Capstone Project

Netflix Movies And TV Shows Clustering

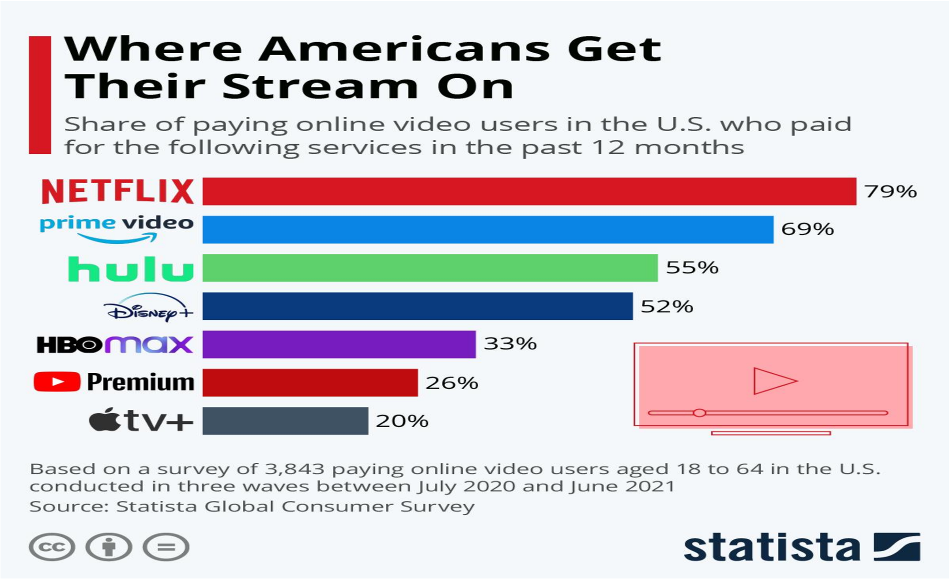
Sanjay Jaiswal

Content:

- Introduction
- Problem Statement
- Data Inspection
- Attribute Information
- Data Cleaning
- Exploratory Data Analysis
- Feature Selection
- Implementing Algorithms
- Conclusion

Introduction:

Netflix, Inc. is an American subscription streaming service and production company. Launched on August 29, 1997, it offers a film and television series library through distribution deals as well as its own productions, known as Netflix Originals.
Netflix was founded on the aforementioned date by Reed Hastings and Marc Randolph in Scotts Valley, California.



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.

Data Inspection:

- This dataset contain **7787 observations** and **12 features**.
- The dataset consists of 11 textual columns and **one numeric column**.
- No Duplicate values.

```
<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
```

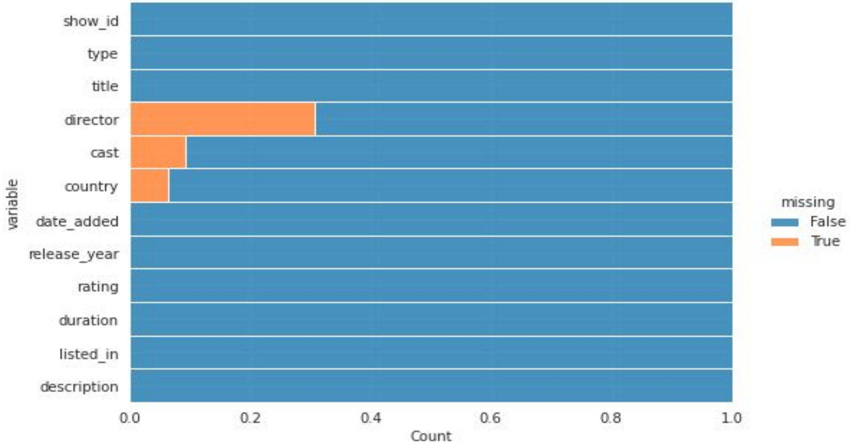
Attribute Information:

- **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
- **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:

- **Director** feature have more than 30.68% of null values. Filling null values by 'unknown'.
- **Country** feature have 6.51% of null values. Filling null values by mode of feature.
- **Cast** feature have 9.22% of null values. Filling null values by 'unknown'.
- **Rating** feature have 0.09% of null values. Filling null values by mode of feature.
- **Date_added** feature have 0.13% of null values Dropping rows corresponding to null values.

	column_name	no.of_missing	missing_percentage
0	show_id	0	0.00
1	type	0	0.00
2	title	0	0.00
3	director	2389	30.68
4	cast	718	9.22
5	country	507	6.51
6	date_added	10	0.13
7	release_year	0	0.00
8	rating	7	0.09
9	duration	0	0.00
10	listed_in	0	0.00
11	description	0	0.00

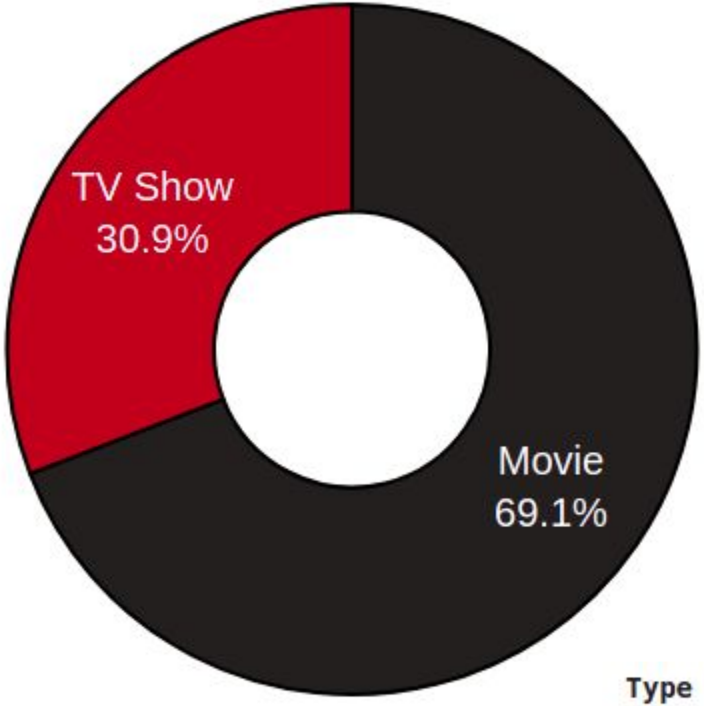


Analysis on Type:

➔ 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.

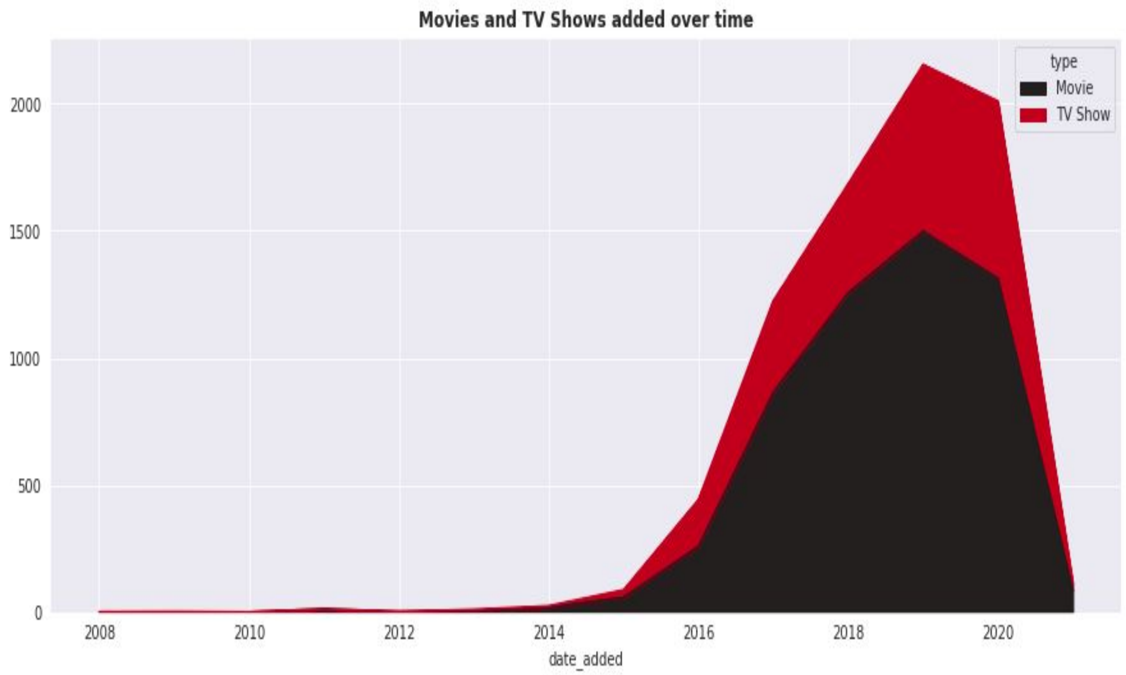
➔ There are about **69%** movies and **31%** TV shows on Netflix.



