Capstone Project-4

Un-Supervised ML on Netflix Movie and TV shows (Clustering)

- Shivaswaroop J P(Pro Flex)
- swaroopjp56@gmail.com

Contents

1. Problem Statement
2. Introduction
3. Data Cleaning and Data Viz
4. Data Modelling and implementation
5. Conclusion

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.

Introduction

- Netflix, Inc. is an American over-the-top content platform and production company headquartered in Los Gatos, California. The company's primary business is a subscription-based streaming service offering online streaming from a library of films and television series, including those produced in-house.

- The streaming platform has increased his catalogue substantially in his last 10 years of existence. Netflix has way more films than all his competitors, such as HBO or Amazon video, which are following Netflix`s steps in order to obtain the same success.

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
- 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 Release year of the movie / show
- rating : TV Rating of the movie / show
- duration : Total Duration - in minutes or number of seasons
- listed_in : Genre
- description: The Summary description

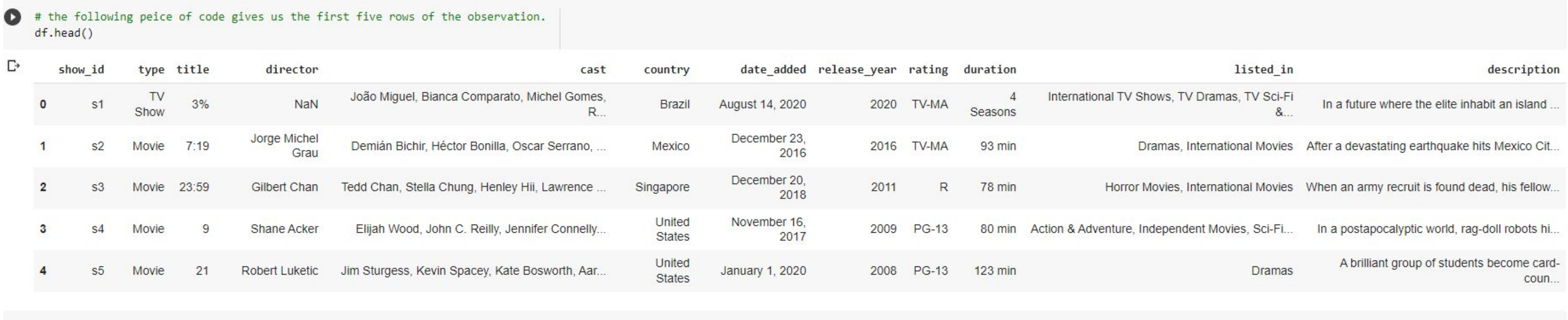
Data Wrangling

- Data wrangling is the process of cleaning and unifying messy and complex data sets for easy access and analysis.

- Our Dataset includes about 12 columns and about 7787 observations

- We don’t have any target variable as this is an unsupervised algorithm.

The snippet of Our dataset looks like:



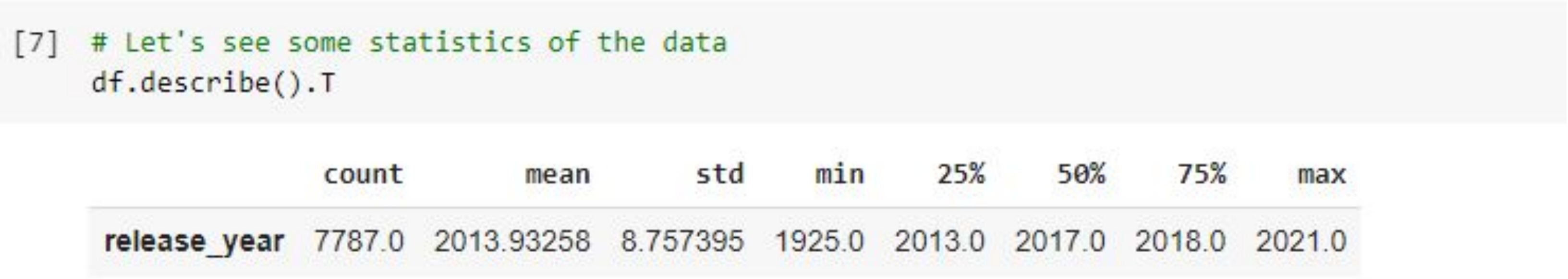
```
# the following peice of code gives us the first five rows of the observation.
df.head()
```

	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, 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...

The df.head() method shows us the first 5 rows of the Dataset.

Let's look at some statistics of the Data

- The Statistics of the data could be found out from an inbuilt function in pandas library called describe() .

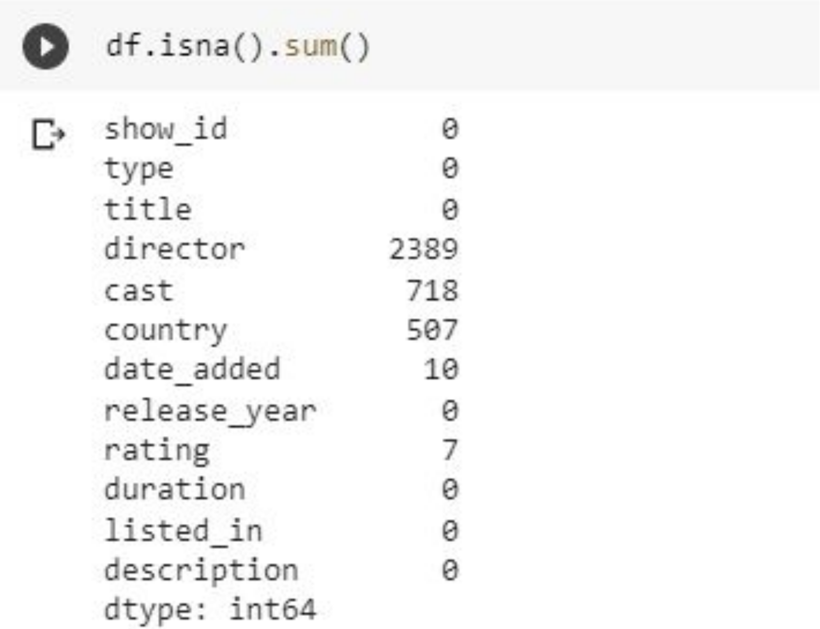


```
[7]  # Let's see some statistics of the data
     df.describe().T
```

	count	mean	std	min	25%	50%	75%	max
release_year	7787.0	2013.93258	8.757395	1925.0	2013.0	2017.0	2018.0	2021.0

- We can see only one row because, all other rows are string values and statistics cannot be defined for string values.

