Capstone Project 4
Netflix Movies & Tv show Clustering

Nitesh Bhowmick

Content

Problem Statement

- Prediction based on Country
- Prediction based on Rating
- Prediction based on production growth
- Prediction based on top 10 genre movies & TV show
- Prediction based on Actor with most content on platform
- Prediction based on Top 15 countries with most contents
- Prediction based on Highest No in One season

Data Summary

- In the Netflix movies & Tv shows project there is a dataset which contains Netflix Tv shows & movies information.

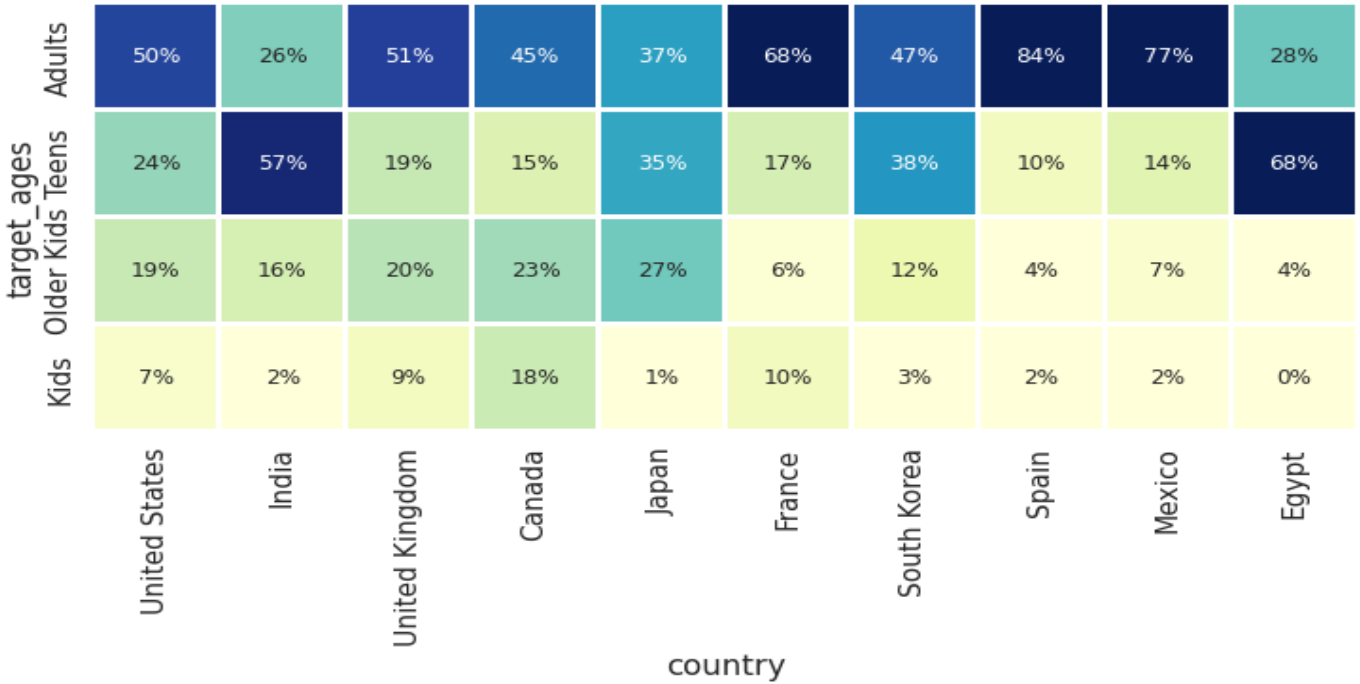
- Netflix has 7787 rows and 12 columns

- Data contain information on rating, duration, Genre, country, director, Title, release year.

Exploratory Data Prediction

- Exploratory Data Prediction is also known as EDA, is the process of interpreting datasets by summarizing their key properties and frequently them
- EDA refers to the critical process of performing initial investigation on datasets so as discover the patterns, to spots anomalies , to hypothesis, and to check assumptions with the help of summary statistics and graphical representation
- In EDA, plotting option include box plot, line plots , scatter plots and many more.

