Capstone Project- 4

Netflix Movies and TV Shows Clustering

Project by- Potdar Vinayak

Contents

Abstract

The project mainly focuses on clustering similar content. The idea comes from the fact that there are people with similar interests and may like similar content to what they have already watched. This is followed by every new platform that we see these days like Spotify and Prime Music, where the user is requested for his/her interests and is delivered similar content.

Problem Statement

In this project we are supposed to build a unsupervised machine learning model using a clustering algorithm which will cluster similar TV shows and Movies from the Netflix Dataset based on the provided text-based features.
The other objective is to carry out EDA and figure out if the platform has been focusing more on TV- Shows than Movies in the recent times.

Understanding Data

```
data.head()
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	s1	TV Show	3%	NaN	JoÃ£o Miguel, Bianca Comparato, Michel Gomes, ...	Brazil	August 14, 2020	2020	TV-MA	4 Seasons	International TV Shows, TV Dramas, TV Sci-Fi &...	In a future where the elite inhabit an island ...
1	s2	Movie	7:19	Jorge Michel Grau	DemiÃ¡n Bichir, HÃ©ctor Bonilla, Oscar Serrano...	Mexico	December 23, 2016	2016	TV-MA	93 min	Dramas, International Movies	After a devastating earthquake hits Mexico Cit...
2	s3	Movie	23:59	Gilbert Chan	Tedd Chan, Stella Chung, Henley Hii, Lawrence ...	Singapore	December 20, 2018	2011	R	78 min	Horror Movies, International Movies	When an army recruit is found dead, his fellow...
3	s4	Movie	9	Shane Acker	Elijah Wood, John C. Reilly, Jennifer Connelly...	United States	November 16, 2017	2009	PG-13	80 min	Action & Adventure, Independent Movies, Sci-Fi...	In a postapocalyptic world, rag-doll robots hi...
4	s5	Movie	21	Robert Luketic	Jim Sturgess, Kevin Spacey, Kate Bosworth, Aar...	United States	January 1, 2020	2008	PG-13	123 min	Dramas	A brilliant group of students become card-coun...

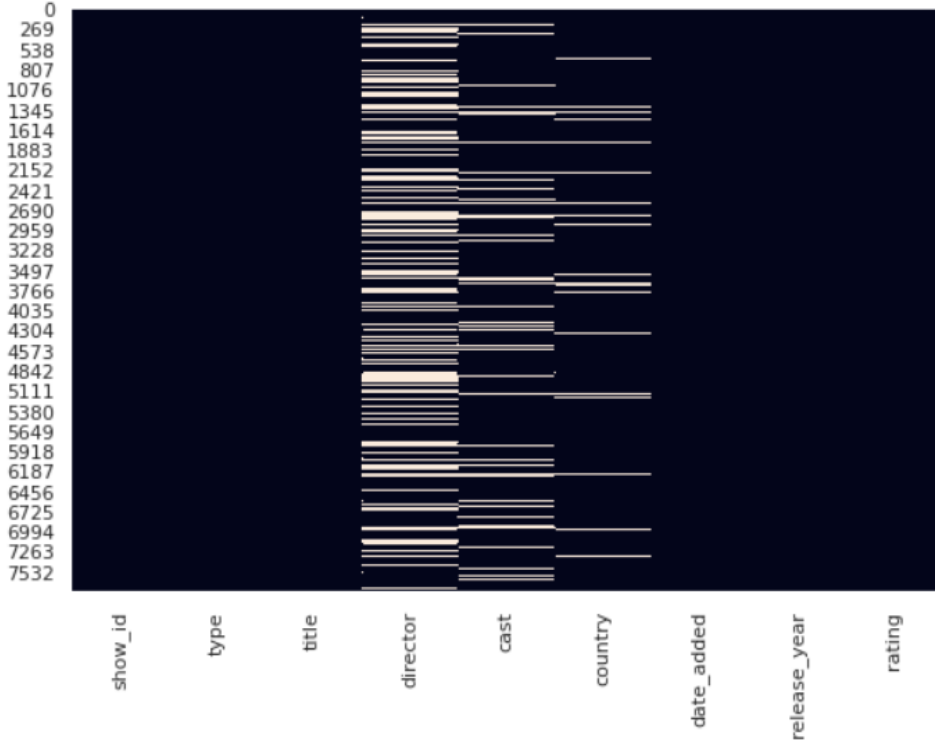
These are the features of the data.

