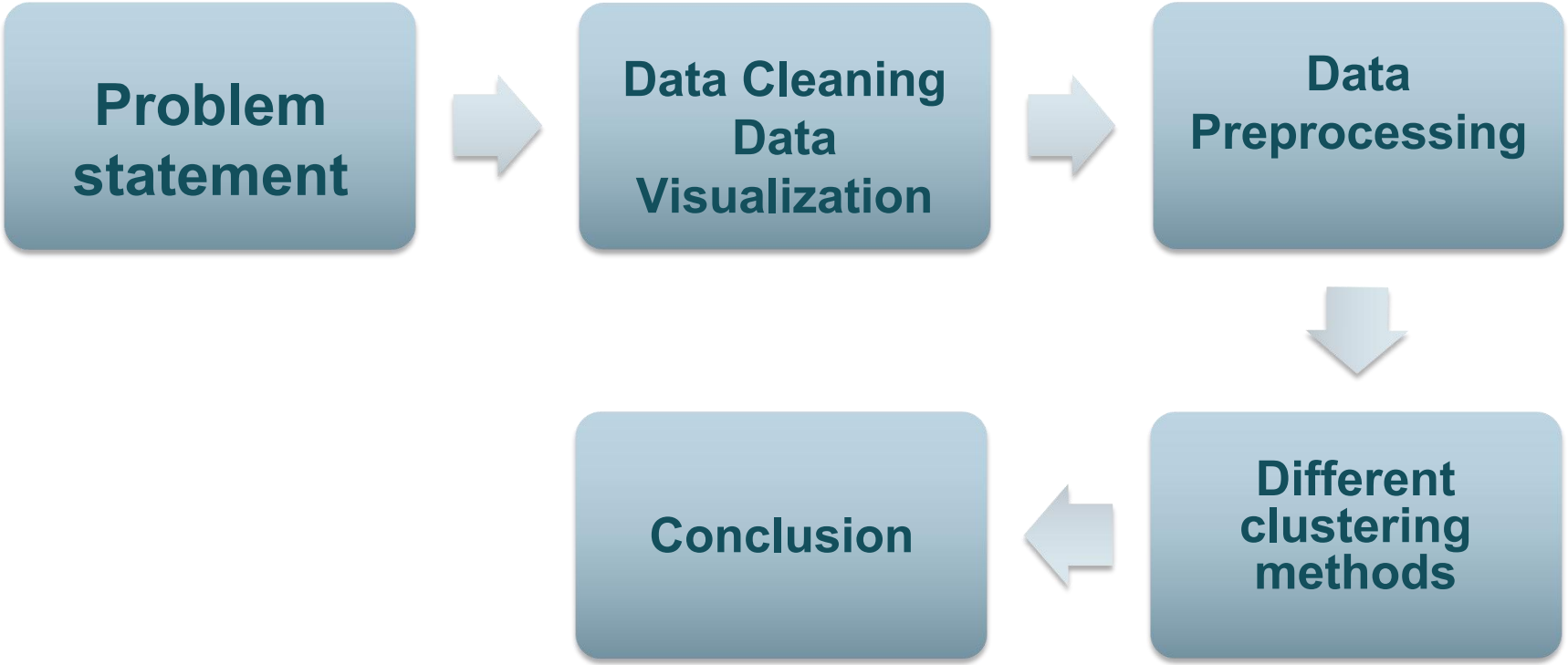
Capstone Project

Netflix Movies and TV Shows Clustering

Unsupervised Machine Learning

Individual Project:
Priyanka Shinde

Contents


Problem statement
Data Cleaning Data Visualization
Data Preprocessing
Different clustering methods
Conclusion

Problem Statement

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

Description of Netflix Dataset

This dataset contains data collected from Netflix of different TV shows and movies from the year 2008 to 2021.

type: Gives information about 2 different unique values one is TV Show and another is Movie

title: Gives information about the title of Movie or TV Show

director: Gives information about the director who directed the Movie or TV Show

cast: Gives information about the cast who plays role in Movie or TV Show

release_year: Gives information about the year when Movie or TV Show was released.

date_added : Date it was added on Netflix

rating: Gives information about the Movie or TV Show are in which category (eg like the movies are only for students, or adults, etc)

duration: Gives information about the duration of Movie or TV Show

listed_in: Gives information about the genre of Movie or TV Show

description: Gives information about the description of Movie or TV Show

country : Country where the movie / show was produced

Data Exploration

- **The dataset consists of eleven textual columns and one numeric column(“release_year”)**

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7787 entries, 0 to 7786
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   show_id       7787 non-null   object
 1   type          7787 non-null   object
 2   title         7787 non-null   object
 3   director      5398 non-null   object
 4   cast          7069 non-null   object
 5   country       7280 non-null   object
 6   date_added    7777 non-null   object
 7   release_year  7787 non-null   int64
 8   rating        7780 non-null   object
 9   duration      7787 non-null   object
 10  listed_in     7787 non-null   object
 11  description   7787 non-null   object
dtypes: int64(1), object(11)
memory usage: 730.2+ KB
```

- **shape of dataset**

