Capstone Project-4

NETFLIX MOVIES AND TV SHOWS CLUSTERING

Pankaj B. Gadge

Point for Discussion

- ❑ Problem Statement
- ❑ Introduction
- ❑ Dataset and Variable Information
- ❑ Dataset Processing
- ❑ Data Encoding
- ❑ EDA
- ❑ Implementation
    - Principal Component Analysis (PCA)
    - K Means Clustering
    - Hierarchical clustering
- ❑ Conclusion

Problem Statement

This dataset consists of tv shows and movies available on Netflix as of 2019. 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.

Understanding what type content is available in different countries

Is Netflix has increasingly focusing on TV rather than movies in recent years.

Introduction

We all watch a lot of TV shows.

Many online streaming services offer a large number of TV shows, which are at our disposal to watch, at the price of a subscription cost. The major online streaming services across the world are Netflix, Prime Video, Hulu, and Disney+. Netflix is a popular entertainment service used by people around the world.

Dataset

The dataset is collected from Flixable which is a third-party Netflix search engine. Integrating this dataset with other external datasets such as IMDB ratings, rotten tomatoes can also provide many interesting findings. It includes over 6,234 entries and 12 columns.

Dataset Variable Information

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

Dataset Variable Information

cast : Actors involved in the movie / show

country : Country where the movie / show was produced

date_added : Date it was added on Netflix

Dataset Variable Information

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 : Genere

description: The Summary description

Dataset Inspection & Processing

After data inspection we found that there are 6,234 entries and 12 columns to work with for this EDA. Few columns that contain null values, “director,” “cast,” “country,” “date_added,” “rating.”

NaN Values Processing

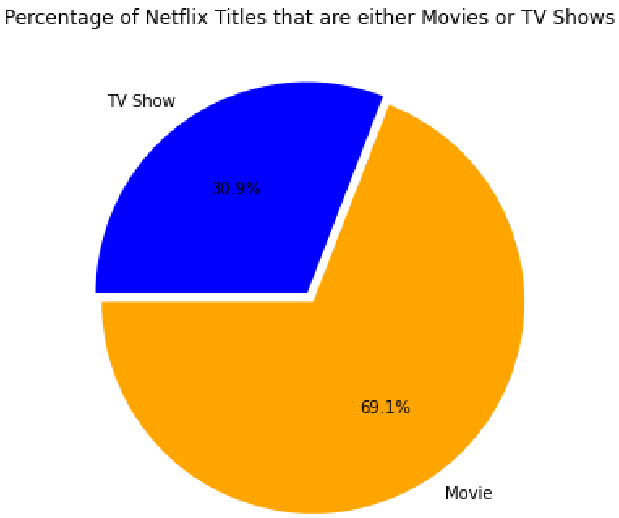
The easiest way to get rid of them would be to delete the rows with the missing data for missing values. However, this wouldn’t be beneficial to our EDA since it is a loss of information. Since “director,” “cast,” and “country” contain the majority of null values, we chose to treat each missing value is unavailable. The other two label “date_added” and “rating” contain an insignificant portion of the data, so it drops from the dataset. Finally, we can see that there are no more missing values in the data frame.

show_id	0	cast	718	rating	7
type	0	country	507	duration	0
title	0	Date_added	10	Listed_in	0
director	2389	Release_year	0	description	0

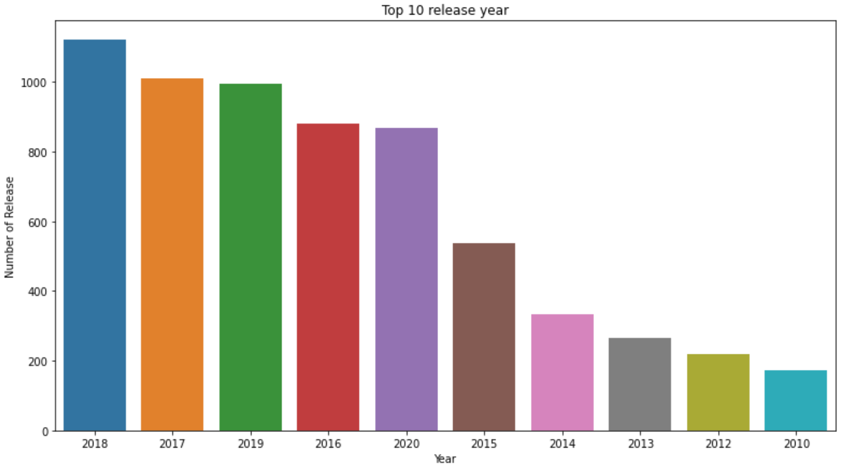
Data Encoding

Label Encoding refers to converting the labels into a numeric form so as to convert them into the machine-readable form. Machine learning algorithms can then decide in a better way how those labels must be operated. It is an important pre-processing step for the structured dataset in supervised learning. LabelEncoder can be used to normalize labels. It can also be used to transform non-numerical labels (as long as they are hashable and comparable) to numerical labels. Here we convert type,country,target_ages,rating,listed_in into numerical values using labelencoder.

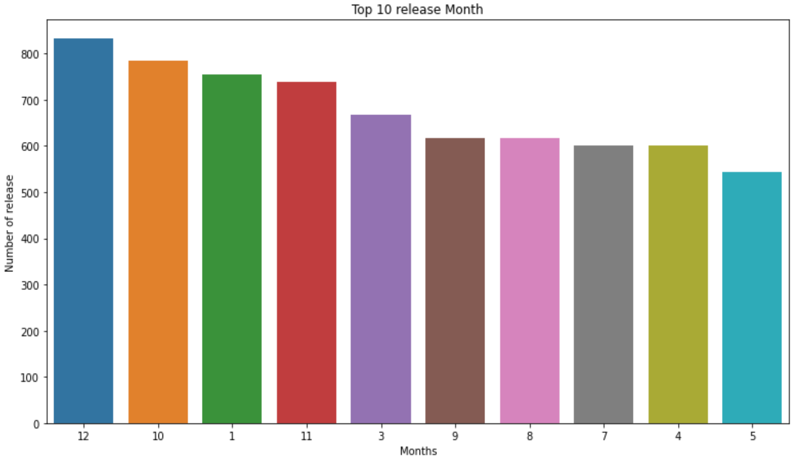
EDA



Percentage of Netflix Titles that are either Movies or TV Shows
TV Show
30.9%
69.1%
Movie

