CAPSTONE PROJECT - 4

Netflix Movies & TV Shows Clustering

COHORT - AUSTIN

Shubham Bhadouria

Contents:

- Introduction.
- Objective
- Dataset Preview.
- Exploratory Data Analysis.
- Data Preprocessing.
- Creating Clusters.
- Conclusions.

Introduction:

Netflix is a media distribution company. It started with DVD distribution via mail but has evolved substantially over the course of its existence. Today, Netflix is focused on streaming video. Some of its content is licensed, and some of the content is produced in-house.

Netflix is an online streaming service that offers a wide variety of award-winning TV shows, movies, anime, documentaries and more on thousands of internet-connected devices. This business is profitable because users make a monthly payment to access the platform. However, customers can cancel their subscriptions at any time. Therefore, the company must keep the users hooked on the platform and not lose their interest. This is where recommendation systems start to play an important role, providing valuable suggestions to users is essential.

Objective:

The objective is to conduct an Exploratory Data Analysis on the given dataset to understand what type of content is available in different countries, genres of the content, frequency of movies/TV shows added on Netflix in various years and use these insights to cluster similar content by matching text-based features and make a simple recommendation system based on cluster data.

Dataset Preview:

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.

Attribute Information

- **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

Dataset summary:

The dataset contains 12 columns and 7787 rows.

There also exist some null values in our data:

null values in director: **2389**

null values in cast : **708**

null values in country : **508**

null values in date_added : **10**

null values in rating : **7**

The null values in above attributes except rating attribute was replaced by 'unknown'.
The null values in rating attribute was replaced by its mode value.

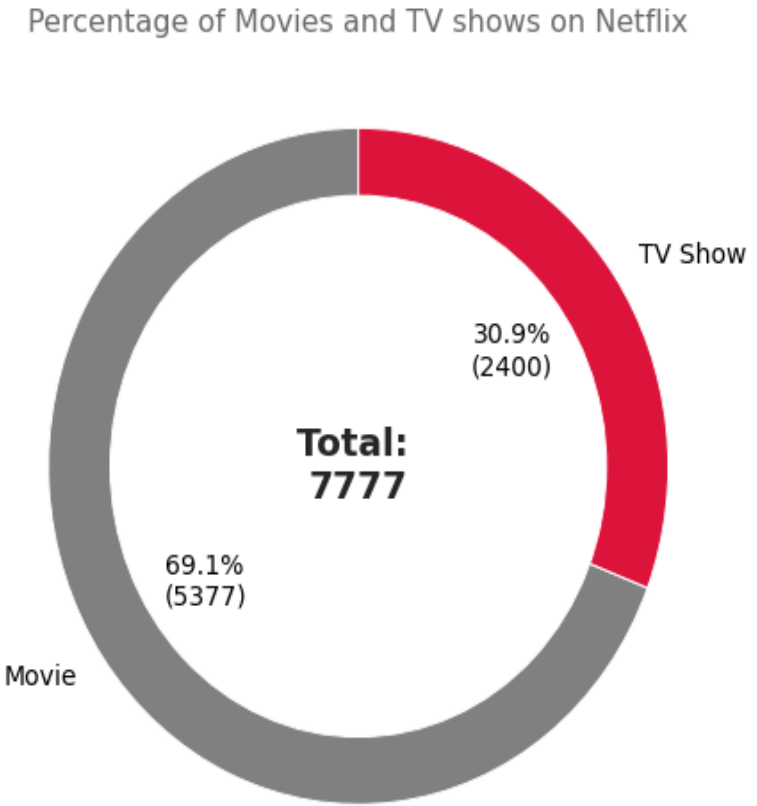
Finally, we will do some feature engineering to create few new variables:

- *Compute year_added, month_added from date_added after converting it into a datetime variable.*

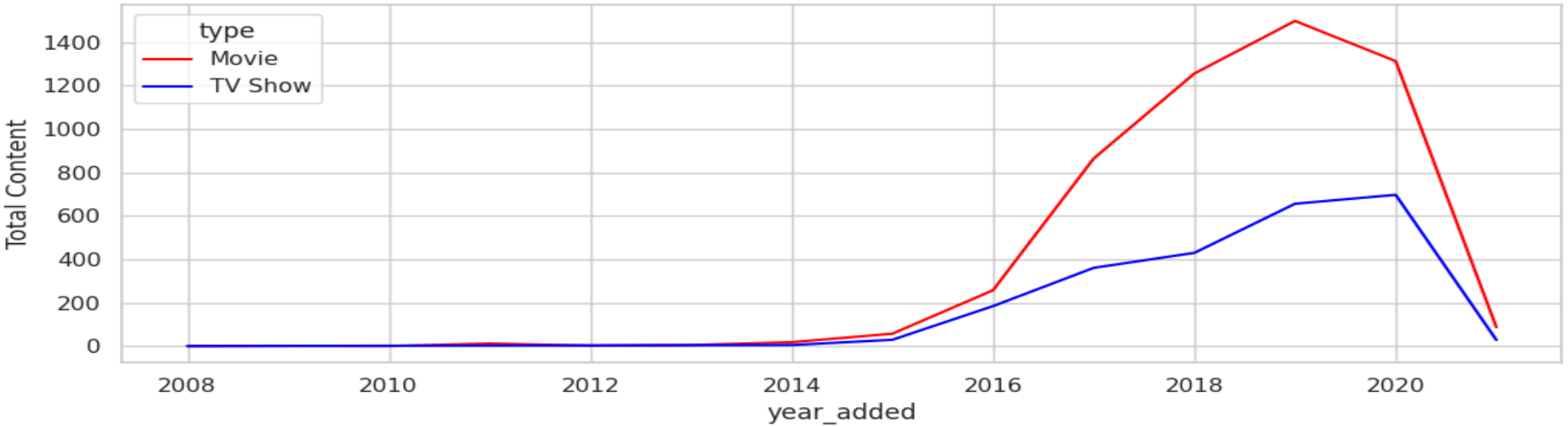
Exploratory Data Analysis:

Type:

69.1% of the content available on Netflix are movies and the remaining 30.9% are TV Shows.



