Unsupervised Machine Learning Capstone Project

Netflix Movies and TV Shows Clustering

Project by:

JUNAID A R

Table of Content

Introduction

Netflix is an American company that revolutionized the watching of movies and television shows. Established in 1997 in California by Reed Hastings (the current CEO) and Marc Randolph. Netflix offers film and television series library through distribution deals as well as its own productions, known as **Netflix Originals**. In 2007 Netflix started subscription based streaming service that allows it's members to watch TV shows and movies on an internet connected device. We can also download TV shows and movies to our iOS, Android, or Windows 10 device and watch without an internet connection.

Their recommendation system make accurate suggestion of movies to the customers and the convenience of access and relatively low price brought Netflix dominated position in the market. They have over 8000 movies and tv shows available on their platform, and as of now, they have over 200M subscribers globally and in 2021 they made $5.1 billion of net profit.

Defining 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

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 similar content by matching text-based features

Dataset Summary

The data was collected from Flixable which is third party Netflix search engine. The dataset consists of movies and TV shows data till 2019. The dataset has 7787 observations and the dataset consists of eleven textual features and one numeric feature.

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

Dataset Summary

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 Release year 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 :** Genres
12. **description:** The Summary description

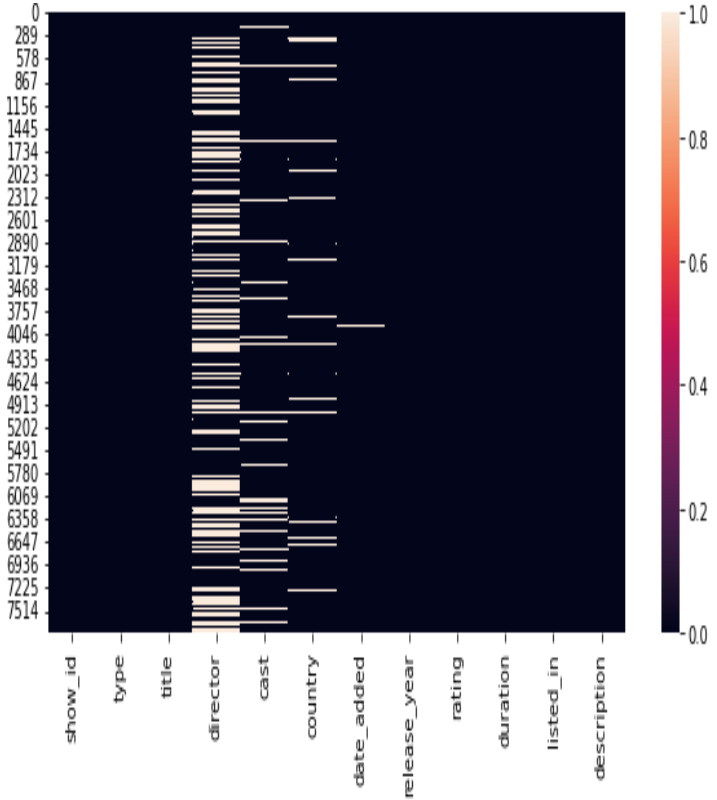
Data Cleaning

Null Value Treatment:

- **Director** feature has **30.67%** of null values.
- **Cast** feature has **9%** of null values.

So we have dropped these feature as they were irrelevant and were containing null values.

- **Country** feature have **6%** of null values so we have replaced null values with mode(United States).
- **date_added and rating** features have less than **1%** of null values so we have simply dropped those attributes.



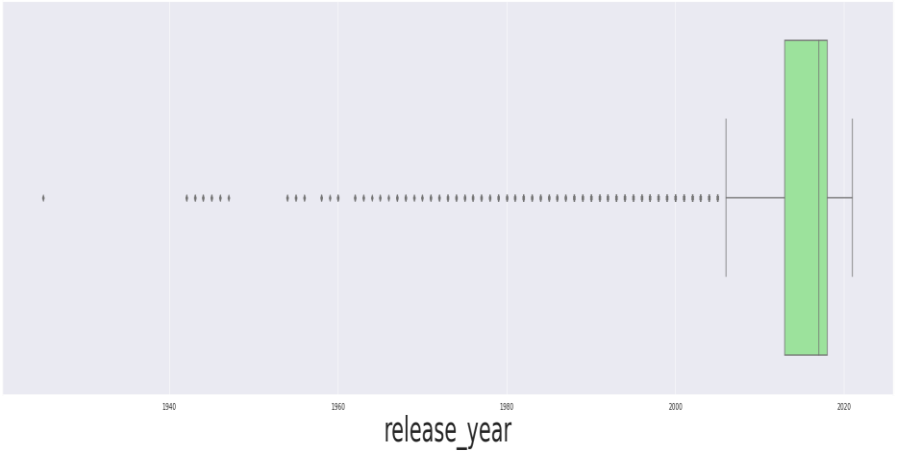
Data Cleaning

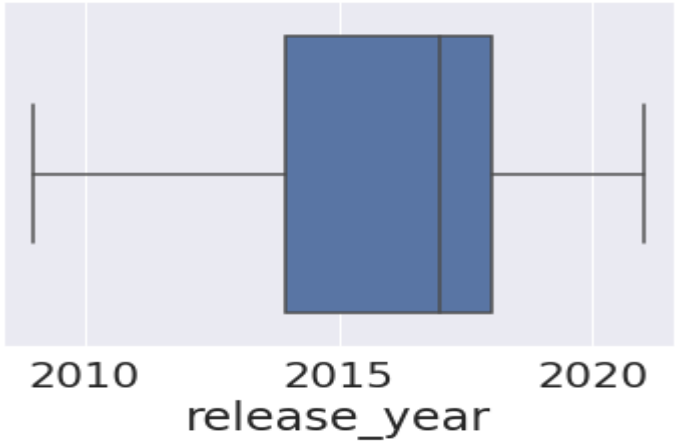
Duplicate Value:

- We didn't found any duplicate attributes in our dataset.

