**Netflix Movies and TV Shows Clustering**

**Pankaj B. Gadge**

**Data science trainee,**

**AlmaBetter, Bangalore**

**Abstract**

As the world’s leading Internet television network with over 160 million members in over 190 countries, members of Netflix enjoy hundreds of millions of hours of content per day, including original series, documentaries and feature films. Netflix invest heavily in machine learning to continually improve member experience and optimize the Netflix service end-to-end.

We are going to work on the Netflix movies and TV Show dataset you can find this dataset on Kaggle. Netflix is an application that keeps growing exponentially whole around the world and it is the most famous streaming platform. This dataset consists of tv shows and movies available on Netflix as of 2019. 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. The proposed experiment is based on a combination of unsupervised machine learning algorithms such as PCA, K means clustering and Hierarchical clustering. In that case clustering similar content by matching text-based features.

***Keywords: machine learning, K Means clustering, PCA, Hierarchical clustering***

1. **Problem Statement**

To explore what all other insights can be obtained from the Netflix dataset.

**Introduction**

We all watch a lot of TV shows.  
Many online streaming services offer a large number of TV shows, which are at our disposal to watch, at the price of a subscription cost. The major online streaming services across the world are Netflix, Prime Video, Hulu, and Disney+. Netflix is a popular entertainment service used by people around the world. 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.

. The models that will be introduced in this project are:

1. PCA
2. K Means Clustering
3. Hierarchical Clustering
4. **Dataset Analysis**

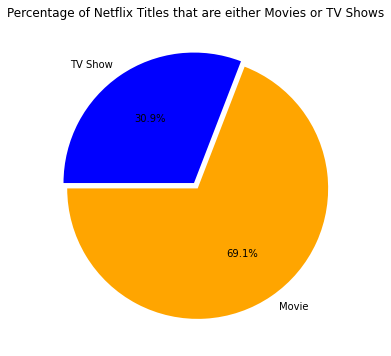
The dataset is collected from Flixable which is a third-party Netflix search engine. Integrating this dataset with other external datasets such as IMDB ratings, rotten tomatoes can also provide many interesting findings. It includes over 6,234 entries and 12 columns.

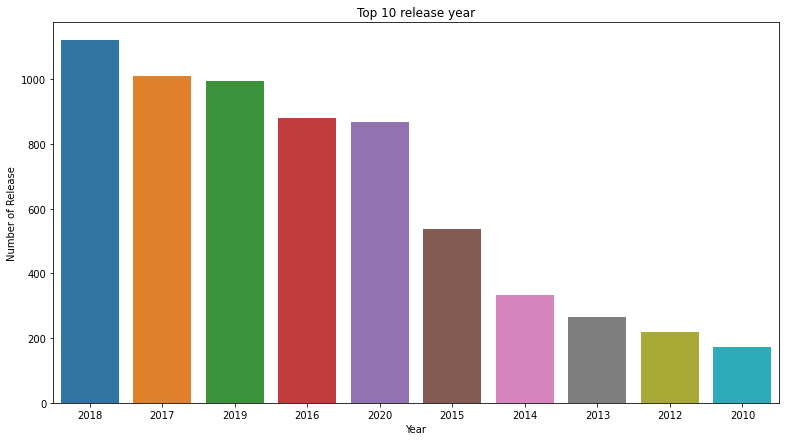
**Steps involved:**

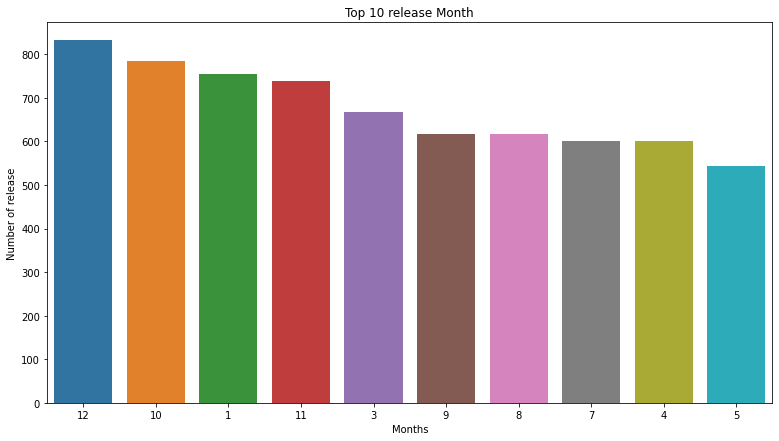
1. **Exploratory Data Analysis**

After loading the dataset, we performed this method by comparing our target variable with other independent variables. Performing an Exploratory Data Analysis allows data scientists to detect errors, debunk assumptions, and much more to ultimately select an appropriate predictive model.

