**Unsupervised Learning**

**NETFLIX MOVIES AND TV SHOWS CLUSTERING**

**By-Sonali Kaushal, Somya Hingorani,**

**Nitesh Verma, Harsh Vardhan, Prateek Gupta**

**Data science trainees,**

**AlmaBetter, Bangalore**

**Abstract:**

Netflix is a subscription-based service to watch tv shows and movies online. They want to provide appropriate movies and tv shows according to their subscriber's interests, so they can attract more subscribers and retain their old ones. For a profitable business, it is crucial to understand the herd mentality, clustering is one of the popular techniques to catch the interest of different groups of people.

**Problem Statement**

This dataset consists of tv shows and movies available on Netflix as of 2019. The dataset is collected from Flexible 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.

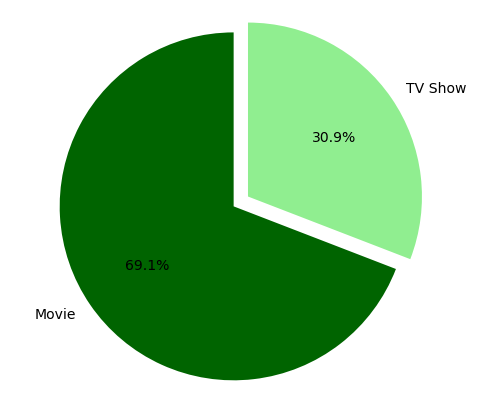
Integrating this dataset with other external datasets such as IMDB ratings, rotten tomatoes can also provide many interesting findings.

**In this project, you are required to do**

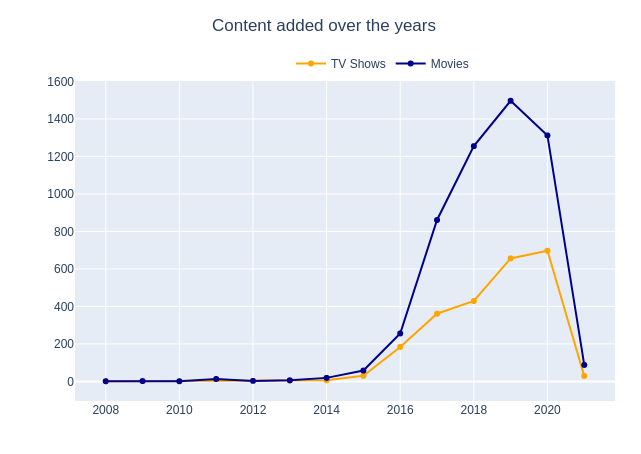
1. Exploratory Data Analysis
2. Understanding what type of content is available in different countries
3. Is Netflix increasingly focusing on TV rather than movies in recent years?
4. Clustering similar content by matching text-based features
5. **Introduction**

Nowadays ott platforms are really popular and Netflix is one of them, it is very important for Netflix to take care of its subscribers and provide according to their demands, to fulfill this demand, it is important to explore the Netflix dataset to find out peoples choice, and work according to their demand to and prevent churning.

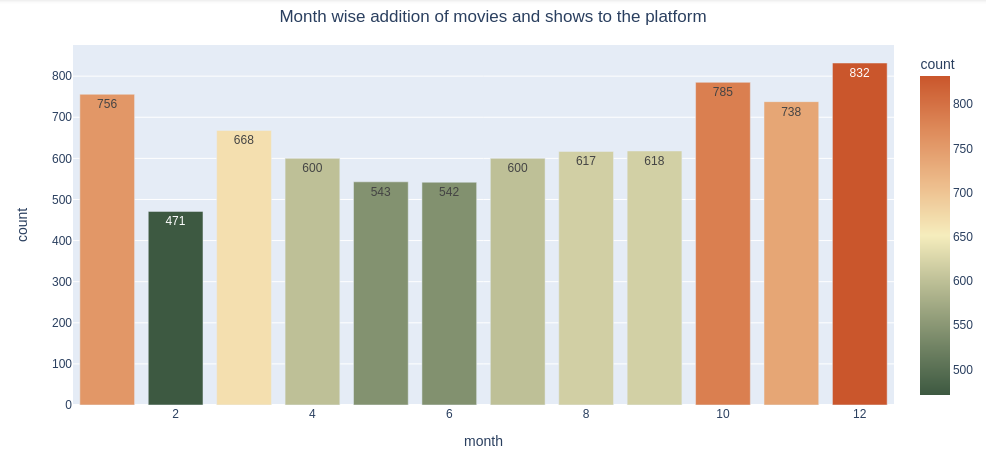
1. **Data Description**
2. show\_id : Unique ID for every Movie / Tv Show
3. type : Identifier - A Movie or TV Show
4. title : Title of the Movie / Tv Show
5. director : Director of the Movie
6. cast : Actors involved in the movie / show
7. country : Country where the movie / show was produced
8. date\_added : Date it was added on Netflix
9. release\_year : Actual Release year of the movie / show
10. rating : TV Rating of the movie / show
11. duration : Total Duration - in minutes or number of seasons
12. listed\_in : Genere
13. description: The Summary description
14. **Data Cleaning**
    1. **Checking duplicate values**
       1. No duplicate values are present in the dataset.
    2. **Handling null values**
       1. Null values present in director, cast, country, date\_added, and rating
       2. Since there are very few null values in date\_added and rating, so we will remove them from the data.
       3. We observed that the variable country had 505 null values in it, we will replace these null values with the most occurring country which is the United States.
15. **Feature Engineering**
    1. **date\_added column**
       1. We extract the day, month, and year from date\_added column
    2. **Listed\_in column**
       1. Genres are extracted and re-defined accordingly. TV and Movie genres are separately defined.
       2. Topics like International TV Shows are removed as it could bring in a bias by displaying content in reference to American movies.
16. **Exploratory Data Analysis**
    1. **Content type on Netflix**

