Capstone Project -4

Netflix Movies and TV Shows Clustering

Submitted by

SANJU KHANRA

Data science trainee, Alma better

Content :

- Introduction
- Abstract
- Problem Statement
- Data Summary
- Data Cleaning
- Data processing
- Exploratory Data Analysis (EDA)
- Hypothesis testing
- Feature Engineering
- Dimensionality Reduction
- Clustering
- Word cloud on Clusters
- Build Recommendation System
- Conclusions

Introduction :

NETFLIX

Netflix is a company that manages a large collection of TV shows and movies, streaming it anytime via online. This business is profitable because users make a monthly payment to access the platform. However, customers can cancel their subscriptions at any time.

METHODOLOGY

Unsupervised Machine Learning (Clustering)

DATABASE

Netflix Movies and TV Shows
7787 rows and 12 columns
Data from last decade

Abstract :

- Netflix is a popular streaming service and production firm.
- According to Statistics, Netflix had approximately **220.67** million paid subscribers world wide as of the second quarter of the shows that are hosted on their platform in order to enhance the user experience for its subscribers.

Problem Statement :

This dataset consists of TV shows and movies available on Netflix as of 2019. The dataset is collected from able 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.

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

Points to discuss

- Data description
- Exploratory data analysis
- Hypothesis testing
- Feature selection
- *Machine learning algorithms(unsupervised)*
- 1. K-mean
- 2.agglomerative clustering
- Model performance

Data Summary :

- **The dataset consists of listings of all the movies and TV shows available on Netflix, along with details such as – cast, directors, ratings, release year, duration.**

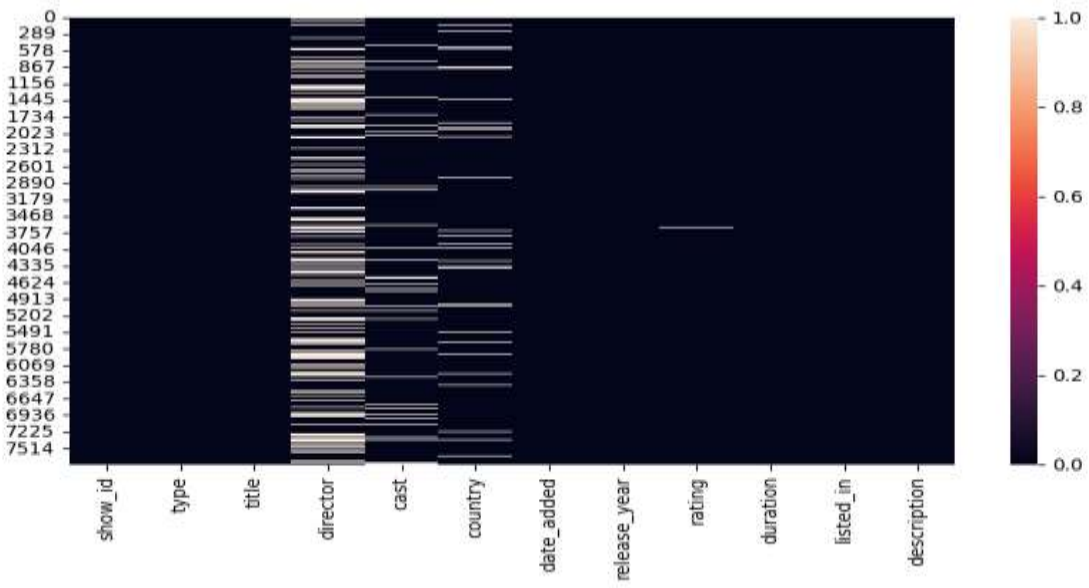
- **Show id** : Unique ID for every Movie / TV Show
- **type** : Identifier – A Movie or TV Show
- **title** : Title of the Movie / TV Show
- **director** : Director of the Movie
- **cast** : Actors involved in the movie / show
- **country** : Country where the movie / show was produced
- **Date added** : 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

Data Cleaning :

Handling Missing values:

- Director(2389),cast(718),and country(507)–replace with **'Unknown'**
- Date added(10)–**dropped**.
- Rating(7)–**mode** imputation.

We have successfully handled all the missing values in the dataset

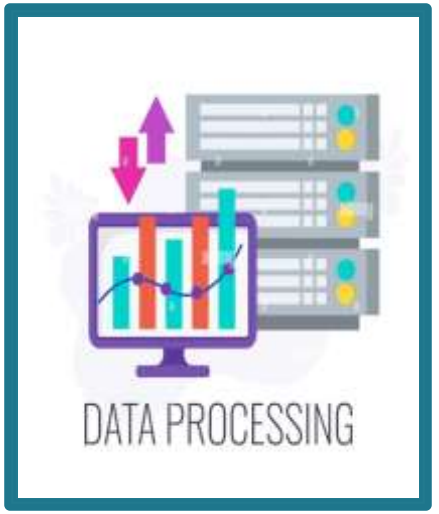


Null Values

Show-id	0
type	0
title	0
director	2389
cast	718
country	507
Date-added	10
Release-year	0
rating	7
duration	0
listed-in	0
description	0

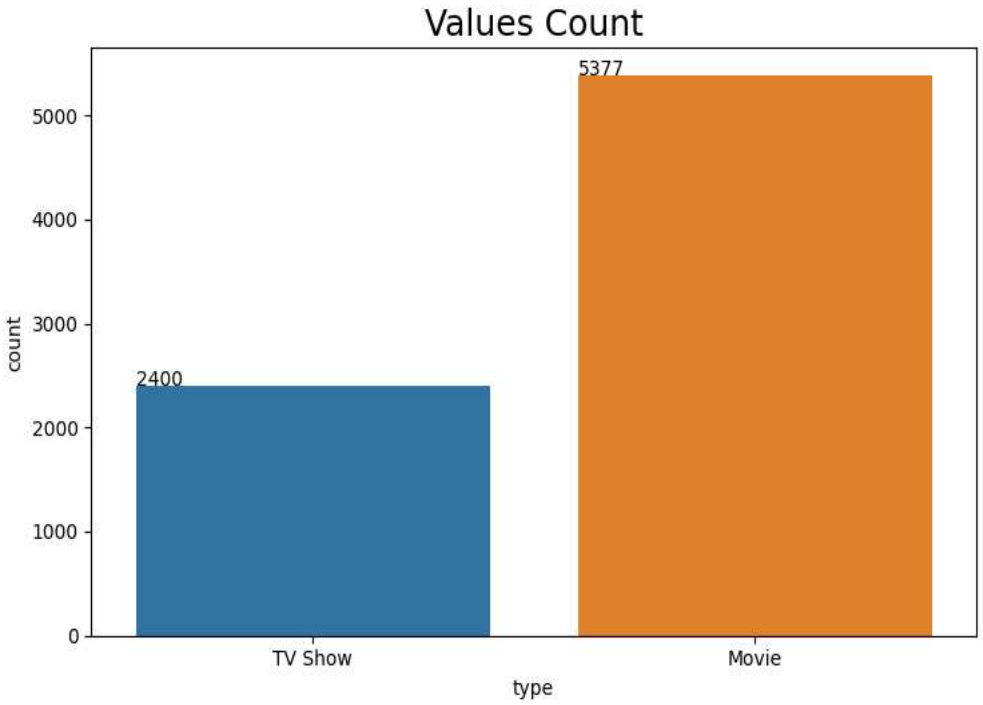
Data processing :

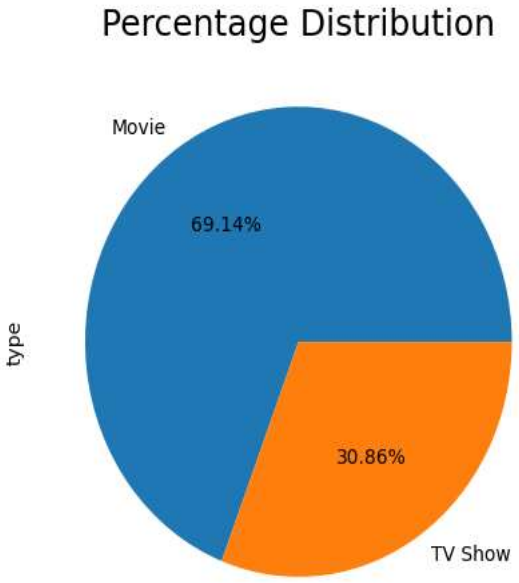
- **Country, listed in:**
  - There are some movies / TV shows that were filmed in multiple countries, have multiple genres associated with it.
  - To simplify the analysis, let's consider only the primary country where that respective movie / TV show was filmed.
  - Also, let's consider only the primary genre of the respective movie / TV show.
- **Typecasting 'duration' from string to integer**
  - Converted the data-type of duration column to INT use apply lambda function.
- **Typecasting 'date-added' from string to date time**
  - The shows were added on Netflix between 1st January 2008 and 16th January 2021, and data-added column year or month divided.
- **The ratings can be changed to age restrictions that apply on certain movies and TV shows.**
  - The data set contained separate age ratings for movies and TV shows and were replaced with values of:'Adults','Teens','YoungAdults','OlderKids','Kids'



EDA(Type column) :