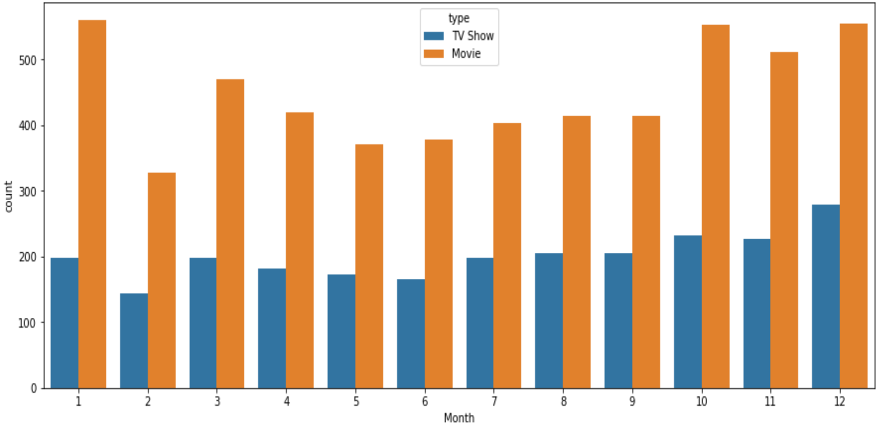


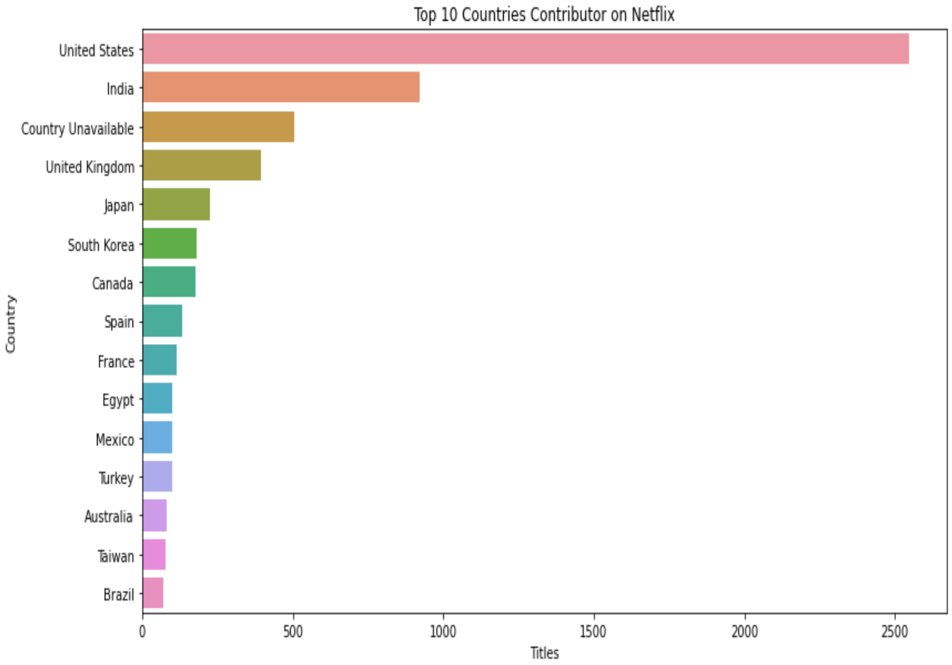
Top 10 release year
Number of Release
0
200
400
600
800
1000
2018
2017
2019
2016
2020
2015
2014
2013
2012
2010
Year

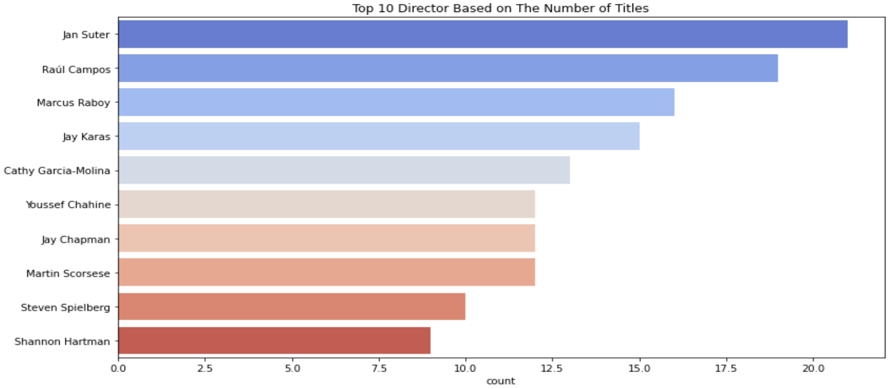


Top 10 release Month
Number of release
0
100
200
300
400
500
600
700
800
12
10
1
11
3
9
8
7
4
5
Months

