Problem Statement

This dataset consists of tv shows and movies available on Netflix as of 2019. The dataset is collected from Fixable 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, rotten tomatoes can also provide many interesting findings.

Inspiration Understanding what content is available in different countries Identifying similar content by matching text-based features Network analysis of Actors / Directors and find interesting insights Is Netflix has increasingly focusing on TV rather than movies in recent years.

Table of Content

- **In this project, we did**
- **Data Overview**
- **Exploratory Data Analysis**
- **Model Building**
- **Overall Conclusion**

In this project, we did

1. Exploratory Data Analysis
2. Understanding what type content is available in different countries
3. Is Netflix has increasingly focusing on TV rather than movies in recent years.
4. Clustering Algorithm (Agglomerative Clustering Algorithm)

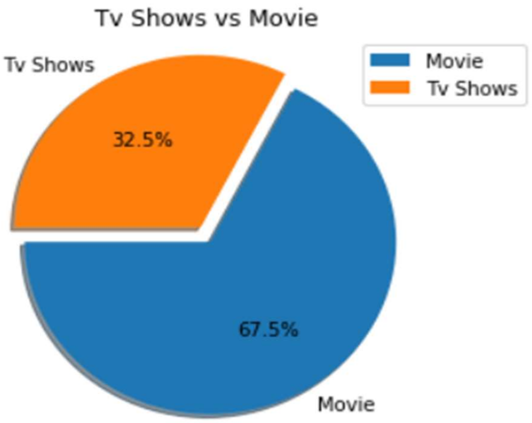
Data Overview:

Attribute Information

1. show_id : Unique ID for every Movie / Tv Show
2. type : Identifier - A Movie or TV Show
3. title : Title of the Movie / Tv Show
4. director : Director of the Movie
5. cast : Actors involved in the movie / show
6. country : Country where the movie / show was produced
7. date_added : Date it was added on Netflix
8. release_year : Actual Releaseyear of the movie / show
9. rating : TV Rating of the movie / show
10. duration : Total Duration - in minutes or number of seasons
11. listed_in : Genere
12. description: The Summary description

