



Netflix Movie and TV Shows Clustering

Technical Document

Mohammad Jibran

Siddhi Thakur

| Data Science Trainees |

ALMABETTER

Abstract

This project is all about recommending a variety of movies and TV shows to viewers, which are managed by the Netflix company. The platform streams the media online. Subscribers can choose between monthly plans and yearly plans based on how much they want to watch Netflix simultaneously. There are three levels of monthly plans on which subscribers can base their subscriptions— Standard, Premium and Ultra HD4 plan memberships; this business is profitable. Customers, however, are free to stop their memberships at any moment. As a result, the business needs to maintain users' interest in the platform. Systems that make useful suggestions to users are crucial in this situation, which is where they start to play a significant role.

In order to comprehend the data in this project, we apply various tools and plot lots of graphs to make proper analysis.also use clustering techniques.

1. 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 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.

2. In this project, you are required to do

1. Exploratory Data Analysis
2. Understanding what type content is available in different countries
3. Is Netflix increasingly focused on TV rather than movies in recent years?
4. Clustering similar content by matching text-based features.

3. Introduction:

Netflix is an online streaming service that allows subscribers to watch television series and movies on their personal computers, mobile devices, and gaming consoles.This is mainly because watching movies at home is cheaper and more convenient than going to the theater. In addition, some people

prefer to watch their favorite shows online as they can easily access the internet at the same time as their show. Netflix was founded in the United States but maintains a global presence. The company is currently valued at \$152 billion. Netflix has over 130 million subscribers worldwide.

In 2009, the chief executive of Qwikster dropped the company's name and launched Netflix with a global reach. Due to this decision, many people thought that the service was originally called "Netflix with a global reach" or "Nwntglobal." However, the name "Netflix" refers to the thin blue line image that appears on screen during movie playback. The name also refers to the fact that streaming is carried out by boats or ships, hence the name "Netflix." In addition, some people believe that Netflix is an acronym for "nulled entertainment network projection system."

In addition to offering streaming for television episodes and movies, Netflix provides movie-download options as well as book rentals via its app store. You can download or stream content from over 130 million titles on Netflix at any given time.

As a globally recognized video-streaming service, Netflix has earned its place among entertainment platforms worldwide. Furthermore, as an international publisher of media content, it attracts subscribers from all 50 states as well as numerous international locations such as France and Japan 6. With over 130 million users globally 7,Netflix is an excellent entertainment choice for individuals looking for uninterrupted viewing of films and TV shows via various platforms 8.

4. Understanding the data

In this dataset there are a total of 7787 rows and 12 columns. The demand for movies and TV shows rise during the winter season.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7787 entries, 0 to 7786
Data columns (total 12 columns):
#   Column          Non-Null Count  Dtype
---  ---
0   show_id         7787 non-null   object
1   type            7787 non-null   object
2   title           7787 non-null   object
3   director        5398 non-null   object
4   cast            7069 non-null   object
5   country         7280 non-null   object
6   date_added      7777 non-null   object
7   release_year    7787 non-null   int64
8   rating          7780 non-null   object
9   duration        7787 non-null   object
10  listed_in       7787 non-null   object
11  description     7787 non-null   object
dtypes: int64(1), object(11)
memory usage: 730.2+ KB
```

In this dataset, we have different types of columns which help to develop efficient models and make accurate predictions, but before it begins we have to analyze our dataset and data types. As we can see date_added columns datatype is object but it is wrong so we change it and extract day, month and year from date_added columns.

```
# Create new features to store date, day, month and year separately.
netflix_cp['date_added'] = pd.to_datetime(netflix_cp['date_added'])
netflix_cp['days'] = netflix_cp['date_added'].dt.day
netflix_cp['months'] = netflix_cp['date_added'].dt.month
netflix_cp['years'] = netflix_cp['date_added'].dt.year
```

```
netflix_cp.drop('date_added',axis = 1,inplace = True)
```

After extracting necessary data we drop the 'date' column from the dataset.

Now we have 14 columns and 7770 rows in datasets.

These are updated columns ----

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 7770 entries, 0 to 7786
Data columns (total 14 columns):
#   Column          Non-Null Count  Dtype
---  -
0   show_id         7770 non-null   object
1   type            7770 non-null   object
2   title           7770 non-null   object
3   director        5394 non-null   object
4   cast            7052 non-null   object
5   country         7265 non-null   object
6   release_year    7770 non-null   int64
7   rating          7770 non-null   object
8   duration        7770 non-null   object
9   listed_in       7770 non-null   object
10  description      7770 non-null   object
11  days            7770 non-null   int64
12  months          7770 non-null   int64
13  years           7770 non-null   int64
dtypes: int64(4), object(10)
memory usage: 910.5+ KB
```

Column Name	Column Description
show_id	A unique identifier of records in the dataset.
type	Verify whether it's Movie or TV show.
title	Title of Movie or TV show.
director	Director of the TV show or movie
cast	The cast of Movie or TV show.
country	The list of the country in which a Movie or TV show is released or watched.
date_added	The date on which the content was on boarded on the Netflix platform.
release_year	Year of the release of the Movie and TV-show
rating	The rating informs about the suitability of the content for a specific group.
duration	Duration is specified in terms of minutes for Movie and and in terms of number of seasons in the case of TV- show.
listed_in	This column specifies the genre of the content.
description	A short summary about the storyline of the content.
days	The day on which the content was on boarded on the Netflix platform.
months	The month on which the content was on boarded on the Netflix platform.
years	The year on which the content was on boarded on the Netflix platform.

5. Data Wrangling

Data wrangling is a term often used to describe data analysis. It helps to transform data from one format into another. The aim is to make data more accessible. These include things like data collection, exploratory analysis, data cleansing, creating data structures and storage.

- **Importing important libraries**

Our major goal in this step was to import all of the necessary libraries to aid us in exploring the issue statement and doing EDA to draw conclusions based on the data collection.

Libraries we used:

1. NumPy

NumPy is the python library used for programming languages, adding support for large, multidimensional arrays and matrices with large collections.

2. Pandas

It is a software library written for python programming, flexible, and expressive data structures designed to make working with relational or labeled data both easy and intuitive. Pandas allows us to access many of Matplotlibs and NumPy's methods with fewer codes.

3. Matplotlib

It is a pyplot collection of functions that make matplotlib work like MATLAB. e.g., creates a figure, lines a plotting area.

4. Seaborn

It is an open-source python library built on top of matplotlib. It is used for data visualization and EDA.seaborn works easily with Data frames and pandas' libraries. The graphs can also be customized.

5. Scikit-learn (Sklearn)

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, and clustering and dimensionality reduction via a consistent interface in Python.

6. WordCloud

Word Cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance.

7. CountVectorizer

It is used to convert a collection of text documents to a vector of term/token counts. It also enables the pre-processing of text data prior to generating the vector representation.

8. TfidfVectorizer

It uses an in-memory vocabulary to map the most frequent words to feature indices and hence compute a word occurrence frequency matrix.

9. cosine_similarity

It is computed as the sum of element-wise products of A and B.

10. KMeans

K Means segregates the unlabeled data into various groups, called clusters, based on having similar features, common patterns.

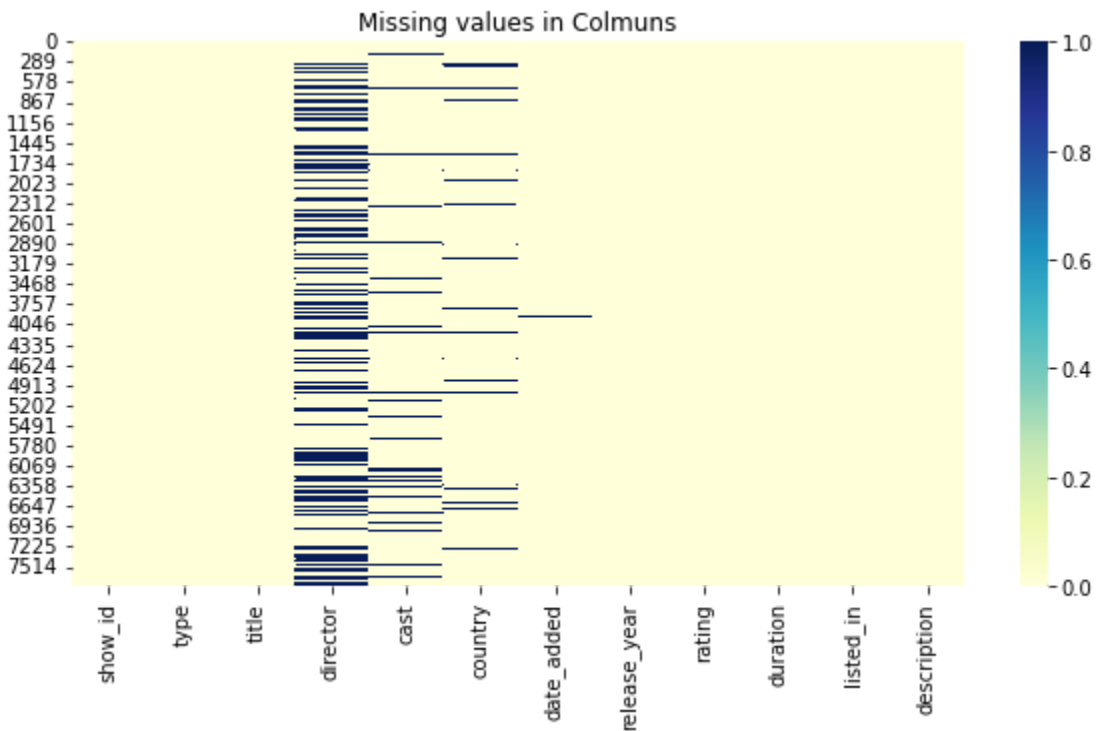
11. silhouette_score

It is used to evaluate the quality of clusters created using clustering algorithms such as K-Means in terms of how well samples are clustered with other samples that are similar to each other.

