Unsupervised ML Project

Netflix Movies & TV Shows Clustering

By- Ranajay Biswas

Content :

- Introduction
- Problem Statement
- Understanding Data & Attributes
- Approach
- Exploratory Data Analysis
- Pre-Processing
- Recommender System
- Topic Modelling
- Cluster Modeling
- Cluster Analysis
- Conclusion

Introduction :

Netflix is a streaming service that offers a wide variety of award-winning TV shows, movies, anime, documentaries and more – on thousands of internet-connected devices. New TV shows and movies are added every week!

Netflix has an extensive library of feature films, documentaries, shows, award-winning Netflix originals, and tons of other contents.

In this unsupervised Machine Learning project, we have to find out different patterns and insights about various movies and shows.

Clustering similar contents and developing a content based recommender system that is capable of recommending new movies and shows to users as per their interests is going to be our main goal.

Performing Exploratory data analysis, insights can be drawn from the data that can be used to better understand the production and demand for different content in different countries. Evaluating genres and content rating for different age groups, we can find out which demographic is dominated by what kind of audience.

Analyzing audiences' taste and finding similar content will help to build intelligent recommendation systems that will be able to satisfy the audience's appetite for content consumption.

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, it we are required to do -

- Exploratory Data Analysis
- Understanding what type content is available in different countries
- If Netflix has increasingly focusing on TV rather than movies in recent years.
- Clustering similar content by matching text-based features

Understanding Data & Attributes :

Data Collection & Summary : Data is in csv format. We used pandas' `read_csv()` function to import the data in our work environment.

Dataset has 7787 rows and 12 columns. And it contains object, integer and float data types...

Attributes :

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

Approach :

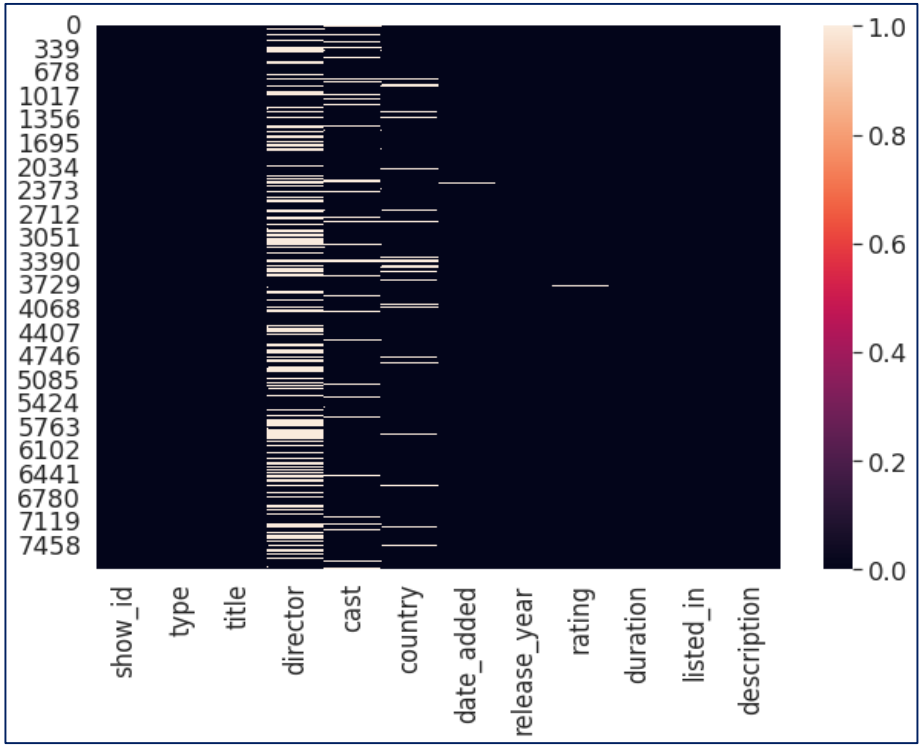
This project is heavily dependent upon analyzing the data and finding significant patterns and trends to find movies and shows which are similar to each other. This way Netflix can be more confident when making recommendations to their audience. It is also going to be helpful to know what type of content is going to be more popular among their platform users. So that Netflix can produce or add those new shows and movies on their platform.

Our approach to solve this problem is going to be -

- Understanding the dataset, different rows and columns.
- During the EDA, we will try to find popular actors and directors, what are the top content producing countries. We will get an idea about different genres and runtimes for different contents. Statistical methods and Visualizations are going to be very helpful in this EDA process.
- In the pre-processing step, we shall filter the most important features, make necessary transformations, create or omit features as needed. Depending on the features we choose, we will find best approaches for text pre-processing. Then we will have the data, ready for implementing clustering algorithms or building a recommender system with this data.
- Using Topic Modelling, we will try to find popular genres. We shall try different clustering algorithms to see which fits the data best. The best performing algorithm will be selected and then we shall analyze different clusters made by the algorithm to find patterns inside those clusters.
- Then we shall conclude the project with an overview and discussion about the observations that we make and how that can prove to be useful for Netflix's streaming services.

Exploratory Data Analysis :

Exploratory Data Analysis (EDA) is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations.



Checking Null Values : There are many null values in the data..

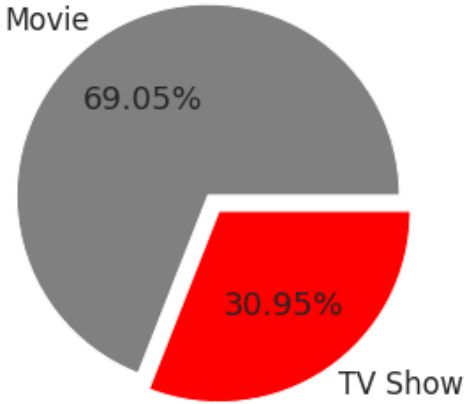
- **Director** column has roughly 31%,
- **Cast** column has 9%,
- **Country** column has 6.51%,
- **Date Added** column has 0.13% and
- **Rating** column has 0.09% null values...

We found no duplicates in the data.

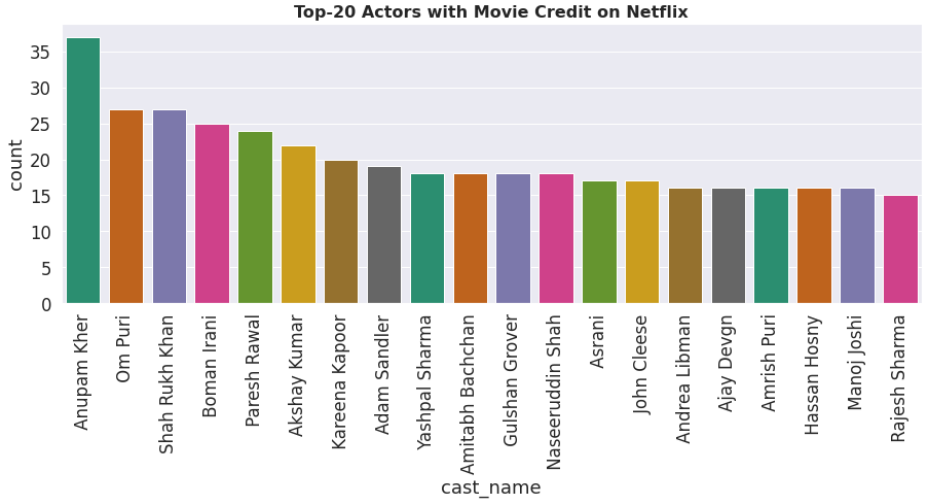
EDA Continued...

Proportion of Movie vs TV Shows :

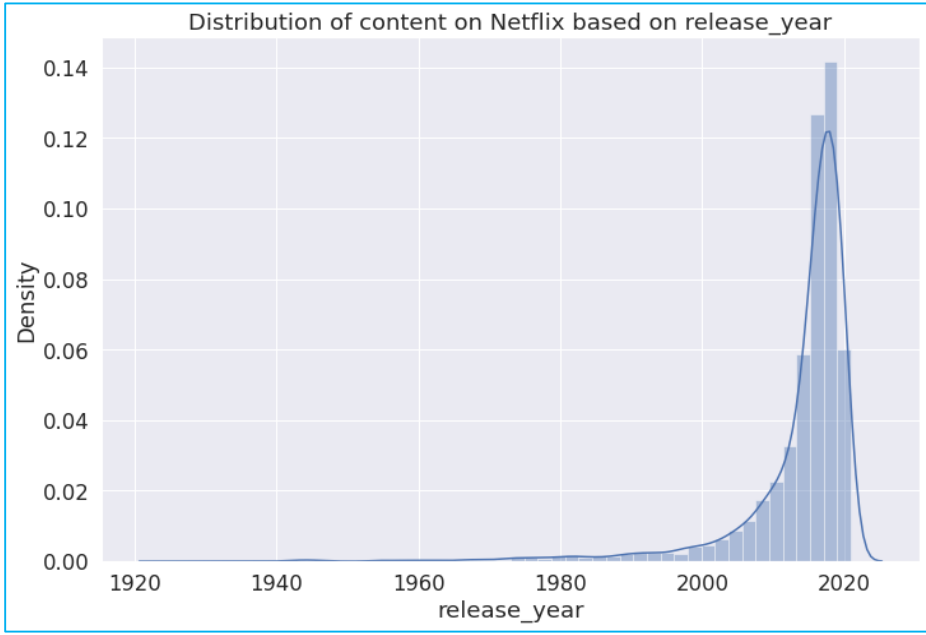
- We have 5377 movies and 2410 TV shows in the data.
- There is roughly 69% movies and 31% TV Shows present in the data.

