Capstone Project-4

Netflix Movies & TV Shows Clustering

Supervised Machine Learning (Classification)

TEAM MEMBERS

Adi Ingrole

Mandar Khatavkar

Will Be Discussing On:

1. Problem Statement
2. Introduction
3. Data Cleaning
4. Data Pre-processing
5. Exploratory Data Analysis
7. Data Pre-processing for Clustering
8. K-Means Clustering
9. Recommender System
10. Conclusion

Probelem statment

1. 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.
2. 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.
3. Integrating this dataset with other external datasets such as IMDB ratings, rotten tomatoes can also provide many interesting findings.
4. 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 Description

- **The data was collected from Flixable which is third party Netflix search engine. The dataset consists of movies and TV shows data till 2019. The dataset has 7787 rows of data.**
- **The dataset consists of eleven textual columns and one numeric column.**

Attribute Information :

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

Data Description

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 Release year 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 :** Genre
12. **description:** The Summary description

3.Data Cleaning

1. Duplicate Values Treatment:

- Duplicate values dose not contribute anything to accuracy of results.
- Our dataset dose not contains any duplicate values.

2. Null Values Treatment:

- Director feature have more than 30% of null values. So, dropping feature director.
- Country feature have 6.51% of null values. Filling null values by mode of feature.
- Cast feature have 9.22% of null values. Filling null values by 'missing'.
- Rating feature have 0.09% of null values. Filling null values by mode of feature.
- Date_added feature have 0.12% of null values. Dropping rows corresponding to null values.

4. Data Pre-processing

1. Data Type Change:

- Features in their appropriate data type provides better understanding and workability on that data.
- Date_added feature have object datatype. Converting to datetime.
- Duration is in combination of integer values and text. Removing text part so as to get integer datatype.

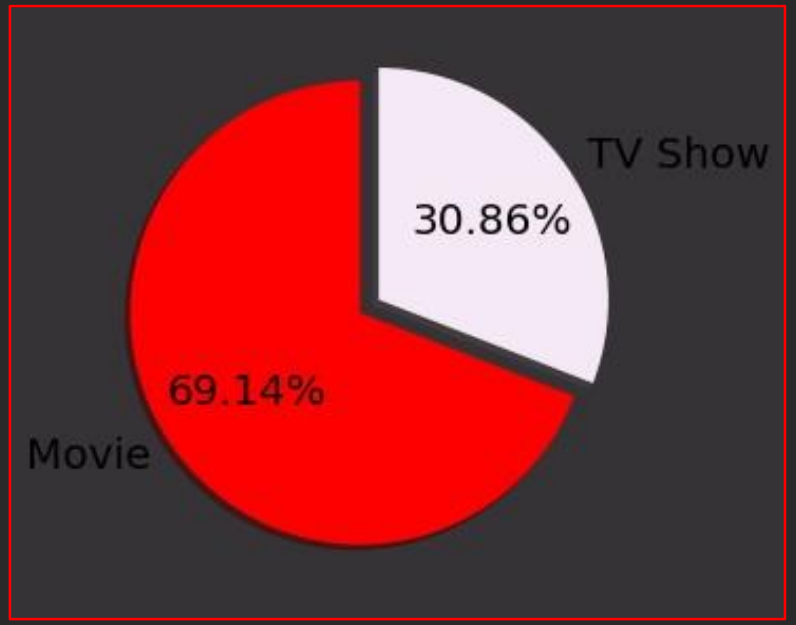
2. New Features:

- From the feature date_ added; extracted year, month and day to form new columns by name of year, month and day respectively.

