

# Capstone Project - 2

## NETFLIX MOVIES AND TV SHOWS CLUSTERING

**Manas Ranjan  
Behera**

## Emmy-winning US TV Shows




## Police Detective TV Dramas



## Critically Acclaimed Witty TV Shows



Dataframe from our Netflix Dataset												
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...

# 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.

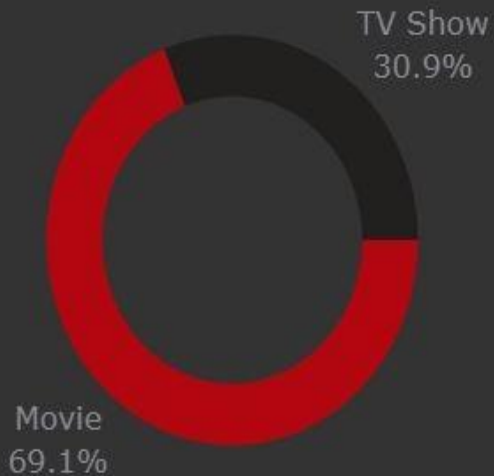
# Closer look at the dataset..!!

- 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

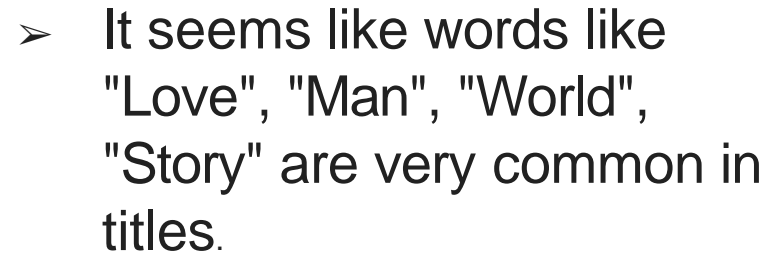
# Exploratory Data Analysis



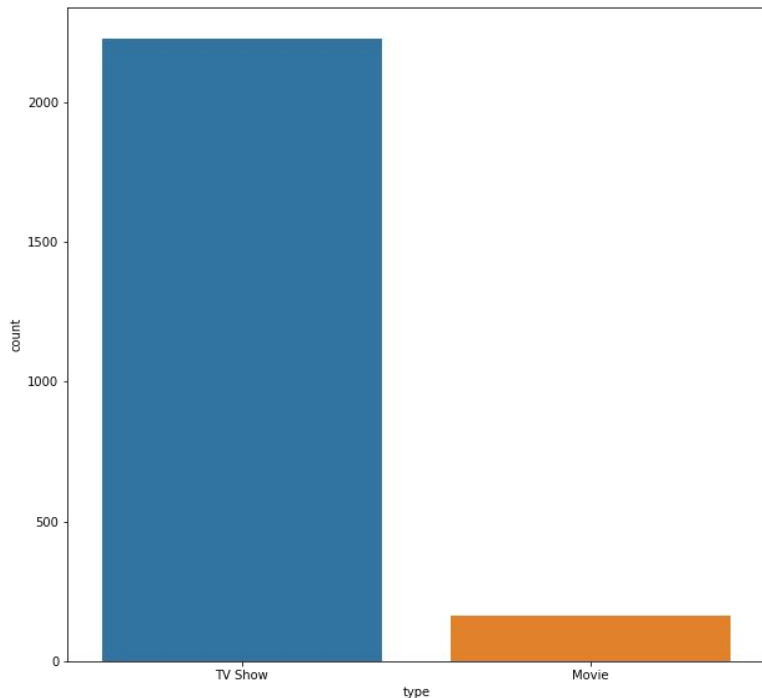
# TV or Movie Shows?



➤ There are 5377 movies and 2410 tv shows.

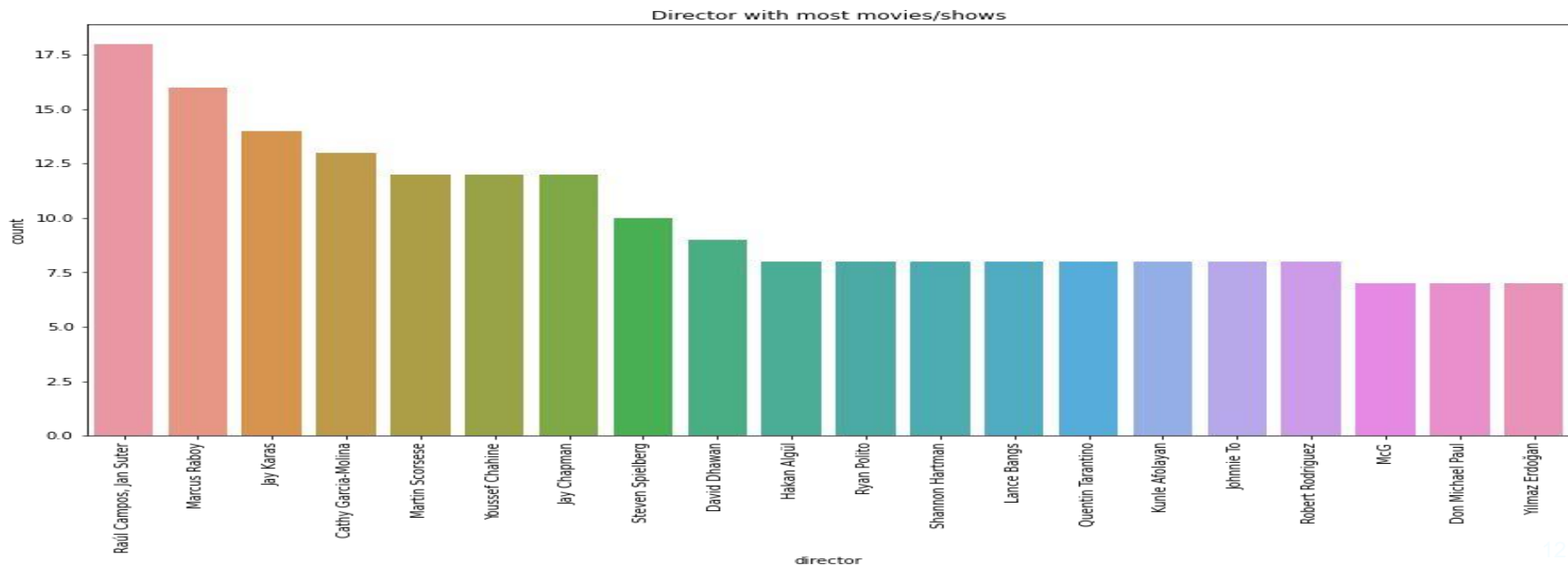


# EDA (Continued..)

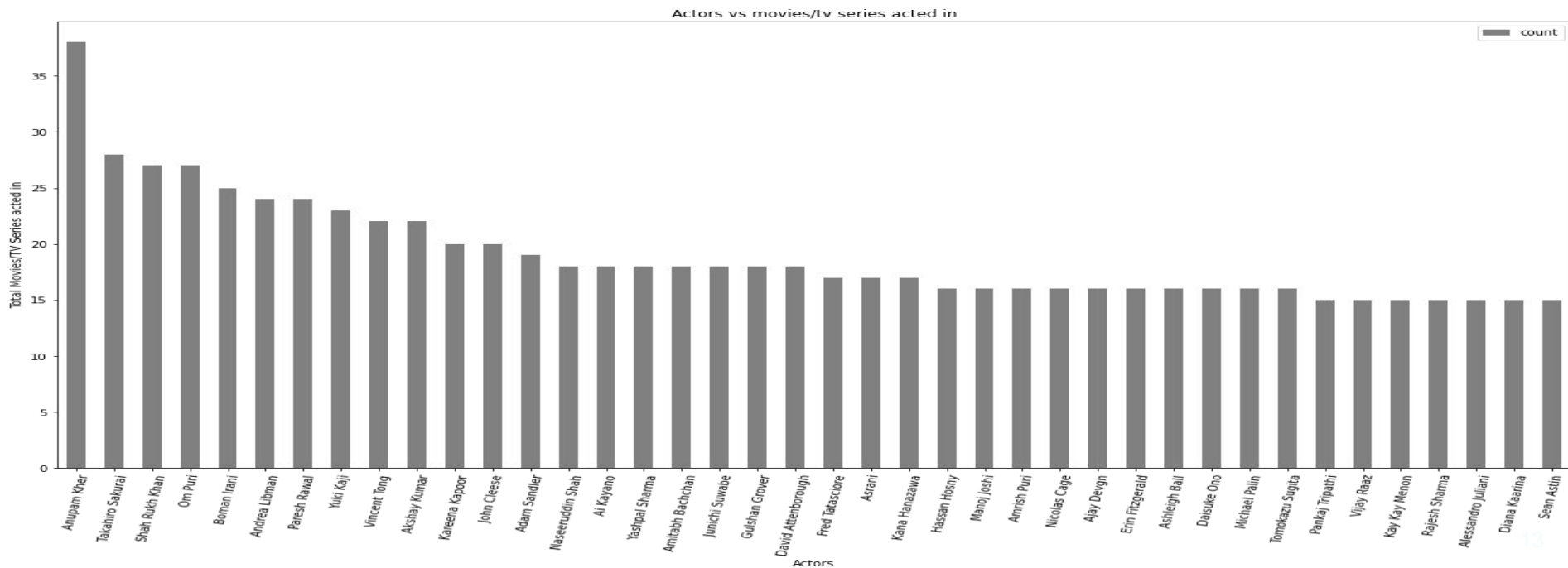


- It looks like most of the missing values of Director is for tv shows.