12. DBSCAN

Density-Based Spatial Clustering of Applications with Noise. Finds core samples of high density and expands clusters from them.

Pre-processing of dataset



In our dataset there are some columns, which have very large missing values. The columns with names: director, cast, country have higher null values and columns date_added and rating have very less null value and some of them have no missing values. We plot a heat-map for better understanding.

Below is a table where we get null values in percentages.

	No Of Total Values	No of NaN values	%age of NaN values
director	7787	2389	30.68
cast	7787	718	9.22
country	7787	507	6.51
days	7787	10	0.13
months	7787	10	0.13
years	7787	10	0.13
rating	7787	7	0.09
show_id	7787	0	0.00
type	7787	0	0.00
title	7787	0	0.00
release_year	7787	0	0.00
duration	7787	0	0.00
listed_in	7787	0	0.00
description	7787	0	0.00

Operation on NULL values

As we can see in the above table the column date_added and rating have very less NULL values,so we decide to clear those NULL values.

Before Removing NULL values.

```
# Number of null values in rating.
netflix_cp.rating.isnull().sum()
```

7

```
netflix_cp.date_added.isnull().sum()
```

10

After removing NULL values.

```
# Remove null values in rating.
netflix_cp.dropna(subset=['rating'], inplace=True)
netflix_cp.rating.isnull().sum()
```

0

```
netflix_cp.dropna(subset=['date_added'], inplace=True)
netflix_cp.date_added.isnull().sum()
```

0

Duplicate values

Our dataset does not contain any duplicate values.

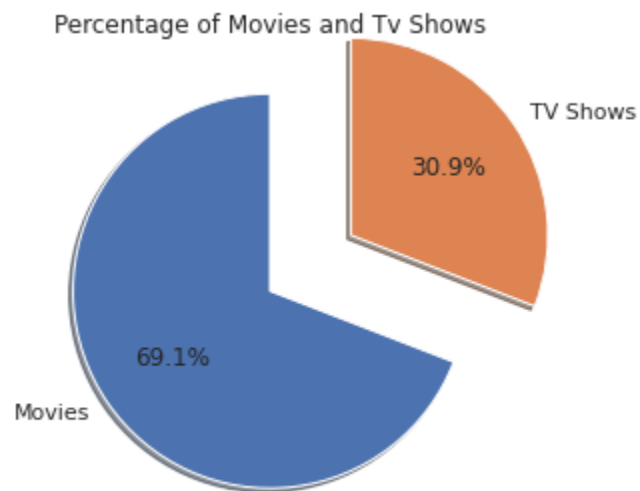
```
# Check for duplicated entries.  
print(f"Total number of duplicated entries : {netflix_cp.duplicated().sum()}")
```

Total number of duplicated entries : 0.

Exploratory Data Analysis

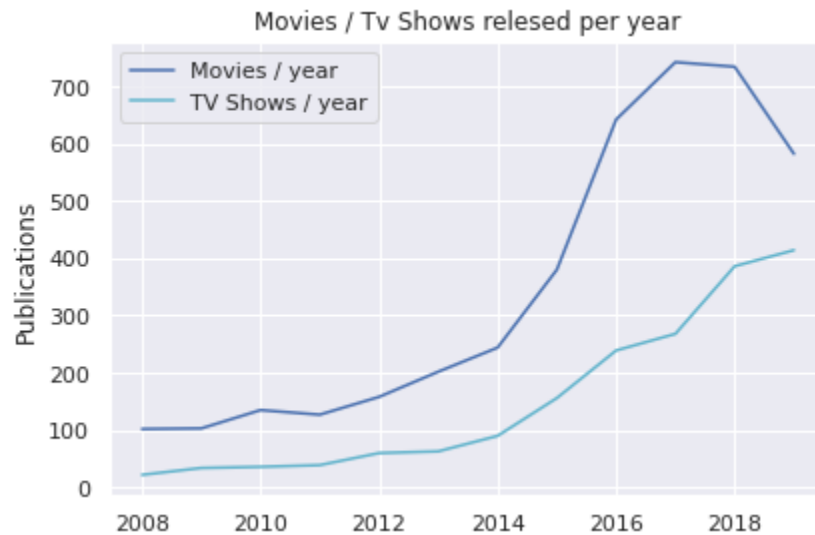
After treating the null values, we start with the EDA. We performed EDA and tried to understand the data by asking some questions.

What type of content is available on Netflix.



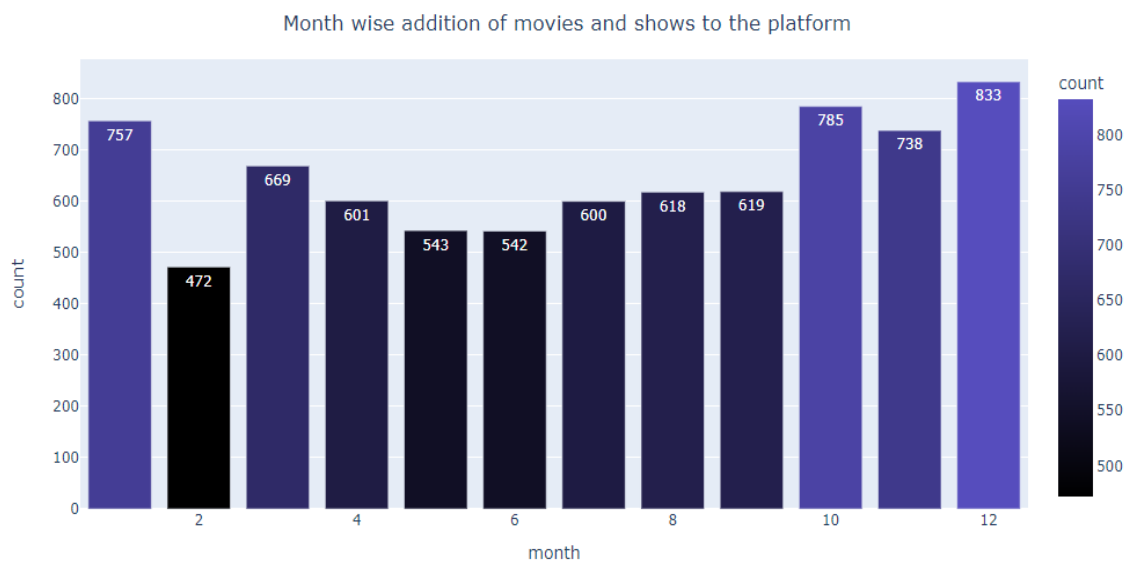
This graph shows the Type of content available on Netflix dataset, we have 69.1% of Movie content and 30.9% of TV-shows content. So we can conclude that Netflix has more movie content than TV-shows.

How many Movie/TV shows are released per year.



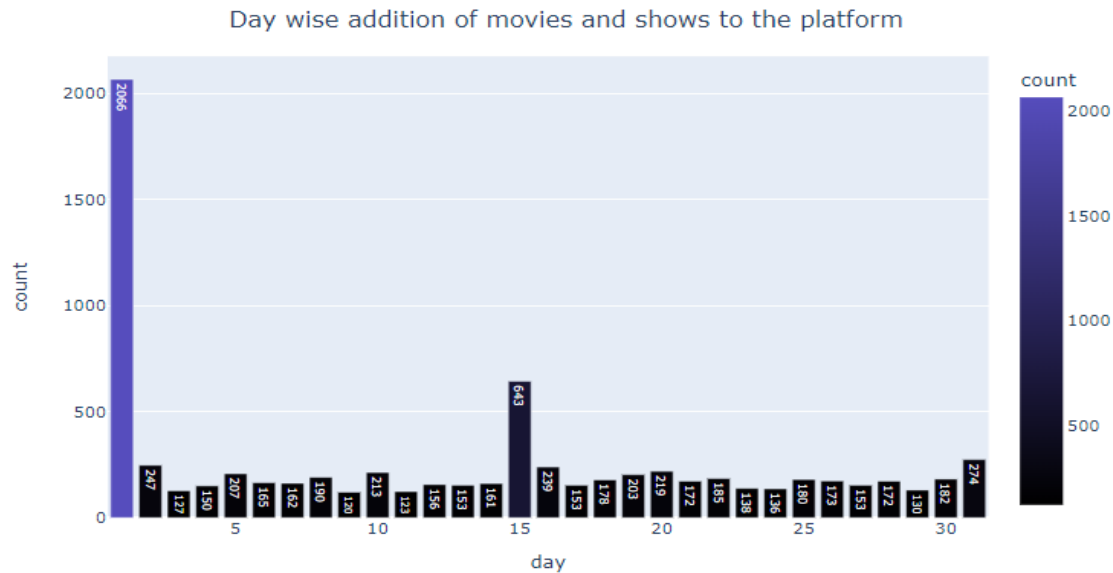
From 2014 to 2019 the release of movies is greater than TV-shows.