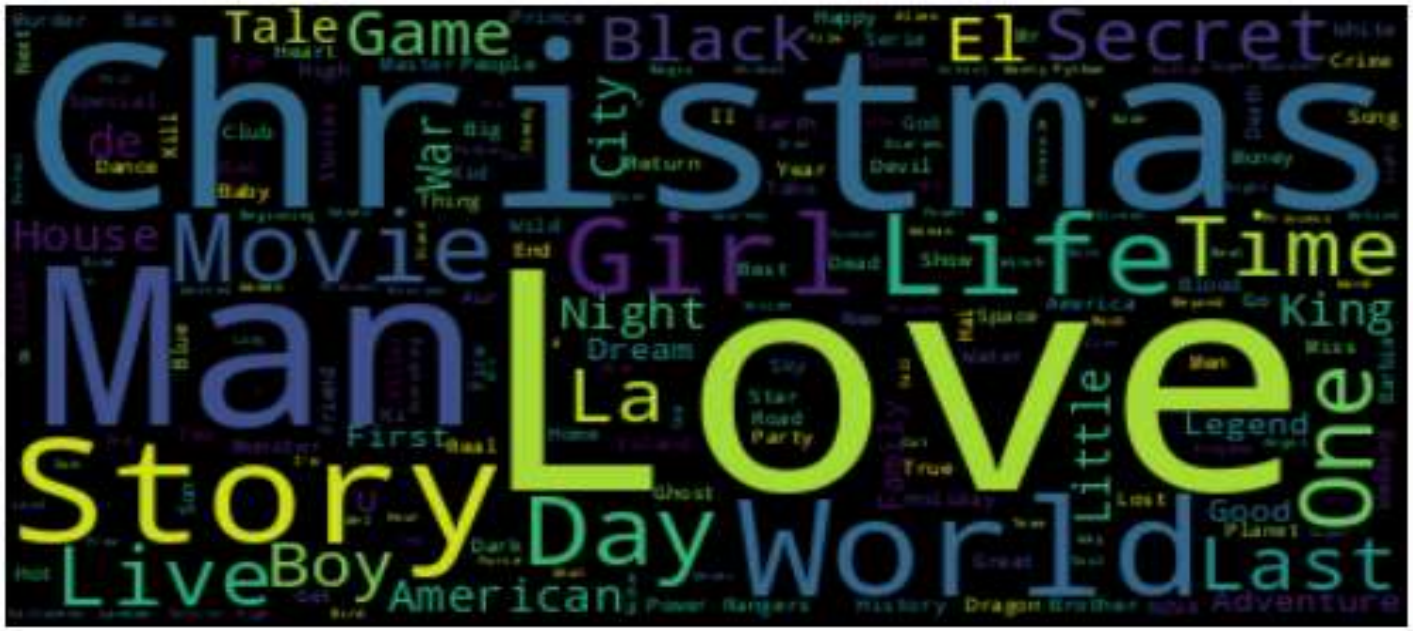
- Netflix has 5372 movies and 2398 TV shows, there are more number movies on Netflix than TV shows. 69% of data belong from Movie class and 31% of data belong from TV shows, Greater number of count belong from movie class than TV show class.





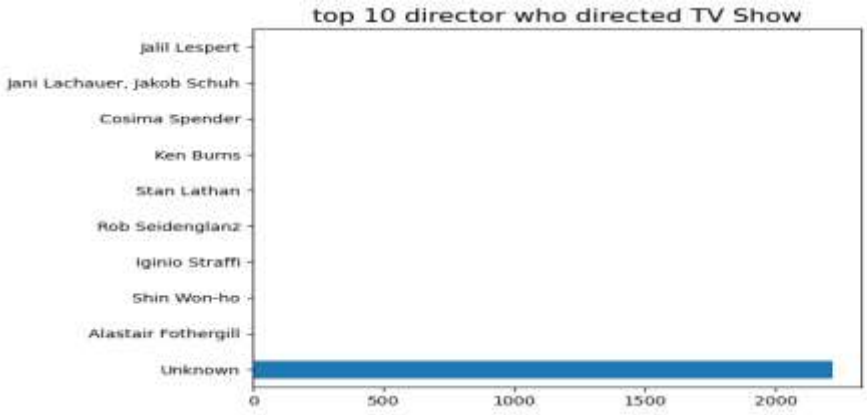
EDA(Title column) :

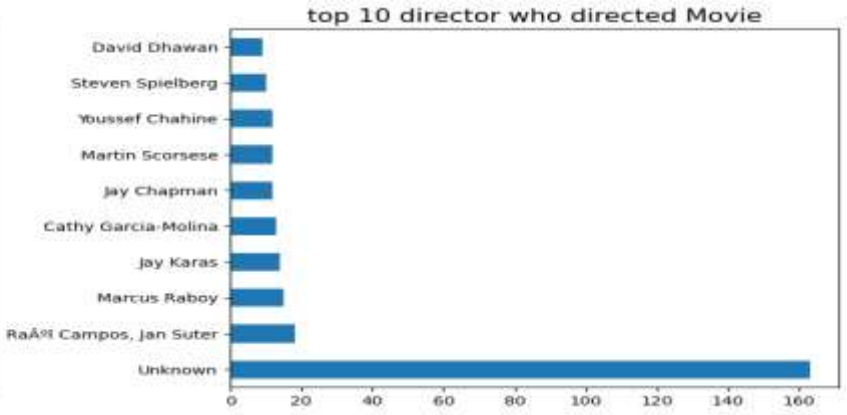
- IM-show(BW) displays the binary image BW in a figure Title column text visualization we can see word like ‘Christmas’, ‘Love’, ‘World’, ‘Man’, ‘Story’ are very common word which are appear most of the time in movie title column.



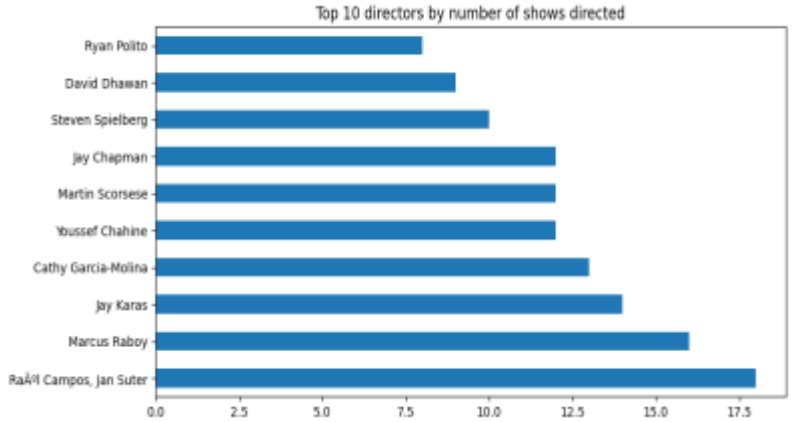
EDA(Director column) :

- In Director column values count after we can see number of movie directed by director is 2400, number of TV show directed by director is 5377





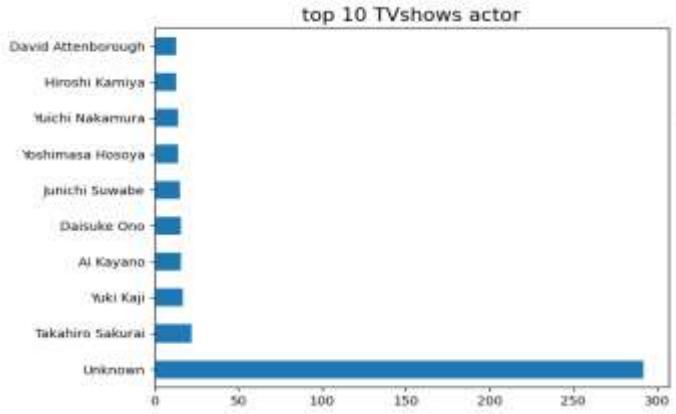
TV & movies unknown Top10 director shows

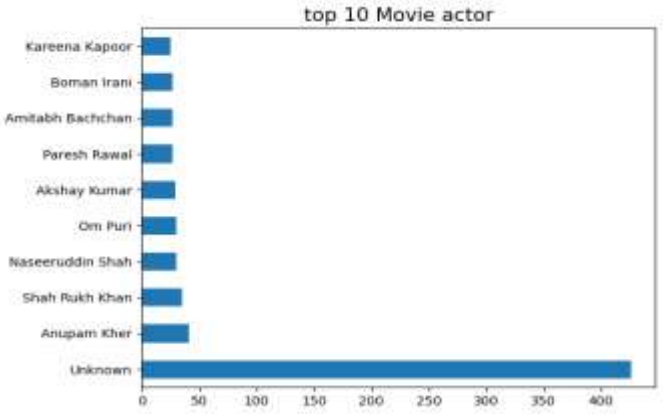


Raul Campos and Jan-Suter together have directed 18 movies / TV shows, higher than anyone in the dataset.

EDA(Cast column):

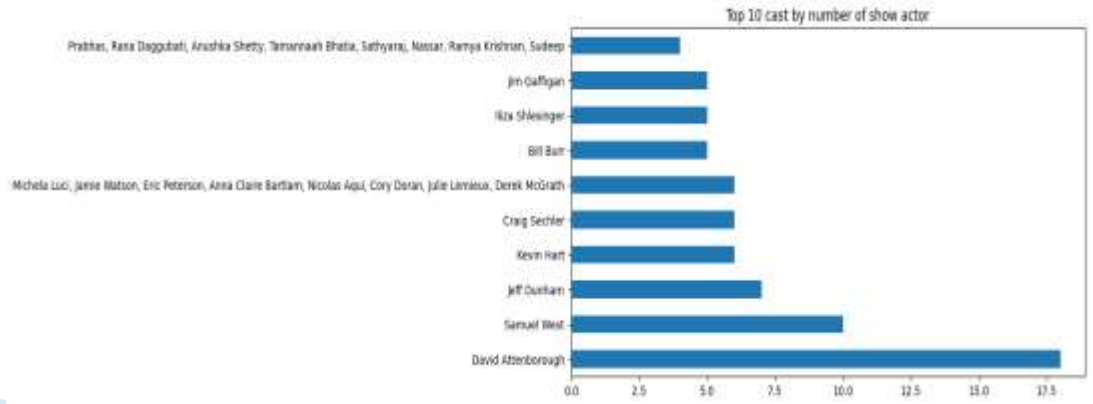
- In Cast column number of TV Shows actor13539 number of Movie actor23050,lets check Top10 TV show and Movies Actor





Takahiro Sakurai, Yuki Kaji, play highest role in the TV shows.
Anupam Kher, Shah rukh Khan, play highest number of role in the movies.