Null Value Analysis and Treatment

The majority of the missing data is from the director feature, which cannot be filled easily using other features, hence I dropped it.

The same is the case with Cast.

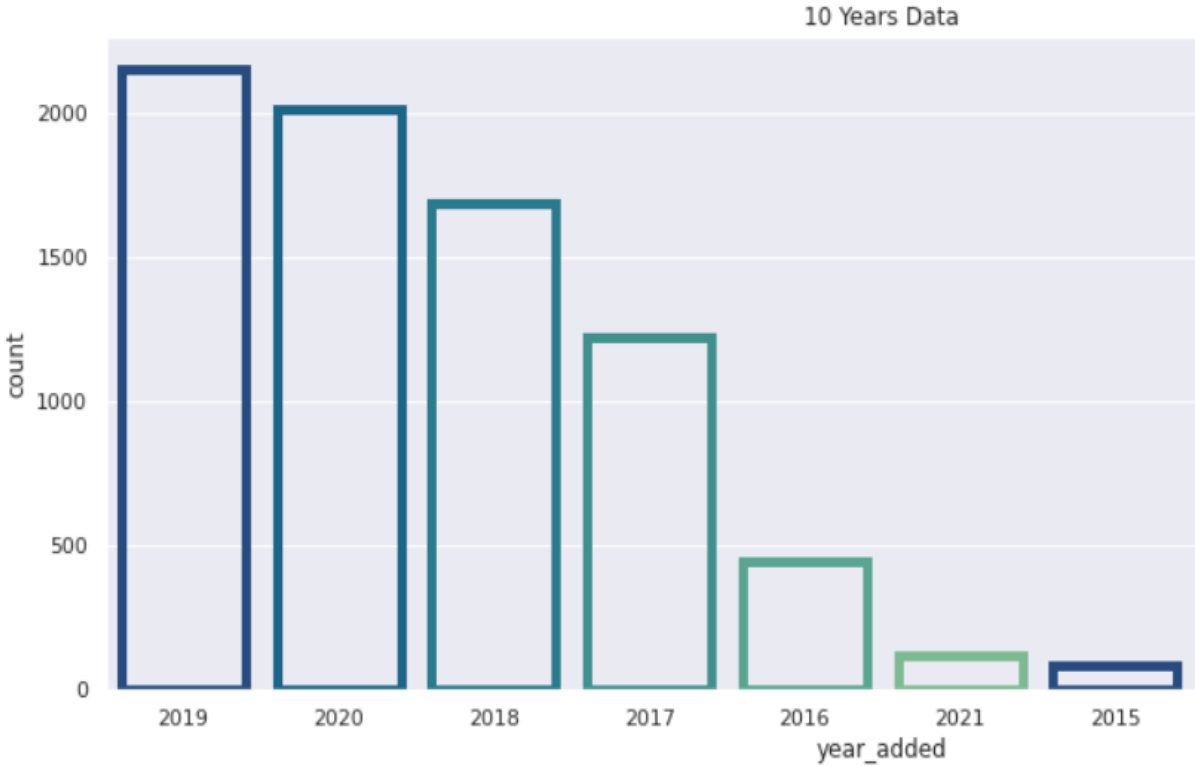
The feature country has been replaced with the mode of the data, which won't be affecting the Clustering process.



Exploratory Data Analysis

Top 10 Most Published Years:
Here, the year 2019 saw the most content being uploaded on the platform, followed by 2020.