In which months do most movies and TV shows are added.



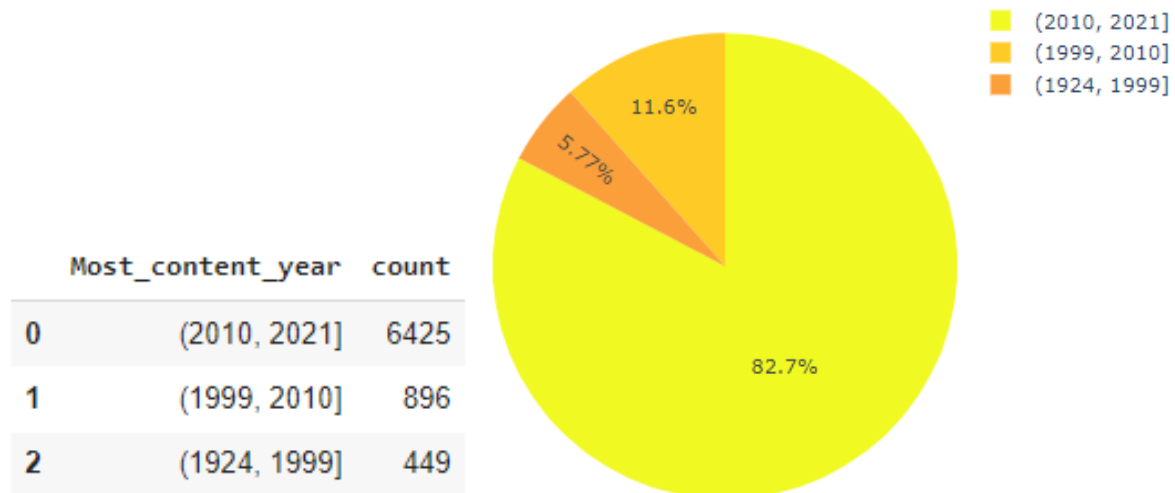
This bar graph shows most of the content is added either by year ending or beginning. As we see the highlighted bars are October, November, December, and January are months in which many shows and movies get uploaded to the platform.

In which Days movies and TV shows are performing Outstanding.



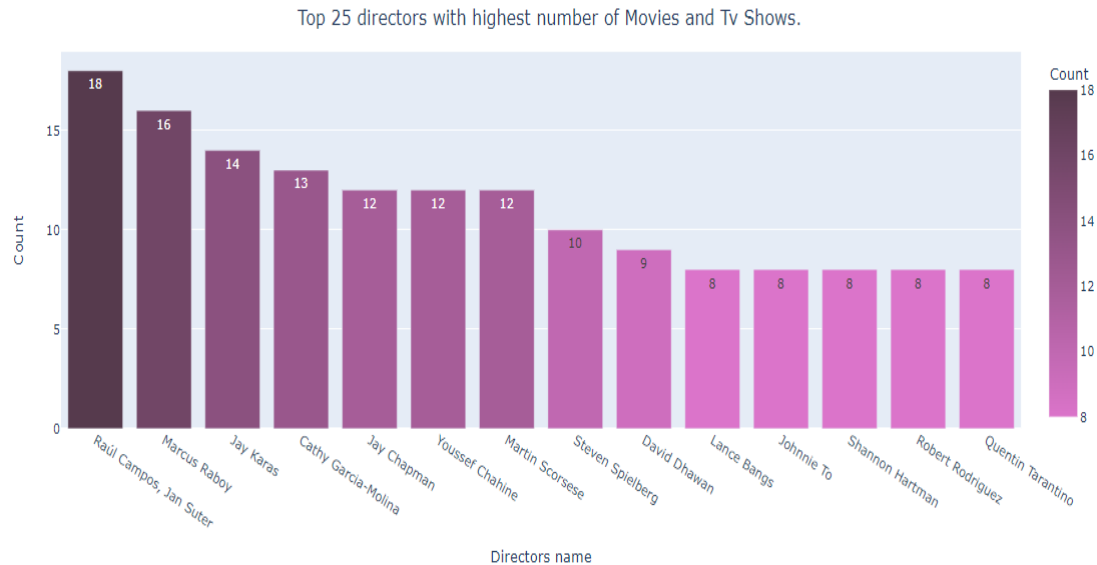
This bar graph shows most of the content is uploaded at the beginning, middle, or the end of a month. Which makes 1st, 15th or 31st of a month more outstanding in getting new tv shows and movies.

In Which year most of the content was added.



We can see that the majority of the content available was between 2010 and 2021 that is 82.7%. Since 1924 to 2010, only 17.28% of the content available was released.

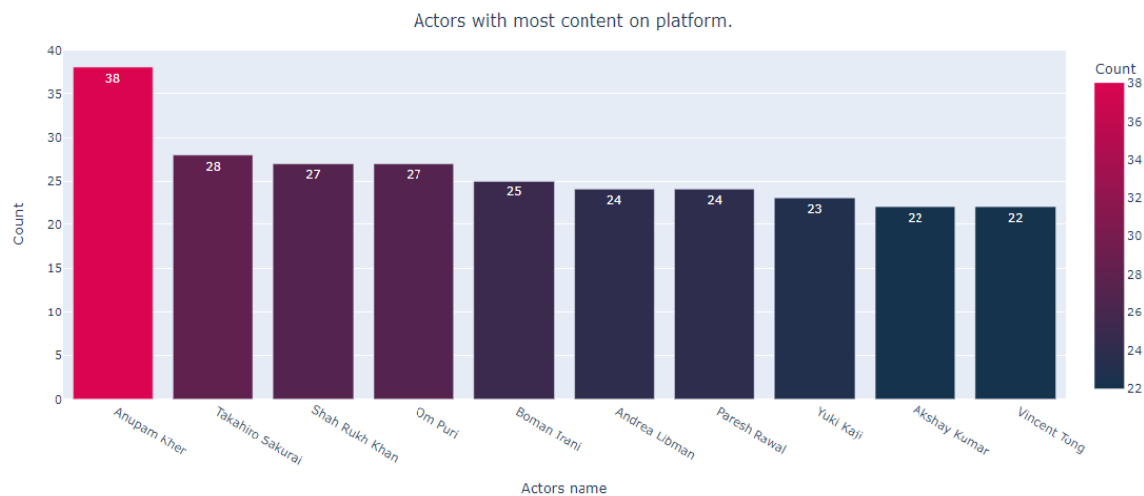
Which director has directed the most movies and TV shows?



This graph shares a report of Top 5 directors who direct the majority of movies and TV-shows.

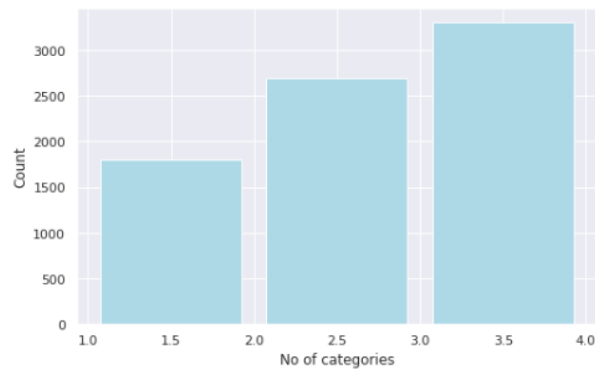
Raúl Campos, Jan Suter, Marcus Raboy, Jay Karas, Cathy Garcia-Molina, Jay Chapman are the top 5 directors.

Which Actor/Actress have been cast in most of the movies and TV shows?