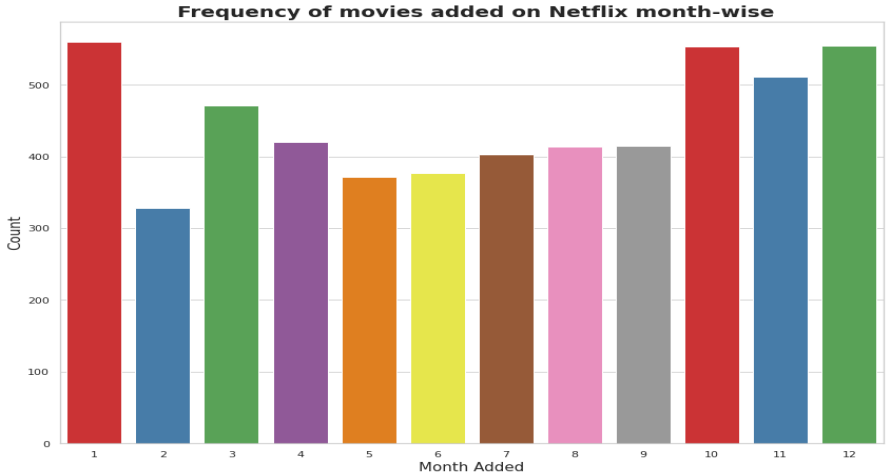
Year_added:



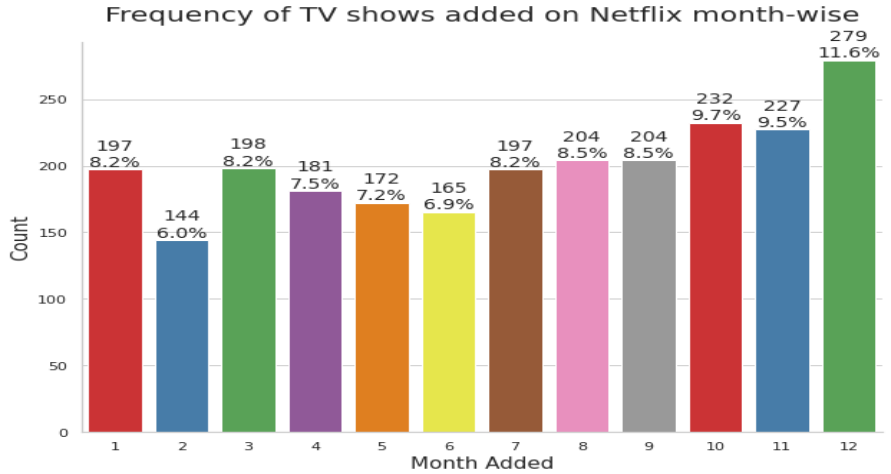
- The number of movies and TV shows added on Netflix has been increasing significantly(almost exponentially) from 2015 to 2019 and after 2019, the number of movies and TV shows added on Netflix has been decreasing significantly due to Covid-19 pandemic
- In 2019, maximum number of TV shows(total 656 shows) and movies(total 1497 movies) has been added on Netflix.
- Increase in the total movies added on Netflix is much higher than the total TV shows added on Netflix.

Month_added and Day_added:

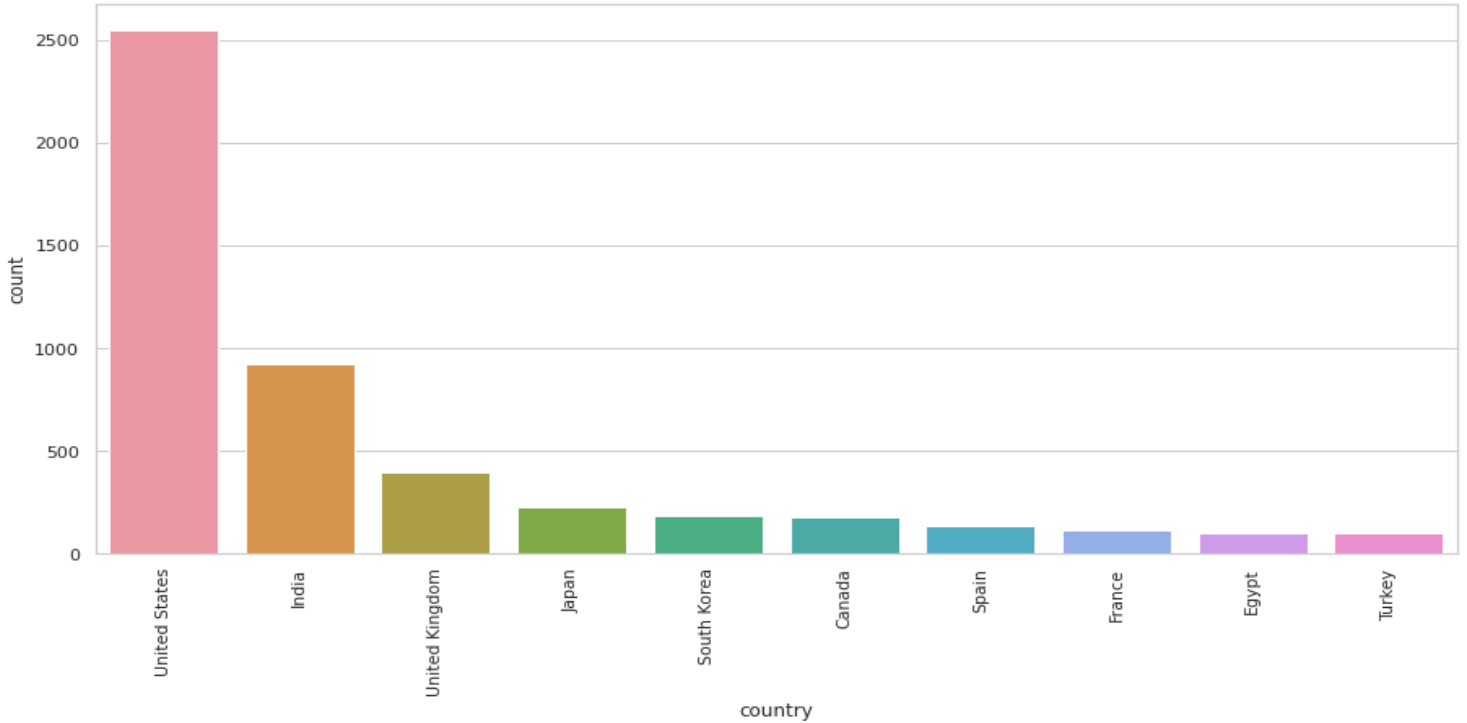
- Most of the movies has been added in the month October, November, December and January.
- Maximum movies has been added in January and minimum movies has been added in the February.



