Capstone Project - 4

Netflix Movies And TV Shows Clustering

By-Dany Chitturi

Table of Contents

1. Defining the problem statement
2. Exploratory Data Analysis
3. Hypothesis Testing
4. Feature Engineering
5. Applying Clustering Models
6. Model Evaluation
7. Conclusion

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, I have done

1. Exploratory Data Analysis

2. Understanding what type of content is available in different countries

3. Is Netflix increasingly focusing on TV rather than movies in recent years.

4. Clustering similar content by matching text-based features

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: The 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

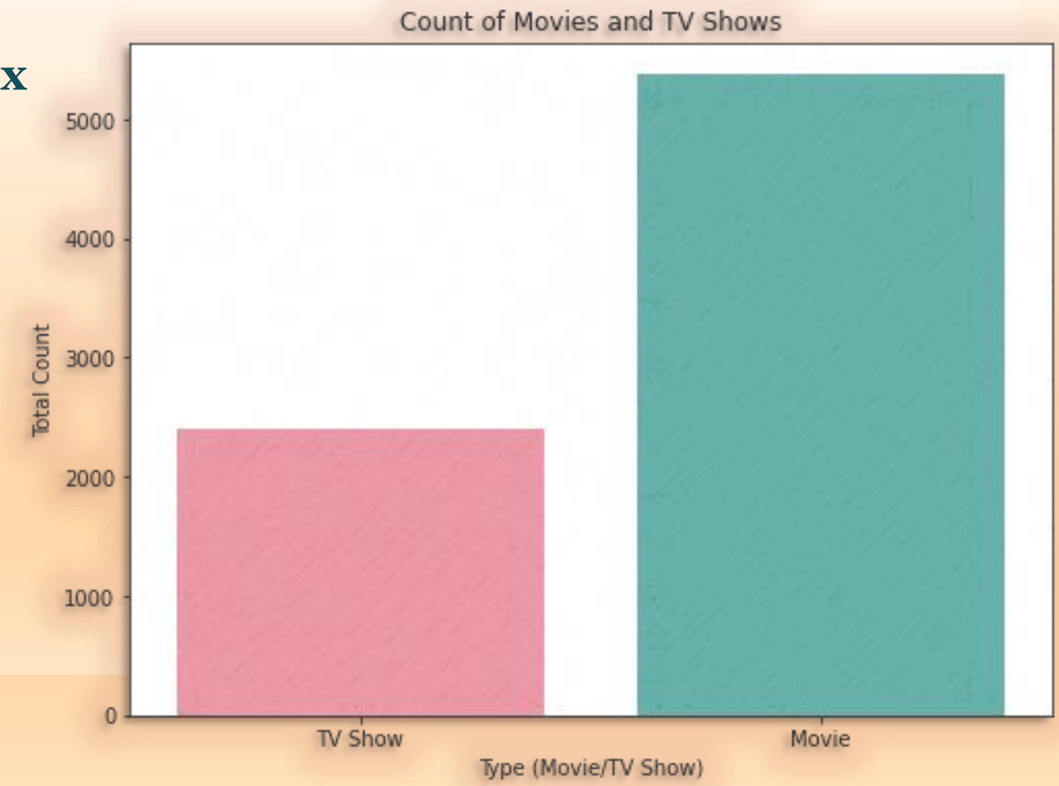
Duration: Total Duration - in minutes or number of seasons

Listed_in: Genre

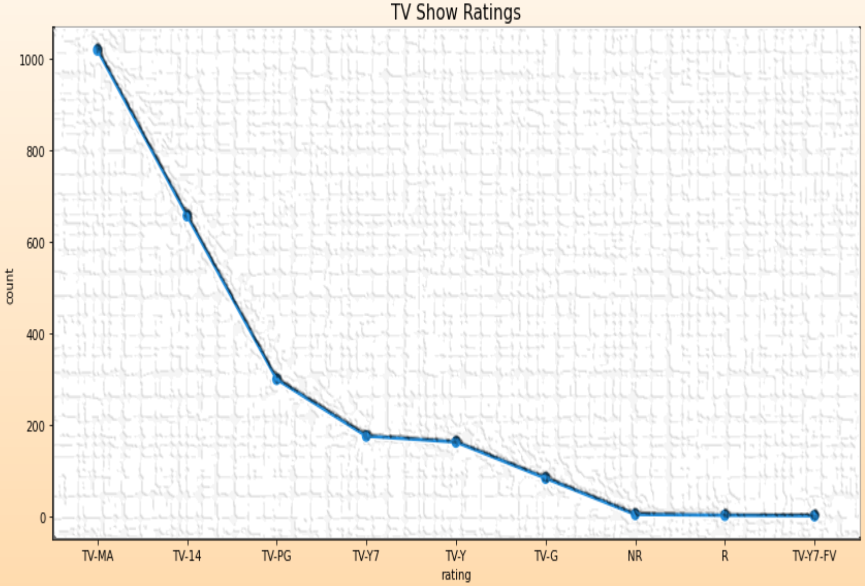
Exploratory Data Analysis

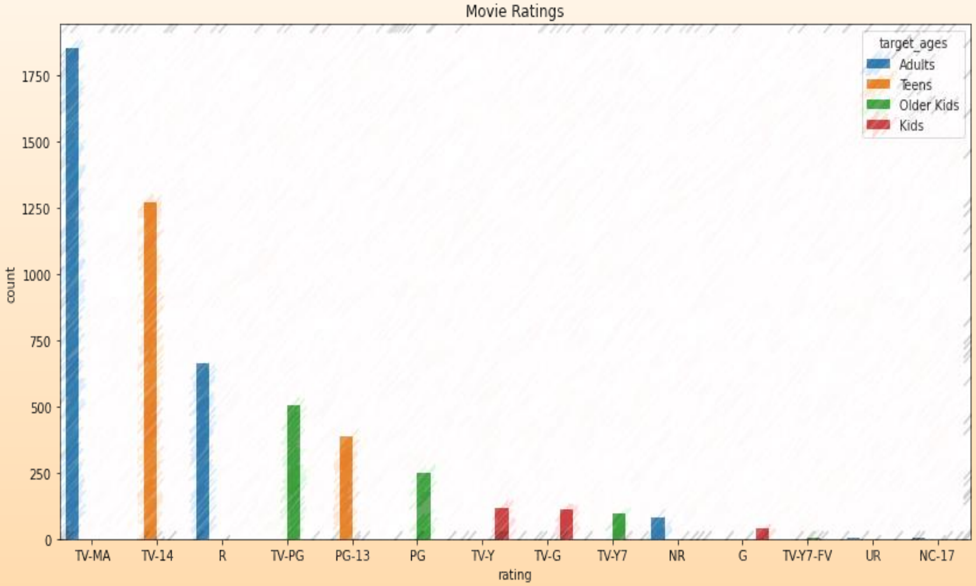
Type of content available on Netflix

- There are more movies on Netflix than TV shows.
- Netflix has 5372 movies, which is more than double the number of TV shows.



Exploratory Data Analysis





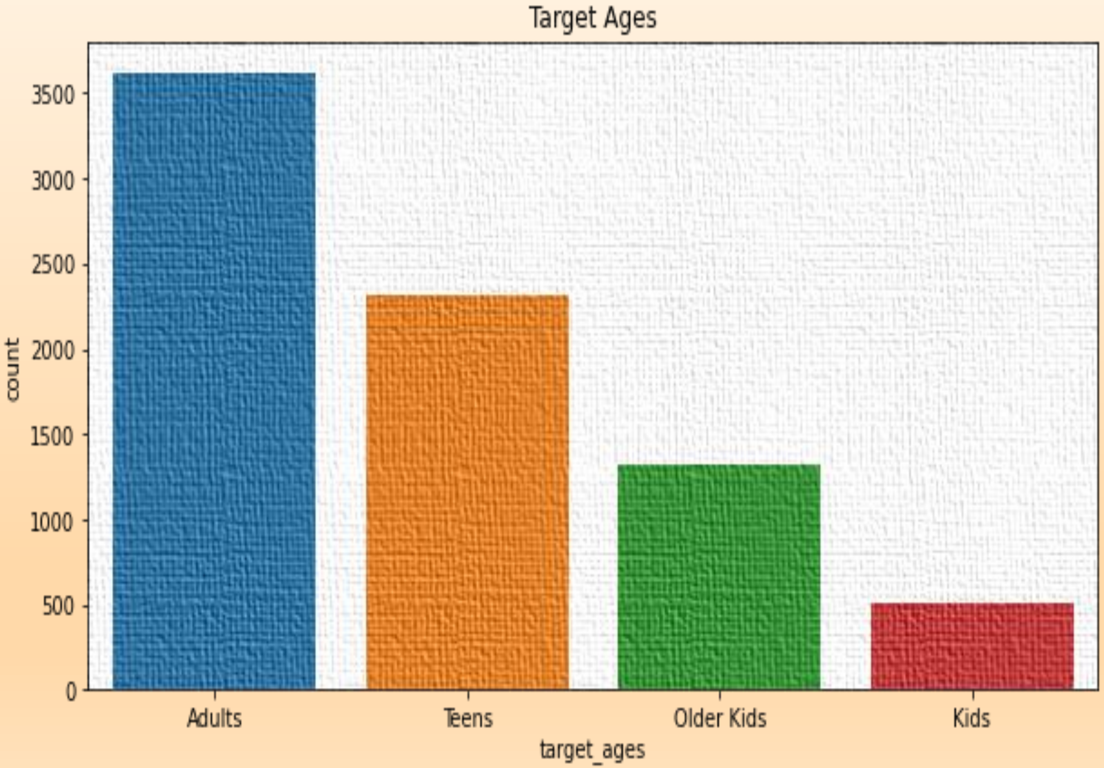
- Most TV shows and Movies are produced for mature audiences who are aged above 18

- Least number of TV shows are produced for kids above 7 years of age and the least number of Movies are rated NC-17 for audiences above 17 years of age.

