Capstone Project - 4

NETFLIX MOVIES AND TV SHOWS CLUSTERING

V Bhavya Reddy

Content:

- Introduction
- Problem Statement
- Data Description
- Null Value
- EDA
- Feature Engineering
- Topic Modelling
- NLP
- K-Means Clustering

Introduction

- Netflix is a subscription-based streaming service that allows its members to watch TV shows and movies without commercials on an internet-connected device.
- Netflix is an American company where the users make a monthly payment to watch movies and TV Shows.

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.
- Integrating this dataset with other external datasets such as IMDB ratings, rotten tomatoes can also provide many interesting findings.

In this project, you are required to do

- Exploratory Data Analysis
- Understanding what type content is available in different countries
- Is Netflix has increasingly focusing on TV rather than movies in recent years.
- Clustering similar content by matching text-based features

Data Description

- The data was collected from Flixable which is third party Netflix search engine. The dataset consists of movies and TV shows data till 2019. The dataset has 7787 rows of data.
- The dataset consists of eleven non- numeric columns and one numeric column.

Attribute Information:

1. **show_id :** Unique ID for every Movie / Tv Show
2. **type :** Identifier - A Movie or TV Show
3. **title :** Title of the Movie / Tv Show
4. **director :** Director of the Movie

Data Description

5. **cast :** Actors involved in the movie / show
6. **country :** Country where the movie / show was produced
7. **date_added :** Date it was added on Netflix
8. **release_year :** Actual Releaseyear of the movie / show
9. **rating :** TV Rating of the movie / show
10. **duration :** Total Duration - in minutes or number of seasons
11. **listed_in :** Genere
12. **description:** The Summary description

Null Value

- 'Director' feature has 30.68% null values.
- 'Cast' feature has 9.22% null values.
- 'Country' feature has 6.51% null values.

Filling NaN values with No Director, No Cast, Country Not Available.

- 'date_added' feature has 0.13% null values.
- 'rating' feature has 0.09% null values.

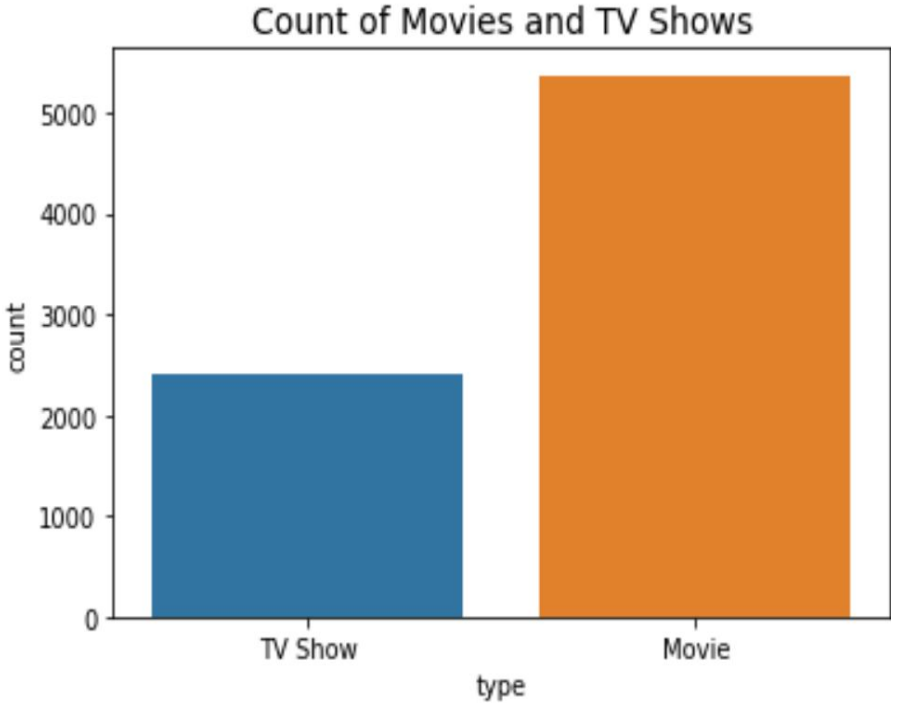
Drop NaN values from date_added, rating

Due to the high number of null values in director, cast, and country, dropping them would lead to imbalanced data and incorrect EDA analysis. Therefore, they are retained.

Exploratory Data Analysis

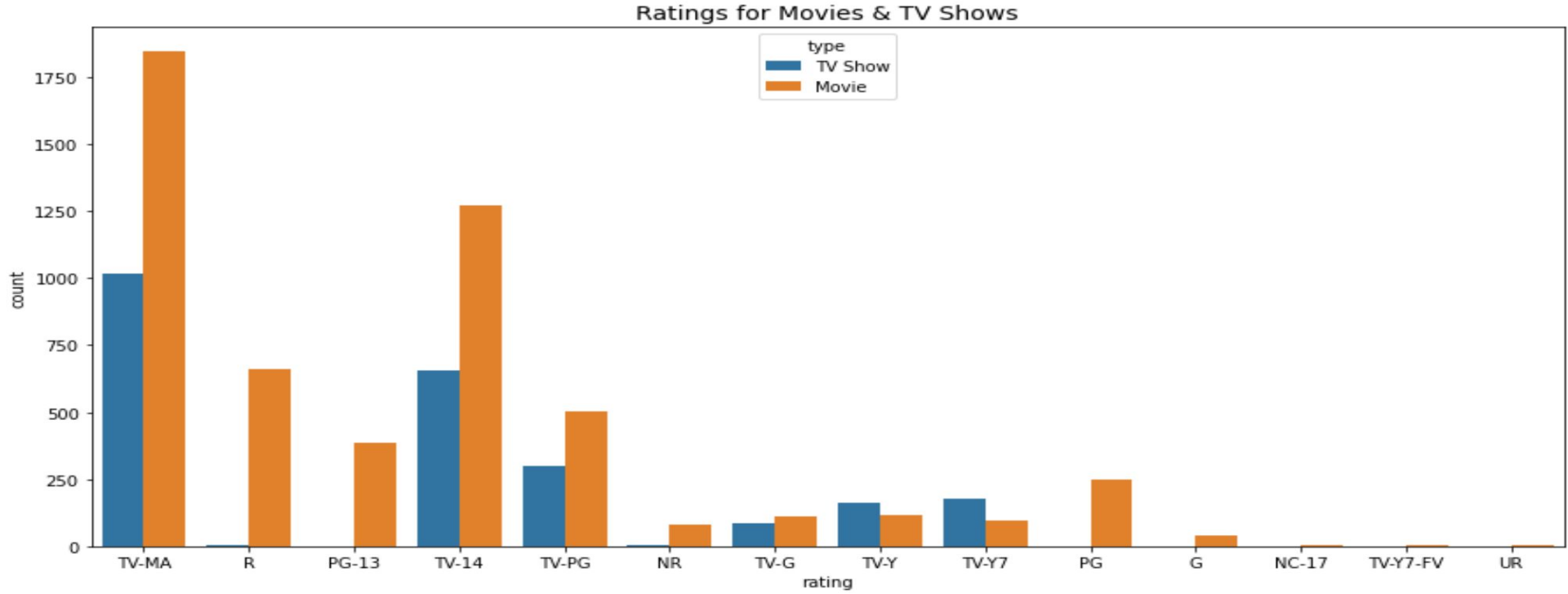
Type:

- It is evident that there are more movies on Netflix than TV shows.
- There are more than 5000+ movies and 2000+ TV shows.It should be noted that a TV Show has at least one season with many episodes, so the TV Shows count being less than Movies is logical.



