CAPSTONE PROJECT 4

NETFLIX MOVIES AND TV SHOWS CLUSTERING

Unsupervised ML Classification Model

Team Members

Anas Mustafa

Chetan Rajput

Sarthak Rastogi

Content

- Introduction
- Defining Problem Statement
- Data Summary
- Abstract
- Handling Null Value
- EDA
- Hypothesis Testing
- Data Preprocessing
- Final Model
- Challenges
- Conclusion

Introduction

- Netflix is a prominent OTT platform with a wide variety of content to view from a variety of nations and genres, so keep an eye on it.
- As of 2019, the dataset contains TV shows and movies available on Netflix. Fixable, a third-party Netflix search engine, provided the data for this study.
- The purpose is to forecast clusters based on similar content by comparing text-based features, in this example, the description column, which is a brief graphic overview of the contents.

PROBLEM STATEMENT

1. 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.

2. 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.

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

ABSTRACT

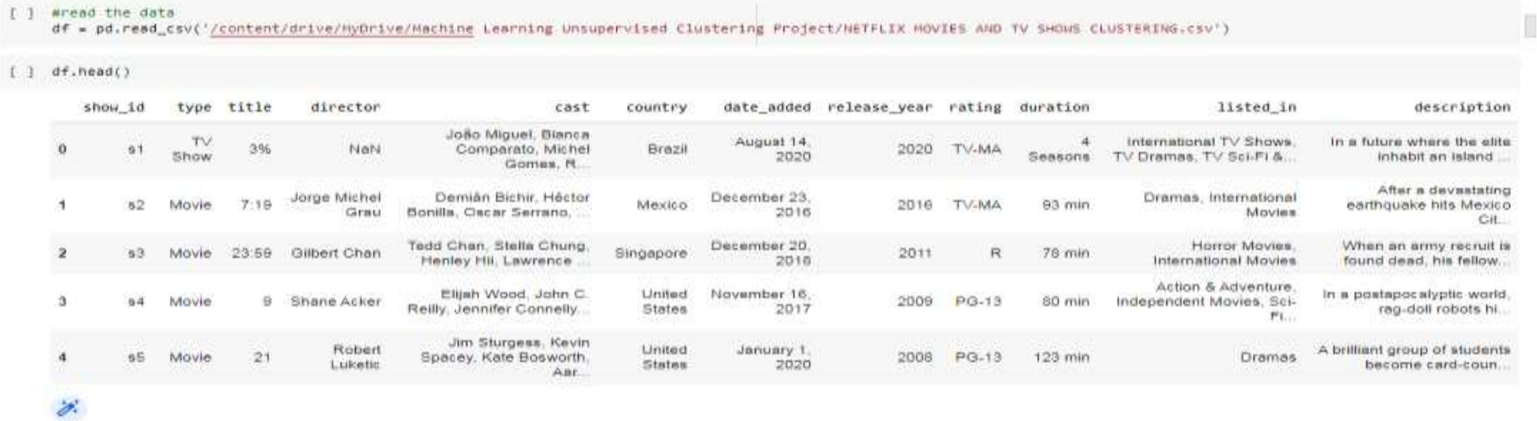
1. The idea was to use text-based variables to anticipate clusters of related content.
2. The dataset is subjected to exploratory data analysis in order to extract insights from it, but the initial null results are ignored.
3. In addition, using EDA's findings, some hypothesis testing was done.
4. After that, our target variable, the description column, must be feature engineered, with NLP operations such as symbol removal, stop words, punctuation, tokenization, and vectorization using TFIDF done on it.
5. All that was left was to discover the clusters, fit our models based on the number of clusters, and evaluate the model using evaluation metrics.

Data Summary

- 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 : Genere.
- Description: The Summary description.

Basic Data Exploration

- This dataset consists of TV shows and movies available on Netflix as of 2019.
- Dataset contains 7787 rows & 12 columns.
- It will be interesting to explore what all other insights can be obtained from the same dataset
- There are null values in four columns.



```
[ ] #read the data
    df = pd.read_csv('/content/drive/MyDrive/Machine Learning Unsupervised Clustering Project/NETFLIX MOVIES AND TV SHOWS CLUSTERING.csv')
```

```
[ ] df.head()
```

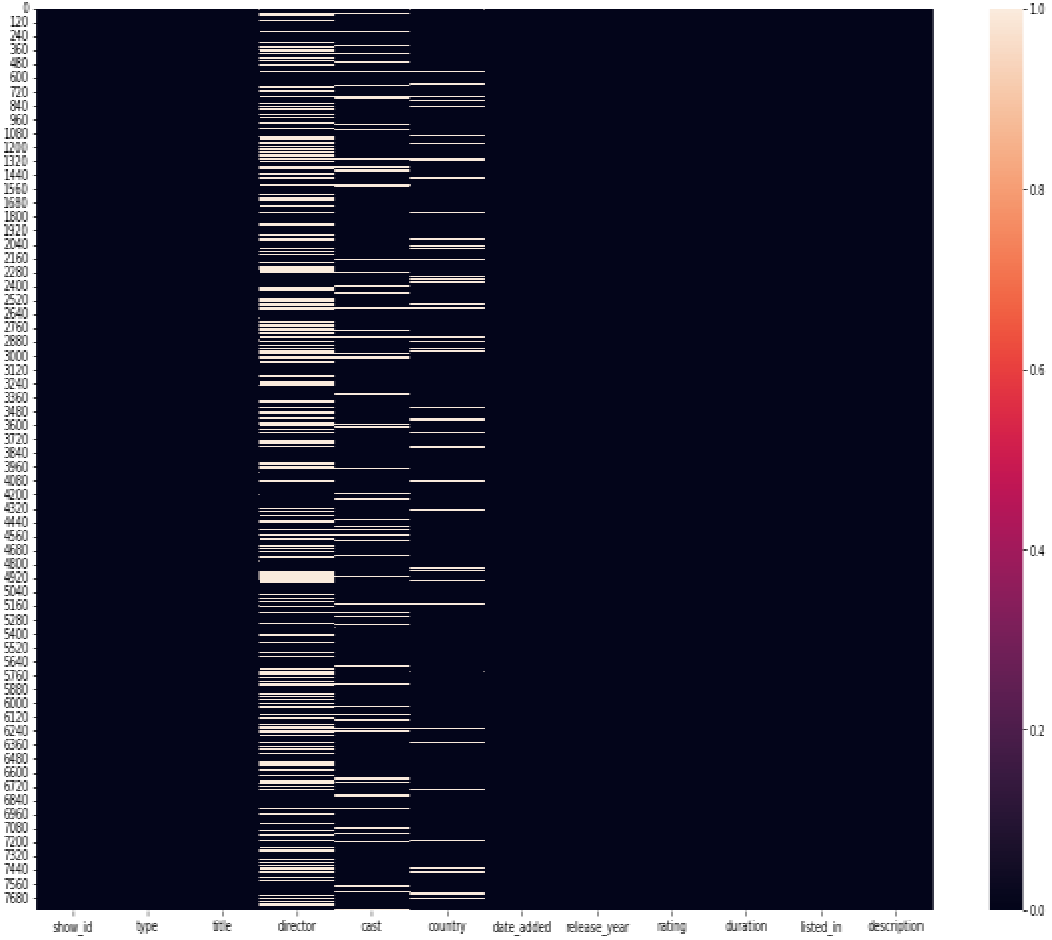
	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	s1	TV Show	3%	NaN	João Miguel, Bianca Comparato, Michel Gomes, R...	Brazil	August 14, 2020	2020	TV-MA	4 Seasons	International TV Shows, TV Dramas, TV Sci-Fi &...	In a future where the elite inhabit an island ...
1	s2	Movie	7:19	Jorge Michel Grau	Demián Bichir, Héctor Bonilla, Oscar Serrano, ...	Mexico	December 23, 2016	2016	TV-MA	93 min	Dramas, International Movies	After a devastating earthquake hits Mexico Cit...
2	s3	Movie	23:59	Gilbert Chan	Tedd Chan, Stella Chung, Henley Hii, Lawrence ...	Singapore	December 20, 2018	2011	R	78 min	Horror Movies, International Movies	When an army recruit is found dead, his fellow...
3	s4	Movie	9	Shane Acker	Elijah Wood, John C. Reilly, Jennifer Connelly...	United States	November 16, 2017	2009	PG-13	80 min	Action & Adventure, Independent Movies, Sci-Fi...	In a postapocalyptic world, rag-doll robots hi...
4	s5	Movie	21	Robert Luketic	Jim Sturgess, Kevin Spacey, Kate Bosworth, Aar...	United States	January 1, 2020	2008	PG-13	123 min	Dramas	A brilliant group of students become card-coun...

Handling Null Values

- We checked for null values after loading the dataset and removed the null values, as well as some unnecessary columns.