Checking for any NULL values in the whole Dataset



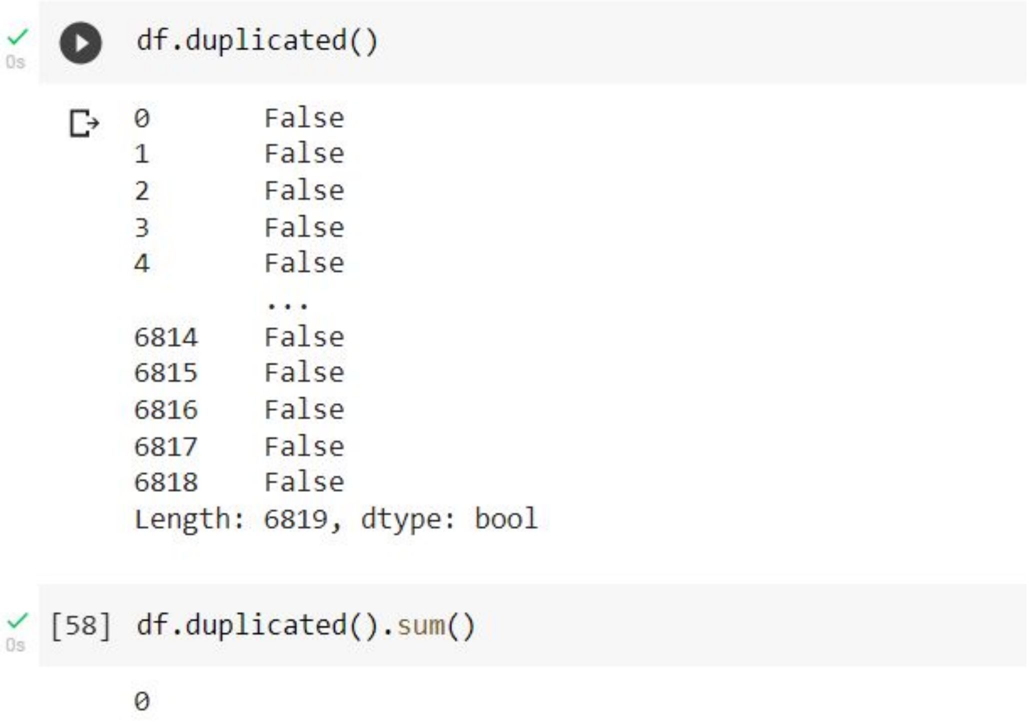
```
df.isna().sum()
```

```
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
```

- There are quite some null Values in our dataset

- There are Null values in the features:
  1. Director
  2. Cast
  3. Country
  4.date_added
  5. rating

Checking for any Duplicate Values:



```
df.duplicated()
```

```
0       False
1       False
2       False
3       False
4       False
        ...
6814    False
6815    False
6816    False
6817    False
6818    False
Length: 6819, dtype: bool
```

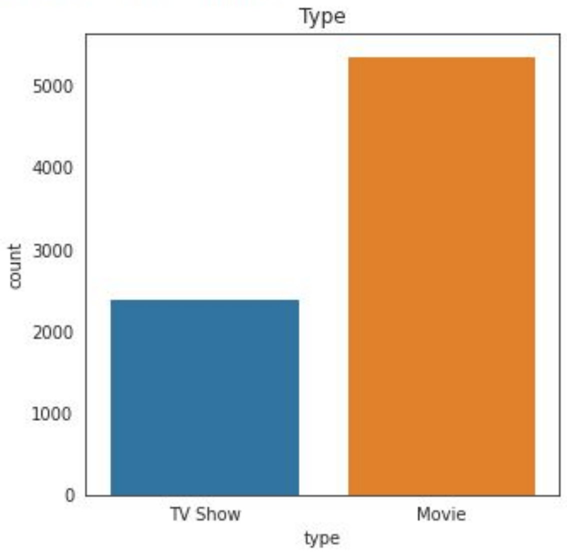
```
[58] df.duplicated().sum()
```

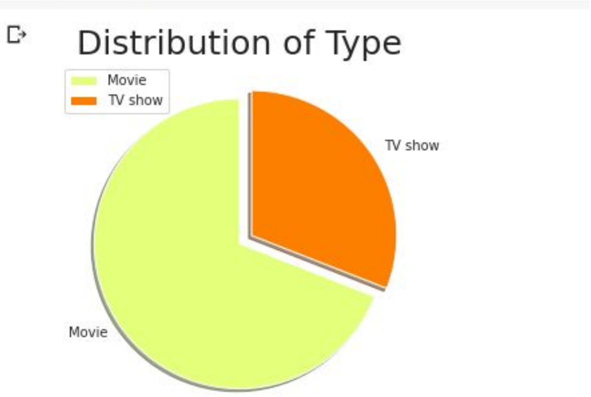
```
0
```

- So, It is evident that there are no duplicate values in our whole dataset.

