Capstone Project-4

Netflix Movies & TV Shows Clustering

(Unsupervised Machine Learning)

BY

Snehal D. Ramteke

Content

- Introduction
- Problem Statement
- Data Summary
- Cleaning Data
- Data Preprocessing
- Exploratory Data Analysis (EDA)
- Data Preprocessing For Clusterin
- K – Means Clustering
- Recommendor System
- Conclusions

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.
- 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, and rotten tomatoes can also provide many interesting findings.

Problem Statement

This project is based on Unsupervised learning and Natural Language processing. We had provided the large dataset in this project and need to perform the following problem statements.

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

Data Summary

- 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 : Genere
- description: The Summary description

Data Cleaning

1. Duplicate Values Treatment:

• Duplicate values do not contribute anything to the accuracy of results.

• Our dataset does not contain any duplicate values.

2. Null Values Treatment:

• Director feature has more than 30% of null values. So, dropping the feature director.

• Country feature has 6.51% of null values. Filling null values by mode of the feature.

• Cast feature has 9.22% of null values. Filling null values by 'missing'.

• Rating feature has 0.09% of null values. Filling null values by mode of the feature.

• Date_added feature has 0.12% of null values. Dropping rows corresponding to null values.

Data Pre-processing

1. Date Type Change:

- Features in their appropriate data type provide better understanding and workability on that data.

• Date_added feature has an object datatype. Converting to DateTime.

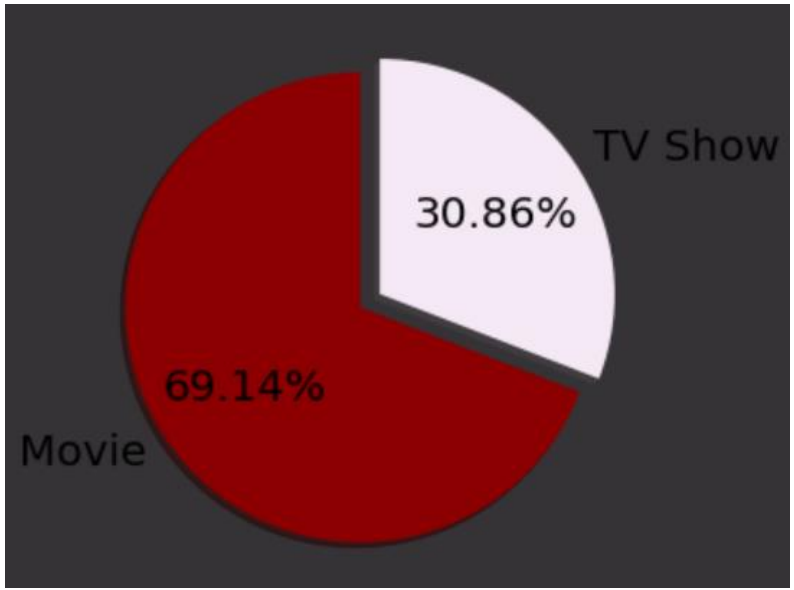
• Duration is in a combination of integer values and text. Removing text part so as to get integer datatype.