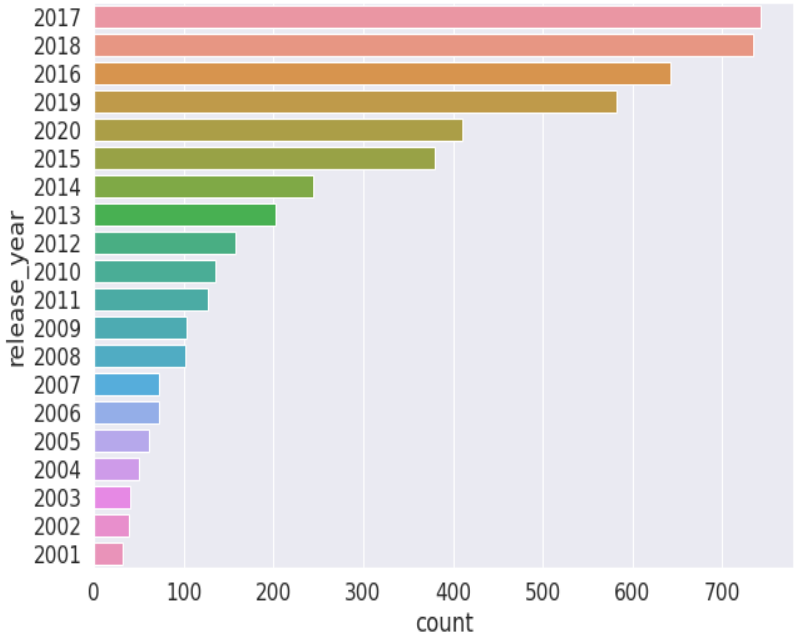
Country

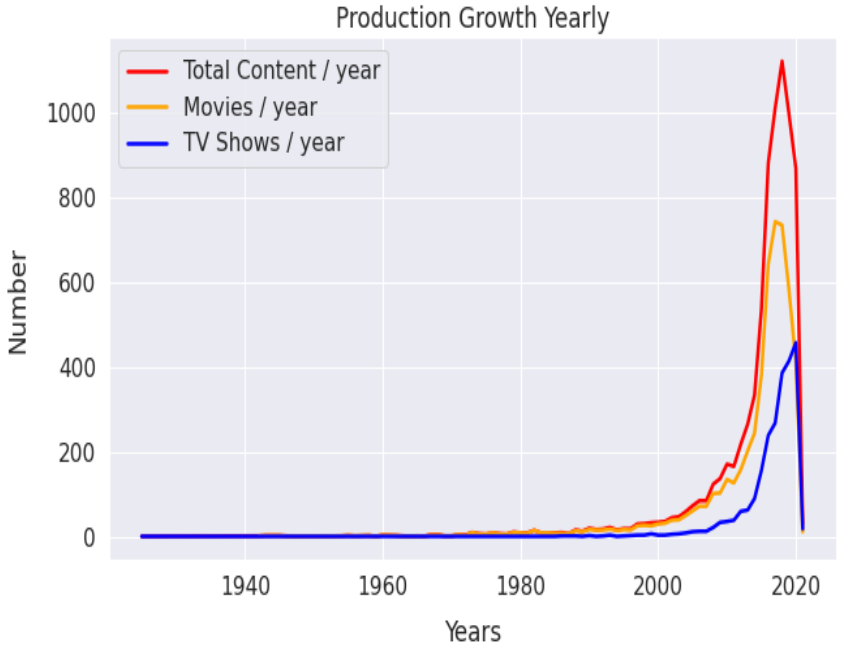


US and UK are closely aligned with their Netflix target ages, but radically different from, example, India or Japan!

