Capstone Project
Netflix Movies and Tv Shows Clustering

Team Members

Pruthvi Raj
Yogesh Reddy

Abstract:

- The dataset consists of tv shows and movies available on Netflix as of 2019. The dataset is collected from Flixable which is 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 services number of movies has decreased by more than 2000 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.
- Here we will perform different EDA techniques and clustering on the data to get ideal clusters for this problem. And also, we will build recommender system.

Problem Statement:

- The data is related to Netflix movies and tv shows till 2019. It is understandable from the data that tv shows number has been tripled from 2010 and the number of movie titles has been reduced by 2000. We have to find several insights from the data using EDA and also, have to cluster the data based on the data

Introduction:

In this project we have to perform the following

1. Exploratory Data Analysis
2. Understanding what type of content is available in different countries
3. Is Netflix focusing more on tv shows rather on movies
4. Cluster similar content by matching text data.

Data Summary:

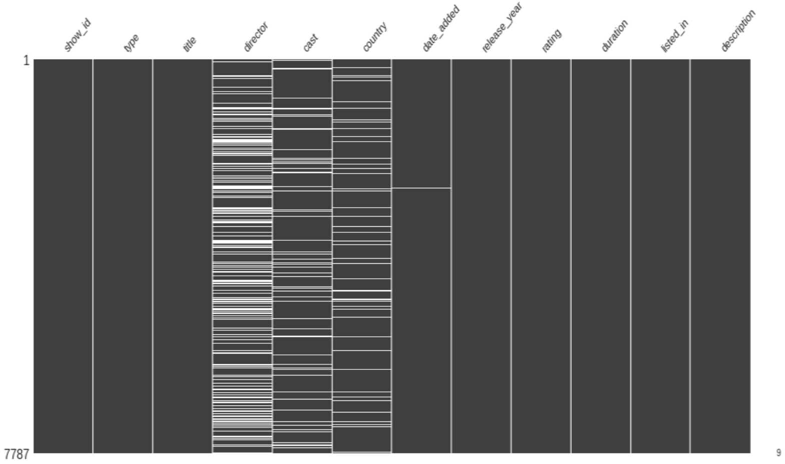
- The Provided data has 7787 observations.
- The dataset contains 7787 rows and 12 columns.

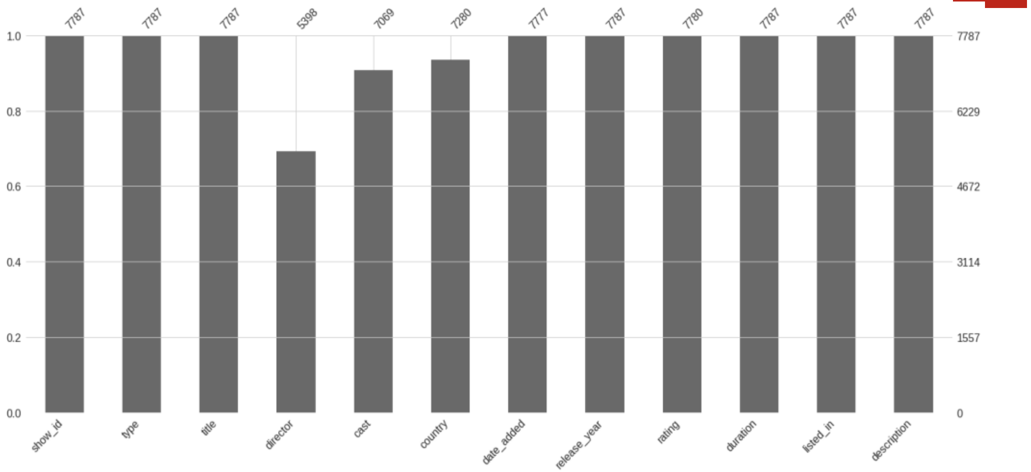
1. show_id: Unique Id for every Movie/Tv show
2. type: Identifier – A movie or tv show
3. title : title of the movie/tv show
4. director : Director of the movie
5. cast : actor involved in the movie / tv show
6. country : Country where the movie/ show was produced
7. date-added: date it was added to Netflix
8. release_year: actual release year of the movie/show.

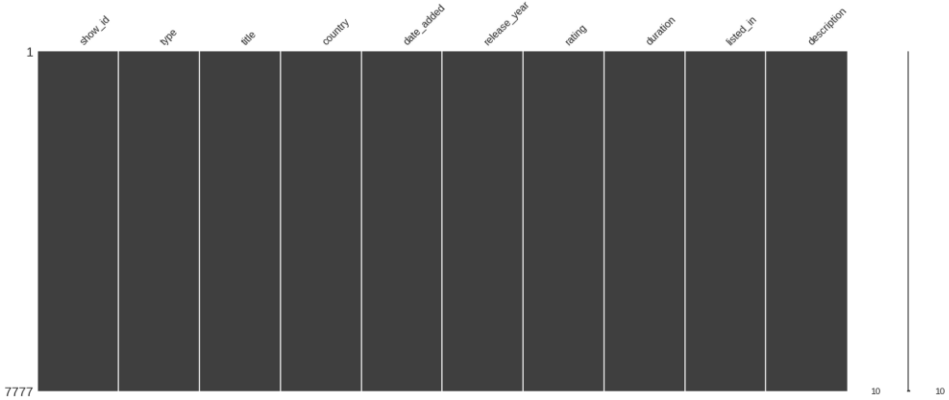
9. rating: rating of the movie/ tv show
10. duration : total duration in minutes or number of seasons
11. listed_in : genre
12. description : the summary description

Exploratory Data Analysis:

- We have performed several EDA techniques to clearly check how each and every feature is behaving with respect to the dependent feature y. We have also checked for null values, duplicate values luckily there were none. We have visualised several plots which gives several insights of all the features and we can see which features are more important.