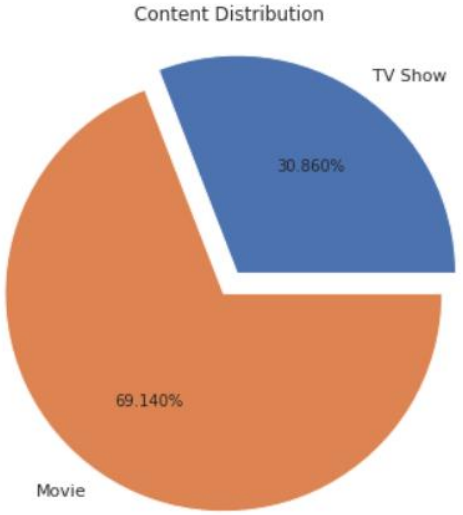
The rise in the content publishing started after the year 2015.

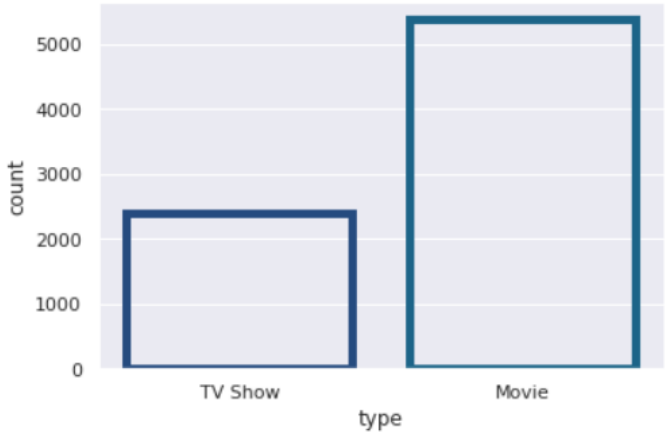


Exploratory Data Analysis

Distribution of TV Shows and Movies on the platform-

The TV- Shows make up 30.86% of the content on the platform whereas, movies make up 69.14% of the whole.



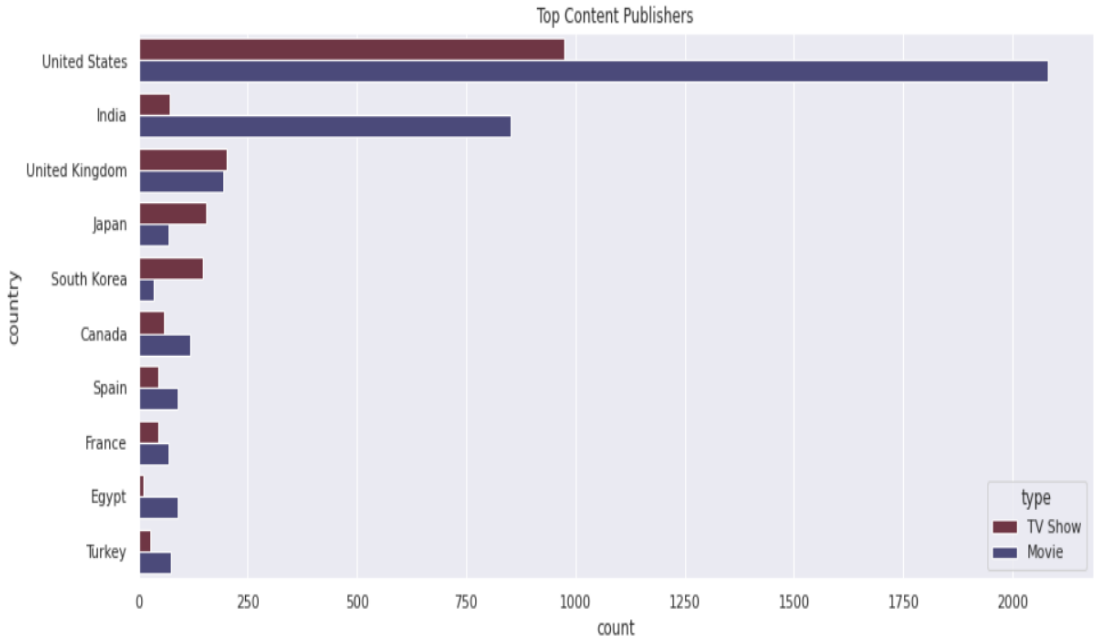


Exploratory Data Analysis

Country wise trend of content publishing-

US, India and UK happen to be the most actively publishing countries on the platform.

