Capstone Project
Netflix Movies & TV Shows Clustering

Akash Salmuthe

Netflix and Chill !!!

NETFLIX

Problem Statement

1. Exploratory Data Analysis

2. Understanding what type of content is available in different countries

3. Is Netflix has increasingly focusing on TV rather than movies in recent years.

4. Clustering similar content by matching text-based features

Introduction

- Netflix is a streaming service with wide variety of content. The idea of this project is to analyze and perform clustering to determine various patterns related to the content available in Netflix.

- The data is gathered from a third party engine.

- Based on the attributes related to the TV shows or movies, we will be implementing different clustering algorithms which comes under unsupervised Machine learning category.

Data Summary

o This dataset consists of TV shows and movies available on Netflix as of 2019.

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

o 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.

Data Description

- **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** : Genre
- **description**: The Summary description

Data Cleaning

1. **Duplicate Values Treatment:**

    • Duplicate values dose not contribute any thing to accuracy of results.

    • Our dataset dose not contains any duplicate values.

Data Cleaning

2. **Null Value 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.

Data Pre-Processing

1. Data Type Change:
- Features in their appropriate data type provides better understanding and work ability on that data.
- ***Date_added*** feature have object data type. Converting to ***datetime***.
- Duration is in combination of integer values and text. Removing text part so as to get integer data type.
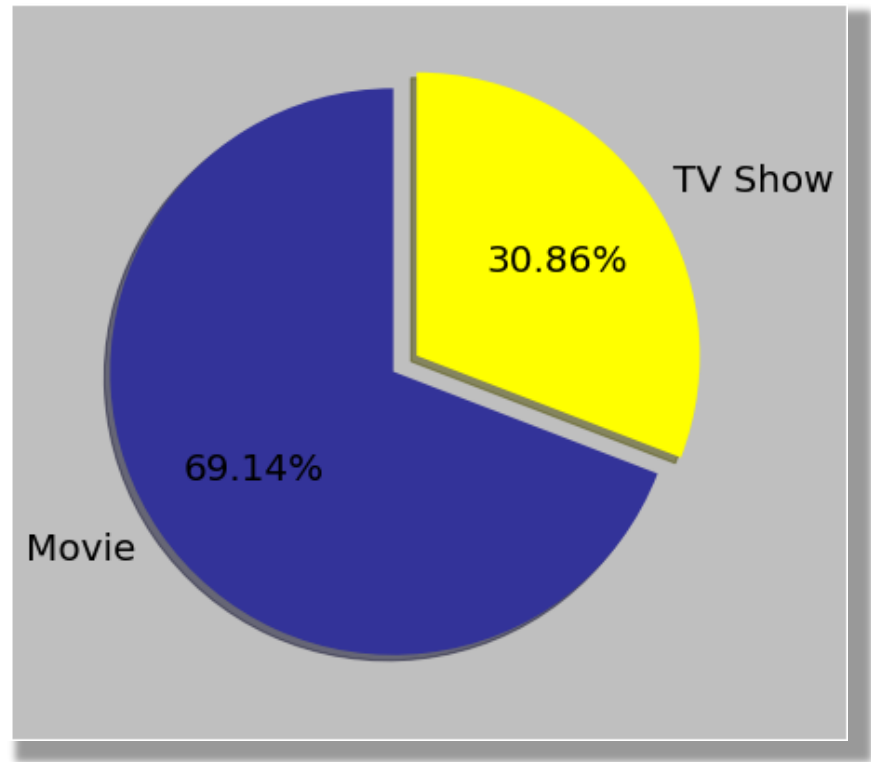
2. New Features:
- From the feature ***date_added***; extracted ***year***, ***month*** and ***day*** to form new columns by name of year, month and day respectively.
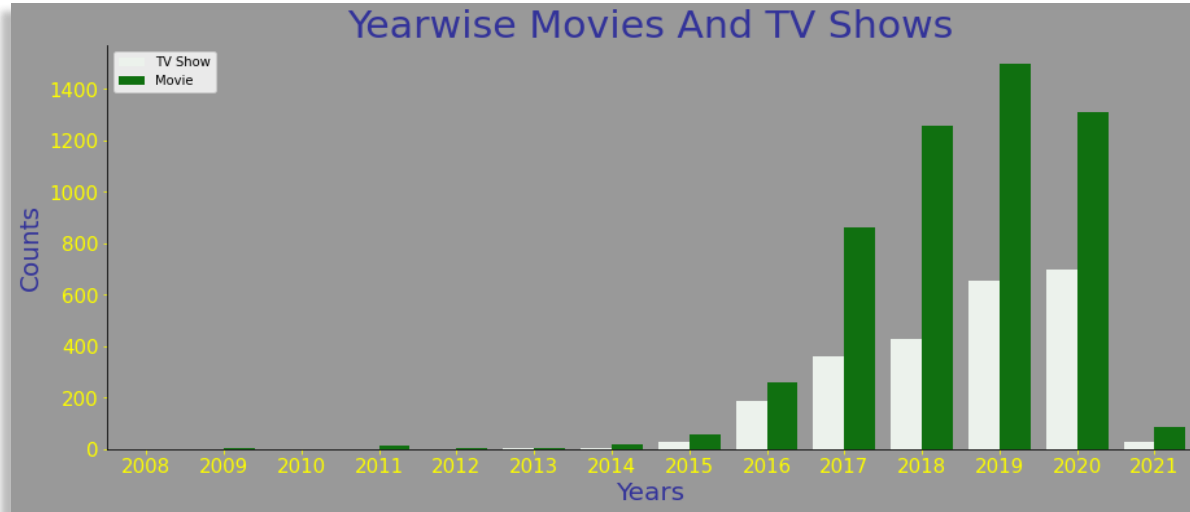
Exploratory Data Analysis

