Capstone Project - 4

NETFLIX MOVIES AND TV SHOWS CLUSTERING

Saurav Kumar
bhadanisaurav7@gmail.com

Index

Problem Statement

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 services 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.

In this project, I have done

- **1**.Exploratory Data Analysis
- 2.Understanding what type 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 pipeline

- **Data Processing :** In this first we removed or replaced or the null and duplicate values as required, and removed the outliers
- **Feature Engineering :** In this we go through each feature, add new feature, changed the column which have object data type, do encoding on categorical variable, dropped the columns which are not necessary.
- **EDA:** In this part we do some exploratory data analysis on the feature which we occupied after feature engineering to explore some useful trends in our data.
- **Create a model :** We have to cluster the similar content based movie and shows. Here I take two approach one by using topic modelling and other by using k nearest cluster.

Data Summary

The dataset consists of listings of all the movies and tv shows available on Netflix, along with details such as - cast, directors, ratings, release year, duration, etc

- **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

Data Summary

- Some important Features in HEALTH INSURANCE CROSS SELL PREDICTION dataset
  - **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 and Feature engineering

For data cleaning:

- We have no duplicates presents in dataset.

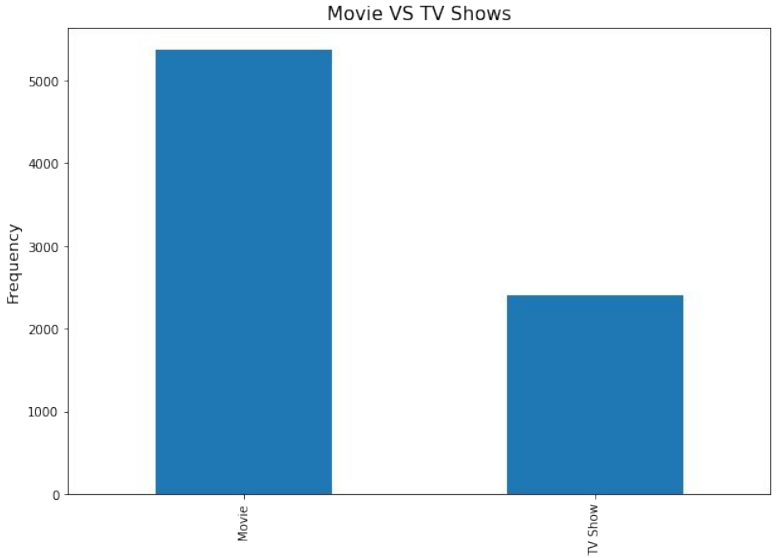
- **Removing null values**
  - We remove the director column from dataset as it has more than 30% of entries unavailable(null).
  - We replace the NaN value from mode value if the frequency of mode in the column is greater than 30% otherwise replace it by unknown.

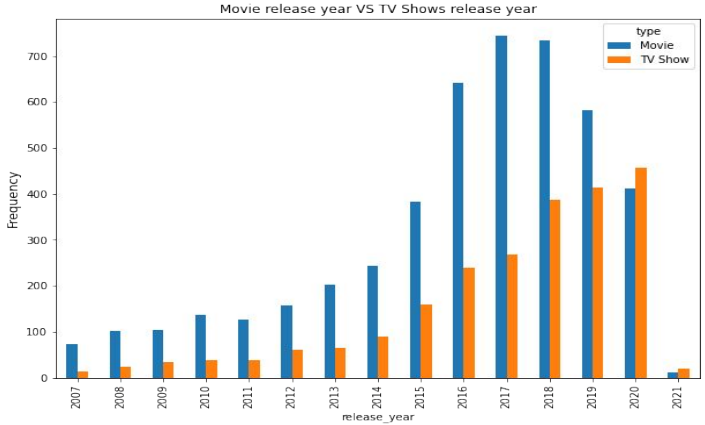
- **Feature engineering**
  - We add new column target_ages using rating column on the basis of which age group person are suitable for watch that movie.
  - We extract the month from the release date and add new column month.
  - Add duration column using min and season column.