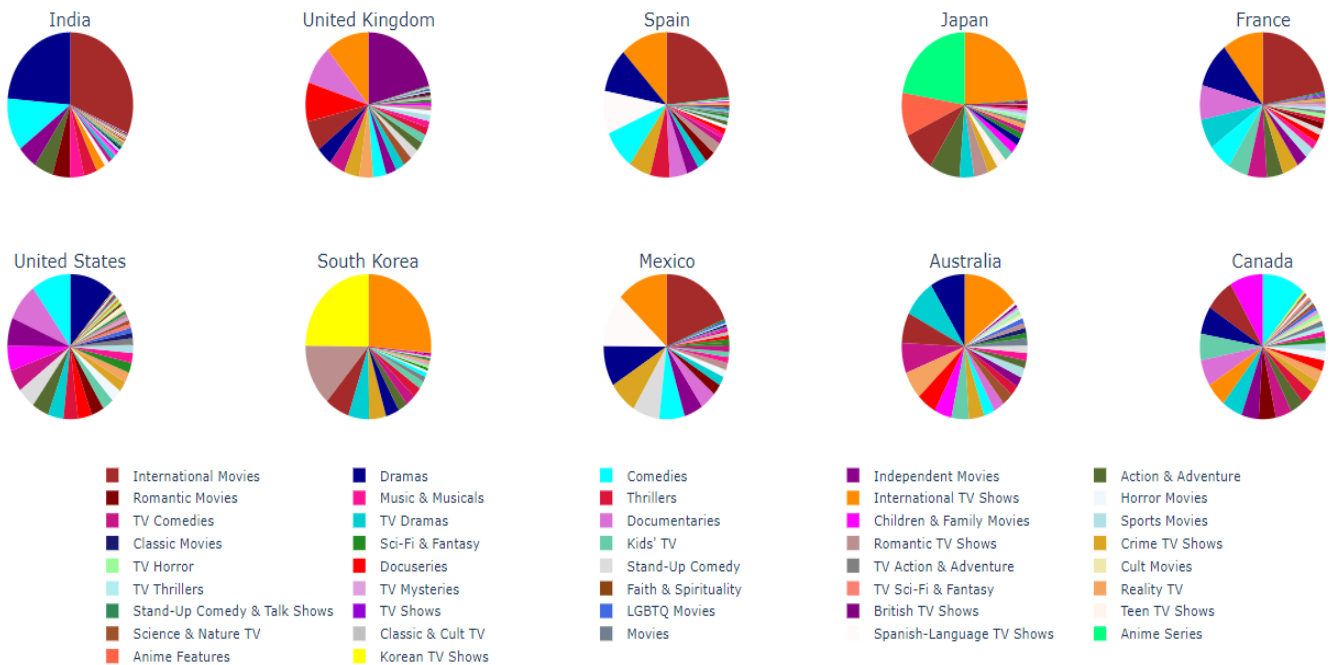
This bar graph presents the count of the Actor/Actress cast in most movies or shows. As we can see Indian Actor Anupam Kher is ranked first with 38 overall appearances in TV shows and films. Also Proud to see Six other Indian actors round up the top ten list.

How many no of categories are present there in each content.



In this bar graph we define categories of genre present in each content. We can see Majority of content belongs to third categories.

Which Genre is more popular in these countries?



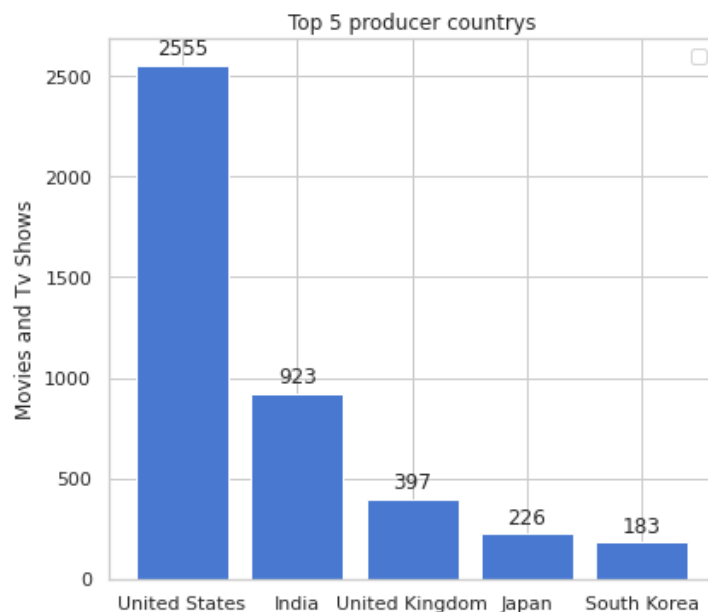
These pie charts define which genre is most popular in which countries and contents, as we can see regional delicacies are more prevalent in some nations, such as anime in Japan and Korean TV shows in South Korea. This makes sense because anime has long been popular in Japan, and the expansion of K-pop culture explains the rise in Korean TV Shows.

In the UK, British and foreign television programming predominates.

Drama, International Movies, and Comedies seem popular choices in most countries.

It's also observed that in the countries where the regional language is not English, International Tv Shows and Movies are more in demand.

Countries producing most no of contents.



Here we did analysis on the content of each country. As we can see in the bar graph, the United States has the most content followed by India with South Korea being the last. It may be because the US and India have a variety of people with different backgrounds.

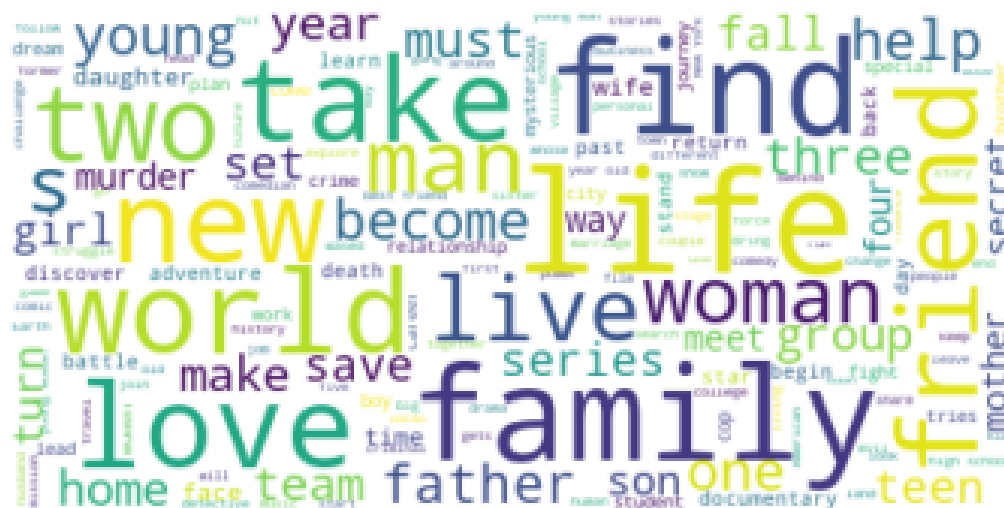
Total number of unique titles present in the title column.



This is called word cloud because Word Cloud creators are used to highlight popular words and phrases based on frequency and relevance.

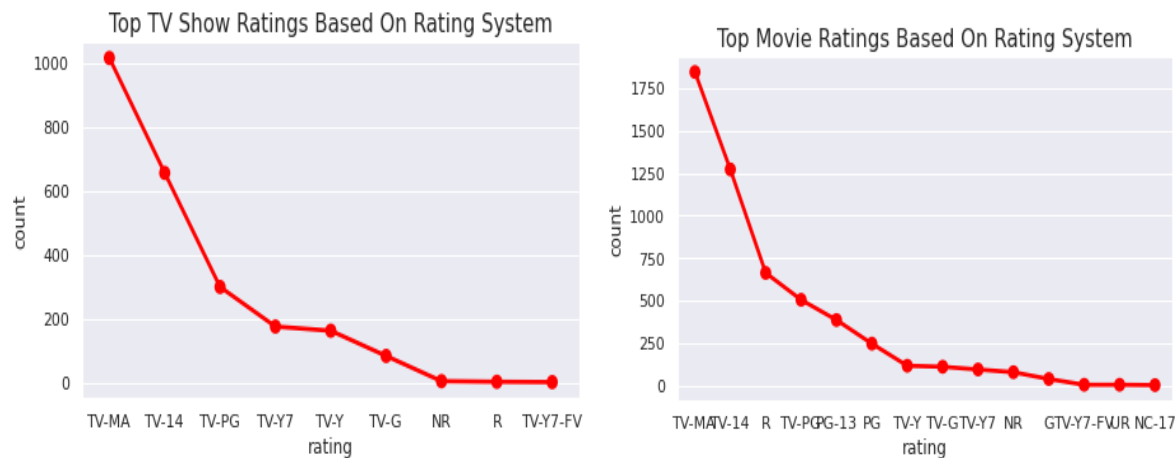
In the above picture we can see Most occurring words in the Title column of the tv shows and movies like `Love, Man, World, Day, Story, Christmas`.Etc.

Total number of unique Description present in the title column.



This is also Picturized by Word Cloud. In the above picture we can see Most occurring words in the description column of the tv shows and movies. Like `family, friend, new, take. Find, etc.`

Most popular TV-Shows\Movie Rating.