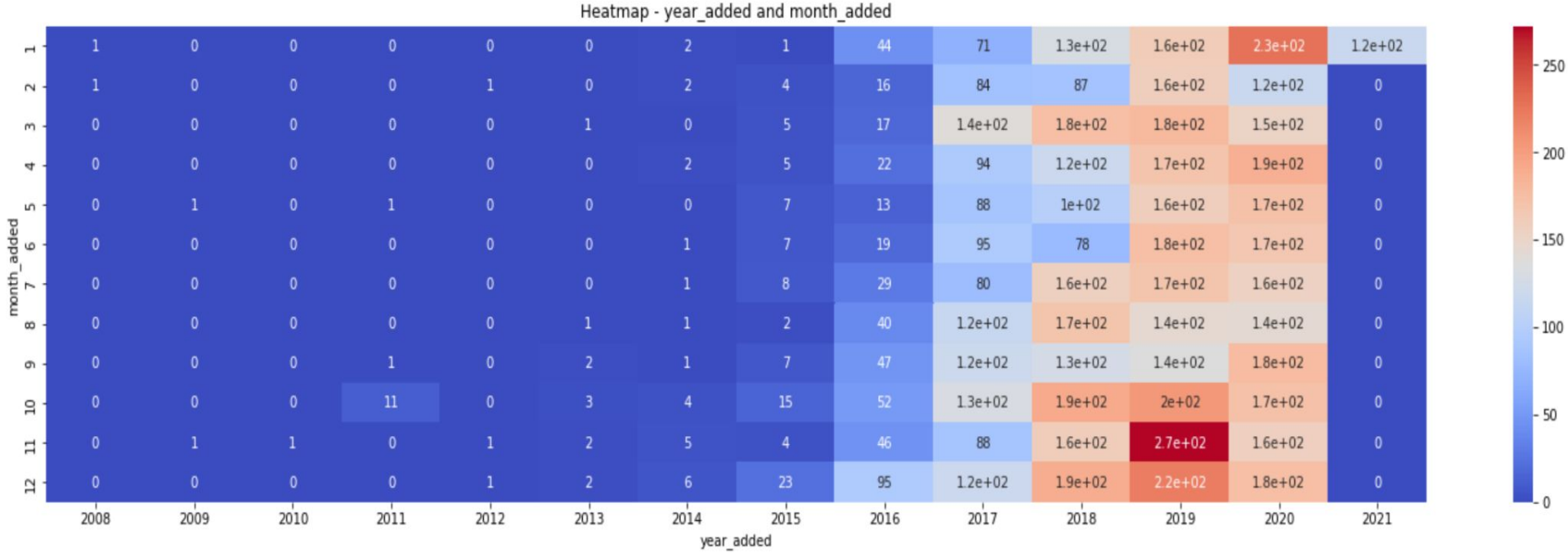
Exploratory Data Analysis

Netflix Film Ratings: The number of mature content movies is greater than the number of mature content TV shows. The majority of TV shows are geared toward younger viewers.



Exploratory Data Analysis

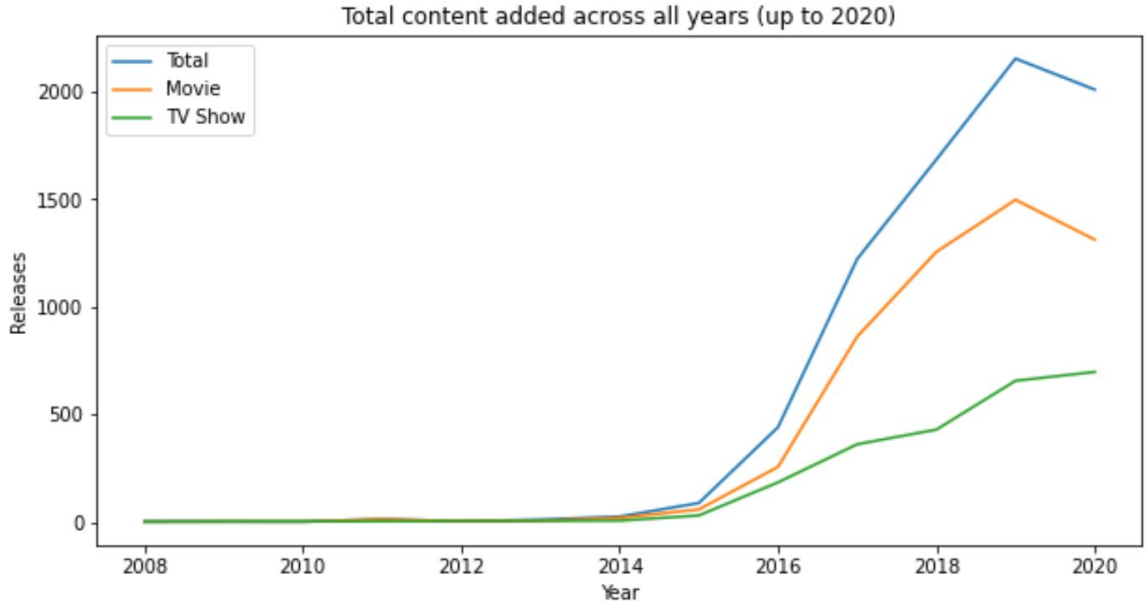
Heatmap - year_added and month_added



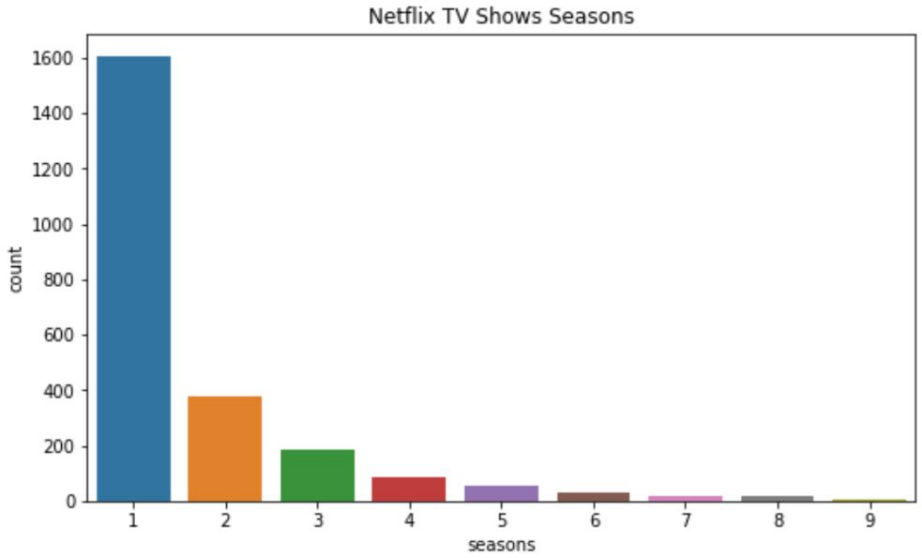
From 2014 the number of shows have increased. The data given to us stops at 2021 January.