EDA

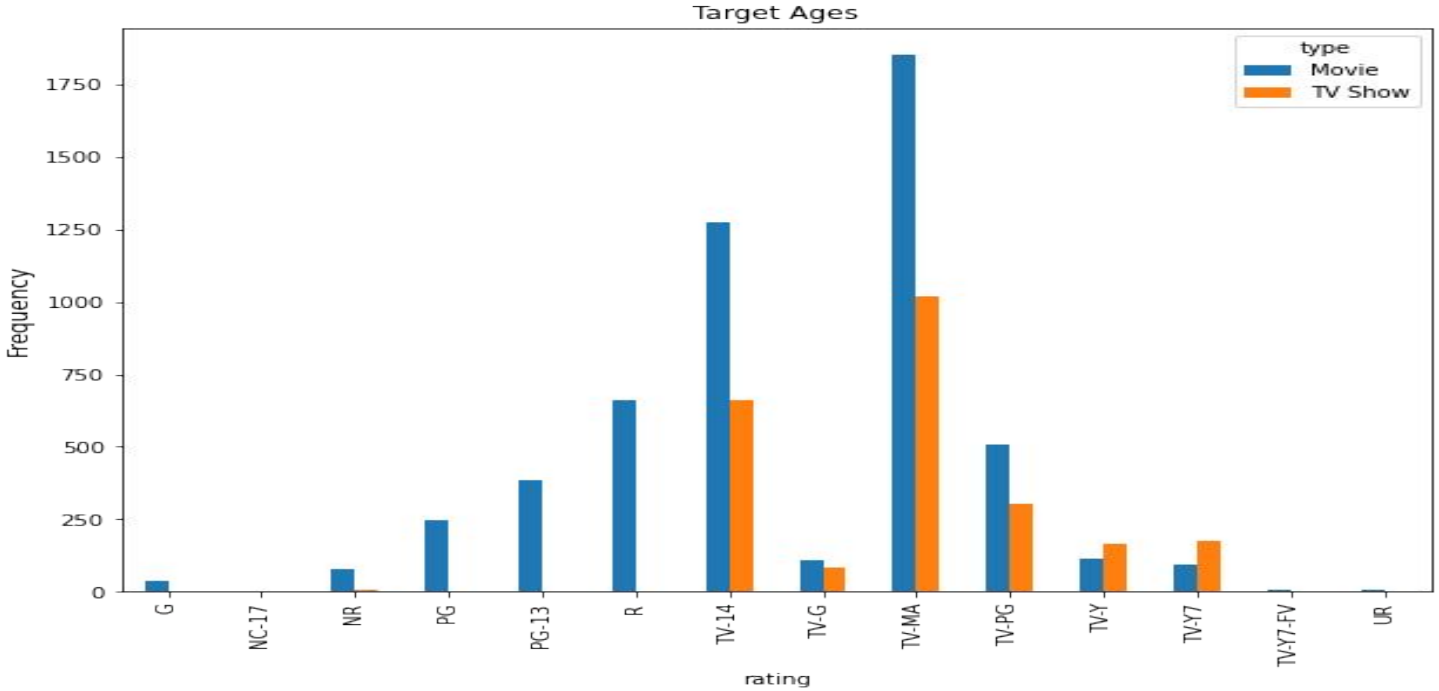
Type

- Netflix has approx 5200 movies and 2200 TV shows, Generally Movies are more released than TV shows.
- Frequency of Tv shows increases exponentially. So Netflix focus more in TV Shows rather than movies in recent years.
- In 2020, Lockdown time netflix releases more TV Shows than movies.



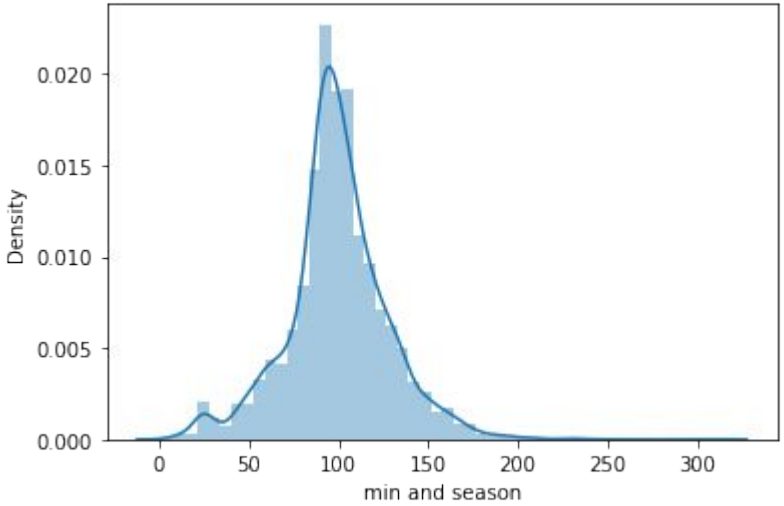


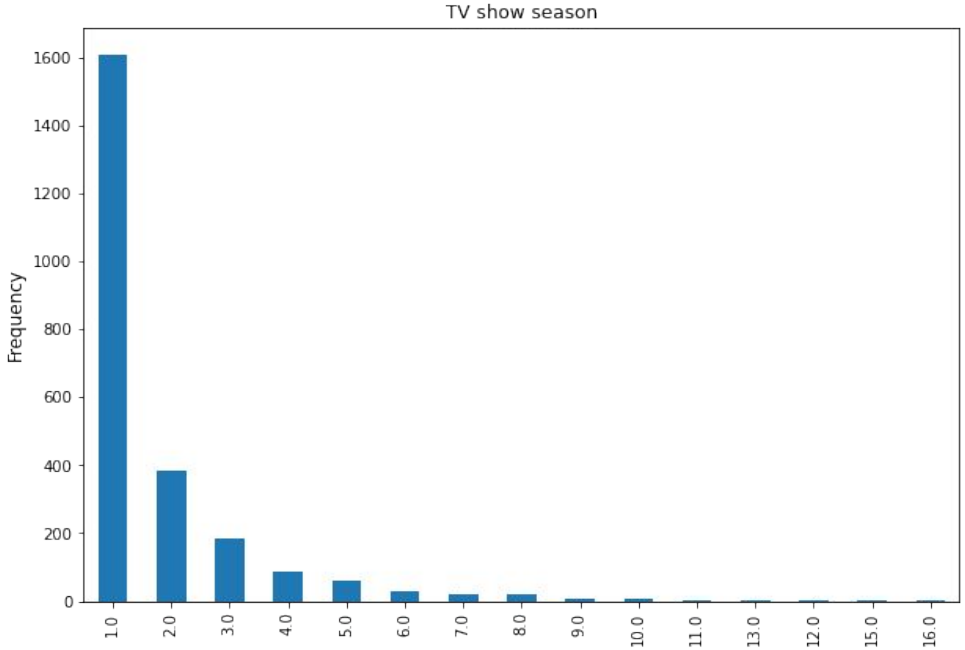
Type



- In both tv series and movies, TV-MA has the highest number of ratings i.e for mature audience.
- There is no any tv shows under rating PG(parental Guidance), PG_13 and R(Restricted under 17)
- In TV-Y(very young audience, age 2-6) and TV-Y7 rating there are more tv shows available than movies.