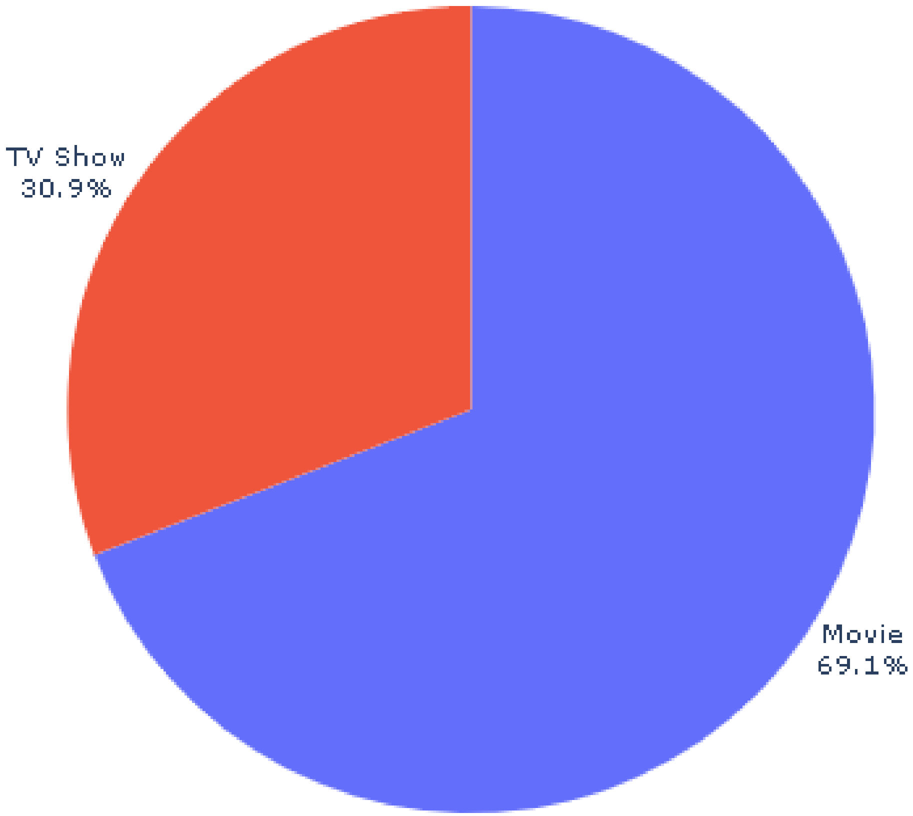




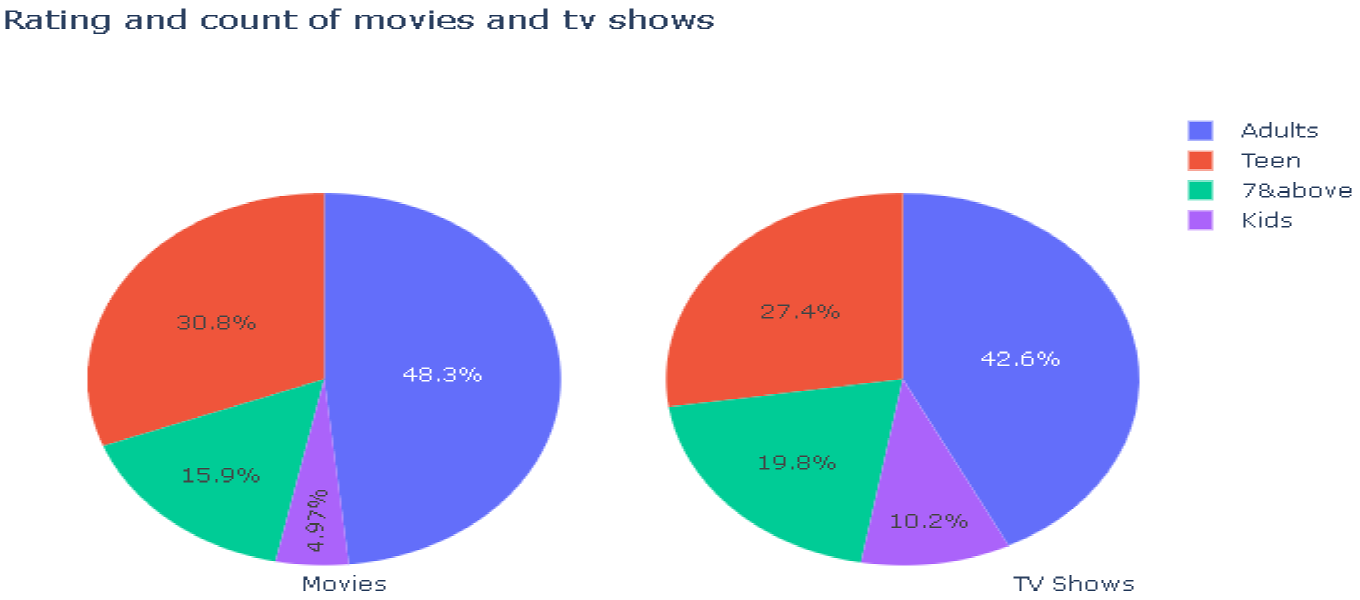
These plots shows the null values in different

features and the last plot shows the final

dataset with zero null values.

