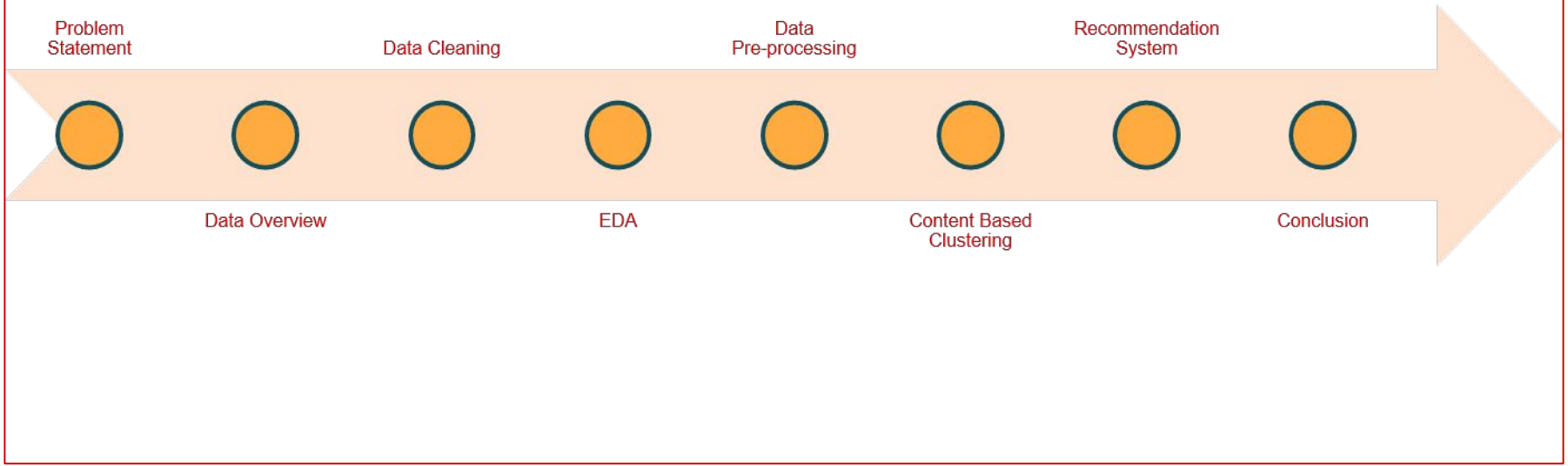
Capstone Project-4

Netflix Movies And TV Shows Clustering

Prepared By
Akshay Nikam

Points for Discussion



Problem Statement

- This dataset consists of tv shows and movies available on Netflix as of 2019. The dataset is collected from Flixable 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 services 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.
- In this project, you are required to do
    - 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 Overview

In this dataset We have 12 columns and 7787 rows.

Columns

1. **show_id :** Unique ID for every Movie / Tv Show
2. **type :** Identifier - A Movie or TV Show
3. **title :** Title of the Movie / Tv Show
4. **director :** Director of the Movie
5. **cast :** Actors involved in the movie / show
6. **country :** Country where the movie / show was produced
7. **date_added :** Date it was added on Netflix
8. **release_year :** Actual Releaseyear of the movie / show
9. **rating :** TV Rating of the movie / show
10. **duration :** Total Duration - in minutes or number of seasons
11. **listed_in :** Genres of content
12. **description:** The Summary description of content

Data Overview

Data Summary:

Data Summary

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7787 entries, 0 to 7786
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   show_id       7787 non-null   object
 1   type          7787 non-null   object
 2   title         7787 non-null   object
 3   director      5398 non-null   object
 4   cast          7069 non-null   object
 5   country       7280 non-null   object
 6   date_added    7777 non-null   object
 7   release_year  7787 non-null   int64 
 8   rating        7780 non-null   object
 9   duration      7787 non-null   object
 10  genres        7787 non-null   object
 11  description   7787 non-null   object
dtypes: int64(1), object(11)
memory usage: 730.2+ KB
```

Data Cleaning

Null values Treatment:

show_id	0
type	0
title	0
director	2389
cast	718
country	507
date_added	10
release_year	0
rating	7
duration	0
genres	0
description	0

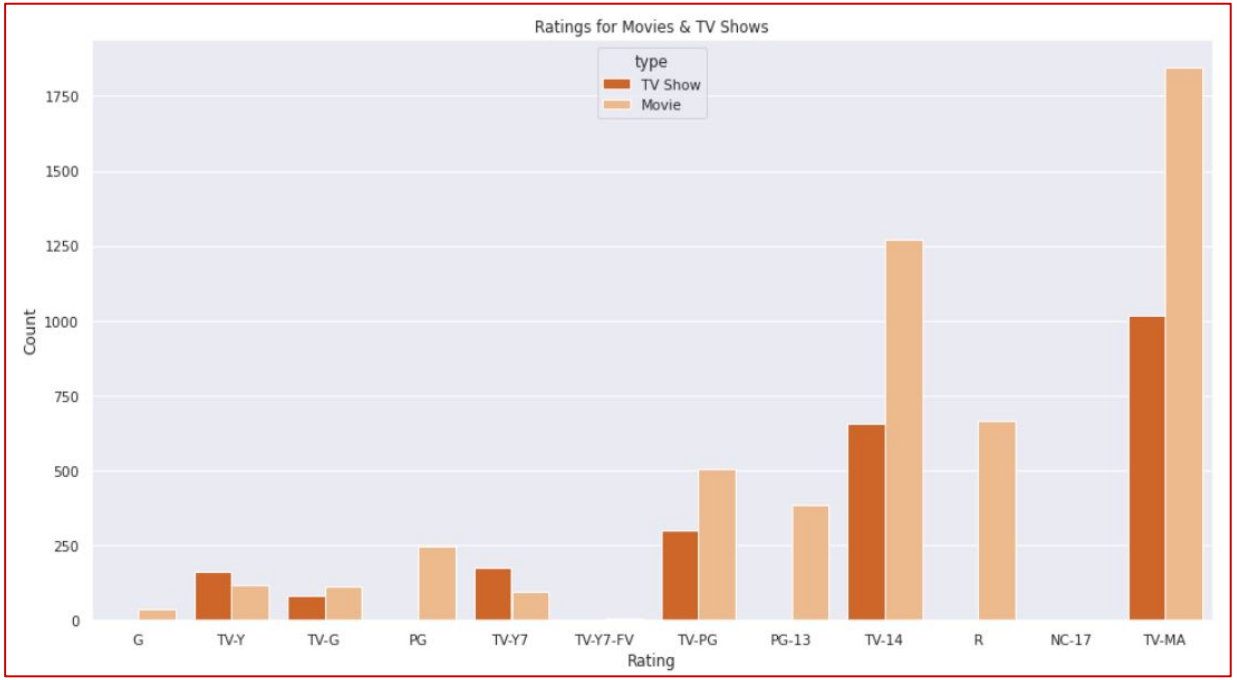
dtype: int64

type	0
title	0
director	0
cast	0
country	0
date_added	0
release_year	0
rating	0
duration	0
genres	0
description	0
added_year	0
added_month	0

dtype: int64

Data Cleaning

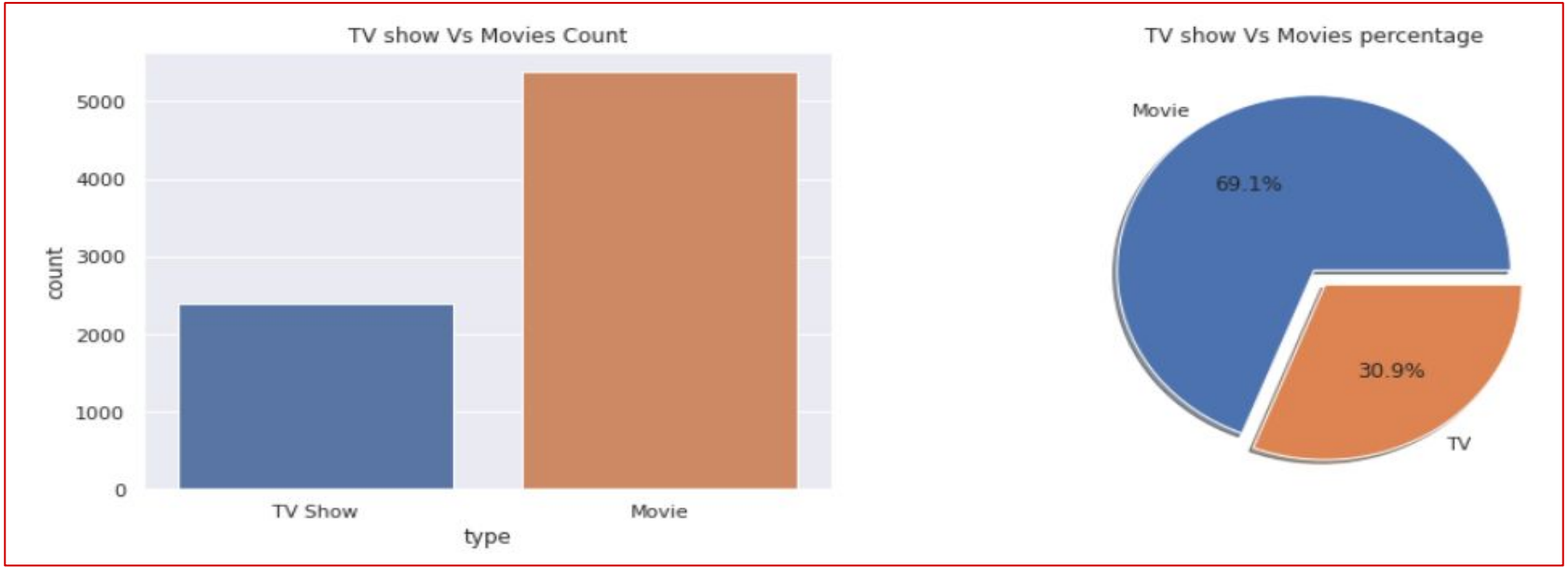
Rating:




```
TV-MA : Adults
    R : Adults
PG-13 : Teens
TV-14 : Young Adults
TV-PG : Older Kids
   NR : Adults
 TV-G : Kids
 TV-Y : Kids
TV-Y7 : Older Kids
   PG : Older Kids
    G : Kids
NC-17 : Adults
TV-Y7-FV : Older Kids
   UR : Adults
```

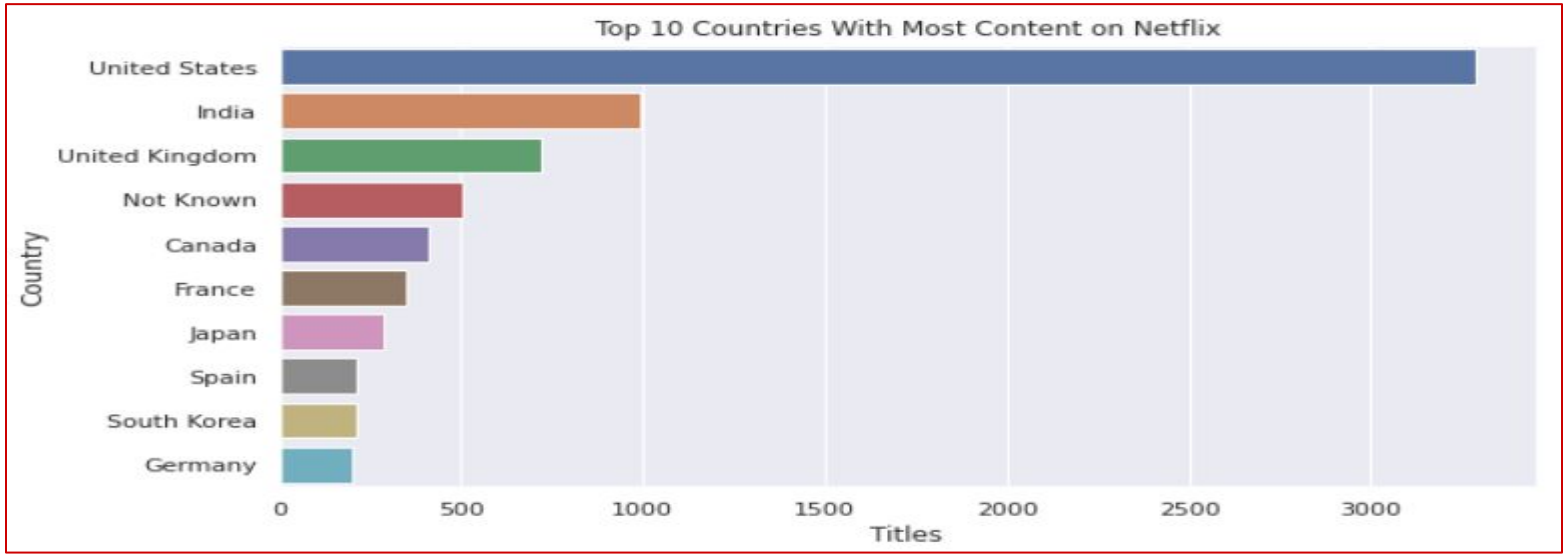
EDA

Univariate Analysis : Type



EDA

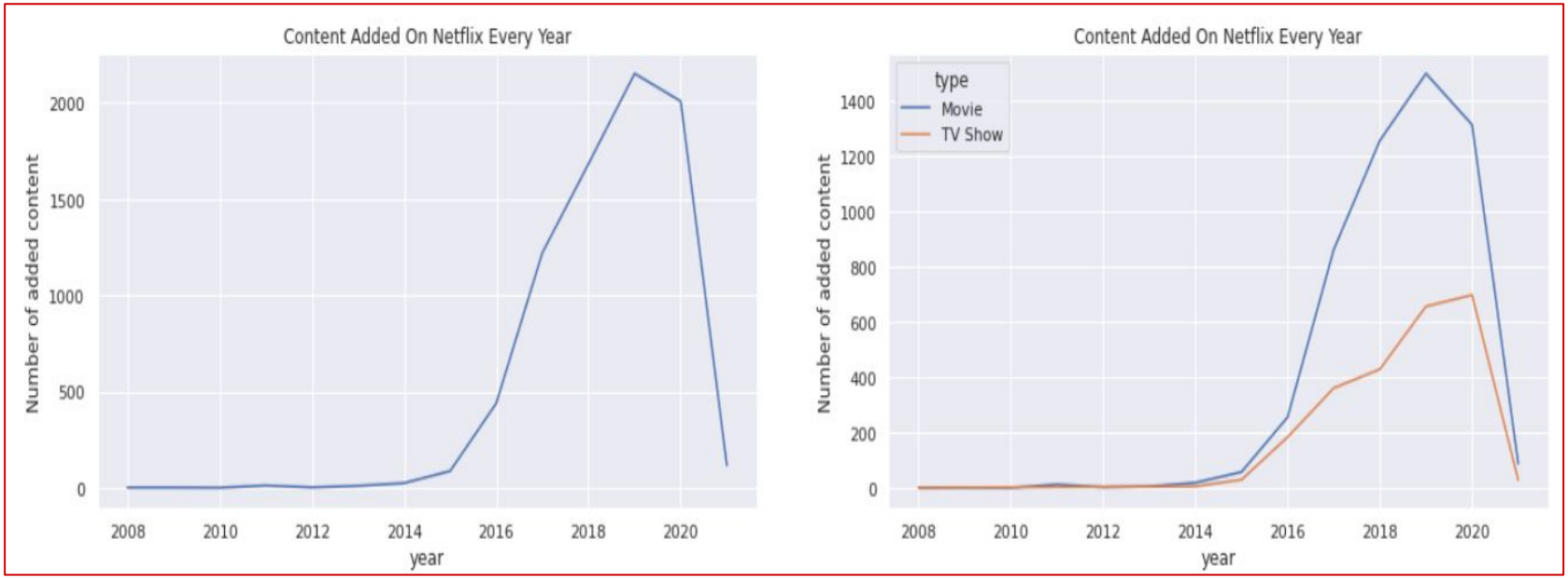
Univariate Analysis : Country



- The top 3 countries together account for about 51% of all movies and TV shows in the dataset.
- This value increases to about 68% for top ten countries.

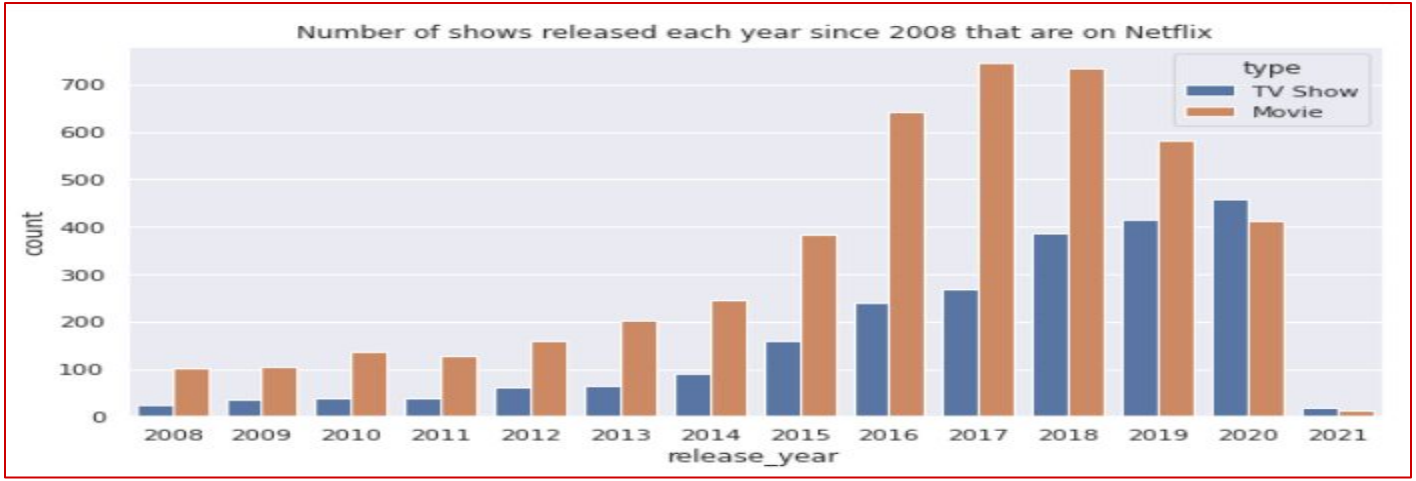
EDA

Univariate Analysis : Year



EDA

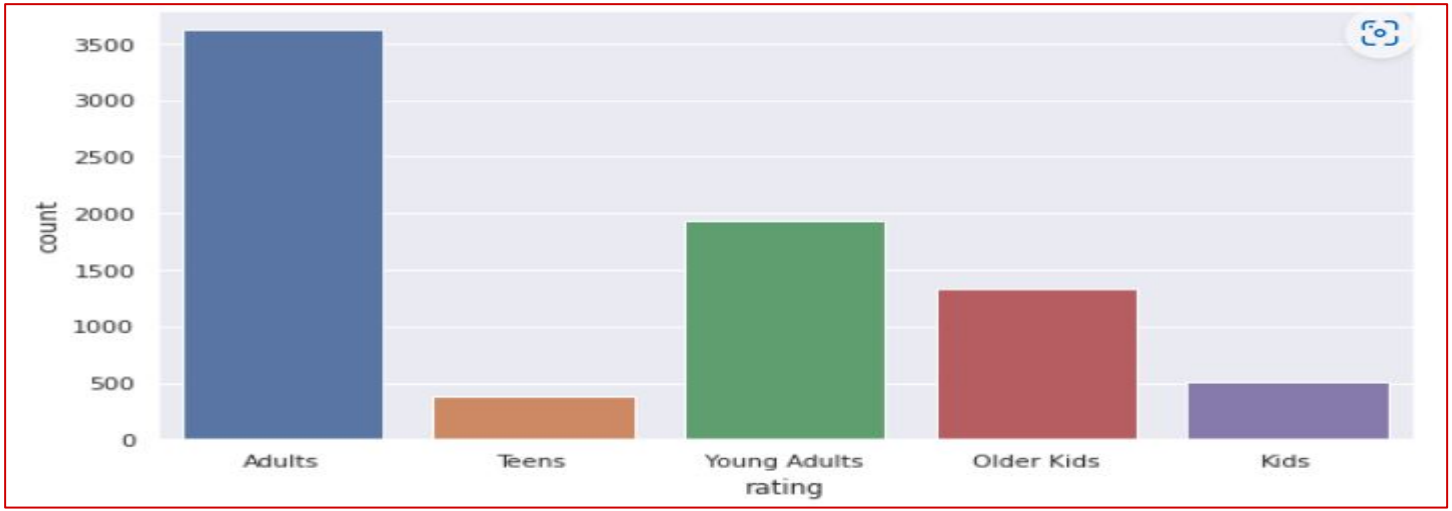
Univariate Analysis : Release Year



- Over the years, Netflix has consistently focused on adding more shows in its platform.
- Though there was a decrease in the number of movies added in 2020, this pattern did not exist in the number of TV shows added in the same year.
- This might signal that Netflix is increasingly concentrating on introducing more TV series to its platform rather than movies.

EDA

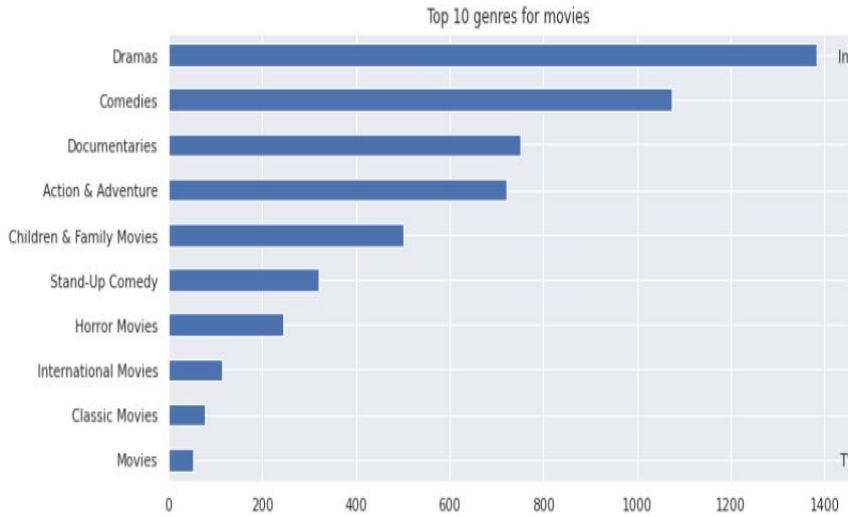
Univariate Analysis : Rating

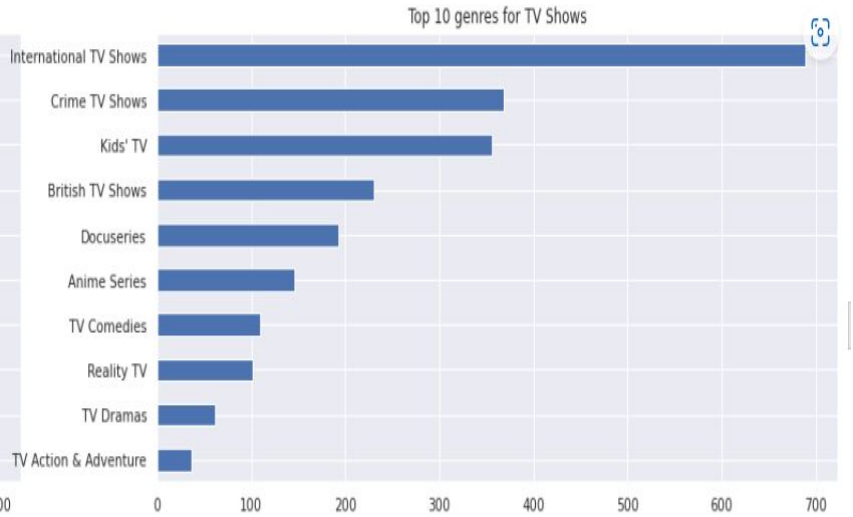


- The majority of the shows on Netflix are catered to the needs of adult and young adult population.
- And this is considerable also beacause now on adults are more in nummber and consuming lots of content on ott than other age groups.

EDA

Bivariate Analysis : Genres





Data Preprocessing

Selection of Attributes for Clustering:

show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
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 ...
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...
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...
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...
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...

- Clustering colunms attributes are choosed on the basis of the texual data variable present in the dataset.
- Problem statement is to do content based clustering so for which Combining all the texual data present in the dataset into a column and the do the clustering

Data Preprocessing

Text Preprocessing: Removing All Non ASCII values