Handling Outliers:

- We found outliers in our numerical feature(release_year) so we replaced the outliers with mean value using quantiles.

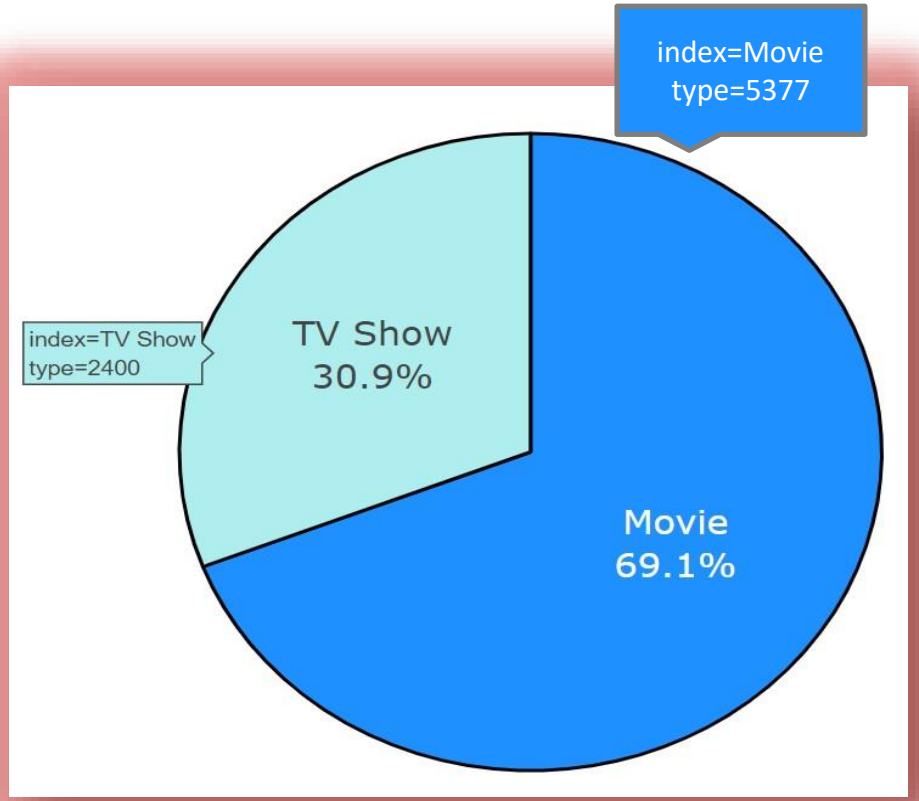




Exploratory Data Analysis

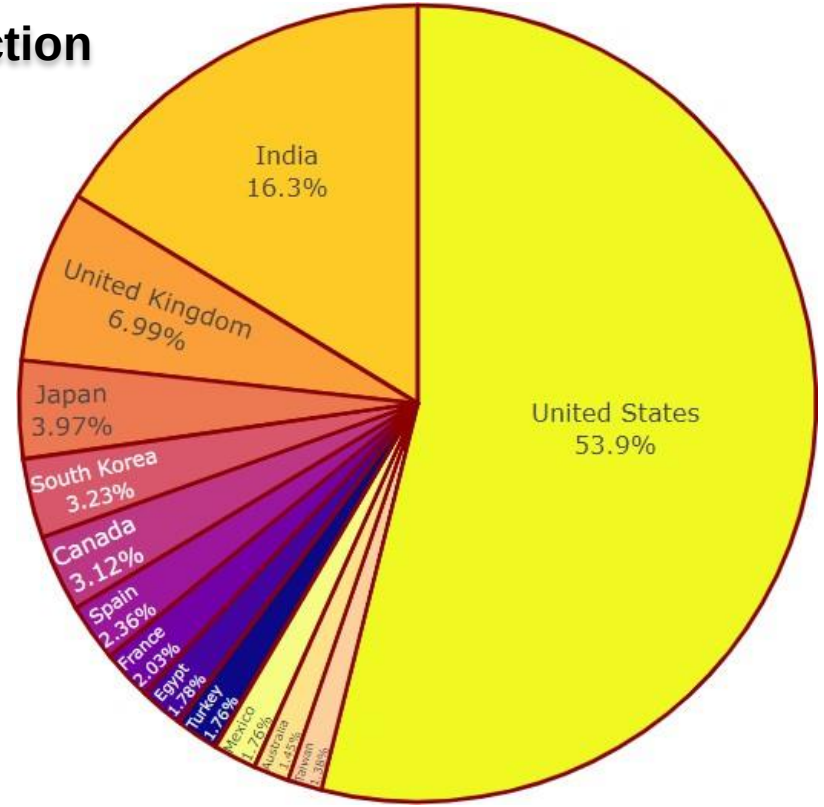
Type of content available on Netflix

- It is evident that there are more movies on Netflix than TV shows.
- Netflix has 5377 movies, which is more than double the quantity of TV shows.