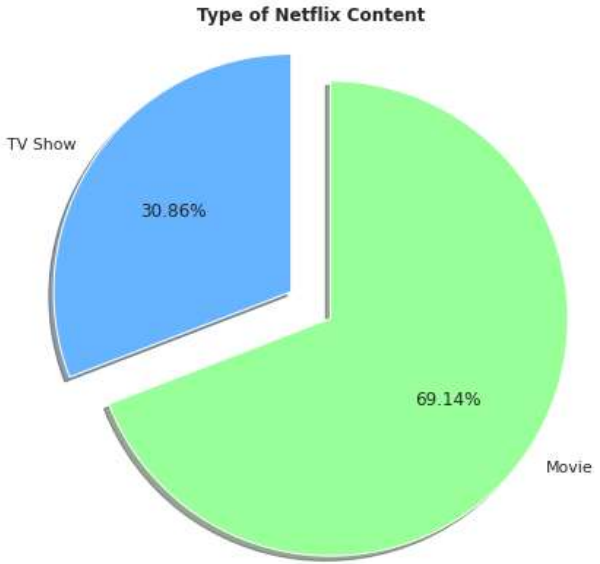
Null Value Treatment

- **RATING & COUNTRY** - Replacing nulls with mode
- **CAST**- Replacing nulls with unknown.
- **DATE** - there are few missing values for date column. so, lets drop missing value rows.
- **DIRECTOR** - Director column has more then 30% null values so we will not use it for our model but will keep it for EDA - Replacing nulls with 'unknown'



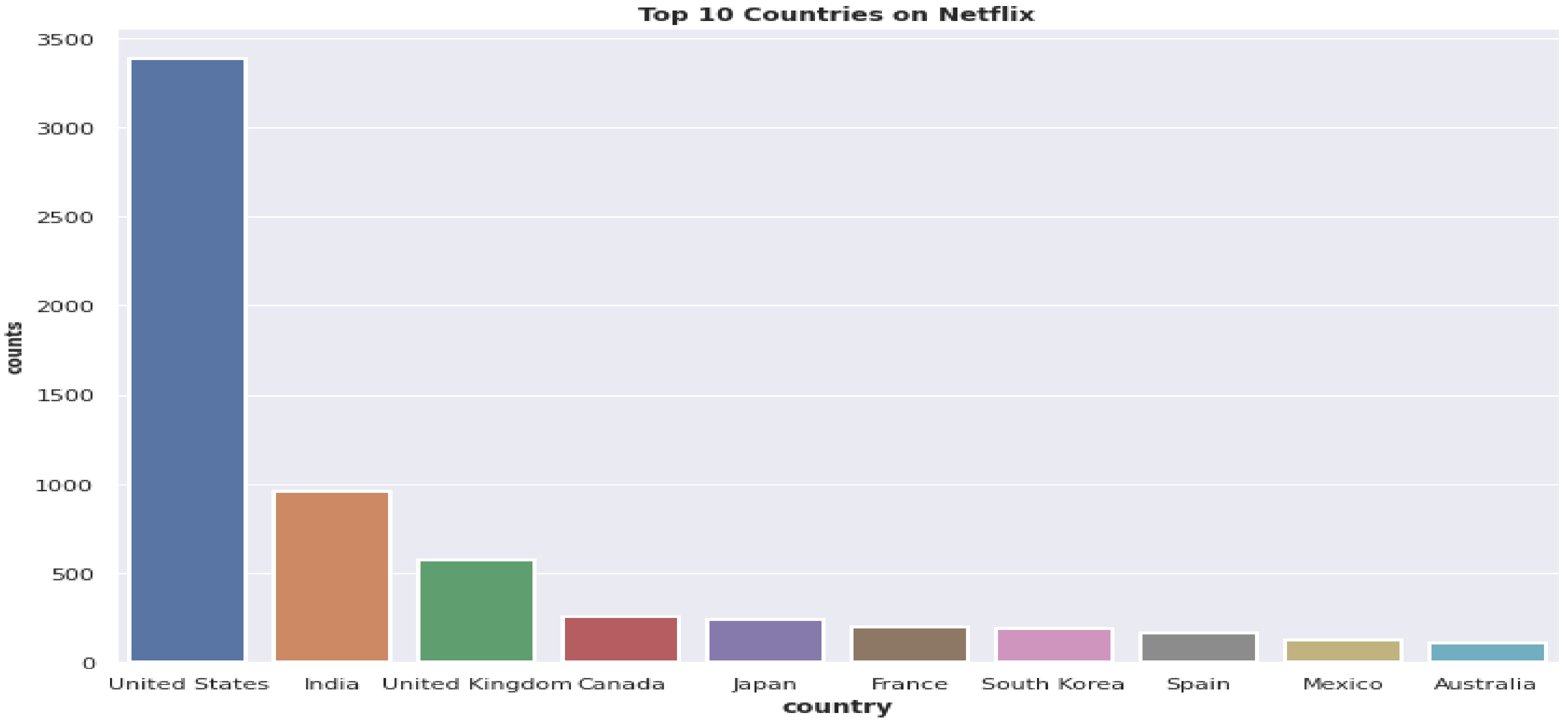
Exploratory Data Analysis

TYPE OF CONTENT



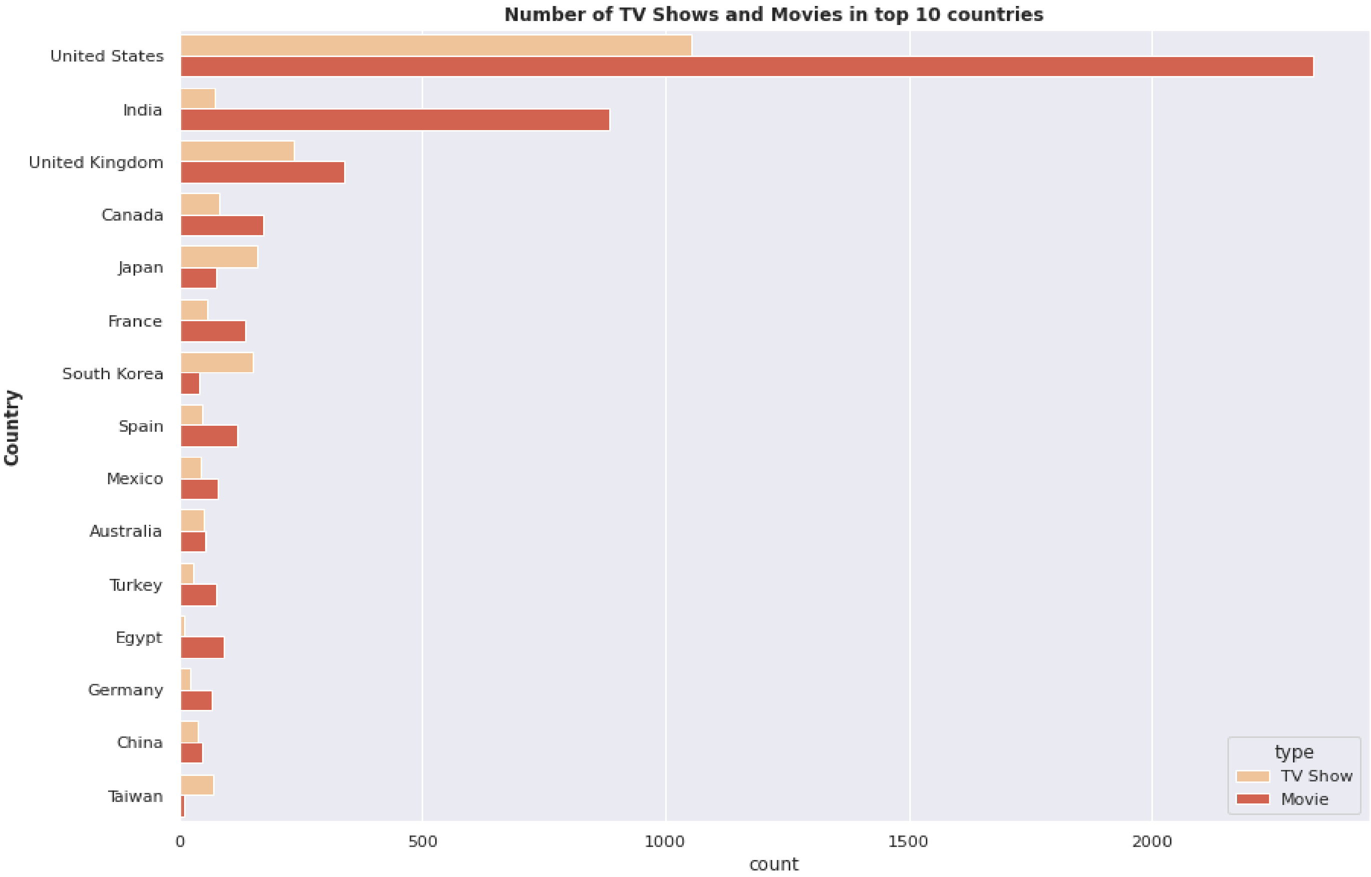
- There are about 70% movies and 30% TV shows on Netflix.