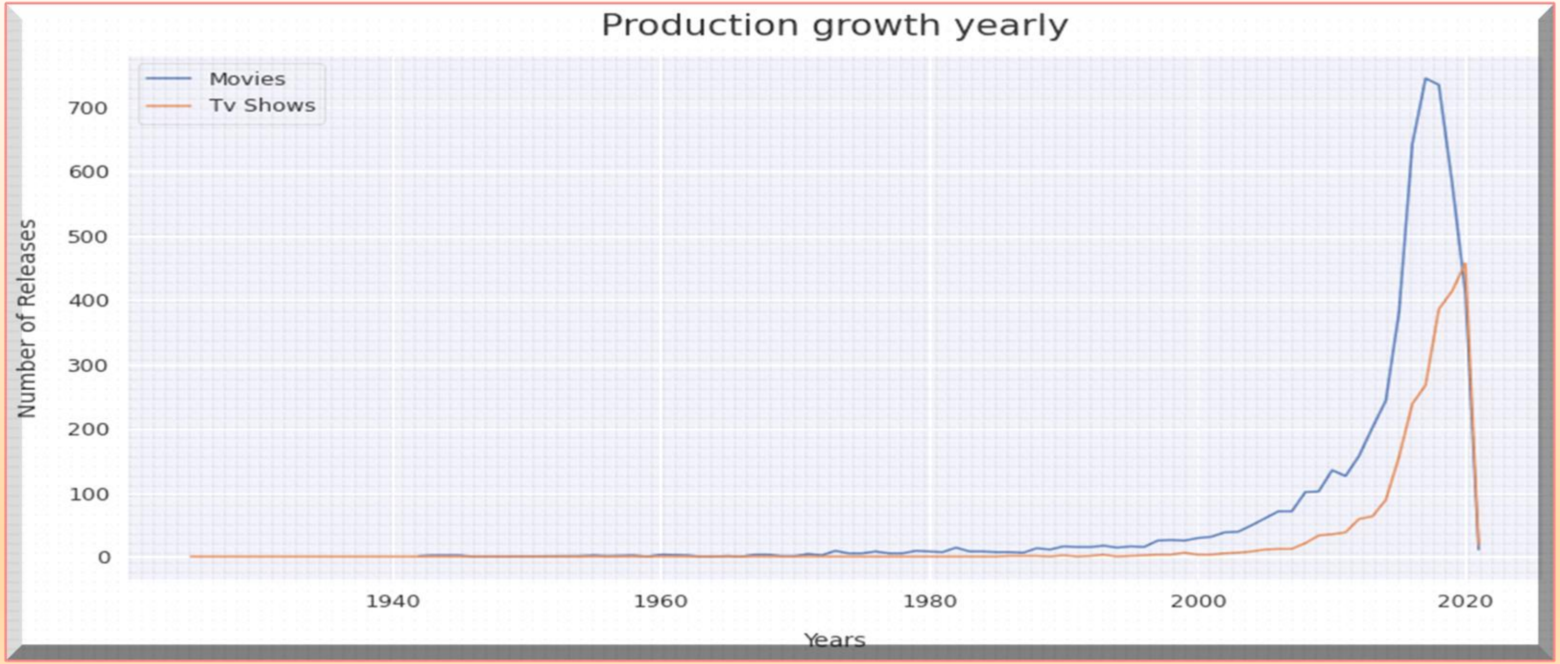
Exploratory Data Analysis

- Around 50% of shows on Netflix are produced for an adult audience, Followed by Teens, older kids, and kids

- Netflix has the least number of shows that are specifically produced for kids than other age groups.



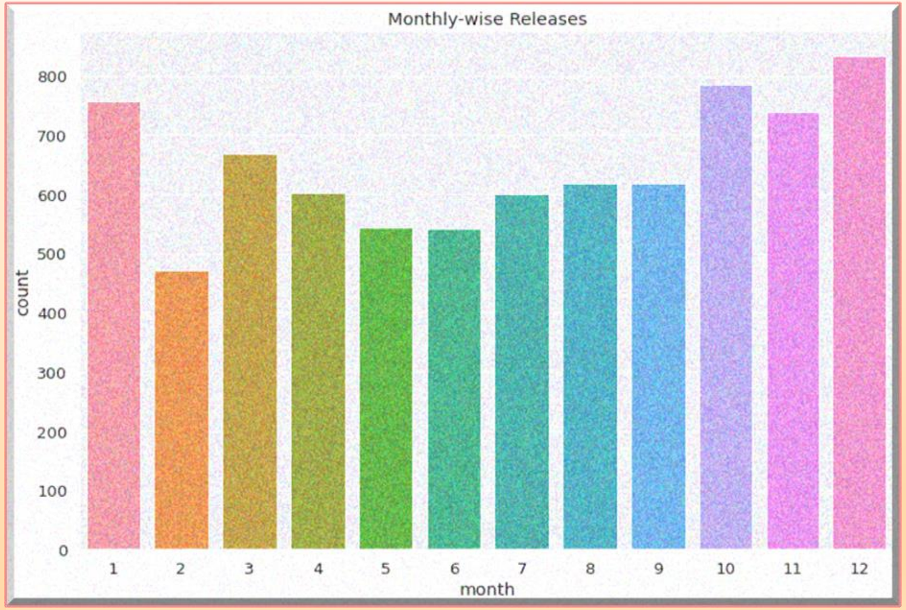
Exploratory Data Analysis

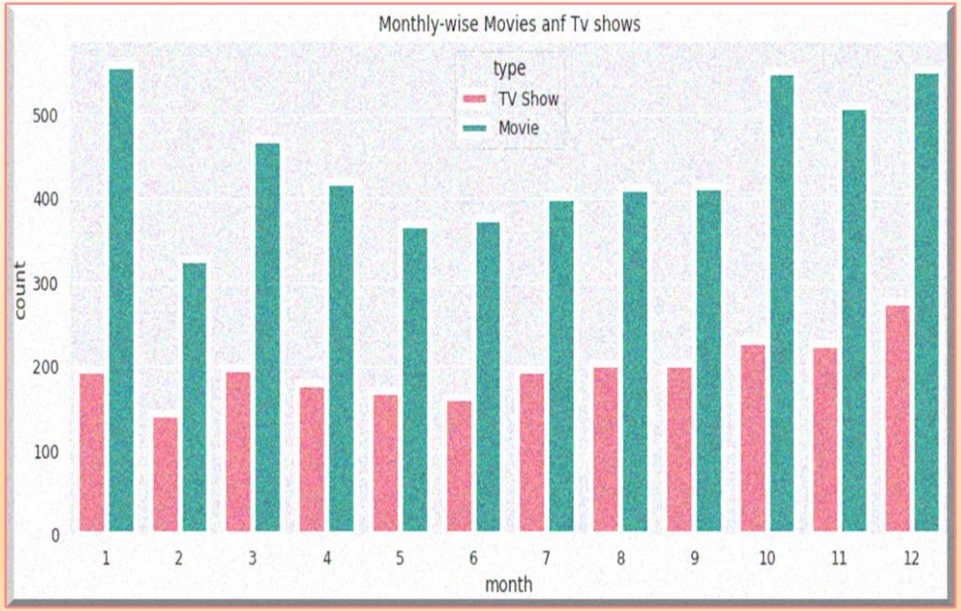


•The number of movies on Netflix is growing significantly faster than the number of TV shows.

• It appears that Netflix has focused more on increasing Movie content than TV Shows. The number of movies has grown much more dramatically than TV shows from the year 2015 to 2020

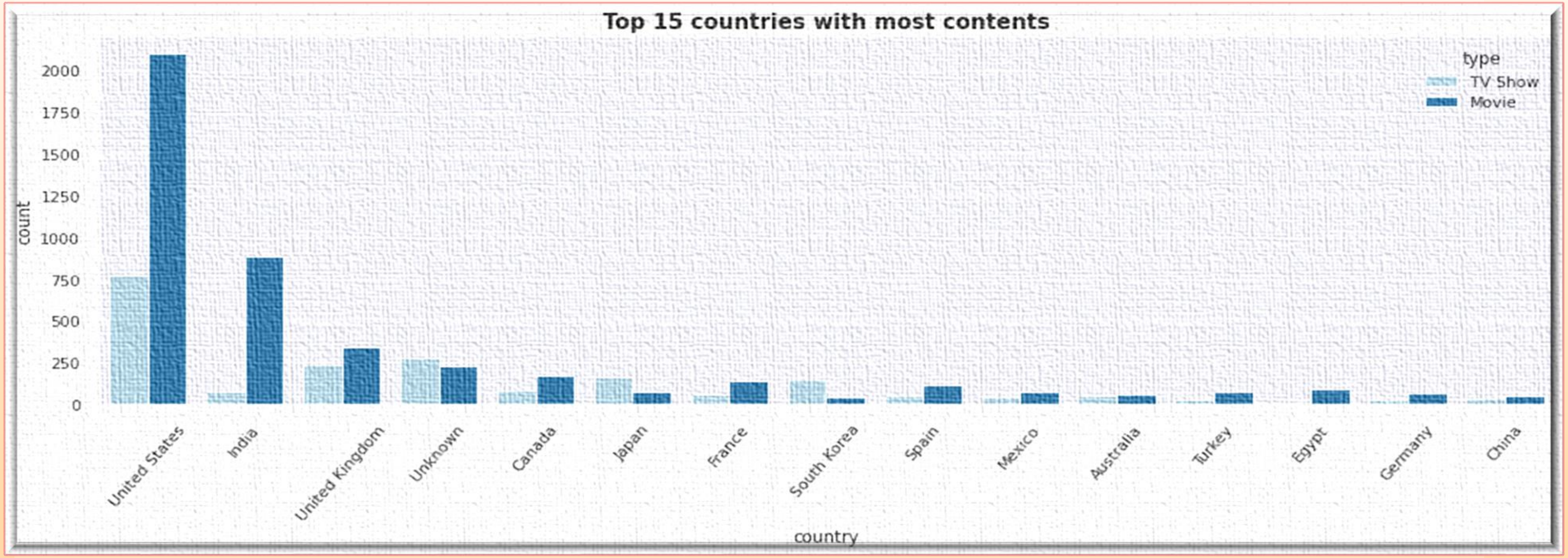
Exploratory Data Analysis





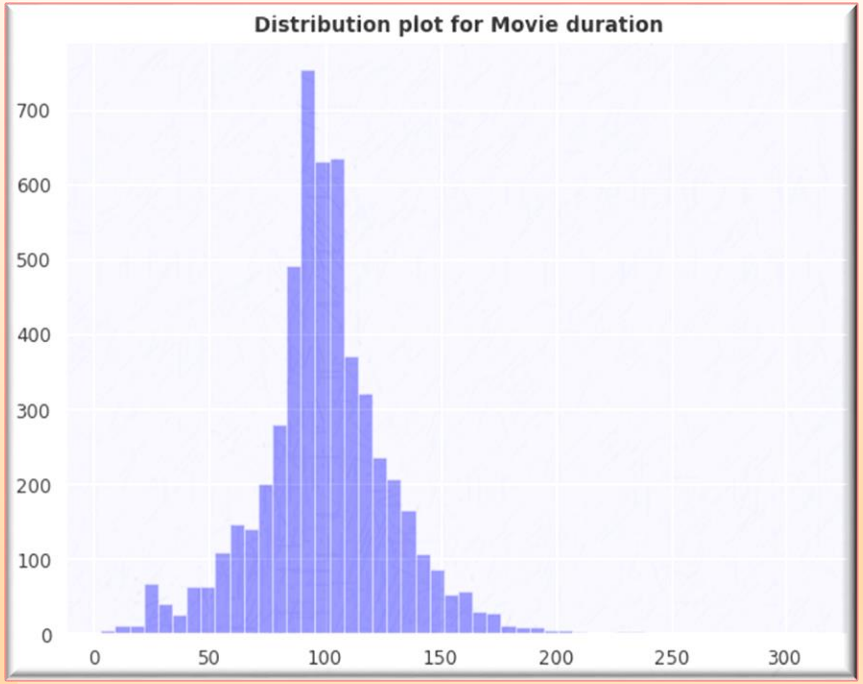
- Most of the content(Movies/Tv shows) is added to Netflix from October to January. We can say that December is the holiday season and it also has Christmas, so in that month most of the content got uploaded.

