CAPSTONE PROJECT

NETFLIX MOVIES AND TV SHOWS CLUSTERING

By – Anita Kumari
(anitaverma4111@gmail.com)

TABLE OF CONTENTS

01 PROBLEM STATEMENT

02 DATA DESCRIPTION

03 DATA PREPARATION AND CLEANING

04 EDA (EXPLORATORY DATA ANALYSIS)

TABLE OF CONTENTS

05 HYPOTHESIS TESTING

06 TEXTUAL DATA PREPROCESSING

07 DIMENSIONALITY REDUCTION

08 MODEL IMPLEMENTATION

09 RECOMMENDER SYSTEM

10 CONCLUSION

01 PROBLEM STATEMENT

Main objective behind the project

PROBLEM STATEMENT

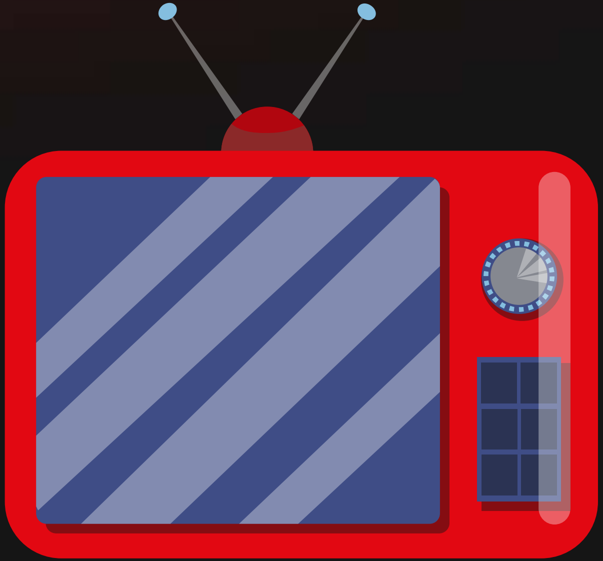
This dataset consists of **tv shows** and **movies** available on **Netflix** as of **2019**. The dataset is collected from Flixable which is a third-party Netflix search engine.

In 2018, they released an interesting **report** which shows that the **number of TV shows** on Netflix has nearly **tripled since 2010**. The streaming 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.

PROBLEM STATEMENT

In this project, 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**.



02

DATA DESCRIPTION

Detailed description of the variables in the dataset

DATA DESCRIPTION

The dataset contains movies and tv shows information like title, cast, director, release year, rating, duration etc.

The features of the dataset are :

- **show_id**: Unique Id number for all the listed rows
- **type**: denotes type of show namely TV Show or Movie
- **title**: title of the movie
- **director**: Name of director/directors
- **cast**: lists the cast of the movie
- **country**: country of the production house
- **date_added**: the date the show was added
- **release_year**: year of the release of the show
- **rating**: show ratings
- **duration**: duration of the show
- **listed_in**: the genre of the show
- **description**: summary/ description of the movie

DATA PREPARATION & CLEANING

To make the data analysis ready i have done the following:

- Filled missing values of **cast** with **Not available**.
- Filled missing values of **country** with **Not Known**.
- Dropped rows of **date_added** missing values.
- Dropped rows of **ratings** missing values.
- Dropped the entire column of **director** as it had much number of missing values.

04

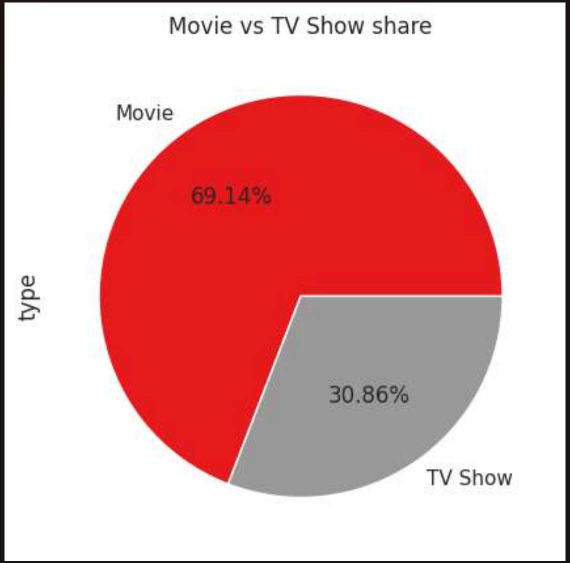
EXPLORATORY DATA ANALYSIS

Analyzing data sets with statistical graphs

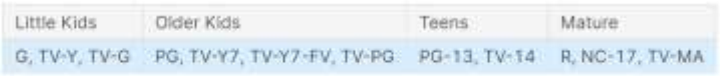
EDA

MOVIE VS TV SHOW SHARE

- Types of shows available in **netflix** is not even with high count for **TV shows**.
- **69.14%** of the data belongs to **movies** and **30.86%** of the data for **TV shows**.



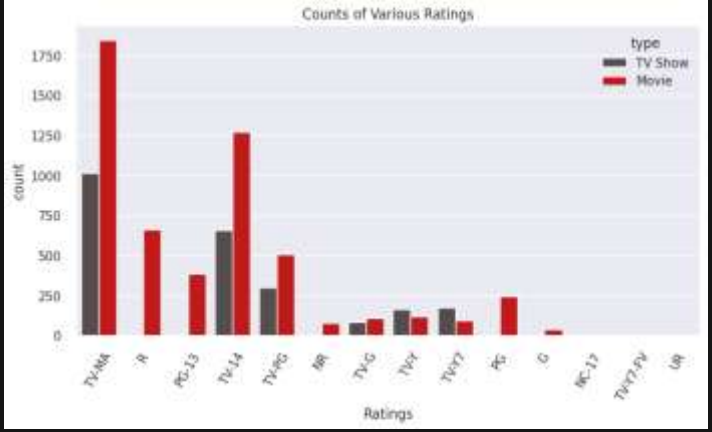
EDA

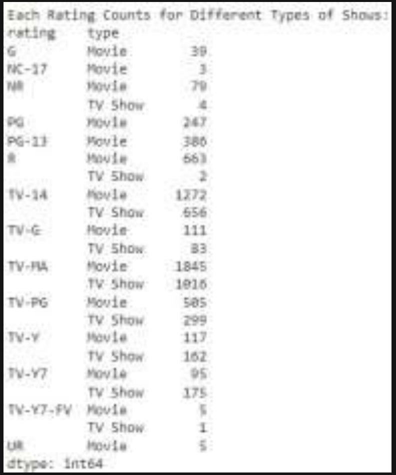


Little Kids	Older Kids	Teens	Mature
G, TV-Y, TV-G	PG, TV-Y7, TV-Y7-FV, TV-PG	PG-13, TV-14	R, NC-17, TV-MA

VARIOUS RATINGS COUNT

- **TV-MA** tops the charts, indicating that **mature content** is **more popular** on **Netflix**.
- This popularity is followed by **TV-14** and **TV-PG**, which are Shows focused on **Teens** and **Older kids**.
- Very few titles with a rating **NC-17** exist. It can be understood since this type of content is purely for the audience **above 17**.