Lets begin with visualizations

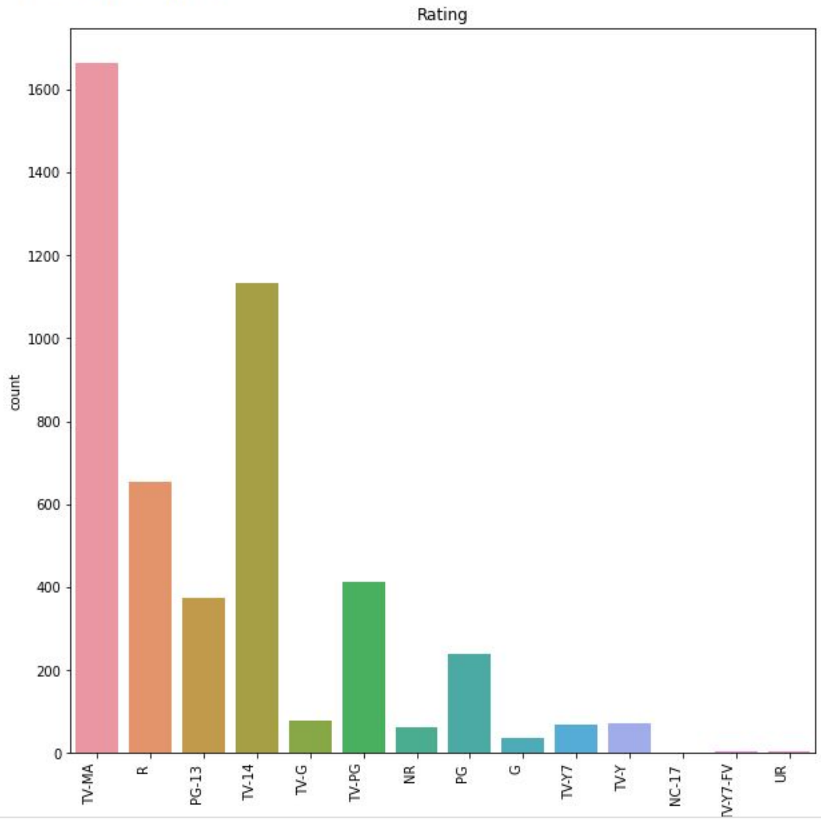
1. Checking the Distribution the variable type(movie or TV Show)





- So, Could be seen around 65 % of the data we have is of the type ‘Movie’ and the rest is of type ‘TV Show’.

2. Lets plot the frequency of different kinds of ratings in our dataset

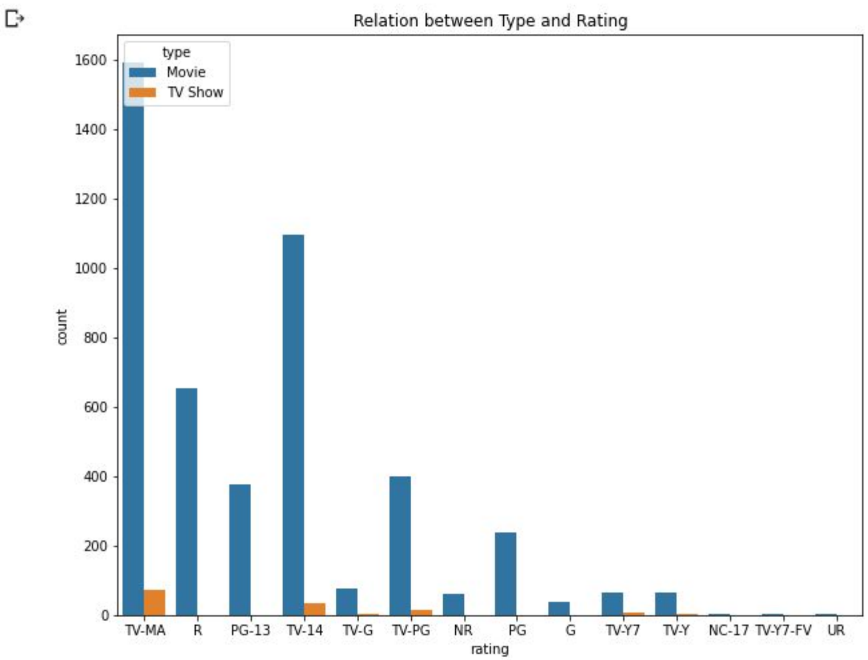


- So, there are 14 different types of ratings in our whole dataset

- On top of the list stands ‘TV-MA’ rating (TV-Mature Audience)

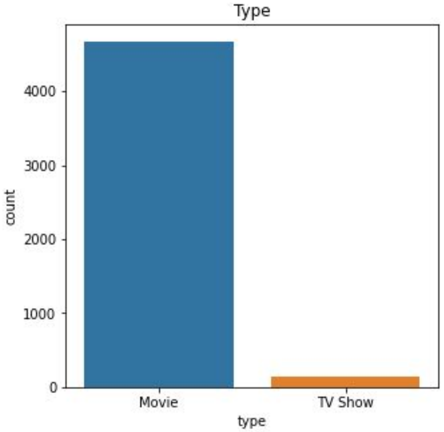
- Last of the list are NC-17 and UR.

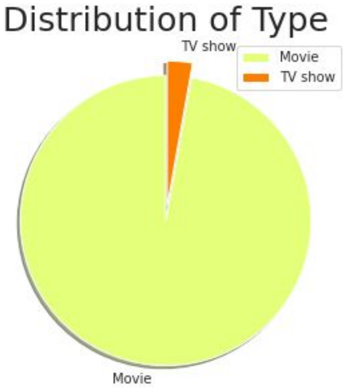
3. Relation between Type and Rating



The above picture shows the relation between type (TV show and Movie) and their respective ratings.