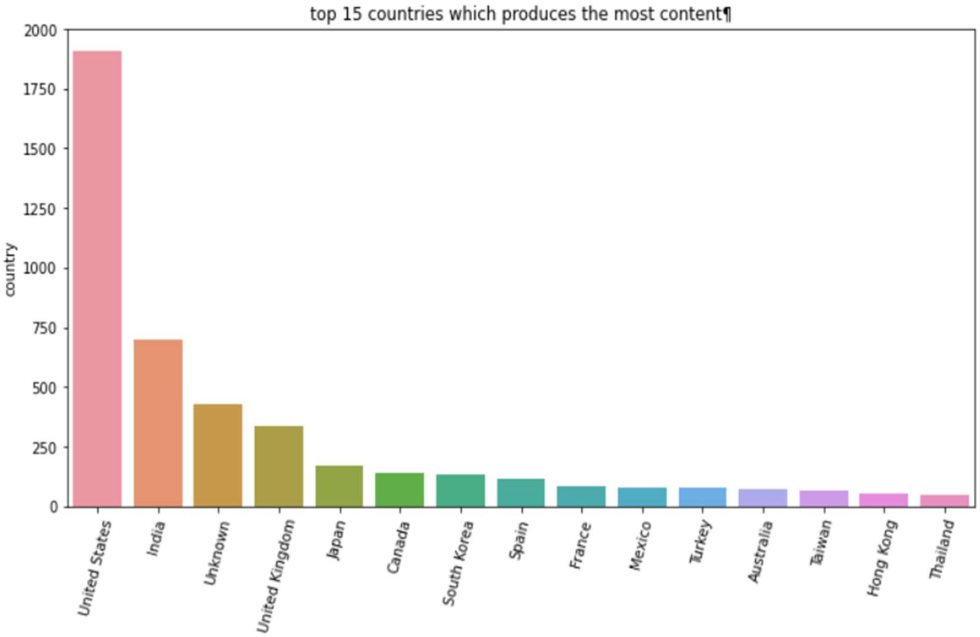
Exploratory Data Analysis

1. Percentage of Movies and Tv shows in this dataset



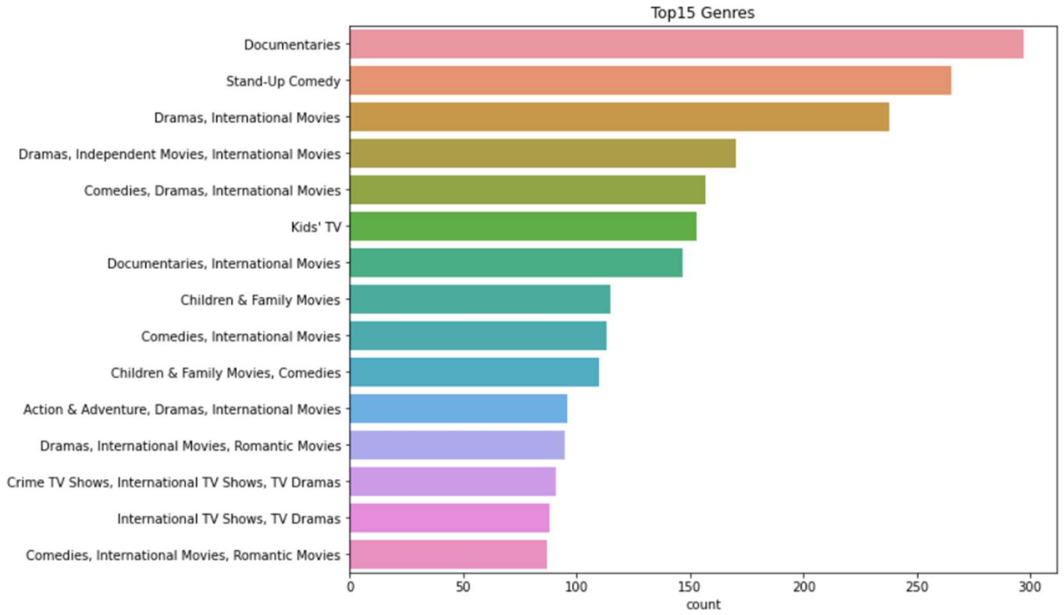
In netflix dataset there are 32.5% Tv shows and 67.5% movies available.

2. Top 15 countries which produces the most content



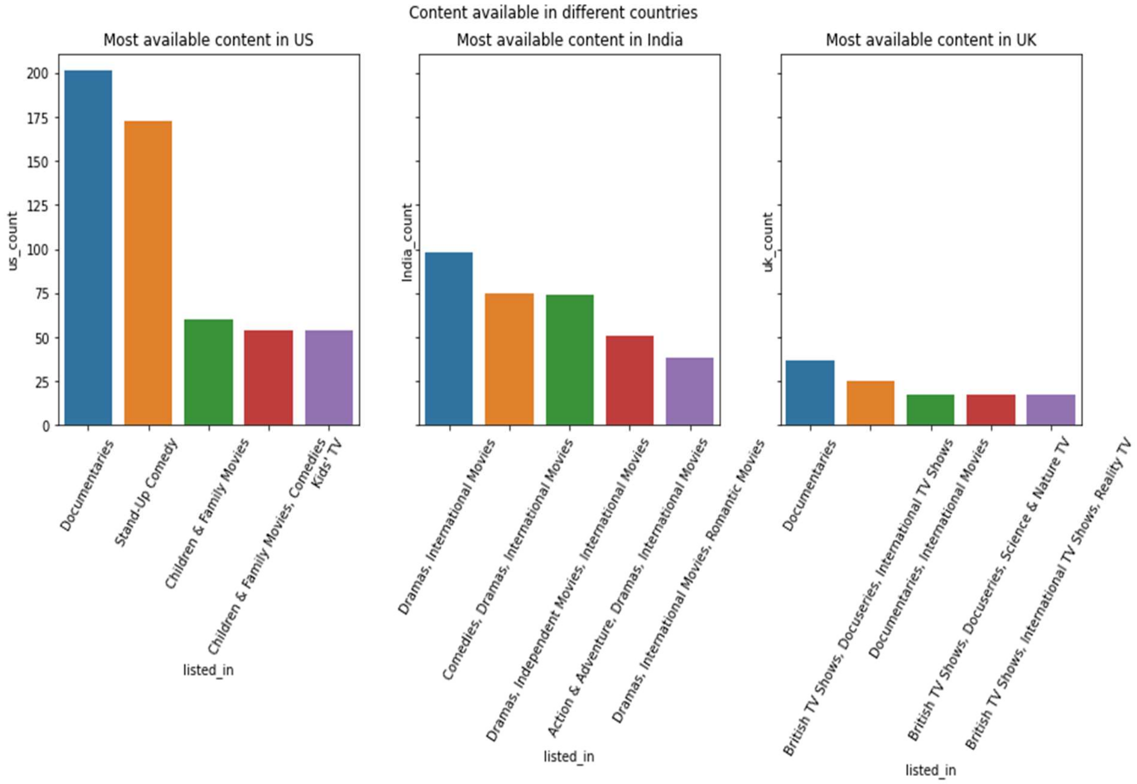
United States produces highest content followed by india second highest.

