Capstone Project

Netflix Movies and TV shows clustering

Member Name

Soumabha Sarkar

Agenda

- Introduction
- Problem Statement
- Data Description
- Null Value
- Exploratory Data Analysis
- Data Cleaning
- Data Pre-processing
- Model Implementation
- K- Means
- Cluster Analysis

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 services 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, you are required to do

1. Exploratory Data Analysis
2. Understanding what type content is available in different countries.
3. Is Netflix has increasingly focusing on TV rather than movies in recent years.
4. Clustering similar content by matching text-based features

Data Description

The data was collected from Flixable which is third party Netflix search engine. The dataset consists of movies and TV shows data till 2019. The dataset has 7787 rows of data.

The dataset consists of eleven textual columns and one numeric column.

Attribute Information :

1. **show_id :** Unique ID for every Movie / TV Show
2. **type :** Identifier - A Movie or TV Show
3. **title :** Title of the Movie / TV Show
4. **director :** Director of the Movie

Data Description

5. **cast :** Actors involved in the movie / show
6. **country :** Country where the movie / show was produced
7. **date_added :** Date it was added on Netflix
8. **release_year :** Actual Release year of the movie / show
9. **rating :** TV Rating of the movie / show
10. **duration :** Total Duration - in minutes or number of seasons
11. **listed_in :** Genre
12. **description:** The Summary description

Null Values

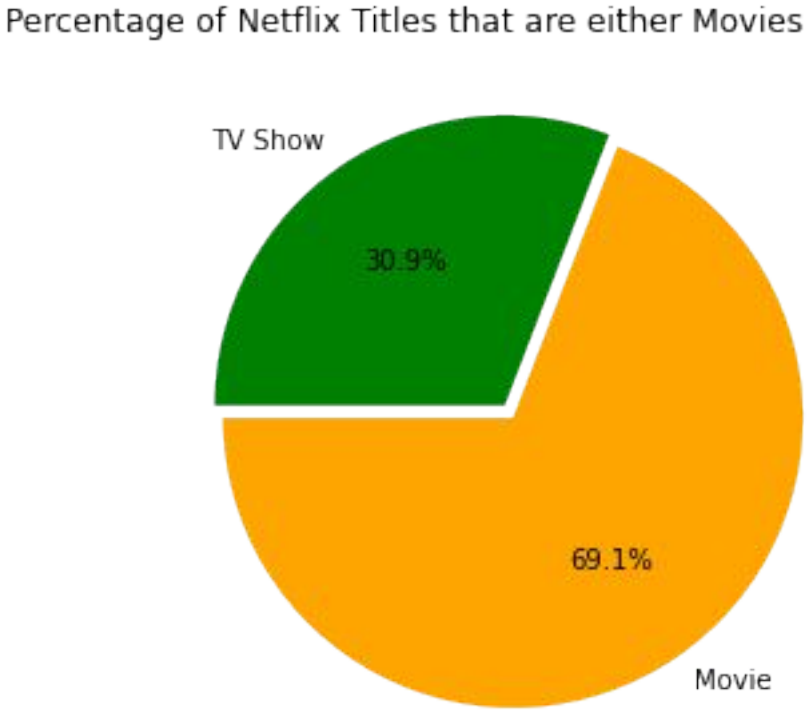
Null Value Treatment:

- ***Director*** feature have more than ***30.68%*** of null values. Filling null values by 'unknown'.
- ***Country*** feature have ***6.51%*** of null values. Filling null values by mode of feature.
- ***Cast feature*** have ***9.22%*** of null values. Filling null values by 'unknown'.
- ***Rating*** feature have ***0.09%*** of null values. Filling null values by mode of feature.
- ***Date_added*** feature have ***0.13%*** of null values. Dropping rows corresponding to null values.

Exploratory Data Analysis

- Type wise analysis.
- Country wise analysis.
- Rating wise analysis.
- Type wise rating analysis.
- Year wise release date analysis.
- Movie release year wise analysis.
- TV shows release year wise analysis.
- Month wise release date analysis.
- Top ten genres in movies.
- Top ten genres in TV shows.
- Movies length distribution.
- TV shows season wise distribution.
- Top ten actors movie wise analysis.
- Top ten actors TV shows wise analysis.

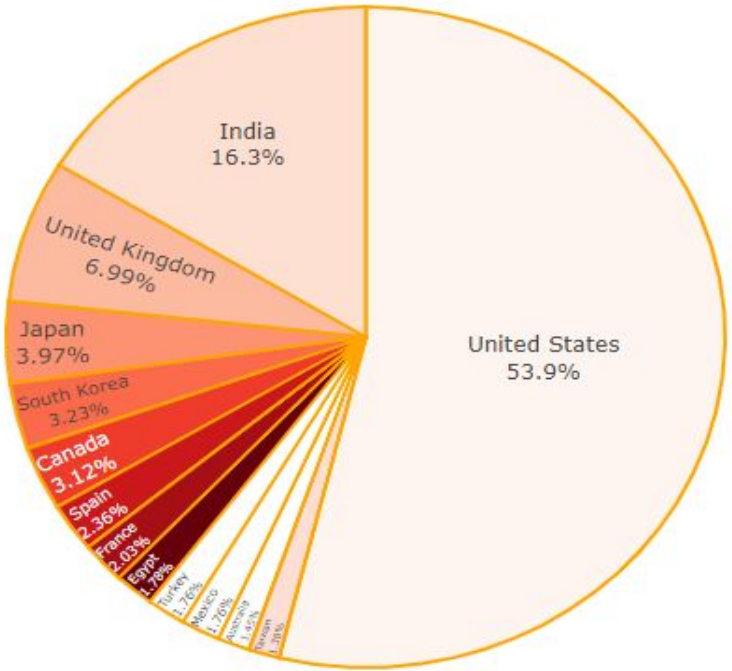
Type wise analysis



Type of content available on Netflix

- It is 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.

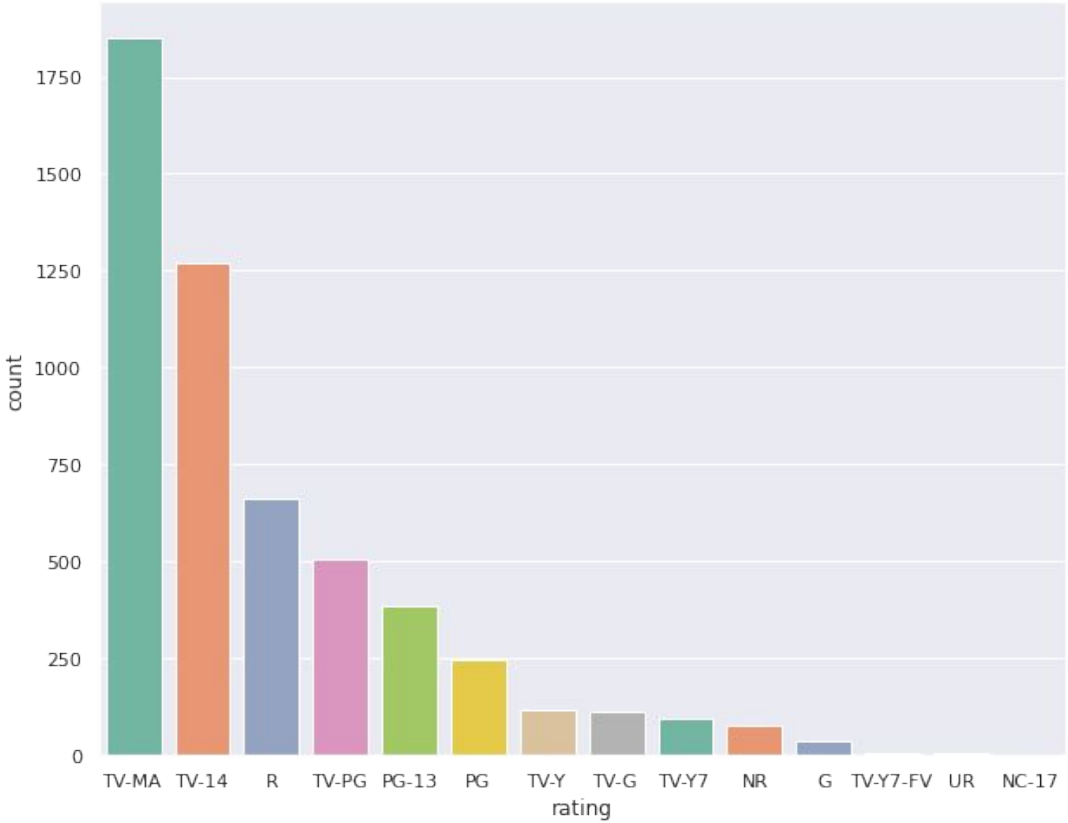
Country wise analysis



Top countries with highest content production.

