Capstone Project - 4

NETFLIX MOVIES AND TV SHOWS CLUSTERING

Team Members

Harisha Chennozwala

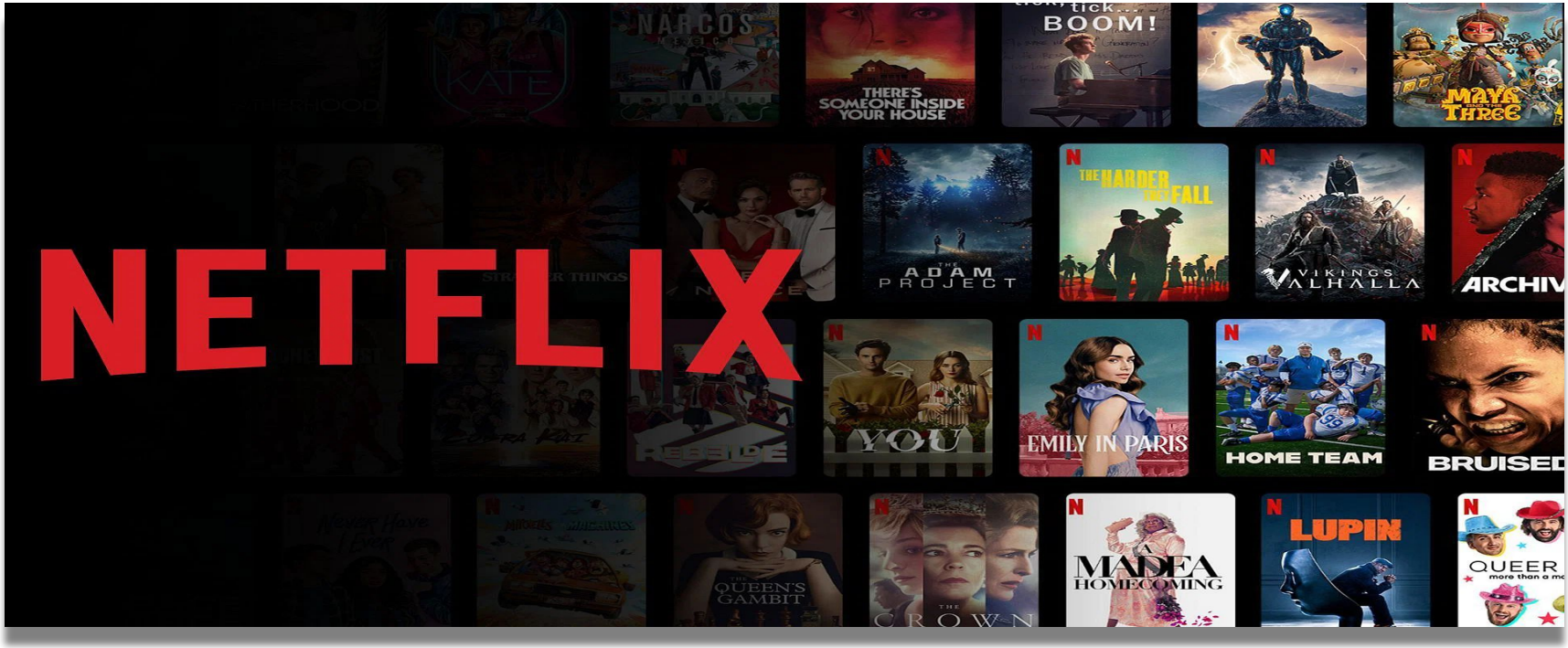
Niharika Soni

Satya Prakash

APPROACH OVERVIEW

Introduction	Objective	Data Acquisition	Data Cleaning	Exploratory Data Analysis
Netflix Movies & TV Shows Clustering and the objectives of this project	• Problem statement • Problem Objective	• Dataset description • Attributes	• Handling Null Values. • Clean data to get it ready for analysis.	• Exam data with visualizations. • Verify findings with statistical test.

Text Features-Clustering	Clustering	Dimensionality Reduction	Recommendation system	Conclusion
• Text Pre-Processing	• Vectorization • TF-IDF • Finding Optimal K • K-Means clustering with optimal K	• PCA • Most common words in each cluster	• Top recommendations	• Future work



Netflix is a streaming service that offers a wide variety of award-winning TV shows, movies, anime, documentaries and more – on thousands of internet-connected devices. You can watch as much as you want, whenever you want, without a single ad, in one simple subscription. There's always something new to discover, and new TV shows and movies are added every week! Streaming in more than 30 languages and 190 countries.

PROBLEM STATEMENT & OBJECTIVE

Exploratory Data Analysis.

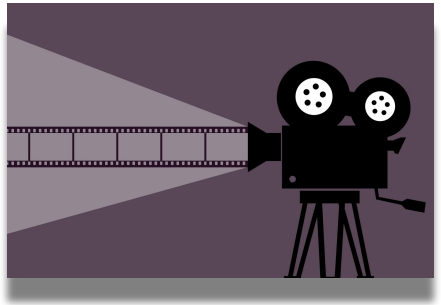
Understanding what type content is available in different countries.

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** (third party Netflix search engine)
- **7787** Rows of data
- Movies & TV Dataset till **2019**
- Dataset consists of **eleven textual** columns and **one numeric** column.



ATTRIBUTES

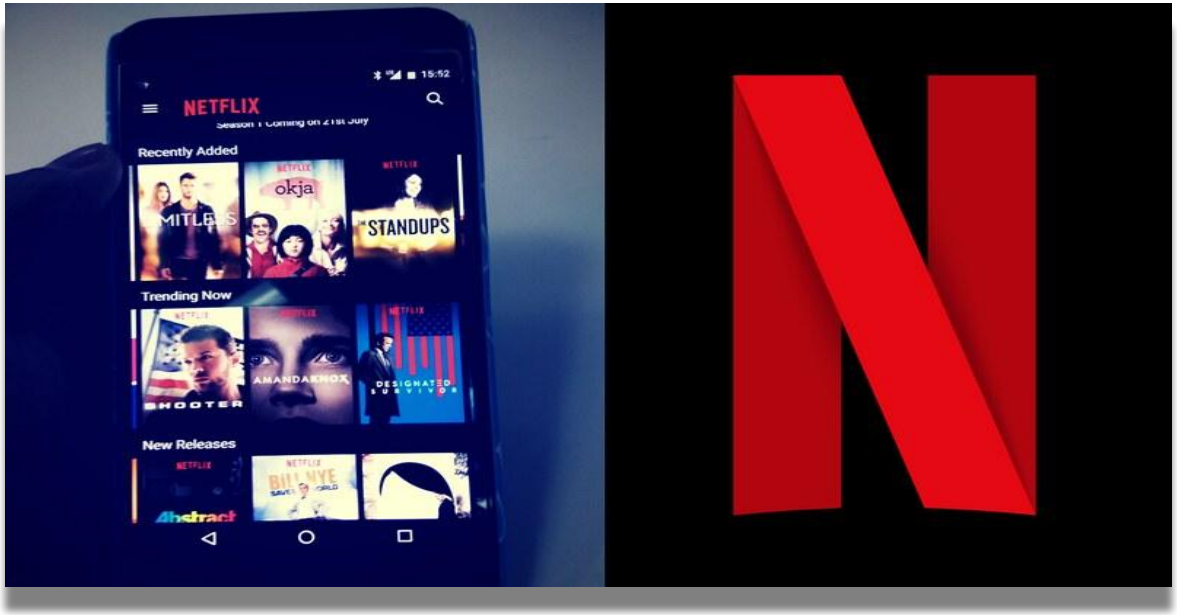
Show_id	Unique ID for every Movie / Tv Show
Type	Identifier - A Movie or TV Show
Title	Title of the Movie / Tv Show
Director	Director of the Movie
Cast	Actors involved in the movie / show
Country	Country where the movie / show was produced
Date_added	Date it was added on Netflix
Release_Year	Actual Release year of the movie / show
Rating	TV Rating of the movie / show
Duration	Total Duration - in minutes or number of seasons
Listed_in	Genre
Description	The Summary description