3. Top15 Genres

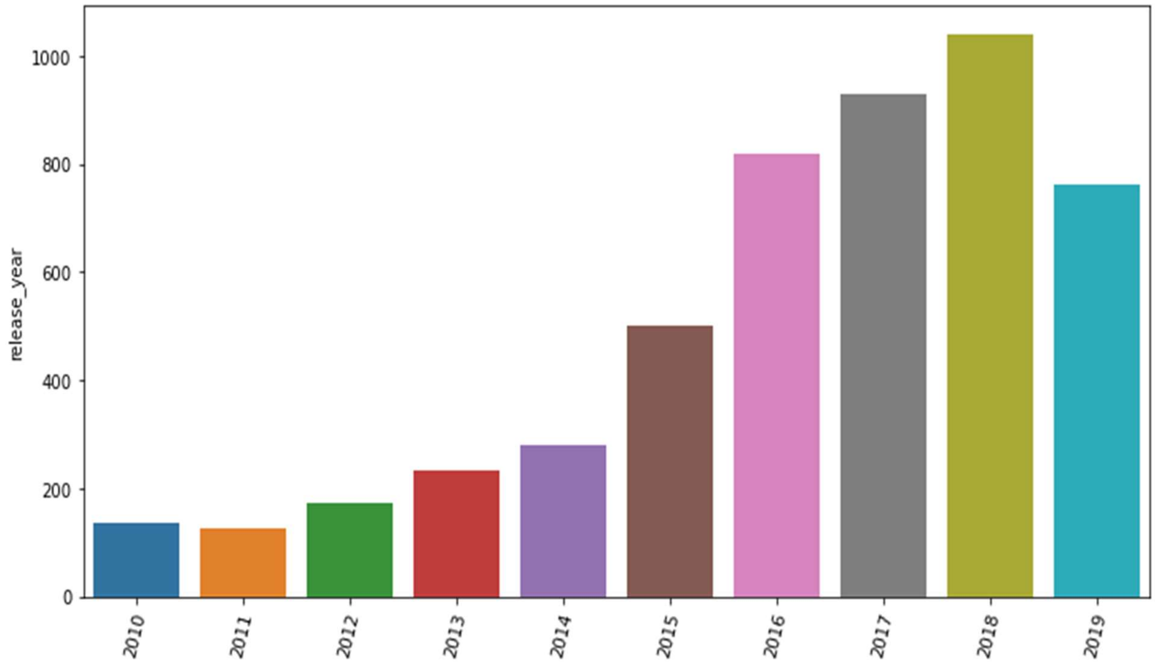


Highest numbers of Documentaries are there followed by Stand-Up Comedy in terms of Genres.

4.Understanding what content is available in different countries.



5. **Movies and Tv shows are released per year**

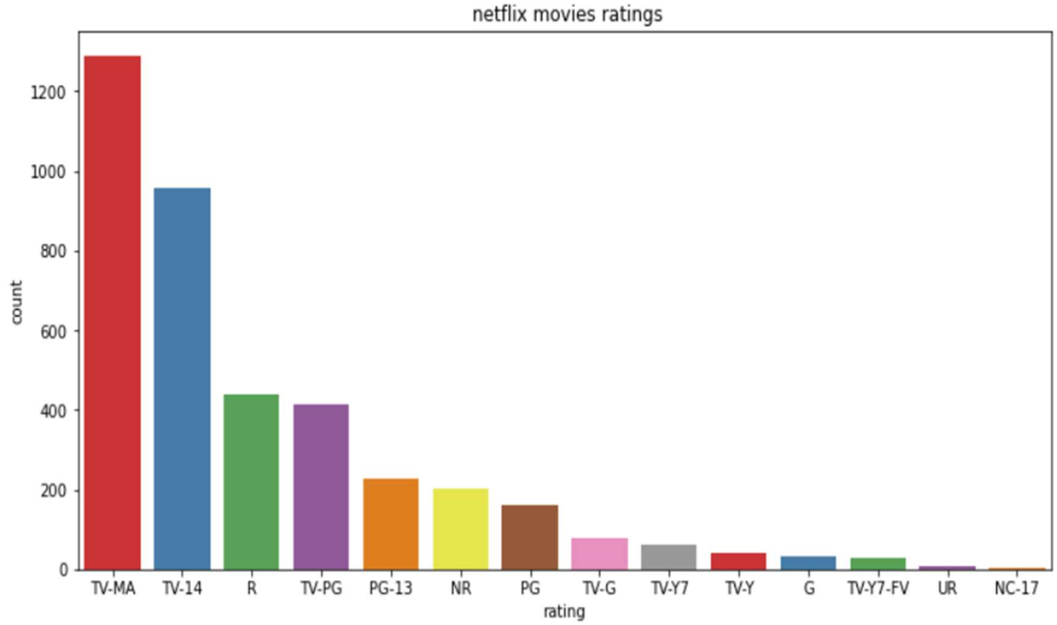


In compare to 2018 there are less no . of tv shows and movies released in the year 2019.