Exploratory Data Analysis

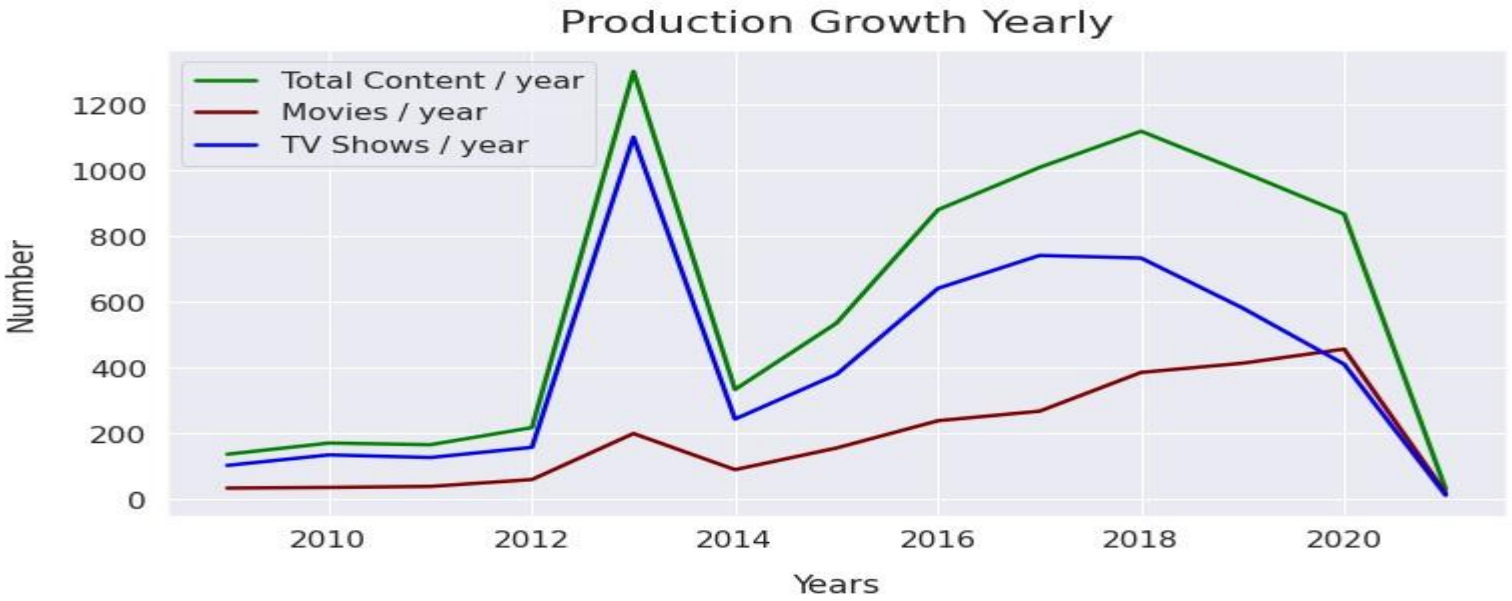
Top countries with highest content production



- United States has the most number of content on Netflix
- India has second highest content on Netflix
- Australia and Taiwan has least number of content on Netflix

Exploratory Data Analysis

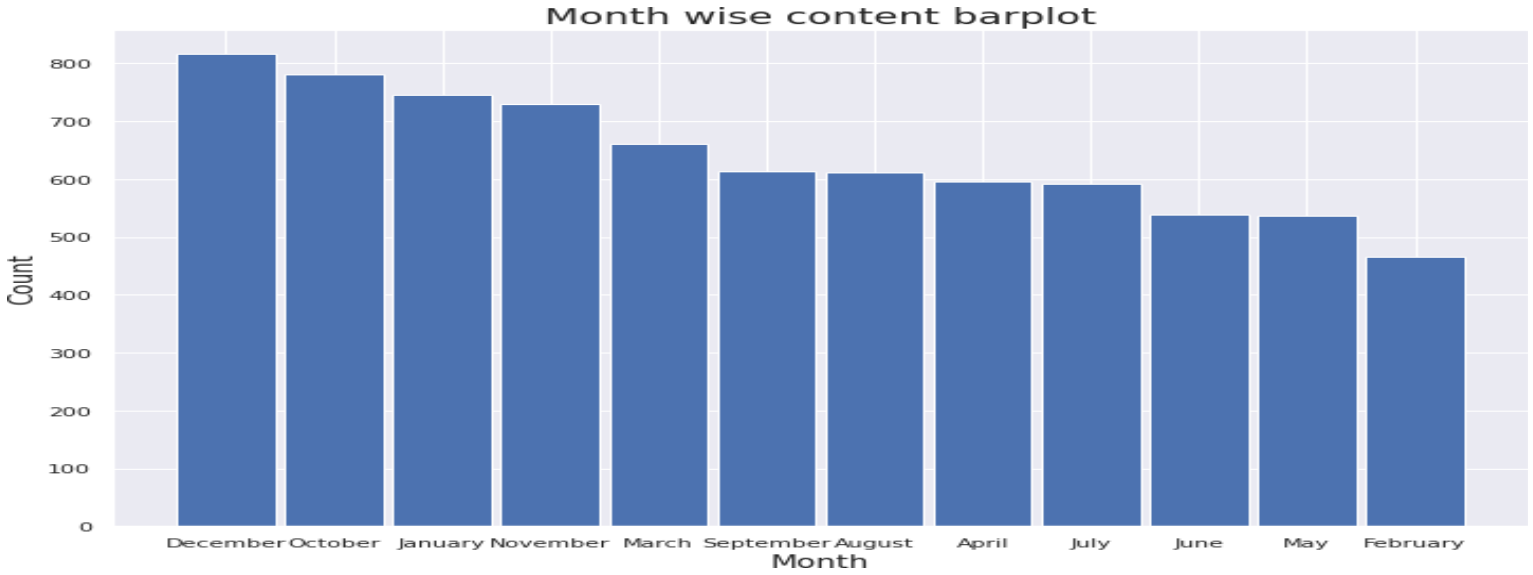
Movies and TV Shows released over the years in Netflix



- The number of release have significantly increased after 2015 and have dropped in 2021 because of Covid 19.