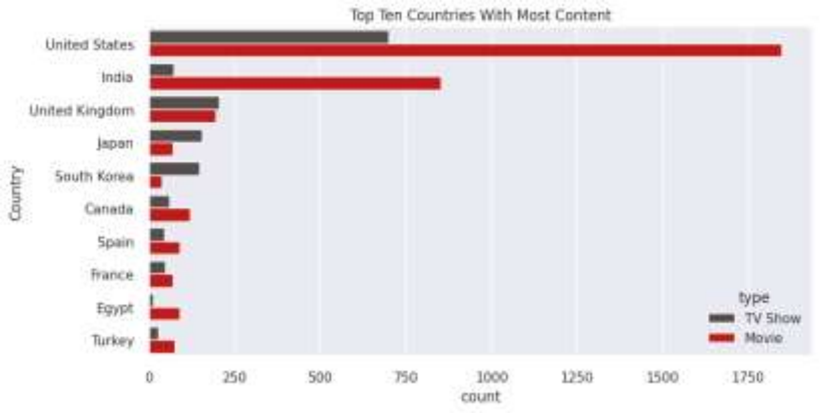


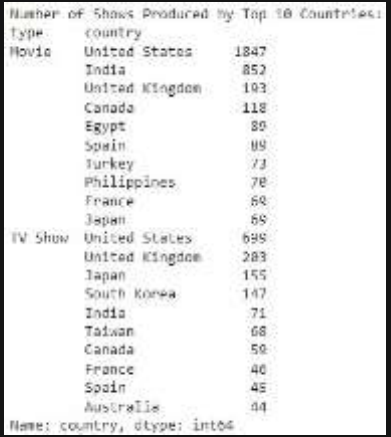
```
Each Rating Counts for Different Types of Shows:
rating     type
G          Movie         39
NC-17      Movie          3
NR         Movie         79
           TV Show        4
PG         Movie        247
PG-13      Movie        386
R          Movie        663
           TV Show        2
TV-14      Movie       1272
           TV Show      656
TV-G       Movie        111
           TV Show       83
TV-MA      Movie       1845
           TV Show     1016
TV-PG      Movie        505
           TV Show      299
TV-Y       Movie        117
           TV Show      162
TV-Y7      Movie         95
           TV Show      175
TV-Y7-FV   Movie          5
           TV Show        1
UR         Movie          5
dtype: int64
```

EDA

TOP 10 COUNTRIES WITH MOST CONTENT

- The **United States** is a **leading producer** of both types of shows (**Movies** and **TV Shows**), this makes sense since **Netflix is a US company**.
- The **influence** of **Bollywood** in **India** explains the type of content available, and perhaps the **main focus** of this industry is **Movies** and **not TV Shows**.
- **TV Shows** are **more frequent** in **South Korea**, which explains the **KDrama** culture nowadays.



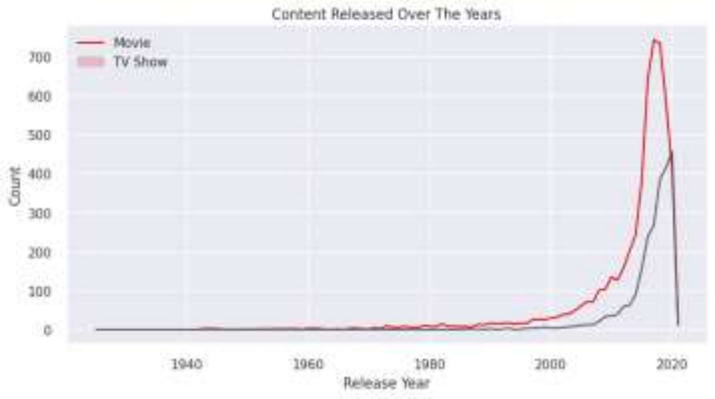


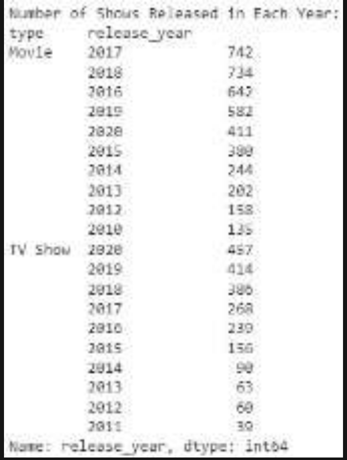
```
Number of Shows Produced by Top 10 Countries:
type     country
Movie    United States     1847
         India              852
         United Kingdom     193
         Canada             118
         Egypt               89
         Spain               89
         Turkey              73
         Philippines         70
         France              69
         Japan               69
TV Show  United States      695
         United Kingdom     283
         Japan              155
         South Korea        147
         India               71
         Taiwan              68
         Canada              59
         France              46
         Spain               45
         Australia           44
Name: country, dtype: int64
```

EDA

CONTENT RELEASED OVER THE YEARS

- Growth in the number of **movies** on Netflix is **much higher** than **tv shows**.
- **Most of the content** available was released between **2010** and **2020**.
- **Highest** number of **movies** got released in **2017** & **2018** and **tv shows** got released in **2019** & **2020**.
- Very few movies, and tv shows got released before the year **2010** and in **2021**. It is due to very little data collected from the year **2021**.