5. Exploratory Data Analysis

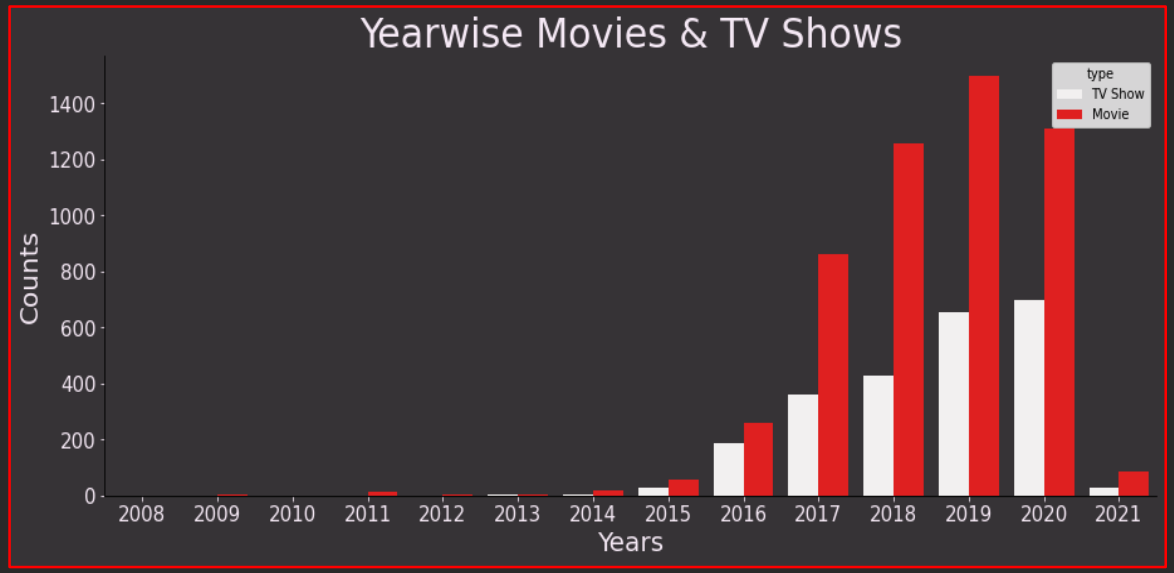
Movies vs. TV Shows

Movies uploaded on Netflix are more than twice the TV Shows uploaded.



EDA (Continued)

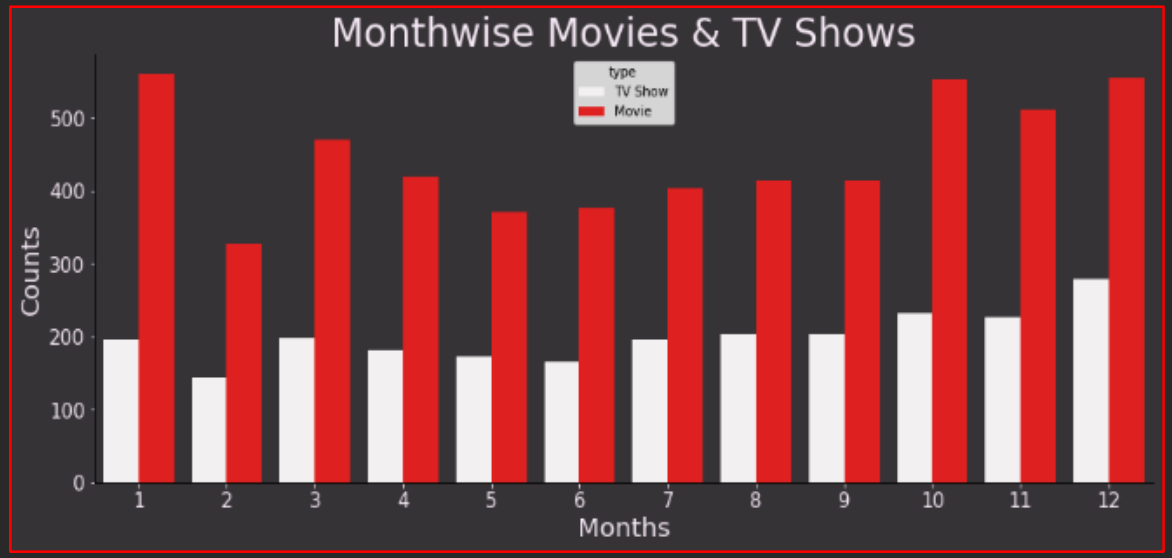
On Year Basis



- TV shows are increasing continuously.
- Movies were increasing continuously but after 2020 there is fall.