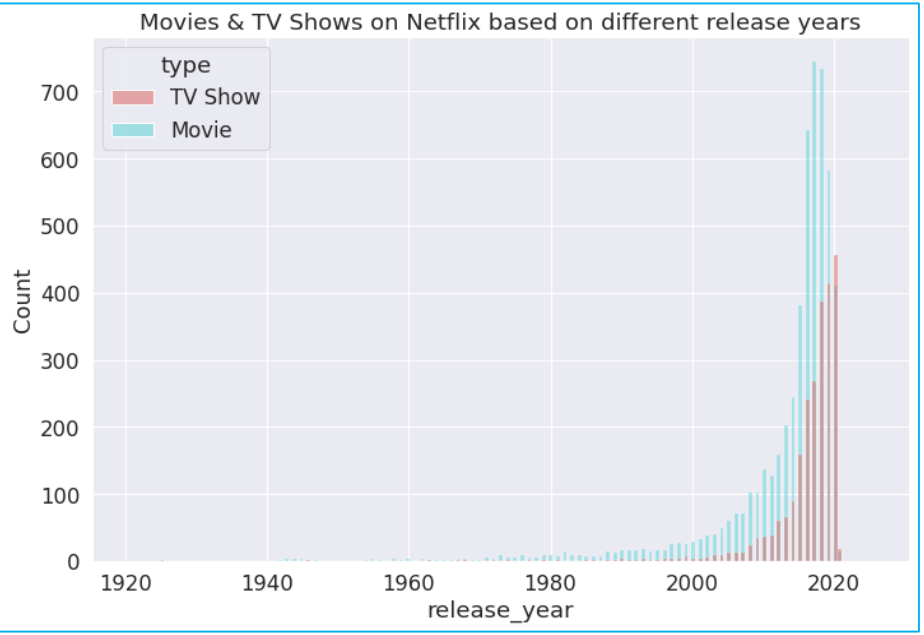


Actors with most movie Credits on Netflix:



Analyzing Based on 'Release Years' of movies and Shows:

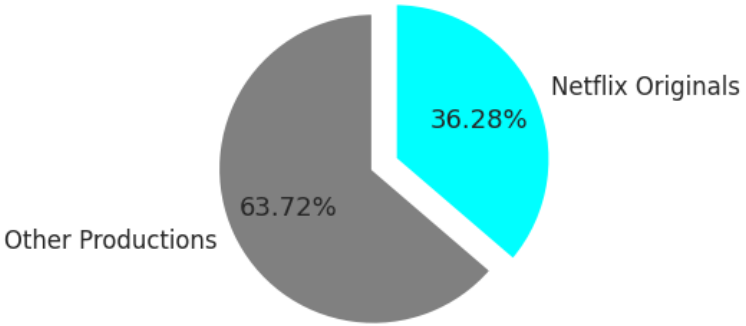


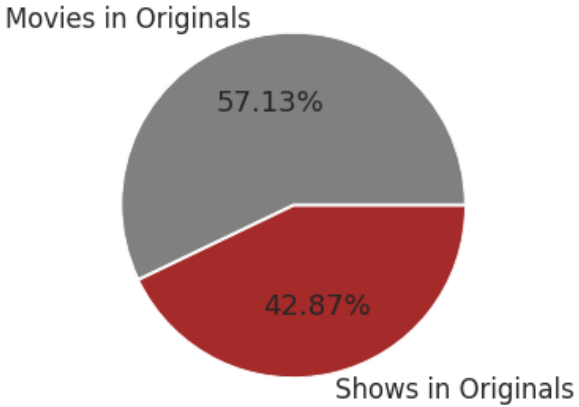


Observation :

- The number of releases of movies and shows kept growing throughout the years.
- After 2019, we see a downfall in the releases. This a direct effect of Covid-19 Lockdown. The production in 2020 and 2021 is also low for the same reason.
- An interesting observations we can make is that in the last 2-3 years, Netflix focused heavily on TV shows compared to movies.

Netflix Originals :



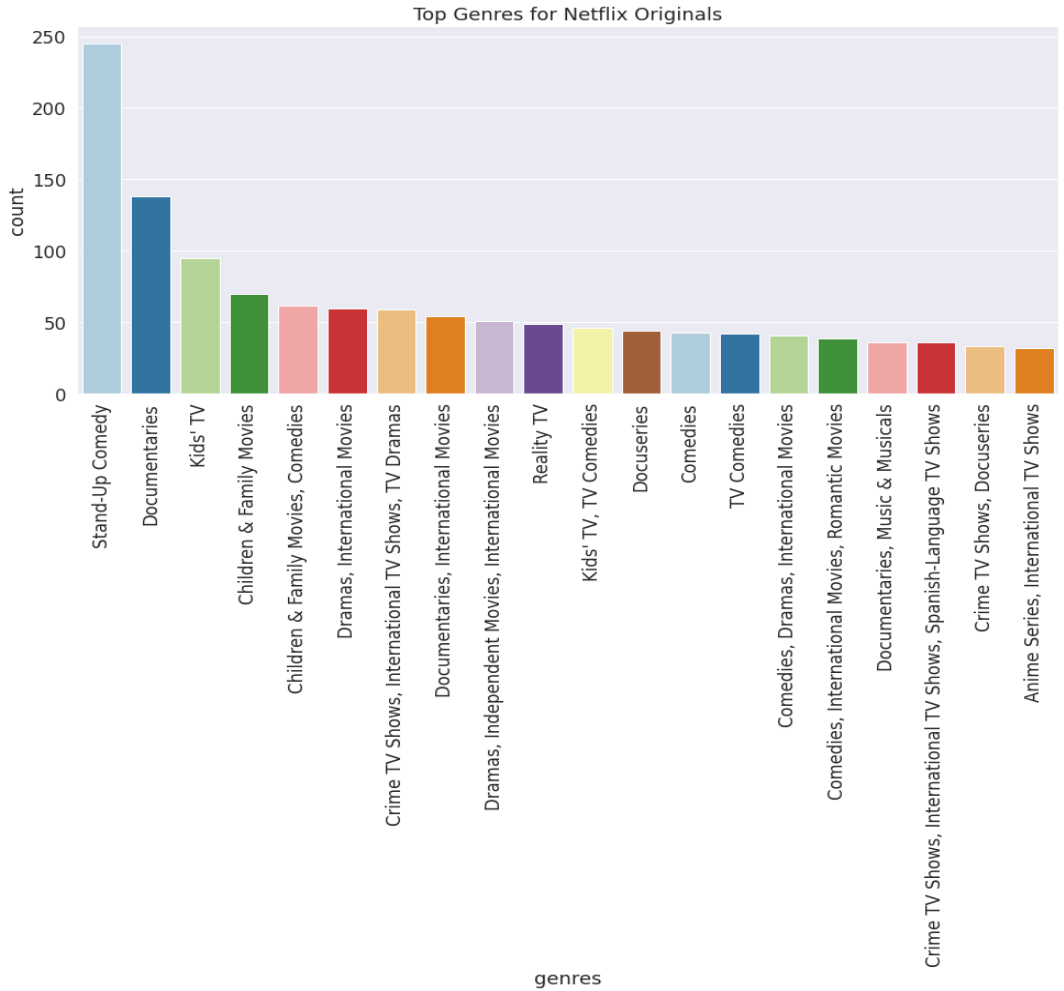


Observations :

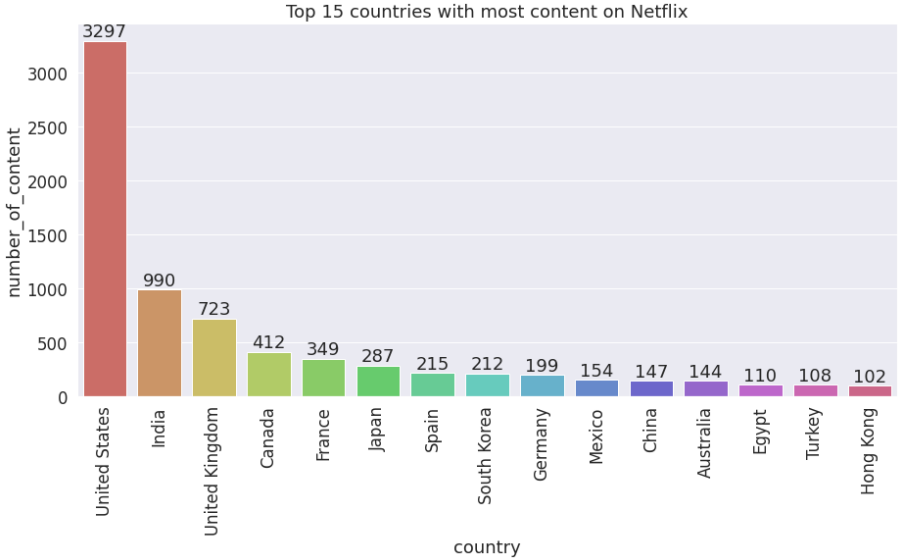
- Of all the total content on Netflix, 36% are Netflix Originals. Rest of the content is made by other productions.
- More movies were produced on Netflix compared to shows.
- 57% of the Originals are Movies and 43 % are TV Shows.