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

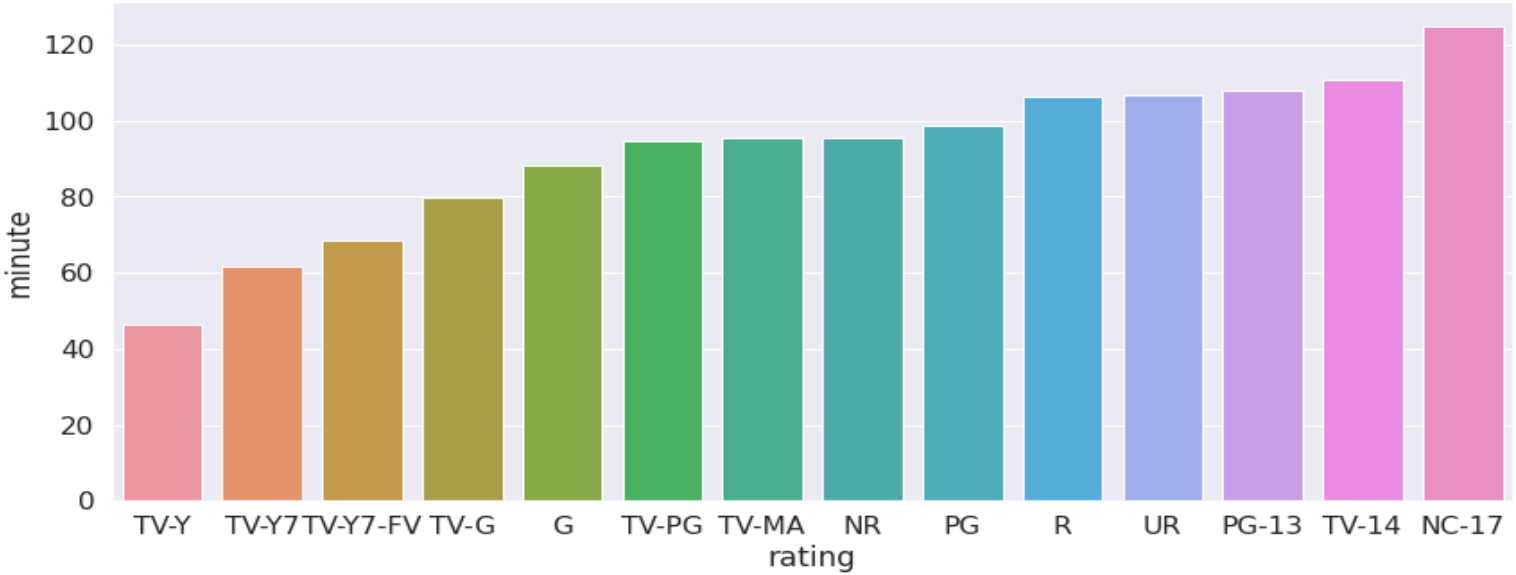
Production Growth





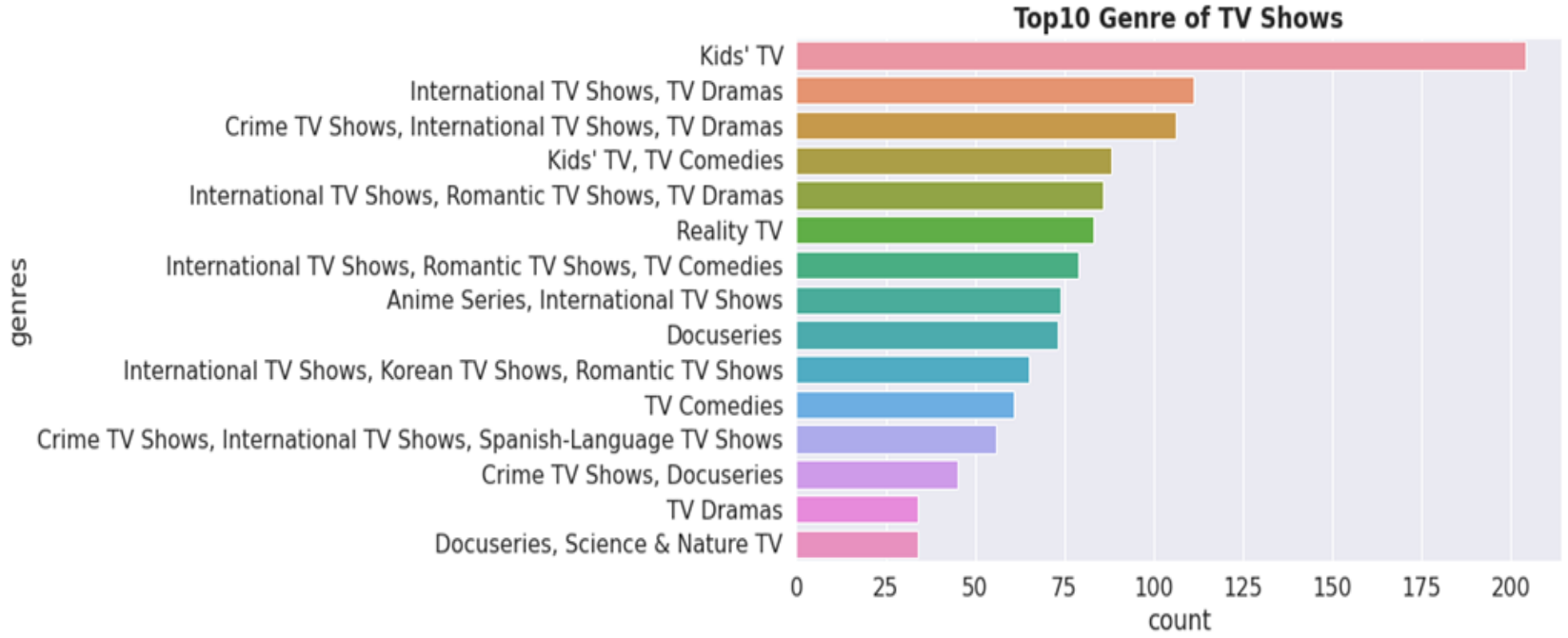
Highest no of movies released in 2017 & 2018

Highest no Rating of single season



