

Capstone Project

Netflix Movies and TV shows Clustering

Submitted by: Kabeer Pande

Table of Content

- **Introduction**
- **Problem Statement**
- **Data Description**
- **Data Cleaning**
- **Exploratory Data Analysis**
- **Feature engineering, Text Cleaning and NLP**
- **Stemming and Vectorization**
- **PCA**
- **Implementing Machine learning Models**
- **Recommendations**
- **Conclusion**

Introduction

Netflix is the most widely used media and video streaming platform. It has around 200 million users World Wide. It offers a large library of movies and TV series that can be accessed at any time via internet services.

Netflix works on a subscription based model, where users get unlimited access to content with a paid subscription and users can cancel their subscription whenever they want that's why Netflix always needs to provide the best content to retain their subscriber because of this, it's essential to have a recommendation system that gives customers helpful suggestions.



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.

- **In this project, you are required to do**
- 1:Exploratory Data Analysis
- 2:Understanding what type content is available in different countries
- 3:Is Netflix has increasingly focusing on TV rather than movies in recent years.
- 4:Clustering similar content by matching text-based features

Data Description

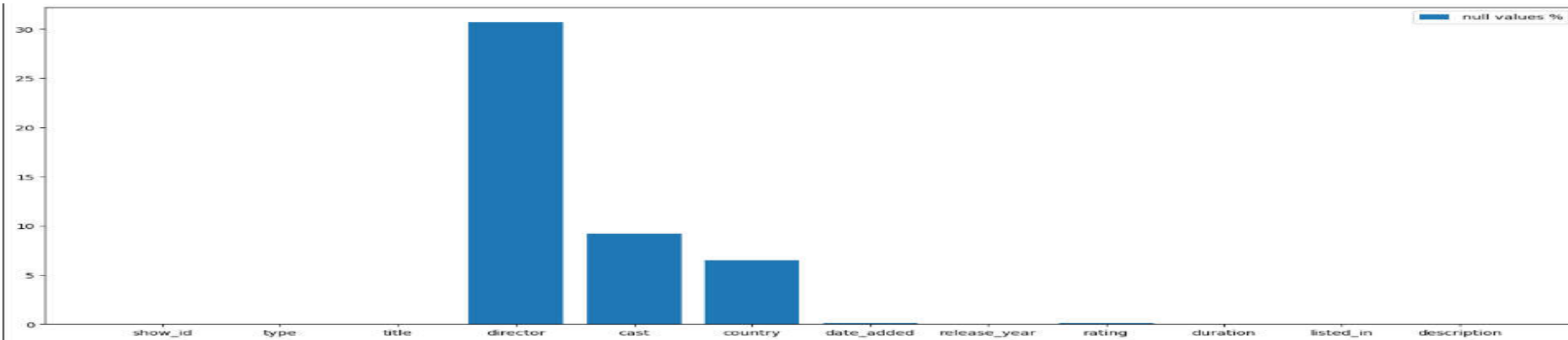
- The dataset have 7787 rows and 12 columns.
- **Attribute Information:**
 - 1. show_id : Unique ID for every Movie / Tv Show
 - 2. type : Identifier - A Movie or TV Show
 - 3. title : Title of the Movie / Tv Show
 - 4. director : Director of the Movie
 - 5. cast : Actors involved in the movie / show
 - 6. country : Country where the movie / show was produced
 - 7. date_added : Date it was added on Netflix
 - 8. release_year : Actual Release year of the movie / show
 - 9. rating : TV Rating of the movie / show
 - 10. duration : Total Duration - in minutes or number of seasons
 - 11. listed_in : Genre
 - 12. description: The Summary description

Data Cleaning

Director ,cast, country, date_added and rating columns contains around 31%,9%,7%,0.13% and 0.09% missing values respectively.