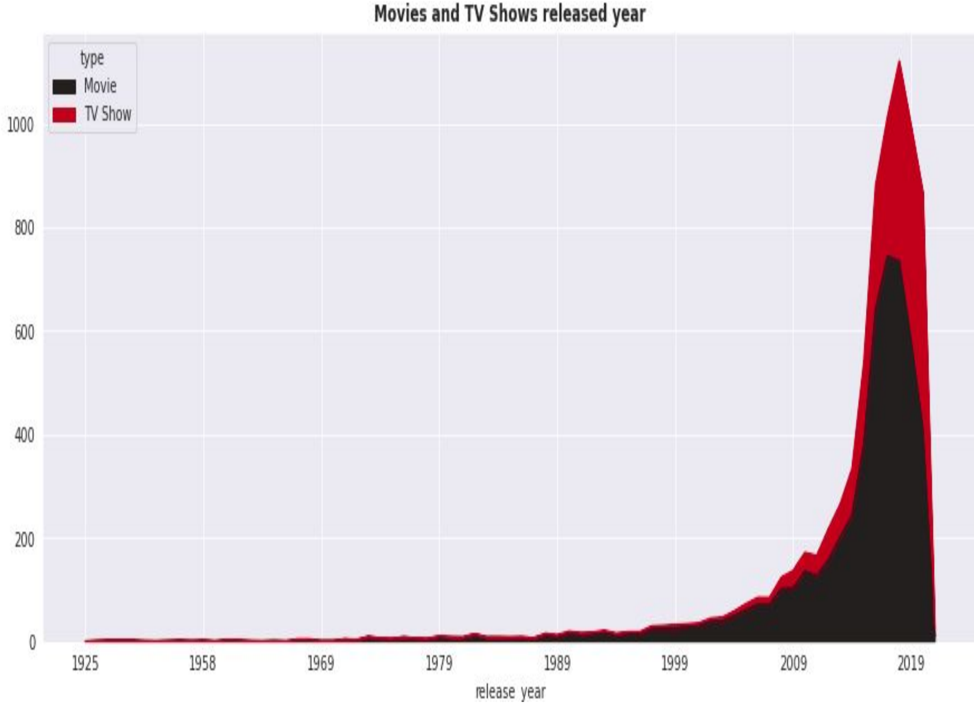
	Type	Count
0	Movie	5377
1	TV Show	2400

Analysis on Movie and TV Shows added over time:

- ➔ We see a slow start for Netflix over several years. Things begin to pick up in 2015 and then there is a rapid increase from 2016.
- ➔ It looks like content additions have slowed down in 2020, likely due to the COVID-19 pandemic.
- ➔ Netflix peak global content amount was in 2019.
- ➔ It appears that Netflix has focused more attention on increasing Movie content that TV Shows.
- ➔ Movies have increased much more dramatically than TV shows.

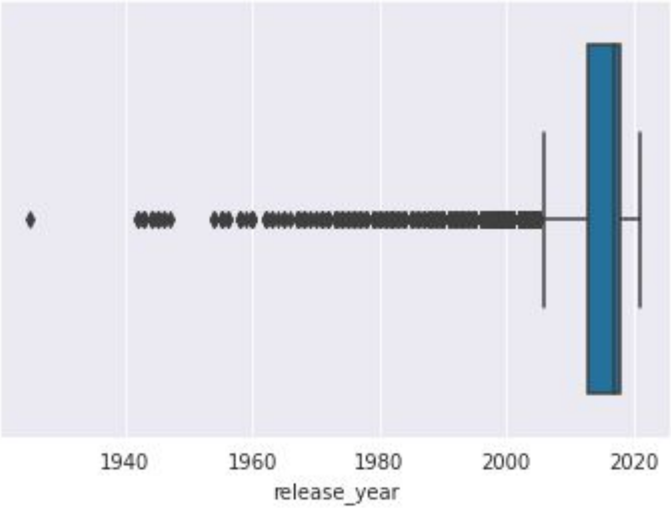


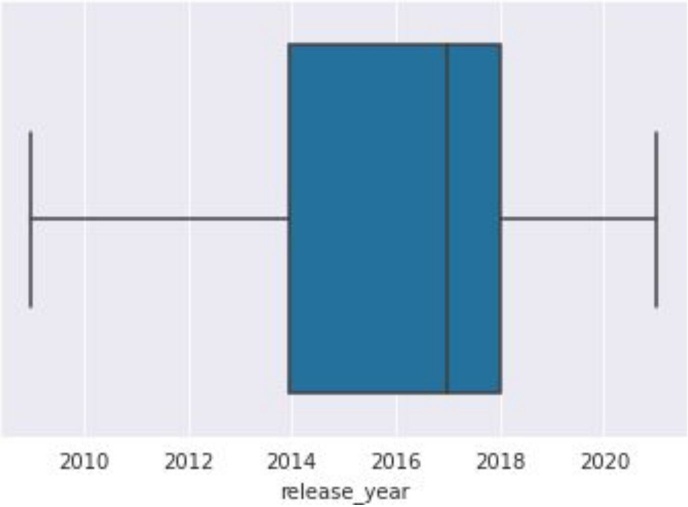
Analysis on Release Year:



- ➔ As we see plot before 2000, movies and tv shows are released very less number and things begin to pick up from 2000 and then there is a rapid increase from 2014.
- ➔ In 2018 maximum number of movies and tv shows are released.
- ➔ In the above area plot it is clearly seen that number of movie released more than TV shows.

Handling Outliers :

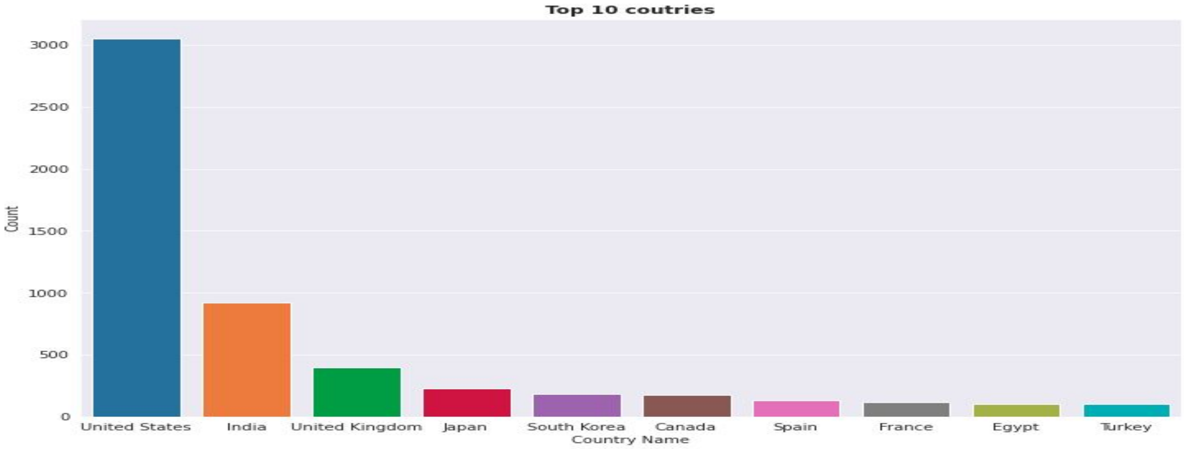


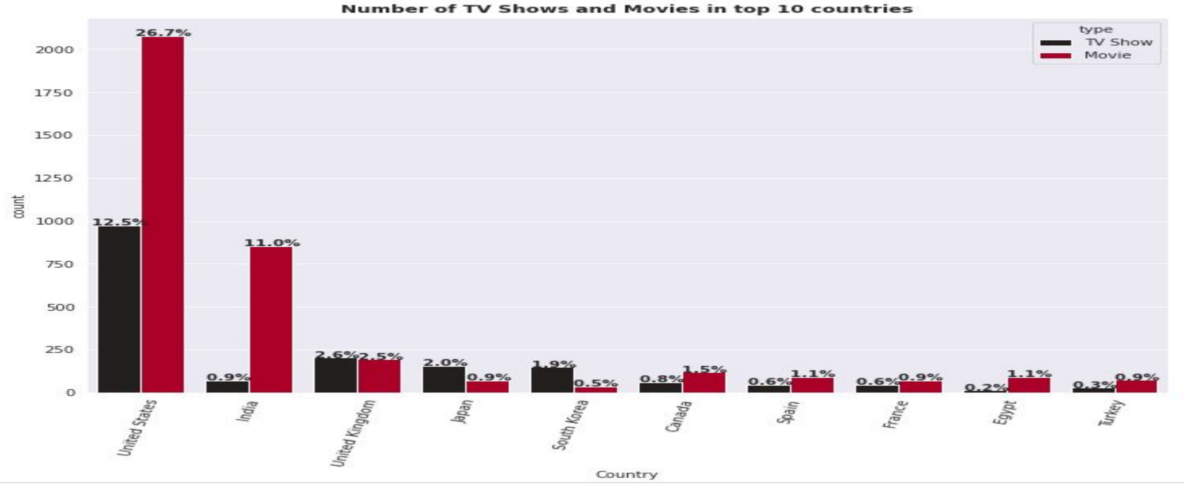


- ➔ The above box plot shows that the outliers in the column **'release_year'.**
- ➔ Which is treated by replacing them with their **mean values.**

