# Capstone Project

## NETFLIX MOVIES AND TV SHOWS CLUSTERING

NITIN GOUR
(nitingour032@gmail.com)

# 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, I have done

- 1.Exploratory Data Analysis
- 2.Understanding what type content is available in different countries
- 3.Is Netflix increasingly focusing on TV rather than movies in recent years.
- 4.Clustering similar content by matching text-based feature

# Points to discuss

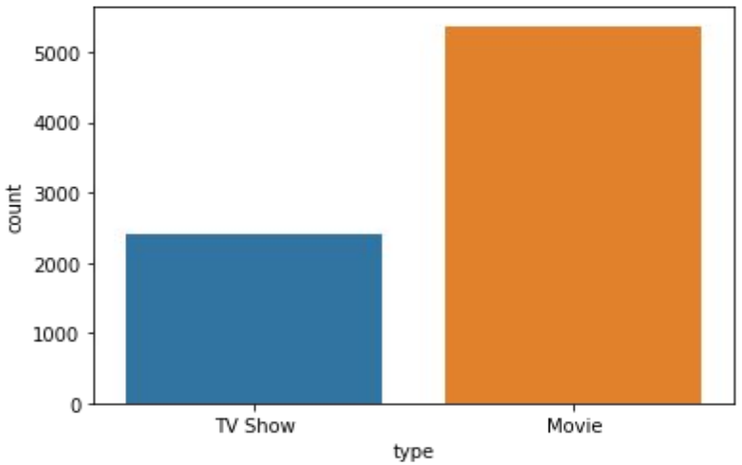
- Data description
- Exploratory data analysis
- Hypothesis testing
- Feature selection
- Machine learning algorithms(unsupervised)
    1. K-mean
    2. agglomerative clustering
- Model performance
- Conclusion

# Data description

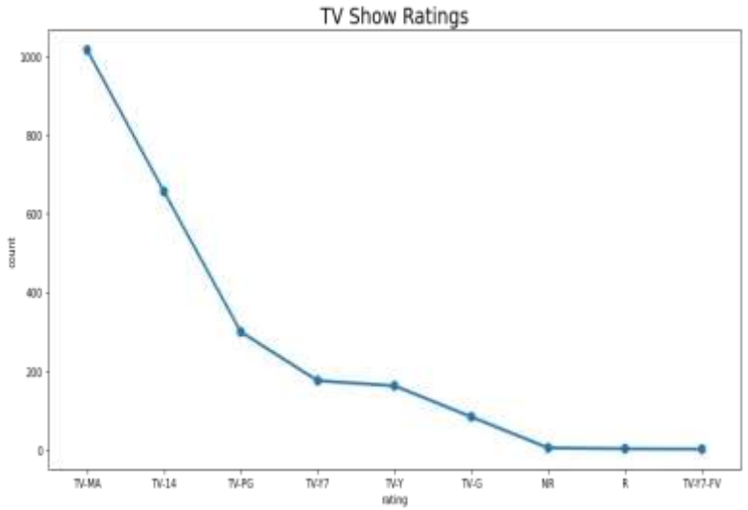
- The dataset consists of listings of all the movies and tv shows available on Netflix, along with details such as - cast, directors, ratings, release year, duration, etc

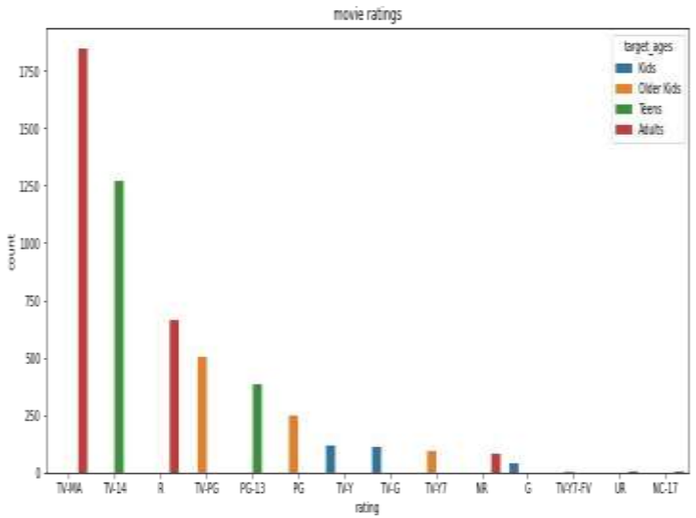
- **show_id** : Unique ID for every Movie / TV Show
- **type** : Identifier - A Movie or TV Show
- **title** : Title of the Movie / Tv Show
- **director** : Director of the Movie
- **cast** : Actors involved in the movie / show
- **country** : Country where the movie / show was produced
- **date_added** : 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