Above point plot describes the rating of TV-shows and Movies based on the Rating system. Most of the content got rating like

- TV-MA (For Mature Audiences)
- TV-14 (May be unsuitable for children under 14)
- TV-PG (Parental Guidance Suggested)
- NR (Not Rated)

Year v\s Types

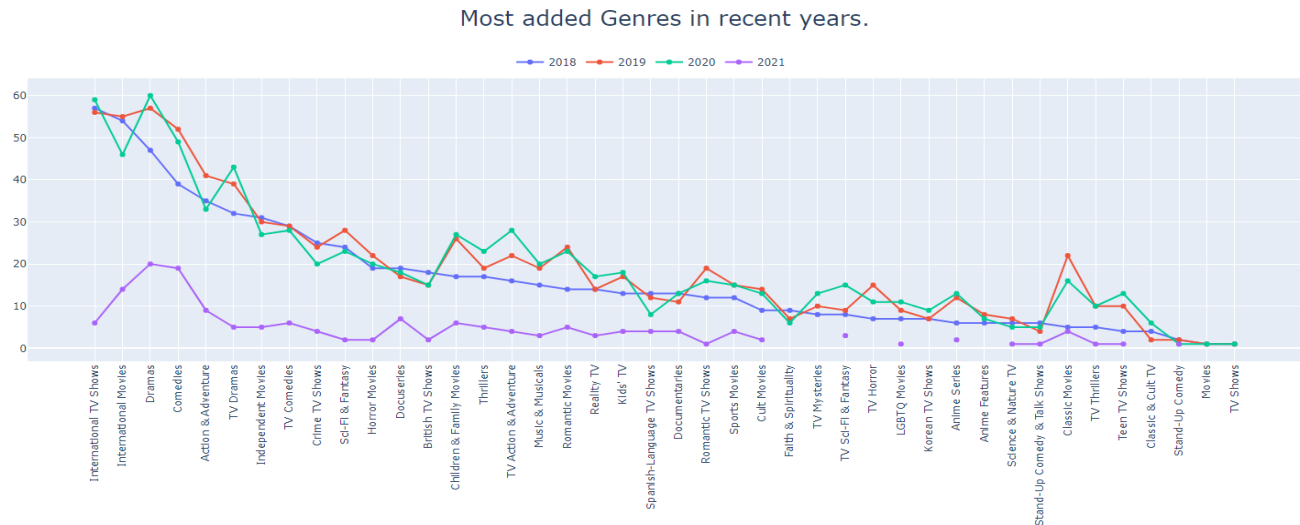
Netflix has increasingly focused on TV rather than movies in recent years.

We have performed hypothesis testing to get the insights if there is any relation between year and type.

- **Null Hypothesis:** year has no impact on the type of content that gets added to the platform.
- **Alternative Hypothesis:** year_added has an impact on the type of content that gets added to the platform.

years	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
type														
Movie	1	2	1	13	3	6	19	58	256	861	1255	1497	1312	88
TV Show	1	0	0	0	0	5	6	30	184	361	429	656	697	29

The p value is smaller than the significance level, so we will reject the null hypothesis and accept the alternative hypothesis.



1. From the EDA we did in the beginning, we saw that there are more Movies than Tv Shows on Netflix, which might be enough to assume that Netflix focuses more on Movies than Tv Shows. But the data proves this assumption wrong.
2. The above line plot shows that Netflix has been adding many International Tv Shows, Tv Dramas, Tv Comedy Shows and many more tv shows in the recent years compared to Movies.
3. From this observation, we can say that Netflix might be shifting slowly towards Tv Shows. Because of covid-19, there is a significant drop in the number of movies and television episodes produced after 2019.

Data Preprocessing

Removing Punctuation: Punctuation has no significance in clustering, therefore deleting it aids in removing noise or unhelpful portions of the data.

Removing stopwords: Stop-words, which don't just exist in English, are essentially a group of frequently used words in all languages. When we eliminate the words that are used a lot in a certain language, we may concentrate on the relevant terms.

Stemming: A word's stem or root is identified by stemming, which is the process of removing a portion of the word. stemming is the process of reducing words to their most fundamental form, or stem, which may or may not be a recognised word in the language.

We use two vectorization techniques and also applied all these data preprocessing on description and Listed in columns.

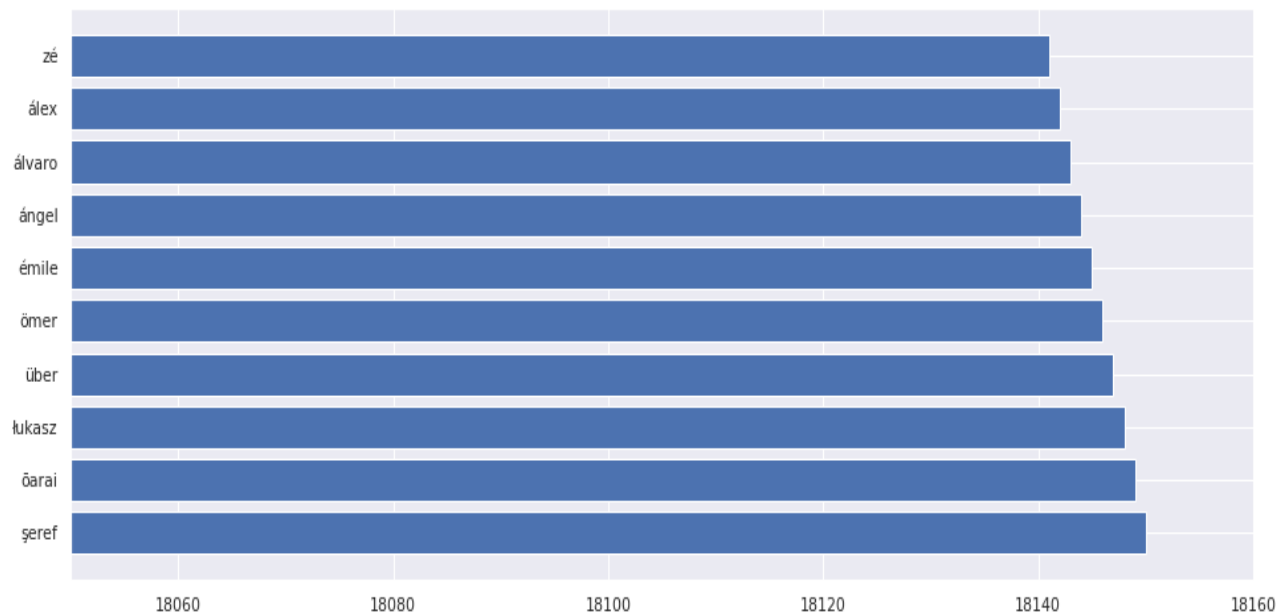
Tfidf vectorization: Transform a count matrix to a normalized tf or tf-idf representation.

Tf means term-frequency while tf-idf means term-frequency times inverse document-frequency. This is a common term weighting scheme in information retrieval that has also found good use in document classification. The goal of using tf-idf instead of the raw frequencies of occurrence of a token in a given document is to scale down the impact of tokens that occur very frequently in a given corpus and that are hence empirically less informative than features that occur in a small fraction of the training corpus.

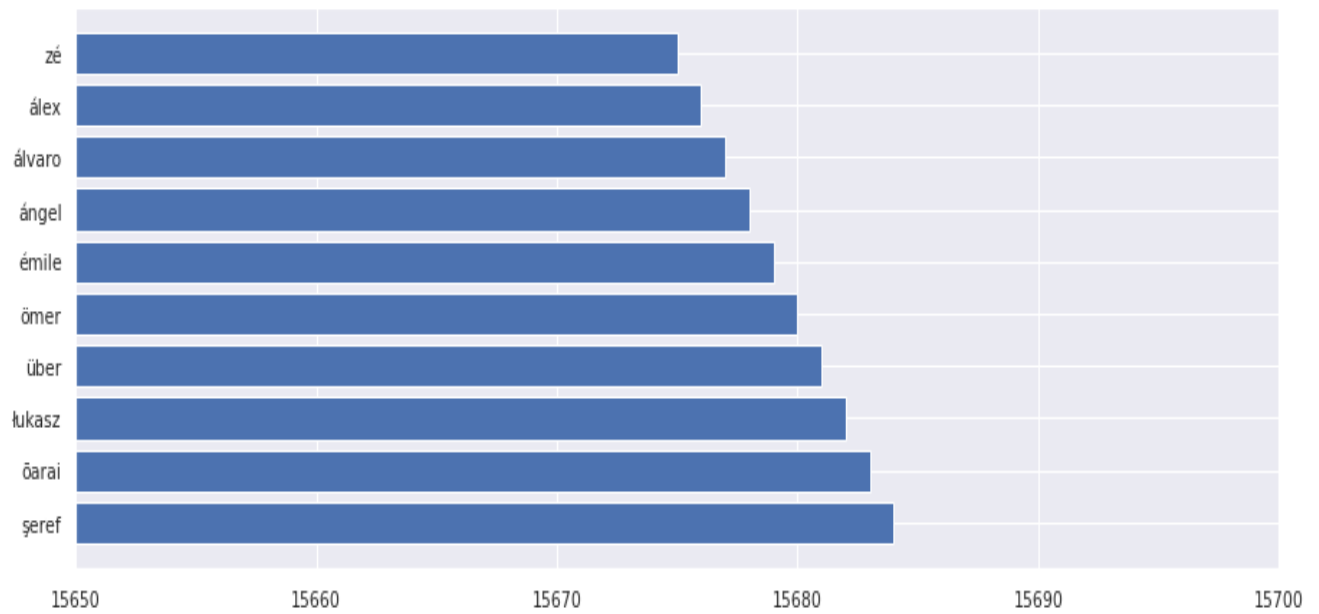
Count Vectorization: Convert a collection of text documents to a matrix of token counts. This implementation produces a sparse representation of the counts using `scipy.sparse.csr_matrix`. If you do not provide an a-priori dictionary and you do not use an analyzer that does some kind of feature selection then the number of features will be equal to the vocabulary size found by analyzing the data.