EDA (Continued)

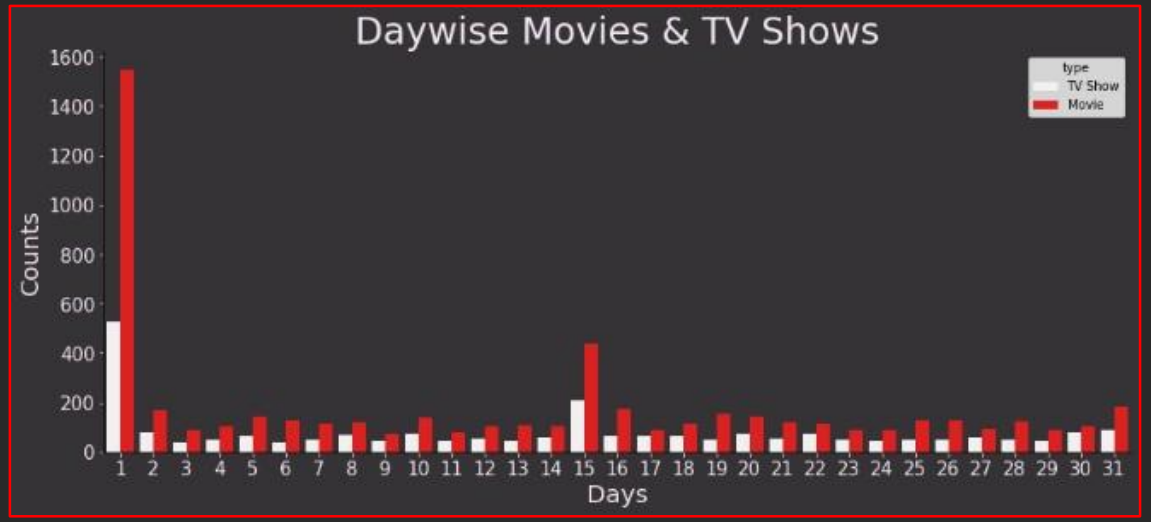
On Month Basis



- From October to January, maximum number of movies and TV shows were added.
- Possible reason for that is, during this period of time events such as Christmas, New Year and several holidays takes place.

EDA (Continued)

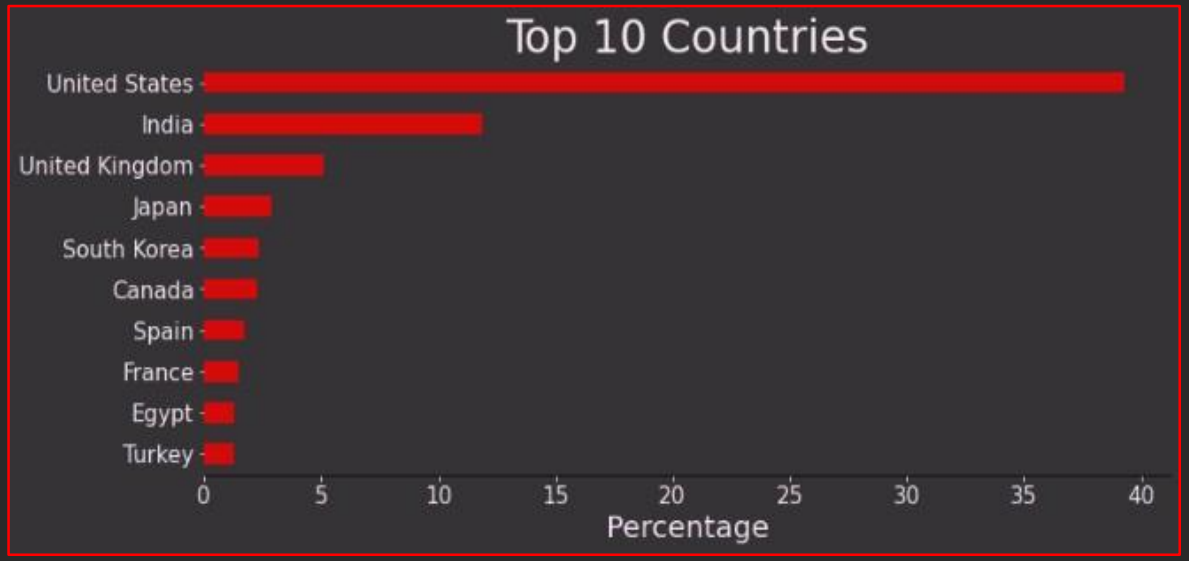
On Day Basis



- Maximum number of movies and TV shows added on start of the month followed by mid of the month.

