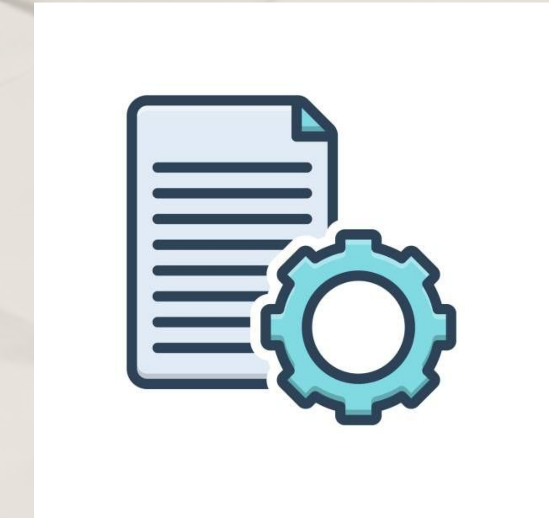
Netflix Analytics

Group – 8

Presented by:
Haritha Jampani

Table of contents

- ➔ Problem Statement
- ➔ Data Description
- ➔ Data Preparation & Cleaning
- ➔ EDA (Exploratory Data Analysis)
- ➔ Textual Data Preprocessing
- ➔ Dimensionality Reduction
- ➔ Model Implementation
- ➔ Recommender System



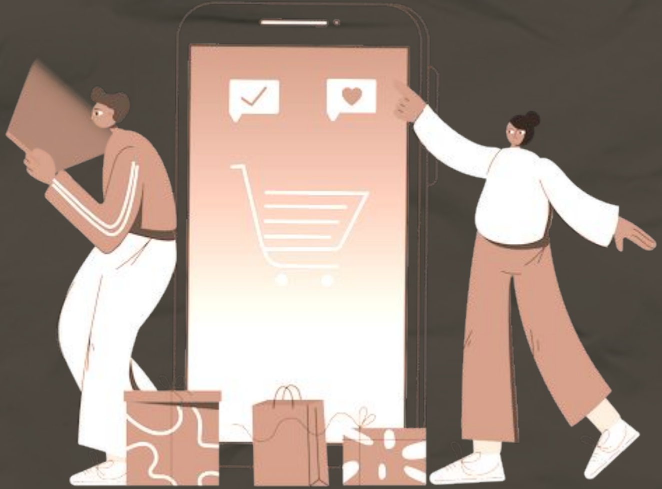
Problem Statement

- Significant increase in TV series and a decline in movies
- Leading global streaming service which grown from $2.7 billion to $269.54 billion within 14 years
- Understanding content availability and learning audience preferences became crucial to maintain and expand its market dominance

Problem Statement Cont.

In this project we will answer:

- Exploratory Data Analysis
- Understanding what content is available in different countries
- Has Netflix been focusing on TV Series more than Movies?
- Clustering similar content by matching text-based features



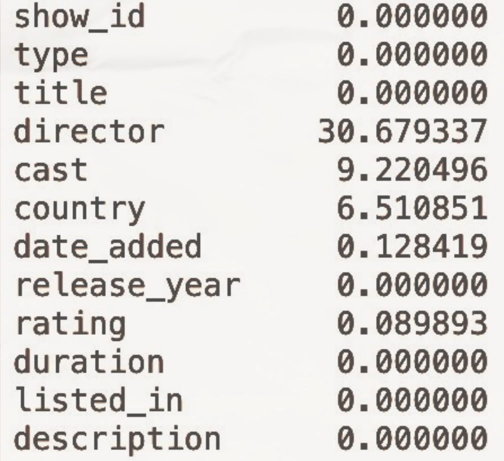
Data Description

The dataset is sourced from Flixable, third party Netflix search engine in 2019

- Show_id: Unique values
- Type: Tv Show/Movie
- Title: Name of the content
- Director: Name of director(s)
- Cast: Name of cast member(s)
- Country: Country the content was produced in
- Date_added: Date added to Netflix
- Release_year: Year the movie was released
- Rating: Abbreviations of ratings
- Duration: Length of the show/movie
- Listed_in: Genre(s) of the movie/show
- Description: Word description of the movie/show