EDA- Top 10 Countries



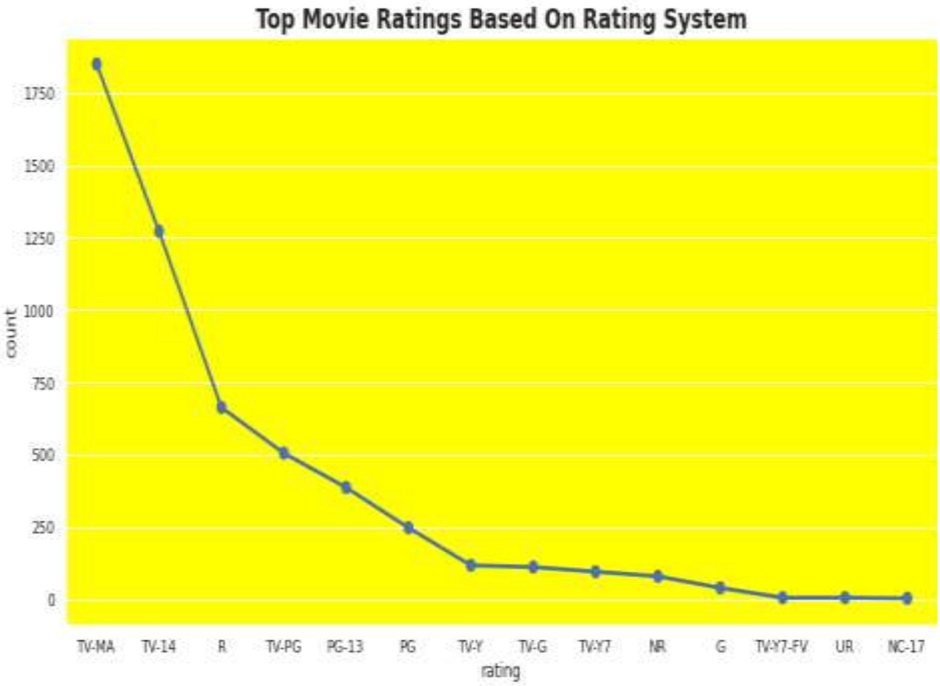
The United States has the most number of content on Netflix by a huge margin followed by India.

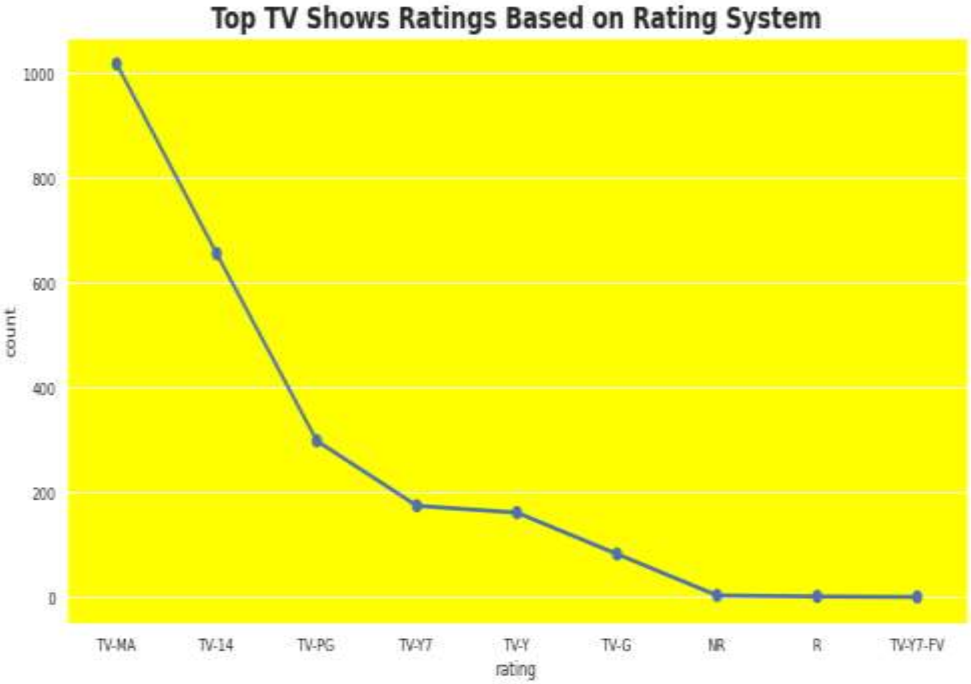
Number Of Tv Shows And Movies In Top 10 Countries



Most of the countries have more movies than TV shows but for South Korea and Japan it's the opposite. It maybe because K-Dramas and Anime are more popular in these two countries respectively.

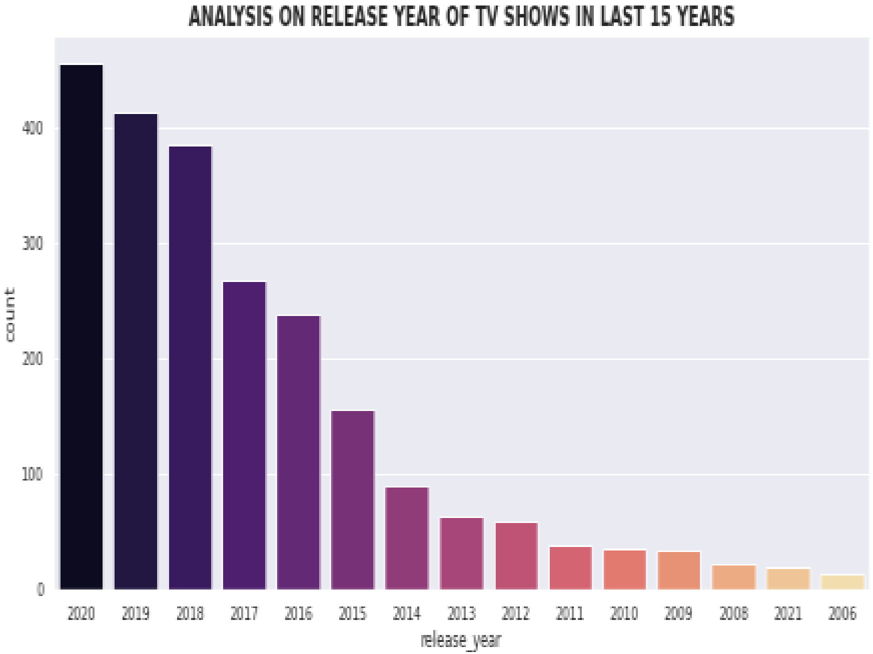
Top Movies & TV Shows Rating Based

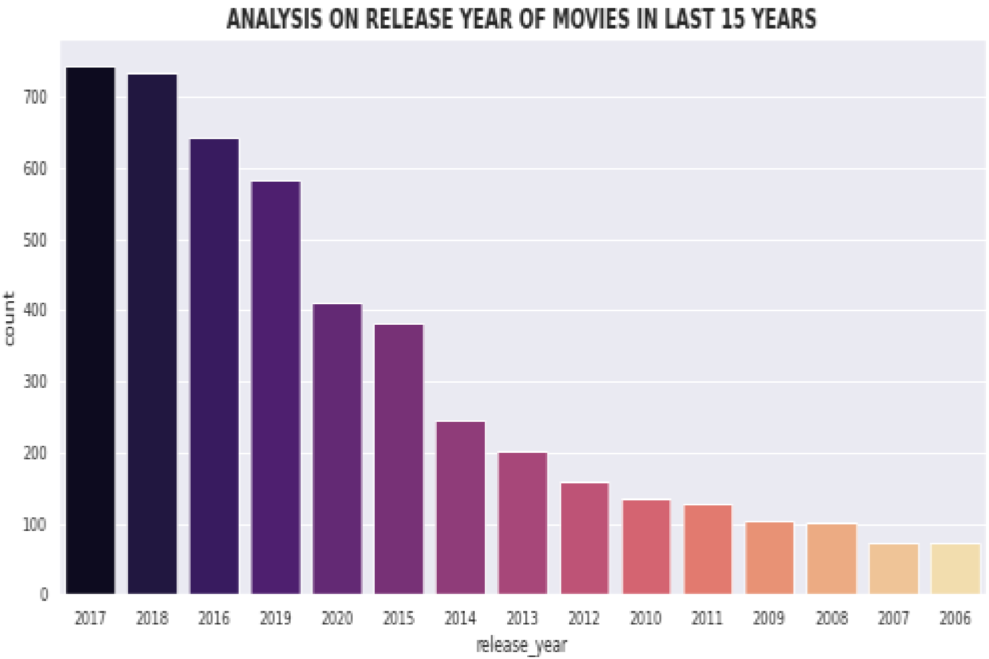




- Most number of movies rated TV-MA i.e. Adult Rating.
- Most number of TV Shows rated TV- MA i.e. Adult Rating.

