Capstone Project

Netflix Movies and TV Shows Clustering

Team Members:

Vidit Ghelani

Mahesh Lakum

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.

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

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

How to Approach:

Approach the problem in three simple steps:

1. Pre- Processing
    - 1.1 View the data
    - 1.2 Clean the data
        - 1.2.1 Find and drop the irrelevant/ redundant features.
        - 1.2.2 Find and deal with the missing values in the dataset.
2. Performing exploratory data analysis (EDA)
3. Performing NLP
4. Model Training
    - 4.1 K-means Clustering
        - 4.1.1. Silhouette Method
        - 4.1.2. Elbow Method
        - 4.1.3. Dendogram
    - 4.2. Aglomerative Clustering
5. Conclusion

Pre-Processing

In just few simple steps:

1. View the data
2. Clean the data
    - 2.1 Remove/ replace data for missing values
    - 2.2 Remove duplicate values (here none)
    - 2.3 Change column names (if needed)
    - 2.4 Find and drop the irrelevant/ redundant features.
    - 2.5 Find and deal with the missing values in the dataset.

Pre-Processing (Cont.):

The dataset contains 12 features.

Features:

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 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 : Genre
12. description: The Summary description

Pre-Processing (Cont.):

Using the basic commands like info, describe, unique, isnull, etc. we try to learn basic stats of the data:

```
df.shape
```

```
(7787, 12)
```

```
df.describe().T
```

	count	mean	std	min	25%	50%	75%	max
release_year	7787.0	2013.93258	8.757395	1925.0	2013.0	2017.0	2018.0	2021.0

```
df.isnull().sum()
```

```
show_id         0
type            0
title           0
director     2389
cast          718
country       507
date_added     10
release_year    0
rating          7
duration        0
listed_in       0
description     0
dtype: int64
```

```
# To check for duplicate entries.
df.nunique()
```

```
show_id       7787
type             2
title         7787
director      4049
cast          6831
country        681
date_added    1565
release_year    73
rating          14
duration       216
listed_in      492
description   7769
dtype: int64
```

```
<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
```

Here, show_id and title have count equal to the number of total rows. Hence, there are no duplicate entries in this dataset.

Pre-Processing (Cont.):

```
# Dealing with the missing values in the dataset
df['director'].fillna('Data unavailable', inplace=True)
df['cast'].fillna('Data unavailable', inplace=True)
df['country'].fillna('Data unavailable', inplace=True)
df.dropna(subset = ['date_added'], inplace=True)
df.dropna(subset = ['rating'], inplace=True)
```

Here, we will do the following:

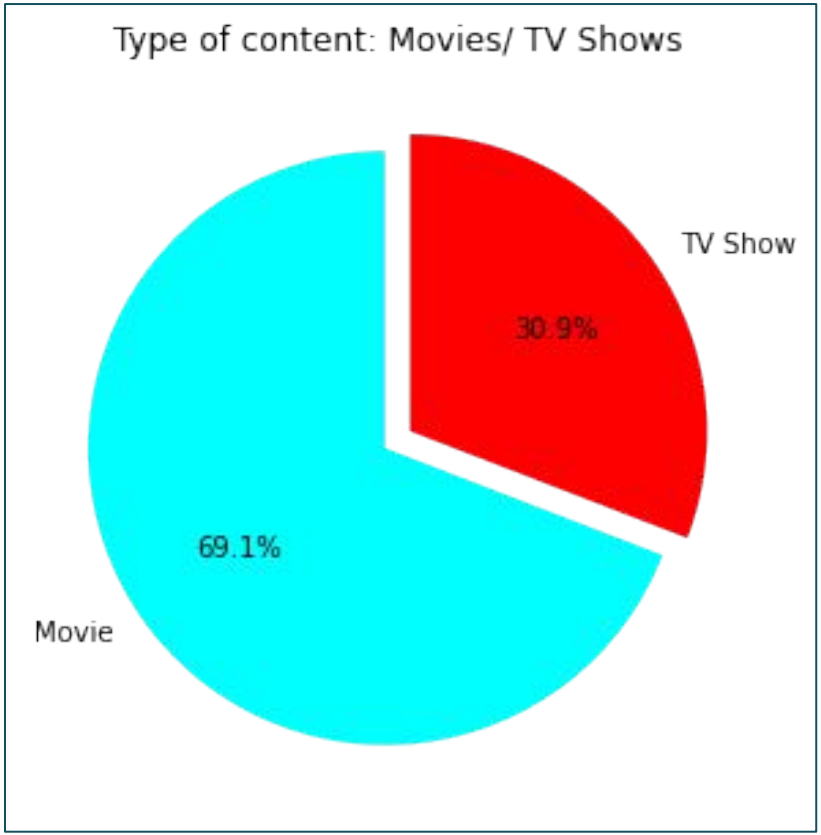
1. Find and drop the irrelevant/ redundant features.
2. Find and deal with the missing values in the dataset.

Hence, we replaced the null data in the 'director', 'cast' and 'country' columns by 'Data unavailable' and dropped the seven null entries in 'rating' and 10 null entries in 'date_added' column.

So, after this shape of the data would be (7770, 12)