Type

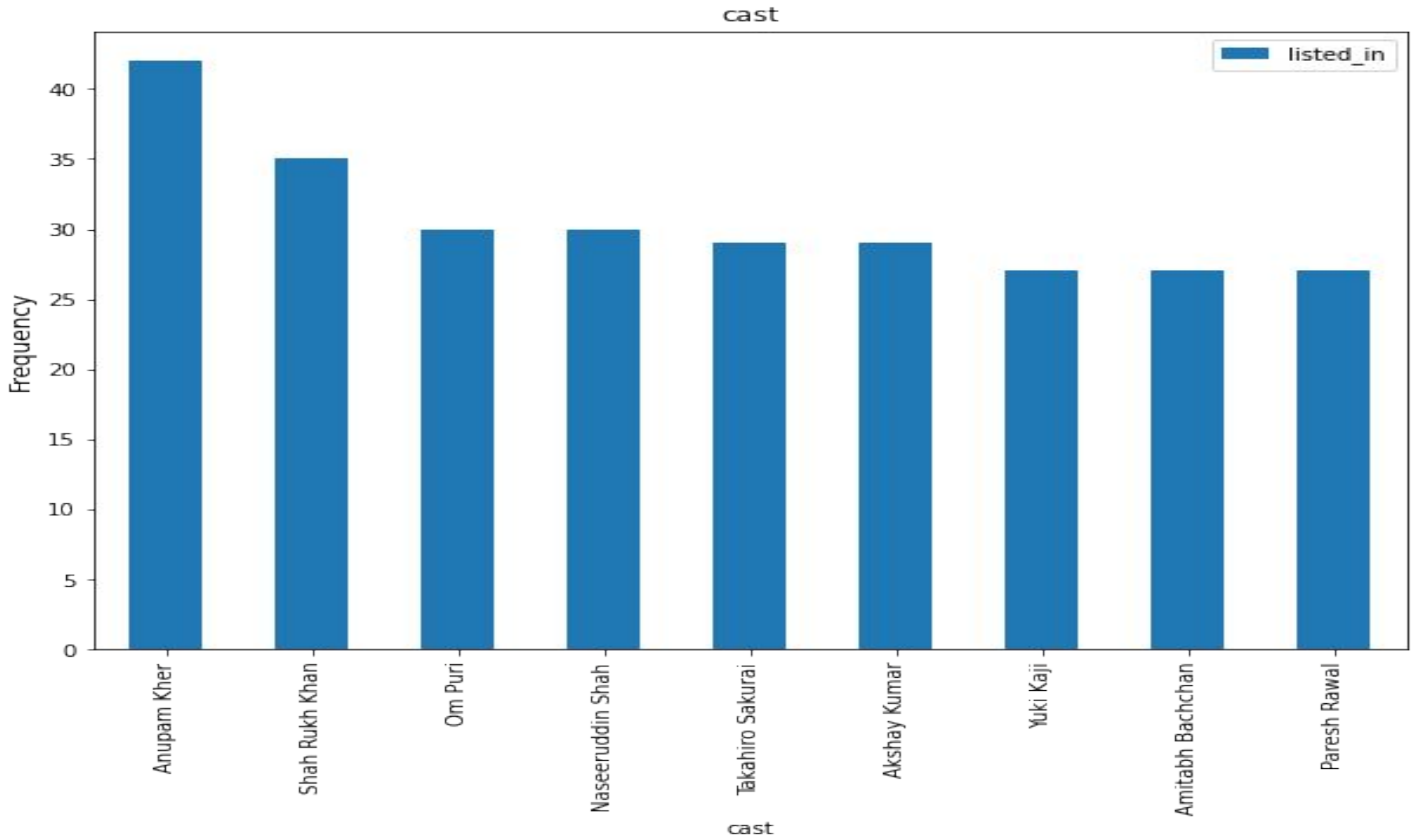




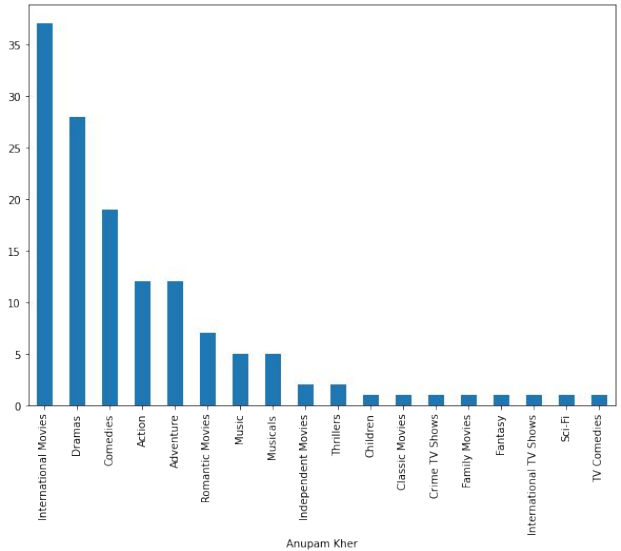
- Most of the movies have duration of between 50 to 150.
- Highest number of tv_shows consisting of single season.
- There are very few TV Shows with 7 seasons or more than 7 seasons.