6.TV Shows and Movies before and after 2010.

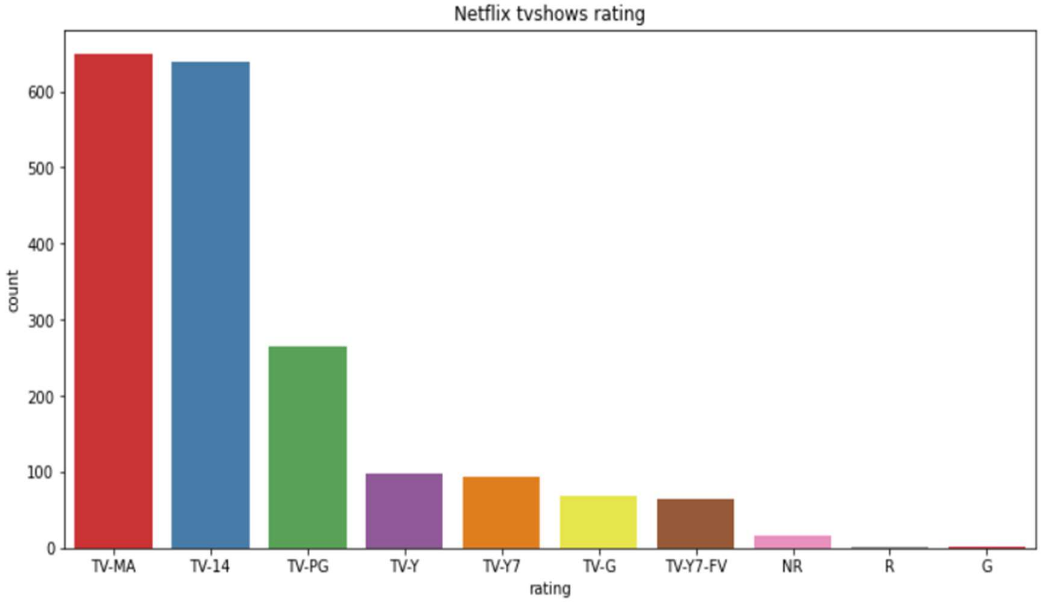
	Tv Shows	Movies
Before 2010	189	783
After 2010	1709	3156
% increase	90%	40%

7. Movies Rating



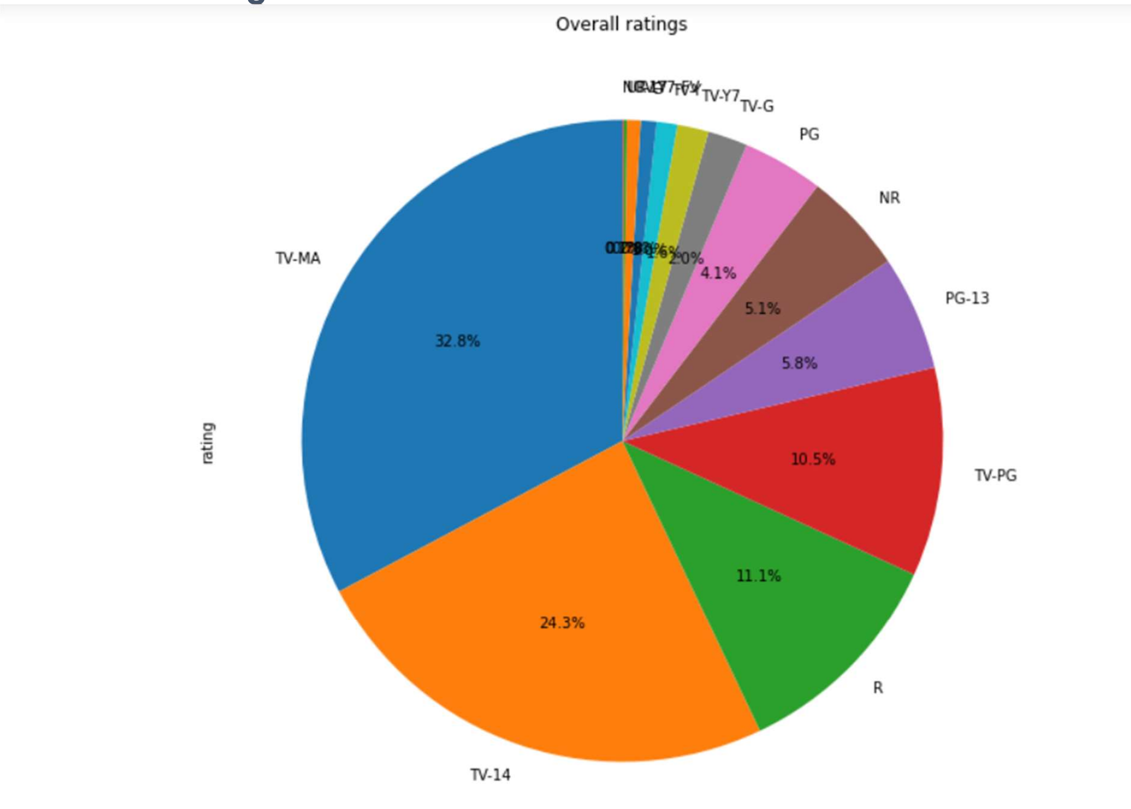
TV-MA and TV-14 are the most common type of rating