- ASCII stands for the "American Standard Code for Information Interchange". It was designed in the early 60's, as a standard character set for computers and electronic devices. ASCII is a 8-bit or 1 bytes character set containing 127 characters
- Non ASCII are those spacial charactets from 128 to 255.
- Words like Gülmez, Taş, Papuççuoğlu contails these charactors which will need to eliminate.

'Muharrem Gulmez Erdem Yener, Ayhan Tas, Emin Olcay, Muharrem Gulmez, Elif Nur Kerkuk, Tark Papuccuoglu, Suzan Aksoy, Doga Konakoglu, Esin Eden, Deniz Ozerman vies The slacker owner of a public bath house rallies his community to save it when a big developer comes to town to close it down and open a new mall.'

Data Preprocessing

Text Preprocessing: Removing Stop Words and Punctuations

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you',
 "you're", "you've", "you'll", "you'd", 'your', 'yours', 'yourself',
 'yourselves', 'he', 'him', 'his', 'himself', 'she', "she's", 'her',
 'hers', 'herself', 'it', "it's", 'its', 'itself', 'they', 'them',
 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom',
 'this', 'that', "that'll", 'these', 'those', 'am', 'is', 'are',
 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had',
 'having', 'do', 'does', 'did', 'doing', 'a', 'an', 'the', 'and',
 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of', 'at',
 'by', 'for', 'with', 'about', 'against', 'between', 'into',
 'through', 'during', 'before', 'after', 'above', 'below', 'to',
 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under',
 'again', 'further', 'then', 'once', 'here', 'there', 'when',
 'where', 'why', 'how', 'all', 'any', 'both', 'each', 'few', 'more',
 'most', 'other', 'some', 'such', 'no', 'nor', 'not', 'only', 'own',
 'same', 'so', 'than', 'too', 'very', 's', 't', 'can', 'will',
 'just', 'don', "don't", 'should', "should've", 'now', 'd', 'll',
 'm', 'o', 're', 've', 'y', 'ain', 'aren', "aren't", 'couldn',
 "couldn't", 'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't",
 'hasn', "hasn't", 'haven', "haven't", 'isn', "isn't", 'ma',
 'mightn', "mightn't", 'mustn', "mustn't", 'needn', "needn't",
 'shan', "shan't", 'shouldn', "shouldn't", 'wasn', "wasn't",
 'weren', "weren't", 'won', "won't", 'wouldn', "wouldn't"],
```