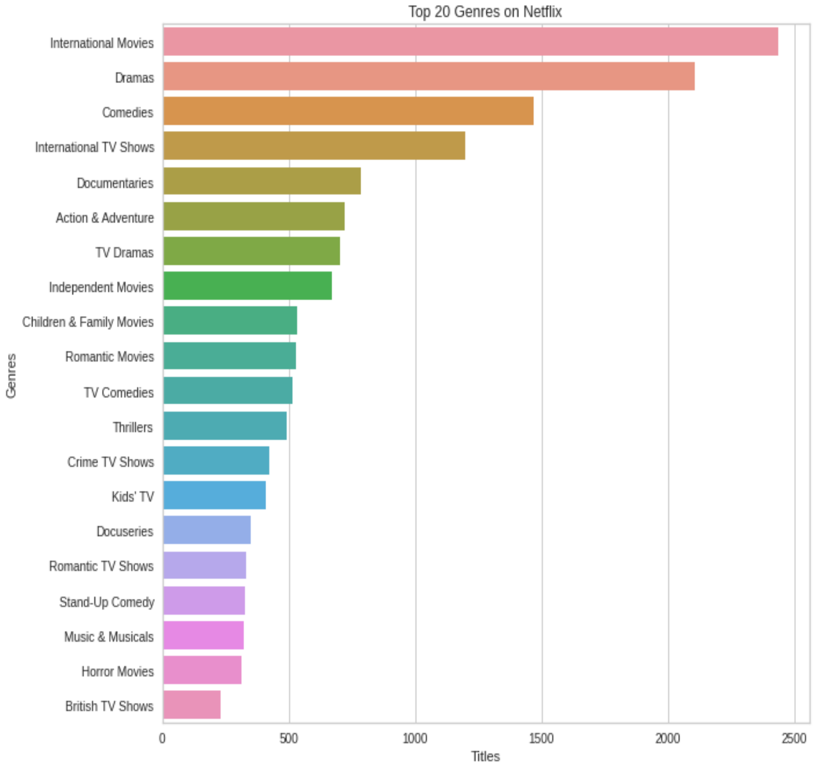
EDA

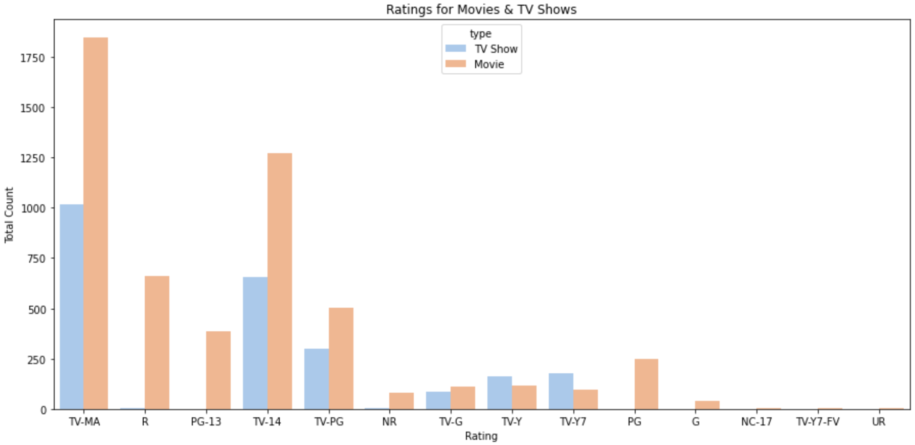


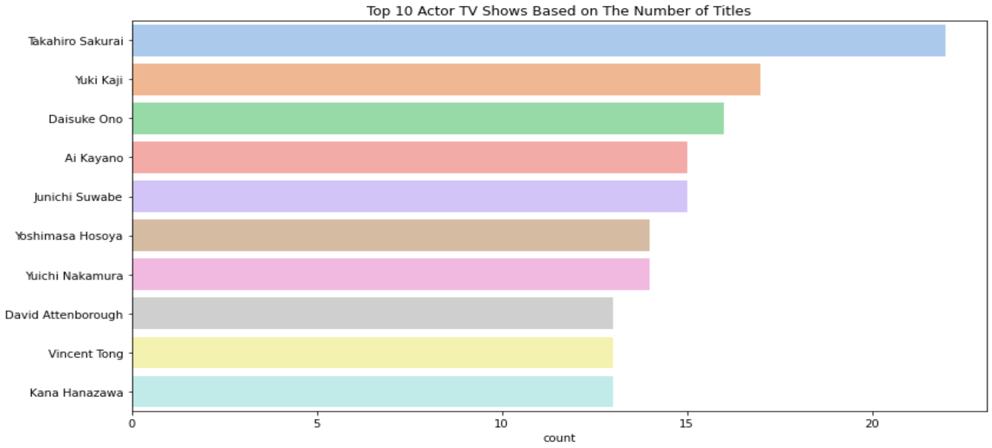

Top 10 Countries Contributor on Netflix


Top 10 Director Based on The Number of Titles

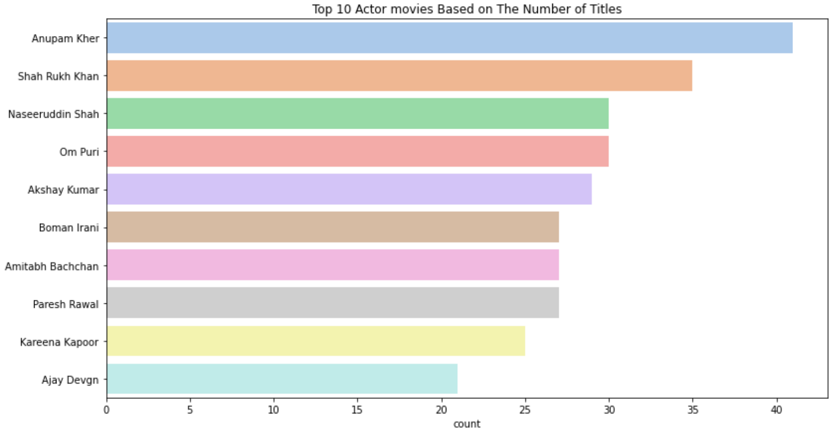
EDA

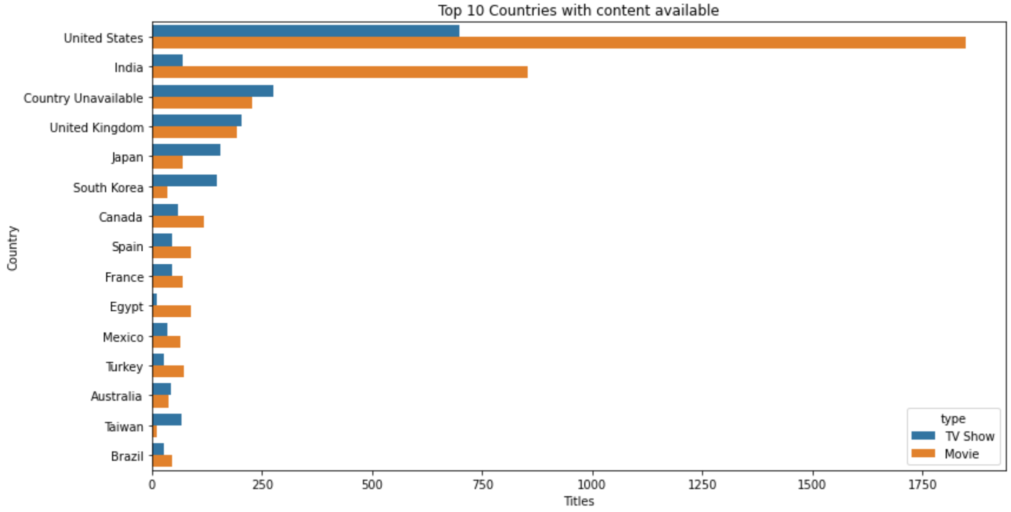

Top 20 Genres on Netflix

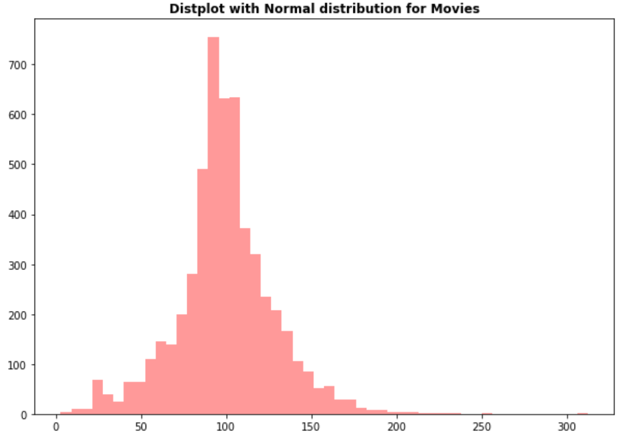

