



NETFLIX MOVIES & TV SHOWS CLUSTERING



CONTENT

- Introduction
- Abstract
- Problem Statement
- Handling Null Values and feature engineering
- EDA
- Hypothesis Testing
- Finding Number of Clusters
- Algorithm
- Model Performance
- Interactive scatterplot of the cluster
- 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.
- 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.

ABSTRACT

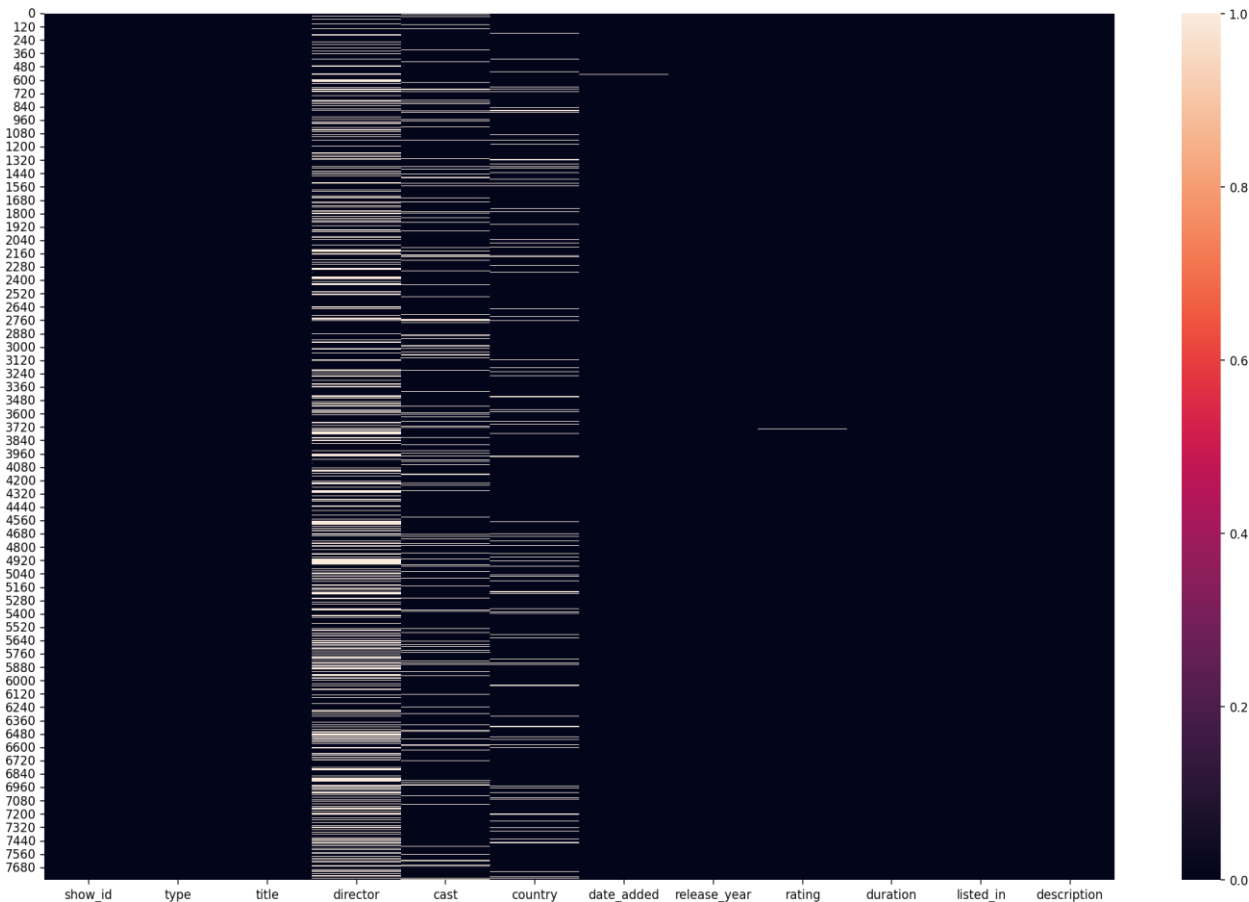
- The idea was to use text-based variables to anticipate clusters of related content.
- The dataset is subjected to exploratory data analysis in order to extract insights from it, but the initial null results are ignored.
- In addition, using EDA's findings, some hypothesis testing was done.
- After 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.
- 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.



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.