Exploratory Data Analysis

Movies and TV Shows released over the months in Netflix



- The number of release have significantly increased during winter season due to festivals like Christmas.

Applying WordCloud on Titles of the Movies and TV Shows

Most occurred words present in Title are:-

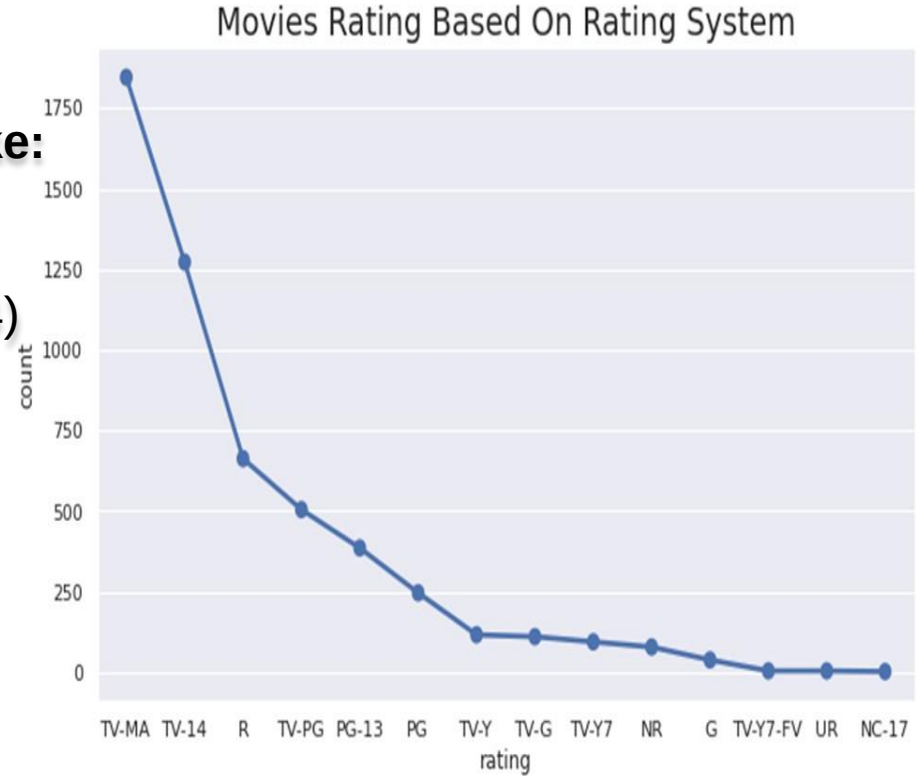
- Love
- Christmas
- Man
- World
- Story
- Girl
- Day

- It is surprising to see that "Christmas" occurred so many time and the reason may be those movies released on the month of December.

Exploratory Data Analysis

Most of the Movies content got ratings like:

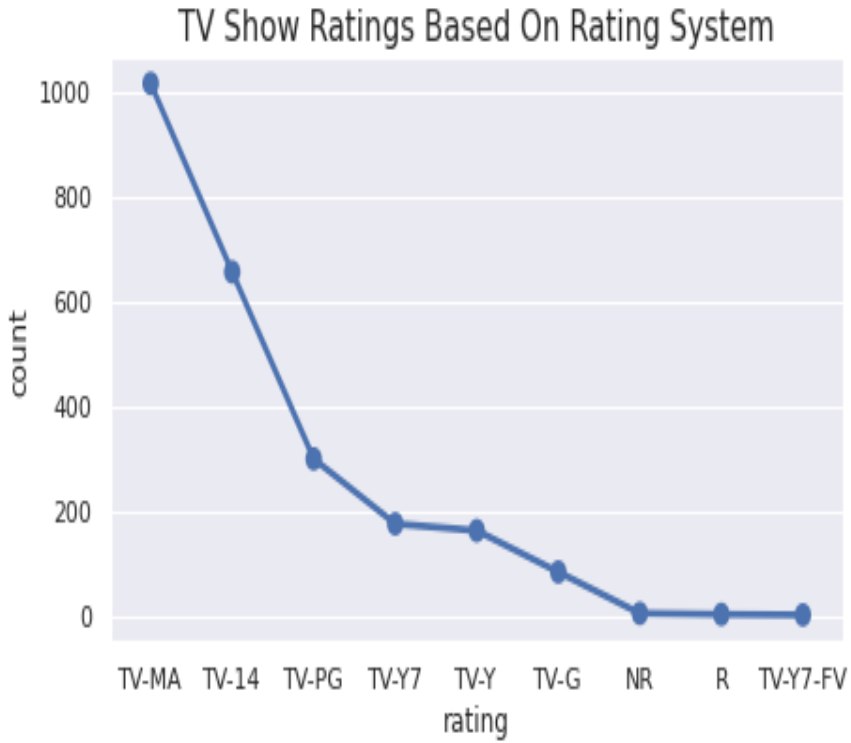
- TV-MA (For Mature Audiences)
- TV-14 (Unsuitable for children under age 14)
- R (Restricted for child)
- TV-PG (Parental Guidance Suggested)



Exploratory Data Analysis

Most of the TV Shows contents got ratings like:

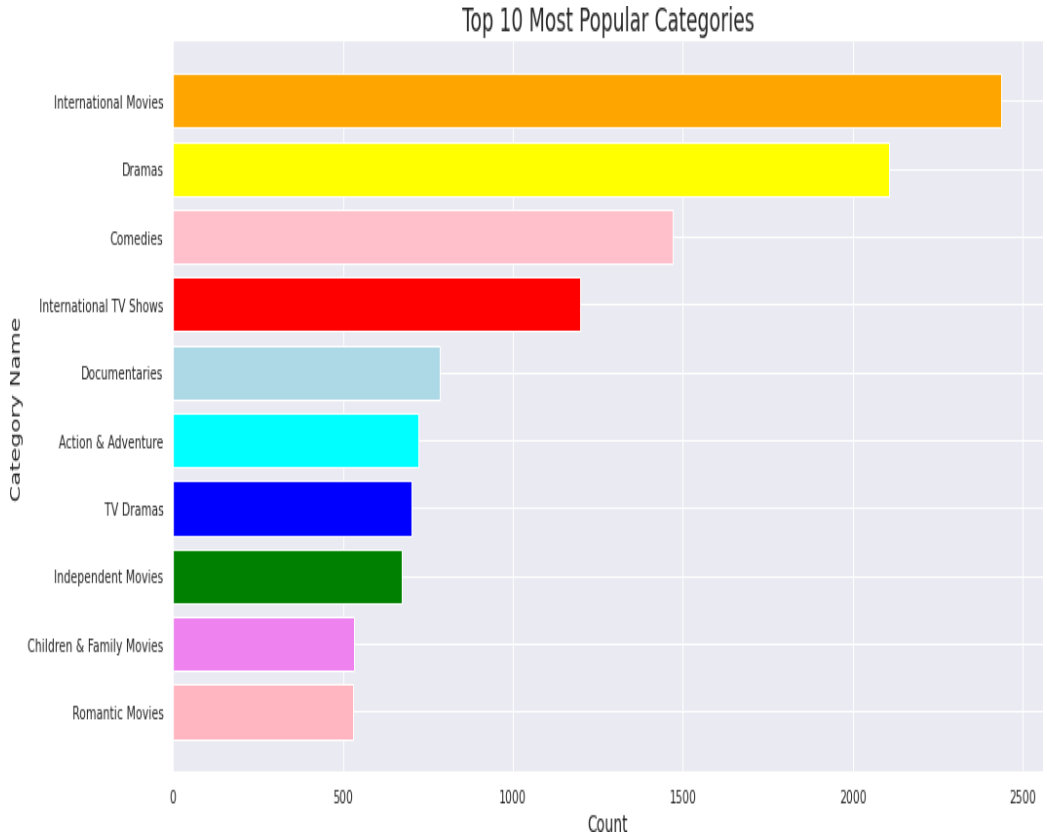
- TV-MA (For Mature Audiences)
- TV-14 (Unsuitable for children under age 14)
- TV-PG (Parental Guidance Suggested



Exploratory Data Analysis

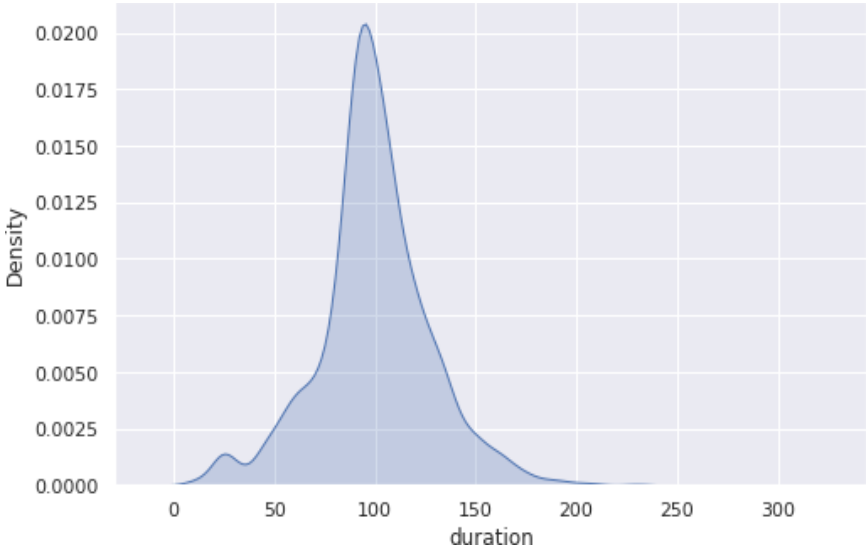
Most popular categories are:

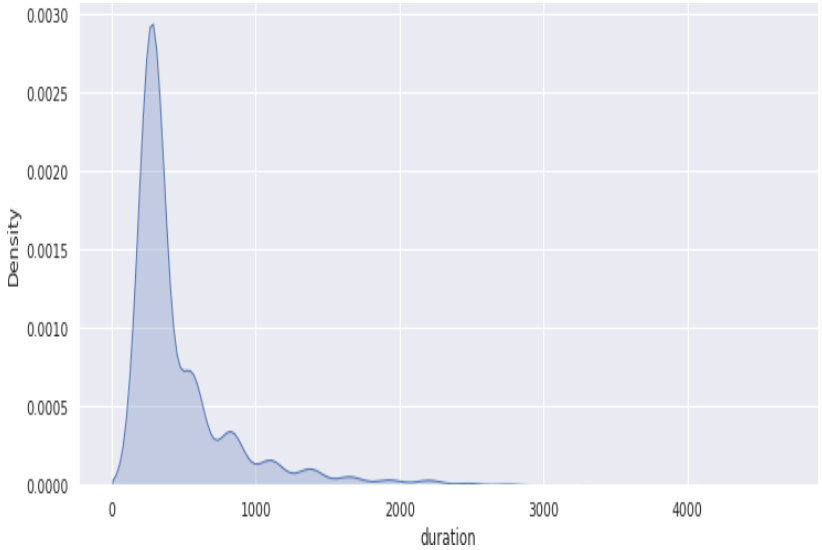
- ❑ International movies
- ❑ Dramas
- ❑ Comedies
- ❑ International TV Shows
- ❑ Documentaries