****

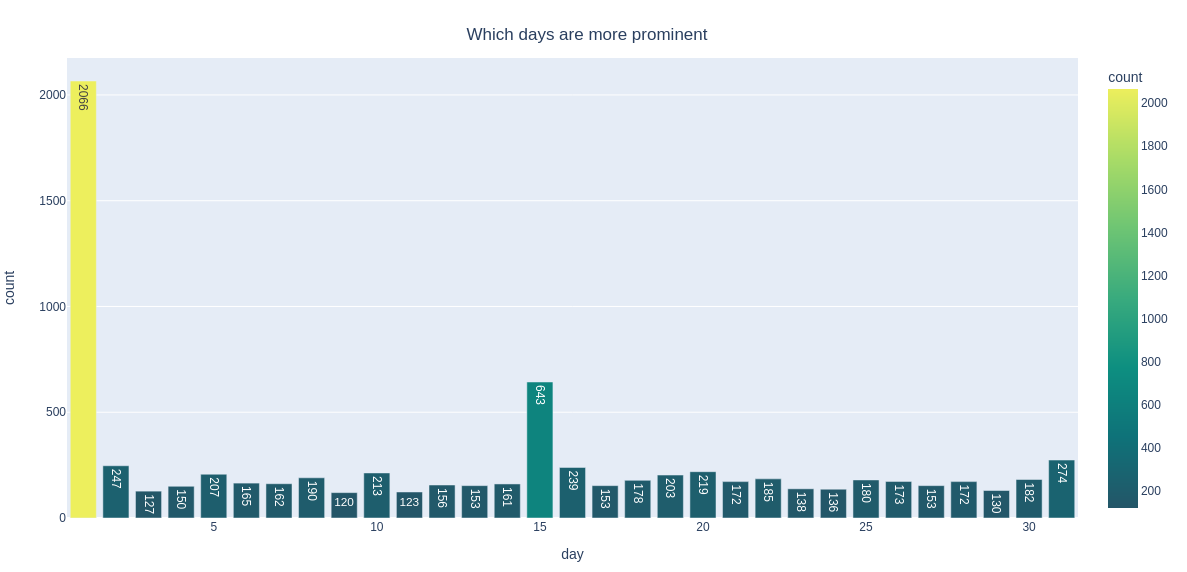
* 69.1% of the content available on Netflix is movies and the remaining 30.9% are TV Shows.
  1. **Content growth over the years**

****

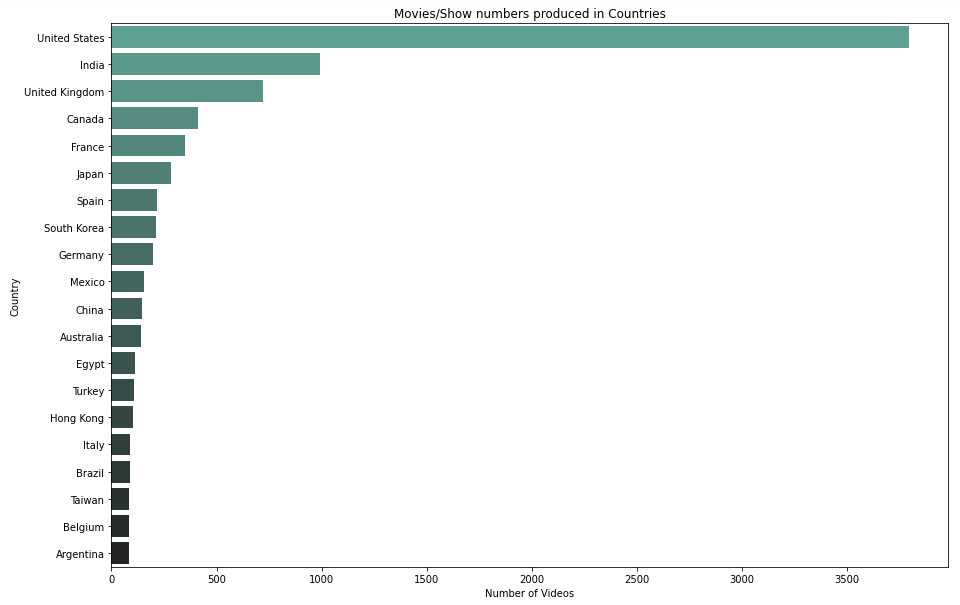
* Given Line Plot shows, growth in the number of movies on Netflix is much higher than tv shows.
* Since 2015 we can see a noticeable growth in the number of movies and tv shows uploaded by Netflix on its platform.
* The highest number of movies and tv shows got added in 2019 and 2020.
* Also, very few movies and tv shows got added in 2021.
  1. **In which month do most movies and tv shows get added?**

****

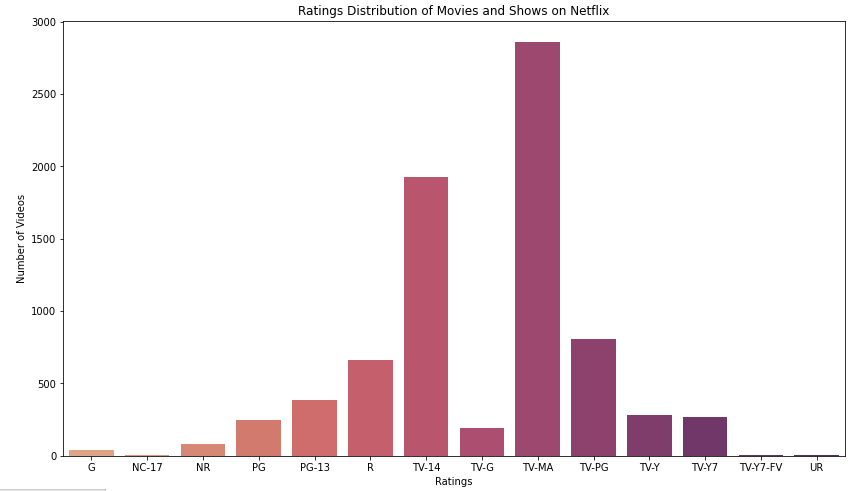
* Most of the content is uploaded either by year ending or beginning.
* October, November, December, and January are months in which many shows and movies get uploaded to the platform.
* It might be due to the winter, as in these months people may stay at home and watch shows and movies in their free time.
  1. **Which days are more prominent?**