Number of TV Shows & Movies Release in last 15 Years

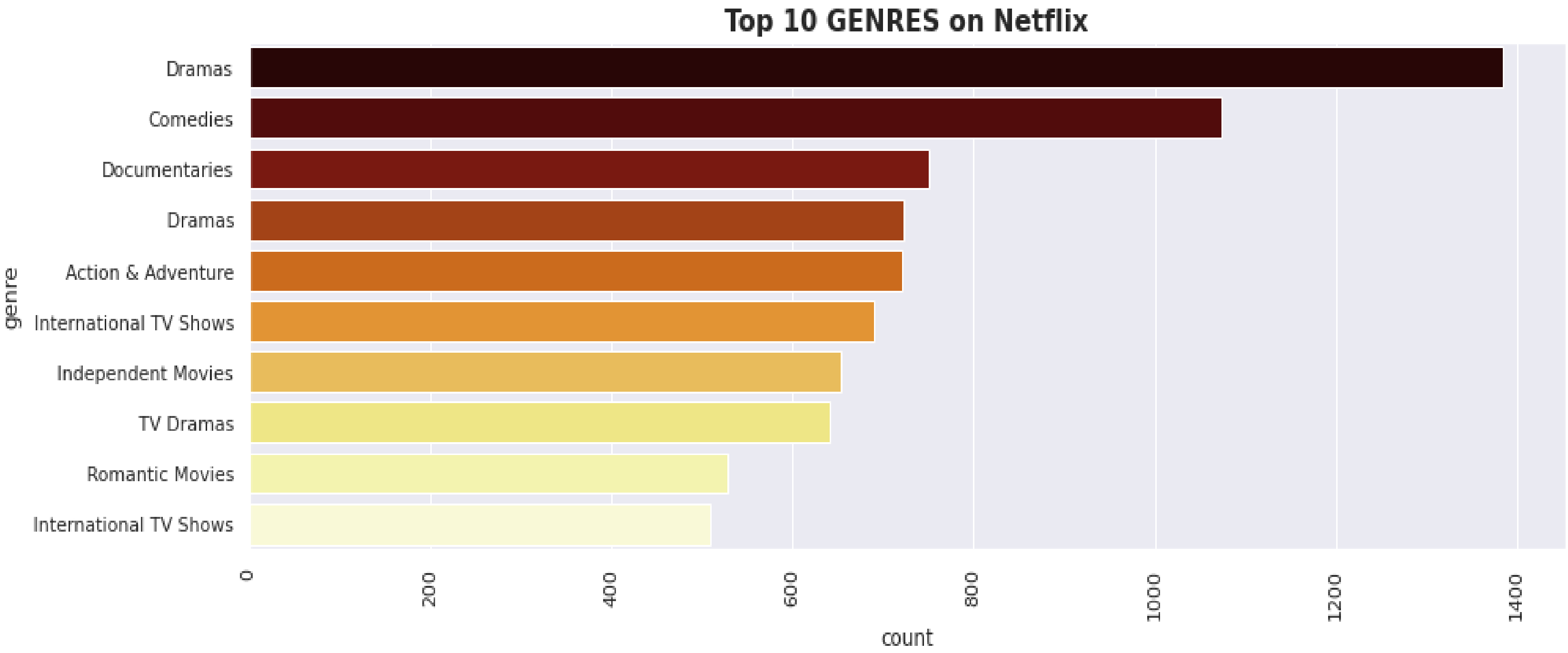




- Year 2020 has the highest number of TV Shows released.

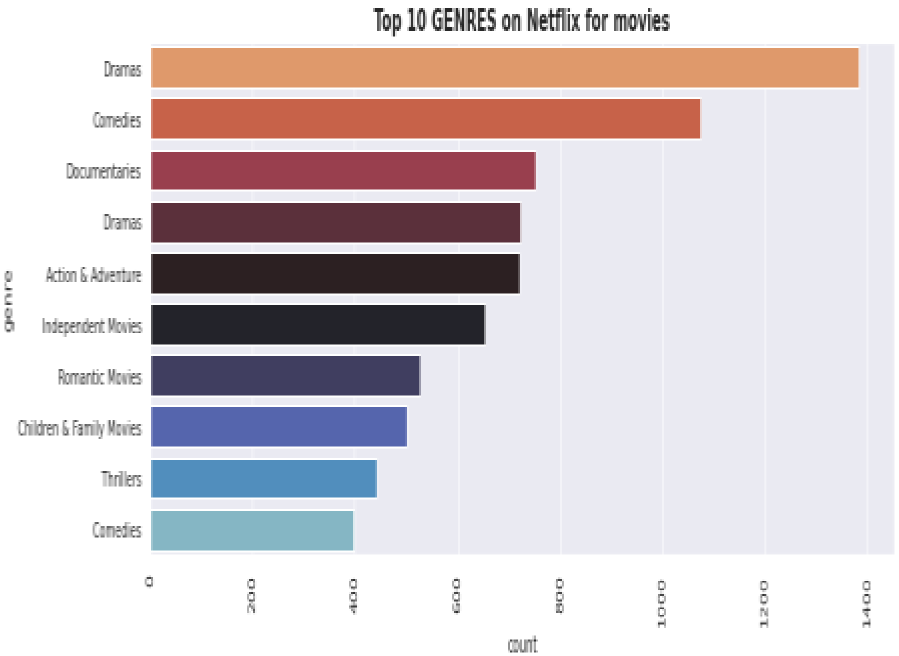
- Year 2017 has the highest movie released.

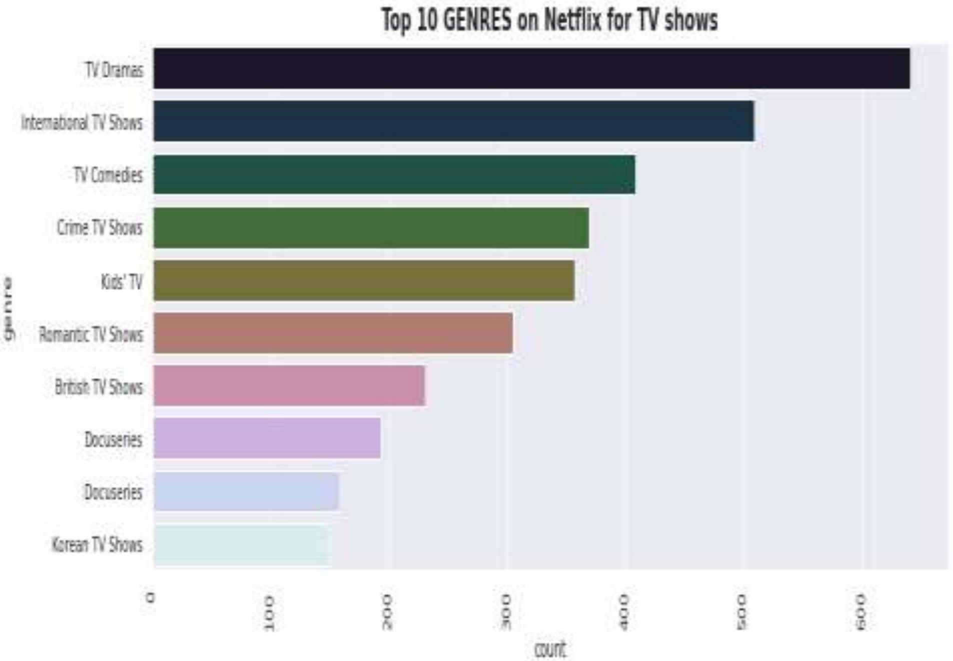
EDA– Top 10 Genre’s On Netflix



- Drama is the most popular genre followed by comedy.
- Romantic Movies & International TV shows have least popularity.

EDA- Top 10 Genres For Tv Shows & Movies





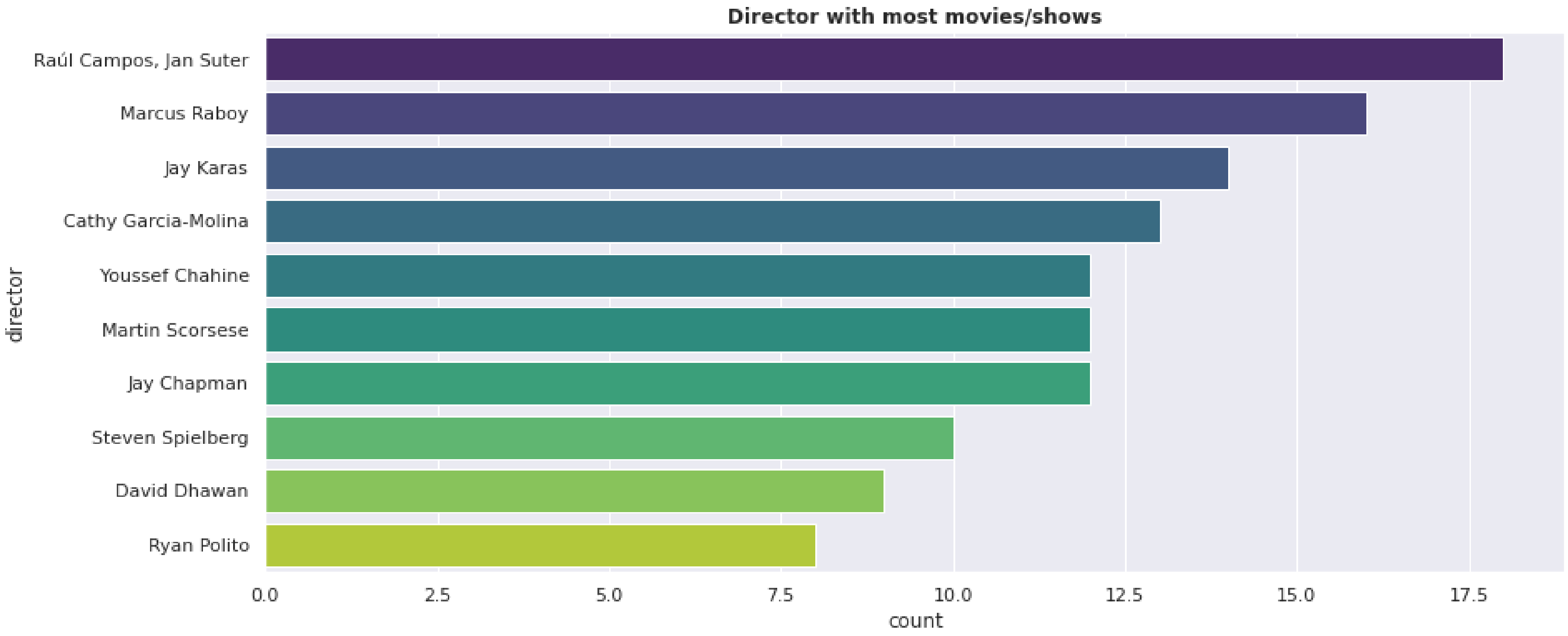
Movies

- Drama is the most popular genre followed by comedy for movies.