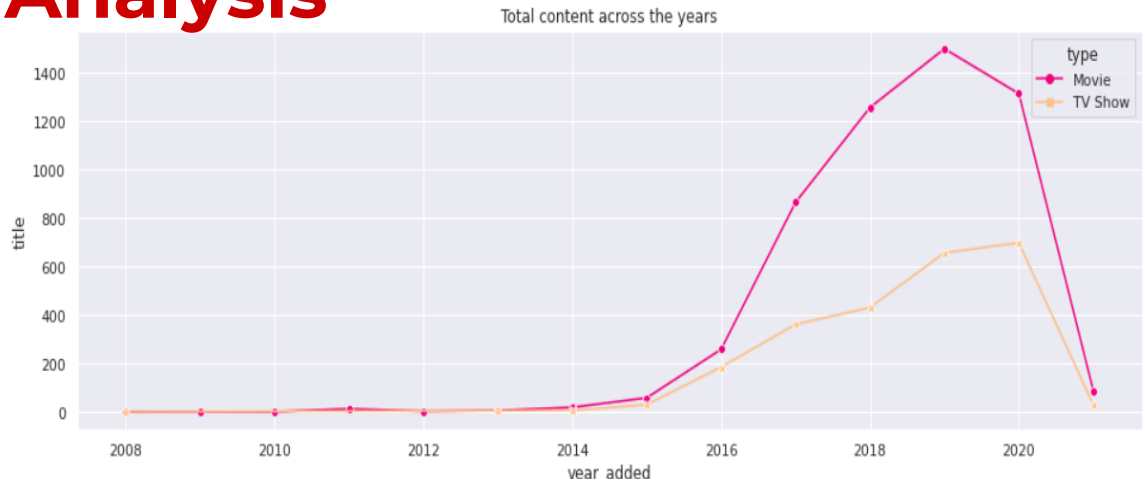
Majorly every country publishes more movies than TV shows except, UK, Japan and South Korea.



Exploratory Data Analysis

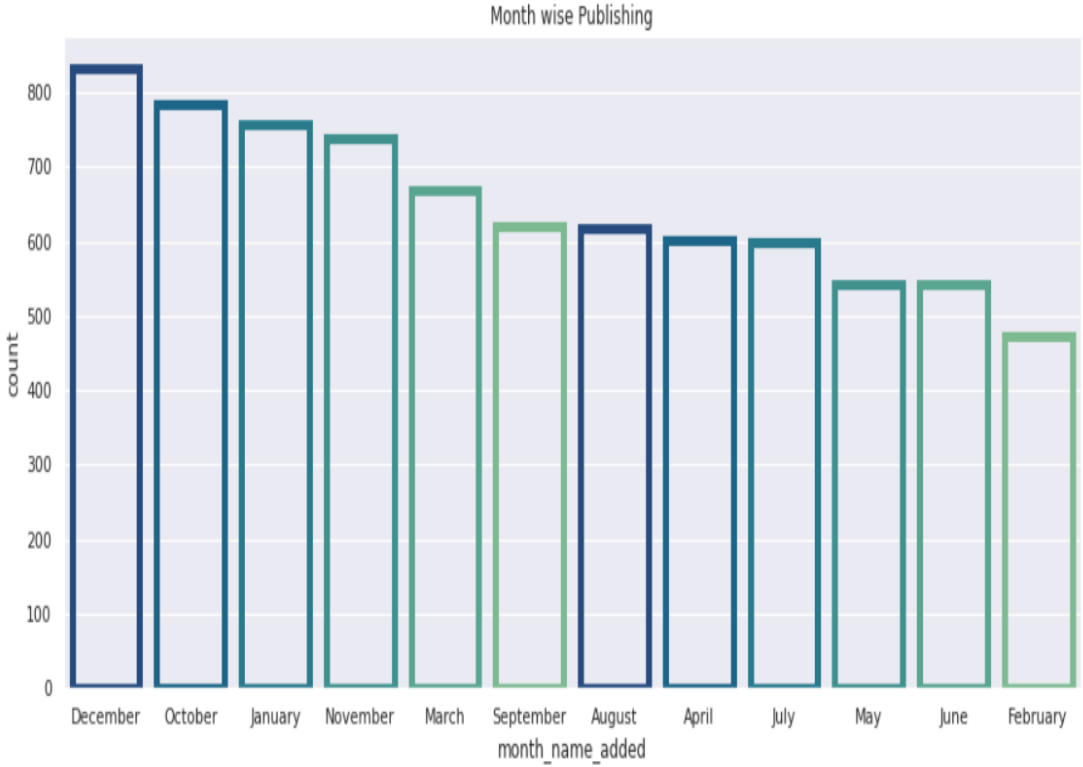
The Trend of Publishing Movies and TV-Shows on the platform-
Movies and TV shows have been increasing exponentially on the platform from 2015.