2. New Features:

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

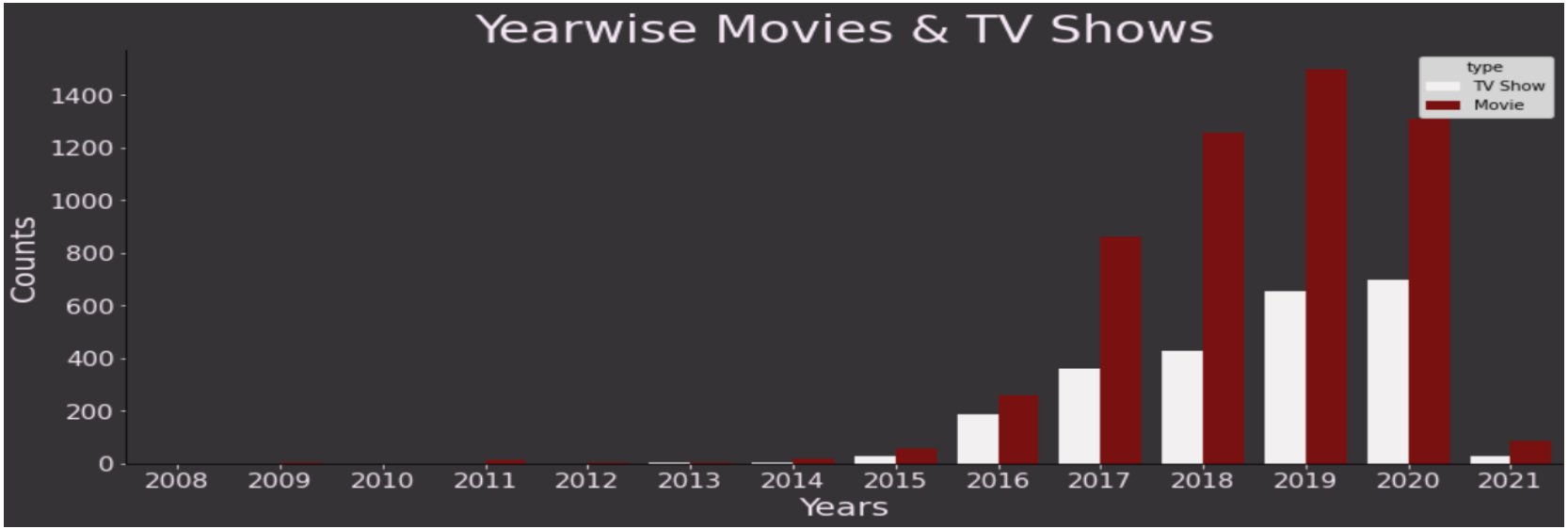
Exploratory Data Analysis (EDA)

Movies Vs TV Show



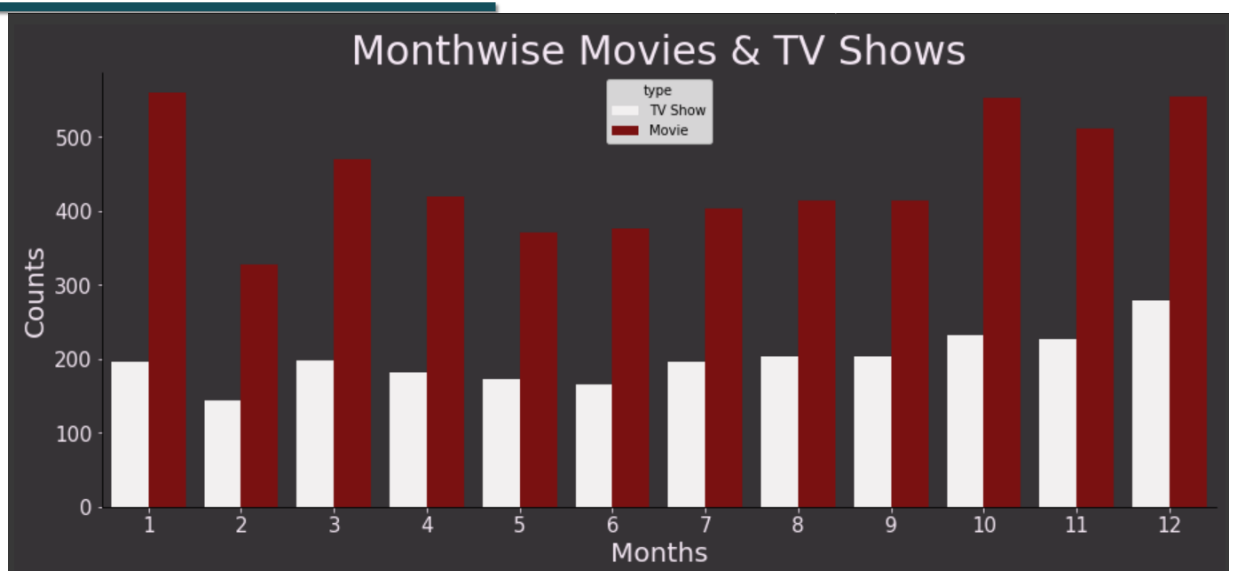
- Movies uploaded on Netflix are more than twice the TV Shows uploaded.
- 30% of movies are released on Netflix.
- 70% of movies added on Netflix were released earlier in a different mode.