Tv Shows

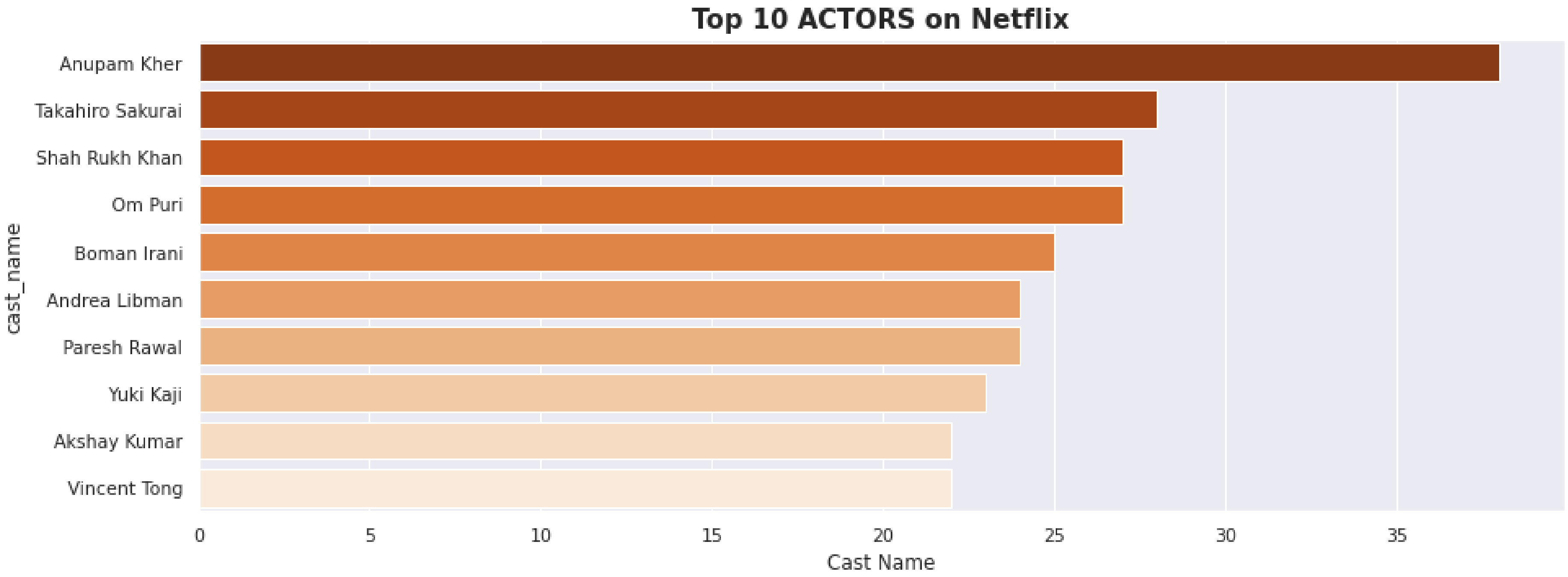
- Drama is the most popular genre followed by International TV shows for movies.

EDA – Director's With Most Movies & Tv Shows



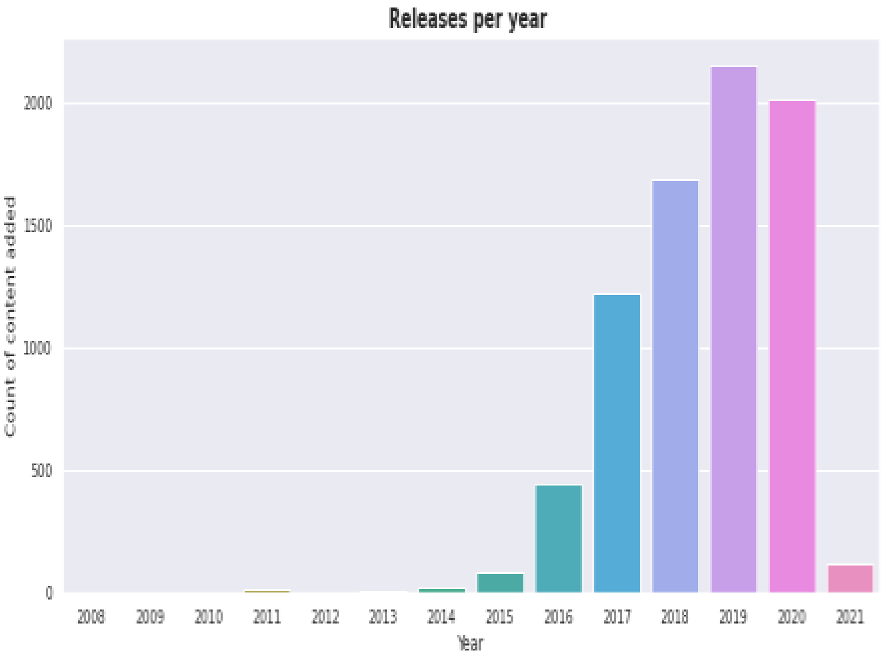
- Raul Campos and Jan Sulter collectively have the most content on Netflix
- Followed by Marcus Rayboy, Jay karas and Others.

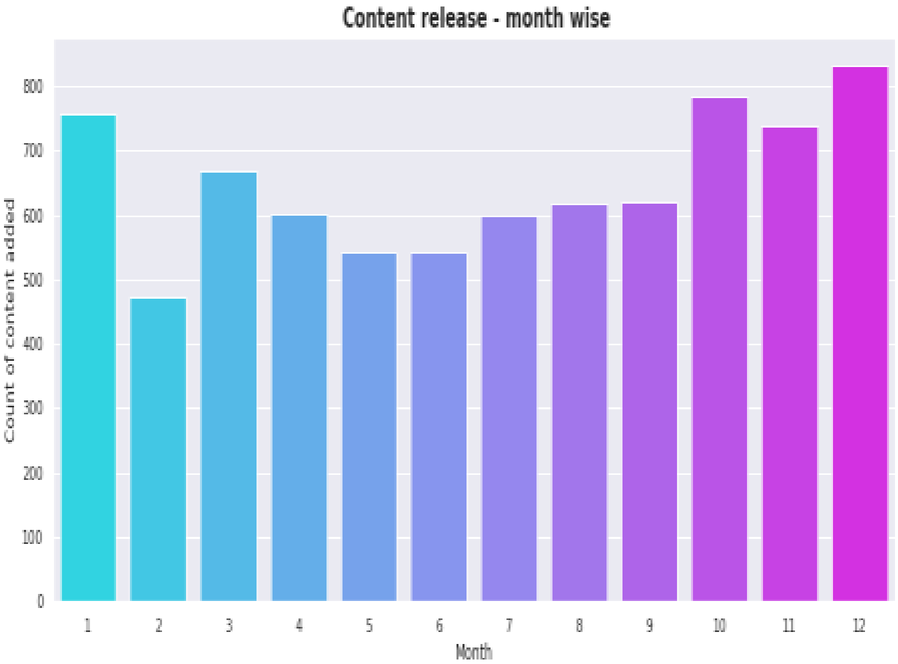
EDA – Top 10 Actor’s On Netflix



- Anupam Kher Have the most number of films on Netflix
- Followed by Takahiro Sakurai, Shah Rukh Khan, Om Puri & Others.