Performing the EDA

Number of Movies & TV Shows

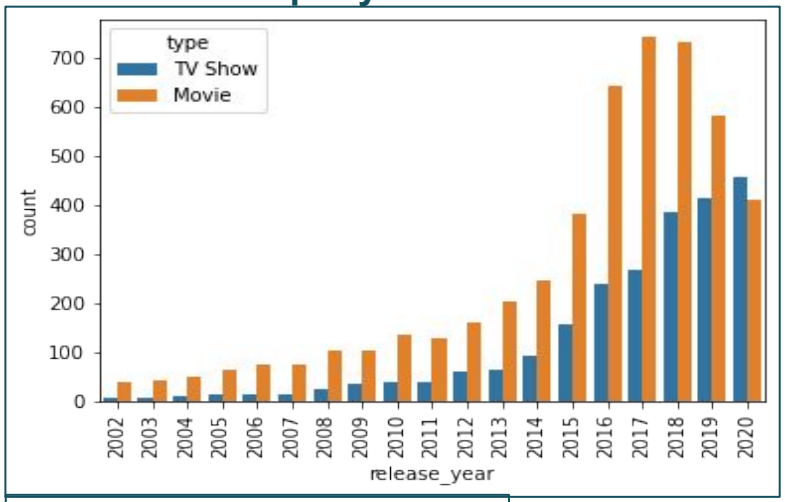


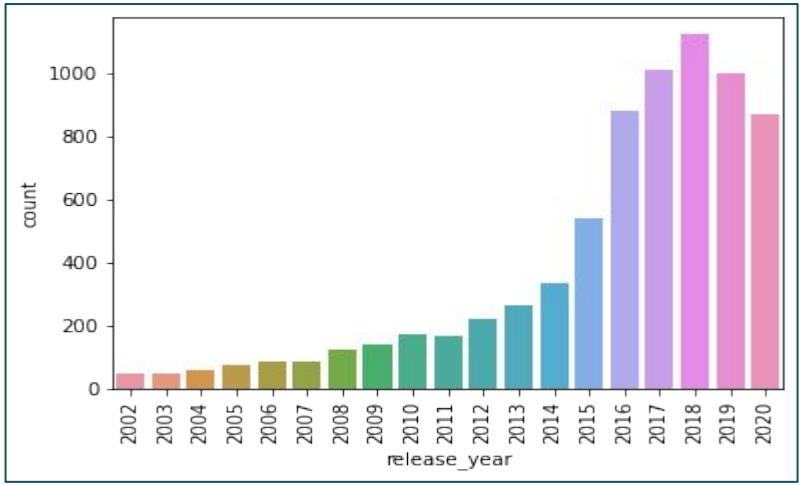
Movie	5372
TV Show	2398

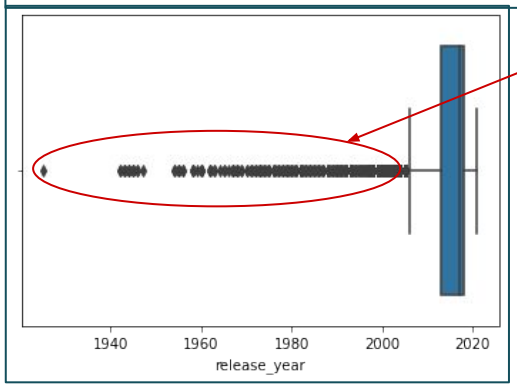
Here, we see that of the total content included in this data, the majority of the data corresponds to Movies. TV Shows have about one third the weightage in the data.

Performing the EDA (Cont.)

Content released per year:





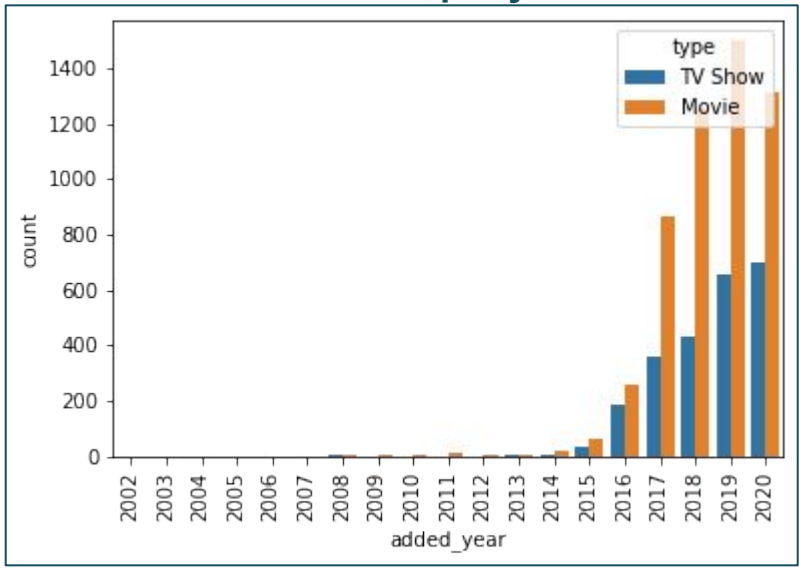


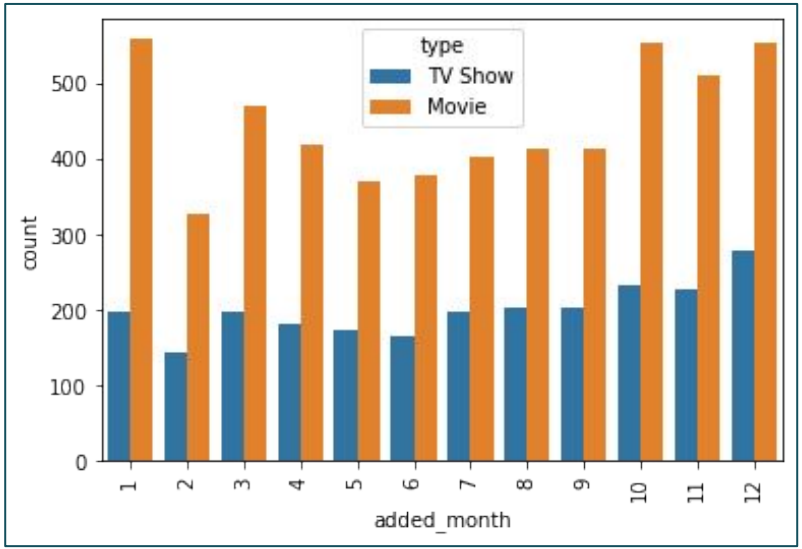
Before 2014 the production growth for Movies & Tv Shows was very less, that's why here it's showing those values(release_year less than 2009) as outliers.

1. From the graph of content released per year, we observe that it follows a steady growth till the year 2014. From the year 2015, it just put out an exponential growth.
2. We also observe that in the same year, the production of TV shows also shot up. Until then it was almost half that of the number of movies produced. Eventually in the year 2020 we see that TV shows surpassed the total count of movies produced.

Performing the EDA (Cont.)

