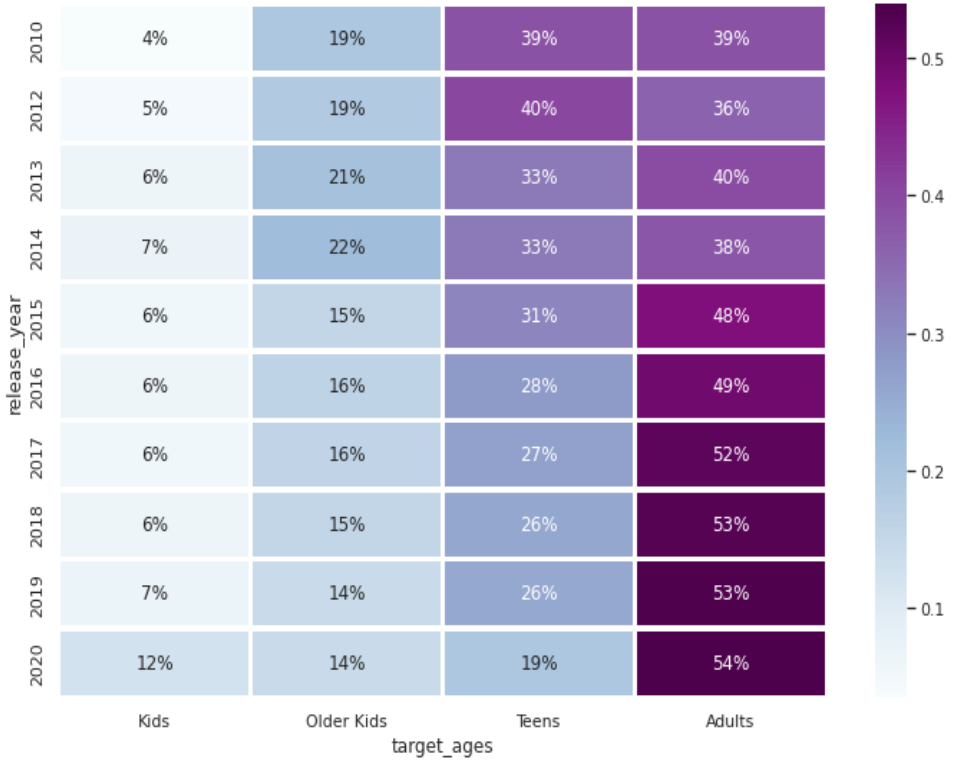
- The content for Adults is mostly made from Spain and France countries
- The content for Teens is mostly made from Egypt and India countries.
- The content for Older Kids is mostly made from Japan country.
- The content for Kids is mostly made from Canada country



Lets observe released year wise content w.r.t age groups

Observations :