12

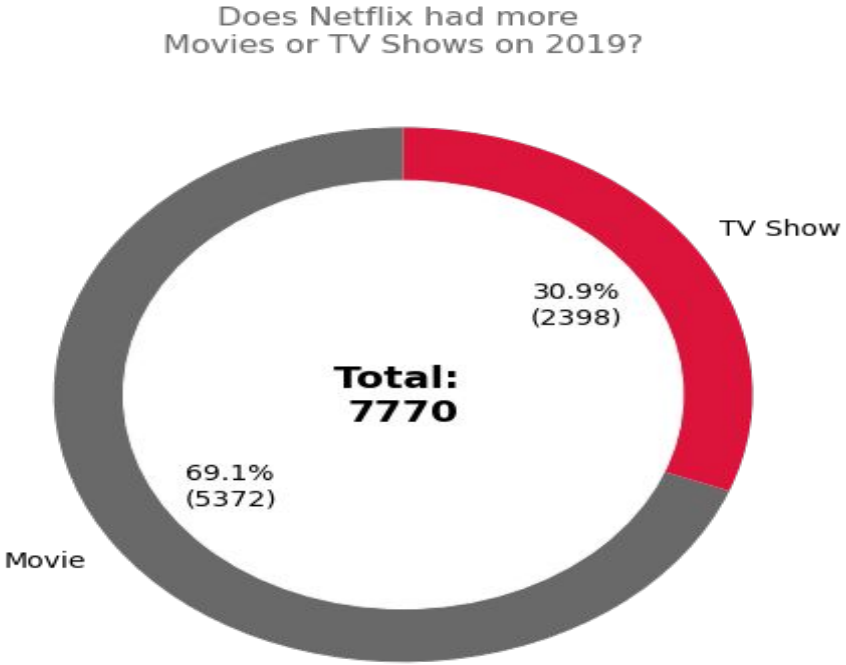
DATA CLEANING

- ★ **Handling Comma & De-limited Values**
- ★ Movies are based on the **duration** of the movie and shows are based on the number of **seasons**. To make EDA easier, convert the values in these columns into **integers** for both the movies and shows datasets.
- ★ **Removing** unnecessary columns like **director, cast.**
- ★ **Dropping** all NAN values containing **date_addded** observations.
- ★ Created 4 new columns :
  No. of categories based on listed_in.
  Date_added_month based on date_added
- ★ **Handling Null Values** - There are a total of **3,631 null values** across the entire dataset with missing points under 'director', 'cast', 'country', 'date_added', and 'rating'.
  - There are **2389** null values in the **Director** column.
  - There are **718** null values in the **cast** column.
  - There are **507** null values in the **country** column.
  - There are **10** null values in the **date added** column.
  - There are **7** null values in the **rating** column.

EXPLORATORY DATA ANALYSIS

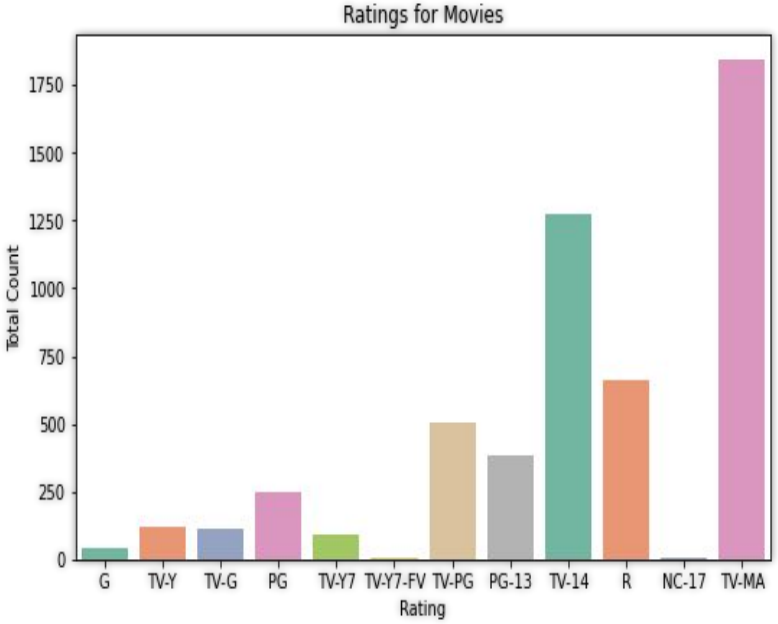


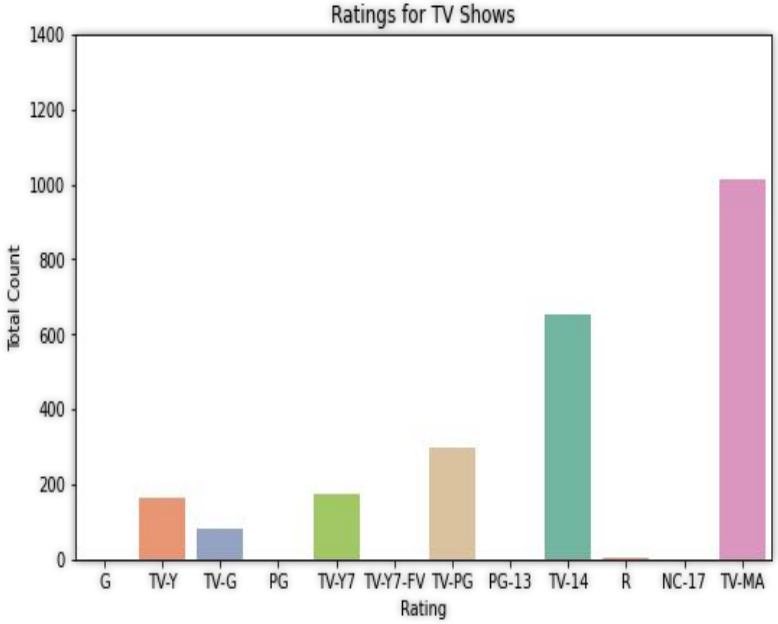
TV SHOWS OR MOVIES?



There are 5,372 movies and 2,398 TV shows. There are far more movie titles (69.1%) that TV shows titles (30.9%) in terms of title.