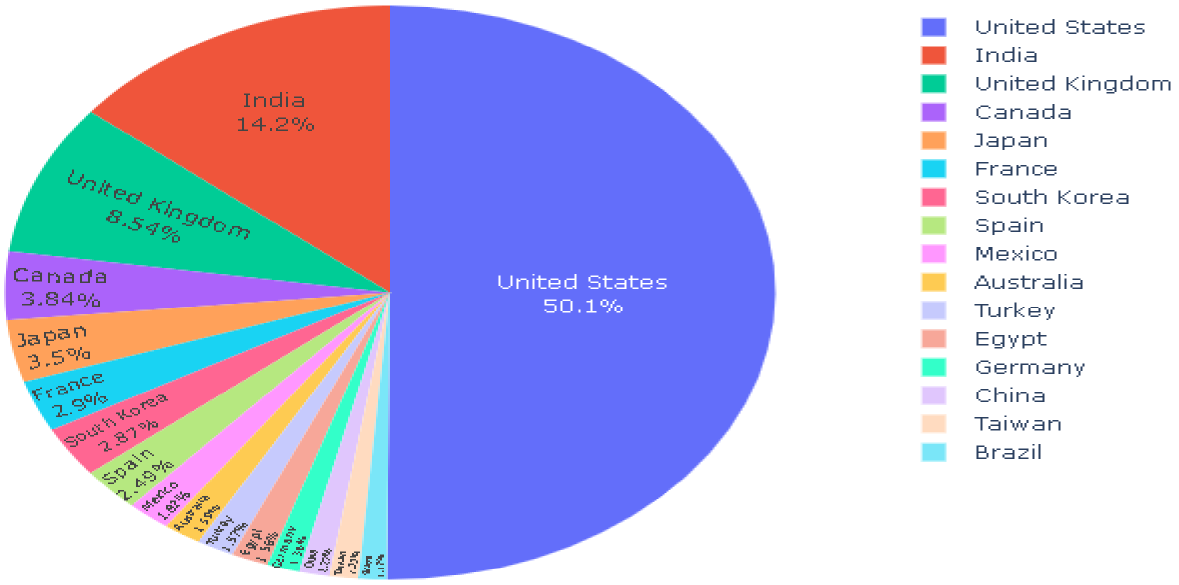


The above pie plot shows the total distribution of movies and tv shows

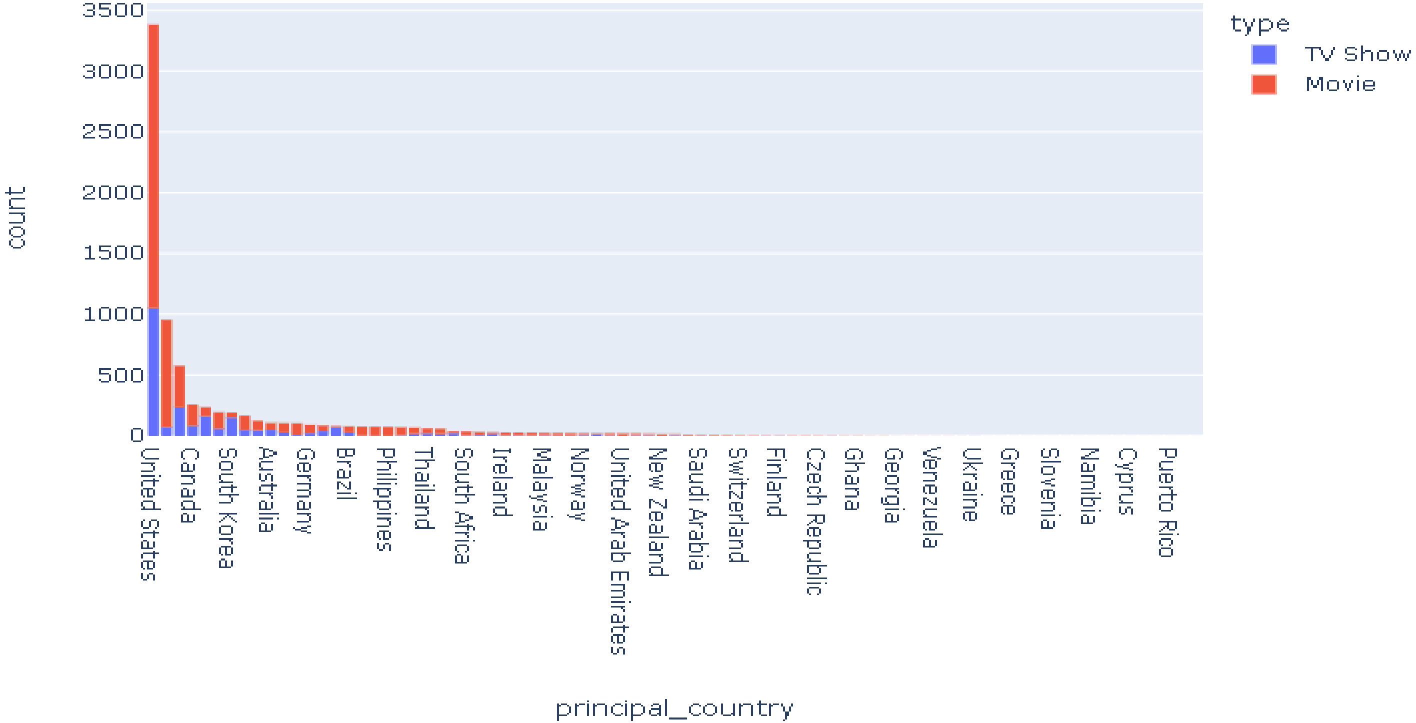


Rating and count of movies and tv shows

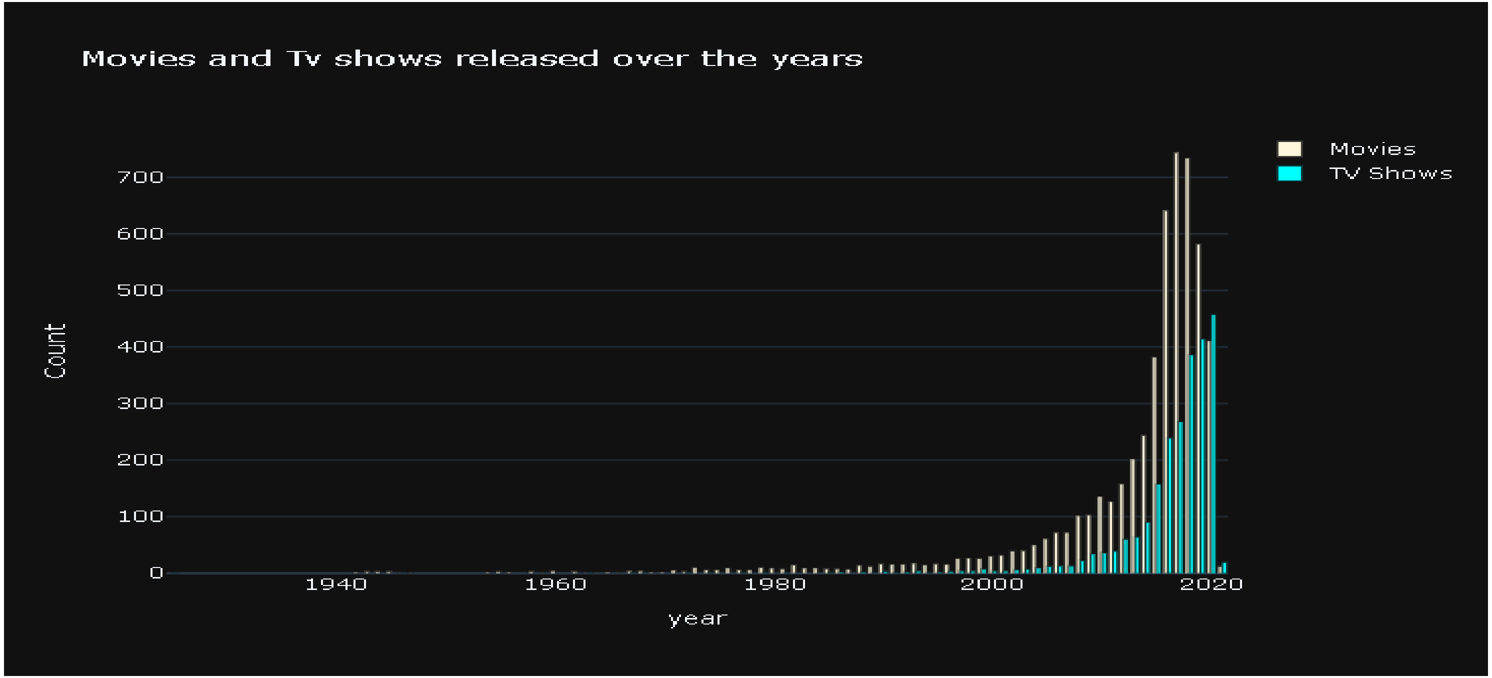
The above pie plot shows that movies has around 50% of adult rated and in tv shows 43% are adult rated.



