Capstone Project - 2

Team: Reality

NETFLIX MOVIES AND TV SHOWS CLUSTERING

Team Members

AYUSH JAIN

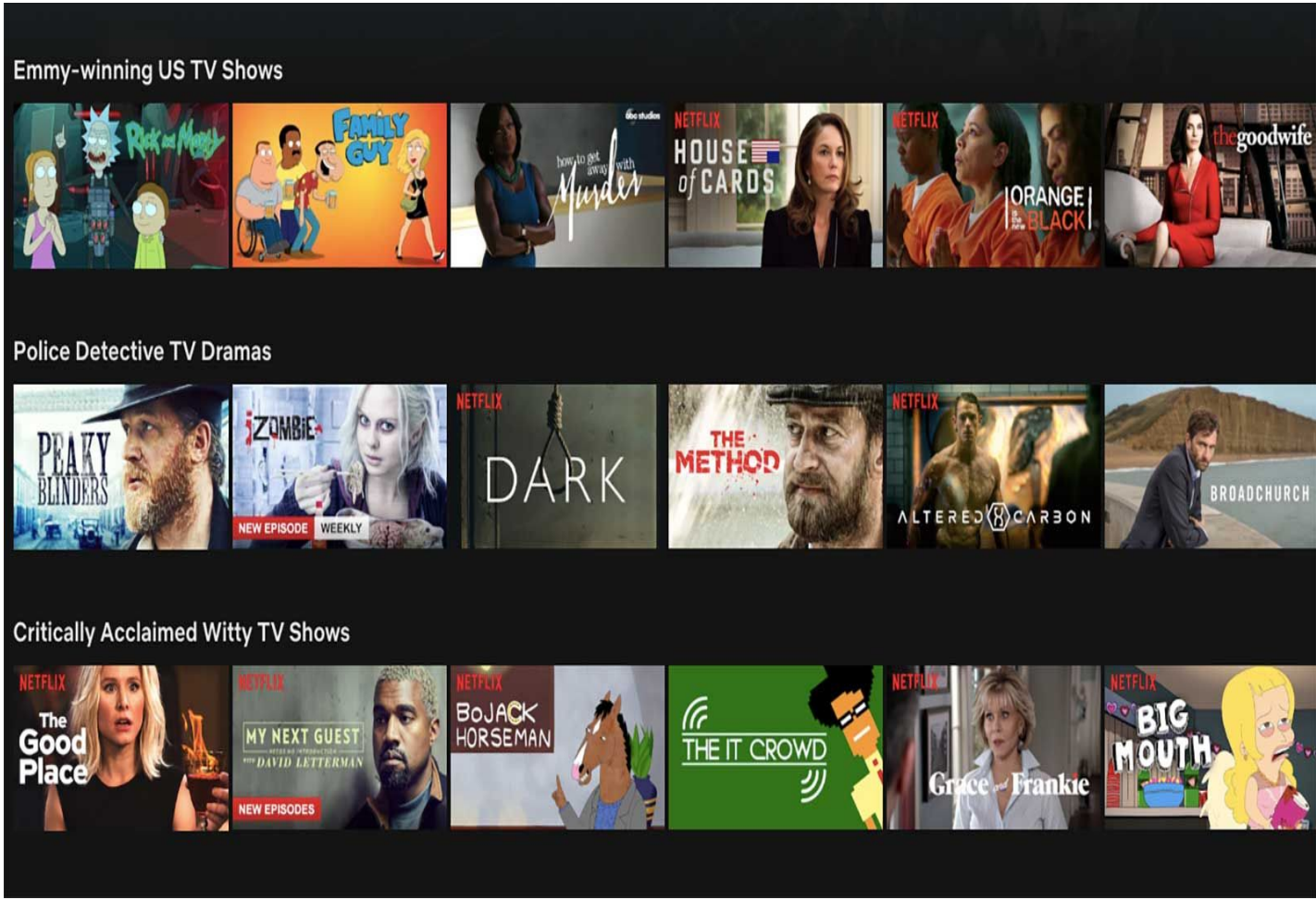
DHONGARI PAVAN

NADEEHA A

ROMALY DAS

SREENIVASAN KV

YASH PATIL


Emmy-winning US TV Shows
Rick and Morty
Family Guy
how to get away with Murder
NETFLIX
HOUSE of CARDS
NETFLIX
ORANGE is the new BLACK
thegoodwife
Police Detective TV Dramas
PEAKY BLINDERS
iZOMBIE
NEW EPISODE
WEEKLY
NETFLIX
DARK
THE METHOD
NETFLIX
ALTERED CARBON
BROADCHURCH
Critically Acclaimed Witty TV Shows
NETFLIX
The Good Place
NETFLIX
MY NEXT GUEST
with DAVID LETTERMAN
NEW EPISODES
NETFLIX
BoJACK HORSEMAN
THE IT CROWD
NETFLIX
Grace and Frankie
NETFLIX
BIG MOUTH

What is expected from this project?

- Exploratory Data Analysis
- Understanding what type of content is available in different Countries
- Is Netflix has increasingly focussed on TV rather than Movies in recent years?
- Clustering similar content by matching text based features.

Closer look at the dataset..!!

- 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 **textual** columns and one **numeric** column.

  Show id:- It is unique for all the movies/tv shows.
  Type :- Type of content i.e movie/tv show.
  Title :- Name of the movie/tv show.
  Director :- Name of the director

Closer look at the dataset..!!

Cast:- Actors involved in the movie/tv show.

Country :- Country where the movie/tv show is produced.

Data added :- Date when movie/tv show is added to Netflix.

Release year :- Year when the movie was released.

Rating :- Content rating.

Listed_in :- Genres of the movie/tv show.

Description :- The summary of the movie/tv show.

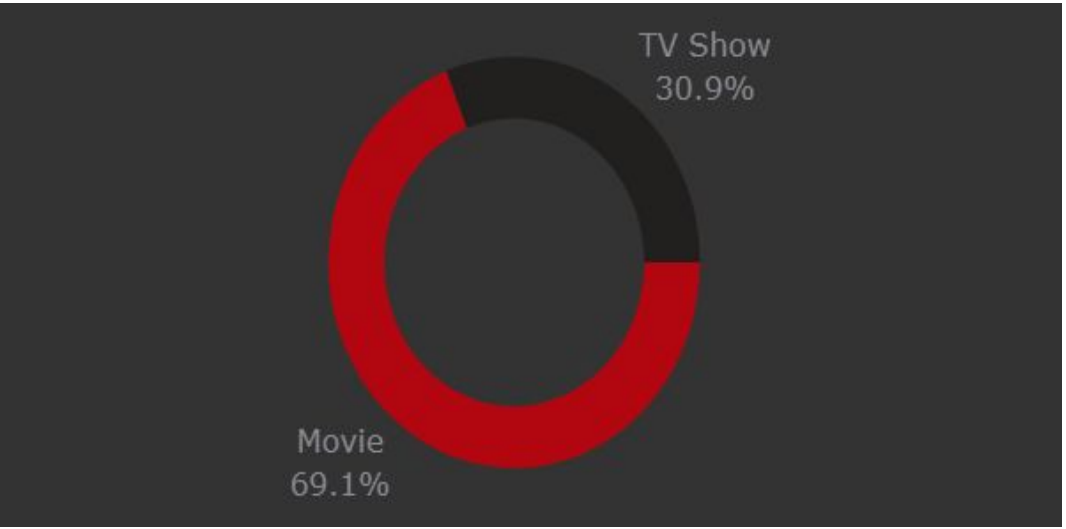
Duration :- total duration of movie in minutes/ seasons for tv shows.

Closer look at the dataset..!!

- There are 2389 null values in **Director** column
- There are 718 null values in **cast** column
- There are 507 null values in **country** column
- There are 10 null values in **date added** column
- There are 7 null values in **rating** column

Exploratory Data Analysis

TV or Movie Shows?

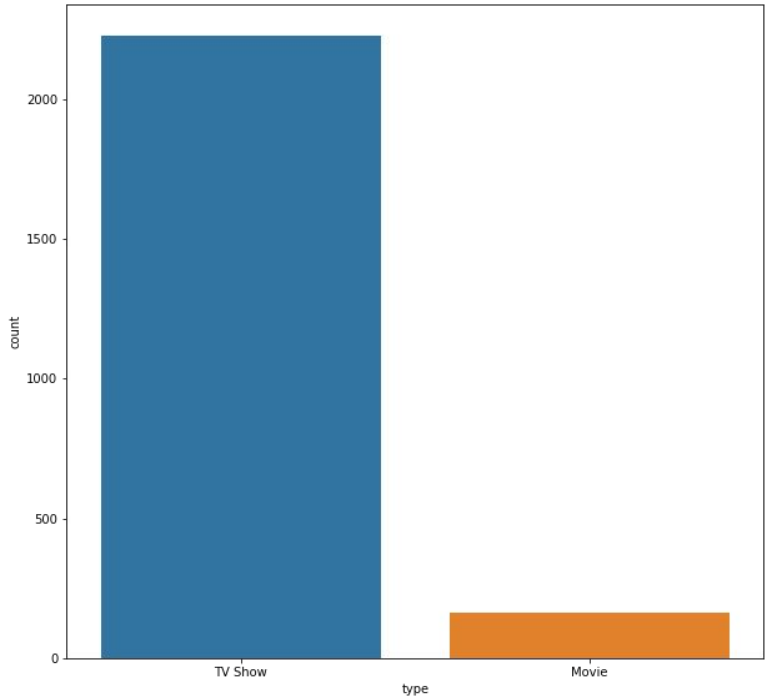


- There are 5377 movies and 2410 tv shows.

Common Titles!