Movies vs. TV Shows

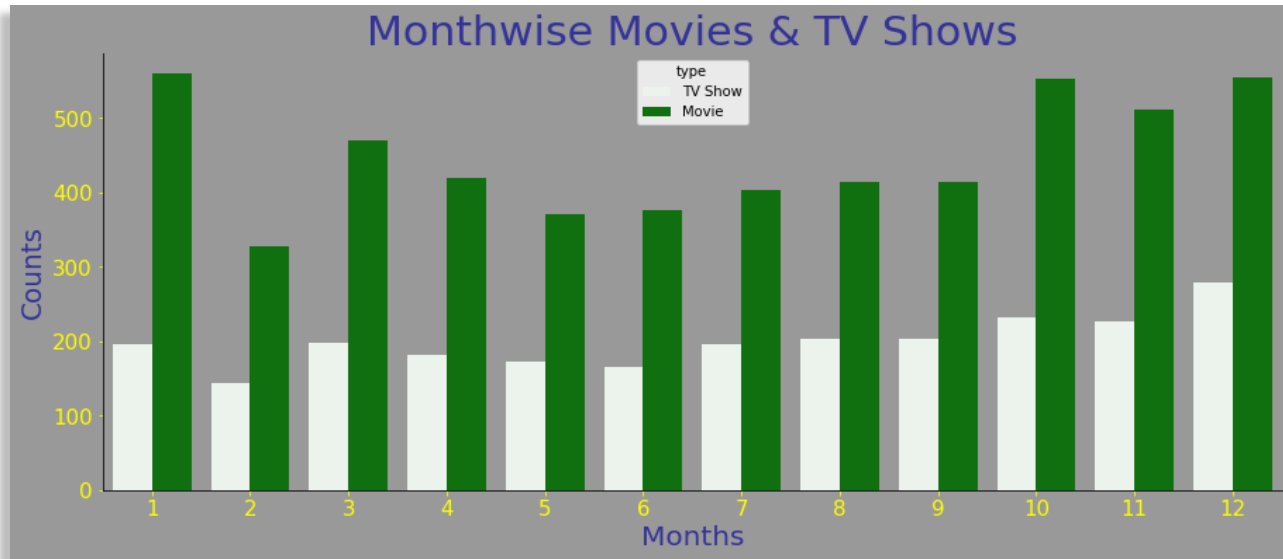- Movies uploaded on Netflix are more than twice the TV Shows uploaded.