Analysis on Top countries with highest content production:

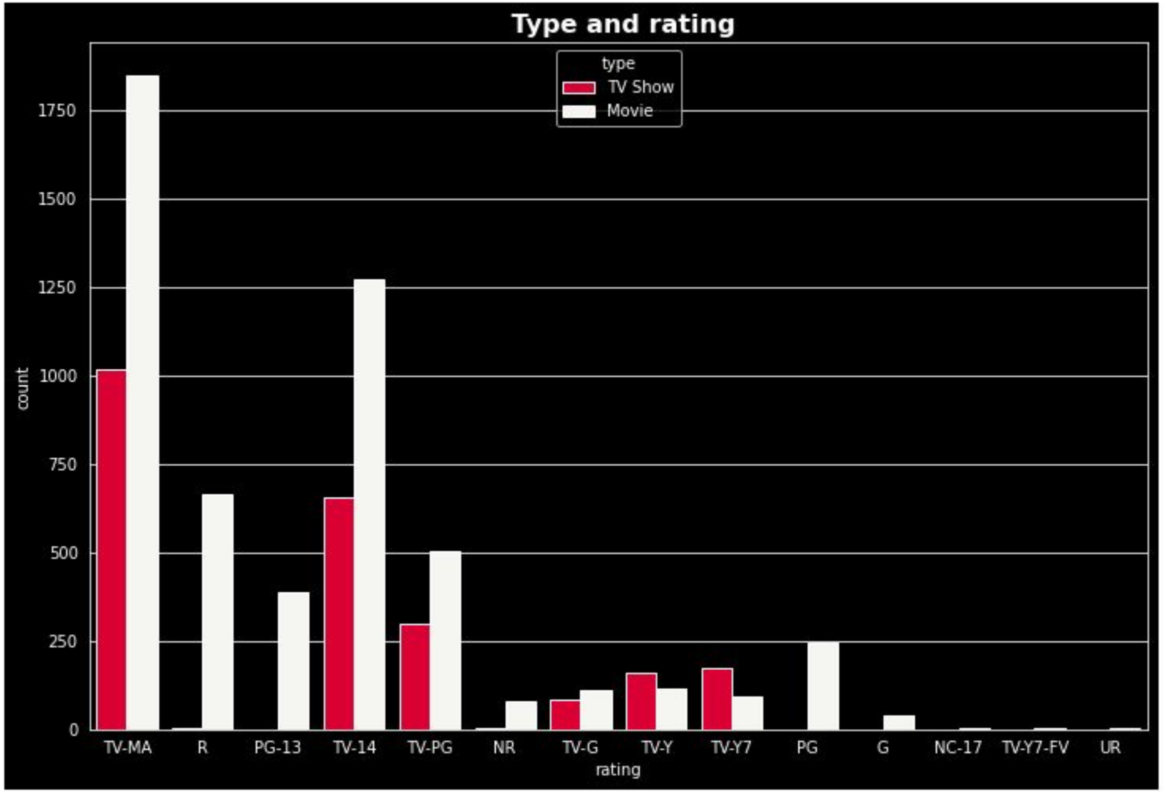
- ➔ **United States has the most number of content on Netflix.**
- ➔ **India has second highest content on Netflix.**
- ➔ **Most of the countries have more movies than TV shows but for South Korea and Japan it's the opposite.**



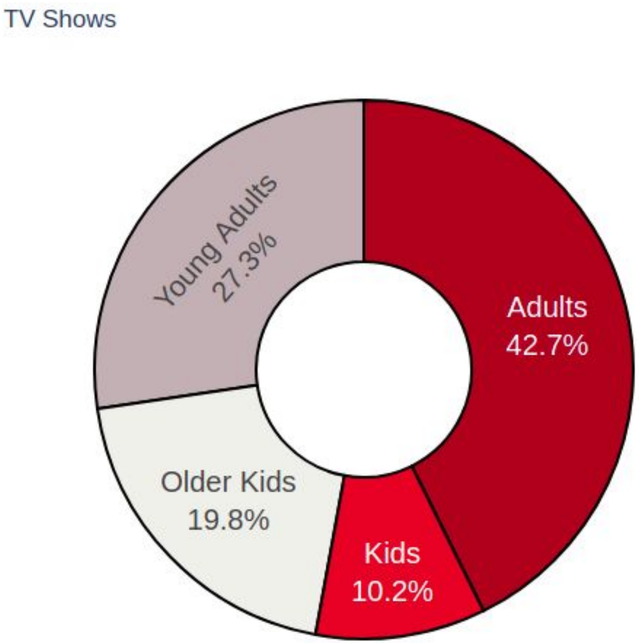


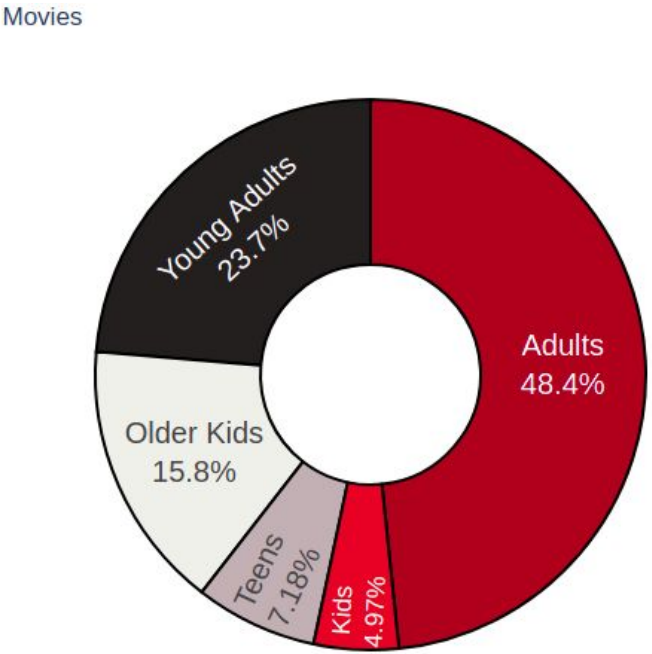
Analysis on Rating:

- ➔ We observe that some ratings are only applicable to Movies. The most common for both Movies & TV Shows are **TV-MA** and **TV-14**
- ➔ Most of the contents got ratings like:
  - ◆ TV-MA (For Mature Audiences)
  - ◆ TV-14 (May be unsuitable for children under 14)
  - ◆ TV-PG (Parental Guidance Suggested)
  - ◆ NR (Not Rated)



Analysis on what age groups are content created:






```
MR_age = {'TV-MA': 'Adults',
          'R': 'Adults',
          'PG-13': 'Teens',
          'TV-14': 'Young Adults',
          'TV-PG': 'Older Kids',
          'NR': 'Adults',
          'TV-G': 'Kids',
          'TV-Y': 'Kids',
          'TV-Y7': 'Older Kids',
          'PG': 'Older Kids',
          'G': 'Kids',
          'NC-17': 'Adults',
          'TV-Y7-FV': 'Older Kids',
          'UR': 'Adults'}
```

➔ **In Movies and TV shows mostly contents are in Adults and Young Adults age group.**