Ratings for Movies & TV Shows


Top 10 Actor TV Shows Based on The Number of Titles

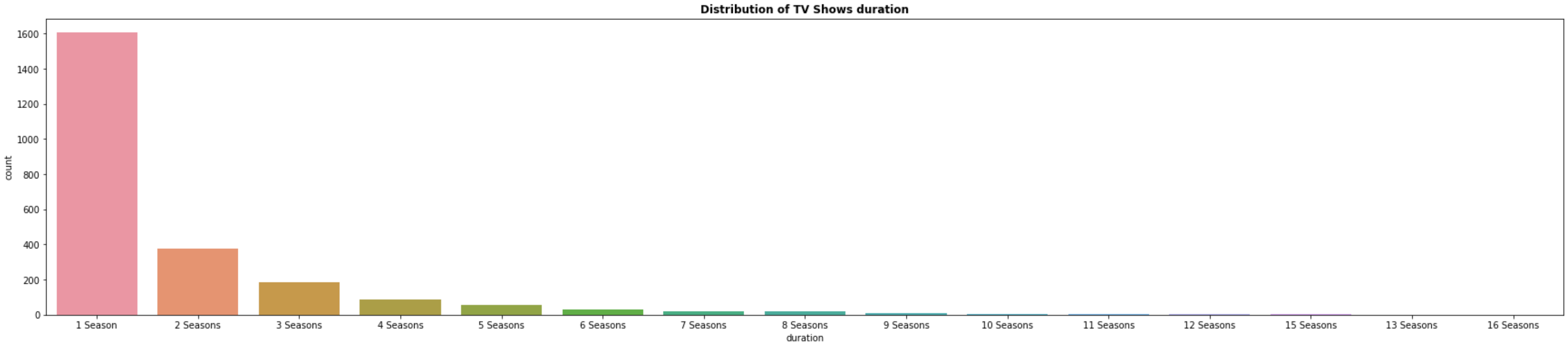
EDA


Top 10 Actor movies Based on The Number of Titles

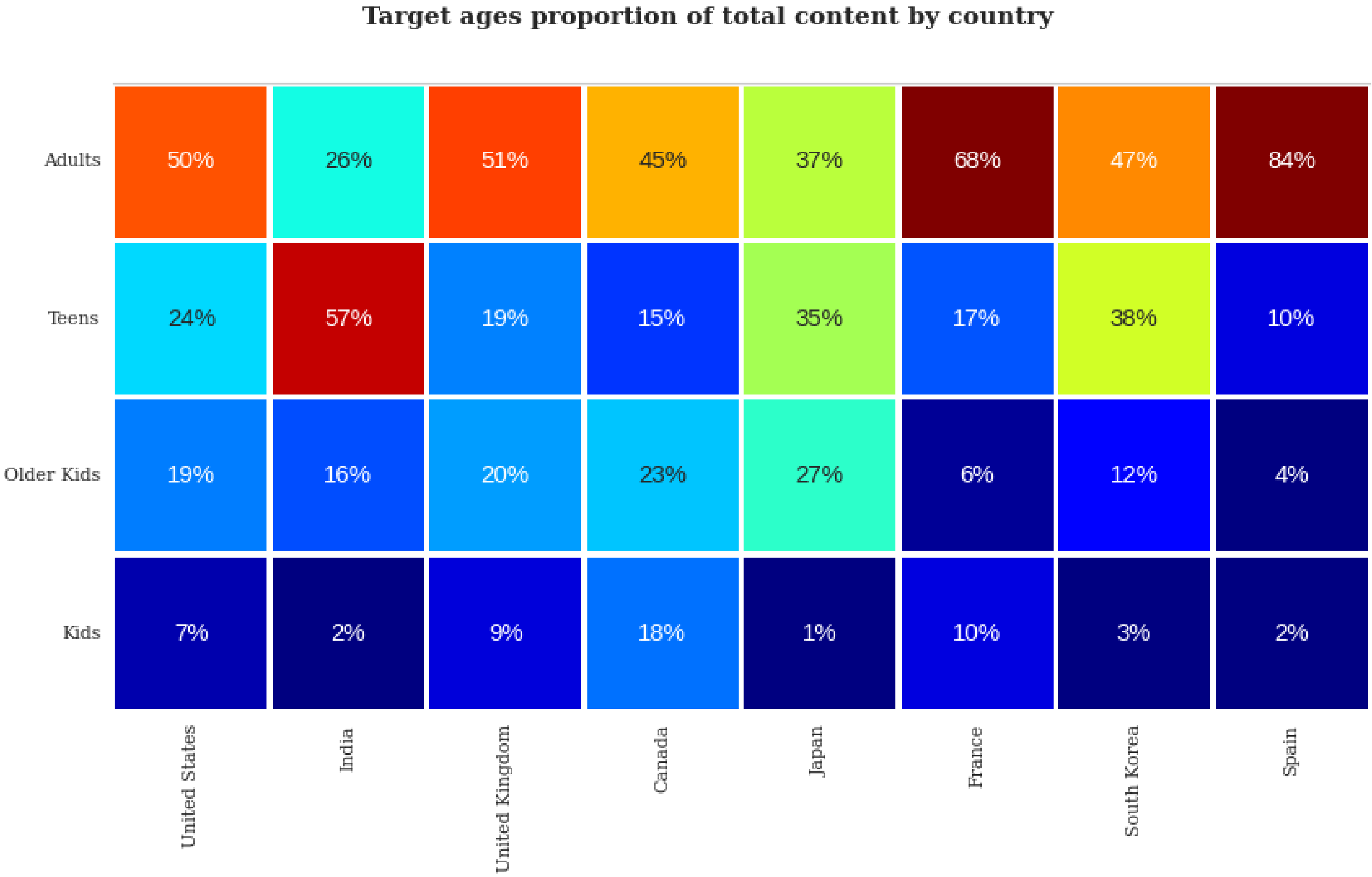

Top 10 Countries with content available


Distplot with Normal distribution for Movies

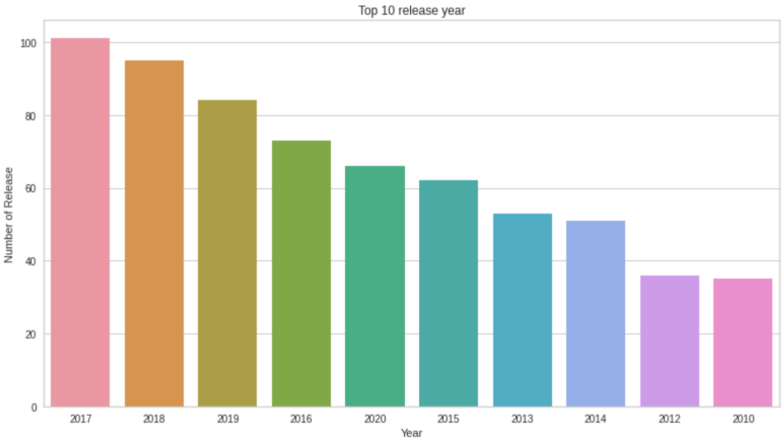
EDA

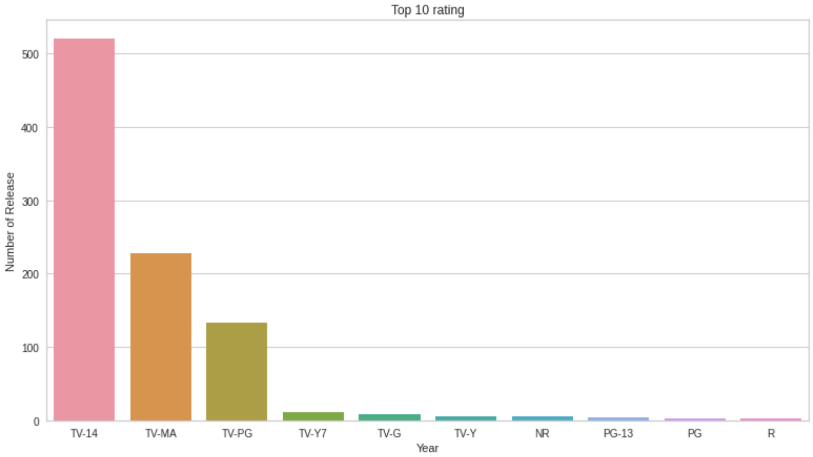