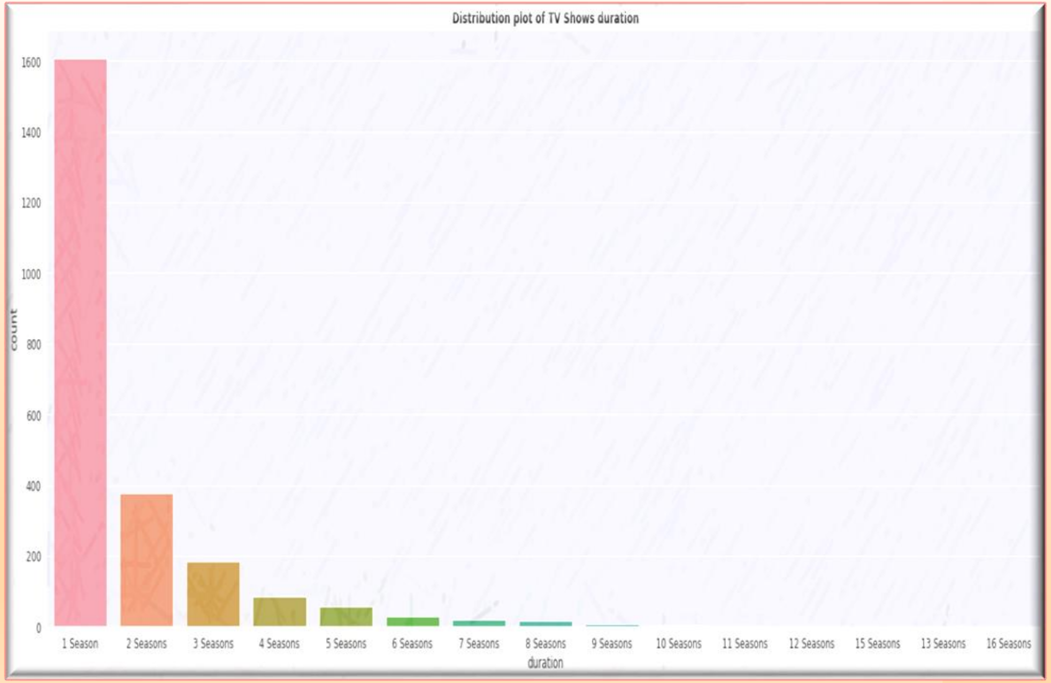
Exploratory Data Analysis



- The highest number of movies / TV shows were based out of the US, followed by India and UK.
- India has the most number of movies when compared to TV shows

Exploratory Data Analysis

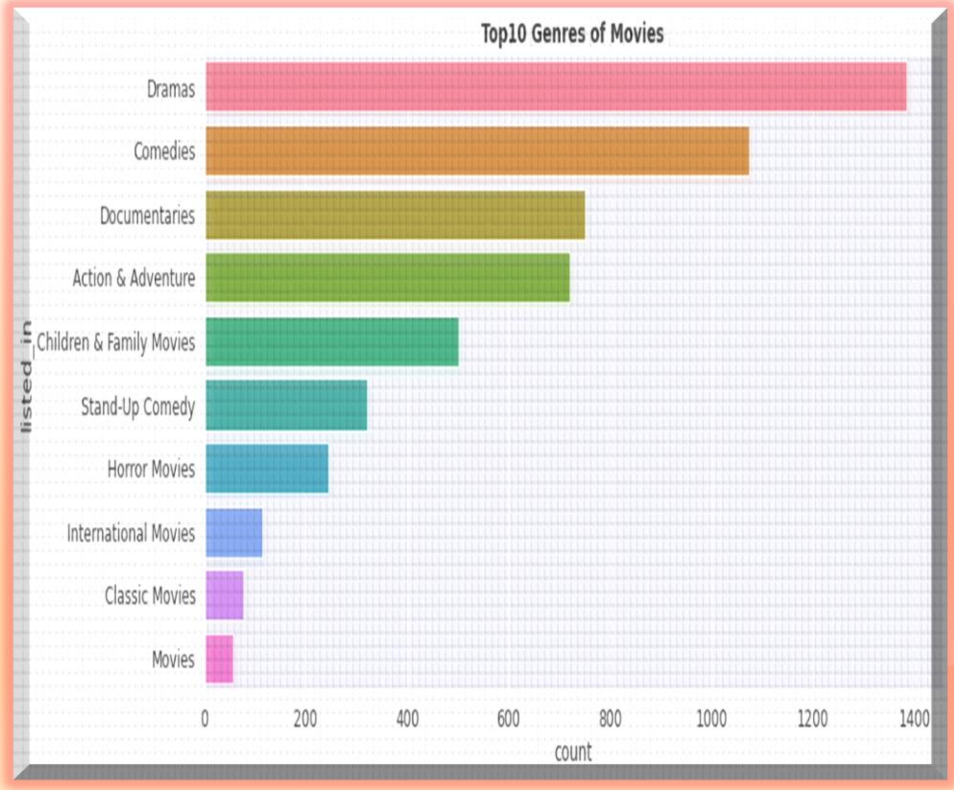


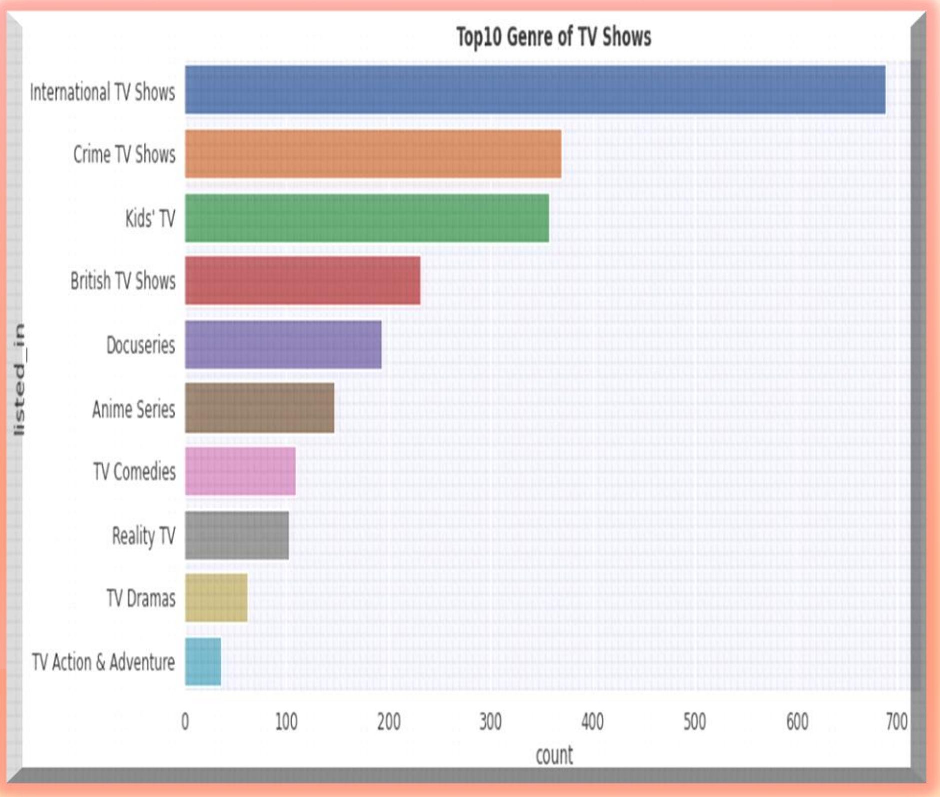


- Most movies have a duration between 70 to 120 minutes

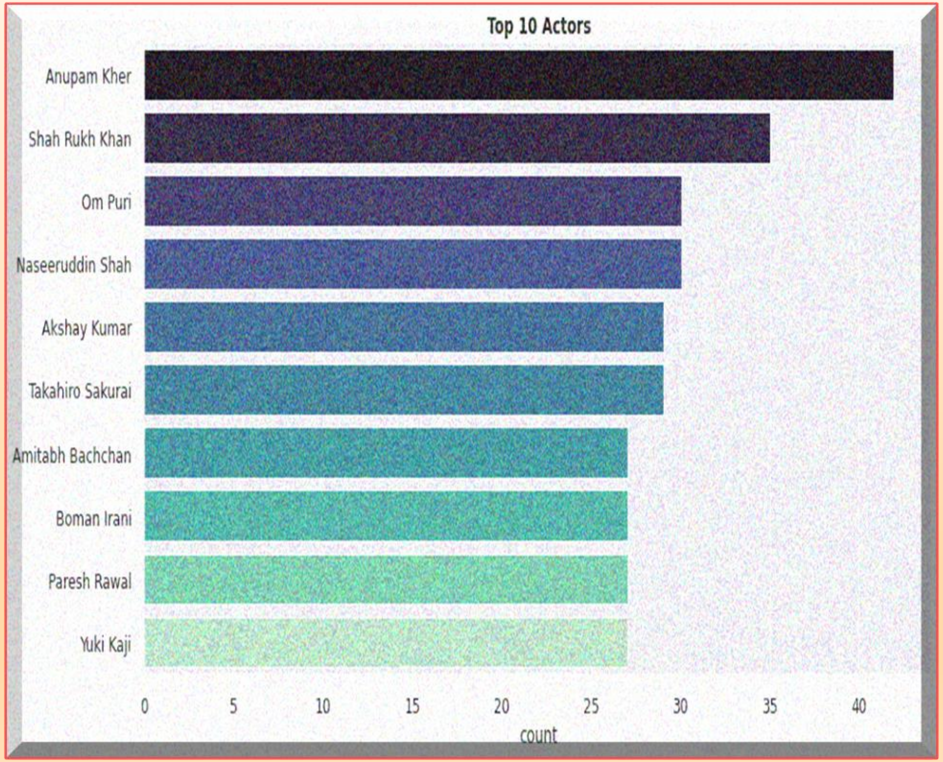
- Highest number of TV shows consist of single season

