Capstone Project 4

Netflix Movies and TV shows Clustering

By:- Viral Bhatu Shewale

Content:

- Problem Statement
- Data Summary
- Data Cleaning
- Exploratory Data Analysis (EDA)
- Data Processing
- Dimensionality Reduction
- Modelling Approach
- Clustering Model
- Conclusion

Abstract:

Although OTT platform such as Netflix is a recent phenomenon, rapid growth has made it a leading competitor in the streaming media and production firm.

Netflix subscriber count as of **2022** is **220.67 million.**

It is crucial that they effectively cluster the shows that are hosted on their platform in order to enhance use experience.

Problem Statement:

- 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.

In this project, we are required to do

- Understanding what type content is available in different countries.
- Is Netflix has increasingly focusing on TV rather than movies in recent years.
- Clustering similar content by matching text-based features.

Data Summary:

There are 12 **different attributes (columns) and** 7787 **rows**

- **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

Data Cleaning:

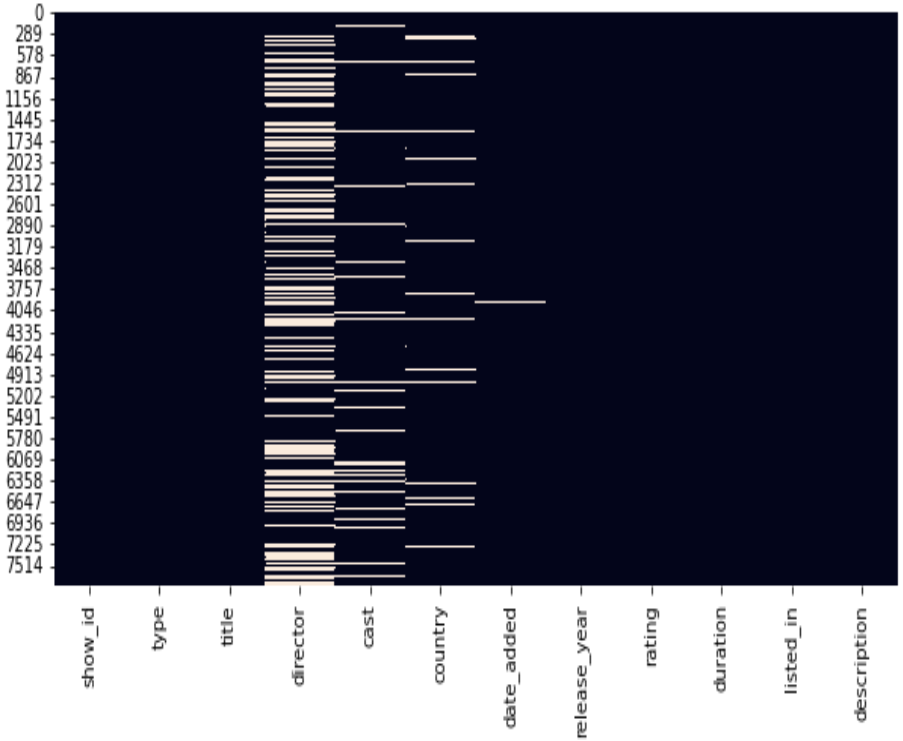
➢ **Missing value columns:**

- Director (2389), cast (718), country (507), date_added (10), rating (7)

➢ **Handling Null values:**

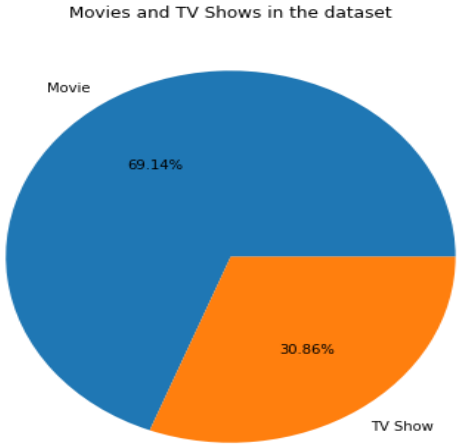
- The missing values in the director, cast, and country attributes replaced with 'Unknown'.
- date_added columns has only 10 missing values hence I dropped that missing values.
- The missing values in the rating column can be imputed with its mode since this attribute is discrete.