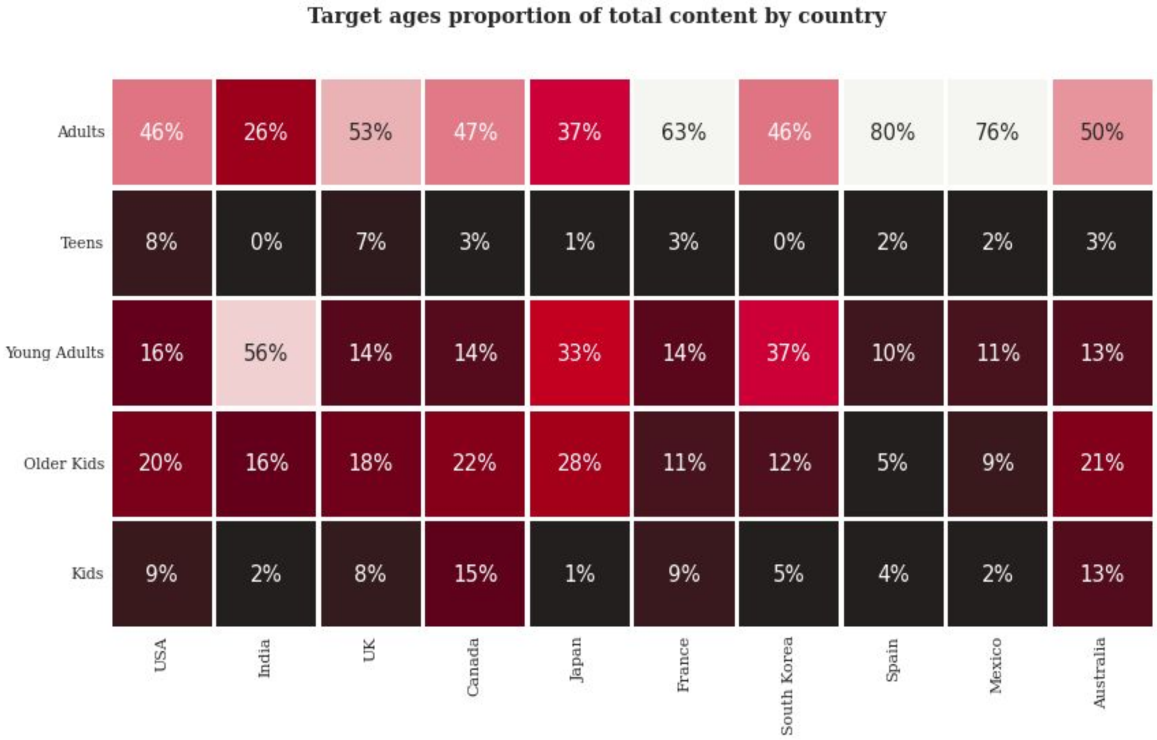
➔ **Very less contents for kids age group.**

** To compare the rating and the age group used information from* prime video

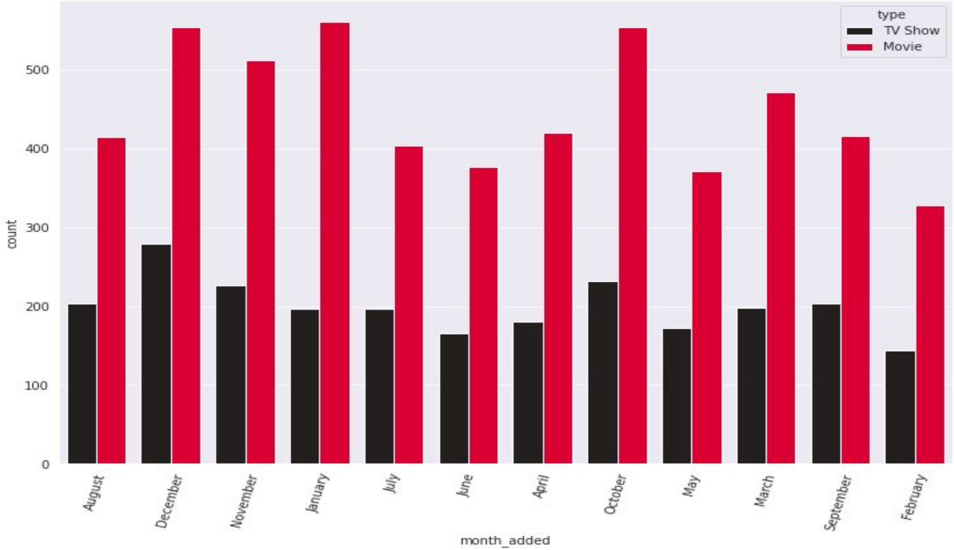
Analysis on Target ages proportion of total content by country:

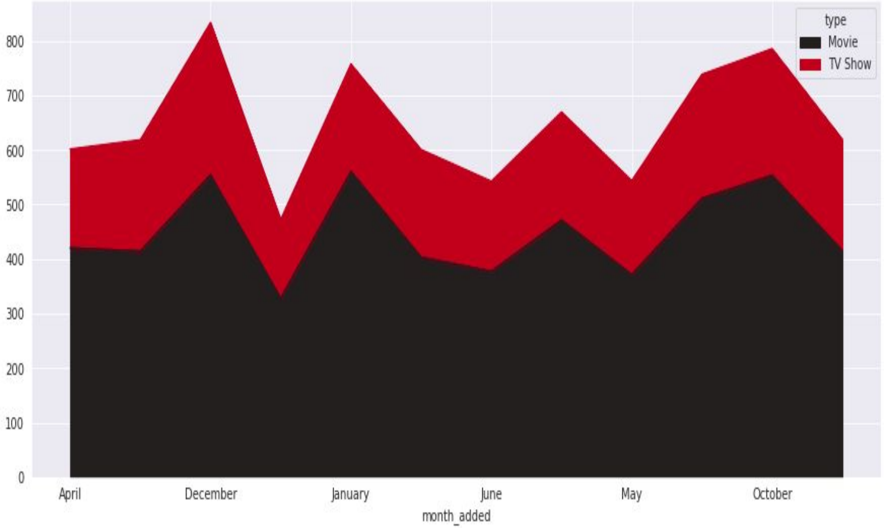
- ➔ It is also interesting to see parallels between culturally comparable nations - the **USA** and **UK** are closely aligned with their Netflix target ages, but radically different from, example, **India** or **Japan!**

- ➔ Also, **Mexico** and **Spain** have similar content on Netflix for different age groups.



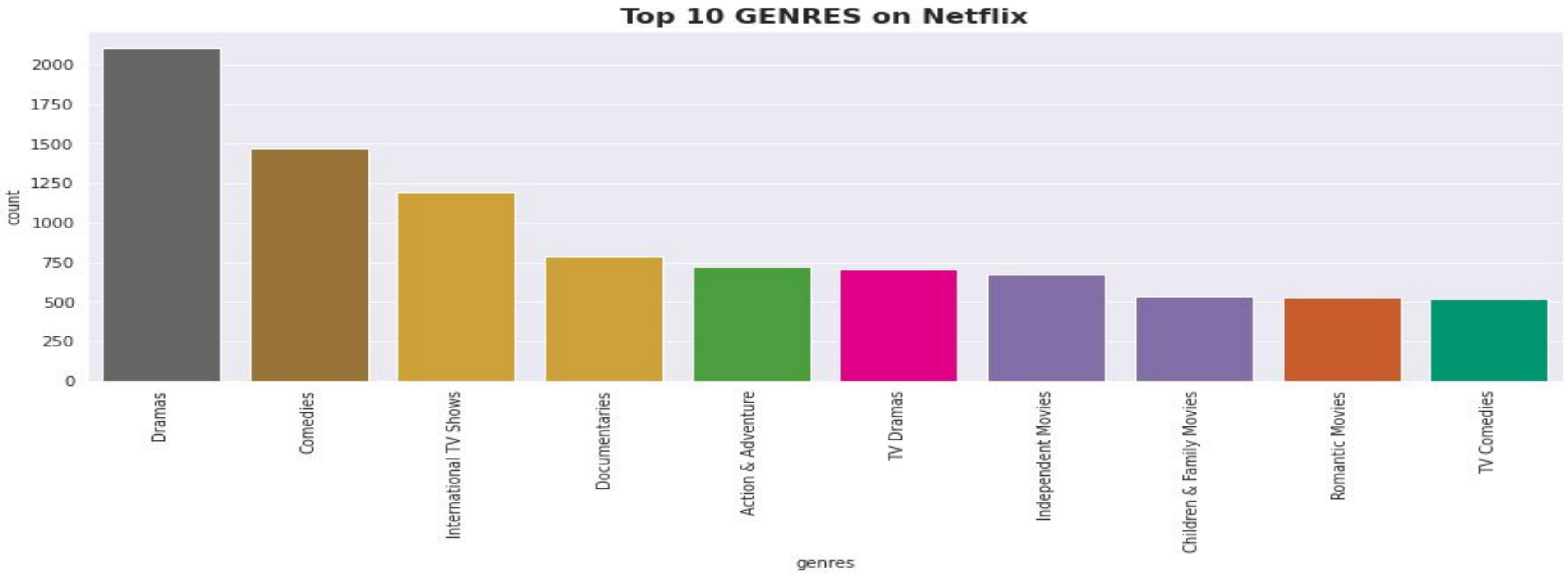
Analysis on content added by month:





- ➔ The end & beginnings of each year seem to be Netflix's preference for adding content.
- ➔ December & January are definitely the best months for new content.
- ➔ December has the highest number of contents followed by october and january reason could be December is the holiday season and it also has Christmas, so there is high possibility that most of the contents upload in this month.
- ➔ February is the worst.