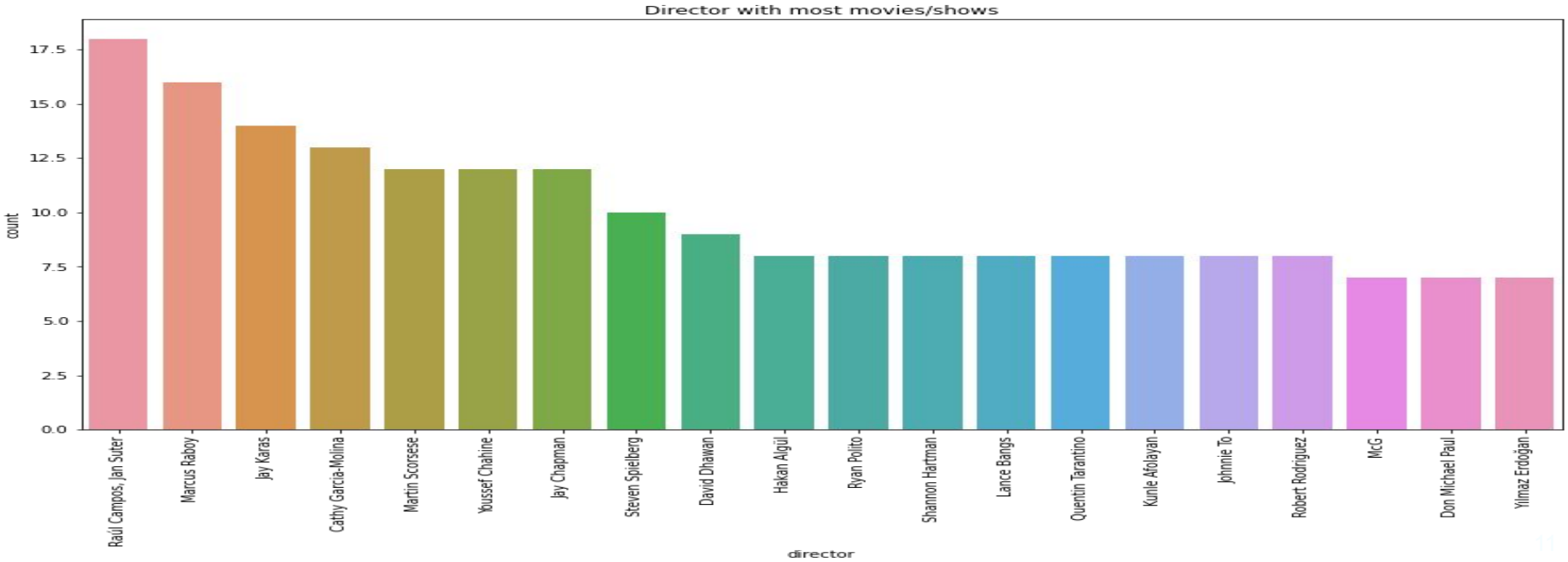
- It seems like words like "Love", "Man", "World", "Story" are very common in titles.

EDA (Continued..)

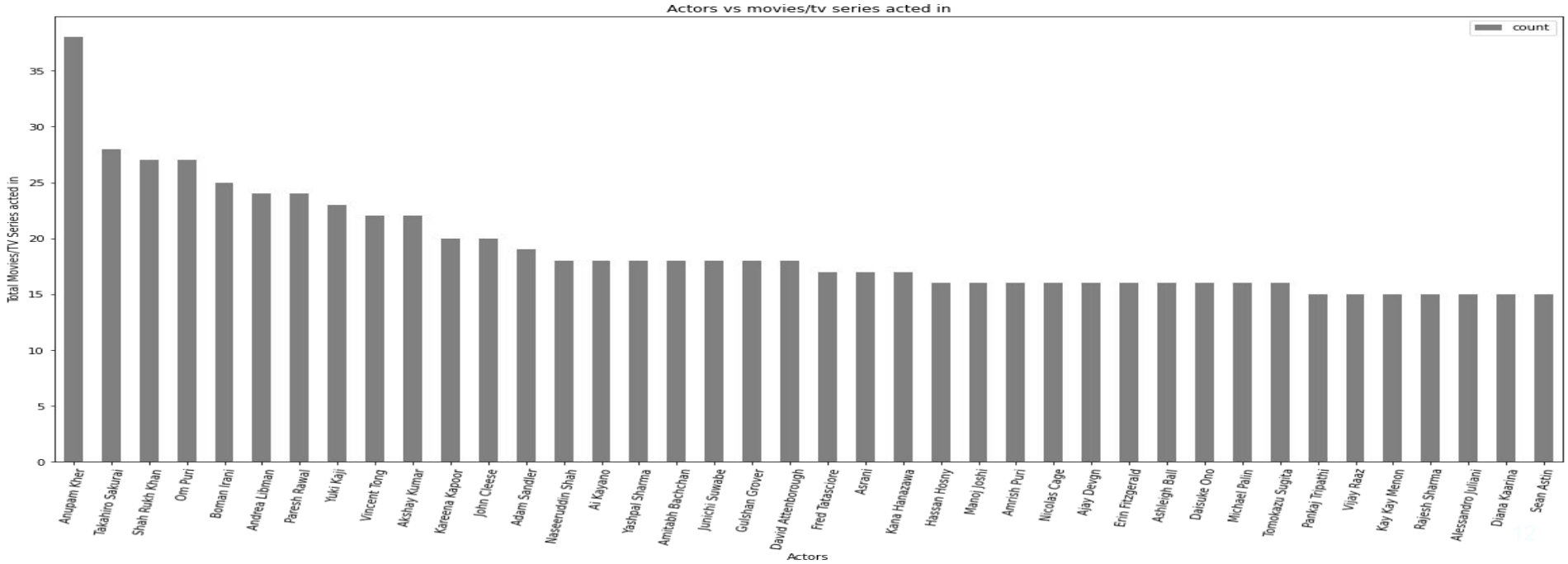


- It looks like most of the missing values of Director is for tv shows.

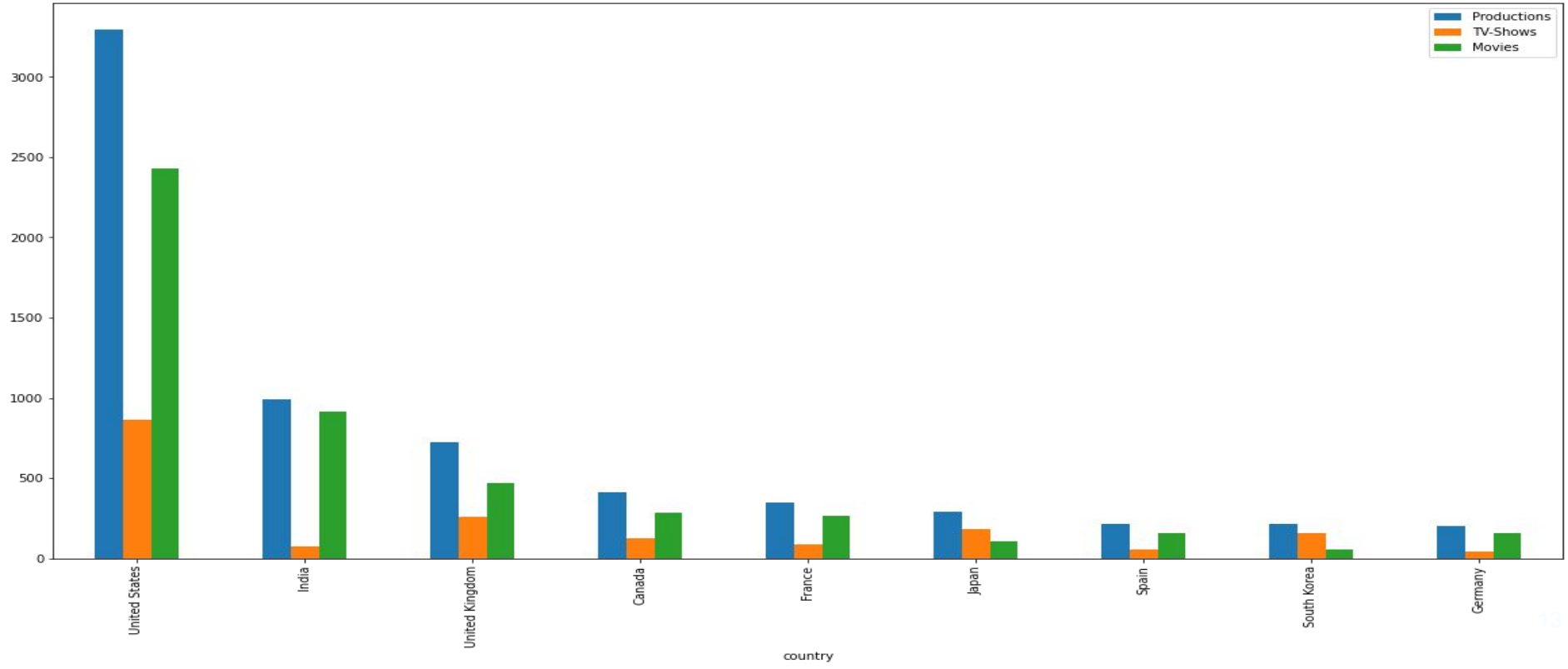
EDA (Continued..)



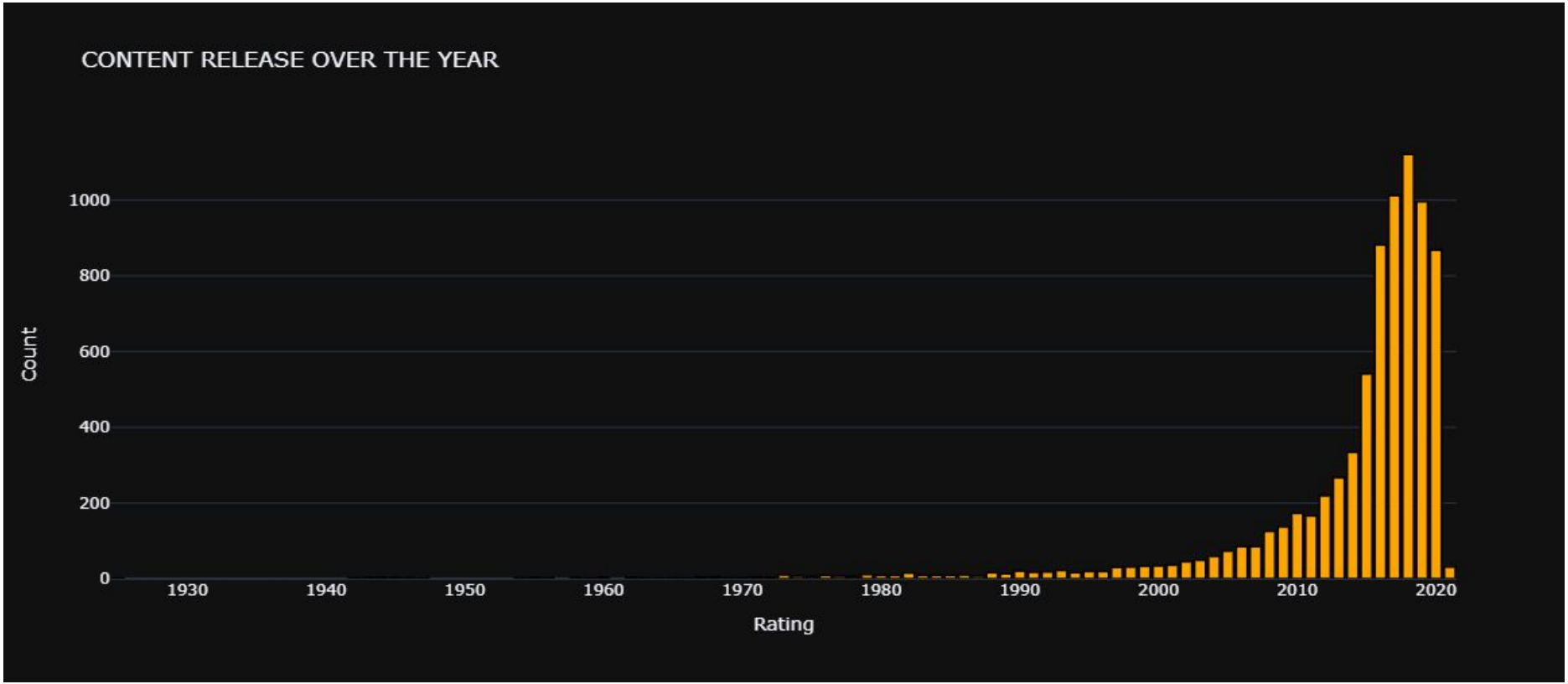
EDA (Continued..)