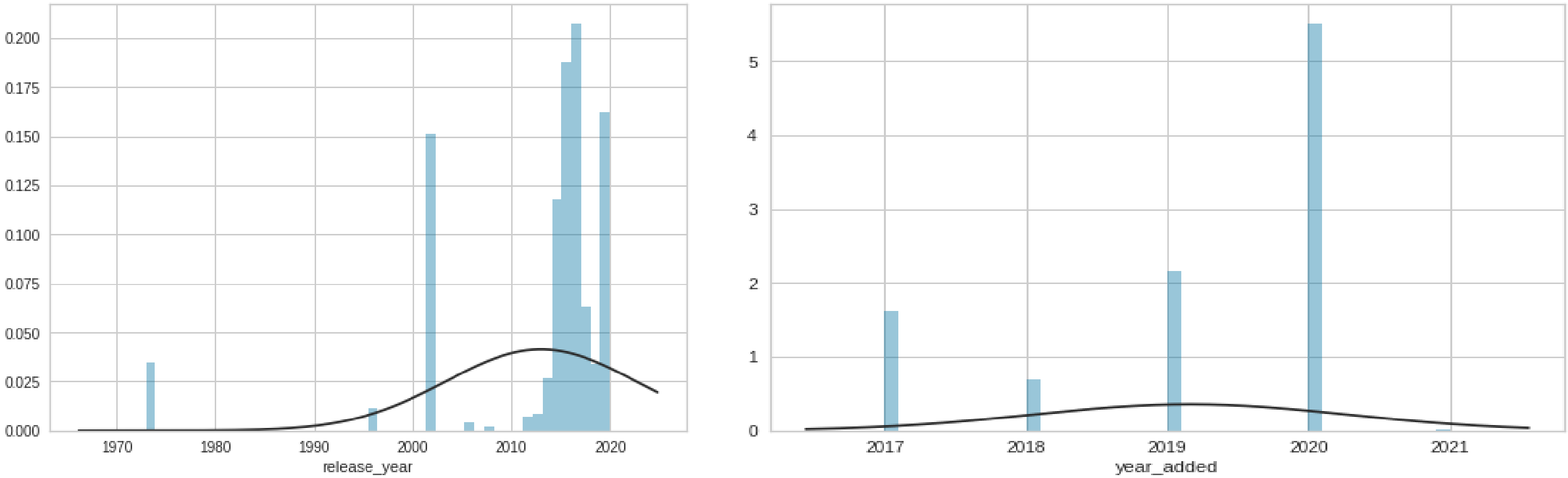
- The above pie plot shows top content producing countries where USA stands first with 51% and followed by India with 14.2%



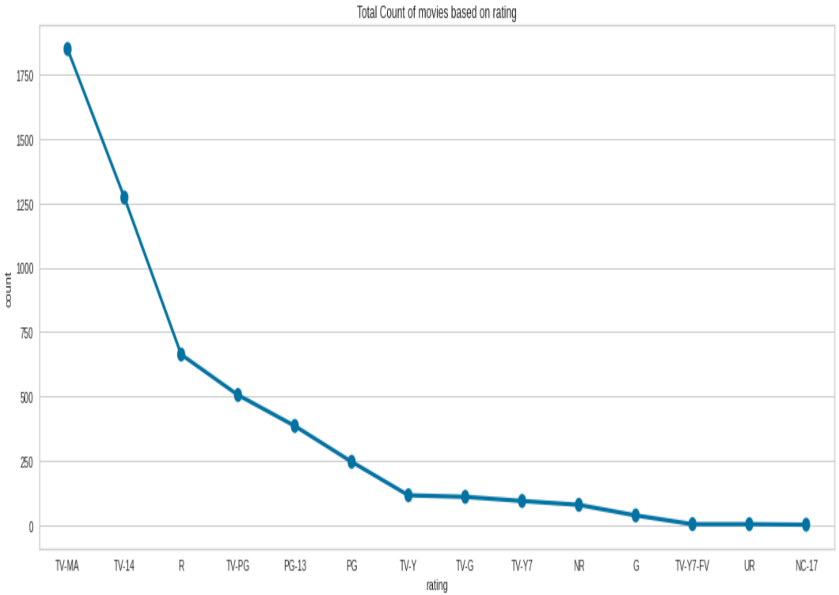
The above histogram shows how many tv shows and movies are produced by each country.



