

Capstone Project
Netflix Movies and TV Shows
Clustering
Team - Dream
Team Members
Indugopal Maity

1. What is “Netflix”
2. Defining Problem Statement
3. Data Pipeline
4. Data Overview
5. EDA
6. Applying Model Clustering
7. Conclusion



What is "Netflix"??

AI

Netflix is a subscription-based streaming service that allows our members to watch TV shows and movies without commercials on an internet-connected device.





Defining Problem Statement

Problem Statement:

Netflix has become dominant company in the on-demand media industry, with 167 million paying subscribers.

We have been provided a dataset collected from “Flixable”, which is a third-party Netflix search engine.

Our job is to:

- **To perform 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 Pipeline

- **Data Pre-processing:** After exploring and understanding our data, we did data cleaning by handling Null/missing values, checking for duplicate values. We further changed “date_added” variable to its appropriate date-time format and created a new variable “year_added” by extracting year from it.
- **EDA:** We performed exploratory analysis of data and find useful insights.
- **Creating Model:** After identifying useful features, we performed text cleaning by removing stopwords, punctuation and doing stemming of words. After calculating clean text lengths, we standardize those values and applied two clustering algorithms K-means and HAC(hierarchal Agglomerative Clustering).

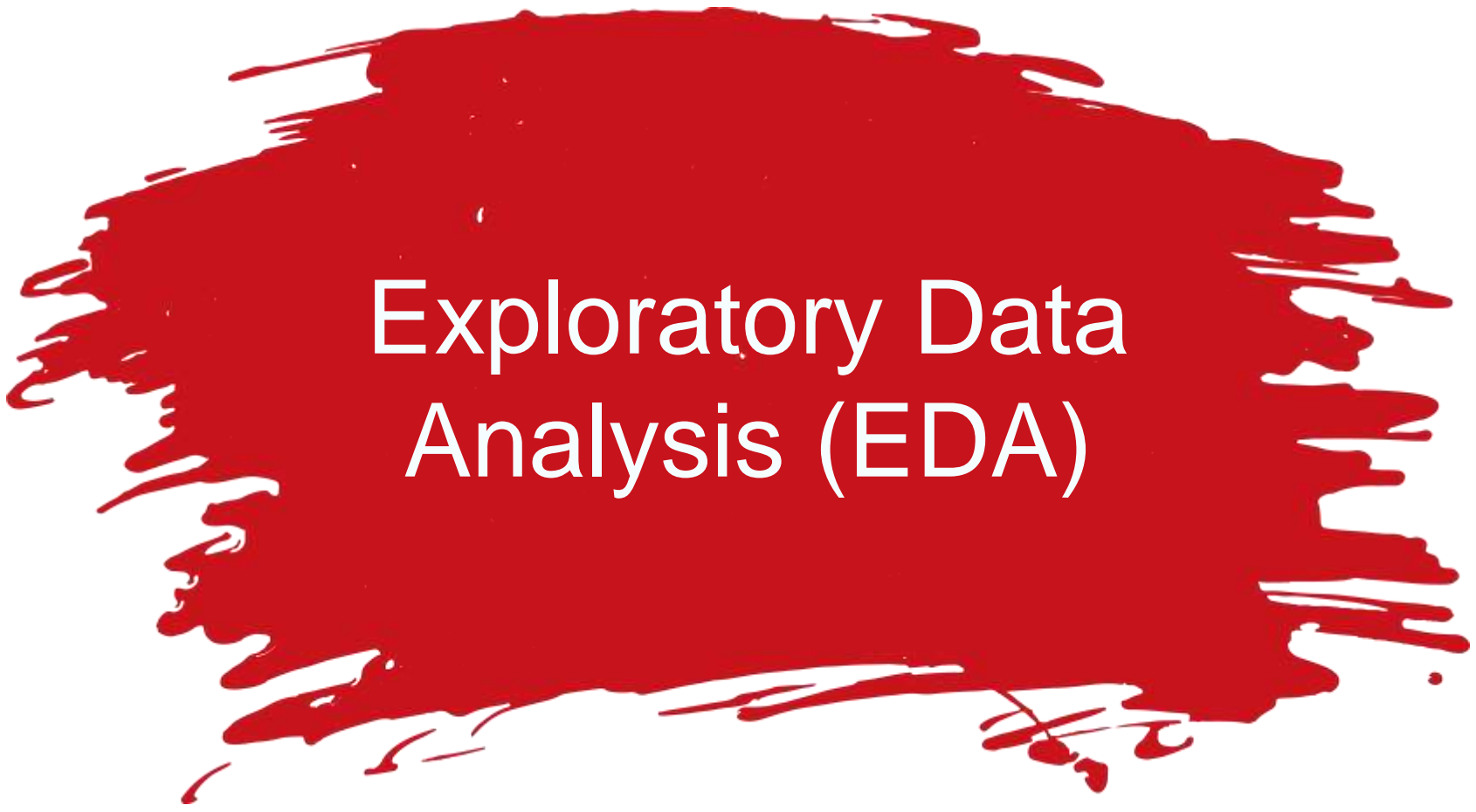


Data Overview

Data Overview:

- Data set obtain from “Netflix Movies and TV Shows Clustering,csv”

