NETFLIX MOVIES AND TV SHOWS CLUSTERING

Amol Kale
Data science
strainee,
AlmaBetter

Abstract:

Netflix is a subscription-based streaming service that allows our members to watch TV shows and movies without commercials on an internet-connected device. One can also download TV shows and movies to your iOS, Android, or Windows 10 device and watch without an internet connection. Netflix content varies by region and may change over time. One can watch from a wide variety of award-winning Netflix Originals, TV shows, movies, documentaries, and more. Netflix have data which were released in the past also. Not all the movies and TV shows are released first on Netflix itself.

Abundant data about movies and TV shows is present on Netflix platform. Clustering of this huge data into clusters may lead develop better recommender system.

Grouping of movies or TV shows in different categories can be achieved by clustering. Which will be useful in understanding of data. To engage user on Netflix, user must get movies and TV shows of same interest of user in past.

Keywords: unsupervised machine learning, Netflix movies and tv shows, k-means clustering, hierarchical clustering, recommender system, EDA.

1.Problem Statement

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. So, we need to verify above made statements are valid or not.

The main objective of this project is clustering of Netflix movies and TV shows in optimum number of clusters.

2. Introduction

This dataset consists of tv shows and movies available on Netflix as of 2019. The dataset is collected from Flexible which is a third-party Netflix search engine. Movies and TV shows can be categorized on several factors such as genres, cast and director of movies or TV shows etc.

Following are descriptions of features given in dataset:

- **show_id:** Unique ID for every Movie and TV Show.
- **type:** Identifier - A Movie or TV Show.
- **title:** Title of the Movie or TV Show

- **director:** Director of the Movie or TV show.
- **cast:** Actors involved in the movie or TV show
- **country:** Country where the movie or TV show was produced.
- **date_added:** Date on which it was added on Netflix
- **release_year:** Actual release year of the movie or TV show.
- **rating:** Content Ratings of the movie and TV shows.
- **duration:** Total Duration of movie in minutes or total number of seasons of TV show.
- **listed_in:** Genres of movies and TV shows.
- **description:** The Summary of the movies or TV show.

3. Steps involved:

- **Data Cleaning**
- **Data Preprocessing**
- **Exploratory Data Analysis**
- **Data Preprocessing for Clustering**
- **K-Means Clustering**
- **Recommender System**
- **Conclusions**

4. Data Cleaning:

4.1 Duplicate Values

Duplicate values dose not contribute anything to accuracy of results. Large duplicate may lead to slower computation and higher space requirement. Our dataset dose not contains any duplicate values.

4.2 Null values Treatment

Null values which might tend to disturb our accuracy hence we dropped or fill them at the beginning of our project in order to get a better result.

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', only for EDA purpose. 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 this null values.

5. Data Preprocessing:

5.1 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 datatype from object datatype.

Duration feature have object datatype. Converting to int datatype from object datatype. Duration is in combination of int values and string. Removing string part so as to get int datatype.

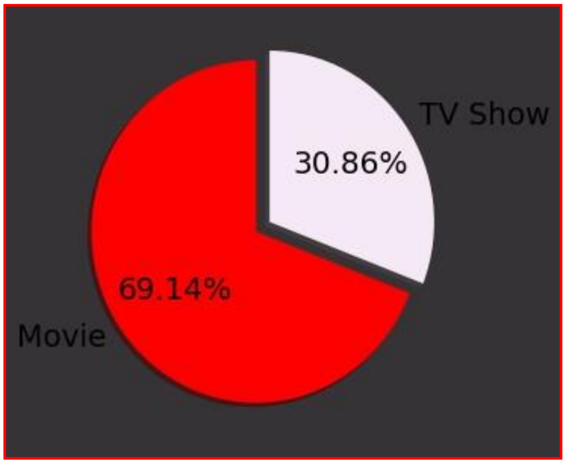
5.2 New Features:

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

6. Exploratory Data Analysis:

6.1 Movies vs. TV Shows

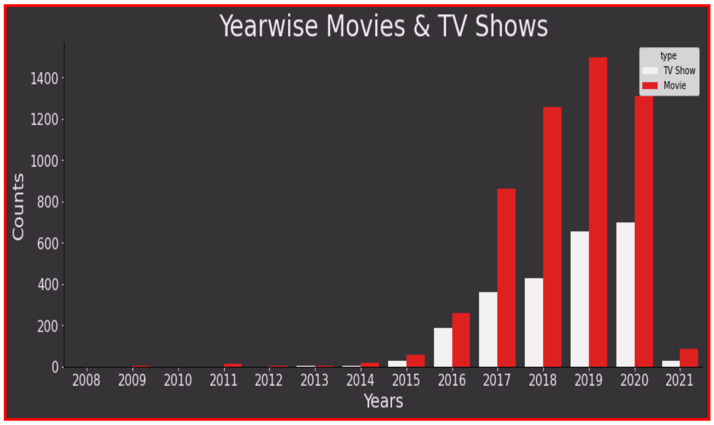
This dataset contains data about movies and TV shows which were added on Netflix. Following pie chart shows percentage of available content,



Movies uploaded on Netflix are more than twice the TV Shows uploaded. This does not implies that movies are more indulging than that of TV Shows. Because TV shows may have several seasons which consist of number of episodes. Duration of TV shows are much more that of movies.

6.2 On Year Basis

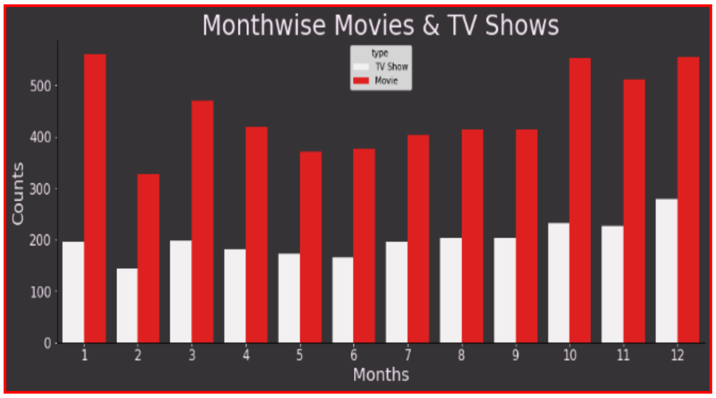
Popularity of Netflix increasing day by day. Following combined bar chart shows year wise number of movies and TV shows,



TV shows are increasing continuously. Movies were increasing continuously but after 2019 there is fall.

6.3 On Month Basis

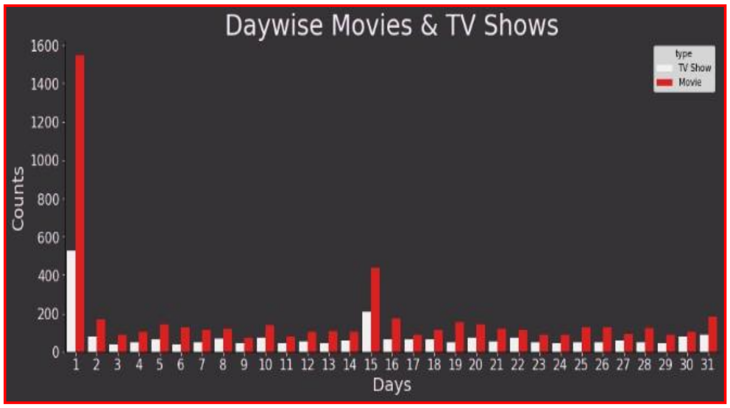
Festivals and holidays might best time add new movies and TV shows. To look for which part of year have the greatest number of movies and TV shows added. Following combined bar chart shows month wise number of movies and TV shows,



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.

6.4 On Day Basis

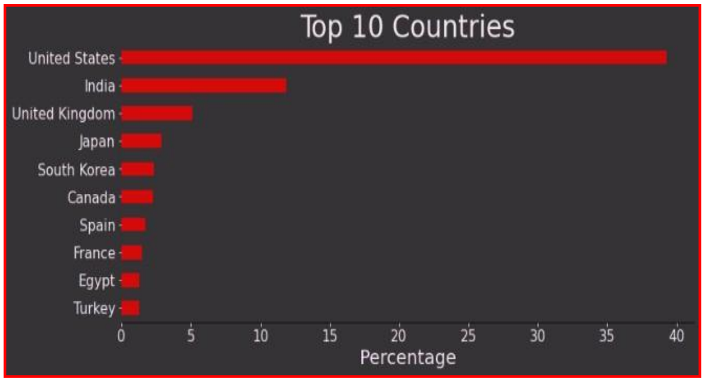
Following combined bar chart shows day wise number of movies and TV shows,



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

6.5 Worldwide Presence

Popularity Netflix is all over the world. To look for its spread over world, created bar plot of country wise with percentage of movies and TV shows in decreasing order. Following bar chart shows top ten countries with maximum number of movies and TV shows.