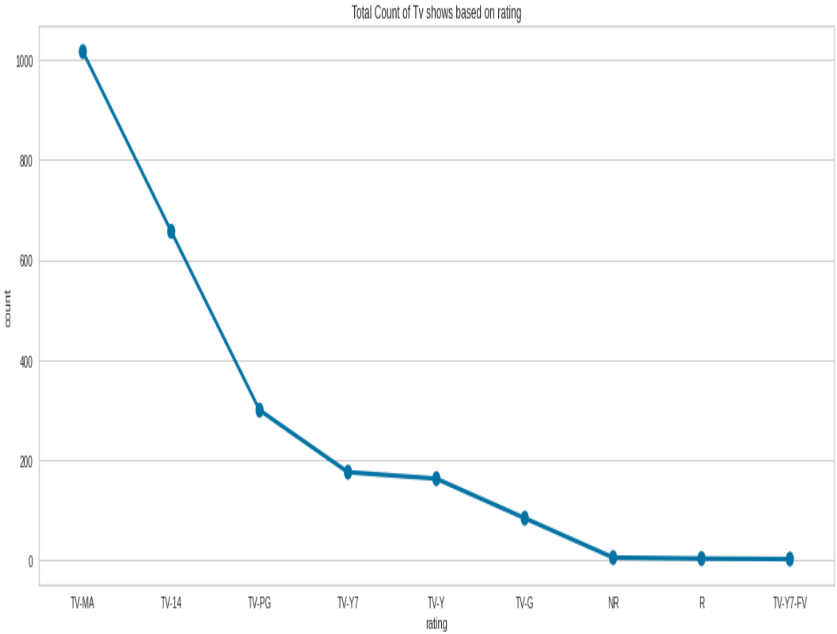
The above bar plot shows number of movies and tv shows are produced each year.



The above plots are the distribution plots of the features release_year and year_added