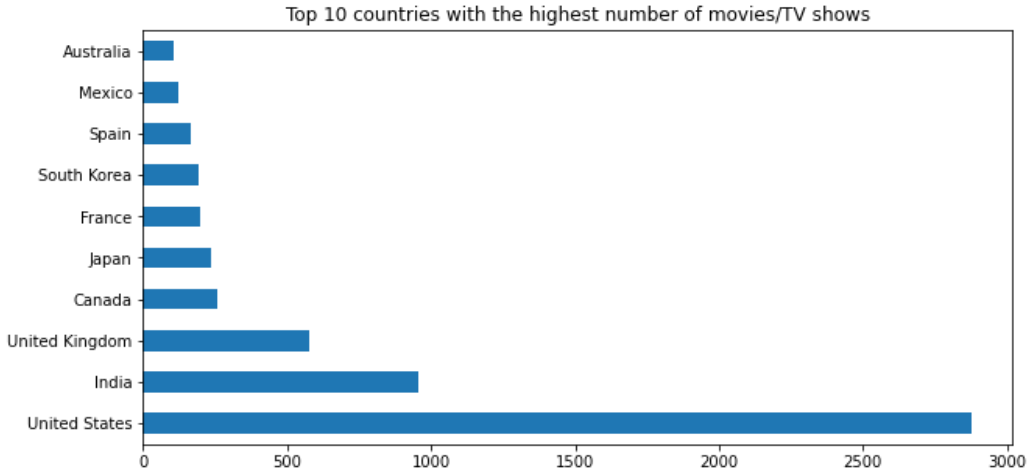
Only primary genre and country were selected to simplify the EDA.



Exploratory Data Analysis:

- There are more **Movies** (**69.14%**) than **TV Show** (**30.86%**).
- The highest number of movies / TV shows were from the **US**, followed by **India** and **UK**.
- The **top 3** countries together account for about **56%** of all movies and TV shows in the dataset.

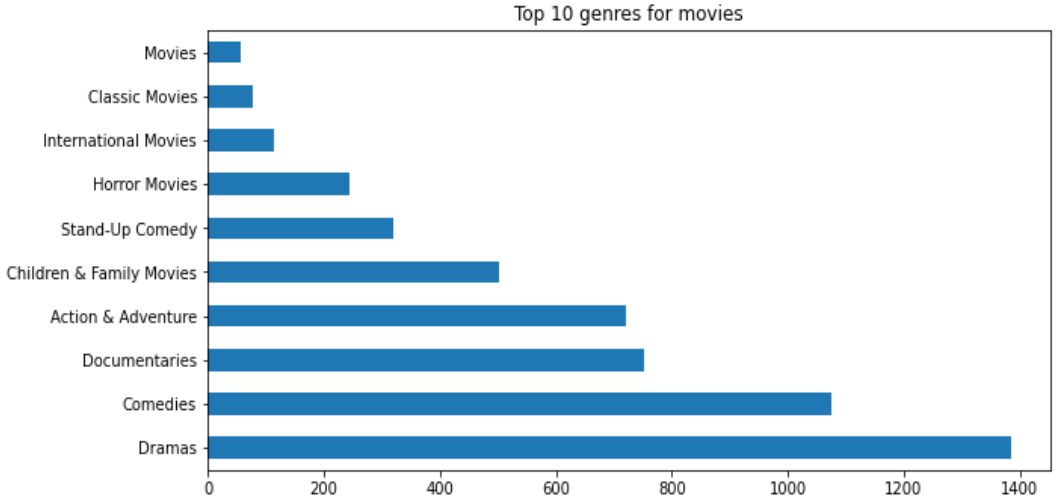


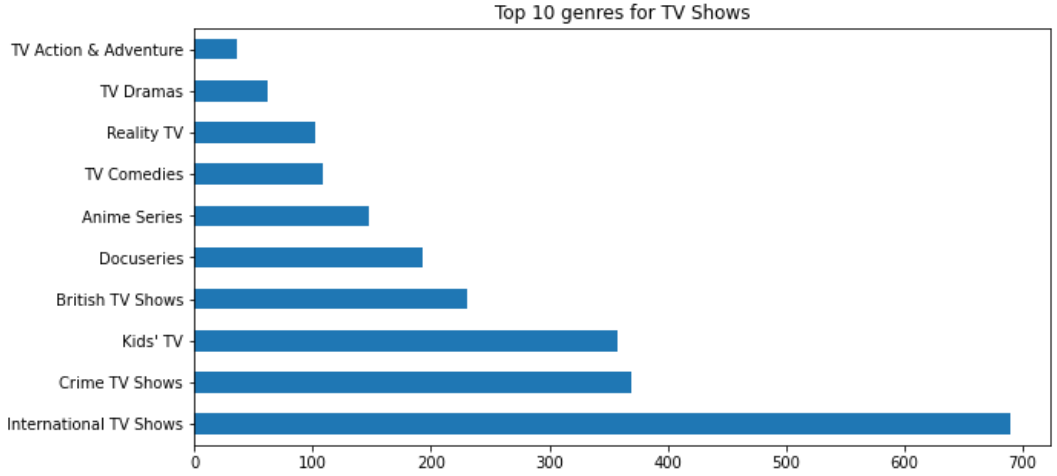


EDA:

- **Dramas**, **comedies**, and **documentaries** are the most popular genre for **movies** on Netflix.