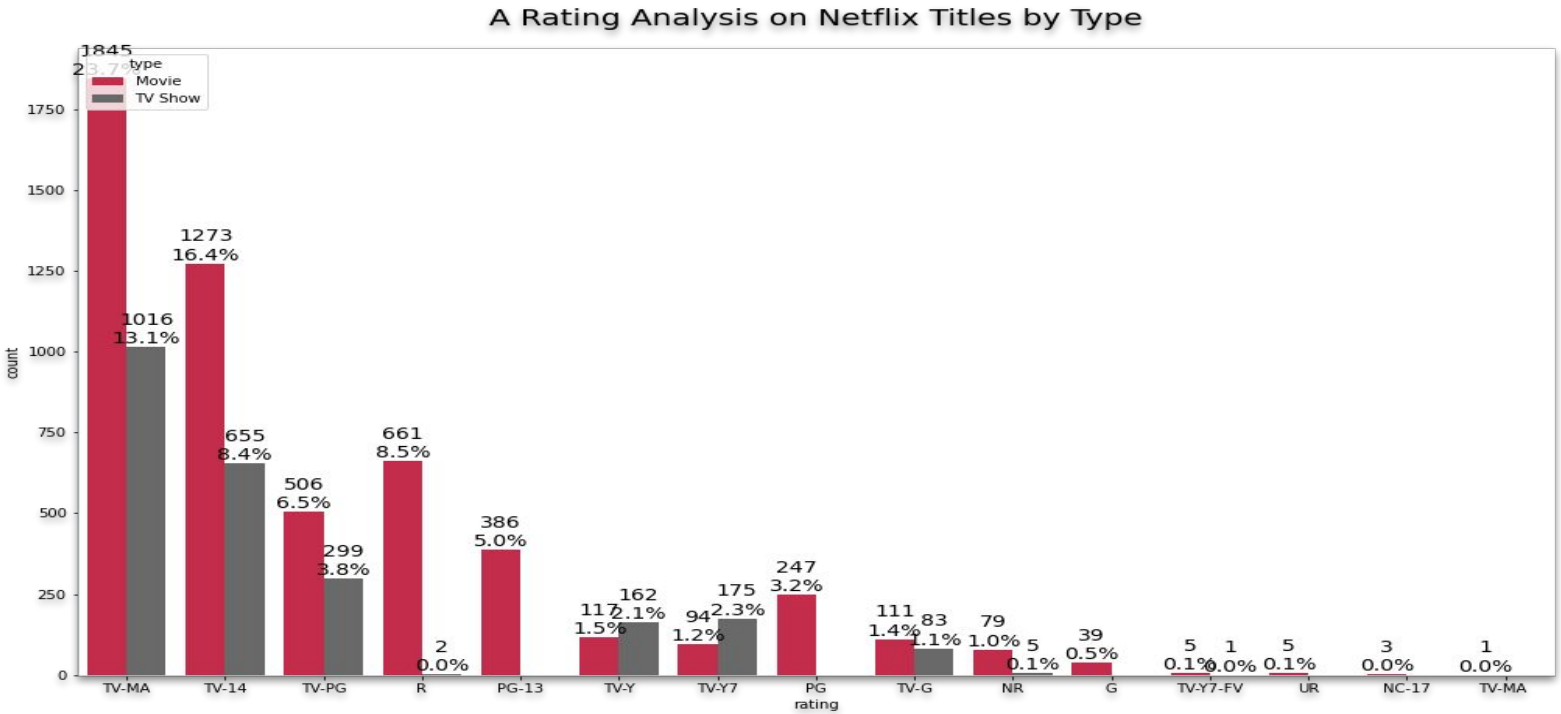
RATINGS ON NETFLIX





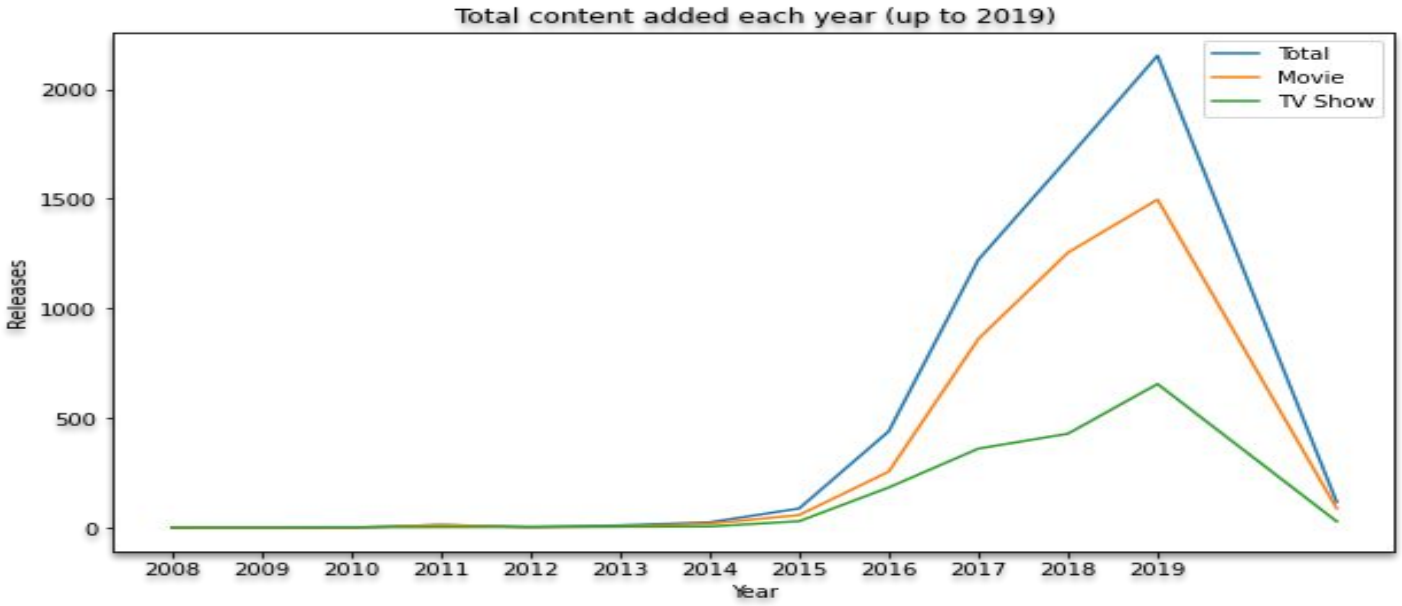
The largest count of Netflix content is made with a “TV-14” rating. But the largest count of TV shows is made with a “TV-MA” rating(“TV-MA” is a rating assigned by the TV Parental Guidelines to a television program designed for mature audiences only).

RATING ANALYSIS ON NETFLIX TITLES BY TYPE



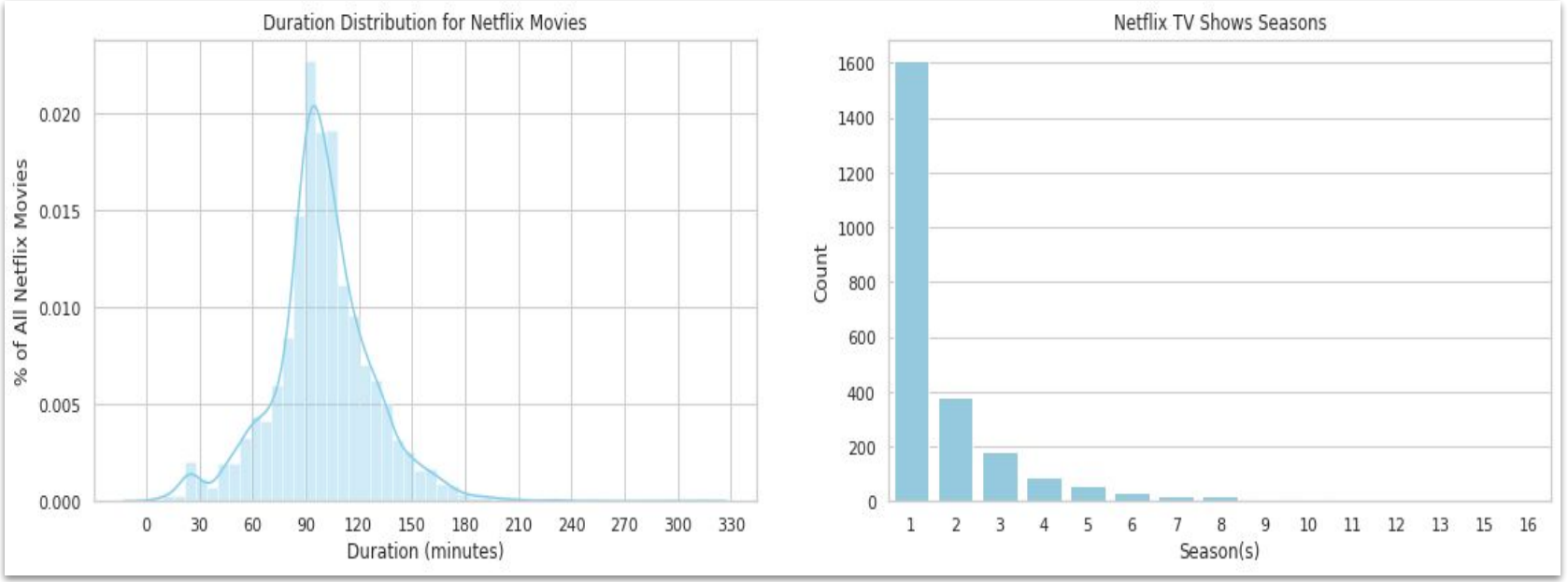
Mature audiences have more Movies content whereas younger(below 17 age) have more TV Shows content.

CONTENT ADDED EACH YEARLY, MONTHLY



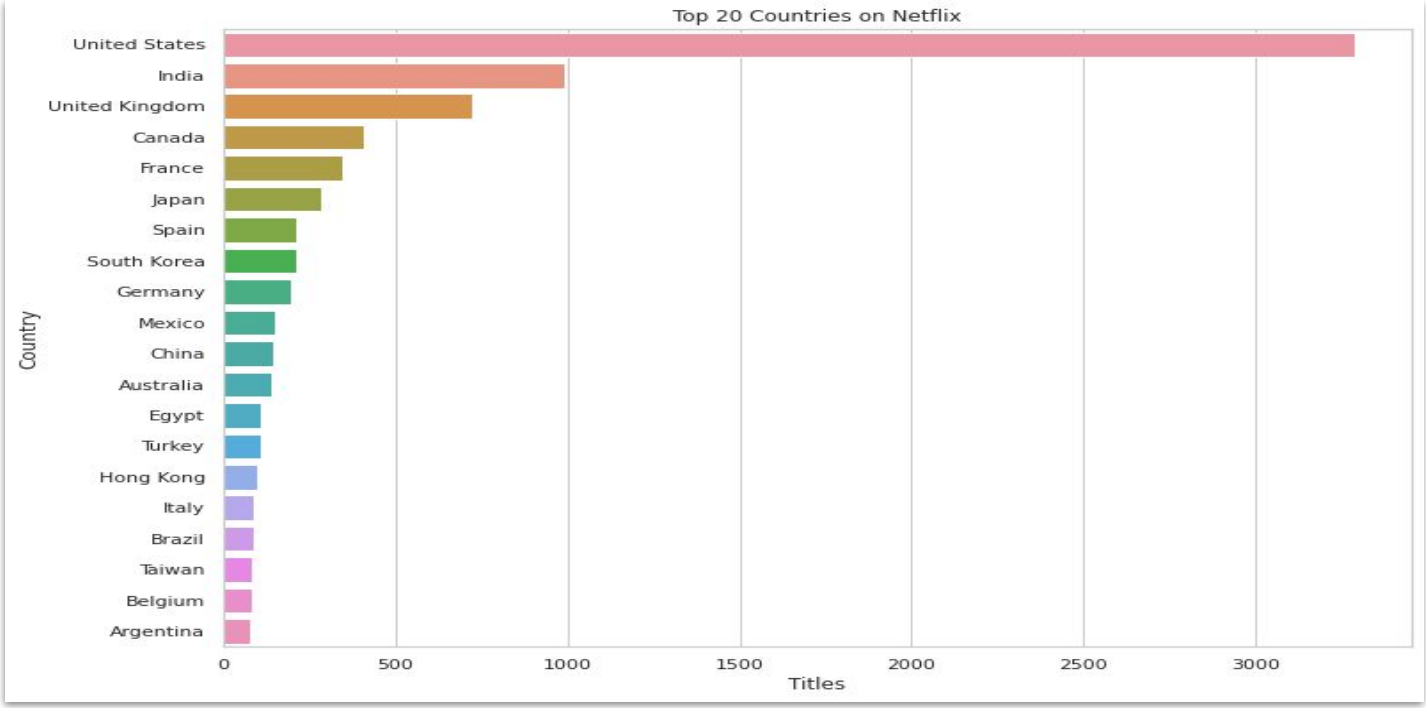
- The popular streaming platform started gaining traction after **2014.**
- There has been a consistent growth in the number of **movies** on Netflix compared to **shows**.