CountVectorizer Before stemming on description columns

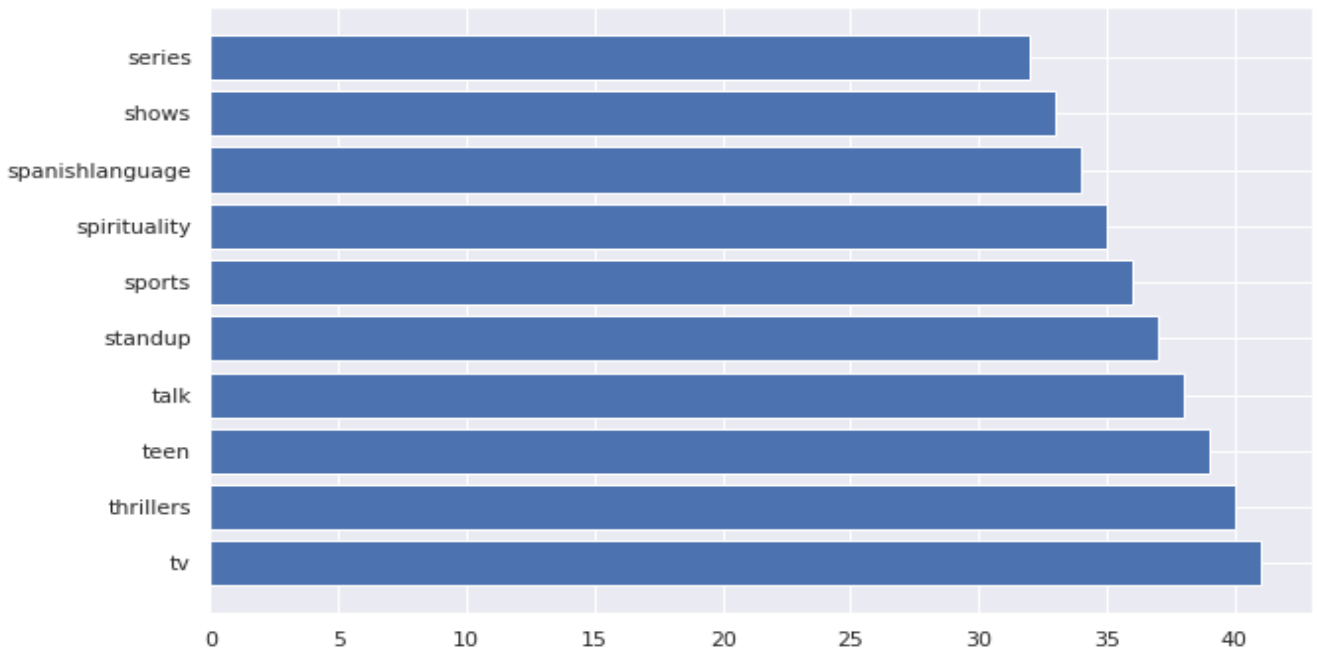
These are the top ten occurring words in description.



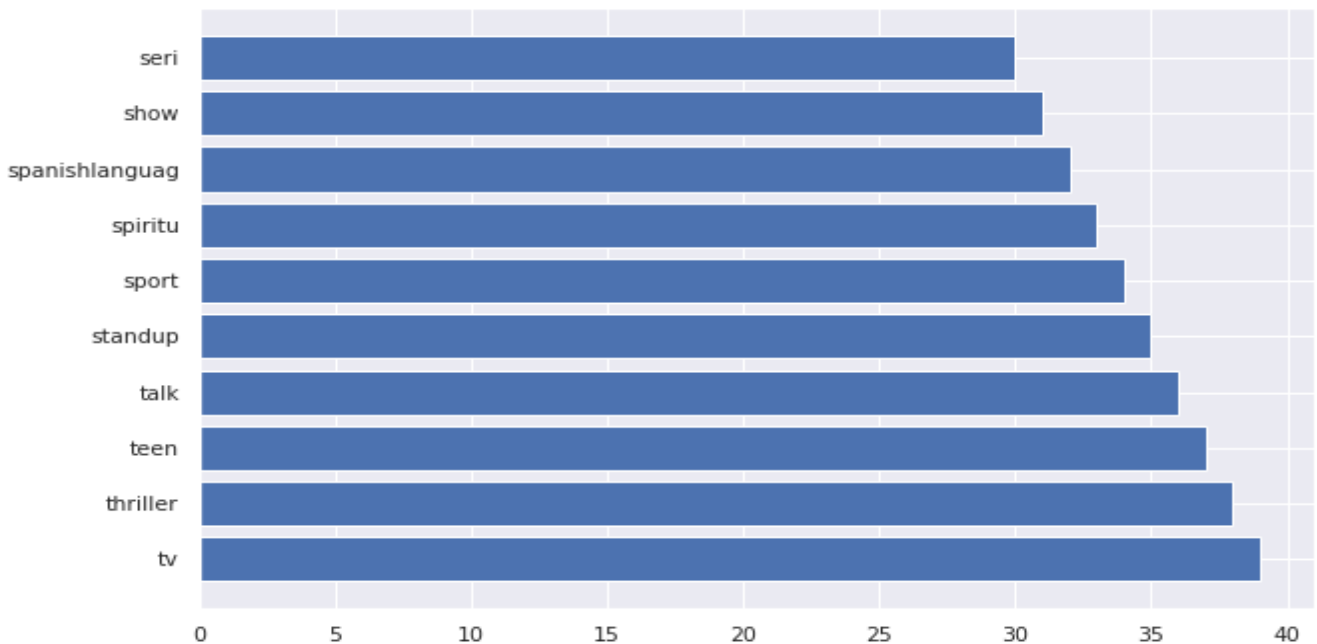
TfidfVectorizer After stemming on description.



CountVectorizer Before stemming on Listed_in columns.



TfidfVectorizer After stemming on Listed_in columns.



Clustering

Clustering or cluster analysis is a machine learning technique, which groups the unlabelled dataset. It can be defined as "A way of grouping the data points into different clusters, consisting of similar data points. The objects with the possible similarities remain in a group that has less or no similarities with another group."

It does it by finding some similar patterns in the unlabeled dataset such as shape, size, color, behavior, etc., and divides them as per the presence and absence of those similar patterns.

It is an unsupervised learning method, hence no supervision is provided to the algorithm, and it deals with the unlabeled dataset.

After applying this clustering technique, each cluster or group is provided with a cluster-ID. ML systems can use this id to simplify the processing of large and complex datasets.

Apply Different clustering Algorithms.

We need to check some conditions before applying silhouette score, there are two things to consider.

1. If we have the ground truth labels (class information) of the data points available with us then we can make use of extrinsic methods like homogeneity score, completeness score and so on.
2. If we do not have the ground truth labels of the data points, we will have to use the intrinsic methods like silhouette score which is based on the silhouette coefficient. We now study this evaluation metric in a bit more detail.

1. Silhouette score:

The silhouette Method is also a method to find the optimal number of clusters and interpretation and validation of consistency within clusters of data. The silhouette method computes silhouette coefficients of each point that measure how much a point is similar to its own cluster compared to other clusters. by providing a succinct graphical representation of how well each object has been classified.