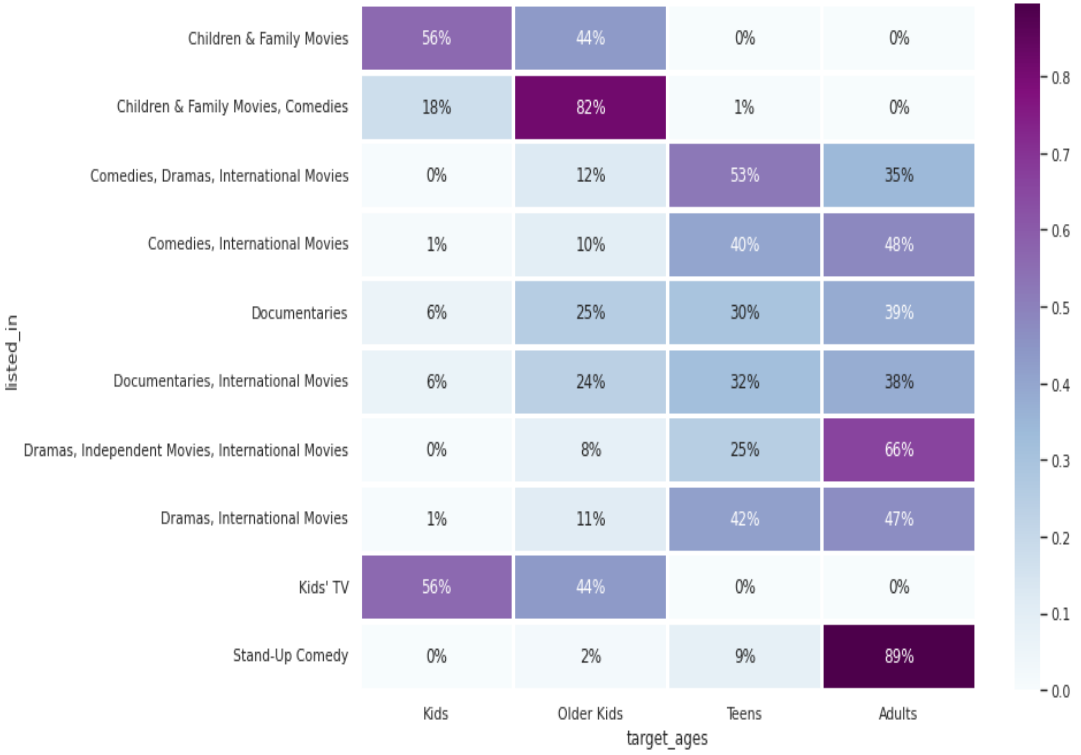
- The content for Adults is mostly released in recent years(2020, 2019, 2018, 2017) .
- The content for Teens is mostly released in the year 2010 and 2012
- The content for Older is Kids released in year of 2014.
- The content for Kids is mostly released in the year of 2020.



Lets observe how genres and age groups are related

Observations :

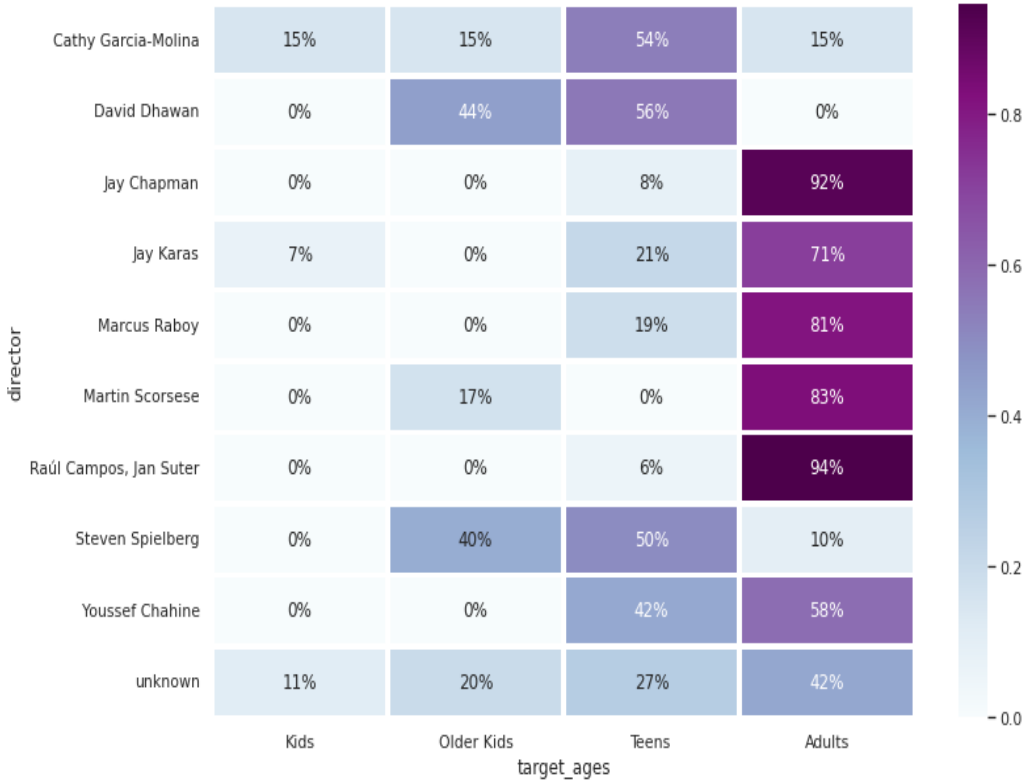
- The content for Adults is mostly from the genre of 'stand up comedy' followed by Dramas.
- The content for Teens is mostly from the genre of 'Comedies, Dramas, international movies'
- The content for Older Kids is mostly from the genre of 'Children, Family movies and Comedies'
- The content for Kids is mostly from the genre of Kids TV and Children and Family Movies



