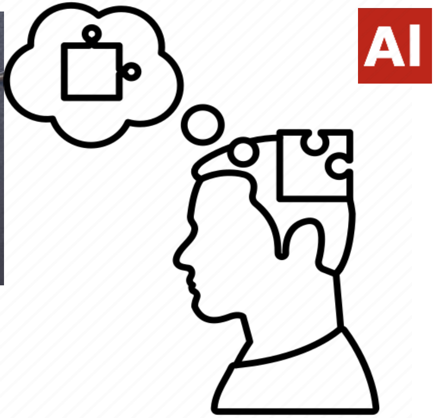
Capstone Project

NETFLIX-MOVIES-AND-TV-SHOWS-CLUSTERING

Mind Benders Team Members

Abdul Aziz
Pooja Yadav
G M Sravya Sree
Abdullah Bin Mohammed



- Introduction
- Problem Statement.
- Presenting Dataset Sample.
- Exploratory Data Analysis.
- Clustering Algorithm.
- Inferences and Conclusions.
- References.

About Netflix

Netflix was founded in 1997 by Reed Hastings and Marc Randolph in Scotts Valley, California. Netflix initially both sold and rented DVDs by mail, but the sales were eliminated within a year to focus on the DVD rental business. In 2007, Netflix introduced streaming media and video on demand.

Netflix is a subscription-based streaming service that allows our members to watch TV shows and movies without commercials on an internet-connected device. You can also download TV shows and movies to your iOS, Android, or Windows 10 device and watch without an internet connection.

As the world’s leading Internet television network with over 160 million members in over 190 countries, our members enjoy hundreds of millions of hours of content per day, including original series, documentaries and feature films.

Problem Statements

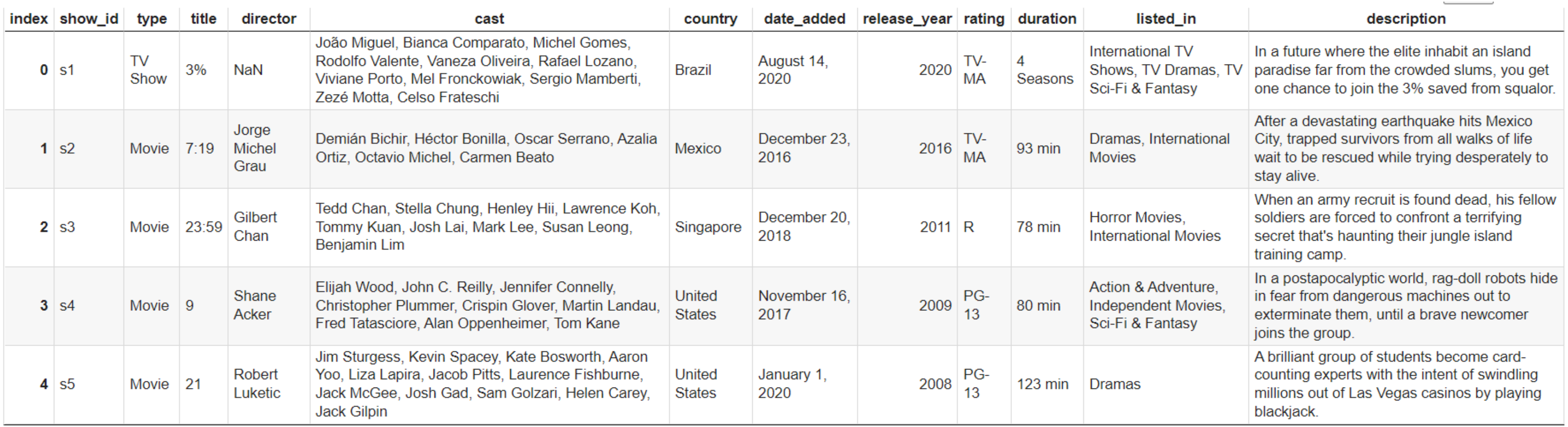
Netflix is all about recommending the next content to its user. The only question they would like to answer is ‘How to personalize Netflix as much as possible to a user?’.

The goal of this project is to find out similarity within groups in people to build a movie recommendation system for users.

We are going to analyze a dataset from Netflix database to explore the characteristics that people share in movies taste.

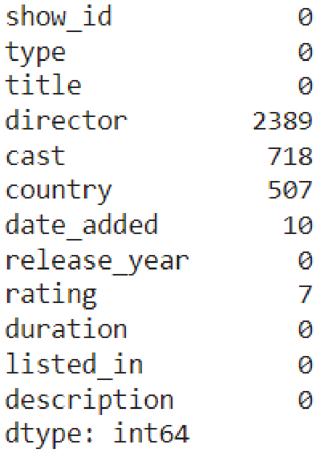
Dataset

This dataset has around 7787 observations in it with 12 columns.



index	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	s1	TV Show	3%	NaN	João Miguel, Bianca Comparato, Michel Gomes, Rodolfo Valente, Vaneza Oliveira, Rafael Lozano, Viviane Porto, Mel Fronckowiak, Sergio Mamberti, Zezé Motta, Celso Frateschi	Brazil	August 14, 2020	2020	TV-MA	4 Seasons	International TV Shows, TV Dramas, TV Sci-Fi & Fantasy	In a future where the elite inhabit an island paradise far from the crowded slums, you get one chance to join the 3% saved from squalor.
1	s2	Movie	7:19	Jorge Michel Grau	Demián Bichir, Héctor Bonilla, Oscar Serrano, Azalia Ortiz, Octavio Michel, Carmen Beato	Mexico	December 23, 2016	2016	TV-MA	93 min	Dramas, International Movies	After a devastating earthquake hits Mexico City, trapped survivors from all walks of life wait to be rescued while trying desperately to stay alive.
2	s3	Movie	23:59	Gilbert Chan	Tedd Chan, Stella Chung, Henley Hii, Lawrence Koh, Tommy Kuan, Josh Lai, Mark Lee, Susan Leong, Benjamin Lim	Singapore	December 20, 2018	2011	R	78 min	Horror Movies, International Movies	When an army recruit is found dead, his fellow soldiers are forced to confront a terrifying secret that's haunting their jungle island training camp.
3	s4	Movie	9	Shane Acker	Elijah Wood, John C. Reilly, Jennifer Connelly, Christopher Plummer, Crispin Glover, Martin Landau, Fred Tatasciore, Alan Oppenheimer, Tom Kane	United States	November 16, 2017	2009	PG-13	80 min	Action & Adventure, Independent Movies, Sci-Fi & Fantasy	In a postapocalyptic world, rag-doll robots hide in fear from dangerous machines out to exterminate them, until a brave newcomer joins the group.
4	s5	Movie	21	Robert Luketic	Jim Sturgess, Kevin Spacey, Kate Bosworth, Aaron Yoo, Liza Lapira, Jacob Pitts, Laurence Fishburne, Jack McGee, Josh Gad, Sam Golzari, Helen Carey, Jack Gilpin	United States	January 1, 2020	2008	PG-13	123 min	Dramas	A brilliant group of students become card-counting experts with the intent of swindling millions out of Las Vegas casinos by playing blackjack.

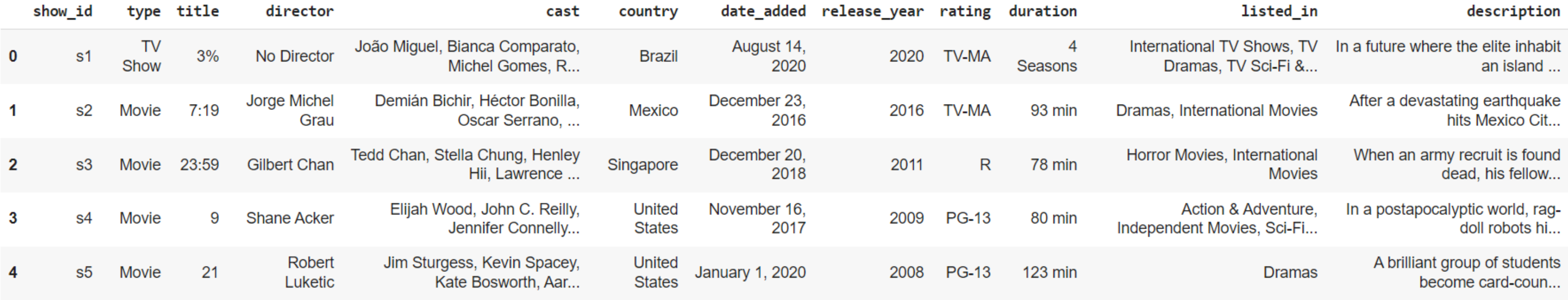
Data Cleaning



```
show_id            0
type               0
title              0
director        2389
cast             718
country          507
date_added        10
release_year       0
rating             7
duration           0
listed_in          0
description        0
dtype: int64
```

Checks the Null Values

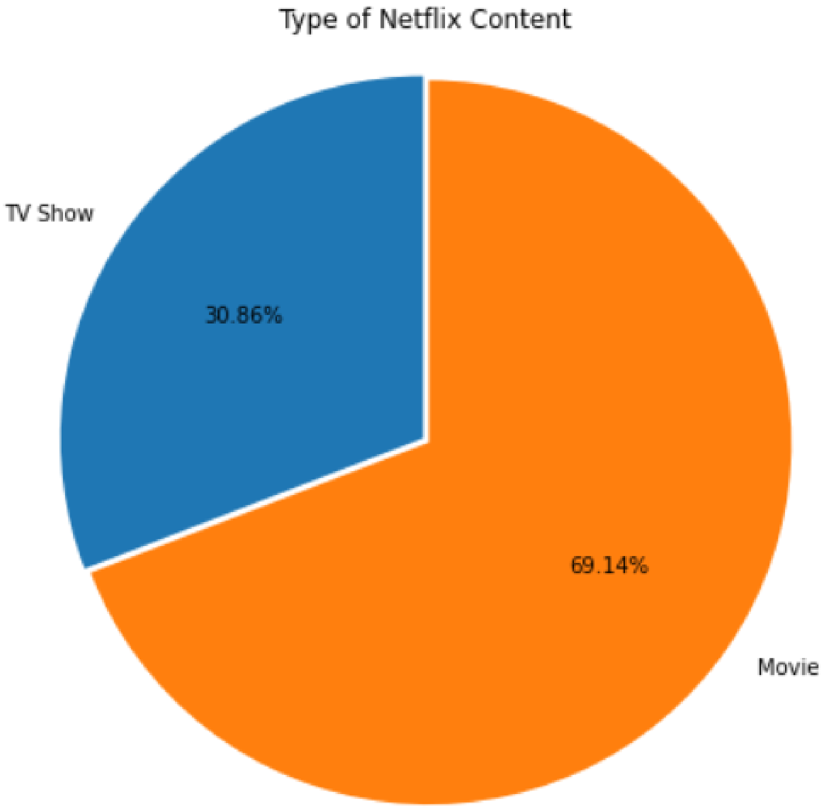
Final Dataset after Imputing missing values



	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	s1	TV Show	3%	No Director	João Miguel, Bianca Comparato, Michel Gomes, R...	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	Demián Bichir, Héctor 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...