This process helped us figuring out various aspects and relationships among the target and the independent variables. It gave us a better idea of which feature behaves in which manner compared to the target variable.

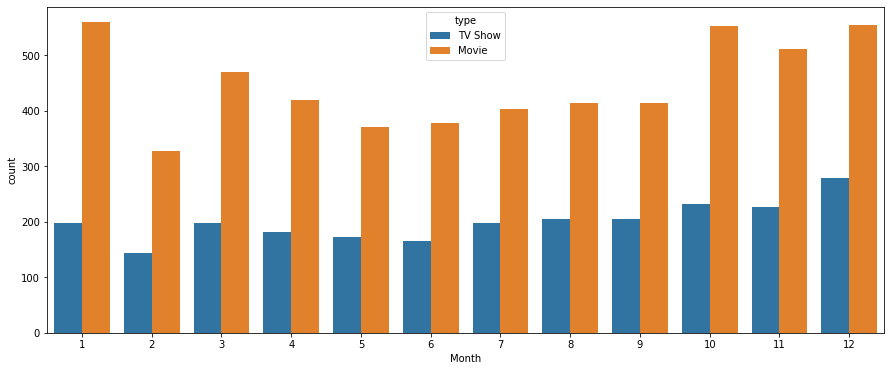
Most of the audience prefers Movies over TV Show as 69.1 percentage of audience like Movies and 30.1 percentage of audience like TV shows.



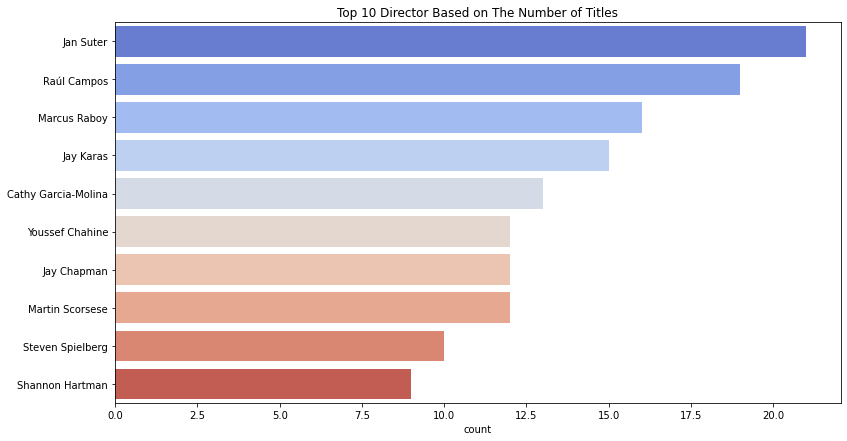


After the year 2019 covid came that badly affects Netflix for producing content. Movies have exponential growth from the start but due to covid, it is going downwards.

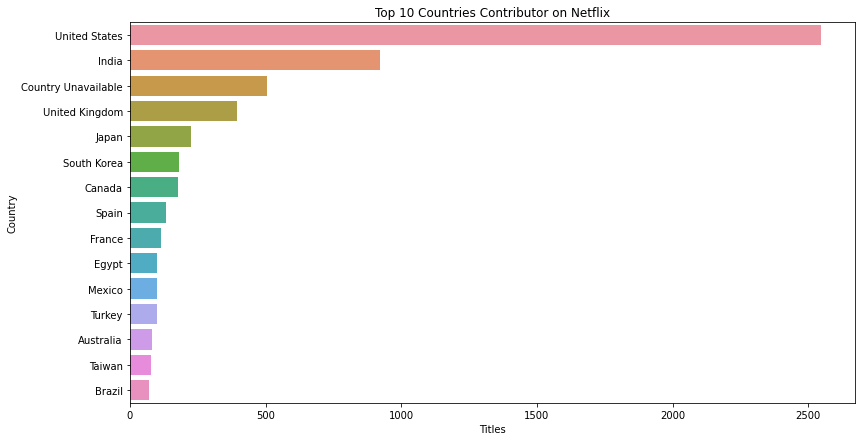
The best month to release content so the producer can gain much revenue. Most of the holidays came in December and Jan month in US and Europe. Most of the holidays came in India in Oct month. So to releases a Movie or TV show in between Oct to Jan is the best way to earn a lot of profit as the whole family will be spending time with each other and watching shows. The best 4 months to release content are October, November, December, and January.