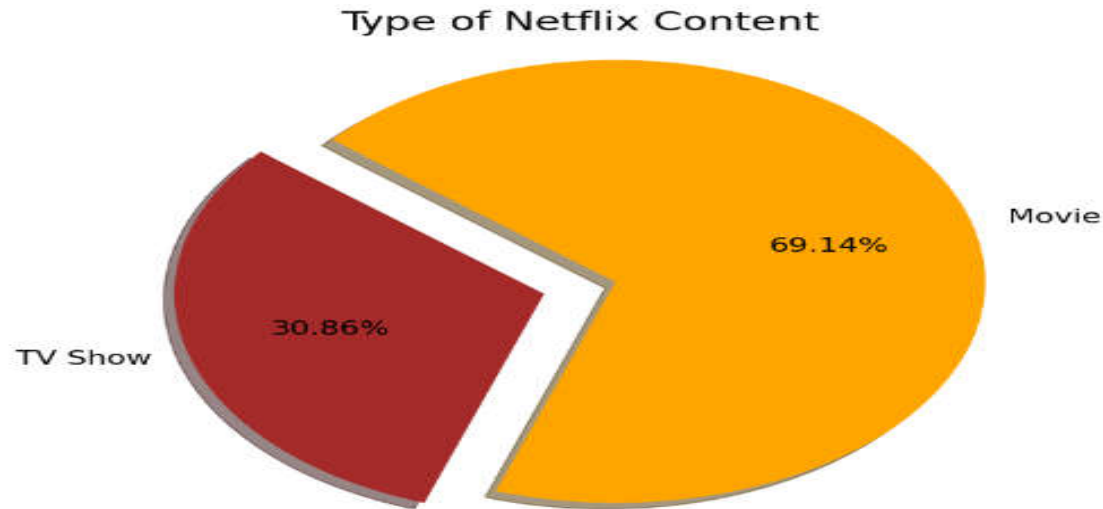
Since the percentage of director, cast and country columns are high, so we cannot drop these values because it can affect our data , therefore we replace the null vales to unavailable. The missing values percentage of columns date_added and rating are minimal so we can simply drop these value.

We added one more feature year_added, extracted from date_added column for EDA.