Analysis on Genres on Netflix:

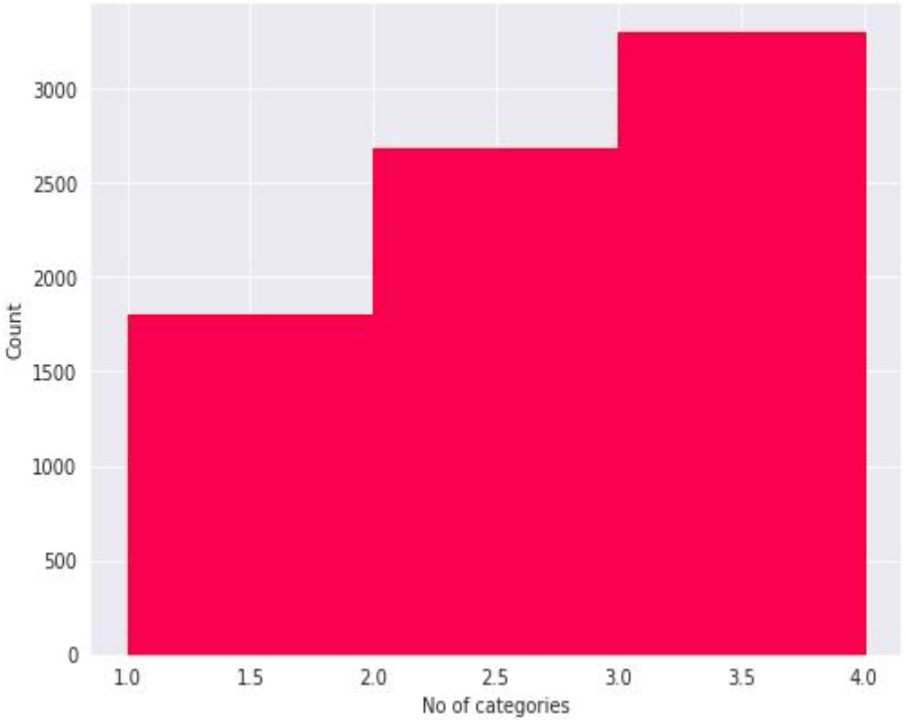


➔ Drama is the most popular genre followed by comedy.

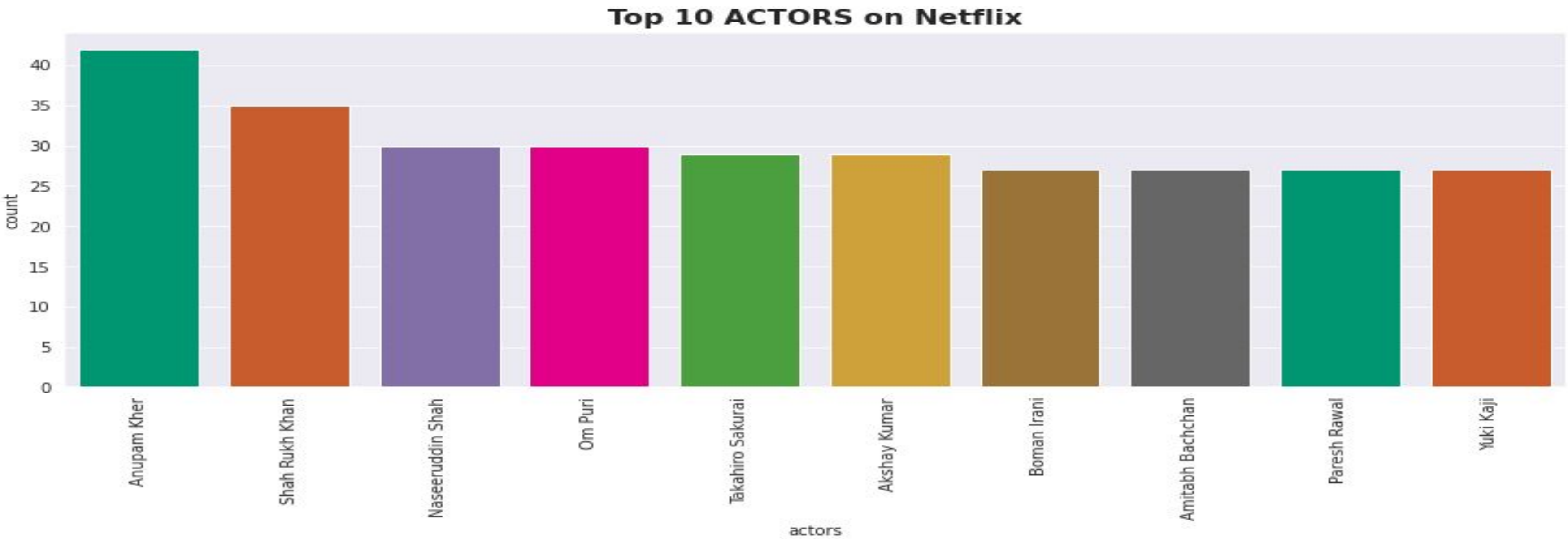
Categories present in each content:

➔ As we see the Histogram graph we can see that there are 3 unique categories of contents with their count values and we bin the values for better clarity, like there are International TV Shows', ' TV Dramas' ' TV Sci-Fi & Fantasy' contents so we bin their count values for better visualization

➔ Most of the movies are belonging to category three.

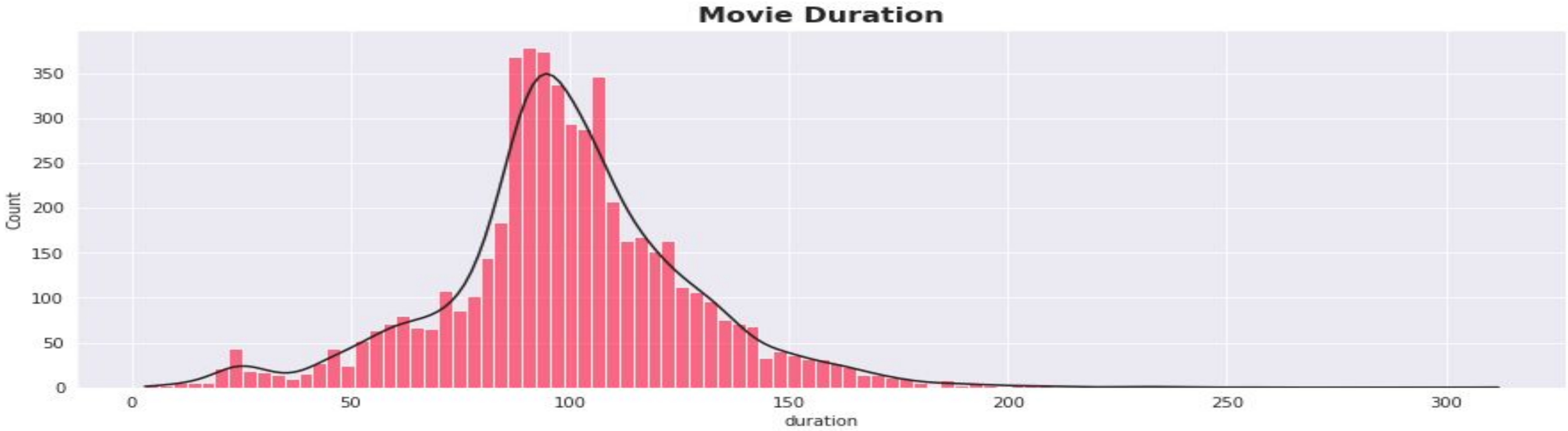


Analysis on Top actors on Netflix:

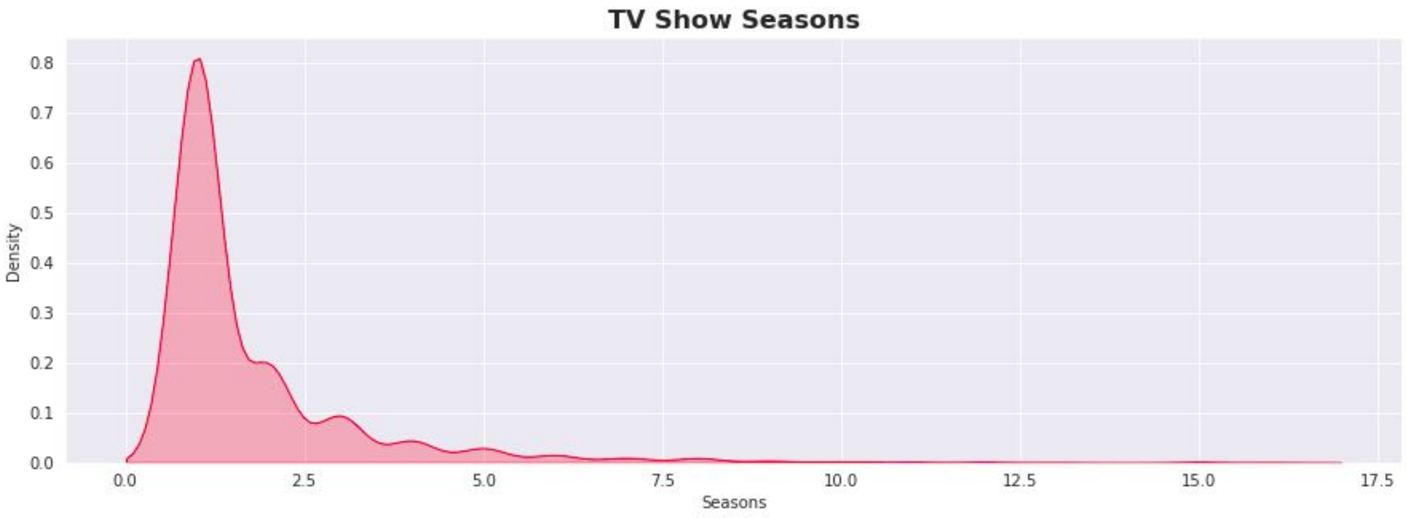


➔ **Anupam Kher Have the most number of films on Netflix.**

Analysis on Movie & TV Show duration:



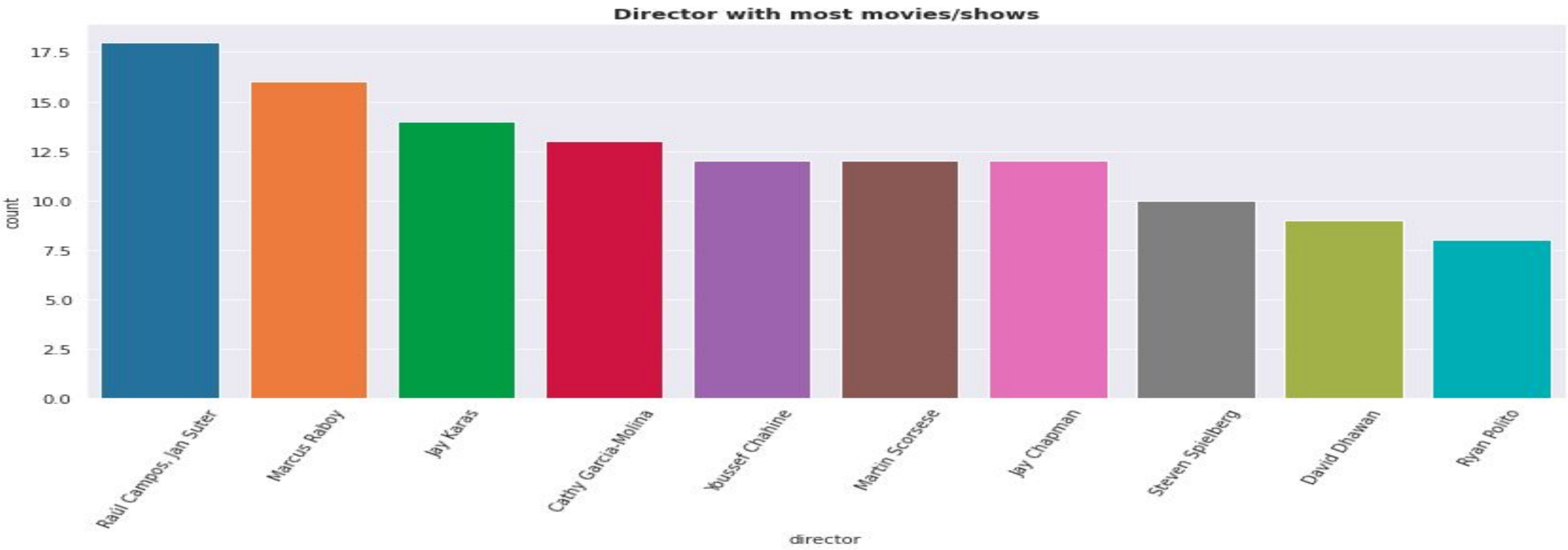
- ➔ Above histogram plot, we can see that the duration for Netflix movies closely resembles a normal distribution with the average viewing time spanning about 90 minutes which seems to make sense.
- ➔ Most content are about 70 to 120 min duration for movies.



	seasons	count
0	1	1608
1	2	378
2	3	183
3	4	86
4	5	57
5	6	30
6	7	19
7	8	18
8	9	8
9	10	5
10	11	2
11	12	2
12	15	2
13	13	1
14	16	1

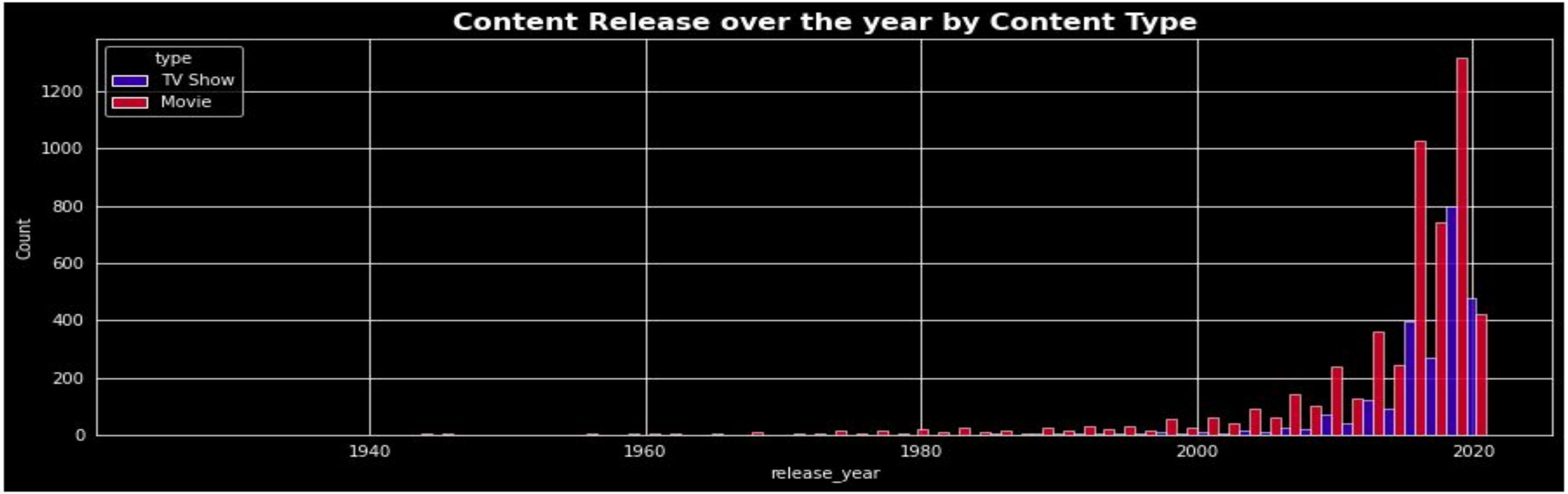
➔ **From above we see that Netflix TV shows on the other hand seems to be heavily skewed to the right or positively skewed where the majority of shows only have 1 season.**

Analysis on Top Directors on Netflix:



➔ **Raul Campos and Jan Suter collectively have the most content on Netflix.**

Is Netflix has increasingly focusing on TV Show rather than movies in recent years ?



Yes, Netflix is increasingly focusing on TV Shows now, which is clear from the graph, in 2020, there were more Shows than Movies. Also, Movie's preference shows a declining graph, while shows are increasing.

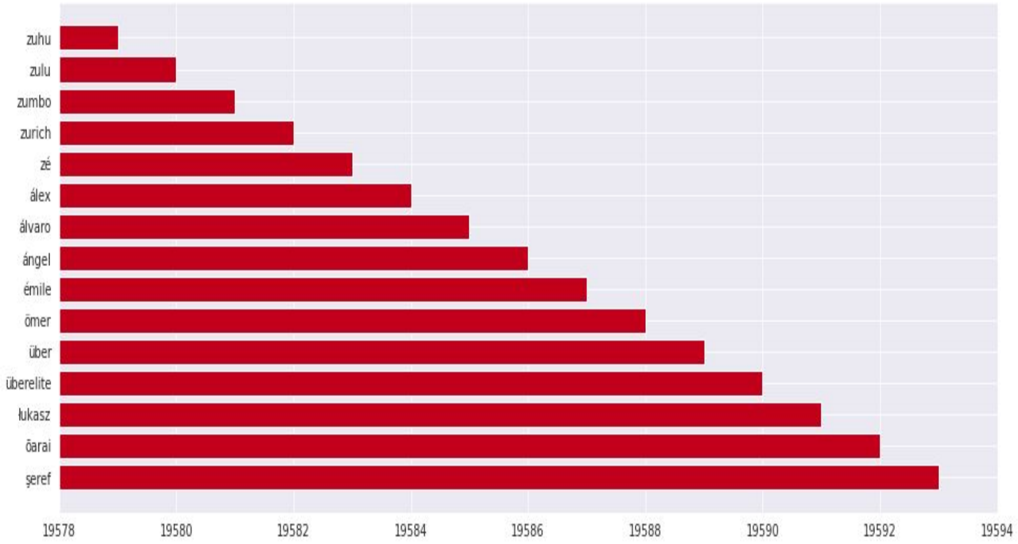
Analysis on Titles:

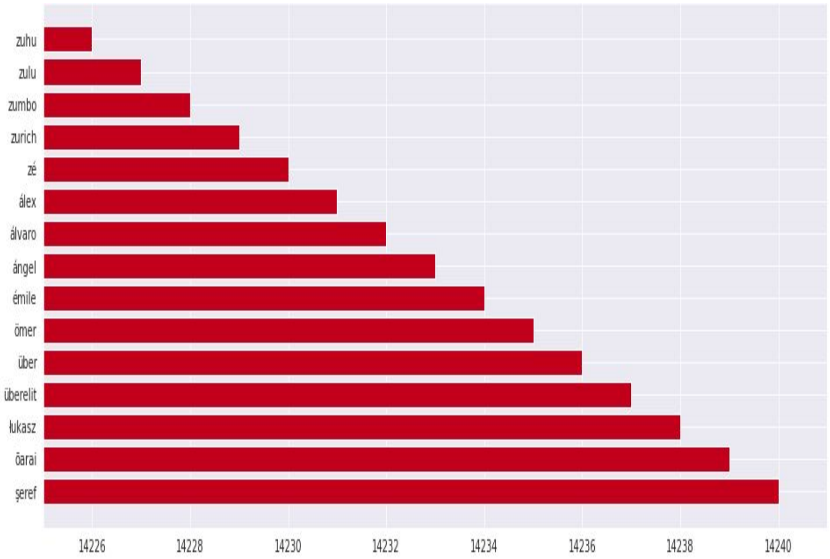
➔ A word cloud (also known as a tag cloud) is a visual representation of words. Cloud creators are used to highlight popular words and phrases based on frequency and relevance.

➔ It seems like words like "Love", "Man", "World", "Story" , "Christmas" are very common in titles.

➔ I have surprised to see "Christmas" occurred so many time . The reason maybe those movies released on the month of december, but I don't have any information about the release month of movies that's why I am not able to check my hypothesis.

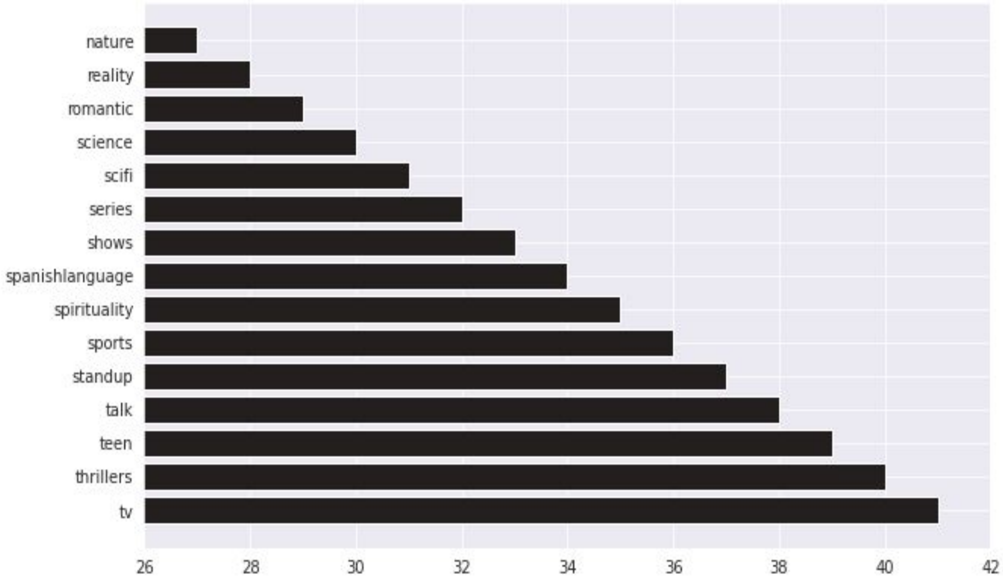
Before & After Stemming most occurred words(description):

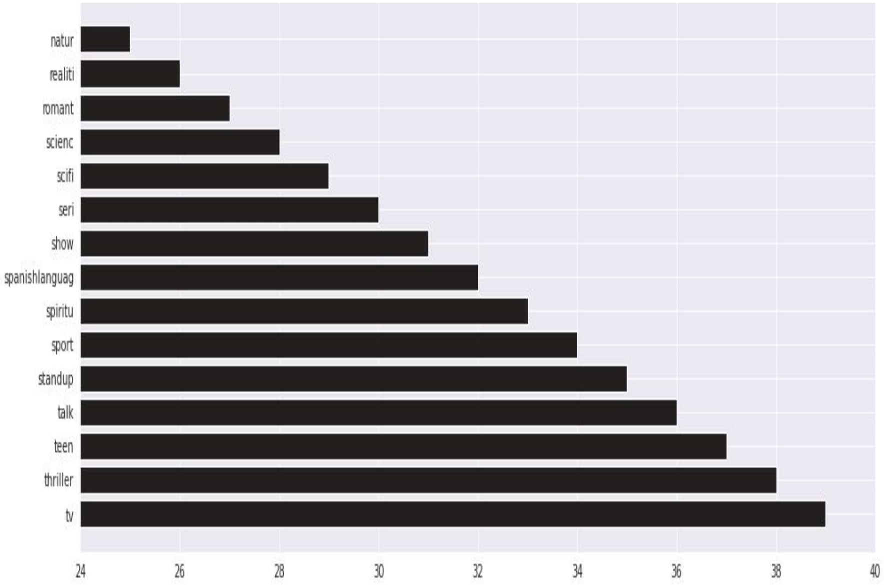




- ➔ Before stemming its ranges between 19578-19594 and after stemming its reduced to 14240.
- ➔ So basically Stemming is a technique used to extract the base form of the words by removing affixes from them. It is just like cutting down the branches of a tree to its stems.

Before & After Stemming most occurred words(listed_in):





➔ **All the inflected words has removed after applying stemming technique.**

Feature Selection & ML algo used:

- **Only selected 3 features , to do CLUSTERING:**
  - no_of_category
  - Length(description)
  - Length(listed-in)

- **Using STANDARDSCALER**

- **Using K-Means and Agglomerative Clustering**

- **Used the following method to find out best k value**
  - Silhouette score
  - Elbow Method
  - Dendrogram

Silhouette Score:

Silhouette score is used to evaluate the quality of clusters created using clustering algorithms such as K Means in terms of how well samples are clustered with other samples that are similar to each other.

Silhouette Coefficient Formula $S = \frac{(b-a)}{max(a,b)}$.

➔ 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.

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.

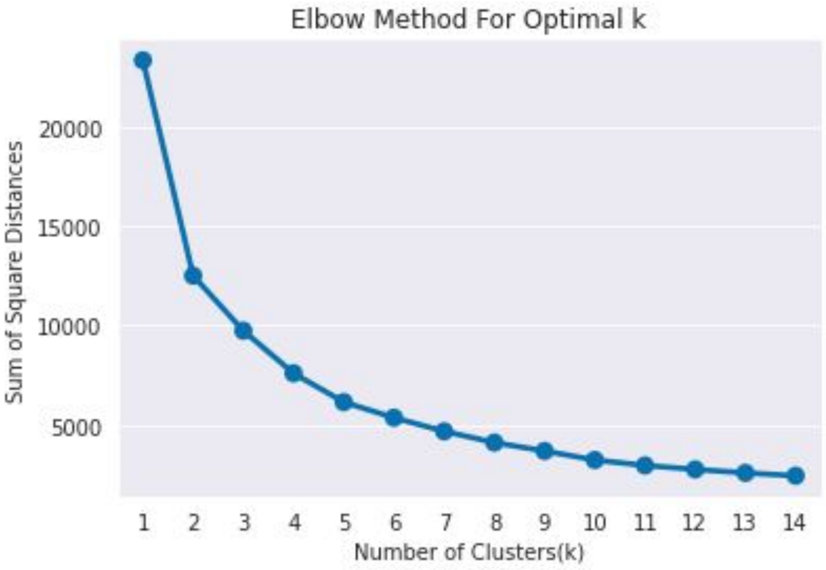
➔ The worst value is -1

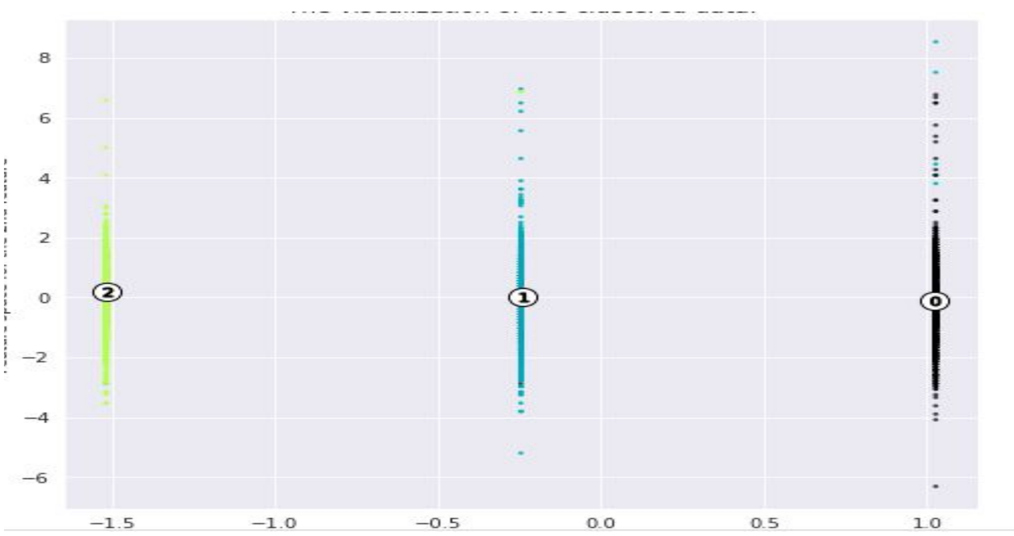
➔ If score is 0 denotes overlapping clusters

n_clusters	silhouette_score
2	0.428
3	0.383
4	0.374
5	0.372
6	0.367
7	0.353
8	0.369
9	0.374
10	0.363
11	0.355
12	0.351
13	0.355
14	0.334
15	0.341