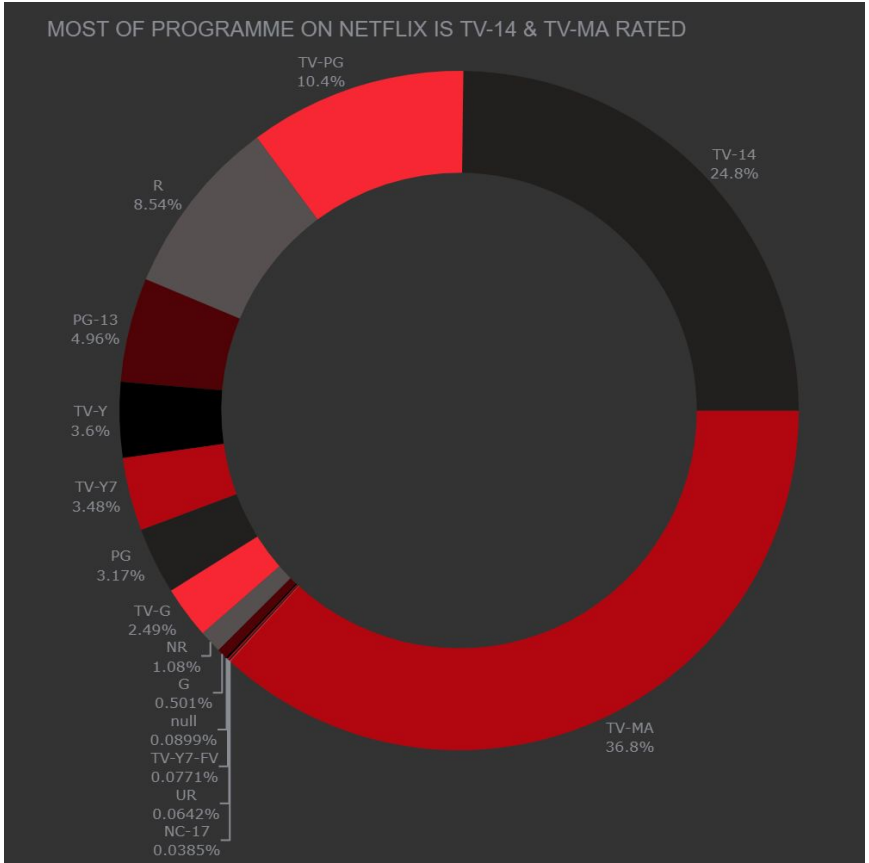
Production split across top countries



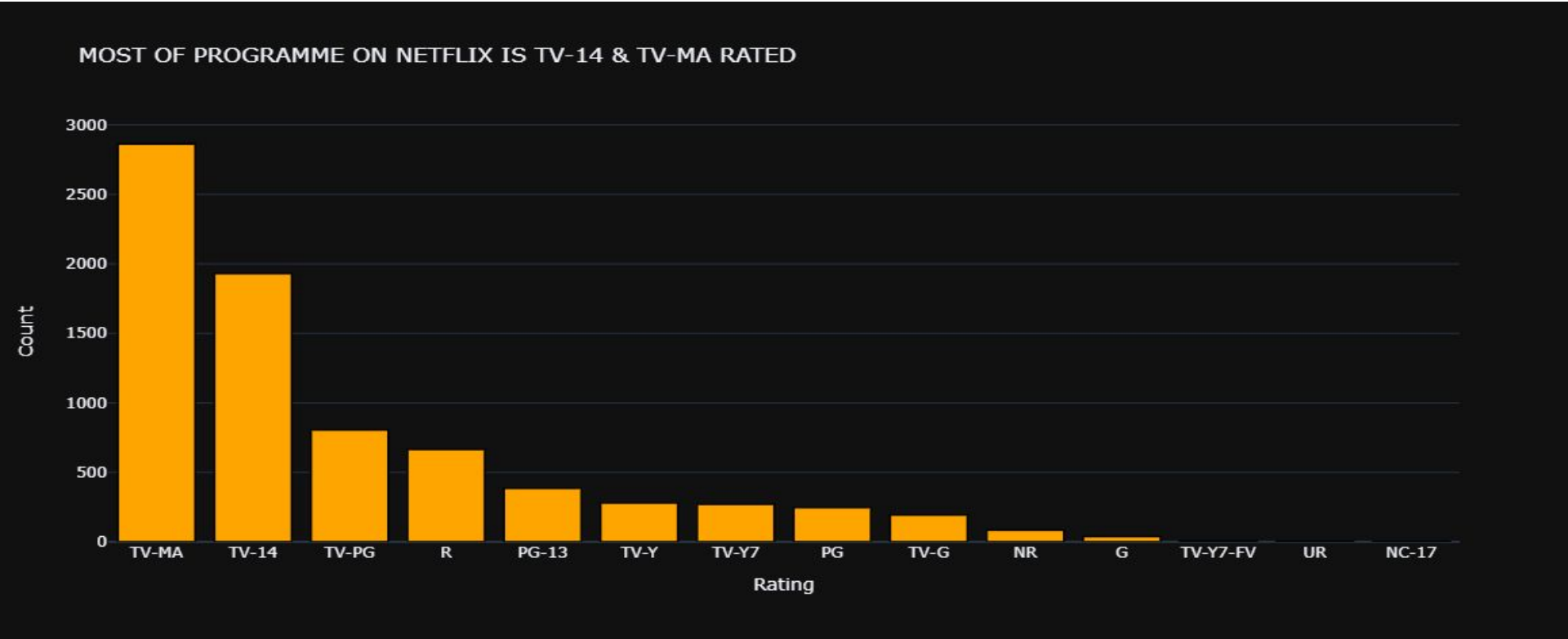
Releases over the year!



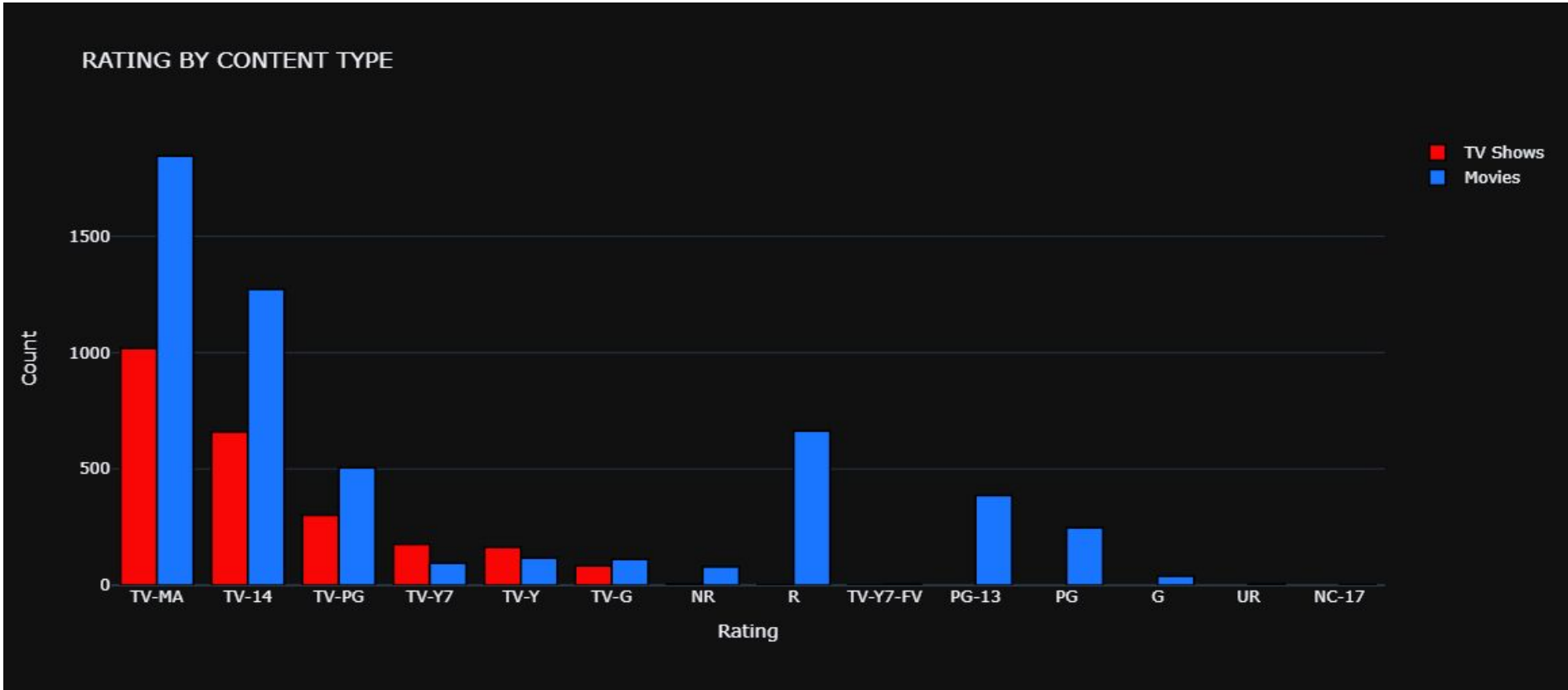
Split of content ratings..