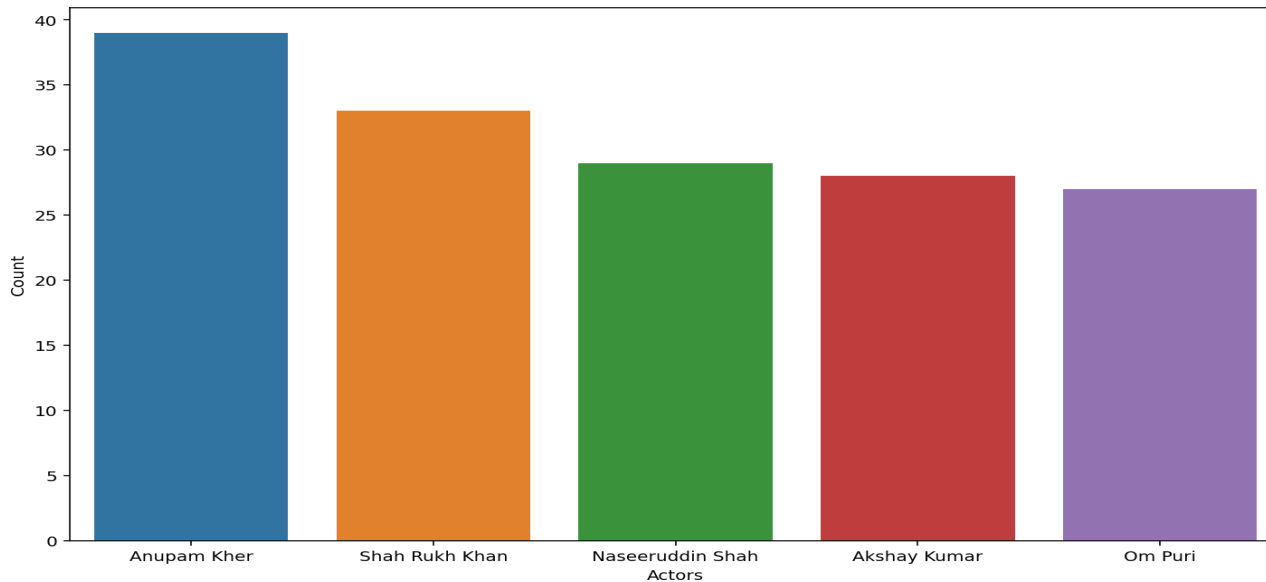
Handling Null Values and feature engineering



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



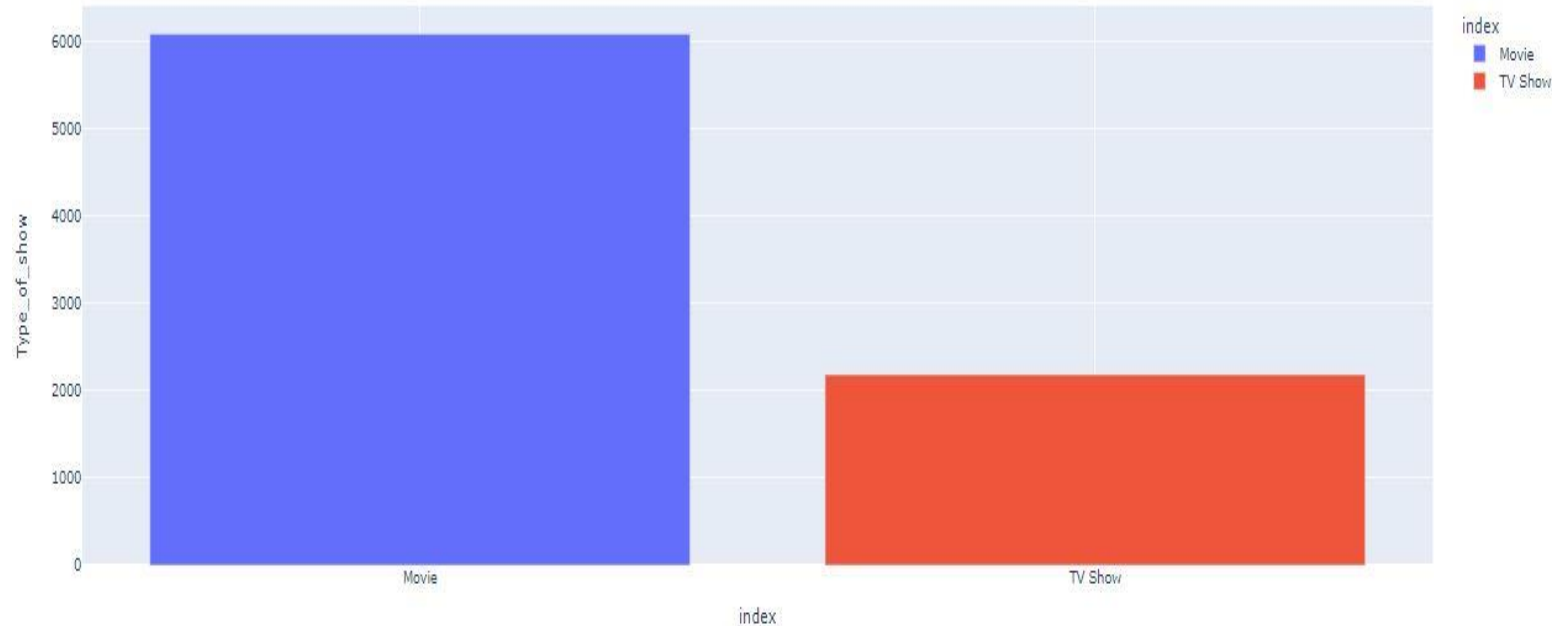
Name of the Actors who acted more times only for Indian Movies?



- As we can see from the plot, Anupam Kher has acted in more Indian films than anyone else.