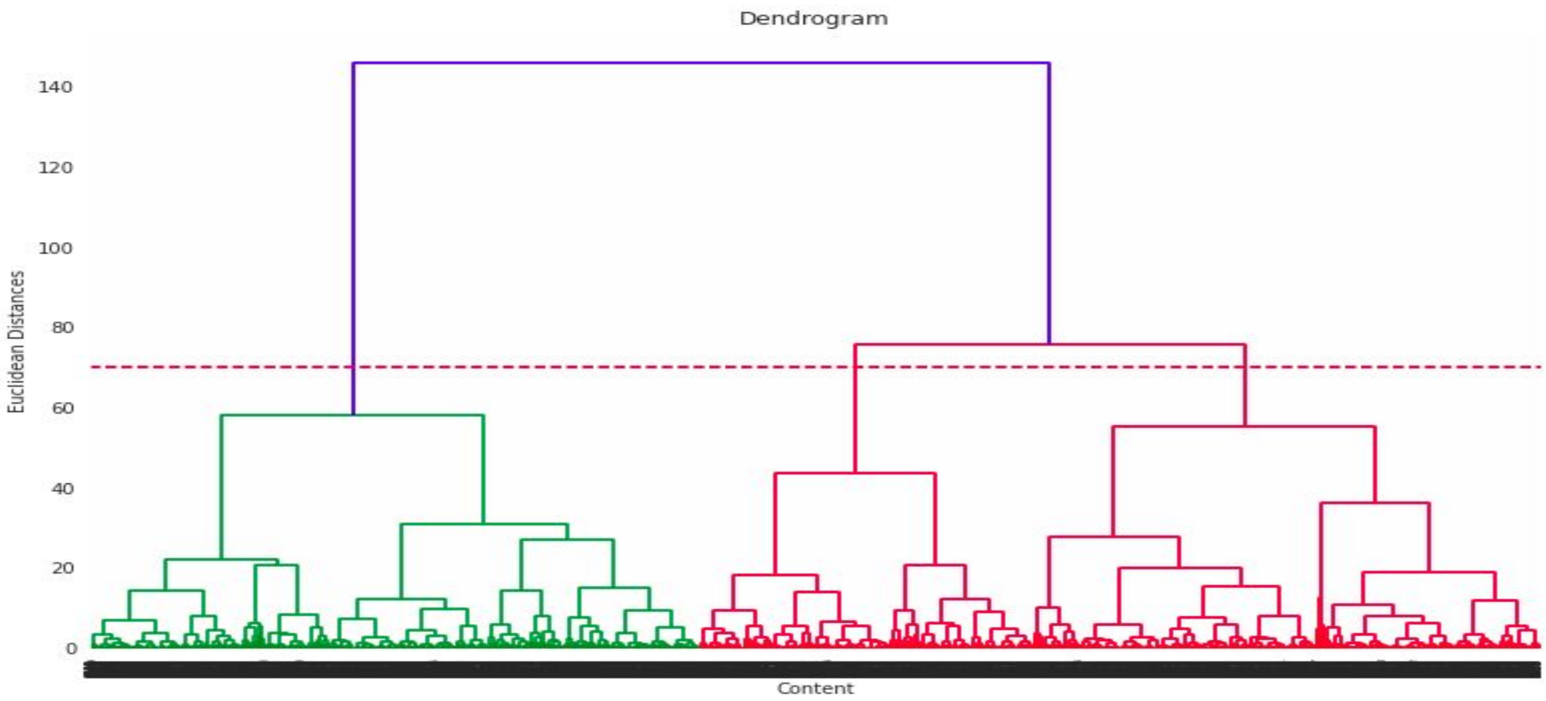
Elbow Method:

- ➔ The Elbow Curve is one of the most popular methods to determine this optimal value of k.
- ➔ The elbow curve uses the sum of squared distance (SSE)to choose an ideal value of k based on the distance between the data points and their assigned clusters.
- ➔ 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.





Dendrogram:

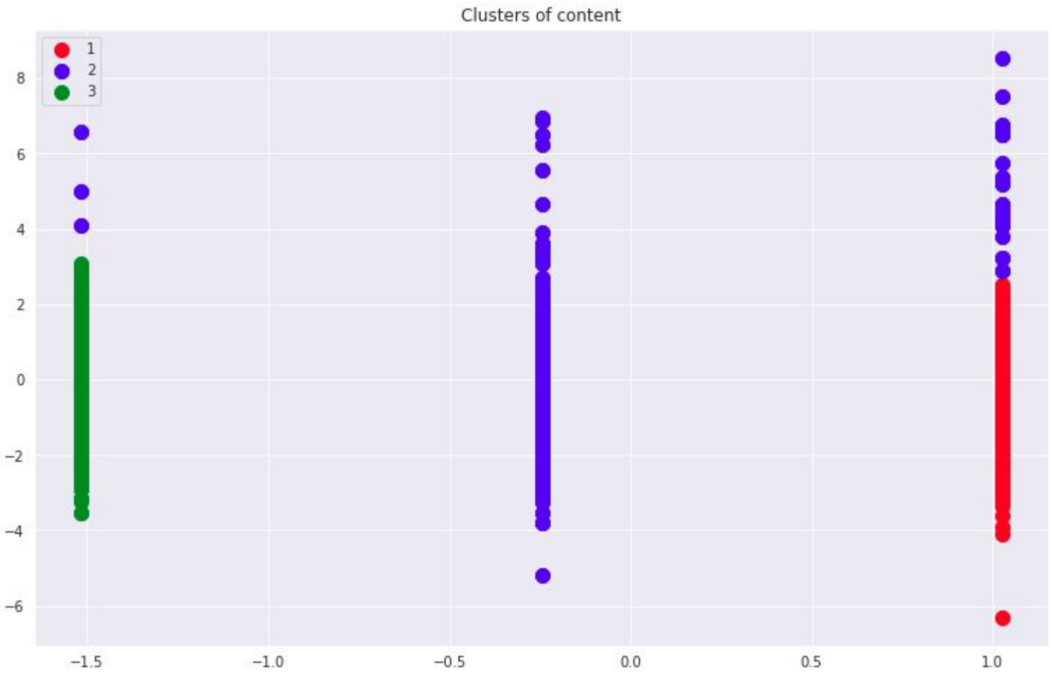


➔ As we can see in the **dendrogram** that horizontal line is cutting 3 vertical lines so, the number of **clusters** we will choose is **3.**

Agglomerative Clustering:

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:

- When we look at the dataset we can clearly see that out of all the given titles 69.1% of them were movies and the rest 30.9% were TV Shows.
- We have reached a conclusion from our analysis from the content added over years that Netflix is more focusing on movies than TV Shows and We see a slow start for Netflix over several years things begin to pick up in 2015 and then there is a rapid increase from 2016.
- We also observe that from release_year(Actual Release Year of the movie / show) that more number of Movies release than TV Shows.
- The most prolific producers of content for Netflix are primarily, the USA, with India and the UK a significant distance behind.
- As I've noted in the insights on the plot, it is really interesting to see how the split of TV Shows and Movies varies by country. South Korea and Japan is dominated by TV Shows and Equally, India is dominated by Movies.
- Looking at the Genres we can see that Drama is the most popular genre followed by comedy.
- The largest count of Netflix content is made with a “TV-MA” rating.
- When we look at the cast for Movies we can see many Indian actors like Anupam Kher,Shah Rukh Khan, Naseeruddin Shah, OM Puri have the most number of films on Netflix
- In text analysis (NLP) I removed stop words, removed punctuations , stemming & TF-IDF vectorizer and other functions of NLP.
- Applied different clustering models like Kmeans, hierarchical 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