```
Number of Rows: 7787
Number of Columns: 12
```

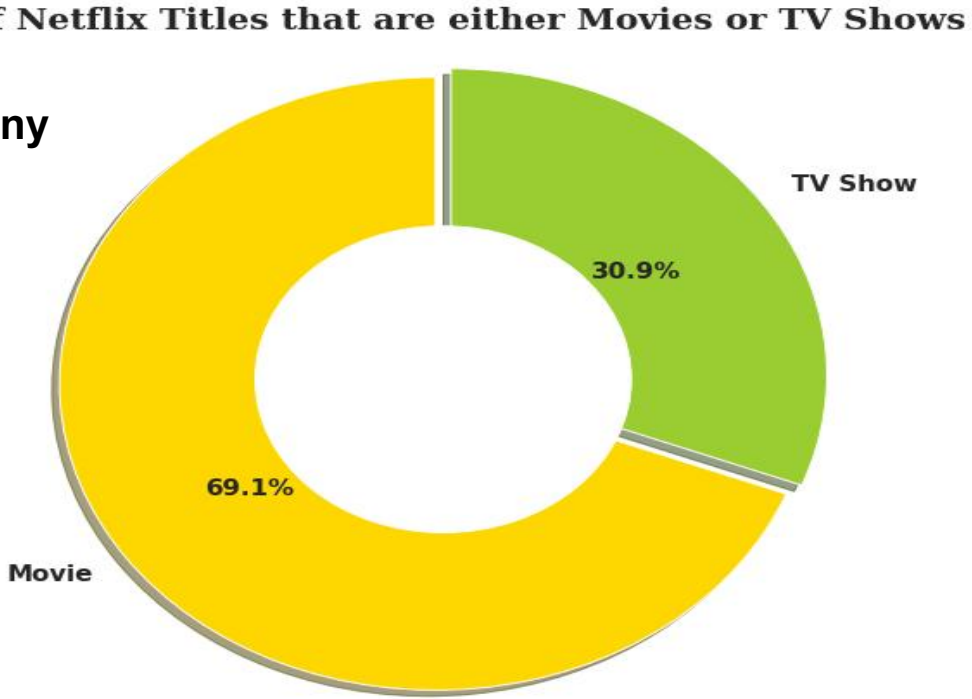
Data Cleaning

- No duplicate values present in this dataset.
- Removing unnecessary row of column 'cast'
- Replace NaN values with “no director”
- For country column replace NaN values with United State.

	No Of Total Values	'Missing values count	%age of NaN values
director	7787	2389	30.68
cast	7787	718	9.22
country	7787	507	6.51
date_added	7787	10	0.13
rating	7787	7	0.09
show_id	7787	0	0.00
type	7787	0	0.00
title	7787	0	0.00
release_year	7787	0	0.00
duration	7787	0	0.00
listed_in	7787	0	0.00
description	7787	0	0.00

Movies Or TV-Shows??

- **Number of Movie is about twice as many as that of TV show for this dataset.**

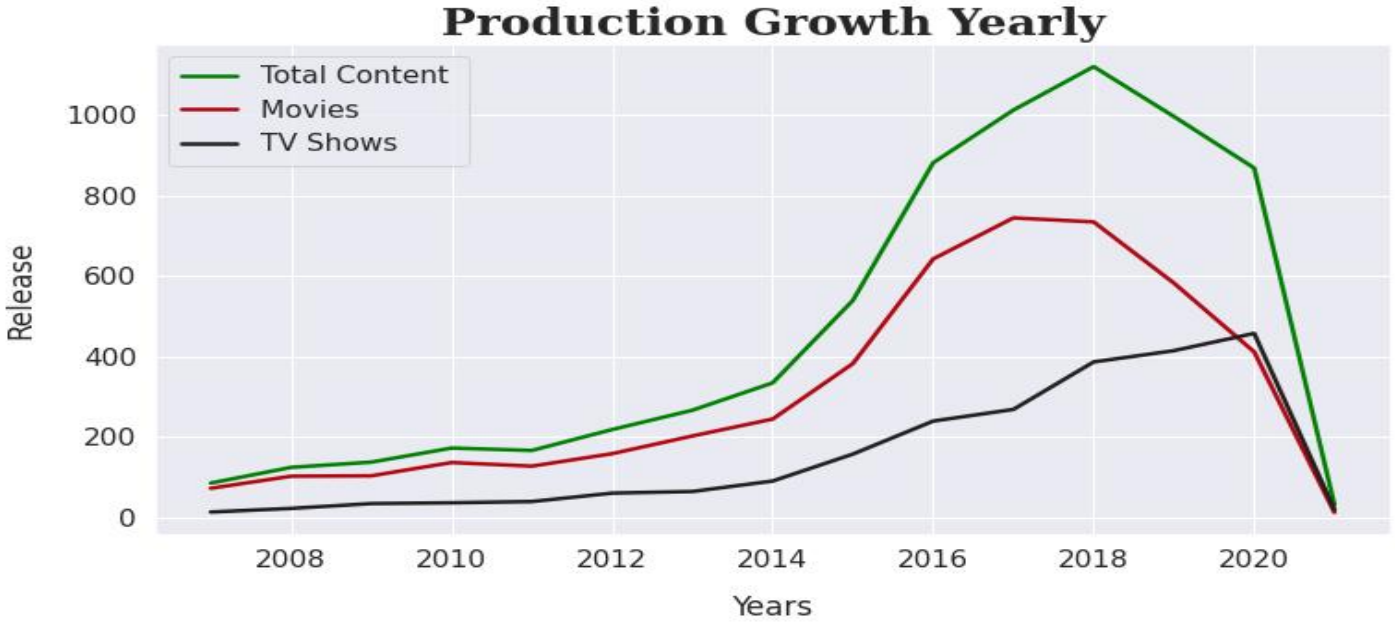


Popular Title words

Most occurred words present in Title are:-