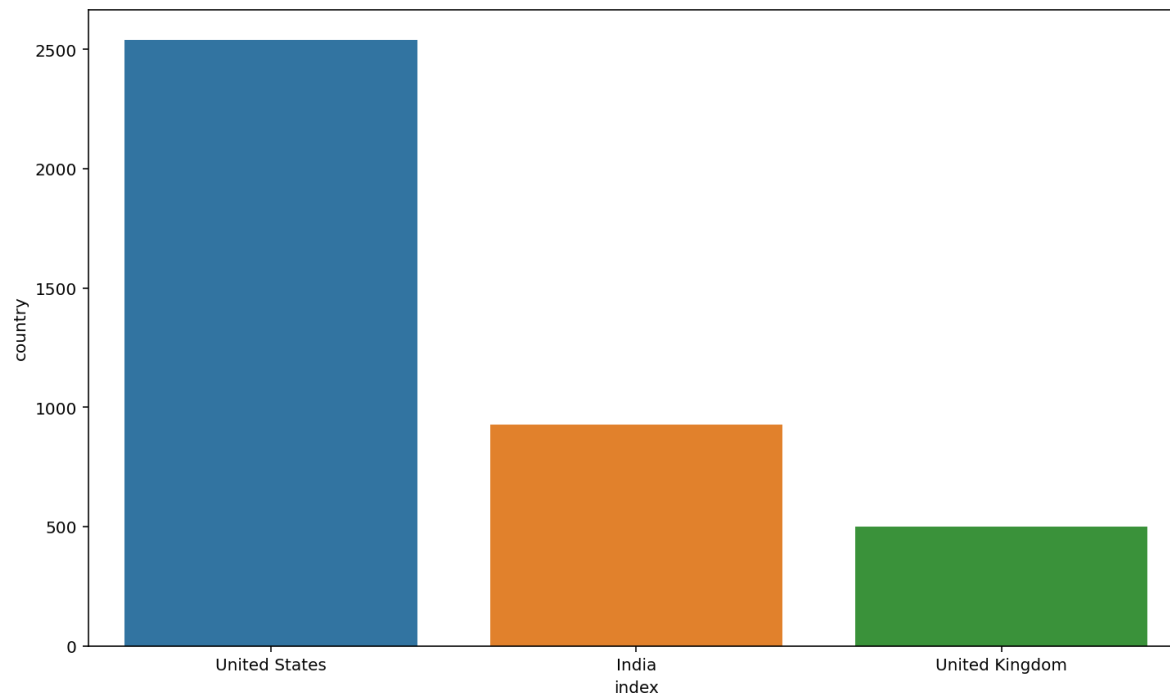
What is more popular on Netflix, movies or TV shows?



- As we can see in the plot, there are more movies available on Netflix compared to TV shows. That means movies are more popular than TV shows.



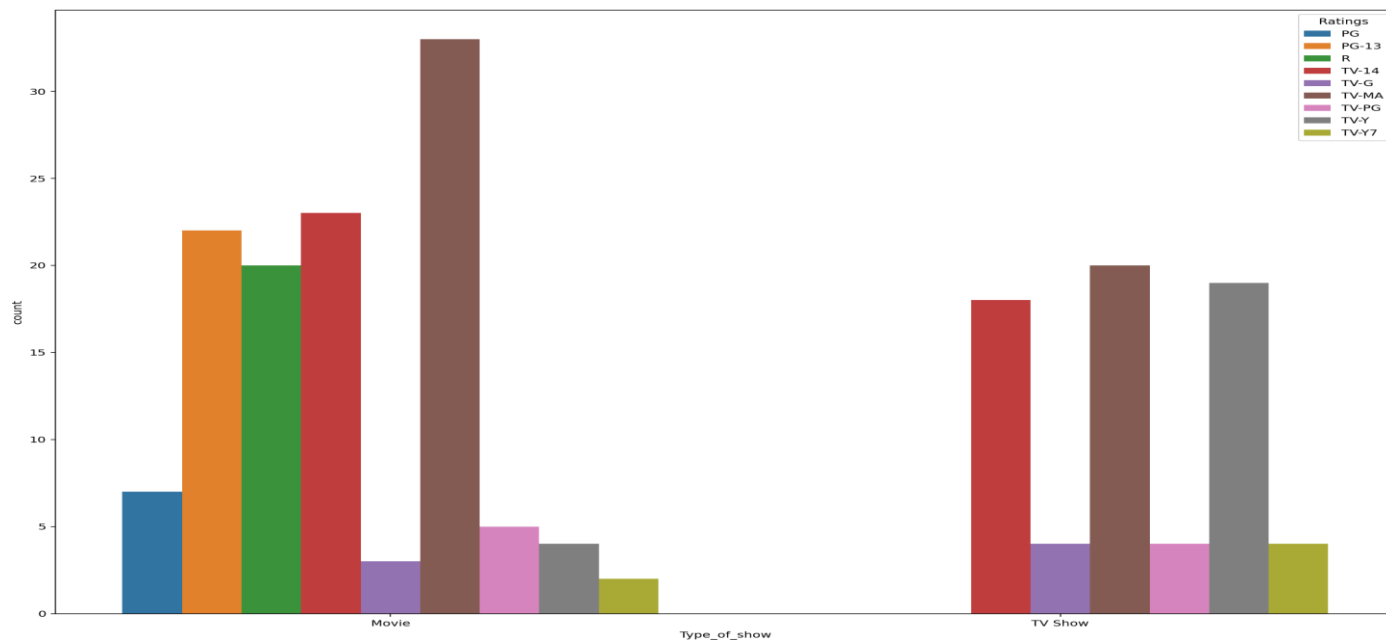
Analysis of the top two countries where Netflix is most popular?