# Type




# Ratings

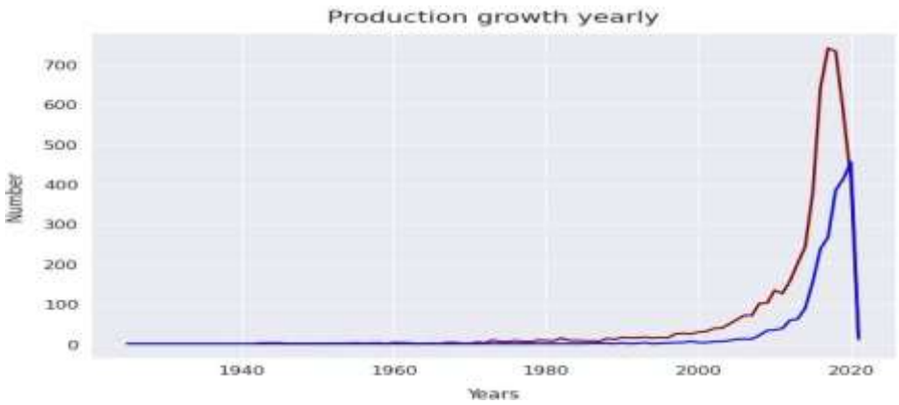



- Netflix has 5372 movies and 2398 TV shows,there are more number movies on Netflix than TV shows.

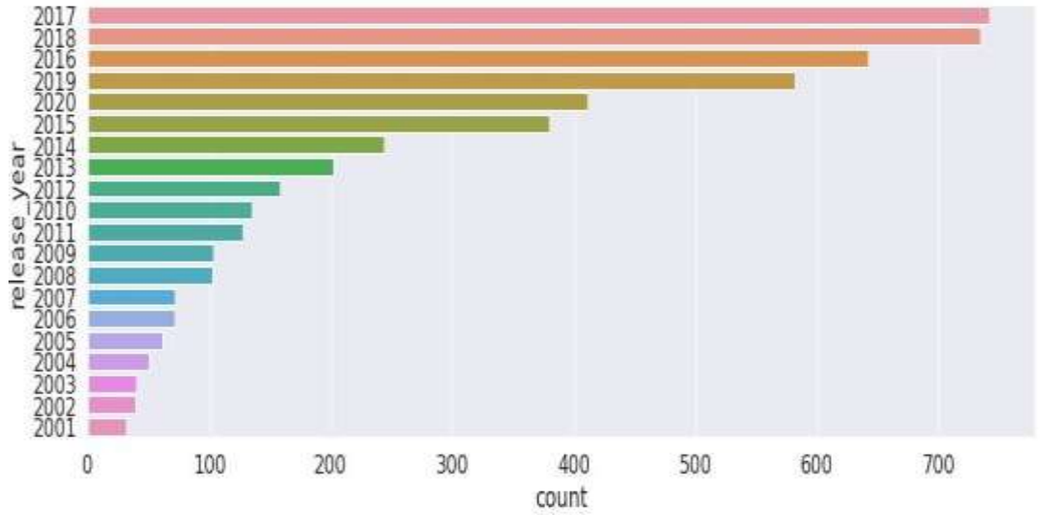
- TV-MA has the highest number of ratings for tv shows i,e adult ratings
- TV-MA has the highest number of ratings for movies i,e adult ratings
- in both the cases TV-MA has the highest number of ratings

# Release year

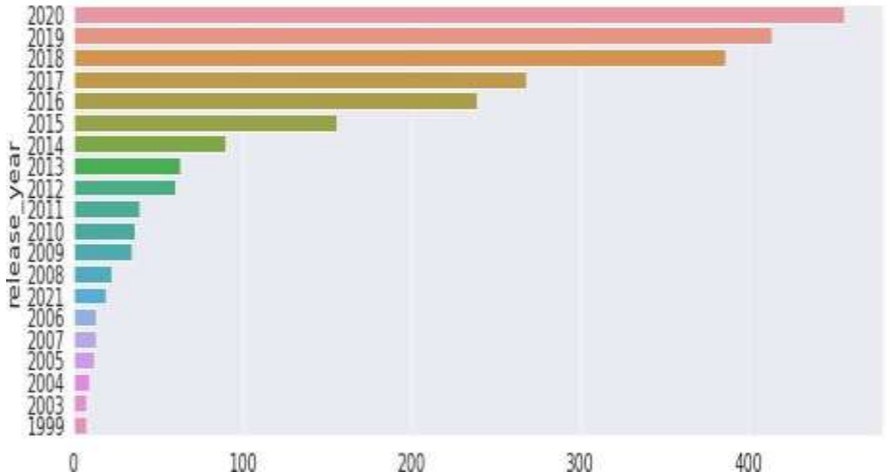



- highest number of movies released in 2017 and 2018
- highest number of movies released in 2020
- The number of movies on Netflix is growing significantly faster than the number of TV shows.
- We saw a huge increase in the number of movies and television episodes after 2015.
- there is a significant drop in the number of movies and television episodes produced after 2020.
- It appears that Netflix has focused more attention on increasing Movie content that TV Shows. Movies have increased much more dramatically than TV shows.