MOVIES & TV SHOWS DURATION(Minutes & Seasons)



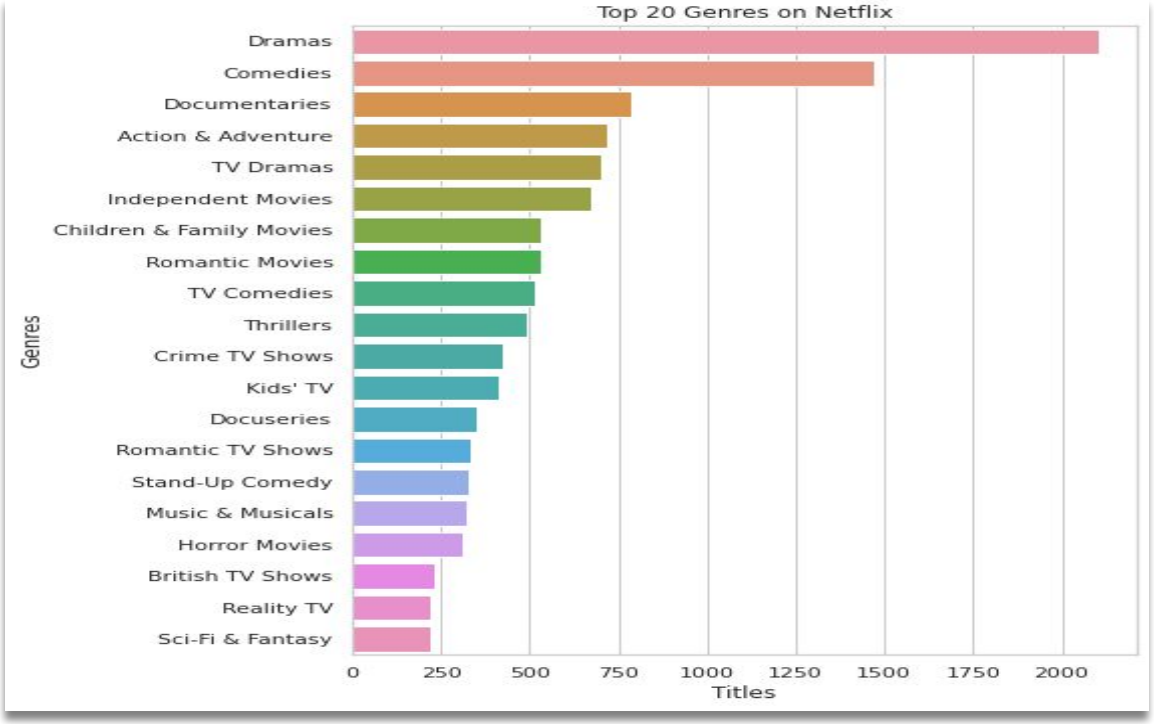
- Duration for **Netflix movies** closely resembles a normal distribution with the average viewing time spanning about **90 minutes.**
- **Netflix TV shows** on the other hand seems to be heavily **skewed to the right** where the majority of shows only have **1 season**.

COUNTRIES WITH MOST CONTENT AVAILABLE



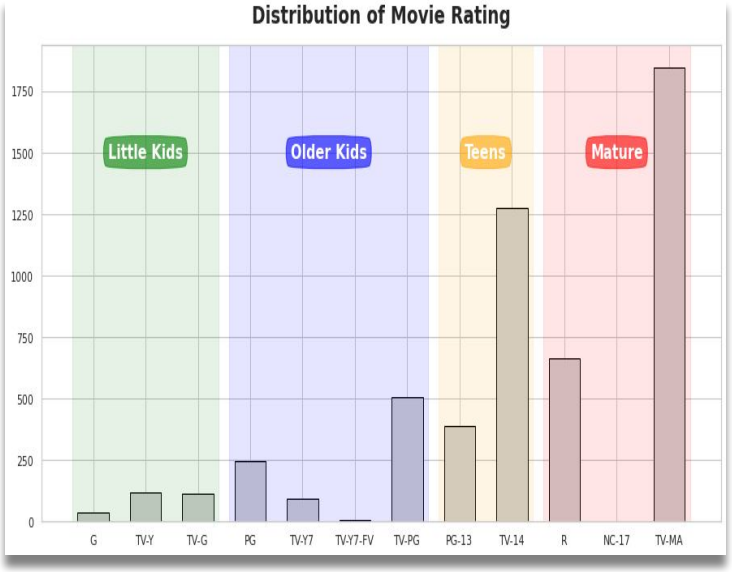
The United States stands out on top. **India** surprisingly comes in second followed by the **UK** and **Canada. China** interestingly is not even close to the top even though it has about 18% of the world's population.

POPULAR GENRES

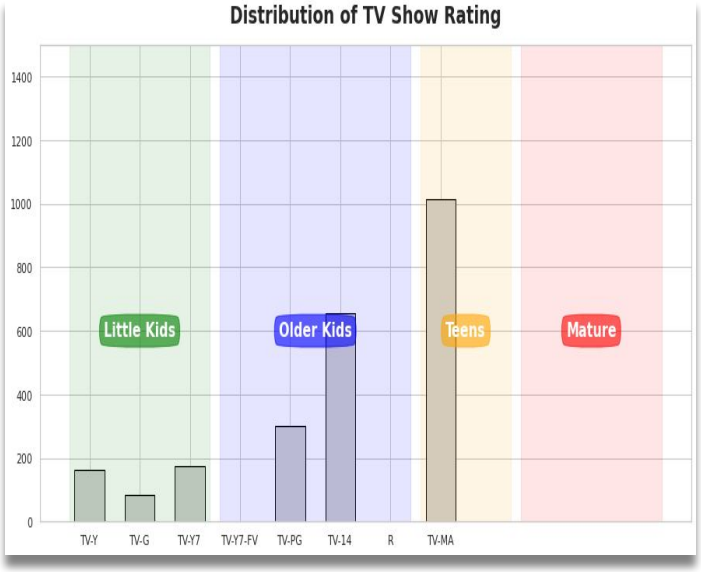


Dramas takes the cake surprisingly followed by **Comedies** and **Documentaries**.

HOW CONTENT IS DISTRIBUTED BASED ON MATURITY LEVEL - KIDS, TEENS & ADULTS-MOVIES & TV SHOW RATING



Less movies for the **kids More** for **teens & adults**.



TV Shows are **more** for the **kids and teens.**

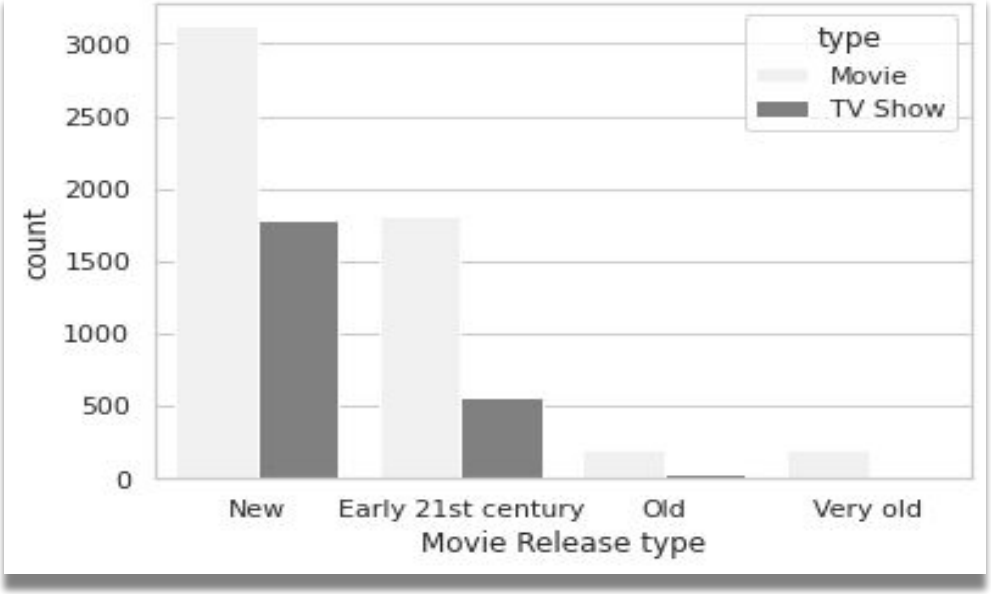
TV-MA- for mature audiences

TV-14 - May be unsuitable for children under 14

TV-PG - Parental guidance suggested.