We can see a drop in the number of movies published in the recent times.



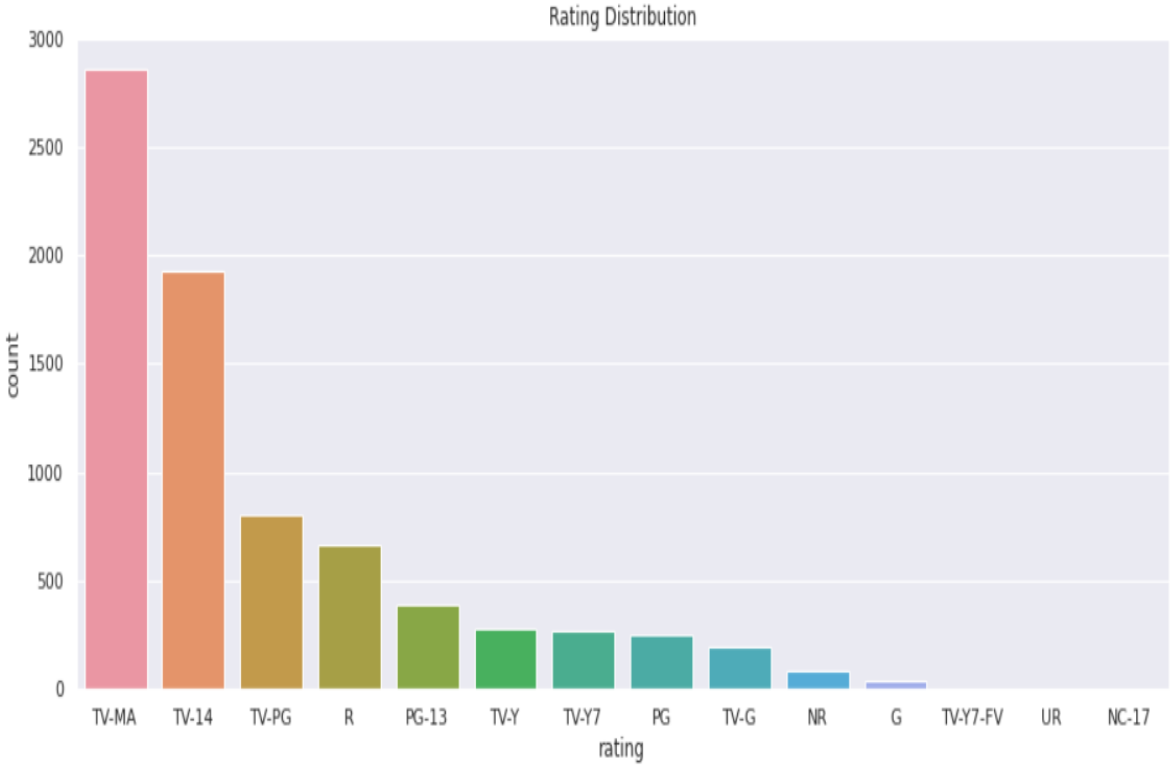
Exploratory Data Analysis

The Trend of Publishing Movies and TV-Shows on the platform-
Most of the content is published towards the end of the year in December, when the holidays are right around the corner.