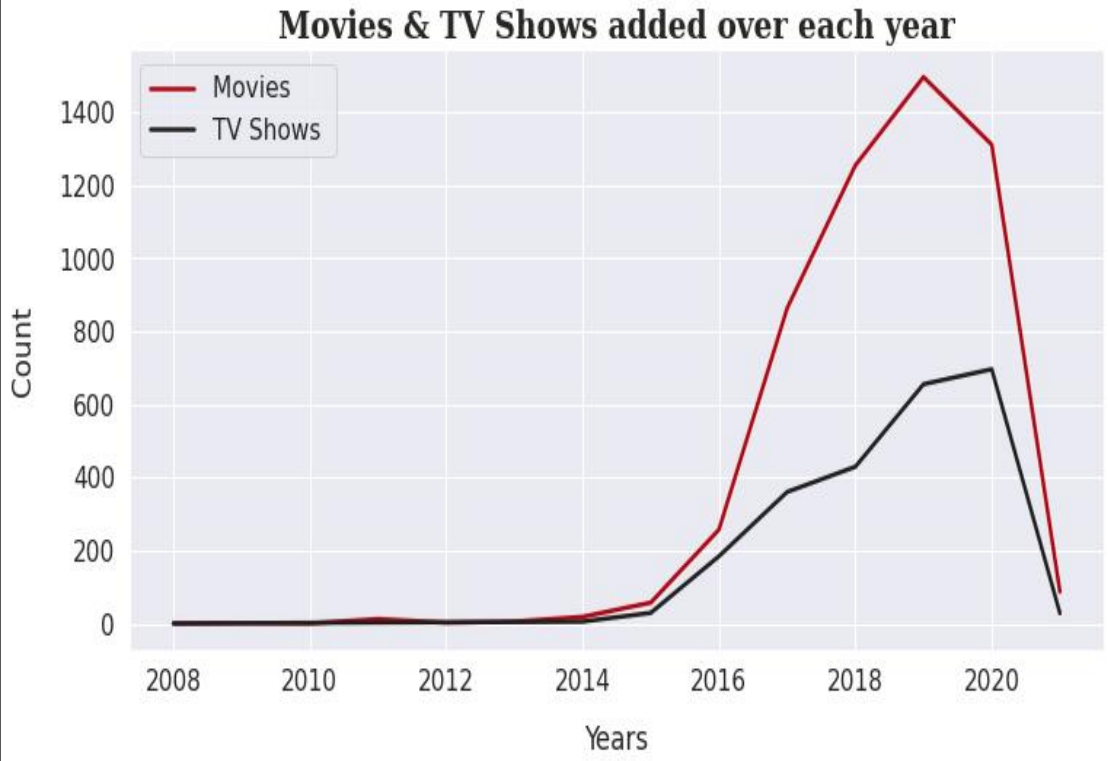
- Christmas
- Man
- World
- Story
- Love
- Girl
- Day

Netflix Content Release Over Yearly



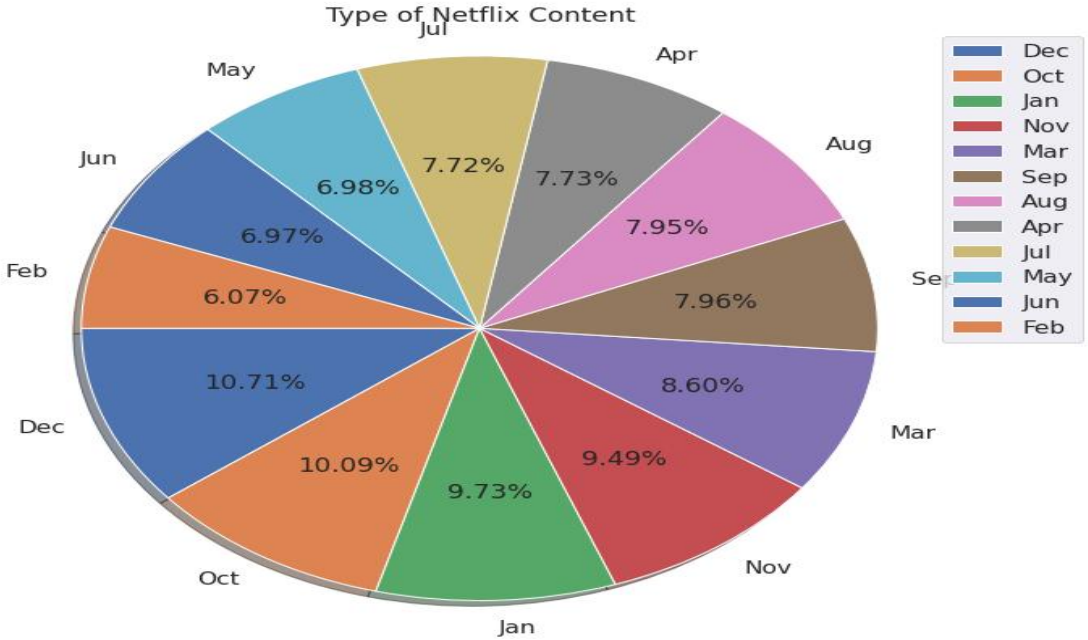
- The most number of TV Shows released in 2020 and for Movies it is 2017

Netflix Content Added Over Yearly



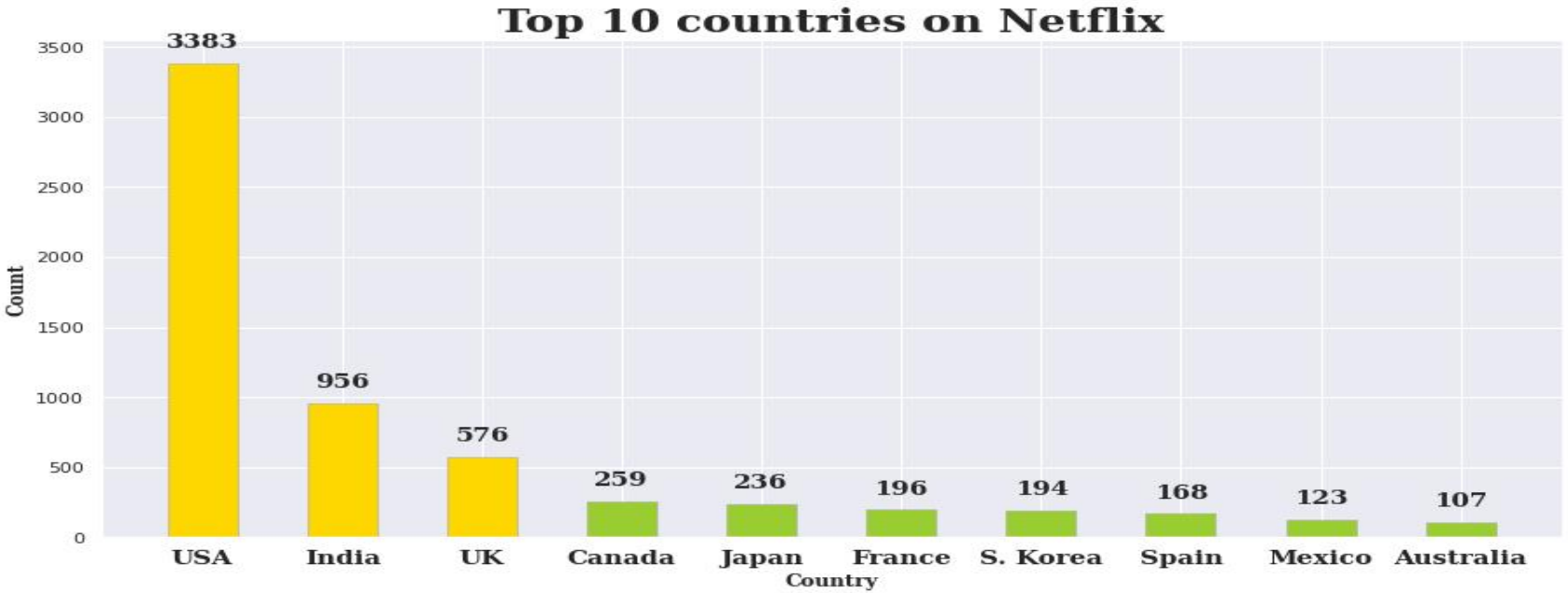
- slow start for Netflix over several years and then there is a rapid increase from 2016.
- content additions have slowed down in 2020, likely due to the COVID-19 pandemic.

Netflix Content Added By Month



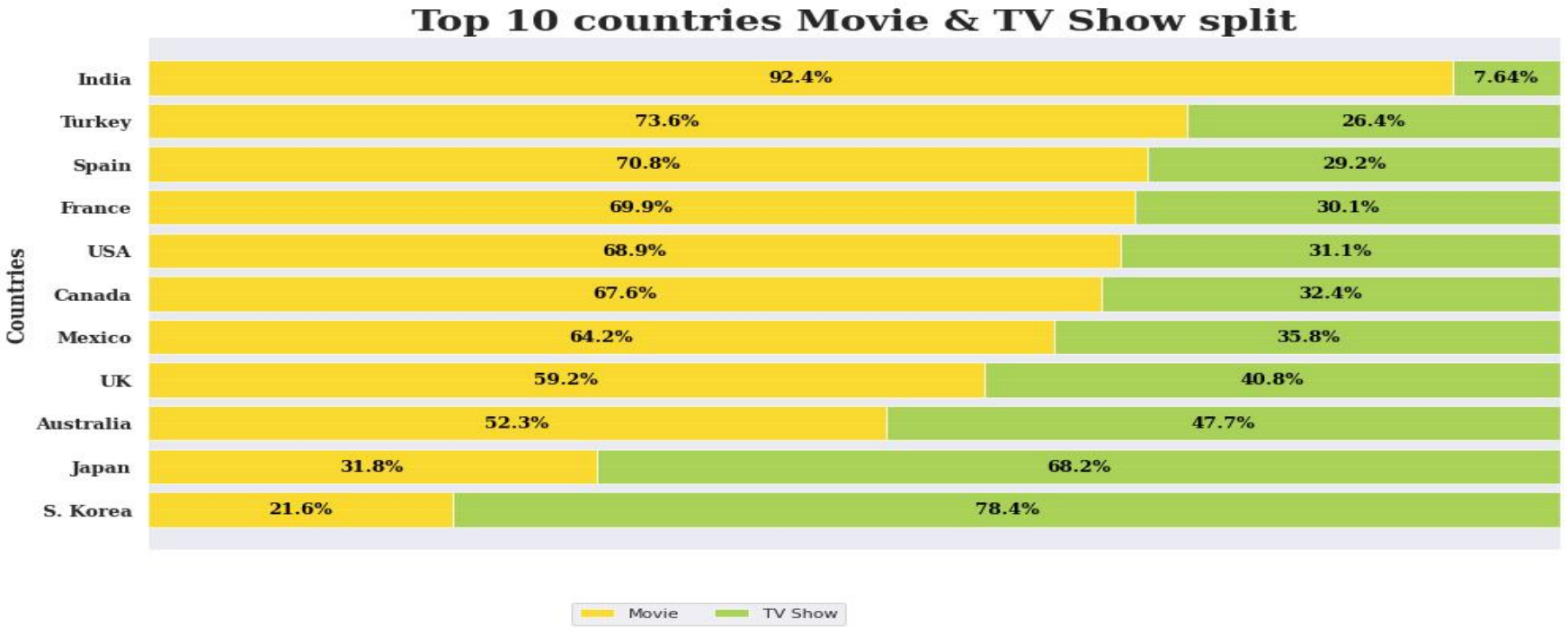
December is the holiday season and it also has Christmas, so in that month most of the content got uploaded

Top countries producing most no of contents



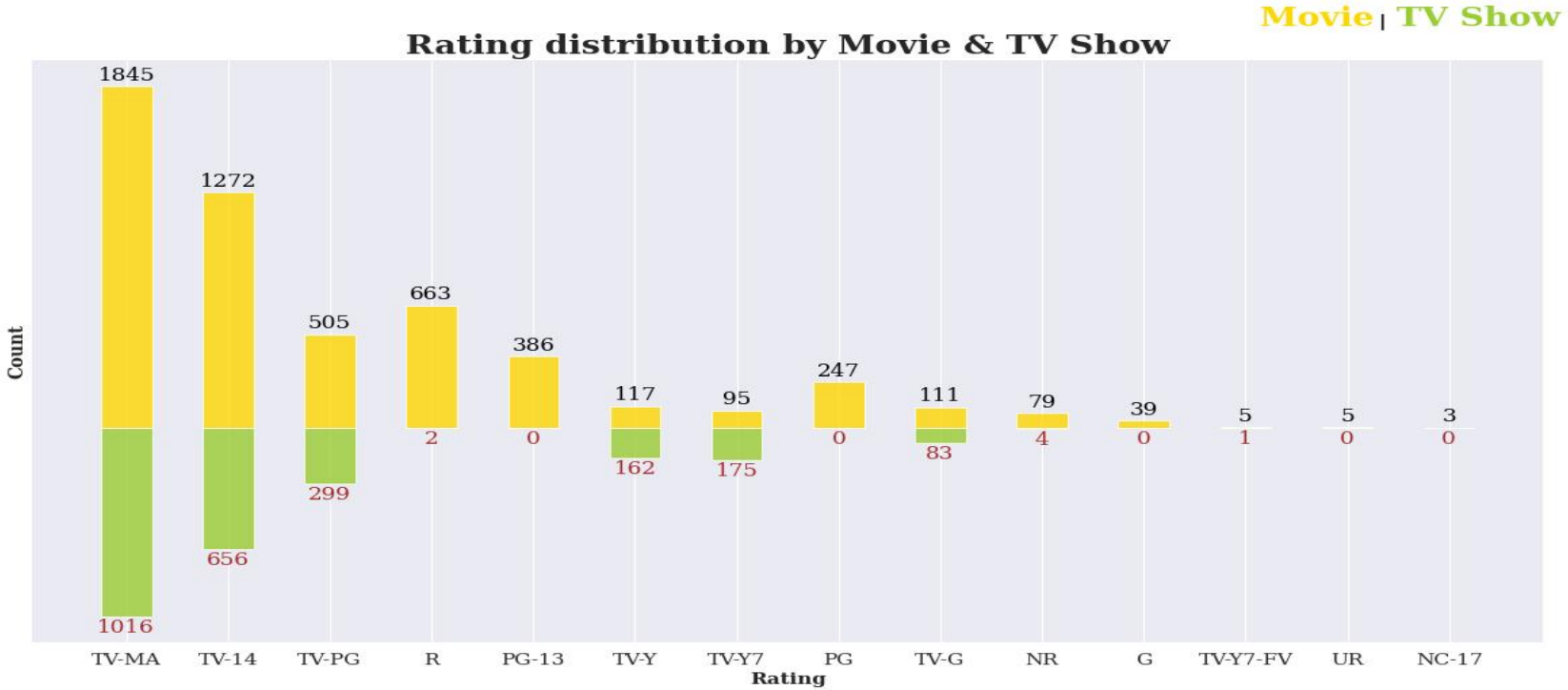
- United States has the most content available.
- India surprisingly comes in second followed by the UK and Canada.

