**Netflix Movie and TV Shows Clustering**

**Praful Gedam**

**Data Science Trainee,**

**Alma Better, Bangalore**

**Abstract**

Netflix is one of the leading over-the-top platform because of its reputation

For offering users a variety of high quality streaming movies as well as TV shows.

The reason why Netflix service are so popular worldwide is that the company uses recent technologies like machine learning, deep learning and artificial intelligence to provide consumers with more appropriate imputative recommendation. This Paper is based on unsupervised clustering Analysis On Netflix movies and TV shows dataset Aim of the project is to form the cluster is based on k mean clustering, Agglomerative clustering and Affinity Propagation

Clustering. We have done data Processing, Text cleaning, Exploratory Data Analysis, Vectorization, implementing clustering Models, Hyper parameter tuning. Data Analysis, understanding what type content is available in different countries,

**Introduction**

It’s fascinating how Netflix applies AI/Data science/ML to running its operation, such as by implementing algorithm to provide movie reconditions and using AI to guarantee high quality streaming even at reduced with bandwidths. The following are some of the numerous applications od airdate Science, and machine learning at Netflix. Improvement in Netflix’s AI integration has made widespread indivilization possible. Simply said, the AI engine keeps an eye on the flow of information.

**Problem statement: -**

This dataset consists of TV shows and movies available on Netflix as of 2019. The dataset is collected from Fixable 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, we are evaluating as below-**

* Exploratory Data Analysis
* Understanding what type content is available in different countries
* Is Netflix increasingly focused on TV rather than movies in recent years?
* Clustering similar content by matching text-based features

**Exploratory Data Analysis**

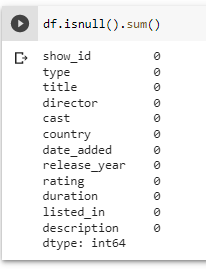
The first step involved in the analysis is to load the dataset into the pandas’ data frame. Before exploring the data using different libraries available in python we should if the dataset is ready to run the operations on it.

**Data Exploration: -** Data Exploration one of the important steps before we starting the models

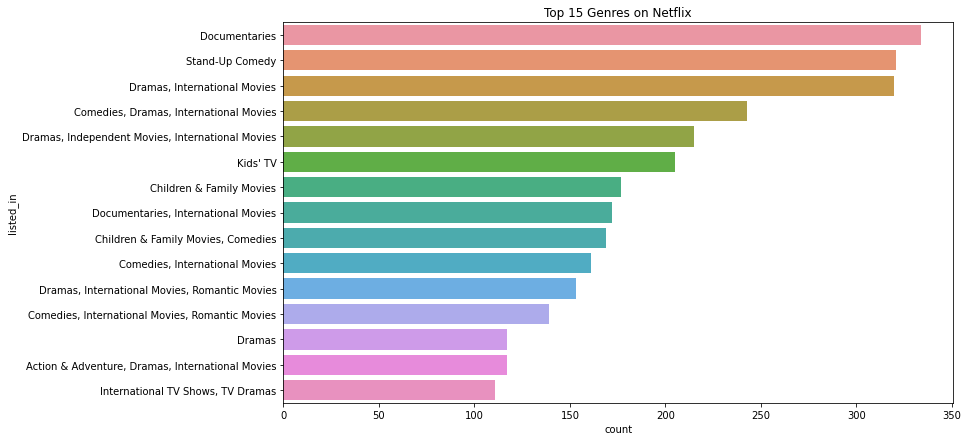
* There are 2389 null values in Director column
* There are 718 null values in cast column
* There are 507 null values in country column
* There are 10 null values in date added column
* There are 7 null values in rating column