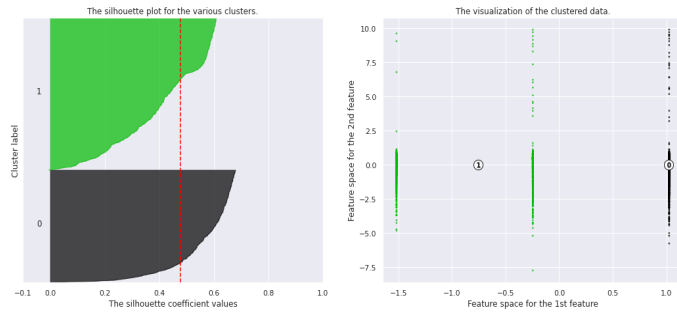
The silhouette value is a measure of how similar an object is to its own cluster (**cohesion**) compared to other clusters (**separation**). The value of the silhouette ranges between $[-1, 1]$, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. If most objects have a high value, then the clustering configuration is appropriate. If many points have a low or negative value, then the clustering configuration may have too many or too few clusters.

mean intra-cluster distance (a) - mean Intra-cluster distance is the distance among members of a cluster, rather than the distance between two different clusters.

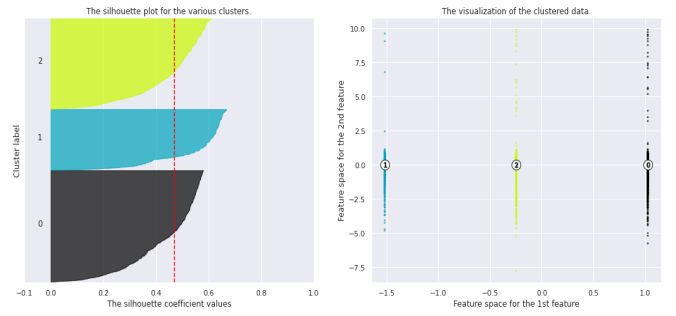
mean nearest-cluster distance (b) - The observation's average distance from every other data point in the subsequent closest cluster. This distance is also referred to as a.

$$s = \frac{b - a}{\max(a, b)}$$

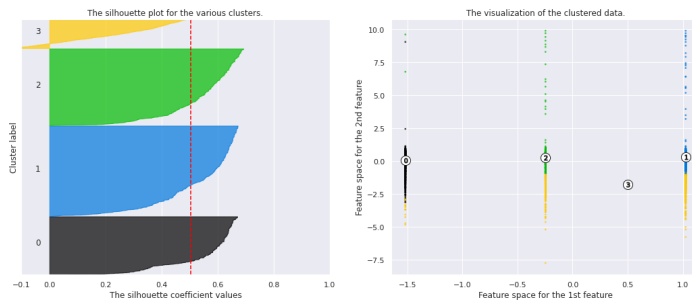
Silhouette analysis for KMeans clustering on sample data with $n_clusters = 2$



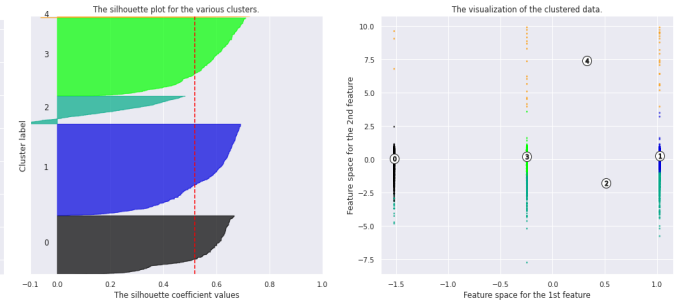
Silhouette analysis for KMeans clustering on sample data with $n_clusters = 3$



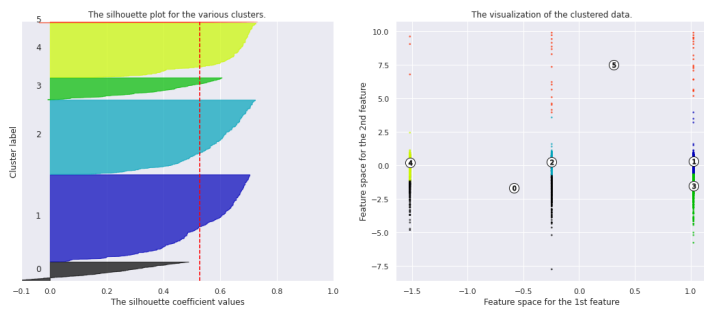
Silhouette analysis for KMeans clustering on sample data with $n_clusters = 4$



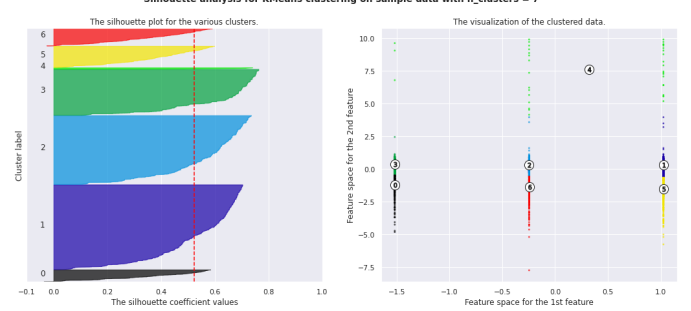
Silhouette analysis for KMeans clustering on sample data with $n_clusters = 5$



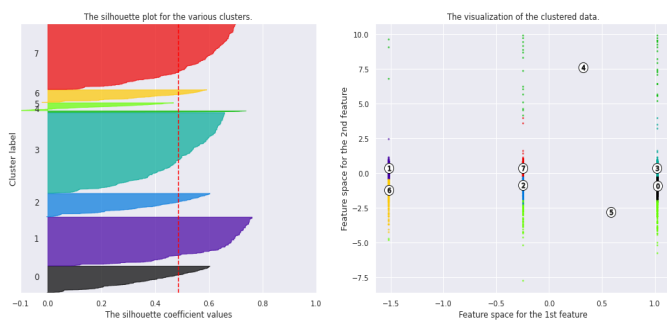
Silhouette analysis for KMeans clustering on sample data with $n_clusters = 6$



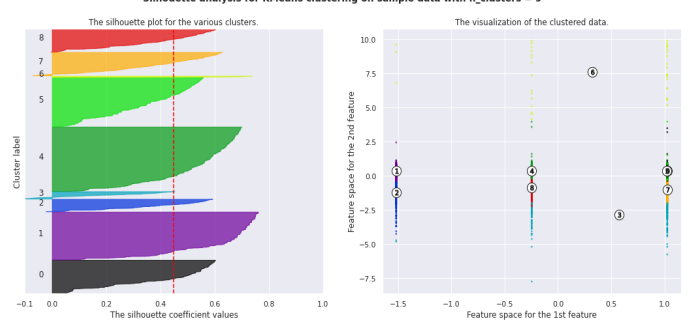
Silhouette analysis for KMeans clustering on sample data with $n_clusters = 7$