Top countries producing most no of contents



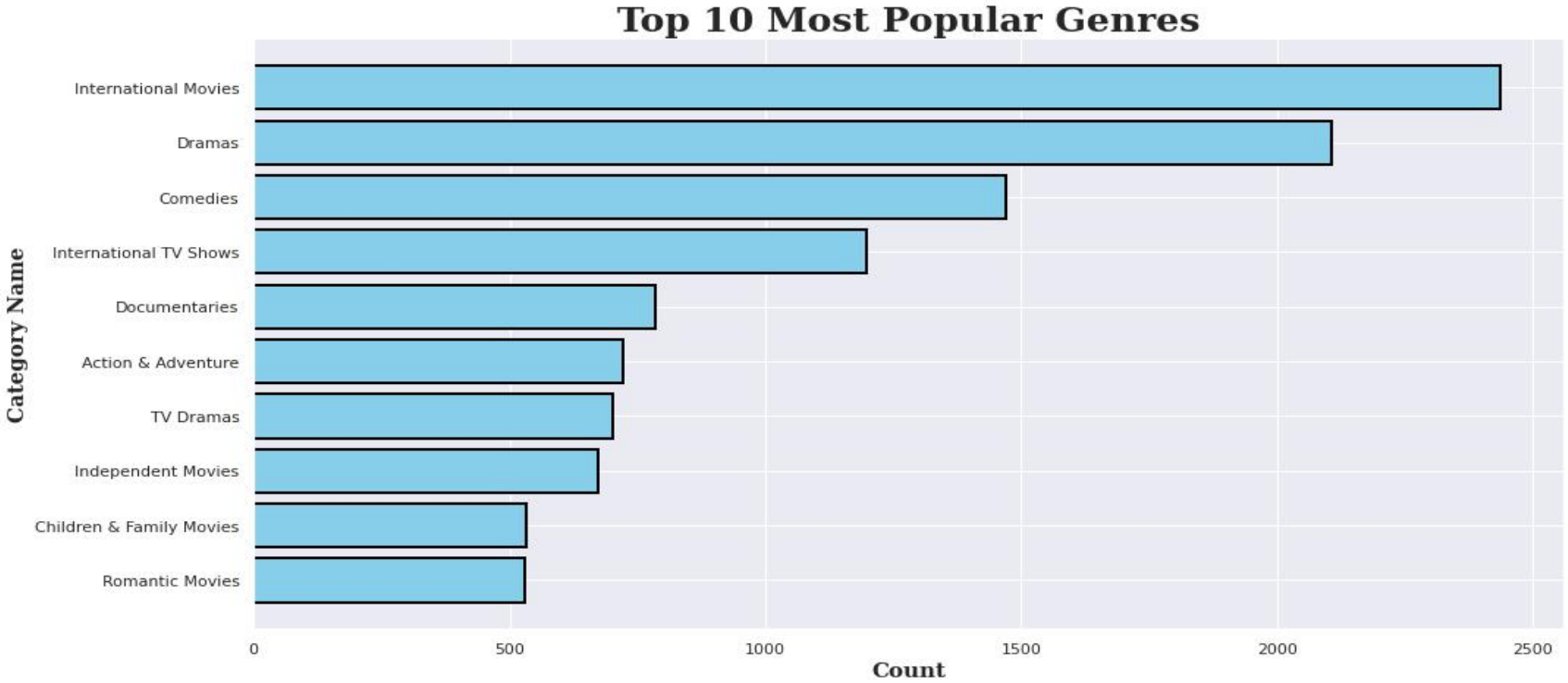
- Interestingly, Netflix in **India** is made up nearly entirely of Movies.
- **South Korean** Netflix on the other hand is almost entirely TV Shows.

Content Based On Rating



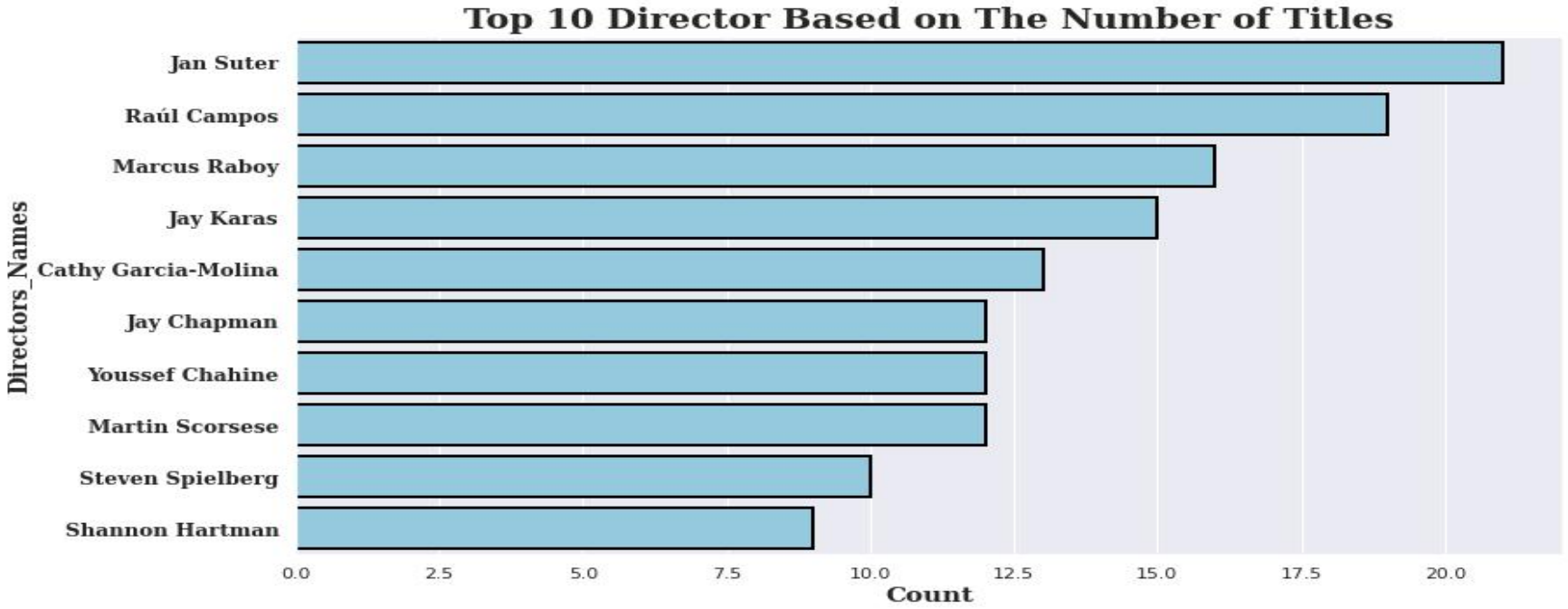
- TV-MA (For Mature Audiences)
- TV-14 (May be unsuitable for children under 14)
- TV-PG (Parental Guidance Suggested)
- NR (Not Rated)

Top 10 Genres

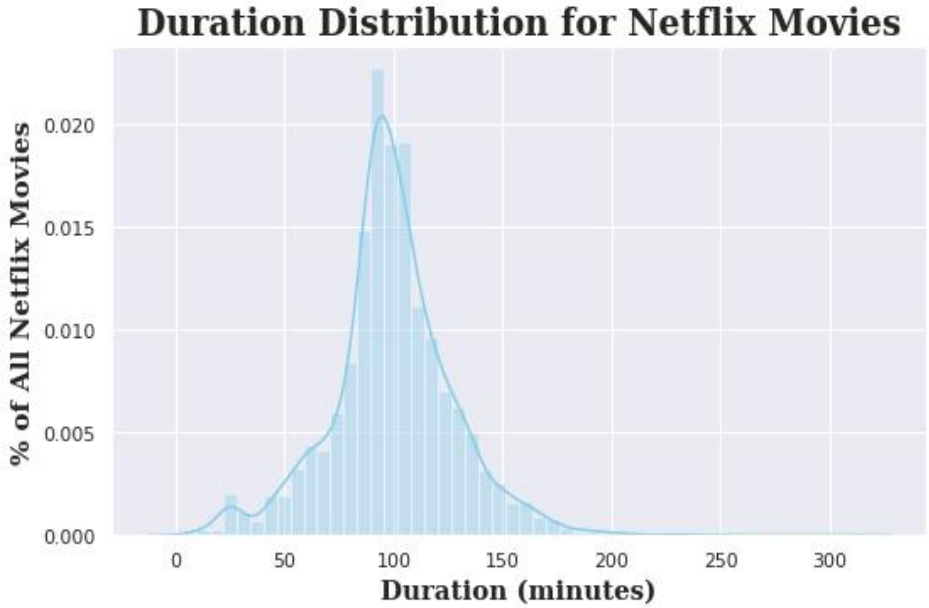


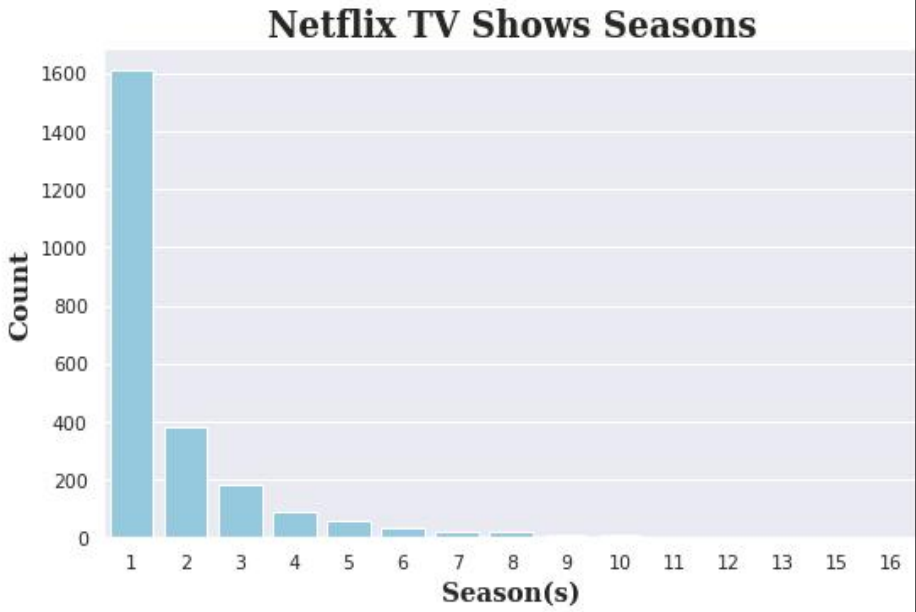
- International movies takes the cake surprisingly followed by dramas and comedies.
- It looks like Netflix has decided to release a ton of international movies.
- The reason for this could be that most Netflix subscribers aren't actually in the United States, but rather the majority of viewers are actually international subscribers.