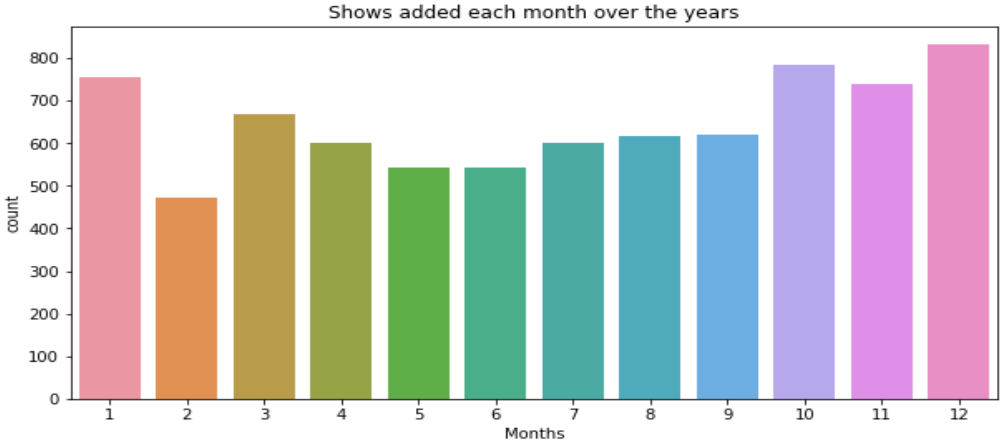
- **International**, **crime**, and **kid** are the most popular genre for TV shows on Netflix.

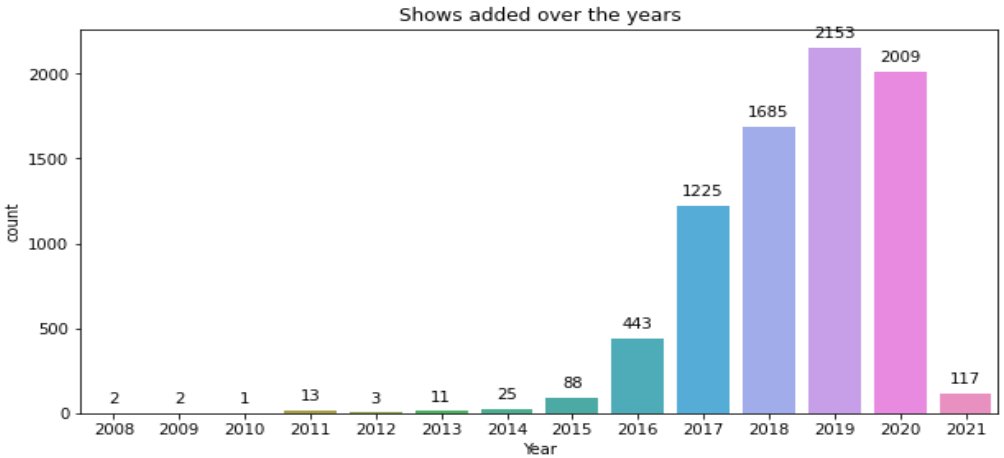




EDA:

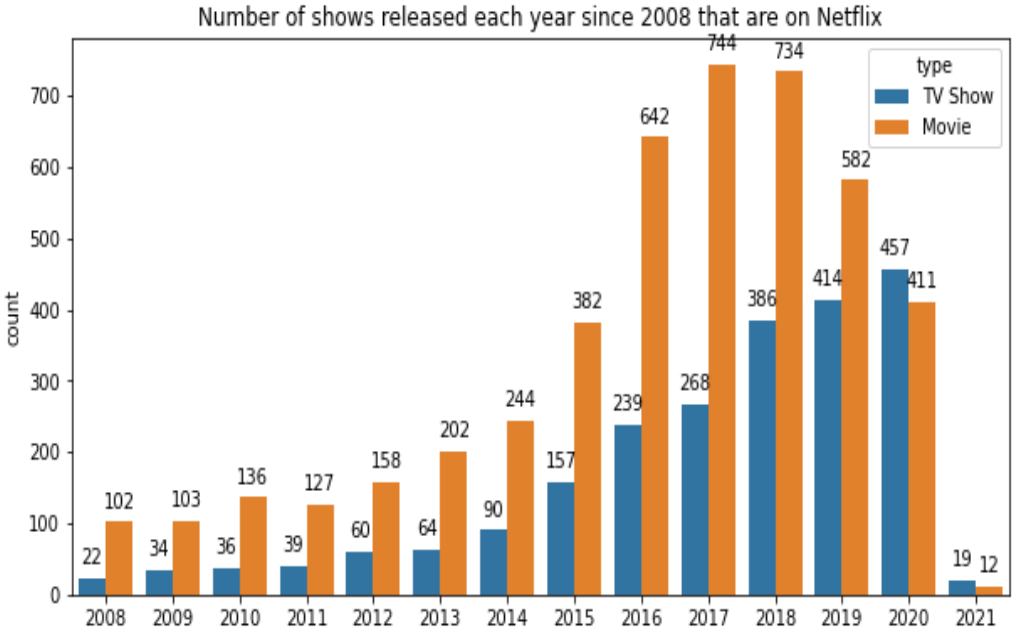
- More movies/TV shows were added to Netflix in **January**, **October** and **December**.
- **Netflix** has continued to add more shows to its platform over the years.
- The year **2020** saw a **drop** in the number of shows added, which can be attributed to the **Covid-19-induced lockdown**, which halted the production of the shows.
- We have Netflix data only up to 16th January 2021.





EDA:

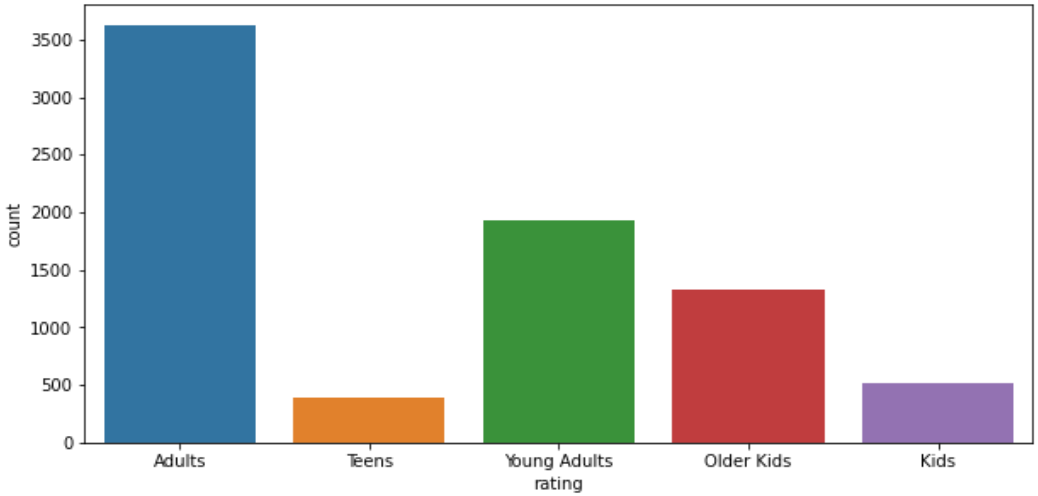
- Over the years, Netflix has consistently focused on adding more shows to its platform.
- Though there was a decrease in the number of movies added in 2020, this pattern did not exist in the number of TV shows added in the same year.
- This might signal that **Netflix is increasingly concentrating on introducing more TV series** to its platform rather than movies.



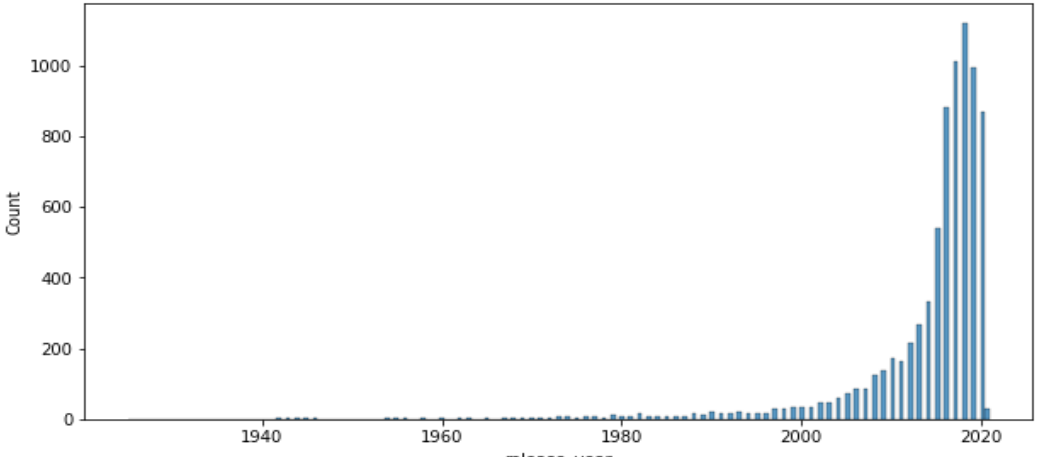
EDA:

- Around **50%** of shows on Netflix are for an **adult audience**, Followed by **young adults**, older kids and kids. Netflix has the least number of shows that are specifically for teenagers than other age groups.