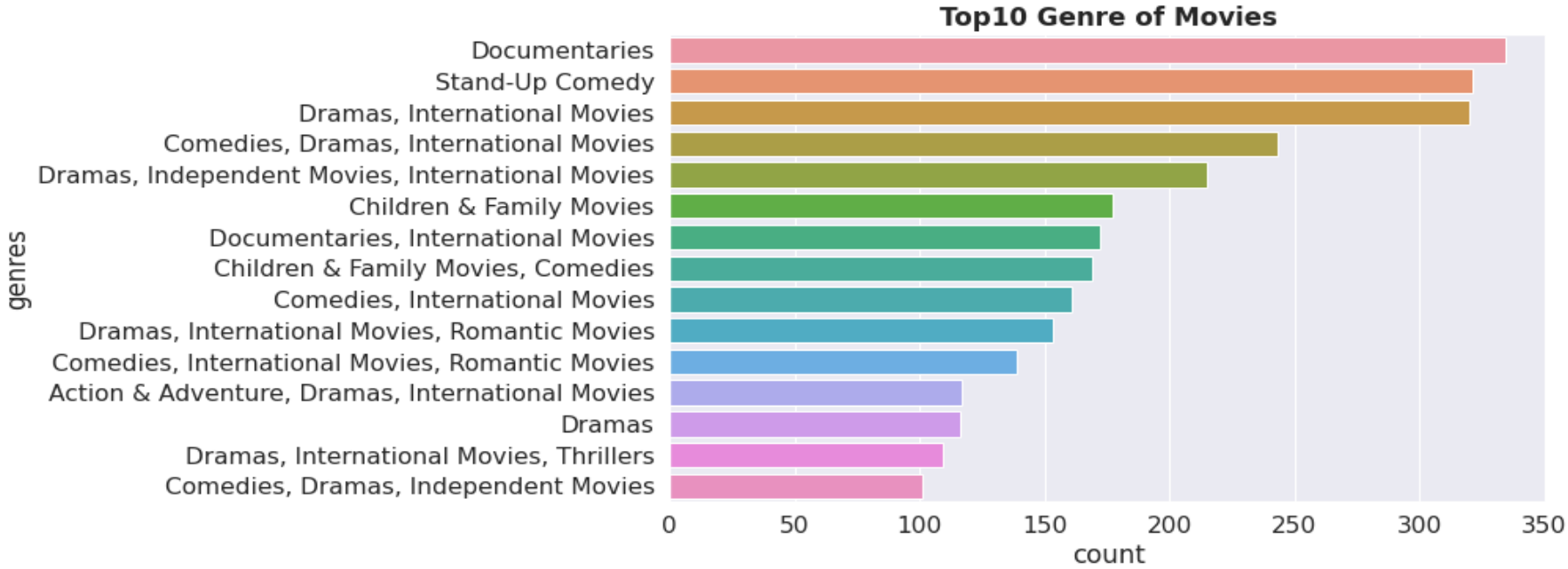
- Those movies that have a rating of Nc-17 means adults, they have the longest average duration
- when it comes to that movies having a TV- Y means kids , they have shortest runtime