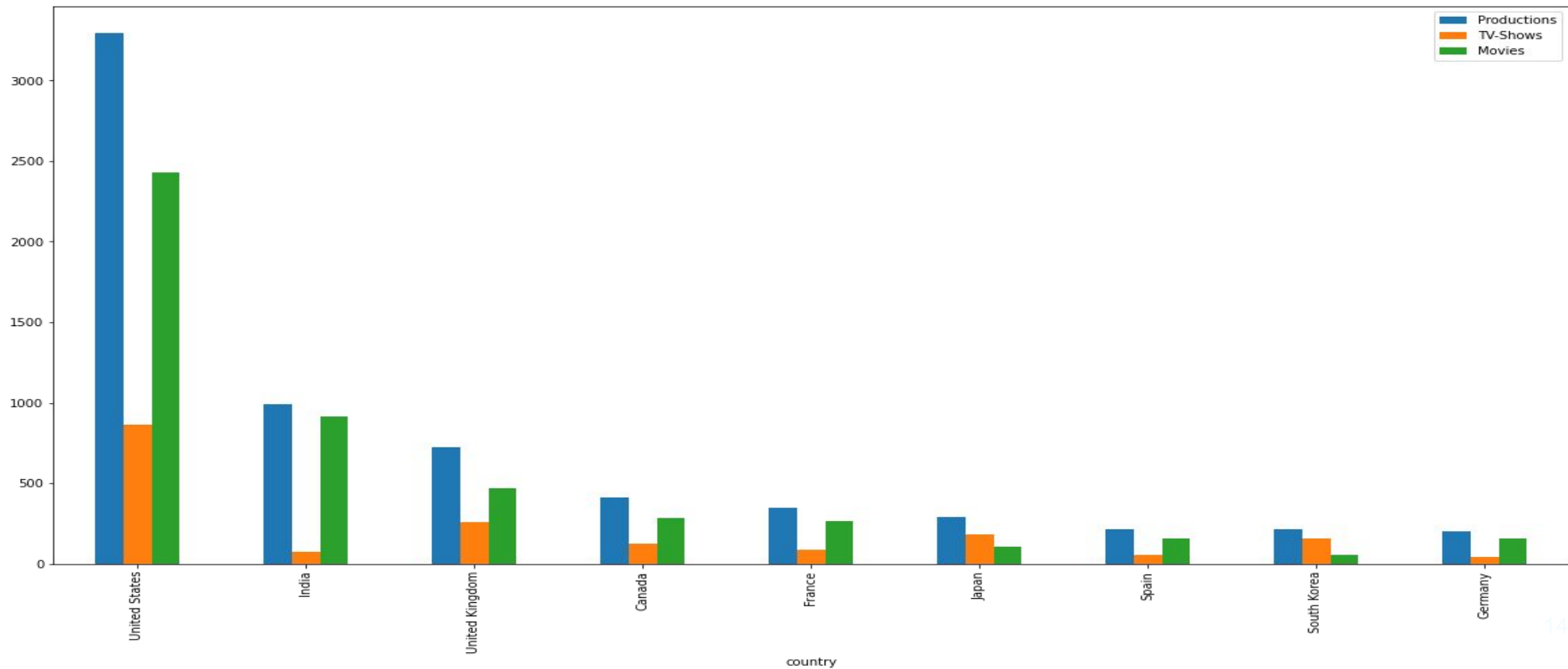
# EDA (Continued..)



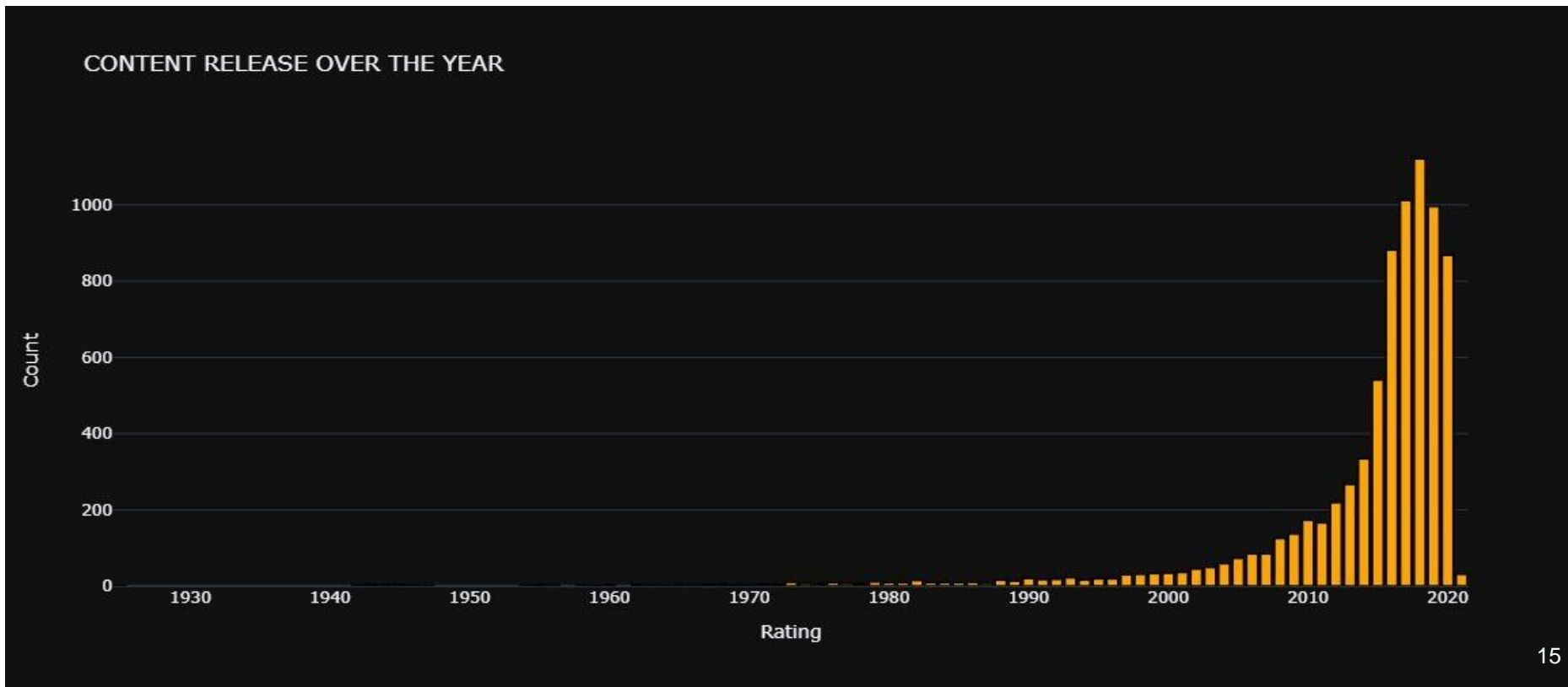
# EDA (Continued..)



