**Netflix Movies and TV Shows Clustering**

**Ritik Vaidande**

**Data Science Practitioner**

**AlmaBetter, Bangalore**

**Abstract:**

As movie industry is evolving into streaming platforms, there’s no doubt that Netflix has become one of the important platforms for streaming. The dataset that we have used for EDA and clustering has been collected by Flixable, a third-party Netflix search engine. There are 12 features and around 7700 observations in the dataset and are mostly textual features

**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 thirdparty 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.

**Dataset Insight**

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 Releaseyear 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 : Genere
12. description: The Summary description

**Steps involved**

* **Cleaning the Dataset**

-There are 2389 null values in Director column.

-There are 718 null values in cast column.

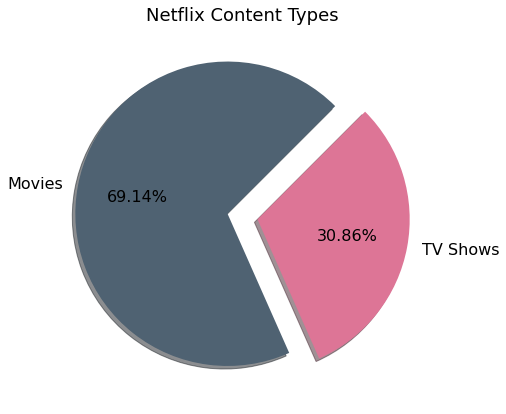
-There are 507 null values in country column.

-There are 10 null values in date added column.

-There are 7 null values in rating column

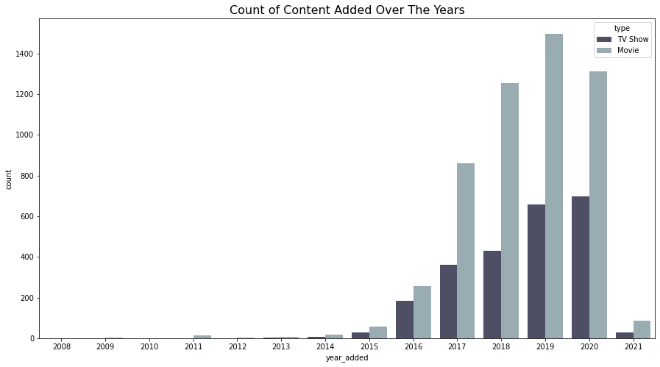
* **Exploratory Data Analysis**

1. **Content analysis:**

Majority of content available on Netflix are Movies. 

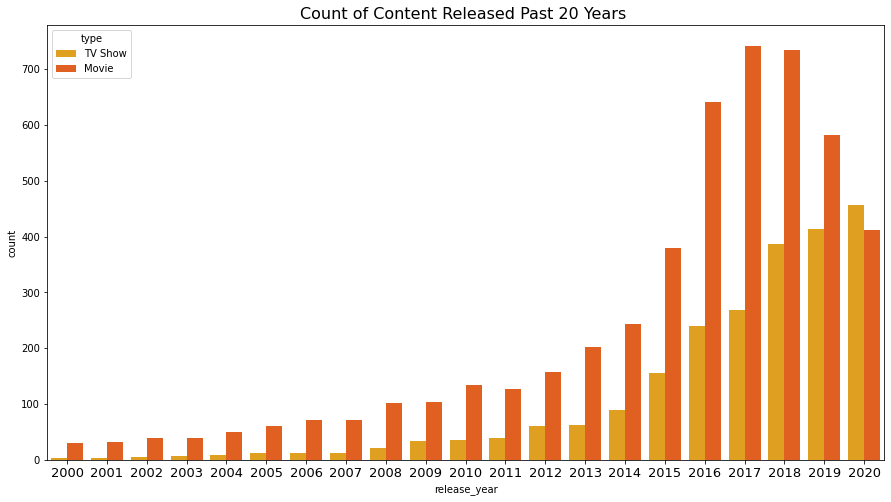
Netflix content type

**2. Content added analysis:** Large number of TV Shows and Movie got added in year 2019 and 2020. Limited amount of data is available for the year 2021



Content added over the years

**3. Release analysis:** Immense amount of TV Shows and Movies were released over past 5 years. TV Shows and Movies are following an consistent ratio among them over the years.