EDA Continued...

- Originals mostly contain Stand-Up Comedies, Documentaries and Kids-TV.
- Children & Family movies and Crime-Drama shows are also popular Netflix productions.

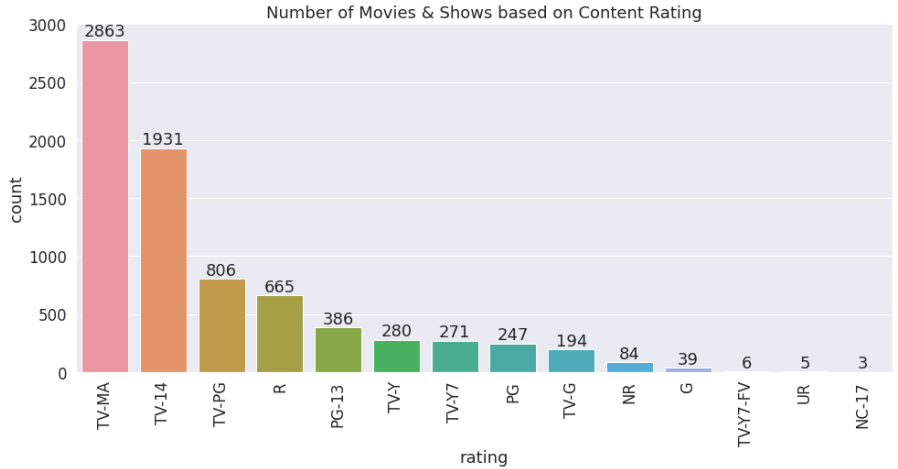


EDA Continued...

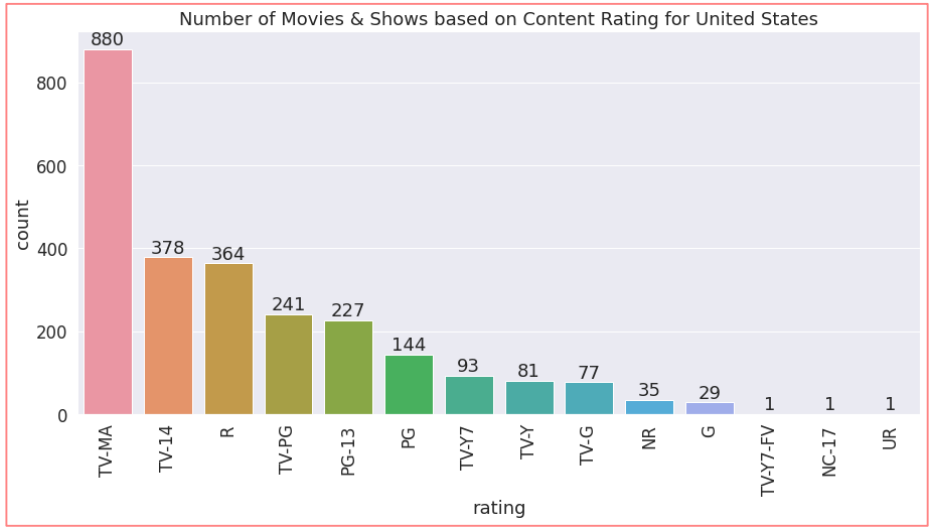


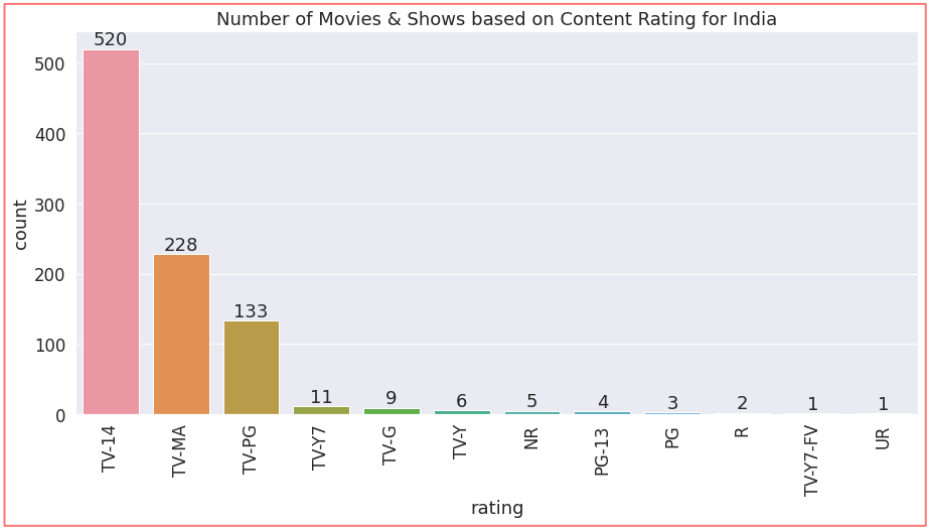
- There is a total of 118 unique countries in the data.
- Most of the content on Netflix comes from United States.
- 2nd & 3rd most content creating countries are India and United Kingdom.

- The unique ratings are - 'TV-MA', 'R', 'PG-13', 'TV-14', 'TV-PG', 'NR', 'TV-G', 'TV-Y', 'TV-Y7', 'PG', 'G', 'NC-17', 'TV-Y7-FV', 'UR‘.
- We can see that TV-MA is the most popular rating type, followed by TV-14, TV-PG & R..



EDA Continued...

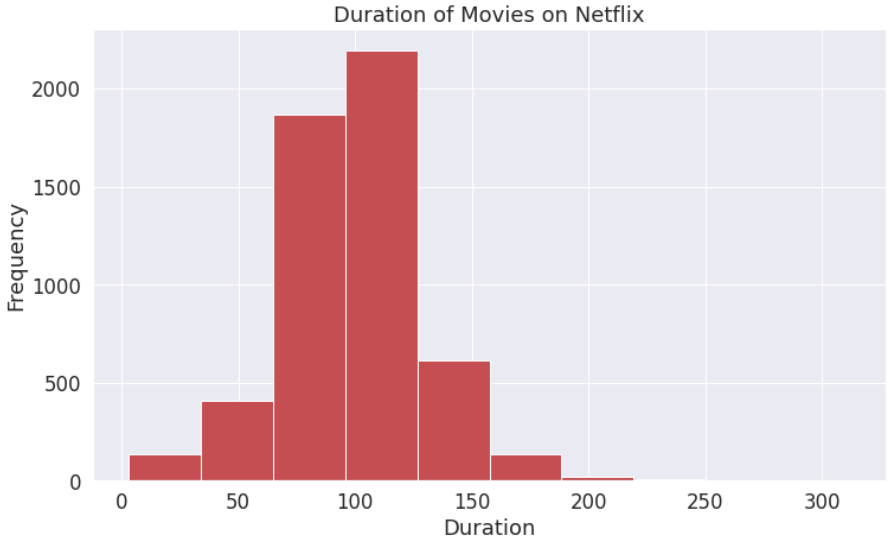


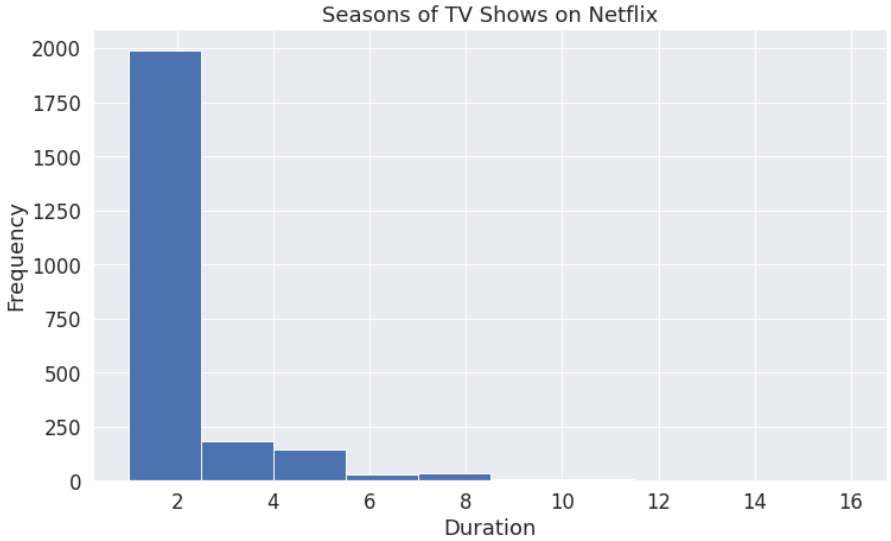


United States and **India** are top content providers for Netflix. But we can see a different for content rating depending on the country -

- For **United States**, top 3 rating types are -TV-MA, TV-14, R.
- For **India**, top 3 content ratings are - TV-14, TV-MA, TV-PG.