Data Preparation & Cleaning

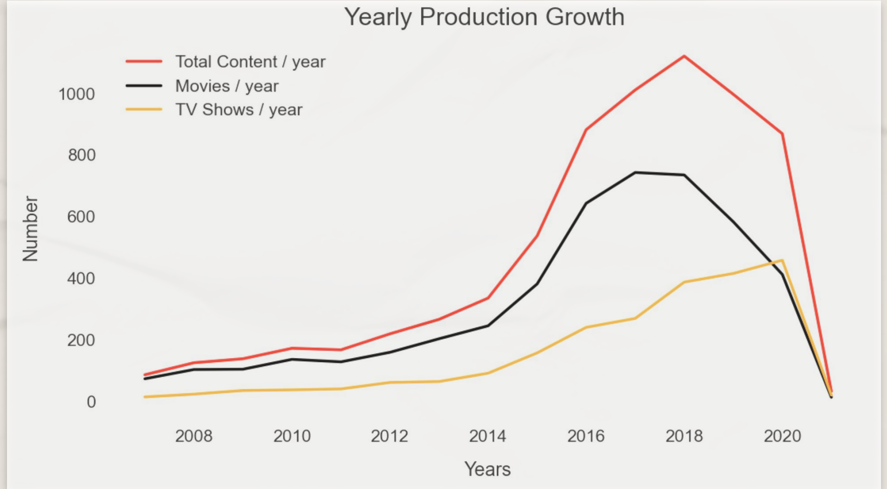
- Except **release_year**, all the other variables are categorical data type
- Filled missing values of **director**, **cast**, **country** as **Unknown**
- Dropped rows of **date_added** missing values
- Dropped rows of **ratings** missing values

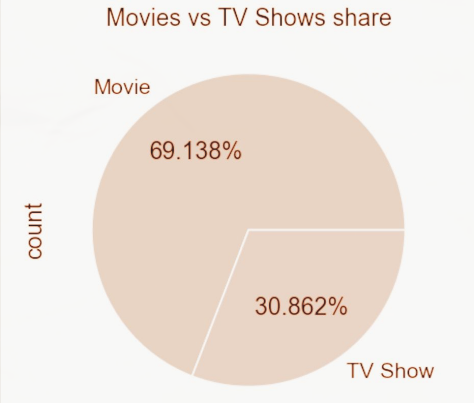


```
show_id          0.000000
type             0.000000
title            0.000000
director        30.679337
cast             9.220496
country          6.510851
date_added       0.128419
release_year     0.000000
rating           0.089893
duration         0.000000
listed_in        0.000000
description      0.000000
```

EDA

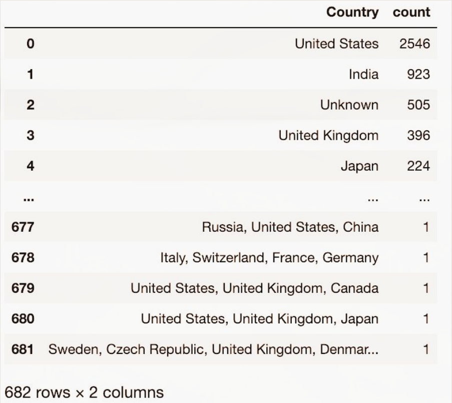
- This time series graph shows how many movie and tv shows are produced over the years and its total
- We can see that more movies are produced compared to TV shows
- Movies account for 69% of the total content available on Netflix and TV Shows account for roughly 31%





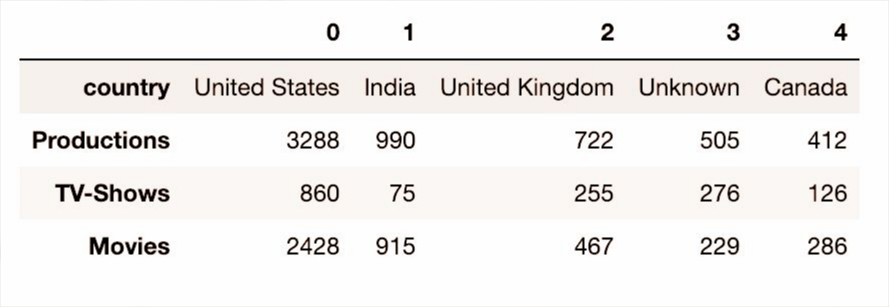
EDA Cont.