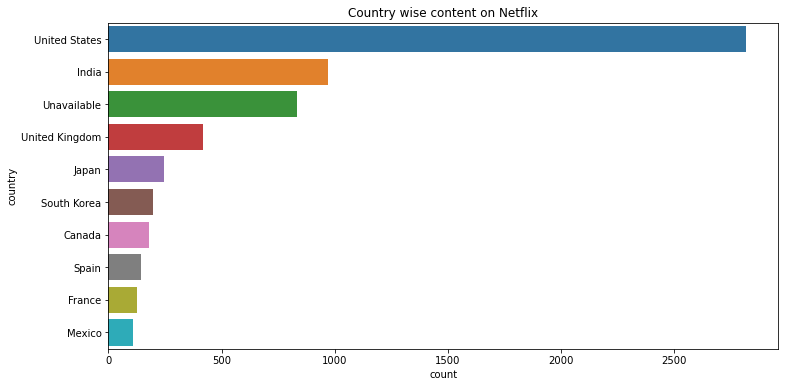
**Handling missing values: -**

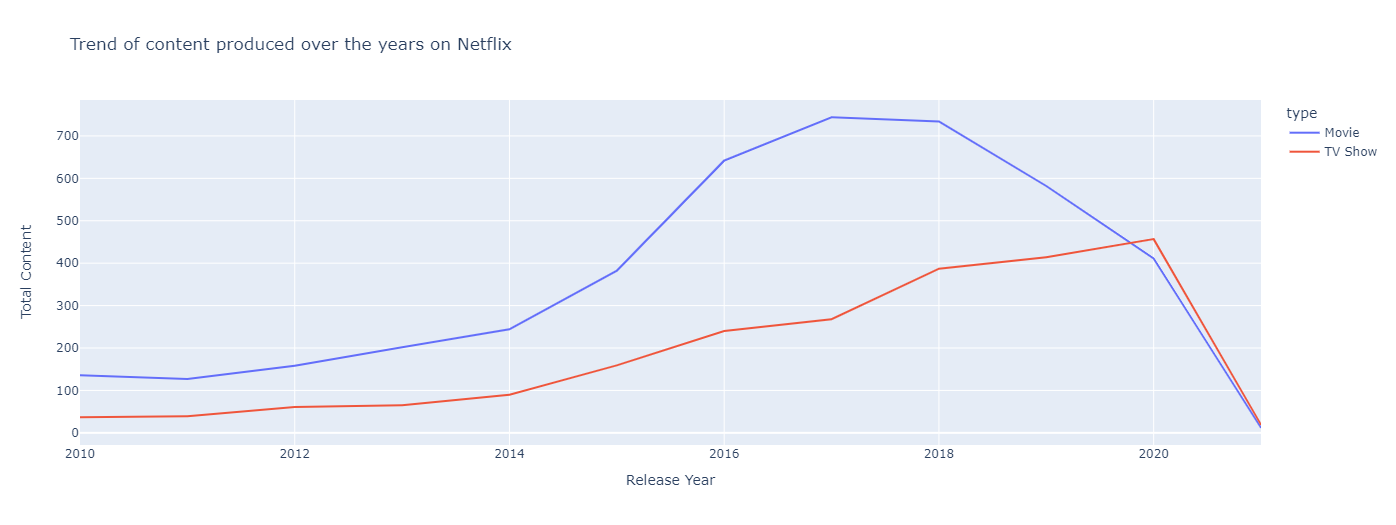
There are no missing values present in the Dataset

****

**Methodology: -**

In this section, we will discuss the framework, extraction and preprocessing features, feature selection, and clustering algorithms.



**Is Netflix increasingly focused on TV rather than Movies in recent years?**

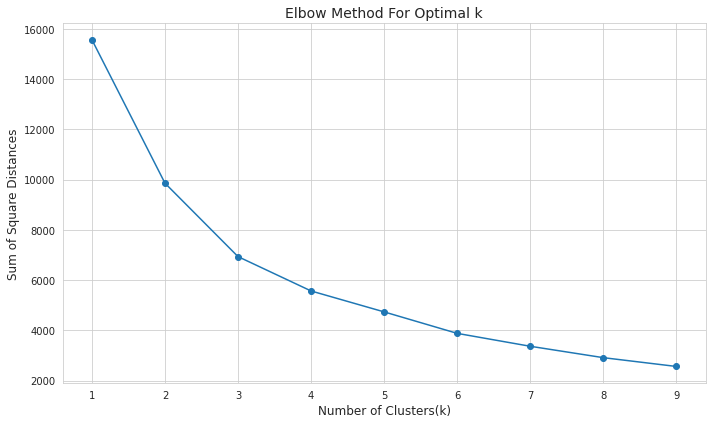
Yes, Netflix is increasingly focusing on TV Shows now, which is clear from the graph, in 2020, there were more Shows than Movies. Also, Movies preference shows a declining graph, while shows are increasing.

**Clustering: -**

1. **K means clustering: -**

K mean clustering is a method of vector quantization, originally from signal processing that aims to partitions aims observation into k cluster in which each observation into k cluster with the nearest means, serving as the prototype of the cluster.

We created the sample data using build blob and used range n cluster to specify the number of cluster



Here we will see the output of 2,3,4 and 5 number of clusters

#### **Silhouette score:**

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 observations belonging to all the clusters.

|  |  |  |  |
| --- | --- | --- | --- |
| For n\_clusters = 2, | silhouette | score | is |
| 0.3551415129065328 |  |  |  |
| For n\_clusters = 3, | silhouette | score | is |
| 0.35586172779109915 |  |  |  |
| For n\_clusters = 4, | silhouette | score | is |
| 0.32858920525532515 |  |  |  |
| For n\_clusters = 5, | silhouette | score | is |
| 0.3348785202036102 |  |  |  |

So in this model the 3 clusters are giving best result.