EDA Continued...



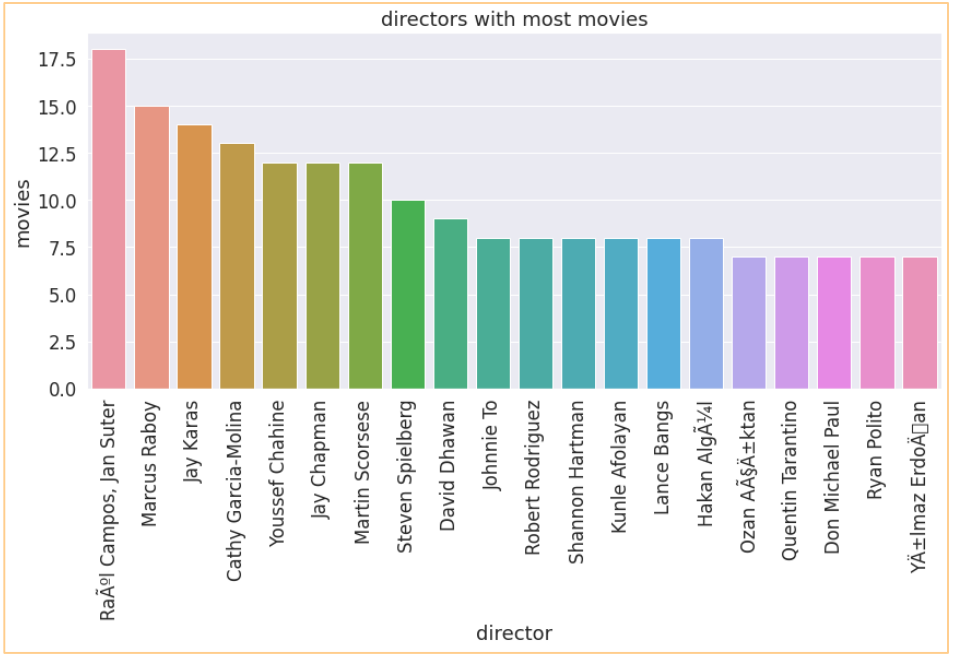


Observation :

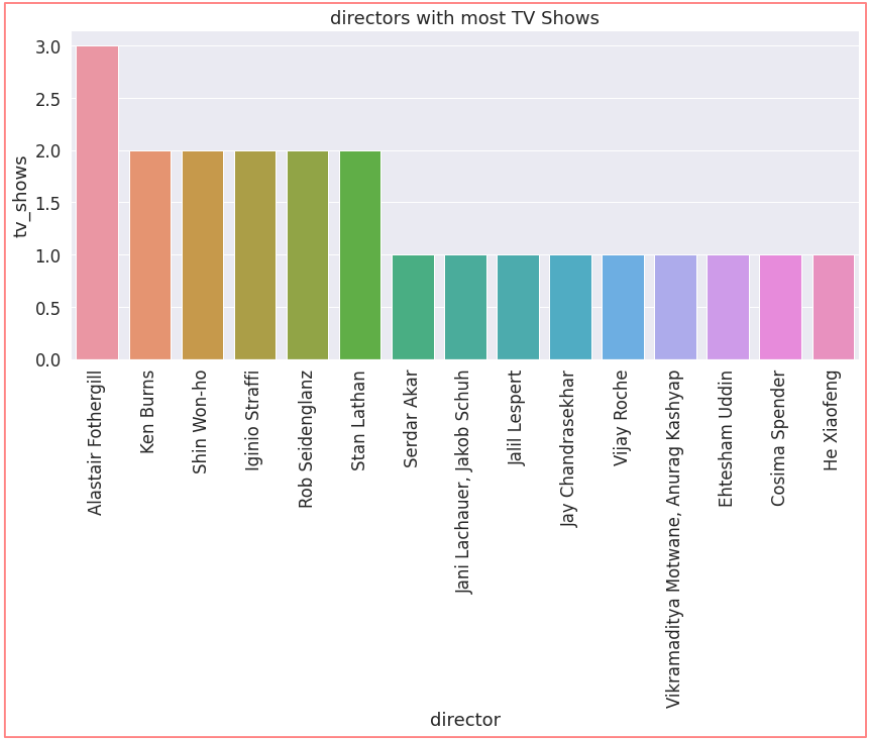
- We see that most movies have a duration of 90 - 120 minutes.
- Mostly Netflix prefers having Shows with 1 or 2 seasons only.

EDA Continued...

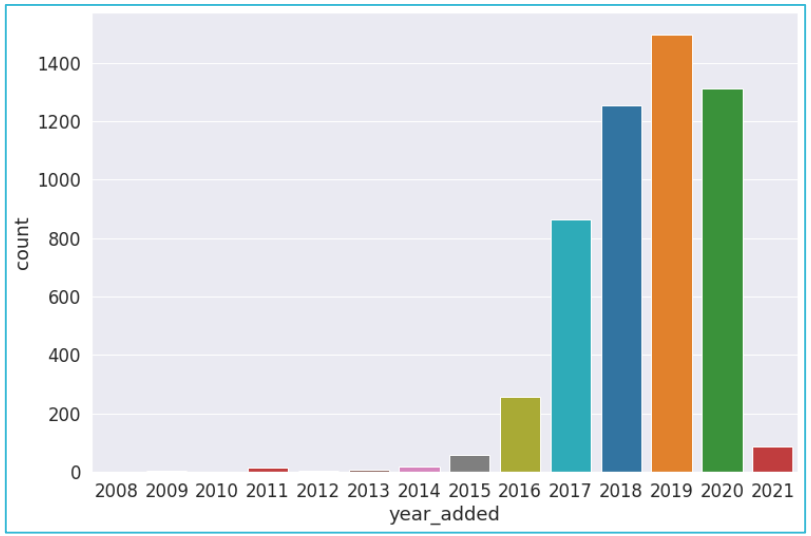
Directors with most Movies

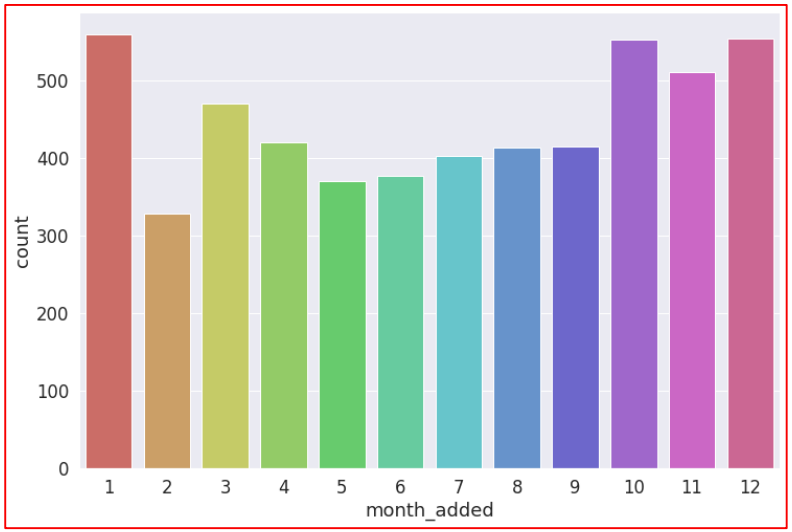


Directors with most TV shows -



Movies added on Netflix based on Years and Months..