Tv Shows Rating



TV-MA and TV-14 are the most common type of rating

8.Overall Rating

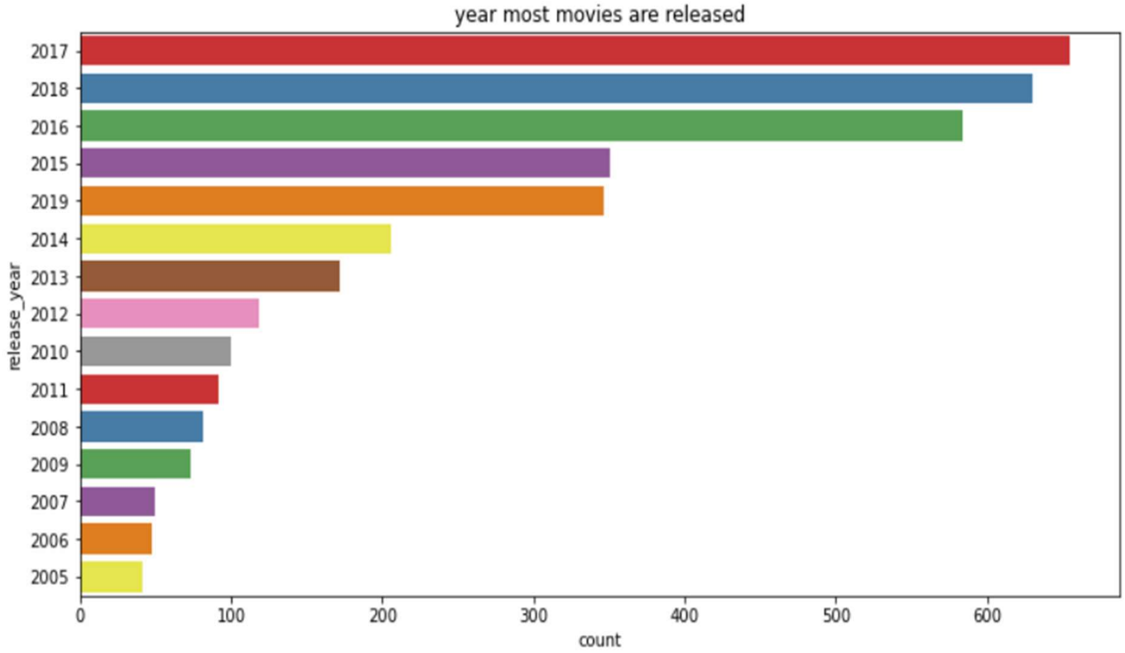


The top3 ratings are TV-MA, TV-14, TV-PG with 32.6%, 27.3%, and 11.3% respecively