Distribution of TV Shows duration
count
1600
1400
1200
1000
800
600
400
200
0
1 Season
2 Seasons
3 Seasons
4 Seasons
5 Seasons
6 Seasons
7 Seasons
8 Seasons
9 Seasons
10 Seasons
11 Seasons
12 Seasons
15 Seasons
13 Seasons
16 Seasons
duration

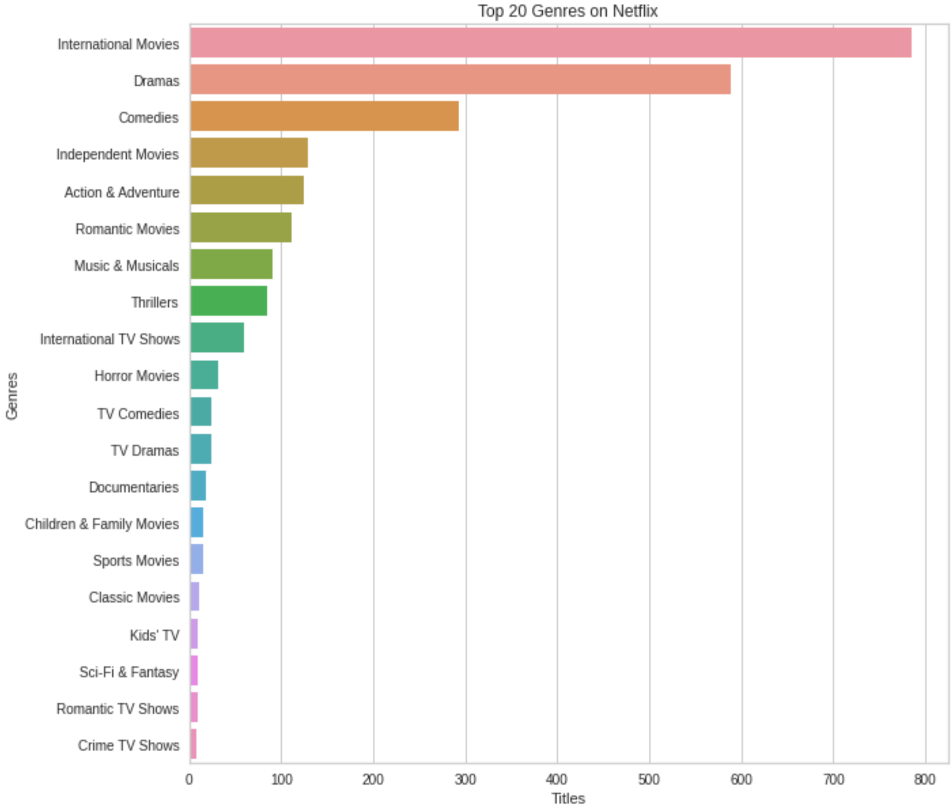
EDA (Correlation Heatmap)



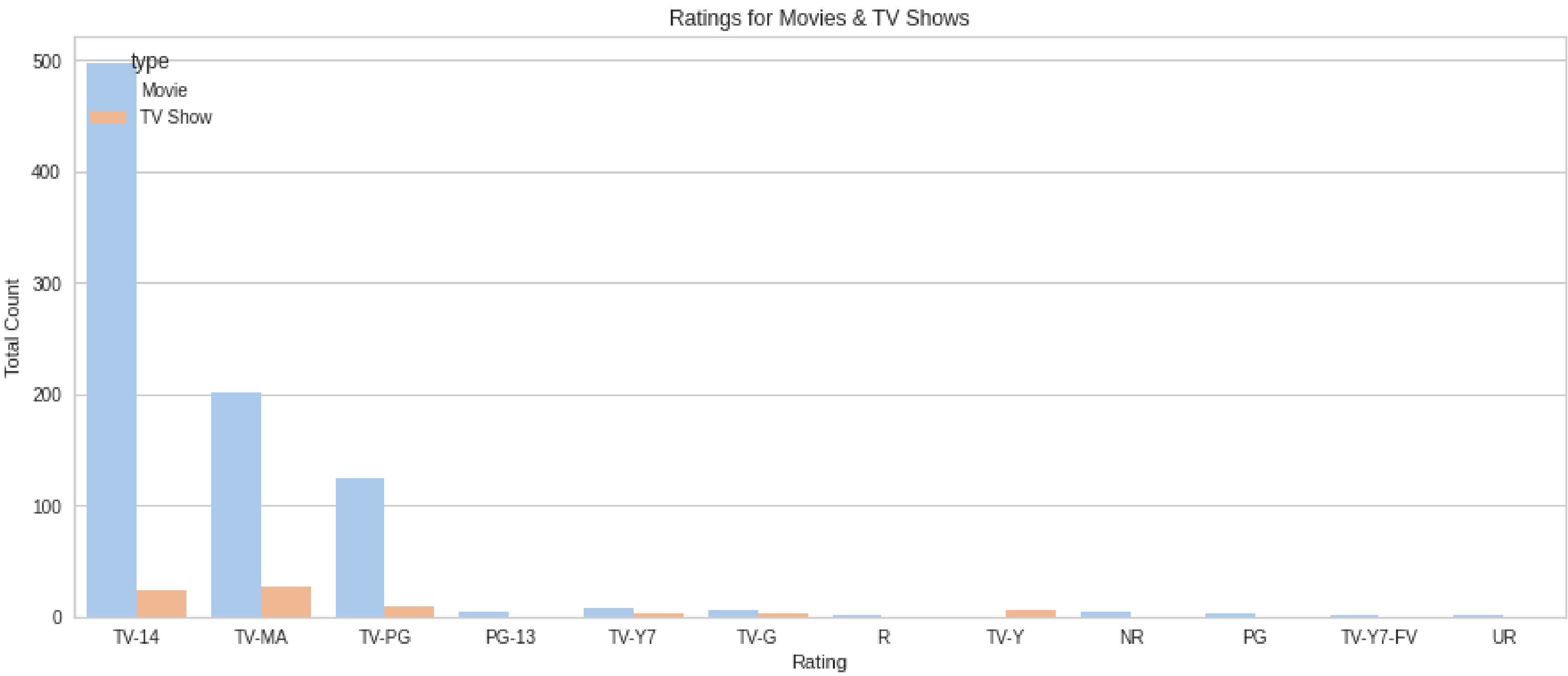
EDA(India)