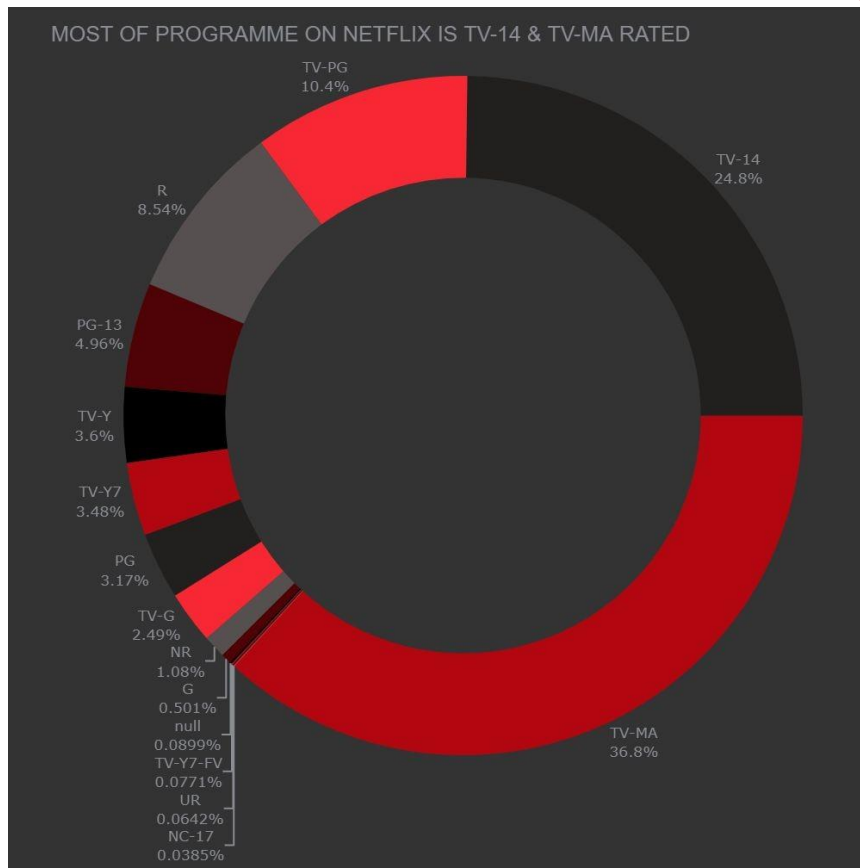
# Production split across top countries



# Releases over the year!



# Split of content ratings..



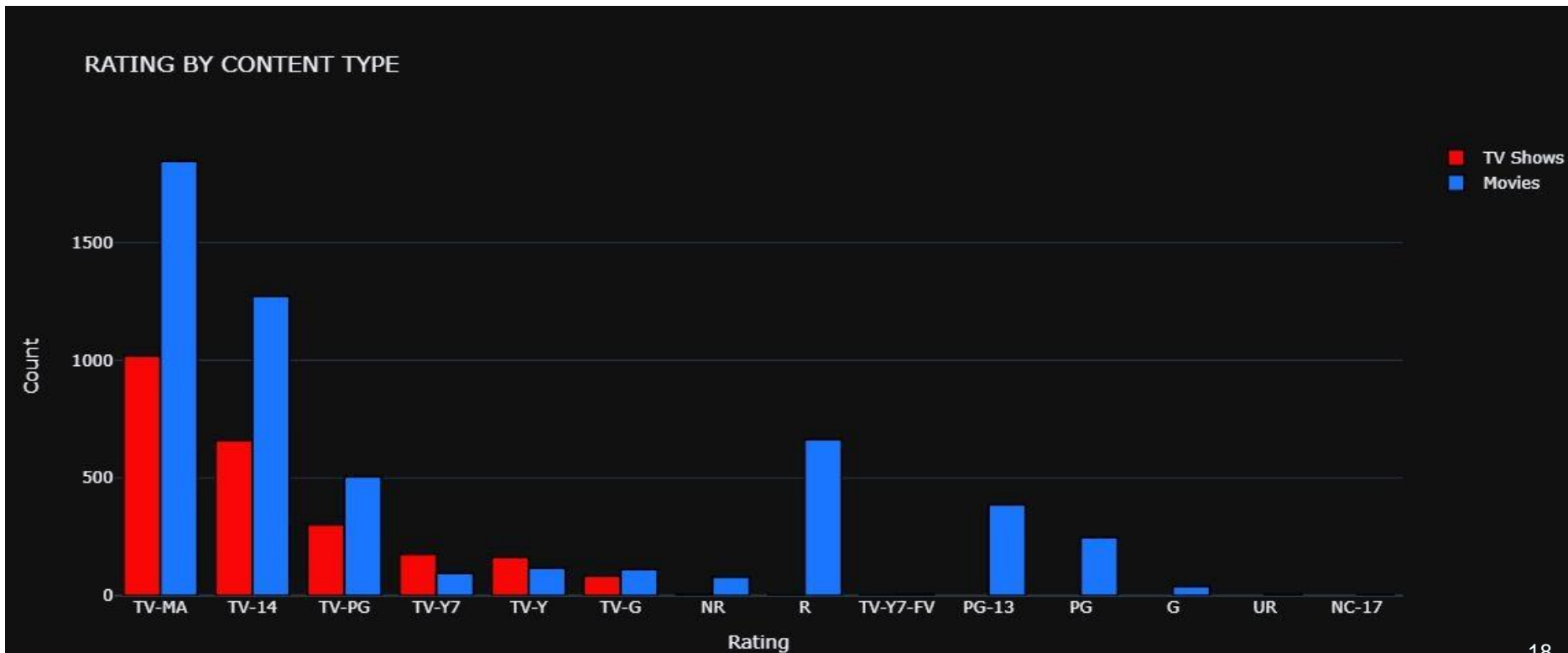


# Split of content ratings..

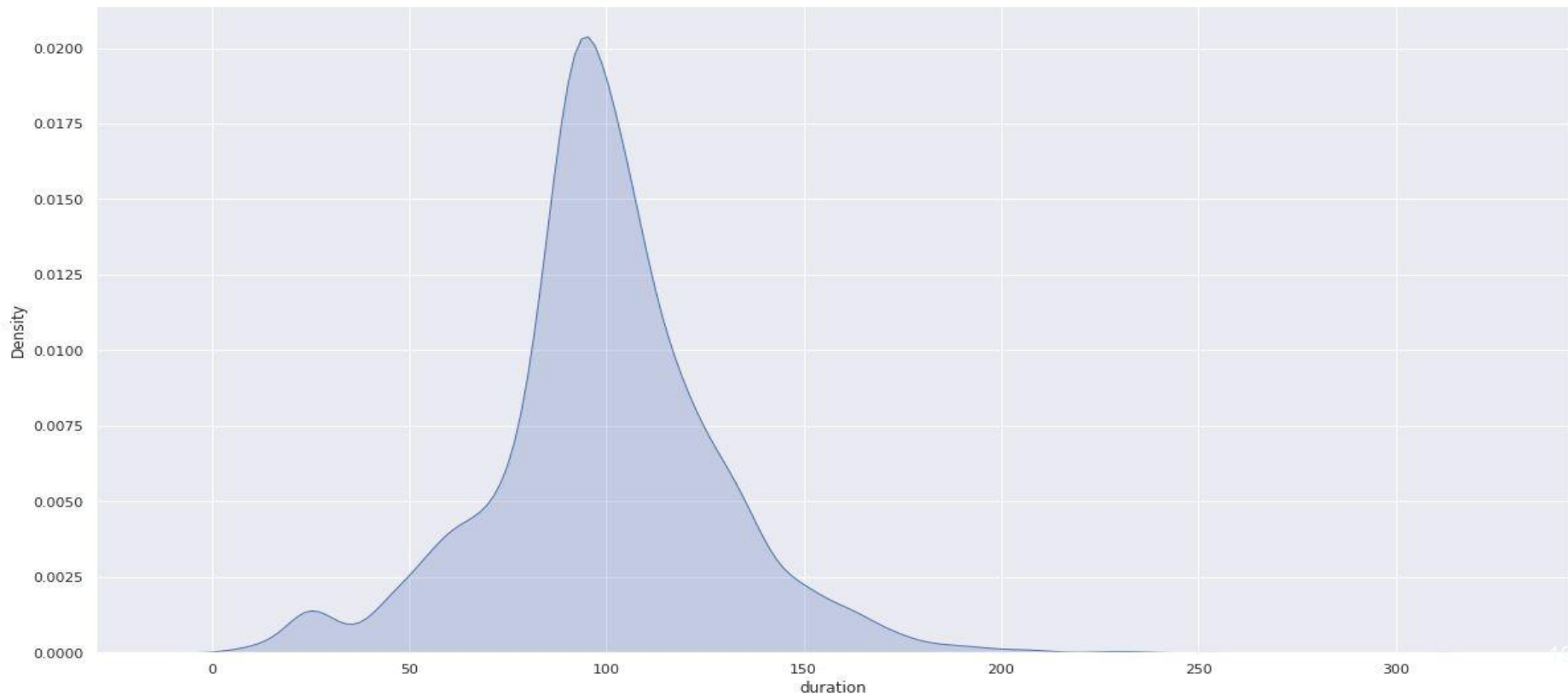
MOST OF PROGRAMME ON NETFLIX IS TV-14 & TV-MA RATED