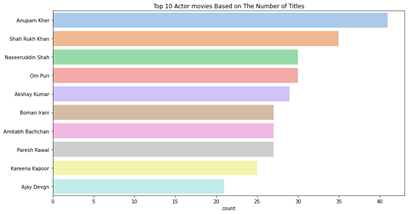
Rate of movies release is more as compared to the TV shows.

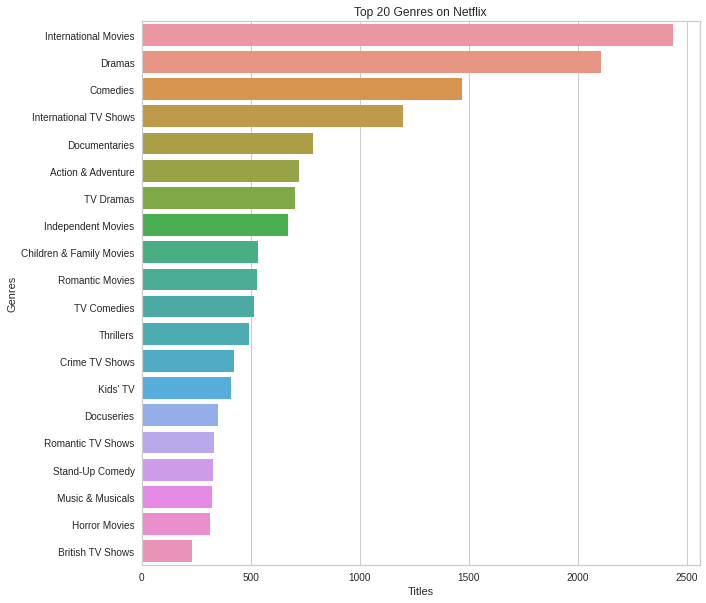


The most popular director on Netflix, with the most titles, is Jan Suter and next is Raul Compos.

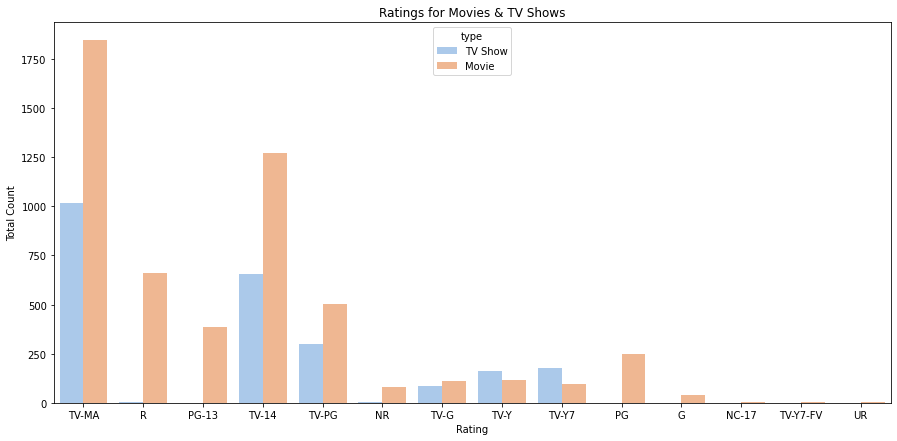


From the above bar graph, we can see the top 15 countries which contributes most on Netflix. The country by the amount of the produces content is the United States.