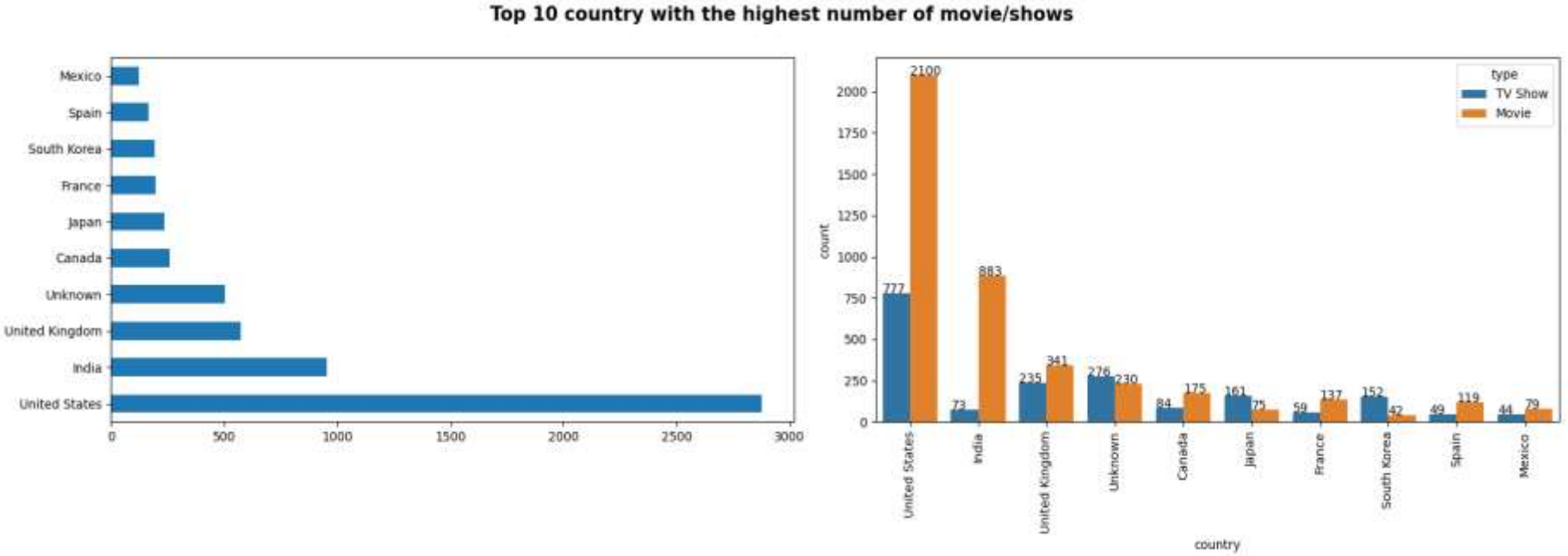
Cast Top10 Unknown actors



Unknown cast actor David Attenboroug, Samuel West play highest role shows or movies

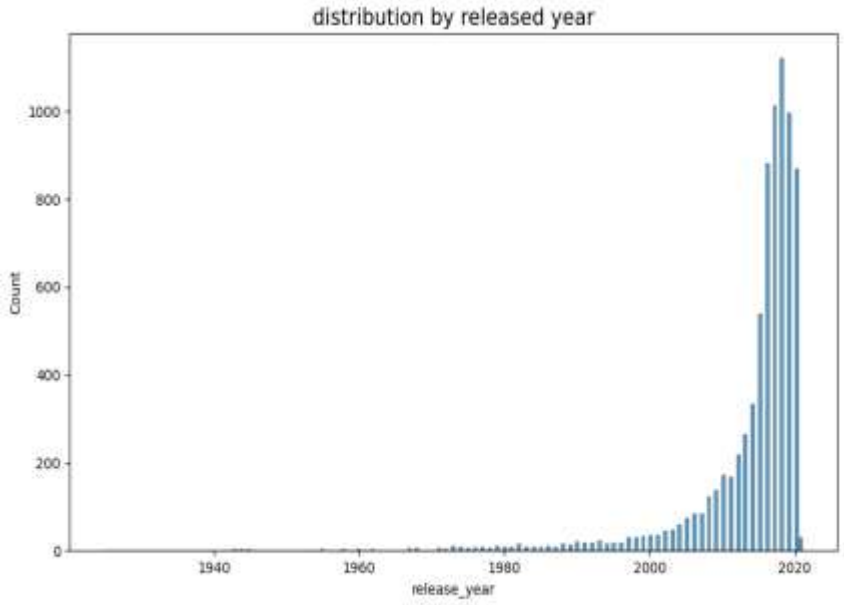
EDA(Country column) :

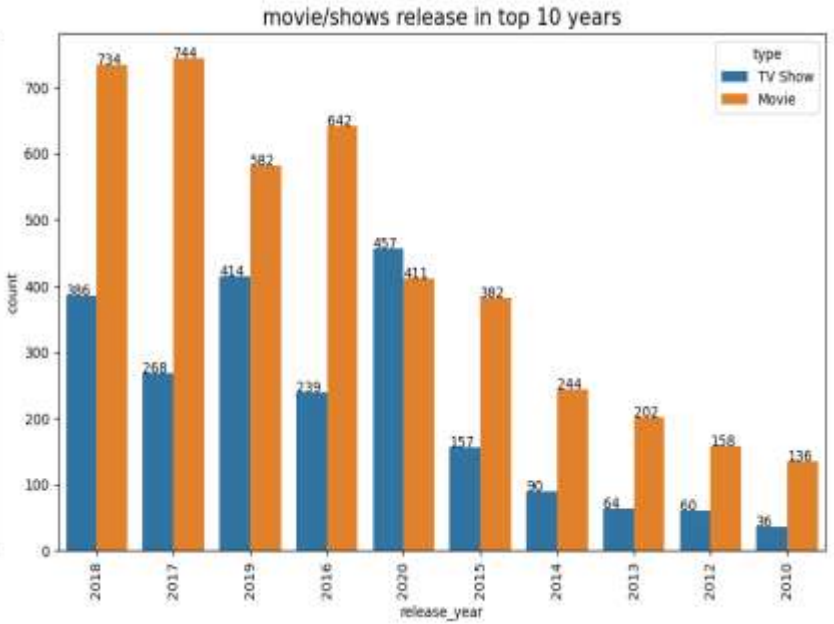
- The top10 country values count of number analysis TV show and Movies ,Than after we can see The highest number of movies / TV shows were based out of the US, followed by India and UK.



EDA(Release-Year column) :

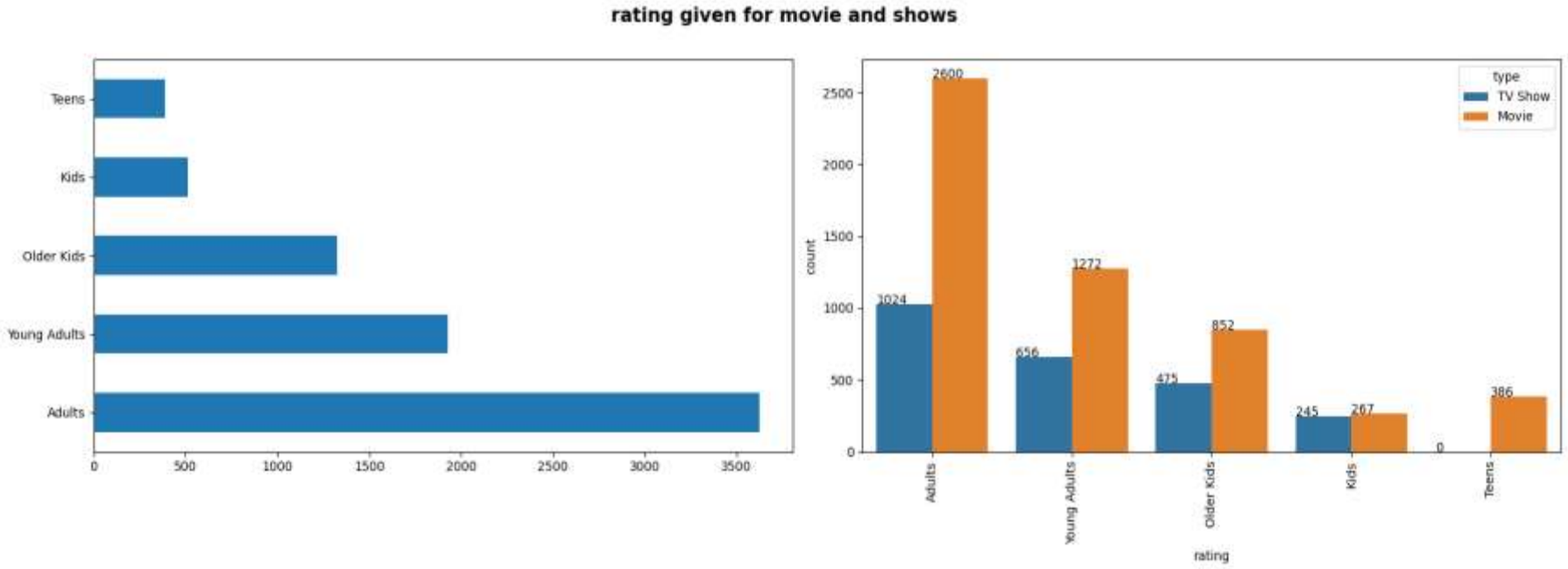
- In this dataset release-year column oldest record of movie/show release year :1925 latest record of movie/show release year : 2021
- Count Top10 movie\shows release year
- 1. Netflix has greater number of new movies / TV shows than the old ones.
- 2. Highest number of movie/shows are released in Netflix in between 2015-2020 and highest number of count belong from 2018 year.