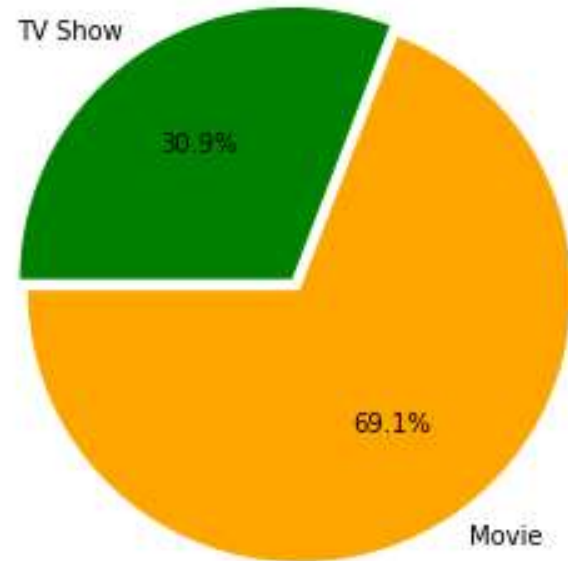
	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	s1	TV Show	3%	NaN	João Miguel, Bianca Comparato, Michel Gomes, R...	Brazil	August 14, 2020	2020	TV-MA	4 Seasons	International TV Shows, TV Dramas, TV Sci-Fi &...	In a future where the elite inhabit 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...



Exploratory Data Analysis (EDA)

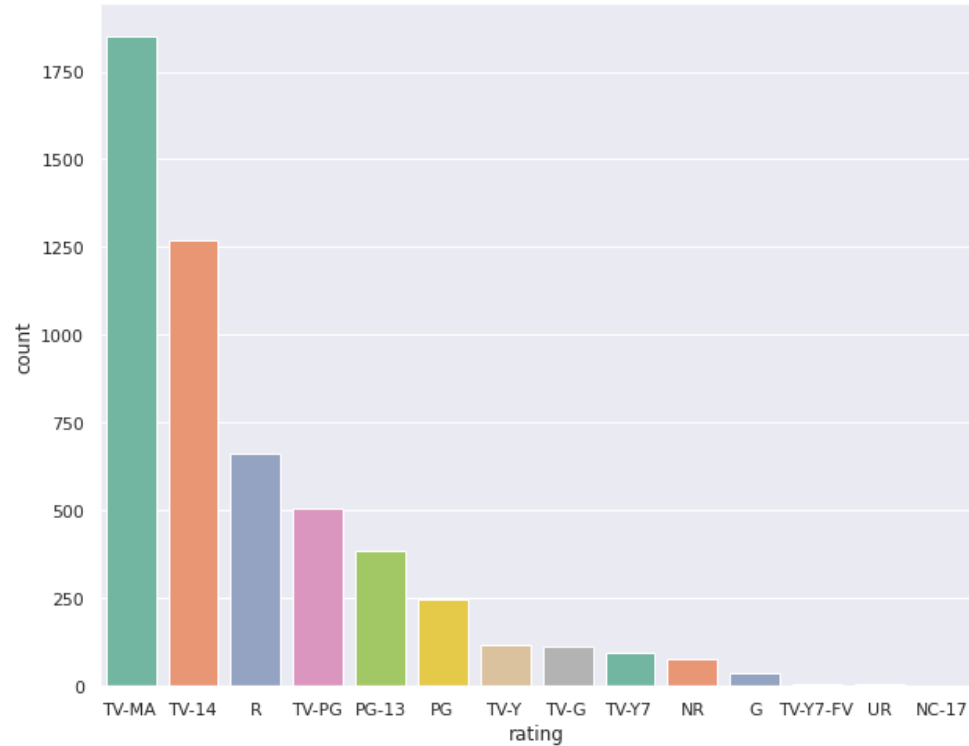
Relation: Movies v/s TV Shows

- Clearly number of Movies on Netflix outnumbered the number of TV Shows.
- Almost 70% content are movies while rest 30% are TV Shows.

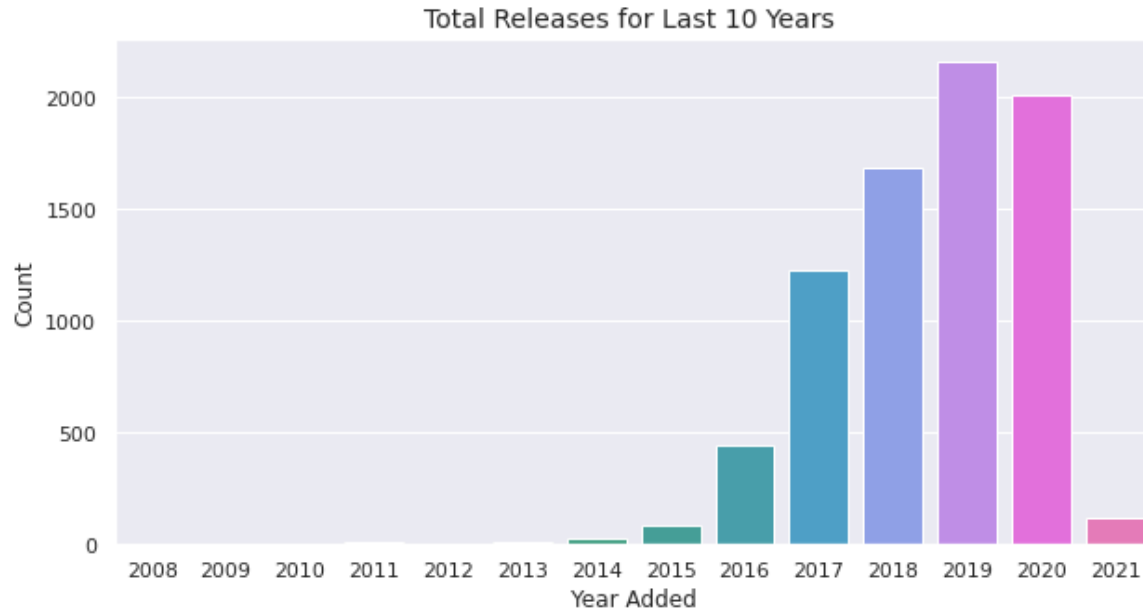


Relation: Rating v/s Count

- **TV-MA, TV-14 and R rating having maximum count with near to 1800, 1250 and 700 in Movies List.**