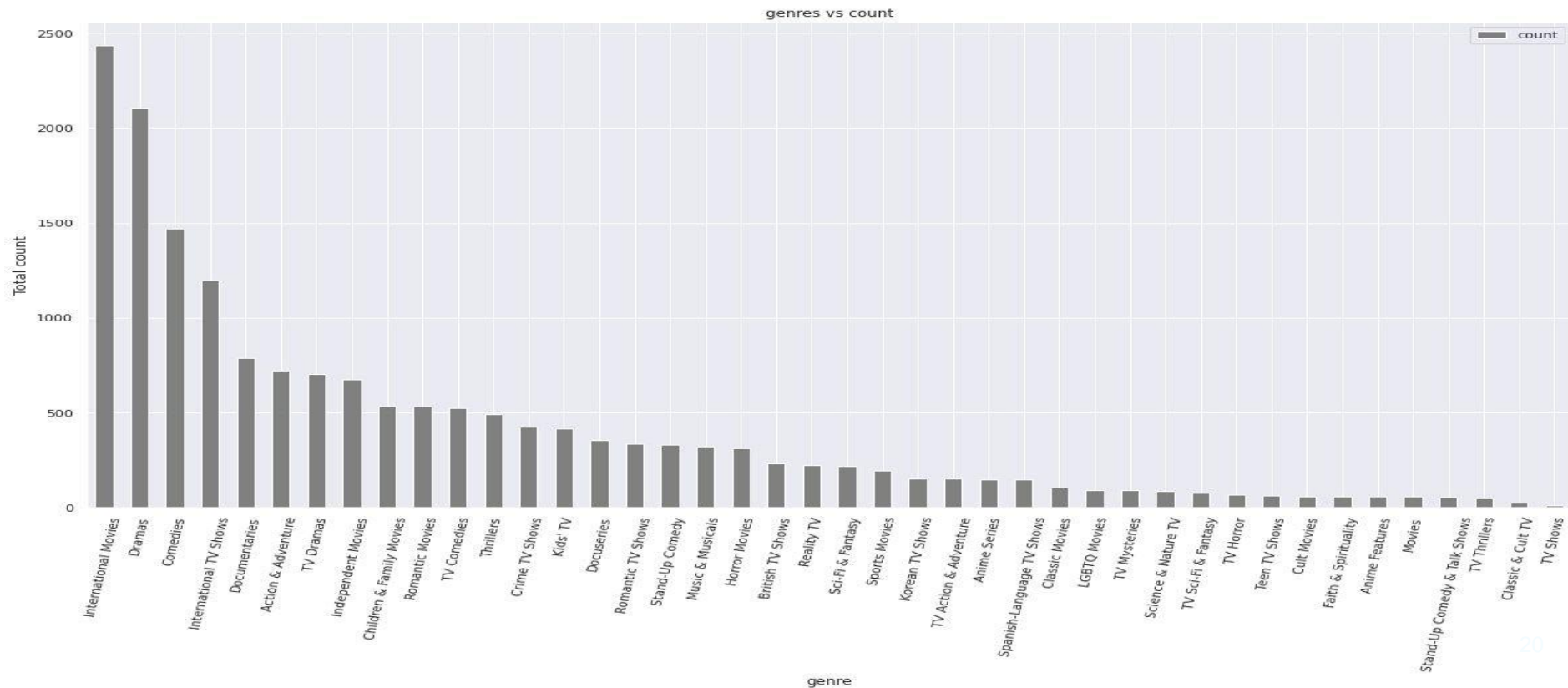
# Split of content ratings..



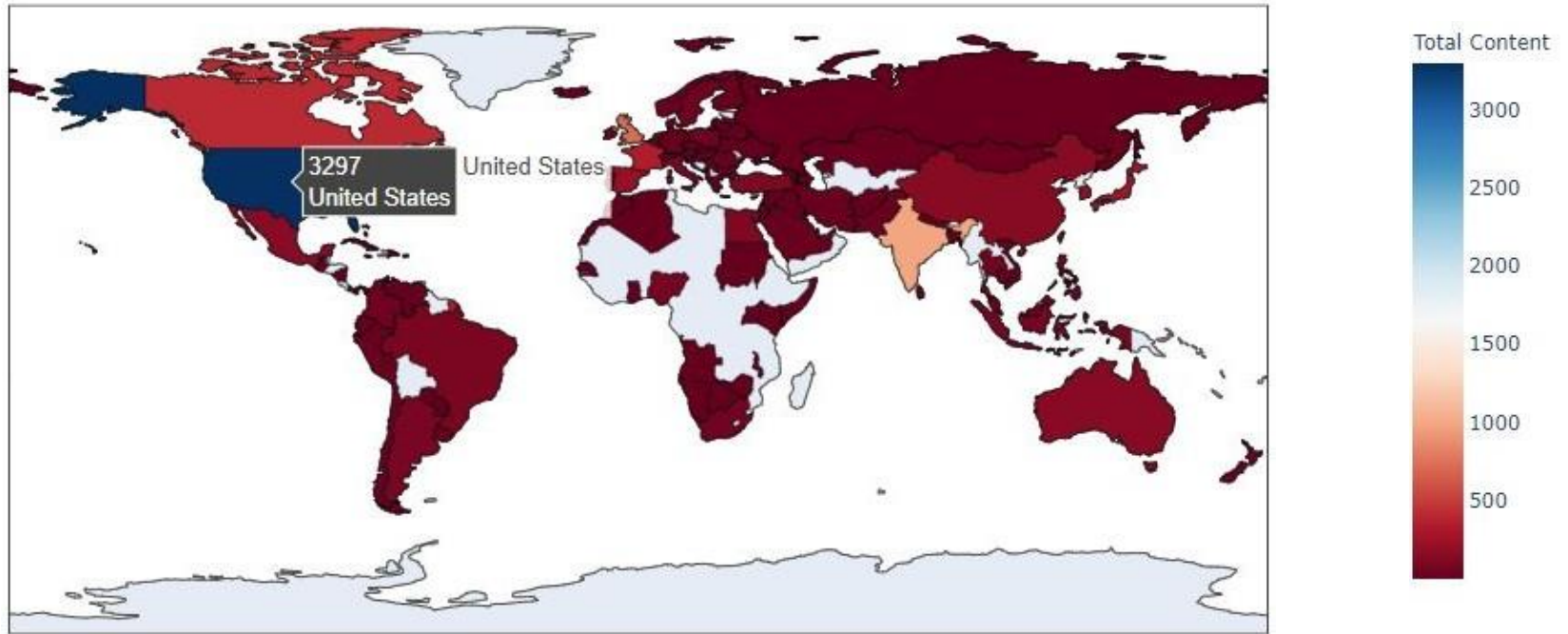
# Duration distribution of Movies



# Top Genre..!!



# Content produced by Countries AI



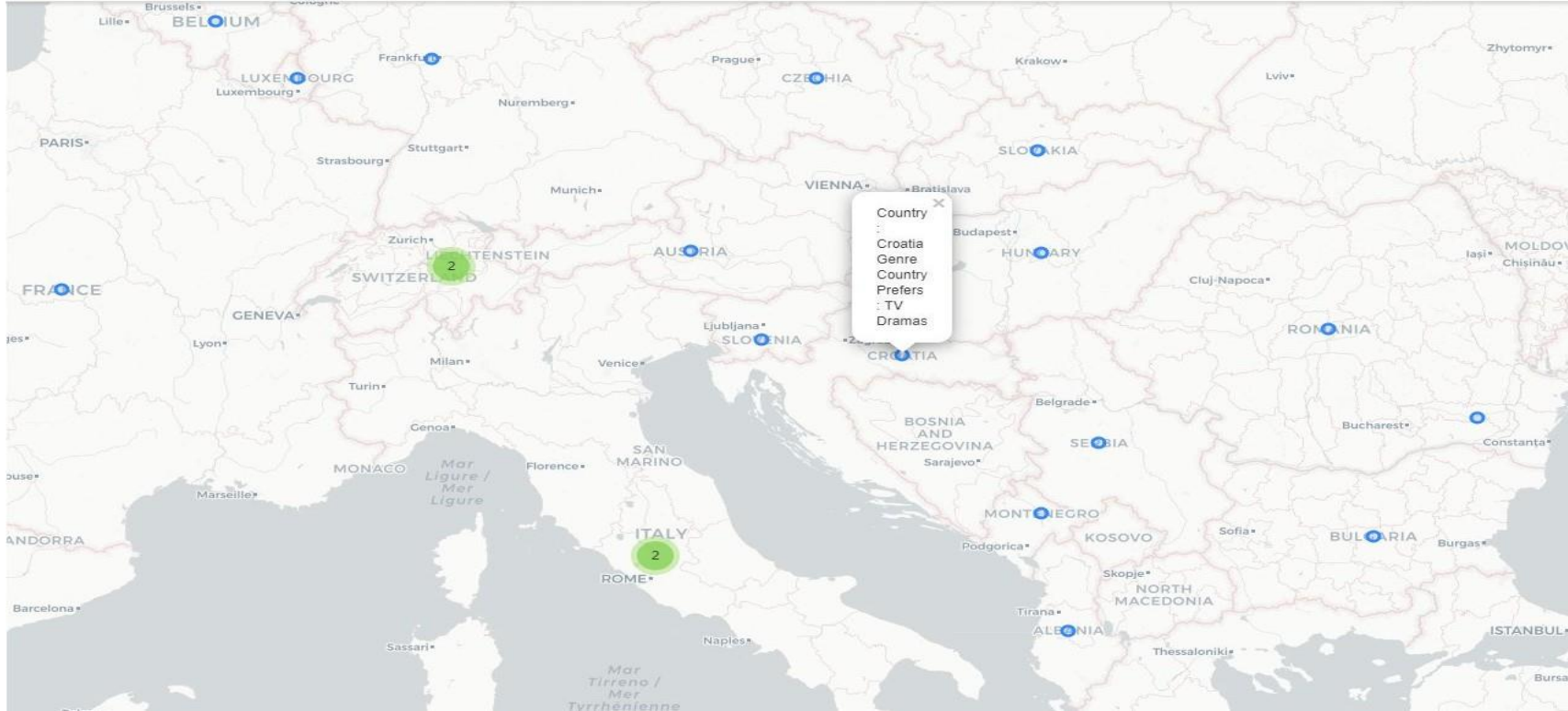
Number of content produced by different countries