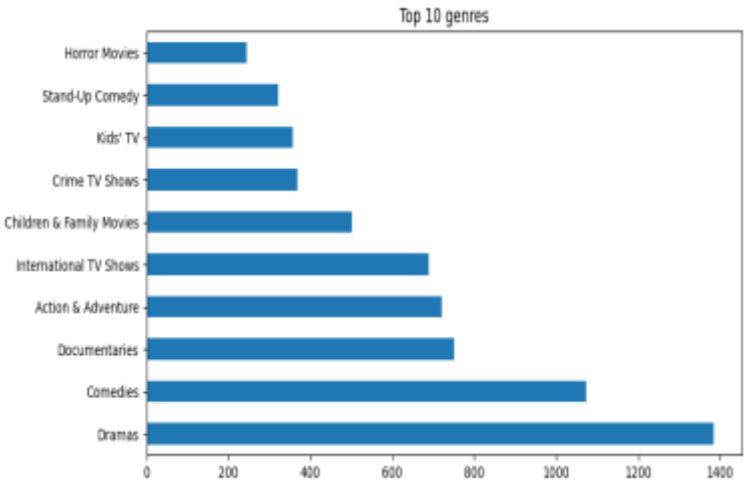
EDA(Rating column) :

- Analyze by the rating given for movies and TV shows than analyze after, **most of the movie and TV shows have rating of Adults (Mature Audience) then followed by Teens (younger audience).**
- **The highest number of rating given for the movies as compared to TV shows it is pretty obvious because of highest number of categories belonging from movie class as we can seen earlier in type column.**

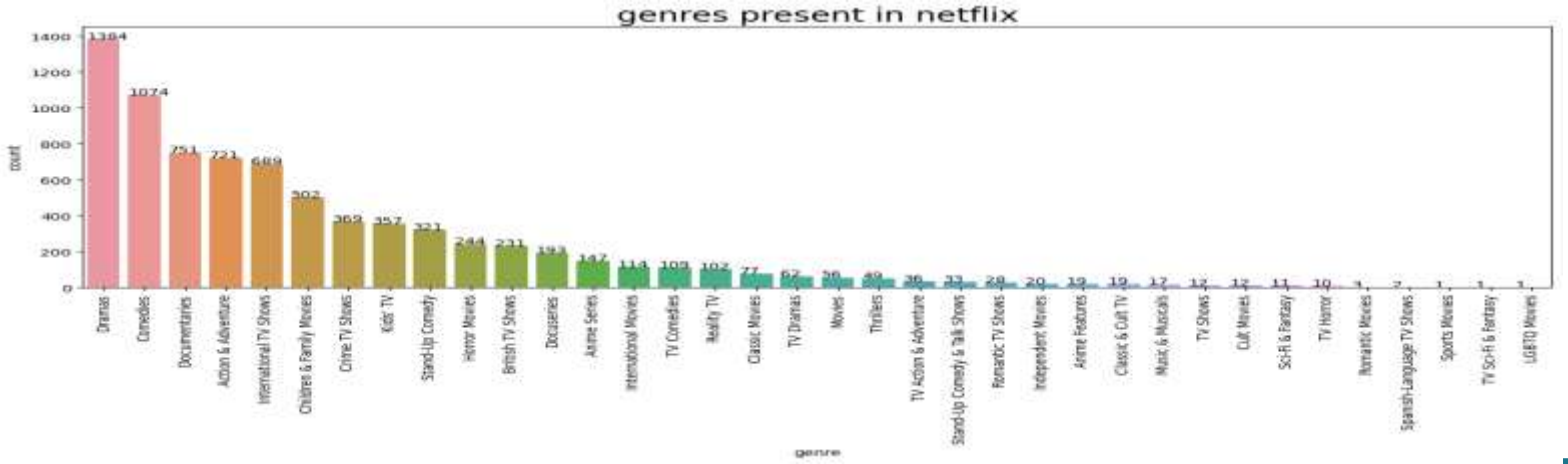


EDA(Listed-In column (Top10 Genres)):

- The dramas is the most popular genre followed by comedies and Documentaries, This value increases to about 82.44% for top 10 genres.



	genres	count
0	Dramas	1384
1	Comedies	1074
2	Documentaries	751
3	Action & Adventure	721
4	International TV Shows	689
5	Children & Family Movies	502
6	Crime TV Shows	369
7	Kids' TV	357
8	Stand-Up Comedy	321
9	Horror Movies	244

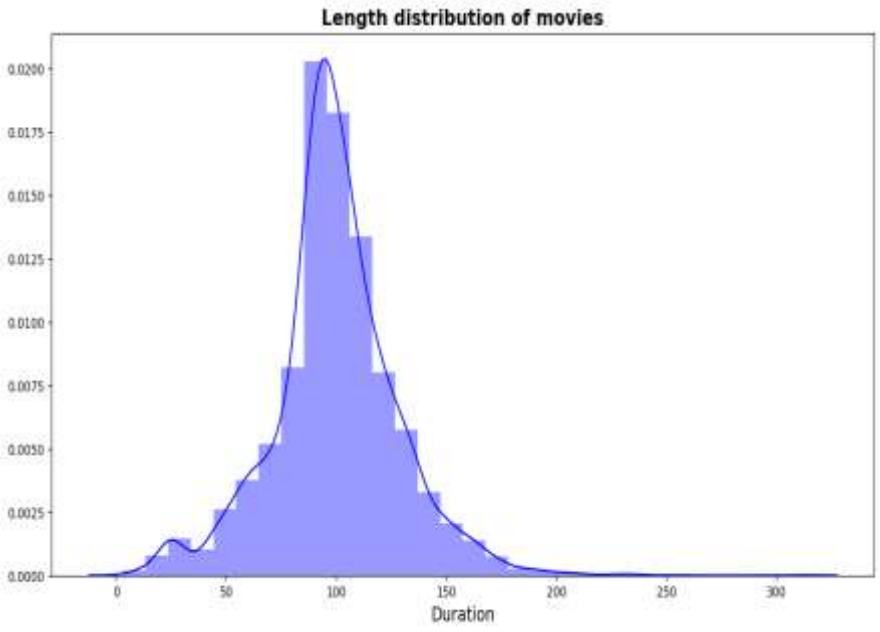


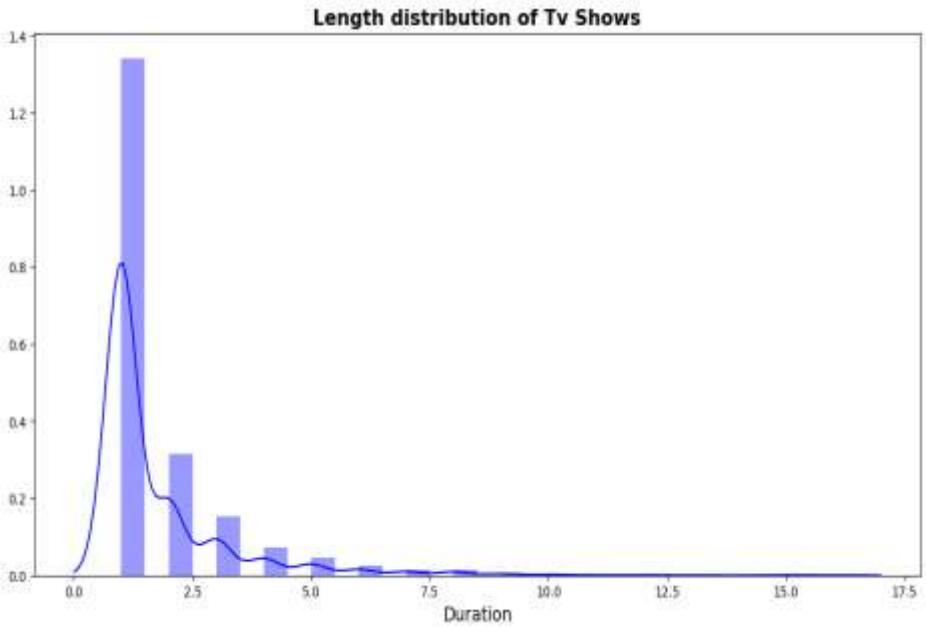
Highest number of genre belong from Dramas, Comedies respectively.
Least number of genre belong from LGBTQ Movies , TV or Sci-Fi & Fantasy TV Shows

EDA(Length of duration in Movies & TV Shows):

- ## Length of duration in Movies ?
  - length of distribution movie 'Black Mirror: Bandersnatch' this duration is 312min This is the longest duration movie and the second longest movies is 'The school of Mischife' duration is 253min

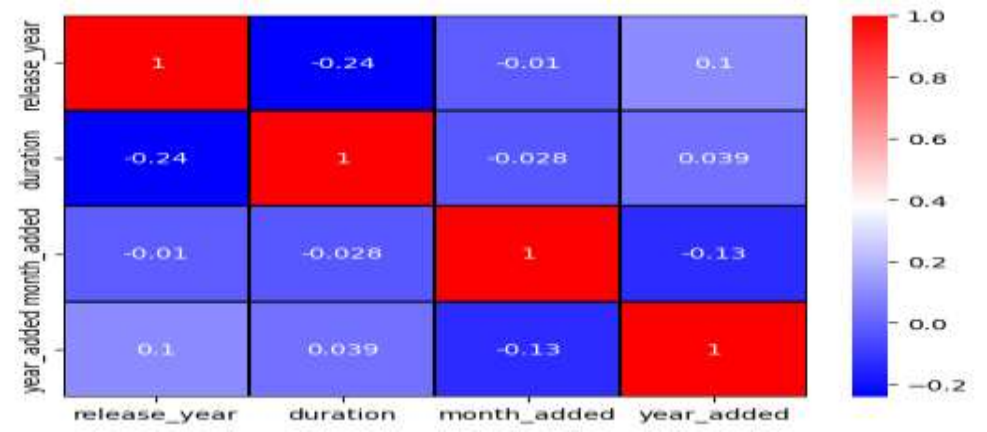
- ## Length of duration in TV Shows ?
  - In TV shows 'Grey's Anatomy' is the longest duration TV show On Netflix This duration is 16min or the second longest TV show is 'NCIS' duration 15min.



