

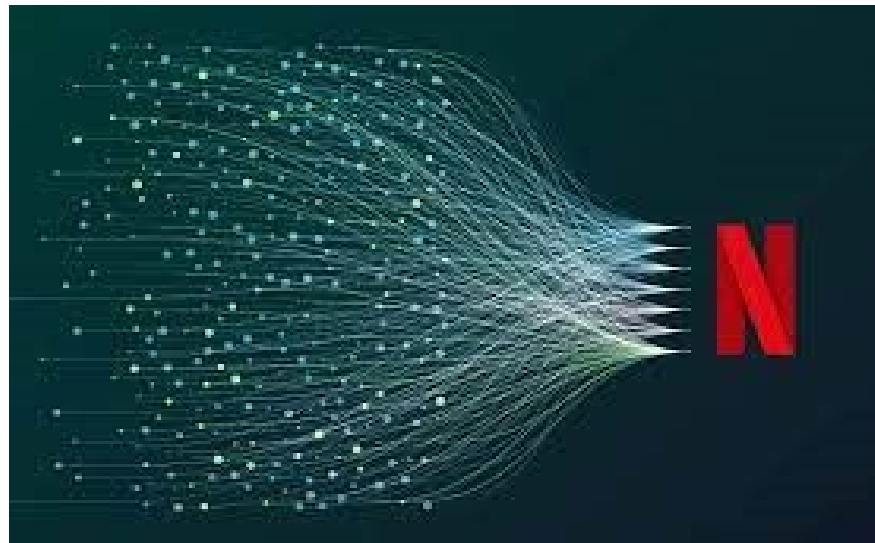
Capstone Project - 4

Individual : Netflix Movies And TV Shows Clustering

Member
Shivam Dhoundiyal

Let's Catch The Defaulters

1. Defining problem statement
2. EDA and feature engineering
3. Dimensionality reduction
4. Clustering
5. Evaluating Clusters
6. Interpretation
7. Visualization



The Dilemma

The objective of this project is to create a predictive model for Netflix movies and TV shows that can estimate their user ratings based on various factors such as title, genre, cast, director, release year, runtime, and production budget. The project is broken down into several parts starting with data collection, where we gather relevant data on Netflix movies and TV shows. The collected data will then be cleaned and transformed into a format that is suitable for analysis.

In the next step, feature selection, we will determine which factors have the most significant impact on the ratings of Netflix movies and TV shows. We will analyze the data to identify which features are correlated with higher or lower ratings and use them as inputs for the predictive model. Additionally, we will use domain knowledge and feature engineering techniques to create new features that can improve the predictive power of the model.

- After selecting the features, we will prepare the dataset for modeling by splitting it into training, validation, and test sets and perform necessary data preprocessing or normalization steps. Then, we will train machine learning algorithms, such as regression or classification models, on the training set to develop a predictive model that can accurately predict Netflix movie or TV show ratings.

- We will evaluate the performance of the model on the validation set and select the best-performing model based on metrics like mean squared error or accuracy. Finally, we will test the final model on the test set to ensure its accuracy and reliability. With this predictive model, Netflix can use the movie and TV show specifications to estimate the expected user ratings and make strategic decisions on content acquisition and production. Additionally, users can use the model to discover new titles that match their preferences or to anticipate the ratings of upcoming releases.

Data Pipeline

The first part of the Netflix movies and TV shows clustering project involved removing unnecessary features that contained null values. In the second part, we encoded categorical features and changed datetime columns to make the data easier to work with and convert categorical data into numerical data.

The third part of the project involved exploratory data analysis (EDA) to identify trends and patterns in the selected features that could be used for clustering.

In the final part of the project, we created a clustering model using algorithms such as K-means or Hierarchical Clustering.

The objective of the clustering model was to group similar movies and TV shows together and separate them from dissimilar ones. Clustering can provide valuable insights for both users and the company and enhance the streaming experience for everyone. The clustering model can help Netflix in content acquisition and production decisions.

Overall, clustering Netflix movies and TV shows can help better understand their characteristics and recommend similar content to users.

Data Summary

NETFLIX

Data Summary

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

Data Summary

date_added : Date it was added on Netflix

release_year : Actual Releaseyear of the movie / show

rating : TV Rating of the movie / show

duration : Total Duration - in minutes or number of seasons

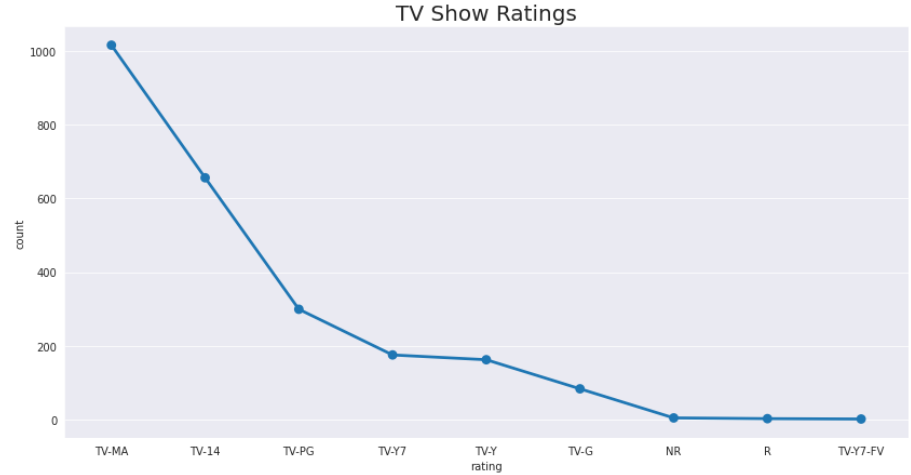
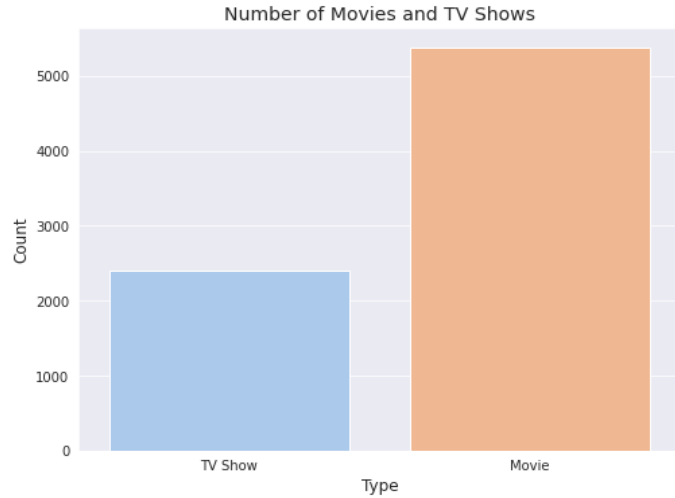
listed_in : Genre

description: The Summary description

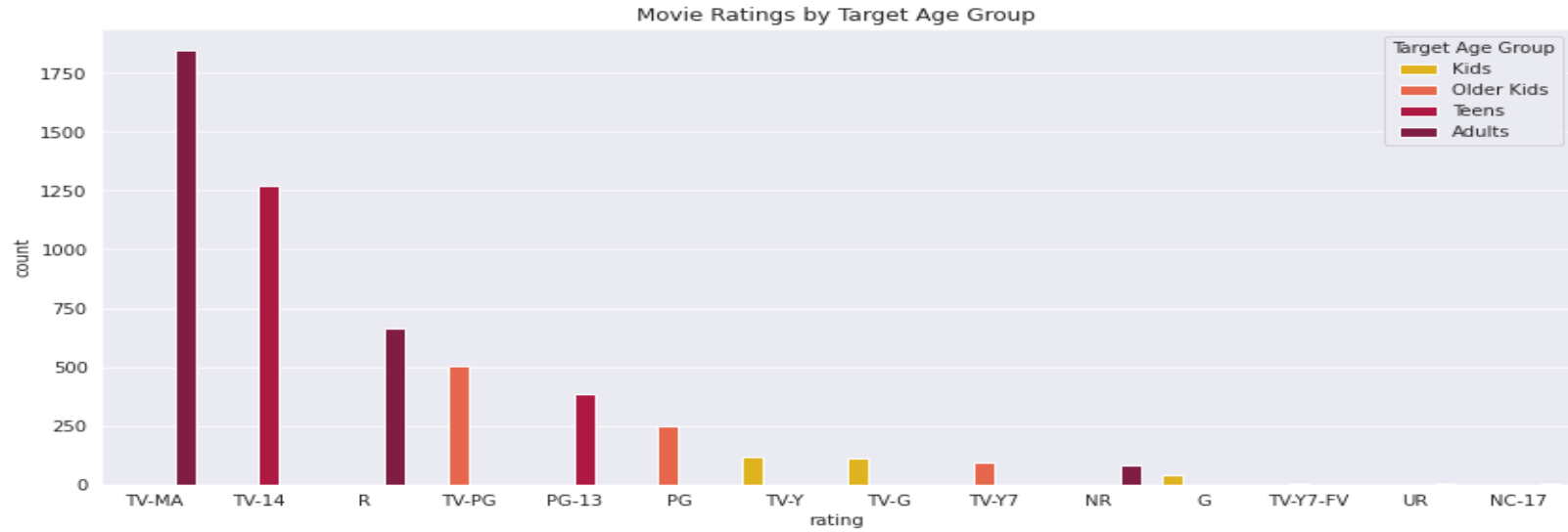
description: The Summary description

.