On Year Basis



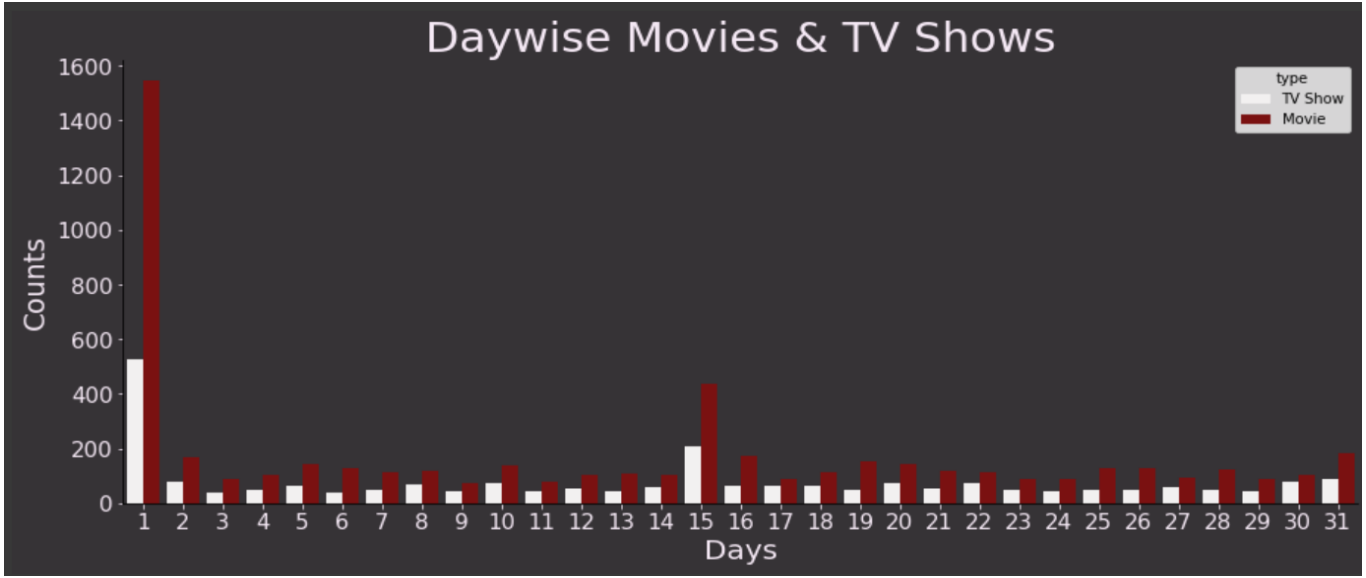
- TV shows are increasing continuously
- The number of releases significantly increased after 2015 and dropped in 2021 because of Covid 19.

On Month Basis



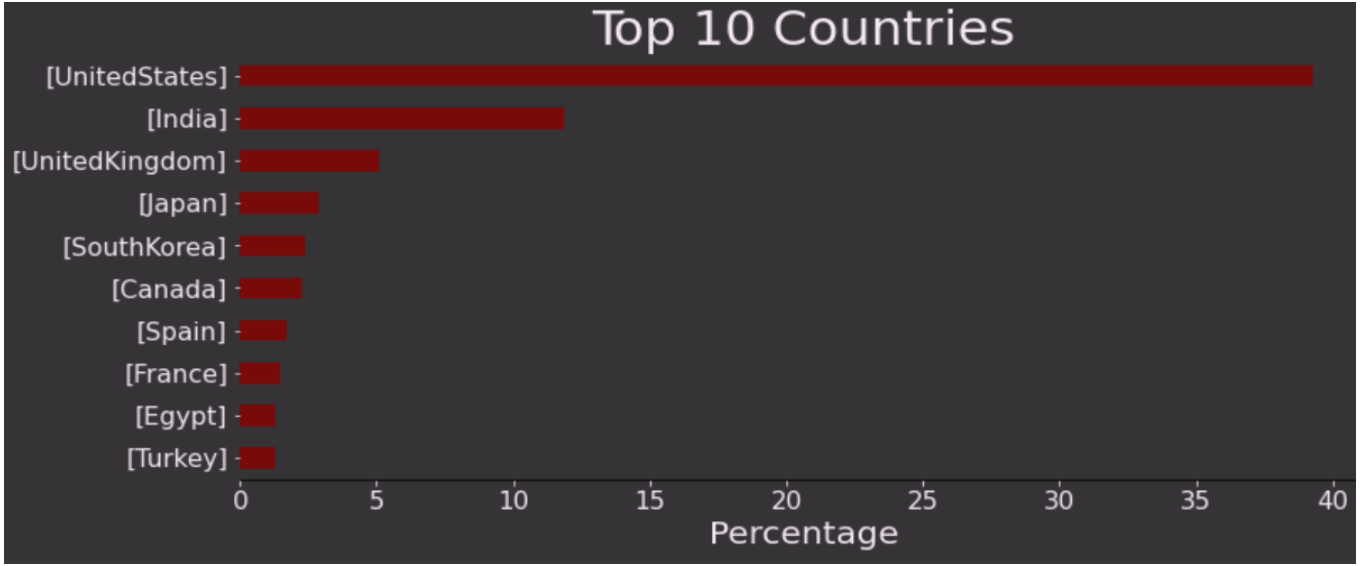
- • From October to January, the maximum number of movies and TV shows were added.
- • Possible reason for that is, during this period of time events such as Christmas, New Year and several holidays take place.