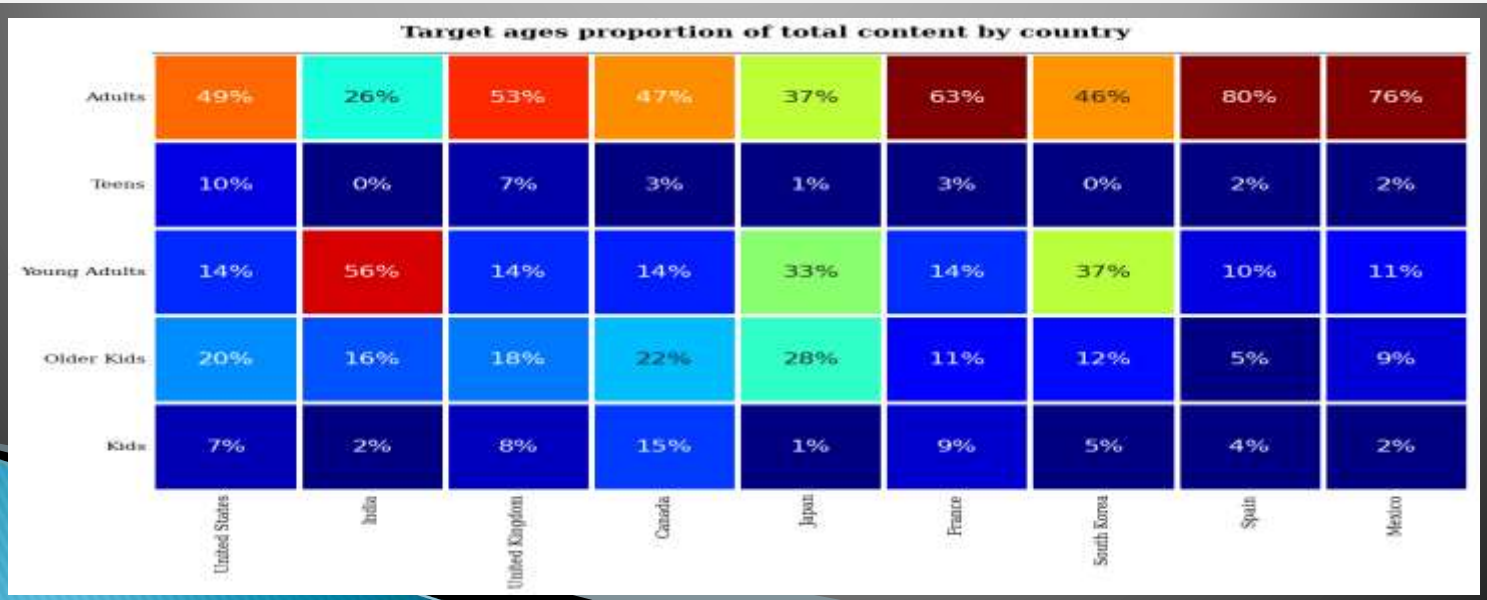


Correlation Heat-map(EDA):

- In this heat-map we can see that release-year or year-added are highly correlated with each other We see a negatively correlation between release-year or duration temperature Release-year and duration negatively correlated

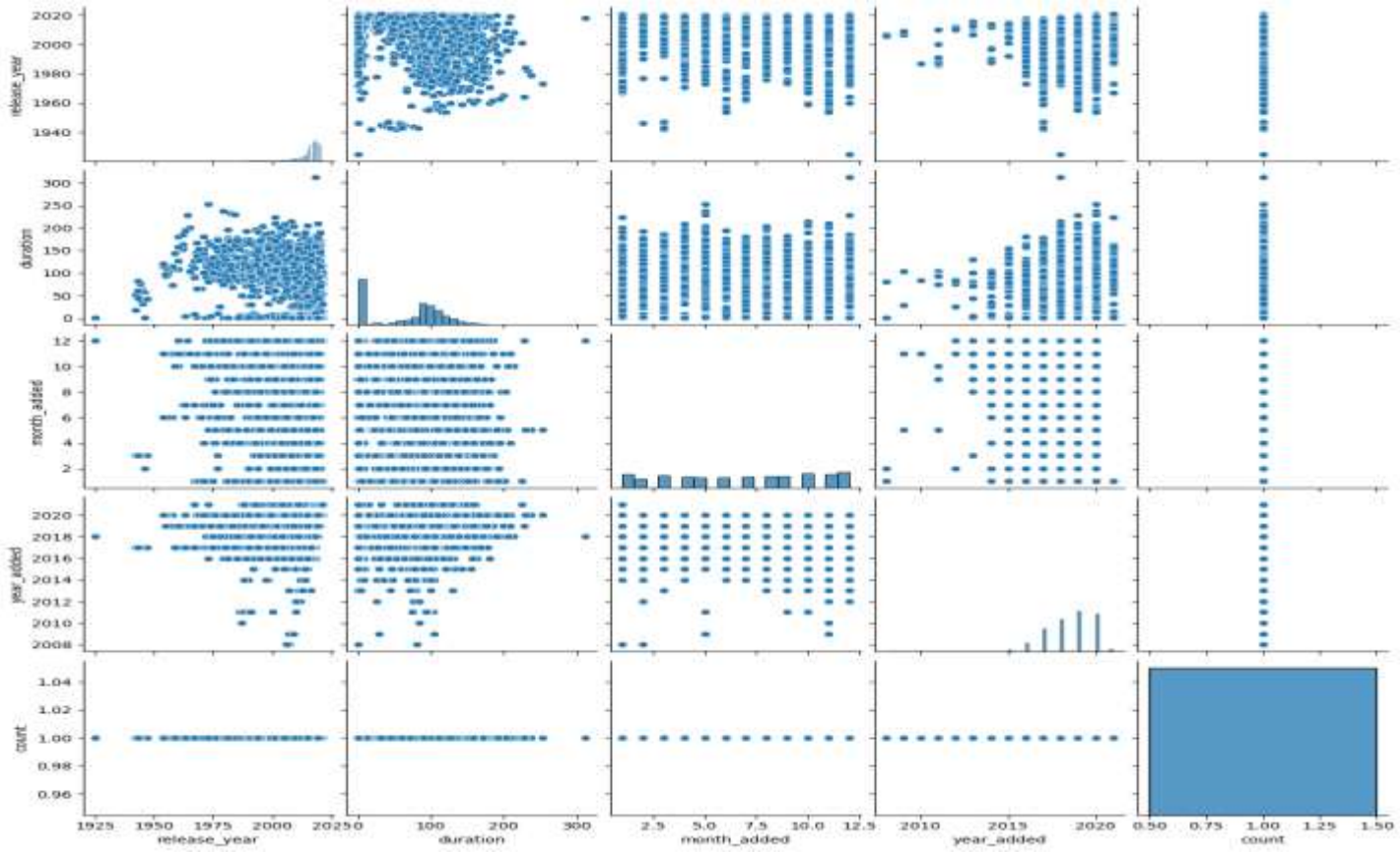


- It is also interesting to see parallels between culturally comparable nations – the US and UK are closely aligned with their Netflix target ages, but radically different from, example, India or Japan!
- Also, Mexico and Spain have similar content on Netflix for different age groups



EDA(Pair Plot) :

- In pair plot release-year or duration is highly pair.
- Or release-year and count good performance compare to month-added and year-add count, month-added and year-add count both are average



Hypothesis Testing :

- Hypothesis testing in statistics refers to analyzing an assumption about a population parameter.

HO : movies rated for kids and older kids are at least two hours long
H1:movies rated for kids and older kids are not at least two hours long.

	Target-ages	duration
0	Kids	66.486891
1	Older-kids	92.024648
2	Teens	107.772021
3	Adults	98.230769

Mean for movies rated for Kids duration 66.486891
Mean for movies rated for older kids duration 92.024648
Standard deviation for movies rated for Older Kids duration 31.182577
Standard deviation for movies rated for kids duration 31.739465

- To perform this method, we first formulate the Null and Alternate Hypotheses.
- The P-value method is used in Hypothesis Testing to check the significance of the given Null Hypothesis. Then, deciding to reject or support it is based upon the specified significance level or threshold.
- **the t-value is not in the range, the null hypothesis is rejected.**
- **As a result, movies rated for kids and older kids are not at least two hours long.**

Hypothesis Testing :

2. H1:The duration which is more than 90mins are movies
HO : The duration which is more than 90mins are NOT movies

	Target-ages	duration
0	Kids	66.486891
1	Older Kids	92.024648
2	Teens	107.772021
3	Adults	98.230769