Top 10 Genre of TV Show



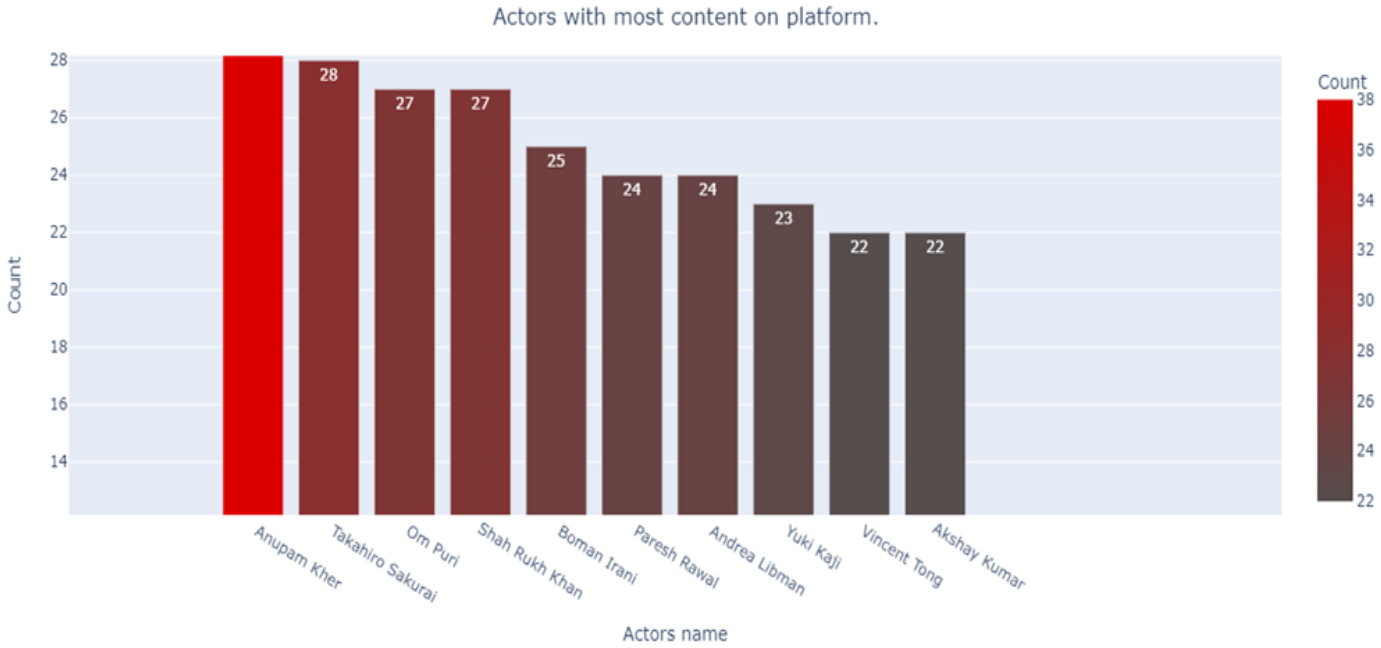
kids Tv is the top most Tv show Genre in Netflix

Top 10 Genre in Movies



Documents are the top most genre in Netflix which is followed by stand up comedy and drama and international movies

Top 10 Actors



- There are 6 of actor in the top 10 of list of number TV shows and movies are from India .