- United States has the most number of content on Netflix.
- India has second highest content on Netflix.
- Australia and Taiwan has least number of content on Netflix.

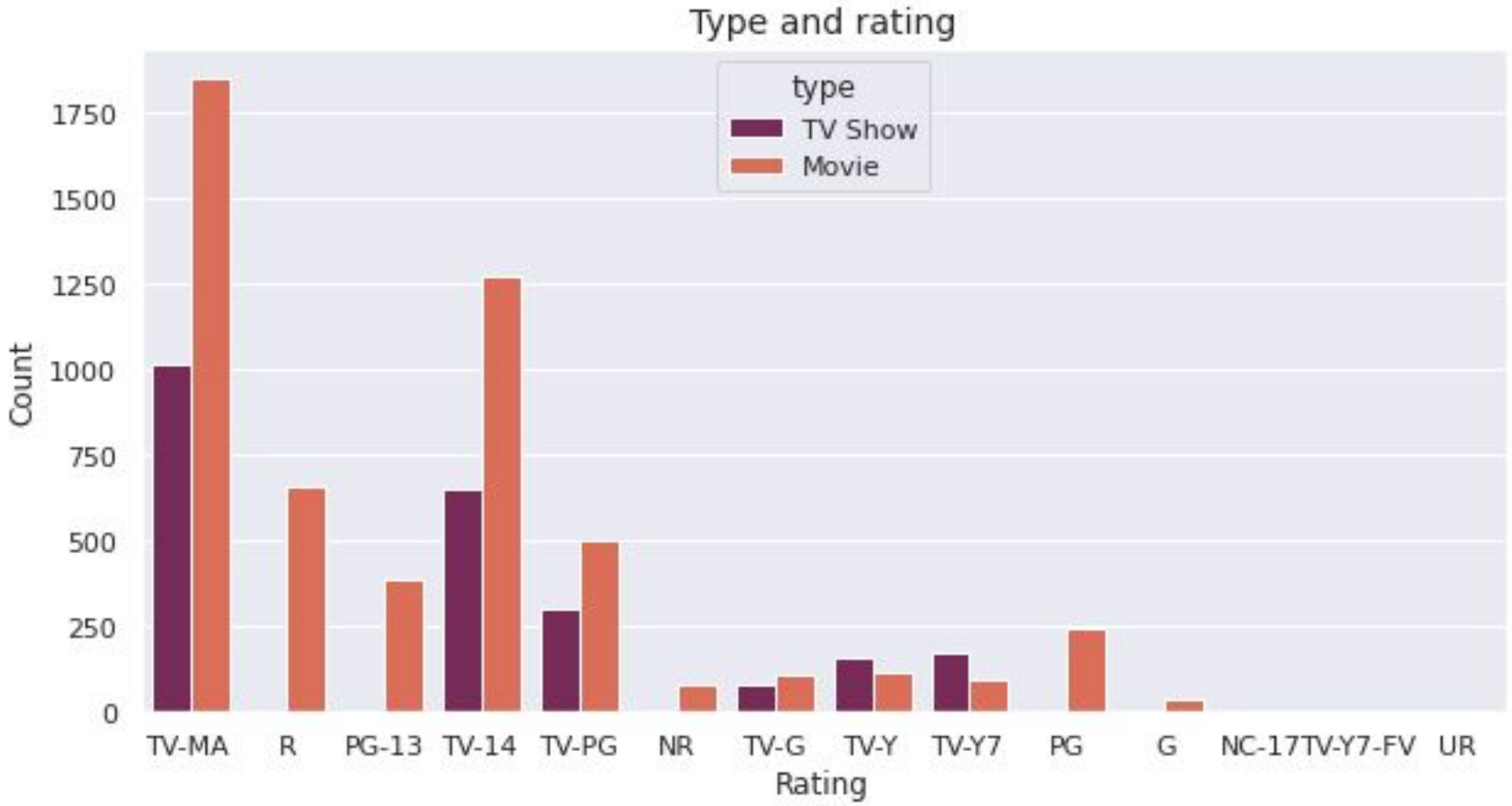
Rating wise analysis



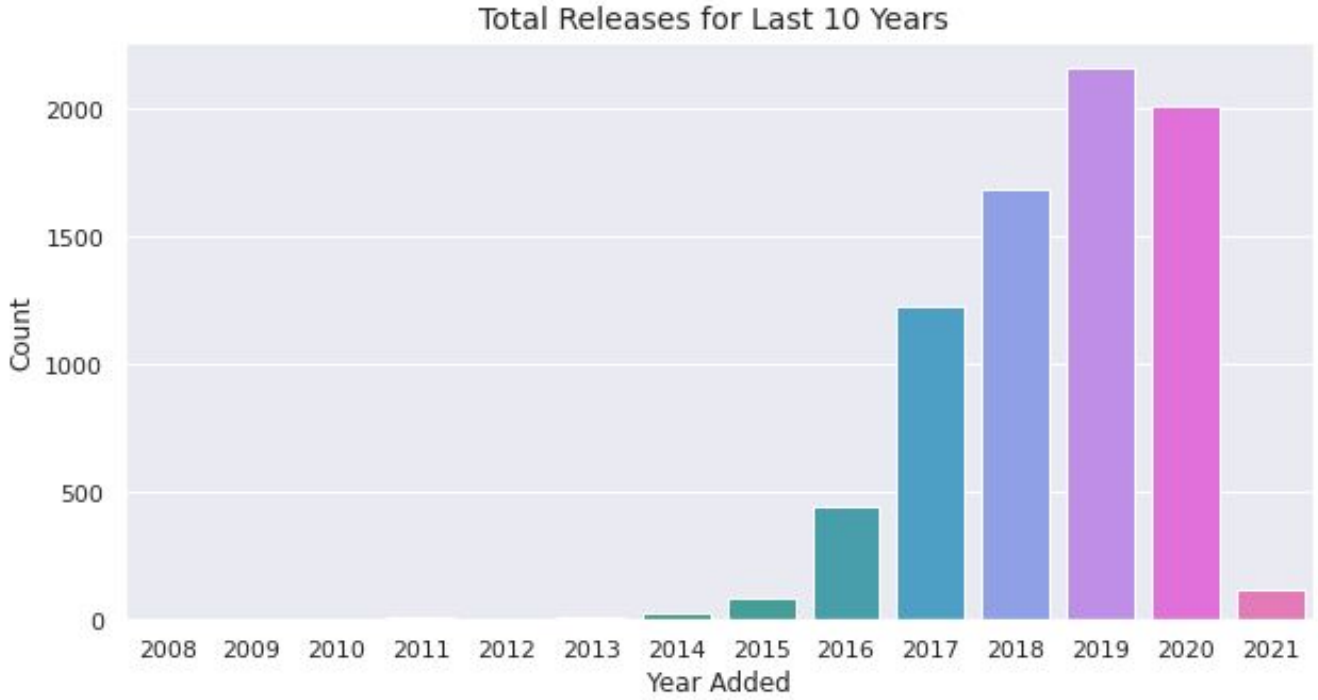
Most of the TV shows and Movies are of TV-MA rating which means most of them are for Mature Audience only.

This is followed by the shows which are strictly for the audience whose age is more than 14 years.