EDA

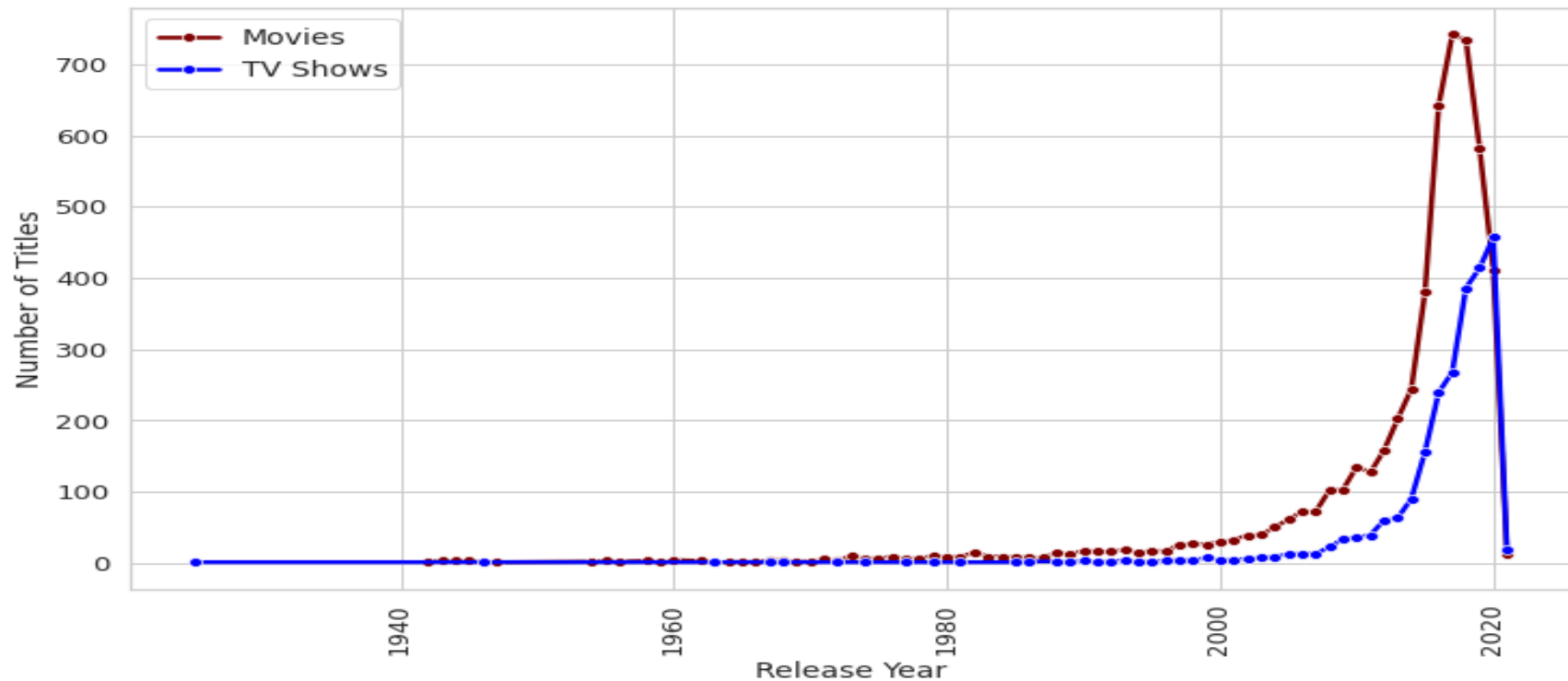


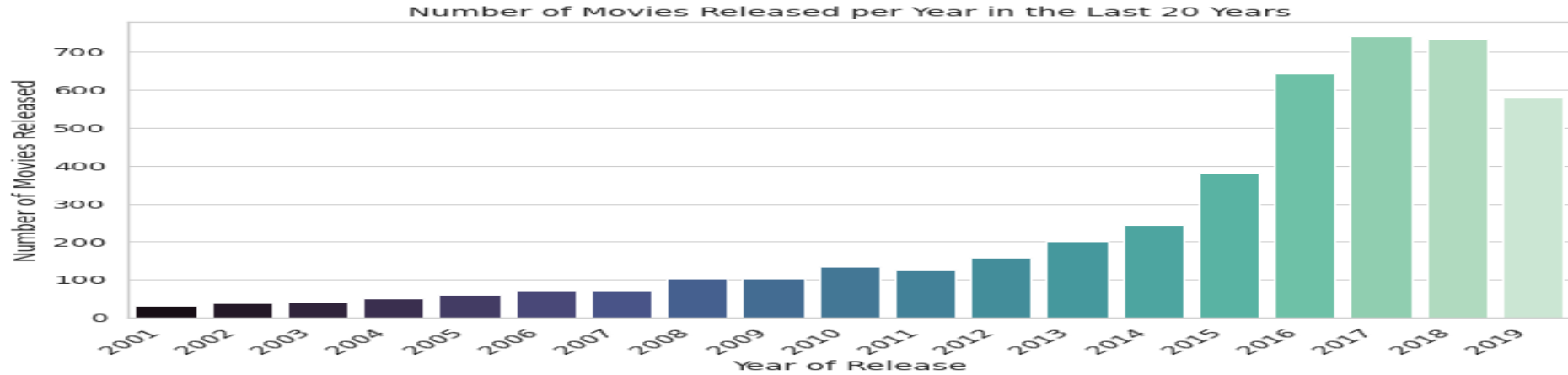
1. The number of movies on Netflix is greater than the number of TV shows, with 5372 movies and 2398 TV shows currently available on the platform.



According to the dataset, TV-MA is the most common rating for TV shows, with the highest number of occurrences in the 'rating' column. This indicates that a significant portion of the TV shows available on Netflix are intended for adult audiences. According to the dataset, TV-MA is the most common rating for both movies and TV shows. This indicates that a significant portion of the content available on Netflix is intended for adult audiences. Specifically, TV-MA has the highest number of occurrences in the 'rating' column for TV shows, while for movies it is also the most common rating. This suggests that Netflix's content caters to a primarily adult demographic, with a focus on mature and potentially controversial themes.

Production Growth Yearly





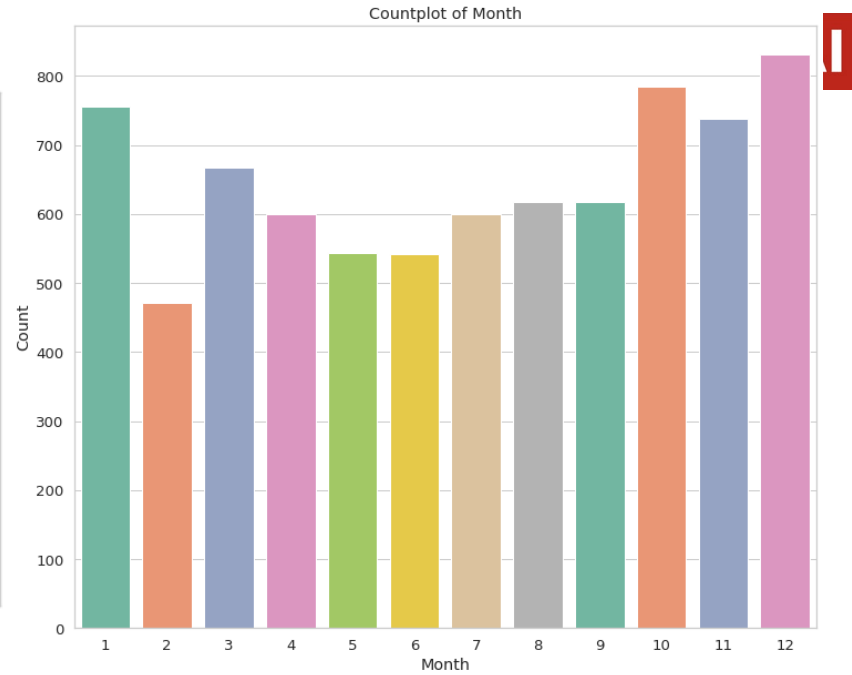
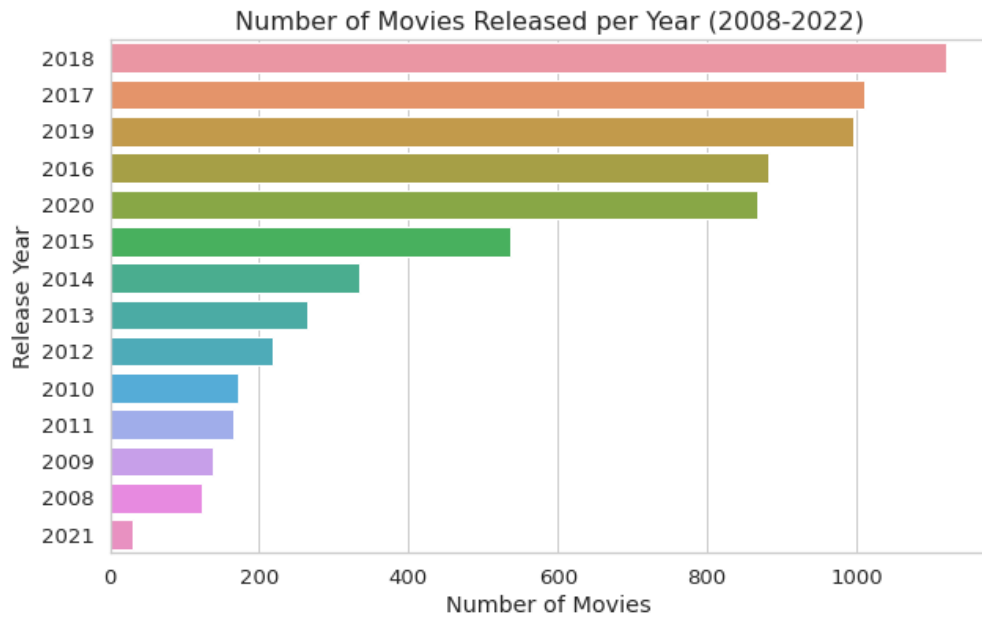
The years 2017 and 2018 had the highest number of movie releases, while 2020 had the highest number of TV show releases.