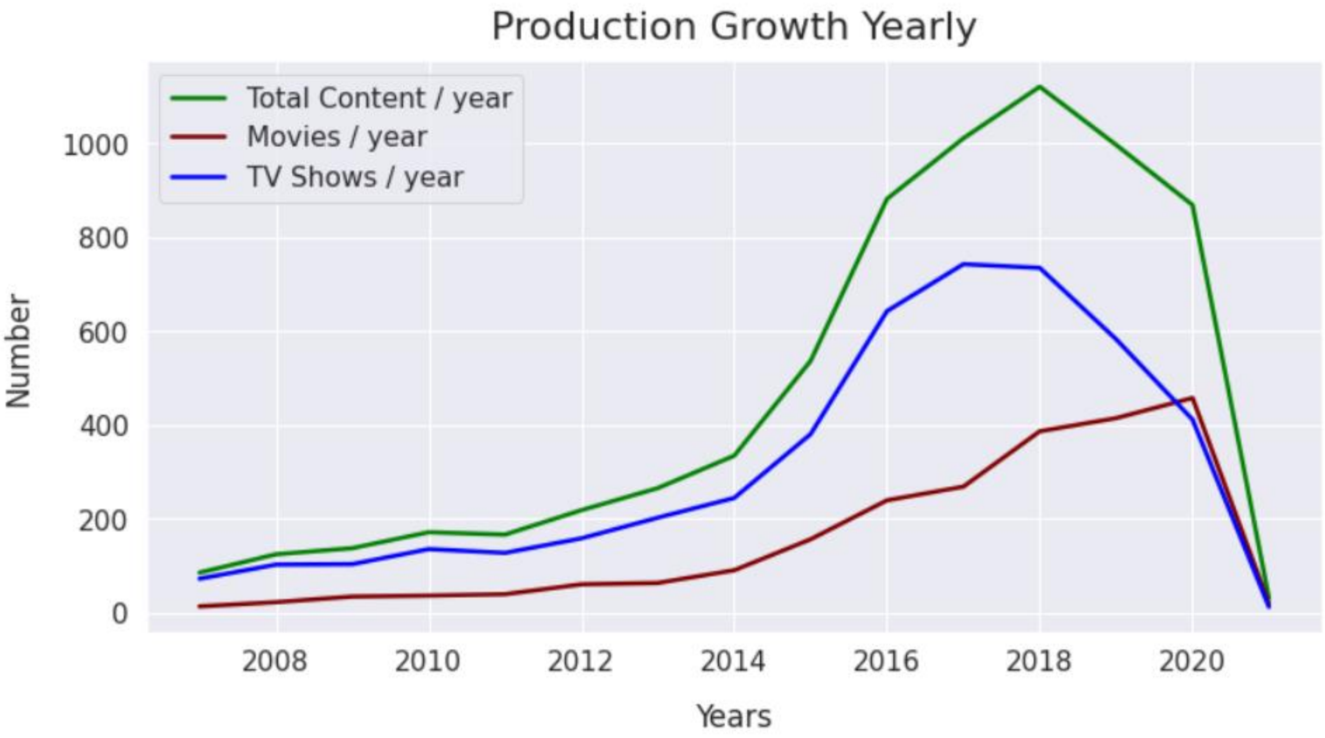
Exploratory Data Analysis

Type of Netflix Content



30.86 % TV Show and 69.6% Movie Content. The content type with most listings on Netflix is movies.

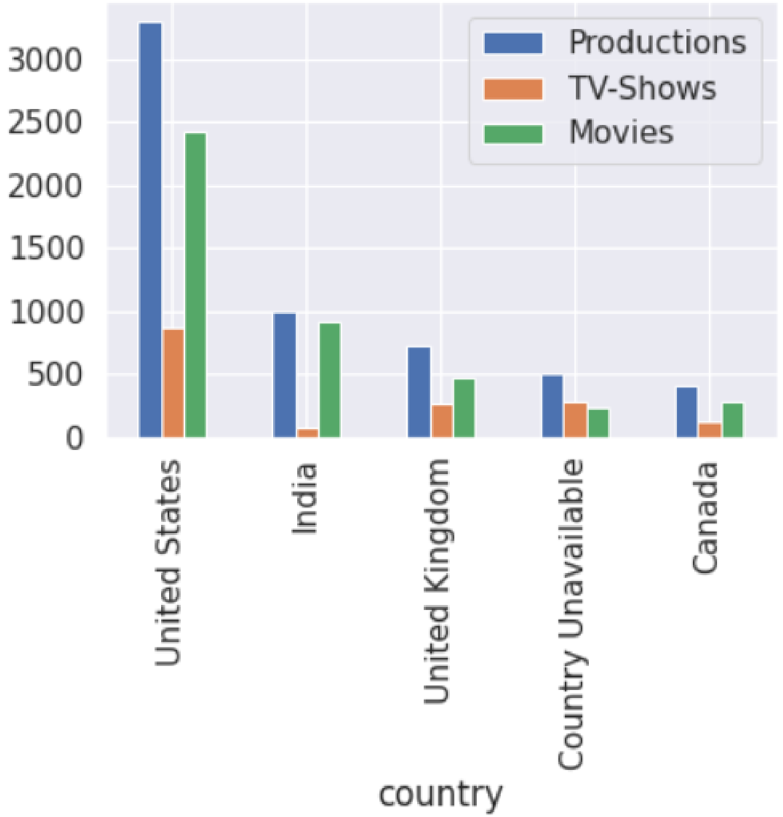
Production Growth Yearly



This plot shows the number of contents uploaded for TV Show or movies. We can see that the number of uploads for both the categories started increasing significantly after 2014.

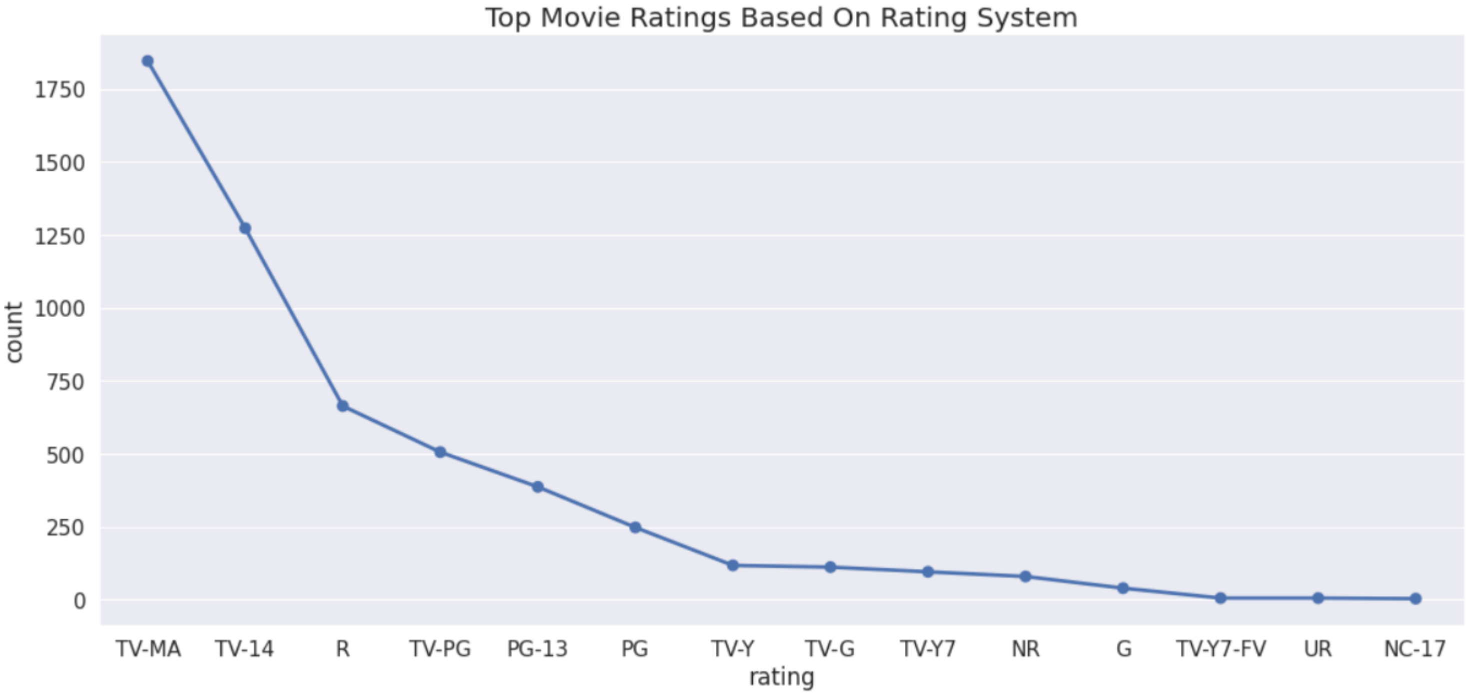
Countries producing most number of content.