Type wise rating

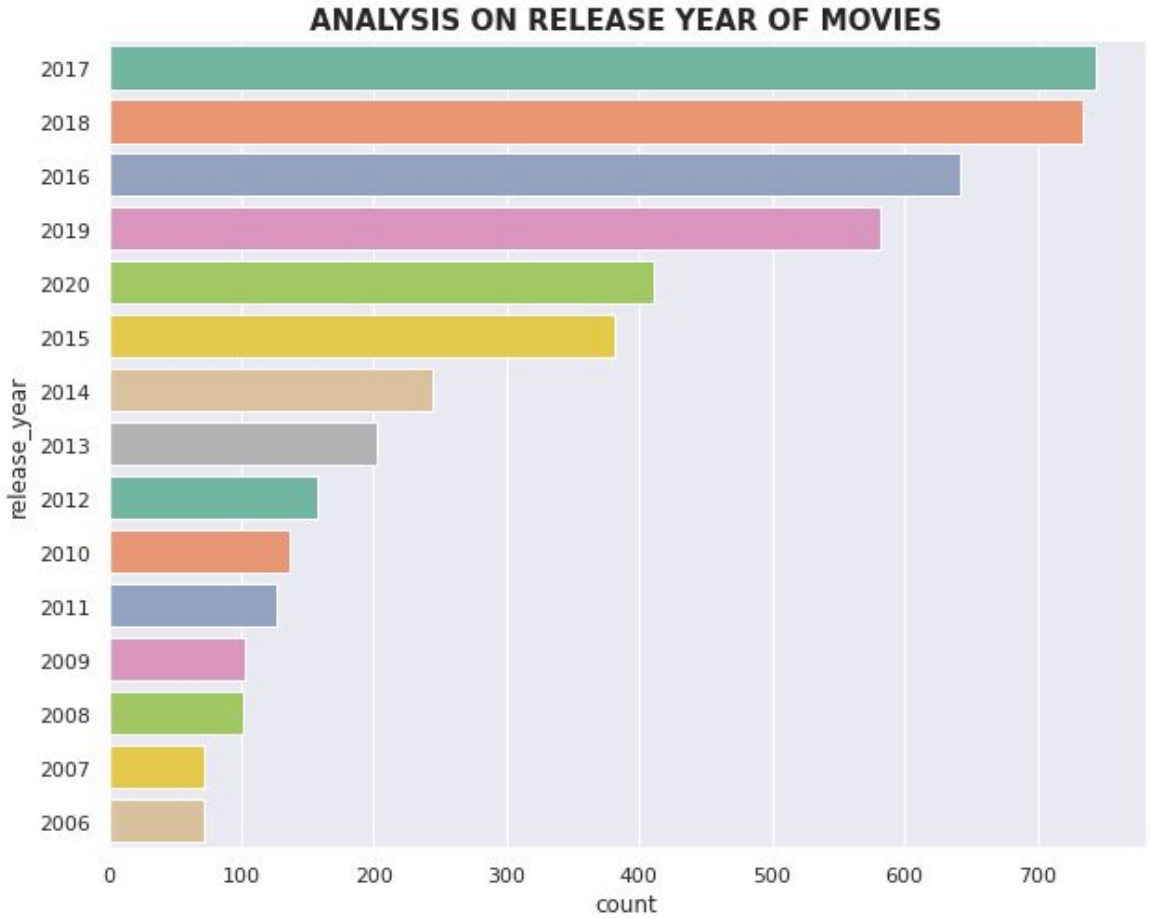


Year wise release date analysis



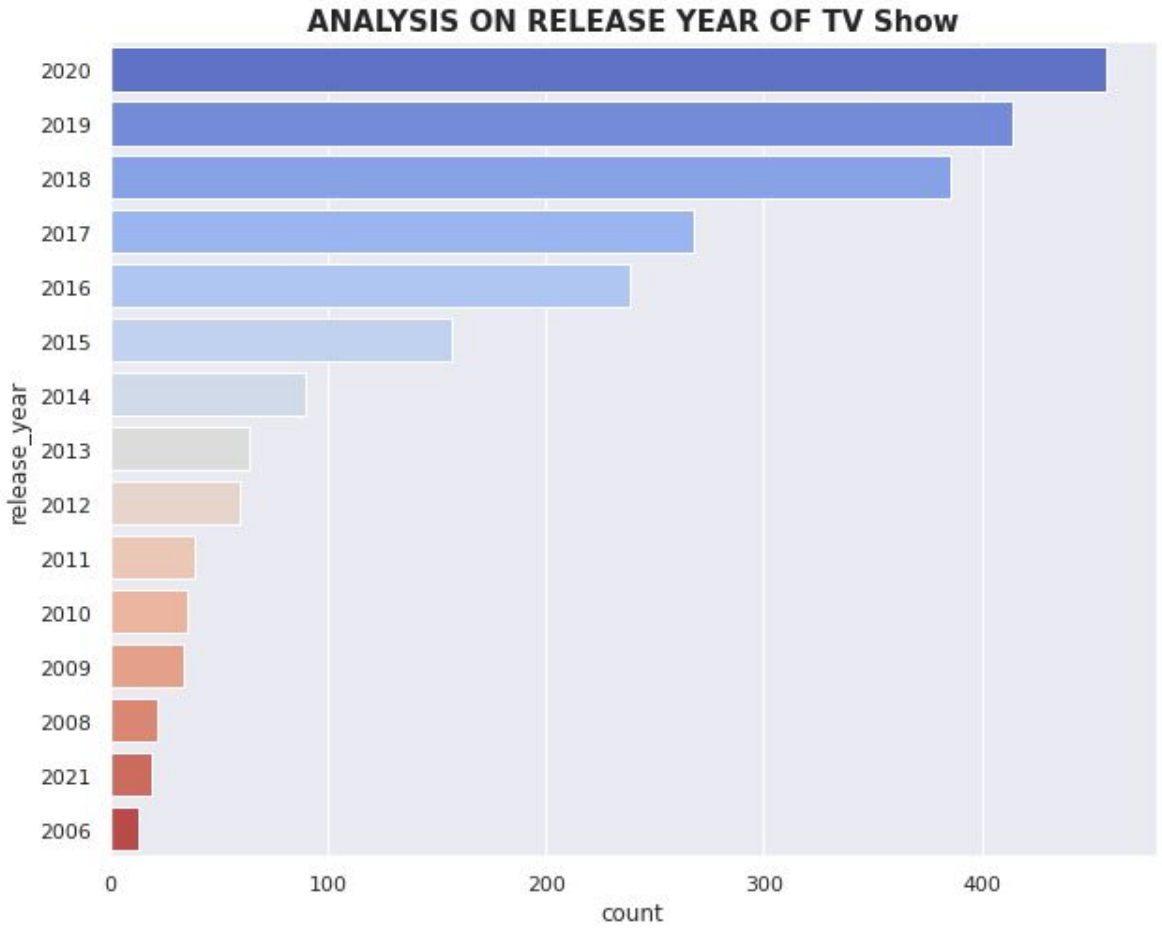
The number of release have significantly increased after 2015 and have dropped in 2021 because of Covid 19.

Movie release year wise analysis



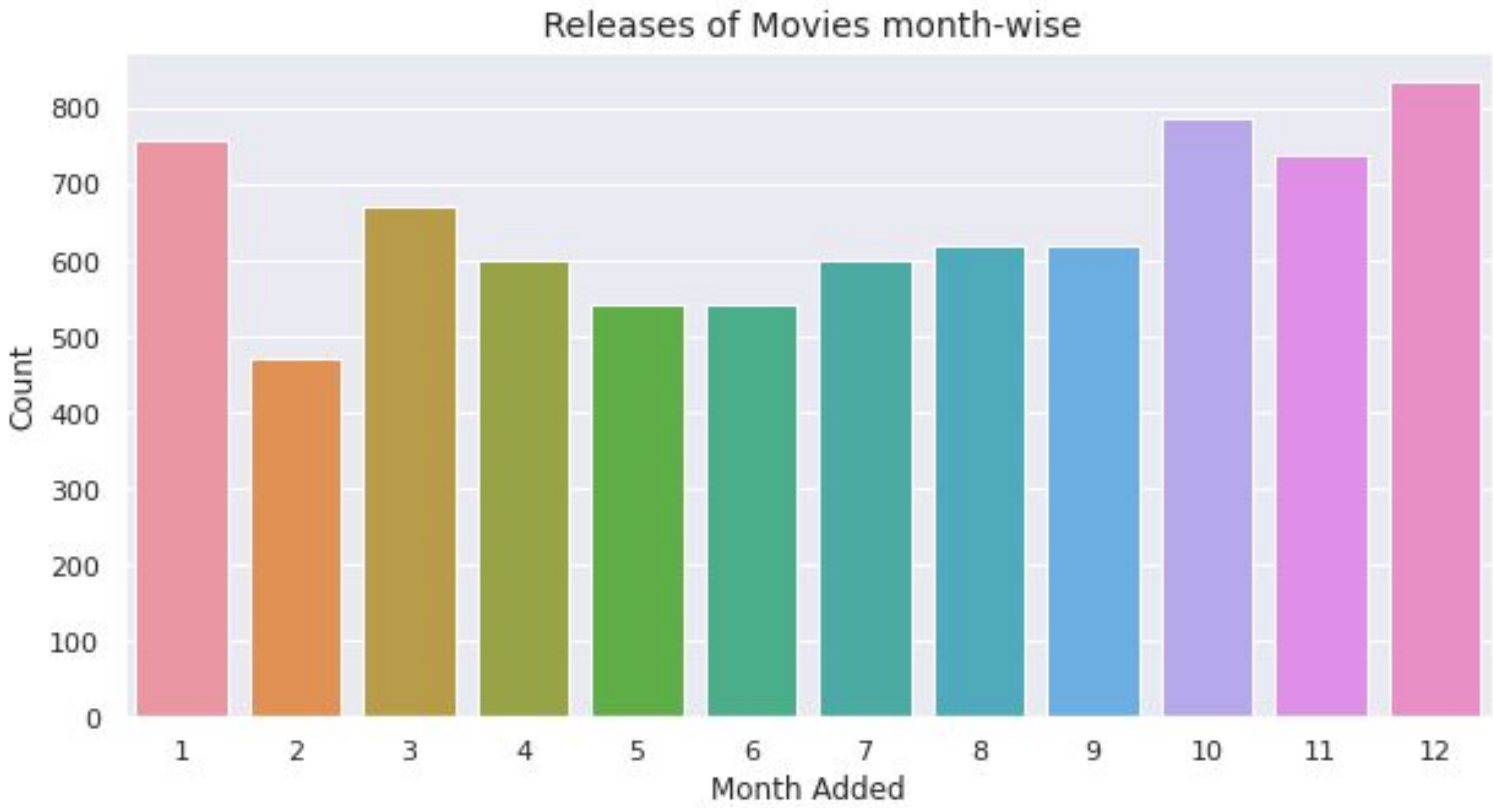
Most of the movies are released in 2017 followed by 2018