Content added to Netflix per year:

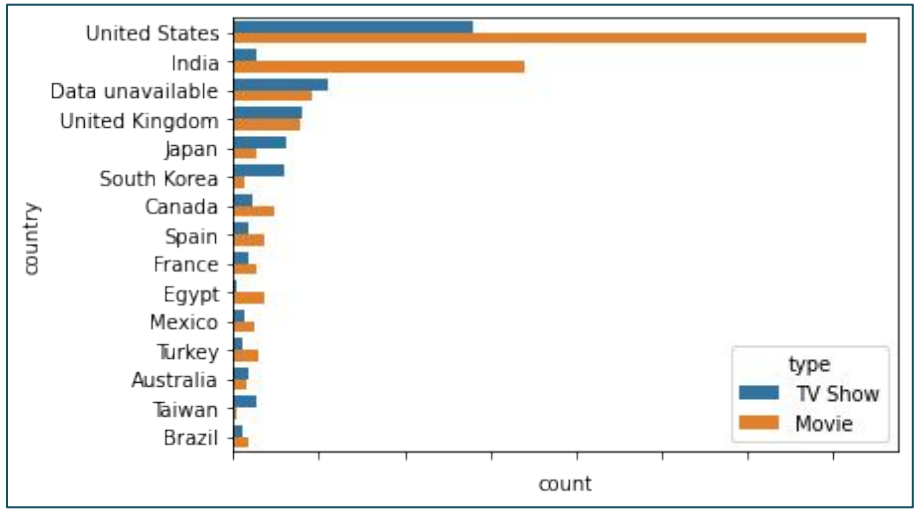




1. Looking at the data in the ‘added_year’ column, we see that till the year 2015, quite less amount of movies were added to Netflix. This started rising from the year 2015. Looking at this and the previous plots we can deduce that Netflix could have sponsored a a few of these productions. Thus, adding a humongous amount of data.
2. Looking at the month wise distribution of the data added to the platform, we observe that the number of movies and TV shows added observed a downward trend from the month of March till July. A reason to speculate could be the fiscal year. The headquarters of Netflix is located in Los Gatos, California, U.S., where the fiscal year is from July to June.

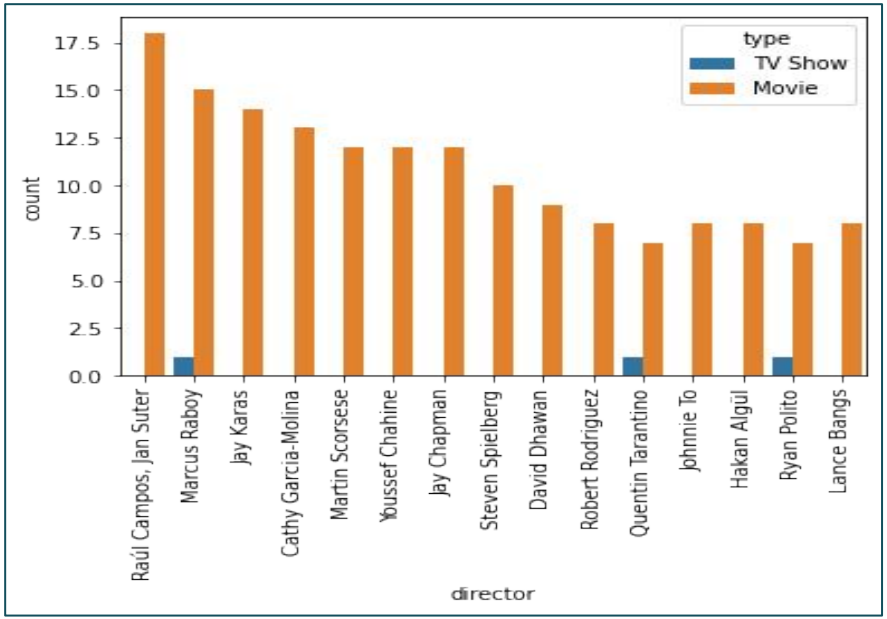
Performing the EDA (Cont.)

Country of Origin:



1. Majority of the content is produced in the USA, followed by India, UK and Japan. Also, a lot of the data seems to lack information on the country of origin hence that ambiguity holds true while still not affecting the outcome of our observation.
2. The UK, Japan and South Korea are the only countries where the total number of TV shows produced is more than the number of movies produced.
3. Of all the top content producer nations, India appears to have the least contribution for TV Shows.

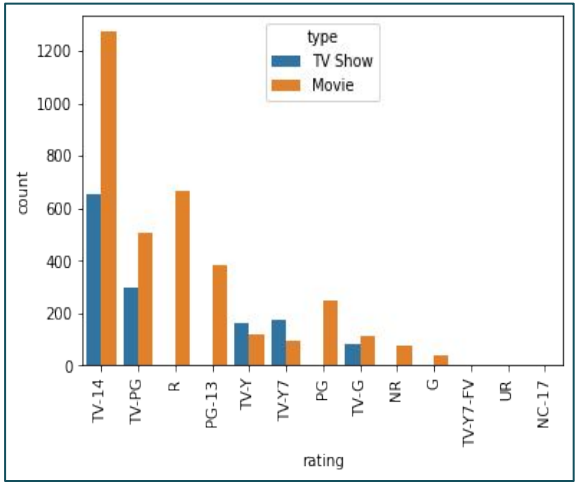
Top Directors:



1. Of the top directors with the majority of content in the dataset, we observe that “Raul Campos” and “Jan Sulter” top the list with their contribution to movies only. Among these top 15 directors only three directors namely, “Marcus Raboy”, “Quentin Tarantino” and “Ryan Polito” are found to have contributed for both movies and TV shows.

Performing the EDA (Cont.)

Top Ratings:



1. TV Shows do no have ratings: 'R', 'PG-13', 'PG', 'NR' and 'G'.
2. The maximum content in the dataset has a rating of 'TV-14' followed by 'TV-PG' & 'R'.
3. 'TV-Y7-FV', 'UR' and 'NC-17' seem to have less or almost no content listed with them.

Meaning of ratings:

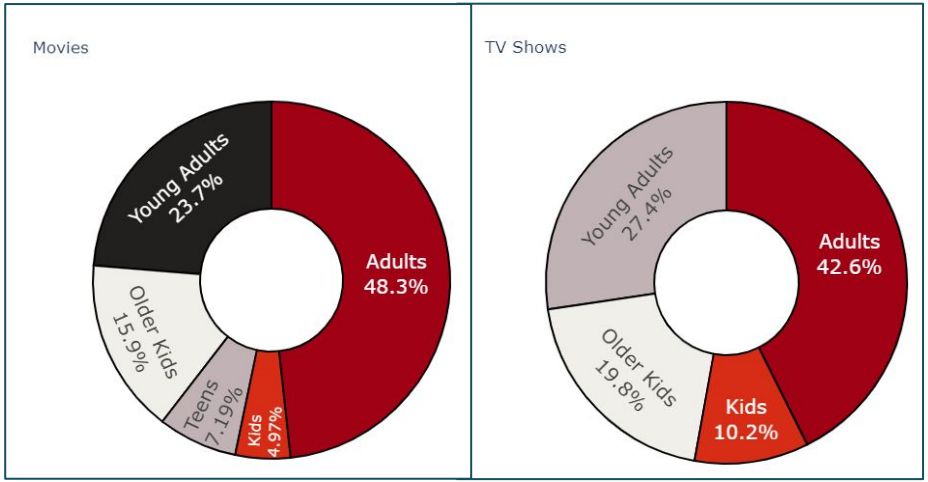
TV-14 → Content unsuitable for children under the age of 14 years.