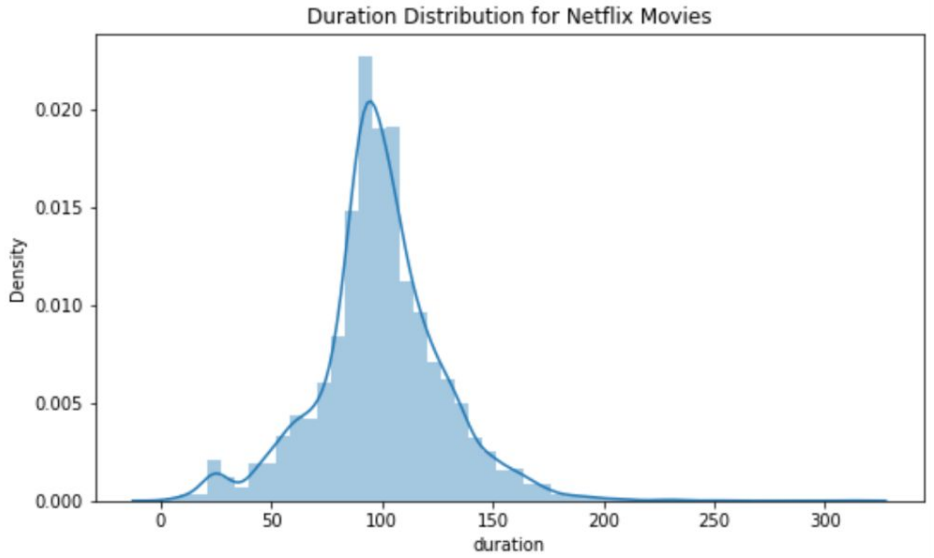
Exploratory Data Analysis

The number of movies in 2020 have reduced compared to its previous year. However the TV shows have increased.



Exploratory Data Analysis



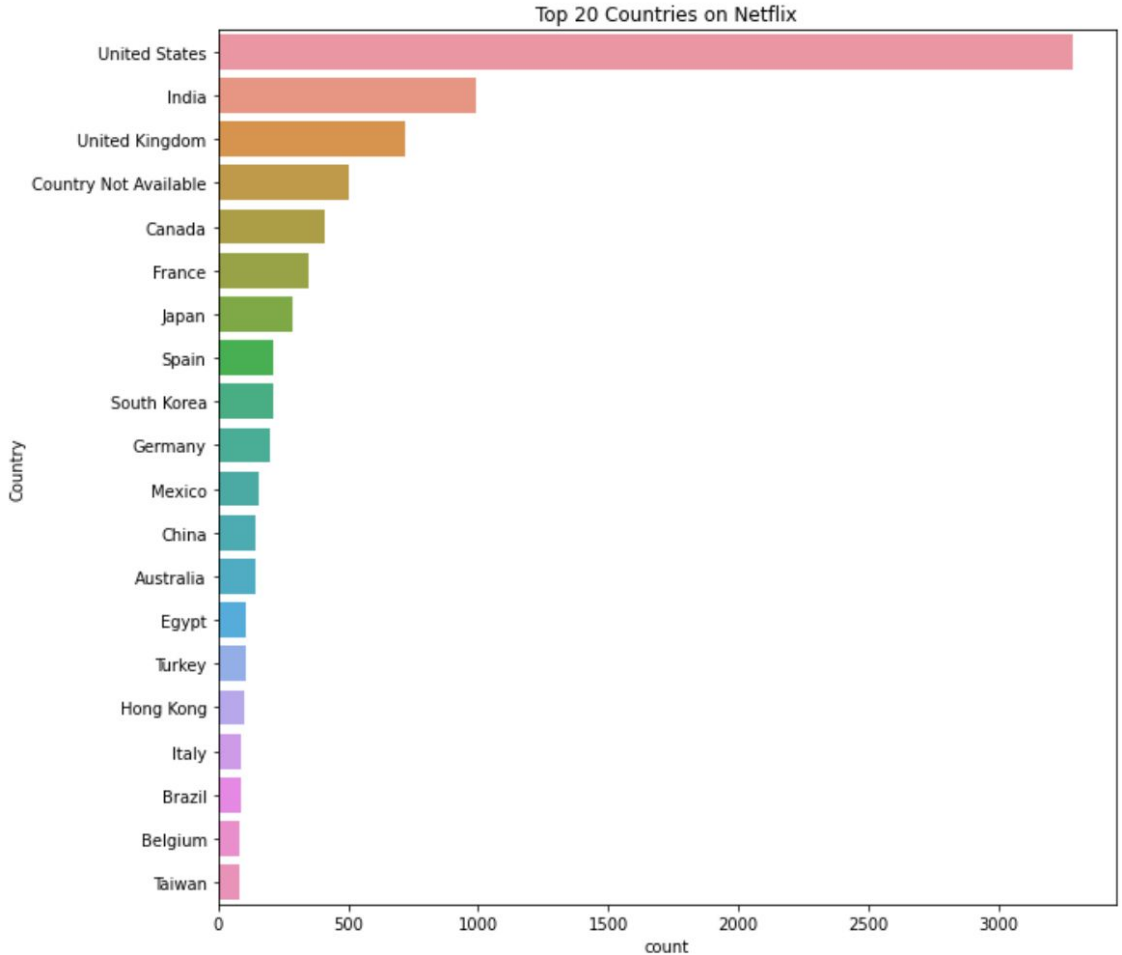


Seasons of Netflix TV shows are right-skewed and most have only one season.

Netflix movies have a normal distribution with a mean of 100 minutes.