Machine learning

- Machine learning is a type of artificial intelligent that allow software application to become more accurate at predicting outcome without being explicitly programmed to do so machine learning algorithms use historical data as input to predict new output value
- Machine learning are a common use case for machine learning other popular uses include fraud detection, business process automation (BPA) and predictive maintenance

Why machine learning is important

- Machine learning is important because it gives enterprises a view of trends in customers behavior and business operational patterns , as well as supports the development of new products. Many of today leading companies such as Facebook, Google and Uber, make machine learning a central part of their operation. Machine learning has become a significant competitive differentiator for many companies

Different types of machine learning

- **Supervised learning :** In this types of machine learning, data scientists supply algorithms with labeled training data and defined the variables they wants the algorithms to assess for correlation.
- **Unsupervised learning :** This types of machine learning involves algorithms that train on unlabeled data.
- **Reinforcement learning:** It is multi step process for which there are clearly defined rules Data scientist program algorithm complete a task

CLUSTERING

Clustering is a type of unsupervised learning method of machine learning .in the unsupervised learning method . The inferences are drawn from the datasets which not contain labelled output variable. It is an exploratory data analysis technique that allow us to analyze he multivariate data sets.

clustering is a task of dividing the data sets into a certain number of cluster in such a manner that the data points belonging to a cluster have similar characteristics clusters are nothing but the grouping of data points such that the distance between the data point s within the clusters is minimal clustering is done to segregate groups with similar traits.

K Mean Clustering.

K-means clustering is an unsupervised learning algorithm. There is no labeled data for this clustering, unlike in supervised learning . K-means performs the division of objects into clusters that share similarities an are dissimilar to the object belonging to another cluster

Advantage & disadvantage:

If the variable are huge , then k means most of the times computationally faster than hierarchical clustering, and on other hand difficult to predict k- value .

K –means produce tighter cluster than hierarchical clustering and other side global cluster , it did not work well .

Dendrogram

A dendrogram is a branching diagram that represents the relationships of similarity among a group of entities & its a type of tree diagram showing hierarchical clustering — relationships between similar sets of data. They are frequently used in biology to show clustering between genes or samples, but they can represent any type of grouped data.
The most common example of a dendrogram is the tiered diagram used to display the playoff games and progress of some sporting event, like hockey, basketball or baseball. Each of the teams that makes the playoffs is listed, along with the games they need to win in order to make it to the finals.

Advantages of Hierarchical Clustering:

- We can obtain the optimal number of clusters from the model itself, human intervention not required.
- Dendrograms help us with clear visualization, which is practical and easy to understand.

Disadvantages of Hierarchical Clustering:

- Not suitable for large datasets due to high time and space complexity.
- There is no mathematical objective for Hierarchical clustering.
- All the approaches to calculating the similarity between clusters have their own disadvantages.

Agglomerative Clustering

The agglomerative clustering is the most well known kind of variable leveled clustering used to gather in bunches based on their comparability ,its otherwise called agglomerative clustering.

Advantage & disadvantage:

- No need any information about how many number of clusters are required And easy to used on other side we can not take a step back in this algorithm and Time complexity is higher at least 0(n^2logn)
- It can produce an ordering of objects, which may be informative for the display.
- The time and space complexity of agglomerative clustering are more than K-means clustering, and in some cases, it is prohibitive.

Challenges

- Understand the column of the dataset.
- Analyze and visualization of the Netflix Tv shows & movies according to feature.
- Find the right chart to show the chart.
- Difficulties to find out the correct model.

Conclusion

- From elbow and silhouette score, optimal of 26 cluster formed, k means is the best for identification than hierarchical as the evaluation metrics also indicates the same .
- Netflix has 5372 movies and 2398 TV shows , there are more number movies on Netflix than TV shows.
- TV_MA has the highest no of rating for TV shows I.e. adult rating
- Highest number of movies released in 2017 and 2018.
- Most content is added to Netflix from October to January
- Kids TV is the top most TV show genre in Netflix
- Most of the movies have duration of between 50 to 150.

Q/A