- Most of the TV shows has been added in the month October, November, December and January.
- Maximum shows has been added in December and minimum shows has been added in the February.

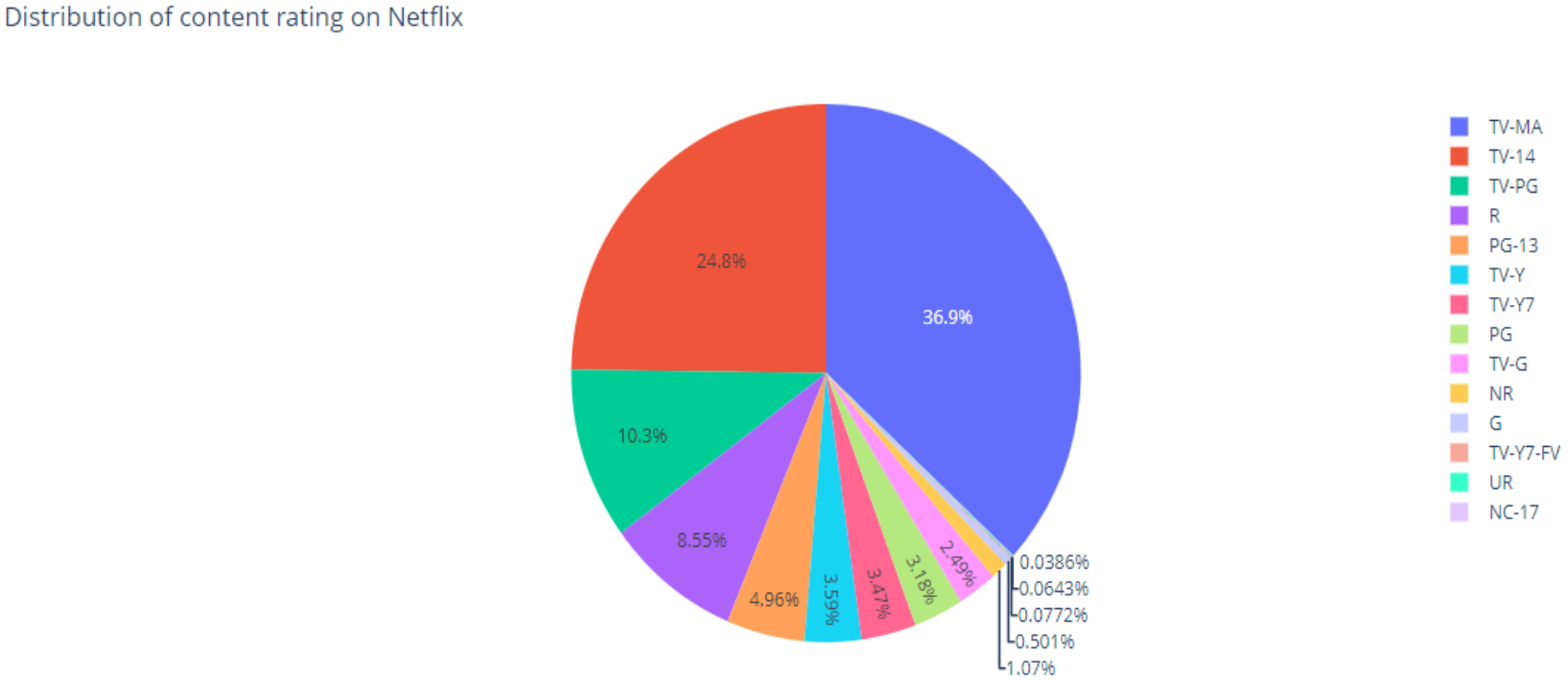


Country:



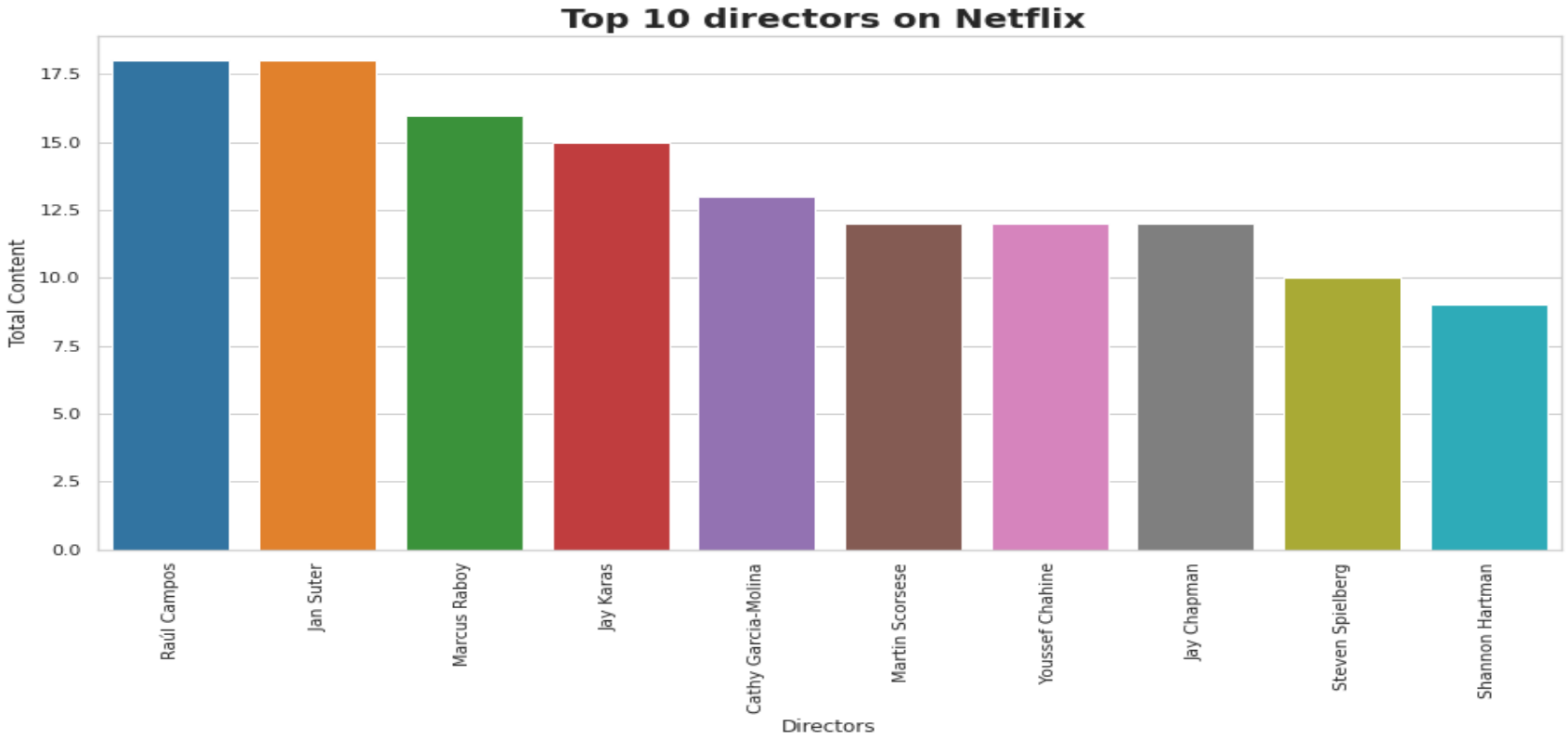
USA, India, and Uk create more than half of the tv shows and movies on the platform.