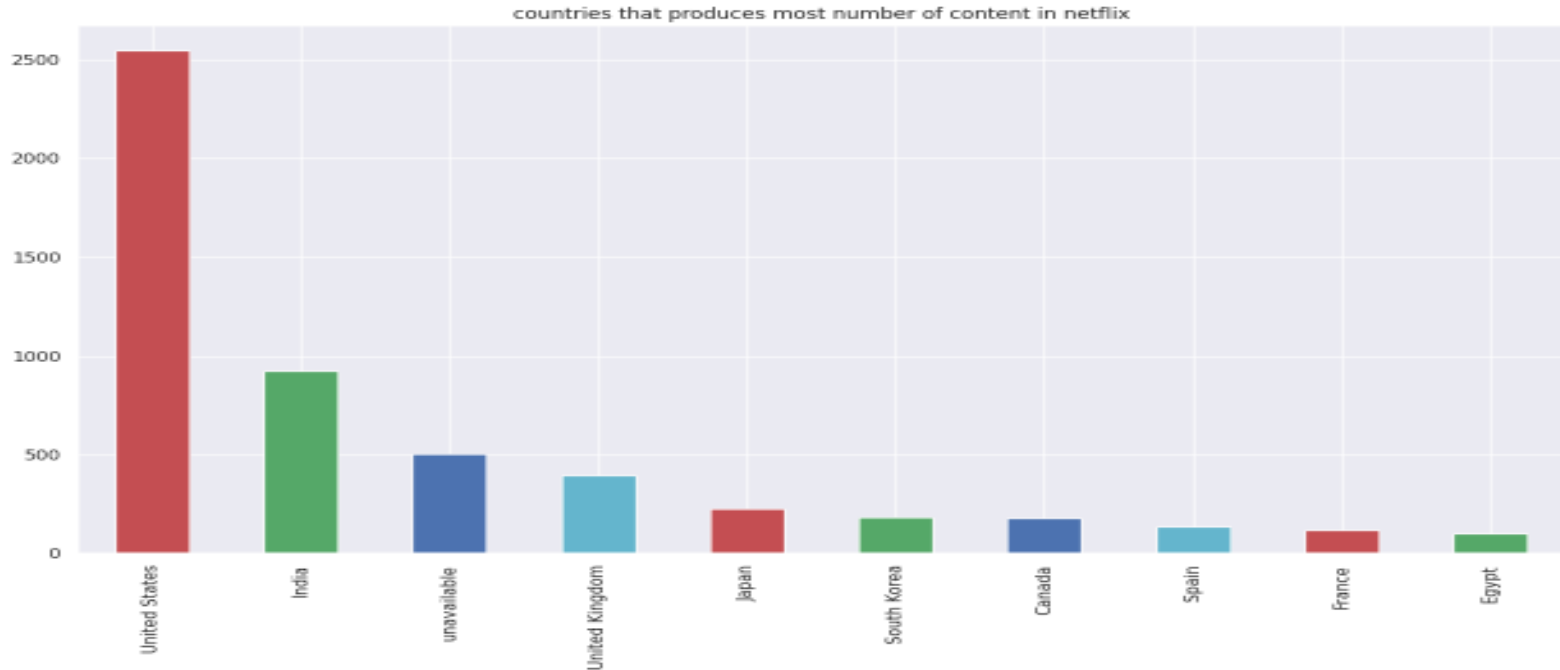
Exploratory Data Analysis

- **Distribution of type of content in the dataset**



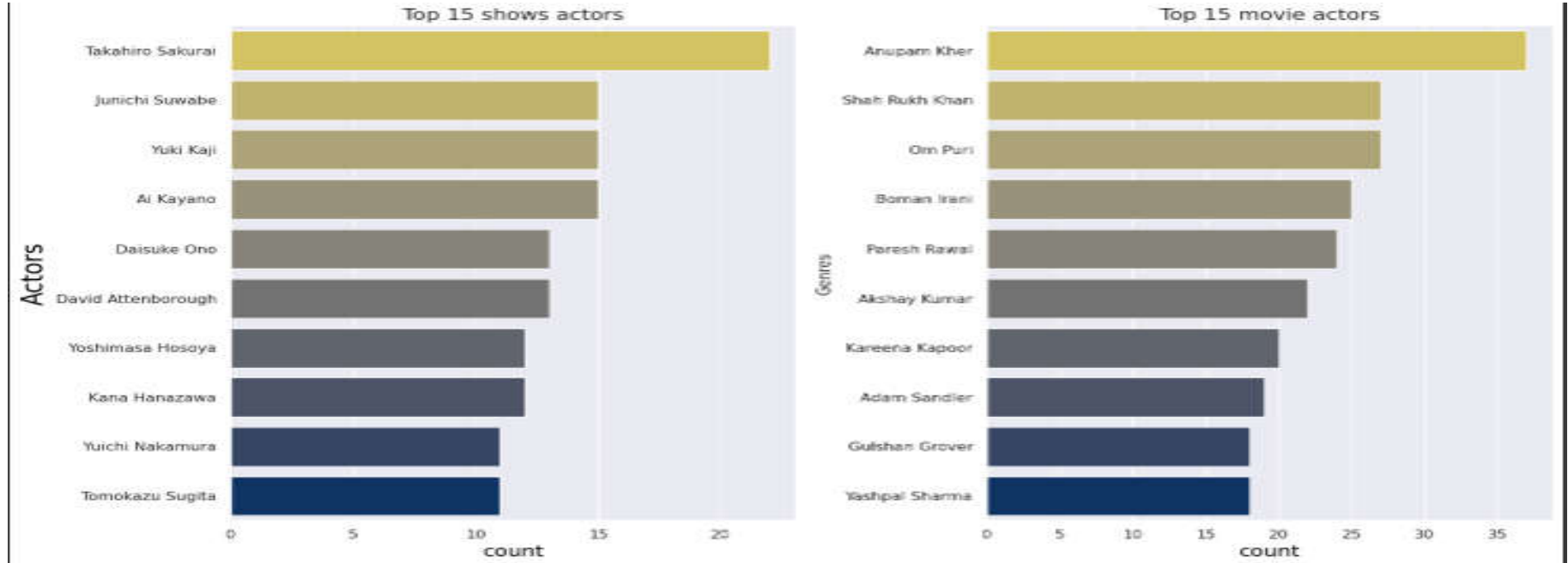
Netflix has around 69% movies and 31% TV shows