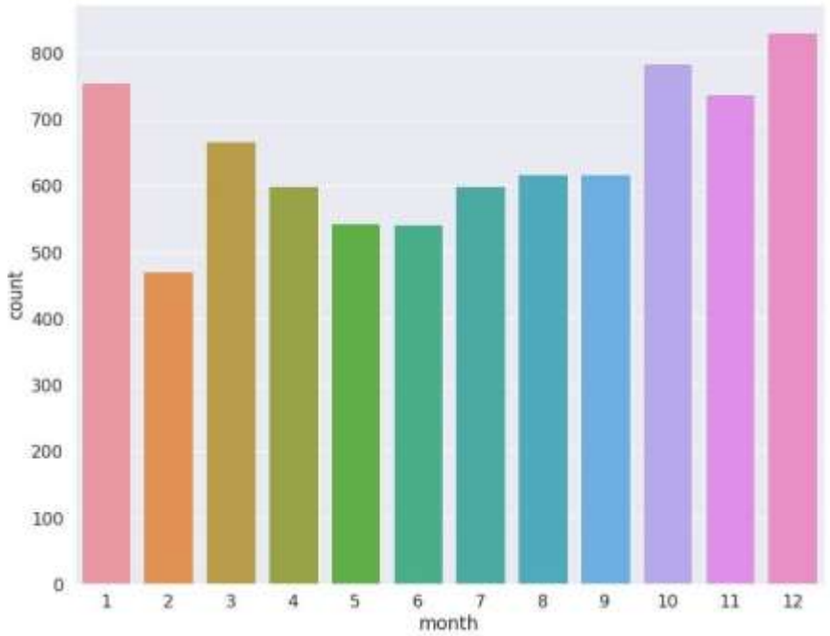
## tv_shows

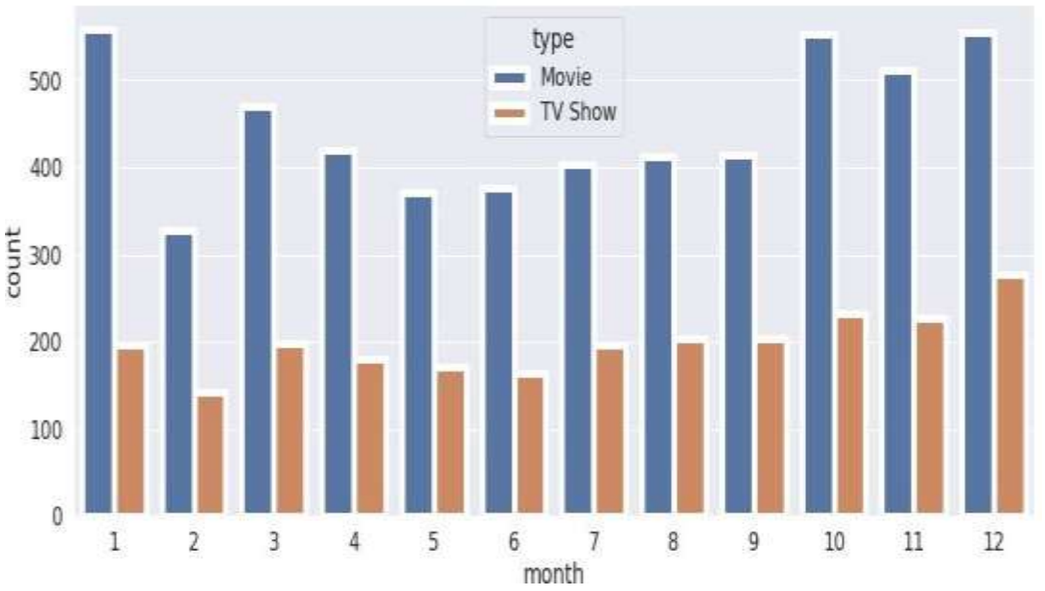



## movies



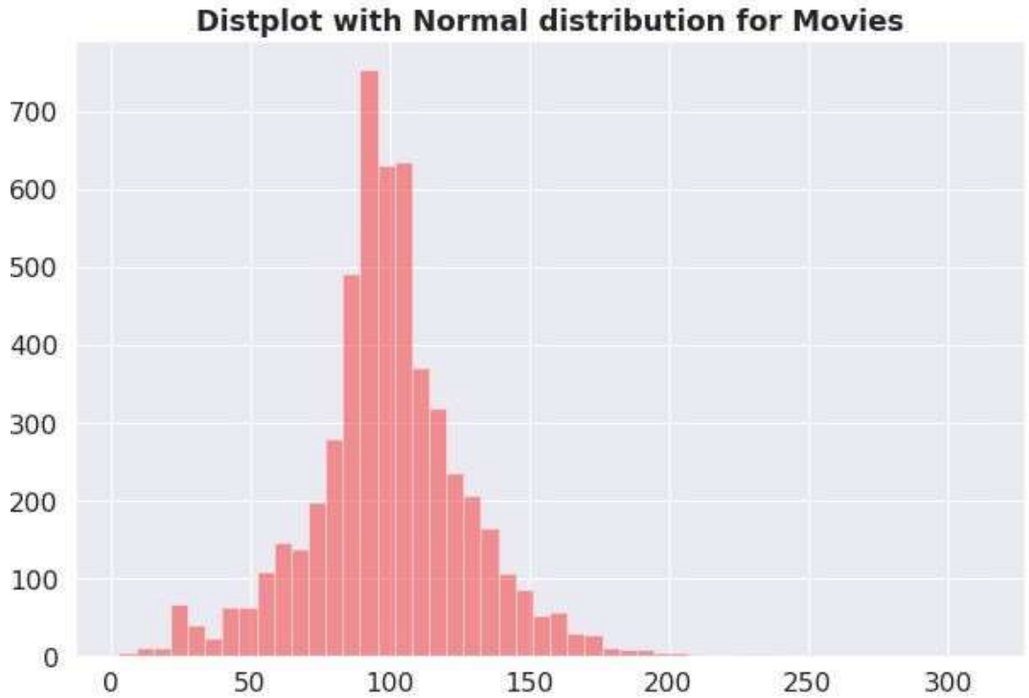
# Release month



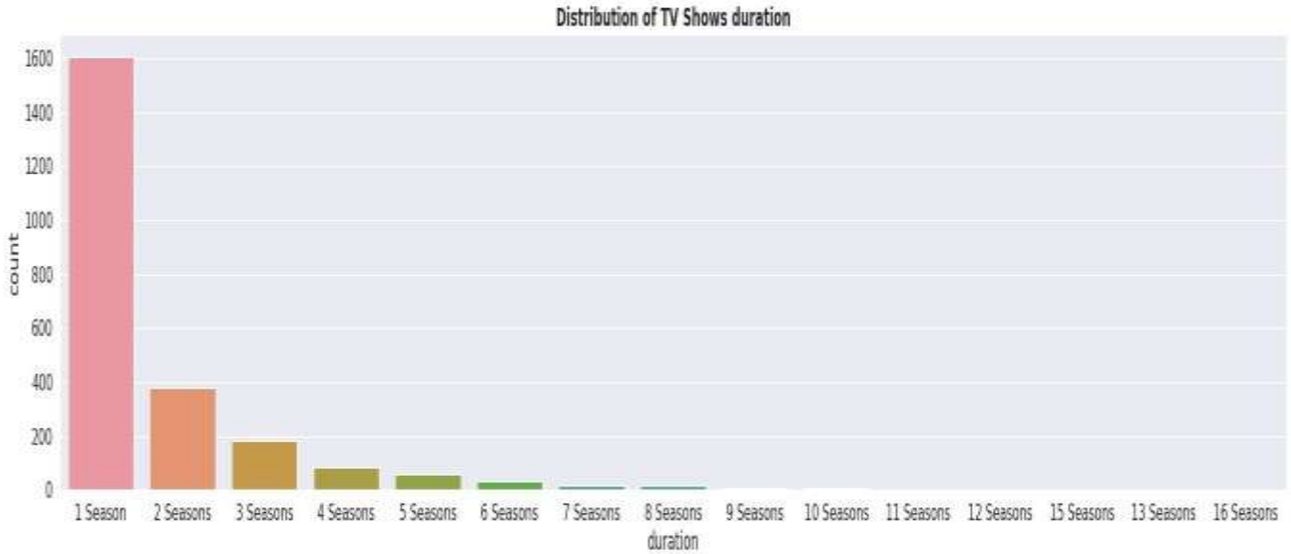

- The above graph shows that in both cases the most content is added to Netflix from october to january