On Year Basis

Yearwise Movies And TV Shows

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

On Month Basis


Monthwise Movies & 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 Halloween Christmas, New Year and several holidays takes place.
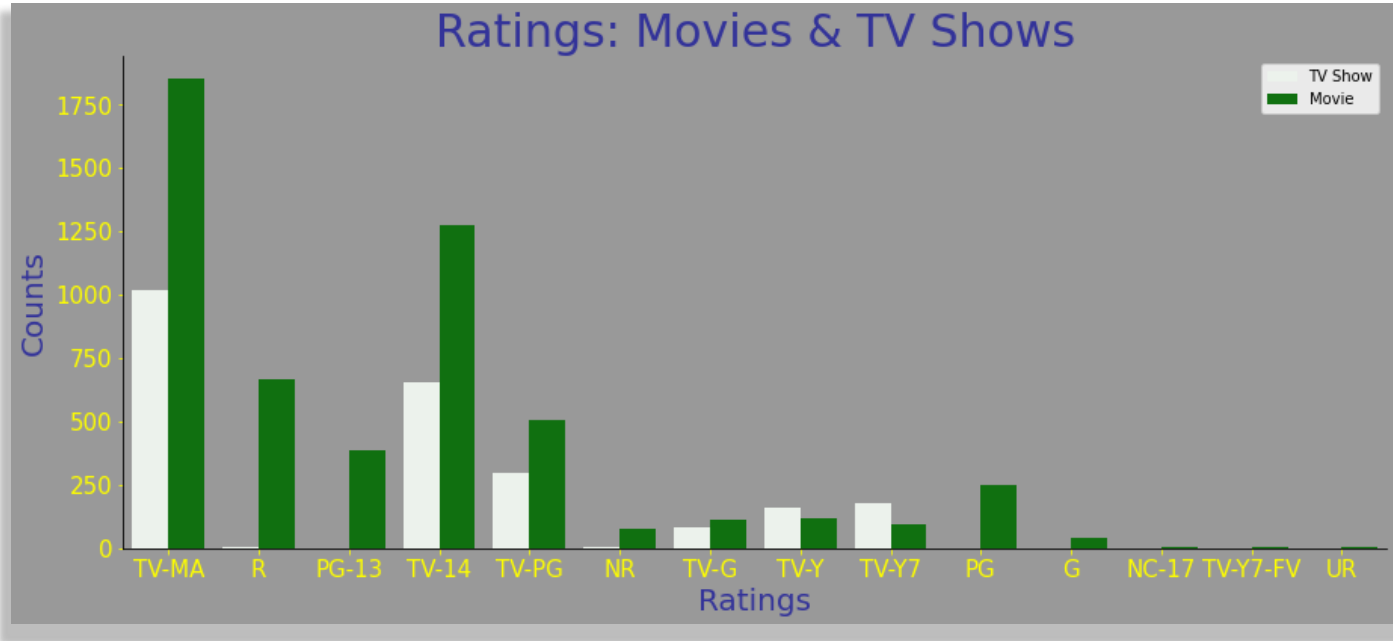
On Day Basis