# Genre across different Countries AI

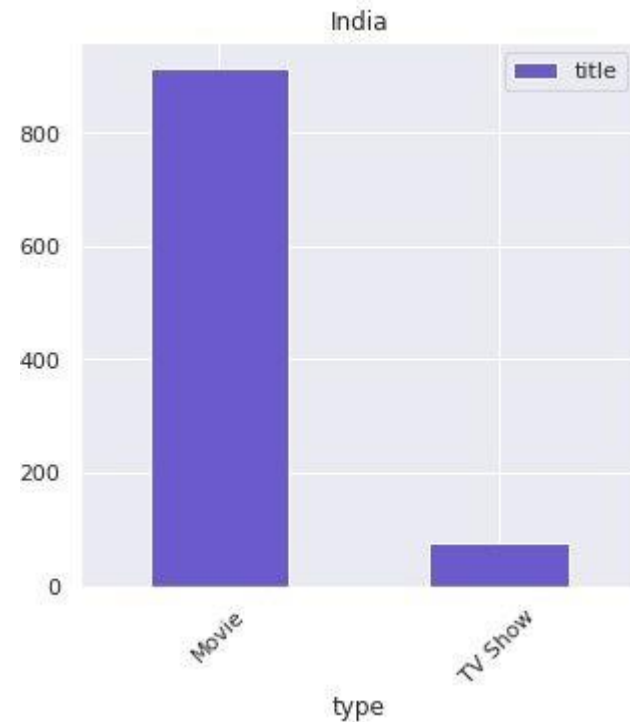
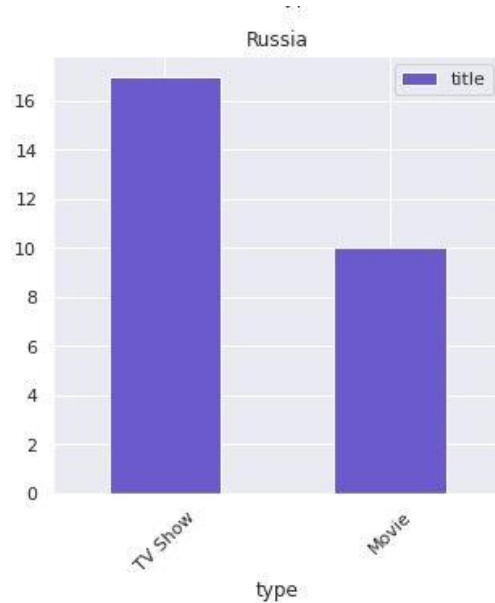
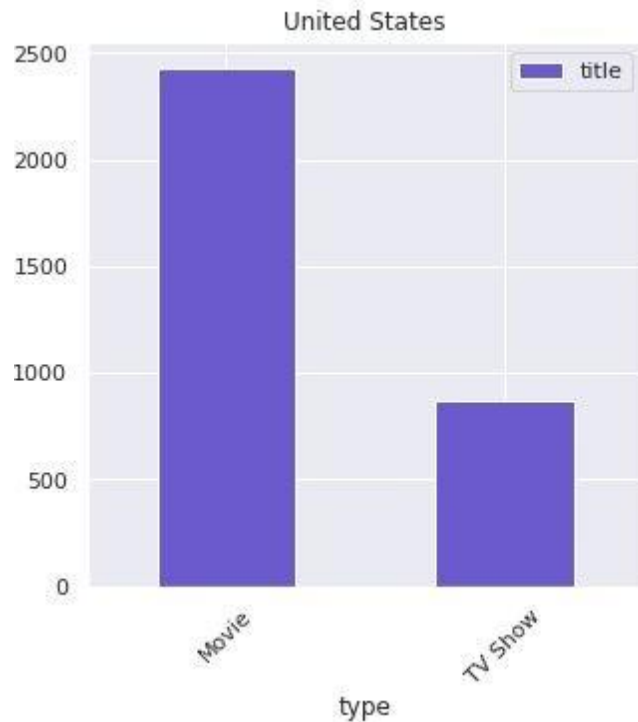


Genre preferred by different countries

# Genre across different Countries



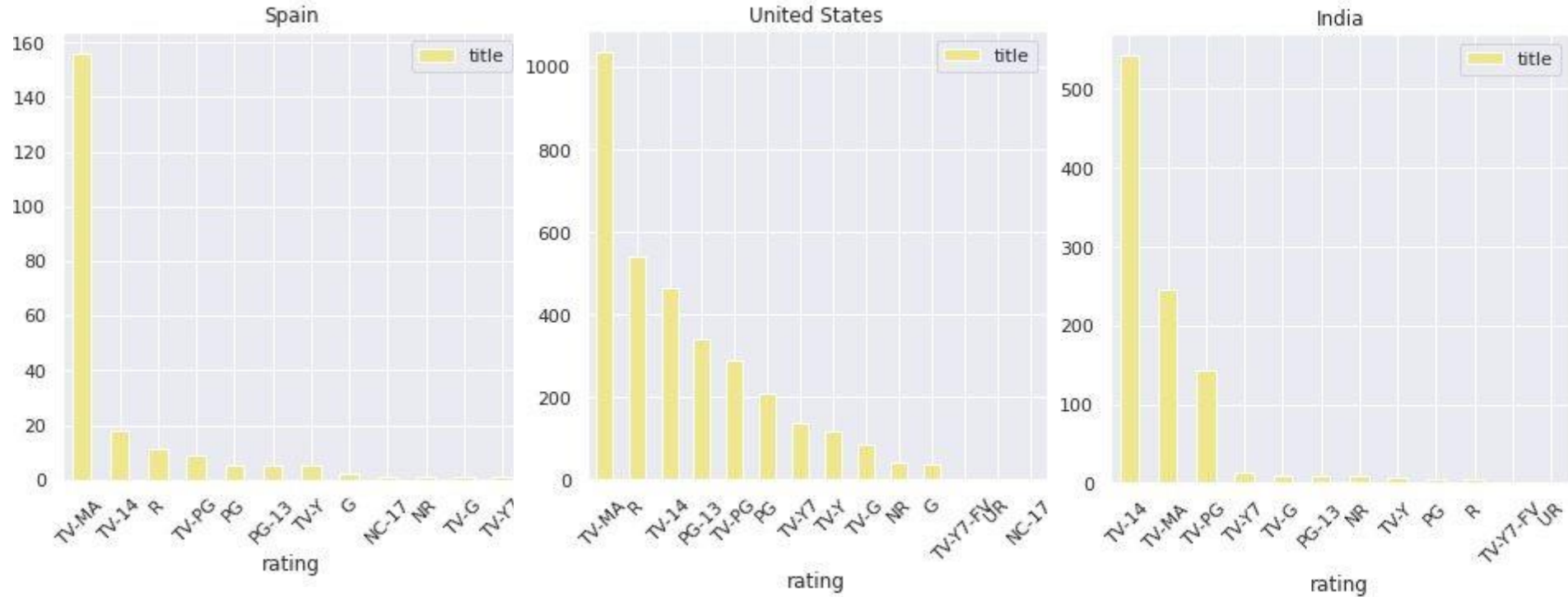
# Russia, US & India: Movies or Shows?