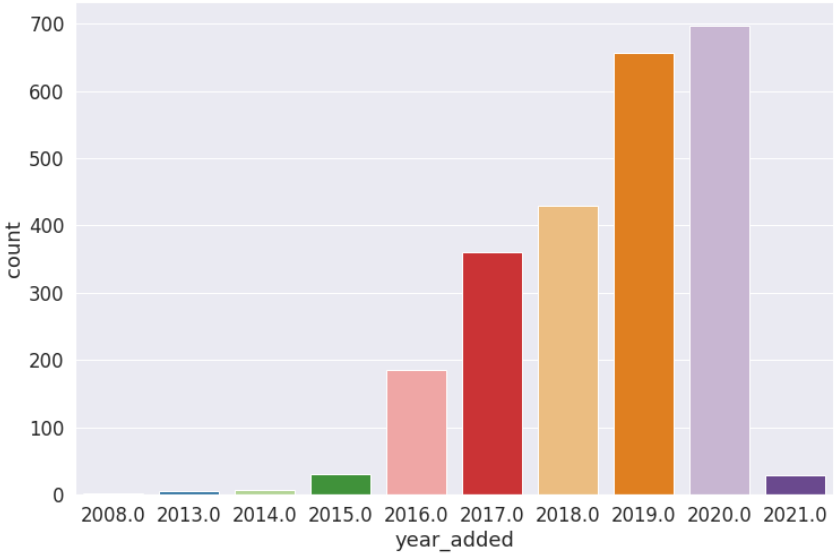


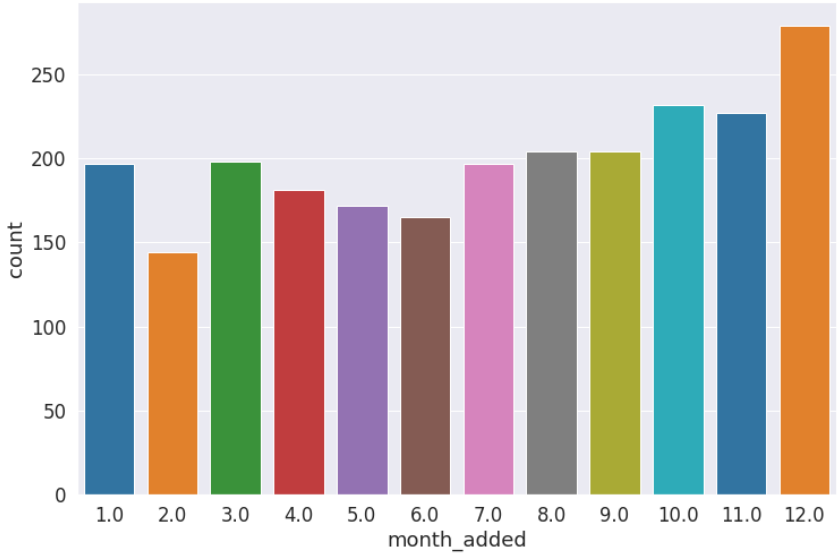


Observations -

- We see the pattern that more and more movies were added on Netflix throughout the years.
- In the year 2019, the numbers peaked.
- A significant decrease in the numbers can be noticed on 2021 as a result of Covid.
- In the month of February, the least number of movies are added.
- From October to December, and also on January, we notice that most number of movies are added on Netflix.

Visualizing Correlation :



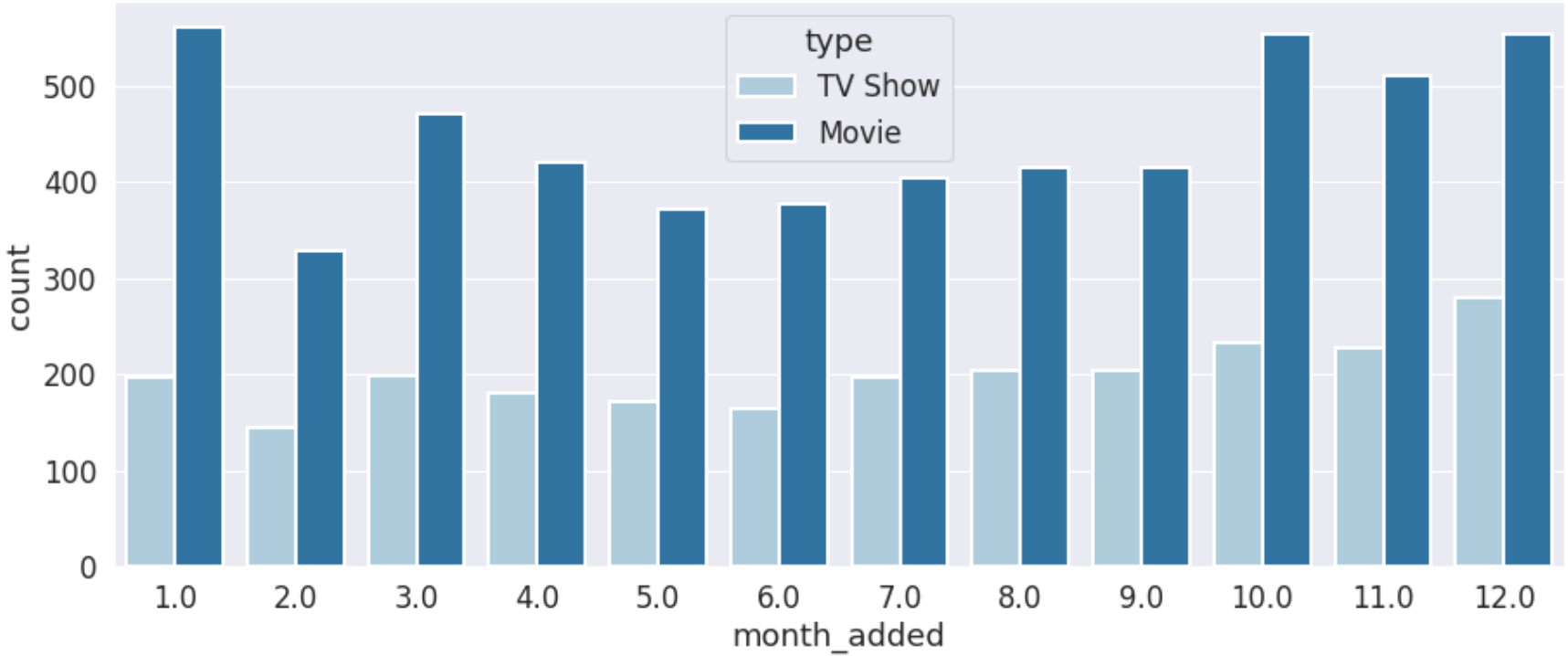


We notice a similar kind of pattern as movies.

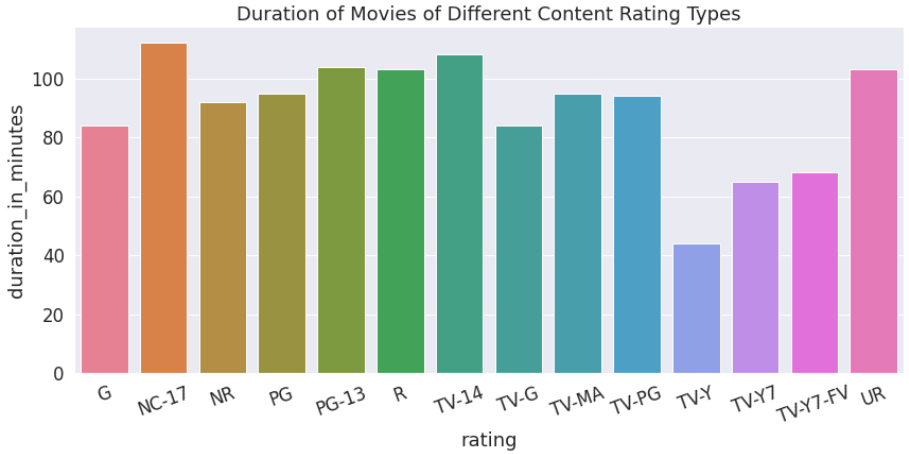
- Year 2021 took a hit because of the pandemic.
- In the month of December, most number of shows are added.

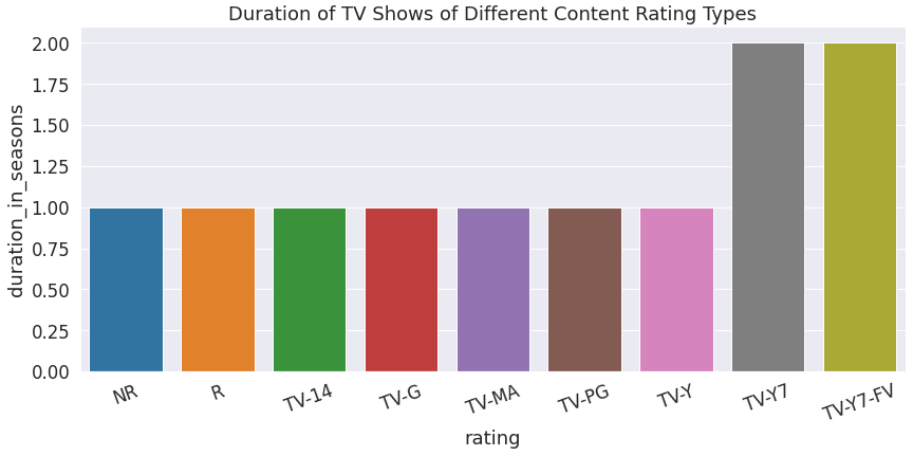
EDA Continued...