Daywise Movies & TV Shows
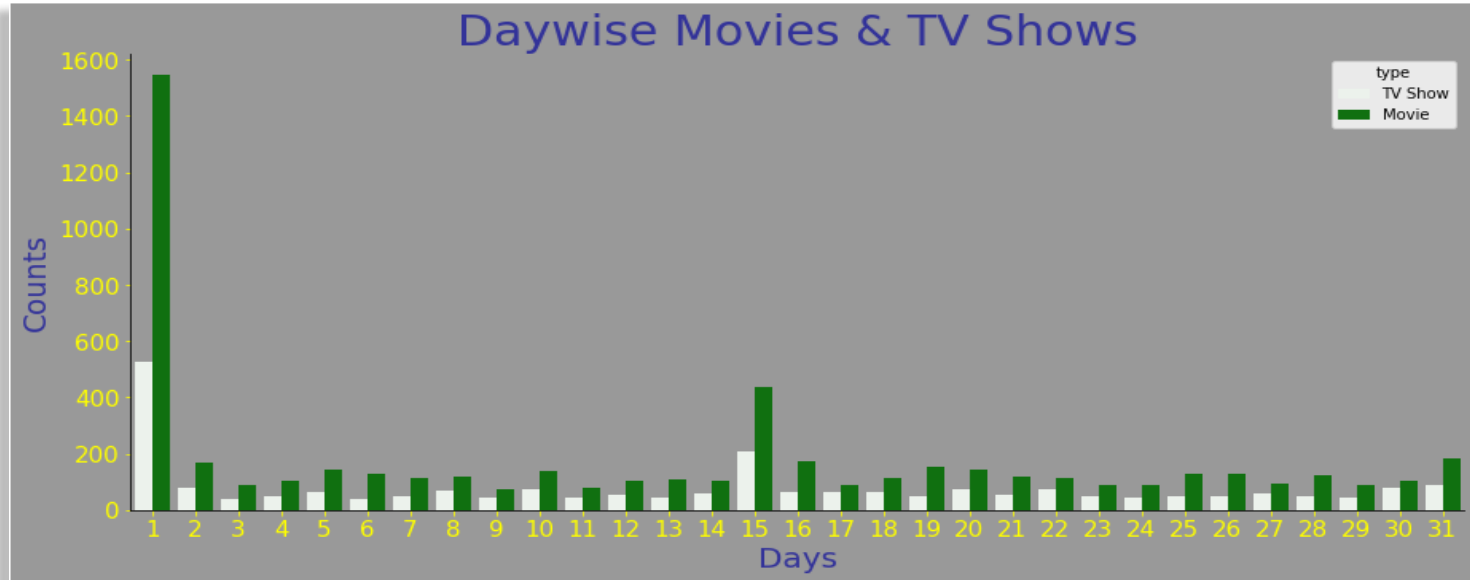
- Maximum number of movies and TV shows added on start of the month followed