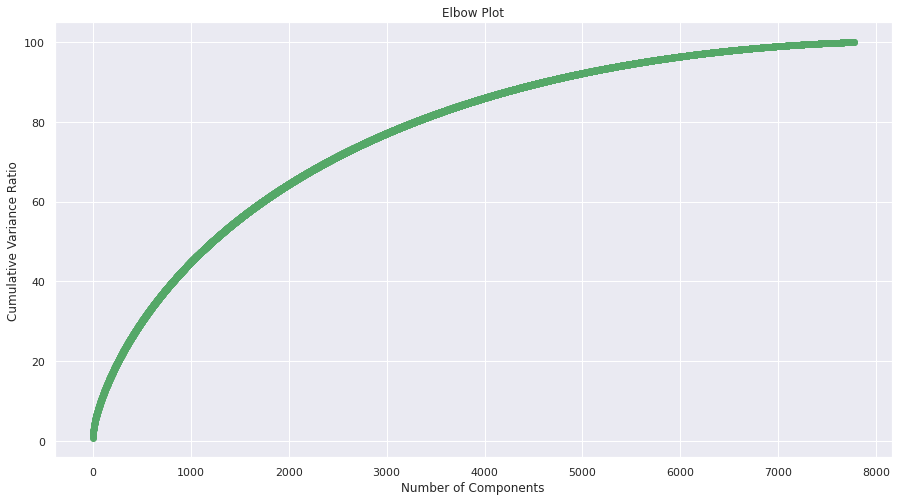


Content released over the years

* **Modelling**

**Feature Preparation**

PCA is an Dimensionality Reduction Technique. It is also a Feature extraction Technique. By PCA we create new features from old (Original) Features but the new features will always be independent of each other. So, its not just Dimensionality Reduction Process, we are even eliminating Correlation between the Variables.



At 6k components our PCA model seems to go Flat without explaining much of a Variance.

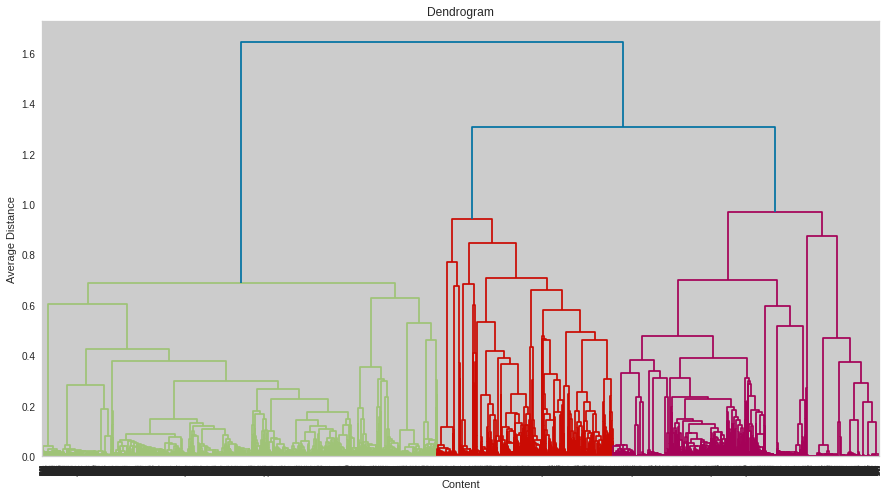
**Clustering**

Clustering is the task of dividing the data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and