****

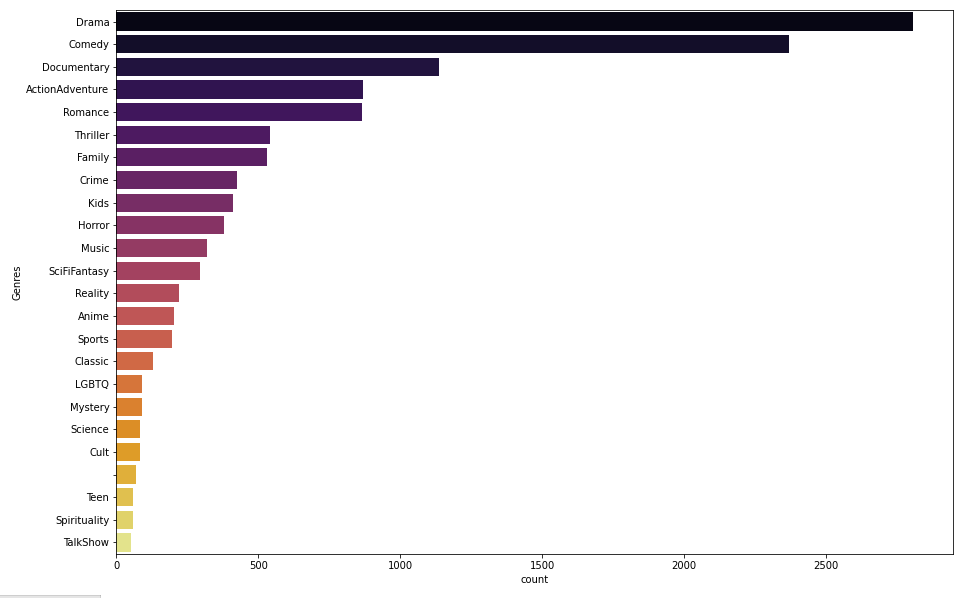
* Most of the content gets uploaded in the beginning and the middle of the month.
  1. **Top 10 Countries that produced content on Netflix**

****

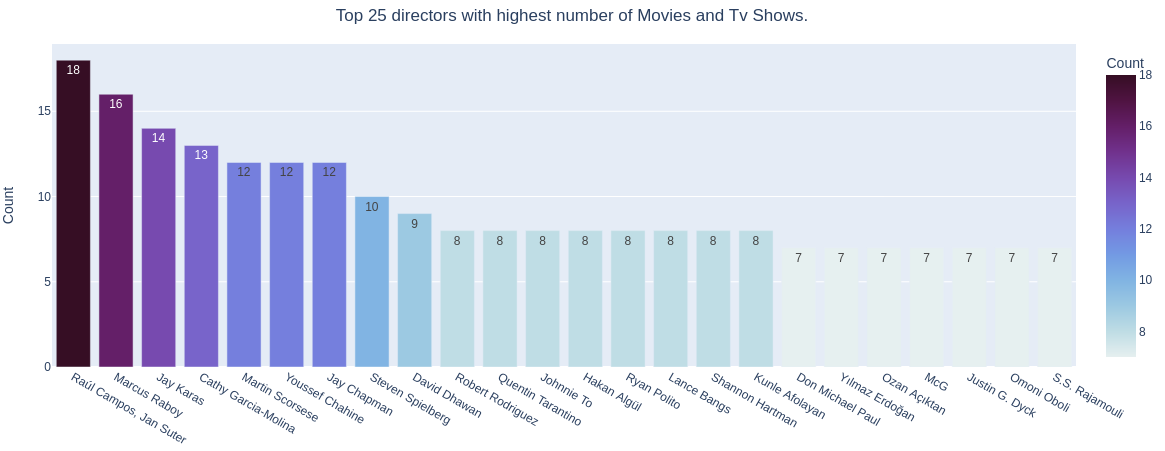
* The majority of the content providers are in the above top-ten countries.
* Among which USA, India, and UK create more than half of the tv shows and movies on the platform.
  1. **Different Ratings on the Platform**

****

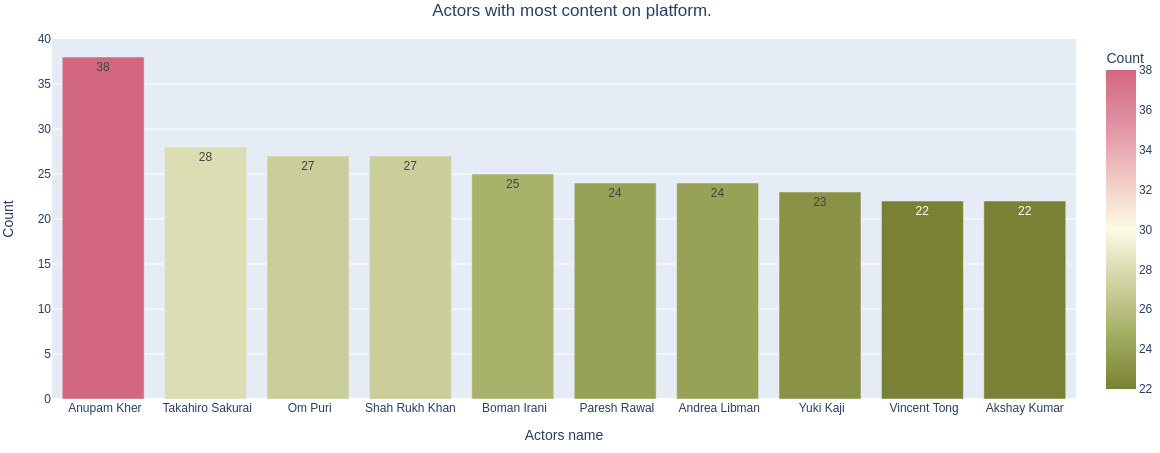
* Most content on Netflix is rated for Mature Audiences and over 14 years old
  1. **Top Genres on the Platform**