Exploratory Data Analysis

Distribution of Ratings- Most of the content published on the platform is TV-MA i.e. the content is for mature audiences only. This is followed by TV-14 which stands for ‘Parents Strongly Cautioned’. This is a bit unwelcoming to the kids.



Hypothesis Testing

The data suggests that Movies are the majority holders of the platform. In order to check if they bring more engagement, we devised a hypothesis.
H0- Movies create more engagement than TV- shows.

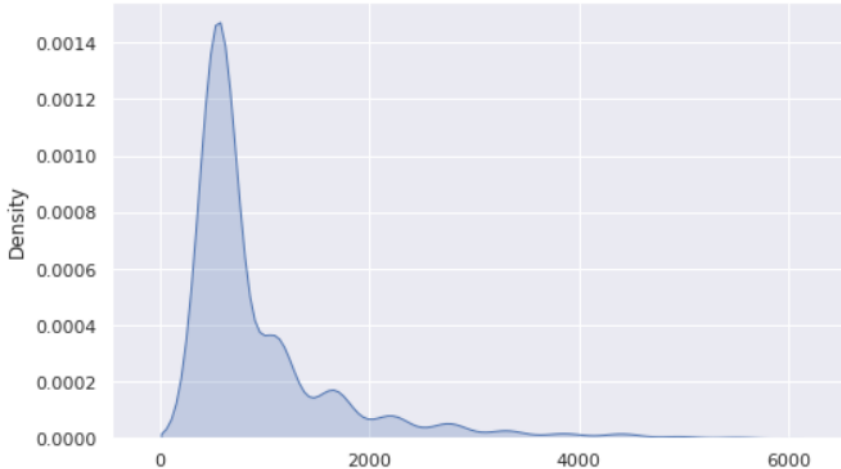
With an average of 55 minutes per episode and 10 episodes per season, we get the value of 1 season as 550 minutes.

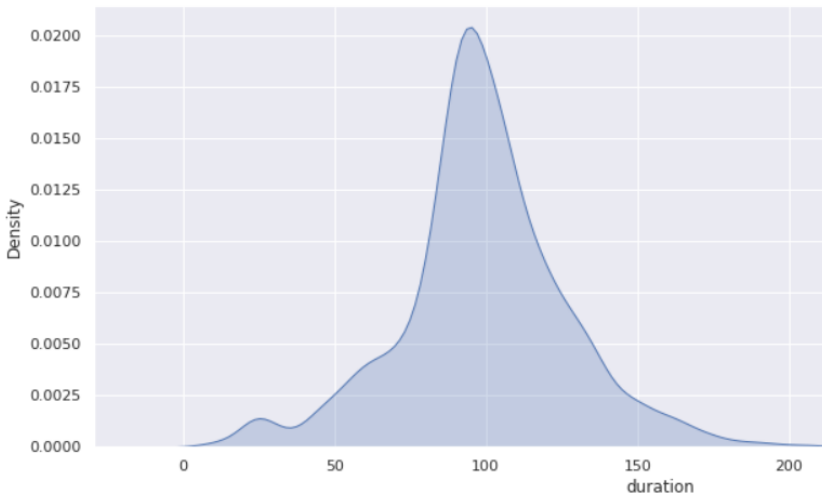
Hypothesis Testing

The first figure shows the TV shows duration where the average time spent can be around 500 minutes.

The second figure shows the Movies duration where the average time spent is around 100 minutes.

Clearly TV- Shows take up more time and engagement than the movies. Which is enough to reject out null hypothesis.