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

6.6 Ratings

For Movies:

G: Kids

PG: Older Kids (7+)

PG-13: Teens (13+)

NC-17, NR, R, Unrated: Adults (18+)

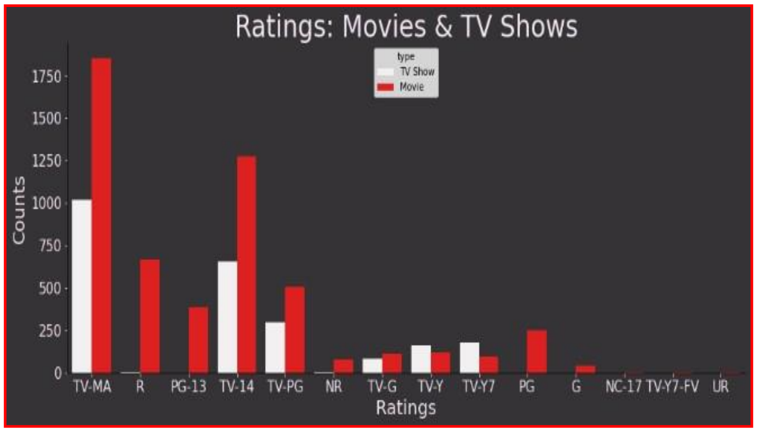
For TV Shows:

TV-G, TV-Y: Kids

TV-Y7/FV/PG: Older Kids (7+)

TV-14: Young Adults (16+)

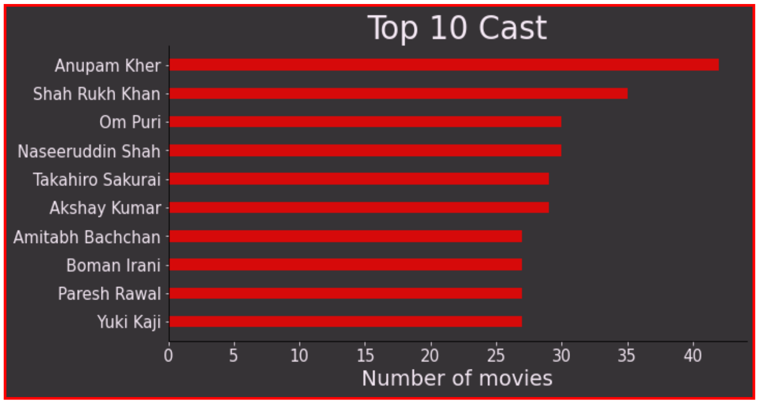
TV-MA: Adults (18+)



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

6.7 Cast

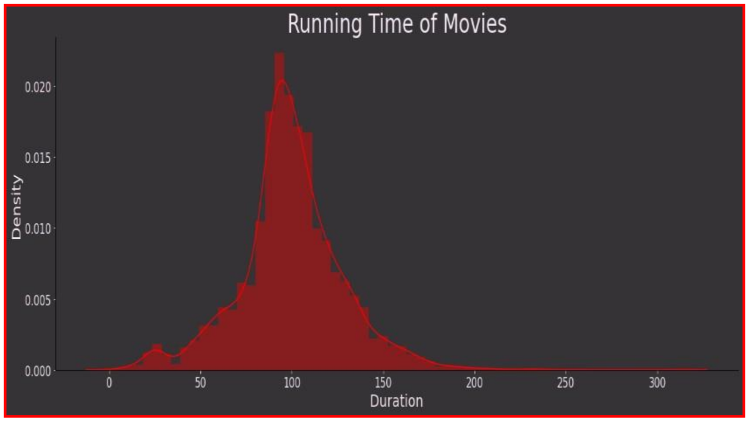
Each movie and TV shows have cast and crew. Popular cast have large number of movies or TV shows. Following plot shows top 10 cast acted in movies or TV shows.