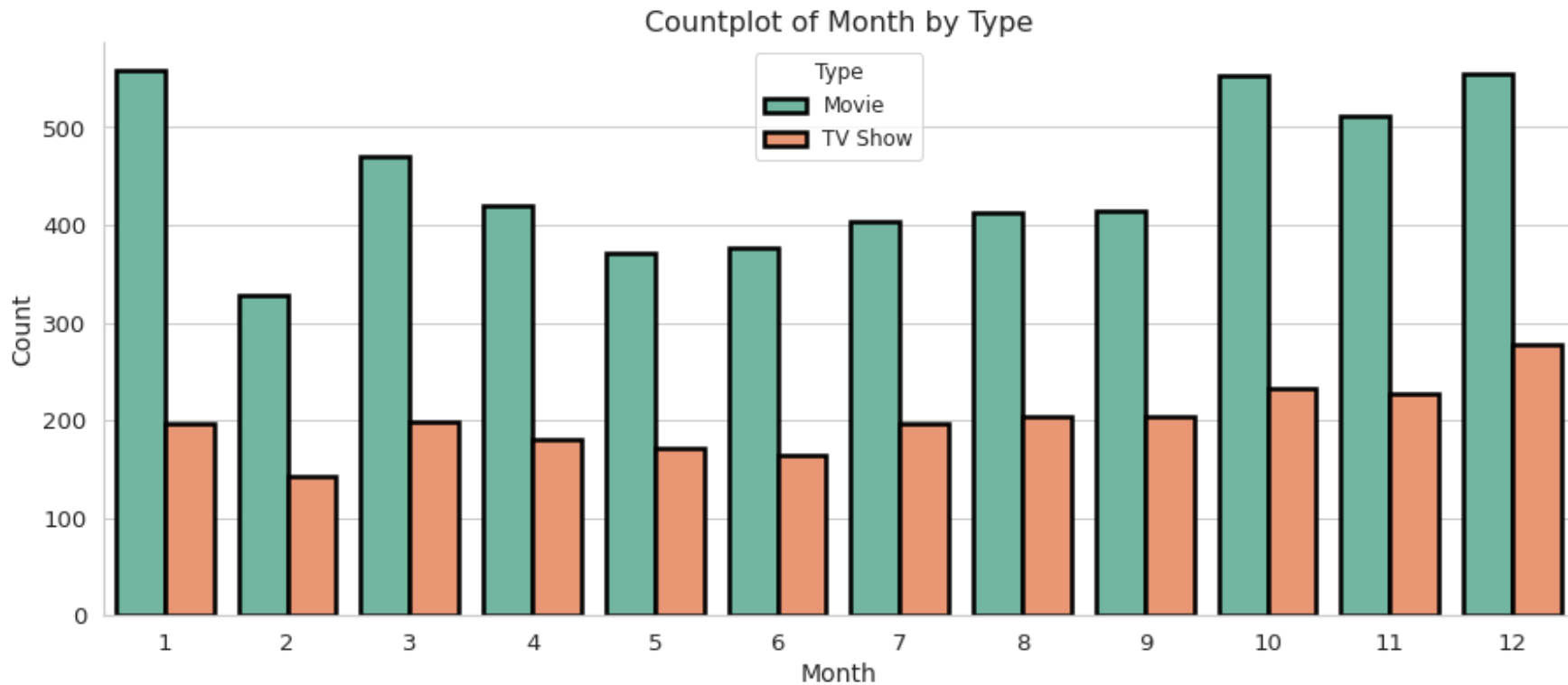
The growth rate of movie releases on Netflix is significantly faster than that of TV shows.

Since 2015, there has been a substantial increase in the number of movies and TV show episodes available on Netflix.

However, there has been a notable drop in the number of movies and TV show episodes produced after 2020.

It appears that Netflix has given more attention to increasing its movie content rather than TV shows, as the growth rate of movies has been much more significant than that of TV shows.

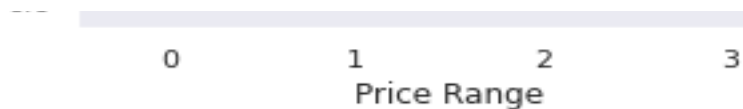
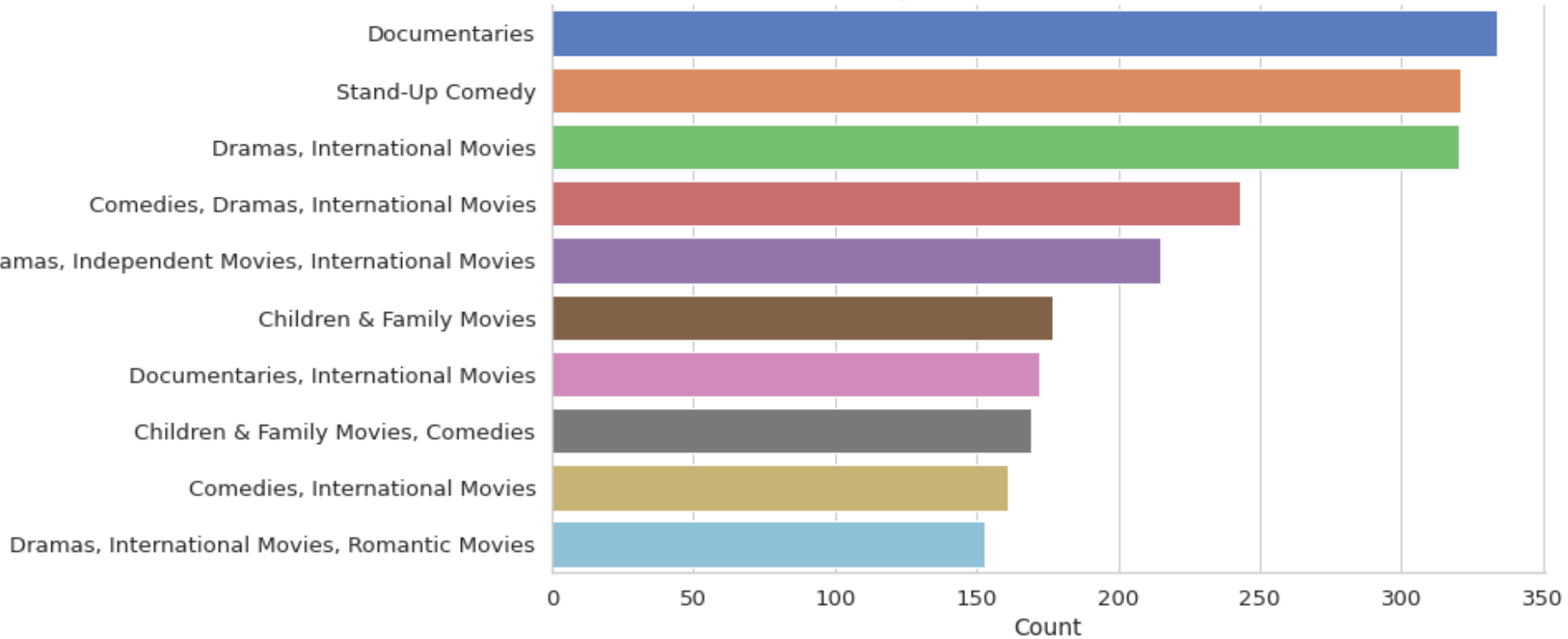