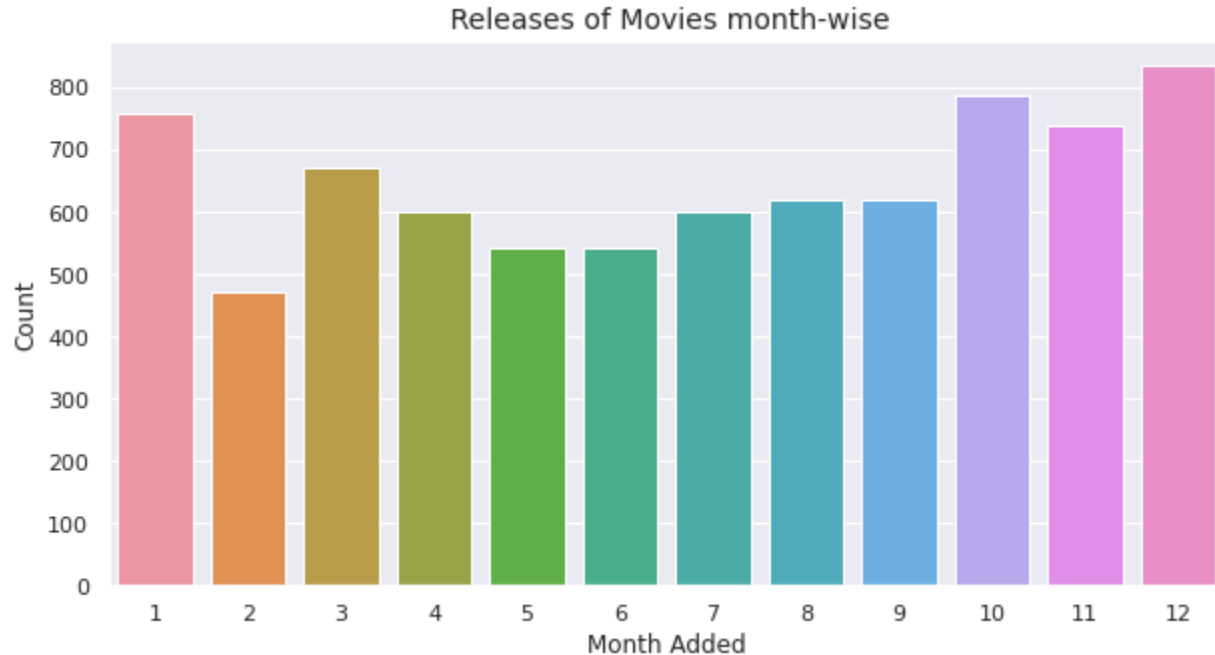


Relation: Year Added v/s Count



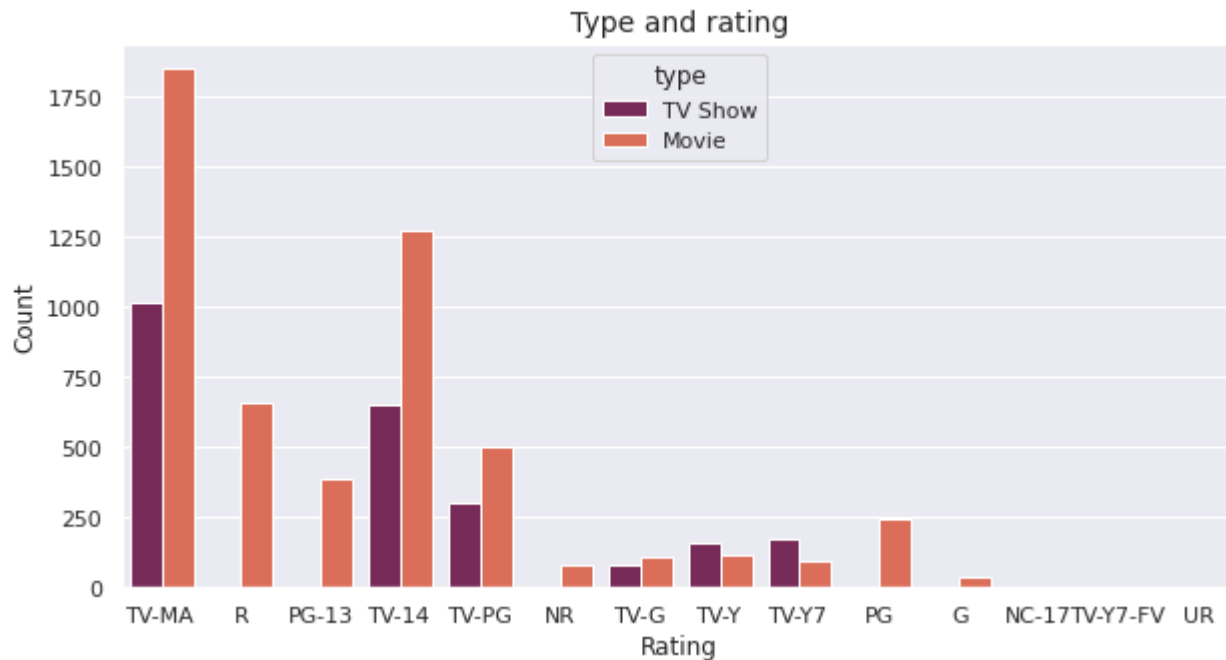
- Maximum released years are 2019, 2020 and 2018 in decreasing order