Comparison of Movies & TV Shows based on how many of each were added every month...



EDA Continued...

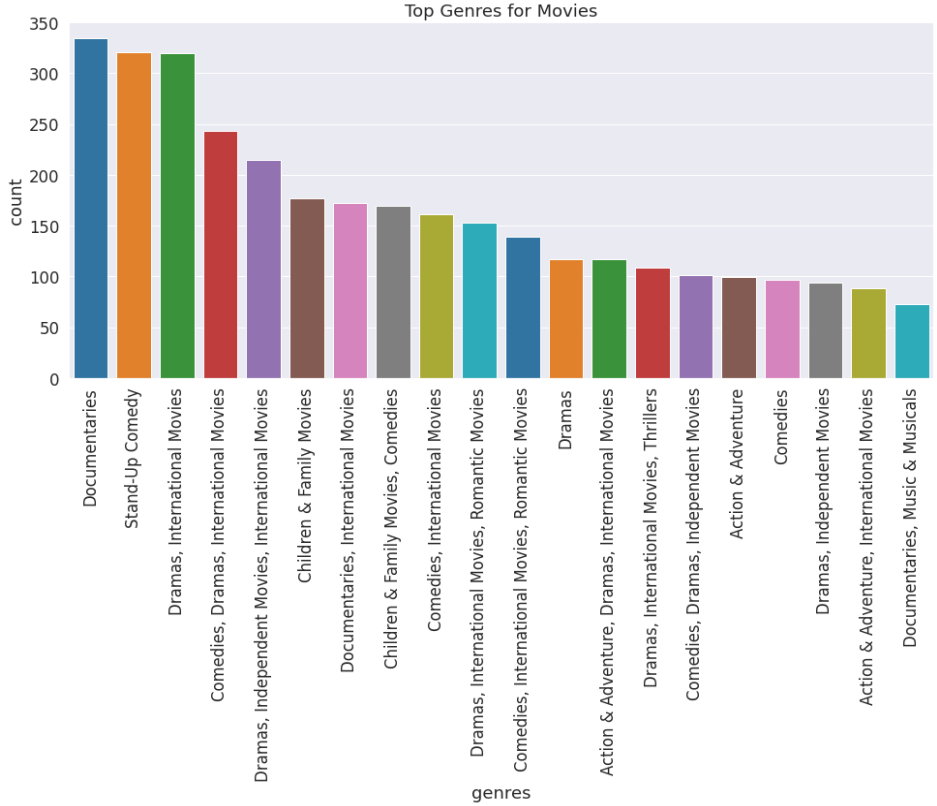


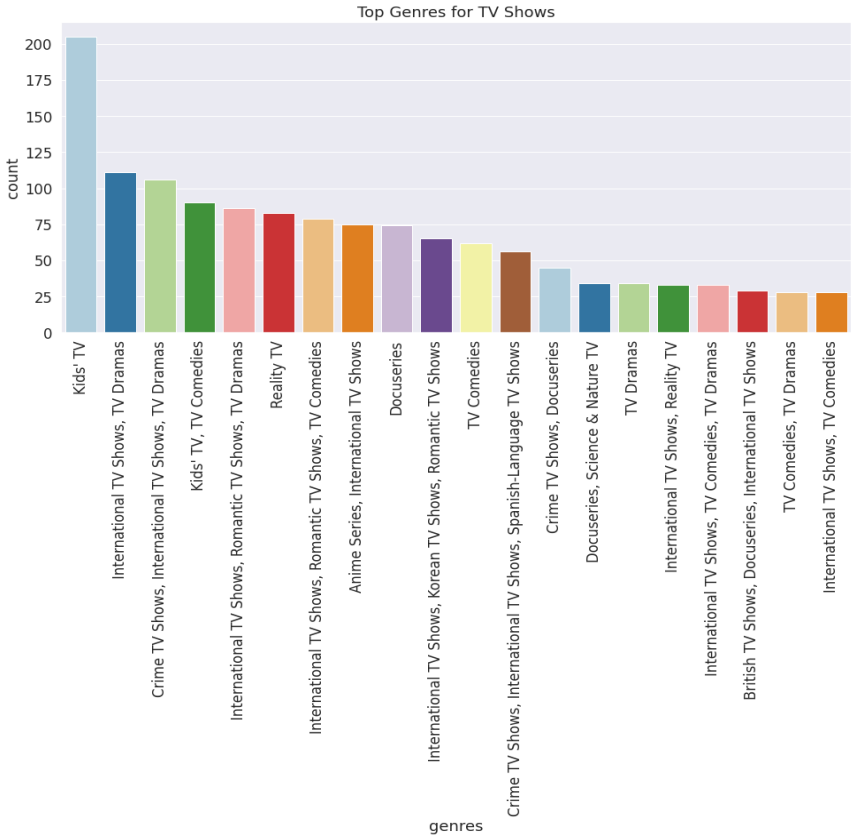


Observations :

- TV-Y rating suggests that the content is for children of all age. And we can see that, movies with this content rating has lowest average(median) runtime.
- NC-17 rating means that the content is not suitable for audience below 17 years of age. For this rating type, the movies have the highest duration or runtime.

EDA Continued...





- Documentaries, Stand-Up Comedy, Drama & International Movies are the top listed movies on Netflix.
- Kids' TV, International TV Shows, TV Dramas are top listed genres for shows.

Pre-Processing :

Feature Selection : To perform clustering and build a recommendation system, we will keep certain columns that will be helpful. So, let's list out the columns that we will need.

We need to cluster similar TV Shows and Movies. Meaning, we will mainly look for content of similar **Description**, **Rating** and **Genres**. Then we can also make the system better by including the **Director** & top **Casts** from these movies and shows.

At this point in time, we are not concerned about differentiating between TV shows and Movies. Rather, we are trying to find similar content regardless of them being a movie or a show.

So, the most important columns are going to be -

1. **Description**
2. **Rating**
3. **Listed In (Genres)**
4. **Director**
5. **Cast**

For Recommendation, **Director** and **Cast** columns have great purpose. But, when trying to cluster similar movies and shows, these two columns might not be that useful. As same directors and actors can work in different kinds of movies. For that reason, what we can do is that, we can form clusters with and without including the **'Director**' and **'Cast**' columns and compare results.

If the clusters formed, does not give good results with these features, then we shall also try forming clusters without the Director and Cast features.

Null values treatment :

We saw that the some of columns that are important for us, have missing values. Columns like **Director**, **Cast** and **Rating** are object type. So, we can not use mean or median. Also, we do not want to use mode for null value imputation as well. That would not make much sense.

We will impute these null values with 'unknown'.

Feature Transformation & Text Pre-processing :

Rating column basically contains all the content ratings for different movies and shows.

Cast and **Director** columns contain the names, **Listed In** column contain the genres. So, in feature engineering, we converted all these features to lowercase, removed the separating characters.

We choose top 4 leading actors from the **Cast** column as we do not want overcrowd the data with all the actors.

Description column contains the short summary of the movie. In order to get the best out of this feature, we will need to perform multiple text-preprocessing step.

The steps we took for this part are –

- **Punctuation Removal :** Presence of signs and punctuations doesn't make sense. It only makes the data noisy. So, we will remove those.
- **Lowercase :** Converting all the text to lowercase is sort of a thumb rule for text preprocessing.
- **Stopwords Removal :** Very commonly occurring words which only contribute for sentence formation are considered stop words. These do not contribute to the meaning that much. So, we will remove that.

Text Pre-Processing continued...

- **Tokenization :** Tokenization is the process for converting words into tokens.
- **Lemmatization :** This helps to bring every word down to it's base form. This step is necessary to identify the same words in the corpus that conveys the same meaning. Another similar substitute for Lemmatization can be Stemming. But usually Lemmatization performs better than Stemming. So, we decided to use Lemmatization for this project.

After the pre-processing is done, we create a feature called '**Tags**' by concatenating all the transformed features. This **Tags** feature will be used to create vectors using the TFIDF method.

TFIDF Vectorizer :

TF-IDF stands for Term Frequency Inverse Document Frequency of records. It can be defined as the calculation of how relevant a word in a series or corpus is to a text. The meaning increases proportionally to the number of times in the text a word appears but is compensated by the word frequency in the corpus (data-set).

Recommender System :

This is going to be **Content Based Recommender System**.

A Content-based filtering uses item features to recommend other items similar to what the user likes, based on their previous actions or explicit feedback.