Top 10 release year


Top 10 rating


Top 20 Genres on Netflix

EDA(India)


Ratings for Movies & TV Shows
type
Movie
TV Show
Total Count
500
400
300
200
100
0
TV-14
TV-MA
TV-PG
PG-13
TV-Y7
TV-G
R
TV-Y
NR
PG
TV-Y7-FV
UR
Rating

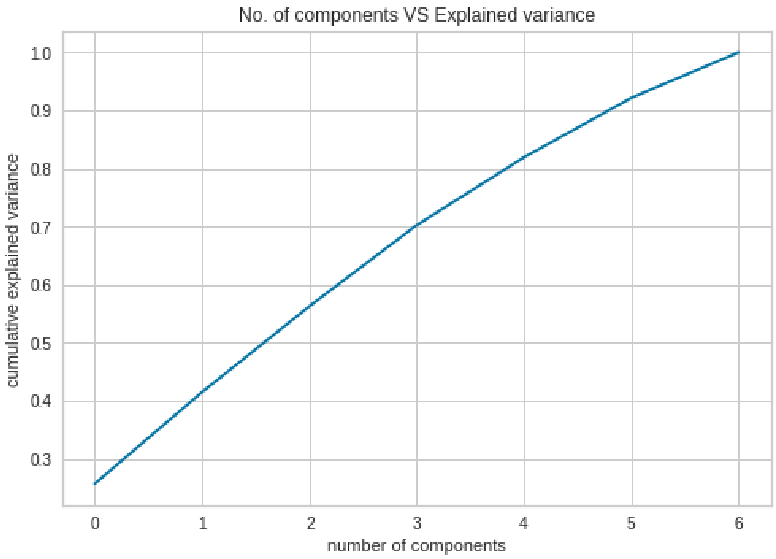
Implementation

Now Lets implement 3 models on our dataset:

1. Principal Component Analysis
2. K Means Clustering
3. **Hierarchical Clustering**

1. Principal Component Analysis (PCA)

PCA is fundamentally a dimensionality reduction algorithm, but it can also be useful as a tool for visualization, for noise filtering, for feature extraction and engineering, and much more.



2. K Means Clustering

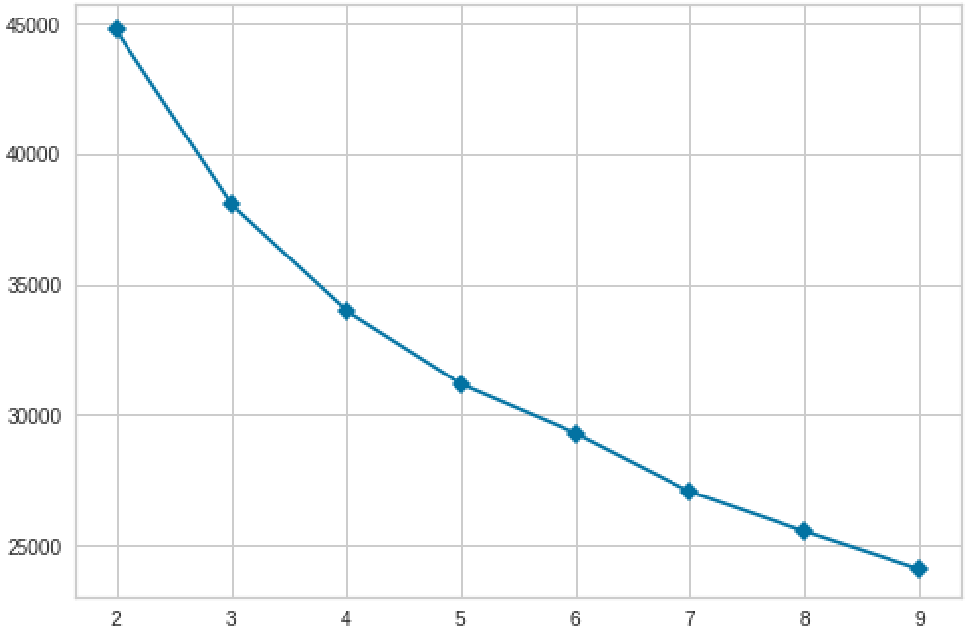
K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process.

Find out how many clusters are used?

K-Elbow Method

Silhouette Score

K-Elbow Method

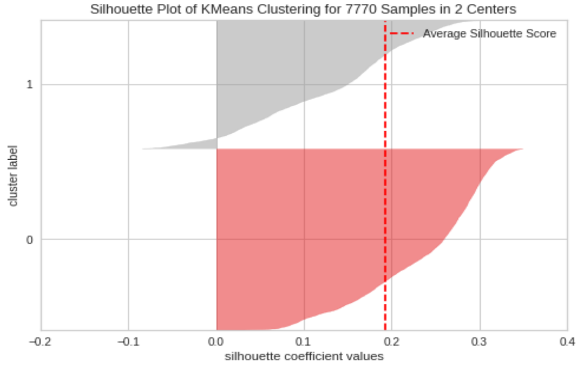


Silhouette Score

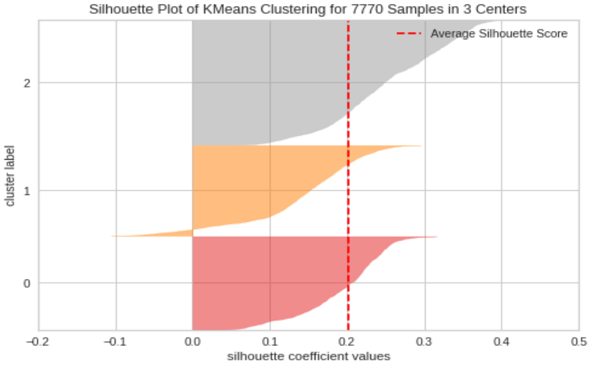
The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like number of clusters visually. This measure has a range of [-1, 1].

Silhouette coefficients (as these values are referred to as) near +1 indicate that the sample is far away from the neighboring clusters. A value of 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters and negative values indicate that those samples might have been assigned to the wrong cluster.