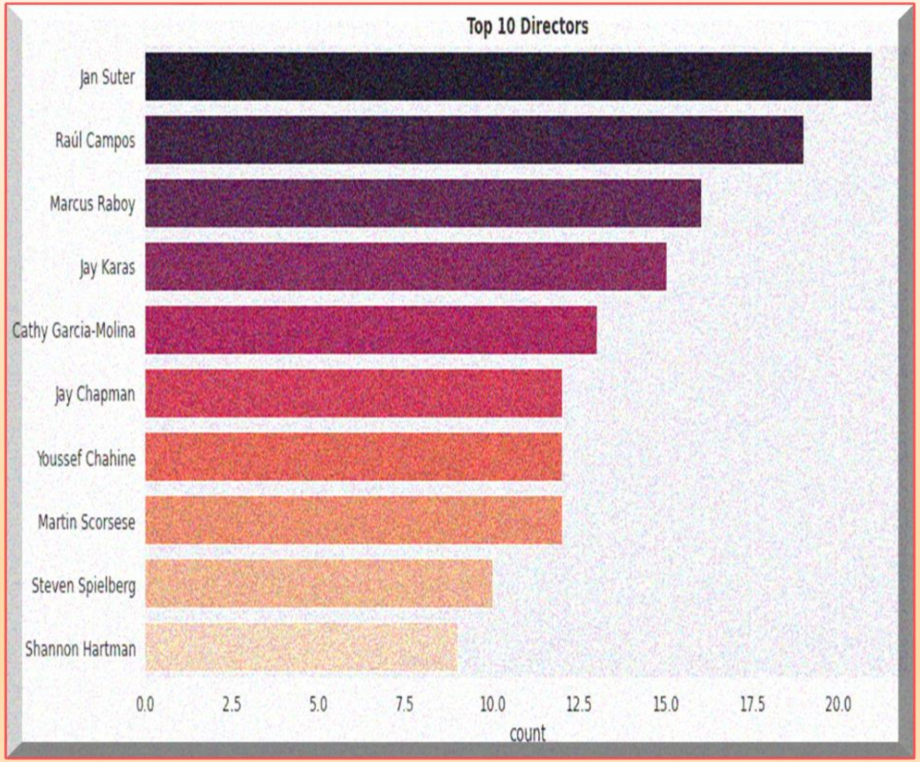
Exploratory Data Analysis





Exploratory Data Analysis



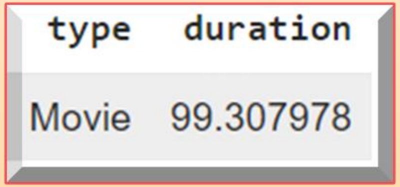


Hypothesis Testing

- ❑ Hypothesis testing in statistics refers to analyzing an assumption about a population parameter.

H0: The average duration of movies on Netflix is equal to 90 minutes

H1: The average duration of movies on Netflix is not equal to 90 minutes



type	duration
Movie	99.307978

T-value= 23.92

95% Confidence Interval= (8.54, 10.07)

P-Value=2.82e-120

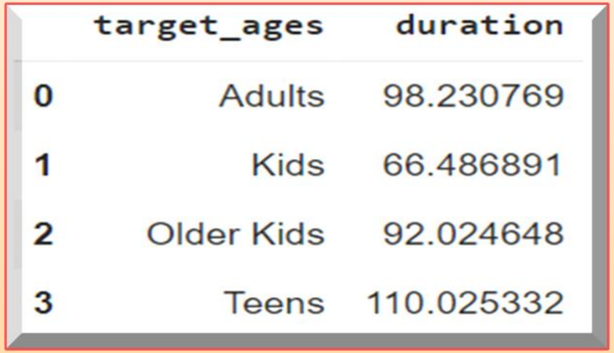
Considering a significance level of α = 0.05, we would reject the null hypothesis of our hypothesis test because this p-value is less than 0.05.

As a result, we concluded that the average duration of movies is not 90 minutes

Hypothesis Testing

2. H0: The duration of movies rated for Teens is less than or equal to Adults.

H1: The duration of movies rated for Teens is greater than for Adults.



	target_ages	duration
0	Adults	98.230769
1	Kids	66.486891
2	Older Kids	92.024648
3	Teens	110.025332

T-value= 14.10
95% Confidence Interval= (10.42, 13.18)
P-Value=1.71e-44

Considering a significance level of α = 0.05, we would reject the null hypothesis of our hypothesis test because the p-value of this right-tailed test is less than 0.05.

As a result, we concluded that the duration for movies rated for teens is greater than Adults

Feature Engineering

- **Clusters are built based on these attributes:** Director, Cast, Country, Rating Listed in (genres), and Description

- **Steps involved in data pre-processing:**

➢ Removing non-ASCII characters

➢ Removing stopwords and converting to lowercase

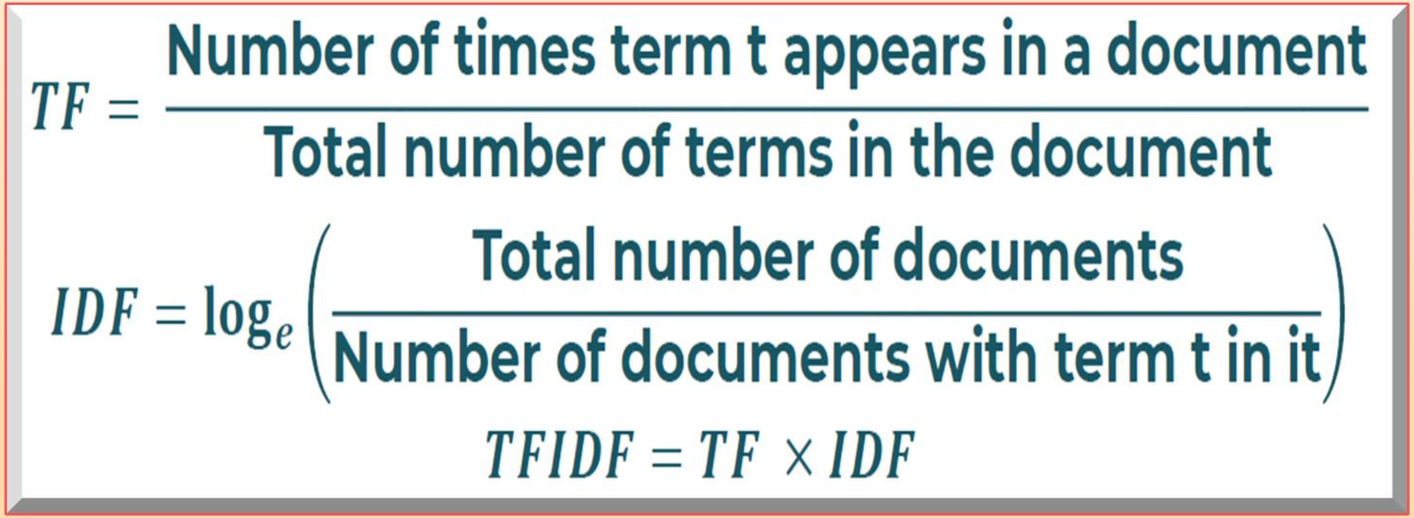
➢ Removing punctuation marks

➢ Lemmatization, tokenization, and text vectorization

➢ Dimensionality reduction using PCA

Feature Engineering

- **TFIDF** (Term Frequency Inverse Document Frequency) vectorizer is used to vectorize the corpus.



$$TF = \frac{\text{Number of times term t appears in a document}}{\text{Total number of terms in the document}}$$

$$IDF = \log_e\left(\frac{\text{Total number of documents}}{\text{Number of documents with term t in it}}\right)$$

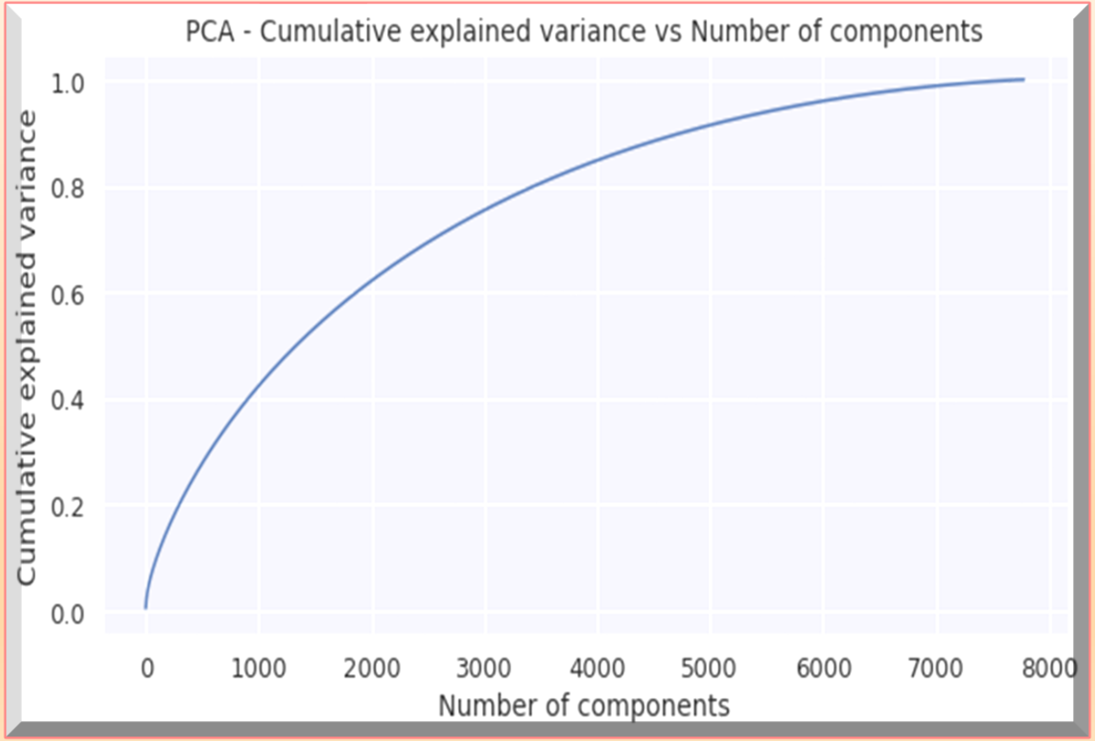
$$TFIDF = TF \times IDF$$

- Maximum number of features was taken as **20000**

Dimensionality Reduction(PCA)

•**100%** of the variance in data is explained by about **~7500** components.