EDA- Year & Month wise Analysis

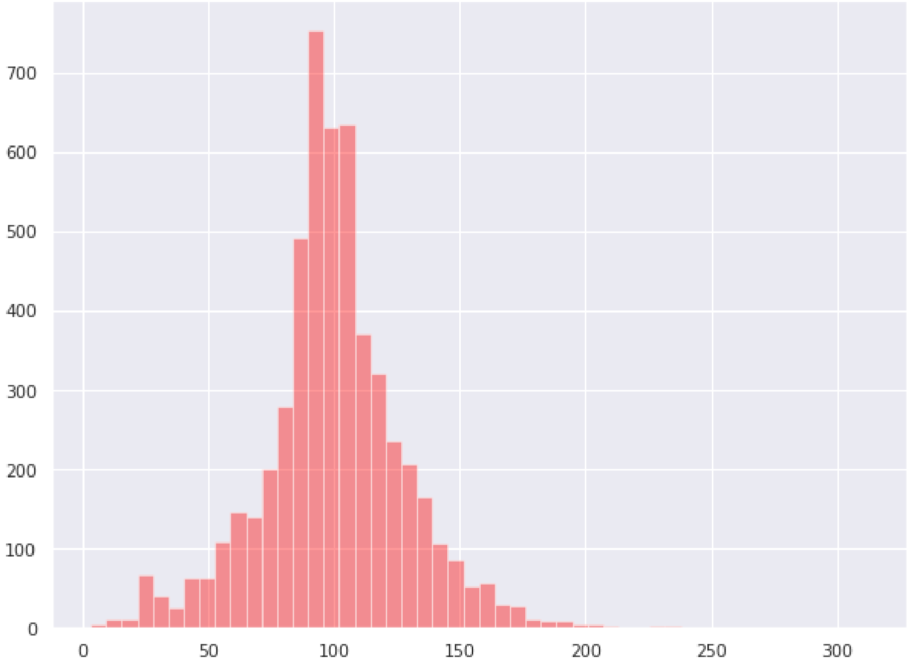


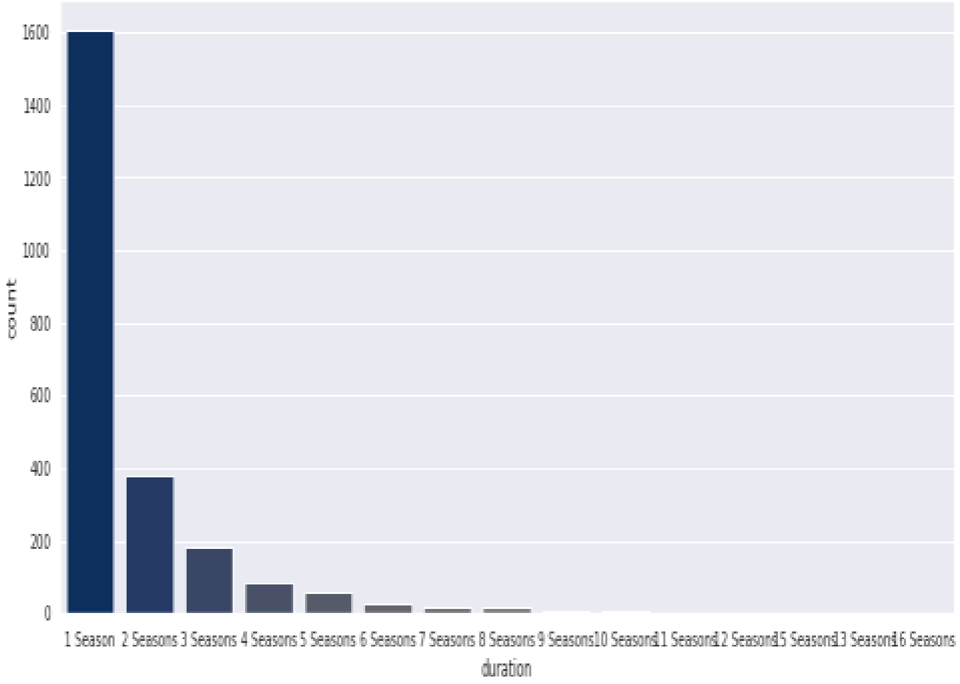


- The number of release have significantly increased after 2015 and have dropped in 2021 because of Covid 19.

- More of the content is released in holiday season - October, November, December and January.

Distribution of Movies & TV Shows Duration


Distplot with Normal distribution for Movies


Distribution of TV Shows duration

- Mainly the movie duration is in b/w 55 to 150 minutes.

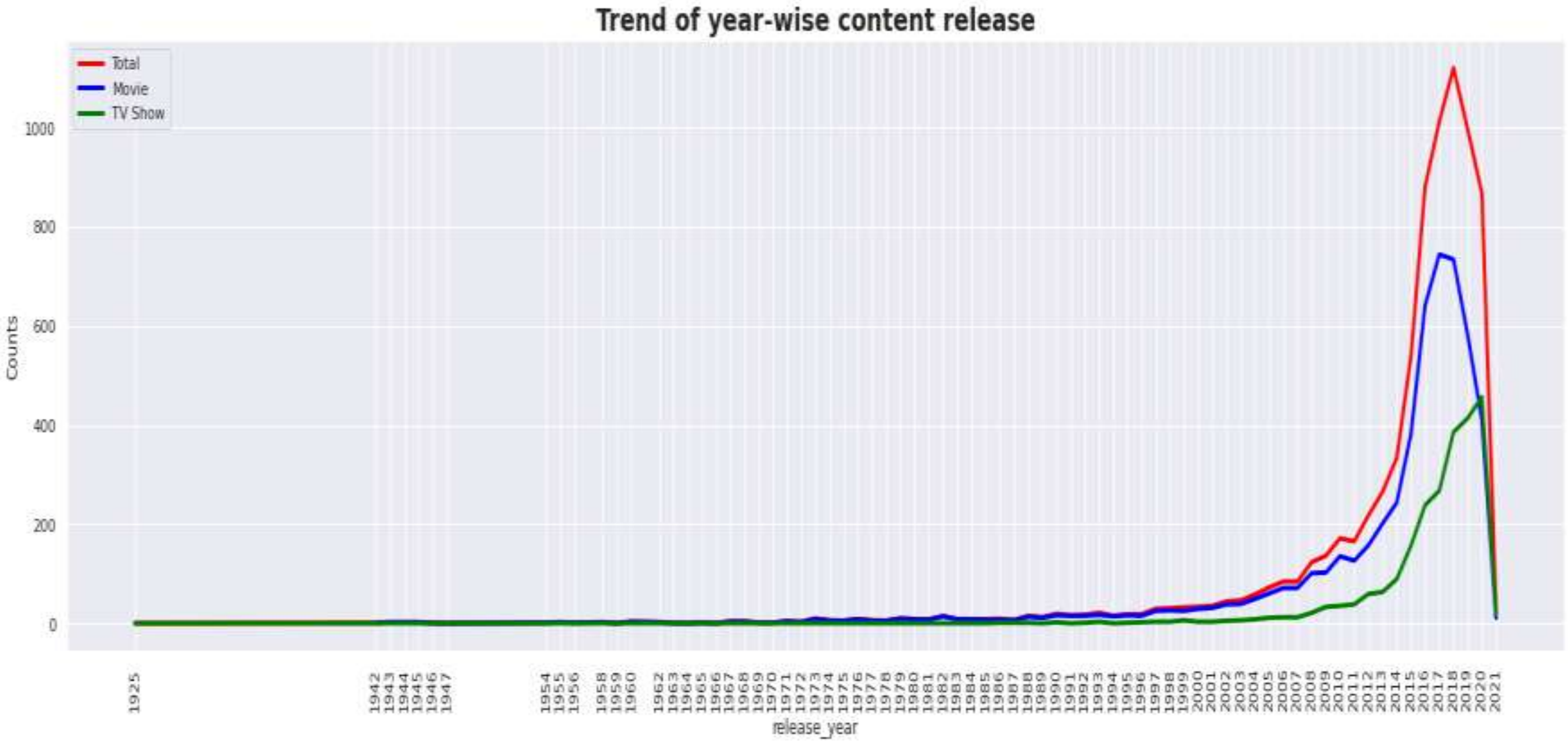
- Mostly every TV Shows has atleast 3 seasons.

HYPOTHESIS TESTING

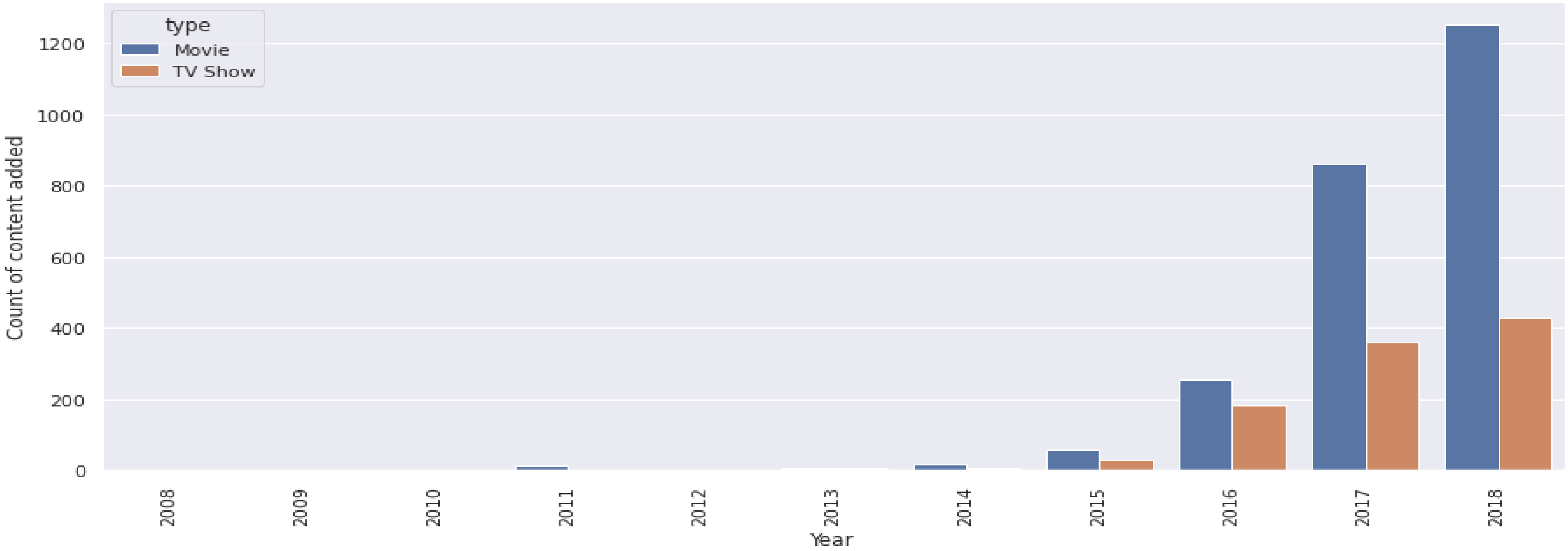
Given: In 2018, they released an interesting report which shows that the number of TV shows on Netflix has nearly tripled since 2010. The streaming service's number of movies has decreased by more than 2,000 titles since 2010, while its number of TV shows has nearly tripled.