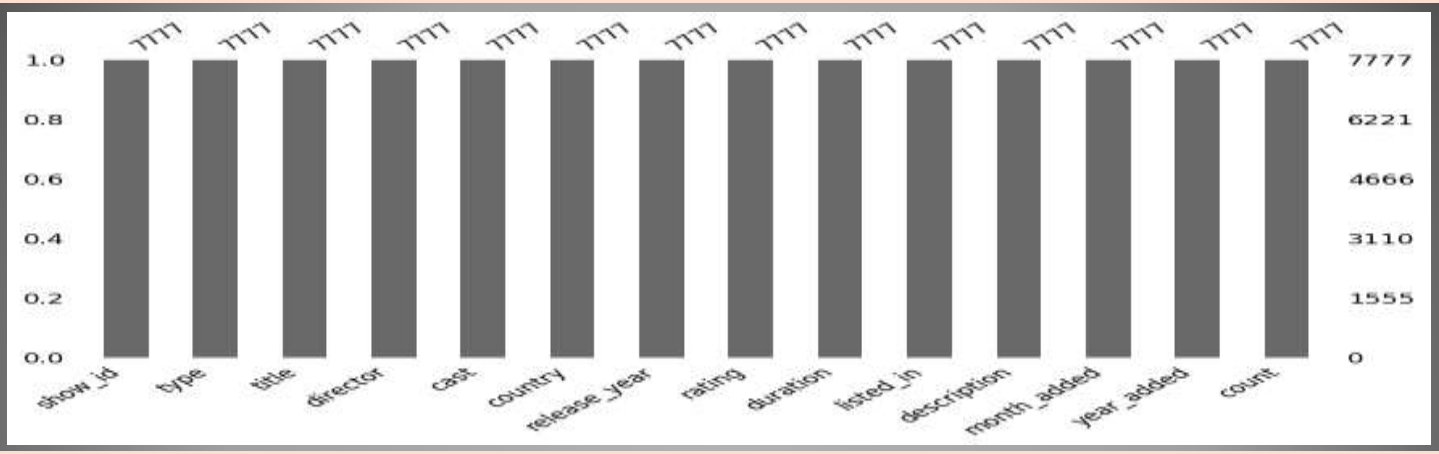
	type	duration
0	Movie	99.307978
1	TV Show	1.760833

Mean for movies rated for Kids duration 99.307978
Mean for movies rated for older kids duration 1.760833
Standard deviation for movies rated for Older Kids duration 1.560603
Standard deviation for movies rated for kids duration 28.530881

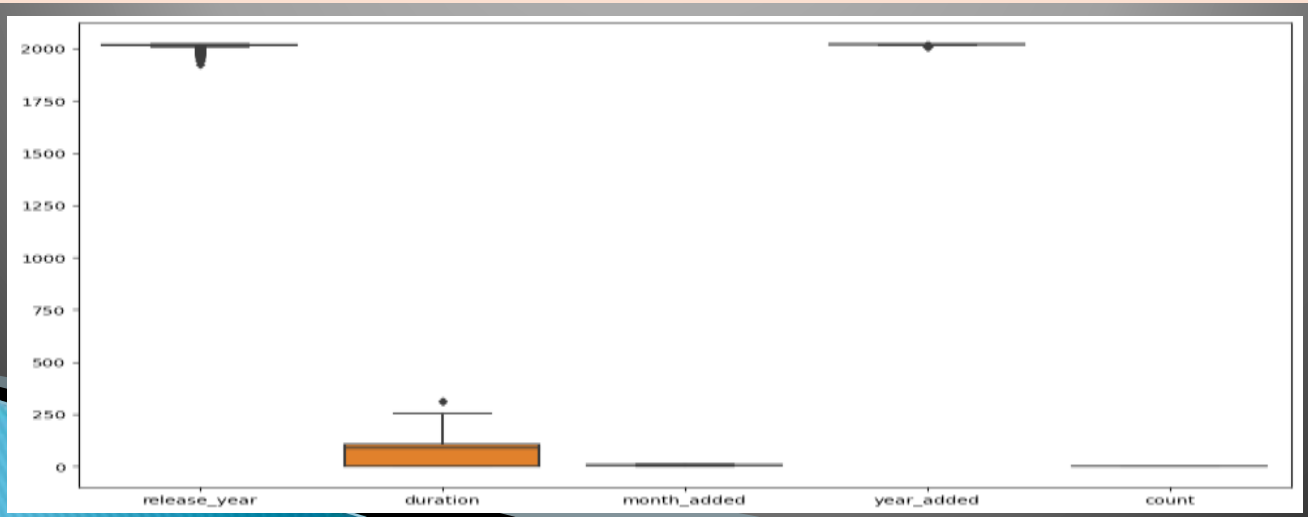
- Because the t-value is not in the range, the null hypothesis is rejected.
- As a result, The duration which is more than 90mins are movies

Feature Engineering:

- Handling Missing Values –
  - we have all ready clear missing values or handling NAN values that why we can see here no missing values , if we drop these nan values it will not affect that much while building the model.



- Handling Outliers–
  - Since, the some of the data present in textual format except release year. and duration is all data clear shows and some unwanted values show. The data that we need to create cluster/building model are present in textual format. So, there is no need to perform handling outlier.



Feature Engineering:

Feature Manipulation & Selection

- There are 4050 directors, 6882 actors / actresses, and 82 countries in the data set which are too many features to include in a K-Means clustering model. Thus, I will reduce the number of features by only taking the primary director, lead actor/actress, and primary country for each movie or TV show. Then, I will count encode each of these features by replacing each categorical value with the number of times it appears in the dataset.

I will use one-hot encoding en coding to encode ratings and genres (listed-in) since there are only 5 ratings and 36 genres. One-hot-encoding creates new columns indicating the presence (1) or absence (0) of each possible value in the data. Since a movie or TV show can belong to more than one genre, I will use a Multi Label Binary for rating.

Categorical Encoding

Unwanted columns drop and 'type', 'country‘ , 'rating', 'listed-in' use label Encoder fit transform

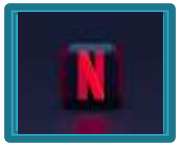
Modeling Approach:

- Select the attributes based on which you want to cluster the shows
- Text preprocessing: Remove all stop words and punctuation marks, convert all textual data to lowercase.
- Stemming to generate a meaningful word out of corpus of words.
- **TFIDF** Word vector
- Lemmatization
- Tokenization
- Dimensionality reduction
- Use different algorithms to cluster the movies, obtain the optimal number of clusters using different techniques
- Build optimal number of clusters and visualize the contents of each cluster using word clouds

Cluster :

We create one cluster column based on the following features:

- Director
- Cast
- Country
- Rating
- Listed in (genres)
- Description



Before clusters implementation we need to pre-process the data. So that we filtered data with following steps :

1.Removing Stop words :

• Stop words are common words like “the”, “and” and “but” do not carry much meaning on their own and are often seen as noise in the data.

2. Lowercasing words :

• Lowercasing the words can also reduce the size of the vocabulary, which can make it easier to work with larger texts or texts in languages with a high number of inflected forms.

3. Removing Punctuation :

• Punctuation marks like periods, commas, and exclamation points can add noise to the data and can sometimes be treated as separate tokens, which can affect the performance of NLP models.

4. Stemming :

• used Snowball Stemmer to generate a meaningful word out of corpus of words.

• For example, the words "run," "runs," "ran," and "running" are all different inflected forms of the same word "run," and a stemmer can reduce them all to the base form "run."

5. Tokenization of corpus and Word vector – TFIDF :

1. This is important in NLP tasks because most machine learning models expect numerical input and cannot work with raw text data directly. Word vector allows you to input the words into a machine learning model in a way that preserves the meaning and context of the words.

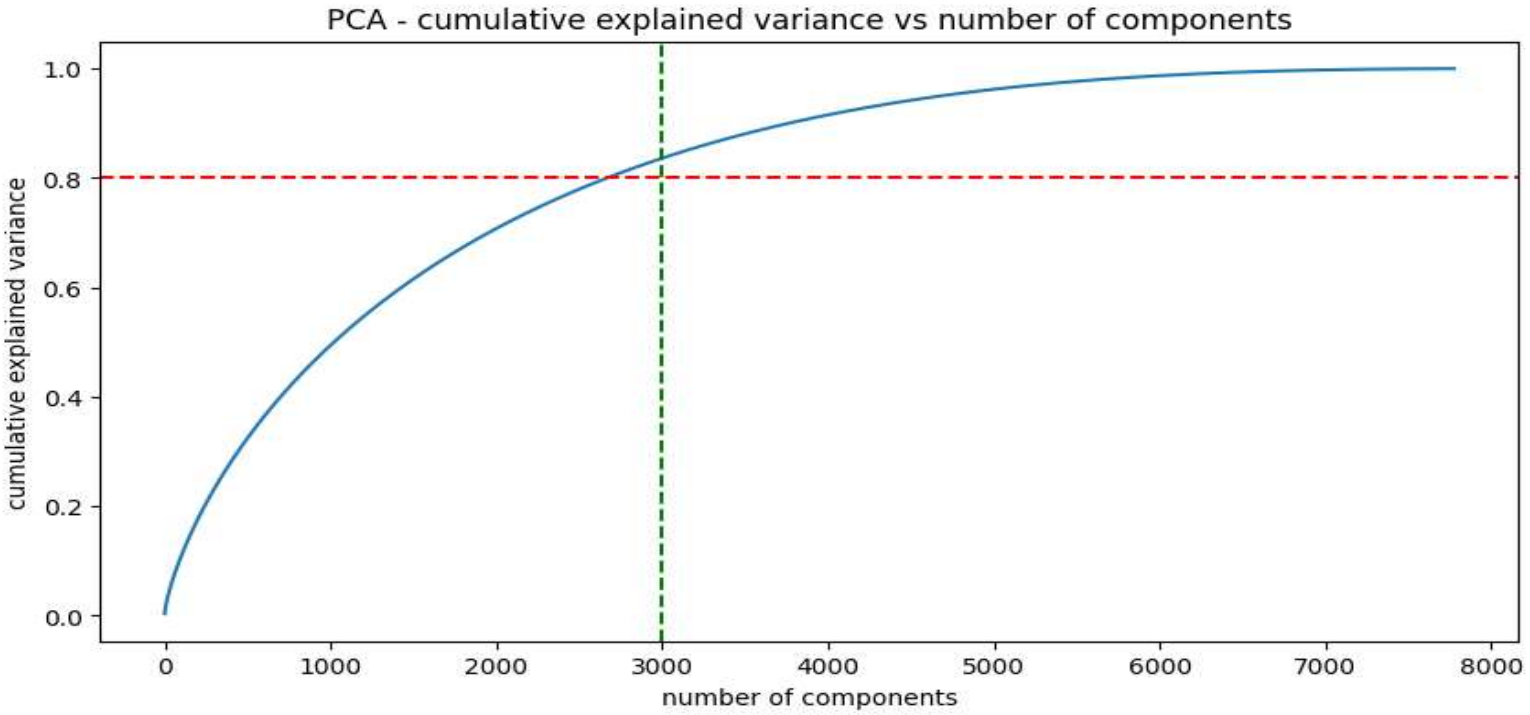
6. Lemmatization :

Lemmatize verbs in list of tokenized words.