•To reduce dimensionality, only the top **4000** components were taken, which will still be able to capture more than **80%** of the variance in the data.



K Means Clustering

The working of the K-Means algorithm is explained in the below steps:

Step-1: Select the number K to decide the number of clusters.

Step-2: Select random K points or centroids. (It can be other from the input dataset).

Step-3: Assign each data point to their closest centroid, which will form the predefined K clusters.

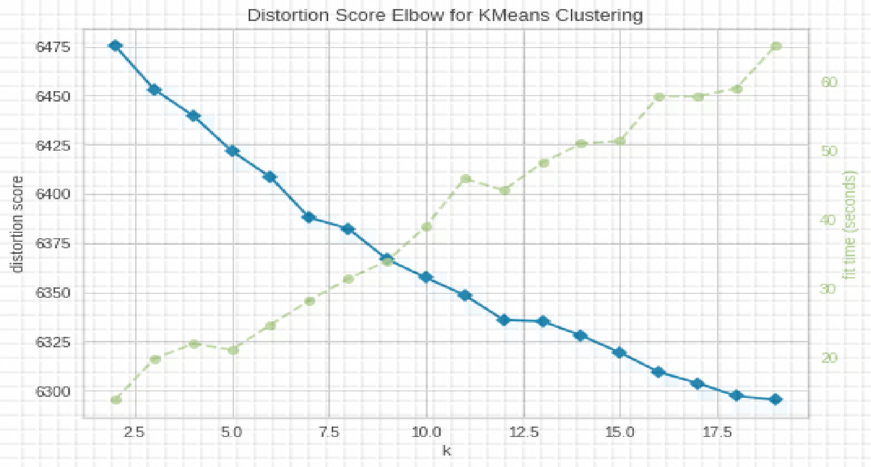
Step-4: Calculate the variance and place a new centroid of each cluster.

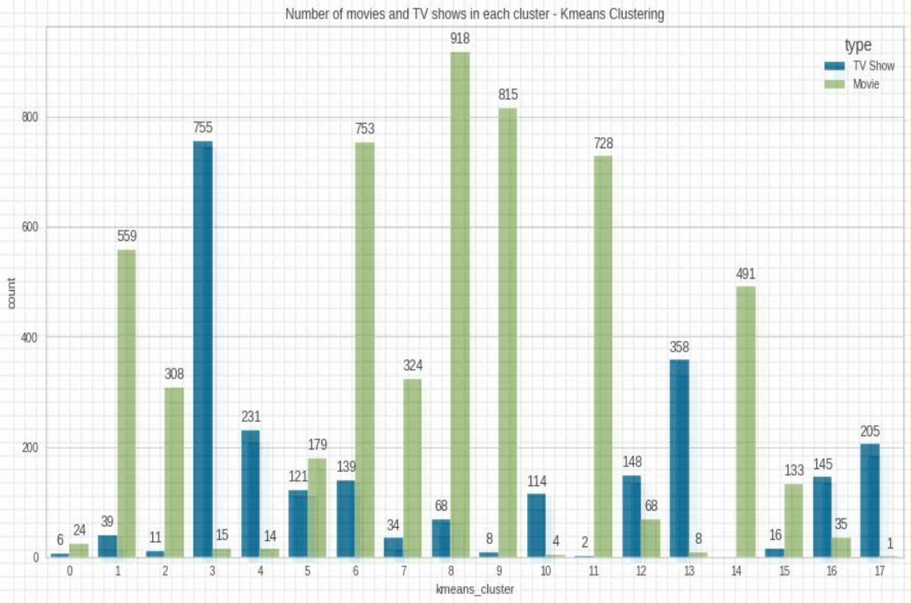
Step-5: Repeat the third steps, which means reassign each datapoint to the new closest centroid of each cluster.

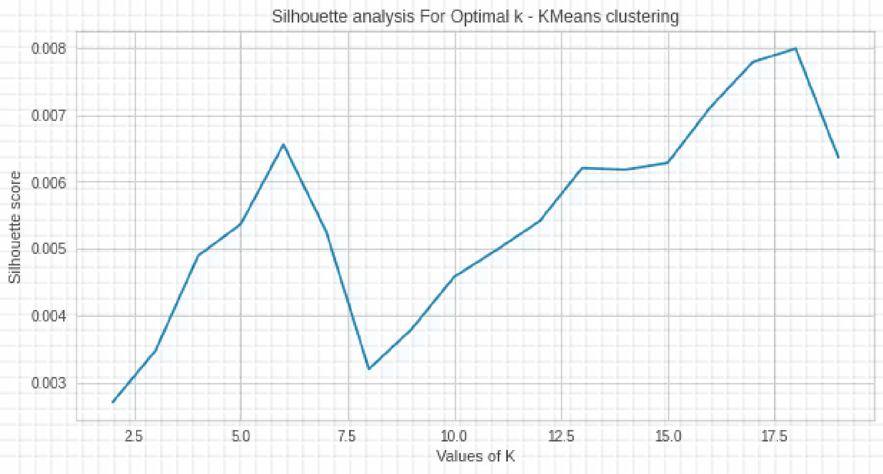
Step-6: If any reassignment occurs, then go to step-4 else go to FINISH.

K Means Clustering

- Distortion: 6296.78
- Silhouette score: 0.0080
- Davies Bouldin’s score:9.484
- Number of clusters: 18

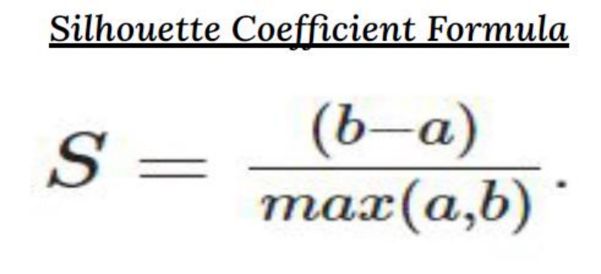






Evaluation Metrics

1. **Silhouette Score:** A metric to evaluate the performance of the clustering algorithm. It uses the compactness of individual clusters(intra-cluster distance) and separation amongst clusters (inter-cluster distance) to measure an overall representative score of how well our clustering algorithm has performed



Silhouette Coefficient Formula

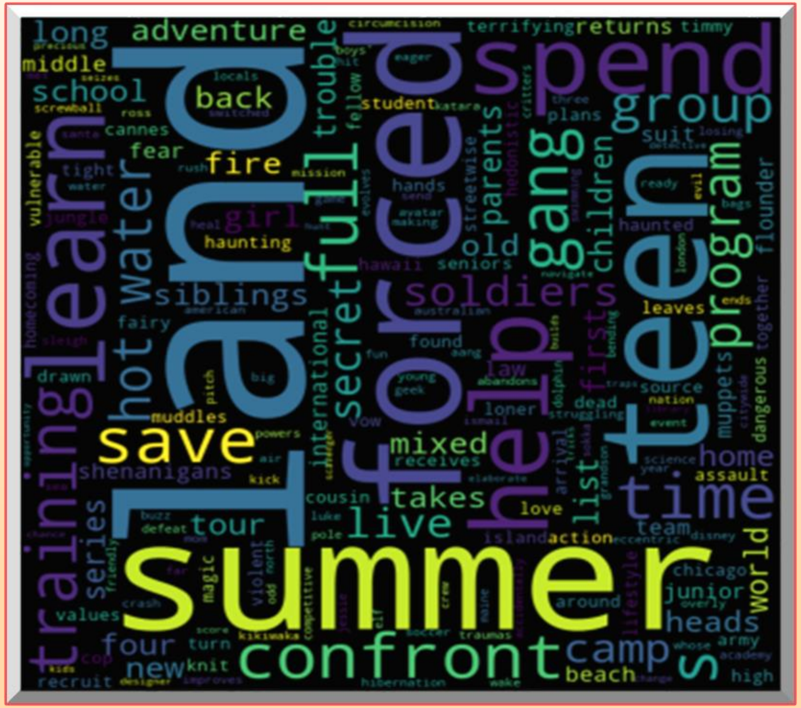
$$S = \frac{(b-a)}{max(a,b)}.$$

- **mean intra-cluster distance(a)**:- Mean distance between the observation and all other data points in the same cluster.
- **mean nearest-cluster distance (b)**:- Mean distance between the observation and all other data points of the next nearest cluster. This distance can also be called a.