Exploratory Data Analysis

Duration Distribution of Movies and TV Shows



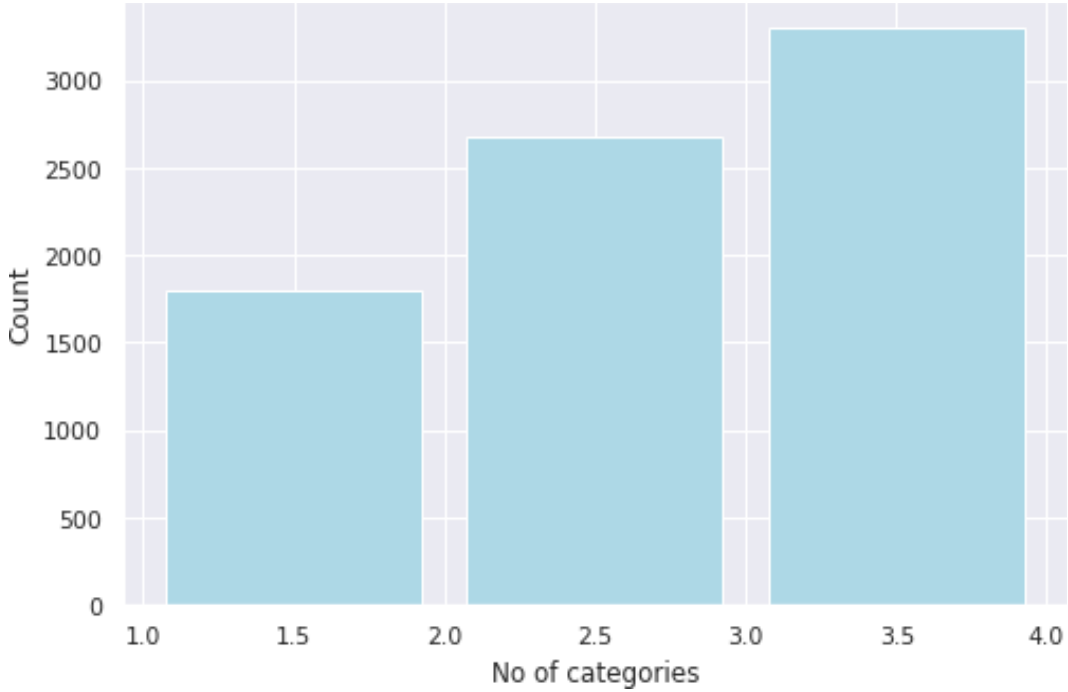


- Most of the Movies content are about 70 to 120 min duration.
- Most of the TV Shows content are around 300 mins duration.

Exploratory Data Analysis

Average number of categories are present there in each content

Most of the movies or TV Shows contains 3-4 categories of genre.



Data Pre-Processing

- **Label Encoding-** We used Label Encoding to converting the labels into a numeric form so as to convert them into the machine-readable form. Machine learning algorithms can then decide in a better way how those labels must be operated.
- **Lemmatisation-** We used Lemmatization, unlike Stemming, to 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.
- **Removing Stop words-** To remove stop words from a sentence, we can divide our text into words and then remove the word if it exits in the list of stop words provided by NLTK.
- **Tf - idf Vectorization-** TF-IDF stands for "Term Frequency — Inverse Document Frequency". This technique helped us to quantify a word in documents, we generally compute a weight to each word which signifies the importance of the word in the document and corpus. This method is a widely used technique in Information Retrieval and Text Mining.
- **Min-max Scaling-** We used MinMaxScaler to scale numerical variables within a given range of 0 to 1 MinMaxScaler subtracts the minimum value in the feature and then divides by the range. It preserves shape of original distribution.

Feature Selection

We only selected 3 features to do clustering

- ❑ no_of_category feature
- ❑ Length(listed-in) feature
- ❑ Length(description) feature

We used 5 algorithms to find out best k value

- ❑ Silhouette Score
- ❑ Elbow Method
- ❑ DBSCAN
- ❑ Dendrogram
- ❑ Agglomerative Clustering

Model Implementation

Silhouette Score

The Silhouette Coefficient Formula:

(b - a) / max(a, b)

- mean intra-cluster distance(a):- Mean distance between the observation and all other data points in the same cluster.
- mean nearest-cluster distance (b) :- Mean distance between the observation and all other data points of the next nearest cluster. This distance can also be called a.

The value of the silhouette coefficient is between [-1, 1]

- If score is 1 denotes the best meaning that the data point is very compact within the cluster to which it belongs and far away from the other clusters.
- The worst value is -1
- If score is 0 denotes overlapping clusters

