Capstone Project-4

Netflix Movies & TV Shows Clustering

Presented By-
Yashwati Patel

Content:-

- Introduction
- Problem Statement
- Data Summary
- Exploratory Data Analysis
- Hypothesis Testing
- Finding Number of Clusters
- Model Performance
- Conclusion

Introduction:-

Netflix is a streaming service that offers a wide variety of award-winning TV shows, movies, anime, documentaries and more – on thousands of internet-connected devices. You can watch as much as you want, whenever you want, without a single ad – all for one low monthly price. There's always something new to discover, and new TV shows and movies are added every week!

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.

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

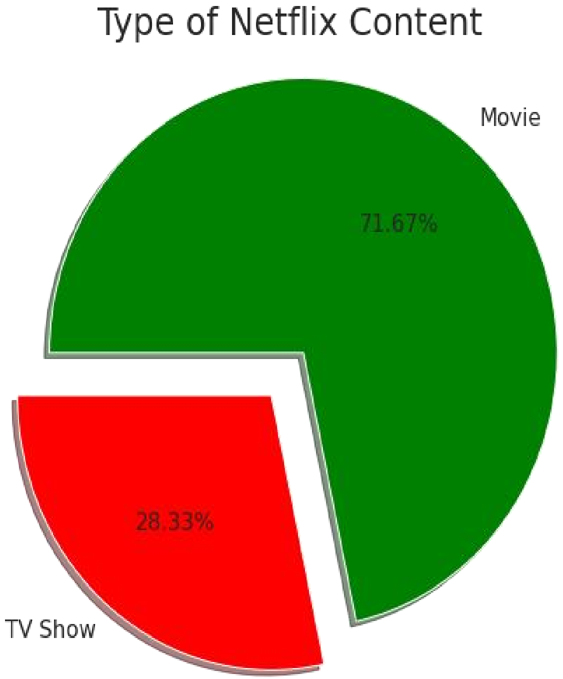
Data Summary:-

- **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 Releaseyear 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

Exploratory Data Analysis

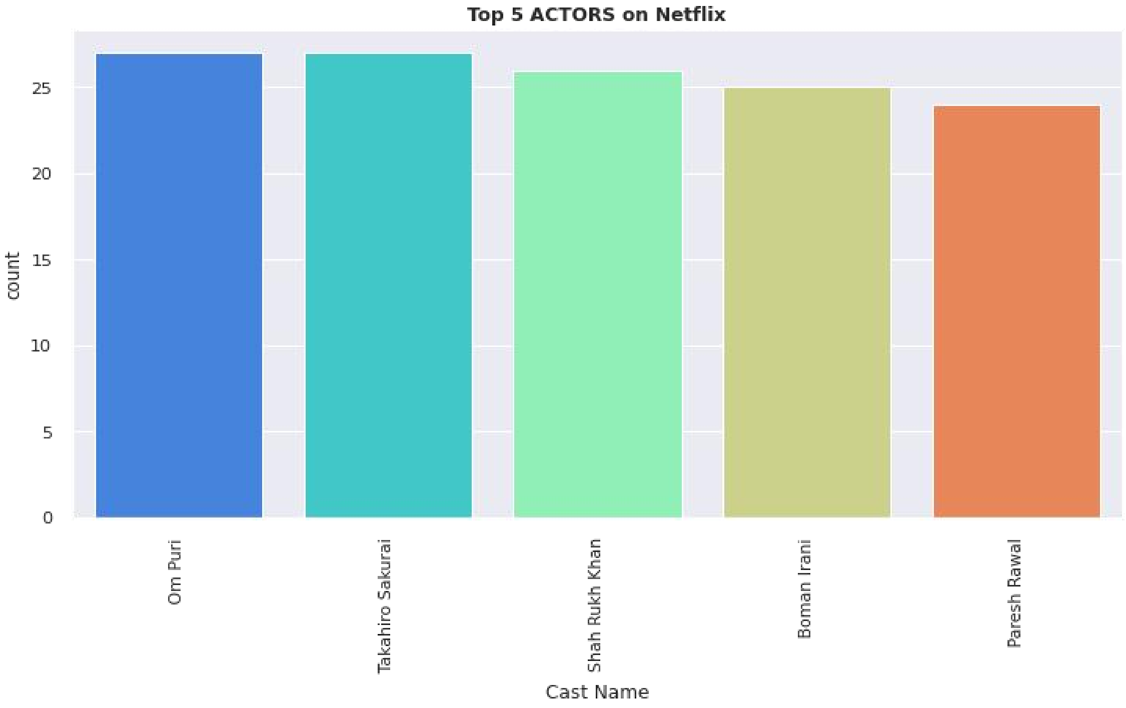
Check type of content:-

From this plot we can see that number of movies are more than TV shows. There are about 70% movies and 30% TV shows on Netflix.