On Day Basis



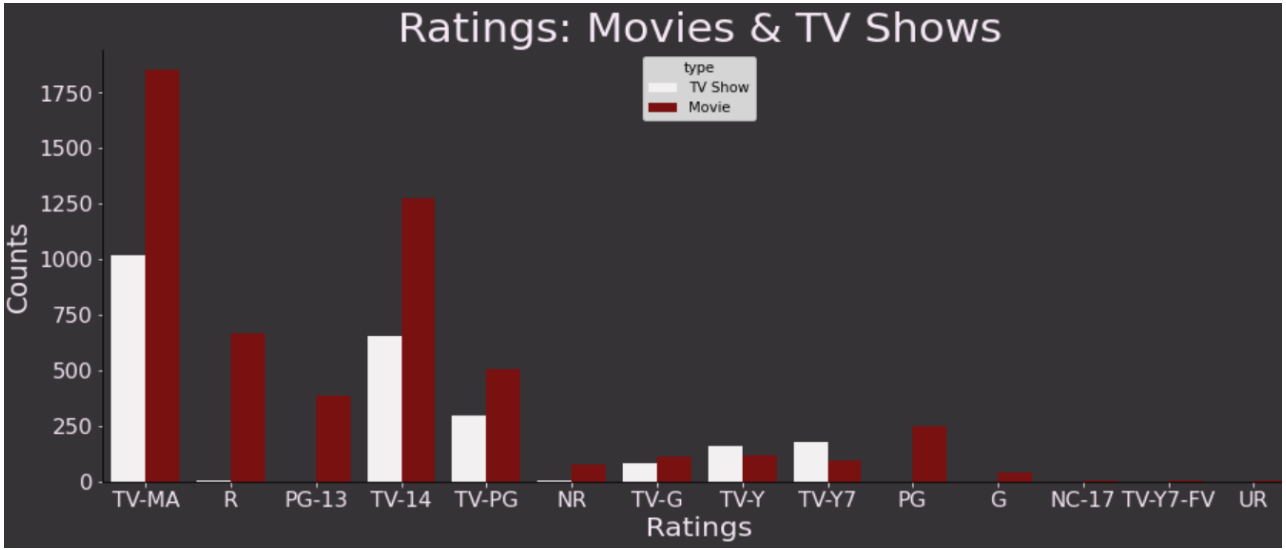
- Maximum number of movies and TV shows added at the start of the month followed by mid of month.