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f9e791cb210>
```



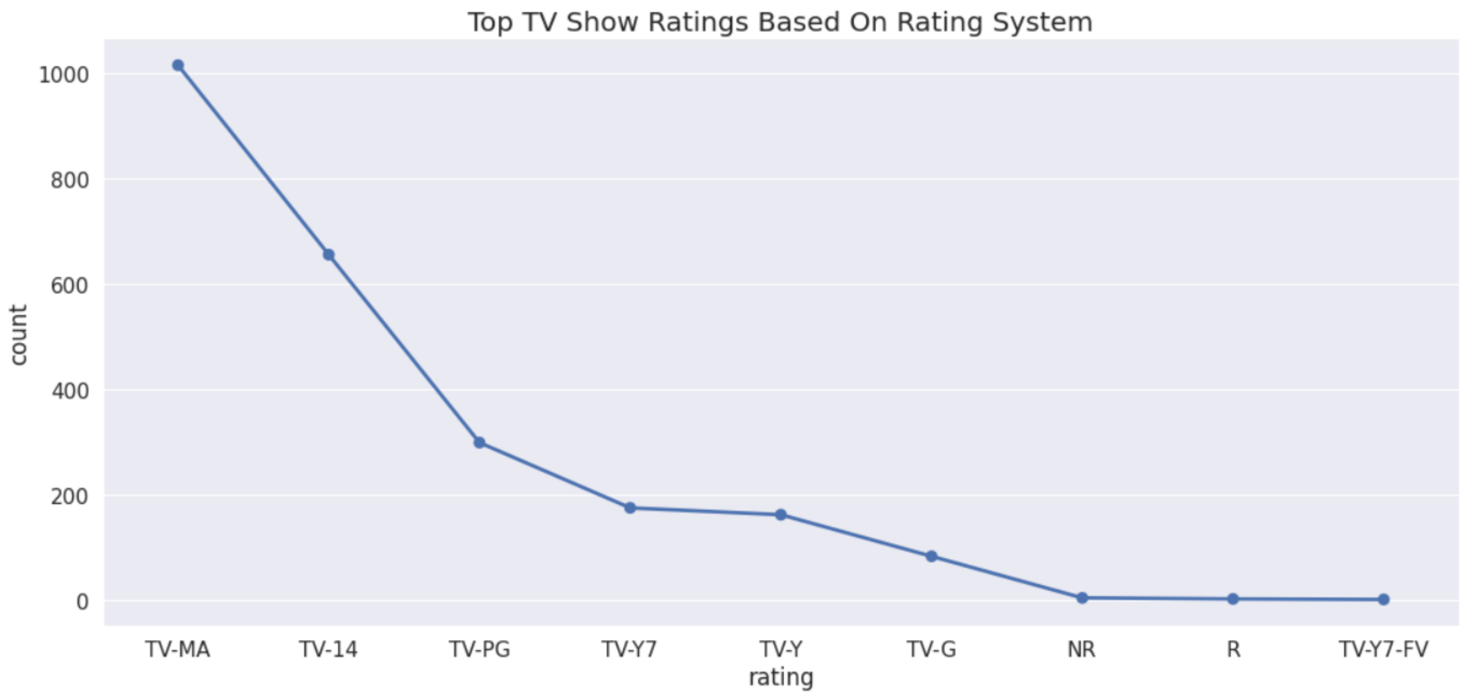
United States is the leader in producing content on Netflix and India is on second position

Top Movie Ratings Based On Rating System



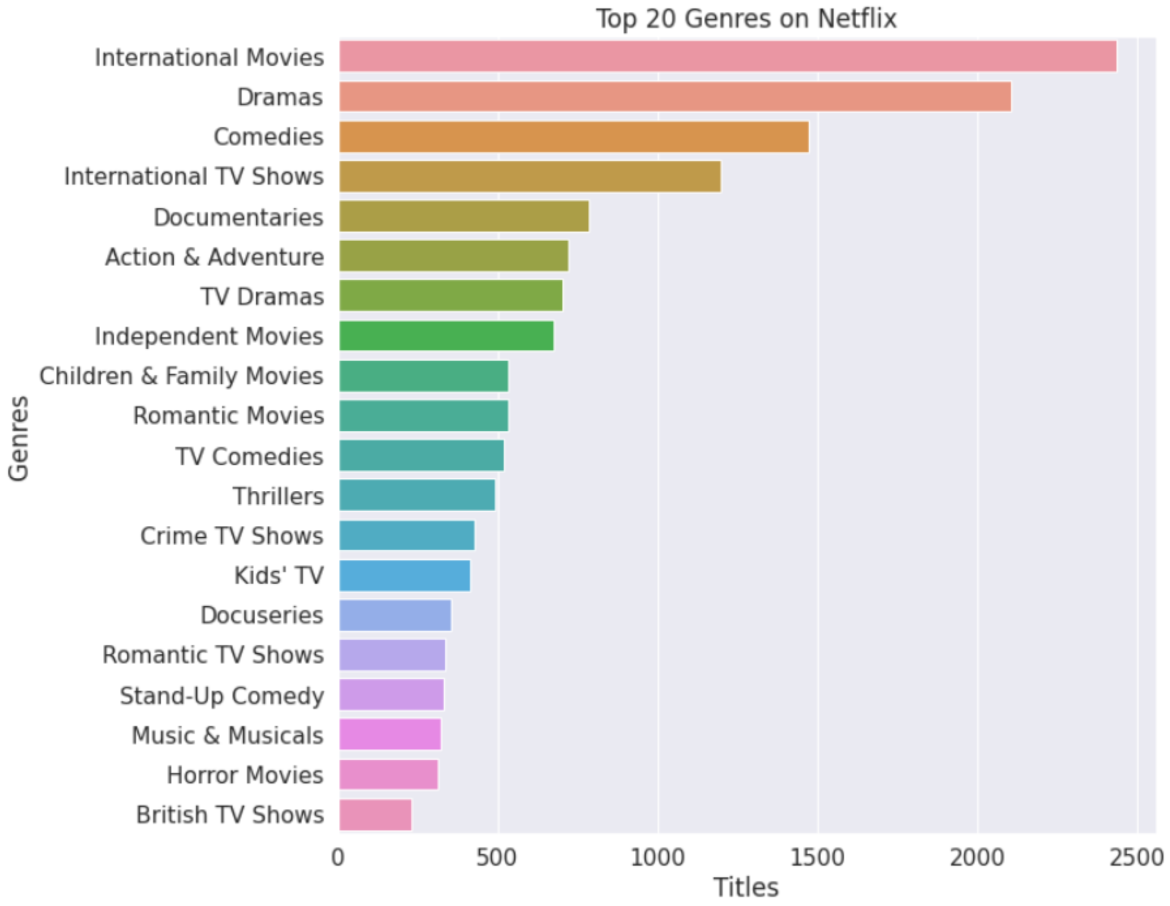
We extract the data for Movies, then we plot the diffrent ratings. Mostly, TV-MA is the rating the users have given followed by TV-14.

Top TV Show Ratings Based On Rating System



Similarly, We extracted the data for TV show, then plot the different ratings. Mostly, TV-MA is the rating the users have given followed by TV-14.

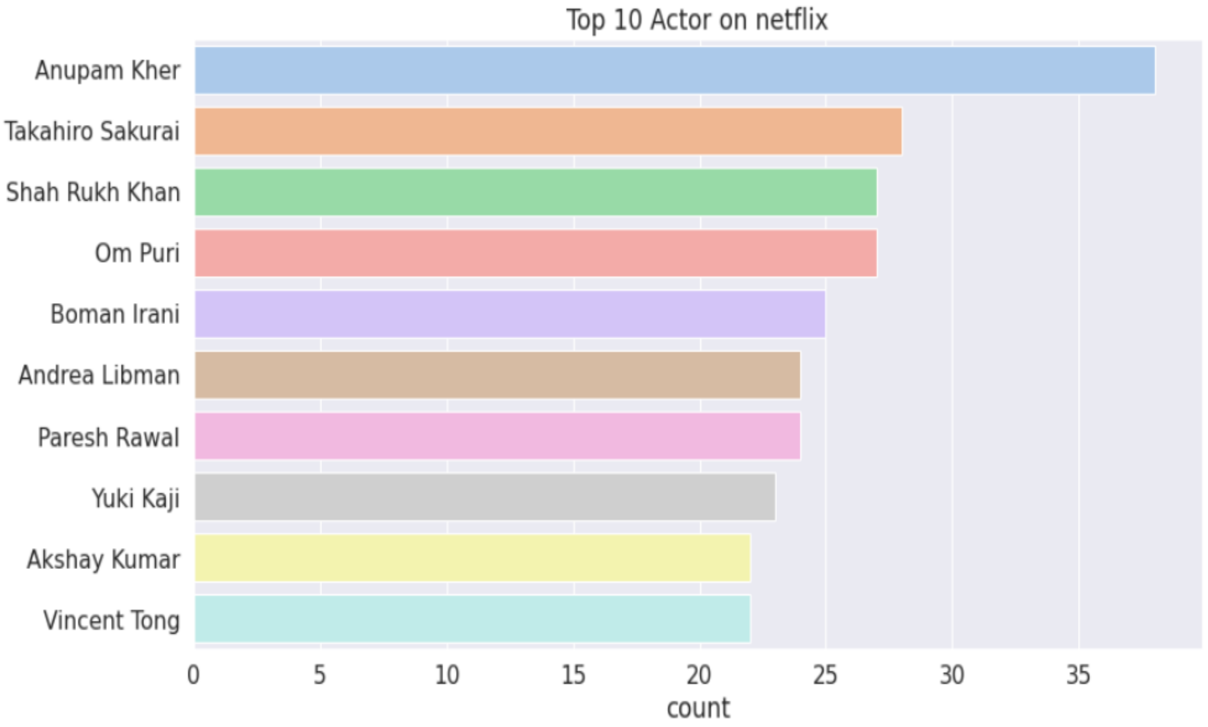
Top 20 Genres on Netflix



This shows the top 20 Genres available in the Netflix dataset.

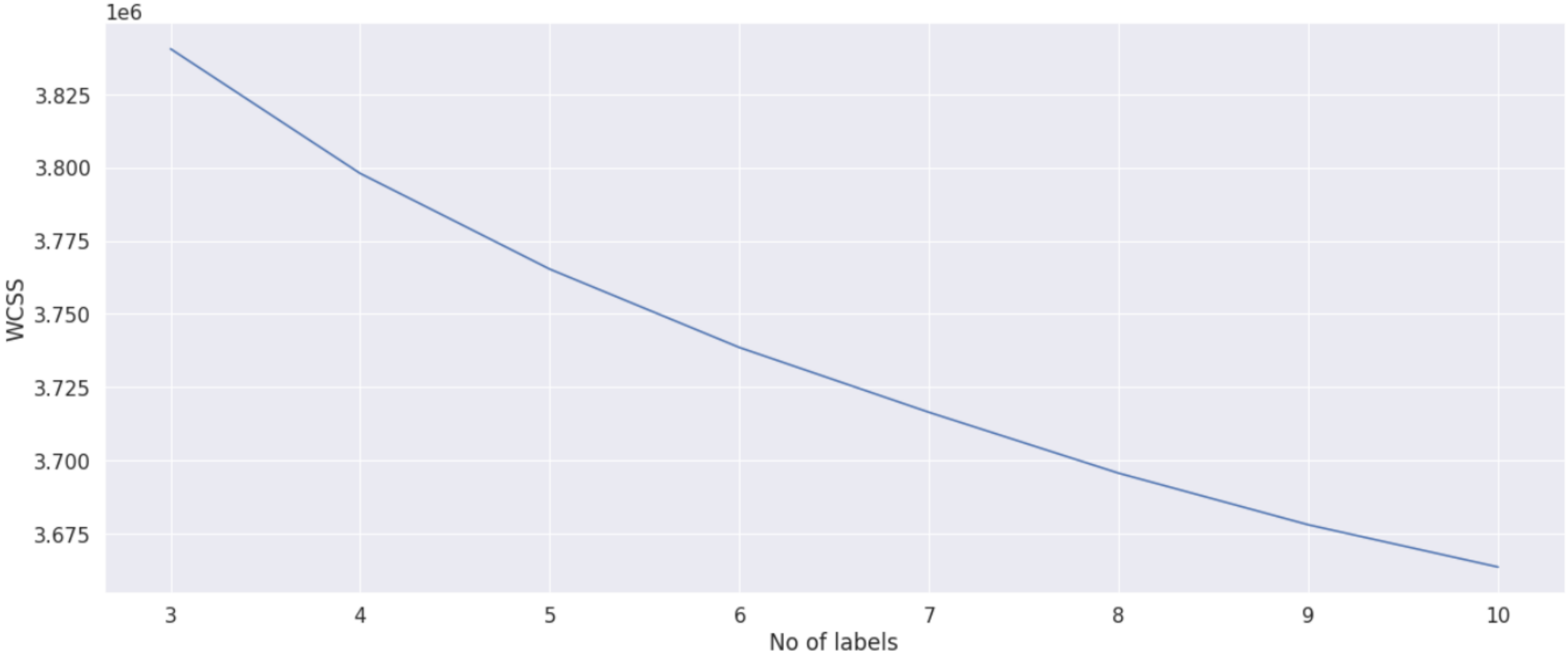