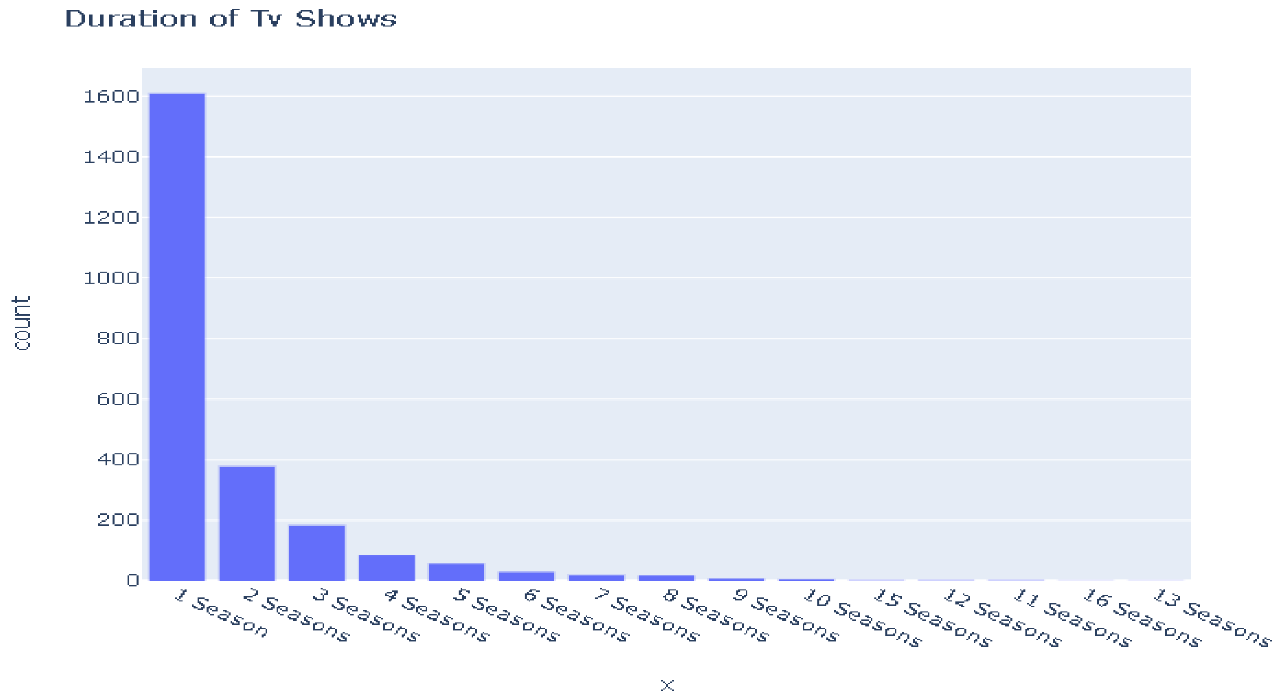


Total Count of movies based on rating

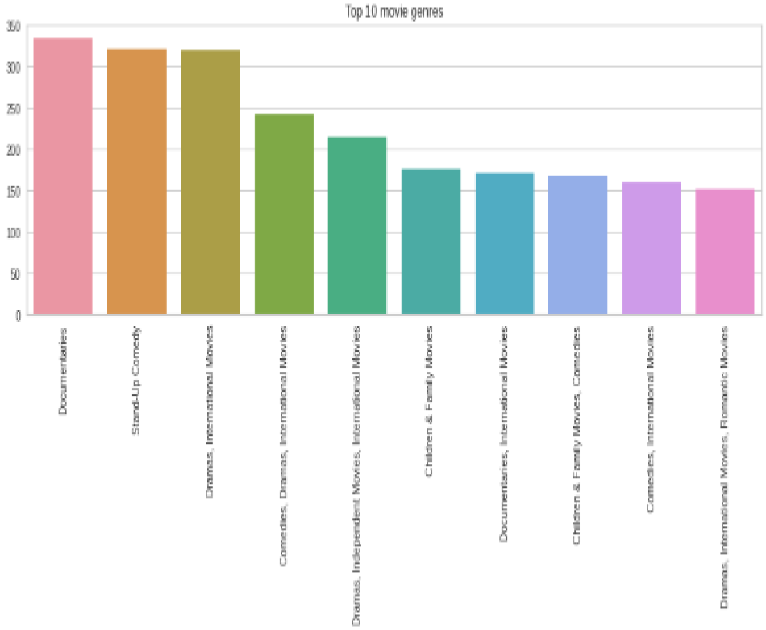


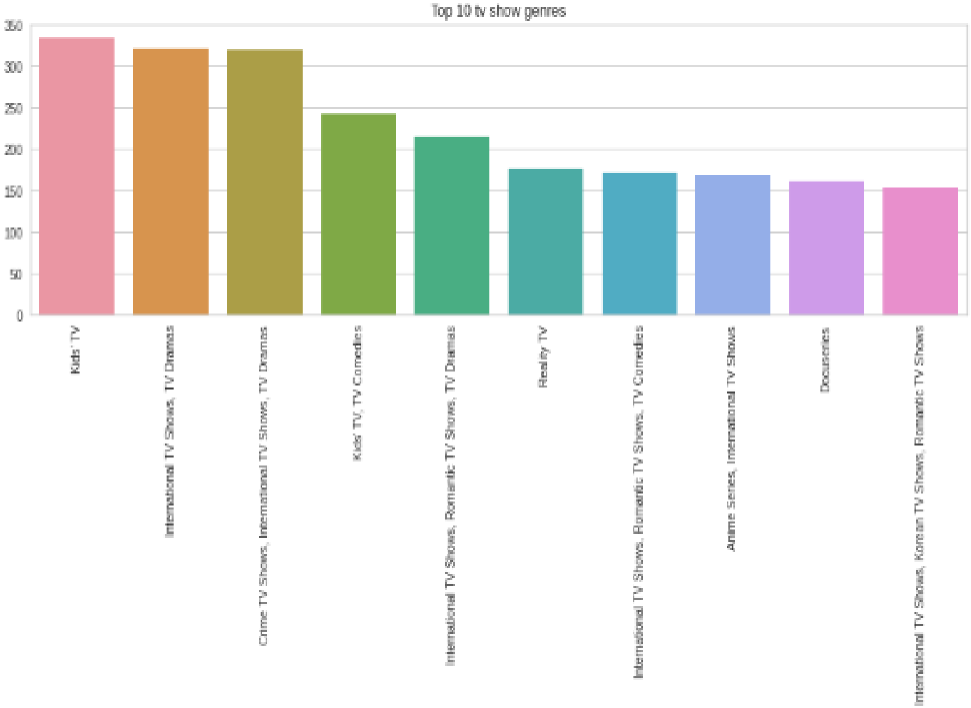
Total Count of Tv shows based on rating

- The above pont plots shows the total count of movies and tv shows for each rating.

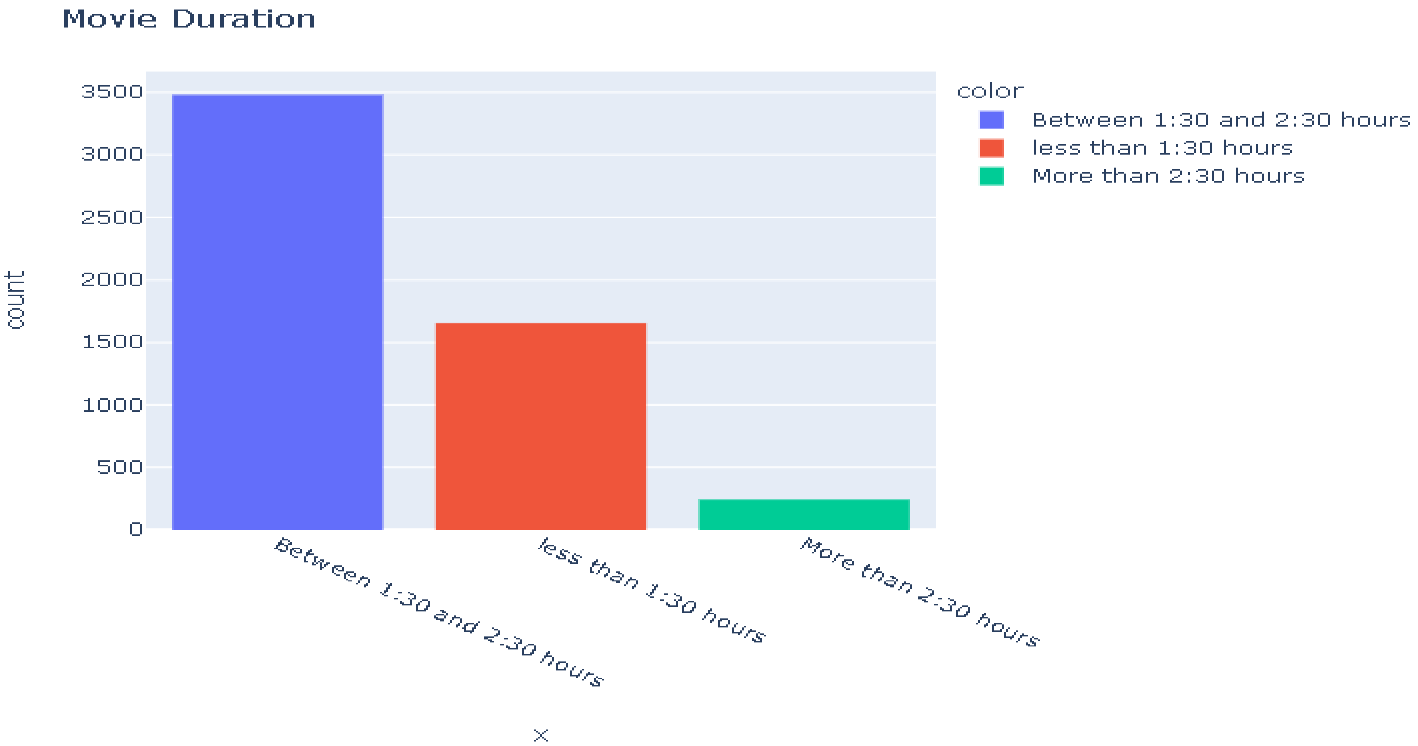


The above bar plot shows the total number of tv shows and their total season count.