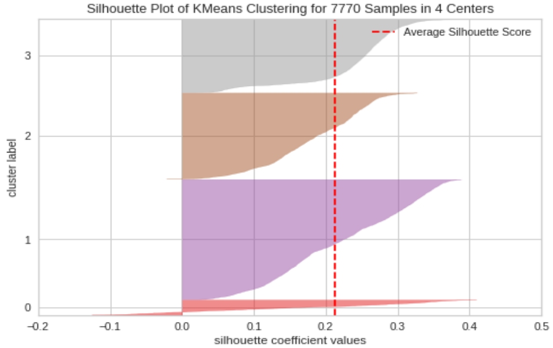
Silhouette Score


Silhouette Plot of KMeans Clustering for 7770 Samples in 2 Centers

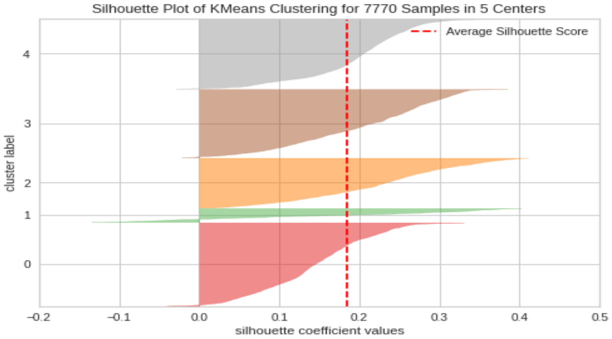
```
For n_clusters = 2, silhouette score
is 0.19261385848221638
```


Silhouette Plot of KMeans Clustering for 7770 Samples in 3 Centers

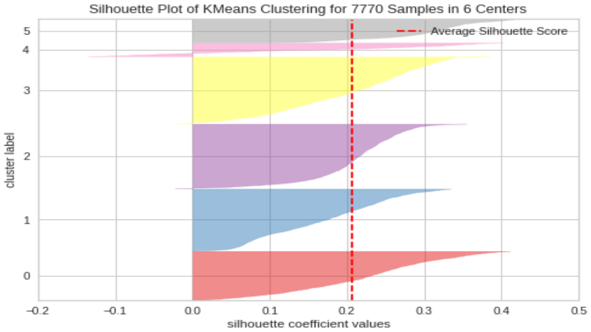
```
For n_clusters = 3, silhouette
score is 0.2011569332714960З
```


Silhouette Plot of KMeans Clustering for 7770 Samples in 4 Centers

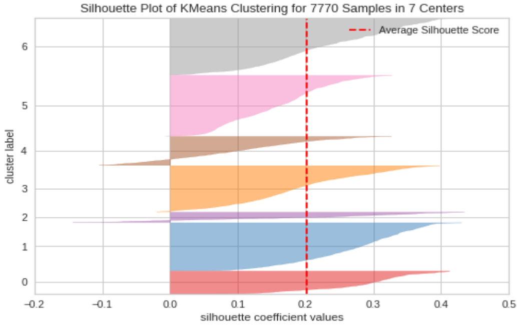
```
For n_clusters = 4, silhouette
score is 0.21208415191245705
```


Silhouette Plot of KMeans Clustering for 7770 Samples in 5 Centers

```
For n_clusters = 5, silhouette score
is 0.18480813970067428
```

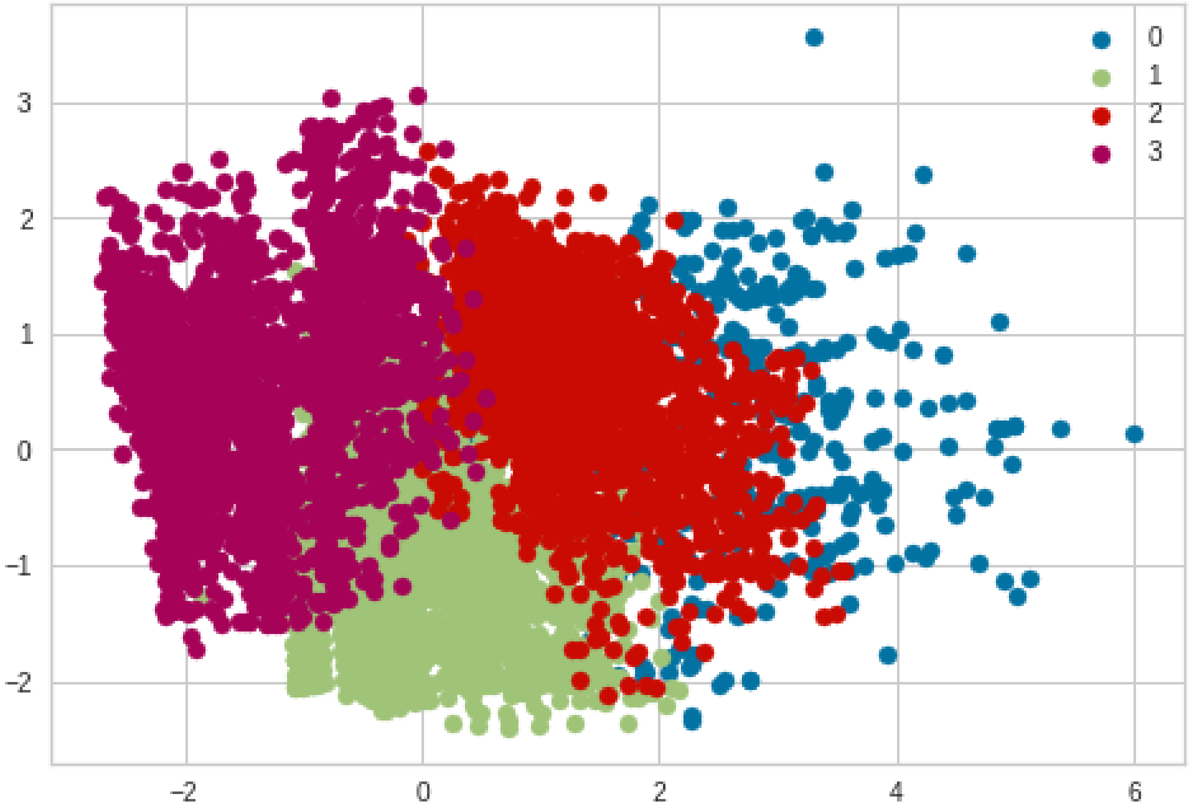

Silhouette Plot of KMeans Clustering for 7770 Samples in 6 Centers

```
For n_clusters = 6, silhouette
score is 0.2056640016509593
```