International Movies is the most famous Genres on Netflix followed by Dramas and comedies.

Top 10 Actor on Netflix



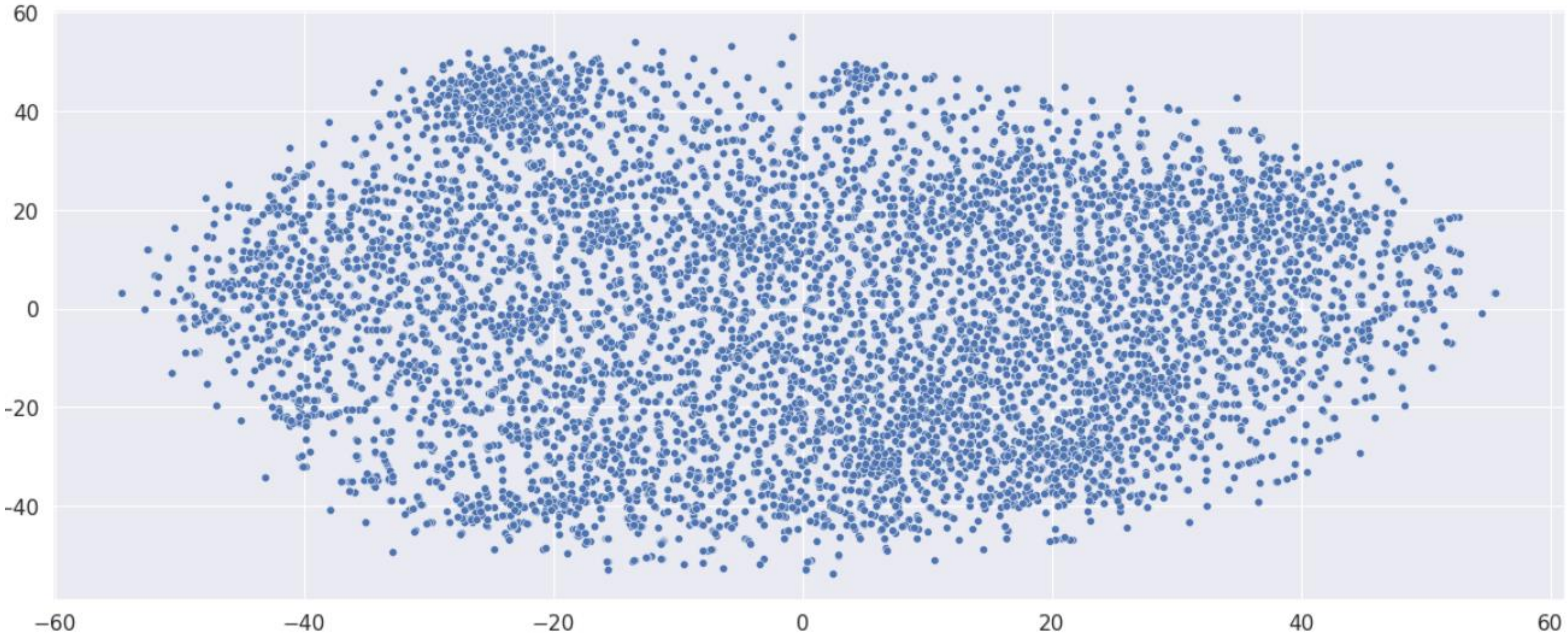
Whatever listings are present in the dataset, out of that Anupam Kher seems to be part of cast in a lot of movies followed by Takahiro Sakurai.

Choose right no of cluster

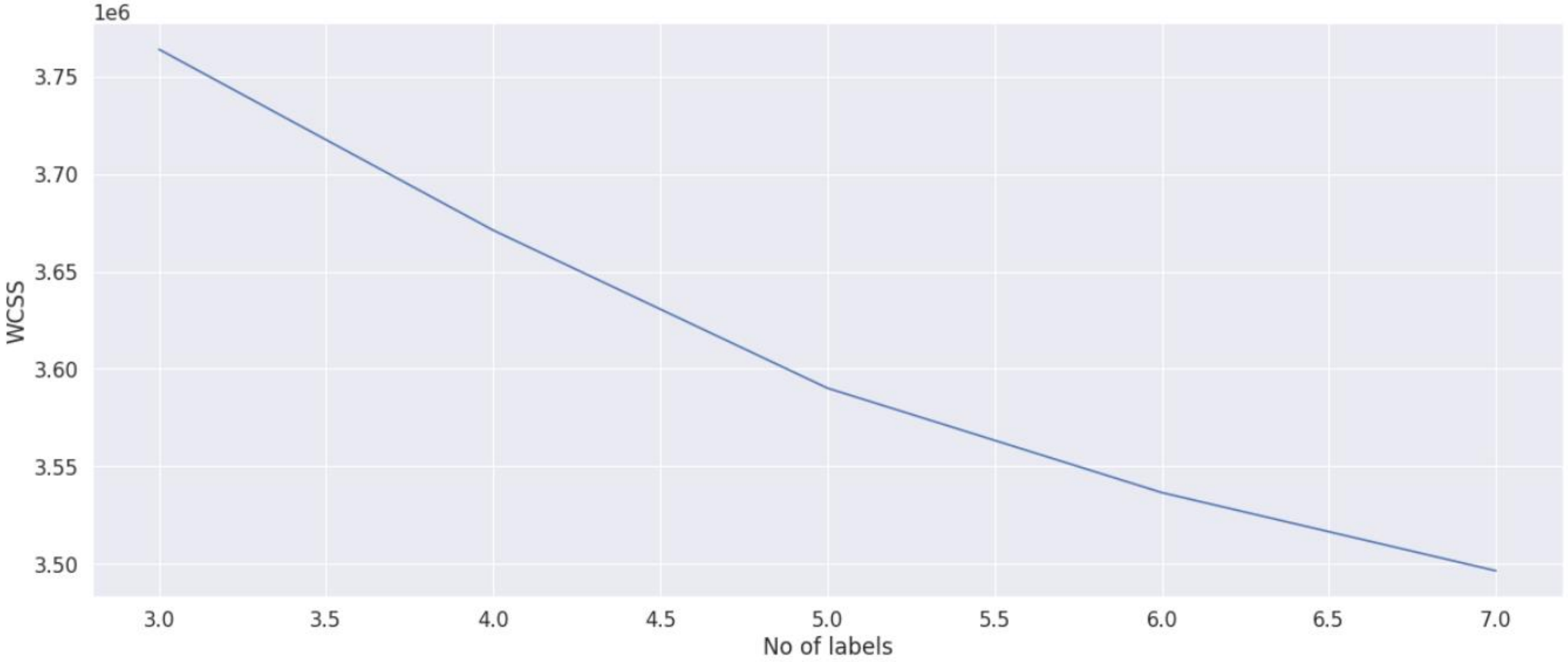


To check the number of clusters possible, we plot the elbow curve for the embeddings matrix for the column description. Not much can be observed from the plot as it is almost a straight line.

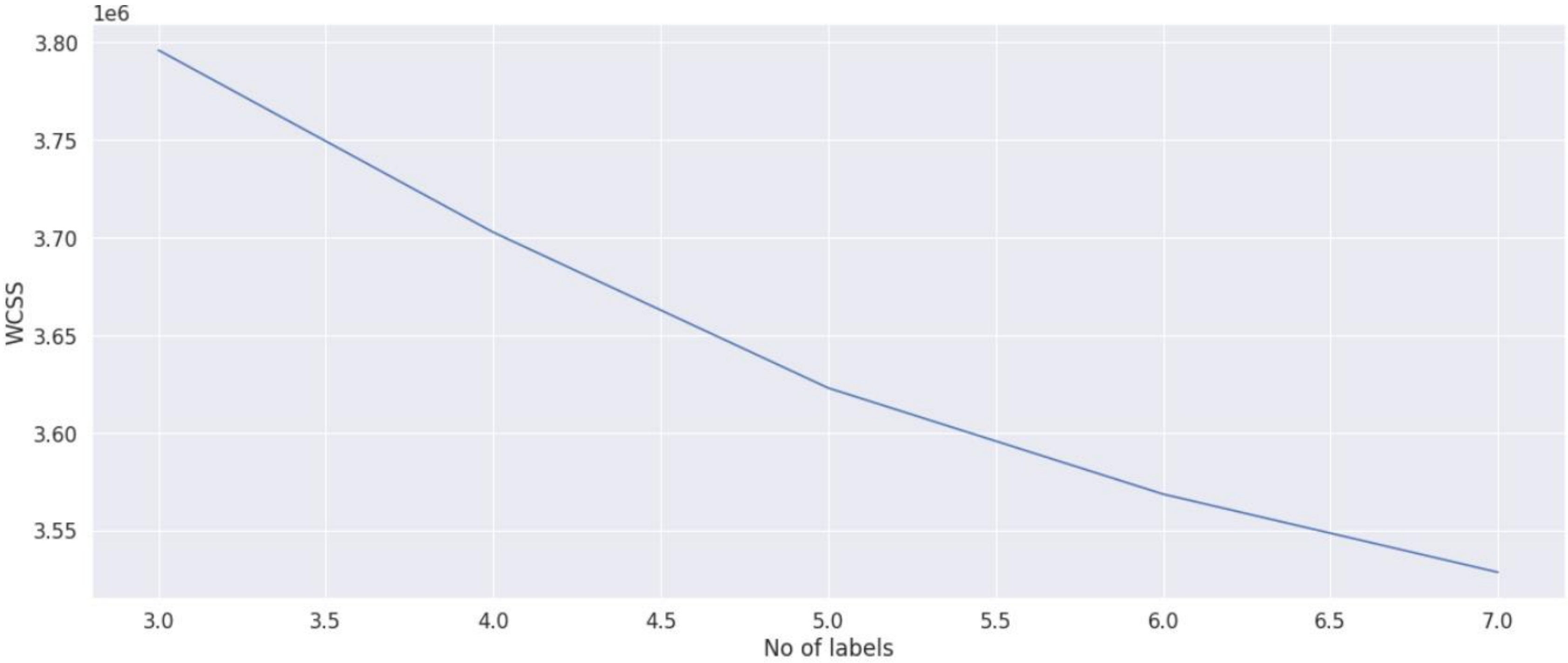
TSNE representation



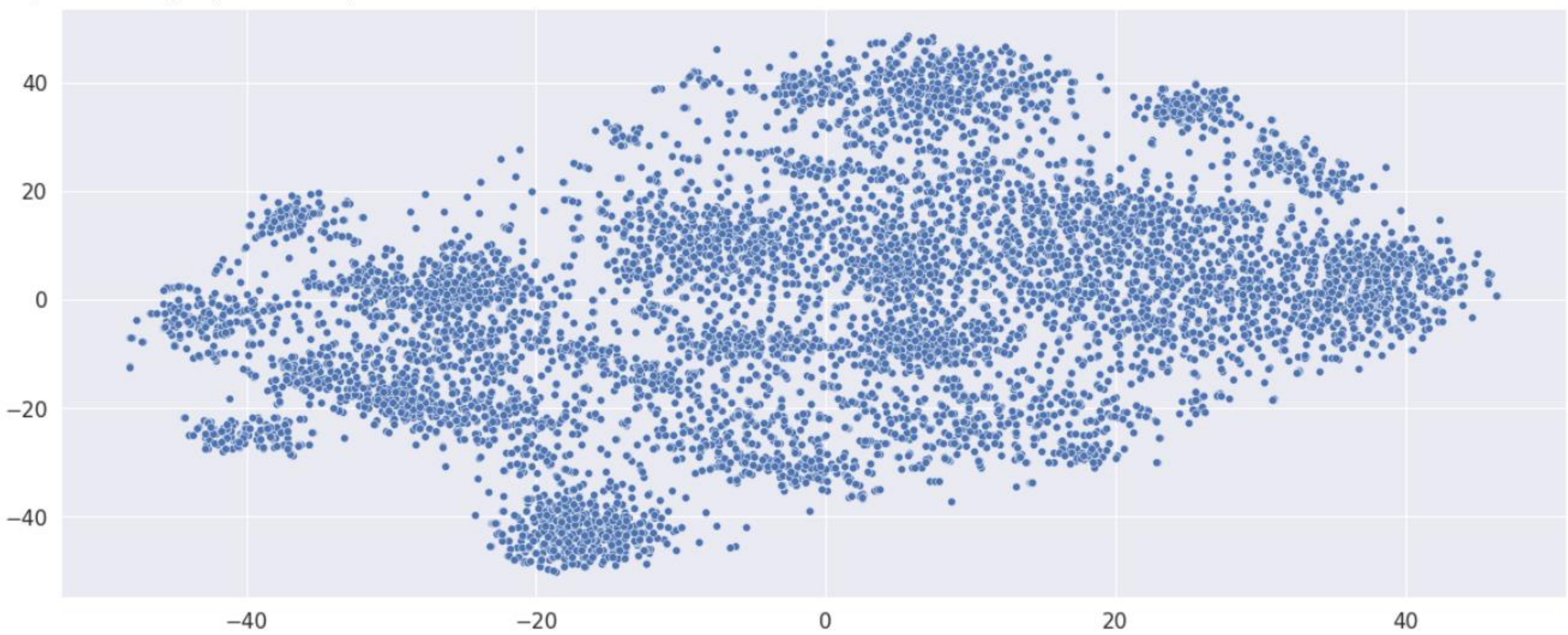
T-sne plot only for the embeddings created