Worldwide Presence



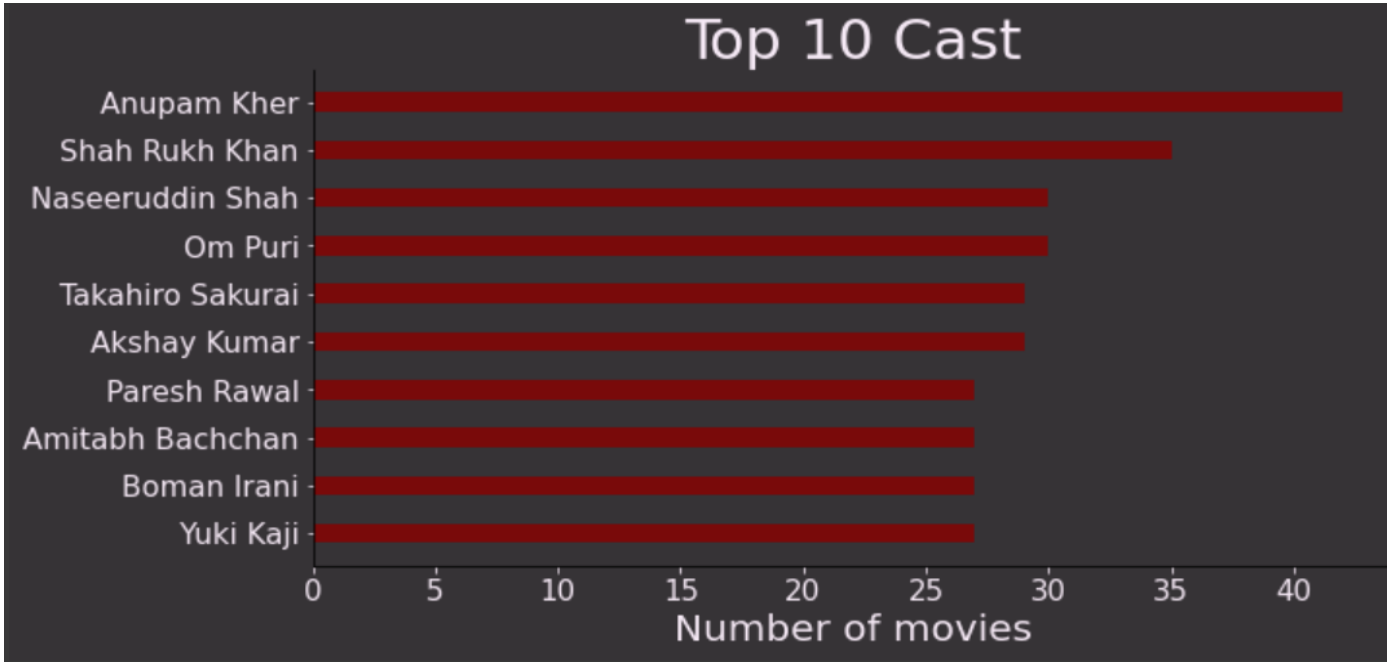
- United States tops in the list of the maximum number of movies and TV shows, followed by India, the UK, and Japan.
- 10^{th} no. of the country is Turkey.

Ratings



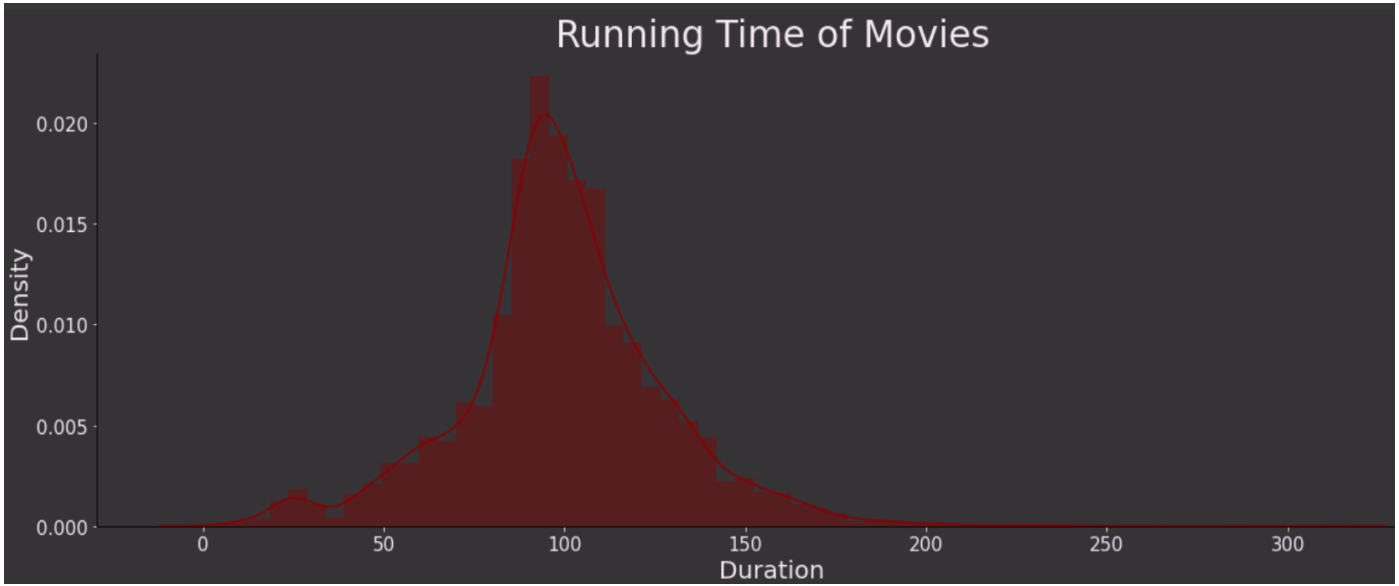
- Maximum of the movies as well as TV shows are for matures only.
- we can observe that the most-rated movie on Netflix is TV-MA and the least-rated movie on Netflix is UR.