3. Let’s plot the distribution of the data after dropping null Values



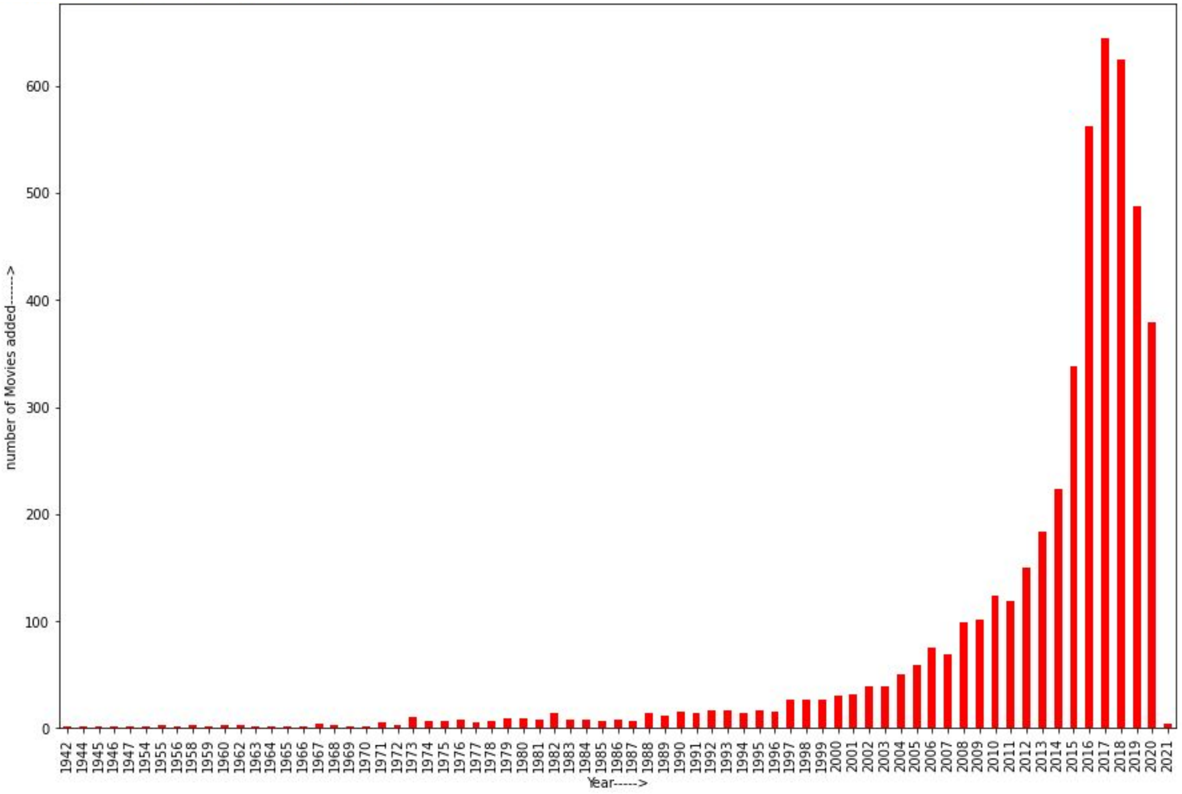


- As we can see, We have lost a lot of data

- Above 2000 rows to be precise

- And also we have lost a lot of data from the type “TV show” which can hamper our future predictions

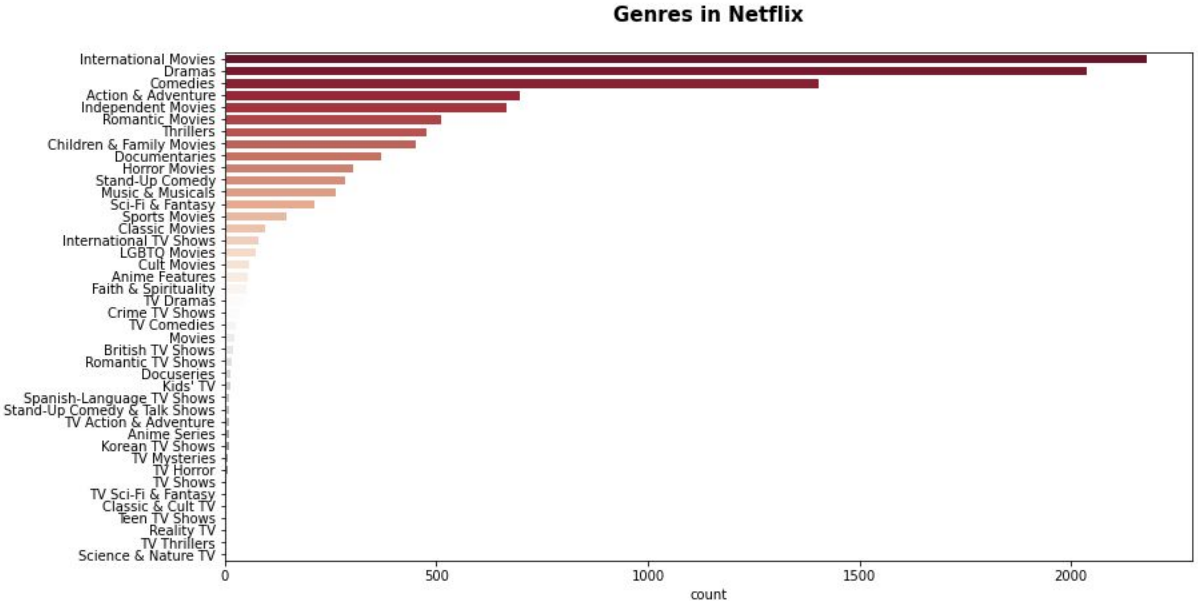
4. Plotting the frequency of number of titles released per year



- The number of titles generated per year increases from the year 1942 till 2017 .