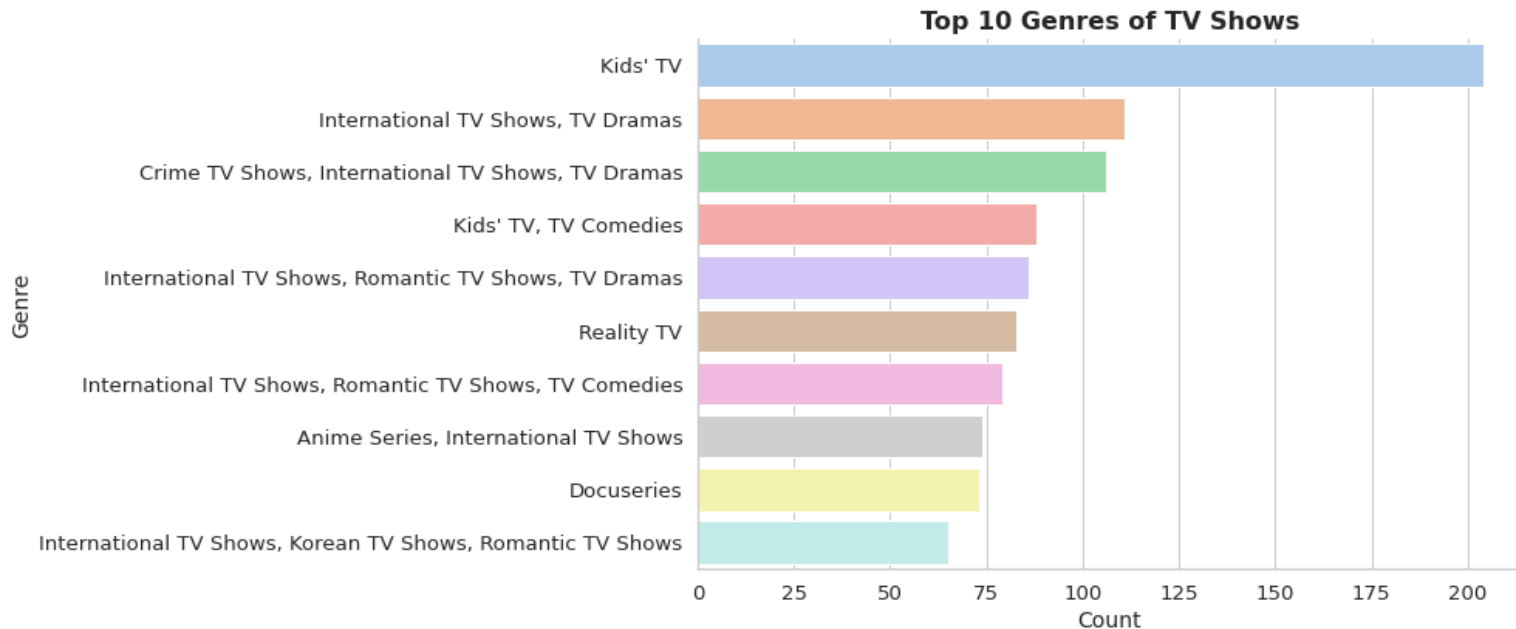


According to the countplot, it appears that Netflix adds the highest number of movies and TV shows during the period between October and January. This period seems to be the busiest time of year for Netflix in terms of adding new content to its platform.

Top 10 Genres of Movies

Genre



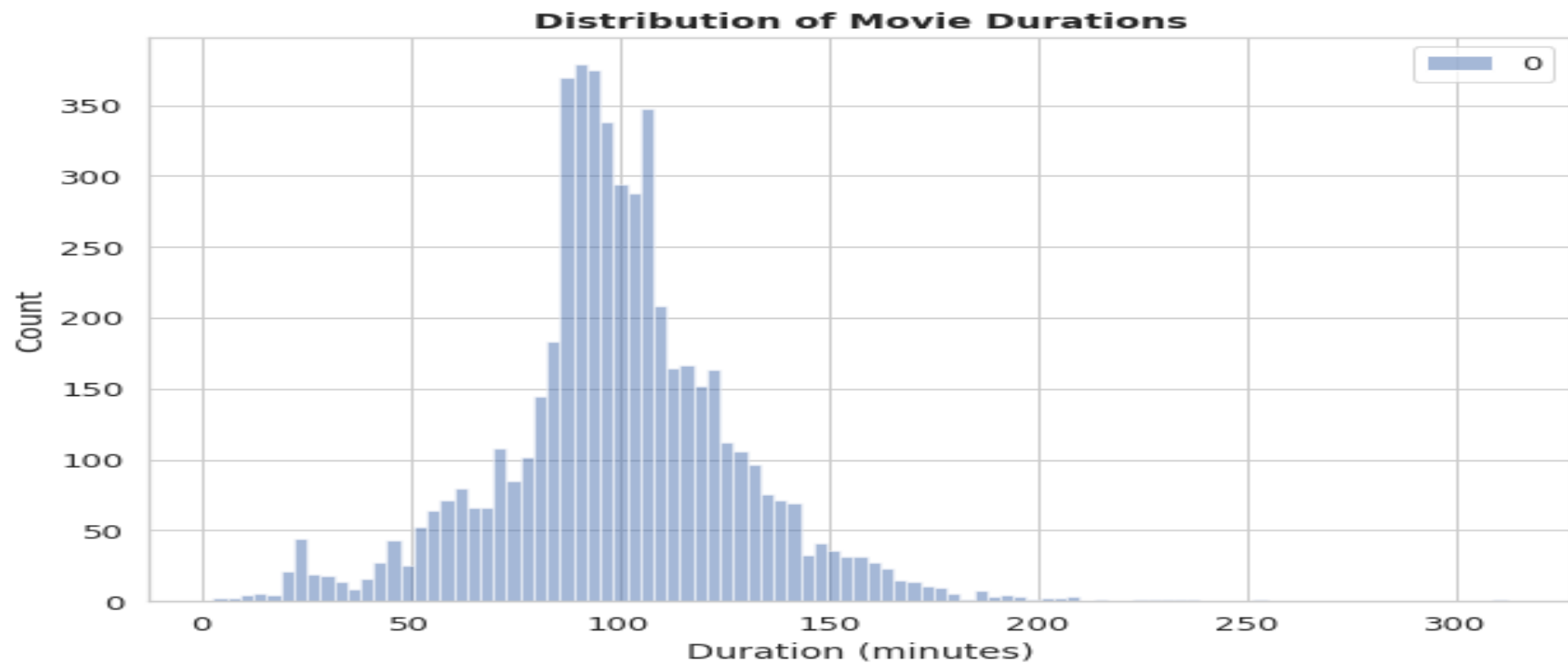


Netflix offers a diverse range of TV show genres, each with its own unique flavor and appeal. However, one genre that stands out as a perennial favorite among viewers of all ages is kids TV.

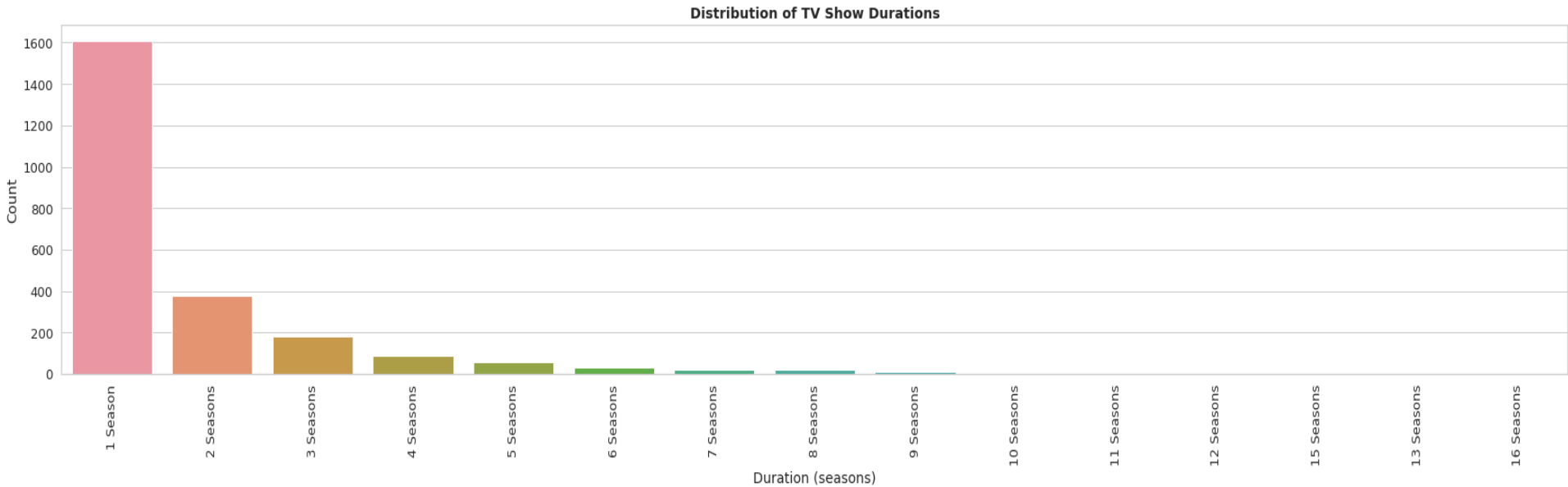
With an impressive selection of animated and live-action shows, Netflix's kids TV category is the perfect destination for families looking for high-quality, entertaining content that is both fun and educational. From beloved classics like SpongeBob SquarePants and Power Rangers to exciting new series like Carmen Sandiego and The Dragon Prince, Netflix's kids TV library has something for every young viewer.

Moreover, Netflix's kids TV category is designed with parents in mind, offering a safe and secure viewing environment that allows them to have peace of mind while their kids enjoy their favorite shows. The parental controls feature allows parents to set age-appropriate content filters, monitor viewing history, and restrict access to certain shows or movies.