EDA (Continued)

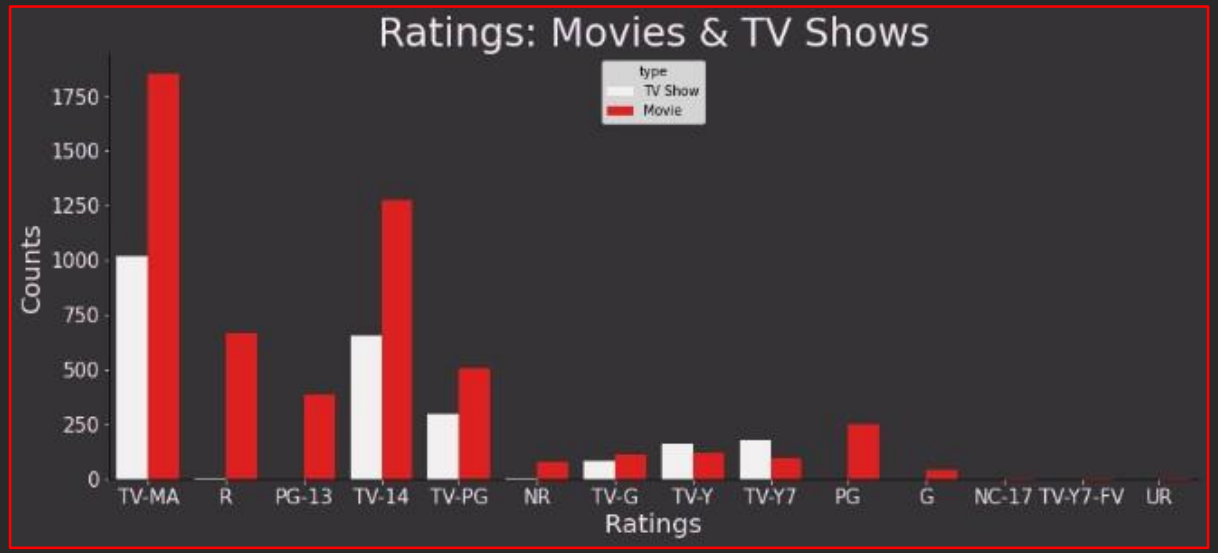
Worldwide Presence



- United State tops in the list of maximum number of movies and TV shows, followed by India, UK and Japan.

EDA (Continued)

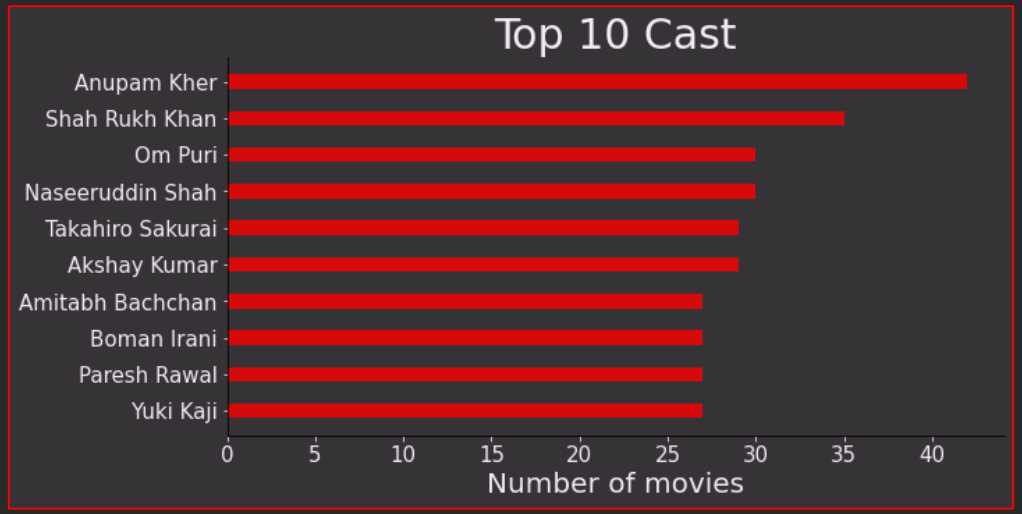
Ratings



- Maximum of the movies as well as TV shows are for matures only.