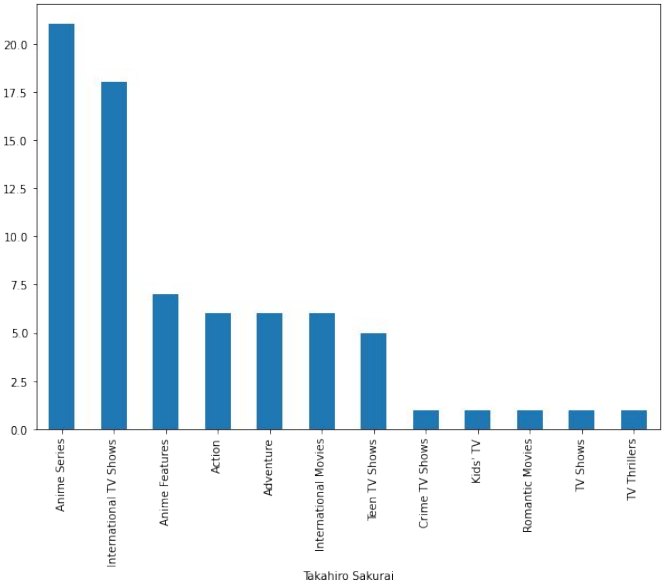
Cast

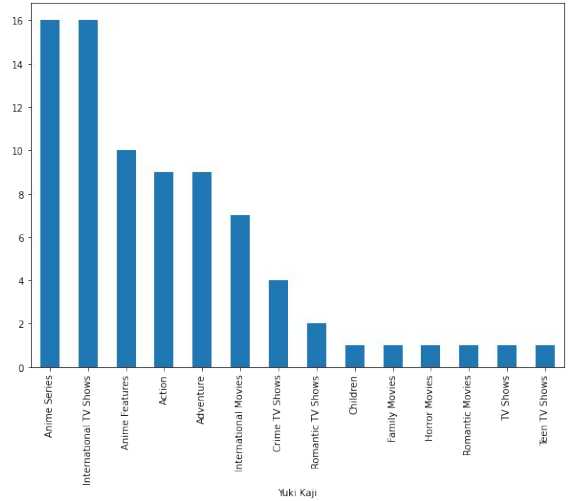
Top 10 Actors on Netflix



Cast Genre

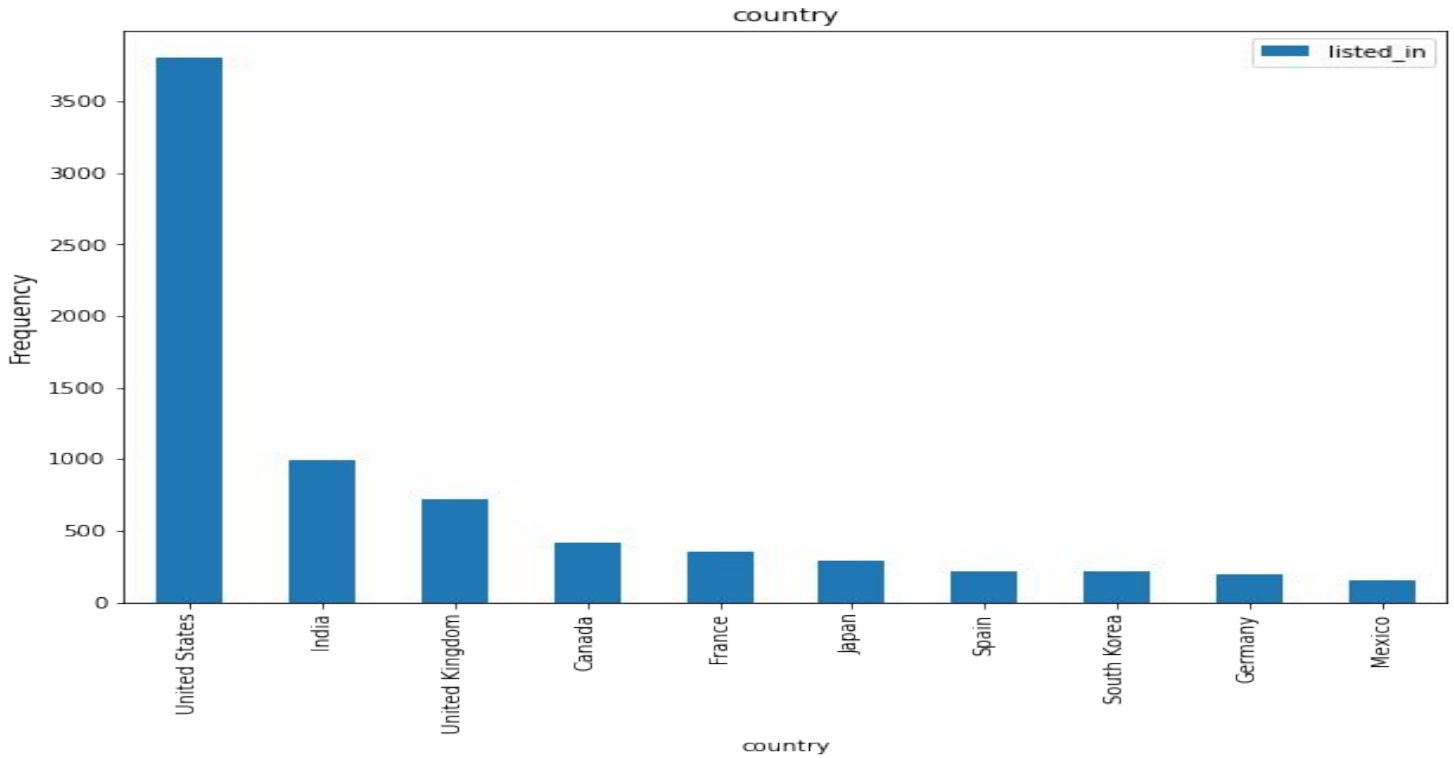






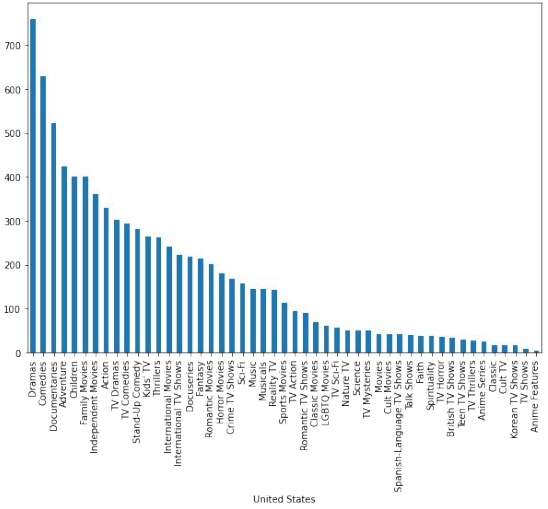
- Anupam Kher and Shah Rukh Khan are worked in most of the movie present our netflix dataset. One can also notice most of indian actor prefer dramas and comedies type of movie.