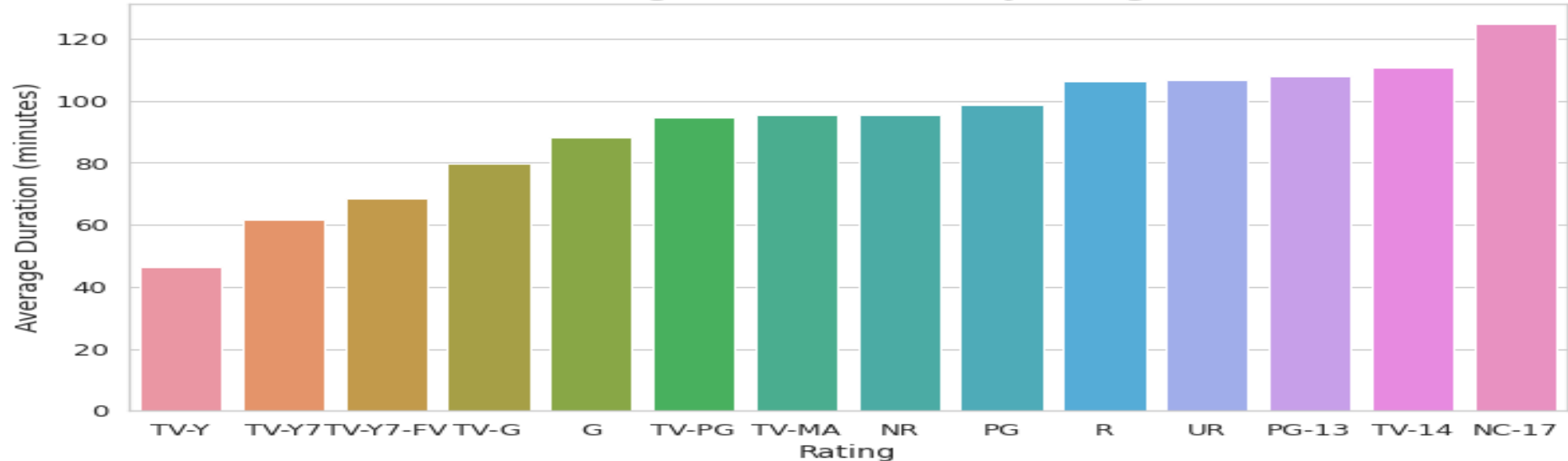
So, whether you're looking for a way to keep your little ones entertained on a rainy day, or just want to bond with your family over a great TV show, Netflix's kids TV category is the perfect place to start. With its vast selection of entertaining and educational content, it's no wonder that kids TV remains one of the top genres on the platform.



EDA (continued)



Average Movie Duration by Rating

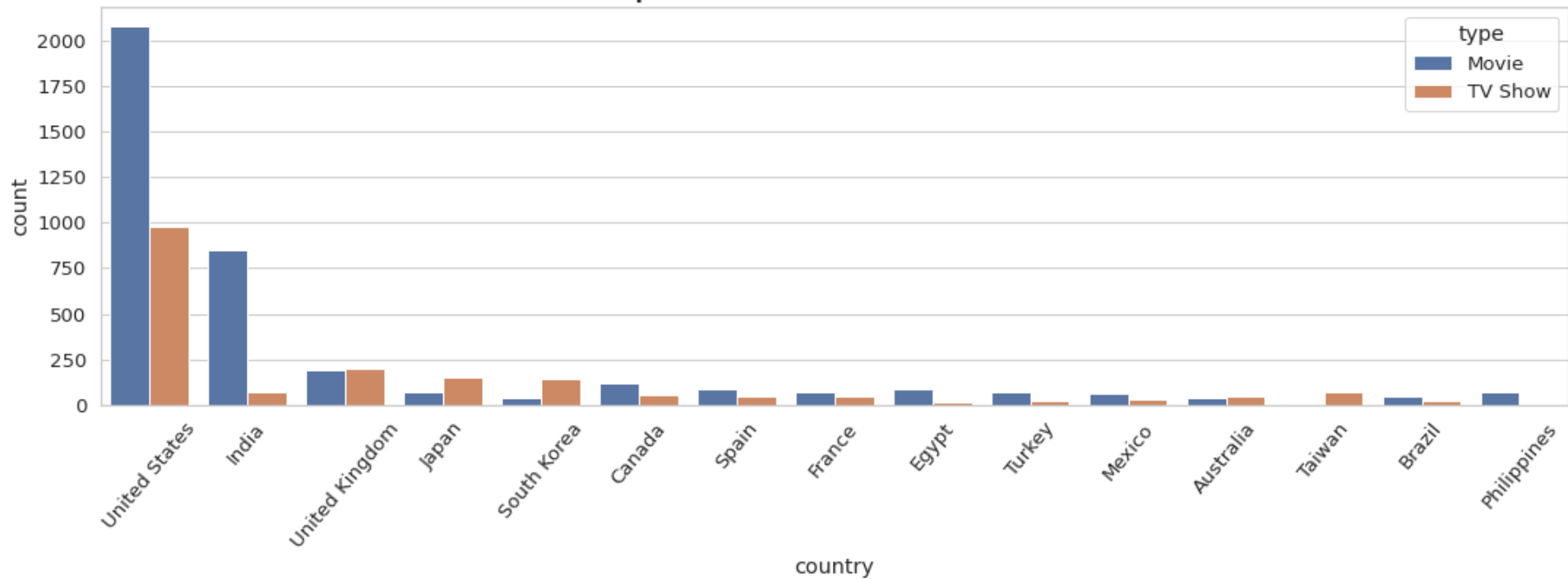


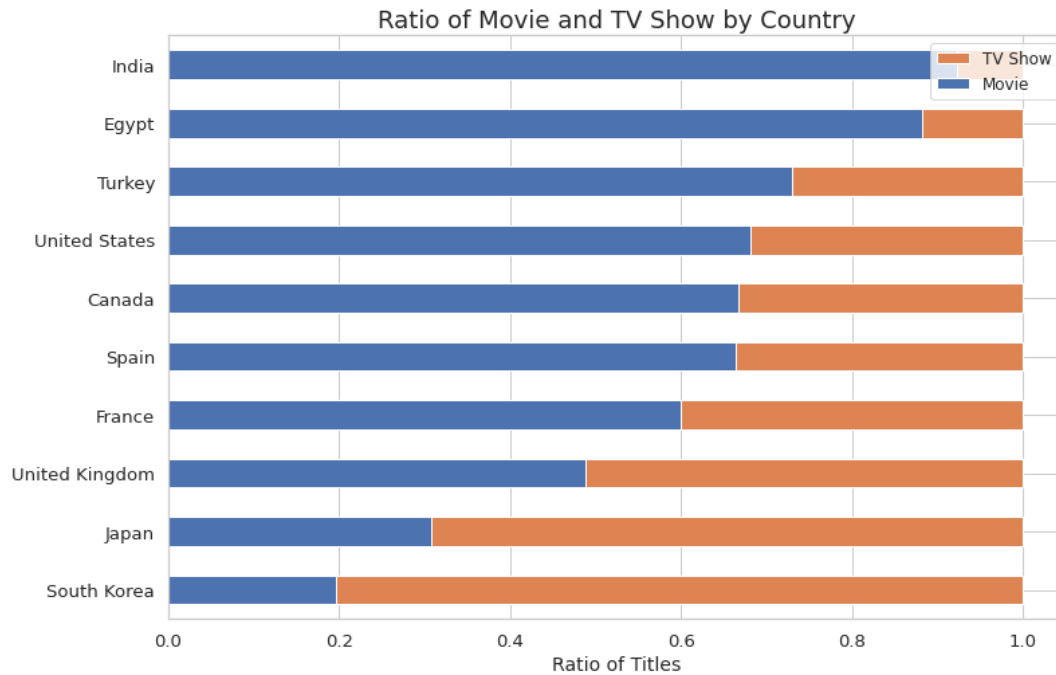
When analyzing the movie durations, it was observed that the majority of the movies have a duration between 50 to 150 minutes. On the other hand, the TV shows have a large number of single-season shows, which indicates that most of the TV shows on Netflix are relatively new.

Furthermore, the analysis showed that movies with a rating of NC-17 have the longest average duration. This might be because the movies with such a rating can explore more mature themes and include more explicit content, which requires a longer runtime to tell a compelling story.