# Duration

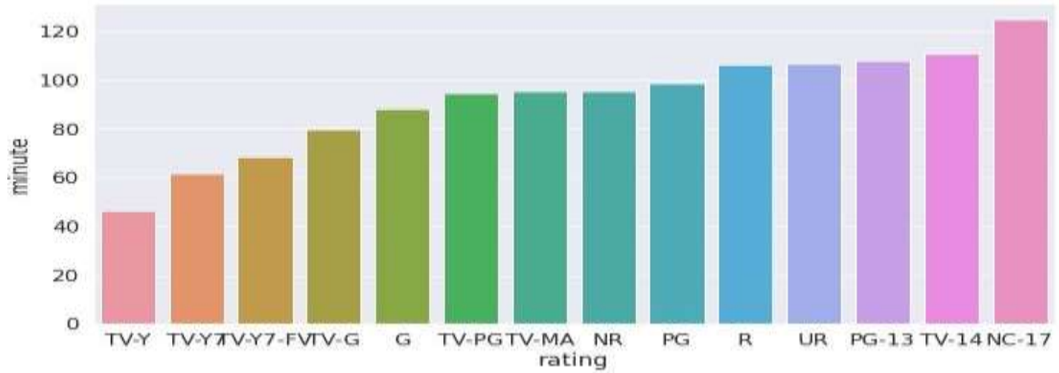



- most of the movies have duration of between 50 to 150

.