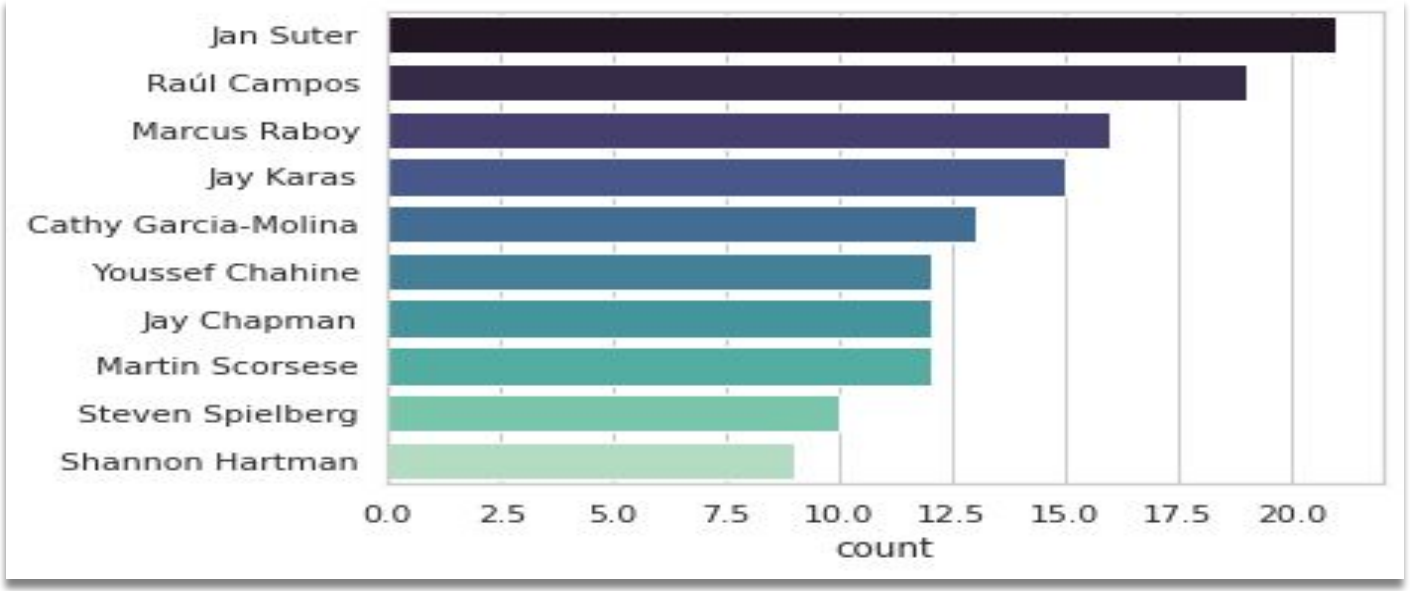
NR - Not rated.

DISTRIBUTION OF RELEASE TYPE



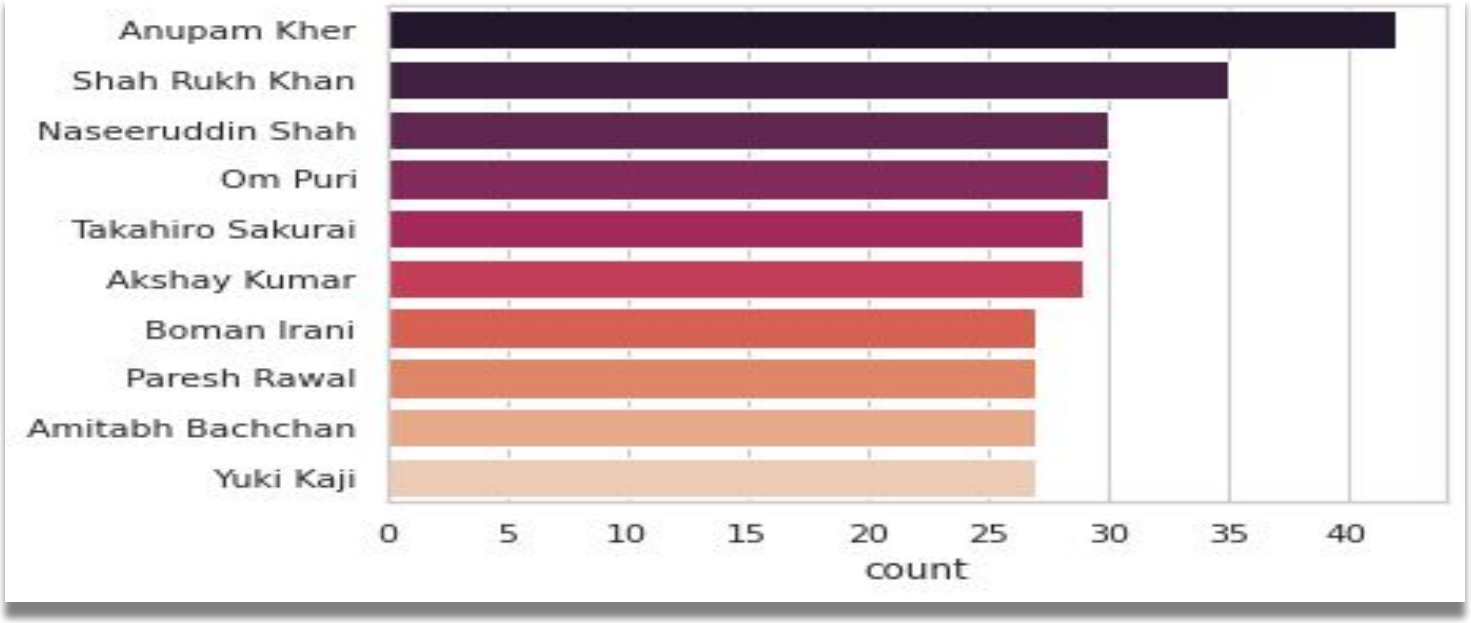
There are more **movies released in recent times** compared to the early 21st century, old and very old.

TOP 10 DIRECTORS ON NETFLIX WITH THE MOST RELEASES?



The most **popular directors** on Netflix with the most titles are mostly **international**.

TOP 10 ACTORS ON NETFLIX MOVIES BASED ON THE NUMBER OF TITLES



The majority of Netflix movies are starring **Indian actors**.