So that we will consider 3 clusters as optimum clusters

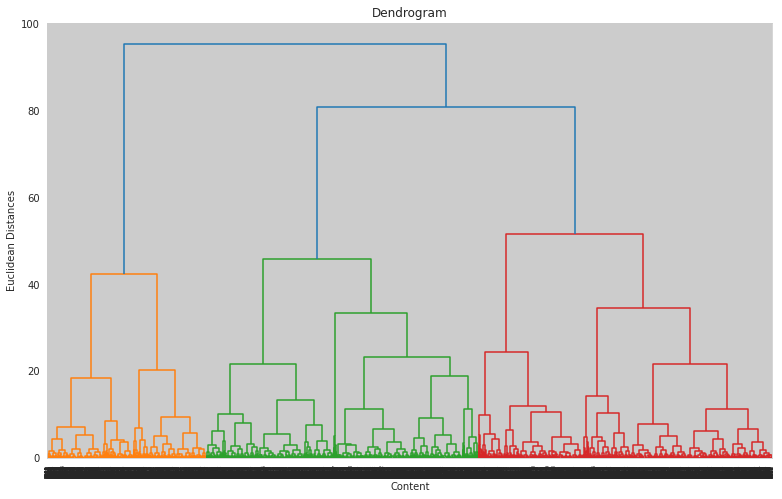
We have also done an interactive clustering visualization in the notebook.

#### **Hierarchical clustering**

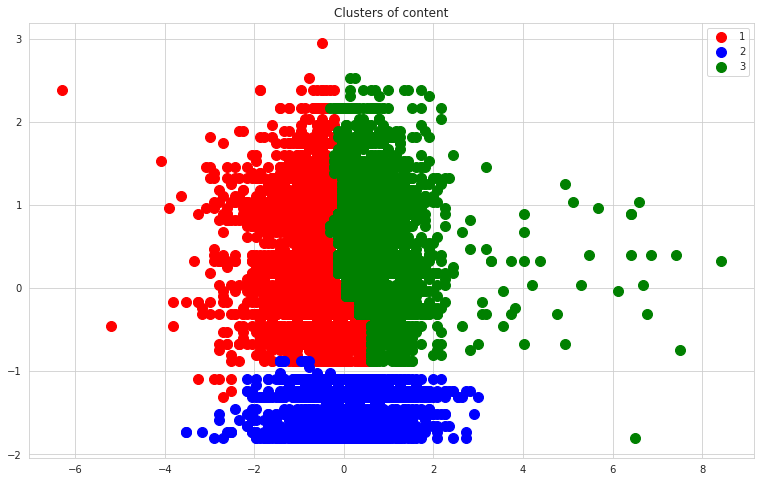
In data mining and statistics, hierarchical clustering (also called hierarchical cluster analysis or HCA) is a method of cluster analysis which seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types:

* + Agglomerative: This is a "bottom-up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
  + Divisive: This is a "top-down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

In general, the merges and splits are determined in a greedy manner. The results of hierarchical clustering are usually presented in a dendrogram. As shown in following figure:



#### **Agglomerative Clustering**

The agglomerative clustering is the most common type of hierarchical clustering used to group object 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 object

**Conclusion: -**

* We started by removing null values and converting the Netflix added date to year, month and day using Proper date times format
* Most films are released in the year 2018,2019,2022
* The months of October, November, December and January had the largest number of films and television series released.
* TV shows account for 2.8 percent of the total, while movies account for

97.2 percent.

* The United States, India, the United Kingdom, Canada, and Egypt are the top five producer countries.
* Netflix has added lot of more movies and TV episode in the previous years, but the number are still low when compared to movies released in last 10 years.
* Dramas International movies are the most popular genres on Netflix
* Affinity Propagation, Agglomerative Clustering, and K-means Clustering were utilized to build the model
* In Affinity Propagation, we had 9 clusters and a Silhouette Coefficient score of 0.340
* A dendrogram was used to determine the number of clusters in Agglomerative Clustering. There were two clusters, with an average silhouette coefficient of 0.5695
* The final model we used was k- means clustering, which consisted of 2,3,4,5,6 clusters. 3 numbers of clusters gives us good fitting
* After clustering, we can say that the number of TV shows launched in the previous years is NOT growing.

# **Reference: -**

1. [https://scikitlearn.org/stable/modules/generated/sklearn.featur e\_extraction.text.CountVectorizer.html](https://scikitlearn.org/stable/modules/generated/sklearn.featur%20e_extraction.text.CountVectorizer.html)
2. <https://help.netflix.com/en/node/2064>
3. [https://scikitlearn.org/stable/modules/generated/sklearn.cluste r.AffinityPropagation.html](https://scikitlearn.org/stable/modules/generated/sklearn.cluste%20r.AffinityPropagation.html)