A new dataframe is created by concatting the columns listed_in and description. This dataframe would be used to create word embeddings. Using the new dataframe we crate the word embeddings again. We use KMeans to compute the wcss and then plot the elbow curve. We observe a curve at 5.

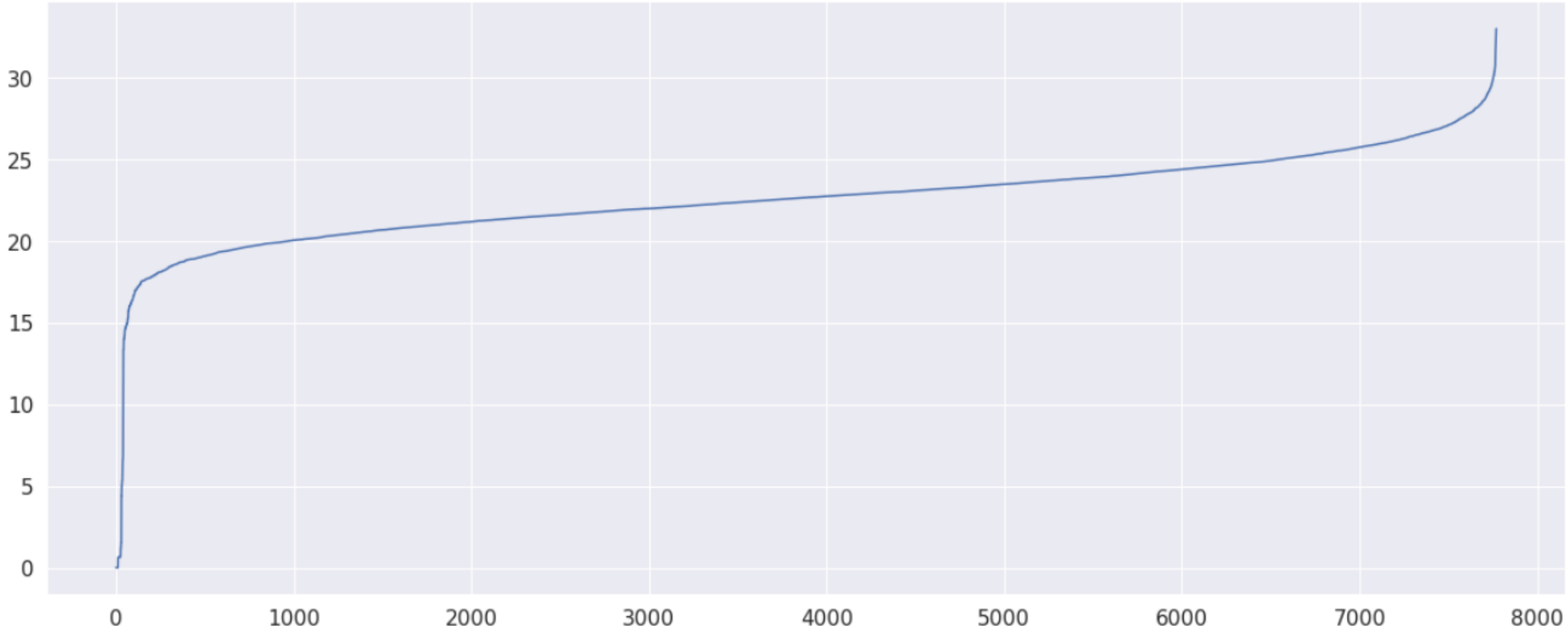


Here we merge the previously encoded columns with the new dataframe for embeddings. We use the index to merge to merge as we do not have any common column. Again with the new dataset we use Kmeans to obtain wcss and plot the elbow curve. We observe a curve at 5. so, we proceed with five number of clusters for our data.



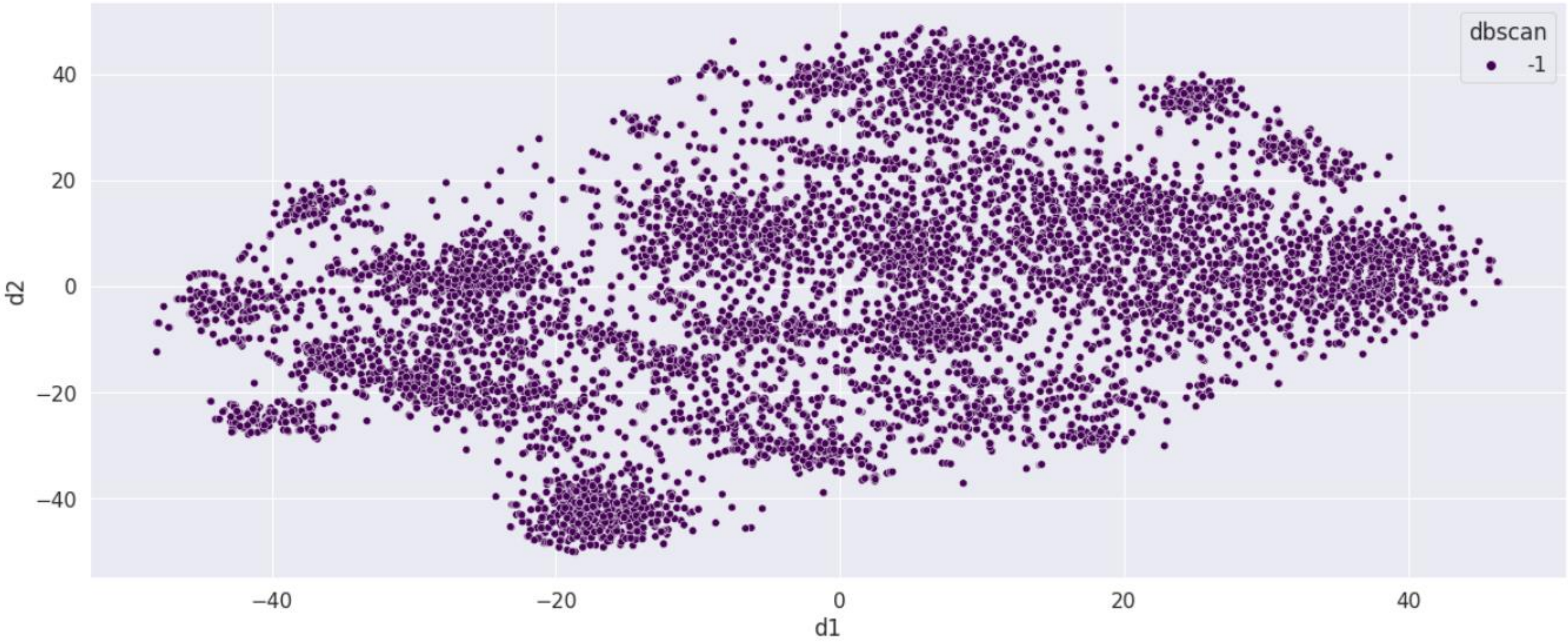
Using TSNE we plot the data points in two dimensions. The plot gives us an idea of the points where the clusters could be.