HYPOTHESIS - Number of TV shows on Netflix have tripled and number of movies have reduced by 2000 between 2010 and 2018.

Year Wise Trend



Year Wise Trend



RESULT: Irrespective of the release years, There is no decline in the number of movies. Also number if movies added has always been more than the number of TV shows added. So with this information, we hereby reject our Hypothesis.

Data Preprocessing

- We have made some changes in data just for EDA so we will start our clustering analysis with fresh data and do the manipulations again.
- Since number of Unique actors are more than our number of rows, it's not going to help much for our analysis, hence we will not use this feature. Director has 30% null values and will not be used by us.

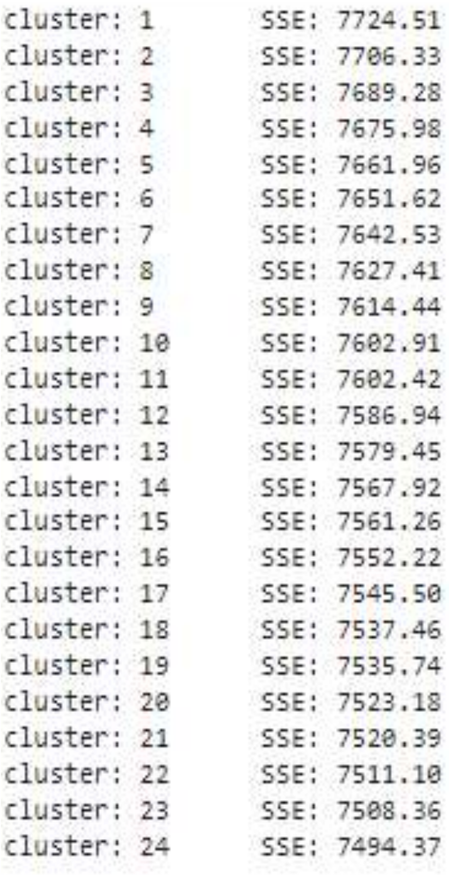
  Then, We Apply:

- **REMOVE STOPWORDS –** It is the most commonly used preprocessing steps across different NLP applications.

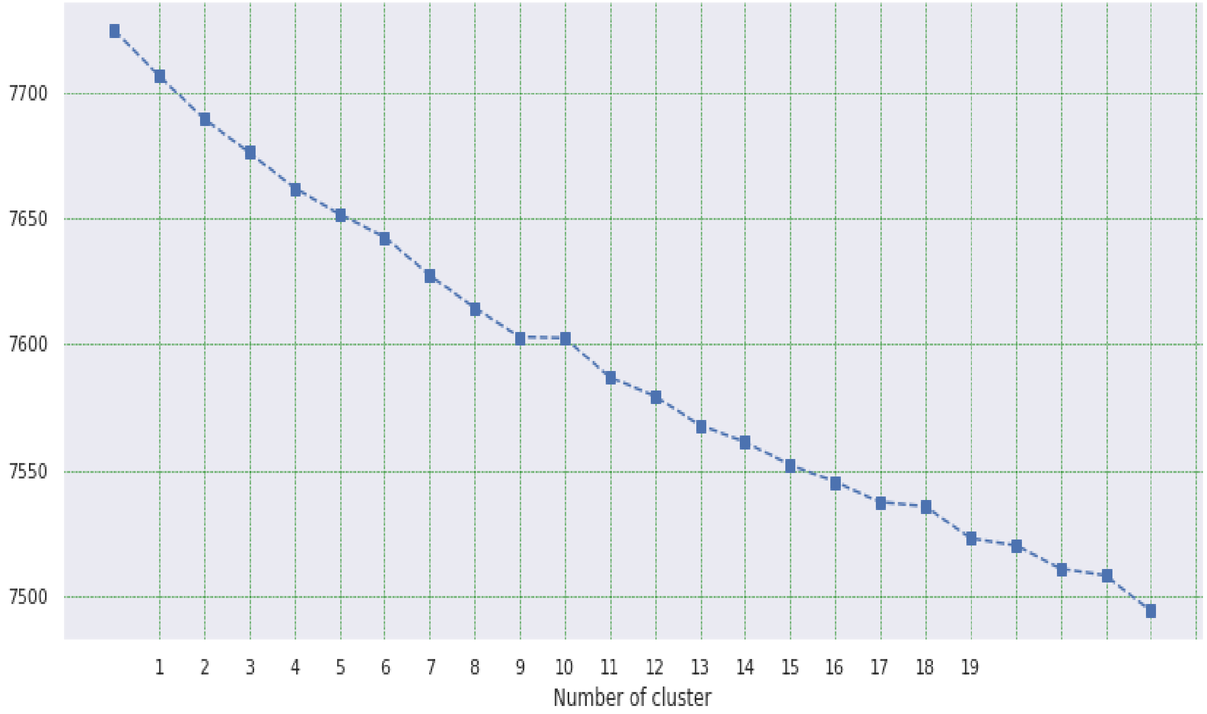
- **STEMMING –** It is a technique that lowers inflection in words to their root forms.

- **VECTORIZATION –** To convert the text data into numerical data, we need some smart ways which are known as vectorization.

Finding Number Of Clusters Using Elbow Method

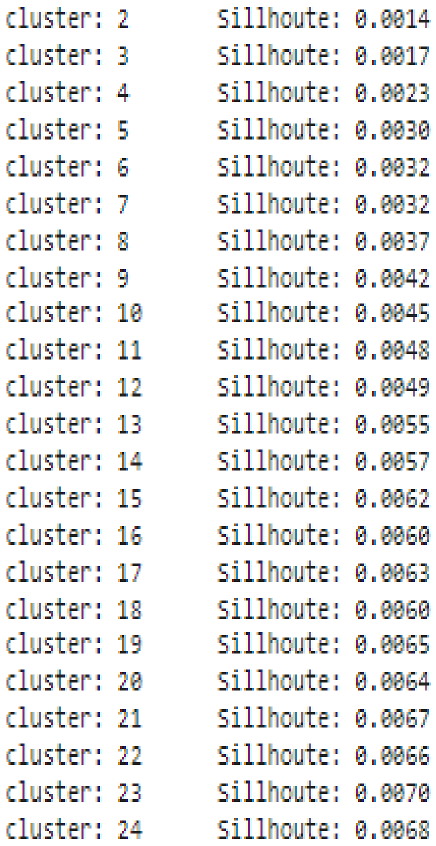


```
cluster: 1         SSE: 7724.51
cluster: 2         SSE: 7706.33
cluster: 3         SSE: 7689.28
cluster: 4         SSE: 7675.98
cluster: 5         SSE: 7661.96
cluster: 6         SSE: 7651.62
cluster: 7         SSE: 7642.53
cluster: 8         SSE: 7627.41
cluster: 9         SSE: 7614.44
cluster: 10        SSE: 7602.91
cluster: 11        SSE: 7602.42
cluster: 12        SSE: 7586.94
cluster: 13        SSE: 7579.45
cluster: 14        SSE: 7567.92
cluster: 15        SSE: 7561.26
cluster: 16        SSE: 7552.22
cluster: 17        SSE: 7545.50
cluster: 18        SSE: 7537.46
cluster: 19        SSE: 7535.74
cluster: 20        SSE: 7523.18
cluster: 21        SSE: 7520.39
cluster: 22        SSE: 7511.10
cluster: 23        SSE: 7508.36
cluster: 24        SSE: 7494.37
```