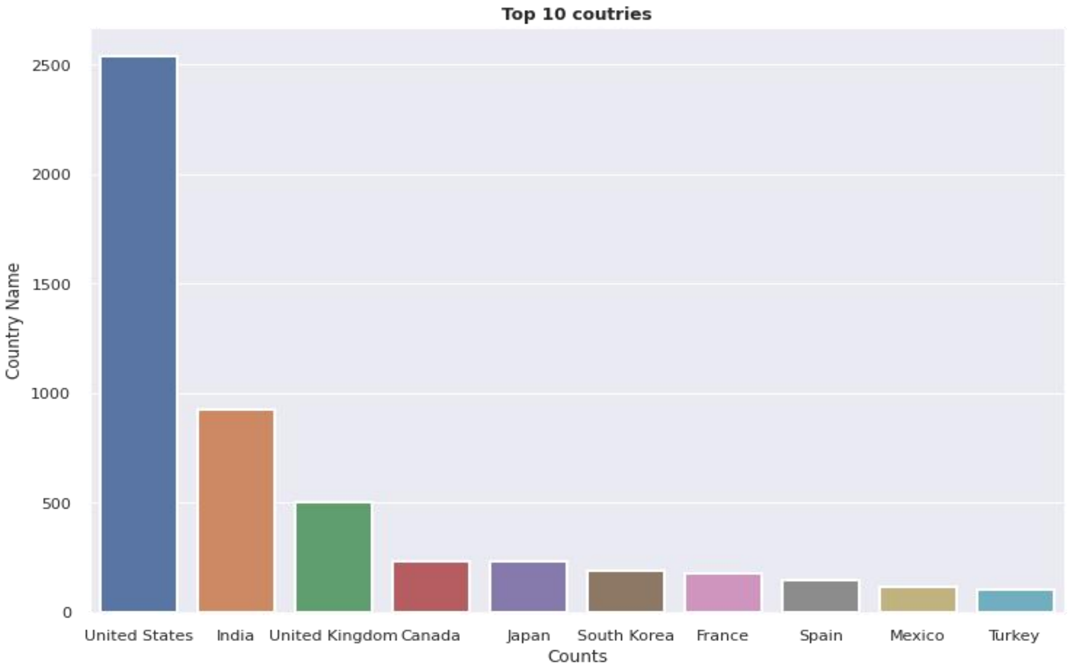
Check top five actors on Netflix:-

From this analysis we can see that in netflix top five actors are Om puri, Takahiro sakurai, Shahrukh khan, Boman irani and Paresh rawal.