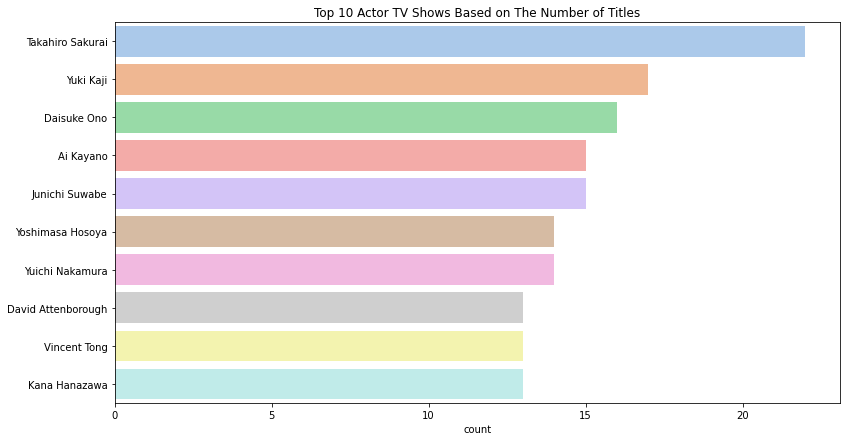




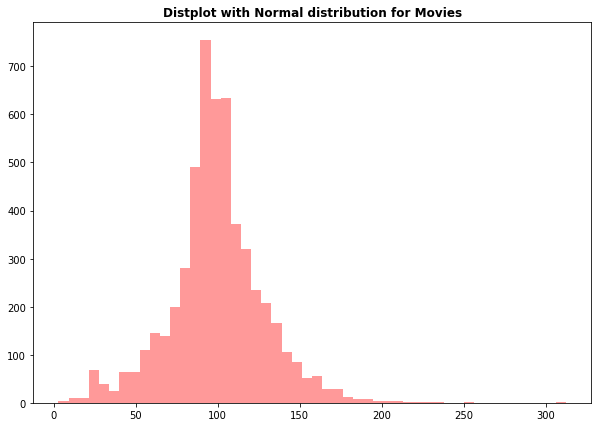
From the graph, we know that International Movies take the first place, followed by dramas and comedies.



The largest count of TV shows is made with a “TV-MA” rating. “TV-MA” is a rating assigned by the TV Parental Guidelines to a television program designed for mature audiences only.

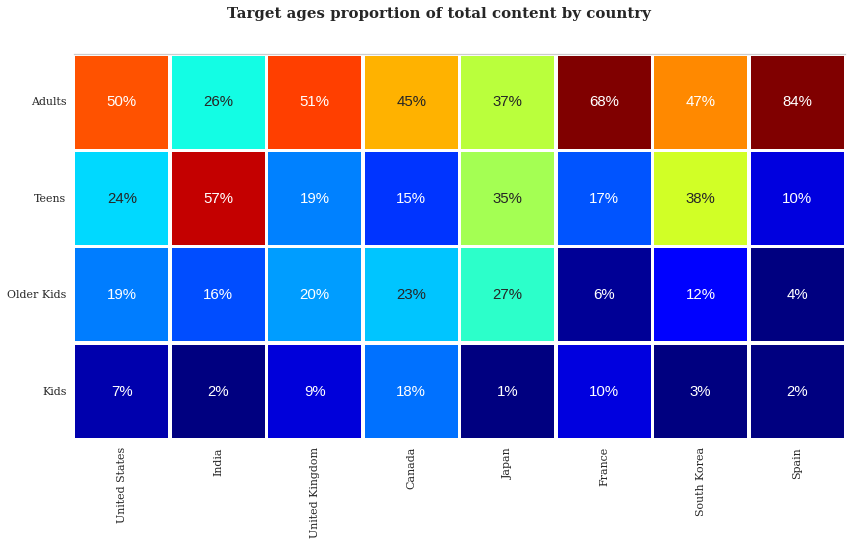


The top actor on Netflix TV Show, based on the number of titles, is Takahiro Sakurai.



Most of the movies have duration of between 50 to 150 min.





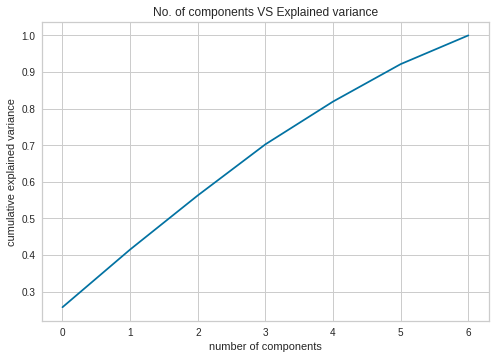
From the corr. graph it is clear that in Spain,France,UK and US mostly adults peoples like to watch moveies and TV shows.

While in India 57 percentage of teens watching movies and TV shows and only 26 percentage of adults like to watching movies and TV shows.

That means in India Netflix need to more concentrate to Teens.

1. **Algorithms:**
2. **Principal Components Analysis (PCA)**

PCA is fundamentally a dimensionality reduction algorithm, but it can also be useful as a tool for visualization, for noise filtering, for feature extraction and engineering, and much more.