Release of TV shows as compare to Movies are more in Russia as compare to India and US

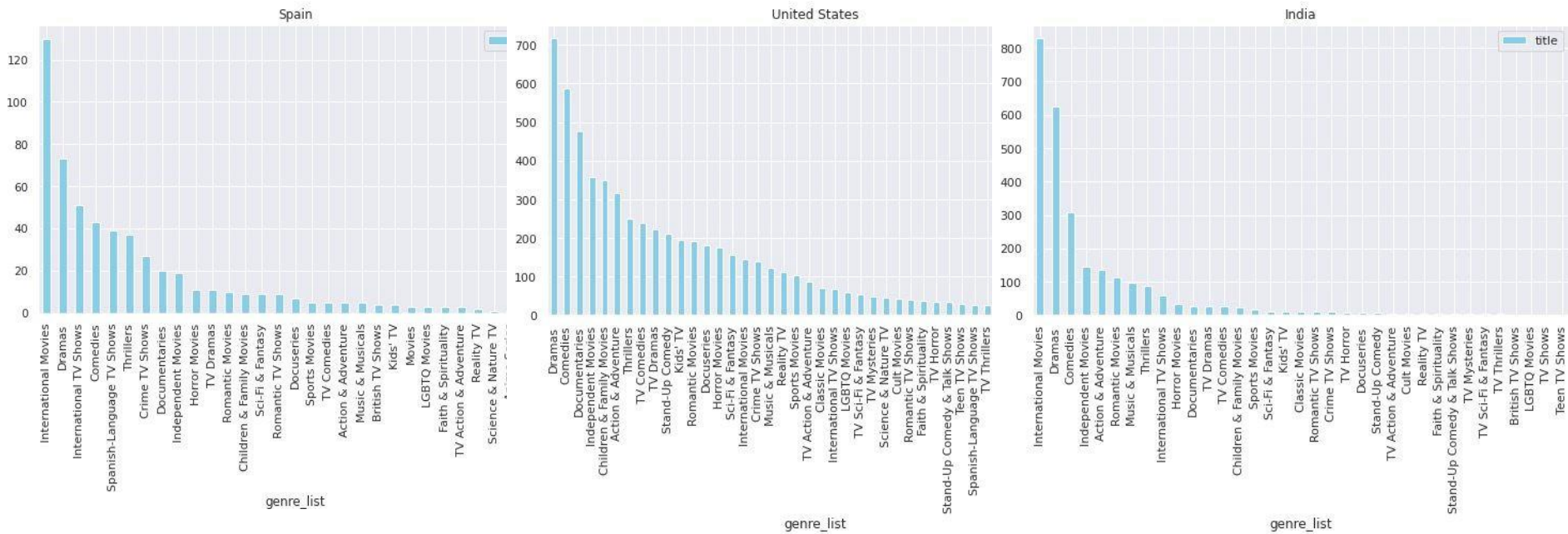


# Spain, US & India: Rating



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.



## AI

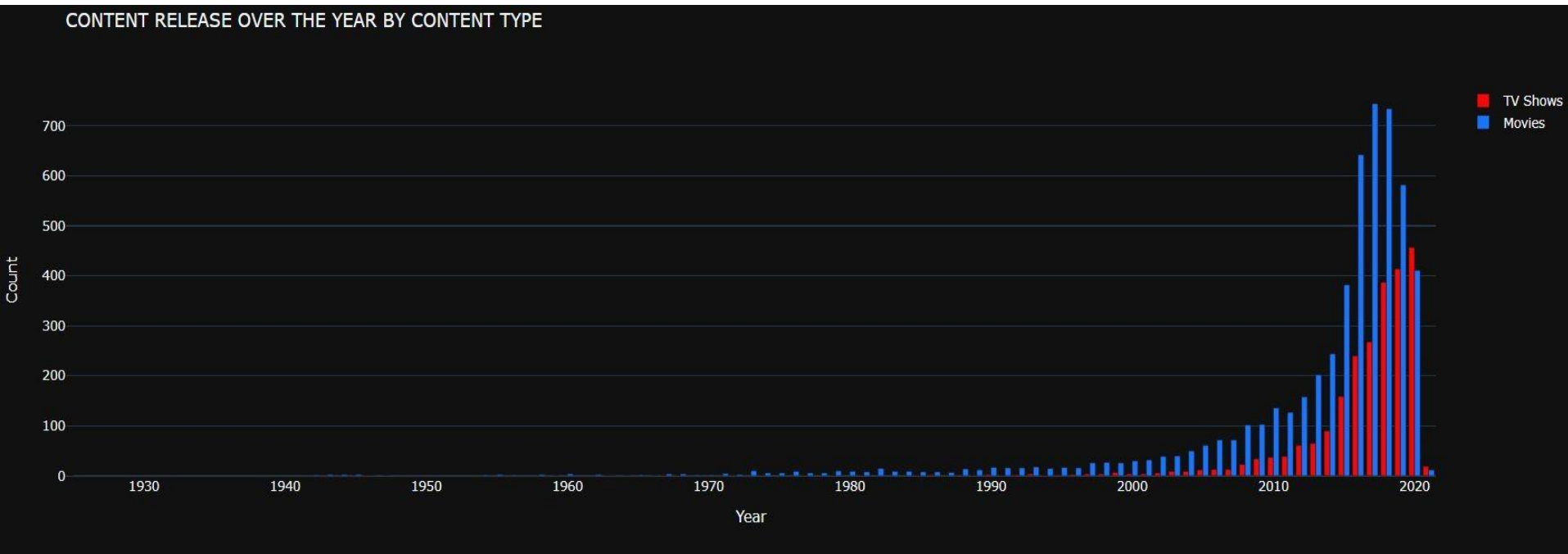








# Is Netflix focusing more on TV Shows/Movies



# 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_tokenize` from nltk

## 4. STEMMING

- Transformed words to roots
- Used Snowball Stemmer

*Everybody stand back, I know regex expressions!*

# Time to Cluster..

Vectorization

Dimensionality Reduction

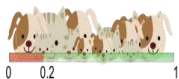
Finding Optimal K

K - Means Clustering with Optimal k

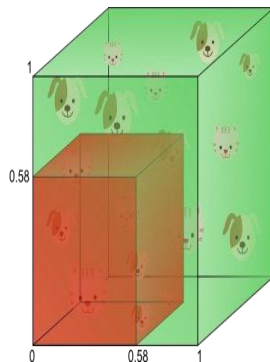
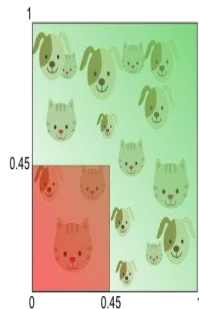
TFIDF Vectorizer

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

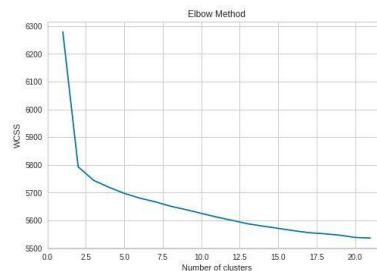
$$\text{idf}(t) = \log \frac{n+1}{\text{df}(d,t)+1} + 1 \quad (2)$$