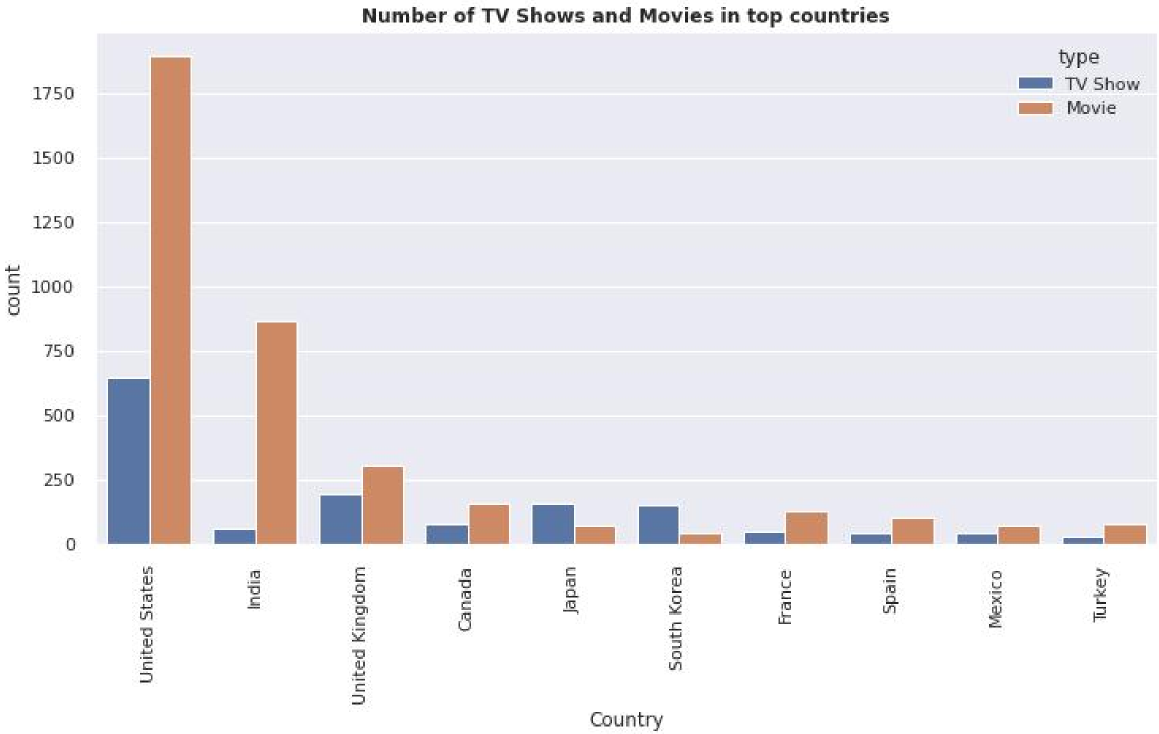
Check top Ten countries Based on content:-

We can see that The United States has the highest number of content on Netflix by a huge margin followed by India.



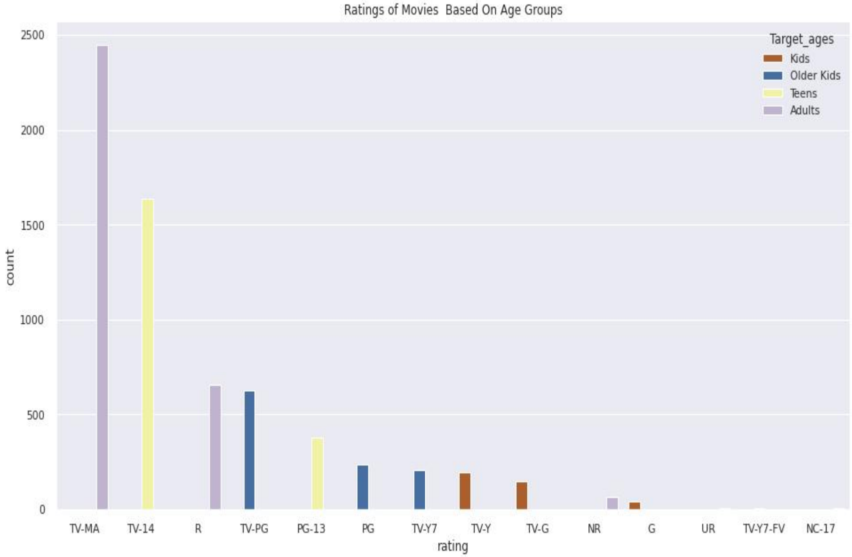
Movies and TV Shows in top countries:-

Here we can see that The United States has the highest number of movies and in india also followed highest number of movies as compare to TV shows.