****

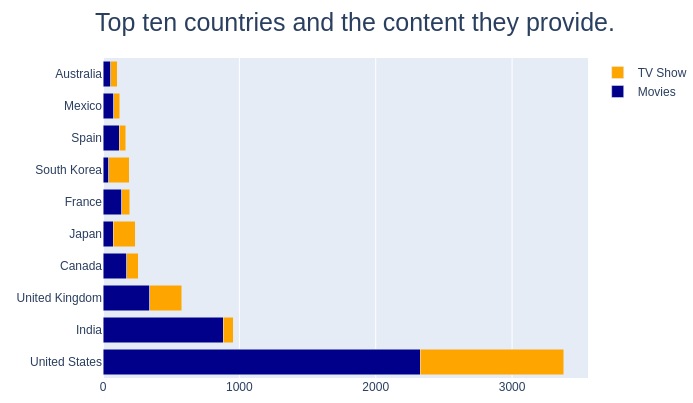
* Top Genres on Netflix are found to be: Drama, Comedy, Documentary, Action and Adventure, Romance, etc.
  1. **Which directors have the most number of movies and tv shows?**

****

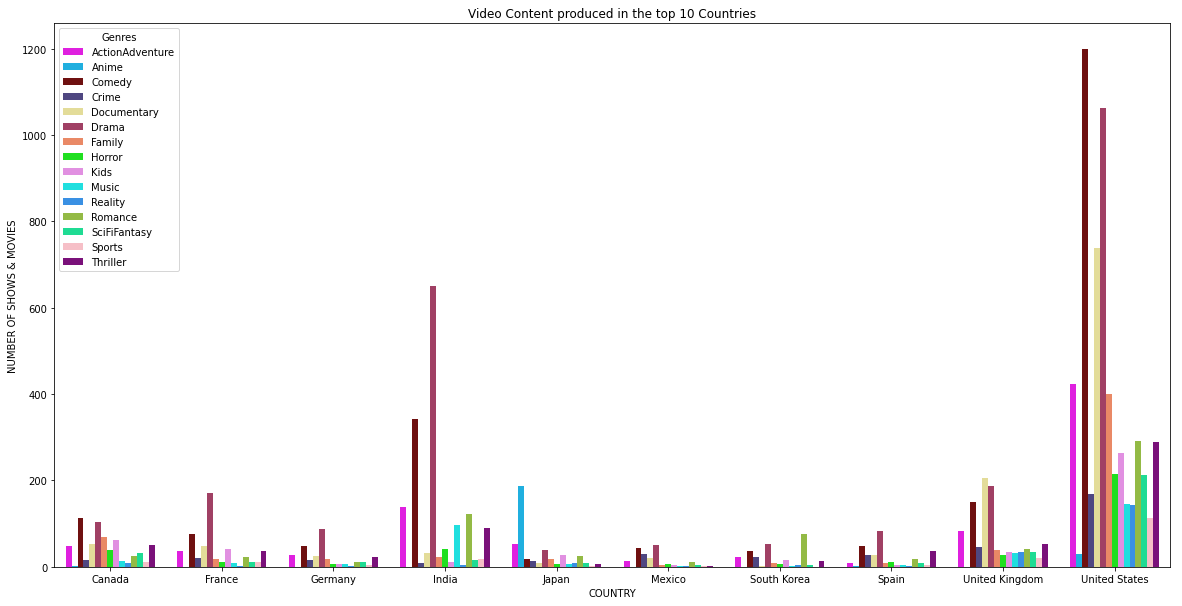
* Raúl Campos, Jan Suter, Marcus Raboy, Jay Karas, Cathy Garcia-Molina, Jay Chapman are the top 5 directors having the most numbers of movies and TV shows.
  1. **Top 10 Actors**

****

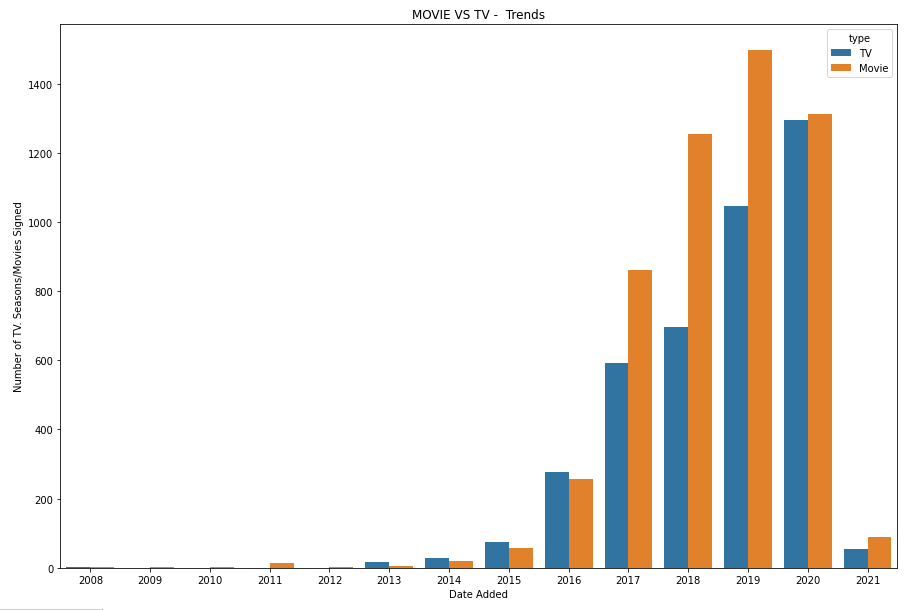
* Anupam Kher is the one who appeared in most of the movies and TV Shows.
* Also, its good to see that 6 of the actors in the top ten list of most numbers of tv shows and movies are from India.
  1. **Content available in different countries**