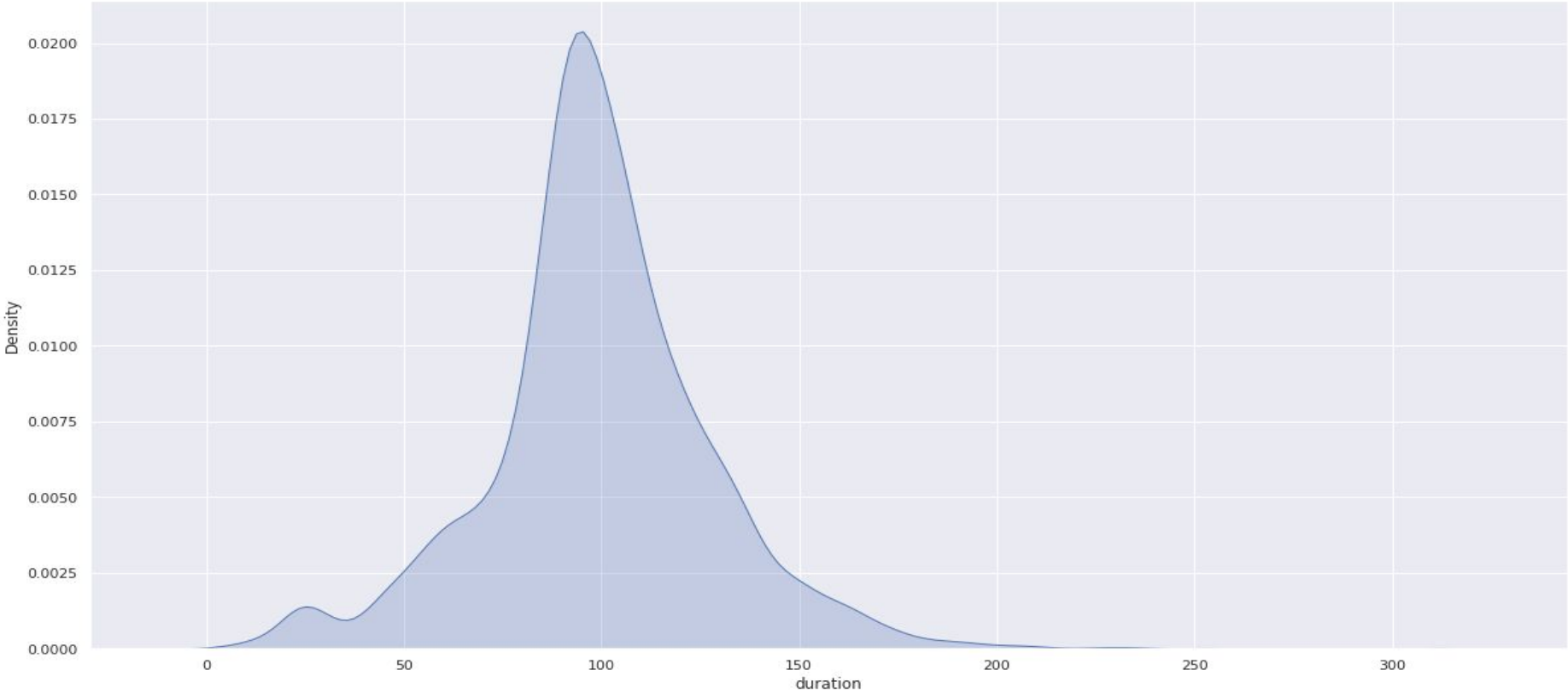
Split of content ratings..



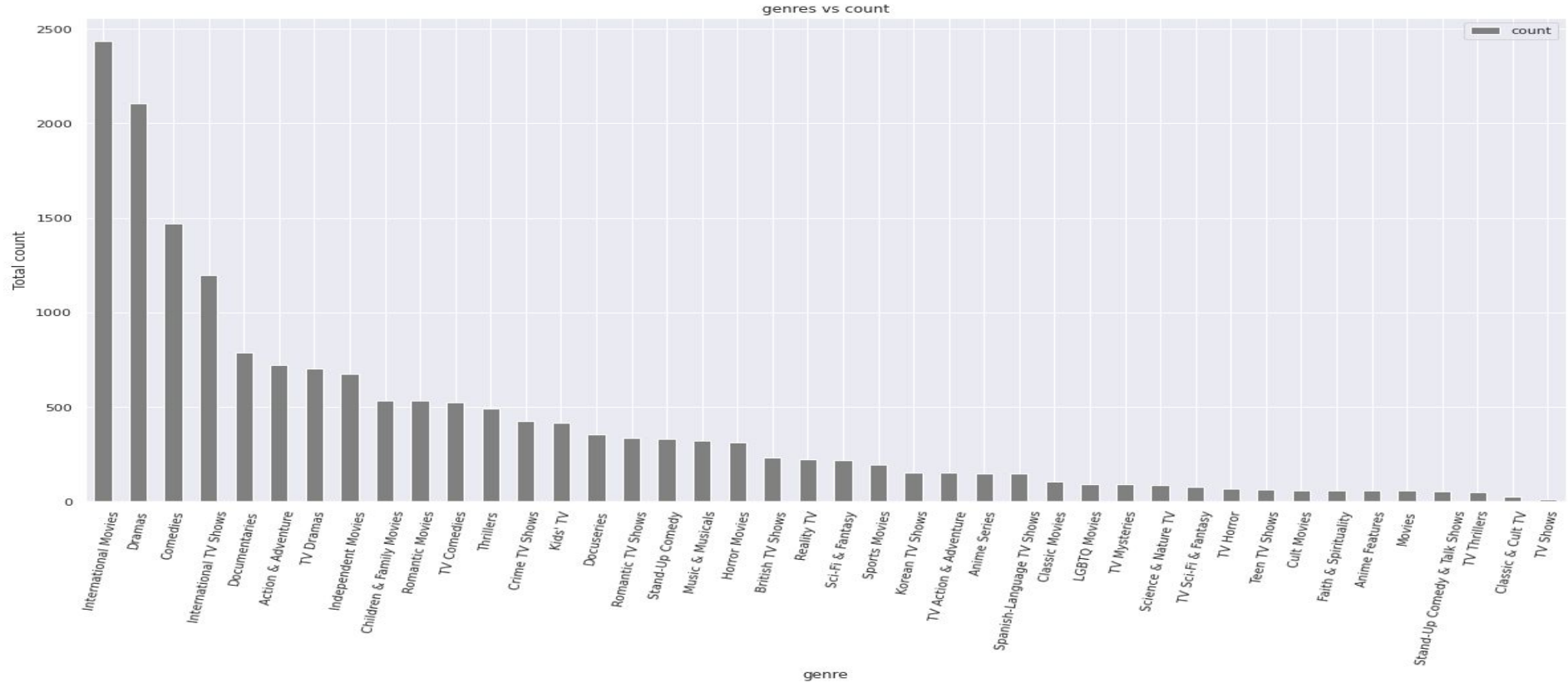
Split of content ratings..



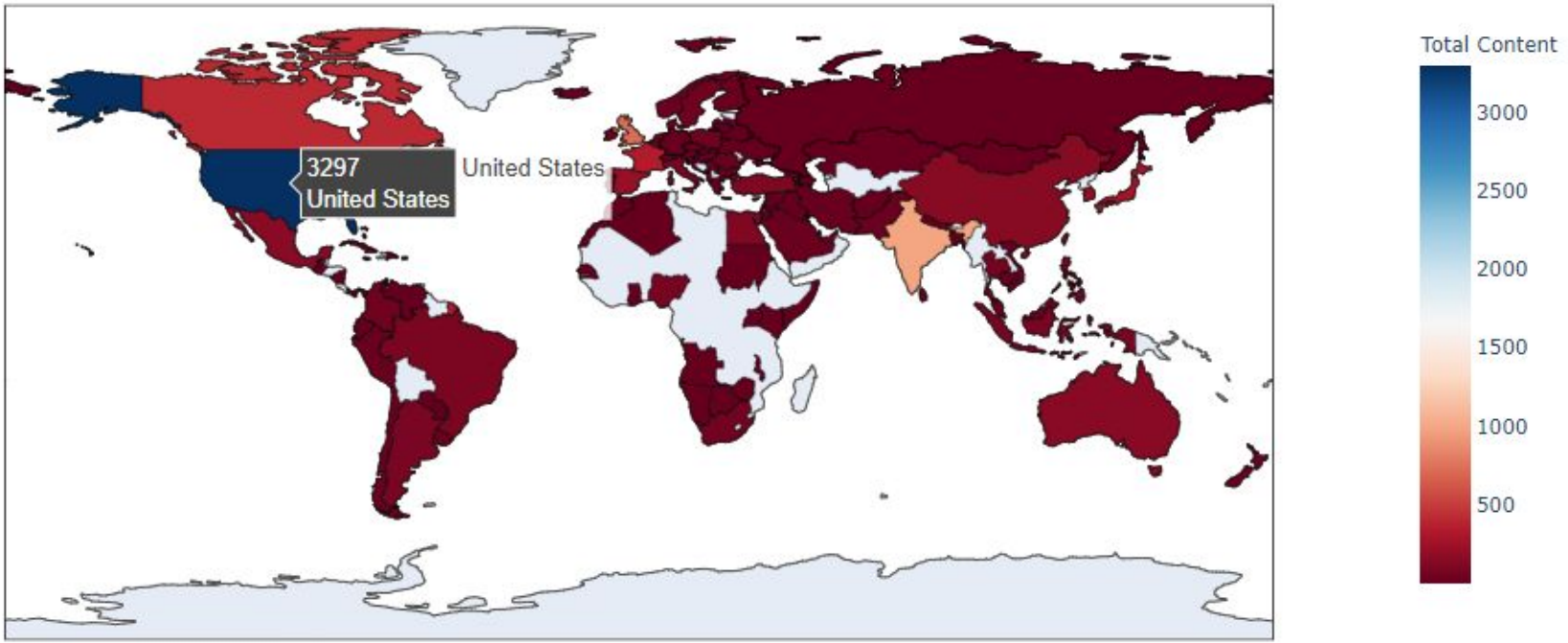
Duration distribution



Top Genre..!!

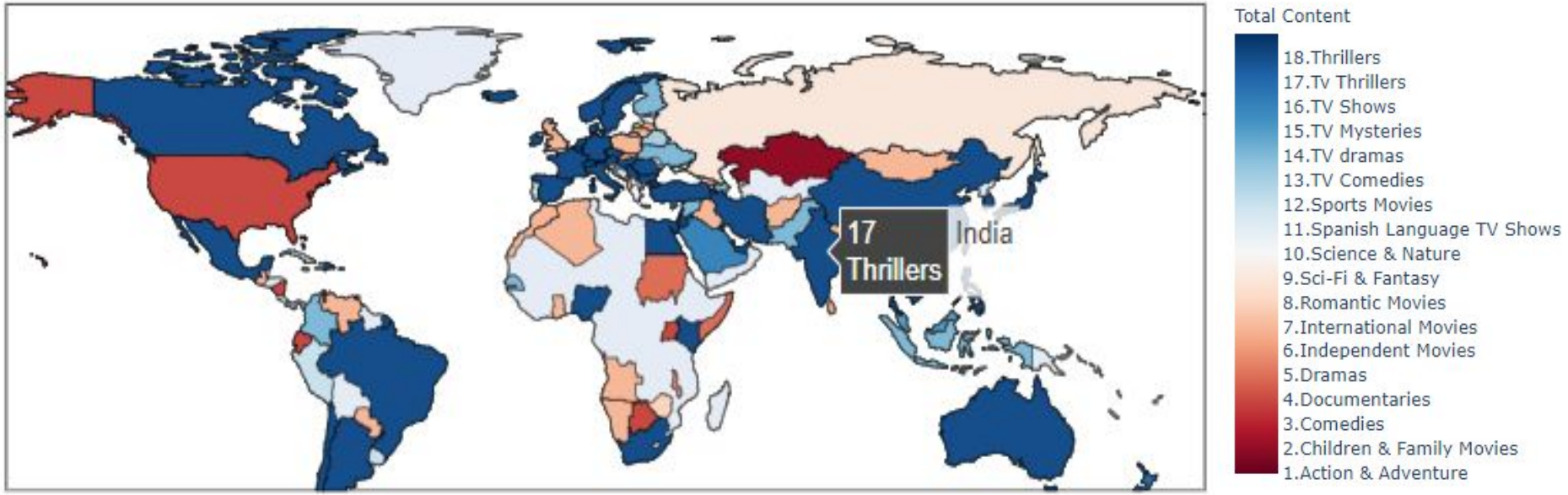

genres vs count
count
Total count
2500
2000
1500
1000
500
0
International Movies
Dramas
Comedies
International TV Shows
Documentaries
Action & Adventure
TV Dramas
Independent Movies
Children & Family Movies
Romantic Movies
TV Comedies
Thrillers
Crime TV Shows
Kids' TV
Docuseries
Romantic TV Shows
Stand-Up Comedy
Music & Musicals
Horror Movies
British TV Shows
Reality TV
Sci-Fi & Fantasy
Sports Movies
Korean TV Shows
TV Action & Adventure
Anime Series
Spanish-Language TV Shows
Classic Movies
LGBTQ Movies
TV Mysteries
Science & Nature TV
TV Sci-Fi & Fantasy
TV Horror
Teen TV Shows
Cult Movies
Faith & Spirituality
Anime Features
Movies
Stand-Up Comedy & Talk Shows
TV Thrillers
Classic & Cult TV
TV Shows
genre