Cast



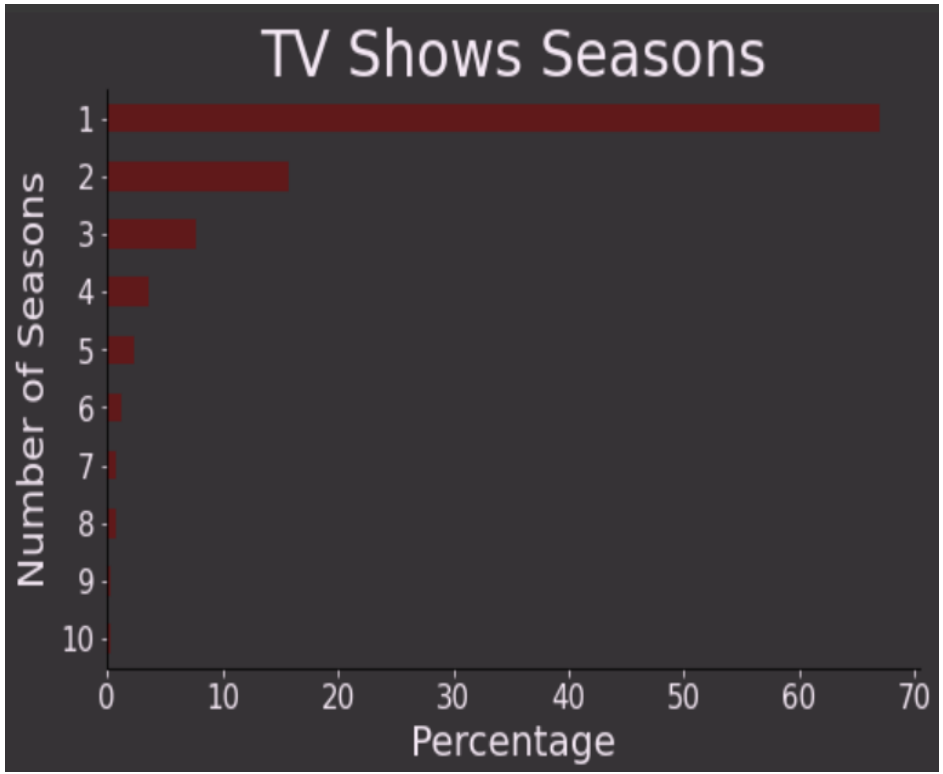
- Anupam Kher top of the list of casts having a maximum number of movies and TV shows.