Typically, we want the explained variance to be between 95–99%.

In this case, to get 99%(most approximate is 100%)of variance explained I need 7 principal components.

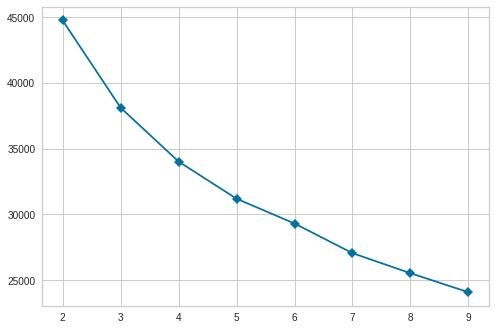
1. **K means Clustering**

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process.

# **Find out how many clusters are used**

1. **K Elbow Method**

The Elbow Method is one of the most popular methods to determine this optimal value of k.

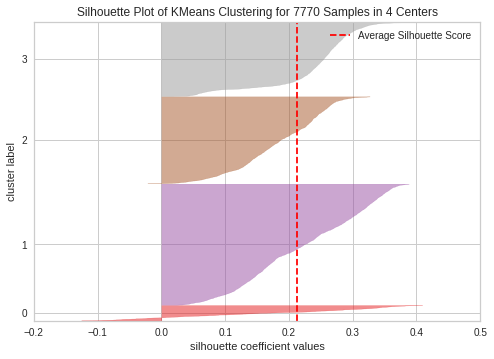


To determine the optimal number of clusters, we have to select the value of k at the “elbow” ie the point after which the distortion/inertia start decreasing in a linear fashion. Thus for the given data, we conclude that the optimal number of clusters for the data is 4.

1. **Silhouette Score**

The silhouette plot displays a measure of how close each point in one cluster is to points in the neighbouring clusters and thus provides a way to assess parameters like number of clusters visually. This measure has a range of [-1, 1].

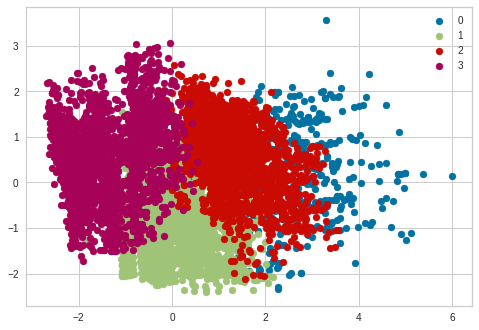
Silhouette coefficients (as these values are referred to as) near +1 indicate that the sample is far away from the neighbouring clusters. A value of 0 indicates that the sample is on or very close to the decision boundary between two neighbouring clusters and negative values indicate that those samples might have been assigned to the wrong cluster.



For n\_clusters = 4, silhouette score is 0.21208415191245705

Here is the Silhouette analysis done on the above plots to select an optimal value for n\_clusters.

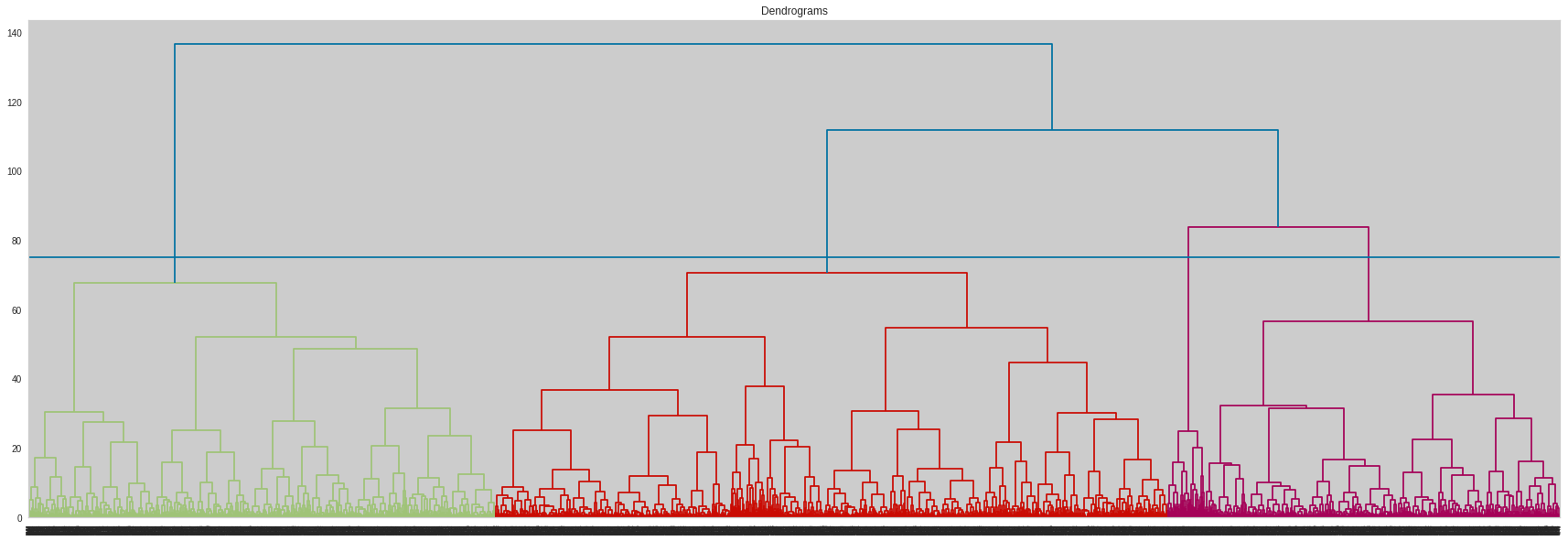
The value of 4 for n\_clusters looks to be the optimal one. The silhouette score for each cluster is above average silhouette scores