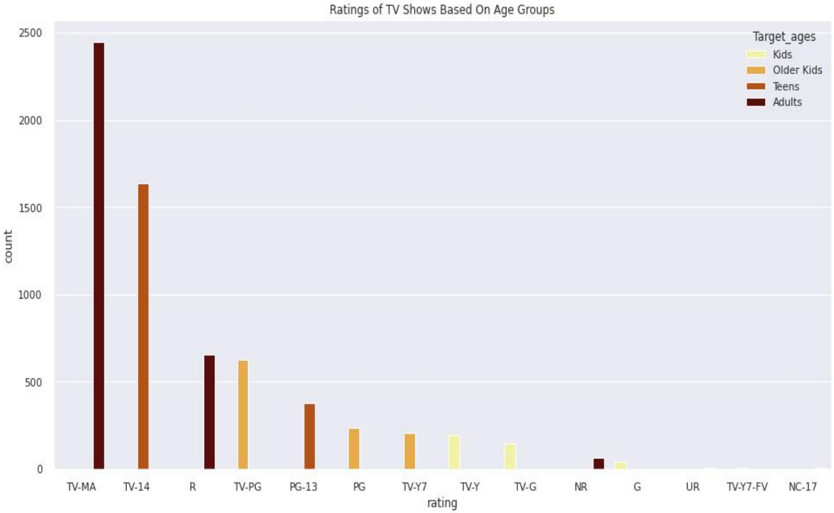


Check content Ratings based on Age group:-

MOVIES



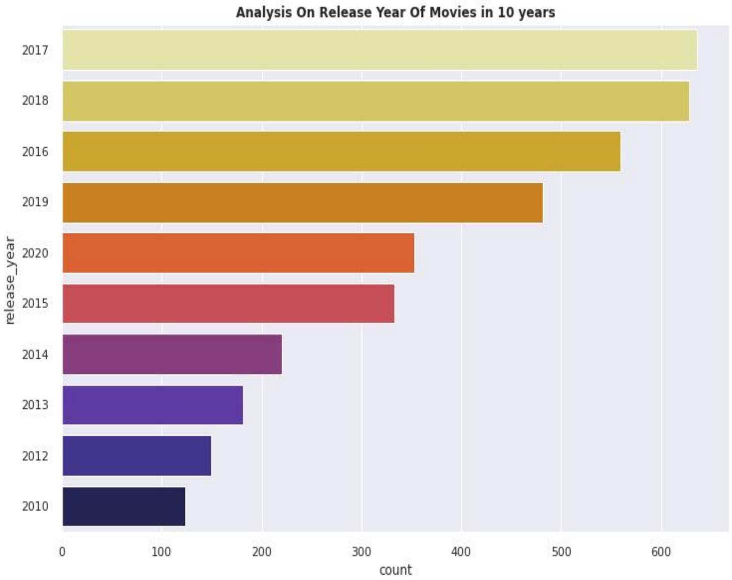
TV SHOWS



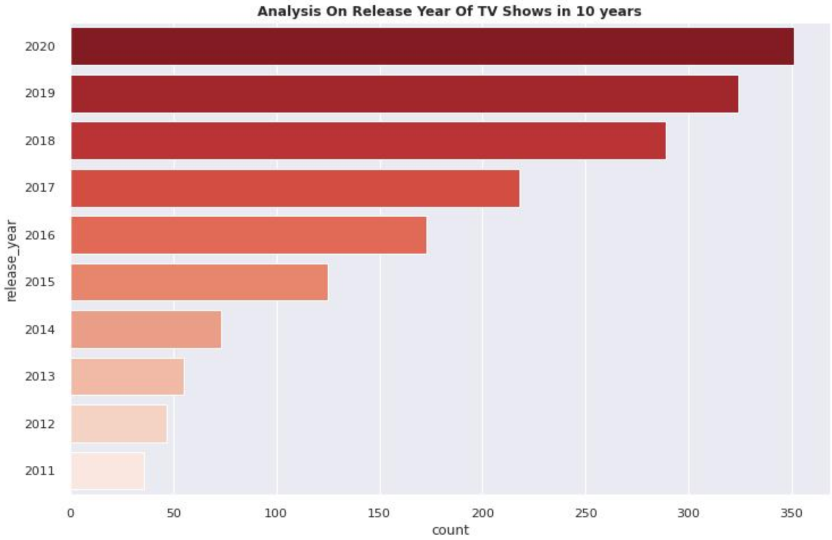
from above plot we can see that most number of movies and TV shows followed by Adults as compare to other Age groups.