TV-PG → Content unsuitable for younger children.

R → Content restricted for viewers under the age of 17 years.

PG-13 → Parental guidance for children under 13.

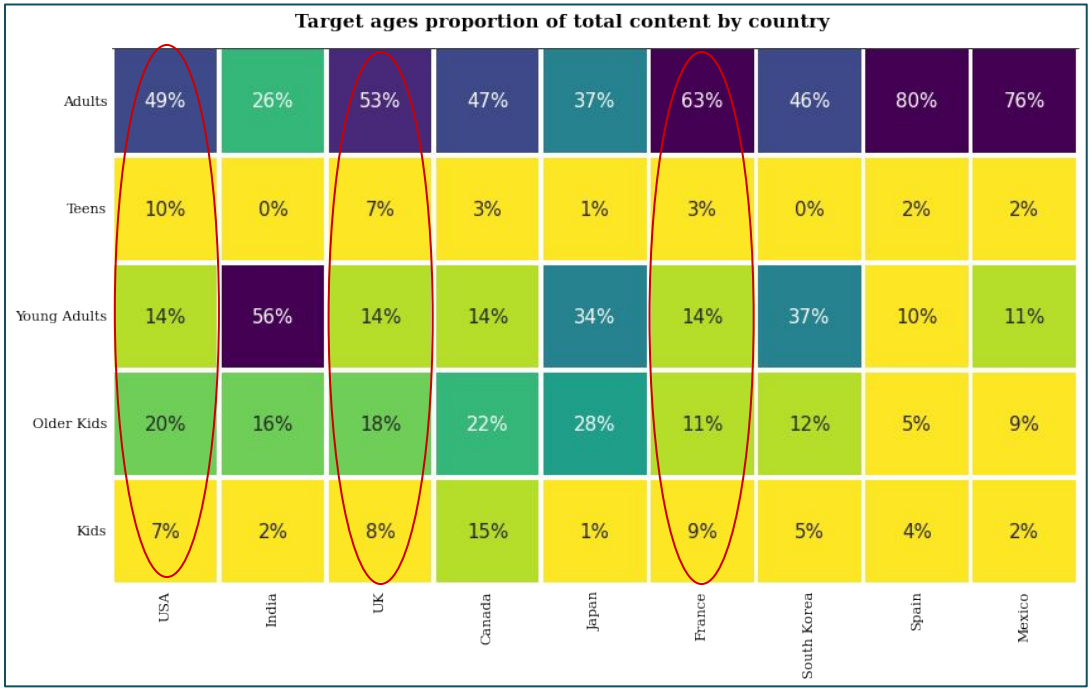
Content vs. Target Age Group:



Using this ratings, we can identify the lower limit of the target audience and based on it we can find out the percentage of content corresponding to different age groups. The above two graphs are plotted for Movies and TV shows for the same.

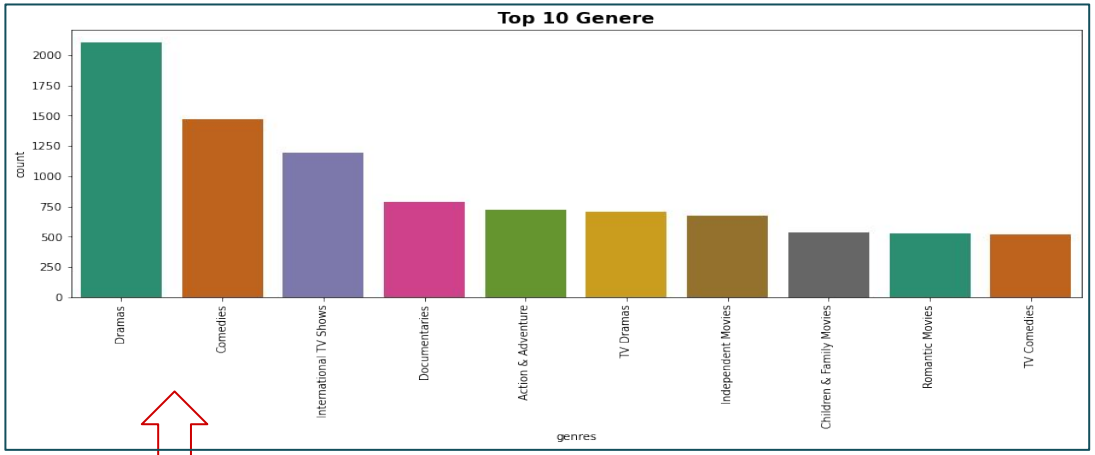
Performing the EDA (Cont.)

Age Group of Target Audience vs. Country of Origin



We see that the proportion of target audience based on the type of content produced by the **USA, UK** and **France** is nearly the same. A similar trait was observed for India and Japan. Apparently the cultural beliefs of these two groups are quite closely similar and this is reflected in the type of content produced.

Performing the EDA (Cont.)