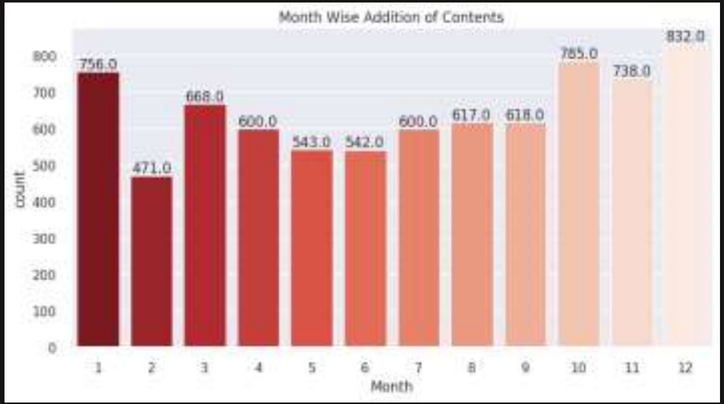


```
Number of Shows Released in Each Year:
type     release_year
Movie    2017            742
         2018            734
         2016            642
         2019            582
         2020            411
         2015            398
         2014            244
         2013            202
         2012            158
         2010            135
TV Show  2020            457
         2019            414
         2018            386
         2017            268
         2016            239
         2015            156
         2014             90
         2013             63
         2012             60
         2011             39
Name: release_year, dtype: int64
```

EDA

CONTENT ADDED OVER THE MONTHS

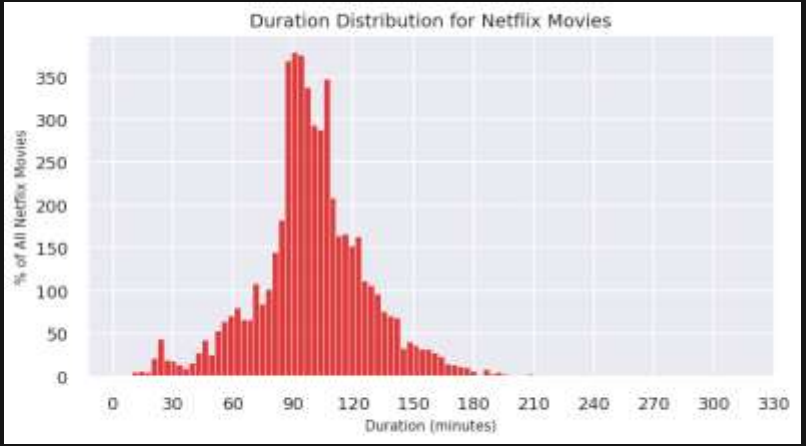
- **October**, **November**, **December**, and **January** are months in which many tv shows and movies get **uploaded** to the platform.
- It might be **due to** the **winter**, as in these months people may **stay at home** and watch **tv shows** and **movies** in their free time.



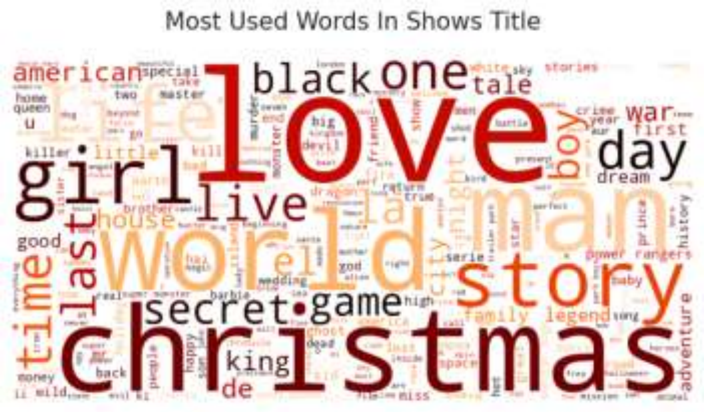
| EDA

NETFLIX MOVIES DURATION

- **Most number of movies** on the Netflix platform are last for **90** to **120 minutes**.
- Very **few movies** are of **length** more than **200 minutes**.



EDA



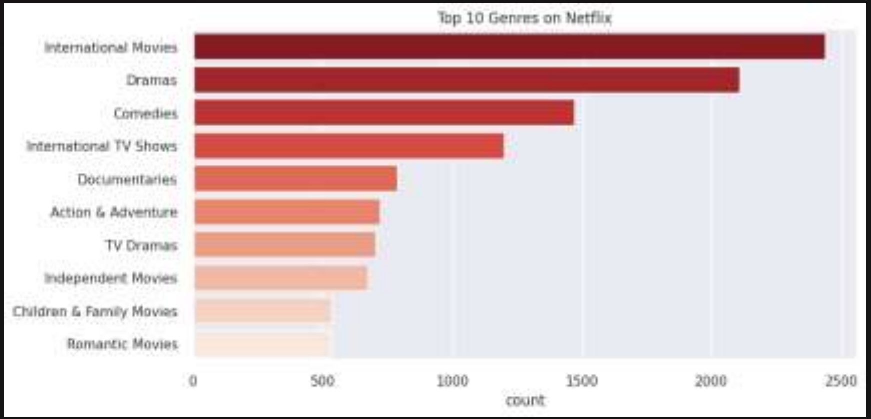
MOST USED WORDS IN SHOWS TITLE

- Most **repeated words** in **title** include **Christmas**, **Love**, **World**, **Man**, and **Story**.
- We saw that most of the movies and tv shows **got added** during the **winters**, which tells why **Christmas** appeared many times in the titles.

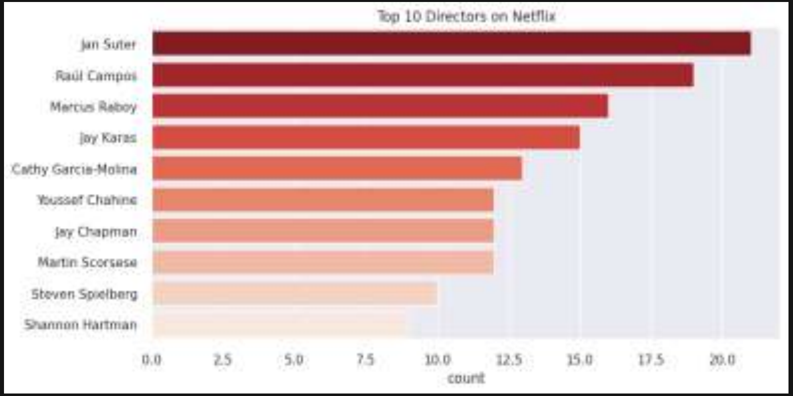
EDA

TOP 10 GENRES

- In terms of genres, **international movies** takes the cake surprisingly followed by **dramas** and **comedies**.
- Even though the **United States** has the **most content** available, it looks like **Netflix** has decided to **release** a ton of **international movies**.