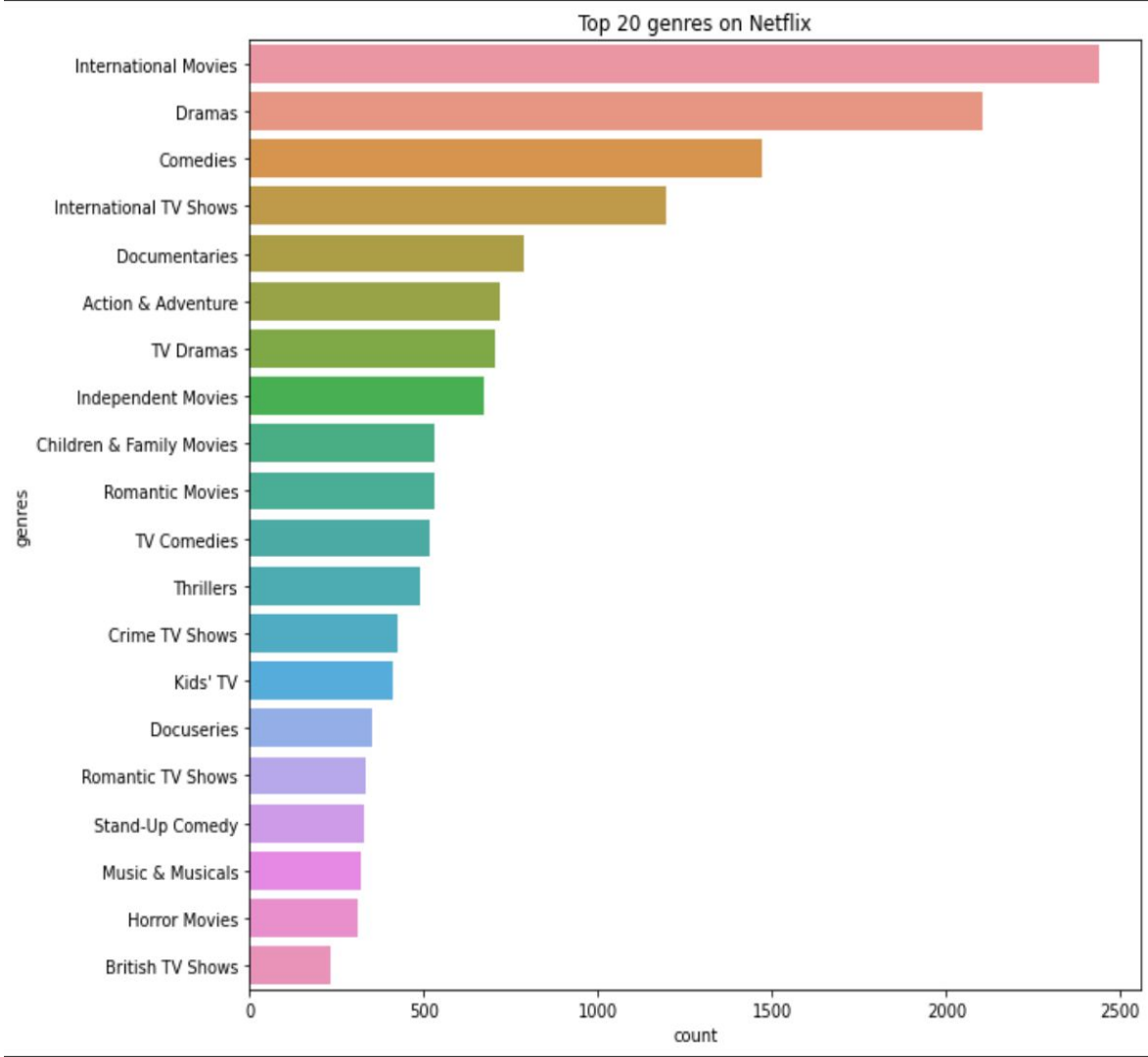
Exploratory Data Analysis

Knowing that Netflix is an American brand, it is no surprise that the United States is the most popular country to watch Netflix in. In second place is India, followed by the UK and Canada



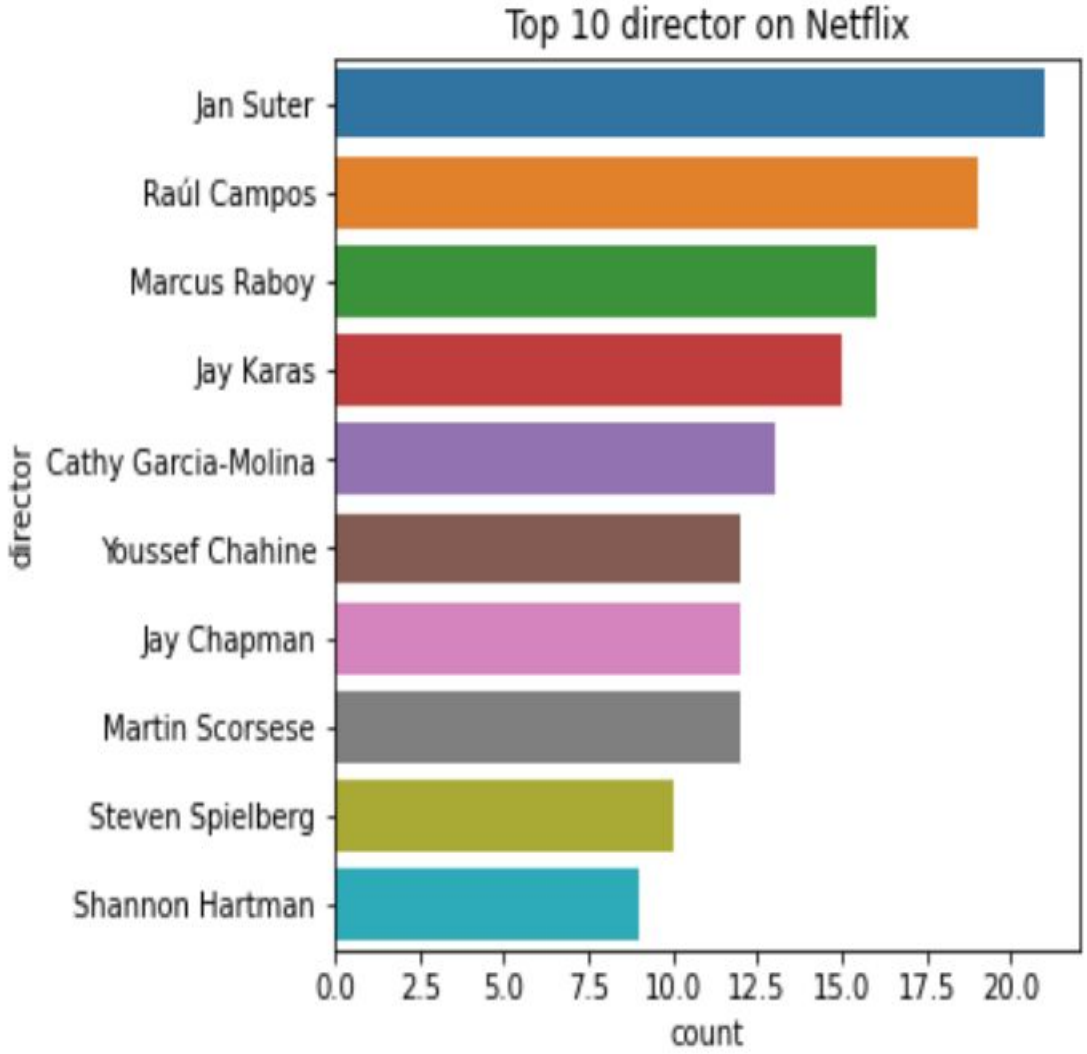
Exploratory Data Analysis

On Netflix, international movies are the most popular genre, followed by dramas and comedies. With Netflix having a lot of international subscribers, it makes sense to have international movies at the top.