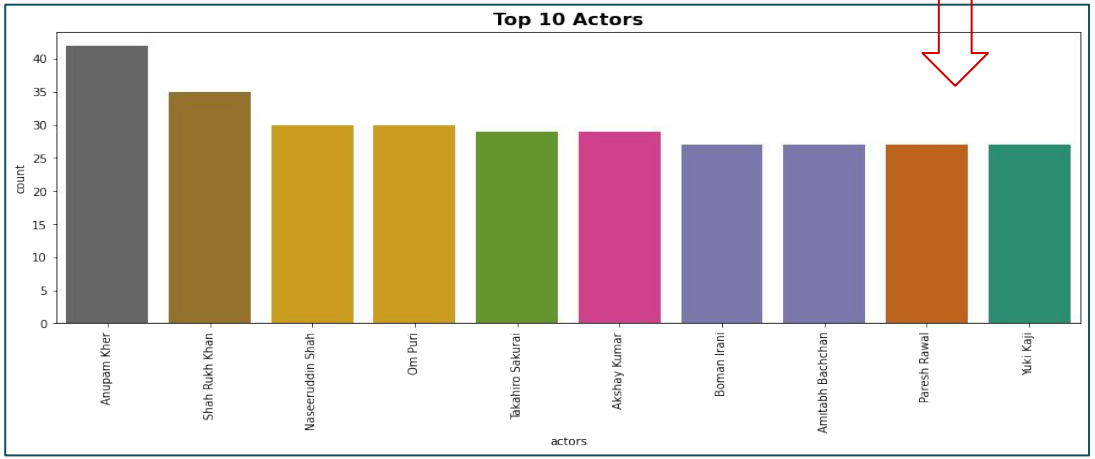


Top Genres:

The most popular genre is ‘Drama’ followed by ‘Comedy’ and ‘International TV Shows’

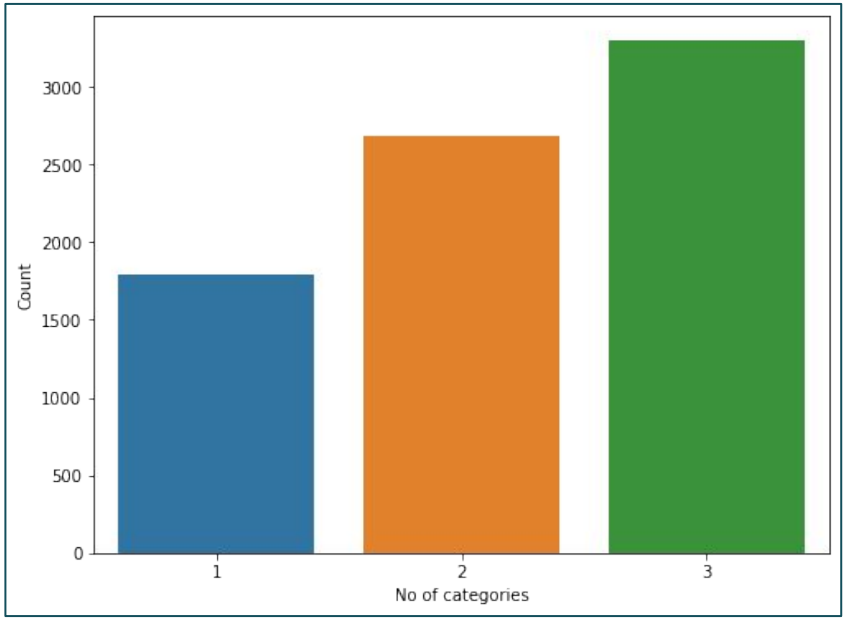
Top Actors:

Top 10 actors based on the number of appearances in movies and TV shows consists of a majority of Indian actors like ‘Anupam Kher’, ‘Shah Rukh Khan’, ‘Naseeruddin Shah’, ‘Om Puri’, ‘Akshay Kumar’, ‘Boman Irani’, ‘Amitabh Bachchan’ and ‘Paresh Raval’.



Performing the EDA (Cont.)

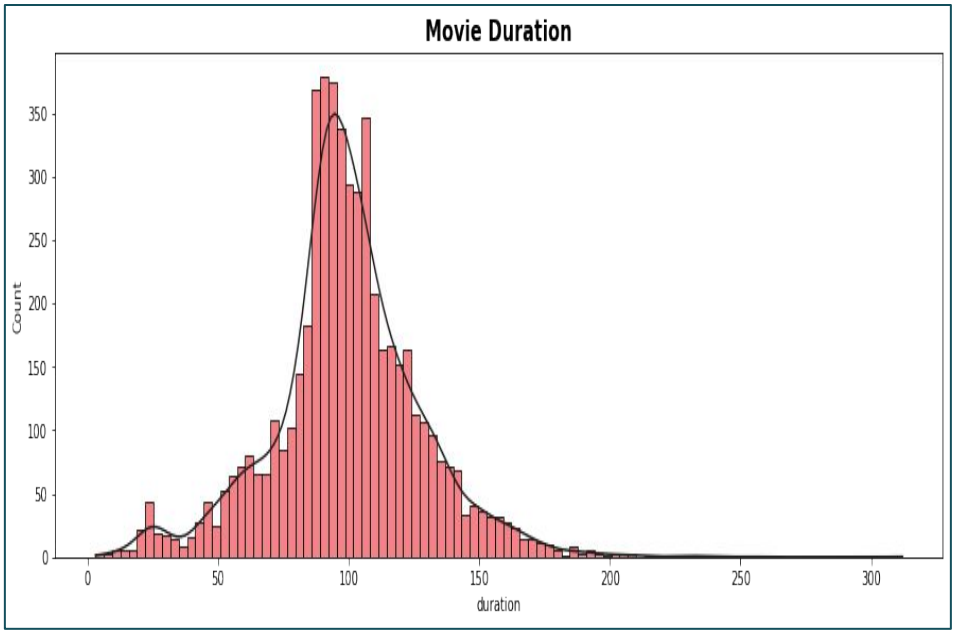
Number of categories each movie holds:



Majority of the content is listed under three different categories.
The content falling under just one category is about one third of the total data.