EDA



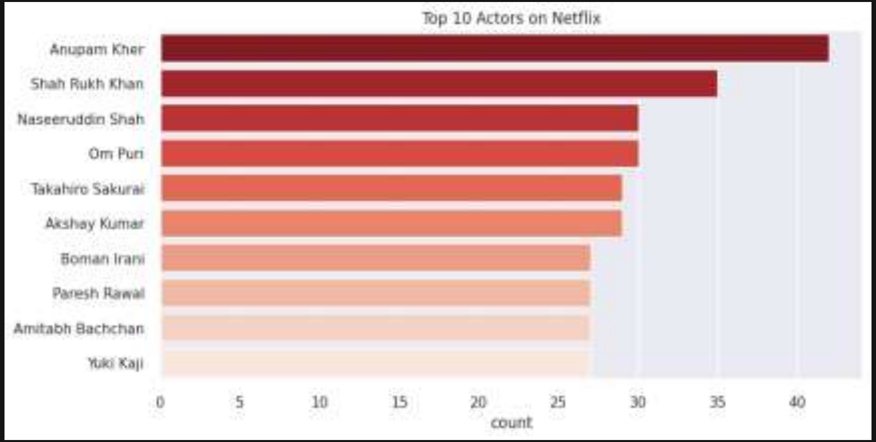
TOP 10 DIRECTORS

- **Jan Suter**, **Raúl Campos**, **Marcus Raboy**, **Jay Karas**, **Cathy Garcia-Molina** are the **top 5 directors** which highest number of movies and tv shows are available in netflix.
- As we stated previously regarding the top genres, it's no surprise that the **most popular directors** on **Netflix** with the most titles are **mainly international** as well.

EDA

TOP 10 ACTORS

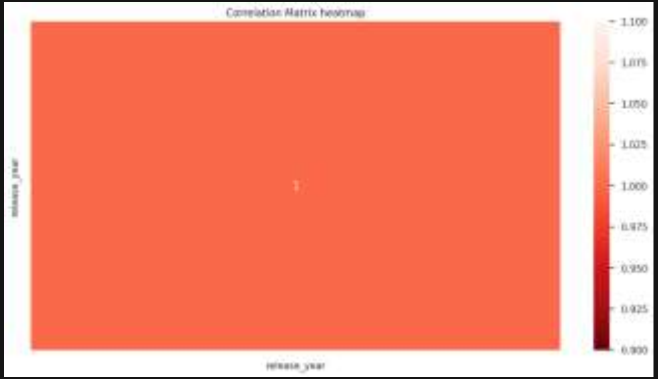
- The **actors** in the **top ten list** of **most numbers tv shows** and **movies** are from **India**.
- **Anupam Kher** and **Shah Rukh Khan** have **30 above content** alone in **netflix**.

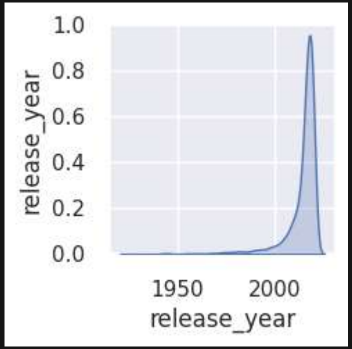


EDA

CORRELATION HEATMAP & PAIR PLOT

- Since there is **only one value** in **dataframe** of **integer type**, we are unable to visualize the **correlation matrix heatmap** and **pair plot** as well.

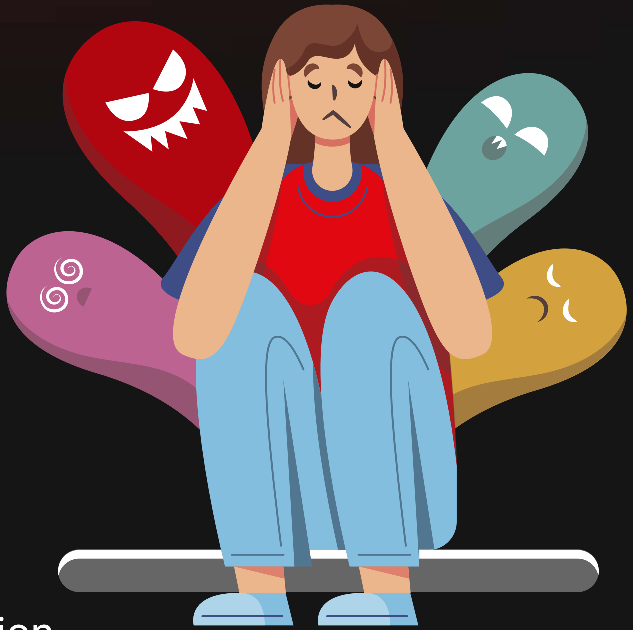




05

HYPOTHESIS TESTING

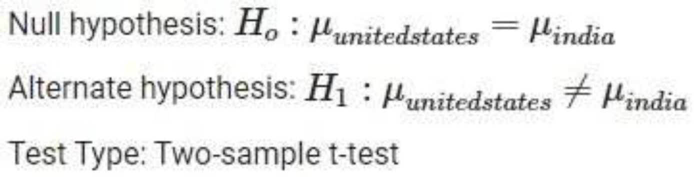
Observation of experiments under given assumption



HYPOTHESIS TESTING

AVERAGE NUMBER OF MOVIES ON NETFLIX IN UNITED STATES IS HIGHER THAN THE MOVIES ON NETFLIX IN INDIA

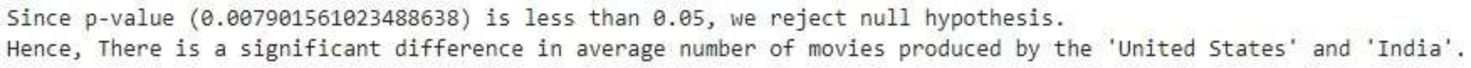
- I selected the **two-sample t-test** for this analysis as it is suitable for **comparing** the **means** of **two independent samples**.
- By applying this test, I was able to calculate the **p-value** and **determine** if there is a **significant difference** in the number of movies **between** the **two countries**.