TV shows release year wise analysis



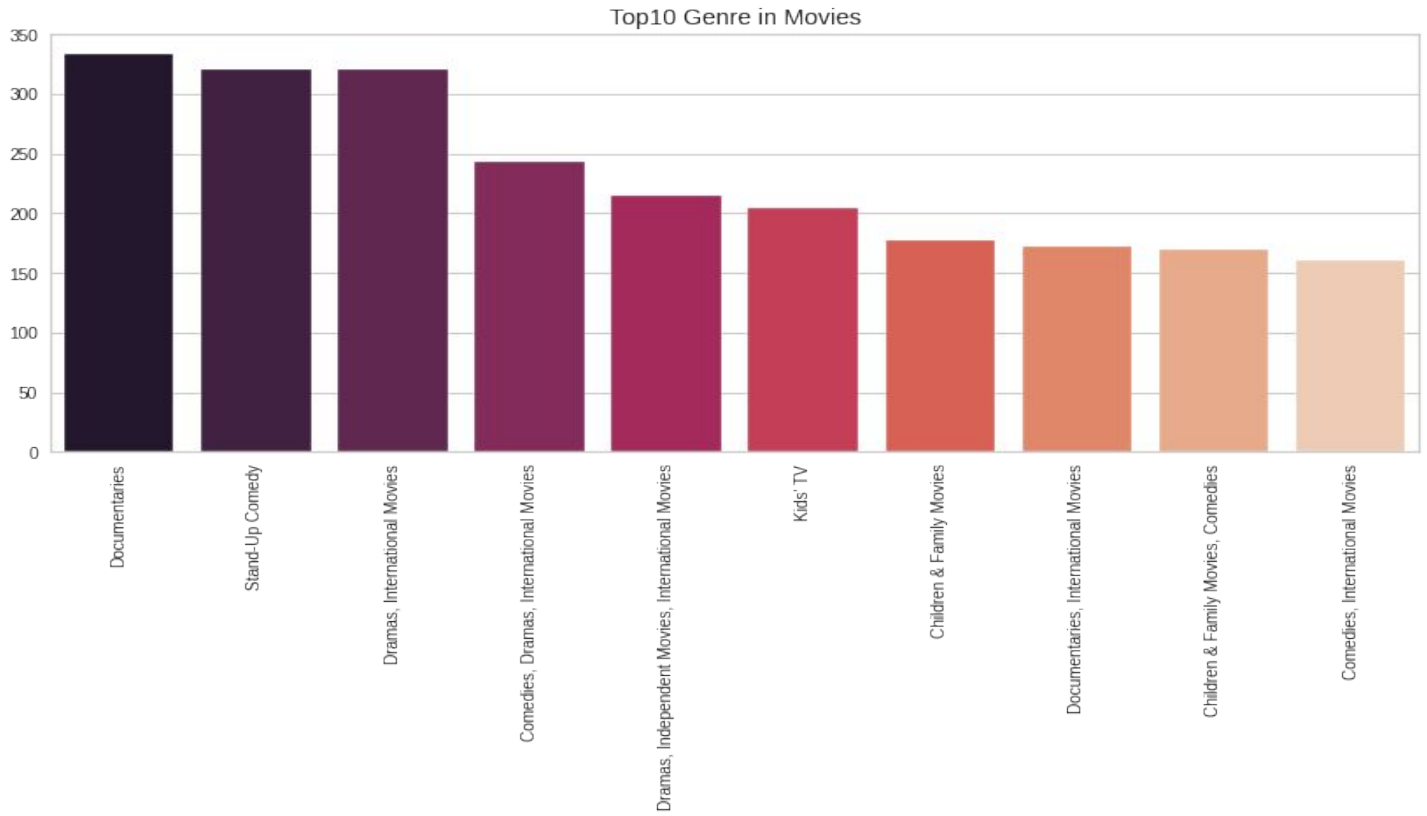
Most of the TV shows are released on 2020.

Month wise release date analysis



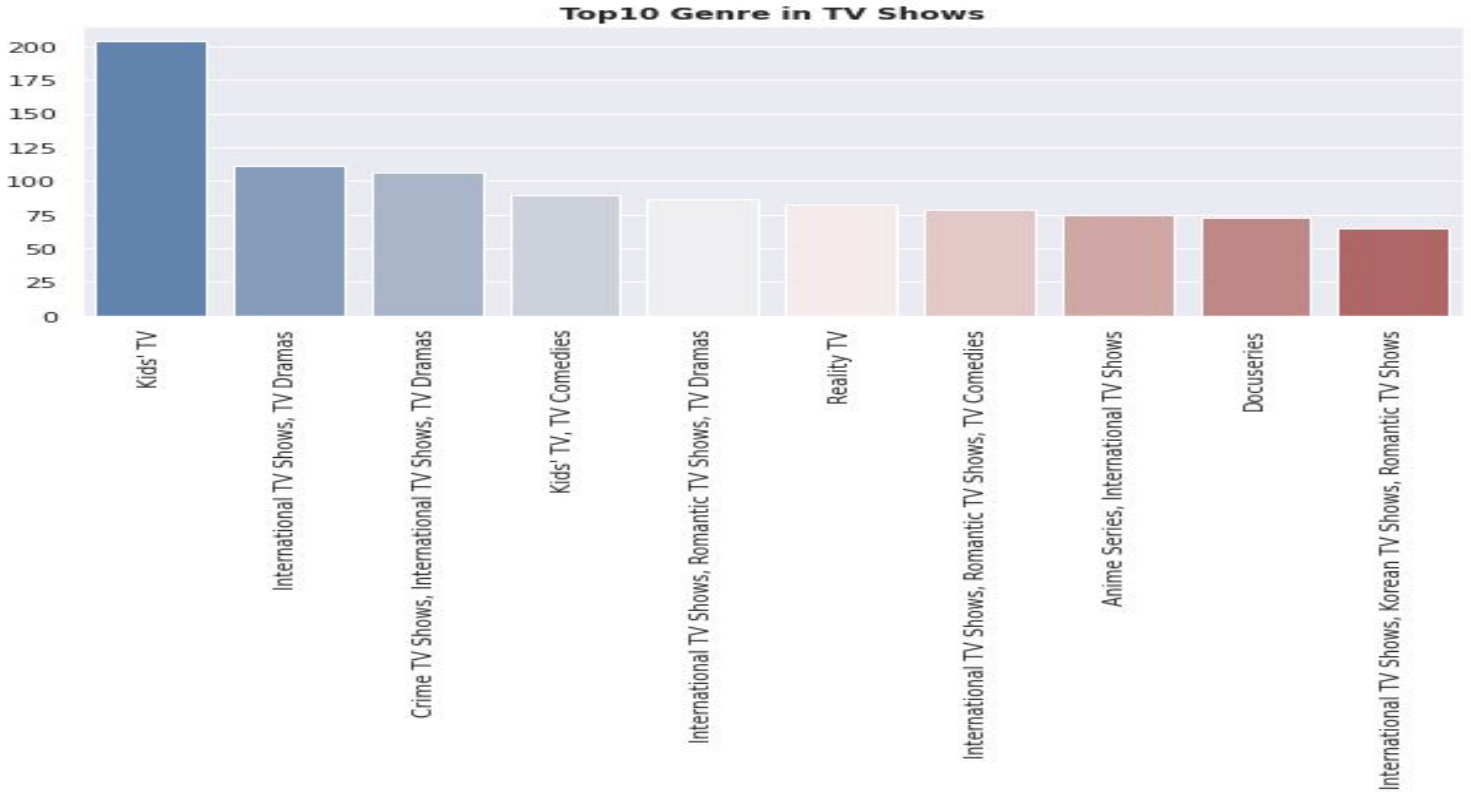
Most of the movies are released in December followed by October and January.