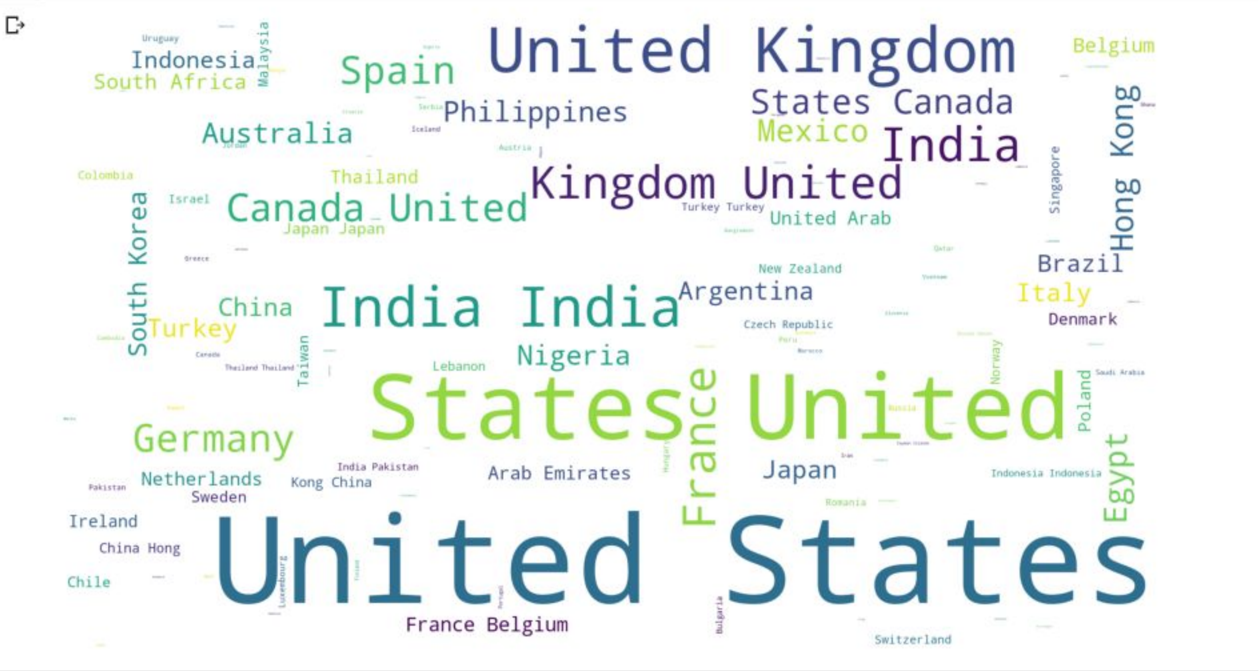
- Then we can see a gradual decrease in the number from 2019 ,whose credit goes to Covid-19.

5. Plotting the frequency of the different Genres in the titles released in Netflix



- Plotting the frequency, the Genre “International Movies” stands on first followed by ‘Dramas” and “Action and Adventures”.

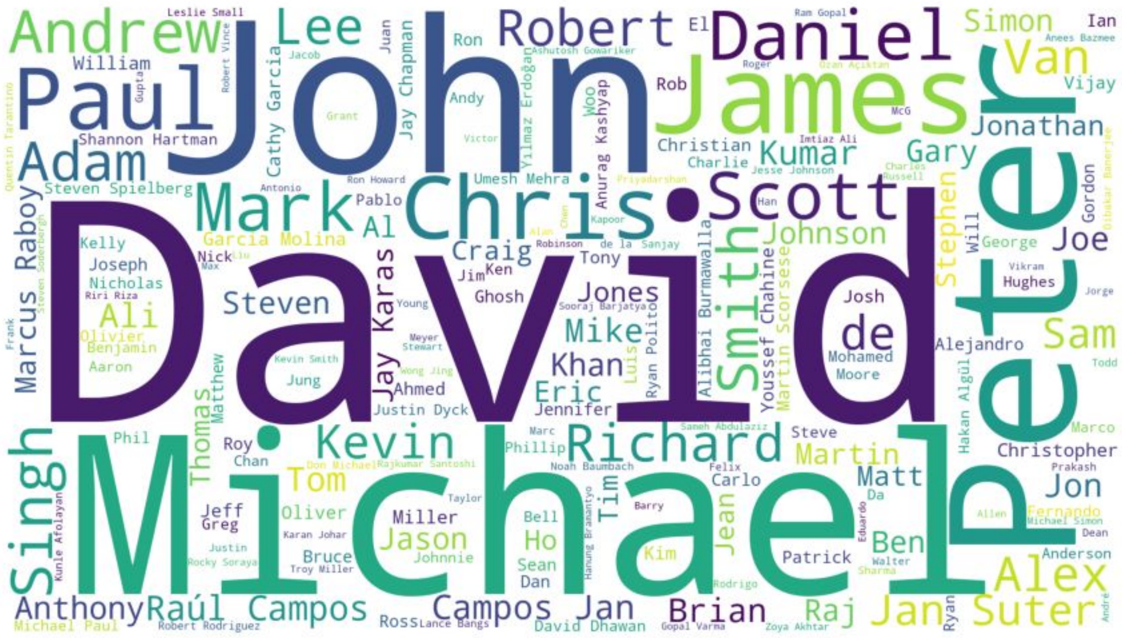
6. Plotting a Word Cloud of different Countries



- A Word Cloud is an image composed of words used in a particular text or subject, in which the size of each word indicates its frequency or importance..

- AS it can be seen from this picture, United states Comes first and next on list is India and United Kingdom

7. Plotting Word Cloud of different Directors



- Plotting a Word Cloud of directors we, find out that David is the most common name that comes up for a director and following that name is John, Michael and peter.

Data Modelling

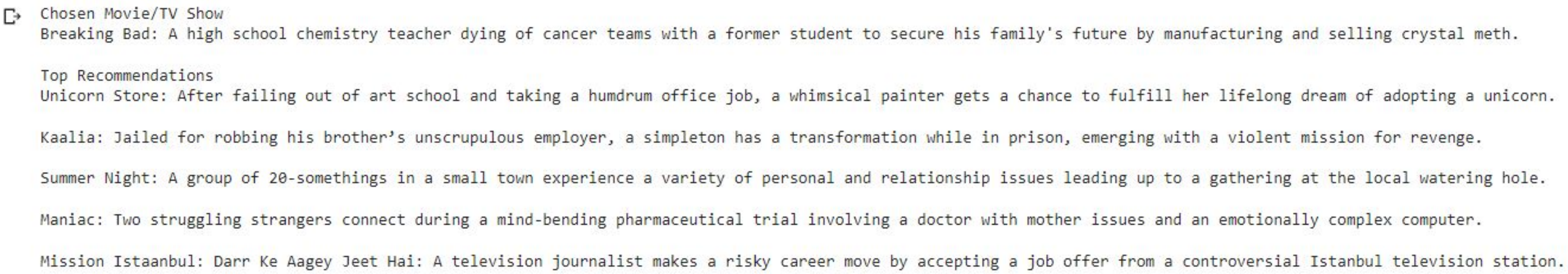
1. Natural Language Processing(NLP):

- Natural language processing (NLP) is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data.