Data Preprocessing / Feature Engineering

Here the features we had to work with were already defined i.e. Text-based data (Listed_in and Description).
I split the data into two parts, one of which consisted of only the text-based data and the other consisted normalized numeric data.

The text-based data was combined and cleaned. I removed the stop-words, punctuations and stemmed it before vectorizing. The vectorizer used was Tf-idf.

This was followed by topic modelling.

Topic Modelling

In topic modelling we try to figure out the hidden theme and meaning of the text corpus. I used LDA and LSA as the two topic modelling techniques.

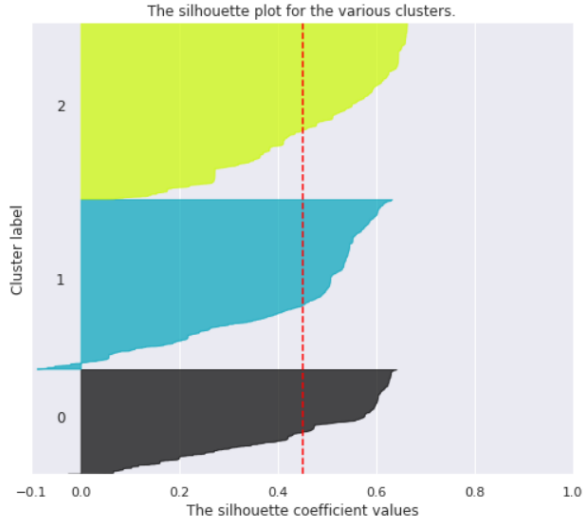
LSA works on distributional hypothesis, meaning similar words appear together frequently.

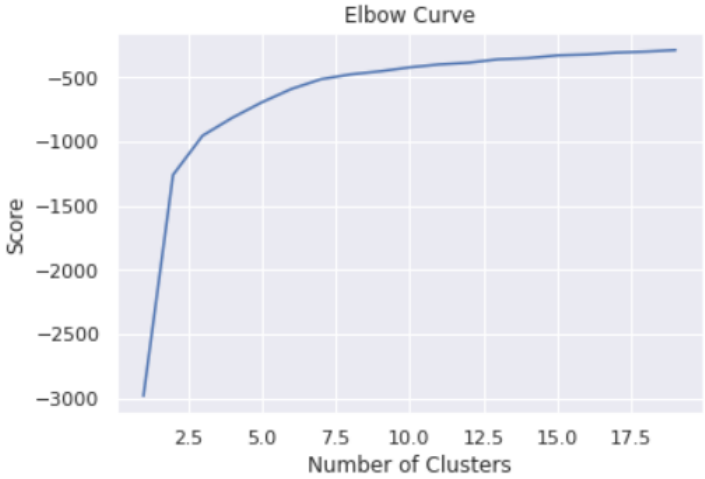
LDA specifically tries to find out the document's topic.

Model Fitting

The model I decided to use for the vectorized text here was Kmeans model. We need to specify the number of the clusters beforehand. In order to get the optimal number of clusters for the data set, I needed to perform some analysis.