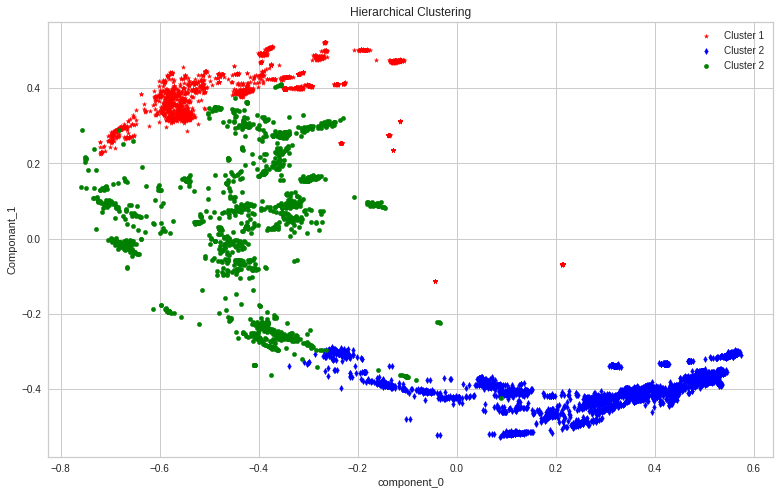
dissimilar to the data points in other groups.

It is basically a collection of objects on the basis of similarity and dissimilarity between them.

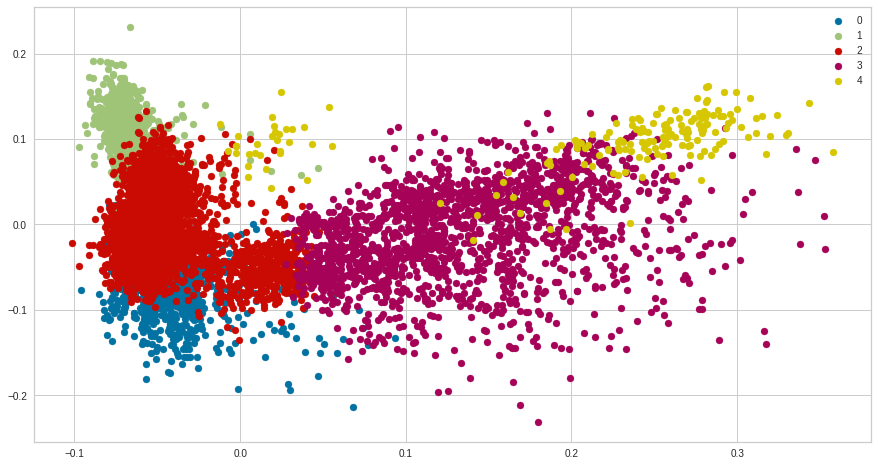
Hierarchical Clustering: Hierarchical clustering separates data into groups based on some measure of similarity, finding a way to measure how they’re alike and different, and further narrowing down the data.



Dendrogram shows that 3 clusters would be suitable for the clustering the data by hierarchical clustering.



K-means Clustering: K-means Clustering Algorithm is the simplest unsupervised learning algorithm that solves clustering problem. It partitions n observations into k clusters where each observation belongs to the cluster with the nearest mean serving as a prototype of the cluster.



Total 5 distinct clusters are created by using K-means Clustering Algorithm.

Data Represented by Each Cluster :

0 - Documentaries

1 - Family and Children Movies

2 - International TV Shows

3 - International Movies and Dramas

4 - Comedy Shows

**Conclusion**

-Majority of content available on Netflix are movies.

-Most of the TV Shows and Movies are added in the month October, November, December and January.

-United States and India are the highest content producing countries.

-Large number of content are for mature audiences.

-Over past 5 years immense amount of TV Shows and Movies were released.

-High percentage of TV-MA rating shows that Mature Content is more popular on Netflix.

-TV Shows rarely go above 5 seasons and average time of a movie is around 90 to 120 minutes.

-It was found that the optimal number of clusters was 5. Therefore total 5 distinct clusters were created using K-means Clustering Algorithm.

-Documentaries, Family and Children Movies, International TV Shows, International Movies and Drama, Comedy Shows are the data represented in the clusters.

**References**

1. Towards Data Science
2. StackOverflow
3. Medium
4. GitHub