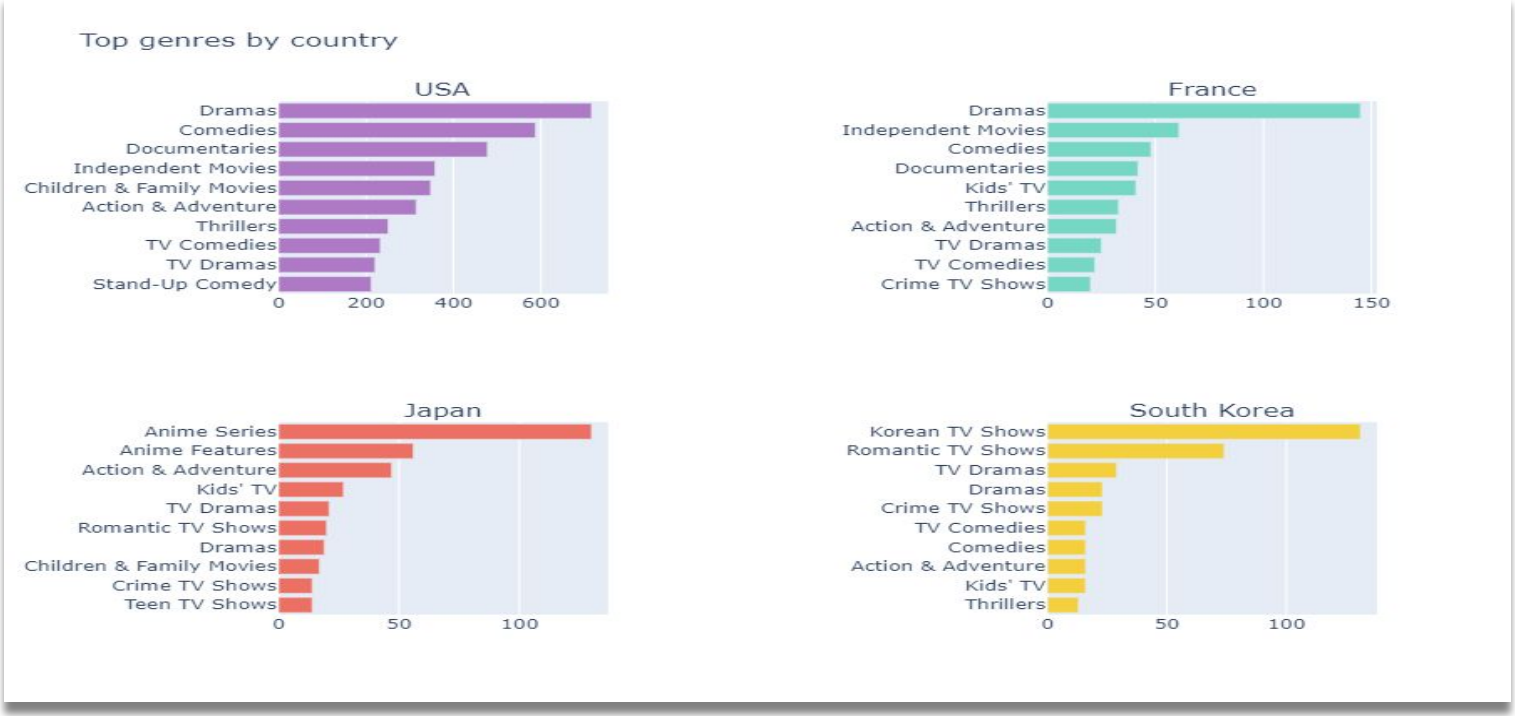
NETFLIX BY TITLES DESCRIPTION ON MOVIES: GENRES

Word Cloud applied on Genres. **Most present movies** are Romantic, Dramas, Independent, Action, Adventure, Children, Family, Comedies.

NETFLIX BY TITLES DESCRIPTION ON TV SHOWS: GENRES

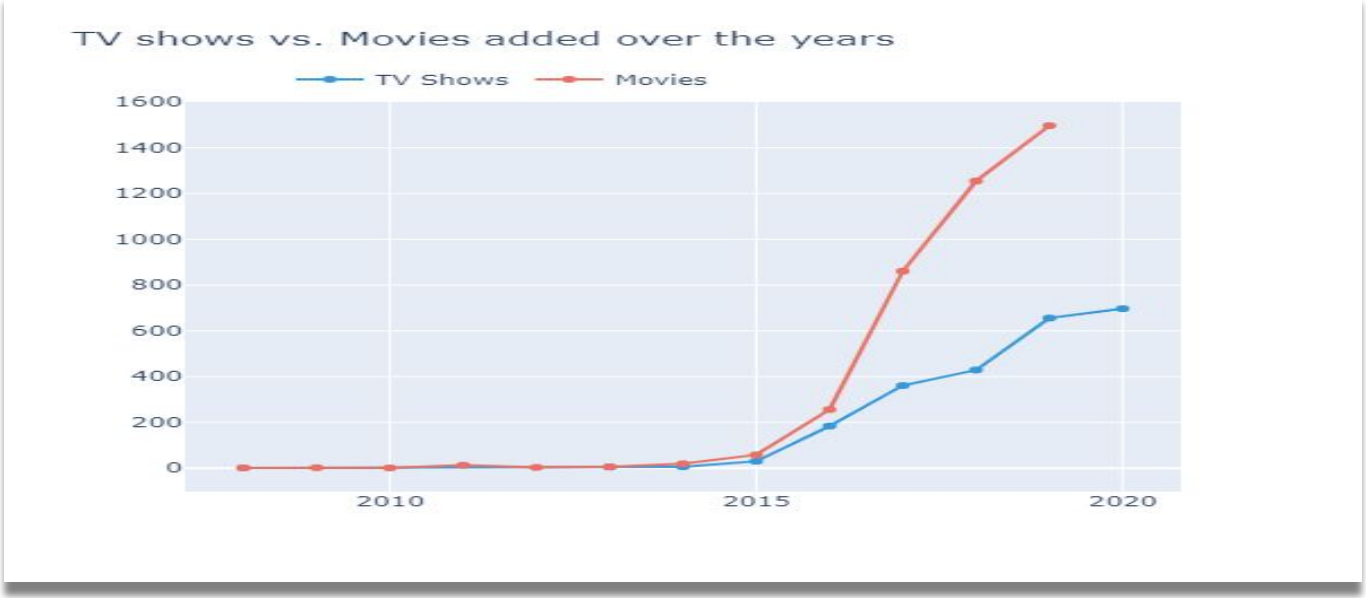
Most present TV Shows are Comedies, crime, romantic, adventure, Korean, action, dramas, british tv, reality shows, Anime & kids.

TYPE OF CONTENT AVAILABLE IN DIFFERENT COUNTRIES



American and French viewers on Netflix may favour Drama and Comedies over Anime series compared to Japanese viewers, whereas viewers in South Korea favour Korean TV shows the most.

DETERMINE IF NETFLIX HAS INCREASINGLY FOCUSING ON TV RATHER THAN MOVIES IN RECENT YEARS



In 2019, Netflix added **1497 movies** and **656 TV shows**. So there is no strong evidence indicating that Netflix has switched focus from movies to TV shows.

CLUSTERING DATASET

TEXT FEATURES

Clustering similar content by matching text-based features

TEXT PRE-PROCESSING

1 CLEANING

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

2 STOP WORDS

- Removed stop words
- Normal English words & problem specific

3 TOKENIZATION

- Splitted sentences to tokens
- Used word_tokenise from NLTK

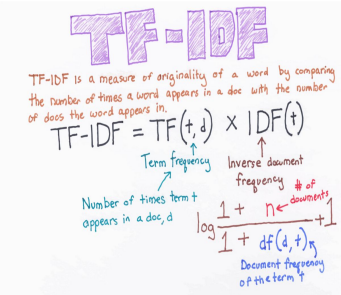
4 STEMMING

- Transformed words to roots
- Used Snowball stemmer

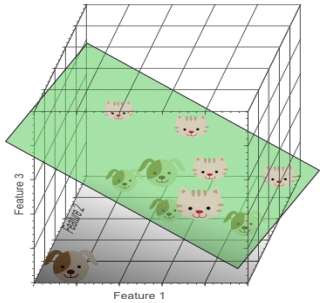
TIME TO CLUSTER....