by mid of the month.
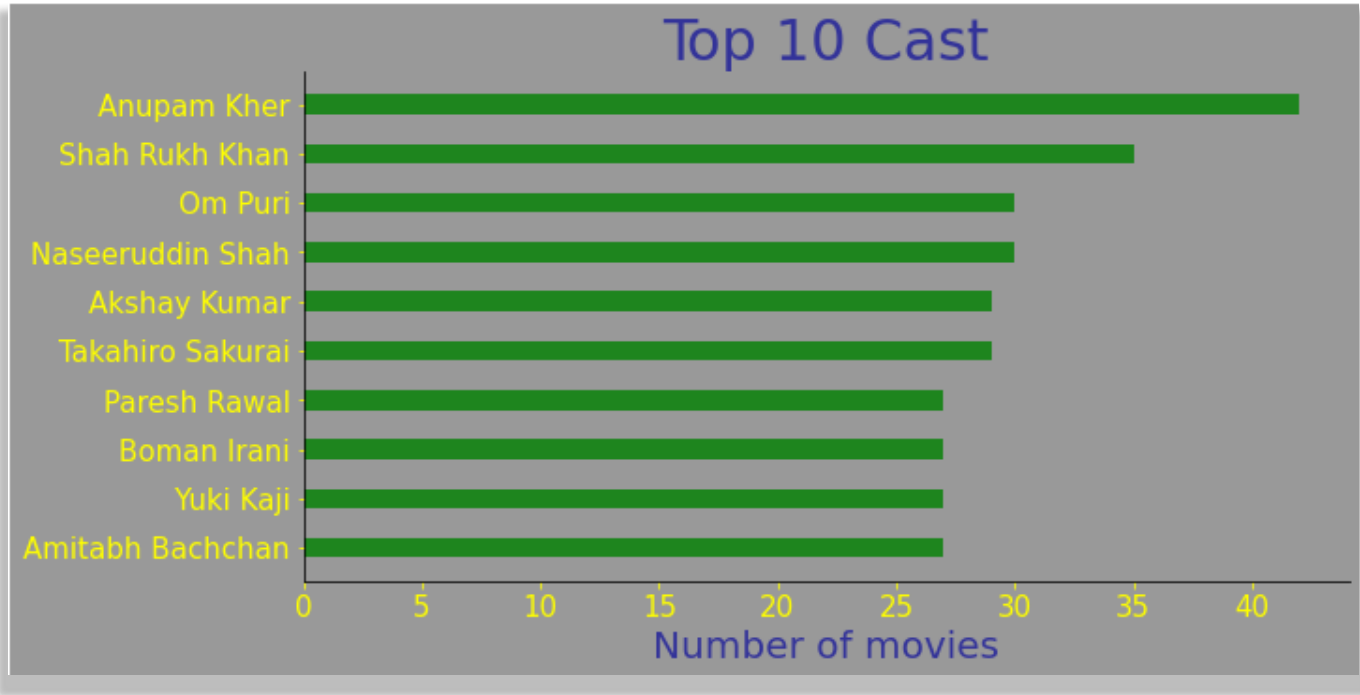
World Wide Presence
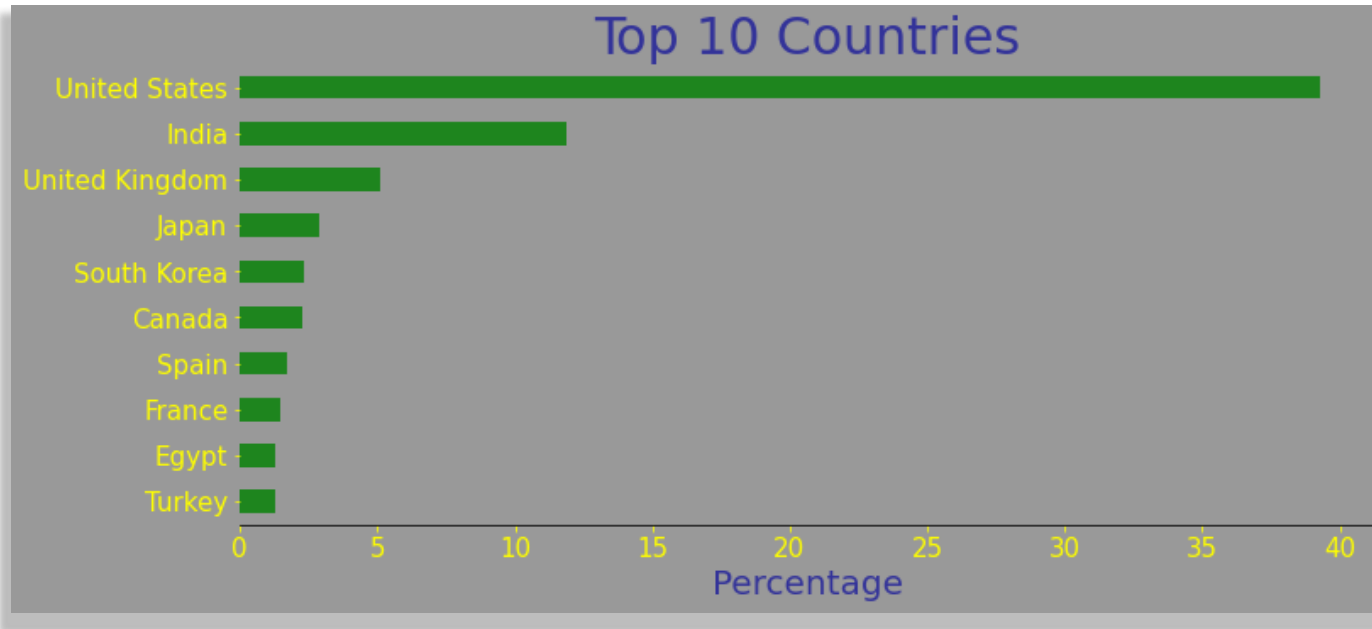
Top 10 Countries