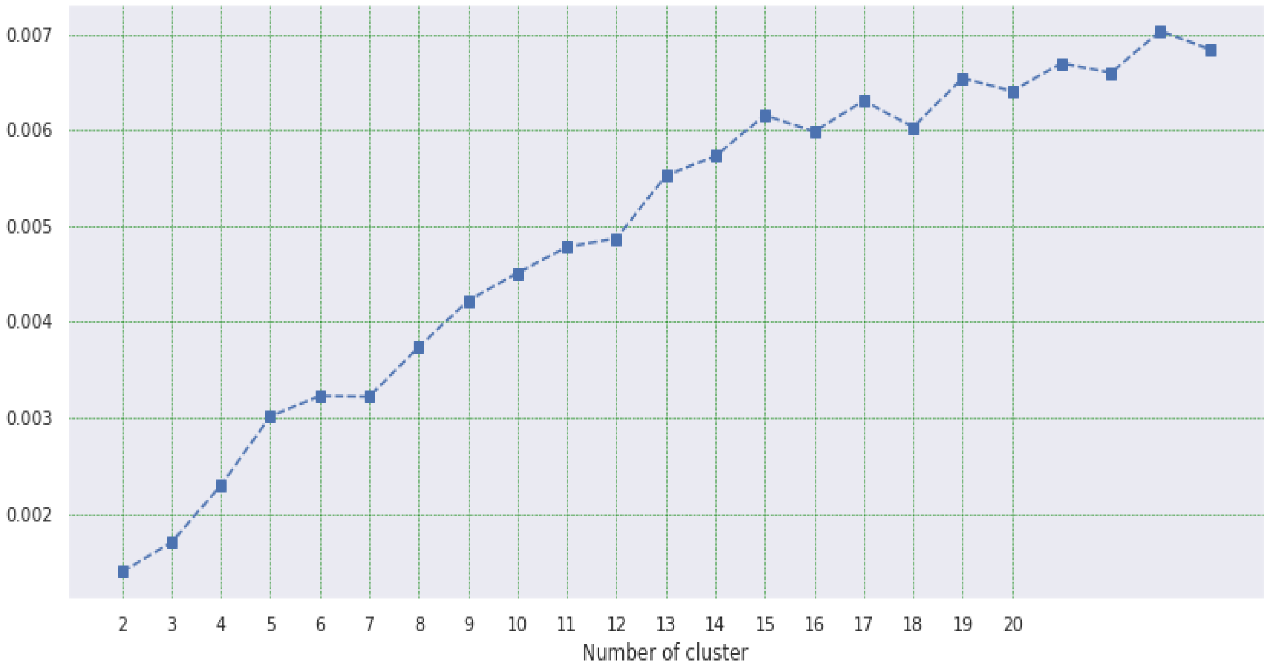


- Looks like we can go with 20 clusters from the visualizations.

Finding number of Clusters from Sillhoute's score



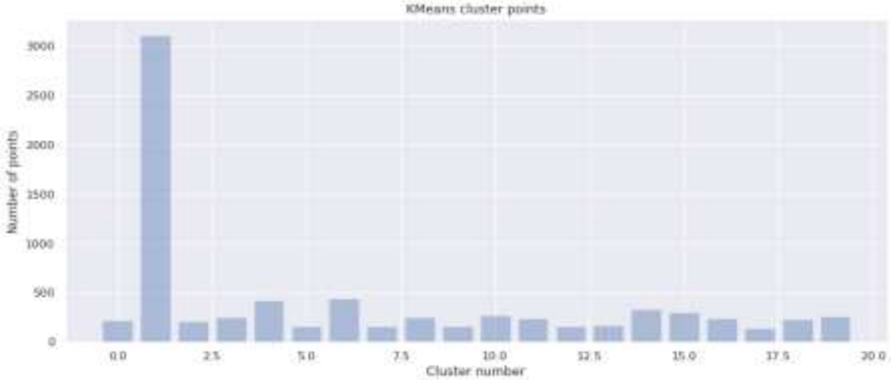
```
cluster: 2        Sillhoute: 0.0014
cluster: 3        Sillhoute: 0.0017
cluster: 4        Sillhoute: 0.0023
cluster: 5        Sillhoute: 0.0030
cluster: 6        Sillhoute: 0.0032
cluster: 7        Sillhoute: 0.0032
cluster: 8        Sillhoute: 0.0037
cluster: 9        Sillhoute: 0.0042
cluster: 10       Sillhoute: 0.0045
cluster: 11       Sillhoute: 0.0048
cluster: 12       Sillhoute: 0.0049
cluster: 13       Sillhoute: 0.0055
cluster: 14       Sillhoute: 0.0057
cluster: 15       Sillhoute: 0.0062
cluster: 16       Sillhoute: 0.0060
cluster: 17       Sillhoute: 0.0063
cluster: 18       Sillhoute: 0.0060
cluster: 19       Sillhoute: 0.0065
cluster: 20       Sillhoute: 0.0064
cluster: 21       Sillhoute: 0.0067
cluster: 22       Sillhoute: 0.0066
cluster: 23       Sillhoute: 0.0070
cluster: 24       Sillhoute: 0.0068
```



- Looks like we can go with 20 clusters from both the visualizations.

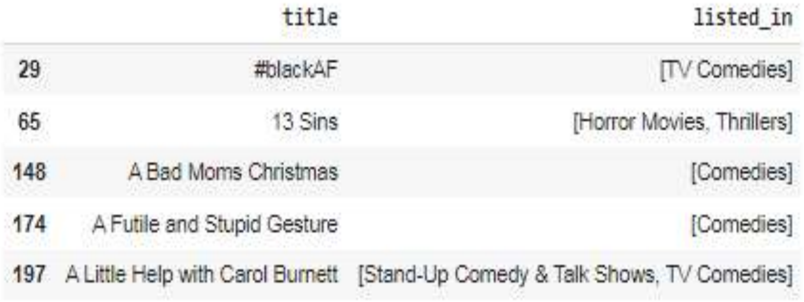
Implementing K-means

Parameters : KMeans(max_iter=100, n_clusters=20, n_init=1)



- Cluster 0 have highest number of cluster points.

Titles available in Cluster 0.



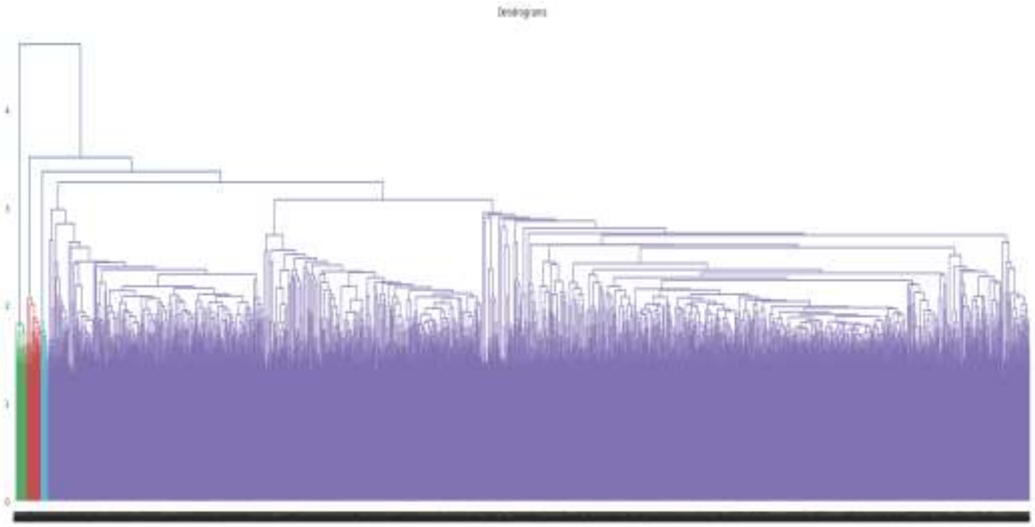
	title	listed_in
29	#blackAF	[TV Comedies]
65	13 Sins	[Horror Movies, Thrillers]
148	A Bad Moms Christmas	[Comedies]
174	A Futile and Stupid Gesture	[Comedies]
197	A Little Help with Carol Burnett	[Stand-Up Comedy & Talk Shows, TV Comedies]

Evalution

- Silhouette Coefficient: 0.00618
- Calinski-Harabasz index: 10.351
- Davies-Bouldin index: 9.650555

Hierarchical Clustering

Finding number of clusters from Dendogram



Evaluation:

- Silhouette Coefficient: -0.002
- Calinski-Harabasz index: 6.6459
- Davies-Bouldin index: 19.0527

- By using Sillhouette's score and Elbow method , we generated optimal of 20 clusters for K Means and from Dendogram , 6 clusters were generated.

Challenges

- Reading the dataset and understanding the problem statement.
- Designing multiple visualizations to summarize the Data points in the dataset and effectively communicating the results and insights to the reader.
- Data preprocessing – Remove stop words, Stemming and vectorization.
- Careful tuning of hyperparameters as it affects accuracy.
- Computation time was a big challenge for us.

Conclusion

1. There are about 70% movies and 30% TV shows on Netflix.
2. The United States has the highest number of content on Netflix by a huge margin followed by India.
3. Raul Campos and Jan Sulter collectively have directed the most content on Netflix.
4. Anupam Kher has acted in the highest number of films on Netflix.
5. Drama is the most popular genre followed by comedy.
6. More of the content is released in holiday season - October, November, December and January.

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

8. NULL HYPOTHESIS -The number of TV shows on Netflix have tripled and number of movies have reduced by 2000 between 2010 and 2018. (REJECTED).

9. By using Sillhouette's score and Elbow method , we generated optimal of 20 clusters for K Means and from Dendogram , 6 clusters were generated.

10. In both the cases, one cluster accounts more than 3000 points whereas in other clusters the points were unevenly distributed. 3.For Tfidf K Means is best for identification than Hierarchical as the evaluation metrics also indicates the same.

THANK YOU!!