Running Time OF movies



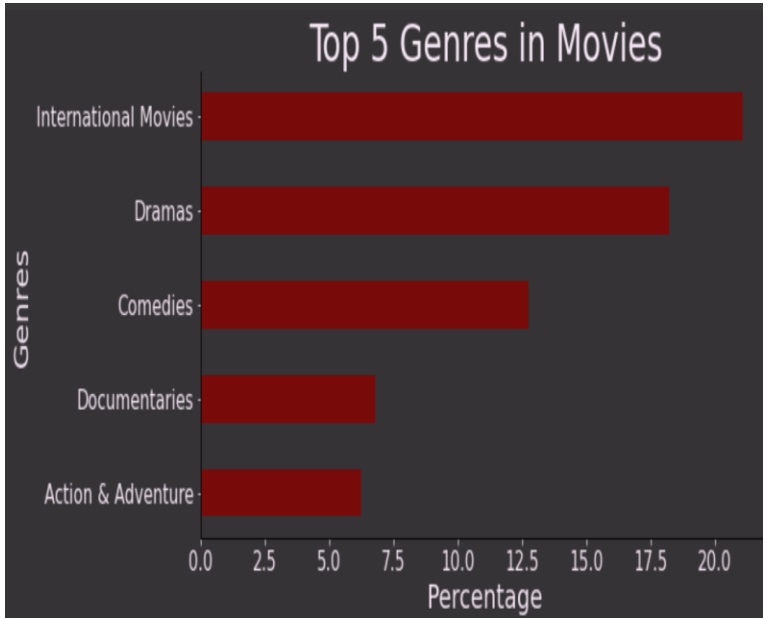
- most of the movies have a duration of between 50 to 150.