7. Dimensionality reduction – PCA :

• Dimensionality reduction is the process of reducing the number of features or dimensions in a dataset while preserving as much information as possible. As high-dimensional datasets can be difficult to work with and can sometimes suffer from the curse of dimensionality

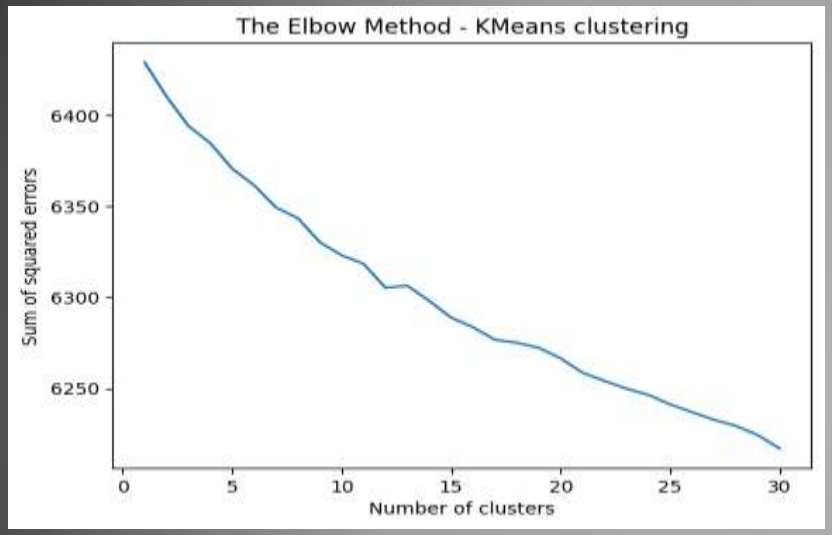
Principle Component Analysis:

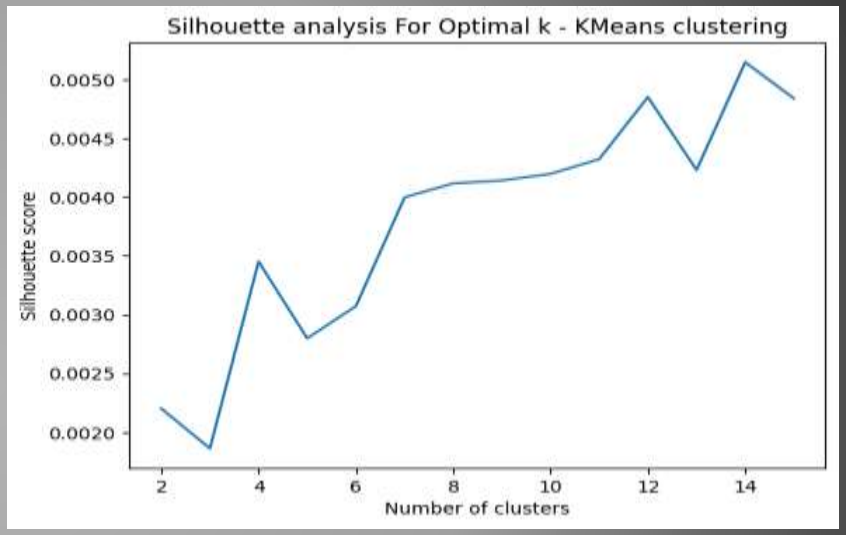


- We find that 100% of the variance is explained by about ~7500 components.
- Also, more than 80% of the variance is explained just by 3000 components.
- Hence to simplify the model, and reduce dimensionality, we can take the top 3000 components, which will still be able to capture more than 80% of variance.

Clusters Model Implementation

- Visualizing the elbow curve and Silhouette score to decide on the optimal number of clusters for K-means clustering algorithm.

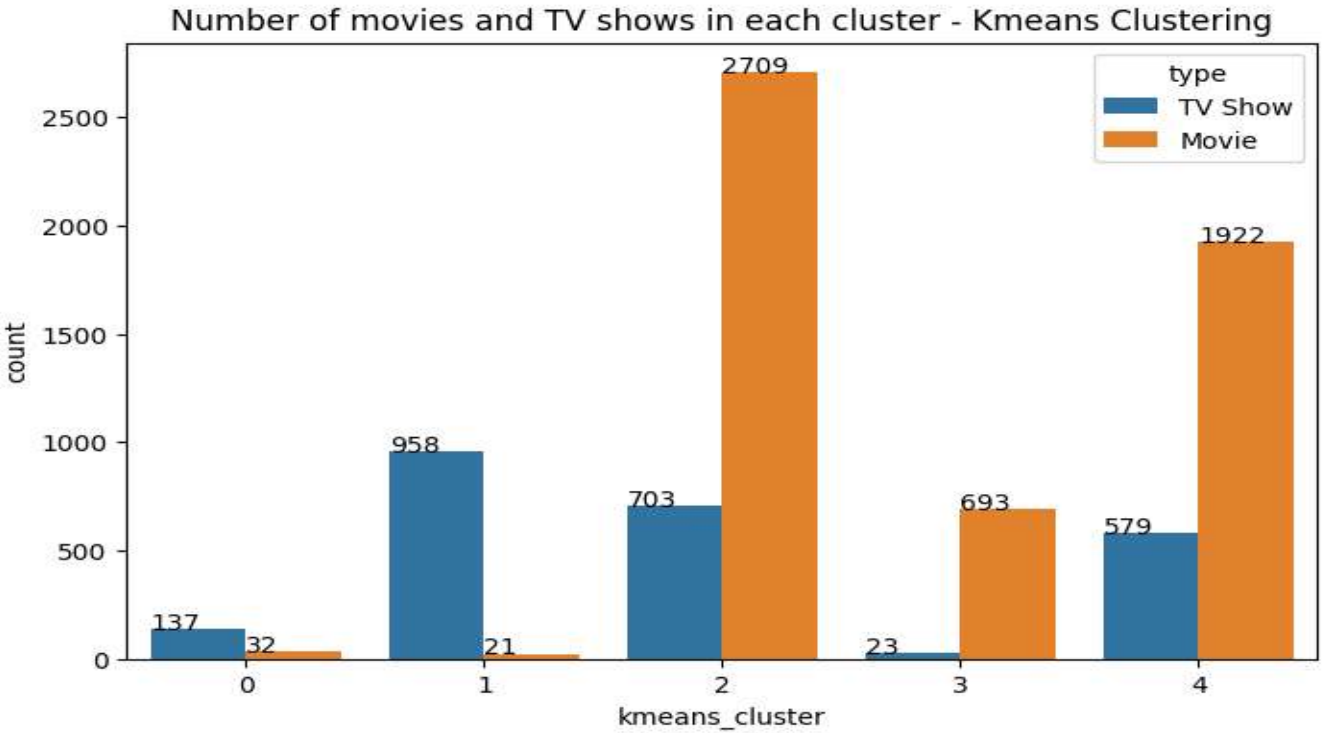




- The sum of squared distance between each point and the centroid in a cluster decreases with the increase in the number of clusters.

- The highest Silhouette score is obtained for 5 clusters.
- Building 5 clusters using the k-means clustering algorithm

K-Means Clusters :



- Successfully built 5 clusters using the k-means clustering algorithm.
- In cluster 0, 1 & 4 highest number of count belong from Movie class.
- Cluster 3 build on TV shows.

Word Cloud :

- In this here Implement and showing only 4 different word cloud cluster

Word Cloud on "description"

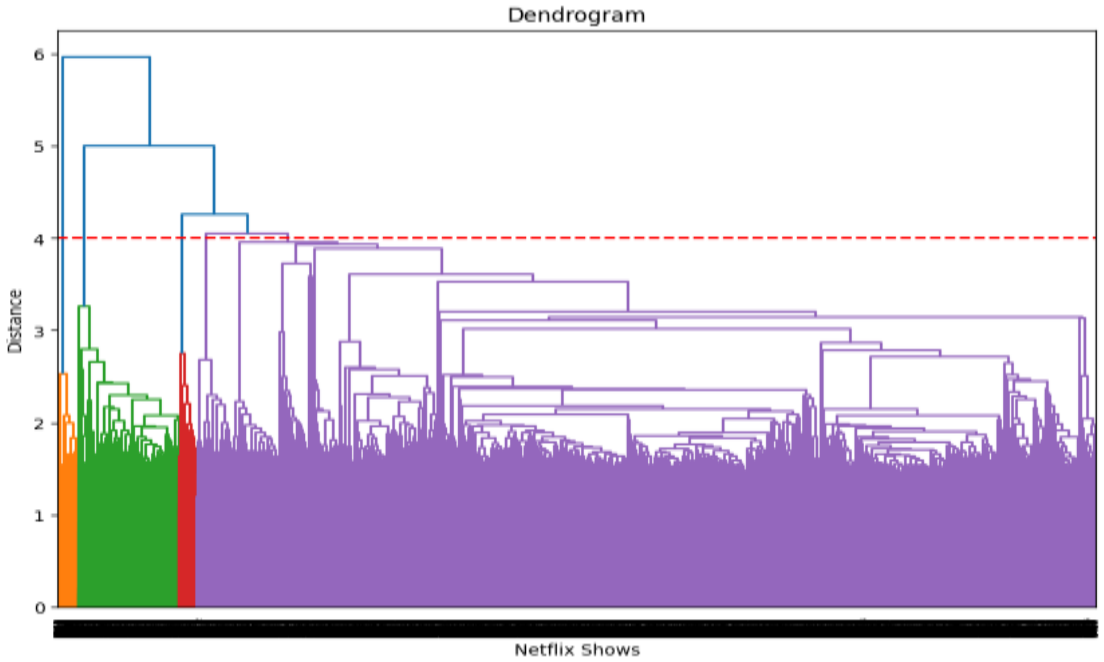
Word Cloud on "cast"