- **As can be seen in the plot above, the United States and India are the two countries where Netflix is most popular.**



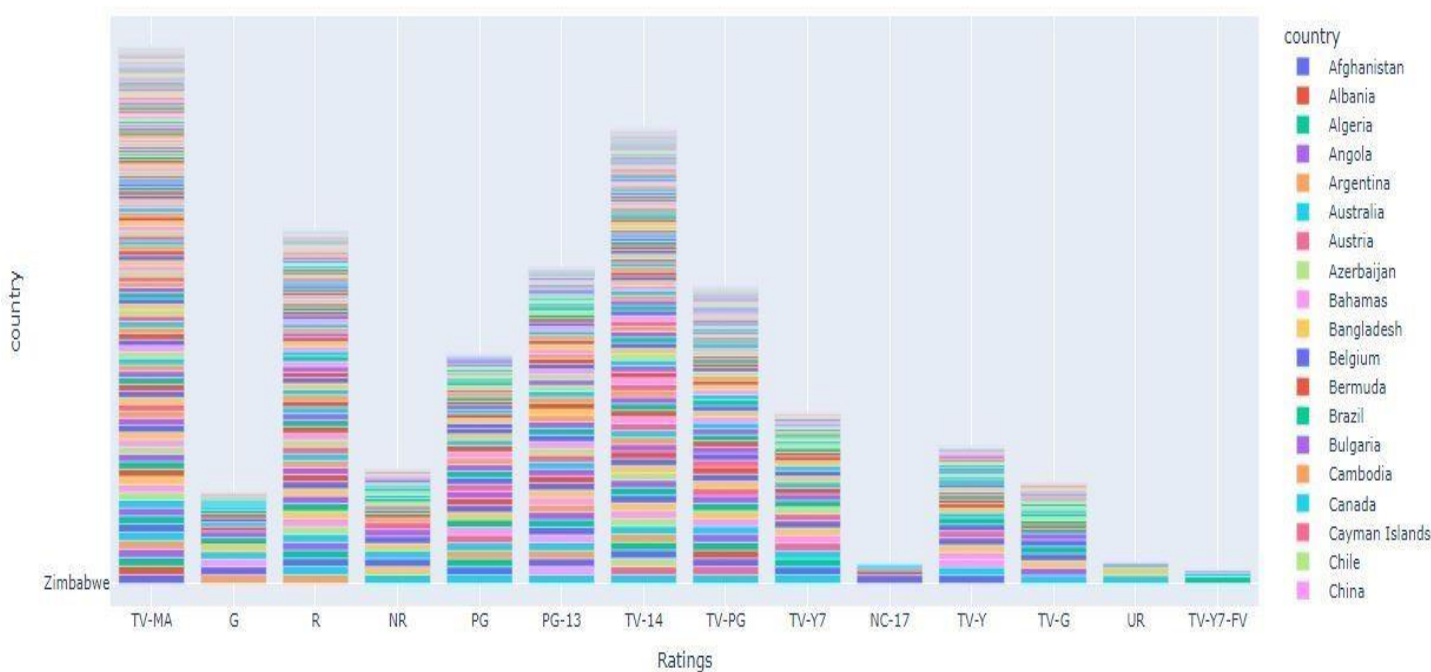
Analysis of type of shows launched and highest rating of movie in December 2020?



- As shown in the plot above, TV-MA ratings in December 2020 are the highest in movies and TV shows, with TV-MA standing for Mature Audience Only. Because this programme is intended for adults, it may not be appropriate for children under the age of 17.



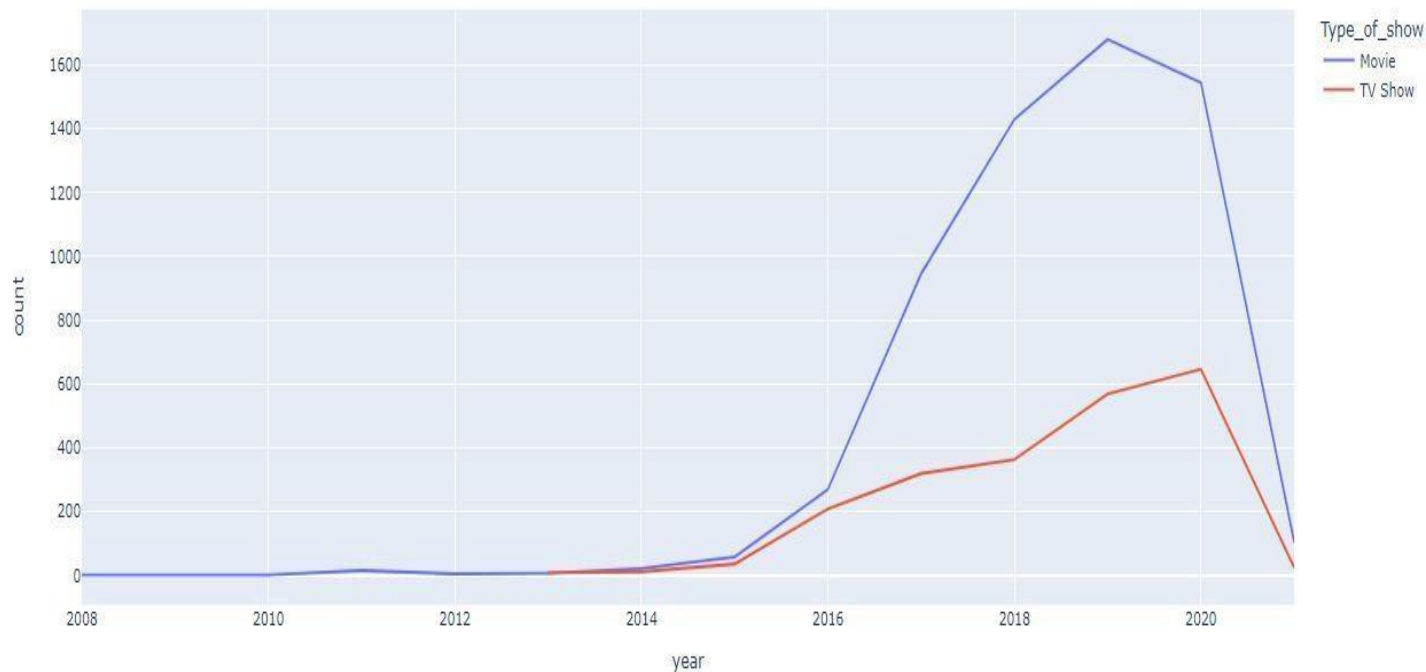
Analysis of type of content available in different countries?