EDA (Continued)

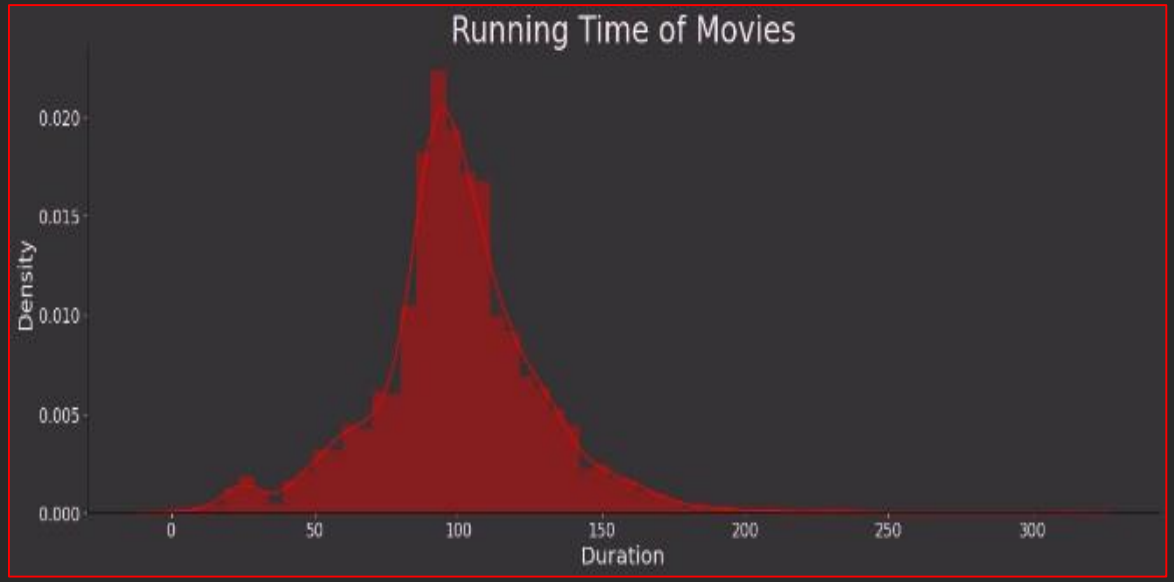
Cast



- Anupam Kher top from the list of casts having maximum number of movies and TV shows.

EDA (Continued)

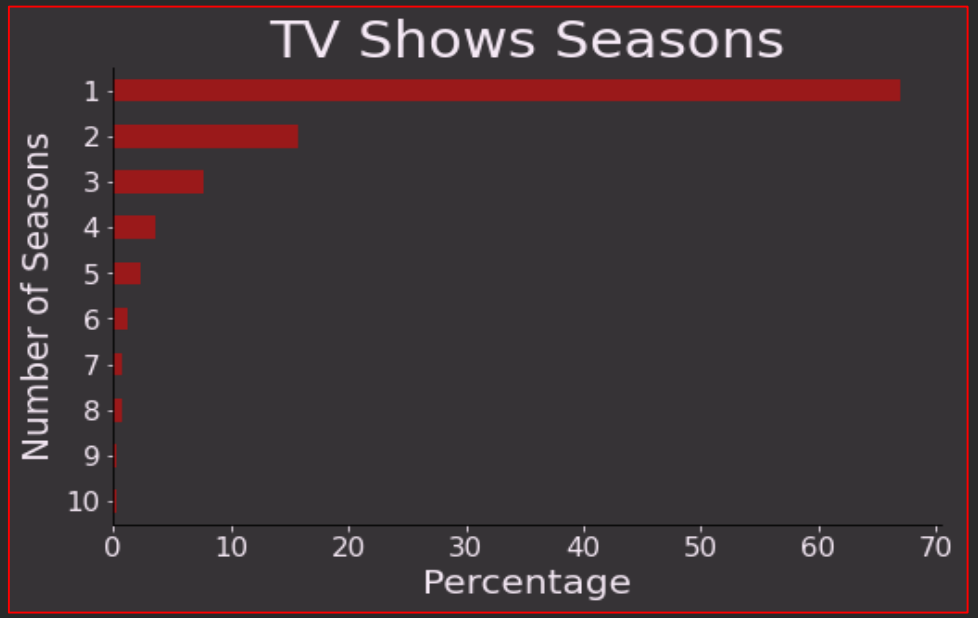
Running Time of Movies



- Majority of movies have running time in between 50 to 150 min.