Lets observe how directors made content w.r.t age groups

Observations :

- 94% content from Raul Campos and Jan Suter is for adults
- Steven Spielberg , David Dhawan and Cathy Garcia-Molina are the directors who makes content mostly for Teens and Older Kids.



Text Pre-processing :

Here we are going to input only text feature to the models, those text features are

- ❖ Title
- ❖ Type
- ❖ Director
- ❖ Cast
- ❖ Country
- ❖ Rating
- ❖ Listed_in
- ❖ Description

so we have to clean and preprocess the above text features. Generally there are 5 steps in which we can prepare the text data as model-ready data. They are

1) Tokenization
2) Lower Casing
3) Removing Stop words
4) Removing Punctuations
5) Stemming

TF-IDF Vectorizer :

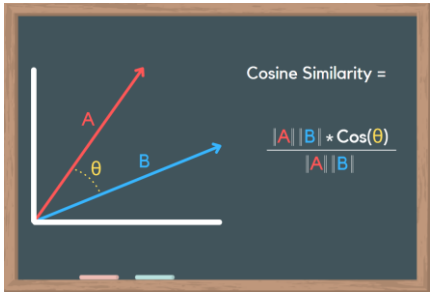
- ❖ TF-IDF is technique in Natural Language Processing for converting words in Vectors and with some semantic information and it gives weighted to uncommon words.
- ❖ After converting words in vectors, the next step is to reduce the dimensionality.
- ❖ Here we can use Principal Component Analysis (PCA) to the reduce the dimensionality.
- ❖ After dimensionality reduction, the data is ready to fit into the models.

Here I experimented clustering using

- ➢ Cosine Similarity
- ➢ K Means CLustering
- ➢ Hierarchical Clustering

Finding Similar Movies Using Cosine Similarity:

Cosine Similarity is a measurement that quantifies the similarity between two or more vectors. The cosine similarity is the cosine of the angle between vectors. The vectors are typically non-zero and are within an inner product space.



By using cosine similarity we can recommend similar movies.

```
similar_movie_3('Breaking Bad', 10)
```

```
The similar movies for movie Breaking Bad are
Breaking Bad
Better Call Saul
The Show
Have You Ever Fallen in Love, Miss Jiang?
My Life My Story
Dancing Angels
Killer Ratings
The School Nurse Files
Pyaar Tune Kya Kiya
Hormones
```

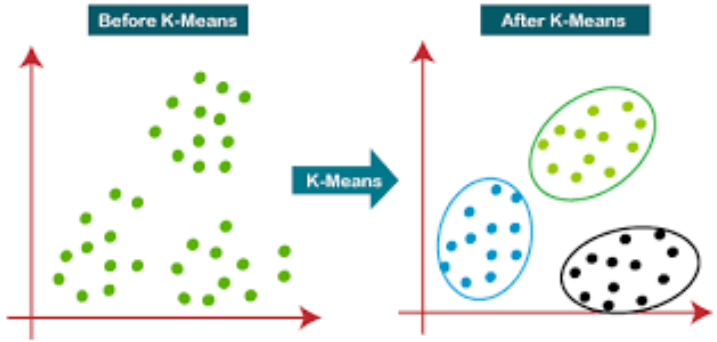
```
similar_movie_3('stranger things', 10)
```

```
The similar movies for movie stranger things are
Stranger Things
Beyond Stranger Things
Prank Encounters
The Umbrella Academy
Kiss Me First
Anjaan: Special Crimes Unit
Reckoning
The OA
Disappearance
The 4400
```

Creating Clusters :