****

* The United States is a leading producer of both types of content which is obvious as Netflix is US Based company.
* It is followed by India where most of the content is in the form of movies.
  1. **Content produced in top 10 countries**

****

* Drama is the most produced genre in a lot of Non-English speaking countries
* Comedy is the most produced genre in English-speaking countries like the United States of America and the United Kingdom and Canada
* Japan is the biggest producer of Anime. Anime is also the most-produced genre in Japan
* Most South Korean content is from the Romance genre
* Documentaries are mainly produced in the United Kingdom and the United States of America
* Drama and Comedy are the most produced genres in the top countries with exceptions of Japan and South Korea
  1. **Is Netflix has increasingly focused on TV rather than movies in recent years.**

****

* We can observe that TV shows signed have been higher than movies in 2016.
* While the no of movies signed was higher, it can be seen that the TV shows signed per year are catching up with the movies signed year by year.

1. **Data Pre-processing**
   1. **Combining attributes**

We combined top cast, director, and listed\_in attributes

* 1. **Word tokenization**

It is the process of splitting a large sample of text into words. This is a requirement in natural language processing tasks where each word needs to be captured and subjected to further analysis like classifying and counting them for a particular sentiment etc.

* 1. **Punctuation removal**

Punctuations does not carry any meaning in clustering, so removing punctuations helps to get rid of unhelpful parts of the data, or noise.

* 1. **Stopwords removal**

Stop-words are basically a set of commonly used words in any language, not just in English. If we remove the words that are very commonly used in a given language, we can focus on the important words instead.

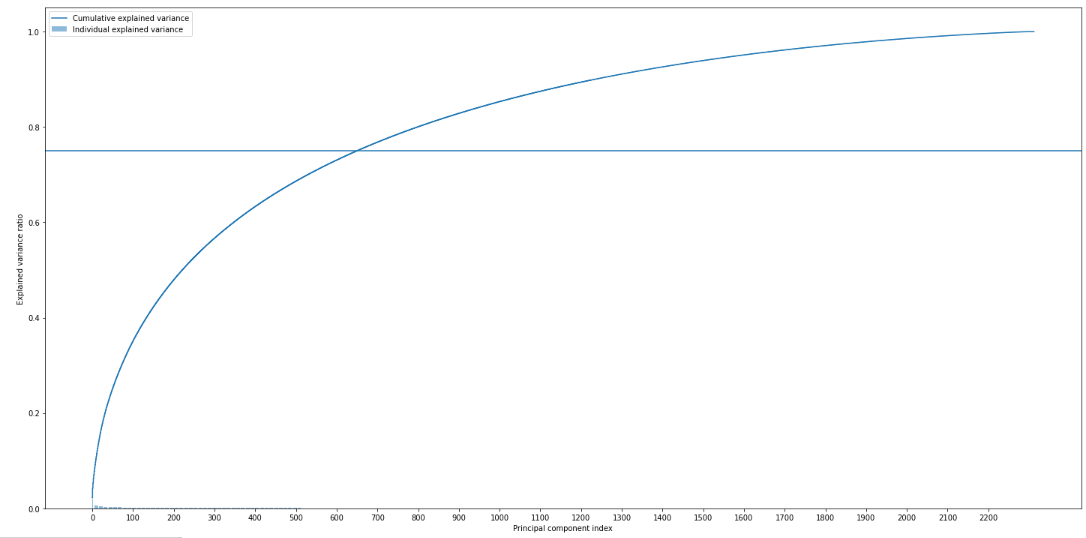
* 1. **Stemming**

Stemming is the process of removing a part of a word, or reducing a word to its stem or root. Applying stemming to reduce words to their basic form or stem, which may or may not be a legitimate word in the language.

* 1. **Text vectorization**

It is the process of converting text into numerical representation.

* 1. **Dimensionality reduction using PCA**

****

Vectorizing the preprocessed attributes Movie Dates and Description has sum total of 2121 + 192 = 2313 dimensions. These dimensions will have to be reduced using PCA, which would result in a loss of information. Alternatively, the two attributes can be used to model the content into topics using Latent Dirichlet Allocation. This would make sure that all the topical information about video content is captured without putting any available information to waste.

* 1. **Encoding attributes**
     1. We encode TV and Movies rating based on their age recommendations.
     2. older\_kids = ['TV-PG', 'PG', 'TV-Y7', 'TV-Y7-FV']

kids = ['TV-G', 'G', 'TV-Y']

teens = ['PG-13','TV-14']

adult = ['R', 'NC-17', 'TV-MA', 'UR', 'NR']

* 1. **Processing movie duration**
  2. **Binning release years based on decades**
  3. **Scaled data**

Scaling data using MinMaxscaler.

1. **Clusters Model Implementation**
   1. **KMeans Clustering**

K-means clustering is one of the simplest and most popular unsupervised machine learning algorithms. Typically, unsupervised algorithms make inferences from datasets using only input vectors without referring to known, or labeled, outcomes.