'muharrem gulmez erdem yener ayhan tas emin olcay muharrem gulmez elif nur kerkuk tark papuccuoglu suzan aksoy doga konakoglu esin eden deniz ozerman turkey comedies international movies slacker owner public bath house rallies community save big developer comes town close open new mall'

Data Preprocessing

Text Preprocessing: Word Lemmatization

The specific discipline of lemmatization is a subcategory of a process called stemming. In natural language processing, stemming allows the computer to group together words according to their various inflections that are tagged with a particular stem. For instance: "walk," "walked" and "walking."

Lemmatization is a bit more complex in that the computer can group together words that do not have the same stem, but still have the same inflected meaning. Grouping the word "good" with words like "better" and "best" is an example of lemmatization. Lemmatization

```
'muharrem gulmez erdem yener ayhan tas emin olcay muharrem gulmez elif nur kerkuk tark papuccuoglu suzan aksoy doga konakoglu esin eden deniz ozerman turkey comedies international
er owner public bath house rallies community save big developer comes town close open new mall'
```

Data Preprocessing

Text Preprocessing: Word Vectorizer

Term Frequency: TF of a term or word is the number of times the term appears in a document compared to the total number of words in the document.

$$TF = \frac{\text{number of times the term appears in the document}}{\text{total number of terms in the document}}$$

Inverse Document Frequency: IDF of a term reflects the proportion of documents in the corpus that contain the term. Words unique to a small percentage of documents (e.g., technical jargon terms) receive higher importance values than words common across all documents (e.g., a, the, and).