K Means Clustering :

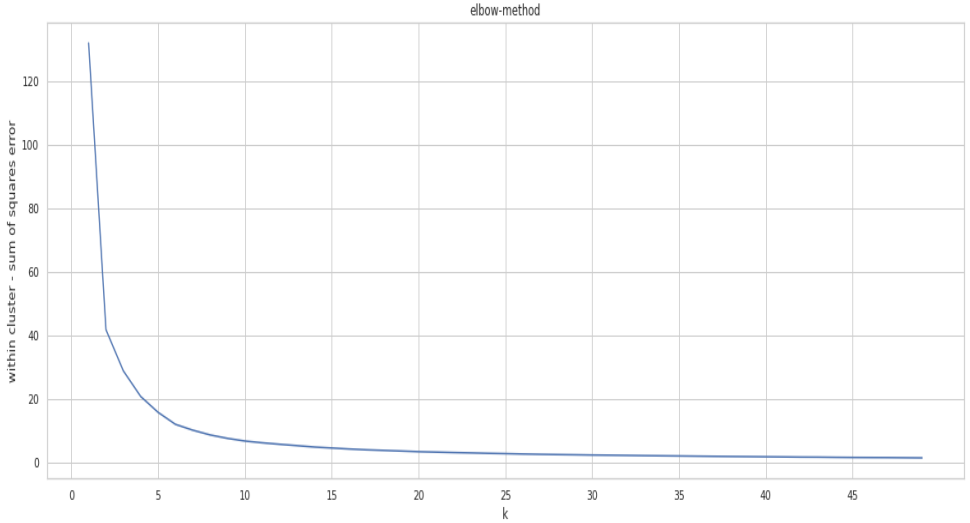
K-means is a centroid-based clustering algorithm, where we calculate the distance between each data point and a centroid to assign it to a cluster. The goal is to identify the K number of groups in the dataset.

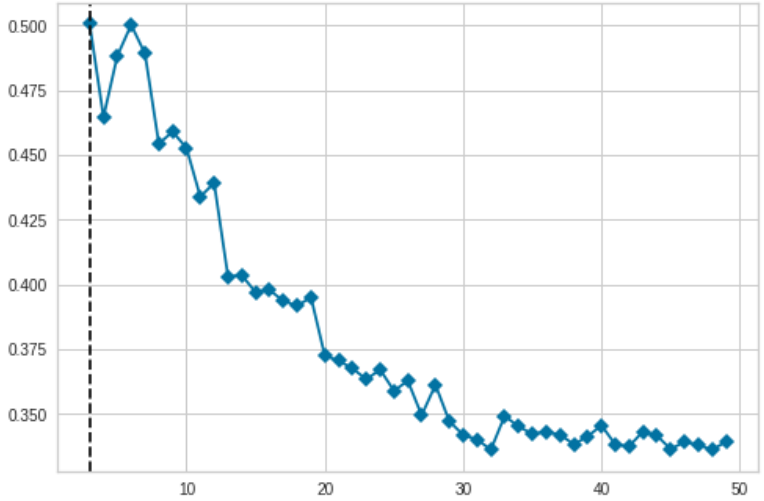


Steps Involved in K Means Clustering:

1) Choosing the number of clusters
2) Initializing centroids
3) Assign data points to the nearest cluster
4) Re-initialize centroids
5) Repeat steps 3 and 4