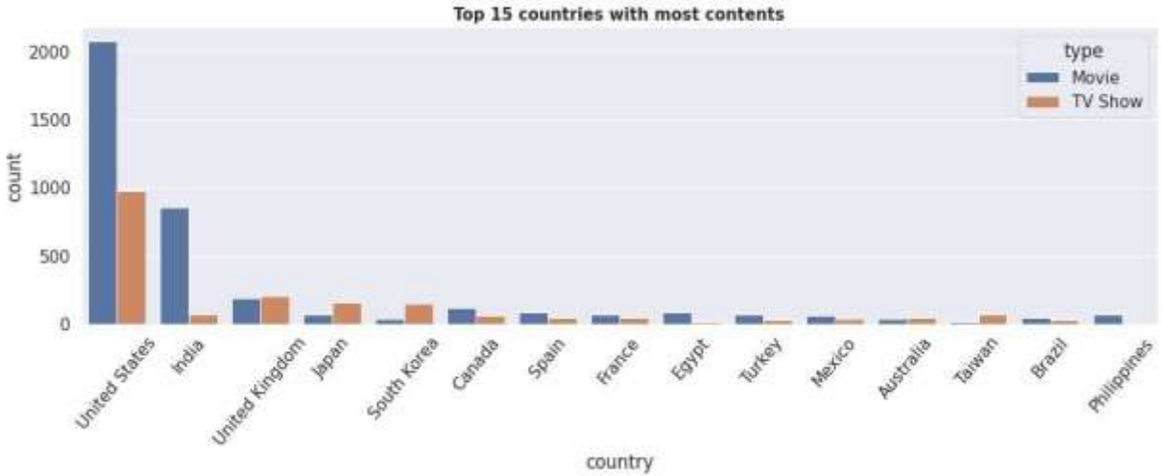


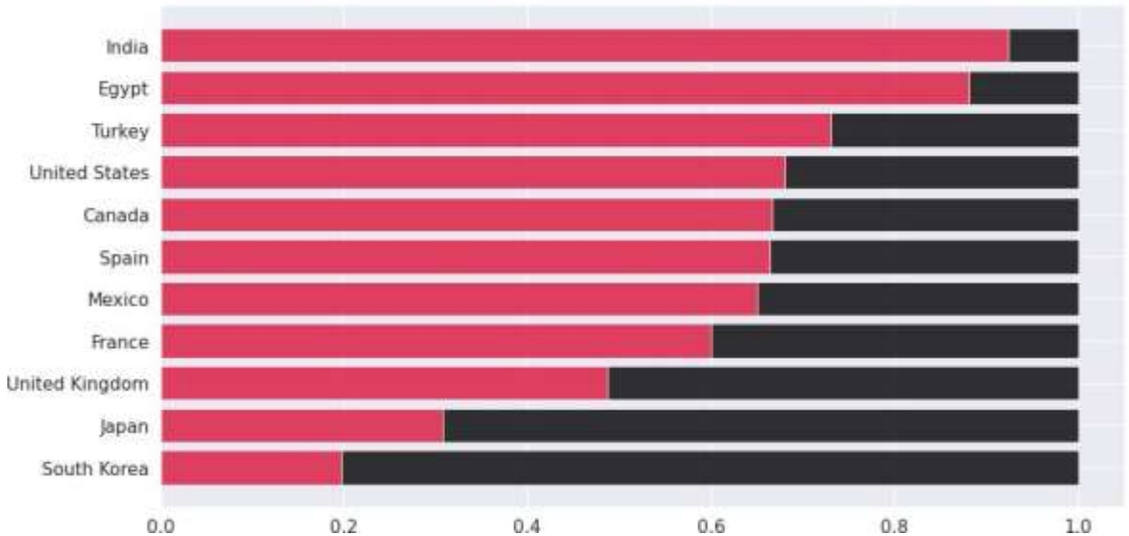

- highest number of tv_shows consistig of single season

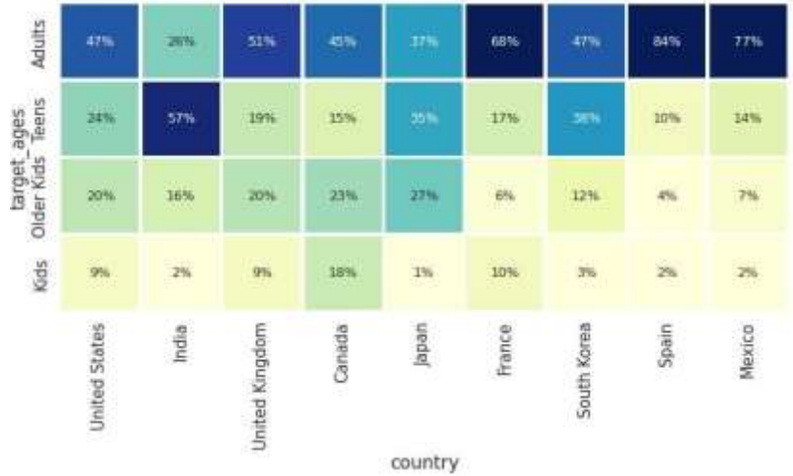



- Those movies that have a rating of NC-17 have the longest average

# Country

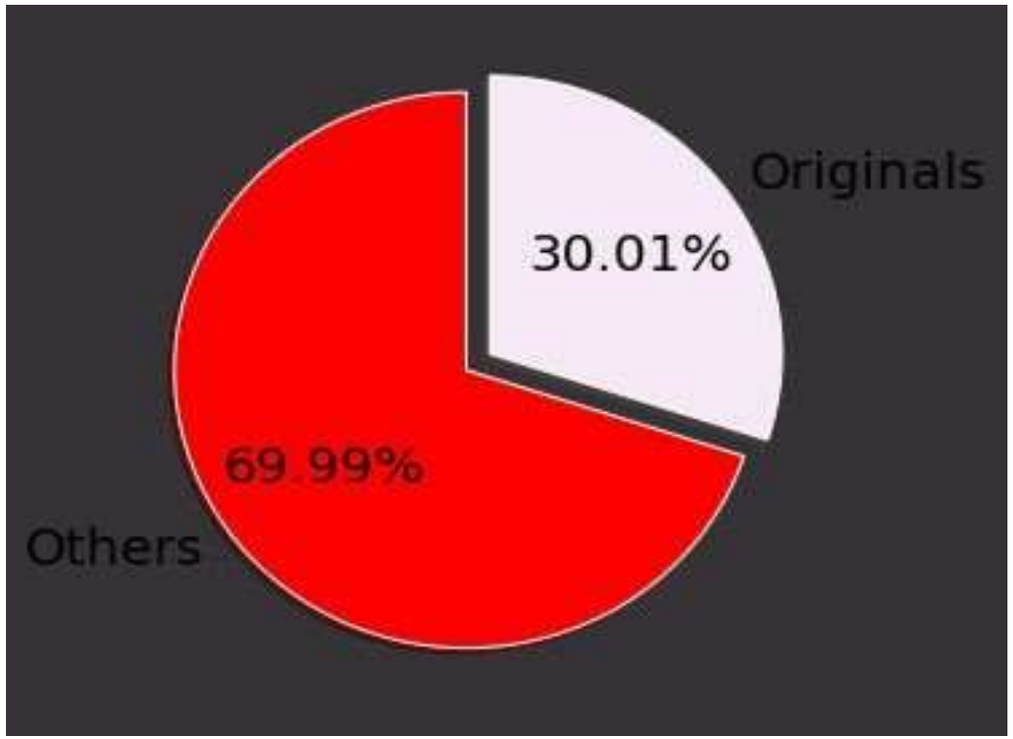



- unitated states has the highest number of content on the netflix ,followed by india
- india has highest number of movies in netflix
- the US and UK are closely aligned with their Netflix target ages, but radically different from, example, India or Japan!
- Also, Mexico and Spain have similar content on Netflix for different age groups.

# Originals




30% movies released on Netflix.

70% movies added on Netflix were released earlier by different mode.

# 1.HYPOTHESIS TESTING

Hypothesis testing in statistics refers to analyzing an assumption about a population parameter.

HO:movies rated for kids and older kids are at least two hours long

H1:movies rated for kids and older kids are not at least two hours long.

| | target_ages | duration |
|---|---|---|
| 0 | Kids | 66.486891 |
| 1 | Older Kids | 92.024648 |
| 2 | Teens | 110.025332 |
| 3 | Adults | 98.275145 |

- the t-value is not in the range, the null hypothesis is rejected.
- As a result, movies rated for kids and older kids are not at least two hours long.