EDA (Continued)

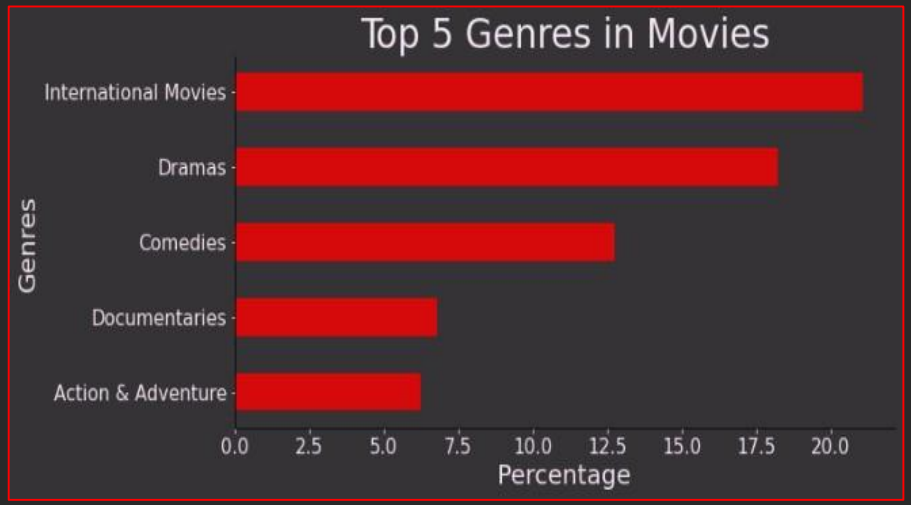
Seasons of TV Shows

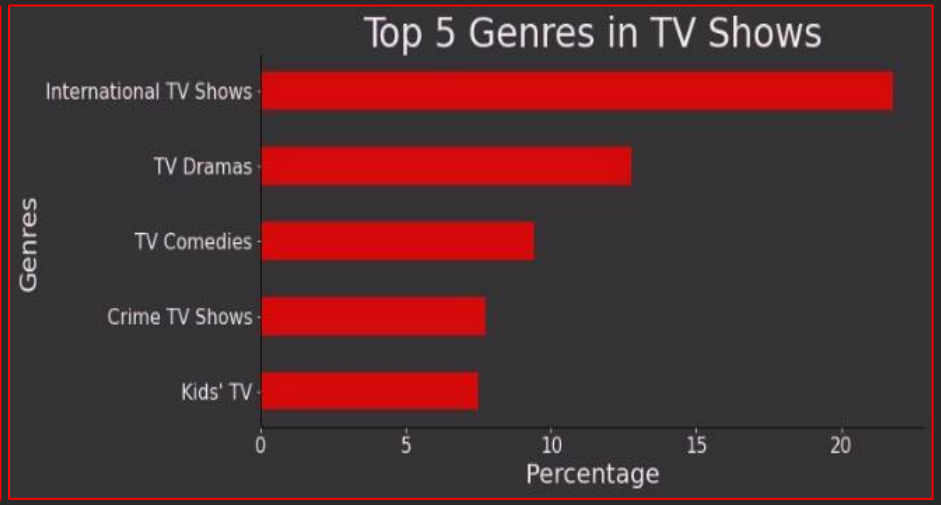


- Almost 68% of TV shows consist of single season only.

EDA (Continued)