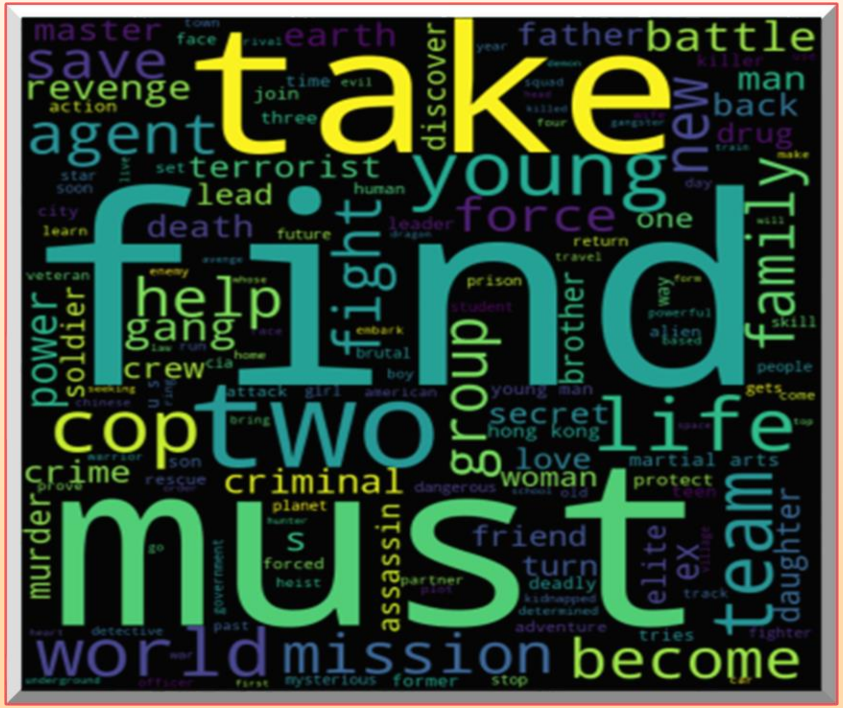
2. **Davies-Bouldin index (DBI**). is most commonly used to evaluate the goodness of split by a K-Means clustering algorithm for a given number of clusters.

- silhouette score would always lie between -1 to 1. ‘1’ represents better clustering
- Silhouette score is 0.0080
- Davies_bouldin_score is 10.44
- so model is performing well

Word Clouds-K Means Clusters

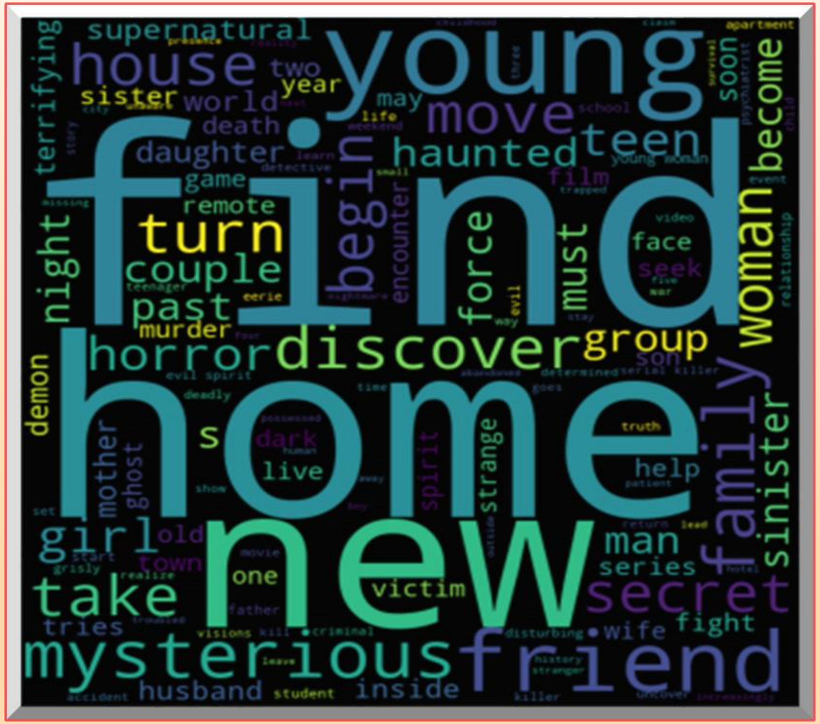


K Means Cluster 0

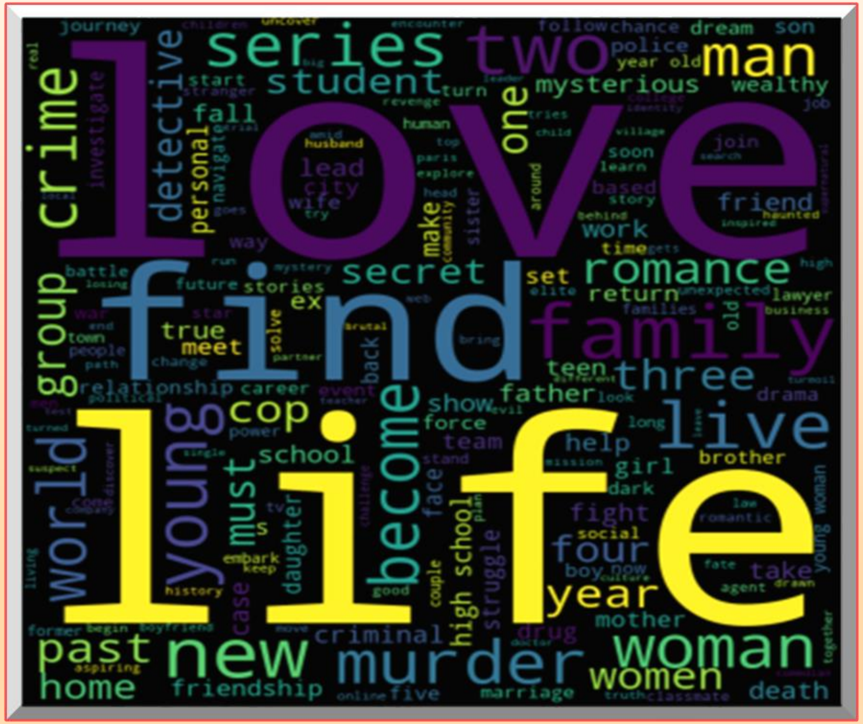


K Means Cluster 1

Word Clouds-K Means Clusters

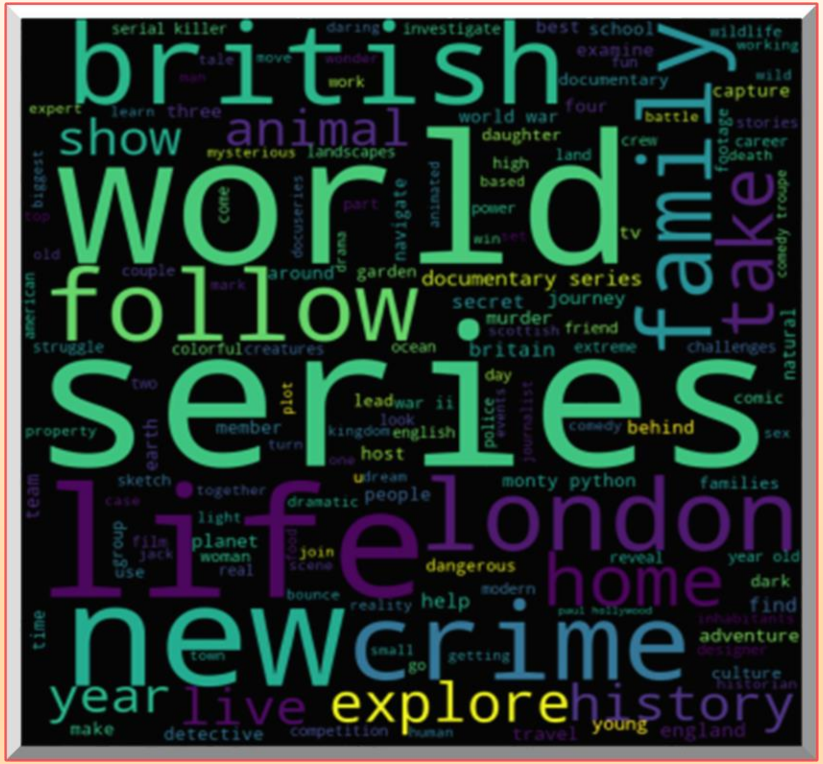


K Means Cluster 2

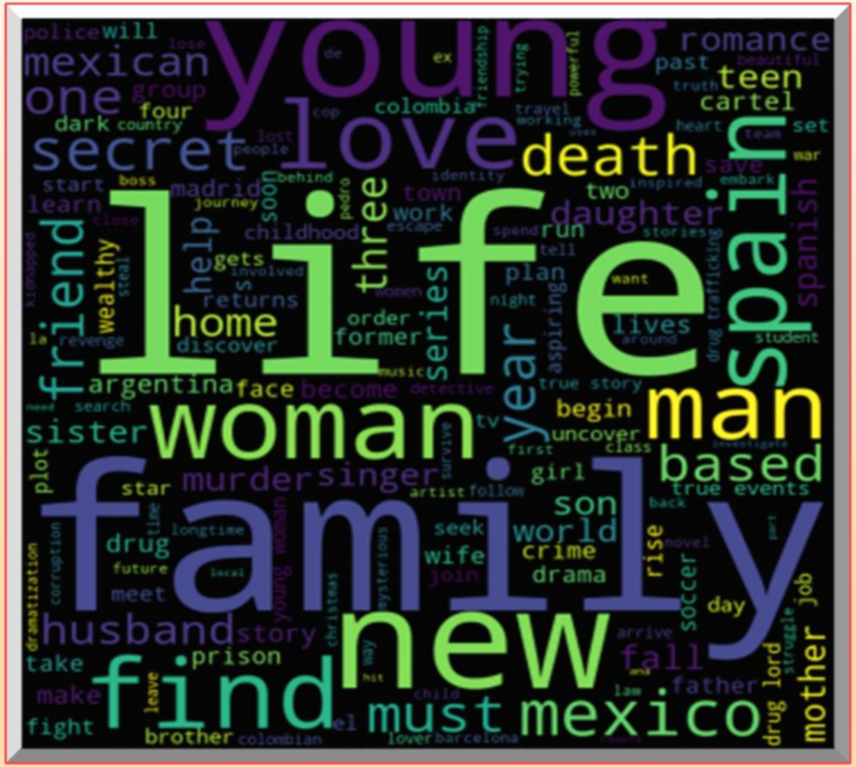


K Means Cluster 3

Word Clouds-K Means Clusters



K Means Cluster 4



K Means Cluster 5

Hierarchical Clustering

The working of the AHC algorithm can be explained using the below steps:

- **Step-1:** Create each data point as a single cluster. Let's say there are N data points, so the number of clusters will also be N.

- **Step-2:** Take two closest data points or clusters and merge them to form one cluster. So, there will now be N-1 clusters.