- We have **682** values for countries and the combinations of countries which produce content for Netflix in the dataset
- Among them we can see that United States, India, and United Kingdom produced a lot more TV Shows and Movies compared to other countries



	Country	count
0	United States	2546
1	India	923
2	Unknown	505
3	United Kingdom	396
4	Japan	224
...	...	...
677	Russia, United States, China	1
678	Italy, Switzerland, France, Germany	1
679	United States, United Kingdom, Canada	1
680	United States, United Kingdom, Japan	1
681	Sweden, Czech Republic, United Kingdom, Denmar...	1

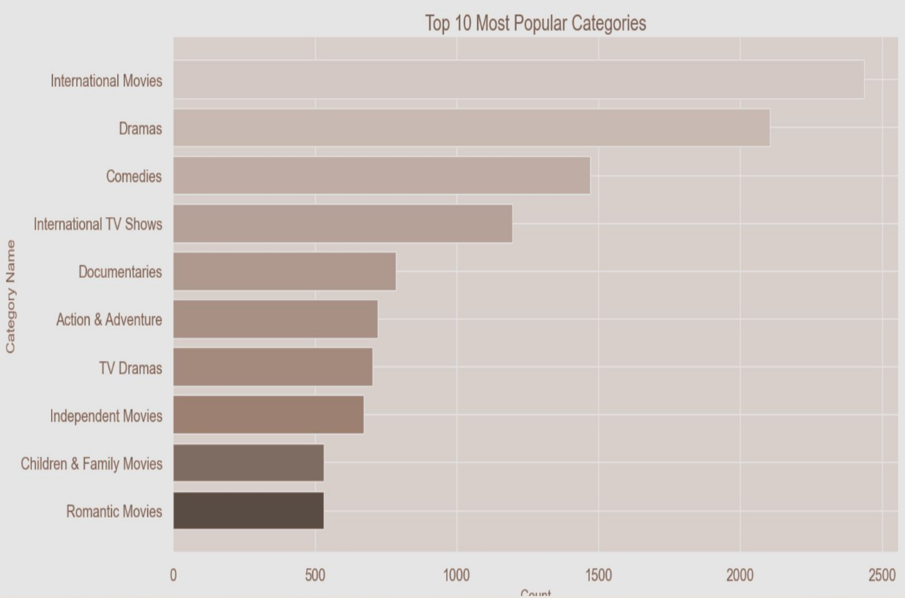
682 rows × 2 columns



	0	**1**	**2**	**3**	**4**
country	United States	India	United Kingdom	Unknown	Canada
Productions	3288	990	722	505	412
TV-Shows	860	75	255	276	126
Movies	2428	915	467	229	286

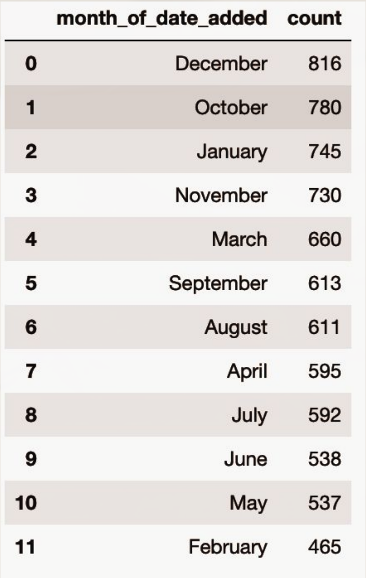
EDA Cont.

- There were 42 unique values for genres after splitting the values in ‘listed_in’ column
- Out of 7770 TV Shows and Movies, we found out that 2437 were listed under International Movies and the second highest genre is Drama with 2105 movies and tv shows combined



EDA Cont.

- We extracted only month from the **date_added** column and calculated the count of content released in each month
- Highest amount of Movies and TV Shows were added in the month of December and the reason behind this is because of the holidays
- Second highest is the month of October and followed by January