Content produced by Countries



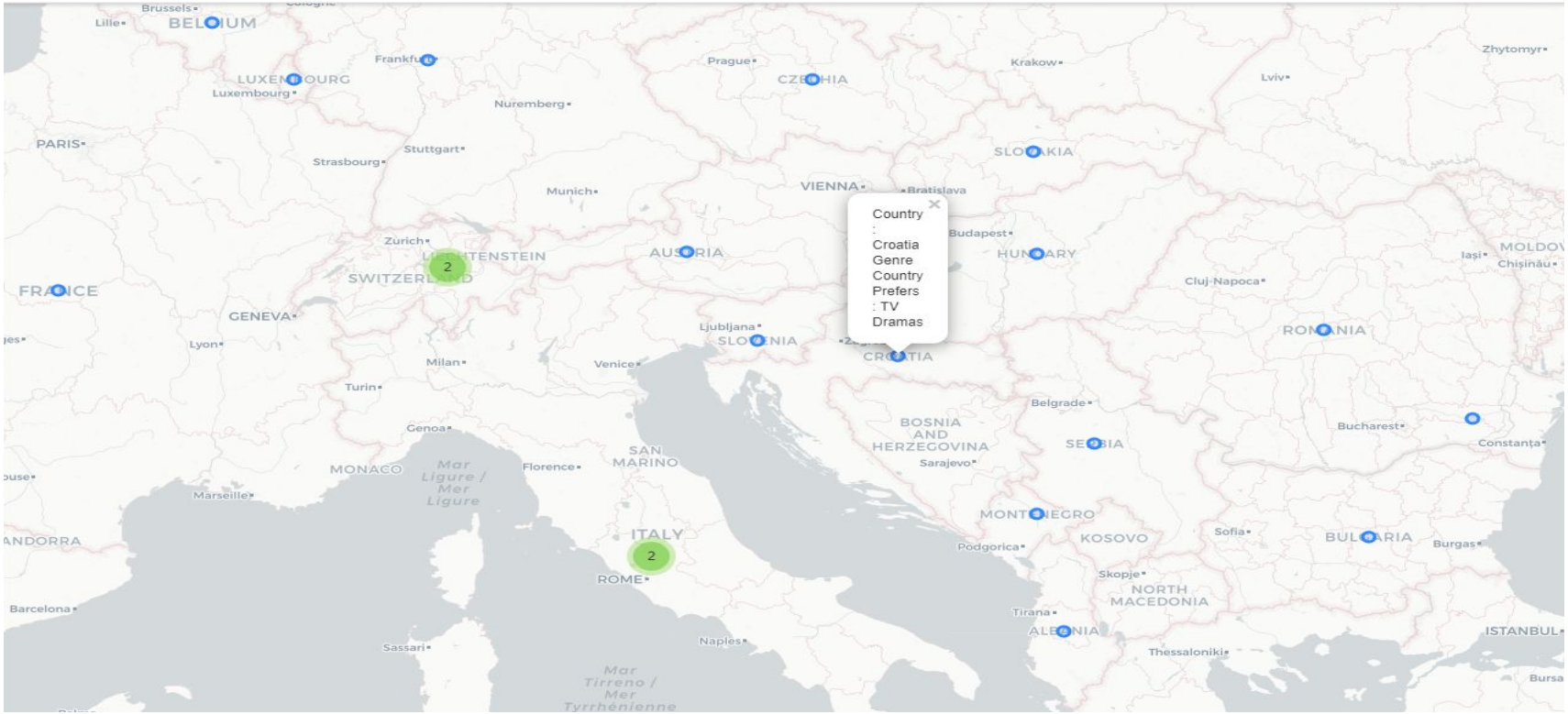
Number of content produced by different countries

Genre across different Countries

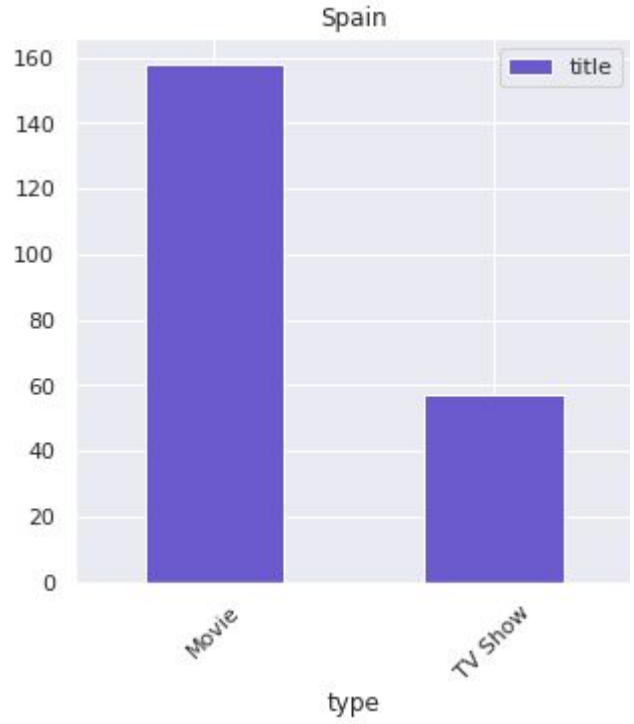


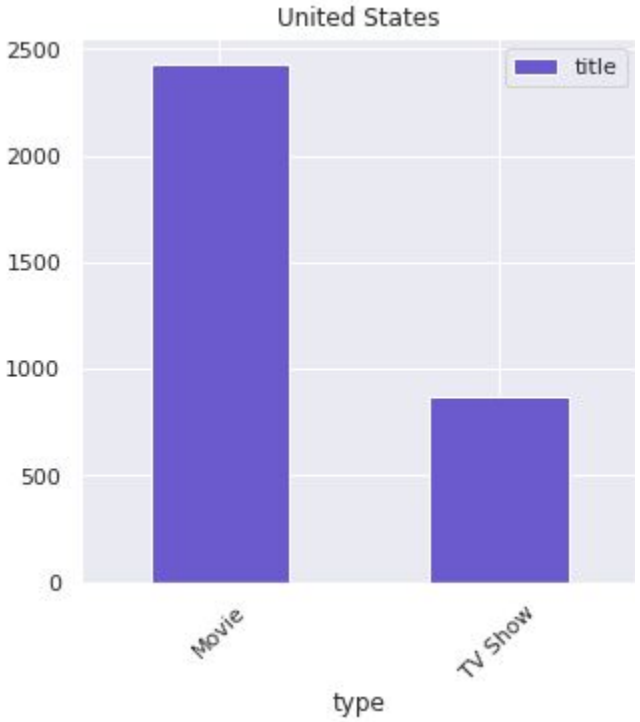
Genre preferred by different countries

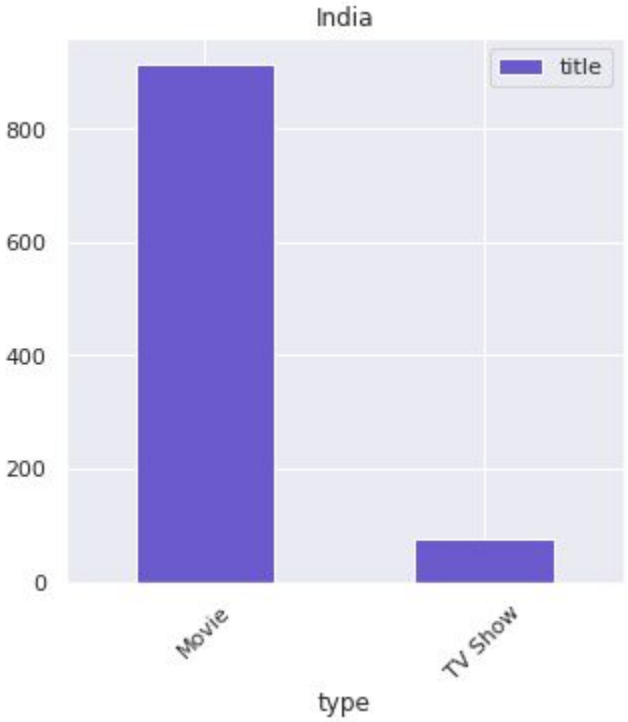
Genre across different Countries



Spain, US & India: Movies or Shows?

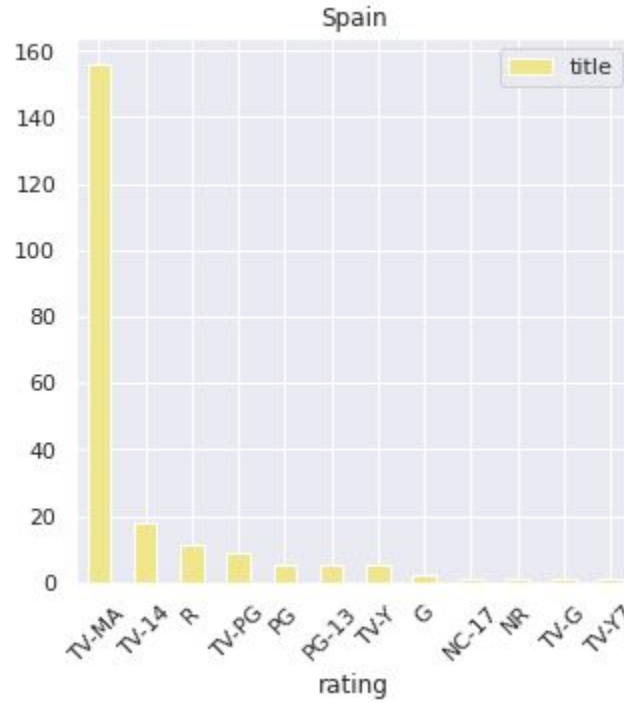


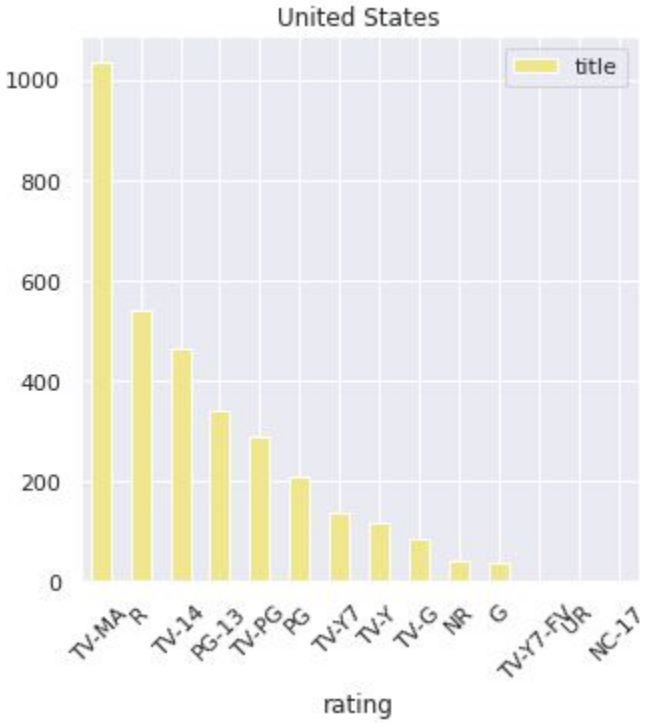




All the countries have preferred Movies over TV shows!!