- As we can see from the plot above, there are various types of content available, but in most countries, TV-MA content is available, and the TV-MA rating you see on many Netflix TV series signifies that the programme is only suitable for mature viewers. A TV show with a TV-MA rating features graphic violence or a combination of brutal violence. So that could be the reason for it, because the Netflix audience enjoys this type of content.



Analysis of Netflix, whether it is focusing on Movies or TV Shows?



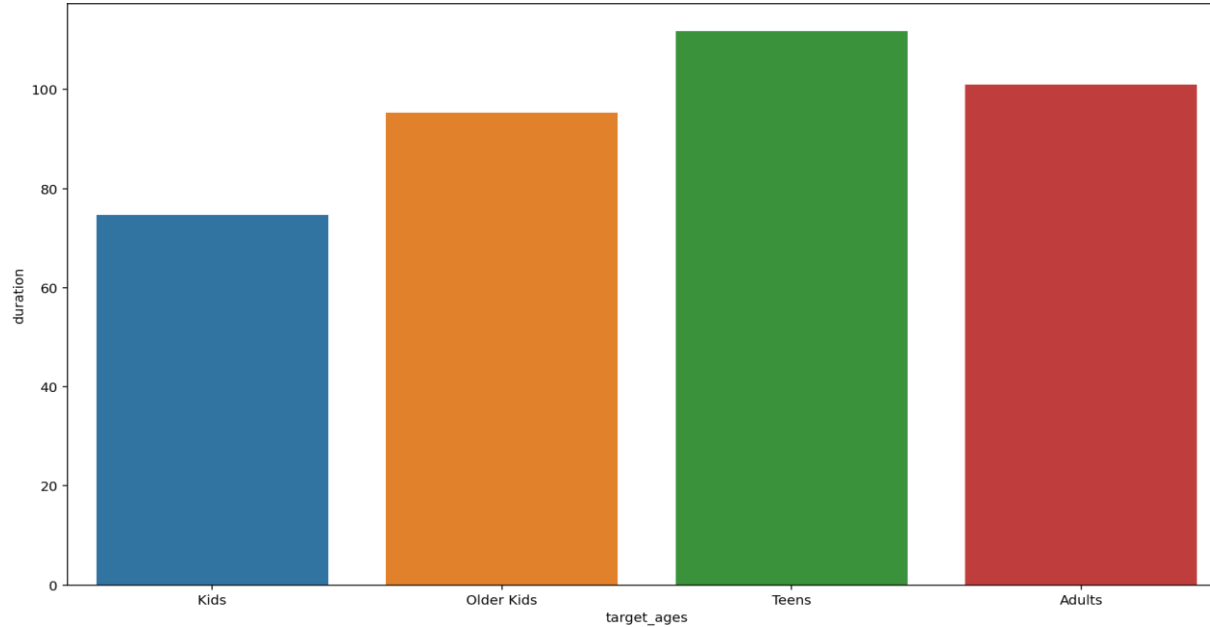
- In the plot above, we can see that Netflix has been increasingly focusing on movies rather than TV shows in recent years, as evidenced by the fact that after 2014, Netflix has relied more on movies than TV shows.