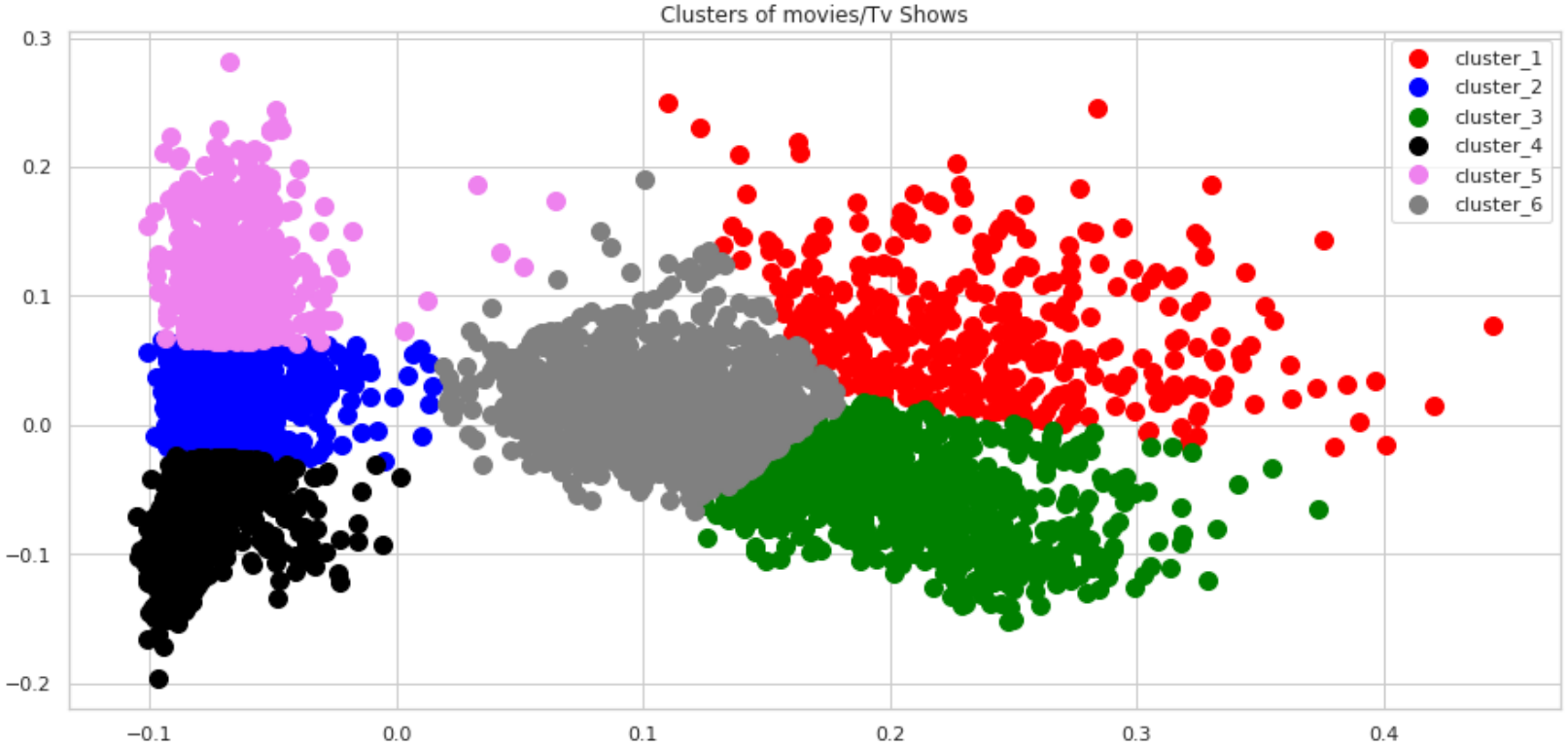
Determining optimal k value :





By using elbow method and silhouette analysis, we can choose optimal k value as 6.

Plotting clusters after fitting the data into K Means algorithm :