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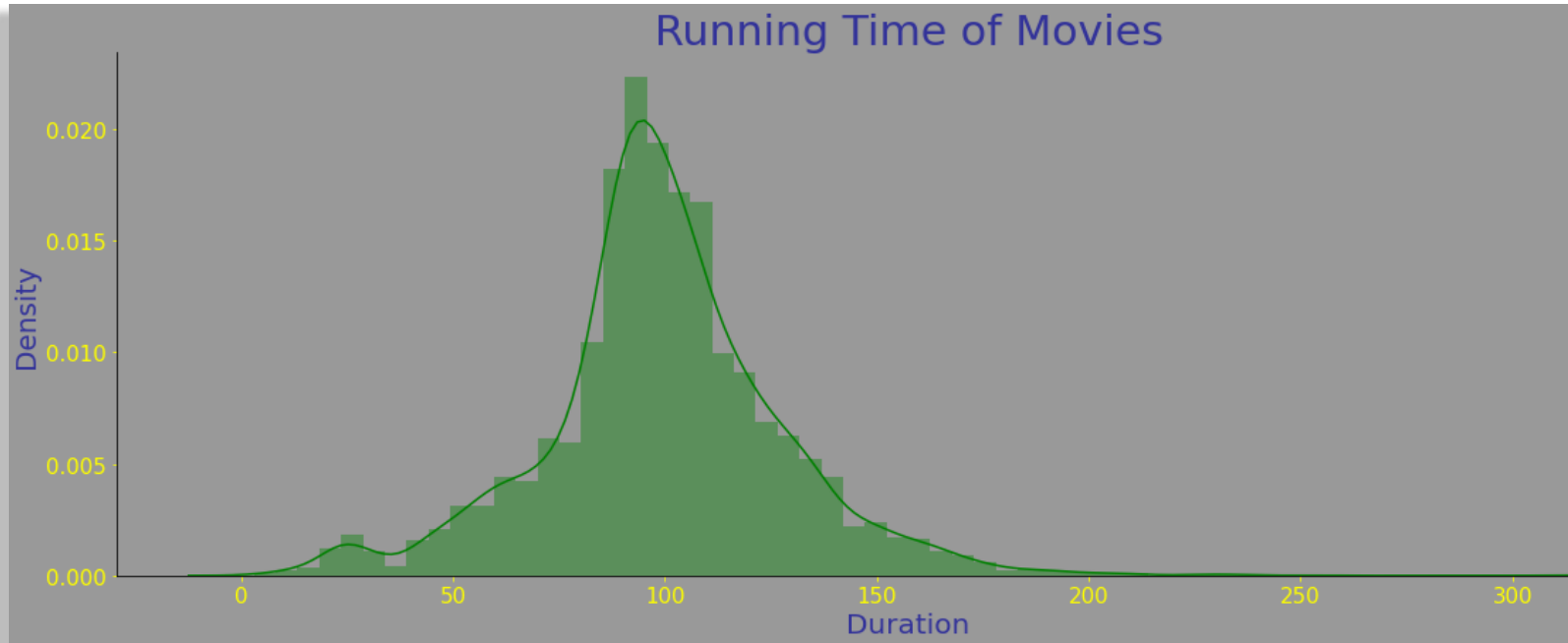
Ratings