Release Analysis in last Ten years:-

MOVIES



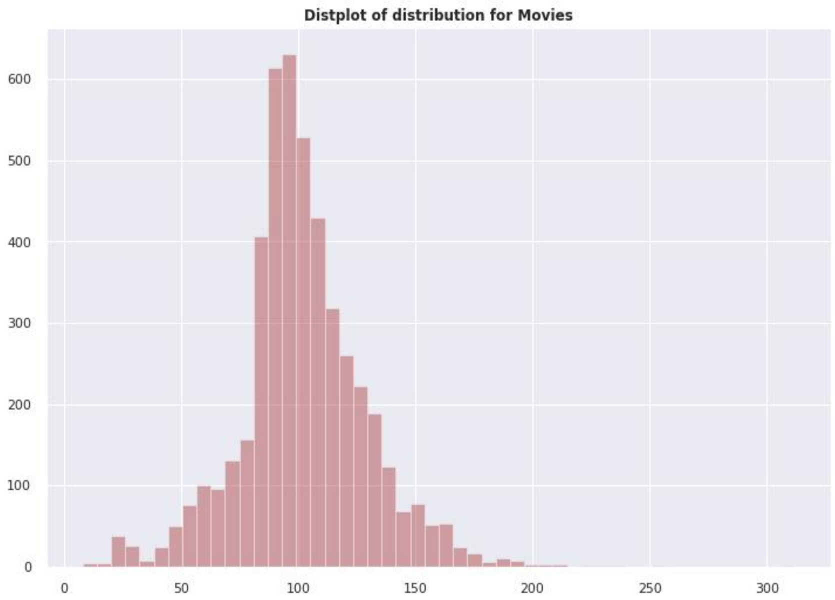
TV SHOWS



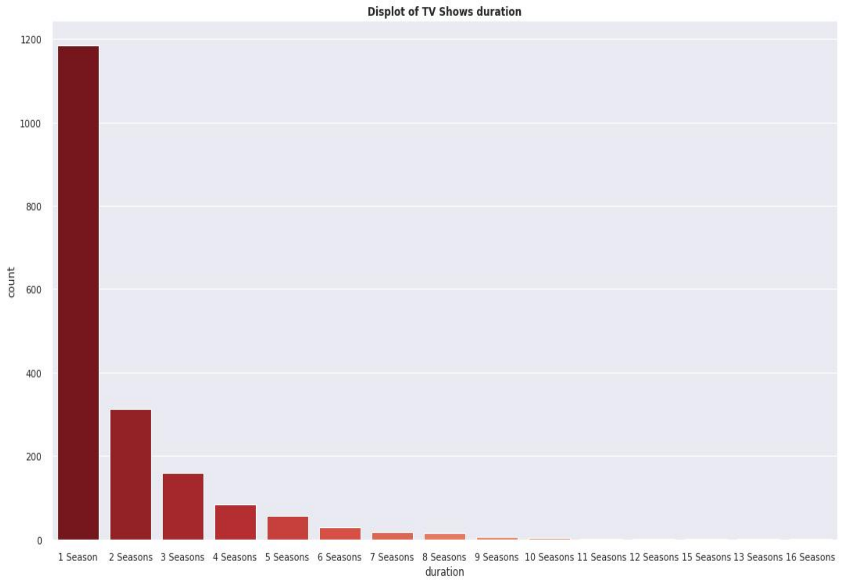
Here we can see that most number of movies released in 2017 and most number of TV shows released in 2020.