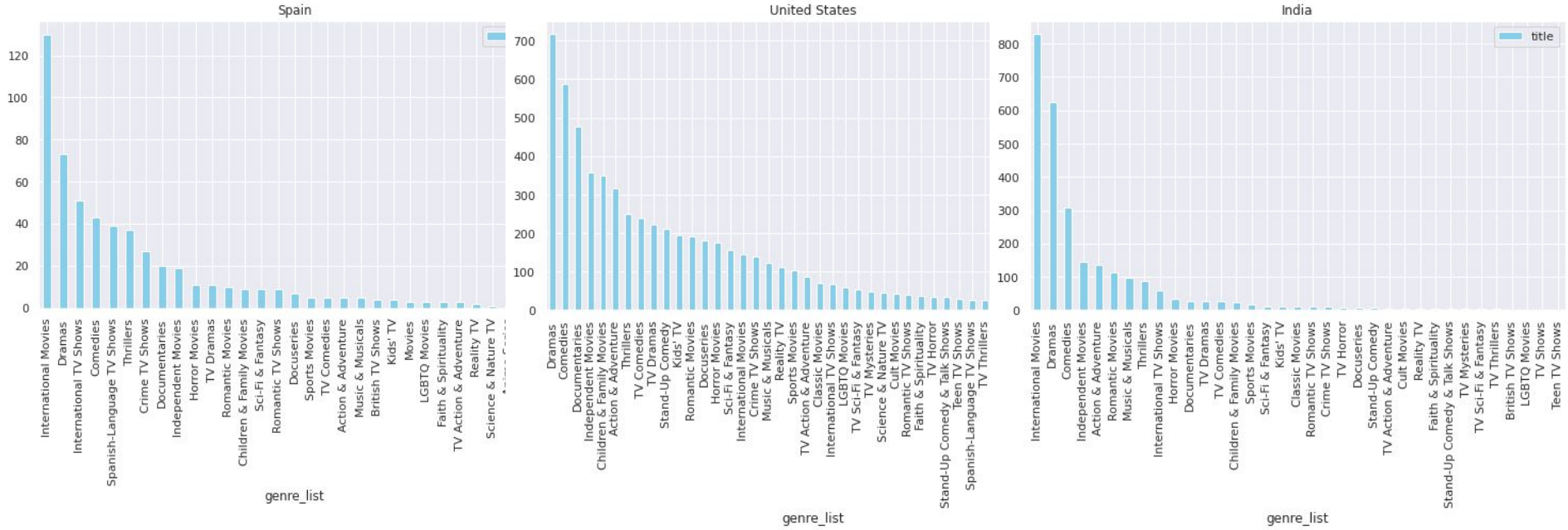
Spain, US & India: Rating





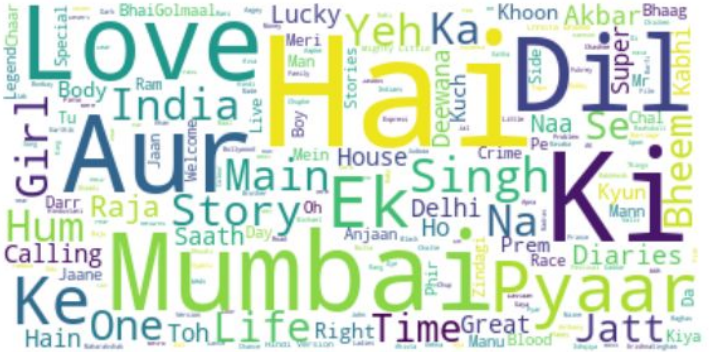
Spain and US prefers TV-Mature content, while India prefers TV -14 (unsuitable for children under the age of 14)

Spain, US & India: Genre

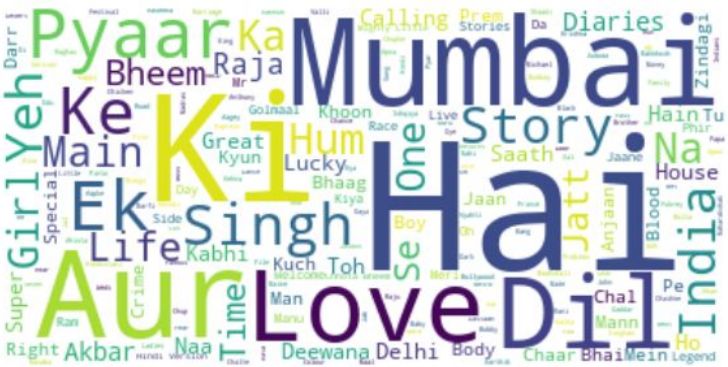


Spain & India has more International Movies, while US has more Drama Content.

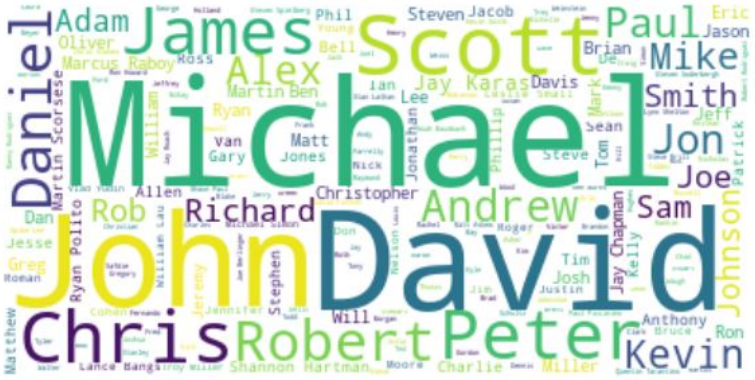
Spain, US & Indian: Titles!



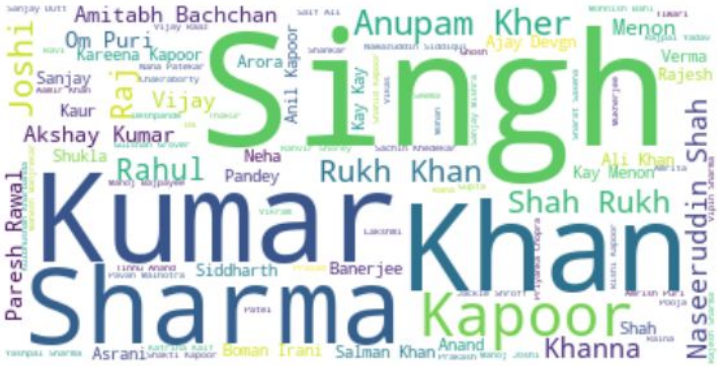
Spain, US & Indian: Description!

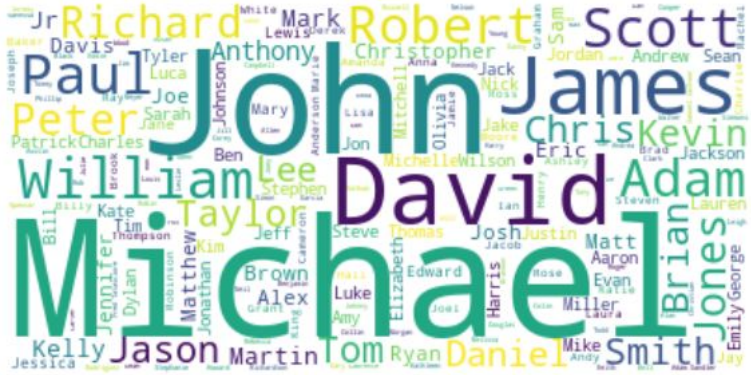


Which Director to choose?

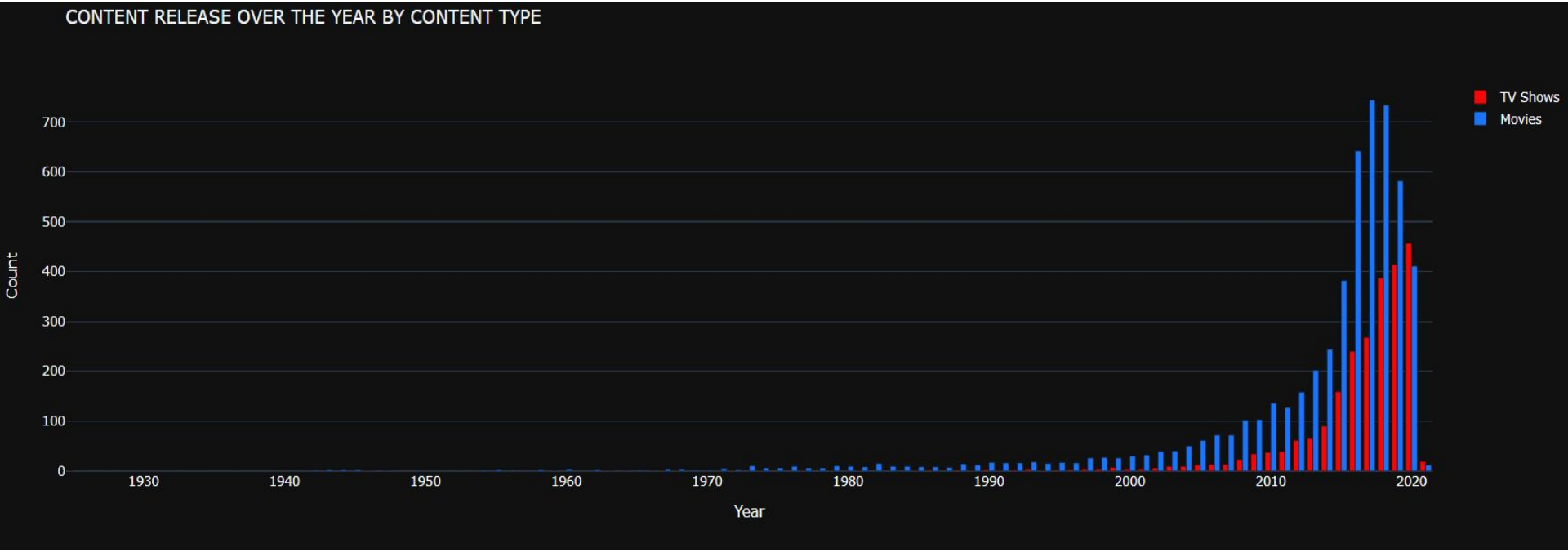


Actors preferred!