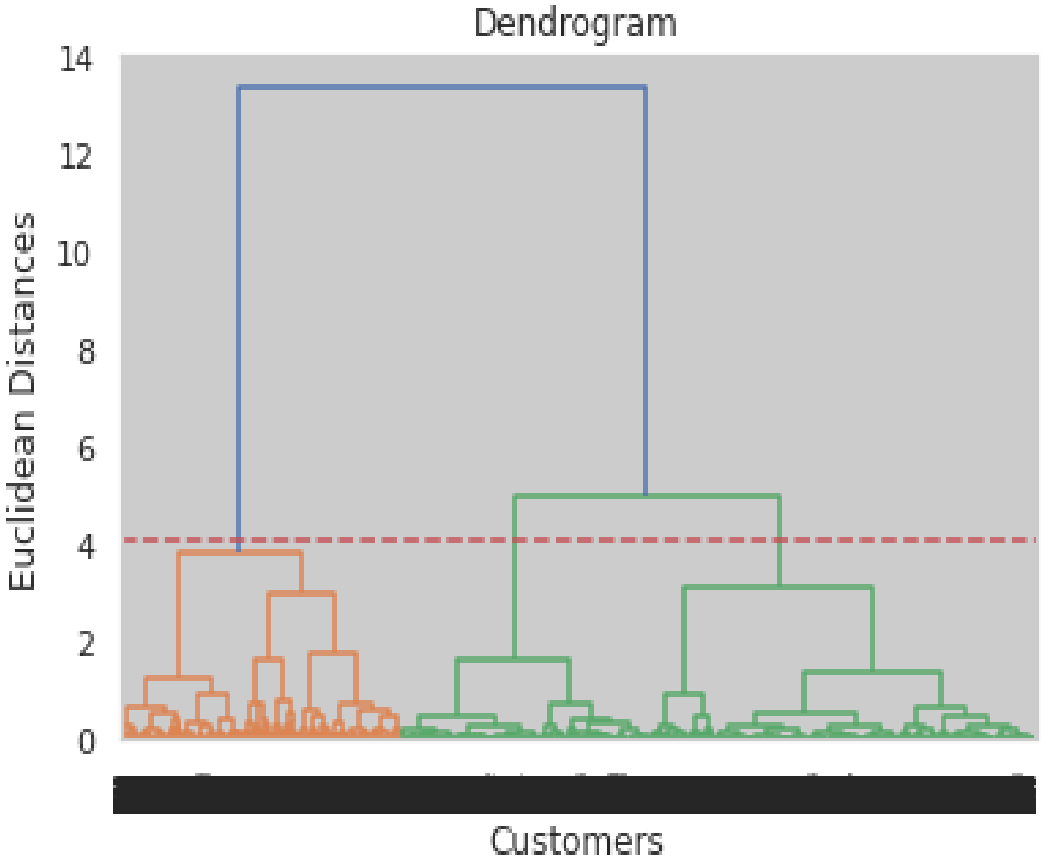


silhouette score is 0.5001880488098 43

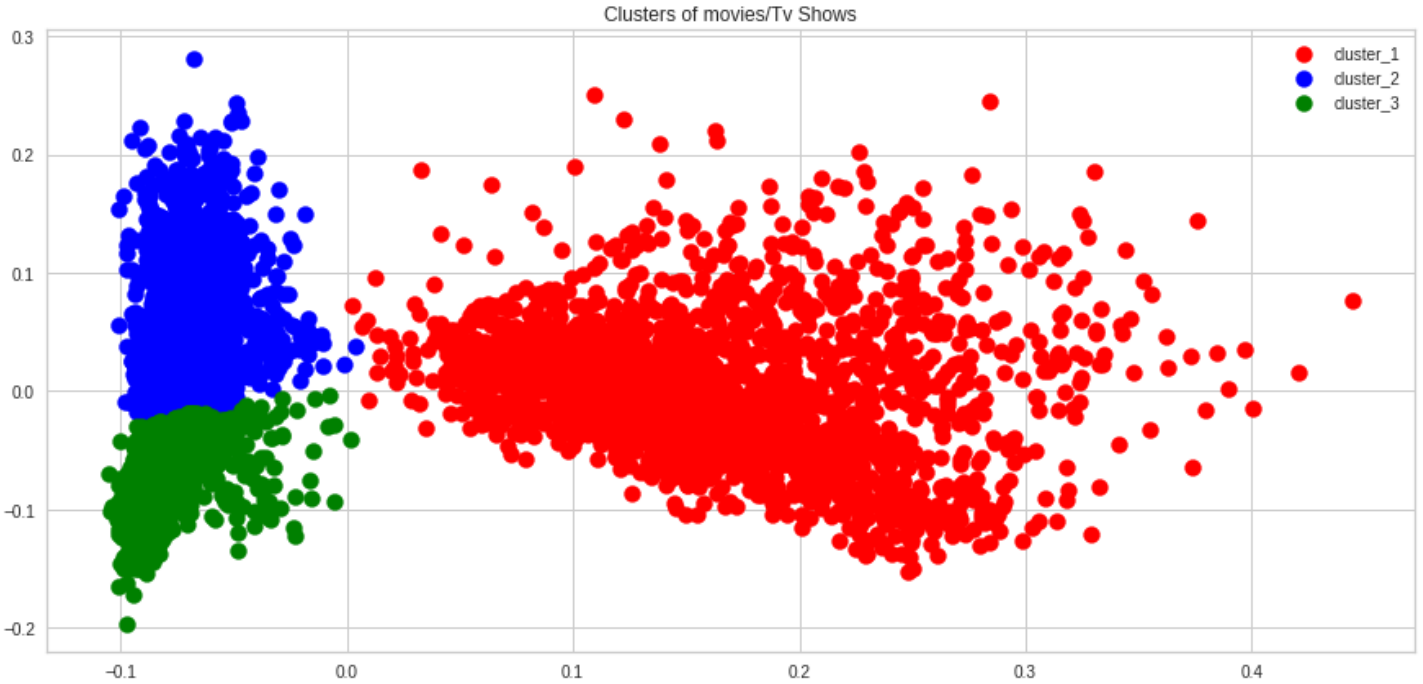
Hierarchical Clustering :