2. H1:The duration which is more than 90 mins are  movies

HO:The duration which is more than 90 mins are NOT movies

| | target_ages | duration |
|---|---|---|
| 0 | Kids | 66.486891 |
| 1 | Older Kids | 92.024648 |
| 2 | Teens | 110.025332 |
| 3 | Adults | 98.275145 |

- the t-value is not in the range, the null hypothesis is rejected.
- As a result, The duration which is more than 90 mins are movies

# Features selection

- Initially Separating the column type into movie and tv shows
- here we are going to do Clustering similar content by matching text-based features so column description is one of the important feature
- Convert the text to lower case
- Tokenize the text
- Removed all the stop words and punctuation
- we use TF-IDF is an abbreviation for Term Frequency Inverse Document Frequency. This is very common algorithm to transform text into a meaningful representation of numbers which is used to fit machine algorithm for prediction

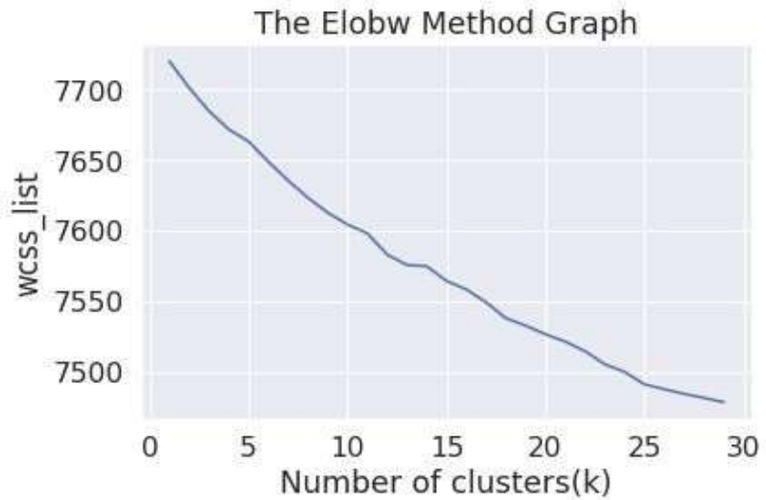
# ML algorithms(unsupervised machine learning)

- 1. K-mean
- 2. agglomerative clustering

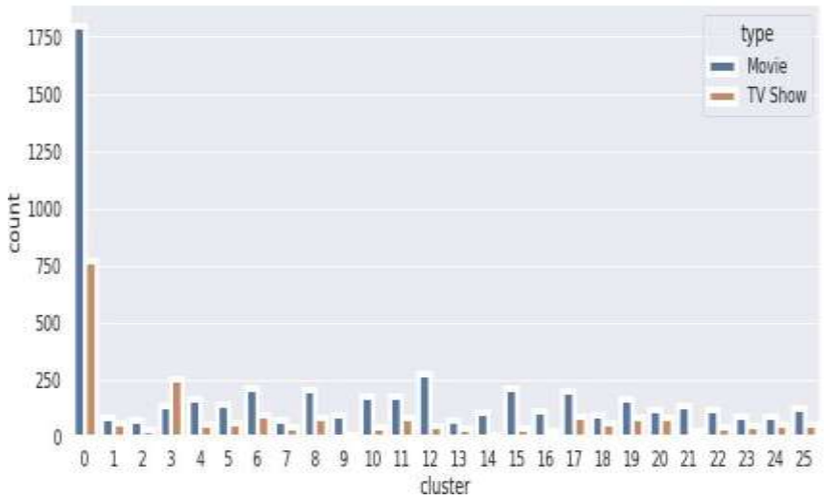
# K-Means:

- K-Means Clustering is an Unsupervised Learning algorithm which groups the unlabeled dataset into different clusters. Here K defines the number of predefined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

**Elbow Method**




- From elbow method generating 26 clusters

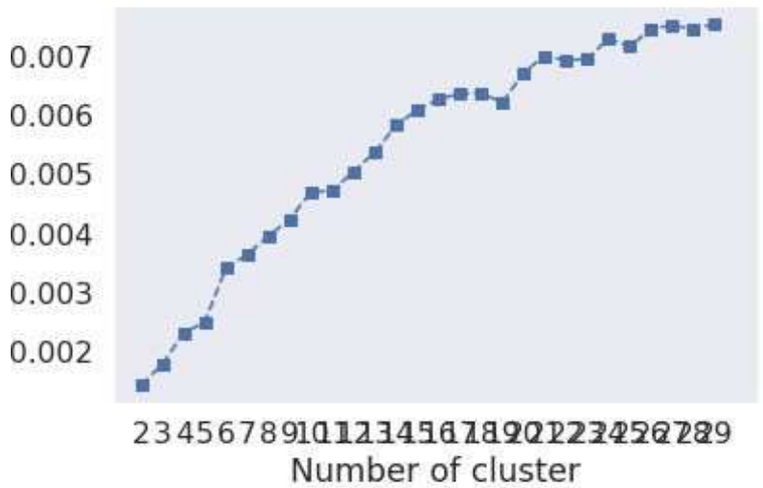



- cluster 0 has the highest number of datapoints and evenly distributed for other cluster