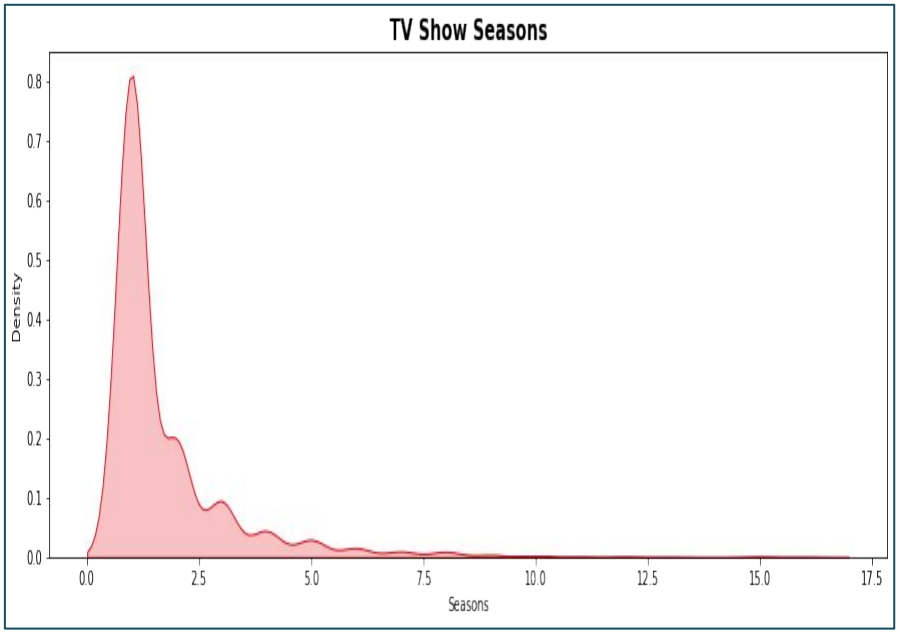
Performing the EDA (Cont.)

Duration of Movies:



In the above histogram plot, we can see that the duration for Netflix movies closely resembles a normal distribution with the average viewing time spanning about 90 minutes which seems to make sense.
Most content are about 70 to 120 min duration for movies.

TV Show - Number of Seasons:



From above we see that the distribution for Netflix TV shows seems to be heavily skewed to the right or say positively skewed where the majority of shows only have 1 season.

Performing the EDA (Cont.)

Wordcloud of Titles:

These words occur the majority of the time in the titles: ‘Movie’, ‘World’, ‘Man’, ‘Story’, ‘Love’, ‘Christmas’, ‘Day’ and ‘Girl’.

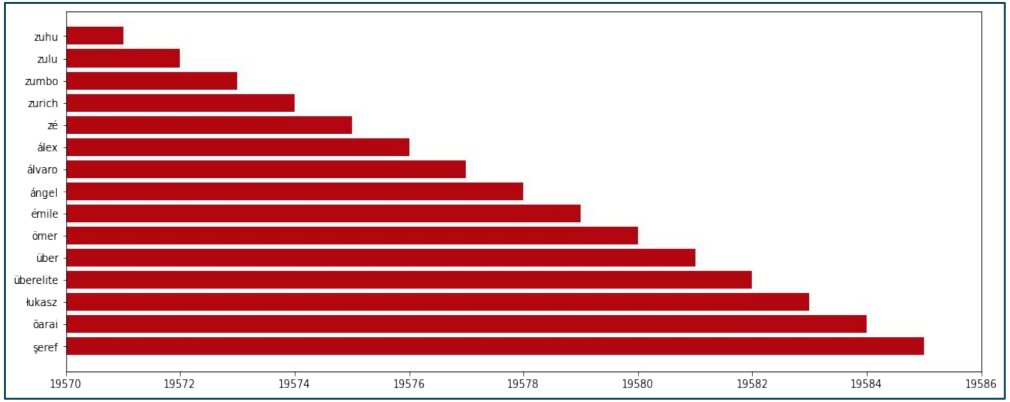
Natural Language Processing (NLP)

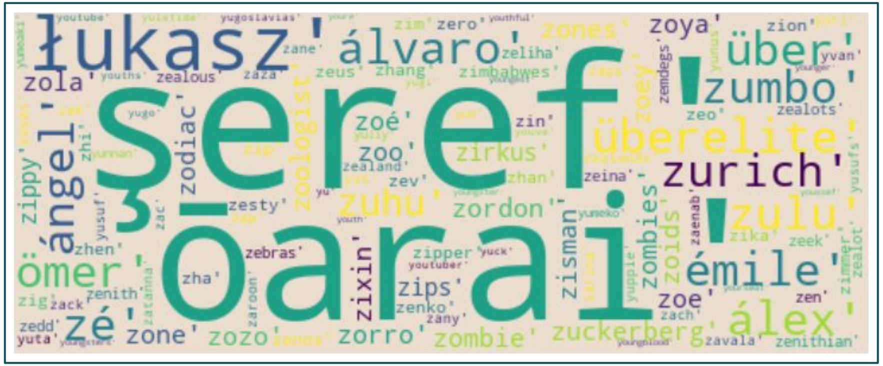
Here we have focused upon the two important features, namely, "description" and "listed_in". Following are the steps followed:

1. Removing the Punctuations
2. Removing the Stopwords
3. Count Vectorizer
4. Most occurred words before stemming
5. Snowball Stemmer
6. TF-IDF Vectorizer
7. Most occurred words after stemming
8. Adding a new feature to hold the length of description

Natural Language Processing (NLP) (Cont.)

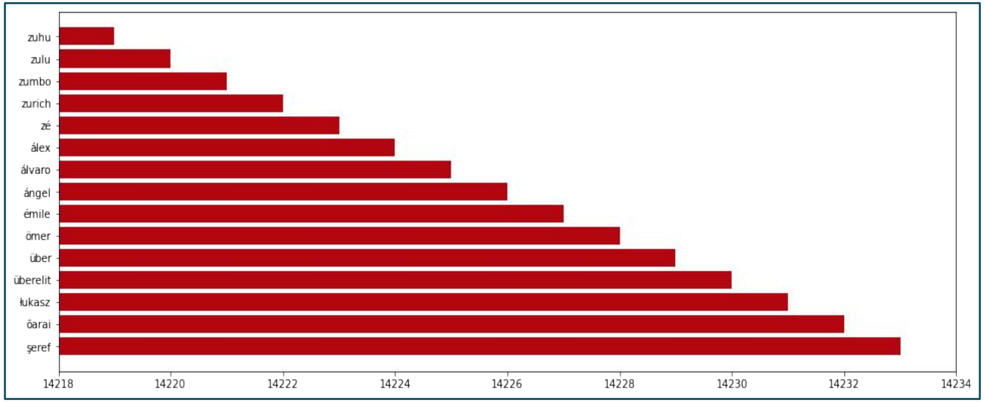
Analyzing the feature: "description":