Top ten genres in movies



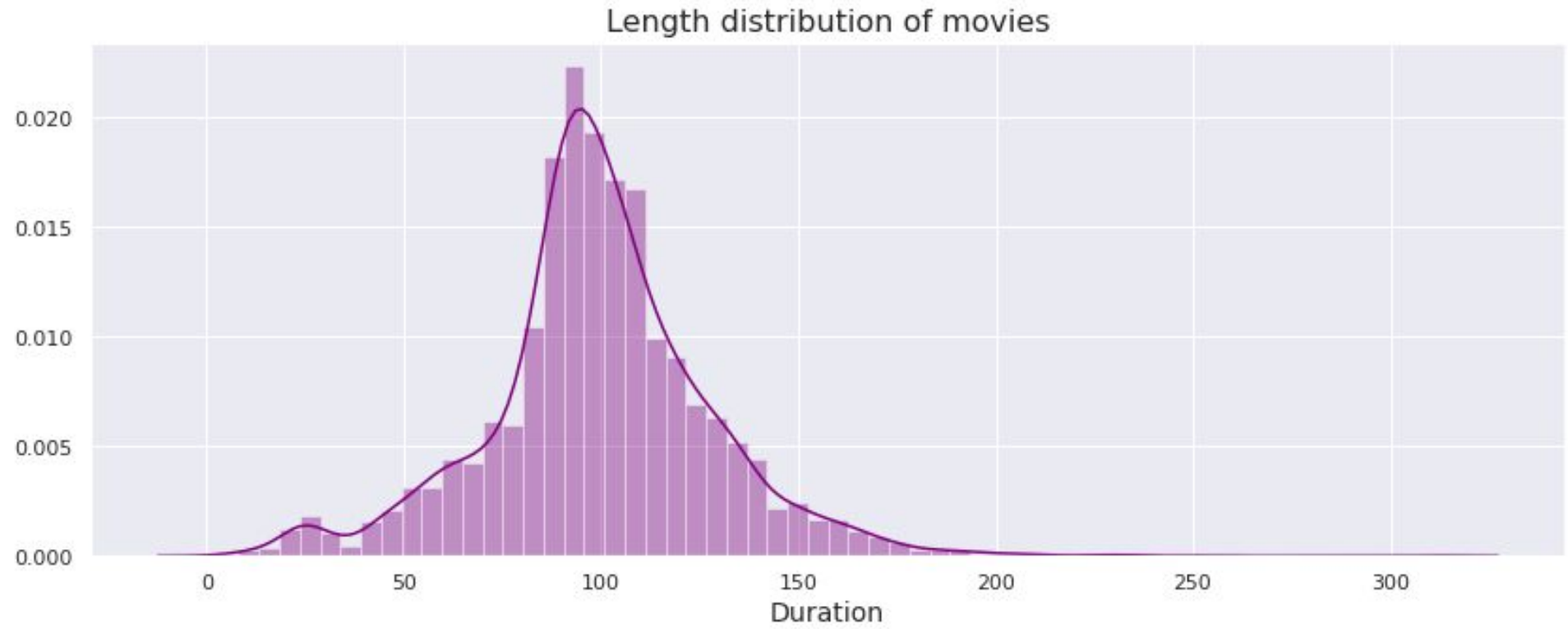
Most of the movies are of Documentary, Stand-Up Comedy and Drama genre.

Top ten genres in TV shows



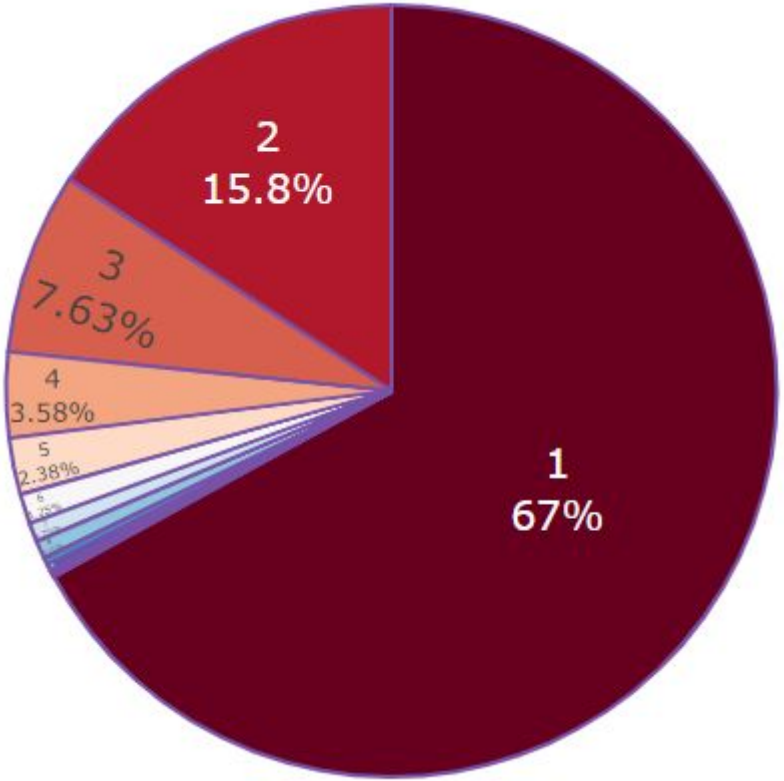
Most of the TV shows are of Kid's TV genre.

Movies length distribution



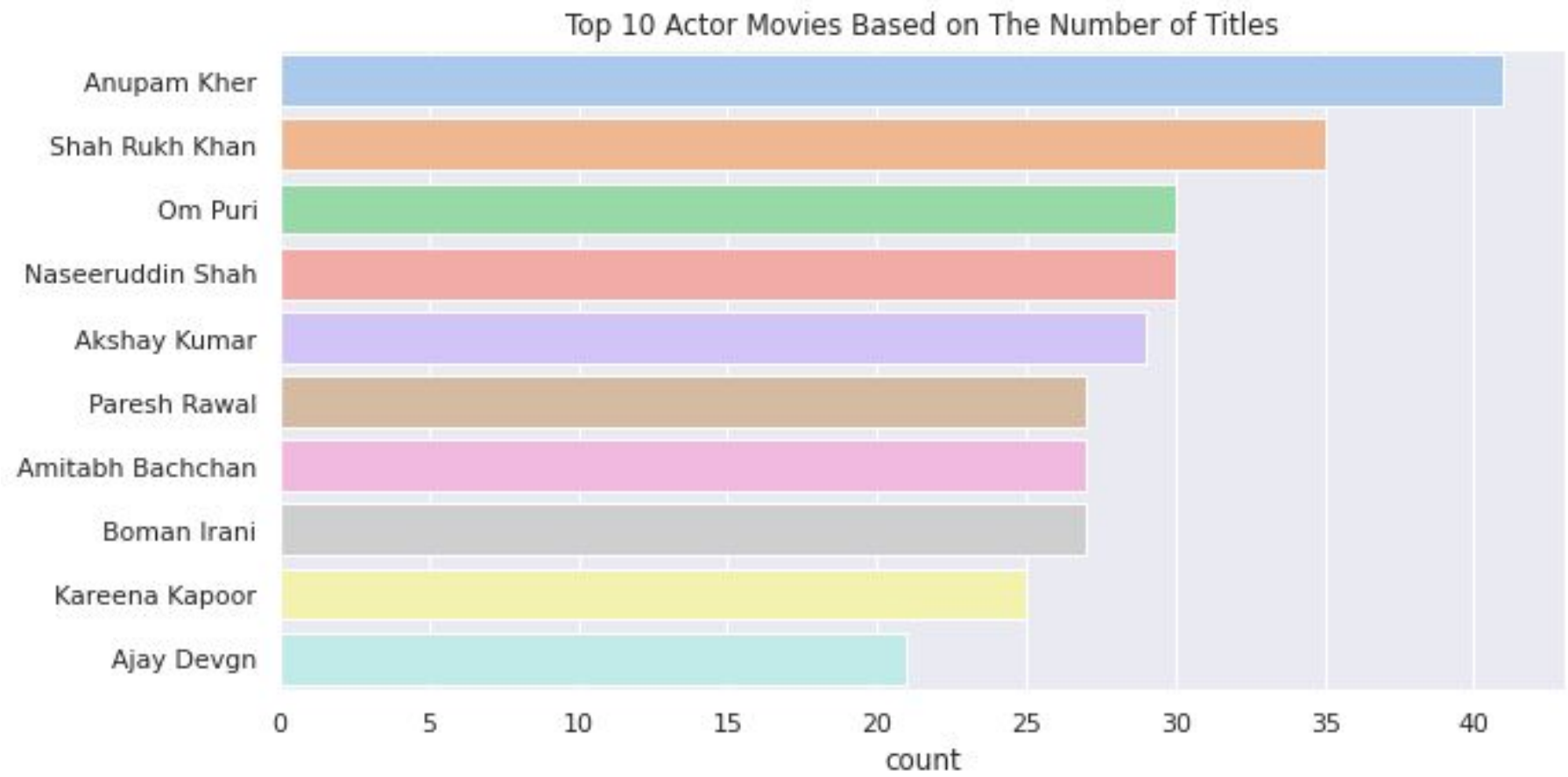
Most of the movies have length of 100 mins.