Null hypothesis: $H_o : \mu_{unitedstates} = \mu_{india}$

Alternate hypothesis: $H_1 : \mu_{unitedstates} \neq \mu_{india}$

Test Type: Two-sample t-test

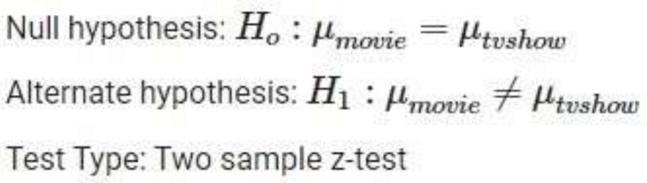


```
Since p-value (0.007901561023488638) is less than 0.05, we reject null hypothesis.
Hence, There is a significant difference in average number of movies produced by the 'United States' and 'India'.
```

HYPOTHESIS TESTING

NUMBER OF MOVIES AVAILABLE ON NETFLIX IS GREATER THAN THE NUMBER OF TV SHOWS AVAILABLE ON NETFLIX

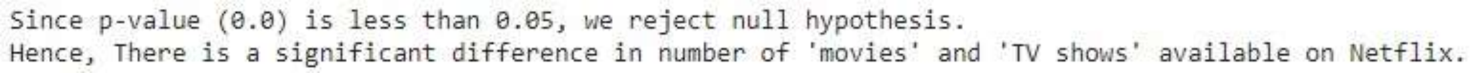
- The **two sample z-test** is used to **determine** if there is a **significant difference** between **two categorical variables**.
- In this case, I wanted to test if there was a **significant difference** between the number of **movies** and **tv shows** available on **Netflix**.



Null hypothesis: $H_o : \mu_{movie} = \mu_{tvshow}$

Alternate hypothesis: $H_1 : \mu_{movie} \neq \mu_{tvshow}$

Test Type: Two sample z-test



```
Since p-value (0.0) is less than 0.05, we reject null hypothesis.
Hence, There is a significant difference in number of 'movies' and 'TV shows' available on Netflix.
```

TEXTUAL DATA PREPROCESSING

06

Transforming text into a clean and consistent form

TEXTUAL DATA PREPROCESSING

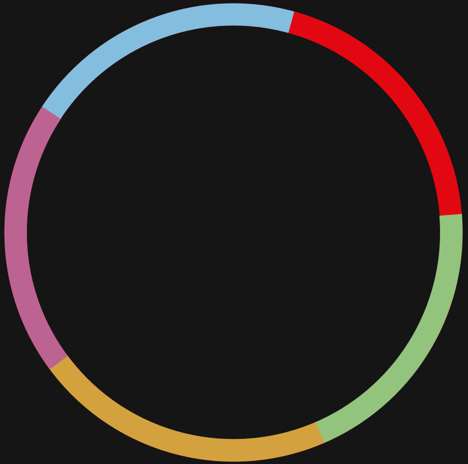
WORK PROCESS

01 **TOKENIZATION**

Replacing **sensitive data** with unique identification **symbols**

02 **TEXT REMOVAL**

Removing **punctuation**, **numbers**, **stopwords** etc.



STEMMING 03

Reducing **words** to their base form (**root form**)

LEMMATIZATION 04

Grouping together **words** with their **root form**

POS TAGGING 05

Process of **finding** the sequence of **tags**

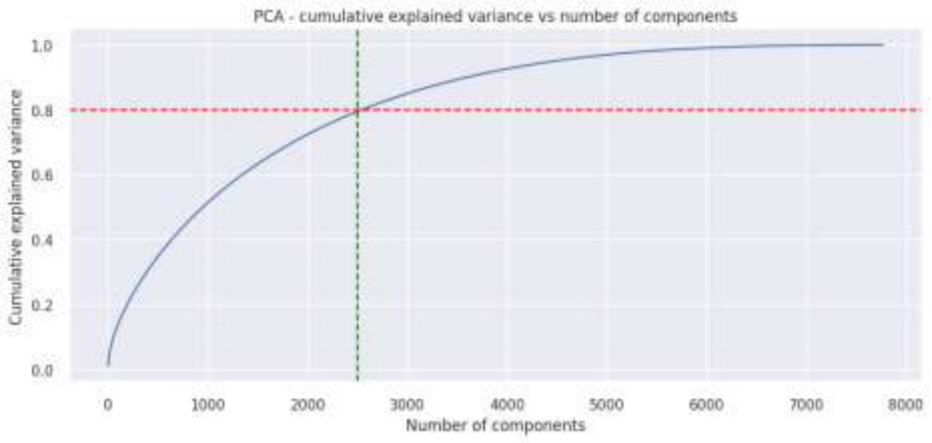
07

DIMENSIONALITY REDUCTION

Reducing the feature set’s dimension

DIMENSIONALITY REDUCTION

- **Principal Component Analysis** (**PCA**) was used to **reduce** the **dimensionality** of data.
- Captured more than **80%** of the **variance** by **reducing** the **components** to **2500**.



08

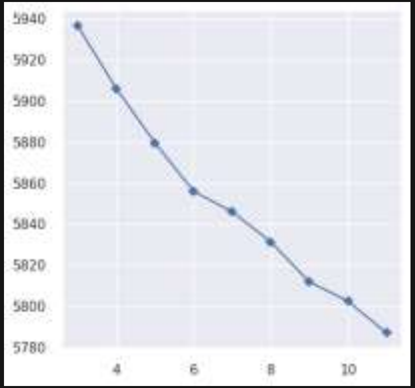
MODEL IMPLEMENTATION

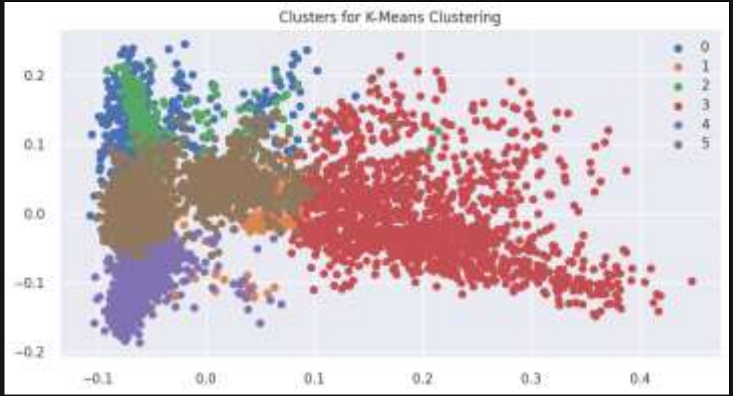
Train ML Algorithms to get best model

MODEL IMPLEMENTATION

K-MEANS CLUSTERING

- **K-means** is a **centroid-based** clustering algorithm, where we **calculate** the **distance** between **each data point** and a **centroid** to assign it to a cluster.
- Here **optimal number** of **clusters** is **6** by using the **elbow method**.