$$IDF = log(\frac{\text{number of the documents in the corpus}}{\text{number of documents in the corpus contain the term}})$$

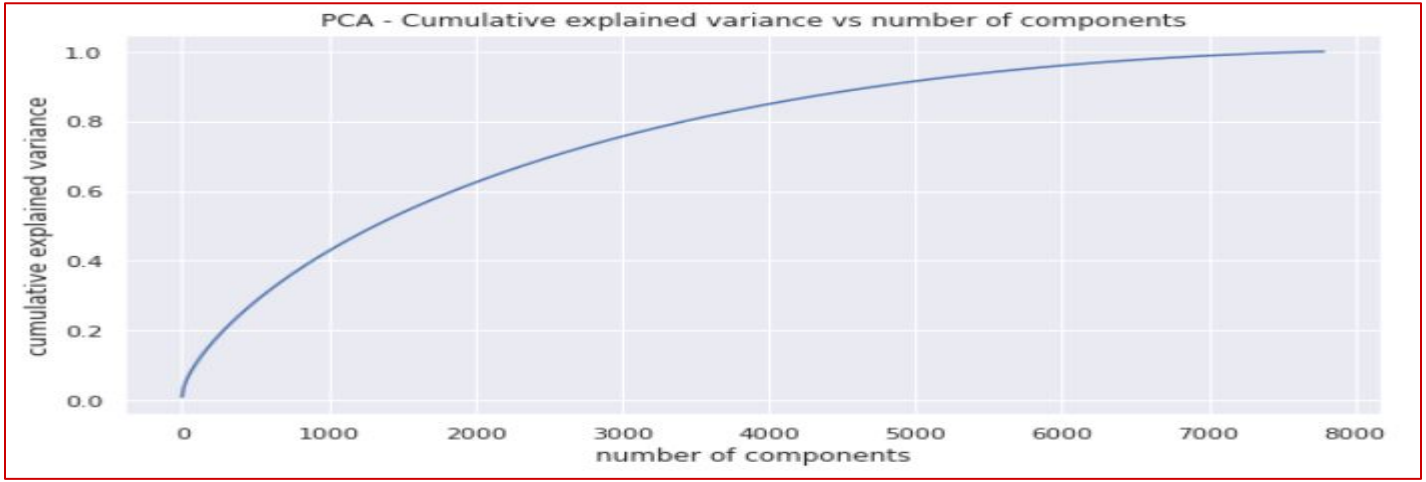
The TF-IDF of a term is calculated by multiplying TF and IDF scores.

$$TF\text{-}IDF = TF * IDF$$

```
(7787, 20000)
```

Data Preprocessing

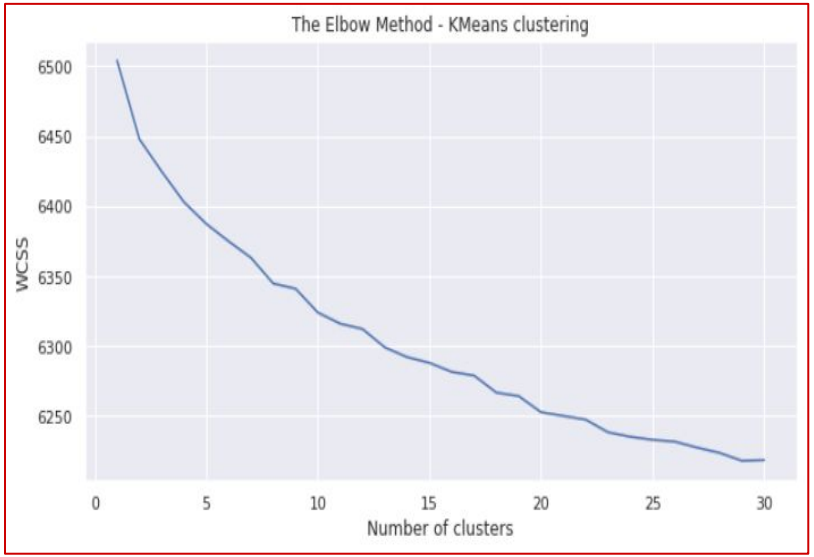
Data Preprocessing: Dimensionality Reduction

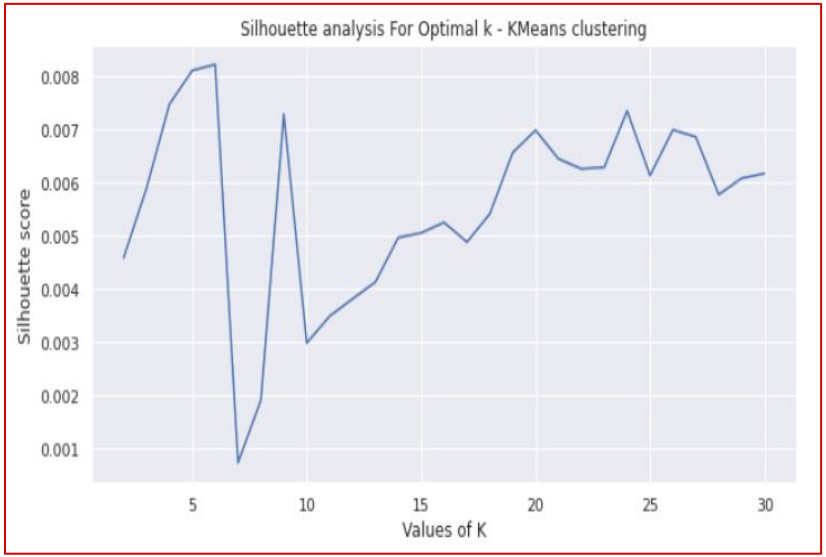


- We find that 100% of the variance is explained by about ~7500 components.
- Also, more than 80% of the variance is explained just by 4000 components.
- Hence to simplify the model, and reduce dimensionality, we can take the top 4000 components, which will still be able to capture more than 80% of variance