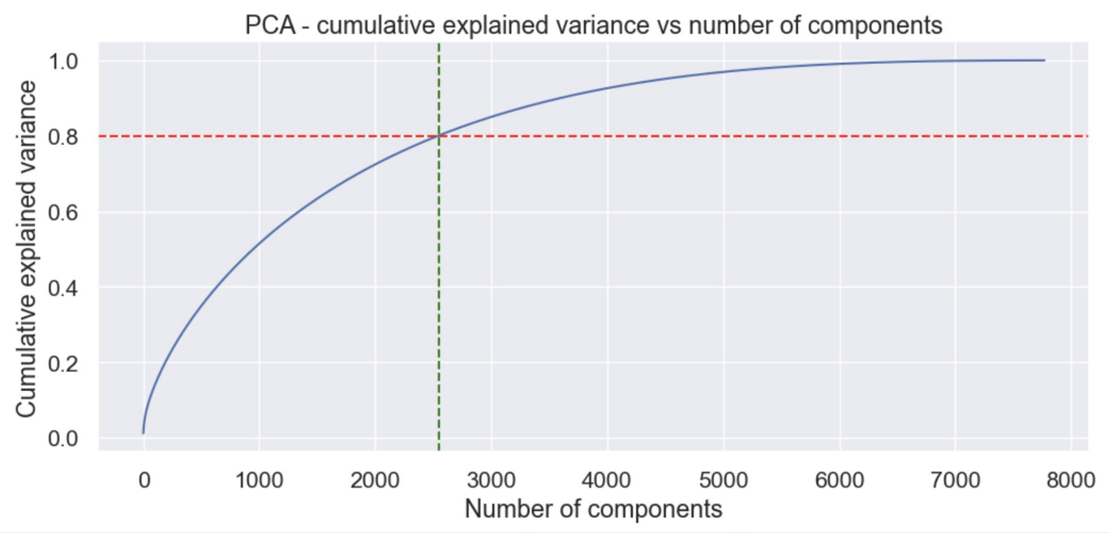
	month_of_date_added	count
0	December	816
1	October	780
2	January	745
3	November	730
4	March	660
5	September	613
6	August	611
7	April	595
8	July	592
9	June	538
10	May	537
11	February	465

Textual Data Preprocessing

1. Created a new column which combine columns; description, rating, country, listed_in, cast
2. Text Removal – removed punctuations, white spaces and stopwords etc
3. Tokenization – converted sequence of text into smaller parts(tokens)
4. Stemming – used Snowball stemmer to reduce words to their root form
5. POS Tagging – helps in effective analysis with grammatical tag
6. Vectorization – used TF-IDF to convert into vectors

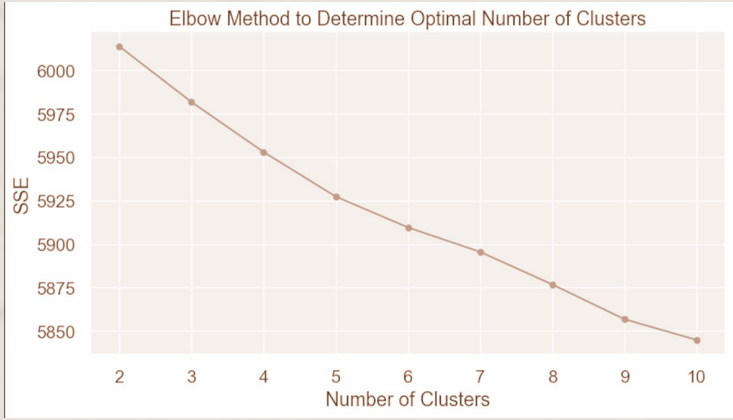
Dimensionality Reduction

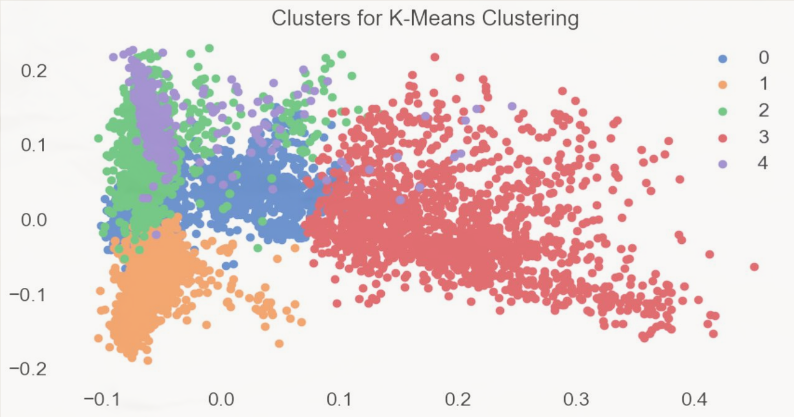
- Used Principal Component Analysis (PCA) to reduce the dimensionality of the data
- Captured 80% of the variance by reducing the components to **2550**