Ratings: Movies & TV Shows
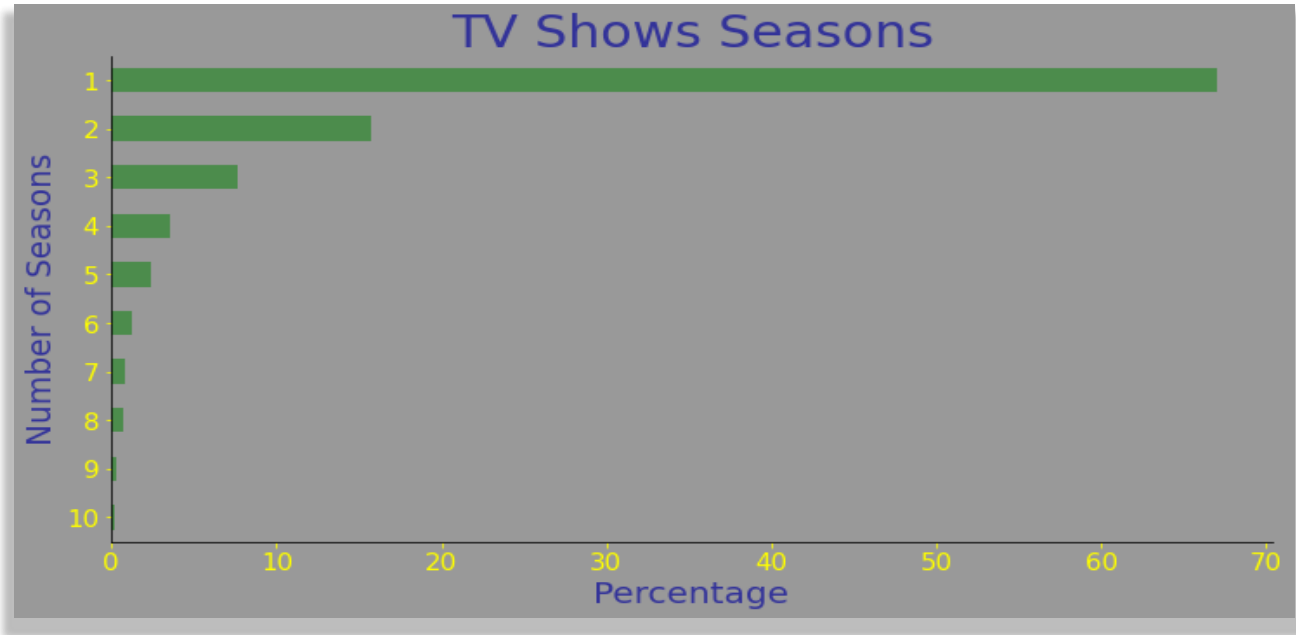
- Maximum of the movies as well as TV shows are for matures only.