# evaluation

**1.Silhouette Score :** is a metric to evaluate the performance of clustering algorithm. It uses compactness of individual clusters(intra cluster distance) and separation amongst clusters (inter cluster distance) to measure an overall representative score of how well our clustering algorithm has performed

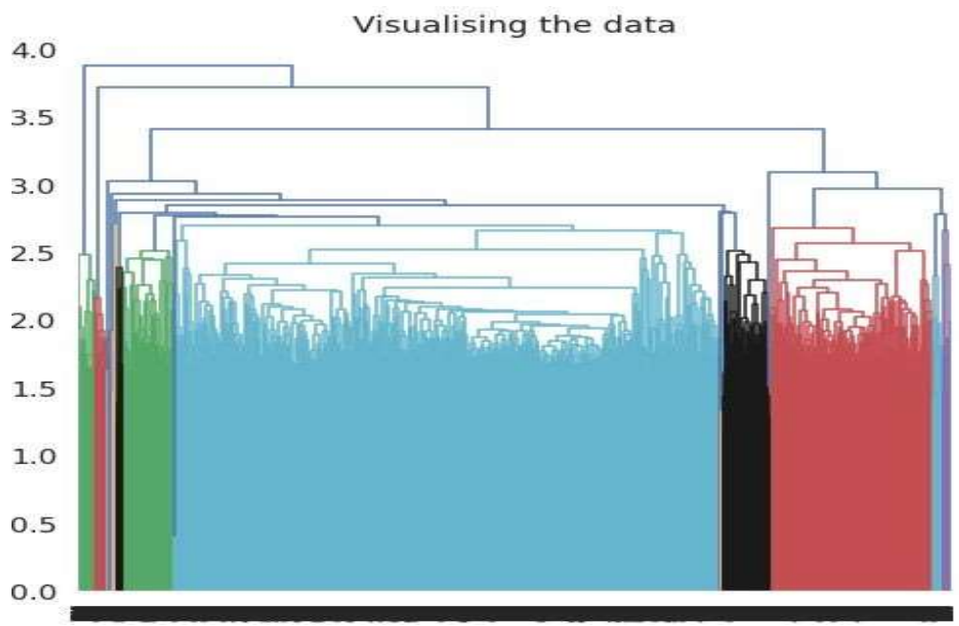
2.**The Davies-Bouldin index (DBI** ).It is most commonly used to evaluate the goodness of split by a K-Means clustering algorithm for a given number of clusters.




- silhouette score would always lie between -1 to 1. 1 representing better clustering
- Silhouette score is 0.007499010681200968
- Davies_bouldin_score is 9.05605194948868
- so model is performing well

# 2.Agglomerative Clustering

- In agglomerative clustering no need to give the value of k beforehand
- The agglomerative hierarchical clustering algorithm is a popular example of HCA
- Here I used ward linkage




- the optimal number of clusters is 4 using the Dendrogram

- **Evaluation**

- Silhouette Coefficient: -0.002
- Davies_bouldin_score is 9.05605
- Comparing with K mean only Davies_bouldin_score is better for hierarchical clustering Model is not performing well

# Conclusion

- from elbow and sillhoute score ,optimal of 26 clusters formed , K Means is best for identification than Hierarchical as the evaluation metrics also indicates the same.In kmean cluster 0 has the highest number of datapoints and evenly distributed for other cluster
- Netflix has 5372 movies and 2398 TV shows, there are more movies on Netflix than TV shows.
- TV-MA has the highest number of ratings for tv shows i,e adult ratings
- highest number of movies released in 2017 and 2018 highest number of movies released in 2020 The number of movies on Netflix is growing significantly faster than the number of TV shows. We saw a huge increase in the number of movies and television episodes after 2015. there is a significant drop in the number of movies and television episodes produced after 2020. It appears that Netflix has focused more attention on increasing Movie content than TV Shows. Movies have increased much more dramatically than TV shows

- the most content is added to Netflix from october to january
- Documentaries are the top most genre in netflix which is followed by standup comedy and Drama and international movies
- kids tv is the top most TV show genre in netflix
- most of the movies have duration of between 50 to 150
- highest number of tv_shows consisting of single season
- Those movies that have a rating of NC-17 have the longest average duration.
- When it comes to movies having a TV-Y rating, they have the shortest runtime on average
- united states has the highest number of content on the netflix ,followed by india
- india has highest number of movies in netflix
- 30% movies released on Netflix. 70% movies added on Netflix were released earlier by different mode