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

6.8 Running Time of Movies

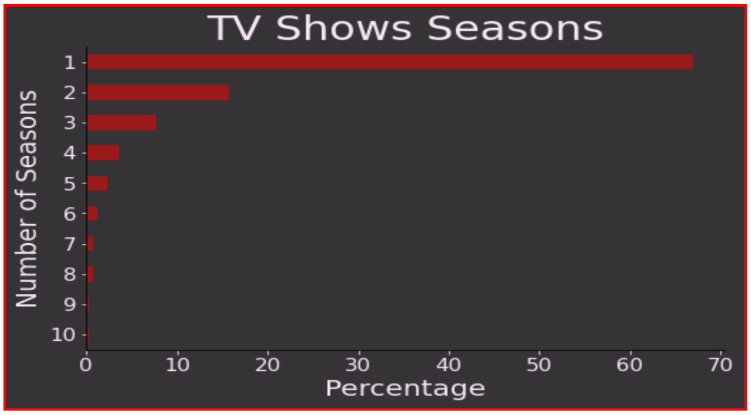
Running time of movies is also most important while engaging audience. So, analysis of running time of movies is very important. Following plot shows distribution of movies as per their running time.



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

6.9 Seasons of TV Shows

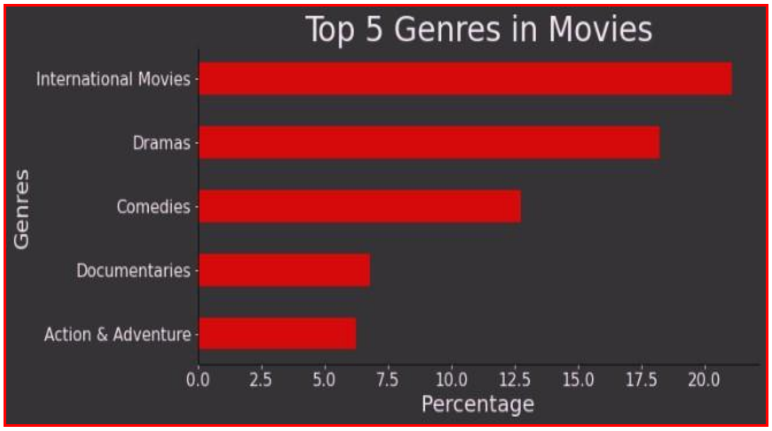
Most of TV shows consist of number of seasons. Following bar chart shows percentage of TV shows with number of seasons in it.



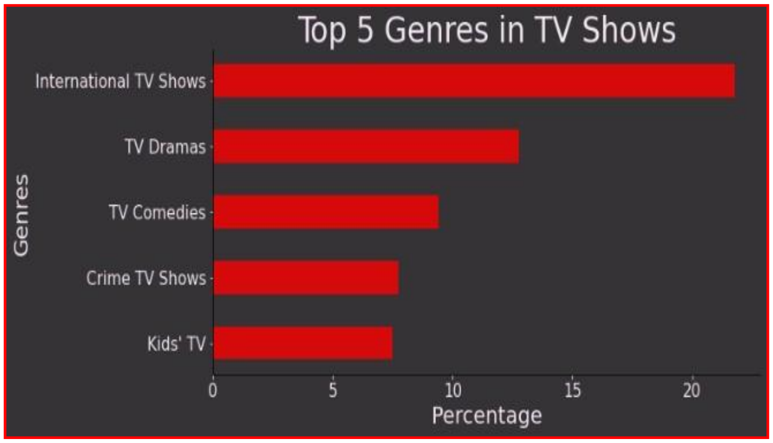
Almost 68% of TV shows consist of single season only.

6.10 Genres

Each and every movie or TV show can be categorized into a genre or set of genres. Following bar chart shows top five genres of movies.



Following bar chart shows top five genres of movies.

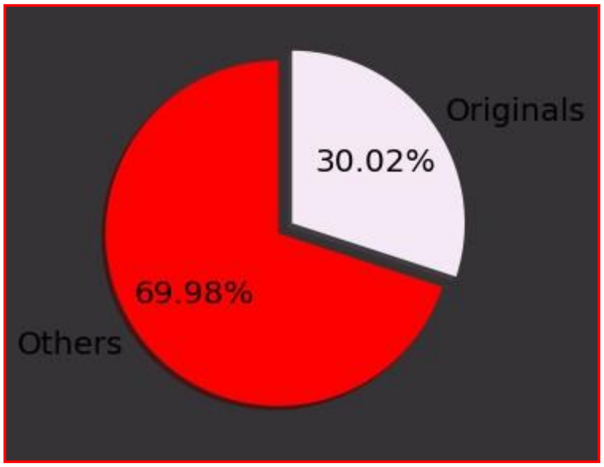


Top 3 genres are exactly same for movies and TV shows. Dramas genres hit all over the world.