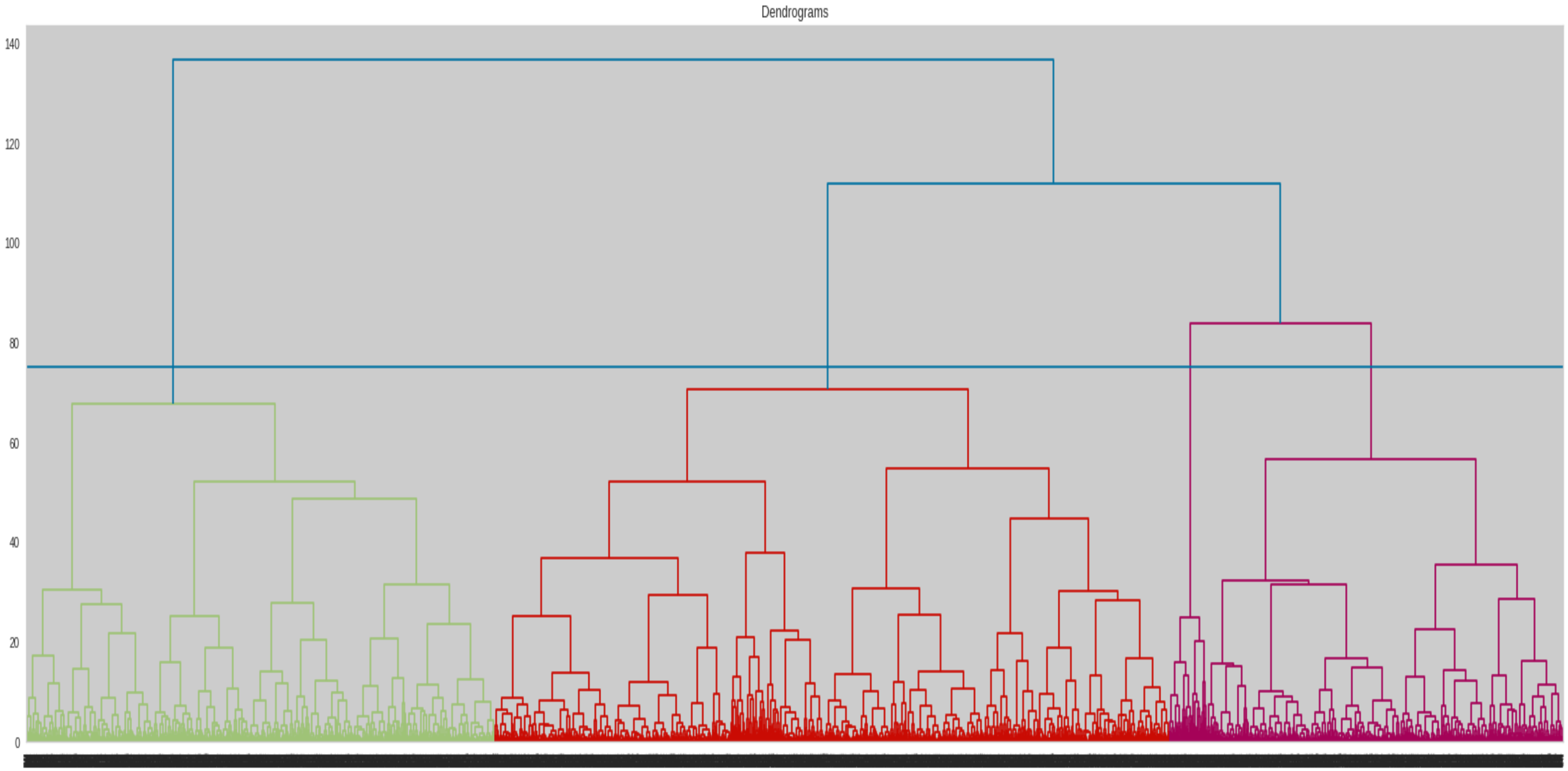

Silhouette Plot of KMeans Clustering for 7770 Samples in 7 Centers

```
For n_clusters = 7, silhouette
score is 0.2023528237249827
```

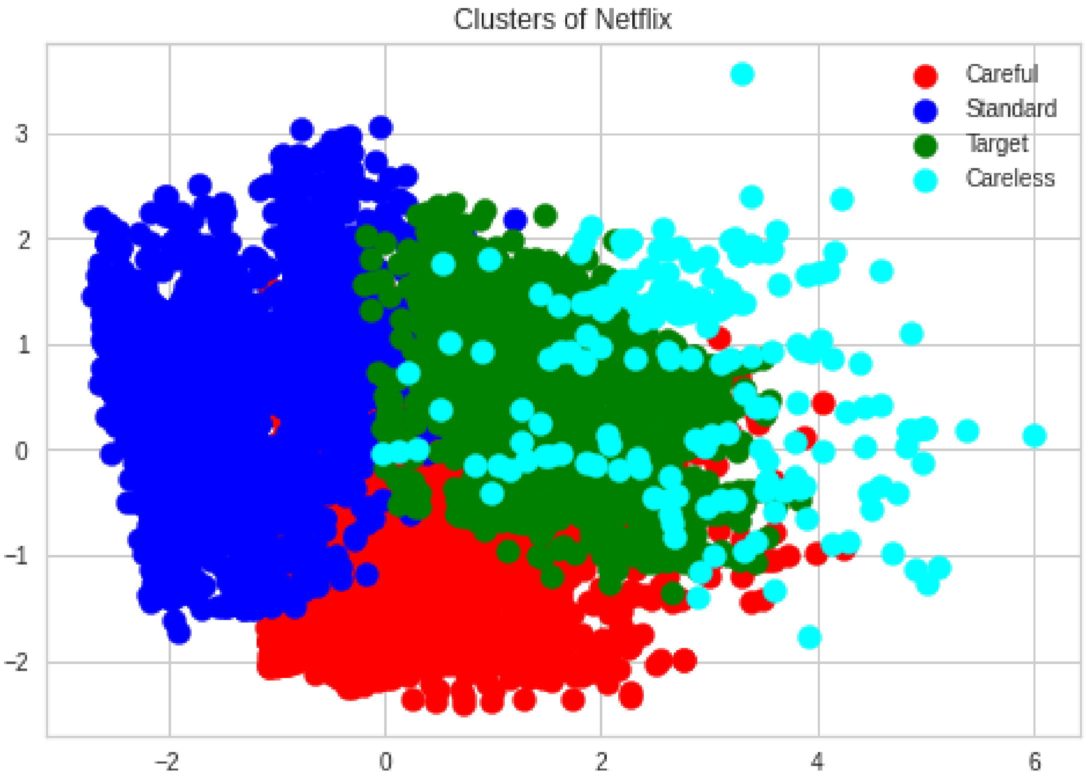
K Means Clustering



Hierarchical clustering



Agglomerative Clustering



Conclusion

We have drawn many interesting inferences from the dataset Netflix titles; here's a summary of the few of them:

The most content type on Netflix is movies. 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 have significantly increased after 2015 and have 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.

Conclusion

More of the content is released in 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. Drama is the most popular genre followed by comedy.

International movies are the top most genre in netflix which is fllowed by standup comedy and Drams.

Most of the movies have duration of between 50 to 150

Highest number of tv_shows consisting of single season

Conclusion

Using correlation heat map we see that in India mostly teens watching netflix so question arises that what content teens watched.

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.

For K Means clustering to find out number of k we used elbow and silhouette score method.

Using both the methods we found k=4 is optimal value of clustering.

Using Hierarchical clustering method again we find out that k=4 is optimal value of clustering.

Thank You