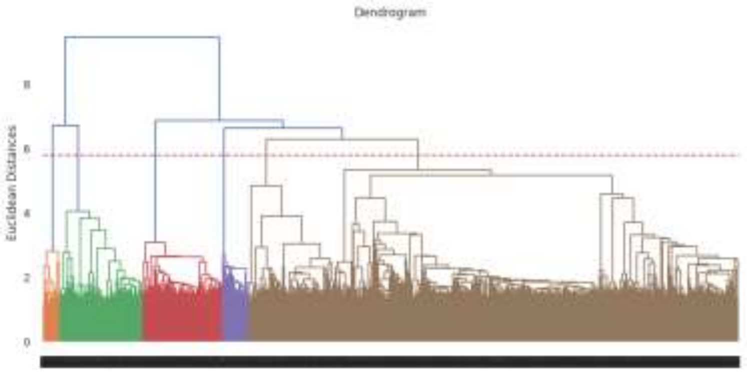


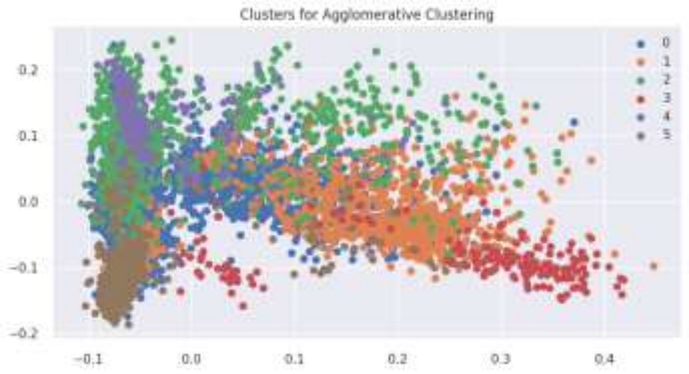


MODEL IMPLEMENTATION

HIERARCHICAL CLUSTERING

- From **dendrogram** we get the **optimal number** of **clusters** is **6**.
- Used **agglomerative clustering** here, which is a type of **hierarchical clustering algorithm**. It helps us to **divides** the **population** into **several clusters**.

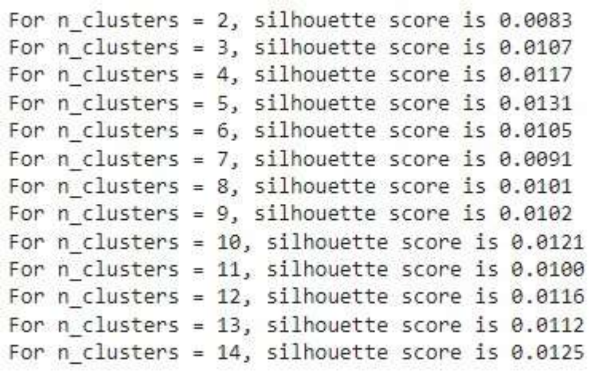




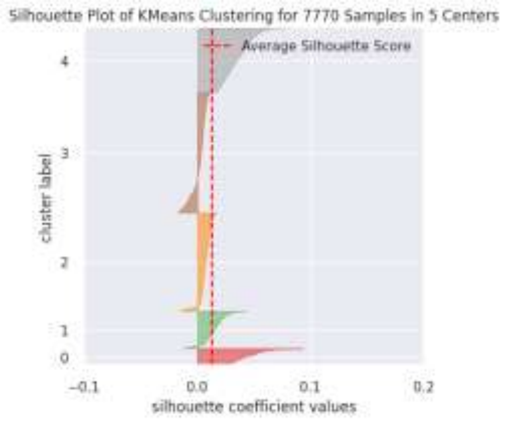
MODEL IMPLEMENTATION

SILHOUETTE SCORE FOR CLUSTERING

- **Silhouette score** is **highest** for the **cluster 5**, so the **optimal number** of **clusters** will be **5**.
- **Silhouette score** is a metric used to calculate the goodness of a clustering technique. Its value ranges from **-1** to **1**.



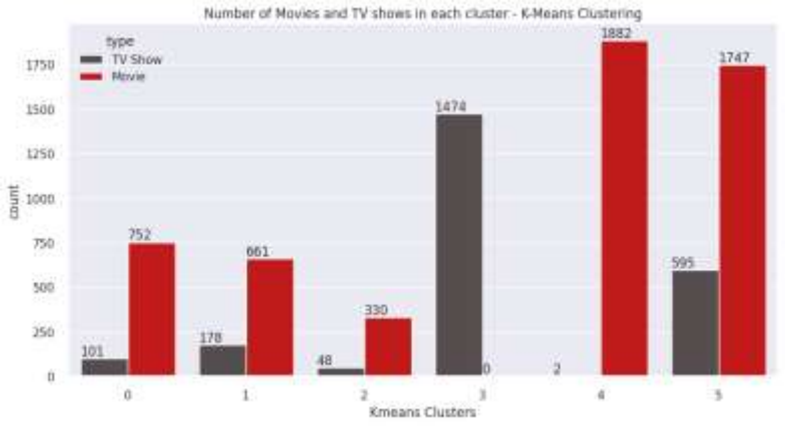
```
For n_clusters = 2, silhouette score is 0.0083
For n_clusters = 3, silhouette score is 0.0107
For n_clusters = 4, silhouette score is 0.0117
For n_clusters = 5, silhouette score is 0.0131
For n_clusters = 6, silhouette score is 0.0105
For n_clusters = 7, silhouette score is 0.0091
For n_clusters = 8, silhouette score is 0.0101
For n_clusters = 9, silhouette score is 0.0102
For n_clusters = 10, silhouette score is 0.0121
For n_clusters = 11, silhouette score is 0.0100
For n_clusters = 12, silhouette score is 0.0116
For n_clusters = 13, silhouette score is 0.0112
For n_clusters = 14, silhouette score is 0.0125
```



MODEL IMPLEMENTATION

FINAL PREDICTION MODEL

- Selected **K-Mean Clustering model** as the **best model** for our data.
- The clusters are **well divided** in this model and through this cluster we can **know** what **type of data** is in **which cluster**.