Our model takes movie as an item then its actors, director, description summary and genre as the most significant features of the movie. Users who watches any movie, then based on these features, other similar movies and shows will be recommended to them by our system.

These recommendations might match the description or which might have similar actors or director or those movies might have similar genre.

In order to find the similar data points, we used Cosine similarity to calculate distance. The 5 closest data points are going to be recommended which will be 5 most similar shows or movies.

```
recommend("Baahubali: The Beginning (Hindi Version)")
```

```
Baahubali: The Beginning (Malayalam Version)
Baahubali: The Beginning (English Version)
Baahubali: The Beginning (Tamil Version)
Baahubali 2: The Conclusion (Malayalam Version)
Baahubali 2: The Conclusion (Hindi Version)
```

```
recommend('Andaz Apna Apna')
```

```
Bodyguard
Taare Zameen Par
Raja Hindustani
Leap Year
Kills on Wheels
```

```
recommend('Bill Burr: Let It Go')
```

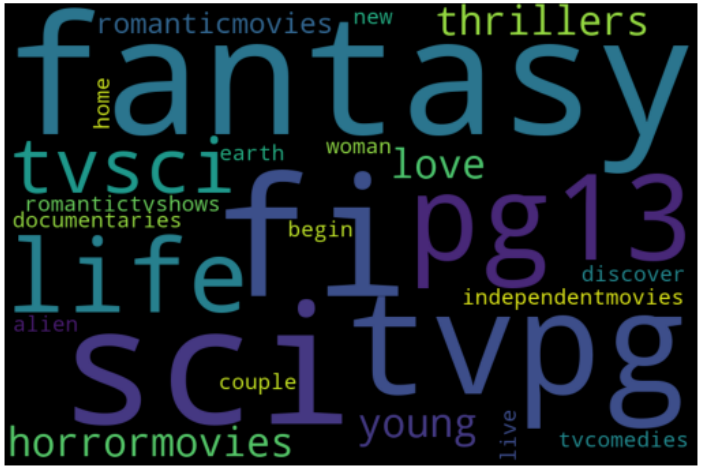
```
W. Kamau Bell: Private School Negro
D.L. Hughley: Clear
Patton Oswalt: Talking for Clapping
Gina Yashere: Laughing to America
Bill Hicks: Sane Man
```

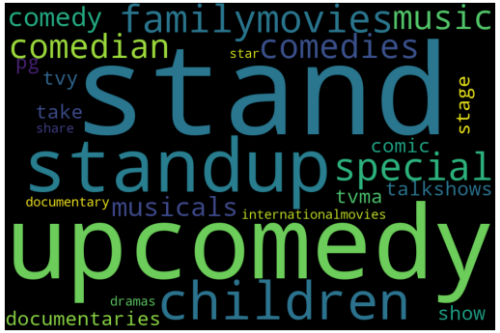
Topic Modelling :

LSA(Latent semantic analysis)

Latent semantic analysis (LSA) is a technique in natural language processing, in particular distributional semantics, of analyzing relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms. LSA assumes that words that are close in meaning will occur in similar pieces of text.

Word Cloud :





Clustering Techniques :

AI

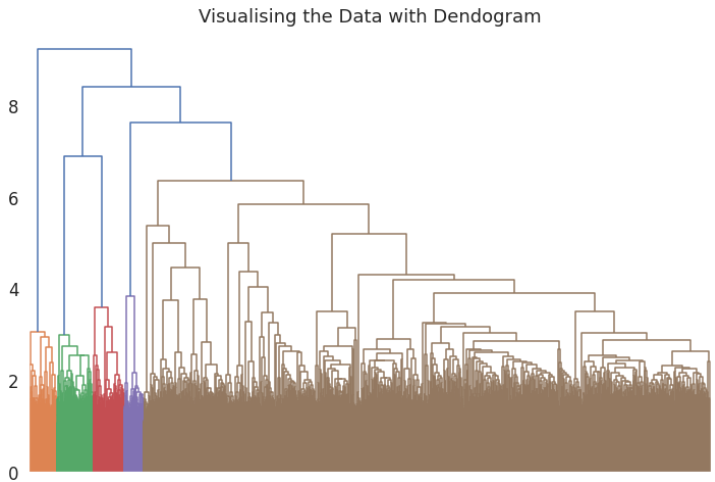
K-Means : K-Means is an iterative algorithm that divides the unlabeled dataset into k different clusters in such a way that each dataset belongs only one group that has similar properties.

It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.

Agglomerative Hierarchical clustering :

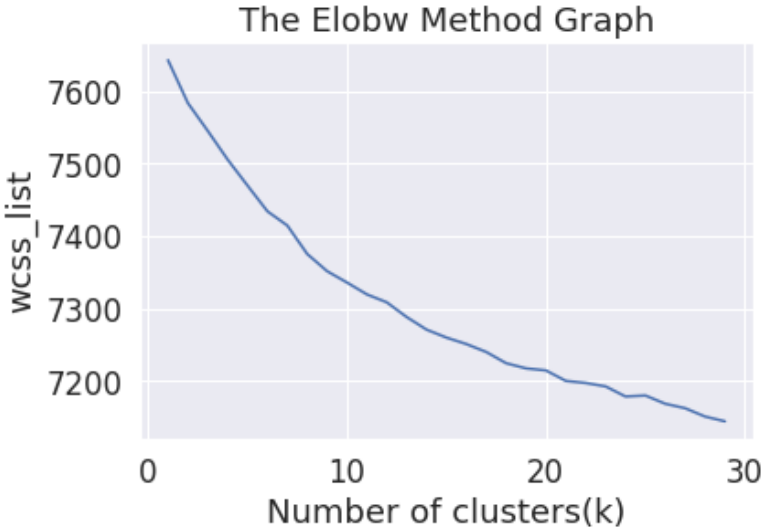
The agglomerative hierarchical clustering algorithm is a popular example of HCA. To group the datasets into clusters, it follows the bottom-up approach. It means, this algorithm considers each datapoint as a single cluster at the beginning, and then start combining the closest pair of clusters together. It does this until all the clusters are merged into a single cluster that contains all the datasets.

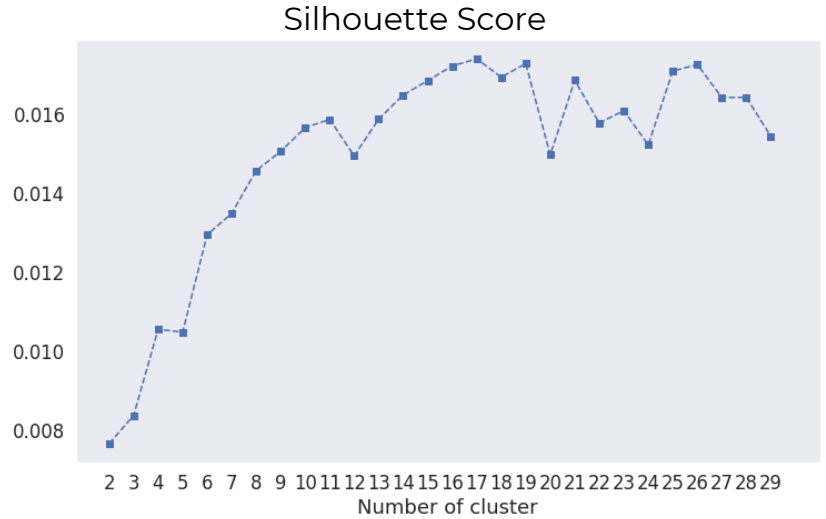
This hierarchy of clusters is represented in the form of the dendogram.



Model Performance of K-Means :

Considering Director and Cast features : We will consider Director & Cast features for clustering.