- Yuki kali and Takahrio sakurai are two actor other than indian in top 10.Both prefer Anime series or Anime movie.
- Very few indian actor work on any Tv shows and series, this means famous indian actor prefer movies more than tv series.

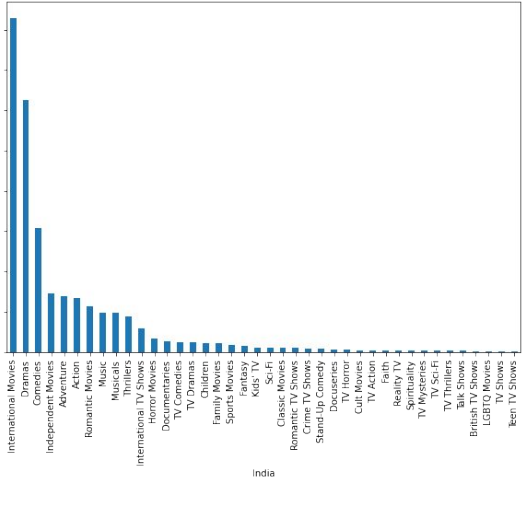
Country

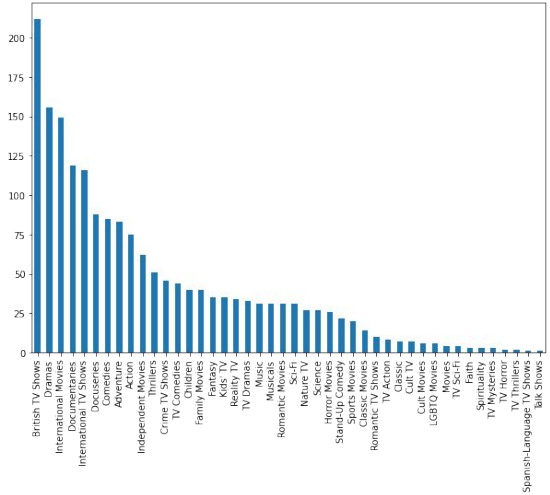


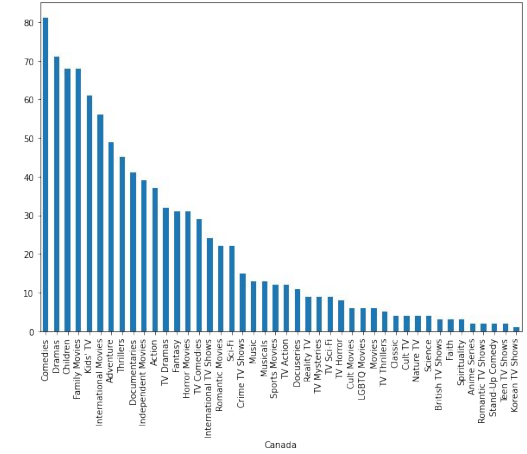
- United state has most number of content present in Netflix, followed by India.