This was done using Silhouette Score, Elbow Method and Dendrogram. The final and optimal number of cluster we got was **3.**



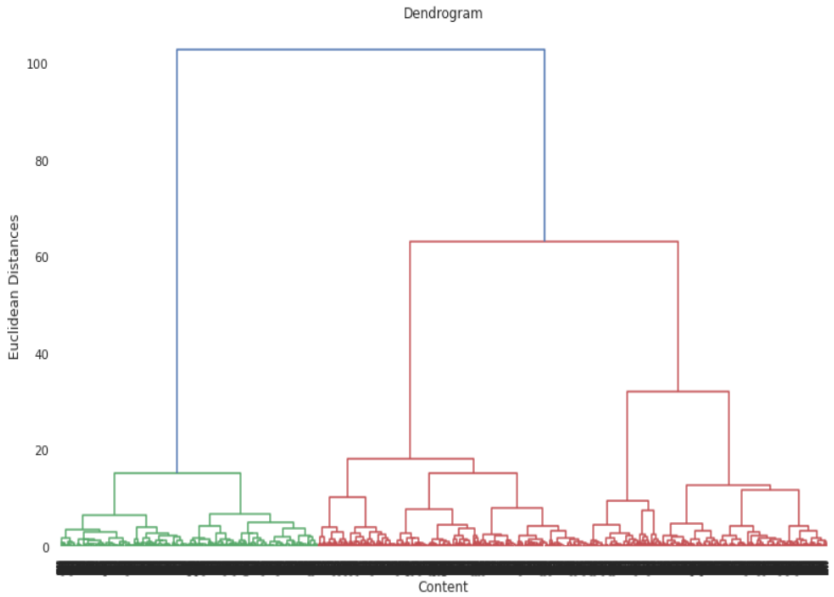


Model Fitting

I then fit the Kmeans model with 3 clusters as the required amount of clusters.

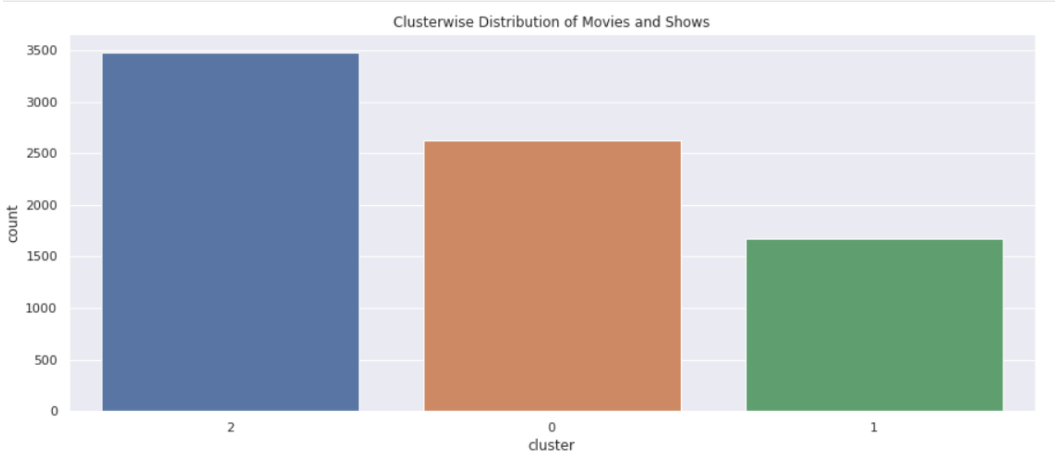
It created 3 clusters with names 0,1,2 as it has no understanding of topics they hold.

I still tried to extract the feature names of the clusters which were prominent near the cluster centers.



Model Fitting

I then fit the Kmeans model with 3 clusters as the required amount of clusters.
It created 3 clusters with names 0,1,2 as it has no understanding of topics they hold.
I still tried to extract the feature names of the clusters which were prominent near the cluster centers.




```
Cluster centroids:

Cluster 0:
 drama
 movi
 intern
 comedi
 independ
 romant
 young
 woman
 horror
 love
-----------------------
Cluster 1:
 tv
 show
 intern
 crime
 british
 romant
 kid
 seri
 korean
 spanishlanguag
-----------------------
Cluster 2:
 documentari
 action
 adventur
 famili
 standup
 children
 intern
 movi
 kid
 world
-----------------------
```

Conclusion

- 2019 was the year with most number of publishing on the platform.
- 69% of the content is movies and the rest is TV-Shows.
- US and India are the top creators of the platform.
- UK, Japan and South Korea are publishing more TV shows than Movies.
- Most of the Movies and shows are published by the end of year.
- Most of the content is for Adult Audience only.
- TV-Shows create more engagement than Movies.
- 3 Clusters was the optimal number of clusters we got.
- We used Kmeans model for the clustering.