VECTORIZATION → **DIMENSIONALITY REDUCTION** → **FINDING OPTIMAL K** → **K-MEANS CLUSTERING WITH OPTIMAL K**

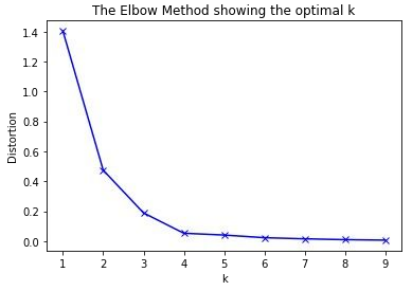
TF-IDF VECTORIZER



PCA

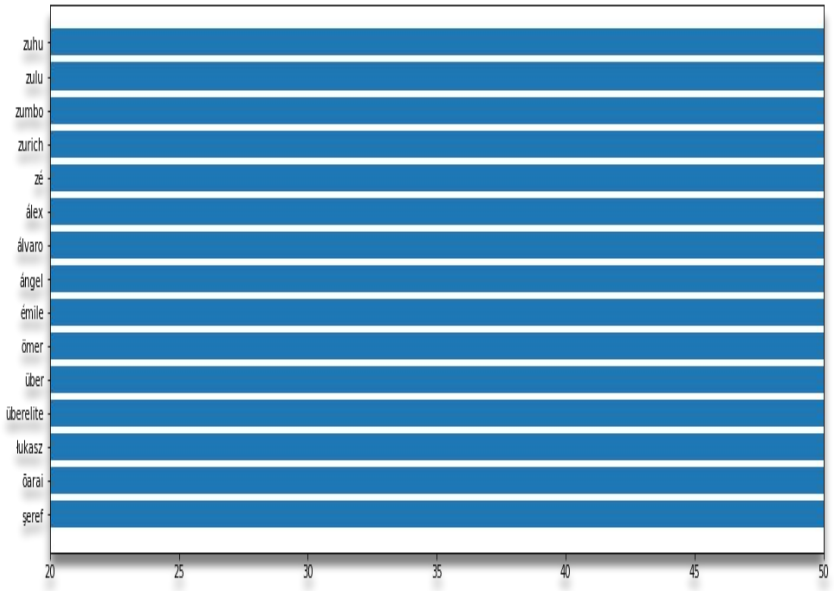


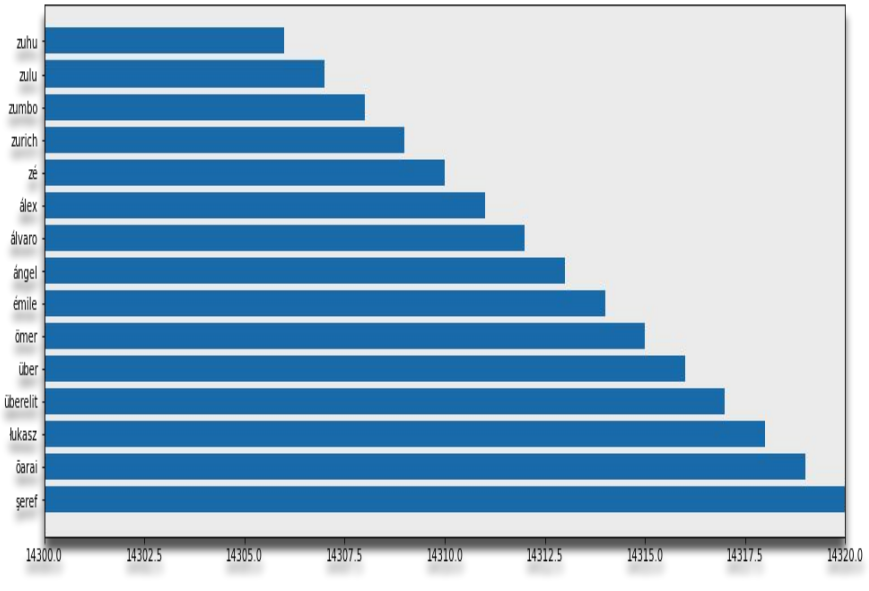
- Silhouette score
- Elbow method
- DBSCAN
- Dendrogram
- Agglomerative Clustering



What could we infer?

BEFORE & AFTER STEMMING MOST OCCURRED WORDS IN DESCRIPTION





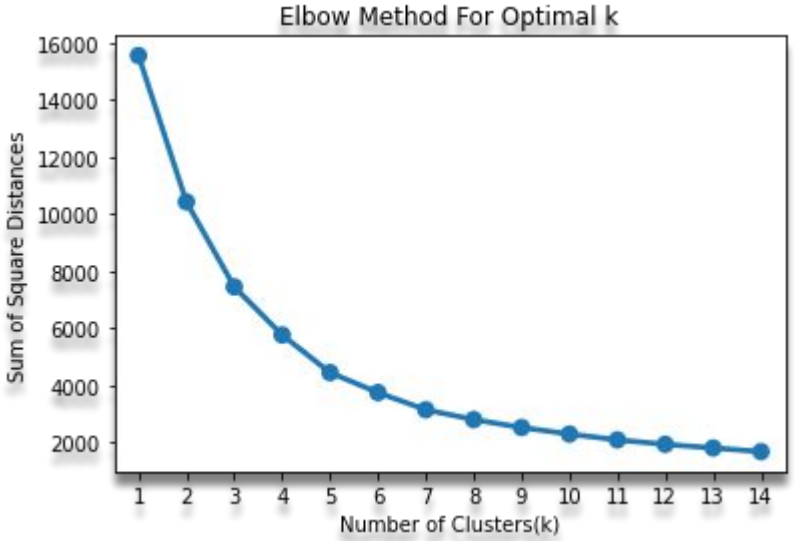
SILHOUETTE SCORE

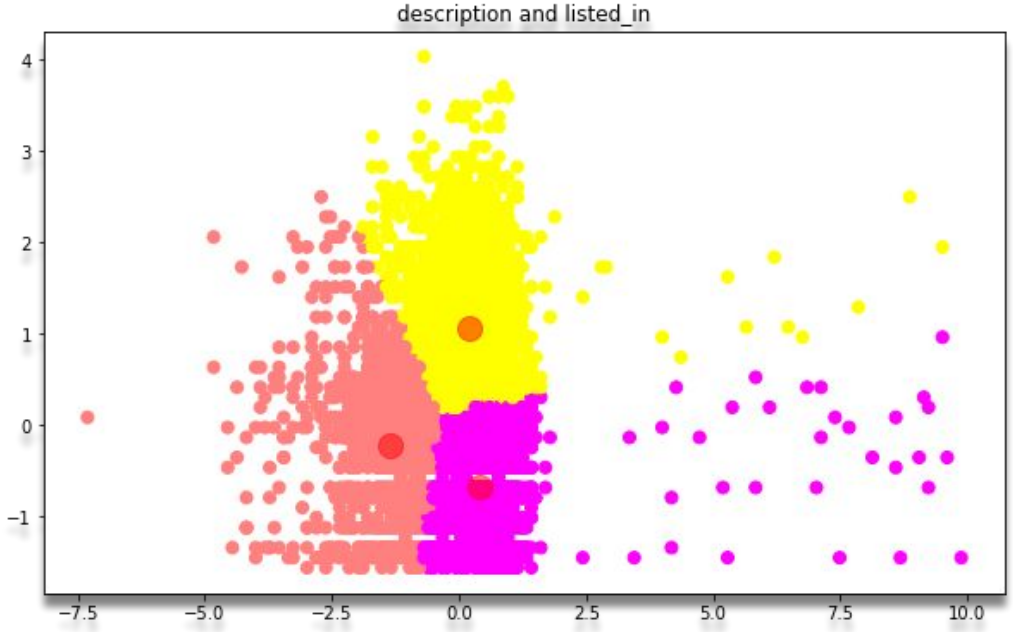
The Silhouette Coefficient formula is $s = \frac{b-a}{\max(a,b)}$

```
For n_clusters = 2 The average silhouette_score is : 0.34938940408423436
For n_clusters = 3 The average silhouette_score is : 0.37281514079850314
For n_clusters = 4 The average silhouette_score is : 0.386406784285601
For n_clusters = 5 The average silhouette_score is : 0.363888917635875
For n_clusters = 6 The average silhouette_score is : 0.34942375790377794
For n_clusters = 7 The average silhouette_score is : 0.3580466702151485
For n_clusters = 8 The average silhouette_score is : 0.3411548103829036
For n_clusters = 9 The average silhouette_score is : 0.3377515907271921
For n_clusters = 10 The average silhouette_score is : 0.3288937518856952
For n_clusters = 11 The average silhouette_score is : 0.3328338625852425
For n_clusters = 12 The average silhouette_score is : 0.3394631252075918
For n_clusters = 13 The average silhouette_score is : 0.3317026106523812
For n_clusters = 14 The average silhouette_score is : 0.33603842193554495
For n_clusters = 15 The average silhouette_score is : 0.3390576163480563
```

Silhouette score method(in range clusters): **Optimum score is 0.377 for n_clusters = 3.**

ELBOW METHOD

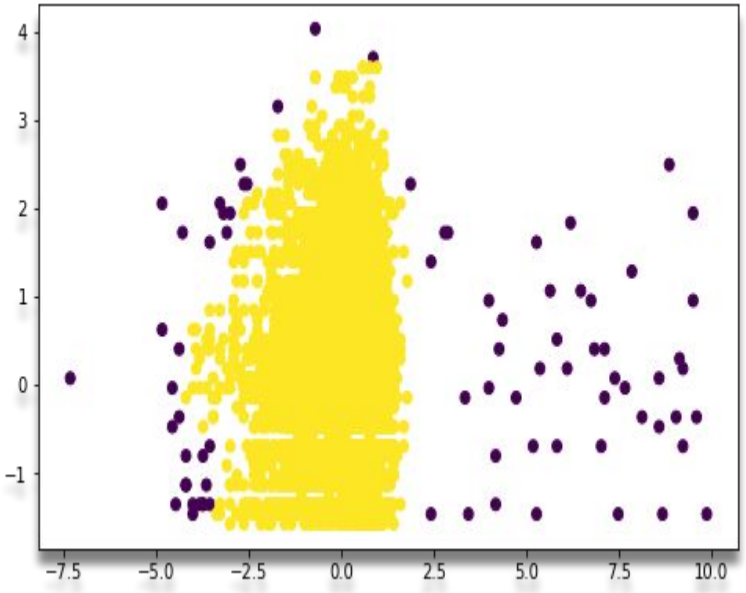




We can observe that the **“elbow”** is the number **6** which is optimal for this case. We can also verify this because there are **6 different genres** so this result was pretty much expected. Clustering is pretty accurate. Applying k-means clustering for K = 3