Relation: Month Added v/s Count



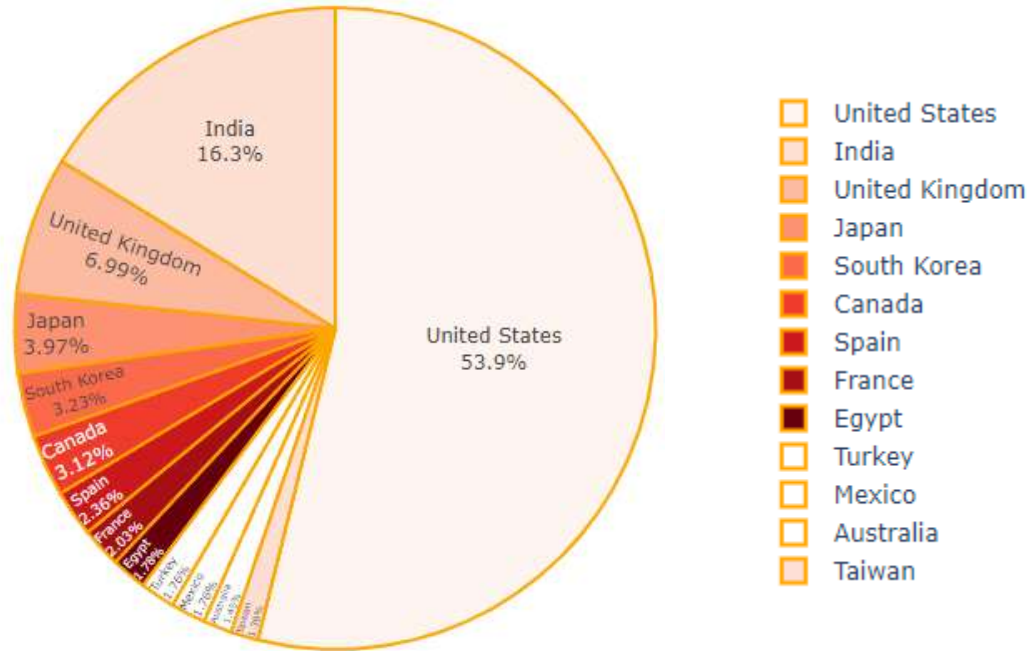
- **Maximum released months are 10, 11, 12 and 1 for Movies**

Relation: TV Shows and Movies v/s Count



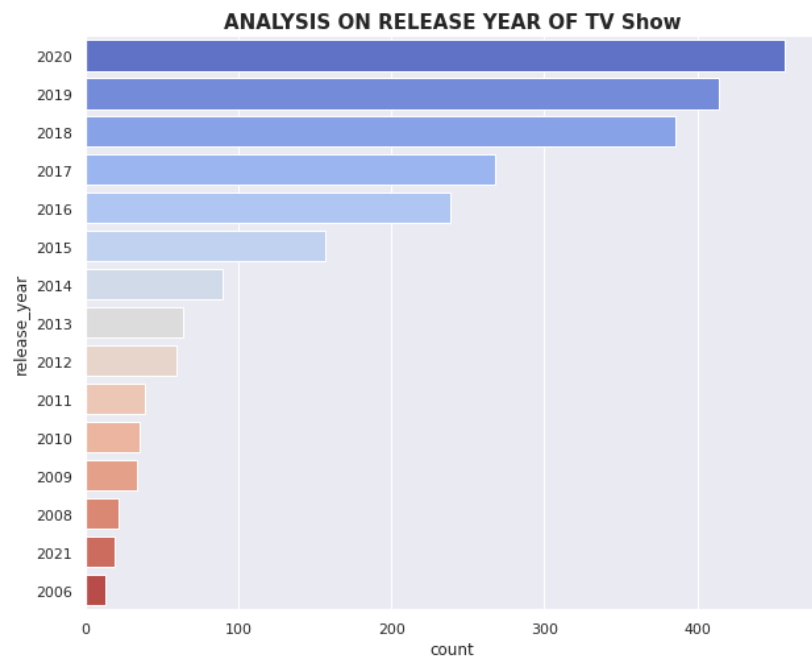
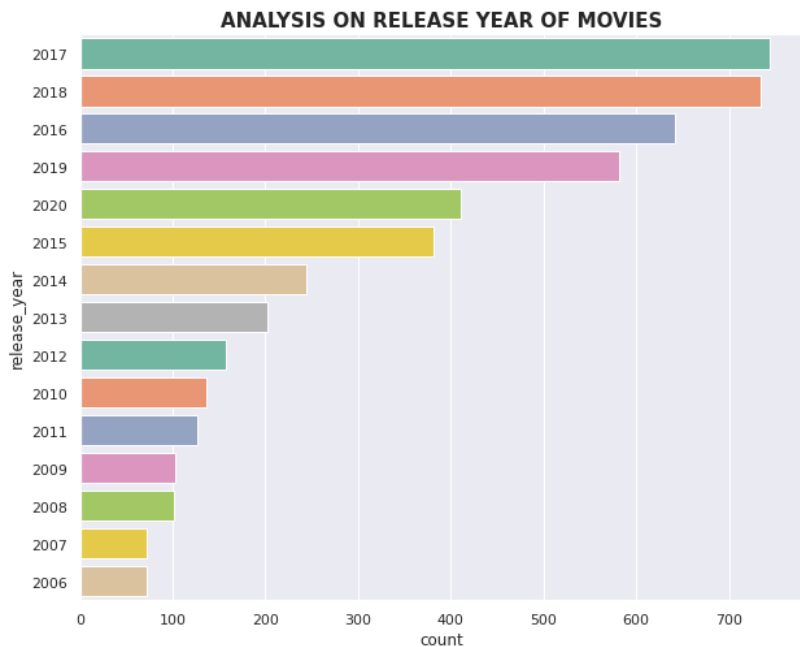
➤ **Rating to TV Show and Movies Count**