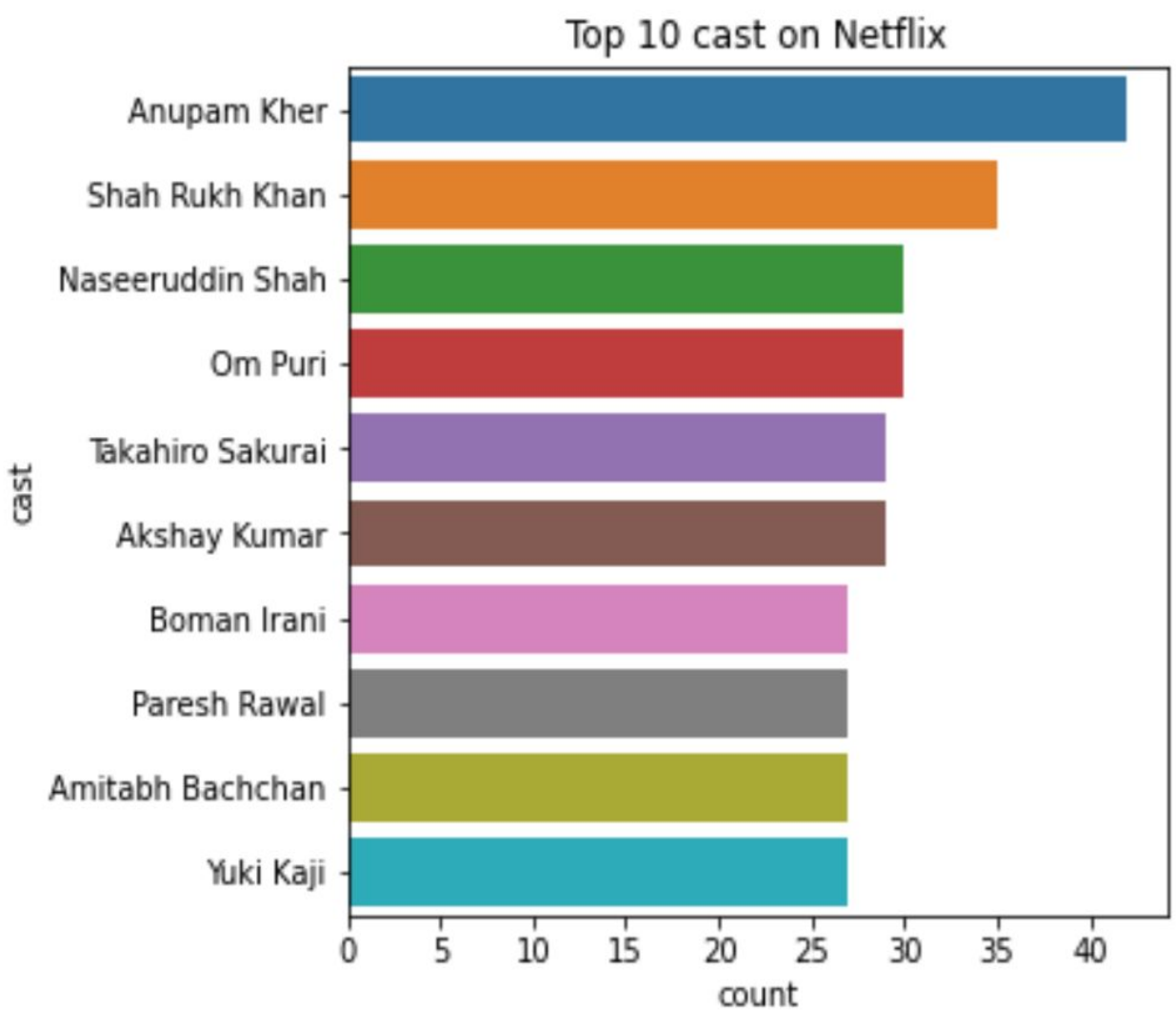
Exploratory Data Analysis

The number of international movies is higher, so it should come as no surprise that the most movie directors are international.



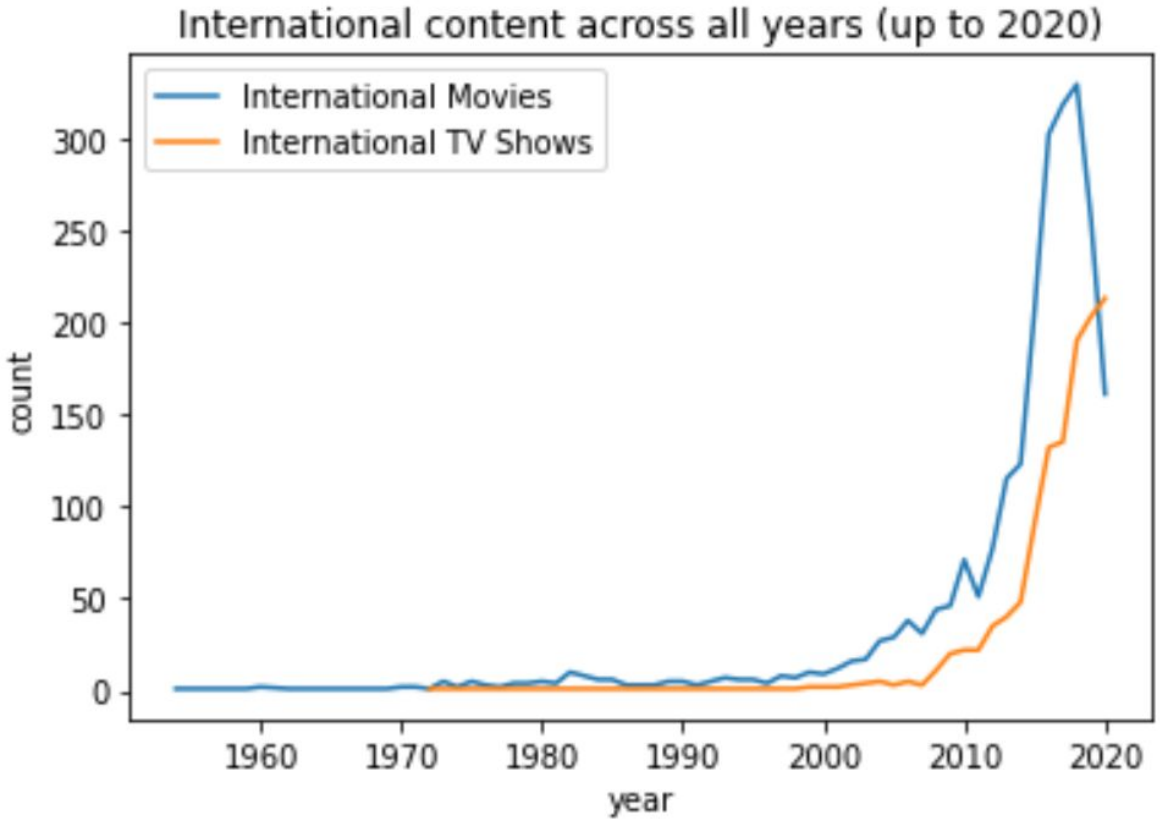
Exploratory Data Analysis

Top actors are mainly international actors implying that Netflix has many international subscribers.



Exploratory Data Analysis

More international movies are released than international television shows. The growth of international movies started to decline in 2018 and international TV shows started to decline in 2019.



Feature Engineering

Grouping ‘rating’ feature -

- From what I understand, 'TV-PG' refers to a TV show that is PG rated whereas just 'PG' refers to a PG rated movie. As a result, I merged them into PG.
- I also believe we can group all the ratings into G, PG, PG-13, R, NC-17, and No Rating(NR).
- These are the rating groups:
  - UR, NR as 0
  - G, TV-Y, TV-G as 1
  - PG, TV-Y7, TV-PG as 2
  - TV-14, PG-14, TV-Y7-FV as 3
  - R as 4
  - NC-17, TV-MA as 5

Feature Engineering

Ordinal Encoding on ‘type’:

- Every show is either a Movie or TV Show.
- So, it is only logical to perform ordinal encoding.

TF-IDF Vectorisation:

- Calculated TF-IDF for director, cast, country, listed_in and then sum every row.
- This leads to director_sum, cast_sum, country_sum, listed_in_sum.
- Used the inbuilt function on description.

StandardScaler on dataset:

- Instead of MaxMinScaler, I used standardisation on the dataset as it is more insensitive to outliers.

Understanding what type content is available in different countries

1. The first thing I did was take the top 20 countries and plot their type features.
2. In the next section, I have plotted all the features in one plot with many subplots. A topic modelling was applied to the description.

OBSERVATION:

1. The number of movies in many countries is higher than the number of TV shows. There are few countries with more TV shows than movies, such as Japan, South Korea, Taiwan, etc.