```
[<matplotlib.lines.Line2D at 0x7fbe8f352d50>]
```


30
25
20
15
10
5
0
0
1000
2000
3000
4000
5000
6000
7000
8000

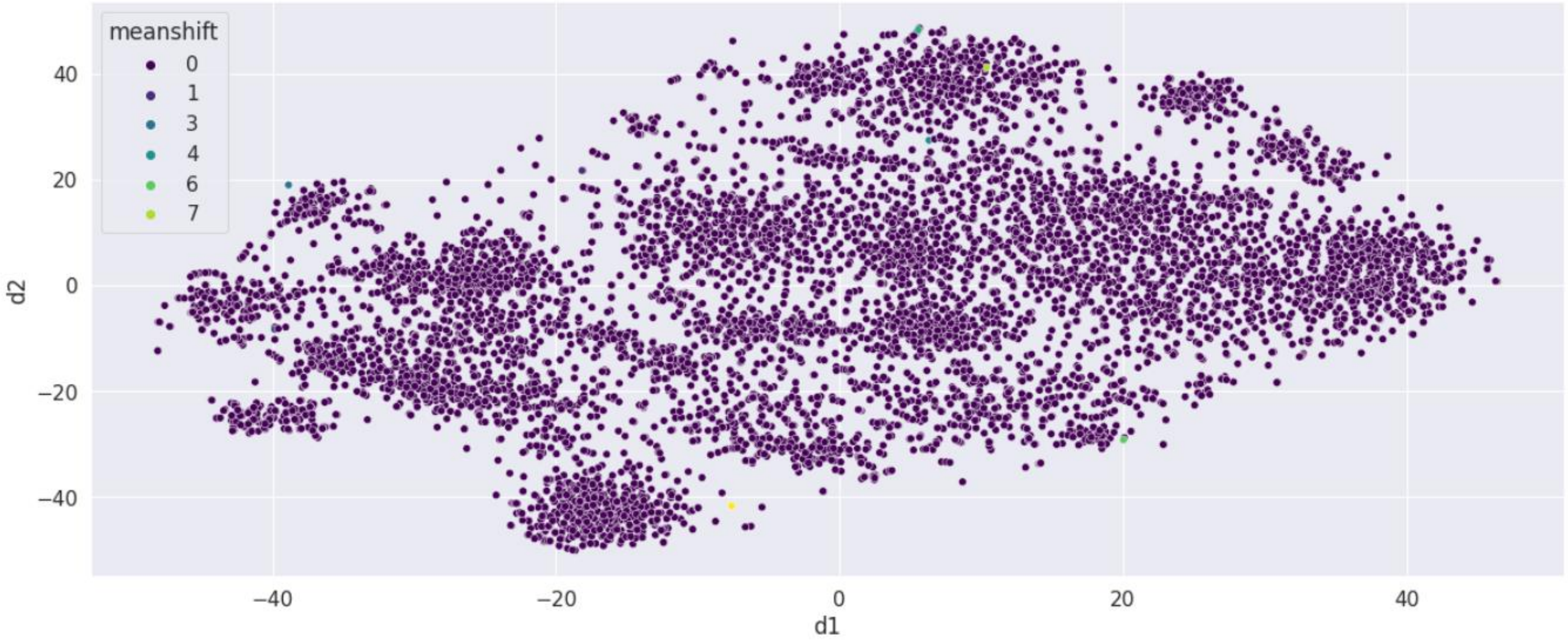
Clustering Algorithm

DBSCAN



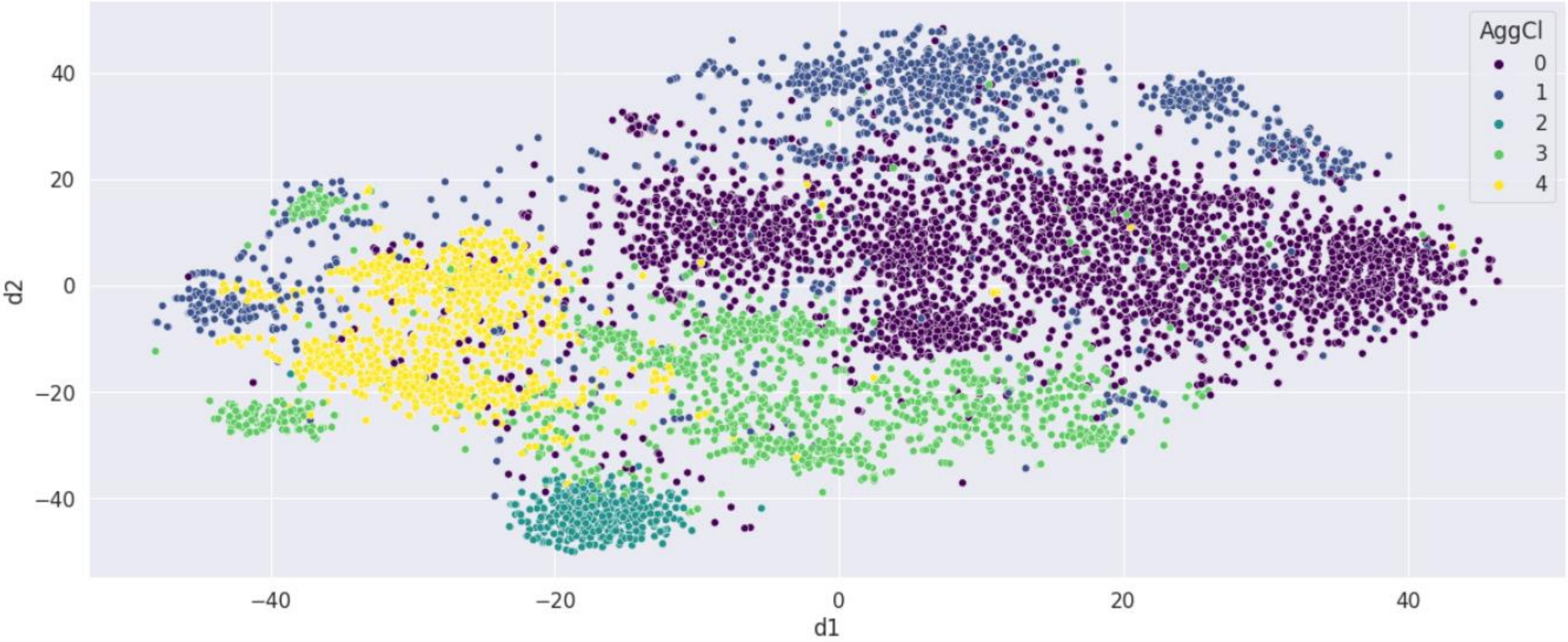
- Using DBSCAN we observe it is not able to mark the clusters

Mean Shift



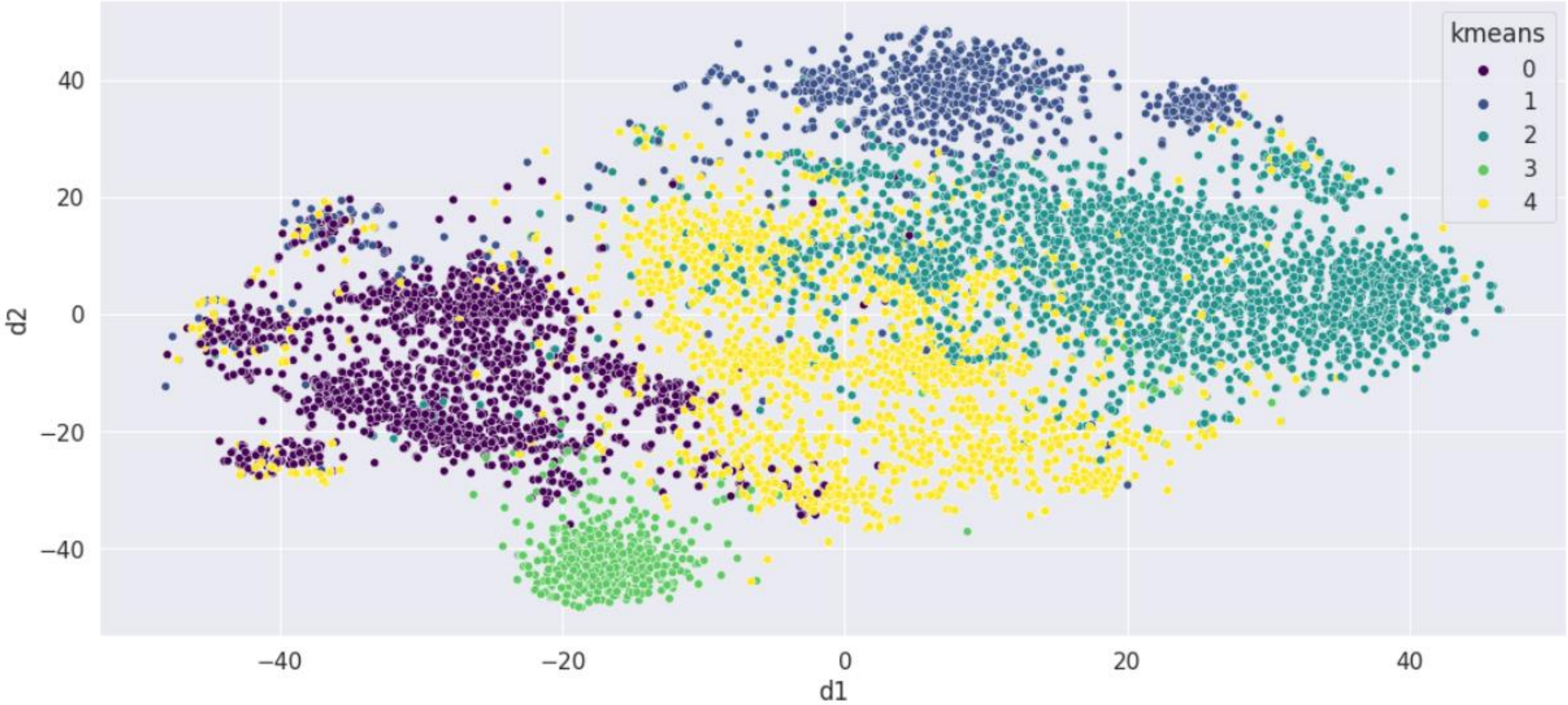
- Mean Shift seems to be not able to properly cluster the data.

Agglomerative



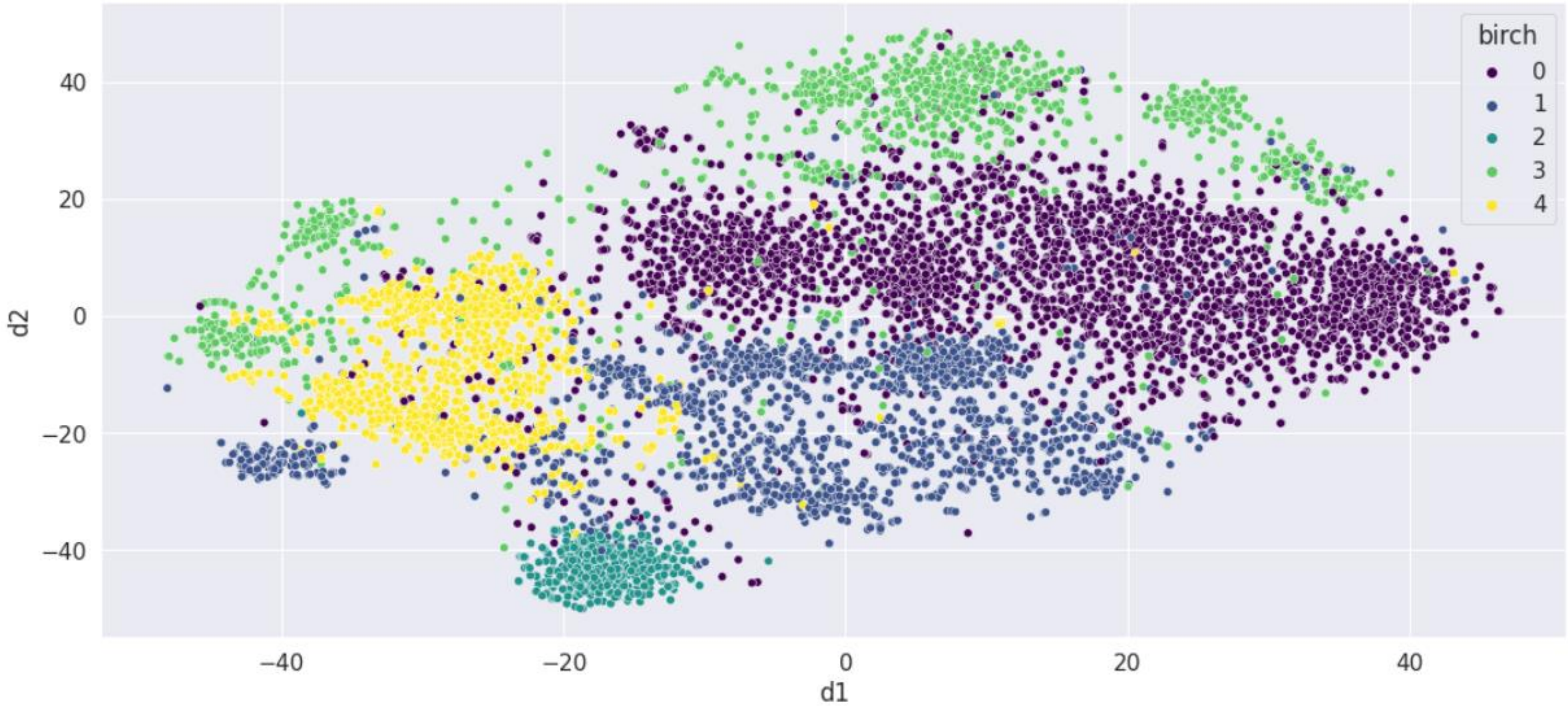
- The Agglomerative clustering gives us a good visual markers for the clusters, but we can observe the green and blue marked points are not well defined in this.

KMeans



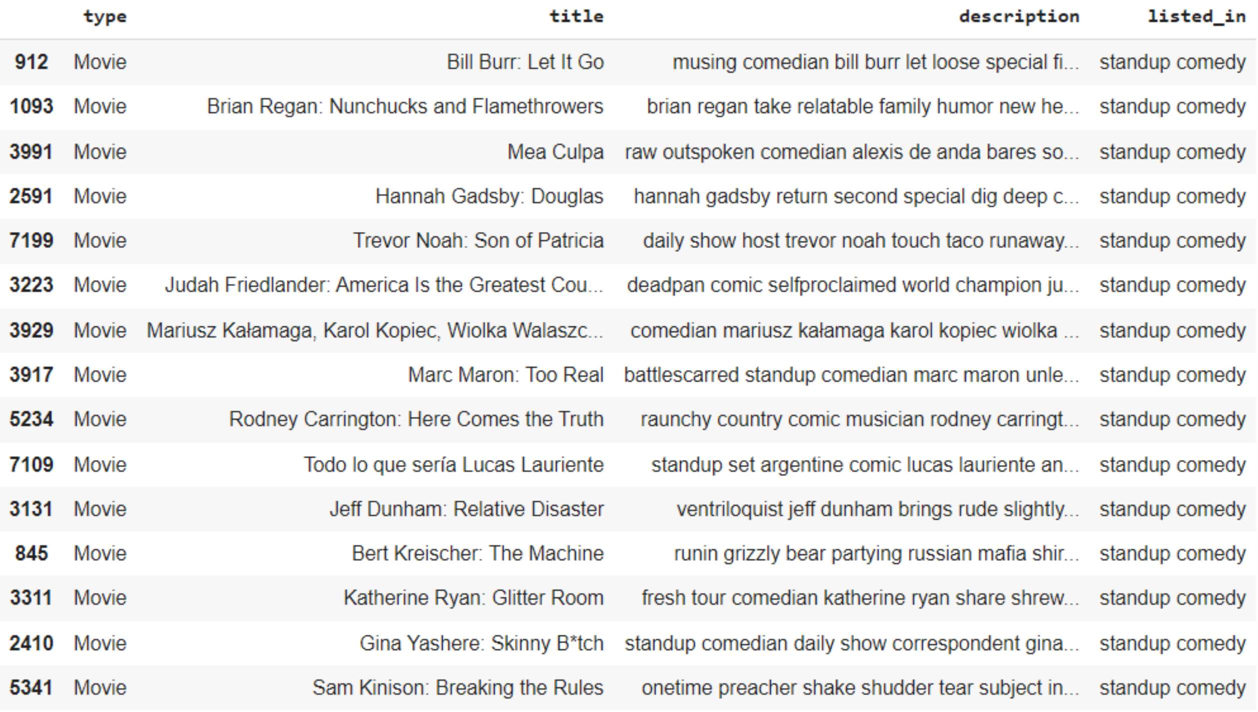
- With Kmeans also we get well separated clusters with some overlap in the points marked in blue and yellow.

BIRCH



The BIRCH model also gives a good estimate of clusters. The points marked in light green is very spread out.

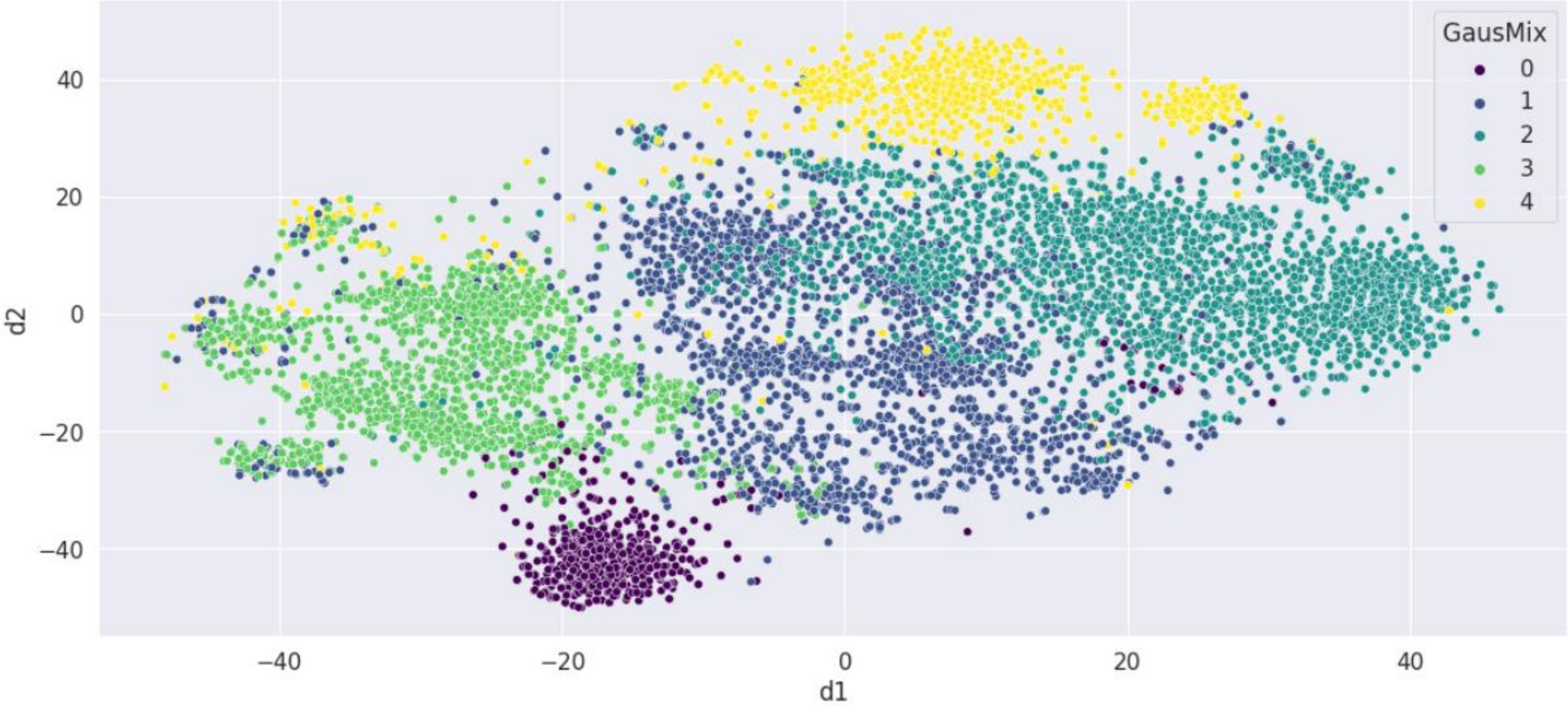
BIRCH



	type	title	description	listed_in
912	Movie	Bill Burr: Let It Go	musing comedian bill burr let loose special fi...	standup comedy
1093	Movie	Brian Regan: Nunchucks and Flamethrowers	brian regan take relatable family humor new he...	standup comedy
3991	Movie	Mea Culpa	raw outspoken comedian alexis de anda bares so...	standup comedy
2591	Movie	Hannah Gadsby: Douglas	hannah gadsby return second special dig deep c...	standup comedy
7199	Movie	Trevor Noah: Son of Patricia	daily show host trevor noah touch taco runaway...	standup comedy
3223	Movie	Judah Friedlander: America Is the Greatest Cou...	deadpan comic selfproclaimed world champion ju...	standup comedy
3929	Movie	Mariusz Kałamaga, Karol Kopiec, Wiolka Walaszc...	comedian mariusz kałamaga karol kopiec wiolka ...	standup comedy