Relation: Countries v/s Percentage Count



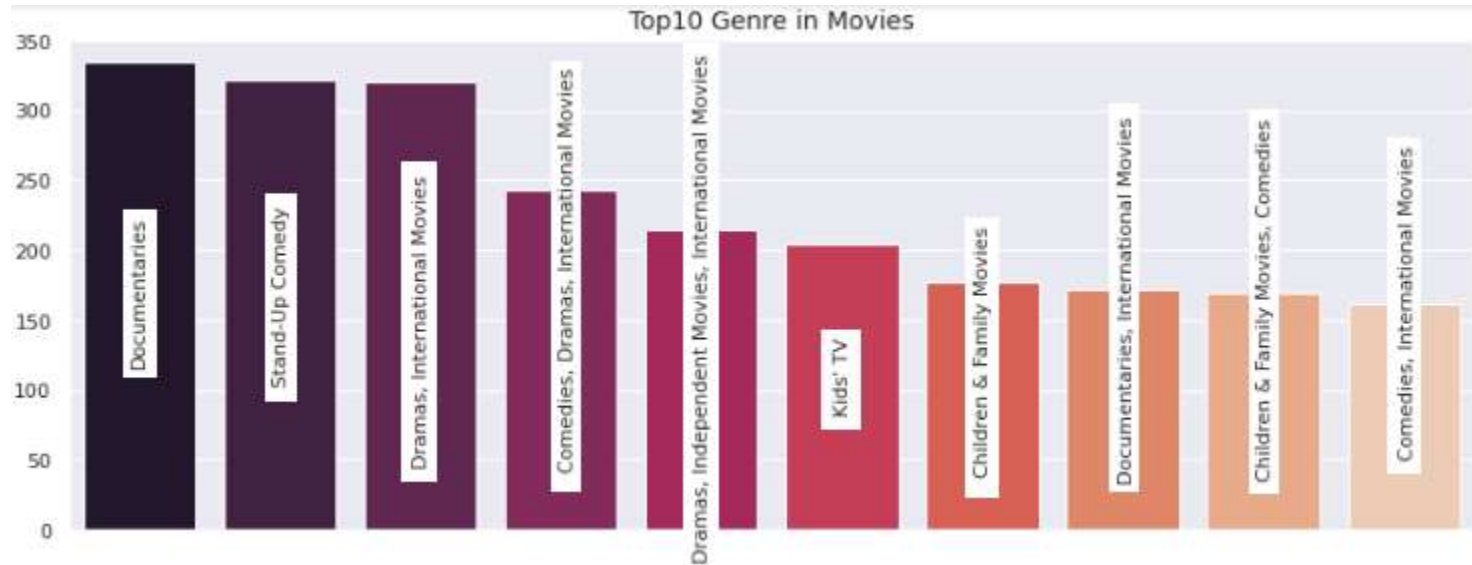
- United State, India and united Kingdom are top three countries produce maximum.

Relation: Movies and TV Shoes v/s Release Year



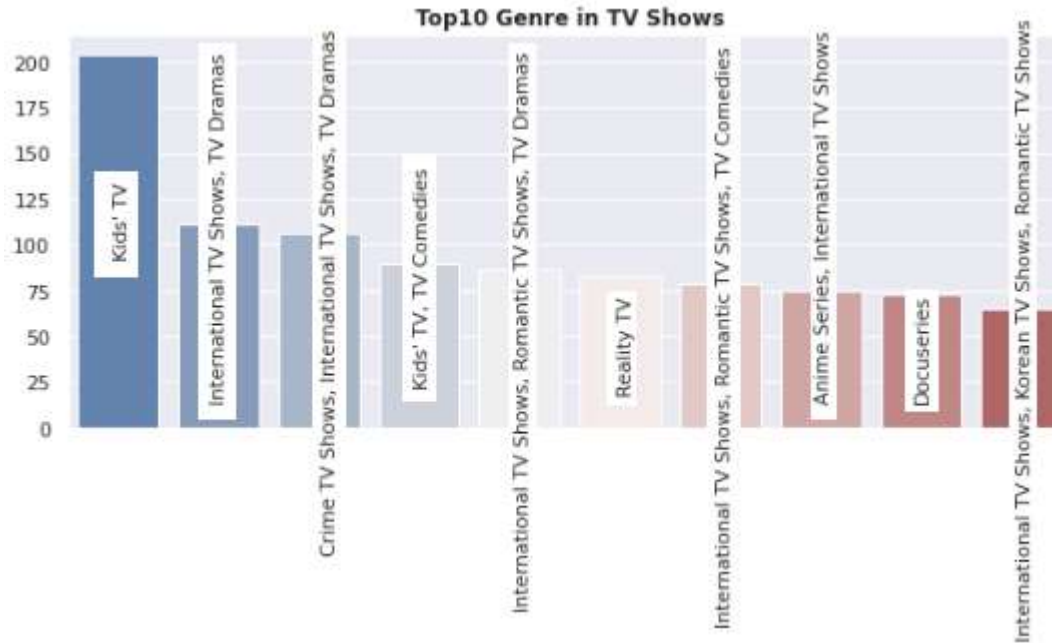
- Movies released maximum in the year “2017” but for TV Shows the year is “2020”.