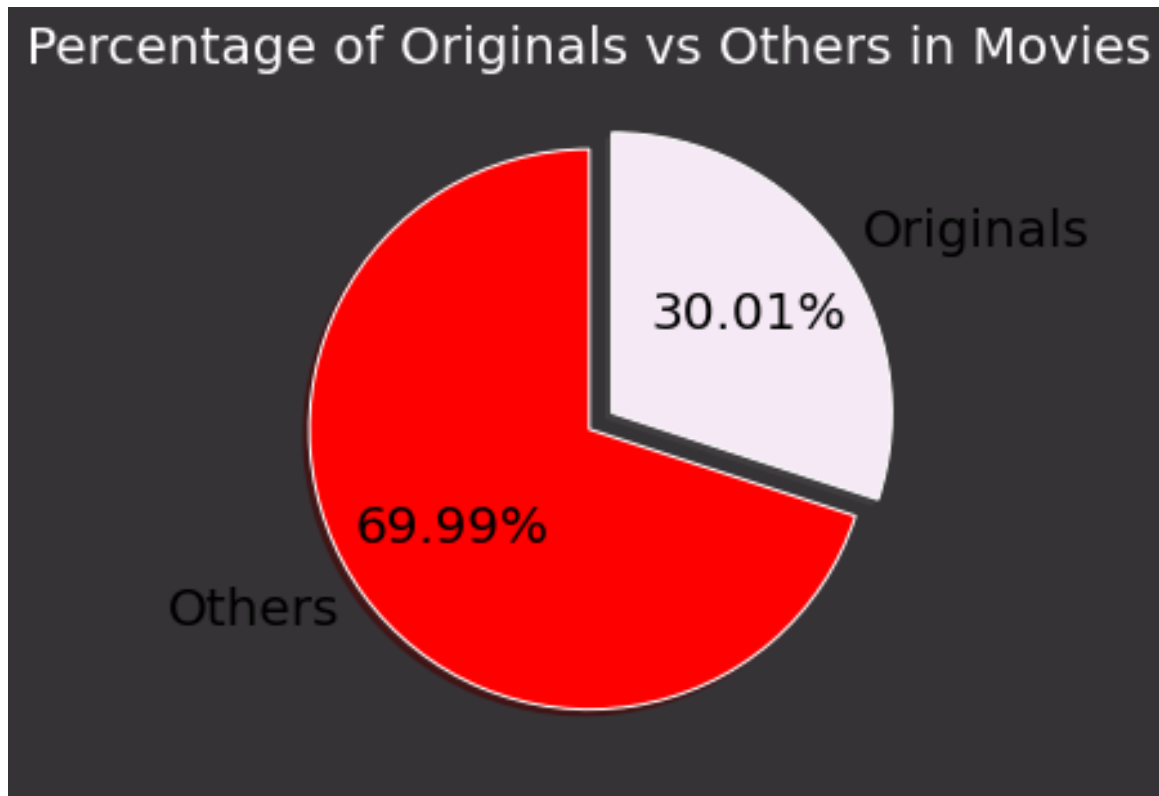
In contrast, the analysis also revealed that movies with a TV-Y rating, which is suitable for all children, have the shortest runtime on average. This suggests that the movies with this rating tend to be shorter and may have simpler plots and themes that are suitable for younger audiences.

Top 15 countries with most contents



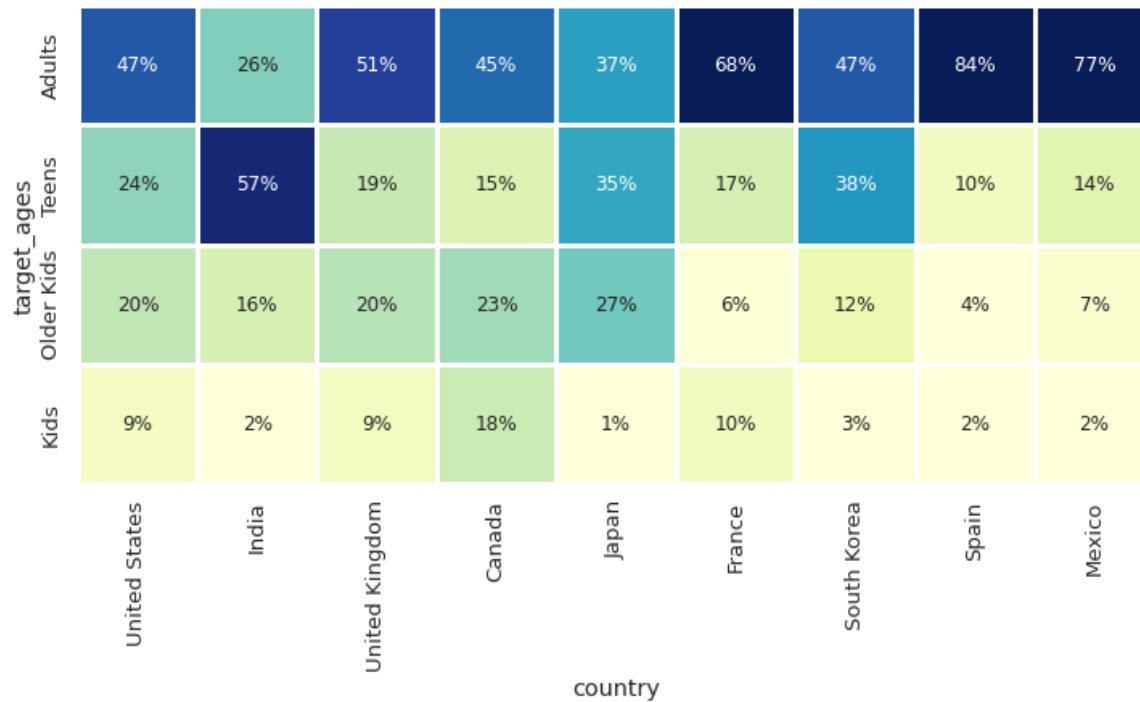


Netflix has the highest number of content in the United States, followed by India. India has the highest number of movies on Netflix.

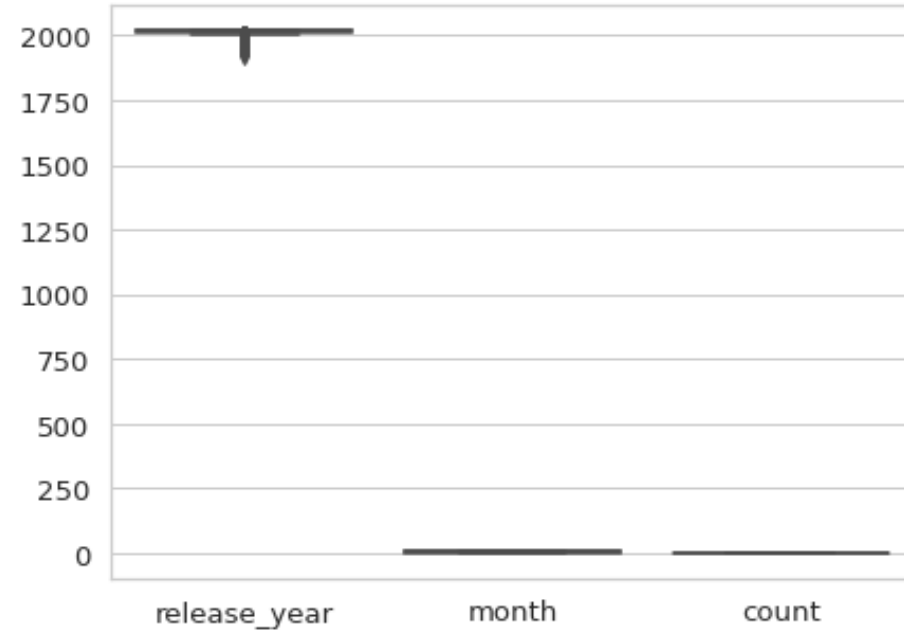
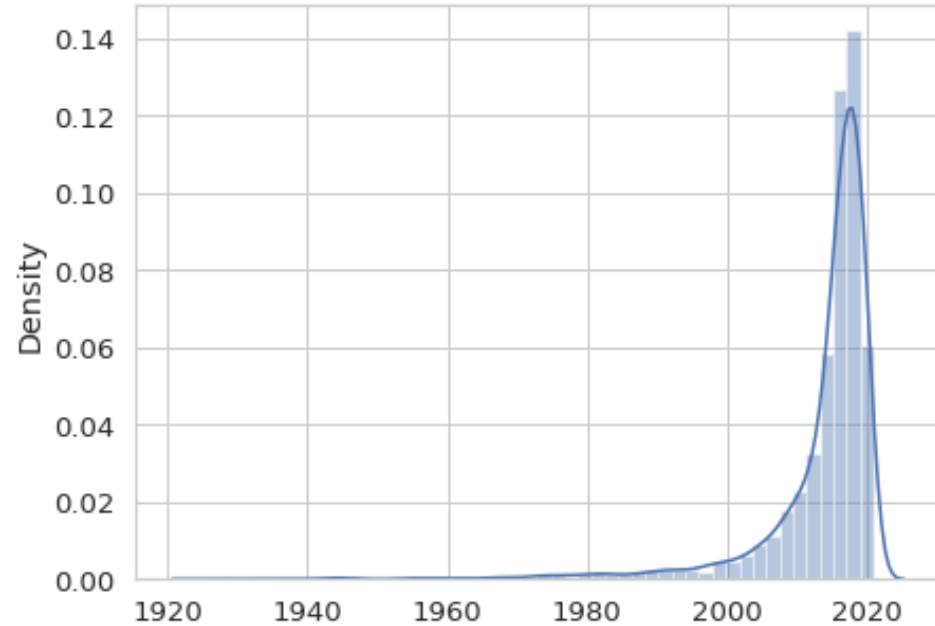


Netflix is known for producing original content, it is interesting to note that only 30% of the movies available on the platform were actually released by Netflix themselves. The remaining 70% of movies were added to Netflix after being released by different modes, such as theaters or other streaming platforms.