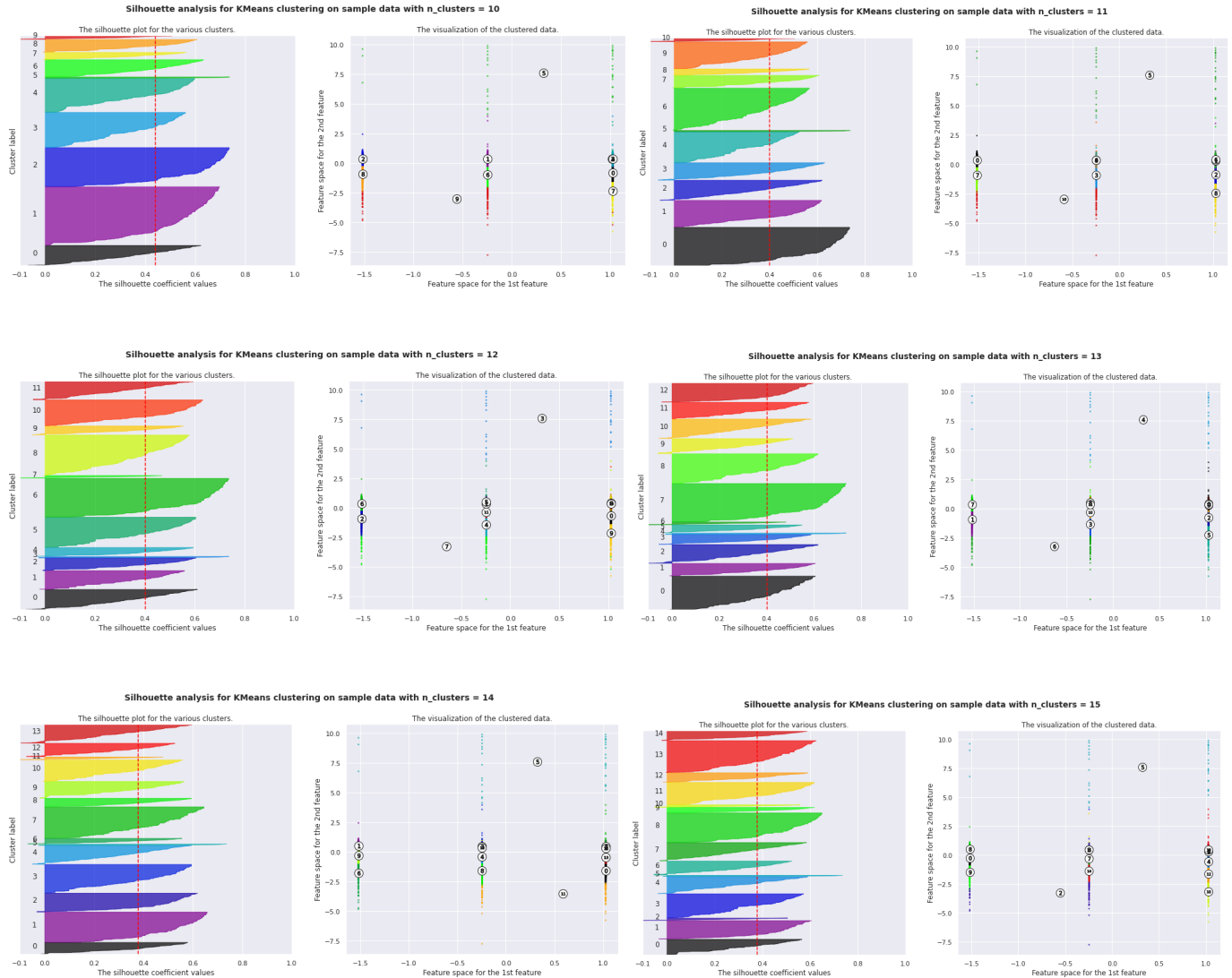


Silhouette analysis for KMeans clustering on sample data with $n_clusters = 8$



Silhouette analysis for KMeans clustering on sample data with $n_clusters = 9$

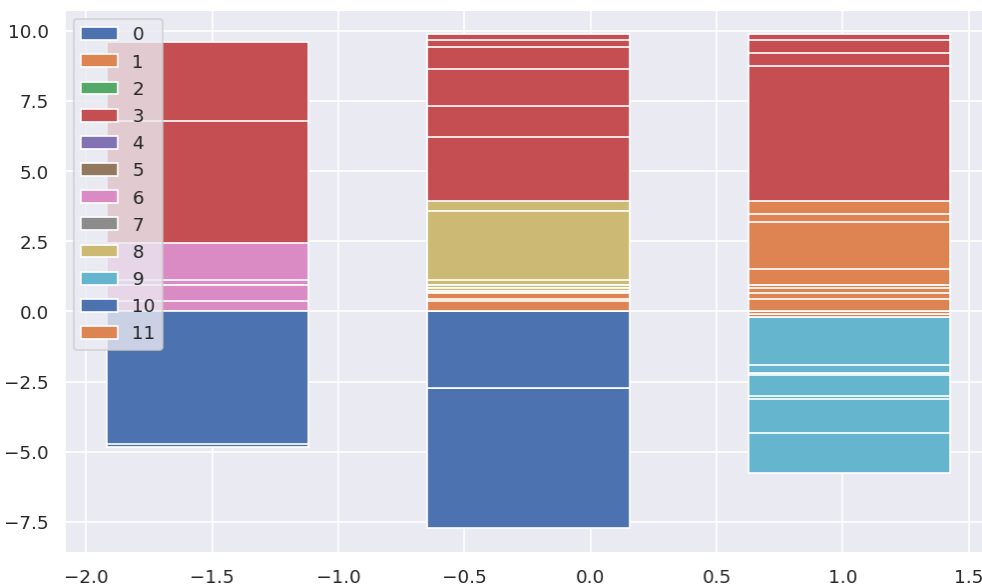
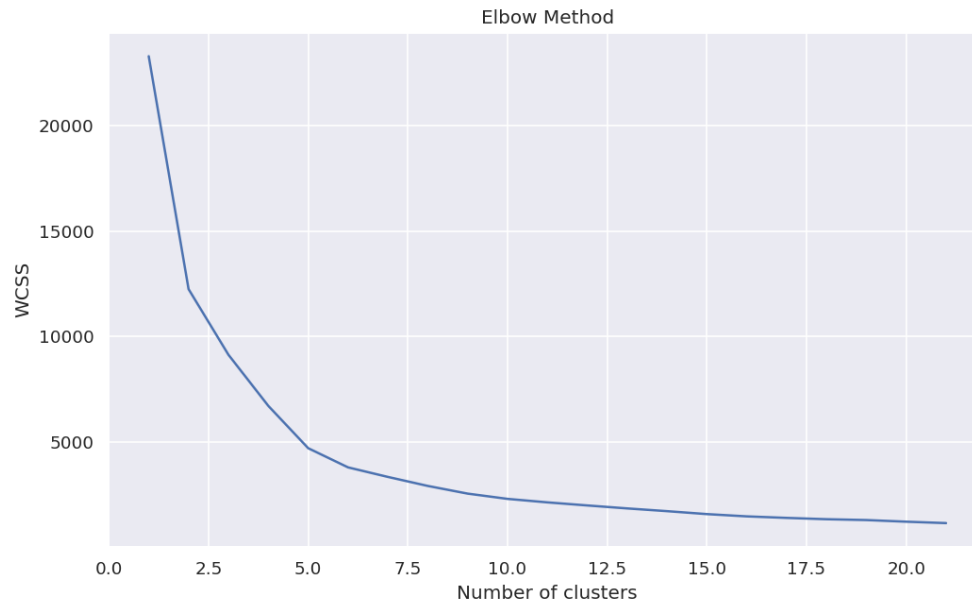




In above all pictures, we can clearly see that plot and score are different according to $n_{cluster}(k)$. So, we can easily choose a high score and number of k via silhouette analysis technique instead of elbow technique.

2. Elbow Method :

The elbow method is used to determine the optimal number of clusters in k-means clustering. The elbow method plots the value of the cost function produced by different values of k . As you know, if k increases, average distortion will decrease, each cluster will have fewer constituent instances, and the instances will be closer to their respective centroids. However, the improvements in average distortion will decline as k increases. The value of k at which improvement in distortion declines the most is called the elbow, at which we should stop dividing the data into further clusters.



3. DBSCAN.

DBSCAN stands for **d**ensity-**b**ased **s**patial **c**lustering of **a**pplications with **n**oise. It is able to find arbitrary shaped clusters and clusters with noise (i.e. outliers).

The main idea behind DBSCAN is that a point belongs to a cluster if it is close to many points from that cluster.

There are two key parameters of DBSCAN:

- **eps**: The distance that specifies the neighborhoods. Two points are considered to be neighbors if the distance between them are less than or equal to eps.
- **minPts**: Minimum number of data points to define a cluster.



As we can see these black dots, these are nothing but noise.


Recommendations

A recommender system aims to suggest relevant content or products to users that might be liked or purchased by them. It helps to find items that the user is looking for and they don't even realize it until the recommendation is displayed. Different strategies have to be applied for different clients and they are determined by available data.

First, we recommend movie titles.

As we can see, the word “**Golden Eye**” is a recommended Top 11 similar Movies title.

```
movie_recommendations = pd.DataFrame(recommendations('GoldenEye'), columns=['Recommendations'])  
movie_recommendations.head(11)
```

	Recommendations 
0	Tomorrow Never Dies
1	The World Is Not Enough
2	Die Another Day
3	The Foreigner
4	Casino Royale
5	Eurovision Song Contest: The Story of Fire Saga
6	Quantum of Solace
7	My Week with Marilyn
8	Bad Boys
9	Remember Me

Second, we recommend TV-show titles.

As we can see, the word “**Twice Upon A Time**” is a recommended Top 11 similar TV-shows title.

```
movie_recommendations = pd.DataFrame(recommendations('Twice Upon A Time'), columns=['Recommendations'])
movie_recommendations.head(11)
```

	Recommendations 
0	Familiar Wife
1	Revolutionary Love
2	Here to Heart
3	Momo Salon
4	Something in the Rain
5	The King: Eternal Monarch
6	Back to 1989
7	Love @ Seventeen
8	Operation Proposal
9	Dead in a Week (Or Your Money Back)

Conclusion

1. The majority of the Netflix content is movies, making it an interesting discovery.
2. There are two different forms of material in this dataset: movies (69.14%) and TV shows (30.86%).
3. We may infer from the dataset insights that the most TV shows were released in 2017, and the most movies were published in 2020.
4. But it has increasingly been concentrating more on television shows.
5. Among the top 5 countries that create all of the content that is made available on the site are the United States and India.
6. In fact, six of the top ten actors with the most content are Indian.
7. In text analysis (NLP) I used stop words, removed punctuations, stemming & TF-IDF vectorizer and other functions of NLP.
8. $k=10$ was found to be an optimal value for clusters using which we grouped our data into 10 distinct clusters.
9. In text analysis (NLP) I used stop words, removed punctuations, stemming & TF-IDF vectorizer and other functions of NLP.

Future Scope

1. Many exciting discoveries can be obtained by combining this dataset with other external datasets, such as IMDB ratings and rotten fruit.
2. A better recommender system might be developed with more time and then put online for users to use.