9. Maximum duration and seasons of movies and TV shows with their names.

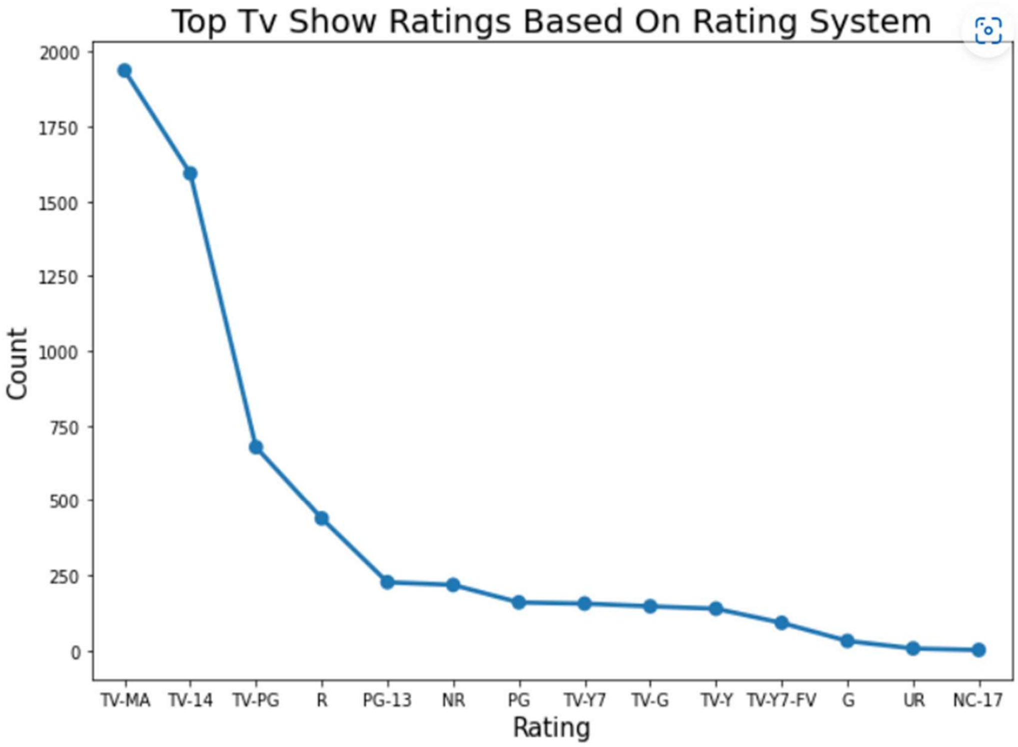
	Duration and seasons	Name of movies and TV shows
maximum duration of the movie	312	Black Mirror: Bandersnatch
Tv shows has the maximum number of seasons	15	Grey's Anatomy ,NCIS

10. year most movies are released

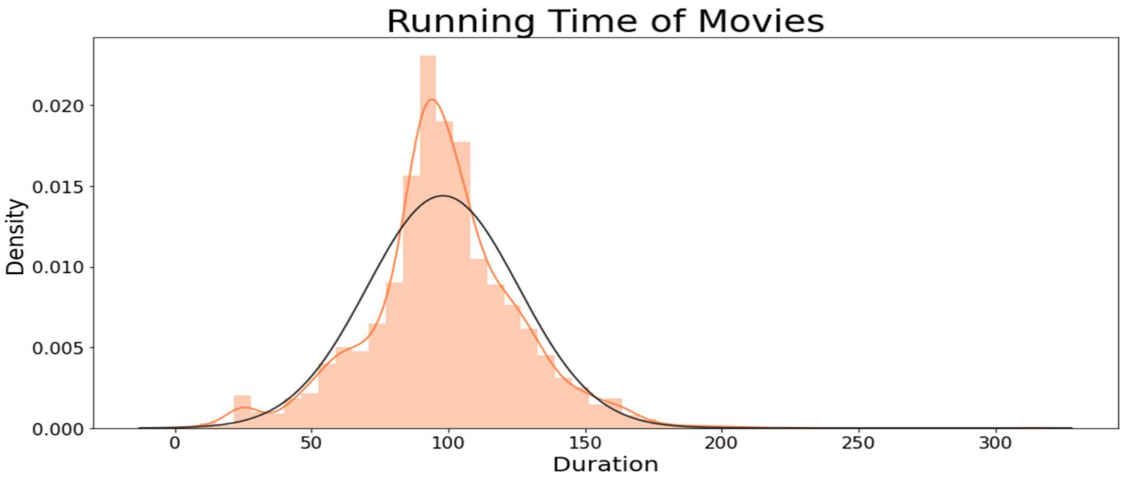


In the year 2018, most of the movies are released.