Top 10 directors on Netflix



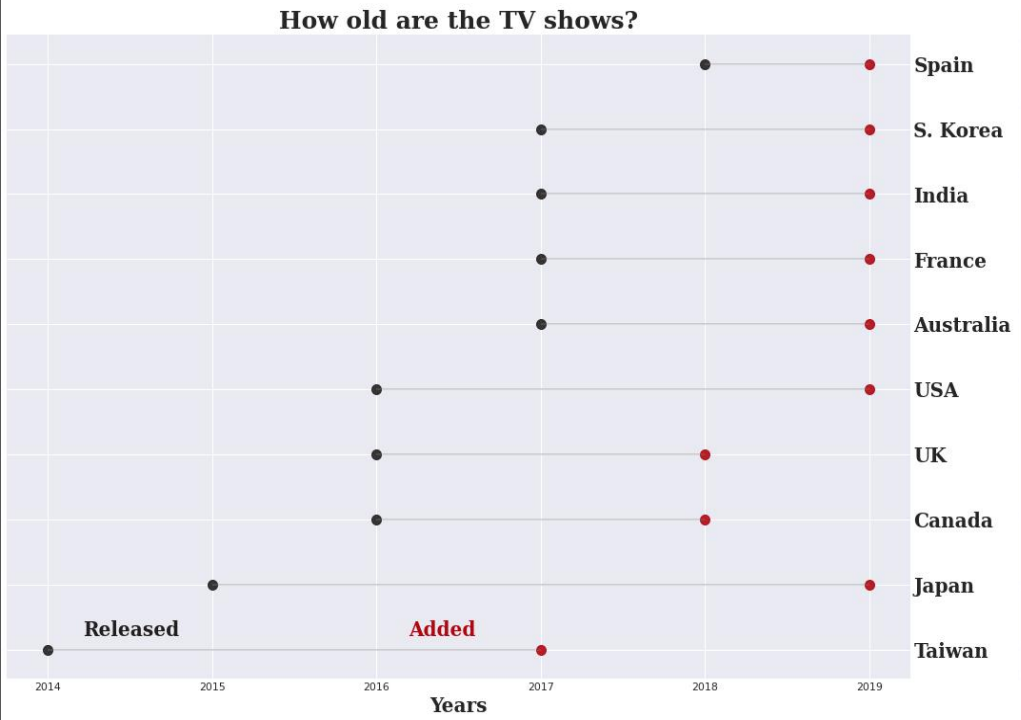
Netflix Content duration

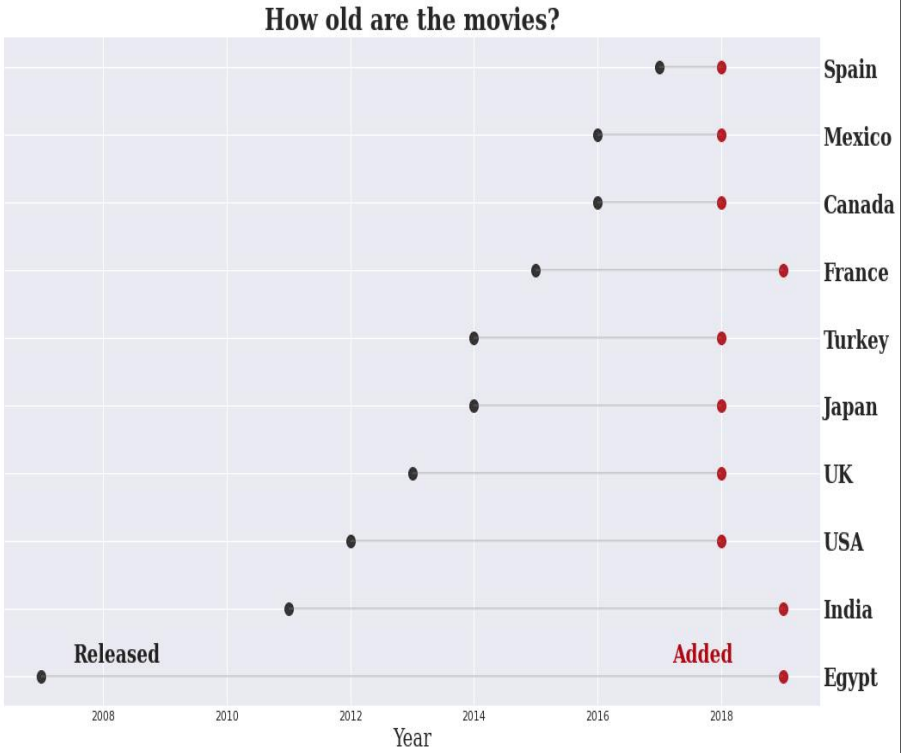




- Most movies are about 70 to 120 min duration
- The majority of shows only have 1 season.

Avg. Gap for Content Released & Added





- Spain have the newest content overall(both Movie & TV-Shows).

Feature Selection & ML Cluster Algorithm

- **Only selected 3 featuresfor clustring**

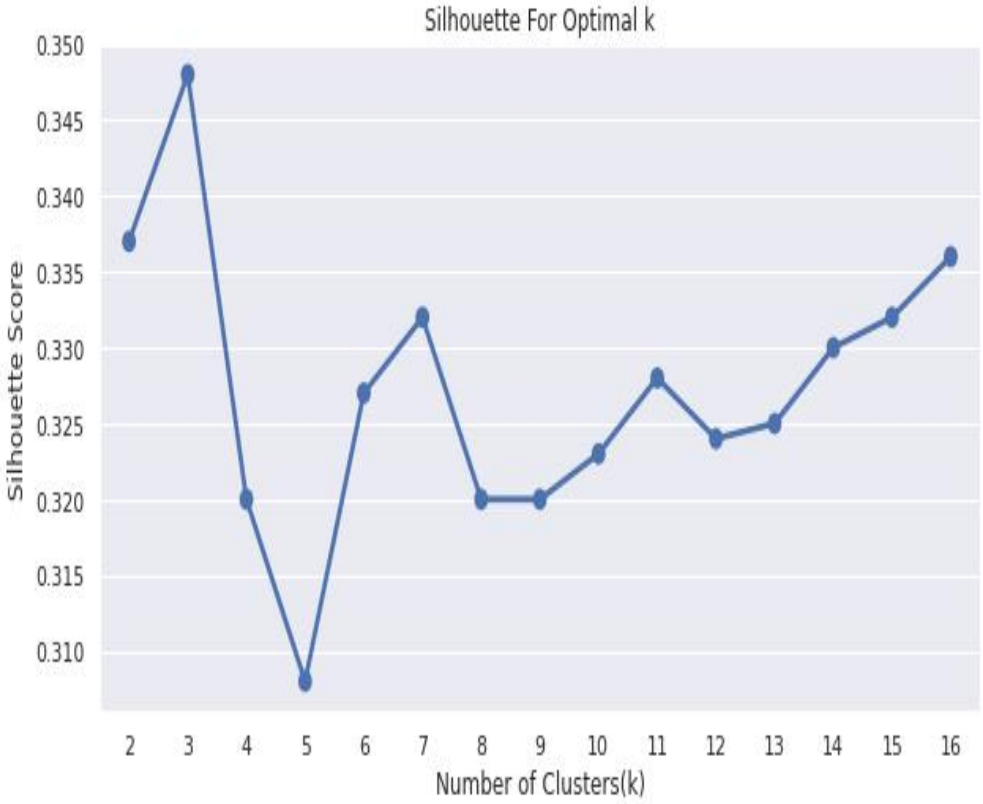
1) Length_description
2) Length_listed_in
3) Category_count

- **For feature Scaling Using StandardScaler**

- **Used 5 algo to find out best k value**

1) Silhouette score
2) Elbow Method
3) DBSCAN
4) Dendrogram
5) AgglomerativeClustering

Silhouette Score



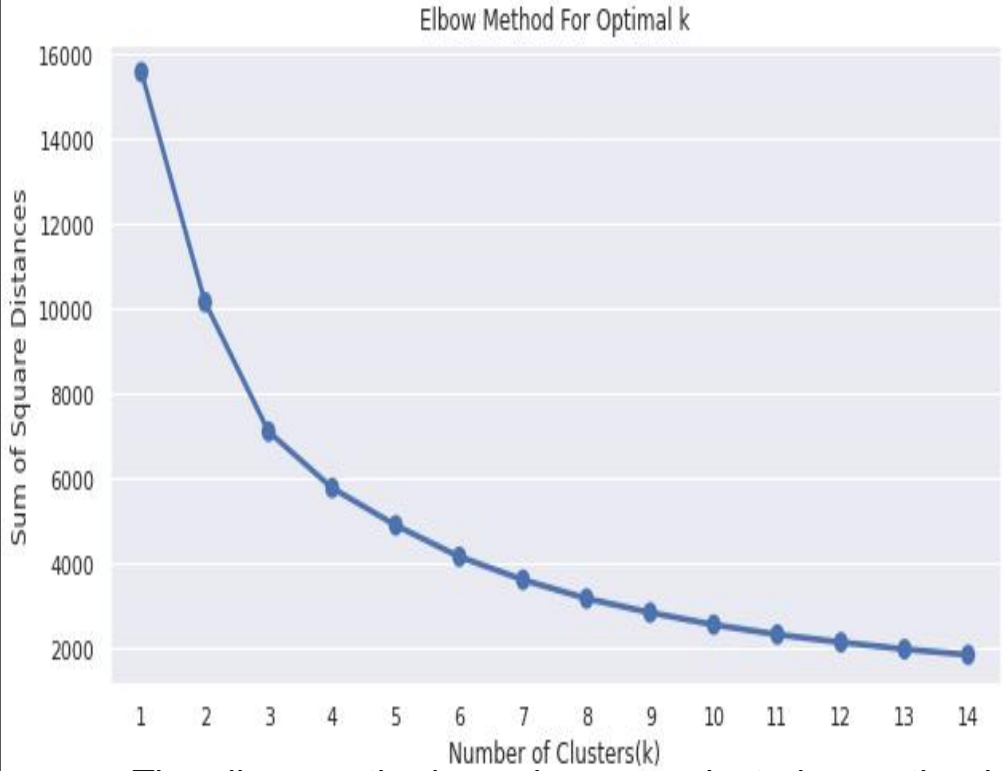