Countries that produce most number of content



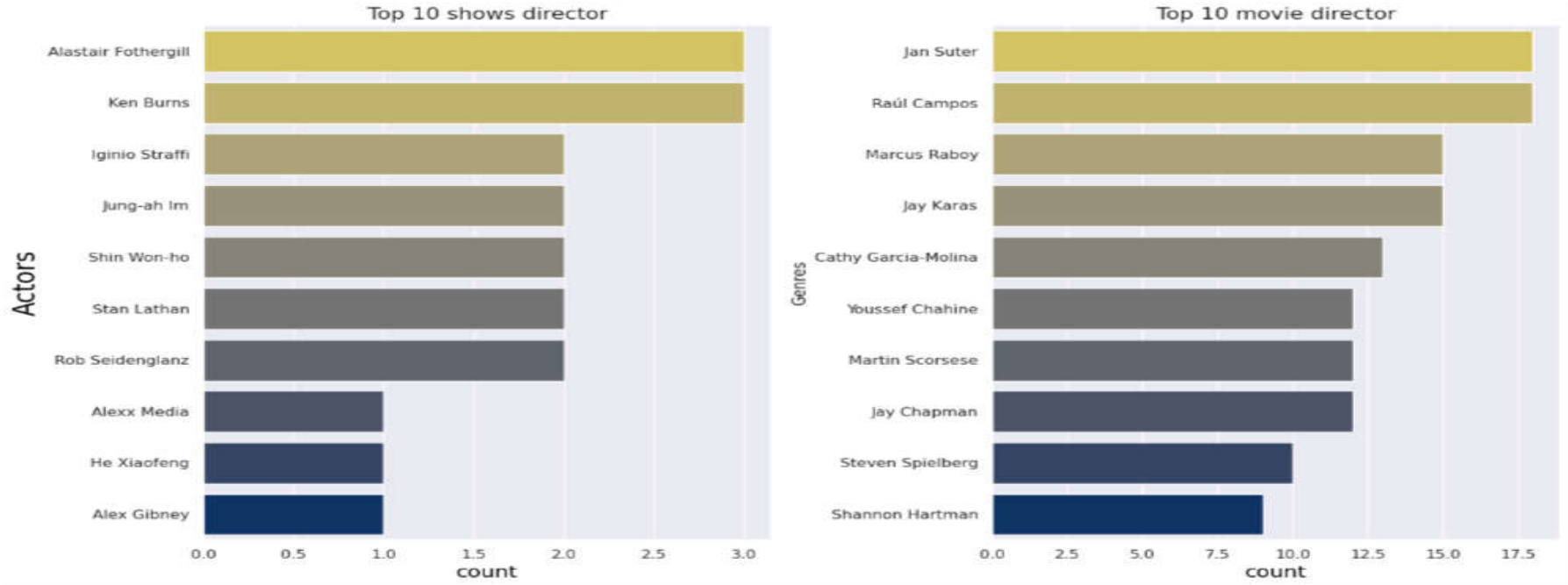
United States contributes maximum content to the platform followed by India