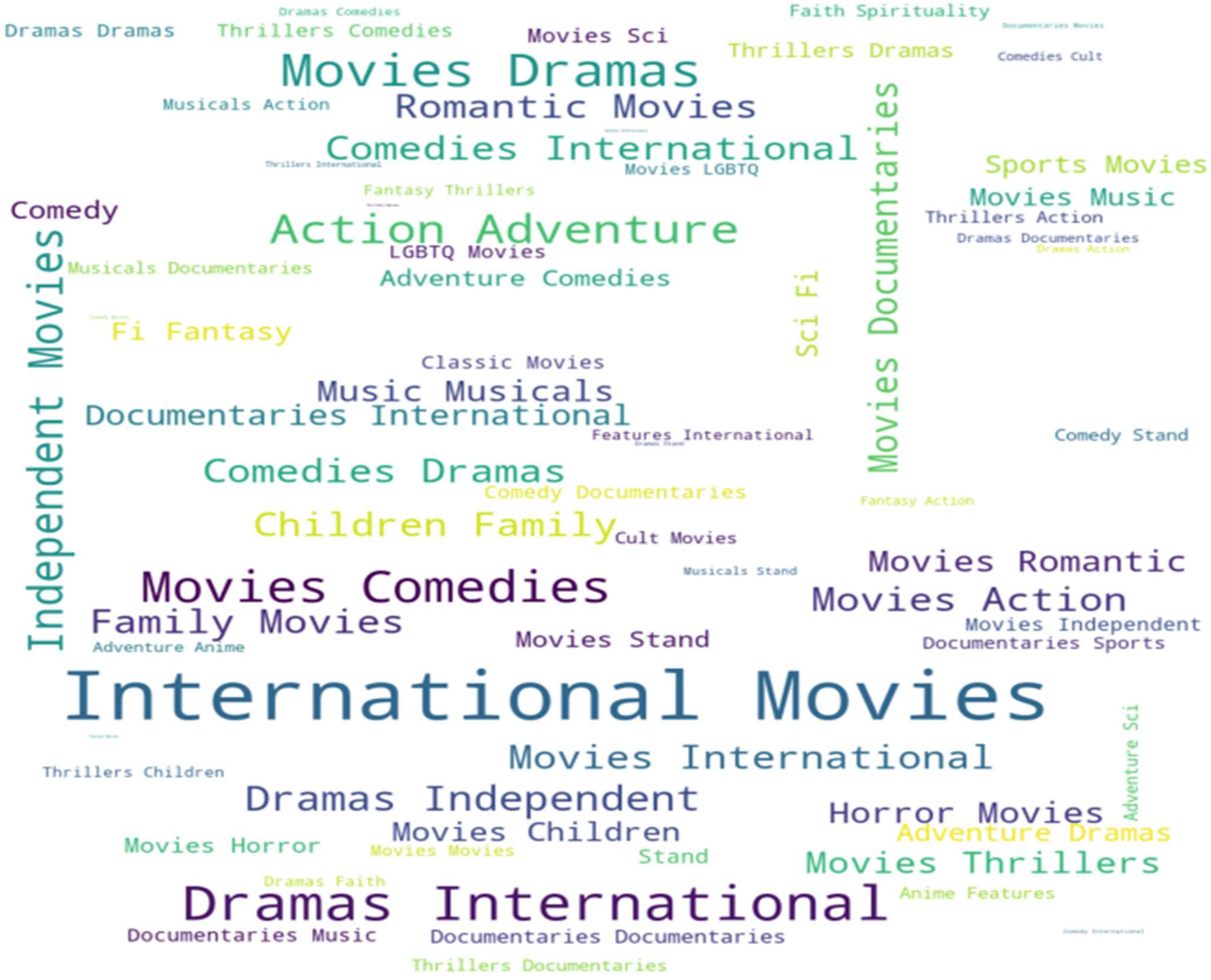
11. Top Tv Show Ratings Based On Rating System



12.Normal distribution of duration of movies.



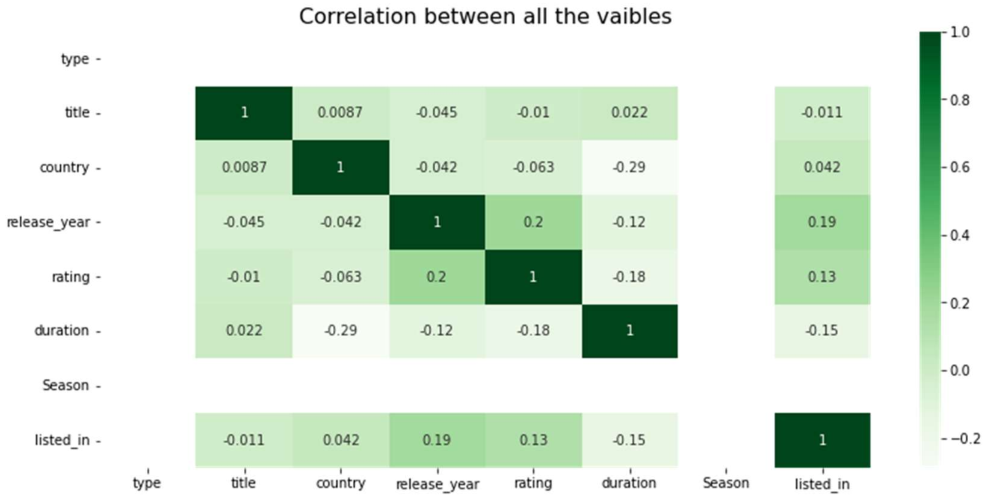
13.Word cloud for Genres



From the above word cloud we can conclude that the genres International Movies is having highest number and drama comedies having lowest number.