Relation: genre of Movies v/s Count



- Top three Genre of Movies are “Documentaries”, “Stand-Up Comedy” and “Drama”

Relation: genre of TV Shows v/s Count



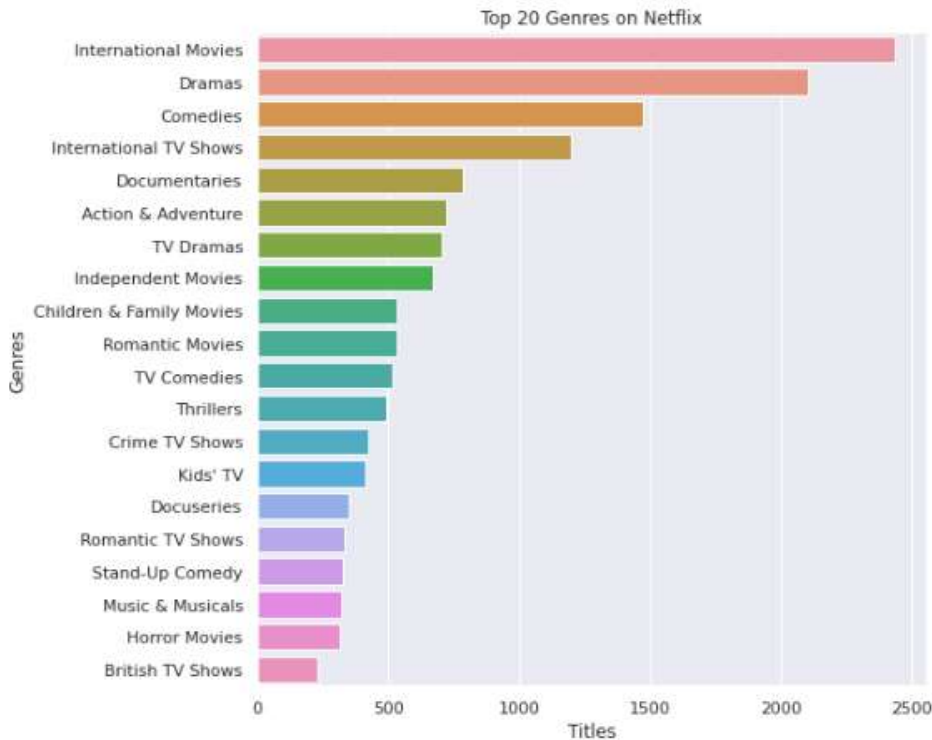
- Top three Genre of TV Shows are “Kid`s TV”, “International TV Shows” and “Crime”

Longest TV Shows Duration List

	title	duration
2538	Grey's Anatomy	16
4438	NCIS	15
5912	Supernatural	15
1471	COMEDIANS of the world	13
1537	Criminal Minds	12
7169	Trailer Park Boys	12
1300	Cheers	11
2678	Heartland	11
1577	Dad's Army	10
1597	Danger Mouse: Classic Collection	10
3592	LEGO Ninjago: Masters of Spinjitzu	10
5538	Shameless (U.S.)	10
5795	Stargate SG-1	10

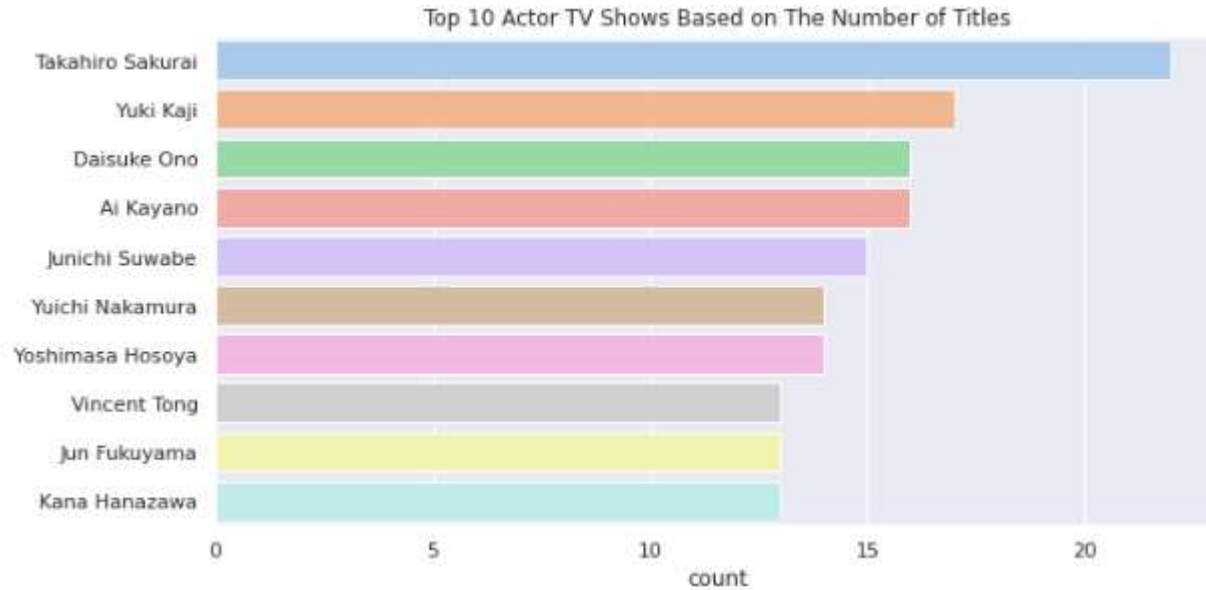
- **“Grey`s Anatomy” is the longest TV Show with duration “16”.**