For n_clusters = 1, score is 0.42843328899854627

For n_clusters = 2, score is 0.3832991558393045

For n_clusters = 3, score is 0.37431547662296216

For n_clusters = 4, score is 0.3720843918336816

For n_clusters = 5, score is 0.36819494916930934

For n_clusters = 6, score is 0.3759188611985892

For n_clusters = 7, score is 0.37050871865079854

For n_clusters = 8, score is 0.37378435200824456

For n_clusters = 9, score is 0.3646974512294472

For n_clusters = 10, score is 0.3552589885907151

For n_clusters = 11, score is 0.34797257402504966

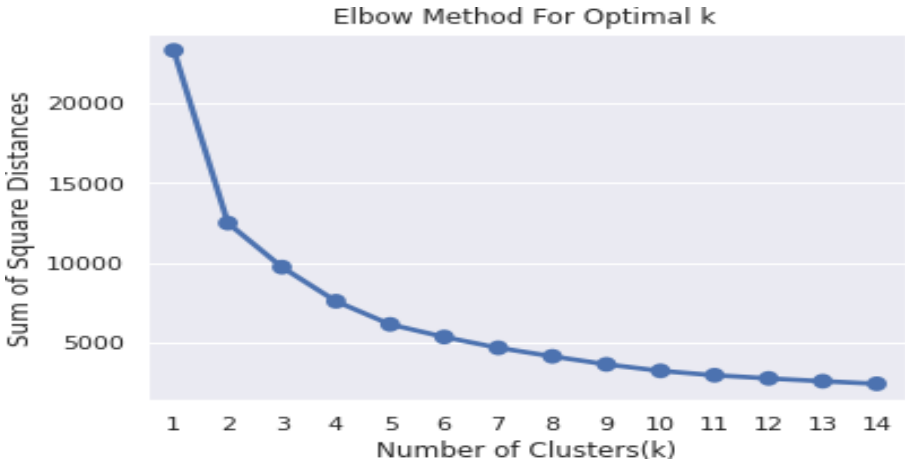
For n_clusters = 12, score is 0.3503371354182111

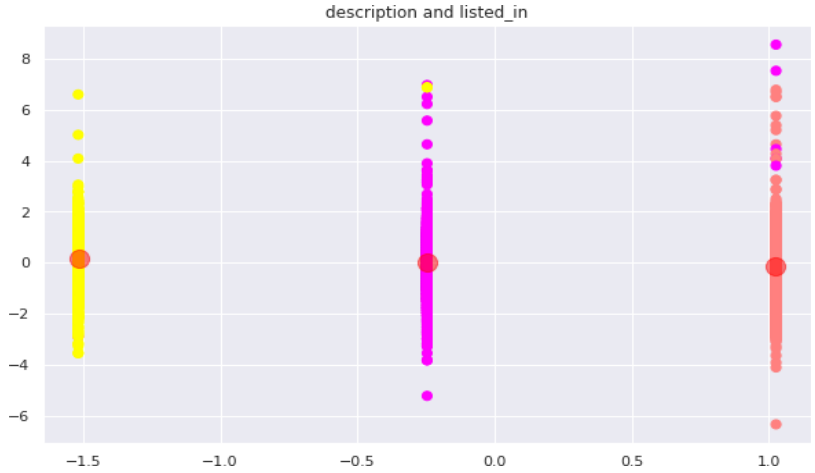
For n_clusters = 13, score is 0.3383078147446296

For n_clusters = 14, score is 0.34180346139272216

Model Implementation

Elbow Method



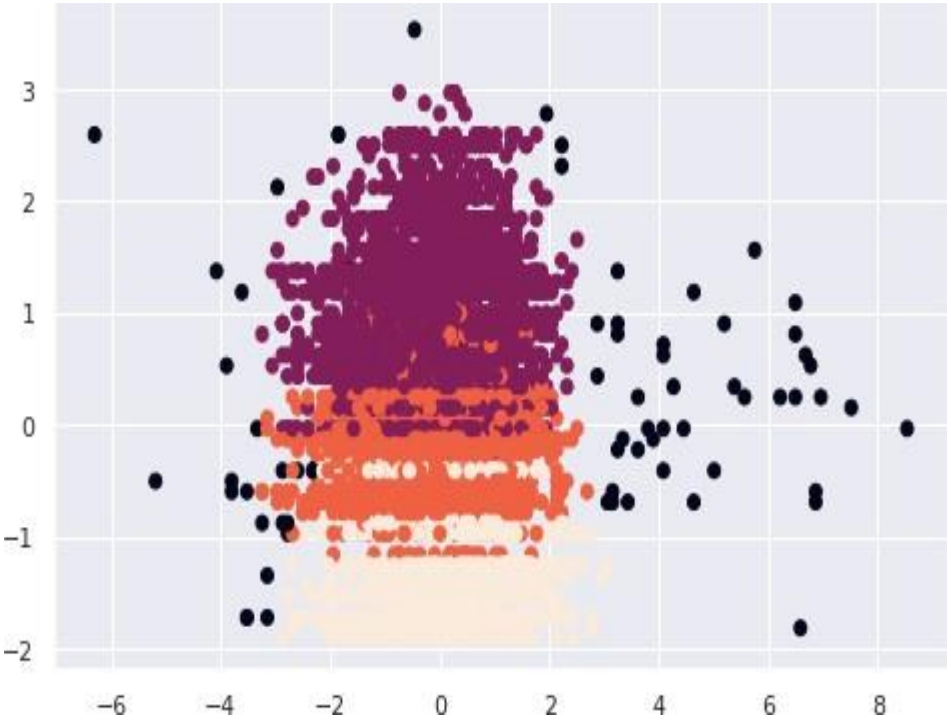


- The elbow method runs k-means clustering on the dataset for a range of values for k (say from 1-15) and then for each value of k computes WCSS value. By default, the distortion score is computed, the sum of square distances from each point to its assigned center.
- We will be using 3 clusters.

Model Implementation

DBSCAN

- The main idea behind DBSCAN is that a point belongs to a cluster if it is close to many points from that cluster.
- If it is away from many data points like the black colour dots in the diagram they are called noise(Outliers).