Actors who worked in most content in movies and TV shows



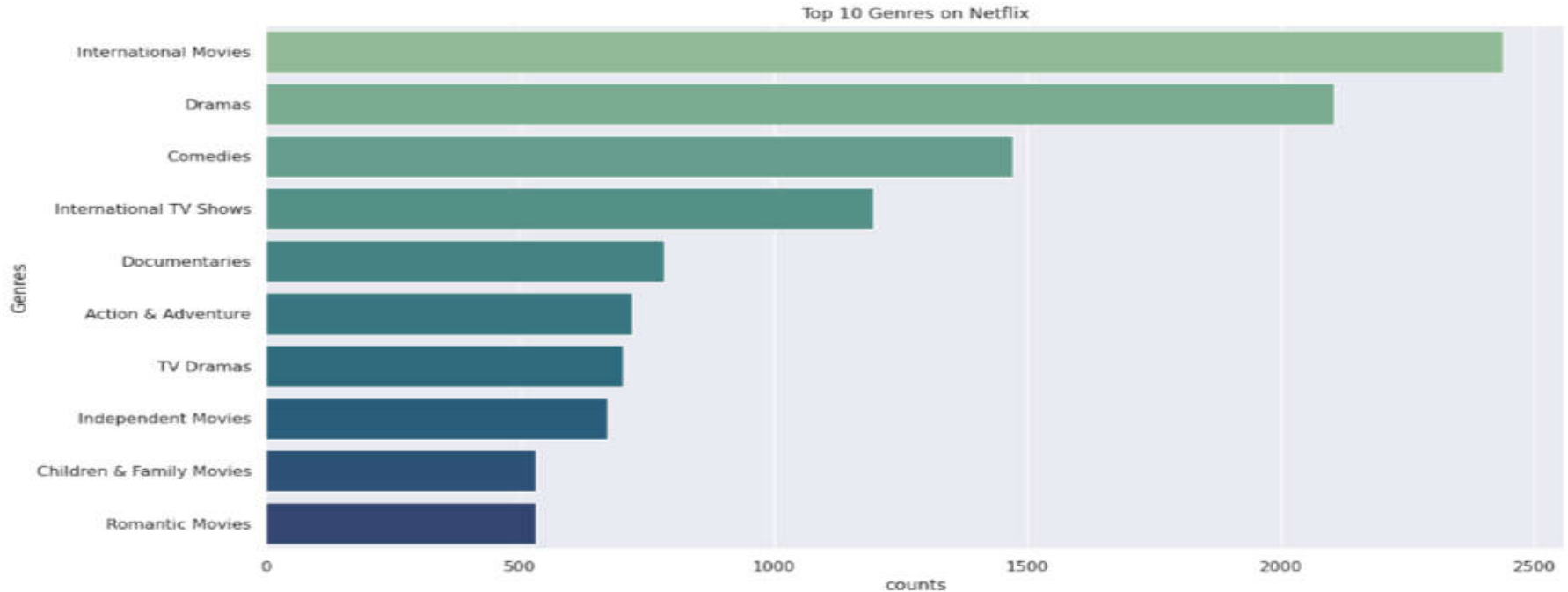
Takahiro Sukurai has worked in maximum TV-Shows and Anupam Kher has worked in maximum movies.

Top director who have worked in most content in movies and TV shows



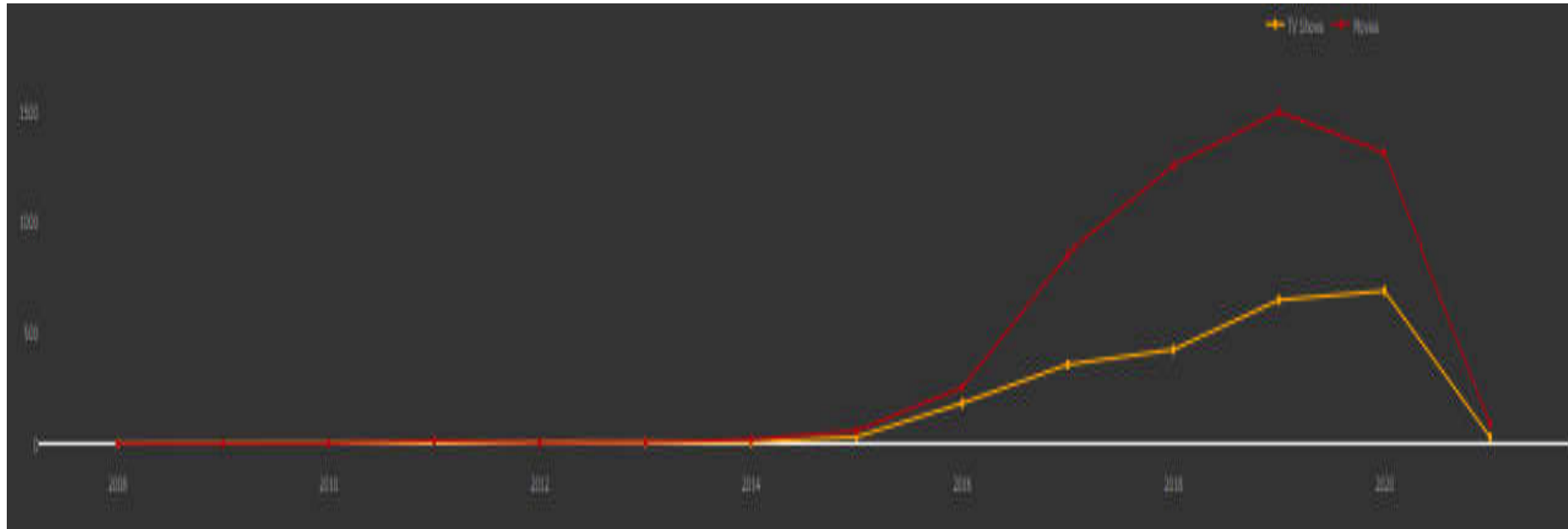
As a Director Alastair Fothergill has worked in most of the TV-Shows and Jan Suter has worked in most of the movies.