The above two bar plots show the top 10 genre of movies and tv shows



The above plot shows the average duration of movies which are divided into three groups between 1:30 to 2:30 hr, less than 1:30hr, more than 2:30hr

Data Preprocessing:

Here we have performed nltk operations on data and done the following:

- Added new feature, in which listed_in, description and country has been added.
- Removed all the stop words
- Performed stemming
- Performed Standard Scaler

Kmeans Clustering:

It is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science. It groups the unlabeled dataset into different clusters. Here k defines the number of pre-defined 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.

The performance of k-means clustering depends upon highly efficient clusters that it forms. But choosing the optimal number of clusters is a big task. There are some different ways to find the optimal cluster size

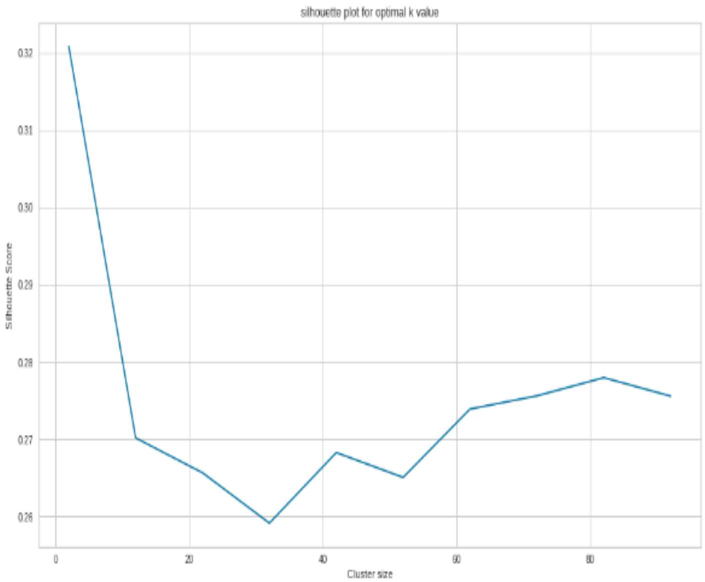
- Elbow method
- Silhouette Score

Silhouette Score:

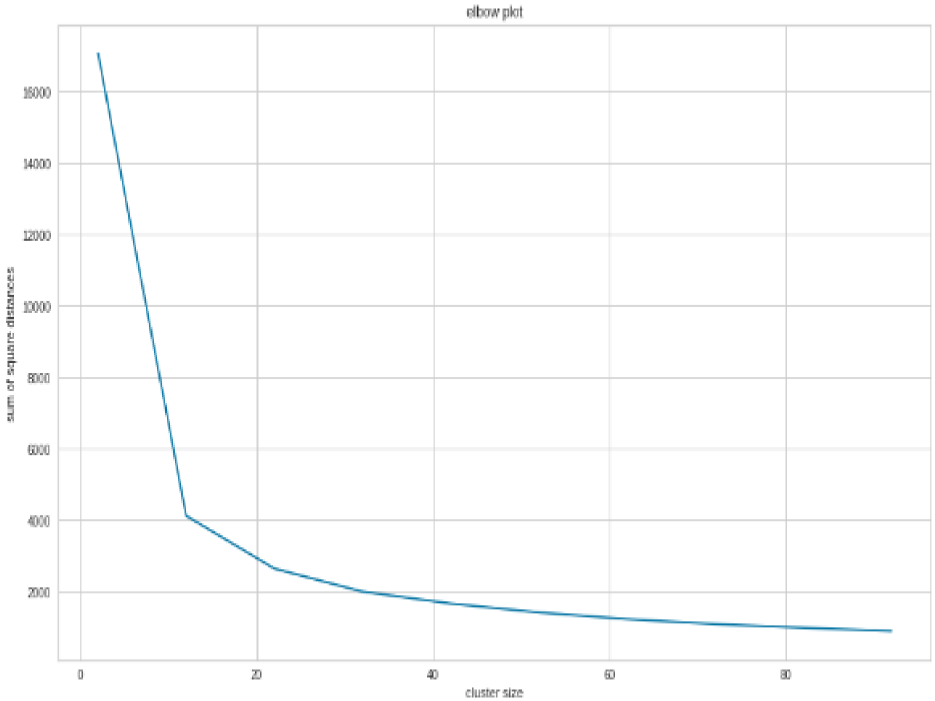
- It is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1.
- 1: means clusters are well apart from each other and clearly distinguished.
- 0: means clusters are indifferent or the distance between the clusters is not significant.
- -1: means clusters are assigned in the wrong way.
- Silhouette Score =(b-a)/max(a,b)
- Where
- a= average intra-cluster distance i.e the average distance between each point within a cluster.
- b= average inter-cluster distance i.e the average distance between all the clusters.