TV shows season wise distribution

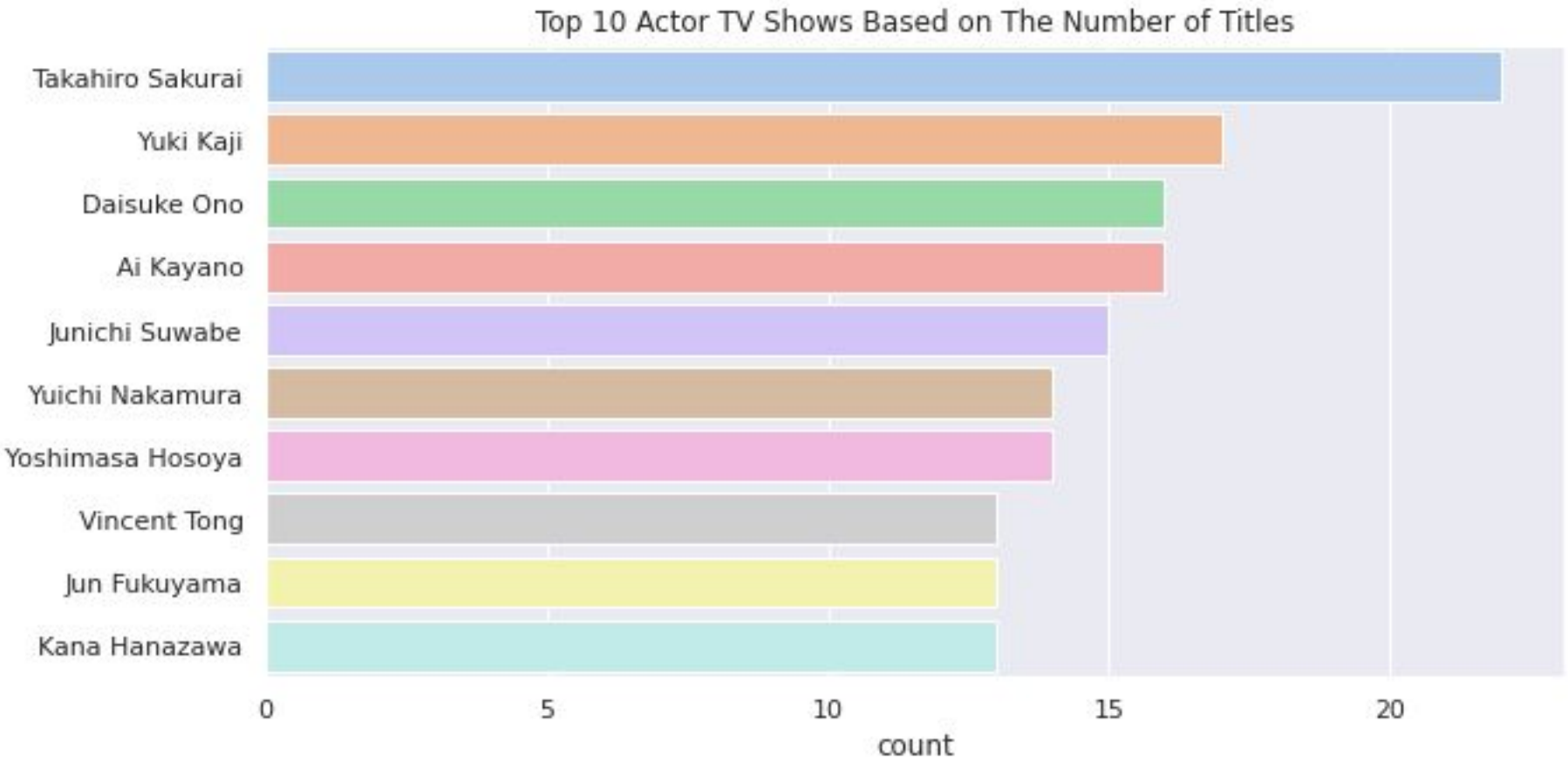


67% TV shows have only one season.
15.8% TV shows have two seasons.
7.63% TV shows have three seasons.

Top ten actors movie wise analysis



Top ten actors TV shows wise analysis



Data cleaning

- **Label encoding**

  We have used LabelEncoder in type, country, rating and listed_in columns. By this process the values of the columns each label is assigned a unique integer based on alphabetical ordering.

- **Standardization**

  We have used StandardScaler to transform the data

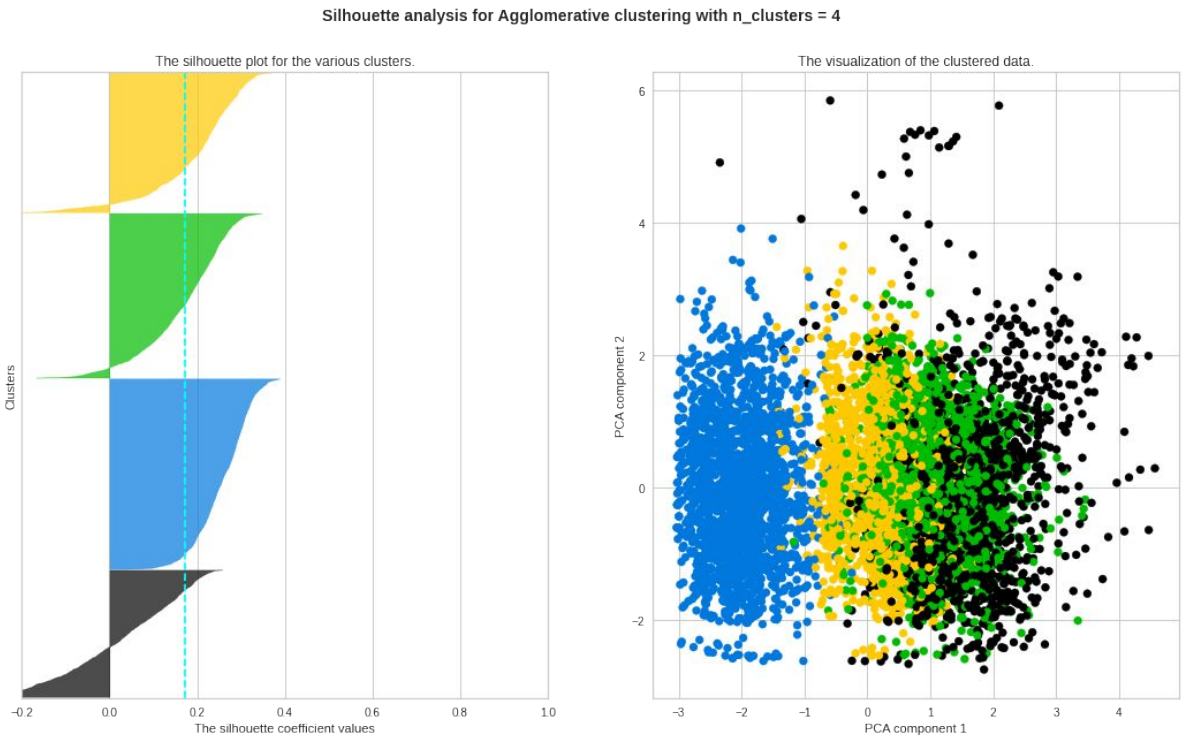
- **PCA**

  We have done PCA to show the clusters.

- **Min-Max Scaling**

  For each value in a feature, MinMaxScaler subtracts the minimum value in the feature and then divides by the range. It preserves shape of original distribution.

Model Implementation



Assume we cut vertical lines with a horizontal line to obtain the number of clusters.
Number of clusters = 4
The average silhouette_score is : 0.17296314851287742

Agglomerative Clustering

K-Means

AI

To process the learning data, the K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids

It halts creating and optimizing clusters when either:

- The centroids have stabilized — there is no change in their values because the clustering has been successful.
- The defined number of iterations has been achieved

K-means clustering

AI

K-means algorithm is an iterative algorithm that tries to partition the dataset into K pre defined distinct non overlapping subgroups where each data point belongs to only one group.