Correlation Heatmap



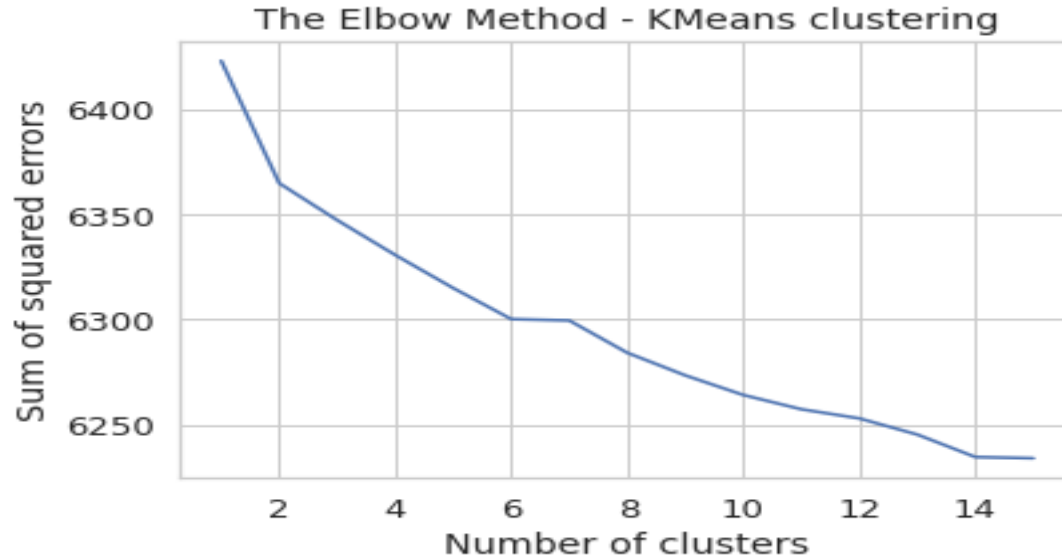
Outliers Handling



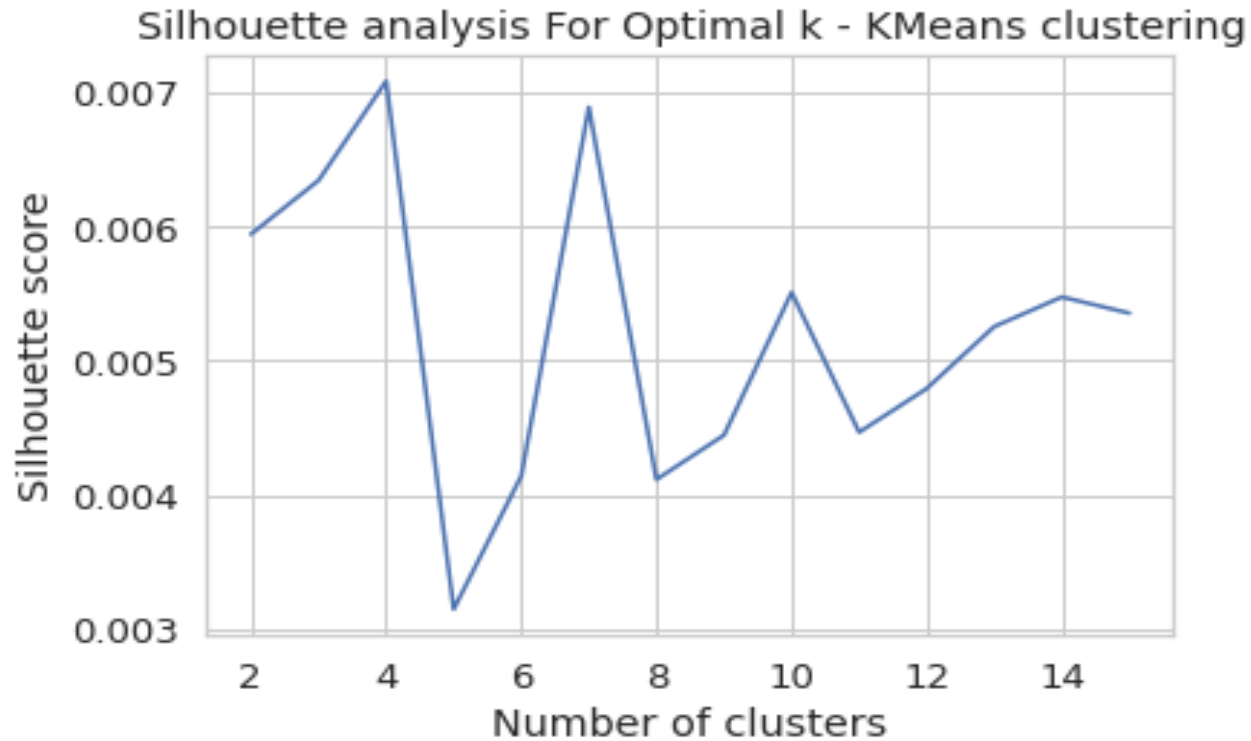
1. Except for the release year, almost all of the data are presented in text format.
2. The textual format contains the data we need to build a cluster/building model. Therefore, there is no need to handle outliers.

Cluster Implementation

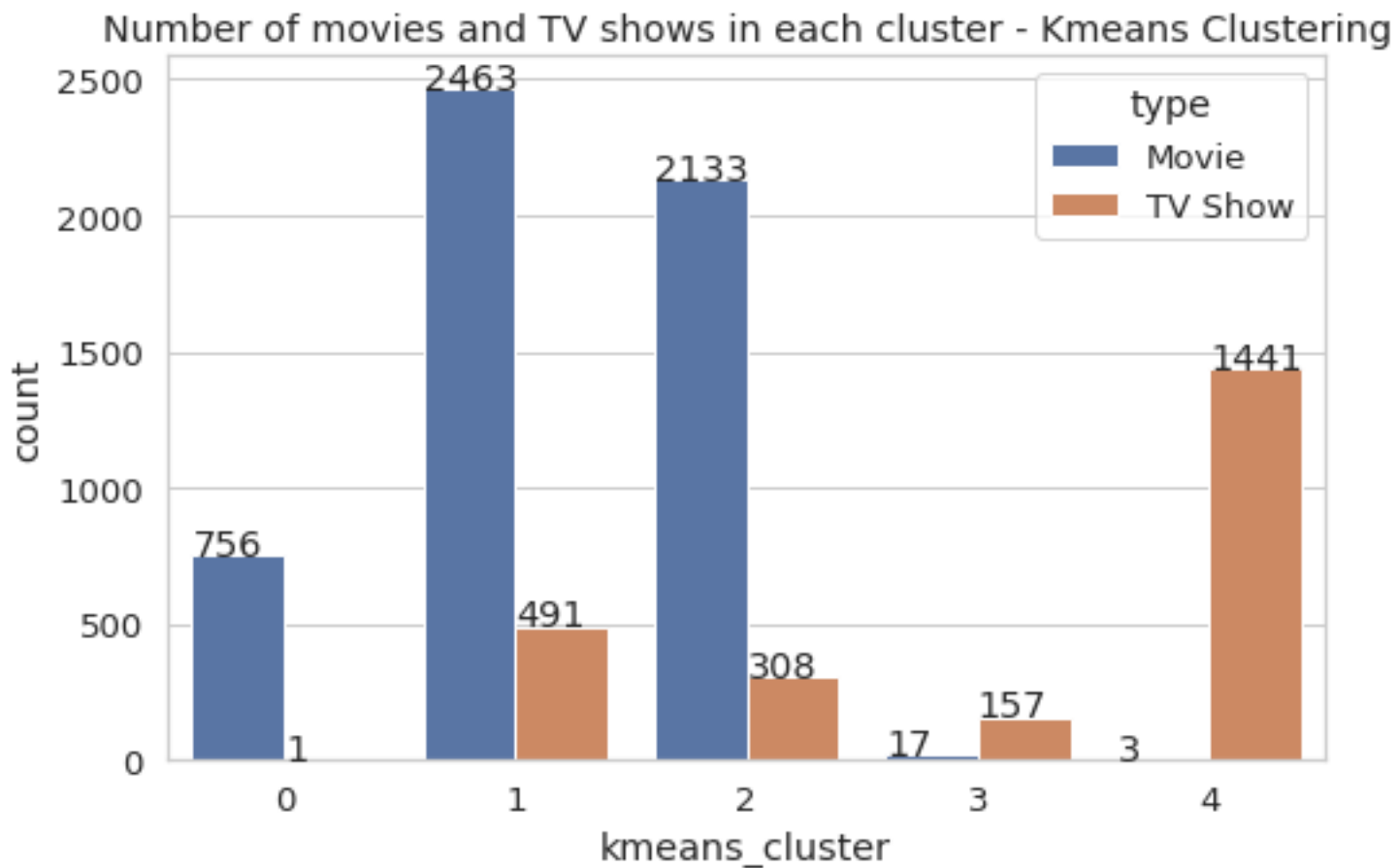
K-Means Clustering



The sum of squared distance between each point and the centroid in a cluster decreases with the increase in the number of clusters.