Rating:



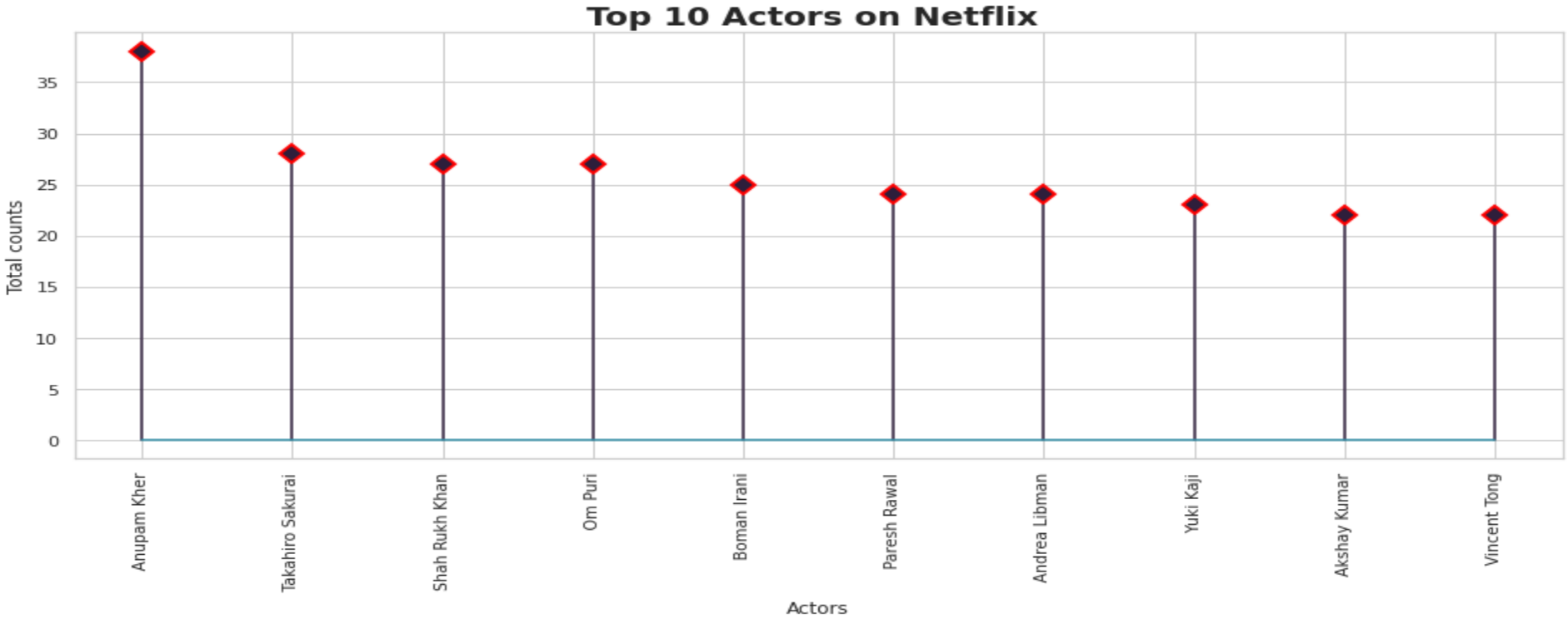
- **Most of the Movie/TV shows has TV-MA ratings(36.9%) followed by TV-14(24.8%)**

Director:



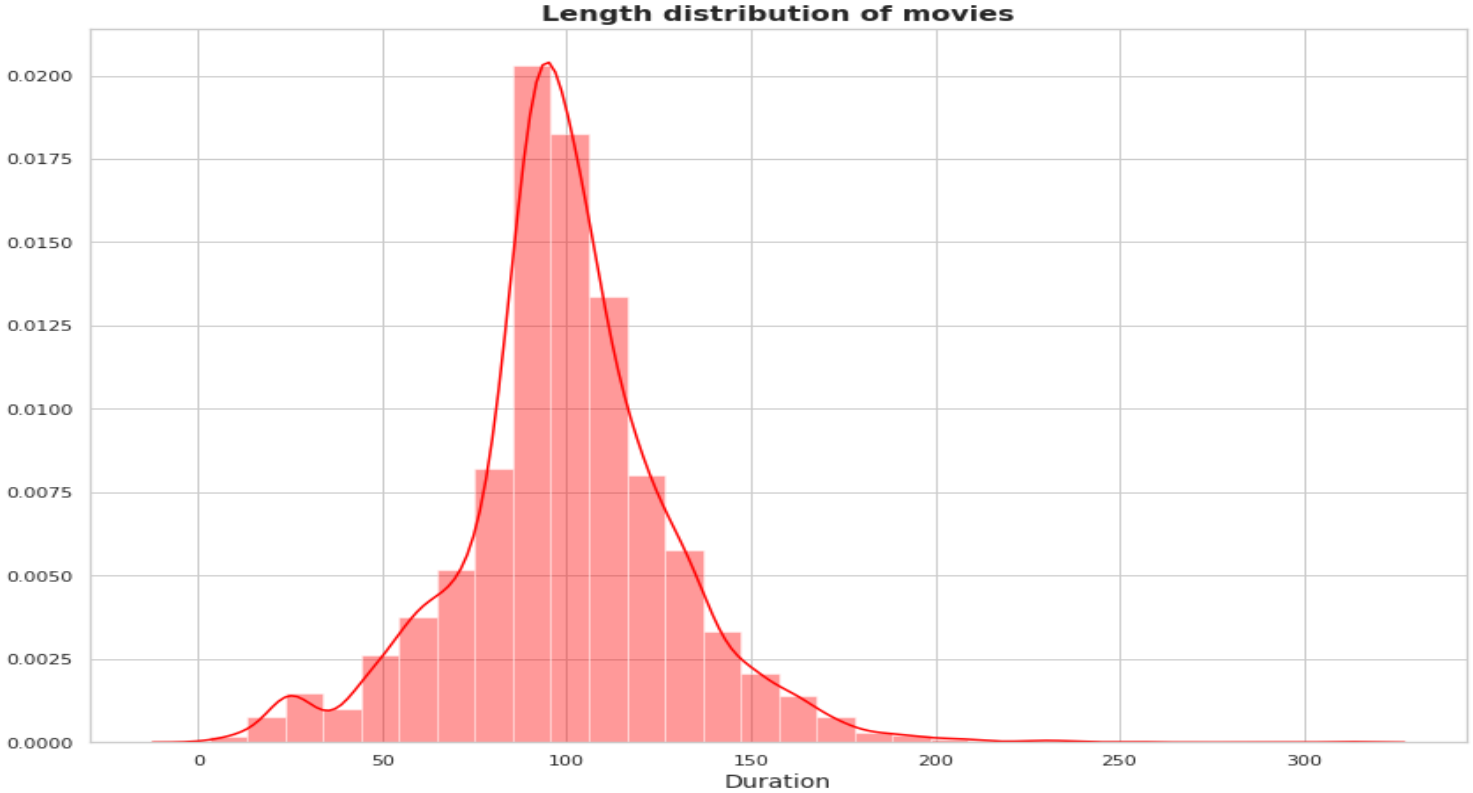
- **Raul Campos and Jan Suter has directed the most number of movies/Tv shows (total 18 both) followed by Marcus Raboy (total 16).**

Cast:



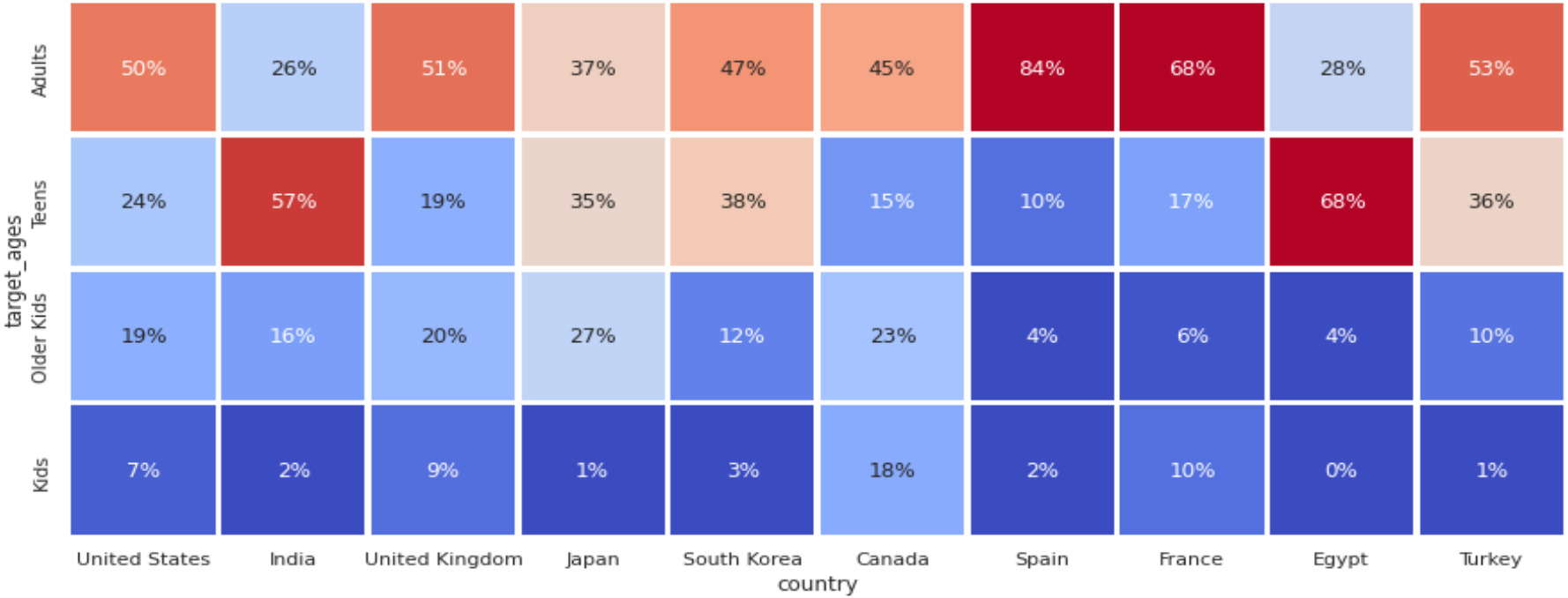
- Six of the actors in the top ten list with most numbers tv shows and movies are from India.
- Anupam Kher has acted in the most number of movies/Tv shows (total 38)

Duration:



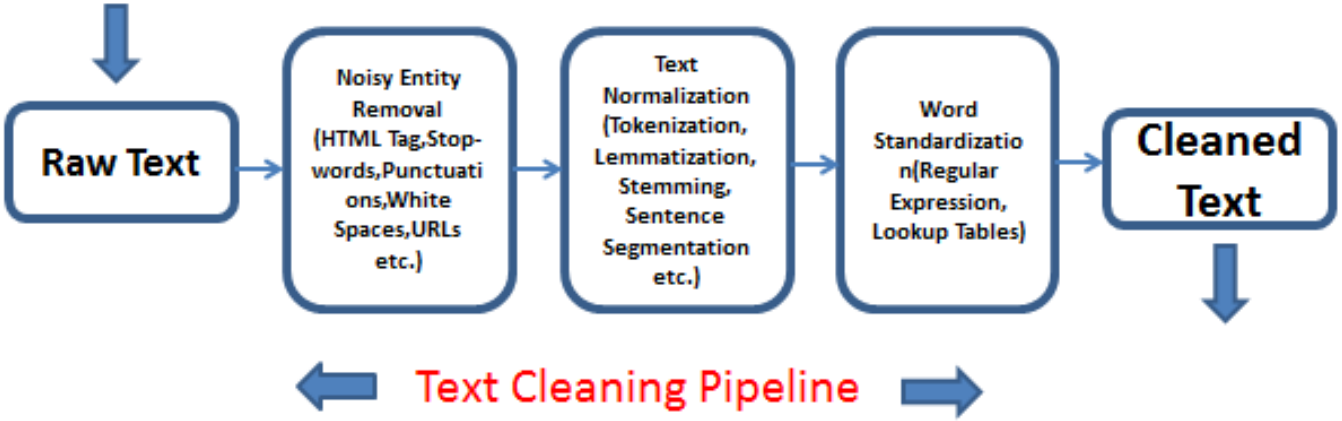
- **Most of the movies last for 90 to 120 minutes.**