PCA



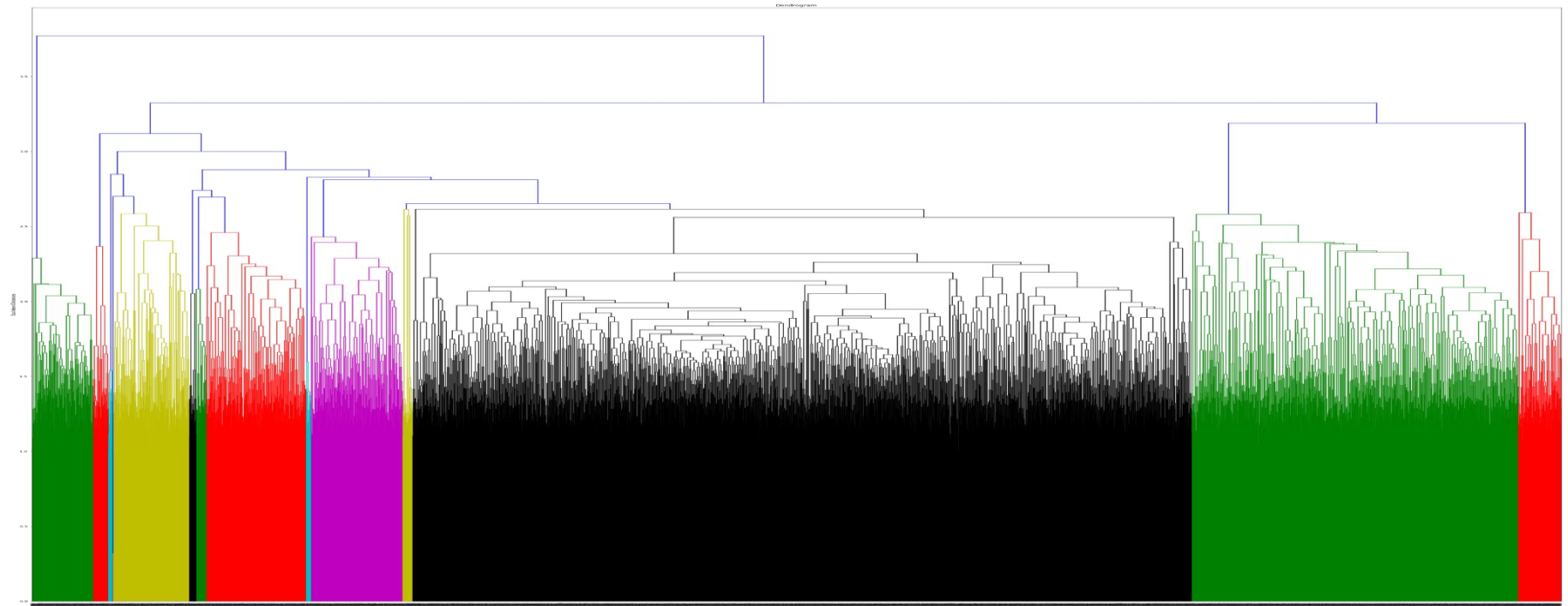
Elbow method:  
WCSS, Silhouette score



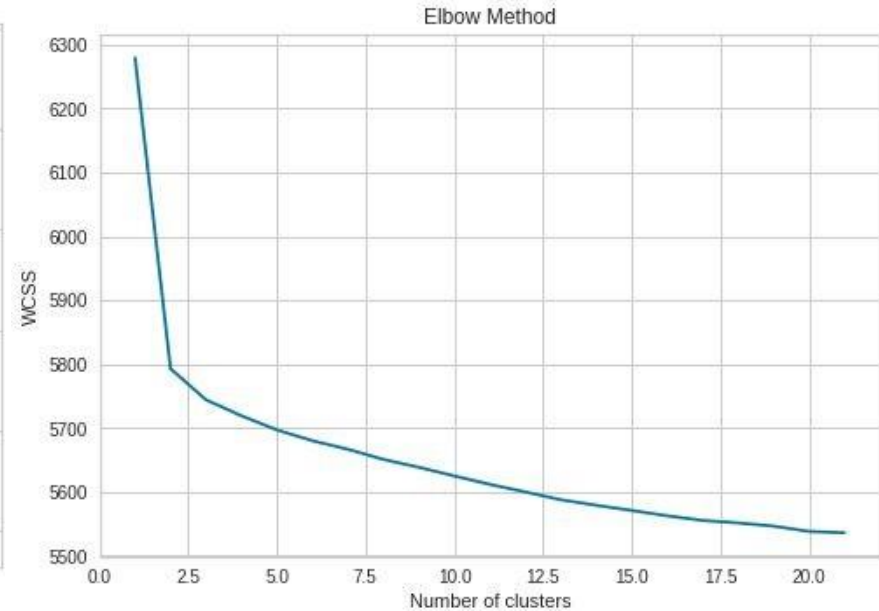
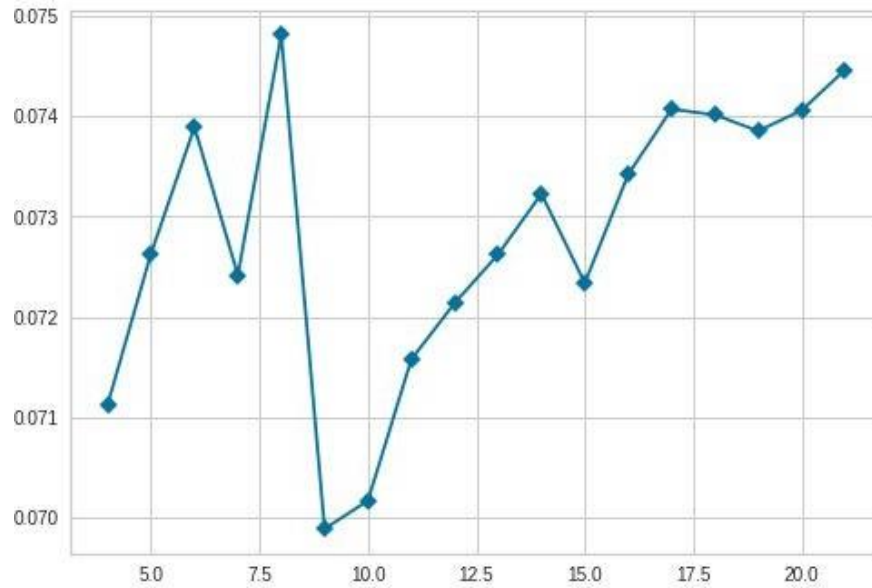
What could we infer?



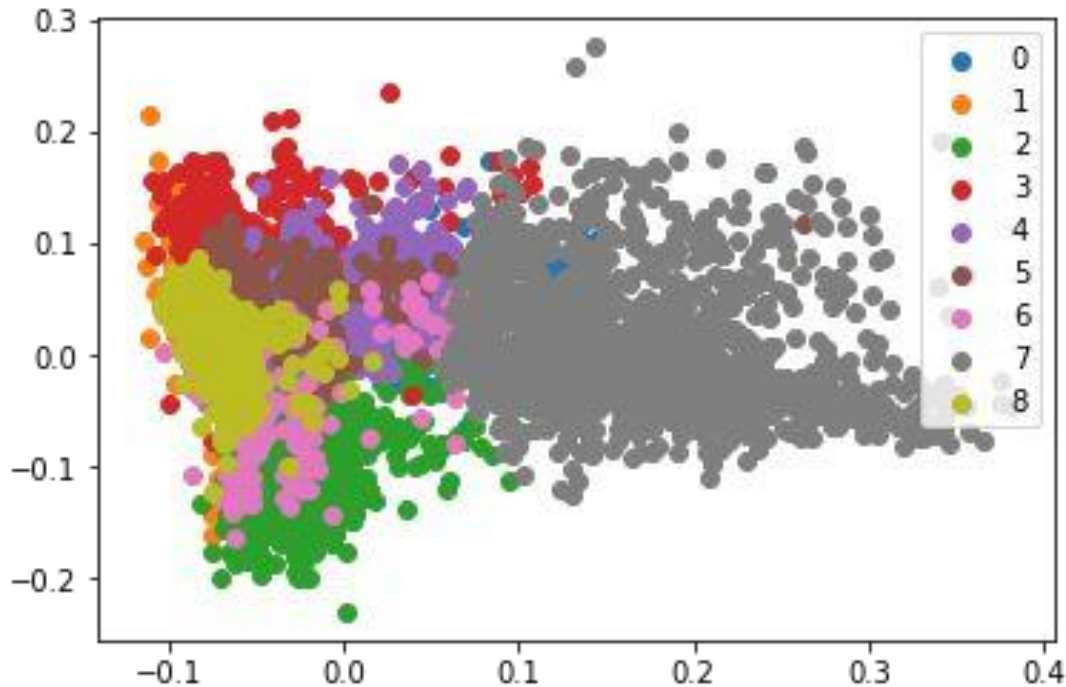
# Dendrogram..



# Finding Optimal K!



# Visualizing Clusters for $k=9$



# 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



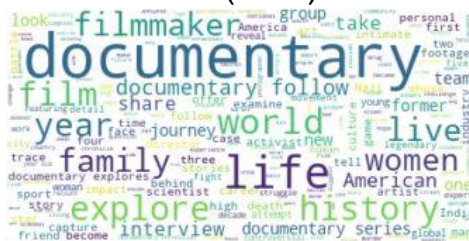
### C-0(Kids)



## C-1 (Musical)



C-2(Drama)



C-3(Documentary)



C-4(American)



### C-5(Comedy)



## C-6(Horror)



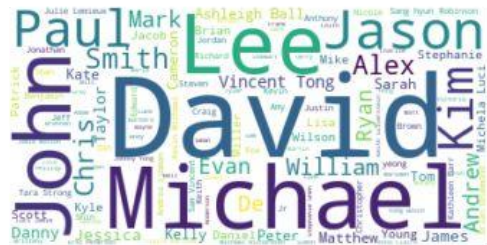
## C-7(TV Shows)



## C-8(Family)



# Cluster: Cast



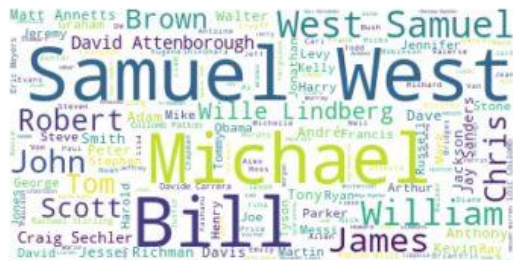
C-0



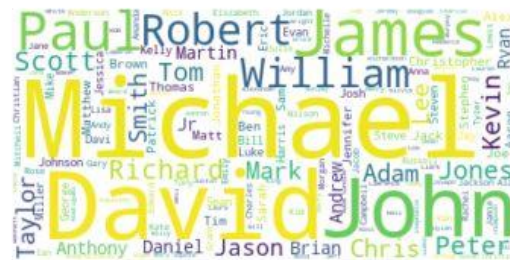
C-1(Indian)