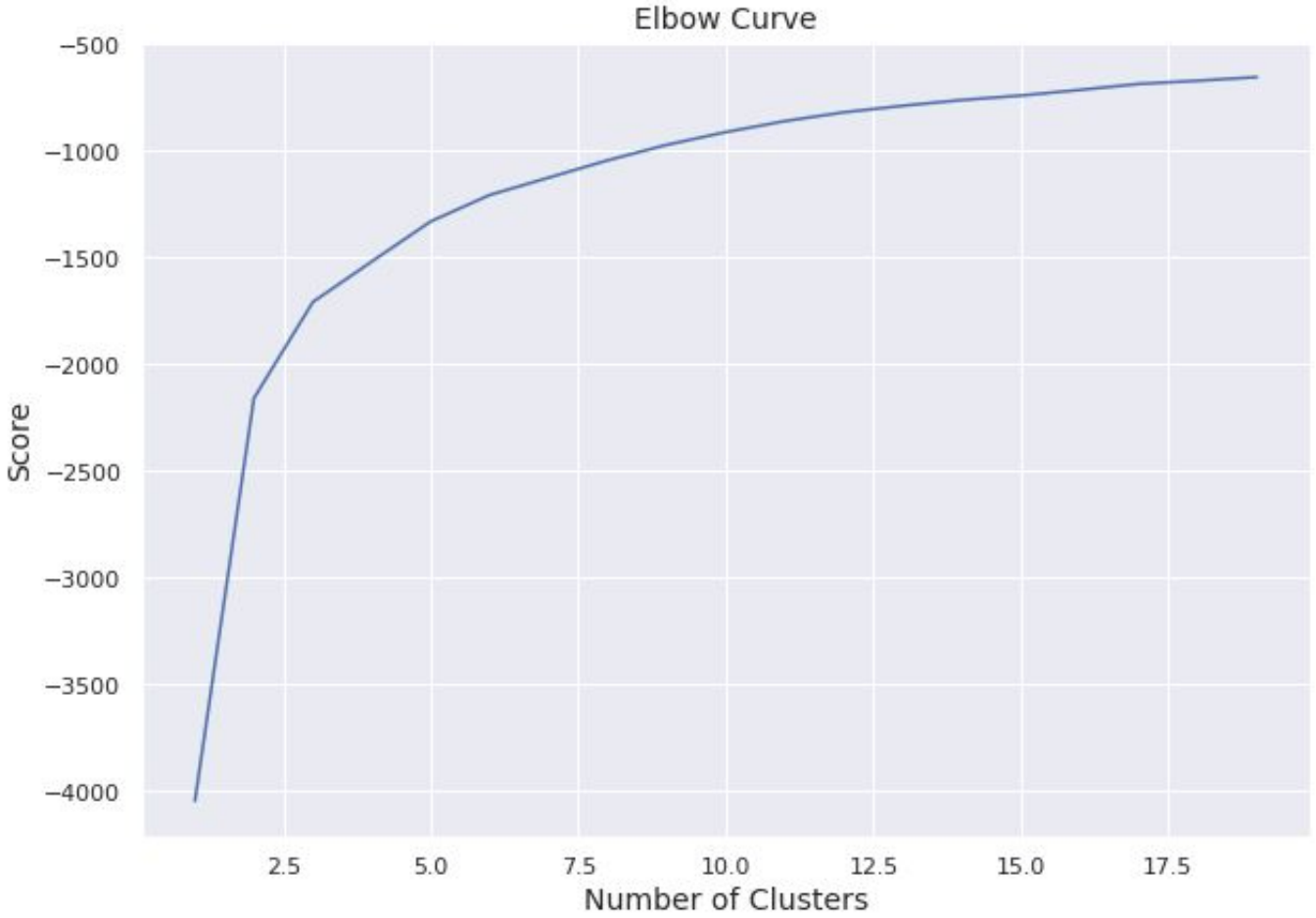
1. Elbow Curve:

- The Elbow Curve is one of the most popular methods to determine this optimal value of k.
- The elbow curve uses the sum of squared distance (SSE)to choose an ideal value of k based on the distance between the data points and their assigned clusters.

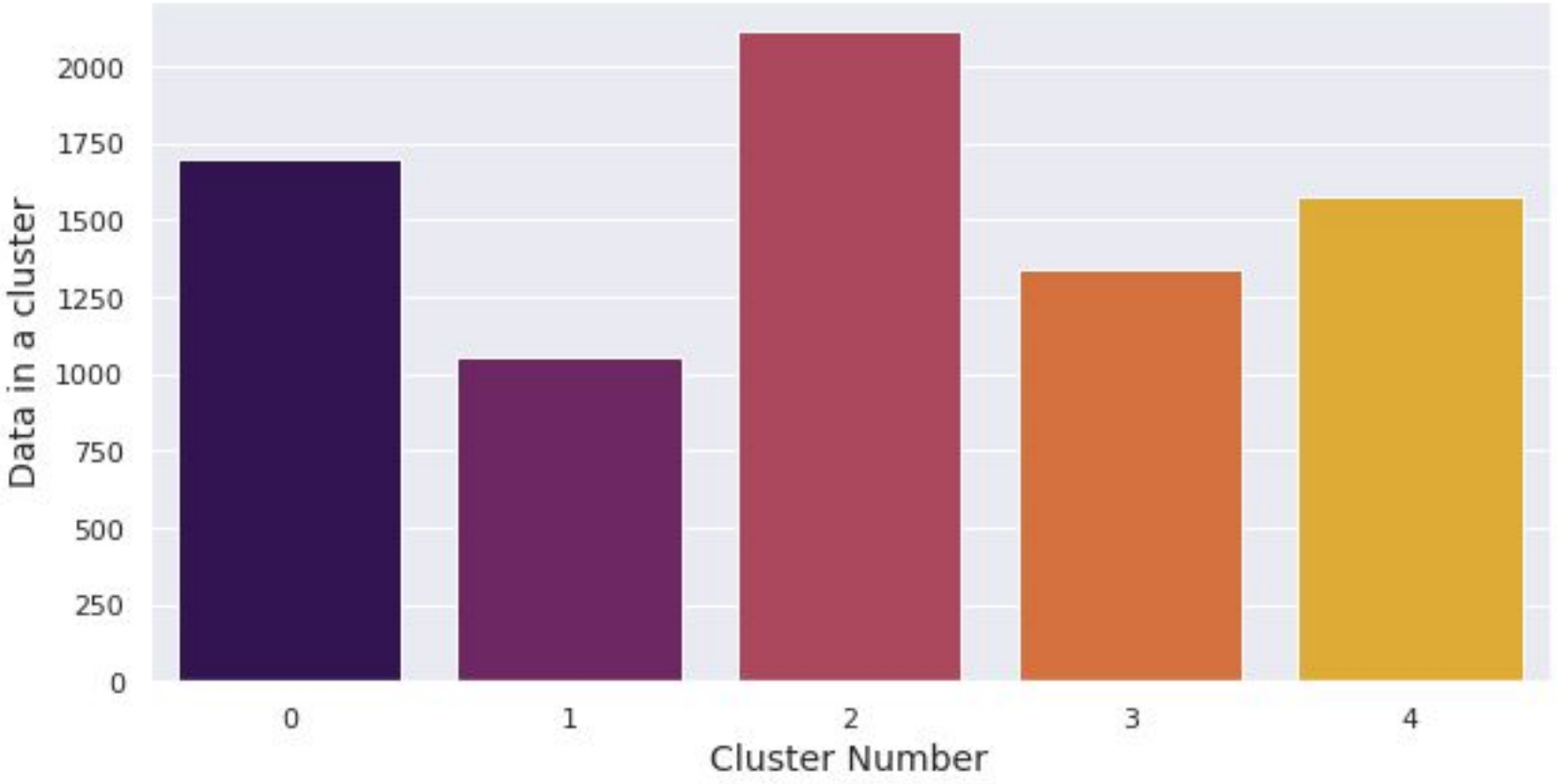
2. Silhouette score :

- Silhouette score is used to evaluate the quality of clusters created using clustering algorithms such as K Means in terms of how well samples are clustered with other samples that are similar to each other.