Top genres on Netflix



International Movies, dramas and Comedies are the top 3 genres of content on Netflix

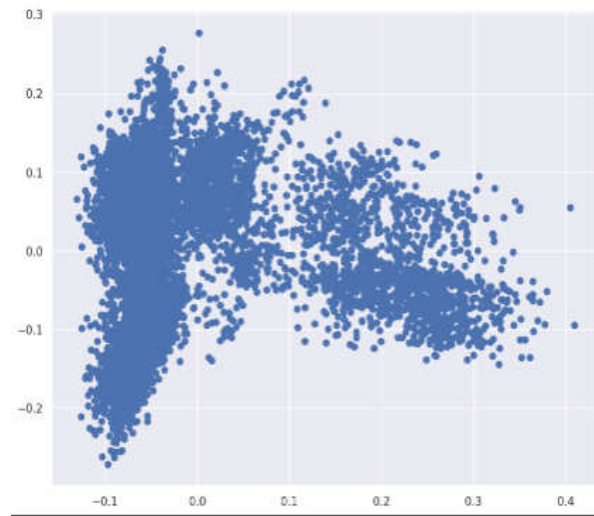
Tv Show and Movies impact over the Year



**Since 2015 there has been a surge in production of TV shows and movies on Netflix.
Post 2019 we can see the graph has dropped due to the pandemic.**

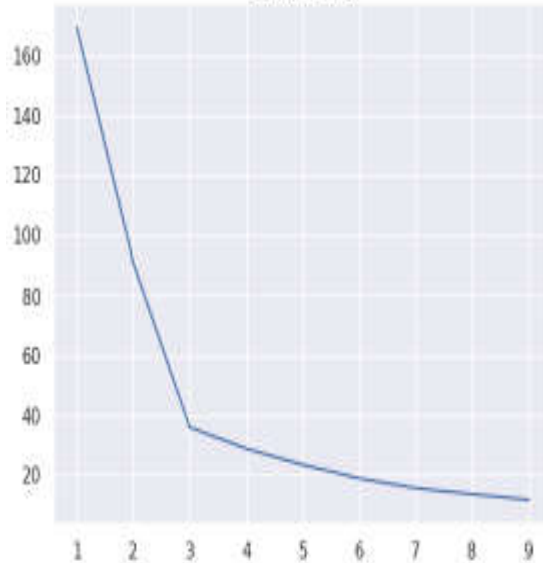
Feature engineering, text cleaning and NLP

- **Step 1** — Combined all important text features in one column.
- **Step 2**- Converted string into lower case and then removed blank spaces, punctuation and stop words followed by stemming.
- **Step 3**- TF-IDF Vectorizer.
- **Step 4**- Applied PCA to reduce dimensions.
- **Step 5**- After doing all the above steps , I visualized the data through the scatter plot.