The value of the silhouette coefficient is between [-1, 1]

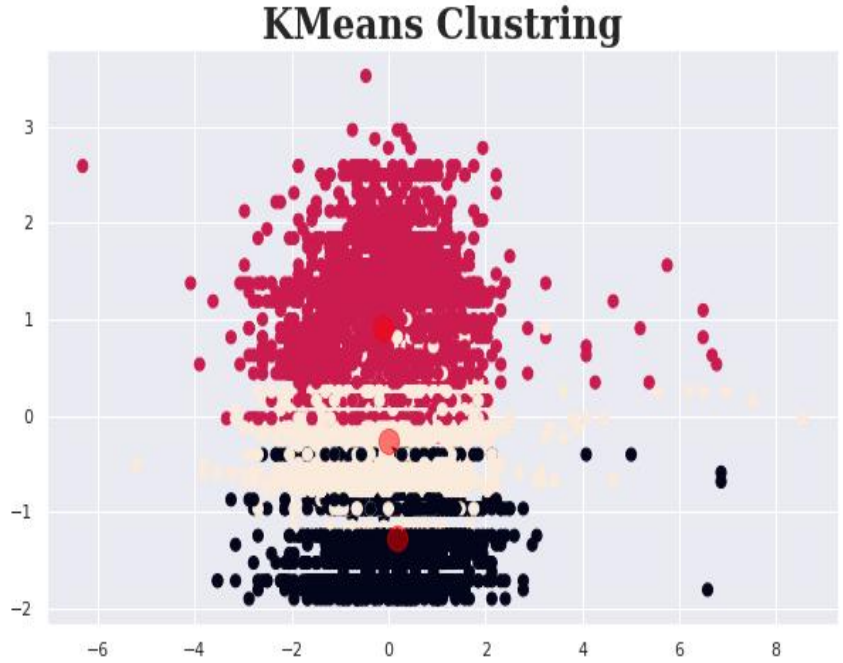
- If score is 1 denotes the best meaning that the data point i is very compact

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

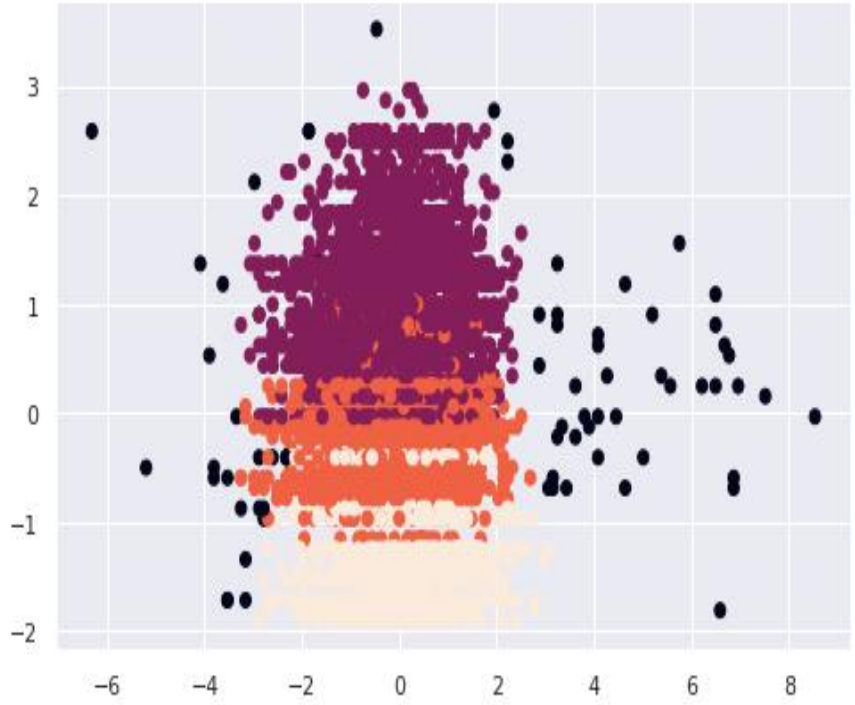
Elbow Method



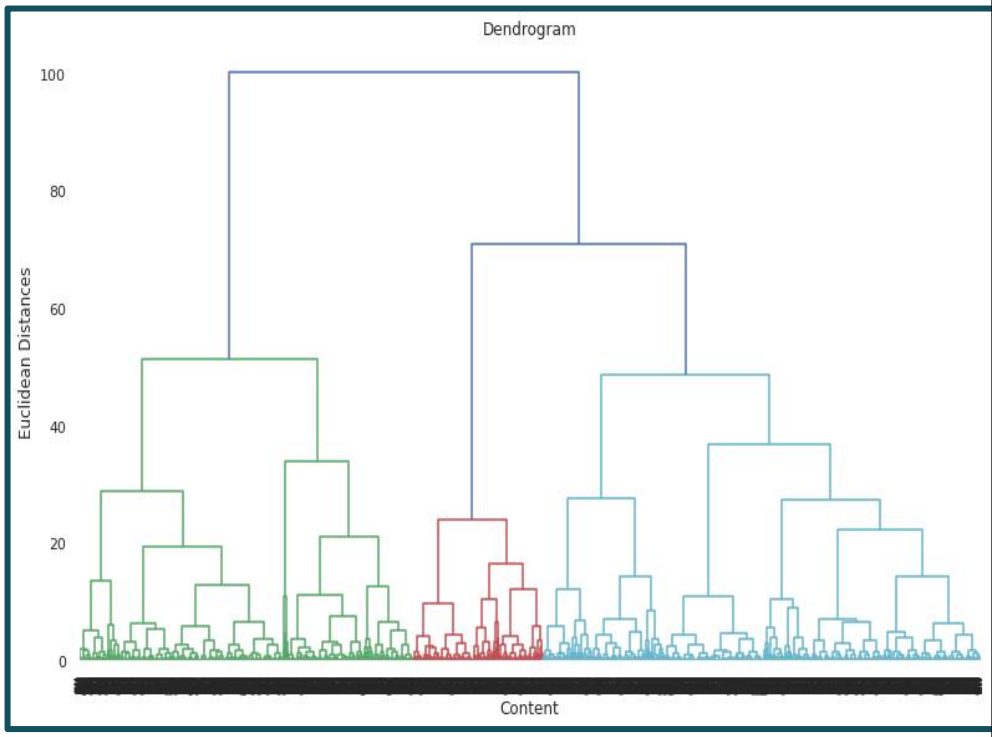


- The elbow method runs k-means clustering on the dataset for a range of values for k (say from 1-15) and then
- for each value of k computes WCSS value . By default, the distortion score is computed, the sum of square
- distances from each point to its assigned center.

DBSCAN & Dendrogram

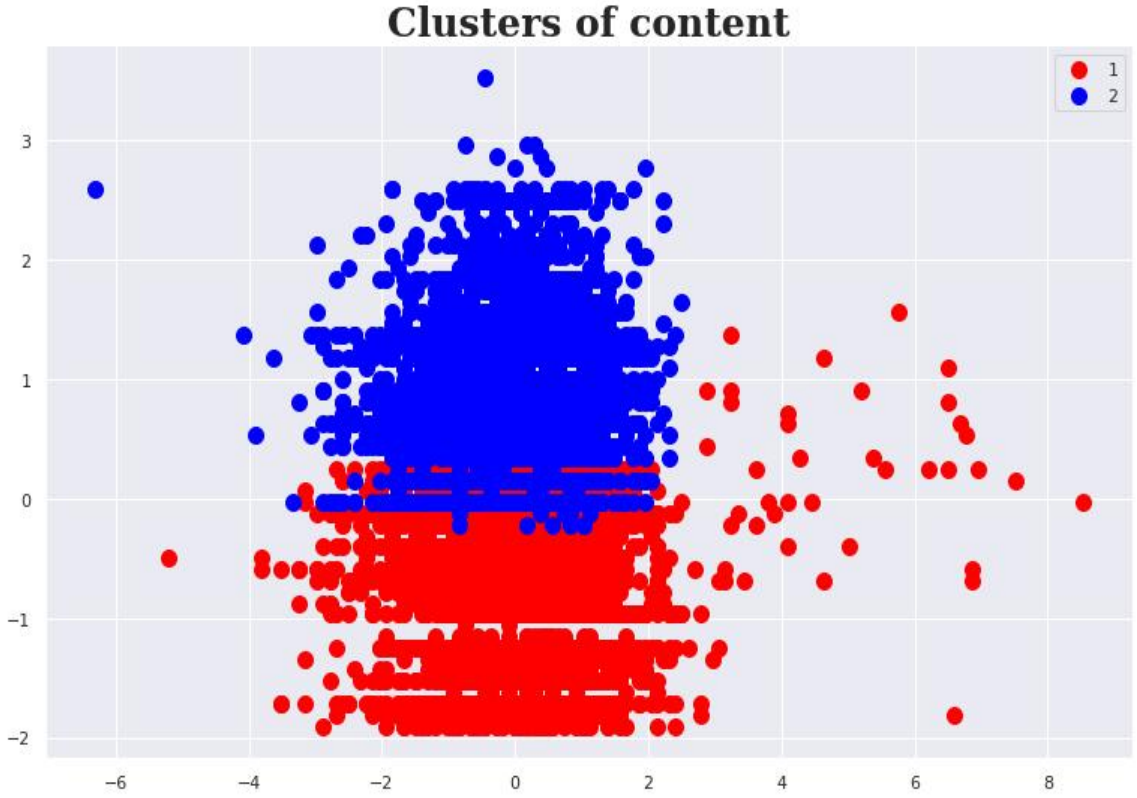


DBSCAN



Dendrogram

AgglomerativeClustering



Steps: -

1. Each data point is assigned as a single cluster.
2. Determine the distance measurement and calculate the distance matrix.