MODEL IMPLEMENTATION

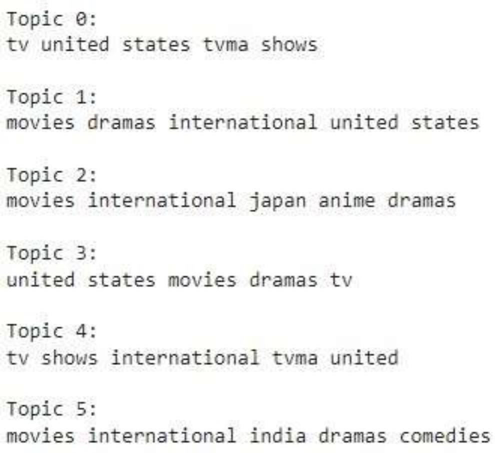
TOPIC MODELING

- Used **topic modeling** instead of feature importance and model explainability.
- We can get **topic wise feature importance**. **Assume** that the **clusters** are **topics**.
- Used **CountVectorizer** process for **Vectorization of data** and **Latent Dirichlet Allocation** for **building a topic**.

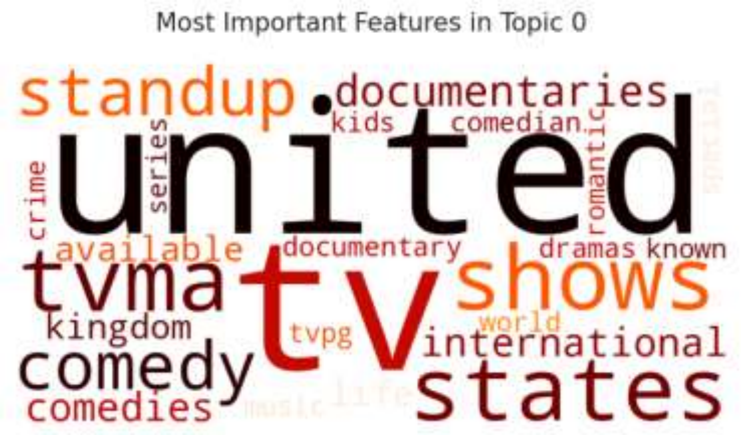
MODEL IMPLEMENTATION

TOPIC MODELING

- Most **important features**, which we are get from each **topics**:



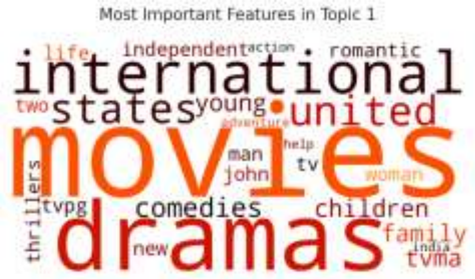
```
Topic 0:
tv united states tvma shows

Topic 1:
movies dramas international united states

Topic 2:
movies international japan anime dramas

Topic 3:
united states movies dramas tv

Topic 4:
tv shows international tvma united

Topic 5:
movies international india dramas comedies
```