- There were more **new movies/TV shows** on Netflix than the old ones**.**

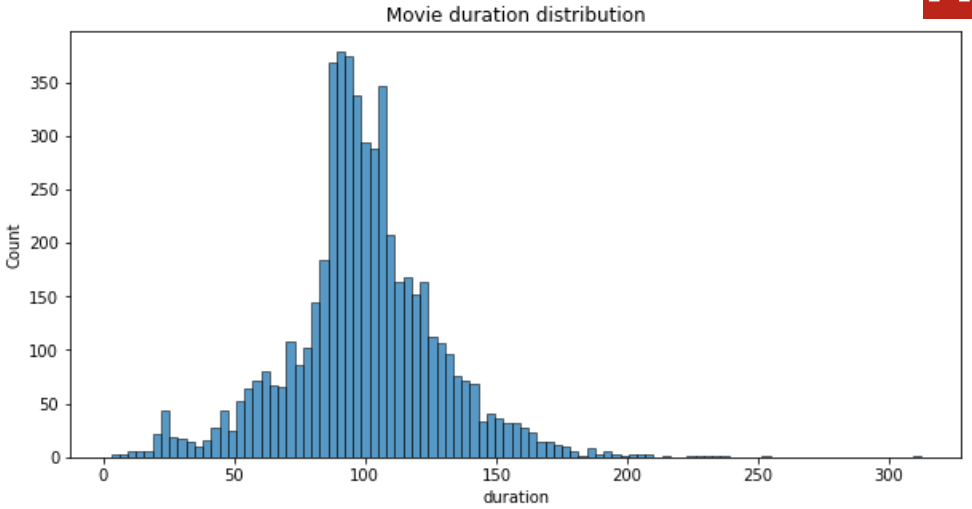


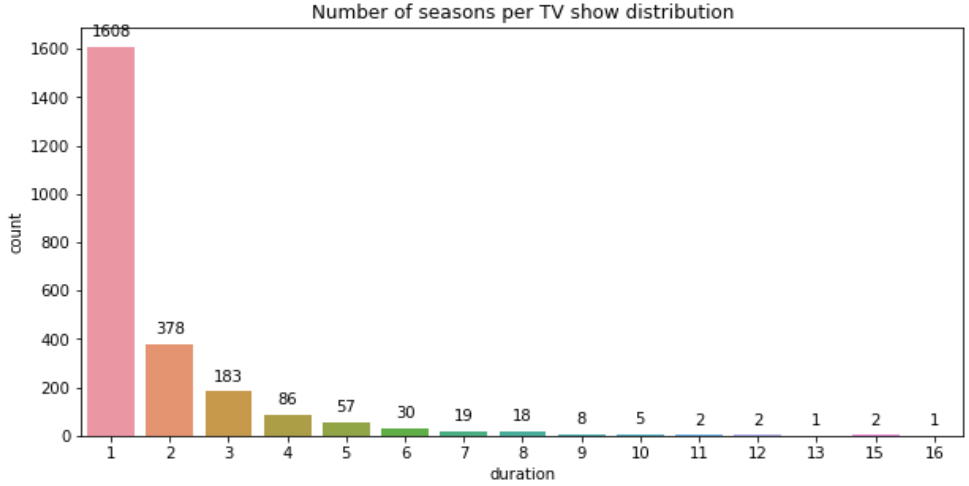
distribution by released year



EDA:

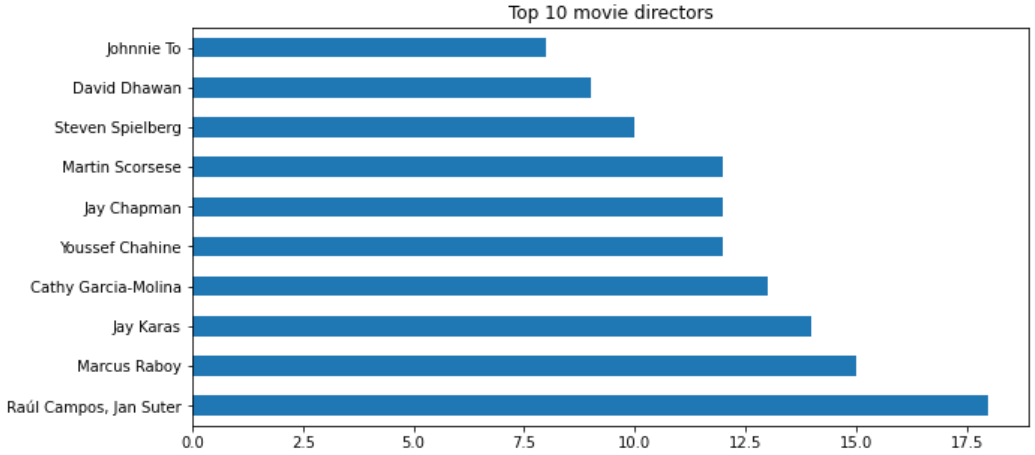
- Length of the movie is **normally distributed**.
- The TV series in the dataset have **up to 16 seasons**. However, many TV shows only have **one season**.

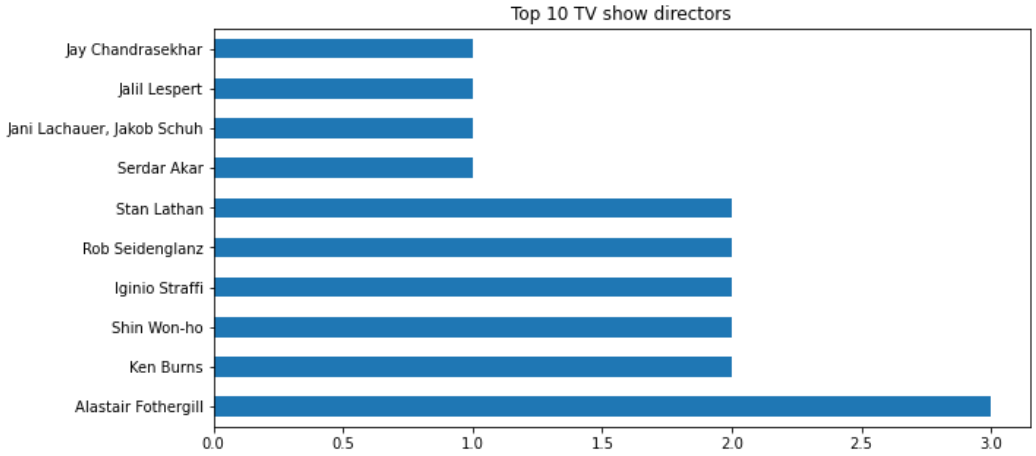




EDA:

- **Raul Campos** and **Jan Suter** have together directed **18 movies**, higher than anyone yet.
- **Alastair Fothergill** has directed **three TV shows**, the most of any director.
- Only six directors have directed more than one television show.





Data Processing:

Feature Engineering:- clusters are built on the attributes Director, cast, country, listed_in (Genre), description.

Steps involved on the data pre-processing:-

1. Text preprocessing: Remove all non-ASCII characters, stop-words and punctuation marks, and convert all textual data to lowercase.
2. Lemmatization generates a meaningful word out of a corpus of words.
3. Tokenization of corpus.
4. Word vectorization.
5. Dimensionality reduction.

Vectorization:

- **TFIDF vectorizer**, where TFIDF stands for - Term Frequency Inverse Document Frequency.

$$TF = \frac{\text{number of times the term appears in the document}}{\text{total number of terms in the document}}$$

$$IDF = log(\frac{\text{number of the documents in the corpus}}{\text{number of documents in the corpus contain the term}})$$

$$TF\text{-}IDF = TF * IDF$$

- TF-IDF is better than Count Vectorizers because it not only focuses on the frequency of words present in the corpus but also provides the importance of the words. (20000 attributes)