Model 1: K_Means

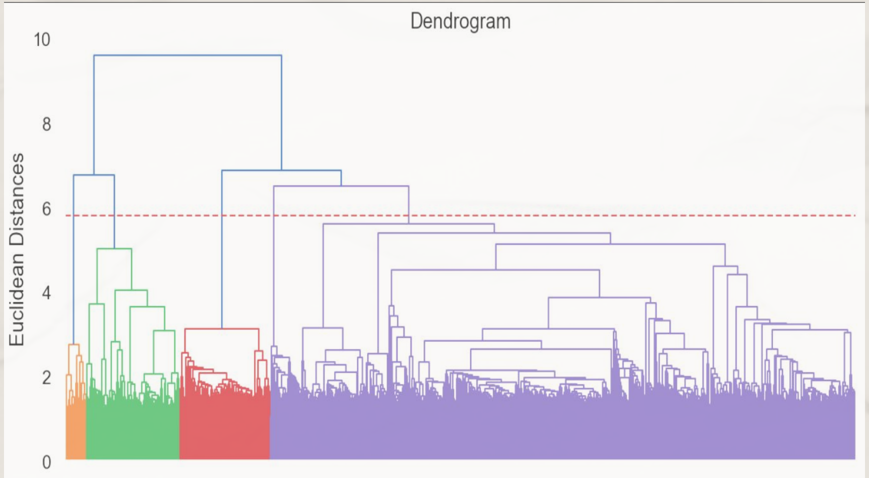
- Performed K-Means Clustering using the vectors found from using PCA
- We used KElbowVisualizer to find the optimal number of clusters

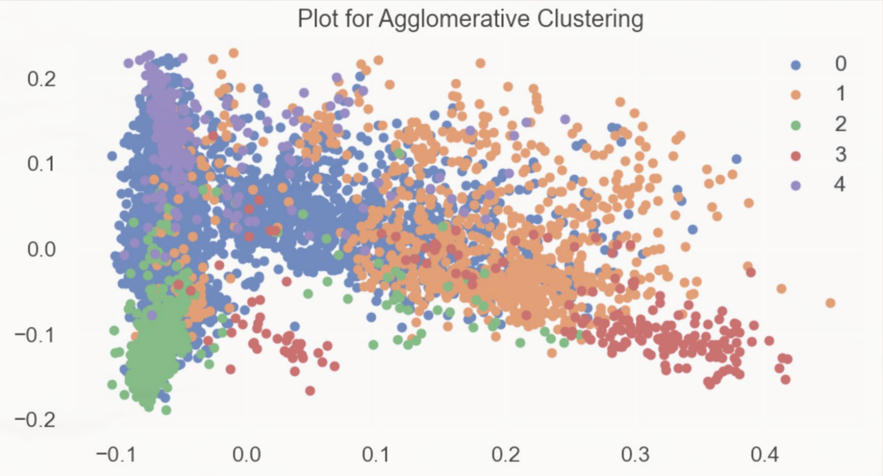




Model 2: Agglomerative Clustering

- Used Dendrogram to decide on the optimal number of clusters using Euclidean distance