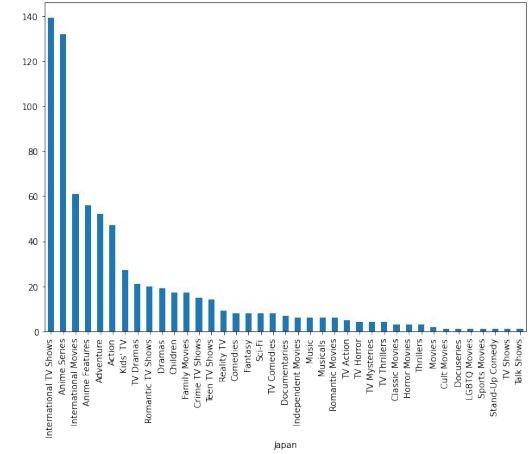
Country Genre

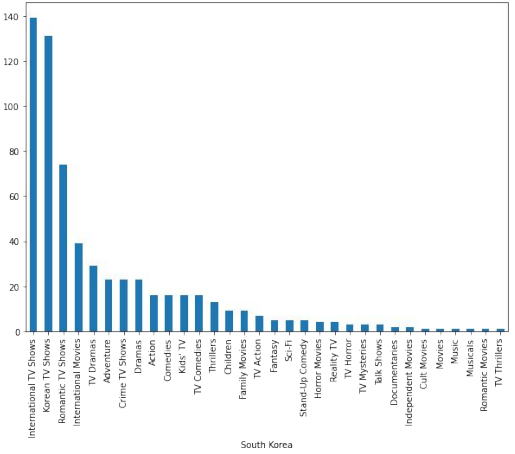










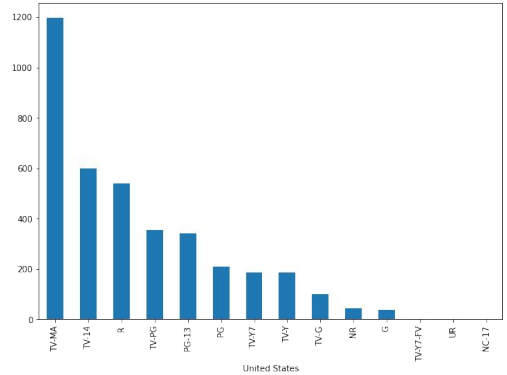


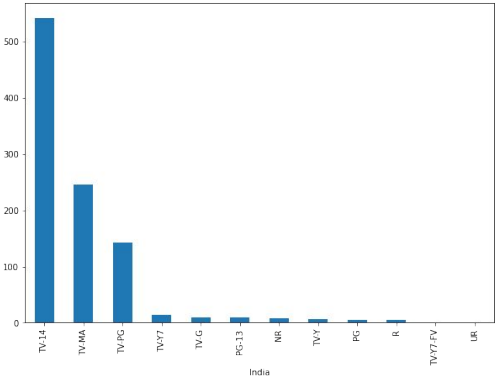
Country Genre

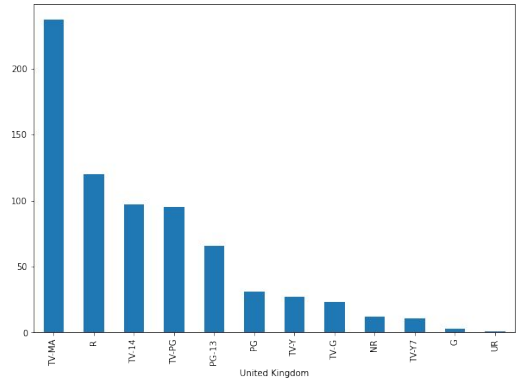
United states has the highest number of content on the netflix ,followed by india.

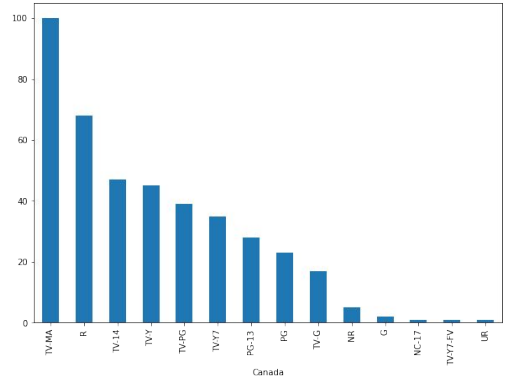
- In united state drama, comedies and documentaries movies are most popular.
- In India, France and Germany drama and comedies content are most popular.
- Children movie and kid's tv shows are most popular in canada, whereas british tv shows are most popular in UK.
- Anime series and feature are most popular in japan. Romantic and korean tv series are most popular in south korea.