Is Netflix focusing more on TV Shows/Movies



Clustering dataset - Text Features

Text Preprocessing!

1. CLEANING

- Cleaned Null values
- All Columns: Only characters selected by regex
- All words to lowercase
- Merged text columns

2. STOPWORDS

- Removed Stop words
- Normal english words & problem specific

3. TOKENIZATION

- Splitted sentences to tokens
- Used word_tokenise from nltk

4. STEMMING

- Transformed words to roots
- Used Snowball Stemmer

Everybody stand back, I know regex expressions!

Time to Cluster..

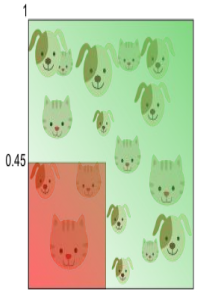
Vectorization → **Dimensionality Reduction** → **Finding Optimal K** → **K - Means Clustering with Optimal k**

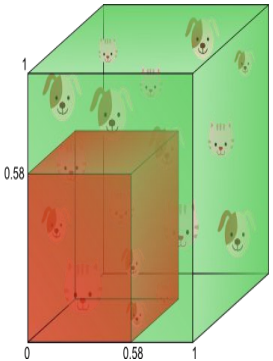
TFIDF Vectorizer

$$\text{tf-idf} = \text{tf} \times \text{idf} \quad (1)$$

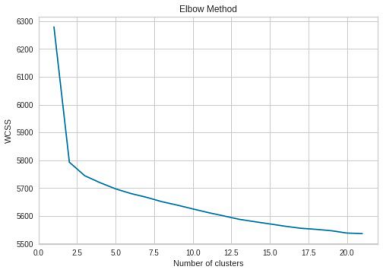
$$\text{idf}(t) = \log \frac{n+1}{\text{df}(d,t)+1} + 1 \quad (2)$$

PCA



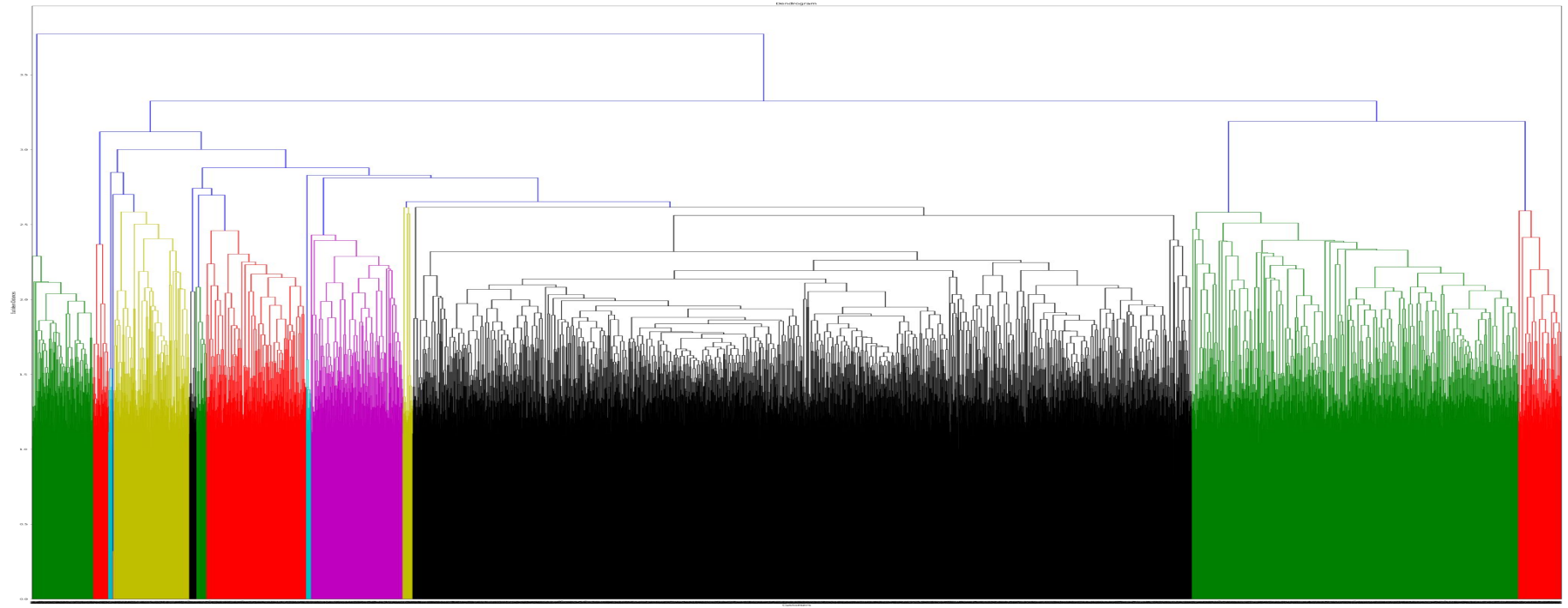


Elbow method: WCSS, Silhouette score

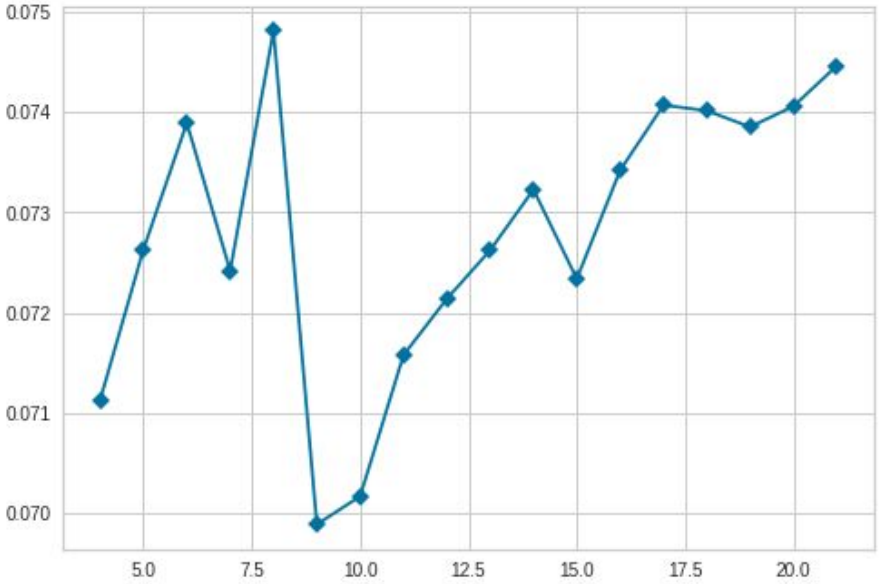


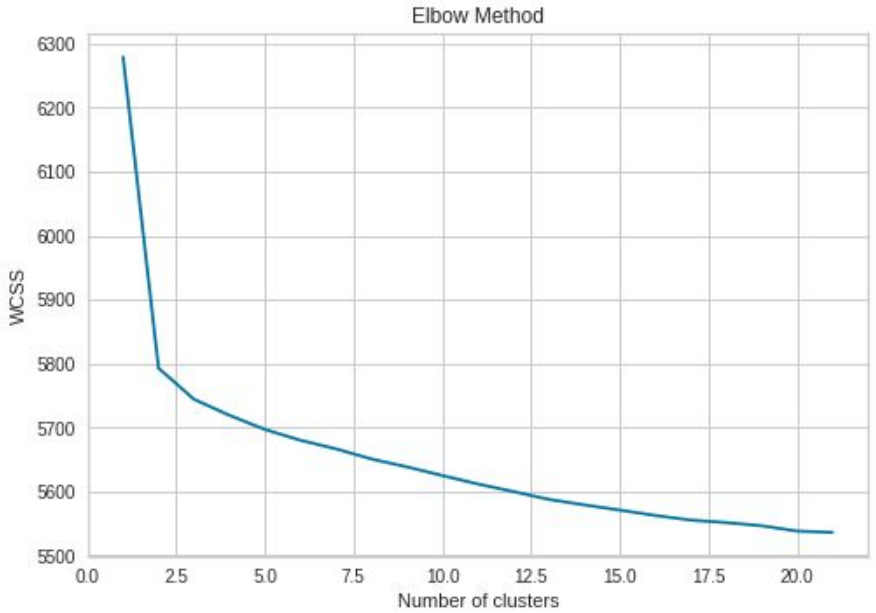
What could we infer?

Dendogram..

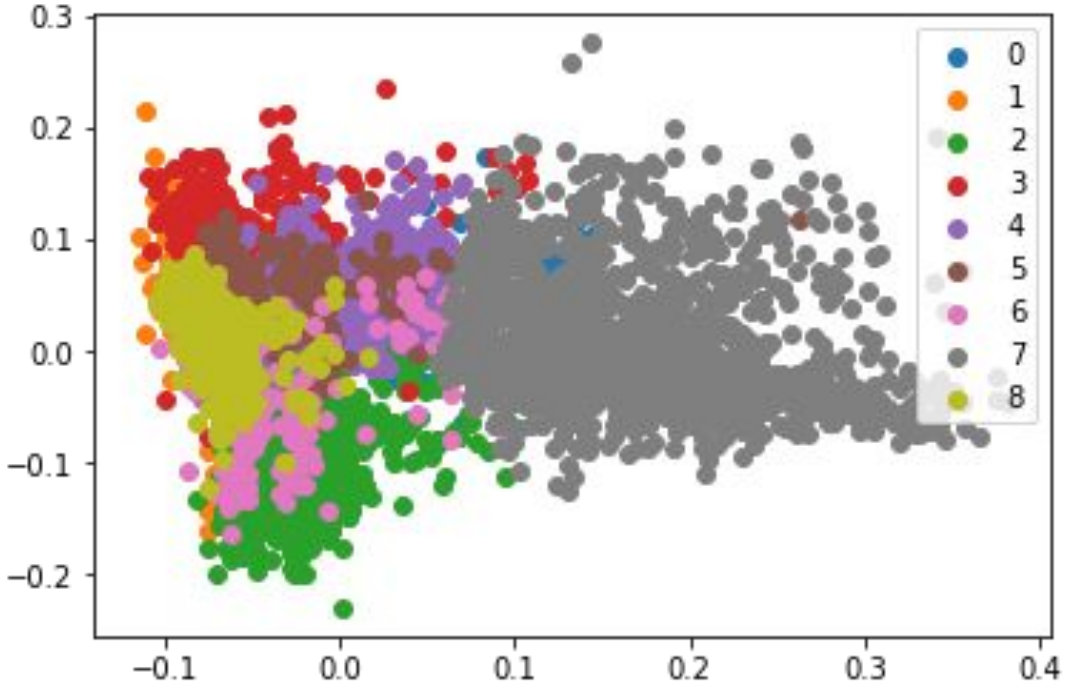


Finding Optimal K!