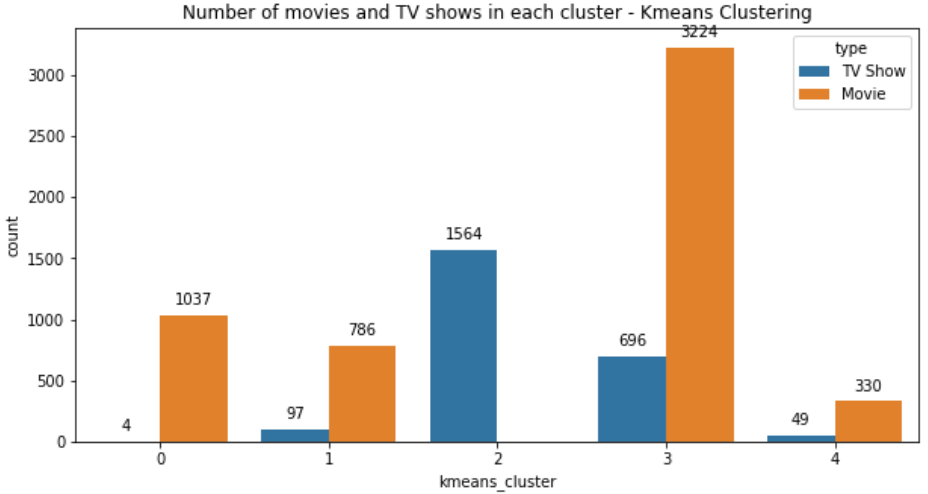
Dimensionality reduction:

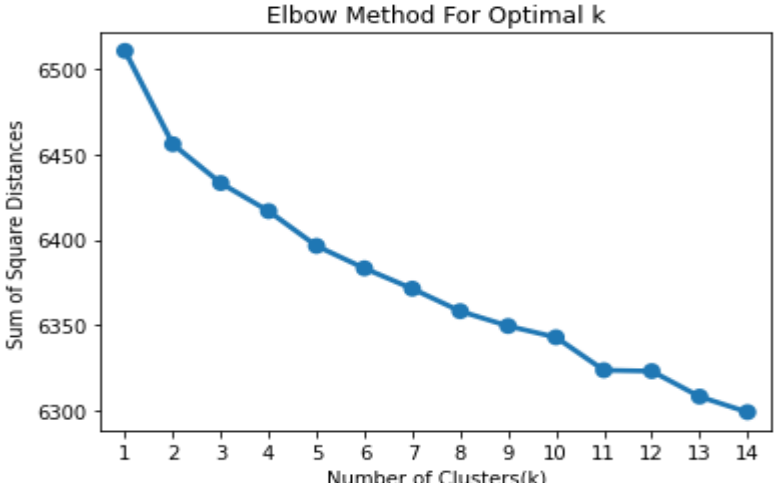
➢ Principal component analysis:

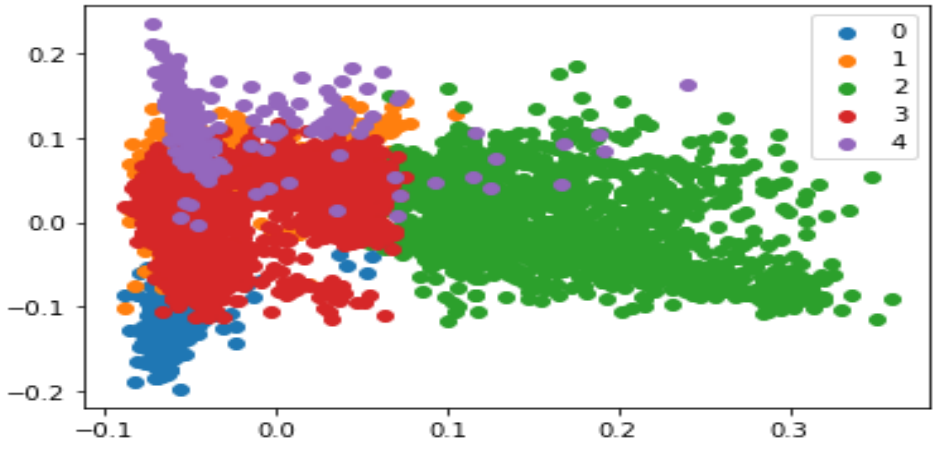
- PCA is a statistical technique for reducing the dimensionality of a dataset.
- **100%** of the variance is explained by **+7500** components.
- **84.57%** of the variance is explained just by **4000 components**. Hence simplify the model, and reduce dimensionality, we can take **the top 4000 components**, which will still be able to capture more than 80% of the variance.
- Shape of the dataset: (**7787, 4000**)

K-Means Clustering:

- Number of clusters = **5**
- Silhouette score =**0.0079**
- Distortion = **6397.62**