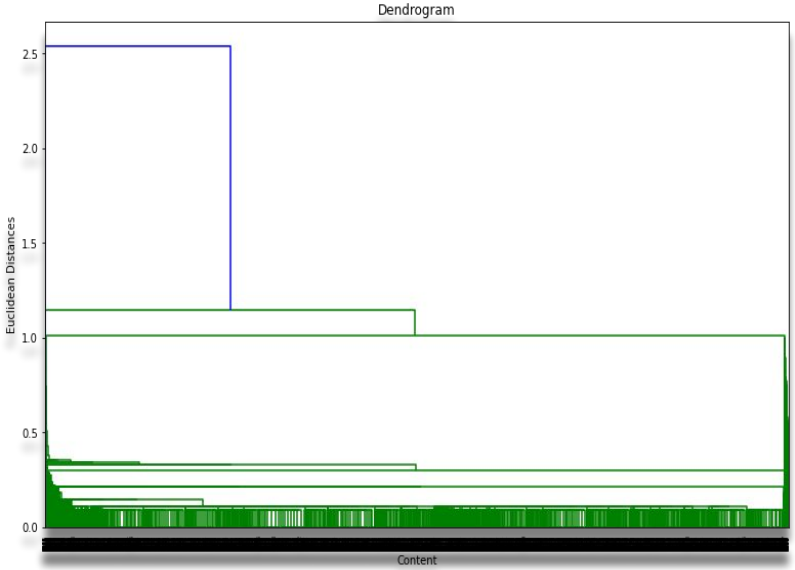
DBSCAN

Density-Based Spatial Clustering Of Applications With Noise



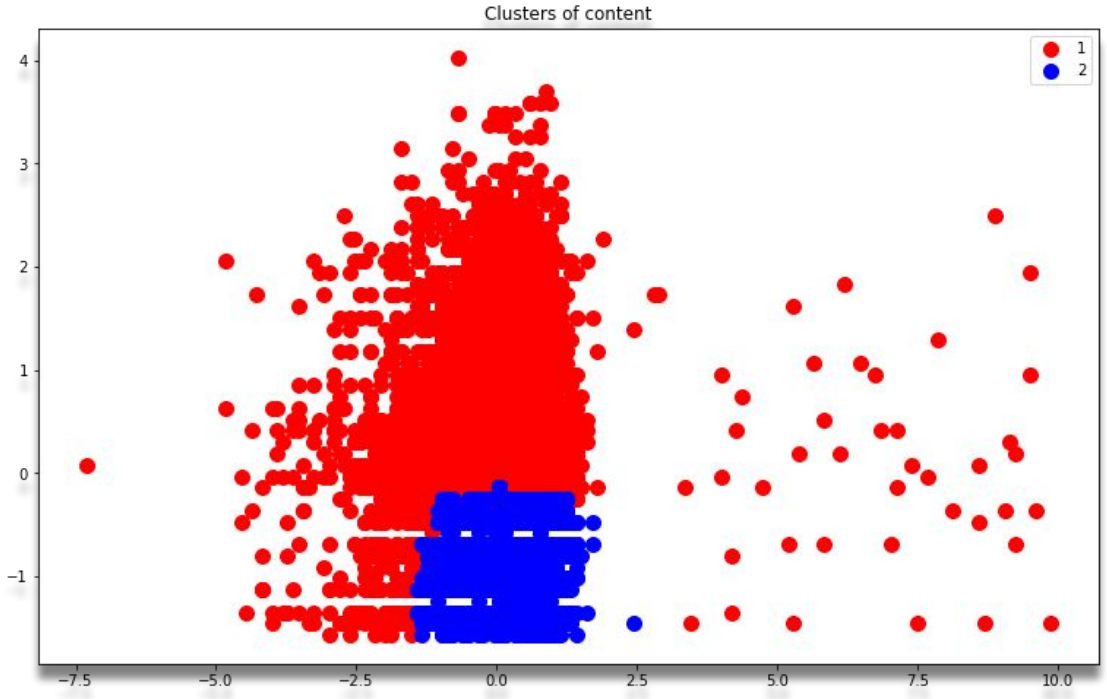
The graph above shows that there are outliers(Black points)-these points do not meet distance and minimum samples requirements to be recognised as a cluster.

DENDROGRAM



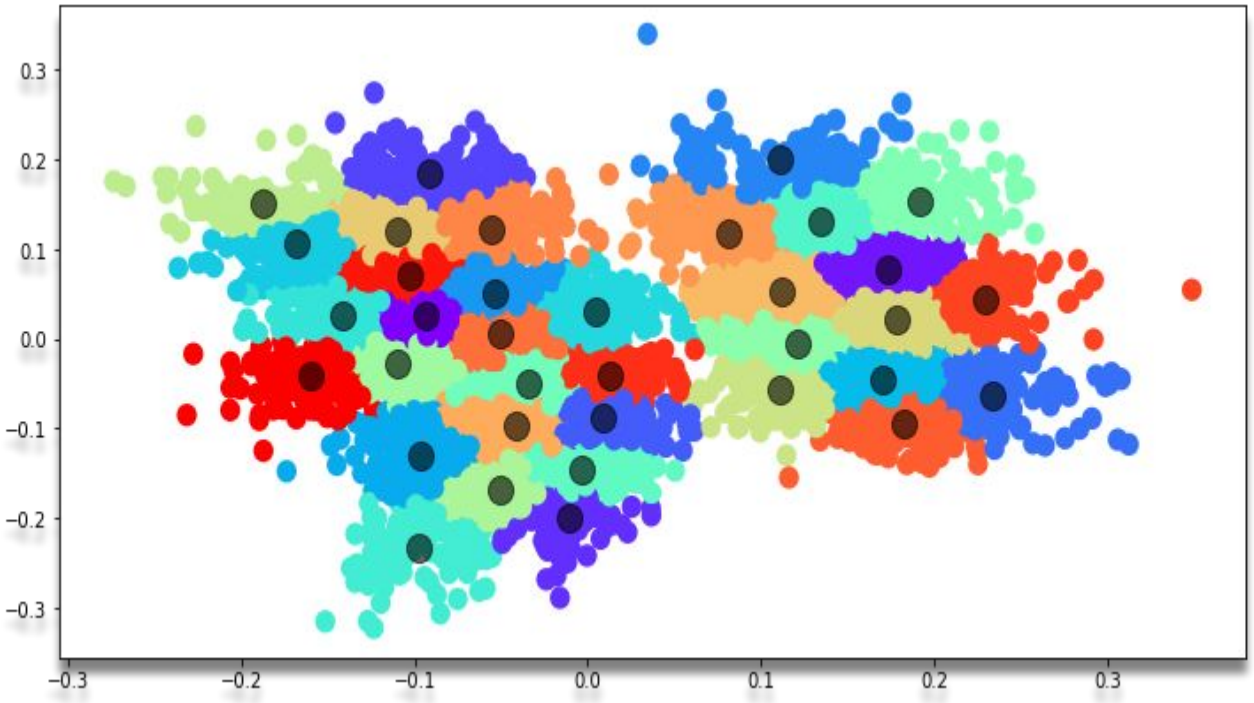
The number of clusters will be the number of vertical lines. It illustrates the arrangement of the clusters produced by the corresponding analyses.

AGGLOMERATIVE CLUSTERING



Applying hierarchical - Clusters of Content. **Optimal number of clusters are 2**

DIMENSIONALITY REDUCTION



After applying Principal Component Analysis(PCA) we have arrived with **Silhouette score 0.344**

RECOMMENDATION SYSTEM FOR NETFLIX.....!!!

- **Chosen Movie/Tv show:** Kapoor & Sons
- **Top Recommendations:**

5293	Rudy Habibie
6521	The Jungle School
99	3 Heroines
3876	Mak Cun
7047	This Earth of Mankind
2702	Her Only Choice
1992	Emma' (Mother)
3755	Love for Sale 2
3216	Jonaki
1274	Chaotic Love Poems

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:

1. Data set contains 7787 rows and 12 columns. In cast and director features contains large number of missing values so we can drop it and we have 10 features for the further implementation.
2. We have two types of content TV shows and Movies (30.9% contains TV shows and 69% contains Movies).
3. By analysing the content added over years we get to know that in recent years netflix is focusing movies than TV shows (movies is increased by 80% and TV shows is increased by 73% compare to 2016 data)
4. The most number of TV shows released in 2017 and movies in 2020 respectively.
5. On Netflix, Dramas genre contains the maximum content among all of the genres and the most of the content added in december month and less content in february.
6. The most content type on Netflix is movies.
7. The country by the amount of the produces content is the United States. And most of the countries preferred to produce movies more than TV shows.
8. International Movies is a genre that is mostly in Netflix.
9. The largest count of Netflix content is made with a “TV-14” rating,
10. The most popular actor on Netflix movie, based on the number of titles, is Anupam Kher.
11. In text analysis with stop words, removed punctuations, stemming & TF-IDF Vectorizer and other functions of NLP.
12. Applied different clustering algorithms like Silhouette score, Dendrogram, DBSCAN, Agglomerative, Elbow methods and got best cluster arrangements.
13. We got best optimal number of cluster is equal to 2.
14. Recommendation System


THANK YOU
“If you torture the data long enough, it will confess to anything” - RONALD COASE