C-2(Indian)



C-3



C-4



C-5



C-6

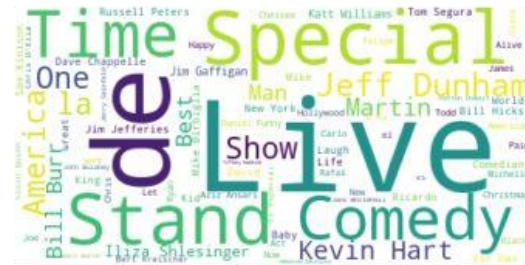
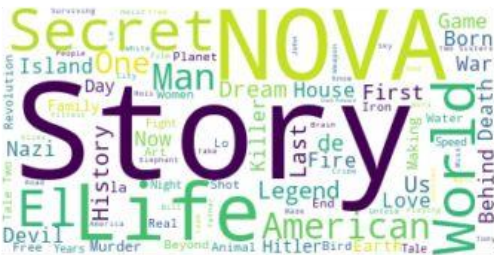


C-7(International)

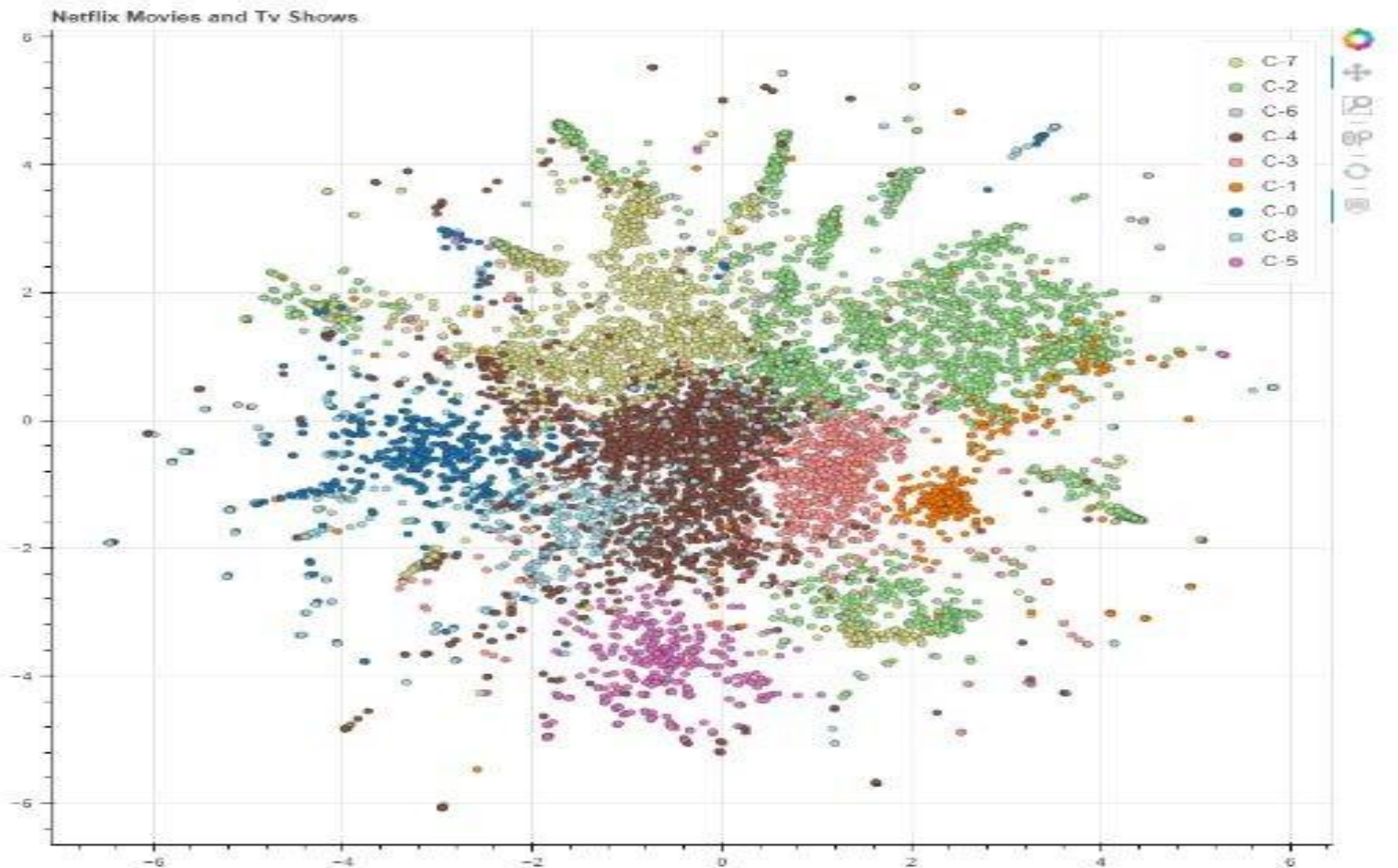


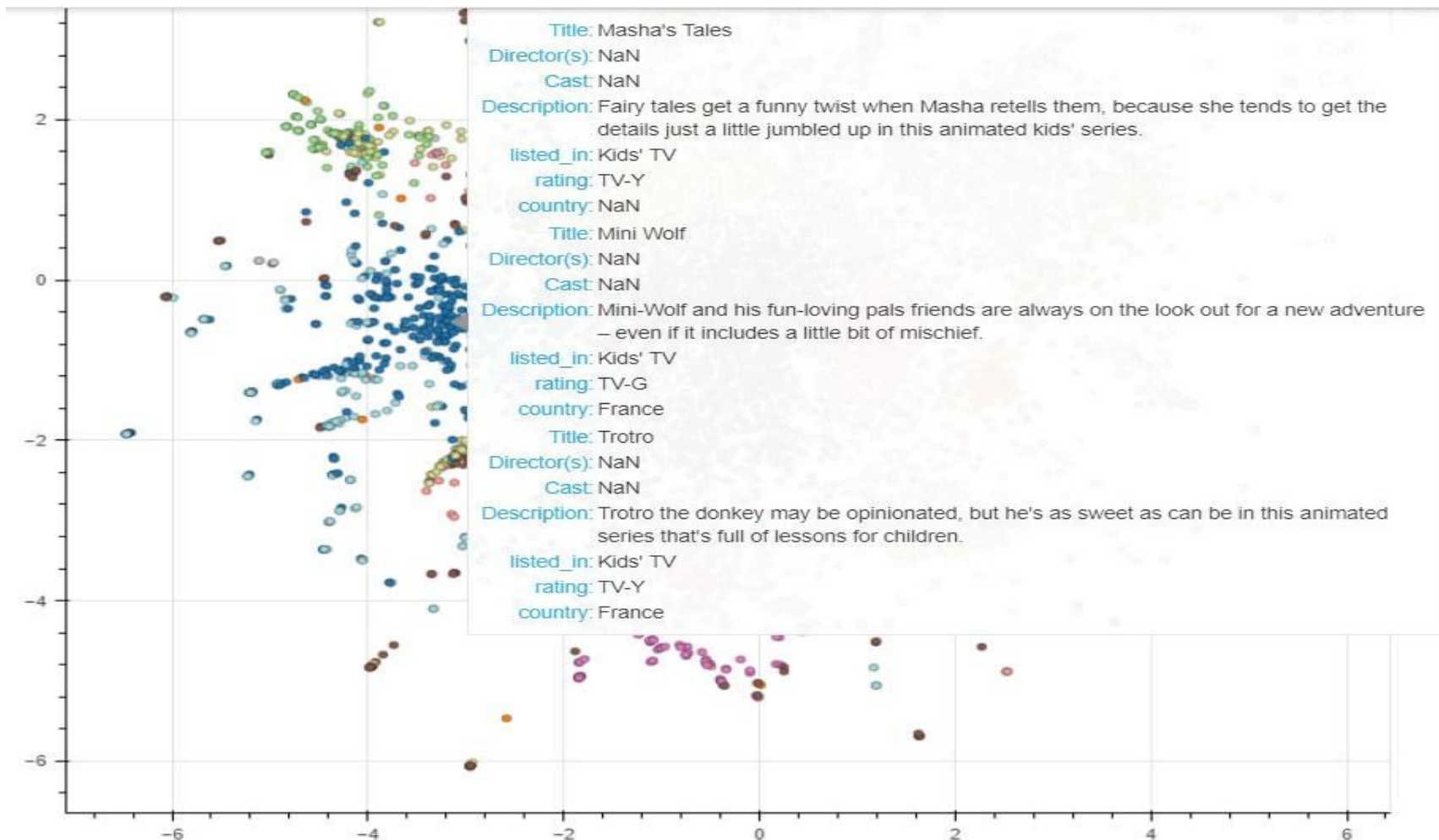
C-8





## Find similar movies / tv shows in corresponding Cluster







# 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!!