Elbow curve

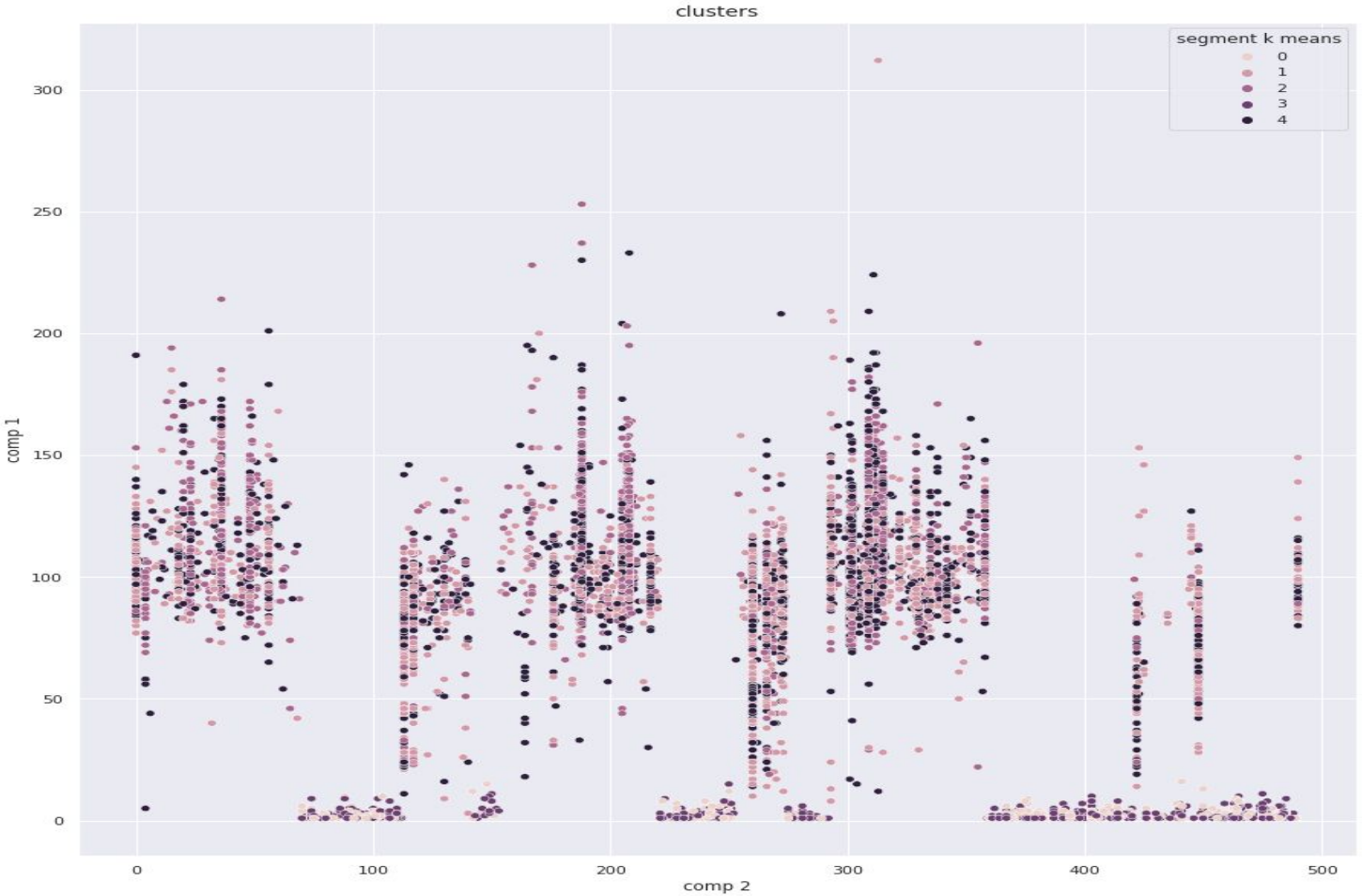

Elbow Curve
−500
−1000
−1500
−2000
−2500
−3000
−3500
−4000
Score
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Number of Clusters

Clusters

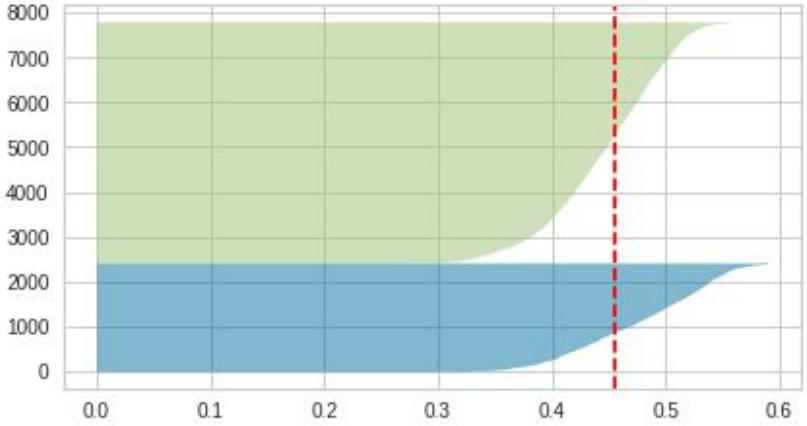


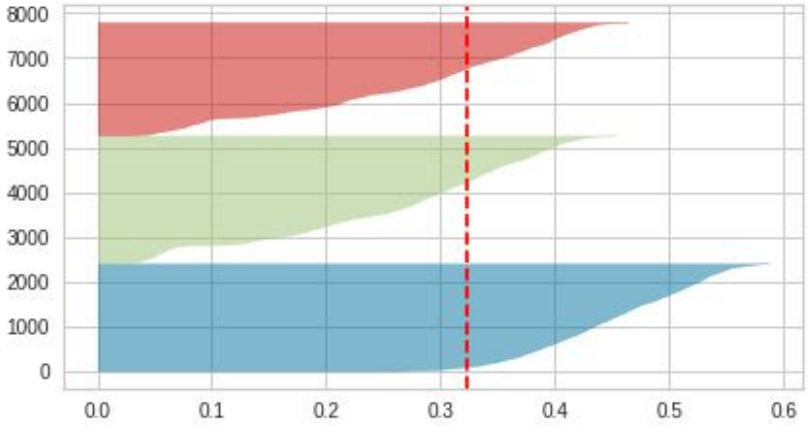
We clearly see that one cluster is the largest and one cluster has the fewest number of movies.

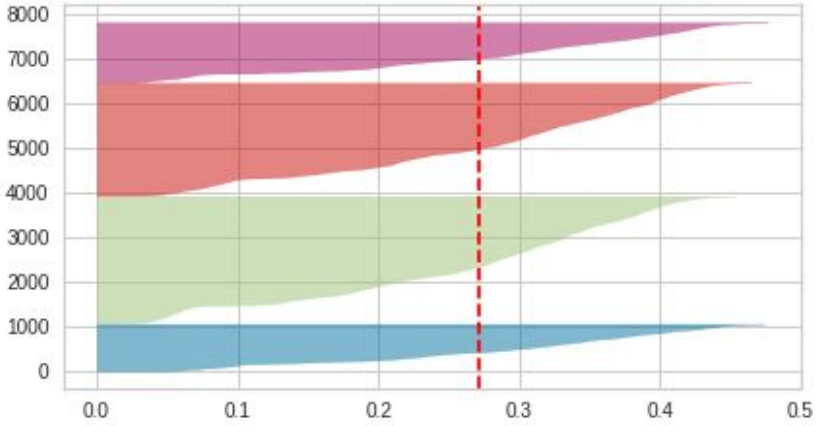
Grouped Clusters

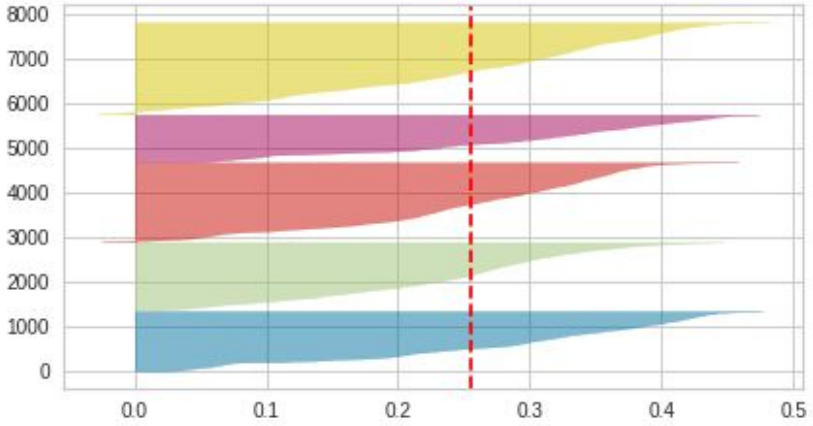


Silhouette Score









Conclusion

- Data set contains 7787 rows and 12 columns in that cast and director features contains large number of missing values so we can drop it.
- We have two types of content TV shows and Movies (30.9% contains TV shows and 69.1% contains Movies).
- Most films were released in the years 2018, 2019, and 2020 and united states have the maximum content on Netflix.
- The months of October, November, December and January had the largest number of films and TV-shows released.
- The USA, India, the United Kingdom, Canada, and Egypt are the top five producer countries.
- For the clustering algorithm, we utilized type, nation, released year, genre, and year.
- Applied different clustering models Kmeans, Hierarchical clustering on data we got the best cluster arrangements.
- We cut vertical lines with a horizontal line to obtain the number of clusters in Agglomerative Clustering. There were four clusters, with an average silhouette score of 0.17296314851287742.
- The final model we used was k-means clustering, which consisted of 2,3,4,5,6 clusters. 4 numbers of clusters gives us good fitting.
- After applying K - means optimal value of number of clusters is 5
- Silhouette score for a set of sample data points is used to measure how dense and well-separated the clusters are.

Thank You