- The goal is a computer capable of "understanding" the contents of documents, including the contextual nuances of the language within them.

Testing NLP:

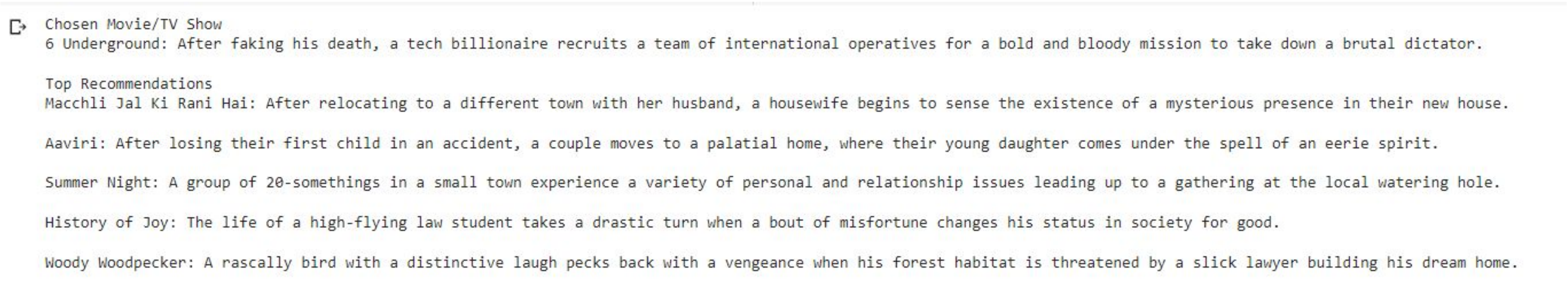
- 1. For the TV Show “Breaking Bad”



```
Chosen Movie/TV Show
Breaking Bad: A high school chemistry teacher dying of cancer teams with a former student to secure his family's future by manufacturing and selling crystal meth.

Top Recommendations
Unicorn Store: After failing out of art school and taking a humdrum office job, a whimsical painter gets a chance to fulfill her lifelong dream of adopting a unicorn.

Kaalia: Jailed for robbing his brother’s unscrupulous employer, a simpleton has a transformation while in prison, emerging with a violent mission for revenge.

Summer Night: A group of 20-somethings in a small town experience a variety of personal and relationship issues leading up to a gathering at the local watering hole.

Maniac: Two struggling strangers connect during a mind-bending pharmaceutical trial involving a doctor with mother issues and an emotionally complex computer.

Mission Istaanbul: Darr Ke Aagey Jeet Hai: A television journalist makes a risky career move by accepting a job offer from a controversial Istanbul television station.
```

- The NLP(Natural Language Processing) gives us the recommendation for the TV Show ‘Breaking Bad’ are: “ Unicorn Store”, “Summer Night”, “Maniac”, “Mission Istanbul” and “Kaalia”.

2. for the movie “6 Underground”



```
Chosen Movie/TV Show
6 Underground: After faking his death, a tech billionaire recruits a team of international operatives for a bold and bloody mission to take down a brutal dictator.

Top Recommendations
Macchli Jal Ki Rani Hai: After relocating to a different town with her husband, a housewife begins to sense the existence of a mysterious presence in their new house.

Aaviri: After losing their first child in an accident, a couple moves to a palatial home, where their young daughter comes under the spell of an eerie spirit.

Summer Night: A group of 20-somethings in a small town experience a variety of personal and relationship issues leading up to a gathering at the local watering hole.

History of Joy: The life of a high-flying law student takes a drastic turn when a bout of misfortune changes his status in society for good.

Woody Woodpecker: A rascally bird with a distinctive laugh pecks back with a vengeance when his forest habitat is threatened by a slick lawyer building his dream home.
```

- The NLP(Natural Language Processing) gives us the recommendation for the movie “6 underground” , are: “machali Jal ki rani Hai”, “aaviri”, “Summer Night”, “History of Joy”, “Woody Woodpecker”

- The Recommendations provided by the model NLP are not up to the mark and lets proceed to our next model “ K-means” Clustering