Relation: genre v/s Count



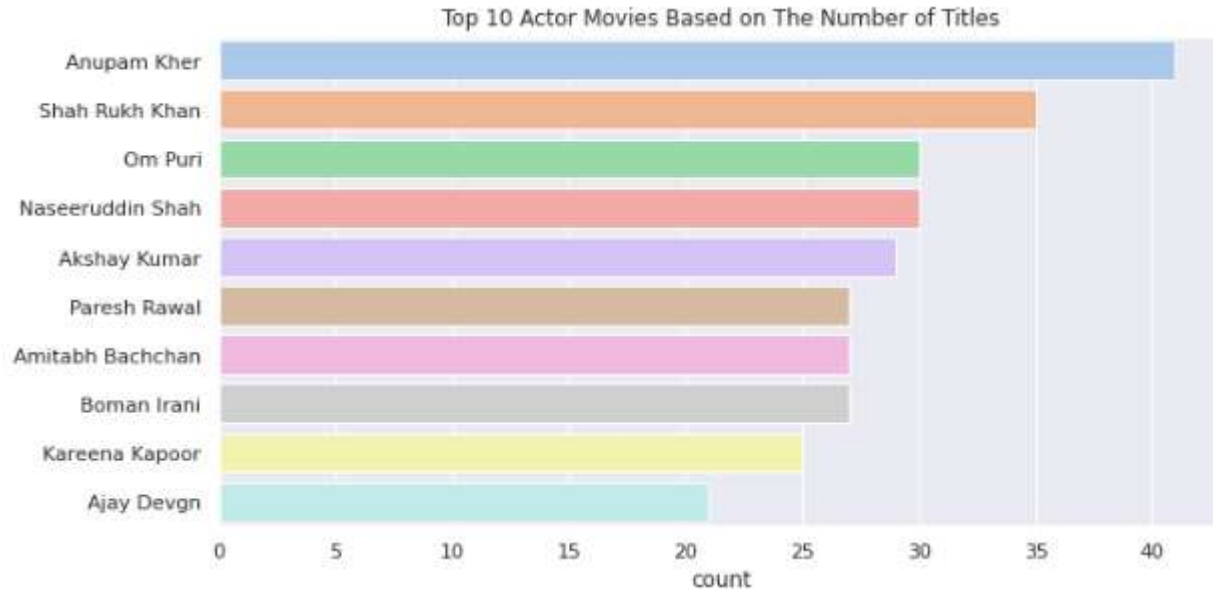
- Top three Genres are “International Movies”, “Dramas” and “Comedies”

Relation: TV Actors v/s Count



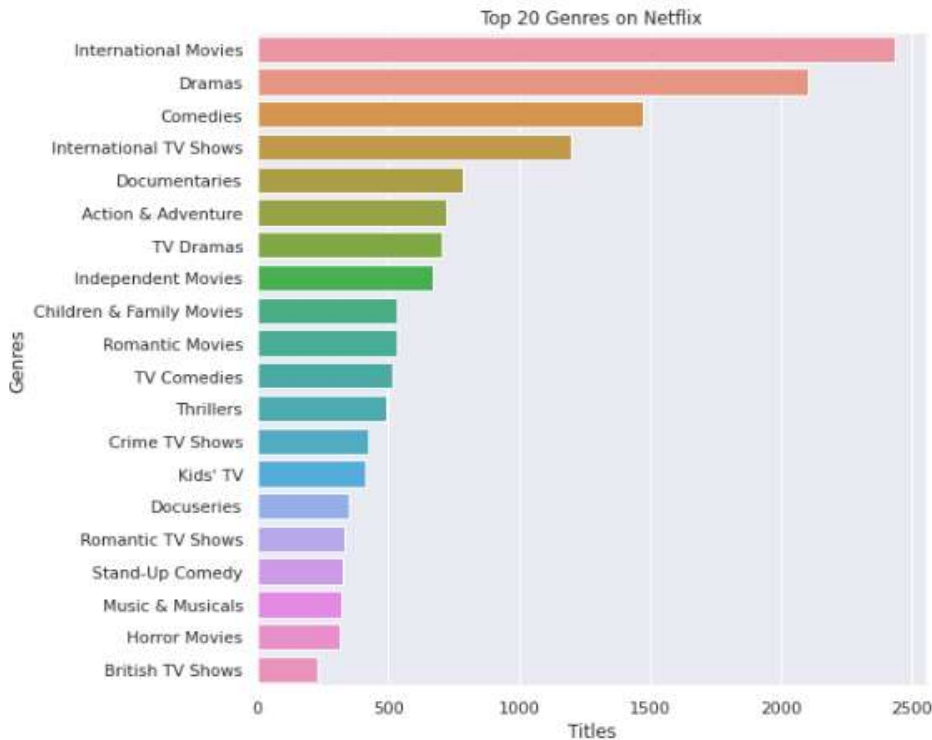
- Top three TV Actors are “Takahiro Sakuai”, “Yuki Kaji” and “Daisuke Ono”