K-means algorithm works:

To process the learning data, the K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then

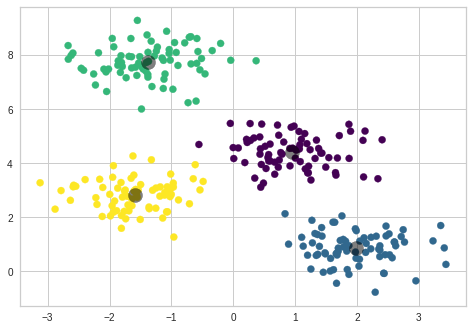
performs iterative (repetitive) calculations to optimize the positions of the centroids. It halts creating and optimizing clusters when either:

• The centroids have stabilized — there is no change in their values because the clustering has been successful.

• The defined number of iterations has been achieved.

K-means algorithm is an iterative algorithm

that tries to partition the dataset into K pre-defined distinct non overlapping subgroups where each data point belongs to only one group.

 ***Figure3.* Ideal clustering**

k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster.

We created the sample data using build blobs and used range n\_clusters to

specify the number of clusters we wanted to utilize in k means.

Silhouette score and visualization

For n\_clusters = 6 The average silhouette\_score is : 0.8241586215209683

For n\_clusters = 7 The average silhouette\_score is : 0.878767417891252

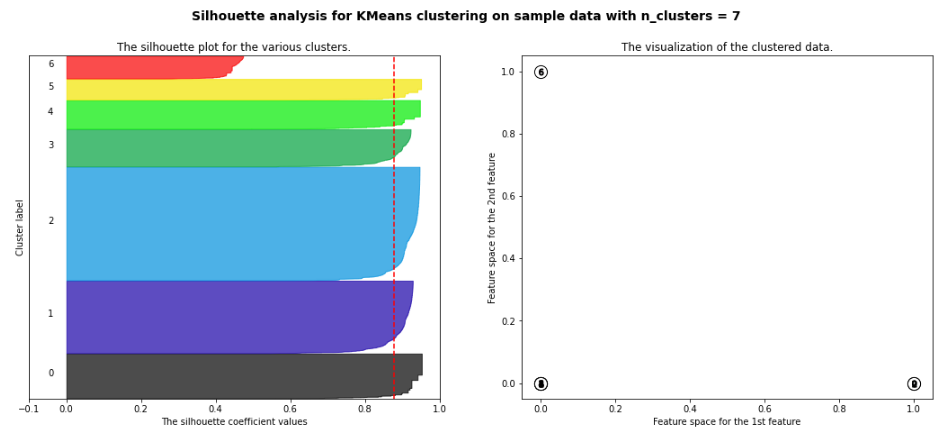
For n\_clusters = 8 The average silhouette\_score is : 0.9088984605920681

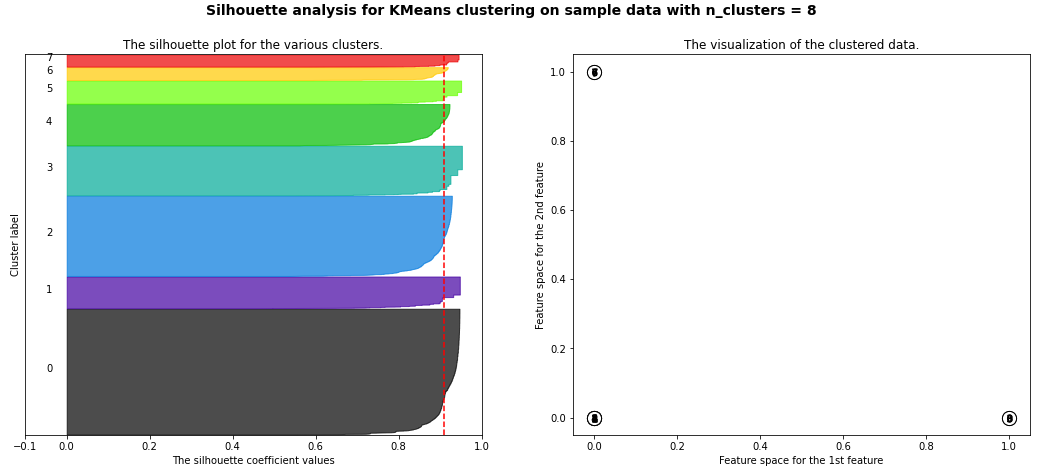
For n\_clusters = 9 The average silhouette\_score is : 0.847653554157447

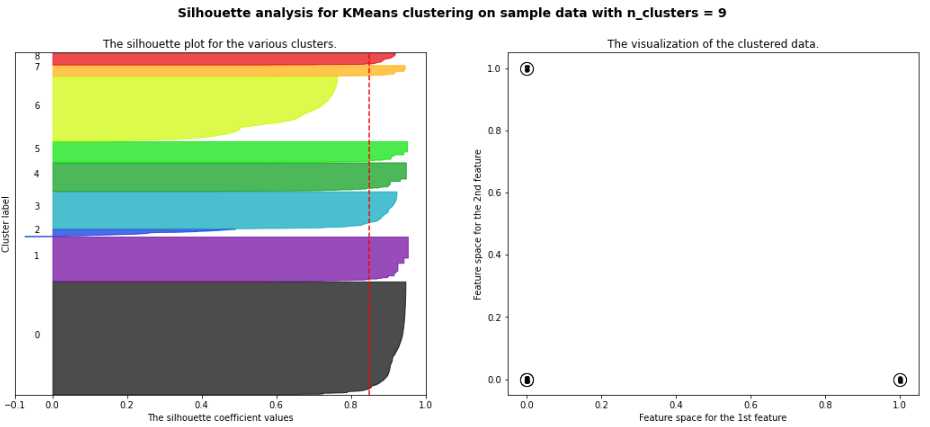
For n\_clusters = 10 The average silhouette\_score is : 0.7575834743509814