Cast

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

Running Time of Movies

Running Time of Movies

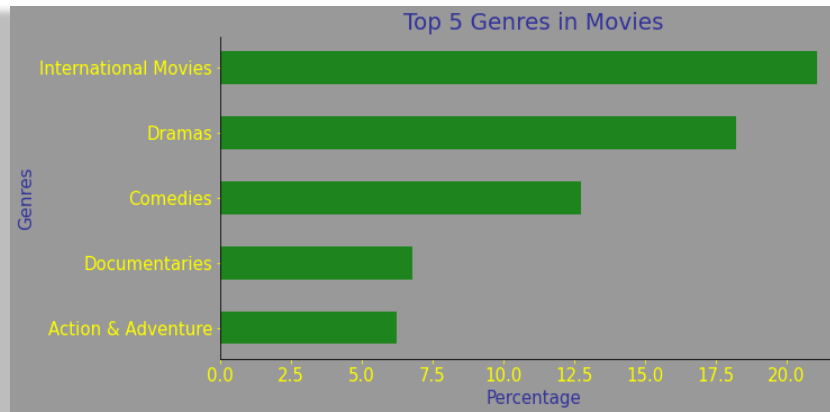- Majority of movies have running time in between 50 to 150 min.

Season of TV Shows

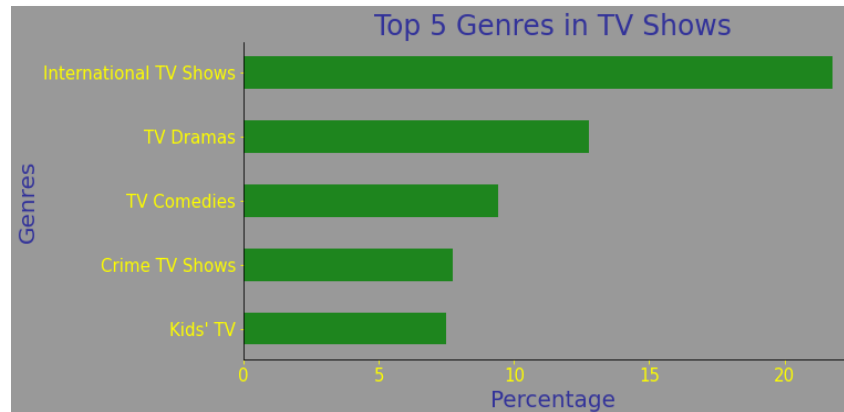- Almost 68% of TV shows consist of single season only.

Top Genres

Movies

Top 5 Genres in Movies

TV Shows

Top 5 Genres in TV Shows
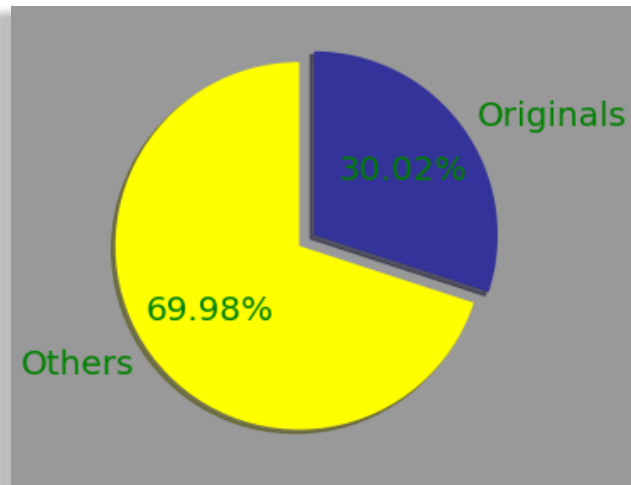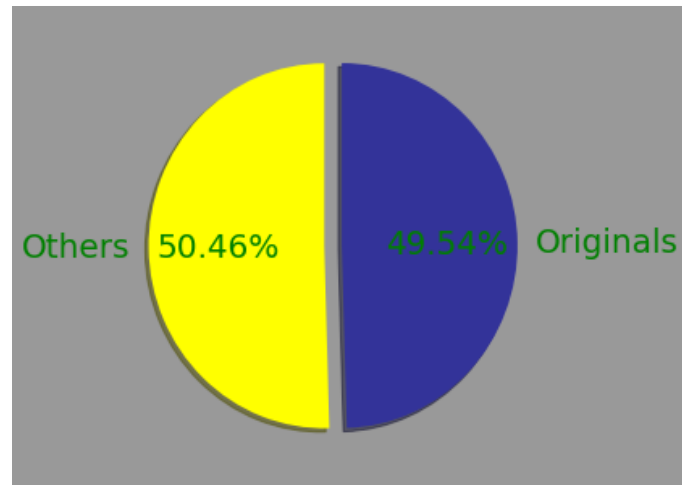
- Top 3 genres are exactly same for movies and TV shows.
- Dramas genres hit all over the world.

Netflix Original

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

Data Pre-processing for Clustering

1. **Removing Punctuation:**
   - • Punctuations does not carry any meaning in clustering.
   - • So, removing punctuations helps to get rid of unhelpful parts of the data, or noise

2. **Removing Stopwords:**
   - • Stopwords are basically a set of commonly used words in any language, not just in 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

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

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.

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.

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.

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.

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 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!