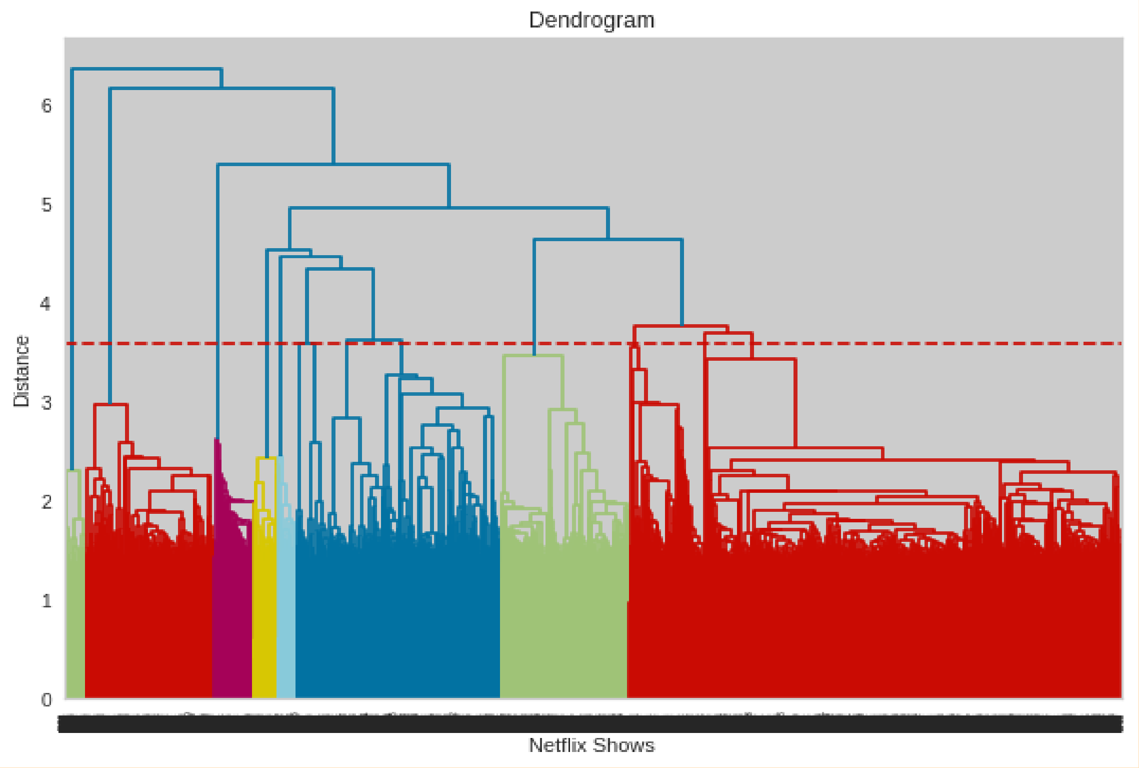
- **Step-3**: Again, take the two closest clusters and merge them together to form one cluster. There will be N-2 clusters.

- **Step-4:** Repeat Step 3 until only one cluster is left.

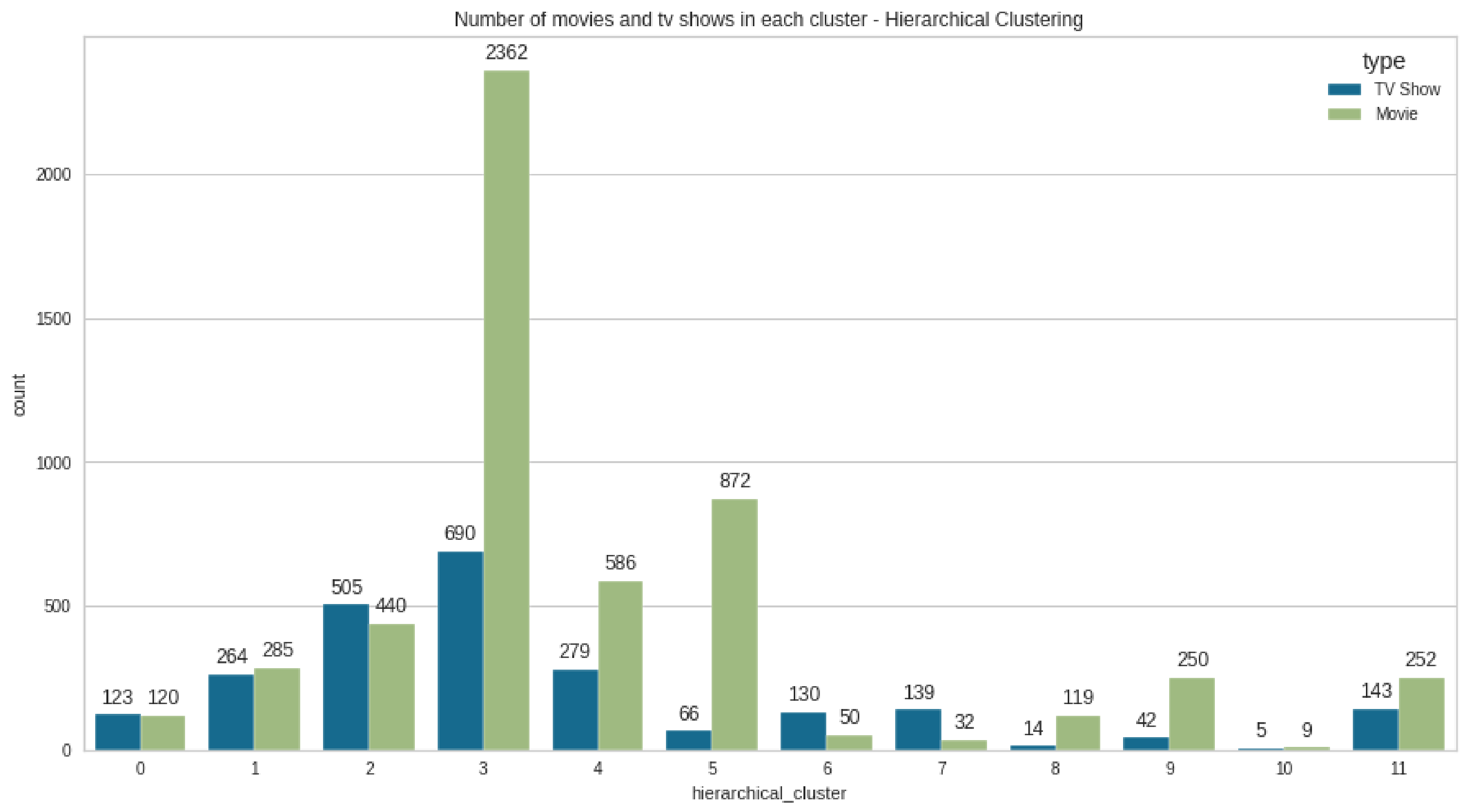
- **Step-5:** Once all the clusters are combined into one big cluster, develop the dendrogram to divide the clusters as per the problem.

Hierarchical Clustering

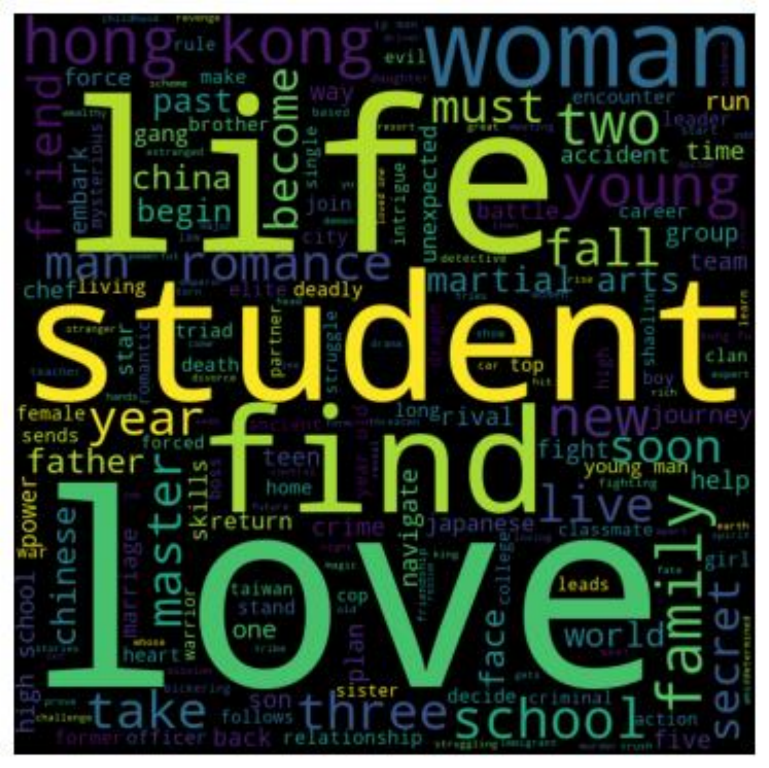
- Agglomerative clustering
- Distance: Euclidean
- Linkage: Ward
- Number of clusters: 12
- Silhouette score: 0.0015
- Davies Bouldin's score:10.44