Understanding what type content is available in different countries

Country Specific Observation:

1. I am understanding the content in the United States.
2. There are five topics in Topic Modelling, and here is what I understand about each:
    a. Topic - 1 : There is a strong focus on high school and teenage stories in this topic.
    b. Topic - 2 : Typically, the topic focuses on family-related comedy series or dramas.
    c. Topic - 3 : This topic focuses on documentaries and series from around the world.
    d. Topic - 4 : It focuses primarily on documentaries.
    e. Topic - 5 : It focuses on New York and star-studded lifestyles.

Understanding what type content is available in different countries

3. The number of movies on Netflix in the United States is greater than the number of TV shows. A majority of movies are aimed at mature audiences, and TV shows are aimed at both mature and younger audiences.

a. Adam Sandler is the most frequently cast actor in the United States.
b. Drama, comedy, and documentary are the most popular genres in the United States.
c. The majority of American movies last for 100 minutes.

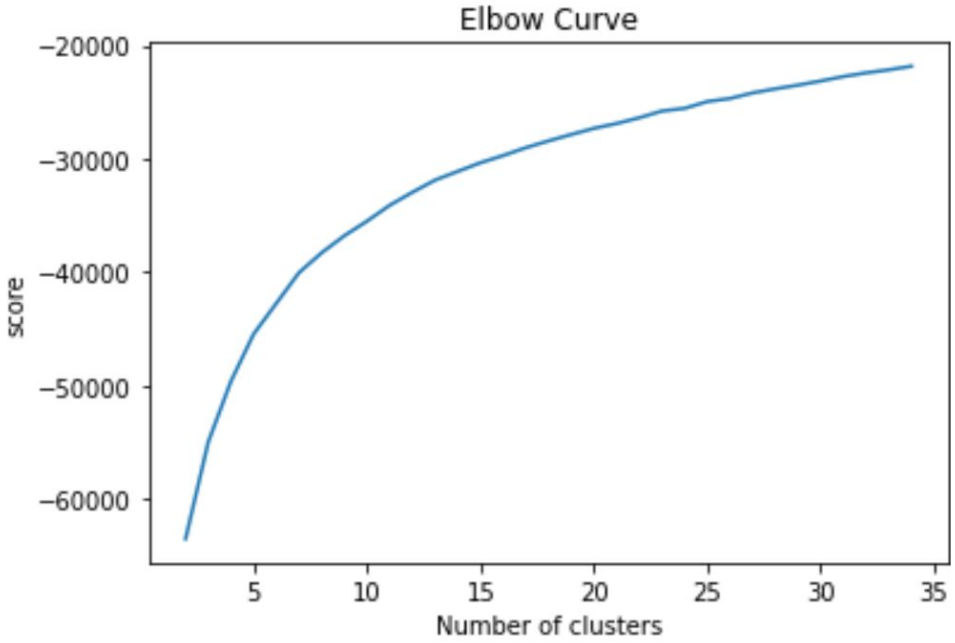
K-Means Clustering:

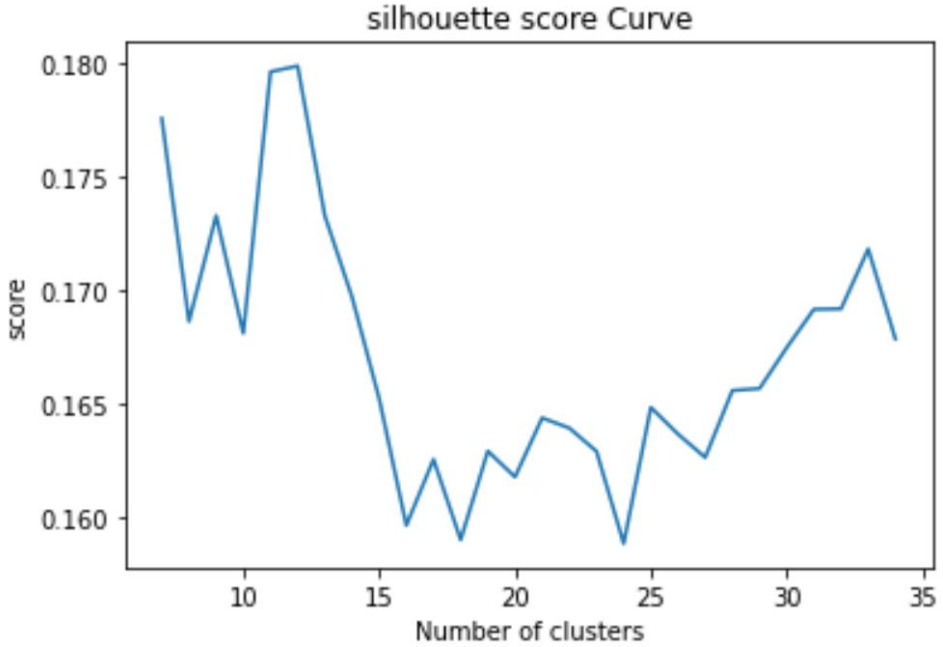
In this segment I am considering the following features - type, director, cast, country, year_added, month_added, release_year, rating, duration, listed_in

Prior to creating the k-means clustering model perform data featuring such as:

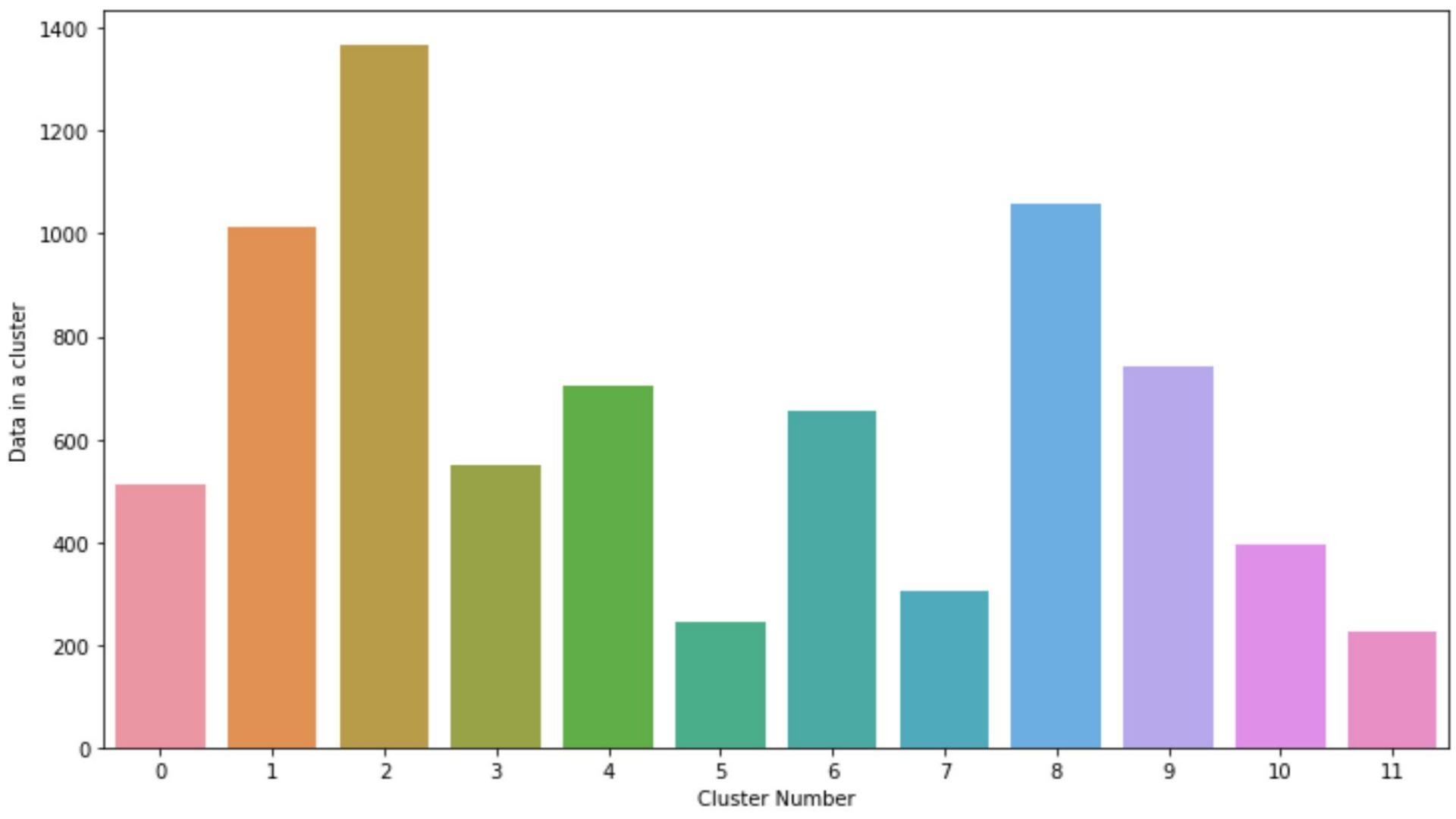
- New features include -
  - Grouping 'rating' feature.
  - Creating - 'Director_sum', 'Cast_sum', 'Country_sum', 'listed_in_sum'
  - Standardizing all the features.
- Dropping 'show_id', 'title', 'director', 'cast', 'country', 'date_added', 'listed_in', 'description', 'duration' features.