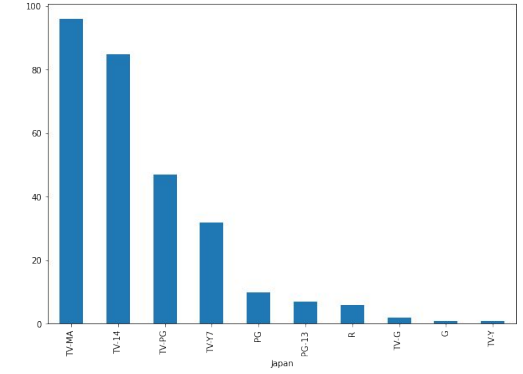
Country Rating

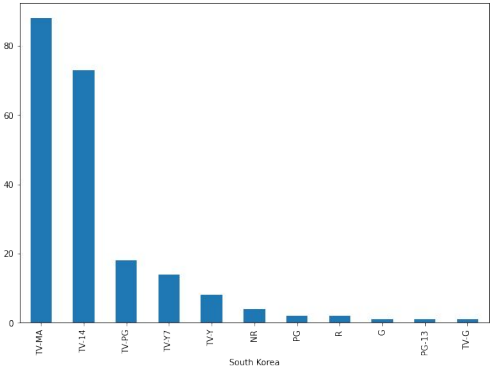






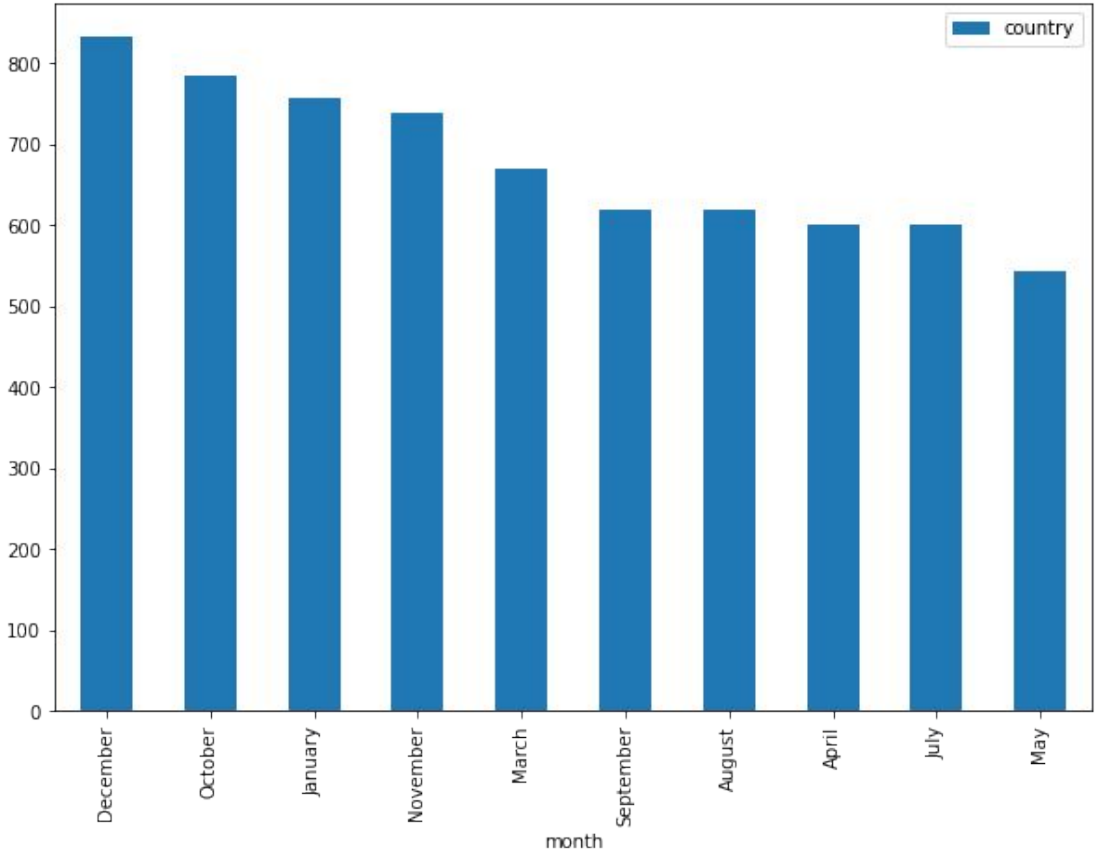






- In Asian countries mostly movies rating are under TV-MA and TV-14
- In Nato countries(European countries, America e.t.c) mostly movies rating are under TV-MA and R rating.

Month



- The most content is added to Netflix from october to january.

Feature Selection

- **Feature selection :** Here we are going to do Clustering similar content by matching text-based features so description column is one of the important feature.
- **Text Preprocessing :** First we convert all word into lower case then we remove the stop words and special character present in the sentence using nltk library and regular expression.
- **Lemmatization :** Lemmatization, unlike Stemming, reduces the inflected words properly ensuring that the root word belongs to the language. In Lemmatization root word is called Lemma. For example, runs, running, ran are all forms of the word run, therefore run is the lemma of all these words.
- **Tf - idf Vectorization :** we use TF-IDF is an abbreviation for Term Frequency Inverse Document Frequency. This is very common algorithm to transform text into a meaningful representation of numbers which is used to fit machine algorithm for prediction.

Topic Modelling (LDA)

Latent Dirichlet Allocation(LDA): In natural language processing, the Latent Dirichlet Allocation (LDA) is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar

Topic Modelling (LDA)

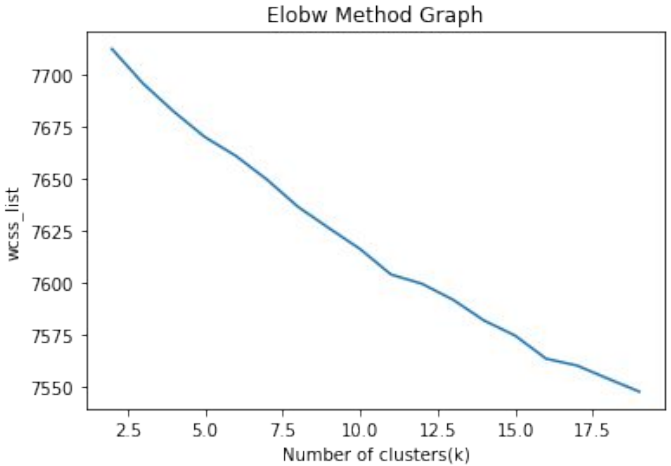
```
Netflix content: 0
  life, new, special, standup, take, series, love, find, band, comedian, family, one, comedy, friend, man
Netflix content: 1
  life, friend, family, world, new, young, school, find, two, father, teen, help, brother, best, man
Netflix content: 2
  life, young, woman, story, man, friend, crime, find, new, take, series, family, love, murder, two
Netflix content: 3
  life, find, world, family, man, young, two, love, one, girl, must, friend, woman, school, get
Netflix content: 4
  life, woman, friend, find, young, love, family, new, get, man, home, father, student, school, year
Netflix content: 5
  new, family, life, love, find, one, young, world, student, woman, series, group, man, documentary, school
Netflix content: 6
  young, life, christmas, get, man, find, woman, world, new, santa, love, family, becomes, mysterious, he
Netflix content: 7
  life, world, documentary, cop, family, woman, man, murder, series, police, crime, set, young, team, story
Netflix content: 8
  life, standup, comedian, love, take, special, documentary, high, woman, friend, new, find, young, family, stage
Netflix content: 9
  young, woman, family, friend, take, year, new, find, life, two, love, must, father, three, secret
```