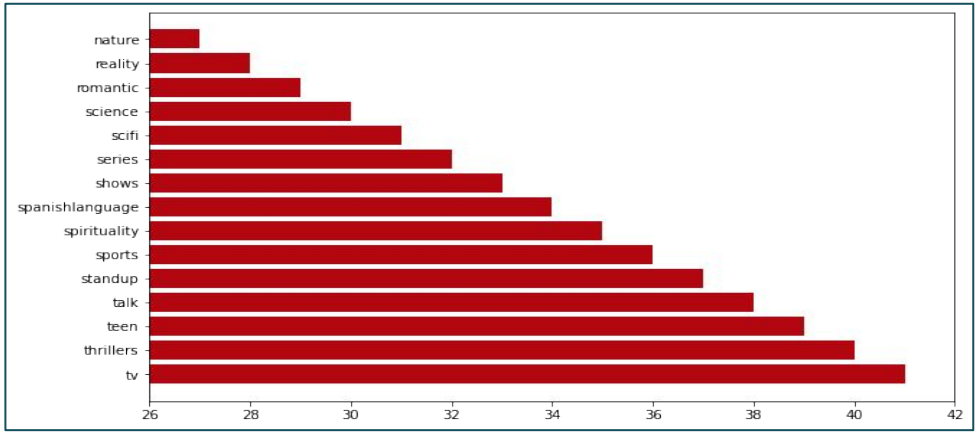
Before Stemming

After Stemming

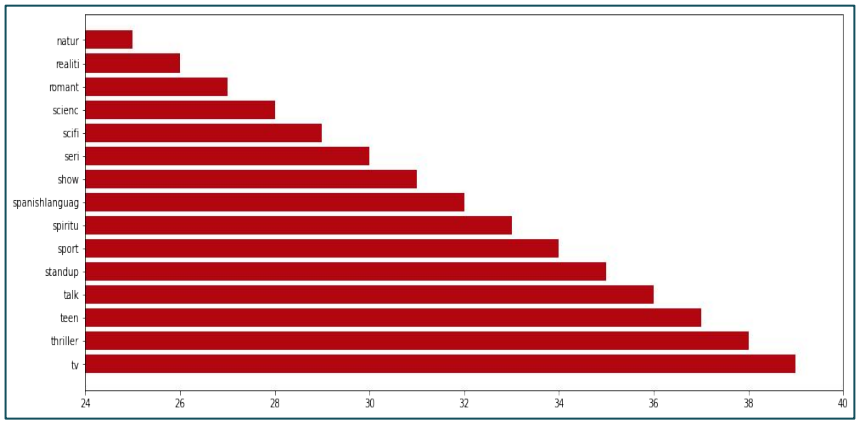


Natural Language Processing (NLP) (Cont.)

Analyzing the feature: "listed_in":



Before Stemming



After Stemming

Applying Clustering Algorithms

Features we are taking for clustering:

1. no_of_category
2. Length (description)
3. Length (listed-in)

Here we have used below clustering method:

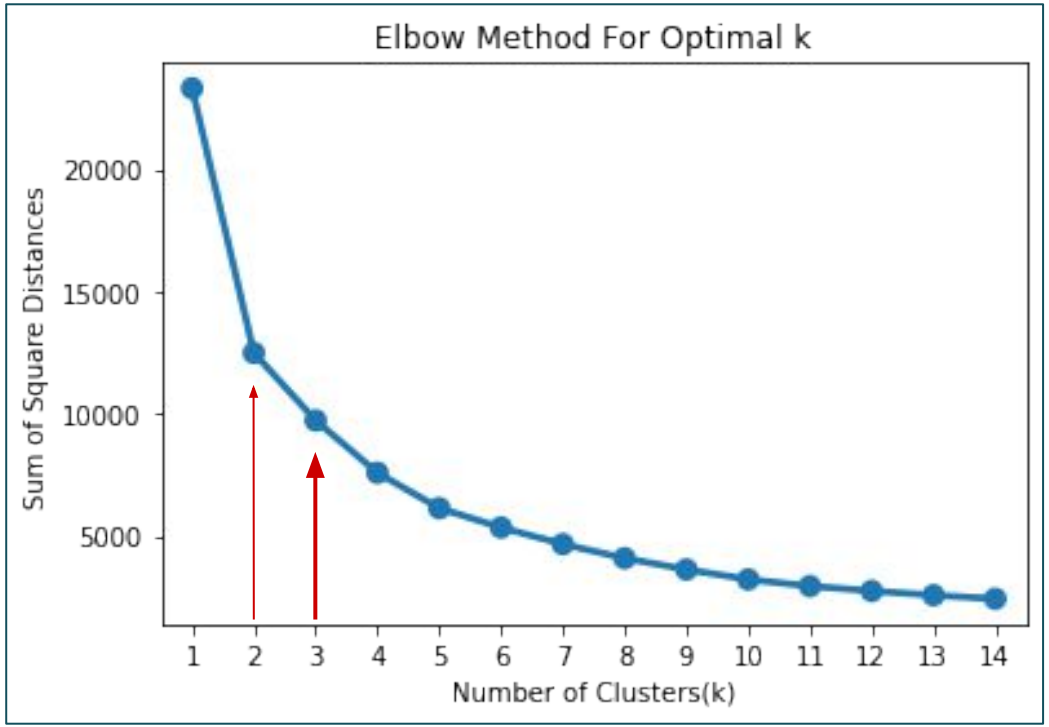
1. K-MEANS clustering
2. Silhouette_score
3. Elbow Method
4. Dendogram
5. AgglomerativeClustering

Applying Clustering Algorithms (Cont.)

Silhouette_score

n_clusters	silhouette_score
2	0.428
3	0.383
4	0.374
5	0.372
6	0.368
7	0.376
8	0.353
9	0.374
10	0.365
11	0.356
12	0.355
13	0.351
14	0.348
15	0.344