Target audience proportion in top 10 countries:



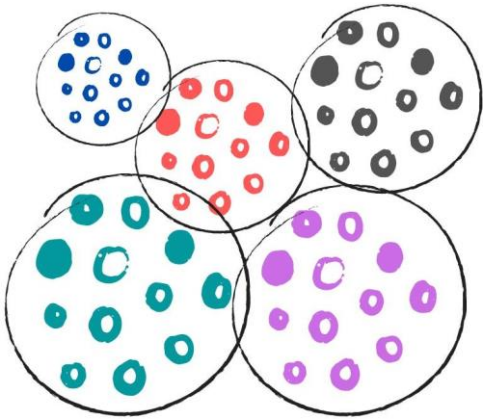
- **Content available for kids is less as compared to other categories and content available for adults is more in almost every country except India**
- **In India, most of the content is available for teens.**

Data Preprocessing.



Creating Clusters:

What is clustering?

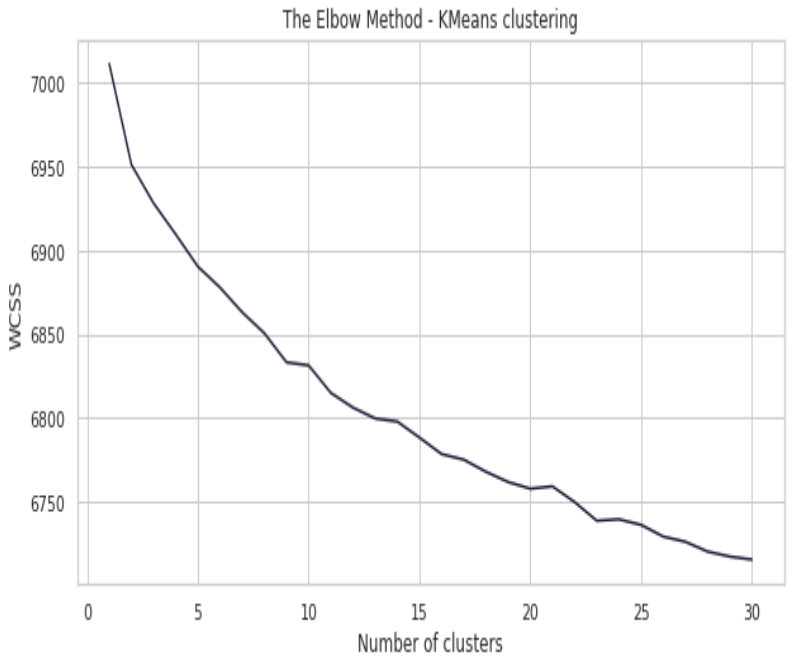


Clustering is the task of dividing the data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them.

How to cluster similar data?

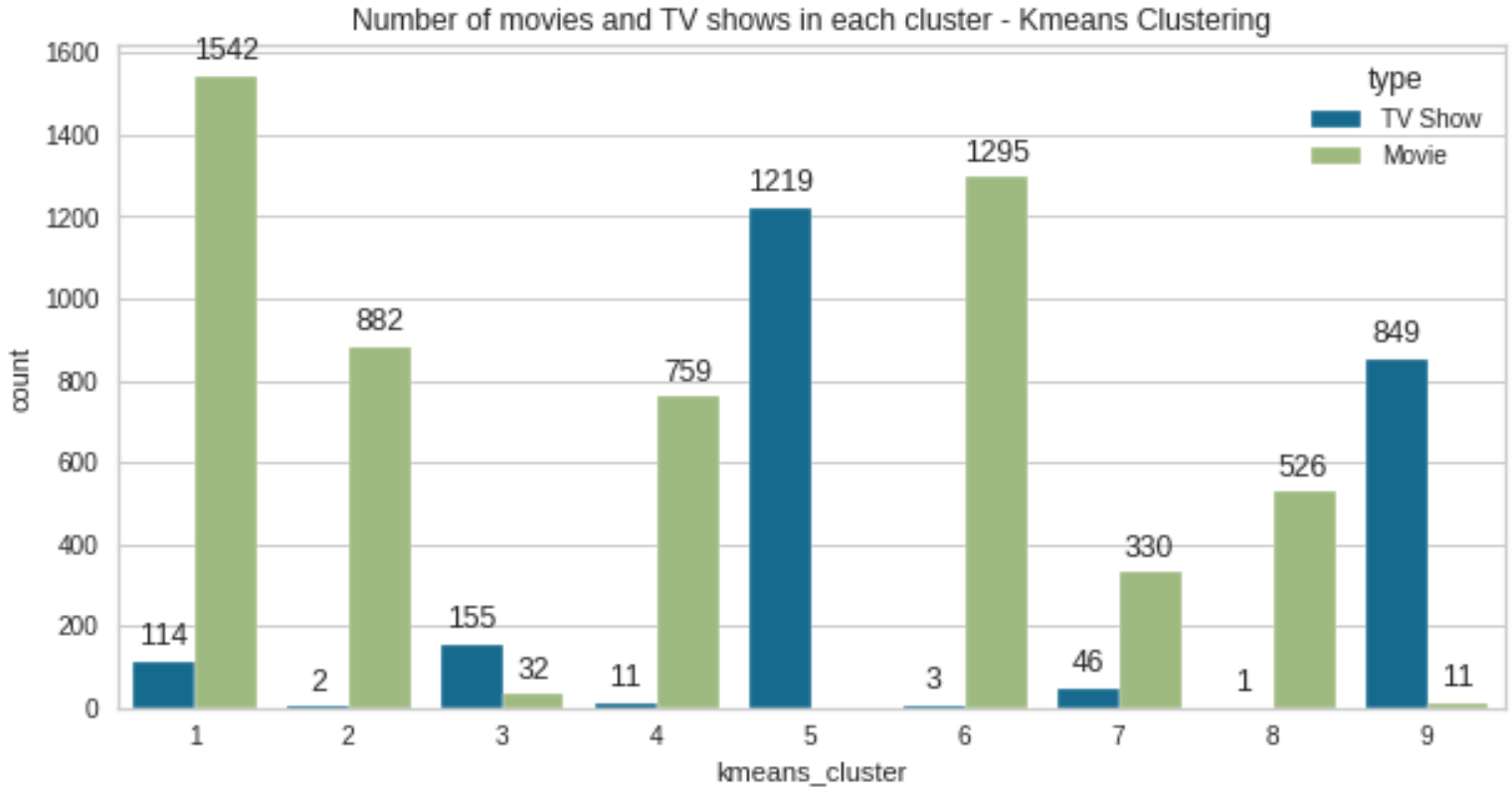
To create clusters, we use the K-Means Clustering and hierarchical clustering.
K-means clustering is a distance-based unsupervised clustering algorithm where data points that are close to each other are grouped in a given number of clusters/groups.
Hierarchical clustering is a method of cluster analysis that seeks to build a hierarchy of clusters.