Genres

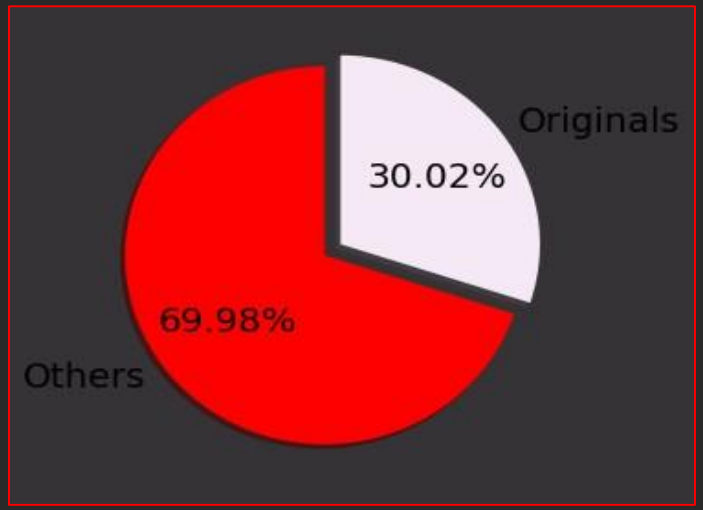


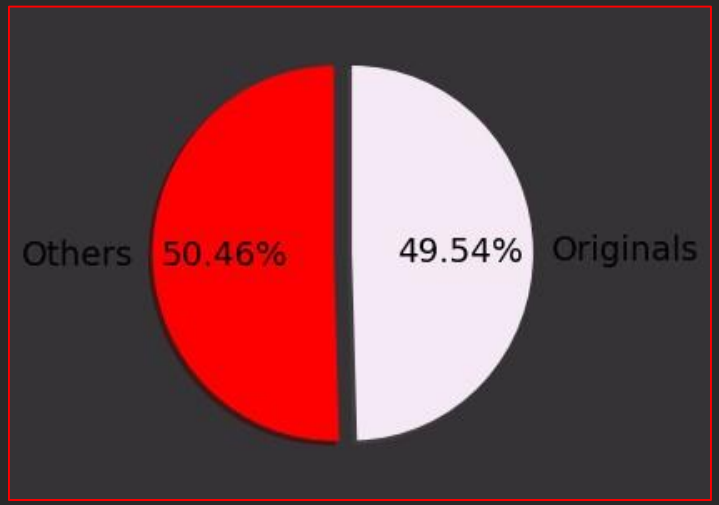


- Top 3 genres are exactly same for movies and TV shows.
- Dramas genres hit all over the world.

EDA (Continued)

Netflix Original





- 30% movies released on Netflix as Netflix originals.
- 50% TV shows originally from Netflix.

6. Data Pre-processing for Clustering:

AI

1. Removing Punctuation

- Punctuations does not carry any meaning clustering.
- So, removing punctuations helps to get rid of unhelpful parts of the data,

or noise.

2. Removing Stop words

- Stop words are basically a set of commonly used words in any language, not just English.
- If we remove the words that are very commonly used in a given

language, we can focus on the important words instead.

AI

3. Stemming

- Stemming is the process of removing a part of a word, or reducing a word to its stem or root.

- Applying stemming to reduce words to their basic form or stem, which may or may not be a legitimate word in the language.

7.K-Means Clustering

AI

K-means algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups where each data point belongs to only one group.

7.1 Vectorization:

- Here we have textual data.
- Clustering algorithms cannot understand textual data.
- So, we use vectorization technique to convert textual data to numerical vectors.

7.2 Elbow Curve:

- The Elbow Curve is one of the most popular methods to determine this optimal value of k.
- The elbow curve uses the sum of squared distance (SSE) to choose an ideal value of k based on the distance between the data points and their assigned clusters.

7.3 Silhouette Score:

- 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.

8. Recommender System:

- Recommender systems are the systems that are designed to recommend things to the user based on many different factors.
- It finds out the match between user and item and imputes the similarities between users and items for recommendation.

8.1 Cosine Similarity:

- Cosine similarity is a metric that measures the cosine of the angle between two vectors projected in a multi-dimensional space.

9. Conclusion

1. Movies uploaded on Netflix are more than twice the TV Shows uploaded.
2. TV shows and movies are increasing continuously but in 2020 there is
   drop in number of movies.
3. From October to January, maximum number of movies and TV shows were added.
4. Maximum number of movies and TV shows were either on start of the month or mid of the month.
5. United State tops in the list of maximum number of movies and TV shows
   followed by India, UK and Japan.
6. Maximum of the movies as well as TV shows are for matures only.

7. Anupam Kher top from the list of casts having maximum number of
   movies and TV shows.
8. Majority of movies have running time of between 50 to 150 min.
9. Almost 68% of TV shows consist of single season only.
10. Top 3 genres are exactly same for movies and TV shows.
11. Dramas genres hit all over the world.
12. 30% movies and 50% TV shows are Netflix Originals.
13. Clustering done by K-Means Clustering, found optimal number of clusters
    equal to 25 with highest Silhouette Score.
14. Recommender system using cosine similarity performs well on data.

THE END!