K-Means Clustering:





The number of clusters is shown to be 12.


1400
1200
1000
800
600
400
200
0
Data in a cluster
0
1
2
3
4
5
6
7
8
9
10
11
Cluster Number

K-Means Clustering:

Observation:

In the cluster containing '13 Reasons Why', the following details can be inferred:

1. The cluster number is 3.
2. Cluster-3 has mostly TV shows and very few movies.
3. In the director feature, this cluster has mostly 'No Director'.
4. In terms of country features, the cluster consists mostly of the US, UK, Canada, and India.
5. The date_added feature is between 2015 and 2021.
   Below are few of the movies recommended for '13 Reasons Why':

'#blackAF', '100 Humans', '13 Reasons Why', '13 Reasons Why: Beyond the Reasons', '3Below: Tales of Arcadia', '7 (Seven)', '9 Months That Made You', 'A Little Help with Carol Burnett', 'A Series of Unfortunate Events', 'A Year In Space', 'A.D. Kingdom and Empire', 'Abstract: The Art of Design', 'Absurd Planet', 'Adam Ruins Everything', 'AJ and the Queen', 'Alexa & Katie', 'All About the Washingtons', 'All American', etc.

Natural Language Processing (NLP) Model

The following actions are performed on ‘description’ feature:

- First, remove the punctuations.
- Next, remove the stopwords.
- Apply snowball stemmer.
- Tfidf vectorizer on the description feature.
- Find the recommended movie for solely the description.

NOTE: In order to provide better recommendations, I combined the TF-IDF scores of the top 10 words of the particular movie's description and the same words scores for all the other movies. Recommendations were given to the closest candidate.

Natural Language Processing (NLP) Model

Top 10 word list in "13 Reasons Why":
tape
choic
receiv
suicid
unravel
tragic
classmat
teenag
mysteri
girl

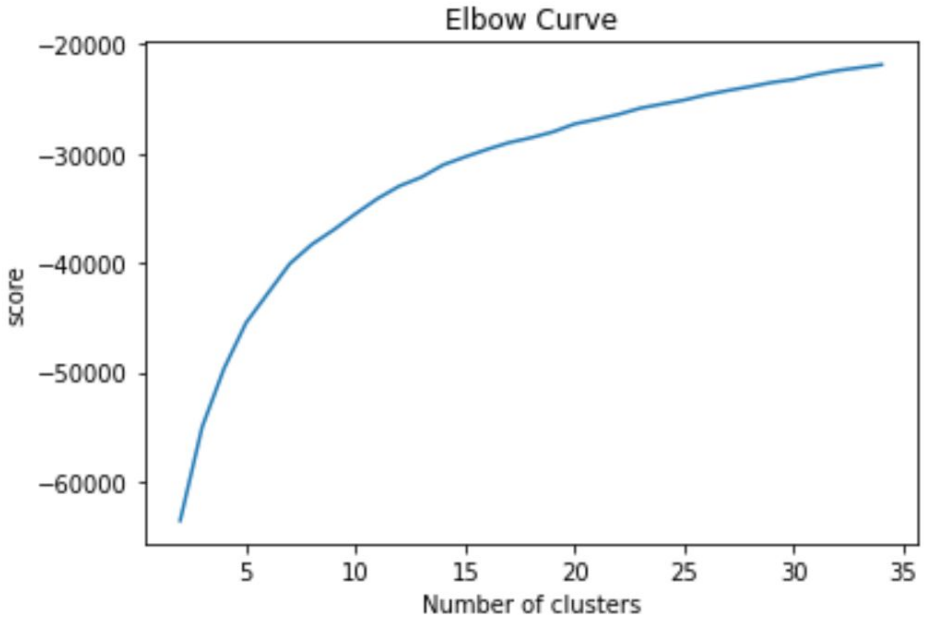
Description for 13 Reasons Why:
After a teenage girl's perplexing suicide, a classmate receives a series of tapes that unravel the mystery of her tragic choice.

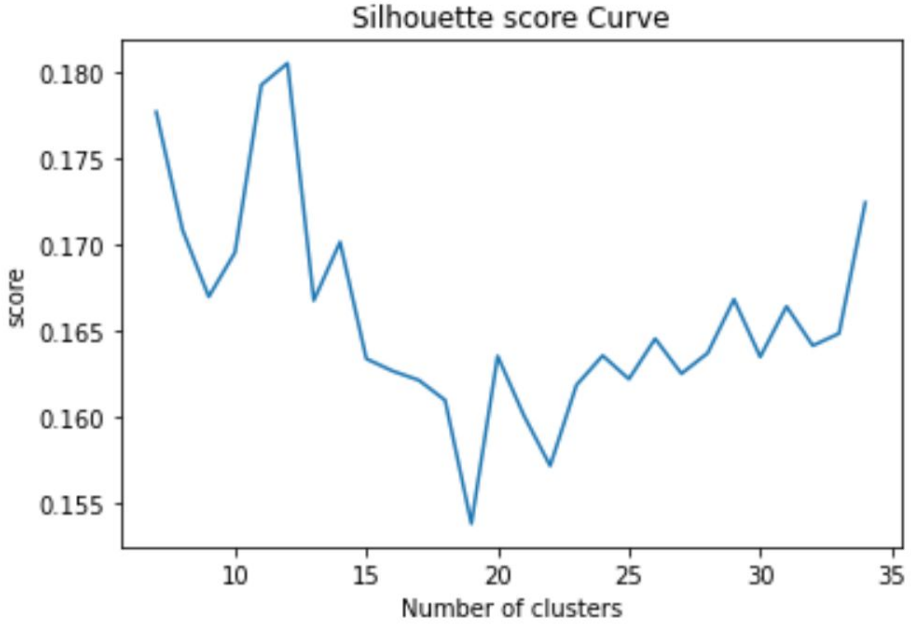
The following are few of the recommended Movies:

Movie - Not Alone :
description - An 18-year-old struggling to understand her best friend's suicide talks to teenagers who have grappled with mental illness and suicidal thoughts.