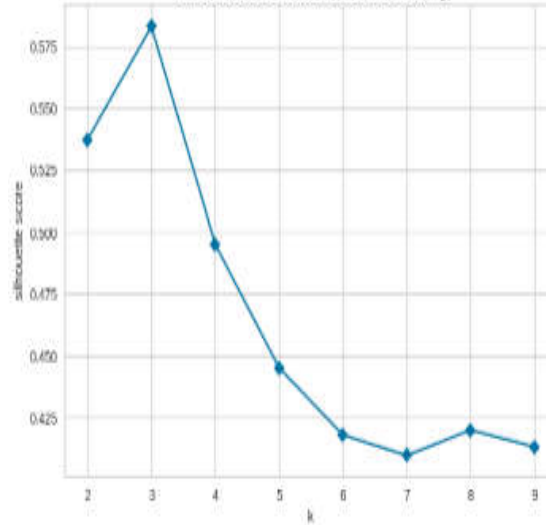


Determine optimal number of clusters

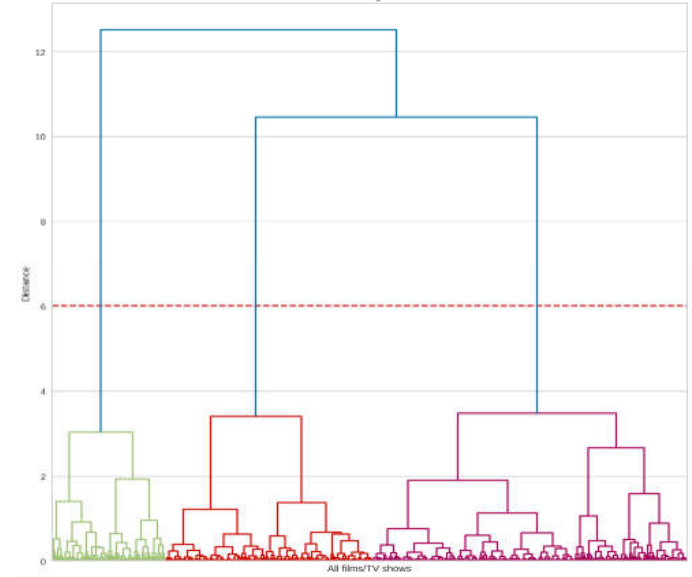
Elbow Curve



Silhouette Score Elbow for KMeans Clustering



Dendrogram



I found optimal number of clusters to be 3 using Elbow Curve, Silhouette score and Dendrogram.

Implementation of ML Models

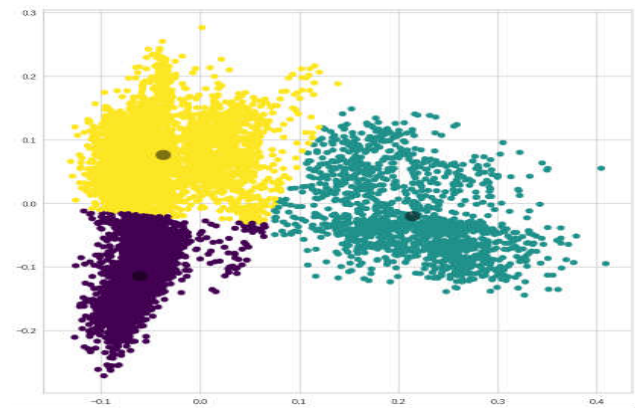
K-Means Clustering

Definition:

K-means clustering is a simple unsupervised learning algorithm that is used to solve clustering problems. It follows a simple procedure of classifying a given data set into a number of clusters, which is fixed beforehand. The clusters are then positioned as points and all observations or data points are associated with the nearest cluster, computed, adjusted and then the process starts over using the new adjustments until a desired result is reached.

Approach:

- Number of clusters are 3.
- After implementing K-MeansClustering, I visualized the clusters .
- I segregated these clusters in variables cluster0, cluster1, cluster2.