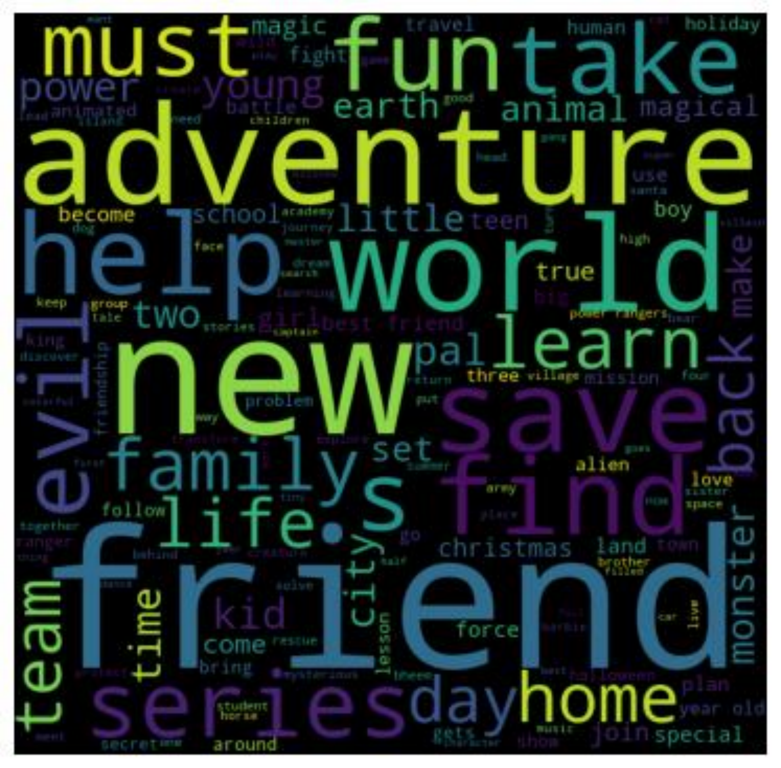
Hierarchical Clustering



Word Clouds: Hierarchical Clusters

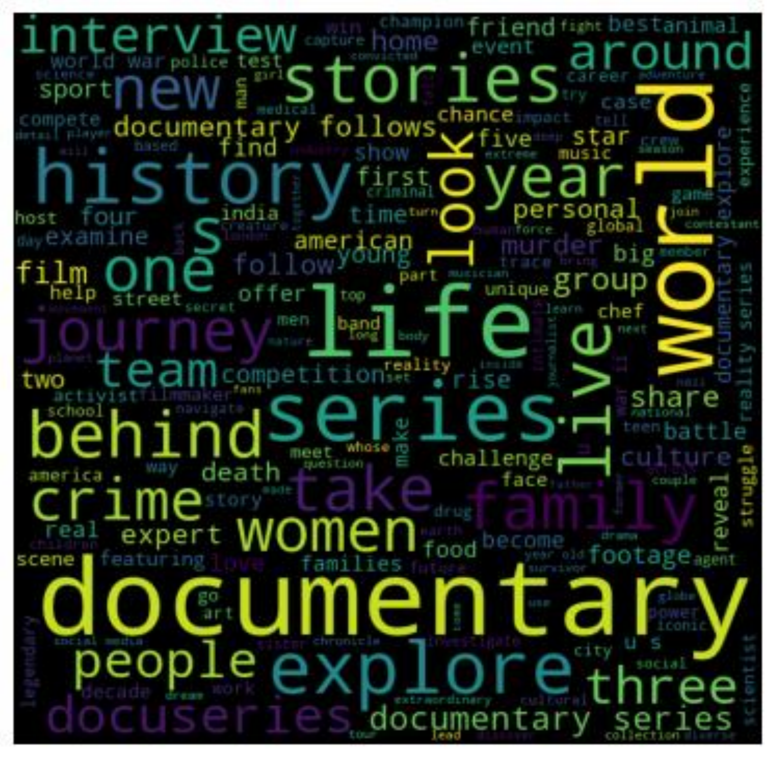


Hierarchical cluster- 0

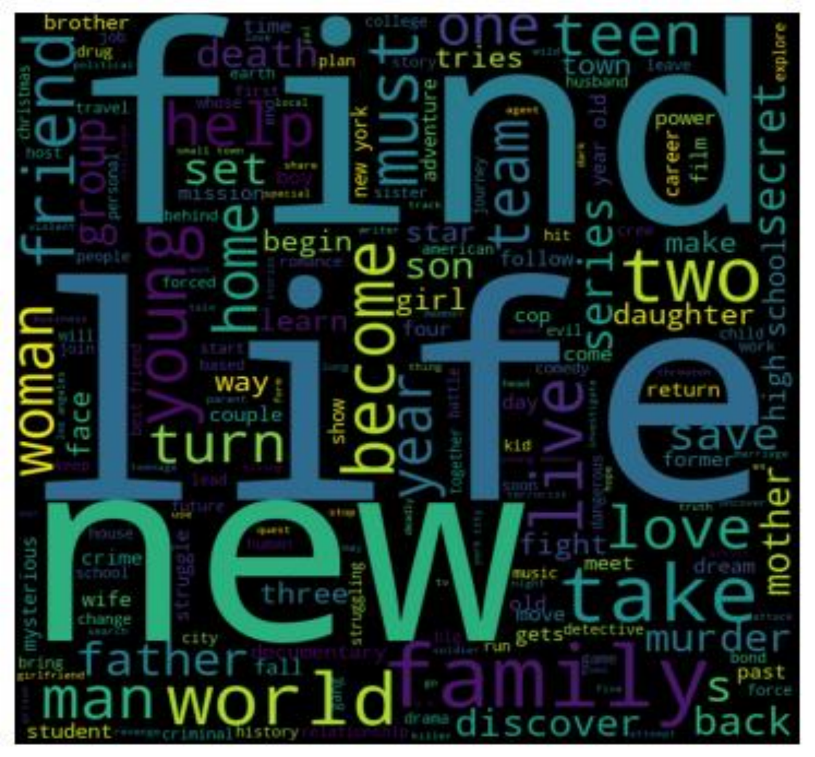


Hierarchical cluster- 1

Word Clouds: Hierarchical Clusters

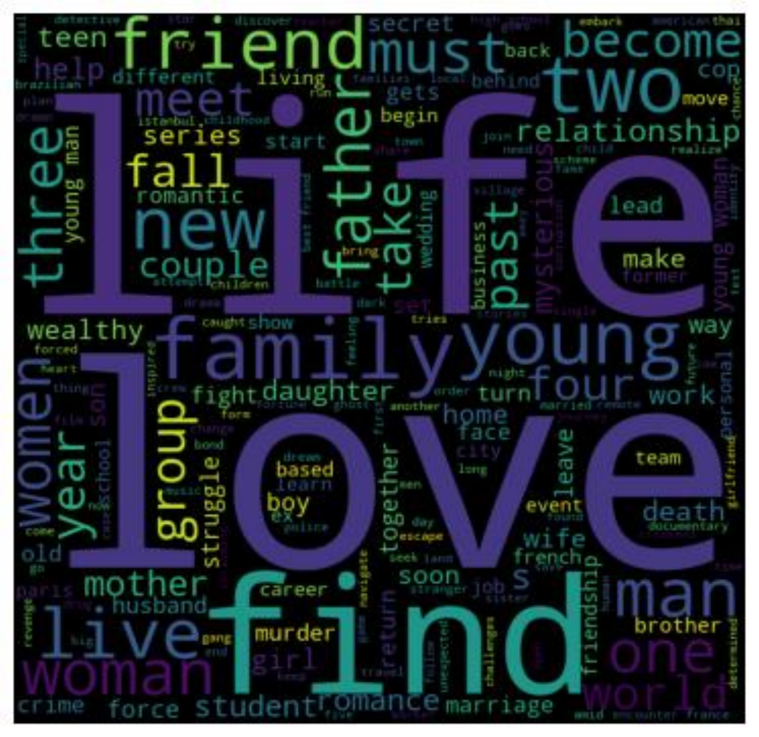


Hierarchical cluster- 2

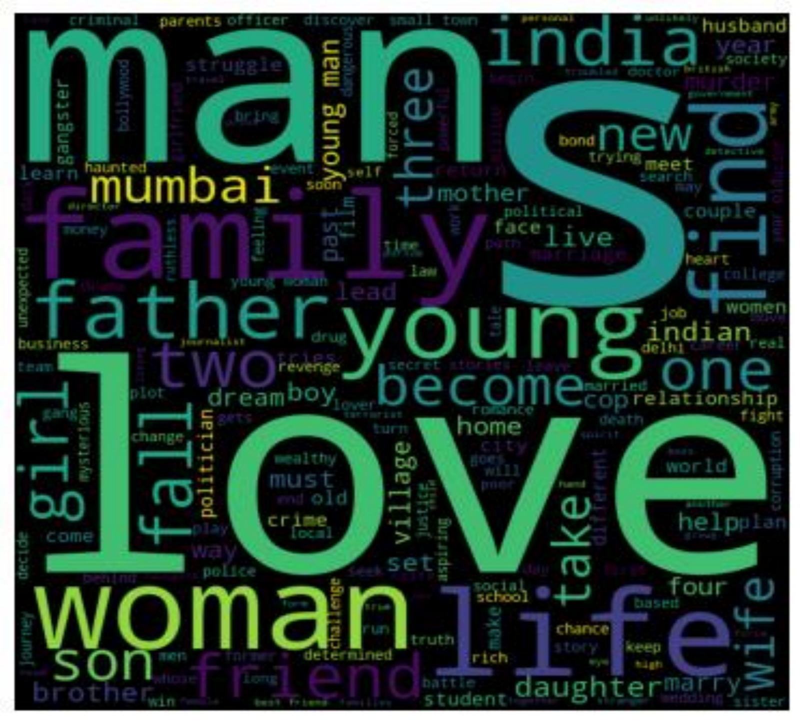


Hierarchical cluster- 3

Word Clouds: Hierarchical Clusters



Hierarchical cluster- 4



Hierarchical cluster- 5

Conclusions

- Netflix has 5372 movies and 2398 TV shows. Around 50% of shows on Netflix are produced for an adult audience, followed by Teens, older kids, and kids. Netflix has the least number of shows for kids than other age groups.

- It was found that Netflix hosts **more movies** than TV shows on its platform, and the total **number of shows added on Netflix is growing exponentially**. Also, most of the shows were produced in the **United States**, followed by **India** and **UK.**

- India has the highest number of movies on Netflix

- Most of the content(Movies/TV shows) is added to Netflix from October to January

- Documentaries are the top genre on Netflix, followed by standup comedy, Drama, and international movies. Crime TV and Kids' TV are the most popular genre for TV shows on Netflix.

- Most movies have a duration between 50 to 150 minutes and the majority of TV shows consist of a single season

- Movies with an NC-17 rating have the highest average duration, and TV-Y-rated movies have the shortest runtime on average.

Conclusions

- The number of movies on Netflix is growing significantly faster than the number of TV shows. We saw a rise in the number of movies and television episodes after 2015 and a significant drop after 2020. It appears that Netflix has focused more on increasing Movie content than TV Shows. The number of movies has grown much more dramatically than TV shows from the year 2015 to 2020

- It was decided to cluster the data based on the attributes: director, cast, country, genre, and description. The values in these attributes were pre-processed, tokenized, and then vectorized using the TFIDF vectorizer.

- Through TFIDF Vectorization, we created a total of 20000 features.

- We used Principal Component Analysis(PCA) to handle the curse of dimensionality. 4000 components were able to
- capture more than 80% of the variance.

- We first built clusters using the k-means clustering algorithm, and the optimal number of clusters came out to be 18. This was obtained through the elbow method and Silhouette score analysis.

- A hierarchical clustering model was built using the Agglomerative clustering algorithm, and the optimal number of clusters came out to be 12. This was obtained after visualizing the dendrogram.

Thank You!