3917	Movie	Marc Maron: Too Real	battlescarred standup comedian marc maron unle...	standup comedy
5234	Movie	Rodney Carrington: Here Comes the Truth	raunchy country comic musician rodney carringt...	standup comedy
7109	Movie	Todo lo que sería Lucas Lauriente	standup set argentine comic lucas lauriente an...	standup comedy
3131	Movie	Jeff Dunham: Relative Disaster	ventriloquist jeff dunham brings rude slightly...	standup comedy
845	Movie	Bert Kreischer: The Machine	runin grizzly bear partying russian mafia shir...	standup comedy
3311	Movie	Katherine Ryan: Glitter Room	fresh tour comedian katherine ryan share shrew...	standup comedy
2410	Movie	Gina Yashere: Skinny B*tch	standup comedian daily show correspondent gina...	standup comedy
5341	Movie	Sam Kinison: Breaking the Rules	onetime preacher shake shudder tear subject in...	standup comedy

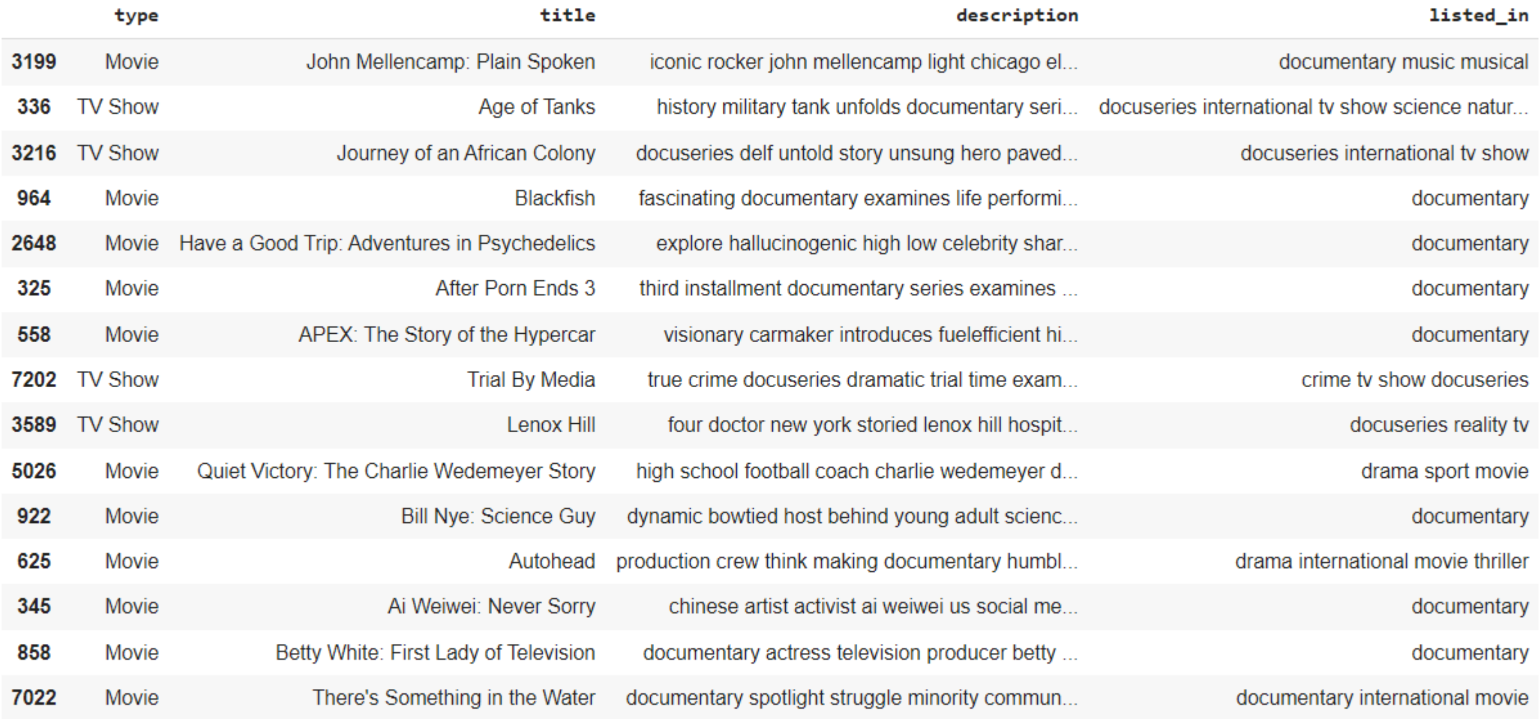
The cluster number 2 from BIRCH model is "Standup comedy"

Gaussian Mixture



The Gaussian Mixture model also does a decent job in indenting the clusters. But, the centrally located points are overlapping.

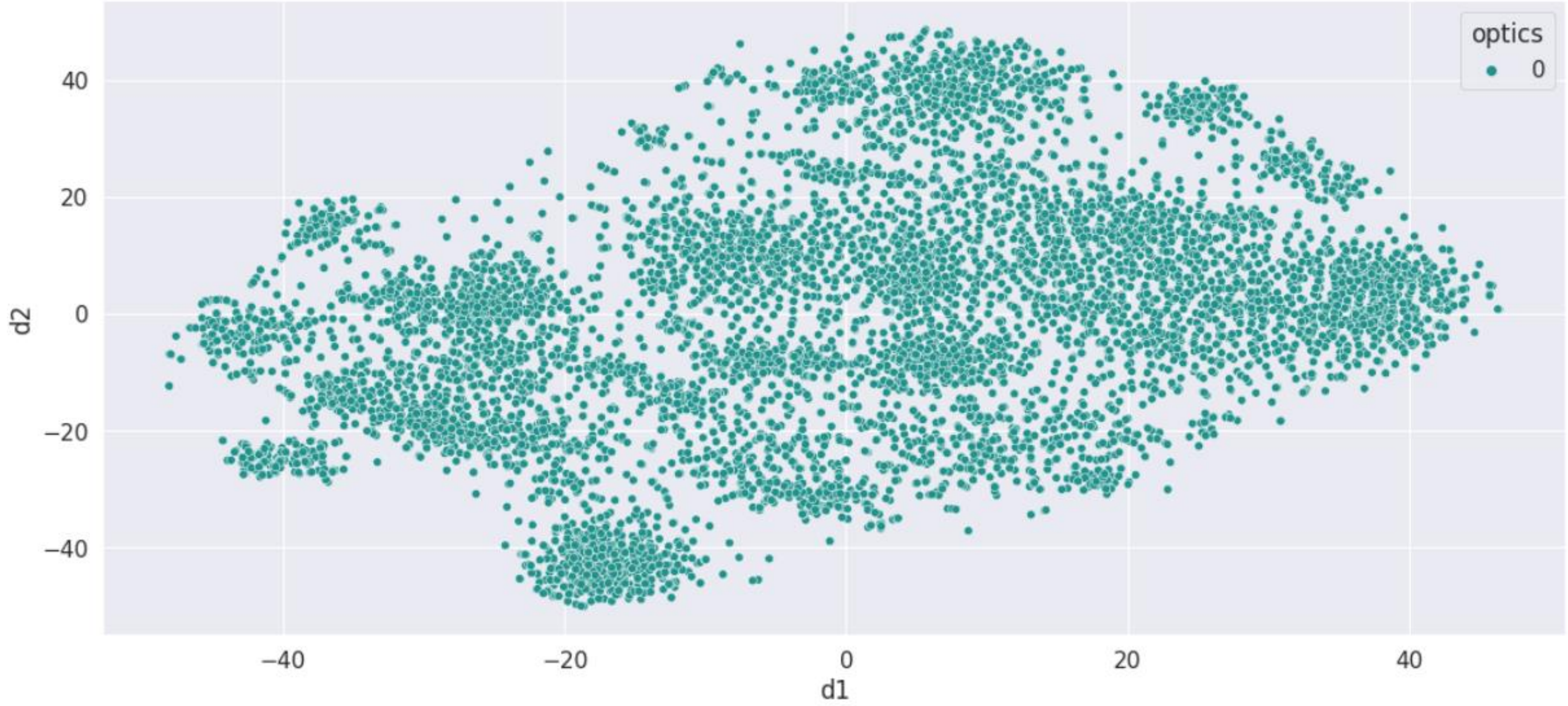
Gaussian Mixture



	type	title	description	listed_in
3199	Movie	John Mellencamp: Plain Spoken	iconic rocker john mellencamp light chicago el...	documentary music musical
336	TV Show	Age of Tanks	history military tank unfolds documentary seri...	docuseries international tv show science natur...
3216	TV Show	Journey of an African Colony	docuseries delf untold story unsung hero paved...	docuseries international tv show
964	Movie	Blackfish	fascinating documentary examines life performi...	documentary
2648	Movie	Have a Good Trip: Adventures in Psychedelics	explore hallucinogenic high low celebrity shar...	documentary