Hierarchical clustering is a popular method for grouping objects. It creates groups so that objects within a group are similar to each other and different from objects in other groups. Clusters are visually represented in a hierarchical tree called a *dendrogram*.

From the dendrogram and silhouette analysis, we can choose optimal k value as 3.



Plotting clusters after fitting the data into Agglomerative Clustering algorithm :



Silhouette score : 0.5049348034764315

Conclusions :

- There are more movies that tv shows on Netflix
- Jan Suter and Raul Campos directed most number of movies.
- Alastair Fothergill and Ken Burns directed most number of TV Shows.
- Anupam Kher acted in most number of movies followed by Shah Rukh Khan.
- Takahiro Sakurai acted in most number of TV Shows followed by Junichi Suwabe
- United Stated produced most number of movies/Tv Shows followed by India.
- Most number of movies are released in the month of December and January.
- In the year 2018 most number of movies and TV Shows are released.
- In the recent years, Netflix has started to increase the TV Shows content as you can see the graph is increasing gradually.
- Most number of movies and TV Shows belongs the rating category of 'TV-MA'
- Most number of movies / TV Shows are from genre of 'International Movies' followed by dramas.
- Most number of movie duration is in the range of 80 to 125 min.
- Most number of TV Shows have only One season.
- Most number of movies and TV Shows are made for adults only.

- The content for Adults is mostly made from Spain and France countries.
- The content for Teens is mostly made from Egypt and India countries.
- The content for Older Kids is mostly made from Japan country.
- The content for Kids is mostly made from Canada country.
- The content for Adults is mostly released in recent years(2020, 2019, 2018, 2017).
- The content for Teens is mostly released in the year 2010 and 2012.
- The content for Older is Kids released in year of 2014.
- The content for Kids is mostly released in the year of 2020.
- The content for Adults is mostly from the genre of 'stand up comedy' followed by Dramas.
- The content for Teens is mostly from the genre of 'Comedies, Dramas, international movies'
- The content for Older Kids is mostly from the genre of 'Children, Family movies and Comedies'

- The content for Kids is mostly from the genre of Kids TV and Children and Family Movies
- 94% content from Raul Campos and Jan Suter is for adults.
- Steven Spielberg , David Dhawan and Cathy Garcia-Molina are the directors who makes content mostly for Teens and Older Kids.
- By using the data, a recommender system was created with cosine similarity.
- By using the data, created a clusters using k means and hierarchical clustering.
- By applying silhouette analysis, for k means, the optimal k value is 6.
- And for k = 6, silhouette score is 0.5001759759427156
- For hierarchical clustering, from dendrogram we can choose optimal k value as 3. And silhouette score is 0.5049348034764315

THANK YOU