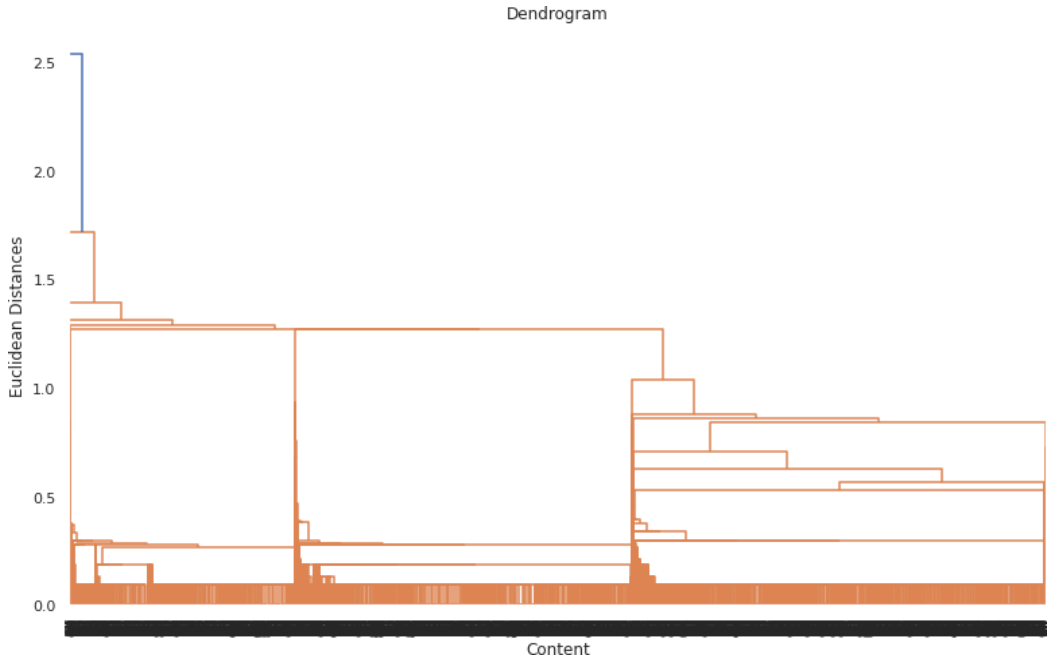


Model Implementation

Dendrogram

The number of clusters will be the number of vertical lines which are being intersected by the line drawn using the threshold.

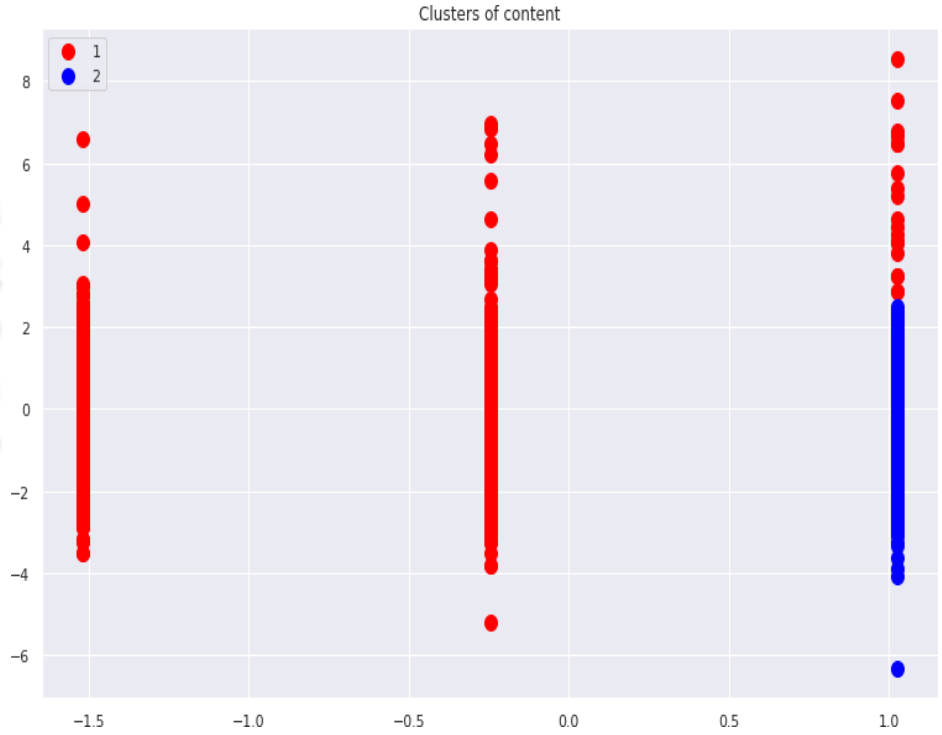
No. of Cluster = 3



Model Implementation

Agglomerative Clustering

In this algorithm, clustering begins with N groups, each containing initially one entity, and then the two most similar groups merge at each stage until there is a single group containing all the data.



Conclusion

- Director and cast attribute contains a large number of null values so we have simply dropped these 2 columns .
- In this dataset there are two types of contents available where 30.86% includes TV shows and the remaining 69.14% includes Movies.
- Maximum number of contents are released in winter season and that might be because of festival season like Christmas and due to which “Christmas” word is more occurred in the title of the content.
- From the dataset insights we can conclude that the most number of TV Shows relea sed in 2017-18 and for Movies it is 2019-20.
- On Netflix, USA has the largest number of contents and mostly countries preferred to produce Movies more than TV shows.
- Most of the movies and TV Shows contain 3 genre categories.

Conclusion

- TOP 3 content categories are International movies , dramas , comedies.
- In text analysis (NLP) I used stop words, removed punctuations , stemming & TF-IDF vectorizer and other functions of NLP.
- Applied different clustering models like Silhouette Score, Elbow Method, DBSCAN, Dendrogram and Agglomerative Clustering on data we got the best cluster arrangements.
- By applying different clustering algorithms to our dataset, we get the optimal number of cluster is equal to 3.

Thank you