Visualizing Clusters

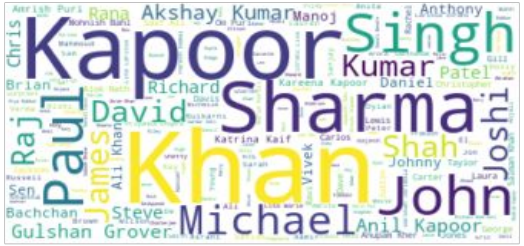


Naming the clusters

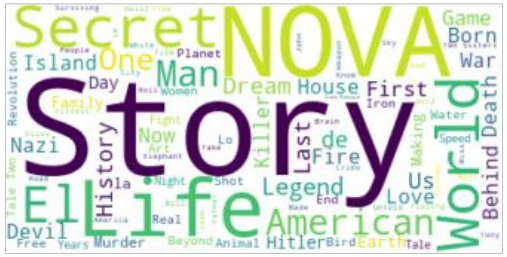
- 0: "Kids, Animation, Anime"
- 1: "Musical International, Indian"
- 2: "Drama, International, Indian"
- 3: "Documentaries, Sports"
- 4: "Drama, American, Adventure"
- 5: "Comedy"
- 6: "Horror"
- 7: "International TV Shows"
- 8: "Family Movies"

Clusters: Description

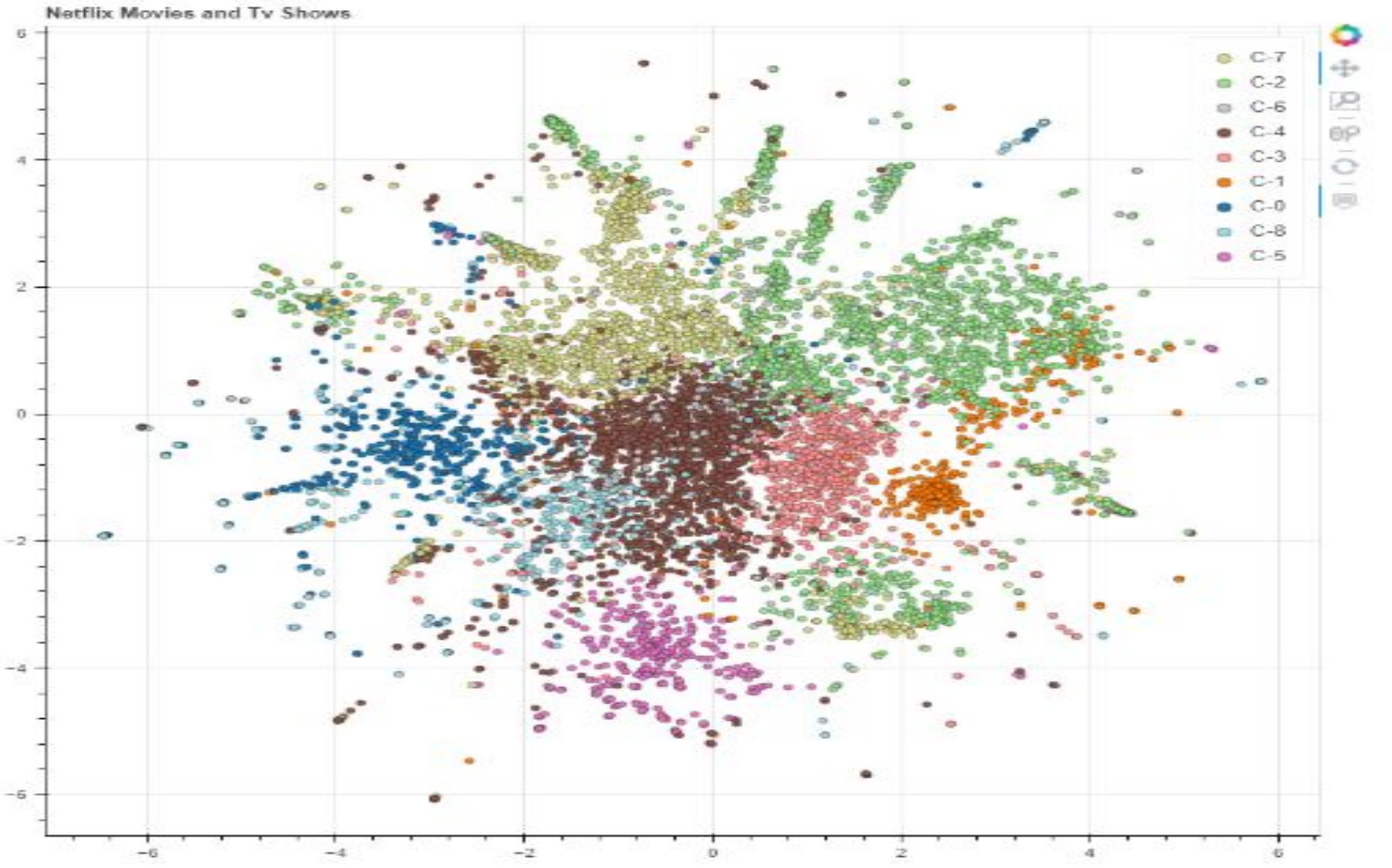
Cluster: Cast

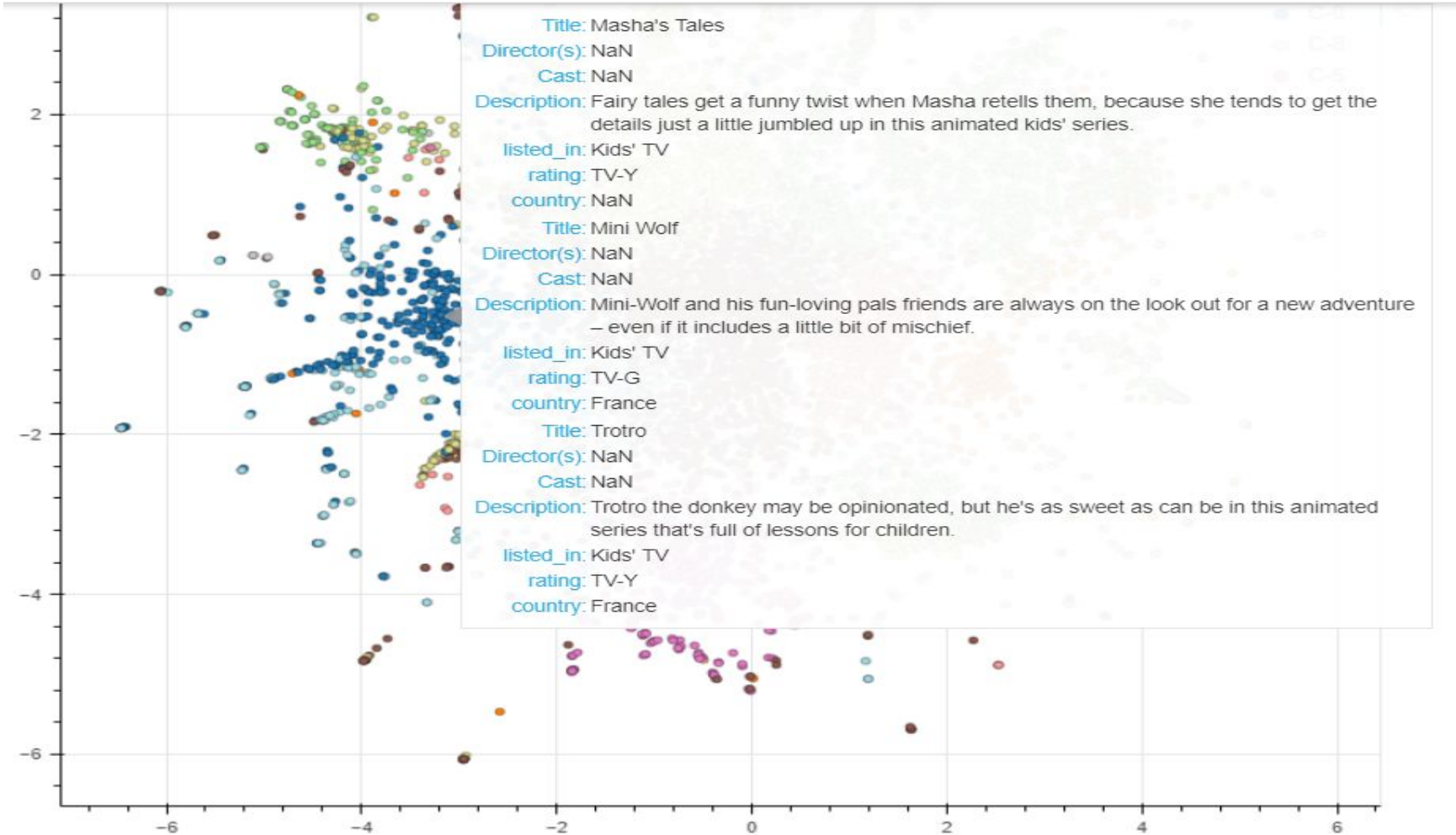


Cluster: Titles!



Find similar movies / tv shows in corresponding Cluster




Title: Masha's Tales
Director(s): NaN
Cast: NaN
Description: Fairy tales get a funny twist when Masha retells them, because she tends to get the details just a little jumbled up in this animated kids' series.
listed_in: Kids' TV
rating: TV-Y
country: NaN
Title: Mini Wolf
Director(s): NaN
Cast: NaN
Description: Mini-Wolf and his fun-loving pals friends are always on the look out for a new adventure – even if it includes a little bit of mischief.
listed_in: Kids' TV
rating: TV-G
country: France
Title: Trotro
Director(s): NaN
Cast: NaN
Description: Trotro the donkey may be opinionated, but he's as sweet as can be in this animated series that's full of lessons for children.
listed_in: Kids' TV
rating: TV-Y
country: France
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6

Recommendation System for Netflix..!!

Chosen Movie/TV Show
Behind Enemy Lines: After dire setbacks in 1940, Winston Churchill commissions a new kind of fighting force: commandos trained to

Top Recommendations
Thunderbolt: A P-47 Thunderbolt squadron is shown in preparation, at play and in bombing raids aiming to halt Nazi supply lines a

A Bridge Too Far: This wartime drama details a pivotal day in 1944 when an Allied task force tried to win World War II by seizing

The Outpost: A group of vastly outnumbered U.S. soldiers at a remote Afghanistan base must fend off a brutal offensive by Taliban

The Siege of Jadotville: Besieged by overwhelming enemy forces, Irish soldiers on a U.N. peacekeeping mission in Africa valiantly

Mission of Honor: As Hitler's Nazis threaten to take command of Britain's skies, a squadron of Polish pilots arrives to aid the R

If only we had more time: Future Scope!

- **Integrate Netflix dataset with other data set and present more insights and clusters.**
- **We could have done some research on recommendation system.**

Conclusion

- Performed **EDA**
- Univariate & multivariate **analysis**
- Visualised Data, inferred **insights**
- **Analysed various trends in Countries**
- **TV Shows or Movies?**
- **Text Based Clustering**
- Identified 9 distinct clusters
- Experimented Interacted Visualisations
- Recommendation System

Suggestions

"Torture the data, and it will confess to anything."

-Ronald Coase, Nobel Prize winner

Time for Q&A!!