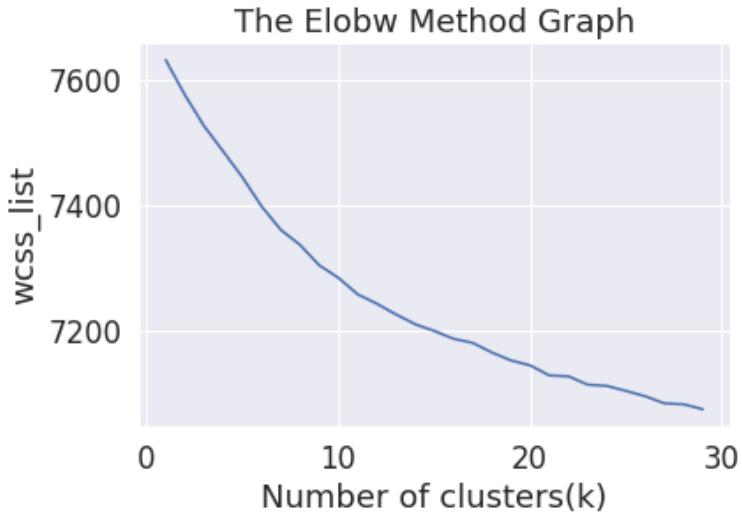


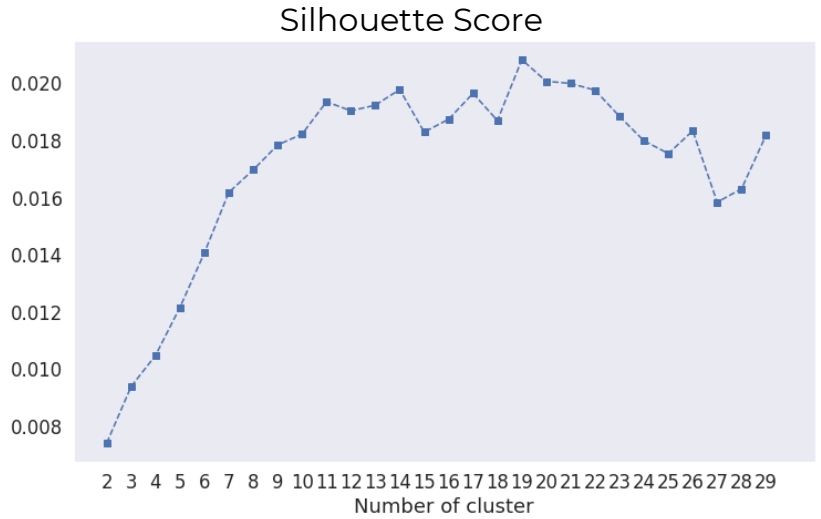


- From the elbow curve and Silhouette score, we see that 17 is the optimal number of clusters.
- The Silhouette score is 0.0174

Model Performance of K-Means :

Not Considering Director and Cast features : This time We will not be considering Director & Cast features for clustering.





- From the elbow curve and Silhouette score, we see that 19 is the optimal number of clusters.
- The Silhouette score is 0.0208

Evaluation Metric :

Silhouette Score : Silhouette Coefficient or silhouette score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1. Positive value Means clusters are well apart from each other and clearly distinguished. With the increasing value of the score, we can be confident to have better clusters.

since, we had better Silhouette score for the second case where we clustered without Director and Cast column, we shall go forward with that approach.

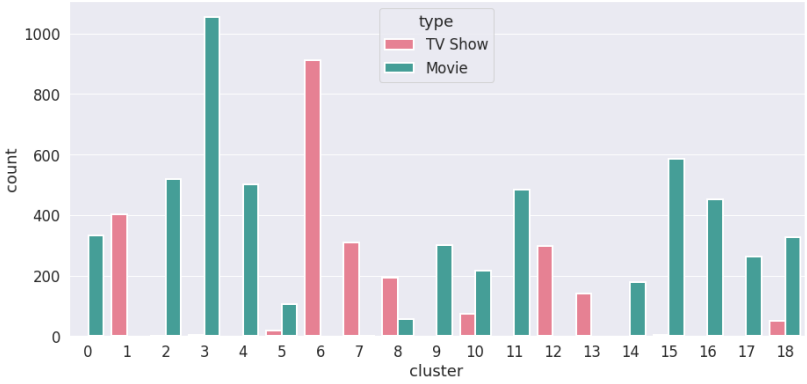
Scores and Model Selection :

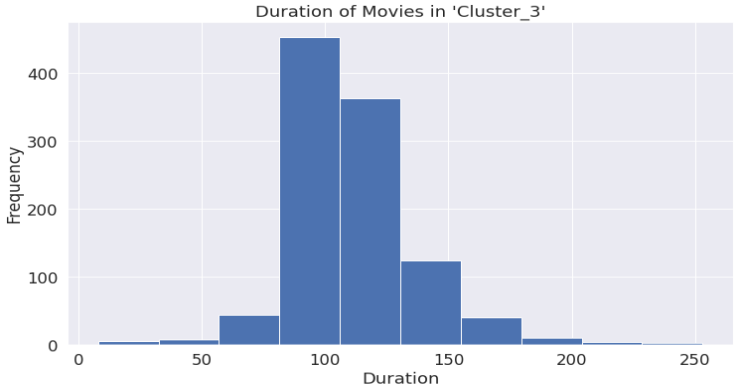
- Scores attained by **K-Means** clustering were the best all around. Clusters are well separated.
- **19** is the optimal number of clusters. We can confirm it by looking at the elbow curve and also from the Silhouette score.
- We also discarded **Director** and **Cast** features before clustering, since it was observed that these features are not helping us to achieve better scores. Also, talking intuitively, same directors and actors can make different kinds of shows. So, it is better to not consider them when clustering.

Clustering Analysis :

Among the clusters that were formed by K-Means algorithm, we find that cluster no. 3 and 6 are having the most data points.

Let's do some analysis on these two clusters -





- *Cluster 3* is dominated by movies whereas *Cluster 6* only contain TV Shows.
- Most contents in both *cluster 3* and *cluster 6* are targeted for mature audience (TV-MA) and children of the older age group (TV-14).
- Most movies in *cluster 3* have 100 - 120 minutes runtime.
- Most popular genre in *cluster 3* are Drama, International Movies and Comedies.
- Most TV shows in the *cluster 6* have genre of – International TV Shows, Dramas, Romantic shows, Comedies.

Conclusion :

Netflix is a streaming service that offers a wide variety of award-winning TV shows, movies, anime, documentaries and more – on thousands of internet-connected devices. New TV shows and movies are added every week!

Netflix has an extensive library of feature films, documentaries, TV shows, award-winning Netflix originals, and more.

In this unsupervised Machine Learning project, we tried to find out different patterns and popularity between various movies and shows. We clustered similar contents and developed a content based recommender system capable of recommending new movies and shows to users as per their interests.

Let's take a brief look at the observations that we made -

Observations :

During the EDA, we found-

- Dataset has 7787 rows and 12 columns. And it contains object, integer and float data types.
- Director, Cast, Country, DateAdded and Rating columns have null values.
- There is roughly 69% movies and 31% TV Shows present in the data.

Conclusion :

- Netflix has content that was released on 1925 to very recent releases like 2021
- The number of releases of movies and shows kept growing throughout the years. After 2019, we see a downfall in the releases. This a direct effect of Covid-19 Lockdown. The production in 2020 and 2021 is also low for the same reason.
- An interesting observations we can make is that in the last 2-3 years, Netflix focused heavily on TV shows compared to movies.
- 36% of all the contents are Netflix Originals. More number of "Original" movies were produced compared to TV shows.
- There is a total 118 unique countries in the data where movies and shows were produced. United States, India and United Kingdom are the top 3 most content producing country for Netflix.
- TV-MA is the most popular rating type, followed by TV-14, TV-PG & R.
- Most movies in the data have a duration of 90 - 120 minutes. Most shows have 1 or 2 seasons only.
- TV-Y rating suggests that the content is for children of all age. And we can see that, movies with this content rating has lowest average(median) runtime. NC-17 rating means that the content is not suitable for audience below 17 years of age. For this rating type, the movies have the highest duration or runtime.
- Documentaries, Stand-Up Comedy, Drama & International Movies are the top listed movies on Netflix. Kids' TV, International TV Shows, TV Dramas are top listed genre for shows.

Conclusion :

For text preprocessing, following methods were used -

- Punctuation Removal
- Lowercasing
- Stopwords removal
- Tokenization
- Lemmatization
- TF-IDF Vectorization

Then We built a content based recommender system that recommends 5 most similar movies and shows considering factors like same actors, directors, genre or contents that share a similar description or theme.

In the process of unsupervised modeling, our objective was to cluster similar movies and shows together based on their similar attributes.
For this, we tried **K-Means Clustering** and **Hierarchical Agglomerative Clustering**. We discarded Director & Cast columns before clustering because it was giving better results.

As evaluation metrics, we used **Silhouette** score and **Davies Bouldin** score.

- We choose **K-Means** to be the better algorithm for this problem.
- We referred the Elbow Curve and the Silhouette Scores to choose the best value for **K** which is the number of clusters.
- 19 was the optimal value for **K** in K-Means clustering.
- Cluster 3 and 6 are the top 2 clusters with highest number of observations.

Conclusion :

Topic Modelling :

We used **LSA**(Latent semantic analysis) to analyze the topics the topics and genres on Netflix. We made visualizations using Word Cloud for top few genres.

Analyzing the most populated clusters, we found -

- Most contents in this cluster are targeted for mature audience and children of the older age group (14 or more).
- Drama is the most repeated genre for this cluster. Other popular ratings are - International movies & shows, Comedies, Romance etc.
- Most movies in this cluster have 100 - 120 minutes runtime.

With that, we have come to an of our unsupervised project.

We hope that our thorough analysis prove useful for future applications like categorizing similar movies and shows. We found interesting insights for contents and countries that can be further looked into.

With the help of the content based filtering and recommendation system, we are able to make recommendations to users with the contents of their taste.

Thank You