Word Cloud on "country"

Word Cloud on "title"

Hierarchical clustering:

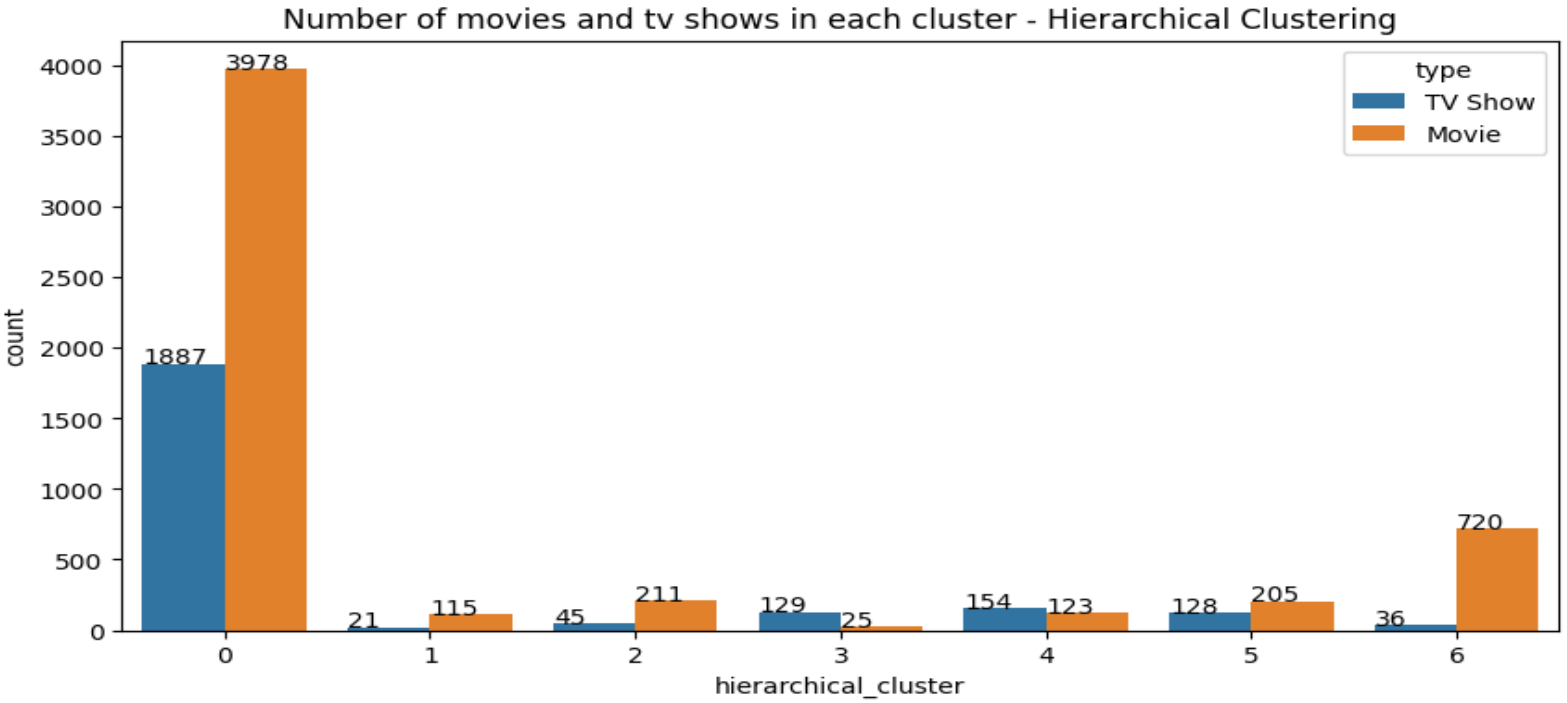
- Building clusters using the **Agglomerative (hierarchical) clustering algorithm.**
  - Agglomerative hierarchical clustering is a method of clustering that is used to build a hierarchy of clusters. It is a bottom-up approach, where each sample is initially treated as a single-sample cluster and clusters are merged together as they are deemed similar.



- Visualizing the Dendrogram to decide on the optimal number of clusters for the agglomerative (hierarchical) clustering algorithm.

- At a distance of 4 units, 7 clusters can be built using the agglomerative clustering algorithm

Hierarchical clusters:



- Agglomerative Hierarchical clusters:

• Successfully built 7 clusters using the Agglomerative (hierarchical) clustering algorithm.

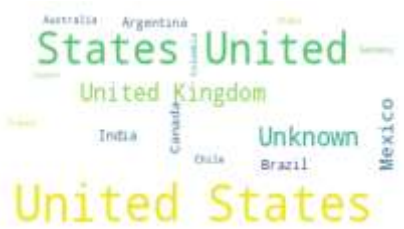
• Highest number of data point build on cluster 0

Word Cloud

- Word Cloud on "listed-in (genre)" & Word Cloud on "country“ This 2 word cloud we can see

Word Cloud on "listed-in (genre)"

Word Cloud on "country“



Content Based Recommendation System :

- We can build a simple content-based recommender system based on the similarity of the movie/shows.
- If a person has watched a show on Netflix, the recommender system must be able to recommend a list of similar shows that s/he likes.
- To get the similarity score of the shows, we can use cosine similarity.
- The similarity between two vectors (A and B) is calculated by taking the dot product of the two vectors and dividing it by the magnitude value. We can simply say that the CS score of two vectors increases as the angle between them decreases.

```
recommend('Naruto')
```

```
If you liked 'Naruto', you may also enjoy:

Naruto ShippÃ»den the Movie: Bonds
Naruto Shippuden: The Movie
Naruto Shippuden : Blood Prison
Naruto the Movie 2: Legend of the Stone of Gelel
Naruto ShippÃ»den the Movie: The Will of Fire
Naruto the Movie 3: Guardians of the Crescent Moon Kingdom
Naruto Shippuden: The Movie: The Lost Tower
DRIFTING DRAGONS
Marvel Anime: Wolverine
Dino Girl Gauko
```

```
recommend('Our Planet')
```

```
If you liked 'Our Planet', you may also enjoy:

Nature's Great Events: Diaries
Nature's Great Events (2009)
Planet Earth: The Complete Collection
Blue Planet II
Africa
Frozen Planet: On Thin Ice
The Making of Frozen Planet
Nature's Weirdest Events
Frozen Planet: The Epic Journey
Moving Art
```

```
recommend('Phir Hera Pheri')
```

```
If you liked 'Phir Hera Pheri', you may also enjoy:

Bhool Bhulaiyaa
Thank You
Ready
Bhagam Bhag
Golmaal: Fun Unlimited
Chup Chup Ke
Khushi
Life in a ... Metro
Hasee Toh Phasee
Humko Deewana Kar Gaye
```

CONCLUSION:

- In this project, we worked on a text clustering problem wherein we had to classify/group the Netflix shows into certain clusters such that the shows within a cluster are similar to each other and the shows in different clusters are dissimilar to each other.
- The dataset contained about 7787 records, and 11 attributes.
- **We began by dealing with the dataset's missing values and doing exploratory data analysis (EDA).**
- Netflix has 5372 movies and 2398 TV shows, there are more number movies on Netflix than TV shows. 69% of data belong from Movie class and 31% of data belong from TV shows.
- Word like Christmas, Love, World, Man, Story are very common word which are appear most of the time in movie title column.
- Raul Campos and Jan Suter together have directed 18 movies / TV shows, higher than anyone in the dataset
- Anupam Kher, Shahrukh Khan, play highest number of role in the movies. Takahiro Sakurai, Yuki Kaji, play highest role in the TV shows. Unknown cast actor David Attenboroug,Samuel West play highest role shows or movies
- The highest number of movies / TV shows were based out of the US, followed by India and UK.
- Highest number of movie/shows are relesed in netflix in between 2015-2020 and highest number of count belong from 2018 year.
- most of the movie and TV shows have rating of Adults (Mature Audiance) then followed by Teens (younger audiance).
- Highest number of genre belong from Dramas, Comedies respectively. Least number of genre belong from TV Sci-Fi & Fantasy or spanish language TV Shows.
- **Hypothesis**
- (i) Because the t-value is not in the range, the null hypothesis is rejected. As a result, movies rated for kids and older kids are not at least two hours long. (ii) Because the t-value is not in the range, the null hypothesis is rejected. As a result, The duration which is more than 90mins are movies
- **Modeling Approach**
- It was decided to cluster the data based on the attributes: director, cast, country, genre, rating and description. The values in these attributes were tokenized, Stemming, Lemmatization, preprocessed, and then vectorized using TFIDF vectorizer.
- Through TFIDF Vectorization, we created a total of 10000 attributes.
- We used Principal Component Analysis (PCA) to handle the curse of dimensionality. 3000 components were able to capture more than 80% of variance, and hence, the number of components were restricted to 3000.
- **Clusters Model**
- We first built clusters using the K-Means Clustering algorithm, and the optimal number of clusters came out to be 5. This was obtained through the elbow method and Silhouette score analysis.
- Then clusters were built using the Agglomerative clustering algorithm, and the optimal number of clusters came out to be 7. This was obtained after visualizing the dendrogram.
- A content based recommender system was built using the similarity matrix obtained after using cosine similarity. This recommender system will make 10 recommendations to the user based on the type of show they watched.


Thank you