Hypothesis testing



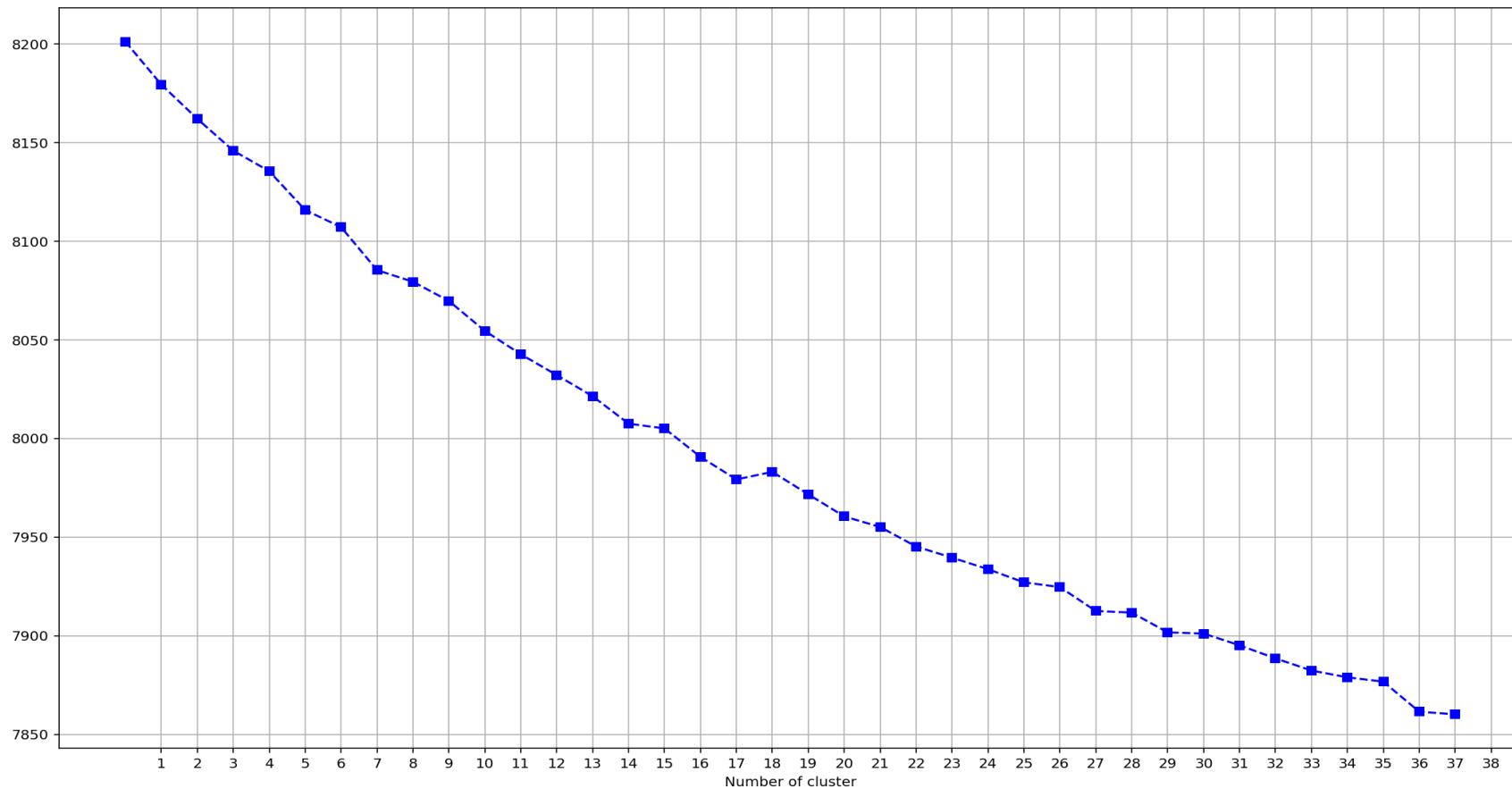
- In the `df_hypothesis` variable, we copied the data frame from the `df_clean_frame` variable.
- Then we did some data augmentation, such as assigning the ratings into grouped categories—
 - `ratings_ages = {`
 - `'TV-PG': 'Older Kids', 'TV-MA': 'Adults',`
 - `'TV-Y7-FV': 'Older Kids', 'TV-Y7': 'Older Kids',`
 - `'TV-14': 'Teens', 'R': 'Adults',`
 - `'TV-Y': 'Kids', 'NR': 'Adults',`
 - `'PG-13': 'Teens', 'TV-G': 'Kids',`
 - `'PG': 'Older Kids', 'G': 'Kids',`
 - `'UR': 'Adults', 'NC-17': 'Adults'}`

Hypothesis testing cont.



- As shown in the graph above, teens have the longest average duration of time for movies, while kids have the lowest.
- Then we made one hypothesis as kids and older kids rated movies are of at least 2 hours long.
- Then we need to reject the null hypothesis because the t-value was not in the range.
- As a result, movies rated for kids and older kids are not at least two hours long.