For n\_clusters = 11 The average silhouette\_score is : 0.702871382837682

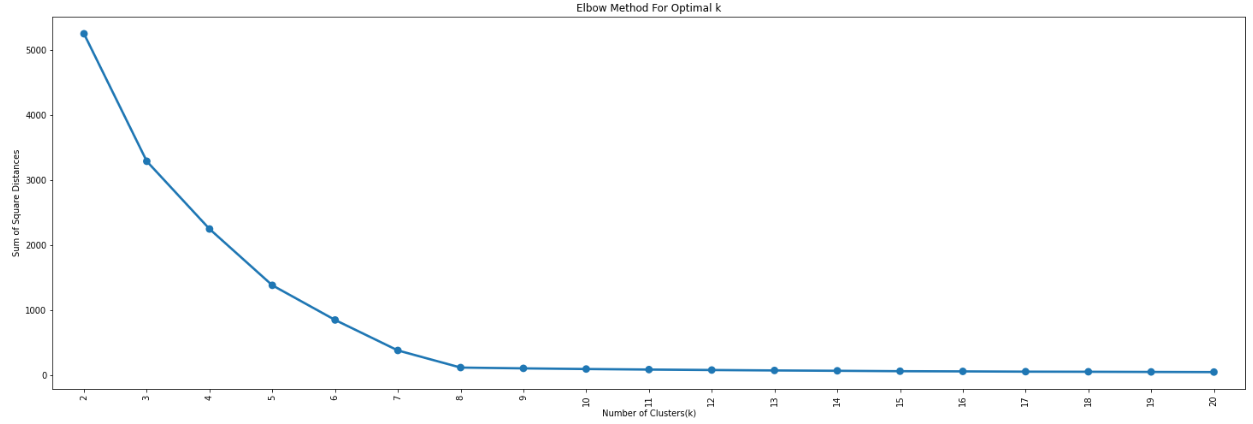
For n\_clusters = 12 The average silhouette\_score is : 0.6445708894336349





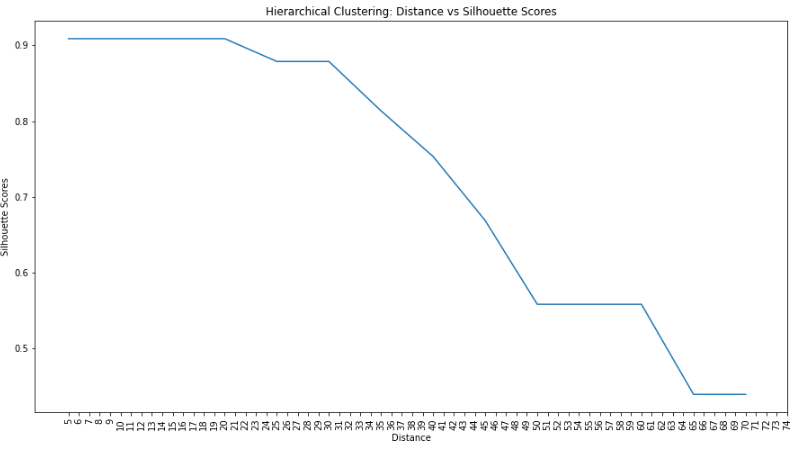


* + 1. **Silhouette Coefficient or silhouette score(meaning)**
       - Silhouette Coefficient or silhouette score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1. 1: Means clusters are well apart from each other and clearly distinguished. ... a= average intra-cluster distance i.e., the average distance between each point within a cluster.
       - If the ground truth labels are not known, the evaluation must be performed utilizing the model itself. The Silhouette Coefficient is an example of such an evaluation, where a more increased Silhouette Coefficient score correlates to a model with better-defined clusters. The Silhouette Coefficient is determined for each sample and comprised of two scores
       - Mean distance between the observation and all other data points in the same cluster. This distance can also be called a mean intra-cluster distance. The mean distance is denoted by a.
       - Mean distance between the observation and all other data points of the next nearest cluster. This distance can also be called a mean nearest-cluster distance. The mean distance is denoted by b.
       - The Silhouette Coefficient s for a single sample is then given as
       - Silhouette score is used to evaluate the quality of clusters created using clustering algorithms such as K-Means in terms of how well samples are clustered with other samples that are similar to each other. The Silhouette score is calculated for each sample of different clusters. To calculate the Silhouette score for each observation/data point, the following distances need to be found out for each observation belonging to all the clusters:
    2. **Elbow Curve:**
       - The Elbow Curve is one of the most popular methods to determine this optimal value of k.
       - The elbow curve uses the sum of squared distance (SSE) to choose an ideal value of k based on the distance between the data points and their assigned clusters.



* 1. **Hierarchical Clustering**
     1. **Agglomerative Clustering**

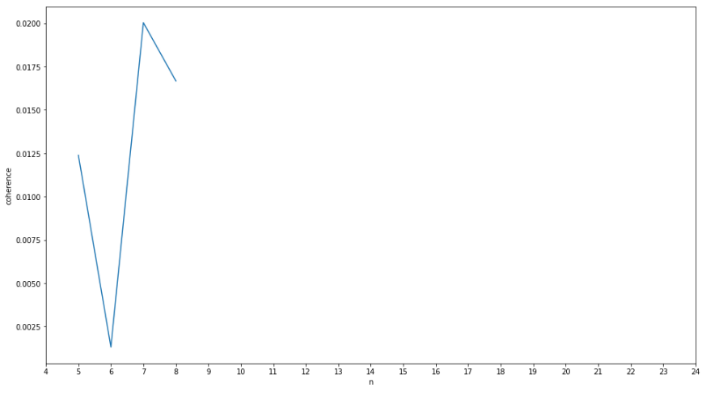
Agglomerative clustering is the most common type of hierarchical clustering used to group objects in clusters based on their similarity. ... Next, pairs of clusters are successively merged until all clusters have been merged into one big cluster containing all objects.