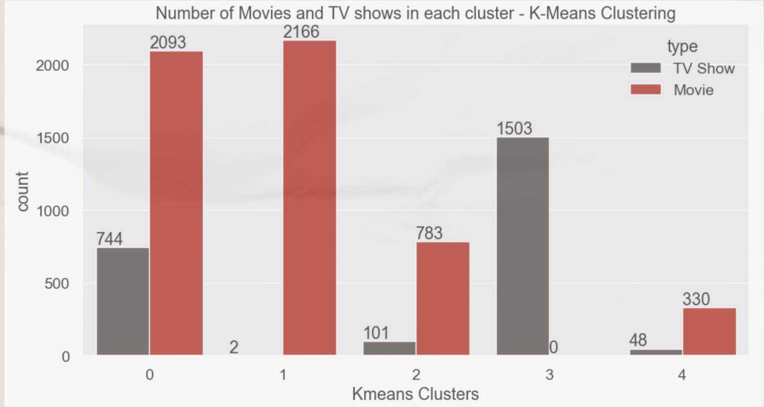


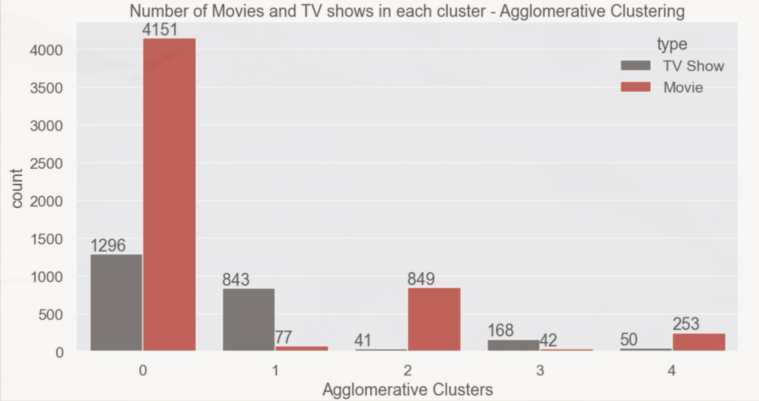


Final Prediction Model

- When compared two models we choose **K-Means** as the **suitable model** for our data.

- Clusters are **well divided** in case of **K-Means** when compared to **Agglomerative** which helps to **what kind of data** is present in **which cluster**

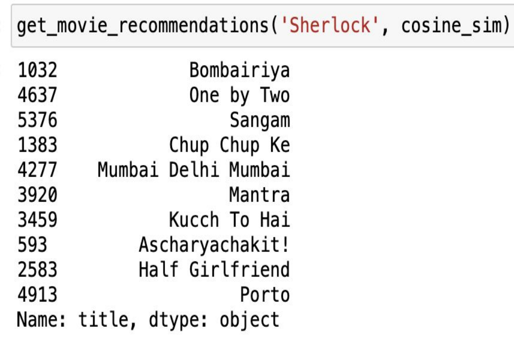




Content Based Recommendation

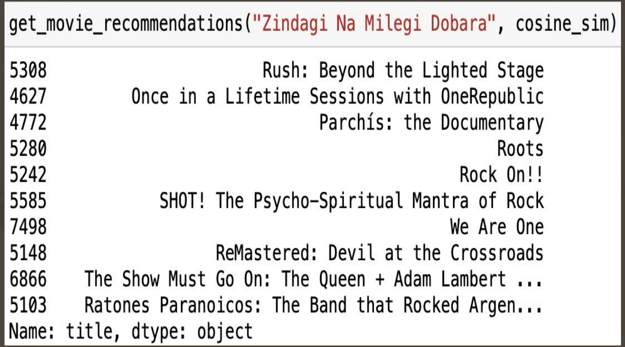
Top 10 Recommended Movies/TV Shows

- Used Cosine Similarity score to build a Content based Recommendation System



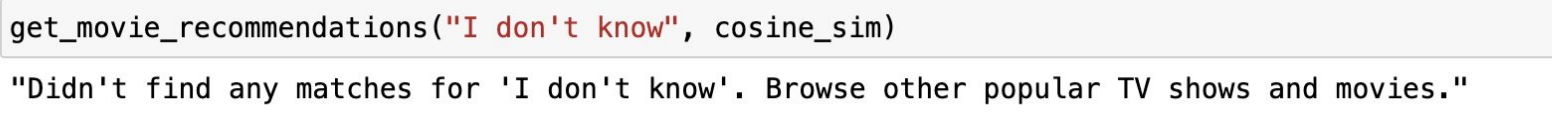
```
get_movie_recommendations('Sherlock', cosine_sim)

1032             Bombairiya
4637             One by Two
5376                 Sangam
1383           Chup Chup Ke
4277    Mumbai Delhi Mumbai
3920                 Mantra
3459           Kucch To Hai
593         Ascharyachakit!
2583        Half Girlfriend
4913                  Porto
Name: title, dtype: object
```



```
get_movie_recommendations("Zindagi Na Milegi Dobara", cosine_sim)

5308                       Rush: Beyond the Lighted Stage
4627         Once in a Lifetime Sessions with OneRepublic
4772                             Parchís: the Documentary
5280                                                Roots
5242                                            Rock On!!
5585            SHOT! The Psycho-Spiritual Mantra of Rock
7498                                           We Are One
5148                  ReMastered: Devil at the Crossroads
6866    The Show Must Go On: The Queen + Adam Lambert ...
5103    Ratones Paranoicos: The Band that Rocked Argen...
Name: title, dtype: object
```



```
get_movie_recommendations("I don't know", cosine_sim)

"Didn't find any matches for 'I don't know'. Browse other popular TV shows and movies."
```

Thank you !!!