Finding Number of Clusters



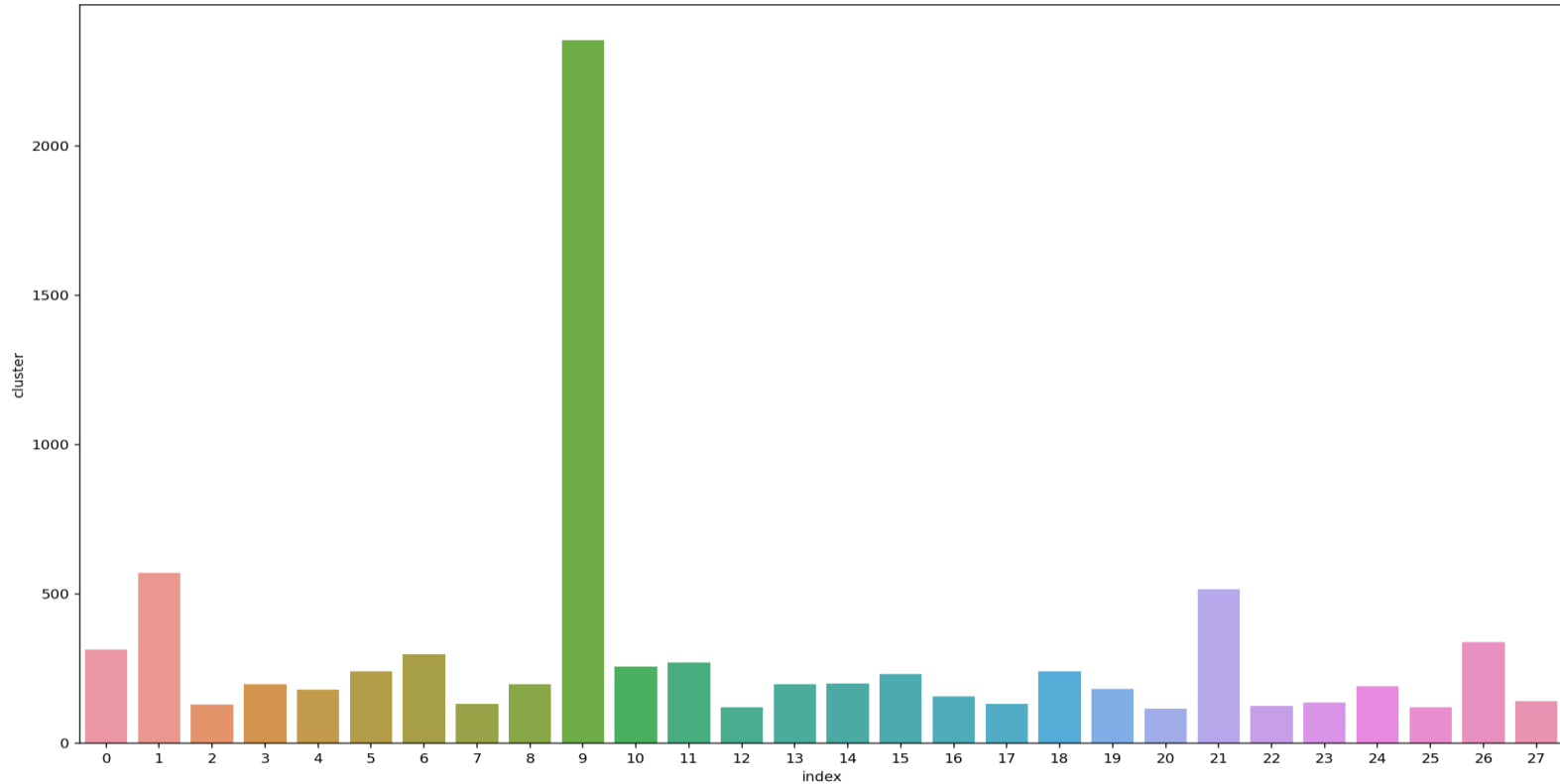


- Clustering is the process of splitting a population or set of data points into many groups so that data points in the same group are more similar than data points in other groups. To put it another way, the goal is to separate groups with similar characteristics and assign them to clusters.
- We used the Elbow method and the Silhouette score to do so, and we chose the 28 clusters based on the results.



Implementation of Kmeans clustering-

- After deciding to use 28 clusters, we used the KMEANS clustering algorithm and then we created another column where each row is assigned to its separate clusters.
- After assigning clusters, most of the points were assigned to cluster number 9, and the rest of the points were not evenly distributed among all the clusters.



- We can observe in the above plot that cluster 9 has the most clusters 2354, while cluster 1 has the second most clusters 569.

Model Performance

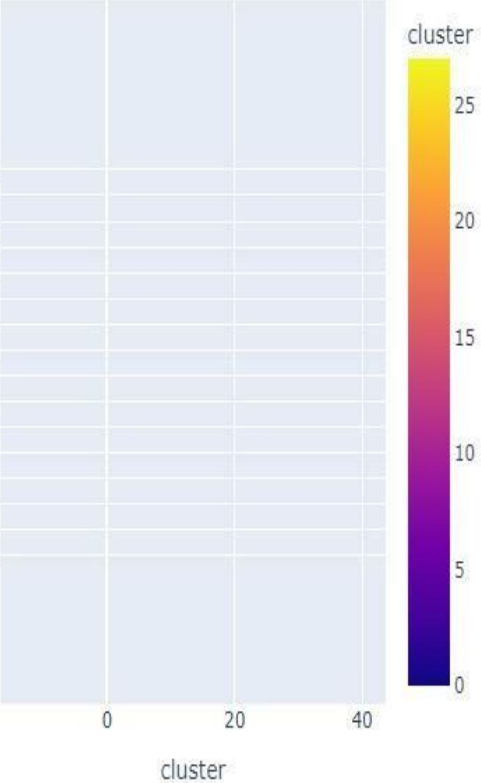
- Evaluation of our K Means model were-
- Silhouette's score was -0.0103
- Calinski Harabasz score -10.5299
- Davies Boulden Index-9.1133
- The Calinski-Harabasz index also known as the Variance Ratio Criterion, is the ratio of the sum of between-clusters dispersion and of inter-cluster dispersion for all clusters, and Compute the Davies-Boulden score The score is defined as the average similarity measure of each cluster with its most similar cluster, where similarity is the ratio of within-cluster distances to between-cluster distances.
- As a result, we can conclude that our cluster is homogeneous within a cluster and heterogeneous with respect to other clusters.