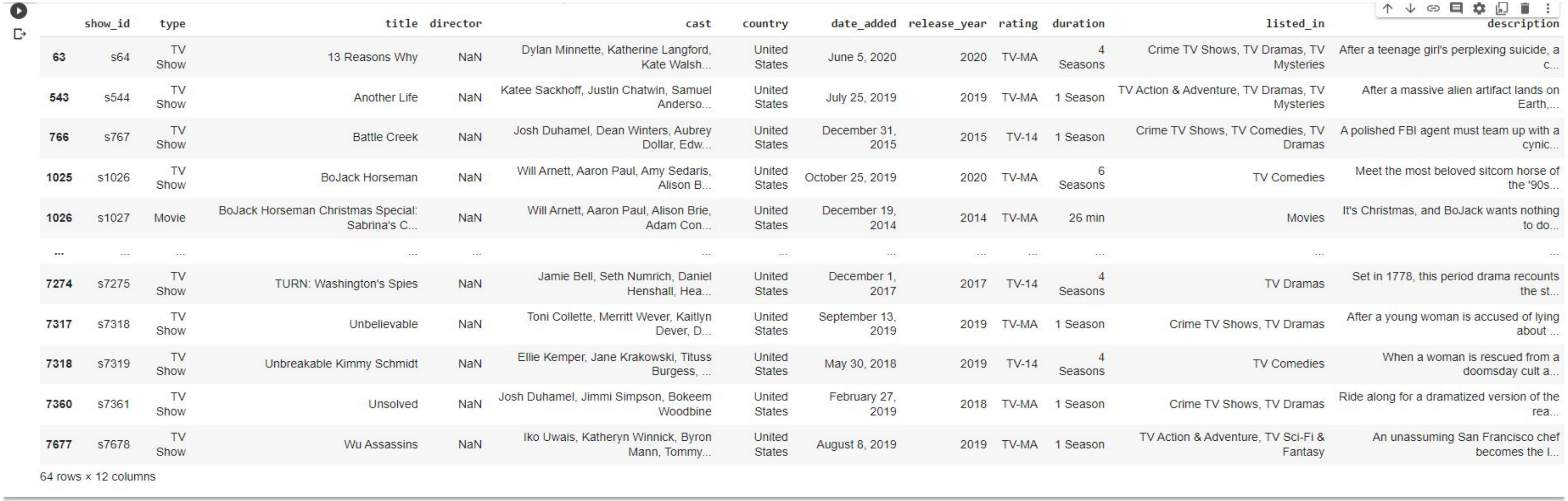
2. K-Means Clustering

- *k*-means clustering is a method of vector quantization, originally from signal processing, that aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster.

- *k*-means clustering minimizes within-cluster variances (squared Euclidean distances)

Testing K-Means Clustering

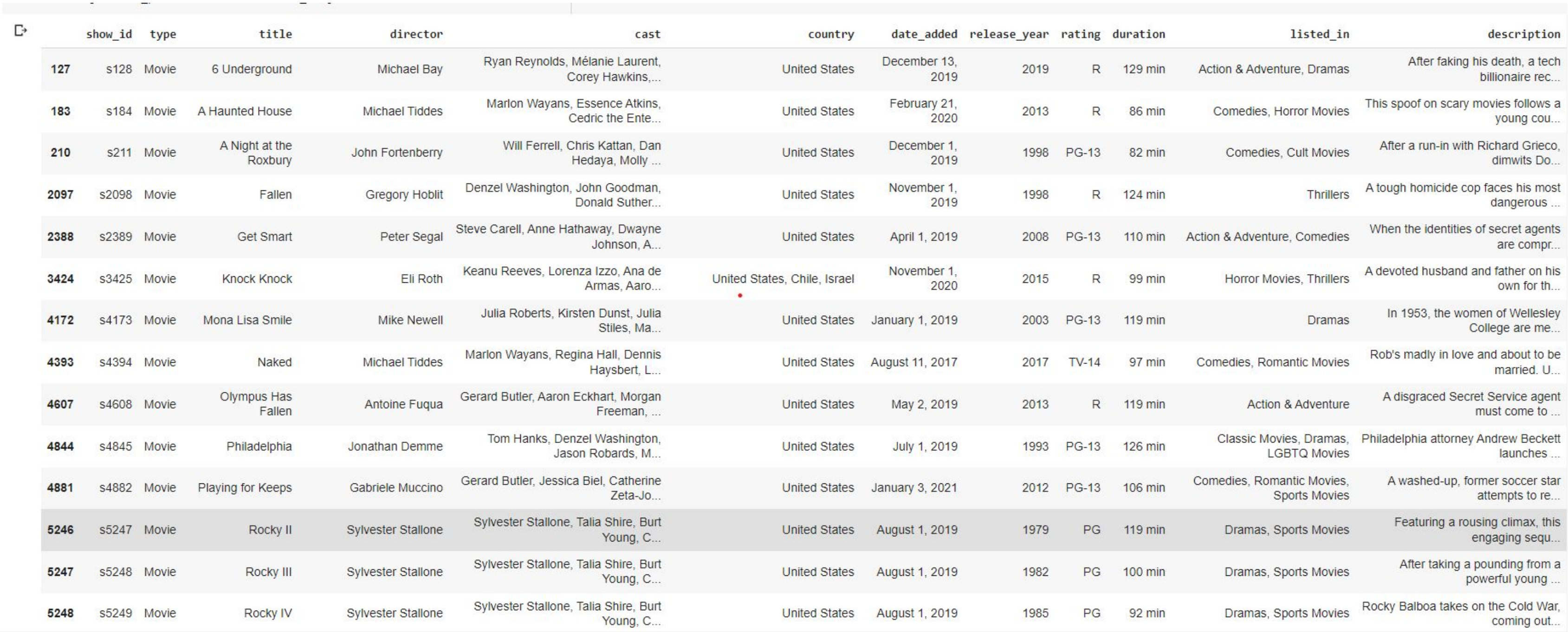
- 1. Testing the model K-means Clustering for the same set of titles