1. **Hierarchical clustering**

Hierarchical clustering is another unsupervised machine learning algorithm, which is used to group the unlabeled datasets into a cluster and also known as hierarchical cluster analysis or HCA.

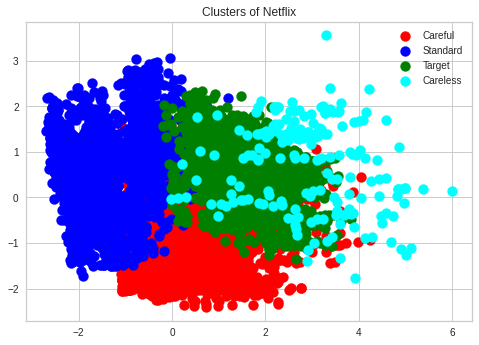
In this algorithm, we develop the hierarchy of clusters in the form of a tree, and this tree-shaped structure is known as the dendrogram.



Assume we cut vertical lines with a horizontal line to obtain the number of clusters. Number of clusters = 4.

1. **Agglomerative Clustering**

The agglomerative hierarchical clustering algorithm is a popular example of HCA. To group the datasets into clusters, it follows the bottom-up approach. It means, this algorithm considers each dataset as a single cluster at the beginning, and then start combining the closest pair of clusters together. It does this until all the clusters are merged into a single cluster that contains all the datasets.



**Conclusion:**

We have drawn many interesting inferences from the dataset Netflix titles; here’s a summary of the few of them:

* The most content type on Netflix is movies.It appears that Netflix has focused more attention on increasing Movie content than TV Shows. Movies have increased much more dramatically than TV shows
* There are about 70% movies and 30% TV shows on Netflix.
* Most films were released in the years 2018, 2019, and 2020.
* The number of releases have significantly increased after 2015 and have dropped in 2021 because of Covid 19.
* The months of October, November, December and January had the largest number of films and television series released.
* More of the content is released in holiday season - October, November, December and January.
* The United States has the highest number of content on Netflix by a huge margin followed by India.
* Raul Campos and Jan Sulter collectively have directed the most content on Netflix.
* Anupam Kher has acted in the highest number of films on Netflix. Drama is the most popular genre followed by comedy.
* International movies are the top most genre in netflix which is fllowed by standup comedy and Drams.
* Most of the movies have duration of between 50 to 150
* Highest number of tv\_shows consistig of single season
* Using correlation heatmap we see that in India mostly teens watching netflix so question arrises that what content teens watched.
* TV-MA has the highest number of ratings for tv shows i,e adult ratings
* In India teens mostly watched international movies.
* Principal Component analysis (PCA)reduced the number of componets as 7 with approximately 99% of variance.
* For K Means clustering to find out number of k we used elbow and sillhoute score method.
* Using both the methods we found k=4 is optimal value of clustering.
* Using Hirarchical clustering method again we find out that k=4 is optimal value of clustering.

**References:**

1. MachineLearningMastery
2. GeeksforGeeks
3. Analytics Vidhya