Content Duration:-

MOVIES

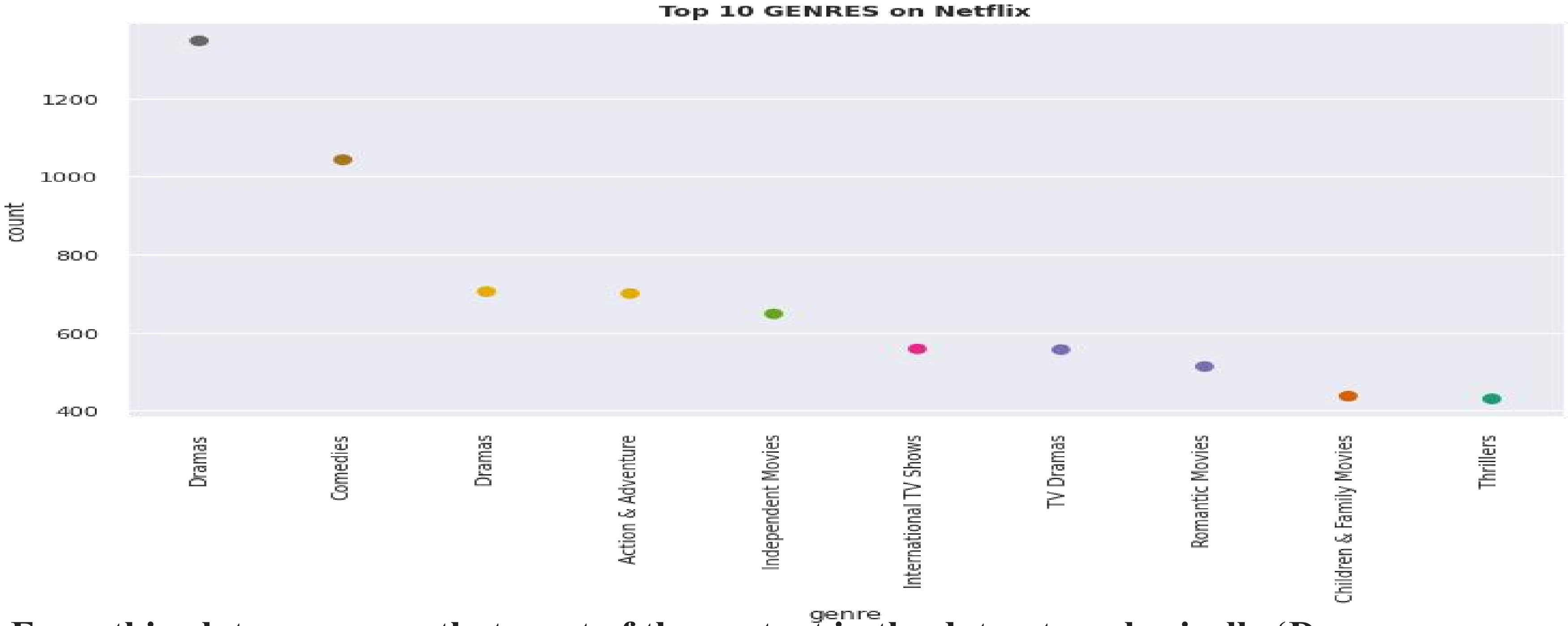


TV SHOWS



From this plot we can see majority of movies have running time of between 50 to 150 min.

Check top TEN genres on Netflix:-

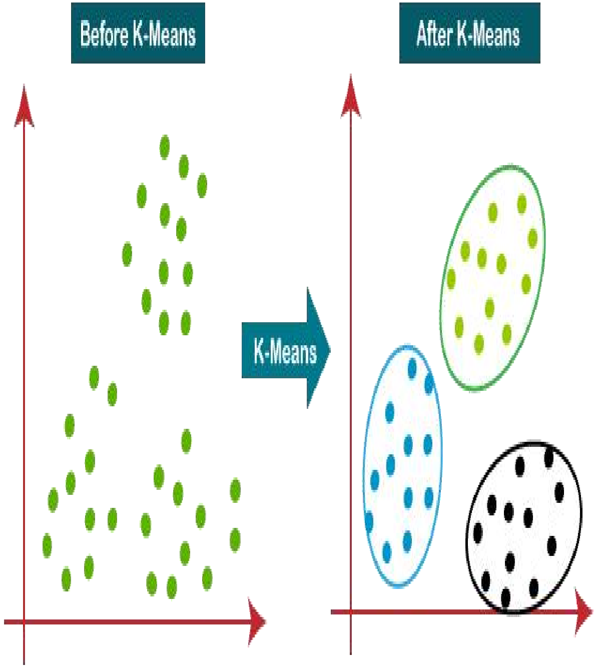


From this plot we can say that most of the content in the dataset are basically ‘Dramas.