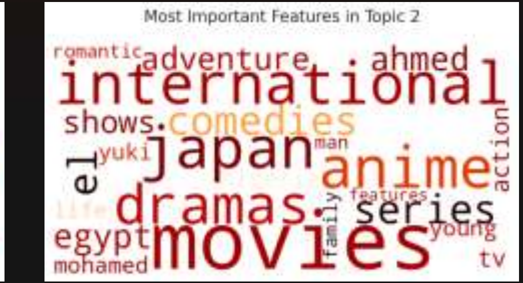


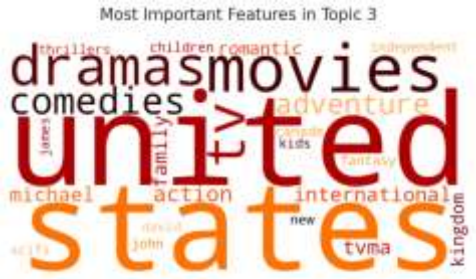
Most Important Features in Topic 0

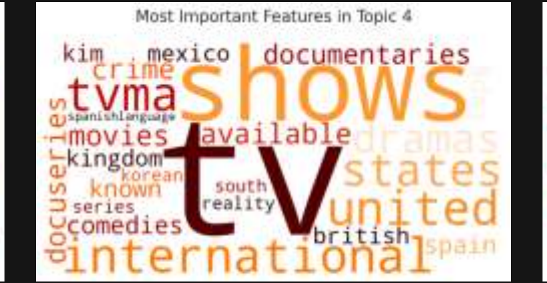
MODEL IMPLEMENTATION

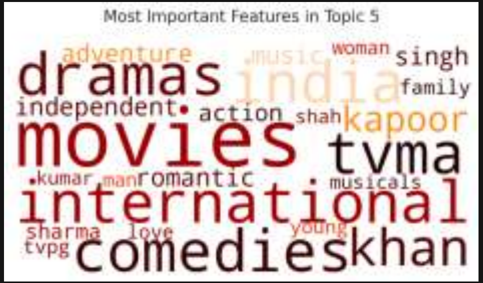
WORDCLOUD FOR OTHER TOPICS



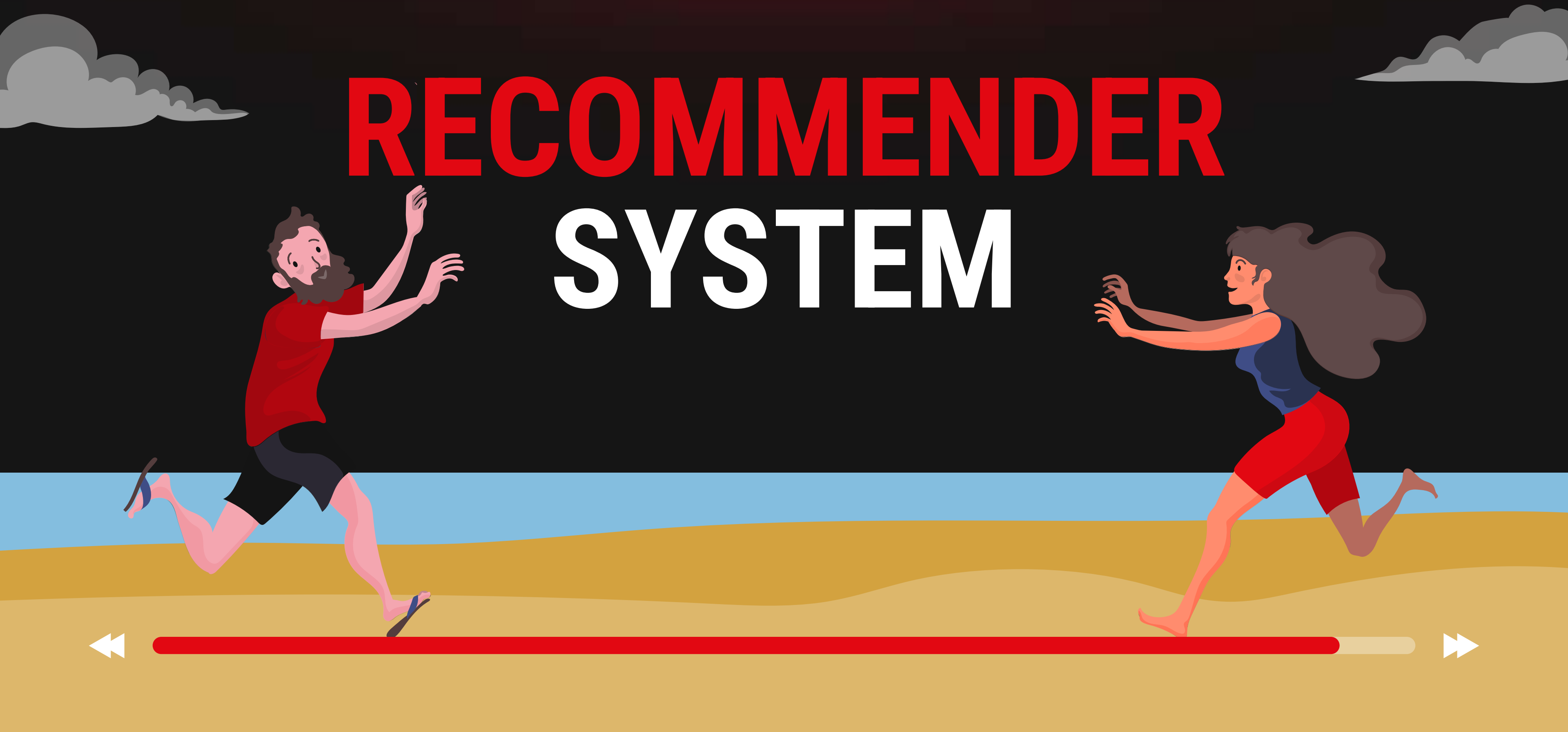








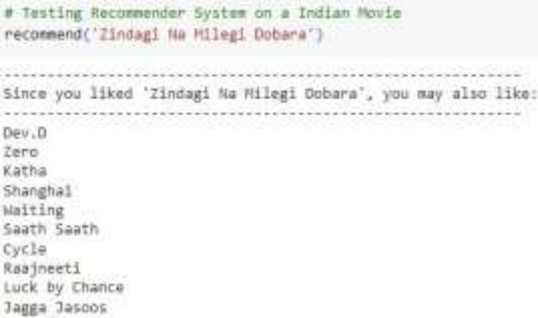
RECOMMENDER SYSTEM



RECOMMENDER SYSTEM

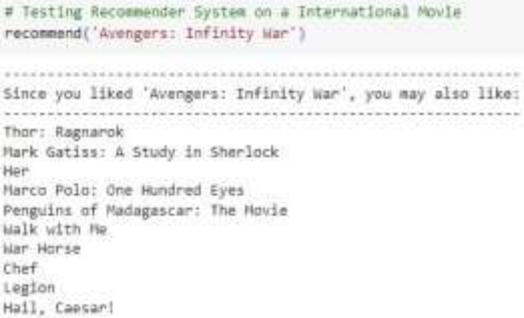
TOP 10 RECOMMENDED MOVIES/ TV SHOWS

- **Content-based recommender system** on the basis of **cosine similarity score**.



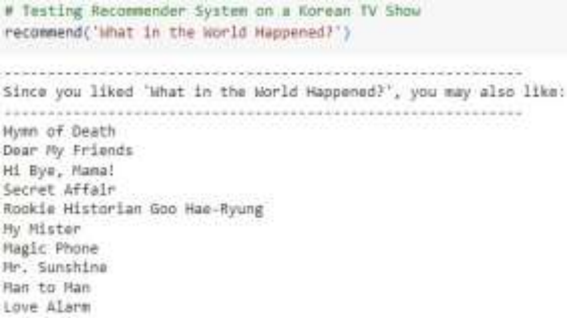
```
# Testing Recommender System on a Indian Movie
recommend('Zindagi Na Milegi Dobara')
```

```
-----------------------------------------------------------------------
Since you liked 'Zindagi Na Milegi Dobara', you may also like:
-----------------------------------------------------------------------
Dev.D
Zero
Katha
Shanghai
Waiting
Saath Saath
Cycle
Raajneeti
Luck by Chance
Jagga Jasoos
```



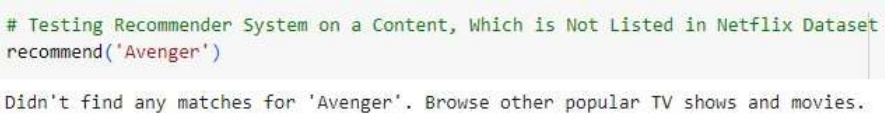
```
# Testing Recommender System on a International Movie
recommend('Avengers: Infinity War')
```

```
-----------------------------------------------------------------------
Since you liked 'Avengers: Infinity War', you may also like:
-----------------------------------------------------------------------
Thor: Ragnarok
Mark Gatiss: A Study in Sherlock
Her
Marco Polo: One Hundred Eyes
Penguins of Madagascar: The Movie
Walk with Me
War Horse
Chef
Legion
Hail, Caesar!
```



```
# Testing Recommender System on a Korean TV Show
recommend('What in the World Happened?')
```

```
-----------------------------------------------------------------------
Since you liked 'What in the World Happened?', you may also like:
-----------------------------------------------------------------------
Hymn of Death
Dear My Friends
Hi Bye, Mama!
Secret Affair
Rookie Historian Goo Hae-Ryung
My Mister
Magic Phone
Mr. Sunshine
Man to Man
Love Alarm
```



```
# Testing Recommender System on a Content, Which is Not Listed in Netflix Dataset
recommend('Avenger')
```

```
Didn't find any matches for 'Avenger'. Browse other popular TV shows and movies.
```

CONCLUSION

- Analysis revealed that Netflix has a **greater number** of **movies** than **TV shows**.
- Clustering **TV shows** and **movies** based on their similarities and differences, created a **content-based recommender system** that recommends **top 10 shows** to users based on their **viewing history**.

CAPSTONE PROJECT

NETFLIX MOVIES AND TV SHOWS CLUSTERING

By - Arindam Paul
(aripaul05@gmail.com)

THANK YOU!