Determining optimal value for k: K mean clustering algorithm

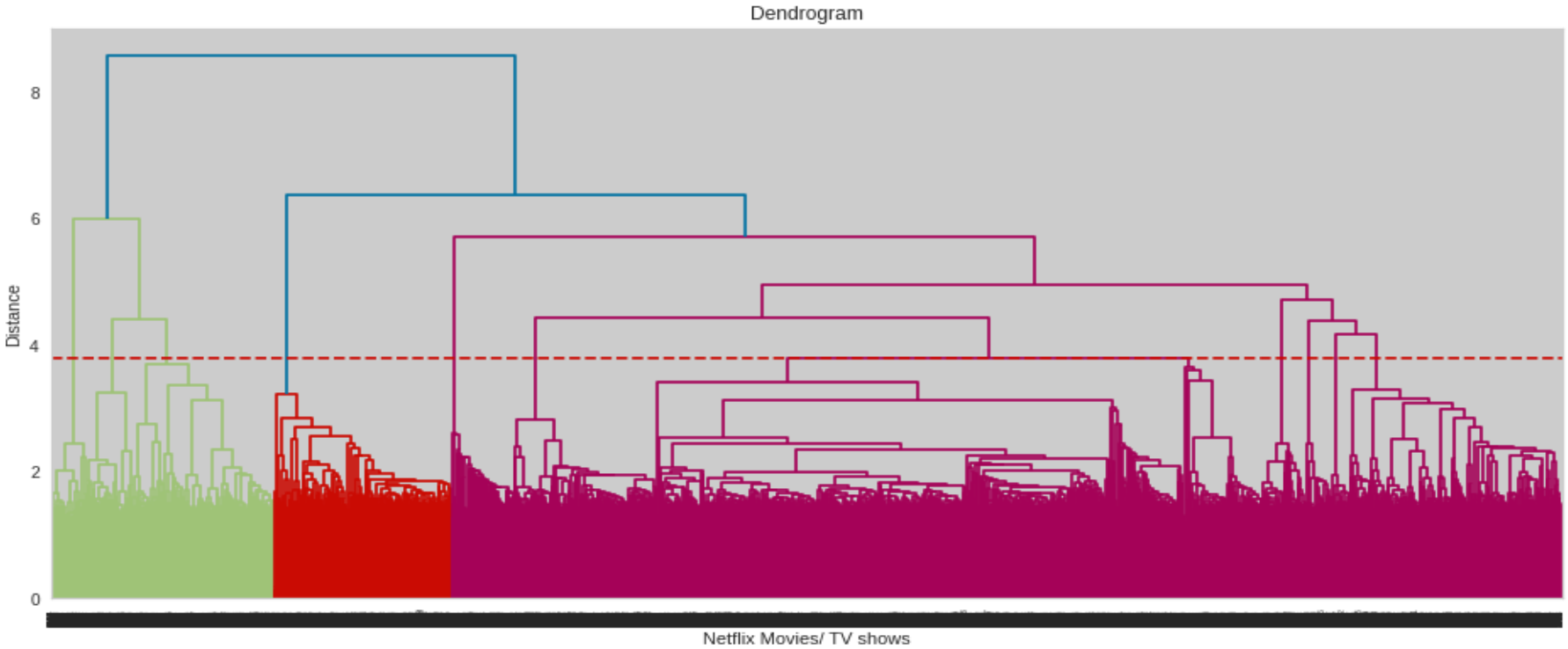


- **Using the Silhouette Score and Elbow Method we select the optimal number of clusters to be 9.**