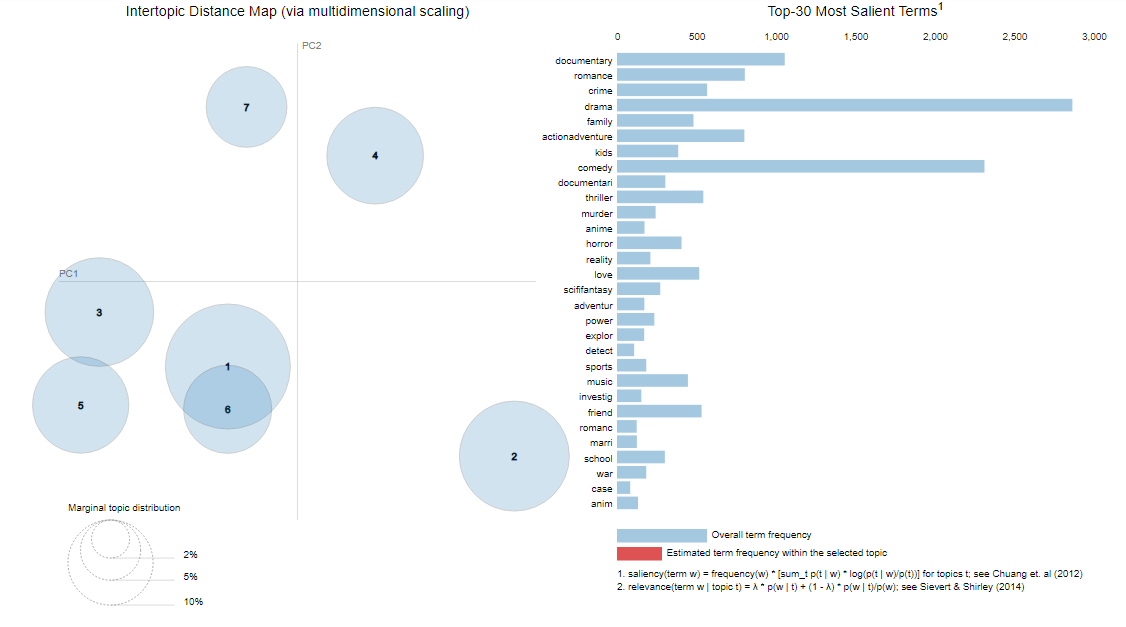


The highest Silhouette Score of 0.90 was achieved at distance = 20 with 8 clusters.

1. **Topic Modelling**
   1. **Latent Dirichlet Allocation (LDA)**

LDA is a generative probabilistic model that assumes each topic is a mixture over an underlying set of words, and each document is a mixture of over a set of topic probabilities.





1. **Conclusion – EDA**
   1. Majority of content available on Netflix is Movies.
   2. Growth in the number of movies on Netflix is much higher than tv shows. The highest number of movies and tv shows got added in 2019 and 2020
   3. October, November, December, and January are months in which many shows and movies get uploaded to the platform.
   4. Most of the content gets uploaded at the beginning and the middle of the month.
   5. The United States and India are the top countries that produce all of the available content on the platform.
   6. TV-MA tops the charts, indicating that mature content is more popular on Netflix
   7. The top Genres on Netflix are found to be: Drama, Comedy, Documentary, Action and Adventure, Romance, etc.

**Analysis of Content produced in different countries**

1. The United States is a leading producer of both types of content which is obvious as Netflix is US Based company. It is followed by India where most of the content is in the form of movies
2. Drama is the most produced genre in a lot of Non-English speaking countries
3. Comedy is the most produced genre in English-speaking countries like the United States of America, United Kingdom, and Canada
4. Drama and Comedy are the most produced genres in the top countries with the exceptions of Japan and South Korea
5. Japan is the biggest producer of Anime. Anime is also the most produced in the genre in Japan

**Is Netflix has increasingly focused on TV rather than movies in recent years.**

1. We have observed that TV shows signed have been higher than movies in 2016.
2. While the no of movies signed was higher, it can be seen that the TV shows signed per year are catching up with the movies signed year by year.
3. **Conclusion – Clustering models**
   1. k=8 is found to be an optimal value for clusters with the highest silhouette score of 0.909 using which we grouped our data into 8 distinct clusters.
   2. Using dendrograms and comparing various distance thresholds, a distance of 20 produced the highest silhouette score of 0.90 with 8 clusters producing the highest silhouette score of 0.90 with 8 clusters

**Topic Modelling**

1. We have observed that TV shows signed have been higher than movies in 2016.
2. While the no of movies signed was higher, it can be seen that the TV shows signed per year are catching up with the movies signed year by year.
3. The Topics and corresponding top words are given down below:

| **Topics** | **Corresponding Top Words** |
| --- | --- |
| Topic 1 | Comedy > Drama > Romance > Love > Life |
| Topic 2 | Documentary > Music > World > Seri > Reality |
| Topic 3 | Drama > Action and Adventure > Thriller > Young > Family |
| Topic 4 | Crime > Drama > Murder > Thriller > Comedy |
| Topic 5 | Family > Comedy > Kids > School > Friend |
| Topic 6 | Action & Adventure > Drama > Crime > Comedy > Team |
| Topic 7 | Comedy > Special > Stand > Anime > Power |

1. **Challenges**
   1. Data pre-processing was tough to tackle
   2. So many outliers are present in the dataset.
   3. Choosing the appropriate algorithm.
   4. Choosing appropriate visualization techniques was difficult.
2. **References**
   1. Python Pandas Documentation ( <https://pandas.pydata.org/pandas-docs/stable>)
   2. Python Matplotlib Documentation (<https://matplotlib.org/stable/index.html>)
   3. Python Seaborn Documentation (<https://seaborn.pydata.org/>)
   4. SK Learn Documentation (<https://scikit-learn.org/>)
   5. Articles of medium.com
   6. Our EDA Project Documentation