ML algorithms(unsupervised)

- K-mean
- Agglomerative clustering

K-Means

- K-Means Clustering is an Unsupervised Learning algorithm which groups the unlabeled dataset into different clusters. Here K defines the number of predefined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.



- From elbow method generating 16 clusters

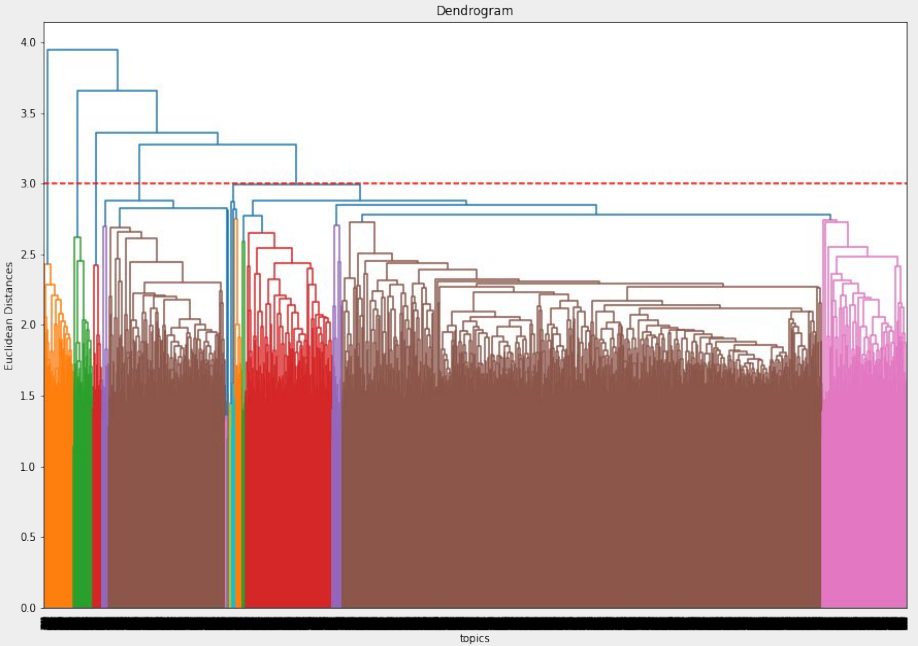
Evaluation

- **Silhouette Score :** It is a metric to evaluate the performance of clustering algorithm. It uses compactness of individual clusters(intra cluster distance) and separation amongst clusters (inter cluster distance) to measure an overall representative score of how well our clustering algorithm has performed.
- Silhouette score would always lie between -1 to 1. Our KNN model silhouette score is `0.0063684465319999405`

Agglomerative Clustering

- A dendrogram is a tree-like diagram that records the sequences of merges or splits.
- The optimal number of clusters can be obtained by the model itself by visualisation with dendrogram.We can set a threshold distance and draw a horizontal line.
- the optimal number of clusters is 5 using the Dendrogram
- Silhouette Coefficient: -0.000912

- K Means clustering perform better than hierarchical clustering Model on the basis of Silhouette score.



Conclusion

- The trend of Tv shows increasing rapidly from 2017, Even in 2021 netflix has more Tv shows than movies. Mostly Tv shows are of 2-3 seasons.
- In both tv series and movies, TV-MA has the highest number of ratings i.e for mature audience. There is no any tv shows under rating PG, PG_13 and R. In TV-Y and TV-Y7 rating there are more tv shows available than movies.
- In India, France and Germany drama and comedies content are most popular.
- Children movie and kid's tv shows are most popular in canada, whereas british tv shows are most popular in UK.
- Anime series and feature are most popular in japan. Romantic and korean tv series are most popular in south korea.
- The months of October, November, December and January had the largest number of films and Tv-shows released.

Conclusion

- In Asian countries mostly movies rating are under TV-MA and TV-14
- In Nato countries(European countries, America e.t.c) mostly movies rating are under TV-MA and R rating.
- Very few indian actor work on any Tv shows and series, this means famous indian actor prefer movies more than tv series
- Out of top 10 country content present in netflix only in Japan and south Korea TV Shows are more popular than movies.
- K means clustering perform better than hierarchical clustering. average Silhouette score for KNN is 0.0063 where as average Silhouette score for hierarchical clustering is -0.0009.from elbow method , for knn optimal of 16 clusters formed.
- We cut vertical lines with a horizontal line to obtain the number of clusters in Agglomerative Clustering. There were five clusters, with an average silhouette score of -0.0009.
- LDA has sorted much more similar title.

Thank you