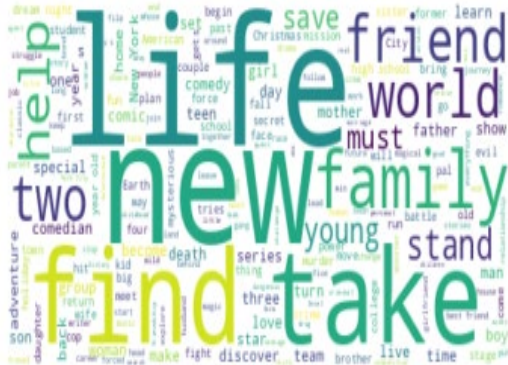


The highest Silhouette score is obtained for 5 clusters.

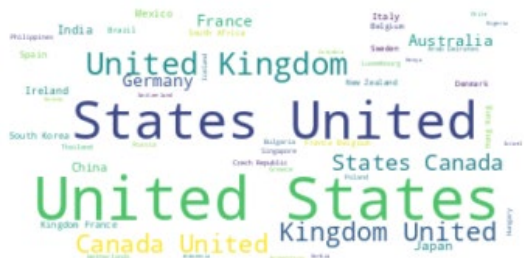
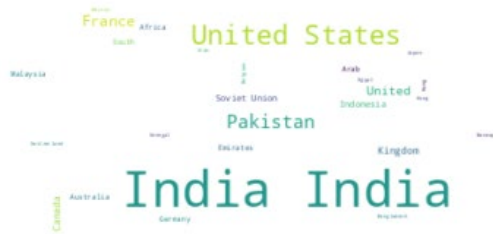


Successfully built 5 clusters using the k-means clustering algorithm.

Description



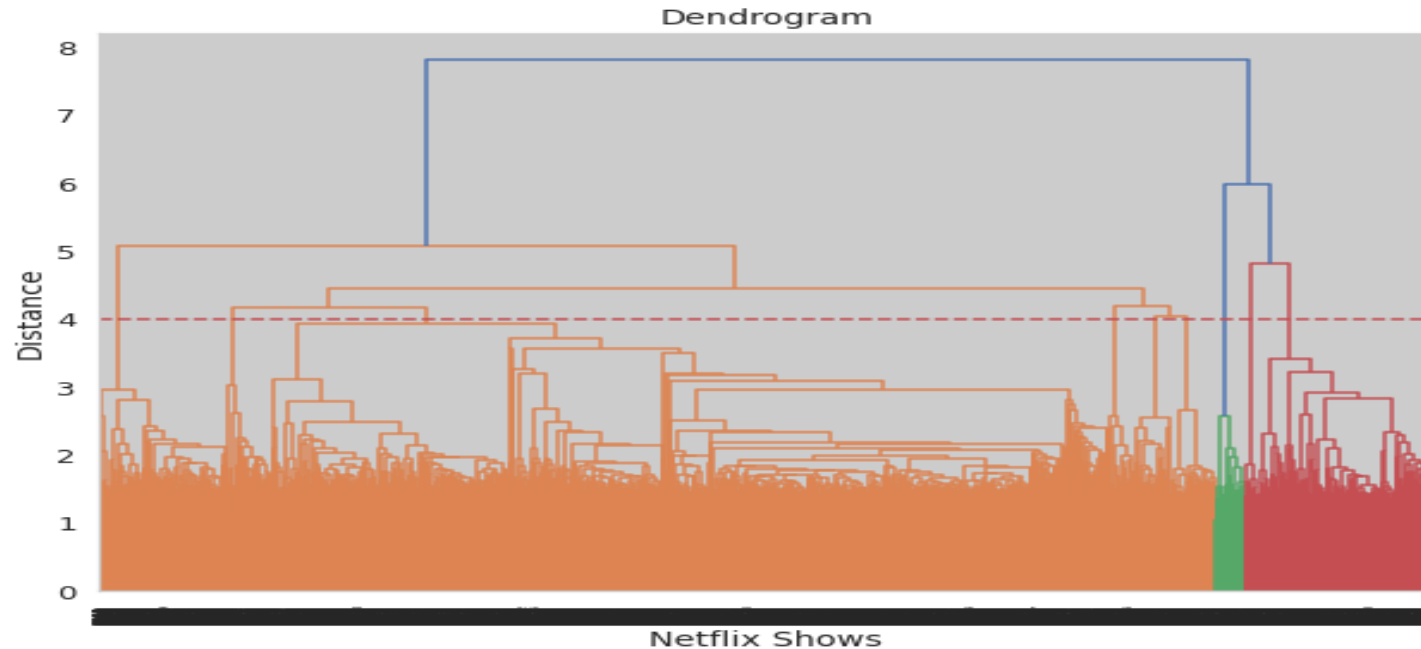




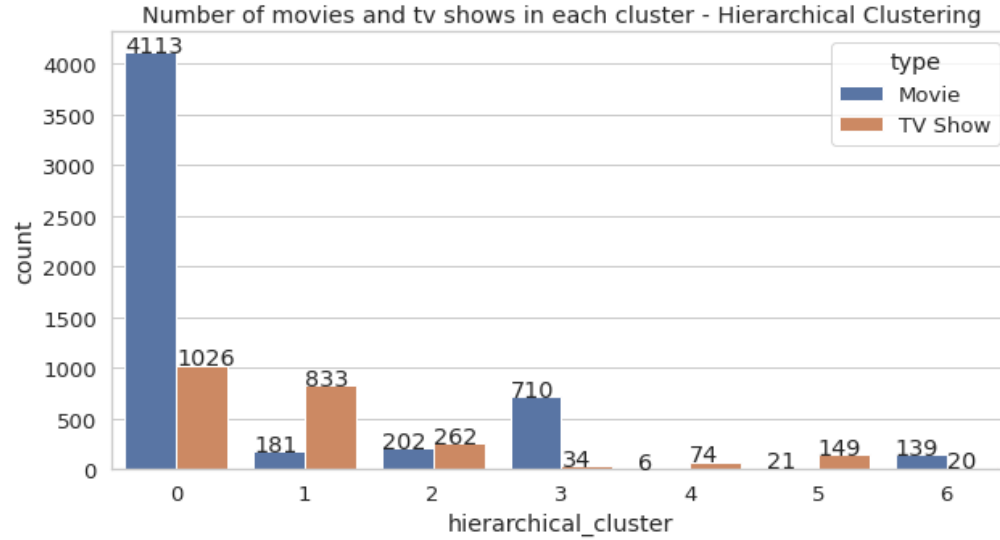




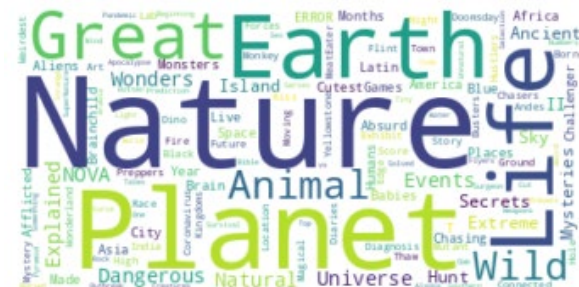
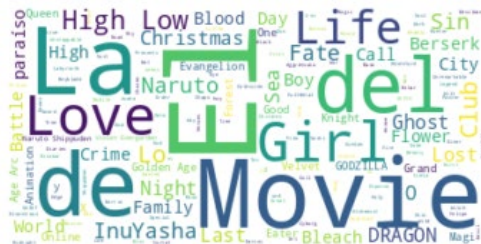
Hierarchical clustering



At a distance of 4 units, 7 clusters can be built using the agglomerative clustering algorithm.



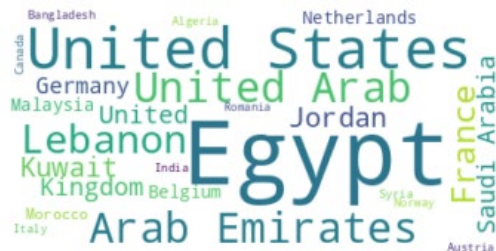
Successfully built 7 clusters using the Agglomerative (hierarchical) clustering algorithm.



Cast



Country



Listed_in



Conclusion

Welcome to our exciting journey of exploring the world of Netflix shows! Our goal was to cluster the shows into groups based on their similarities and differences, ultimately creating a content-based recommender system that suggests 10 shows based on the user's viewing history.

With over 7787 records and 11 attributes, we began our adventure by delving into the dataset's missing values and performing exploratory data analysis (EDA). Our findings revealed that Netflix boasts more movies than TV shows, with a rapidly growing collection of shows from the United States.

To cluster the shows, we focused on six key attributes: director, cast, country, genre, rating, and description. We transformed these attributes into a 10000-feature TFIDF vectorization, then used Principal Component Analysis (PCA) to tackle the curse of dimensionality. By reducing the components to 3000, we were able to capture more than 80% of the variance.

Next, we used two clustering algorithms, K-Means and Agglomerative clustering, to group the shows. K-Means determined that the optimal number of clusters was 5, as confirmed by the elbow method and Silhouette score analysis. Meanwhile, Agglomerative clustering suggested 7 clusters, which we visualized using a dendrogram.

But we didn't stop there. We then created a content-based recommender system using the similarity matrix obtained through cosine similarity. This system provides personalized recommendations based on the type of show the user has watched, giving them 10 top-notch suggestions to explore.

Join us in discovering the diverse world of Netflix shows, and let our recommender system guide you to your next binge-worthy obsession.