Interactive scatterplot of the cluster

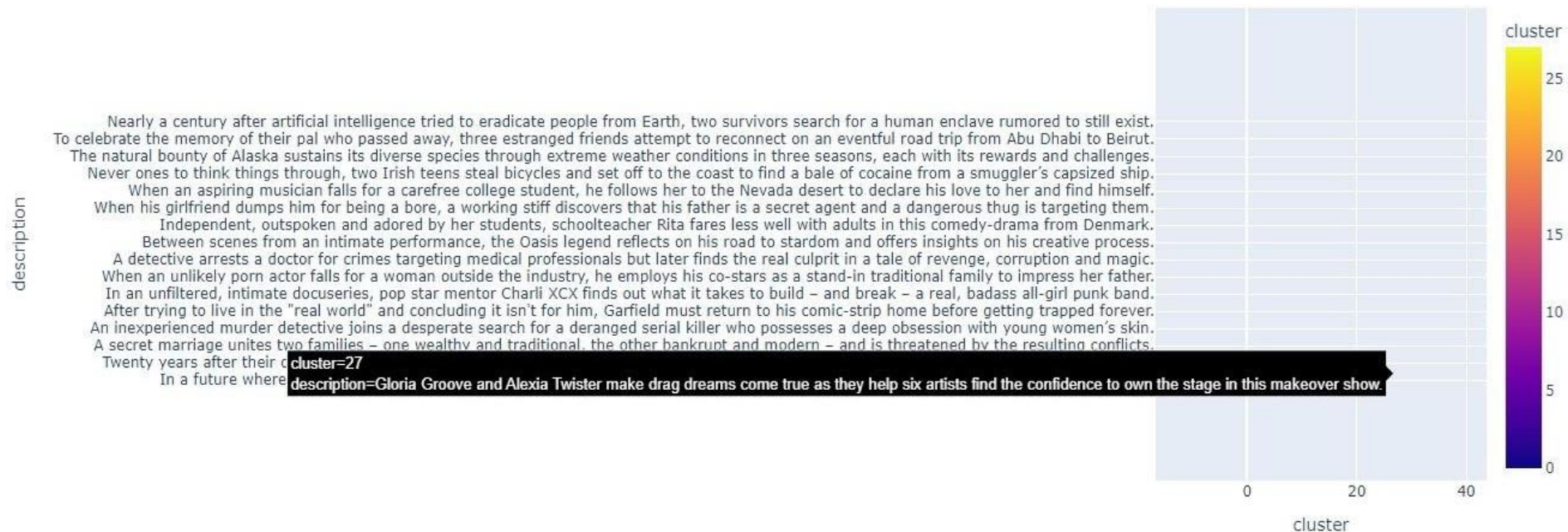


description

Nearly a century after artificial intelligence tried to eradicate people from Earth, two survivors search for a human enclave rumored to still exist.
To celebrate the memory of their pal who passed away, three estranged friends attempt to reconnect on an eventful road trip from Abu Dhabi to Beirut.
The natural bounty of Alaska sustains its diverse species through extreme weather conditions in three seasons, each with its rewards and challenges.
Never ones to think things through, two Irish teens steal bicycles and set off to the coast to find a bale of cocaine from a smuggler's capsized ship.
When an aspiring musician falls for a carefree college student, he follows her to the Nevada desert to declare his love to her and find himself.
When his girlfriend dumps him for being a bore, a working stiff discovers that his father is a secret agent and a dangerous thug is targeting them.
Independent, outspoken and adored by her students, schoolteacher Rita fares less well with adults in this comedy-drama from Denmark.
Between scenes from an intimate performance, the Oasis legend reflects on his road to stardom and offers insights on his creative process.
A detective arrests a doctor for crimes targeting medical professionals but later finds the real culprit in a tale of revenge, corruption and magic.
When an unlikely porn actor falls for a woman outside the industry, he employs his co-stars as a stand-in traditional family to impress her father.
In an unfiltered, intimate docuseries, pop star mentor Charli XCX finds out what it takes to build – and break – a real, badass all-girl punk band.
After trying to live in the "real world" and concluding it isn't for him, Garfield must return to his comic-strip home before getting trapped forever.
An inexperienced murder detective joins a desperate search for a deranged serial killer who possesses a deep obsession with young women's skin.
A secret marriage unites two families – one wealthy and traditional, the other bankrupt and modern – and is threatened by the resulting conflicts.
Twenty years after their debut, join the beloved members of Arashi on a new journey as they showcase their lives, talents and gifts to the world.
In a future where the elite inhabit an island paradise far from the crowded slums, you get one chance to join the 3% saved from squalor.



Interactive scatterplot of the cluster cont.



- In the above interactive scatterplot of the cluster, the number of clusters is on the x-axis, and the description feature is on the y-axis, and we may interact with clusters with similar content.

Conclusion



- We've done null value treatment, feature engineering, and EDA since loading the dataset, and then we've completed some tasks that were assigned to us.
- In this context, we've noticed that Netflix is increasingly focusing on movies rather than TV shows, especially after 2014.
- We've also found that different types of content are available in different countries, but TV-MA is the content that is available in the majority of countries. This could be because it shows that it is just for adult audiences, and the Netflix audience enjoys content like this.
- We've also defined different clusters based on their content; we've defined 28 clusters and implemented the KMEANS clustering algorithm. And then we determined that cluster number nine has the most clusters; we've also plotted a scatter plot in which we may interact with similar content in connection to that cluster.

Thank you