Clustering

Selection of K values (Number of Clusters): Elbow Plot and Silhouette Score

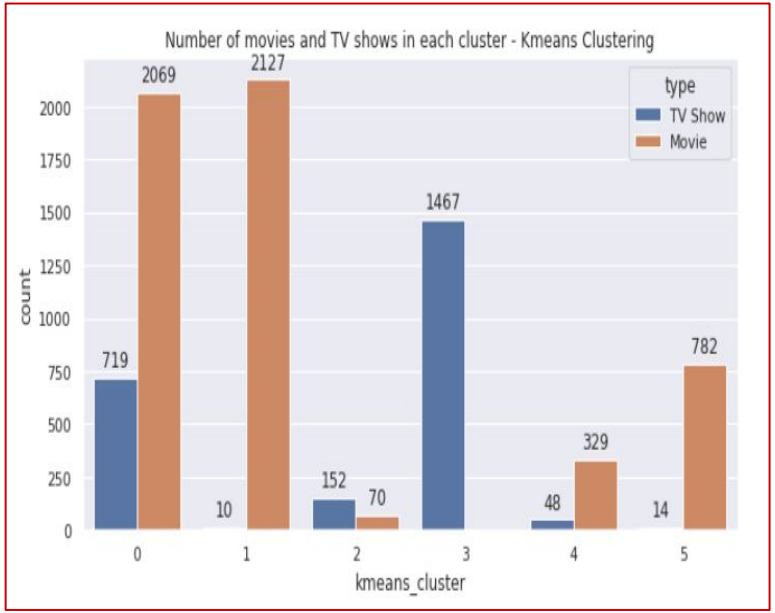




- From above plots, will form 6 clusters using Hierarchical Clustering.

Clustering

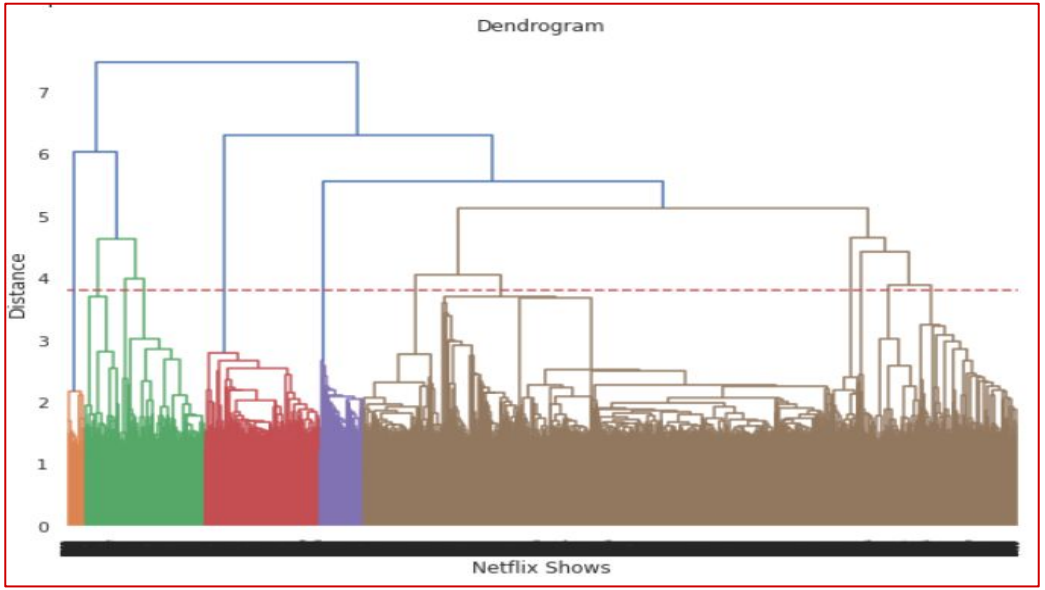
Clustering: K - Means Clustering



- Keywords observed in cluster 0 : life, new, family, friend, save, help, discover, home, teen
- Keywords observed in cluster 1: life, love, family, father, young, girl, man, woman, friend, daughter.
- Keywords observed in cluster 2: young, world, girl, mysterious, humanity, life, student, school, battle, demon, force.
- Keywords observed in cluster 3: love, life, family, romance, crime, murder, world, adventure.
- Keywords observed in cluster 4: comedian, special, stand, comic, stage, sex, joke.
- Keywords observed in cluster 5: documentary, world, life, filmmaker, american, life.

Clustering

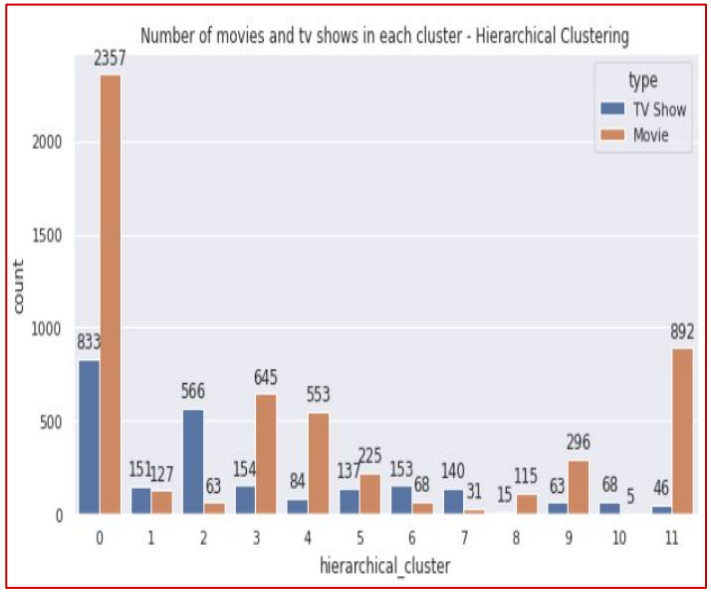
Selection of Number of Clusters: Dendrogram



- From above Dendrogram, will form 12 clusters using K - Means Clustering.