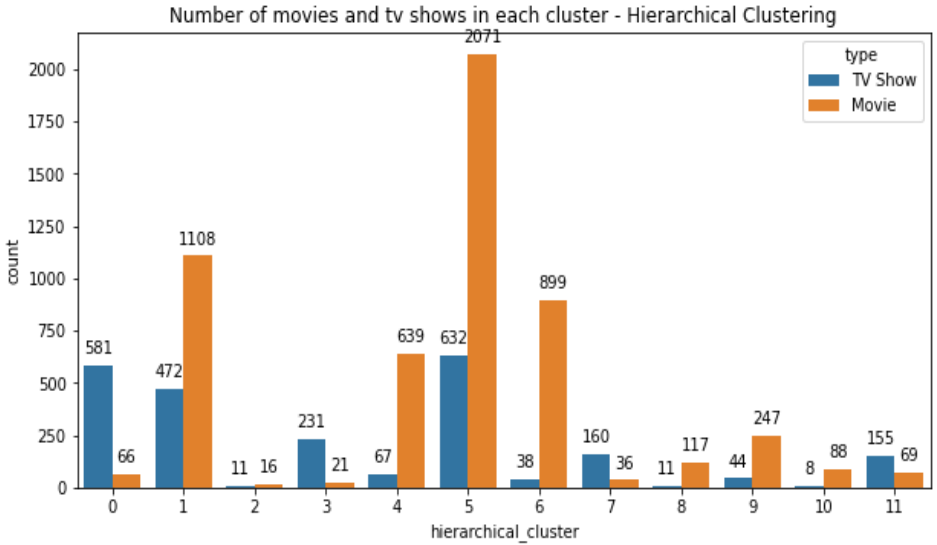


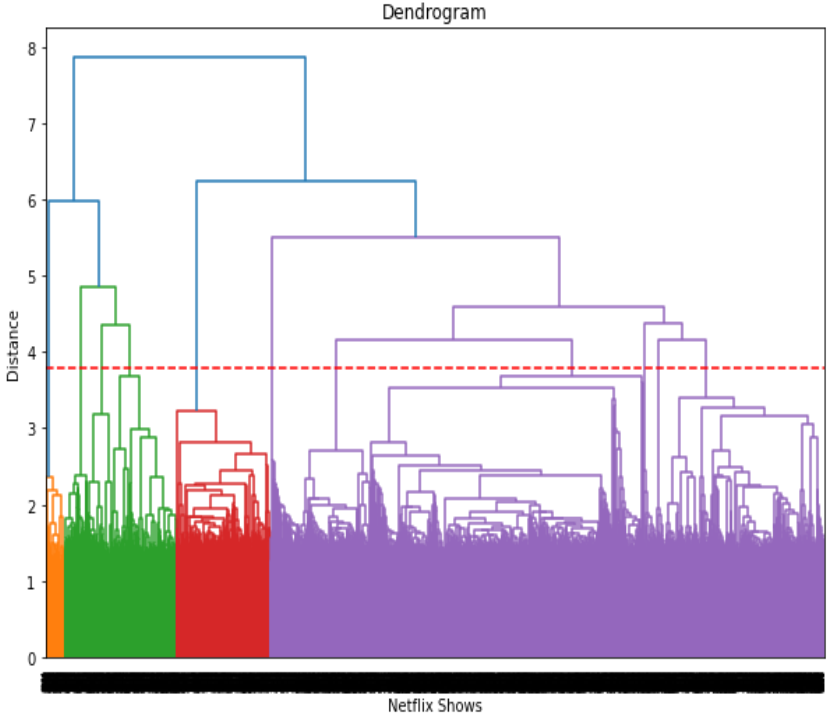




Hierarchical Clustering:

- **Agglomerative clustering**
- Number of clusters = **12**
- Linkage =**ward**
- Distance = **Euclidean**





Conclusion:

- In this project, I worked on a text clustering problem wherein I had to classify/group the Netflix shows into different clusters where each data point belongs to only one group.
- The dataset contained 7787 records and 11 attributes.
- I started by dealing with missing values of datasets and doing exploratory data analysis (EDA).
- Netflix hosts more movies than TV shows on its platform, and the total number of shows added on Netflix is growing exponentially.
- United state produces the highest number of shows.
- Cluster the data based on the attributes: director, cast, country, genre, and description. The values in these attributes were tokenized, preprocessed, and then vectorized using the TFIDF vectorizer. Through TFIDF Vectorization, I created a total of 20000 attributes.
- I used principal component analysis to reduce the dimensionality of the data. 84.57% of the variance is explained only by 4000 components, so the number of components was restricted to 4000.

- The Elbow method and Silhouette score were used to decide the optimal number of clusters for the K-means clustering algorithm, and the obtained number of the Clusters was 5 (k=5), and then the K-Means clustering was built.
- Then clusters were created using the agglomerative clustering algorithm, and the optimal number of Clusters turned out to be 12 after looking at the dendrogram.