325	Movie	After Porn Ends 3	third installment documentary series examines ...	documentary
558	Movie	APEX: The Story of the Hypercar	visionary carmaker introduces fuelefficient hi...	documentary
7202	TV Show	Trial By Media	true crime docuseries dramatic trial time exam...	crime tv show docuseries
3589	TV Show	Lenox Hill	four doctor new york storied lenox hill hospit...	docuseries reality tv
5026	Movie	Quiet Victory: The Charlie Wedemeyer Story	high school football coach charlie wedemeyer d...	drama sport movie
922	Movie	Bill Nye: Science Guy	dynamic bowtied host behind young adult scienc...	documentary
625	Movie	Autohead	production crew think making documentary humbl...	drama international movie thriller
345	Movie	Ai Weiwei: Never Sorry	chinese artist activist ai weiwei us social me...	documentary
858	Movie	Betty White: First Lady of Television	documentary actress television producer betty ...	documentary
7022	Movie	There's Something in the Water	documentary spotlight struggle minority commun...	documentary international movie

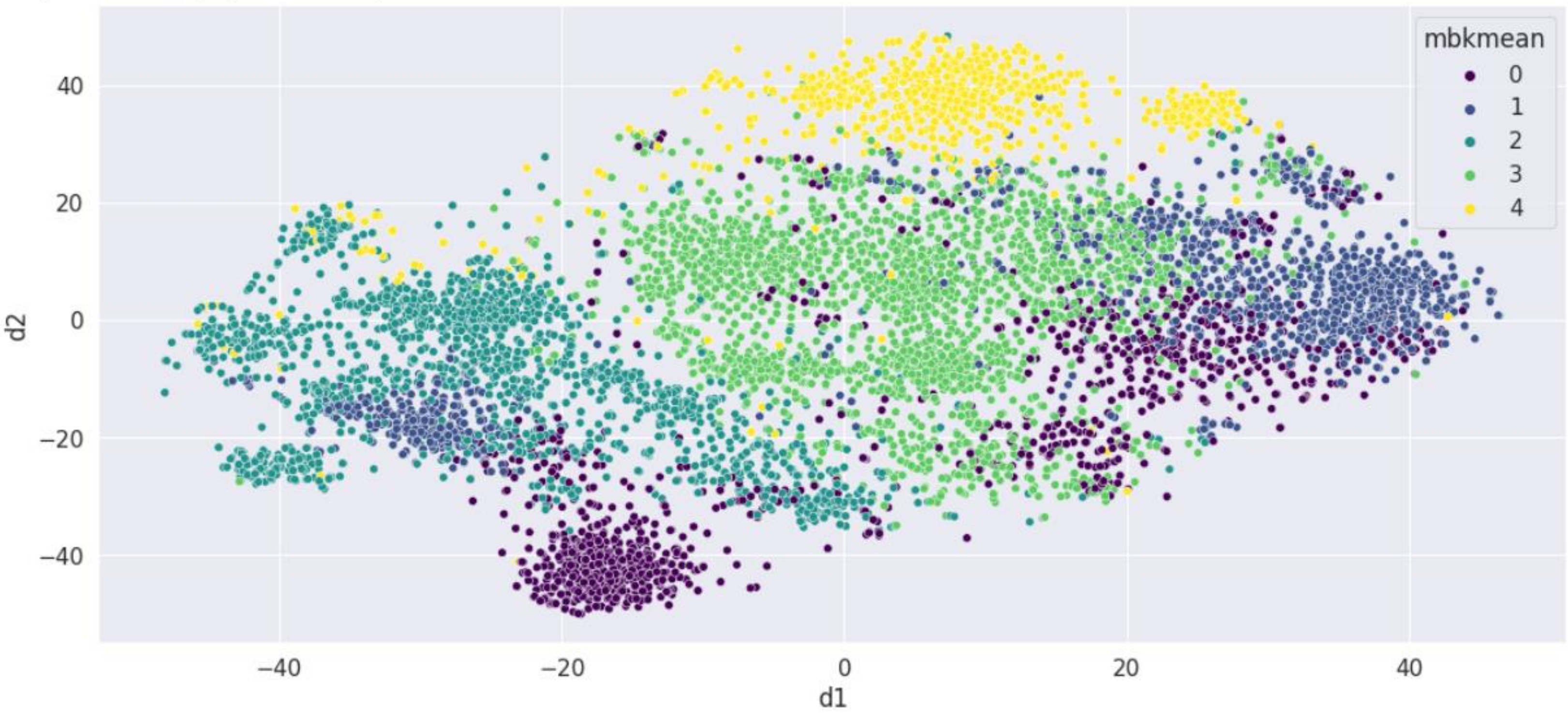
The cluster number 3 has mostly TV dramas.

Optics



With OPTICS we observer the model is not able to segregate the clusters.

Mini-Batch Kmeans



MiniBatch Kmeans also identifies the clusters but the points marked in point seem to overlap with other cluster.

Conclusion:

- ❑ From the different clustering algorithms we trained our data on, DBSCAN and Mean Shift seems to be not able to properly cluster the data.
- ❑ Agglomerative Clustering, BIRCH, KMeans and Gaussian mixture does a good job on identifying the clusters. So these models can be used for Future work to further tune and produce better results.

References:

- Kaggle competition
- Analytics vidhya

Thank You!!