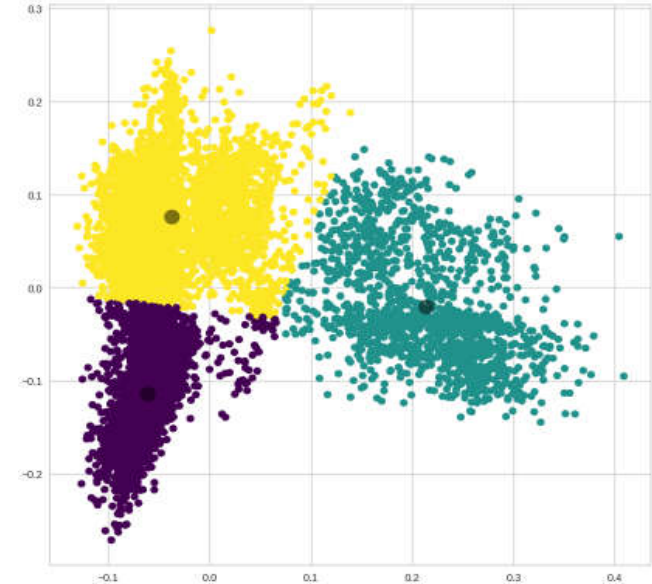
•Agglomerative clustering

Definition:

Agglomerative clustering is the most common type of hierarchical clustering, and it is used to group objects into clusters based on their similarity. The algorithm starts by treating each object as a singleton cluster. Then, one by one, pairs of clusters are merged until all clusters have been merged into one large cluster containing all objects. The dendrogram that results is a tree-based representation of the objects.

Approach:

- Number of clusters are 3.
- After implementing Agglomerative Clustering, I visualizes the clusters .
- I segregated these clusters in variables `agglo_cluser0`, `agglo_cluser1`, `agglo_cluser2`.



Cluster 0 is for those audience who loves to watch Family and Kids related content.

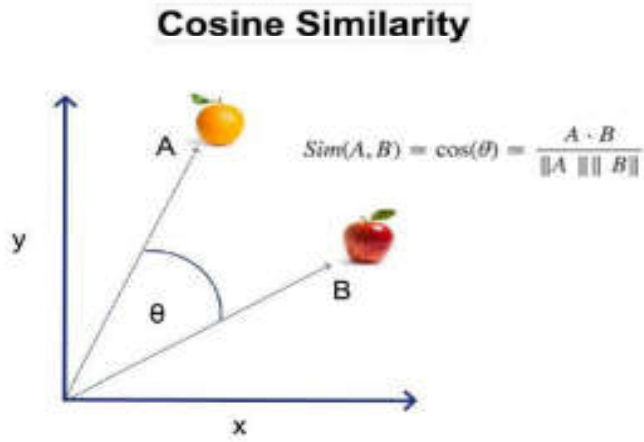


Cluster 1 is for those audience who loves to watch TV shows.

Cluster 2 is for those audiences who loves to watch international content.

Recommendations

We obtained recommendations for Movies and TV- Shows using Cosine similarity.



```
# content recommending  
recommend("Breaking Bad")
```

	Recommend programme	similarity(0-1)
0	Better Call Saul	0.29
1	Dare Me	0.26
2	Hormones	0.24
3	Tiny Pretty Things	0.22
4	The Assassination of Gianni Versace	0.21
5	The Bad Kids	0.21
6	Kakegurui	0.20
7	MINDHUNTER	0.20
8	Ozark	0.20
9	The Blacklist	0.20

Conclusion

- 1: Netflix has around 69% movies and 31% TV shows.
- 2: USA ,India and The United kingdom are top three producer countries on Netflix
- 3:Takahiro Sukurai has worked in most TV-Shows and Anupam Kher has worked in most movies.
- 4: As a Director Alastair Fothergill has worked in most of the TV-Shows and Jan Suter has worked in most of the movies.
- 5: International Movies, Dramas and Comedies are the top three genres on Netflix.
- 6: Netflix added most no of content in year 2018,2019,2020.
- 7: After the year 2019 covid came that badly affected Netflix for producing content.
- 8: cluster 0 and aggro_cluster0 is for those audience who loves to watch family & Kids related content.
- 9: Cluster 1 and aggro_cluster1 is for those audience who loves to watch TV_ shows.
- 10: Cluster 2 and aggro_cluster2 is for those audience who loves international content.
- 11: At last we make a recommendation system ,that can help audience to find there favorite content.