3. Determine the linkage criteria to merge the clusters.
4. Update the distance matrix.
5. Repeat the process until every data point become one cluster.

Conclusion

1) Applied different clustering models like Kmeans, hierarchical, Agglomerative clustering, DBSCAN on data we got the best cluster arrangements.
2) By applying different clustering algorithms to our dataset .we get the optimal number of cluster is equal to 3.
4) In text analysis (NLP) I used stop words, removed punctuations , stemming & TF-IDF vectorizer and other functions of NLP.
5) TOP 3 content categorirs are international movies takes the cake surprisingly followed by dramas and comedies. Even though the United States has the most content available, it looks like Netflix has decided to release a ton of international movies. The reason for this could be that most Netflix subscribers aren't actually in the United States, but rather the majority of viewers are actually international subscribers.
6) On Netflix USA has the largest number of contents. And most of the countries preferred to produce movies more than TV shows.
7)From the dataset insights we can conclude that the most number of TV Shows released in 2020 and for Movies it is 2017.
8) We observe that the content added over years that Netflix is focusing movies and TV shows (Fom 2016 data we get to know that Movies is increased by 80% and TV-shows is increased by 73% compare)
9) In dataset there are two types of contents where 30.86% includes TV shows and the remaining 69.14% carries Movies.
10) The end & beginnings of each year seem to be Netflix's preference for adding content.Least number of contents are added in the month of February.


THANK
YOU