Relation: Movie Actors v/s Count



- Top three Movie Actors are “Anupam Kher”, “Shah Rukh Khan” and “Om Puri”

Relation: genre v/s Count



- Top three Genres are “International Movies”, “Dramas” and “Comedies”



Applying Model Clustering

K-Means Clustering:

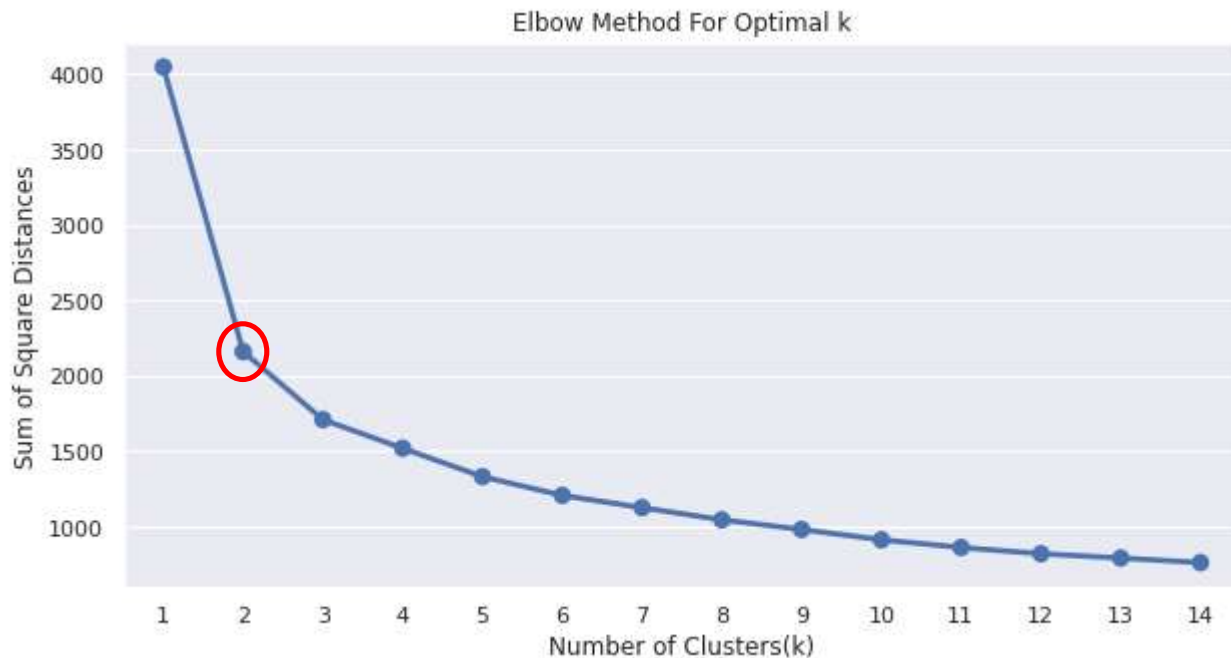
The **K-Means** algorithm searches for a predetermined number of clusters within an unlabelled multidimensional dataset.

The **Elbow Method** is one of the most popular methods to determine this optimal value of k number of clusters.

To determine the optimal number of clusters, we have to select the value of k at the "elbow" i.e. the point after which the distortion/inertia start decreasing in a linear fashion.

Thus from this chart we need to check, which would be the best number of clusters from 2,3,4,5, and 7.

We found elbow formation at $k=2$.



Silhouette Score for K-Means:

- Silhouette score 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.
- The Silhouette score is calculated for each sample of different clusters.

- For n_clusters = **2**, silhouette score is **0.2656**045956669093
- For n_clusters = **3**, silhouette score is **0.2153**8517767329393
- For n_clusters = **4**, silhouette score is **0.1729**6314851287742
- For n_clusters = **5**, silhouette score is **0.1900**239619775974
- For n_clusters = **6**, silhouette score is **0.1595**8106090265092
- For n_clusters = **7**, silhouette score is **0.1500**915091981548
- For n_clusters = **8**, silhouette score is **0.1325**983580463935

A large, dark, textured brushstroke background, resembling a thick application of black paint with visible bristles and some lighter areas, creating a sense of movement and depth.

Conclusion

- We started this project with the intention to obtain some useful insights related to the type of Netflix content. For this, we performed exploratory data analysis on our data after cleaning and making it easy to analyze. This analysis helped us to understand the data.
 - We found that most of the content on Netflix are TV-MA and TV-14
 - USA and India are two countries producing the maximum number of content
 - Documentaries and Stand-Up are top genre in terms of number of contents they have on platform. Further we found number of TV-Shows on Netflix outnumbered Movies
- Our next job was to make an unsupervised clustering model. For this, we processed our text by removing useless characters like – stopwords, punctuation and did stemming. After getting the length for each text feature we rescaled them for generalization and started applying algorithms.
- We first used K-means clustering. In order to find appropriate cluster number we used elbow method and finally got the best silhouette score of around 0.26. Next, we applied Hierarchical Agglomerative Clustering for which we made dendrogram.

Q & A