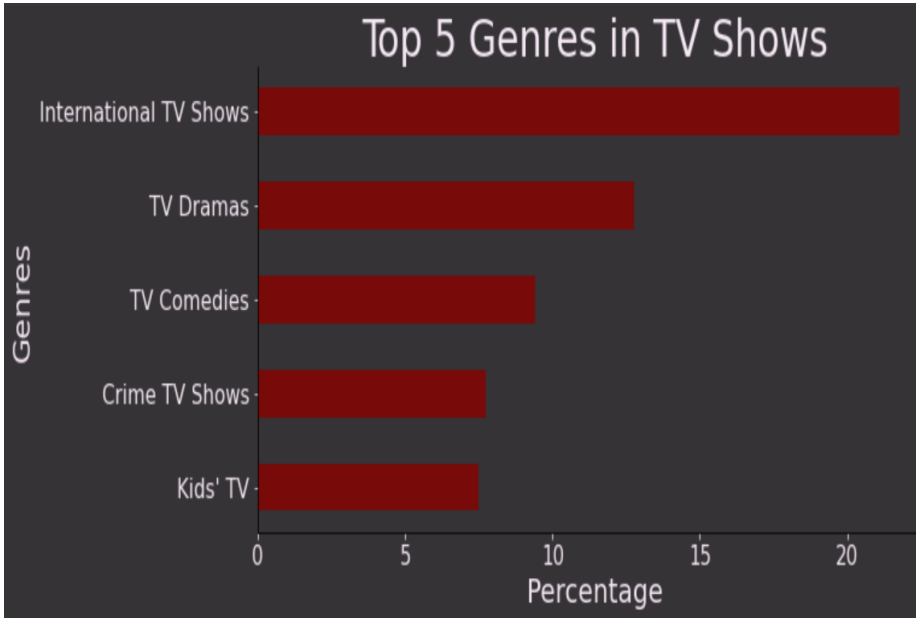
Seasons of TV Shows



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

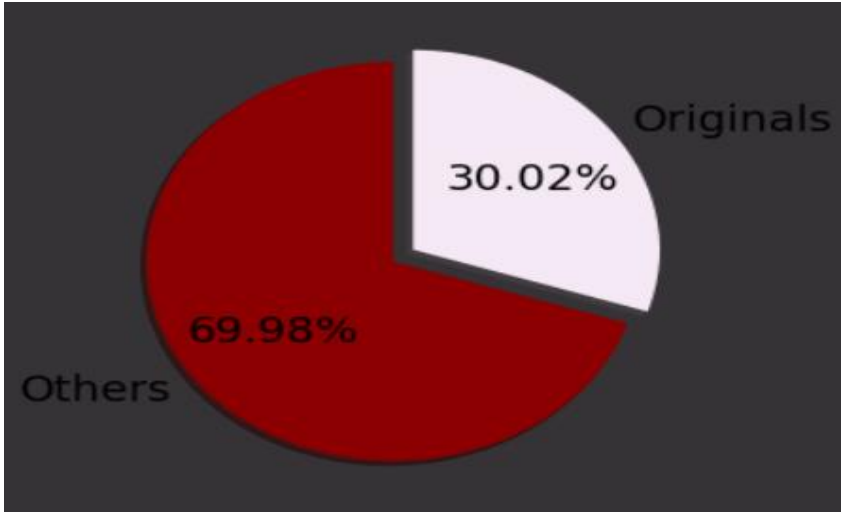
Genre

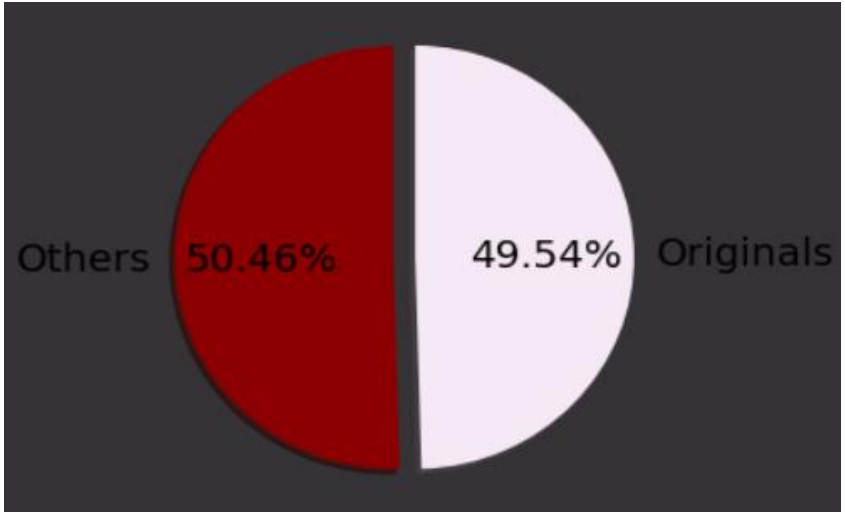




- Top 3 genres are exactly the same for movies and TV shows.
- Dramas genres hit all over the world.
- International Movies are the top most genre on Netflix which is filled with Dramas and comedies.

Netflix Original





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

Data Pre-processing for Clustering

1. Removing Punctuation:

- Punctuations do 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.

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.

K – Means Clustering

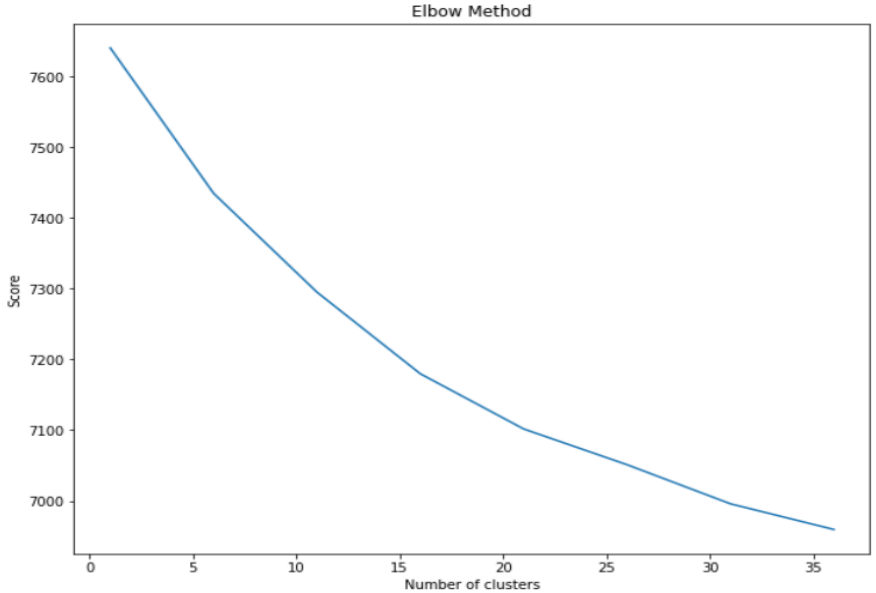
AI

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.

Vectorization:

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

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.

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 ranges from −1 to +1, where a high value indicates that the object is well-matched to its own cluster and poorly matched to neighboring clusters.
- Maximum value of Silhouette score is for k equals to 24.

Recommender System

AI

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

Cosine Similarity:

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

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 a drop in the number of movies.
3. From October to January, a maximum number of movies and TV shows were added.
4. The maximum number of movies and TV shows was either at the start of the month or mid of month.
5. The United States tops in the list of the maximum number of movies and TV shows followed by India, the UK, and Japan.
6. Maximum of the movies as well as TV shows are for matures only.

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


THANK YOU