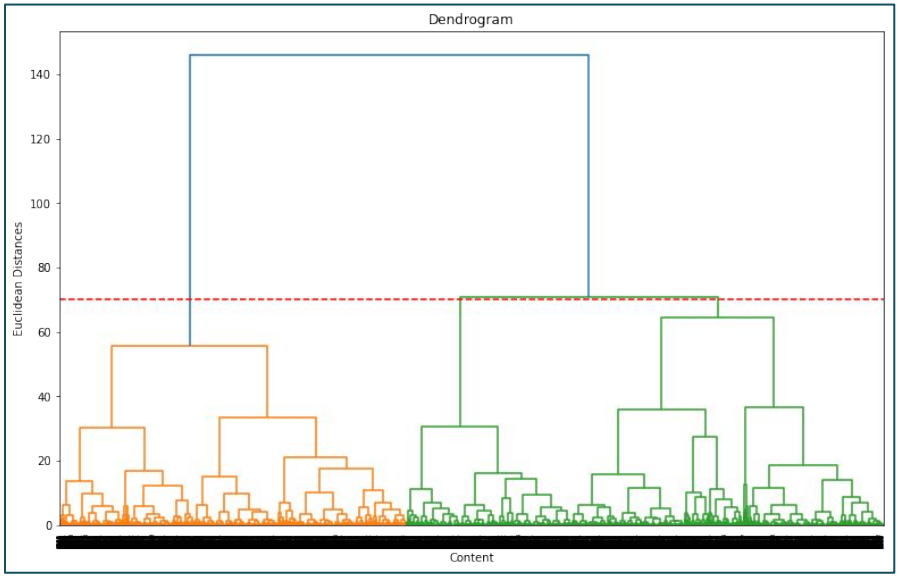
Elbow Method



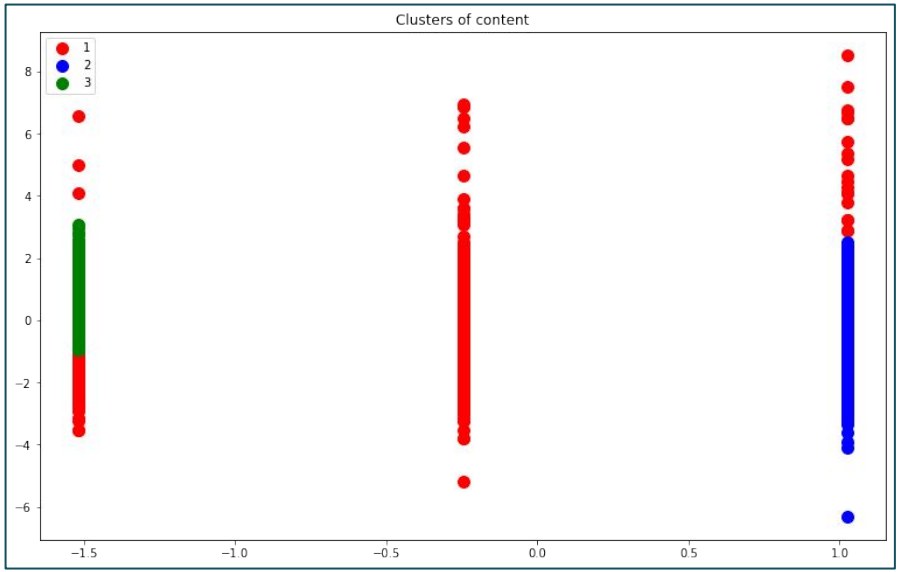
Here, the elbow appears to give two options, i.e. 2 and 3 clusters.

Applying Clustering Algorithms (Cont.)

Dendogram



Agglomerative Clustering



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

Content per cluster:

Cluster 1 (Red) : 3275

Cluster 2 (Blue) : 2937

Cluster 3 (Green) : 1558

Conclusions:

1. The available data consists of 69.1% of titles corresponding to movies and the rest 30.9% to TV Shows. Thus, movies as a category of content dominate the quantity in this dataset.
2. Looking at the data from release year 2001 to 2018, the number of movies and TV shows released has observed an exponential rise, with a major break through observed in the year 2014-2015
3. We can also observe that the amount of movies released has been about 2 to 3 times the amount of TV shows released. However, it is evident that this ratio started to reduce from the year 2019. Thus the demand for the TV shows started to peak up from the year 2017 and impacted the production criteria by the year 2019.
4. However, the content started to get on the Netflix platform in mass from the year 2015 - 2016.
5. Looking at the month wise distribution of the data added to the platform, we observe that the number of movies and TV shows added, observed a downward trend from the month of March till July. A reason to speculate could be the fiscal year. The headquarters of Netflix is located in Los Gatos, California, U.S., where the fiscal year is from July to June.
6. Majority of the content producers were from the USA, followed by India, UK and Japan. Also, a majority of the data seems to lack information on the country of origin hence that ambiguity holds true while still not affecting the outcome of our observation.
7. The UK, Japan and South Korea are the only countries where the total number of TV shows produced is more than the number of movies produced.
8. Of all the top content producer nations, India appears to have the least contribution for TV Shows.
9. Of the top directors with the majority of content in the dataset, we observe that “Raul Campos” and “Jan Sulter” top the list with their contribution to movies only. Among these top 15 directors only three directors namely, “Marcus Raboy”, “Quentin Tarantino” and “Ryan Polito” are found to have contributed for both movies and TV shows.

Conclusions:

10. Observation for Ratings:
    TV Shows do not have ratings: 'R', 'PG-13', 'PG', 'NR' and 'G'. The maximum content in the dataset has a rating of 'TV-14' followed by 'TV-PG' & 'R'. 'TV-Y7-FV', 'UR' and 'NC-17' seem to have less or almost no content listed with them.
10. We could say that the majority of the content here is for adults and young adults. Very little content is available for kids.
11. We see that the proportion of target audience based on the type of content produced by the USA, UK and France is nearly the same. A similar trait was observed for India and Japan. Apparently the cultural beliefs of these two groups are quite closely similar and this is reflected in the type of content produced.
12. The most popular genre is ‘Drama’ followed by ‘Comedy’ and ‘International TV Shows’.
13. Top 10 actors based on the number of appearances in movies and TV shows consists of a majority of Indian actors like ‘Anupam Kher’, ‘Shah Rukh Khan’, ‘Naseeruddin Shah’, ‘Om Puri’, ‘Akshay Kumar’, ‘Boman Irani’, ‘Amitabh Bachchan’ and ‘Paresh Raval’.
14. The duration of movies appears to be a near normal distribution with the mean duration of 90 minutes. Looking at the number of seasons for the TV shows, it appears that the majority of the shows have only one season. Hence, the plot is heavily skewed towards the left.
15. These words occur the majority of the time in the titles: ‘Movie’, ‘World’, ‘Man’, ‘Story’, ‘Love’, ‘Christmas’, ‘Day’ and ‘Girl’.
16. We applied two clustering algorithms namely K- Means and Agglomerative clustering algorithm. The best cluster arrangement we obtained was three.

Thank you

References of Images:

1. Vector image and graphics of the muse: shutterstock.com.