	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
63	s64	TV Show	13 Reasons Why	NaN	Dylan Minnette, Katherine Langford, Kate Walsh...	United States	June 5, 2020	2020	TV-MA	4 Seasons	Crime TV Shows, TV Dramas, TV Mysteries	After a teenage girl's perplexing suicide, a c...
543	s544	TV Show	Another Life	NaN	Katee Sackhoff, Justin Chatwin, Samuel Anderso...	United States	July 25, 2019	2019	TV-MA	1 Season	TV Action & Adventure, TV Dramas, TV Mysteries	After a massive alien artifact lands on Earth,...
766	s767	TV Show	Battle Creek	NaN	Josh Duhamel, Dean Winters, Aubrey Dollar, Edw...	United States	December 31, 2015	2015	TV-14	1 Season	Crime TV Shows, TV Comedies, TV Dramas	A polished FBI agent must team up with a cynic...
1025	s1026	TV Show	BoJack Horseman	NaN	Will Arnett, Aaron Paul, Amy Sedaris, Alison B...	United States	October 25, 2019	2020	TV-MA	6 Seasons	TV Comedies	Meet the most beloved sitcom horse of the '90s...
1026	s1027	Movie	BoJack Horseman Christmas Special: Sabrina's C...	NaN	Will Arnett, Aaron Paul, Alison Brie, Adam Con...	United States	December 19, 2014	2014	TV-MA	26 min	Movies	It's Christmas, and BoJack wants nothing to do...
...	...	...	...	...	...	...	...	...	...	...	...	...
7274	s7275	TV Show	TURN: Washington's Spies	NaN	Jamie Bell, Seth Numrich, Daniel Henshall, Hea...	United States	December 1, 2017	2017	TV-14	4 Seasons	TV Dramas	Set in 1778, this period drama recounts the st...
7317	s7318	TV Show	Unbelievable	NaN	Toni Collette, Merritt Wever, Kaitlyn Dever, D...	United States	September 13, 2019	2019	TV-MA	1 Season	Crime TV Shows, TV Dramas	After a young woman is accused of lying about ...
7318	s7319	TV Show	Unbreakable Kimmy Schmidt	NaN	Ellie Kemper, Jane Krakowski, Tituss Burgess, ...	United States	May 30, 2018	2019	TV-14	4 Seasons	TV Comedies	When a woman is rescued from a doomsday cult a...
7360	s7361	TV Show	Unsolved	NaN	Josh Duhamel, Jimmi Simpson, Bokeem Woodbine	United States	February 27, 2019	2018	TV-MA	1 Season	Crime TV Shows, TV Dramas	Ride along for a dramatized version of the rea...
7677	s7678	TV Show	Wu Assassins	NaN	Iko Uwais, Katheryn Winnick, Byron Mann, Tommy...	United States	August 8, 2019	2019	TV-MA	1 Season	TV Action & Adventure, TV Sci-Fi & Fantasy	An unassuming San Francisco chef becomes the I...

64 rows × 12 columns

- The above picture shows prediction for the TV show “ Breaking Bad”



	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
127	s128	Movie	6 Underground	Michael Bay	Ryan Reynolds, Mélanie Laurent, Corey Hawkins,...	United States	December 13, 2019	2019	R	129 min	Action & Adventure, Dramas	After faking his death, a tech billionaire rec...
183	s184	Movie	A Haunted House	Michael Tiddes	Marlon Wayans, Essence Atkins, Cedric the Ente...	United States	February 21, 2020	2013	R	86 min	Comedies, Horror Movies	This spoof on scary movies follows a young cou...
210	s211	Movie	A Night at the Roxbury	John Fortenberry	Will Ferrell, Chris Kattan, Dan Hedaya, Molly ...	United States	December 1, 2019	1998	PG-13	82 min	Comedies, Cult Movies	After a run-in with Richard Grieco, dimwits Do...
2097	s2098	Movie	Fallen	Gregory Hoblit	Denzel Washington, John Goodman, Donald Suther...	United States	November 1, 2019	1998	R	124 min	Thrillers	A tough homicide cop faces his most dangerous ...
2388	s2389	Movie	Get Smart	Peter Segal	Steve Carell, Anne Hathaway, Dwayne Johnson, A...	United States	April 1, 2019	2008	PG-13	110 min	Action & Adventure, Comedies	When the identities of secret agents are compr...
3424	s3425	Movie	Knock Knock	Eli Roth	Keanu Reeves, Lorenza Izzo, Ana de Armas, Aaro...	United States, Chile, Israel	November 1, 2020	2015	R	99 min	Horror Movies, Thrillers	A devoted husband and father on his own for th...
4172	s4173	Movie	Mona Lisa Smile	Mike Newell	Julia Roberts, Kirsten Dunst, Julia Stiles, Ma...	United States	January 1, 2019	2003	PG-13	119 min	Dramas	In 1953, the women of Wellesley College are me...
4393	s4394	Movie	Naked	Michael Tiddes	Marlon Wayans, Regina Hall, Dennis Haysbert, L...	United States	August 11, 2017	2017	TV-14	97 min	Comedies, Romantic Movies	Rob's madly in love and about to be married. U...
4607	s4608	Movie	Olympus Has Fallen	Antoine Fuqua	Gerard Butler, Aaron Eckhart, Morgan Freeman, ...	United States	May 2, 2019	2013	R	119 min	Action & Adventure	A disgraced Secret Service agent must come to ...
4844	s4845	Movie	Philadelphia	Jonathan Demme	Tom Hanks, Denzel Washington, Jason Robards, M...	United States	July 1, 2019	1993	PG-13	126 min	Classic Movies, Dramas, LGBTQ Movies	Philadelphia attorney Andrew Beckett launches ...
4881	s4882	Movie	Playing for Keeps	Gabriele Muccino	Gerard Butler, Jessica Biel, Catherine Zeta-Jo...	United States	January 3, 2021	2012	PG-13	106 min	Comedies, Romantic Movies, Sports Movies	A washed-up, former soccer star attempts to re...
5246	s5247	Movie	Rocky II	Sylvester Stallone	Sylvester Stallone, Talia Shire, Burt Young, C...	United States	August 1, 2019	1979	PG	119 min	Dramas, Sports Movies	Featuring a rousing climax, this engaging sequ...
5247	s5248	Movie	Rocky III	Sylvester Stallone	Sylvester Stallone, Talia Shire, Burt Young, C...	United States	August 1, 2019	1982	PG	100 min	Dramas, Sports Movies	After taking a pounding from a powerful young ...
5248	s5249	Movie	Rocky IV	Sylvester Stallone	Sylvester Stallone, Talia Shire, Burt Young, C...	United States	August 1, 2019	1985	PG	92 min	Dramas, Sports Movies	Rocky Balboa takes on the Cold War, coming out...

- The above picture shows the recommendation for the movie “6 underground” from the model K-means Clustering.

- The recommendations from the K-means Clustering are very close to the movie titles we produced.

- So, K-means Clustering is the model we would like to choose as final model for further predictions.

Conclusion:

- The Experiment I chose is to Cluster the movie recommendations from Netflix Movie and TV shows dataset.
- The dataset has 12 features on offer .
- We applied 2 Machine Learning Algorithms namely :
  1. Natural Language Processing(NLP)
  2. K-means Clustering

- According to the recommendations seen from both the models, the results from the K-means model were very close to the films in terms of the description of the movie.