Elbow Method:

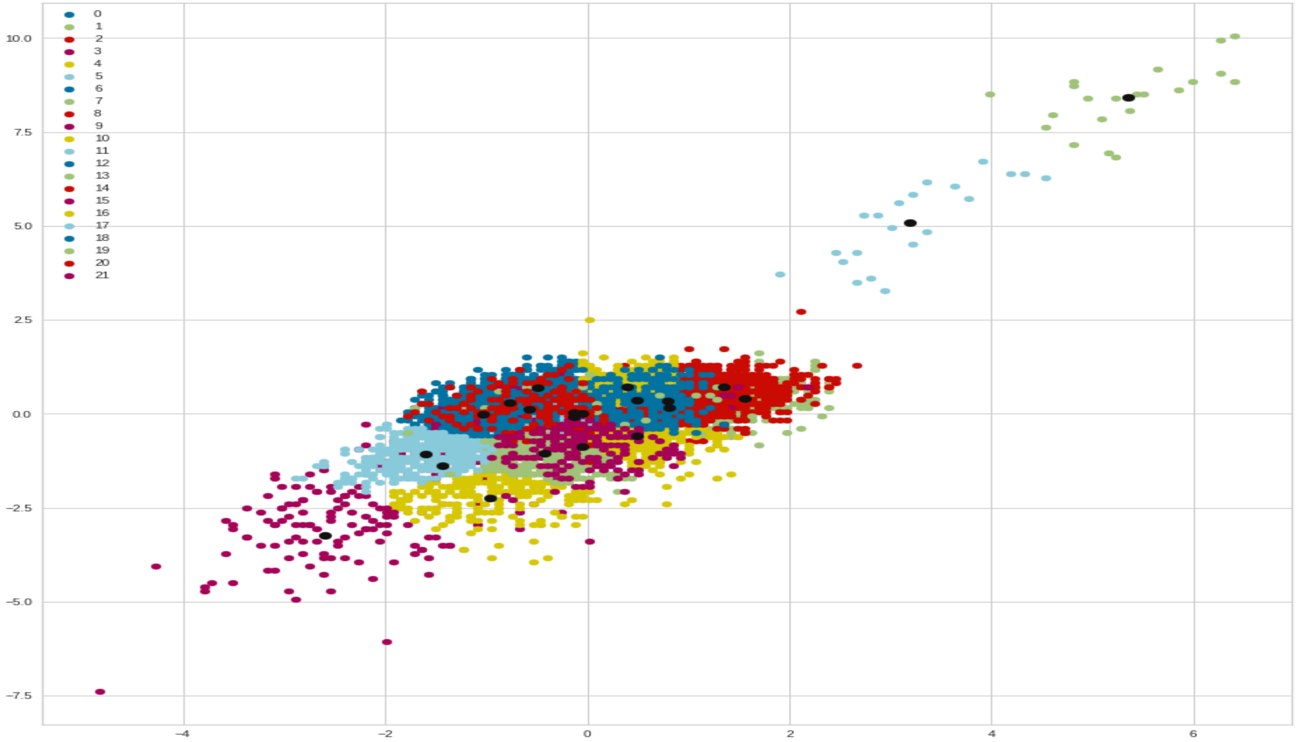
The elbow method is used to determine the optimal number of clusters in k-means clustering. The elbow method plots the value of the cost function produced by different values of k. As you know, if k increases, average distortion will decrease, each cluster will have fewer constituent instances, and the instances will be closer to their respective centroids. However, the improvements in average distortion will decline as k increases. The value of k at which improvement in distortion declines the most is called the elbow, at which we should stop dividing the data into further clusters.



Silhoutte Score plot



Elbow method plot



The above are the clusters formed using Kmeans clustering.

There are 21 clusters, which are ideal for this problem.

Recommender System:

- Algorithms that are used to provide the list of suggestions are called recommendation systems or engines. The recommendation system in its core algorithm uses a fundamental mathematical metric called ‘similarity’, which compares and quantifies the similarity between two items user selected vs rest of items in the catalog. The list of items with high similarity values to the ones that the user selected are recommended as ‘you may also like’.
- There are many similarity metrics used, here we have ‘cosine similarity’.
- Cosine Similarity ranges from -1 to 1. 1 indicates the items are same whereas -1 represents the items are dissimilar.

Conclusion:

- After Implementing very good analysis and most of the visualisations we have a pretty good idea of what this dataset represents. Dataset contains 7787 Rows and 12 Columns, there are around 2800 null values in director column, 718 on cast and 507 in country.

- The main content providers for Netflix are the United States which contributes 50% of the content, followed by India it contributes 14% and United Kingdom comes third of 8.5% and the rest follows.

- This Netflix dataset has been categorised into two items:

- Movies
- TV Shows

- Movies is the majority here which covers around 69% and TV Shows covers the rest which is 31% of the dataset.

- In the Movies section 48% of the movies are Adult rated, 31% rated as Teen, 15.9% as 7&above, and the rest as Kids.

- In the TV Shows section 41% of the shows are Adult rated, 27% rated as Teen, 20% as 7&above and the rest 10% as kids Rated.

- The year 2017 saw the highest content ever on Netflix for movies where 744 movies were released and the yesr 2020 saw the highest for 457 TV Shows.

- Our analysis on movies duration showed that movies ranging between 90mins to 150mins were the highest with 3481 movies followed by movie duration less than 90mins with 1653 movies and finally movies with more than 150mins duration has 243 movies.

- Coming to the main part processing the text data using NLTK library and performing those actions on description feature. After that we have removed all the stop words and applied stemming to it (Stemming is the process of reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of words known as a lemma). After that finding lengths of features and storing new features for clustering.

- Standardized all the data before performing Ml Model in this case we used Kmeans algorithm because after seeing the analysis done we thought this is the algorithm can balance all the features. Plotted Silhouette for the optimal Kvalues and Elbow plot for cluster size and we found out the optimal clusters can be used are 21.

- Performed Recommender systems using cosine similarity because The cosine similarity is beneficial because even if the two similar data objects are far apart by the Euclidean distance because of the size, they could still have a smaller angle between them. Smaller the angle, higher the similarity.

- The recommendations after all the tuning and model selection we are getting pretty good range of recommendations and satisfied with the recommendations.
-
- The given Netflix dataset has been successfully analysed and visualised by various plots and charts and the final result recommendations are very positive.

THANK YOU