6.11 Netflix Original

Some movies and TV shows were actually released in the past and were added later on Netflix. But some movies and TV shows were released on Netflix itself. Named those as Netflix Originals.

Following pie chart shows percentage of Netflix originals in movies



30% movies released on Netflix as Netflix originals. Remaining 70% movies added on Netflix were released earlier by different mode. May be after buying rights of old released movies and then adding all the movies on Netflix.

7. Data Preprocessing for Clustering:

7.1 Removing Punctuation

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

7.2 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.

7.3 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.

8. K-Means Clustering:

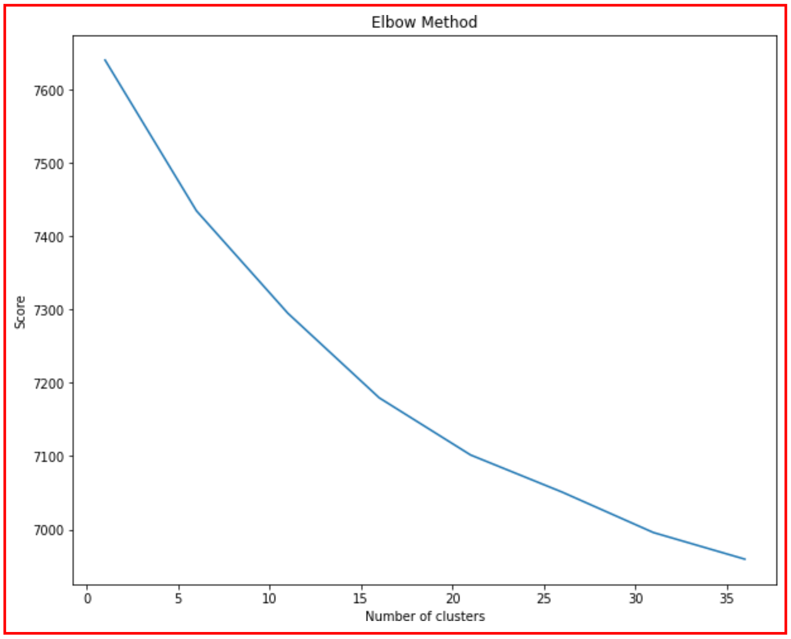
K-means algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups where each data point belongs to only one group. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid is at the minimum. The less variation we have within clusters, the more homogeneous the data points are within the same cluster.

8.1 Vectorization:

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

8.2 Elbow Curve:

The Elbow Method is one of the most popular methods to determine this optimal value of k. The elbow method 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.

We perform the hyperparameter tuning to choose the best value of k.

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, for the given data, we conclude that the optimal number of clusters for the data is around 25.

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

The value of the Silhouette score varies from -1 to 1. If the score is 1, the cluster is dense and well-separated than other clusters. A value near 0 represents overlapping clusters with samples very close to the decision boundary of the neighboring clusters. A negative score [-1, 0] indicates that the samples might have got assigned to the wrong clusters.

We calculated silhouette score for number of clusters from 20 to 30. Highest silhouette score comes to be at 25. So, optimal number of clusters chosen to be 25.

9. Recommender System:

Recommender systems are the systems that are designed to recommend things to the user based on many different factors. The recommender system deals with a large volume of information present by filtering the most important information based on the data provided by a user and other factors that take care of the user’s preference and interest. It finds out the match between user and item and imputes the similarities between users and items for recommendation.

9.1 Cosine Similarity:

Cosine similarity is a metric that measures the cosine of the angle between two vectors projected in a multi-dimensional space. As the cosine similarity measurement gets closer to 1, then the angle between the two vectors A and B becomes smaller. In this case, A and B are more similar to each other.

10. Conclusions:

- Movies uploaded on Netflix are more than twice the TV Shows uploaded.
- TV shows and movies are increasing continuously but in 2019 there is drop in number of movies.
- From October to January, maximum number of movies and TV shows were added.
- Maximum number of movies and TV shows were either on start of the month or mid of the month.
- United State tops in the list of maximum number of movies and TV

shows followed by India, UK and Japan.

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

References-

1. StackOverFlow
2. GeeksforGeeks