Clustering

Clustering: Hierarchical Clustering

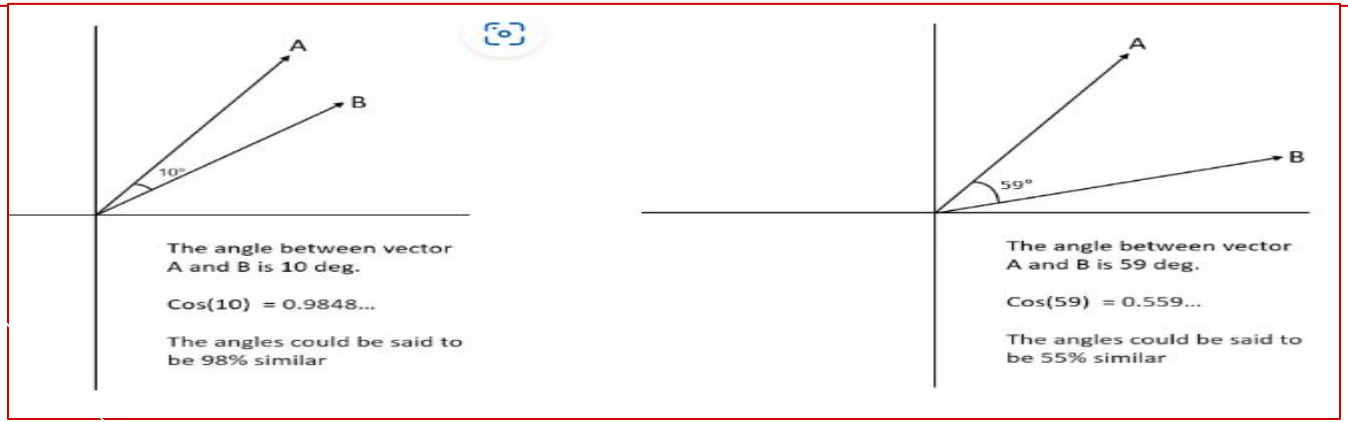


- Keywords observed in cluster 0: life, new, find, family, save, friend, young, teen, adventure.
- Keywords observed in cluster 1: love, family, life, student, romance, school, woman, master, father
- Keywords observed in cluster 2: life, new, series, crime, world, murder, history, detective
- Keywords observed in cluster 3: family, life, love, friend, teen, woman, man, young, world, wedding, secret
- Keywords observed in cluster 4: documentary, music, world, team, interview,history, family, career, battle, death
- Keywords observed in cluster 5: family, life, mexico, young, new, woman, man, secret, spain, death, singer
- Keywords observed in cluster 6: young, life, girl, world, friend, mysterious, demon, student, school, father
- Keywords observed in cluster 7: love, life, woman, new, student, family, korea, secret, detective, young
- Keywords observed in cluster 8: woman, man life, egypt, wealthy, money, young, love, revolution, struggling
- Keywords observed in cluster 9: comedian, stand, life, comic, special, show, live, star, stage, hilarious, stories
- Keywords observed in cluster 10: animal, nature, explore, planet, species, survive, natural, life, examine, earth
- Keywords observed in cluster 11: love, man, woman, india, father, friend, girl, mumbai, city, learn, young

Recommendation System

Recommendation System: Cosine Similarity

- We will build a simple content based recommender system based on the similarity score between shows.
- If a person has watched a show on Netflix, the recommender system must be able to recommend a list of similar shows that he/she 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 Cosine Similarity score of two vectors increases as the angle between them decreases.



Recommendation System

Recommendation System:

```
# Recommendations for 'A Man Called God'
recommend_10('A Man Called God')
```

```
If you liked 'A Man Called God', you may also enjoy:

['Mr. Sunshine',
 'One Spring Night',
 'Rugal',
 'The King: Eternal Monarch',
 'My Mister',
 'My Little Baby',
 'Reply 1994',
 'Extracurricular',
 'My Secret Romance',
 'Chef & My Fridge']
```

```
# Recommendations for 'Stranger Things'
recommend_10('Stranger Things')
```

```
If you liked 'Stranger Things', you may also enjoy:

['Beyond Stranger Things',
 'Prank Encounters',
 'The Umbrella Academy',
 'Haunted',
 'Scream',
 'Warrior Nun',
 'Nightflyers',
 'Zombie Dumb',
 'Kiss Me First',
 'The Vampire Diaries']
```

Conclusion

Conclusion :

- It was found that Netflix hosts overall more movies than TV shows on its platform. Also, majority of the shows were produced in the United States, and the majority of the shows on Netflix were created for adults and young adults age group.
- Over the years, Netflix has consistently focused on adding more shows in its platform. Though there was a decrease in the number of movies added in 2020, this pattern did not exist in the number of TV shows added in the same year. This might signal that Netflix is increasingly concentrating on introducing more TV series to its platform rather than movies
- It was decided to cluster the data based on the attributes: director, cast, country, genre, and description. The values in these attributes were tokenized, preprocessed, and then vectorized using TFIDF vectorizer. Through TFIDF Vectorization, we created a total of 20000 attributes.
- We used Principal Component Analysis (PCA) to handle the curse of dimensionality. 4000 components were able to capture more than 80% of variance, and hence, the number of components were restricted to 4000.
- We first built clusters using the k-means clustering algorithm, and the optimal number of clusters came out to be 6. 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 12. 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 and it is working accuratetly.

Thank You