Number of movies and tv shows in each cluster

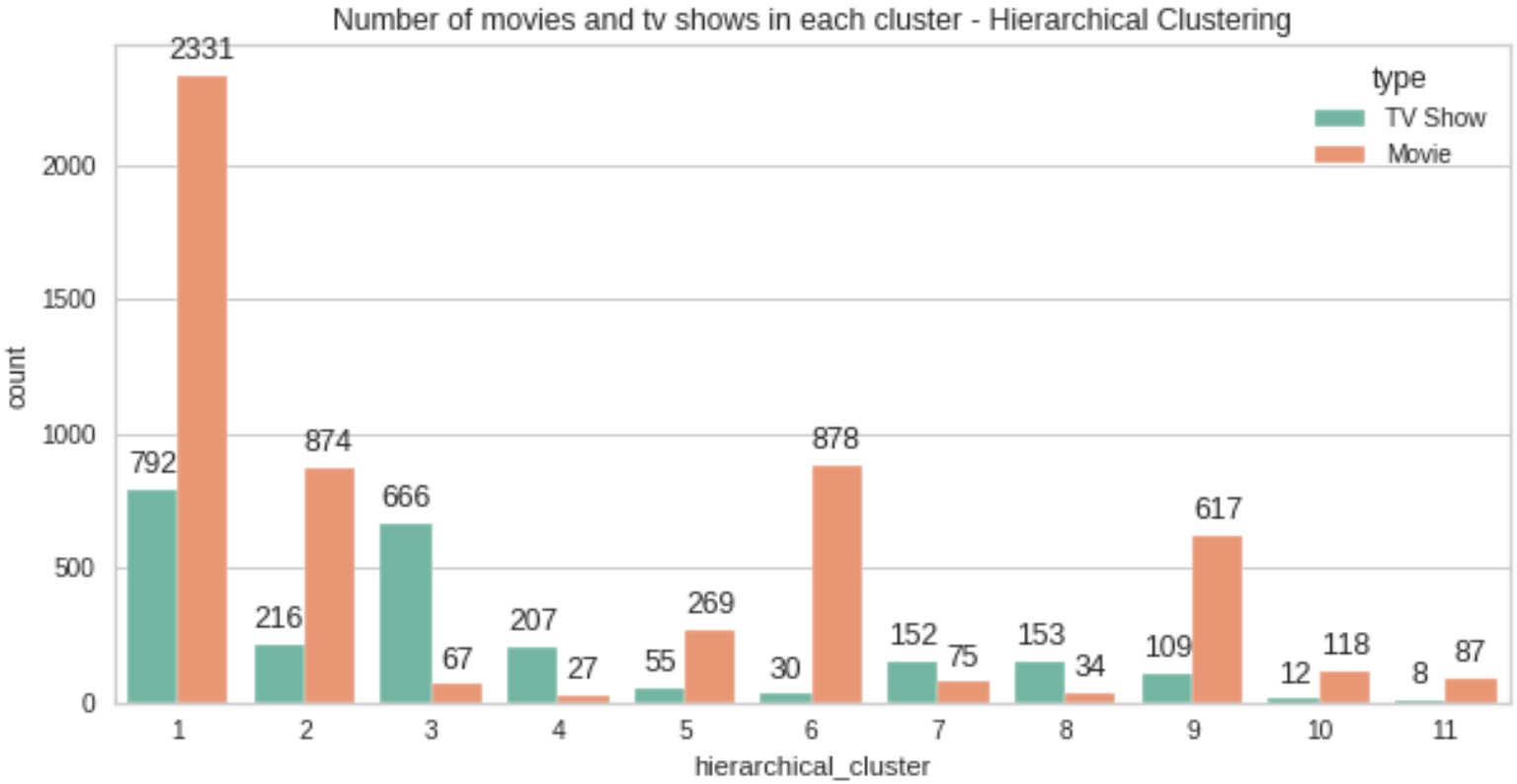


Determining optimal value for k: Hierarchical clustering algorithm



- The red dotted horizontal line cuts the 11 vertical lines in the dendrogram, therefore optimum number of clusters will be 11.

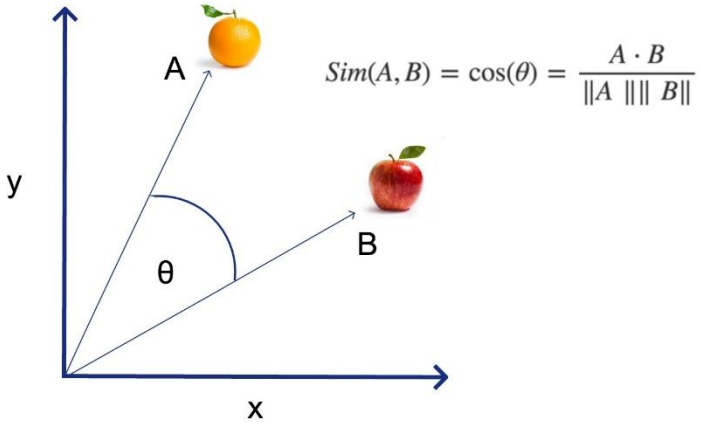
Number of movies and tv shows in each cluster



Getting Recommendations:

We obtained recommendations for Movies and Tv- Shows using Cosine similarity.

Cosine Similarity



$$Sim(A,B) = \cos(\theta) = \frac{A \cdot B}{\|A\| \, \|B\|}$$

```
If you liked 'Indiana Jones and the Last Crusade', you may also enjoy:
```

	Recommendations
0	Indiana Jones and the Raiders of the Lost Ark
1	Indiana Jones and the Kingdom of the Crystal S...
2	Indiana Jones and the Temple of Doom
3	Searching for Bobby Fischer
4	A Bridge Too Far
5	Santa Girl
6	Dragons: Dawn of the Dragon Racers
7	Logan's Run
8	Tellur Aliens
9	Monty Python and the Holy Grail

```
If you liked 'Zoids Wild', you may also enjoy:
```

	Recommendations
0	Magi: The Labyrinth of Magic
1	JoJo's Bizarre Adventure
2	March Comes in Like a Lion
3	K
4	Haikyu!!
5	Devilman Crybaby
6	Magi: Adventure of Sinbad
7	Kuroko's Basketball
8	Code Geass: Lelouch of the Rebellion
9	Levius

Conclusions:

- **Overall Netflix has more movies than the TV shows in a percentage of 69.1% against 30.9%. Netflix has 5377 movies, which is more than double the quantity of TV shows. Increase in the total movies added on Netflix in different years is much higher than the total TV shows added on Netflix.**

- **Raul Campos and Jan Suter has directed the most number of movies/Tv shows (total 18 both) followed by Marcus Raboy (total 16).**

- **Anupam Kher has acted in the most number of movies/Tv shows (total 38) followed by Takahiro Sakurai (total 28).**

- **Majority of the Movies/TV shows were produced in the United States, and the majority of the shows on Netflix were created for adults and teens means mature content is more popular on Netflix.**

Conclusions:

- Clusters are made using the k-means clustering algorithm and Agglomerative clustering algorithm.
    - With k-means clustering algorithm, optimal number of clusters came out to be 9. This was obtained through the elbow method and Silhouette score analysis.
    - With Agglomerative clustering algorithm, optimal number of clusters came out to be 11. This was obtained after visualizing the dendrogram.

- A simple content based recommender system was created using cosine-similarity which makes 10 recommendations to the user based on the type of movie/show they watched.

Future Scope:

- Building a recommendation system is a very challenging and time consuming process and there is always the chance of improvement, so more time could be given in making a better recommender system, which later can be deployed on web for usage.

- Integrating this dataset with other external datasets such as IMDB ratings, rotten tomatoes can also provide many interesting findings.


thank
you