Implementation of K-Means clustering-

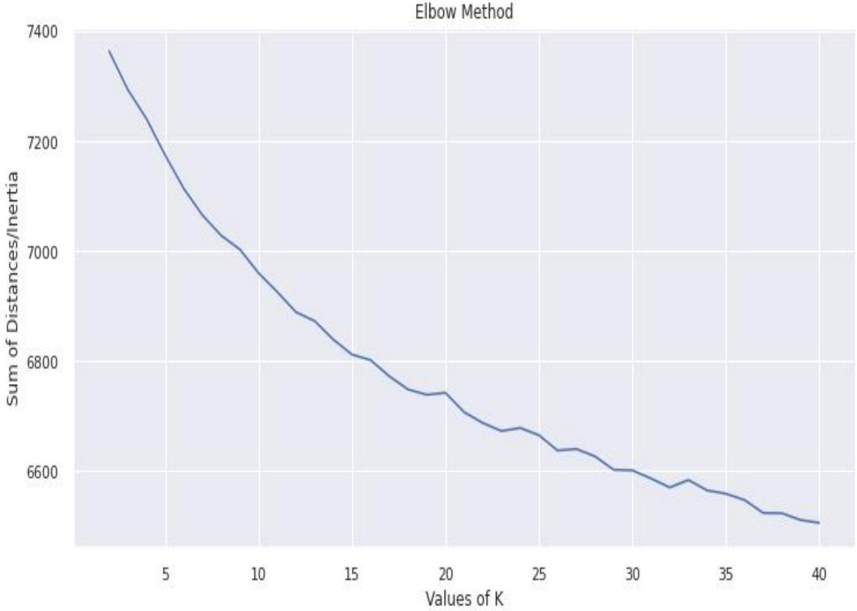
K-means clustering is one of the simplest and popular unsupervised machine learning algorithms.Typically, unsupervised algorithms make inferences from datasets using only input vectors without referring to known, or labelled, outcomes.

To process the learning data, the K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids.

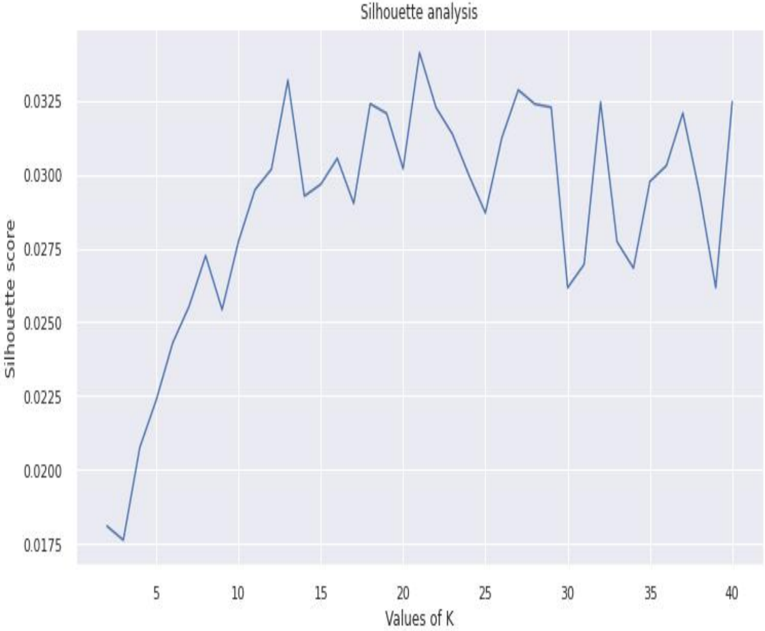


Finding Clusters:-

Elbow Method



Silhouette analysis



Elbow method:- The ‘Elbow method’ is a heuristic method of interpretation and validation of consistency within cluster analysis designed to find the appropriate number of clusters.

Silhouette method:- Silhouette refers to a method of interpretation and validation of consistency within clusters of data. The technique provides a succinct graphical representation of how well each object has been classified.

- **To find the number of clusters we used elbow method and Silhouette's score. then we found that the best optimal number of clusters was 40. from silhouette method we found silhousttes score 0.0325.**

Conclusion:-

- ✓ In this project We done Data Wrangling , feature engineering, and EDA since loading the dataset, and then we have completed some tasks that were assigned to us.
- ✓ We found that Movies uploaded on Netflix are more than twice the TV Shows uploaded, There are about 70% movies and 30% TV shows on Netflix.
- ✓ We can see that in top actors Anupam Kher has acted in the highest number of films on Netflix and Drama is the most popular genre followed by comedy.
- ✓ In top countries we found that The United States has the highest number of content on Netflix by a huge margin followed by India.
- ✓ Majority of movies have running time of between 50 to 150 min. From this analysis we can see that more number of movie release in 2017.
- ✓ We have also defined different clusters based on their content and implemented the KMEANS clustering algorithm. And then we determined some clusters we have plot Silhouette and Elbow plot in which we may interact with similar content in connection to that cluster.

THANK YOU!