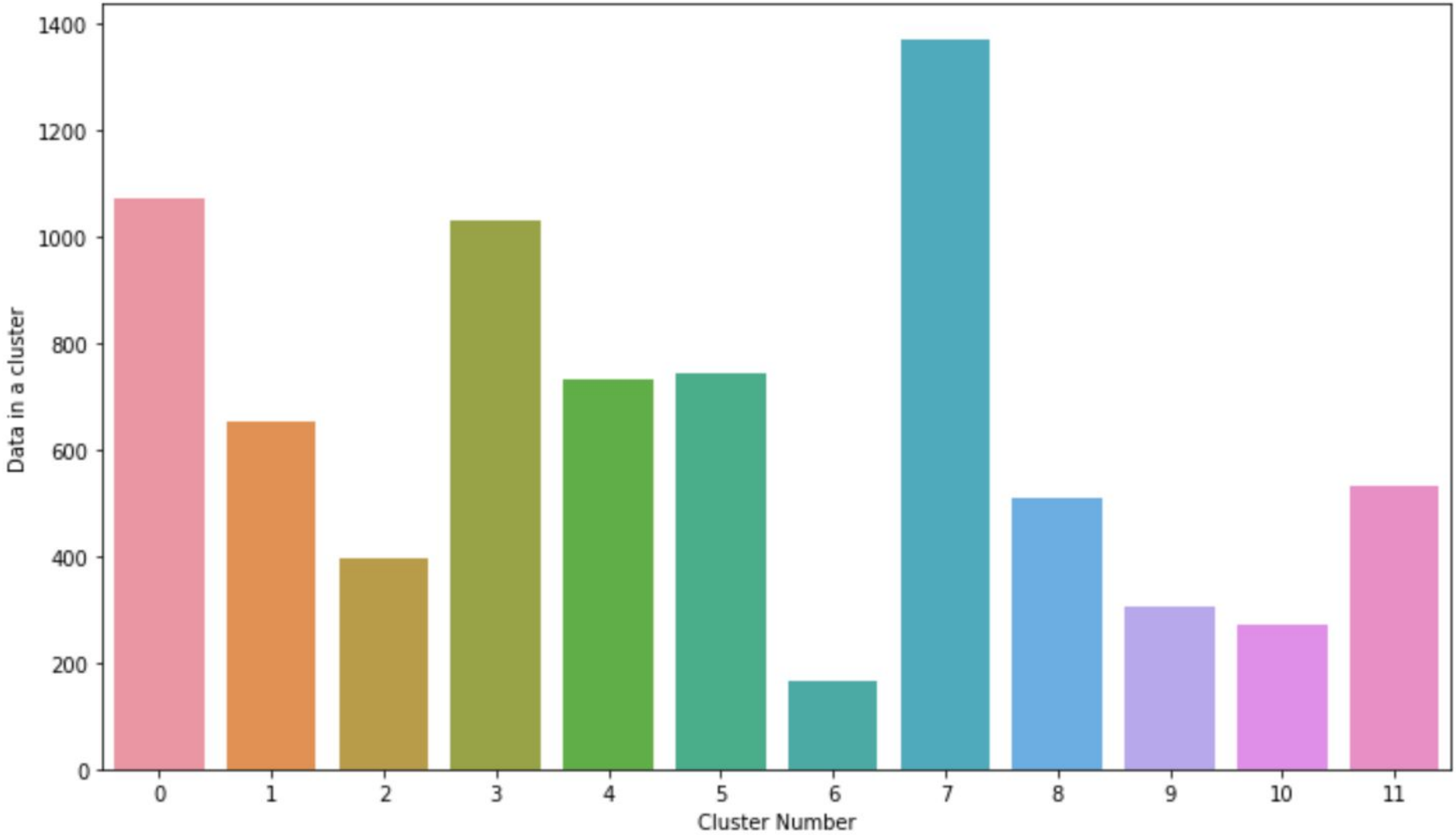
Movie - Devil's Bride :
description - On a small Finnish island in 1666, a teenage girl in love with a married fisherman becomes the center of a tragic witch hunt and power struggle.

NLP + K-Means Clustering





The number of clusters is chosen to be 12.


1400
1200
1000
800
600
400
200
0
Data in a cluster
0
1
2
3
4
5
6
7
8
9
10
11
Cluster Number

NLP + K-Means Clustering

This section differs from the K-Means Clustering section by including description obtained from NLP section.

Below are few of the movies recommended for '13 Reasons Why' -

'#blackAF', '100 Humans', '13 Reasons Why: Beyond the Reasons', '3Below: Tales of Arcadia', '7 (Seven)', '9 Months That Made You', 'A Little Help with Carol Burnett', 'A Series of Unfortunate Events', etc.

VERDICT:
The suggestions have slightly changed.

Conclusion:

1. We started by replacing Nan values with No Director, No Cast, and Country Not Available for director, cast, and country respectively. Nan values were dropped from date_added and rating features.
2. 'date_added' feature is used to obtain the 'month_added' and 'year_added' features.
3. As a first step, we compare the top 20 countries by type - Movies vs TV Shows. From the dataset, we select country-specific data and compute LDA and Document Term Matrix(DTM). Several plots were taken to evaluate the available content for specific countries. There were several conclusions made regarding the United States, as an example.
4. In the feature engineering - the rating values were reassigned, ordinal encoding was used on the type, the string value was dropped from the duration, etc.
5. K-Means Clustering was performed on type, director, cast, country, year_added, month_added, release_year, rating, duration, listed_in features. The silhouette_score was used to calculate the number of clusters. A test on '13 Reasons Why' was then conducted to determine the recommendations.
6. In NLP, we used only the description to determine recommendations.
7. A K-means clustering dataset with TF-IDF vectoriser from NLP is then combined for clustering with a little tweaking in the features.

Conclusion:

1. A recommendation system with the description column works well.
2. In the case of K-means, the optimal number of clusters are 12.
3. When K-means is applied to the description sum column, the optimal number of clusters was also 12.
4. Clustering with the description column did make a few changes in the earlier recommendations.
5. The optimal number of clusters was calculated using silhouette_score.

Conclusion:

1. The most content type on Netflix is movies.
2. The largest count of Netflix content is made with a 'TV-MA' rating.
3. After 2014 the amount of content added has been increasing significantly.
4. The number of movies in 2020 have reduced compared to the previous year. However, the number of TV shows has increased.
5. While most TV seasons have only 1 season, movie lengths follow a normal distribution with a mean of 100 minutes.
6. According to the amount of content produced, the United States is the top country.
7. International Movies are a genre mostly found on Netflix.
8. In terms of titles, Jan Suter is the most popular director on Netflix.
9. Anupam Kher is the most popular Netflix cast member, according to number of movies made.
10. In 2018, 2019, and 2020, the majority of films were released.
11. A large number of movies and TV Shows were released in October, November, December, and January.