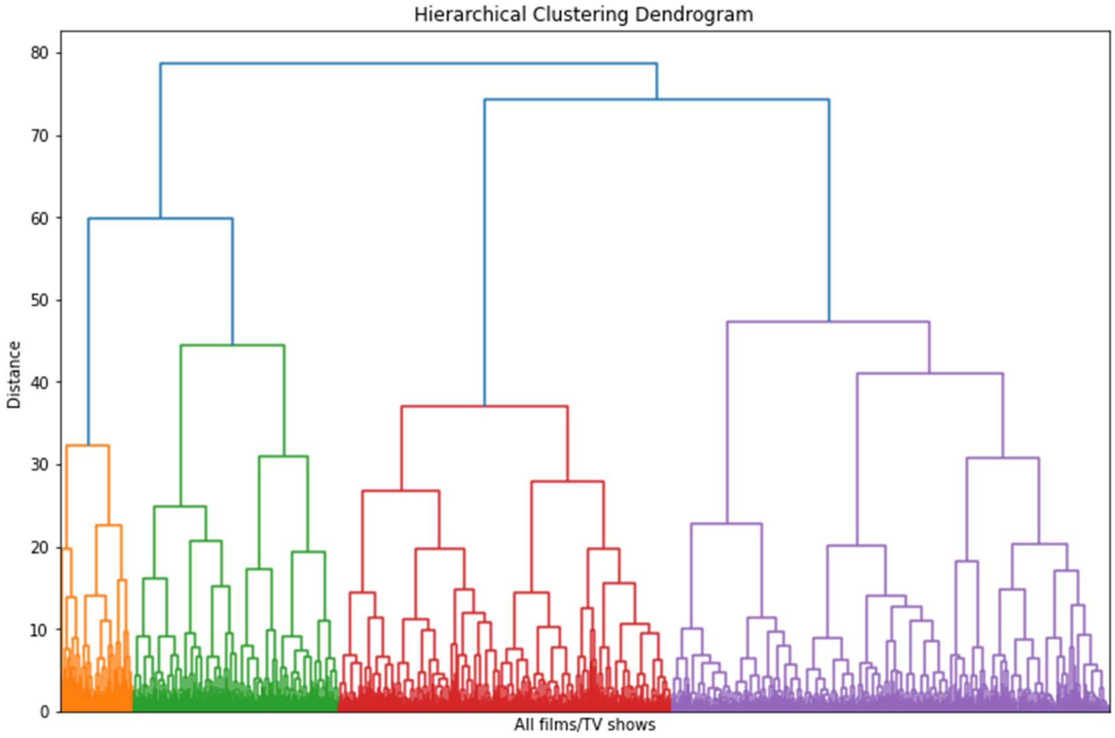
Model Building

Co-relation Between the variables



From the above graph we can say that all variables are independent.

Clustering Agglomerative algorithm



Assume we cut vertical lines with a horizontal line to obtain the number of clusters.
Number of clusters = 4

Each level of dendrogram has a subtle meaning to the relationship between its data members. In a regular relationship chart, one may interpret that at the top lies grandparents or the first generation, the next level corresponds to parents or second generation and the final level belongs to children or third generation. Likewise, in every branching procedure of dendrogram, all the data points having the membership at each level belongs to a certain class.

In this algorithm, we start with considering each data point as a subcluster. We define a metric to measure the distance between all pairs of subclusters at each step and keep merging the nearest two subclusters in each step. We repeat this procedure till there is only one cluster in the system.

Inferences and Conclusion

Movie : 3939 , TV Show : 1898

On average the movie lasts 97 minutes

Maximum number of seasons in TV-Shows is 15

Mostly movies and TV-shows are made in America

TV-MA and TV-14 are the most common type of rating

From the word cloud graph we can conclude that the genres International Movies is having highest number and drama comedies having lowest number.

In year 2018 most of the movies and tv shows released.

Top 3 ratings are TV-MA,TV-14 ,TV-PG.